In [ ]:
```
✅ TASK 1: Exploratory Data Analysis (EDA)
```

In [2]:
```
import pandas as pd
df = pd.read_csv('Titanic-Dataset.csv')
df.head()
```

Out[2]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

In [3]:
```
print(df.columns.tolist())
```

```
['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp', 'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked']
```

In [5]:
```
!pip install scipy
```

```
Collecting scipy
  Downloading scipy-1.15.3-cp310-cp310-win_amd64.whl.metadata (60 kB)
Requirement already satisfied: numpy<2.5,>=1.23.5 in c:\users\asim ali\anaconda3\envs\fresh_env\lib\site-packages (fr
om scipy) (1.26.4)
Downloading scipy-1.15.3-cp310-cp310-win_amd64.whl (41.3 MB)
   ---------------------------------------- 0.0/41.3 MB ? eta -:--:--
   ---- ----------------------------------- 5.0/41.3 MB 27.4 MB/s eta 0:00:02
   ---------- ------------------------------ 11.3/41.3 MB 30.6 MB/s eta 0:00:01
   ---------------- ----------------------- 17.8/41.3 MB 30.3 MB/s eta 0:00:01
   ---------------------- ----------------- 23.6/41.3 MB 29.8 MB/s eta 0:00:01
   ----------------------------- ---------- 30.4/41.3 MB 30.1 MB/s eta 0:00:01
   ------------------------------------ ---- 36.7/41.3 MB 29.9 MB/s eta 0:00:01
   ---------------------------------------  41.2/41.3 MB 30.1 MB/s eta 0:00:01
   ---------------------------------------  41.2/41.3 MB 30.1 MB/s eta 0:00:01
   ---------------------------------------  41.2/41.3 MB 30.1 MB/s eta 0:00:01
   ---------------------------------------- 41.3/41.3 MB 21.0 MB/s eta 0:00:00
Installing collected packages: scipy
Successfully installed scipy-1.15.3
```

In [6]:

```python
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from scipy import stats

sns.set(style="whitegrid")
plt.rcParams['figure.figsize'] = (10, 6)

# Load the dataset
df = pd.read_csv('Titanic-Dataset.csv')

print("=== Dataset Overview ===")
print(f"Shape: {df.shape}")  # (rows, columns)
print("\nFirst 5 rows:")
print(df.head())
print("\nData types and non-null counts:")
print(df.info())
print("\nStatistical summary:")
print(df.describe(include='all'))

print("\n=== Missing Values ===")
print(df.isnull().sum().sort_values(ascending=False))
```

```
plt.figure(figsize=(10, 4))
sns.heatmap(df.isnull(), cbar=False, cmap='viridis')
plt.title('Missing Values Heatmap')
plt.show()


plt.figure()
sns.countplot(x='Survived', data=df)
plt.title('Survival Count (0 = Died, 1 = Survived)')
plt.show()

print(f"\nSurvival Rate: {df['Survived'].mean():.2%}")

plt.figure()
sns.countplot(x='Pclass', data=df)
plt.title('Passenger Class Distribution')
plt.show()

plt.figure()
sns.countplot(x='Sex', data=df)
plt.title('Gender Distribution')
plt.show()

plt.figure()
sns.histplot(df['Age'].dropna(), kde=True, bins=30)
plt.title('Age Distribution')
plt.show()

plt.figure()
sns.histplot(df['Fare'], kde=True, bins=30)
plt.title('Fare Distribution')
plt.show()

# Embarkation Port
plt.figure()
sns.countplot(x='Embarked', data=df)
plt.title('Embarkation Port Distribution')
plt.show()


# Survival by Passenger Class
```

```
plt.figure()
sns.barplot(x='Pclass', y='Survived', data=df, ci=None)
plt.title('Survival Rate by Passenger Class')
plt.ylabel('Survival Rate')
plt.show()

# Survival by Gender
plt.figure()
sns.barplot(x='Sex', y='Survived', data=df, ci=None)
plt.title('Survival Rate by Gender')
plt.ylabel('Survival Rate')
plt.show()

# Age vs Survival
plt.figure()
sns.boxplot(x='Survived', y='Age', data=df)
plt.title('Age Distribution by Survival')
plt.show()

# Fare vs Survival
plt.figure()
sns.boxplot(x='Survived', y='Fare', data=df)
plt.title('Fare Distribution by Survival')
plt.show()

# Pclass, Sex, and Survival
plt.figure()
sns.catplot(x='Pclass', y='Survived', hue='Sex', kind='bar', data=df, ci=None)
plt.title('Survival Rate by Class and Gender')
plt.ylabel('Survival Rate')
plt.show()

# Age and Fare relationship
plt.figure()
sns.scatterplot(x='Age', y='Fare', hue='Survived', data=df, alpha=0.6)
plt.title('Age vs Fare by Survival Status')
plt.show()

# Family Size Analysis
df['FamilySize'] = df['SibSp'] + df['Parch'] + 1
plt.figure()
sns.countplot(x='FamilySize', data=df)
```

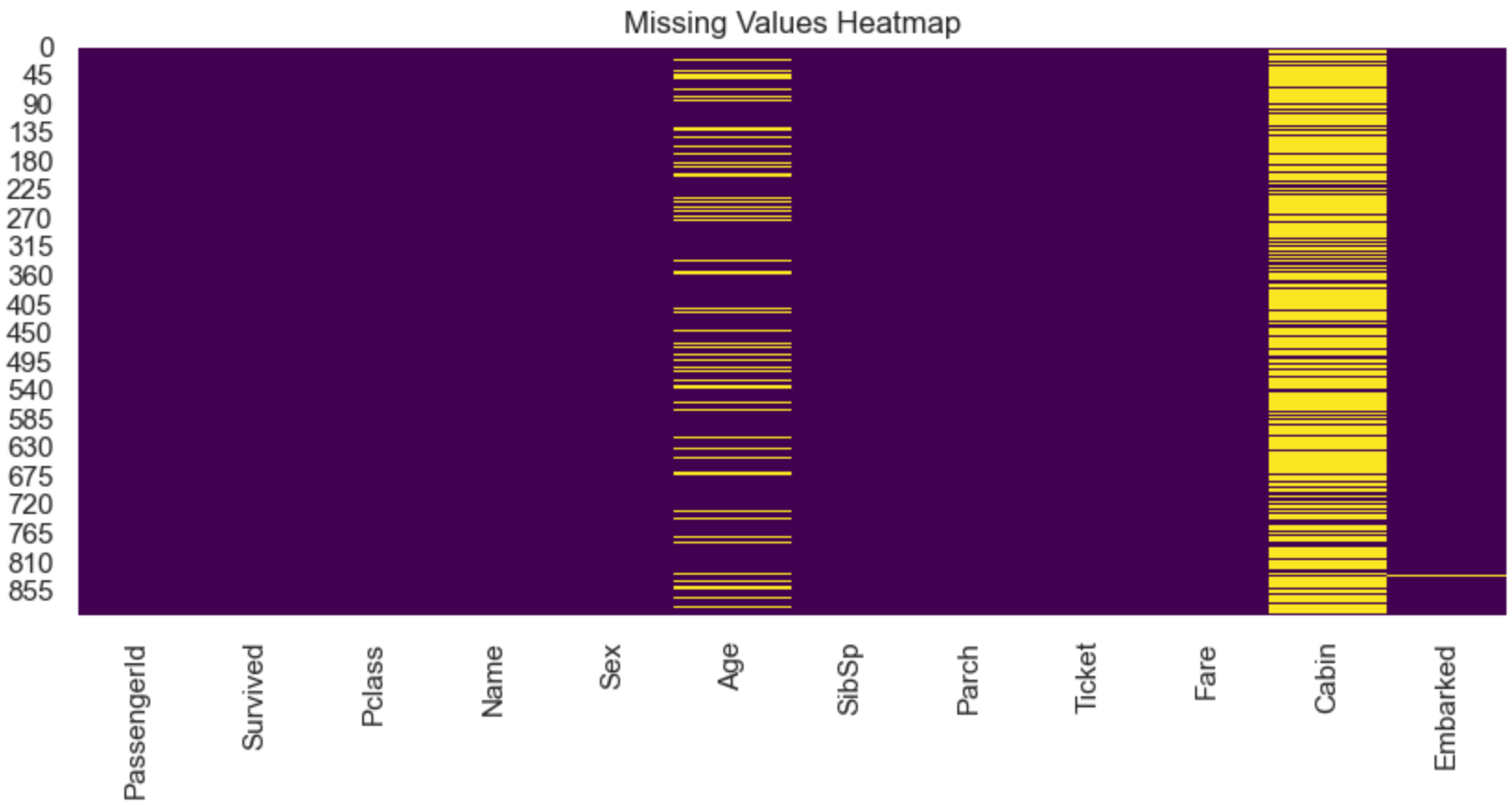
```
plt.title('Family Size Distribution')
plt.show()

plt.figure()
sns.barplot(x='FamilySize', y='Survived', data=df, ci=None)
plt.title('Survival Rate by Family Size')
plt.ylabel('Survival Rate')
plt.show()

# Correlation Analysis
numeric_cols = df.select_dtypes(include=[np.number]).columns
corr_matrix = df[numeric_cols].corr()

plt.figure()
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', center=0)
plt.title('Correlation Matrix')
plt.show()

## Outlier Detection
# For Age
plt.figure()
sns.boxplot(x=df['Age'])
plt.title('Age Boxplot for Outlier Detection')
plt.show()

# For Fare
plt.figure()
sns.boxplot(x=df['Fare'])
plt.title('Fare Boxplot for Outlier Detection')
plt.show()

## Hypothesis Testing

# Hypothesis 1: Higher class passengers had better survival rates
print("\n=== Hypothesis Testing ===")
contingency_table = pd.crosstab(df['Pclass'], df['Survived'])
chi2, p, dof, expected = stats.chi2_contingency(contingency_table)
print(f"Chi-square test for Pclass vs Survival: p-value = {p:.4f}")

# Hypothesis 2: Females had higher survival rates than males
female_survival = df[df['Sex'] == 'female']['Survived'].mean()
male_survival = df[df['Sex'] == 'male']['Survived'].mean()
```

