AI & ML INTERNSHIP

NAME : K R LAHARI

Task 2: Exploratory Data Analysis (EDA)

Objective: Understand data using statistics and visualizations.

Tools: Pandas, Matplotlib, Seaborn, Plotly

CODE:

```
import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt


# Load the dataset

data = pd.read_csv('Titanic-Dataset.csv')


# 1. Generate summary statistics

print("Summary Statistics:")

print(data.describe())

print("\nMedian values:")

print(data.median(numeric_only=True))


# 2. Create histograms and boxplots for numeric features

numeric_cols = data.select_dtypes(include=['int64', 'float64']).columns


for col in numeric_cols:

    # Histogram

    plt.figure(figsize=(5, 3))

    plt.hist(data[col], bins=20, color='lightblue', edgecolor='black')
```

```python
    plt.title(f'Histogram of {col}')

    plt.xlabel(col)

    plt.ylabel('Frequency')

    plt.tight_layout()

    plt.show()


    # Boxplot

    plt.figure(figsize=(5, 2))

    sns.boxplot(x=data[col], color='lightgreen')

    plt.title(f'Boxplot of {col}')

    plt.tight_layout()

    plt.show()


# 3. Pairplot and correlation matrix

# Pairplot

sns.pairplot(data[numeric_cols])

plt.suptitle("Pairplot of Numeric Features", y=1.02)

plt.show()


# Correlation matrix

plt.figure(figsize=(10, 6))

sns.heatmap(data[numeric_cols].corr(), annot=True, cmap='coolwarm', fmt=".2f")

plt.title("Correlation Matrix")

plt.show()
```
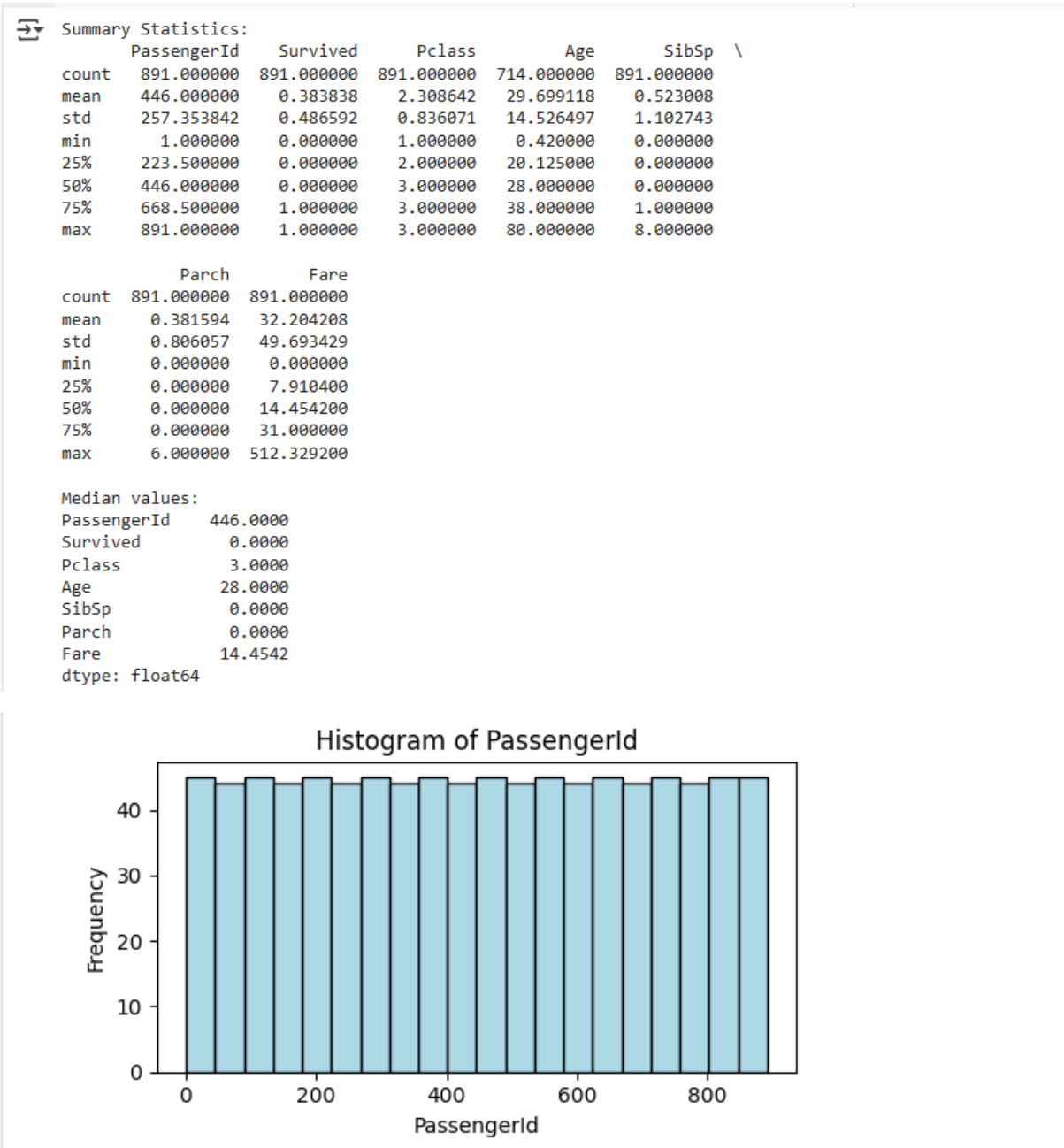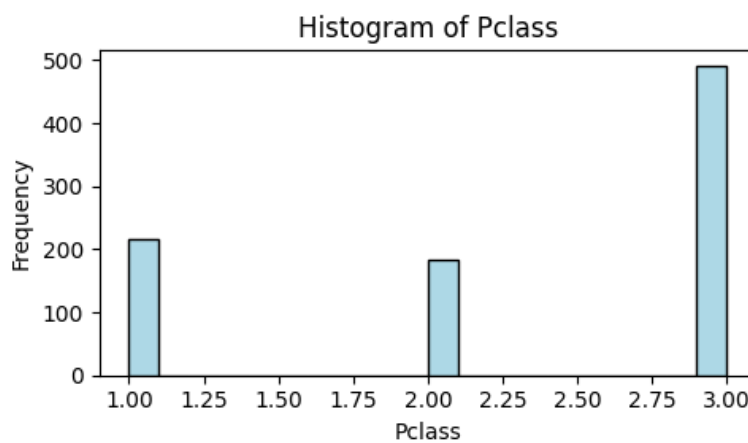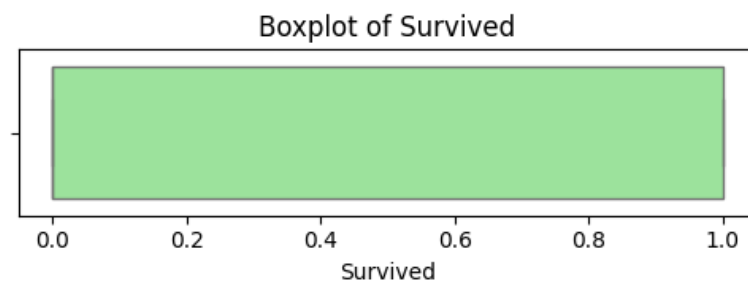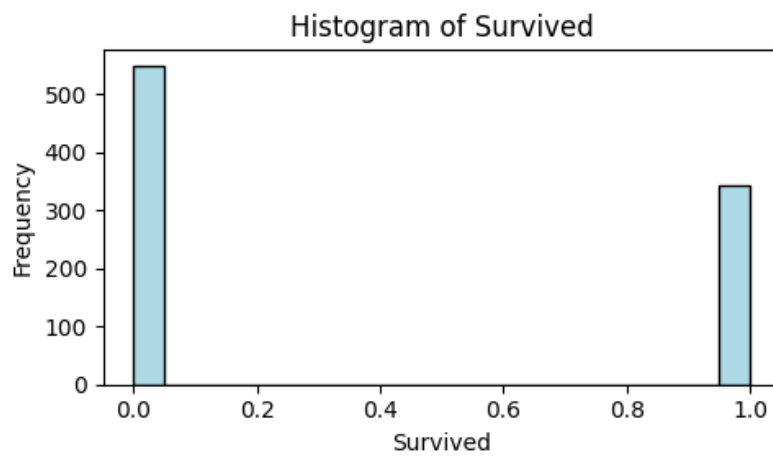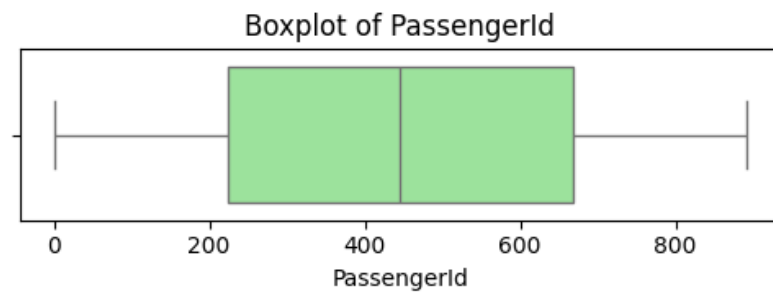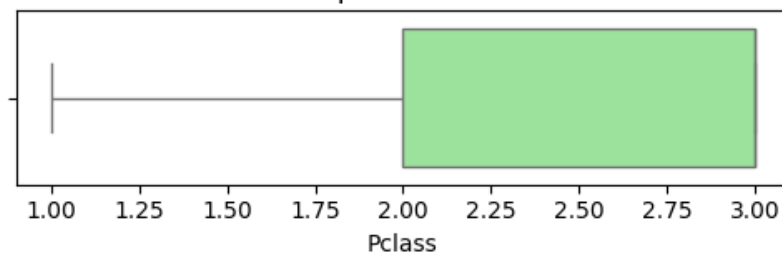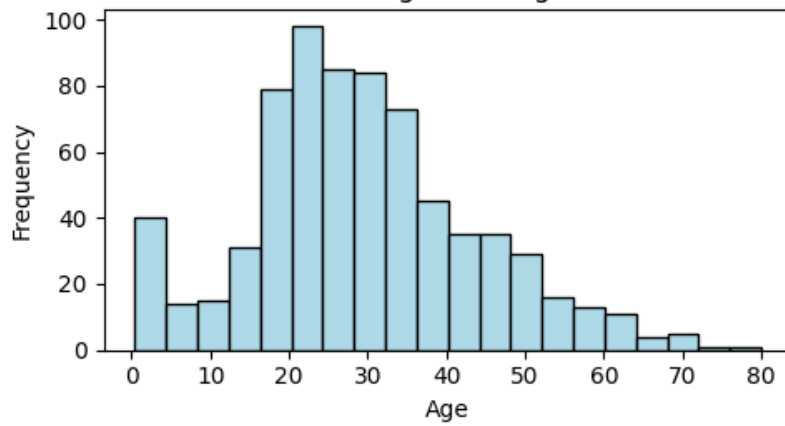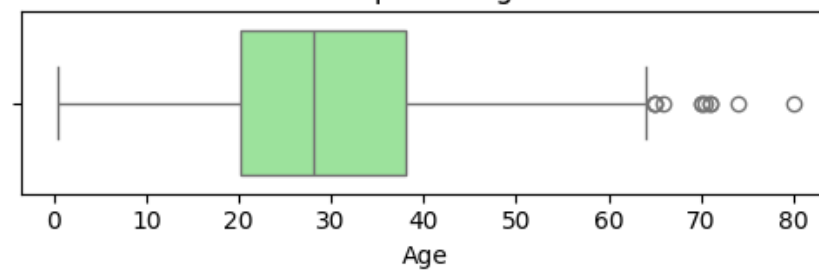
OUTPUT:

```
Summary Statistics:
       PassengerId    Survived      Pclass         Age       SibSp  \
count   891.000000  891.000000  891.000000  714.000000  891.000000
