**Assignment No. 1**

**Problem Statement:** Exploring data analysis (Various operations on dataset).

**Objective:** To perform Exploratory Data Analysis (EDA) and Preprocessing on a dataset to understand its structure, detect anomalies, and prepare it for machine learning models. The process includes handling missing data, analyzing correlations, applying encoding techniques, and visualizing data using charts and heatmaps.

**Prerequisite :**

1. A Python environment set up with libraries like pandas, xml.etree.ElementTree, and requests (for web access).
2. Internet connection (for reading datasets from the web).
3. Text editor and basic knowledge of python and EDA

**Theory :**

**Steps for EDA and Preprocessing**

1. **Understanding the Dataset**

Before performing any analysis, it is crucial to explore the dataset structure and its contents. This helps in identifying potential issues and determining the necessary preprocessing steps. The key aspects to check include:

* **Number of Rows and Columns**

1. The dataset size is checked using .shape, which gives the count of rows (samples) and columns (features).
2. A large dataset may need feature selection to avoid overfitting, while a small dataset may require augmentation techniques.

* **Data Types of Columns**

1. Different columns may have numerical (integer/float) or categorical (string/object) values.
2. The .info() function provides an overview of data types, which helps determine if encoding is required.

* **Missing Values**

1. Missing values can cause biases in model predictions.
2. They are detected using .isnull().sum(), which counts the number of missing values per column.

* **Basic Statistical Measures**

1. Measures like mean, median, and standard deviation (.describe()) provide insights into data distribution.
2. Skewness in distributions may indicate the need for transformations such as log scaling.

2. **Handling Missing Data**

Missing values must be addressed to prevent biased model training. There are two main strategies:

* **Removal of Missing Data**

1. If a column has more than **50-60% missing values**, it may be dropped as it lacks sufficient information.
2. Rows with missing values may also be removed, but only if their number is small.

* **Imputation Techniques**

1. **Numerical Data:** Replace missing values with:
   1. **Mean** (if data is normally distributed).
   2. **Median** (if data is skewed).
2. **Categorical Data:** Replace with the **mode** (most frequent category).

**3. Correlation Analysis**

Correlation measures the relationship between numerical features. It helps in identifying redundant features that may lead to **multicollinearity**, negatively impacting model performance.

1. **Pearson’s Correlation Coefficient**

Values range from **-1 to +1**:

* **+1**: Strong positive correlation (as one increases, the other increases).
* **-1**: Strong negative correlation (as one increases, the other decreases).
* **0**: No correlation.

1. **Heatmap Visualization**

A heatmap helps identify highly correlated features, which can be removed or merged.

**4.Encoding Categorical Features**

Since machine learning models only work with numerical data, categorical features must be converted into numerical representations.

* **Encoding Techniques**

1. **Label Encoding:** Assigns an integer to each category. Used for **ordinal** data (e.g., low < medium < high).
2. **One-Hot Encoding (OHE):** Creates binary columns for each category. Suitable for **nominal** data (e.g., gender, cities).

**5.Data Visualization**

* Key Types of Plots for EDA

1. **Histograms:** Show the distribution of numerical variables.
2. **Boxplots:** Identify outliers.
3. **Scatter plots:** Show relationships between two numerical variables.

**6. Feature Scaling and Normalization**

Feature scaling ensures uniformity in numerical features, improving model performance.

* **Standardization (Z-score Normalization)**

1. Transforms values to **zero mean** and **unit variance**.
2. Formula: X′=X−μσX' = \frac{X - \mu}{\sigma}X′=σX−μ​
3. Suitable for models like **linear regression, logistic regression, and PCA**.

