

<b>S.No.</b>	<b>CONTENT</b>	<b>PG.NO.</b>
<b>1.</b>	<b>What is Data?</b>	
<b>2.</b>	<b>Introduction to Data Analytics</b>	
	<b>2.1 Big Data and Data Science</b>	

# What is Data?

## Definition of Data:

- **Data:** Raw facts and figures that are collected, stored, and processed by a computer. These facts can be numbers, words, measurements, observations, or even descriptions of things.

## Types of Data:

### 1. Structured Data:

- **Characteristics:**
  - Organized in a defined manner, usually in tabular format.
  - Easily searchable and can be managed by traditional databases.
- **Examples:**
  - Spreadsheets (Excel)
  - SQL databases
  - CSV files

### 2. Unstructured Data:

- **Characteristics:**
  - No predefined format or organization.
  - Harder to process and analyze with traditional tools.
- **Examples:**
  - Text documents
  - Emails
  - Social media posts
  - Images, videos, and audio files

### 3. Semi-Structured Data:

- **Characteristics:**
  - Contains both structured and unstructured elements.
  - Does not fit neatly into tables but has some organizational properties.
- **Examples:**
  - JSON files
  - XML files
  - NoSQL databases

## Sources of Data:

### 1. Primary Data:

- Collected directly from the source or original users.
- **Examples:** Surveys, interviews, experiments.

### 2. Secondary Data:

- Collected from existing sources, previously gathered for other purposes.
- **Examples:** Research papers, government reports, company records.

Truck No.	Trip No.	Truck Enters	Service Completed	Delay Time	Truck Exits System	Service Time		Truck Reenters	Back Cycle Time
		Systems	Before			Queue Yes	Queue No	System	
1	1	0.00	0.00	0.00	1.83	1.83	N/A	45.05	43.22
2	1	0.00	1.83	1.83	4.23	2.40	N/A	54.00	49.77
3	1	0.00	4.23	4.23	6.92	2.68	N/A	55.88	48.97
4	1	3.75	6.92	3.17	8.47	1.55	N/A	58.02	49.55
5	1	4.72	8.47	3.75	11.37	2.90	N/A	60.80	49.43
6	1	14.80	11.37	(3.43)	16.38	N/A	1.58	63.58	47.20
7	1	21.12	16.38	(4.73)	22.77	N/A	1.65	65.77	43.00
8	1	23.20	22.77	(0.43)	24.75	N/A	1.55	66.05	41.30
9	1	39.85	24.75	(15.10)	41.67	N/A	1.82	83.03	41.37
10	1	43.32	41.67	(1.65)	44.72	N/A	1.40	86.20	41.48
1	2	45.05	44.72	(0.33)	46.68	N/A	1.63	89.00	42.32
2	2	54.00	46.68	(7.32)	55.58	N/A	1.58	93.82	38.23
3	2	55.88	55.58	(0.30)	57.58	N/A	1.70	95.35	37.77
4	2	58.02	57.58	(0.43)	59.75	N/A	1.73	97.93	38.18

Figure 1.1 – Data Formats

## Data Attributes:

- **Volume:** The amount of data.
- **Variety:** The different types and sources of data.
- **Velocity:** The speed at which data is generated and processed.
- **Veracity:** The accuracy and reliability of data.
- **Value:** The usefulness of the data for decision-making.

## Importance of Data:

- **Decision Making:** Provides a factual basis for making informed decisions.
- **Understanding Trends:** Helps in identifying patterns and trends in various domains.
- **Improving Efficiency:** Enables optimization of processes and operations.
- **Innovation:** Fuels new product development and service enhancements.
- **Competitive Advantage:** Offers insights that can lead to a competitive edge.

## Data Life Cycle:

### 1. Data Collection:

- Gathering data from various sources.
- Methods: Surveys, sensors, transaction records.
- 2. **Data Storage:**
  - Storing data in databases, data warehouses, or data lakes.
  - Technologies: SQL databases, NoSQL databases, cloud storage.
- 3. **Data Processing:**
  - Cleaning, transforming, and organizing data.
  - Techniques: Data cleansing, ETL (Extract, Transform, Load).
- 4. **Data Analysis:**
  - Applying statistical and computational methods to extract insights.
  - Tools: Excel, R, Python, SAS.
- 5. **Data Visualization:**
  - Presenting data insights in graphical formats for better understanding.
  - Tools: Tableau, Power BI, Matplotlib.
- 6. **Data Archiving:**
  - Storing data that is no longer in active use but may be needed for future reference.
  - Methods: Tape storage, cloud archiving.
- 7. **Data Deletion:**
  - Safely disposing of data that is no longer needed.
  - Techniques: Data wiping, shredding, degaussing.

### **Challenges in Data Management:**

- **Data Quality:** Ensuring accuracy, completeness, and consistency.
- **Data Security:** Protecting data from unauthorized access and breaches.
- **Data Integration:** Combining data from different sources and formats.
- **Data Privacy:** Complying with regulations to protect personal information.
- **Data Volume:** Managing large volumes of data efficiently.

### **Summary:**

- Data is a fundamental asset in the modern world, driving innovation, efficiency, and informed decision-making across various fields. Understanding the nature, types, and lifecycle of data is essential for leveraging its full potential.

