

Prepared by Dr. Wesam Ahmed

IT393
Business Intelligence

Lecture 6



Predictive Analytics I: Data Mining

Lecture 6

IT393

Business Intelligence

Prepared by
Dr. Wesam Ahmed

01

Learning Objectives:

- 1.1 Using Excel for data analytics
- 1.2 Understand the importance of LLM for data exploration and visualization
- 1.3 Define data mining as an enabling technology for business analytics
- 1.4 Understand the objectives and benefits of data mining

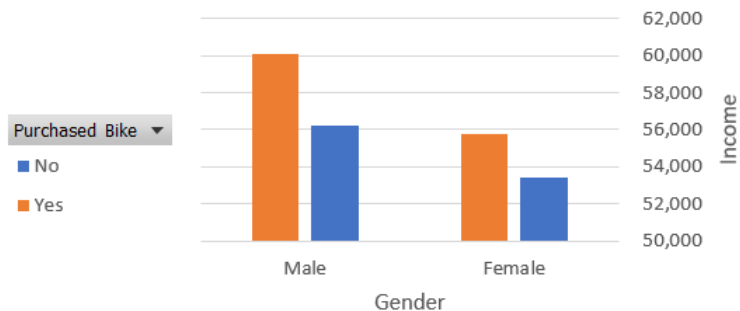
02

Using Excel for data analytics:

Bike Dashboard

Average of Income

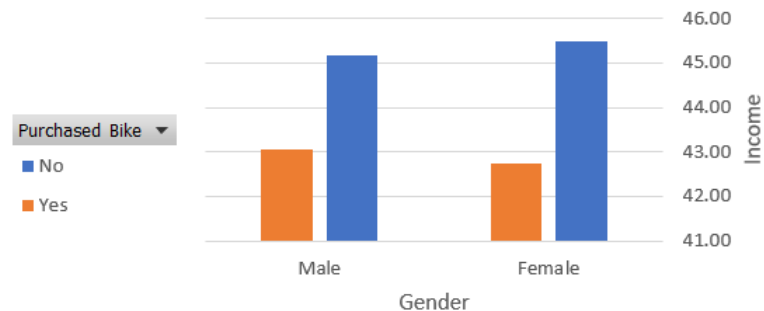
Average Income per Purchase



Gender ▼

Average of Age

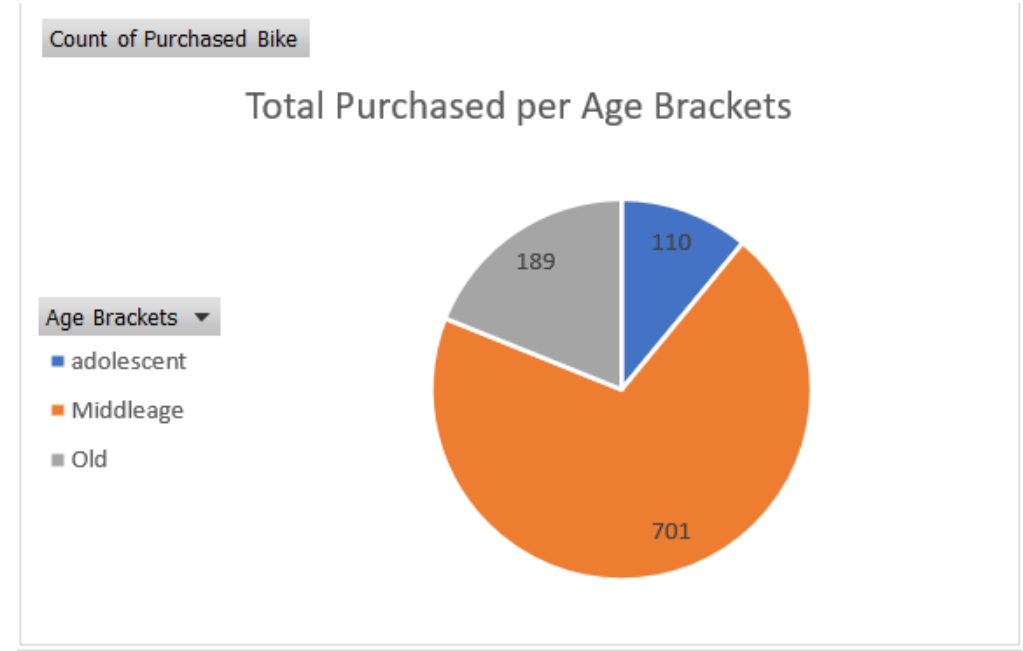
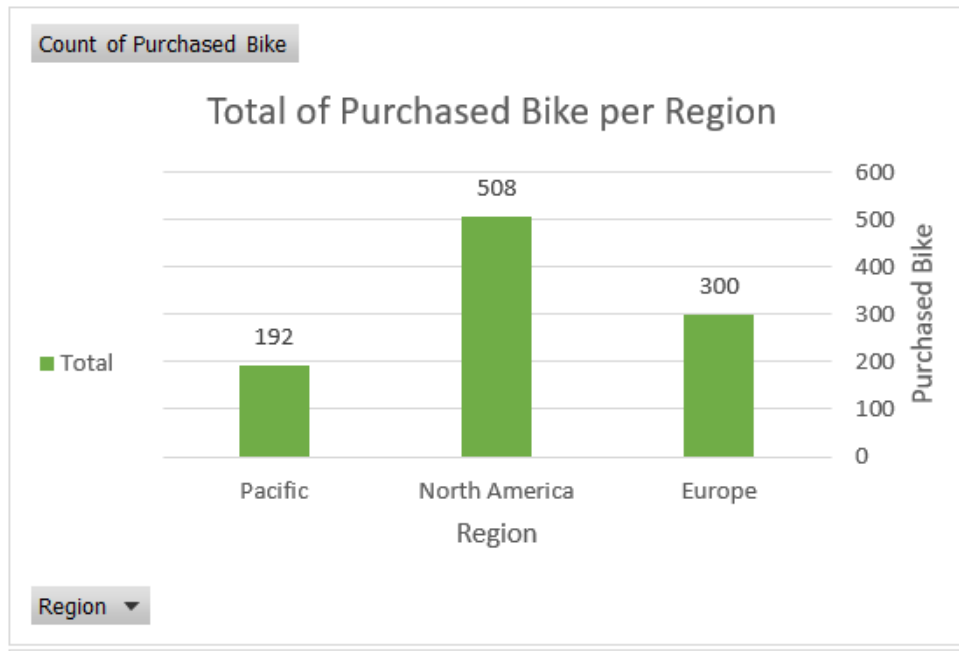
Avg Age of Purchase



