MACHINE LEARNING

Shoaib Farooq

Department of Computer Science

Instructor Information

Instructor	Shoaib Farooq	E-mail	m.shoaib1050@gmail.com
 GitHub	https://github.com/SHOAIB1050		
 YouTube	https://www.youtube.com/c/pakithub		
 LinkedIn	https://www.linkedin.com/in/shoaib-farooq-b5b190105/		

Let's Start .....

Lecture #3

Goals

This Lecture Will Cover:

- Data And Types
- Structured Data
- Unstructured Data
- Semi-Structured Data
- What Is Data Preprocessing?
- Exploratory Data Analysis (EDA)
- Importance Of EDA
- Steps In EDA
- Mean, Median, Mod

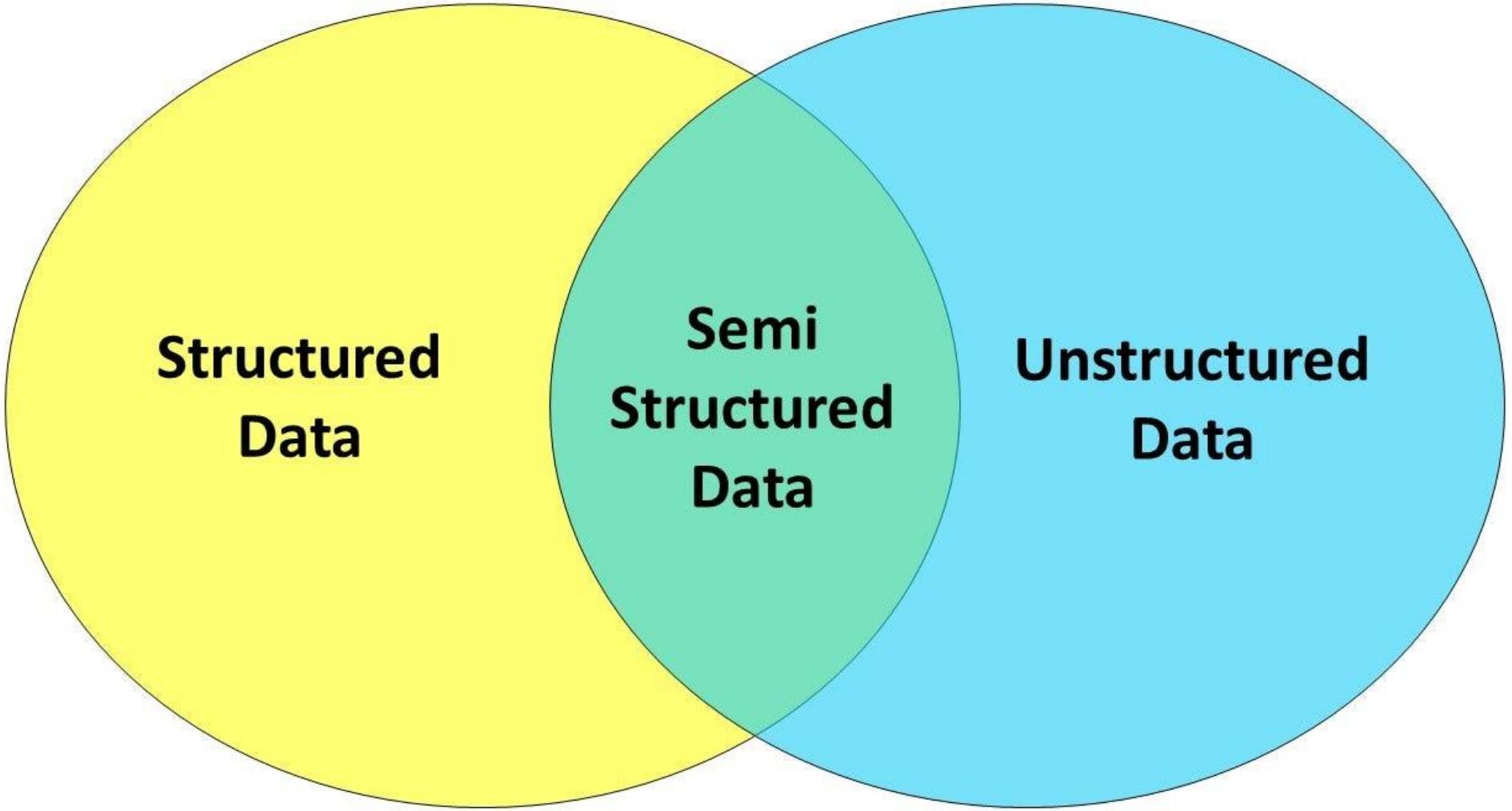
The Machine Learning Process

DATA AND TYPES

the term "data" refers to raw facts, observations, measurements, or information that can be collected and analyzed. Data is the foundation of any data-driven decision-making process and is essential for generating insights, making predictions, and uncovering patterns or trends.

