Task 2) Perform data cleaning and exploratory data analysis (EDA) on a dataset of your choice, such as the Titanic dataset from Kaggle. Explore the relationships between variables and identify patterns and trends in the data.

Ans)

SOURCE CODE:

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

# Load the Titanic dataset from an Excel file

file\_path = r'C:\Users\ANIRBAN CHOWDHURY\Downloads\TITANIC.csv'

titanic\_df = pd.read\_csv(file\_path)

print(titanic\_df.head())

print(titanic\_df.info())

print(titanic\_df.describe(include='all'))

# Data Cleaning

# Check for missing values

print(titanic\_df.isnull().sum())

titanic\_df['Age'].fillna(titanic\_df['Age'].median(), inplace=True)

# Fill missing 'Embarked' values with the most common port

titanic\_df['Embarked'].fillna(titanic\_df['Embarked'].mode()[0], inplace=True)

# Drop the 'Cabin' column as it has too many missing values

titanic\_df.drop(columns=['Cabin'], inplace=True)

# Handling Incorrect Data Types

titanic\_df['Sex'] = titanic\_df['Sex'].astype('category')

titanic\_df['Embarked'] = titanic\_df['Embarked'].astype('category')

titanic\_df['Survived'] = titanic\_df['Survived'].astype('category')

titanic\_df['Pclass'] = titanic\_df['Pclass'].astype('category')

# Exploratory Data Analysis (EDA)

#univariate

# Distribution of 'Age'

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

sns.histplot(titanic\_df['Age'], kde=True)

plt.title('Age Distribution')

plt.xlabel('Age')

plt.ylabel('Frequency')

plt.grid(True)

plt.show()

# Distribution of 'Sex'

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

sns.countplot(data=titanic\_df, x='Sex')

plt.title('Sex Distribution')

plt.xlabel('Sex')

plt.ylabel('Count')

plt.grid(True)

plt.show()

# Distribution of 'Survived'

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

sns.countplot(data=titanic\_df, x='Survived')

plt.title('Survival Distribution')

plt.xlabel('Survived')

plt.ylabel('Count')

plt.grid(True)

plt.show()

# Bivariate

# Survival rate by sex

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

sns.countplot(data=titanic\_df, x='Sex', hue='Survived')

plt.title('Survival Rate by Sex')

plt.xlabel('Sex')

plt.ylabel('Count')

plt.grid(True)

plt.show()

# Survival rate by passenger class

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

sns.countplot(data=titanic\_df, x='Pclass', hue='Survived')

plt.title('Survival Rate by Passenger Class')

plt.xlabel('Passenger Class')

plt.ylabel('Count')

plt.grid(True)

plt.show()

# Age distribution by survival status

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

sns.histplot(data=titanic\_df, x='Age', hue='Survived', kde=True, multiple='stack')

plt.title('Age Distribution by Survival Status')

plt.xlabel('Age')

plt.ylabel('Frequency')

plt.grid(True)

plt.show()

# Pair plot to explore relationships between numerical variables

sns.pairplot(titanic\_df, hue='Survived', diag\_kind='kde', markers=["o", "s"])

plt.show()

# Multivariate

# Heatmap of correlation matrix

plt.figure(figsize=(12, 8))

correlation\_matrix = titanic\_df.corr(numeric\_only=True)

sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm', linewidths=0.5)

plt.title('Correlation Matrix')

plt.show()

