**Name: Sahil Vishwas Shendkar**

**Roll No: TYITB118**

**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. Instead, it lays the groundwork for further analysis by helping analysts familiarize themselves with the data and formulate hypotheses that can later be tested using formal modeling techniques.**

**DATABASE:**

**SALARY DATABASE - A database refers to a structured collection of data that is stored and accessed electronically. In the context of Exploratory Data Analysis, the database is the source from which the dataset is extracted for analytical purposes. The database often resides in formats such as SQL servers, CSV files, Excel sheets, or cloud-based platforms like Google BigQuery, MongoDB, or Amazon RDS.**

**In this assignment, the Salary Dataset is assumed to be extracted from a relational database, where each employee's record represents a row and features like salary, department, age, and education level are represented as columns. Databases ensure data integrity, efficient retrieval, and secure storage — which are critical for any type of analysis. Before performing EDA, the data must often be cleaned and transformed after retrieval from the database.**

**Dataset Description – Salary Dataset**

**In this assignment, we consider a Salary Dataset, which typically contains features like:**

* **EmployeeID – Unique identifier for each employee**
* **Department – Department in which the employee works (e.g., IT, HR, Finance)**
* **Experience (Years) – Number of years of professional experience**
* **Education Level – Qualification category (e.g., Bachelor's, Master's, PhD)**
* **Gender – Male or Female**
* **Age – Age of the employee**
* **Salary – Monthly or annual income of the employee**

**This dataset allows for exploratory analysis to identify patterns such as:**

* **How experience affects salary**
* **Whether there are salary disparities by gender**
* **If higher education levels lead to higher salaries**
* **Which departments offer the highest average pay**

**The salary variable (target) is continuous and serves as the focus for understanding central tendency, spread, and outlier behavior.**

**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. It represents the average value in the dataset and is sensitive to outliers.  
● 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. It is a robust measure and is not affected by extreme 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 to better suit data with specific characteristics.**

**Example from Salary Dataset:  
The mean salary gives us a general sense of income level.  
The median salary helps assess the central salary when data is skewed (e.g., due to a few very high-paying jobs).  
The mode may help identify commonly offered salary brackets.**

**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\text{IQR} = Q3 - Q1IQR=Q3−Q1**

**where Q1 (first quartile) is the 25th percentile and Q3 (third quartile) is the 75th percentile of the dataset. IQR is useful for identifying outliers and assessing skewness in the data.**

**Example from Salary Dataset:  
A high standard deviation in salary may suggest income inequality within the company.  
A large IQR could indicate that mid-level employees have varying pay scales based on department or experience.**

**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 Q3 or below Q1.  
● Z-Scores: Standardizing data points using Z-scores helps in identifying values that deviate significantly (typically beyond ±3 standard deviations from the mean).**

**Example from Salary Dataset:  
A few executives earning exceptionally high salaries will appear as outliers in a salary boxplot.  
Identifying these is crucial to prevent distortion in statistical analysis and model training.**

**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. In the case of the Salary dataset, EDA helps uncover pay distribution, detect outliers like excessively high or low salaries, and understand how factors like experience or education relate to income. The database from which this dataset is retrieved plays a vital role in ensuring data accuracy and consistency. Mastering EDA techniques is essential for effective data-driven insights and decision-making.**