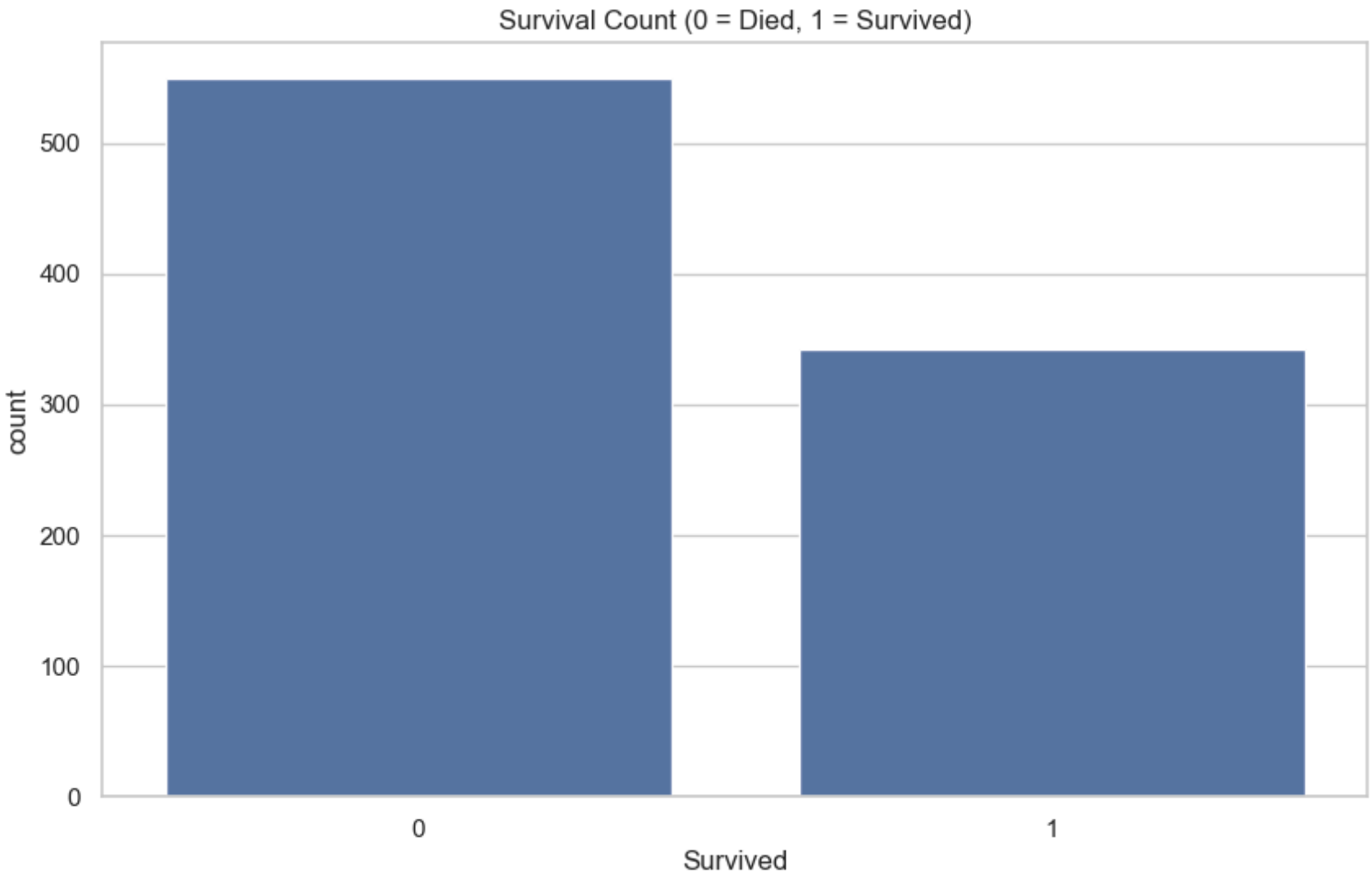
```python
print(f"\nFemale survival rate: {female_survival:.2%}")
print(f"Male survival rate: {male_survival:.2%}")

# T-test for age difference between survivors and non-survivors
survived_age = df[df['Survived'] == 1]['Age'].dropna()
died_age = df[df['Survived'] == 0]['Age'].dropna()
t_stat, p_val = stats.ttest_ind(survived_age, died_age)
print(f"\nT-test for age difference: p-value = {p_val:.4f}")

## Interesting Observations
df['Title'] = df['Name'].str.extract(' ([A-Za-z]+)\.', expand=False)
print("\n=== Title Analysis ===")
print(pd.crosstab(df['Title'], df['Sex']))

# Survival by title
plt.figure(figsize=(12, 4))
sns.countplot(x='Title', hue='Survived', data=df)
plt.xticks(rotation=45)
plt.title('Survival Count by Title')
plt.show()
```

```
=== Dataset Overview ===
Shape: (891, 12)

First 5 rows:
   PassengerId  Survived  Pclass  \
0            1         0       3
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3

                                                Name     Sex   Age  SibSp  \
0                            Braund, Mr. Owen Harris    male  22.0      1
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1
2                             Heikkinen, Miss. Laina  female  26.0      0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                           Allen, Mr. William Henry    male  35.0      0

   Parch            Ticket     Fare Cabin Embarked
0      0         A/5 21171   7.2500   NaN        S
1      0          PC 17599  71.2833   C85        C
2      0  STON/O2. 3101282   7.9250   NaN        S
3      0            113803  53.1000  C123        S
4      0            373450   8.0500   NaN        S

Data types and non-null counts:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
```

