**Exploratory Data Analysis (EDA)**

### 1. What is EDA?

Exploratory Data Analysis (EDA) is the first and most essential step in data science. It involves examining datasets to summarize their main characteristics, often using visual methods.

* **Purpose of EDA:**
* Understand structure and content of data
* Detect outliers, missing values, and anomalies
* Identify relationships between variables
* Prepare for modeling
* **Why it Matters:**
* Ensures that data is clean, complete, and ready for analysis
* Helps avoid incorrect assumptions
* Supports better decision-making
* **Common Questions EDA Helps Answer:**
* What types of variables exist in the dataset?
* Are there missing or duplicated values?
* Are features skewed or normally distributed?
* Are there any variables that correlate strongly with each other?

### 2. Summary Statistics

* **Mean**: The average value.
* Formula: **Mean = (Sum of values) / (Number of values)**
* Indicates the central tendency of the data
* **Variance**: Measures how far each number is from the mean.
* Formula: **Variance = sum((x - mean)^2) / (n - 1)**
* High variance means data is spread out
* **Standard Deviation**: **Square root of variance.**
* Expresses deviation in the same unit as the data
* Useful for understanding spread in a more interpretable way
* **Real-world Use Cases:**
* Finance: Measuring volatility of stock prices
* Machine learning: Normalizing features
* Quality control: Monitoring consistency in production

### 3. Correlation Analysis

* **Correlation** measures how two variables move together.

Ranges from -1 to +1

+1: Strong positive (both increase together)

0: No correlation

-1: Strong negative (one increases, other decreases)

**Types of Correlation:**

* Pearson: Measures linear correlation (most common)
* Spearman: Ranks the data before correlation
* **Heatmap**: A color-coded matrix showing correlation values between variables.
* Darker = stronger relationship
* Visual way to detect multicollinearity
* **Applications:**
* Feature selection in ML
* Understanding variable relationships for hypothesis building

### 4. Identifying Outliers

**Outlier**: A value significantly different from others. Can distort summary statistics and affect model performance.

#### IQR (Interquartile Range) Method

The **Interquartile Range (IQR)** is the range where the **middle 50%** of the data lies. It helps us find out if a number is too far from most of the data (an outlier).

IQR = Q3 - Q1 (range of middle 50% of data)

Outlier if:

Value < Q1 - 1.5 \* IQR

Value > Q3 + 1.5 \* IQR

**Box Plot**: Visual tool to show distribution and outliers

Displays min, Q1, median, Q3, max, and outliers

#### Z-Score Method

Z-Score = (value - mean) / standard deviation

Indicates how many standard deviations a value is from the mean

|Z| > 3 is typically considered an outlier

#### Scatter Plot

Visual tool to examine data points and identify deviations

Effective for bivariate outlier detection

### 5. Categorical Data Analysis

**Categorical Data**: Data with labels (e.g., Gender, Fruit)

**Value Counts**: Shows how many times each category occurs **Frequency Table**: Shows category distribution in percentage

#### Visual Tools:

* **Bar Plot**: Displays counts of categories with bars
* **Count Plot**: Similar to bar plot but often with color and styling (Seaborn)

#### Encoding Techniques:

**Label Encoding**: Categories → Numbers (e.g., Apple → 0)

Can be misleading if categories are unordered

**One-Hot Encoding**: Creates binary columns for each category

Preferred for nominal variables

### Automated EDA Tools :

Automated **Exploratory Data Analysis (EDA)** tools help generate quick insights into datasets with minimal coding. They provide summaries, distributions, correlations, missing values, and visualizations in one report.

#### Pandas-profiling / ydata-profiling

* Auto-generates comprehensive reports
* Includes summary statistics, correlations, missing value matrix, histograms, and warnings

#### Sweetviz

* Creates high-contrast visual reports
* Allows comparison between training and testing sets
* Highlights features that impact target variable

#### D-Tale

* Interactive GUI for Pandas DataFrames
* Allows sorting, filtering, statistics, and visualizations directly in browser