## **2. Introduction to Data Analytics**

### **Definition:**

- **Data Analytics:** The science that analyze crude data to extract useful knowledge (patterns) from them.
- This process can also include data collection, organization, pre-processing, transformation, modeling and interpretation.

## Importance:

- **Decision Making:** Helps businesses and organizations make informed decisions by providing insights and trends.
- **Efficiency:** Improves operational efficiency by identifying bottlenecks and optimizing processes.
- **Competitive Advantage:** Provides a competitive edge by uncovering patterns and predicting future trends.

## Types of Data Analytics:

1. **Descriptive Analytics:**
  - Summarizes past data to understand what has happened.
  - Tools: Reports, dashboards, and scorecards.
  - Example: Monthly sales reports.
2. **Diagnostic Analytics:**
  - Examines data to understand the reasons behind past outcomes.
  - Tools: Drill-down, data discovery, and correlations.
  - Example: Identifying reasons for a decline in sales.
3. **Predictive Analytics:**
  - Uses historical data to predict future events.
  - Tools: Statistical models, machine learning algorithms.
  - Example: Forecasting future sales based on past data.
4. **Prescriptive Analytics:**
  - Recommends actions to achieve desired outcomes.
  - Tools: Optimization, simulation algorithms.
  - Example: Suggesting marketing strategies to increase sales.

## Data Analytics Process:

1. **Data Collection:**
  - Gathering raw data from various sources like databases, spreadsheets, sensors, etc.
2. **Data Cleaning:**
  - Ensuring data quality by removing inaccuracies, inconsistencies, and duplicates.
3. **Data Transformation:**
  - Converting raw data into a suitable format for analysis.
  - Includes normalization, aggregation, and scaling.
4. **Data Analysis:**
  - Applying statistical and machine learning techniques to analyze data.
  - Tools: R, Python, SAS, SQL.

## 5. Data Visualization:

- Presenting data insights in graphical formats.
- Tools: Tableau, Power BI, Matplotlib, Seaborn.

## 6. Interpretation and Reporting:

- Interpreting results to provide actionable insights and recommendations.
- Creating reports and dashboards for stakeholders.

## Tools and Technologies:

- **Programming Languages:** Python, R
- **Database Management Systems:** SQL, NoSQL
- **Data Visualization:** Tableau, Power BI
- **Big Data Technologies:** Hadoop, Spark
- **Machine Learning Libraries:** Scikit-learn, TensorFlow, Keras

## Applications of Data Analytics:

- **Business:** Market analysis, customer segmentation, and sales forecasting.
- **Healthcare:** Predictive diagnosis, personalized treatment, and operational efficiency.
- **Finance:** Risk management, fraud detection, and algorithmic trading.
- **Retail:** Inventory management, personalized marketing, and demand forecasting.
- **Sports:** Performance analysis, injury prevention, and game strategy.

## Challenges in Data Analytics:

- **Data Quality:** Ensuring the accuracy and completeness of data.
- **Data Privacy:** Protecting sensitive information.
- **Integration:** Combining data from various sources.
- **Skill Gap:** Need for skilled professionals with expertise in analytics.

## Future of Data Analytics:

- **Artificial Intelligence and Machine Learning:** Enhanced predictive and prescriptive capabilities.
- **Real-Time Analytics:** Instant insights for immediate decision-making.
- **Advanced Visualization:** Interactive and immersive data visualizations.
- **Integration with IoT:** Analyzing data from interconnected devices.

## Summary:

- **Summary:** Data Analytics is a powerful tool that transforms raw data into actionable insights, driving informed decision-making and strategic planning.
- **Future Trends:** Embrace emerging technologies and continuous learning to stay ahead in the field of data analytics.

## 1.1 Big Data and Data Science

### Introduction to Big Data:

- **Big Data:** Refers to large and complex data sets that traditional data-processing software cannot manage efficiently.
  - **Three Vs:**
    - **Volume:** The enormous amount of data generated and stored, requiring advanced storage solutions.
    - **Variety:** The different types and sources of data (e.g., structured, unstructured, semi-structured) that need to be integrated and analyzed.
    - **Velocity:** The rapid speed at which data is generated and needs to be processed, often in real-time.

### Evolution of the Vs:

- Additional Vs that further define Big Data:
  - **Veracity:** The quality and accuracy of data.
  - **Value:** The usefulness and importance of the data for decision-making.
  - **Variability:** The inconsistency of data over time, affecting data quality and analysis.

### Big Data Technologies:

- **MapReduce:** A programming model for processing large data sets with a parallel, distributed algorithm.
- **Hadoop:** An open-source framework that allows for distributed processing of large data sets across clusters of computers.
- **Spark:** A unified analytics engine for large-scale data processing with modules for streaming, SQL, machine learning, and graph processing.
- **Storm:** A real-time computation system for processing data streams.

### Applications of Big Data:

- **Finance Transaction Processing:** Efficiently handling large volumes of financial transactions.
- **Web Data Processing:** Managing and analyzing data from web activities.
- **Georeferenced Data Processing:** Handling and analyzing geographical information for various applications.

### Introduction to Data Science:

- **Data Science:** An interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data.
  - Focuses on creating models to find patterns in complex data and applying these models to solve real-life problems.
  - **Key Components:**
    - **Statistics:** Provides the theoretical foundation for data analysis.
    - **Data Mining:** The process of discovering patterns in large data sets.
    - **Machine Learning:** Developing algorithms that learn from and make predictions on data.
    - **Visualization:** Presenting data insights in graphical formats for better understanding and decision-making.

