# **Exploratory Data Analysis (EDA)**

## **What is EDA?**

Exploratory Data Analysis (EDA) involves summarizing and visualizing data to understand its main characteristics. It helps in identifying patterns, trends, relationships, and anomalies, making data interpretation easier and aiding in better decision-making.

### **Key Objectives of EDA:**

* **Understanding data distribution** through summary statistics.
* **Identifying missing values and anomalies** in the dataset.
* **Visualizing relationships** between variables.
* **Detecting patterns, trends, and outliers** in the data.
* **Facilitating better decision-making** by simplifying complex data.

## **Data Summaries**

### **1. Descriptive Statistics:**

Descriptive statistics help summarize and describe the dataset using numerical measures:

* **Measures of Central Tendency**:
  + **Mean**: The average value of a dataset.
  + **Median**: The middle value when data is sorted.
  + **Mode**: The most frequently occurring value in the dataset.
* **Measures of Dispersion**:
  + **Variance**: Measures how far the data points are from the mean.
  + **Standard Deviation**: The square root of variance; indicates data spread.
  + **Range**: Difference between the maximum and minimum values.
  + **Interquartile Range (IQR)**: The range of the middle 50% of data (Q3 - Q1).

## **Types of Analysis in EDA**

### **1. Univariate Analysis**

* Examines a **single** variable.
* Helps understand the distribution, central tendency, and spread of data.
* Common visualizations:
  + Histograms
  + Boxplots
  + Bar charts
  + Pie charts

### **2. Bivariate Analysis**

* Examines **two** variables together.
* Helps identify relationships and correlations between variables.
* Common visualizations:
  + Scatter plots (numerical vs. numerical)
  + Line charts (time series trends)
  + Boxplots (categorical vs. numerical)
  + Bar charts (categorical vs. numerical)

### **3. Multivariate Analysis**

* Examines **more than two** variables simultaneously.
* Helps detect complex interactions and dependencies between variables.
* Common visualizations:
  + Pair plots
  + Heatmaps (correlation matrices)
  + Bubble charts
  + 3D scatter plots

## **Data Visualization Techniques**

### **1. Line Chart**

* Shows trends over time.
* Used to compare two numerical variables.

### **2. Histogram**

* Analyzes the distribution of a **single numerical variable**.
* Helps identify skewness, modality, and outliers.

### **3. Bar Chart**

* Used to compare categorical and numerical data.
* Displays frequency or proportion of categories.

### **4. Boxplot (Box-and-Whisker Plot)**

* Used to analyze the distribution of a **single variable**.
* Shows the median, quartiles, and outliers.

### **5. Pie Chart**

* Represents proportions of different categories in a dataset.
* Best for showing percentage distributions.

### **6. Scatter Plot**

* Used for **bivariate analysis** to examine relationships between two numerical variables.
* Helps detect correlations and trends.

### **7. Pair Plot**

* A grid of scatter plots showing relationships between multiple variables.
* Helps in multivariate analysis.

### **8. Bubble Chart**

* Similar to a scatter plot but with a third variable represented by bubble size.
* Useful for multivariate analysis.

## **Importance of EDA**

* Helps **understand data quality** (missing values, duplicates, outliers).
* **Validates assumptions** and identifies potential data errors.
* Supports **feature selection** and engineering for modeling.
* Enhances **data-driven decision-making** by summarizing key insights visually.