

HUMAN RESOURCE ANALYTICS

Dr. Rekha N. Patil
Ms. Shilpa S. Hiregoudar

- BASICS OF BUSINESS ANALYTICS
- RISE OF HUMAN RESOURCE (HR) ANALYTICS
- APPLICATIONS OF HR METRICS AND CREATING HR DASHBOARDS
- CORRELATION AND REGRESSION FOR HR ANALYTICS
- HR ANALYTICS APPLICATIONS USING T-TEST AND ANOVA
- HR ANALYTICS APPLICATIONS USING REGRESSION

Himalaya Publishing House

ISO 9001:2015 CERTIFIED

HUMAN RESOURCE ANALYTICS

Dr. Rekha N. Patil
Assistant Professor,
Department of MBA,
VTU-RRC, Kusnoor Road,
Kalaburagi.

Smt. Shilpa S. Hiregoudar
Assistant Professor
Department of MBA,
Godutai Engineering College for Women,
Sharnbasva University,
Kalaburagi.



Himalaya Publishing House

ISO 9001:2015 CERTIFIED

© **Authors**

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording and/or otherwise without the prior written permission of the publisher.

FIRST EDITION : 2023

Published by	: Mrs. Meena Pandey for HIMALAYA PUBLISHING HOUSE PVT. LTD. , “Ramdoot”, Dr. Bhalerao Marg, Girgaon, Mumbai - 400 004. Phone: 022-23860170, 23863863; Fax: 022-23877178 E-mail: himpub@bharatmail.co.in; Website: www.himpub.com
Branch Offices	:
New Delhi	: “Pooja Apartments”, 4-B, Murari Lal Street, Ansari Road, Darya Ganj, New Delhi - 110 002. Phone: 011-23270392, 23278631; Fax: 011-23256286
Nagpur	: Kundanlal Chandak Industrial Estate, Ghat Road, Nagpur - 440 018. Phone: 0712-2721215, 2721216
Bengaluru	: Plot No. 91-33, 2nd Main Road, Seshadripuram, Behind Nataraja Theatre, Bengaluru - 560 020. Phone: 080-41138821; Mobile: 09379847017, 09379847005
Hyderabad	: No. 3-4-184, Lingampally, Besides Raghavendra Swamy Matham, Kachiguda, Hyderabad - 500 027. Phone: 040-27560041, 27550139
Chennai	: No. 34/44, Motilal Street, T. Nagar, Chennai - 600 017. Mobile: 09380460419
Pune	: “Laksha” Apartment, First Floor, No. 527, Mehunpura, Shaniwarpet (Near Prabhat Theatre), Pune - 411 030. Phone: 020-24496323, 24496333; Mobile: 09370579333
Cuttack	: Plot No. 5F-755/4, Sector-9, CDA Market Nagar, Cuttack - 753 014, Odisha. Mobile: 09338746007
Kolkata	: 3, S.M. Bose Road, Near Gate No. 5, Agarpara Railway Station, North 24 Parganas, West Bengal - 700109. Mobile: 9674536325
DTP by	: Sudhakar Shetty (On behalf of HPH Pvt. Ltd.)
Printed at	: M/s. Giriraj Printers, Hyderabad. (On behalf of HPH Pvt. Ltd.)

PREFACE

We decided to write this book because we saw need of the new syllabus and updated knowledge to the student and faculty. Human Resource Analytical is an essential part of any company's ability to maintain success and ensure the achievement of standards within an organization. As the world economy experiences significant improvements in managing changes due to rapid growth of the technology, it has become prior important to know and understand in depth knowledge about HR Analytical for HR Professional. This book helps to gain basic knowledge and advanced tools in HR Analytical, and the tools that will help them make their HR Analytical practices more efficient.

The book highlights Human Resource Analytical concept and in-depth knowledge. It consists of the following chapters:

Chapter 1. Basic of Business Analytics

Chapter 2. Rise of Human Resources Analytics

Chapter 3. Applications of HR Metrics and Creating HR Dashboards

Chapter 4. Correlation and Regression for HR Analytics

Chapter 5. HR Analytics Applications using t-test and ANOVA

Chapter 6. HR Analytics Applications using Regression

The composition of the handbook is as per the prescribed syllabus of VTU, Bangalore University and Private University. It shall also prove valuable to HR Analytical for in-depth references and study for enhancing their conceptual understanding.

We assure you this book will help a lot to student and teaching fraternity to understand and gain knowledge regarding HR analytical concept and even it will help the HR practitioners to understand practical application of HR analytical concept.

We thank all student and teaching fraternity, and specially thank the family members and Himalaya Publishing House Pvt. Ltd.

Any suggestion regarding improvement in book content or errors, if any, will be gratefully acknowledged.

AUTHORS

DETAILED CONTENTS

MODULE 1: BASICS OF BUSINESS ANALYTICS

1 – 23

- 1.1 Introduction
- 1.2 Evolution of Business Analytics
- 1.3 The Need or Motivation for Studying Business Analytics
- 1.4 Major Challenges in Business Analytics
- 1.5 Use Cases of Business Analytics
- 1.6 Scope of Business Analytics
- 1.7 Tools of Business Analytics
- 1.8 Business Analytical Process
- 1.9 Advantages of Business Analytics
- 1.10 Levels of Analytical Maturity
- 1.11 Emergence of Business Analytics
- 1.12 Analytics in Different Domain of Business Applications
- 1.13 Applications of Analytics in Different Fields
- 1.14 Managing Business Analytics Projects

MODULE 2: RISE OF HUMAN RESOURCE (HR) ANALYTICS

24 – 79

- 2.1 History of HR Analytics
- 2.2 Definition and Meaning of HR Analytics
- 2.3 Role of Analytics
- 2.4 Four Levels of Analytics
- 2.5 Purpose of HR Analytics
- 2.6 Pitfalls of People Analytics
- 2.7 Areas of HR Analytics
- 2.8 Applying HR Analytics
- 2.9 Tools for HR Analytics
- 2.10 Benefits of HR Analytics
- 2.11 Uses of HR Analytics
- 2.12 Challenges of HR analytics
- 2.13 Three Waves to Creating Lasting Value

- 2.14 HR Measurement Journey in Tune with HR Maturity Journey
- 2.15 The Five Lean Principles Explained
- 2.16 HR Decision-making
- 2.17 Importance and Significance of Human Resource Analytics
- 2.18 Critical HR Decision-making and HR Analytics
- 2.19 HR Analytics and Changing Role of HR Managers
- 2.20 Conducting HR Analytics
- 2.21 The Basics of Text Analytics for HR Analytics
- 2.22 Future of HR Analytics
- 2.23 Understanding Future Human Resources
- 2.24 Predictive Analytics for HR in Action

MODULE 3: APPLICATIONS OF HR METRICS AND CREATING HR DASHBOARDS

80 – 215

- 3.1 History of HR Analytics
- 3.2 Definition and Meaning of HR Analytics
- 3.3 Metrics for Training and Development Function
- 3.4 Major Categories of Human Resources Metrics
- 3.5 HR Metrics and Types of Data
- 3.6 HR Dashboards
- 3.7 HR Analytics as a Better Tool for HR Decisions
- 3.8 Statistical Software Used for HR Analytics
- 3.9 Data Visualisation Tools
- 3.10 Staff or Skills Matrix
- 3.11 Developer TAB
- 3.12 Index
- 3.13 Form
- 3.14 Excel Dashboard
- 3.15 Storyboard
- 3.16 HR Scorecard
- 3.17 Insight into Data-driven HRA
- 3.18 Cleaning Data (Exercise)
- 3.19 Basics of HR Analytics

- 3.20 Evolution of HR Analytics
- 3.21 Analytic Value Chain
- 3.22 Analytical Model
- 3.23 Practical Application of HR Analytics
- 3.24 Key Performance Indicators in Action
- 3.25 HR Analytics Tools

MODULE 4: CORRELATION AND REGRESSION FOR HR ANALYTICS 216 – 288

- 4.1 Correlation Analysis
- 4.2 Methods of Computing Correlation
- 4.3 The Case of Outlier
- 4.4 Software for Statistical Analysis 1 – GNU PSPP
- 4.5 Software for Statistical Analysis 2 – R and R Commander
- 4.6 Regression Analysis
- 4.7 Cause-and-effect Relationship
- 4.8 Multiple Regression Analysis
- 4.9 Interaction Effects

MODULE 5: HR ANALYTICS APPLICATIONS USING T-TEST AND ANOVA 289 – 342

- 5.1 Hypothesis
- 5.2 Hypothesis Testing
- 5.3 Important Terms
- 5.4 Types of Hypothesis Testing
- 5.5 One-sample t-test
- 5.6 Paired Sample t-test
- 5.7 Independent SAMPLE t-test
- 5.8 ANOVA (Analysis of Variance)

MODULE 6: HR ANALYTICS APPLICATIONS USING REGRESSION 343 – 429

- 6.1 What is HR Analytics?
- 6.2 Regression Analysis
- 6.3 Causation

- 6.4 Logistic Regression
- 6.5 Multiple Predictors
- 6.6 Factor Analysis
- 6.7 Cluster Analysis
- 6.8 Tasks in Data Mining
- 6.9 Outlier Analysis
- 6.10 Clustering in *R*

CASE STUDIES

430 – 432

LIST OF CASES

Case Study 1: A Roadmap towards Predictive HR Analytics	430
Case Study 2: Using HR Analytics to Support Managerial Decisions: A Case Study	430
Case Study 3: Walmart – HR Analytics	431
Case Study 4: Net Income on a Spreadsheet	431
Case Study 5: HR Metrics	431

LIST OF TABLES AND FIGURES

Table 4.1: Degree of Correlation	219
Table 4.2: Calculation of Coefficient of Correlation between the Age of Husband and Wives	227
Table 4.4: Difference between Correlation and Regression Analysis	252
Table 4.5: Mileage and Groove Depth of a Car Tire	267
Table 4.6: Comparison between Simple and Multiple Regressions	278
Table 5.1: Hypothesis Decision Table	291
Table 5.2: Difference between Large and Small Sample Tests	298
Table 5.3: Calculation for Sample Mean and SD	304
Table 5.4: $\alpha = 2.5\%$ Level of Significance	308
Fig. 1.1: Business Analytical Process	8
Fig. 2.1: People Analytics is the Intersection of People, Strategy, Science, Statistics, and Systems	30
Fig. 2.2: The Five Lean Principles	51
Fig. 2.3: Seven Steps in the HR Analytics Cycle	57
Fig. 3.1: Participant Data	85
Fig. 3.2: Framework for Ethics of Data and AI	88
Fig. 3.3: Ethics in People Analytics: Four Dimensions of Trust	89
Fig. 3.4: Continuum of Human Capital Analytics	90
Fig. 3.5: An Example of Descriptive Statistics Output Generated by SAS	96
Fig. 3.6: R Programming	99
Fig. 3.7: RStudio	100
Fig. 3.8: Tableau	103
Fig. 3.9: Example of a Skills Matrix	106
Fig. 3.10: A Competencies Matrix	109
Fig. 3.11: Skill Assessment	111
Fig. 3.12: Employees' Skill Comparison by Sum of Skill Weights	112
Fig. 3.13: Department's Average Skill Weight	113
Fig. 3.14: Skills Matrix Template	114

Fig. 3.15(a): How to Add the Developer Tab to the Excel Ribbon?	115
Fig. 3.15(b): How to Add the Developer Tab to the Excel Ribbon?	116
Fig. 3.15(c): How to Add the Developer Tab to the Excel Ribbon?	117
Fig. 3.16: Index Structure	119
Fig. 3.17: B-tree structure of an index	120
Fig. 3.18: Clustered and Non-Clustered Index	121
Fig. 3.19: Tree Structure of Index	123
Fig. 3.20: Index Seek (Non-Clustered)	126
Fig. 3.21: Data Form	128
Fig. 3.22: Summary of Form Controls	129-131
Fig. 3.23: Summary of ActiveX Controls	132-134
Fig. 3.24: Data Required by HR Analytics Tools	180
Fig. 3.25: Data Cleaning Checklist	195
Fig. 4.1: Types of Correlation	217
Fig. 4.2: Methods of Computing Correlation	220
Fig. 4.3: Scatter Diagram	220
Fig. 4.4(a): Perfect Positive Correlation	221
Fig. 4.4(a): Perfect Positive Correlation	221
Fig. 4.6: Calculation of Karl Pearson's Coefficient of Correlation	223
Fig. 4.7: PSPPIRE Data Editor	235
Fig. 4.8: Regression Lines	253
Fig. 4.9(a): Two-dimensional Scattered Diagram	278
Fig. 4.9(b): Three-dimensional Scatter Diagram	279
Fig. 6.1: Linear Regression	347
Fig. 6.2: Regression with Outlier	349
Fig. 6.3: Sigmoid Function	390
Fig. 6.4: Clustering of a Set of Objects Based on the k -Means Method	399
Fig. 6.5: Four Cases of the Cost Function for k -Medoids Clustering	400
Fig. 6.6: Applications of Clustering in R	411
Fig. 6.7: Types of Clustering Algorithms in R	412

MODULE 1

BASICS OF BUSINESS ANALYTICS

Learning Objectives:

1. Acquire knowledge regarding basic concepts of Business Analytics.
2. Helps to learn how to manage a Business Analytics Project.
3. Understand the various Business Analytics concepts to solve day-to-day corporate problems.
4. Study the challenges of Business Analysts.
5. Understand different levels of Analytics Maturity.

1.1 INTRODUCTION

Business Analytics provides the models and procedures to BI (Business Intelligence). It also involves tracking data and then analyzing them for competitive advantages. Business analysts started to get employed to assist managers and to take on some analytics roles, especially to make reports. The tools and techniques of industrial engineering and quality control, statistics and operations research, were developed and used by a diverse set of practitioners who provided advice to organizations, Business analytics practitioners treat this as their professional lineage.

The functions of planning, decision-making, providing direction, motivating, monitoring and control are part of managers' job. Business analytics makes extensive use of data, statistical and quantitative analysis, explanatory and predictive modeling, and fact-based management to drive decision-making. It is therefore closely related to management science. Analytics may be used as input for human decisions or may drive fully automated decisions.

1.2 EVOLUTION OF BUSINESS ANALYTICS

BA in the 1800s – The Need to Stay Ahead

The first use of data to stay ahead of his competitor's dates back to 1865. During this time, Mr. Richard Miller Devens described in his book how Sir Henry Furnese, a banker, was always one step ahead by actively gathering information and acting on it before any of his competitors.

This makes it clear that professionals such as Sir Furnese relied more on data and empirical evidence, rather than gut instinct.

BA in the Late 1800s – The Advent of Scientific Management

During this time, Frederick Taylor introduced the first-ever system of business analytics in the United States of America, and he called it scientific management. The purpose of this system was to analyze the production techniques and labourers body movements to identify greater efficiencies.

BA in the Early 1900s – The Transformation of the Manufacturing Industry

Frederick Taylor's scientific management system inspired Henry Ford, who hired Taylor as his consultant. Ford was willing to measure the time each component of his Ford Model T took to complete on his assembly line. This analysis transformed his work and the manufacturing industry across the globe.

BA in the 1950s – The First Hard Drive Disk by IBM

Computers were not accessible in the early 1900s but had a massive demand during World War-II. As they were still rudimentary, punch cards or tapes were used to store information. However, in 1956, the tech giant, IBM invented the first hard disk drive. This allowed users to save a vast amount of data with better flexibility.

BA in the Late 1900s – The Emergence of Business Intelligence

Owing to the lower prices for storage space and better databases, the next generation of business intelligence solutions was all set to step in. By now, there was a considerable amount of data available but not a centralized place to store it. To address this problem, Ralph Kimball and Bill Inmon proposed similar strategies to build data warehouses (DW).

BA in the New Millennium – Availability of Different Analytical Solutions

By this time, medium- and large-sized businesses had already realized the value of business intelligence solutions. Companies such as IBM, Microsoft, SAP and Oracle were at the forefront of offering such solutions to change the way businesses function.

BA in 2005 – Accessibility of Data for the Common People

Considering the extensive usage of data, companies started directing their efforts on improving the speed at which the information was available. New business analytics tools were introduced to ensure technical as well as non-technical people were able to mine the data and gain insights.

Around this time, the increasing interconnectivity of the business world led to the need for real-time information. This was when Google Analytics was introduced. Google wanted to provide a free and accessible way for users to analyze their website data.

BA from 2005 to 2020 – The Bread and Butter for Companies Globally

With the internet available to almost everyone and the increasing data, companies needed better solutions to store and analyze all the information. Building computers with more storage capacity and better speed was not possible for many. So, companies resorted to using several machines at the same time. This was the beginning of cloud computing.

Since the last decade, Big Data, Cloud Computing, and Business Analytics have become integral for almost all companies. The new advancements have made these technologies even better. Now, Data Analytics and Science are known to be the future. From advertising and marketing to recruiting and planning operational activities, these terms are tossed around in every field.

1.3 THE NEED OR MOTIVATION FOR STUDYING BUSINESS ANALYTICS

To better understand why business analytics has become a buzzword in the corporate world and why you should consider specializing in the field, let's first define what business analysts do.

Business analysts are the people that have the needed knowledge, skills, and sources of information to decide on the direction the business needs to take to succeed in the future.

Their specialties include choosing the opportunities with the highest potential of success, calculating the strategy that would prove most fruitful for the business, or preparing the company for the upcoming change.

1.4 MAJOR CHALLENGES IN BUSINESS ANALYTICS

We generate a large amount of data within a minute in this digitalized environment. The volume of data produced in minute makes it difficult to archive, process, utilize, and interpret the data. Yet big companies fail to find means of making use of this massive volume of info. Currently, as stated earlier, the volume of data generated by major companies is rising at a pace of 40% to 60% per annum. Simply processing this huge volume of data won't be all that helpful and that's why companies are looking at solutions such as data lakes and large data analytics software that will help them manage big data to a great degree. The following are the major challenges in Business Analytics.

The major challenges in Business Analytics are as follows:

- 1. Increase in Number of Sources:** When datasets become bigger and more complex, bringing them into an analytics framework poses a huge challenge in business analytics. If this is ignored, it creates gaps that result in incorrect communications and observations.
- 2. Shortage of Talent for Data Analytics:** Data processing is essential in order to render usable this voluminous volume of data that is generated every minute. The immense demand for big data scientists and big data analysts has been generated on the market with the rapid growth in data. It is essential for business organizations,

because the role of a data scientist is multidisciplinary, to employ a data scientist with qualifications that are versatile.

3. **Data Secrecy:** Gaining valuable insights from Big Data analytics is crucial for business enterprises and it is therefore critical that this intelligence is obtained only through the appropriate team. A major challenge in business analytics posed by the Big Data analytics firms is to successfully bridge this large void.
4. **Handling Large Volumes of Data:** It is not shocking that for every day that passes data is through. It clearly means that business entities need to manage a great regular amount of info. The volume and quality of data available these days will overpower every computer engineer and this is why it is deemed important for brand owners and managers to make data accessibility quick and convenient.
5. **Changing Technological Landscape:** Each day, new technology and businesses are being built with the growth of Big Data. Nonetheless, a major challenge in business analytics posed by big data analytics firms is to figure out which technologies can better match them without new challenges and future threats being added.
6. **Quality of Storage and Retrieving Data:** Enterprise companies are increasing rapidly. With the exponential development of the companies and major business organizations, the volume of data generated increases. Storing a large volume of data is now a huge challenge in business analytics for us. Popular data storage solutions such as data lake/warehouses are widely used to capture and preserve vast volumes of unstructured, organized data in their format.

1.5 USE CASES OF BUSINESS ANALYTICS

1. **Finance:** Business Analytics can help financial organizations to optimize budgeting, determine creditworthiness in case of a loan, and also suggest the chances of a customer defaulting on a loan. In a business that grapples with fraud, Business Analytics plays an important role in raising red flags in time.
2. **Credit Card Business:** Business Analytics helps in extracting crucial information hidden behind the credit and debit transactions and lets the business know, the spending habits, lifestyle preferences, and financial standing, raising red flags wherever there is a probability of loss of business. Credit card companies can decide on the fly which customers they can extend a line of credit to and by how much.
3. **CRM:** Business analytics built into today's CRM systems enable businesses to gain deep insights into demographics, socio-economic information and lifestyle of their customer groups and what would be the best fit strategy to retain and increase the customer base.
4. **Manufacturing:** The manufacturing business is typically a low margin, highly volatile business. It becomes crucial to stay on top of things, so as to cover enough to stay protected against equipment downtime, delays in raw material supply, the inventory

levels to maintain, and the maintenance expense of machines among others. Business analytics helps to decide on the optimum levels of inventory to maintain and how much to make up for equipment downtime and keep production at optimum levels and much more. Also, business analytics also encompasses continuous improvement, identifying corners to cut and helping streamline the business and nimble.

5. **Marketing:** Business analytics plays an important role in determining the effectiveness of marketing campaigns by generating insights on which kind of campaign is most effective, which one is most penetrative in the market, and how much each type of campaign should be invested into gain maximum benefits and cut losses.
6. **e-Retailing:** Today, e-retailing business is expanding like never before with more and more people preferring to order online than visit brick-and-mortar stores with COVID pandemic attenuating it further. There are many players in the market and it becomes necessary for the e-retailer to keep a hawk's eye on inventories to maintain with suppliers and keep the pricing competitive while cutting losses. Business analytics, and more recently, Data Science comes to the rescue. The better a business applies its strategy basis the outcomes of Business Analytics, the better it will fare in the market.

1.6 SCOPE OF BUSINESS ANALYTICS

Business analytics has a wide range of application and usages. It can be used for descriptive analysis in which data is utilized to understand past and present situation. This kind of descriptive analysis is use to assess current market position of the company and effectiveness of previous business decision.

It is used for predictive analysis, which is typically used to assess previous business performance. Business analytics is also used for prescriptive analysis, which is utilized to formulate optimization techniques for stronger business performance. For example, business analytics is used to determine pricing of various products in a departmental store based on past and present set of information. Following are the typical business problems where business analytics are applied:

1. Helping Organizations Know their Customers

For the Chief Marketing Officer or Director of Sales, understanding customers can give one a competitive advantage. Companies that understand why their customers are their customers can fine-tune their products/services, promotions, pricing policies, store locations and other marketing levers to retain and grow their installed base, enhance customer loyalty, target the profitable customers, and improve customer satisfaction.

Retailers find that customer analytics data is not only valuable for their internal planning purposes, but can be sold to generate additional income. For example, product manufacturers may find the details that retailers have on their customers to be of interest. This could include the

demographics of the consumers of their products, as well as a comparison of these consumers with the demographics of their competitor's consumers.

2. Reputation Management

Another means by which companies collect data consumers is social media. Marketers collect information that people post publicly on sites such as Reddit, StumbleUpon, Pinterest, Facebook, and Twitter. With this information, companies can detect negative opinions being expressed about their company or products and take proactive steps to address the problems and to build on positive opinions.

1.7 TOOLS OF BUSINESS ANALYTICS

These are different types of application software that retrieve data from various business systems and consolidate it in a repository. This repository is called a data warehouse, which is then reviewed and analyzed to understand customers better. Any organization can **use more than one business analytics tool to get a 360 degree overview of the company**. It can then use the information to provide key insights into the business to **make meaningful decisions**.

1. **SAS:** SAS is a statistical software suite developed by the SAS Institute for Data Management in North Carolina for data management, advanced analytics, and predictive analytics. It now finds its wide usage in the analytics industry, thanks to its robust nature and easy-to-learn methodologies.

This top-of-the-line business analytics tool simplifies the process of analyzing large datasets in real time. Its high-end platform allows users to convert raw data into meaningful information which can help in drawing relevant insights. Over the span of a few years, SAS has added more modules to derive the best results for SAS Analytics useful for IoT, SAS anti-money laundering, and more.

2. **Excel:** Now, this might come as a surprise to many. Yes, Excel is one of the least costly and widely used data analytics tools all over the world.

Even if you are an expert in SAS or Tableau, the usage of Excel is imperative. It becomes even more important when there is a requirement of analytics on client's internal data. Here, Excel does the complex task of summarizing the data with a preview of pivot tables that help in data filtering based on the client's requirements. Excel has advanced business analytics options that aid in modeling capabilities that have pre-built options like time grouping and automatic relationship detection.

3. **KXEN:** Products offered by KXEN are a company that produces products that are based on automated analytics. Most of their products are based on algorithms that were created by Vladimir Vapnik, a mathematician of Russian origin. A tool that is relatively easy to use and relatively fast in nature, KXEN business analytics tools are able to work with vast amounts of data in an effective fashion. As this tool works like a black box, many users may not like the tool so much.

Decision Tree Algorithm is one of the popular supervised type machine learning algorithms that is used for classifications. This algorithm generates the outcome as the optimized result based upon the tree structure with the conditions or rules.

4. **SPSS Modeller (Clementine):** SPSS Modeller is a data mining software tool by SPSS Inc., an IBM company. It was originally named SPSS Clementine. This tool has an intuitive GUI and its point-and-click modeling capabilities are very comprehensive.
5. **Salford Systems:** It provides a host of predictive analytics and data mining tools for business. The software is easy to use and learn.
6. **MATLAB:** Developed by MathWorks, MATLAB is statistical software based on computers that are used for plotting of functions and data, matrix manipulation, the creation of user interfaces and implementation of algorithms. With many add-on toolboxes, MATLAB is able to address major areas of functionality like finance, image processing, and bioinformatics among other things. Though this is not a free software, there are many other business analytics tools that offer the same functionality.
7. **R:** R is the latest trend that has been introduced into the realm of Business Analytics. This has also been observed to surpass the usage of one of the commercial business analytics tools SAS, even though it is a cost-effective solution. R has grown more versatile and improved over the years and has got a nice facelift after Microsoft took over from its earlier counterparts.

Over the last couple of years, there are more than 1800 models added to R and making its available count of models for usage at an enormous figure of 8000. The count has added a lot of capabilities set to R and also provides seamless integration with Big Data which has proven of great success over the years now.

1.8 BUSINESS ANALYTICAL PROCESS

Business analytics is a 7-step process.

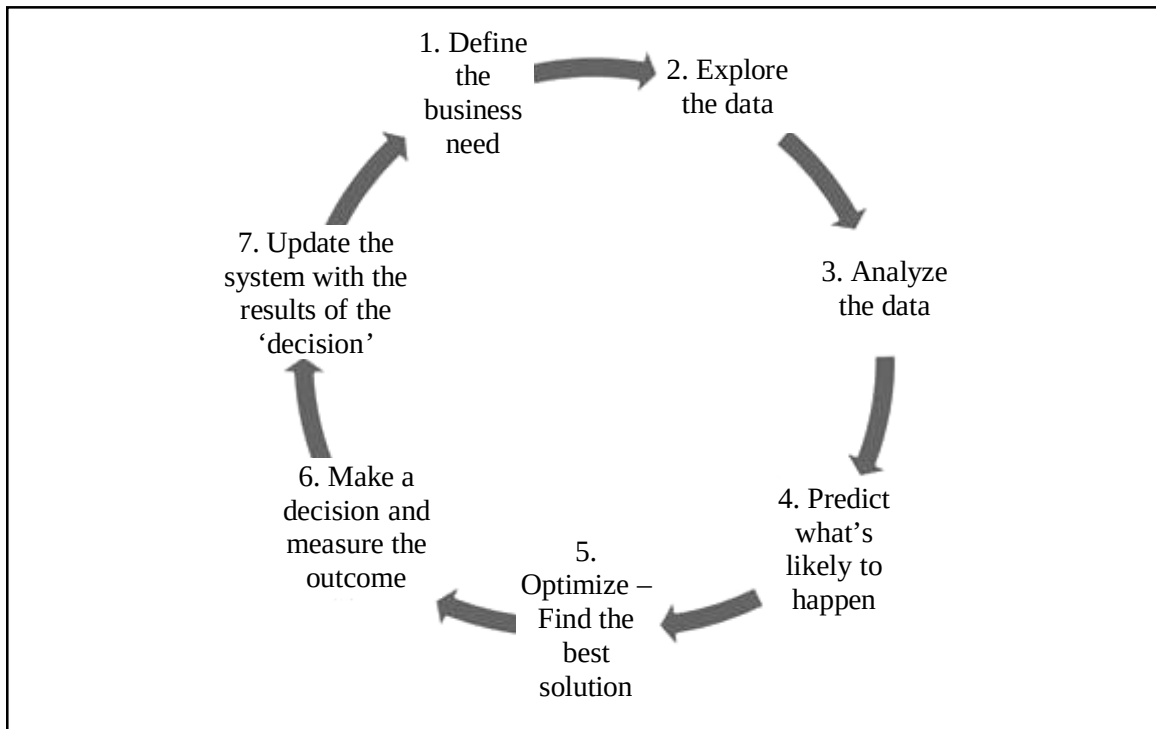


Fig. 4.1: Business Analytical Process

Step 1. Defining the business needs:

The first stage in the business analytics process involves understanding what the business would like to improve on or the problem it wants solved. Sometimes, the goal is broken down into smaller goals. Relevant data needed to solve these business goals are decided upon by the business stakeholders, business users with the domain knowledge and the business analysts. At this stage, key questions such as “What data is available?”, “How can we use it?” and “Do we have sufficient data?” must be answered.

Step 2. Explore the data:

This stage involves cleaning the data, making computations for missing data, removing outliers, and transforming combinations of variables to form new variables. Time series graphs are plotted as they are able to indicate any patterns or outliers. The removal of outliers from the dataset is a very important task as outliers often affect the accuracy of the model if they are allowed to remain in the dataset.

Once the data has been cleaned, the analyst will try to make better sense of the data. The analyst will plot the data using scatterplots (to identify possible correlation or non-linearity). He

will visually check all possible slices of data and summarize the data using appropriate visualization and descriptive statistics (such as mean, standard deviation, range, mode and median) that will help provide a basic understanding of the data. At this stage, the analyst is already looking for general patterns and actionable insights that can be derived to achieve the business goal.

Step 3. Analyze the data:

At this stage, using statistical analysis methods such as correlation analysis and hypothesis testing, the analyst will find all the factors that are related to the target variable. The analyst will also perform simple regression analysis to see whether simple predictions can be made. In addition, different groups are compared using different assumptions and these are tested using hypothesis testing. Often, it is at this stage that the data is cut, sliced and diced and different comparisons are made while trying to derive actionable insights from the data.

Step 4. Predict what is likely to happen:

Business analytics is about being proactive in decision-making. At this stage, the analyst will model the data using predictive techniques that include decision trees, neural networks and logistic regression. These techniques will uncover insights and patterns that highlight relationships and ‘hidden evidences’ of the most influential variables. The analyst will then compare the predictive values with the actual values and compute the predictive errors. Usually, several predictive models are run and the best performing model is selected based on accuracy and outcomes of the model.

Step 5. Optimize (find the best solution):

At this stage, the analyst will apply the predictive model coefficients and outcomes to run ‘what-if’ scenarios, using targets set by managers to determine the best solution, with the given constraints and limitations. The analyst will select the optimal solution and model based on the lowest error, management targets and his intuitive recognition of the model coefficients that are most aligned to the organization’s strategic goal.

Step 6. Make a decision and measure the outcome:

The analyst will then make decisions and take action based on the derived insights from the model and the organizational goals. An appropriate period of time after this action has been taken; the outcome of the action is then measured.

Step 7. Update the system with the results of the decision:

Finally, the results of the decision, action and the new insights derived from the model are recorded and updated into the database. Information such as ‘Was the decision and action effective?’, ‘How did the treatment group compare with the control group?’ and ‘What was the return on investment?’ are uploaded into the database. The result is an evolving database that is continuously updated as soon as new insights and knowledge are derived.

1.9 ADVANTAGES OF BUSINESS ANALYTICS

1. It Helps to Measure Your Progress

Every business has a mission statement, which shows their values to their customers either as a basis of checking in their development or their marketing plan. Mission statements allow companies to retain and even promote employees as they offer an idea of the business's values as guidelines.

While it is crucial to determine who helps your business success, it would not be strategic to leave the matter there. Your business's values must be quantified and conveyed in such a manner that they generate more profits for your organization.

Having quantified values helps a business improve its analytical process because it gives them a definite goal to follow. When your company has a clearly defined plan, everyone in your organization can strive to achieve it.

Your employees can evaluate your business's quantified values to get a clearer idea of what to expect from them. The better they understand the objectives of your business, the better they will be able to perform. This is one of the most significant benefits of business analytics.

2. It Supports Data-driven Decision-making

When your business has access to crucial data, it allows you to make accurate decisions that give better results. Business analytics provides your organization with meaningful and relevant data, fostering smart, data-driven decisions.

Moreover, business analytics makes it easier for you to share your insights and discuss the pathway ahead with your employees and stakeholders. Thus, it promotes collaboration between everyone involved in your business, helping to materialize your goals faster and more efficiently.

3. It Delivers Clear Insights

One of the primary benefits of business analytics is it helps to gain clear insights through data visualization. Data visualization refers to representing data and its insights through visuals or graphics such as pie charts, graphs, etc.

Using comprehensive graphs and charts can make your decision-making process much more efficient. You can extract data from them easily, *identify trends* in your data, and generate relevant and practical insights with better clarity.

Data visualization makes it easier for you share your insights with a non-technical audience that does not know much about business analytics. It is pivotal to business analytics because even though your audience might not be familiar with business analytics' inner workings, their input might be significant in the decision-making process.

The non-technical audience could be your employees or stakeholders. With *visualized data*, you can share your insights with them very easily and make them a part of the overall process, reducing the chances of errors and wrong decisions.

4. It Keeps You Updated

The modern consumer has a short attention span. They move from one business to another very quickly and forget about the previous one instantly. However, you can prevent this from happening, thanks to business analytics.

Marketing analytics is a prominent branch of business analytics, and it focuses on identifying patterns in customer behaviour to understand how your target audience behaves and thinks. It gives you valuable insight into how they choose a particular business, why they prefer one business over another, what products/services interest them, their pain points, and so on.

Such insights can help to make better marketing plans and strategies to promote your products (or services). When you understand your audience, you can focus on their particular needs and gain their attention.

It is a well-established fact that the present industry is highly dynamic. Although such dynamism is very spontaneous, business analytics will ensure that you stay on top of your game by making accurate future predictions.

For example, many businesses adapted to the changing dynamics of a post-pandemic world. Those who adapted to these changes were able to grow their business even when others were lagging. Business analytics helps to predict such changes in the market and mould your growth strategies accordingly.

1.10 LEVELS OF ANALYTICAL MATURITY

The journey to organizational maturity can be approached through five steps, outlined below.

1. Know what you are looking to achieve

Take a step back and think about the business issues your organization is trying to solve. Ensure the project is small enough to be manageable because the scope of a project will be critical to managing time and resources. Having a well-crafted analytics plan with clear goals makes it easier to:

- Craft a promotion.
- Understand the message to deliver.
- Know what components need to be included.
- Identify what the competition is doing.

Invest in datasets that provide the foundation necessary to accurately answer the questions you need answered. While it might be interesting to get data showing that doctor A clicked on a website more frequently than doctor B, if that is not going to factor into key decision-making, then there is no point investing the time and money to uncover this metric. Understand that it takes discipline to ask what the agenda of the analytical organization is. What do we want them to be able to tell us? Equally, it is important to identify the stakeholders: Who will be affected by the

initiative and who are the potential saboteurs? Use analytics to support decision-making, not to replace anecdotal information and intuition.

2. Commit resources and encourage ownership

Resources have become more constrained in the industry, making executive buy-in and the ability to tie the project back to the broader business vision critical. Establish a commitment to quality data. While analytics teams may conduct data extrapolation and analysis, unless the owners of the data provide good quality data and ensure it can be accessed, the analysis will be stymied. So, for example, if the marketing team has held a program with a group of doctors, it is important to ensure that list of doctors is readily accessible within the central data repository. The issue of ensuring that data gets captured properly is a huge problem within the industry, but IT and analytics teams can help by being very specific about what needs to be in the repository. When it comes to data generated through an outsourcing partner, companies need to think about who controls the data and where it resides. The danger is that without due consideration, a company can lose control of its data and insights. One way to manage this issue is to ensure procurement writes into contracts that vendors are required to return the data.

3. Communicate regularly and in a structured way

Communication has to be formalized, with regularly scheduled meetings, to discuss where the organization is in the implementation process, where the potential or existing barriers are, and how those are to be overcome. Companies may also want to consider other methods of communicating updates and process enhancements, including shared portal sites, training and education, and newsletters. *Ad hoc* communication creates uncertainty. The next question is: Who should be communicating with whom and when? The earlier in development that the R&D and commercial sides of the business begin communicating with one another, the better. For example, communication helps R&D to determine priorities and how much need there is in the marketplace for a new compound in a particular therapeutic indication; it gives the commercial side of the business time to develop a commercialization strategy that is more likely to resonate in the market. Conversely, good communication also enables the commercial side of the business to inform R&D if, for example, it would be valuable to have another compound in a particular indication to enable more efficient use of resources, or which therapeutic categories the organization is developing a capability in. It is often advisable to have a data steward at the table to help negotiate and manage the conversations around the data. Data stewards also help mediate the relationship between business and IT.

4. Invest in analytics for the long term in a disciplined way

Turnover within departments and teams is generally high, which can mean while analytics was a priority for one set of team leaders, others might value analytics less. Without organizational discipline around analytics, it is difficult to achieve one version of the truth, and data can be twisted to make a situation look better than it is. Unless companies take a disciplined approach, there is a risk of data paralysis with analytics teams spending their time spinning data and creating numerous reports, all repeating the same thing and each having to be reconciled and

digested by decision-makers. It is important, therefore, to balance an analytics team with data science people alongside those with line experience, such as former district managers. In doing so, it not only creates a team with experience in analytics, but also one that can determine what insights are useful versus simply a point of interest. Technology has made it easier to understand what analytics can do for an organization, but it is important to realize technology won't solve the larger business issues; it is the supporting piece – not the resolving piece. What technology can do is make it possible for business users to access data, and implement and interpret the insights that analytics enables, and not just rely on data scientists or programmers to do this for them. This vastly increases the pervasiveness and power of analytics.

An analytically mature company invests time in data management to ensure the integrity and quality of the data, but there also needs to be governance around how data is used, disseminated and shared across an organization. It is important to determine where the analytical applications are to be stored and provide stakeholders with the training and background to learn about analytics and perform analytics in a shared environment and space that is monitored and controlled.

5. Develop the capabilities to do predictive analytics

Predictive analytics is the branch of data mining used to predict future probabilities and/or trends. One or more predictors is observed among subjects in past outcomes, and then applied to future subjects to predict their responses. Predictive analytics can help companies determine where they should be investing dollars, saving time and money. For example, if a marketing team runs a campaign against a broad audience, without the benefit of predictive analytics, it will cost more and achieve generally the same result as running the campaign against those who are predicted to have a higher probability of responding. It is important to have a more mathematically rigorous way to predict future probabilities and trends.

1.11 EMERGENCE OF BUSINESS ANALYTICS

Today, business analytics has become a buzzword for companies around the globe. Every business, irrespective of its size, is on a lookout for different ways to make sense of the vast amount of raw data available. This is because business analytics has been transforming the way companies function for over a decade now. From targeting the right customers and increasing sales to helping HR personnel select the right candidates and reducing overhead costs, there is hardly any sector where data analytics has failed to tap in. By now, we are sure that you might have a brief idea of why business analytics plays such a vital role. Now, let us take a step back and analyze how did this all start? Why has this field gained popularity just recently? Has the way businesses use data evolved over the years? What did business analytics look like 20 years ago? To help you find answers to all such questions, we have created an in-depth blog. By the end of this blog, you will have a brief idea of how business analytics has evolved over the years. That is not all that we have to offer.

1.12 ANALYTICS IN DIFFERENT DOMAIN OF BUSINESS APPLICATIONS

1. Finance

Business Analytics is of utmost importance to the finance sector. Data Scientists are in high demand in investment banking, portfolio management, financial planning, budgeting, forecasting, etc.

Example: These days, companies have a large amount of financial data. Use of intelligent Business Analytics tools can help use this data to determine the products' prices. Also, on the basis of historical information, Business Analysts can study the trends on the performance of a particular stock and advise the client on whether to retain it or sell it.

2. Marketing

Studying buying patterns of consumer behaviour and analyzing trends help in identifying the target audience, employing advertising techniques that can appeal to the consumers, forecast supply requirements, etc.

Example: Use Business Analytics to gauge the effectiveness and impact of a marketing strategy on the customers. Data can be used to build loyal customers by giving them exactly what they want as per their specifications.

3. HR

HR professionals can make use of data to find information about educational background of high performing candidates, employee attrition rate, number of years of service of employees, age, gender, etc. This information can play a pivotal role in the selection procedure of a candidate.

Example: HR manager can predict the employee retention rate on the basis of data given by Business Analytics.

4. CRM

Business Analytics helps one analyze the key performance indicators, which further helps in decision-making and make strategies to boost the relationship with the consumers. The demographics and data about other socio-economic factors, purchasing patterns, lifestyle, etc. are of prime importance to the CRM department.

Example: The company wants to improve its service in a particular geographical segment. With data analytics, one can predict the customer's preferences in that particular segment, what appeals to them, and accordingly, improve relations with customers.

5. Manufacturing

Business Analytics can help you in supply chain management, inventory management, measure performance of targets, risk mitigation plans, improve efficiency on the basis of product data, etc.

Example: The Manager wants information on performance of machinery which has been used past 10 years. The historical data will help evaluate the performance of the machinery and decide whether costs of maintaining the machine will exceed the cost of buying new machinery.

6. Credit Card Companies

Credit card transactions of a customer can determine various factors such as financial health, life style, preferences of purchases, behavioural trends, etc.

Example: Credit card companies can help the retail sector by locating the target audience. According to the transactions reports, retail companies can predict the choices of the consumers, their spending pattern and preference over buying competitor's products, etc. This historical as well as real-time information helps them direct their marketing strategies in such a way that it hits the dart and reaches the right audience.

Other than above, Business Analytics is helpful in various areas such as the fraud detection, healthcare, defence, sales, etc. Like Mohit, if you are fascinated by numbers, and want a challenging job role, you should definitely have a career in the field of business analytics. Data can be an intelligent tool to foster your brand name in the economic market. Business Analytics lets you exploit this opportunity to make better business decisions.

1.13 APPLICATIONS OF ANALYTICS IN DIFFERENT FIELDS

Not just one or two, the use of data analytics is in every field you can see around. Be it from online shopping, or high tech industries, or the government, everyone uses data analytics to help them in decision-making, budgeting, planning, etc. The data analytics are employed in various places like:

1. Transportation

Data analytics can be applied to help in improving transportation systems and intelligence around them. The predictive method of the analysis helps you to find transport problems like traffic or network congestions. It helps you to synchronize the vast amount of data, and use them to build and design plans and strategies to plan alternative routes, reduce congestions and traffics, which in turn reduces the number of accidents and mishappenings. Data Analytics can also help to optimize the buyer's experience in the travels through recording the information from social media. It also helps the travel companies to fix their packages and boost the personalized travel experience as per the data collected.

Example: During the wedding season or the holiday season, the transport facilities are prepared to accommodate the heavy number of passengers travelling from one place to another using prediction tools and techniques.

2. Logistics and Delivery

There are different logistic companies like DHL, FedEx, etc. that uses data analytics to manage their overall operations. Using the applications of data analytics, they can figure out the best shipping routes, approximate delivery times, and also can track the real-time status of goods

that are dispatched using GPS trackers. Data Analytics has made online shopping easier and more demandable.

Example of Use of Data Analytics in Logistics and Delivery: When a shipment is dispatched from its origin, till it reaches its buyers, every position is tracked which leads to the minimizing of the loss of the goods.

3. Web Search or Internet Web Results

The web search engines like Yahoo, Bing, DuckDuckGo, and Google use a set of data to give you when you search a data. Whenever you hit on the search button, the search engines use algorithms of data analytics to deliver the best-searched results within a limited time frame. The set of data that appears whenever we search for any information is obtained through data analytics.

The searched data is considered as a keyword and all the related pieces of information are presented in a sorted manner that one can easily understand.

Example: When you search for a product on Amazon, it keeps showing on your social media profiles or to provide you with the details of the product to convince you by that product.

4. Manufacturing

Data analytics helps the manufacturing industries maintain their overall working through certain tools like prediction analysis, regression analysis, budgeting, etc. The unit can figure out the number of products needed to be manufactured according to the data collected and analyzed from the demand samples, and likewise, in many other operations, increasing the operating capacity as well as the profitability.

5. Security

Data analytics provides utmost security to the organization; Security Analytics is a way to deal with online protection zeroed in on the examination of information to deliver proactive safety efforts. No business can foresee the future, particularly where security dangers are concerned, yet by sending security investigation apparatuses that can dissect security occasions, it is conceivable to identify danger before it gets an opportunity to affect your framework and main concern.

6. Education

Data analytics applications in education are the most needed data analyst in the current scenario. It is mostly used in adaptive learning, new innovations, adaptive content, etc. It is the estimation, assortment, investigation, and detailing of information about students and their specific circumstance, for reasons for comprehension and streamlining learning and conditions in which it happens.

7. Healthcare

Applications of data analytics in healthcare can be utilized to channel enormous measures of information in seconds to discover treatment choices or answers for various illnesses. This won't

just give precise arrangements dependent on recorded data yet may likewise give accurate answers for exceptional worries for specific patients.

8. Military

Military applications of data analytics bring together an assortment of specialized and application-situated use cases. It empowers chiefs and technologists to make associations between information investigation and such fields as augmented reality and psychological science that are driving military associations around the globe forward.

9. Insurance

There is a lot of data analysis taking place during the insurance process. Several data, such as actuarial data and claims data, help insurance companies realize the risk involved in insuring the person. Analytical software can be used to identify risky claims and bring them before the authorities for further investigation.

10. Digital Advertisement

Digital advertising has also been transformed as a result of the **application of data science**. Data analytics and data algorithms are used in a wide range of advertising mediums, including digital billboards in cities and banners on websites.

11. Fraud and Risk Detection

Detecting fraud may have been the first **application of data analytics**. They applied data analytics because they already had a large amount of customer data at their disposal. Data analysis was used to examine recent spending patterns and customer profiles to determine the likelihood of default. It eventually resulted in a reduction in fraud and risk.

12. Travel

Data analysis applications can be used to improve the traveller's purchasing experience by analyzing social media and mobile/weblog data. Companies can use data on recent browse-to-buy conversion rates to create customized offers and packages that take into account the preferences and desires of their customers.

13. Communication, Media, and Entertainment

When it comes to creating content for different target audiences, recommending content, and measuring content performance, organizations in this industry analyze customer data and behavioural data simultaneously. Data analytics is applied to collect and utilize customer insights and understand their pattern of social media usage.

14. Energy and Utility

Many firms involved in energy management use **data analysis applications** in areas such as smart-grid management, energy distribution, energy optimization, and automation building for other utility-based firms.

1.14 MANAGING BUSINESS ANALYTICS PROJECTS

Business analytics projects are often characterized by uncertain or changing requirements — and a high implementation risk. So, it takes a special breed of project manager to execute and deliver them. Managers have used business analytics to inform their decision-making for years. Numerous studies have pointed to its growing importance, not only in analyzing past performance but also in identifying opportunities to improve future performance. (1) As business environments become more complex and competitive, managers need to be able to detect or, even better, predict trends and respond to them early. (2) Companies are giving business analytics increasingly high priority in hopes of gaining an edge on their competitors. Few companies would yet qualify as being what management innovation and strategy expert Thomas H. Davenport has dubbed “analytic competitors,” but more and more businesses are moving in that direction.

Against this backdrop, we set out to examine what characterizes the most experienced project managers involved in business analytics projects. Which best practices do they employ, and how would they advise their less experienced peers? Our goal was to fill in gaps in management’s understanding of how project managers involved in analytics projects can contribute to the new intelligent enterprise. (See “About the Research.”) We found that project managers’ most important qualities can be sorted into five areas: (1) having a delivery orientation and a bias toward execution, (2) seeing value in use and value of learning, (3) working to gain commitment, (4) relying on intelligent experimentation, and (5) promoting smart use of information technology.

1. Having a Delivery Orientation and a Bias toward Execution

As a starting point, it is important to understand what makes experienced business analytics project managers tick. The vast majority of our interviewees do not consider themselves different from other project managers. Like other focused project managers, they want to deliver their projects on time and on budget, and they have a strong delivery orientation.

But unlike many traditional project managers, they do not have a plan bias. Instead, they have a strong bias toward execution. (See “Learning from Experience.”) Although our interviewees do not question the importance of initial planning, their focus is on project execution and delivery as opposed to adherence to the plan. In fact, they start with the assumption that the initial plan will have to change as the project progresses. This is what we mean by “a bias toward execution.”

Why do analytics project managers have this execution bias? Many say it is because of the inherent complexity of the projects themselves, and they cite three reasons. First, analytics projects are typically characterized by uncertain or changing requirements. Project sponsors and users will often have a vision of what they seek to accomplish with analytics, *e.g.*, to improve direct marketing response, reduce inventory or increase service quality and customer satisfaction while controlling costs. But how they will achieve those goals is often unclear and involves further exploration. Second, the technology or models for meeting the uncertain requirements are often not known; they may be new to the team, or they may not even exist. This adds to the

exploratory nature of analytics projects. Third, users of business analytics applications expect responsiveness. So, the applications, by nature, should be highly responsive to user interaction. The challenge, then, is to find a balance between responsiveness and robustness.

Traditional project management methods tend to focus primarily on planning or *a priori* risk management (as opposed to managing and mitigating risk during execution). However, the uncertainty associated with analytics projects calls for a different approach. A growing body of literature on project management emphasizes the importance of adapting management and processes to the project characteristics. So while there may be a set of general-purpose tools for managing projects, different projects call for different managerial approaches. On the one hand, production-oriented and specifications-based approaches emphasize detailed early planning and requirements specification with minimal ongoing change and exploration. On the other, experimentation-based approaches emphasize less-specific early planning, good-enough requirements, and experimental and evolutionary design with significant ongoing learning and change. The latter, more adaptive approach, interviewees say, is better suited to analytics projects.

2. Seeing Value in Use and Value of Learning

There is increasing awareness in the project management community that sticking to the original plan does not necessarily provide value. Instead, value comes from a focus on execution and delivery. Experienced analytics project managers say they approach return on investment as a process rather than as a control metric. By focusing on execution, they seek to add value throughout the project's life cycle, not just at the end of the project.

Our interviewees are guided by the concept of "value in use," which measures the value in terms of how a given asset provides benefits to a specific owner under a specific use. The idea is that the assets themselves have no inherent value; they generate value only when they offer specific benefits to their owners or users (*e.g.*, by allowing them to do their work differently). Consequently, only when an analytical model or application is actually used, its real benefits (and costs) can be identified.

We found that project managers involved in analytics projects usually want to assess the value of the project quickly and accurately. Interviewees explained how they try to capture value both early in a project and throughout (for example, by using iterative feature-based delivery or rapid prototyping). Indeed, capturing value early and often can significantly improve a project's ROI. For the assessment of value to be accurate, it needs to be carried out with a certain degree of rigor — which, as we have noted and will discuss later, is what our interviewees do.

Many project managers have learned through experience that they cannot expect to be right the first time. A bias toward execution is essential, interviewees report, because it is better to attempt to execute good ideas quickly than to attempt to impose the "perfect" plan. This implies that the focus is not on explaining discrepancies between the plan and actual results but on learning something new in the course of implementation that might justify altering the plan. Similarly, the iterative, incremental delivery described by interviewees assumes that each iteration provides learning inputs.

As a result, project benefits can be expressed in terms of “value in use” and “value of learning” that accrue during the project. Many analytics project managers have adopted project management approaches that tie in with the project management methods that are being developed to support highly complex projects. Adaptive methods assume that there is a need to gather information and learn as to go along. These methods typically emphasize rapid delivery of prototypes and require that those involved be allowed to experiment during the project.

The success of an analytics project is a function of the user’s acceptance of the model or the application. Our data make a convincing case for the value of continuous exposure to user feedback. As a project manager at a European Financial Services Group explained, “Ideally, analysts and users are physically in the same room, or in close proximity.” The design environment and the operating environment should be closely linked, with the analytics project managers facilitating continuous interaction between them.

3. Working to Gain Commitment

Experienced project managers are unequivocal about the importance of engaging business users and other stakeholders as much as possible, as opposed to merely informing them after the fact. As an enterprise-business-intelligence architect at an international transport solutions provider put it: “We don’t want to develop a model just like that. If the business processes aren’t aligned with the model, or if the business doesn’t understand the definitions used in the model, then it simply won’t be used.”

The importance of explaining or clarifying the thinking behind a decision — or, in this case, the analytical model or application — cannot be overestimated. Indeed, interviewees say that one of the major risks of analytics projects is that the decision-makers won’t be savvy enough to understand the analysis or the model’s underlying assumptions, and they will try to apply it where it is not applicable. Explanation is also crucial in gaining trust, as one project manager at a financial institution notes: “Gut feeling and intuition still take precedence over analytics. No matter how transparent analytical models are, they are inevitably statistically complex. That’s why users find it difficult to put their faith in quantitative data and methods.” And this is why analytics project managers should be pedagogical experts and help open up the black box of analytic models.

Ultimately, our interviewees agree that expectation management should not be overlooked. Setting the right expectations at the beginning (*e.g.*, regarding the quality of the data and the applicability of the models) and managing them as the project progresses increases both acceptance and the chances that the project will be successful.

The process described above bears a strong resemblance to what W. Chan Kim and Renée Mauborgne recognized thought leaders and authorities on business strategy, innovation and wealth creation, have described as “fair process.” Process fairness has proven its worth in diverse management contexts as a way to gain stakeholder commitment to decisions and change.

4. Relying on Intelligent Experimentation

A key element that emerges from the interview data is the importance of experimentation. Many of the analytics project managers we spoke to consider experimentation fundamental to the learning process. This is consistent with leading research by Harvard University professor and innovation management authority Stefan H. Thomke, who defines experimentation as a fundamental innovation-process activity, consisting of iterative trial and error and directed by insight. The execution of an experiment, then, follows a four-step cycle: design, build, run and analyze.

The quality of the experimentation process has a strong bearing on the extent to which the project succeeds. Interviewees tend to be strong advocates of “good experimentation,” which is consistent with the scientific method. Well-designed experiments need clear goals and objectives, which is why the first step of the scientific method devote significant time and effort to observation, to specifying the questions the experiment is intended to answer and to background research. Project managers need to invest time upfront examining the analytics project and setting the objectives. Good experiments need measurable hypotheses about the expected outcomes and controlled testing of these hypotheses. Interviewees reported spending significant amounts of time setting up the experiments and analyzing the results. What they learn forms the basis for improvements and for the next batch of experiments.

However, most interviewees recognize that a laboratory-style scientific approach is neither appropriate nor practical. They take a more pragmatic approach to experimentation. Still, their process has enough rigors to allow a sufficiently accurate assessment of a project’s benefits. Interviewees acknowledge that business analytics is still a relatively new area and that some companies are still learning to incorporate analytics into their business. As a business analytics coordinator at an international food retailer put it: “Our policy is clear: Not using a control group is no option. But the reality is sometimes different. Many of our people are still trying to get to grips with analytics. Conflicts won’t help our case.” Successful, experienced project managers will try to advance the learning curve and coach the stakeholders into cultural change.

5. Promoting Smart Use of Information Technology

So far, information technology has been conspicuously absent from this discussion. Yet intelligent use of IT can allow for frequent and faster iterations between the design and operating environments, and this can improve experimentation efficiency. MIT professor and global IT expert Erik Brynjolfsson, who coined the phrase “IT productivity paradox,” has noted that leading companies leverage IT to revolutionize the way in which they innovate by playing on four dimensions simultaneously: measurement, experimentation, sharing and replication. The big advantage of IT-based experimentation, he argues, is that it can trace causality in a way that would be impossible with pure measurement and observation.

IT capabilities are the key to helping companies to explore as well as exploit their full potential in turbulent markets. But while business has embraced IT capabilities, it is often far less positive about the IT department. Many of the analytics project managers we interviewed identify

with the business side of their organizations even if they report to the IT department. They often trace their frustration with IT to negative experiences. As one explained: “Whenever IT is involved, business analytics projects cost more and take more time than planned. They are hideously inflexible. It’s virtually impossible to go ahead with anything at short notice. And not only that, they just speak another language. They really lack the sense of urgency and pragmatism to find on the business side.”

IT departments that ignore complaints from the business side risk are being circumvented. In fact, some of the analytics teams we encountered built valuable models and applications independently from the IT department. Still, bypassing the IT department altogether can be counterproductive, especially when the focus is on delivering enterprise value rather than locally optimized solutions and functional value. Certainly with enterprise value, the most appropriate *modus operandi* would be to approach analytics projects as partnerships between the business side and IT.

Research suggests that best-in-class CIOs have realized that IT and business need to find better ways to work together. By proposing pragmatic solutions and pointing out the consequences of infrastructural decisions, CIOs can become constructive partners, enabling their businesses to make smarter choices. This means that IT should, where possible, pursue opportunities to deliver faster implementation cycles, maintaining just enough process and architectural hygiene to ensure quality and professional support.

But what is just enough process and infrastructure? Enterprise infrastructure remains a long-term investment. The big challenge is to develop a process that provides for flexible infrastructure even as the process itself — the way applications and infrastructure are built and modified — remains stable. This will require infrastructure architects with mindsets much like those of the analytics project managers we interviewed. Indeed, these architects will need to have a strong bias toward execution, so that IT solutions and infrastructure are rooted in the present without mortgaging the future.

Two things are certain. First, the boundaries between functional domains are blurring within organizations, requiring cross-functional collaboration. Second, it will take experience-based negotiation, not theoretical design, to create just enough process and infrastructure. This is a vital area where experienced analytics project managers can put their interpersonal skills to good use.

SUMMARY

- Frederick Taylor established the first-ever business analytics system in the United States of America at this time, which he named scientific management. The goal of this system was to examine manufacturing procedures and labourers’ body movements in order to find ways to improve efficiency.
- Taking in and processing historical business data is what business analytics is all about. Identifying trends, patterns, and core causes through analyzing the data. Using those insights to make data-driven business decisions.

- Frederick W. Taylor created the first organized method of business analytics in the United States in the late 1800s.
- Businesses can use business analytics to make better, more informed decisions and achieve a range of objectives. Businesses can better understand consumer behaviour by harnessing data. Get a better understanding of their competitors.
- Data analysis not only improves productivity, but it also aids in the discovery of new business prospects that could otherwise go unnoticed, such as untapped client categories. As a result, the potential for growth and profit becomes limitless, as well as more intelligence-based.

QUESTIONS

1. Explain the evolution of Business Analytics.
2. Explain the need/motivation for studying Business Analytics.
3. List out the advantages of Business Analytics.
4. Explain the challenges of Business Analytics.
5. Explain the understanding of Business Analytics.
6. Write a short note on Levels of Analytics of Maturity.

REFERENCES

- <https://www.asmbmr.edu.in/blog/the-history-of-the-evolution-of-business-analytics/>
- <https://www.oracle.com/in/business-analytics/what-is-business-analytics/>
- <https://www.cyfe.com/blog/history-evolution-business-analytics/>
- <https://taxandbusinessonline.villanova.edu/blog/the-importance-of-business-analytics/>
- <https://www.michiganstateuniversityonline.com/resources/business-analytics/how-business-analytics-can-help-tor-business/>



MODULE 2

RISE OF HUMAN RESOURCE (HR) ANALYTICS

Learning Objectives:

1. Study the basic concepts of HR Analytics.
2. Understand the history of HR Analytics.
3. Apply the concept of HR Analytics.
4. Evaluate the scope of Big Data in HR Analytics.
5. Understand the future of HR Analytics.

2.1 HISTORY OF HR ANALYTICS

To understand the history and process of development of HR analytics, it is essential at the outset to mention about various HR valuation models, which are in use from early 1990s onwards. Such valuation models not only try to measure human resources for organizations but also legitimize the use of various ratios and values for a critical HR decision-making process. For example, balanced scorecard pioneered by Kaplan and Norton (1992) can not only measure performance data but can also provide valuable insights into employees' potentiality, which can substantially influence talent management related decision-making. Similar efforts on monetary valuation models pioneered by Sullivan (2000), intellectual asset valuation by Anderson and McLean (2000), return on assets and economic value added (EVA) by Stewart (1997) and so on help us not only in approximate valuation of human resources in organizations but also in terms of using vital data for critical HR decisions.

BRIS is a descriptive analytics used by HR managers as a reporting tool for the past and the current happenings. It is not forward looking. Hence, it cannot visualize the future outcomes or future decisional implications. As a result, with HRIS, HR managers cannot align the HR decision-making process with organizational strategies and business goals. With business complexities, the pattern of HR decision-making has been changed. It has now become more dynamic, forward looking and strategic encompassing all functional areas of organizations. Another inherent problem with descriptive decision-making tools like HRIS is that these can only see the situation from the macro perspective, while the modern HR decision-making process

requires in-depth analysis even at micro levels for predicting the behaviour of different variables, their interrelationships and how it impacts business of the organization. This can only be done with predictive analytics through the process of identifying the unexplored pattern of the data with appropriate models.

Measures in any form, at the first instance, help us to compare or benchmark. For human resources, obviously measures help us to understand the relative value of data, which ensures better approximation of our decision-making process. With analytics, we can perform multiple measures interpreting interrelationships between data. In the process, our HR decisions also become more encompassing, sophisticated, and strategic and business aligned. HR analytics over the years not only helps in measuring but also leads to associating. Over the years, it transcended from causation to prediction.

The history or process of development of HR analytics over the years traced back first with various HR valuation models and various measurement tools based on data collection, primarily with the use of statistics. Initial focus was on descriptive methods, surveys and collated data from HR functions, based on which HR managers used to develop metrics and use the same for HR decision-making. With Fitz-enz and Mattox II's (2014) revolutionary publication on 'Predictive Analysis for Human Resources', we could understand how HR decision-making processes can be made more encompassing, scientific, and strategic and business aligned. Although HR analytics roots are credited to Fitz-enz (1984), its use and applicability has been popularized only after 2014 publications. Today, HR analytics have become a legitimate strategic tool for HR decision-making process, which can substantially reduce the chances of wrong decisions that may have a potential negative effect on achieving the strategic intents and business goals.

Many organizations use the terms HR analytics and talent analytics interchangeably. Today, however, more appropriately, it is called people analytics. People analytics require use of Big Data in HR as people or employees of the organization have enduring effect on holistic organizational performance, and their assessment requires cross-functional data. Because of this, in people analytics, organizations involve not only HR professionals but also statisticians, engineers and mathematicians.

Bersin (Deloitte) suggested a people analytics model which integrates HR functions with sales, customer retention, accidents, frauds and quality issues, and then performs the management of data to get new insights into better decision-making.

Major technology vendors such as IBM, SAP, Oracle, Visier and Evolv today offer HR analytics solutions to organizations, which can be customized by the HR managers, pacing with their needs, to get edge in decision-making and to emerge as a legitimate business partner for the organization.

Human resource is a people-oriented function and is so perceived by most people. But for those who think that the HR team's contributions are limited to extending offer letters and onboarding new hires, human resource analytics (HR analytics) can prove them wrong. When used strategically, analytics can transform how HR operates, giving the team insights and allowing it to actively and meaningfully contribute to the organization's bottom line.

2.2 DEFINITION AND MEANING OF HR ANALYTICS

Analytics

It is defined as the interpretation of data patterns that aid decision-making and performance improvement. HR analytics is defined as the process of measuring the impact of HR metrics, such as time to hire and retention rate, on business performance.

Analytics is the scientific process of discovering and communicating the meaningful patterns.

HR Analytics

To understand the essence of HR analytics and to explain how it impacts business performance, we asked Mick Collins, Global Vice President, Workforce Analytics and Planning Solution Strategy and Chief Expert at SAP Success Factors, to break it down for us.

“HR analytics is a methodology for creating insights on how investments in human capital assets contribute to the success of four principal outcomes: (a) generating revenue, (b) minimizing expenses, (c) mitigating risks, and (d) executing strategic plans. This is done by applying statistical methods to integrated HR, talent management, financial, and operational data,” says Collins in an exclusive discussion with HR Technologist.

HR Analytics focuses primarily on the HR function and is not – as is largely believed – exactly interchangeable with people analytics or workforce analytics.

HR practices enabled by information technology that uses descriptive, visual and statistical analyses of data related to HR processes, human capital, organizational performance, and external economic benchmarks to establish business impact and enable data-driven decision-making.

The CIPD (2016) outlines that an HR analytics strategy should meet three aims:

- To connect HR data with business data in order to explore how the HR function influences outcomes at the organizational level.
- To enable HR leaders to design and implement HR management activity in an efficient and effective manner.
- To allow the organization to measure the effectiveness of HR in delivering against its objectives.

2.3 ROLE OF ANALYTICS

Analytics is especially useful in areas which record a lot of data or information not just in the world of business, but also in science, sports, healthcare and just about any field where vast amounts of data are collected. It provides us with meaningful information which may otherwise be hidden from us within large quantities of data.

In the world of business, organizations would usually apply analytics in order to describe, predict and then improve the business performance of the company, specifically in the following

areas like Web Analytics, HR Solutions, Fraud Analysis, Risk Analysis, Advertisement and Marketing, Enterprise Decision Management, Market Optimization, Market Modeling, etc.

2.4 FOUR LEVELS OF ANALYTICS

When we think about data trends, we think about the big catchphrases like Machine Learning, Big Data, Artificial Intelligence and the like. But at the core of it, data is all about helping to make smarter, more well-informed decisions.

What would be the point of things like Predictive Algorithms and Big Data if they did not lead organizations to make smarter, better, well-informed decisions? But it is not just access to data that helps to make smarter decisions, it is the way to analyze it. That is why it is important to understand the four levels of analytics: descriptive, diagnostic, predictive and prescriptive.

1. Descriptive Analytics

Descriptive (also known as observation and reporting) is the most basic level of analytics. Many times, organizations find themselves spending most of their time in this level. Think about dashboards and why they exist: to build reports and present on what happened in the past. This is a vital step in the world of analytics and decision-making, but it is really only the first step. It is important to get beyond the initial observations and dive into insights, which is the second level of analytics.

2. Diagnostic Analytics

Diagnostic analytics is where we get to the why. We move beyond an observation (like whether the chart is trending up or down) and get to the “what” that is making it happen. This is where the ability to ask questions about the data and tie those questions back to objectives and business imperatives is most important.

Imagine going to a doctor where the only thing they do is look at you, make the observation that “Oh, yeah, you look sick,” and then leave the room. That is not going to do much for your health. We need to be able to understand what is causing the sickness. The doctor should make the observation, diagnose you and then give you a treatment plan to help you feel better. It is the same thing with analytics: to make an observation, identify the descriptive analysis and move forward to the diagnosis.

3. Predictive Analytics

Predictive analytics allows organizations to predict different decisions, test them for success, find areas of weakness in the business, make more predictions — and so forth. This flow allows organizations to see how the first three levels can work together.

Predictive analytics involves technologies like machine learning, algorithms, and artificial intelligence, which gives it power because this is where the data science comes in. Now, when we incorporate the importance of not just predicting, but using data science, statistics, and the third-level of analytics combined with the first two levels, organizations truly can see success with their data and analytical strategies.

However, the reality is that currently most of your organization isn't spending a lot of time with predictive analytics. Leaders are spending most of their time in descriptive and diagnostic, but predictive is a very important part of the puzzle. Every organization needs a workforce that can speak the language of data, and the language of predictive analytics.

4. Prescriptive Analytics

Prescriptive analytics exist at a very advanced level and is the most powerful and final phase, and truly encompasses the "why" of analytics. It is when the data itself prescribes what should be done. Data-driven decision making is tied most closely to predictive and prescriptive analytics, even though these are the most advanced.

Think of prescriptive analytics as taking all other levels of analytics to prescribe things you should be doing; the data and analytics show you the way. Thomas Matthew, chief product officer at Zoomph describes it well:

"Prescriptive analytics builds on predictive by informing decision-makers about different decision choices with their anticipated impact on a specific key performance indicator. Think of traffic navigation app, Waze. Pick an origin and destination, and a multitude of factors get mashed together, and it advises you on different route choices, each with a predicted ETA. This is everyday prescriptive analytics at work."

Think of the first three levels of analytics: you have your description of what has happened; followed by diagnosing why, and then you end with predicting what will happen. Now, imagine you allow the data and analytics to inform you what action to take. That is powerful and why it matters for businesses.

2.5 PURPOSE OF HR ANALYTICS

HR analytics can help enterprise leaders to develop and improve recruiting methods, make general and specific hiring decisions, and keep the best workers with the company. In addition, workforce analytics can help management personnel:

1. To identify the need for new departments and positions,
2. To determine which departments or positions can be reassigned or eliminated,
3. To identify and quantify physical risks to employees in specific positions,
4. To assign and delegate responsibility for tasks and goals,
5. To investigate the effectiveness of performance management or performance-related pay in improving performance,
6. To evaluate the effectiveness of learning and development activities,
7. To measure the impact of organizational development activities,
8. To improve employee retention rates,
9. To reduce absenteeism of employees,

10. To evaluate the effectiveness of different sources of recruits,
11. To provide the guidance on future HR strategy.

2.6 PITFALLS OF PEOPLE ANALYTICS

- Recognizing the most common mistakes
- Avoiding the biggest messes.
- Proceeding on for people analytics journey while mitigating risk.

In the chapter '*In the article Ten Pitfalls of People Analytics*', Mike West shares ten of the most common pitfalls he has seen people analytics efforts succumb to over the years. These pitfalls can prepare for self and your teammates to steer clear of danger.

Pitfall 1: Changing People is Hard

People analytics can change the entire nature of managing people – and nothing gets people stirred up quite like a change. Regardless of the scope of people analytics you are implementing, it is likely that you will encounter some level of difficulty if you are trying to change the way people think, the way people make decisions, and the way people do what they do.

Quite often, people talk about resistance to change. Ambivalence is more often encountered than resistance. Ambivalence can be worse than resistance because everyone may outwardly cheer you on, however indirectly, they do not provide the support you need to be successful. The difficulty of change materializes in many different ways. Here are just a few:

- You need someone to give access to a system, and there is a lengthy delay.
- You need to get others to explain data definitions so that they can claim, either don't reply or they tell they are too busy to get to it.
- You need to talk to someone about a report that are building for them, and you might not get time on their calendar.
- You need someone to change how others are inputting data into systems and help us with the data clean-up, and nobody gives the time for the day.
- You need to proactively build awareness of the benefits of change and address the difficulty of change to try to get everyone pitching in. Otherwise, what starts as ambivalence can quickly turn into full-blow failure, leaving your movement toward data-informed in tatters?

Here are some ways that can proactively improve the probability of the success.

- Set up a cross-functional people analytics taskforce to help you be more aware of the needs of others and create an umbrella of support. Whenever you encounter problems, talk it over at a monthly meeting. Not only will you get a valuable input, but also buy-in.

- Build lasting relationships with IT, data management, and human resource information technology (HRIT) folks. You need their blessing, their input, their support, and their friendship. As with any significant relationship, you usually get out what we put in.
- Get a project sponsor. If, when you communicate with others, you can drop that the head honcho is closely watching the outcome of the project, people tend to take more heed.
- Three words: communicate, communicate, and communicate. It is a good idea to send notes to all stakeholders periodically – reinforcing the importance of the people analytics, revisiting the benefits of the project, and giving them updates on status.

Here, the Four S People Analytics Framework – also known as the People Analytics Intersection is referred for introducing people analytics, *i.e.*, people analytics is the new thing that is created at the intersection of people strategy, science, statistics, and systems. Figure 2.1 illustrates how people analytics joins the four broad people S capabilities (strategy, science, statistics, and systems) to create some innovation that didn't exist before: that thing is people analytics. Most companies will start people analytics with strengths in one S and deficiencies in one or more others. It is essential to recognize that different strengths and weaknesses will produce various pitfalls. It is the component that is most deficient that will define the trap that will materialize. By clarifying capability related blind spots below, you can avoid these pitfalls entirely.

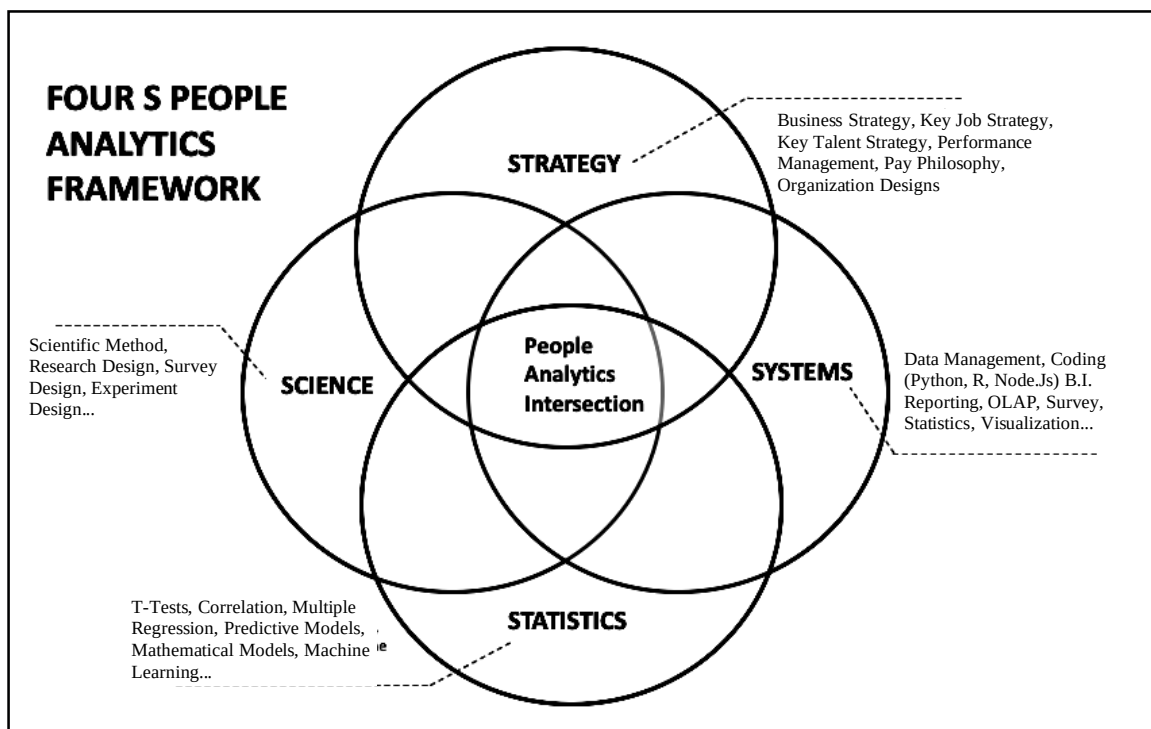


Fig. 2.1: People Analytics is the Intersection of People, Strategy, Science, Statistics, and Systems

Pitfall 2: Missing the People Strategy Part of the People Analytics Intersection

People analytics is only beneficial if it is aligned with your strategy and informs decision-making. Anything else is just going in circles. When data analysis is not linked to strategy, the company is wasting time and money in all this activity that is never going to be used.

In the absence of a linkage of people analytics to company strategy, most companies either attempt to measure everything or measure only what everyone else is measuring, because they don't know what else to do.

Measuring What is Easy to Measure

This is by far the biggest mistake that people make — they work out what is easy to measure and measure everything easy to measure regardless of its relevance to the business.

HR executives will often brainstorm measures and analyses. Usually, the ones they end up with will be the ones they have heard of, read about in a management journal, or are metrics they have seen before. This is not the best way to develop a plan because the resulting list is overwhelmingly long, is not relevant to the strategy of the business, and does not answer the most critical questions.

Measuring What Everyone Else is Measuring

A trap that many businesses fall into is looking to other companies to see what they are measuring, and then just doing the same.

For example, the HR leader may notice that many businesses are conducting annual engagement surveys and think that they'd better do that too. Rather than taking a step back and working out what questions they need to ask, they look elsewhere at the questions other businesses are asking and follow suit. Instead, the ones they should be measuring are the ones that are directly relevant.

- If you start with an ill-defined problem, your work turns into a fishing expedition as to cast to line into the data lake and hope something hooks itself.
- If you start on a slight problem, our efforts turn into a trivial pursuit, where you might win the game, but the fact that you won doesn't matter a bit when done.
- If you try to start without even specifying a problem, you make yourself into an atmospheric scientist: You're analyzing the air.

The best insights come from focused projects. You need to understand the problem you are trying to solve and have a clear notion of what you need to analyze to get the results you are looking for.

Start with an idea, form a hypothesis, and look to confirm, refine, or reject the theory with data.

Focus matters because, unless you have a large team and unlimited resources, you do not have time or money to do everything. Strategy helps drive focus.

Pitfall 3: Missing the Statistics Part of the People Analytics Intersection

There are many reasons to learn about statistics, but perhaps its most important purpose is to make better decisions in a world of uncertainty. Increasingly, the world is also full of data. In theory, we look at data to reduce uncertainty. However, the more data the more uncertainty. Statistics can help to make sense of data, and in so doing, make more sense of the world.

People unfamiliar with statistics expect to be able to see the answers to their questions in a line graph or bar chart. However, visual patterns can mislead you. Just because two bars on a graph are different sizes doesn't say that the difference is significant or meaningful. Just because a line seems to increase overtime doesn't mean that your conclusions about why it is rising are the actual causes. Only substantial differences among straightforward comparisons present themselves obviously in visualizations. Overreliance on visualization leads to simple observations that are not up to the task of producing answers to complex questions. The real world is complicated: many factors push and pull in different directions at the same time. These don't translate readily to visualization.

Statistics can be an intimidating subject for many people, but ultimately, it is a subject involving certain logic and procedures that can be learned by anyone. Relationships are much more complicated and uncertain; to manage relationships every day!

Pitfall 4: Missing the Science Part of the People Analytics Intersection

The typical capabilities organizations look for in analysts are knowledge of technology tools like Python, R, or Machine Learning. In my experience, though, nothing could be more misplaced. About the only things in common among the best analysts I've met in people analytics are curiosity, imagination, and knack to get to the heart of problems. Besides this, it is also noticed that they tend to have studied some form of behavioural science. In particular psychology, sociology, operations science, or economics, for example.

Science is everywhere in today's world, so much so now that we hardly notice it. Science has impacted nearly every aspect of our daily lives, from what we eat, to how we dress, to how we get from point A to point B. Why not work too? Advances in science are transforming our world at an incredible pace. Our children's future will be filled with leaps we can only imagine. No one can escape the significance of science in our world. Still, not everyone understands the importance of science, has been taught to think critically, or been provided with the tools to analyze and test a problem in the ways people who study science have. Science offers highly valuable skills for the workplace. The application of science to people at work is a new frontier that I'm proud to be a part of, and you can be part of this too!

The most important thing about science is that it is not a dictionary of facts. It can generate facts, but more importantly, it is a method you can use to discover new facts and understand the world around you. Science is a process for observing, wondering why, and investigating your ideas.

The beauty of science is that it's self-correcting. Science is coming up with an idea, testing that idea, and then observing to decide if it works or if you have to throw that idea out. If a plan is

wrong, you have to get rid of it, and you have to bring in a new one. Science is a way of looking at anything you want to understand and saying how does this work, why does it work, and how can you know about this?

Pitfall 5: Missing the System Part of the People Analytics Intersection

Today, it could probably go without saying that companies will have technology systems supporting the many transactional functions of human resources. You will certainly have a system for payroll so you can pay people. After this, you will most likely have a human resource information system and some combination of other systems to facilitate all of the specialized activities of human resources: applicant tracking, performance management, compensation planning, employee relations, learning management, and other specialized operational HR activities or activities facilitated by HR. Systems are rarely the missing ingredient.

The operational systems described in the paragraph above are designed primarily to serve the transactional needs of HR, not the reporting or analysis needs of people analytics. Sure, each system may have a front-end reporting interface to provide access to data in the systems directly; however, these leave much to be desired. In most cases, the standard reports available from transactional systems are just lists of people or facts, and most of the time, you have too little control over what goes on those reports. These lists of facts are necessary but not sufficient. Transactional systems are not designed to perform the core tasks of analytics.

Here are just a few examples of the tasks necessary for analytics that transactional systems are generally not designed to do:

- Provide control over workflow functions like extract, transform, load (ETL) from and to other data environments, *e.g.*, how to move data to or from different data sources and join them.
- Provide control over how you to add or remove data elements on a report and how you group data.
- Provide control over what calculations to perform and how you perform calculations.
- Allow you to construct custom datasets to perform statistics like correlation, chi-square, multiple regression, and t-tests.
- Provide control of the design of the graphs.
- Provide business users a central location for all of the information relevant for them to manage their teams. You don't want to tell business users they have to go to four different systems to pick up data on different topic areas.

<Remember> The sub-functions of human resource management (HRM) are Talent Acquisition, Total Rewards (Compensation and Benefits), Learning and Development, Talent Management and Organisational Effectiveness, Diversity, and Employee Relations). The exact division of responsibility and names of sub-functions of human resources vary from company to company as do the constellation of systems they use to perform each function.

As a result of the reporting and analysis deficiencies of the transactional HR systems, data-minded professionals serving sub-functions of HR create makeshift reports, dashboards, and analysis in desktop tools like Tableau or Excel to serve needs not met currently by the transactional systems. For example, if the reporting needs of recruiting have not been picked up by a centralized HR reporting team, then the Talent Acquisition team may hire their own analysts to build reports on the recruiting process. The stories the analysts in Talent Acquisition makes may speak only to the data from the applicant tracking system (ATS), which serves the operational needs of the recruiting function. The Talent Acquisition analyst may have no access to data in other HR systems and will likely have no insight or interest in how all the data from different environments fit together. This leaves a lot to be desired.

The pitfall here is that you want to understand with data may cross over functional boundaries – the splintering makes it so that you cannot see the most important insights. Inevitably, problems may be resolved in one part of the company and just pushed to another, or a solution may remain elusive to you forever because a universal data perspective has not been brought together.

In addition, if the analytics activity is splintered, then duplicative tasks are being performed in different places of the company. Often, the same tasks are repeated by various people that could be more efficiently handled in one standard data environment and then split out for their needs or modification. Inefficiency is the problem you are most likely to find if you have missed the systems part of the four S's of people analytics.

People analytics can be cobbled together and performed using borrowed systems and scraps of data; however, this can only be a short-term solution. If you have many people performing many tasks on data in an inefficient (and overall ineffective) manner, your chances of long-term success are slim. United you stand, divided you fall. This is why you should try to get your systems house in order by creating a centralized people analytics data environment that brings together multiple data sub-domains and data management functions into one common area.

Pitfall 6: Not Involving Other People in the Right Ways

What if you built a data dashboard, and nobody showed up to use it? The users are what will make your data dashboard a success or failure. There are lots of reasons why users may not flock to the tools you roll out. First and foremost, if the data dashboard doesn't add value to their job, they won't use it; it may be as simple as that. It also may be that they can use the information from time to time but not all the time. They don't log into the system frequently enough to remember how it could be helpful to them or be comfortable operating in that environment to enjoy it.

Rather than start with data that may or may not have any value, start by getting away from your desk to understand better the world of the people you support. Ask questions about production, sales, and other processes. To help other people, to need to understand what those people do, what their pains are, and what success looks like to them. Armed with that information, you can connect them with an analysis or a report to help them do what they do better.

While it is good to meet with end-users early in the process, at the same time, you must realize that you can't expect people to be able to translate their needs to you into the language of dashboard and analysis design. You should not simply walk in and ask them what they want to see on a dashboard. Most often, they will describe something they have seen before. Unfortunately, this may not be the best report or analysis to help them solve their problems. You won't find this out until you build what they asked for and discover they aren't using it. Contemplating what went wrong, you may think, "but this is what they asked me for." Absent a deliberate strategy of interaction with end-users from beginning to end; this disconnect will happen a lot.

<Remember> People analytics is new. Remember that most people don't know what people analytics is (or is capable of doing), let alone what kind of analysis could help them right now. Figuring out what other people need should be your area of expertise. Your job is to elicit an understanding of those to support — what their job is and what business problems they are solving right now and then design a reporting or analysis solution that will help them do it better.

After you have met with people to talk about what they do, the next thing you should do is create a prototype analytics solution. A prototype is a subset of the total solution. The scope is narrowed down to a few of the essential data elements and company segments. At its core, a prototype usually consists of a limited dashboard or analysis solution combined with a stripped-down version of one or more end-user tools that visualize the data. A prototype is a great way to get a reaction, flush out uncertainties, model the challenges that you will find working with a similar dataset at full-scale. A prototype also helps you get people involved early.

Pitfall 7: Underfunding People Analytics

If you are hired to head up people analytics and find out you have no budget, don't feel bad. You aren't the first, and you won't be the last. Whether it makes sense to quit now or ride it out and fail later depends on your circumstances. One thing is sure if you don't communicate clearly and honestly about the level of expectations others have for what they hope to achieve from people analytics versus the level of support and resources provided to achieve those results, then your efforts will fail.

Full-time people analytics professionals have a notoriously high incidence of burnout, because people analytics crosses every department in the company, every sub-function of human resources (Recruiting, Compensation, Benefits, Payroll, Employee Relations, Learning and Development, Organization Design, Diversity, ...) and requires data from many different systems (ATS, HRIS, Compensation Planning, Performance Management, ...). Before you know it, you are getting requests from the CHRO, division vice presidents, mid-level managers, head of HR sub-functions, and people you haven't even heard of before.

If you can't get resources to build a scalable data environment, and you can't keep up with report demand — something doesn't reconcile. Clearly, what you are doing must be important, or you wouldn't be getting hit with all these report requests, but how then can it not be significant enough to invest in the proper data, systems, and support? Having too many people coming to you

for reports is a better problem you have than not having any! Yet, you will need to navigate this situation carefully.

Don't let enthusiasm for the work of people analytics get in the way of making rational decisions about what you can reasonably deliver. It is a natural tendency to get excited that others are asking for your help and say yes to everything. Resist this tendency. It would be best if you thought long and hard before you take on a project. Here are some considerations:

- Does this project offer enough business value above the effort?
- Are you considering the relative importance of this project to other projects that will inevitably receive less of your time and attention if you take on the new project?
- Do you truly have what you need to be successful in this new project?
- What is the person asking for the analysis or report from you willing to help you to be successful?

There are some things you can do that will help you:

- Be blunt about the time and resource realities you face.
- Create a transparent prioritization system. Point to this. Use it.
- Request that the people who need something from you do something for you too. This could be as simple as becoming an advocate for people analytics and you in the company. Become a project sponsor or member of the people analytics task force. It could be accessible to members of their team for data entry, data definitions, data cleanup, or data training.

Pitfall 8: Garbage-In-Garbage-Out

Garbage in, garbage out (GIGO) describes the concept that flawed, or nonsense input data produces nonsense output or "garbage." If you have questionable data feeding into your dashboards and analysis, then their production won't be worth a pile of rotten tomatoes.

If there were ever a poster child for a garbage-in-garbage-out statement, it's the realm of HR data. Data quality is an ongoing problem in all analytics but especially when there's people data involved. People and their reporting relationships are in a constant state of change; so are the data entities that represent them. That can create confusion and headaches.

For example, if you implement a data dashboard to work with the latest organization hierarchy definition, what will happen when you need to look at historical records when the company had a different organizational hierarchy structure? Nothing can destroy the credibility of a reporting or analysis initiative faster than not being able to explain the numbers you get back.

Source data is not a fix-it-and-forget element. Managerial and organizations changes are not just a problem in past data. Organizational changes are continually happening. So, you have to continue to be on the lookout for changes that are occurring anywhere in the company: reorganizations, managers changes, acquisitions, divestitures, and name changes can happen at any time. How do you recognize when you are getting bad results? And if you do identify a

problem, can you trace it to its source? Imagine if one of the operational data sources feeding your data environment breaks and a field that your extract, transform, and load (ETL) process uses as a primary key fails to come through. Suddenly, everything is wrong on downstream reports. If there is no data check alert designed in the data process, you'll have no way of knowing that this problem is happening until someone opens up a report in a meeting and says, "this is wrong." A meeting is not a place where anyone wants to find a problem they can't explain.

You might be surprised to learn this, but most executives only know the units they manage. This means they have no visibility into the complexities of the company beyond this and no awareness of the history of changes across the groups of the company for which they do not have direct responsibility. It may be that nobody before has attempted to report on all of the units of the company as a complete, coherent picture over time. You also may find that some divisions define organization units by the manager, some by financial cost centers, others by location, and others by something else entirely. You have to facilitate agreement on a method of arranging organizational structure to report on the company and its many pieces that represent how the company operates, but that also maintains data integrity and efficient data processing as a whole.

On top of the problem of complexity and change endemic to people, you also have system complexity. Most companies use different systems for applicant tracking, human resource information systems, performance management, compensation planning, and other transactional human resource management needs. Even if all of those transactions are performed in one system, you will still likely have several payroll and benefits partners, which each also have their systems to contend with. Different system owners apply different business rules and often take no interest to maintain data entry and corrections for data elements that have little need for. Since their job is transactions, they are on to the next one – they could not care less about the data for those they have completed in the past.

Piecing together the data picture as a whole – gathering the entirety of current business rules and naming conventions — is a challenging non-trivial exercise. For this, three pieces of advice are offered.

- Pick a topic that matters and start slowly. Make sure you can do a few things well before you sign up to do more things.
- Plan ahead for confusion and change by creating visibility and flexibility in our master set of metric definitions, data hierarchies, and data relationships.
- Whenever planning for a project, you need to include time to audit, analyze, and (if necessary) fix the data at the source if at all possible.

Perhaps, most importantly, you need to carefully pick the metrics and analysis for which you are going to focus, based on what will offer the company the most value. If you try to mine the whole of HR data to hope to find something, you are going to end up in a world of pain fast. It is just impossible to keep everything perfect all the time and even if you could do that, it will be a waste because most of the data you have perfected will go unused anyway.

Pitfall 9: Skimping on New Data Development

Data is the lifeblood of people analytics: Transactional HR systems are designed to speed up operational HR processes, and these systems, by coincidence, produce some data. Data from transactional systems is necessary for people analytics. However, data produced by these transactions is limited to only a specific range of insights. Failing to understand this fact will limit the value you can derive from people analytics. Here are a few tips that will help you create new data to crank out new insights.

Budget for new data development: Neither HR or IT (or even specialized Analytics departments) do an excellent job at budgeting for investment in collecting new data or in acquiring a subscription to new external data sources. Consider having a slice of your budget set aside just for new data development. You might include sending people to conferences to look for ideas for new data sources or interview other people working in the field of people analytics about the data that have proved most valuable to them. Regardless, have a budget – and use it.

Never miss a chance to get relevant new data: Creating data is something to keep in the back of your mind at all times. Since there are so many different ways to approach this, some examples are listed as follows: If you have people interview candidates and make some decisions about them, why not record this information in a structured way so over the long-term it can be analyzed too? If you have a new hire fill out a form for some information required by rule of law, why not also collect a few pieces of information that will be valuable to you for analysis reasons too? If it takes them only 30 seconds longer to provide you with a few additional details and they already had to stop what they are doing – why not? You have employee address for other reasons, why not use this data in conjunction with a public data source to calculate commute time? Speaking of open data, did you know that the US Department of Labour maintains a database of the number of people in a particular zip code by job classification by ethnicity and by job classification by gender? You may talk about diversity, but do you know if the percentage of people you hire by job classification is proportional to what you can expect in the population? Did you even know you could do that? When you deliberately think about the types of data you can use in people analytics you will find there is an endless variety of surveys, personality inventories, tests, subscription benchmark data sources, and so on, which represent new data sources that can augment your existing people analytics efforts to produce new insights.

Reinvest time: By tradition or lack of imagination, most of the investment in people analytics capability is focused on producing more efficient outputs. For example, data warehousing tools, data management tools, and data visualization tools are all geared towards more efficiently creating insights that people have previously produced manually in inferior applications, like Excel. The movement to more scalable and efficient analytics systems is valuable in itself. Still, it cannot by itself produce new insights if it is merely answering the same questions with the same data. You should take some of the time or money saved by moving to more efficient tools and set it aside for new data development. Look at the time people used to spend in Excel before implementing the new automated reporting environment for the same metrics. Allocate some of the time or money saved to obtain or design new data collection

instruments that can produce new metrics. Deliberately putting time and money into developing new data sources to go into your people analytics effort will drive more value out of your people analytics effort in the end.

Pitfall 10: Not Getting Started at All

The list of possible pitfalls for people analytics is long, as is the list of excuses used to not getting started.

Anything worth doing requires hard work and carries a certain amount of risk. What you have to decide is whether you believe that your company can make better people-related decisions continuing with the old way or whether trying to make things better with data is more likely to give to predictable repeatable success.

The reality is that both options come with pitfalls and risks. Regardless of whether you choose to use or ignore data in your decision-making, your company can't avoid making decisions. The part you can control is how well you equip yourself to make good ones. People analytics will serve you well, but to do so, you must get started. So, get to it.

2.7 AREAS OF HR ANALYTICS

The common areas that need to be tracked by managers in HR.

1. Employee Turnover

This is probably the most basic and the most critical area that needs to be tracked by every HR professional. Hiring employees, training them, and then finally inducting them into the workforce as professional's needs resources. That's why leaders need to track the churn rate of their organization and compare it with the industry standards.

Understanding **employee turnover** rates also help HR professionals to predict future exits from the organizations. These insights help avoid a few entries and prepare the organization for the unavoidable ones. Managers can look at previous data, KPIs such as employee engagement level, employee satisfaction scores, and staff advocacy score to identify and minimize employee turnover.

2. Recruitment Analytics

Recruitment analytics is probably the second most popular HR analytics area after employee turnover. Managers need to identify the channels and places that optimize recruitment. They need to determine and track from where the talents are coming and what worked with the last batch of successful hires and try to do the same.

Recruitment analytics is an offset of employee turnover analytics. A great manager identifying patterns in the recruitment data can decrease the turnover rates of the organization.

3. Performance Analytics

Every organization needs high performers in their team for their business to grow and thrive. So, it becomes extremely important to invest in **employee performance analytics** to assess individual employee performance.

Managers can identify the top performers and provide assistance or training for the employees who need to up their game. This makes it easier for employees to max out their performance and deliver competitive results.

But focusing on one aspect of a job is usually not desirable for optimum analysis of performance. For example, focus on only the sales numbers. However, it is the most crucial aspect of a sales executive. It might not be the ideal way to get a complete insight into the employees' performance. Modern data capture techniques make it possible to analyze performance holistically. This is ideal because focusing on a tiny part of a job might lead to employees skewing the results in ways likely to put themselves in the limelight.

4. Employee Engagement Analytics

This is one of the essential and most difficult areas to measure. Employee engagement is a critical criterion in determining the engagement and productivity of employees in a workforce. **Disengaged employees** tend to be less productive and tend to bring a few more people down too. So, it becomes essential to have proper tools and techniques at your disposal to analyze **employee engagement** levels properly.

But how do you measure something as subjective as engagement? One of the obvious answers is surveys. But traditional surveys aren't going to do the job. Traditional surveys tend to be dry, and employees don't put thought and effort into giving proper feedback through the survey.

What comes to the rescue is the new age pulse surveys. Pulse surveys are quick, frequent, and short survey systems away with complex questions and aim to increase feedback levels. It is usually done weekly or once every few weeks. It gives a proper idea of the health of productivity and engagement in the workforce, thus the name "pulse."

Having a proper insight into the employee engagement of your workforce can be a crucial differentiating factor when it comes to the success of your organization.

5. Corporate Culture Analytics

Corporate culture is probably the only area that is more difficult than employee engagement to collect and analyze data. Corporate culture is basically the both, the explicit and unspoken rules, systems, procedures, and patterns of behaviour in your organization. This subjective nature of its important aspects makes it extremely difficult to collect corporate culture data and analyze them.

One way to assess corporate culture is through the analysis of the formal internal **conversation** and also the customer service conversations. This provides a great deal of data

about corporate culture. One more way is through specialized feedback and the participation of the same. Using unconventional techniques becomes handy in identifying corporate culture data.

2.8 APPLYING HR ANALYTICS

- **Revenue per employee**, *i.e.*, calculating the average revenue generated by the employee.
- **Offer acceptance rate**, *i.e.*, calculating the rate at which a firm's job offers are accepted during a certain period.
- **Training expenses per employee**, *i.e.*, calculating the expense extended by the firm on each employee's training.
- **Training efficiency**, *i.e.*, calculating the efficiency of employees post training by assessing their performance and test scores.
- **Voluntary turnover rate**, *i.e.*, calculating the rate at which employees exit the organization voluntarily to identify the gaps in employee experience.
- **Involuntary turnover rate**, *i.e.*, calculating the rate at which employees exit the organization involuntarily to implement changes in the recruitment strategy.
- **Time to fill**, *i.e.*, calculating the number of days between advertising for a job opening and hiring someone to fill it.
- **Time to hire**, *i.e.*, calculating the number of days between approaching a candidate for a job offer and the candidate accepting it.
- **Absenteeism**, *i.e.*, calculating the number of days an employee is absent from work to get an insight into the overall employee health and happiness.
- **Human capital risk**, *i.e.*, calculating the employee-related risks such as lack of a particular skill to fill a new type of job.

2.9 TOOLS FOR HR ANALYTICS

There are different tools for HR Analytics which are as follows:

- **Optimize the Interview Process with HireVue Insight:** At one time, hiring was fairly subjective. Committees would sit down, listen to anecdotal interviewer analysis of candidates, and decisions would be made for better or worse. HireVue Insights does away with all of that. Using Iris-HireVue's patented data engine, interviews are analyzed for over 15,000 interview attributes and candidates showing characteristics similar to the business top performers are recommended based on analysis.
- **Better Define Your Benchmarks with Bersin by Deloitte:** While not a tool so much as a membership, Bersin by Deloitte offers access to industry leading insights (provided by Bersin) that are backed by the power and accuracy of Deloitte's research. With access to these insights, the company will have access to optimized models of

success, frameworks, and reasonable growth benchmarks designed to increase the business impact of the HR department.

- **Increase Retention of Staff Learning Initiatives with SkillPort:** Though we may all have certain affection for the employee training videos, we have seen 1,000 times and the outdated, awkward learning modules developed for us in the early 2000s that “demonstrate employee understanding of HR-based learning initiative”, the truth is that they are entirely ineffective. A quick survey of education and educational psychology over the last 30 years confirms the fact that not everyone learns the same way, and that one-size-fits-all attempts at education fail. That’s where SkillPort comes in. By offering a variety of learning modalities and connecting learners across social and community portals, SkillPort increases the likelihood of information retention and provides concise, actionable feedback to administration.
- **Intelligently Manage the Talent with Saba:** Saba allows employees to create profiles, connect with one another, and access a variety of educational materials. But that’s not where the magic happens. Instead, it happens in Saba’s intelligent core, which tracks all employee interactions, document views, etc., and uses the data to intelligently recommend reading to further professionalize each employee on a personal level. Add to that Saba’s ability to take all the data it has and make recommendations about which internal employees would be a good fit for new positions, successors, and even what adequate compensation would be given their skillset, and company have a truly invaluable analytic tool.

2.10 BENEFITS OF HR ANALYTICS

1. Improved Hiring Decisions

HR analytics helps HR professionals make better choices based on historical data. Through HR analytics, the recruiter and hiring team can make data-driven hiring decisions rather than “trusting their intuition” or “gut” and hoping for the best.

An effective HR analytics tool can enable HR professionals to easily determine the best candidates and hire them. For example, if you hired 10 candidates and 4 out of them belonging to a particular university could not qualify beyond a level, perhaps you will not hire someone from the same university again.

HR analytics prevent you from making such mistakes. Moreover, it also allows the recruiter to learn more about candidates through online databases, applications, social media profiles, job portals, records, etc.

2. Reduced Talent Scarcity

Structured and stat-backed recruiting strategies provide the insights needed to cast a more accurate hiring net to ultimately build a better pipeline. With the help of HR analytics, it is easier to know whom to target for a particular position when a requirement arises in the company.

It is easier to understand the type of person required, the knowledge required, along other aspects that might be missed in a regular talent acquisition process.

When talent is scarce for a specific industry or role, it might seem that talent isn't there at all. The talent is there but finding it would require you to employ new methods of attracting or searching for it.

3. Improved Hiring Process

Over a period of time, once the company has enough data to break down into patterns, the company can then use it to make its recruitment process quicker and efficient.

Hiring managers will know which positions in the company require what type of talent, people, background, and experience, along with the type of people that will not be suitable. Then it becomes much more elementary to target a specific talent pool.

4. Good Training

Training is a necessity for any organization to constantly grow. Gaps in the process of training can increase the cost to the company and give rise to unwanted challenges. Various professional development programs offered by companies, if not handled properly, can deplete a major chunk of the company's budget.

That is why it is highly suggested that enterprises invest in training by first collecting big data and learning from the patterns. HR analytics can enable enterprises to buy or implement accurate training programs.

It can help organizations to identify if employees are making full use of the opportunities and knowledge given to them during training programs and if it is even relevant for them or not.

5. Better Employee Insights

HR analytics help the company look at an employee's professional life by tracking, sharing, and analyzing performance-related data. This is why companies track and record the behaviour of their employees with customers, co-workers, and how they spend their time.

Furthermore, the employee's performance data could be used by the hiring manager to identify great talent. This data not only provides more insights about the employees but also shapes the strategies to boost employee morale, retention, and engagement.

6. Reduced Attrition Rate

HR analytics can easily help identify and demystify the causes and patterns of attrition among employees. It tells you the reason why employees leave and why they stay.

The reasons could range from under-performance, lack of skills, overworking, to low compensation, management issue, unsuitable work environment, and more.

Understanding the exact reason behind for employees leaving your organization sooner than expected can help you rectify the challenge and connect with your employees better.

HR analytics helps in finding the gap areas where employees are finding difficulty in meeting goals and targets. HR professionals can use various methods like employee surveys, team assessments, exit and stay interviews to find out the reason for attrition and map out strategies to retain them.

7. Improved Candidate Experience

Recruitment analytics help companies understand exactly what kind of talent they are looking to hire. Since this understanding is backed with credible data, this puts those organizations way ahead in the game to narrow down and attract the right talent in a shorter span of time.

A shorter hiring process automatically gives a good impression to candidates and gives them a better hiring experience. Simultaneously, this helps improve your brand image, giving your enterprise an edge over competitors.

8. More Productive Workforce

Since enterprises have a grip on the kind of people to hire for a job, it becomes easier to expect and better predict performance and productivity from employees.

With analytics, you get an understanding of the workforce and, hence, know what kind of environment, policies, and teams will make them work hard and optimize company performance.

2.11 USES OF HR ANALYTICS

Human Resources Analytics has become an umbrella term that encompasses a wide range of activities and processes. HR Analytics are used in various ways and forms.

- **Reporting:** A substantial amount of effort in the study and practices of analytics has focused on reporting. Reporting incorporates decisions about what analytics will be reported; how analytics will be packaged and how, when and whom they should be reported?
- **Dashboards:** Dashboards are enriched component of reporting. Dashboards contain business unit analyses to permit managers to drill down to examine analytics on several levels of the organization.
- **Data Mining:** Data mining refers to efforts to identify patterns that exist within data and that may identify unrecognized causal mechanisms that can be used to enhance decision-making. To identify these causal mechanisms, data mining used correlation and multiple regression methods to identify patterns of relationships.
- **Better Problem Solving and Decision-making:** In organization, decisions result in tactical decision. Making these decisions require understanding the outcomes that one is attempting to achieve, understanding the factors that influence those outcomes and their current states, and knowing available tactical options and their costs.

2.12 CHALLENGES OF HR ANALYTICS

- **Lack of Data Analytics Skills:** The organization often comes across predicaments while attempting to find people with the right skill set for the gathering, managing, and reporting of data.
- **Data Cleansing:** It is crucial for data to be collected, cleansed, merged, and evaluated from numerous departments as well as a multitude of business functions, making the need for experts who can properly gather and organize the data alongside evaluating it highly imperative.
- **Excess of Data:** Often, excessive data is collected leading to too much data to parse and a lack of clarity on which data is more imperative.
- **Data Privacy and Compliance:** As data is collected regarding an employee or potential employee, the privacy of the same needs to be kept in mind as violating the privacy of its employees could get the company in trouble.
- **Deciding the Tools:** There may be a lack of clarity while identifying the most suitable HR technologies for keeping track of the data.

2.13 THREE WAVES TO CREATING LASTING VALUE

Lasting Value

Lasting value creation means your organization starts with a cycle that leads to success.

We speak of lasting value if the value is created for all stakeholders:

- Shareholder value – higher profit and a better price earnings ratio.
- Customer value – better service, greater customer satisfaction, better connection between the customer and the organization, and higher profit per customer.
- Employee value – employees connecting better with themselves, with colleagues and the organization in relation to better performance.
- Social value – connecting better with society and making a contribution to society.

Organizations with this approach to stakeholder value perform in the long run 10 times better than the average!

But, creating value for all stakeholders is not easy. On the contrary, it requires a great deal of work on various fronts. In the book, *Creating Lasting Value*, the authors Jeroen Geelhoed, Salem Samhoud and Nur Hamurcu identified the three waves that lead to creating lasting value.

The First Wave: Lead the Value

The first wave concerns the inside of an organization. When you enter an organization that has incorporated this first *Lead the Value* wave, you can sense it immediately. You notice it when you speak to employees. You notice there is energy, a sense of urgency or a sense of excitement

about getting down to work and doing things together. There is also a clear and inspiring vision and strategy that gives people meaning and direction.

As a result, the employees work with inspiration and passion and have their customers' interests at heart. The leadership is characterized by ownership, charting the course and demonstrating solidarity with the organization.

Such organizations also have a well-designed business model and a clear target group. This target group has a clear offering and an intelligent revenue model providing a basis for economic success.

Organizations that excel in *Lead the Value* have a healthy culture, which on the one hand provides unity but at the same time harnesses the power of complementarity and diversity.

It is logical that the first wave is at the center of the model. After all, it generates the other waves. Yet the first wave alone is completely insufficient to create sustainable value, because results must also be delivered.

Check your organization on the first wave, *Lead the Value*.

- Is there energy, a sense of urgency and excitement about creating value for all stakeholders?
- Is there an inspiring vision?
- Is there a clear strategy?
- Is there a sound business model?
- Is there a management team that takes ownership?

The Second Wave: Manage the Value

The second wave of value creation relates to “make it happen”: executing the strategy and demonstrating performance.

Organizations that are good at *Manage the Value* show results. They score highly customer satisfaction rankings. Customers recommend them and employees have high productivity. Such organizations make a profit, year after year. These organizations are also able to improve themselves and offer new products and services.

The second wave concerns an organization's execution power. *Manage the Value* is about living the vision in day-to-day life, but also bringing in the voice of the customer, conducting performance dialogues and optimizing value streams.

And, last but not the least, by empowering entrepreneurship within the organization. Check the organization on the second wave: *Manage the Value*. How does your organization perform on:

- Customer satisfaction or Net Promoter Score?
- Employee Engagement and productivity?
- Profitability?

- Having frequent performance dialogues with the teams on all levels of the organizations?
- Living the vision and values in daily practice?
- Bringing in the voice of the customer?

The Third Wave: Market the Value

Market the Value is about showing and sharing the value to create and celebrating success. It is about inspiring others.

We make a distinction between two mainstreams: Showing yourself and Sharing your value.

Show yourself involves branding to track record, publicizing what you stand for, being open to others and participating in rankings. It is important that you show the world the progress you have made and celebrate it. The world must be able to see it and experience it! It will also inspire others to build a better future.

2.14 HR MEASUREMENT JOURNEY IN TUNE WITH HR MATURITY JOURNEY

HR Measurement Journey

HR Measurement are metrics that track actively, Metrics are simply measurements. Human Resource (HR) metrics are measurements used to determine the value and effectiveness of HR initiatives, typically including such areas as turnover, training, return on human capital, costs of labour, and expenses per employee.

For example, “the company average engagement level is 80%.” “Annual turnover is 50%” “average performance score is 60%.” Metrics alone do not show what affects engagement, what causes turnover and what drives performance.

Human resource analytics examine the effect of HR metrics on organizational performance. In more general terms, analytics look for patterns of similarity between metrics.

For example, do high performances leave at a higher rate than low performers and if so, what is driving that turnover? Statistical tests are necessary to get to those answers through analytics.

HR Maturity Journey

As an organization begins to transform, understanding the current maturity level is crucial to creating better employee experiences. This can be done by relating technological adoption and integrating learning and management processes.

Learning and Development are the two distinct processes and a critical function in the HR domain. It plays a crucial role in attracting, building and retaining talent. Hence, it has become essential for HR professionals today to integrate processes, people flows and systems in order to deliver a significant impact to business goals.

Level 1: Compliance-driven HR Services

Compliance-driven HR Services is the Level 1, the primary level of maturity.

The “compliance-driven HR” organization focuses primarily on fulfilling mandatory HR requirements. Leadership is often shared with an Operations or Finance role, and the philosophy concentrates on the local legal requirements concerning hiring and employment. These organizations have no documents have no documented HR strategy and line managers assume many of the responsibilities for recruiting, managing and motivating employees.

Level 2: Fundamental HR Services

At Level 2, a “fundamental HR services” organization has an HR strategy focused on improving processes and creating HR depth. All efforts are made to make the Core HR services standardized and efficient. In addition, the organization begins to focus on talent management services, such as recruiting, succession planning or training. These organizations typically have organized HR structures for functions, such as recruiting, development, compensation and workforce planning, with separate, automated systems for these processes.

Level 3: Strategic HR Department

At Level 3, a “strategic HR department” is defined by its approach to business partnership. These organizations are focused on creating deep connections with the business goals, while maintaining an optimized HR department. Leadership is seen as a critical attribute of the executive team. The focus for this level is on integrating processes, data capturing and business advising. This function includes more strategic talent management, as well as flexible frameworks for core HR functions that meet both legal and business needs, some core HR services may be outsourced or automated, while strategic HR functions are improved. System integration and communication are critical to this level.

Level 4 – Business-integrated HR

At the highest level of maturity (Level 4), the “business-integrated” HR organization helps to drive the business through workplace strategies and people data. These organizations are focused on optimizing the workforce to meet business needs, and offering insights that will transform business goals.

The HR leader has a strategic role in the organization and is closely connected with business executives. HR systems are tightly linked connected to business systems, enabling HR and business leaders to access the necessary people data to make decisions. Major initiatives at this level involve data analysis and trending, providing business guidance, and continuous improvement of HR processes.

The continuous improvement aspect is key among business-integrated (Level 4) HR organizations. Even though these organizations have reached the highest stage of maturity, they are always in a constant state of evolution. The focus is on continuous adaptation, ensuring that the HR structure, team, programs and systems are optimized and aligned with the needs of the business.

Every Lean organization is unique, but there are some core qualities that they tend to have in common.

- **Focused:** Lean organizations are obsessed with customer value. They recognize that, by establishing customer value as the number one priority for every employee, manager, and executive, other business metrics will fall in line. That's why every decision involves an assessment of how it will impact the customer.
- **Aligned:** In a Lean organization, every member of the team understands the mission, values, and strategic priorities of the business. This is achieved through open lines of communication between frontline staff and senior leaders, and the development of a culture in which individuals are empowered to help the organization achieve those goals. This results in an army of people focused on improving the organization.
- **Humble:** One of the central tenants of Kaizen, the philosophy at the heart of Lean, is that there is always room for improvement. Even when the organization is doing well, leaders encourage people to find ways to drive further success. After all, continuous improvement is not something you can ever finish. A Lean organization recognizes that there are endless opportunities to improve.
- **Collaborative:** Silos have no place in the practice of Lean. Cross-functional collaboration and communication are necessary to tackling the biggest challenges. Organizations that provide people with the tools they need break down silos and share knowledge and experiences across the organization drive company-wide improvement.
- **Tenacious:** Lean organizations do not opt for workarounds or simple answers to complex questions. They seek out and address root causes by empowering the people doing the work to speak up, identify problems, and improve them. Lean organizations are not interested in applying band-aids and short-term fixes.
- **Engaged:** Lean is a business process methodology that leverages the skills, input, and observations of every employee. Employees who are invested in the success of the organization and engaged in improvement are essential. There is a ton that leaders can do to increase their employee engagement in improvement, and we write about it a lot on this blog. Be sure to subscribe at the bottom of this post to ensure you don't miss anything.
- **Methodological:** Improvement work is most effective when intentional practices such as PDSA, A3, Catchball, and Hoshin Kanri are applied. Whatever methodology your organization has chosen to propel continuous improvement, it is important to remember that it must be simple, habitual, and consistent across the organization. The methodology behind your improvement work is what drives long-term success in a Lean organization.
- **Proactive:** Lean organizations inspect processes, not products to prevent problems and waste before they happen. How do they do that? By engaging employees in the work. Managers don't oversee every process, so while they might notice failures in the end

product (or service), it is up to the people doing the work to notice ways to improve the process. Lean organizations have successfully created a culture in which employees feel safe speaking up and taking the time to make improvements along the way.

- **Documented:** Standard work is documented, accessible, and continually reviewed in Lean organizations. This creates uniform processes across the organization, and allows everyone to start at the same elevated level as they improve, rather than constantly reinventing the wheel. It is important to point out here, though, that just because a process is documented does not mean that the improvement process is over. It just means that the current best practice has been identified and shared; in a Lean organization, people are expected to use this as the starting point for future improvement.
- **Resilient:** Resilience in Lean organizations is the ability to anticipate trouble spots and improvise when the unexpected occurs. The organization must be able to identify errors for correction while at the same time innovating solutions. This is done by empowering employees to make autonomous decisions whenever possible, so that they can respond with speed and agility.
- **Progressive:** Most organizations that successfully apply the Lean way realize that the improvement process requires the support of enabling technology. After all, there are a lot of people working together on improvement, and a limited amount of time on which to work on improvement. To ensure that they get the most done in the least amount of time, they reduce their administrative and communicative burdens by relying on continuous improvement software solutions that help move the improvement process forward with alerts, notifications, and improvement impact reporting.
- **Grateful:** In order for Lean culture to take hold, employees must feel valued and recognized for their contributions to improvement. Smart leaders realize this and broadcast employee successes at every opportunity.

2.15 THE FIVE LEAN PRINCIPLES EXPLAINED

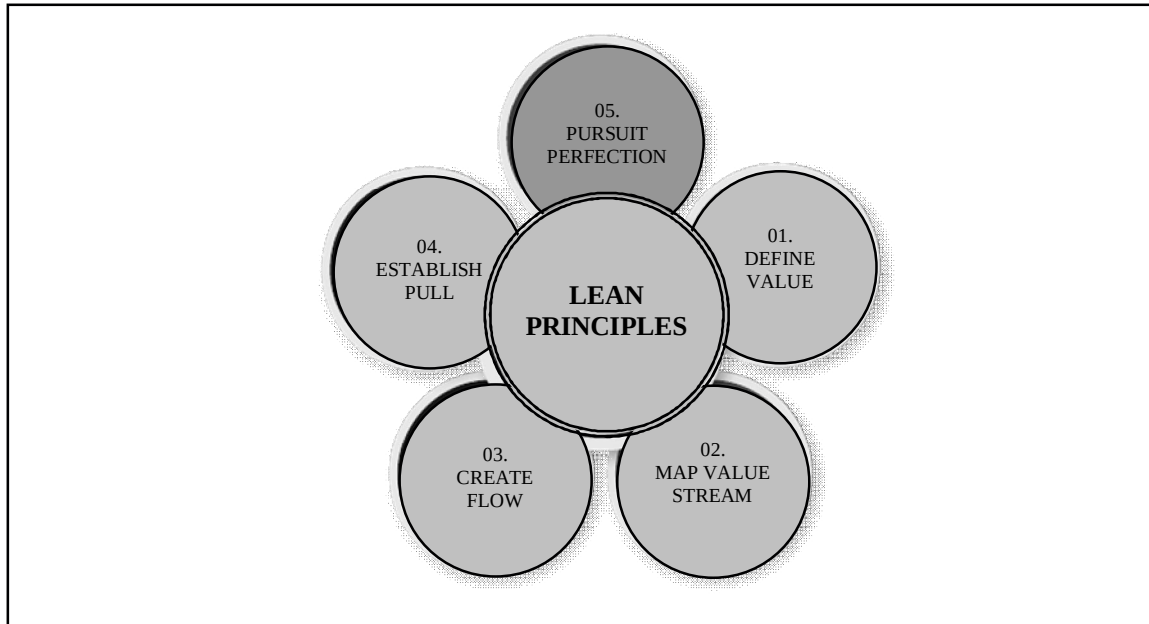


Fig. 2.2: The Five Lean Principles

1. Define Value

To better understand the first principle of defining customer value, it is important to understand what value is. Value is what the customer is willing to pay for. It is paramount to discover the actual or latent needs of the customer. Sometimes, customers may not know what they want or are unable to articulate it. This is especially common when it comes to novel products or technologies. There are many techniques such as interviews, surveys, demographic information, and web analytics that can help to decipher and discover what customers find valuable. By using these qualitative and quantitative techniques, you can uncover what customers want, how they want the product or service to be delivered, and the price that they afford.