### **Role of Data Science:**

- **Knowledge Extraction:** Creating frameworks to extract meaningful and useful knowledge from data.
- **Real-Life Applications:** Using models to address practical problems in various domains, such as healthcare, finance, marketing, and more.

### **Relationship to Other Fields:**

- **Analytics:** Systematic computational analysis of data.
- **Pattern Recognition:** Identifying patterns and regularities in data.
- **Data Engineering:** Designing and constructing systems for collecting, storing, and analyzing data.

### **Summary:**

- **Big Data:** Provides the technological infrastructure for data storage, management, and processing.
- **Data Science:** Utilizes the data to discover new and useful knowledge, creating value through analytics and model creation.
- Together, they form a comprehensive approach to handling and extracting insights from the vast amounts of data generated in today's digital world.



## 1.2 Big Data Architectures

### Introduction

- **Scalability:** The capacity of a system to grow and manage increased data by adding resources.
  - Achieved by distributing processing tasks across multiple computers, forming clusters.

### Clusters of Computers

- **Clusters:** Groups of interconnected computers working together to process large data sets.
  - Not to be confused with clusters from clustering techniques in analytics.

### Challenges with Conventional Distributed Systems

- Traditional distributed systems struggle to efficiently distribute data among processing and storage units.
- New software tools and techniques are necessary to manage the requirements of big data.

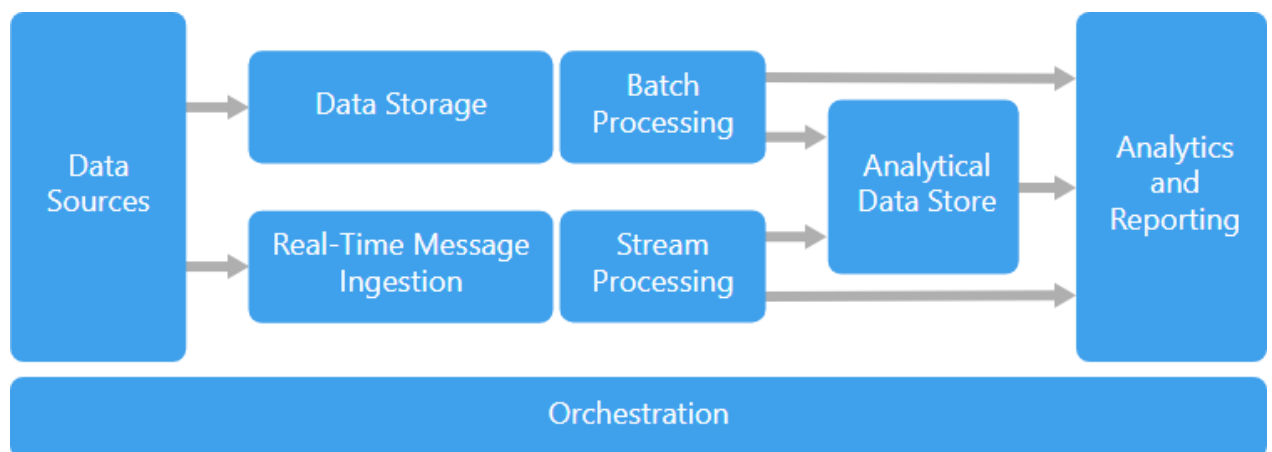


Figure 1.2 Big Data Architecture

### MapReduce

- **MapReduce:** A programming model for processing large data sets using distributed algorithms.
  - **Map Step:** Divides the data set into smaller chunks and distributes them across the cluster.
  - **Reduce Step:** Aggregates results from each chunk to produce the final output.
  - **Example:**

- Calculating the average salary of 1 billion people using a cluster of 1000 computers.
- Each computer processes the salary data of 1 million people and calculates the average.
- The final average is computed by averaging the results from all computers.

## Requirements for Efficient Big Data Systems

### 1. Data Integrity

- Ensure no data chunk is lost.
- If a computer fails, its task and data chunk must be reassigned to another computer.

### 2. Redundancy

- Duplicate tasks and data chunks across multiple computers.
- If one computer fails, a redundant computer can continue the task.

### 3. Fault Tolerance

- Computers that have failed can rejoin the cluster once repaired.

### 4. Flexibility

- Easily add or remove computers from the cluster as processing demands change.

## User Transparency

- The solution should abstract the complexities from the data analyst.
- Analysts should not need to worry about how data chunks and tasks are distributed among the cluster computers.

## Summary:

- Big data architectures must be scalable and robust to handle large volumes, high velocities, and diverse varieties of data.
- Technologies like MapReduce and Hadoop are designed to meet these needs, enabling efficient distributed processing of big data.

## 1.3 Small Data

### Introduction to Small Data

- **Small Data:** Refers to data sets that are small enough in volume and format to be processed and analyzed by an individual or a small organization.
  - Contrasts with big data, which involves large, complex data sets that require significant computing resources for processing.

### Characteristics of Small Data

- **Personal and Subjective:** Focuses on more personalized, subjective analysis of data.
- **Manageable Size:** Small enough to be fully understood and processed by a person or small group.
- **Partitioning Problems:** Breaks down larger problems into smaller, manageable parts that can be analyzed independently.

