Unit II

Data: Data Types, Data Collection, Data Pre-Processing, Data Analysis and Analytics, Descriptive Analytics, Diagnostic Analytics, Predictive and Perspective Analytics, Explorative Analysis, Mechanistic Analysis.

1. Structured vs. Unstructured Data

- **Structured Data** – Organized in a tabular format (e.g., relational databases, Excel sheets).
- **Unstructured Data** – Does not follow a fixed format (e.g., text, images, videos, emails).

2. Quantitative vs. Qualitative Data

- **Quantitative Data (Numerical)** – Data that can be measured or counted.
  - **Discrete Data** – Whole numbers, countable (e.g., number of students, product count).
  - **Continuous Data** – Measured and can take any value within a range (e.g., temperature, height).
- **Qualitative Data (Categorical)** – Data that describes qualities or characteristics.
  - **Nominal Data** – No natural order (e.g., gender, colors, names).
  - **Ordinal Data** – Has a meaningful order (e.g., survey ratings, education levels).

3. Primary vs. Secondary Data

- **Primary Data** – Collected firsthand for a specific purpose (e.g., surveys, experiments).
- **Secondary Data** – Pre-existing data collected by others (e.g., research papers, reports).

1. Types of Data Collection Methods

A. Primary Data Collection (First-Hand Data)

Collected directly from sources for a specific purpose.

- **Surveys & Questionnaires** – Structured forms to gather responses from individuals.
- **Interviews** – One-on-one discussions for in-depth insights.
- **Observations** – Monitoring behaviors or events in real-time.
- **Experiments** – Controlled studies to test hypotheses.
- **Sensors & IoT Devices** – Real-time data from physical devices.
- **Web Scraping** – Automated extraction of data from websites.

B. Secondary Data Collection (Pre-Existing Data)

Data gathered from existing sources for analysis.

- **Databases & Repositories** – Government, corporate, or public datasets (e.g., Kaggle, UCI).
- **APIs (Application Programming Interfaces)** – Real-time access to online data (e.g., weather, social media).
- **Open Data Portals** – Publicly available datasets (e.g., World Bank, WHO).
- **Research Papers & Reports** – Academic and industry studies.
- **Logs & Transactions** – System-generated records (e.g., server logs, purchase histories).

2. Data Collection Challenges

✔️ **Data Quality** – Ensuring accuracy, consistency, and completeness.
✔️ **Ethical Considerations** – Respecting privacy and security (e.g., GDPR, HIPAA).
✔️ **Volume & Scalability** – Handling large datasets efficiently.

3. Tools & Technologies for Data Collection

SQL & Databases – MySQL, PostgreSQL, MongoDB
Web Scraping – BeautifulSoup, Scrapy
APIs – REST APIs, JSON, Postman
Data Pipelines – Apache Kafka, Airflow
Survey Tools – Google Forms, Qualtrics

1. Steps in Data Preprocessing

A. Data Cleaning (Handling Missing & Noisy Data)

🔹 Handling Missing Values

- Remove missing data (if minimal impact).
- Fill missing values using **mean, median, mode**
- Use predictive models (e.g., regression, k-NN imputation).

🔹 Handling Noisy Data

- Remove duplicate or irrelevant data.
- Apply **smoothing techniques** (e.g., moving average, binning).
- Use **outlier detection** methods (e.g., Z-score).

B. Data Transformation (Standardization & Encoding)

- **Scaling & Normalization**

- **Standardization (Z-score Normalization):** Transforms data to have a mean of 0 and a standard deviation of 1.