- Structured Data
- Unstructured Data
- Semi-Structured Data

DATA AND TYPES



STRUCTURED DATA

- Structured data is organized and follows a **predefined format**, usually stored in **databases** or **spreadsheets**.
- **Example:** A customer database with columns for name, age, email, and purchase history.

  - Ordinal data
  - Nominal data
  - Numerical data

STRUCTURED DATA

STRUCTURED DATA

- **Ordinal Data:**
  - Ordinal data represents categories with a meaningful order or ranking.
  - The intervals between the categories are not uniform or measurable.
  - Examples include education levels (e.g., high school, college, graduate school) or customer satisfaction ratings (e.g., "low," "medium," "high").

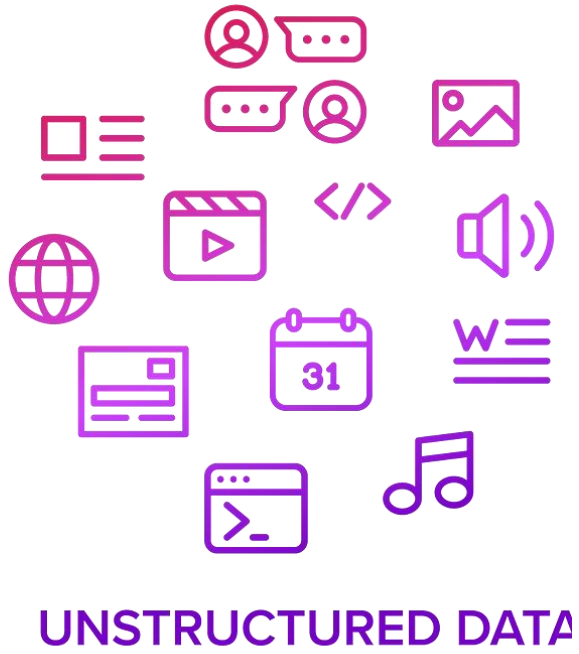
- **Nominal Data:**
  - Nominal data represents categories or labels with no inherent order or ranking among them.
  - **Examples** include gender, color, or types of fruits.

STRUCTURED DATA

- **Numerical Data:**
  - Numerical data, also known as quantitative data, consists of numerical values that represent measurable quantities.
  - Numerical data can be further divided into two subtypes: **discrete and continuous.**
  - **Discrete numerical data** consists of separate, distinct values, often counted in whole numbers (e.g., the number of cars in a parking lot).
  - **Continuous numerical data** can take any value within a given range and can have decimal or fractional parts (e.g., height, weight, temperature).
  - Numerical data allows for mathematical operations such as addition, subtraction, multiplication, and division.

UNSTRUCTURED DATA

- Unstructured data is **not organized** and lacks a predefined format, often in the form of **text, images, audio, or video**.
- **Example:** Social media posts, customer reviews, or images from a surveillance camera.



Semi-Structured Data

- Semi-structured data has some organization but does not adhere to a **strict schema**, often **containing tags** or **labels.**
- **Example:** Emails, JSON files that contain data with tags or key-value pairs.

Structured Data

Often numbers or labels, stored in a structured framework of columns and rows relating to pre-set parameters.

- ID CODES IN DATABASES
- NUMERICAL DATA GOOGLE SHEETS
- STAR RATINGS

Semi-unstructured Data

Loosely organized into categories using meta tags

- EMAILS BY INBOX, SENT, DRAFT
- TWEETS ORGANIZED BY HASHTAGS
- FOLDERS ORGANIZED BY TOPIC

Unstructured Data

Text-heavy information that's not organized in a clearly defined framework or model.

- MEDIA POSTS, EMAILS, ONLINE REVIEWS
- VIDEOS, IMAGES
- SPEECH, SOUNDS

WHAT IS DATA PREPROCESSING?

