```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
# Load the data
train_df = pd.read_csv('/content/train.csv')
test_df = pd.read_csv('/content/test.csv')
# Combine the training and testing data
df = pd.concat([train_df, test_df])
# Data Cleaning
print(df.isnull().sum())
```

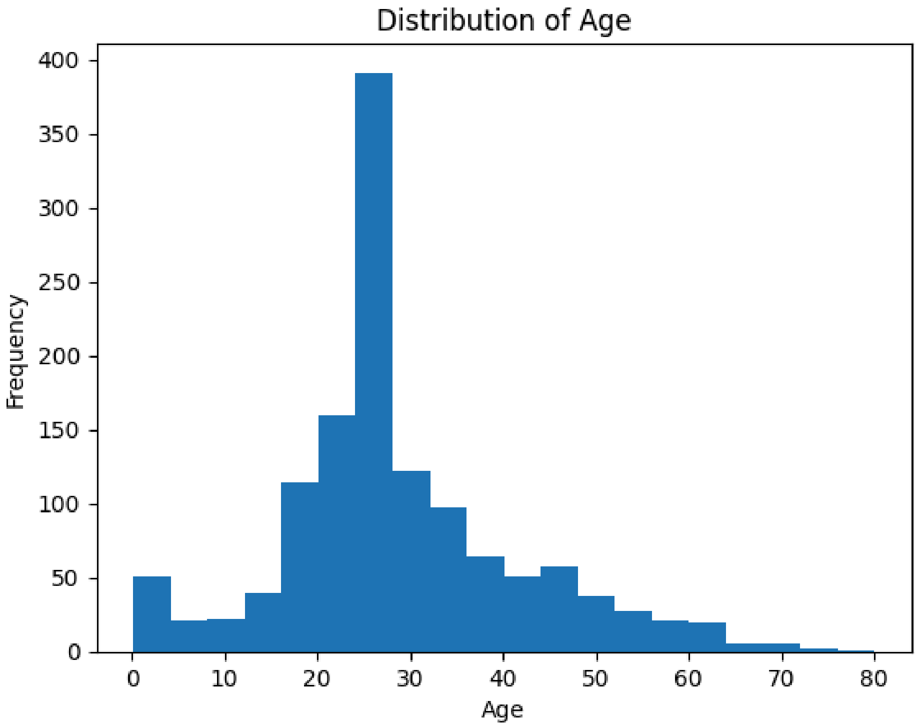
```
PassengerId       0
Survived        418
Pclass            0
Name              0
Sex               0
Age             263
SibSp             0
Parch             0
Ticket            0
Fare              1
Cabin          1014
Embarked          2
dtype: int64
```

```
# Fill missing values in Age with median age
df['Age'].fillna(df['Age'].median(), inplace=True)

# Fill missing values in Embarked with most frequent value
df['Embarked'].fillna(df['Embarked'].mode()[0], inplace=True)

# Convert categorical variables to numerical variables
df['Sex'] = df['Sex'].map({'male': 0, 'female': 1})
df['Embarked'] = df['Embarked'].map({'S': 0, 'C': 1, 'Q': 2})

# Exploratory Data Analysis (EDA)
# Summary statistics
print(df.describe())
```

