**Name: Sumit Avinash Solanke**

**Roll No: TYITB121**

**Assignment No. 1**

**AIM:** Exploring Data Analysis

**PREREQUISITE:** Statistics and Python programming

**THEORY:**

Exploratory Data Analysis (EDA) is an essential initial step in data analysis that helps in understanding the underlying patterns, detecting anomalies, and summarizing the key characteristics of the dataset. The primary objectives of EDA include:

* Detecting mistakes in data collection and processing.
* Checking assumptions related to data distribution and relationships.
* Selecting appropriate statistical models.
* Identifying relationships between variables.
* Assessing the size and direction of relationships between variables.

EDA primarily involves graphical and statistical methods, but it does not focus on formal statistical inference.

**Measures of Central Tendency**

The central tendency of a dataset refers to its middle or typical values. The most commonly used measures of central tendency are:

* **Mean:** The arithmetic mean is calculated as the sum of all values divided by the number of observations.
* **Median:** The median is the middle value in an ordered dataset. If the number of observations is even, the median is the average of the two middle values.
* **Mode:** The mode is the most frequently occurring value in the dataset. A dataset may have one mode (unimodal), multiple modes (multimodal), or no mode at all.

In addition to these, specialized means like the geometric mean, harmonic mean, and trimmed mean are used in certain statistical applications.

**Measures of Variability**

The variability or dispersion of a dataset indicates how spread out the values are. Some common measures include:

* **Variance (s²):** Variance quantifies the average squared deviation from the mean: It provides insight into how much the data points deviate from the central value.
* **Standard Deviation (s):** The standard deviation is the square root of variance and retains the same unit as the original data, making it more interpretable: It helps in understanding the spread of data, particularly in normally distributed datasets.
* **Interquartile Range (IQR):** The IQR measures the spread of the middle 50% of the data and is calculated as: IQR=Q3−Q1 | QR = Q3 - Q1 where Q1Q1 (first quartile) is the 25th percentile and Q3Q3 (third quartile) is the 75th percentile of the dataset. IQR is useful for identifying outliers and assessing skewness in the data.

**Outlier Identification**

Outliers are extreme values that differ significantly from the rest of the data. They can be detected using:

* **Boxplots:** A boxplot visually represents the dataset’s distribution, showing the median, quartiles, and potential outliers. Outliers are typically identified as values lying beyond 1.5 times the IQR above Q3Q3 or below Q1Q1.
* **Z-Scores:** Standardizing data points using Z-scores helps in identifying values that deviate significantly (typically beyond ±3 standard deviations from the mean).

**Dataset Description:**

**Dataset Name:** Employee\_Salary\_Dataset.csv  
**Total Records:** *(Example: 100 entries — you can update this based on your actual data)*  
**Purpose:** To analyze the relationship between employee attributes like experience, age, gender, and their corresponding salary. This dataset is suitable for salary prediction and trend analysis.

**Columns and Their Descriptions:**

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| ID | Integer | Unique identifier for each employee |
| Experience\_Years | Integer | Total number of years of professional experience |
| Age | Integer | Current age of the employee in years |
| Gender | Categorical | Gender of the employee (Male, Female, or other values if present) |
| Salary | Float | Current annual salary of the employee in INR or USD (based on your dataset) |

**Sample Records (Preview):**

| **ID** | **Experience\_Years** | **Age** | **Gender** | **Salary** |
| --- | --- | --- | --- | --- |
| 1 | 3 | 25 | Male | 450000.00 |
| 2 | 7 | 32 | Female | 850000.00 |
| 3 | 1 | 22 | Male | 300000.00 |
| 4 | 10 | 40 | Female | 1200000.00 |
| 5 | 5 | 29 | Male | 650000.00 |

**Possible Exploratory Questions:**

* What is the average salary across the company?
* Is there a strong correlation between Experience\_Years and Salary?
* Do older employees tend to have higher salaries?
* What is the salary distribution by Gender?
* Are there any outliers in salary based on experience?
* What is the most common age group in the dataset?

**CONCLUSION:**

Exploratory Data Analysis is a crucial step in understanding and preparing data for modeling and decision-making. By using measures of central tendency and variability, along with visualization techniques like boxplots, we can detect patterns, assess relationships, and identify anomalies in datasets. Mastering EDA techniques is essential for effective data-driven insights and decision-making.