2. Map the Value Stream

The second Lean principle is identifying and mapping the value stream. In this step, the goal is to use the customer's value as a reference point and identify all the activities that contribute to these values. Activities that do not add value to the end customer are considered waste. The waste can be broken into two categories: (i) non-valued-added but necessary and (ii) non-value and unnecessary. The latter is pure waste and should be eliminated while the former should be reduced as much as possible. By reducing and eliminating unnecessary processes or steps, you can ensure that customers are getting exactly what they want while at the same time reducing the cost of producing that product or service.

3. Create Flow

After removing the wastes from the value stream, the following action is to ensure that the flow of the remaining steps run smoothly without interruptions or delays. Some strategies for ensuring that value-adding activities flow smoothly includes breaking down steps, reconfiguring the production steps, leveling out the workload, creating cross-functional departments, and training employees to be multi-skilled and adaptive.

4. Establish Pull

Inventory is considered one of the biggest wastes in any production system. The goal of a pull-based system is to limit inventory and work-in-process (WIP) items while ensuring that the requisite materials and information are available for a smooth flow of work. In other words, a pull-based system allows for just-in-time delivery and manufacturing where products are created at the time that they are needed and in just the quantities needed. Pull-based systems are always created from the needs of the end customers. By following the value stream and working backwards through the production system, you can ensure that the products produced will be able to satisfy the needs of customers.

5. Pursue Perfection

Wastes are prevented through the achievement of the first four steps: (i) identifying value, (ii) mapping value stream, (iii) creating flow, and (iv) adopting a pull system. However, the fifth step of pursuing perfection is the most important among them all. It makes Lean thinking and continuous process improvement a part of the organizational culture. Every employee should strive towards perfection while delivering products based on the customer needs. The company should be a learning organization and always find ways to get a little better each and every day.

2.16 HR DECISION-MAKING

HR decisions are characteristically different from non-HR decisions in organizations, both in terms of its scope and complexities. Cases of organizational failure are largely attributable to wrong HR decisions. A good example is decisions on talent recruitment and selection. Despite adoption of multiple approaches for making the selection process in true sense a successive hurdle technique, we observe selected talent become unproductive, unfit for team work, unable to integrate with the culture of the organization and so on. Those who are the right-fit decide to quit the organization unpredictably, causing increased cost of talent replacement.

Prima facie a decision-making process is our judgement for any action. An HR decision-making process, therefore, can be defined as HR managers' judgemental thoughts on an action. As most of the HR decisions are encompassing all stakeholders of the organizations, it is always prudent to reinforce HR decisions with supportive data and information. Obviously, for this reason today, we observe increasing acceptance of HR Analytics as an important tool for HR decision-making.

An effective HR decision-making process requires the integration of critical thoughts and information. HR managers develop critical thoughts using their professional skill and competence,

while business information are made available through Big Data Analysis. Critical thoughts of HR managers are also moderated by evidence-based HR decision-making (Rousseau and Barends, 2011). In the process of emulating best practices, HR managers get influenced by organizational experiences, more specifically in terms of cognizance to decisional consequences. Data-driven HR decision-making is in practice in organizations for quite some time. In fact, researchers such as Locke (2009) and Charlier, *et al.* (2011) could observe that an HR decision-making process, across organizations, has the richest scientific evidence. This signifies that organizations have already embedded a data-driven HR decision-making process, for their sustenance. Evidence-based HR decision-making, in its current stage of development, signifies factoring of managerial practices and decision-making with couple of factors such as evidence-based indicators and metrics, procedural justice and ethical issues, stakeholders' interests and so on. In all such cases, the use of Big Data and HR Analytics can help.

HR managers largely feel constrained in their decision-making process for couple of reasons, such as organizational systems, regulatory influences, social influences and stakeholders' responses. Most of the HR decisions in organizations often neglect the stakeholders' interests, resulting gradual decline in their performance and brand value. For example, 'say on pay' is a critical issue for shareholders (Bhattacharyya, 2013). In India, it is yet to gain popularity, but in Europe, America and a couple of other industrially advanced countries, this issue is gaining importance. 'Say on pay' means exertion of shareholders' right to vote against or for the proposed executive compensation packages of organizations, which have been approved by the compensation committee, and put up for vetting before the shareholders' annual general meeting. Shareholders today are more informed for the information availability from various shareholders' advisory services. Even though shareholders' voting effect is non-binding for the company, but when they vote against, companies dissuade them to implement the increased compensation package, fearing decrease in market capitalization and dilution of their brand value.

2.17 IMPORTANCE AND SIGNIFICANCE OF HUMAN RESOURCE ANALYTICS

For organizations, HR Analytics provide scientific facts to human resources. Reviewing the framework of HR analytics, introduced in several organizations, we find that it starts at the outset with the identification of key business concerns, focus on strategy and long-term sustainability issues. After this, we go forward with assessing the available knowledge and competencies' set in the organizations. Gaps, if any, need to be met through appropriate interventions before we proceed to the next step, *i.e.*, the analysis of organizational data. At the end, results are communicated, examining which, HR managers decide for appropriate interventions, well in advance, to improve the decisional outcomes or to reduce the adverse effect of decisional outcomes. This is a working model of HR analytics.

For operational ease, in many organizations, HR managers at the beginning start with descriptive analytics, *i.e.*, the use of metrics, to understand the current syndrome; gradually, HR managers start the use of statistics and research methodology for better analysis of decisional outcomes; and finally, HR managers use HR analytics to predict how HR decisions influence

business results and align with the strategies of organizations. With prediction results, HR managers can alter or improve HR decisions.

Although in the corporate world, it is largely believed that HR analytics can significantly alter and improve HR decision-making, and bring holistic change in organizations, some HR managers still believe that such assumptions are more overpitched. With standalone use of HR analytics, HR managers believe they can influence the HR decision-making process, establishing a link between HR decisions and employees. More appropriately, HR Analytics help in diagnosing HR issues, which can facilitate in taking prevention actions through appropriate interventions.

We can understand the importance and significance of HR analytics, once we create a business case, and then analyze how it can facilitate in holistic decision-making. Maersk, a Denmark-based large conglomerate, made use of HR analytics to study how its employee engagement programmes can enhance managerial commitment, improve training effectiveness and workplace safety. For Maersk, workplace safety is a business-critical issue.

Why HR analytics is important for organizations? We do not have any simple answer to this question. But reviewing organizational experiences, when we document the typical trends in HR decision-making process, we can truly understand the significance of HR analytics. Review indicates that, in most of the organizations, HR decisions are made in silos.

When decisions on investments in people are made in silos, and even without any data, organizations may fail to understand the benefits of investments, like the RoI on training.

However, all HR decisions are not measurable in terms of business outcome. A good example of this is employee engagement. Hence, if a company feels that its investment in employee engagement will enhance the sales revenue and expect HR analytics to provide such information, it would perhaps be more aspirational. A better way to look into this, perhaps, is measuring the change in the attitude of the employees. It is like improved employee engagement can contribute to the development of positive employees' attitude, which leads to enhanced sales revenue as sales improve with improved customers' relationships. Hence, the significance or importance of HR analytics needs to be understood in terms of the analysis of its effects on the HR decision-making process.

Again, the influence of HR decisions is very much culture-specific. Hence, what could be seen in an organization need not be universalistic. Thus, if we feel what Google could do with HR analytics, we also can, it is certainly not logical and realistic. Therefore, the other significant characteristic of HR analytics is that it is time-consuming, as connecting people data with business may not yield quick results. HR analytics require cross-functional knowledge. Today's HR managers need to understand that interpreting people data in silos take them nowhere. Rather, people data require interpretation in alignment with other functional areas of organizations. With HR analytics, it is also possible to fix responsibilities and make people accountable for their work.

2.18 CRITICAL HR DECISION-MAKING AND HR ANALYTICS

In critical HR decision-making processes, namely talent acquisition or recruitment and selection, many organizations come out with their detailed decisional guidelines. Such guidelines ensure decisional consistency and so also can immune HR decision-makers from any future liability, including legal hassles. Moreover, recruitment and selection are serious HR functions, first because of choosing the right candidate and then for the need of predicting their retention probabilities. Hence, from organizations' point of view, recruitment is investment.

What could be the appropriate decisional guidelines for the decision-making largely depends on the organizations, the level of recruitment, the prevailing legal norms and restrictions, and so on. A tentative decisional guideline for the recruitment of a hypothetical organization for managerial positions can be listed as follows:

1. Ensure your recruitment decisions are impersonal, and for the best interest for the common good and justice.
2. Ensure your recruitment decisions are fair, objective and transparent.
 - Stereotyping
 - Prejudice
 - Mirroring
 - Laying emphasis on personalities, rather than candidates' capabilities
 - Decisions are not holistic (based on assessment single or some attributes)
 - Spur of the moment decisions
 - Decisions vitiated by halo and horn effects
 - Poor verification of antecedents

The earlier described list of don'ts in recruitment for managerial positions can to a great extent ensure appropriate recruitment and selection processes of any hypothetical organization. Most of the parameters can be objectively measured and assessed, and their effects on the organizations can be predicted using HR analytics.

2.19 HR ANALYTICS AND CHANGING ROLE OF HR MANAGERS

Characteristically, human resources today widely vary from the past. Today's human resources are diverse, tough and look for more challenging workplaces. The changing pattern of human resources obviously redefined the role of HR managers in organizations. For HR managers, additionally the critical challenge is managing talent. Attracting, developing and retaining talent are not only challenging for HR managers, the requirements, along with other strategic issues, today literally changed the dynamics of HRM functions in organizations. For effective talent management alone, getting answers to questions such as why people leave an organization within a short time, who will be the top performers, what training will be required, and who will benefit from training and so on is very important. To get answers to these questions requires HR

managers to first develop critical insights and then analyze the data and information using HR analytics. HR decisions, being more strategic with HR analytics, HR managers today can predict the trend on critical HR issues, and accordingly, can design their interventions to avoid any catastrophe. Hence, today's HR functions are no longer transactional; rather these are more data-driven and involve detailed business analysis for better HR insights and predictive HR decisions for strategic HR outcome.

Today's HR managers, therefore, are required to understand which data can give them better HR insights. HR managers also need to understand the availability of such data inside and outside organizations, the integration of such data for developing actionable insights and then making use of these data for strategic and predictive HR decision-making process. With HR analytics, HR managers can take operational HR decisions at the highest level of accuracy and precision. Also, HR managers can create his/her dashboards for better analytical, predictive and strategic decision-making. Depending on the level of IT knowledge, HR managers can create simple dashboards such as bar charts, pie charts, frequency tables and so on or more complex dashboards like scatter plots. Whether simple or complex, dashboards when linked with organizational processes and performances, can give better insights into decision-making.

The nature and extent of HR decision-making process that can be facilitated by HR analytics depend on HR managers. However, common HR decision-making processes which can be facilitated by HR analytics are HRP, performance management, training and development, compensation and benefits programme design and management, talent acquisition, retention, development, mobility of human resources and so on.

A typical job description of HR managers today even includes performing employee-related research, developing models to serve business needs and also focusing on human resources. The collection and analysis of HR data for actionable analysis and reporting to managers and executives are more important part of HR managers' functions in today's organizations. Hence, the responsibilities of HR managers today are building HR intelligence function, talent acquisition, development and retention with talent analytics, evaluating the effectiveness of any form of HR investment, functioning as HR business partners, providing coaching and training support and so on. However, such responsibilities may change from time to time, depending on the nature of the organization and the scope of HR activities.

2.20 CONDUCTING HR ANALYTICS

The HR analytics cycle involves seven steps (refer to Figure 2.1) to help organizations develop HR research and analytics capabilities (Falletta, 2014). More importantly, it serves as a proactive and systematic process for establishing a comprehensive portfolio of strategic HR analytics practices, projects and priorities that enable HR strategy, evidence-based decision-making and the execution of the overall business strategy.

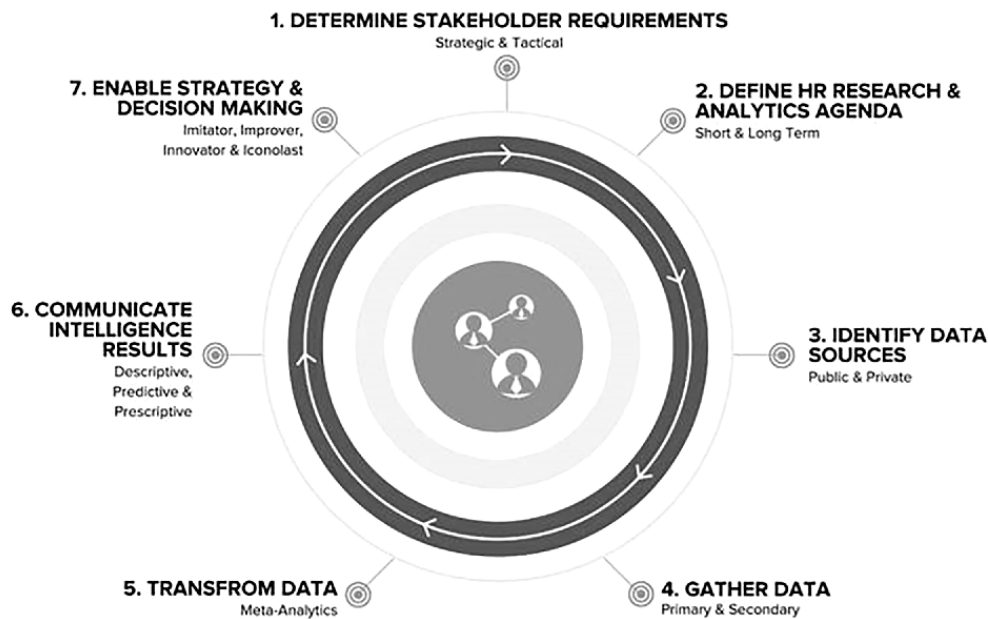


Fig. 2.3: Seven Steps in the HR Analytics Cycle

Step 1: Determine Stakeholder Requirements

Determining stakeholder requirements is vital to the overall success of any HR research and analytics initiative. It is much more than meeting with a few influential or vocal stakeholders each year to formulate the annual HR research and analytics agenda. It is about establishing and cultivating a partnership and becoming a legitimate player by adding value to the business. With respect to the HR analytics cycle, an ongoing and proactive partnership with key stakeholders is an essential role to ensure trust and upfront legitimacy to obtain an accurate picture of the most pressing organizational problems.

Who are the stakeholders? A stakeholder, in the broadest sense, is anyone who is directly or indirectly affected by HR analytics work. Stakeholders in the context of HR analytics include executives, line managers, senior HR leaders, HR business partners, employees and, in some cases, human capital technology vendors. Each stakeholder has a different perspective and set of concerns regarding HR analytics practices and activities. Whereas line managers are usually most interested in the data visualization and reporting of key metrics and analytical insights, executives and senior HR leaders are generally more interested in how HR analytics enables HR strategy and execution, critical workforce decisions and other important business outcomes.

During this step, many interesting questions are likely to arise. Carl Sagan, in his book, *The Demon-Haunted World: Science as a Candle in the Dark*, explains “There are naive questions, tedious questions, ill-phrased questions, questions put after inadequate self-criticism. But every question is a cry to understand the world. There is no such thing as a dumb question” (Sagan, 1996, p. 323). While this is a very popular axiom and largely true, we’ve encountered a fair share of unusual and potentially unethical questions by some well-intended stakeholders in our work.

To be fair, well-intended stakeholders are also bombarded with the latest management fads and trends (Barends and Rousseau, 2018; Rice and Boudreau, 2015). Hence, those leading HR analytics in organizations have an important role and responsibility in assisting stakeholders with framing HR research and analytics questions that are realistic, measurable, valued-added and more importantly — evidence-based and ethical. Determining stakeholders' requirements is important to formulate and frame HR research and analytics questions, identify strategic and tactical HR research and analytics priorities, secure stakeholder commitment and support, and provide communication on the ongoing progress of key HR research and analytics initiatives.

Step 2: Define HR Research and Analytics Agenda

Once the stakeholder needs and expectations are identified, it is time to define the HR research and analytics agenda. An HR research and analytics agenda may be long-term or short-term. The constantly evolving nature of business and the “future of work” are redefining what is considered long-term versus short-term. In our age of on demand data aggregation and visualization, algorithms, artificial intelligence and automation – long-term is no longer 3-5 years out. One year is considered the long-term norm currently in virtually all industries. Conversely, short-term requirements tend to coincide with the organization's quarterly results, sometimes monthly. It important to note that short term does not necessarily mean tactical or reactive, nor is long term equated with strategic. Short-term and long-term research requirements can be both strategic and tactical in nature.

Tips for Establishing the HR Analytics and Research Agenda:

- Organize the general stakeholder requirements by theme or major topic;
- Pose broad research questions for each theme or major topic and use stakeholders' language and terminology to the extent possible;
- Under each of the broad research questions, begin generating more targeted questions and hypotheses that lend themselves to measurement and consider the ethicality of the questions and hypotheses;
- Identify both the long-term and short-term requirements of the overall HR research and analytics agenda;
- Share the HR research and analytics agenda with key stakeholders and go through an iterative process of refinement; and
- Lastly, strive for a balanced HR research and analytics agenda in terms of reactive and proactive work. A relatively recent HR analytics study conducted across Fortune 1,000 firms revealed that on average nearly 40% of HR analytics priorities were determined by the HR research and analytics team while approximately 60% were driven by stakeholders (Falletta, 2014). Hence, the HR analytics team should devote some resources to research projects that are strategic in nature, aside and apart from what was generated during the stakeholder requirements step.

Step 3: Identify Data Sources

Once the HR research and analytics agenda is established, the next step is to identify the sources of data that will help to answer the HR research and analytics questions and hypotheses. Data sources may be either public or private. Public data reside in university libraries and governmental databases (*e.g.*, UK's Office of National Statistics, US Department of Labour Statistics, etc.). Private data includes an organization's internal employee data housed in its HRIS as well as external benchmarking data from "best-in-class" organizations. Research reports and results gathered by membership-based consortia (*e.g.*, Global Centre for Work-Applied Learning, Centre for Evidence-based Management, Gartner CEB, The Conference Board and the i4CP) and academic think tanks (*e.g.*, Institute for Employment Studies in the UK, Cornell's Center for Advanced Human Resource Studies and University of Southern California's Center for Effective Organizations) are credible sources of private data and information. Sources of data may or may not exist depending on your organization's current HR research and analytics practices.

While reviewing data sources, questions may arise as to whether the organization's HR research and analytics practices are still useful to the business (*e.g.*, some practices may have become institutionalized over the years). Therefore, some tough decisions may need to be made with respect to modifying existing HR research and analytics practices, discontinuing outmoded or symbolic practices and adopting new ones.

In terms of adopting new HR analytics practices and solutions, we caution organizations to do their homework before plowing into uncharted territory or chasing the next human capital management (HCM) technology or artificial intelligence (AI) platform. Too often, HR practitioners are too enamored with HCM technologies designed to select, manage, measure, track, monitor, quantify, nudge, engage and coach people in the workplace. These novel technologies, while promising, typically rely on a proprietary algorithm and unexplainable "black box" without fully understanding the underlying mechanisms at work (*i.e.*, showing which factors or reasons led to human capital prediction or decision and how). According to Angrave *et al.* (2016), analytics must be rooted in an understanding of the data to be used and the context under which data were collected if any meaningful insight is to be gained.

In our view, if an HCM technology vendor is unwilling to crack open the proverbial "black box" – don't work with them. Although HR analytics does involve the use of technology to collect, manipulate and report data (Marler and Boudreau, 2017), it is critically important to weigh the potential rewards and risks including the legal and ethical implications associated with various HCM technologies.

Step 4: Gather Data

This step of the HR analytics cycle involves the actual collection of data through primary research, secondary research or mining the organization's internal data in the HRIS. Primary research is new or original research that addresses a specific research question or set of questions (*e.g.*, a research project or experiment to identify which factors enable or inhibit employee engagement and performance, selection research using validated personality instruments for

leadership succession, employee and organizational surveys for action planning and change, organizational network analysis to determine the level of collaboration by specific jobs and roles). Primary research can be done in-house if an organization has the HR research and analytics capabilities or in partnership with credible think tanks or universities (Simón and Ferreiro, 2018). Secondary research is data and information available through existing sources (*e.g.*, literature review of journal articles and reports from credible institutes, HR benchmarking, labour market databases, etc.). The most reliable and trustworthy secondary sources of information for evidence-based decisions and practice come from scientific research studies (*e.g.*, systematic reviews, meta-analyses, literature reviews, etc.) (Briner and Barends, 2016). Theory-driven (deductive) and data-driven (inductive) approaches to mining and modeling data from the organization's HRIS, SaaS-based platforms (*e.g.*, data aggregation and visualization products, employee engagement tools, and social or organizational network technologies) and other external data sources are another way to gather, query and analyze data about the workforce, provided it is done ethically and responsibly.

Step 5: Transform Data

Transforming data into useful and meaningful insights is arguably the most important, yet most challenging, step. Many advancements and innovations have been made by leading-edge software firms (*e.g.*, Oracle, Salesforce, SAP, SAS, and Workday) that incorporated HR analytical capabilities within their suite of products such as predictive analytics, process analytics, text and sentiment analytics and real-time analytics, to name a few (Strohmeier *et al.*, 2015). Similarly, several promising data aggregation and visualization platforms have entered the market. However, none of these enterprise platforms and SaaS-based tools can magically codify, analyze, visualize and interpret all the disparate "Big Data" at our disposal (Angrave *et al.*, 2016). This is especially true for evidence-based findings and insights derived through scholarly research as well as applied HR research and experimentation projects performed in organizations.

In the context of HR strategy, much of this work is still done manually by qualified HR researchers, analysts and data scientists (Falletta, 2014). It is suggested that organizations start small and build their HR analytical capabilities over time. Organizations should perform a meta-analysis (*i.e.*, an analysis of analysis) across a handful of targeted data sources. We refer to the term meta-analysis loosely and not in the traditional sense in which scholars use advanced statistical procedures to combine the results and examine the effect sizes of multiple scientific studies (Barends and Rousseau, 2018). Therefore, meta-analysis in the context of HR analytics is a practical approach to explore and understand multiple data sources in relation to each other. For example, to what extent are the results from individual 360-degree assessments consistent with your employee survey data, exit survey data, or actual turnover? Are high-potential, emerging leaders leaving the organization for the same reasons year over year (*e.g.*, little to no advancement and promotion opportunities, few organizational leadership opportunities, lack of decision rights, and low base pay relative to the market)? Performing a meta-analysis enables to answer these questions, and more importantly, codify and make sense of disparate data sources to glean critical workforce insights. Performing the meta-analysis can be simple or complex. This largely depends

on the nature of the data gathered, sophistication and competency of the HR researcher, analyst or data scientist and the amount of resources and time to conduct the analysis.

Step 6: Communicate Intelligence Results

The sixth step of the HR analytics cycle involves communicating intelligence results. Real HR analytics capabilities place more effort and emphasis on telling a story about the data and visualizing the data in the context of the organization's most pressing problems and successes (Minbaeva, 2018). Storytelling can be a powerful approach in communicating data-driven insights, both in words and visually, because it stimulates emotions (*i.e.*, the brain processes emotions differently than facts and data) (Waters *et al.*, 2018; Welbourne, 2015). Storytelling, however, should not be a guise or pretext for telling executives what they want to hear or "cherry-picking" the data and insights. For example, Rasmussen and Ulrich state that "HR analytics can be misused to maintain the status quo and drive a certain agenda (*i.e.*, when to know what story to want to tell, and to then go look for data to support same" (2015, p. 237). Moreover, we need to consider the "veracity of the story" and the ethicality of the way in which data-driven insights are derived, communicated and used (Rotolo and Church, 2015). The dissemination of inaccurate or misleading insights will invariably lead to bad workforce decisions and big organizational consequences (Church and Dutta, 2013). Hence, communicating and reporting HR analytical insights involves not only some ethical interpretation on the part of HR analytics team, but also speaking truth to power.

Considerable advances have been made in linking HR to the business strategy and organizational performance (*i.e.*, a value creation framework) through the use of the balanced scorecard (Kaplan and Norton, 1992), HR scorecard (Becker *et al.*, 2001) and the workforce scorecard (Huselid *et al.*, 2005). However, as alluded to earlier, not all forms of internal and external data and information can be expressed in terms of a simple metric or indicator. For example, some HR research and analytics activities (*e.g.*, advanced research in the context of social, behavioural and organizational sciences) require advanced statistical analysis, meta-analytic, multivariant or causal modeling procedures as well as expert interpretation and insight. Nonetheless, scorecards and dashboards are effective means to communicate strategic metrics and are best used as communication and reporting tools for strategy execution once an agreed-upon strategy is in place.

Step 7: Enable Strategy and Decision-making

The final step of the HR analytics cycle is to enable HR strategy creation and evidence-based decision-making. We have heard the proverbial mantra that behind every successful organization is a strategy that works. But what exactly is strategy? Strategy is a multidimensional concept that can be defined in many ways. Mintzberg describes strategy in terms of a plan, ploy, pattern, position and perspective. As a plan, strategy relates to leaders establishing the overall direction for the organization. As a ploy, strategy is all about maneuvering and outwitting competitors. As a pattern, strategy involves engaging in specific behaviours and consistent action to effectively implement the strategy. Strategy is also a position in terms of how an organization

differentiates itself in the competitive marketplace. Lastly, strategy is a perspective that reflects an organization's culture and character (*i.e.*, does the organization see itself as an imitator, improver or innovator when it comes to its people, products and services) (Mintzberg *et al.*, 2005). In summary, strategy involves asking intelligent questions, identifying strengths, weaknesses, opportunities and threats (SWOT), knowing the right things at the right time, scenario planning, evidence-based decision-making, establishing priorities and goals and effectively managing execution. Despite these various methods and processes, strategy in the context of HR "refers to the processes, decisions, and choices organizations make regarding how they manage their people" (Cascio and Boudreau, 2014, p. 79). A human resource strategy tends to focus on aligning people policies, practices and processes with the overall business strategy in order to achieve the organization's goals and objectives. In a perfect world, HR strategy creation should be done in concert with the overall business strategy, although this rarely occurs in practice.

A human resource strategy also involves making smarter HR decisions, and there are no shortage of models, frameworks and guidelines on the topic of HR decision-making. For example, Boudreau and Ramstad (2007) introduced their HR decision science approach to improve human capital decisions in organizations. Similarly, evidence-based HR involves making HR decisions through the conscientious, explicit and judicious use of the best available evidence from multiple sources of information (Briner and Barends, 2016). Another intriguing approach is Dulebohn and Johnson's (2013) classification framework for HR decision-making in the context of HR information systems and analytics that is based on the seminal work of Gorry and Scott Morton (1971). The framework is organized by two dimensions: (i) management decision-making level (*i.e.*, strategic planning, management control and operational control) and (ii) decision-making structure (*i.e.*, structured, semi-structured and unstructured). Their framework describes the data needs, decision characteristics and concomitant metrics across the operational, managerial and strategic levels. Specifically, this framework is a useful decision support tool for categorizing a wide range of HR practices, activities and choices in a meaningful way (Dulebohn and Johnson, 2013).

In short, these models and frameworks represent a complementary approach for enabling HR strategy and decision-making in the context of HR analytics. In our view, the primary purpose of HR analytics is to enable HR strategy and decision-making. The resultant data and insights derived from HR analytics play a central role in influencing HR strategy, decision-making and the strategic choices organizational leaders make when it comes to adopting evidence-based HR practices (Falletta, 2014). Further, we introduce the HR strategy axis as a useful framework to guide strategic HR choices with respect to human capital decision-making and investments in organizations.

2.21 THE BASICS OF TEXT ANALYTICS FOR HR ANALYTICS

Organizations with a strong *data culture* regularly use data to make better and more efficient decisions. Perhaps, you've already decided to try and build a stronger data culture at your company. But are you using data in all the ways you could be?

When you think of organizational HR data, you may think of the quantitative data companies routinely gather via employee engagement surveys or performance reviews. You may also think of other types of HR data, such as employee attrition rates, attendance statistics, or legally-mandated *DEI statistics*. But there is one form of organizational data that is frequently overlooked, although most companies have access to it. Your business may be collecting large amounts of this data without even realizing it.

This “hidden data source” is text-based data or written information. Think of all the textual information most HR departments use each day — performance reviews, interview notes, onboarding documents, survey responses, and more. All these text provides a wealth of data for potential analysis.

Below, we’ll give an overview of what text analysis is, and how it can be used to generate valuable insights for your business.

2.21.1 What is Text Analysis?

Text analysis is also referred to as “text mining”, “text data mining”, or “text analytics”. It is the process of extracting information from text using data science techniques, tools, and analytics methods, such as natural language processing and artificial intelligence.

To get started with text analysis, you will first need to gather a set of written resources. These written resources are known as the corpus. Once you have the corpus put together, you can answer a number of questions using analytics techniques. Here are some key factors you can address:

1. Word Frequency

There may be situations where you need to analyze the frequency of particular words in your corpus. For example, if your organization is large, you may have a significant amount of written survey responses that need to be analyzed. Perhaps, you’d like to know how employees feel about a specific subject — such as a new team-building programme. In that case, understanding how frequently the word “team building” and related words show up in the survey responses would be a useful insight.

You can easily visualize word frequency by using a word cloud. Word clouds show you what the key words in your corpus are, and how often they are being used. Large words in the word cloud indicate frequent use, while smaller words indicate less frequent use. Here’s what a word cloud looks like:

3. Word Association

Understanding the association between different words will add an additional level of depth to your analysis. Returning to the team-building example, perhaps the word “team-building”, is often found alongside words like “instructor”, “program”, “lunch hour”, “fun”, or “frustrating” in your company surveys. By seeing which words are associated with one another, you can better understand the tone and content of the text. Visualizations of word associations can provide

valuable insights. Here's an example from the University of Cincinnati's Center for Business Analytics, using text from the Harry Potter book series:

3. Employee Sentiment Analysis

Employee sentiment analysis is about evaluating the tone of written texts to understand the way employees perceive your organization. Using AI and natural language processing, analytics professionals can figure out if the overall theme of a piece of text is positive, negative, or neutral. Sentiment analysis can also help to drill down to more specific emotions, such as "confusion" or "concern". This is useful, because it gives you a more objective way of evaluating employee data (such as survey results) without needing to rely on the reader's interpretations. In addition, sentiment analysis can help you quickly get a general idea of overall employee sentiment when you have gathered a large amount of data, without having to sift through all of it manually.

2.21.2 How Does Text Analysis Help with HR Analytics?

There are many different applications of text analysis in *HR analytics*. In the "Word Frequency" section above, we gave one important example: text analysis can help large organizations review their written employee engagement survey results more efficiently. Here are a few other possible uses for it:

1. Reviewing Internal Materials

Text analysis can help you understand how your internal documents might be perceived by employees and prospective employees.

For example, the onboarding materials you present to new employees send a certain message about the company and the type of environment the person is joining. It is important to make sure these documents are striking the right tone for a successful entry into the organization. Reviewing the text using sentiment analysis can help you reach this goal.

2. Applying Text Analysis for DE&I

Another interesting application for text analysis is in DEI (Diversity, Equity, and Inclusion). When aiming to meet their DEI goals, companies often find themselves combating subtle forms of discrimination, such as unconscious bias. These forms of bias can be difficult to identify and deal with effectively.

To understand if individuals from different backgrounds are having a consistent experience at your workplace, you could perform a text analysis of narrative performance assessments. If managers are assessing employees formally or informally through text, this would provide a resource for to draw from. A text analysis will help you identify if the frequency of certain words, types of words, or tone used in performance assessments is consistently the same across all individuals in your company.

What Can You Evaluate With Text Analysis?

There are two broad types of text analysis: quantitative and qualitative analysis. Each can help you understand and evaluate your textual corpus.

1. Quantitative Text Analysis

Quantitative text analysis is the process of analyzing text using statistics and data science methods. Usually, quantitative analysis is performed using a software environment such as R, or a programming language like Python.

With quantitative analysis, you will be using some of the techniques previously mentioned, such as word clouds and employee sentiment analysis. You may also be looking at quantitative variables like word count. Perhaps, some of your written sources are longer or shorter than others.

2. Qualitative Text Analysis

When performing a qualitative analysis, analytics professionals take a more general thematic approach. This type of analysis may involve human coders as well as algorithms.

The goal of qualitative/thematic analysis is to understand which themes are mentioned in a body of text. For example, imagine a restaurant has these three reviews online:

- “The food was excellent, but the room was a little cold.”
- “A great venue for fine dining. The only downside is the poor indoor climate control.”
- “The ambiance and heating in the main hall were just perfect. I also really enjoyed my dish.”

In this case, the themes all three reviews have in common are the **temperature of the restaurant, and the quality of the food.**

Qualitative text analysis uses similar techniques as employee sentiment analysis. In the above example, a sentiment analysis would show that the tone of the first two reviews is partially negative, while the third review is entirely positive.

2.22 FUTURE OF HR ANALYTICS

Introduction

Future HR managers will be equally thorough with core business operations, will make use of data and analytics for business-aligned HR decisions and will manage human resources as a more specialized function with analytics. Future HR functions will require in-depth knowledge in data analysis. Technology will be an important enabler for future HR functions. Technology-enabled HR decision-making would be evident in all spheres of HR functions. With technology support, human resources will be more matured to drive business outcomes of organizations. Social media also will influence the way human resources attracts, hires, engages and retains talent in organizations. This would require human resources to focus on employer branding, collaborating with the marketing function, so that in social media, organizations become compelling news with all positive manifestation.

Future HR analytics will not limit its utilization within HR department; rather, it would be integrated with strategic-level business analytics. Obviously, in future, human resources will also become a business function. This could make HR analytics more technology enabled in future

and would require specialized skill sets. It is apprehended unless future generations of HR managers are more equipped with the power of data analysis, spread of data being expanding day by day, encompassing social media and various other external data sources; apart from organization-wide datasets, companies may require to outsource the HR analytics function (Inostix, 2015). Obviously, this will be a great future challenge for human resources. Challenge will be not only to identify proper vendor but also to ensure data security to the organization.

Globally, we have examples of many organizations, such as Google, Wells Fargo, Xerox, 3M, Ericsson and so on, that could benefit from HR analytics. Some of these known benefits are Xerox's lowering of attrition for effective hiring through HR analytics, 3M's productivity improvement through scientific succession management, Wells Fargo's achievement of high employee engagement and enhanced performance and so on.

Boudreau (2015) could identify five forces which may change the future of HR functions. These are exponential technology change, social and organizational reconfiguration, truly connected world, all-inclusive global talent market, and human-machine collaboration. All these forces will bring changes in HR functions, requiring more evidence-based HR decision-making using HR analytics and predictive decision models. For example, exponential technology change would require successful engagement of the workers with automation. Social and organizational reconfiguration will influence talent sourcing and employee engagement processes. So also, a truly connected world and all-inclusive global talent market would have effect on the talent management function. Finally, human and machine collaboration will enhance the use of analytics, algorithms, big data for decision-making and so on.

2.23 UNDERSTANDING FUTURE HUMAN RESOURCES

In future, HR analytics can substantially do many HR jobs, thus relieving HR managers from drudgeries of lengthy work processes. For example, in future, HR analytics can facilitate employees' performance evaluations more holistically, encapsulating apparently extraneous and exogenous data. Another example is the power of HR analytics to match manpower scheduling based on demand, so as to ensure the availability of right manpower at times of need. This is particularly important in retail and hospitality industries where revenue models encapsulate optimum deployment of manpower. All these would be substantial time-saver, thus leaving ample time for HR managers to focus on employee development and many other strategic and business functions. With the increased spread of digital work environment, HR managers in future are likely to face the challenge of acute talent attrition, and in the process, talent attraction, development and retention would become more challenging jobs, which itself may substantially pre-occupy them. They need to address talent attrition issues engaging employees, and also with early signals, they plan for suitable intervention.

Technology-enabled human resources will be able to answer our age-old questions such as why top talent will work for the company, what would be the future experience of customers with our employees, what can contribute to employees' motivation, how can an organization better relate to employees' previous experience with their job success and so on. The social behaviour of

employees will be the determinant for their workplace behaviour and can exert influence on their ethical values, confidentiality, information sharing and so on. All these can help in the better understanding of organizational dynamics and can also enable human resources to make the work more satisfying to employees. Automation of HR jobs for advanced analytics, on the one hand, will make future jobs more talent-driven, and on the other hand, it will make HR tasks more personal, facilitating collaboration, showing concern for people, championing for changes and so on.

HR analytics will gradually penetrate the entire organizational structure and will provide new insights into competitive advantage making use of data. With data-driven informed decision-making in all critical areas of human resources such as recruitment and selection, performance management and so on, future human resources will have direct impact on business growth and the bottom line. With built-in cognitive capabilities, HR analytics solutions increasingly become more user-friendly and will not require specialized knowledge, training or expertise to make the HR decision-making process data-driven. A data-driven approach to HRM will gradually expand to entire HR functions, and it would become a language for human resources and the other top managerial functionaries of the organization.

HR functions today are no longer engaged in doing traditional soft HR functions but in helping organizations to gain sustainable competitive advantages. Obviously, this requires a thorough understanding of the impact of everything they do in the business and strategy context, and legitimize their role as a strategic business partner.

With HR analytics, future HR functions will witness growth opportunities. Some of the future changes predicted by experts indicate that routine HR jobs will decrease and companies will go for more outsourcing of HR functions. With new technology, HR functions will be more inclusive and participative, and will become a business process. With self-service computer systems, employees across all the functions will do their part of data entry, thus rendering a large chunk of HR jobs redundant. Such redundancy would be more evident at entry-level HR jobs, and also in routine or transactional HR tasks, as this will be outsourced. Contrarily, some HR tasks like compensation and benefits will demand more expertise from future HR professionals, as this function will become more complex with new employment-related regulations.

We can understand such future trend analyzing people operations team of Google, the company known for its extensive use of people analytics in all HR decisions. People operations team of Google require the following types of employees:

- Staffing (with focus on talent attraction and recruitment)
- Analyst (to support growth and management strategies)
- Programme manager
- HR business partner
- Strategy and operations
- Learning and development
- Administrative support

Similarly, future HR tasks would demand more strategic thinking, as human resources will increasingly become a strategic business partner. Here also, HR analytics and technology-enabled HR process could make this possible. Human resources can enhance strategic business value to an organization only when HR managers build their future capabilities to project the future trend based on the understanding of the current business situation. Obviously, this requires strategic planning, which can only rest on in-house HR capabilities (as this cannot be outsourced). Such strategic focus may even create new job opportunities in future for human resources, like HR business partners. Future HR managers' expertise would require not only holistic understanding of human resources but also a thorough understanding of business, operations and strategy, more focus on HR specialist role. Quantum jump in using Big Data and Analytics in Human Resources will also exert influence on future HR tasks, such as management of remote workforces, talent attraction with organizational brand building, social networking and so on.

Generic Future HR Skill Sets and Knowledge

Although it is difficult to list generic skills of HR managers in future, especially in the context of increasing the use of HR analytics and predictive decision-making models and algorithms, we can list some common skills that are essentially required for future HR managers.

Right in the beginning, HR managers of future need to possess thorough knowledge of human resources and business. HR knowledge, namely reasons for attrition and talent management, need to be analyzed in the context of business capturing data on sales, customers' satisfaction, sales margins, key sales objectives (KSOs) and so on. Such knowledge, on the one hand, can help in identifying a cluster of employees who are likely to leave the company and important reasons for their leaving; on the other hand, with business knowledge, HR managers can also predict KSOs at risk and so also can predict the RoI from business.

Statistical knowledge is important for data interpretation. Statistical knowledge today encompasses various technology-driven techniques like machine learning. Core knowledge, however, centres on the mechanism of data collection, survey design and designing of experiments (particularly for finding cause-and-effect relationships). Statistical and mathematical knowledge can help in developing decision algorithms based on which behavioural modeling for future decision-making can be done.

With programming skills, HR managers can analyze the data using various available HR analytics solutions. Most of the HR analytics solutions are user-friendly; hence, today these do not require complex programming skills. Like programming skills, data visualization skills are also important. With these skill sets, HR managers can help in understanding the right story from the data and accordingly can persuade people to change their behaviour.

With all these skills sets and knowledge, future HR managers would be able to manage future workplaces which would be more analytics driven.

Ethical Issues in HR Analytics

HR functions by nature require a lot of confidential information, which employees provide to the organization. These are personal to the employees, even though organizations are

custodians. In terms of employment contract with the employees, organizations make use of employees' information for managing the business. For HR analytics, employees' data are used for modeling the HR decision-making process extensively and even at times for violating the ethical norms. While the best practice is to take employees' consent for making use of their personal data and information for decision-making, operationally, it may not be always feasible. Technology-enabled HR processes require storing of employees' data about which employees may not be aware. Such data even at times go beyond employees' work spheres, like data from employees' social media interaction. Even data from employees' work spheres also may include some typical behavioural inputs, such as their food choice in staff canteen, their preference for beverages and so on. These require organizations to draw a line between ethical and unethical use of employees' data that are privy to them.

Often analyzing employees' data, which may be privy to them, help organizations to get deep insights into future HR decision-making. For example, relating employees' food choice in the canteen to their performance, organizations can come out with ideal meal plan. But again, this could be a critical issue of ethics. Unlike survey, where employees' reserve the right to respond to their own fashion and rights, direct and indirect data of employees are collected without their knowledge. This is often construed as managerial prerogative. Good practice, however, calls for communicating the employees' organizational intention to make use of their information for modeling future HR decisions, so as to ensure both people and organizations deliver their best in achieving business goals.

Organizations at times also face security challenges when HR analytics support is outsourced. Here, it is very important to ensure data security with clear ethical guidelines. Issues such as stereotyping, prejudice and discrimination are a critical ethical concern for employees' data. Although organization-wise ethical standards in using employees' data or HR analytics widely varies, CIPD, UK (2012) could come out with some specific guidelines in this respect. These ethical guidelines have been given in general for HR professionals, rather than specifically for HR analysts, but adhering to these guidelines can ensure compliance with the ethical issues pertaining to the use of employees' data also. CIPD emphasizes the need to:

- Establish, maintain and develop business relationships based on confidence, trust and respect;
- Exhibit and defend professional and personal integrity and honesty at all times;
- Demonstrate sensitivity for the customs, practices, culture and personal beliefs of others;
- Advance employment and business practices that promote equality of opportunity, diversity and inclusion and support human rights and dignity;
- Safeguard all confidential, commercially sensitive and personal data acquired as a result of business relationships and not use them for personal advantage or the benefit or detriment of third parties.

A more specific form of human resources ‘analyst ethical standards’ based on Schwartz (2011) can be listed as follows:

1. Compliance with the legal, cultural and social norms.
2. Safeguard against the interest of all stakeholders.
3. Accountability for the HR analytics process.
4. Safeguard against the security of information.
5. Concern for information sensitivity.

HR managers while making use of HR analytics need to ensure that they are not ignoring the employees’ privacy without their consent. However, we have conflicting claims on ethical issues in HR analytics, more specifically in predictive HR analytics. This was evident in the debate on SHRM (2015). For example, it is claimed that the real beneficiaries of predictive HR analytics are employees, as through this, an organization can develop its structures, develop its performance management systems and rewards systems, can retain its talent and work on succession planning more effectively. All these can substantially contribute in creating an enabling environment for employees to give their best. Another claim is that predictive HR analytics while developing decision algorithms and models based on employees’ data often violate ethical norms. Many companies often collect data on employees’ attitudes, preferences and values, which are in general privacy to the employees. This requires certain code of ethics to ensure predictive HR analytics are not prejudicing the privacy of the employees and that they are adequately informed about the possible use of their data.

HR Feel More Empowered with HR Analytics

Empowerment of HR function can be possible once HR managers can manage HR activities with HR analytics. HR analytics can help in performing many transactional HR jobs at a much faster pace, thus leaving HR managers to focus more on strategic HR activities to justifiably align human resources with the business. True empowerment of human resources is possible when we are able to bring HR functions closer to business. This can also make human resources responsible for the bottom line and the ROI for the organization.

With HR analytics, HR managers can significantly improve the performance of a company, can build engaged human resources, increase their productivity, and in the process, can achieve the goals of the company. Many critical HR decisions when taken based on HR analytics, human resources feel more empowered, as HR managers can get insights into the issues and take decisions more accurately. With predictive analytics, HR managers can even go beyond in managing HR functions with a forward-looking approach. Also, predictive analytics can help HR managers in driving HR strategies with future perspectives.

What Does the Future Look Like?

In order to look toward the future, it is valuable to look toward four other fields of study for insights that will enhance human capital data. They are:

1. **Finance:** Business standards and valuation of organizations
2. **Mathematics:** Chaos theory
3. **Information Technology:** Big Data
4. **Automated Processes:** Decision support

Finance: Business Standards and Valuation of Organizations

As discussed in previous chapters, organizations can use the Talent Development Reporting Principles (TDRP) framework to gather and report standard measures about HR that business leaders can easily understand and use for decision-making. In the near future, more organizations will begin to adopt this framework to help make sense of HR in measurable ways. The framework is simple in concept and broad enough to apply across industries. Moreover, it is a standardized approach that allows for benchmarking and comparison.

For more than a decade, finance and accounting specialists on Wall Street have been pushing for a way to quantify the value of intangibles. Traditional valuation focuses on the brick and mortar of a company to determine its value. The tangible aspects of a business — buildings, equipment, products, inventory, and pipeline — were given the most weight when determining the value of a company. As the first world has shifted toward a knowledge-based economy, the intangible aspects of a business now represent from 60% to 80% of the value of some enterprises. Patents, research and development, brand, and employees are the intangibles that are propelling progress and creating competitive advantage. For accountants, inventory is easy to count and value. It is not so easy to assess the value of intellectual property, especially when it is trapped within the brain trust of the workforce. Organizations like the Human Capital Management Institute are striving to create taxonomy of human capital measures that aid accountants in determining the financial value of intangibles. In 1992, Gary Becker won the Nobel Prize for Economics for applying macroeconomics to human behaviour like discrimination, educational pursuits and family interactions. At the time, the stock markets valued tangible assets. The company valuation is now more heavily influenced by human capital. The TDRP framework is one possible method to assess the intangible value of human capital. For this reason, as Wall Street and other financial centers assess the market value of companies, TDRP may be used pervasively in the future.

Mathematics: Chaos Theory

In 1987, James Gleick wrote a watershed book on mathematical theory called ‘Chaos: Making a New Science’. In one volume, he integrated seemingly disparate mathematical ideas of fractal geometry, strange attractors, and the butterfly effect. Michael Crichton, the author of the novel Jurassic Park, made these complex mathematical ideas accessible to the general public. Gleick drove home the thesis that there are measurable patterns hidden within seemingly chaotic datasets.

Patterns in chaos. That’s almost as ironic as the notion that change is constant. If you saw the 2001 movie, *A Beautiful Mind*, you saw Russell Crowe portraying John Forbes Nash, a

brilliant mathematician, and you may recall the scene where he is standing in front of a blackboard covered with data. In time, he sees a pattern in the data that no one else can see. Unless we can channel John Forbes Nash, we must rely on analytics to find the data patterns.

An example that mixes HR and waterwheels might help. HR executives face daunting challenges when managing the human capital of an organization. While trying to maintain a productive workforce and addressing a myriad of socio-demographic challenges, leaders have to:

- Capture and share knowledge from an aging or retiring workforce;
- Manage the needs of a multigenerational workforce;
- Communicate with a multilingual workforce;
- Optimize productivity among a dispersed, international, and virtual workforce;
- Coordinate operations with offshore/outsourcing partner.

Add to these challenges the numerous external factors that influence change — such as the fluctuating economy, the war for talent, and ever-evolving attitudes among Gen X, Gen Y, and Millennials — and plotting the proper course forward becomes even more difficult. It is no wonder the ebb and flow of employees often seems chaotic and uncontrollable. This is where Chaos Theory becomes practical.

Take a moment to search the Internet for water wheels. We suggest using the terms “chaotic Lorenz water wheel.” You will likely find a link to a YouTube video of a simple water wheel. This is not the standard water wheel used at a gristmill or a sawmill hundreds of years ago. This mechanism uses buckets that pass below a water source, gaining weight and speed as they fill up. However, the buckets also have holes in their bottoms. So, the inflowing water is also released slowly. This combination of inflow and outflow creates a chaotic rhythm on the water wheel that seems to have no cycle. Watch the wheel for some time. (It is hard not to get hypnotised.) It rolls forward with great speed, then slows and even reverses direction.

What does this have to do with HR? Everything. Consider the water wheel as a metaphor for the revolving door that brings the workforce into a company and escorts them out. There is a stream of incoming new hires and an outflow of retirees, transfers (across business units), and terminations (voluntary and involuntary). The stream is often a deluge of new hires. It can also be a mass involuntary exodus due to downsizing. The important distinction is that the revolving door does not spin at a constant rate. In fact, the variation is seemingly inexplicable.

Modeling is the Only Way to Define the Situation Accurately

Gleick states that the goal of mathematics is to reveal “the fine structure hidden within a disorderly stream of data.” While the term “chaos” implies that there is infinite complexity that cannot be modelled, the scientific study of seemingly chaotic processes has the end goal of finding and defining the hidden patterns. Once the pattern or structure is defined, it can be explained. This is true for any system including the recruiting cycle within an organization. Then, when the contributing factors are explained, they can be controlled. By controlling the inputs, the outputs can be influenced. The wild and frenzied hiring and firing seasons that lead to

overstaffing and understaffing can be attenuated. By controlling the extreme swings in the workforce, organizations control costs and increase efficiency, effectiveness, and profitability.

A case in point is the 1990s trial of a major electronics company that got its staffing and sales curves out of phase: Hiring while sales were falling and *vice versa*. Over a period of four years, it spent about \$1 billion over hiring and laying off. Eventually, it fell behind the curve for its primary product line (where it was No. 2 in the world) and had to sell it off at fire-sale prices. The company subsequently fell behind the innovation race in a second key product and never recovered.

How do we go about modeling a water wheel and, in turn, recruiting? Thankfully, mathematicians have been modeling this system and many others for decades. Edward Lorenz, who is famous for studying weather patterns and proffering the “butterfly effect” — the notion that dynamic systems like weather can change dramatically even when the inputs to the system are very small — is also famous for creating a mathematical model for convection patterns of fluid being heated in a container. The initial patterns when graphed in two-dimensional space showed no pattern at all. However, when Lorenz graphed the equations in three-dimensional space (X, Y, and Z axes), a pattern emerged.

The pattern became a model known as a Lorenz Attractor in his honour. Coincidentally, the man who gave us the butterfly effect also gave us a three-dimensional model that looks like a set of butterfly wings. The model clearly shows a pattern. Another search of the Internet using the terms “Lorenz attractor in motion” will show how the pattern evolves over time. The graphic in motion represents the movement of the water wheel. It spins slowly, then faster in a small tight loop, and then when it reverses, it crosses over to begin a small tight loop on a different dimension. Until the multiple dimensions are shown together in this video, it is hard to fully understand the complexity of the underlying pattern. Consider how this pattern might affect your attempts to hire, develop, and retain talent.

We hope you have begun to wonder what the hiring pattern looks like within your organization. Is it a butterfly-like Lorenz attractor? It may be. Or it may not be. Regardless, it is worthwhile investigating. Gather the data and create a mathematical model. The underlying pattern might be simple and linear. Or it might have a logarithmic or a dynamic chaotic structure. Regardless of the actual pattern, it certainly wouldn’t be boring to investigate, model, and define the inflow and outflow pattern of employees. To say you don’t have time to pay attention to these patterns leaves your enterprise behind the market.

Mathematical models help us define and understand our world. Sometimes, advanced math, like chaos theory, is required to uncover the underlying structure of our data. Other times, the pattern can be uncovered easily through observation or simple analytics. Most importantly, an investigation into the data patterns, particularly the inflow and outflow of employees, can provide insights and actions that can improve organizational performance, product quality, and efficient staffing processes.

Information Technology: Big Data

In 2012 and 2013, Big Data became the hot topic in business literature. The October 2012 editions of American Management Association's journal *M World* and *Harvard Business Review* are dedicated to the topic.

(Jac was commissioned to write "Predictive Analytics and Big Data" for *M World*). The book *Big Data: A Revolution that Will Transform How We Live, Work and Think*, is widely available and explains the issue well with meaningful and readable examples. According to Amazon.com (August 2013), it was the No. 2-selling business book about information management and the No. 1622-selling book overall.

While the topic is seemingly simple in name, the issue is more complex. The industry standard definition of Big Data consists of three data characteristics:

1. **Volume:** The amount of data being created on a daily basis is overwhelming.
2. **Velocity:** The speed and rate of change at which data is collected and processed is difficult to fathom.
3. **Variety:** The range of data types and sources is expanding, making standard analytics difficult to enact.

In the HR area, two things are driving the accumulation of Big Data sets: the automation of HR practices and the consolidation of HR systems. Information technology (IT) systems have made business processes extremely efficient:

- Learning management systems store courses, serve up e-learning, track compliance, manage registrations, issue certifications, and even administer tests.
- Talent systems manage candidate applications, employee demographics, salary information, performance reviews, promotions, and many other things.

The by-product of these efficient systems is a massive source of data. The HR software industry has seen multiple changes:

2004: PriceWaterhouseCoopers acquires Saratoga Institute

2010: Success Factors acquires Info HRM and Plateau, a learning management system

2011: SAP acquires Success Factors

2012: IBM acquires Kenexa

In the near future, we can expect there will be a verifiable Grand Unified Theory of Human Capital. It will explain the driving factors that lead to hiring, promoting, and retaining the best talent available. More importantly, the theory will be tested, confirmed and modified by billions of HR records. Big Data will allow analysts to ignore samples and focus on the big picture — the really big picture.

Due to the individual systems and consolidation of datasets through acquisitions, HR data proliferates. Today, many HR organizations have an opportunity that existed only in the realms of

science fiction for most large companies — an integrated view of HR data from cradle to grave for employees.

In the near future, companies will become more sophisticated with their analytic techniques. Understanding what is happening in an organization will not be sufficient. Leaders will want to understand why things are happening. Correlation will still be useful, but it will be a gateway to more sophisticated techniques such as multiple linear regression, structural equation modeling, logistic regression, linkage analysis and discriminant analysis. Three things will support this change:

1. **Thirst for Information among Leaders:** The C suite will quickly get bored with descriptive statistics, dashboards, and correlations that simply define the current state. They will want predictive analytics, and they will track future performance against predictions.
2. **Tools will be Readily Available:** Advanced statistical tools like SAS, R, SPSS, Minitab, and even some features of MS Excel will enable analysts to dig past the what to the why. These tools are currently available, but you can expect that they will be packaged with systems like Success Factors, Salesforce.com, or Saba so that analysts can conduct statistical tests without downloading the data to an offline system.
3. **Resources will be Abundant:** In October 2012, Davenport and Patil openly stated that the sexiest job of the twenty-first century will be data scientists (business analysts, statisticians, quants, and data jockeys). You may be asking, “Really?” Well, there is a business need, and because of their article, it has been publicly marketed. Additionally, universities are responding by offering analytics degrees. Henschen published a list of the *Big Data Analytics Master’s Degrees: 20 Top Programs* in InformationWeek. Newly minted data scientists should be available to fill positions soon.

Automated Processes: Decision Support

Athletes get better by practicing. They perform, observe their behaviour, and adjust. If they cannot observe their own behaviour, they hire a coach. Performance is measured and goals are set. In this way, athletes create a feedback loop to drive performance improvement.

During a TED Talk in 2013, Raffaello D’Andrea demonstrated how he and a team of scientists have made quadcopters (small helicopters with four rotors) perform like athletes. A quadcopter can orient in space, balance a pole, carry a glass of water, hit a ball back to Raffaello, work in tandem to catch a ball in a net, and even do a backflip. Like the water wheel, this TED video is fascinating to watch. The science behind it is mathematically complex, but the process is simple: Use mathematics to model the flying behaviour, set control parameters around the performance, monitors the performance, and provide a continuous feedback loop in real time.

The continuous improvement process is not new, and organizations currently have various types of monitoring, feedback and action loops in place. The use of a Kaplan and Norton balanced scorecard is a perfect example of such a feedback loop. It begins by selecting key performance indicators along four critical business dimensions: financial, internal business processes, learning

and growth, and customer. Data is gathered, analyzed, and reported in a scorecard. Leaders review the scorecards quarterly and take actions to improve metrics. The TDRP framework is also a viable structure for gathering and reporting data in a continuous improvement loop.

Bringing It All Together

In the near future, the following processes will happen within organizations:

- IT departments will help HR integrate talent systems so that organizations can link data from various talent processes.
- HR will create standard scorecards that gather the right data (efficiency, effectiveness, and outcomes) to demonstrate the influence of HR on the business.
- Data scientists within HR will use advanced analytic approaches to assess what and why to provide insights to business leaders.
- HR departments will set up continuous, automated feedback loops that put information in the hands of decision-makers to increase the efficiency and effectiveness of HR processes.

2.24 PREDICTIVE ANALYTICS FOR HR IN ACTION

The conversation between employees and HR will change dramatically. Involuntary turnover will disappear for all practical purposes because talented employees will be retained. The future of predictive analytics in HR might look something like the situation described next.

Imagine for a moment that you are a newly hired recruiter and you are attending a one-on-one termination meeting with your coach. She has brought you in to learn by shadowing her. The employee who is being terminated, Joe, enters the room. He knows he is being let go but is reluctant to leave. He asks several questions about benefits and then leaves the meeting and the organization. Although you are new to the organization, you know him. He has a reputation as is a solid performer. He is friendly in a professional way and is considered an asset to his team. When he leaves, your coach asks, “Do you have any questions?”

“Just one,” to answer. “Why did you fire him? He is a good performer.”

“Yes,” she says, “He is a good performer. But he is not a great performer. He used to be great, and that’s why he has been here for seven years.” As she talks, she types on her computer. She is opening a file for you to review. “His stellar performance lasted through year five when he was promoted and received a large salary increase. During the last two years, he has plateaued as an average performer.”

“So, average performance in a new role is a reason to fire someone?”

“No, but this file shows us that we have many options. I received an alert last month from the HR system that there is a 75% probability that Joe is likely to leave for a new position that opened at a competitor. That percentage jumped to 95% when he requested a day off last Monday — likely for a job interview. If he had taken a day off on Wednesday, I would not have been concerned. Mondays and Fridays are usually interview days.”

“You said there are options. Many options.”

“Yes.” She points to the screen. “I have an account with an online sourcing system. I entered the job profile for Joe last week, and there are 100 candidates that match the requisition. 90% of them are in the metro area. 20 of them make 25% less than Joe, and have 25% more experience. 15 are men, 5 are women, and 2 of them are minorities. The colleagues on Joe’s team indicated they would like to have a female leader. This is also a chance to expand the diversity of the team.” She points to a name on screen. “This is my top candidate. New hires like to prove themselves, so she is likely to work harder than Joe, and we expect that she will be 21% more productive. If we offer her the job at a salary 19% lower than Joe’s, there is an 85% chance she will take it.”

“But she lives outside the metro area. That will be long commute.”

“The probability jumps to 87% if we give her one flex day per week. It jumps to 99% if we let her work from home and only require her to come to the office for critical meetings.”

Then to ask, “When do we interview her?”

“In 15 minutes.”

“I better get ready then,” you say as you get up to leave the office.

“You’ll find her resume in your e-mail. Just sent it to you.”

Before you exit, you stop at the door and ask, “Can I see the report you had about me prior to my job interview?”

She laughs. “Yes, I’ll share it with you during your performance review. There’s a 90% chance you’ll stay with us long enough to have one.”

Example Measures of Efficiency, Effectiveness, and Outcomes

Talent Development Reporting Principles (TDRP) is mentioned several times throughout this book as a valuable approach for structuring and reporting results to business leaders. More information is available from the Center for Talent Reporting at www.centerfortalentreporting.org. Currently, the center maintains a comprehensive set of more than 600 metrics organized into three categories, as shown next. The goal of using TDRP is to find and measure business outcomes, the value-adding objectives of metrics and analytics. Several metrics that align with the TDRP are shared in Chapters 4, 5, and 6. The next list is more extensive but still is only a small sample of what is available. The critical step for any organization is to define the most important metrics for success and to begin gathering them for monitoring and reporting.

<i>Efficiency</i>	<i>Effectiveness</i>	<i>Outcome</i>
Number of open requisitions	Quality of hire	Length of stay
Number of hires	Quality of the recruiting process	Monthly productivity
Hires by level	Quality of service from recruiter	Improved cycle time in job
Hires by business unit	Fit with the job	Contributions to customer satisfaction

Average cost to hire	Fit with culture	Contributions to quality
Average salary per hire	Performance rating	Increased sales
Amount of training attended	High/low potential status	Increased margins
Amount of coaching received	Employee engagement	Improved project quality
Performance improvement plan	Employee loyalty	Improved on-time projects
Job rotation learning path	Project management capability	Improved on budget projects
International assignment	Ability to coach	Improved efficiency
Mobile workforce	Ability to be coached	Increased cost saving
Hours/job task	Develops others	Increased innovation
Cost/job task	Creates standard procedures	Faster product development cycle

SUMMARY

- Human resource analytics, also known as people analytics, workforce analytics, or talent analytics, entails collecting, analyzing, and reporting HR data. It allows your company to assess the influence of a variety of HR KPIs on overall business performance and make data-driven decisions. HR analytics, in other terms, is a data-driven approach to Human Resources Management.
- The obstacles of exploring the data once it has been examined include preserving the information for simple access and sharing it while maintaining privacy. For HR experts, these issues lead to one conclusion: In today's company world, having a good HR system is essential.
- Text analytics is the act of automatically converting vast amounts of unstructured text into numerical data in order to identify insights, trends, and patterns. This technique, when combined with data visualization tools, allows businesses to comprehend the story behind the numbers and make better decisions.

QUESTIONS

1. Define HR Analytics.
2. Explain the pitfalls of HR Analytics.
3. Explain the different levels of Analytics.
4. Illustrate conducting of HR Analytics.
5. Explain the future of HR Analytics.
6. Illustrate the scope of Big Data in HR Analytics.
7. Explain the scope of Text Analytics in HR Analytics.
8. Who are applying HR Analytics?
9. What is not HR Analytics?

REFERENCES

- <https://www.aihr.com/blog/what-is-hr-analytics/>
- <https://www.aberdeen.com/hcm-essentials/6-ways-big-data-will-change-hr/>
- <https://monkeylearn.com/blog/what-is-text-analytics/#:~:text=Text%20analytics%20is%20the%20automated,numbers%20and%20make%20better%20decisions.>



MODULE 3

APPLICATIONS OF HR METRICS AND CREATING HR DASHBOARDS

Learning Objectives:

1. Understand the basic concept of HR Metrics and how to create HR Dashboards.
2. Develop a plan for Excel formulas and its usefulness for creating Dashboards.
3. Understand the basic applications of VLOOKUP, INDEX, SUMIF, AVERAGEIF and COUNTIF.
4. Understand the Story boarding.
5. Study how to connect the Dots and integrate the findings.

3.1 HISTORY OF HR ANALYTICS

For measuring HR performances, HR metrics are used. HR metrics can report present, ongoing and year-on-year information on HR performances, even in strategic areas such as employee attrition, employee engagement and performance, incentives and employees' performance, training and employees' performance and so on. HR metrics can be developed for all HR operational and strategic areas, interpreting which HR managers can take effective HR decisions. It helps in setting targets for human resources and measuring the performance of an HR department. While developing HR metrics, it is important to see which metrics can be better aligned with the business goals and strategies of the organization. Such metrics can report meaningful HR information, based on which appropriate HR decisions can be taken. For example, human resources can assess the trend in training expenditures, and accordingly, can try to analyze how such training expenditures relate to employees' performances.

Most of the HR metrics that we see today can be credited to Fitz-enz (1984) and later to Fitz-enz and Davidson (2002). All these metrics were developed jointly by Saratoga Institute and Society for Human Resource Management. Kaplan and Norton (1996), however, refurbished our thought process on HR metrics introducing a balanced scorecard. Balanced scorecards design performance indicators from four generic perspectives, *i.e.*, financial, customer, business process, and learning and growth. We will not discuss balanced scorecards in detail, as it is not within the scope of this book.

During the same period, Huselid (1995) introduced the concept of high-performance work systems that could establish how systematic HRM can make significant differences in organizational effectiveness. Further, Becker *et al.* (2001) could improve this concept to develop HR scorecards, which we have discussed in detail in the succeeding paragraph.

Most of the HR metrics focus on data on cost and benefits, rather than intelligently reporting on insights on HR functions that can help in taking critical business and strategic decisions. Such common metrics aggregate the observation to the level of organization. This can only be improved when HR managers take an initiative to align HR metrics to business goals and strategies of organizations and see the information from a holistic perspective.

HR metrics and HR analytics are not one and the same. HR metrics are data that can provide some descriptive details about some HR processes, *e.g.*, the success of a training programme. Therefore, characteristically, HR metrics can report on organizational HR activities. HR analytics, on the other hand, combines data with metrics and can examine the relationships or the changes in HR metrics. From this perspective, HR analytics has significant influence on decision-making. Also, HR analytics can help in predicting the future outcome of a decision, which helps in taking precautionary steps, when decisional negativity has potential adverse effects on an organization.

3.2 DEFINITION AND MEANING OF HR ANALYTICS

HR metrics are simply a measurement of any function that has to do with HR. Measuring the number of staff sick days over a certain period of time would be an example of an HR metric. Tracking HR metrics (or data) can help employers identify trends, which can in turn help with HR planning.

Human resources metrics are different measurements that are used to determine the value and effectiveness of HR initiatives, typically including areas such as turnover, training, return on human capital, costs of labour and expenses per employee that the human resources function provides to the organization. These measurements demonstrate how effective the efforts of the human resources department are to the overall success of the organization.

Why Should Organizations Adopt HR Metrics/Need for HR Metrics?

- HR metrics are a vital method for quantifying the impact and cost of HR processes and employee programs. It is also a strategy for measuring the progress or collapse of HR actions.
- HR metrics can uncover a business's strengths and vulnerabilities and facilitate an understanding of the areas requiring focus or improvement as well as those ready for capitalization.
- From fundamental HR capability to revealing the precise value of each new worker, HR metrics are priceless for evaluating your business and devising future approaches. The most intuitive, user-friendly HR administration programs can make employing HR metrics easy, straightforward and uncomplicated.

- HR metrics should be collected over several months or years so that comparisons can be made over time, highlighting trends that may otherwise go unnoticed. Using employee turnover can be an example of a trend concerning retention or recruitment concerns and analyzed to know why this might be happening.
- Tracking HR metrics (data) with an element of quantity, quality, cost and time will result in a better understanding of how the organization is changing over time (trends).
- Enhanced HR metrics have allowed companies to analyze data within their HR function. HR metrics are invaluable data and alert us to changes or potential problems that should be addressed. Metrics do not show why things are happening, but are a signal to start asking questions.
- HR metrics can help pinpoint where human capital issues exist and allow HR staff to monitor trends. Information from appropriate HR metrics can be used to assess, plan and improve HR as well as to contribute to achieving business strategies and objectives. For example, an increase in turnover rates could be due to a number of factors, including job dissatisfaction, better wages/benefits/working conditions in other sectors, etc.

3.3 METRICS FOR TRAINING AND DEVELOPMENT FUNCTION

In managing the training and development function, we use various metrics. These metrics help in measuring the efficiency of the training and development function in an organization. Some of these metrics are internal accession rate, internal addition and replacement rate, headcount trained rate and so on. All these metrics can be combined to build the total career path ratio, i.e., employee promotion as the percentage of employee movement within an organization. Other than promotion of employees, movement can also happen due to transfer, redeployment and so on. Another important training and development metrics is regular employees' headcount trained ratio, which indicates the number of employees who are being trained as the percentage of regular headcount. Apart from these, we have the training investment factor ratio (represented in terms of average training cost per employee), the training transfer ratio (indicated in terms of the rate of utilization of training inputs for an actual job) and so on.

HR metrics, therefore, provide some standards of measurement, based on which the HR performance can be measured and the measurement results can be used for better HR decision-making.

3.4 MAJOR CATEGORIES OF HUMAN RESOURCES METRICS

There are three major categories of human resources metrics:

1. Metrics that measure the effectiveness of the human resources function.
2. Metrics that measure the efficiency of the human resources department.
3. Metrics that measure the effectiveness of the employees within the organization.

HR Metrics, Data Sources and Requirements

As we know that HR Metrics are measurements that help in tracking key areas in HR data. The most important areas are as follows:

(a) Organizational performance

- Turnover percentages.
- Per cent of regretted loss.
- Statistics on why personnel is leaving.
- Absence percentages and behaviour
- Recruitment (time to fill, number of applicants, recruitment cost, etc.).

(b) HR operations

- HR efficiency (*e.g.*, time to resolving HR self-service tickets).
- HR effectiveness (*e.g.*, perception of HR service quality).

(c) Process optimization

- Process optimization helps to analyze the processes (of how we do what we do) in Human Resource Management.
- The HR metrics and analytics in this area focus on changes in HR efficiency and effectiveness over time.
- These HR metrics and analytics are then used to re-engineer and reinvent the current status in HR which helps to optimize the Human Resource delivery process.
- Process optimization metrics are to a large extent very superior as they require a very high level of both data maturity and analytics maturity.

3.5 HR METRICS AND TYPES OF DATA

Human resources metrics or tracking employee and employer data is a key to effective HR management. Many forms of data are required to be maintained by employers, while other data is helpful to delivering efficient services to employees and determining HR staffers' abilities to meet the needs of the workforce. More than 60% of organizations track turnover rates. The types of data that are tracked are as follows:

- **Employee Count:** The employee base, or the number of workers employed, is critical data for HR to track. Everything from the ratio of HR staffers to employees to the organization's ability to negotiate group health insurance rates depends on the number of employees.

For example, HR best practices suggest a ratio of one full-time HR staffer for every 100 employees. Companies in their growth stages should track employee base numbers to determine how many HR staffers they should hire.

- **Voluntary Equal Employment Opportunities:** Government contractors are required to track applicant and employee data such as race, age, sex, disability and veteran status to use when analyzing their companies' diversity representation. Applicants and employees voluntarily provide this information; however, some organizations report the data based on observation. Regardless of the method, HR tracks this data to demonstrate equal employment opportunity practices and to determine if the company should embark upon specific outreach measures to attract a diverse pool of qualified applicants.
- **Wages:** HR should track current and historical compensation data. Current salaries and wages are obviously tracked for accurate payroll processing; however, tracking historical data on the company's wages is equally important. Historical compensation data is essential for record-keeping purposes, such as the company being required to justify wage-setting practices or investigation by the US Department of Labour Wage and Hour Division for minimum wage and overtime pay practices. Compensation data is also important for conducting wage surveys and comparison to similar businesses to determine if the company is paying competitive wages.
- **Training Expenses:** Tracking data concerning employee training can assist in illustrating the return on investment in employees' professional development. A return on investment can be measured by the number of employees who take advantage of professional development opportunities versus the number who apply what they have learned to their current jobs or use the training to advance within the company. Many organizations track this data according to the average amount expended per employee or the aggregate amount for the HR department's fiscal year spending. Data related to training expense is also useful in determining whether to conduct in-house training to outsource it.
- **Employee Eligibility:** The US Citizenship and Immigration Services enforce the Immigration Reform and Control Act of 1986 which requires that employer's document proof of employees' eligibility to work. The Act requires employers to maintain records, such as Forms 1-9 that contain employees' identification and proof of eligibility, for three years. Tracking this data is imperative and failure to do so can result in serious consequences for employers. In addition, documentation related to foreign workers is important for justifying the employer's sponsorship of a foreign worker.
- **Recruitment and Retention:** Recruitment data may include applicant counts, number of interviews granted and hiring process outcomes for measuring the effectiveness of the company's recruitment and selection methods. Cost-per-hire metrics require this data, and in some cases, analyzing selection data can determine whether hiring managers are conducting appropriate and effective interviews. Retention data, as in the number of employees who stay with the company, and turnover data are valuable data for calculating cost per hire as well.

Tying Datasets Together

One of the most crucial tasks in an analysis is combining data from different sources for analysis. Database people talk about “unique identifiers” — pieces of information that can tie together data sources. Examples include social security numbers, e-mail addresses, and employee ID numbers. Note that there is exactly a one-to-one mapping between a person and each of these items, at least in theory. In practice, there may be data clean-up that needs to occur. For example, an e-mail address might be read by multiple people (*e.g.*, sales@foo.com) or a company may issue a new employee ID when someone transitions between full-time and part-time.

To make connections between datasets, your data analysts will need one or more unique identifiers. Sometimes, these efforts are fairly straightforward, but they can also be challenging. With people, the identifiers are clean, efficiently stored, and unambiguous. They also protect the privacy of the participants. Proper names are messy identifiers. Not only is there the privacy issue, but there are variants of names — maiden and married, with suffixes attached or not, middle names omitted entirely or truncated to an initial, and so forth. If your participants are something other than people (*e.g.*, branch offices), they will still require an identifier. Sometimes, your data analyst will need to make some sort of transformative magic happen to tie things together.

For example, we have seen performance management systems that use an employee ID, a training system that uses proper names, and other systems that use an e-mail address, all within the same organization. We will not get into how common mapping can be created, but a good data analyst will be able to manage this. The circles in the figure represent where information about particular participants in the study may reside. Human resource data usually encompasses other types of data, although the overlap may not be perfect.

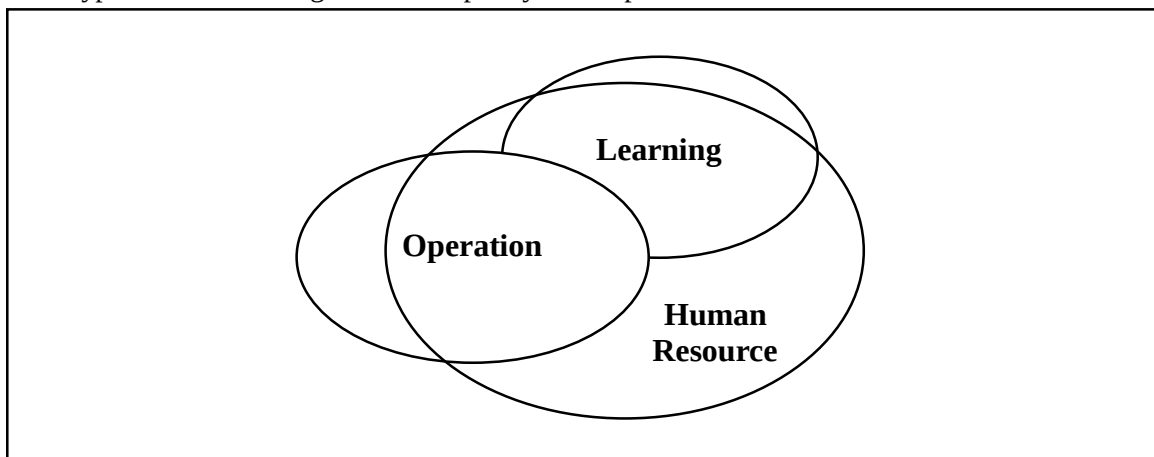


Fig. 3.1: Participant Data

You may notice that figure has areas where data exist only in operation but not in human resources, and so forth. There are often cases where this occurs in practice, although it should not occur in theory. Common explanations for missing data are:

- Many personnel data sources have different start and end points and may include records on terminated employees.

- Systems may have different criteria for including and excluding employees, such as terminated employees, summer interns, contract and temporary workers, and more.
- Some systems may use a convention for identifying employees that does not exist elsewhere (*e.g.*, e-mail addresses are common identifiers in some systems, but not all employees have them).
- Employees may have slightly different identifiers in different systems. For example, proper names can be problematic in very big employee bases; are “Robert M. Johnson,” “Bob Johnson Jr.,” and “R.M. Johnson” one person or three? Proper names should be avoided where possible. Loss rates of 5% to 10% of the entire dataset are not uncommon under these circumstances, although other information, such as birth date or job title, may be used to salvage some of the data.
- Not all employees belong in all datasets. For example, administrative employees may not be recorded in operational datasets.
- Identifiers may change over time. For example, people may change surnames when married or divorced or may be issued a new employee ID or e-mail address when they change job status.
- Human resource data requires revisiting the important questions about participants and how they are identified on a regular basis.

Difficulties in Obtaining Data

As Tolstoy noted about families like happy families are all alike; every unhappy family is unhappy in its own way, unsuccessful data pulls, etc. – all have their own character, and it is sometimes impossible to anticipate all of the possible difficulties. The Tolstoy quote has been generally appropriated by scientists as shorthand for situations where many conditions need to be true simultaneously, and a singly failure creates an overall failure.

Sometimes, the problem is not too few systems containing data but too many. In one study we did with Sun Microsystems, the company needed to consolidate five different LMS systems, which existed due to acquisitions and continual development of decentralized learning tools.

In these types of situations, the data are not always comparable between different systems, and it is not always possible to consolidate the systems prior to pulling data. Many difficulties can be described as ‘politics’. Here are some of the examples of situations which have witnessed in major corporations:

1. People were sent multiple datasets initially, including some sensitive HR files. The senior leader of HR demanded that the information be deleted, which delayed the project for a full year until the proper permissions were obtained. One of the partners on the client side had been a little bit too entrepreneurial in acquiring data for the project and had not obtained the proper approvals prior to sending us the HR data.
2. A highly placed person at a client organization, who was in charge of negotiating the contract, tried to bargain in terms of how much data we would receive. The negotiator

was in the habit of pushing to give the vendor as little of everything as possible. After explaining very carefully that reducing the amount of data was of advantage to no one, we were able to convince the negotiator to change his strategy. The negotiator was able to convince the negotiator to change his strategy. The negotiator was a lawyer, which should surprise no one.

3. Some divisions of a company own certain parts of the business data and share them about as readily as toddlers share toys. Actually, that is not really fair, because toddlers are sometimes in sharing moods and can eventually be taught to behave better.
4. Some stakeholders express apprehension about the results. If they are not very optimistic about what a study will show, they resist getting involved in a project and sharing their data.

Ethics of Measurement and Evaluation in HR Analytics

People Analytics have taken off and the workplace has become a highly instrumented place. Companies use surveys and feedback tools to get our opinions, new tools monitor e-mails and our network of communications, organizational network analysis (ONA), we capture data on travel, location, and mobility, and organizations now have data on our well-being, fitness, and health. And added to this is a new stream of data which includes video (every video conference can be recorded and more than 40% of job interviews are recorded), audio (tools that record meetings can sense mood), and image recognition that recognizes faces wherever we are.

In the early days of HR analytics, companies captured employee data to measure span of control, the distribution of performance ratings, succession pipeline, and other talent related topics. Today, with all this new information entering the workplace virtually everywhere you click at work is stored somewhere, the domain of people analytics is getting very personal. However, it comes accompanied by ethical issues which we need to consider. One such issue is the Risk of Data Abuse, which can be seen when using a new employee engagement tool or “retention risk predictor” from the HR software company. These new systems bring risk. When you buy the system, you really don’t know how it works. So, every decision, recommendation, or suggestion it makes becomes your organization’s problem.

While I know HR professionals take the job of ethics and safety seriously, I’d like to point out some ethical issues we need to consider.

Suppose, for example, you use Pymetrics, HireVue, or another advanced assessment technology to assess job candidates. While these vendors work hard to remove racial, gender, and generational bias from their tools, if you implement them and a job candidate sues you, your company is responsible. And this happens all the time. (Read how Amazon inadvertently created its own gender-biased recruitment system).

Another example is when a company that turned on the “retention predictor” from their HCM system had their managers do all sorts of strange things when they see flight risk. Some managers actually stop talking to these people and reduce the support they get at work because “they’re thinking about leaving anyway.” Obviously, this is not good management, but if we

don't use this data well, people can use it incorrectly. And of course, there are other things that could go wrong. If you have access to employee health data and you use it to assess or discuss an employee's performance, a legal case can be slapped on.

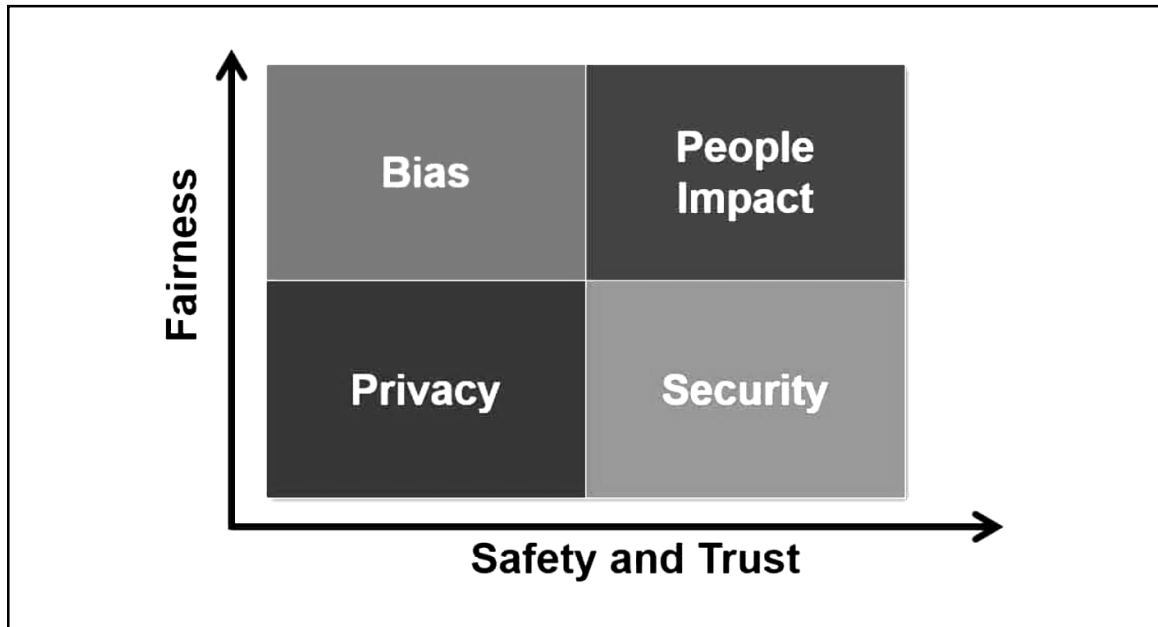


Fig. 3.2: Framework for Ethics of Data and AI

As you can see from this framework, there are two dimensions to ethics.

First, is the data and algorithm you're using fair? Does it accurately reflect the performance or productivity data you want without excluding, discriminating, or inadvertently biasing the result? This is tricky and I will discuss it below.

Second, is the data system and algorithm safe? Are we protecting privacy, confidentiality and security? Who has access and how do we audit its use and path through the company? This is a well-known problem in IT but now one we have to deal with in HR.

When you look at these two dimensions, you essentially find that there are four dimensions to trust.

PRIVACY	SECURITY	BIAS	PEOPLE IMPACT
- Is my information private? - Do I know what information is captured? - Can I see the information? - Can I edit or repair information that is incorrect? - Can I be forgotten? - What happens if it is leaked?	- Is the information encrypted and secure? - Who has access to seeing it? - What are the cyber protections? - What rights do I have for damages if lost? - Does the organization hold itself accountable?	- How will my age, gender, race, education, degree and other demographic information be used? - What safeguards and processes are there to remove bias? - Is the AI auditable? - What studies and tools can I see?	- Is this information being used to save money or reduce costs? - How will this make people's work life better? - How will this improve customer satisfaction? - How will this make the workplace or team better?
Facebook CVS Yahoo Vizio (TV Sets) Ashley Madison	Marriott: 500+ Million Equifax: 147 Million Under Armor: 150 Million Target: 110 Million Quora: 100 Million Facebook: 50 Million	Amazon – Hiring Credit Scores – Experian Police Images – Implicit bias Mortgages	Everyone?

Fig. 3.3: Ethics in People Analytics: Four Dimensions of Trust

1. Privacy: The first ethical issue to consider is privacy. As the chart above shows, companies like Facebook, CVS, Yahoo, and many others have gotten in trouble here. When an employee joins your company, they give the rights to collect a lot of data, but we as employers do not have the rights to expose this data, share it, or link it with personally identified information.

2. Security: The sister of privacy is security. Is the data stored and protected in a place where others cannot find it? Do you have password policies, encryption, and other data protection practices in place so that an employee cannot take the data home, send it to a third party, or accidentally release it into the internet? These are IT issues which all companies have to deal with, and when we have sensitive information like pay, job history, healthcare data, and other personal information we have to protect it well.

3. Bias: The third and most difficult (and new) problem we have in People Analytics is bias. Whether you are analyzing the data yourself or buying an AI tool from a vendor, we have to remember that all algorithmic systems are based on existing data. And if the existing data is biased, the predictions and recommendations will be biased. This is a very difficult problem to solve, and many organizations are working on it. (IBM Research has an excellent video on this topic.) For example:

- Systems that try to assess fair pay will compare an employee to peers but may not understand issues of race, location, and age.
- Systems that predict retention may discriminate against minorities or others who leave the company for cultural reasons.
- Systems that assess fit to a job may institutionalize old, discriminating hiring practices that are embedded into hiring history.
- Systems that use organizational network analysis to identify performance may not realize that gender or age plays a huge role in trust and relationships.
- Systems that predict high performers will be biased toward existing highly rated individuals (who may be white men).

Every predictive analytics system you buy or build is going to have bias built-in. (The word “bias” means “influence based on the past,” which is really what AI is trying to do.)

4. People Impact: As GDPR rules clearly state, it’s not ok to capture data to “see what it might tell us.” If employees believe they are being monitored for the wrong reasons, the impact will be negative.

Most of the vendors have the best of intentions.

- Phenom People’s new talent experience platform uses AI to help job candidates find the right opening, helps internal candidates find the right job, and helps the chatbot ask you intelligent questions to understand your job needs.
- Glint’s new Manager Concierge uses AI to recommend behaviour changes and courses to help you become a better leader. ADP’s Compass tool and Zugata by Culture Amp does the same. Humu is doing this for team and operational performance.
- Watson Candidate Assistant from IBM uses your resume to identify your skills as a job seeker and finds you the best job, dramatically improving quality of hire and time to hire.
- EdCast, Valamis, Fuse, and Volley are using AI to recommend learning content, and BetterUp uses AI to find you the best coach.
- Oracle, Workday, and SuccessFactors use AI for many features. Oracle HCM recommends salary adjustments and even customizes the screens you see based on your own role and behaviour, simplifying the system itself.
- And vendors like Spring Health now use AI to diagnose your mental health and recommend the right tips, counsellors, or doctors.

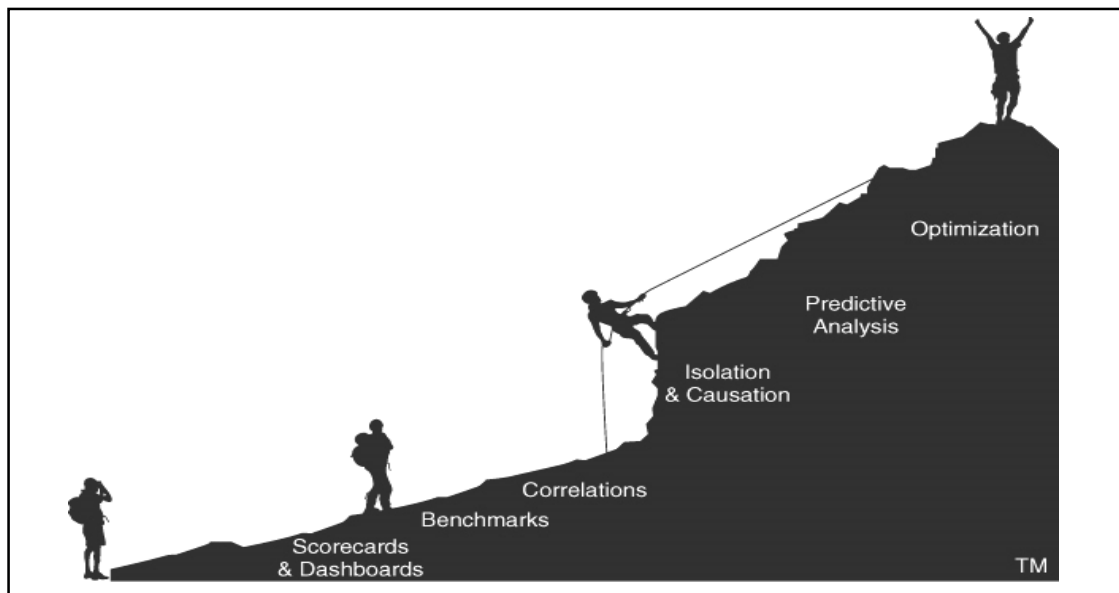


Fig. 3.4: Continuum of Human Capital Analytics

The continuum is based on what we have seen in our work, starting with simple, commonly used techniques. Viewing the continuum as a mountain, we suggest that similar to mountain climbing, things become more difficult but the view improves as you reach higher ground.

- The ascent begins with **anecdotes or storytelling**. This well-thought-out interview has important applications in clinical psychology, anthropology, expert systems, and many other areas. Scorecards and dashboards are other important areas.
- **Scorecards**, most notably Balanced Scorecards, are a strategic performance management tool that can leverage automated surveys to track how an organization executes strategy and the consequences arising from business processes, most commonly referred to as activity metrics. Scorecards characteristically have a mixture of financial and non-financial measures, each compared to its targets, all within a single concise report. They are an important step on the continuum, because this is where you must lay out the basic assumptions: what are the strategies, and what are the various ways in which you will measure them?
- **A dashboard** is a distillation of the most important key performance indicators of a company that an executive can view at a glance. Dashboards might be an *ad hoc* effort put together on spreadsheets or even lower-tech tools, or they may involve special-purpose programming.
- **Benchmarks** are another step on the continuum. Benchmarking has long been used as a standard tool; the idea is that studying the best-run companies in a specific area can be very beneficial in terms of setting things such as salary, training levels, desired turnover rates, and so forth.
- **Correlations and causations** are the next two stages in the continuum.

Correlations are used to describe the descriptive statistics that might occur on a sophisticated dashboard. For example, Where are sales highest? Did the trained employees outperform the untrained ones? These are rich data mines for understanding business resources and human capital.

Causation is the next level beyond correlation. Two events must show a clear and statistically significant connection. One event must precede the other and all other probable causes must be ruled out and the final stage is viewed as optimization, the holy grail of HR measurement.

- **Optimization** is having the intelligence to understand where the impact is occurring. Optimization is closely wrapped up with causation. Without understanding all of the factors that control impact, it is impossible to be sure that you have correctly assessed impact. The really positive side of having assessed and measured the different factors that control and mediate impact is that you can use them to control future impact and improve outcomes. If you have controlled and assessed the impact of the various factors, you are able to use them prescriptively. If you know how much tenure controls performance and how the various tenure levels benefit from training, the logical

consequence is that you can now specify where training should be focused and what new programs need to be created for areas that do not show the benefits.

On the basis of a typical survey, the audience comes to know where its members are on this continuum. Based on the feedback, most organizations are on the lower left of the continuum, obtaining information from their dashboards and scorecards.

So, human capital analytics continuum enables to optimize the human capital investments in the world of predictive analytics. There are, of course, other ways to describe the continuum of human capital measurement. Several different measurement methodologies are particular to learning and development but have limitations when we apply advanced analytics to them.

In developing an organization's continuum, a broader view of all human capital investments, independent of learning and development use is strived for which is but a necessity for HR.

The application of predictive analytics allows to incredible insights into the workforce and enables to improve HR investments, both financially and for the employees.

3.6 HR DASHBOARDS

From an HR scorecard, we can deduce HR dashboards primarily for monitoring and measurement of HR activities based on various operating metrics. Hence, dashboard *per se* is not an HR scorecard, nor is it aligned with the business and strategies of organizations. It primarily focuses on short-term or operational goals, helps in visualization of the performance, and reports immediately the deviation for interventions. With charts and diagrams, HR managers can track and monitor the performance. Dashboard development requires metrics in KPIs format.

A HR dashboard is an advanced analytics tool that displays important HR metrics using interactive data visualizations. It helps the HR department to improve recruiting processes, to optimize the workplace management as well as to enhance the overall employee performance.

The Human Resource Dashboard and HR Report are an important part of managing human resources. Both tools are part of HR reporting and form the basics of informed decision-making. Why?

When done right, it offers three key benefits for both HR and management:

- HR monitoring
- Management information
- Track problem areas.

An automated HR dashboard enables the HR to keep a close view on the key aspects of an organization by tracking workforce metrics. S/he can track emerging trends, opportunities and well as potential problems so that they can be addressed in advance.

HR dashboards have come to play a critical role for data-driven HR departments. Besides providing stakeholders with important information at a glance, HR dashboards also give HR personnel the quick insights they need to make important talent management decisions, through things such as data visualizations. The most popularly used dashboards are as follows:

- BambooHR
- Dayforce
- Gusto
- Paycor's Perform
- UltiPro
- WebHR

3.7 HR ANALYTICS AS A BETTER TOOL FOR HR DECISIONS

HR Analytics today is the most important strategic tool for the HR decision-making process. While HR metrics and tools such as HR scorecards and HR dashboards can monitor and control HR operational functions, HR analytics, making use of HR metrics, can develop algorithms for HR decisions in alignment with the business goals and strategies of organizations. HR analytics gather HR data from multiple sources and then develop algorithms for predictive HR decisions. Today, most of the organizations across the globe manage their human resources or human capital with strategic focus using HR analytics. Some of the strategic benefits of HR analytics are: managing talent attrition, anticipating performance, effectively managing compensation and benefits programme, improving employee motivation and morale and so on.

Beginning of HR analytics can be credited to Iac Fitz-enz, when in 1984 he could develop ideas of measuring HRM services with metrics. Although our initial focus was on measuring employee performances, with the passage of time, HR Analytics has developed (rather evolved) for managing the entire gamut of HRM or HC management strategically. Organizations using HR analytics as a strategic tool can not only measure but also improve critical HR issues such as employee satisfaction, retention, compensation design, workforce planning, performance management and identification of employees for a future leadership role. HR analytics gather workforce data, prepare algorithms and feed the same in advanced computer models to get insights into the past, present and future trends for critical decision-making. For example, a typical decision to change incentive plan may be tested in advance using HR analytics, in terms of its positive or negative effect on talent retention. Such impact analysis in advance, therefore, ensures a better HR decision-making process. With improved HR decisions, organizations also benefit, as human resources can be made more business aligned.

With HR analytics, HR managers can analyze data to draw hiring strategies, highlight business opportunities and forge the best career paths for top performers. Today, HR analytics can gauge talent availability, predict talent retention and can objectively determine the implications of all HR decisions in advance. In August 2012, IBM bought talent management software provider Kenexa for a whopping \$1.3 billion. SAP (with Success Factors) and Oracle (with Taleo) also made acquisitions to enter this field while smaller players such as Visier and Evolv gain ground with highly scalable, cloud-based tools. Therefore, HR analytics is an evolving and expanding field of HR decision-making process.

3.8 STATISTICAL SOFTWARE USED FOR HR ANALYTICS

There are some statistical software used for HR Analytics as follows:

1. Excel

Excel is where most of us started. Whenever you manually extract data from any of your HR systems, it most likely comes out in the form of a comma-separated value (CSV) file. These files can easily be opened and edited using Excel.

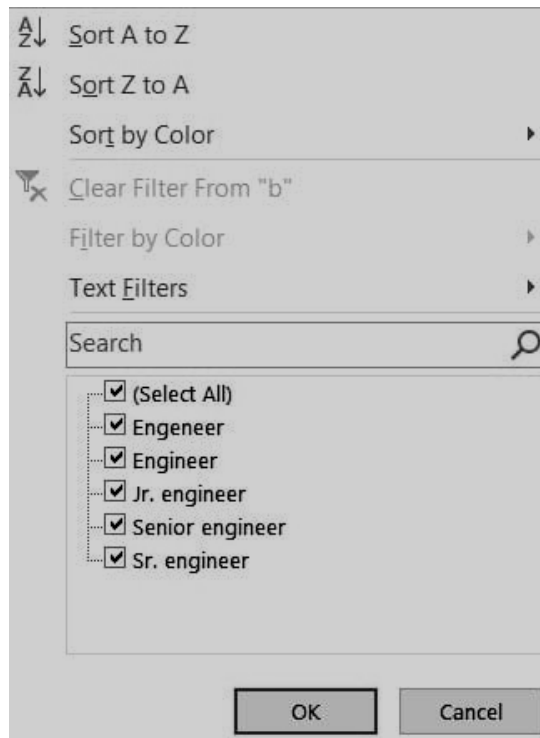
The good thing about Excel is that it's very intuitive to most of us HR data geeks and therefore easy to use.

For example, if you wanted to check how clean your data is, you can quickly transform a dataset into a table and check each column's data range for outliers.

This way, if you select the age column, you can quickly check the minimum and maximum ages. You wouldn't expect anyone below 16 to work at your company, nor would you expect anyone over the age of 80 to work for you. You can find these outliers in a single click.

Some quick tips on how to use Excel for HR analytics purposes:

- If you want to run advanced analyses, load the Analysis ToolPak in Excel. This package enables you to do advanced analytics, including correlation and linear regression.
- When you work with large files, transform them into Tables. Excel can work much more efficiently when data is structured in a table.
- Don't use Excel formulas in large datasets. When you calculate a column using an Excel formula, transform the outcome to a numeric value. Formulas recalculate every time you make a change in the dataset. This places a significant and unnecessary burden on to computer's memory and processing speed – and bogs down Excel.
- Categorical variables (Gender: Male, Female) are easy to check in a table. Select the table column and check for errors or inconsistencies. Can you spot the discrepancies in columns in this picture?



- If you want to merge datasets, the 'VLOOKUP' function is your best friend. It makes connecting two separate datasets very easy.
- Pivot tables do a great job in summarizing large quantities of data. Pivot tables and the VLOOKUP function practically enable you to do HR analytics in Excel.

2. SPSS

Whereas Power BI, Tableau, and Qlik are mostly used to aggregate data, SPSS is used to actually analyze data.

SPSS is one of the most commonly used HR analytics tools in the social sciences. Thanks to its user-friendly interface, you are able to analyze data without having extensive statistical knowledge. Because SPSS is often used in the social sciences, a lot of HR professionals know how to use it – especially the ones interested in data analysis.

This is also the reason why we put SPSS on the list and not its biggest competitor, SAS. SAS has more users outside of the social science field. However, SAS has a steeper learning curve. SPSS also shares many similarities with Excel, which makes it easier to work with.

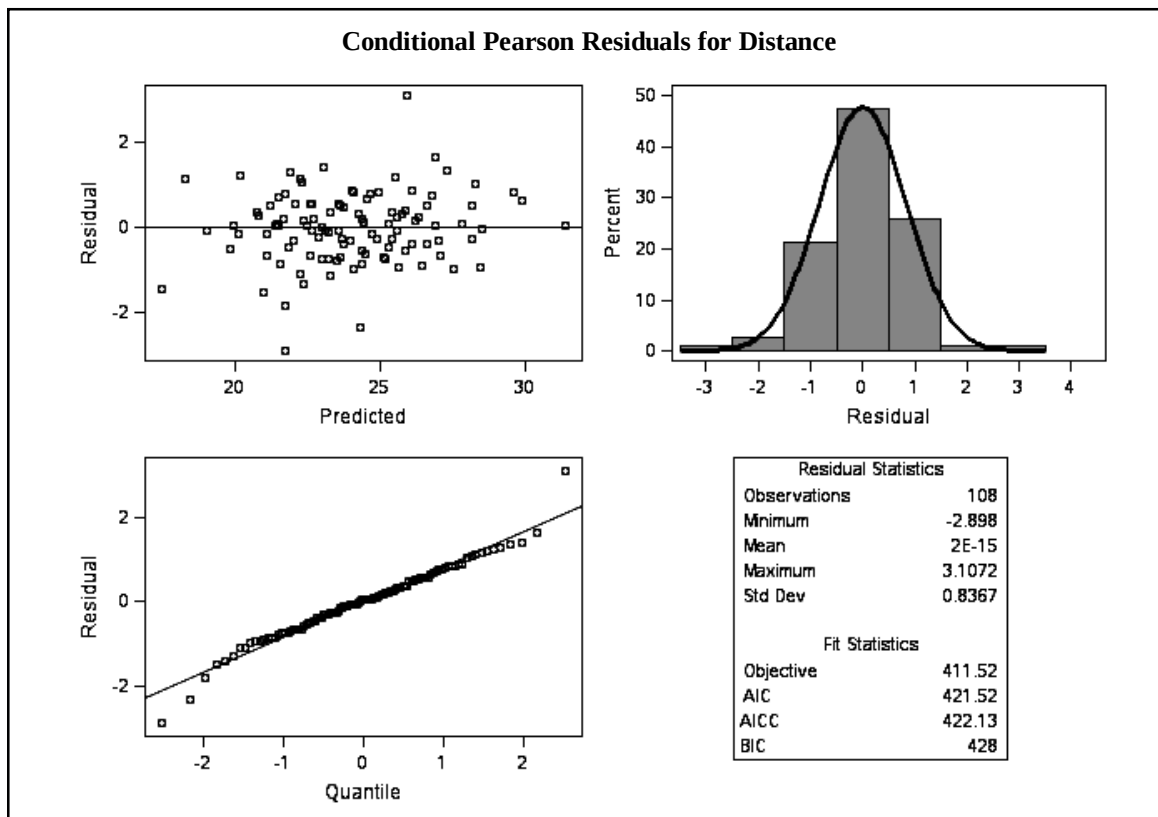


Fig. 3.5: An Example of Descriptive Statistics Output Generated by SAS

Consider SPSS an easy stepping stone for companies with less mature analytical capabilities. SPSS makes it easy to do an exploratory correlation analysis or a quick regression analysis. For more complicated (machine learning) algorithms, R is the better candidate.

3. AMOS

It is statistical software and it stands for analysis of a moment structures. AMOS is an added SPSS module, and is specially used for structural equation modeling, path analysis, and confirmatory factor analysis. It is also known as analysis of covariance or causal modeling software. AMOS is a visual program for structural equation modeling (SEM). In AMOS, we can draw models graphically using simple drawing tools. AMOS quickly performs the computations for SEM and displays the results.

Construction of Model in AMOS

First, we have to run AMOS. By clicking the “start” menu and selecting the “AMOS graphic” option, we can run the program. The moment AMOS starts running, a window appears called the “AMOS graphic”. In this window, we can manually draw our SEM model.

- **Attaching Data:** By selecting a file name from the data file option, we can attach data in AMOS for SEM analysis. This option also appears if we click on the “select data” icon.
- **Observed Variable:** A rectangle icon is used to draw the observed variable.
- **Unobserved Variable:** A circle icon is used to draw the unobserved variable.
- **Cause-and-effect Relationship:** A single-headed arrow in AMOS is used to draw the cause-and-effect relationship between the observed and unobserved variables.
- **Covariance:** A double-headed arrow is used to draw the covariance between variables.
- **Error Term:** In AMOS, the error term icon is next to the unobserved variable icon, and it is used to draw the latent variable.
- **Naming the Variable:** When we right click on a variable in a graphical window, the first option, “object properties” is used to give the name of the variable in AMOS.

4. SAS

SAS is a leader in business analytics. Through innovative analytics, it caters to business intelligence and data management software and services. SAS transforms data into insight which can give a fresh perspective on business.

Unlike other BI tools available in the market, SAS takes an extensive programming approach to data transformation and analysis rather than a pure drag drop and connect approach. That makes it stand out from the crowd as it gives much finer control over data manipulation. SAS has a very large number of components customized for specific industries and data analysis tasks.

SAS stands for **Statistical Analysis Software**. It was created in the year 1960 by the SAS institute. From 1st January 1960, SAS was used for data management, business intelligence, predictive Analysis, Descriptive and Prescriptive Analysis, etc. Since then, many new statistical procedures and components were introduced in the software.

With the introduction of JMP (Jump) for statistics, SAS took advantage of the **Graphical User Interface** which was introduced by the Macintosh. Jump is basically used for the applications like Six Sigma, designs, quality control and engineering and scientific analysis.

SAS is platform independent which means you can run SAS on any operating system either Linux or Windows. SAS is driven by SAS programmers who use several sequences of operations on the SAS datasets to make proper reports for data analysis.

Over the years, SAS has added numerous solutions to its product portfolio. It has solution for Data Governance, Data Quality, Big Data Analytics, Text Mining, Fraud Management, Health Science, etc. We can safely assume SAS has a solution for every business domain.

To have a glance at the list of products available, you can visit SAS Components.

Why We Use SAS?

SAS is basically worked on large datasets. With the help of SAS software, you can perform various operations on the data like:

- Data Management
- Statistical Analysis
- Report Formation with Perfect Graphics
- Business Planning
- Operations Research and Project Management
- Quality Improvement
- Application Development
- Data Extraction
- Data Transformation
- Data Updation and Modification

SAS Component and their Usage

If we talk about the components of SAS, then more than 200 components are available in SAS.

- **Base SAS:** It is a core component which contains data management facility and a programming language for data analysis. It is also the most widely used.
- **SAS/GRAPH:** Create graphs, presentations for better understanding and showcasing the result in a proper format.
- **SAS/STAT:** Perform Statistical analysis with the variance analysis, regression, multivariate analysis, survival analysis, psychometric analysis and mixed model analysis.
- **SAS/OR:** Operations research.
- **SAS/ETS:** Econometrics and Time Series Analysis
- **SAS/IML:** Interactive matrix language in C
- **SAS/AF:** Applications facility
- **SAS/QC:** Quality control
- **SAS/INSIGHT:** Data mining
- **SAS/PH:** Clinical trial analysis.
- **SAS/Enterprise Miner:** Data mining.

5. R Programming

R is the most used HR analytics tool. It is great for statistical analysis and visualization. It is well-suited to explore massive datasets. It enables you to analyze and clean datasets with millions of rows of data. It also lets you visualize your data and analysis.

The most often used Integrated Development Environment, or IDE, for R is RStudio. An IDE is software that provides additional facilities for software development and data analysis. This makes the software more user-friendly.

Simply put, RStudio does everything that R does but more and better. The RStudio interface contains a code editor, the R console, an easily accessible workspace, a history log, and room for plots and files. The picture below shows all these elements.

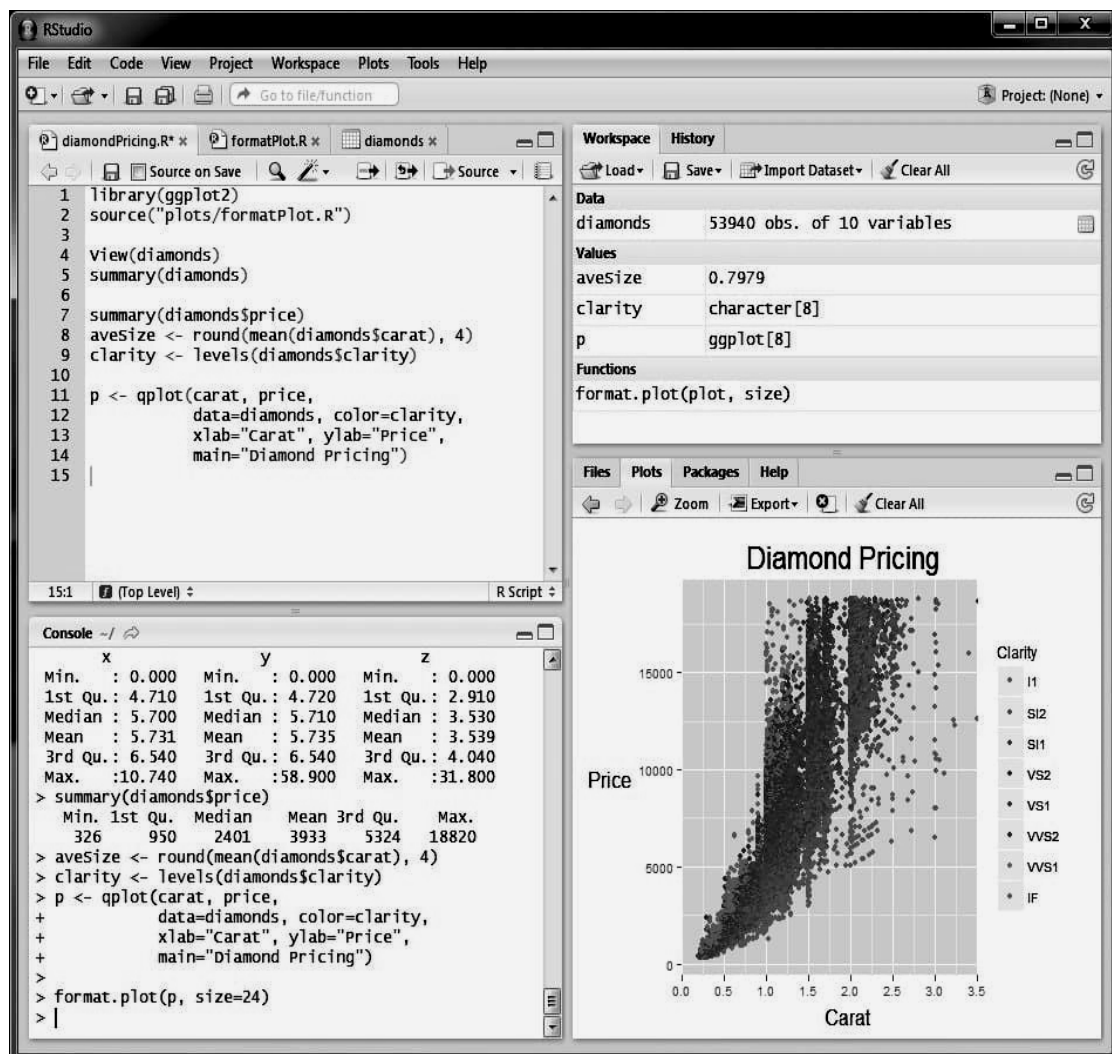


Fig. 3.6: R Programming

As previously stated, R is useful because it enables you to work with much larger datasets compared to, *e.g.*, Excel. Furthermore, R has a very extensive library with R packages.

These packages are easy to install and allow you to run virtually all statistical analyses and create beautiful visualizations. Take, for example, the **caret package**. This package enables you to split data into training and testing sets to train algorithms using cross-validation.

Another example of an R package is **ggplot**, which helps you to visualize graphs. In an article on **R Churn Analytics**, Lyndon showed the distribution of employee turnover for a large Canadian company as seen in the following chart using ggplot.

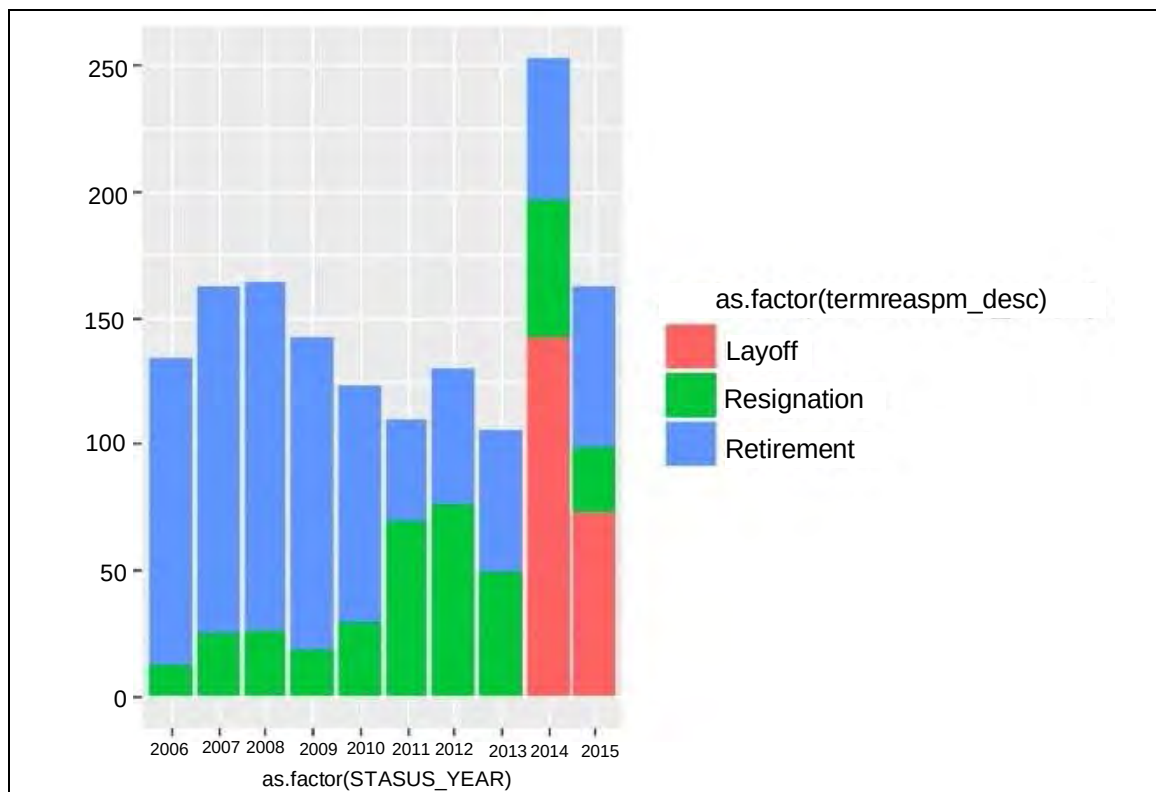


Fig. 3.7: RStudio

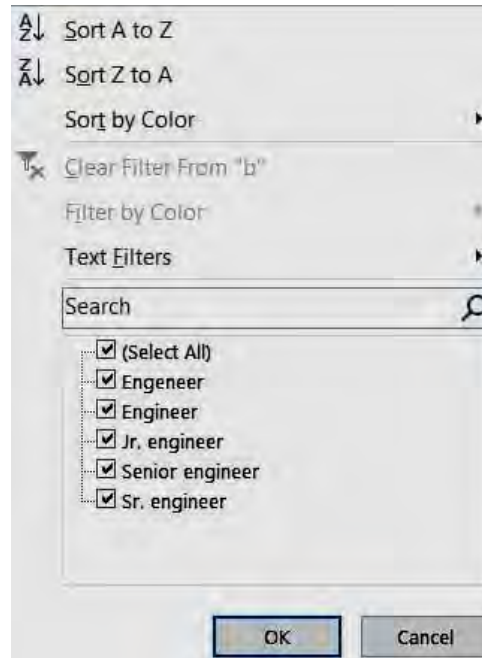
All in all, R is an excellent tool for analyzing and visualizing vast amounts of data.

3.9 DATA VISUALIZATION TOOLS

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

Common General Types of Data Visualization



- Charts
- Tables
- Graphs
- Maps
- Infographics
- Dashboards

More specific examples of methods to visualize data:

- Area Chart
- Bar Chart
- Box-and-Whisker Plots
- Bubble Cloud
- Bullet Graph
- Cartogram
- Circle View
- Dot Distribution Map
- Gantt Chart
- Heat Map

- Highlight Table
- Histogram
- Matrix
- Network
- Polar Area
- Radial Tree
- Scatter Plot (2D or 3D)
- Streamgraph
- Text Tables
- Timeline
- Treemap
- Wedge Stack Graph
- Word Cloud
- And any mix-and-match combination in a dashboard!