### Sources of Small Data

- **Daily Activities:** Data generated from everyday actions such as:
  - Navigating the web
  - Shopping transactions
  - Medical examinations
  - Mobile app usage
- When these data are collected and stored in large data servers, they transition into big data.

### Differences Between Small Data and Big Data

- **Volume and Complexity:**
  - **Big Data:** Large, complex data sets from multiple sources and formats, generated at high velocities.
  - **Small Data:** Smaller, simpler data sets that are easier to process and analyze.
- **Analysis Goals:**
  - **Big Data:** Seeks correlations and patterns to understand customer behavior, product performance, and service efficiency.
  - **Small Data:** Aims to uncover causality relationships, helping individuals understand their own behavior and preferences.

### Applications of Small Data

- **Personal Insights:** Helps individuals gain insights into their own activities and behaviors.
- **Small Organizations:** Enables small businesses and groups to perform data analysis without the need for extensive computing resources.
- **Distributed Analysis:** Allows different people or small teams to analyze different parts of a problem, facilitating collaborative and integrated solutions.

### Summary

- **Small Data:** Provides a complementary approach to big data, focusing on more personalized, manageable, and causality-driven analysis.
- **Importance:** While big data helps organizations understand their customers and optimize services, small data empowers individuals to gain insights into their own actions and behaviors.

## 1.4 What is Data?

### Definition of Data

- **Data:** In the information age, data are a collection of bits encoding various forms of information, such as numbers, text, images, sounds, and videos.
  - By themselves, data are meaningless and need to be processed to become information and eventually knowledge.

### Transformation of Data

- **Data to Knowledge:** Data become meaningful (knowledge) when information is added to them.
  - **Steps:** Data go through several steps of organization, gaining information before becoming knowledge.

### Example: Contact List Data

- **Tabular Data:** A common format where data are organized in rows and columns.
  - **Rows:** Represent instances or objects.
  - **Columns:** Represent attributes or features.

### Attributes and Instances

- **Attribute/Feature:** Characteristics of the data instances.
  - Example: In a contact list, attributes could be "Contact," "Age," "Educational Level," and "Company."
- **Instance/Object:** Examples of the concept being characterized.
  - Example: Each contact in a list is an instance, represented by a row in the table.

### Tabular Data Example

- **Table 1.1:** Represents a dataset of a private contact list.

Contact	Age	Educational Level	Company
Andrew	55	1.0	Good
Bernhard	43	2.0	Good
Carolina	37	5.0	Bad
Dennis	82	3.0	Good
Eve	23	3.2	Bad
Fred	46	5.0	Good
Gwyneth	38	4.2	Bad

Contact	Age	Educational Level	Company
Hayden	50	4.0	Bad
Irene	29	4.5	Bad
James	42	4.1	Good
Kevin	35	4.5	Bad
Lea	38	2.5	Good
Marcus	31	4.8	Bad
Nigel	71	2.3	Good

## Relational Data

- **Relational Datasets:** Sometimes data need to be split into multiple tables to represent relationships between them.
  - **Example:** Family relationships between contacts in a contact list.

Friend	Father	Mother	Sister
Eve	Irene	Andrew	Andrew
Hayden	Irene	Eve	

## Key Points

- **Single Table Representation:** Not all data can be represented in a single table. When relationships between data points exist, multiple tables are used.
- **Relational Databases:** These datasets are easily managed using relational databases, though this book will primarily use simpler forms of relational data.

## Summery:

- **Understanding Data:** Data needs to be organized and contextualized to transform into useful information and knowledge.
- **Formats:** Data can be represented in various formats, such as tabular or relational, to facilitate understanding and analysis.

## 1.5 A Short Taxonomy of Data Analytics

### Overview

- **Data Analytics:** The field involves various techniques and methods to analyze data and extract valuable information.
- **Natural Taxonomy in Data Analytics:**
  - **Descriptive Analytics:** Summarizes or condenses data to extract patterns.
  - **Predictive Analytics:** Extracts models from data to make future predictions.

### Descriptive Analytics

- **Purpose:** Directly apply algorithms to data to obtain results such as statistics, plots, or groups of similar instances.
- **Method or Technique:**
  - A systematic procedure to achieve a specific goal.
  - Example: Calculating the average age of contacts.
- **Algorithm:**
  - A step-by-step set of instructions to implement a method.
  - **Example Algorithm:** Calculate the average age of contacts.

### Example Algorithm: Calculate Average Age

1. **Input:** A vector A of size N containing ages.
2. **Initialize:** Set sum S to 0.
3. **Iteration:** Loop through each age.
  - Add each age to S.
4. **Calculation:** Divide sum S by N.
5. **Output:** Return the average age A.

### pseudo

Copy code

Algorithm to calculate average age:

```
1: INPUT: A: vector of size N with ages
2: S ← 0
3: for i = 1 to N do
4:   S ← S + Ai
5: A ← S / N
6: return A
```

- **Alternative Expression:** Using a formula instead of an algorithm.
  - Average (A) =  $\Sigma(A_i) / N$ , where  $\Sigma(A_i)$  is the sum of ages.

### Predictive Analytics

- **Purpose:** Describes methods to generate models for making predictions on new data.
- **Model:**
  - A generalization obtained from data.
  - Used to generate predictions for new instances.