Gender ▼

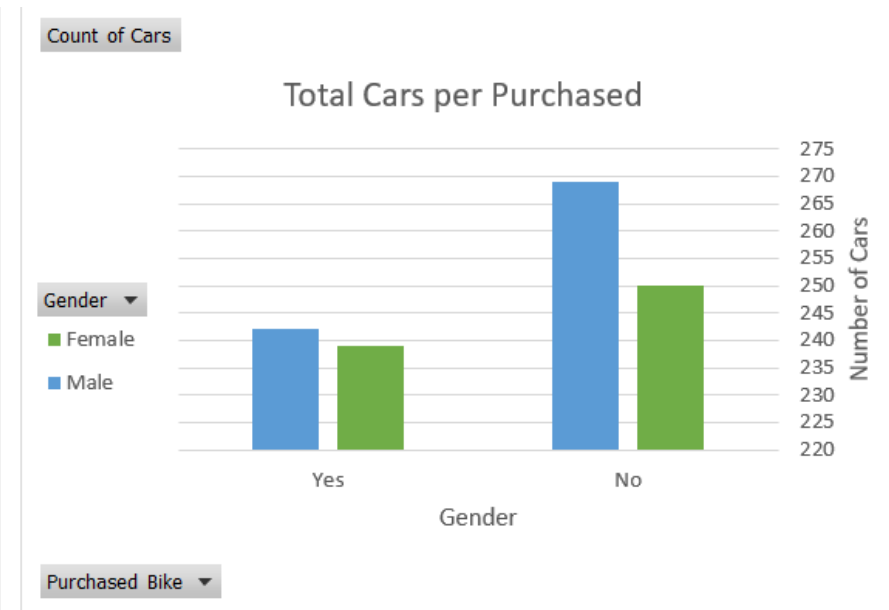
03

Using Excel for data analytics:



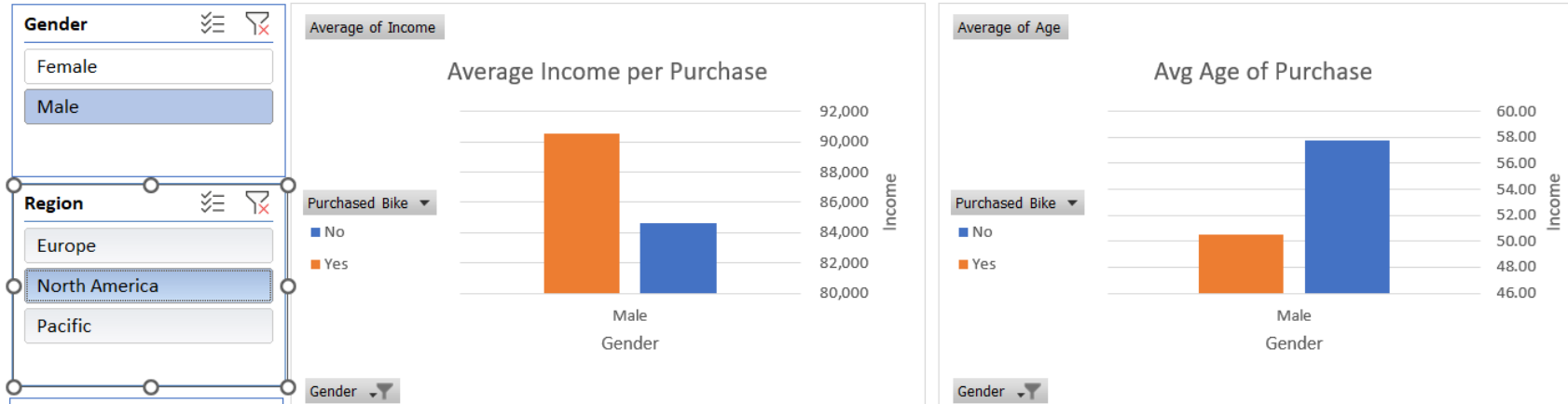
04

Using Excel for data analytics:



05

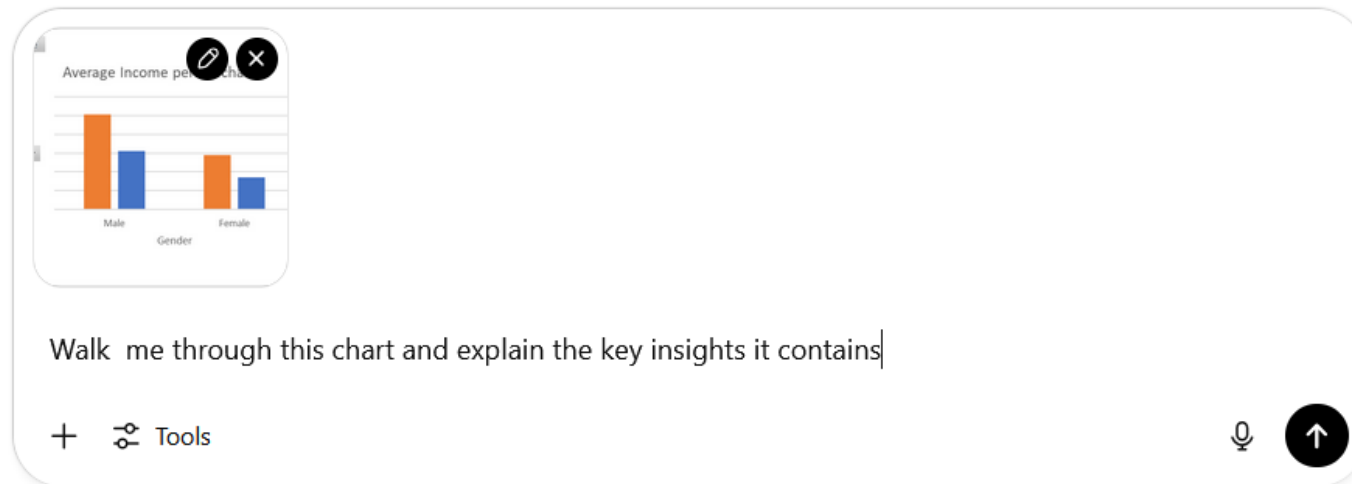
Using Excel for data analytics:



06

Using Excel and LLM for data analytics:

Ready when you are.



07

Using Excel and LLM for data analytics:

This chart titled "**Average Income per Purchase**" presents the average income of individuals segmented by **Gender** and whether they **purchased a bike** or not. Here's a breakdown of its components and insights:

**Chart Breakdown**

- **X-Axis (Horizontal):** Represents **Gender** (Male, Female).
- **Y-Axis (Vertical):** Represents **Income** (in a range from \$50,000 to \$62,000).
- **Legend (Color Code):**
 - **Orange bar:** Individuals who **purchased a bike**.
 - **Blue bar:** Individuals who **did not purchase a bike**.

08

Using Excel and LLM for data analytics:

Key Observations

1. Income & Purchase Behavior by Gender

- **Males:**
 - Those who purchased a bike had the **highest average income**, around **\$61,000**.
 - Those who did not purchase had a lower average income, around **\$56,000**.
 - **◆ Insight:** Among males, higher income correlates with a higher likelihood of bike purchase.
- **Females:**
 - Purchasers had an average income of about **\$55,000**.
 - Non-purchasers had an average income around **\$53,000**.
 - **◆ Insight:** For females, there is also a positive correlation between income and bike purchase, but the income gap is smaller compared to males.

2. Gender Comparison

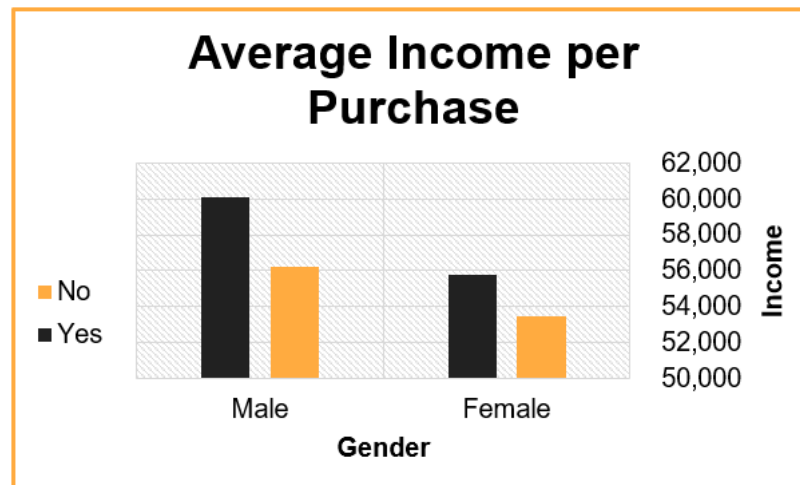
- Males generally have **higher average incomes** than females in both purchasing and non-purchasing categories.
- The **income gap between purchasers and non-purchasers** is more pronounced in males (\$5,000) than in females (\$2,000).



09

Using Excel and LLM for data analytics:

What is the average income for each gender among people who purchased a bike?



This chart titled "**Average Income per Purchase**" presents the average income of individuals segmented by **Gender** and whether they **purchased a bike** or not.

In the males, Those who purchased a bike had the highest average income, around \$61,000. Those who did not purchase had a lower average income, around \$56,000. Among males, higher income correlates with a higher likelihood of bike purchase.

In the females, **Females:** Purchasers had an average income of about \$55,000. Non-purchasers had an average income around \$53,000. For females, there is also a positive correlation between income and bike purchase, but the income gap is smaller compared to males.

Males generally have **higher average incomes** than females in both purchasing and non-purchasing categories.