1. Tableau

Tableau is very similar to Power BI in that it enables the aggregation and visualization of various data sources. Founded in 2003 as a commercial outlet for research produced at Stanford University, the software has taken the visualization world by storm.

Tableau is arguably the best business intelligence tool (BI tool) out there when it comes to visualization. It has been recognized in the Gartner Magic Quadrant for seven consecutive years between 2012 and 2019.

The image below shows an employee referral dashboard. The dashboard shows how long different recruiters take to resolve referrals. By sharing this visualization with recruiters, average response time to these referrals dropped significantly (see below, Q1, Q2, and Q3).

The disadvantage of Tableau is that it is more expensive compared to competitor Power BI.

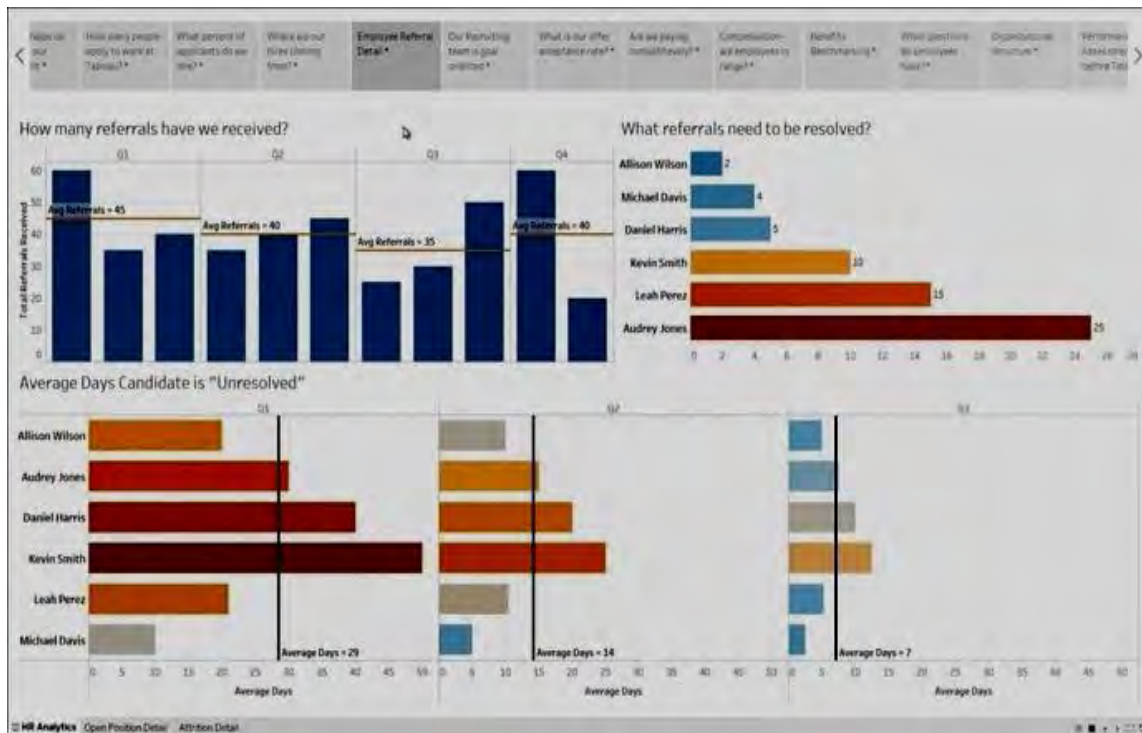


Fig. 3.8: Tableau

2. Qlik

Similar to Visier, Qlik aggregates data. The main difference is that Visier focuses on actionable insights on workforce data, whereas Qlik is a more general data aggregation tool with extensive dashboarding capacity. This means that Qlik is very suitable for more general data aggregation, warehousing, and dashboarding.

Just like Power BI and Tableau, Qlik has been a leader in the Gartner Magic Quadrant for several years.

3. FusionCharts

This is a very widely used, JavaScript-based charting and visualization package that has established itself as one of the leaders in the paid-for market. It can produce 90 different chart types and integrates with a large number of platforms and frameworks giving a great deal of flexibility.

One feature that has helped make fusion charts very popular is that rather than having to start each new visualization from scratch, users can pick from a range of “live” example templates, simply plugging in their own data sources as needed.

Features of FusionCharts

- Completely free of cost.
- PHP Wrapper class for quick and easy embedding of charts.
- Gantt and Funnel Charts.
- Code samples including PHP and ASP.NET.

Advantages of Using FusionCharts

- **Animated Charts add a “wow” factor to your applications and presentations:** It is rightly said that “A picture is worth a thousand words”. With FusionCharts, you can take it a step further by animating it. By giving an exciting face to the boring data tables, you will easily be able to sweep the audience off their feet. Different chart types support different forms of customizable animation.
- **Interactive options that image-based charting solutions cannot offer:** Since FusionCharts is based on Adobe Flash Platform, it offers a lot of interactive options that traditional image-based charting solutions cannot offer. Few examples are scroll charts, self-updating charts, drag-able charts, visual data selection and manipulations features, AJAX compatibility, etc.
- **No installation hassles:** To use FusionCharts, to do not have to install anything on your server. All you need to do is copy paste the SWF files (the core files of FusionCharts) to your server, just like you would do to any image files – and you are ready to go! So, even on those servers which do not allow installation of ActiveX or any other form of components, FusionCharts can run without any hassles.
- **Easy to use:** FusionCharts makes the chart creation process a painless experience for you. Since it uses XML as its data interface, all you need to do is convert your data into XML using a programming language or using the visual GUI provided in FusionCharts — and that’s all what is required to create interactive and animated charts.
- **Runs on a variety of platforms:** Irrespective of what server-side scripting language you are using, FusionCharts can be used to create charts in that. Since Fusion Charts uses XML as the data interface, you can run it on any server and against any scripting language. Also, to view the charts, your users just need to have Adobe Flash Player 8, which is one of the most used browser plugins on the planet.
- **Reduces load on the servers:** In traditional image-based charting systems, the charts are generated as image at the server side. So, for each chart that you need to serve to a user, you will have to build complex images on the server and then stream it to the client. When the need of hour is high, this can be an expensive resource on the server, as image creation takes a lot of toll on the server. FusionCharts brings you great relief – all the charts are rendered at client side using the widely installed Adobe Flash Platform. The server is just responsible for streaming the pre-built SWF files and your

XML data files to the end viewers. From thereon, Flash Player takes the onus of rendering the charts. Also, the chart SWF Files can be cached so that you can just update the data, and not send chart SWF files every time.

- **A plethora of chart types:** FusionCharts offers you a plethora of chart types. From the basic bar, column, line, pie etc. to the advanced combination and scroll charts, you can build all the charts with the same ease of use.
- **Clear licensing policies resulting in a major cost savings:** FusionCharts offers to a clear-cut nonsense licensing model. We do not charge for per-testing server or per-production server. Once you buy a license of FusionCharts, you can host it on multiple servers, if all those servers are hosting the same website. We offer unlimited development servers, unlimited test servers and unlimited CPUs on a given server with each license of FusionCharts. Moreover, if you are using multiple technologies for the same application – like say, ASP for a few modules and PHP for a few modules, you can use FusionCharts for both of them, without having to pay us anything additional. This saves you a lot on licensing cost, and renders a generic look and feel to all the charts in your application.
- **Easy yet powerful AJAX/JavaScript integration:** FusionCharts offers advanced options to integrate charts with AJAX applications or JavaScript modules. You can update charts on client side, invoke JavaScript functions as hotspot links, or make dynamic calls for XML data to server without involving any page refreshes.
- **Open-source product:** FusionCharts is an open-source commercial product. Our open source policy buys you peace of mind, as you can modify or change the source code whenever and however you wish. Moreover, you can use our framework to create new types of charts for your own applications.
- **Personalized support:** As a FusionCharts customer, you will get personal support from our proven support team. Additionally, we have Online Forums and Knowledge Base that can provide ready-made solutions to most of the technical issues.

3.10 STAFF OR SKILLS MATRIX

What is a Staff or Skills Matrix?

A **skills matrix** is a framework used to map employees' skills and their levels. It is a grid that contains information about available skill and their evaluation.

It is used to manage, plan, and monitor existing and desired skills for a role, team, department, project, or an entire company.

Sometimes, a skills matrix is also called a competency matrix. We will cover the difference down below, but in general, they are the same thing.

A simple table with employees' names and their skills might be called a skill matrix.

Example of a Skills Matrix

As we mentioned before, the simplest skill matrix is a table with employees' names and their skills or competencies with their evaluation.

Team Member	Communication	Campaign Management	Content Marketing	Webinar Management	Social Media Marketing	Design Skills
John	8/10	7/10	9/10	7/10	3/10	2/10
Elizabeth	4/10	5/10	7/10	6/10	8/10	3/10
Robert	6/10	9/10	5/10	8/10	5/10	7/10
Victoria	8/10	9/10	7/10	9/10	7/10	6/10
Edward	7/10	4/10	4/10	3/10	4/10	8/10
Anne	7/10	4/10	6/10	7/10	9/10	3/10

Fig. 3.9: Example of a Skills Matrix

Even a simple matrix can allow you to discover any missing skills that are needed or determine who needs some training.

Why is a Skills Matrix Important?

Skills matrices are essential tools for any company that is driven by data, particularly for the HR department and project management team leaders. Matrices are ideal for keeping track of the staff's skills, qualifications, certifications, and competencies throughout the whole organization. When used correctly, they work really well in practice and improve efficiencies within teams, as well as increasing the bottom line.

A skill matrix is important because when it is set up well, it can help you organize your teams easily, identifying any skill sets that you need based on the project requirements. With a good skill matrix, the project is completed efficiently and by the most qualified and skilled team members.

It is also helpful to the HR department, in that if they need to hire extra staff, they will know what skills to advertise for. At the same time, it can be used to schedule vacation periods for employees, making sure you have enough people with needed skills on site.

What is the Difference between a Skills Matrix and a Competency Matrix?

The concept behind them is totally the same. The only difference is usage and level. A skills matrix, as we mentioned, mostly contains employees' skills evaluation, but this is a small part of the knowledge that they have. A competency matrix includes knowledge, skills, and attitudes. So, it should even encompass a skill matrix within. A competency matrix is mainly created and used by management to understand available knowledge in the company, identify gaps, and plan accordingly.

Benefits of Creating a Skills Matrix

1. Identifying the right people

A skills matrix allows you to select the right people for the job, task, or project. It will help managers to form better and more productive teams and fill positions with the employees who are the best fit for that role.

2. Identify missing competencies

Skills matrices help to determine what skill set to are missing, whether within a team, department, or the company as a whole.

If the staff lacks certain knowledge or competencies that are necessary for the business and the competitors possess them, then to a company is at a huge disadvantage.

Also, when you know what skill set you need to start a new project, campaign, or business, it is easier to plan for it, meet deadlines, and achieve successful results.

For team members, a competency matrix can give insight into the strengths of the team as a whole and also show the areas where they are lacking skills and expertise.

3. Identify the gap between employees, teams, and departments

“A chain is only as strong as its weakest link” — Thomas Reid, *“Essays on the Intellectual Powers of Man,”* 1786.

Knowing where you are lacking the skills can save time and money for our business.

A skills matrix can help you identify employees with knowledge or skill gaps, and provide the needed training, as you rotate employees between crucial projects or teams.

4. Track employee development

Skills matrices give the L&D department information that they can use to determine what training is needed for employees. By using the skills matrix, they can identify training opportunities.

It also helps employees to understand what their gaps are, so they can take action to gain the skill set they need to excel in their position.

Also, the skill matrix can be used as a template in an employee’s career planning and development. Knowing what skills you need for the next level or promotion helps to create a learning path.

5. Help HR to find the right candidate

If someone leaves your organization, it is easier to understand what skill set you lost and who you need to hire when you are using a skills matrix.

It speeds up the hiring process and helps the HR department to hire staff with needed skills, making it more efficient, and providing a better end result.

Also, it can be used for the future planning of new hires.

6. Track the key staff members

With a competency matrix, it is easy to find the employees who have the highest value skill set in your organization, as well as to track the development of key people. These are valuable employees, and it is good to know exactly who they are and are they satisfied.

Also, it can be used for Identifying staff for promotion and recognition.

How to Create a Skill Matrix?

1. Create a skill database

Skill database is the list of skills required for the role, department, or project.

At this stage, you need to define required skills by gathering and grouping them into different categories and even subcategories, e.g.:

- Communication
- Management
- Technical skills
- Technologies
- Problem-solving skills
- Data analysis
- Graphic design
- Marketing, etc.

Be as specific as possible when determining what skills are needed, and remember that **soft skills** such as communication, teamwork, and reliability are just as important as the **hard skill sets** such as knowing the technology or techniques.

We recommend starting with the most essential skills in the beginning. They should be related to the role or business in general. Although, keep in mind that the more data you have, the better decisions you can make.

One good way to create a specific skill set for a role is to ask more experienced employees about it. For example, you can consult highly skilled employees or department heads and ask them to write a list of skills required for different levels: beginner, senior, intermediate, and advanced. This will help you save time and will highlight specific skills you might not have considered.

One way to automate this is to use a learning platform with such functionality. At Valamis, we created a competencies matrix with vast editorial capabilities.

Competencies Administration

Categories

Competences

Level of competence

Skills

Select categories to change

Soft Skills

Hard Skills

Hard Skills > Databases

Hard Skills > Programming

Add category

Fig. 3.10: A Competencies Matrix

It makes it easier to create and maintain multiple categories and subcategories, edit, and delete skills. Moreover, you can assign numbers to each category to prioritize them according to the needs.

However, you can use Google or Excel spreadsheets; they will also do the job.

2. Create a grading system

A grading system is used for ranking the competence or skill level of users.

When you have a list of skills, you need to define how you will grade them. For example, for the IT industry, Junior, Middle, or Senior could be levels of competence.

Consider that, for different businesses and industries, grading systems might be different, as well as for different roles and departments. You need to determine what works best for your organization.

For example, at Valamis, we are using the following grading system:

- **Awareness:** A person has a basic knowledge and understanding of the topic but hasn't yet applied the skill at work.
- **Novice:** A person has applied the skill, but might still need support.
- **Professional:** A person has experience of utilizing the skill and can work independently. Solves problems proactively. Can guide and support more junior people performing related tasks.

- **Expert:** A person has a lot of experience in applying the skill in practice. Can communicate with customers about the topic and can coach others. Follows the trends of the field and shares knowledge with colleagues. Might have an official certification in the skill.
- **Leading-edge Expert:** A person understands the bigger picture. Can discuss details with the client and advise the clients. Shares knowledge proactively, by, *e.g.*, writing blogs or presenting at conferences or in Dev days. Follows the field and keeps the skill up-to-date.

Depending on the company size and diversity of the business, you might consider having several grading systems for different areas or departments.

3. Evaluate your employees' skills

To do so, you can use several techniques; each has its own pros and cons. We tried to order them based on speed and accuracy.

- **Self-evaluation:** An employee grades each skill personally.
- **Manager Evaluation:** Manager grades employees' skills.
- **Team or Client Evaluation:** You can collect team or client feedback and use it to evaluate employees' skills.
- **Skills Assessment:** Employees need to complete the tests or quizzes to determine their skills level.
- **Certification:** A fast and accurate method to evaluate specific skills, but not all employees have it nor can all skills be certified.

For a fast result, we recommend that you start with self-evaluation and certificates. Offer your employees the opportunity to assess their own skills based on the grading system and add their certificates.

If you need a more accurate evaluation, use skill assessment, or several methods at once.

Tip: Add options for employees to select their favourite skills or skills that they would like to improve.

Q Search user or skill

My Skill Profile

[< Back](#)

Edit your skills

Save

	Awareness	Novice	Professional	Expert	Leading-edge expert	Favourite
14 Marketing						
14.5 Marketing						
Blog Writing				✓		♥
Content Marketing				✓		♥
Copywriting				✓		
Customer Case Writing		✓				
Editing			✓			
Infographics		✓				
SEO Copywriting				✓		
Storytelling		✓				
Web Content Writing			✓			
Webinar management	✓					
White Paper Writing	✓					
14.6 Social Media						
Facebook business manager (includes Instagram)		✓				
LinkedIn Advertising			✓			

Fig. 3.11: Skill Assessment

4. Visualize data and reveal insights

Even though this part can be optional, we strongly recommend visualizing the data.

This will help you discover quite insightful information right away.

Here are some ideas:

(a) Calculate the sum of skill weights

Skill weight is numeric value for each skill level (*e.g.*, novice – 1, professional – 2, expert – 3, leading-edge expert – 4).

When you have this parameter calculated for each employee, you can determine the most skilled people in the department or for the role, identify people who might need training or might be promoted, and gain many other insights.

On the screenshot, we filtered the software development team so you can straightaway determine different groups of employees: team leads most skilled group on the top right corner, medium and senior developers in the middle, and several junior developers in the left bottom corner.

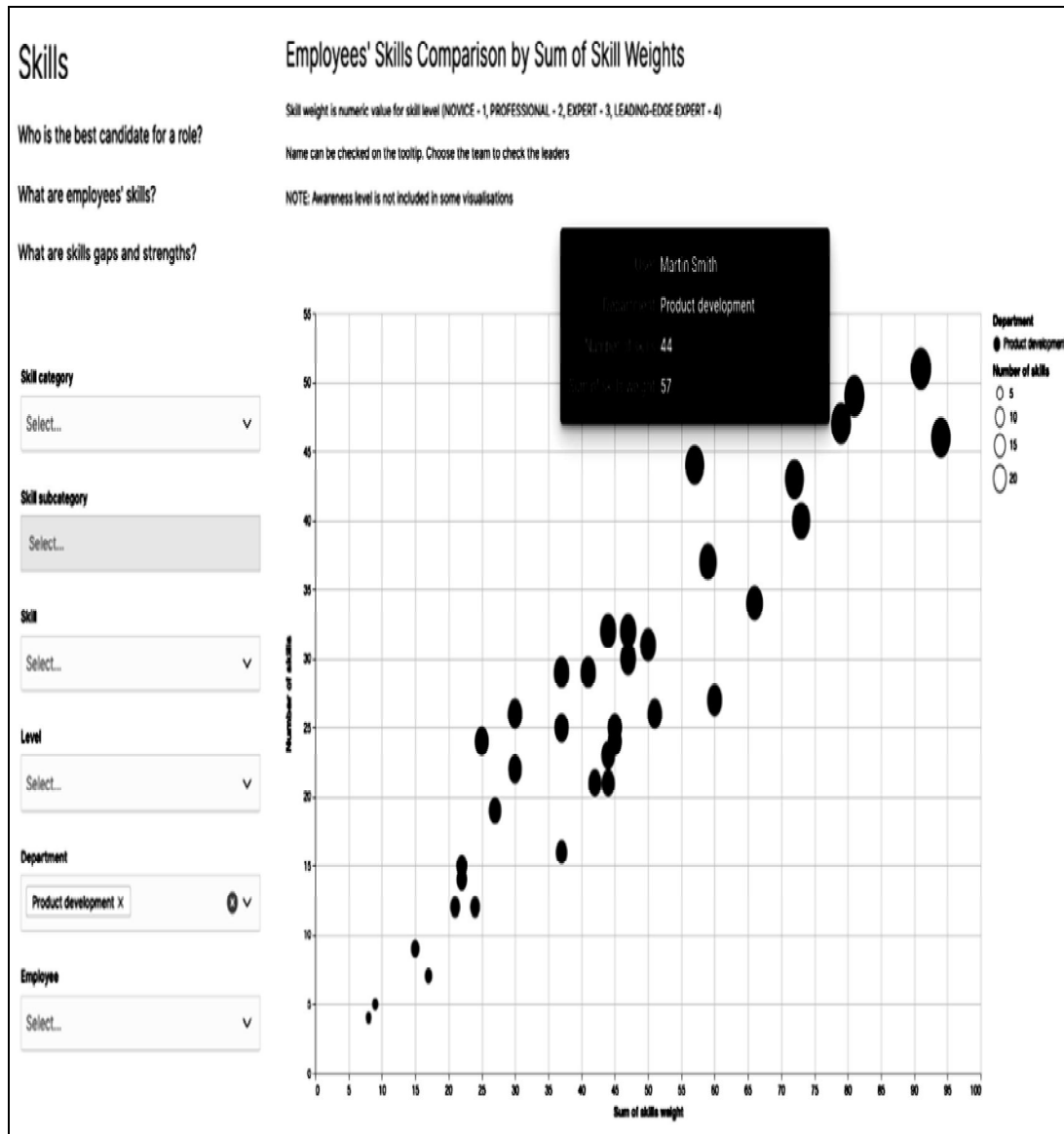


Fig. 3.12: Employees' Skill Comparison by Sum of Skill Weights

You can do the same for a specific skill or skill set to find the perfect candidate for a role and form a team.

(b) Define a potent skill loss threat

Such visualization is really valuable for determining skills that might be lost if employees leave the company. Usually, such a case happens when you have one or only a few employees with this skill. This is especially crucial if this skill is important for the department.

In this example, you can see that video and audio production skills, as well as media relations, are in a tight spot. Those skills are quite important for marketing. So, losing people with this knowledge will require some time to recover. It would, therefore, be good to train more employees in these areas to handle it and be prepared for any potential departures.

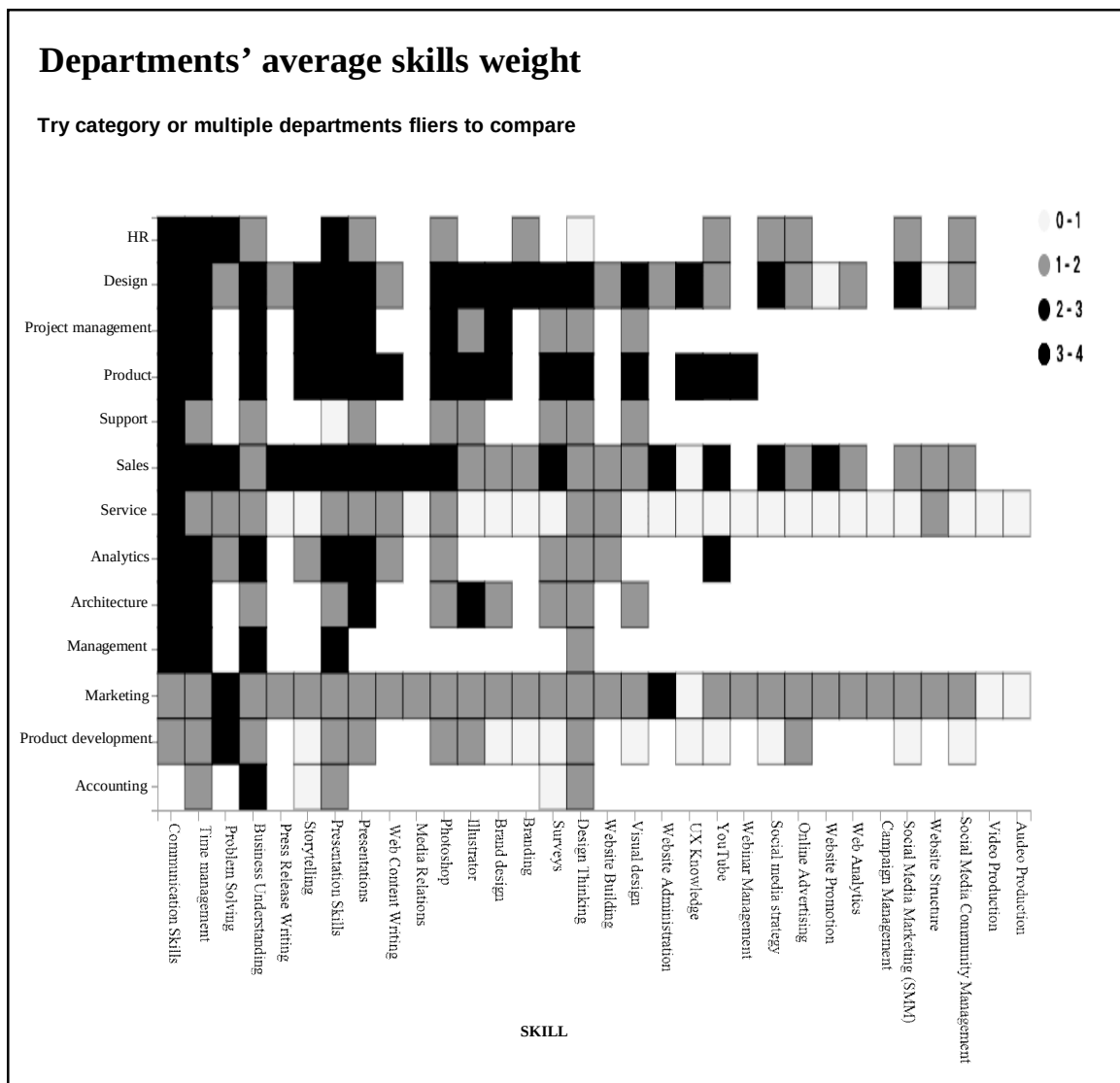


Fig. 3.13: Department's Average Skill Weight

Skills Matrix Template

We prepared a simple skills matrix template that can help you start building your own skill matrix. You can use it as it is, or build your own, based on the example. Make a copy and experiment with it.

		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R		
		Skills Matrix																			
Department					0	Awareness															
Total Number of Staff Members					1	Novice															
Total Number of skills / competences					2	Professional															
					3	Expert															
					4	Leading-edge expert															

Fig. 3.14: Skills Matrix Template

3.11 DEVELOPER TAB

What is the Developer Tab?

The Developer Tab, which is a built-in tab in Excel, provides the features needed to use Visual Basic for Applications (VBA) and perform a macro operation. The tab is disabled by default. It must be enabled first in the Options section on the File menu to make it visible on the toolbar at the top of the Excel window.

Importance of Developer Tab in Excel

Microsoft Excel is one of the most important Microsoft Office applications that are available to finance professionals. It helps in creating reports, analyzing large volumes of data, generating reports, etc. The spreadsheet application makes it easy for financial analysts, investment bankers, and portfolio managers to make decisions with ease.

Finance professionals can evaluate the financial performance of companies, simulate expected outcomes, and even compare investment scenarios. When the Developer Tab is activated in Excel, they carry out more advanced functions such as designing forms, writing code, creating macros, as well as importing and exporting XML file formats.

How to Add the Developer Tab to the Excel Ribbon?

The Developer Tab is usually hidden by default, and the user needs to activate it from the Options section. Follow the following steps to make the Developer Tab visible:

1. Click on the File menu, and scroll to the Options menu. Click Options from the drop-down menu to open the Excel Options window as follows:

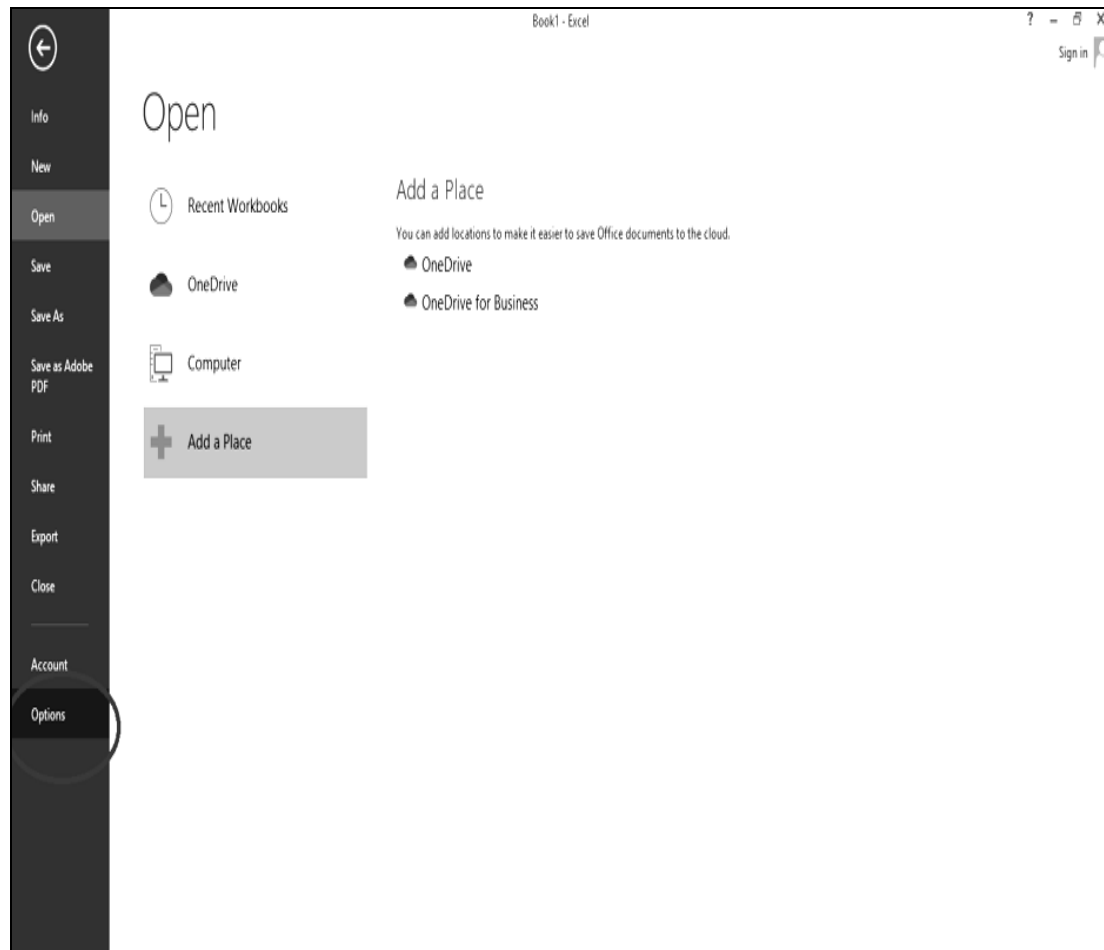


Fig. 3.15(a): How to Add the Developer Tab to the Excel Ribbon?

2. Click the Customize Ribbon option in the Excel Options window. Then, on the right side under the list of main tabs, click on the Developer checkbox to tick the box. Click the OK button.

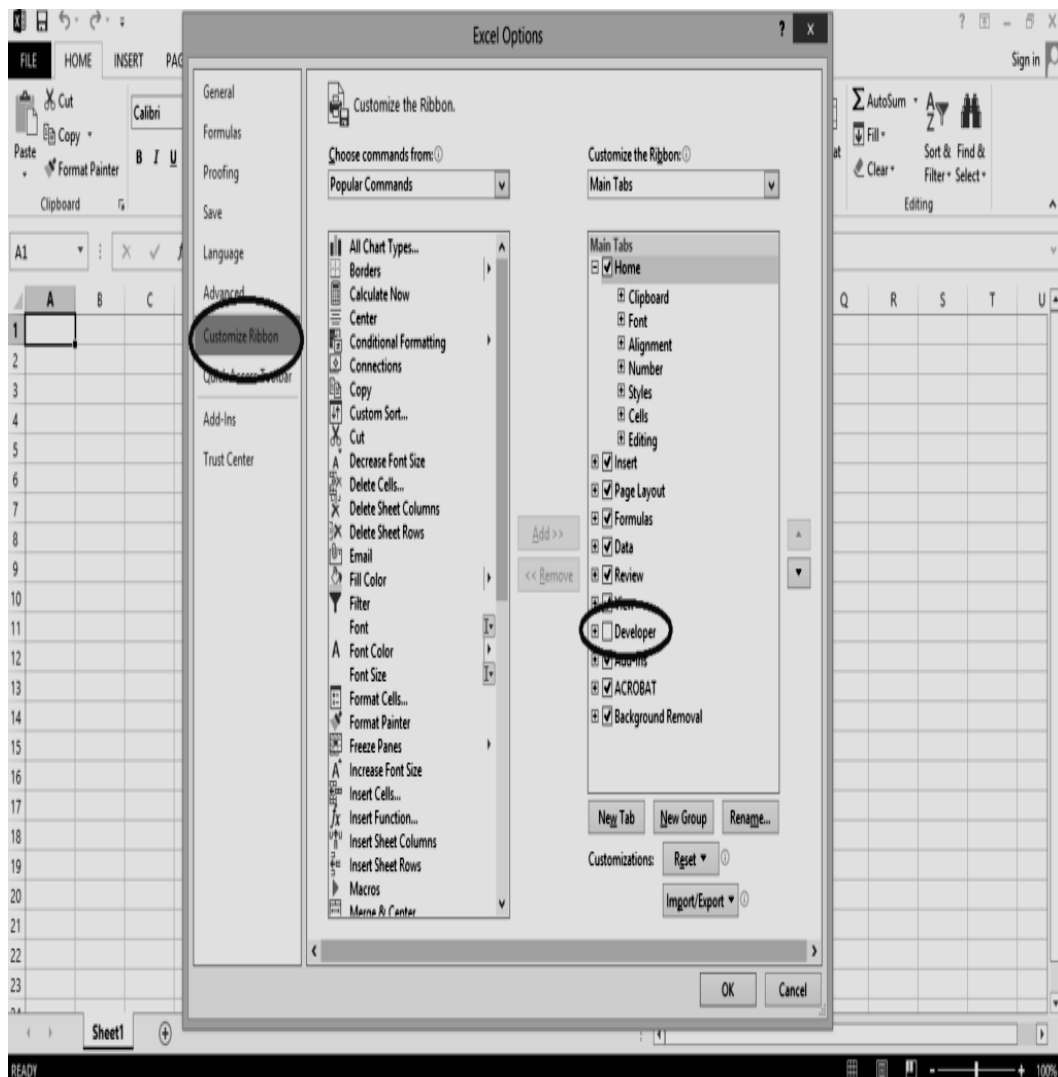


Fig. 3.15(b): How to Add the Developer Tab to the Excel Ribbon?

3. The Developer Tab will appear on the default tabs at the top of the Excel window. The tab contains a list of all features that you will need for advanced functions. It will remain visible unless you uncheck the Developer option or uninstall the Microsoft Office suite.

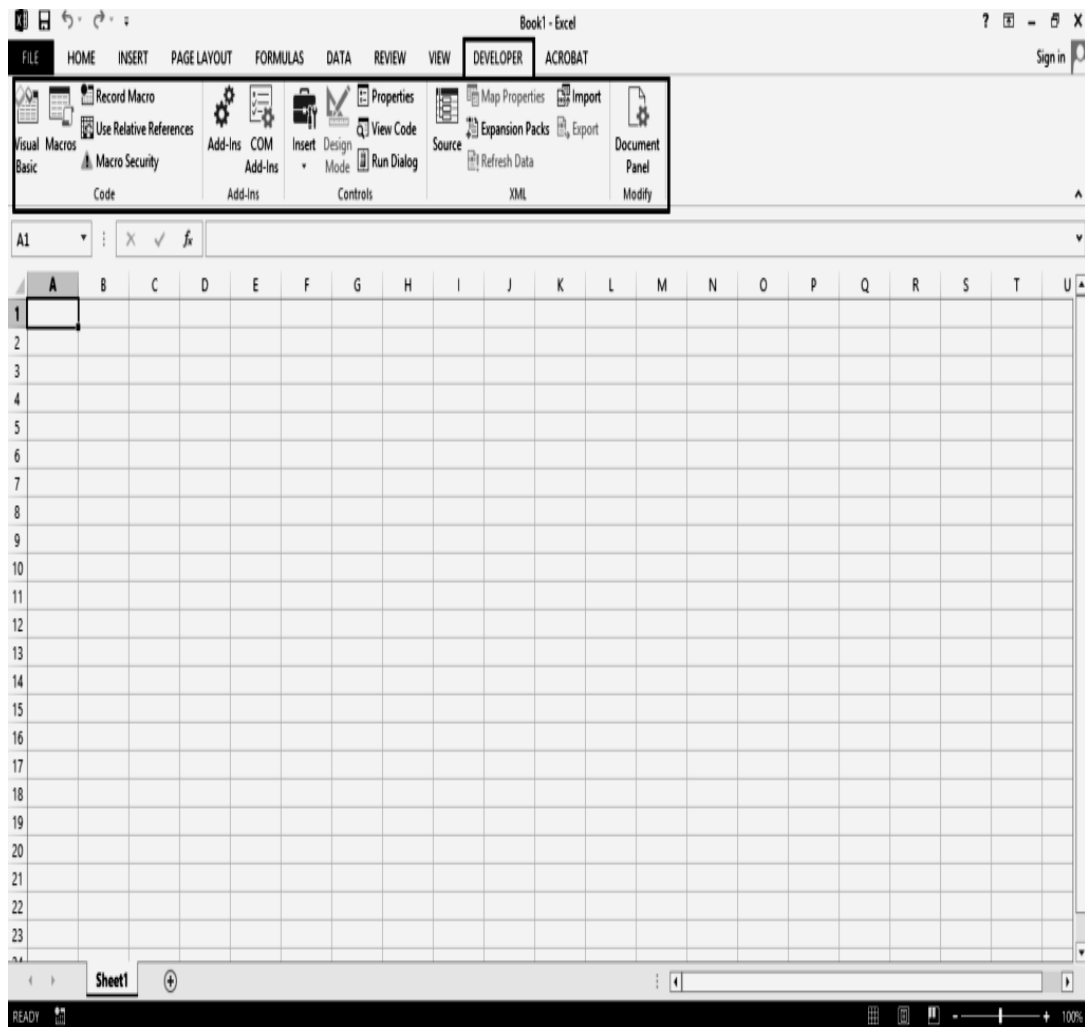


Fig. 3.15(c): How to Add the Developer Tab to the Excel Ribbon?

3.11.1 Options in the Developer Tab

The following are some of the useful options available in the Excel Developer Tab:

1. Macros

The Macros option is the code group in the Developer Tab, and it gives users access to a list of macros that they can work with. Users can record new macros, run macros, and even manage already recorded macros. A macro is a set of instructions that a user can record and get them repeated automatically once a specific action is carried out.

Once you click on the “Record Macro” option under the Developer Tab, Excel will automatically start recording the user’s actions in the program. Clicking the “Record Macro” again will stop recording the macro, and save it for later use.

Under the Macro options, users can also open the Visual Basic for Application (VBA) Editor to create, edit, and record macros that can perform frequent or repetitive tasks in MS Office applications. While users can still create macros in the traditional manner, writing a macro from the VBA Editor provides more flexibility, and users can complete more complicated tasks using VBA. The Windows shortcut for launching the VBA Editor is Alt + F11.

2. Add-ins

Add-ins are additional programs that users can install to get additional features, functionalities, and commands in Excel. When a user installs Excel for the first time, add-ins are not immediately available, and users need to install and activate them first.

Some of the Excel add-ins are built within the application, such as Analysis Toolpak and Solver, while others are developed by solution providers to perform specific functions. Users can also create an add-in by writing or recording the code in a module, saving it as an Excel add-in, and then putting the add-in to the Quick Access toolbar.

3. Interactive Controls

There are two types of controls in Excel worksheets that are used on worksheet forms. They include Form Controls and ActiveX Controls. Form Controls are simple to use. On the other hand, ActiveX is more advanced, and it provides a flexible design that users can use in VBA.

Users can insert the controls in a worksheet, edit the VBA code for control, and switch the design mode to on or off. Users can also use the controls to insert interactive tools such as radio buttons, checkboxes, and combo boxes.

4. XML Commands

The Developer Tab also includes the XML Source task panel that allows users to import and export XML data that is created from other applications or databases. The section also helps manage expansion packs that may be included in an active worksheet. Users can also exchange XML data between different applications.

3.12 INDEX

What is an index?

An index is a structure within SQL that is used to quickly locate specific rows within a table. It can be useful to imagine an index at the back of a textbook when thinking about SQL indexes. They both serve the same purpose – to find specific information quickly.

General Structure

An index is defined on one or more columns, called key columns. The key columns (also referred to as the index key) can be likened to the terms listed in a book index. They are the values that the index will be used to search for. As with the index found at the back of a text book (see Fig. 3.16), the index is sorted by the key columns.

If an index is created with more than one key column, it is known as a composite index.

The general structure of an index is that of a balanced tree (B-tree). The index will have a single root page, zero or more intermediate levels and then a leaf level. A page is an 8 kilobyte chunk of the data file, with a header and footer and is identified by a combination of File ID and Page number.

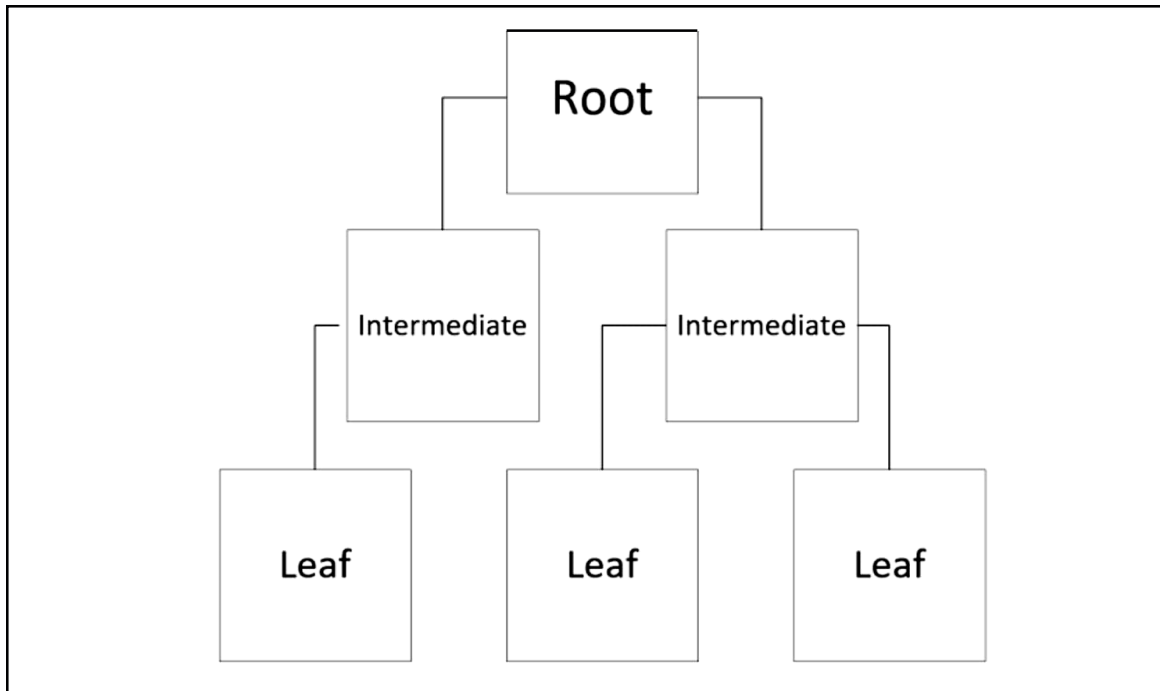


Fig. 3.16: Index Structure

Note: Commonly, the root page is shown at the top of the tree diagram and the leaf pages at the bottom. Think of it as an inverted tree.

In the leaf level, there is one entry for each row in the index. The entries in the index are ordered logically in the order of the index key.

The non-leaf levels of the index contain one row per page of the level below, referencing the lowest index key value on each page. If all of those rows fit onto a single page, then that page is considered the root and the index is only two levels deep. If all of those rows will not fit on a single page, then one (or more) intermediate levels are added to the index.

The number of levels in an index is referred to as the depth of the index. This is an important consideration for evaluating the efficiency of the index. The index illustrated in Fig. 3.7 has a depth of 3.

- With the exception of SQL 2008's filtered indexes, an index will have the same number of rows at the leaf level as the table.

- I'm using the phrase 'logically ordered' because the index does not necessarily define the physical storage of the rows. The rows are stored in a way that SQL can retrieve them ordered.

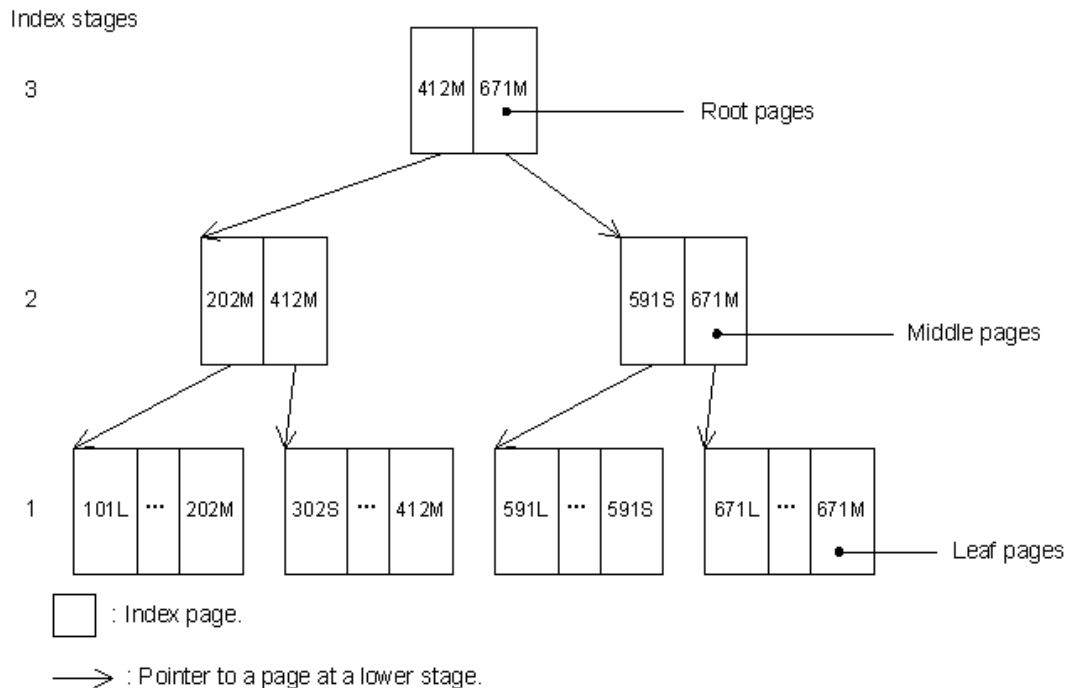


Fig. 3.17: B-tree structure of an index

Explanation

The index is created in Step 3 on the basis of the *product code* column, PCODE is the key, and 101L to 671M are key values.

Organization of this Subsection

- Single column index and multi-column index
- Optimizing based on cost
- Index definition

Clustered and Non-clustered

There are two main types of indexes in SQL Server, the clustered index and the non-clustered index.

Clustered indexes define the logical order of the table. The leaf level of the clustered index has the actual data pages of the table. Because of this, there can only be one clustered index per table. A table that does not have a clustered index is referred to as a heap.

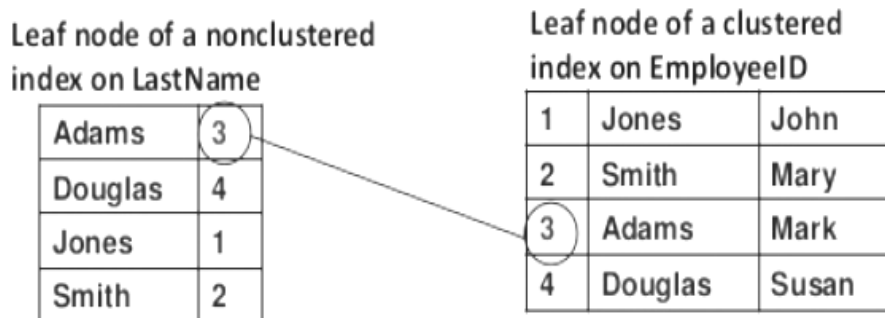


Fig. 3.18: Clustered and Non-Clustered Index

Non-clustered indexes are separate from the table. The leaf level of a non-clustered index has a pointer as part of each index row. That pointer is either the clustered index key in the cases where the base table has a clustered index or the RID (Row Identifier) in the cases where the table is a heap. The RID is an 8-byte structure comprised of File ID, Page Number and Slot Index, and will uniquely identify a row in the underlying heap. Either way, the each row of a non-clustered index has a reference to the complete data row.

Index Limits

There are a number of built-in limitations on indexes.

Key Size

The size of an index key is limited to a maximum of 900 bytes and a maximum of 16 columns. This is definitely a limit, not a goal, as the larger the index key gets, the more pages in the index and the deeper the index tree. As the number of pages and the depth of the tree increases so the index becomes less efficient to use. Larger indexes also use more storage space and result in less efficient use of SQL's data cache.

Number of Indexes

In SQL 2005 and earlier, there was a limitation of 250 indexes per table, one clustered and 249 non-clustered. In SQL 2008, with the addition of filtered indexes, that limitation was increased to 1000, one clustered and 999 non-clustered indexes.

Both of these limits are very high and there are few circumstances where a well-designed system should approach that limit.

The reason for this is two-fold.

1. As the number of indexes increases, the total size occupied by the table (with all of its indexes) also increases. Sure, hard drives are cheap and storage is abundant but increasing the size of a database has other effects. Maintenance operations (backups, restores, consistency checks and index rebuilds) all take longer as the size of a database increases.

2. Indexes have to be kept up-to-date as data changes and the more indexes there are on a table, the more places the data has to be changed. If there are 10 non-clustered indexes on a table, an insert must be done in 11 places (the table and each of those non-clustered indexes). On databases that are mostly read-only (decision support, data warehouses) may be acceptable. On databases that have frequent inserts, updates and deletes (OLTP systems), the overhead imposed by multiple indexes may not be acceptable.

How SQL Uses Indexes?

If a table does not have index, the only way to find all occurrences of a value within a table is to read the entire table. If a table has an index, it speeds up the locating of values within that index in two ways:

1. The index is sorted in the order of the key columns. This means that once all the matching values have been found, the remaining portion of the table can be ignored. This is the same as a telephone directory, where once all entries with a particular surname have been found, the rest of the book can be ignored as no further matches are possible.
2. The tree structure of the index allows a divide-and-conquer approach to locating rows, where large portions of the table can be quickly excluded from the search. This is illustrated in Figure 3.9.
3. SARGable is a made-up word, constructed from the phrase Search ARGument. It refers to a predicate that is of a form that SQL can use for an index seek.

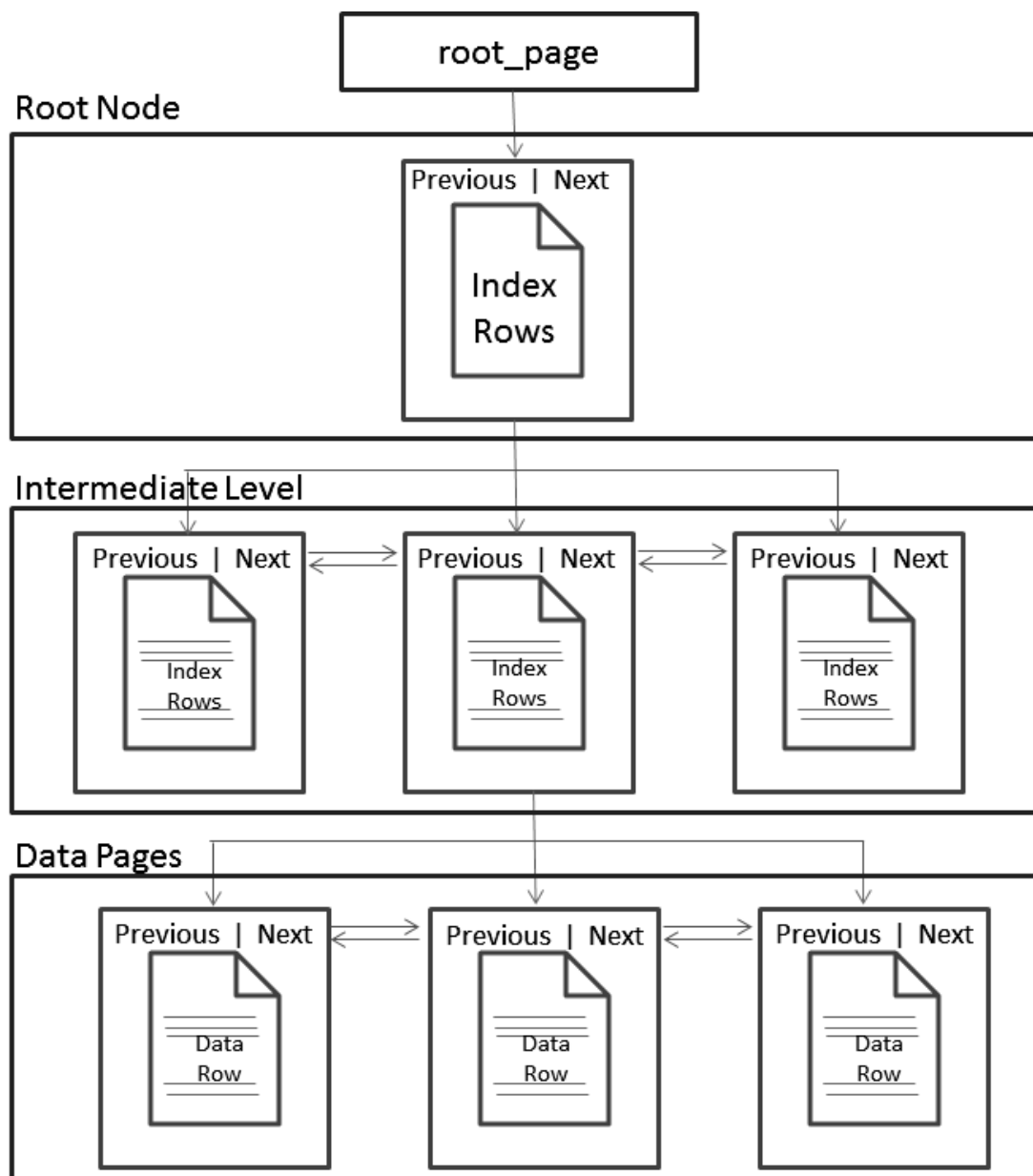


Fig. 3.19: Tree Structure of Index

There are four basic operations that SQL can do on an index. It can scan the index, it can seek on the index, it can do look-ups to the index and it can update the index.

Scans

An index scan is a complete read of all of the leaf pages in the index. When an index scan is done on the clustered index, it is a table scan in all but name.

When an index scan is done by the query processor, it is always a full read of all of the leaf pages in the index, regardless of whether all of the rows are returned. It is never a partial scan.

A scan does not only involve reading the leaf levels of the index, the higher level pages are also read as part of the index scan.

Seeks

An index seek is an operation where SQL uses the B-tree structure to locate either a specific value or the beginning of a range of value. For an index seek to be possible, there must be a SARGable predicate specified in the query and a matching (or partially matching) index. A matching index is one where the query predicate used a left-based subset of the index columns.

The seek operation is evaluated starting at the root page. Using the rows in the root page, the query processor will locate which page in the next lower level of the index contains the 1st row that is being searched for. It will then read that page. If that is the leaf level of the index, the seek ends there. If it is not the leaf, then the query processor again identifies which page in the next lower level contains the specified value. This process continues until the leaf level is reached.

Once the query processor has located the leaf page containing either the specified key value or the beginning of the specified range of key values, then it reads along the leaf pages until all rows that match the predicate have been returned.

If the index contains all the columns that the query needs, then the index is said to be covering for that query. If the index does not contain all the columns then SQL will do a lookup to the base table to fetch the other columns in order to process the query.

Lookups

Lookups occur when SQL uses an index to locate rows affected by a query but that index does not contain all the columns required to satisfy the query, that is the index is not covering for that query. To fetch the remaining columns, SQL does a lookup to either the clustered index or heap.

A lookup to a clustered index is essentially a single row clustered index seek and it is always a single row seek. So, if a lookup is needed for 500 rows, that involves 500 individual clustered index seeks.

Index Seek (NonClustered)	
Scan a particular range of rows from a nonclustered index.	
Physical Operation	Index Seek
Logical Operation	Index Seek
Actual Execution Mode	Row
Estimated Execution Mode	Row
Storage	RowStore
Number of Rows Read	2585
Actual Number of Rows	2585
Actual Number of Batches	0
Estimated Operator Cost	0.0273606 (100%)
Estimated I/O Cost	0.0234614
Estimated Subtree Cost	0.0273606
Estimated CPU Cost	0.0038993
Estimated Number of Executions	1
Number of Executions	1
Estimated Number of Rows	2545.7
Estimated Number of Rows to be Read	2545.7
Estimated Row Size	15 B
Actual Rebinds	0
Actual Rewinds	0
Partitioned	True
Actual Partition Count	7
Ordered	True
Node ID	0
Object	
[WideWorldImportersDW].[Fact].[Order].	
[IX_SQLShack_StockItemKey]	
Output List	
[WideWorldImportersDW].[Fact].[Order].Stock Item Key,	
[WideWorldImportersDW].[Fact].[Order].Salesperson Key	
Seek Predicates	
Seek Keys[1]: Start: PtnId1001 >= Scalar Operator((1)), End:	
PtnId1001 <= Scalar Operator((7)), Seek Keys[2]: Prefix:	
[WideWorldImportersDW].[Fact].[Order].Salesperson Key =	
Scalar Operator((83))	

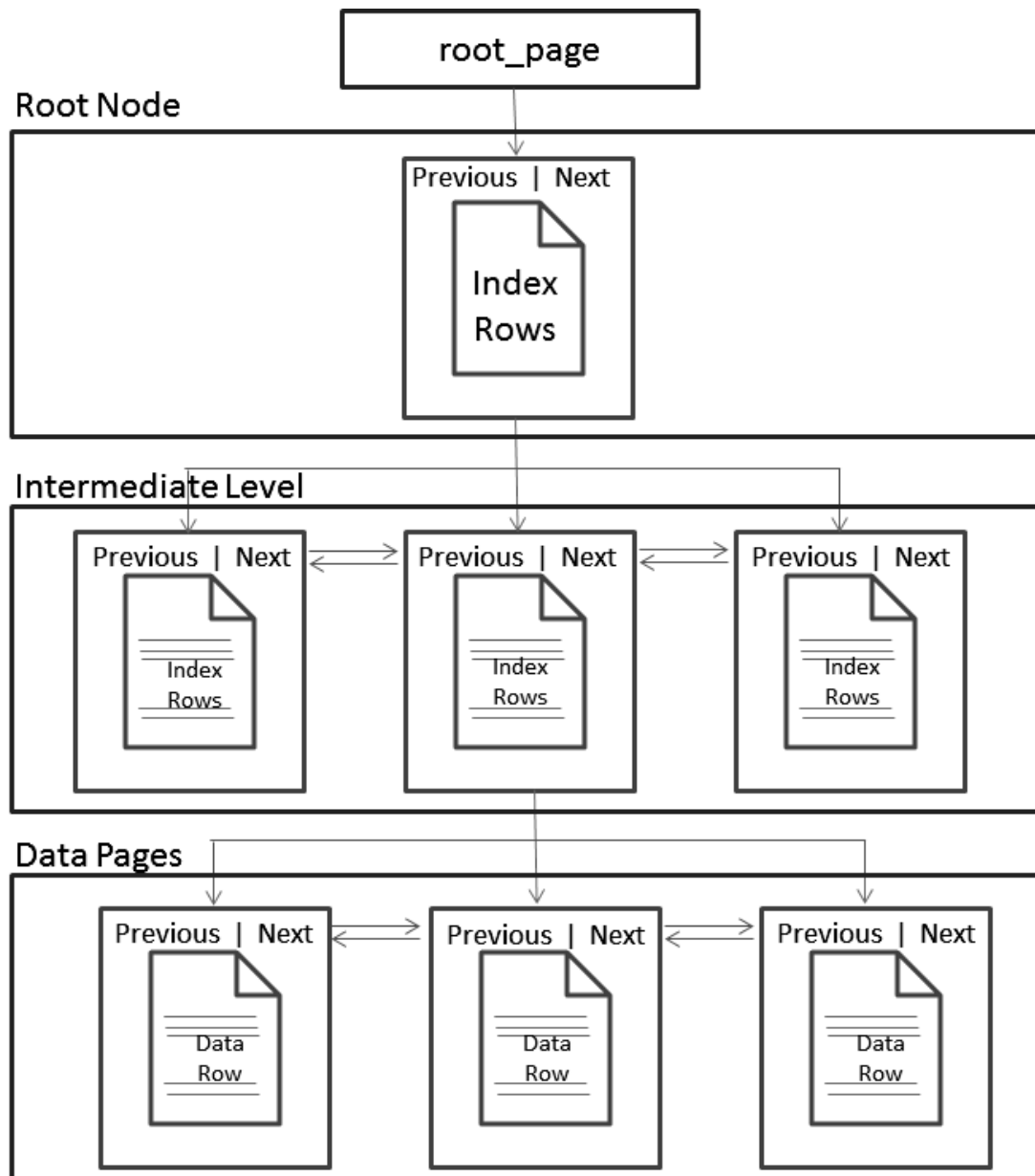


Fig. 3.20: Index Seek (Non-Clustered)

Updates

Anytime that a row is changed, those changes must be made not only in the base table (clustered index or heap) but also in any index that contains the columns that were affected by the change. This applies to insert, update and delete operations.

Considerations for Creating Indexes

More detail on considerations for indexes in the next two parts, but in general, are as follows:

- Clustered index should be narrow because the clustering key is part of all non-clustered indexes.
- Composite non-clustered indexes are generally more useful than single column indexes, unless all queries against the table filter on one column at a time.
- Indexes should be no wider than they have to be. Too many columns wastes space and increase the amount of places that data must be changed when an insert/update/delete occurs.
- If an index is unique, specify that it is unique. The optimizer can sometimes use that information to generate more optimal execution plans.
- Be careful of creating lots of indexes on frequently modified tables as it can slow down data modifications.

3.13 FORM

What is a Form?

A form, whether printed or online, is a document designed with a standard structure and format that makes it easier to capture, organize, and edit information.

- Printed forms contain instructions, formatting, labels, and blank spaces for writing or typing data. You can use Excel and Excel templates to create printed forms.
- Online forms contain the same features as printed forms. In addition, online forms contain controls. Controls are objects that display data or make it easier for users to enter or edit data, perform an action, or make a selection. In general, controls make the form easier to use. Examples of common controls include list boxes, option buttons, and command buttons. Controls can also run assigned macros and respond to events, such as mouse clicks, by running Visual Basic for Applications (VBA) code.

You use Excel in several ways to create printed and online forms.

Types of Excel Forms

There are several types of forms that you can create in Excel: data forms, worksheets that contain Form and Active X controls, and VBA User Forms. You can use each type of form by itself, or you can combine them in different ways to create a solution that is right for you.

1. Data Form

A data form provides a convenient way to enter or display one complete row of information in a range or table without scrolling horizontally. You may find that using a data form can make data entry easier than moving from column to column when you have more columns of data than can be viewed on the screen. Use a data form when a simple form of text boxes that list the

column headings as labels is sufficient and you do not need sophisticated or custom form features, such as a list box or spin button.

The screenshot shows a dialog box titled "Customers" with a close button (X) in the top right corner. It contains several text input fields with labels to their left: "Customer ID:" (value: LINOD), "Company Name:" (value: LINO-Delicatesses), "Contact Name:" (value: Felipe Izquierdo), "Contact Title:" (value: Owner), "Address:" (value: Ave. 5 de Mayo Porlamar), "City:" (value: I. de Margarita), "Region:" (value: Nueva Esparta), "Postal Code:" (value: 4980), "Country:" (value: Venezuela), "Phone:" (value: (8) 34-56-12), and "Fax:" (value: (8) 34-93-93). To the right of these fields is a vertical scrollbar. Further to the right are several buttons: "New", "Delete", "Restore", "Find Prev", "Find Next", "Criteria", and "Close". In the top right corner of the dialog, it says "47 of 91".

Fig. 3.21: Data Form

Excel can automatically generate a built-in data form for your range or table. The data form displays all column headers as labels in a single dialog box. Each label has an adjacent blank text box in which you can enter data for each column, up to a maximum of 32 columns. In a data form, you can enter new rows, find rows by navigating or (based on cell contents), update rows and delete rows. If a cell contains a formula, the formula result is displayed in the data form, but you cannot change the formula by using the data form.

2. Worksheet with Form and ActiveX Controls

A worksheet is a type of form that enables you to enter and view data on the grid, and there are several control-like features already built-in to Excel worksheets, such as comments and data validation. Cells resemble text boxes in that you can enter and format them in a variety of ways. Cells are often used as labels, and by adjusting cell height and width and merging cells, you can make a worksheet behave like a simple data entry form. Other control-like features, such as cell comments, hyperlinks, background images, data validation, conditional formatting, embedded charts, and AutoFilter can make a worksheet behave like an advanced form.

For added flexibility, you can add controls and other drawing objects to the drawing canvas of a worksheet, and combine and coordinate them with worksheet cells. For example, you use a list box control to make it easier for a user to select from a list of items, or you use a spin button control to make it easier for a user to enter a number.

Because controls and objects are stored on the drawing canvas, you can display or view controls and objects alongside associated text that is independent of row and column boundaries without changing the layout of a grid or table of data on your worksheet. Most of the time, many of these controls can also be linked to cells on the worksheet and do not require VBA code to make them work. You can set properties that determine whether a control floats freely or moves

and resizes together with a cell. For example, you might have a check box that you want to move together with its underlying cell when the range is sorted. However, if you have a list box that you want to keep in a specific location at all times, you probably do not want it to move together with its underlying cell.

Excel has two types of controls: Form Controls and ActiveX Controls. In addition to these sets of controls, you can also add objects from the Drawing tools, such as a AutoShapes, WordArt, SmartArt graphic, or text boxes.

The following sections describe these controls and drawing objects, and also explain how to work with these controls and objects in more detail.

Form Controls

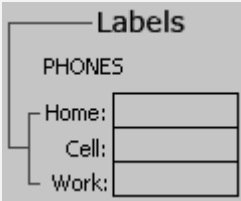
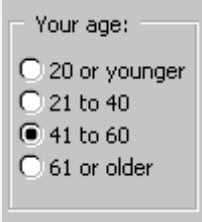
Form controls are the original controls that are compatible with earlier versions of Excel, starting with Excel version 5.0. Form controls are also designed for use on XLM macro sheets.

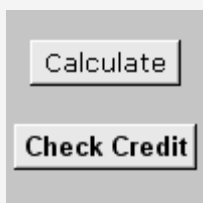
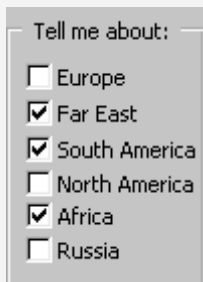
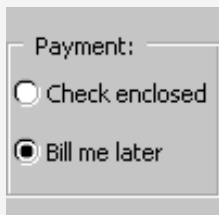
You can use Form controls when you want to easily reference and interact with cell data without using VBA code, and when you want to add controls to chart sheets. For example, after you add a list box control to a worksheet and linking it to a cell, you can return a numeric value for the current position of the selected item in the control. You can then use that numeric value in conjunction with the **INDEX** function to select different items from the list.

You can also run macros by using Form controls. You can attach an existing macro to a control, or write or record a new macro. When a user of the form clicks the control, the control runs the macro.

However, these controls cannot be added to UserForms, used to control events, or modified to run Web scripts on Web pages.

Summary of Form Controls

Button Name	Example	Description
	Label	 Identifies the purpose of a cell or text box, or displays descriptive text (such as titles, captions and pictures) or brief instructions.
	Group box	 Groups related controls into one visual unit in a rectangle with an optional label. Typically, option buttons, check boxes, or closely related contents are grouped.

Button Name	Example	Description	
	Button		Runs a macro that performs an action when a user clicks it. A button is also referred to as a push button.
	Check box		Turns on or off a value that indicates an opposite and unambiguous choice. you can select more than one check box on a worksheet or in a group box. A check box can have one of three states: selected (turned on), cleared (turned off), and mixed, meaning a combination of on and off states (as in a multiple selection).
	Option button		Allows a single choice within a limited set of mutually exclusive choices; an option button is usually contained in a group box or a frame. An option button can have one of three states: selected (turned on), cleared (turned off), and mixed, meaning a combination of on and off states (as in a multiple selection). An option button is also referred to as a radio button.
	List box		Displays a list of one or more items of text from which a user can choose. Use a list box for displaying large numbers of choices that vary in number or content. There are three types of list boxes: • A single-selection list box enables only one choice. In this case, a list box resembles a group of option buttons, except that a list box can handle a large number of items more efficiently. • A multiple-selection list box enables either one choice or contiguous (adjacent) choices. • An extended-selection list box enables one choice, contiguous choices, and non-contiguous (or disjointed) choices.

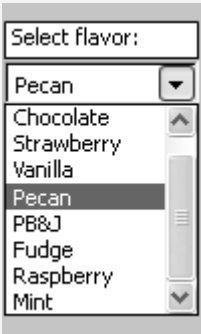
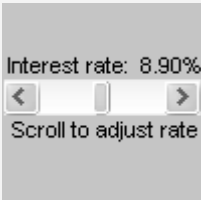
Button Name	Example	Description	
	Combo box		Combines a text box with a list box to create a drop-down list box. A combo box is more compact than a list box but requires the user to click the down arrow to display the list of items. Use a combo box to enable a user to either type an entry or choose only one item from the list. The control displays the current value in the text box, regardless of how that value is entered.
	Scroll bar		Scrolls through a range of values when to click the scroll arrows or drag the scroll box. In addition, you can move through a page (a preset interval) of values by clicking the area between the scroll box and either of the scroll arrows. Typically, a user can also type a text value directly into an associated cell or text box.
	Spin button		Increases or decreases a value, such as a number increment, time, or date. To increase the value, click the up arrow; to decrease the value, click the down arrow. Typically, a user can also type a text value directly into an associated cell or text box.

Fig. 3.22: Summary of Form Controls

Note: The following controls are unavailable in Office Excel 2007 workbooks. These controls can be used in Excel version 5.0 dialog sheets only.

Button Name
 Text Field
 Combo List – Edit
 Combo Drop – Down Edit
 Run Dialog

ActiveX Controls

ActiveX controls can be used on worksheet forms, with or without the use of VBA code, and on VBA UserForms. In general, use ActiveX controls when you need more flexible design requirements than those provided by Form controls. ActiveX controls have extensive properties that you can use to customize their appearance, behaviour, fonts, and other characteristics.

You can also control different events that occur when an ActiveX control is interacted with. For example, you can perform different actions, depending on which choice a user selects from a list box control, or you can query a database to refill a combo box with items when a user clicks a button. You can also write macros that respond to events associated with ActiveX controls. When a user of the form interacts with the control, your VBA code then runs to process any events that occur for that control.

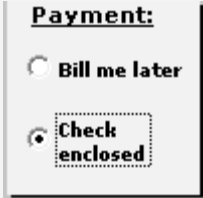
Your computer also contains many ActiveX controls that were installed by Excel and other programs, such as Calendar Control 12.0 and Windows Media Player.

Important: Not all ActiveX controls can be used directly on worksheets; some can be used only on Visual Basic for Applications (VBA) UserForms. If you try to add any one of these particular ActiveX controls to a worksheet, Excel displays the message “Cannot insert object.”

However, ActiveX controls cannot be added to chart sheets from the user interface or to XLM macro sheets. You also cannot assign a macro to run directly from an ActiveX control the same way you can from a Form control.

Summary of ActiveX Controls

Button Name	Example	Description
	Check box	 Turns on or off a value that indicates an opposite and unambiguous choice. You can select more than one check box at a time on a worksheet or in a group box. A check box can have one of three states: selected (turned on), cleared (turned off), and mixed, meaning a combination of on and off states (as in a multiple selection).
	Text box	 Enables you to, in a rectangular box, view, type, or edit text or data that is bound to a cell. A text box can also be a static text field that presents read-only information.
	Command button	 Runs a macro that performs an action when a user clicks it. A command button is also referred to as a push button.

Button Name	Example	Description
	Option button	 Allows a single choice within a limited set of mutually exclusive choices usually contained in a group box or frame. An option button can have one of three states: selected (turned on), cleared (turned off), and mixed, meaning a combination of on and off states (as in a multiple selection). An option button is also referred to as a radio button.
	List box	 Displays a list of one or more items of text from which a user can choose. Use a list box for displaying large numbers of choices that vary in number or content. There are three types of list boxes: • A single-selection list box enables only one choice. In this case, a list box resembles a group of option buttons, except that a list box can handle a large number of items more efficiently. • A multiple selection list box enables either one choice or contiguous (adjacent) choices. • An extended-selection list box enables one choice, contiguous choices, and non-contiguous (or disjointed) choices.
	Combo box	 Combines a text box with a list box to create a drop-down list box. A combo box is more compact than a list box, but requires the user to click the down arrow to display the list of items. It is used to allow a user to either type an entry or choose only one item from the list. The control displays the current value in the text box, regardless of how that value is entered.
	Toggle button	 Indicates a state, such as Yes/No, or a mode, such as On/Off. The button alternates between an enabled and disabled state when it is clicked.

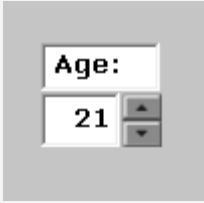
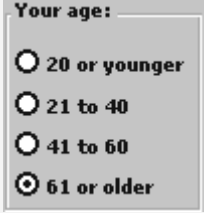
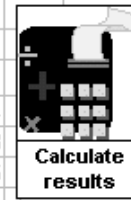
Button Name	Example	Description
	Spin button 	Increases or decreases a value, such as a number increment, time, or date. To increase the value, click the up arrow; to decrease the value, click the down arrow. Typically, a user can also type a text value into an associated cell or text box.
	Scroll bar 	Scrolls through a range of values when you click the scroll arrows or drag the scroll box. In addition, you can move through a page (a preset interval) of values by clicking the area between the scroll box and either of the scroll arrows. Typically, a user can also type a text value directly into an associated cell or text box.
	Label 	Identifies the purpose of a cell or text box, displays descriptive text (such as titles, captions, pictures), or provides brief instructions.
	Image 	Embeds a picture, such as a bitmap, JPEG, or GIF.
	Frame control 	A rectangular object with an optional label that groups related controls into one visual unit. Typically, option buttons, check boxes, or closely related contents are grouped in a frame control. Note: The ActiveX frame control is not available in the ActiveX Controls section of the Insert command. However, you can add the control from the More Controls dialog box by selecting Microsoft Forms 2.0 Frame .
	More Controls	Displays a list of additional ActiveX controls available on your computer that you can add to a custom form, such as Calendar Control 12.0 and Windows Media Player. You can also register a custom control in this dialog box.

Fig. 3.23: Summary of ActiveX Controls

Drawing Tool Objects

Blue Sky Airlines Stock History		
Click here to go to current stock quote Web page.		
Date	High	Low
3-Apr	56 3/8	55 1/4
10-Apr	56	54 1/8
17-Apr	56 3/8	56
24-Apr	55 7/8	54 3/4



You may also want to include SmartArt graphics, Shapes, WordArt, and text boxes on your form. You can resize, rotate, flip, colour, and combine these objects to create even more complex shapes. When you type text directly in a Shape or text box object, the text becomes part of that object — if you rotate or flip the object, the text rotates or flips with it. Unlike ActiveX controls, you can assign different attributes, such as font size and font style, to individual words and characters in the object. You can assign macros and add hyperlinks to these objects. You can even link text in a Shape or text box object to a worksheet cell and dynamically display updated values in those objects.

Working with Controls and Objects on the Worksheet Form

After adding forms and ActiveX to a worksheet form, you usually want to fine-tune and rearrange the controls in a variety of ways to create a well-designed, user-friendly form. Common tasks include the following:

- Controlling the display of gridlines while you work with the controls, and deciding whether to display the gridlines to the user on the final worksheet form.
- Selecting and deselecting controls so that you can specify properties or make additional adjustments.
- Editing text in a control, such as the caption or label.
- Grouping, copying, moving, and aligning controls to organize the layout of the worksheet form.
- Resizing and formatting controls to obtain the appearance that you want.
- Positioning or sizing a control with a cell.
- Protecting controls and linked cells according to your specific data protection needs.
- Enabling or disabling the printing of controls when the worksheet form is printed.
- Deleting unused controls.

You can design a worksheet form with or without cell gridlines in the background. For example, you might want to turn off cell gridlines and then format all the cells with the same colour or pattern, or even use a picture as a sheet background. To hide or show the gridlines, on the **View** tab, in the **Show/Hide** group, clear or select the **Gridlines** check box.

Determining the Type of Control That is on Your Worksheet

Because there are three different types of controls and objects that you can modify uniquely, you might not know for sure which type of control it is just by looking at it. To determine the type of control (Form or ActiveX), select and right-click the control, and then display the shortcut menu:

- If the shortcut menu contains the command **Properties**, the control is an ActiveX control, and you are in design mode.
- If the shortcut menu contains the command **Assign Macro**, the control is a Form control.

Tip: To display the correct shortcut menu for the group box Form control, make sure that you select the perimeter instead of the interior of the group box.

- If the shortcut menu contains the command **Edit Text**, the object is a Drawing object.

3. VBA UserForms

For maximum flexibility, you can create UserForms, which are custom dialog boxes, that usually include one or more ActiveX controls. You make UserForms available from VBA code that you create in the Visual Basic Editor. The high-level steps for creating a UserForm are as follows:

- Insert a UserForm into your workbook's VBAProject. You access a workbook's VBAProject by first displaying the Visual Basic Editor (press ALT + F11) and then, on the **Insert** menu, clicking **UserForm**.
- Write a procedure to display the UserForm.
- Add ActiveX controls.
- Modify properties for the ActiveX controls.
- Write event-handler procedures for the ActiveX controls.

By using UserForms, you can also utilize advanced form functionality. For example, you can programmatically add a separate option button for each letter of the alphabet or you can add a check box for each item in a large list of dates and numbers.

Before creating a UserForm, consider using built-in dialog boxes available from Excel that might fit your needs. These built-in dialog boxes include the VBA **InputBox** and **MsgBox** functions, the Excel **InputBox** method, **GetOpenFilename** method, **GetSaveAsFilename** method, and the Dialogs object of the **Application** object, which contains all the built-in Excel dialog boxes.

How to Write Excel SUMIFS Function?

Using SUMIFS, you can find the sum of values in the data that meet multiple conditions.

So, to get the sum of all the **Blow Torches** sold in **North**, we just write:

=SUMIFS(D3:D16, B3:B16,"Blow Torch",C3:C16,"North")

Similarly, to find the podgun sales in East, just write:

Product	Region	Sales
Pod Gun	North	127
Blow Torch	West	139
Blow Torch	West	103
Spit Bomb	North	168
Blow Torch	North	122
Spit Bomb	West	200
Spit Bomb	South	177
Pod Gun	South	170
Blow Torch	South	177
Blow Torch	North	170
Spit Bomb	West	144
Pod Gun	East	172
Pod Gun	East	167
Spit Bomb	East	100

SUMIFS formula makes it really easy to sum on multiple conditions

What is total sales of "Pod Gun" in "East"?

=SUMIFS(B14:B27,"Pod Gun","East")

(c) chandoo.org

SUMIFS Function – Syntax and Explanation

SUMIFS formula takes a range for summing the values and at least one criteria range and criteria. You can specify as many as 127 conditions for summing your data.

SUMIFS Formula Syntax

SUMIFS(sum_range, criteria_range1, criteria1, criteria_range2, criteria2, ...)

Mandatory Parameters Optional Criteria

Imagine asking “How many spit bombs Hansolo sold in North region of Planet Naboo between long long ago and long ago that resulted in more than 25% profits?” and getting an instant answer.

The beauty of SUMIFS formula is that it works with wildcards too, just like its siblings – SUMIF and COUNTIF. So, formulas can be written as:

=SUMIFS(D3:D16,B3:B16,"Spit Bomb",C3:C16,"*th") to get sum of spit bombs sold in North and South.

Using SUMIFS() with Tables

You can write SUMIFS function on either a range of data or on a table. When using with tables, you can simply apply **structural references** – i.e., Table Name [Column Name] notation to specify the *criteria* columns. See this example:

Sales for Blow Torch in West

\$242

=SUMIFS(acme[Sales], acme[Product], "Blow Torch", acme[Region], "West")

Product	Region	Sales
Pod Gun	North	127
Blow Torch	West	139
Blow Torch	West	103
Spit Bomb	North	168
Blow Torch	North	122
Spit Bomb	West	200
Spit Bomb	South	177
Pod Gun	South	170
Blow Torch	South	177
Blow Torch	North	170
Spit Bomb	West	144
Pod Gun	East	172
Pod Gun	East	167
Spit Bomb	East	100

Table named ACME

SUMIFS Examples

Let's say you have a table named ACME as pictured above. See these examples to understand how the function works.

- **Sales for Blow Torch in West**
 - =SUMIFS(acme[Sales], acme[Product], "Blow Torch", acme[Region], "West")
- **Total Sales above 150 in East**
 - =SUMIFS(acme[Sales], acme[Sales], ">150", acme[Region], "East")
- **Sales of North for all excluding Pod Gun**
 - =SUMIFS(acme[Sales], acme[Region], "North", acme[Product], "<>Pod Gun")
- **Sales of all products that contain letter B**
 - =SUMIFS(acme[Sales], acme[Product], "*B*")

Using SUMIFS() with Date and Time Values

When you have a column of dates, you can apply special operators like >, <, =, <> to specify a date range.

For example, to count total sales between March 2018 and May 2018, we can use:

```
=SUMIFS(acme[Sales], acme[Sales Date], ">=1-Mar-2018", acme[Sales Date], "<=31-May-2018")
```

You can either type the date in the formula or bring it from a cell. If you have two cells containing start and end date for your window of dates, you can use this formula:

```
=SUMIFS(acme[Sales], acme[Sales Date], ">=" & start_date_cell, acme[Sales Date], "<=" & end_date_cell)
```

Replace *start_date_cell* and *end_date_cell* with actual cell references or names.

Bonus

Just like SUMIFS, there is COUNTIFS and AVERAGEIFS too in Excel. Once you know SUMIFS(), you can use all these other functions with ease.

What is the SUMIFS Function in Excel?

We all know the SUMIF function allows us to sum the data given based on associated criteria within the same data. However, the SUMIFS Function in Excel allows applying multiple criteria.

Formula Used for the SUMIFS Function in Excel

```
"SUMIFS (sum_range, criteria_range1, criteria1, [criteria_range2, criteria2, criteria_range3, criteria3, ... criteria_range_n, criteria_n] )
```

where,

Sum_range = Cells to add.

Criteria_range1 = Range of cells that we want to apply criteria1 against.

Criteria1 = Used to determine which cells to add. Criteria1 is applied against criteria_range1.

Criteria_range2, criteria2, ... = The additional ranges along with their associated criteria.

The SUMIFS Function in Excel allows us to enter up to 127 range/criteria pairs for this formula. Remember:

- SUMIFS will return a numeric value.
- Rows and columns should be the same in the criteria_range argument and the sum_range argument.

Let's take an example to understand it. Suppose we use data about flowers and their cost per dozen for different regions.

Similarly, if I wish to find out the total cost of Daffodils for the North region, I can use the same formula:

The SUMIFS function can use comparison operators like '=', '>', '<'. If we wish to use these operators, we can apply them to an actual sum range or any of the criteria ranges. Also, we can create comparison operators using them:

- '<=' (less than or equal to)
- '>=' (greater than or equal to)
- '<>' (less than or greater than/not equal to)

Let's take an example to understand this in detail.

Using the sales figures per region of different salespersons, we wish to find out the:

- Sales of North region greater than 100
- Sales greater than 500

Use of Wildcards

Wildcard characters such as '*' and '?' can be used within the criteria argument when using the SUMIFS function. Using these wildcards will help us in finding matches that are a similar but not accurate match.

1. **Asterisk (*):** It matches any sequence of characters. It can be used after, before or surrounding criteria to allow partial search criteria to be used.

For example, if we apply the following criteria in SUMIFS function:

- N* – It implies all cells in the range that start with N.
 - *N – It implies all cells in the range that ends with N.
 - *N* – It implies cells that contain N.
2. **Question Mark (?):** It matches any single character. Suppose I apply N?r as the criteria. Here "?" will take the place of a single character. N?r will match with North, Nor, etc. However, it will not take into consideration Name.

What if the given data contain an asterisk or an actual question mark?

In this case, we can use "tilde (~)". We need to type "~" in front of the question mark in that scenario.

Using Named Ranges with SUMIFS Function

Named range is the descriptive name of a collection of cells or range in a worksheet. We can use named ranges while using the SUMIFS function.

SUMIF vs. SUMIFS

The SUMIF function can be only used for adding a single continuous range based on a single specified range with a single criterion, whereas, SUMIFS can be applied over multiple continuous ranges.

- When using SUMIF, we can evaluate only one condition, whereas different criteria can be evaluated under SUMIFS formula. This is the primary difference between the two Excel functions.
- SUMIFS is available from MS Excel 2007.

- The SUMIF function can be only used for adding a single continuous range based on a single specified range with a single criterion, whereas, SUMIFS can be applied over multiple continuous ranges.

AVERAGEIF

What is the AVERAGEIF Function?

The AVERAGEIF Function is an Excel statistical function, which calculates the average of a given range of cells by a specific criterion. Basically, AVERAGEIF calculates central tendency, which is the location of the center of a group of numbers in a statistical distribution. The function was introduced in Excel 2007.

Formula:

=AVERAGEIF (range, criteria, [average_range])

The AVERAGEIF function uses the following arguments:

- **Range (required argument):** This is the range of one or more cells that we want to average. The argument may include numbers or names, arrays, or references that contain numbers.
- **Criteria (required argument):** Criteria determines how the cell will be averaged. It can be in the form of an expression, number, cell reference, or text that defines which cells are averaged. For example, 12, "<12", "Baby" or C2.
- **Average_range (optional argument):** This is the actual set of cells that we wish to average. If a user omits it, the function will use the range given.

How to Use the AVERAGEIF Function in Excel?

This is a built-in function that can be used as a worksheet function in Excel. To understand the uses of the AVERAGEIF function, let's consider a few examples:

Step by Step Example of AVERAGEIF

Generally, AVERAGEIF is commonly used in finding the average of cells that are an exact match of a given criterion. For example, from the table below, we wish to find out the sales of chocolate pastries. We are given the following data:

We get the result below:

C20								
	A	B	C	D	E	F	G	H
2		AVERAGEIF Function						
3								
4		Item	Sales					
5		Chocolate pastries	1,000					
6		Pineapple pastries	500					
7		Lemon Tarts	650					
8		Pancakes	600					
9		Pineapple pastries	350					
10		Lemon Tarts	450					
11		Pancakes	350					
12		Chocolate pastries	250					
13		Chocolate pastries	350					
14		Plain cakes	500					
15		Chocolate Cakes	350					
16		Lemon Tarts	200					
17		Chocolate pastries	150					
18		Chocolate Cakes	200					
19								
20		Chocolate Pastries	438					

If we wish to avoid typing the condition in the formula, we can type in a separate cell and provide a reference to the cell in the formula, as shown below:

IRR								
	A	B	C	D	E	F	G	H
1								
2		AVERAGEIF Function						
3								
4		Item	Sales					
5		Chocolate pastries	1,000					
6		Pineapple pastries	500					
7		Lemon Tarts	650					
8		Pancakes	600					
9		Pineapple pastries	350					
10		Lemon Tarts	450					
11		Pancakes	350					
12		Chocolate pastries	250					
13		Chocolate pastries	350					
14		Plain cakes	500					
15		Chocolate Cakes	350					
16		Lemon Tarts	200					
17		Chocolate pastries	150					
18		Chocolate Cakes	200					
19								
20		Chocolate Pastries	438					
21			=AVERAGEIF(B5:B18, B20, C5:C18)					

Example: How to find the average, with wildcard characters!

Let's now see how we can use wildcards in the AVERAGEIF function. Using the data above, but now I wish to find out the average sales of all types of pastries sold.

As in our search, the keyword is to be preceded by some other character, so we will add an asterisk in front of the word. Similarly, we can use it at the end of the word if the keyword will be followed by some other character.

Now, the formula we will use would be =AVERAGEIF(B5:B18, "*Pastries", C5:C18).

IRR X ✓ fx =AVERAGEIF(B5:B18, "*Pastries", C5:C18)						
	A	B	C	D	E	F
2		AVERAGEIF Function				
3						
4		Item	Sales			
5		Chocolate Pastries	1,000			
6		Pineapple pastries	500			
7		Lemon Tarts	650			
8		Pancakes	600			
9		Pineapple pastries	350			
10		Lemon Tarts	450			
11		Pancakes	350			
12		Chocolate pastries	250			
13		Chocolate pastries	350			
14		Plain cakes	500			
15		Chocolate Cakes	350			
16		Lemon Tarts	200			
17		Chocolate pastries	150			
18		Chocolate Cakes	200			
19						
20						
21		All Pasteries	=AVERAGEIF(B5:B18, "*Pastries", C5:C18)			

We will get the result below:

C23							
	A	B	C	D	E	F	G
1							
2							
3							
4							
5							
6							
7							
8							
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							
19							
20							
21							
22							
23							
24							

Suppose we wish to find out the average of all items except pastries, we will use the following formula =AVERAGEIF(A12:A15, "<>*(Pastries)", B12:B15).

SUM		✕ ✓ <i>fx</i>		=AVERAGEIF(B5:B18, "<>*(Pastries)", C5:C18)	
	A	B	C	D	
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
21					
22					
23					
24					

AVERAGEIF Function	
Item	Sales
Chocolate Pastries	1,000
Pineapple pastries	500
Lemon Tarts	650
Pancakes	600
Pineapple pastries	350
Lemon Tarts	450
Pancakes	350
Chocolate pastries	250
Chocolate pastries	350
Plain cakes	500
Chocolate Cakes	350
Lemon Tarts	200
Chocolate pastries	150
Chocolate Cakes	200

All Pastries	433
--------------	-----

=AVERAGEIF(B5:B18, "<>*(Pastries)", C5:C18)

Example: How to find an average, with empty or not empty criteria?

Sometimes when doing data analysis in Excel, we may need to find an average of numbers that correspond to either empty or non-empty cells.

So, in order to include blank cells in the formula, we will enter "=", i.e., the "equals" sign. The cell will not contain anything that is not a formula or zero-length string.

C10		✕ ✓ <i>fx</i>		=AVERAGEIF(B5:B7, "=", C5:C7)		
	A	B	C	D	E	F
1						
2		AVERAGEIF Function				
3						
4		Item	Total cost			
5		Pastries	100			
6			100			
7		Dark Chocolates	200			
8						
9						
10			100			

For example, we are given the total cost of preparation of three items. Using `=AVERAGEIF(B5:B7, "=", C5:C7)` formula, Excel will calculate an average of cell B5:B7 only if a cell in Column A in the same row is empty, as shown below:

IRR	:	X	✓	<i>f_x</i>	<code>=AVERAGEIF(B5:B7, "=", C5:C7)</code>		
	A	B	C	D	E	F	G
1							
2		AVERAGEIF Function					
3							
4		Item	Total cost				
5		Pastries	100				
6			100				
7		Dark Chocolates	200				
8							
9							
10			<code>=AVERAGEIF(B5:B7, "=", C5:C7)</code>				

Suppose we wish to average values that correspond to blank cells and include empty strings that are returned by any other function. In such a scenario, we will use `"<>"` in the given criteria.

For example: `=AVERAGEIF(B5:B7, "<>", C5:C7)`.

C10	:	X	✓	<i>f_x</i>	<code>=AVERAGEIF(B5:B7, "<>", C5:C7)</code>		
	A	B	C	D	E	F	
1							
2		AVERAGEIF Function					
3							
4		Item	Total cost				
5		Pastries	100				
6			100				
7		Dark Chocolates	200				
8							
9							
10			100				

Now, suppose we wish to find the average of only those values that correspond to empty cells. In this scenario, we will use `"<>"` in the criteria.

So, now using the same data, the formula to use is given below:

IRR		✕ ✓ <i>f_x</i>		=AVERAGEIF(B5:B7, "", C5:C7)		
	A	B	C	D	E	F
1						
2	AVERAGEIF Function					
3						
4	Item		Total cost			
5	Pastries		100			
6			100			
7	Dark Chocolates		200			
8						
9						
10			100			
11			=AVERAGEIF(B5:B7, "", C5:C7)			

The result will be the average of Pastries and Dark Chocolates, i.e., 150.

C12		✕ ✓ <i>f_x</i>		=AVERAGEIF(B5:B7, "<>", C5:C7)		
	A	B	C	D	E	F
1						
2	AVERAGEIF Function					
3						
4	Item		Total cost			
5	Pastries		100			
6			100			
7	Dark Chocolates		200			
8						
9						
10			100			
11			100			
12			150			

Things to Remember about the AVERAGEIF Function

1. #DIV0! error – Occurs when:
 - No cells in the range meet the criteria
 - Range is a blank or text value.
2. When the given criteria is empty, AVERAGEIF treats it as 0 value.

3. The use of wildcard characters such as the asterisk (*), the tilde (~) in the function as criteria, greatly enhances the search criteria.
4. The function will ignore cells that contain TRUE or FALSE.

COUNTIFS

What is the COUNTIFS Function?

The COUNTIFS function is categorized under Excel statistical functions. COUNTIFS will count the number of cells that meet a single criterion or multiple criteria in the same or different ranges.

The difference between COUNTIF and COUNTIFS is that COUNTIF is designed for counting cells with a single condition in one range, whereas COUNTIFS can evaluate different criteria in the same or different ranges.

When doing **financial analysis**, COUNTIF helps in doing a quick analysis. For example, we are given a list of tasks to be completed by a department, and the dates and priority of each task. In such a scenario, we can prepare a table showing the date, count of each task, and their priority using the COUNTIFS function.

Formula:

=COUNTIFS(criteria_range1, criteria1, [criteria_range2, criteria2]...)

The COUNTIFS function uses the following arguments:

- **Criteria_range1 (required argument):** This is the first range that would be evaluated with the associated criteria.
- **Criteria1 (required argument):** The conditions to be tested against the values. The criteria can be in the form of a number, expression, cell reference, or the text that defines which cells shall be counted. For example, criteria can be expressed as 2, ">2," A4, "Mangoes," or "32."
- **Criteria_range2, criteria2, ... (optional argument):** Additional ranges and their associated criteria. The function allows up to 127 range/criteria pairs.

The given criteria can be any of the following:

- It is a numeric value. It can be an integer, decimal, date, time, or logical value.
- A text string. It may include wildcards. Wildcards can be a ? (question mark) or an * (asterisk). A ? matches any single character, whereas * matches any sequence of characters. If we wish to actually find the ? (question mark) or * (asterisk) character, we need to type the ~ symbol before this character in our search.

How to Use the COUNTIFS Function?

As a worksheet function, the COUNTIFS function can be entered as part of a formula in a cell of a worksheet.

To understand the uses of the function, let's consider a few examples:

Example 1: Let's assume we are given the following data:

	A	B	C	D	E
1					
2		COUNTIFS Function			
3					
4		Product	In Stock	Sold	
5		Cheese	25	5	
6		Brown Bread	30	0	
7		Cucumber	25	10	
8		Beetroot	20	5	
9		Onion	30	15	
10		Potato	40	2	
11					

We want to get a count of items that are in stock (value in column C is greater than 0) but remain unsold (value in column D is equal to 0).

IRR

:

X

✓

fx

=COUNTIFS(C5:C10,">0", D5:D10,"=0")

	A	B	C	D	E	F	G
1							
2		COUNTIFS Function					
3							
4		Product	In Stock	Sold			
5		Cheese	25	5			
6		Brown Bread	30	0			
7		Cucumber	25	10			
8		Beetroot	20	5			
9		Onion	30	15			
10		Potato	40	2			
11							
12							
13		=COUNTIFS(C5:C10,">0", D5:D10,"=0")					
14							

Here, as we can see, it is only Brown Bread where no stock is sold. The result we get is:

Product	In Stock	Sold
Cheese	25	5
Brown Bread	30	0
Cucumber	25	10
Beetroot	20	5
Onion	30	15
Potato	40	2

Formula bar: `=COUNTIFS(C5:C10,>0, D5:D10,\"=0\")`

Result: 1 Count

Example 2: Let's create count by month. For that, we need to use two functions – the COUNTIFS function and the EDATE function with two criteria.

Suppose we are given the following data:

Issue	Date	Priority
Getting Machine repaired	01/12/17	1
Meeting with Vendors	01/12/17	1
New laptop order	02/12/17	3
Issuing Salary checks	03/12/17	1
Making MIS reports	03/12/17	2
Identifying defective		
Inventory	03/12/17	1
Reorder Raw materials	03/12/17	1

The list of issues is in Column B. Each issue includes a date (column C) and priority (column D).

The formula generates the correct count for each date. For priorities, we used the formula below:

IRR : X ✓ fx =COUNTIFS(\$C\$6:\$C\$12,">="&\$B16,\$C\$6:\$C\$12,"<"&EDATE(\$B16,1),\$D\$6:\$D\$12,D\$15)

	A	B	C	D	E	F	G	H	I	J	K	L
1												
2		COUNTIFS Function										
3												
4												
5		Issue	Date	Priority								
6		Getting Machine repaired	01/12/17	1								
7		Meeting with Vendors	01/12/17	1								
8		New laptop order	02/12/17	3								
9		Issuing Salary checks	03/12/17	1								
10		Making MIS reports	03/12/17	2								
11		Identifying defective										
12		Inventory	03/12/17	1								
13		Reorder Raw materials	03/12/17	1								
14												
15												
16		Date	Count	1	2	3						
17		01/12/17	2	=COUNTIFS(\$C\$6:\$C\$12,">="&\$B16,\$C\$6:\$C\$12,"<"&EDATE(\$B16,1),\$D\$6:\$D\$12,D\$15)								

The formula returns a count by date but broken down by priority, which is picked up directly from row 6.

The end result is the number of tasks to be done in terms of priority-wise:

D16 : X ✓ fx =COUNTIFS(\$C\$6:\$C\$12,">="&\$B16,\$C\$6:\$C\$12,"<"&EDATE(\$B16,1),\$D\$6:\$D\$12,D\$15)

	A	B	C	D	E	F	G	H	I	J
1										
2		COUNTIFS Function								
3										
4										
5		Issue	Date	Priority						
6		Getting Machine repaired	01/12/17	1						
7		Meeting with Vendors	01/12/17	1						
8		New laptop order	02/12/17	3						
9		Issuing Salary checks	03/12/17	1						
10		Making MIS reports	03/12/17	2						
11		Identifying defective								
12		Inventory	03/12/17	1						
13		Reorder Raw materials	03/12/17	1						
14										
15										
16		Date	Count	1	2	3				
17		01/12/17	2	2	0	0				
18		02/12/17	1	0	0	1				
19		03/12/17	4	3	1	0				

3.14 EXCEL DASHBOARD

How to Create an Excel Dashboard?

Before creating a dashboard in Excel, **keep in mind the main objective.**

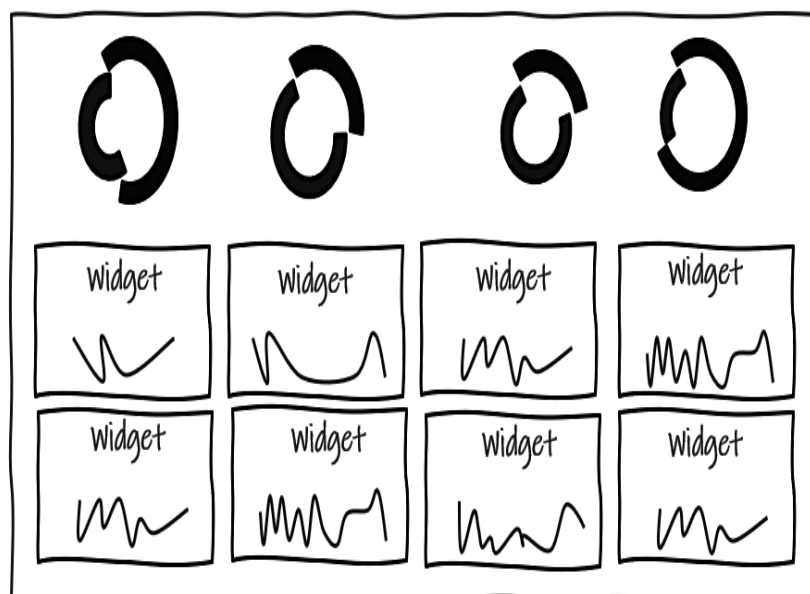
This section will help you create an Excel dashboard to track **HR activities**. Your goal is to show the monthly data on your main charts. Then, build a scorecard to compare the selected month and the past month.

The core of every Excel dashboard is **a one-page layout**. Why? Keep it in mind: a CEO doesn't always interested in the details.

1. Create a Layout for the Excel Dashboard

Create a proper draft! You can use paper and pencil, but we prefer Microsoft Excel to create mockups. We have used simple, grouped shapes.

Note: Let us review the effect of the Excel Dashboard UI mockup. In the figure below, we are showing a layout. First, select the type of grid dashboard layout that you will use. Then pick a colour scheme and font type and assign it to the report. Finally, make a wireframe that contains the following style, colour codes, and font types. You can prevent most issues if you work with structured data and data tables.



How can to create a logical workbook structure? What this mean? Open an Excel workbook and create three sheets.

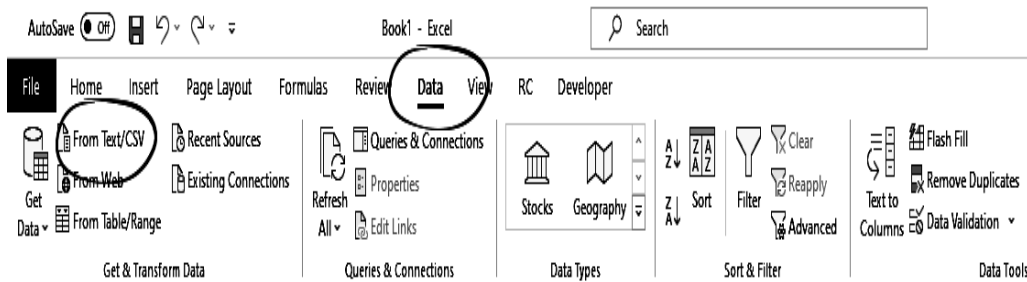
The parts of the workbook structure: Mostly, you use three worksheets for an Excel dashboard.

- **Data:** You can store the raw data tables here.
- **Dashboard Tab:** The main dashboard Worksheet.
- **Calculation:** Make the calculations on this Worksheet.

Your wireframe and structure are ready. Let's start creating a dashboard in Excel!

2. Get the Data into Excel

To create an Excel Dashboard, you need to choose data sources. If the data is present in Excel, you are lucky and can jump to the next step. If not, you have to use external data sources.



Go to the **Data tab** and **pick one of the import options**. It is easy to import data into an Excel workbook. In the example, you are **using a CSV file** to create the initial dataset for our dashboard.

3. Clean Raw Data

Our raw data is in Excel. Now, you can start the data cleansing process. **There are many tricks to clean and consolidate data.**

- Sort data to see extremes and peaks.
- Remove duplicates to avoid errors.
- Change the text to lower, upper or proper case.
- Remove leading and trailing spaces.

How to remove leading and trailing spaces from raw data? First, go to the Formula bar and **apply the TRIM function**. Now, copy the formula down. Finally, use **data cleansing tools** to avoid issues and clean your data faster and easier.

Note: Apply simple sorting in Excel to find errors! You'll find the peaks in a range (highest and lowest data) using sorted data. To do that, right-click on the first cell and select the '**Sort Largest to Smallest**' option from the context menu.

4. Use an Excel Table and Filter Data

You don't have cleaned input in this phase, but you already have data on a worksheet. What will be the next step?

First, you must **check if the required information is in a tabular format**. The tabular format means that every data point lives in one cell, *e.g.*, the name of the city, address, or phone

number. If it is in a tabular format, to should **convert it into an Excel table and select the data range.**

On the Insert Tab, choose Table or use the insert table shortcut, Ctrl + T.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1				OrderDate	Region	Rep	Item	Units	Unit Cost	Total				
2				1/6/19	East	Jones	Pencil	95	1.99	189.05				
3				1/23/19	Central	Kivell	Binder	50	19.99	999.50				
4				2/9/19	Central	Jardine	Pencil	36	4.99	179.64				
5				2/26/19	Central	Gill	Pen	27	19.99	539.73				
6				3/15/19	West	Sorvino	Pencil	56	2.99	167.44				
7				4/1/19	East	Jones	Binder	60	4.99	299.40				
8				4/18/19	Central			75	1.99	149.25				
9				5/5/19	Central			90	4.99	449.10				
10				5/22/19	West			32	1.99	63.68				
11				6/8/19	East			60	8.99	539.40				
12				6/25/19	Central			90	4.99	449.10				
13				7/12/19	East			29	1.99	57.71				
14				7/29/19	East			81	19.99	1,619.19				
15				8/15/19	East			35	4.99	174.65				
16				9/1/19	Central	Smith	Desk	2	125.00	250.00				
17				9/18/19	East	Jones	Pen Set	16	15.99	255.84				
18				10/5/19	Central	Morgan	Binder	28	8.99	251.72				

In this case, we don't have headers. Excel will automatically insert headers into the first row. If you need more data, you can only expand the table and not lose the formulas.

5. Analyze, Organize, Validate and Audit the Data

You took you through the method that converts raw data into a structure capable of creating a dashboard.

Ask yourself:

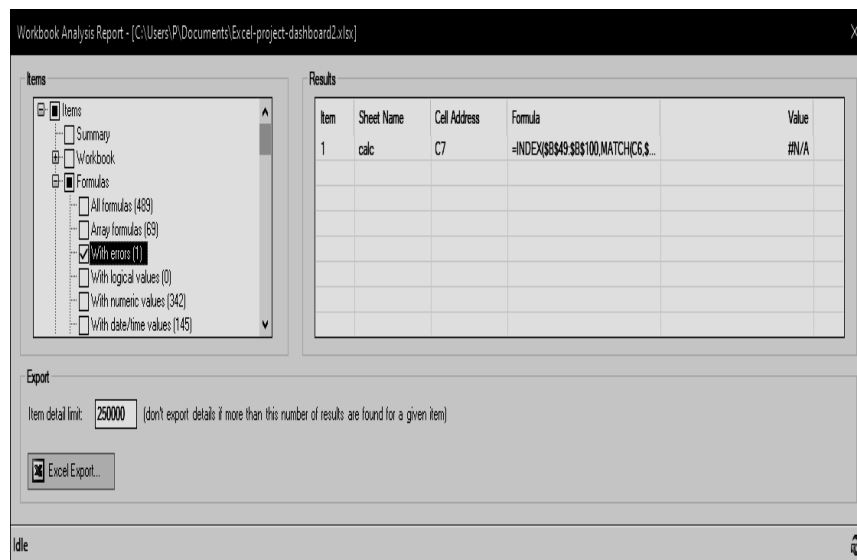
- Do you have to display all the data at once?
- Is it necessary to remove some data?

You can use **Excel formulas** and various methods that help us move forward from now on. However, you need creativity rather than knowing all the formulas to make a useful dashboard.

So you'll use these **functions** and tools to build the Excel Dashboard in Excel: **XLOOKUP, IF, SUMIF, COUNTIF, ROW, NAME MANAGER.**

Excel grants great auditing tools to help you find and fix Workbook or Worksheet issues.

Use **Microsoft Excel Inquiry** to visualize which cells in your Worksheet contribute to a formula error. This step should cut down the time spent on the usual validation procedures.



So, before we start creating a chart, you have to validate the data.

If you want to analyze the data quickly, use **the Quick Analysis Tool**.

6. Choose the Right Chart Type for the Excel Dashboard

Now you have an organized, cleaned, and error-free dataset, it is time to **choose the proper chart**.

Data is useless without the ability to visualize it. Strike a balance between great looking Excel Dashboard and its function. First, you can choose what graphs are best for different goals.

- **Compare Values:** Their characteristic is that they merely show high or low values. Recommended types for charting are a Column, Mekko, Bar, Line, **Panel Chart**, and **Bullet Charts**. Don't forget to check how the radial bar chart work.
- **Composition:** How can you display different sales results in different regions? The most fitting charts are Pie, Stacked Bar, Mekko, Stacked Column Chart, Area, and Waterfall Charts. We prefer the **heat map** also.
- **Analyzing Trends:** To analyze the result of a dataset in a given period, use the following charts: Line, Dual-Axis Line, and Column Charts. If to want to **create a quick forecast in Excel**, check this example.
- Show the differences between budget and actual values; use variance charts.
- **Performance Measurement:** Use **gauge charts** to see how far you are from reaching a goal. It displays a single value.
- **SPARKLINES ARE TINY GRAPHS** in a worksheet cell that offers a visual representation of the dataset. Use sparklines to show trends in a series of values. Another useful thing: you can highlight maximum and minimum values easily. So, the greatest impact of sparklines: you can position the chart near its data source.

- **Dynamic Charts** are essential if we want to create interactive charts to refresh the chart based on the user's choice.

Your goal is to show the % of job seekers who accepted a job offer each month on a chart. In this section, **you'll use custom combination charts** using doughnut charts – progress circle charts – for displaying key performance indicators.

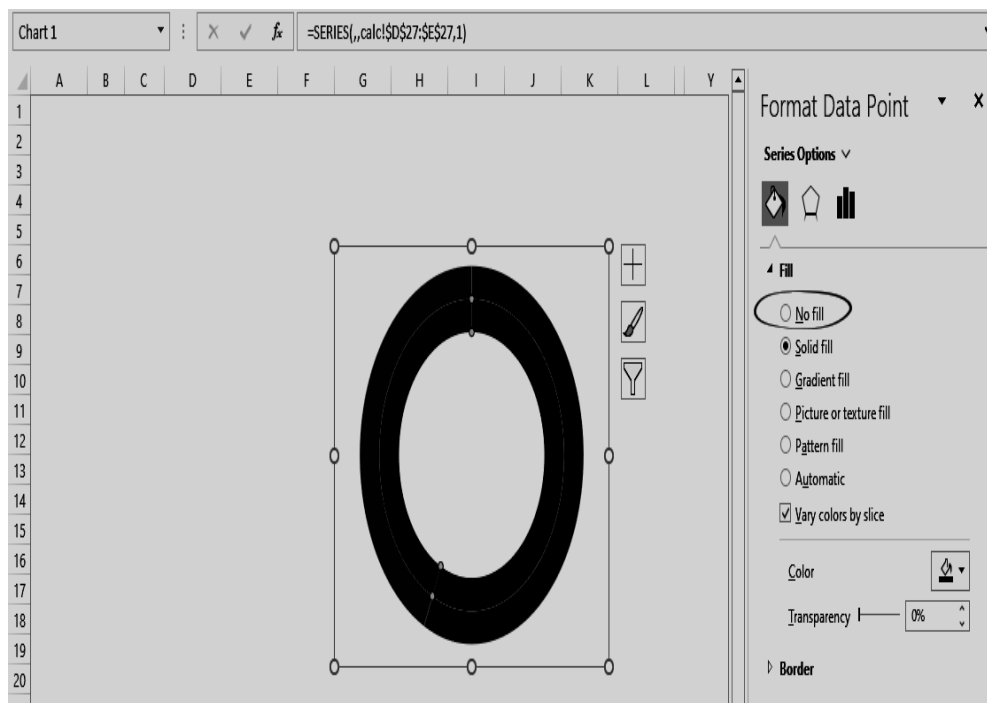
Note: Just a few words about the pie charts. Pie charts are the most overused graphs in Excel. It is one of the worst ways to present data. In other words, if you want to create a better dashboard, get rid of the pie charts!

7. Select the Data and Build Your Chart

We have cleaned and grouped data in this phase, and we've just picked the chart or graph for the data. It's time to select the data! As to learn before, **the combo chart requires two doughnut charts** and a simple formula.

Select the '**Calculation**' tab (which contains filtered data and calculated fields). Highlight the range of what you want to display. In the example, you use two values to show the **Acceptance Ratio**.

The actual value comes from the **Data tab**. After that, then calculate the **reminder value** using this simple formula. **In this case, 75%**. Next, select the 'Calculation' tab. Cell E23 will show the actual value. The second cell, E24, **contains a simple formula and displays the reminder value to 100%**.

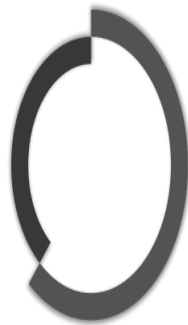


Make sure that the value in the source cell is in percentage format! Okay, now we select the ‘Actual Value’ and ‘Reminder Value’ data. Next, open the ‘Insert Chart’ dialog to create a custom combo chart to preview and choose different chart types. Furthermore, you can move the data series to the secondary axis.

Select the inserted chart and press **Control + C** to duplicate the chart.

8. Improve the Charts

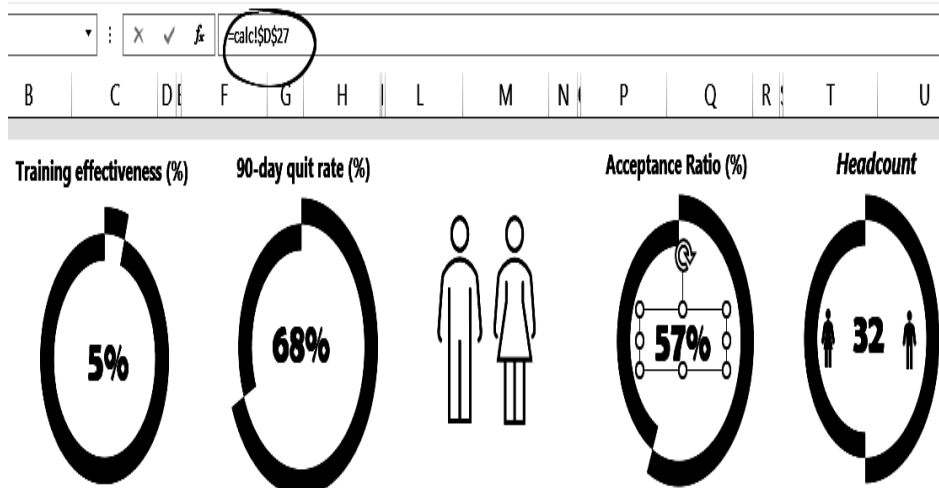
Now, you have a chart that’s fit your data. It looks great, but you can improve it to take the Excel dashboard to the next level! First, clean up the chart to remove the background, title, and borders from the chart area. Next, select the reminder value section of the outer ring.



Right-click, then choose **Format Data Point**. Use the ‘No fill’ option. Let’s see the inner ring, select the actual value section, and apply the ‘No fill’ option. Adjust the doughnut hole size if you want. Insert a Text Box and remove the background and border.

Link the actual value to the text box.

To do that, select Text Box. Next, go to the formula bar and press “=.” Next, select the actual value and click enter. Once Text Box is linked with actual value, format the text box.



Repeat the process for the other data! For example, **a typical Excel dashboard contains various charts to display data**. Next, repeat the chart insertion and data validation steps for other essential metrics, like the quit rate.

Keep the source data in the Data tab and do not remove or hide it. If further calculations are necessary, **use the Calculation Worksheet**. If you want to replace the source data, use the Calculation sheet, not the Data Worksheet.

Note: If you are uncertain about which types of charts are good for you, don't hesitate to choose **'Recommended Charts'**. In this case, you will get a custom set that Excel thinks will fit best with the data.

9. Create a Dashboard Scorecard

Your Excel dashboard is almost ready. You need only a few components to create a scorecard:

- Label,
- Actual value,
- Annual trendline,
- Variance (between the selected and the last month).

Absence rate	Absence cost	Benefits satisfaction	Productivity rate	Satisfaction index
12%	\$1,180	77%	74%	94%
				
141.0% vs. previous month	-2.2% vs. previous month	4.1% vs. previous month	7.2% vs. previous month	10.6% vs. previous month

Because you need a little bit more space, merge the cells. Select the cells to place the components and click on the 'merge cells' button. Now, link the label name from the 'Data' sheet. If you change the name of the value on the 'Data' sheet, the widget label will reflect. Now, link the data from the 'Data' sheet to a 'Dashboard' sheet.