mean    446.000000    0.383838    2.308642   29.699118    0.523008
std     257.353842    0.486592    0.836071   14.526497    1.102743
min       1.000000    0.000000    1.000000    0.420000    0.000000
25%     223.500000    0.000000    2.000000   20.125000    0.000000
50%     446.000000    0.000000    3.000000   28.000000    0.000000
75%     668.500000    1.000000    3.000000   38.000000    1.000000
max     891.000000    1.000000    3.000000   80.000000    8.000000

            Parch        Fare
count  891.000000  891.000000
mean     0.381594   32.204208
std      0.806057   49.693429
min      0.000000    0.000000
25%      0.000000    7.910400
50%      0.000000   14.454200
75%      0.000000   31.000000
max      6.000000  512.329200

Median values:
PassengerId    446.0000
Survived         0.0000
Pclass           3.0000
Age             28.0000
SibSp            0.0000
Parch            0.0000
Fare            14.4542
dtype: float64
```
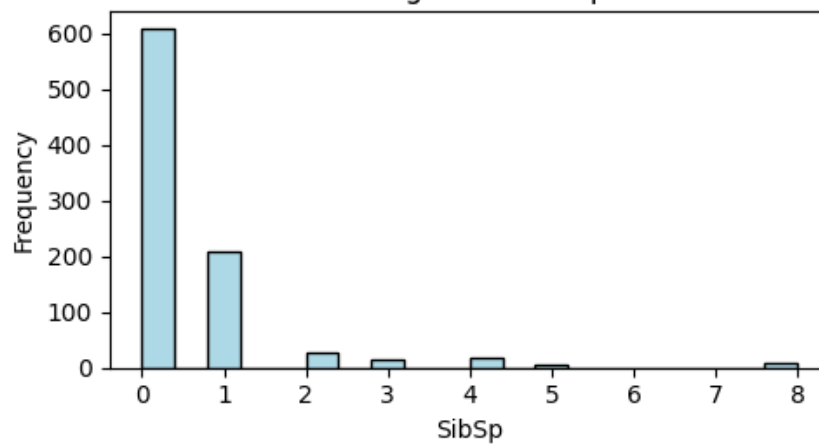