# Survival rate by sex and passenger class

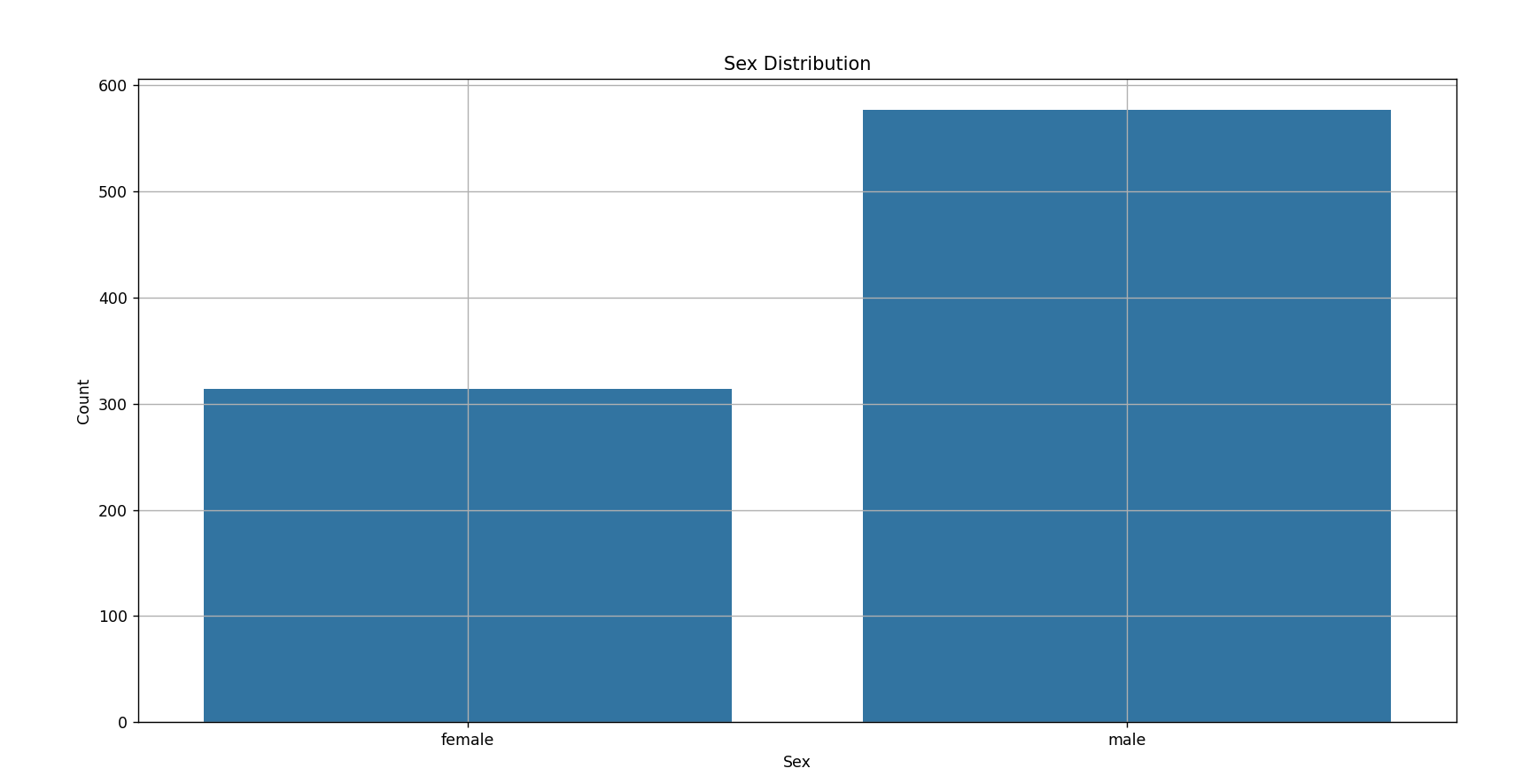
plt.figure(figsize=(12, 8))

sns.catplot(data=titanic\_df, x='Pclass', hue='Sex', col='Survived', kind='count', height=5, aspect=1)