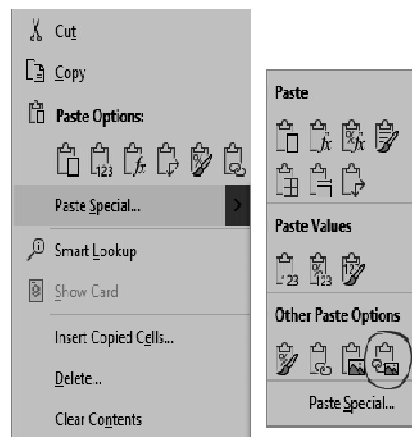
Go to the formula tab, enter an equal sign, and select the 'Data' sheet value. Next, use yearly data on the 'Data' sheet and insert a line chart to create a trendline. To highlight the variance, use a little trick. Go to the 'Calculation' sheet and create a helper table.

Create three new conditional formatting rules.

	C	D	F	G	H	I	J	K	L	M	N	O	P
KPI		Jun	May	Var.									
Absence rate		12.0%	5.0%	141.0%									
Absence cost		1180	1207	-2.2%									
Benefits satisfaction		77.0%	74.0%	4.1%									
Productivity rate		74.0%	69.0%	7.2%									
Satisfaction index		94.0%	85.0%	10.6%									
Engagement index		49.0%	33.0%	48.5%									
Internal promotion rate		8.5%	11.9%	-29.0%									
Net Promoter Score		7	6	16.7%									

Select the cell which contains variance and copy it. Then, navigate to the ‘Dashboard’ sheet and apply the ‘Paste Special’ option. Next, choose the ‘Paste as linked Picture’ option. Working with linked pictures is easy.

Check the steps in the picture below:



We want to add dynamic text to the main sheet to indicate the changes in key metrics. You link a text to the object you have inserted into the main Excel dashboard from now on. Then, if you change the value on the source sheet, the target cell will show the refreshed value. What a nice feature! You can apply this trick for textboxes or charts, like sparklines.

Best Practices for Creating Visually Effective Excel Dashboards

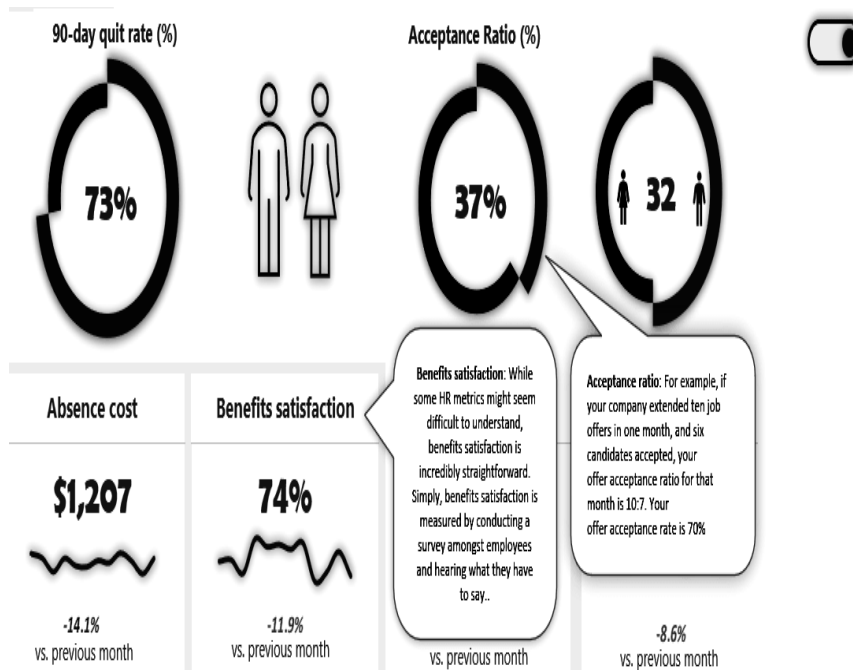
- A **Drop-down List** is a space-saving solution that is of great value when you create one-page dashboards. You can use data validation to control the type of data or the values that users type into a cell. To build the list of options is to type them on a

worksheet. You can do this method on the sheet with the drop-down menus or a different worksheet.

- **Conditional Formatting** is the right choice if you want to highlight cells based on any condition or rule. Of course, you can use other methods besides colours. You can achieve splendid results using icons, bars, shapes, colour scales, indicators, and ratings.
- **Named Ranges:** You can call selected cells with any given name. Highlight a range that contains data. In the name box, write the chosen name, *e.g.*, 'sales.' From this point, you can save the time working with cells or ranges.
- Use a **Scroll Bar** to save space on your Excel dashboard.
- **Data Validation:** Restrict what users can write in a single cell. Just imagine that ten users in 10 Excel workbooks write phone numbers. If you do not restrict the format of the phone numbers with the help of data validation when summing up the spreadsheets, there might be mistakes.
- **Data Entry using UserForm and VBA:** A manual data input data always carry errors. Instead, use the UserForm and write a short macro for it. You can create a user-friendly form that is easy to customize. Active report parts like form controls or pivot table slicers suggest playing with the chart.
- **Excel Pivot Tables** are the most potent weapon in Excel when we are working with large datasets. It is easy to use with only a few clicks; we can summarize data and drill down the data into any chosen structure.

Improve the Excel Dashboard

Here are two great tips for dashboard designers: To create **interactive screen tips**, visit our guide! Then, by clicking **on the toggle button**, you can show or hide the text.



3.15 STORYBOARD

Making a video can be a complicated undertaking. you've got to come up with a storyline, manage lots of small details, and decide how to edit the finished product. Luckily, there's a low-tech tool that can make every aspect of video creation easier: **a storyboard**.

So, what is a storyboard?

Creating a storyboard simplifies every stage of making a video. You'll be able to plan more effectively, pitch the idea to other people more clearly, and avoid problems during production. If you make any kind of business or marketing videos, storyboarding is an essential skill for you. By the time to finish reading this article, you'll have a clear understanding of **what a storyboard is** as well as all the knowledge you need to start storyboarding the next video.

What is a Storyboard?

A storyboard is a visual outline for your video. It is made up of a series of thumbnail images that convey what happens in your video, from beginning to end. It also includes notes about what is happening in each frame. A finished storyboard looks like a comic strip.

Storyboards can be simple or complex. They are usually hand-drawn, although some people prefer to use storyboarding software to create their images. A storyboard is similar to a script, but the two are not quite the same – storyboards are visual, while scripts are text-based.

			
Scene 1 Wide to medium zoom in on door	Scene 2 Medium of mom, dad, & kids leaving and saying goodbye	Scene 3 Mom & kids get on escalator	Scene 4 Mixed shots on escalator and calm elevator music
		Are you going to extremes to keep your family safe on winter roads?	Make it easy this winter www.abctires.com
Scene 5 Medium shot stepping off escalator into truck	Scene 6 Long shot of monster truck driving away	V.O. Same ↑	V.O. get the #1 rated all season tires at ABC Tires
		Scene 7 Slide	Scene 8 Slide

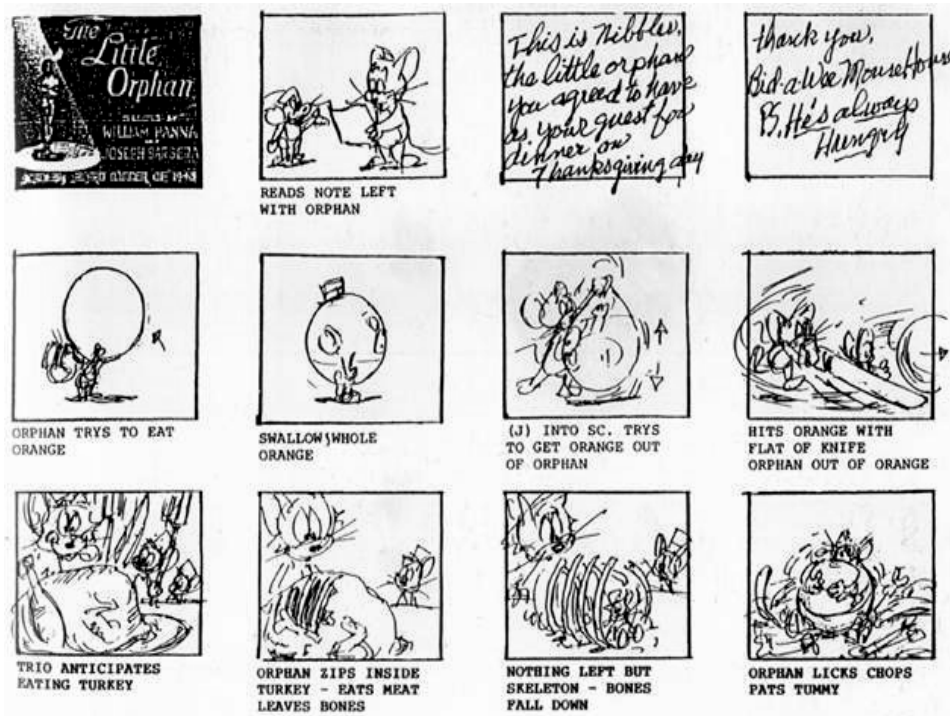
Types of Storyboards with Examples

Storyboards have always been a part of filmmaking and video production, and it has evolved over the years. There are now different types of storyboards that you can use depending on your skill level and type of video. Here are some different types of storyboards to take the video from script to screen:

- Traditional Storyboards:** Traditional storyboards are a series of pencil or ink drawings that help to visualize the video before filming begins. They can be sequentially arranged on a wall or in a spiral-bound book for easy reference. The drawing itself can be as detailed as you want it, and you can use it to sketch each shot to a few scenes in your script.

Many traditional storyboards also have written notes that describe what is going on in the scene. The best thing about this type of storyboard is that it is easy to edit.

Here is a sample of a traditional storyboard from Tom and Jerry:



- **Thumbnail Storyboards:** Thumbnail storyboards involve making small sketches – the size of a postage stamp – of the shots in your video on a few pieces of paper. Due to the small size, they are less detailed and take less time than traditional storyboards. You can use it to quickly draft your ideas before making drawing a more detailed version, like a traditional storyboard.

Check out this thumbnail storyboard example from Moulin Rouge:



- Animated Storyboards:** Technology has made it possible to create more complex storyboards that are animated and have some dialogue and music. Essentially, animated storyboards are different sketches that are filmed to show the proposed timing, pacing, and overall visual flow of the scene.

Such details may be hard to convey with sketches. Making animated storyboards does not have to be complex. You can use Wave.video to combine the photos of your sketches into one video within a few minutes. You won't need to download or learn new software to make your drawing come to life.

Why Do You Need a Storyboard?

Storyboarding adds an extra step to your video creation process, but it's well worth the time and effort. Making a storyboard pays off in a number of important ways. So, if you are still asking yourself "What is a storyboard? How can I benefit from it when making my videos?", here are a few of the top benefits:

- **Organize the Thoughts:** Making a storyboard helps you condense all the ideas bouncing around your head into one coherent, fleshed-out vision. You can use the storyboards as a reference during production to remember all your great ideas.

Apart from that, you can use storyboards to test different ideas, camera angles, and cuts without losing your vision. By the time you are ready to begin filming, you will have a concrete idea that you can easily communicate.

- **Plan Better:** It helps you plan your video more effectively. Video requires coordinating a lot of small but important details. Because of this, it is difficult (if not downright impossible) to create a good video on the fly. The more thoroughly you can plan the video, the better the finished product will turn out.
- **Share Your Ideas:** It helps you communicate your ideas to other people. You might be able to envision exactly how you want your video to look, but it can be hard to convey your ideas to other people using only words. A storyboard bridges that gap for you, helping your colleagues or clients envision the kind of video you have in mind.

Storyboards are a great way to secure executive interest and get buy-in from stakeholders. It makes them less likely to resist your idea down the line.

- **Identify Problems Early:** Much like dress rehearsals, storyboards get you one step closer to perfection in your final video. It reveals problems and weaknesses that you may not have noticed until production or post-production – at which points it might be too expensive to fix. For example, you may realize that you need more props, visual effects, or equipment than you anticipated.
- **Simplify Everything:** It simplifies everything that comes later. Investing some **time and effort up-front** in a storyboard can save you much more energy, and frustration later on. If you've got a good storyboard before you start shooting, you'll be able to bring the video to life more efficiently, and you'll be less likely to make expensive mistakes or waste time on ideas that don't pan out. Having a storyboard also makes editing your video more straightforward.
- **Time and Budget Savings:** Storyboards are essential to making effective marketing videos that stay on budget. You can avoid shooting useless footage or doing extra

work to shoot more footage after production. Time is money, and every minutes spent on set costs real dollars in salaries and equipment rentals. Storyboards reduce unnecessary work and cut right to the chase.

3.16 HR SCORECARD

An HR scorecard provides a framework for measuring human resources. It is designed with the balanced scorecard measurement framework. This framework, as we know, is developed based on four perspectives, *i.e.*, financial, internal process, customer, and learning and growth. Often, organizations hurriedly go for measurement. But quantifying HR measurable primarily requires understanding what needs to be measured. That will justify the logic behind developing a scorecard. Before we design an HR scorecard, it is also important to draw tentative lists of operating metrics, so that the scorecard can effectively help in the HR decision-making process: It is unwise to put all measures in one scorecard, rather it has to be structured using a sequence such as value-based HR metrics, HR outcome measures operational HR metrics and HR analytics. With this sequence, we can use an HR scorecard meaningfully and provide support to the HR decision-making process.

Some of the other contextual issues for an HR scorecard are: value drivers, value propositions, human resources' value, its contribution, measurable HR functions, activity- and value-driven metrics, human capital metrics and so on. All these issues are directly attributable to organizational business. Hence, designing an HR scorecard factoring such contextual issues can make it more meaningful. We try to get results from an HR scorecard in three important dimensions including the business values of human resources, HR processes and employees to an organization. Preparing an HR scorecard with value-based metrics may not be a straightforward journey, but it needs to be scaled up gradually, foremost with HR operational and efficiency metrics.

It is said an HR scorecard is an organic design, as it evolves through the changes in business – one that changes with the business. Obviously for this reason, a scorecard has to be continuously reviewed and updated, ensuring it generates reports with relevant metrics. Old metrics are gradually discontinued. This can make an HR scorecard relevant to organizational business needs, rather than reporting of some operational and efficiency metrics.

Therefore, an HR scorecard aligns HR support with organizational business and strategies. It demonstrates HR value contribution in achieving organizational business and strategic intents. It helps in strategic alignment of HR functions by mapping critical HR supports such as performance management systems, innovation, training and development and so on. An HR scorecard creates strategic alignment between human resources and the organization at all levels.

For developing an HR scorecard, we need to follow a certain sequence of actions as follows:

- (a) Develop a strategy.
- (b) Map the strategy.
- (c) Establish measures for each strategy.

- (d) Cascade it to a scorecard.
- (e) Implement the scorecard.

All the earlier listed actions can help us in developing a suitable HR scorecard that can benefit organizations in multiple ways. Some of these benefits are presented as follows:

- Helps in framing an HR strategy in alignment with the organizational strategy.
- Helps in getting clarity on the vision and mission of the organization.
- Creates a sense of ownership of the strategy.
- Improves communication of strategy.
- Prioritizes HR initiatives aligning activities to business goals.
- Supports functions to identify and communicate unique strategies.
- Creates a framework for prioritization of initiatives.
- Aligns measurement with business goals.
- Measures human resources' strategic contribution in candid terms.
- Provides real-time graphical display of KPIs.

Assessing HR Program

A comprehensive and effective evaluation plan is a critical component of any Successful training program. It should be structured to generate information of the impact of training on the reaction; on the amount of learning that has taken place; on the trainees' behaviour; and its contribution to the job/organization. Therefore, evaluation of training is a measure of how well it has met the need of its human resources.

“Any attempt to obtain information (feedback) on the effect of training program and to assess the value of training in the light of that information for improving further training.”

— Hamblin

“The systematic collection of descriptive and judgmental information necessary to make effective training decisions related to the selection, adoption, value, and modification of various instructional activities.”

Assessing the effectiveness of the training program in terms of the benefits to the trainees and the company must look into the following aspects:

- process of collecting outcomes to determine if the training program was effective.
- from whom, what, when, and how information should be collected.

Purpose of Assessment

- Determine accomplishment of specific training objectives.
- Determine cost-effectiveness of training.
- Explain program failure, if it occurs.

- Enhancement of the credibility of training and development if tangible benefits to an organization are proved.
- Determine whether the program is meeting the intended objectives.
- Identify strengths and weaknesses.
- Determine cost-benefit ratio.
- Identify who benefited most or least.
- Determine the future participants.
- Provide information for improving HRD programs.
- Reinforce major points to be made.
- Gather marketing information.
- Determine if training program is appropriate.
- Establish management database.

Data Collection for HRD Assessment

There are several techniques of data collection. They are as follows:

- Observation
- Questionnaire
- Interview
- Written Test
- Simulation/Performance Test

Advantages and Disadvantages of HRD Techniques

<i>Methods</i>	<i>Advantages</i>	<i>Disadvantages</i>
Interview	• Flexible • Opportunity for clarification • Depth possible • Personal contact	• High reactive effects • High cost • Face-to-face threat potential • Trained observers needed
Questionnaire	• Low cost • Anonymity possible • Respondent sets pace • Variety of options	• Possibly inaccurate data • Return rate beyond control
Observation	• Non-threatening • Excellent way to measure behavioural change	• Reactive effect • Probable • Trained observers needed
Written Test	• Low purchase cost • Readily scored and quickly processed • Easily administered • Wide sampling possible	• Possible low relation to job performance • Reliance on norms may distort individual performance • Cultural bias

Simulation/ Performance Test	• Reliable • Objective • Close relation to job performance	• Time-consuming • Simulation often difficult and costly
------------------------------------	--	---

Employee Engagement and HR Scorecard

Engaged employees play a crucial role in the success of any organization. Employees who are actively invested in the company's goals do not just perform consistently, but they also contribute new ideas, support morale, and look for ways to help the business grow. Employee engagement does not just result in a positive workplace culture; it impacts the bottom line and affects how well the company is achieving work objectives. That's why Human Resources teams must look for ways to measure the employee engagement within their organizations.

The following are the three Key Performance Indicators (KPIs) that will help assess engagement.

1. Turnover Rate

Measuring the turnover rate gives idea about employee engagement levels in an organization. HR teams can measure employee engagement by regularly monitoring the company's turnover rate. Specifically, it helps to try and link the turnover rate to particular interventions the company has taken to boost workplace engagement. For example, if an organization implements a new strategy to boost employee satisfaction, monitor the turnover rate to see if it has a positive effect. On the other hand, if the turnover rate during one year (or any other clearly defined time period) is worse than expected, HR teams should try to identify if any changes at the company may be to blame.

2. Suggestion Boxes

Employees tend to be more engaged with their work when they feel their voices are being heard. In other words, something as simple as a suggestion box or a **digital feedback tool** could not only help the organization get a better sense of how engaged their employees are, but also create an environment in which engagement among workers is much more common. A company's HR department should coordinate with all office managers and relevant team supervisors to develop a simple, easy-to-implement procedure for accepting employee suggestions and feedback.

3. Engagement Surveys

An employee survey can yield useful insights when implemented properly. Create a scoring key for the survey to determine which survey results represent the feelings of an engaged employee, which represent the feelings of a neutral employee, and which represent those of a disengaged worker. A pulse survey (which focus on a specific issue) tends to be short and usually consists of no more than 10 questions. This allows your HR team to distribute them as often as once a week if you decide that would be helpful. Analyzing the data and evaluating the results would be like taking steps to address this problem which will yield major rewards in the long run.

Finding Money in Analytics

Analytics capabilities are being used to help companies optimize people, processes, and technology. From staffing to inventory to raw materials, analytics can give companies visibility into their precious resources and help ensure that those resources deliver optimal value.

For analytics to help the organization save or make money, careful implementation of analytics initiatives is critical. The “test and learn” approach, outlined in *Analytics – All about The Money*, helps ensure that small-scale changes are tested thoroughly before they are rolled out to the larger organization, minimizing costs and risk while improving outcomes. Saving money with analytics can also mean reducing investment in the organization’s analytics infrastructure. This involves taking a closer look at the enterprise data warehouse. A careful assessment of what decisions the warehouse will (and will not) support can help contain infrastructure costs.

Organization can make money with analytics through incremental revenue gains from sales, marketing, and new product development activities. Analytics can help businesses sell more products, receive higher profits for each product sold, develop products customers want to buy, and through effective targeting and promotions. To save and make money with analytics, companies should turn hype into quantifiable results. Big data “finishing expeditions” have to be replaced by careful, productive analysis of business problems, and buzzwords exchanged for effective cost-benefit analysis.

Making money from big data depends on a company’s ability to effectively leverage the data that is available for analysis. Here are five examples of ways companies can use big data to their financial advantage:

1. Make time to compare different variables to see which perform or convert better. Details as minor as colour or font can influence consumers.
2. Obtain data from the broadest range of sources possible to get a clear understanding of the customer experience.
3. Process data in real time, enabling rapid innovation that meets shifting demands.
4. Remember to adapt to trends, continue to test for optimization to see where performance could be improved.
5. Utilize a combination of qualitative research and quantitative research, rather than relying too heavily on one or the other, to get a more complete picture.

Linking HR Data to Operational Performance

HR data plays a critical role in measuring the success of an organization’s HR transformation, gauging employee engagement and satisfaction, or calculating the inputs to help solve broader strategic business objectives.

The key is determining the requisite “Measures that Matter” and right-sizing the organization’s reporting tools and capabilities to provide insight into the achievement of those objectives organizations should develop a comprehensive HR measurement and reporting strategy

that is vertically aligned with the overall business strategy (leveraging the “top-down” approach) and can provide varying levels of information as needed.

There are three categories of HR measurements, each with different objectives and uses:

1. HR Reports:

- HR Reports are the least mature of the three categories of HR metrics and measures, and are focused on transactions.
- HR Reports consist of a formatted and organized presentation of data relating to the performance of daily operations.
- These operationally focused metrics utilize standard data extracts to provide a detailed analysis about a particular aspect of the organization.
- HR Reports enable more immediate, short-term decision-making in support of optimizing the specific operations being measured.

2. HR Dashboards:

- HR Dashboards provide a more advanced way to assess metrics and Key Performance Indicators (KPIs), allowing organizations to present information in a more interactive and user-focused fashion.
- HR Dashboards are designed to be dynamic and visual, including charts and graphs that illustrate key trends and insights and enable stakeholders to filter information according to their needs.
- HR Dashboards help “manage the business” and are used to monitor and drive performance improvements across HR processes in support of broader leadership and organizational objectives.

3. HR Scorecards:

- HR Scorecards, the most mature and innovative method of visualizing HR metrics and measures, use KPI’s focused on areas of critical performance to help “direct the business.”
- An HR Scorecard provides a snapshot of overall HR performance against strategic goals at a particular point in time, allowing executives and business leaders to monitor and manage the results of HR’s key objectives.
- This strategically focused category of HR metrics and measures should be updated periodically, based on the timeline of an organization’s strategic goals.

Within the three primary categories of HR measurements outlined above, there are four key performance areas:

1. Customer Metrics:

- Customer Metrics quantify a company's internal "customer" satisfaction with HR services and HR's ability to meet its users' needs.
- These customer-focused metrics are effective at identifying improvement opportunities to enhance the quality of HR services and range from general (*e.g.*, overall employee satisfaction) to specific (*e.g.*, the percentage of leaders citing leadership training as a driver of satisfaction) depending on the broader business questions these figures seek to answer.

5. Process Metrics:

- Process Metrics measure the output and efficiency of HR processes and address high-level issues such as the effectiveness of process execution, the identification of opportunities to improve efficiency or reduce errors, and determining the overall volume of transactions being measured.
- Process Metrics include insights such as job offer acceptance rate, number of HR data errors, and variance from compensation guidelines.
- This type of metric is especially helpful in identifying opportunities to improve HR processes and delivery mechanisms.

6. Talent Metrics:

- Talent Metrics help to identify competency gaps in a company's workforce, recognize the implications of different hiring, attrition and promotion strategies, and seek to understand which talent characteristics foster high performance.
- Talent Metrics include succession plan promotion rates, retention of high performers or critical segments, and the share of new hires who receive top performance ratings or leave within six months.
- Talent Metrics quantify the strength of the talent pipeline and identify improvement opportunities that will ultimately help with overall retention and employee productivity.

7. Financial Metrics:

- Financial Metrics quantify the cost and impact of HR processes and programs, and include information such as the cost of turnover, training spend per employee, return on investment of an HR initiative, and the realization of business case savings from an HR Transformation project.
- These insights answer important business questions involving the monetary success of a program relative to industry benchmarks or previously determined goals.

HR Data and Stock Performance

There are several brave consulting companies and individuals out there that have attempted to prove a statistical correlation between stock prices and HR metrics. Trying to correlate stock

performance with the amount of money a company invests in its human capital is one example. However, given the large swings in stock prices witnessed in recent years due to the impact of global events and instabilities, it is clear that external forces beyond the control of corporations are having more of an impact on stock price than internal events. For that reason, it is up to individual companies to determine whether such a study on their own stock price over time would yield a return on the investment required for this type of project. It is more likely that an RoI can be substantiated on programs that are internal and where the impact can be isolated and measured more easily.

To date, it is more often the academic and research world that has taken on this challenge rather than the corporate world. Some impressive work on this topic certainly exists and is worth reading, but for the HR leader, it may need more development before it starts to become a mainstream HR activity.

There is also the issue of applying general result to one's own company. As an example, it seems that there are HR articles released every day providing professionals with guiding results. An article reported that companies had seen a 10% increase in turnover. Before acting on this information, one question should be kept in mind.

Creating HR Scorecard

Tracking and monitoring Human Resource (HR) functions in a business organization typically involves creating a scorecard or report that lists key performance indicators, such as benefit costs as a percentage of operating expenses. By aligning HR strategic goals with the company's, HR professionals ensure that their efficiency contributes to the company's overall profitability. The scorecard, usually a spreadsheet of data, helps these managers prove that effective HR strategies represent a competitive advantage due to cost savings or cost avoidance.

Like any department, an organization's HR department needs the right framework to understand their goals. Without it, "success" is a vague target, and understanding goals can be difficult. Since 1992, balanced scorecards have played an important role in the business philosophy of many companies. The concept was first introduced by the Harvard Business Review, which coined the phrase "balanced scorecard" as a set of BI (Business Intelligence) data points used to make, measure, and meet goals. Successful balanced scorecards are specific and from a human resources perspective, along with the internal perspective, a holistic perspective will help the HR department find the right areas to focus on in the future.

HR scorecard will need the following perspectives:

- 1. Financial perspective:** The financial goals relate to the organization's larger financial perspective which include:
 - Reducing costs related to HR
 - Reducing turnover rates
 - Employee absenteeism
 - Workers' compensation

- Employee satisfaction
- Employee engagement
- Employee benefits

By tracking these measurable KPIs, the HR department can create a baseline of balanced scorecard metrics that can then be correlated to the overall financial health of the organization.

2. **Internal Perspective:** Finding specific and measurable objectives for this perspective tend to involve factors that are more qualitative than financial numbers. Identifying “success” from this perspective can include factors like turnover rates, manager performance and ratings, overall employee satisfaction, and recruitment/employee retention data. It may also include tracking recruitment channels and relating them to employee success. These might include referrals, online applications (through to organization’s website), LinkedIn referrals, etc.
3. **Learning and Growth Perspective:** The learning and growth perspective focuses on communication and performance improvements, maybe by implementing best practices like –
 - To make sure everyone is aligned with your organization’s overall goals.
 - To ensure every department, team, and employee are aware of the data they need.
 - Can they access and understand the data easily?

The answer to these questions gives a picture of where the company’s learning and growth strategy is at and helps to pinpoint opportunities for improvement.

4. **Customer Perspective:**
 - Goal setting does not have to be solely interdepartmental because the organizations depend on departments working together towards common goals.
 - In order to increase overall customer satisfaction, this means consistent customer communication and relationships for starters. Other specific goals might include customer retention and return customer rates.
 - Generally speaking, customer satisfaction correlates with lower turnover rates, which will influence employee satisfaction.
 - Additionally, factors like employee productivity can affect turnover rate, etc. In the end, the customer service goals for HR should include direct actions that relate to finances, internal factors, and growth.

Developing a HR Measurement System

An HR scorecard measurement system allows a business to tangibly monitor HR performance. In performance-driven organizations, performance metrics are often used to express the value and success each department contributes. The efforts of human resources cannot be expressed as a direct, bottom line number contributing to sales, although the value of effective

HR practices may make significant contributions toward a company's success. The steps involved in developing a HR Scorecard Measurement system are:

- **Assessing the Corporate Culture:** Before designing a scorecard, the HR professional should understand top-level management goals and strategies for the enterprise. A manufacturing business in an industry with much competition may need low-cost labour to meet necessary margins, while a cutting-edge tech startup seeks innovative programmers, regardless of cost. How a company utilizes its workforce plays a role in how HR performance can be assessed? A particular business or industry may have its own priorities for HR assessment. A factory, for example, may focus on safety initiatives, while a call center watches turnover.
- **Operational or Outcome: The Scorecard Focus:** A business that applies metrics to elements of production often works in measures of efficiency, such as orders per hour or waste per 100 units. A simple implementation of metrics on a personnel level would simply transfer these operational measures from department to worker. While a natural extension of metrics practices, these have little value in assessing HR function. An HR scorecard is often better expressed in terms of outcome, and this may require long-term, broad picture approach. For example, if an HR objective is to grow a stable workforce with improved productivity, orders per hour performance is still valid, but across an entire department, and over a defined period of time. That information combines with turnover statistics for the same period to establish a scorecard metric.
- **Compiling the Metrics:** Referring to the strategic objectives of the company and matching these with outcome-based HR goals provides the basis for the HR scorecard. SMART methodology provides an outline for each scorecard item to be specific, measurable, achievable, relevant and time-constrained. For example, using the SMART acronym might develop a scorecard objective of "increased training" into "Health and safety training hours increased to 10 hours or more per employee over the next two quarters".

Implementing the HR Scorecard

HR Scorecard can be implemented across the organization by linking the HR activities with the strategic goals as follows:

- **Define Your Business Strategy:** Effective human resource professionals establish specific, measurable, achievable, realistic and timely objectives. For example, if your company wants to reduce its health costs per employee, you need to identify the cost of employee benefits and then define specific ways of reducing that cost, such as choosing another plan provider or offering different services.
- **Identify the Business Indicators:** Leading indicators include data, such as the number of building permits, which reflects current economic conditions. Lagging indicators, such as unemployment rates, confirm trends. In the past, human resource professionals functioned primarily in an administrative capacity handling benefits and payroll. More

recently, HR professionals have started to be seen as business partners enabling the hiring and developing of critical personnel, arranging for appropriate compensation for exceptional accomplishments and promoting the training and development of employees directly responsible for innovative business improvements.

- **Create the Measurement System for the HR Scorecard:** Typically, HR professionals need to show how they control costs by eliminating inefficiencies. For example, HR professionals strive to minimize employee turnover, which results in the need to recruit, interview and train new employees at considerable expense to the company. HR professionals create presentations, spreadsheets and other documents to list data. They distribute these through e-mail and websites.
- **Conduct regular meetings to review progress:** By reviewing the Scorecard on a regular basis with sponsors and stakeholders, such as other leaders in the business organization, HR professionals ensure their results align with the company's strategic efforts. When deficiencies occur, HR professionals identify the performance gaps. By providing skills training to employees and giving opportunities to them to enhance their competency in specific areas, such as project management or quality management, HR professionals help companies attract and retain the most talented employees to help fulfill strategic goals.

Strategic Map of How HR Creates Value for an Organization

- **Identifying HR Deliverables:** The two most important HR deliverables in any organization are a stable, high talent staffing in the R&D function and optimal staffing levels. These deliverables constitute many competencies that combine cutting-edge technical knowledge influencing growth and contributing to revenue growth as well.
- **High-performance Work System:** Once the HR deliverables have been clearly defined, we can begin, identify and measure the foundational High-Performance Work System (HPWS) elements that help to generate those deliverables. We term these elements a HPWS to represent the fact that they have been selected specifically with the intent of implementing strategy through HR deliverables.
- **Identifying HR System Alignment:** Identification of HR system elements that needs to reinforce one another so as to produce the two HR deliverables. In the case of stable, high-talent staffing in R&D, it can be assumed that the firm has developed a validated competency model. To achieve optimal staffing in manufacturing, HR must keep its recruiting cycle time short. The appropriate alignment measure can be a fourteen day recruiting cycle time which would reflect progress toward that objective.
- **Identifying HR Efficiency Measures:** The organization must be able to identify cost per hire as a strategic efficiency measure. For both deliverables, cost per hire will probably be higher than average. But, the benefits of those hiring processes will also be well above average.

An HR Scorecard for the entire company would include many more entries. More importantly, the scorecard is presented in such a way that it reinforces the required competencies.

3.17 INSIGHT INTO DATA-DRIVEN HRA

Introduction

Data-driven HRA, or intelligent HRA, is about using this data explosion in a smart way and extracting insight that not only improve the performance of people within the company (including its HR team), but also contribute to the overall success of the organization. HR teams can use data to make better HR decisions, better understand and evaluate the business impact of people, improve leadership's decision-making in people-related matters, make HR processes and operations more efficient and effective, and improve the overall well-being and effectiveness of people, which all can have a significant impact on a company's ability to achieve its strategic aims. Right now, this idea of the data-driven HR team does not exist in many organizations, outside of perhaps the very largest or most innovative companies, but it is certainly gathering pace. HR and people management are undergoing a revolution, and starting to be taken over by this wave of data and analytics. This part of business functionality that has traditionally focused on softer elements like people, culture, learning and development, and employee engagement is becoming increasingly driven by data analysis, and employee engagement is becoming increasingly driven by data analysis, and with good reason. HR is one of the most data-rich functions in most organizations. Employers have been using data and analytics to some extent or other for a long time now to understand metrics like staff satisfaction. The rise of Big Data has accelerated this practice, as well as taking it in exciting new directions.

The idea behind data-driven HRA is all about making HR smarter in every possible way and making smarter decisions is a huge part of this. Data can help HR professionals make better decisions about their own activities (such as smarter recruitment and performance reviews), and HR data also can be used to report to elsewhere in the company and support wider company decisions. After all, leadership teams need HR and people-related data to make their own decisions, and the intelligent HR team is well equipped to support this process.

Typical Data Sources

Broadly, the data required by an HR analytics tool is classified into internal and external data. One of the biggest challenges in data collection is the collection of the right data and quality data.

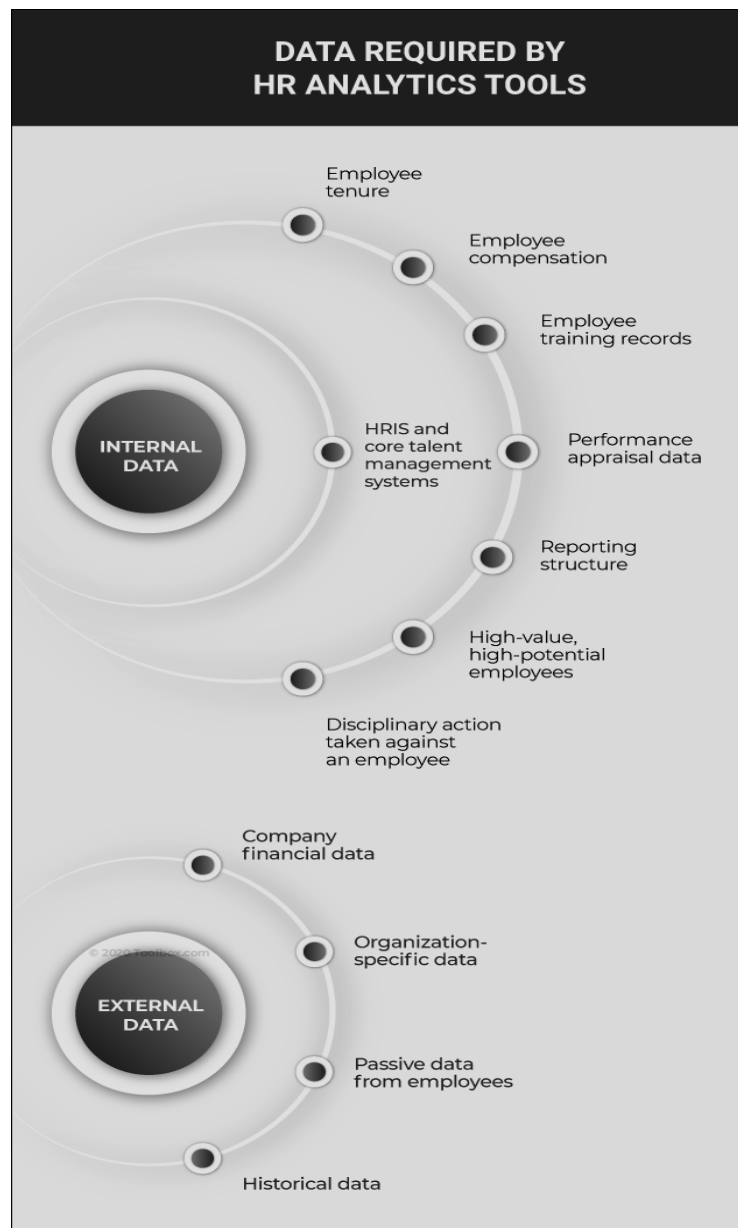


Fig. 3.24: Data Required by HR Analytics Tools

1. Internal Data

Internal data specifically refers to data obtained from the HR department of an organization. The core HR system contains several data points that can be used for an HR analytics tool. Some of the metrics that an HRIS system contains includes:

- Employee tenure
- Employee compensation

- Employee training records
- Performance appraisal data
- Reporting structure
- Details on high-value, high-potential employees
- Details on any disciplinary action taken against an employee.

The only challenge here is that sometimes, this data is disconnected and so may not serve as a reliable measure. This is where the data scientist can play a meaningful role. They can organize this scattered data and create buckets of relevant data points, which can then be used for the analytics tool.

2. External Data

External data is obtained by establishing working relationships with other departments of the organization. Data from outside the organization is also essential, as it offers a global perspective that working with data from within the organization cannot.

- **Financial data:** Organization-wide financial data is key in any HR analysis to calculate, *e.g.*, the revenue per employee or the cost of hire.
- **Organization-specific data:** Depending on the type of organization and its core offering (product or service), the type of data that HR needs to supplement analytics will vary. For example, says Collins, “HR leaders at a global retailer should power their analytics engine with store revenue and costs, and customer experience data whereas HR at a construction company might pursue operational – health and safety – data and data related to contingent labour costs.”
- **Passive data from employees:** Employees continually provide data that is stored in the HRIS from the moment they are approached for a job. Additionally, data from their social media posts and shares and from feedback surveys can be used to guide HR data analysis.
- **Historical data:** Several global economic, political, or environmental events determine patterns in employee behaviour. Such data can offer insights that limited internal data cannot. For example, the recession in 2008 was a global event that changed the way employees perceived jobs or “work.” The freelance, start-up and gig economies took off as people continued to lose their jobs. Data from such a critical historical event can help predict how the workforce may react to similar shifts in the future. It can then be used to identify trends in the current workforce and predict voluntary and involuntary turnover.

Typical Questions Faced (Survey)

As the field of people analytics expands to incorporate data gathered through relational analytics, organizational network analysis and other streams, the employee survey remains a critical tool to support a comprehensive people analytics practice.

Passive monitoring of email traffic, social media connections, proximity data, and other sources of data can indicate where there are strong and weak connections within organizations, but like demographic data or organizational hierarchy, these are descriptive statistics that do not measure the employee experience or how individuals perceive their workplace. Responses to employee survey questions provide qualitative data representing the opinions and perceptions of employees throughout the organization. Without this data, HR can at best describe the workforce; it cannot get at how employees feel about the company or their work, or the why of employee sentiment regarding the company, leadership, management, or culture.

Organizations seek to strike a relatively stable balance between culture and engagement. Leaders expect a constant flow of daily feedback and hence the need for information must be balanced against the employee point of view. However, there is no one-size-fits-all approach to surveying; the employee survey should be considered in terms of both people analytics and functionality, balancing the organization's need for information against what is asked, how it is measured, and how often. Most importantly, it is difficult to infer what leaders can do to improve the business when they are limited to quantitative descriptive statistics.

The design and conduct of surveys have a big impact on the quality of feedback received from employees. A survey strategy that relies solely on annual census surveys will likely miss critical moments in the employee experience and deprive leaders of information needed for managing change. Instead of a rigid survey schedule, a more agile approach can capture the most critical information at the most appropriate moment, while also reflecting changes and trends over time. Every employee survey should be designed with these three important points about survey design and frequency which are as follows:

1. Surveys represent a two-way conversation.
2. Survey design needs to align with the company's overarching strategy.
3. Survey design and strategy needs to be purposeful and flexible.

The typical questions asked in the HR Analytics survey may include (or can be customized depending on the situation) — the age, structure, size and strategy of the organization, its competitive intensity, labour-capital ratio, degree of application of HR Analytics in the organization like basic HR reporting (data from ERP and HCM systems), HR value-added metrics, talent management metrics, HR business driver analytics, etc.

Typical Data Issue

Some of the typical data issues are as follows:

1. The amount of data being collected

With today's data-driven organizations and the introduction of big data, risk managers and other employees are often overwhelmed with the amount of data that is collected. An organization may receive information on every incident and interaction that takes place on a daily basis, leaving analysts with thousands of interlocking datasets.

There is a need for a data system that automatically collects and organizes information. Manually performing this process is far too time-consuming and unnecessary in today's environment. An automated system will allow employees to use the time spent processing data to act on it instead.

2. Collecting meaningful and real-time data

With so much data available, it is difficult to dig down and access the insights that are needed most. When employees are overwhelmed, they may not fully analyze data or only focus on the measures that are easiest to collect instead of those that truly add value. In addition, if an employee has to manually sift through data, it can be impossible to gain real-time insights on what is currently happening. Outdated data can have significant negative impacts on decision-making.

A data system that collects, organizes and automatically alerts users of trends will help solve this issue. Employees can input their goals and easily create a report that provides the answers to their most important questions. With real-time reports and alerts, decision-makers can be confident they are basing any choices on complete and accurate information.

3. Visual representation of data

To be understood and impactful, data often needs to be visually presented in graphs or charts. While these tools are incredibly useful, it is difficult to build them manually. Taking the time to pull information from multiple areas and put it into a reporting tool is frustrating and time-consuming.

Strong data systems enable report building at the click of a button. Employees and decision-makers will have access to the real-time information they need in an appealing and educational format.

4. Data from multiple sources

The next issue is trying to analyze data across multiple, disjointed sources. Different pieces of data are often housed in different systems. Employees may not always realize this, leading to incomplete or inaccurate analysis. Manually combining data is time-consuming and can limit insights to what is easily viewed.

With a comprehensive and centralized system, employees will have access to all types of information in one location. Not only does this free up time spent accessing multiple sources, it allows cross-comparisons and ensures data is complete.

5. Inaccessible data

Moving data into one centralized system has little impact if it is not easily accessible to the people that need it. Decision-makers and risk managers need access to all of an organization's data for insights on what is happening at any given moment, even if they are working off-site. Accessing information should be the easiest part of data analytics.

An effective database will eliminate any accessibility issues. Authorized employees will be able to securely view or edit data from anywhere, illustrating organizational changes and enabling high-speed decision-making.

6. Poor quality data

Nothing is more harmful to data analytics than inaccurate data. Without good input, output will be unreliable. A key cause of inaccurate data is manual errors made during data entry. This can lead to significant negative consequences if the analysis is used to influence decisions. Another issue is asymmetrical data: when information in one system does not reflect the changes made in another system, leaving it outdated.

A centralized system eliminates these issues. Data can be input automatically with mandatory or drop-down fields, leaving little room for human error. System integrations ensure that a change in one area is instantly reflected across the board.

7. Pressure from the top

As risk management becomes more popular in organizations, CFOs and other executives demand more results from risk managers. They expect higher returns and a large number of reports on all kinds of data.

With a comprehensive analysis system, risk managers can go above and beyond expectations and easily deliver any desired analysis. They will also have more time to act on insights and further the value of the department to the organization.

8. Lack of support

Data analytics cannot be effective without organizational support, both from the top and lower-level employees. Risk managers will be powerless in many pursuits if executives do not give them the ability to act. Other employees play a key role as well: if they do not submit data for analysis or their systems are inaccessible to the risk manager, it will be hard to create any actionable information.

Emphasize the value of risk management and analysis to all aspects of the organization to get past this challenge. Once other members of the team understand the benefits, they are more likely to cooperate. Implementing change can be difficult, but using a centralized data analysis system allows risk managers to easily communicate results and effectively achieve buy-in from multiple stakeholders.

9. Confusion or anxiety

Users may feel confused or anxious about switching from traditional data analysis methods, even if they understand the benefits of automation. Nobody likes change, especially when they are comfortable and familiar with the way things are done.

To overcome this HR problem, it is important to illustrate how changes to analytics will actually streamline the role and make it more meaningful and fulfilling. With comprehensive data

analytics, employees can eliminate redundant tasks like data collection and report building, and spend time acting on insights instead.

10. Budget

Another challenge risk managers regularly face is budget. Risk is often a small department. So, it can be difficult to get approval for significant purchases such as an analytics system.

Risk managers can secure budget for data analytics by measuring the return on investment of a system and making a strong business case for the benefits it will achieve.

11. Shortage of skills

Some organizations struggle with analysis due to a lack of talent. This is especially true in those without formal risk departments. Employees may not have the knowledge or capability to run in-depth data analysis.

This challenge is mitigated in two ways: by addressing analytical competency in the hiring process and by having an analysis system that is easy to use. The first solution ensures skills are on hand, while the second will simplify the analysis process for everyone. Everyone can utilize this type of system, regardless of skill level.

12. Scaling data analysis

Finally, analytics can be hard to scale as an organization and the amount of data it collects grows. Collecting information and creating reports becomes increasingly complex. A system that can grow with the organization is crucial to manage this issue.

Connecting HR Analytics to Business Benefit (Case Study)

One of the biggest trends in business over the last few years has been the rise of data-driven HR. While most other functions have been using data and analytics for years, HR only recently moved to this approach. Collecting and analysing data can help in various areas of HR – from spotting employee trends over time to improving recruiting practices and strategies. Be it small and medium-sized business (SMB), organizations or the larger organizations, both can gain significant business advantages and success by implementing HR analytics into their business strategy to help lower turnover, increase retention, improve hiring and ultimately lead to business growth. The benefits of HR Analytics maybe seen in the following ways:

- **Measure the Success of Employee Development:** An engaging L&D program plays an important role in recruiting and retention, which has major implications for a business. This is where collecting, measuring, and analysing L&D data can be extremely beneficial. Not only can HR analytics help to identify the right courses for a workforce, but they can also help to uncover the number of employees who are taking and completing various courses.
- **Identify Workplace and Employee Trends:** The most important impact data and analytics has for a business is the ability to identify trends (short and long term) over time. HR is no exception for the importance of trends, as they can be extremely helpful

for both employers and employees. Tracking certain people analytics can help to show trends in employee sentiment, happiness, productivity, engagement, and morale just to name a few. Being able to identify trends and put efforts in place to address it can help businesses resolve workplace issues, improve employee happiness and productivity, and achieve greater business success.

- **Improve Recruiting Strategies and Results:** Recruiting is still one of the biggest challenges facing businesses today and employers are looking for ways to improve their overall results. In recent years, moving to a data-driven approach means collecting and interpreting recruiting data which helps employers to apply the recruiting metrics to ensure the likelihood of hiring the best talent for their organization and avoid making a bad hire (which can cost a lot of money for a business). HR analytics addresses issues like these which in turn help businesses build a competitive advantage in hiring for any business.
- **Boost the Employee Experience:** HR analytics can assist focusing significant amounts of time and resources in improving employee engagement and alternatively look at initiatives to address the employee experience by creating, sending and analysing employee surveys. Using HR analytics to identify ways to improve these important employee trends can help an organization in many ways leading to improved engagement, happiness, productivity and much more – all of which are critical to the overall employee experience and the health of the organization.
- **Measure Employee Turnover and Retention:** Talent Acquisition, Employee Retention and Turnover are critical for the health of an organization. Using HR analytics can help identify reasons for employee turnover, which can then be fixed to prevent further employees from leaving the organization. HR Analytics helps keep track of data around retention and turnover which helps the employers to understand the drivers that cause a great employee to leave or which can enable business owners to address and correct these issues or otherwise appreciate for the exemplary work done which in the process can save an employer a lot of money and prevent disruptions to a business that occurs when an employee leaves.

A Caselet on Connecting HR Analytics to Business Benefit

A large company named XYZ, employs, at any given point of time, around 4000 employees. However, every year, around 15% of its employees leave the company and need to be replaced with the talent pool available in the job market. The management believes that this level of attrition (employees leaving, either on their own or because they got fired) is bad for the company, because of the following reasons:

- The former employees' projects get delayed, which makes it difficult to meet timelines, resulting in a reputation loss among consumers and partners.
- A sizeable department has to be maintained for the purposes of recruiting new talent.

- More often than not, the new employees have to be trained for the job and/or given time to acclimatize themselves to the company.

Hence, the management has planned to implement HR analytics practices to understand what factors they should focus on, in order to curb attrition. In other words, they want to know what changes they should make to their workplace, in order to get most of their employees to stay. Also, they want to know which of these variables is most important and needs to be addressed right away.

Techniques of Establishing Questions

The various types of questioning techniques are as follows:

1. Open Questions

Open questions are an essential part of questioning techniques, and they deal in the broader discussion, explanations, and elaboration. These are framed in a manner of conversation between two people where questions help to describe the situation.

Open questions help in a better understanding of the topic under discussion as it allows endless questions and answers. It is useful in productive talks so that you can delve deep and extract more information.

When you are asking open questions, the person at the other end feel the interest, and this creates a strong bond. This method of questioning techniques is best for individuals or small groups, not large ones, as it will become challenging to extract information and give personal attention to everyone.

Examples of open questions are:

- What happened in the class today?
- What did you do in the office today?
- Who was present at the time of the incident?
- Explain, why it happened?
- What are your thoughts about the incident?
- Why did it happen?

2. Closed Questions

Closed questions are quick and easy to answer because they deal in short and often one-word answers. These are often yes-or-no answers asked for affirmations, agreements, disagreements, and understanding concepts.

These are not part of a conversation as asking a close question can put an end to it. Closed questions are easy to compare during statistical analysis.

The limitation of this questioning technique is that it does not allow the respondent to express himself or put forth his views, nor does it facilitate a better understanding between the people involved in the question and answers. Examples of closed questions are:

- Do you like hard drinks?
- Do you like soft drinks?
- Did you drive to the office today?
- Are you happy in the office?
- Where do you live?

Closed-ended questions which are aimed at collecting accurate statistical data can be classified into the following types:

- (i) **Leading Questions:** A question forcing the target audience to opt for a specific kind of answer is called a leading question. All answers for a leading question are almost similar. Leading questions are usually prepared to derive audience opinion within a set of limited words.
- (ii) **Likert Questions:** The degree to which respondents agree to a specific statement can be ascertained using Likert questions. Customers' feelings about a topic, product or service can be easily gauged by asking them these questions.
- (iii) **Dichotomous Questions:** Questions that make respondents answer with a simple "yes" or "no" are called dichotomous questions. These questions carry one disadvantage that there is no other way of analysing the answer between a "yes" and "no". A middle perspective is not possible.
- (iv) **Bipolar Questions:** Questions that have two answers with different levels of extremities, written at opposite ends of a scale, are called bipolar questions. Respondents have to mark their response anywhere between these two extremities, showing their opinion.
- (v) **Rating Scale Questions:** Questions that ask respondents to provide a rating on a specific matter on a scale of 1 to 10 or on a scale of "poor" to "good" are called rating scale questions. Normally, these questions have an even number of choices, so as to prevent respondents to choose a middle way out.

3. Rhetorical Questions

Rhetorical questions are used to engage the audiences. It encourages people to think out of the box and come up with innovative ideas. This type of questioning technique is not looking for answers.

These are statements that are phrased as questions to engage the other party and draw him into agreeing with to. Rhetorical questions make the conversation exciting and engaging, and are used to persuade people to their way of thinking. Examples of rhetorical questions are:

- Isn't it nice working in this office?
- Aren't all the team members cooperative?
- Isn't this a perfect fit?

4. Leading Questions

Leading questions are also known as reflective questions because of their nature. Here an environment is created by leading the respondent towards a specific route. It also encourages people to agree with you and say yes.

This type of questioning technique is useful for closing a deal, building positive rapport, and directing the conversation towards an outcome that you desire. Sometimes, leading questions are also used for manipulating situations to the advantage.

- Do you have any issues in the workplace?
- Do you enjoy working with the colleagues?
- Did anyone say anything to you?
- Did anyone misbehave with you?

5. Probing Questions

As the name suggests, probing questions are used to probe and extract information. It is looking for elaborate answers to clear all the related doubts. This type of questioning technique is most useful for encouraging others to open up and provide more information.

It involves a series of questions that encourage others to talk and give a full picture for better understanding. Examples of probing questions are:

- When do you need this information?
- What is this information needed for?
- Where will you use this information?
- Who requires this information and details?

6. Funnel Questions

Funnel questions are called by this unique name because of the type of questioning involved. A funnel is full at the mouth and gradually narrows down to the bottom, and so is the concept in this type of questioning techniques.

The person starts with lots of general questions about a specific topic, and with time, narrows it down to one point to arrive at the result. The funnel type of questioning techniques is generally used by people interested in research and investigation, *e.g.*, by a policeman investigating a case or a scientist researching his paper.

They ask funnel questions to collect useful information that is later arrowed down to an obvious conclusion. Examples of funnel questions are:

- When did you last see him?
- What was he wearing?
- Was he disturbed?
- What did you talk about?
- Did he say anything specific to you?

7. Clarifying Questions

As the name suggests, clarifying questions are used to verify specific information. In this type of questioning techniques, things are finalized at the end to confirm the matter that was under discussion. Examples of clarifying questions are:

- Am I right in believing that you all broke the hostel curfew?
- Just to confirm that you have asked for an unlimited package for the internet?

8. Loaded Questions

The loaded questions are often closed questions that seem pretty straightforward. There is a twist in loaded question as it includes an assumption about the other person.

This type of questioning technique is considered tricky. It is used mainly by journalists and lawyers who want to trick the other person in giving answers that they usually would not provide.

Loaded questions are considered fact-finders as they are useful in discovering facts that anyone would be reluctant to share. Examples of loaded questions are:

- Have you stopped overeating?
- Have you stopped cheating during an examination?

9. Recall and Process Questions

The recall questioning technique is used when you want the respondent to remember a particular fact. For instance, a teacher asks the student what is 5 multiplied by 5, and the student will have to recall and answer 25.

Process questions, on the other hand, encourage the respondent to recall as well as add their opinion and then answer. It tries to test their understanding and knowledge about a specific subject.

The recall and process questions are most useful in developing critical thinking in individuals. They are used during the in-depth evaluation of a topic during discussions, interviews, and tests. Examples of recall and process questions are:

- What is your account password?
- Why do you consider to yourself the right person for the job?

Building Support and Interest

Data, when analyzed, creates insight. Insight alone rarely results in action. For insight to result in action, the value of the insight needs to be defined and communicated. Action, in isolation, rarely in turn leads to sustained change. To create a reason to change, trust is needed. Trust is built through delivering and measuring the value that was promised. This trust then helps fuel more analysis which, when acted on and validated, supports even greater change. This cycle of change, when aligned to a strategic vision, helps create competitive differentiation. To create a reason to act and support ongoing change, every business analytics initiative needs to:

- Define the value, thereby building support and avoiding a lack of interest.
- Communicate the value, thereby building understanding and avoiding confusion.
- Deliver the value, thereby enabling tactical gains as well as long-term competitive differentiation.
- Measure the value, thereby supporting ongoing transformation rather than one-off successes.

They require different skills to traditional analytics. Where data scientists are excellent at extracting insight from data, “value architects” are experts at using those insights to justify change.

- **Defining the value:** Without defining the value, it is impossible to justify action. Organizations are economically driven, politically motivated entities. Every organization, regardless of how well funded or profitable it is, has a limited budget. Investments need to be prioritized to maximize return. Prioritization requires comparison. For most decision-makers, the benchmark is expected value, any outcome that has a real and measurable positive impact on an individual or the organization. In this sense, value is not an abstract concept. It needs to be realized. Recommendations or predictions might well be accurate. However, they create no value until they are acted on.
- **Communicating the value:** Knowing the value is only the start. The next step is convincing people that the value is worth the actions necessary to realize the value. Building support involves making sure the case for change is understood. Organizations are often referred to in the abstract as entities in their own right. This is misleading; organizations are pluralistic entities populated by individuals with their own views, agendas, and opinions. Truth is subjective and dependent on one’s point of view. Because of this, the case for change needs to be tailored to others’ preferences. Communicating value comes from presenting information in a way that eases understanding. By tailoring one’s message to an audience’s preferred communication style, relating it to their understanding of value, and going into an appropriate level of detail, change agents build a stronger case for action.
- **Delivering the value:** One-off benefits are valuable. And, actually delivering the value of business analytics is essential. However, these tactical benefits rarely lead to competitive differentiation. Competitive advantage comes creating and sustaining deep operational differences. This is more than just a successful project; it is about creating and delivering a vision. It requires being able to effectively link limited resources to both tactical activities and strategic gains. Every benefit needs to map into a longer-term plan that progressively moves towards true differentiation. Having this vision and building trust through delivering the value helps create and sustain ongoing change.
- **Measuring the value:** Finally, maintaining momentum and investment requires demonstrating continuous successes over time. We trust that which we believe is

“safe”. It is impossible can be measured and proven. Debate without data is nothing more than speculation. At best, further investments will be delayed until political consensus can be achieved. At worst, entire programs of work can be delayed, undermined, or outright cancelled. The fastest path to success in business analytics comes from demonstrating ongoing successes.

Obtaining Data

The HR data normally needs to come from documents, legacy systems or other means such as localized PC-based applications. The data could come from multiple sources. For example, you are implementing a new absence management HRIS module and many of the locations maintain decentralized leaves and absence details. As a result, you will need to get the data such as current leaves balances or leave transactions in current year from multiple locations. While you get this data from multiple locations, you will need to format and clean the data. Each location may be following their own format for this data and it needs to be cleaned and brought into standardized format required for a central HRIS application for absence management. Leave and absence data will have details such as type of leave that the employee has taken, duration of leave as well as in some cases reasons for the leave. The employee definitely wishes to keep such details private, and therefore, users and consultants who may handle this data will need a level of authorization as well as security while handling the data.

3.18 CLEANING DATA (EXERCISE)

This process is used to determine inaccurate, incomplete, or unreasonable data and then improving the quality through correction of detected errors and omissions.

HR data is oftentimes dirty. Dirty data is any data record that contains errors. This can happen for different reasons. The simplest one being missing data, other examples of dirty data are different labels for one and the same job functions, multiple records for the same people in one system, non-matching records in different systems, and so on. Cleaning and ordering this data can be a time-consuming process. Indeed, aggregating data from all these different data sources and making them compliant can take weeks or even months. This holds especially true for international companies. These often use different systems in different countries to record the same data.

The problem with data is that it is easy to get dirty. As soon as data collection procedures differ in the slightest, the data will become inconsistent. As a company, you can decide to clean the data at once. Some companies opt for this strategy. However, this can take tremendous amounts of time. It is hence much smarter to clean only the data you need to perform a specific analysis. This approach will prevent a lot of unnecessary work and produce results faster. Based on the outcomes of the first analysis, you can determine which extra data you need to clean to run the next analysis.

Data cleaning helps to run a smooth analysis. It also helps normal HR reporting as clean data can be fed back into the HR systems. This will help improve the data quality and is

extremely beneficial for later data analyses and data aggregation efforts. Data cleaning is thus a necessary step in the HR analytics process.

The Data Cleaning Process

When cleaning HR data, there are two things you need to understand. The first is data validity and the second is data reliability.

When data is not valid or reliable, it may tell you something different than what you are looking for. It is important to understand these two terms:

1. Validity

Validity is whether actually measuring what you need to measure. Does the appraisal system only measure individual performance, or does it (also) measure who is best liked by his/her manager? Is data collected evenly throughout the organization, or is it skewed in one way or another?

For example, the city of Boston made an app that their drivers could install on their smartphone. The app would measure bumps in the road and report their location via GPS. These bumps were then recorded and the city road service would fix them. According to a spokesperson, “the data provides the city with real-time information it uses to fix problems and plan long-term investments”.

Unfortunately, not everyone benefited equally from this system. The app was mainly used by the young and in more affluent communities. Meanwhile, the poorer communities did not have equal access to smartphones and mobile data. This is a significant bias in the data.

Questions you can ask to yourself to check for validity are:

- Does the data represent what we want to measure?
- Are there any biases in the way we measured our data?
- Was the data collected in a clear and consistent way?
- Are there outliers in the data?

2. Reliability

Reliability is about measuring the same thing over and over again and achieving the same result.

When you measure someone’s engagement in the morning, you want to have a similar result as when you measure it again in the afternoon. This is because engagement is a trait that is relatively stable over time.

The same holds true for different raters. If you ask both Bill and Jim to rate Wendy’s engagement, you want both Bill and Jim to give Wendy the same rating. However, when the scales that are used to rate Wendy are vague and open to different interpretations, Bill and Jim will likely give Wendy different ratings. This is called a rater bias and is best avoided.

This might sound obvious but it is not. Often times, reported data depends on other factors like the instructions that are given, and the mood of the person who gives the rating. This is the big question when we talk about reliability: Are the same scores achieved when the same data is measured in the same way by different people and at different times of the day/week?

Procedures play an important role in this process. In rating performance, if one manager considers an employee's performance over the last six months, while another only thinks back over the last two weeks, performance ratings will likely differ and be unreliable. Clearly documented procedures would help different managers' measure performance the same way.

Questions you should ask yourself in this context are:

- Did we consistently produce the same results when the same thing was measured multiple times?
- Did we use clearly documented data collection methods?
- Were data collection instructions followed each time?

The Data Cleaning Checklist

The previous questions on validity and reliability help you to analyze whether your input data is sufficiently accurate to yield reliable and valid results. There are several other criteria for data needs to comply with. For example, the data needs to be up-to-date.

Data that is outdated will produce potentially irrelevant results and can potentially mess up your results. Additionally, you need to check if you have all the relevant data: records are oftentimes missing. Depending on how to analyze your data, this may or may not cause problems. Some methods of analysis allow for missing data while other algorithms struggle when data is missing.

Missing data will narrow your population. Plus, there is a real chance that there are shared similarities between the people whose data is missing. For instance, if one department still uses an outdated performance management system that omits certain questions, it would mean that you'd lack data of all the people working in that department. This can seriously skew your results towards the other departments and threaten the generalizability of the results.

This is a practical checklist with six steps for data cleaning:

1. **Check if the data is up-to-date.**
2. **Check for recurring unique identifiers.** Some people hold more than one positions. Systems often create separate records for each position. These people thus end up having multiple records in a single database. Depending on the situation, these records may be condensed.
3. **Check data labels across multiple fields and merged datasets and see if all the data matches.**
4. **Count missing values.** When missing values are over-represented in specific parts of the organization, they may skew your results. We saw this in the previous example. In addition, an analysis with too many missing values (*i.e.*, insufficient data) runs the risk of becoming inaccurate. This also impacts the generalizability of the results.

5. **Check for numerical outliers.** Calculate the descriptive statistics and the values of the quartiles. These enable you to calculate potential outliers. The minimum and maximum values are a good starting of the point.

In addition, you can calculate the interquartile range. You can do this by multiplying the difference between quartile 3 (Q3) and Q1 by 1.5. The result can be added to Q3 and subtracted from Q1. Values outside this range are assumed to be outliers.

6. **Define valid data output and remove all invalid data values.** This is useful for all data. Character data is clearly defined. For example, gender is defined by M or F. These are the valid data values. Any other values are presumed to be invalid. This data can be easily flagged for inspection.

Numeric data is often limited in range (*e.g.*, working age is between 15 and 100). Numeric data that falls outside the predefined range can be flagged the same way.

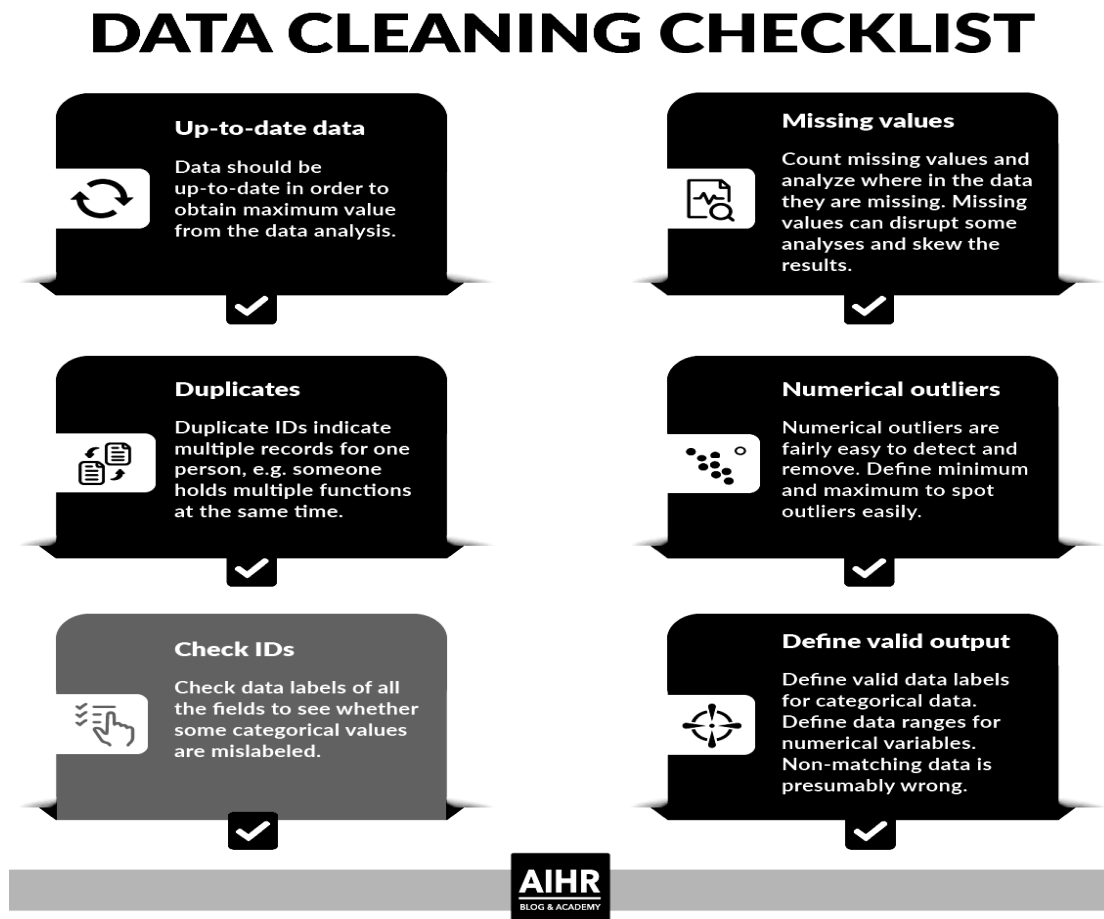


Fig. 3.25: Data Cleaning Checklist

Supplementing Data

To achieve the organizational objectives efficiently and effectively, it is important that management at different levels in an organization has access to appropriate information, including human resource information to enable it to make informed decisions. Ideally, government objectives should be driving an organization's strategic directions, including the human resource initiatives adopted by the organization. This, in turn, should be determining the human resource data that is collected, analyzed and reported. The organizations should report human resource data monthly, quarterly or six monthly to executive management committees as part of a consolidated package of management information. However, although there may be some exceptional data such as staff commencements and separations, staff numbers by divisions/regions, and leave balances. Sometimes, this is supplemented with additional data such as various types of leave taken.

To minimize the data burden on state officials, the study obtained most of the demographic data for such variables as the concentration of low-income families and local unemployment rates from other sources and merged this information with the data supplied by the states. Specifically, the study used data from the following sources:

1. The elementary and secondary general information survey 1978-79, conducted by the National Center for Education Statistics: information on the size of LEAs (ADM and ADA), county and SMSA codes, secondary enrolment, racial, and ethnic 'data'.
2. The Higher Education General Information Survey (HEGIS), 1973-79, conducted by the National Center of Education Statistics: information rotation enrolments, finances, county and SMSA codes, racial and ethnic data.

3.19 BASICS OF HR ANALYTICS

Over the last few years, companies have been increasing their focus on HR analytics. Whether this is due to the need you have a better handle on where talent is within the company or to prove that the HR organization adds value to the corporation, HR organization are forming analytics groups that focus entirely on identifying through structured, deep analysis-opportunities for plan design, process, or practice changes that lead to service improvements and/or cost savings.

HR analytics is the go-to approach whenever leaders need accurate statistics or fact-based predictions in order to make better business decisions. In a growing number of organizations, this analytic approach is beginning to permeate every aspect of the HR function.

By establishing systems and methodologies for understanding and linking measures on the people side of the business to key performance indicators, the potential for identifying how HR policies, procedures, systems, and interventions drive organizational performance is greatly enhanced. But at the same time, this process may reveal deficiencies in an organization's current HR strategies and programmes.

What is Analytics?

It is concerned with turning raw data into insight for making better decisions. Analytics relies on the application of statistics, computer programming, and operations research in order to quantify and gain insight to the meanings of data. It is especially useful in areas which record a lot of data or information.

Association of networking continues its explosive growth; the percentage of unstructured data will necessarily expand. In practice, structured and unstructured data can be merged into a mixture, amalgamation, or fusion. In short, it will be what some now call hybrid data. While hybrid data will be essential for future analysis, it will also make the process much more complicated.

In recent times, analytics has gained importance due to four important factors: Analytics can be fragmented into four basic types, based on the business intelligence it provides:

1. Availability of large amounts of data in various forms, both structured and unstructured.
2. Easier, cheaper storage and retrieval of this data.
3. A need to take decisions at a faster pace to stay competitive in the market.
4. A need to get the most out of every dollar spent.

Analytics is divided into three levels:

1. **Descriptive:** Traditional HR metrics are largely efficiency metrics (turnover rate, time to fill, cost of hire, number hired and trained, etc.). The primary focus here is on cost reduction and process improvement. Descriptive HR analytics reveal and describe relationships, and current and historical data patterns. This is the foundation of your analytics effort. It includes dashboards and scorecards, workforce segmentation, data mining for basic patterns, and periodic reports.
2. **Predictive:** Predictive analytics covers a variety of techniques (statistics, modeling and data mining) that use current and historical facts to make predictions about the future. It is about probabilities and potential impact. It involves models used for increasing the probability of selecting the right people to hire, train, and promote.
3. **Prescriptive:** Prescriptive analytics goes beyond predictions and outlines decision options and workforce optimization. It is used to analyze complex data to predict outcomes, provide decision options, and show alternative business impacts. It involves models used for understanding how alternative learning investments impact the bottom line (rare in HR).

The process starts with the simple reporting of HR metrics and goes all the way up to prescriptive modeling of business practices. Although financial capital (cash) and economic capital (intangible assets) are the lifeblood of a business, it is human capital (people) that apply cash and leverage intangible assets to drive business performance.

3.20 EVOLUTION OF HR ANALYTICS

The evolution of HR analytics has come a long way since it was introduced and organizations today have the ability to track all kinds of metrics to ensure that their people analytics, talent analytics, and workforce analytics are accurate.

From gut-feel to optimization, we have come a long way indeed. As we go up the value chain, we need to train ourselves and the organization to which we belong, in the statistical techniques, analytical tools, and assimilation of the business scenario and interpretation of the same.

- Initially standard static reports and charts to know what happened.
- Dynamism (ad hoc) was brought in to these reports to know how much, how frequently or where this happened.
- OLAP (online analytical processing) tools were used to slice and dice the information in any way which we wanted and used alerts to inform us when something went wrong, so that corrective action could be taken or necessary escalations made.
- The data was subject to statistical analysis, to know the root cause of the problem, in order to prevent it from occurring again.
- Next, forecasting techniques were used to study the trends in data because there were seasonal variations and repetitive happenings.
- Predictive models based on statistical algorithms using probabilities for the happening of an event or otherwise were adopted.
- Operations research (OR) technique of optimization was used to get the best possible outcomes.

Analytical Capabilities

Data can be viewed in two ways — structured and unstructured. Structure data is similar to financial data, and unstructured data typically economic or less tangible data. As social networking continues its explosive growth, the percentage of unstructured data will necessarily expand. In practice, structured and unstructured data can be merged into a mixture, amalgam, or fusion. In short, it will be what some now call hybrid data. While hybrid data will be essential for future analysis, it will also make the process much more complicated. This is precisely why analysis is essential.

When a situation is a complex mixture of objective facts and subjective beliefs, there is way other than through logical inquiry and statistical treatment for us to comprehend what is not readily apparent. Excuse the effort for making this point again, but analysis has its genesis in descriptive data that tells us what has happened up to the present. But prediction and prescription are concerned with what can or preferably could happen and how to make it happen. All three levels are necessary. Throughout the twentieth century, we took descriptive data and mentally extrapolated it into the future through presumed trends. This was somewhat acceptable when the market was stable. However, today and tomorrow, the market are anything but stable.

Therefore, prediction is vitally necessary for any company that desires to sustain or grow market share and profitability in the new millennium. Not-for-profit organizations need it even more, because so much of their “business” involves subjective data such as a humanistic mission and the satisfaction of its constituency. In the HR or human capital realm, change is just as constant as it is in financial, production, or marketing arenas. Labour availability and costs, skill development, leadership, engagement, and retention issues constantly evolve as business needs and market dynamics change.

Last year’s recruitment programme can be rendered ineffective overnight by market dynamics, competitors’ actions, new technology, and capricious customers. In the first decade of the twenty-first century, the dot-com crash and the liquidity crisis totally disrupted the labour model. Yesterday’s skill requirements are made obsolete by continuing advances in technology as well as new customer demands or changes in government regulations. As the economy improves, acquisition and retention of mission-critical skills become a problem. These and other human capital management dynamics demand that we apply analytics to reconfigure our labour scenario and predict out next best move. If we wait to see what the competition is doing, we lose competitive advantage and market share.

3.21 ANALYTIC VALUE CHAIN

A journey to optimising value from data. Analytics Value Chain is the key to timely outcomes in The Internet of Things (IoT).

With the tremendous volume of data in the IoT, Analytics offers great potential to create new business value. Analytics must be executed in real time across the Analytics Value Chain (streaming, historical, predictive, and prescriptive analytics) with relevant contextual and situational data. This capability paired with the next best action creates the greatest value.

- **Collect:** Connect and integrate data sources from inside and outside the organization. Merge history with real-time updates to provide actionable insights, understood within the context of your business.
- **Prepare:** Aligning all data involved in decision-making against common metrics and hierarchies is the beginning of effective organizational decision-making.
- **Visualize:** Visualization tools integrated into existing workflows and applications can improve efficiency and engage teams in process improvement.
- **Describe:** Report and analyze results within the context of the business to tell the story of what’s happening. Share evolving business trends with colleagues to align on actionable insights.
- **Diagnose:** Understanding root cause can mean a huge difference to develop data-driven decisions as they are empowered through strong data platforms. It becomes easy to discover the missed opportunities and spot potential issues early.

- **Predict:** Forecasting can be best described as leveraging existing data to establish models that can help business users to accurately plan for potential outcome and move ahead with the right decision-making in any organization.
- **Prescribe:** Data-driven decision-making is its ability to leverage data, machine learning and artificial intelligence to provide a competitive advantage. The organizations employ advanced models and methodology to define and optimize appropriate courses of action and ensure the best possible outcome.

3.22 ANALYTICAL MODEL

Analytical model defines structure to analytics, *i.e.*, more than simply running a statistical analysis. The first step to analysing any problem is to ask questions. The five-step process of data analytic is as follows:

1. **Organizing:** The steps start with collecting and organizing human capital data which is often a major problem. HR department should work principally with employee records in the beginning.
2. **Displaying:** After collecting basic data, the companies have to develop dashboards that format data for various internal customers. The dashboards are capable of reporting degree of performance that are colour coded – red, yellow and green. This is an extension of descriptive data which shows the current state and trends up to this point. It is the basis for reporting the past. It does not speak to the future except through vague, unsubstantiated inferences. The trends are not extendable because underlying conditions and future assumptions are not included. The primary value of dashboards is that perceptive and creative users can see possibilities for attempting hoped improvements.
3. **Relating:** At this level, the focus turned outward from the data that has been hacked to its relationships to other data and phenomena. The most common practice today is to benchmark your data to that of other companies. While benchmarking with other companies, one has to be doubly sure that it is similar to the benchmarking company.
4. **Modeling:** At this stage, a model has to be developed depending on the change needed. The model should predict certain patterns or relationships that interact to obtain desired outcome.
5. **Evaluating:** The model developed will probably connect people, policies, products and processes to achieve some improvement in performance. At this stage, the HR department should be able to monitor or measure the amount or degree of change attained. In addition, the model gives a new routine that should sustain or continue to improve outcomes.

3.23 PRACTICAL APPLICATION OF HR ANALYTICS

The HR analytics examples are slowly edging towards practicality and it finds its use in below mentioned areas:

1. Employee Retention

According to the stats from the US employers, the average employee replacement cost is 200% of the annual salary they draw.

When an employee leaves the organization, the cost of onboarding cost of the recruitment process and lost productivity adds to the total loss to the organization.

It becomes critical for the HR department to contain the attrition rate and it is achievable only when the HR department adapts to a data-driven approach.

Some of the aspects that drive impact iteration analysis are:

- Churn rates
- Per department attrition rates
- Onboarding experience
- Employee interview data
- Employee performance data

HR analytics helps identify the reasons behind attrition, and develop policies and training programs to dampen the impact of attrition.

2. Employee Performance

Around 45% of HR professionals believe that performance reviews are not accurate which makes it difficult to retain the talented employees.

Thus, accurate performance evaluation is critical for retention. HR analytics tools are intelligent and leverage the employee data to identify the key players based on multiple performance parameters.

As the performance measurement and career progression are interdependent, an accurate data mining of the two could help HR professionals identify the employee expectations.

For larger organizations, HR leaders can analyze the promotion expectation and average promotion wait time to determine how the employees can be motivated to perform better and stay.

3. Employee Recruitment

The talent shortage is the biggest nightmare for enterprises. Almost 42% of the employers recruiting today are worried about the selection of not-so-appropriate candidate.

The HR recruitment team is primarily responsible for finding out the right CVs from the pool, but before that they need to develop a powerful ideal candidate portfolio.

This is where what is HR analytics is answered and its role is identified. The data collected from hiring managers and the performance data of previous hires for the same role are fed to the HR analytics tools to create an optimized skill set which is desired.

The data which is considered is:

- Identify the average number of the applicant after analyzing the applicant pool.
- Number of interview rounds.
- Offer acceptance statistics.

4. Employee Development

HR plays a critical in employee development as a skill gap always exists with new recruitment.

According to American Employers, 40% of the recruited resource is not ideal for the job but with HR's employee development programs, the skill gap can be covered substantially.

The HR Analytics tools help human resource management assess the skill needs, train the employees accordingly and allocate the right resources to the teams. This increases the agility of the organization as well as enhances employee satisfaction.

Employee development programs are running currently also, but digging out the right requirement is still a challenge. In light of HR analytics, more refined employee development programs can be led.

5. Employee Engagement

Attracting the best talent to the organization is an art and HR strive hard to achieve it. Having appropriate employee engagement is critical for an organization to attract and retain the employees.

As critical as it is to identify the factors that drive employee engagement, it is equally difficult to find the right metrics.

The HR needs to do statistical analysis of employee engagement data surveys to identify the data which leads to better employee engagement.

6. Developing Compensation Programs

Employee compensation is known to contribute 33% to employee retention and performance. As it still remains the biggest investment by the business expense, its appropriation and accuracy needs go without saying.

Both internal and external factors impact the compensation plans. That is why, this area requires more precise automation.

The HR needs to analyze what the competitors are offering their resources and what kind of compensation is inducing higher retention.

Predictive Analytics

Predictive analytics is a branch under advanced analytics primarily used to make predictions about the uncertain future events. Predictive analytics makes use of statistics, modeling, data mining, artificial intelligence, and machine language to work on the current set of data provided as instructions and predict the future events.

Through the usage of the historical data and transactional data patterns, risks are identified in addition to exploring opportunities for future needs. Through the usage of predictive analysis, businesses can effectively interpret the data available to them for their benefits.

Data mining and text analytics along with statistics enables business users to create a predictive intelligence by unravelling relationships among the data patterns both in the structured and unstructured data. Certain data can be used readily for commencing the analysis like the age, gender, income, sales, etc. These data are referred to as structured data. In addition to the structured data, we also have information available around us like the social media, website information and any other information from which data could be picked and then structured for further analysis. This is called unstructured data. The data collected from such activities are then extracted and used in model building process.

Through predictive analytics, organizations tend to get more proactive and forward looking in terms of planning and apt decision-making rather than going by emotions or a gutsy feeling!

Process Involved in Predictive Analytics

1. **Project Definition:** Identify what shall be the outcome of the project, the deliverables, business objectives, and based on that, go towards gathering those datasets that are to be used.
2. **Data Collection:** This is more of the Big Basket where all data from various sources are binned for usage. This gives a picture about the various customer interactions as a single view item.
3. **Analysis:** Here, the data is inspected, cleansed, transformed and modelled to discover if it really provides useful information and arriving at conclusion ultimately.
4. **Statistics:** This enables to validate if the findings, assumptions and hypothesis are fine to go ahead with and test them using statistical model.
5. **Modeling:** Through this, accurate predictive models about the future can be provided. From the options available, the best option could be chosen as the required solution with multimodal evaluation.
6. **Deployment:** Through the predicative model deployment, an option is created to deploy the analytics results into everyday effective decision. In this way, the results, reports and other metrics can be taken based on modeling.
7. **Monitoring:** Models are monitored to control and check for performance conformance to ensure that the desired results are obtained as expected.

Applications of Predictive Analytics

1. **CRM:** Through predictive analytics, marketing campaigns, sales and customer services are objectively achieved. This can be used in analytical customer relationship management throughout the customer life cycle right from the acquisition, relationship growth, retention and customer win back can be better planned and strategically addressed for retaining customers and addressing them more clearly.
2. **Healthcare:** Usage of predictive analytics in the healthcare domain can aid to determine and prevent cases and risks of those developing certain health-related complications like diabetics, asthma and other life -threatening ailments. Through the administering of predictive analytics in healthcare, better clinical decisions can be made.
3. **Collection Analytics:** These applications optimize the allocation of collection resources by identifying collection agencies, contact strategies to reach out to them, legal actions to increase recovery and cost reduction of collection.
4. **Cross-sell:** Through predictive analytics, applications attached to various touchpoints connected to the customers, a detailed analysis on the customer spends, usage pattern of certain purchases they make regularly, customer behaviour can be obtained which ultimately leads to efficient cross sales or selling additional products to customers. In this way, organizations dealing with multiple products can effectively increase its sales volume and profits ultimately.
5. **Fraud Detection:** Predictive Analytics can aid to spot inaccurate credit application, deviant transactions leading to frauds both online and offline, identity thefts and false insurance claims saving financial and insurance institutions of lots of security issues and damages to their operations.
6. **Risk Management:** The best portfolio prediction to maximize returns on the capital invested, probabilistic risk assessment to yield accurate forecasts are some of the important benefits of using predictive analytics.
7. **Direct Marketing:** Predictive Analytics also aids in identifying the most effective combination of product versions, marketing material, communication channels and the timing to be used to target a given consumer in the current environment where the dynamics are constantly changing and gets challenging for a business to compete and run successfully.
8. **Underwriting:** Perhaps, one of the biggest benefits that can infiltrate into underwriting is providing information about the likelihood of illness, default of loan/insurance and bankruptcy. Predictive Analytics streamline the process of customer acquisition by closely predicting the future risk behaviour of a customer through the application data.

Determine Key Performance Indicator

A Key Performance Indicator is a measurable value that demonstrates how effectively a company is achieving key business objectives. Organizations use KPIs at multiple levels to evaluate their success at reaching targets. High-level KPIs may focus on the overall performance of the business, while low-level KPIs may focus on processes in departments such as sales, marketing, HR, support and others.

Using KPIs as Part of Your Performance Management Frameworks

The most common elements between most performance management frameworks are setting objectives, measuring performance, and managing all related activities.

According to classic old adage, Goodhart's Law, "any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes."

Charles Goodhart was an economist in 1975 whose research was used in helping criticize government decision-making processes, specifically with regards to monetary policy. This concept was then made mainstream by **Marilyn Strathern**, "when a measure becomes a target, it ceases to be a good measure."

A performance indicator or key performance indicator is just one type of performance measurement. There are many performance management frameworks that are both similar yet different. Each of these frameworks bring forward elements that can be pulled together to help drive success backed by data. Let's dig in.

Step 1: Aligning Business Strategy

A popular theme in start-ups these days is the **One Metric That Matters (OMTM)**. The key takeaway from this simple, yet extremely powerful tool is that you have a thorough understanding of your business model in order to hone in on that metrics and get the entire organization aligned.

Many will argue that sales is the most important metric when it comes to measuring the success of a business. The challenge with this metric is the measured outcome.

Ask yourself: What is the one metric that would help drive more sales?

One answer to this question could be tracking the number of customers who have integrated your product with three other applications. This measure would be indicative of level of engagement, and their probability of churning would likely be reduced.

The reason being that once customers are locked in, they churn less which then creates the right unit economics for the company to grow. So, in this case, instead of looking at sales numbers, we would only count a customer, if and only if, they are connected with three apps.

This is merely an example, and doesn't mean that there is only one metric you should care about! This framework helps with keeping everyone focused on the one thing they should care about most.

Step 2: Cover All of Your Bases

With business comes trade-offs.

You have probably heard the saying, “You have a cheap, good, or fast. But to only pick 2”.

Let’s start with a classic framework that helps to navigate these trade-offs. The Balanced Scorecard (BSC) helps to break down the key areas of your business (perspectives) where activities need to be monitored.

The four perspectives that need to be in balance are:

1. Financial Perspective,
2. Customer Perspective,
3. Internal Business Process Perspective, and
4. Learning and Growth Perspective.

These four key areas of your business are intertwined and all must be aligned. When one is impacted, there is impact on another. In other words, there will be a trade-off.

Step 3: Putting Your BSC Strategy Framework into Action with OKRs

The Balanced Scorecard (BSC) strategy suggests that for each perspective to develop objectives, measures (KPIs), set targets (goals), and initiatives (actions). A more recent framework that is getting popularized is the **OKR Framework**. Popularized by its use at Google, the OKR (objectives and key results) framework is used to define and track objectives and their outcomes. Many would argue that this framework sits in-between a KPI strategy and the Balanced Scorecard approach.

OKRs are used as a performance tool that sets, communicates, and monitors goals in an organization so that all employees are focused in the same direction. The system encourages employee success through clear work objectives and desired key results. The beauty of the system is that it provides a simple, practical, and straightforward framework for defining, tracking, and measuring goals, both as something to aspire to and as something that can be measured.

Step 4: Monitoring with a KPI Dashboard

A **KPI dashboard** provides you with an at-a-glance view of your business performance in real-time so that you can get a better picture on how the entire organization is doing.

Common terms found in these frameworks that are worth understanding include:

- **Key risk indicator (KRI):** It is a measure used in management to indicate how risky an activity is. Key risk indicators are metrics monitored by organizations to provide an early warning of increasing risk exposures in various areas of the business.
- **Critical success factor (CSF):** It is a management term for an element that is necessary for an organization to achieve its mission. Critical success factors should not be confused with success criteria. Success criteria are most commonly used in project

management to determine if the project was a success or not. Success criteria are defined with the objectives and can be quantified by using KPIs.

- **Performance metrics:** It measures an organization's behaviour, activities, and performance at the individual level and not organizational level. For example, an individual who works in a call centre may have performance metrics such as number of calls answered, average wait time, number of successful calls processed, and average length of call.

Creating good KPIs for your organization is an iterative process.

10 Criteria to Consider When Designing Key Performance Measures

Consider this list of criteria when building out your key business performance measurement systems:

1. Be based on quantities that can be influenced, or controlled, by the user alone or in cooperation with others
2. Be objective and not based on opinion
3. Be derived from strategy and focus on improvement
4. Be clearly defined and simple to understand
5. Be relevant with an explicit purpose
6. Be consistent (in that they maintain their significance as time goes by)
7. Be specific and relate to specific goals/targets
8. Be precise – be exact about what is being measured
9. Provide timely and accurate feedback
10. Reflect the “business process”, *i.e.*, both the supplier and customer should be involved in the definition of the measure

3.24 KEY PERFORMANCE INDICATORS IN ACTION

KPI Report

Whether you share a KPI report daily, weekly, monthly, quarterly, annually or all of the above, setting up a good KPI report platform is key to your success. At Klipfolio, we monitor a few KPIs but then track more deeply all the measures and activities that can affect that KPI.

For example, if we track Monthly Recurring Revenue (MRR), we know that # of quality leads, # of trials started, # of successful onboards and many other measures will impact the success of MRR. So, we track a daily number of new leads created with an email report every morning at 8 a.m. We have a dashboard to track several key activities to ensure the product trial starts are going smoothly in real time and we track monthly the number of onboards completed successfully by the customer success team.

KPI Dashboard

With **KPI dashboards** becoming more and more prevalent in today's fast moving organizations such as SaaS and cloud-based businesses, they usually represent a consuming format where an individual can review their data in real-time whereas reports tend to be specific snapshots in a moment of time.

One of the most common uses cases of KPI dashboard tools are in start-ups who share their core organizational performance measures to get alignment from all the employees. When you walk around their offices, TVs will be placed near specific teams highlighting the results in real time such as number of support tickets resolved today or number of new wins.

Analyzing and Reporting Data

Once the dataset is structured into a usable file format, the process of data analysis begins. Statistical analysis is generally divided into two types:

1. **Descriptive Statistics:** Descriptive statistics describe the data by using statistical terms that have become common in day-to-day language, such as the number of responses, the mean or average, the standard deviation, or frequency distribution. Descriptive analysis is necessary to understand the data.
2. **Inferential Statistics:** Inferential statistics search for relationships among variables using techniques like correlation and regression. Inferential statistics also test for differences between groups using t-tests and analysis of variance, among other techniques. Based on the results of the statistical test, the analyst infers that the relationship (or difference) is true for the cases in the sample but also generalizable to the entire population. Predictive analytics is an extension of inferential statistics because inferential techniques are used to predict future values

The **table below** shows a basic analytics plan for each of the measures in our example. The majority of the work outlined in this table focuses on descriptive statistics – computing N counts, averages, or 9-box ratings for each business unit or level. These results are necessary to describe the current state of the organization using the KPIs.

<i>Measures</i>	<i>Descriptive</i>	<i>Display – Graphs for Dashboards</i>	<i>Inferential</i>
Efficiency			
Number of open requisitions	N count of open requisition by business unit or level	Line chart by business unit or level	ANOVA – compare across business units or levels
Positions filled per months	N count of filled positions per month by business unit or level	Line chart by business unit or level	ANOVA – compare across business units or levels
Time to fill open positions	Average amount of time in days; also display outliers high and low	Bar chart of average time to fill by business unit or level	ANOVA – compare across business units or levels
Salary associated with	Average salary (possibly standard deviation) by	Box and whisker chart displaying	ANOVA – compare across

positions	business unit and level	average salary and standard deviation or quartiles	business units or levels
Cost to hire the new resource	Average cost to hire (possibly standard deviation) by business unit and level	Box and whisker chart displaying average cost and standard deviation or quartiles	ANOVA – compare across business units or levels
Effectiveness			
Performance rating at 90 days and 36 days	Percentage within each of the 9 box	9-box (3 × 3) display	ANOVA – compare across business units or levels
Identification of high potentials	Percentage of 6-9 ratings	Bar chart by business unit or level	ANOVA – compare across business units or levels
Assessment results	Average of assessment scores by competency; highlight strengths and gaps	Bar chart by business unit or level	ANOVA – compare across business units or levels
Speed to competency	Average amount of time to demonstrate competence	Bar chart by business unit or level	ANOVA – compare across business units or levels
Sponsor satisfaction – leader input	Average response to each survey question	Bar chart by leader	ANOVA – compare across leader
Exit survey results	Average response to each survey question	Bar chart by business unit or level	ANOVA – compare across business units or levels
Outcomes			
Engagement survey results	Average response to each survey question	Bar chart by leader or business unit	ANOVA – compare across business units or levels
Productivity	Average number of chargeable hours (or widgets produced)	Bar chart by leader or business unit	ANOVA – compare across leaders or business units
Retention/turnover within 90 or 365 days	Average turnover percentage at 90 and 365 days	Bar chart by business unit or level	ANOVA – compare across business units or levels
Profitability	Average Revenue	Bar chart by business unit or level	ANOVA – compare across business units or levels.

Inferential statistics are also noted in the table. For all cases, the analytic technique is analysis of variance (ANOVA). It is used to compare the average across groups, such as business units or levels. While statistical tests are not always necessary, they can be useful. When the results are shared with stakeholders, someone will likely ask, “Is that difference among groups statistically significant?” If the ANOVA test has been completed, that question can be answered yes or no with authority. It is also worth noting that practical significance is also important. A statistically significant difference may not be meaningful.

Interpreting the Results

Statistical techniques provide valuable results, but they require interpretation before they become useful.

Correlation

Correlation analysis provides good insights into the relationships among the key variables in the dataset. Before we examine the relationships, it is important to formulate a hypothesis – at least to establish in our own minds our expectations about relationships among our measures.

We can expect that several metrics will be closely related to each other, such as the performance ratings at 90 and 365 days. Additionally, the new hire's salary and the cost to hire should be related because the salary is a component of the cost to hire. We might also expect that high potential status, the competency assessment score, and speed to competency will be similar. The key concern is this — how do these measures relate to individual productivity and in turn profitability? The Table below shows the relationships among variables using the statistic *r*. The values range from –1.00 to +1.00. The strongest relationships are closest to +1.00 or –1.00.

KPI	Time to fill	Salary	Cost to hire	Performance rating 90 days	Performance rating 365 days	Assessment results	Speed to competency	Sponsor satisfaction	Exit survey rating	Engagement survey	Productivity	Profitability
Time to fill	1	.052	.065	.633	.556	.398	–.326	.201	.587	.140	.537	.386
Salary	.052	1	.999	–.170	–.046	.105	–.024	.070	–.568	.029	–.006	.514
Cost to hire	.065	.999	1	–.168	–.046	.100	–.022	.062	–.569	.024	–.002	.512
Performance rating 90 days	.633	–.170	–.168	1	.915	.695	–.746	.449	.851	.341	.848	.468
Performance rating 365 days	.556	–.046	–.046	.915	1	.714	–.686	–.686	.838	.334	.864	.548
Assessment results	.398	.105	.100	.695	.714	1	–.749	.726	.809	.492	.747	.585
Speed to competency (days)	–.326	–.024	–.022	–.746	–.686	–.749	1	–.555	–.865	–.264	–.750	–.503
Sponsor satisfaction	.201	.070	.062	.449	.492	.726	–.555	1	.732	.339	.572	.467
Exit survey rating	.587	–.568	–.569	.851	.838	.809	–.865	.732	1	.525	.799	.714
Engagement survey rating	.140	.029	.024	.341	.334	.492	–.264	.339	.525	1	.316	.207
Productivity	.537	–.006	–.002	.848	.864	.747	–.750	.572	.799	.316	1	.617
Profitability	.386	.514	.512	.468	.548	.585	–.503	.467	.714	.207	.617	1

Italicized values in this figure are significantly correlated. For example, the relationship between cost to hire (third row) and salary (second column) is .999, almost perfect. The correlation between time to fill (first row) and salary (second column) is positive, slight (close to zero) and not significant. Last, the two values that are underlined are nearly significant, close to the .05 significance level.

A diagonal line from the top left to the bottom right of the figure shows a value of 1 at the intersection of each measure with itself. Each measure is perfectly correlated with itself. The triangle of values below this diagonal line of 1s mirrors the triangle of values above the diagonal.

Here is a simple interpretation of the values in the figure:

1. **Time to fill:** It is correlated with all measures except salary, cost to hire, and the engagement survey rating. Notably, it is negatively correlated with speed to competency. As time to fill increases, the speed to competency decreases.
2. **Salary:** It is correlated with only three variables – cost to hire, exit survey rating, and profitability. The cost to hire metric and profitability are algebraic variations of salary. So, we expect these to be related. Surprisingly, salary is negatively correlated with the exit survey rating.
3. **Cost to hire:** It is correlated with only two variables – the exit survey rating (negatively) and profitability. Cost to hire is a function of salary. Logically, it does not have much influence on exit survey ratings. It is related to profitability because costs impact profitability, and for this reason, cost has a significant correlation. Profitability increases as salary decreases.
4. **Performance rating at 90 days:** These are related to all variables except salary and cost to hire. Speed to competency is negatively correlated. As speed decreases (fewer days required demonstrating competency), performance ratings increase.
5. **Performance rating at 365 days:** These are related to all variables except salary and cost to hire. Speed to competency is negatively correlated. As speed decreases (fewer days required to demonstrate competency), performance rating increase.
6. **Speed to competency:** It is correlated with all variables except salary and cost to hire; all correlations are negative, *i.e.*, faster speed to competency (fewer days required to demonstrate competence) is correlated better performance.
7. **Sponsor satisfaction:** It is correlated with all variables except salary and cost to hire. It correlates negatively with speed to competency.
8. **Exit survey rating:** It is correlated with all ratings except the engagement survey rating. Notably, the correlation value is moderate at .525, which means there is a relationship here. An examination of the significance test shows the p-value to be .054, meaning it missed being included as a statistically significant relationship by four thousandths of a point. Practically, we can consider this a meaningful relationship – positive engagement ratings often yield positive exit survey scores.
9. **Engagement survey rating:** These are correlated with all items except time to fill, salary, cost to hire, and the exit survey rating.
10. **Productivity:** It is correlated with all other measures except salary and cost of hire. The correlation with speed to competency is negative.

- 11. Profitability:** It is correlated with all measures and negatively related to speed to competency.

The correlation analysis describes which measures are related. But it does not always offer clarity. A lot of information is included in the list, and it is not easy to discern which relationships are most important. This is clearly a drawback of correlation analysis, especially when many measures are involved.

Predicting the Future

Assume for a moment that HR wants an early indicator of the quality of the candidates who have been hired. Using what we know from the regression analysis of productivity, we see that two early indicators are – speed to competency and time to hire. At 90 days of post hire, we will not have the other two measures, performance rating at 365 days and sponsor satisfaction. So, we should not consider them as predictors. Let's modify the prediction equation by dropping the performance rating and sponsor satisfaction metrics. When we rebuild the equation, it looks like this:

$$\text{Productivity} = 83.23 - .30 (\text{speed to competency}) + .42 (\text{time to fill})$$

Using the prediction equation and the data we have for five new hires, we can predict their future productivity. Table below shows the values for five new hires at 90 days post hire.

<i>New Hire</i>	<i>Speed to Competency</i>	<i>Time to Fill</i>	<i>Predicted Productivity</i>
A	50 days	67 days	96.37
B	45 days	50 days	90.73
C	66 days	55 days	86.53
D	Not declared competent	12 days	No value
E	Not declared competent	28 days	No value

New hire A to C look promising. In fact, they should be billable 86% of the time or more, based on our prediction. New hires D and E have no predicted score because they have not been declared competent at 90 days. This in itself should be a good indicator that these two employees will not be stellar performance and provide development opportunities or terminate employment.

Table extends our predictive capabilities to show the estimated profitability of the three high-performing candidates.

<i>New Hire</i>	<i>Speed to Competency</i>	<i>Time to Fill</i>	<i>Predicted Productivity</i>	<i>Predicted Profitability</i>
A	50 days	67 days	96.37	\$38,446
B	45 days	50 days	90.73	\$34,543
C	66 days	55 days	86.53	\$31,636

The predicted profitability for each new hire exceeds \$30,000. Since salary and benefits are included in the profitability metric, these values represent revenue that goes directly to the bottom line.

Throughout this analysis, we have focused on all cases across all business units. The same analysis can be repeated within each business unit to determine if relationships are different across units. In this case, the relationships are not substantially different.

You may wonder why you should go through all of these efforts to demonstrate something that is generally common knowledge – that speed to competence is a good indicator of productivity. The value comes from the ability to predict performance and manage employees as soon as possible. New hires D and E may need some performance support from training, coaching, job aids, or other means. Alternatively, they may need to be terminated. Such early intervention can improve performance, increase productivity, save a career, and continue to improve organizational performance.

3.25 HR ANALYTICS TOOLS

Metrics

Metrics are measures of quantitative assessment commonly used for assessing, comparing, and tracking performance or production. Generally, a group of metrics will typically be used to build a dashboard that management or analysts review on a regular basis to maintain performance assessments, opinions, and business strategies.

Executives use them to analyze corporate finance and operational strategies. Analysts use them to form opinions and investment recommendations. Portfolio managers use metrics to guide their investing portfolios. Furthermore, project managers also find them essential in leading and managing strategic projects of all kinds.

Overall, metrics refer to a wide variety of data points generated from a multitude of methods. Best practices across industries have created a common set of comprehensive metrics used in ongoing evaluations. However, individual cases and scenarios typically guide the choice of metrics used.

Types of Metrics

Metrics, in general, are of three types:

1. **Efficiency metrics:** Efficiency metrics measure performance of the HR department, *e.g.*, performing basic administrative jobs.
2. **Effectiveness metrics:** Effectiveness metrics measure the effect of HR programmes on employees, *e.g.*, measuring skill development through trainings instead of number of trainings attended.
3. **Impact metrics:** Impact metrics measure to what extent HR can provide a sustainable competitive advantage, *e.g.*, defect reduction after relevant training.

On the other hand, Mayo had proposed seven metrics for HR analytics:

1. Workforce statistics.
2. Financial ratios relating to people and productivity.

3. Measures of people's values.
4. Measures of people's engagement.
5. Measures of efficiency of the HR function.
6. Measures of effectiveness of people processes.
7. Measures of investment in one-off initiatives and programmes.

After data is collected using the above-mentioned metrics, statistical tools and techniques can be used to show the casual relationship between HR practices and organizational performance metrics like profitability, customer satisfaction, sales, etc. which forms the basis of HR analytics.

Metrics Used for HR Analytics

The metrics used for human resource analytics are:

1. Monthly turnover rate = $(\text{Number of separations during month} / \text{Average number of employees during month}) \times 100$.
2. Revenue per employee = $\text{Total revenue} / \text{Total number of employees}$.
3. Yield ratio = Percentage of applicants for a recruitment source that make it to a determined stage of the application process.
4. Human capital cost = $\text{Pay} + \text{Benefits} + \text{Contingent labour cost} / \text{full-time equivalents}$.
5. HR to staff ratio = $\text{Employees} / \text{Human resource team members}$.
6. Cost to hire = $\text{Training} + \text{Advertising} + \text{Lost productivity} + \text{Time/employee salary}$ (ranges from 50-250%).

SUMMARY

- HR metrics, often known as human resource metrics, are crucial figures that assist firms in tracking their human capital and determining the effectiveness of their human resource activities. Turnover, cost per hire, benefits participation rate, and other metrics are examples of this type of information (we'll go over more of them later).
- Both executives and HR experts agree that assessing HR and its influence on an organization is vital, according to a recent research. Many executives, in fact, desire to hear from their HR departments more frequently and receive more extensive HR reports.
- HR metrics give you information to help you make better decisions. They are necessary for process optimization. It isn't tough to use analytics, but it does take time and perseverance: Obtain information about your employees.
- Human Resource teams can measure, analyze, and report on HR KPIs using an HR dashboard, which is a business intelligence tool. Modern, interactive dashboards make use of an HR analytics platform that makes it simple to mix data from several systems and study it in depth right within the dashboard.

- A dashboard serves as a single hub for all pertinent and up-to-date information that employees require to stay productive. An employee dashboard on a mobile platform organizes and prioritizes data so that employees do not waste time looking for it.

QUESTIONS

1. What is HR metrics?
2. Explain the different types of HR metrics.
3. What is staffing metrics?
4. Write a short note Dashboards.
5. Explain VLOOKUP, INDEX, SUMIF, AVERAGEIF and COUNTIF.
6. Write a short note on storyboarding.
7. Explain few key Excel Add-ins.
8. Illustrate the functions to help create dashboards.
9. Explain connecting the dots and integrating the findings.

REFERENCES

- <https://www.bamboohr.com/blog/key-hr-metrics/#:~:text=HR%20metrics%2C%20or%20human%20resources,into%20more%20of%20them%20later/>
- <https://www.workforcehub.com/blog/what-is-the-importance-of-hr-metrics-analysis/>
- <https://www.qlik.com/us/dashboard-examples/hr-dashboard#:~:text=An%20HR%20dashboard%20is%20a,data%20directly%20within%20the%20dashboard>
- <https://www.beekeeper.io/blog/employee-dashboard/>



MODULE 4

CORRELATION AND REGRESSION FOR HR ANALYTICS

Learning Objectives:

1. Understand Correlation Analysis and its objectives.
2. Establish effective output of Correlation Analysis.
3. Understand how to develop and structure Software for Statistical Analysis R&R Commander.
4. Learn the conceptualization of the Simple Linear Regression Analysis in PSPP.
5. Demonstrate the advantages of Free OSS over Closed Software.

4.1 CORRELATION ANALYSIS

4.1.1 Meaning and Definition

Study of the linear relationship between two variables is known as correlation. If the changes in the value of one variable will affect the value of the other variable, then both the variables are correlated. Correlation is a statistical tool used to measure the relationship between two sets of variables and express each in a precise manner.

According to **Croxton** and **Cowden**, “The appropriate statistical tool for discovering and measuring the relationship of quantitative nature and expressing it in brief formula is known as correlation”.

According to **A.M. Tuttle**, “An analysis of the covariance of two or more variables is usually called correlation”.

Correlation analysis is used to determine the degree of relationship between two variables. For example, an HR specialist may wish to know whether the mechanical ability of employees is related to their job performance. If a relationship exists, then a mechanical ability test given to job applicants might be useful in deciding who should be hired.

4.1.2 Types of Correlation

Categorization of correlation is done on various bases such as on the direction of change of the variable, consistency of the ratio of change between the variable, etc.

Types of Correlation

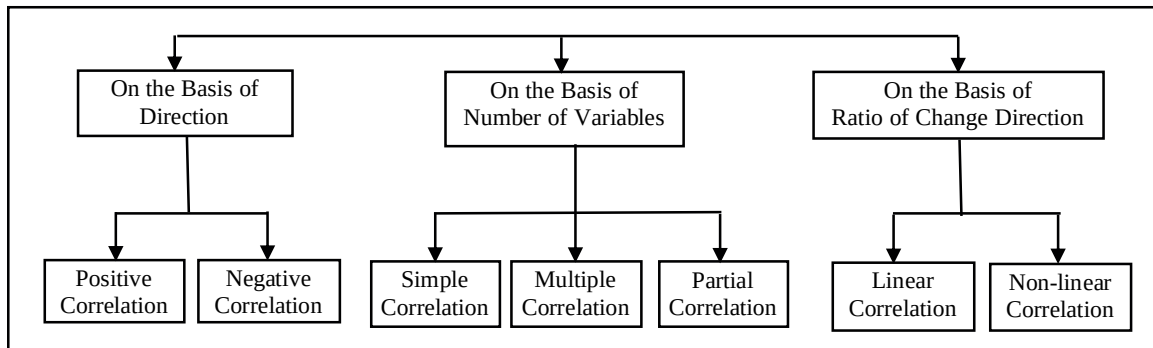


Fig. 4.1: Types of Correlation

4.1.2.1 Positive Correlation and Negative Correlation

Depending on the direction of change of the variables, correlation may be positive or negative.

- 1. Positive Correlation:** Correlation is positive when two variables vary in the same direction.

For example, consider the correlation between sales and expenses of firm (table below).

Firms	1	2	3	4	5	6
Sales (in ₹)	50	60	70	80	90	100
Expenses (in ₹)	11	15	19	23	27	31

- 2. Negative Correlation:** Two variables are said to be negatively correlated when both the variables vary in the opposite direction. When one variable increases, the other variable decreases and *vice versa* is also a negative correlation.

For example, consider the correlation between production and price of crop.

Production (in kg)	100	200	300	400	500	600	700	800
Sales (in ₹)	10	9	8	7	6	5	4	3

4.1.2.2 Linear Correlation and Non-linear (Curvilinear) Correlation

On the basis of consistency of the ratio of change between the variable, correlation may be linear or non-linear:

- 1. Linear Correlation:** In a linear correlation, change in the values of one variable has a fixed ratio to the variation in the values of other variable. When these variables are plotted on a graph, the plotted points would fall on a straight line.

For example, consider the following relationship in the table.

Mustard (kg)	10	20	30	40	50	60	70	80
Oil (kg)	3	6	9	12	15	18	21	24

2. **Non-linear (Curvilinear) Correlation:** In a non-linear correlation, change in values of one variable does not have a fixed ratio to the variation in the value of other variable. When these variables are plotted on a graph, the plotted points would fall on a curve.

For example, consider the following relationship shown in the table.

Use of Fertilizer (in kg)	1	2	3	4	5
Production of Rice (in kg)	4	6	9	17	28

4.1.2.3 Simple, Partial and Multiple Correlations

Depending on the number of variables studied, it can be categorized into three types:

1. **Simple Correlation:** When we measure the linear relationship between two variables, then this interpretation is known as simple correlation, *e.g.*, relationship between sales and expenses, income and consumption, etc.
2. **Partial Correlation:** If we have various related variables and try to find out the relationship between two variables, then it is known as partial correlation. For example, consider the two variables height and weight, which are partially correlated because of the effect of third variable 'age' on height and weight. Relationship between height and weight, known as partial correlation.
3. **Multiple Correlation:** It is defined as the measurement of the effects of multiple variables on one variable. For example, if we try out to find the relationship of rainfall and temperature on the yield of wheat, then this is known as multiple correlation.

4.1.3 Degrees of Correlation

On the basis of coefficient of correlation, the degrees of correlation are determined as follows:

1. **Perfect Correlation:** If two variables change in the same direction and in the same proportion, the correlation between the two variables is called perfect positive correlation. In this case, the coefficient of correlation is +1. If two variables change in the same proportion, the correlation between the two variables is called perfect negative correlation. In this case, the coefficient of correlation is -1.
2. **Absence of Correlation:** If two variables show no relation between them or change in one variable does not lead to a change in the other variable, then it means that there is no relationship between variables. In this case, the coefficient of correlation is zero.
3. **Limited Degree of Correlation:** If two variables are not perfectly correlated or there is an absence of perfect correlation, then it is referred to as limited correlation. It may be positive, negative or zero but lies with the limits ± 1 . High degree, moderate degree and low degree are the three categories of this kind of correlation.

If the value of coefficient of correlation (represented by 'r') lies between ± 0.75 to ± 1 , then it is known as high degree of correlation. When the value of coefficient of correlation lies between ± 0.25 to ± 0.75 , it is known as moderate degree of correlation.

When the value of coefficient of correlation lies between 0 and ± 0.25 , low degree of correlation is said to exist (Table 4.1).

Table 4.1: Degree of Correlation

<i>Degree</i>	<i>Positive Correlation Coefficient</i>	<i>Negative Correlation Coefficient</i>
1. Perfect	± 1	-1
2. Limited		
(i) High	Between $+0.75$ to $+1$	Between -0.75 to -1
(ii) Moderate	Between $+0.25$ to $+0.75$	Between -0.25 to 0.25
(iii) Low	Between 0 to $+0.25$	Between 0 to -0.25
(iv) Absence	Zero	Zero

4.1.4 Correlation Coefficient

The extent of measurement as to how much one number can expect to be influenced by changes in another number is known as correlation coefficient. For example, measuring the value that how much the share fluctuations in the index. If the two numbers are perfectly correlated, then it means that value of correlation coefficient is $+1$.

If one grows, the other also grows, and the change in one is a multiple of the change in the other. If the numbers are perfectly or inversely correlated, then it means that value of correlation coefficient is -1 . If one grows, the other falls. The growth in one variable is a negative multiple of the growth in the other variable. If the two variables are not related, then the correlation coefficient is zero.

When numbers are related and the coefficient is either $+1$ or -1 and other influences and relationship between two numbers is not fixed, then this is known as non-zero correlation coefficient. If one number is known, the other number can be calculated, but with uncertainty. Uncertainty increases when correlation coefficient is more towards zero. But this relationship is not useful, if the correlation is very low.

4.1.5 Applications of Correlation

1. The nature, direction and degree of relationship between two or more variables are determined by the use of correlation analysis.
2. It is used for estimating the change in value of one variable occurs due to the changes in the value of other variable.
3. Under some specific conditions, it helps the researchers to understand the behaviour of certain events. For example, one can find that the various factors affect the production of rice.
4. Correlation analysis provides the platform of decision-making to business word and reduces the factor of uncertainty from the decision-making.
5. Correlation analysis is helpful in making predictions.

4.2 METHODS OF COMPUTING CORRELATION

The various methods of computing linear correlation can be shown diagrammatically as follows:

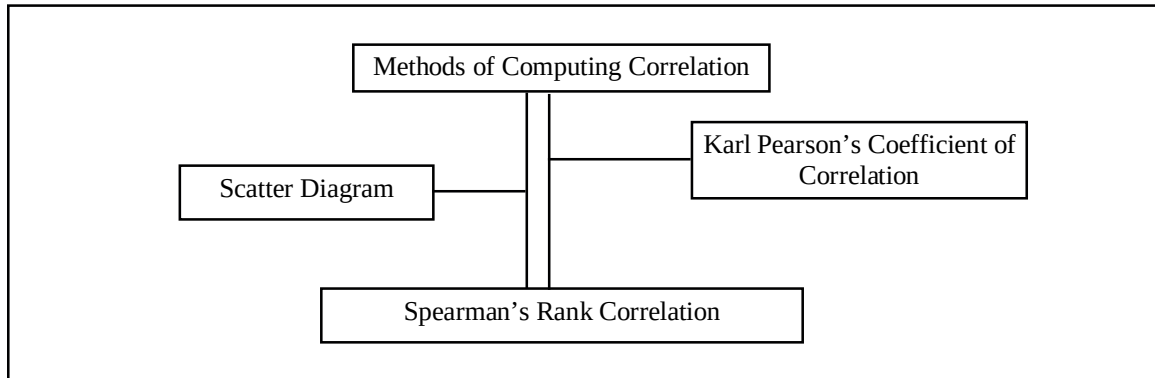


Fig. 4.2: Methods of Computing Correlation

4.2.1 Scatter Diagram

Graphical presentation of the relationship between two variables calculated on the same set of individual is known as **scatter diagram**.

Note: Scatter diagram is a dot chart specially used to show the correlation.

For example, if we plot a dot chart for some predefined value of two variables and get some trends (either upwards or downwards) in the chart, then the trend shows a correlation between both the variables.

The idea will be clear from the Fig. 4.3.

