Assignment No. 1

AIM: Assignment of exploring data analysis.

PREREQUISITE: Basic knowledge of statistics
Python programming (libraries like Pandas, NumPy, Matplotlib, Seaborn, scikit-learn)

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 an important step that helps in summarizing and understanding the main characteristics of the dataset before applying any machine learning model.
In this assignment, the Titanic dataset was loaded, and basic information such as number of rows, columns, data types, and missing values was analyzed.

Handling missing values is critical because machine learning models cannot work properly with incomplete data.
Here, missing numerical values were replaced with the mean, and missing categorical values were replaced with the mode to maintain the dataset's integrity.

To understand the relationships between different numerical features, a correlation matrix was created.
A heatmap was plotted based on this matrix because heatmaps give a quick visual summary of how features are correlated.
Using a heatmap, one can easily identify which variables have strong positive or negative relationships, which can be useful for feature selection and multicollinearity detection.

Label Encoding was applied to convert categorical data into a numerical format required for machine learning algorithms.
After encoding, StandardScaler was used to scale numerical features, ensuring that each feature contributed equally without being dominated by larger values.

Visualizations were an essential part of the EDA:

- Histograms were used to check the distribution of numerical features.

- Boxplots helped detect outliers in the data.

- Pair plots allowed the exploration of pairwise relationships between variables.
  By performing these steps, the dataset was cleaned, structured, and made suitable for further predictive analysis.

  Dataset Name: `Titanic Dataset`

Algorithm:

1.Start

2.Import necessary libraries such as pandas, numpy, matplotlib, seaborn, and sklearn for preprocessing.

3.Load the dataset into a pandas DataFrame.

4.Display basic information about the dataset:

- Shape of the dataset
- Column names
- Data types
- Check for missing values

5.Handle missing data:

- Fill missing numerical values with the mean of the respective columns.
- Fill missing categorical values with the mode (most frequent value).

6.Analyze correlation between numerical features:

- Compute the correlation matrix.
- Visualize the correlation matrix using a heatmap.

7.Encode categorical variables:

- Apply Label Encoding to convert categorical columns into numerical format.

8.Scale numerical features:

- Apply StandardScaler to standardize the numerical columns.

9.Visualize the dataset:

- Generate histograms for understanding feature distributions.
- Generate box plots for detecting outliers.
- Generate pair plots to analyze relationships between variables.

10.Display the processed dataset.

11.End

Conclusion:

Exploratory Data Analysis and preprocessing are crucial for preparing data before machine learning.In this assignment, missing values were handled, categorical data was encoded, numerical features were scaled, and key patterns were identified through visualizations.Correlation analysis and heatmaps revealed important feature relationships.Through systematic EDA, the Titanic dataset was cleaned, structured, and made ready for accurate predictive modeling.