10

Using Excel and LLM for data analytics:



Walk me through this chart and explain the key insights it contains

11

Using Excel and LLM for data analytics:

This chart titled "**Avg Age of Purchase**" compares the **average age** of individuals who **did or did not purchase a bike**, broken down by **gender**.



Chart Breakdown

- **X-Axis (Horizontal):** Represents **Gender** (Male, Female).
- **Y-Axis (Vertical):** Although incorrectly labeled as "Income", it actually represents **Age** based on the title and context.
- **Legend (Color Code):**
 - **Orange bars:** People who **purchased a bike**.
 - **Blue bars:** People who **did not purchase a bike**.

12

Using Excel and LLM for data analytics:

 Key Observations

1. Age & Bike Purchase Behavior by Gender

- **Males:**
 - Purchasers have an **average age around 43 years**.
 - Non-purchasers have a **higher average age, about 45 years**.
 - **◆ Insight:** Younger males are more likely to purchase a bike.
- **Females:**
 - Purchasers average about **42.5 years**.
 - Non-purchasers average about **45.5 years**.
 - **◆ Insight:** The same pattern holds — younger females are more likely to buy bikes.

2. Cross-Gender Comparison

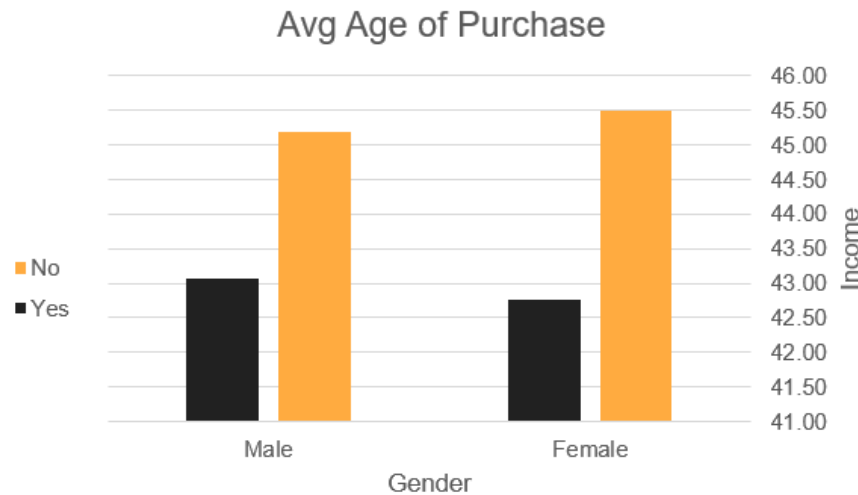
- In both genders, the **average age of those who purchased bikes is lower** than those who did not.
- The **age gap** between purchasers and non-purchasers is roughly:
 - ~2 years for males
 - ~3 years for females



13

Using Excel and LLM for data analytics:

What is the average age of individuals who purchased a bike?



In the males: Purchasers have an average age around 43 years. Non-purchasers have a higher average age, about 45 years. Younger males are more likely to purchase a bike.

In the females: Purchasers average about 42.5 years. The non-purchasers average about 45.5 years. The same pattern holds — younger females are more likely to buy bikes.

In both genders, the average age of those who purchased bikes is lower than those who did not.

14

Using Excel and LLM for data analytics:



Excel Project Dataset.xlsx
Spreadsheet

what is the dataset about?

Key Features in the Dataset

Column Name	Description
ID	Unique identifier for each individual
Marital Status	Whether the person is Single or Married
Gender	Male or Female
Income	Annual income in dollars
Children	Number of children in the household
Education	Education level (e.g., Bachelors, Partial College)
Occupation	Job category (e.g., Clerical, Professional)
Home Owner	Whether the individual owns a home (Yes/No)
Cars	Number of cars owned
Commute Distance	Daily commuting distance to work
Region	Geographic re ↓ (e.g., Europe, Pacific)

15

Using Excel and LLM for data analytics:

what is the number of features and rows in the dataset?

The dataset contains:

- ✓ 213 rows (individual records)
- ✓ 14 features (columns)

16

Using Excel and LLM for data analytics:

Are there any missing values?

If you're using Excel, you can check for missing values by doing the following:

1. **Select the dataset.**
2. Press `Ctrl + G` → click **Special** → choose **Blanks** → click **OK**.
3. Excel will highlight all blank cells.

17

Using Excel and LLM for data analytics:

Help me visualize the average income for each gender who purchased bike?

Create a Bar Chart for Average Income by Gender (Bike Purchasers)

1. Filter the Data:

- Select your dataset.
- Apply a filter (via `Data` tab → `Filter`).
- Filter `Purchased Bike` column to show only `Yes`.

2. Create a Pivot Table:

- Go to `Insert` → `PivotTable`.
- Use the filtered data.
- Set:
 - **Rows:** `Gender`
 - **Values:** `Income` (use "Average" instead of "Sum")

3. Insert a Bar Chart:

- Select the pivot table.
- Go to `Insert` → `Column or Bar Chart` → choose a bar chart.

4. Customize (Optional):

- Add chart title: "Average Income by Gender (Bike Purchasers)"
- Label axes appropriately.