Problem 1: Min-Max Normalization

Question:

Given the dataset:

$$X = [10, 20, 30, 40, 50]$$

Normalize the values using **Min-Max Scaling** to a range of **[0,1]** using the formula:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Solution:

Step 1: Identify min and max values

- Min value $X_{\min} = 10$
- Max value $X_{\max} = 50$

Step 2: Apply Min-Max Formula

For each value in X:

$$X' = \frac{X - 10}{50 - 10} = \frac{X - 10}{40}$$

Original (X)	Min-Max Normalized (X')
10	$\frac{10-10}{40} = 0.00$
20	$\frac{20-10}{40} = 0.25$
30	$\frac{30-10}{40} = 0.50$
40	$\frac{40-10}{40} = 0.75$
50	$\frac{50-10}{40} = 1.00$

Final Answer:

$$X' = [0.00, 0.25, 0.50, 0.75, 1.00]$$

Problem 2: Z-Score Normalization

Question:

Given the dataset:

$$X = [10, 20, 30, 40, 50]$$

Normalize the values using **Z-Score Normalization** using the formula:

$$Z = \frac{X - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation.

Solution:

Step 1: Calculate the Mean (μ)

$$\mu = \frac{10 + 20 + 30 + 40 + 50}{5} = \frac{150}{5} = 30$$

Step 2: Calculate the Standard Deviation (σ)

Standard deviation formula:

$$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}}$$

$$\sigma = \sqrt{\frac{(10-30)^2 + (20-30)^2 + (30-30)^2 + (40-30)^2 + (50-30)^2}{5}}$$

$$= \sqrt{\frac{(-20)^2 + (-10)^2 + 0^2 + 10^2 + 20^2}{5}}$$

$$= \sqrt{\frac{400 + 100 + 0 + 100 + 400}{5}}$$

$$= \sqrt{\frac{1000}{5}} = \sqrt{200} \approx 14.14$$

Step 3: Apply Z-Score Formula

For each value in X:

$$Z = \frac{X - 30}{14.14}$$

Original (X)	Z-Score Normalized (Z)
10	$\frac{10-30}{14.14} = \frac{-20}{14.14} \approx -1.41$
20	$\frac{20-30}{14.14} = \frac{-10}{14.14} \approx -0.71$
30	$\frac{30-30}{14.14} = 0$
40	$\frac{40-30}{14.14} = \frac{10}{14.14} \approx 0.71$
50	$\frac{50-30}{14.14} = \frac{20}{14.14} \approx 1.41$

Final Answer:

$$Z = [-1.41, -0.71, 0.00, 0.71, 1.41]$$

Summary of Results

Original (X)	Min-Max Normalized (X')	Z-Score Normalized (Z)
10	0.00	-1.41
20	0.25	-0.71
30	0.50	0.00
40	0.75	0.71
50	1.00	1.41

Data Analysis

Definition:

Data Analysis is the process of **cleaning, processing, and interpreting** raw data to extract meaningful **insights, patterns, and trends**. It is often used for understanding past data and making data-driven decisions.

Key Characteristics:

Descriptive & Diagnostic – Focuses on "What happened?" and "Why did it happen?"
Exploratory Approach – Identifies patterns, relationships, and anomalies in data.

Uses Statistical Methods – Mean, median, standard deviation, correlation, etc.
Visualization Techniques – Charts, graphs, histograms, heatmaps.

Example Use Case:

A retail company analyzes **sales data** to determine which products performed well last quarter.

Data Analytics

Definition:

Data Analytics goes a step further than analysis by applying **advanced techniques** (such as predictive modeling, machine learning, and business intelligence) to **find actionable insights and make future predictions**.

Key Characteristics:

Predictive & Prescriptive – Focuses on "What will happen?" and "How can we make it happen?"
Data-Driven Decision Making – Uses algorithms and AI to extract deeper insights.
Machine Learning & AI Techniques – Regression, clustering, deep learning.
Business Intelligence (BI) Applications – Dashboards, KPI tracking, forecasting models.

Example Use Case:

A bank uses **data analytics** to predict which customers are at risk of **loan default** based on their transaction history and credit score.