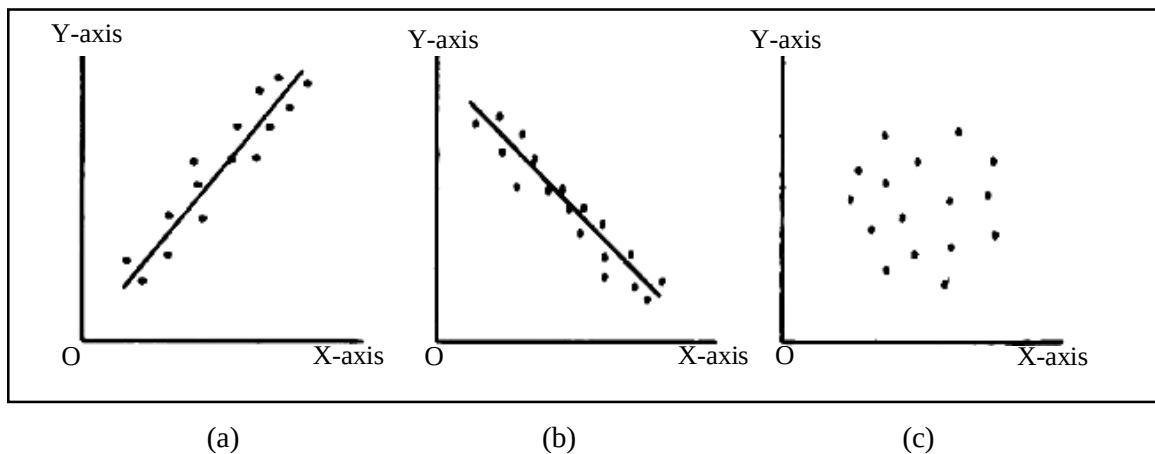


Fig. 4.3: Scatter Diagram

- (a) Highly positive correlation
- (b) Highly negative correlation
- (c) Absence of correlation

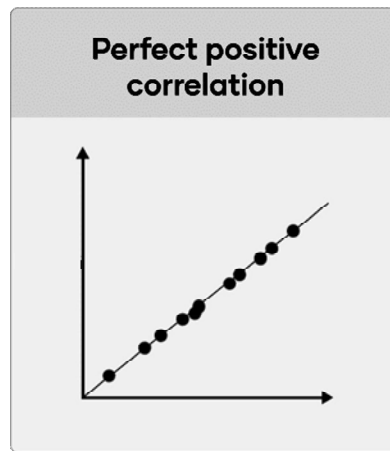


Fig. 4.4(a): Perfect Positive Correlation

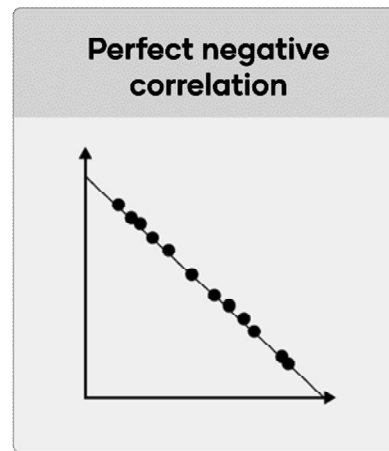


Fig. 4.4(b): Perfect Negative Correlation

In Figure 4.3(a), the value of the two variables move in the upward direction, hence the correlation is positive.

In Figure 4.3(b), we find negative correlation as the values move in reverse direction.

In Figure 4.3(c), the value of the two variables moves in the same direction and when $r = +1$, it shows perfect positive correlation, and when $r = -1$, it shows perfect negative correlation.

4.2.1.1 Advantages of Scatter Diagram

1. It is very simple and non-mathematical technique.
2. It is not influenced by the size of extreme item.
3. It is the very basic step to find out the relationship between two variables.

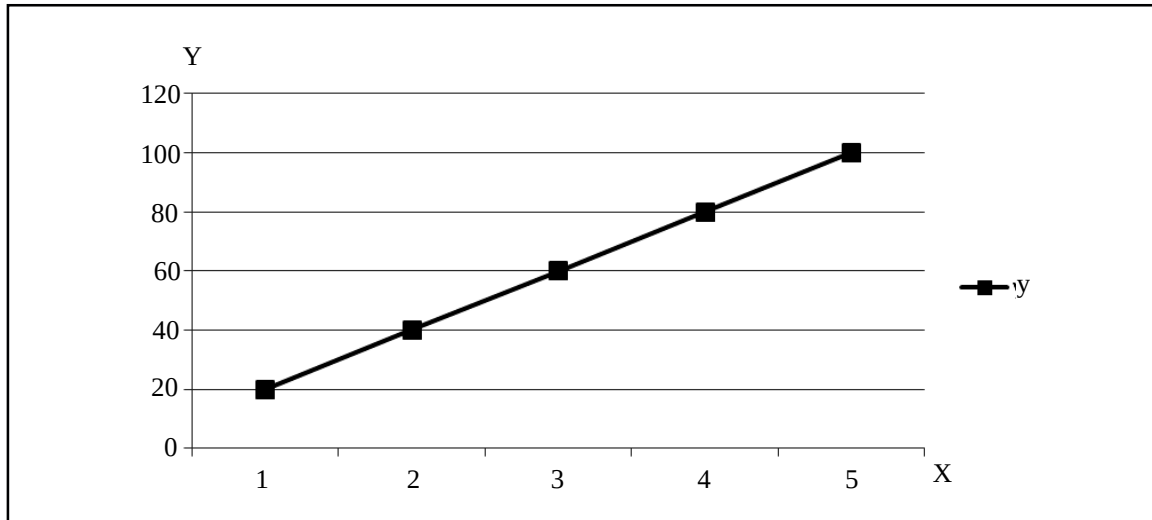
4.2.1.2 Disadvantages of Scatter Diagram

The main disadvantage of this technique is that it cannot find out an exact degree of correlation between two variables. We can only view the visual form of correlation and direction (positive or negative) on the chart.

Example 1: Draw a scatter diagram for the following and indicate whether correlation is positive or negative.

X	10	20	30	40	50
Y	20	40	60	80	100

Solution:



4.2.2 Karl Pearson's Coefficient of Correlation

To calculate the magnitude of linear relationship between two variables, Karl Pearson (biometrician and statistician) gave a quantitative technique. This technique is known as Pearson's Coefficient of Correlation (r) and is extensively used in practice.

4.2.2.1 Properties of Karl Pearson's Coefficient of Correlation

1. Ideal measure of correlation and independent of the units of the variables.
2. Free from change of origin and scale.
3. Based on all observations.
4. Lies between -1 and $+1$ and when:
 - (i) $r = -1$, shows a perfect negative correlation between variables.
 - (ii) $r = 0$, shows no correlation between variables.
 - (iii) $r = +1$, shows a perfect correlation between variables.

4.2.2.2 Advantage of Karl Person's Coefficient of Correlation

The main advantage of the Karl Person's coefficient of correlation is that it gives the result in one value and also summarizes the degree of correlation and direction.

4.2.2.3 Disadvantages of Karl Person's Coefficient of Correlation

1. Every time, it assumes only a linear relationship between variables.
2. Interpreting the value of correlation coefficient (r) is difficult.
3. Significance of correlation coefficient (r) is affected by the extreme values.
4. It is a time-consuming method.
5. It does not convey the cause-and-effect relationship.

4.2.2.4 Calculation of Karl Pearson's Coefficient of Correlation

The various formulae used to calculate coefficient of correlation (r) are:

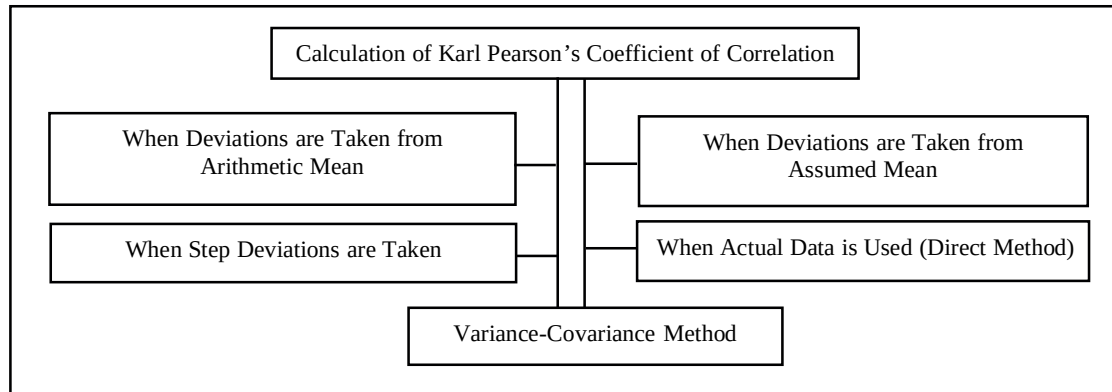


Fig. 4.6: Calculation of Karl Pearson's Coefficient of Correlation

4.2.2.5 When Deviations are Taken from Arithmetic Mean

When deviation is taken from arithmetic mean, then r is calculated as:

$$r = \frac{\sum xy}{\sqrt{\sum x^2} \sqrt{\sum y^2}} \text{ or } \frac{\sum xy}{\sqrt{\sum x^2} \sqrt{\sum y^2}} \text{ or } \frac{\sum xy}{No_x o_y}$$

$$\text{or } r = \frac{\sum (x - \bar{X})(y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2} \sqrt{\sum (Y - \bar{Y})^2}}$$

where, $X = x - \bar{X}$ (deviation from Arithmetic Mean (AM) of X series)

$Y = y - \bar{Y}$ (deviation from Arithmetic Mean (AM) of Y series)

$\sigma_x =$ Standard Deviation (SD) of X series

$\sigma_y =$ Standard Deviation (SD) of Y series

and $N =$ Number of observations

Example 2: Calculate Karl Pearson's Coefficient of Correlation from the following data:

X	42	52	55	60	66	68	65	60	58	34
Y	11	13	18	22	26	40	31	27	24	18

Solution:

X	Y	$x = X - \bar{X}$	x^2	$y = Y - \bar{Y}$	y^2	xy
42	11	-14	196	-12	144	168
52	13	-4	16	-10	100	40
55	18	-1	1	-5	25	5
60	22	4	16	-1	1	-4
66	26	10	100	3	9	30
68	40	12	144	17	289	204
65	31	9	81	8	64	72

60	27	4	16	4	16	16
58	24	2	4	1	1	2
34	18	-22	484	-5	25	110
$\Sigma X = 560$	$\Sigma Y = 230$		$\Sigma x^2 = 1058$		$\Sigma x^2 = 674$	$\Sigma xy = 643$

$$\bar{X} = \frac{\Sigma x}{N} = 56; \quad \bar{Y} = \frac{\Sigma y}{N} = \frac{230}{10} = 23$$

$$\text{Coefficient of correlation } (r) = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} = \frac{643}{\sqrt{1058 \times 674}} = 0.76$$

4.2.2.6 When Deviations are Taken from Assumed Mean

When deviations are taken from assumed mean, the following formula is used to calculate the value of r .

$$r = \frac{N \times \Sigma dxdy - \Sigma dx \Sigma dy}{\sqrt{N \cdot \Sigma dx^2 - (\Sigma dx)^2} \sqrt{N \cdot \Sigma dy^2 - (\Sigma dy)^2}}$$

where, $dx = X - A$ and $dy = Y - A$

A = Assumed Mean.

Example 3: Find the coefficient of correlation from the following data:

X	1	2	3	4	5	6	7	8	9	10
Y	62	56	48	41	36	28	21	16	12	8

Solution: Let the assumed mean for X be 5 and assumed mean for Y be 40.

X	Y	$dx = (X - 5)$	$dy = (Y - 40)$	dx^2	dy^2	$dxdy$
1	62	-4	22	16	484	-88
2	56	-3	16	9	256	-48
3	48	-2	8	4	64	-16
4	41	-1	1	1	1	-1
5	36	0	-4	0	16	0
6	28	1	-12	1	144	-12
7	21	2	-19	4	361	-38
8	16	3	-24	9	576	-72
9	12	4	-28	16	784	-112
10	8	5	-32	25	1024	-160
Total		$\Sigma dx = 5$	$\Sigma dy = -72$	$\Sigma dx^2 = 85$	$\Sigma dy^2 = 3710$	$\Sigma dxdy = -547$

$$r = \frac{N \times \Sigma dxdy - \Sigma dx \Sigma dy}{\sqrt{N \cdot \Sigma dx^2 - (\Sigma dx)^2} \sqrt{N \cdot \Sigma dy^2 - (\Sigma dy)^2}}$$

$$= \frac{10(-547) - 5(-72)}{\sqrt{10 \times 85 - (5)^2} \times \sqrt{10 \times 3710 - (-72)^2}}$$

The variable X and Y are perfectly negatively correlated.

4.2.2.7 When Step Deviations are Taken

To calculate r , the following formula is used when step deviation is taken:

$$r = \frac{\sum dx'dy' - \frac{\sum dx'dy'}{N}}{\sqrt{\sum dx'^2 - \frac{(\sum dx')^2}{N}} \sqrt{\sum dy'^2 - \frac{(\sum dy')^2}{N}}}$$

where, $dx' = \frac{dx}{C}$, $dy' = \frac{dy}{C}$

Example 4: Calculate the correlation between the expenses and saving of any family.

Expenses (₹)	0	15	25	35	75
Saving (₹)	70	55	35	25	15

Solution: Let the assumed mean for X be 25 and assumed mean for Y be 45.

X	$dx = X - 25$	$dx' = \frac{dx}{5}$	$(dx')^2$	Y	$dy = (Y - 45)$	$dy' = \frac{dy}{5}$	$(dy')^2$	$dx'dy'$
0	-25	-5	25	70	25	5	25	-25
15	-10	-2	4	55	10	2	4	-4
25	0	0	0	35	-10	-2	4	0
35	10	2	4	25	-20	-4	16	-8
75	50	10	100	15	-30	-6	36	-60
$N = 5$		$\sum dx' = 5$	$\sum dx'^2 = 133$	$N = 5$		$\sum dy' = -5$	$\sum dy'^2 = 85$	$\sum dx'dy' = -97$

$$r = \frac{\sum dx'dy' - \frac{\sum dx'dy'}{N}}{\sqrt{\sum dx'^2 - \frac{(\sum dx')^2}{N}} \sqrt{\sum dy'^2 - \frac{(\sum dy')^2}{N}}}$$

$$= \frac{-97 - \left(\frac{-97}{5}\right)}{\sqrt{133 - \left(\frac{25}{5}\right)} \sqrt{-85 - \left(\frac{25}{5}\right)}} = \frac{-77.6}{\sqrt{128} \sqrt{80}} = \frac{-77.6}{\sqrt{10240}} = -0.767$$

4.2.2.8 When Actual Data is Used (Direct Method)

In case of actual data, the following formula is used to calculate the coefficient r .

$$r = \frac{N \times \sum XY - (\sum X)(\sum Y)}{\sqrt{N \cdot \sum X^2 - (\sum X)^2} \sqrt{N \cdot \sum Y^2 - (\sum Y)^2}}$$

Example 5: Calculate the Karl Pearson's coefficient of correlation for the following data:

X	4	7	11	14	19	15
Y	18	16	17	19	19	21

Solution:

X	Y	X ²	Y ²	XY
4	18	16	324	72
7	16	49	256	112
11	17	121	289	187
14	19	196	361	266
19	19	361	361	361
15	21	225	441	315
ΣX = 70	ΣY = 110	ΣX ² = 968	ΣY ² = 2032	ΣXY = 1313

$$r = \frac{N \sum XY - (\sum X)(\sum Y)}{\sqrt{N \sum X^2 - (\sum X)^2} \sqrt{N \sum Y^2 - (\sum Y)^2}}$$

$$= \frac{6 \times 1313 - 70 \times 110}{\sqrt{6 \times 968 - (70)^2} \sqrt{6 \times 2032 - (110)^2}} = \frac{178}{\sqrt{83536}} = \frac{178}{289} = 0.616$$

4.2.2.9 Variance-Covariance Method

Following formula is used to calculate the value of r when variance and covariance are given:

$$r = \frac{\text{Cov.}(X, Y)}{\sqrt{\text{Var.}(X)} \sqrt{\text{Var.}(Y)}}; \quad \text{where, } r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2} \sqrt{\sum (Y - \bar{Y})^2}}$$

Divide numerator and denominator by N , we have

$$r = \frac{\frac{\sum (X - \bar{X})(Y - \bar{Y})}{N}}{\sqrt{\frac{\sum (X - \bar{X})^2}{N}} \sqrt{\frac{\sum (Y - \bar{Y})^2}{N}}} = \frac{\sum (X - \bar{X})(Y - \bar{Y}) / N}{\sqrt{\frac{\sum (X - \bar{X})^2}{N}} \times \sqrt{\frac{\sum (Y - \bar{Y})^2}{N}}}$$

$$= \frac{\text{Cov.}(X, Y)}{\sqrt{\text{Var.}(X)} \sqrt{\text{Var.}(Y)}}$$

Example 6: If covariance between X and Y variables is 12.5 and the variance of X and Y are respectively 16.4 and 13.8, find the coefficient of correlation between them.

Solution: Covariance of $XY = 12.5$, Variance of $X = 16.4$, Variance of $Y = 13.8$

$$r = \frac{\text{Cov.}(X, Y)}{\sqrt{\text{Var.}(X)} \sqrt{\text{Var.}(Y)}} = \frac{12.5}{\sqrt{16.4} \sqrt{13.8}} = \frac{12.5}{\sqrt{226.32}} = \frac{12.5}{15.04} = 0.83$$

4.2.2.10 Correlation in Continuous Series

When we have to find coefficient of correlation from a bivariate grouped data, the following formula is used:

$$r = \frac{N \times \sum fd_x d_y - \sum fd_x \cdot \sum fd_y}{\sqrt{N \cdot \sum fd_x^2 - (\sum fd_x)^2} \sqrt{N \cdot \sum fd_y^2 - (\sum fd_y)^2}}$$

where, $d_x = X - A_x$

$d_y = Y - A_y$

A_x = assumed mean of X

A_y = Assumed mean of Y

Example 7: From the table given below calculate the Coefficient of Correlation between the ages of husbands and wives:

Age of Wives (Y-Series)	Age of Husbands (X-Series)					
	20-30	30-40	40-50	50-60	60-70	Total
15-25	5	9	3	—	—	17
25-35	—	10	25	2	—	37
35-45	—	1	12	2	—	15
45-55	—	—	4	16	5	25
55-65	—	—	—	4	2	6
Total	5	20	44	24	7	100

Solution:

Table 4.2.: Calculation of Coefficient of Correlation between the Age of Husband and Wives

Age of Wives (Y)				Age of Husbands Wives (X)								
				Age group 20-30	30-40	40-50	50-60	60-70				
d_x d_y				m.p.25	35	45	55	65				
				-20	-10	0	+10	+20	Total			
				-2	-1	0	+1	+2	f	fd_y	fd_y^2	fd_xd_y
Age group 15-25	m.p. 20	-20	-2	20 (fd_xd_y) 5	18 9	0 3			17	-34	68	38
25-35	30	-10	-1		10 10	0 12	-2 2		37	-37	37	8
35-45	40	0	0		0 1	0 12	0 2		15	0	0	0
45-55	50	+1	+1			0 4	16 16	10 5	25	+25	25	26
55-65	60	+20	+2				8 4	8 2	6	+12	24	16
Total f				5	20	44	24	7	100	$\sum fd_y = -34$	$\sum fd_y^2 = 154$	$\sum fd_xd_y = 88$

fd_x	-10	-20	0	+24	+14	$\sum fd_x = 8$			
fd_x^2	20	20	0	24	28	$\sum fd_x^2 = 92$			
$fd_x d_y$	20	28	0	22	18				

$$d_x = \frac{X - A_x}{h}, d_y = \frac{Y - A_y}{h}$$

Substituting the above values in Karl Pearson's formula:

$$r = \frac{\sum fd_x d_y - n \left(\frac{\sum fd_x}{n} \right) \left(\frac{\sum fd_y}{n} \right)}{\sqrt{\sum fd_x^2 - \left(\frac{\sum fd_x}{n} \right)^2} \sqrt{\sum fd_y^2 - \left(\frac{\sum fd_y}{n} \right)^2}}$$

$$\text{we get, } r = \frac{88 - 100 \left(\frac{8}{100} \right) \left(\frac{-34}{100} \right)}{100 \sqrt{\frac{92}{100} - \left(\frac{8}{100} \right)^2} \sqrt{\frac{154}{100} - \left(\frac{-34}{100} \right)^2}}$$

$$= \frac{9072}{\sqrt{9136 \times 14244}} = \frac{9072}{11407.59} = 0.79$$

4.2.3 Spearman's Rank Correlation

Spearman's Rank Correlation Coefficient, named after Charles Edward Spearman's and denoted by R , is a technique to find the correlation between the ranks of two series. This technique is used when the value of variable cannot be calculated quantitatively. To apply the Spearman's rank correlation technique, we need to first arrange the value of variable in serial order. We apply this method when we deal with qualitative characteristics.

For example, when we test the IQ of a person with the number of hours spent in front of TV per week, etc. Spearman's rank correlation coefficient (R) is defined as:

$$R = 1 - \frac{6 \sum D^2}{N(N^2 - 1)}$$

where, R = Rank coefficient of correlation

D = Difference between two ranks ($R_1 - R_2$)

N = Number of pair of observations

Rank Correlation Coefficient (R) always exists between -1 and $+1$ and also used when the measurements are given for both the series.

4.2.3.1 Advantages of Spearman's Rank Method

1. Easy to understand and simple to calculate.
2. When data are qualitative (*e.g.*, intelligence, efficiency, etc.) in nature, it is very useful.
3. It also applies when actual data are given.

4.2.3.2 Disadvantages of Spearman's Rank Method

1. It is not useful in frequency distribution.
2. If we take a large sample ($n > 30$), then calculation becomes more difficult and takes too much time.

4.2.3.3 Calculation of Rank Correlation

Various conditions for calculating the Rank Correlation are as follows:

1. When Ranks are Given: When the actual ranks (say R_1 and R_2) are given, the following steps are used to calculate the rank:

- (i) Calculate the difference of two ranks ($R_1 - R_2$) and assign it to D , i.e.,

$$D = (R_1 - R_2).$$

- (ii) Then calculate the square of D and summate the D as ΣD^2 .

- (iii) In the last, put all the calculated values in the formula.

Example 8: Following are the rank obtained by 10 students in two subjects, English and History. To what extent the knowledge of the students in the two subjects is related?

English	1	2	3	4	5	6	7	8	9	10
History	2	4	1	5	3	10	9	6	7	8

Solution: Calculation of Rank Correlation Coefficient.

Rank of English (R_1)	Rank of History (R_2)	$D = (R_1 - R_2)$	D^2
1	2	-1	1
2	4	-2	4
3	1	+2	4
4	5	-1	1
5	3	+2	4
6	10	-4	16
7	9	-2	4
8	6	+2	4
9	7	+2	4
10	8	+2	4
			$\Sigma D^2 = 46$

$$\begin{aligned}
 \text{Rank correlation coefficient (R)} &= 1 - \frac{6 \Sigma D^2}{N(N^2 - 1)} = 1 - \frac{6 \times 46}{10(10^2 - 1)} \\
 &= 1 - \frac{276}{990} = 1 - 0.28 = +0.72
 \end{aligned}$$

2. When Ranks are Not Given: If no rank is given, then we first calculate the rank of the given data (actual data).

- (i) To assign the rank, we take the lowest number of rank as 1, we take the next lowest number as 2, and so on as well as assign the rank lightest number.
- (ii) Calculate the difference of two ranks ($R_1 - R_2$) and assign it to D , i.e., $D = (R_1 - R_2)$.
- (iii) Then calculate the square of D and summate the D as ΣD^2 .
- (iv) In the last, put all the calculated values in the formula.

Example 9: Calculate coefficient of correlation by rank method:

X	54	58	85	75	65	90	80	50
Y	120	134	150	115	110	140	142	100

Solution: Calculation of Rank Correlation Coefficient

X	Rank (R_1)	Y	Rank (R_2)	$D = (R_1 - R_2)$	D^2
54	2	120	4	-2	4
58	3	134	5	-2	4
85	7	150	8	-1	1
75	5	115	3	+2	4
65	4	110	2	+2	4
90	8	140	6	+2	4
80	6	142	7	-1	1
50	1	100	1	0	0
					$\Sigma D^2 = 22$

Coefficient of Rank Correlation

$$R = 1 - \frac{6 \Sigma D^2}{N(N^2 - 1)} = -\frac{6 \times 22}{8(64 - 1)} = 1 - \frac{132}{504} = \frac{504 - 132}{504} = \frac{372}{504} = 0.738$$

3. Tied Rank/Equal or Repeated Ranks: When the values are repeated, then ranks assigned to these values are similar. In this case, calculate the average of rank and assign it to the repeated values. For example, if two persons are positioned on the fifth place, then each of them are given the rank $\frac{5+6}{2} = 5.5$ which is common rank to be assigned; and the next rank will be 7.

In case if three are ranked equal at the fifth place, then they all are given the rank $\frac{5+6+7}{3} = 6$

which is the common rank to be assigned to each; and the next rank will be 8, in this case. In this (Tied Ranks/Equal or Repeated Ranks) case, to perform calculation, the following formula is used:

$$R = 1 - \frac{6 \left(\Sigma D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) \dots \right)}{N^2 - N}$$

where, m = Number of items having common ranks.

Example 10: Obtain the rank correlation coefficient for the following data:

X	68	64	74	50	64	80	74	40	55	64
Y	62	58	67	45	81	60	67	48	50	70

Solution: Calculation of Rank Correlation

x	Rank (R ₁)	Y	Rank (R ₂)	D = (R ₁ - R ₂)	D ₂
68	7	62	6	1	1
64	5	58	4	1	1
74	8.5	67	7.5	1	1
50	2	45	1	1	1
64	5	81	10	-5	25
80	10	60	5	5	25
74	8.5	67	7.5	1	1
40	1	48	2	-1	1
55	3	50	3	0	0
64	5	70	9	-4	16
					$\Sigma D^2 = 72$

In series C, the value 64 occurs thrice ($m_1 = 3$), i.e., at 4th, 5th and 6th rank.

Hence, all three values are given the average rank, i.e., $\frac{4+5+6}{3}$ th = 5th rank.

Also, in series X, the value 74 occurs twice ($m_2 = 2$), i.e., at 8th and 9th rank.

Hence, values are given the average rank, i.e., $\frac{8+9}{2}$ th = 8.5th rank.

While, in series Y, the value 67 occurs both at the 7th and 8th rank, ($m_3 = (2)$). Hence, both are given the average rank, i.e., $\frac{7+8}{2}$ th = 7.5th rank.

$$\text{Coefficient of Rank Correlation (R)} = 1 - \frac{6 \left[\sum D^2 + \frac{m_1^3 - m_1}{12} + \frac{m_2^3 - m_2}{12} + \frac{m_3^3 - m_3}{12} \right]}{N(N^2 - 1)}$$

$$= 1 - \frac{6 \left[72 + \frac{3^3 - 3}{12} + \frac{2^3 - 2}{12} + \frac{2^3 - 2}{12} \right]}{10(10^2 - 1)}$$

$$R = 0.5455$$

4.2.4 Output of Correlation Analysis

Correlation estimates the strength of the linear relationship between two (and only two) variables. Correlation coefficients ranges from -1.0 (a perfect negative correlation) to positive 1.0

(a perfect positive correlation). The closer correlation coefficients get to zero, the weaker the correlation is between the two variables. Ordinal or ratio data (or a combination) must be used.

Use the Pearson's correlation coefficient to examine the strength and direction of the linear relationship between two continuous variables:

1. **Strength:** The correlation coefficient can range in value from -1 to $+1$. The larger the absolute value of the coefficient, the stronger the relationship between the variables. For the Pearson's correlation, an absolute value of 1 indicates a perfect linear relationship. A correlation close to 0 indicates no linear relationship between the variables.
2. **Direction:** The sign of the coefficient indicates the direction of the relationship. If both variables tend to increase or decrease together, the coefficient is positive, and the line that represents the correlation slopes upward. If one variable tends to increase as the other decreases, the coefficient is negative, and the line that represents the correlation slopes downward.

Consider the following points when to interpret the correlation coefficient:

1. It is never appropriate to conclude that changes in one variable cause changes in another based on correlation alone. Only properly controlled experiments enable to determine whether a relationship is causal.
2. The Pearson's correlation coefficient is very sensitive to extreme data values. A single value that is varying different from the other values in a dataset can greatly change the value of the coefficient. You should try to identify the cause of any extreme value. Correct any data entry or measurement errors. Consider removing data values that are associated with abnormal, one-time events (special causes). Then repeat the analysis.
3. A low Pearson's correlation coefficient does not mean that no relationship exists between the variables. The variables may have a non-linear relationship. To check for non-linear relationships graphically, create a scatterplot or use simple regression.

4.3 THE CASE OF OUTLIER

Outlier is a data point in the dataset that differs significantly from the other data or observations. The scatterplot can be used to identify possible outliers within the dataset. The outlier could have undue influence on the value of the correlation and regression coefficients estimated in fitting a model to the most other values in a sample of data. It is important to identify outliers as just a few extreme values may alter the results considerably.

Sometimes, outliers are legitimate values; other times, they are a result of a data entry error. If it is the former, we go with the data given. It might give an odd, but perhaps right model. If it is the latter, that can be rosily fixed by consulting the data owners to help determine what the right number should be.

How do we know if any one point is truly an outlier? The best approach is to make a probabilistic analysis:

Step 1: Examine the distribution of data prior to statistical analysis.

Step 2: As a rule of thumb, any data point that lies $\pm$ three standard deviations away from the mean of your data is likely to be an outlier. In a normal distribution, the probability that any point in the data sample will fall three standard deviations away from the mean is 99%. Thus, any data point that falls $\pm$ three standard deviations from the mean only has a 1% probability of less of being in any other sample that you may do in the future. It is possible then to confidently get rid of the outlier because it is unlikely to occur again.

It is worth noting that, in the day-to-day work of HR analytics, the normal bell curve (the S-curve) hardly ever happens. Bell curves are found more often in naturally occurring things, such as height, weight, amount of rainfall, amount of sunshine, etc. The things we deal with in HR analytics – pay rates, turnover rates, performance appraisals, even intelligence – will not follow a normal distribution.

The main reason is that HR processes are usually biased, from a statistical point of view. The data are generally skewed because we set it out to be skewed. We hardly do anything that would follow a random pattern — we hire people that we think are smart; we promote people who are more capable; we slow down pay increases over time. Mostly, everything we do is going to be skewed. However, for purposes of determining outliers, the three standard-deviation rules of thumb is nonetheless useful.

4.4 SOFTWARE FOR STATISTICAL ANALYSIS 1 – GNU PSPP

4.4.1 Introduction

GNU PSPP is a program for statistical analysis of sampled data. It is an alternative for IBM SPSS Statistics and is written in C. It has a graphical user interface and conventional command-line interface. PSPP functionality includes descriptive statistics, t-tests, ANOVA, linear and logistic regression, cluster analysis, reliability and factor analysis, non-parametric tests and more.

PSPP is a program for statistical analysis of sampled data. PSPP is free statistical program being developed as an alternative to the proprietary statistical package SPSS. PSPP is almost same to SPSS functionality and interface, but the advantage is PSPP is a free source software. Despite being in development since 1995 now, PSPP has not hit a 1.0 release yet, and its current stable version is at 0.8.0.

PSPP is GNU software and part of the GNU System. PSPP will always work best on a free operating system. PSPP is a program for data such as descriptive statistics, t-tests, linear regression and non-parametric tests.

PSPP can generate high quality plots to help with visualization of the distribution of data. Among the type of plots which can be displayed are box-and-whisker plots, normal probability plots and histograms. These complement descriptive statistics and help determine the most appropriate type of analysis for the data, and/or what transformations are necessary. The data selection capabilities of PSPP make it simple to generate plots from a subset of variables or from data which match only certain criteria.

Plots and graphs created by PSPP are formatted in standard file formats such as postscript or PNG, so as to allow easy import into reports or other documents.

4.4.2 Features of PSPP

A brief of some of the PSPP's features is as follows:

1. Support for over 1 billion cases.
2. Support for over 1 billion variables.
3. Syntax and data files which are compatible with those of SPSS.
4. A choice of terminal or graphical user interface.
5. A choice of text, postscript, PDF, open document or html output formats.
6. Interoperability with Gnumeric, LibreOffice, OpenOffice.org and other free software.
7. Easy data import from spreadsheets, text files and database sources.
8. The capability to open analyses and edit two or more datasets concurrently. They can also be merged, joined or concatenated.
9. A user interface supporting all common character sets and which has been translated to multiple languages.
10. Fast statistical procedures, even on very large datasets.
11. No license fees.
12. No expiration period.
13. No unethical "end-user license agreements."
14. A fully indexed user manual.
15. Freedom ensured; it is licensed under the GPT v3 or later.
16. Portability; runs on many different computers and many different operating systems (GNU or GNU/Linux are the preferred platforms, but we have had many reports that it runs well on other systems too.)

4.4.3 Started with PSPP

On windows, PSPP can be started by clicking on the Start Menu button on the lower left corner of your left corner of your display. PSPP will be available with the "P" programs. Open the PSPP folder, and then click on the PSPP icon to launch the program, the program can be started in a similar manner on other computer systems.

PSPP also has a command line option. This is more useful for advanced users. The remaining icons in the PSPP folder are mostly documentation.

4.4.4 PAPP Environment

PAPP for windows provides a powerful statistical analysis and data management system in a graphical environment, using descriptive menus and simple dialogue boxed to do most of the work for you. Most of the tasks can be accomplished simply by pointing and PSPP consists of three windows:

1. PSPPIRE Data Editor
2. PSPPIRE Output Viewer (Table View and Chart View)
3. PSPPIRE Syntax Editor.

There are two major types to run PSPP which is going through windows of PSPPIRE Data Editor and PSPPIRE Syntax Editor.

For **PSPPIRE Data Editor** Window:

Go to **File > New > Click once at Data.**

For **PSPPIRE Syntax Editor** Window:

Go to **File > New > Click once at Syntax.**

4.4.4.1 PSPPIRE Data Editor

The PSPPIRE Data Editor (PSPPIRE) appears as a Spreadsheet-like interface which allows the user to enter data into cells like in Excel and other Spreadsheet applications. It allows easy entry and editing of data. It includes two views, called “Data View” and “Variable View”, which are seen as tabs at the bottom-left of the Spreadsheet window that can be used to switch between these modes. PSPP can open SPSS data files as well as CSV files (see Fig. 4.7).

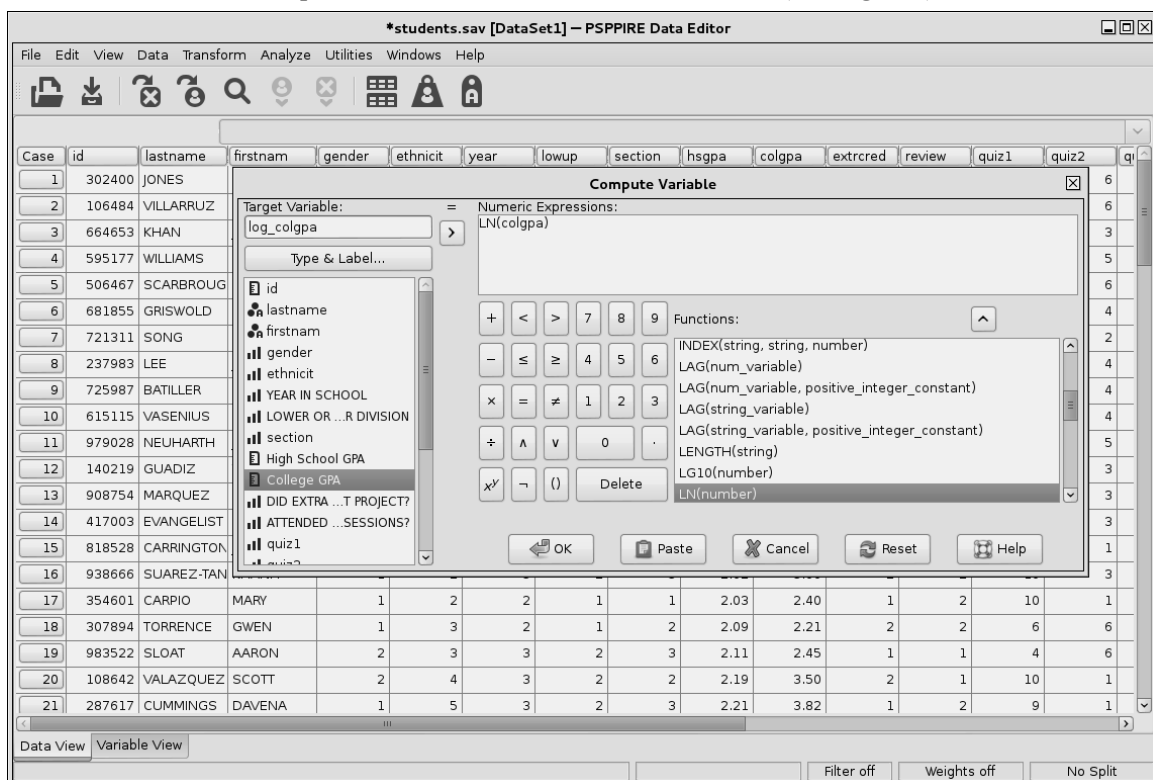


Fig. 4.7: PSPPIRE Data Editor

PSPP typically uses two different windows, which is just like SPSS.

The “data view” is the default that you will see when you start PSPP. The data view will display your data in a table that looks similar like a spreadsheet, although it doesn’t look much like a spreadsheet yet. When an analysis is done, a new, second window will be opened to display the output.

An important difference between PSPP and Spreadsheets is that PSPP only uses numbers, letters, words, or other data in the data view. There are no formulas entered into the cells in the data view. The analyses instead will be done from the “Analyze” dropdown menu.

As seen above, the Data View shows the actual values of the cells for every row and column of data. This is the mode for editing and manipulating the actual values of observations. The Variable View displays characteristics of each variable and allows the user to edit these. You can add or delete (clear) cases or variables with the right click on the variable name (column name) or case number (row number).

4.4.4.2 PSPPIRE Output Viewer

PSPP output is currently text with table and graphs implemented in the graphical interface (PSSPIRE) in this version.

4.4.4.3 PSPPIRE Syntax Editor

Although most tasks can be accomplished by simple pointing and clicking, PSPP also provides a powerful command language in syntax window, which allows you to save and automate many common tasks.

4.4.5 Menus in PSPP

1. Menus in PSPPIRE Data Editor: PSPP is a menu-driven interface and has a variety of pull-down menus available for the user. The main menu bar in the PSPPIRE Data Editor contains nine menus – File, Edit, View, Data, Transform, Analyze, Utilities, Windows and Help.

Note: SPSS has 10 main menus; PSPP has nine same main menus with SPSS, but no graphs.

- (i) **File:** The File menu allows you to create new files, open existing files (SAV file) or syntax file, import delimited data, save the file (in SAV format only) or display data file information, open recently used files, save files and also put the name as a dataset.
- (ii) **Edit:** From the Edit menu, the user can insert variable, go to case, or cut, copy and paste data. Search option enables user to search for a specific value in a selected variable.
- (iii) **View:** From the View menu, the user can switch between data view and variable view, show or hide the status bar (area at the bottom, where information about filtering, weights and splits is displayed), show or hide gridlines, change fonts to display data, and show value labels in data view.
- (iv) **Data:** From the data menu, the user can do several data manipulations:

- (a) **Sort Cases:** This dialog box sorts cases. You can sort cases in ascending or descending order.
- (b) **Transpose:** Transpose creates a new data file in which the rows and columns in the original data file are transposed so that cases (rows) become variables and variables (columns) become cases. Transpose automatically creates new variable names and displays a list of the new variable names.
- (c) **Split File:** Split file splits the data file into separate groups for analysis based on the values of one or more grouping variables. If you select multiple grouping variables, cases are grouped by each variable within categories of the preceding variable on the groups based on the list.

Cases should be sorted by values of the grouping variables, and in the same order, that variables are listed in the Groups Based on List. If the data file is not already sorted, select Sort the File by Grouping variables.

 - **Compare Groups:** Split-file Groups are shown together for comparison purposes. For pivot tables, a single pivot table is created and each split-file variable can be moved between table dimensions. For Charts, a separate chart is created for each split-file group and the charts are displayed together in the Viewer.
 - **Organize Output by Groups:** All results for each procedure are displayed separately for each split-file group.
- (d) **Select Cases:** Select Cases provide several methods for selecting a sub-group of cases on criteria that include variables and complex expressions. You can also select a random sample of cases.
- (e) **Weight Cases by Weight Variable:** Sometimes, we need to let some cases have more “weight” in the analyses because we undersampled or oversampled from a group. We can determine a set of weights, one weight in each case so that the groups would more closely resemble the proportions we had hoped to sample.
- (v) **Transform:** From Transform menu, the user can compute new variables, rank cases and do recoding.
- (vi) **Analyze:** Analyze menu is the most important menu in PSPP. It contains entries for several statistical analysis of data. For now, the following analyses are supported:
 - Descriptive statistics (frequencies, descriptive, explore crosstabs)
 - Compare means (one sample t-test, independent samples t-test, paired samples t-test and one-way ANOVA)
 - Reliability
 - Linear regression
 - Non-parametric statistics (Chi-square and binomial).

- (vii) **Utilities:** In Utilities menu, the user can view information about variables in the data file (searching among variable from a list is also enabled) and also data file comments.
- (viii) **Windows:** The Windows menu allows you to arrange, select and control the attributes of the various windows. Using this menu, you can move efficiently between data, syntax and output windows.
- (ix) **Help:** The Help menu allows you to access information link on how to use the various features of PSPP, and also information about the author and license.

2. Menus in PSPPIRE Output Viewer: The main menu bar in the PSPPORE Output Viewer contains four menus, namely File, Edit, Windows and Help.

- (i) **File:** Data exporting is allowed in PSPP. **File** → **Print** prints the file in the Output Viewer window.
- (ii) **Edit:** The Edit menu allows to copy text from the output and to select all text or data in the Output Viewer window.
- (iii) **Windows and Help:** Menu Windows and Help in the Output Viewer window is the same function with the menu in the Data Editor window.

4.4.6 Toolbar

The toolbar, located just below the menu bar, provides quick, easy access to many frequently used features. The toolbar contains tools that are available when a particular type of window is active. Each window has its own toolbar. When you put the mouse cursor on a tool in an active window, a brief description of that tool is displayed.

Tools in the Data Editor include:

1. **File Open:** Allows a particular data file to be opened for analysis.
2. **File Saves:** Saves the file in the active window.
3. **Go to Variable:** Allows the typing in of the variable name you want to go to and it will be found in the data file.
4. **Go to Case:** Allows the typing in of the number of the case you want to go to and it will be found in the data file.
5. **Find:** Allows data to be found easily within the Data Editor.
6. **Insert Cases:** Inserts a case above the case containing the active cell.
7. **Insert Variable:** Inserts a variable to the left of the variable containing the active cell.
8. **Split File:** Splits the data file into separate groups for analysis based on the values of one or more grouping variables.
9. **Weights cases:** Gives cases different weights for statistical analysis.
10. **Value labels:** Allows switching between values in the Data Editor.

4.4.7 Creating a Dataset

Step 1: Entering Variables: Upon opening PSPP, you will see a blank data table. Because no variables have yet been defined, your column headings will all read “var,” while the row headings will be ordered numerically.

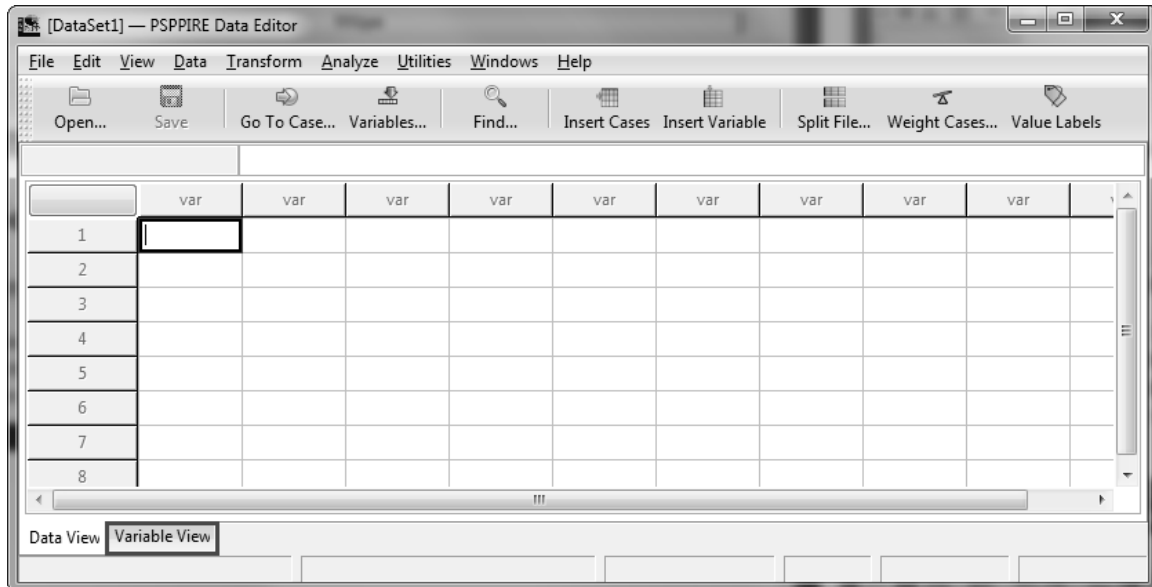


Fig. 4.7.1

In order to enter the variables for your dataset, you will need to switch from Data View to Variable View. You can do this by locating the **Variable View** tab at the bottom left-hand corner of your screen, and clicking on it.

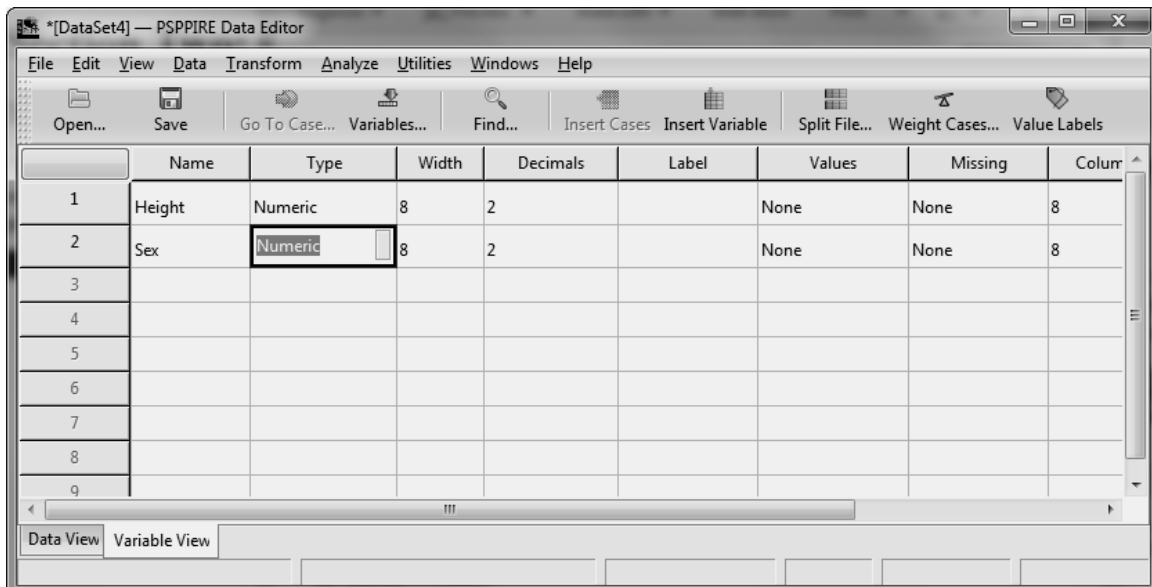


Fig. 4.7.2

Once you are in Variable View, you will be able to enter the variables for your dataset. You can enter names of the variables under the **Name** column, and press **Tab**. This will cause the rest of the fields to propagate in accordance with default settings.

For the purpose of this tutorial, you will create a simple dataset with two variables: the first variable will be (1) **height**, and the second (2) **sex**. Click on the first cell under the **Name** column, and enter “Height.” Press **Tab**, and then click on the second cell under the **Name** column, entering the variable “Height.” Press **Tab** again. Now, the names of your variables have been successfully entered.

Step 2: Assigning Value Labels for Variables: Now, you may enter the value labels for your variables. There are two values under the variable “sex,” namely “male” and “female.” In assigning a value label, as the name implies, we are simply associating a label (*i.e.*, “male” or “female”) to a particular value (*i.e.*, “1” or “2”). This is similar to how such values might appear on a survey:

1. Height (enter in inches): _____
2. Sex (circle “1” or “2”):
 - (i) Male
 - (ii) Female

In the above example, under the second (2) variable “Sex,” the value “1” is assigned the value label “Male,” while the value “2” is assigned the value label “Female”.

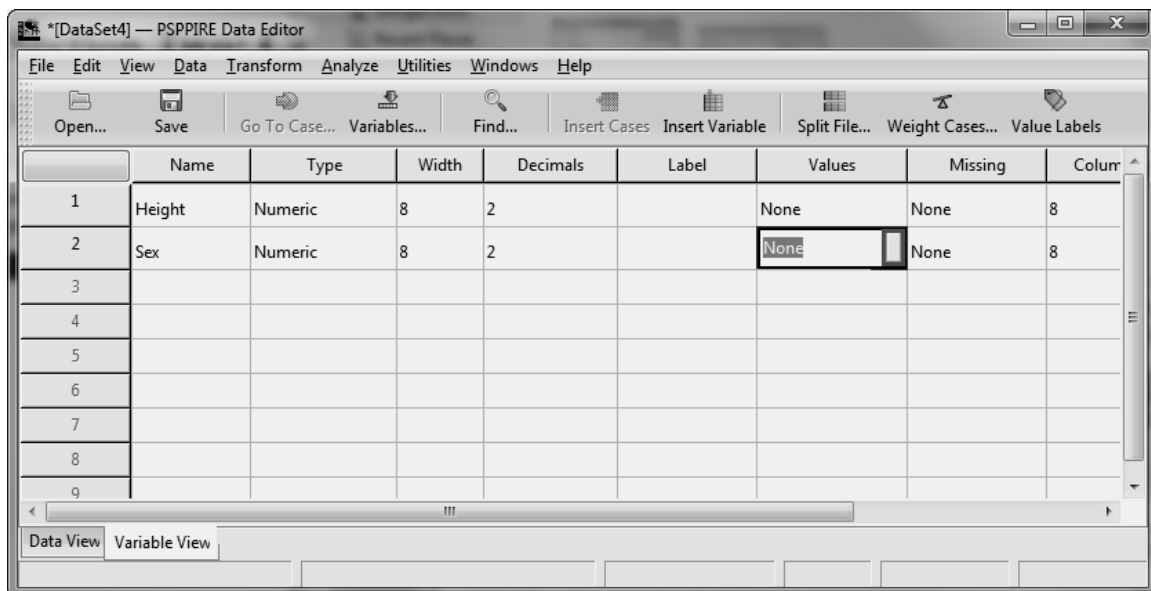


Fig. 4.7.3

To assign values for a particular variable, click on the cell under the **Values** column, and click on the grey box in the right-hand corner of the cell.

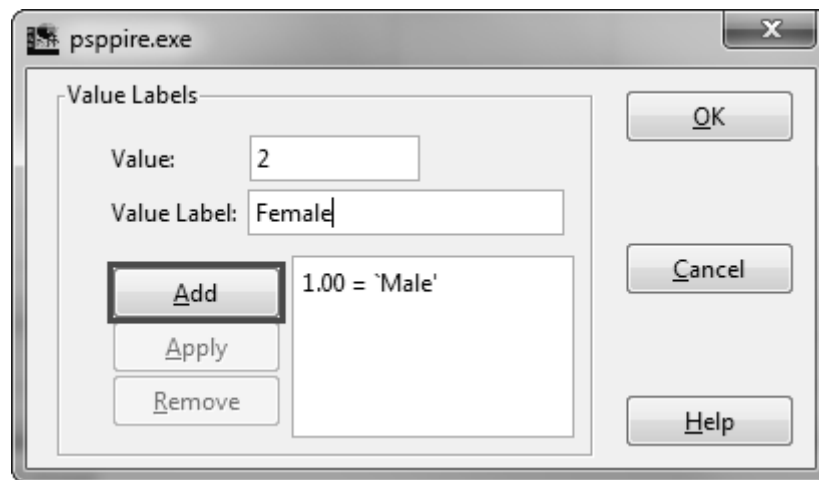


Fig. 4.7.4

A window will appear, prompting you to enter the value and the value label. For the first value label, enter “1” in the value field, and then the first label (in this case “Male”) in the value label field. For the second value label, enter “2” in the value field, entering the second label (in this case “Female”) in the value label field.



Fig. 4.7.5

After entering all your value labels, click OK. After returning to the main screen, your Variable View should look like Fig. 4.7.6.

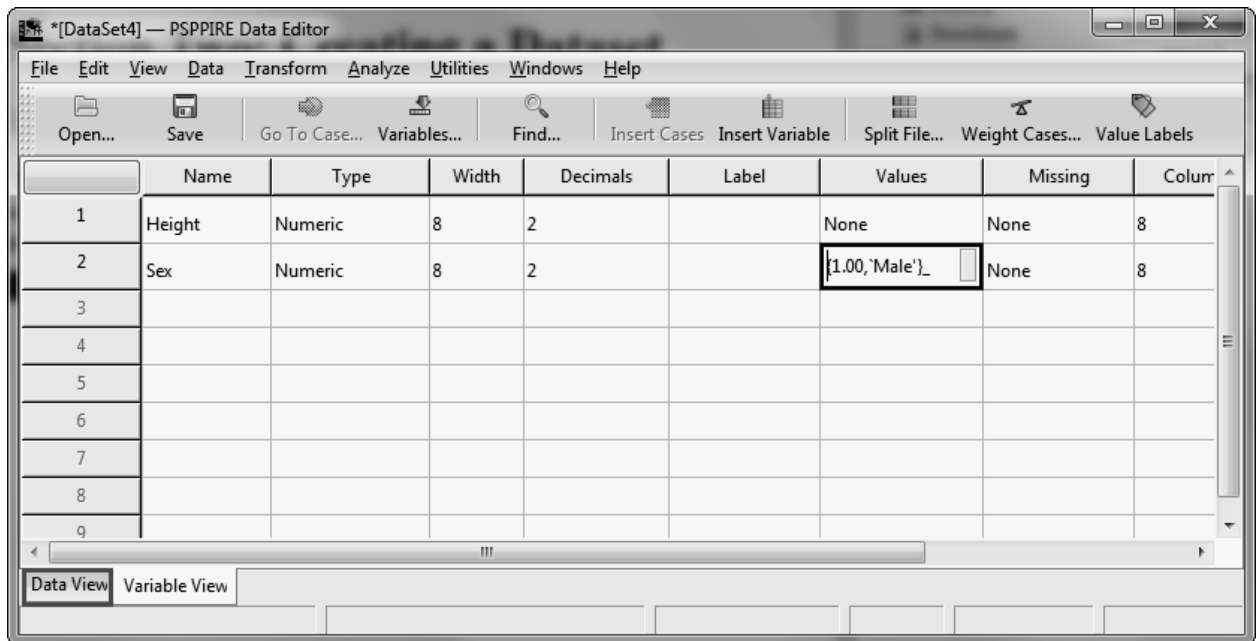


Fig. 4.7.6

Now that the value labels for the variable “Sex” have been properly assigned, we can begin entering our data by returning to the Data View. You can do this by clicking on the Data View tab at the bottom left-hand corner of your screen.

Step 3: Entering Data: After clicking on the Data View tab, you will return to the same screen that appeared when you first opened the program, except now instead of the first and second columns having the heading “var,” they will have as headings “Height” and “Sex” respectively.

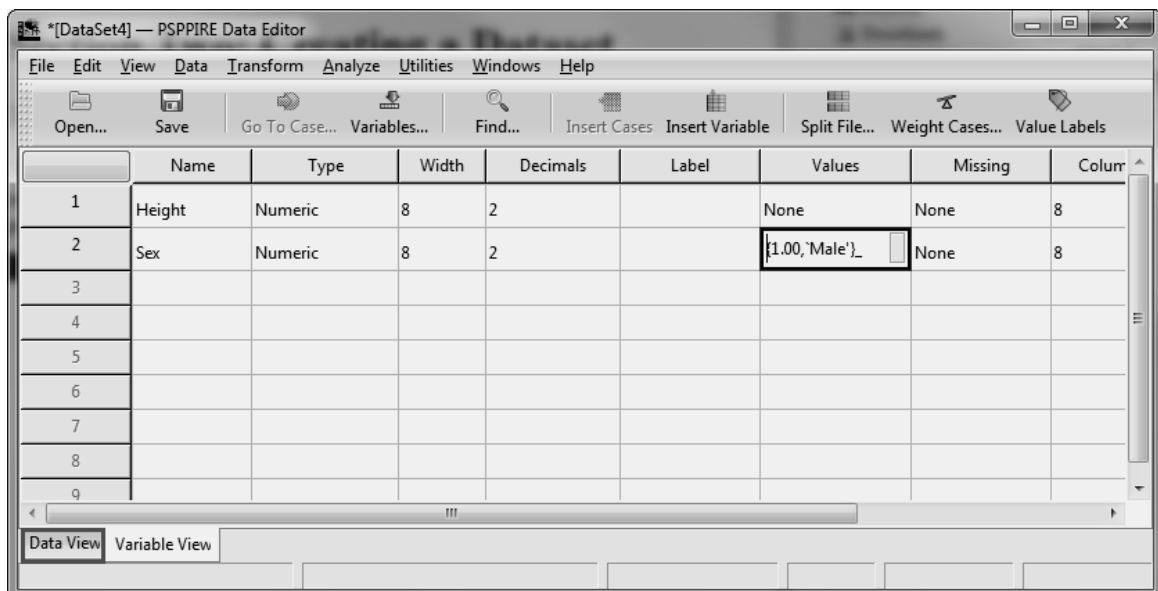


Fig. 4.7.7

Now, you can begin entering your data. In this example dataset, we are entering our data in inches under the first column, and the values “1” and “2” for “Male” and “Female” under the second column.

Step 4: Saving a Dataset: Now that you have entered your data, you can save your dataset.

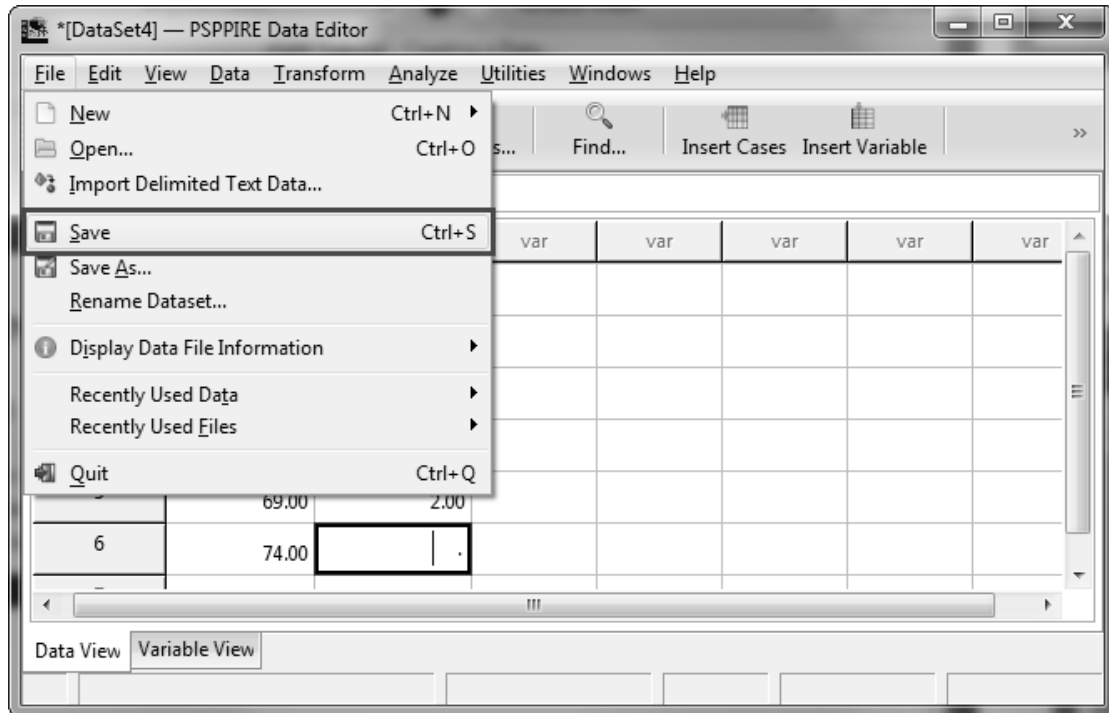


Fig. 4.7.8

To save, open the **File** menu and click on **Save**.

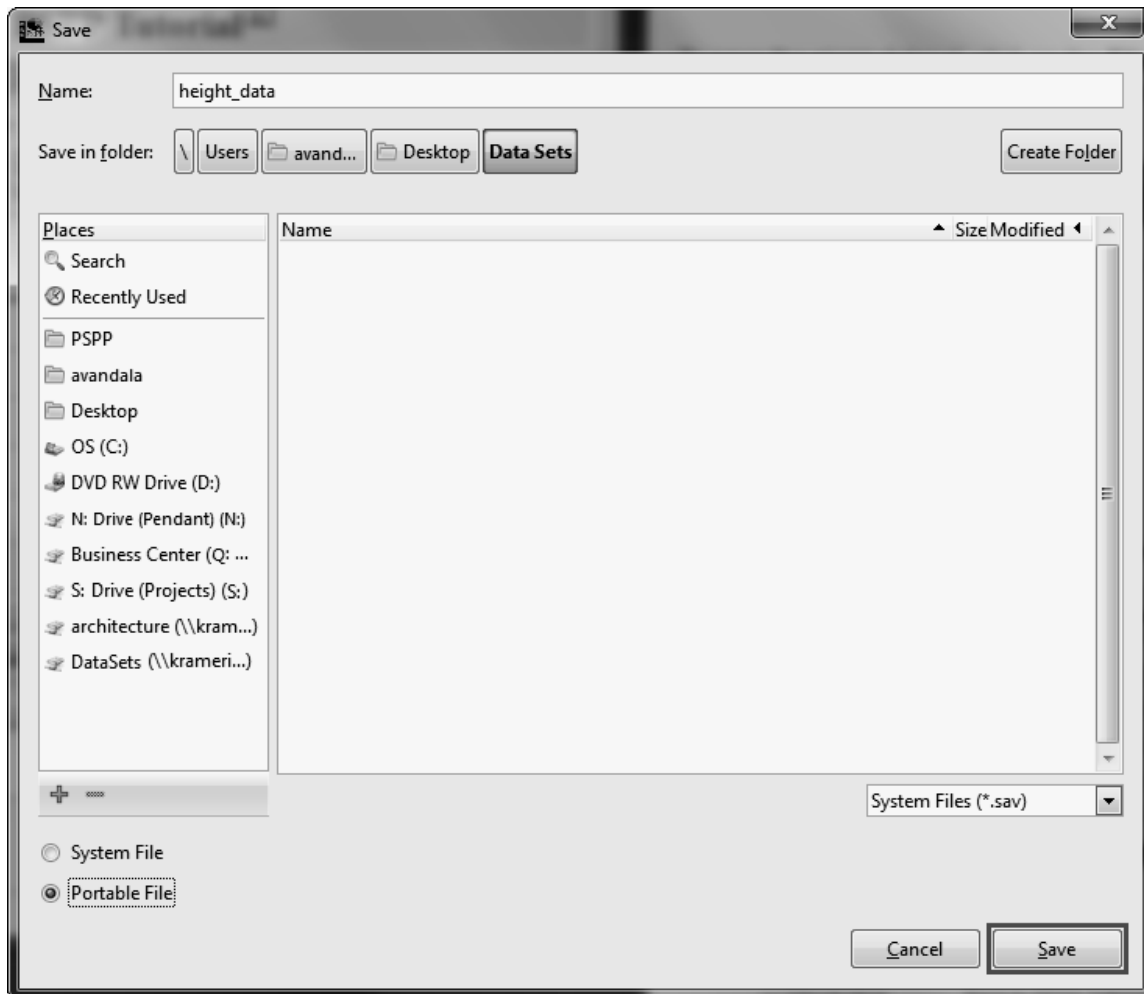


Fig. 4.7.9

Navigate to the location where you want to save your dataset, and then click on the **Save** button.

4.4.8 Plotting Scatter Plot in PSPP

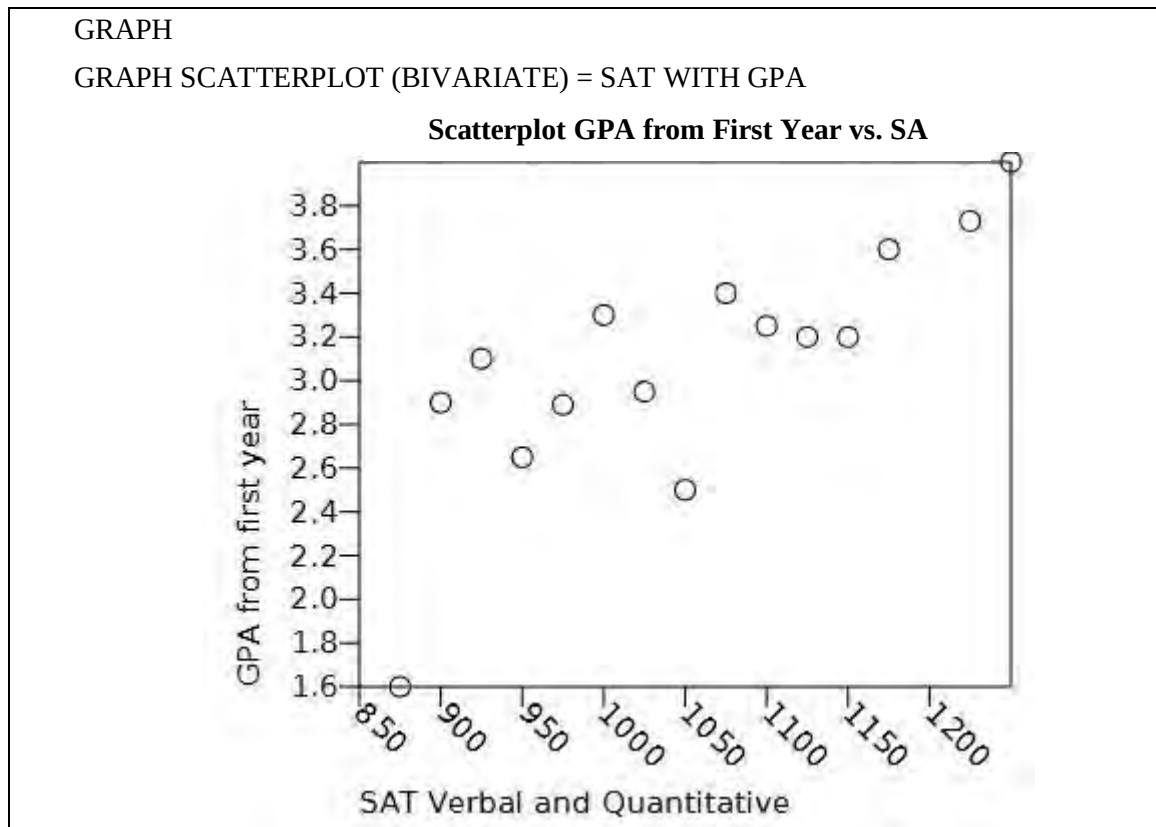
A descriptive graphing approach to examining relationships between variables is the scatterplot. This graph is made from data where each person has two measured variables. This kind of data will be the commonly used textbook example for screening college applicants. Each student has a standardized test score (SAT, ACT, etc.) and grade point average from their first year of college. The relationships from these past institutional data can be used to help screen new applicants who are applying to college.

The example data file has fifteen students, with a SAT score and a first year college GPA for each student. This data file can be downloaded as relationship Data Save.

Start by going to the Graphs menu and then select the Scatterplot command.

The Scatterplot dialog box has fields for the *X* and *Y* variables that must be selected. It is a matter of judgement about which variable should be *X* or *Y*. A useful guide is to select the potential causal variable or first variable for the *X*-axis. For scatterplot, we will select SAT because this test was taken before the first year grades were available.

The scatterplot is displayed in the output viewer, like this:



Each individual is represented by one data point. The intersection of the *X* and the *Y* variable for each person is represented by a dot. The pattern in this example tends to move from the lower left to the upper right. This shows a positive relationship in which people with higher SAT scores tend to have higher GPA scores. Both variables tend to increase together, making this a positive or direct relationship.

4.4.9 Conducting Correlation in PSPP

The most common analysis for relationships are to conduct correlation coefficients. This will give us a score between -1.0 and $+1.0$. The score sign tells the kind of relationship with positive relationships, meaning that both variables increase together. Negative relationships occur in situations where an increase in one variable is connected to a decrease in the other variable. In other words, the variables are headed in opposite directions. The numerical value of the correlation coefficient indicates stronger relationships as it gets farther from zero, in either the

positive or negative direction. The PSPP feature for doing simple correlation coefficients is Bivariate Correlation. The prefix “bi” refers to two. So, this analysis will be correlations between two variables. This command will compute Pearson’s r , which is the most commonly used correlation coefficient.

Like most analyses, we have to choose the variables that we want to correlate. Drag these variables to the open field on the right. It is possible to add more than two variables. PSPP will make a table that will calculate all of the possible correlations between the variables.

The “test of significance” and “flag” features are for computing the statistical significance of these variables. A two-tailed test is for situations with open-ended predictions. One-tailed test is for situations with specific predictions.

The output has a table that shows the correlation coefficients for all of the possible comparisons between the selected variables. The intersections of the rows and columns show the comparisons for the variables. Pearson r for the relationship between GPA and SAT in this example is $+.78$, which is strong for behavioural data. The p -value is in the “sig. (2-tailed)” row with “sig.” being short for statistical significance.

CORRELATIONS			
CORRELATION			
/VARIABLES = SAT GPA			
/PRINT = TWO TAIL SIG.			
Correlations			
		SAT Verbal and Quantitative	GPA from First Year
SAT Verbal and Quantitative	Pearson Correlation	1.00	.78
	Sig. (2-tailed)		0.01
	N	15	15
GPA from First Year	Pearson Correlation	.78	1.00
	Sig. (2-tailed)	0.001	
	N	15	15

Notice that each variable has a correlation with itself of $+1.0$, a perfect relationship. This is not a useful analysis. So, just ignore it.

APA style has specific recommendations for reporting statistics for correlations. The format is $r(\text{degrees of freedom}) = r$ score, $p =$ probability. The degrees of freedom for correlations are the total number of score pairs (N) minus two.

For this example, the degrees of freedom would be $df = 15 - 2 = 13$.

Also not that r and p are in italic font. The probability is shown in the sig. (2-tailed) field of the output. Altogether, the proper APA style for reporting this result is $r(13) = .78, p = .001$.

4.5 SOFTWARE FOR STATISTICAL ANALYSIS 2 – R AND R COMMANDER

4.5.1 Introduction

R Commander is a graphical user interface for R, a statistical environment comparable to SPSS or SAS. R and R Commander are shareware applications.

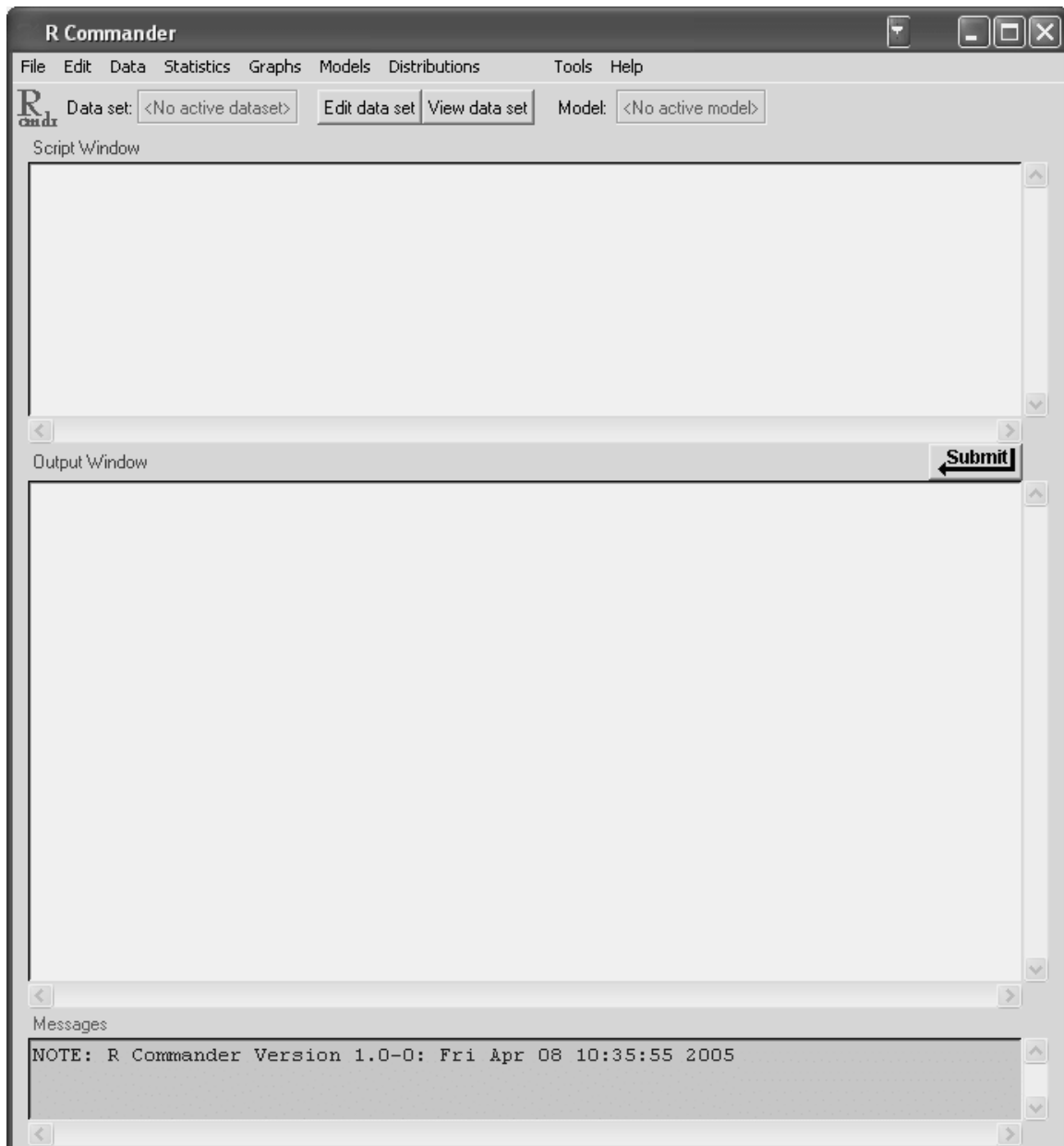
R Commander is a basic graphical user interface (GUI) for R. It provides a series of menus that allow running lots of statistic tests and creating graphics without typing a line of code.

R is highly capable, free open-source statistical software. Although it is hard to know with any certainty how many people use R, it is – for example, judging by Internet traffic – possibly the most popular statistical software in the world. R is in any event, very widely used, and its use is growing rapidly.

4.5.2 Starting R Commander

Once R is running, simply loading the Rcmdr package by typing the command library (“Rcmdr”) into the R Console starts the R Commander graphical user interface (GUI). To function properly under Windows, the R Commander requires the single-document interface (SDI) to R.

After loading the package, R Commander window should appear more or less as in Fig. 4.7.10.

**Fig. 4.7.10**

The *R Commander* and *R Console* windows float freely on the desktop. You will normally use the menus and dialog boxes of the *R Commander* to read, manipulate, and analyze data.

1. *R* commands generated by the *R Commander* GUI appear in the upper window (labelled Script window).
2. Printed output appears by default in the second text window (labelled Output window).

3. The lower, gray window (labelled Messages Window) displays error messages, warnings, and some other information (“notes”), such as the start-up message in Fig. 4.7.10.
4. When you create graphs, these will appear in a separate Graphics Device Window, outside of the main *R* Commander Window.

There are several menus along the top of the *R* Commander window:

1. **File:** Menu items for loading and saving script files; for saving output and the *R* workspace; and for exiting.
2. **Edit:** Menu items (Cut, Copy, Paste, etc.) for editing the contents of the scripts and output windows. Right clicking in the script or output window also brings up an edit “context” menu.
3. **Data:** Sub-menus containing menu items for reading and manipulating data.
4. **Statistics:** Sub-menus containing menu items for a variety of basic statistical analyses.
5. **Graphs:** Menu items for creating simple statistical graphs.
6. **Models:** Menu items and sub-menus for obtaining numerical summaries, confidence intervals, hypothesis tests, diagnostics, and graphs for a statistical model, and for adding diagnostic quantities, such as residuals, to the dataset.
7. **Distributions:** Cumulative probabilities, probability densities or masses, quantiles, and graphs of standard statistical distributions (to be used, for example, as a substitute for statistical tables) and samples from these distributions.
8. **Tools:** Menu items for loading *R Packages* unrelated to the *Rcmdr Package* (e.g., to access data saved in another (package), and for setting some options.
9. **Help:** Menu items to obtain information about the *R Commander*. As well each *R Commander* dialog box has a *Help* button.

4.5.3 Advantages of Free OSS over Closed Software

1. **Flexibility and Agility (Quickness):** IT leaders must fundamentally provide flexibility and agility for their enterprise. If you cannot compete on agility, you are going to get left behind by the competition. Open source enables technology agility, typically offering multiple ways to solve problems. Open source helps keep your organization from getting blocked because a particular capability is not available from a vendor. Instead of waiting for the vendor to deliver that capability, you can create it for yourself.
2. **Speed:** Open source enables speed. A great advantage of open source is the ability to take the community versions, get started, understand whether they can solve business problem, and begin to deliver value right away. Once to make that determination, professional support and services are increasingly available for Open source products.

3. **Cost-effectiveness:** Open source is generally much more cost-effective than a proprietary solution. Not only are open source solutions typically much more inexpensive in an enterprise environment for equivalent or superior capability but they also give enterprises the ability to start small scale. Given that enterprises are often budget challenged, it just makes financial sense to explore open source solutions.
4. **Ability to Start Small:** With Open source, you can start small and quickly with community versions, and then migrate to a commercially-supported solution as your business requirements drive you there. If the project does not require support, you can continue on the community version indefinitely. You have the option to try the various alternatives, pick the one that's going to work, and then scale up with a commercial solution.
5. **Solid Information Security:** Commercial open source has a solid information security record in a dangerous world. Obviously, it is difficult to claim security superiority for any solution and it is a challenging environment for all of us, but the responsiveness of the open source community and vendors relative to information security problems has been very good.
6. **Attract Better Talent:** Open source gives enterprises the ability to attract better talent. Most professional technologists are well aware of open source and many believe it is where the industry is headed. Many enjoy creating their own projects and having the ability to interact with others outside their enterprise to develop solutions. Giving developers flexibility and freedom can be an important tool in attracting better talent.

4.6 REGRESSION ANALYSIS

4.6.1 Meaning and Definition

Regression analysis is a statistical tool used to calculate a continuous dependent variable from various independent variables and is commonly used for prediction and forecasting. For example, if the agriculturist knows that the yield of rice and rainfall are closely related, then he will want to know the amount of rain required to achieve a certain production. In this situation, he will use the regression analysis.

This technique is used for modeling and analysis of mathematical data using the value of a dependent variable and one or more independent variables. Dependent variable, also known as response variable, is the single variable predicted by the regression model. Independent variable, also known as predictor variable, is used to predict the values of dependent variable. Regression analysis is the process to calculate the unknown value of one variable from the known values of the other variable.

According to **Blair**, "Regression is the measure of the average relationship between two or more variable in terms of the original units of the data."

According to **Taro Yamane**, “One of the most frequently used techniques in economics and business research, to find a relation between two or more variables that are related causally, is regression analysis.”

Regression analysis makes an equation to describe the statistical relationship between one or more predictor variables and the response variable. It helps to determine the nature and strength of a relationship between independent and dependent variables. Regression may be classified as follows:

1. Simple Regression
2. Multiple Regressions
3. Linear Regression
4. Non-linear Regression

4.6.2 Assumptions of Regression Analysis

We make some assumptions (which are related to the bivariate population) to make valid statistical inference using regression analysis.

Applications of simple linear regression models are based upon these assumptions. These assumptions are:

1. There exists a linear relationship between independent variable (x) and dependent variable (Y). The general linear regression equation $Y = a + bx + c$ is used to describe the relationship between the variables x and y and its desired value for a given value of x .
2. There exists an expected value of variable y and these are normally distributed, for each value of variable y and these are normally distributed values will be located on the regression line.
3. The values of variable x are fixed (not be random) while the values of variable y is a continuous random variable.
4. It is assumed that sampling error associated with mean values of y be an independent random variable and distributed normally with mean constant standard deviation and zero.
5. For every value of x lying within the sample data range, the standard deviation and variance of expected values of y about the regression line are constant.
6. The value of y cannot be determined for the value of x (which is outside the boundary of the sample data).

4.6.3 Applications of Regression Analysis

Following are the basic uses of regression analysis:

1. It is used to know the relationship between different (one or more) variables.
2. It is used to find out the value of dependent variable from the value of independent variables.

3. It is used to find out the coefficients of correlation (r) and coefficients of determination.
4. In corporate sector, it is useful to check the quality.
5. It is also very useful to determine the statistical curve (demand, supply, etc.).

4.6.4 Difference between Correlation and Regression Analysis

Table 4.4 shows the difference between Correlation and Regression Analysis:

Table 4.4: Difference between Correlation and Regression Analysis

<i>Basis</i>	<i>Correlation</i>	<i>Regression</i>
Meaning	It is the study of the linear relationship between two variables.	It is a statistical tool used to calculate a continuous dependent variable from various independent variables.
Variables	Random variables are used.	One variable is random variable and other is fixed variable. It may be possible that sometimes both the variables are random.
Relationship	It does not indicate the cause-and-effect relationship between variables and gives variables	

4.6.5 Regression Lines

The lines of the best fit between two variables and stating the common average relationship are known as regression lines. There are two regression lines for two variables and Y as independent variables. The straight lines is called as the regression line of x and y and vice versa.

The regression line of Y on X gives the finest possible values of Y for the given values of X . So, this line is the sum of squares of deviations of the 'calculated values of Y ' and 'observed values of Y is minimum', *i.e.*,

$$\Sigma(Y_c - Y_0)^2 = \text{minimum}$$

where, Y_c = Calculated value of Y

Y_0 = Observed value of Y

Now, for the regression line of X on Y :

$$\Sigma(X_c - X_0)^2 = \text{minimum}$$

where, X_c = Calculated value of X

X_0 = Observed value of X

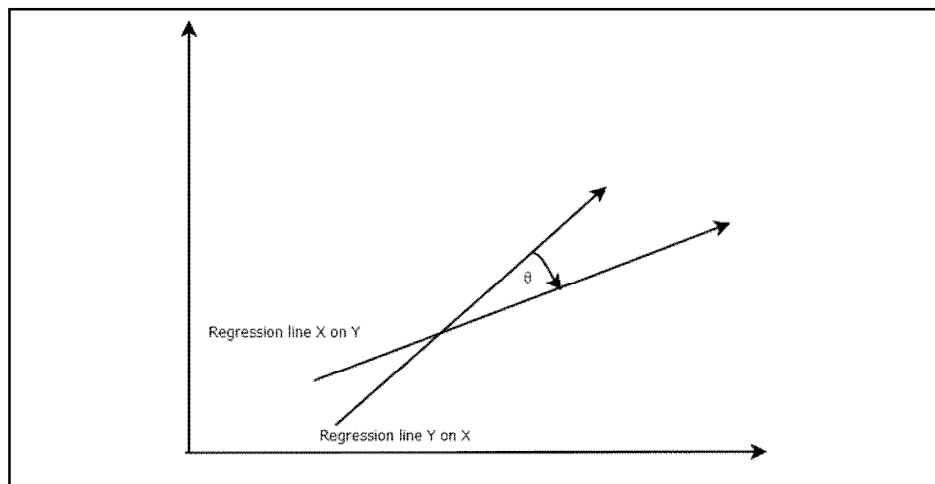


Fig. 4.8: Regression Lines

This is why, to understand the concept of regression line, two regression lines are required. If there is only one regression line existing between the two variables, then it means $r = \pm 1$. Now, if the value of r is 0, then both the variables are independent and the two lines will cut each other at a right angle (i.e., the two lines will be parallel to X-axis and Y-axis) (Fig. 4.8).

4.6.6 Regression Equations

Regression equation is used to show the regression line and there are two regression equations because there are two regression lines.

So, if there are two variables, say X and Y , then the two regression equations are written as:

1. **Regression Equation of X on Y :** The regression equation of ' X on Y ' is written as follows:

$$X = a + bY$$

To calculate the values of a and b , the following normal equations are solved simultaneously:

$$\Sigma X = Na + b\Sigma Y^2$$

$$\Sigma XY = a\Sigma Y + b\Sigma Y^2$$

2. **Regression Equation of Y on X :** The regression equation of ' Y on X ' is written as follows:

$$Y = a + bX$$

To get the values of a and b , the following normal equations are to be solved simultaneously:

$$\Sigma Y = Na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

In the normal equation, ΣX , ΣY , ΣXY , ΣX^2 and ΣY^2 gives the summations and can be determined from the observed pairs of values of two variables 'X and Y'. N is the number of observed pairs of values.

4.6.7 Regression Coefficient

If 'b' is the slope of regression line of 'Y on X', then b_{yx} is known as coefficient of regression of 'Y on X'.

For the regression line of 'X on Y',

$$\text{So, } b_{yx} = \text{Regression Coefficient of Y on X} = r \frac{\sigma_Y}{\sigma_X}$$

Similarly, for the regression line of 'X on Y',

$$b_{xy} = \text{Regression Coefficient of X on Y} = r \frac{\sigma_X}{\sigma_Y}$$

4.6.8 Properties of Regression Coefficients

1. Coefficient of correlation is calculated by determining the geometric mean of the two regression coefficients, i.e.,

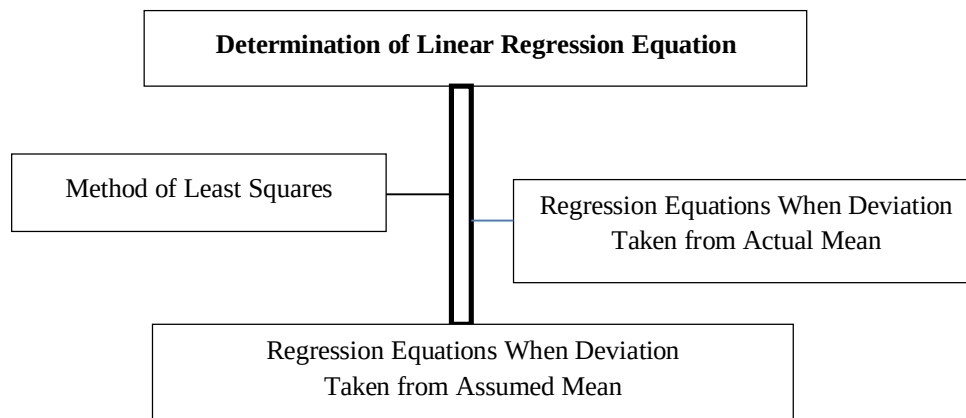
$$r = \pm \sqrt{b_{xy} \times b_{yx}}$$

where, r = coefficient of correlation and its value is positive when both the regression coefficients are positive. Since r lies between +1 to -1, the value of regression coefficient is not more than one.

2. It is independent from origin but not from scale.
3. If one regression coefficient is greater than unity, then remaining must be smaller than unity.
4. Both the regression coefficients have the same sign (i.e., if one is positive, then the other is also positive and if one is negative, then the other is also negative).
5. The value of coefficient of correlation is less than the mean value of regression coefficients.

4.6.9 Interpretation of the Output of Simple Liner Regression Analysis

To determine the parameters of a fitted regression equation, the following methods are used:



4.6.9.1 Method of Least Squares Method

(1) Let $X + a + bY$ be the line of X on Y . To determine the values of a and b , the following two normal equations are to be solved simultaneously:

$$\Sigma X = Na + b\Sigma Y$$

$$\Sigma XY = a\Sigma Y + b\Sigma Y^2$$

Example 11: Consider the following data to obtain the two regression equations:

X	5	3	10	8	9
Y	8	12	6	5	14

Solution: Calculation of Regression Equations

X	Y	X^2	Y^2	XY
5	8	25	64	40
3	12	9	144	36
10	6	100	36	60
8	5	64	25	40
9	14	81	196	126
$\Sigma X = 35$	$\Sigma Y = 45$	$\Sigma X^2 = 279$	$\Sigma Y^2 = 465$	$\Sigma XY = 302$

1. Regression equation of X on Y : $X = a + bY$ and the two normal equations are:

$$\Sigma X = Na + b\Sigma Y$$

$$\Sigma XY = a\Sigma Y + b\Sigma Y^2$$

Substituting the values, we get

$$35 = 5a + 45b \quad \dots (1)$$

$$302 = 45a + 465b \quad \dots (2)$$

Multiplying equation (1) by 9, we get

$$315 = 45a + 405b \quad \dots (3)$$

Subtracting equation (2) from (3), we get

$$13 = -60b$$

$$\text{or } b = -0.22$$

Substituting the value of b in equation (1), we get

$$35 = 5a + 45(-0.22)$$

$$\Rightarrow 35 = 5a - 9.9$$

$$\Rightarrow 5a = 35 + 9.9 = 44.9$$

$$\therefore a = 8.98$$

Putting the value of a and b in the equation, we get

The regression equation of X on Y :

$$X = 8.98 - 0.22Y$$

2. Let $Y = a + bX$ be the line of Y on X . To determine the values of a and b . The following two normal equations are to be solved simultaneously:

$$\Sigma X + Na + b\Sigma$$

$$\Sigma XY = a\Sigma Y + b\Sigma Y^2$$

Substituting the values, we get

$$45 = 5a + 35b \quad \dots (1)$$

$$302 = 35a + 279b \quad \dots (2)$$

Multiplying equation (1) by 7, we get

$$315 = 35a + 245b \quad \dots (3)$$

Substituting the value of b in equation (1), we get

$$45 = 5a + 35(-0.38)$$

$$5a = 45 + 13.3 = 58.3 \text{ or } a = 11.66$$

Putting the values of a and b in the equation, we get

Regression equation of Y on X :

$$Y = 11.66 - 0.38X$$

4.6.9.2 Regression Equations When Deviation Taken from Actual Mean

Following formulae are used to find out the deviation of X and Y series when deviation is taken from the actual mean.

1. Regression Equation of X on Y :

$$X - \bar{X} = r \frac{\sigma_X}{\sigma_Y} (Y - \bar{Y})$$

where, $\bar{X}$ = Mean of X Series and $\bar{Y}$ = Mean of Y Series

$$r \frac{\sigma_X}{\sigma_Y} = \text{Regression Coefficient of } X \text{ on } Y \text{ or } b_{xy}$$

$$b_{xy} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\Sigma(Y - \bar{Y})^2} = \frac{\Sigma xy}{\Sigma y^2} \text{ or } \frac{\Sigma x - \frac{(\Sigma x)(\Sigma y)}{N}}{\Sigma y - \frac{(\Sigma y)^2}{N}}$$

where, $x = (X - \bar{X})$; $y = (Y - \bar{Y})$

2. Regression Equation of Y on X

$$Y - \bar{Y} = r \frac{\sigma_Y}{\sigma_X} (X - \bar{X})$$

where, $\bar{X}$ = Mean of X Series and $\bar{Y}$ = Mean of Y Series

$$r \frac{\sigma_X}{\sigma_Y} = \text{Regression Coefficient of } X \text{ on } Y \text{ or } b_{yx}$$

$$b_{yx} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\Sigma(X - \bar{X})^2} = \frac{\Sigma xy}{\Sigma x^2} \text{ or } \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{N}}{\Sigma x^2 - \frac{(\Sigma x)^2}{N}}$$

where, $x = (X - \bar{X})$; $y = (Y - \bar{Y})$

Example 12: Using the following data, obtain two regression equations.

X	15	19	24	21	25	22	16	20	27	21
Y	31	36	48	37	50	45	33	41	39	20

Solution: Calculation of Regression Equations

X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	y^2	xy
15	31	-6	-7	36	49	42
19	36	-2	-2	4	4	4
24	48	3	10	9	100	30
21	37	0	-1	0	1	0
25	50	4	12	16	144	48
22	45	1	7	1	49	7
16	33	-5	-5	25	25	25
20	41	-1	3	1	9	-3
27	39	6	1	36	1	6
21	20	0	-18	0	324	0
$\Sigma X = 210$	$\Sigma Y = 380$	$\Sigma x = 0$	$\Sigma y = 0$	$\Sigma x^2 = 128$	$\Sigma y^2 = 706$	$\Sigma xy = 1$

$$\bar{X} = \frac{210}{10} = 21 \text{ and } \bar{Y} = \frac{380}{10} = 38$$

$$b_{xy} = \frac{\Sigma xy}{\Sigma y^2} = \frac{159}{706} = 0.225;$$

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{159}{128} = 1.242$$

Regression Equation of X on Y is given by:

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

$$\Rightarrow X - 21 = 0.225(Y - 38)$$

$$\Rightarrow X = 0.225Y - 38 \times 0.225 + 21$$

$$\Rightarrow X = 0.225Y - 8.55 + 21$$

$$\Rightarrow X = 12.45 + 0.225Y$$

Regression Equation of Y on X is given by:

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

$$\Rightarrow Y - 38 = 1.242(X - 21)$$

$$\Rightarrow Y = 1.242X - 21 \times 1.242 + 38$$

$$\Rightarrow Y = 1.242X - 26.1 + 38$$

$$\Rightarrow Y = 11.9 + 1.242X$$

4.6.9.3 Regression Equations when Deviation Taken from Assumed Mean

If we take the deviation from the assumed mean, then the following formulae are used to solve the regression equation.

Note: Assumed mean is taken because value of actual mean is in the fraction form.

Regression Equation of X on Y is:

$$X - \bar{X} = r \frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$$

Now, find out the value of $r \frac{\sigma_x}{\sigma_y}$.

By applying the following formula, we get

$$b_{xy} \text{ or } r \frac{\sigma_x}{\sigma_y} = \frac{\Sigma dx dy - \frac{\Sigma dx \times \Sigma dy}{N}}{\Sigma dy^2 - \frac{(\Sigma dy)^2}{N}} \text{ or } \frac{N \Sigma dx dy - \Sigma dx \times \Sigma dy}{N \Sigma dy^2 - (\Sigma dy)^2}$$

Regression Equations of Y on X is:

$$Y - \bar{Y} = r \frac{\sigma_X}{\sigma_Y} (X - \bar{X})$$

Now, find out the value of $r \frac{\sigma_Y}{\sigma_X}$ by applying the following formula:

$$b_{yx} \text{ or } r \frac{\sigma_Y}{\sigma_X} = \frac{\sum dxdy - \frac{\sum dx \times \sum dy}{N}}{\sum dx^2 - \frac{(\sum dx)^2}{N}} \text{ or } \frac{N\sum dxdy - \sum dx \times \sum dy}{N\sum dx^2 - (\sum dx)^2}$$

where, $d_x = X - A_x$; $d_y = Y - A_y$

Example 13: From the following data, find the regression equations, and estimate the likely value of Y when X = 100.

X	74	98	76	82	58	76	94	88	51
Y	124	131	117	131	96	121	136	97	85

Solution: Let assume that the mean for X-series is 80 and for Y-series is 120.

X	Y	$d_x = (X - A_x)$	$d_y = (Y - A_y)$	$dxdy$	dx^2	dy^2
74	124	-6	4	-24	36	16
98	131	18	11	198	324	121
76	117	-4	-3	12	16	9
82	131	2	11	22	4	121
58	96	-22	-24	528	484	576
76	121	-4	1	-4	16	1
94	136	14	16	224	196	256
88	97	8	-23	-184	64	529
51	85	-29	-35	1015	841	1225
$\Sigma X = 697$	$\Sigma Y = 1038$	$\Sigma dx = -23$	$\Sigma dy = -42$	$\Sigma dxdy = 1787$	$\Sigma dx^2 = 1981$	$\Sigma dy^2 = 2854$

$$\bar{X} = \frac{697}{9} = 77.44 \text{ and } \bar{Y} = \frac{1038}{9} = 115.33$$

$$b_{xy} = \frac{N\sum dxdy - \sum dx \sum dy}{N\sum dy^2 - (\sum dy)^2} = \frac{9 \times 1787 - (-23)(-42)}{9 \times 2854 - (-42)^2} = \frac{16083 - 966}{17829 - 529} = \frac{15117}{17300} = 0.874$$

$$b_{yx} = \frac{N\sum xy - \sum dx \sum dy}{N\sum dx^2 - (\sum dx)^2} = \frac{9 \times 1787 - (-23)(-42)}{9 \times 1981 - (-23)^2} = \frac{16083 - 966}{17829 - 529} = \frac{15117}{17300} = 0.874$$

Regression equation of X on Y is given by:

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

$$\Rightarrow X - 77.44 = 0.632(Y - 115.33)$$

$$\Rightarrow X - 77.44 = 0.632Y - 115.33 \times 0.632$$

$$\Rightarrow X - 77.44 = 0.632Y - 72.88$$

$$\Rightarrow X = 0.632Y - 72.88 + 77.44$$

$$\Rightarrow X = 4.56 + 0.632Y$$

When $X = 100$,

$$Y = 47.65 + 0.874 \times 100 = 47.65 + 87.4 = 135.05$$

Regression equation of Y on X is given by:

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

$$\Rightarrow Y - 115.33 = 0.874(X - 77.44)$$

$$\Rightarrow Y - 115.33 = 0.874X - 77.44 \times 0.874$$

$$\Rightarrow Y - 115.33 = 0.874X - 67.68$$

$$\Rightarrow Y = 0.874X - 67.68 + 115.33$$

$$\Rightarrow Y = 47.65 + 0.874X$$

4.6.10 Relationship between Correlation and Regression Coefficients

The correlation coefficient is the geometric mean between the regression coefficients, i.e.,

$$\begin{aligned} r^2 &= b_{yx} \times b_{xy} \\ \Rightarrow r &= \sqrt{b_{yx} \times b_{xy}} \end{aligned} \quad \dots(1)$$

Let us suppose that the two lines of regression are:

$$a_1X + b_1Y + c_1 = 0 \quad \dots(2)$$

$$a_2X + b_2Y + c_2 = 0 \quad \dots(3)$$

Let us suppose that equation (2) is the line of regression of Y on X and equation (3) is the line of regression of regression of X on Y.

From equation (2), we get

$$\text{or } a_1X + b_1Y + c_1 = 0$$

$$b_1Y = -a_1X - c_1$$

$$\text{or } Y = \frac{-a_1}{b_1}X - \frac{c_1}{b_1} \Rightarrow b_{yx} = \frac{-a_1}{b_1}$$

From equation (3), we get

$$\text{or } a_2X + b_2Y + c_2 = 0$$

$$a_2X = -b_2Y - c_2$$

$$\text{or } X = \frac{-b_2}{a_2}Y - \frac{c_2}{a_2} \Rightarrow b_{xy} = \frac{-b_2}{a_2}$$

The correlation coefficient r between X and Y can now be obtained by this formula:

$$r^2 = b_{xy}b_{yx} = \left(\frac{-a_1}{b_1} \right) \times \left(\frac{-b_2}{a_2} \right) = \frac{a_1b_2}{b_1a_2}$$

$$\therefore r = \pm \sqrt{\frac{a_1 b_2}{a_2 b_1}} \quad \dots(4)$$

Example 14: If the two lines of regression are $4X - 5Y + 30 = 0$ and $20X - 9Y - 107 = 0$, which of these is the line of regression of X on Y and Y on X ? Find r_{xy} and σ_y when $\sigma_x = 3$.

Solution: We are given the regression lines as:

$$4X - 5Y + 30 = 0 \quad \dots (1)$$

$$\text{and } 20X - 9Y - 107 = 0 \quad \dots (2)$$

Let (1) be the equation of the line of regression of X on Y and (2) be the equation of the line of regression of Y on X .

$$\text{From (1), we get, } X = \frac{5}{4}Y - \frac{30}{4} \Rightarrow b_{xy} = \frac{5}{4}$$

$$\text{From (2), we get } Y = \frac{20}{9}X - \frac{107}{9} \Rightarrow b_{yx} = \frac{20}{9}$$

$$\therefore r = b_{xy} = \frac{20}{9} \times \frac{5}{4} = 2.7778$$

Since $r^2 > 1$, our supposition is wrong. [$\because$ We always have $0 \leq r^2 \leq 1$.]

Hence, (1) is the line of regression of Y on X and (2) is the line of regression of X on Y .

Rewriting (1), we get

$$5Y = 4X + 30 \Rightarrow Y = \frac{4}{5}X + 6 \Rightarrow b_{yx} = \frac{4}{5}$$

Rewriting (2), we get

$$20X + 9Y + 107 \Rightarrow \bar{X} = \frac{9}{20}Y + \frac{107}{20} \Rightarrow b_{xy} = \frac{9}{20}$$

$$\therefore r = b_{yx} b_{xy} = \frac{4}{5} \times \frac{9}{20} = 0.36$$

$$\Rightarrow r = \pm \sqrt{0.36} = \pm 0.6$$

Since both the regression coefficients are positive, r must be positive. Hence, we take, $r = r_{xy} = 0.6$.

We are given, $\sigma_x = 3$

We have

$$b_{yx} = r \frac{\sigma_y}{\sigma_x} \Rightarrow \sigma_y = \frac{b_{yx} \sigma_x}{r} = \frac{4}{5} \times \frac{3}{0.6} = 4$$

Correlation Coefficient

The correlation coefficient is the under-root of the two regression coefficients, i.e.,

$$r = \sqrt{b_{xy} \times b_{yx}}$$

$$b_{xy} = 0.95; b_{yx} = 0.95$$

$$\text{Hence, } r = \sqrt{0.95 \times 0.95} = 0.95$$

Example 15: In a partially destroyed laboratory record of an analysis of correlations data, the following results are legible:

Variance of $X = 9$

Regression equations: $4x - 5y + 33 = 0$

$$22x - 15y = 108$$

Find out:

- (i) Mean values of x and y
- (ii) Standard deviation of y
- (iii) Coefficient of correlation between x and y

Solution:

(i) Calculation of Mean Values of x and y :

The given equations are:

$$4x - 5y + 33 = 0 \quad \text{or} \quad 4x - 5y = -33 \quad \dots (1)$$

$$22x - 15y = 108 \quad \dots (2)$$

Since the mean of x and y lie on the regression line, we have

$$4\bar{X} - 5\bar{Y} = -33 \quad \dots (3)$$

$$22\bar{X} - 15\bar{Y} = 108 \quad \dots (4)$$

Multiplying 3 in equation (3), we get

$$12\bar{X} - 15\bar{Y} = -99 \quad \dots (5)$$

Subtracting equation (5) from (4), we get

$$10\bar{X} = 207$$

$$\Rightarrow \bar{X} = \frac{207}{10} = 20.7$$

Putting the value of $\bar{X}$ in equation (3), we get

$$4 \times 20.7 - 5\bar{Y} = -33$$

$$\Rightarrow 82.8 - 5\bar{Y} = -33$$

$$\Rightarrow -5 - 5\bar{Y} = -33 - 828$$

$$\Rightarrow \bar{Y} = 172.2$$

Hence, mean are $\bar{X} = 20.7$ and $\bar{Y} = 172.2$.

(ii) Calculation of Coefficient of Correlation (r_{xy})

Let us assume that the first equation be regression equation of y on x .

$$4x - 5y + 33 = 0$$

$$\Rightarrow 5y = 4x + 33$$

$$\Rightarrow y = \frac{33}{5} + \frac{4}{5}x$$

$$\Rightarrow y = 6.6 + 0.8x$$

$$\therefore b_{yx} = 0.8$$

Assuming second equation as regression equation of x on y , we get

$$22x - 15y = 108$$

$$\Rightarrow 22x = 108 + 15y$$

$$\Rightarrow x = \frac{108}{22} + \frac{15}{22}y$$

$$\Rightarrow x = 4.9 + 0.682y$$

$$\therefore b_{xy} = +0.682$$

$$\text{Now, } r = \sqrt{b_{xy} \cdot b_{yx}} = \sqrt{0.682 \times 0.8} = +0.74$$

(iii) Calculation of σ_y :

Given, variance of $x = \sigma_x^2 = 9 \Rightarrow \sigma_x = 3$

$$b_{xy} = 0.682, r = +0.74$$

$$\text{We know, } b_{xy} = r \cdot \frac{\sigma_x}{\sigma_y}$$

Substituting the value, we get

$$0.682 = 0.74 \times \frac{3}{\sigma_y}$$

$$\Rightarrow \sigma_y = \frac{2.22}{0.682} = 3.25$$

Hence, $\bar{X} = 10.7$, $\bar{Y} = 172.2$, $r = +0.74$, $\sigma_y = 3.25$

4.6.11 Standard Error in Regression Analysis

In regression analysis, dependent variables are calculated by regression equation while value of independent variable is given. It is almost impossible to calculate 100% accurate estimate.

To check the accuracy and reliability, one calculates the standard error of estimate. Standard error is calculated to find out that to what extent the estimated values are closer to actual values.

There are two standard error of estimate – one is for regression of X on Y and second is regression of Y on X :

1. **Standard Error of Estimate of Y on X (σ_{yx}):** Following formula is used to calculate the standard error of estimate of Y on X (denoted by σ_{yx}).

(i) First Formula:

$$\sigma_{yx} = \sqrt{\frac{\sum(Y - Y_c)^2}{N}}$$

where, Y = actual values and Y_c = estimated values.

(ii) Second Formula:

$$\sigma_{yx} = \sqrt{\frac{\sum Y^2 - a \sum Y - b \sum XY}{N}}$$

where, a and b are to be obtained from normal equations, a = intercept and b = slope of line.

(iii) Third Formula:

$$\sigma_{yx} = \sigma_y \sqrt{1 - r^2}$$

where, σ_y = SD of Y

r = Coefficient of correlation between X and Y

The above formula is suitable when the values of correlation coefficient (r) and standard deviations (σ_x and σ_y) are given.

Example 16: Find the ‘Standard error of the estimates’:

$\sigma_x = 4.4$, $\sigma_y = 2.2$ and $r = 0.8$.

Solution: Given, $r = 0.8$, $\sigma_x = 4.4$ and $\sigma_y = 2.2$

We can find out the ‘standard error of estimates’ as follows:

$$\sigma_{yx} = \sigma_y \sqrt{1 - r^2} \quad \dots (1)$$

$$\sigma_{xy} = \sigma_x \sqrt{1 - r^2} \quad \dots (2)$$

Putting the given values in equations (1) and (2), we get

$$\sigma_{yx} = 2.2 \times \sqrt{1 - (0.8)^2} = 2.2 \times \sqrt{1 - 0.64} = 2.2 \times \sqrt{0.36} = 2.2 \times 0.6 = 1.32$$

$$\sigma_{xy} = 4.4 \times \sqrt{1 - (0.8)^2} = 4.4 \times \sqrt{1 - 0.64} = 4.4 \times \sqrt{0.36} = 4.4 \times 0.6 = 2.64$$

Example 17: The expert of hospital administration told that number of FTFs (full-time employees) of hospital can be determined by calculating the total number of beds in the hospital. The business researcher of healthcare decided to make a regression model to calculate the number of FTFs of a hospital by the total number of beds. She took the data of 12 hospitals as shown in table below. Data are obtained in sequence, according to total number of beds.

No. of Beds	23	29	29	35	42	46	50	54	64	66	76	78
FTFs	69	95	102	118	126	125	138	178	156	184	176	225

Calculate the sum of squares of error and the standard error of the estimate.

Solution: Let us first determine the value of Σx , Σy , Σx^2 and Σxy using the following table:

Hospital	Number of Beds x	FTFs y	x^2	xy
1	23	69	529	1587
2	29	95	841	2755
3	29	102	841	2958
4	35	118	1225	4130
5	42	126	1764	5292
6	46	125	2116	5750
7	50	138	2500	6900
8	54	178	2916	9612
9	64	156	4096	9984
10	66	184	4356	12144
11	76	176	5776	13376
12	78	225	6084	17550
	$\Sigma x = 592$	$\Sigma y = 1692$	$\Sigma x^2 = 33044$	$\Sigma xy = 92038$

By above values, the healthcare researcher solved for the sample slope (b), the sample y and intercept (a):

$$\sigma_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} = 92038 - \frac{(592)(1692)}{12} = 8566$$

$$\sigma_{yx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 33044 - \frac{(592)^2}{12} = 3838.667$$

$$b = \frac{\sigma_{xy}}{\sigma_{yx}} = \frac{8566}{3838.667} = 2.232$$

$$a = \frac{\sum y}{n} - b \frac{\sum x}{n} = \frac{1692}{12} - (2.232) \frac{592}{12} = 30.888$$

A least square equation of the regression line is as follows:

$$\hat{y} = 30.888 + 2.232x$$

Slope is $b = 2.232$. This shows that for all unit increase of x (every bed), y (number of FTFs) is projected to increase by value 2.232.

Hospital	Number of Beds x	FTFs y	Residuals $y - \hat{y}$	$(y - \hat{y})^2$
1	23	69	-13.22	174.77
2	29	95	-0.62	-0.38
3	29	102	6.38	40.70
4	35	118	8.99	80.82
5	42	126	1.37	1.88
6	46	125	-8.56	73.27
7	50	138	-4.49	20.16
8	54	178	26.58	706.50
9	64	156	-17.74	314.71
10	66	184	5.80	33.64
11	76	176	-24.52	601.23
12	78	225	20.2	400.80
	$\Sigma x = 592$	$\Sigma y = 1692$	$\Sigma(y - \hat{y}) = -.01$	$\Sigma(y - \hat{y})^2 = 2448.86$

Error Sum of Square (SSE) = 2448.86

$$\text{Standard Error of Estimate, } s_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{2448.86}{10}} = 15.65$$

The 15.65 FTFs is estimated standard error. The analysis of the residuals for this problem tells that eight of 12 (67%) are within $\pm 1s_e$ and 100% are within $\pm 2s_e$.

4.6.12 Conducting Simple Linear Regression Analysis in PSPP

The last basic relationship analysis is simple linear regression. This is used to make predictions from an established linear relationship. The Linear Regression command is in the Regression folder of the **Analyze** menu.

PSPPIRE Data Editor

The Regression dialog box has fields for Independent (the X-axis) and Dependent (the Y-axis) variables. The variable used for making the predictions should be used for the X variable.

We will be using SAT to make predictions about GPA. SAT is used for the X variable. College GPA will be dependent or outcome variable that we are interested in making predictions about.

The “Statistics” button has some additional options for choosing the analyses.

The output file looks like this. Of particular importance is the “Unstandardized coefficients, B” column. This shows the weights for the regression formula.

REGRESSION

/VARIABLES=SAT

/DEPENDENT=GPA

/METHOD=ENTER

/STATISTICS=COEFF ANOVA

Model Summary (GPA from First Year)

<i>R</i>	<i>R Square</i>	<i>Adjusted R Square</i>	<i>Standard Error of the Estimate</i>
.78	.61	.59	.37

ANOVA (GPA from First Year)

	<i>Sum of Squares</i>	<i>df</i>	<i>Mean Square</i>	<i>F</i>	<i>Sig.</i>
Regression	2.79	1	2.79	20.74	.001
Residual	1.75	3	.13		
Total	4.54	14			

Coefficients (GPA from First Year)

	<i>Unstandardized Coefficients</i>		<i>Standardized Coefficients</i>	<i>t</i>	<i>Sig.</i>
	<i>B</i>	<i>Standard Error</i>	<i>Beta</i>		
(Constant)	-.93	.89	.00	-1.05	.312
SAT Verbal and Quantitative	.00	.00	.78	4.55	.001

4.6.13 Conducting Simple Linear Regression Analysis in R Commander

Linear regression is a method for modeling the relationship between two variables – one independent (x) and one dependent (y). The history and mathematical models behind regression analysis can be found out.

Scientists are typically interested in getting the equation of the line that describes the best least-squares fit between two datasets. They may also be interested in the coefficient of determination (R^2) which describes the proportion of variability in a data that is accounted for by the linear model.

Let us look at following table:

Table 4.5: Mileage and Groove Depth of a Car Tire

<i>Mileage (in 1000 Miles)</i>	<i>Groove Depth (in Miles)</i>
0	394.33
4	329.50
8	291.00
12	255.17
16	229.33
20	204.83

24	179.00
28	163.83
32	150.33

```
X → Mileage ← c(0, 4, 8, 12, 16, 20, 24, 28, 32)
```

```
Y ← Groove_Depth ← c(394.33, 329.50, 291.00, 255.17, 229.33, 204.83, 179.00, 163.83, 150.33)
```

```
p1 ← plot(x, y)
```

```
title[main="Scatter Plot of Groove Depth vs. Mileage",
```

```
xlab="Mileage(in 1000 miles)", ylab = "Groove Depth (in miles)"]
```

```
#The above commands draw the scatterplot of our data in x, y, including the (from the table value) caption, and labels for the x-axis and the y-axis#
```

```
Mod ← 1m(y ~ x)
```

```
#Fit the least squares linear regression model between y and x, where, y is the response variable, and x the regression#
```

Description of Function 1m:

1m is used to fit linear models. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance (although aov may provide a more convenient interface for these, you can see instructions of 1m() and aov() in your R.

Usage:

```
1m(formula, data, subset, weights, na.action, method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE, Singular.ok = TRUE, Contrasts = NULL, offset, ....)
```

```
abline(mod)
```

```
#You can add the regression line to the scatterplot by abline()#
```

ANOVA Tables

Description:

Compute analysis of variance (or deviance) tables for one or more fitted model objects.

Usage:

```
anova(object, ...)
```

Arguments:

Object an object containing the results returned by a model fitting function (e.g., 1m or g1m).

..... additional objects of the same type

```
ANOVA(mod)
```

Analysis of Variance Table

Response: y

	<i>df</i>	<i>Sum Sq</i>	<i>Mean Sq</i>	<i>F value</i>	<i>Pr(>F)</i>
<i>x</i>	1	50887	50887	140.71	6.871e-06***
<i>Residuals</i>	7	2532	362	—	—
<i>Signif. Codes</i>	0 '***'	0.001 '**'	0.01 '*'	0.05	'.' 0.1 ' ' 1

#df means the degrees of freedom, $Pr(>F)$ is the p-value we need. In this case, $Pr(>F) < 0.05$ (alpha). So, we reject H_0 and accept H_a that x has strong linear relationship with y .#

Now, let us check the four assumptions for linear regression model.

Checking for Model Assumptions:

1. Checking for Linearity
2. Checking for Constant Variance
3. Checking for Normality
4. Checking for Independence

```
Par (mfrow = c(2,2))
```

```
#this code let R show the plots in 2*2 format#
```

```
Plot (mod)
```

```
#we obtain the following plots, based on which you can judge if all related assumptions hold.
```

```
Summary(mod)
```

```
#Get details of the regression#
```

```
Call:
```

```
1m (formula = y ~ x )
```

Residuals:

<i>Min</i>	<i>IQ</i>	<i>Median</i>	<i>3Q</i>	<i>Max</i>
-18.099	-11.392	-6.902	7.051	33.693

Coefficients:

	<i>Estimate</i>	<i>Std. Error</i>	<i>t value</i>	<i>Pr(> t)</i>
<i>(Intercept)</i>	360.6367	11.6886	30.85	9.70e-09***
<i>x</i>	-7.2806	0.6138	-11.86	6.87e-06***
<i>Signif. Codes</i>	0 '***'	0.001 '**'	0.01 '*'	0.05 ' . ' 0.1 ' ' 1

Residual standard error: 19.02 on 7 degrees of freedom

Multiple R-squared: 0.9526, Adjusted R-squared: 0.9458

F-statistic: 140.7 on 1 and 7 DF, p -value: 6.871e-06

#You can get the estimates of coefficients of intercept and x . Both p -value of them is very small, which means we reject H_0 and accept H_a that they have strong linear relationship with y . You also can find that p -value here is as same as the p -value in ANOVA table before. Finally, coefficient of determination (r^2) is shown as R -squared. In this case, R -squared is close to 1, which means our linear regression works well#

4.7 CAUSE-AND-EFFECT RELATIONSHIP

4.7.1 Introduction

Cause and effect is a popular way to organize information in different texts, including academic papers in human resources management. Cause and effect refers to relationship between two phenomena in which one phenomenon is the reason behind the other. For example, eating too much fast food without any physical activity leads to weight gain. Here, eating without any physical activity is the “cause” and weight gain is the “effect.” Cause-and-effect Relationship is a relationship in which one event (the cause) makes another event happen (the effect).