- Example: Decision tree model to predict if a contact is a good company based on age.

#### *Example: Decision Tree Model*

- **Purpose:** Predict whether a contact is a good company based on their age.
- **Explanation:**
  - People older than 38 are typically good company.
  - People aged 38 or younger are typically bad company.
- **Usage:** Predict the company quality of a new contact based on their age.

#### **Summary**

- **Descriptive Analytics:** Focuses on summarizing data.
- **Predictive Analytics:** Focuses on building models to predict future data points.
- **Algorithms and Models:** Essential tools in both types of analytics for processing and interpreting data.

## **1.6 Examples of Data Use**

### **Introduction**

- **Purpose:** Introduce real-world problems from different areas.
- **Focus Areas:** Medicine (breast cancer detection) and economics (company insolvency prediction).
- **Relevance:** These problems will be solved in the project chapters of the book.

#### **1.6.1 Breast Cancer in Wisconsin**

- **Problem:** Detection of breast tumors, mainly affecting women.
- **Detection Technique:** Biopsy technique known as fine-needle aspiration.
- **Dataset:** Images of breast mass samples obtained through fine-needle aspiration.
- **Objective:** Detect different patterns of breast tumors for diagnostic purposes.

#### **1.6.2 Polish Company Insolvency Data**

- **Problem:** Prediction of economic wealth of Polish companies.
- **Prediction Target:** Identifying companies likely to become insolvent in the next five years.
- **Relevance:** Important for institutions and shareholders to make informed decisions.

### **Conclusion**

- **Relevance of Data:** Data analysis plays a crucial role in solving complex real-world problems.

- **Impact:** Solutions derived from data analytics can have significant implications in various industries.

## 1.7 A Project on Data Analytics

### Introduction to Project Planning

- **Importance of Planning:** Every project needs a methodology for preparation.
- **Elements of a Project:** Understanding the problem, defining objectives, data collection, preparation, method selection, hyper-parameter tuning, result analysis, and iteration.

### Hyper-parameters and Parameters

- **Hyper-parameters:** Set by the user or an external optimization method.
- **Parameters:** Set by the modeling or learning algorithm internally.
- **Example:** Number of layers and activation function (hyper-parameters) vs. weights found by backpropagation (parameters) in a neural network.

### Methodology Overview

- **Organized Operations:** Describes the systematic approach to project planning and development.
- **Hyper-parameter Tuning:** Important for optimizing model performance.
- **Methodologies Covered:** KDD (from academia) and CRISP-DM (from industry).

#### 1.7.1 A Little History on Methodologies for Data Analytics

- **Development in the 1990s:** Strong development in machine learning and knowledge discovery.
- **Academic Methodology:** KDD process by Fayyad, Piatetsky-Shapiro, and Smyth.
- **Industry Methodology:** CRISP-DM, conceived in 1996, widely used in various corporations and industries.

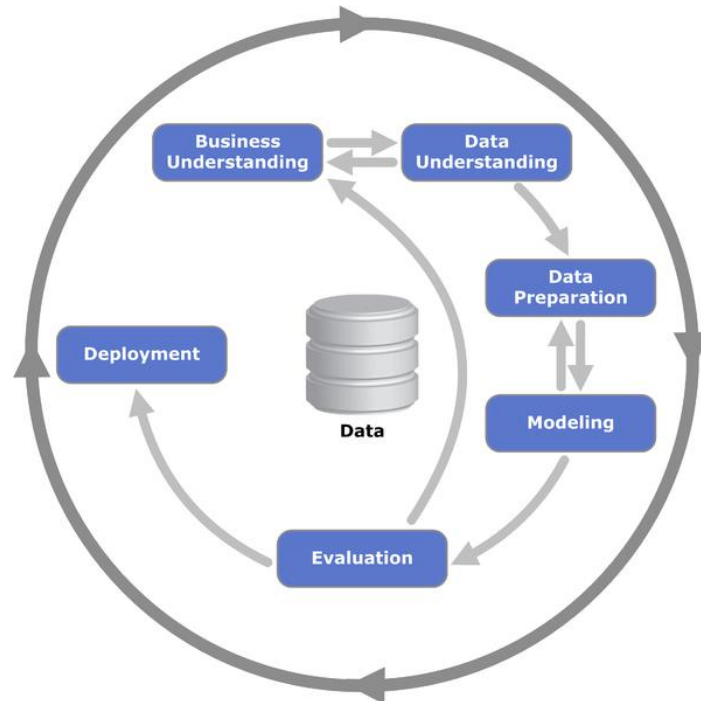
#### 1.7.2 The KDD Process

- **Nine Steps:** Learning application domain, creating target dataset, data cleaning, reduction, choosing mining function and algorithm, mining, interpretation, and using discovered knowledge.
- **Flexibility:** Allows going back to previous steps if needed.

#### 1.7.3 The CRISP-DM Methodology



- **Six Phases:** Business understanding, data understanding, data preparation, modeling, evaluation, and deployment.
- **Iterative Process:** Viewed as a perpetual process, used in successive iterations throughout the life of a company.



### Summery:

- **Comprehensive Planning:** Methodologies provide a structured approach to extract knowledge from data.
- **Adaptability:** Both methodologies allow flexibility and iteration, crucial for addressing complex data analytics problems.

