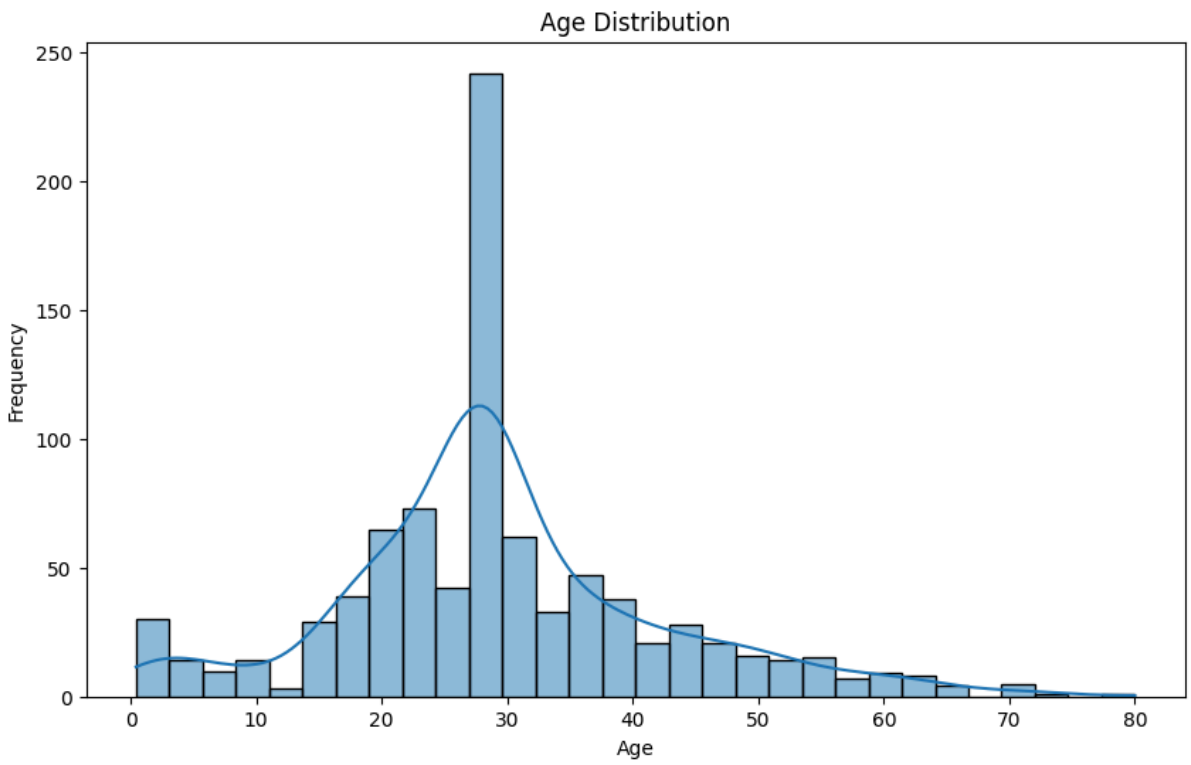
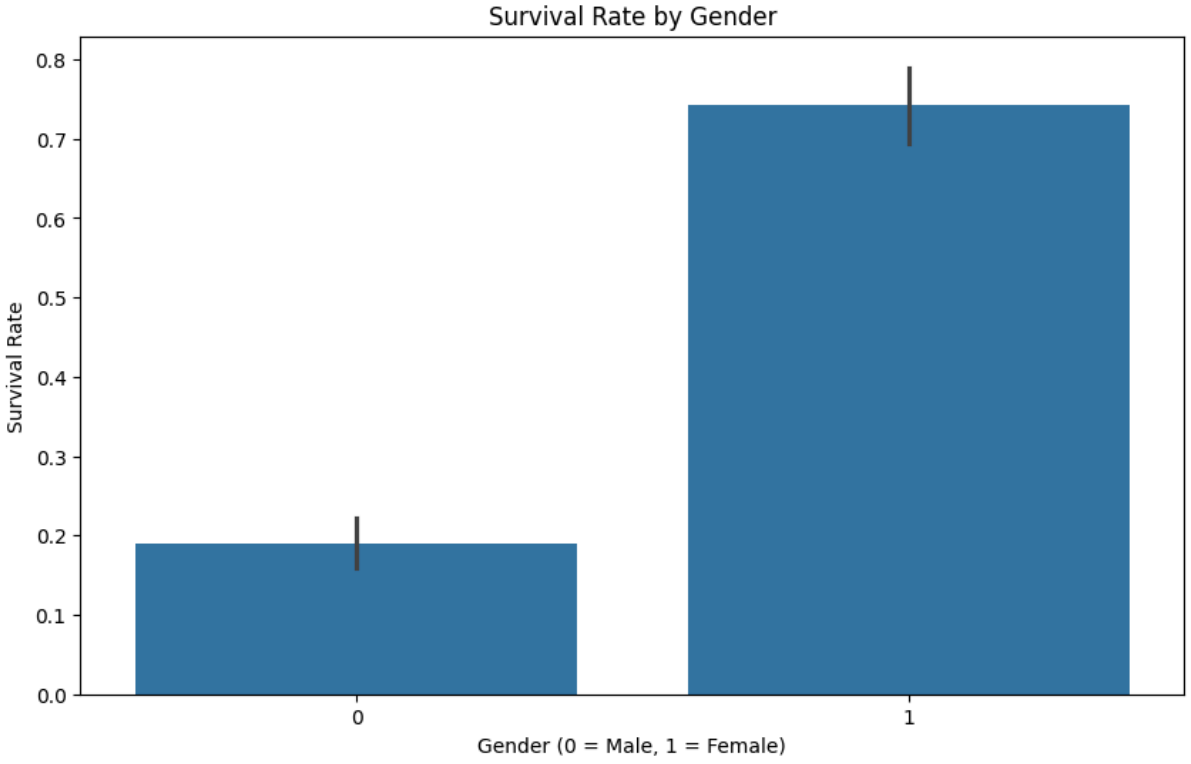
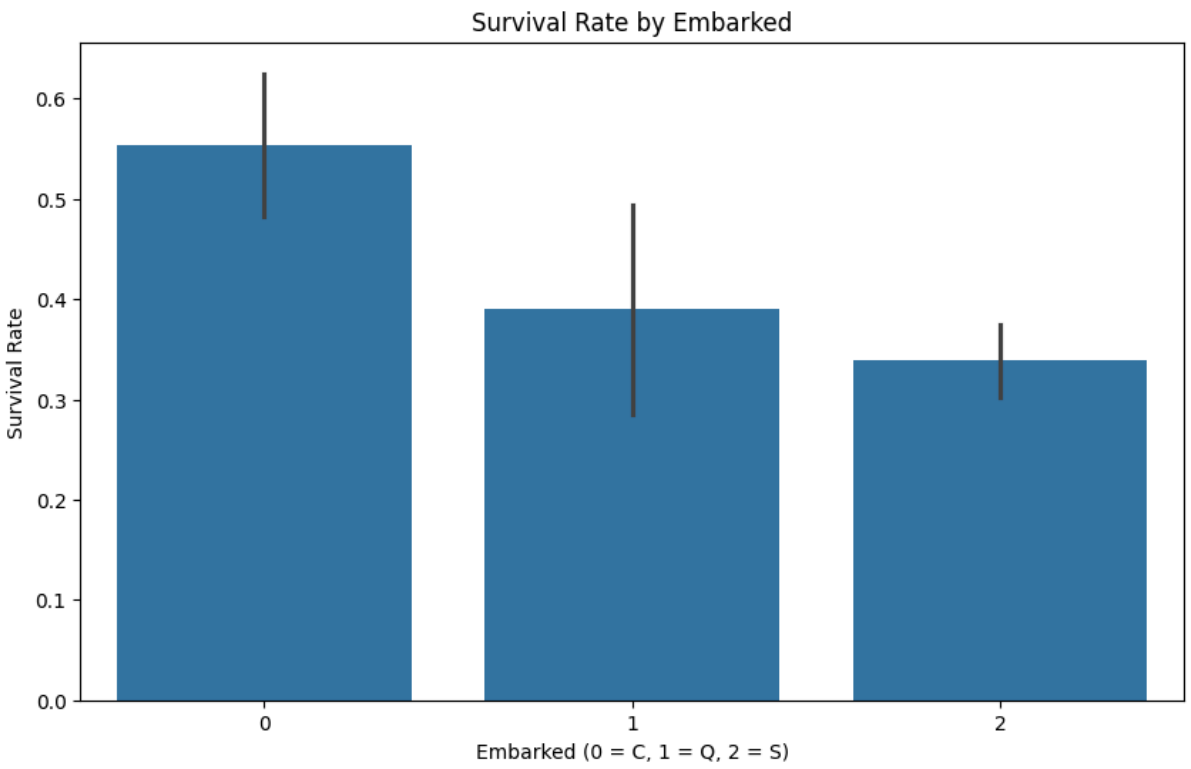
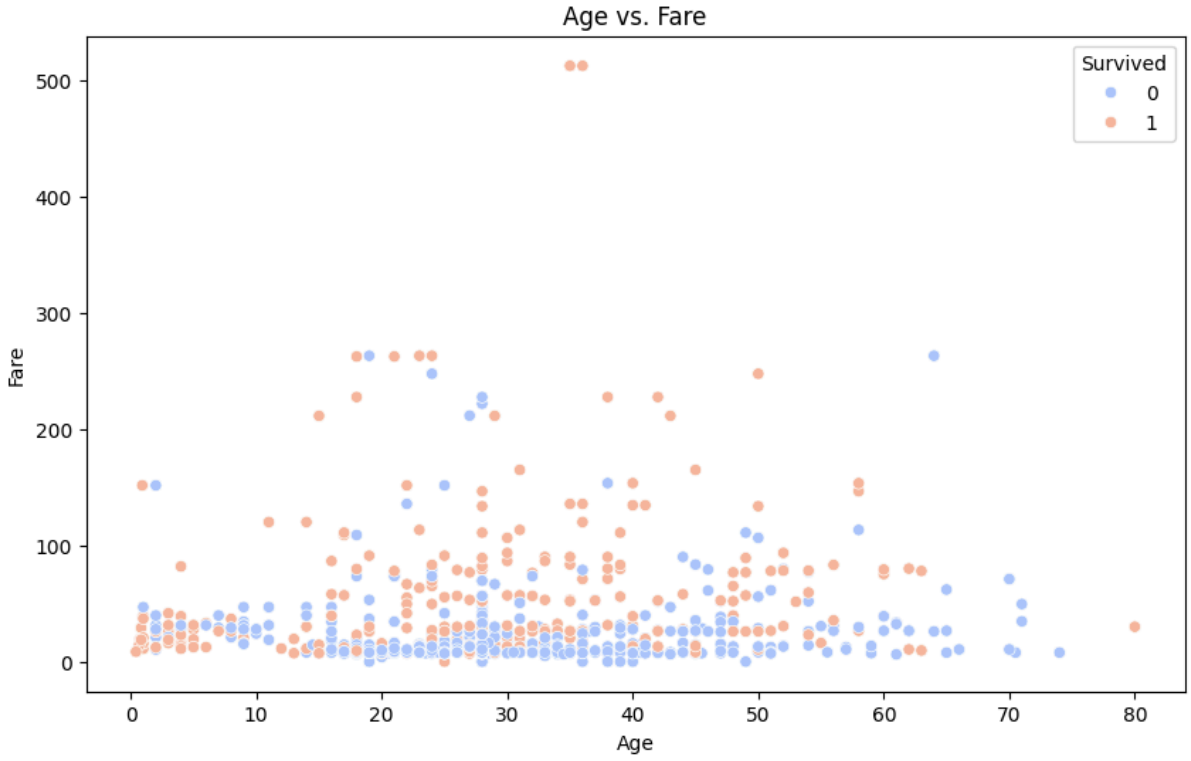
```
# Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Load the Titanic dataset
url = 'https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv'
df = pd.read_csv(url)

# Display the first few rows of the dataset
print("First few rows of the dataset:")
print(df.head())

# Data Cleaning

# Check for missing values
print("\nMissing values in each column:")
print(df.isnull().sum())

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

# Fill missing Embarked with 'S' (the most frequent value)
df['Embarked'].fillna('S', inplace=True)