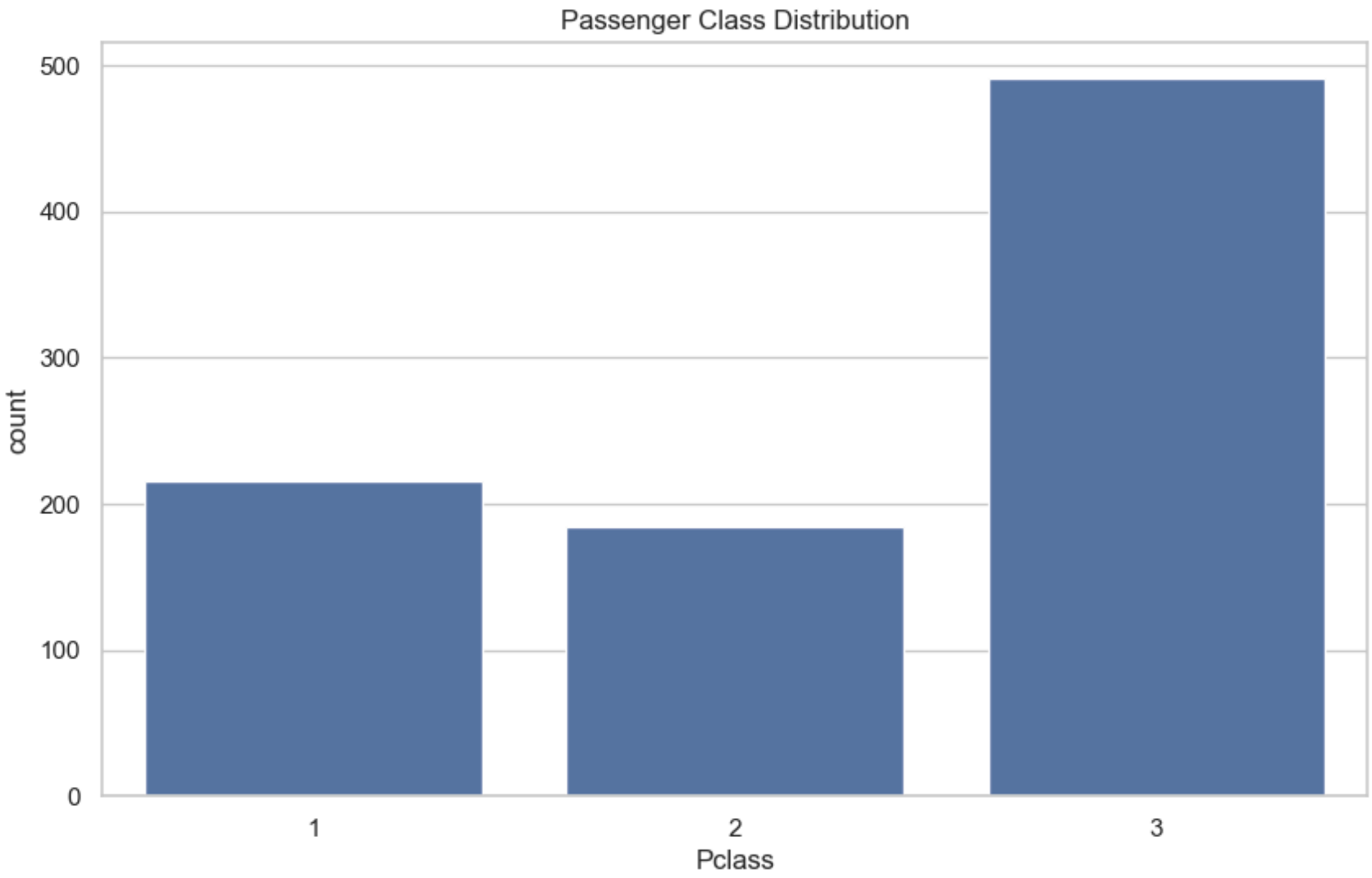
```
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
None

Statistical summary:
        PassengerId    Survived      Pclass                     Name   Sex  \
count    891.000000  891.000000  891.000000                      891   891
unique          NaN         NaN         NaN                      891     2
top             NaN         NaN         NaN  Braund, Mr. Owen Harris  male
freq            NaN         NaN         NaN                        1   577
mean     446.000000    0.383838    2.308642                      NaN   NaN
std      257.353842    0.486592    0.836071                      NaN   NaN
min        1.000000    0.000000    1.000000                      NaN   NaN
25%      223.500000    0.000000    2.000000                      NaN   NaN
50%      446.000000    0.000000    3.000000                      NaN   NaN
75%      668.500000    1.000000    3.000000                      NaN   NaN
max      891.000000    1.000000    3.000000                      NaN   NaN

               Age       SibSp       Parch  Ticket        Fare    Cabin  \
count   714.000000  891.000000  891.000000     891  891.000000      204
unique         NaN         NaN         NaN     681         NaN      147
top            NaN         NaN         NaN  347082         NaN  B96 B98
freq           NaN         NaN         NaN       7         NaN        4
mean     29.699118    0.523008    0.381594     NaN   32.204208      NaN
std      14.526497    1.102743    0.806057     NaN   49.693429      NaN
min       0.420000    0.000000    0.000000     NaN    0.000000      NaN
25%      20.125000    0.000000    0.000000     NaN    7.910400      NaN
50%      28.000000    0.000000    0.000000     NaN   14.454200      NaN
75%      38.000000    1.000000    0.000000     NaN   31.000000      NaN
max      80.000000    8.000000    6.000000     NaN  512.329200      NaN

       Embarked
count       889
unique        3
top           S
freq        644
mean        NaN
std         NaN
min         NaN
25%         NaN
50%         NaN
```

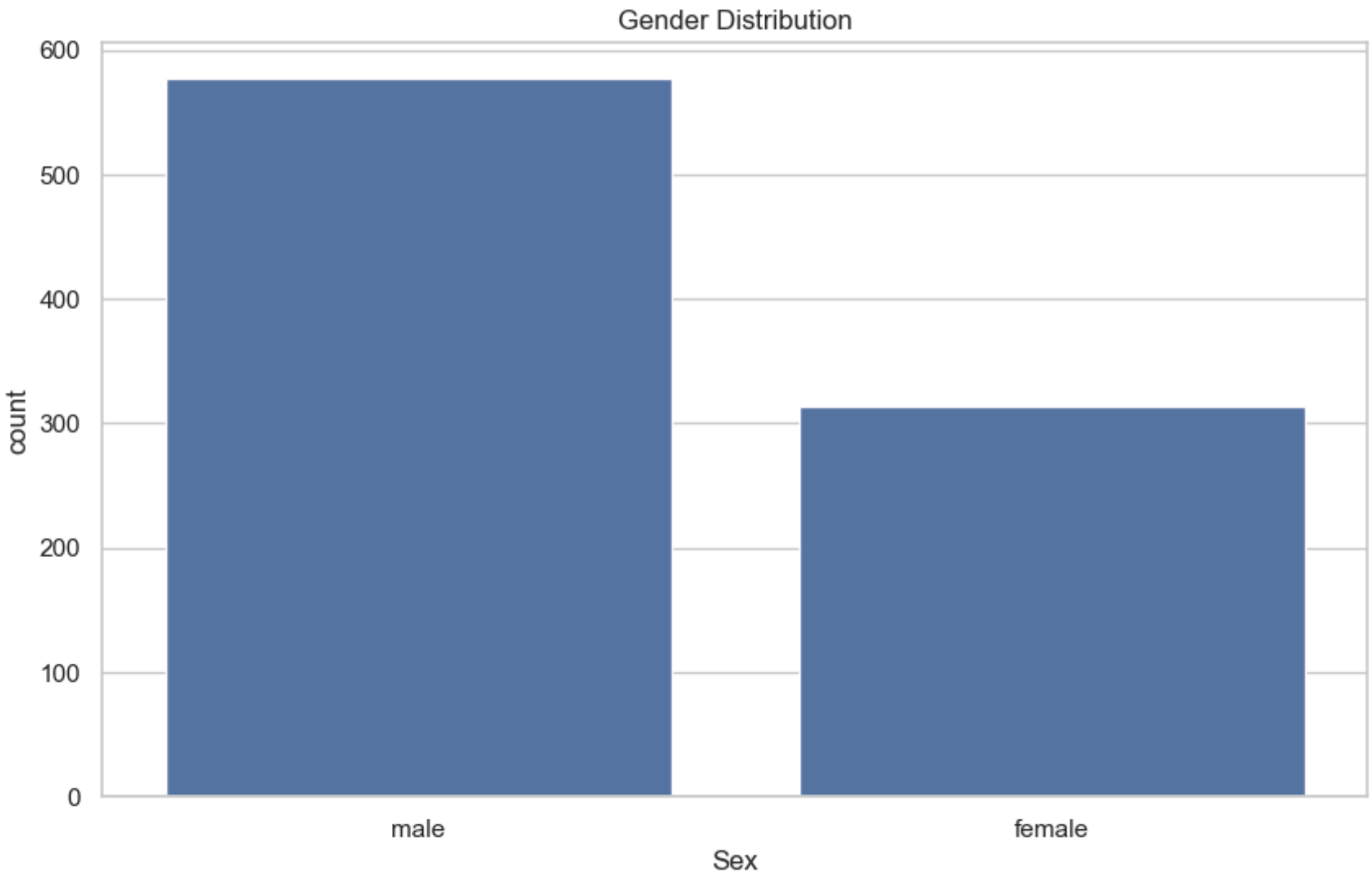
```
75%         NaN
max         NaN

=== Missing Values ===
Cabin          687
Age            177
Embarked         2
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
SibSp            0
Parch            0
Ticket           0
Fare             0
dtype: int64
```