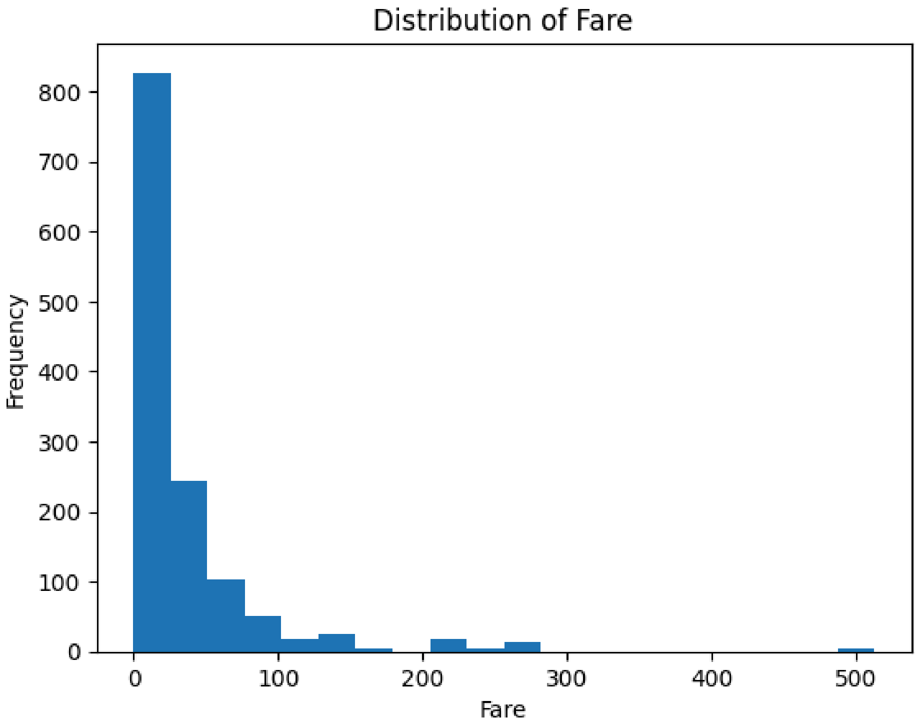
```
       PassengerId    Survived       Pclass          Sex          Age  \
count  1309.000000  891.000000  1309.000000  1309.000000  1309.000000
mean    655.000000    0.383838     2.294882     0.355997    29.503186
std     378.020061    0.486592     0.837836     0.478997    12.905241
min       1.000000    0.000000     1.000000     0.000000     0.170000
```

```
25%     328.000000    0.000000     2.000000     0.000000    22.000000
50%     655.000000    0.000000     3.000000     0.000000    28.000000
75%     982.000000    1.000000     3.000000     1.000000    35.000000
max    1309.000000    1.000000     3.000000     1.000000    80.000000

             SibSp        Parch         Fare     Embarked
count  1309.000000  1309.000000  1308.000000  1309.000000
mean      0.498854     0.385027    33.295479     0.394194
std       1.041658     0.865560    51.758668     0.653499
min       0.000000     0.000000     0.000000     0.000000
25%       0.000000     0.000000     7.895800     0.000000
50%       0.000000     0.000000    14.454200     0.000000
75%       1.000000     0.000000    31.275000     1.000000
max       8.000000     9.000000   512.329200     2.000000
```

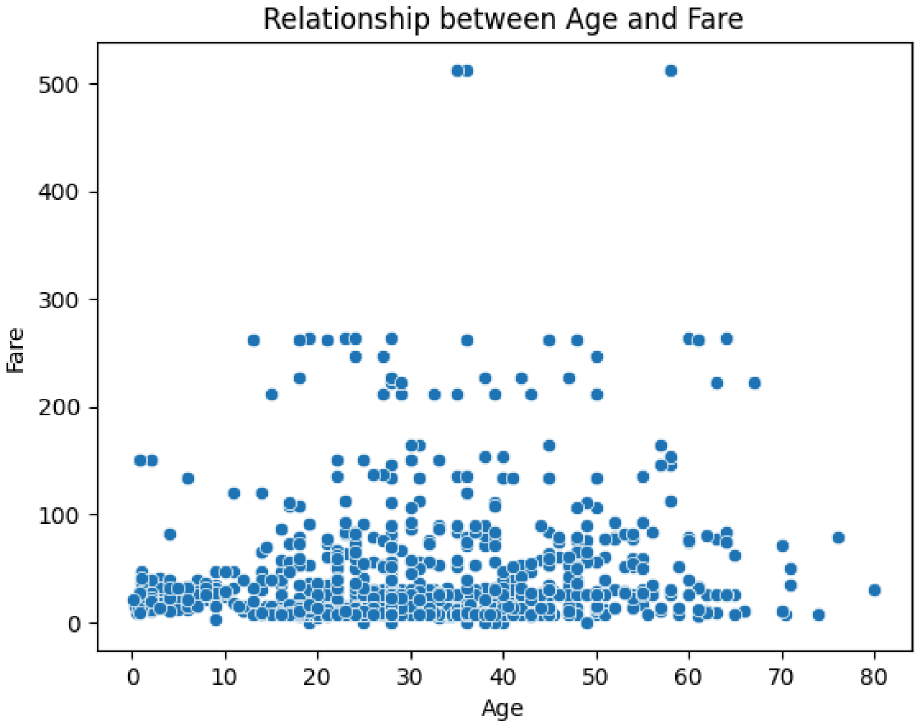
```
# Distribution of Age
plt.hist(df['Age'], bins=20)
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.title('Distribution of Age')
plt.show()
```



