Assignment No. 1

AIM: Assignment of exploring data analysis.

PREREQUISITE: Statistics and Python programming

Objective:

- To perform exploratory data analysis on a given dataset.
- To handle missing values and preprocess data effectively.
- To analyze feature relationships using correlation matrices.
- To apply encoding techniques to categorical variables.
- To scale numerical data for better model performance.
- To visualize data distribution and patterns using various graphical representations.

Theory:

Exploratory Data Analysis (EDA) is the first step in data analysis, where the dataset is examined for patterns, anomalies, missing values, and relationships among variables. The key objectives of EDA include:

- Understanding dataset structure.
- Identifying missing data and handling it appropriately.
- Detecting relationships between variables using correlation.
- Encoding categorical variables for model compatibility.
- Standardizing and normalizing numerical features.
- Visualizing data distribution, trends, and patterns.

Steps for EDA and Preprocessing:

1. Load the Dataset: Read data from CSV or other formats.
2. Basic Information: Display dataset overview, check missing values.
3. Handle Missing Data: Impute missing values using mean, median, or mode.
4. Analyze Correlation: Use correlation matrices to detect relationships.
5. Encoding Techniques: Convert categorical data into numerical form.
6. Scaling Data: Normalize or standardize numerical features.
7. Visualizations: Generate histograms, box plots, scatter plots, and heatmaps.

Algorithm:

1. Start

2. Load the dataset from a CSV or other file format.
3. Display basic information about the dataset, including shape, columns, and missing values.
4. Handle missing data:
    - Fill missing numerical values using mean/median.
    - Fill missing categorical values using mode.
5. Perform correlation analysis:
    - Compute correlation between numerical features.
    - Visualize using a heatmap.
6. Encode categorical variables:
    - Apply Label Encoding for ordinal data.
    - Apply One-Hot Encoding for nominal data.
7. Scale numerical data:
    - Apply StandardScaler or MinMaxScaler for feature scaling.
8. Visualize the dataset using:
    - Histograms for distribution analysis.
    - Box plots for detecting outliers.
    - Pair plots to analyze relationships between variables.
9. Display processed dataset.
10. End

Conclusion:

EDA and preprocessing are essential steps before applying machine learning models. Proper handling of missing values, feature encoding, scaling, and visualizing data helps in better model performance and decision-making.