## 2.Descriptive Statistics

### Introduction to Descriptive Statistics:

- Descriptive statistics is a fundamental branch of statistics that deals with summarizing and describing important aspects of data.
- It provides tools and techniques to analyze and interpret data sets, enabling researchers to draw meaningful insights.

### Importance of Sampling:

- Sampling is essential when it is impractical or impossible to survey an entire population.
- By analyzing a subset of the population, we can estimate characteristics of the entire population, such as proportions or averages.

### Deduction vs. Induction:

- Deduction involves drawing specific conclusions based on general principles or the population.
- Induction involves generalizing from specific observations or samples to make broader conclusions about the population.

### Types of Descriptive Statistics:

1. **Univariate Analysis:** Focuses on a single variable, describing its central tendency, dispersion, and shape.
2. **Bivariate Analysis:** Examines the relationship between two variables, often using measures like correlation and regression.
3. **Multivariate Analysis:** Studies the relationships between multiple variables simultaneously, providing a more comprehensive understanding of the data.

### Data Visualization Techniques:

- **Histograms:** Visual representation of the distribution of a single variable, showing the frequency of each value.
- **Box Plots:** Summarizes the distribution of a dataset by displaying the minimum, first quartile, median, third quartile, and maximum values.
- **Scatter Plots:** Used to explore the relationship between two continuous variables, showing how one variable changes in relation to another.

### Scale Types:

- **Qualitative Scales:** Categorize data into distinct groups without any inherent order (nominal) or with an order (ordinal).

- **Quantitative Scales:** Measure data with meaningful numerical values, including absolute (ratio) scales with a true zero point and relative (interval) scales without a true zero.

### Conclusion:

- Descriptive statistics plays a crucial role in data analysis, providing meaningful summaries and visualizations that facilitate a deeper understanding of datasets.
- By utilizing various statistical techniques, researchers can effectively interpret data and make informed decisions.

## Scale Types in Descriptive Statistics

### Qualitative Scales:

- **Nominal Scale:** Categorizes data into distinct groups without any inherent order. Examples include names, gender, or colors.
- **Ordinal Scale:** Orders data based on a certain characteristic, allowing for ranking. For example, levels of satisfaction (e.g., good, better, best) or educational attainment (e.g., high school, bachelor's, master's).

### Quantitative Scales:

- **Absolute (Ratio) Scale:** Includes a true zero point, where zero indicates the absence of the measured quantity. Examples include weight, height, and temperature in Kelvin.
- **Relative (Interval) Scale:** Lacks a true zero point, where zero does not indicate the absence of the measured quantity. Examples include temperature in Celsius or Fahrenheit.

### Converting Between Scales:

- Data expressed in a more informative scale can be converted to a less informative scale, but this results in a loss of information.
- Converting from a less informative scale to a more informative one is also possible, but the level of information obtained will be limited by the original scale.

### Example Conversions:

1. **Weight (Absolute Scale):**
  - **Relative Scale:** Subtracting 10 from all weights, changing the reference point but maintaining the order.
  - **Ordinal Scale:** Classifying weights as "fat," "normal," or "thin," based on predefined thresholds.

## 2. Choosing Data Types:

- In software, data types (text, character, integer, etc.) are chosen based on the scale type of the attribute.
- For example, numeric data types (integer, real, float) are used for quantitative scales, with discrete or continuous values.

### Note:

- Attributes expressed as numbers may not necessarily have a quantitative scale type; they could be ordinal or nominal.

# Descriptive Univariate Analysis

## Definition

- **Descriptive Statistics:** Focuses on describing and summarizing the features of a dataset.
- **Univariate Analysis:** Examines a single variable in isolation.

## Methods

1. **Frequency Tables:**
  - **Absolute Frequency:** Number of times each value occurs.
  - **Relative Frequency:** Proportion of total occurrences.
2. **Statistical Measures:**
  - **Central Tendency:** Mean, median, mode.
  - **Variability:** Range, variance, standard deviation.
3. **Plots:**
  - Histograms, bar charts, box plots.
  - Helps visualize the distribution of data.

## Frequency Tables

- **Absolute Frequency:** Counts how many times each value appears.
- **Relative Frequency:** Expresses the absolute frequency as a percentage of the total.

### Example:

- **Absolute Frequency** of "Good" in the "Company" attribute is 7.
- **Relative Frequency** of "Good" is 50%.

## Cumulative Frequencies

- **Absolute Cumulative Frequency:** Counts the total occurrences up to a given value.

- **Relative Cumulative Frequency:** Expresses the absolute cumulative frequency as a percentage of the total.

**Example:**

- **Absolute Cumulative Frequency** for 175 cm in "Height" is 8.
- **Relative Cumulative Frequency** for 175 cm is 57.13%.

## Distribution Functions

- **Empirical Frequency Distribution:** Describes how data are distributed in a sample.
- **Empirical Cumulative Distribution Function:** Describes how data are distributed up to a given value in a sample.

**Example:**

- The "rel. freq." column in a table shows an empirical frequency distribution.
- The "rel. cum. freq." column shows an empirical cumulative distribution function.

## Population Distribution Functions

- **Probability Distribution Functions:** Used for populations with discrete attributes.
- **Probability Density Functions:** Used for populations with continuous attributes.

**Note:**

- In a continuous space, the probability of an exact value is zero.
- Probability density functions have an area of one, representing 100%.