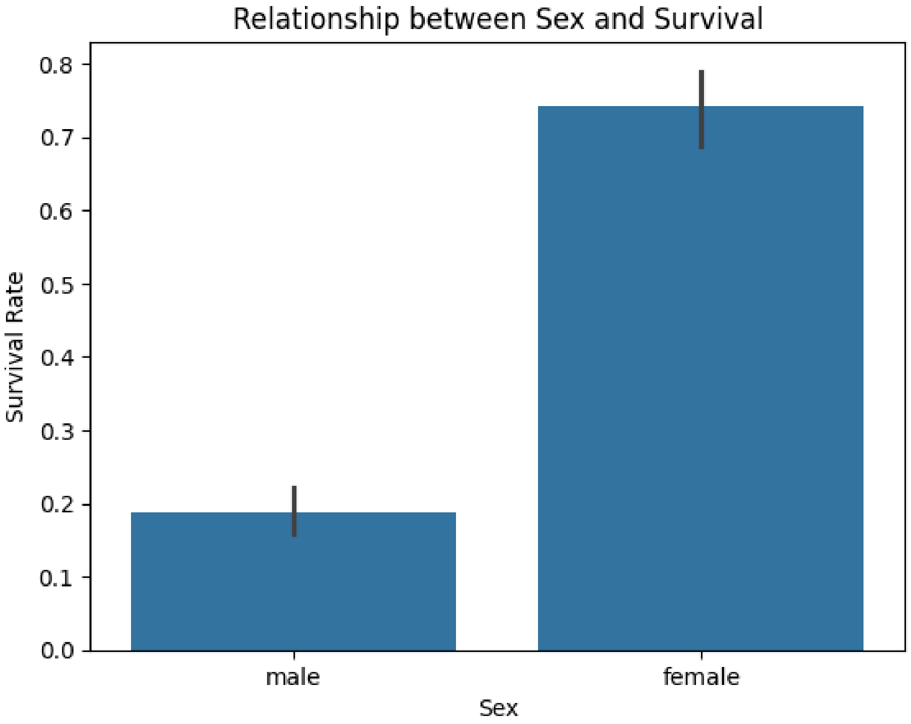
```
# Distribution of Fare
plt.hist(df['Fare'], bins=20)
plt.xlabel('Fare')
plt.ylabel('Frequency')
plt.title('Distribution of Fare')
plt.show()
```