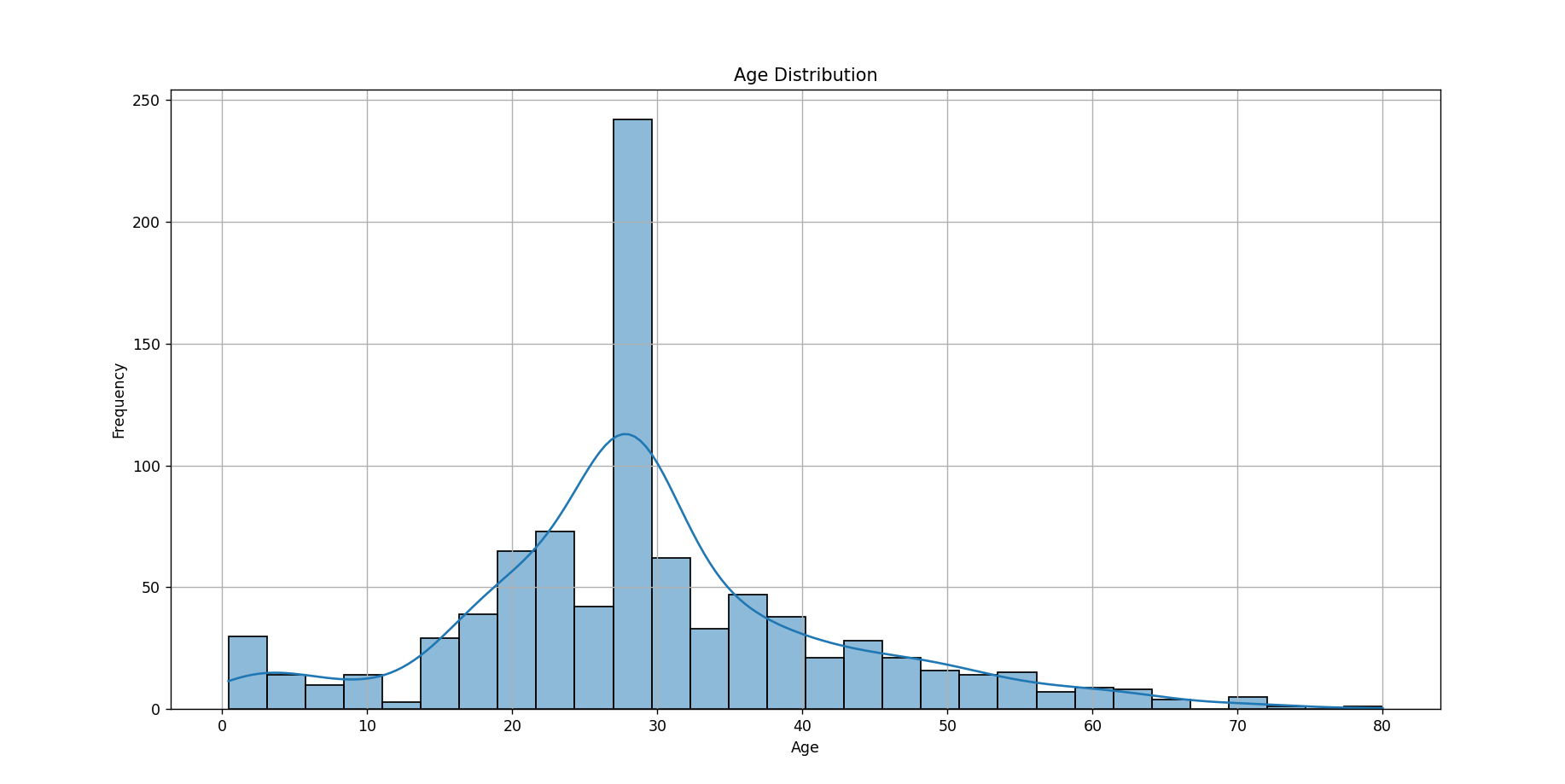
plt.show()

