# Unit 1: Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a critical step in understanding and analyzing data before any modeling or formal hypothesis testing. EDA involves summarizing the main characteristics of data, identifying patterns, detecting anomalies, and testing assumptions. Below are detailed notes on the sub-topics for a better understanding of EDA.

## *1. Introduction to Single Variables*

When analyzing a single variable, we are interested in summarizing and understanding the characteristics of the data associated with that variable. This is done using basic statistical tools like mean, median, mode, range, etc.

- **Definition**: A single-variable analysis focuses on exploring and summarizing one variable from the dataset.
- **Common Measures**:
  - **Mean**: The average value of the data points.
    - **Formula**: $Mean = \frac{\sum xi}{n}$
    - **Example:** For the values 4, 5, 7, the mean is $\frac{4+5+7}{3} = 5.33$
  - **Median**: The middle value when the data is ordered.
    - Example: For the values 3, 7, 9, the median is 7.
  - **Mode**: The most frequent value in the data.
    - Example: In the dataset 2, 3, 3, 4, 3 is the mode.

## *2. Distribution Variables*

A distribution represents how frequently values occur within a dataset. Understanding the shape of the distribution is key to analyzing the behavior of the data.

- **Definition**: The distribution of a variable shows how its values are spread or clustered across different intervals.
- **Types of Distributions:**
  - **Symmetrical**: Data is evenly distributed (e.g., Normal Distribution).
  - **Skewed**: Data is concentrated on one side.

        - Right-skewed: Long tail on the right.
        - Left-skewed: Long tail on the left.
    - **Kurtosis**: Describes the "tailedness" of the distribution.

- **Example**: If we examine the income of individuals, we might find that most people earn around a certain amount, but a few individuals earn significantly more, leading to a right-skewed distribution.

### *3. Numerical Summaries of Level and Spread*

Numerical summaries are essential in summarizing the central tendency (level) and the spread of data.

- **Central Tendency** (Level): Describes where the center of the data lies.
    - **Mean**, **Median**, **Mode** (as explained earlier).
- **Spread**: Describes how far data points are from the center.
    - **Range**: The difference between the maximum and minimum values.
        - Formula: **Range = Max – Min**
    - **Variance:** Measures the spread of data points around the mean
        - Formula**:** $\boldsymbol{Variance = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}}$
        - Example: For the data 2, 4, 6, 8, the mean is 5. Variance:

$$\frac{(2-5)^2+(4-5)^2+(6-5)^2+(8-5)^2}{4}=5$$

    - **Standard Deviation**:

      The square root of the variance, providing a measure of the typical distance of data points from the mean.

$$Standard\ Deviation = \sqrt{\text{Variance}}$$

      **Example**:

      For the data 2, 4, 6, 8, the standard deviation is: $\sqrt{5} \approx 2.24$

**Example (Numerical Summaries)**:

Given the data 55, 60, 65, 70, 75, calculate the mean, range, variance, and standard deviation:

- Mean = 65
- Range = 75 - 55 = 20
- Variance = 50
- Standard Deviation = $\sqrt{50} \approx 7.07$

## *4. Scaling and Standardizing*

**Definition**:
Scaling and standardizing are preprocessing techniques used to adjust the values of numerical variables to make them comparable, especially when different variables have different units or scales.

**Scaling**:

Scaling adjusts the range of data to a specific range, usually between 0 and 1. This is done using **Min-Max Scaling**.

- **Min-Max Scaling Formula**: $x' = \frac{x - Min(x)}{Max(x) - Min(x)}$
- **Example**:
  If data has a range from 10 to 100, and the value is 40, the scaled value is:

$$\frac{40-10}{100-10} = 0.33$$

**Standardizing:**

Standardizing transforms data so that it has a mean of 0 and a standard deviation of 1. This is useful when comparing variables with different units or ranges.

- **Z-score Standardization Formula**: : $z = \frac{x-\mu}{\sigma}$

- **Example**:

  For a value of 50, with a mean of 40 and a standard deviation of 10, the z-score is: $z = \frac{50-40}{10} = 1$ This means the value is 1 standard deviation above the mean.

### *5. Inequality*

**Definition**:

Inequality measures how evenly or unevenly values are distributed, often used in economics to describe income or wealth distribution.

**Key Metrics**:

- **Gini Coefficient**:

  A measure of inequality that ranges from 0 (perfect equality) to 1 (perfect inequality). It is often used to measure income or wealth inequality.

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n^2\mu}$$

  **Example**:

  A high Gini coefficient (e.g., 0.7) indicates high inequality, while a low Gini coefficient (e.g., 0.2) indicates more equality.

- **Lorenz Curve**:

  A graphical representation of inequality. It plots the cumulative percentage of total income or wealth against the cumulative percentage of the population.

**Example Question**

"Given the dataset of test scores, calculate the mean, median, mode, range, variance, and standard deviation. Then, apply a 3-point moving average to smooth the data and comment on the results."