18

Predictive Analytics : Data Mining

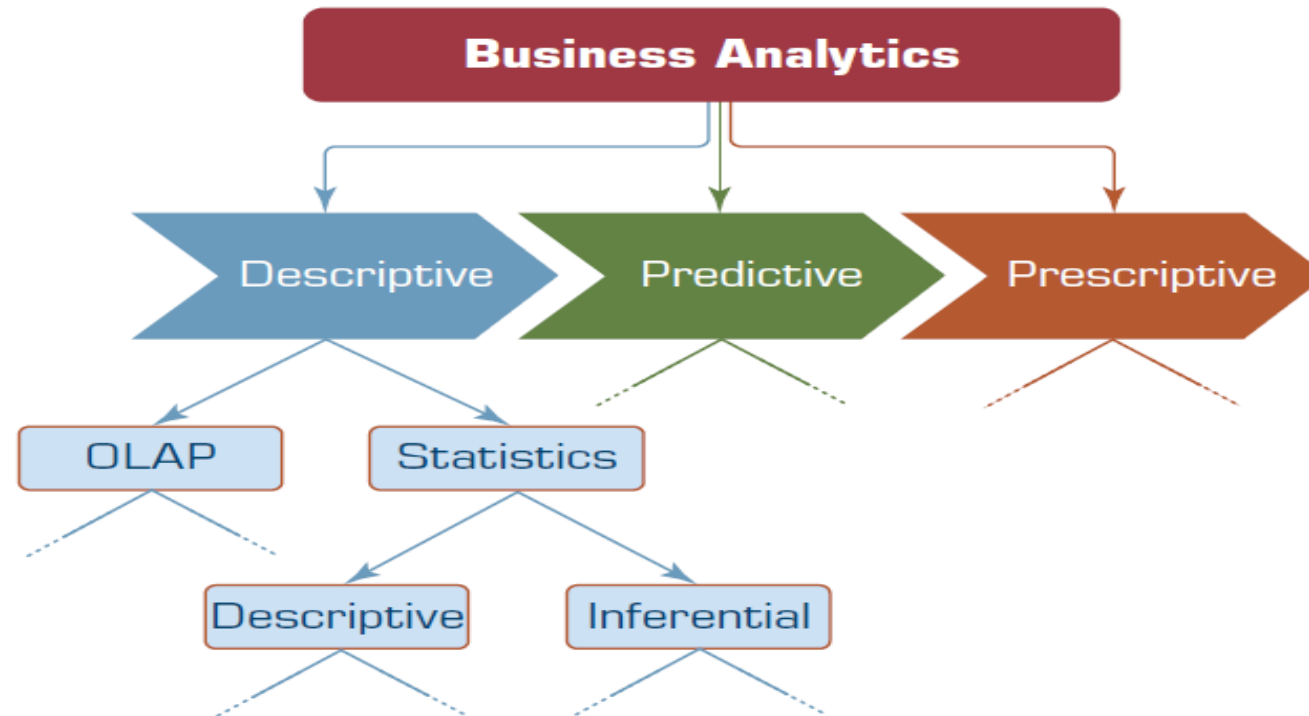


Figure 6.1: Types of business analytics.

19

Predictive Analytics : Data Mining

Predictive analytics is a branch of **advanced analytics** that makes predictions about future outcomes using historical data combined with statistical modeling, data mining techniques and machine learning.



20

Predictive Analytics : Data Mining

Companies employ predictive analytics to find patterns in this data to identify risks and opportunities. Predictive analytics is often associated with big data and data science. Today, companies today are inundated with data from log files to images and video. All of this data resides in disparate data repositories across an organization. To gain insights from this data, data scientists use deep learning and machine learning algorithms to find patterns and make predictions about future events. Some of these statistical techniques include logistic and linear regression models, neural networks and decision trees. Some of these modeling techniques use initial predictive learnings to make additional predictive insights.

21

Predictive Analytics : Data Mining

Predictive analytics can be deployed in across various industries for different business problems:

- Banking: Financial services use machine learning and quantitative tools to make predictions about their prospects and customers. With this information, banks can answer questions like who is likely to default on a loan, which customers pose high or low risks, which customers are the most lucrative to target resources and marketing spend.
- Healthcare: Predictive analytics in health care is used to detect and manage the care of chronically ill patients, as well as to track specific infections. Geisinger created a predictive model based on health records for more than 10,000 patients who had been diagnosed with sepsis in the past. The model yielded impressive results, correctly predicting patients with a high rate of survival.

22

Predictive Analytics : Data Mining

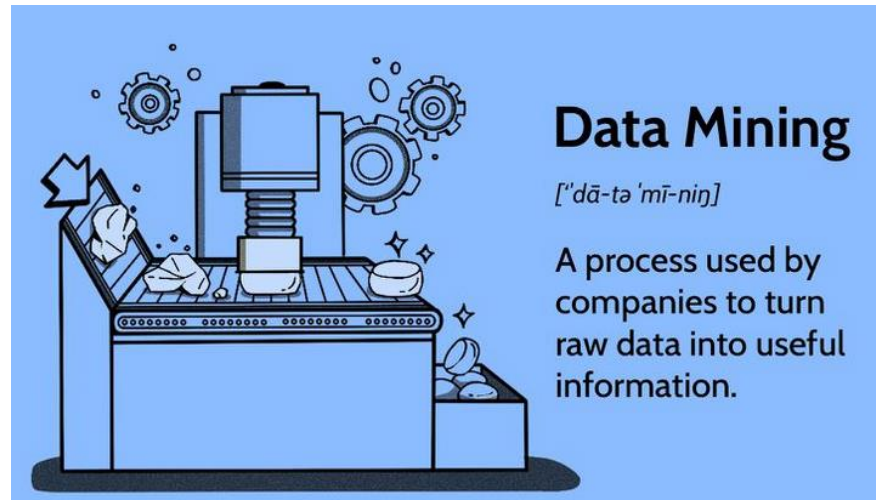
Predictive analytics can be deployed in across various **industries for different business problems:**