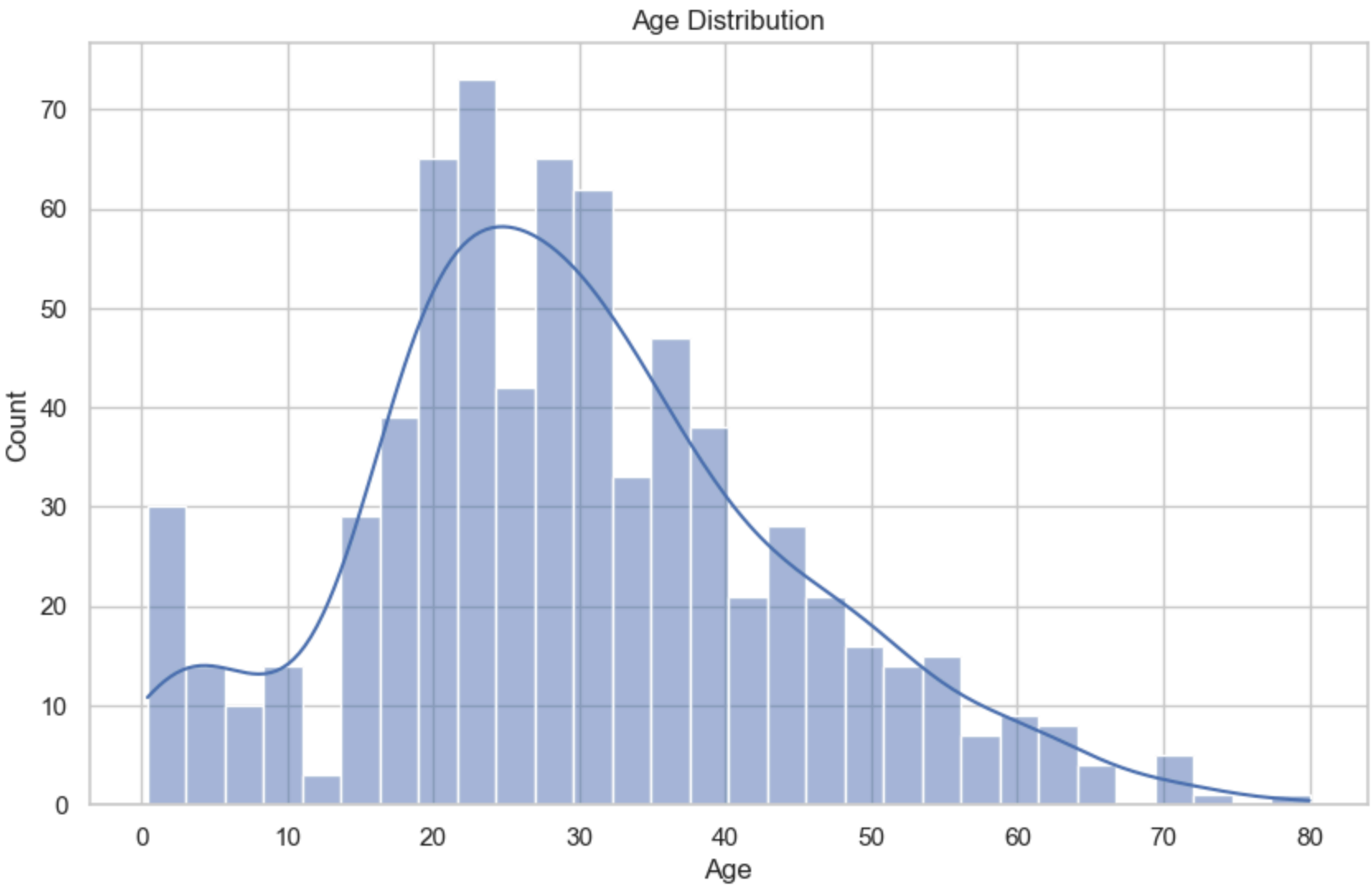

Missing Values Heatmap
0
45
90
135
180
225
270
315
360
405
450
495
540
585
630
675
720
765
810
855
PassengerId
Survived
Pclass
Name
Sex
Age
SibSp
Parch
Ticket
Fare
Cabin
Embarked