# Drop columns that won't be used or have too many missing values
df.drop(columns=['Cabin', 'Ticket'], inplace=True)

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

# Exploratory Data Analysis (EDA)

# Descriptive statistics
print("\nDescriptive statistics:")
print(df.describe())

# Distribution of Age
plt.figure(figsize=(10, 6))
sns.histplot(df['Age'], bins=30, kde=True)
plt.title('Age Distribution')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
```

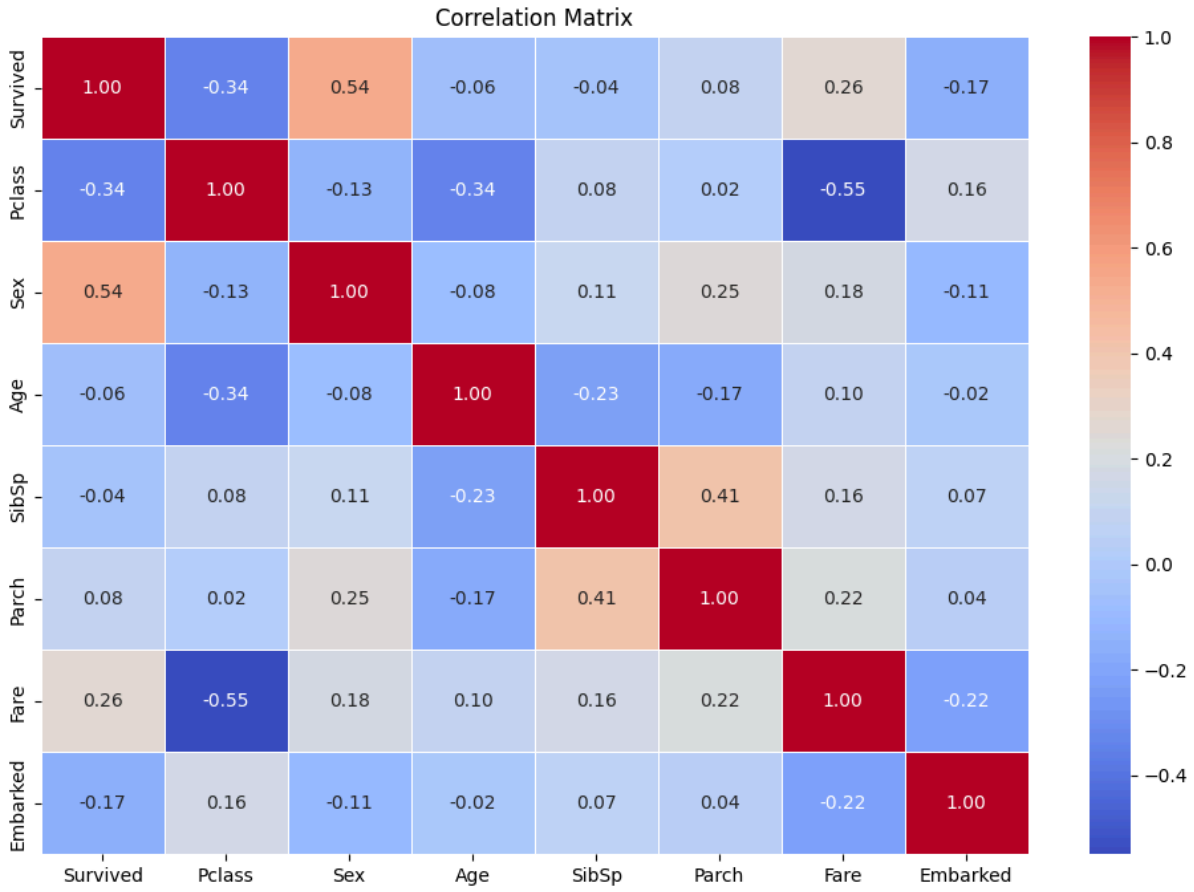
```
# Survival Rate by Gender
plt.figure(figsize=(10, 6))
sns.barplot(x='Sex', y='Survived', data=df)
plt.title('Survival Rate by Gender')