- **Human resources (HR):** HR teams use predictive analytics and employee survey metrics to match prospective job applicants, reduce employee turnover and increase employee engagement. This combination of quantitative and qualitative data allows businesses to reduce their recruiting costs and increase employee satisfaction, which is particularly useful when labor markets are volatile.
- **Supply chain:** Businesses commonly use predictive analytics to manage product inventory and set pricing strategies. This type of predictive analysis helps companies meet customer demand without overstocking warehouses. It also enables companies to assess the cost and return on their products over time.

23

Predictive Analytics : Data Mining

Data mining is a way to **develop intelligence** (i.e., actionable information or knowledge) from data that an organization collects, organizes, and stores. **A wide range of data mining techniques** are being used by organizations to **gain a better understanding of their customers and their operations and to solve complex organizational problems.**

**Data Mining***['dā-tə 'mī-niŋ]*

A process used by companies to turn raw data into useful information.

24

Predictive Analytics : Data Mining

Large portion of “understanding the customer” can come from analyzing the vast amount of data that a company collects. The cost of storing and processing data has decreased dramatically in the recent past, and, as a result, the amount of data stored in electronic form has grown at an explosive rate. With the creation of large databases, the possibility of analyzing the data stored in them has emerged. The term *data mining* was originally used to describe the process through which previously unknown patterns in data were discovered.

Although the term data mining is relatively new, the ideas behind it are not. Many of the techniques used in data mining have their roots in traditional statistical analysis and artificial intelligence work done .

25

Predictive Analytics : Data Mining

Why, then, has it suddenly gained the attention of the business world? Following are some of most pronounced reasons:

- More intense competition at the global scale driven by customers' ever-changing needs and wants in an increasingly saturated marketplace.
- General recognition of the untapped value hidden in large data sources.
- The exponential increase in data processing and storage technologies.
- Significant reduction in the cost of hardware and software for data storage and processing.

26

Predictive Analytics : Data Mining

Data Mining Applications:

Data generated by the Internet is increasing rapidly in both volume and complexity. Large amounts of genomic data are being generated and accumulated all over the world. **Disciplines such as astronomy and nuclear physics** create huge quantities of data on a regular basis.

Medical and pharmaceutical researchers constantly generate and store data that can then be used in data mining applications to identify better ways to accurately diagnose and treat illnesses and to discover new and improved drugs.

On the commercial side, perhaps the most common use of data mining has been in **the finance, retail, and healthcare sectors**.

27

Predictive Analytics : Data Mining

Data mining is used to detect and reduce fraudulent activities, especially in insurance claims and credit card use to identify customer buying patterns to reclaim profitable customers to identify trading rules from historical data; and to aid in increased profitability using market-basket analysis.

Data mining is already widely used to better target clients, and with the widespread development of e-commerce, this can only become more imperative with t



28

Predictive Analytics : Data Mining

Data mining is a term used to describe discovering or “**mining**” knowledge from large amounts of data.

data mining has become the choice of the community. Many other names that are associated with data mining include knowledge extraction, pattern analysis, data archaeology, information harvesting, pattern searching, and data dredging.

data mining is a process that uses **statistical, mathematical, and artificial intelligence techniques** to **extract and identify useful information and subsequent knowledge** (or patterns) from large sets of data.

These patterns can be in the form of **business rules, affinities, correlations, trends, or prediction models** .

Data mining is tightly positioned at the intersection of many disciplines, including **statistics, artificial intelligence, machine learning, management science, information systems (IS), and databases** (see Figure 6.2).

29

Predictive Analytics : Data Mining

Using advances in all of these disciplines, data mining strives to **make progress in extracting useful information and knowledge from large databases.**