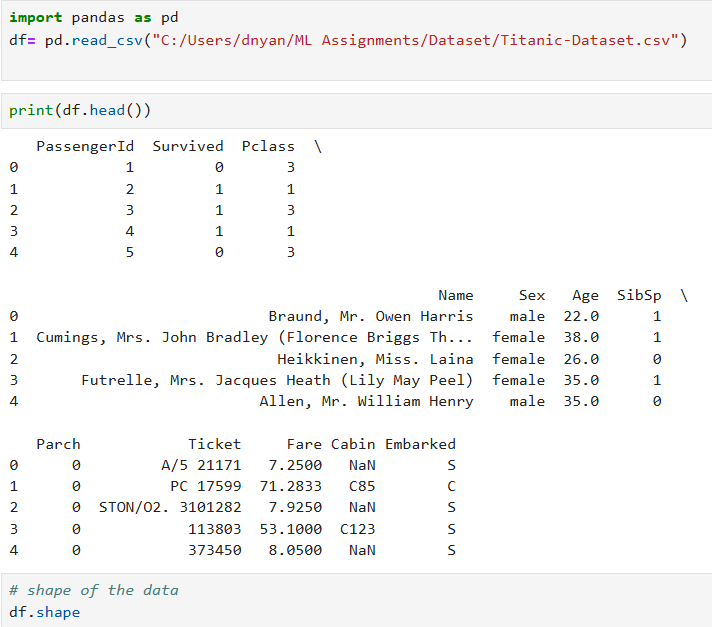
* **Min-Max Scaling**

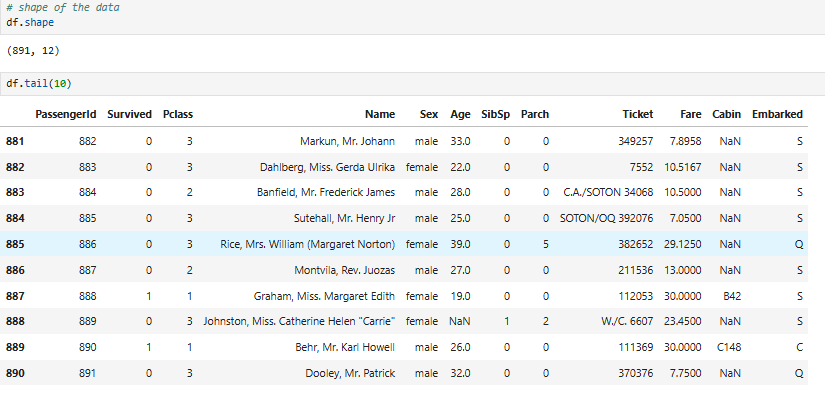
1. Scales values between **0 and 1**.
2. Formula: X′=X−XminXmax−XminX' = \frac{X - X\_{min}}{X\_{max} - X\_{min}}X′=Xmax​−Xmin​X−Xmin​​
3. Used for models like **KNN and neural networks**.

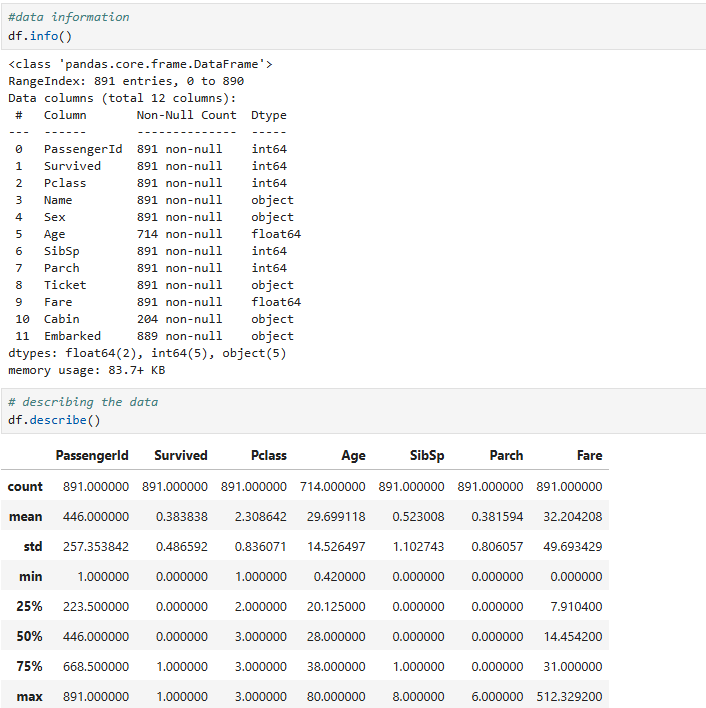
* **Robust Scaling**

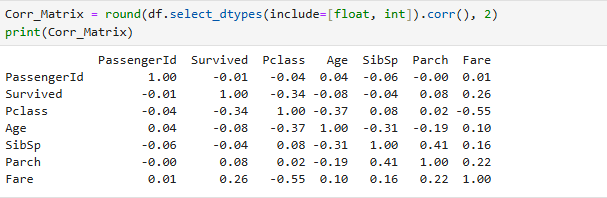
1. Uses **median and IQR** to handle outliers.
2. Formula: X′=X−MedianIQRX' = \frac{X - Median}{IQR}X′=IQRX−Median​
3. Best for **datasets with extreme values**.