Data preprocessing is the process of **cleaning**, **transforming**, and **organizing raw** data to make it suitable for analysis and machine learning models.

Importance: High-quality data preprocessing is crucial for accurate and meaningful insights.

COMMON STEPS IN DATA PREPROCESSING

- Importing the Data
- Handling Missing Data
- Handling Duplicate Data
- Handling Outliers
- Encoding Categorical Variables
- Scaling and Normalization

COMMON STEPS IN DATA PREPROCESSING

- **Importing the Data:**
  - Load the dataset into your programming environment.
  - Use libraries like Pandas to handle data structures.

- **Handling Missing Data:**
  - Identify and handle missing values in the dataset.
  - Options include removal, imputation (the process of replacing missing data with substituted values), or using default values.

- **Handling Duplicate Data:**
  - Check for and remove duplicate entries in the dataset.
  - Duplicate data can deform analysis and machine learning models.

COMMON STEPS IN DATA PREPROCESSING

- **Handling Outliers:**

Outliers are data points that differ significantly from other observations.

It may be due to:

Errors (data entry mistakes, sensor issues, etc.)

True variability (e.g., very rich person's income in salary dataset).

Example:

Heights in cm: [160, 162, 165, 170, 400] => 400 is an **outlier**.

COMMON STEPS IN DATA PREPROCESSING

- **Handling Outliers:**

Why Outliers Matter?

Mislead ML models (especially models based on distance: k-NN, clustering, regression).

Increase error rate and reduce model performance.

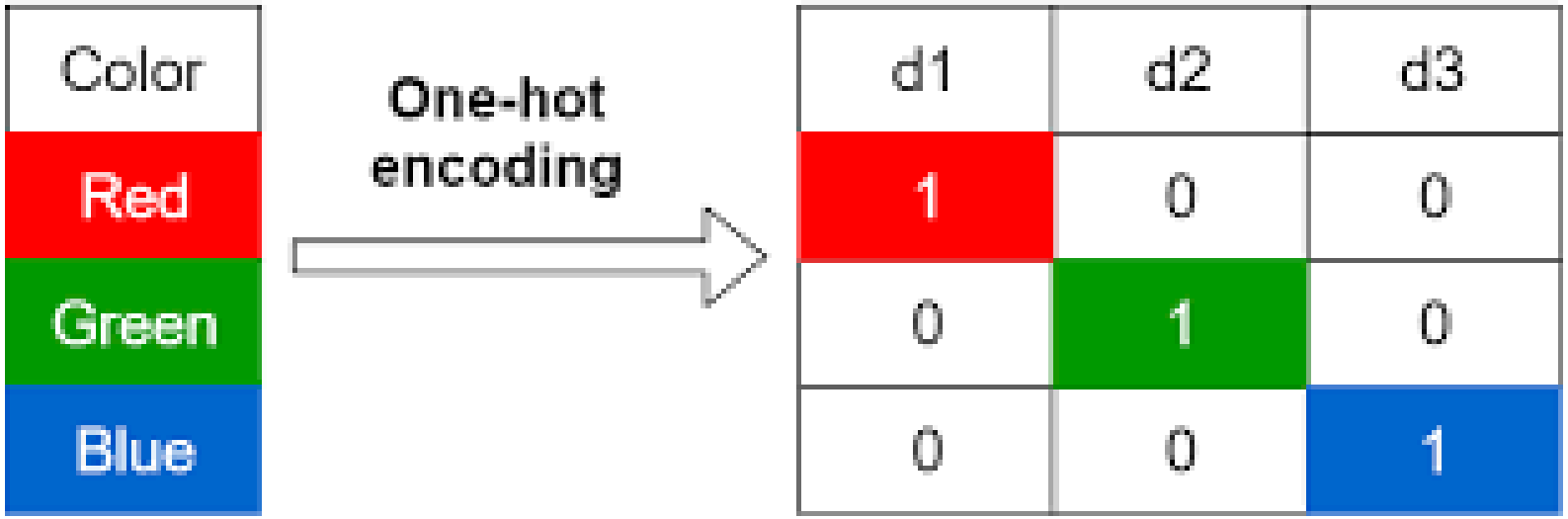
Methods for Handling Outliers

- **Z-score method**:

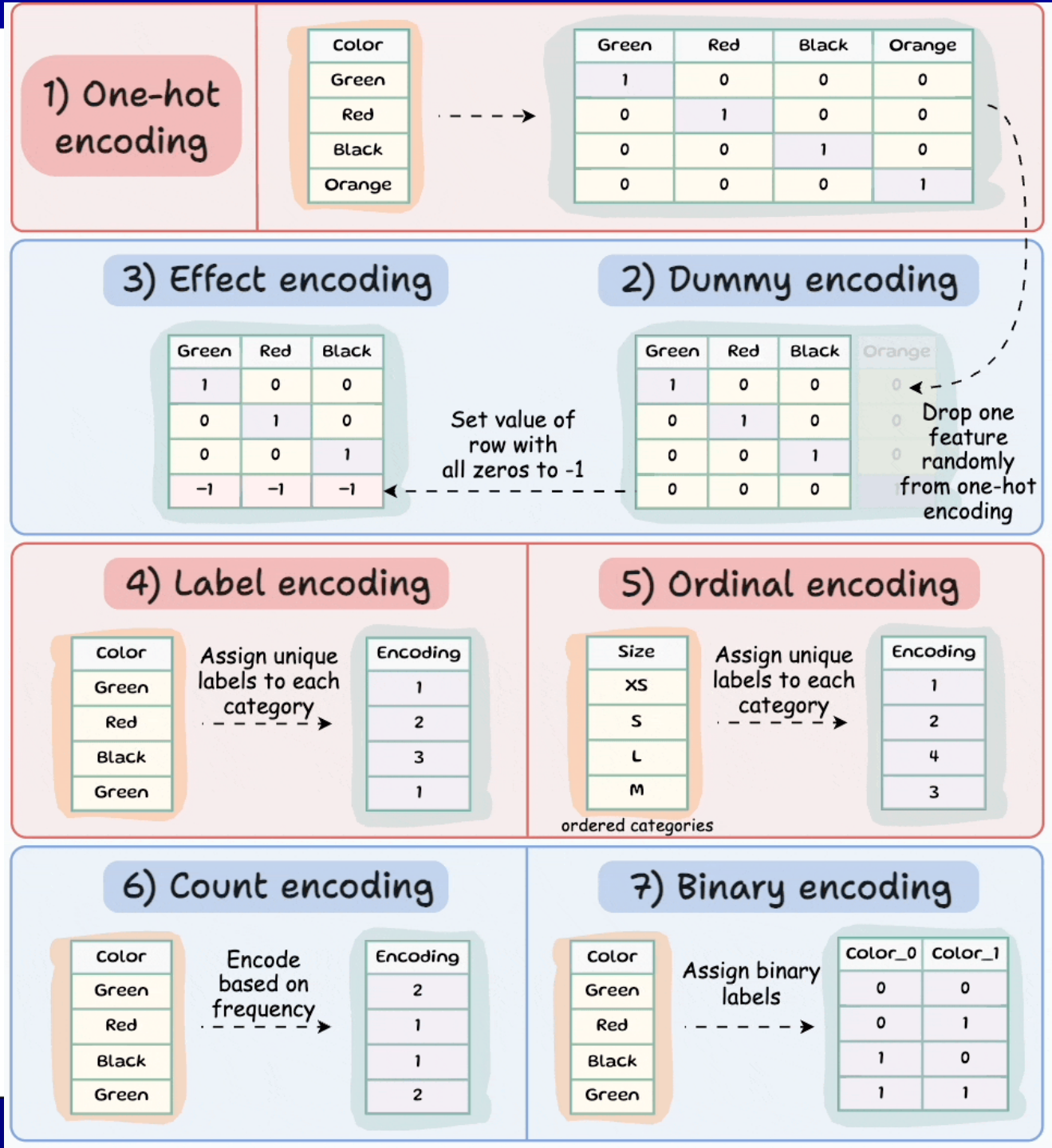
$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \Rightarrow \quad \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \quad \Rightarrow \quad |Z| = \frac{X - \mu}{\sigma} > 3$$

COMMON STEPS IN DATA PREPROCESSING

- **Encoding Categorical Variables:**
  - Convert categorical variables into numerical format.
  - Utilize techniques like one-hot encoding or label encoding.



Encoding Categorical Variables



COMMON STEPS IN DATA PREPROCESSING

➢ Scaling and Normalization:

When working with **machine learning models**, our input features (data) often have very different **ranges** and **units**.