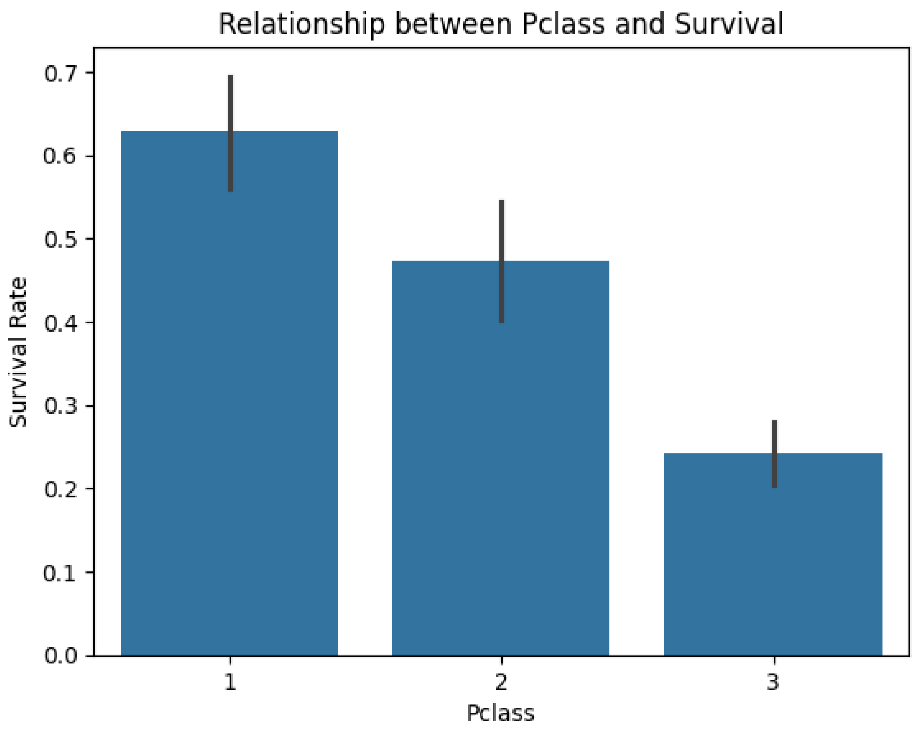
```
# Relationship between Age and Fare
df_reset = df.reset_index(drop=True)
sns.scatterplot(x='Age',y='Fare', data=df_reset)
plt.xlabel('Age')
plt.ylabel('Fare')
plt.title('Relationship between Age and Fare')
plt.show()
```



```
# Relationship between Sex and Survival
sns.barplot(x='Sex', y='Survived', data=train_df)
plt.xlabel('Sex')
plt.ylabel('Survival Rate')
plt.title('Relationship between Sex and Survival')
plt.show()
```

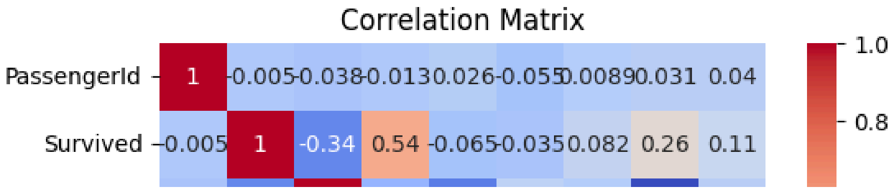


```
# Relationship between Pclass and Survival
sns.barplot(x='Pclass', y='Survived', data=train_df)
plt.xlabel('Pclass')
plt.ylabel('Survival Rate')
plt.title('Relationship between Pclass and Survival')
plt.show()
```



```
# Correlation matrix
numerical_df = df.select_dtypes(include=['number'])
corr_matrix = numerical_df.corr()
print(corr_matrix)

# Heatmap of correlation matrix
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', square=True)
plt.title('Correlation Matrix')
plt.show()
```

```
             PassengerId  Survived    Pclass       Sex       Age     SibSp  \
PassengerId     1.000000 -0.005007 -0.038354 -0.013406  0.025799 -0.055224   
Survived       -0.005007  1.000000 -0.338481  0.543351 -0.064910 -0.035322   
Pclass         -0.038354 -0.338481  1.000000 -0.124617 -0.377908  0.060832   
Sex            -0.013406  0.543351 -0.124617  1.000000 -0.053663  0.109609   
Age             0.025799 -0.064910 -0.377908 -0.053663  1.000000 -0.189972   
SibSp          -0.055224 -0.035322  0.060832  0.109609 -0.189972  1.000000   
Parch           0.008942  0.081629  0.018322  0.213125 -0.125851  0.373587   
Fare            0.031428  0.257307 -0.558629  0.185523  0.179256  0.160238   
Embarked        0.040143  0.106811  0.038875  0.120423  0.018654 -0.073461   

                Parch      Fare  Embarked  
PassengerId  0.008942  0.031428  0.040143  
Survived     0.081629  0.257307  0.106811  
Pclass       0.018322 -0.558629  0.038875  
Sex          0.213125  0.185523  0.120423  
Age         -0.125851  0.179256  0.018654  
SibSp        0.373587  0.160238 -0.073461  
Parch        1.000000  0.221539 -0.095523  
Fare         0.221539  1.000000  0.061126  
Embarked    -0.095523  0.061126  1.000000  
```


Correlation Matrix
PassengerId
Survived