**Name: Abhishek Rajaram Walkoli**

**PRN: 123B2F155**

**Assignment No. 1**

**Aim**

Exploring Data Analysis and Preprocessing

**Objective**

The aim of the process is to perform a thorough Exploratory Data Analysis (EDA) and preprocessing on a dataset so as to have profound insights into its pattern, identify any anomalies in it, and prepare it well for machine learning models. It includes missing data handling, correlation analysis, categorical variable encoding, and data visualization through charts, graphs, and heatmaps in order to find patterns and trends.

**Prerequisites**

To effectively conduct EDA and preprocessing, you require:

1. A Python environment with the necessary libraries like pandas, numpy, matplotlib, seaborn, sklearn, and xml.etree.ElementTree for handling structured data.

2.Access to the internet to download datasets from the web if needed.

3. A text editor or Jupyter Notebook and a basic knowledge of Python and data exploration methods.

**Steps for Exploratory Data Analysis (EDA) and Preprocessing**

**1. Understanding the Dataset**

Prior to performing any analysis, it is important to inspect the structure and content of the dataset in order to recognize possible problems and preprocessing requirements. This includes:

●Verifying the Number of Rows and Columns:

○Use.shape to find out dataset dimensions (number of rows and columns).

○Big data can be optimized using feature selection to enhance efficiency, whereas smaller data can utilize augmentation methods to enhance sample size.

●Checking Data Types:

○Apply.info() to know the types of data involved (numerical, categorical, object/string values).

○This aids in establishing whether encoding methods such as label encoding or one-hot encoding will be needed for categorical data.

●Missing Values Detection:

○Apply.isnull().sum() to identify missing values in every column.

○Missing values can cause machine learning model biases, and it is important to handle them accordingly.

● Basic Statistical Analysis:

○ Apply .describe() to calculate mean, median, standard deviation, and other stats metrics.

○ Detect skewness and required transformations (e.g., log transformation if the data is significantly skewed).

**2. Dealing with Missing Data**

Dealing with missing data is important to avoid biased model training and provide reliable predictions. Two main approaches are:

● Removing Missing Data:

○If a feature has more than 50-60% missing values, it may be dropped due to insufficient information.

○Entire rows with missing values can be removed if they form a small percentage of the dataset.

●Imputation Techniques:

○For Numerical Data:

■Replace missing values with the mean (if data follows a normal distribution).

■Use the median if the data is skewed to avoid the effect of outliers.

○For Categorical Data:

■ Replace missing values with the mode (most common category).

**3. Correlation Analysis**

Knowledge of the relationship between numerical features assists in feature selection and preventing multicollinearity problems.

● Pearson's Correlation Coefficient:

○ Values between -1 and +1:

■ +1: Strong positive correlation (as one feature goes up, the other goes up).

■ -1: Strong negative correlation (as one feature goes up, the other goes down).

■ 0: No correlation.

● Heatmap Visualization:

○ Highly correlated features can be visualized using a seaborn heatmap.

○ Features with very high correlation (>0.8) can be dropped or combined to eliminate redundancy.

**4. Encoding Categorical Features**

Numerical input is needed for machine learning models, so categorical data needs to be encoded properly.

● Encoding Techniques:

○ Label Encoding: Maps integer values to categorical labels, suitable for ordinal data (e.g., low < medium < high).

○ One-Hot Encoding (OHE): Maps categorical variables into binary columns, suitable for nominal data (e.g., gender, city names).

**5. Data Visualization**

Visualizing data is important to see patterns, distributions, and possible anomalies.

●Key Plots for EDA:

○Histograms: Display the distribution of numerical variables.

○Boxplots: Detect outliers in numerical data.

○Scatter Plots: Display relationships between two numerical features.

**6. Feature Scaling and Normalization**

Feature scaling provides consistency in numerical features, enhancing model performance.

●Standardization (Z-score Normalization):

○Converts values to zero mean and unit variance.