For example:

- ❑ Age: 18–90
- ❑ Salary: 20,000–500,000
- ❑ Height: 1.5–2.1

If you feed such data directly into a model, features with large values (like salary) may **dominate** features with small values (like height).

- ➢ Standardize numerical features to the same scale.
- ➢ Methods include Min-Max scaling or Z-score normalization.

COMMON STEPS IN DATA PREPROCESSING

1. Feature Scaling

Scaling means **adjusting the range** of numerical features so they are on a similar scale.

Types of Scaling

a) Min-Max Scaling (Normalization to [0,1])

Formula:

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

Scales values between 0 and 1.

Sensitive to outliers.

Min-Max Scaling : when you need data between **0–1**, often in **neural networks**.

COMMON STEPS IN DATA PREPROCESSING

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

b) Standardization (Z-score Normalization)

- Mean = 0, Standard Deviation = 1
- Useful when data follows **Gaussian distribution**.
- Less sensitive to outliers compared to min-max.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Standardization (Z-score) : when data is normally distributed, good for **SVM, Logistic Regression, PCA**

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

COMMON STEPS IN DATA PREPROCESSING

c) MaxAbs Scaling

Formula:

$$X_{scaled} = \frac{X}{|\ X_{max}\ |}$$

- Scales values between **-1 and 1**.
- Works well for sparse data (a dataset where most values are zeros, nulls, or other indicators of missing information, meaning the data is not fully populated.)

EXPLORATORY DATA ANALYSIS (EDA)

The process of **exploring and summarizing** the main characteristics of the data to uncover **patterns, relationships, and trends.**

It helps in formulating questions and making data-driven decisions.

IMPORTANCE OF EDA

- Provides an **initial understanding** of the dataset.
- Helps in identifying data **quality issues**, such as **missing values**, **outliers**, and **inconsistencies**.
- Guides the selection of appropriate statistical techniques and **models**.
- Helps in feature engineering and **variable selection**.
- Enables the discovery of **meaningful insights** and actionable conclusions.

STEPS IN EDA

- Data Cleaning
- Descriptive Statistics
- Data Visualization
- Data Distribution
- Correlation Analysis
- Outlier Detection
- Data Transformation

STEPS IN EDA

- **Data Cleaning:**
  - Identify and address errors, inconsistencies, and missing values in the dataset.
  - Ensure data integrity and reliability for analysis.

- **Descriptive Statistics:**
  - Summarize and describe key characteristics of the dataset.
  - Include measures like mean, median, standard deviation, etc.

- **Data Visualization:**
  - Represent data visually using charts, graphs, or plots.
  - Enhance understanding and identify patterns in the data.

STEPS IN EDA

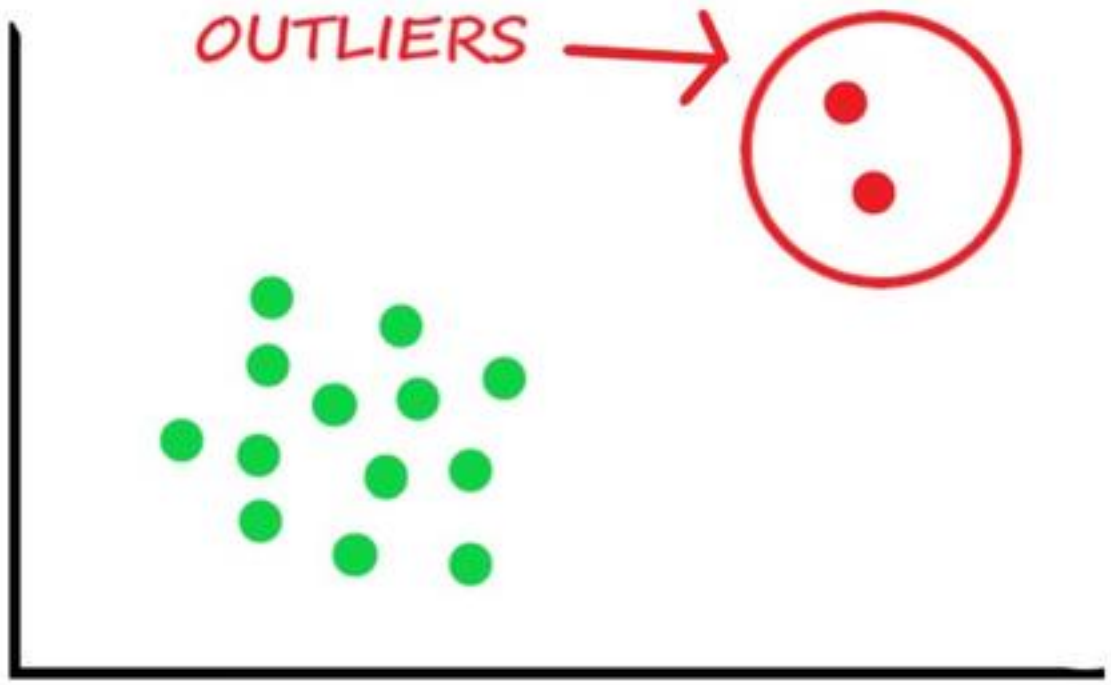
- **Data Distribution:**
  - Analyze the distribution of data points.
  - Understand the spread and central tendency of the dataset.