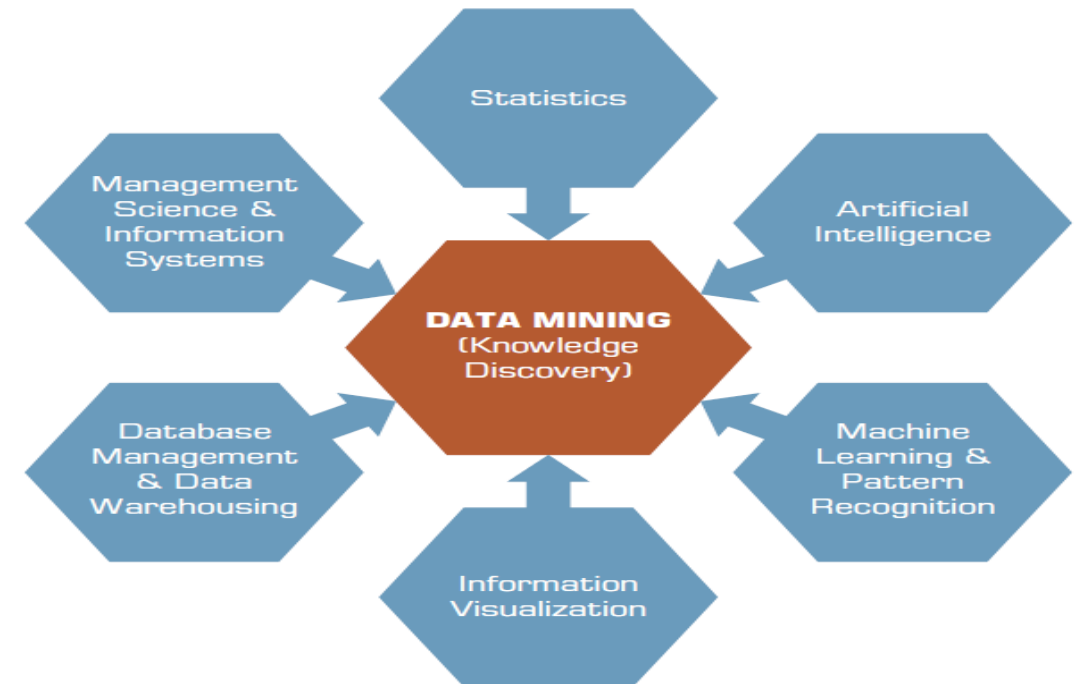


Figure 6.2 : Data Mining Is a Blend of Multiple Disciplines.

30

Predictive Analytics : Data Mining

The following are the major characteristics and objectives of data mining:

- Data are often buried deep within very large databases, which sometimes contain data from several years. In many cases, the data are cleansed and consolidated into a data warehouse. Data may be presented in a variety of formats .
- The data mining environment is usually a client/server architecture or a Web-based IS architecture.
- Sophisticated new tools, including advanced visualization tools, help to remove the information ore buried in corporate files or archival public records. Finding it involves massaging and synchronizing the data to get the right results.
- Because of the large amounts of data and massive search efforts, it is sometimes necessary to use parallel processing for data mining.

31

Predictive Analytics : Data Mining

What Are Benefits of a Data Mining?

- It's an efficient, cost-effective solution compared to the other data applications.
- It helps to businesses make profitable production and also operational adjustments.
- It helps to businesses make informed decisions.
- It helps to detect credit risks and fraud.
- It helps to data scientists easily **analyze the enormous amounts of data quickly.**
- Data scientists can use information to detect fraud, build risk models, and improve product safety.
- It helps to data scientists quickly initiate automated predictions of behaviors and trends and also discover hidden patterns.

32

Predictive Analytics : Data Mining

company that effectively leverages data mining tools and technologies can acquire and maintain a strategic competitive advantage. Data mining offers organizations an indispensable decision-enhancing environment to exploit new opportunities by transforming data into a strategic weapon.

How Data Mining Works:

Using existing and relevant data obtained from within and outside the organization, data mining builds models to discover patterns among the attributes presented in the data set.

Models are the mathematical representations (simple linear relationships/affinities and/or complex and highly nonlinear relationships) that identify the patterns among the attributes of the things (e.g., customers, events) described within the data set.

33

How Data Mining Works:

Some of these patterns are **explanatory** (explaining the interrelationships and affinities among the attributes), whereas others are predictive (foretelling future values of certain attributes).

In general, data mining seeks to identify four major types of patterns:

1. **Associations** find the **commonly co-occurring groupings of things**.

Market Basket Analysis:

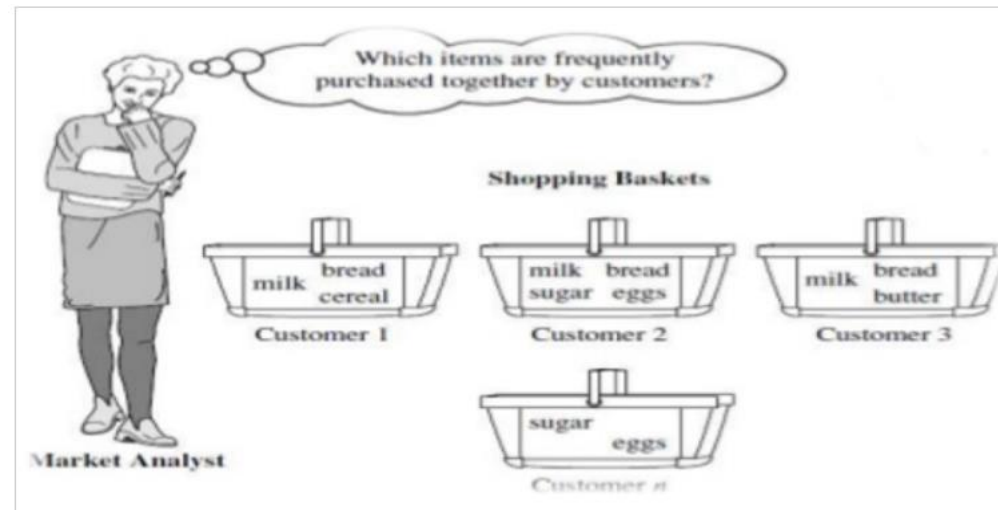
In transactional data, each case is connected with a set of objects. In principle, **the list may contain all possible data items in the collection**. For example - in a single market-basket analysis, goods with related items may be bought. However, only a small subset of all potential goods are present in a given set; **only a small fraction of the items available for sale in the shop reflect the items in the market basket**.

34

How Data Mining Works:

A common example of a regular pattern (item set) mining for association rules is market basket analysis. Business basket, the research analyzes the purchasing patterns of consumers by identifying correlations with the multiple items carried in their shopping baskets by customers.