plt.xlabel('Gender (0 = Male, 1 = Female)')
plt.ylabel('Survival Rate')
plt.show()

# Survival Rate by Embarked
plt.figure(figsize=(10, 6))
sns.barplot(x='Embarked', y='Survived', data=df)
plt.title('Survival Rate by Embarked')
plt.xlabel('Embarked (0 = C, 1 = Q, 2 = S)')
plt.ylabel('Survival Rate')
plt.show()

# Age vs. Fare
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Age', y='Fare', hue='Survived', data=df, palette='coolwarm')
plt.title('Age vs. Fare')
plt.xlabel('Age')
plt.ylabel('Fare')
plt.show()

# Correlation matrix
plt.figure(figsize=(12, 8))
correlation_matrix =  df.drop(columns=['Name','PassengerId']).corr()
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f',linewidths=0.5)
plt.title('Correlation Matrix')
plt.show()

# Analyze Survival by Pclass
plt.figure(figsize=(10, 6))
sns.barplot(x='Pclass', y='Survived', data=df)
plt.title('Survival Rate by Pclass')
plt.xlabel('Pclass')
plt.ylabel('Survival Rate')
plt.show()
```

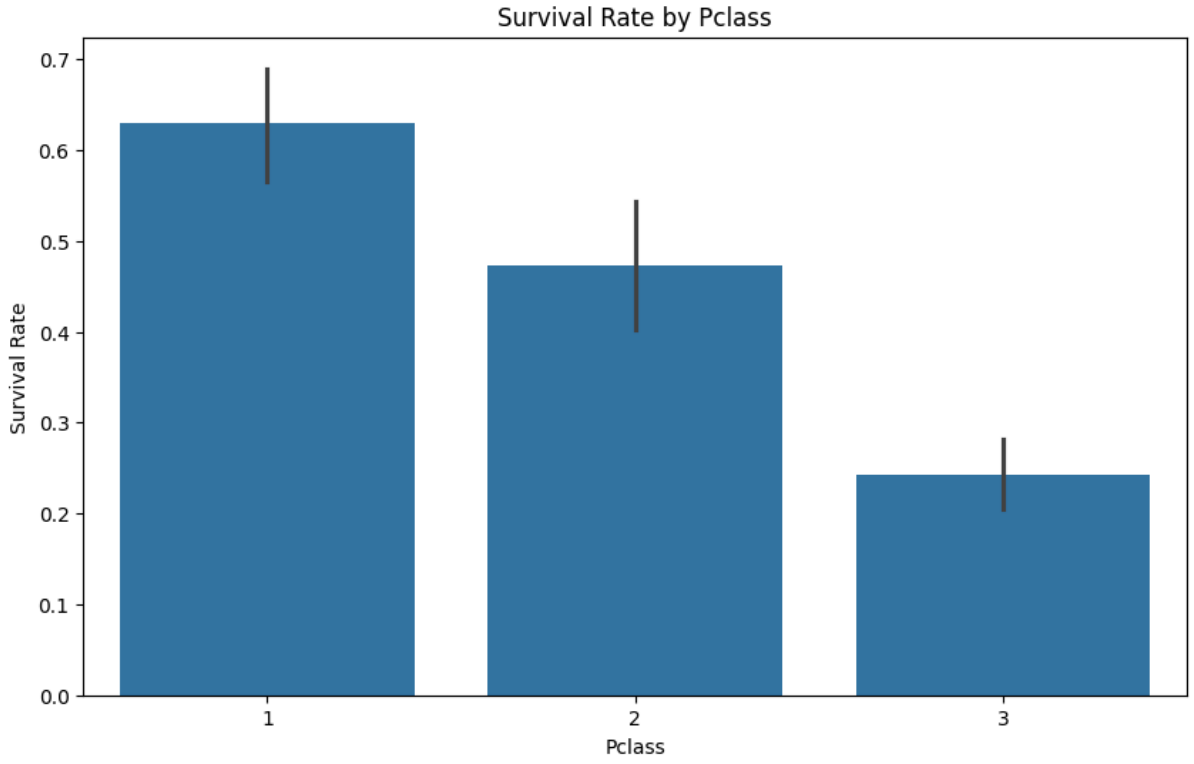

Age Distribution
Frequency
Age
0
50
100
150
200
250
0
10
20
30
40
50
60
70
80


Survival Rate by Gender
Survival Rate
Gender (0 = Male, 1 = Female)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
1


Survival Rate by Embarked
Survival Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
1
2
Embarked (0 = C, 1 = Q, 2 = S)


Age vs. Fare
Survived
0
1
Fare
0
100
200
300
400
500
0
10
20
30
40
50
60
70
80
Age


Correlation Matrix
Survived
Pclass
Sex
Age
SibSp
Parch
Fare
Embarked
1.00 -0.34 0.54 -0.06 -0.04 0.08 0.26 -0.17
-0.34 1.00 -0.13 -0.34 0.08 0.02 -0.55 0.16
0.54 -0.13 1.00 -0.08 0.11 0.25 0.18 -0.11
-0.06 -0.34 -0.08 1.00 -0.23 -0.17 0.10 -0.02
-0.04 0.08 0.11 -0.23 1.00 0.41 0.16 0.07
0.08 0.02 0.25 -0.17 0.41 1.00 0.22 0.04
0.26 -0.55 0.18 0.10 0.16 0.22 1.00 -0.22
-0.17 0.16 -0.11 -0.02 0.07 0.04 -0.22 1.00
Survived
Pclass
Sex
Age
SibSp
Parch
Fare
Embarked
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


Survival Rate by Pclass
Survival Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
1
2
3
Pclass