## Summery

Descriptive univariate analysis helps in understanding the distribution and characteristics of data in a dataset, providing insights for further analysis.

## Descriptive Univariate Analysis

### Frequency Tables

Frequency tables are used to summarize the distribution of a single variable. They provide a way to organize and display the counts or percentages of different values.

### Example: Company Attribute

Consider the following dataset of contacts:

Name	Company
Andrew	Good
Bernhard	Good
Carolina	Bad
Dennis	Good
Eve	Bad
Fred	Good
Gwyneth	Bad
Hayden	Bad
Irene	Bad
James	Good
Kevin	Bad
Lea	Good
Marcus	Bad
Nigel	Good

*Absolute and Relative Frequencies for the "Company" Attribute*

**Company Absolute Frequency Relative Frequency**

Good	7	50%
Bad	7	50%

In this example, there are 7 contacts with a "Good" company rating and 7 contacts with a "Bad" company rating. The relative frequency indicates that each category makes up 50% of the total.

**Cumulative Frequencies**

Cumulative frequencies show the running total of frequencies as you move through the data.

**Example: Height Attribute**

Continuing with our dataset, let's look at the heights of the contacts:

<b>Name</b>	<b>Height (cm)</b>
Andrew	175
Bernhard	195
Carolina	172
Dennis	180
Eve	168
Fred	173
Gwyneth	180
Hayden	165
Irene	158
James	163
Kevin	190
Lea	172
Marcus	185
Nigel	192

*Absolute and Relative Cumulative Frequencies for the "Height" Attribute*

<b>Height (cm)</b>	<b>Abs. Cum. Freq.</b>	<b>Rel. Cum. Freq.</b>
158	1	7.14%
163	2	14.28%
165	3	21.42%
...	...	...
195	14	99.98%

The absolute cumulative frequency for each height value represents the total number of instances up to that height. The relative cumulative frequency expresses this as a percentage of the total instances.

## 2.2.2 Univariate Data Visualization

### *Pie Chart*

- **Applicability:** Typically used for nominal scales.
- **Usage:** Not advisable for scales with order (ordinal and quantitative scales), but can be used.
- **Comparison:** Bar charts are often considered better for comparing values between different classes, as it's easier to perceive differences in bar lengths than in pie slice sizes.

### *Bar Chart*

- **Applicability:** Typically used for qualitative scales.
- **Order Consideration:** When there's an order, classes should be displayed in horizontal bars in increasing order of magnitude.
- **Comparison:** Considered better for comparing values between different classes than pie charts.
- **Additional Usage:** Can be used with quantitative scales when the possible values are limited (e.g., counting occurrences of each face of a die or number of students with a specific mark on a 0-20 scale).

### *Histogram*

- **Applicability:** Used for quantitative scales.
- **Usage:** Displays the distribution of a continuous variable.
- **Bars:** Bars represent intervals of values (bins) rather than individual values.
- **Frequency:** Height of each bar represents the frequency of values within that interval.

### *Box Plot (Box-and-Whisker Plot)*

- **Applicability:** Primarily used for quantitative scales.
- **Representation:** Provides a visual summary of the data's distribution.
- **Elements:** Includes a box representing the interquartile range (IQR), a line within the box indicating the median, and "whiskers" extending from the box to the minimum and maximum values (outliers may be shown as points beyond the whiskers).

### *Line Chart*


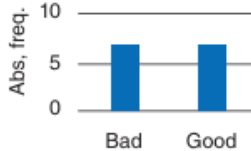
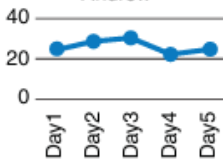
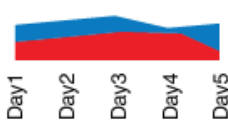
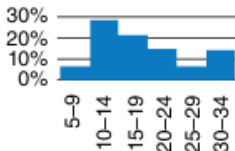
- **Applicability:** Typically used for quantitative scales, especially for time series data.
- **Representation:** Shows data points connected by straight line segments.
- **Usage:** Effective for showing trends or changes over time.



## Scatter Plot

- **Applicability:** Used for quantitative scales, specifically to show relationships between two variables.
- **Representation:** Each point represents a single data point, with one variable on the x-axis and the other on the y-axis.
- **Usage:** Useful for identifying correlations or patterns in data.

Table 2.4 Univariate plots.

Plot	Qualitative	Quantitative	Observation	Plot draft
Pie	Yes	No	Company relative frequency	<p>Company</p>  <p>■ Bad ■ Good</p>
Bar	Yes	Not always	Company absolute frequency	<p>Company</p>  <p>Abs. freq.</p> <p>Bad Good</p>
Line	No	Yes	Andrew's 5-day max. temperatures	<p>Andrew</p>  <p>Day1 Day2 Day3 Day4 Day5</p>
Area	No	Yes	Andrew and Eve 5-day max. temperatures	<p>Andrew &amp; Eve</p>  <p>Day1 Day2 Day3 Day4 Day5</p> <p>■ Andrew ■ Eve</p>
Histogram	No	Yes	Max. last day temperatures of the 14 contacts	<p>Max.temp.</p>  <p>30% 20% 10% 0%</p> <p>5-9 10-14 15-19 20-24 25-29 30-34</p>

## Area Charts

- **Usage:** Compare time series and distribution functions.
- **Insights:** Understanding data distributions provides insights into the concentration of data values.