**UNIVARIATE DATA ANALYSIS**:

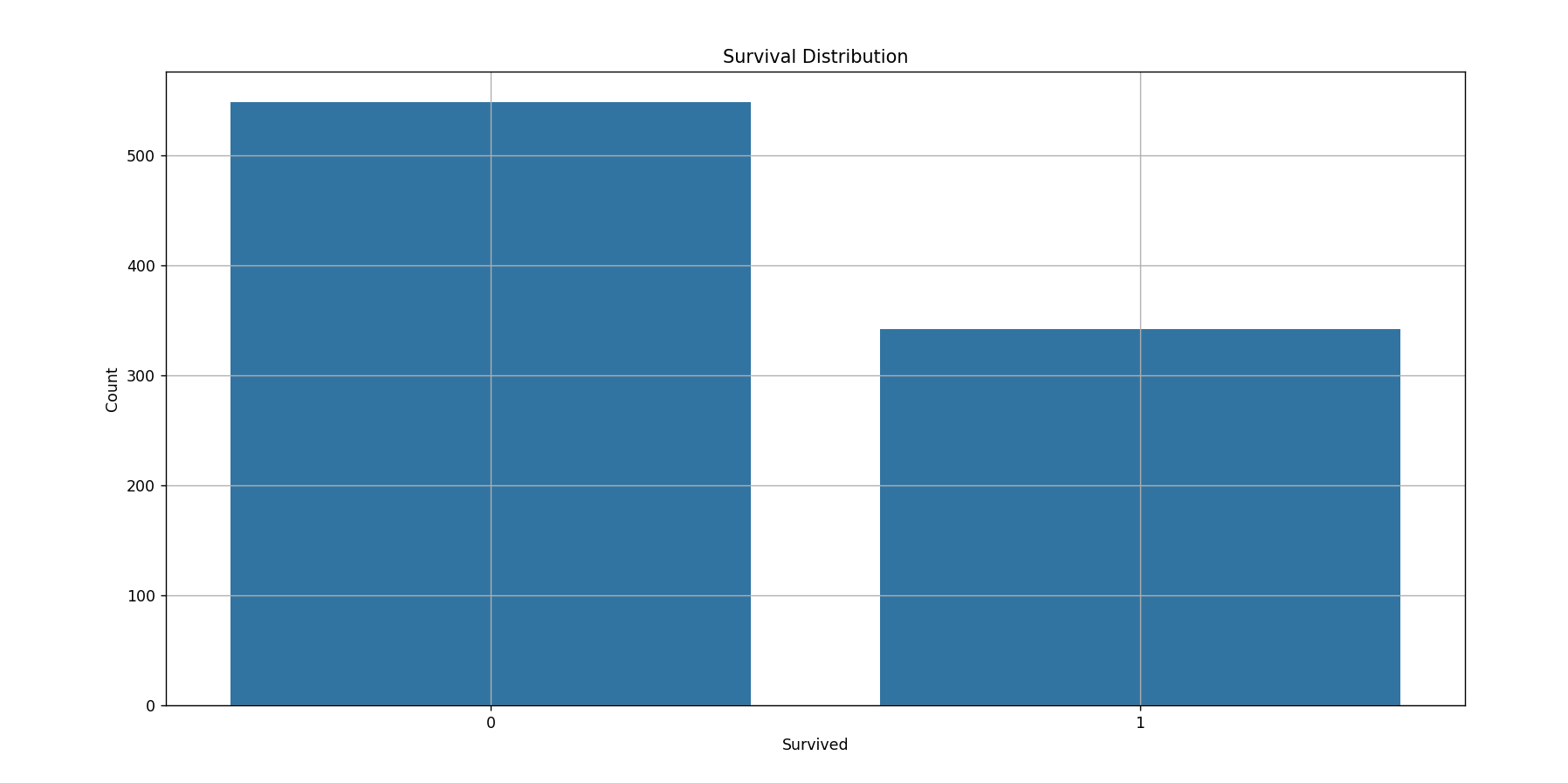
**DISTRIBUTION OF SEX**:



**DISTRIBUTION OF AGE**:

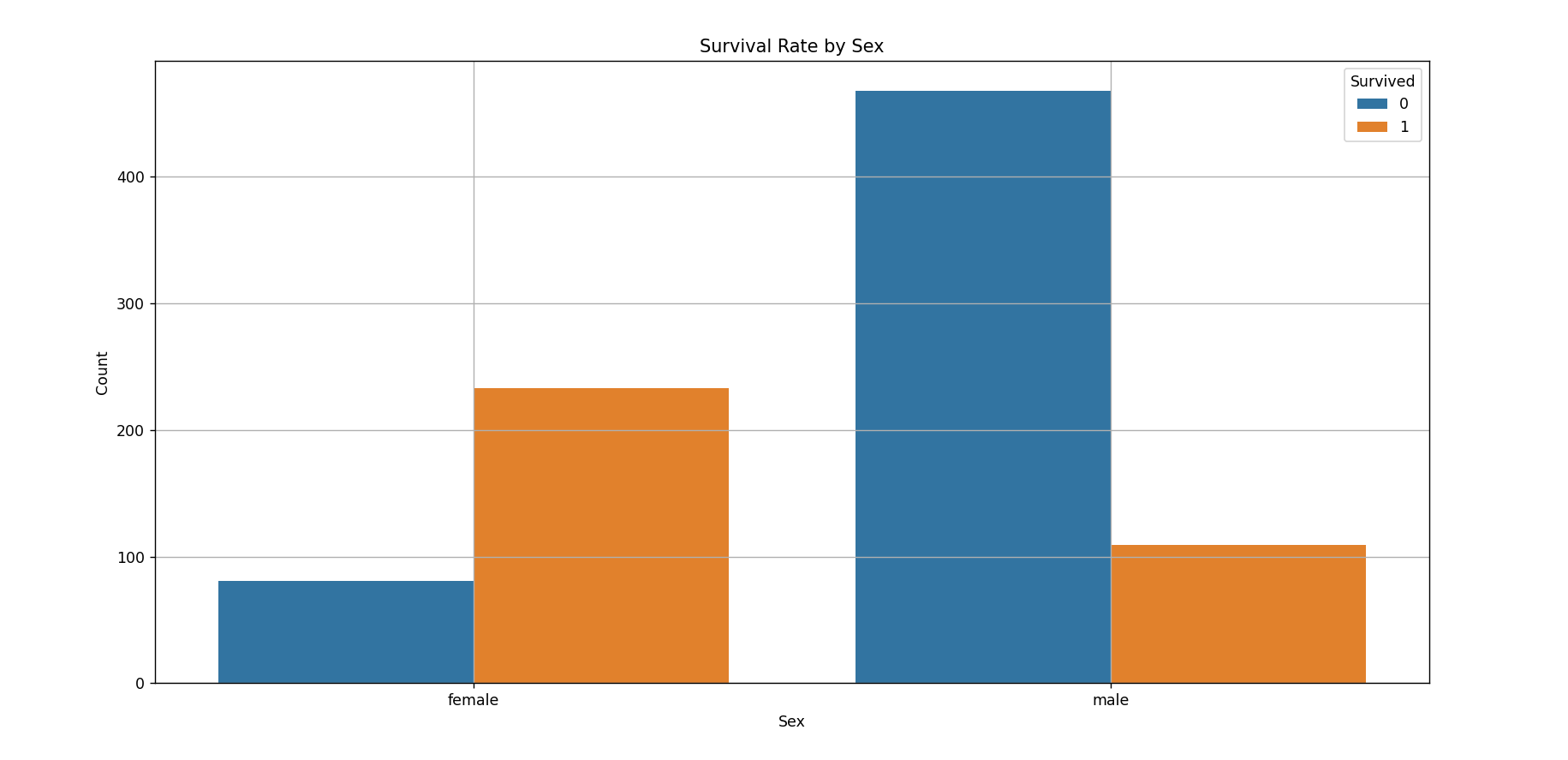


**DISTRIBUTION OF SURVIVAL**:

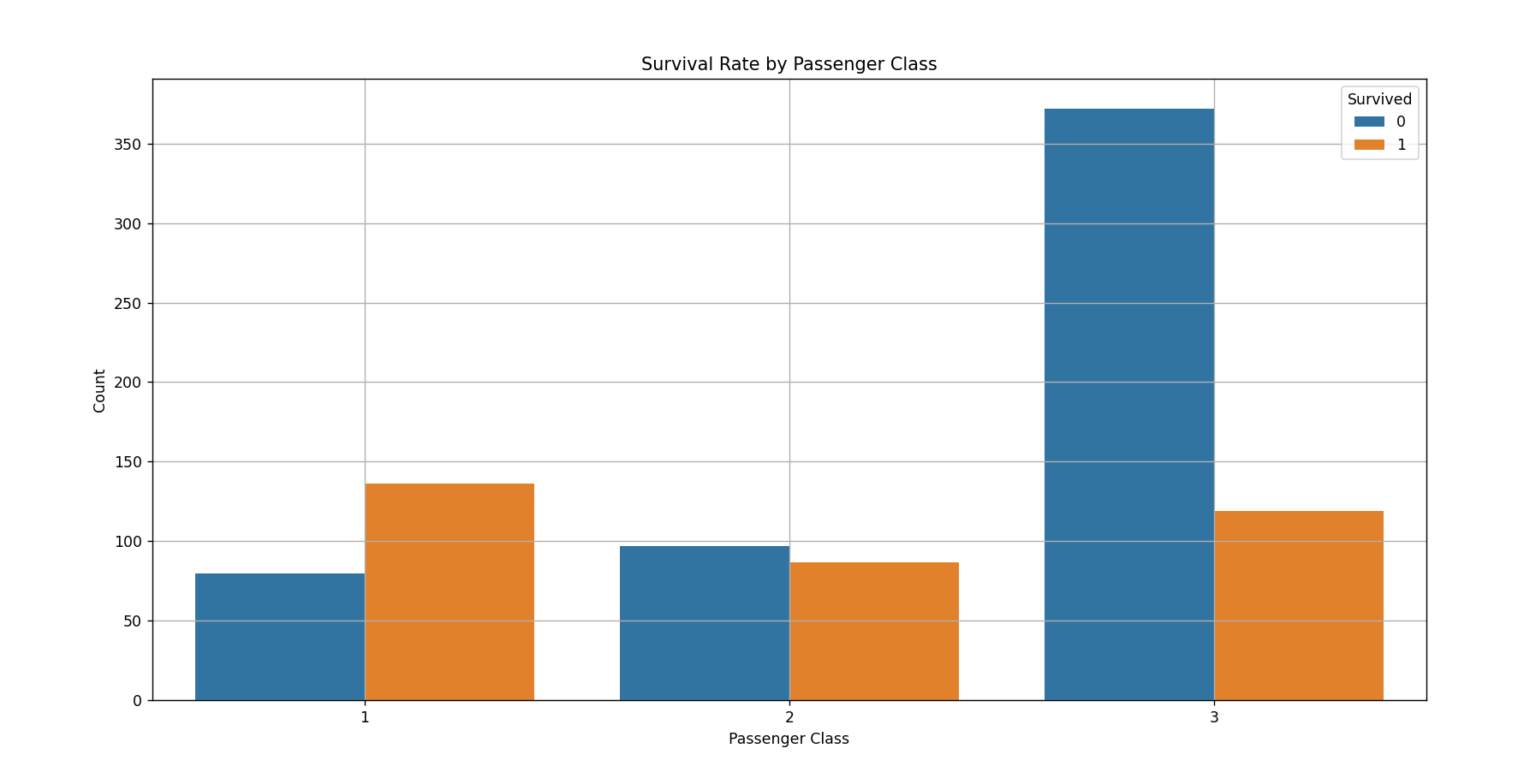