○Formula: X′=X−μσX' = \frac{X - \mu}{\sigma}

○Applicable for models such as linear regression, logistic regression, and PCA.

● Min-Max Scaling:

○ Scales the values to 0 to 1.

○ Formula: X′=X−XminXmax−XminX' = \frac{X - X\_{min}}{X\_{max} - X\_{min}}

○ Best for models such as KNN and neural networks.

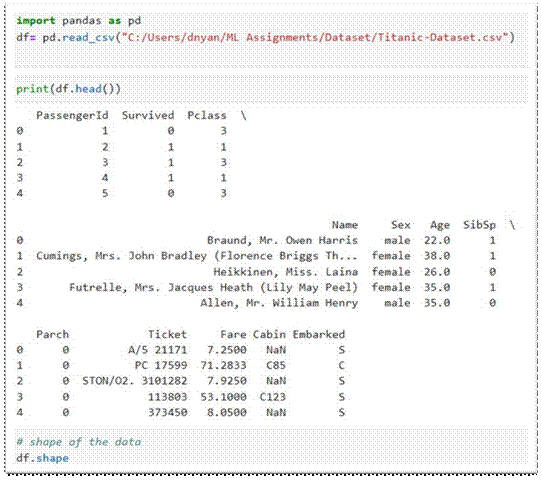
● Robust Scaling:

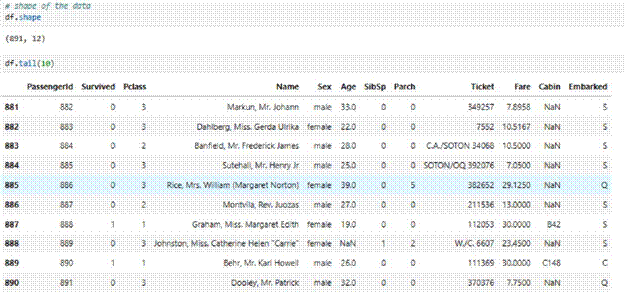
○ Utilizes median and interquartile range (IQR) to manage outliers.

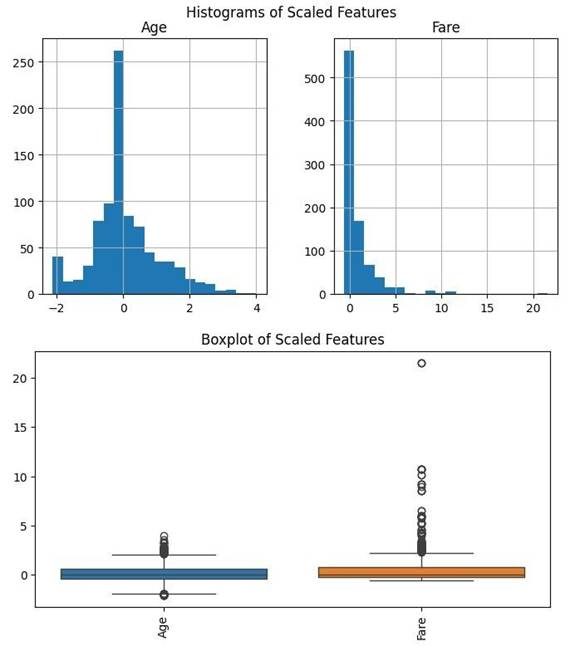
○ Formula: X′=X−MedianIQRX' = \frac{X - \text{Median}}{IQR}

○ Best suited for datasets with extreme values.

Code & Output :

****





**Conclusion**

The exercise of Exploratory Data Analysis (EDA) and preprocessing is critical prior to machine learning model training. Getting a grip on the structure of the dataset, treating missing data values, assessing correlation, coding categorical features, displaying data for visual understanding, and using techniques for feature scaling improves data quality as well as increases the forecasting power of the model.

Doing it this way keeps the dataset organized, cleansed, and primed for further use of complex machine learning processes.