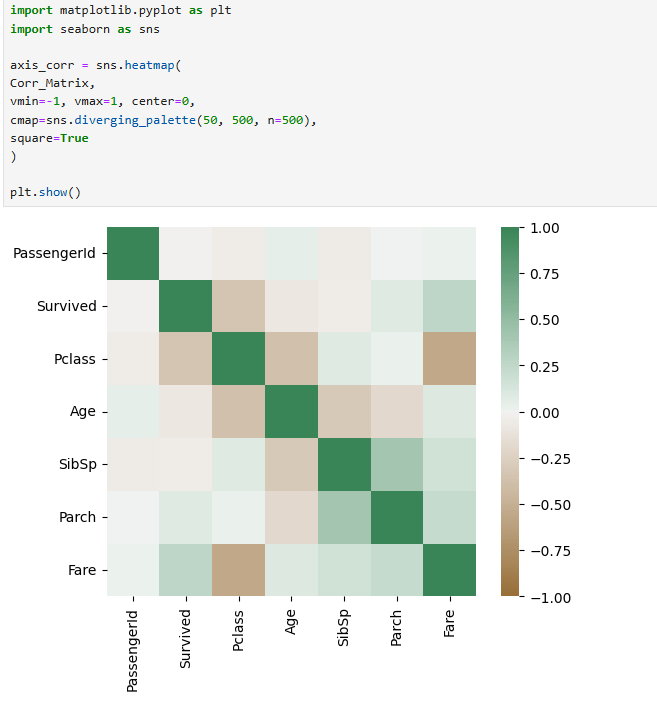
**Code & Output :**

****

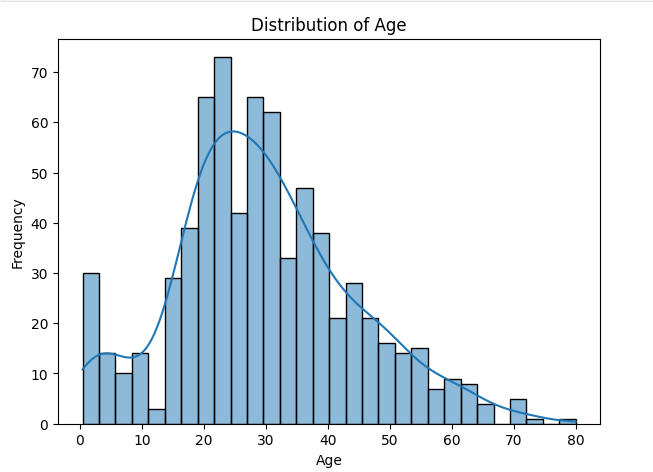
****

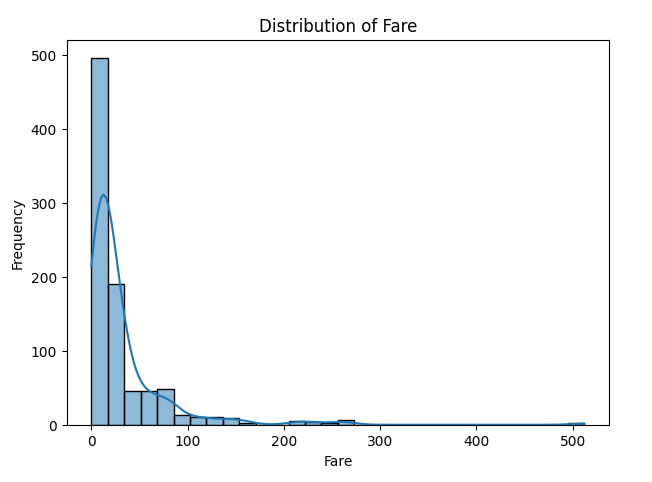
****

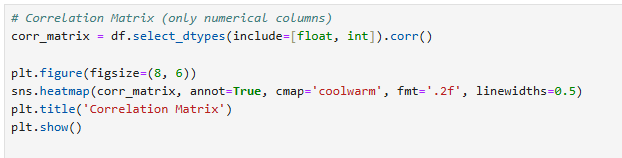
****

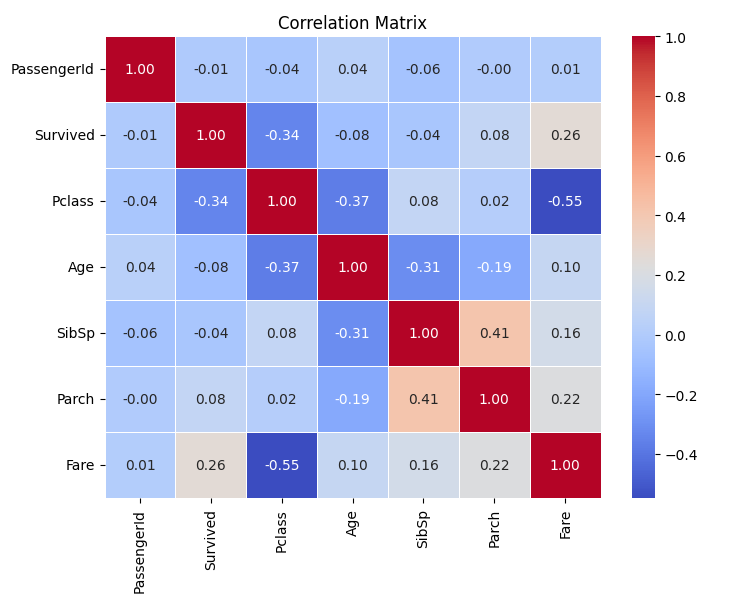
****

****

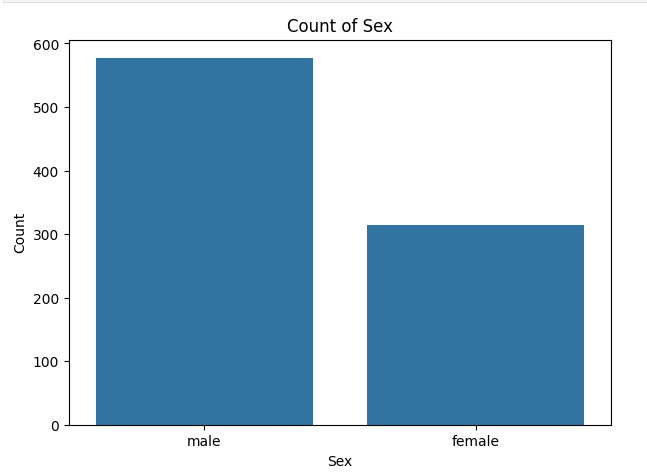