**BIVARIATE DATA ANALYSIS:**

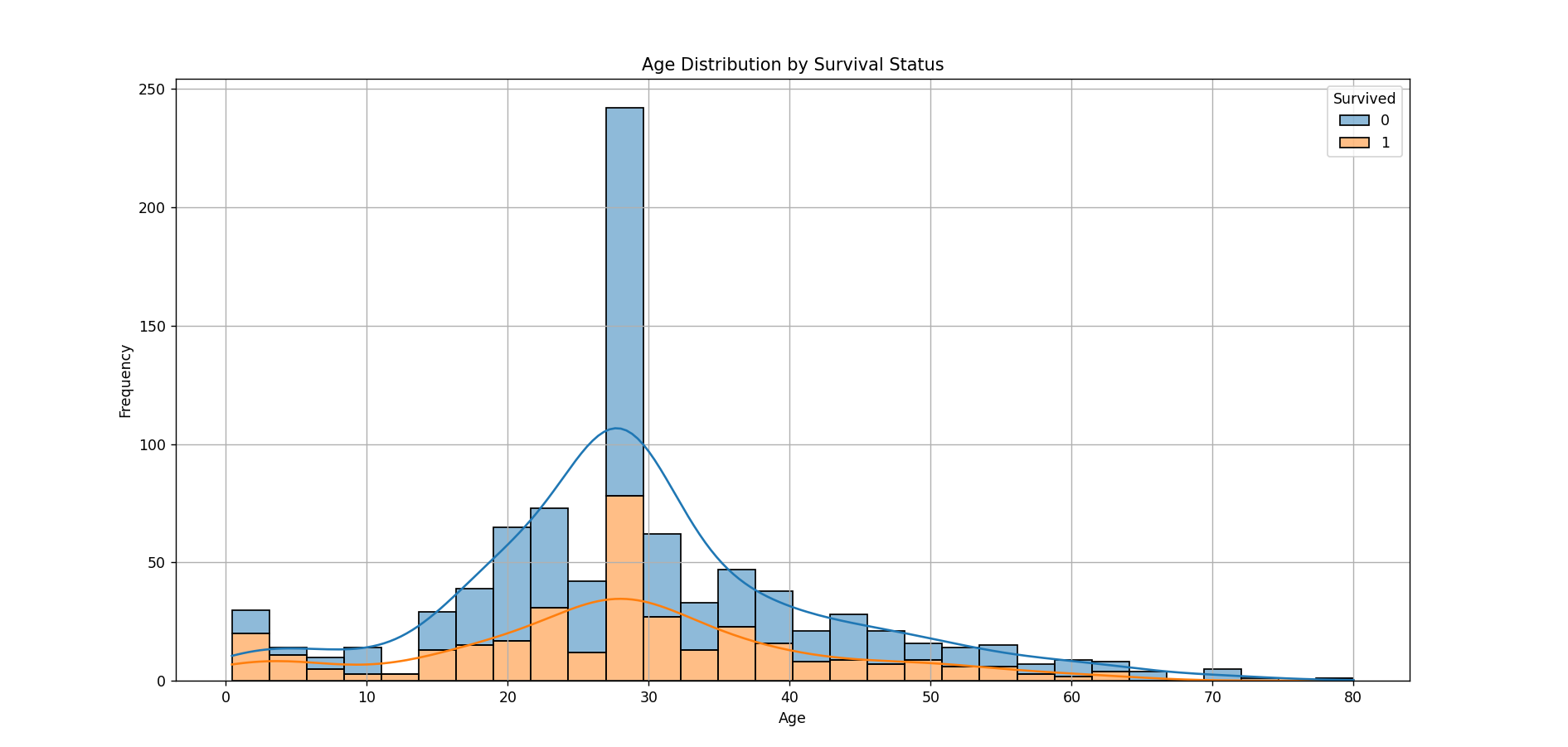
**SURVIVAL RATE BY SEX:**



**SURVIVAL RATE BY PASSENGER CLASS:**

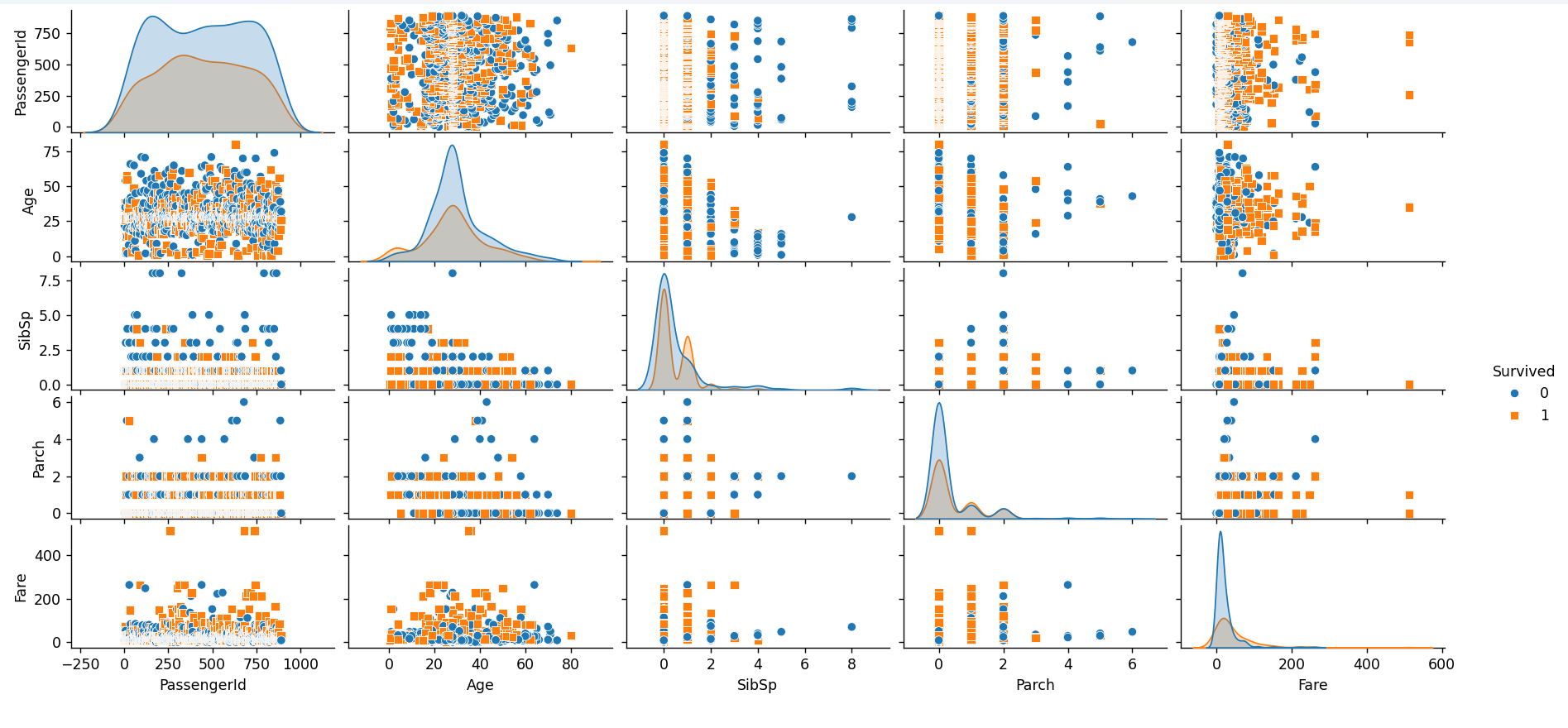


