

Prepared by Dr. Wesam Ahmed

IT393
Business Intelligence

Lecture 3



Descriptive Analytics I:, Statistical Modeling , and Data Analytics

Lecture 3

IT393

Business Intelligence

Prepared by
Dr. Wesam Ahmed

01

Learning Objectives:

- 1.1** Learn about descriptive and inferential statistics
- 1.2** descriptive and inferential statistics using Excel
- 1.3** Define business reporting
- 1.4** The data analytics lifecycle

02

Measures of dispersion (May Also Be called measures of spread) :

Measures of dispersion are a representation of the numerical spread of a given data set.

To describe this dispersion, a number of statistical measures are developed; the most notable ones are range, variance, and standard deviation .

One of the main reasons why the measures of dispersion/spread of data values are important is the fact that it gives us a framework within which we can judge the central tendency: gives us the indication of how well the mean (or other centrality measures) represents the sample data. If the dispersion of values in the data set is large, the mean is not deemed to be a very good representation of the data.

For instance, consider a salary dataset: [15k, 18k, 16k, 14k, 15k, 15k, 12k, 17k, 90k, 95k]. The mean is approximately 30.7k, heavily influenced by the two high outliers (90k, 95k). This makes the mean unrepresentative of the typical salary, which lies mostly in the 12k–18k range.

03

Measures of dispersion (May Also Be called measures of spread) :

It is often perceived as a positive sign to see a small variation within each data sample, as it may indicate homogeneity, similarity, and robustness within the collected data. For instance, ages in a narrow range (e.g., 39–52) indicate similarity and robustness, making the data reliable for analysis.

Range: It is the difference between the largest and the smallest values in a given data set (i.e., variables).

$$(\text{range} = \text{maximum} - \text{minimum})$$

Example: For example, let us consider the following data set: 23 56 45 65 59 55 62 54 85 25

Sol: The maximum value is 85 and the minimum value is 23. This results in a range of 62. The range can be used to detect any errors when entering data. For example, if you have recorded the age of school children in your study and your range is 7 to 123 years old you know you have made a mistake!

04

Measures of dispersion (May Also Be called measures of spread) :

Variance: It is a method used to calculate the deviation of all data points in a given data set from the mean as shown in Eq.2.1.

The larger the variance, the more the data are spread out from the mean and the more variability one can observe in the data sample.

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

Where n is the number of samples, $\bar{x}$ is the mean of the sample and x_i is the i th value in the data set.

05

Measures of dispersion (May Also Be called measures of spread) :

Example with Small Variance: Consider the dataset: 28, 29, 30, 31, 32. The mean is 30, and the variance is 2, showing the data points are closely clustered, indicating low variability.

Example with Large Variance: Now, take 10, 20, 30, 40, 100. The mean is 40, and the variance is 980, reflecting high variability due to the outlier (100), showing the data is widely spread.

Dataset	Mean	Variance	Interpretation
28, 29, 30, 31, 32	30	2	Low variance, data points closely clustered, low variability
10, 20, 30, 40, 100	40	980	High variance, data points widely spread, high variability

06

Measures of dispersion (May Also Be called measures of spread) :

Standard Deviation : The standard deviation is also **a measure of the spread of values within a set of data.**

The standard deviation is calculated by **simply taking the square root of the variations.**

The **following formula** shows the calculation of standard deviation from a given sample of data points.

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

07

Measures of dispersion (May Also Be called measures of spread) :

Let's calculate the variance and STD of the following data set:

Data Point	Value	Difference from Mean (Value - 6.6)	Squared Difference
1	2	-4.6	21.16
2	7	0.4	0.16
3	3	-3.6	12.96
4	12	5.4	29.16
5	9	2.4	5.76
Sum of Squared Differences			69.20
Variance			13.84

Standard Deviation
= 3.72

08

How to do analysis regression in Excel with Analysis ToolPak

In your Excel, click **File > Options**.

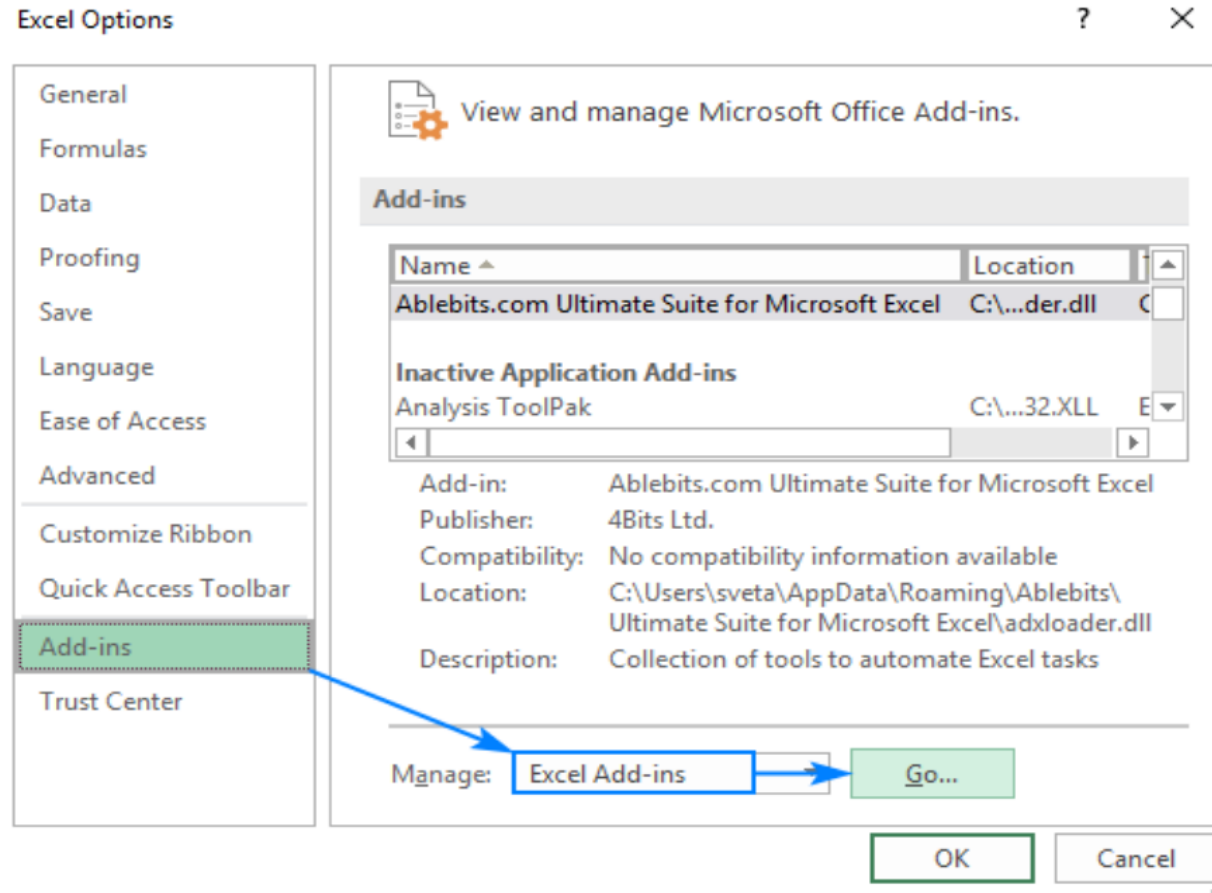
In the Excel Options dialog box, select **Add-ins** on the left sidebar, make sure **Excel Add-ins** is selected in the **Manage box**, and click **Go**