****

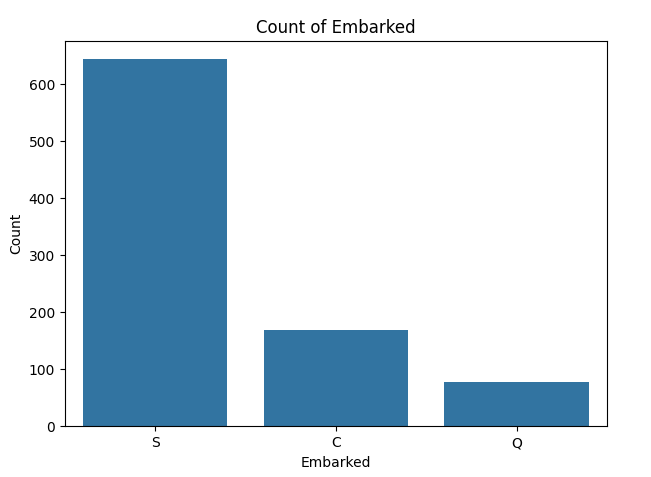
****

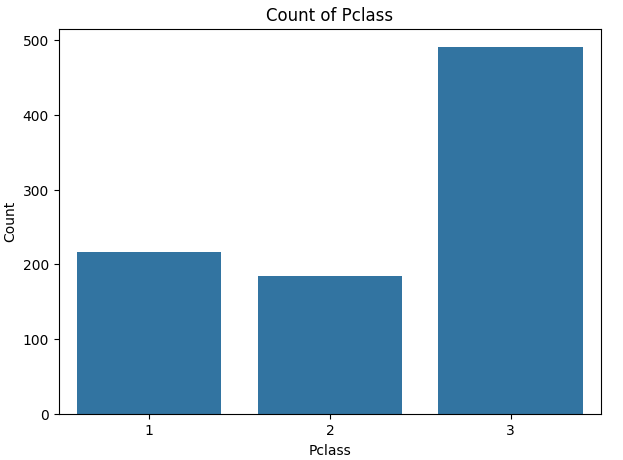
****

****

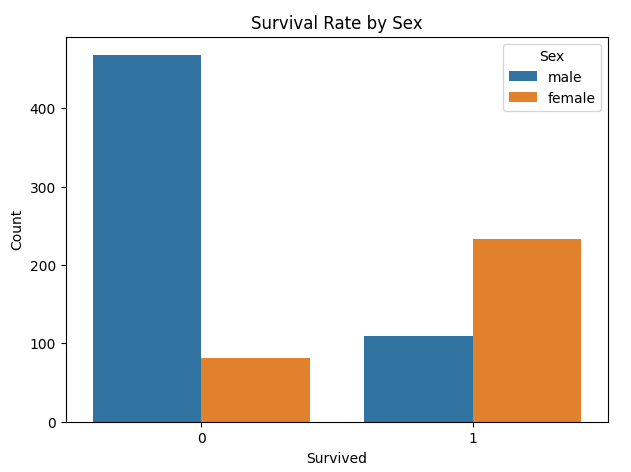
****

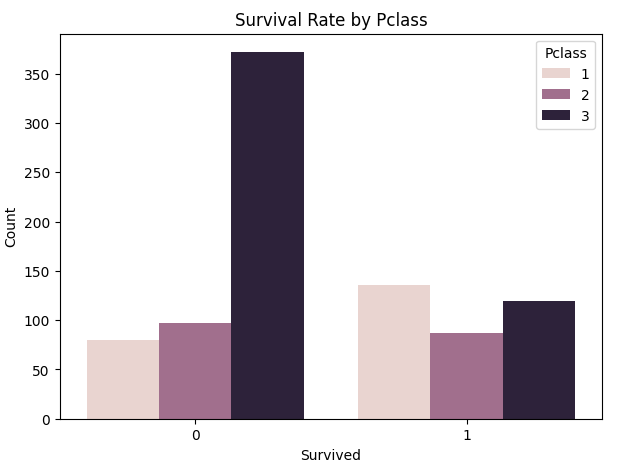
****

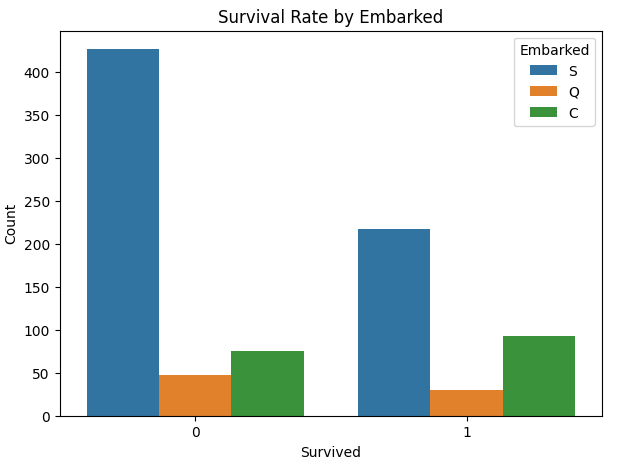
****

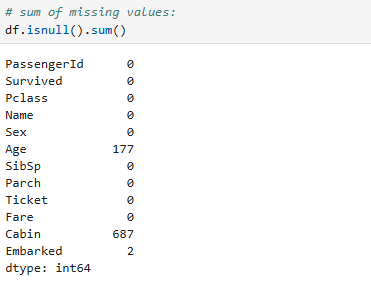
****

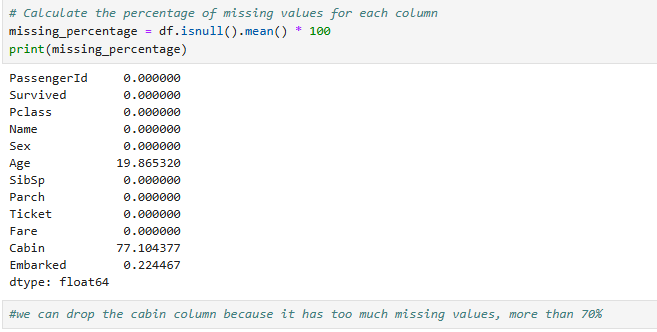