- **Correlation Analysis:**
  - Evaluate relationships between variables.
  - Use correlation coefficients to measure the strength and direction of associations.

STEPS IN EDA

- **Outlier Detection:**
  - Identify and handle data points that deviate significantly from the norm.
  - Consider statistical methods or visualization techniques.
- **Data Transformation:**
  - Modify the structure or values of the data.
  - Examples include scaling, encoding, or creating derived features.

OUTLIER



MEAN

The mean is the **average** of a set of numbers.

Calculated by summing all the numbers in the dataset and dividing by the total count.

Formula: Mean = (Sum of all numbers) / (Total count)

Example: For the dataset **[5, 7, 10, 12, 15],** the mean is **(5 + 7 + 10 + 12 + 15) / 5 = 9.8.**

Sensitive to outliers

Example: For the dataset **[5, 7, 10, 12, 150],** the mean is **(5 + 7 + 10 + 12 + 150) / 5 = 36.8.**

MEDIAN

The median is the **middle value** in a **sorted** dataset.

If the dataset has an odd number of values, the median is the middle number.

If the dataset has an even number of values, the median is the average of the two middle numbers.

Example 1: For the dataset **[3, 5, 6, 8, 9]**, the median is **6.**

Example 2: For the dataset **[2, 4, 6, 8]**, the median is **(4 + 6) / 2 = 5.**

MOD

The mode is the value that appears **most frequently** in a dataset.

A dataset can have one mode, more than one mode (multimodal), or no mode (no value repeats).

Example 1: For the dataset **[3, 4, 4, 6, 8]**, the mode is **4.**

Example 2: For the dataset **[1, 2, 3, 4, 5]**, there is **no mode**.

QUARTILES AND INTERQUARTILE RANGE (IQR)

Quartiles: Before understanding the IQR, we need to grasp the concept of quartiles.

Quartiles divide a dataset into **four equal parts**, each representing a quarter **(25%)** of the data.

The first quartile **(Q1)** is the value below which **25%** of the data falls.

The second quartile **(Q2)** is the value below which **50%** of the data falls.

It's actually **Median.**

The third quartile **(Q3)** is the value below which **75%** of the data falls.

QUARTILES AND INTERQUARTILE RANGE (IQR)

To find the quartiles, first, arrange the dataset in **ascending order**.

If the dataset has an odd number of values, the median is the second quartile (Q2).

If the dataset has an even number of values, calculate the median of the lower and upper halves to find Q2.