In the Add-ins dialog box, **tick off Analysis Toolpak**, and click **OK**:

This will add the Data Analysis tools to the Data tab of your Excel ribbon

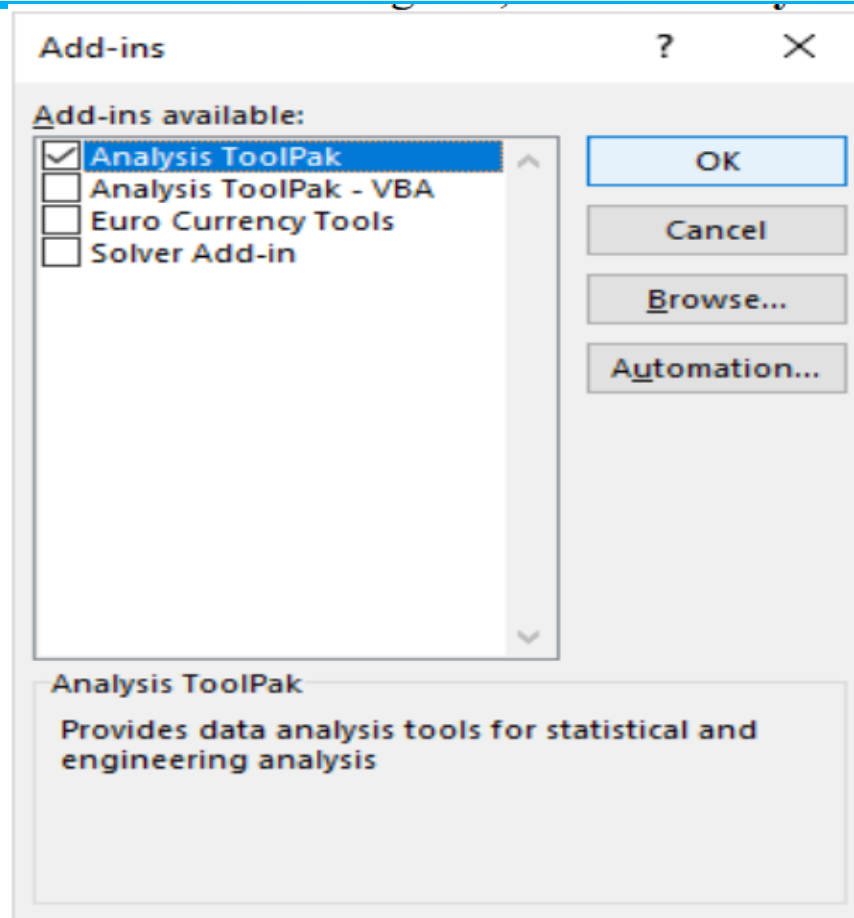
09

How to do analysis in Excel with Analysis ToolPak



10

How to do analysis in Excel with Analysis ToolPak



11

Measures of dispersion (May Also Be called measures of spread) :

Let's calculate **descriptive statistics** of the follow data set using **Excel**:

Scores
28
29
40
50
60
70
80
120
130
130

Excel Dataset

Statistics	
Mean	73.7
Standard Error	12.69649996
Median	65
Mode	130
Standard Deviation	40.14985817
Sample Variance	1612.011111
Kurtosis	-1.426043123
Skewness	0.466620704
Range	102
Minimum	28
Maximum	130
Sum	737
Count	10

Statistics of the Dataset

- Go to the **Data** tab.
- Click on **Data Analysis** in the Analysis group.
- In the **Data Analysis** dialog box, scroll to and select **Descriptive Statistics**.

12

Regression Modeling for Inferential Statistics:

Regression, especially linear regression, is perhaps the most widely known and used analytics technique in statistics. Regression has become the statistical technique for characterization of relationships between explanatory (input) variable(s) and response (output) variable.

Regression aims to capture the functional relationship between and among the characteristics of the real world and describe this relationship with a mathematical model, which may then be used to discover and understand the complexities of reality—explore and explain relationships or forecast future occurrences.

In prediction, regression identifies additive mathematical relationships (in the form of an equation) between one or more explanatory variables and a response variable. Once determined, this equation can be used to forecast the values of the response variable for a given set of values of the explanatory variables.

13

Regression Modeling for Inferential Statistics:

Simple versus Multiple Regression : If the regression equation is built **between one response variable and one explanatory variable**, then it is called **simple regression**.

For instance, the regression equation built to predict/explain the relationship between **a height of a person (explanatory variable)** and **the weight of a person (response variable)** is a good example of **simple regression**.

Multiple regression is the extension of simple regression where the **explanatory variables are more than one**.

For instance, in the previous example, if we were to include not only the height of the person but **also other personal characteristics (e.g., BMI, gender, ethnicity)** to **predict the weight of a person**, then we would be performing **multiple regression analysis**. In both cases, **the relationship** between the response variable and the explanatory variable(s) **are linear and additive in nature**.

14

Regression Modeling for Inferential Statistics:

To understand the relationship between two variables, the simplest thing that one can do is to **draw a scatter plot**, where the **y-axis represents the values of the response variable** and the **x-axis** represents the values of the **explanatory variable** (see Figure 3.1).