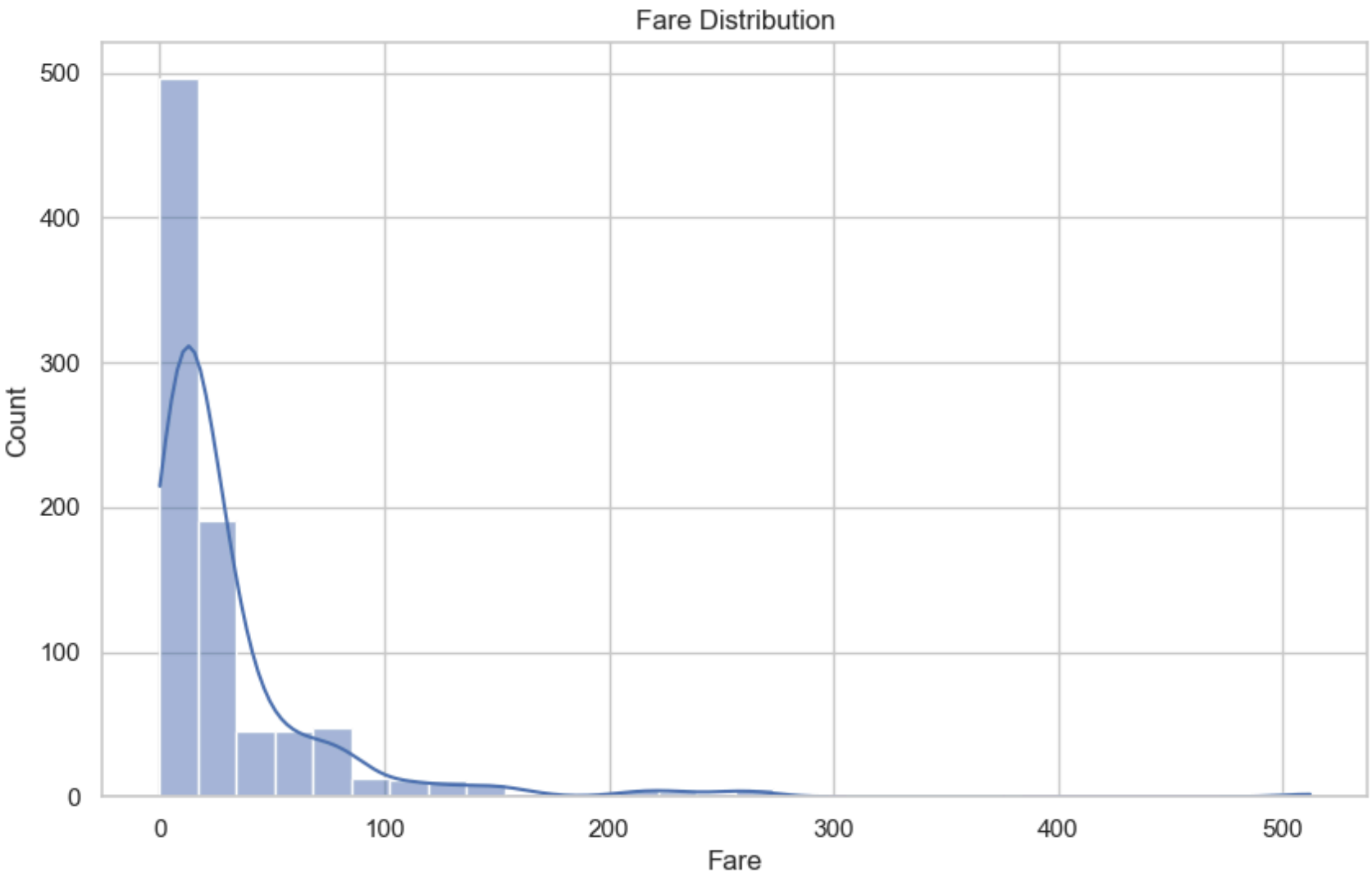


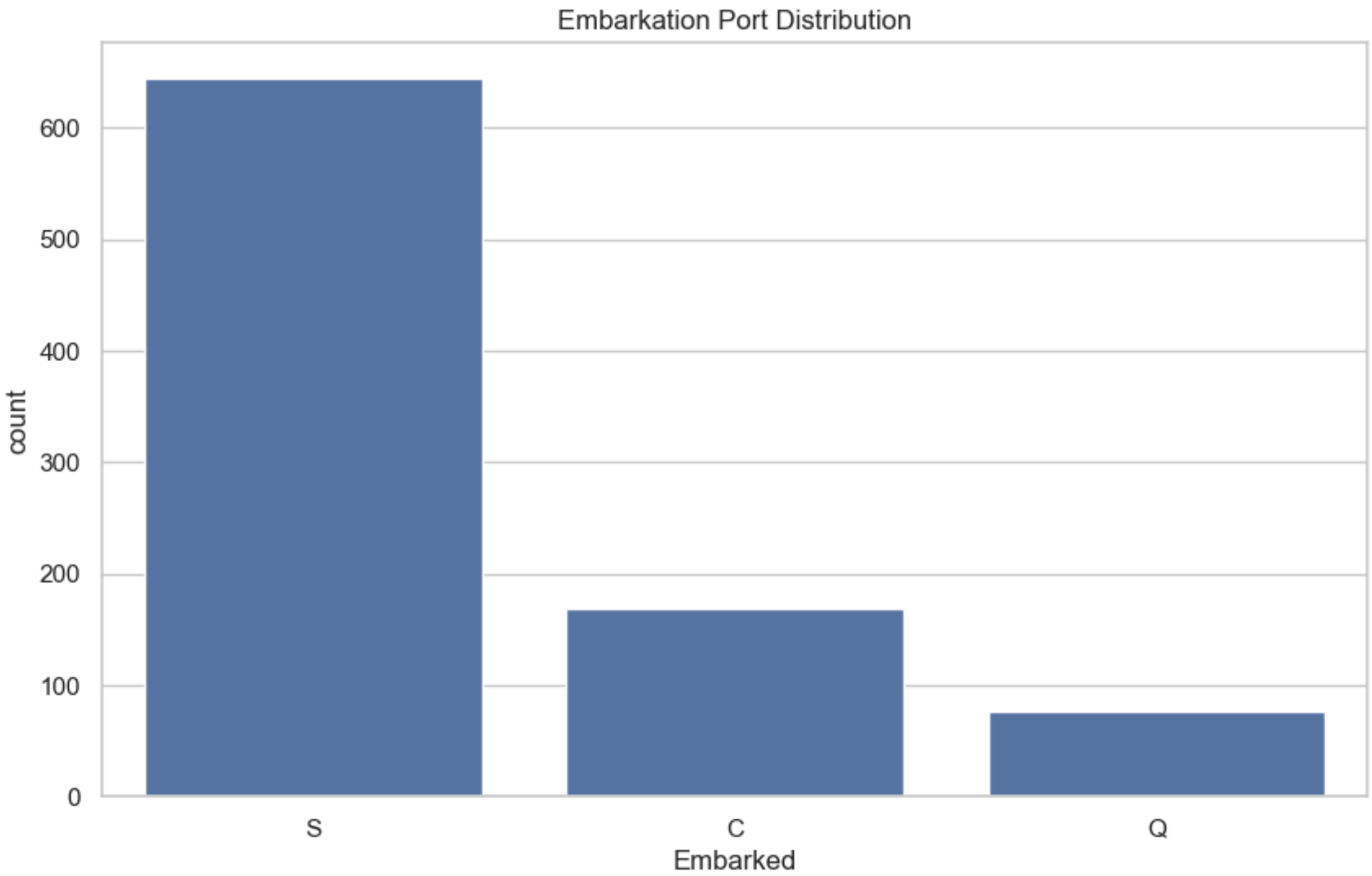

```
Survival Rate: 38.38%
```


Passenger Class Distribution
500
400
300
200
100
0
count
1
2
3
Pclass

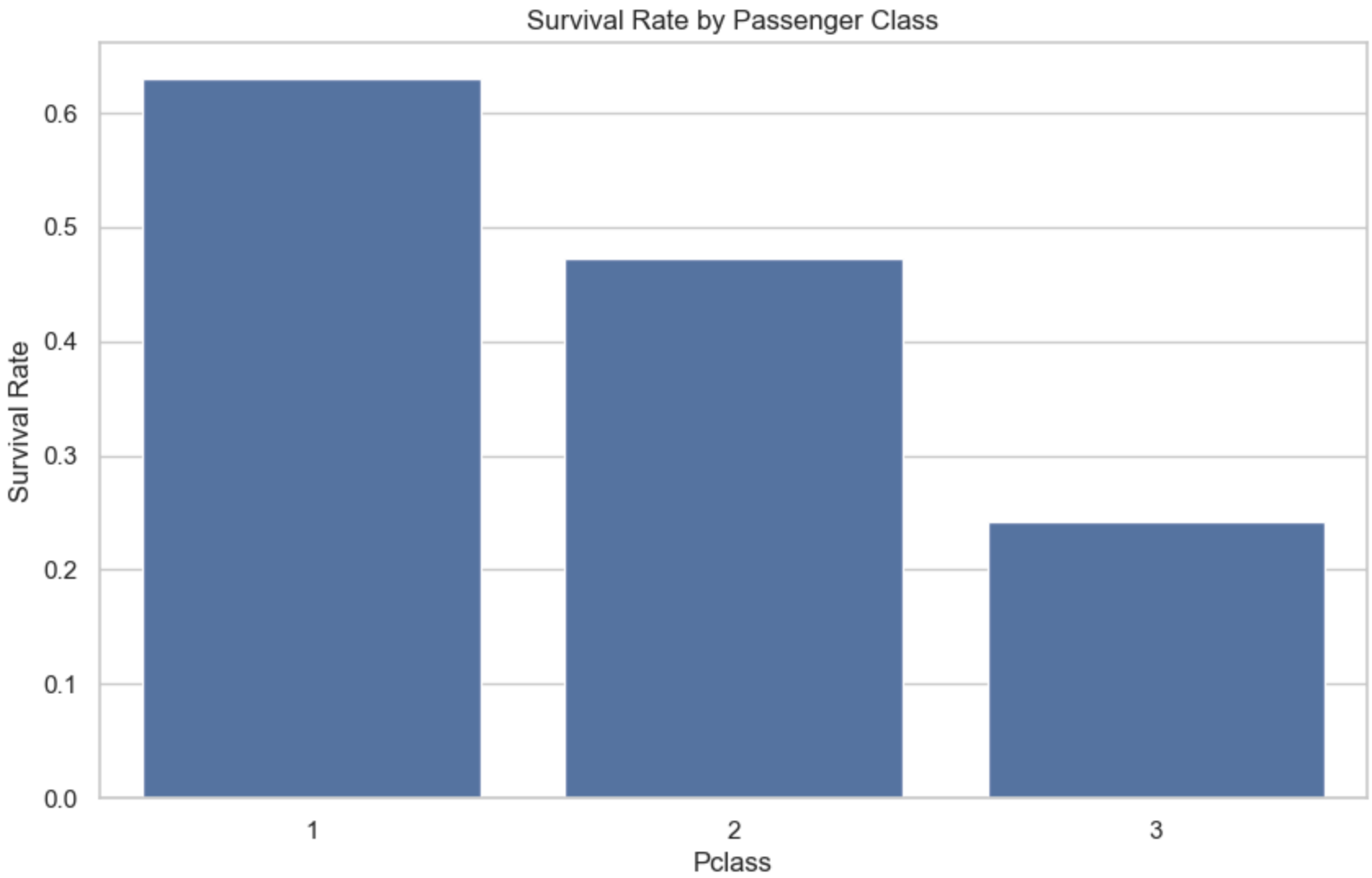

Gender Distribution
600
500
400
300
200
100
0
count
male
female
Sex


Age Distribution
Count
0
10
20
30
40
50
60
70
0
10
20
30
40
50
60
70
80
Age

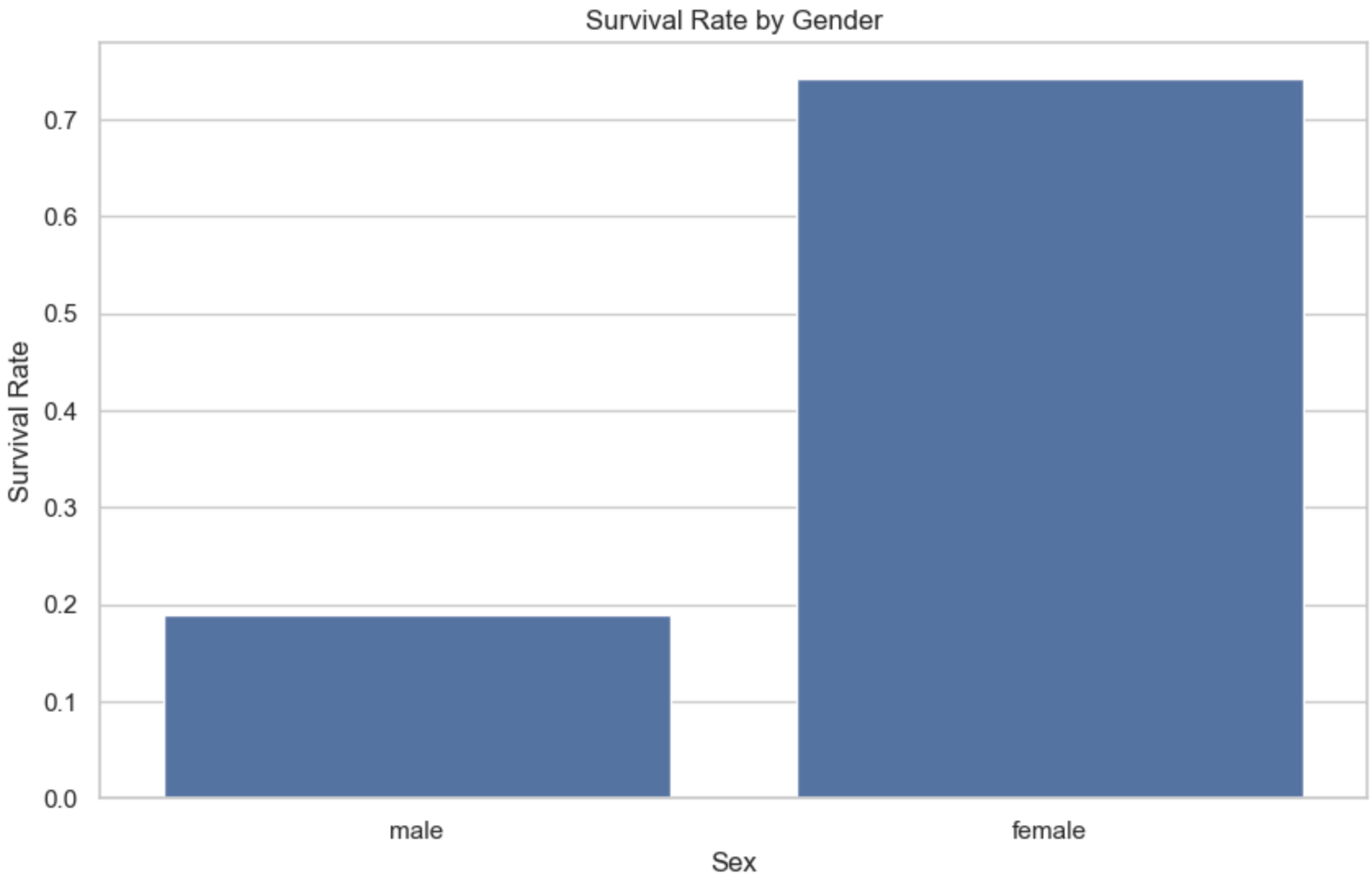