5. Tools Used in Data Analysis & Data Analytics

Category	Data Analysis Tools	Data Analytics Tools
Programming	Python (Pandas, NumPy), R	Python (Scikit-learn, TensorFlow)
Databases	SQL, PostgreSQL, MySQL	BigQuery, Apache Spark
Visualization	Tableau, Power BI, Matplotlib	Tableau, Looker, BI dashboards
Analytics	Excel, Jupyter Notebook	Google Analytics, AI/ML tools

Descriptive analytics is like a rear-view mirror, giving us a clear picture of past data to understand what has happened. It's the simplest form of data analysis and is usually the first step in data processing.

it involves:

1. **Summarizing Data:** It involves collecting historical data and presenting it in a readable and understandable format, using measures like mean, median, mode, etc.
2. **Data Visualization:** Use of charts, graphs, histograms, and other visual aids to make data easily digestible.
3. **Identifying Trends:** Helps in identifying patterns and trends over time, such as sales trends, customer behavior, and operational performance.

Tools commonly used in descriptive analytics:

- **Excel:** Popular for its pivot tables, charts, and graphs.
- **Tableau:** Great for interactive data visualization.
- **Power BI:** Microsoft's tool for data visualization and business intelligence.
- **SQL:** Used to query and manage data in databases.

Diagnostic analytics is like playing detective with your data. It goes a step further than descriptive analytics by not only looking at what happened but also trying to understand why it happened. This is crucial for uncovering root causes and making better decisions moving forward.

Here's what it involves:

1. **Identifying Anomalies:** Look for outliers or unusual patterns that deviate from the norm.
2. **Drill-Down Analysis:** Dig deeper into data subsets to get more detailed insights and identify the underlying factors.
3. **Correlation Analysis:** Assess the relationships between different variables to understand how they influence one another.
4. **Hypothesis Testing:** Form and test hypotheses to determine the causes of specific outcomes.

Tools commonly used in diagnostic analytics:

- **R and Python:** These programming languages are popular for their powerful statistical and data analysis packages.
- **SQL:** Helpful for querying detailed data sets and performing complex joins.
- **Tableau and Power BI:** These tools allow you to drill down into visual data representations and explore the factors behind the trends.

Prescriptive analytics is a type of data analytics that goes beyond descriptive and diagnostic analytics. While descriptive analytics tells you what happened and diagnostic analytics explains why it happened, prescriptive analytics recommends actions you can take to achieve desired outcomes. It leverages advanced techniques such as optimization algorithms, machine learning, and simulation to provide actionable insights and guidance on the best course of action.

Predictive analytics involves using historical data, machine learning algorithms, and statistical techniques to predict future outcomes. It aims to forecast trends, behaviors, and events by analyzing patterns found in existing data. This technique is widely applied across various industries, including finance, marketing, healthcare, and more.

Here are some key aspects of predictive analytics:

- **Data Collection**: Gathering relevant historical data from various sources.
- **Data Cleaning**: Ensuring the data is accurate and free of inconsistencies or errors.
- **Data Analysis**: Applying statistical methods and machine learning algorithms to identify patterns and relationships in the data.
- **Model Building**: Creating predictive models that can forecast future events based on the analyzed data.
- **Validation**: Testing the models to ensure their accuracy and reliability.
- **Deployment**: Implementing the models into real-world scenarios to make predictions and informed decisions.

Steps in Exploratory Data Analysis

1. **Data Collection**: Gather raw data from various sources.
2. **Data Cleaning**: Handle missing values, duplicates, and incorrect data entries.
3. **Descriptive Statistics**: Calculate basic statistics like mean, median, mode, variance, and standard deviation.
4. **Data Visualization**: Create plots and graphs (e.g., histograms, scatter plots, box plots) to visualize distributions and relationships.
5. **Identifying Patterns**: Look for trends, correlations, and outliers that might indicate underlying relationships.
6. **Hypothesis Testing**: Formulate and test hypotheses about the data.

Mechanistic analysis is a method used to understand the underlying mechanisms or processes that drive a system's behavior. It involves breaking down complex systems into their individual components and studying how each part interacts to produce the observed outcomes. This type of analysis is commonly used in fields like biology, chemistry, physics, and engineering.

Aspect	Data Analytics	Data Analysis
Scope	Broad, includes various techniques	Narrower, focuses on specific techniques
Goal	To provide actionable insights for decision-making	To uncover patterns and trends within the data
Tools	Advanced tools and software	Statistical tools and basic software
Application	Business intelligence, marketing, finance, etc.	Research, academic studies, operations