A scatter plot would **show the changes in the response variable as a function of the changes in the explanatory variable**.

In the case shown in Figure 3.1, there seems to be **a positive relationship between the two**; as the explanatory variable values increase, so does the response variable. Simple regression analysis aims to find **a mathematical representation of this relationship**.

15

Regression Modeling for Inferential Statistics:

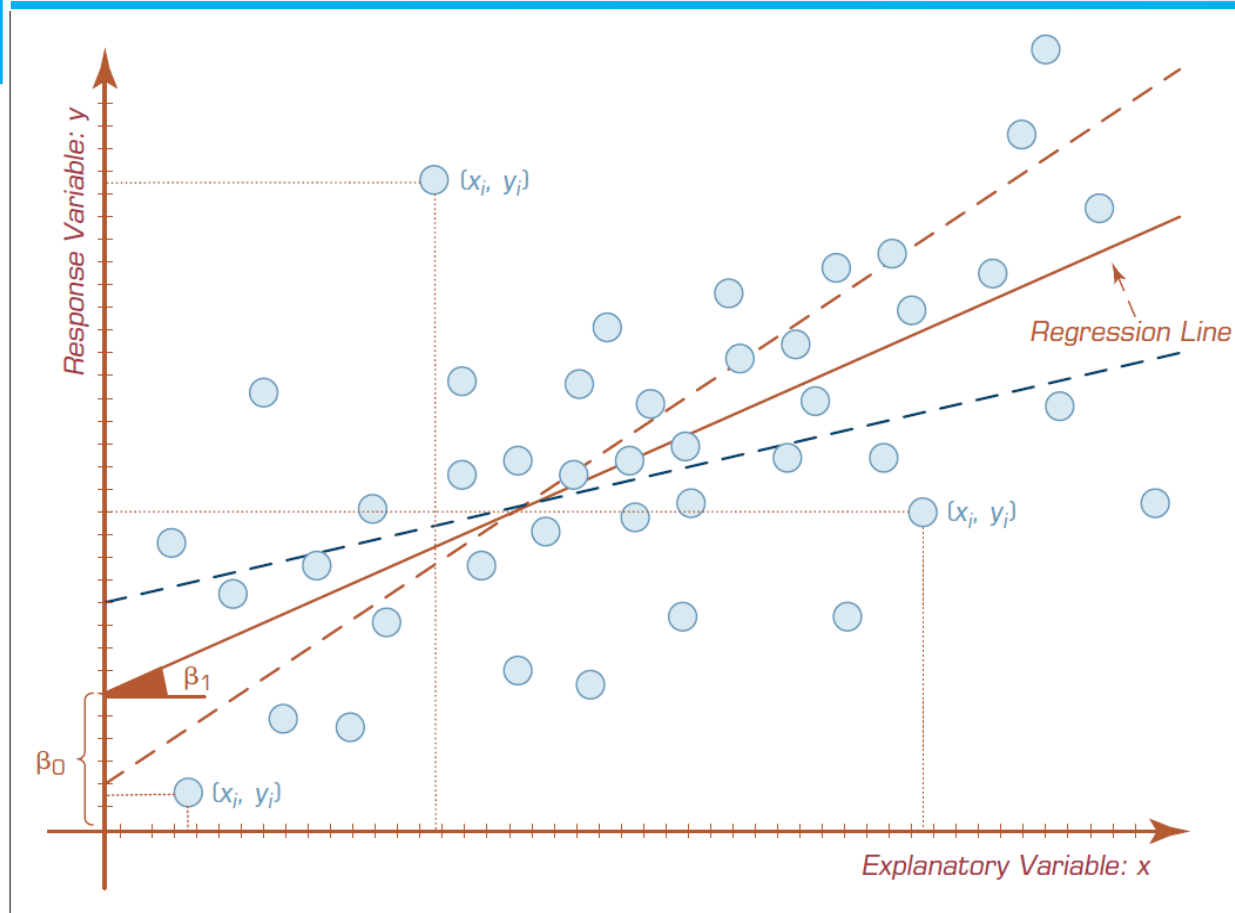


Figure 3.1 A Scatter Plot and a Linear Regression Line.

16

Regression Modeling for Inferential Statistics using Excel:

Month	Rainfall	Umbrella sold
January	78	120
February	64	95
March	92	135
April	48	70
May	35	60
June	20	40
July	15	30

In this example, we are going to do a simple linear regression in Excel. What we have is a list of **average monthly rainfall for some months in column B**, which is our independent variable (predictor), and **the number of umbrellas sold in column C**, which is the dependent variable.

17

Regression Modeling for Inferential Statistics using Excel:

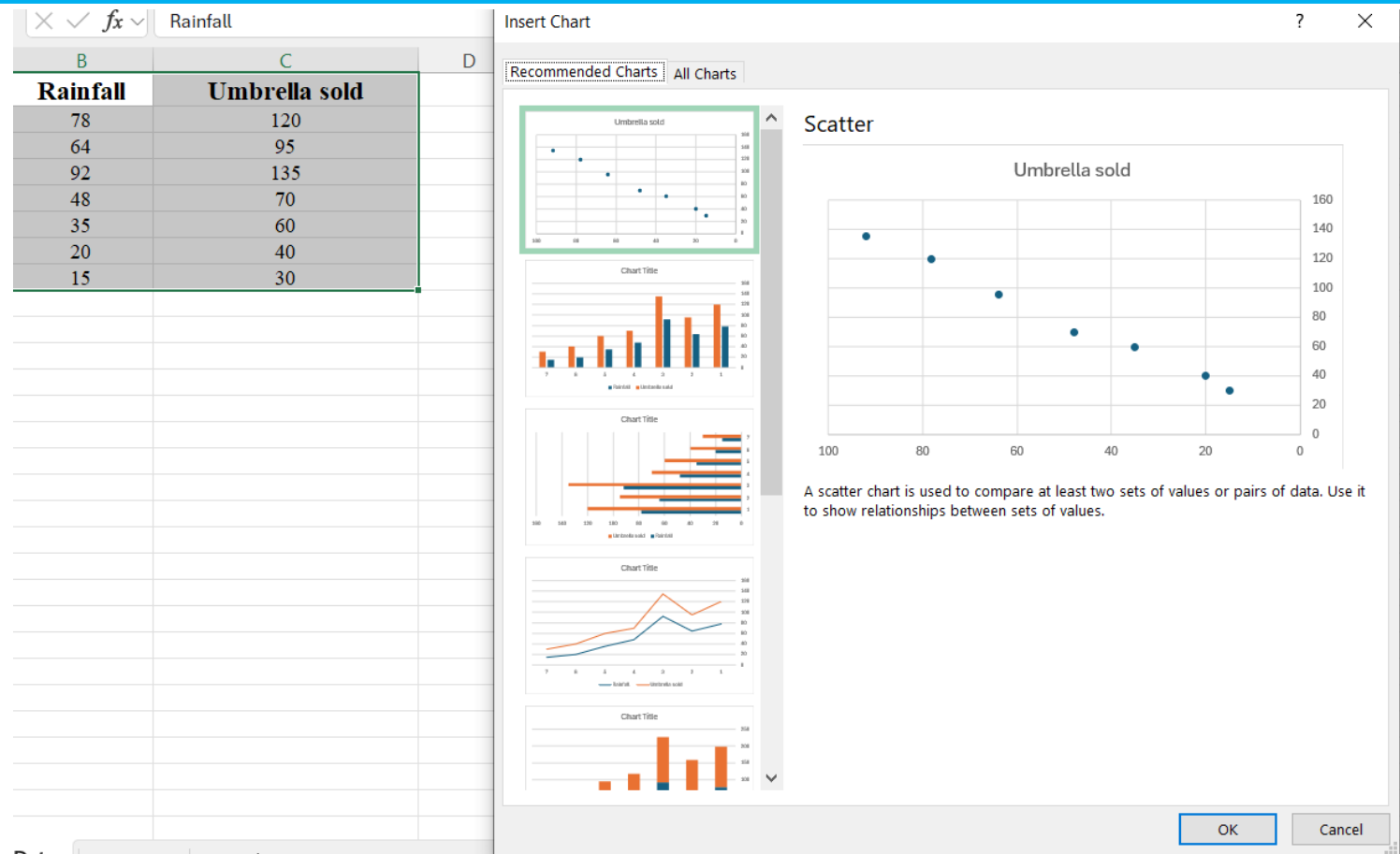
If you need to quickly visualize the relationship between the two variables, draw a linear regression chart. That's very easy! Here's how:

Select the two columns with your data, including headers.

On the **Inset** tab, in the **Charts** group, click the **Scatter** chart icon, and select the **Scatter** thumbnail (the first one):

18

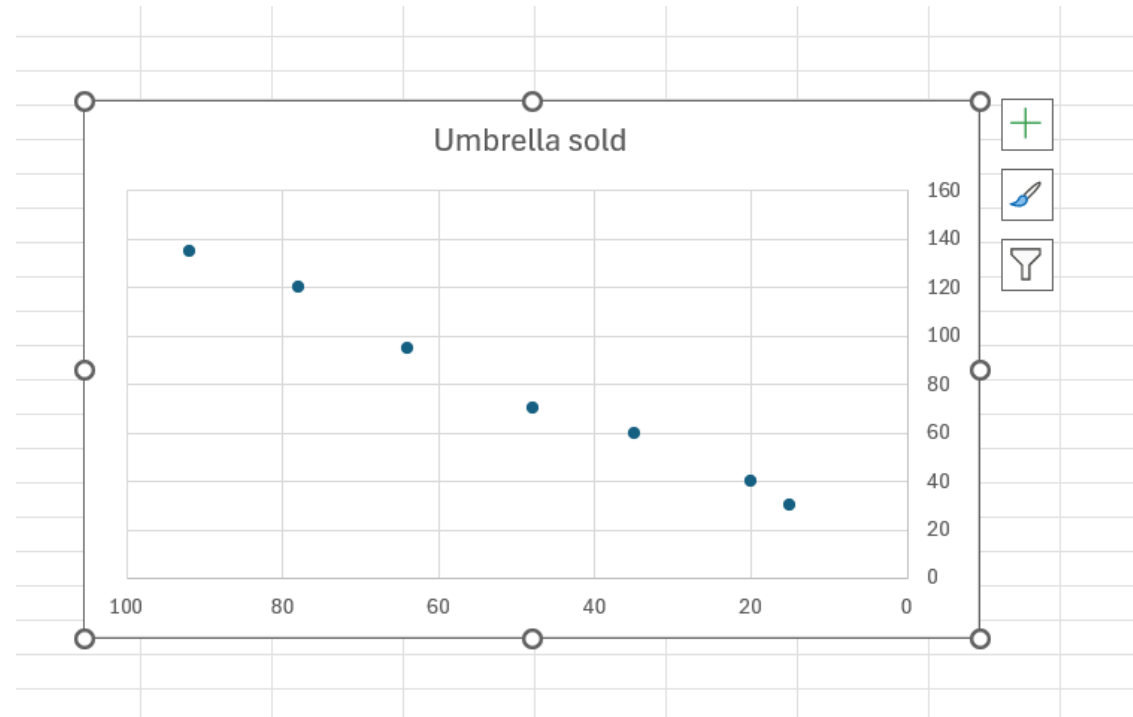
Regression Modeling for Inferential Statistics using Excel:



19

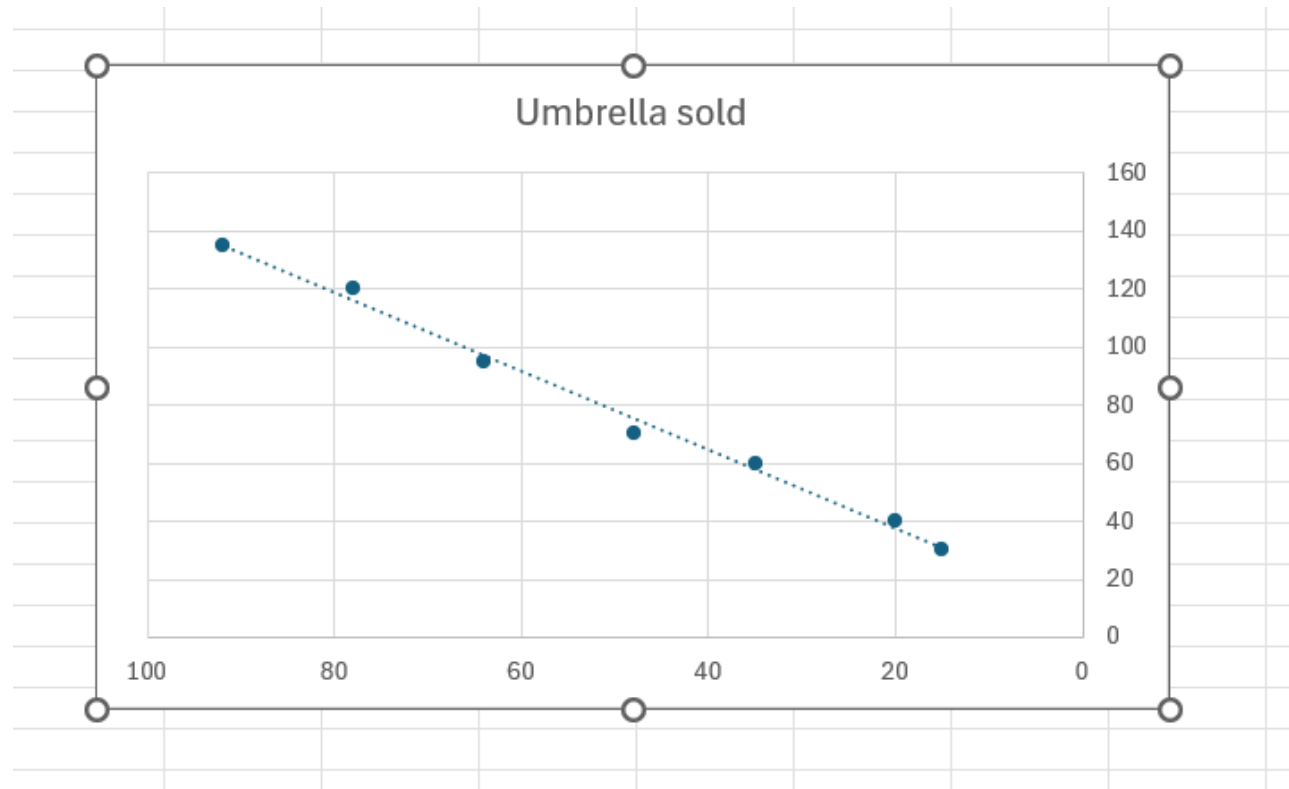
Regression Modeling for Inferential Statistics using Excel:

This will insert a scatter plot in your worksheet, which will resemble this one:



20

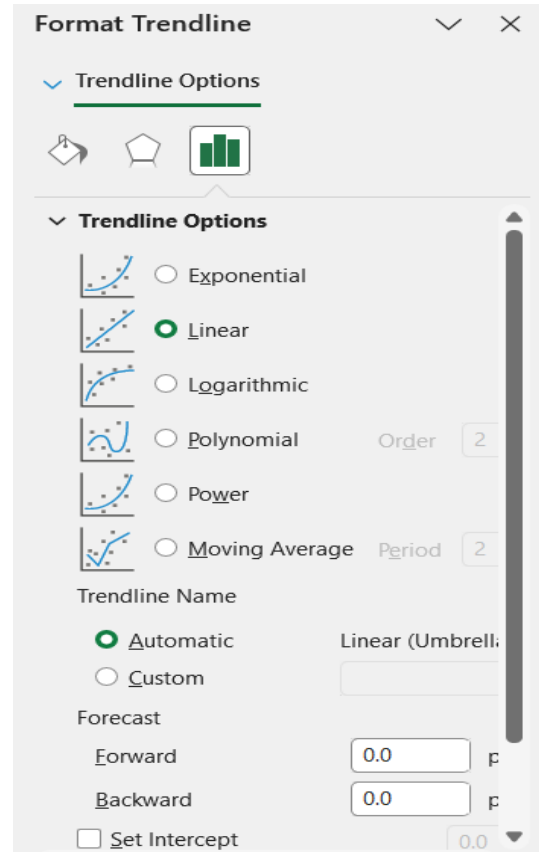
Regression Modeling for Inferential Statistics using Excel:



Now, we need to draw the least squares regression line. To have it done, right click on any point and choose **Add Trendline...** from the context menu.

21

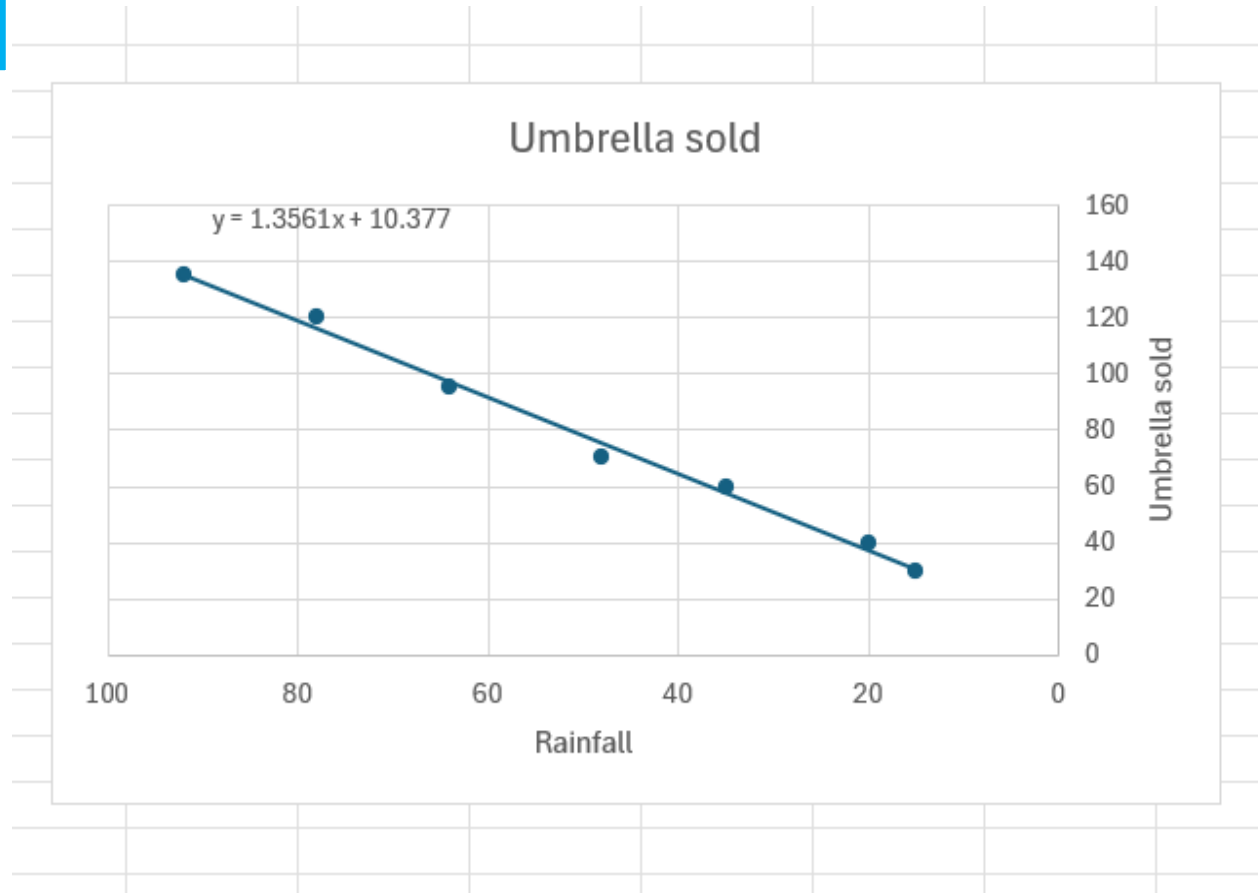
Regression Modeling for Inferential Statistics using Excel:



On the right pane, select the **Linear** trendline shape and, optionally, check **Display Equation on Chart** to get your regression formula:

22

Regression Modeling for Inferential Statistics using Excel:



You may want to make a few more improvements:

Drag the equation wherever you see fit.

Add axes titles (*Chart Elements* button > *Axis Titles*).

23

Business Reporting:

Decision makers are in need of **information to make accurate and timely decisions**. Information is essentially the contextualization of data. In addition to statistical means that were explained in the previous section. The information is usually provided to the decision makers in the form of a written report (digital or on paper), although it can also be provided orally.

Simply put, **a report is any communication artifact prepared with the specific intention of conveying information** in a digestible form to whoever needs it, whenever and wherever they may need it.

It is usually **a document that contains information (usually driven from data) organized in a narrative**, graphic, and/or tabular form, prepared periodically (recurring) or on an as-needed (ad hoc) basis, referring to specific time periods, events, occurrences, or subjects.

24

Business Reporting:

Business reports can fulfill many different (but often related) functions. Here are a few of the most prevailing ones:

- To ensure that all departments are functioning properly
- To provide information
- To provide the results of an analysis

Business reporting (also called BI) is an essential part of the larger drive toward improved, evidence-based, optimal managerial decision making. The foundation of these business reports is various sources of data coming from both inside and outside the organization .

25

Business Reporting:

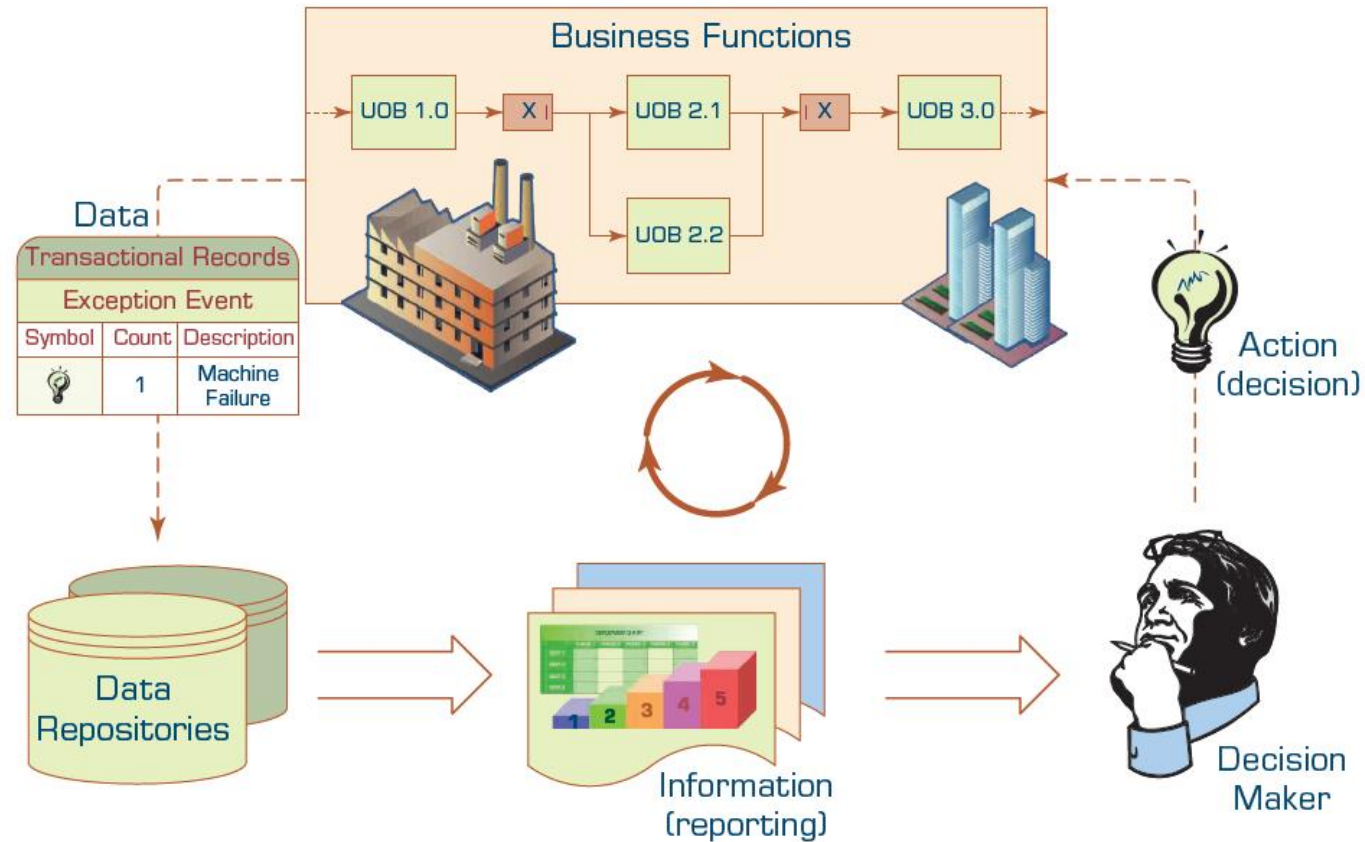


Figure 3.2 The Role of Information Reporting in Managerial Decision Making.