Fare Distribution
Count
500
400
300
200
100
0
0
100
200
300
400
500
Fare

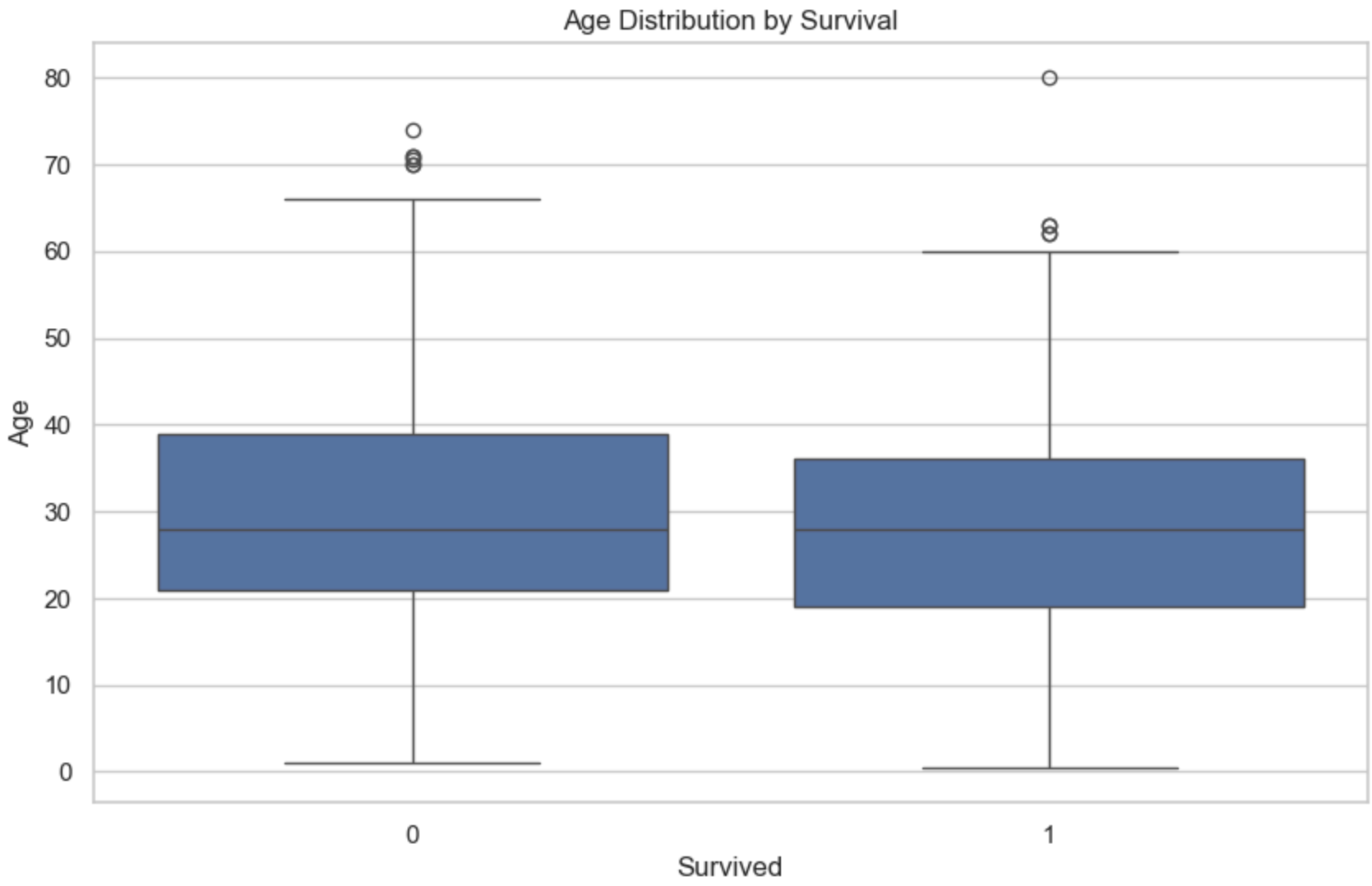


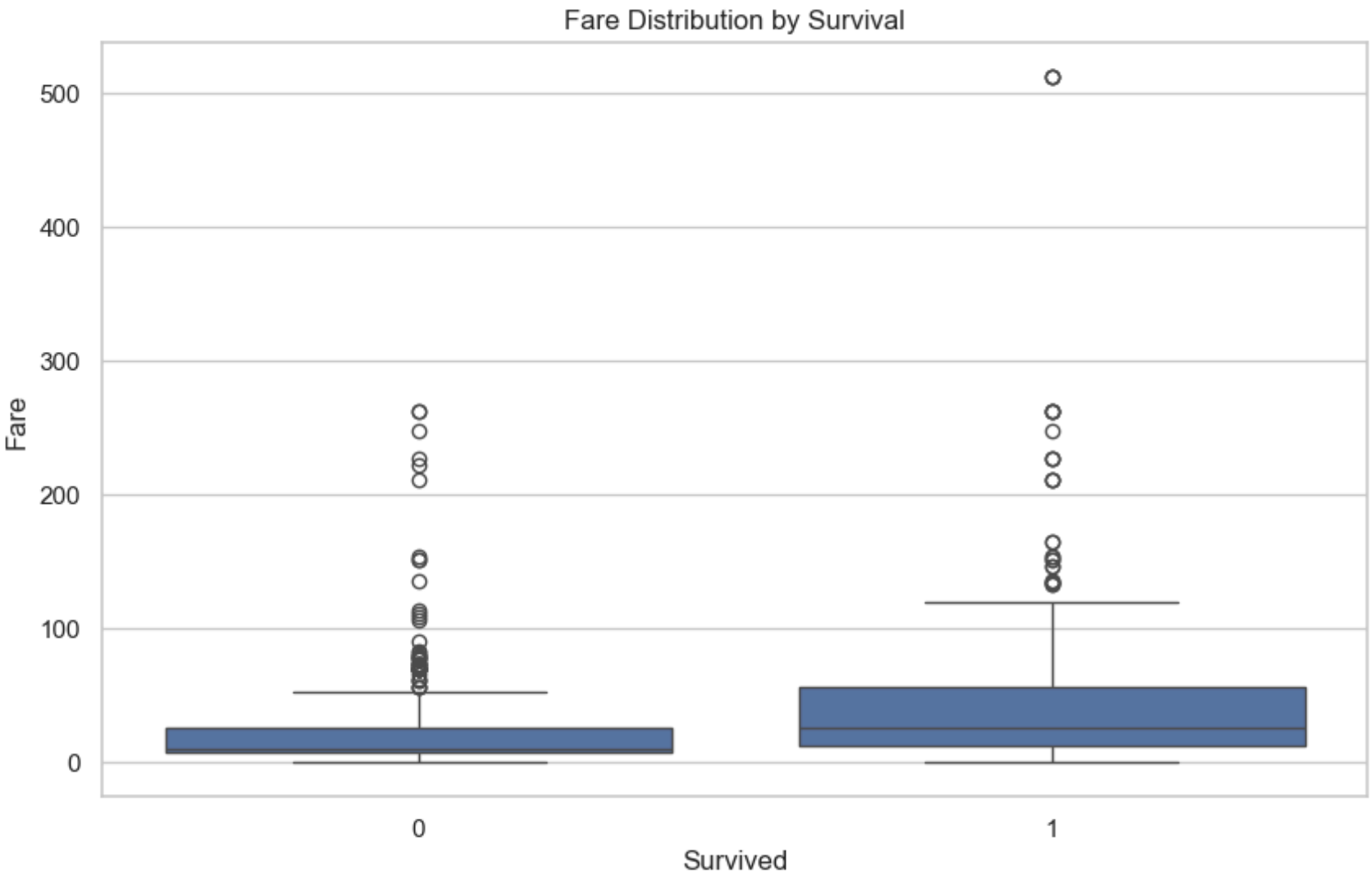

```
C:\Users\ASIM ALI\AppData\Local\Temp\ipykernel_9864\17829154.py:78: FutureWarning:

The `ci` parameter is deprecated. Use `errorbar=None` for the same effect.

  sns.barplot(x='Pclass', y='Survived', data=df, ci=None)
```