The purpose of cause-and-effect relationship in HR management is to demonstrate your understanding of how employee recruitment, training, compensation, and other activities can help businesses in improving organizational performance and achieving their strategic goals. Cause-and-effect relationship explains the reasons of why some processes, events or phenomena happen or the effects of something. A business exists to solve specific problems. Business analysis is typically done to identify difficulties, the underlying causes and the subsequent effects that are experienced if the problems are not dealt with.

A business analyst will always look to establish a cause-effect relationship to propose viable solutions. The cause-and-effect relationship in management involves analysis of past problems and mistakes. The idea is to identify areas of concern in the past and avoid them in the present.

4.7.2 Steps of Cause-and-Effect Relationship

Following are some of the steps to analyze a cause-and-effect relationship:

1. **Identify the Problem:** We need to identify the problem of interest and describe adequately the magnitude and nature of the problem. It is important to determine which part of the business is affected and their possible reasons.
2. **Identify the Root Cause:** The root cause of the problem can be effectively identified by analysing an obstacle to procedures equipment, materials and people in the business. We have to come up with as many potential causes as we can.
3. **Identify the Sub-causes:** There are minor or additional ordeals and events that contribute to the occurrence of the major cause. We have to identify these ‘micro problems and find all the possible sub-causes contributing to the barriers. Customer dissatisfaction can result from poor product quality or average after-sales service. Similarly, employee performance can be affected by inadequate salary or poor

workplace conditions. To correctly and fully resolve an issue, the aim should be to identify the exact factor that is behind it.

4. **Analyze and Validate the Issue:** We must review each aspect to check the validity of the needs and the concerns that were stated and determine if they are admissible. Ascertain which factors need real-time action and which are critical for the business, sort problems according to urgency and priority, and delve deeper into the causes.
5. **Develop an Action Plan:** Once the problem, its causes and sub-causes are identified and the priority and legitimacy are ascertained, we have to come up with an action plan. An action plan embodies the remedies that are crucial for resolving the problem a business is facing. It helps to come up with authoritative business solutions.

The process can be effective only if professionals executing it are able and astute. It is important that managers fully understand what a cause-and-effect relationship is to effectively use it as a problem-solving tool or invite assistance from able practitioners accustomed to applying it.

4.7.3 Examples of Cause-and-Effect Relationship

Signal words such as why, how, because, what and analyze help in determining a cause-and-effect relationship. Here are some examples of cause-and-effect relationships:

1. **Reef Bleaching:** Global warming is threatening the Great Barrier Reef, affecting flora and fauna. Subsequently, it is becoming less colourful and increasingly prone to severe diseases.
Effect: Reef bleaching.
Root Cause: Global warming.
Sub-causes: Burning of fossil fuel, plastic, industry chemicals reverse boring, excessive livestock breeding.
Action Plan: Improve practices to reduce global temperature by 1.5 degrees Celsius.
2. **Poor Employee Performance:** An organization has been suffering significant losses due to incomplete tasks and delayed deliveries. Under-performing employees to reduce global temperature by 1.5 degrees Celsius.
Effect: Customer dissatisfaction and losses.
Root Cause: Poor employee performance.
Sub-causes: Faulty workplace machinery, poor Internet connectivity, low salaries, poor channels of communication between employees and managers.
Action Plan: Replace faulty machines with newer ones, install high-speed Internet and Wi-Fi, review salaries and incentives for each employee level, and establish seamless communication channels with managers by using collaborative platforms.

Outcomes in an organization are mostly the consequence of multiple factors combining in unpredictable ways and acting in concert. Establishing a cause-and-effect relationship is essential for identifying problems and coming up with resolutions. Managers must use it to come up with

the most viable solutions, prepare business visualizations of existing problems and improve the process of problem-solving within an organization.

Harappa's Executing Solutions course will help professionals stand out as growth-driven and ambitious. Solve problems and make crucial decisions that impact both the team and organization. Get things done quickly, plan projects and create a roadmap for success.

4.7.4 Conditions of Cause-and-Effect Relationship

When the objective of a research project is to confirm or test a hypothesized causal relationship, the research design requires special attention (Shadish, Cook and Campbell, 2002). The selection of the research design affects the level of rigor in making claims of causality based on study results. Generally speaking, in order to establish a causal relationship between *A* and *B*, we need to meet the following three conditions:

1. **Temporal Precedence:** Changes in *A* precede the observed changes in *B*.
2. **Covariation of the Cause and Effect:** Changes in *B* are related in a systematic way to changes in *A*.
3. **No Plausible Alternative Explanation:** No other factors are responsible for the observed changes in *B*.

4.7.4.1 Temporal Precedence

When you attempt to establish that *A* causes *B*, one of the minimum conditions you need to meet is that the changes in *A* happened before the changes in *B*.

A change that occurs prior to an event cannot be claimed to be caused by it. In Emily's case, the change in employees' experience due to the diversity training (*A*) needs to precede any observed changes in the level of cultural competence (*B*₁) and level of workplace conflict (*B*₂) observed in comparison to employees who did not receive the training.

4.7.4.2 Covariation of the Cause and Effect

Another condition you need to meet the claim that *A* cause *B* is that changes in *A* are systematically related to changes in *B*. If the changes in *B* happen at random, regardless of the presence of changes in *A*, then you cannot make a claim that *A* caused *B*. In other words, *A* and *B* need to have a relationship.

If you observe that whenever *A* is present, then *B* is present, or whenever *A* is absent, so is *B*, there is a transitive relationship between *A* and *B*. This relationship is typically described as a syllogism:

If *A*, then *B*.

If not *A*, then not *B*.

In Emily's case, any changes in the level of cultural competence (*B*₁) and the level of workplace conflict (*B*₂) need to be systematically related to the change introduced by the diversity

training (A). Putting the example of cultural competence in the syllogism illustrates why Emily needs a control group to establish a systematic relationship:

1. If diversity training (A) is offered, then there is outcome in cultural competence (B_1).
2. If diversity training (A) is not offered, then there is no outcome in cultural competence (B_1).

A cause-and-effect relationship is not always binary (yes/no; present/absent). In some cases, you may be looking for a situation where a different amount of A leads to a different amount of B . This relationship is described in a slightly different syllogism:

1. If more A , then more (or less) B .
2. If less A , then less (or more) B .

In Emily's case, she might later be interested in examining if any observed changes in cultural competence from the first training increased still more for employees who took additional diversity training. In this case, she would be looking for a relationship between the quantity of training and the quantity of improved cultural competence.

4.7.4.3 No Plausible Alternative Explanations

Emily's Case: Emily, HR Director at the city of Westlawn, brought her research team together to share what she discussed with Ahmed, the Community Foundation Program Officer. The team included training manager, Mei Lin, and a graduate student intern, Leo. Emily explained that she now had two research questions that would help them focus their evaluation of the training – “Does the training improve people's cultural competences?” and “Does the training decrease workplace tension?” She shared her idea to measure cultural competence and workplace tension before and after the training to assess the impact. She mentioned the idea of splitting each department into two groups, so half would participate in the training before the others. Then the team could measure the level of cultural competence and workplace tension, and compare the two groups at the point. She tasked Leo to find as much literature as possible that discusses training evaluation, measuring cultural competence, and workplace conflict. She asked Mei Lin to identify multiple scenarios for rolling out the training. They decided to have a weekly meeting to discuss how to implement the project.

Emily told Mei Lin and Leo, “It looks like this is going to be a lot of work, but I really want to do this right. I don't want to be doing the training for the sake of training without knowing what kind of impact it has on our employees. I believe focusing on the research design before we launch the training is important. I appreciate both of your help on this.”

Once you establish a relationship between A and B by temporal precedence and covariation, you then need to make sure the observed cause is not really due to some other factor C that is also systematically related to A and B . To be certain that A is the cause of B , you will need to eliminate all plausible alternative explanations for the changes in B . In Emily's case, she will need to show that no other factors other than diversity training (A) are responsible for any observed changes in the level of cultural competence (B_1) and workplace conflict (B_2).

When you claim a cause-and-effect relationship exists between *A* on *B*, the extent to which your claim is valid is referred to as internal validity. The presence of plausible alternative explanations are threats to internal validity. Following are eight common threats to the internal validity:

1. **History Threat:** An external event can be a threat to the causal argument. Recall Emily's case, when she was thinking of administering a survey and measuring the level of cultural competence and workplace conflict before and after the diversity training. She expected that any observed increase in cultural competence would be due to the training. Ahmed, the Community Foundation Program Officer, then countered that external events, such as the president giving a speech on a race relations or a work picnic, might occur at the same time and influence the results. Ahmed was raising the possibility of a historical effect that could pose a plausible threat to validity.
2. **Maturation Threat:** People change over time. They learn and mature from their daily experiences. They also grow physically and get older. This natural maturation can impact the outcomes you are observing in your research and could be a plausible alternative to your causal argument. This possibility makes it important to consider, for example, age or work experience differences among individuals in groups you are measuring.
3. **Instrumentation Threat:** Researchers use a variety of instruments to measure the phenomenon of interest. Instrumentation threat refers to a case when the instrument itself could be influencing the result. In Emily's case, consider the survey she intends to use to measure cultural competence. If she uses a different survey before and after the diversity training, with different wording or order for the same questions, her results could be affected by differences in the instrument.

A common example of the instrumentation threat to validity appears in face-to-face interviews on sensitive topics, where a respondent may shape answers to avoid a negative appearance. The instrument can also refer to the people who collect the data. Over time, an observer might get bored and pay less attention, or might learn from experience and change the way observations are interpreted.

4. **Testing Threat:** Similar to the instrumentation threat, the testing threat operates when measurement takes place more than once. Here, instead of the instrument itself, the issue relates to a learning effect by the subjects being measured.

In Emily's case, if she uses the same survey before and after the diversity training, there is a possibility that some employees will have thought about their earlier responses and decided to change answers to a "right" answer that they think Emily wants to hear. Similarly, if students are given the same math exam a second time, they might show improvement that reflects experience with the particular questions on the exam, not improved skill in the math involved.

5. **Mortality or Attrition Threat:** During a research study, participants will often drop out. The term mortality is used metaphorically (usually) to refer to the attrition of the study participants. Participants may drop out for particular reasons, perhaps because they performed poorly in a baseline assessment or for other reasons that distinguish them from participants who stay in the study. If attrition is random, there may be no consequence, but a systematic change in the people in your study is likely to affect your results. It will at least make it plausible that there is a threat to validity, and you will need to address the issue to avoid criticism of the results.

In Emily's case, suppose a number of employees refuse to take the cultural competence survey following the training. If the results of the survey show improvement, she will need to consider the possibility that those who dropped out do not endorse the idea of embracing diversity in the organization and were responsible for lower average scores on the initial survey. Dropping out can be a form of protest among individuals who are systematically different from those who continue to participate.

6. **Regression Threat:** The regression threat is also known as regression artifact or regression to the mean. It refers to a statistical phenomenon that occurs when the mean (scores of the data) from a non-random sample is measured twice, it will move closer to the population mean. Scores from a non-random sample may include extreme scores in the first measurement. However, when the same sample is measured twice, it is less likely that the extreme scores will persist. In other words, even if you do nothing to the sample, an extreme score is likely to move closer to the mean when measured a second time. This threat to validity is dependent on the level of variation possible in the value being measured.
7. **Selection Threat:** Comparing two groups is a common procedure researchers take to establish causality for an intervention offered to one of the groups. If an outcome changes for the group with the intervention (called the experimental group), but the outcome does not change for the group with no intervention (called the control group), then you have a basis to argue that the intervention caused the change in outcome. This approach is called an experimental design.

When two or more groups are compared, attention needs to be given to the possibility that the composition of the groups are systematically different from each other and are not comparable. If the groups are different to begin with (selection bias), then any difference in results observed following an intervention could be due to the original difference in the groups and not from the intervention.

Notice in Emily's case, when she decided with Ahmed to offer the training first to half of the employees in each department to compare to the other half who would not take the training, the issue arose how the employees would be selected. If Emily allows people to sign up for the training (self-selection), then a selection bias could occur.

Very possibly, those employees most interested in the diversity training would sign up first. In that case, improved survey results on cultural competence following the training could be due to their interest and predisposition to be influenced by the training. Emily and Ahmed agreed that the employees would need to be randomly assigned to take the training to avoid this kind of selection bias in the composition of the groups.

Note that even when groups are selected by random assignment, researchers usually examine the resulting composition of the groups by age and other factors to assess any differences that might have occurred in the selection process.

8. Selection Interaction Threats: The selection threat to validity can also interact with other threats to internal validity. Variations are described below:

- (i) A selection-history threat could occur if individuals in two groups experience an external event differently; *e.g.*, due to differences in a pre-existing attitude or different reporting of the event.
- (ii) A selection-maturation threat could occur if two groups mature differently; *e.g.*, due to gender or socio-economic differences.
- (iii) A selection-instrumentation threat could occur, *e.g.*, when responses from two groups are measured with two different survey instruments.
- (iv) A selection-testing threat could occur if differences between two groups influence the way they respond or learn from exposure to repeated testing, *e.g.*, due to perceived burden and inattention or learning to avoid stigma by finding the “right” answer.
- (v) A selection-regression threat could occur, in one example, whenever an undetected selection bias occurs in the composition of two groups. At first measurement, the groups could appear similar and only later appear different, due to regression toward different original conditions.

Random assignment should control for this possibility. The threat is more common in situations where researchers select extreme cases for an intervention and then find improvement occurs. If variation is possible in the value that was used to select the sample, then part of the improvement could be attributable to regression to the mean.

4.8 MULTIPLE REGRESSION ANALYSIS

4.8.1 Introduction

Three or more variables, such as X_1 , X_2 , and X_3 , are analyzed in multiple regression analysis. Here, X_1 is taken as a dependent variable and its relative movement is found out, according to the movement in both independent variables X_2 and X_3 . Thus, the effect of two or more independent variables on one dependent variable is studied in the ‘multiple regression analysis’. The procedure

for studying multiple regressions is similar to simple regression, but there is a single difference that the independent variable is more than one. If there are three variables, X_1 , X_2 and X_3 , the multiple regressions will take the following form:

$$X_1 = a_{1.23} + b_{12.3}X_2 + b_{13.2}X_3 \quad \dots (1)$$

In the above equation, there are three constraints. The subscript after the dot indicates the variables, which are held constant.

If X_2 and X_3 are 0, then $X_1 = a_{1.23}$

For multiple regression models with two independent variables, the result is determined with the use of three simultaneous equations with three unknown (b_0 , b_1 and b_2) values. Statistical software, such as SPSS, MINITAB, MS-EXCEL, etc. are used in multiple regression models because this analysis involves lengthy calculations.

4.8.2 Assumptions of Multiple Regression Analysis

Following are the assumptions of multiple regression analysis:

1. The dependent variable should be a random variable whereas the independent variable need not be random variable.
2. The relationship between various independent variables and dependent variables is linear.
3. The variances of the conditional distributions of the dependent variable and various combinations of values of the independent variables are all equal. The conditional distributions for the dependent variable follow the normal probability distribution.

4.8.3 Steps in Multiple Regression

Step 1: State the research hypothesis.

Step 2: State the null hypothesis.

Step 3: Collect the data.

Step 4: Calculate each variable separately (calculate measures of central tendency and dispersion, frequency distributions and graphs). Check whether the variables are normally distributed or not.

Step 5: Calculate the relationship between the dependent variable and the independent variables (find the correlation coefficient and obtain a scatterplot). Check whether the variables are linearly correlated or not.

Step 6: Calculate the relationship between all the independent variables with each other (calculate the correlation coefficient matrix for all the independent variables).

Step 7: Calculate the regression equation from the data.

Step 8: Calculate and examine appropriate measures of association and test of statistical significance for each coefficient as well as for the equation as a whole.

Step 9: Accept or reject the null hypothesis.

Step 10: Reject or accept the research hypothesis.

Step 11: Explain the practical implications of the findings.

4.8.4 Difference between Simple and Multiple Regressions

The difference between simple and multiple regressions are shown in below Table 4.6.

Table 4.6: Comparison between Simple and Multiple Regressions

<i>Simple Regressions</i>	<i>Multiple Regressions</i>
An dependent variable Y predicted from an independent variable, X .	An dependent variable, Y predicted from the independent variables set $(X_1, X_2, \dots, X_k)$.
Simple Regression has only one regression coefficient.	For each independent variable, there is a regression coefficient.
r^2 is the proportion of variation in dependent variable Y predicted from X .	R^2 is the proportion of variation in dependent variable Y predicted by the independent variables set of X 's.

4.8.5 Methods of Calculation

1. Multiple Regression Equations

In case of simple regression, there is a regression line approximating a set of N points (X and Y) in a two-dimensional scatter diagram [Fig. 4.9(b)]. Similarly, in case of multiple regression, there exists a least square regression plane fitting a set of N points (X_1, X_2 and X_3) in a three-dimensional scatter diagram [Fig. 4.9(a)]:

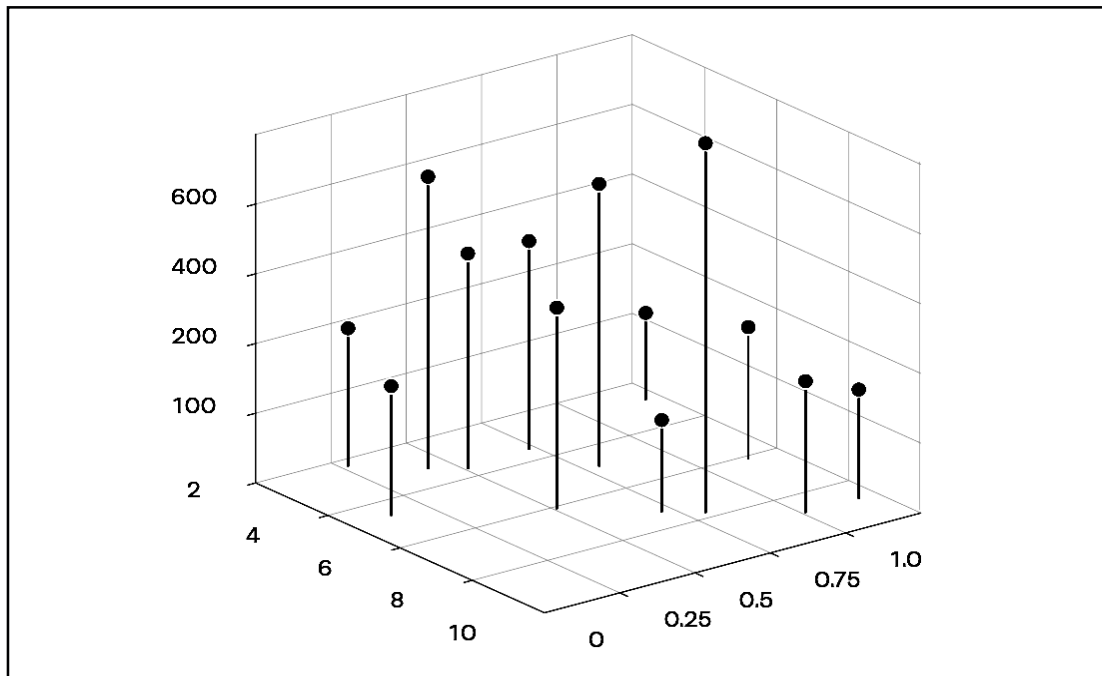


Fig. 4.9(a): Two-dimensional Scattered Diagram

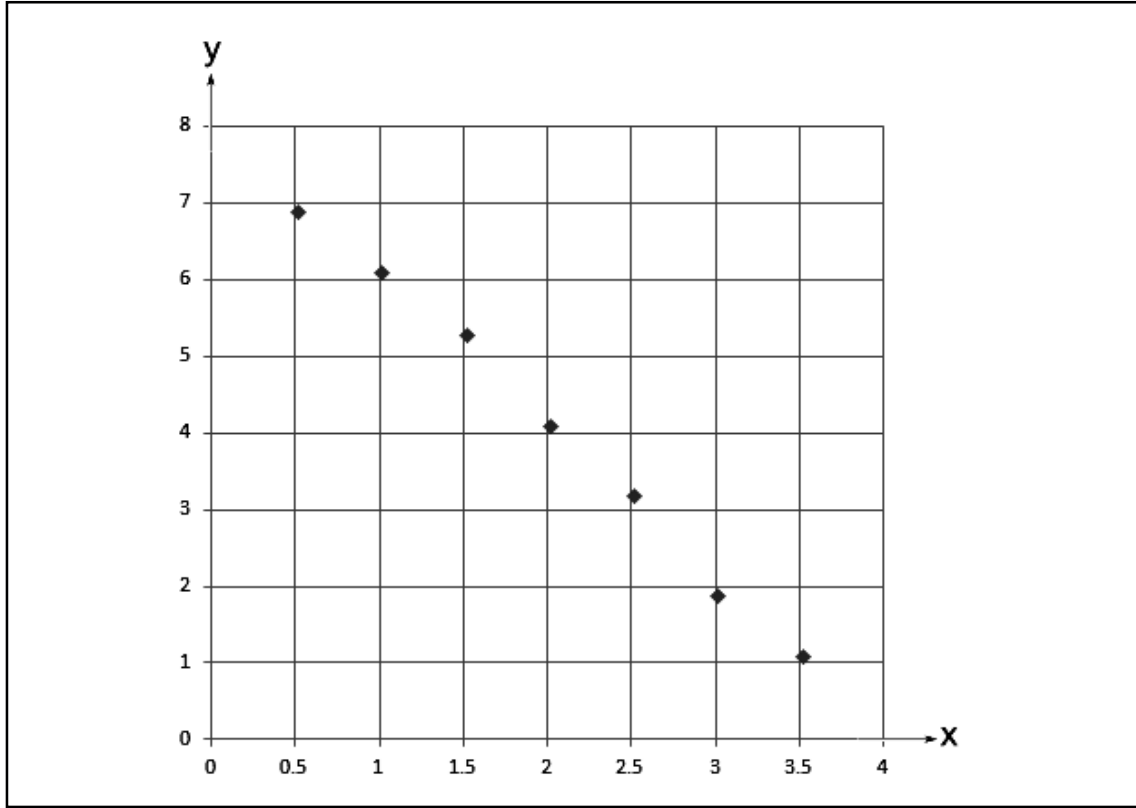


Fig. 4.9(b): Three-dimensional Scatter Diagram

- (i) The regression plane of X_1 on X_2 and X_3 is represented by the equation (1):

$$X_1 = a_{1.23} + b_{12.3}X_2 + b_{13.2}X_3 \quad \dots (1)$$

The values of $b_{12.3}$ and $b_{13.2}$ are determined by solving simultaneously the following three normal equations:

$$\sum X_1 = Na_{1.23} + b_{12.3}\sum X_2 + b_{13.2}\sum X_3$$

$$\sum X_1X_2 = a_{1.23}\sum X_2 + b_{12.3}\sum X_2^2 + b_{13.2}\sum X_2X_3$$

$$\sum X_1X_3 = a_{1.23}\sum X_3 + b_{12.3}\sum X_2X_3 + b_{13.2}\sum X_3^2$$

On multiplying X_2 and X_3 on both sides of equation (1), the last two normal equations are obtained.

The regression equation of X_2 on X_1 and X_3 would be as follows:

$$X_2 = a_{2.13} + b_{21.3}X_1 + b_{23.1}X_3 \quad \dots (2)$$

The normal equations for fitting the above equation are:

$$\sum X_2 = Na_{2.13} + b_{21.3}\sum X_1 + b_{23.1}\sum X_3$$

$$\sum X_1X_2 = a_{2.13}\sum X_1 + b_{21.3}\sum X_1^2 + b_{23.1}\sum X_1X_3$$

$$\sum X_2X_3 = a_{2.13}\sum X_3 + b_{21.3}\sum X_1X_3 + b_{23.1}\sum X_3^2$$

(iii) The regression equation of X_3 on X_1 and X_2 would be as follows:

$$X_3 = a_{3.12} b_{31.2} X_1 b_{32.1} X_2 \quad \dots (3)$$

The normal equations for fitting the above equation are:

$$\Sigma X_3 = Na_{3.12} + b_{31.2} \Sigma X_1 + b_{32.1} \Sigma X_2$$

$$\Sigma X_1 X_3 = a_{3.12} \Sigma X_1 + b_{31.2} \Sigma X_1^2 + b_{32.1} \Sigma X_1 X_2$$

$$\Sigma X_2 X_3 = a_{3.12} \Sigma X_2 + b_{31.2} \Sigma X_1 X_2 + b_{32.1} \Sigma X_2^2$$

Generalization for more than three variables is as follows:

$$X_1 = a_{1.234} + b_{12.34} X_2 + b_{13.24} X_3 + b_{14.23} X_4$$

The four normal equations are determined by multiplying the above equation on both sides by X_1 , X_2 , X_3 and X_4 to find the values of $a_{1.234}$, $b_{12.24}$, $b_{12.34}$ and $b_{14.23}$.

2. Deviations from the Mean

If the deviations from the respective means of the various variables are taken into account, then the calculations for solving these equations can be reduced. Thus, if $X_1 = (X_1 - \bar{X}_1)$, $X_2 = (X_2 - \bar{X}_2)$ and $X_3 = (X_3 - \bar{X}_3)$, then the multiple regression equation can be expressed as follows:

$$X_1 = b_{12.3} X_2 + b_{13.2} X_3$$

Following two normal equations are required to be solved, for determining the values of $b_{12.3}$ and $b_{13.2}$:

$$\Sigma X_1 X_2 = b_{12.3} \Sigma X_2^2 + b_{13.2} \Sigma X_2 X_3$$

$$\Sigma X_1 X_3 = b_{12.3} \Sigma X_2 X_3 + b_{13.2} \Sigma X_3^2$$

The values of $b_{12.3}$ and $b_{13.2}$ can also be determined as follows:

$$b_{12.3} = r_{12.3} \times \frac{S_{1.23}}{S_{2.13}} \quad b_{13.2} = r_{13.2} \frac{S_{13.2}}{S_{3.12}}$$

Similar calculations are done for the regression equation of X_2 on X_1 and X_3 and for regression equation of X_3 on X_1 and X_2 .

3. Multiple Regression Equation of X_1 on X_2 and X_3

It can be expressed as follows:

$$X - \bar{X}_1 = \frac{\partial_1}{\partial_2} \left[\frac{r_{12} r_{13} r_{23}}{1 - r_{23}^2} \right] (X_2 - \bar{X}_2) + \frac{\partial_1}{\partial_3} \left[\frac{r_{12} r_{13} r_{23}}{-r_{23}^2} \right] (X_3 - \bar{X}_3)$$

Similarly, the multiple regression equation of X_2 on X_3 and X_1 can be expressed as follows:

$$X_2 - \bar{X}_2 = \frac{\partial_2}{\partial_3} \left[\frac{r_{23} r_{12} r_{13}}{1 - r_{13}^2} \right] (X_3 - \bar{X}_3) + \frac{\partial_2}{\partial_1} \left[\frac{r_{12} r_{13} r_{23}}{-r_{23}^2} \right] (X_1 - \bar{X}_1)$$

Lastly, the multiple regression equation of X_3 on X_1 and X_2 would be expressed as follows:

$$X_3 - \bar{X}_3 = \frac{\partial_3}{\partial_3} \left[\frac{r_{23}r_{12}r_{13}}{1-r_{13}^2} \right] (X_3 - \bar{X}_3) \left[\frac{r_{13}r_{23}r_{12}}{-r_{12}^2} \right] (X_1 - \bar{X}_1)$$

In this method, determination of regression equations is simple in comparison when the normal equations are required to be solved for determining the regression equations. Here, we only need the following values:

$X_1, X_2, X_3, \partial_1, \partial_2, \partial_3$ and r_{12}, r_{13}, r_{23}

Example 18: Find the multiple regression equation of X_1 on X_2 and X_3 from the data relating to three variables given below:

X_1	4	6	7	9	13	15
X_2	15	12	8	6	4	3
X_3	30	24	20	14	10	4

Solution: The regression equation of X_1 on X_2 and X_3 is:

$$X_1 = a_{1.23} + b_{12.3}X_2 + b_{13.2}X_3$$

The three normal equations to get the values of $a_{1.23}$, $b_{12.3}$ and $b_{13.2}$ are:

$$\Sigma X_1 = Na_{1.23} + b_{12.3}\Sigma X_2 + b_{13.2}\Sigma X_3$$

$$\Sigma X_1X_2 = a_{1.23}\Sigma X_2 + b_{12.3}\Sigma X_2^2 + b_{13.2}\Sigma X_2X_3$$

$$\Sigma X_1X_3 = a_{1.23}\Sigma X_3 + b_{12.3}\Sigma X_2X_3 + b_{13.2}\Sigma X_3^2$$

X_1	X_2	X_3	X_1X_2	X_1X_3	X_2X_3	X_2^2	X_3^2
4	15	30	60	120	450	225	900
6	12	24	72	144	288	144	576
7	8	20	56	140	160	64	400
9	6	14	54	126	84	36	196
13	4	10	52	130	40	16	100
15	3	4	45	60	12	9	16
$\Sigma X_1 = 54$	$\Sigma X_2 = 48$	$\Sigma X_3 = 102$	$\Sigma X_1X_2 = 339$	$\Sigma X_1X_3 = 720$	$\Sigma X_2X_3 = 1034$	$\Sigma X_2^2 = 494$	$\Sigma X_3^2 = 2188$

Substituting the values in the normal equations:

$$6a_{1.23} + 48b_{12.3} + 102b_{13.2} = 45 \quad \dots (1)$$

$$48a_{1.23} + 494b_{12.3} + 1034b_{13.2} = 339 \quad \dots (2)$$

$$102a_{1.23} + 1034b_{13.2} = 720 \quad \dots (3)$$

Multiplying equation (1) by 8, get,

$$48a_{1.23} + 384b_{12.3} + 816b_{13.2} = 432 \quad \dots (4)$$

Subtracting equation (2) from equation (4), we get

$$110b_{12.3} + 218b_{13.2} = -93 \quad \dots (5)$$

Multiplying equation (1) by 17, we get

$$02a_{1.23} + 816b_{12.3} + 1734b_{13.2} = 918 \quad \dots (6)$$

Subtracting equation (3) from equation (6), we get

$$218b_{12.3} + 454b_{13.2} = -198 \quad \dots (7)$$

Multiplying equation (5) by 109, we get

$$11990b_{12.3} + 23762b_{13.2} = -10137 \quad \dots (8)$$

Multiplying equation (7) by 55, we get

$$11990b_{12.3} + 24970b_{13.2} = -10890 \quad \dots (9)$$

Subtracting equation (8) from equation (9), we get

$$1208b_{13.2} = 753 \text{ or } b_{13.2} = \frac{753}{1208} = -0.623$$

Substituting this value of $b_{13.2}$ in equation (5), we get

$$110b_{12.3} + 218(-0.623) = -93$$

$$110b_{12.3} = 135.814 - 93$$

$$b_{12.3} = \frac{42.814}{110} = +0.389$$

Substituting the values of $b_{12.3}$ and $b_{13.2}$ in equation (1), we get

$$6a_{1.23} + 48(0.389) + 1021(-0.623) = 45$$

$$6a_{1.23} = 54 + 63.546 - 18.672$$

$$6a_{1.23} = 98.708774$$

$$a_{1.23} = 16.476$$

Thus, the required regression equation is: $X_1 = 16.479 + 0.389 + 0.389X_2 - 0.623X_3$.

4.9 INTERACTION EFFECTS

An interaction occurs when an independent variable has a different effect on the outcome depending on the values of another independent variable. Interaction effects occur when the effect of one variable depends on the value of another variable. Interaction effects are common in regression analysis, ANOVA, and designed experiments. In statistical designs the problems to be faced if do not include them in model.

Interaction Effects in Equations

In a regression equation, an interaction effect is represented as the product of two or more independent variables. For example, here is a typical regression equation without an interaction:

$$Y = b_0 + b_1X_1 + b_2X_2$$

where, Y is the predicted value of a dependent variable, X_1 , and X_2 are independent variables, and b_0 , b_1 , and b_2 are regression coefficients. And here is the same regression equation with an interaction:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_1X_2$$

Here, b_3 is a regression coefficient, and X_1X_2 is the interaction. The interaction between X_1 and X_2 is called a two-way interaction because it is the interaction between two independent variables. Higher-order interactions are possible, as illustrated by the three-way interaction in the following equation:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_4X_1X_2 + b_5X_1X_3 + b_7X_1X_2X_3$$

Analysts usually steer clear of higher-order interactions, like $X_1X_2X_3$, since they can be hard to interpret.

Suppose that there is a cholesterol lowering drug that is tested through a clinical trial. Suppose we are expecting a linear dose-response over a given range of drug dose:

This is a standard simple linear model. Now, however, suppose that we expect men to respond at an overall higher level compared to women. There are various ways that this can happen.

On the other hand, if men, have a less steeper dose-response curve compared to women, we would expect like this:

Of these four graphs, the first indicates no difference between men and women, the second illustrates that there is a difference, but since it is constant, there is no interaction term. The third and fourth graphs represent the situation with an interaction of the effect of the drug, depending on whether it is given to men or women.

Of these, the first indicates no difference between men and women, the second illustrates that there is a difference, but since it is constant, there is no interaction term. The third and fourth graphs represent the situation with an interaction of the effect of the drug, depending on whether it is given to men or women.

In terms of regression equations, we have:

No effect of sex:

$$Y = \alpha + \beta_1 * dose (+0 \times sex + 0 \times dose * sex)$$

where, Y represents the outcome (amount of cholesterol lowering), β_1 represents the effect of the drug (presumed here to be non-zero), and all other coefficients for the rest of the terms (effect of sex and interaction term) are zero.

Sex has an effect, but no interaction:

$$Y = \alpha + \beta_1 * dose + \beta_2 \times sex (+0 \times dose * sex)$$

Sex has an effect with an interaction:

$$Y = \alpha + \beta_1 * dose + \beta_2 \times sex + \beta_3 \times dose * sex$$

Let us consider how to report the effects of sex and dose in the presence of interaction terms. If we consider the first of the above models, without any effect of sex, it is trivial to report. There is no effect of sex, and the coefficient β_1 provides the effect of dose. In particular, β_1 represents the amount by which cholesterol changes for each unit change in dose of the drug.

If we consider the second model, where there are effects of both dose and sex, interpretation is still straightforward. Since it does not depend on which sex is being discussed (effect is the same in males and females), β_1 still represents the amount by which cholesterol changes for each unit change in dose of the drug. Similarly, β_2 represents the effect of sex, which is “additive” to the effect of dose, because to get the effect of both together for any dose, we simply add the two individual effects.

Now, consider the third model with an interaction term. Things get a bit more complicated when there is an interaction term. There is no longer any unique effect of dose because it depends upon whether you are talking about the effect of dose in males or females. Similarly, the difference between males and females depends on the dose.

Consider first the effect of dose. The question of “what is the effect of dose” is not answerable until one knows which sex is being considered. The effect of dose is β_1 for females (if they are coded as 0, and males coded as 1, as was the case here). This is because the interaction term becomes 0 if sex is coded as 0. So, the interaction term “disappears”.

On the other hand, if sex is coded as 1 (males), the effect of dose is now equal to $\beta_1 + \beta_3$. This means, in practice, that for every one unit increase in dose, cholesterol changes by the amount $\beta_1 + \beta_3$ in males (compared to just β_1 for females).

All of the above models have considered a continuous variable combined with a dichotomous (dummy or indicator) variable. We can also consider interactions between two dummy variables, and between two continuous variables. The principles remain the same, although some technical details change.

SUMMARY

- One of the most widely used models for data analysis is regression analysis. In this module, you will learn how to perform a regression analysis with the help of several real-world examples and a business case.
- Correlation and linear regression are the most often used approaches for examining the relationship between two quantitative variables. The strength of a linear link between two variables is measured by correlation, whereas regression expresses the relationship as an equation.
- Regression is most commonly used to construct models/equations that predicts a key response (Y) from a group of predictor (X) variables. Correlation is generally used to convey the direction and degree of links between two or more quantitative variables quickly and succinctly.
- PSPP is a program that analyses sampled data statistically. It is a free alternative to the proprietary SPSS application. T-tests, ANOVA, GLM, linear and logistic regression,

reliability analysis, clustering, factor analysis, non-parametric tests, and other statistical functions are all supported by PSPP.

- R has an integrated development environment (IDE) called R Studio. It is essentially a R front-end with a console, a scripting window, a graphics window, and a R workspace, among other features. R Commander is a simple graphical user interface (GUI) for the R programming language.
- Interaction effects reflect the influence of a third variable on the connection between two independent variables. Because the relationship between an independent and dependent variable changes based on the value of a third variable, statisticians say these variables interact in this case.

QUESTIONS

Very Short Answer Type Questions

1. What is linear correlation?
2. Mention various types of correlation.
3. What is coefficient of correlation?
4. What is meant by rank correlation?
5. What is the interpretation of Correlation coefficient (p) = 0?
6. What are multiple regressions?
7. What is temporal precedence?
8. What are the PSPP and R Commander?
9. What are interaction effects?

Short Answer Type Questions

1. What is the rank difference method of computing coefficient of correlation?
2. Define the term “Correlation”. Explain the concept of positive and negative correlation.
3. Discuss the merits and demerits of Rank Correlation Coefficient.
4. What is correlation? What are its uses?
5. Discuss the cause-and-effect relationship in detail.
6. Discuss about the PSPP environment in detail.
7. Write the steps of Cause-and-effect Relationship.
8. Discuss about the menu and toolbars in PSPP.
9. Discuss how to plot scatter diagram in PSPP.
10. Explain how to conduct correlation in PSPP.

Long Answer Type Questions

1. Explain the significance of correlation and scatter diagram method.
2. Explain the meaning and significance of the concept of correlation. Give the general rules for interpreting its coefficient.
3. What is meant by correlation? What is its significance in statistical analysis? Discuss briefly the different measures of correlation.
4. Explain the interpretation of correlation with the help of scatter diagrams.
5. What do you mean by Correlation? Distinguish between Positive and Negative Correlation. What are different methods of studying correlation?
6. What are the assumptions of multiple regression? Also write the steps of multiple regressions.
7. Explain how to conduct simple linear regression analysis using PSPP and R Commander.
8. From the following data, calculate: (i) Correlation and (ii) SD of Y .

$$X = 0.854Y, Y = 0.89X \text{ and SD of } X = 3$$

[Ans: $r = 0.87$, SD of $Y = 3.06$]

9. If $b_{xy} = -0.4$ and $b_{yx} = -0.9$, then find r . [Ans: -0.6]
10. The annual sales of a company are given below. Estimate the sales for the year 2017.

Year	Sales in Lakh (₹)
2014	23.4
2015	24.2
2016	25.8
2017	?

11. Find Karl Pearson's coefficient of correlation from the following index numbers and interpret it:

Wages	100	101	103	102	104	99	97	98	96	96
Cost of Living	98	99	99	97	95	02	95	94	90	91

[Ans: $r = 0.85$]

12. Find the correlation by Karl Pearson's method between the two kinds of assessment of postgraduate student's performance (marks out of 100):

Roll No. of Students	1	2	3	4	5	6	7	8	9	10
Internal Assessment	45	62	67	32	12	38	47	67	42	85
External Assessment	39	48	65	32	20	35	45	77	30	62

[Ans: $r = 0.88$]

13. Find the coefficient of correlation between price and sales from the following data:

Price (₹)	103	98	85	92	90	84	88	90	93	95
Sales (Units)	500	610	700	630	670	800	800	750	700	680

Also, calculate probable error and point out whether the coefficient of correlation is significant or not.

[Ans: $r = 0.85$, P.E. = 0.0877]

14. Calculate the coefficient of correlation between the values of X and Y given below:

X	78	98	96	69	59	79	68	61
Y	125	137	156	112	107	136	123	108

[Ans: $r = 0.955$]

15. Calculate the Karl Pearson's Coefficient of Correlation between age and playing habits from the data given below. Comment on the value:

Age	20	21	22	23	24	25
Number of Students	500	400	300	240	200	160
Regular Players	400	300	180	96	60	24

[Ans: $r = -0.991$]

16. Find the correlation coefficient between age and playing habits of the following students:

Age	20	21	22	23	24	25
Number of Students	500	400	300	240	200	160
Number of Players	200	150	90	48	30	12

[Ans: $r = -0.991$]

17. Find Rank correlation coefficient:

Marks in Economics	80	45	55	56	58	60	65	68	20	75	85
Marks in Mathematics	82	86	50	48	60	62	64	65	70	74	90

[Ans: $r = -0.58$]

18. Find Rank correlation coefficient:

Scores in Test X	21	25	35	24	16	18	25	16	16	38
Scores in Test Y	12	15	24	20	8	15	20	20	11	26

[Ans: $r = -0.755$]

19. Find the Karl Pearson's coefficient of correlation and explain the answer.

Statistics	10	25	13	25	22	11	12	25	21	20
Economics	12	22	16	15	18	18	17	23	24	17

[Ans: $r = 0.462$]

20. Calculate regression coefficients for the following data:

X	2	4	6	8	10	12	14
Y	4	2	5	10	4	11	12

[Ans: $b_{xy} = 0.85$, $b_{yx} = 0.73$]

21. Find Karl Pearson's coefficient of correlation between sales and expenses of the following ten firms:

Firm	Sales in Thousand Units	Expenses in Thousand ₹
1	50	11
2	50	13

3	55	14
4	60	16
5	65	16
6	65	15
7	65	15
8	60	14
9	60	13
10	50	13

[Ans: 0.786]

22. Calculate the Coefficient of Correlation by Karl Pearson's method from the following data:

X	12	9	8	10	11	13	7
Y	14	8	6	9	11	121	3

[Ans: 0.949]

23. Calculate Karl Pearson's coefficient of correlation between price and supply of a commodity from the following data:

Price (₹)	17	18	19	20	21	22	23	24	25	26
Supply (kg)	38	27	38	33	32	33	34	29	26	33

[Ans: $r = (-)0.42$]

24. Calculate Karl Pearson's coefficient of correlation between the age of husbands and wives from the following data:

Husband's Age	32	35	27	38	25	40
Wife's Age	28	32	26	35	24	38

[Ans: 0.982]

REFERENCES

- <https://www.aihr.com/blog/hr-analytics-techniques-regression-analysis/>
- <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC374386/>
- <https://www.graphpad.com/support/faq/what-is-the-difference-between-correlation-and-linear-regression/>
- <https://savannah.gnu.org/projects/pspp/#:~:text=PSPP%20is%20a%20program%20for,tests%20C%20and%20other%20statistical%20features.>
- <https://www.andrewheiss.com/blog/2012/04/17/install-r-rstudio-r-commander-windows-osx/>
- <https://statisticsbyjim.com/regression/interaction-effects/#:~:text=Interaction%20effects%20indicate%20that%20a,value%20of%20a%20third%20variable.>



MODULE 5

HR ANALYTICS APPLICATIONS USING T-TEST AND ANOVA

Learning Objectives:

1. Understand the basic concepts of Hypotheses, ANOVA and t-test.
2. Develop t-test in R Commander and PSPP.
3. Study how to develop and conduct Independent ANOVA in R Commander and PSPP.
4. Learn how to conceptualize the Simple Linear Regression Analysis in PSPP.
5. Demonstrate the advantages of Free OSS over Closed Software.

5.1 HYPOTHESIS

5.1.1 Meaning and Definition of Hypothesis

It is derived from a Greek word “*hyposthenia*” which means “to put under” or “to suppose”. Hypothesis is the process of taking and using a sample statistics to make an inference for a population parameter.

According to **Prof. Morris Hamburg**, “A hypothesis in statistics is simply a quantitative statement about a population”.

According to **Palmer O. Johnson**, “A hypothesis is islands in the uncharted seas of thought to be used as bases for consolidation and recuperation as we advance into the unknown”.

Hypothesis generally occurs when we perform inductive reasoning. In inductive reasoning, researcher takes a series of observations to perform a theory. So, hypothesis is the first step of any research process. It is very important for any research and formulated in various ways. Its basic function is the prediction of the final result of the investigation.

5.1.2 Qualities of a Good Hypothesis

1. **Clear and Precise:** Hypothesis should be clear and precise. If the hypothesis is not clear and precise, the inferences drawn on its basis cannot be taken as reliable.
2. **Ability to Test:** Hypothesis should be capable of being tested.

3. **State Relationship between Variables:** Hypothesis state the relationship between variables, if it happens to be a relational hypothesis.
4. **Limited in Scope:** Hypothesis should be lifted in scope and must be specific. A researcher must remember that narrower hypotheses are generally more testable and he should develop such hypotheses.
5. **Simple:** Hypothesis should be stated as far as possible in most simple terms to that the same is easily understandable by all concerned. But one must remember that simplicity of hypothesis has nothing to do with its significance.
6. **Consistent:** Hypothesis should be consistent with most known facts, *i.e.*, it must be consistent with a substantial body of established facts. In other words, it should be one which judges **accept** as being the **most likely**.
7. **Logical:** If there are two or more hypothesis derived from the same basic theory, then they should not contradict each other.
8. **Amenable:** Hypothesis should be amenable to testing within a reasonable time. One should not use even an excellent hypothesis, if the same cannot be tested in reasonable time as one cannot spend a lifetime collecting data to test it.
9. **Subjectivity:** Researcher's subjectivity or his biased judgement should be eliminated from the hypothesis.

5.2 HYPOTHESIS TESTING

5.2.1 Introduction

Hypothesis testing plays an important role in the application of statistics to real-life problems. It helps the researchers to make the decisions by using sampled data of unknown distribution of a population and its parameters.

For example, consider a manufacturing company which makes the geyser. The average life of the geyser is 1000 hour when manufactured through old process and the average life is 1500 hours when the geyser is manufactured by the new process. So, in this case, we can make three hypothesis:

1. New process is better than old process.
2. Old process is better than new process.
3. There is no difference between both the processes.

5.2.2 Purpose of Hypothesis Testing

1. Determine the focus and direction for a research effort.
2. Development of a hypothesis forces the researcher to clearly state the purpose of the research activity.

3. Determine what variables will not be considered in a study, as well as those that will be considered.

5.2.3 Types of Hypothesis

Below are some of the important types of hypothesis:

1. **Research Hypothesis:** It is directional hypotheses which specify the projected direction of the result. Using earlier research or theory, the direction is implied. It is a statement which is generated by the researchers when they make a guess on the result of an experiment or research. When one needs to test that a data supports the research hypothesis, then he/she requires to transform the data into statistical hypothesis.
2. **Statistical Hypothesis:** It is a hypothesis that can be verified statistically. It is not necessary that the statement is logical; it may be illogical (if statistics verifies it, then this is statistical hypothesis). It is a statistical hypothesis that shows statement with one or more parameters (like k-means, percentage, etc.) which are population measure under study.
3. **Substantive Hypothesis:** The objective is this hypothesis to provide the answer to the raised questions and answer to the felt problem of a research study in a significant way.

According to **Best and Kalm**, a substantive hypothesis is “a formal affirmative statement predicting a single research outcome, a tentative explanation of the relationship between two or more variables”.

According to **Stangor**, “a substantive or research hypothesis can be defined as a specific and falsifiable prediction regarding the relationship between or among two or more variables, the research hypothesis states the existence of a relationship between the variables of the interest and the specific direction of that relationship”.

5.2.4 Hypothesis Decision Table

These tables are the combination of rows and columns and every row corresponds to a single rule. Columns in the table explain the action and condition for the rules. When every condition is valid, the action of a given rule is performed. Table 5.1 shows a hypothesis decision table.

Table 5.1: Hypothesis Decision Table

<i>Decision</i>	<i>Null Hypothesis</i>	
Accept H_0	Correct Decision	Type II Error
Reject H_0	Type I Error	Correct Decision

5.2.5 Element of Hypothesis Testing

Following are the two basic elements of hypothesis testing.

1. **Test Statistic:** It is determined from the sample. To test the statistic, the researchers assume that the hypothesis of population parameter is true and then compares the

actual value with the hypothetical value of the parameter. If the difference is small, then the hypothesis is accepted, else rejected.

A statistic on which the decision can be based, whether to accept or reject a hypothesis, is called Test Statistic (T.S.). Following are the examples of test statistic:

$$T.S.Z. = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} - \frac{\sigma_2^2}{n_2}}} \text{ (for Z-test, for two samples)}$$

In case of one sample, test statistic (T.S.) is given as below:

$$T.S.Z. = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \text{ where } \frac{\sigma}{\sqrt{n}}$$

- 2. Rejection Region:** The rejection region is a subset of the sample space and is associated with the statistics test. If the observed sample falls only within this set, then the researcher can reject the null hypothesis and choose its alternative.

Hypothesis Test of the Form:

$$H_0 : \beta = \beta^+, H_1 : \beta \geq \beta^+$$

Example 1: 50 smokers were questioned about the number of hours they sleep each day. One wants to test the hypothesis that the smokers need less sleep than the general public which needs an average of 7.7 hours of sleep. The steps below are followed:

- (i) Compute a rejection region for a significance level of 0.05.
- (ii) If the sample mean is 7.5 and the population standard deviation is 0.5, what can be concluded?

Solution: The null and alternative hypothesis are as follows:

$$H_0 : \mu = 7.7$$

$$H_{01} : \mu \leq 7.7$$

This is a left-tailed test. The z-score that corresponds to 0.05 is -1.645 . The critical region is the area that lies to the left of -1.645 . If the x-value is less than -1.645 , then one will reject the null hypothesis and accept the alternative hypothesis. If it is greater than -1.645 , then one will fail to reject the null hypothesis and say that the test was not statistically significant.

$$Z = \frac{7.5 - 7.7}{0.5/\sqrt{50}} = -2.83$$

Since -2.83 are to the left of -1.645 , it is the critical region. Hence, the null hypothesis is rejected and the alternative hypothesis is accepted. One can conclude that smokers need less sleep.

5.2.6 Steps/Procedure of Hypothesis Testing

Step 1. Set up a Hypothesis: To set up the hypothesis, first set the null hypothesis (denoted by H_0). Null hypothesis is tested for the possible rejection under the assumption that it is true. It is expressed in the form of equation.

Null hypothesis is given as follows:

$$H_0 : \mu = \mu_0$$

where, μ = Population means

μ_0 = Hypothesized value of the population mean

For example, to test whether a population mean is equal to 140, the null hypothesis can be set as “population means is equal to 140”.

$$H_0 : \mu = 140$$

Now, set the alternative hypothesis which is rationally opposite to the null hypothesis and is denoted by H_1 .

Alternative hypothesis is given as:

$$H_1 : \mu \neq \mu_0 = \mu_0 = 0$$

Therefore, $H_1 : \mu < \mu_0$

$$H_0 : \mu < \mu_0$$

For the above example, the alternative hypothesis can be set as “population mean is not equal to 140”.

So, $H_1 : \mu \neq 140$

This gives the other alternative hypothesis:

$H_1 : \mu = 140$, Indicates, *i.e.*, the population mean is less than 140.

$H_1 : \mu = 140$, Indicates, *i.e.*, the population mean is greater than 140.

Step 2. Set up a Suitable Significance Level: Significance level is related to the null hypothesis and denoted by α . The size of the rejection or critical region is also known as the level of significance. Ensure that the obtained result is free from decision-maker’s biases of choice. For this, the significance level must always be determined before we draw samples.

0.01, 0.05 and 0.10 are the most commonly used levels of significance by researchers.

Step 3. Test Statistic: In this step, one can select the suitable statistical test for the analysis. On the basis of the type, number and level of data, statistical test is decided. To select the statistical test, statistics (mean, proportion, etc.) used in the study is also considered to get the good result.

Step 4. Doing Computations: After selecting the statistical test, the researcher tests the performance of different calculation from random sample of size n . Here, the researcher tests the

performance of different calculation from random sample of size n . Here, the researcher also computes the standard error.

Step 5. Making Decision: In this step, the researcher takes the decision that the null hypothesis is to be rejected or accepted.

Following conditions are checked to accept and reject a null hypothesis:

- (i) When the calculated value is less than the critical value, the test statistics falls in the acceptance region and null hypothesis is accepted.
- (ii) When the calculated value is greater than the critical value, the test statistics falls in the rejection region and null hypothesis is rejected.

Generally, in the testing of hypothesis, 5% level of significance ($\alpha = 0.05$) is used to take the decision.

5.2.7 Advantages of the Tests of Hypothesis

1. It is suitable for establishing the focus and directions in a research effort.
2. It helps the researcher in the development of a hypothesis force which shapes the purpose of the research movement.
3. It determines which variables will not be measured in a study and which will be measured.

5.2.8 Disadvantages of the Tests of Hypothesis

1. The tests should not be used in a mechanical fashion. They should keep in mind that testing is not decision-making itself. The tests are only useful aids for decision-making. Hence, “proper interpretation of statistical evidence is important for intelligent decisions.”
2. Test does not explain the reason as to why the differences exist, say between the means of the two samples. They indicate whether the difference is due to fluctuations of sampling or because of other reasons but the tests do not tell us as to which is/are the other reason(s) causing the difference.
3. Results of significant test are based on probabilities and cannot be expressed with full certainty. When a test shows that a difference is statistically significant, then it simply suggests that the difference is probably not by chance.
4. Statistical inferences based on the significant tests cannot be said to be entirely correct evidences concerning the truth of the hypothesis. This is especially in case of small samples where the probability of drawing erring inferences happens to be generally higher. For greater reliability, the size of sample is sufficiently enlarged.

5.3 IMPORTANT TERMS

5.3.1 Null Hypothesis

A statistical hypothesis which is stated for the purpose of possible acceptance is called null hypothesis. It is usually denoted by the symbol H_0 . For example, the null hypothesis may be expressed symbolically as:

According to **Prof. R.A Fisher**, “Null hypothesis is the hypothesis which is tested for possible rejection under the assumption that it is true.” The following may be borne in mind in setting the null hypothesis:

$$H_0: \mu = 162 \text{ cms}$$

1. If we want to test the significance of the difference between a statistic and the parameter or between two sample statistics, then we set up the difference of a null hypothesis that it is not significant. This means that the difference is just due to fluctuations of sampling:

$$H_0: \mu = \bar{X}$$

2. If we want to test any statement about the population, then we set up the null hypothesis that it is true. For example, if we want to find whether the population mean has specified value μ_0 , then we set up the null hypothesis,

$$H_0: \mu = \mu_0$$

5.3.2 Alternative Hypothesis

Any hypothesis which is complementary to the null hypothesis is called an alternative hypothesis and is usually denoted by H_1 or H_a . For example, if we want to test the null hypothesis that the average height of the soldiers is 162 cms, *i.e.*,

$$H_1: \mu = 162 \text{ cms} = \mu_0 \text{ (say)}$$

Then the alternative hypothesis could be:

$$H_1: \mu \neq \mu_0 \text{ (i.e., } \mu > \mu_0 \text{ or } \mu < \mu_0)$$

$$H_1: \mu > \mu_0$$

$$H_1: \mu < \mu_0$$

5.3.3 Errors in Hypothesis Testing

Two types of errors occur in hypothesis testing:

1. Type I Error
2. Type II Error

Following possibilities occur during the hypothesis testing.

1. Test leads to rejection while hypothesis is true. (Type I Error)
2. Test leads to acceptance while hypothesis is false. (Type II Error)

3. Test leads to acceptance and hypothesis is also true. (No Error)
4. Test leads to rejection and hypothesis is false. (No Error)

The first two possibilities lead to errors. If we reject a hypothesis when it should be accepted (possibility 1), then the error is known as type I error, and if we accept a hypothesis when it should be rejected (possibility 2), then the error is known as type II error.

Consider the following table which provides the basic idea of type I and type II error.

	<i>Accept H_0</i>	<i>Reject H_0</i>
H_0 is true	No error	Type I error
H_0 is false	Type II error	No error

5.3.3.1 Type I Error

If we reject the correct hypothesis, then it comes under Type I Error and denoted by α .

So, α = Probability of Type I Error = Probability of rejecting H_0 when H_0 is true.

When the difference of the two means is zero and the test indicates the rejection of the null hypothesis, then it is known as Type I Error.

5.3.3.2 Type II Error

If we accept the wrong hypothesis, then it comes under the Type II Error and denoted by β .

So, β = Probability of Type II Error = Probability of accepting H_0 when H_0 is not true (false).

When the difference of the two means is not zero and the test indicates the acceptance of the null hypothesis, then it is known as Type II Error.

5.3.4 Level of Significance

We check the validity of the null hypothesis (H_0) against the alternative hypothesis (H_1) at a certain level of significance to set up the hypothesis. Researchers generally use the 5% and 1% level of significance in the research.

Factors affecting the level of significance:

1. Size of the difference between sample means,
2. Total size of the sample,
3. Variability of measurements within the sample, and
4. Direction (directional or non-directional) of the difference.

Note: A directional hypothesis predicts the direction of the difference between means and difference between proportions, etc.

5.3.5 Degree of Freedom

In statistics, the number of elements in the calculation which are able to vary is known as degree of freedom.

For example, consider that the mean of the data is 30 and that the values are 20, 15, 45, and one value is unknown. To find the mean of a list of data, add all the data and divide it by the total number of values.

Now, if the unknown value is x , then using the mean formula:

$$= \frac{20 + 15 + 45 + x}{4} = 30$$

So, $x = 40$.

Now, take the situation where two values are missing and are denoted by x and y . So, using the mean formula, we find that

$$= \frac{20 + 15 + 45 + x + y}{5} = 30; \frac{80 + x + y}{5} = 30; x + y = 70$$

So, $x = (70 - y)$.

It shows that, when we select a value for x , then the value for y is determined. This shows that here the degree of freedom is one. For n samples, the degree of freedom will be $(n - 1)$.

For example, for the sample size of 15, the degree of freedom will be 14. For the $s \times t$ size contingency table, the degree of freedom will be $((s - 2)(t - 1))$

where, s = Number of columns and

t = Number of rows

Now, if the contingency table is of 2×2 , then the degree of freedom is:

$$(2 - 1)(2 - 1) = 1.$$

5.3.6 One-tailed and Two-tailed Tests

When the population parameter is known in a hypothesis test and it falls to the right or the left of the center of the sampling distribution, then it is known as one-tailed test. This test is used to check the increase and decrease in the parameter. So, it is used to test the null hypothesis.

For example, junior class students will not score significantly greater than higher class students in an IQ test; women are not significantly stronger than men. In every case, the null hypothesis predicts the direction of the expected difference. When hypothesis predicts that one value is dissimilar from another value and does not predict which is higher, then such hypothesis is known as two-tailed (non-directional) test.

5.3.7 Power of Statistical Test/Measuring the Power of Hypothesis Test

The power of a test denotes the capability of the test to make a right decision, *i.e.*, to reject the null hypothesis in case it is false. The power of a test can also be described as the probability of not making a type 2 error. Mathematically, it can be expressed as follows:

$$\text{Power} = 1 - p(\text{Type II Error}) = 1 - \beta$$

The maximum power a test can have is 1, the minimum is 0. Ideally, we want a test to have high power, close to 1. A power curve is the result of plotting $1 - \beta$ for possible value of the population parameter. Absolute care should be taken not to reject the null hypothesis (α) when it is correct.

5.4 TYPES OF HYPOTHESIS TESTING

Hypothesis testing can be categorized as follows:

1. On the Basis of Size of Sample

(i) Test of Significance for Small Sample: When the sample size (n) is less than 30 ($n < 30$), then it is known as small sample. In case of small sample, following tests are used:

- (a) t-test
- (b) z-test
- (c) F-test

(ii) Test of Significance for Large Sample: When the sample size (n) is greater than 30 ($n > 30$), then it is known as large sample. Generally, large sample is concerned with the variables such as height, weight, etc. Following are the most prominent large sample tests:

- (a) z-test
- (b) χ^2 test

Table 5.2: Difference between Large and Small Sample Tests

<i>Basis of Difference</i>	<i>Large Sample</i>	<i>Small Sample</i>
Sample Size	It is greater than 30, i.e., $n > 30$.	It is less than 30, i.e., $n < 30$.
Assumption for Test of Significance	The random sampling distribution of a statistic is approximately normal.	Assumption of normality is made unless otherwise stated.
Comparison from Population	Different values of samples are close to population values.	Sample values and population values differ significantly.
Standard Error	This concept is used while testing for significance.	Generally, this concept is not used.
Tests of Hypothesis	z-test and Chi-square or χ^2 test.	t-test, z-test and F-test.

2. On the Basis of Parameter Used

(i) Parametric Tests: Parametric tests are based on models with some assumptions. If the assumptions hold well, these tests may offer a more powerful tool for analysis. Parametric tests make certain assumptions about the population from which the samples are taken. Generally, non-parametric tests make far lesser assumptions and can be utilized for testing different types of data. However, parametric tests are preferred more because non-parametric tests are more susceptible to errors on account of not being less sensitive in identifying variations in samples or in ascertaining the affect that an independent variable may have on a dependent variable.

(ii) Non-parametric Tests: The practical data may be abnormal and/or it may not be possible to estimate the parameter(s) of the data. The tests which are used in such situations are called non-parametric tests. As the data is free from being classified into different parameters, these tests are called non-parametric tests. The level of calculation required in non-parametric tests is quite less as the computation of the parameters is not required. These tests can be applied on a very small sample size especially while conducting a pilot study. As the interval data is typically converted into ordinal data, there is no need to make any assumption about the parent population. The biggest advantage of using non-parametric tests is that they can handle ordinal level of data.

5.5 ONE-SAMPLE T-TEST

5.5.1 Introduction

A particular frequency distribution is obtained when a very large number of small samples are taken from a population and the mean of these samples is used to plot the frequency distribution. It is called **Student's t-distribution**. One-sample t-test is used to compare the mean of one sample to a known standard (or theoretical/hypothetical) mean (μ).

Sir William Gossett and **R.A. Fisher** are considered to be the great contributors towards the theory of small samples. Under the name 'students', Gossett published his findings in the year 1905 and thus this theory is commonly known as 'students' distribution', 't-test' or 'students' t-distribution'. The t-distribution can be implemented in situations when the sample size is 30 or less and no information is available with respect to the population standard deviation.

The formula is:

$$t = \frac{(\bar{X} - \mu)}{\sigma} \sqrt{n}$$

$$\text{where, Sample Standard Deviation } (\sigma) = \sqrt{\frac{\sum(X - \bar{X})^2}{n-1}}$$

If sample SD is given without using $n - 1$ as denominator, then

$$T = \frac{(X - \mu)}{\sigma / \sqrt{n-1}}$$

5.5.2 Properties of t-Distribution

Different properties of t-test are as follows:

1. Similar to normal distribution, a bell-shaped frequency curve is also associated with t-distribution which is symmetrical in nature. Degree of freedom is the only parameter which determines the shape of the curve. The shape of curve changes with the change in degree of freedom.
2. Similar to normal curve, the mean of t-distribution is zero.

3. The value of variance of t-distribution is usually greater than 1 and distribution becomes normal distribution in case of sample size being very large.
4. It can also be used for large samples.
5. The value of t-distribution ranges from -1 to $+1$.

It is clear that when the degree of freedom changes, the shape of the t-distribution also changes.

1. A t-distribution is lower at the mean and higher at the tail than a normal distribution.
2. The t-distribution has proportionately greater area in its tail than the normal distribution. Interval widths from t-distributions are therefore wider than those based on the normal distribution.

5.5.3 Assumptions for One-sample t-test

1. The parent population from which the sample is drawn is normal.
2. The sample observations are random, i.e., the given sample is drawn by random sampling method.
3. The population standard deviation σ is not known.

Example 2: Is the temperature required to damage a laptop on an average is less than 110 degrees? Because of the price of testing, 30 laptops were tested to see what minimum temperature will damage the laptop. The damaging temperature averaged 108 degrees with a standard deviation of 3 degrees. (Use $\alpha = 0.05$.)

Solution: The test hypothesis is:

$$H_0: \mu = 110$$

$$H_1: \mu < 110$$

The t-statistic is:

$$t = \frac{108 - 110}{3/\sqrt{30}} = -\frac{2}{0.545} = -3.6$$

This is a one-tailed test. So, one can go to t-table with 19 degrees of freedom to find that:

$$t_c = 1.69$$

$$\text{Since } -3.67 < -1.69$$

It is seen that the test statistic falls in the critical region. One accepts the null hypothesis and concludes that there is sufficient evidence to suggest that the temperature required to damage a laptop on an average is less than 110 degrees.

5.5.4 Types of t-test

There are three types of T-tests:

1. One-sample t-test
2. Independent Sample t-test
3. Paired Sample t-test

5.5.5 Interpreting the Output of One-sample t-test

If the variance of a normal population is unknown and one wants to determine the mean of a sample which is drawn from the population, and it deviates significantly from a stated value (the hypothetical value of the population mean), then the following statistic is used:

$$t = \frac{(\bar{X} - \mu)\sqrt{n}}{\sigma}$$

where, $\bar{X}$ = the mean of the sample,

μ = the actual or hypothetical mean of the population,

N = the sample size,

σ = the standard deviation of the sample.

$$\sigma = \sqrt{\frac{\sum(X - \bar{X})^2}{n-1}} \text{ or } \sigma = \sqrt{\frac{\sum d^2 - n(d)^2}{n-1}} = \sqrt{\frac{1}{n-1} \left[\sum d^2 - \frac{(\sum d)^2}{n} \right]}$$

where, d = deviation from the assumed mean.

Confidence or Fiducial Limits for μ

For given degrees of freedom v and for a given level of confidence α , the fiducial limits for μ are given by:

$$\bar{X} \pm t_{\alpha}(v) X \frac{8}{\sqrt{n}}$$

Thus, 95% confidence limits (or 5% level of significance or $\alpha = .05$) for μ are:

$$\bar{X} \pm t_{.05} X \frac{8}{\sqrt{n}} \text{ and}$$

99% confidence limits (or $\alpha = .01$) for μ are:

$$(461) \pm t_{.01} X \frac{8}{\sqrt{n}}$$

Procedure:

Step 1. Set up the null hypothesis.

Step 2. Then we find standard deviation from the following formula:

(i) When deviation is taken from actual mean:

$$\sigma = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}$$

(ii) When deviation is taken from assumed mean:

$$\sigma = \sqrt{\frac{\sum d^2 - n(d)^2}{n-1}} = \sqrt{\frac{1}{n-1} \left[\sum d^2 - \frac{(\sum d)^2}{n} \right]}$$

Step 3. Degree of freedom = $v = n - 1$.

Step 4. Then put this value of standard deviation and find the value or t .

Step 5. If $t_{cal} < \text{table value}$ at the significance level, then the null hypothesis is accepted. If $t_{cal} > \text{table value}$ at the significance level, then the null hypothesis is rejected and one accepts alternative hypothesis.

Example 3: The mean weekly sale of soap bars in departmental stores was 146.3 bars per store. After an advertising campaign, the mean weekly sales in 22 stores for a typical week increased to 153.7 and showed a standard deviation of 17.2. Was the advertising campaign successful? (Given tabulated of t for 21 d.f. at 5% level of significance = 1.72.)

Solution: Given, $n = 22$, $\bar{X} = 153.7$, $\sigma = 17.2$

Null Hypothesis:

$H_0 = \mu = 146.3$ bars, i.e., the difference between $\bar{X}$ and μ is not significant.

In other words, the advertising campaign is not successful.

Alternative Hypothesis:

$H_1: > 146.3$ bars (right-tailed)

Test Statistics:

Under the null hypothesis, the test statistic is:

$$t = \frac{\bar{X} - \mu}{s / \sqrt{n-1}}$$

$$\text{Now, } t = \frac{153.7 - 146.3}{17.2 / \sqrt{21}} = \frac{7.4\sqrt{21}}{17.2} = 1.9716$$

Tabulated value of t for 21 degree of freedom at 5% level of significance level for single-tailed test is 1.72.

Since the calculated value is greater than the tabulated value, it is significant and we reject the null hypothesis. Hence, the advertising campaign was successful.

Example 4: Ten workers of a factory are selected at random. The number of units produced by them on a working day was as follows: 71, 72, 73, 75, 76, 77, 78, 79, 79, 80.

On the basis of the given data, is it reasonably correct to say that the mean number of units produced by them is 78 (for $v = 9$, $t_{0.05} = 2.262$)?

Solution:

Calculation from Mean and Standard Deviation

X	$x = (X - \bar{X})$	x^2
71	-5	25
72	-4	16
73	-3	9
75	-1	1
76	0	0
77	1	1
78	2	4
79	3	9
79	3	9
80	4	16
$\Sigma X = 760$		$\Sigma x^2 = 90$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{760}{10} = 76$$

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{90}{10}} = \sqrt{9} = 3$$

- (i) **Null Hypothesis H_0 :** $\mu = 78$, i.e., the difference between $\bar{X}$ and μ is not significant or $\bar{X} = \mu$ (Two-tailed Test).

Alternative Hypothesis H_1 : $\mu \neq 78$

- (ii) Data given $\bar{X} = 76$, $\mu = 78$, $\sigma = 3$, $n = 10$
 (iii) Test statistic of the value of t is:

$$t = \frac{\bar{X} - \mu}{s / \sqrt{n-1}} = \frac{76 - 78}{3 / \sqrt{10-1}} = \frac{-2}{3/3} = -2$$

For $v = 9$, $p = 0.05$ (or 5%). The critical value of t for two-tailed test = 2.262. The calculated value is less than the table value. Thus, null hypothesis is accepted and it is concluded that the mean number of units produced by workers is 78.

Example 5: The random sample of 10 boys had the following IQs: 70, 120, 110, 101, 88, 83, 95, 98, 107, and 100. Can these data support the assumption of a population mean IQ values of samples of 10 boys exist?

Solution: Let us consider that the null hypothesis H_0 shows that the data are consistent with the assumption of a mean IQ of 100 in the population, i.e., $\mu = 100$ and H_1 is the alternative hypothesis, $\mu \neq 100$.

Under H_0 , the test statistic is:

$$t = \frac{(\bar{X} - \mu)}{\sqrt{\frac{\sigma^2}{n}}} \sim t_{(n-1)}$$

where, $\bar{X}$ and σ^2 are to be estimated from the sample value of IQs.

Table 5.3: Calculation for Sample Mean and SD

X	$(X - \bar{X})$	$(X - \bar{X})^2$
70	-27.2	739.84
120	22.8	519.84
110	12.8	163.84
101	3.8	14.44
88	-9.2	84.64
83	-14.2	201.64
95	-2.2	4.84
98	0.8	0.64
107	9.8	96.04
100	2.8	7.84
Total	972	1833.60

Therefore, $n = 10$, $\bar{X} = \frac{972}{10} = 97.2$ and $\sigma^2 = \frac{1833.60}{9} = 203.73$

$$\therefore |t| = \frac{197.2 - 1001}{\sqrt{\frac{203.73}{10}}} = \frac{2.8}{\sqrt{20.37}} = \frac{2.8}{4.514} = 0.62$$

Tabulated value $t_{0.05}$ for $(10 - 1)$, i.e., 9 degree of freedom (d.f.) for two-tailed test is 2.262.

Conclusion: Because calculated $t < t_{0.05}$ for 9 d.f., H_0 may be accepted at 5% level of significance and we may determine that the data are consistent with the assumption of mean IQ of 100 in the population. The 95% of confidence limits in which the mean IQ values of samples of 10 boys will exist are given by following formula:

$$\bar{X} \pm t_{0.05} \frac{\sigma}{\sqrt{n}} = 97.2 \pm 2.262 \times 4.514 = 97.2 \pm 10.21 = 107.41 \text{ and } 86.99$$

Therefore, 95% of confidence interval is [86.99, 107.41].

5.5.6 Conducting One-sample t-test in PSPP

One-sample t-tests are used for comparing data from a sample to population data.

For example, the local school system might want to compare standardized test scores from local students to a state or national average. This would address the question of whether local

students are the same (null hypothesis) or different (alternative hypothesis, non-directional) than the population.

Let us pretend that the data in the one-sample column of the test example. Save file are the standardized math test scores from 10 ninth grade students at the local high school. These can be compared to a national average, which we say is 100. The null hypothesis is that our local students are equal or similar to the national average – the average, being either higher or lower. Begin this analysis by going to the Analyze menu, then the Compare Means folder. Select the One-sample t-test command.

The first step is to choose the desired data variable from the list on the left by moving it to the Test Variable(s) field. The “Math Scores” label represents the One-sample variable. Move this variable right to the Test Variable(s) field. The second step is that we must tell PSPP the population means value. This is the mean that our sample will be compared to. Enter 100 in the Test Value field to represent the population mean. If no mean is given, PSPP will assume that the population mean is zero.

The output is a table with descriptive statistics and the t-test results.

T-TEST
T-TEST/TESTVAL=100
/VARIABLES=OneSample/MISSING=ANALYSIS
/CRITERIA=CI(0.95).

One-sample Statistics

	<i>N</i>	<i>Mean</i>	<i>SD</i>	<i>SE Mean</i>
Math score for local students	10	100.40	5.46	1.72

One-Sample Test

	<i>Test Value = 100,000000</i>					
	<i>T</i>	<i>d. f.</i>	<i>Sig. (2-tailed)</i>	<i>Mean Difference</i>	<i>95% Confidence Interval of the Difference</i>	
					<i>Lower</i>	<i>Upper</i>
Math score for local students	.23	9	.822	.40	-3.51	4.31

The one-sample t-test will compare these two means to see if these are similar or different. The mean values in this example are quite similar. The *p*-value is shown in the Sig. (2-tailed) field, where “Sig.” is short for statistical significance.

This *p*-value is much larger than .05. So, we will keep the null hypothesis. This difference is not statistically significant. In other words, our sample and the population mean are essentially the same. Our local school kids are similar to national averages on this standardized test. The APA style for reporting t-test statistics is *t* (degrees of freedom) = *t* score, *p* = probability value. The symbols for *t* and *p* should be in italic font. For example, the proper APA style would be *t*(9) = .23, *p* = .822.

5.5.7 Conducting One-sample t-test in R Commander

The single-sample t-test tests a null hypothesis that the population mean is equal to a specified value. If the value is zero (or not entered), then the confidence interval for the sample mean is given.

1. Statistics → Means → Single-sample t-test.
2. Select the variable of interest.
3. Typically, the confidence level of 0.95 is used.
4. Three alternative hypothesis are possible:
 - (i) The mean is less than the specified value.
 - (ii) The mean is more than the specified value.
5. OK.
6. Interpretation: If the p -value is below the significance threshold, then the means is not equal to the specified value

5.6 PAIRED SAMPLE T-TEST

5.6.1 Introduction

The paired sample t-test is a statistical test that compares the means and standard deviation of two related groups to find whether there is a significant difference between them. It is also known as dependent or correlated t-test. The significant difference happens when the difference between two groups are unlikely to be due to sampling error. A group can be related to similar group of persons or the same items.

A paired sample t-test are more better than unpaired t-test as using the similar participants or item eliminates variation among the samples that could be caused by anything other. To study the effect of some action or process, this test has to be used. For example, when a stimulus is given to some patients and their blood pressure after and before giving the stimulus is tested, and then with the help of paired t-test, it is analyzed whether the stimulus is effective or not. Also, the sales of some departmental stores are recorded for definite period after and before some sales advertising procedure and then it is analyzed whether the sales advertising procedure is effective in increasing the sales. In similar manner, the performance of a batch of students is recorded before and after training and it is checked if training is effective or not.

Assume that $x_1, x_2, \dots, x_n$ denotes the sales of a product in n departmental stores for a specific period before sales advertising and $y_1, y_2, \dots, y_n$ the corresponding sales of the same departmental stores for the same period length after applying certain sales advertising procedure. Then (x_1, y_2) , (x_2, y_2) , $\dots$, (x_n, y_n) are sets of sales data before and after the sales advertising in the n departmental stores.

We apply paired t-test to test the significance of the difference between the sales x and y . Suppose $d_i = x_i - y_i$ or $y_i - x_i$, $i = 1, 2, \dots, n$.

The differences are $x_i - y_i$ or $y_i - x_i$. So, Σd_i is positive. In case, Σd_i is negative, we take positive value of d , i.e., $|d|$.

Often, the use of dependent or paired samples will enable us to perform a more precise analysis, because they will allow us to control the extraneous factors. The t-test based on paired observations is defined by the following formula:

$$t = \frac{\bar{d} - 0}{S} \times \sqrt{n} \text{ or } t = \frac{\bar{d} \sqrt{n}}{S}$$

where, $\bar{d}$ = Mean of the difference

S = Standard deviation of the differences

The value of S is calculated as follows:

$$S = \sqrt{\frac{\Sigma(d - \bar{d})^2}{n - 1}} \text{ or } \sqrt{\frac{\Sigma d^2 - n(\bar{d})^2}{n - 1}}$$

It should be noted that t is based on $n - 1$ degree of freedom.

5.6.2 Hypothesis of a Paired Sample t-test

1. **H_0 (Null Hypothesis):** “There is no significant difference between the means of two groups.”
2. **H_1 (Alternative Hypothesis):** “There is a significant difference between two population means, and this difference is unlikely to be caused by sampling error.”

5.6.3 Assumptions of a Paired Sample t-test

1. The dependent variable is distributed normally.
2. The observations are sampled independent.
3. The dependent variable is measured on an incremental level like ratios.
4. Independent variables must consist of two matched pairs.

5.6.4 Procedure for Testing Paired Sample t-test

Step 1. Set up the null hypothesis.

Step 2. Then we find value of standard deviation from the formula:

$$S = \sqrt{\frac{\Sigma(d - \bar{d})^2}{n - 1}} \quad \bar{d} = \text{mean of difference}$$

Step 3. Degree of freedom = $n - 1$ obtained the table value at the level of significance.

Step 4. Putting this value of standard deviation, find the value of t .

Step 5. If $t_{cal} < \text{table value}$ at the level of significance, then null hypothesis is accepted.

If $t_{cal} >$ table value at the level of significance, then null hypothesis is rejected and alternative hypothesis is accepted.

Example 6: An Aquarius Health Club has been promoting a demanding programme for body conditioning. The club claims that after 1 month, the average participant should be more able to do least eight push-ups in 2 minutes than he or she could do at the beginning. Does the Club's claim is right on the random sample of 10 programme participants given below? Take the 0.025 level of significance.

Participant	1	2	3	4	5	6	7	8	9	10
Before	38	11	34	25	17	38	12	27	32	29
After	45	24	41	39	30	44	30	39	40	41

Solution: H_0 (Null Hypothesis): $\mu_1 - \mu_2 = 8$

H_1 (Alternative Hypothesis): $\mu_1 - \mu_2 > 8$ (The difference is more than 8.)

Table 5.4: $\alpha = 2.5\%$ Level of Significance

Participant	Before	After	Difference = d	(Difference) ² = $(d)^2$
1	38	45	+7	49
2	11	24	+13	169
3	34	41	+7	49
4	25	39	+14	196
5	17	30	+13	169
6	38	44	+6	36
7	12	30	+18	324
8	27	39	+12	144
9	32	40	+8	64
10	29	41	+12	144
Total			110	1344

$$\bar{d} = \frac{\sum d}{n} = \frac{+110}{10} = +11;$$

$$S.E. = \sqrt{\frac{n \sum d^2 - (\sum d)^2}{n \sqrt{n-1}}} = \sqrt{\frac{10 \times 1344 - (110)^2}{10 \times 3}} = 1.22$$

One-tailed t-test is used here, as the Club needs to prove that the increase is more than 8. An appropriate t-value at 2.5% significance level with 9 d.f. is the t-value for two-tailed t-test at 0.05 significance level with 9 d.f. and is equal to 2.262.

$$\text{The upper limit} = 8 + 2.262(1.22) = 10.76$$

As the mean difference 11 lies outside the acceptance region, hence null hypothesis H_0 is rejected. This result supports the Club's claim.

Example 7: A random sample of 8 companies was taken from companies having at least 15% dividends. Then it has to test whether the statement of dividends has any effect on the market price of a share of a company. A data of share prices of 8 companies are as follows:

	1	2	3	4	5	6	7	8
Market Price 10 days before dividends were declared	70	65	112	58	25	147	95	68
Market Price 10 days after dividends were declared	80	85	110	64	32	159	100	70

1. Can we predict that the increase in average market price after declaration of dividends is significant at 5% level of significance? Consider that population is normal distributed.
2. List the all possible inadequate in the methodology accepted to test the validity of the above statement 1.

Solution: This question is based on paired sample t-test:

1. H_0 : The increase in market price after declaration of dividends is not significant.

H_1 : The increase in market price after declaration of dividends is significant.

Let us suppose that X and Y indicate the market price 'before' and 'after' declaration of dividends. Then,

$$H_0: \mu_1 - \mu_2$$

[μ_1 and μ_2 is the average market price before and after the declaration of the dividend]

$$H_1 = \mu_1 > \mu_2$$

i.e.,

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1 = \mu_1 > \mu_2$$

Assume $d = Y - X$, where

X	Y	d	d^2
70	80	10	100
65	85	20	400
112	110	-2	4
58	64	6	36
25	32	7	49
147	159	12	144
95	100	5	25
68	70	2	4
		60	762

$$\bar{d} = \frac{60}{8} = 7.5; \quad t = \frac{\left| \bar{d} \right| \sqrt{n-1}}{\sqrt{n \sum d^2 - (\sum d)^2}} = \frac{7.5 \times 8 \times \sqrt{7}}{\sqrt{8 \times 762 - (60)^2}} = \frac{158.75}{50} = 3.18$$

The experiential t value is $t = \frac{7.5}{2.3569} = 3.18$ (with 7 degree of freedom).

The value 1.895 is the tabulated value of t at 5% level of significance for 7 degree of freedom.

Tabulated value of t (1.895) < calculated t -value (3.18)

Thus, we discard H_0 .

Hence, after declaration of dividends, there is significant increase in prices.

2. Inadequacies are as follows:

- (i) These have too small sample size.
- (ii) The circumstances prevailing on these dates may be different as the dividends may be declared on different dates.
- (iii) Although conclusion is that dividends affect market price positively, but there is not certain that the dividend is the only cause behind increase.
- (iv) A group of companies that have declared 15% or more dividends are considering only. All the companies which do not follow this group are excluded from the study. Their prices might have increment during the same period.

Example 8: A similar test was given to 12 participants to verify that whether a course in quantitative analysis improved performance. Their scores both before and after the course are shown in the table below:

Candidate	1	2	3	4	5	6	7	8	9	10	11	12
Before	44	40	61	52	32	44	70	41	67	72	53	72
After	53	38	69	57	46	39	73	48	73	74	60	78

Suppose these 12 candidates as a sample from a population and test at 5% level of significance in case the course was useful in terms of presentation on the test.

Solution: We apply here the sample t -test.

Candidate	1	2	3	4	5	6	7	8	9	10	11	12
Score before the Course	44	40	61	52	32	44	70	41	67	72	53	72
Score after the Course	53	38	69	57	46	39	73	48	73	74	60	78
Difference (d)	9	-2	8	5	14	-5	3	7	6	2	7	6
d^2	81	4	64	25	96	25	9	49	36	4	49	36

This is the problem of one-tailed t -test at 5% significance and 11 d.f.,

$t_{0.05} = 1.796$ (table value):

1. **Null Hypothesis:** There is no improvement due to course.

2. **Alternate Hypothesis:** Course has better performance.

$$\Sigma d = 60$$

$$\frac{\Sigma d}{n} = \frac{9 - 2 + 8 + 5 + 14 - 5 + 3 + 7 + 6 + 2 + 7 + 6}{12} = \frac{50}{12} = 5$$

$$\Sigma d^2 = 81 + 4 + 64 + 25 + 96 + 25 + 9 + 49 + 36 + 4 + 49 + 36 = 578$$

$$t = \frac{\left| \bar{d} \right| n \sqrt{n-1}}{\sqrt{n \sum d^2 - (\sum d)^2}} = \frac{5 \times 12 \times \sqrt{11}}{\sqrt{12 \times 578 - (60)^2}} = \frac{199}{\sqrt{3336}} = \frac{199}{57.758} = 3.445$$

As the calculated value of t (3.445) > table value (1.796), hence Null Hypothesis is rejected, i.e., the course has better performance.

Example 9: Before given a drug, the Systolic blood pressures of six hypertensive patients were 183, 179, 165, 190, 175 and 180 mm of Hg. After given a particular drug for one week, the blood pressures were recorded as 187, 175, 150, 180, 180 and 170 mm of Hg respectively. Determine could such differences arise due to chance?

Solution:

Null Hypothesis: $H_0 : \mu_1 = \mu_0$ or $\bar{D} = 0$

Alternative Hypothesis: $H_1 : \mu_1 \neq \mu_2$.

Calculation for $\bar{D}$ and estimated variance are shown in table below:

Patients	B.P. before Drug (X_1)	B.P. after Drug (X_2)	Difference $D = (X_1 - X_2)$	D^2
1	183	187	-4	16
2	179	175	4	16
3	165	150	15	225
4	190	180	10	100
5	175	180	-5	25
6	180	170	10	100
			$\Sigma D = 30$	$\Sigma D^2 = 482$

$$\text{Mean Difference, } \bar{D} = \frac{\Sigma D}{n} = \frac{30}{6} = 5$$

$$\text{Estimated Variance, } \sigma^2 = \frac{1}{n-1} \left[\Sigma D^2 - \left(\frac{\Sigma D}{\sqrt{n}} \right)^2 \right] = \frac{1}{6-1} \left[482 - \frac{(30)^2}{6} \right]$$

$$= \frac{1}{5} \left[482 - \frac{900}{6} \right] = \frac{1}{5} (482 - 150)$$

$$\therefore \sigma^2 = \frac{1}{5} \times 332 = 66.4 \text{ or } \sigma = \sqrt{66.4} = 8.15$$

$$\text{Standard Error of } \bar{D} = \frac{\sigma}{\sqrt{n}} = \frac{8.15}{2.45} = 3.33;$$

$$t = \frac{\bar{D}}{S.E. \text{ of } \bar{D}} \text{ or } t = \frac{5}{3.33} \Rightarrow \therefore t = 1.5$$

As the calculated value of $|t| = 1.5 < t_{0.04, 5} = 4.03$ (i.e., table value of t and 1% of significance difference with 5 d.f.). Thus, the null hypothesis is accepted and we conclude that there is no significance difference between the averages of before and after observations.

Example 10: In order to specify the result of specific oral contraceptive on weight gain, 9 healthy females were weighed before the start of its use and again after ending of 3 months' period.

Subject No.	Initial Weight (kg.)	Weight after 3 months (kg.)
1	48.0	49.2
2	56.4	57.2
3	52.0	56.0
4	60.0	58.0
5	54.0	56.0
6	56.0	57.2
7	48.0	47.2
8	56.0	56.4
9	52.0	52.8

Test and provide sufficient evidence that females experience gain in weight after 3 months of the oral contraceptive use.

Solution:

Subject No.	Initial Weight (kg.)	Weight after 3 months (kg)	Difference (D)	D^2
1	48.0	49.2	-1.2	1.44
2	56.4	57.2	-0.8	0.64
3	52.0	56.0	-4.0	16.00
4	60.0	58.0	2.0	4.00
5	54.0	56.0	-2.0	4.00
6	56.0	57.2	-1.2	1.44
7	48.0	47.2	0.8	0.64
8	56.0	56.4	-0.4	0.16
9	52.0	52.8	-0.8	0.64
Total	—	—	$\Sigma D = -7.6$	$\Sigma D^2 = 28.96$

$$\text{Mean difference } \bar{D} = \frac{\Sigma D}{n} = \frac{-7.6}{9} = -0.84$$

$$\text{Estimated Variance, } \sigma^2 = \frac{1}{n-1} \left[\Sigma D^2 - \left(\frac{\Sigma D}{\sqrt{n}} \right)^2 \right] = \frac{1}{9-1} \left[28.96 - \frac{57.76}{9} \right]$$

$$= \frac{1}{8} [28.96 - 6.42] = \frac{1}{8} \times 22.54$$

$$\therefore \sigma^2 = 2.82 \text{ or } \sigma = \sqrt{2.82} = 1.68$$

Null Hypothesis, H_0 : There is no weight gain after 3 months, *i.e.*, $\mu_1 - \mu_2$.

Alternate Hypothesis, H_1 : There is weight gain after 3 months of the oral contraceptive use, *i.e.*, $\mu_1 \neq \mu_2$ (two-tailed test).

Standard Error of Difference, $S.E.(\bar{D}) = \frac{\sigma}{\sqrt{n}}$

$$\therefore S.E.(\bar{D}) = \frac{1.68}{\sqrt{9}} = \frac{1.68}{3} = 0.56$$

$$\therefore t = \frac{\bar{D}}{S.E.(\bar{D})} = \frac{0.84}{0.56}$$

$$\therefore t = -1.5$$

Since calculated value $|t| = 1.5 < \text{table value of } t_{0.01, 8} = 3.3$, this implies that the null hypothesis is accepted, *i.e.*, there is no weight gain after 3 months of the oral contraceptive use.

5.6.5 Conducting Paired Sample t-test in PSPP

The most common research situation would be to compare a dependent variable measured before and after an experimental manipulation or a clinical treatment. These kinds of comparisons are commonly done in education and medical research. For example, an educator might want to compare students before and after instruction on a particular topic to see if the test scores have increased. A second example is that a medical doctor might want to compare patients before and after a medical intervention to see if their symptoms have improved.

It is important to keep in mind that these before-after experiments fall short of being true, pure experiments. The problem is that time is a variable in these designs that is changing along with other experimental treatments, which can make the interpretation of the outcomes uncertain. For example, consider the situation where a patient improves after a medical treatment. Is the change due to the medical intervention or the passage of time? It is possible that the patient would have gotten better on their own over time without any medical intervention. This ambiguity can obscure the interpretation of tracking people for changes over time.

Sometimes, groups can be related in ways that do not necessarily involve a before-after comparison. An important example is the study of identical twins to see how similar or different they are. The twins are related, quite literally, because they have the same genes. It would be appropriate to compare these twins with a test that takes into account that the comparison is not between two randomly selected individuals.

The t-test used for this situation is the paired sample t-test. This test is sometimes called the related-samples t-test or a dependent-samples t-test. We will call it a paired sample t-test because this is the terminology that PSPP uses.

The data for a paired sample t-test needs two variables. The most common form would be variables before (pre) and after (post) an intervention. In this example, we will assume that an

educator is comparing the math test scores for a class before and after instruction of a particular topic. The same people are tested repeatedly. So, a paired sample t-test would be a good choice.

<i>RepSamplePre</i>	<i>RepSamplePost</i>
101.00	140.00
102.00	145.00
99.00	132.00
98.00	155.00
103.00	165.00
105.00	148.00
104.00	147.00
99.00	151.00
101.00	145.00
100.00	147.00

In this example, math test scores are shown before instruction (*RepSamplePre*) and after instruction (*RepSamplePost*). The *Rep* in the variable name is short for repeated, which means that these people have been tested at least two times. The labels for this variable are “Math Pre Score” and “Math Post Score”.

The paired variables need to be drag from the list at the left to the Var 1 and Var 2 fields in the center. Be careful to select just the paired variables, which are Math Pre Score and Math Post Score in this case. When finished, press the OK button to conduct the analysis.

Period Sample Statistics				
	<i>Mean</i>	<i>N</i>	<i>SD</i>	<i>S.E. Mean</i>
Pair 1 Math Pre Score	101.20	10	2.30	.73
Math Post Score	147.50	10	8.72	2.76
Paired Sample Correlations				
	<i>N</i>	<i>Correlation</i>	<i>Sig.</i>	
Pair 1 Math Pre Score and Math Post Score	10	.19	.591	

The output begins with descriptive statistics for each group. This analysis shows a large difference between groups, with an increase of about 46 points on this math test from before instruction to after instruction. The right side of the bottom table of the output shows the t-test results:

<i>t</i>	<i>d.f.</i>	<i>Sig. (2-tailed)</i>
-17.07	9	.000

The APA style for reporting this result is $t(9) = -17.07, p = .001$. This test result is statistically significant because $p < .05$. It would be best to report this result as $p < .001$ rather than $p = .000$ because p -values of zero are not really possible.

Note that the t-test result is negative, which suggests a decrease, but the real result is that an increase occurred between the pre-score and the post-score measurement. This contradiction occurs because the paired t-test computations subtract the second variable from the first variable. One strategy for dealing with this contradiction would be to enter the post-variable as variable 1 and the pre-variable as variable 2. The resulting change scores would then reflect the direct of the change. For example, an increase would be reflected as a positive *t*-score. This approach is not absolutely necessary, but it might be useful in some situations to decrease potential confusion.

5.6.6 Conducting Paired Sample t-test in R Commander

In a paired experiment, there is a one-to-one correspondence between the values in the two samples. (*e.g.*, before and after treatment, paired subjects, *e.g.*, twins).

A paired approach is considered more sensitive as it is looking for a treatment difference excluding initial biological differences. As such, the null hypothesis for this statistical test is that the average difference (second measurement minus first) is zero.

Note: Data File Format

Need two columns; one that contains the first number in each dataset pair (*e.g.*, “before” data) and another column that contains the second number in each dataset pair (*e.g.*, “after” data). Pairs of numbers must be in the same row.

Example 11:

Subject	Weight Before	Weight After
1	60	57
2	75	73
3	67	66

1. Statistics → Means → Paired t-test.
2. Select the first variable.
3. Select the second variable.
4. Under the Options tab, there are two things to consider:
 - (i) The alternative hypothesis can be two-sided where the change in mean difference can be either an increase or a decrease or defined as an increase or a decrease.
 - (ii) The confidence level – the threshold at which to reject the null hypothesis and accept the alternate hypothesis. Typically, this is 0.05, which is a 5% risk that the difference is a sampling effect rather than a true population difference.
5. OK.
6. Interpretation:
 - (i) If the *p*-value is below the significance threshold, then the mean difference is not equal to 0.

- (ii) The mean of the difference indicates the average difference (variable 1 – variable 2).
- (iii) The 95% confidence interval is the confidence interval around the mean difference.

5.7 INDEPENDENT SAMPLE T-TEST

5.7.1 Introduction

By using a technique known as **Pooled t-test**, we will compare the means of the two independent samples. The unpaired t-test is a statistical technique that compares means of two independent or unrelated groups to find whether there is a significant variance between the two. It is also known as **independent t-test**.

If the two independent samples of size n_1 and n_2 with means $\bar{X}_1$ and $\bar{X}_2$ and standard deviation S_1 and S_2 are given and one wants to test the hypothesis that the samples come from the same normal population, then to carry out the test, we calculate the statistic as follows:

- (i) The test statistics which is followed in this case is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S} \times \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

where, $\bar{X}_1$ = mean of the first sample

$\bar{X}_2$ = mean of the second sample

n_1, n_2 = number of observations in the first and second sample

S = combined standard deviation

- (ii) The value of S is calculated by the following formula:

$$S = \sqrt{\frac{\Sigma(X_1 - \bar{X}_1)^2 + \Sigma(X_2 - \bar{X}_2)^2}{n_1 + n_2 - 1}}$$

The deviations should be taken from assumed means when the actual means are in fraction. In such a case, the combined standard deviation is obtained by applying the following formula:

$$S = \sqrt{\frac{\Sigma(X_1 - A_1)^2 + \Sigma(X_2 - A_2)^2 - n_1(\bar{X}_1 - A_1)^2 - n_2(\bar{X}_2 - A_2)^2}{n_1 + n_2 - 2}}$$

where, A_1 = Assumed mean of the first sample

A_2 = Assumed mean of the second sample

$\bar{X}_1$ = Actual mean of the first sample

$\bar{X}_2$ = Actual mean of the second sample

The degrees of freedom = $(n_1 + n_2 - 2)$

- (iii) When the number of observations and the standard deviation of the two samples are given, then the pooled estimate of standard deviation can be obtained as follows:

$$S = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

If the calculated value of t is greater than the tabulated value at any specified significance level (5% or 1%), then the difference between the sample means is said to be significant at specified level of significance. Otherwise, the data is said to be consistent with the hypothesis.

5.7.2 Hypotheses of an Unpaired t-test

1. **H_0 (Null Hypothesis):** “There is no significant difference between the means of two groups.”
2. **H_1 (Alternative Hypothesis):** “There is a significant difference between two population means, and this difference is unlikely to be caused by sampling error.”

5.7.3 Assumptions for Testing the Difference of Means

1. The universe from which the samples are drawn are normally distributed.
2. The two samples are random samples and have been drawn independently.
3. The variances of the two populations are equal. It follows that the standard deviations of the two universe or σ_1 and σ_2 are also equal.
4. Population variances are not known.

5.7.4 Procedure for Testing the Difference of Means

Step 1. Set up the null hypothesis.

Step 2. Then we find standard deviation from the following formula:

- (i) When deviation is taken from actual mean:

$$S = \sqrt{\frac{\Sigma(X_1 - \bar{X}_1)^2 + \Sigma(X_2 - \bar{X}_2)^2}{n_1 + n_2 - 1}}$$

- (ii) When deviation is taken from assumed mean:

$$S = \sqrt{\frac{\Sigma(X_1 - A_1)^2 + \Sigma(X_2 - A_2)^2 - n_1(\bar{X}_1 - A_1)^2 - n_2(\bar{X}_2 - A_2)^2}{n_1 + n_2 - 2}}$$

- (iii) When standard deviation of two samples is given:

$$S = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

Step 3. Degree of freedom = $n_1 + n_2 - 2$.

Step 4. Then put this value of standard deviation and find the value of t .

Step 5. If $t_{cal} < \text{table value of significance level}$, then null hypothesis is accepted.

If $t_{cal} > \text{table value of significance level}$, then null hypothesis is rejected and alternative hypothesis is accepted.

Example 12: Suppose that we compared TV watching hours for female and male students of a college and cannot discard the hypothesis that they had the same population means. A sample size for males was only about $\frac{1}{2}$ than females, but the standard deviation was larger for males. Following summary is given below:

	Sample	Size	Mean	Standard Deviation	Variance
Male	n_1	59	2.37	1.87	3.50
Female	N_2	116	1.95	1.51	2.28

A difference in means is given as 0.42 hours, the unpooled standard error is 0.281, 1.49 is test statistics, and 0.14 is resulting p -value. Examine the data and apply pooled t-test.

Solution: Here are the computations for the pooled t-test:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = 2.69, \quad S_p = \sqrt{2.69} = 1.64$$

$$\text{Pooled S.E. } \bar{X}_1 - \bar{X}_2 = \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}} = S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 1.64 \sqrt{\frac{1}{59} + \frac{1}{116}} = 0.262$$

$$t = \frac{\text{Sample Statistics} - \text{Null Value}}{\text{Pooled Standard Error}} = \frac{0.42}{0.262} = 1.60$$

Degree of Freedom (d.f.) = $59 + 116 - 2 = 173$; p -value = 0.11

While the p -value for pooled test still would not lead to rejection of null hypothesis, it is remarkably smaller than the p -value of 0.14 for unpooled test.

Example 13: In a mutation breeding experiment, gamma-irradiation effect was evaluated on 200 seed weight in grams per plant of a mungbean variety in M_2 generation. The experimenter obtained the following results. Analyze the data using the t-test and give your inference as regards the effect of gamma-irradiation.

200-seed Weight (Grams) per Plant

Control	2.9	3.2	3.5	3.4	3.1	4.0	3.8	3.0	4.1	4.0
Treated	2.9	2.8	3.0	3.5	3.7	3.3	3.0	3.1	2.9	2.8

Solution: Null Hypothesis, H_0 : There is no significant difference between the mean 10-seed weight per plant in the control and treatment.

200-seed Weight (Grams)

Control		Treated	
X_1	$(X_1)^2$	X_2	$(X_2)^2$
2.9	8.41	2.9	8.41
3.2	10.24	2.8	7.84
3.5	12.325	3.0	9
3.4	11.56	3.5	12.25
3.1	9.61	3.7	13.69
4.0	16	3.3	10.89
3.8	14.44	3.0	9
3.0	9	3.1	9.61
4.1	16.81	2.9	8.41
4.0	16	2.8	7.84
$\Sigma X_1 = 35$	$\Sigma(X_1)^2 = 124.32$	$\Sigma X_2 = 31$	$\Sigma(X_2)^2 = 96.94$

$$\bar{X}_1 = \frac{\Sigma X_1}{n_1} = \frac{35}{10} = 3.5$$

$$\bar{X}_2 = \frac{\Sigma X_2}{n_2} = \frac{31}{10} = 3.1$$

$$\Sigma X_1^2 - \frac{(\Sigma X_1)^2}{n_1} = 124.32 - \frac{(35)^2}{10} = 124.32 - 122.5 = 1.82$$

$$\Sigma X_2^2 - \frac{(\Sigma X_2)^2}{n_2} = 96.94 - \frac{(31)^2}{10} = 96.94 - 96.1 = 0.84$$

Degree of Freedom, d.f. = $(n_1 + n_2 - 2) = (10 + 10 - 2) = 18$

$$\text{Pooled Variance, } S_1^2 = \frac{1.82 + 0.84}{18} = \frac{2.66}{18} = 0.147$$

$$\text{Pooled Standard Deviation, } S_d = \sqrt{S_d^2} = \sqrt{0.147} = 0.38$$

$$\therefore t = \frac{\bar{X}_1 - \bar{X}_2}{S_d \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{|3.5 - 3.1|}{0.38 \sqrt{\frac{1}{10} + \frac{1}{10}}} = \frac{0.4}{0.38 \sqrt{\frac{1}{5}}} = \frac{0.4}{0.38 \times 0.45} = \frac{0.4}{0.17} = 2.35$$

Conclusion: Calculated t -value = 2.25

Tabulated t -value = 2.10 ($p = 0.01$)

d.f. = 18

The calculated ' t ' value for the difference of the two sample means is greater than the tabulated ' t ' value at 5% level of probability with 18 d.f. The difference between the two sample means is significant. The null hypothesis stating that the two sample means are not different is

rejected at $p < 0.05$. Therefore, gamma-irradiation has a very significant effect on 200-seed weight.

5.7.5 Conduction Independent — Sample t-test in PSPP

The example, which compared a local sample to a population mean, is not a true experiment. It is an interesting comparison for local educators. We could no, however, make strong conclusions about whether a particular kind of instruction was effective or not.

For this next example, we will do a true experiment. Let us say that there is a new method of math instruction that is believed to be superior to older methods of instruction. We will form two different classes. The control group will get the older, standard instructional method. Both groups will be measured on a standardized math test at the end of the semester. The results will be compared to see if the new method is effective.

In PSPP, we will need a variable that represents the math scores (dependent variable). A second categorical variable will be needed to code for which group each student is in, experimental or control. Here is what the data might look like in our example.

<i>IndSample</i>	<i>IndSampleGroups</i>
99.00	.00
98.00	.00
105.00	.00
102.00	.00
103.00	.00
120.00	1.00
125.00	1.00
122.00	1.00
133.00	1.00
128.00	1.00

The variable *IndSample* is the dependent variable with the math test scores. The variable *IndSample Groups* has a 0 or 1 coding scheme for representing the group membership. The 0 represents people in the control group. The 1 value is for people in the experimental group.

We will test for differences between the experimental group and the control group by using the independent-samples t-test. Choose the independent-samples t-test command by going to Analyze > Compare means > Independent Samples t-test.

The first step is to choose the variables. Move the “math experiment” variable right to the Test Variable(s) field. This field is for the dependent variable. Move the “math groups” variable to the Grouping Variable field. This field represents the independent variable and group membership.

The second step is to inform PSPP how the groups are represented. Click on the Define Groups button. Enter the coding scheme values, which are 0 and 1 for our example. Note that

PSPP has a feature for taking a continuously scaled variable and form two groups from a cut-off point if desired. This is represented by the “cut-off point” option.

Press the Continue and OK buttons to perform the analysis. The output will look like this:

T-TEST

T-TEST/VARIABLES = IndSample

/GROUP = IndSampleGroups(0.1) / MISSING=ANALYSIS

/GRITERIA = CI(0.95)

Group Statistics

Math Groups		N	Mean	SD	S.E. Mean
Math experiment	Control	5	101.40	2.88	1.29
	Experimental	5	125.60	5.13	2.29

Independent Samples Test

		Levene’s Test for Equality of Variances		t	d.f.	Sig. (2-tailed)
		F	Sig.			
Math experiment	Equal variances assumed	1.48	.259	−9.20	8.00	.000
	Equal variances not assumed			−9.20	6.30	.000

The upper table shows some useful group statistics like means and standard deviations. The lower table has the t-test results. This gets a little tricky because there are two versions of the test – equal variances assumed and equal variances not assumed (see the rows). For most situations, we will use the equal variances assumed version.

How can you tell which version to use? The Levene's test part of the output will give you the answer. When this test is non-significant, use the equal variances assumed version. The Levene's test p -value for this example is $p = .259$.

Our t-test result is $t(8) = 9.20, p < .001$ when expressed in APA style. We will reject the null hypothesis because this p -value is less than .05. This is a statistically significant difference. The students who get the new math teaching approach are doing better than the students who received the older teaching approach.

Note that the p -value (see Sig. 2-tailed) that PSPP gives is .000. Having a zero p -value is not really possible, but PSPP cannot make a judgement about something impossible. It would be best to report this result as $p < .001$ rather than $p = .000$.

A common mistake is to report the results from the Levene's test as being the t-test results. The Levene's test just helps us decide whether or not we need the equal variances assumed version of the independent samples t-test. The Levene's test is not the actual t-test.

5.7.6 Conducting Independent Samples t-test in R Commander

The two-sample student's t-test is used to determine if two population means are equal.

1. Statistics → Means → Independent Samples t-test.
2. Select the grouping variable, *e.g.*, genotype and the response variable, *e.g.*, Bone Area.
3. Under the Options tab, there are three things to consider:
 - (i) **Alternative Hypothesis:** Typically, you select a two-sided hypothesis; this means the change in mean can be either an increase or a decrease.
 - (ii) **Confidence Level:** The threshold at which you reject the null hypothesis and accept the alternate hypothesis. Typically, this is 0.05, which is a 5% risk that the difference is a sampling effect rather than a true population difference.
 - (iii) **Assume Equal variances:** Yes or No. As variance is a measure of the spread of the data, if the spread of the data is different between the two groups then for the statistical test to work reliable you need to select No. This can be assessed visually by looking at boxplots of the data and statistically. If you do not assume equal variance, this test is equivalent to the Welch t-test and is considered more robust. Small departures from equal variance significantly affect the robustness of results.
4. OK.
5. Interpretation: If the *p*-value is below the significance threshold, then there is a significant difference in the mean scores for each of the two groups.

5.8 ANOVA (ANALYSIS OF VARIANCE)

5.8.1 Introduction

The Analysis of Variance is similar to the Regression Analysis. It is abbreviated as ANOVA, developed by R.A. Fisher.

ANOVA

ANOVA is widely used by the researchers to investigate the underlying relationship between a dependent variable and various independent variables.

It is widely used in the areas of business, industries, economics, psychology, sociology, biology, etc. The researchers prefer the ANOVA when there are multiple cases of sample. It is a statistical tool which is used to analyze the variations among the data and calculate the inequality among the population means.

ANOVA also checks the level of homogeneity among the various groups of samples from different populations. The use of ANOVA by the researcher helps them to divide the total variation of the various variations of the various samples in two groups:

1. The first variation is that which occurs due to chance, and
2. The second variation is that which occurs due to any specific cause.

ANOVA helps the researcher to split the variance for investigative purpose. With the use of ANOVA, the researchers can examine '*n*' number of factors influencing the dependent variables.

In simpler words, analysis of variance (ANOVA) is a statistical technique which is used for exploring the variation of a dependent variable with respect to independent variables. In the ANOVA technique, when single factor is taken and researcher analyzes the difference amongst its various categories which have numerous possible values, then this ANOVA technique is known as one-way ANOVA. When analysis is done on two factors at the same time, then this ANOVA is known as two-way ANOVA.

5.8.2 Characteristics of ANOVA

1. ANOVA research tool is helpful in the study of analysis of variance.
2. One of the most important characteristics of ANOVA is that it splits the variance into chance variance and cause variance.
3. ANOVA is used to the examination of the various comparative experiments.
4. ANOVA results are free from any type of constant bias and scaling errors.
5. It describes the true and complete characteristics of given populations.
6. It examines the variance of more than two groups of samples.

5.8.3 Assumptions of ANOVA

Analysis of variance is done on the basis of some assumptions. These assumptions are given below:

1. Normally distributed universe used to take the sample for the study.
2. Samples are selected randomly and they are also homogeneous and independent of each other.
3. It is assumed that there is no significant difference amongst the variances of the different universes from which the samples have been drawn.
4. Starts with null hypothesis, i.e., $V_1 = V_2 = V_3 = \dots = V_n$ (where, $V_1, V_2, \dots, V_n$ (are variance)).
5. It is assumed that the critical values of the variance ratio (F), as estimated by Snedecor, for different degrees of freedom at different levels of significance, viz., 5%, 1%, etc. hold good in testing the significance of the calculated values of F that represents the variance ratio of the samples.

5.8.4 Applications of ANOVA

1. With the help of the ANOVA, an organic scientist or researcher can elaborate whether different varieties of fertilizers, seeds or soils differs significantly or not. This helps the researchers to draft the policy accordingly to the outcomes.
2. If the medical researchers or scientists are interested to study the impact of the medicine on different age groups of the human beings, the ANOVA is very useful.

3. The top management of an organization may examine the performance of their sales managers working on different population. Using ANOVA, management can come to a decision whether the performance or the efficiency of the sales managers differ significantly or not.
4. The statistical tool ANOVA can help the researcher to study whether the purchase intentions of group of people differs significantly or not.

5.8.5 Basic Principle of ANOVA

The underlying principle of the ANOVA is to compare the differences in the different means of population, by studying amount of the variations within the samples, with respect to amount of the variation present between samples of the population. In case of the variations which occur inside given population, generally, it is presumed by the researcher that this is because of the random effect in the population. In simpler words, because of the random effect within the population, the value of X_{ij} (a random value) differs with mean of that population.

It has been observed that there exist some influences on the (X_{ij}) , which cannot be explainable. On the other hand, the differences which occur between different populations are due to some 'specific factor'. This is the reason why the researchers assume using the ANOVA technique; the samples have been selected from the normal population. Also, each population has same degree of the variance. It is assumed by the researcher that there are few variables which cannot be controlled, but the rest of variables are easily controllable. So, the researcher assumes that absence of various factors may affect the results regarding the factor(s) to be studied. While estimating the ANOVA, the researcher needs to estimate the two population variances:

1. Based on within sample variance, and
2. Between the sample variance.

The value of the above two estimates of the population are compared with the F-test. The F-test is calculated as follows:

$$F = \frac{\text{Estimation of the population variance based on the sample variance}}{\text{Estimation of the population variance based on within sample variance}}$$

The above calculated value of the F-test is then compared with the limit of the F-test for a given degree of freedom. If the calculated value of the above equation is more than or equal to F-limit, then there exists a significant change between the sample means.

5.8.6 ANOVA Techniques

Following are the two ANOVA techniques:

1. Variance which occurs due to one variable, and
2. Variance which occurs due to two variables.

The above stated techniques of the variances are discussed in detail which is as follows:

1. Variance Which Occurs Due to One Variable: In this case, the change or the variance that occurs in a variable is because of only a single variable. Or we can say that there is only one independent variable and we try to study the impact of that independent variable on the other dependent variable.

For example, if the researcher wants to study the impact of one variable (*e.g.*, any one amongst types of seed or fertilizer or soil) on the other variable (*i.e.*, amount of harvest of a crop), then this is known as one-way ANOVA or variance due to the one variable. Here, type of soil, fertilizer and seeds are three independent variables. In this case, the researcher will study the impact of one independent variable (from seed, fertilizer or soil) on the dependent variable (harvesting of crop). This type of relationship can be given as:

A (Independent Variable)	B (Independent Variable)
11	14
12	16
17	11
16	10
14	15

2. Variance Which Occurs Due to Two Variables: In this case, the researcher wants to study the impact of multiple independent changes or the variation in two or more independent variables is studied in different condition and also the values corresponding dependent variable is also observed. For example, a researcher wants to study the impact of two independent variables, say season and effort of the salesman, on the total sale of corresponding salesman. This type of situation is best example of the variation due to the impact of two variables. The above example can be best described with the help of below stated format.

Samples of Sales in 000

Seasons	Salesmen				Total
	P	Q	R	S	
Summer	29	28	26	25	108
Monsoon	34	33	34	19	120
Winter	25	28	27	28	108
Total	88	89	87	72	336

On the basis of the above classification, there are two ANOVA techniques:

1. Analysis of the variance for one-way classification (One-way ANOVA)
2. Analysis of the variance for two-way classification (Two-way ANOVA).

5.8.7 One-way ANOVA

In the concept of the one-way ANOVA, there is an analysis of variance of a continuous dependent variable with a single classification variable (Independent Variable). One-way analysis

of variance (ANOVA) always tests equality of the population means when there is classification only by one variable. After calculations, we check whether there exists any difference within the factor or not.

5.8.7.1 Steps to Analyze the Output of the One-way ANOVA

1. Calculate the means of all samples individually, then find:

$$\bar{X}_1, \bar{X}_2, \bar{X}_3, \dots, \bar{X}_k$$

Here, K = sample observation

2. Calculate the mean of the sample:

$$\bar{X} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \dots + \bar{X}_k}{\text{No. of Samples } (k)}$$

3. Then calculate the deviation of the 'sample means' from the 'means of the sample means'. Then each deviation is squared and multiplied by the total number of the items in corresponding sample.

Each item is then added up and the output so obtained is called as sum of squares (SS).

$$SSB = SS \text{ between} = n_1(\bar{X}_1 - \bar{X})^2 + n_2(\bar{X}_2 - \bar{X})^2 + \dots + n_k(\bar{X}_k - \bar{X})^2$$

4. Then we need to divide the output that is obtained from third step by degrees of the freedom. Then, we get the means square between samples (MS). It is calculated as follows:

$$MSB = MS \text{ between} = \frac{SSB}{(k-1)}$$

Here, the MSB is the mean square between the samples and $(k-1)$ denotes degree of freedom (d.f.).

5. The researcher gets the value of the deviations for all of the sample items from their means. Then calculate the square of all the deviations and sum it up. The value so obtained after addition is called as sum of the squares of the variance within sample, abbreviated as (SS within). It can be written as:

$$SSW = SS \text{ within} = \sum (X_{1i} - \bar{X}_1)^2 + \sum (X_{2i} - \bar{X}_2)^2 + \dots + \sum (X_{ki} - \bar{X}_k)^2$$

6. Then divide the 'output of fifth step' by 'degree of the freedom' within the samples, in order to get the 'mean square' also abbreviated as MS (mean square within sample). It is written as:

$$MSW = MS \text{ within} = \frac{SSW}{(n-k)}$$

where, $(n-k)$ = degrees of freedom within samples

N = total number of items in all the samples, i.e., $n_1 + n_2 + \dots + n_k$

k = number of samples.

7. To cross check, the researcher can calculate the sum of the squares of the deviation for the total variance by adding up the squares of all the deviations. This can be done when the deviations of the each item from each individual sample is determined using 'mean of sample means'. This is represented as:

$$SST = SS \text{ for total variance} = \sum (X_{ij} - \bar{X})^2; i = 1, 2, 3, \dots; j = 1, 2, 3, \dots$$

This value of SST is equal to the total of the results of the step three and five. In each case, the degree of the freedom is calculated by subtracting 1 from the total number of items in each sample, i.e., $(n - 1)$. The degree of the freedom in the case of within and between samples should add to degree of the freedom for the complete variance. This can be represented as:

$$(n - 1) = (k - 1) + (n - k)$$

This fact describes the additive property of the ANOVA.

8. Finally, the F-ratio can be represented as:

$$F\text{-ratio} = \frac{MSB}{MSW}$$

The researcher uses this ratio to know if the differences among various samples mean is just because of the sampling fluctuation or it has some significance and cannot be ignored. The researchers consult the table and put the F -value for given 'degrees of the freedom' at different significance levels. If the calculated value of F is less than table value of F , then the differences are insignificant and can be considered that it is occurring just by chance and the researcher accepts the Null Hypothesis. If the calculated value of F is more than or equal to the table value of F , then the results are significant and the researcher accepts the alternate hypothesis. If calculated F is greater than the table value of F , then the researcher becomes sure for about his conclusions and outcomes.

Example 14: Four machines A , B , C and D are used to produce a certain kind of cotton fabrics. Samples of size 4 with each unit as 100 square meters are selected from the outputs of the machines at random, and the number of flaws in each 100 square meters are counted, with the following result:

A	B	C	D
11	7	12	16
9	9	11	20
14	10	18	23
16	4	9	21

Do you think that there is a significant difference in the performance of the four machines?