**AGE DISTRIBUTION BY SURVIVAL CLASS:**

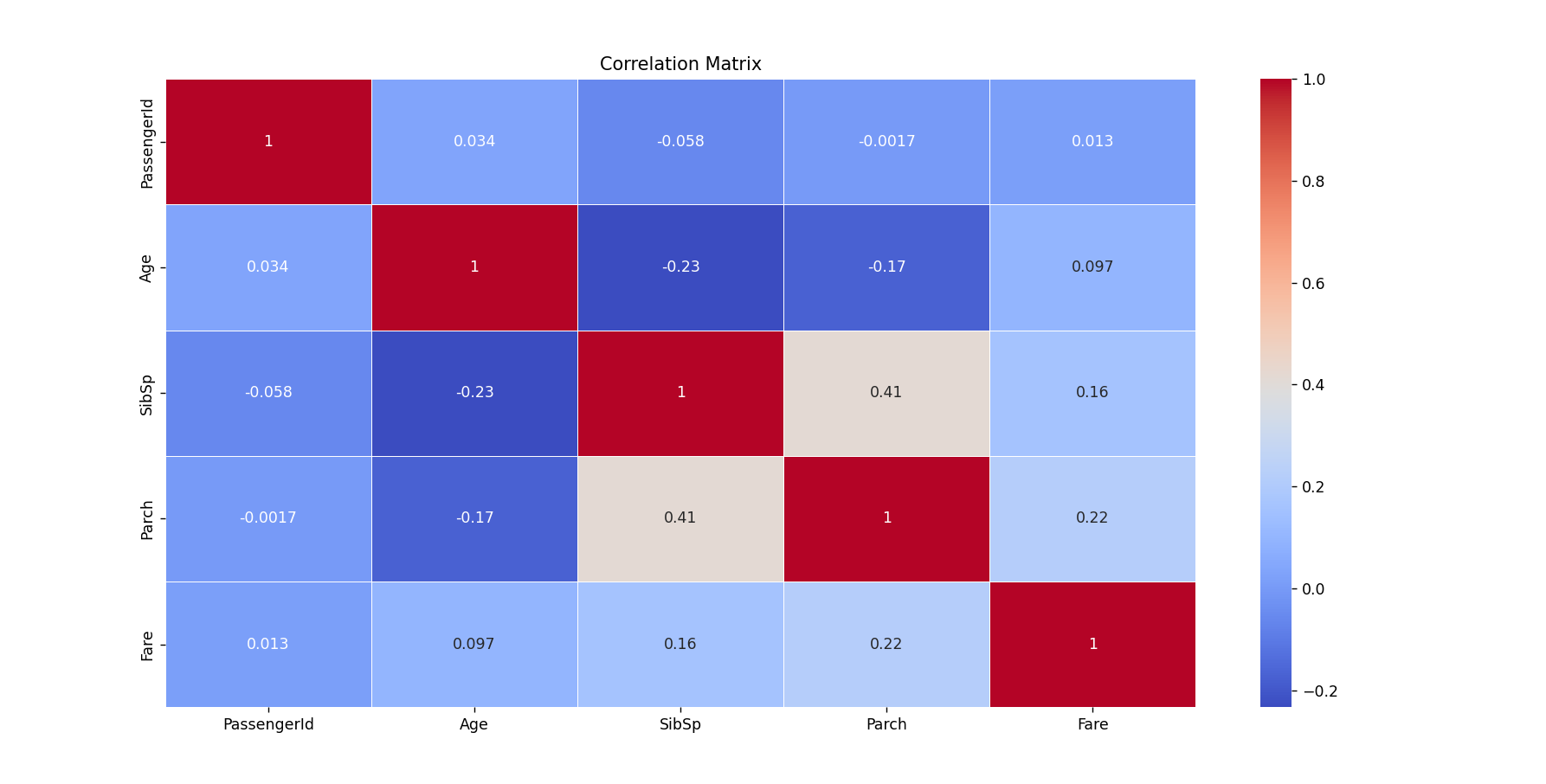


**RELATIONSHIP BETWEEN VARIABLES**

**PAIR PLOT TO EXPLORE RELATIONSHIPS BETWEEN NUMERICAL VARIABLES:**



**HEATMAP OF CORRELATION MATRIX:**



**MULTIVARIATE DATA ANALYSIS:**

**SURVIVAL RATE BY SEX AND PASSENGER CLASS:**