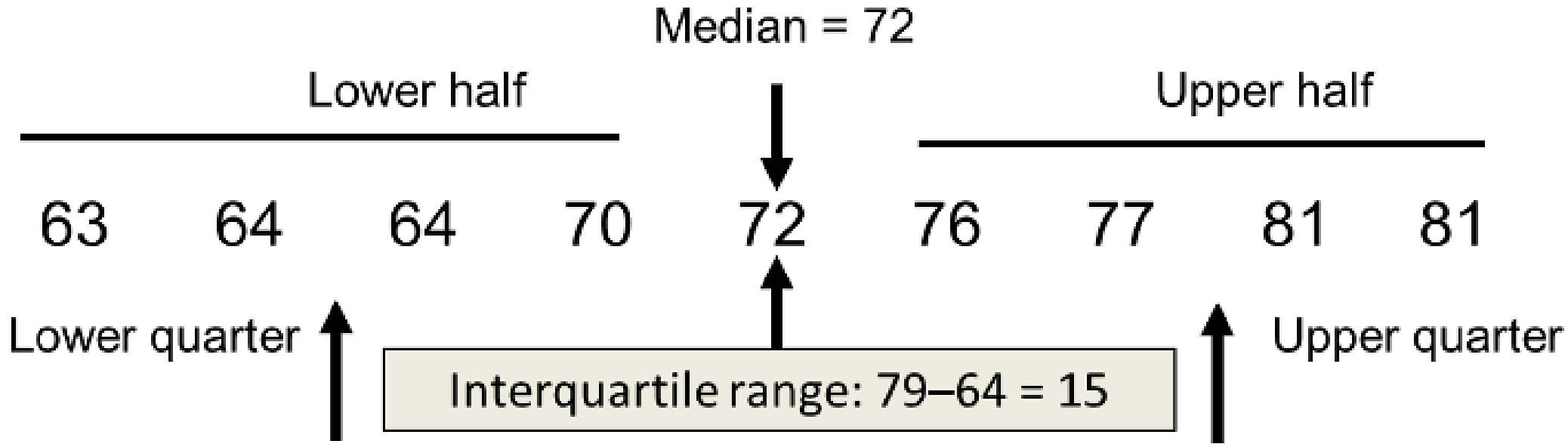
QUARTILES AND INTERQUARTILE RANGE (IQR)

The IQR represents the difference between the first and third quartiles(Q1 and Q3).

It indicates the spread of the middle 50% of the data.

Formula: IQR = Q3 - Q1

QUARTILES AND INTERQUARTILE RANGE (IQR)

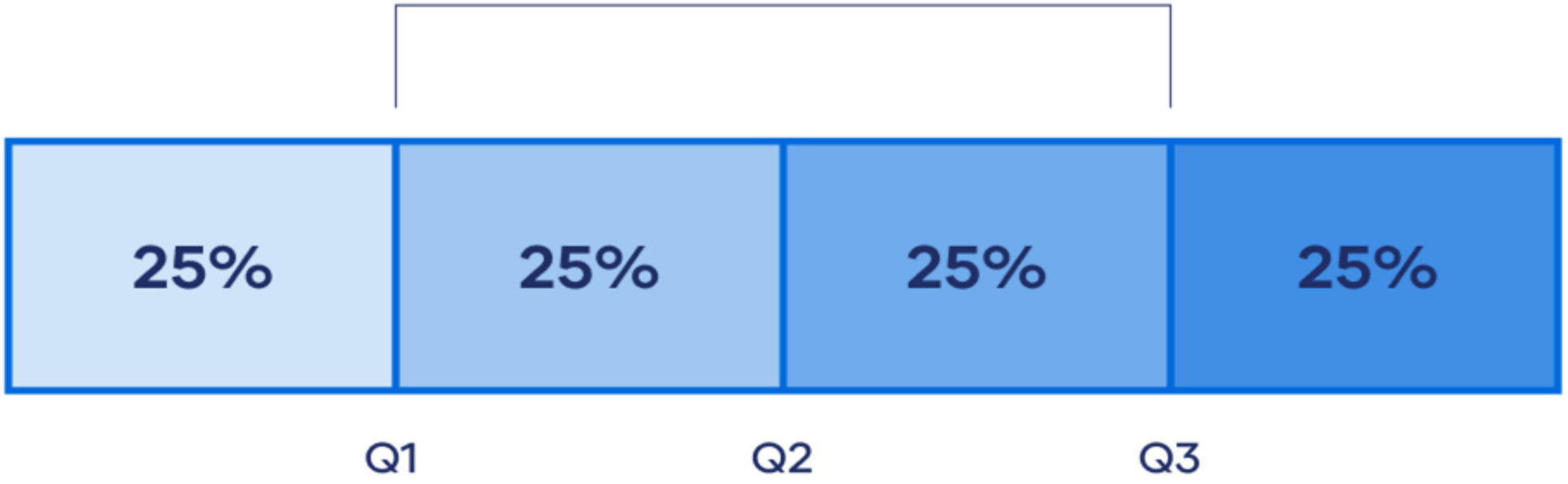


$Q_1 = (64+64)/2=64$

$Q_3 = (77+81)/2=79$

QUARTILES

Quartiles: Quartiles divide a dataset into **four equal parts**, each representing a quarter **(25%)** of the data.



Thank You ☺