26

Business Reporting:

Figure 3.2 shows the continuous cycle of data acquisition → information generation → decision making → business process management.

Perhaps the most critical task in this cyclical process is the reporting (i.e., information generation)—converting data from different sources into actionable information.

Key to any successful report are clarity, brevity, completeness, and correctness.

27

Business Reporting:

Even though there are a wide variety of business reports, the ones that are often used for managerial purposes can be grouped into **two major categories**:

Metric Management Reports: In many organizations, **business performance is managed through outcome-oriented metrics**. For internal management, they are **key performance indicators (KPIs)**.

KPIs are specific, measurable values that indicate how effectively a company is achieving key business objectives. They are a subset of metrics that matter most for tracking success.

Example KPIs:

- Customer Satisfaction Score (CSAT)
- Employee Turnover Rate
- Net Profit Margin

28

Business Reporting:

Dashboard -Type Reports: A popular idea in business reporting in recent years has been to present a range of different performance indicators on one page, like a dashboard in a car.

Typically, **dashboard vendors would provide a set of predefined reports with static elements and fixed structure**, but also allow for customization of the dashboard widgets, views, and set targets for various metrics.

It's common to have **color-coded traffic lights defined for performance** (red, orange, green) to draw management's attention to particular areas.

29

Business Reporting:



Figure 3.3:
Example of Dashboard -Type
Reports

30

The data analytics lifecycle:

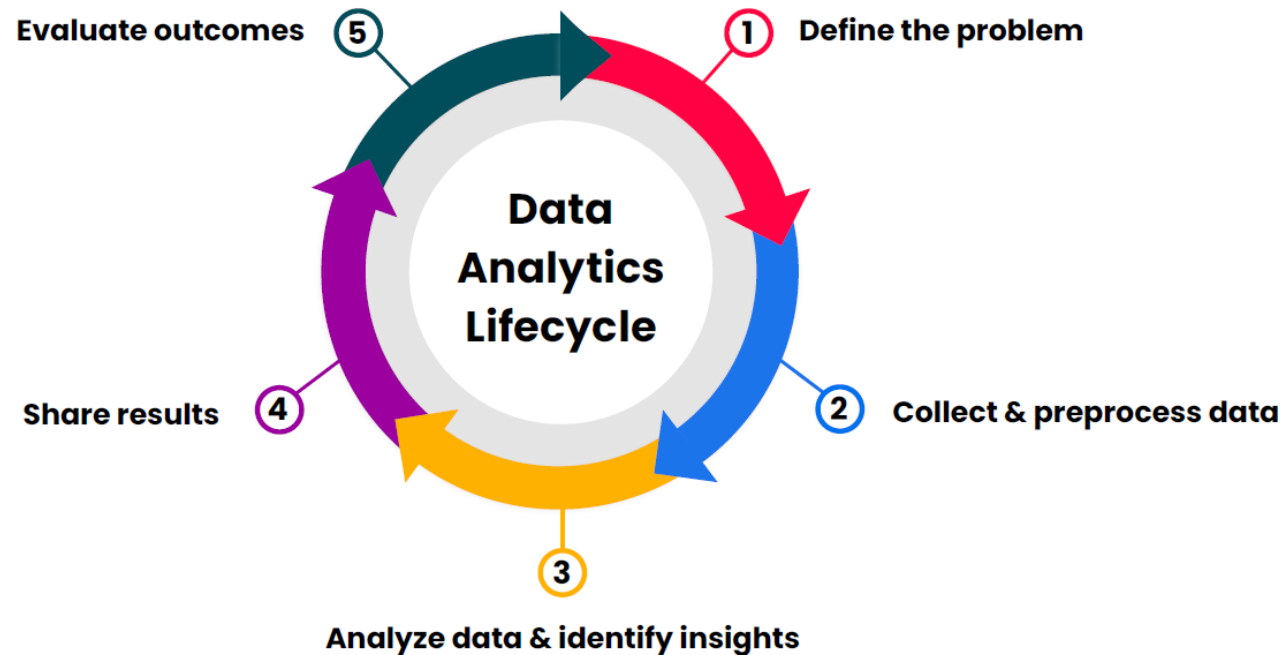


Figure 3.4: The data analytics lifecycle

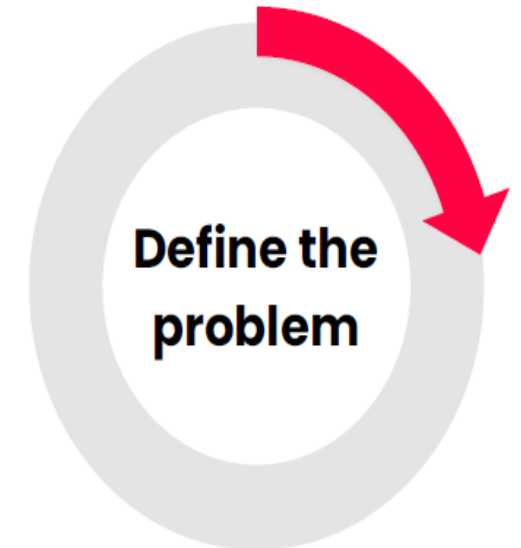
31

The data analytics lifecycle:

The first phase of the analytics cycle is about **understanding the problem that needs to be solved**. This includes identifying the **business objectives**, the **available data**, and **the resources needed to solve the problem**.

Key questions:

- What are the business goals?
- What decisions need to be made to achieve those goals?
- Who are the stakeholders making those decisions, and what are their needs?



32

The data analytics lifecycle:

① What are the business goals?

Most businesses:



Increase sales



Reduce costs



Improve satisfaction

More specific objective:

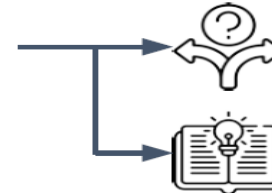


Better product recommendations

② Who are the stakeholders and what are their needs?



Who will be using your results?



Decisions

Information

③ What are the key unknowns?