## Histograms

- **Applicability:** Represent empirical distributions for attributes with a quantitative scale.
- **Grouping:** Group values into cells to reduce sparsity common in quantitative scales.
- **Cell Definition:** Number of cells is problem-dependent, often around the square root of the number of values.
- **Cell Configuration:** No space between columns to preserve the idea of continuity; equal-sized cells are common, adjusting height proportionally to width changes.

## Bar Chart vs. Histogram

- **Comparison:** Histograms are more informative than bar charts for quantitative scales.
- **Representation:** Histograms show distribution, while bar charts compare values between different classes.

## Cumulative Distribution Functions

- **Empirical vs. Probability Distributions:** Empirical distributions are based on samples, while probability distributions are about populations.
- **Understanding:** The step-wise nature of empirical cumulative probability distributions is due to limited sample values and predefined precision levels.

## Stacked Bar Plot

- **Usage:** Represent frequencies for combined values of two different attributes in a single chart.
- **Example:** Figure 2.6 illustrates frequencies for the target value of "company" split by gender, showing a stacked bar plot.

## Summery:

- **Selection:** Choose the appropriate plot based on the scale and type of data to be analyzed.
- **Clarity:** Ensure that the plot is clear and easily interpretable for the audience.
- **Context:** Provide context and labels to help viewers understand the data being presented.

These visualizations help in gaining insights into data distributions, identifying patterns, and comparing values, which are crucial for data analysis and decision-making.

# Univariate Statistics

## Location Statistics

- **Definition:** Descriptors that identify a value in a specific position.
- **Examples:**
  - Minimum: the lowest value.
  - Maximum: the largest value.
  - Mean: the average value.
  - Mode: the most frequent value.
  - Quartiles: values that divide the data into quarters.

## Measures of Central Tendency

- **Mean, Median, Mode:** These are measures that represent the center of a dataset.
- **Use Cases:**
  - Mean: Suitable for quantitative scales.
  - Median: More robust against extreme values or strongly skewed distributions.
  - Mode: Not useful when data are very sparse.

## Dispersion Statistics

- **Definition:** Measures how distant different values are from each other.
- **Common Statistics:**
  - Amplitude: Difference between maximum and minimum values.
  - Interquartile Range: Difference between the third and first quartiles.
  - Mean Absolute Deviation (MAD): Mean absolute distance between observations and the mean.
  - Standard Deviation: Typical distance between observations and their mean.

## Formulas

**MAD for Population:**

$$MAD_x = \frac{\sum_{i=1}^n |x_i - \mu_x|}{n}$$

**Standard Deviation for Population:**

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu_x)^2}{n}}$$

**MAD for Sample:**

$$MAD_x = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n-1}$$

**Standard Deviation for Sample:**

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

## Graphical Representations

- **Box-Plots:** Present the minimum, first quartile, median, third quartile, and maximum.
- **Usage:** Describe how symmetric/skewed the distribution of an attribute is.

### Likert Scale

- **Definition:** An ordered scale used in surveys, often interpreted as a quantitative scale.

### Conclusion

- **Application:** Choose the appropriate statistic based on the type of scale and the nature of the data.
- **Robustness:** Consider the robustness of each statistic against extreme values and sparsity.

These statistics and measures help in summarizing and understanding the characteristics of a dataset, providing valuable insights for further analysis and decision-making.

## Common Univariate Probability Distributions

### Uniform Distribution

- **Definition:** The uniform distribution describes a situation where all outcomes are equally likely within a specified range. For example, if you roll a fair six-sided die, each number (1-6) has an equal probability of 1/6.
- **Notation:** Denoted as  $X \sim U(a, b)$ , where  $a$  and  $b$  are the minimum and maximum values of the range, respectively.
- **Probability Density Function (PDF):** The PDF of a uniform distribution is constant within the interval  $[a, b]$  and zero elsewhere. It is given by:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

- **Mean and Variance:** For a uniform distribution, the mean  $\mu$  and variance  $\sigma^2$  are calculated as:

$$\mu = \frac{a + b}{2}, \quad \sigma^2 = \frac{(b - a)^2}{12}$$

### Normal Distribution (Gaussian Distribution)

- **Definition:** The normal distribution is a continuous probability distribution that is symmetric around its mean, forming a bell-shaped curve. Many natural processes follow this distribution.
- **Notation:** Denoted as  $X \sim N(\mu, \sigma^2)$ , where  $\mu$  is the mean and  $\sigma^2$  is the variance.
- **Probability Density Function (PDF):** The PDF of the normal distribution is given by the formula:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

This function describes the relative likelihood of observing a value  $x$ .

- **Mean and Variance:** For the normal distribution, the mean  $\mu$  locates the center of the distribution, and the standard deviation  $\sigma$  determines the spread or width of the distribution.

### *Central Limit Theorem*

- The central limit theorem states that the distribution of the sum (or average) of a large number of independent, identically distributed random variables approaches a normal distribution, regardless of the original distribution of the variables. This theorem is fundamental in statistics, as it explains why the normal distribution is so common in nature and in statistical analysis.

Understanding these fundamental probability distributions is crucial for various fields, including statistics, data science, and machine learning, as they form the basis for modeling and analyzing data.