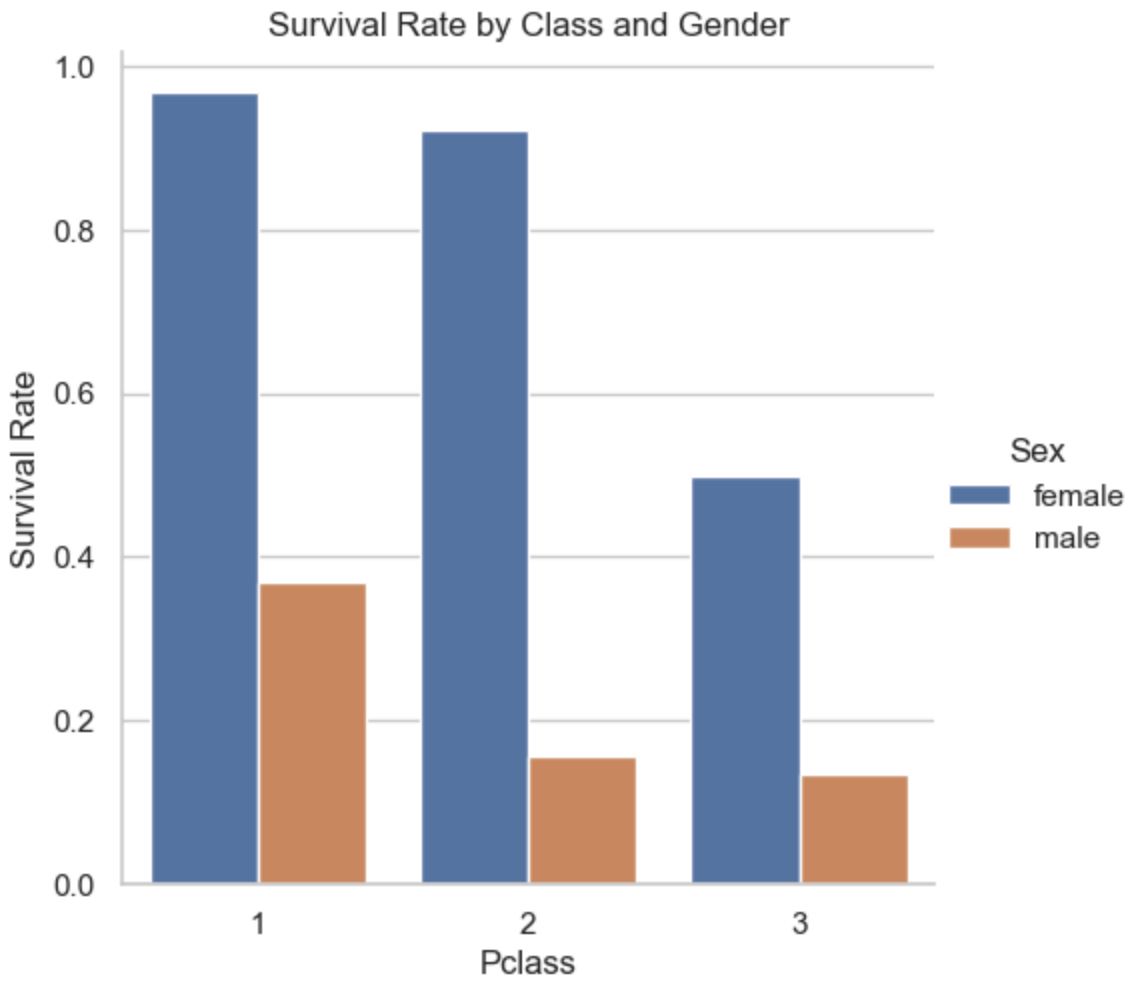

```
C:\Users\ASIM ALI\AppData\Local\Temp\ipykernel_9864\17829154.py:85: FutureWarning:

The `ci` parameter is deprecated. Use `errorbar=None` for the same effect.

  sns.barplot(x='Sex', y='Survived', data=df, ci=None)
```

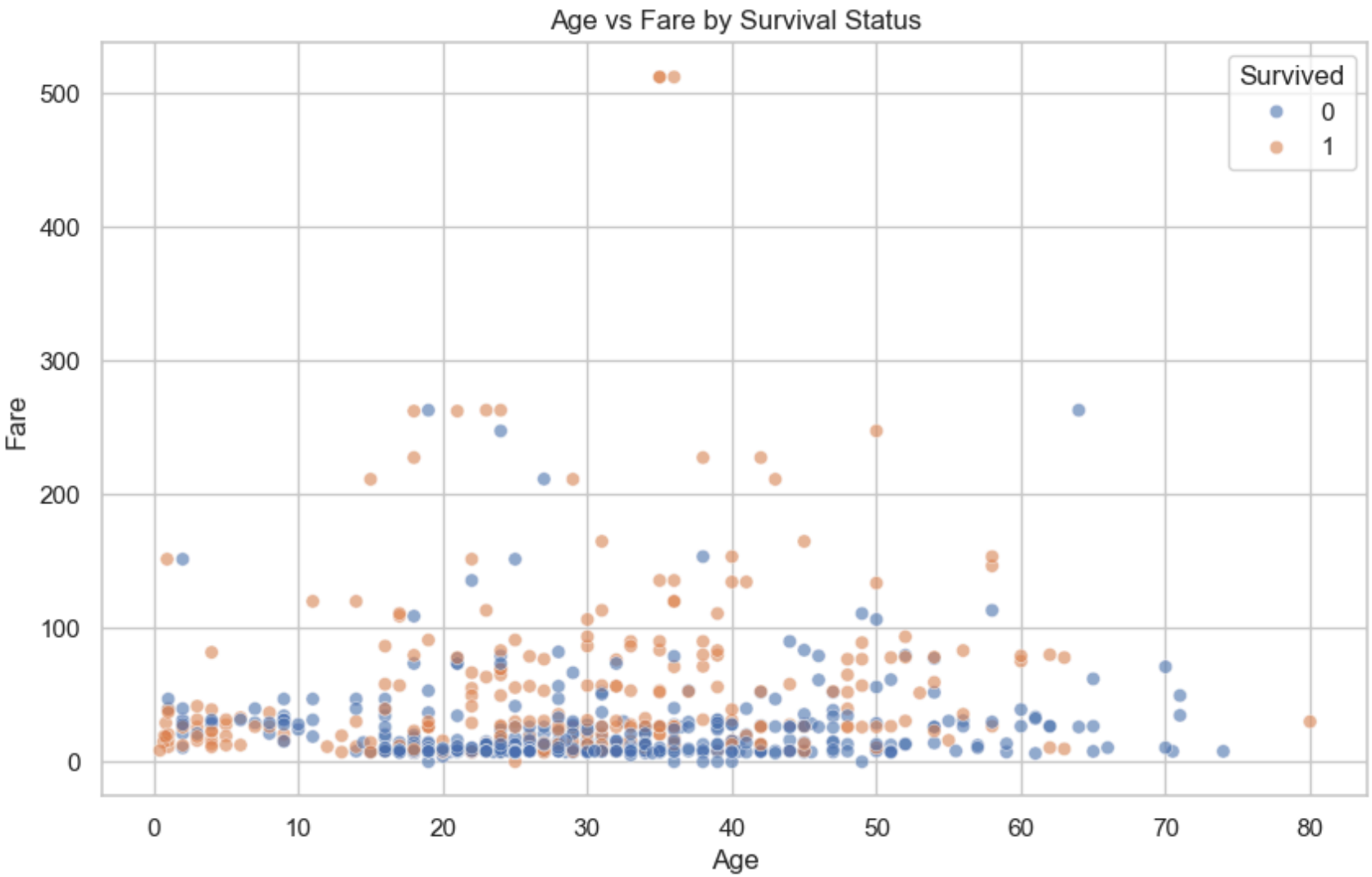

Survival Rate by Gender
Survival Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
male
female
Sex