Examples:



Most effective marketing channels



Cancellation factors

33

The data analytics lifecycle:

Descriptive Analytics Example:

Let's look at an example. Let's say you want to **understand your company's revenue drivers**. You perform descriptive analytics on your sales data, answering questions such as:

- **How many sales were made in the last month? (The count of sales.)**
- **How much is the average transaction value? (The mean sale price.)**
- **How much variation is there between total sales per store? (The standard deviation of total sales by store.)**
- **What were the total sales per day? (A line plot of total sales versus date.)**

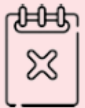
34

The data analytics lifecycle:

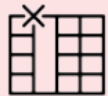
Stage 2: Collect & preprocess data

- Transform raw data into **usable information**
- Most data requires **some transformation**

Challenges



Missing
information



Poor
formatting



Unavailable
data

Quality of **data**



Quality of **insights**



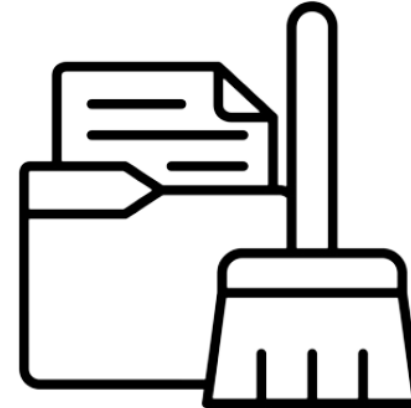
35

The data analytics lifecycle:



Step 1: Data collection

Identify sources of information



Step 2: Data preprocessing

Transforming your data for analysis

Figure 3.4 Stage 2: Collect & preprocess data.

36

The data analytics lifecycle:



Figure 3.5 Strategies for data collection

37

The data analytics lifecycle:

Raw data



**Data ready
for analysis**

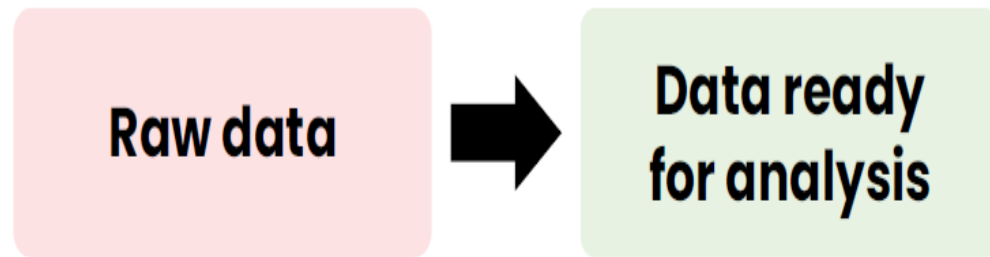
2021/06/15
May 20, 2020
2019-08-10

15-06-2021
20-05-2020
10-08-2019

Figure 3.6 Data preprocessing
(Formatting)

38

The data analytics lifecycle:



John Doe	25	2021/06/15
John Doe	25	2021/06/15

John Doe	25	2021/06/15
----------	----	------------

Figure 3.7 Data preprocessing
(Cleaning)

39

The data analytics lifecycle:

Raw data



**Data ready
for analysis**

Emily	28	UNKNOWN
-------	----	---------

Emily	28	UNKNOWN
------------------	---------------	--------------------

Figure 3.8 Data preprocessing
(Handling missing values)

40

The data analytics lifecycle:

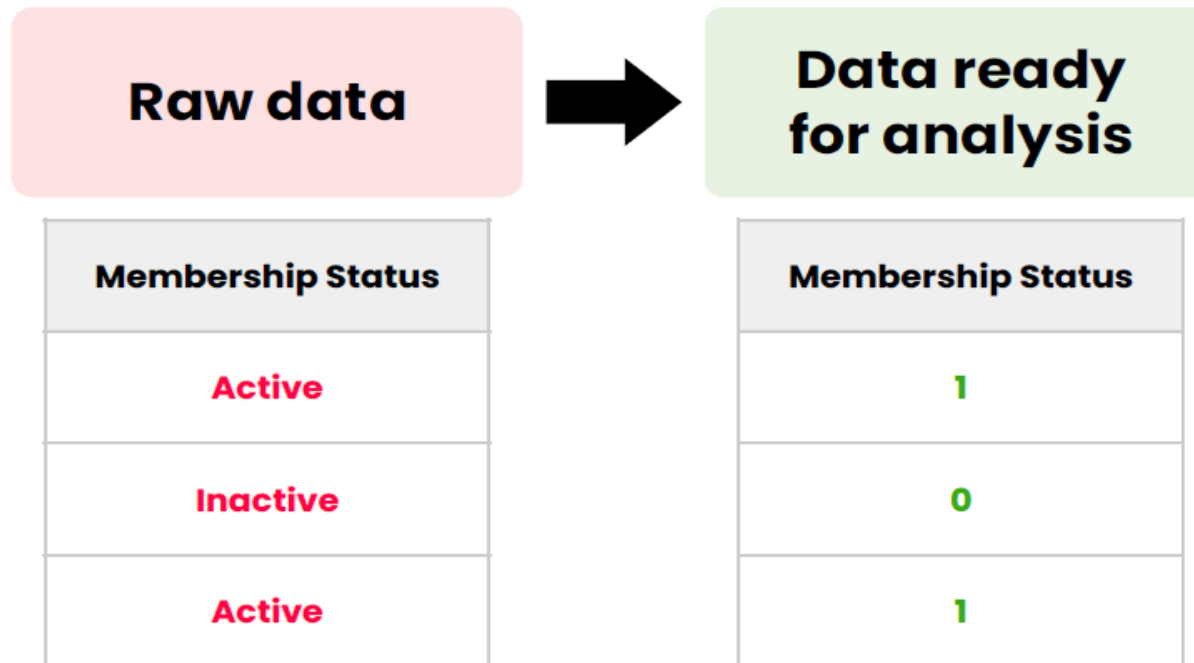


Figure 3.9 Data preprocessing
(Transformation)

Summary

1. Descriptive statistics methods can be used to measure central tendency, dispersion, or the shape of a given data set.
2. Regression, especially linear regression, is perhaps the most widely known and used analytics technique in statistics.
3. A report is any communication artifact prepared with the specific intention of conveying information in a presentable form.
4. A business report is a written document that contains information regarding business matters.
5. The data analytics lifecycle.



Thank You