****

****

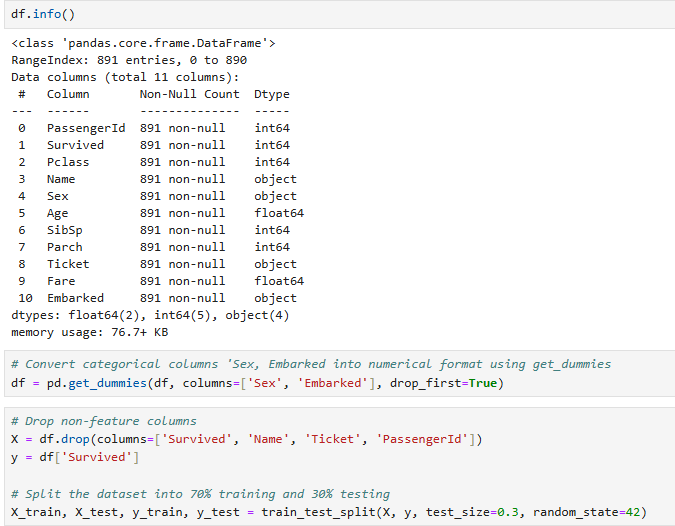
****

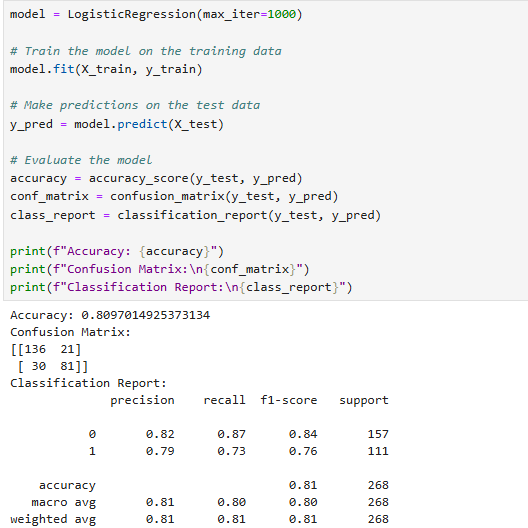
****

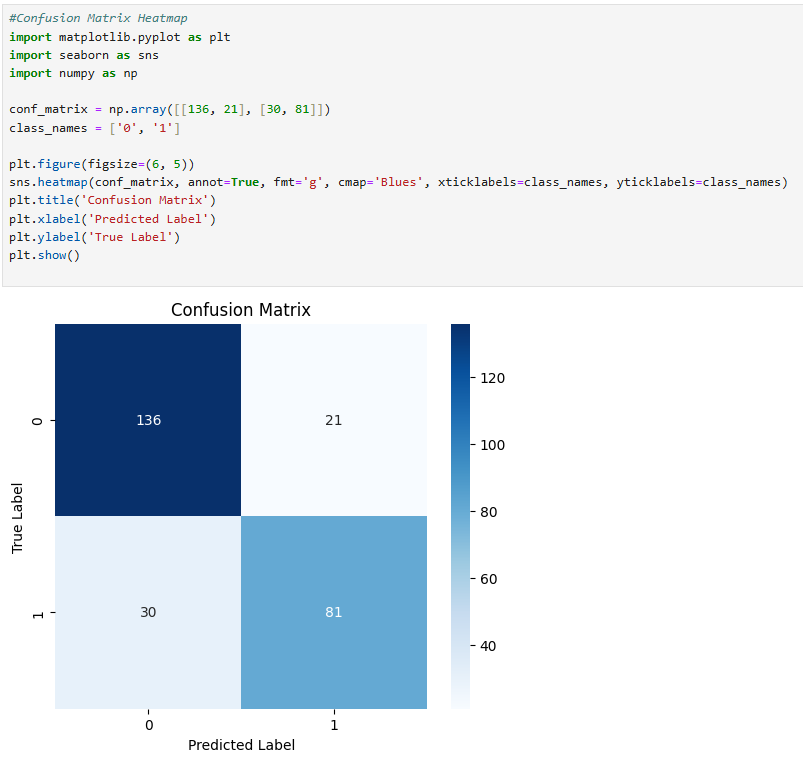
****

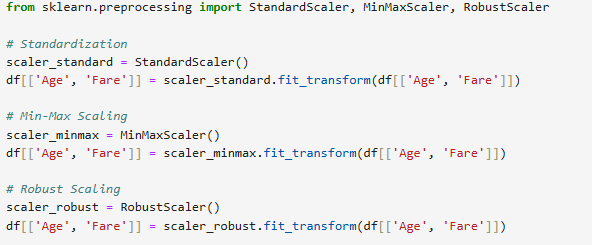
****

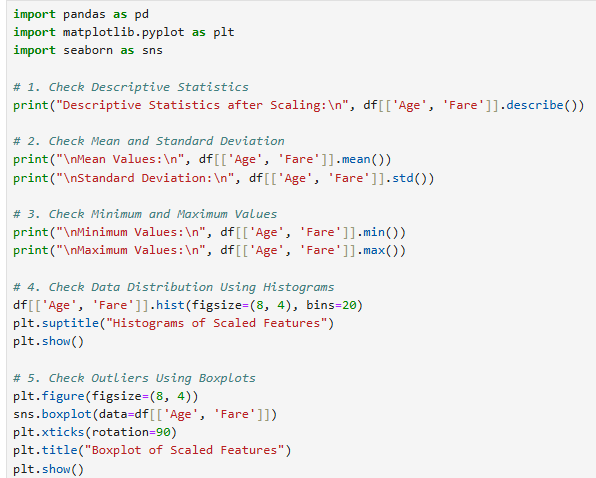
****

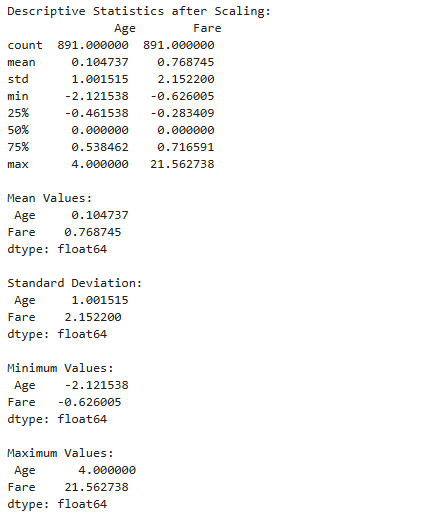
****

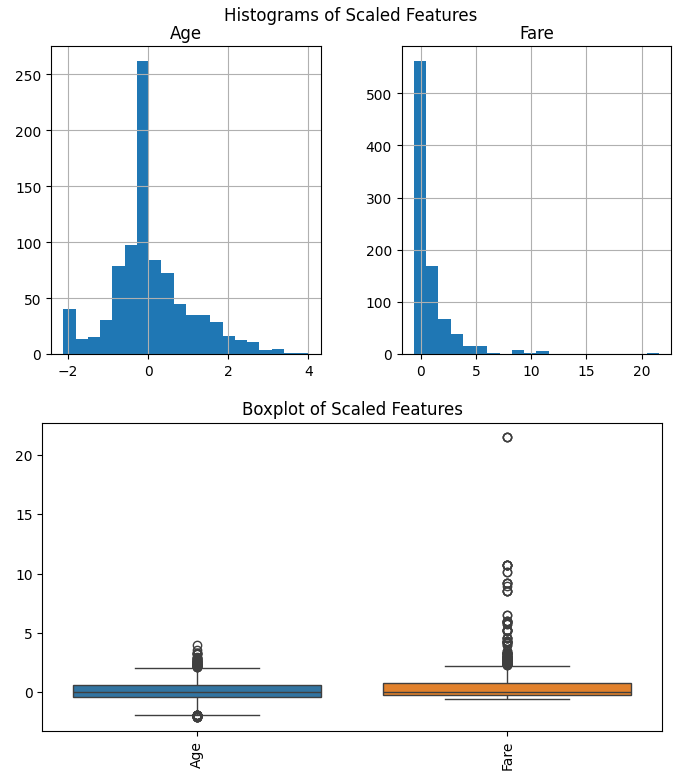
****

****

****

****

****

****

**Github :-** [**https://github.com/dnyaneshwardhere/ML**](https://github.com/dnyaneshwardhere/ML)

**Conclusion:**

This EDA task covered data inspection, missing value handling, categorical feature encoding, correlation analysis, and feature scaling. Missing values were handled, categorical variables were encoded, and duplicate features were detected. Various scaling methods were used, but incorrect overwriting corrupted the results. Proper transformations are necessary for improved model performance. The process generally enhanced data quality and got it machine learning ready.