Solution: Let us take the null hypothesis that the machines do not differ significantly in performance, i.e., $H_0: u_1 = u_2 = u_3 = u_4$.

X_1	X_1^2	X_2	X_2^2	X_3	X_3^2	X_4	X_4^2
11	121	7	49	12	144	16	256
9	81	9	81	11	121	20	400
14	196	10	100	18	324	23	529
16	256	4	16	9	81	21	441
$\Sigma X_1 = 50$	$\Sigma X_1^2 = 654$	$\Sigma X_2 = 28$	$\Sigma X_2^2 = 246$	$\Sigma X_3 = 50$	$\Sigma X_3^2 = 670$	$\Sigma X_4 = 80$	$\Sigma X_4^2 = 1626$

$$T = \Sigma X_1 + \Sigma X_2 + \Sigma X_3 + \Sigma X_4 = 50 + 28 + 50 + 80 = 208$$

$$\text{Correction factor} = \frac{T^2}{N} = \frac{(2008)^2}{16} = \frac{44521}{16} = 2704$$

$$SST = \Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 + \Sigma X_4^2 - \frac{T^2}{N} = 654 + 246 + 670 + 1626 - 2704 = 3196 - 2704 = 492$$

$$SSB = \frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3} + \frac{(\Sigma X_4)^2}{n_4} - \frac{T^2}{N}$$

$$= \frac{(50)^2}{4} + \frac{(28)^2}{4} + \frac{(50)^2}{4} + \frac{(80)^2}{4} - 2704 = 625 + 196 + 625 + 1600 - 2704 = 342$$

$$SSW = SST - SSB = 492 - 342 = 150$$

ANOVA Table

Source of Variation	Sum of Squares	Degree of Freedom	Mean Square
Between samples	342	3	114
Within samples	150	12	12.5
Total	492	15	

$$F_{(3, 12)} = \frac{114}{12.5} = 9.12$$

The table value for $F_{(3, 12)}$ at 1% level of significance is 5.95. The calculated value of F is greater than the table value. Hence, we reject the null hypothesis and conclude that there is a significant difference in the performance of the four machines.

Example 15: A common test was performed on a number of patients taken from the four hospitals randomly to measure the significance of possible difference in performance of an antibiotic after its administration to TB patients. The data are given table below. Give an analysis of variance of data.

Hospital			
A	B	C	D
8	12	18	13
10	11	12	9
12	9	16	12
8	14	6	16
7	4	8	15

Solution: Null Hypothesis, H_0 : There is no significant difference in the average effects of antibiotics on TB patients.

Alternate Hypothesis, H_1 : There is significant difference in the average effects of antibiotics on TB patients.

A	B	C	D
8	12	18	13
10	11	12	9
12	9	16	12
8	14	6	16
7	4	8	15
$T_1 = 45$	$T_2 = 50$	$T_3 = 60$	$T_4 = 65$

$$\text{Grand Total (G)} = T_1 + T_2 + T_3 + T_4 = 45 + 50 + 60 + 65 = 220$$

$$\text{Correction Factor (C.F.)} = \frac{G^2}{n} = \frac{(220)^2}{20} = 2420$$

$$\begin{aligned} \text{Total Addition of Squares (TSS)} &= \sum_{i=1}^k \sum_{j=1}^m x_{ij}^2 - C.F. \\ &= (8)^2 + (10)^2 + (12)^2 + \dots + (16)^2 + (15)^2 - C.F. \\ &= 64 + 100 + 144 + \dots + 256 + 225 - 2420 \\ &= 2678 - 2420 = 258 \end{aligned}$$

Sum of Squares among the Samples (SST)

$$= \sum_{i=1}^3 \frac{T_i^2}{n_i} - C.F. = \frac{(45)^2}{5} + \frac{(50)^2}{5} + \frac{(60)^2}{5} + \frac{(65)^2}{5} - 2420 = 2470 - 2420 = 50$$

Sum of squares within the sample (Error)

$$SSE = TSS - SST = 258 - 50 = 208$$

The ANOVA table is shown below:

ANOVA Table

Source of Variation	d.f.	Sum of Squares	Mean Sum of Squares	Variance Ratio F
Between Samples	$4 - 1 = 3$	50	$\frac{50}{3} = 16.7$	$F = \frac{16.7}{13.0} = 1.28$
Within Samples	$20 - 4 = 16$	208	$\frac{208}{16} = 13.0$	
Total	$20 - 1 = 19$			

Table value of $F_{0.05}$ for (3, 16) degree of freedom (d.f.) is 3.24. Hence, the calculated value of F is less than the tabulated value of F , null hypothesis is accepted at 5% level of significance. Thus, the average effect of antibiotics on TB patients in various hospitals of the city is not significant.

5.8.7.2 Conducting Independent One-way ANOVA in PSPP

The t-test is limited to comparing two groups or conditions. A different kind of statistical test must be used if there are three or more groups that need to be compared.

Analysis of Variance, or ANOVA, was invented by Sir Ronald Fisher to simultaneously compare the results from several groups. The most basic version of ANOVA compares groups that vary in a single dimension, or factor. This is called one-way ANOVA.

A more advanced version of ANOVA is available for experiments that have two or more factors. In PSPP, a one-way ANOVA must have a dependent variable for the data. A second variable is needed to represent the group membership. This approach is similar to the setup for an independent-samples t-test.

In this example, let us pretend that a drug is being tested to treat high blood pressure. Hypertensive individuals are randomly assigned to groups that get no drug (control), a low dose, medium dose, or high dose. Part of the example data are shown below, with some data omitted to save space. The BP variable represents systolic blood pressure. The Group variable has group membership represented by 0, 1, 2, or 3 for a zero low, medium, or high dose, respectively. The data file is available for downloading.

<i>Case</i>	<i>BP</i>	<i>Group</i>
1	155	0
2	166	0
3	176	0
4	157	0
5	163	0
6	172	0
7	164	0
8	159	0
9	163	0
10	182	0
11	152	1
12	155	1
13	162	1
14	143	1

The one-way ANOVA test is started by selecting Analyze, Compare Means and One-way ANOVA.

PSPP also provides check boxes for “descriptive” and “homogeneity”. This is useful information you have. So, let us check these boxes. The output begins with some basic descriptive statistics, such as the mean and standard deviation for each group.

Descriptives				
		<i>N</i>	<i>Mean</i>	<i>Std. Deviation</i>
Blood pressure	Control	10	165.40	8.24
	Low dose	10	149.90	5.82
	Medium dose	10	134.10	8.45
	High dose	10	123.10	5.63
	Total	40	143.12	17.61

The next part of the output shows the homogeneity of variance test. ANOVA assumes that the groups have variances that are similar. If the results of this test are significant, it would suggest that the group variances are significantly different. So, ANOVA might not be an appropriate test. Fortunately, the homogeneity test is not significant for this current example. It will be okay to use an ANOVA test.

Test of Homogeneity of Variances

	<i>Levene Statistic</i>	<i>d.f.1</i>	<i>d.f.2</i>	<i>Sig.</i>
Blood Pressure	1.41	3	36	.257

The final part shows the ANOVA results.

ANOVA

ANOVA						
		<i>Sum of Squares</i>	<i>d.f.</i>	<i>Mean Square</i>	<i>F</i>	<i>Sig.</i>
Blood Pressure	Between Groups	10245.28	3	3415.09	66.70	.000
	Within Groups	1843.10	36	51.20		
	Total	12088.38	39			

The *F*-score statistic and the *p*-value are shown on the right side. This result would be statistically significant because $p < .05$. The sums of squares and means squares results are sometimes used for further analyses, such as calculating effect sizes. The format for reporting ANOVA results in APA style is $F(\text{degrees of freedom} - \text{between}, \text{degrees of freedom} - \text{within}) = F\text{-score}, p = p \text{ value}$. For this example, the results would be reported as $F_{(3, 36)} = 66.7, p < .001$.

5.8.7.3 Conducting Independent One-way ANOVA in R Commander

The test is used when you wish to compare the mean scores of more than two groups. Analysis of variance is so-called because it compares the variance (variability in scores) between the different groups (believed to be due to the grouping variable) with the variability within each of the groups (believed to be due to chance). The ratio of the variance is converted to a *p*-value which assesses the chance that this difference in variance arises from sampling affects.

A significant p -value indicates that we can reject the null hypothesis which states that the population's means are equal. It does not, however, tell us which of the groups are different. If a significant score is obtained in the one-way ANOVA, then post-hoc testing is used to tell where the difference arose. The software uses Tukey Post-hoc Comparison Procedure which is essential like a student's t -test. However, the test takes into account the risk of accumulating false positives as multiple tests are being conducted.

The steps are as follows:

1. Statistics $\rightarrow$ Means $\rightarrow$ One-way ANOVA
2. Enter a name for the model.
3. Select a response variable as the variable of interest.
4. Select the grouping variable.
5. OK.
6. Interpretation: If the p -value is below the significance threshold, then somewhere, there is a statistically differences in the means of two or more groups.
7. If the p -value is significant, repeat the analysis with the pair-wise comparisons of means ticked. This repeats the analysis with the groups being compared to each other group using Tukey contrasts.
8. Interpretation:
 - (i) For each possible pair-wise comparison, the software calculates the mean difference for each comparison and tests it with the null hypothesis that the difference should be zero and returns a p -value on this statistical test.
 - (ii) It also calculates the 95% confidence interval on the observed mean differences for each pair-wise comparison.
 - (iii) Graphically, it visualizes these calculated values. You are looking for comparisons where the mean difference confidence interval does not span zero indicating a statistically significant difference between these groups.

5.8.8 Two-way ANOVA Advanced Concepts

The two-way Analysis of Variance (ANOVA) technique is widely used by the researcher when data is categorized into two factors. The two-way ANOVA can be better explained by the following example:

1. A business manager may be interested in analysing the sales data of their sales manager on the basis of the efforts done by them or the area in which they are working.
2. An agricultural researcher or scientist may analyze the rice yield on the basis of the type of seed or fertilizer which is used by the farmer.
3. In a manufacturing unit, the project manager may analyze the output of a machine on the basis of the experience of the worker and the quality of the raw material.

In all the cases, the researcher takes the observation at regular interval and makes a note of it. Sometimes, the researcher needs to take some repeated values also.

Types of Two-way ANOVA Techniques

The ANOVA technique is little different in case of repeated measurement where we also compute the interaction variation. Two types of ANOVA techniques are:

1. ANOVA technique in the context of two-way designs when repeated values are not there: In the case when repeated values are missing, the researcher cannot compute sum of the squares directly within the sample. So, the residual or the error variation is calculated by subtracting the sum of variance between varieties of a treatment and variance between varieties of other treatment from the total sum of squares for total variance.

Steps:

- (i) The researcher needs to use a coding device, to make the work simpler.
- (ii) After this, the researcher has to take total of all the individual values (or the coded values) of the entire sample, and name it as T .
- (iii) Then apply correction factor stated below:

$$\text{correction factor} = (T)^2/n$$

- (iv) Calculate the square value of all the items (or square of the coded values) taken individually and then sum it up. Then subtract the correction factor from the summation to get the square of deviations for the total variance:

$$\text{Sum of Squares of Deviation for Total Variance or Total SS} = \sum X_{ij}^2 - \frac{(T)^2}{n}$$

- (v) After the above step, calculate the total of all the different columns, and then calculate the square of sum of every column. Then divide the square value of the each column by the corresponding number of items in that column and again get the sum of that particular column. Now, subtract the above calculated correction factor from this total. So, the output received is the sum of the squares of deviation of the variance between the columns (SS between columns).
- (vi) In this step, the researcher calculates the total rows and then finds out the square of sum of every row. This squared sum is then divided by the number of items of that specific row to obtain the sum of that entire row. Now, correction factor is subtracted from the calculated sum and the researcher gets sum of the square of the deviation for variance between the columns (SS between columns).
- (vii) Here, by subtracting the sum of 5th and 6th step from 4th step, the researcher can calculate the sum of the squares of the deviation for the error or residual variance.

$$\text{Total SS} - (\text{SS between columns} + \text{SS between rows}) = \text{SS for residual or error variance}$$

Degrees of freedom (d.f.) can be worked out as under:

d.f. for Total Variance = $(c.r - 1)$

d.f. for Variance between Columns = $(c - 1)$

d.f. for variance between Rows = $(r - 1)$

d.f. for Residual Variance = $(c - 1)(r - 1)$

where, c = Number of columns

r = Number of rows

(viii) ANOVA table can be set up as:

Analysis of Variance Table for Two-way ANOVA

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (d.f.)	Mean Square Value (MS)	F-ratio Value
Between the columns treatment	$\sum \frac{(T_j)^2}{n_j} - \frac{(T)^2}{n}$	$(c - 1)$	$\frac{SSB(columns)}{(c - 1)}$	$\frac{MSB(columns)}{MSR}$
Between the rows treatment	$\sum \frac{(T_i)^2}{n_i} - \frac{(T)^2}{n}$	$(r - 1)$	$\frac{SSB(rows)}{(r - 1)}$	$\frac{MSB(rows)}{MSR}$
Residual or error	Total SS – (SS between the columns + SS between the rows)	$(c - 1)(r - 1)$	$\frac{SSR}{(c - 1)(r - 1)}$	
Total	$\sum X_{ij}^2 - \frac{(T)^2}{n}$	$(c.r - 1)$		

where, c = Number of columns

r = Number of rows

SSR (SS residual) = Total SS – (SS between the columns + SS between the rows).

The basis for F-ratio is provided by the Mean Square Residual (MSR) or residual variance. It is concerned with the variations that are between the column and the row calculation. The fluctuations in the sampling always cause the Mean Square Residual. The two F-ratios are then compared with the related table values for a given degree of the freedom at specified significance level. If the calculated value for the F-ratio is greater than or equal to the table value, in that case, difference of column means is treated significant. In the same way, the F-ratio related to the variations between the rows can be calculated.

ANOVA technique in the context of two-way design when repeated values are there:

The researcher gets the separate value of the individual smallest change or variation, in the case of the two-way ANOVA, when there are some repeated values. The degree of freedom and sum of the squares are calculated for the variance within the sample same as in case of one-way ANOVA. The total sum of the square (SS), which are between the columns and the rows, can be

calculated as stated above. Later, the researcher finds out the 'left over sums of the square' and also the 'left over degree of freedom' which is known as the 'interaction variation'. (Interaction variation is the value of the relationship present between two different classifications). After all the calculations have been done, ANOVA table can be set up for observing various inferences.

Example 16: Five doctors each test 5 treatments for a certain disease and observe the number of days each patient takes to recover. The results are given in the table below. Discuss the difference between —

1. The doctors, and
2. The treatment.

Use $\alpha = 0.05$

Doctors	Treatments				
	1	2	3	4	5
1	10	14	23	18	20
2	11	15	24	17	21
3	9	12	20	16	19
4	8	13	17	17	20
5	12	15	19	15	22

Solution: We subtracted 15 from the given values and work out with the new values of x_{ij} .

Doctors	Treatments					
	1	2	3	4	5	Total
1	-5	-1	8	3	5	10
2	-4	0	9	2	6	13
3	-6	-3	5	1	4	1
4	-7	-2	2	2	5	0
5	-3	0	4	0	7	8
Total	-25	-6	28	8	27	32

$$\text{Correction factor} = \frac{T^2}{N} = \frac{(32)^2}{25} = 40.96$$

(where number of observations or N is 25)

Sum of Squares between Columns (Treatments)

This will be obtained by squaring the treatments totals, dividing each total by the number of observations included in it, adding these figures and then subtracting the correction factor from them. Thus, sum of squares between salesman is:

$$= \frac{(-25)^2}{5} + \frac{(-6)^2}{5} + \frac{(28)^2}{5} + \frac{(8)^2}{5} + \frac{(27)^2}{5} - \frac{T^2}{N}$$

$$= 125 + 7.2 + 156.8 + 12.8 + 145.8 - 145.8 - 40.96 = 406.64$$

$$\text{Degrees of Freedom } (v) = (5 - 1) = 4$$

Sum of Squares between Rows (Doctors)

This will be obtained by dividing the squares of the doctors totals by the number of observations that make up each total, adding all such figures and subtracting these from the correction factor. Thus, sum of squares between doctors is:

$$= \frac{(10)^2}{5} + \frac{(13)^2}{5} + \frac{(1)^2}{5} + \frac{(0)^2}{5} + \frac{(8)^2}{5} - \frac{T^2}{N} = 66.8 - 40.96 = 25.84$$

$$\text{Degrees of Freedom } (v) = (5 - 1) = 4$$

Total Sum of Squares

This will be obtained by adding the squares of all the observations in the table and subtracting the correction factor therefrom. Thus, total sum of squares is:

$$\begin{aligned} &= (-5)^2 + (-4)^2 + (-6)^2 + (-7)^2 + (-3)^2 + (-1)^2 + (0)^2 + (-3)^2 + (-2)^2 + (0)^2 + (8)^2 + (9)^2 + \\ &\quad (5)^2 + (2)^2 + (4)^2 + (3)^2 + (2)^2 + (1)^2 + (2)^2 + (0)^2 + (5)^2 + (6)^2 + (4)^2 + (5)^2 + (7)^2 - \frac{T^2}{N} \\ &= 25 + 16 + 36 + 49 + 9 + 1 + 0 + 9 + 4 + 0 + 64 + 81 + 25 + 4 + 16 + 9 + 4 + 1 + 4 + 0 \\ &\quad + 25 + 36 + 16 + 25 + 49 - 40.96 \\ &= 508 - 40.96 \\ &= 467.04 \end{aligned}$$

The above information is presented in the following table:

ANOVA Table

Source of Variation	Sum of Squares	d.f.	Means Square	F Ratio
Between Columns (Treatments)	406.64	4	101.66	$\frac{101.66}{2.16} = 47.06$
Between Rows (Doctors)	25.84	4	6.46	$\frac{6.46}{2.16} = 2.99$
Residual	34.56	16	2.16	
Total	467.04	24		

The table value of F for 4 and 16 degrees of freedom at 5% level of significance is 3.01.

1. Table value of F is greater than the calculated value for doctors. Hence, there is no difference between doctors.
2. Table value of F is less than the calculated value for doctors. Hence there is highly significant difference between treatments.

Example 17: There are four different drugs which have been developed for the cure of a certain disease. These drugs are experimented on patients of three different hospitals. The

numbers of cases of recovery from the disease per 100 people are given in the table below. Analyze the variance and interpret results.

Hospital	Drugs			
	A	B	C	D
H_1	24	20	24	17
H_2	20	25	30	9
H_3	13	18	31	13

Solution: Null Hypothesis, H_0 : There is no result on the drugs in the cure of disease of patients.

Alternate Hypothesis, H_1 : There is a result on the drugs in the cure of disease of patients.

Hospital	Drugs				
	A	B	C	D	Total
H_1	24	20	24	17	85
H_2	20	25	30	9	84
H_3	13	18	31	13	75
Total	57	63	85	39	244

Grand total (G) = 244

$$\text{Correction factor (CF)} = \frac{G^2}{N} = \frac{(244)^2}{12} = \frac{59536}{12} = 4961.33$$

$$\begin{aligned} \text{Total Sum of Squares (TSS)} &= \sum \sum y_{ij}^2 - C.F. \\ &= (24)^2 + (20)^2 + \dots + (13)^2 - C.F. \\ &= 5470 - 4961.33 \\ &= 508.67 \end{aligned}$$

$$\begin{aligned} \text{Sum of Squares due to Rows (SSR)} &= \frac{(85)^2}{4} + \frac{(84)^2}{4} + \frac{(75)^2}{4} - C.F. \\ &= 1806.25 + 1764 + 1406.25 - 4961.38 \\ &= 4976.50 - 4961.33 \\ &= 15.17 \end{aligned}$$

$$\begin{aligned} \text{Sum of Squares due to Columns (SSC)} &= \frac{(57)^2}{3} + \frac{(63)^2}{3} + \frac{(85)^2}{3} + \frac{(39)^2}{3} - C.F. \\ &= 1083 + 1323 + 2408.33 + 507 - 4961.33 \\ &= 5321.33 - 4961.33 \\ &= 360 \end{aligned}$$

$$\begin{aligned}
 \text{Sum of Squares due to Error (SSE)} &= \text{TSS} - \text{SSR} - \text{SSC} \\
 &= 508.67 - 15.17 - 360 \\
 &= 508.67 - 375.17 \\
 &= 133.5
 \end{aligned}$$

Analysis of Variance Table (ANOVA)

Source of Variation	d.f.	Sum of Squares (SS)	Mean Sum of Squares (MSS)	Variance Ratio F
Between Hospitals (Rows)	$3 - 1 = 2$	15.17	$\frac{15.17}{2} = 7.58$	$F_{(2,6)} = \frac{7.58}{22.25} = 0.34$
Between Drugs (Columns)	$4 - 1 = 3$	360	$\frac{360}{3} = 120$	$F_{(3,6)} = \frac{120}{22.25} = 5.39$
Error	$(3 - 1)(4 - 1) = 6$	133.5	$\frac{133.5}{6} = 22.25$	
Total	$12 - 1 = 11$	508.67		

- 1. Tabulated Values:** At 5% level of significance, $F_{(2,6)} = 5.14$ and $F_{(3,6)} = 4.76$.
- 2. F for Between Drugs:** The calculated value of $F >$ tabulated value of F . Thus, null hypothesis is rejected. This shows that the drugs have significant means.
- 3. F for Between Hospitals:** The calculated value of $F <$ tabulated value of F . Hence, the null hypothesis is accepted. In other words, the hospitals have no significant means.

QUESTIONS

Very Short Answer Type Questions:

1. What is hypothesis?
2. Give the purpose of hypothesis testing.
3. State and define standard error of mean.
4. Write the characteristics of ANOVA.
5. Give an application of ANOVA.
6. What is one-sample t-test?
7. What is independent sample t-test?

Short Answer Type Questions:

1. Discuss the qualities of a good hypothesis.
2. Give a definition of null hypothesis and alternative hypothesis.
3. What are the limitations of hypothesis testing?
4. Explain the hypothesis testing of one mean.

5. Define formulation of hypothesis.
6. Explain the hypothesis testing of two means.
7. What are the differences between parametric and non-parametric tests?
8. Write the properties of t-distribution.
9. Explain the paired t-test.
10. What is ANOVA (analysis of variance)? Write the assumptions of ANOVA.
11. Discuss how to conduct independent sample t-test in PSPP and R Commander?
12. Discuss how to conduct independent one-way ANOVA in PSPP and R Commander?

Long Answer Type Questions:

1. What do you mean by hypothesis testing? And explain the various types of hypothesis.
2. Give brief explanations on errors in hypothesis testing and level of significance.
3. Explain the following terms:
 - (a) Degree of freedom
 - (b) One-tailed and two-tailed tests
4. Discuss how to interpret the output of one-sample t-test?
5. Discuss how to conduct paired sample t-test in PSPP and R Commander?
6. Discuss the two-way ANOVA in detail.
7. Two horses, A and B, were tested according to time (in seconds) to run a particular race with the following results:

Horse A	28	30	32	33	33	29	34
Horse B	29	30	30	24	27	29	

Test whether you can discriminate between the two horses.

(Given $t_{0.05}(11) = 2.2$)

[Ans: Hypothesis is rejected at 5%.]

8. Two independent samples of 8 and 7 items respectively had the following values of the variable (weight in ounces).

Sample I	9	11	13	11	15	9	12	14
Sample II	10	12	10	14	9	8	10	

Is the difference between the means of the sample significant?

(Given $t_{0.05}(13) = 2.16$)

[Ans: No]

9. 10 persons were appointed in an electrical position in an office. Their performance was noted by giving a test and means recorded out of 50. They were given 6 months' training and again they were given a test and marks were recorded out of 50.

Employees	1	2	3	4	5	6	7	8	9	10
Before Training	25	20	35	15	42	28	26	44	35	48
After Training	26	20	34	13	43	40	29	41	36	46

By applying t-test, can it be concluded that the employees have benefited by training ($t_{0.05}(9) = 2.262$).

[Ans: No]

10. Samples of two types of electric light bulbs were tested for length of life and following data were obtained:

	Type I	Type II
Sample Number	$N_1 = 8$	$N_2 = 7$
Sample Means	$X_1 = 1234$ hours	$X_2 = 1036$ hours
Sample SDs	$S_1 = 36$ hours	$S_2 = 40$ hours

Is the difference in the means sufficient to warrant that *Type I* is superior to *Type II* regarding length of life?

[Ans: Hypothesis is rejected, $t_{0.05}(13) = 2.16$]

11. Two horses, *A* and *B* were tested according to time (in seconds) to run a particular race with the following results:

Horse A	28	30	32	33	33	29	34
Horse B	29	30	30	24	27	29	

Test whether you can discriminate between the two horses.

(Given $t_{0.05}(11) = 2.2$)

[Ans: Hypothesis is rejected at 5%]

12. Two independent samples of 8 and 7 items respectively had the following values of the variable (weight in ounces).

Sample I	9	11	13	11	15	9	12	14
Sample II	10	12	10	14	9	8	10	

Is the difference between the means of the sample significant?

(Given $t_{0.05}(13) = 2.16$)

[Ans: No]

13. There are three main brands of a certain powder. A set of its 120 sales is examined and found to be allocated among four groups (*A*, *B*, *C* and *D*) and brands (*I*, *II* and *III*) as shown below:

Brands	Groups			
	A	B	C	D
<i>I</i>	0	4	8	15
<i>II</i>	5	8	13	6
<i>III</i>	18	19	11	13

Is there any significant difference in brands preference? Answer at 5% level using one-way ANOVA.

[Ans: No significant difference in brands]

SUMMARY

- The one-sample t-test compares the mean of a single sample to a specified value to see if it is significantly more or less than it. The independent sample t-test compares the means of two unique groups.
- When we want to know if our sample originates from a specific population, but do not have complete population information, we use the one-sample t-test. For example, we might wish to know if a specific group of college students is representative of all college students or not.
- The null hypothesis states that there is no relationship between the two phenomena under investigation or that there is no association between two groups. An assertion that there is a relationship between two selected variables in a study is referred to as an alternate hypothesis.
- Begin by selecting the Scatterplot command from the Graphs menu. Fields for the X and Y variables must be selected in the Scatterplot dialogue box. Which variable should be X or Y is a question of opinion. Selecting the potential causal variable or initial variable for the X-axis is a good starting point.
- A regression model that uses a straight line to evaluate the connection between one independent variable and one dependent variable is known as simple linear regression. Both variables should have a numerical value.
- The principle of temporal precedence states that the cause (*i.e.*, independent variable) comes before the consequence (*i.e.*, outcome); proving that the cause and effect are connected and/or correlated; and demonstrating that there are no other credible explanations.
- ANOVA examines groups that differ in a single dimension, or factor, in its most basic form. This is referred to as one-way ANOVA. For studies with two or more components, a more advanced version of ANOVA is provided.
- A dependent variable for the data is required in a one-way ANOVA in PSPP. To express group membership, a second variable is required. This method is comparable to how an independent samples t-test is set up.

REFERENCES

- <https://www.statisticssolutions.com/free-resources/directory-of-statisticalanalyses/one-sample-t-test-2/>
- <http://www.psychology.emory.edu/clinical/bliwise/Tutorials/TOM/meanstests/tone.htm>

- <https://microbenotes.com/null-and-alternative-hypothesis/#:~:text=Definition-,The%20null%20hypothesis%20is%20a%20general%20statement%20that%20states%20that,selected%20variables%20in%20a%20study.>
- <https://www.garyfisk.com/pspp/35scatterplot.html>
- <https://www.scribbr.com/statistics/simple-linear-regression/#:~:text=What%20is%20simple%20linear%20regression,Both%20variables%20should%20be%20quantitative.>
- [https://www.sciencedirect.com/topics/psychology/temporal-precedence#:~:text=1.-,temporal%20precedence%2C%20which%20is%20establishing%20that%20the%20cause%20\(i.e.%2C,3.](https://www.sciencedirect.com/topics/psychology/temporal-precedence#:~:text=1.-,temporal%20precedence%2C%20which%20is%20establishing%20that%20the%20cause%20(i.e.%2C,3.)
- <https://www.garyfisk.com/pspp/43ANOVA.html#:~:text=The%20most%20basic%20version%20of,dependent%20variable%20for%20the%20data.>



MODULE 6

HR ANALYTICS APPLICATIONS USING REGRESSION

Learning Objectives

1. Analyze the Logistic Regression with Single Nominal Predictor Concept.
2. Learn the Logistic Regression Analysis in PSPP.
3. Understanding the steps conducting Logistic Regression using Rattle Package.
4. Study the Conducting Factor Analysis in PSPP and R.
5. Interpret the Output of Cluster Analysis and Advanced Concepts.

6.1 WHAT IS HR ANALYTICS?

HR Analytics does not collect data about how the employees are performing at work. Instead, its sole aim is to provide better insight into each of the human resource processes, gathering related data and then using this data to make informed decisions on how to improve these processes.

6.2 REGRESSION ANALYSIS

Regression analysis is a statistical tool used to calculate a continuous dependent variable from various independent variables and is commonly used for prediction and forecasting. For example, if agriculturist knows that the yield of rice and rainfall are closely related, then he will want to know the amount of rain required to achieve a certain production. In this situation, he will use the regression analysis.

This technique is used for modelling and analysis of mathematical data using the value of a dependent variable and one or more independent variables. Dependent variable, also known as response variable, is the single variable predicted by the regression model. Independent variable, also known as predictor variable, is used to predict the values of dependent variable.

Regression analysis is the process to calculate the unknown value of one variable from the known values of the other variable.

According to **Blair**, “Regression is the measure of the average relationship between two or more variable in terms of the original units of the data.”

According to **Taro Yamane**, “One of the most frequently used techniques in economics and business research, to find a relation between two or more variables that are related casually, is regression analysis.”

Regression analysis makes an equation to describe the statistical relationship between one or more predictor variables and the response variable.

It helps to know how to determine the nature and strength of a relationship between independent variables.

Regression may be classified as follows:

1. Simple regression
2. Multiple regression
3. Linear regression
4. Non-linear regression

6.2.1 Objectives of Regression Analysis

The first objective of regression analysis is to build a regression model (or predictive equation) that can be used to describe, predict and control the dependent variable on the basis of the independent variable.

The second objective of regression analysis is to explain variability in dependent variable by means of one or more of independent or control variables.

The third objective of regression analysis is to obtain a measure of the error involved in using the regression line as a basis for estimation. For this purpose, standard error of estimate is obtained. This helps in understanding the correlation existing between X and Y .

And the last one is that regression analysis will produce information about the nature and strength of influence of value determining factors, as well as to estimate the correct weights for adjustment, and to produce an estimate of the correct value.

6.2.2 Uses of Regression Analysis

Regression analysis is a branch of statistical theory that is widely used in almost all the scientific disciplines. In economics, it is the basic technique for measuring or estimating the relationship among economic variable that constitute the essence of economic theory and economic life.

The uses of regression are not limited to economics and business field only. Its applications are extended to almost all the natural, physical and social sciences. The uses of regression analysis are as follows:

1. Regression analysis studies the relationship of two or more variables.
2. It helps to find out the value of dependent variables from the value of independent variables.

3. It is useful for calculating coefficients of correlation (r) and coefficients of determination.
4. It is widely useful for quality control in corporate sector.
5. It is useful for estimation of statistical curve for demand, supply, price, consumption and cost.

6.3 CAUSATION

Causation, or causality, is the capacity of one variable to influence another. The first variable may bring the second into existence or may cause the incidence of the second variable to fluctuate.

Causation is often confused with correlation, which indicates the extent to which two variables tend to increase or decrease in parallel. However, correlation by itself does not imply causation. There may be a third factor that is responsible for the fluctuations in both variables.

It is the combination of descriptive and prescriptive analytics that we call causal analysis. Professor Nick Bontis, is one of the earlier and most effective practitioners of human capital causal analysis, sometimes called causal modelling. With causal modelling, we can find the hidden root cause of a problem or make a business proposition for a human capital investment. In Bontis learning impact model, he points out that the success of training is not a function only of the courseware and trainer delivery, but it is very much the result of the perceived value of the training by the trainee. One way to find and express training's business value is to survey the trainees, asking them to score their perceived value on a scale of 1% to 100%; the courseware, from 1 to 5 points; and the delivery, 1 to 5 points. Then to put these factors into an equation such as the one below:

Perceived value		Courseware		Delivery		Cumulative score
.80	×	4	×	5	=	16

(Maximum possible is 25: $1.0 \times 5 \times 5$)

As any of the rating change, the cumulative score changes as well. The various uses for this simple causal model are broad indeed, including how people are chosen for the training, how the perceived value of the training changes from before to after, how the training department's performance can be tracked, and on and on. When causal analysis is combined with predictive management, the power is truly remarkable. Predictive management is the model and mindset that was developed by the Predictive Initiative. Bontis was a key contributor to the initiative and has published several examples of his work on causal modelling over the past decade.

6.3.1 Need of Causation

The need of causation depends on what are we going to do with the data. For example, if we are going to just recommend a product based on the data, chances are that correlation is enough. However, if we are taking a life changing decision or make a major policy decision, we might need causality.

Let us investigate both type of cases.

Correlation is enough when stakes are low, or we can later verify our decision. Following are few examples.

1. When stakes are low (*e.g.*, marketing, recommendations) – when showing an advertisement or recommending a product to buy, one has more freedom to make an error.
2. As a starting point for an investigation, correlation is never enough to prove someone is guilty. However, it can show us useful places to start digging.
3. Sometimes, it is hard to know what things are connected, but easily verify the quality given a choice. For example, if you are trying to match candidates to a job or decide good dating pairs, correlation might be enough. In both these cases, given a pair, there are good way to verify the fit.

There are other cases where causality is crucial. Following are few examples.

1. Find a cause for disease.
2. Policy decisions (Would \$15 minimum wage be better? Would free health care be better?)
3. When stakes are too high (shutting down a company, passing a verdict in court, sending a book to each kind in the state, etc.).
4. When we are acting on the decision (firing an employee).

6.3.2 Difference between Causation and Correlation

<i>Causation</i>	<i>Correlation</i>
When an event causes the occurrence of another event, then a causal relationship between those two events can be said to exist.	A degree or measure of relationship between two variables is known as correlation. There can be positive correlation, negative correlation or no correlation at all.
One independent variable and two or more groups are involved in causal studies.	One group and two or more variables are involved in correlational studies.
The two variables must be correlated if a causal relationship exists between them.	Causation is not implied by just a correlation between two variables.
Causation is an area that is often misinterpreted and it can be extremely challenging to infer causation between two variable without undertaking a randomized controlled experience.	Correlation can be a helpful measure but has certain restrictions as it is generally related to measuring a linear relationship. However, causation is not implied by such correlation.

6.4 LOGISTIC REGRESSION

Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, True or False, etc., but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving regression problems, whereas logistic regression is used for solving the classification problems. In logistic regression, instead of fitting a regression line, we fit an 'S' shaped logistic function, which predicts two maximum values (0 or 1).

The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc. Logistic regression is a significant.

That means Logistic regression is usually used for Binary classification problems.

Binary Classification refers to predicting the output variable that is discrete in **two** classes.

A few examples of binary classification are Yes/No, Pass/Fail, Win/Lose, Cancerous/Non-cancerous, etc.

6.4.1 Types of Logistic Regression

- **Simple Logistic Regression:** A single independent variable is used to predict the output.
- **Multiple Logistic Regression:** Multiple independent variables are used to predict the output.

6.4.2 Extensions of Logistic Regression

Although it is said that Logistic Regression is used for Binary Classification, it can be extended to solve multi-class classification problems.

- **Multinomial Logistic Regression:** The output variable is discrete in three or more classes with no natural ordering.

Food Texture: Crunchy, Mushy, Crispy, etc.

Hair Colour: Blonde, Brown, Brunette, Red, etc.

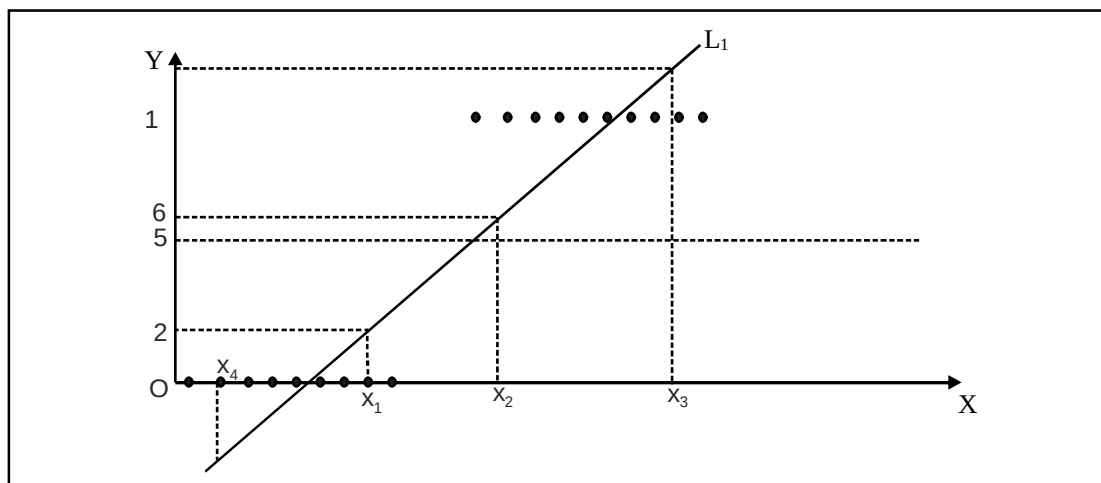


Fig. 6.1: Linear Regression

- **Ordered Logistic Regression:** Also known as a Ordinal regression model. The output variable is discrete in three or more classes with the ordering of the levels.

Customer rating: Extremely dislike, dislike, neutral, like, extremely like, etc.

Income level: Low income, middle income, high income, etc.

6.4.3 Use Linear Regression for Classification

Now, let us try if we can use linear regression to solve a binary class classification problem. Assume we have a dataset that is linearly separable and has the output that is discrete in two classes (0, 1).

In Linear regression, we draw a straight line (the best fit line) L_1 such that the sum of distances of all the data points to the line is minimal. The equation of the line L_1 is $y = mx + c$, where, m is the slope and c is the y-intercept.

We define a threshold $T = 0.5$, above which the output belongs to class 1 and class 0 otherwise.

$y = mx + c$, Threshold $T = 0.5$

$$y = \begin{cases} 1, & mx + c \geq 0.5 \\ 0, & mx + c < 0.5 \end{cases}$$

Case 1: The predicted value for x_1 is ≈ 0.2 which is less than the threshold. So, x_1 belongs to class 0.

Case 2: The predicted value for the point x_2 is ≈ 0.6 which is greater than the threshold. So, x_2 belongs to class 1.

Case 3: The predicted value for the point x_3 is beyond 1.

Case 4: The predicted value for the point x_4 is below 0.

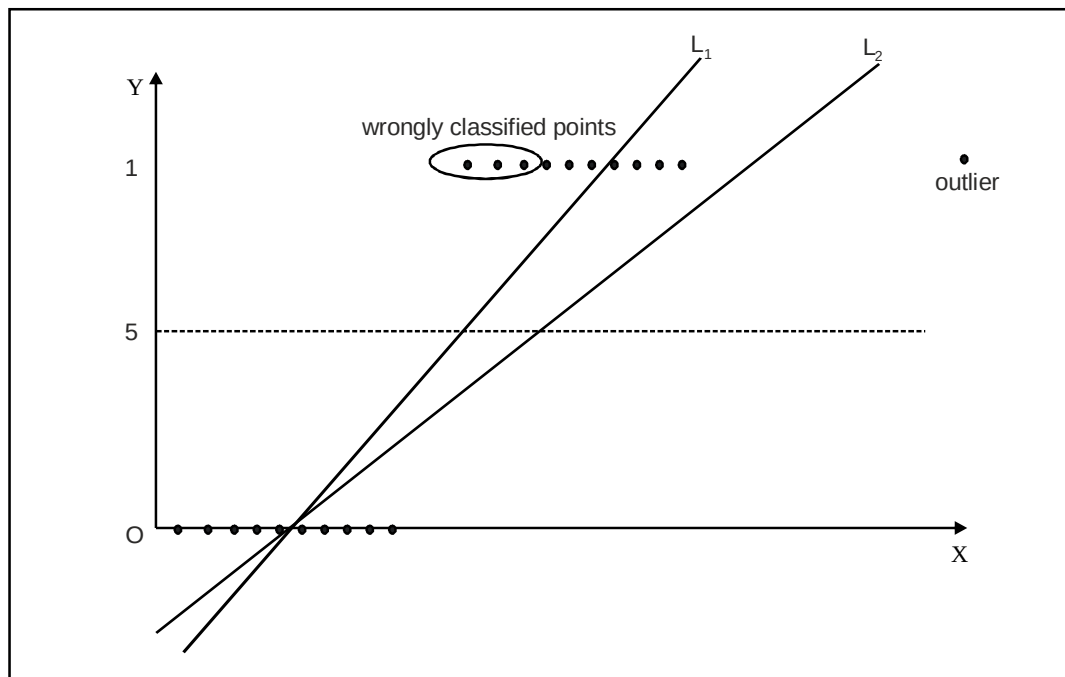


Fig. 6.2: Regression with Outlier

The predicted values for the points x_3, x_4 exceed the range (0,1) which does not make sense because the probability values always lie between 0 and 1. And our output can have only two values either 0 or 1. Hence, this is a problem with the linear regression model.

Now, introduce an outlier and see what happens. The regression line gets deviated to keep the distance of all the data points to the line to be minimal.

L_2 is the new best-fit line after the addition of an outlier. Seems good till now. But the problem is, if we closely observe, some of the data points are wrongly classified.

Certainly, it increases the error term. This again is a problem with the linear regression model.

The two limitations of using a linear regression model for classification problems are:

- the predicted value may exceed the range (0, 1)
- error rate increases if the data has outliers

There definitely is a need for Logistic regression here.

6.4.4 Logistic Regression Equation

The logistic regression equation is quite similar to the linear regression model.

Consider we have a model with one predictor “ x ” and one Bernoulli response variable “ $\hat{y}$ ” and p is the probability of $\hat{y} = 1$. The linear equation can be written as:

$$p = b_0 + b_1x \quad \dots (1)$$

The right-hand side of the equation ($b_0 + b_1x$) is a linear equation and can hold values that exceed the range (0, 1). But we know that probability will always be in the range of (0, 1).

To overcome that, we predict odds instead of probability.

Odds: The ratio of the probability of an event occurring to the probability of an event not occurring.

$$\text{Odds} = p/(1 - p)$$

The equation 1 can be rewritten as:

$$p/(1 - p) = b_0 + b_1x \quad \dots (2)$$

Odds can only be a positive value. To tackle the negative numbers, we predict the **logarithm of odds**.

$$\text{Log of odds} = \ln (p/(1 - p))$$

The equation (2) can be rewritten as:

$$\ln (p/(1 - p)) = b_0 + b_1x \quad \dots (3)$$

To recover p from equation (3), we apply exponential on both sides.

$$\exp (\ln (p/(1 - p))) = \exp (b_0 + b_1x)$$

$$e^{\ln (p/(1 - p))} = e^{(b_0 + b_1x)}$$

From the inverse rule of logarithms,

$$p/(1 - p) = e^{(b_0 + b_1x)}$$

Simple algebraic manipulations

$$p = (1 - p) * e^{(b_0 + b_1x)}$$

$$p = e^{(b_0 + b_1x)} - p * e^{(b_0 + b_1x)}$$

Taking p as common on the right-hand side, we get

$$p = p * ((e^{(b_0 + b_1x)})/p - e^{(b_0 + b_1x)})$$

$$p = e^{(b_0 + b_1x)} / (1 + e^{(b_0 + b_1x)})$$

Dividing numerator and denominator by $e^{(b_0 + b_1x)}$ on the right-hand side, we get

$$p = 1/(1 + e^{-(b_0 + b_1x)})$$

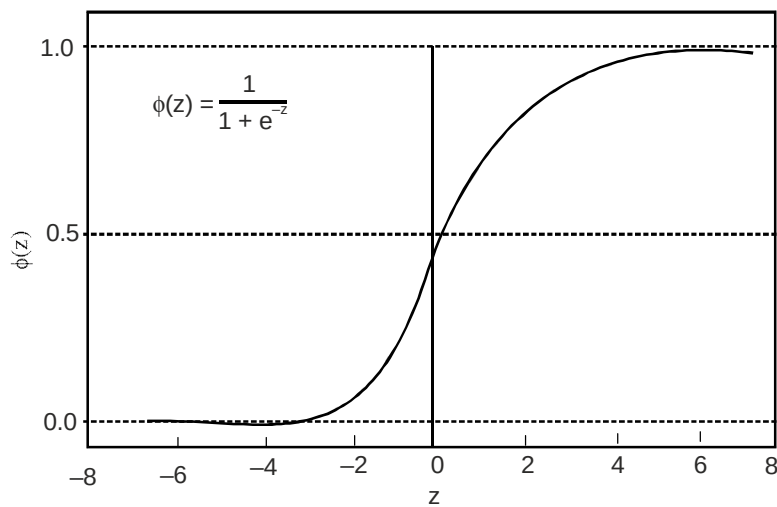
Similarly, the equation for a logistic model with 'n' predictors is as below:

$$p = 1 / (1 + e^{-(b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_nx_n)})$$

The right side part looks familiar, isn't it? Yes, it is the **sigmoid function**. It helps to squeeze the output to be in the range between 0 and 1.

Sigmoid Function

The sigmoid function is useful to map any predicted values of probabilities into another value between 0 and 1.



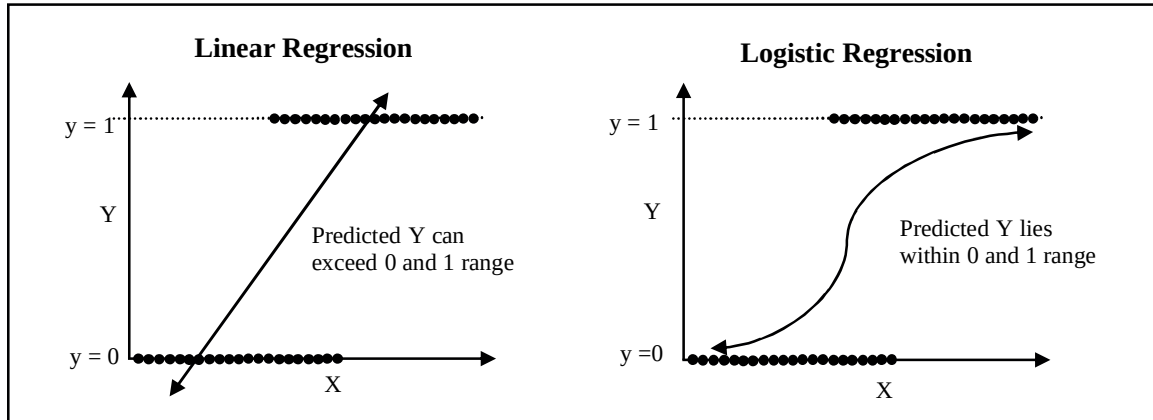
We started with a linear equation and ended up with a logistic regression model with the help of a sigmoid function.

Linear model: $\hat{y} = b_0 + b_1x$

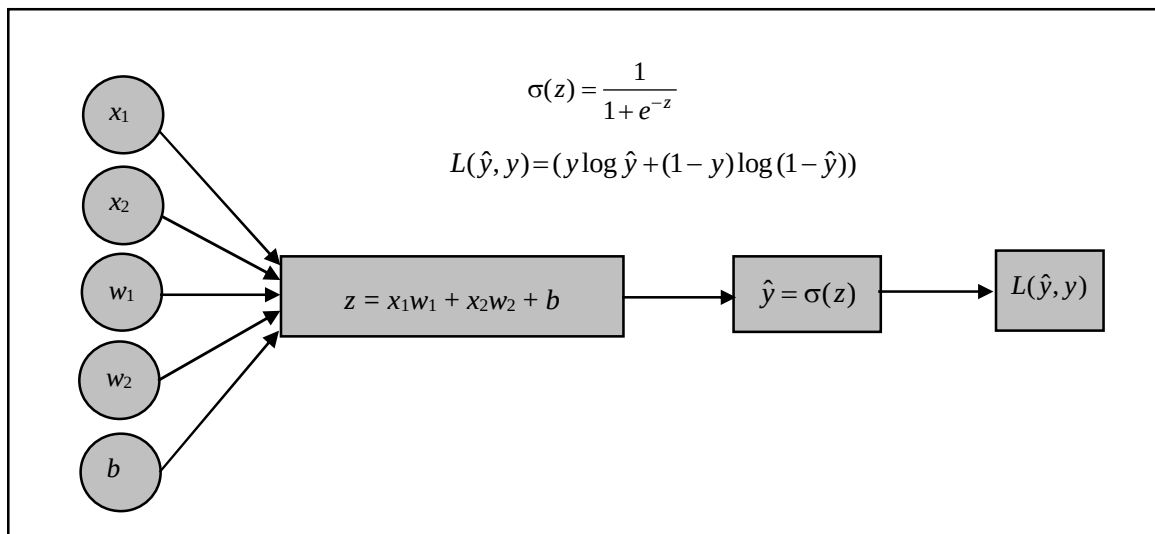
Sigmoid function: $\sigma(z) = 1/(1 + e^{-z})$

Logistic regression model: $\hat{y} = \sigma(b_0 + b_1x) = 1/(1 + e^{-(b_0 + b_1x)})$

So, unlike linear regression, we get an 'S' shaped curve in logistic regression.



The image that depicts the working of the Logistic regression model.



A linear equation (z) is given to a sigmoidal activation function (σ) to predict the output ($\hat{y}$).

To evaluate the performance of the model, we calculate the loss. The most commonly used loss function is the mean squared error.

But in logistic regression, as the output is a probability value between 0 or 1, mean squared error wouldn't be the right choice. So, instead, we use the cross-entropy loss function.

The cross-entropy loss function is used to measure the performance of a classification model whose output is a probability value.

Implementation in Python using Scikit-Learn Library

For classification, using a popular Fish dataset from Kaggle, add the dataset to GitHub repository for easy access.

Define the dataset URL:

```
dataset_url = "https://raw.githubusercontent.com/harika-bonthu/02-linear-regression-fish/master/datasets_229906_491820_Fish.csv"
```

Create a p and as data frame from the fish dataset

```
import pandas as pd
fish = pd.read_csv(dataset_url, error_bad_lines=False)
fish.head()
```

	Species	Weight	Length1	Length2	Length3	Height	Width
0	Bream	242.0	23.2	25.4	30.0	11.5200	4.020
1	Bream	290.0	24.0	26.3	31.2	12.4800	4.3056
2	Bream	340.0	23.9	26.5	31.1	12.3778	4.6961
3	Bream	363.0	26.3	29.0	33.5	12.7300	4.4555
4	Bream	430.0	26.5	29.0	34.0	12.4440	5.1340

Checking unique categories of the target feature. Our dataset has information about 7 fish species

```
fish['Species'].unique()  
array(['Bream', 'Roach', 'Whitefish', 'Parkki', 'Perch', 'Pike', 'Smelt'], dtype=object)
```

Check if there are any null values. Our dataset has no null values.

```
fish.is null().sum()  
  
Species      0  
Weight       0  
Length1      0  
Length2      0  
Length3      0  
Height       0  
Width        0  
dtype: int64
```

Defining input and target variables:

```
x = fish.iloc[:, 1:]  
y = fish.loc[:, 'Species']
```

Scaling the input features using MinMaxScaler:

```
from sklearn.preprocessing import MinMaxScaler  
scaler = MinMaxScaler()  
scaler.fit(X)  
X_scaled = scaler.transform(X)
```

Label encoding the target variable using LabelEncoder:

```
from sklearn.preprocessing import LabelEncoder  
label_encoder = LabelEncoder()  
y = label_encoder.fit_transform(y)
```

Splitting into train and test datasets using train_test_split:

```
learn.model_selection import train_test_split  
, X_train, X_test, y_train, y_test= train_test_split(X_scaled, y, test_size=0.2, random_state=42)
```

Model Building and Training:

```
from sklearn.linear_model import LogisticRegression  
clf = LogisticRegression()
```



```
# training the model
clf.fit(X_train, y_train)
```

Predicting the output:

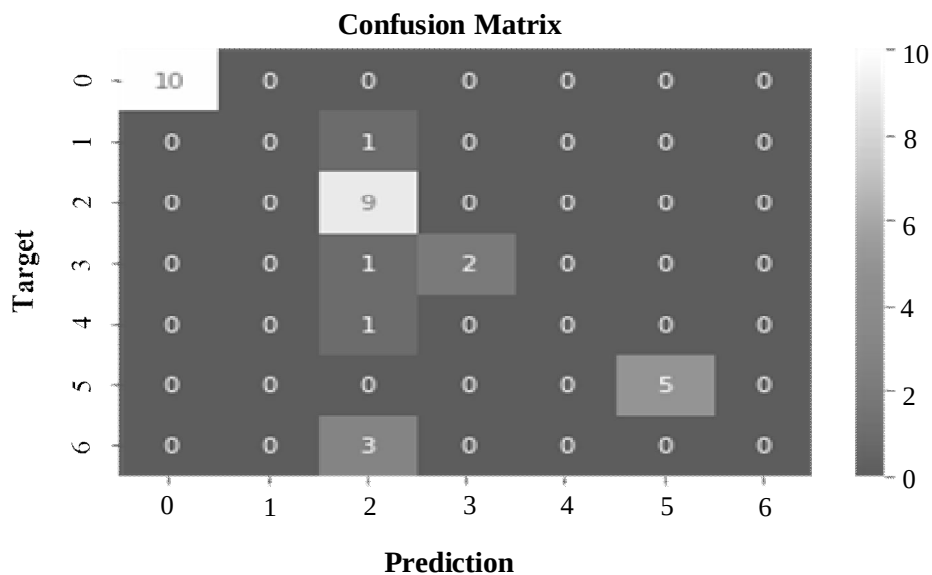
```
y_pred = clf.predict (X_test)
```

Computing the accuracy:

```
from sklearn.metrics import accuracy_score
accuracy = accuracy_score(y_test, y_pred)
print ("Accuracy: {:.2f}%".format(accuracy * 100))
```

Confusion Matrix

```
from sklearn.metrics import confusion_matrix
cf = confusion_matrix(y_test, y_pred)
plt.figure()
sns.heatmap(cf, annot=True)
plt.xlabel('Prediction')
plt.ylabel('Target')
plt.title('Confusion Matrix')
```



6.4.5 Assumptions for Logistic Regression

- The response variable Y is a binomial random variable with a single trial and success probability π . Thus, $Y = 1$ corresponds to “success” and occurs with probability π , and $Y = 0$ corresponds to “failure” and occurs with probability $1 - \pi$.
- The set of predictor or explanatory variables $x = (x_1, x_2, \dots, x_k)$ are fixed (not random) and can be discrete, continuous, or a combination of both. As with classical regression, two or more of these may be indicator variables to model the nominal categories of a single predictor, and others may represent interactions between two or more explanatory variables.
- Together, the data is collected for the i^{th} individual in the vector $(x_{1i}, \dots, x_{ki}, y_i)$, for $i = 1, \dots, n$. These are assumed independent by the sampling mechanism. This also allows us to combine or group the data, which we do below, by summing over trials for which π is constant.

Logistic regression is a method that we can use to fit a regression model when the response variable is binary.

Before fitting a model to a dataset, logistic regression makes the following assumptions:

Assumption #1: The Response Variable is Binary

Logistic regression assumes that the response variable only takes on two possible outcomes. Some examples include:

- Yes or No
- Male or Female
- Pass or Fail
- Drafted or Not Drafted
- Malignant or Benign

How to check this assumption?: Simply count how many unique outcomes occur in the response variable. If there are more than two possible outcomes, you will need to perform ordinal regression instead.

Assumption #2: The Observations are Independent

Logistic regression assumes that the observations in the dataset are independent of each other, *i.e.*, the observations should not come from repeated measurements of the same individual or be related to each other in any way.

How to check this assumption?: The easiest way to check this assumption is to create a plot of residuals against time (*i.e.*, the order of the observations) and observe whether or not there is a random pattern. If there is *not* a random pattern, then this assumption may be violated.

Assumption #3: There is No Multicollinearity among Explanatory Variables

Logistic regression assumes that there is no severe multicollinearity among the explanatory variables.

Multicollinearity occurs when two or more explanatory variables are highly correlated to each other, such that they do not provide unique or independent information in the regression model. If the degree of correlation is high enough between variables, it can cause problems when fitting and interpreting the model.

For example, suppose you want to perform logistic regression using **max vertical jump** as the response variable and the following variables as explanatory variables:

- **Player height**
- **Player shoe size**
- **Hours spent practicing per day**

In this case, **height** and **shoe size** are likely to be highly correlated since taller people tend to have larger shoe sizes. This means that multicollinearity is likely to be a problem if we use both of these variables in the regression.

How to check this assumption?: The most common way to detect multicollinearity is by using the variance inflation factor (VIF), which measures the correlation and strength of correlation between the predictor variables in a regression model.

Assumption #4: There are No Extreme Outliers

Logistic regression assumes that there are no extreme outliers or influential observations in the dataset.

How to check this assumption?: The most common way to test for extreme outliers and influential observations in a dataset is to calculate Cook's distance for each observation. If there are indeed outliers, you can choose to – (1) remove them, (2) replace them with a value like the mean or median, or (3) simply keep them in the model but make a note about this when reporting the regression results.

Assumption #5: There is a Linear Relationship between Explanatory Variables and the Logit of the Response Variable

Logistic regression assumes that there exists a linear relationship between each explanatory variable and the logit of the response variable. Recall that the logit is defined as:

$\text{Logit}(p) = \log(p/(1 - p))$, where, p is the probability of a positive outcome.

How to check this assumption?: The easiest way to see if this assumption is met is to use a Box-Tidwell test.

Assumption #6: The Sample Size is Sufficiently Large

Logistic regression assumes that the sample size of the dataset is large enough to draw valid conclusions from the fitted logistic regression model.

How to check this assumption?: As a rule of thumb, you should have a minimum of 10 cases with the least frequent outcome for each explanatory variable. For example, if you have 3 explanatory variables and the expected probability of the least frequent outcome is 0.20, then you should have a sample size of at least $(10 \times 3) / 0.20 = 150$.

Assumptions of Logistic Regression vs. Linear Regression

In contrast to linear regression, logistic regression does not require:

- A linear relationship between the explanatory variable(s) and the response variable.
- The residuals of the model to be normally distributed.
- The residuals to have constant variance, also known as homoscedasticity.

We all have used one of the following supervised learning algorithms for predictive analysis:

1. Logistic Regression
2. Ridge Regression
3. LASSO Regression
4. Linear Discriminant Analysis (LDA)
5. k -Nearest Neighbors (KNN)
6. Naive Bayes (NB)
7. Support Vector Machine (SVM)
8. Decision Tree
9. Random Forest (RF)
10. Gradient Boosting

Pros or Cons

1. Logistic Regression:

Pros:

- (a) It is used when the data is linearly separable.
- (b) It is easier to implement, interpret and very efficient to train.
- (c) It gives the measure of the importance of a predictor both in positive or negative direction.

Cons:

- (a) It can overfit in high-dimensional datasets.
- (b) It does not support non-linear relationship between the predictor and the outcome.

2. Ridge Regression:

Pros:

- (a) It prevents overfitting in higher dimensions.

- (b) It balances bias-variance trade-off. Sometimes having higher bias than zero can give better fit than high variance and zero bias.

Cons:

- (a) It increases bias.
- (b) We need to select optimal alpha (hyper parameter).
- (c) Model interpretability is low.

3. LASSO Regression:**Pros:**

- (a) It performs feature selection by shrinking coefficients towards zero.
- (b) It avoids overfitting.

Cons:

- (a) Selected features can be highly biased.
- (b) For $n < p$ (n – number of data points, p – number of features), LASSO selects at most n features.
- (c) For different bootstrapped data, the features selected can be very different.

4. Linear Discriminant Analysis (LDA):**Pros:**

- (a) It is simple, fast and portable algorithm. It still beats some algorithms (logistic regression) when its assumptions are met.

Cons:

- (a) It requires normal distribution assumption on features/predictors.
- (b) Sometimes, it is not good for few categories variables.

5. k -Nearest Neighbours (KNN)**Pros:**

- (a) It is the simplest algorithm to implement with just one parameter, number of neighbours k .
- (b) One can plug in any distance metric even defined by the user. This allows working with complex objects like time series, graphs, geographical coordinates, and basically anything you can define distance metric for.
- (c) This algorithms can be used for classification, ranking, regression (using neighbour's average or weighted average), recommendations, missing value imputation, etc.

Cons:

- (a) KNN is a lazy learner because it does not learn a model weights or function from the training data but “memorises” the training dataset instead. Hence, it takes longer time for inference than training.
- (b) It is a distance based-approach. Hence, the model can be badly affected by outliers. In other words, it is prone to overfitting.
- (c) Model size grows with the new data incorporated.
- (d) The model suffers from the curse of dimensionality.

6. Naive Bayes (NB)**Pros:**

- (a) It is easy and fast to predict class of test dataset. It also performs well in multiclass prediction.
- (b) When assumption of independence holds, a NB classifier performs better compared to other models like logistic regression and you need less training data.

Cons:

- (a) If categorical variable has a category (in test dataset), which was not observed in training dataset, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as “Zero Frequency”. To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is called Laplace estimation.
- (b) There is a strong set of assumptions on the distribution of the features such as normal, multinomial, etc.
- (c) There is also the assumption of independence in predictors. In real life, it is almost impossible that we get predictors which are completely independent.

7. Support Vector Machine (SVM):**Pros:**

- (a) It works really well with a clear margin of separation.
- (b) It is effective in high-dimensional spaces.
- (c) It uses a subset of training points in the decision function (called support vectors). So, it is also memory efficient.
- (d) It is also efficient in non-linear model fitting.

Cons:

- (a) It does not perform well when we have large dataset because the training becomes time-consuming.

- (b) It also does not perform very well, when the dataset has more noise, *i.e.*, target classes are overlapping.
- (c) SVM does not directly output probability. Other methods need to be used to convert the output from SVM to probability.

8. Decision Tree:

Pros:

- (a) Easy to understand and interpret; perfect for visual representation.
- (b) It requires little data pre-processing, *i.e.*, no need for one-hot encoding, standardisation and so on.
- (c) It is a non-parametric model. Hence, no assumptions is required about the distribution of the data.
- (d) Feature selection happens automatically. So, unimportant features will not influence the result.

Cons:

- (a) It tends to overfit.
- (b) It is very sensitive. Small change in the data can affect prediction greatly (High variance).

9. Random Forest (RF):

Pros:

- (a) It is robust to outliers. It lowers risk of overfitting.
- (b) It also works well with non-linear data.
- (c) It runs efficiently on a large dataset.
- (d) Usually, it gives better accuracy than other classification algorithms.

Cons:

- (a) Random forests are found to be biased while dealing with categorical variables.
- (b) Slow training.
- (c) It is not suitable for linear methods with a lot of sparse features.

10. Gradient Boosting:

Pros:

- (a) Boosting comes with an easy to read and interpret algorithm, making its prediction interpretations easy to handle.
- (b) Boosting is a resilient method that curbs overfitting easily.

Cons:

- (a) It is sensitive to outliers since every classifier is obliged to fix the errors in the predecessors. Thus, the method is too dependent on outliers.
- (b) Boosting is almost impossible to scale up. This is because every estimator bases its correctness on the previous predictors, thus making the procedure difficult to streamline.

Single Categorical Predictor

In this section, we will focus on a single explanatory variable x .

The model is expressed as:

$$\log(\pi_{i1} - \pi_i) = \beta_0 + \beta_1 x_i$$

Or, by solving for π_i , we have the equivalent expression

$$\pi_i = \exp(\beta_0 + \beta_1 x_i) / (1 + \exp(\beta_0 + \beta_1 x_i))$$

To estimate the parameters, we substitute this expression for π_i into the joint pdf (joint probability density function) for $y_1, \dots, y_n$.

$$\Pi_{i=1}^n = \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$$

to give us the likelihood function $L(\beta_0, \beta_1)$ of the regression parameters. By maximizing this likelihood over all possible β_0 and β_1 , we have the maximum likelihood estimates (MLEs): $\hat{\beta}_0$ and $\hat{\beta}_1$. Extending this to include additional explanatory variables is straightforward.

Working with Grouped Data

It is worth noting that when the same value of the predictor x occurs more than once, the success probability is constant for all Y responses associated with that x value, and we can work with the sum of such Y s as a binomial with the number of trials equal to the number of times that x value occurred. Our example with parent and student smoking illustrates this. Recall the data for the two-way table:

	Student smokes	Student does not smoke
1-2 parents smoke	816	3203
Neither parent smokes	188	1168

We have two approaches to work with. The first is for the individual or ungrouped data: $x_i = 1$ if the i^{th} student's parents smoke (0 otherwise), $Y_i = 1$ if the i^{th} student smokes, and $Y_i \sim \text{Binomial}(1, \pi_i)$, for $i = 1, \dots, 5375$. The second is for the grouped data: $x_i = 1$ if the i^{th} table row corresponds to parents smoking (0 otherwise), $Y_i =$ number of students who smoke for the i^{th} row, and $Y_i \sim \text{Binomial}(n_i, \pi_i)$, for $i = 1, 2$ and $(n_1, n_2) = (4019, 1356)$.

With either approach, π is the probability that a student smokes, and its estimate will be the same as well. Even though the index has a much greater range in the ungrouped approach, there are still only two unique values for π : (i) that for the group whose parents smoke and (ii) that for

the group whose parents don't. The likelihood function is simplified for the grouped data, however, and becomes,

$$\Pi_{i=1}^2 = \frac{n_i!}{y_i!(n_i - y_i)} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}$$

The range for i in the grouped approach is denoted by N to distinguish it from the individual total n (so, $N = 2$ in this example). Whenever possible, we will generally prefer to work with the grouped data because, in addition to having a more compact likelihood function, the larger binomial counts can be approximated by a normal distribution and will lend themselves to more meaningful residual diagnostics and large-sample significance tests.

Grouping is not always possible, however. Unless two responses have the same predictor value (or combination of predictor values if two or more predictors are present), their success probabilities would not be the same, and they cannot be added together to give a binomial variable. This is a common issue when lots of predictors and/or interactions are included in the model or with continuous predictor types because their values tend to be unique.

Interpretation of Parameter Estimates

One source of complication when interpreting parameters in the logistic regression model is that they are on the logit or log-odds scale. We need to be careful to convert them back before interpreting the terms of the original variables.

- $\exp(\beta_0)$ = the odds that the success characteristic is present for an individual for which $x = 0$, *i.e.*, at the baseline. If multiple predictors are involved, all would need to be set to 0 for this interpretation.
- $\exp(\beta_1)$ = the multiplicative increase in the odds of success for every 1-unit increase in x . This is similar to simple linear regression but instead of an additive change, it is a multiplicative change in rate. If multiple predictors are involved, others would need to be held fixed for this interpretation.
- If $\beta_1 > 0$, then $\exp(\beta_j) > 1$, indicating a positive relationship between x and the probability and odds of the success event. If $\beta_j < 0$, then the opposite holds.

Conducting Logistic Regression using RATTLE

Many a time, situations arise where the dependent variable is not normally distributed, *i.e.*, the assumption of normality is violated. For example, think of a problem when the dependent variable is binary (Male/Female). Will you still use Multiple Regression? Of course not! Why? We'll look at it below.

Let's take a peek into the history of data analysis.

So, until 1972, people did not know how to analyze data which has a non-normal error distribution in the dependent variable. Then, in 1972, came a breakthrough by **John Nelder** and **Robert Wedderburn** in the form of Generalized Linear Models.

Generalized Linear Models are an extension of the linear model framework, which includes dependent variables which are non-normal also. In general, they possess three characteristics:

1. These models comprise a linear combination of input features.
2. The mean of the response variable is related to the linear combination of input features via a link function.
3. The response variable is considered to have an underlying probability distribution belonging to the family of exponential distributions such as binomial distribution, Poisson distribution, or Gaussian distribution. Practically, binomial distribution is used when the response variable is binary. Poisson distribution is used when the response variable represents count and Gaussian distribution is used when the response variable is continuous.

Logistic Regression belongs to the family of generalized linear models. It is a binary classification algorithm used when the response variable is dichotomous (1 or 0). Inherently, it returns the set of probabilities of target class. But, we can also obtain response labels using a probability threshold value. Following are the assumptions made by Logistic Regression:

1. The response variable must follow a binomial distribution.
2. Logistic Regression assumes a linear relationship between the independent variables and the link function (logit).
3. The dependent variable should have mutually exclusive and exhaustive categories.

In *R*, we use **glm() function** to apply Logistic Regression. In Python, we use **sklearn.linear_model** function to import and use Logistic Regression.

Note: We do not use Linear Regression for binary classification because its linear function results in probabilities outside [0,1] interval, thereby making them invalid predictions.

6.4.6 What are the Types of Logistic Regression Techniques?

Logistic Regression is not just limited to solving binary classification problems. To solve problems that have multiple classes, we can use extensions of Logistic Regression, which includes **Multinomial Logistic Regression** and **Ordinal Logistic Regression**. Let's get their basic idea:

1. Multinomial Logistic Regression: Let's say our target variable has $K = 4$ classes. This technique handles the multiclass problem by fitting $K - 1$ independent binary logistic classifier model. For doing this, it randomly chooses one target class as the reference class and fits $K - 1$ regression models that compare each of the remaining classes to the reference class.

Due to its restrictive nature, it is not used widely because it does not scale very well in the presence of a large number of target classes. In addition, since it builds $K - 1$ models, we would require a much larger dataset to achieve reasonable accuracy.

2. Ordinal Logistic Regression: This technique is used when the target variable is ordinal in nature. Let's say, we want to predict years of work experience (1, 2, 3, 4, 5, etc). So, there

exists an order in the value, *i.e.*, $5 > 4 > 3 > 2 > 1$. Unlike a multinomial model, when we train $K - 1$ models, Ordinal Logistic Regression builds a single model with multiple threshold values.

If we have K classes, the model will require $K - 1$ threshold or cut-off points. Also, it makes an imperative assumption of proportional odds. The assumption says that on a logit (S shape) scale, all of the thresholds lie on a straight line.

Note: Logistic Regression is not a great choice to solve multiclass problems. But, it is good to be aware of its types. We'll focus on Logistic Regression for binary classification task.

6.4.7 How Does Logistic Regression Work?

Now comes the interesting part!

As we know, Logistic Regression assumes that the dependent (or response) variable follows a binomial distribution. Now, you may wonder, what is binomial distribution? Binomial distribution can be identified by the following characteristics:

1. There must be a fixed number of trials denoted by n , *i.e.*, in the dataset, there must be a fixed number of rows.
2. Each trial can have only two outcomes, *i.e.*, the response variable can have only two unique categories.
3. The outcome of each trial must be independent of each other, *i.e.*, the unique levels of the response variable must be independent of each other.
4. The probability of success (p) and failure (q) should be the same for each trial.

Let's understand how Logistic Regression works. For Linear Regression, where the output is a linear combination of input feature(s), we write the equation as:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

In Logistic Regression, we use the same equation but with some modifications made to Y . Let's reiterate a fact about Logistic Regression: we calculate probabilities. And, probabilities always lie between 0 and 1. In other words, we can say:

1. The response value must be positive.
2. It should be lower than 1.

First, we'll meet the above two criteria. We know the exponential of any value is always a positive number. And, any number divided by number + 1 will always be lower than 1. Let's implement these two findings:

$$P(Y = 1|X) = \frac{e^{(\beta_0 + \beta_1 x)}}{e^{(\beta_0 + \beta_1 x)} + 1}$$

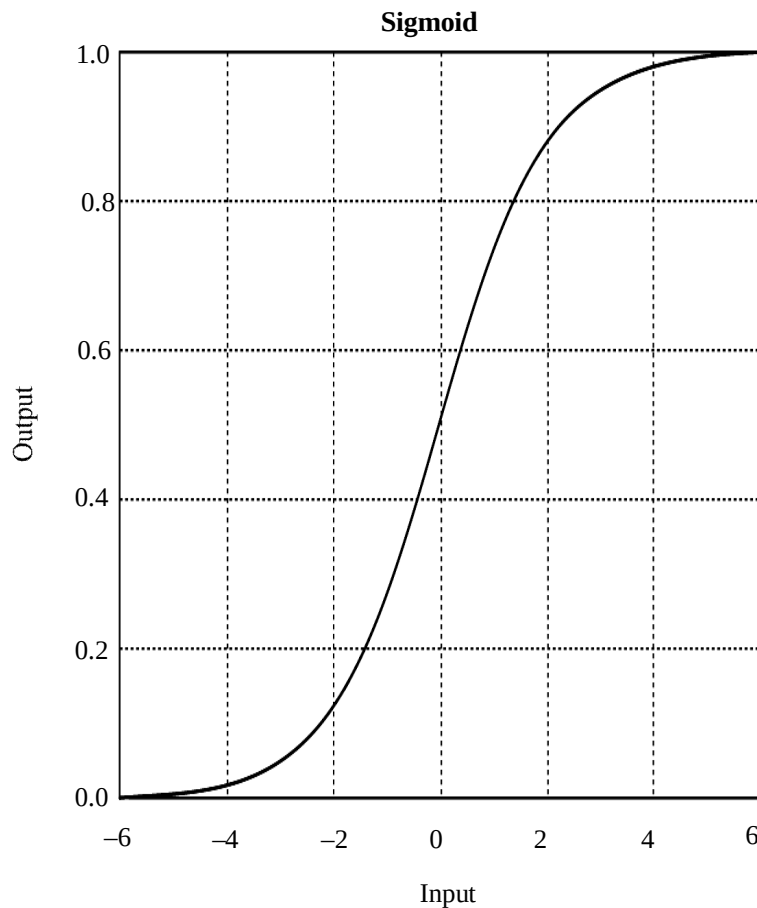
This is the logistic function.

Now, we are convinced that the probability value will always lie between 0 and 1. To determine the link function, follow the algebraic calculations carefully. $P(Y=1|X)$ can be read as

“probability that $Y = 1$ given some value for x .” Y can take only two values, 1 or 0. For ease of calculation, let’s rewrite $P(Y=1|X)$ as $p(X)$.

$$\begin{aligned}
 \Rightarrow p(X) &= \frac{e^{(\beta_0 + \beta_1 x)}}{e^{(\beta_0 + \beta_1 x)} + 1} \\
 \Rightarrow p(e^{(\beta_0 + \beta_1 x)} + 1) &= e^{(\beta_0 + \beta_1 x)} \\
 \Rightarrow p \cdot e^{(\beta_0 + \beta_1 x)} + p &= e^{(\beta_0 + \beta_1 x)} \\
 \Rightarrow p &= e^{(\beta_0 + \beta_1 x)} - p \cdot e^{(\beta_0 + \beta_1 x)} \\
 \Rightarrow p &= e^{(\beta_0 + \beta_1 x)}(1 - p) \\
 \Rightarrow \frac{p}{1 - p} &= e^{(\beta_0 + \beta_1 x)} \\
 \Rightarrow \ln\left(\frac{p}{1 - p}\right) &= \beta_0 + \beta_1 x
 \end{aligned}$$

As you might recognize, the right side of the (immediate) equation above depicts the linear combination of independent variables. The left side is known as the **log-odds** or **odds ratio** or **logit** function and is the link function for Logistic Regression. This link function follows a sigmoid (shown below) function which limits its range of probabilities between 0 and 1.



Until here, I hope you have understood how we derive the equation of Logistic Regression. But how is it interpreted?

We can interpret the above equation as, a unit increase in variable x results in multiplying the odds ratio by e to power β . In other words, the regression coefficients explain the change in $\log(\text{odds})$ in the response for a unit change in predictor. However, since the relationship between $p(X)$ and X is not straight line, a unit change in input feature does not really affect the model output directly but it affects the odds ratio.

This is contradictory to Linear Regression where, regardless of the value of input feature, the regression coefficient always represents a fixed increase/decrease in the model output per unit increase in the input feature.

In Multiple Regression, we use the **Ordinary Least Square (OLS)** method to determine the best coefficients to attain good model fit. In Logistic Regression, we use **maximum likelihood method** to determine the best coefficients and eventually a good model fit.

Maximum likelihood works like this: It tries to find the value of coefficients (β_0, β_1) such that the predicted probabilities are as close to the observed probabilities as possible. In other words, for a binary classification (1/0), maximum likelihood will try to find values of β_0 and β_1 such that the resultant probabilities are closest to either 1 or 0. The likelihood function is written as:

$$l(\beta_0, \beta_1) = \prod_{i: y_i=1} p(x_i) \prod_{i: y_i'=0} (1 - p(x_i'))$$

6.4.8 How Can You Evaluate Logistic Regression Model Fit and Accuracy?

In Linear Regression, we check adjusted R^2 , F Statistics, MAE, and RMSE to evaluate model fit and accuracy. But, Logistic Regression employs all different sets of metrics. Here, we deal with probabilities and categorical values. Following are the evaluation metrics used for Logistic Regression:

1. Akaike Information Criteria (AIC)

You can look at AIC as counterpart of adjusted r square in multiple regression. It is an important indicator of model fit. It follows the rule: Smaller the better. AIC penalizes increasing number of coefficients in the model. In other words, adding more variables to the model would not let AIC increase. It helps to avoid overfitting.

Looking at the AIC metric of one model would not really help. It is more useful in comparing models (model selection). So, build 2 or 3 Logistic Regression models and compare their AIC. The model with the lowest AIC will be relatively better.

2. Null Deviance and Residual Deviance

Deviance of an observation is computed as -2 times $\log$ likelihood of that observation. The importance of deviance can be further understood using its types: Null and Residual Deviance.

Null deviance is calculated from the model with no features, *i.e.*, only intercept. The null model predicts class via a constant probability.

Residual deviance is calculated from the model having all the features. On comparison with Linear Regression, think of residual deviance as residual sum of square (RSS) and null deviance as total sum of squares (TSS). The larger the difference between null and residual deviance, better the model.

Also, you can use these metrics to compared multiple models: whichever model has a lower null deviance, means that the model explains deviance pretty well, and is a better model. Also, lower the residual deviance, better the model. Practically, AIC is always given preference above deviance to evaluate model fit.

3. Confusion Matrix

Confusion matrix is the most crucial metric commonly used to evaluate classification models. It is quite confusing but make sure you understand it by heart. If you still do not understand anything, ask me in comments. The skeleton of a confusion matrix looks like this:

	1 (Predicted)	0 (Predicted)
1 (Actual)	True Positive	False Negative
0 (Actual)	False Positive	True Negative

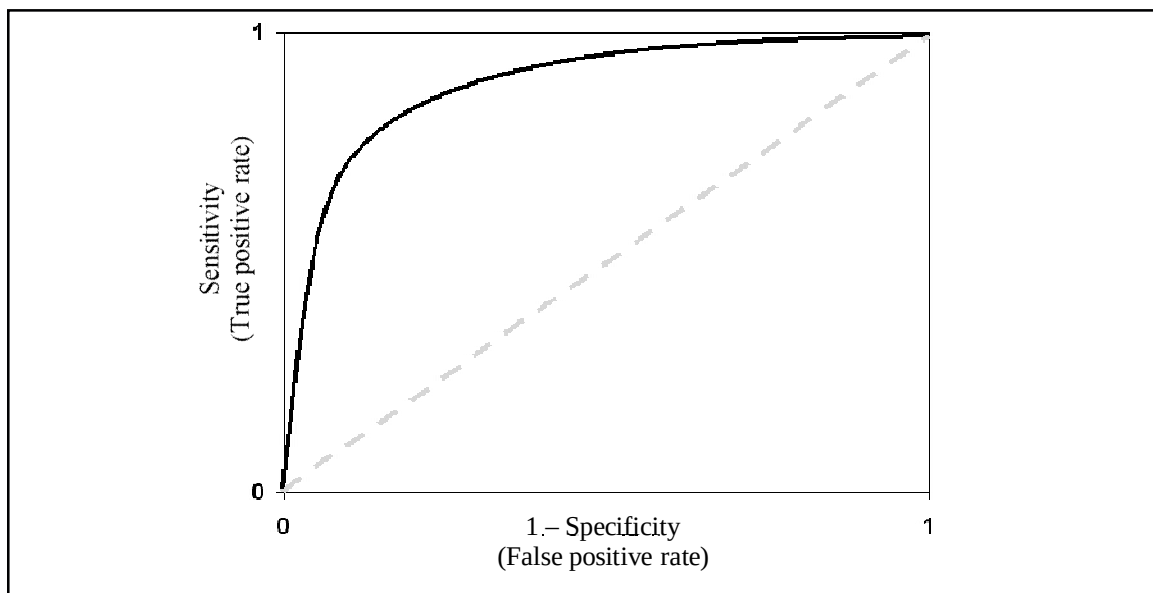
As you can see, the confusion matrix avoids “confusion” by measuring the actual and predicted values in a tabular format. In table above, Positive class = 1 and Negative class = 0. Following are the metrics, we can derive from a confusion matrix:

- **Accuracy:** It determines the overall predicted accuracy of the model. It is calculated as $\text{Accuracy} = (\text{True Positives} + \text{True Negatives}) / (\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives})$.
- **True Positive Rate (TPR):** It indicates how many positive values, out of all the positive values, have been correctly predicted. The formula to calculate the true positive rate is $(\text{TP} / \text{TP} + \text{FN})$. Also, $\text{TPR} = 1 - \text{False Negative Rate}$. It is also known as **Sensitivity** or **Recall**.

- **False Positive Rate (FPR):** It indicates how many negative values, out of all the negative values, have been **incorrectly predicted**. The formula to calculate the false positive rate is $(FP/FP + TN)$. Also, $FPR = 1 - \text{True Negative Rate}$.
- **True Negative Rate (TNR):** It indicates how many negative values, out of all the negative values, have been **correctly predicted**. The formula to calculate the true negative rate is $(TN/TN + FP)$. It is also known as **Specificity**.
- **False Negative Rate (FNR):** It indicates how many positive values, out of all the positive values, have been incorrectly predicted. The formula to calculate false negative rate is $(FN/FN + TP)$.
- **Precision:** It indicates how many values, out of all the predicted positive values, are actually positive. It is formulated as: $(TP/TP + FP)$.
- **F Score:** F score is the harmonic mean of precision and recall. It lies between 0 and 1. Higher the value, better the model. It is formulated as $2((\text{precision} * \text{recall}) / (\text{precision} + \text{recall}))$.

4. Receiver Operator Characteristic (ROC)

ROC determines the accuracy of a classification model at a user-defined threshold value. It determines the model's accuracy using Area Under Curve (AUC). The area under the curve (AUC), also referred to as index of accuracy (A) or concordant index, represents the performance of the ROC curve. Higher the area, better the model. ROC is plotted between True Positive Rate (Y axis) and False Positive Rate (X Axis). In this plot, our aim is to push the curve (shown below) toward 1 (left corner) and maximize the area under curve. Higher the curve, better the model. The dashed line represents the ROC curve at 0.5 threshold. At this point, sensitivity = specificity.



Why Do We Use the Sigmoid Function?

Sigmoid Function acts as an activation function in machine learning which is used to add non-linearity in a machine learning model. In simple words, it decides which value to pass as output and what not to pass. There are mainly several types of Activation Functions which are used in machine learning and deep learning.

Conducting Logistic Regression Analysis using R Commander

R makes it very easy to fit a logistic regression model. The function to be called is `glm()` and the fitting process is similar the one used in linear regression.

The data which has been used is Bankloan. The dataset has 850 rows and 9 columns (age, education, employment, address, income, debtinc, creddebt, othdebt, default). The dependent variable is default (Defaulted and Not Defaulted).

Let's first load and check the head of data.

```
bankloan<-read.csv("bankloan.csv")  
head(bankloan)
```

Now, making the subset of the data with 700 rows.

```
mod_bankloan<- bankloan[1:700,]
```

Setting a seed of 1000 (meaning picking random numbers from 1000 as starting point)

```
set.seed(500)
```

Let's have a sample of 500 values. So, creating a variable of training data of 700 rows.

```
>train<-sample(1:700, 500, replace=FALSE)
```

Creating training as well as testing data.

```
>training data<- mod_bankloan [train,]  
>testing data<- mod_bankloan [-train,]
```

Now, let's fit the model. Be sure to specify the parameter family = binomial in the `glm()` function.

```
model1<-glm(default~.,family=binomial(link='logit'),data=trainingdata)  
>summary( model1)
```

The summary will also include the significance level of all the variables. If the p -value is less than 0.05, then the variables are significant. We can also remove the insignificant variables to make our accurate.


```

Call:
glm(formula = default ~ ., family = binomial(link = "logit"),
    data = trainingdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1205  -0.4437   0.2687   0.6506   1.9248

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.33645    0.73335   1.822 0.068398 .
age           -0.04664    0.02118  -2.202 0.027631 *
edDoctorate   -1.46871    1.80181  -0.815 0.415000
edGraduate    -0.09121    0.39746  -0.229 0.818483
edHigh School  0.51355    0.30506   1.683 0.092286 .
edPost Graduate 0.30883    0.53201   0.581 0.561571
employ        0.25192    0.04014   6.277 3.46e-10 ***
address        0.10382    0.02819   3.683 0.000231 ***
income         0.01843    0.01367   1.348 0.177629
debtinc       -0.06922    0.03908  -1.771 0.076521 .
creddebt      -0.61456    0.13987  -4.394 1.11e-05 ***
othdebt       -0.12168    0.10123  -1.202 0.229351
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 585.24  on 499  degrees of freedom
Residual deviance: 392.04  on 488  degrees of freedom
AIC: 416.04

Number of Fisher Scoring iterations: 6

```

In our model, only age, employment, addresses and creddebt seems to be significant. So, building another model with only these variables.

```

model12<glm(default~age+employ+address+creddebt,family=binomial(link='logit'),data=trainingdat

```

```

Call:
glm(formula = default ~ age + employ + address + creddebt, family = binomial(link = "logit"),
    data = trainingdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6638  -0.4807   0.3252   0.6957   1.9128

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.08409    0.54600   1.986 0.047087 *
age         -0.04213    0.01987  -2.120 0.034006 *
employ       0.28142    0.03418   8.234 < 2e-16 ***
address      0.10041    0.02655   3.782 0.000156 ***
creddebt     -0.80048    0.10136  -7.897 2.85e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 585.24  on 499  degrees of freedom
Residual deviance: 422.15  on 495  degrees of freedom
AIC: 432.15

Number of Fisher Scoring iterations: 6

```

Let's now predict the model with the training data.

```
pred1<-predict(model12,newdata=trainingdata, type="response")
```

Now looking at the probability with 0.5% flight delayed or on-time.

```
predicted_class<-ifelse(pred1<0.5, "Defaulted", "Not Defaulted")
```

Creating a table to see the same.

```
Table(training data $ default, predicted_class)
```

	predicted_class	
	Defaulted	Not Defaulted
0	64	72
Defaulted	64	72
Not Defaulted	36	328

This is also known as **confusion matrix**. It is a tabular representation of Actual vs. Predicted values. This helps us to find the accuracy or error of the model and avoid overfitting.

There are 64 customers who actually defaulted and our model also predicted the same. However, 72 customers defaulted but model predicted them as Not Defaulted. Also, 36 customers actually Not Defaulted where the model mentioned them as defaulted. Let's now find out the error rate.

```
err_rate<-1-sum((training data $ default ==predicted_class))/500>err_rate0.344
```

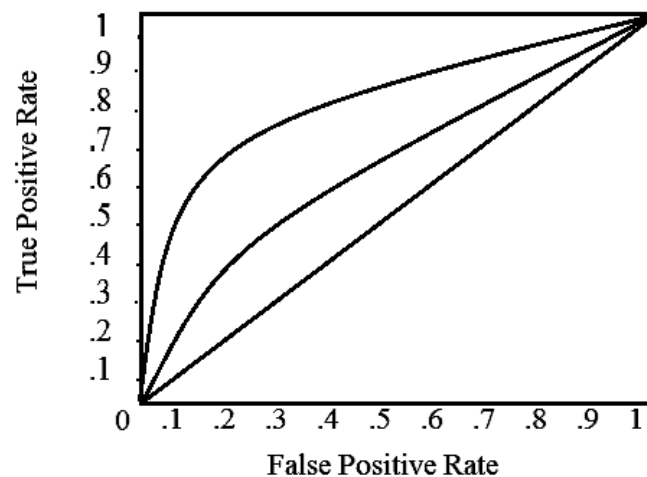
Which is 34%.

Going ahead, lets test the model on testing data.

```
pred2<-predict(model12, newdata=testing data,type="response")
predicted_class2<-ifelse(pred2<0.5, "Defaluted", "Not Defaulted")
table(testing data $ default, predicted_class2)err_rate<-1-sum((testing data $ default
==predicted_class2))/200err_rate0.31
```

Here, the error rate is 31%.

Now, we can plot this in **Receiver Operating Characteristics Curve** (commonly known as ROC curve). In R, it can be done by downloading a package called ROCR. An output of the plot is given below.



ROC traces the percentage of true positives accurately predicted by a given logit model as the prediction probability cut-off is lowered from 1 to 0. For a perfect model, as the cut-off is lowered, it should mark more of actual 1's as positives and lesser of actual 0's as 1's. The area under curve, known as index of accuracy, is a performance metric for the curve. Higher the area under curve, better the prediction power of the model.

Multiclass Logistic Regression

Instead of $y = 0, 1$, we will expand our definition so that $y = 0, 1, \dots, n$. Basically, we rerun binary classification multiple times, once for each class.

Procedure

1. Divide the problem into $n + 1$ binary classification problems (+1 because the index starts at 0?).
2. For each class...
3. Predict the probability the observations are in that single class.
4. prediction = $\max(\text{probability of the classes})$

For each sub-problem, we select one class (YES) and lump all the others into a second class (NO). Then we take the class with the highest predicted value.

Softmax Activation

The softmax function (softargmax or normalized exponential function) is a function that takes as input a vector of K real numbers, and normalizes it into a probability distribution consisting of K probabilities proportional to the exponentials of the input numbers, *i.e.*, prior to applying softmax, some vector components could be negative, or greater than one; and might not sum to 1; but after applying softmax, each component will be in the interval $[0, 1]$, and the components will add up to 1, so that they can be interpreted as probabilities. The standard (unit) softmax function is defined by the formula:

$$\sigma(z_i) = \frac{e^{z_{(i)}}}{\sum_{j=1}^K e^{z_{(j)}}} \quad \text{for } i=1, \dots, K \text{ and } z = z_1, \dots, z_K$$

In other words, we apply the standard exponential function to each element z_i of the input vector z and normalize these values by dividing by the sum of all these exponentials. This normalization ensures that the sum of the components of the output vector $\sigma(z)$ is 1.

Scikit-Learn Example

Let's compare our performance to the LogisticRegression model provided by scikit-learn.

```
import sklearn
from sklearn.linear_model import LogisticRegression
from sklearn.cross_validation import train_test_split

# Normalize grades to values between 0 and 1 for more efficient computation
normalized_range = sklearn.preprocessing.MinMaxScaler(feature_range=(-1,1))

# Extract Features + Labels
labels.shape = (100,) #scikit expects this
features = normalized_range.fit_transform(features)

# Create Test/Train
features_train, features_test, labels_train, labels_test =
train_test_split(features, labels, test_size=0.4)

# Scikit Logistic Regression
scikit_log_reg = LogisticRegression()
```

```

scikit_log_reg.fit(features_train,labels_train)

#Score is Mean Accuracy
scikit_score = clf.score(features_test,labels_test)
print 'Scikit score:', scikit_score

#Our Mean Accuracy
observations, features, labels, weights = run()
probabilities = predict(features, weights).flatten()
classifications = classifier(probabilities)
our_acc = accuracy (classifications,labels.flatten())
print 'Our score:', our_acc

```

Conducting Logistic Regression Analysis in PSPP

```

LOGISTIC REGRESSION [VARIABLES =] dependent_var WITH predictors
    [/CATEGORICAL = categorical_predictors]
    [{/NOCONST | /ORIGIN | /NOORIGIN }]
    [/PRINT = [SUMMARY] [DEFAULT] [CI(confidence)] [ALL]]
    [/CRITERIA = [BCON(min_delta)] [ITERATE(max_interations)]
        [LCON(min_likelihood_delta)] [EPS(min_epsilon)]
        [CUT(cut_point)]]
    [/MISSING = {INCLUDE|EXCLUDE}]

```

Bivariate Logistic Regression is used when to want to explain a dichotomous dependent variable in terms of one or more predictor variables.

The minimum command is:

```
LOGISTIC REGRESSION y WITH x1 x2 ... xn.
```

Here, y is the dependent variable, which must be dichotomous and $x_1 \dots x_n$ are the predictor variables whose coefficients the procedure estimates.

By default, a constant term is included in the model. Hence, the full model is:

$$\{bfy\} = b_0 + b_1 \{bfx_1\} + b_2 \{bfx_2\} + \dots + b_n \{bfx_n\}$$

Predictor variables which are categorical in nature should be listed on the /CATEGORICAL sub-command. Simple variables as well as interactions between variables may be listed here.

If you want a model without the constant term b_0 , use the keyword /ORIGIN. /NOCONST is a synonym for /ORIGIN.

An iterative Newton-Raphson procedure is used to fit the model. The /CRITERIA sub-command is used to specify the stopping criteria of the procedure and other parameters. The value of *cut_point* is used in the classification table. It is the threshold above which predicted values are considered to be 1. Values of *cut_point* must lie in the range [0,1]. During iterations, if any one of the stopping criteria are satisfied, the procedure is considered complete. The stopping criteria are:

- The number of iterations exceeds *max_iterations*. The default value of *max_iterations* is 20.
- The change in the all coefficient estimates are less than *min_delta*. The default value of *min_delta* is 0.001.
- The magnitude of change in the likelihood estimate is less than *min_likelihood_delta*. The default value of *min_delta* is zero. This means that this criterion is disabled.
- The differential of the estimated probability for all cases is less than *min_epsilon*. In other words, the probabilities are close to zero or one. The default value of *min_epsilon* is 0.00000001.

The PRINT sub-command controls the display of optional statistics. Currently, there is one such option, CI, which indicates that the confidence interval of the odds ratio should be displayed as well as its value. CI should be followed by an integer in parentheses to indicate the confidence level of the desired confidence interval.

The MISSING sub-command determines the handling of missing variables. If INCLUDE is set, then user-missing values are included in the calculations, but system-missing values are not. If EXCLUDE is set, which is the default, user-missing values are excluded as well as system-missing values. This is the default.

6.5 MULTIPLE PREDICTORS

The use of predictors **multiple times in the tree helps to detect complex interactions in the data.** (From: Handbook of Statistical Analysis and Data Mining Applications.)

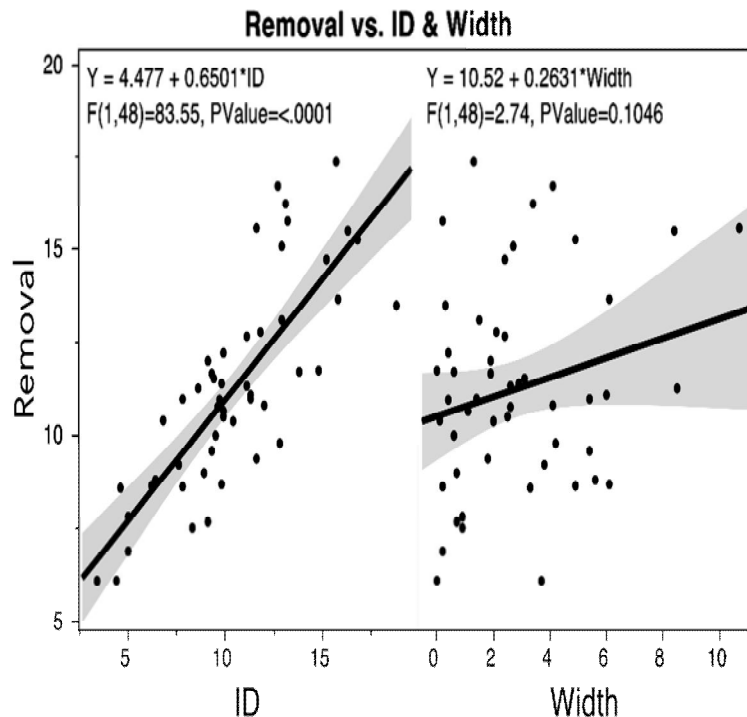
Multiple regression analysis is a powerful technique used for predicting **the unknown value of a variable from the known value of two or more variables** also called the predictors.

6.5.1 Linear Regression with Multiple Predictors

Recall that *simple linear regression* can be used to predict the value of a response based on the value of one continuous predictor variable. Depending on the context, the response and predictor variables might be referred to by other names. For simplicity, we'll generally stick with the terms *response and predictor* throughout this discussion.

Let's return to an earlier example. Using the Cleaning data, we used simple linear regression to model the relationship between the response, **Removal**, and the predictor, **OD**. We found that a 1-unit increase in **OD** is associated with a 0.53-unit increase in **Removal**. But, what do we do if

we have more than one predictor variable? For the cleaning example, we have three potential predictors, **OD**, **ID**, and **Width**. How can we extend our analysis of **Removal** to account for the additional predictors? One option would be to fit separate regression models for the different predictors.



Continuing with this example, we learn that there is a significant relationship between **Removal** and **ID**, and that for a 1-unit increase in **ID**, **Removal** increases by 0.65 units on average. We also learn that there is not a significant relationship between **Removal** and **Width**. In other words, there is no association between changes in **Width** and changes in **Removal**.

However, fitting simple linear regression models for each predictor ignores the information in the other variables.

- Because we can make predictions based on only one model at a time, we lose potentially valuable information in the other predictors. This means that we cannot make predictions with a model based on one predictor using the information from other predictors.
- Also, because each simple model ignores the other variables when the coefficients are estimated, the coefficients might be misleading, especially if the predictors are correlated with one another.

Instead of fitting separate models for each predictor, we can include multiple predictors in the same model. When more than one predictor is used, the procedure is called multiple linear regression.

Recall the unknown or true, linear regression model with one predictor:

$$\mu_{Y/X} = \beta_0 + \beta_1 X_1$$

This equation describes how the mean of Y changes for given values of X . We can also write the equation in terms of the observed values of Y , rather than the mean. Because the individual data values for any given value of X vary randomly about the mean, we need to account for this random variation, or error, in the regression equation. We add the Greek letter epsilon to the equation to represent the random error in the individual observations:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

When we fit a multiple linear regression model, we add a slope coefficient for each predictor. For the Cleaning example, with **OD** and **ID** as predictors, the model has slope coefficients for both predictors.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Each coefficient represents the average increase in Removal for every one unit increase in that predictor, holding the other predictor constant. What if we have more than two predictors? As a generalization, let's say that we have p predictors. The multiple linear regression model can be extended to include all p predictors.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$$

number of predictors

Linear regression models can also include functions of the predictors such as transformations, polynomial terms, and cross-products, or interactions. And later, we'll see that linear models can also be fit with categorical predictors.

$$Y = \beta_0 + \boxed{\beta_1 \sqrt{X_1}} + \boxed{\beta_2 \text{Log}(X_2)} + \dots + \beta_p X_p + \varepsilon$$

$$Y = \beta_0 + \beta_1 X_1 + \boxed{\beta_2 X_2} + \boxed{\beta_{12} X_1 X_2} + \dots + \beta_p X_p + \varepsilon$$

A challenge when fitting multiple linear regression models is that we might need to estimate many coefficients. Although modern statistical software can easily fit these models, it is not always straightforward to identify important predictors and interpret the model coefficients. In the sections that follow, we talk about fitting and interpreting multiple linear regression models and some of the challenges involved.

6.5.2 When to Use Multiple Regression?

You use multiple regression when you have three or more measurement variables. One of the measurement variables is the dependent (YY) variable. The rest of the variables are the independent (XX) variables. The purpose of a multiple regression is to find an equation that best predicts the YY variable as a linear function of the XX variables.

6.5.3 Multiple Regressions for Prediction

One use of multiple regressions is prediction or estimation of an unknown YY value corresponding to a set of XX values. For example, let's say you are interested in finding a suitable habitat to reintroduce the rare beach tiger beetle, *Cicindela dorsalis dorsalis*, which lives on sandy beaches on the Atlantic coast of North America. You have gone to a number of beaches that already have the beetles and measured the density of tiger beetles (the dependent variable) and several biotic and abiotic factors, such as wave exposure, sand particle size, beach steepness, density of amphipods and other prey organisms, etc. Multiple regressions would give you an equation that would relate the tiger beetle density to a function of all the other variables. Then, if you went to a beach that did not have tiger beetles and measured all the independent variables (wave exposure, sand particle size, etc.), you could use the multiple regression equation to predict the density of tiger beetles that could live there if you introduced them.



Atlantic Beach Tiger Beetle: This is the Atlantic beach tiger beetle (*Cicindela dorsalis dorsalis*), which is the subject of the multiple regression study in this atom.

6.5.4 Multiple Regressions for Understanding Causes

A second use of multiple regression is you try to understand the functional relationships between the dependent and independent variables, you try to see what might be causing the variation in the dependent variable. For example, if a regression of tiger beetle density on sand particle size by itself, you would probably see a significant relationship. If a regression of tiger beetle density on wave exposure by itself, you would probably see a significant relationship. However, sand particle size and wave exposure are correlated; beaches with bigger waves tend to have bigger sand particles. Maybe sand particle size is really important, and the correlation between it and wave exposure is the only reason for a significant regression between wave exposure and beetle density. Multiple regression is a statistical way you try to control for this; it can answer questions like, "If sand particle size (and every other measured variable) were the same, would the regression of beetle density on wave exposure be significant? "

6.5.5 Null Hypothesis

The main null hypothesis of a multiple regression is that there is no relationship between the X variables and the Y variables. In other words, that the fit of the observed Y values to those predicted by the multiple regression equation is no better than what you would expect by chance.

As you are doing a multiple regression, there is also a null hypothesis for each X variable, meaning that adding that X variable to the multiple regression does not improve the fit of the multiple regression equation any more than expected by chance.

6.5.6 Estimating and Making Inferences about the Slope

The purpose of a multiple regression is to find an equation that best predicts the Y variable as a linear function of the X variables.

You use multiple regression when you have three or more measurement variables. One of the measurement variables is the dependent (Y) variable. The rest of the variables are the independent (X) variables. The purpose of a multiple regression is to find an equation that best predicts the Y variable as a linear function of the X variables.

6.5.7 How It Works?

The basic idea is that an equation is found like this:

$$Y_{\text{exp}} = a + b_1X_1 + b_2X_2 + b_3X_3 + \dots$$

The Y_{exp} is the expected value of Y for a given set of X values. b_1 is the estimated slope of a regression of Y on X_1 , if all of the other X variables could be kept constant. This concept applies similarly for b_2 , b_3 , etc. a is the intercept. Values of b_1 , etc. (the “partial regression coefficients”) and the intercept are found so that they minimize the squared deviations between the expected and observed values of Y .

How well the equation fits the data is expressed by R^2 , the “coefficient of multiple determination”. This can range from 0 (for no relationship between the X and Y variables) to 1 (for a perfect fit, *i.e.*, no difference between the observed and expected Y values). The p -value is a function of the RR^2 , the number of observations, and the number of X variables.

6.5.8 Importance of Slope (Partial Regression Coefficients)

When the purpose of multiple regression is prediction, the important result is an equation containing partial regression coefficients (slopes). If you had the partial regression coefficients and measured the X variables, you could plug them into the equation and predict the corresponding value of Y . The magnitude of the partial regression coefficient depends on the unit used for each variable. It does not tell you anything about the relative importance of each variable.

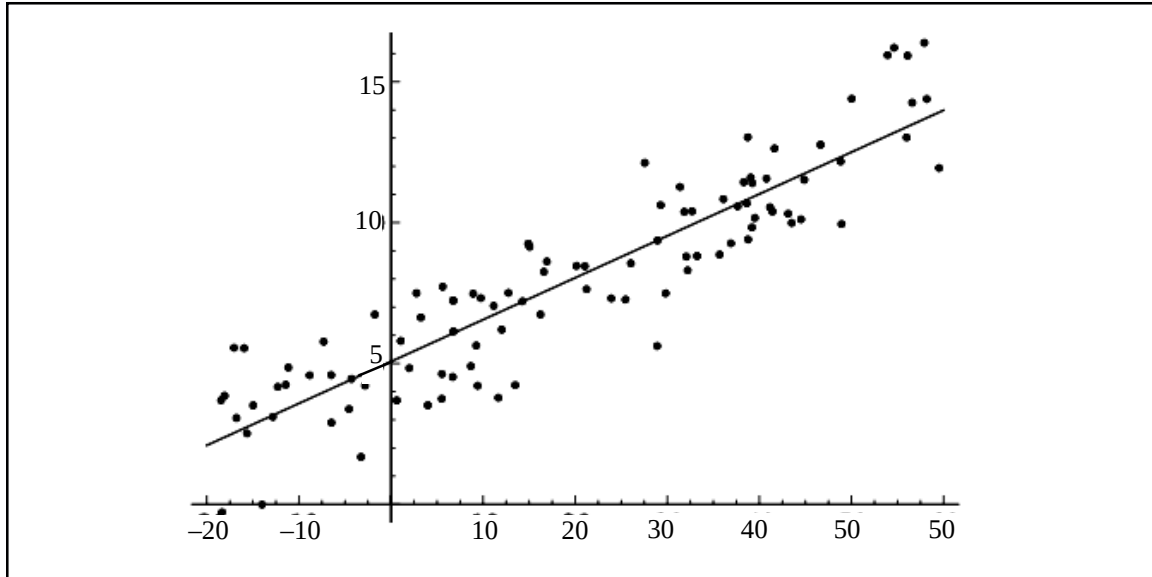
When the purpose of multiple regressions understands functional relationships, the important result is an equation containing standard partial regression coefficients, like this:

$$Y'_{\text{exp}} = a + b'_1x'_1 + b'_2x'_2 + b'_3x'_3 + \dots$$

where, b'_1 is the standard partial regression coefficient of y on X_1 . It is the number of standard deviations that Y would change for every one standard deviation change in X_1 , if all the other X variables could be kept constant. The magnitude of the standard partial regression coefficients tells you something about the relative importance of different variables; X variables with bigger standard partial regression coefficients have a stronger relationship with the Y variable.

6.5.9 Evaluating Model Utility

The results of multiple regressions should be viewed with caution.



6.5.10 Using Multiple Regression for Prediction

Standard multiple regression is the same idea as simple linear regression, except now, we have several independent variables predicting the dependent variable. Imagine that we wanted to predict a person's height from the gender of the person and from the weight. We would use standard multiple regression in which gender and weight would be the independent variables and height would be the dependent variable. The resulting output would tell us a number of things. First, it would tell us how much of the variance in height is accounted for by the joint predictive power of knowing a person's weight and gender. This value is denoted by R^2 . The output would also tell us if the model allows the prediction of a person's height at a rate better than chance. This is denoted by the significance level of the model. Within the social sciences, a significance level of 0.05 is often considered the standard for what is acceptable. Therefore, in our example, if the statistic is 0.05 (or less), then the model is considered significant. In other words, there is only a 5 in a 100 chance (or less) that there really is not a relationship between height, weight and gender. If the significance level is between 0.05 and 0.10, then the model is considered marginal. In other words, the model is fairly good at predicting a person's height, but there is between a 5-10% probability that there really is not a relationship between height, weight and gender.

In addition to telling us the predictive value of the overall model, standard multiple regression tells us how well each independent variable predicts the dependent variable, controlling for each of the other independent variables. In our example, the regression analysis would tell us how well weight predicts a person's height, controlling for gender, as well as how well gender predicts a person's height, controlling for weight.

To see if weight is a “significant” predictor of height, we would look at the significance level associated with weight. Again, significance levels of 0.05 or lower would be considered significant, and significance levels between 0.05 and 0.10 would be considered marginal. Once we have determined that weight is a significant predictor of height, we would want to more closely examine the relationship between the two variables. In other words, is the relationship positive or negative? In this example, we would expect that there would be a positive relationship. In other words, we would expect that the greater a person’s weight, the greater the height. (A negative relationship is present in the case in which the greater a person’s weight, the shorter the height). We can determine the direction of the relationship between weight and height by looking at the regression coefficient associated with weight.

A similar procedure shows us how well gender predicts height. As with weight, we would check to see if gender is a significant predictor of height, controlling for weight. The difference comes when determining the exact nature of the relationship between gender and height, *i.e.*, it does not make sense to talk about the effect on height as gender increases or decreases, since gender is not a continuous variable.

Conclusion

As mentioned, the significance levels given for each independent variable indicate whether that particular independent variable is a significant predictor of the dependent variable, over and above the other independent variables. Because of this, an independent variable that is a significant predictor of a dependent variable in simple linear regression may not be significant in multiple regression (*i.e.*, when other independent variables are added into the equation). This could happen because the covariance that the first independent variable shares with the dependent variable could overlap with the covariance that is shared between the second independent variable and the dependent variable. Consequently, the first independent variable is no longer uniquely predictive and would not be considered significant in multiple regression. Because of this, it is possible to get a highly significant R^2 , but have none of the independent variables be significant.

Multiple Regression: This shows data points and their linear regression. Multiple regression is the same idea as single regression, except we deal with more than one independent variables predicting the dependent variable.

6.6 FACTOR ANALYSIS

6.6.1 Introduction

As for principal components analysis, factor analysis is a multivariate method used for data reduction purposes. Again, the basic idea is to represent a set of variables by a smaller number of variables. In this case, they are called factors. These factor can be thought of as underlying constructs that cannot be measured by a single variable (*e.g.*, happiness).

6.6.2 Assumptions

Factor analysis is designed for interval data, although it can also be used for ordinal data (e.g., scores assigned to Likert scales). The variables used in factor analysis should be linearly related to each other. This can be checked by looking at scatterplots of pairs of variables. Obviously, the variables must also be at least moderately correlated to each other. Otherwise, the number of factors will be almost the same as the number of original variables, which means that carrying out a factor analysis would be pointless.

6.6.3 Steps in Factor Analysis

The factor analysis model can be written algebraically as follows. If you have p variables $X_1, X_2, \dots, X_p$ measured on a sample of n subjects, then variable i can be written as a linear combination of m factors $F_1, F_2, \dots, F_m$ where, as explained above, $m < p$. Thus,

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \dots + a_{im}F_m + e_i$$

where, the a_i s are the factor loadings (or scores) for variable i and e_i is the part of variable X_i that cannot be 'explained' by the factors.

There are three main steps in a factor analysis:

1. Calculate initial factor loadings

This can be done in a number of different ways; the two most common methods are described very briefly below:

- **Principal Component Method:** As the name suggests, this method uses the method used to carry out a principal components analysis. However, the factors obtained will not actually be the principal components (although the loadings for the k^{th} factor will be proportional to the coefficients of the k^{th} principal component).
- **Principal Axis Factoring:** This is a method which tries to find the lowest number of factors which can account for the variability in the original variables that is associated with these factors (this is in contrast to the principal components method which looks for a set of factors which can account for the total variability in the original variables).

These two methods will tend to give similar results if the variables are quite highly correlated and/or the number of original variables is quite high. Whichever method is used, the resulting factors at this stage will be uncorrelated.

2. Factor Rotation

Once the initial factor loadings have been calculated, the factors are rotated. This is done to find factors that are easier to interpret. If there are 'clusters' (groups) of variables — i.e., subgroups of variables that are strongly interrelated — then the rotation is done to try to make variables within a subgroup score as highly (positively or negatively) as possible on one particular factor while, at the same time, ensuring that the loadings for these variables on the remaining factors are as low as possible. In other words, the object of the rotation is to try to ensure that all variables have high loadings only on one factor.

There are two types of rotation method, **orthogonal** and **oblique** rotation. In orthogonal rotation, the rotated factors will remain uncorrelated whereas in oblique rotation, the resulting factors will be correlated. There are a number of different methods of rotation of each type. The most common orthogonal method is called varimax rotation; this is the method that many books will recommend.

3. Calculation of Factor Scores

When calculating the final factor scores (the values of the m factors, $F_1, F_2, \dots, F_m$ for each observation), a decision needs to be made as to how many factors to include. This is usually done using one of the following methods:

- Choose m such that the factors account for a particular percentage (*e.g.*, 75%) of the total variability in the original variables.
- Choose m to be equal to the number of eigen values over 1 (if using the correlation matrix). [Different criteria must be used if using the covariance matrix.]
- Use the screen plot of the eigen values. This will indicate whether there is an obvious cut-off between large and small eigen values.

In some statistical packages (*e.g.*, SPSS), this choice is actually made at the outset. The second method, choosing eigen values over 1, is probably the most common one.

The final factor scores are usually calculated using a regression-based approach.

6.6.4 Carrying Out Factor Analysis in SPSS

- Analyze
- Data Reduction
- Factor
- Select the variables to want the factor analysis to be based on and move them in to the **Variable(s)** box.
- In the Descriptives window, you should select **KMO and Bartlett's Test of Sphericity**. KMO is a statistic which tells whether to have sufficient items for each factor. It should be over 0.7. Bartlett's test is used to check that the original variables are sufficiently correlated. This test should come out significant ($p < 0.05$) — if not, factor analysis will not be appropriate. Click on **Continue**.
- In the **Extraction** window, you can select the extraction method you want to use (*e.g.*, principal components, etc.). Under **Analyze**, ensure that **Correlation Matrix** is selected (this is the default). The default is also to extract eigen values over 1 but if you want to extract a specific number of factors, you can specify this. Click on **Continue**.
- In the **Rotation** window, you can select your rotation method (as mentioned above, **Varimax** is the most common). You can also ask SPSS to display the rotated solution. Once to have selected this, click on **Continue**.

- In the **Scores** window, you can specify whether you want SPSS to save factor scores for each observation (this will save them as new variables in the dataset). Under **Method**, choose Regression. You can also ask SPSS to display the factor score coefficients (the a_i s). Click on **Continue**.
- **OK**

6.6.5 Factor Analysis in R

Factor analysis (FA) or **exploratory factor analysis** is another technique to reduce the number of variables to a smaller set of factors. FA identifies the relationships among a set of variables and narrows it down to a smaller set.

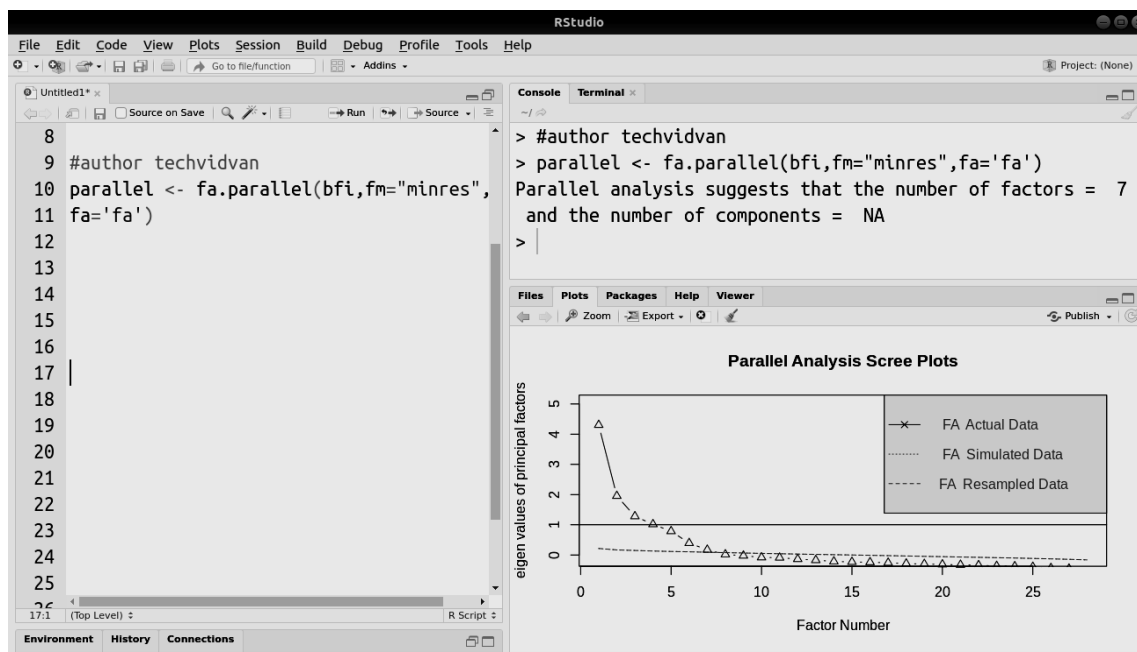
We will be using the **bfi** dataset, which is a built-in dataset provided in R. It comprises 25 different personality factors. We will require the **psych** and the **GPArotation** packages. So, install and load them into the library. Then we will use the following command to find how many factors we need.