Histogram of PassengerId

Boxplot of PassengerId

Histogram of Survived

Boxplot of Survived

Histogram of Pclass

Boxplot of Pclass

Pclass

Histogram of Age

Age

Boxplot of Age

Age

Histogram of SibSp

SibSp

Boxplot of SibSp

SibSp
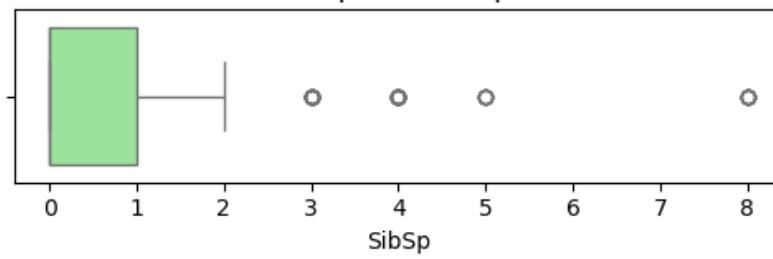
Histogram of Parch

Frequency
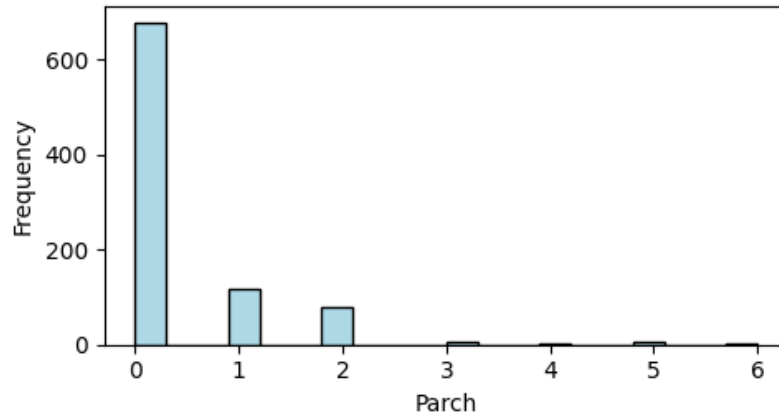
Parch

Boxplot of Parch

Parch

Histogram of Fare

Frequency
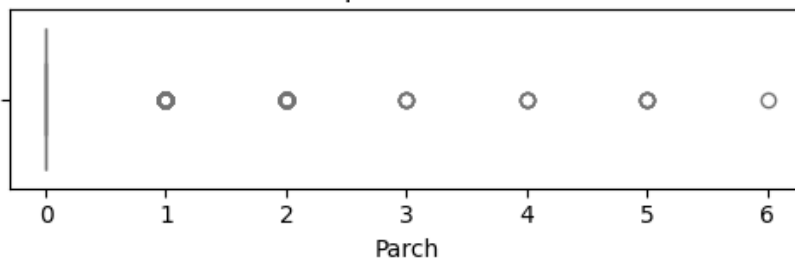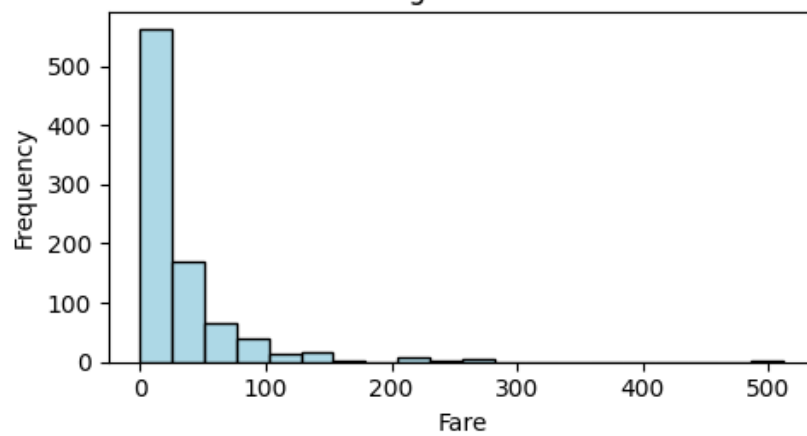
Fare

Boxplot of Fare

Pairplot of Numeric Features

Correlation Matrix