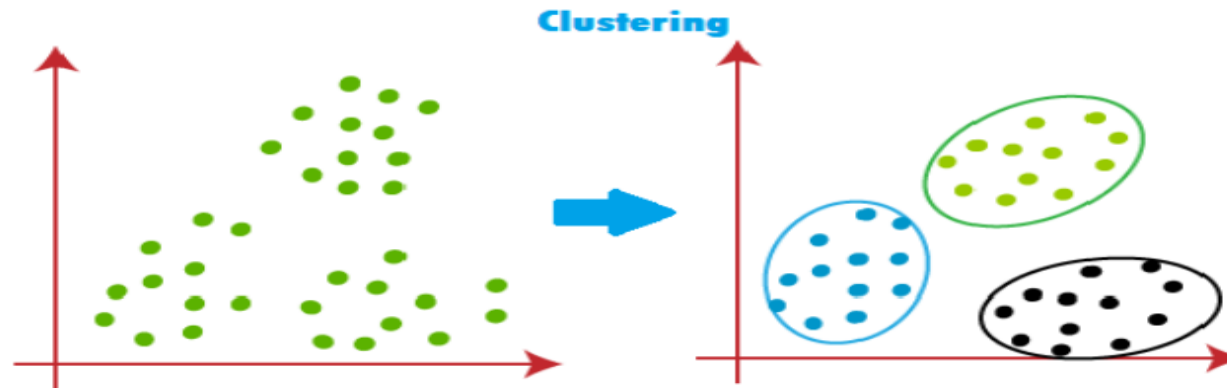
An example of association rule - milk, bread In a shop, if a shopkeeper sales milk then it is a probability to sell bread because a customer who is buying milk may also purchase bread. **So it is showing that milk and bread are correlated with one another.**



35

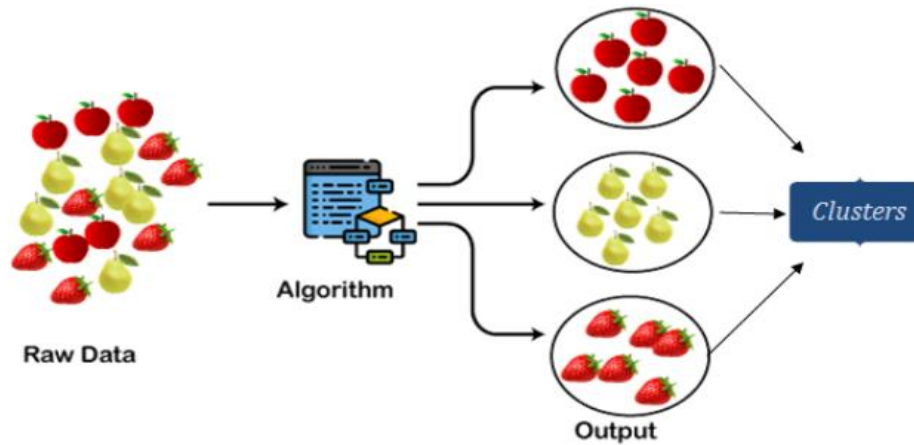
How Data Mining Works:

2. *Predictions*: tell the nature of **future occurrences of certain events based on what has happened in the past**, such as predicting the winner of the Super Bowl or forecasting the absolute temperature of a particular day.
3. *Clusters* identify natural groupings of things based on their known characteristics, such as assigning customers in different segments based on their demographics and past purchase behaviors.



36

How Data Mining Works:

**Cluster Analysis
Example**

4. *Sequential relationships* discover time-ordered events, such as predicting that an existing banking customer who already has a checking account will open a savings account followed by an investment account within a year.

37

How Data Mining Works:

Generally speaking, data mining tasks can be classified into three main categories: prediction, association, and clustering.

Based on the way in which the patterns are extracted from the historical data, the learning algorithms of data mining methods can be classified as either supervised or unsupervised.

With supervised learning algorithms, the training data includes both the descriptive attributes (i.e., independent variables or decision variables) as well as the class attribute (i.e., output variable or result variable).

In contrast, with unsupervised learning the training data includes only the descriptive attributes.

38

How Data Mining Works:

User ID	Gender	Age	Salary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	1
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	1
15728773	Male	27	58000	1
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	1
15727311	Female	35	65000	0
15570769	Female	26	80000	1
15606274	Female	26	52000	0
15746139	Male	20	86000	1
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1

Example of supervised learning algorithms

39

A Taxonomy for Data Mining

Data Mining Tasks & Methods	Data Mining Algorithms	Learning Type
Prediction		
Classification	Decision Trees, Neural Networks, Support Vector Machines, kNN, Naive Bayes, GA	Supervised
Regression	Linear/Nonlinear Regression, ANN, Regression Trees, SVM, kNN, GA	Supervised
Time Series	Autoregressive Methods, Averaging Methods, Exponential Smoothing, ARIMA	Supervised
Association		
Market-basket	Apriory, OneR, ZeroR, Eclat, GA	Unsupervised
Link analysis	Expectation Maximization, Apriory Algorithm, Graph-based Matching	Unsupervised
Sequence analysis	Apriory Algorithm, FP-Growth, Graph-based Matching	Unsupervised
Segmentation		
Clustering	K-means, Expectation Maximization (EM)	Unsupervised
Outlier analysis	K-means, Expectation Maximization (EM)	Unsupervised

A Simple Taxonomy for Data Mining Tasks, Methods, and Algorithms.

40

Summary

1. Using Excel for doing dashboard.
2. Using LLMS for data exploration.
3. Define data mining as an enabling technology for business analytics.



Thank You