Code:

```
parallel <- fa.parallel(bfi, fm="minres", fa='fa')
```

Output:

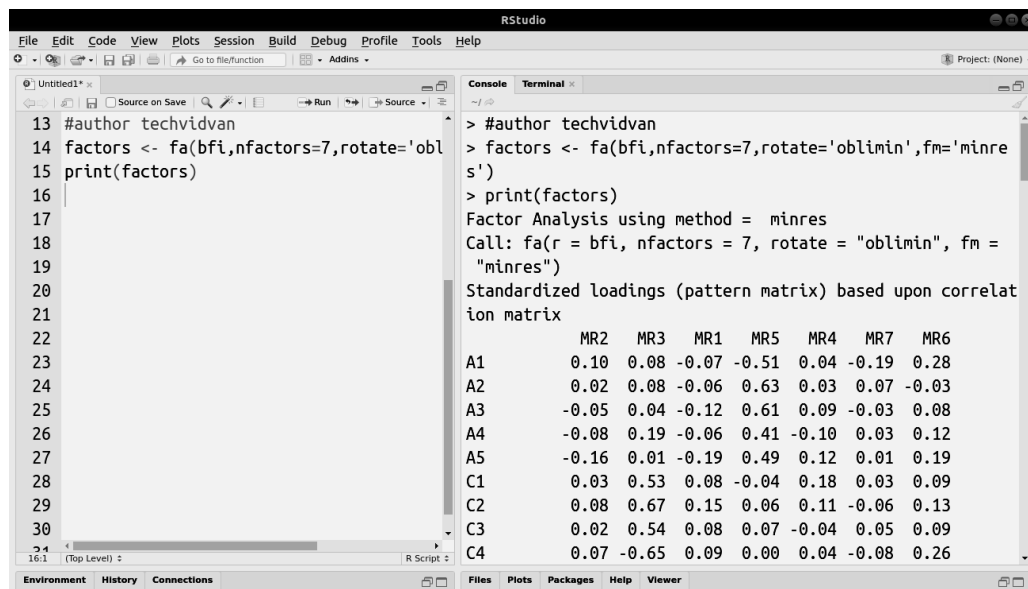
Parallel analysis suggests that the number of factors = 7 and the number of components = NA.



Now that we know how many factors we need, we can perform the factor analysis using the **fa()** function.

Code:

```
factors <- fa(bfi,nfactors=7,rotate='oblimin',fm='minres')
print(factors)
```

Output:


```
> #author techvidvan
> factors <- fa(bfi,nfactors=7,rotate='oblimin',fm='minres')
> print(factors)
Factor Analysis using method = minres
Call: fa(r = bfi, nfactors = 7, rotate = "oblimin", fm = "minres")
Standardized loadings (pattern matrix) based upon correlation matrix
```

	MR2	MR3	MR1	MR5	MR4	MR7	MR6
A1	0.10	0.08	-0.07	-0.51	0.04	-0.19	0.28
A2	0.02	0.08	-0.06	0.63	0.03	0.07	-0.03
A3	-0.05	0.04	-0.12	0.61	0.09	-0.03	0.08
A4	-0.08	0.19	-0.06	0.41	-0.10	0.03	0.12
A5	-0.16	0.01	-0.19	0.49	0.12	0.01	0.19
C1	0.03	0.53	0.08	-0.04	0.18	0.03	0.09
C2	0.08	0.67	0.15	0.06	0.11	-0.06	0.13
C3	0.02	0.54	0.08	0.07	-0.04	0.05	0.09
C4	0.07	-0.65	0.09	0.00	0.04	-0.08	0.26

6.6.6 Uses of Factor Analysis

Factor analysis is used to uncover the latent structure of a set of variables. It reduces attribute space from a large number of variables to a smaller number of factors and as such is a non-dependent procedure.

Factor analysis could be used for any of the following purpose:

1. **To reduce a large number of variable to a smaller number of factors for modeling purposes**, where the large number of variables precludes modeling all the measures, individually. As such factor analysis is integrated in structural equation modeling, helping create the latent variables modeled by SEM (structure equation model).
2. **To select a subset of variables from a large set** based on which original variable have the highest correlations with the principal component factors.
3. **To create a set of factors to be treated as uncorrelated variable** as one approach to handling multicollinearity regression.

6.6.7 Assumptions of Factor Analysis

Factor analysis makes use of several assumptions in order to produce the outcomes:

- There will not be any outliers in the data.
- The sample size will be greater than the size of the factor.

- Since the method is interdependent, there will be no perfect multicollinearity between any of the variables.
- When in a sequence of random variables, all the variables have the same finite variance, known as being homoscedastic. Since factor analysis works as a linear function, it will not need homoscedasticity between variables.

There is the assumption of linearity. This means that even non-linear variables can be used, but once transferred, they become linear variables.

There is also the assumption of interval data

Factor analysis is a part of the **Multiple General Linear Hypothesis (MGLH)**, family of procedures and makes many of the same assumptions as multiple regressions, linear relationship interval or near-interval data, untruncated data, proper specification (relevant variable included extraneous are excluded), lack of high multicollinearity and multivariable normality for purpose of significant testing.

Factor analysis generates a table on which the rows and the observed row indicator variables and the columns are the factor or latent variables which explain as much of the variable in those variables as possible. The cells in this table are factor loadings and the meaning of the factors must be induced from seeing which variables are most heavily loaded on which factors. This inferential process can be fraught with difficulty as diverse researchers impute different tables.

6.6.8 Methods of Factor Analysis

There are several different types of factor analysis:

1. Principal component method
2. Principal axes method
3. Summation method
4. Centroid method

6.6.8.1 Principal Component Method

In this method, factors are selected one at a time such that each factor best fits the data. The first fraction is created such that it represents the most highly correlated set of variables.

Each subsequent selected factor explains less variance than its predecessor.

This procedure is continued till all the factors are selected. All the factors selected explain the largest amount of residual variance in the entire set of standardized response scores.

6.6.8.2 Centroid Method

1. Obtain the correlation matrix.
2. Obtain grand matrix sum, row sum and column sum.
3. Calculate $N = 1 \frac{1}{\sqrt{\text{Grand Total}}}$

4. Multiply each column sum with N , which gives the first factor loading.
5. To find the second factor loading, find the cross product matrix of factor 1 by testing first factor loading horizontally and vertically, and then multiplying corresponding rows and columns.
6. Find the first factor residual matrix given as:

$$\begin{aligned}\text{Residuals} &= \text{Total Variations} - \text{Explained variations} \\ &= r_{ij} - l_{ij}\end{aligned}$$

7. **Reflection:** Reflection means that each test vector retains its length but extends in opposite directions. The major purpose of reflection is to get a reflected cost matrix having the highest possible grand total. This step is taken due to the reason that some of the factors loading is with grand total. The method of reflection is by trial and error. This can be done by changing the signs of the variables from positive to negative and negative to positive column-wise and row-wise. The outcome we get is a reflective residual correlation matrix.
8. Repeat steps from (3) to (7).

6.6.9 Assumptions of Factor Analysis Model

1. Measurement error has constant variance and is on average zero, *i.e.*,
 $\text{var}(e) = s^2$
 $E[e_i] = 0$
2. No association between the factor and measurement error, $\text{cov}(F, e_j) = 0$
3. No association between errors, $\text{cov}(e_j, e_k) = 0$
4. Local (*i.e.*, conditional independence): Given factor, observed variables are independent of one another, $\text{cov}(X_j, X_k | F) = 0$

6.6.10 What are the Prerequisites before Running an Exploratory Factor Analysis?

Are there any other things I have to take into consideration other than:

- sample size
- recoding items (if necessary)
- Gaussian distribution
- reliability/commonality
- items per factor

I would be immensely grateful for further aspects or refinements of my list so far.

6.6.11 Steps in Exploratory Factor Analysis

1. Collect data: choose relevant variables.
2. Extract initial factors (via principal component).

3. Choose number of factors to retain.
4. Choose estimation method, *i.e.*, estimate model.
5. Rotate and interpret.
6. (a) Decide on changes need to be made (*e.g.*, drop items include items)
(b) Repeat steps from (4) and (5).
7. Construct scales and use on further analysis.

6.6.12 Methods for Finding Number of Factors to be Extracted

1. **Thumb Rule:** All the interrelated factors must explain at least as much as variances as an average variable. Check, if a variable is under a factor, then the percentage of variable explaining variance should be less than the percentage of factor explaining.
2. **Eigen Value Index:** When the eigen value of a factor is less than 1, it explains less variance. Then the variables included in the factor itself. Such a factor should not be considered.

3. **Fruckter Formula:**

$$\text{Number of factors} = \frac{(2n-1) - \sqrt{8n+1}}{2}$$

where, n is the number of variables included in the study.

4. **Residual Correlation Matrix Method:** In this method, the residual correlation is observed and if it is soon that most of the correlation coefficient in this matrix are zero, then the extraction of the factor can be determinate.
5. **Screen Plot Test:** This method is to decide about the number of factors to be retained from the extracted factors. The test determines which of the extracted factors are actually contributing variance and does not measure random errors. The number of factors is plotted against the proportion of variance. It extracts in the order of the extracted factors.

- **Standardization of Responses:**

$$\hat{X}_i = \frac{X_i - X}{\sigma}$$

where, X_i is a value corresponding to a response and σ is the variance.

- **Factor Loading:** It is the correlation between a factor and a variable. It helps interpret the meaning of a factor by indicating how well the factor fits the standardised responses to a variable. **The greater the value of factor loading, the better is the fit of the factor** to the data from the concerned statement. All variables load on all factors, but they load highly on some specific factors. Range is ± 1 .

There may be some variables which may be loading highly to more than two factors, decide in which factors which variables are to be kept.

- **Eigen Values:** It is the measurement of the amount of variants explained by a factor. A factor eigen value is the sum of the square of its factor loading.
- **Communality:** It indicates the proportion of variance in the responses to the statement which is explained by the identified factors.
- **Percentage of Variance:**

$$= \frac{\text{Eigen value of the factor}}{\text{Sum of all eigen values}} \times 100$$

If any variable is not combined in any of the groups, then it can be left or can be considered as another factor. To remove variable which is totally different use rotation, *i.e.*, we are basically changing its direction.

6.6.13 Rotation of Factors

For the purpose of simplifying the interpretation of obtained factors and to increase the number of high and low positive loadings in the columns of a factor, factor rotation is used. There are two methods for this:

1. **Orthogonal Rotation/Variance Rotation:** Here, the factors are rotated such that the original factors as well as rotated factors are orthogonal. The line between the factors axis remains 90.
2. **Promax Rotation:** The factors are rotated such that the line between original and rotated factors is more than or less than 90.

6.6.14 Advantages of Factor Analysis

1. Both objective and subjective attributes can be used.
2. It can be used to identify the hidden dimensions or constraints which may or may not be apparent from direct analysis.
3. It is not extremely difficult to do and at the same time its inexpensive and gives accurate results.
4. There is flexibility in naming and using dimensions.

6.6.15 Disadvantages of Factor Analysis

1. The usefulness depends on the researcher's ability to develop a complete and accurate set of product attributes. If important attributes are missed, the value of procedure is reduced accordingly.
2. Naming of the factors can be difficult. Multiple attributes can be highly correlated with no apparent reasons.
3. If the observed variables are completely unrelated, the factor analysis is unable to produce meaningful pattern.

4. It is not possible to know factors actually represents. Only theory can help inform the researchers on this.

What is the Sigmoid Function?

It is a mathematical function having a characteristic that can take any real value and map it to between 0 to 1 shaped like the letter “S”. The sigmoid function also called a logistic function.

$$Y = 1 / 1 + e^{-z}$$

 analyticSteps
www.analyticsteps.com

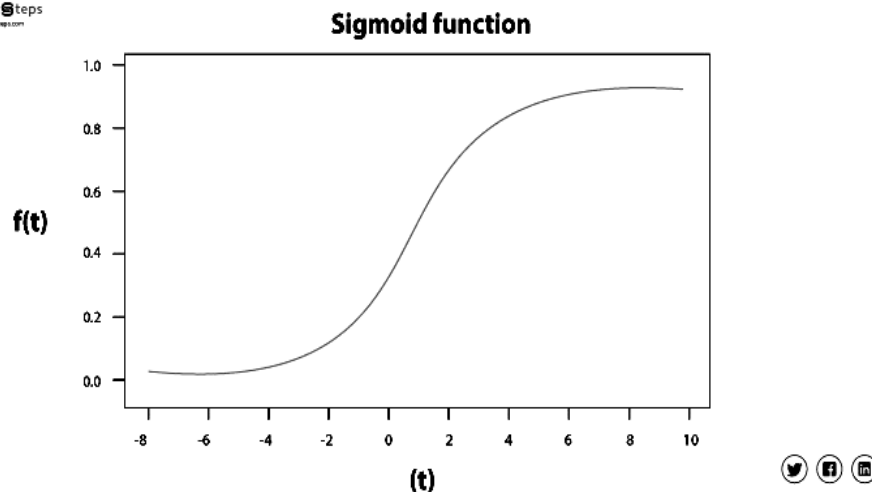


Fig. 6.3: Sigmoid Function

So, if the value of z goes to positive infinity, then the predicted value of y will become 1 and if it goes to negative infinity, then the predicted value of y will become 0. And if the outcome of the sigmoid function is more than 0.5, then we classify that label as class 1 or positive class, and if it is less than 0.5, then we can classify it to negative class or label as class 0.

6.7 CLUSTER ANALYSIS

6.7.1 Meaning of Cluster Analysis

- The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering.
- A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters.
- A cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression.
- Cluster analysis tools based on k-means, k-medoids, and several methods have also been built into many statistical analysis software packages or systems, such as S-Plus, SPSS, and SAS.

6.7.2 Applications of Cluster Analysis

- Cluster analysis has been widely used in numerous applications, including market research, pattern recognition, data analysis, and image processing.
- In business, clustering can help marketers discover distinct groups in their customer bases and characterize customer group's based on purchasing patterns.
- In biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionality, and gain insight into structures inherent in populations.
- Clustering may also help in the identification of areas of similar land use in an earth observation database and in the identification of groups of houses in a city according to house type, value, and geographic location, as well as the identification of groups of automobile insurance policy holders with a high average claim cost.
- Clustering is also called data segmentation in some applications because clustering partitions large datasets into groups according to their similarity.
- Clustering can also be used for outlier detection. Applications of outlier detection include the detection of credit card fraud and the monitoring of criminal activities in electronic commerce.

6.7.3 Typical Requirements of Clustering in Data Mining

- **Scalability:** Many clustering algorithms work well on small datasets containing fewer than several hundred data objects. However, a large database may contain millions of objects. Clustering on a sample of a given large dataset may lead to biased results.
Highly scalable clustering algorithms are needed.
- **Ability to deal with different types of attributes:** Many algorithms are designed to cluster interval-based (numerical) data. However, applications may require clustering other types of data, such as binary, categorical (nominal), and ordinal data, or mixtures of these data types.
- **Discovery of clusters with arbitrary shape:** Many clustering algorithms determine clusters based on Euclidean or Manhattan distance measures. Algorithms based on such distance measures tend to find spherical clusters with similar size and density.
However, a cluster could be of any shape. It is important to develop algorithms that can detect clusters of arbitrary shape.
- **Minimal requirements for domain knowledge to determine input parameters:** Many clustering algorithms require users to input certain parameters in cluster analysis (such as the number of desired clusters). The clustering results can be quite sensitive to input parameters. Parameters are often difficult to determine, especially for datasets containing high-dimensional objects. This not only burdens users, but it also makes the quality of clustering difficult to control.

- **Ability to deal with noisy data:** Most real-world databases contain outliers or missing, unknown, or erroneous data. Some clustering algorithms are sensitive to such data and may lead to clusters of poor quality.
- **Incremental clustering *r* and insensitivity to the order of input records:** Some clustering algorithms cannot incorporate newly inserted data (*i.e.*, database updates) into existing clustering structures and instead, must determine a new clustering from scratch. Some clustering algorithms are sensitive to the order of input data.

That is, given a set of data objects, such an algorithm may return dramatically different clusterings depending on the order of presentation of the input objects.

It is important to develop incremental clustering algorithms and algorithms that are insensitive to the order of input.

- **High-dimensionality:** A database or a data warehouse can contain several dimensions or attributes. Many clustering algorithms are good at handling low-dimensional data, involving only two to three dimensions. Human eyes are good at judging the quality of clustering for up to three dimensions. Finding clusters of data objects in high-dimensional space is challenging, especially considering that such data can be sparse and highly skewed.
- **Constraint-based clustering:** Real-world applications may need to perform clustering under various kinds of constraints. Suppose that the job is to choose the locations for a given number of new automatic banking machines (ATMs) in a city. To decide upon this, you may cluster households while considering constraints such as the city's rivers and highway networks, and the type and number of customers per cluster. A challenging task is to find groups of data with good clustering behaviour that satisfy specified constraints.
- **Interpretability and usability:** Users expect clustering results to be interpretable, comprehensible and usable, *i.e.*, clustering may need to be tied to specific semantic interpretations and applications. It is important to study how an application goal may influence the selection of clustering features and methods.

6.7.4 Assumptions of Cluster Analysis

There are always two assumptions in it.

- It is assumed that the sample is a representative of the population.
- It is assumed that the variables are not correlated. Even if variables are correlated remove correlated variables or use distance measures that compensate for the correlation.

Conducting Cluster Analysis using RATELL

R is the most common platform for predictive analysis. Rattle library is an extension of *R* which takes the predictive analysis to another level.

This blog is focused towards people who have some experience in R.

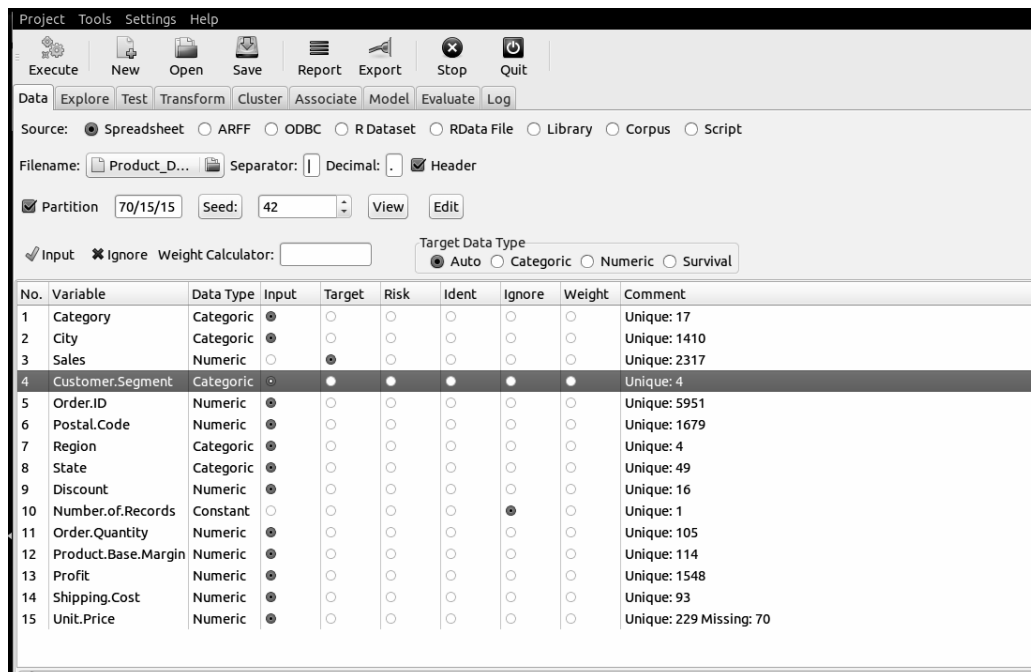
Rattle is the library provided for R language that is used for data mining process, where you can apply different types of clustering, classification types algorithm. Here, you can transform the data as per requirements, you can evaluate the relation between the datasets like correlation, distributions, etc. It provides a very fine-tuned user interface to perform a large amount of data mining operations.

Steps to Install/Use the Rattle:

- `install.packages("rattle");`
- `rattle();`

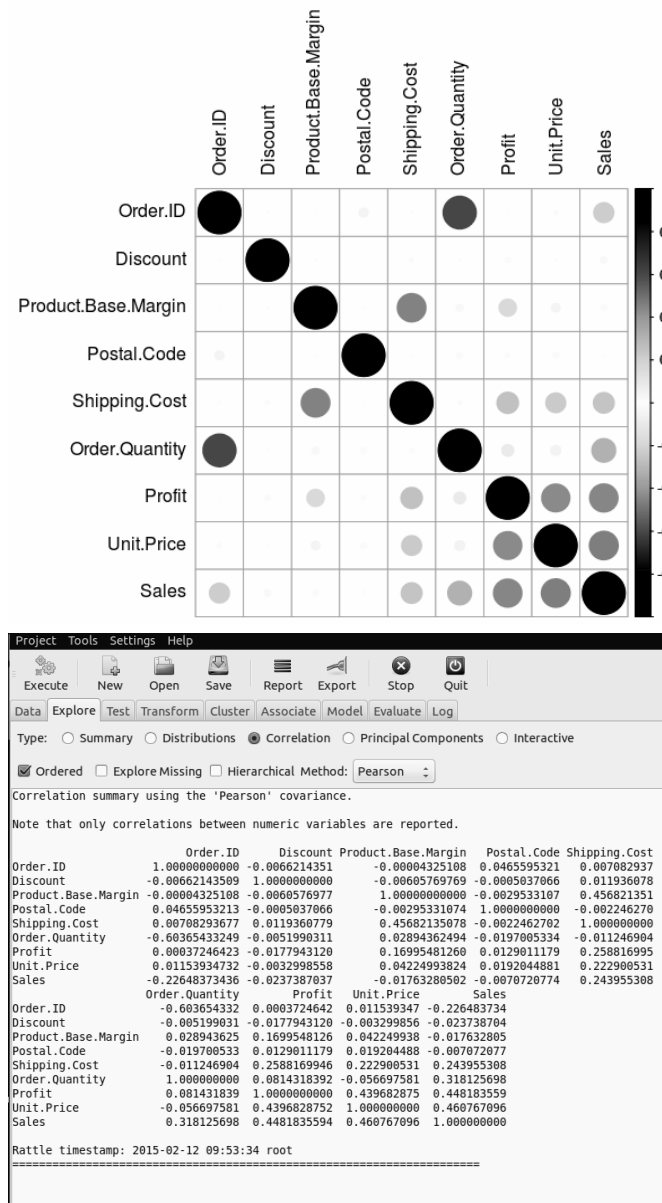
Use Case: We have datasets of sales which contains different types of field like price, quantity, shipping cost, item name, discount and many more. We can apply different types of algorithms as per our target variable. Let's go through Rattle UI to perform different types of operations.

- **Datasets:** We need to upload the datasets using Rattle. We can upload data in different format, partitioning the data like training, test datasets, etc.



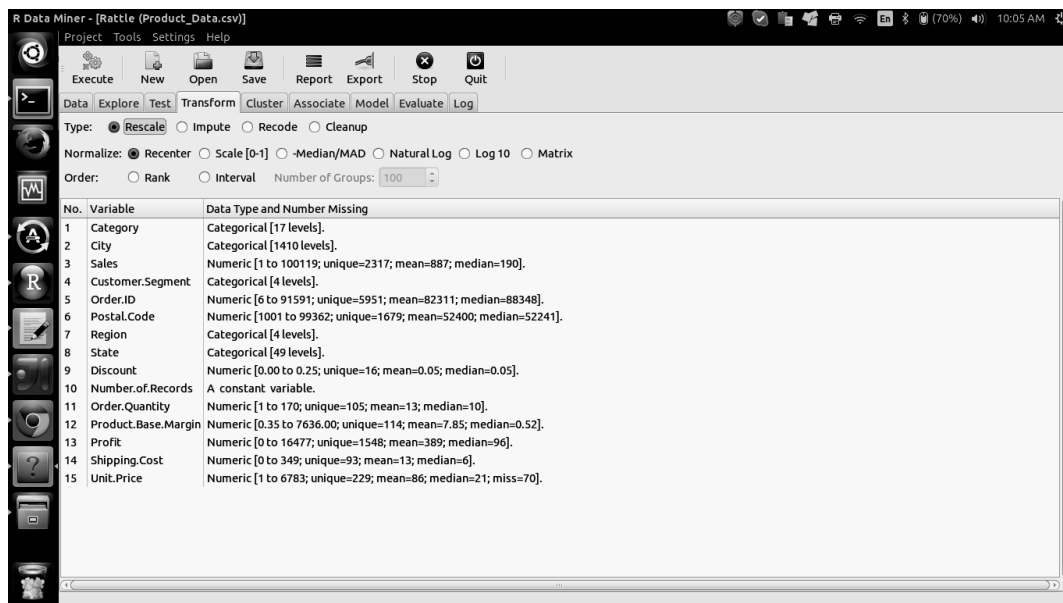
Input type variables, target type and their data type are automatically detected by Rattle. You can adjust it as per the requirements.

- **Explore:** You can explore the data like summarized datasets, check correlations, apply principle components, etc. Here, major one, you check the correlation of our datasets to remove the less importance variable from the input datasets.

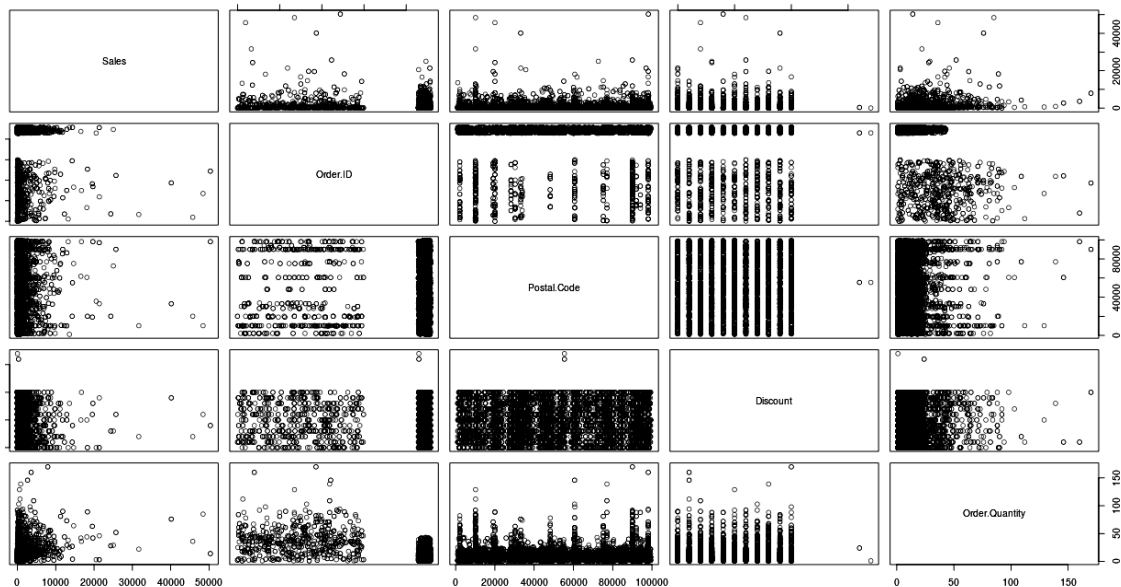


Here, you can see the correlation between different types of variables, to can adjust the input variable as checking by correlations.

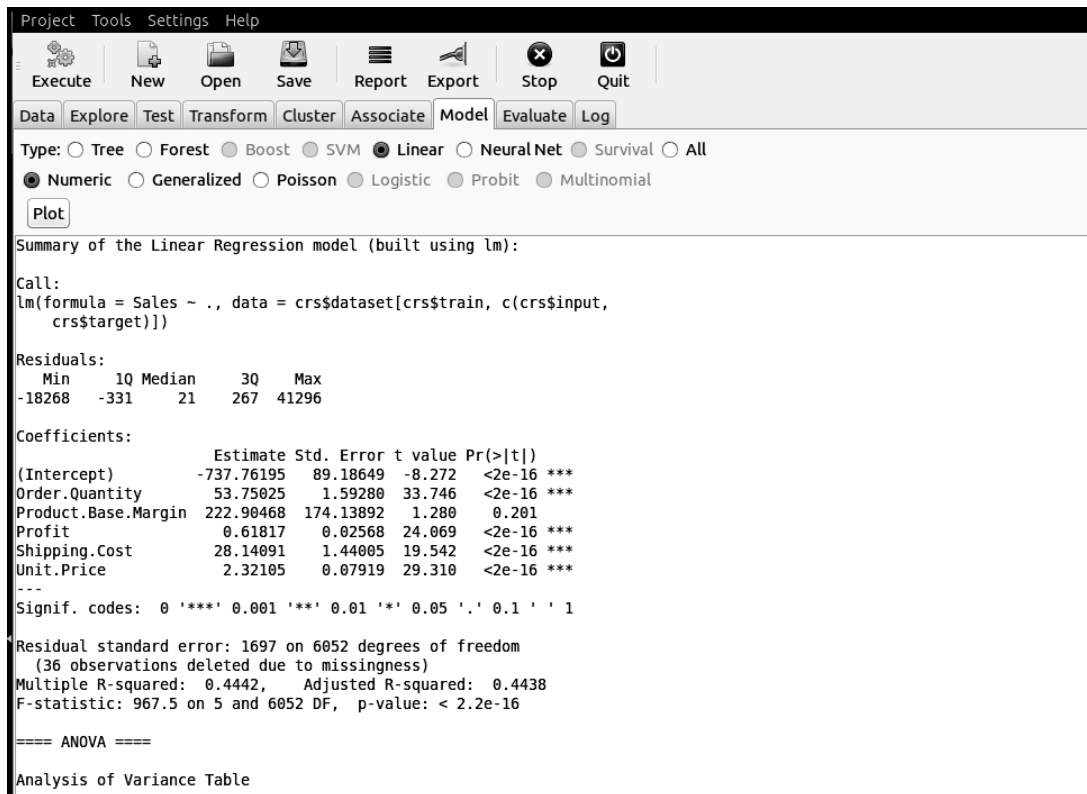
- Transform:** Sometimes, we need to transform the data like remove the observations that having missing data, to clean-up data, rescale the data using different types of filters, etc. All these features are provided in Rattle UI under Transform tab.



- **Clustering:** Apply different types of cluster like k -means, Bi Cluster, plot clusters, etc. and check the stats of the data using cluster algorithm.



- **Models:** Use different types of model like tree, random forest, linear regression and neural networks. Choose the best one as per the use of case. Our use case is to forecast the sales, so for this linear regression may provide the better results.



The screenshot shows the Rattle software interface. The 'Model' tab is selected, displaying a summary of a Linear Regression model built using the `lm()` function. The interface includes a menu bar (Project, Tools, Settings, Help) and a toolbar with icons for Execute, New, Open, Save, Report, Export, Stop, and Quit. Below the toolbar are tabs for Data, Explore, Test, Transform, Cluster, Associate, Model, Evaluate, and Log. The 'Model' tab shows the following details:

Type: ☐ Tree ☐ Forest ☐ Boost ☐ SVM ☒ Linear ☐ Neural Net ☐ Survival ☐ All

Numeric: ☒ Generalized ☐ Poisson ☐ Logistic ☐ Probit ☐ Multinomial

Plot button

Summary of the Linear Regression model (built using lm):

Call:
`lm(formula = Sales ~ ., data = crs$dataset[crs$train, c(crs$input, crs$target)])`

Residuals:

Min	1Q	Median	3Q	Max
-18268	-331	21	267	41296

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-737.76195	89.18649	-8.272	<2e-16 ***
Order.Quantity	53.75025	1.59280	33.746	<2e-16 ***
Product.Base.Margin	222.90468	174.13892	1.280	0.201
Profit	0.61817	0.02568	24.069	<2e-16 ***
Shipping.Cost	28.14091	1.44005	19.542	<2e-16 ***
Unit.Price	2.32105	0.07919	29.310	<2e-16 ***

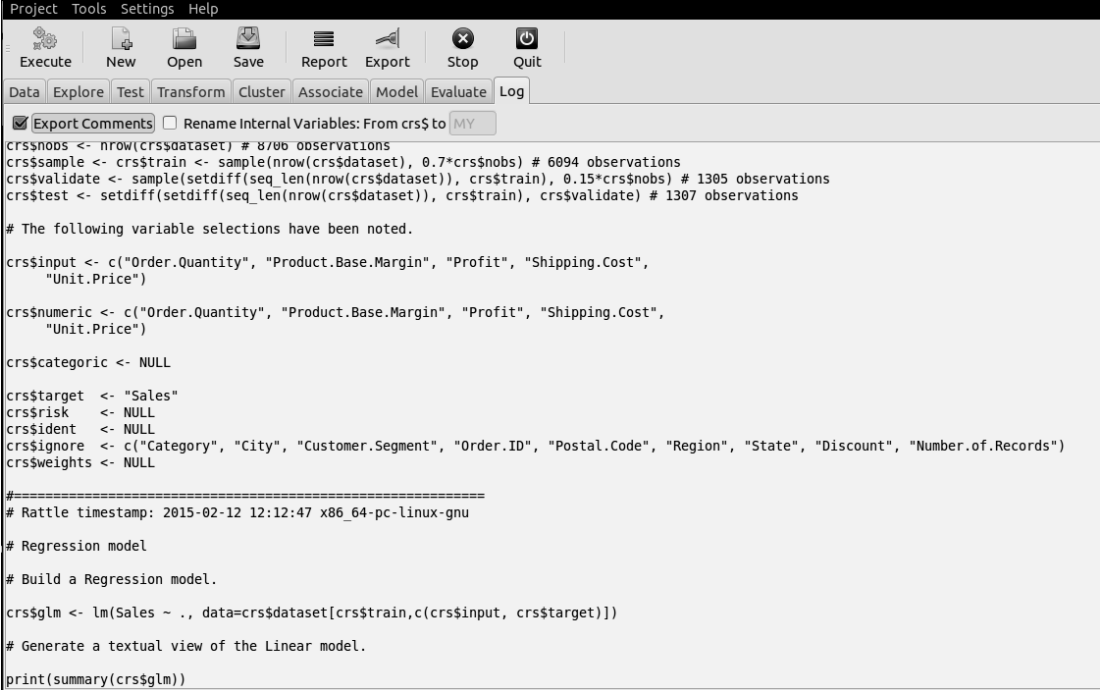
 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1697 on 6052 degrees of freedom
 (36 observations deleted due to missingness)
 Multiple R-squared: 0.4442, Adjusted R-squared: 0.4438
 F-statistic: 967.5 on 5 and 6052 DF, p-value: < 2.2e-16

==== ANOVA ====

Analysis of Variance Table

- **Logs:** Whatever actions you are performing in Rattle, it will automatically generate the source code of the action in Log tab which you can use later once you have final data, model and then use same code for the production purpose.



```

Project Tools Settings Help
Execute New Open Save Report Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
☒ Export Comments ☐ Rename Internal Variables: From crs$ to MY
crs$nobs <- nrow(crs$dataset) # 8706 observations
crs$sample <- crs$train <- sample(nrow(crs$dataset), 0.7*crs$nobs) # 6094 observations
crs$validate <- sample(setdiff(seq_len(nrow(crs$dataset)), crs$train), 0.15*crs$nobs) # 1305 observations
crs$test <- setdiff(setdiff(seq_len(nrow(crs$dataset)), crs$train), crs$validate) # 1307 observations

# The following variable selections have been noted.

crs$input <- c("Order.Quantity", "Product.Base.Margin", "Profit", "Shipping.Cost",
              "Unit.Price")

crs$numeric <- c("Order.Quantity", "Product.Base.Margin", "Profit", "Shipping.Cost",
                "Unit.Price")

crs$categoric <- NULL

crs$target <- "Sales"
crs$risk <- NULL
crs$ident <- NULL
crs$ignore <- c("Category", "City", "Customer.Segment", "Order.ID", "Postal.Code", "Region", "State", "Discount", "Number.of.Records")
crs$weights <- NULL

=====
# Rattle timestamp: 2015-02-12 12:12:47 x86_64-pc-linux-gnu
# Regression model
# Build a Regression model.

crs$glm <- lm(Sales ~ ., data=crs$dataset[crs$train,c(crs$input, crs$target)])

# Generate a textual view of the Linear model.

print(summary(crs$glm))

```

You can see how powerful Rattle is. You can perform various types of data mining operation. You do not need to worry about the coding part, major part is what is the use-case, what operation is needed to be performed, which model is best suited for you.

6.7.5 Major Clustering Methods

- Partitioning Methods
- Hierarchical Methods
- Density-based Methods
- Grid-based Methods
- Model-based Methods

6.7.5.1 Partitioning Methods

A partitioning method constructs k partitions of the data, where each partition represents a cluster and $k \leq n$, i.e., it classifies the data into k groups, which together satisfy the following requirements:

- Each group must contain at least one object, and
- Each object must be long to exactly one group.

A partitioning method creates an initial partitioning. It then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another.

The general criterion of a good partitioning is that objects in the same cluster are close or related to each other, whereas objects of different clusters are far apart or very different.

Classical Partitioning Methods

The most well-known and commonly used partitioning methods are:

- The k -means Method
- k -medoids Method

1. The k -means Method (Centroid-based Technique)

The k -means algorithm takes the input parameter, k , and partitions a set of n objects into k clusters so that the resulting intra-cluster similarity is high but the inter-cluster similarity is low. Cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed as the cluster's centroid or center of gravity.

The k -means algorithm proceeds as follows:

- First, it randomly selects k of the objects, each of which initially represents a cluster mean or center.
- For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean.
- It then computes the new mean for each cluster.
- This process iterates until the criterion function converges.

Typically, the square-error criterion is used, defined as,

$$E = \sum_{i=1}^k \sum_{p \in C_i} |p - m_i|^2$$

where, E = the sum of the square error for all objects in the dataset

p = the point in space representing a given object

m_i = the mean of cluster C_i

The k -means Partitioning Algorithm

Input:

k : the number of clusters,

D : a dataset containing n objects

Output: A set of k clusters.

Method:

1. Arbitrarily choose k objects D as the initial cluster centers.
2. Repeat.

3. (Re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster.
4. Update the cluster means, *i.e.*, calculate the mean value of the objects for each cluster.
5. Until no change.

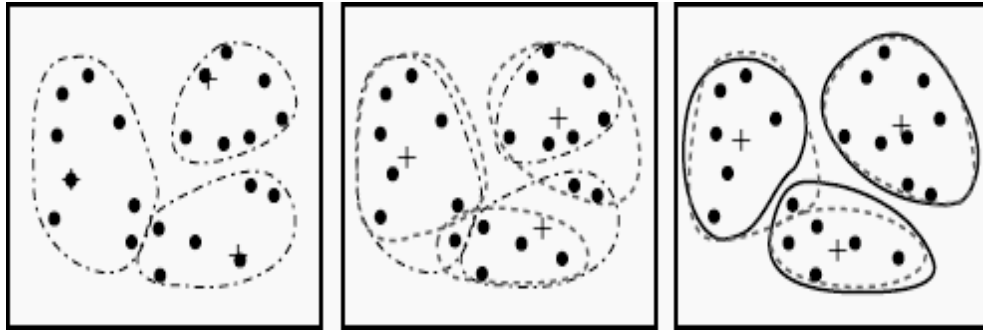


Fig. 6.4: Clustering of a Set of Objects Based on the *k*-Means Method

2. The *k*-Medoids Method

- The *k*-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.
- The *k*-means algorithm is sensitive to outliers because an object with an extremely large value may substantially distort the distribution of data. This effect is particularly exacerbated due to the use of the squared-error function.
- Instead of taking the mean value of the objects in a cluster as a reference point, we can pick actual objects to represent the clusters, using one representative object per cluster. Each remaining object is clustered with the representative object to which it is the most similar.
- The partitioning method is then performed based on the principle of minimizing the sum of the dissimilarities between each object and its corresponding reference point, *i.e.*, an absolute-error criterion is used, defined as:

$$E = \sum_{j=1}^k \sum_{p \in C_j} |p - O_j|$$

where, E = the sum of the absolute e error for all objects in the dataset

p = the point in space representing a given object in cluster C

O_j = the representative object of C_j

- The initial representative objects are chosen arbitrarily. The iterative process of replacing representative objects by non-representative objects continues as long as the quality of the resulting clustering is improved.

- This quality is estimated using a cost function that measures the average dissimilarity between an object and the representative object of its cluster.
- To determine whether a non-representative object, O_{random} , is a good replacement for a current representative object, O_j , the following four cases are examined for each of the non-representative objects.

Case 1:

p currently belongs to representative object, O_j . If O_j is replaced by O_{random} as a representative object and p is closest to one of the other representative objects, O_i , $i \neq j$, then p is reassigned to O_i .

Case 2:

p currently belongs to representative object, O_j . If O_j is replaced by O_{random} as a representative object and p is closest to O_{random} , then p is reassigned to O_{random} .

Case 3:

p currently belongs to representative object, O_i , $i \neq j$. If O_j is replaced by O_{random} as a representative object and p is still closest to O_i , then the assignment does not change.

Case 4:

P currently belongs to representative object, o_i , $i \neq j$. If O_j is replaced by O_{random} as a representative object and p is closest to O_{random} , then p is reassigned to O_{random} .

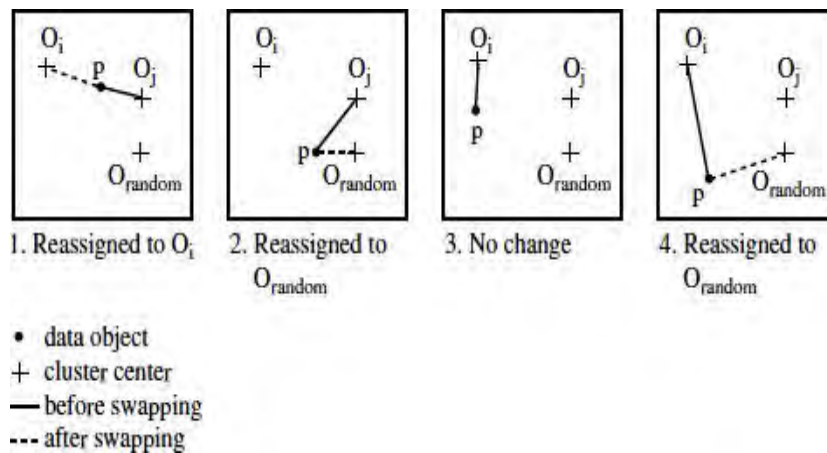


Fig. 6.5: Four Cases of the Cost Function for k -Medoids Clustering

The k -Medoids Algorithm:

The k -medoids algorithm for partitioning based on medoid or central objects.

Input:

- k : the number of clusters.
- D : a dataset containing n objects.

Output: A set of k clusters.

Method:

1. Arbitrarily choose k objects in D as the initial representative objects or seeds.
2. Repeat.
3. Assign each remaining object to the cluster with the nearest representative object.
4. Randomly select a non-representative object, O_{random} .
5. Compute the total cost, S , of swapping representative object, O_j with O_{random} .
6. If $S < 0$, then swap O_j with O_{random} to form the new set of k representative objects.
7. Until no change.

The k -medoids method is more robust than k -means in the presence of noise and outliers, because a medoid is less influenced by outliers or other extreme values than a mean. However, its processing is more costly than the k -means method.

6.7.5.2 Hierarchical Methods

A hierarchical method creates a hierarchical decomposition of the given set of data objects. It can be classified as being either agglomerative or divisive, based on how the hierarchical decomposition is formed.

- The agglomerative approach, also called the bottom-up approach, starts with each object forming a separate group. It successively merges the objects or groups that are close to one another, until all of the groups are merged into one or until a termination condition holds.
- The divisive approach, also called the top-down approach, starts with all of the objects in the same cluster. In each successive iteration, a cluster is split up into smaller clusters, until eventually each object is in one cluster, or until a termination condition holds.

Hierarchical Clustering Methods

- A hierarchical clustering method works by grouping data objects into a tree of clusters.
- The quality of a pure hierarchical clustering method suffers from its inability to perform adjustment once a merge or split decision has been executed, *i.e.*, if a particular merge or split decision later turns out to have been a poor choice, the method cannot back track and correct it.

Hierarchical clustering methods can be further classified as either agglomerative or divisive, depending on whether the hierarchical decomposition is formed in a bottom-up or top-down fashion.

Agglomerative hierarchical clustering:

- This bottom-up strategy starts by placing each object in its own cluster and then merges these atomic clusters into larger and larger clusters, until all of the objects are in a single cluster or until certain termination conditions are satisfied.
- Most hierarchical clustering methods belong to this category. They differ only in their definition of inter-cluster similarity.

Divisive hierarchical clustering:

- This top-down strategy does the reverse of agglomerative hierarchical clustering by starting with all objects in one cluster.
- It subdivides the cluster into smaller and smaller pieces, until each object forms a cluster on its own or until it satisfies certain termination conditions, such as a desired number of clusters is obtained or the diameter of each cluster is within a certain threshold.

6.7.5.3 Density-based Methods

- Most partitioning methods cluster objects based on the distance between objects. Such methods can find only spherical-shaped clusters and encounter difficulty at discovering clusters of arbitrary shapes.
- Other clustering methods have been developed based on the notion of density. Their general idea is to continue growing the given cluster as long as the density in the neighbourhood exceeds some threshold; *i.e.*, for each data point within a given cluster, the neighbourhood of a given radius has to contain at least a minimum number of points. Such a method can be used to filter out noise (outliers) and discover clusters of arbitrary shape.
- DBSCAN and its extension, OPTICS, are typical density-based methods that grow clusters according to a density-based connectivity analysis. DENCLUE is a method that clusters objects based on the analysis of the value distributions of density functions.

6.7.5.4 Grid-based Methods

- Grid-based methods quantize the object space into a finite number of cells that form a grid structure.
- All of the clustering operations are performed on the grid structure, *i.e.*, on the quantized space. The main advantage of this approach is its fast processing time, which is typically independent of the number of data objects and dependent only on the number of cells in each dimension in the quantized space.
- STING is a typical example of a grid-based method. Wave Cluster applies wavelet transformation for clustering analysis and is both grid-based and density-based.

6.7.5.5 Model-based Methods

- Model-based methods hypothesize a model for each of the clusters and find the best fit of the data to the given model.
- A model-based algorithm may locate clusters by constructing a density function that reflects the spatial distribution of the data points.
- It also leads to a way of automatically determining the number of clusters based on standard statistics, taking noise or outliers into account, and thus yielding robust clustering methods.

6.8 TASKS IN DATA MINING

- Clustering High-dimensional Data
- Constraint-based Clustering

1. Clustering High-dimensional Data

- It is a particularly important task in cluster analysis because many applications require the analysis of objects containing a large number of features or dimensions.
For example, text documents may contain thousands of terms or keywords as features, and DNA micro array data may provide information on the expression levels of thousands of genes under hundreds of conditions.
- Clustering high-dimensional data is challenging due to the curse of dimensionality.
- Many dimensions may not be relevant. As the number of dimensions increases, the data become increasingly sparse so that the distance measurement between pairs of points becomes meaningless and the average density of points anywhere in the data is likely to be low. Therefore, a different clustering methodology needs to be developed for high-dimensional data.
- CLIQUE and PROCLUS are two influential subspace clustering methods, which search for clusters in sub-spaces of the data, rather than over the entire dataspace.
- Frequent pattern-based clustering, another clustering methodology, extracts distinct frequent patterns among subsets of dimensions that occur frequently. It uses such patterns to group objects and generate meaningful clusters.

2. Constraint-based Clustering

- It is a clustering approach that performs clustering by incorporation of user-specified or application-oriented constraints.
- A constraint expresses a user's expectation or describes properties of the desired clustering results, and provides an effective means for communicating with the clustering process.

- Various kinds of constraints can be specified, either by a user or as per application requirements.
- Spatial clustering employs with the existence of obstacles and clustering under user-specified constraints. In addition, semi-supervised clustering employs for pair-wise constraints in order to improve the quality of the resulting clustering.

6.8.1 Constraint-based Cluster Analysis

Constraint-based clustering finds clusters that satisfy user-specified preferences or constraints. Depending on the nature of the constraints, constraint-based clustering may adopt rather different approaches.

There are a few categories of constraints.

- **Constraints on individual objects:** We can specify constraints on the objects to be clustered. In a real estate application, for example, one may like to spatially cluster only those luxury mansions worth over a million dollars. This constraint confines these to f objects to be clustered. It can easily be handled by pre-processing after which the problem reduces to an instance of unconstrained clustering.
- **Constraints on the selection of clustering parameters:** A user may like to set a desired range for each clustering parameter. Clustering parameters are usually quite specific to the given clustering algorithm. Examples of parameters include k , the desired number of clusters in a k -means algorithm; or e , the radius and the minimum number of points in the DBSCAN algorithm. Although such user-specified parameters may strongly influence the clustering results, they are usually confined to the algorithm itself. Thus, their fine-tuning and processing are usually not considered a form of constraint-based clustering.
- **Constraints on distance or similarity functions:** We can specify different distance or similarity functions for specific attributes of the objects to be clustered, or different distance measures for specific pairs of objects. When clustering sports men, for example, we may use different weighting schemes for height, body weight, age, and skill level. Although this will likely change the mining results, it may not alter the clustering process *per se*. However, in some cases, such changes may make the evaluation of the distance function non-trivial, especially when it is tightly intertwined with the clustering process.
- **User-specified constraints on the properties of individual clusters:** A user may like to specify desired characteristics of the resulting clusters, which may strongly influence the clustering process.
- **Semi-supervised clustering based on partial supervision:** The quality of unsupervised clustering can be significantly improved using some weak form of supervision. This may be in the form of pair-wise constraints (*i.e.*, pairs of objects labelled as belonging to the same or different cluster). Such a constrained clustering process is called semi-supervised clustering.

6.9 OUTLIER ANALYSIS

- There exist data objects that do not comply with the general behaviour or model of the data. Such data objects, which are grossly different from or inconsistent with the remaining set of data, are called outliers.
- Many data mining algorithms try to minimize the influence of outliers or eliminate them altogether. This, however, could result in the loss of important hidden information because one person's noise could be another person's signal. In other words, the outliers may be of particular interest, such as in the case of fraud detection, where outliers may indicate fraudulent activity. Thus, outlier detection and analysis is an interesting data mining task, referred to as outlier mining.
- It can be used in fraud detection, for example, by detecting unusual usage of credit cards or telecommunication services. In addition, it is useful in customized marketing for identifying the spending behaviour of customers with extremely low or extremely high incomes, or in medical analysis for finding unusual responses to various medical treatments.

Outlier mining can be described as follows:

Given a set of n data points or objects and k , the expected number of outliers, find the top k objects that are considerably dissimilar, exceptional, or inconsistent with respect to the remaining data. The outlier mining problem can be viewed as two sub-problems:

- Define what data can be considered as inconsistent in a given dataset, and
- Find an efficient method to mine the outliers so defined.

6.9.1 Types of Outlier Detection

- Statistical Distribution-based Outlier Detection.
- Distance-based Outlier Detection.
- Density-based Local Outlier Detection.
- Deviation-based Outlier Detection.

6.9.1.1 Statistical Distribution-based Outlier Detection

The statistical distribution-based approach to outlier detection assumes a distribution or probability model for the given dataset (*e.g.*, a normal or Poisson distribution) and then identifies outliers with respect to the model using a discordancy test. Application of the test requires knowledge of the dataset parameters knowledge of distribution parameters such as the mean and variance, and the expected number of outliers.

A statistical discordancy test examines two hypotheses:

1. A working hypothesis.
2. An alternative hypothesis.

Working Hypothesis: A working hypothesis, H , is a statement that the entire dataset of n objects comes from an initial distribution model, F , i.e.,

$$H: O_i \in F, \text{ where, } i = 1, 2, \dots, n$$

The hypothesis is retained if there is no statistically significant evidence supporting its rejection. A discordancy test verifies whether an object, O_i , is significantly large (or small) in relation to the distribution F . Different test statistics have been proposed for use as a discordancy test, depending on the available knowledge of the data. Assuming that some statistic, T , has been chosen for discordancy testing, and the value of the statistic for object O_i is V_i , then the distribution of T is constructed. Significance probability, $SP(v_i) = \text{Prob}(T > V_i)$, is evaluated. If $SP(V_i)$ is sufficiently small, then O_i is discordant and the working hypothesis is rejected.

Alternative Hypothesis: An alternative hypothesis, H , which states that O_i comes from another distribution model, G , is adopted. The result is very much dependent on which model F is chosen because O_i may be an outlier under one model and a perfectly valid value under another. The alternative distribution is very important in determining the power of the test, i.e., the probability that the working hypothesis is rejected when O_i is really an outlier.

There are different kinds of alternative distributions.

- **Inherent alternative distribution:** In this case, the working hypothesis that all of the objects come from distribution F is rejected in favour of the alternative hypothesis that all of the objects arise from another distribution, G :

$$H: O_i \in G, \text{ where, } i = 1, 2, \dots, n$$

F and G may be different distributions or differ only in parameters of the same distribution.

There are constraints on the form of the G distribution in that it must have potential to produce outliers. For example, it may have a different mean or dispersion, or a longer tail.

- **Mixture alternative distribution:** The mixture alternative states that discordant values are not outliers in the F population, but contaminants from some other population, G . In this case, the alternative hypothesis is:

$$H: O_i \in (1 - \lambda)F + \lambda G, \text{ where } i = 1, 2, \dots, n$$

- **Slip page alternative distribution:** This alternative states that all of the objects (apart from some prescribed small number) arise independently from the initial model, F , with its given parameters, whereas the remaining objects are independent observations from a modified version of F in which the parameters have been shifted.

There are two basic types of procedures for detecting outliers:

- (i) **Block procedures:** In this case, either all of the suspect objects are treated as outliers or all of them are accepted as consistent.

- (ii) **Consecutive procedures:** An example of such a procedure is the inside-out procedure. Its main idea is that the object that is least likely to be an outlier is tested first. If it is found to be an outlier, then all of the more extreme values are also considered outliers. Otherwise, the next most extreme object is tested, and so on. This procedure tends to be more effective than block procedures.

6.9.1.2 Distance-based Outlier Detection

The notion of distance-based outliers was introduced to counter the main limitations imposed by statistical methods. An object, o , in a dataset, D , is a distance-based (DB) outlier with parameters pct and $dmin$, i.e., a DB (pct ; $dmin$)-outlier, if at least a fraction, pct , of the objects in D lie at a distance greater than $dmin$ from o . In other words, rather than relying on statistical tests, we can think of distance-based outliers as those objects that do not have enough neighbours, where neighbours are defined based on distance from the given object. In comparison with statistical-based methods, distance-based outlier detection generalizes the ideas behind discordancy testing for various standard distributions. Distance-based outlier detection avoids the excessive computation that can be associated with fitting the observed distribution into some standard distribution and in selecting discordancy tests.

For many discordancy tests, it can be shown that if an object, o , is an outlier according to the given test, then o is also a DB (pct , $dmin$)-outlier for some suitably defined pct and $dmin$. For example, if objects that lie three or more standard deviations from the mean are considered to be outliers, assuming a normal distribution, then this definition can be generalized by a DB(0.9988, 0.13s) outlier.

Several efficient algorithms for mining distance-based outliers have been developed.

- **Index-based algorithm:** Given a dataset, the index-based algorithm uses multidimensional indexing structures, such as R -trees or k - d trees, to search for neighbours of each object o within radius $dmin$ around that object. Let M be the maximum number of objects within the $dmin$ -neighbourhood of an outlier. Therefore, once $M + 1$ neighbours of object o are found, it is clear that o is not an outlier. This algorithm has a worst-case complexity of $O(n_2k)$, where n is the number of objects in the dataset and k is the dimensionality. The index-based algorithm scales well as k increases. However, this complexity evaluation takes only the search time into account, even though the task of building an index in itself can be computationally intensive.
- **Nested-loop algorithm:** The nested-loop algorithm has the same computational complexity as the index-based algorithm but avoids index structure construction and tries to minimize the number of I/Os. It divides the memory buffer space into two halves and the dataset into several logical blocks. By carefully choosing the order in which blocks are loaded into each half, I/O efficiency can be achieved.
- **Cell-based algorithm:** To avoid $O(n^2)$ computational complexity, a cell-based algorithm was developed for memory-resident datasets. Its complexity is $O(c^k + n)$, where, c is a constant depending on the number of cells and k is the dimensionality.

In this method, the data space is partitioned into cells with a side length equal to $\frac{d_{\min}}{2\sqrt{k}}$. Each cell has two layers surrounding it. The first layer is one cell thick, while the second is $\lfloor 2\sqrt{k-1} \rfloor$ cells thick, rounded up to the closest integer. The algorithm counts outliers on a cell-by-cell rather than an object-by-object basis. For a given cell, it accumulates three counts—the number of objects in the cell; in the cell and the first layer together; and in the cell and both layers together. Let's refer to these counts as cell count, cell + 1 layer count, and cell + 2 layers count, respectively.

Let M be the maximum number of outliers that can exist in the $d_{\min}$ -neighbourhood of an outlier.

An object, o , in the current cell is considered an outlier only if cell +1 layer count is less than or equal to M . If this condition does not hold, then all of the objects in the cell can be removed from further investigation as they cannot be outliers.

If cell_+ 2_layers_count is less than or equal to M , then all of the objects in the cell are considered outliers. Otherwise, if this number is more than M , then it is possible that some of the objects in the cell may be outliers. To detect these outliers, object-by-object processing is used where, for each object, o , in the cell, objects in the second layer of o are examined. For objects in the cell, only those objects having no more than M points in their $d_{\min}$ -neighbourhoods are outliers. The $d_{\min}$ -neighbourhood of an object consists of the object's cell, all of its first layer, and some of its second layer.

A variation to the algorithm is linear with respect to n and guarantees that no more than three passes over the dataset are required. It can be used for large disk-resident datasets, yet does not scale well for high dimensions.

6.9.1.3 Density-based Local Outlier Detection

Statistical and distance-based outlier detection both depend on the overall or global distribution of the given set of data points, D . However, data are usually not uniformly distributed. These methods encounter difficulties when analyzing data with rather different density distributions.

To define the local outlier factor of an object, we need to introduce the concepts of k -distance, k -distance neighbourhood, reachability distance, and local reach ability density.

These are defined as follows:

- The **k -distance** of an object p is the maximal distance that p gets from its k -nearest neighbours. This distance is denoted as $k\text{-distance}(p)$. It is defined as the distance, $d(p, o)$, between p and an object $o \in D$, such that: (1) for at least k objects, $o' \in D$, it holds that $d(p, o') \leq d(p, o)$, i.e., there are at least k objects in D that are as close as or closer to p than o , and (2) for at most $k - 1$ objects, $o'' \in D$, it holds that $d(p, o'') < d(p, o)$, i.e., there are at most $k - 1$ objects that are closer to p than o . You may be wondering at this point how k is determined. The LOF method links to density-based clustering in that it

sets k to the parameter *MinPts*, which specifies the minimum number of points for use in identifying clusters based on density. Here, *MinPts* (as k) is used to define the local neighborhood of an object, p .

- The **k -distance neighbourhood** of an object p is denoted as $N_{k_distance(p)}(p)$, or $N_k(p)$ for short. By setting k to *MinPts*, we get $N_{MinPts}(p)$. It contains the *MinPts*-nearest neighbours of p , i.e., it contains every object whose distance is not greater than the *MinPts*-distance of p .
- The **reachability distance** of an object p with respect to object o (where o is within the *MinPts*-nearest neighbors of p), is defined as $reach_dist_{MinPts}(p, o) = \max\{MinPts_distance(o), d(p, o)\}$.

Intuitively, if an object p is far away, then the reachability distance between the two is simply their actual distance. However, if they are sufficiently close (i.e., where p is within the *MinPts*-distance neighbourhood of o), then the actual distance is replaced by the *MinPts*-distance of o . This helps to significantly reduce the statistical fluctuations of $d(p, o)$ for all of the p close to o .

The higher the value of *MinPts* is, the more similar is the reachability distance for objects within the same neighbourhood.

- Intuitively, the **local reachability density** of p is the inverse of the average reachability density based on the *MinPts*-nearest neighbours of p . It is defined as:

$$lrd_{MinPts}(P) = \frac{|N_{MinPts}(p)|}{\sum_{o \in N_{MinPts}(p)} reach_dist_{MinPts}(p, o)}$$

- The local outlier factor (LOF) of p captures the degree to which we call p an outlier. It is defined as:

$$LOF_{MinPts}(p) = \frac{\sum_{o \in N_{MinPts}(p)} \frac{lrd_{MinPts}(o)}{lrd_{MinPts}(p)}}{|N_{MinPts}(p)|}$$

- It is the average of the ratio of the local reachability density of p and those of p 's *MinPts*-nearest neighbors. It is easy to see that the lower p 's local reachability density is, and the higher the local reachability density of p 's *MinPts*-nearest neighbours are, the higher $LOF(p)$ is.

6.9.1.4 Deviation-based Outlier Detection

Deviation-based outlier detection does not use statistical tests or distance-based measures to identify exceptional objects. Instead, it identifies outliers by examining the main characteristics of objects in a group. Objects that “deviate” from this description are considered outliers. Hence, in this approach, the term deviations is typically used to refer to outliers. Sequential Exception Technique (first technique given) but OLAP data cube approach not given.

Sequential Exception Technique

The sequential exception technique simulates the way in which humans can distinguish unusual objects from among a series of supposedly like objects. It uses implicit redundancy of the data. Given a dataset, D , of n objects, it builds a sequence of subsets, $\{D_1, D_2, \dots, D_m\}$, of these objects with $2 \leq m \leq n$ such that

$$D_{j-1} \subset D_j, \text{ where, } D_j \subseteq D$$

Dissimilarities are assessed between subsets in the sequence. The technique introduces the following key terms:

- **Exception set:** This is the set of deviations or outliers. It is defined as the smallest subset of objects whose removal results in the greatest reduction of dissimilarity in the residual set.
- **Dissimilarity function:** This function does not require a metric distance between the objects. It is any function that, if given a set of objects, returns a low value if the objects are similar to one another. The greater the dissimilarity among the objects, the higher the value returned by the function. The dissimilarity of a subset is incrementally computed based on the subset prior to it in the sequence. Given a subset of n numbers, $\{x_1, \dots, x_n\}$, a possible dissimilarity function is the variance of the numbers in the set, that is,

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2,$$

where $\bar{x}$ is the mean of the n numbers in the set. For character strings, the dissimilarity function may be in the form of a pattern string (*e.g.*, containing wildcard characters that is used to cover all of the patterns seen so far. The dissimilarity increases when the pattern covering all of the strings in D_{j-1} does not cover any string in D_j that is not in D_{j-1} .

- **Cardinality function:** This is typically the count of the number of objects in a given set.
- **Smoothing factor:** This function is computed for each subset in the sequence. It assesses how much the dissimilarity can be reduced by removing the subset from the original set of objects.

6.10 CLUSTERING IN R

Clustering is one of the most popular and commonly used classification techniques used in machine learning. In clustering or cluster analysis in R , we attempt to group objects with similar traits and features together, such that a larger set of objects is divided into smaller sets of objects. The objects in a subset are more similar to other objects in that set than to objects in other sets.

Clustering is not an algorithm, rather it is a way of solving classification problems. There are multiple algorithms that solve classification problems by using the clustering method. These

algorithms differ in their efficiency, their approach to sorting objects into the various clusters, and even their definition of a cluster.

Imagine you have a dataset containing n rows and m columns and that we need to classify the objects in the dataset. We can see this dataset as n points in an m -dimensional space. In a clustering algorithm, we calculate the distance between these objects and put the objects nearest to each other into separate clusters.

6.10.1 Applications of Clustering in R

There are many classification problems in every aspect of our lives today. Machine learning helps to solve most of them. One of the multitudes of clustering algorithms helps to solve these problems. Let us look at a few of the real-life problems that are solved using clustering.

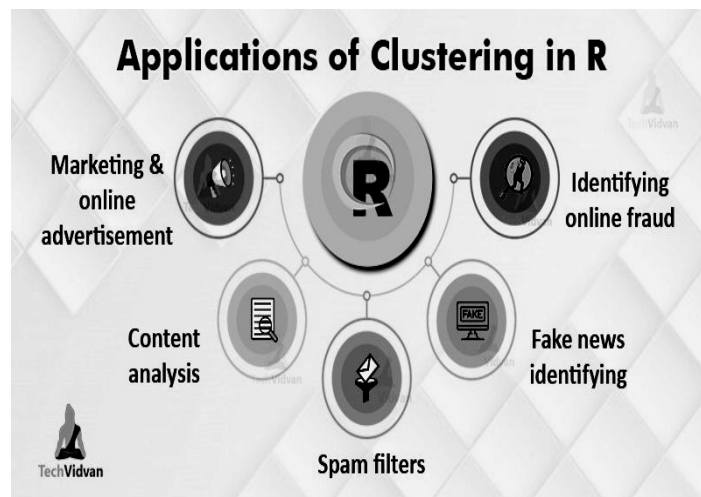


Fig. 6.6: Applications of Clustering in R

1. **Marketing and Online Advertisement:** Identifying customers that are more likely to respond to the product and its marketing is a very common classification problem these days. Everybody is online these days. This also poses a problem as not everybody is going to be receptive to a marketing campaign. Clustering algorithms are used to classify various customers according to their interests which helps with targeted marketing.
2. **Content Analysis:** Clustering algorithms are used to classify content based on various factors like key terms, sources, and subjects. Many search engines and custom search services use clustering algorithms to classify documents and content according to their categories and search terms.
3. **Fake News Identifying:** The internet is full of fake news and advice. These fake facts are not only misleading but they can also be dangerous for many people. Clustering algorithms are helpful to match news, facts and advice with verified sources, and classify them as truths, half-truths, and lies.

6.10.2 Types of Clustering Algorithms in R

Types of Clustering Algorithms in R

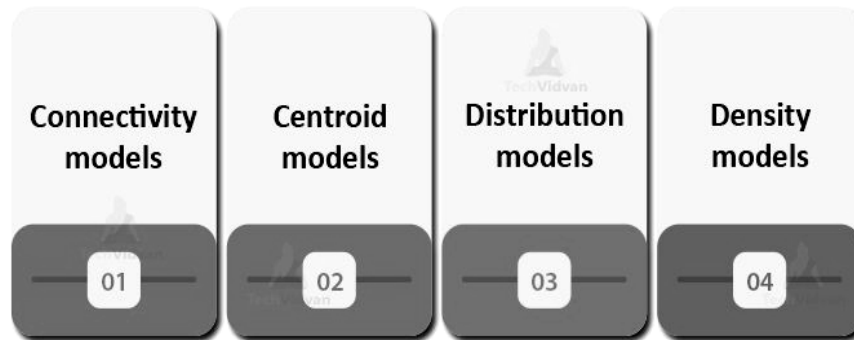


Fig. 6.7: Types of Clustering Algorithms in R

1. **Connectivity Models:** Connectivity models classify data points into clusters based on the distance between them. The basic principle behind these models is that objects closer to each other are likely to be more similar to each other than objects that are farther away.

Connectivity models may have two different approaches – the first being to divide all points into clusters and then aggregating them as the distance increases. The second approach is to put all points in a single cluster and then divide them into separate clusters as the distance increases. The distance measure can also vary from algorithm to algorithm with Euclidean and Manhattan distance being most common.

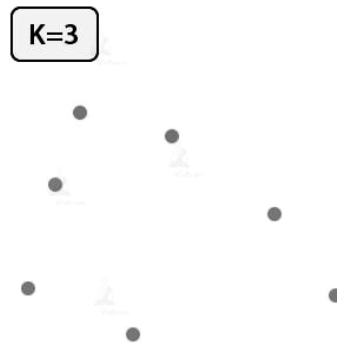
2. **Centroid Models:** Centroid models are iterative clustering algorithms. Their idea of similarity is derived from the distance from the centroid of the cluster. These algorithms require the number of clusters beforehand. With each iteration, they correct the position of the centroid of the clusters and also adjust the classification of each data point.
3. **Distribution Models:** Distribution models differ from the other types of clustering algorithms in that they consider the probability of a data point belonging to a certain cluster instead of any distance measure. These models often suffer from overfitting.
4. **Density Models:** Density models consider the density of the points in different parts of the space to create clusters in the subspaces with similar densities. They isolate various subspaces based on the density of the data point present in them and assign the data points to separate clusters.

There are hundreds of different clustering algorithms available to choose from. But here, we are going to study the two most popular clustering algorithms.

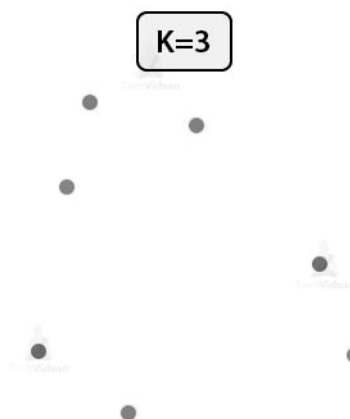
6.10.3 k -means Clustering in R

k -means is a centroid model or an iterative clustering algorithm. It works by finding the local maxima in every iteration. The algorithm works as follows:

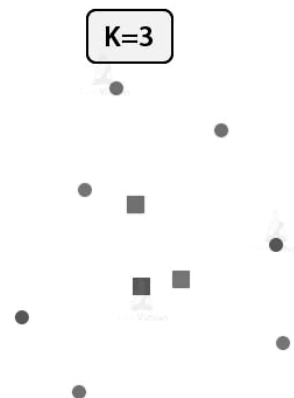
1. **Specify the number of clusters required denoted by K .** Let us take $K = 3$ for the following seven points. This means that two clusters shall exist.



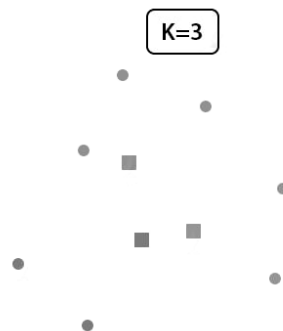
2. **Assign points to clusters randomly:** Let us divide the points among the three clusters randomly. The three different clusters are denoted by three different colours.



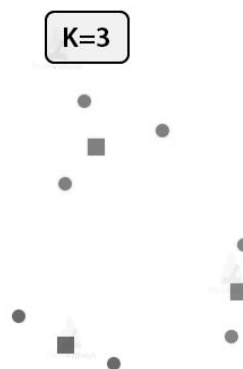
3. **Find the centroids of each cluster:** Now, we have to find the centroids for each of the clusters. We mark them as crosses in the diagram below.



4. **Reassign points according to their closest centroid:** Once we have the centroids, we will reassign points to the centroid they are the closest to. As you can see in the example below, three points have been reassigned to different clusters.



5. **Readjust the positions of the cluster centroids:** Adjust the positions of the cluster centroids according to the new points in the clusters.



6. **Repeat steps 4 and 5 until no further changes are there:** This gives us the final set of clusters with each point classified into one cluster.

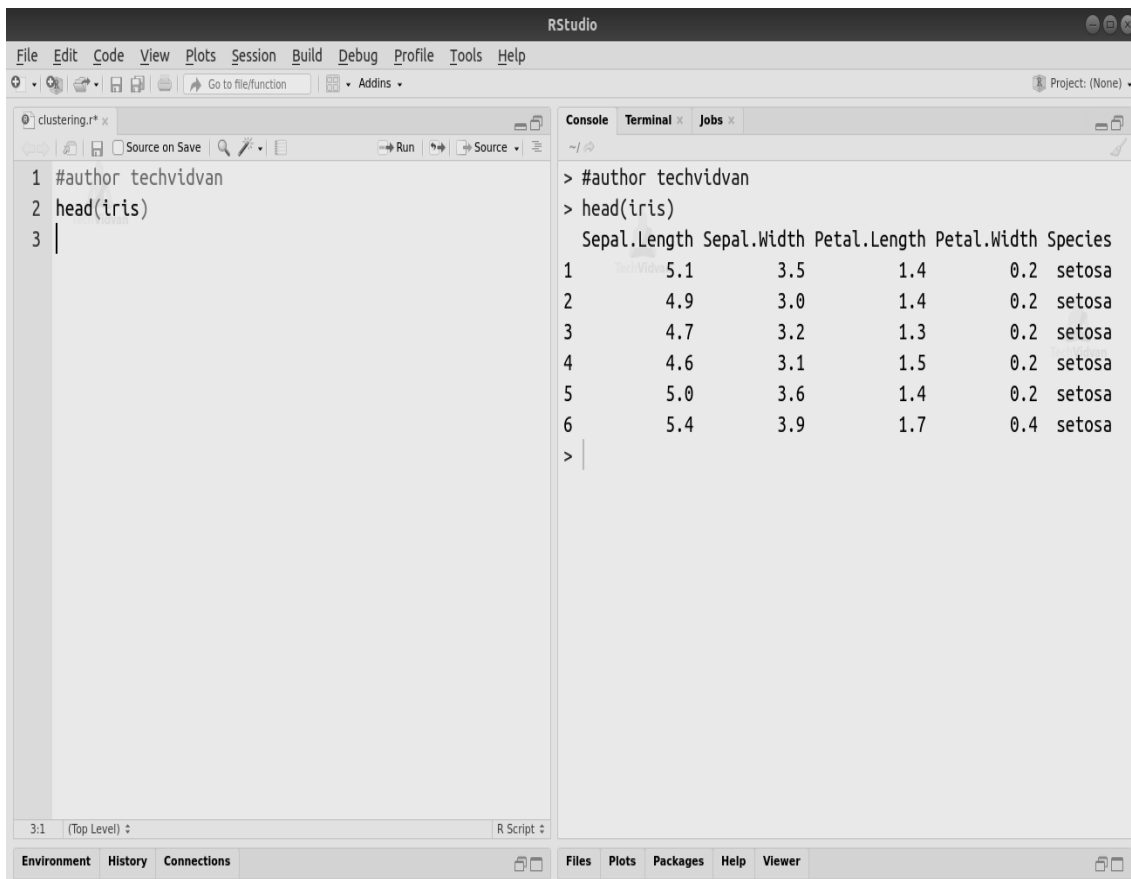
6.10.4 Practical Implementation of Cluster Analysis in R

Let us implement a clustering algorithm in R. We will be implementing the k-means clustering algorithm on the iris dataset that is inbuilt in R. We will also need the ggplot2 package to plot the graphs.

1. Let us explore the data and get familiar with it first.

```
head(iris)
```

Output:



The screenshot shows the RStudio interface. The script editor on the left contains the following code:

```
1 #author techvidvan  
2 head(iris)  
3 |
```

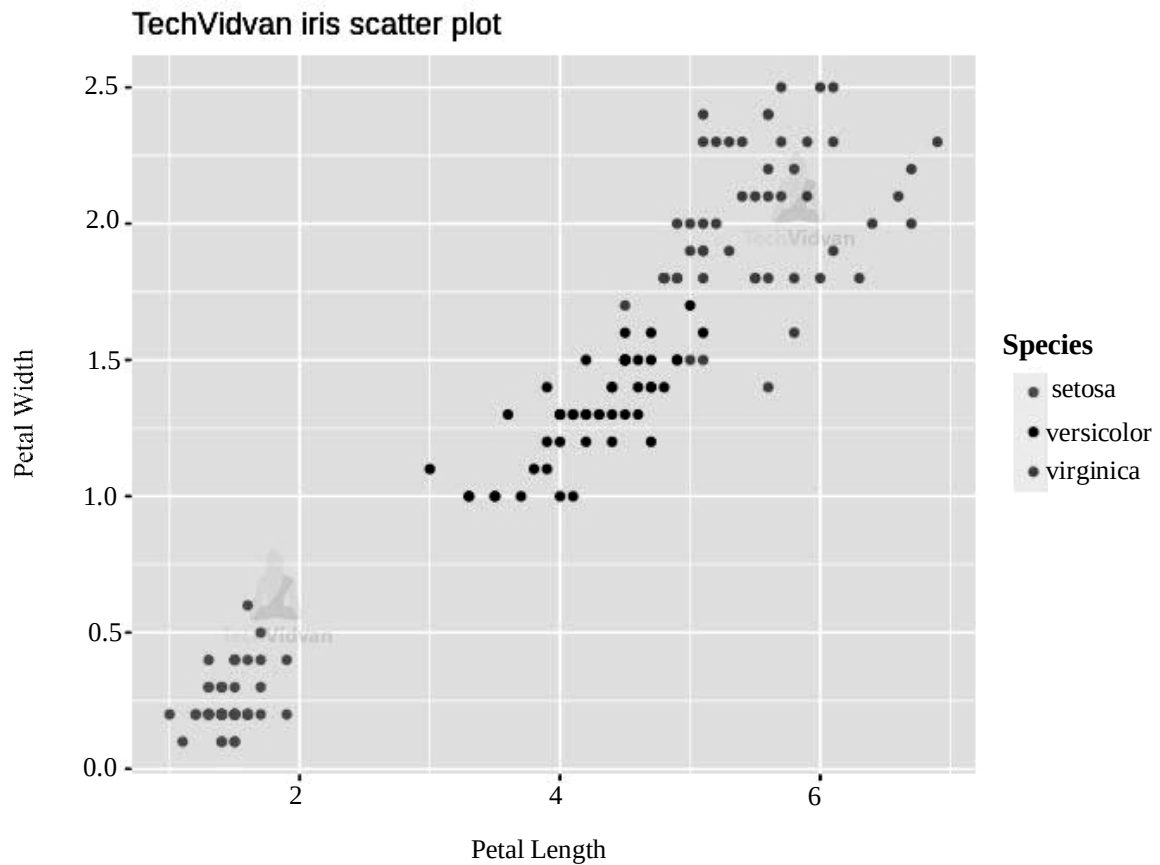
The console on the right shows the output of the `head(iris)` command:

```
> #author techvidvan  
> head(iris)  
      Sepal.Length Sepal.Width Petal.Length Petal.Width Species  
1      5.1         3.5         1.4         0.2   setosa  
2      4.9         3.0         1.4         0.2   setosa  
3      4.7         3.2         1.3         0.2   setosa  
4      4.6         3.1         1.5         0.2   setosa  
5      5.0         3.6         1.4         0.2   setosa  
6      5.4         3.9         1.7         0.4   setosa  
> |
```

2. As we know, the iris dataset contains the sepal and petal length as well as the width of three different variants of the iris flower. The Petal.Length and Petal.Width is similar for flowers of the same variety but vastly different for flowers of different varieties. This is clear in the following plot.

```
library(ggplot2)  
ggplot(iris, aes(Petal.Length, Petal.Width, col=Species)) + geom_point() + ggtitle  
("TechVidvan iris scatterplot")
```

Output:



3. Now, we can use the *k* means() function to form the clusters.

```
set.seed(55)
cluster.iris <- kmeans(iris[, 3:4], 3, nstart = 20)
cluster.iris
```

Output:

The screenshot shows the RStudio interface with a script editor on the left and a console on the right. The script editor contains the following R code:

```

12 #author techvidvan
13 set.seed(55)
14 cluster.iris <- kmeans(iris[, 3:4], 3, nstart = 20)
15 cluster.iris
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32

```

The console shows the output of the R code:

```

> set.seed(55)
> cluster.iris <- kmeans(iris[, 3:4], 3, nstart = 20)
> cluster.iris
K-means clustering with 3 clusters of sizes 52, 48, 50

Cluster means:
  Petal.Length Petal.Width
1   4.269231   1.342308
2   5.595833   2.037500
3   1.462000   0.246000

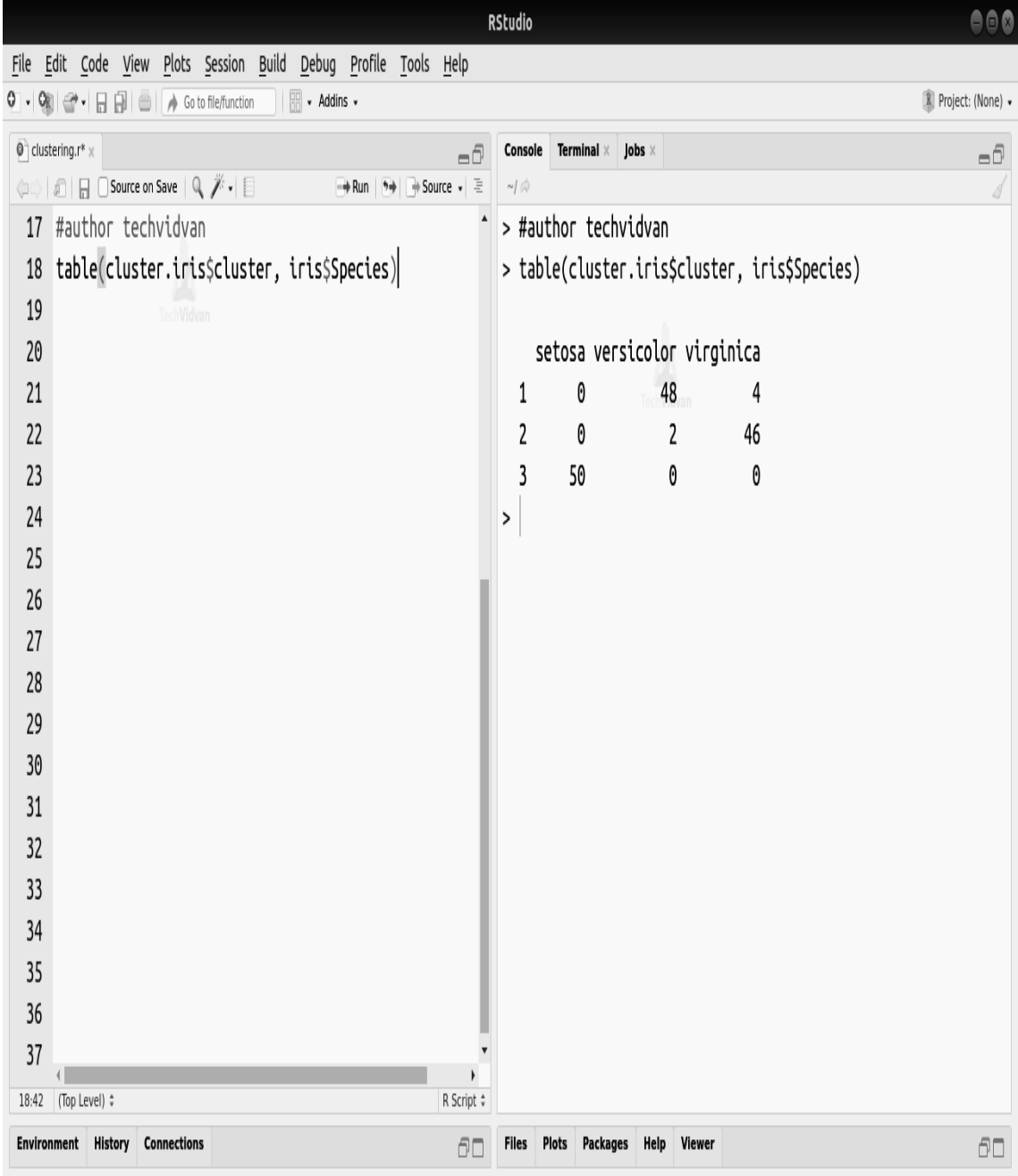
Clustering vector:
 [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[27] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[53] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[79] 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2
[105] 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[131] 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2

Within cluster sum of squares by cluster:
 [1] 13.05769 16.29167  2.02200
(between_SS / total_SS =  94.3 %)

```

4. We can compare the results by forming a table with the species column of the original data.

```
table(cluster.iris$cluster, iris$Species)
```


Output:

The screenshot shows the RStudio interface. The script editor on the left contains the following R code:

```
17 #author techvidvan
18 table(cluster.iris$cluster, iris$Species)
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
```

The console on the right shows the output of the code:

```
> #author techvidvan
> table(cluster.iris$cluster, iris$Species)

      setosa versicolor virginica
1         0          48         4
2         0           2        46
3        50           0         0
```

The status bar at the bottom indicates the cursor is at line 18, column 42, on the 'Top Level' of an 'R Script'.

As we can see, all 50 points of the Setosa variety were put in the 3rd cluster. The Versicolor points were placed in the first cluster but two points of this variety were classified incorrectly. The points in the Virginica variety were put into the second

cluster but four of its points were classified incorrectly. The accuracy of the model can be calculated as:

$$A = (50 + 48 + 46)/150 = 0.96$$

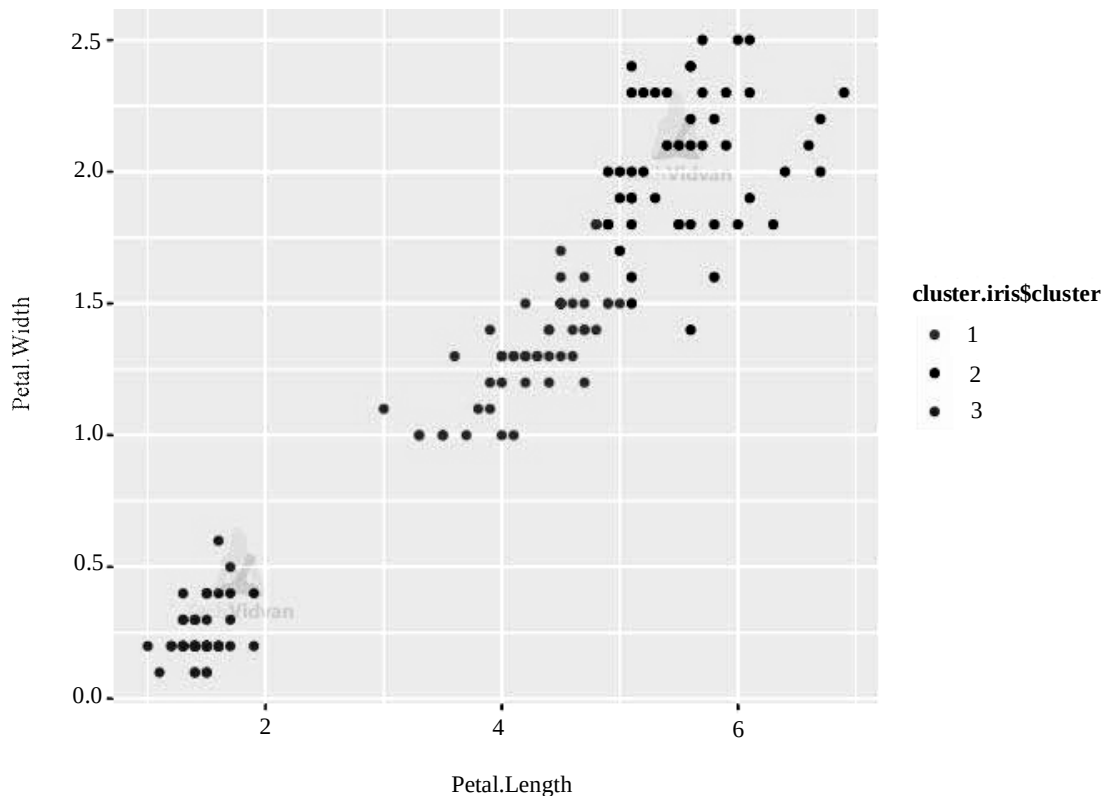
The accuracy is 96%.

5. We can plot the cluster using ggplot. As the accuracy is high, we expect the plot to look very much like the original data plot.

```
cluster.iris$cluster <- as.factor(cluster.iris$cluster)
ggplot(iris, aes(Petal.Length, Petal.Width, color=cluster.iris$cluster)) + geom_point() +
ggtitle("TechVidvan iris cluster plot")
```

Output:

TechVidvan Iris Cluster Plot



Practical – Who Survived the Titanic Disaster?

For illustration, we will be working on one of the most popular datasets in machine learning: **Titanic**. It is fairly small in size and a variety of variables will give us enough space for creative feature engineering and model building. This dataset has been taken from **Kaggle**. I hope you know that model building is the last stage in machine learning. Therefore, we will be doing quick

data exploration, pre-processing, and feature engineering before implementing Logistic Regression.

```
#set working directory
> path <- "~/Manish/Data/Titanic"
> setwd(path)
#load libraries and data
> library (data.table)
> library (plyr)
> library (stringr)
> train <- fread("train.csv",na.strings = c("",",",NA, "NA"))
> test <- fread("test.csv",na.strings = c("",",",NA, "NA"))
```

After we have loaded the data in *R*, we will quickly look at the variables.

```
> head(train)
> head(test)
> str(train)
```

This dataset comprises a mix of character and numeric variables. Among these variables, Survived is the dependent variable. To get a deeper understanding of these variables, let's do a quick data exploration. Discoveries from this step will help us determine what to do next:

```
#check missing values
> colSums(is.na(train))
> colSums(is.na(test))
```

We see that variable Age and Cabin have missing values. Interestingly, the variable Fare has one missing value only in the test set. Let's start exploring each variable individually. For numeric variables, we will use the summary function. For character/factor variables, we will use table for exploration:

```
#Quick Data Exploration
> summary(train$Age)
> summary(test$Age)

> train[,.N/nrow(train),Pclass]
> test[,.N/nrow(test),Pclass]
```

```
> train [,N/nrow(train),Sex]
> test [,N/nrow(test),Sex]

> train [,N/nrow(train),SibSp]
> test [,N/nrow(test),SibSp]

> train [,N/nrow(train),Parch]
> test [,N/nrow(test),Parch]#extra 9

> summary(train$Fare)
> summary(test$Fare)

> train [,N/nrow(train),Cabin]
> test [,N/nrow(test),Cabin]

> train [,N/nrow(train),Embarked]
> test [,N/nrow(test),Embarked]
```

Following are the insights we can derive from the data exploration above:

1. The variable **Fare** is skewed (right) in nature. We will have to log transform it such that it resembles a normal distribution.
2. The variable **Parch** has one extra level (9) in the test set as compared to the train set. We will have to combine it with its mode value.

A smart way to make modifications in train and test data is by combining them. This way, you will save yourself from writing some extra lines of code. I suggest following every line of code carefully and simultaneously checking how every line affects the data.

```
#combine data
> alldata <- rbind(train,test,fill=TRUE)
```

After looking at the data, I could figure out to create some new variables. Didn't you? If not, I request you to look deeper into the data; check the distribution of the dependent variable with predictor variables. You will discover various interesting insights about probability of surviving passengers.

For now, we will create two new variables. First, title of the passengers. This information can be extracted from **Name**. Second, I suspect that **Ticket** notation could give us some

information. For example, some ticket notation starts with alphanumeric, while others only have numbers. We will capture this trend using a binary coded variable.

```
#New Variables
#Extract passengers title
> alldata[,title := strsplit(Name,split="[,.]")]
> alldata[,title := ldply(.data = title,.fun =function(x) x[2])]
> alldata[,title := str_trim(title,side ="left")]

#combine titles
> alldata[,title := replace(title, which(title %in%
c("Capt","Col","Don","Jonkheer","Major","Rev","Sir")), "Mr"),by=title]
> alldata[,title := replace(title, which(title %in% c("Lady","Mlle","Mme","Ms","the
Countess","Dr","Dona")), "Mrs"),by=title]

#ticket binary coding
> alldata[,abs_col := strsplit(x =Ticket,split ="")]
> alldata[,abs_col := ldply(.data = abs_col,.fun =function(x)length(x))]
> alldata[,abs_col := ifelse(abs_col >1,1,0)]
```

Next, we will impute missing values, transform **Fare** variable and remove an extra level from **Parch** variable. This will make our data ready for machine learning.

```
#Impute Age with Median
>for(i in"Age")
set(alldata,i = which(is.na(alldata[[i]])),j=i,value = median(alldata$Age,na.rm = T))

#Remove rows containing NA from Embarked
> alldata <- alldata[!is.na(Embarked)]

#Impute Fare with Median
>for(i in"Fare")
set(alldata,i = which(is.na(alldata[[i]])),j=i,value = median(alldata$Fare,na.rm = T))

#Replace missing values in Cabin with "Miss"
> alldata[is.na(Cabin),Cabin:="Miss"]
```

```
#Log Transform Fare
> alldata$Fare <- log(alldata$Fare +1)

#Impute Parch 9 to 0
> alldata [Parch==9L,Parch:=0]
```

The method of using for-set loop for imputing missing values works blazing fast on large datasets. In our case, the dataset is small. Hence, it is difficult to note the difference. Now, our dataset is ready. Let's implement Logistic Regression and check our model's accuracy.

```
#Collect train and test
> train <- alldata[!(is.na(Survived))]
> train [,Survived:=as.factor(Survived)]

> test <- alldata[is.na(Survived)]
> test [,Survived:= NULL]

#Logistic Regression
> model <- glm(Survived~., family = binomial(link = 'logit'), data = train[, -c("PassengerId",
"Name", "Ticket")])
> summary(model)
```

In R, you can implement Logistic Regression using the glm function. Now, let's understand and interpret the crucial aspects of summary:

1. The glm function internally encodes categorical variables into $n - 1$ distinct levels.
2. Estimate represents the regression coefficients value. Here, the regression coefficients explain the change in log(odds) of the response variable for one unit change in the predictor variable.
3. Std. Error represents the standard error associated with the regression coefficients.
4. Z-value is analogous to t -statistics in multiple regression output. Z-value > 2 implies the corresponding variable is significant.
5. P -value determines the probability of significance of predictor variables. With 95% confidence level, a variable having $p < 0.05$ is considered an important predictor. The same can be inferred by observing stars against p -value.

In addition, we can also perform an ANOVA Chi-square test to check the overall effect of variables on the dependent variable.

```
#run anova
> anova(model, test = 'Chisq')
```

```
Analysis of Deviance Table

Model: binomial, link: logit
Response: Survived
Terms added sequentially (first to last)
```

		Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL				888	1182.82	
Pclass	1	100.179		887	1082.64	< 2.2e-16 ***
Sex	1	255.814		886	826.82	< 2.2e-16 ***
Age	1	21.842		885	804.98	2.961e-06 ***
SibSp	1	14.300		884	790.68	0.0001559 ***
Parch	1	0.434		883	790.25	0.5098483
Fare	1	7.296		882	782.95	0.0069107 **
Cabin	146	160.582		736	622.37	0.1934723
Embarked	2	1.757		734	620.61	0.4153892
title	3	54.459		731	566.16	8.958e-12 ***
abs_col	1	0.363		730	565.79	0.5468450

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Following are the insights we can collect for the output above:

In the presence of other variables, variables such as Parch, Cabin, Embarked, and abs_col are not significant. We will try building another model without including them.

The AIC value of this model is 883.79.

Let's create another model and try to achieve a lower AIC value.

```
> model2 <- glm(Survived~Pclass+Sex+Age+SibSp+Fare+ title, data = train,family =
binomial(link="logit"))
> summary(model2)
```

```

1 Call:
2 glm(formula = Survived ~ Pclass + Sex + Age + SibSp + Fare +
3   title, family = binomial(link = "logit"), data = train)
4
5 Deviance Residuals:
6     Min       1Q   Median       3Q      Max
7  -2.5270  -0.5646  -0.3918   0.5474   2.6789
8
9 Coefficients:
10      Estimate Std. Error z value Pr(>|z|)
11 (Intercept)  6.98914    1.40390   4.978 6.41e-07 ***
12 Pclass      -0.98570    0.16835  -5.855 4.77e-09 ***
13 Sexmale     -3.00290    0.96105  -3.125 0.00178 **
14 Age         -0.02453    0.00875  -2.803 0.00506 **
15 SibSp       -0.69604    0.13105  -5.311 1.09e-07 ***
16 Fare         0.25841    0.14903   1.734 0.08292 .
17 titleMiss   -3.40649    1.08782  -3.131 0.00174 **
18 titleMr     -3.33433    0.53258  -6.261 3.83e-10 ***
19 titleMrs    -2.97026    1.05776  -2.808 0.00498 **
20 ---
21 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
22
23 (Dispersion parameter for binomial family taken to be 1)
24
25 Null deviance: 1182.8 on 888 degrees of freedom
26 Residual deviance: 735.8 on 880 degrees of freedom
27 AIC: 753.8
28
29 Number of Fisher Scoring iterations: 5

```

As you can see, we have achieved a lower AIC value and a better model. Also, we can compare both the models using the ANOVA test. Let's say our null hypothesis is that second model is better than the first model. $p < 0.05$ would reject our hypothesis, and in case $p > 0.05$, we will fail to reject the null hypothesis.

```
#compare two models
```

```
> anova(model,model2,test ="Chisq")
```

```

1 Analysis of Deviance Table
2
3 Model 1: Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Cabin +
4   Embarked + title + abs_col
5 Model 2: Survived ~ Pclass + Sex + Age + SibSp + Fare + title
6   Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
7 1      730      565.79
8 2      880      735.80 -150      -170    0.126

```

With $p > 0.05$, this ANOVA test also corroborates the fact that the second model is better than first model. Let's predict on unseen data now. Since, we cannot evaluate a model's performance on test data locally, we will divide the train set and use model 2 for prediction.

```
#partition and create training, testing data
```

```
> library(caret)
```

```
> split <- createDataPartition(y = train$Survived,p =0.6,list = FALSE)
```



```

> new_train <- train[split]
> new_test <- train[-split]

#model training and prediction
> log_model <- glm(Survived~Pclass+Sex+Age+SibSp+Fare+ title, data = new_train[, -
c("PassengerId", "Name", "Ticket")], family = binomial(link="logit"))
> log_predict <- predict(log_model, newdata = new_test, type = "response")
> log_predict <- ifelse(log_predict > 0.5, 1, 0)

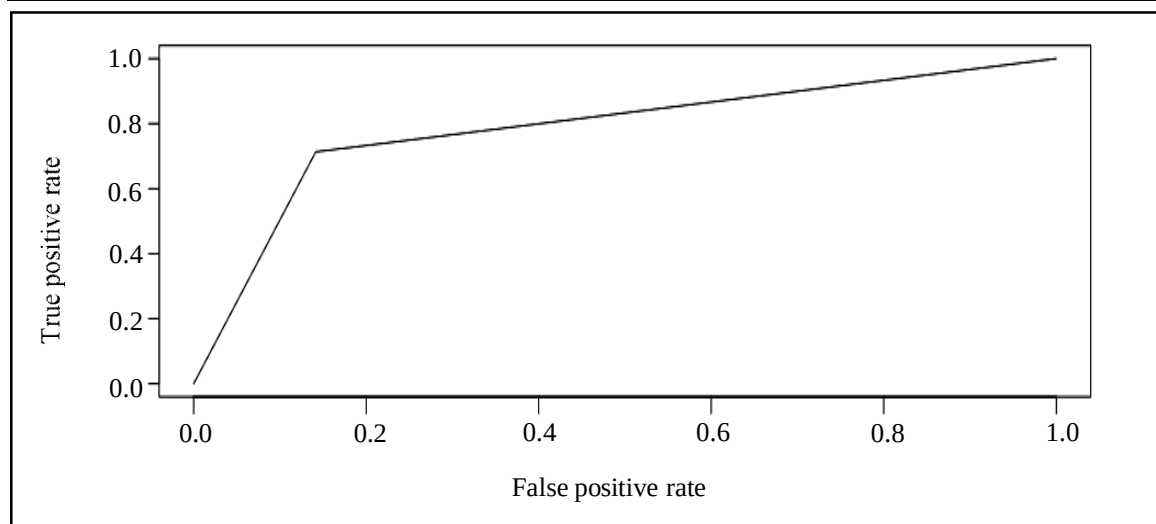
```

For now, I have set the probability threshold value as 0.5. Let's get the flavour of our model's accuracy. We will use AUC-ROC score to determine model fit. Higher the score, better the model. You use confusion matrix to determine accuracy using Confusion Matrix function from caret package.

```

#plot ROC
> library(ROCR)
> library(Metrics)
> pr <- prediction(log_predict, new_test$Survived)
> perf <- performance(pr, measure = "tpr", x.measure = "fpr")
> plot(perf) > auc(new_test$Survived, log_predict) #0.76343

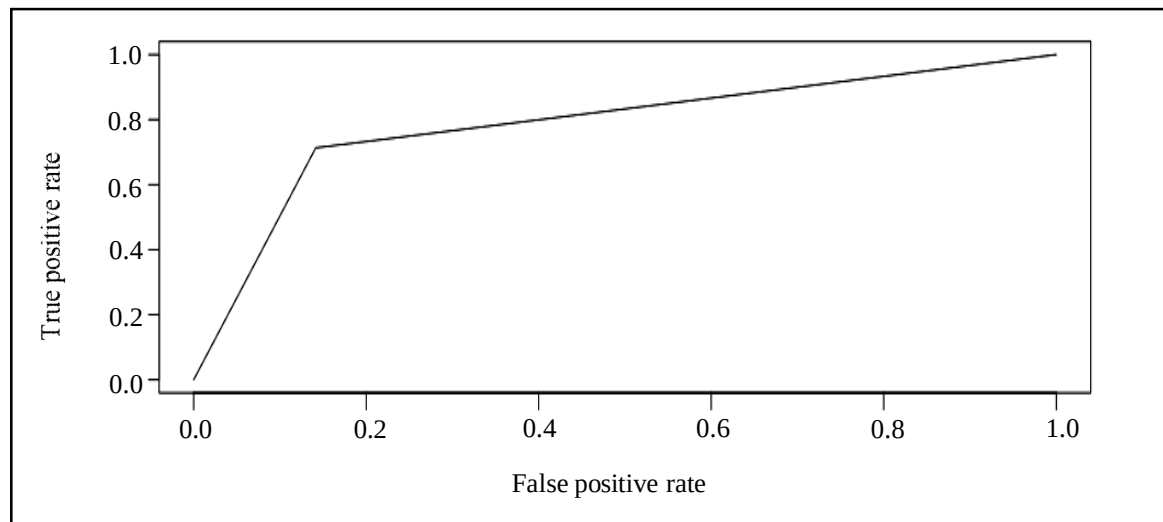
```



Our AUC score is 0.763. As said above, in ROC plot, we always try to move up and top left corner. From this plot, we can interpret that the model is predicting more negative values incorrectly. To move up, let's increase our threshold value to 0.6 and check the model's performance.

```
> log_predict <- predict(log_model,newdata = new_test,type ="response")
> log_predict <- ifelse(log_predict >0.6,1,0)

> pr <- prediction(log_predict,new_test$Survived)
> perf <- performance(pr,measure ="tpr",x.measure ="fpr")
> plot(perf)
> auc(new_test$Survived,log_predict)` ` #0.8008
```



SUMMARY

- A link between predictor factors and a categorical response variable is modelled using logistic regression. For example, we could use logistic regression to forecast if a crack larger than 10 mils will occur by modelling the link between several measurements of a manufactured specimen (such as size and chemical composition) (a binary variable: either yes or no). Given a collection of predictors, logistic regression can help us estimate the likelihood of falling into a specific level of categorical response. Depending on the nature of the categorical answer variable, we can use one of three types of logistic regression: Binary, Nominal and Ordinal logistic regression.
- Binary Logistic Regression (BLR) is a type of logistic regression used to predict the relationship between predictors and predicted variable.
- When the response is binary, this expression is used (*i.e.*, it has two possible outcomes). Binary logistic regression would be used in the cracking scenario above. Passing or failing a test, answering yes or no on a survey, and having high or low blood pressure are all examples of binary responses.
- Nominal Logistic Regression (NLR) is a type of logistic regression that is used to predict when there are three or more categories and there is no natural order to the levels,

this term is used. Departments inside a company (*e.g.*, marketing, sales, HR), the type of search engine utilized (*e.g.*, Google, Yahoo!, MSN), and colour are all examples of nominal replies (black, red, blue, orange).

- Ordinal Logistic Regression (OLR) is a type of logistic regression that is used to predict the dependent variable with ordered multiple categories and independent variables.
- When there are three or more categories with a natural ordering to the levels, but the ranking of the levels does not always imply that the intervals between them are equal, this term is used. Ordinal replies can include how students rank the effectiveness of a college course on a scale of 1 to 5, the flavours of hot wings, and medical conditions (*e.g.*, good, stable, serious, critical).
- Interaction is the result of many predictors (multiple predictors). The influence of one predictor is dependent on the degree of another predictor, which is known as interaction.
- Factor analysis is a statistical method for describing variability among associated variables in terms of a smaller number of unobserved variables known as factors.
- Cluster analysis, often known as clustering, is the problem of arranging a set of items so that objects in the same group (called a cluster) are more comparable (in some sense) to those in other groups (clusters).
- The *R* Commander is a graphical user interface (GUI) for the *R* statistical programme, which is free and open source. The *R* Commander is implemented as a *R* package called Rcmdr, which is accessible for free on CRAN (the *R* package archive).
- Rattle (Williams, 2011) is a *R* programme that provides a graphical user interface to a large number of other *R* packages that provide data mining capability. Rattle is reliant on over 40 other *R* packages as well as a few of non-*R* software applications/libraries.

QUESTIONS

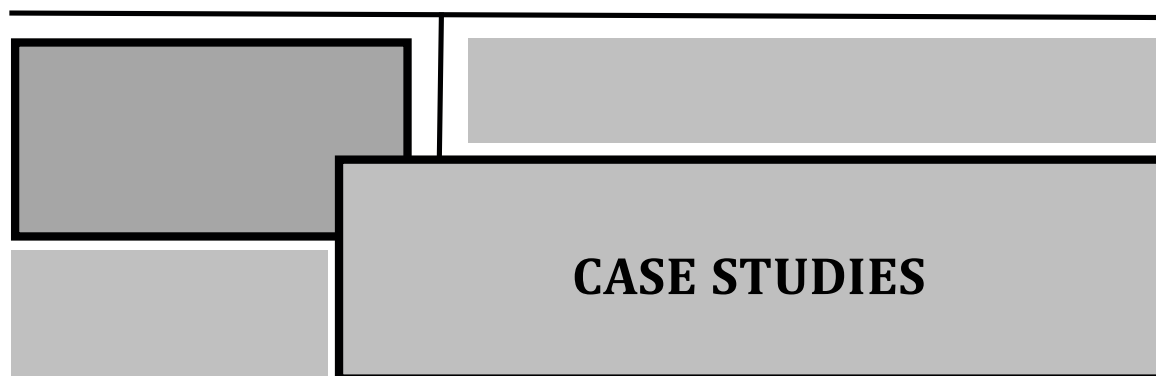
1. Explain logistic regression with single nominal predictor.
2. Explain the assumptions of logistic regression analysis.
3. Explain conducting logistic regression analysis in PSPP.
4. Explain conducting logistic regression analysis in *R* Commander.
5. Explain output of logistic regression analysis.
6. What is multiple predictors?
7. Explain pros and cons of logistic regression as a supervised learning algorithm.
8. What is factor analysis?
9. What is cluster analysis?
10. Explain the assumptions of factor analysis.
11. Explain conducting factor analysis in PSPP and *R* Commander.
12. Explain interpretation of output of factor analysis.
13. Explain the assumptions of cluster analysis.

14. Explain conducting cluster analysis in PSPP and in *R* Commander.
15. Explain interpretation of output of cluster analysis.
16. Explain conducting cluster analysis in Rattle.

REFERENCES

- <https://online.stat.psu.edu/stat501/lesson/15/15.1>
- <https://www.sciencedirect.com/topics/mathematics/multiple-predictor>
- https://en.wikipedia.org/wiki/Factor_analysis#:~:text=Factor%20analysis%20is%20a%20statistical,of%20unobserved%20variables%20called%20factors.
- [https://en.wikipedia.org/wiki/Cluster_analysis#:~:text=Cluster%20analysis%20or%20clustering%20is,in%20other%20groups%20\(clusters\).](https://en.wikipedia.org/wiki/Cluster_analysis#:~:text=Cluster%20analysis%20or%20clustering%20is,in%20other%20groups%20(clusters).)
- <https://cran.r-project.org/web/packages/rattle/vignettes/rattle.pdf>
- <https://socialsciences.mcmaster.ca/jfox/Misc/Rcmdr>





CASE STUDY 1

A ROADMAP TOWARDS PREDICTIVE HR ANALYTICS

HR Analytics demonstrates or predicts the direct impact of people data on business outcomes.

Leading-edge companies are predominantly and progressively embracing sophisticated and futuristic approaches of scrutinizing and interpreting employee data to magnify their competitive advantage manifold. The corporate giants like Google and others need to ensure the highest productivity, engagement, and retention of top talent, and then replicating their successes. HR Analytics is the only way by which we will be able to forecast how much better performance we expect from our top employees — who are perhaps our greatest asset and our largest expense—we'll do well to favoring analytics over our gut instincts.

Data are the new bacon of enterprise business. But as it was difficult in collecting and analyzing HR data and also the existing HRIS systems unable to satisfy the business need, the previous research also evidence to invest in human capital (Fred & Kinange, 2015) as the HRIS suggested. Also, in 2009, Google worked on finding out the best competent traits that are needed to be an effective manager and doing this Google developed “Project Oxygen” which caused a prodigious transition from traditional HR measurements to HR analytics. Google too pointed out the perks of using HR analytics in ameliorating organisational performances. The HRIS has many great features but as HRIS workflow (Achary, 2018) is based on hard-coded business rules, industries are in favour of HR analytics which are based on statistical modelling and machine learning for better connect HR need.

CASE STUDY 2

USING HR ANALYTICS TO SUPPORT MANAGERIAL DECISIONS: A CASE STUDY

More and more organizations are becoming employee centric in the 21st century. The employment and workforce industry are increasingly relevant because the value of human capital is directly linked to an organization's profitability. Human Resource (HR) Analytics enables HR managers to make strategic contributions and support managerial decisions. However, there are

several challenges with integrating analytics into HR: the data is significant, messy, and imbalanced, it is hard to harness both structured and unstructured data, most HR managers lack data mining skills, and there is a lack of related empirical research that gives detailed analytics guidelines. The contribution of this study is that we develop a framework to support an industrial aluminum company to make the decisions and to improve strategy execution.

CASE STUDY 3

WALMART – HR ANALYTICS

Walmart is the world's largest privately owned company with 2.2 million global employee. Its HR analytics team was created by Elpide Orfanidou in 2009 and now numbers over 70 people with a diverse array of skills, capabilities and backgrounds. The team is organized into four pillars: Modelling and Data Mining, Research and Social Media, Visualization and Prototype, and Test and Learn. As far as turnover, absence, mobility and leadership development is concerned, there is a strong focus on monitoring customer experience and sales – linking HR data to wider business metrics.

Questions:

1. Why HR Analytics is important for Walmart?
2. How reporting, dashboards and data mining helps the Walmart?

CASE STUDY 4

NET INCOME ON A SPREADSHEET

The ESCON Limited is one of leading companies in IT industry. Its financial year statement calculation includes income and expenditure of the sales and its products every year are shown below. The company's calculation of net income is based on the following formulas which are given as mentioned below:

1. Gross Profit = Sales – Cost of Goods Sold
2. Operating Expenses = Administrative Expenses + Selling Expenses + Depreciation Expenses
3. Net Operating Income = Gross Profit – Operating Expenses
4. Earnings before Taxes = Net Operating Income – Interest Expense
5. Net Income = Earnings before Taxes – Taxes

Use above information and prepare spreadsheet through HR analytics concept.

CASE STUDY 5

HR METRICS

Regional Hospital is a 500-bed hospital and several associated clinics in a major East Coast metropolitan area. It has been an aggressive adopter of computing technologies in efforts to decrease costs and improve operational efficiencies. A critical challenge facing the hospital is

meeting its ongoing challenges to staff the hospital and allied clinics effectively, given the ongoing shortage of nurses; uncertainty in healthcare legislation; emphasis on shortening hospital stays to reduce costs, which causes the daily census (numbers of patients in various departments) to vary dramatically from day to day and shift to shift; the continued aging of the population in its primary care area; and the unending competition for employees with key skill sets. Employee expenses represent more than 80% of the overall costs of operation for the hospital. So, identifying ways to match optimal skills and numbers of employees to the appropriate shifts is critical to achieving consistent success. However, individual shift managers struggle to make effective staffing decisions, resulting in consistent overstaffing or understaffing of shifts and departments. These staffing problems potentially increase the high costs of varied levels of patient care and satisfaction and potentially increase the risk that staff turnover may escalate because of dissatisfaction with the continuing inability of managers to match staffing needs to demand. Company managers recognize the potential that HR metrics and analytics might have for their organization, and they have come to for help. They are hearing from their peers in other hospitals that metrics can help in this area but not quite sure where to start. They are looking for offering guidance on how to do HR metrics and workforce analytics.

Questions:

1. Do you believe that a program of HR metrics and workforce analytics might be useful in Regional Hospital? If so, why?
2. What opportunities do you see regarding where and how metrics and analytics might be applied in this organization?
3. Identify three analysis and associated metrics you think might be useful for Regional Hospital to consider.

REFERENCE BOOKS

- *Predictive HR Analytics Mastering the HR Metric by Martin Edwards* by Kirsten Edwards, 2nd Edition, Publisher: Kogan Page, Year 2019, ISBN 9780749484453.
- *Fundamentals of HR Analytics: A Manual on Becoming HR Analytical*, by Fermin Diez, Mark Bussin and Venessa Lee, Publisher: Emerald Publishing Limited, Year 2019, ISBN: 9781789739633.
- *HR Analytics* by Dipak Kumar Bhattacharyya.
- *Practical Applications of HR Analytics: A Step by Step Guide* by Pratyush Banerjee and Manish Gupta, Publisher: Sage Publications India Pvt. Ltd., Year July 2019, ISBN: 9789353282967.
- *Predictive HR Analytics: Mastering the HR Metric* by Dr. Martin Edwards and Kirsten Edwards, Paperback – 3rd March 2016.
- *People Analytics for Dummies* by Mike West, Released March 2019, Publisher: For Dummies, ISBN: 9781119434764.



ABOUT THE BOOK

Data about individuals are analysed using an approach called HR analytics. Based on HR analytics, more qualified candidates are found quickly, and better hiring choices are made, which aids in the expansion of the business. Additionally, an HR analytics specialist can choose better options for moving people to various functions and enhancing the organization's future potential based on previous data. The books on HR Analytics go into many methods for hiring, evaluating, and spotting possible promotion candidates, among other things. The HR Analytics course will also develop new skills, practical experience with various projects, and much more.

This book on HR Analytics is an excellent resource with circumstances and examples that make it simple to understand. Thus, it is a simple introduction to HR Analytics. Additionally, it is an ideal location for most HR professionals to start. However, this offers a comprehensive practical approach to employing data, dealing with actual HR difficulties, and elucidates analytics with appropriate guidelines and recommendations.

ABOUT THE AUTHORS



Dr. Rekha N. Patil is working as an Assistant Professor in VTU Regional Centre, Kalaburagi, Karnataka, India, and has 16 years of teaching experience for post-graduate and undergraduate courses. She has published more than 40 papers in National, International journals, and in SCOPUS. She has authored six books like International Human Resource Management, Recruitment and Selection, etc. as well as published five patents to her credit.



Mrs. Shilpa S. Hiregoudar was working as an Assistant Professor in the Department of MBA (Exclusively for Women) in Sharnbasva University Kalaburagi. She has obtained B.E. in Computer Science from B.L.D.E.A. Engineering College, Bijapur. She has pursued MBA from R.V. Institute of Management, Bengaluru, Karnataka, India. She has been teaching core subjects in Management Studies and in the area of Human Resource Management and Marketing Management for the last eight years. Currently, she is pursuing Ph.D. in Management from Visvesvaraya Technological University, Belagavi. She has published over 26 research papers in various National, International journals, SCOPUS, conferences and One Patent. Additionally, she has attended more than 30 workshops, Faculty Development Programs, and webinars at National and International levels. As a dedicated researcher, she strives for excellence and quality in the academic endeavours.

www.himpub.com



ISBN: 978-93-5693-618-8

ECH 0317