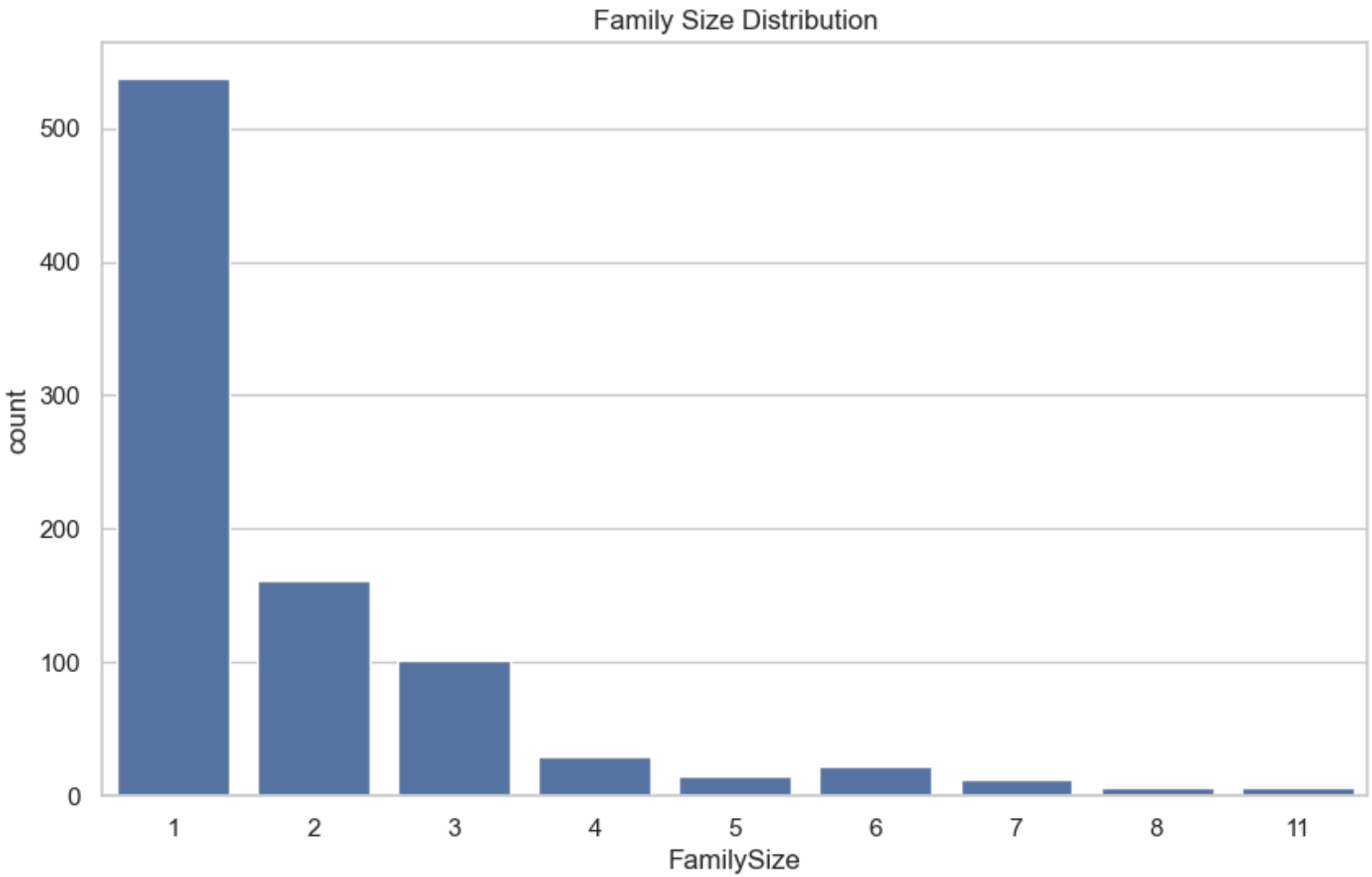

Age Distribution by Survival
Age
80
70
60
50
40
30
20
10
0
0
1
Survived



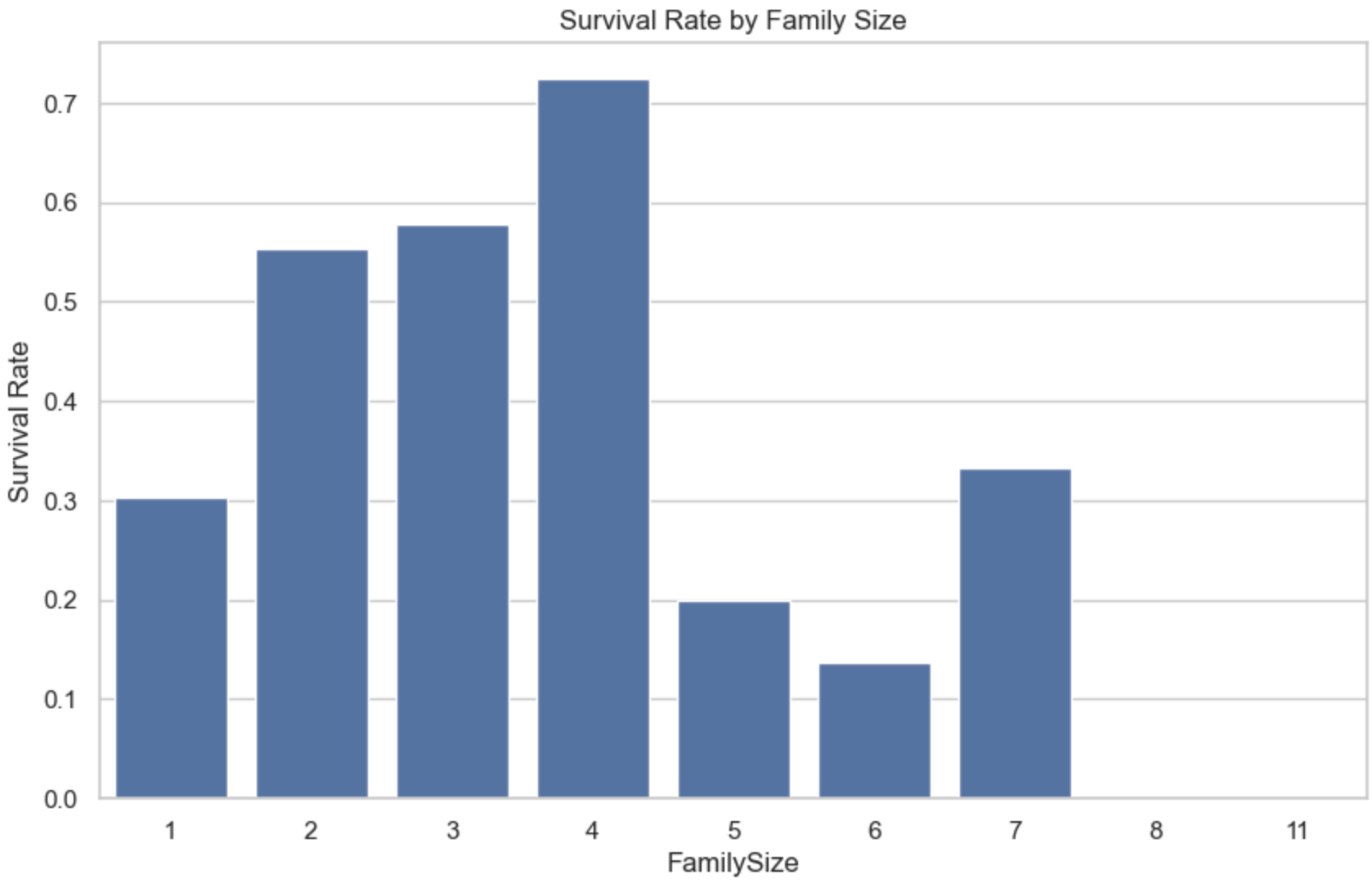

```
C:\Users\ASIM ALI\AppData\Local\Temp\ipykernel_9864\17829154.py:104: FutureWarning:

The `ci` parameter is deprecated. Use `errorbar=None` for the same effect.

  sns.catplot(x='Pclass', y='Survived', hue='Sex', kind='bar', data=df, ci=None)
<Figure size 1000x600 with 0 Axes>
```

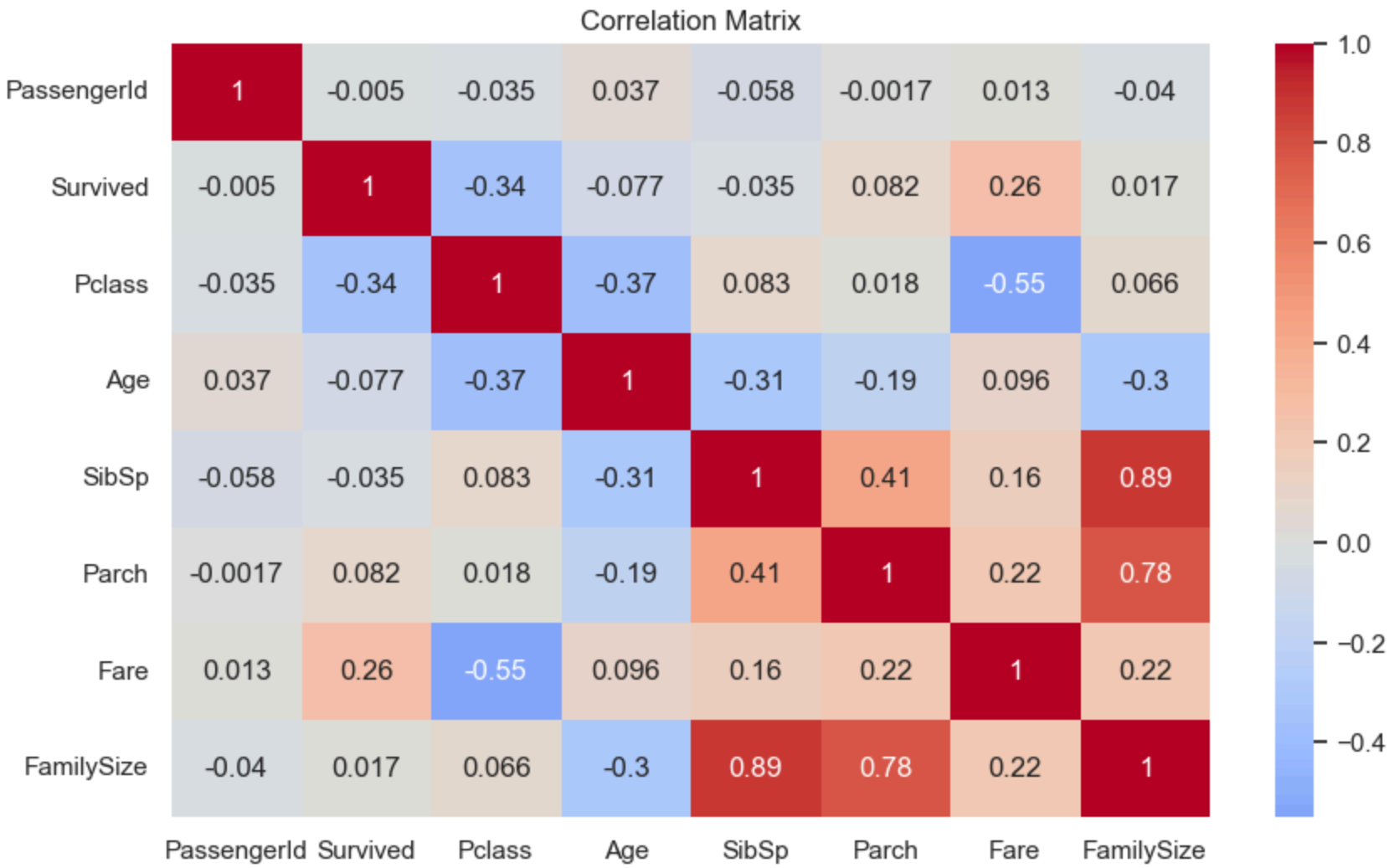

Survival Rate by Class and Gender
1.0
0.8
0.6
0.4
0.2
0.0
Survival Rate
1
2
3
Pclass
Sex
female
male

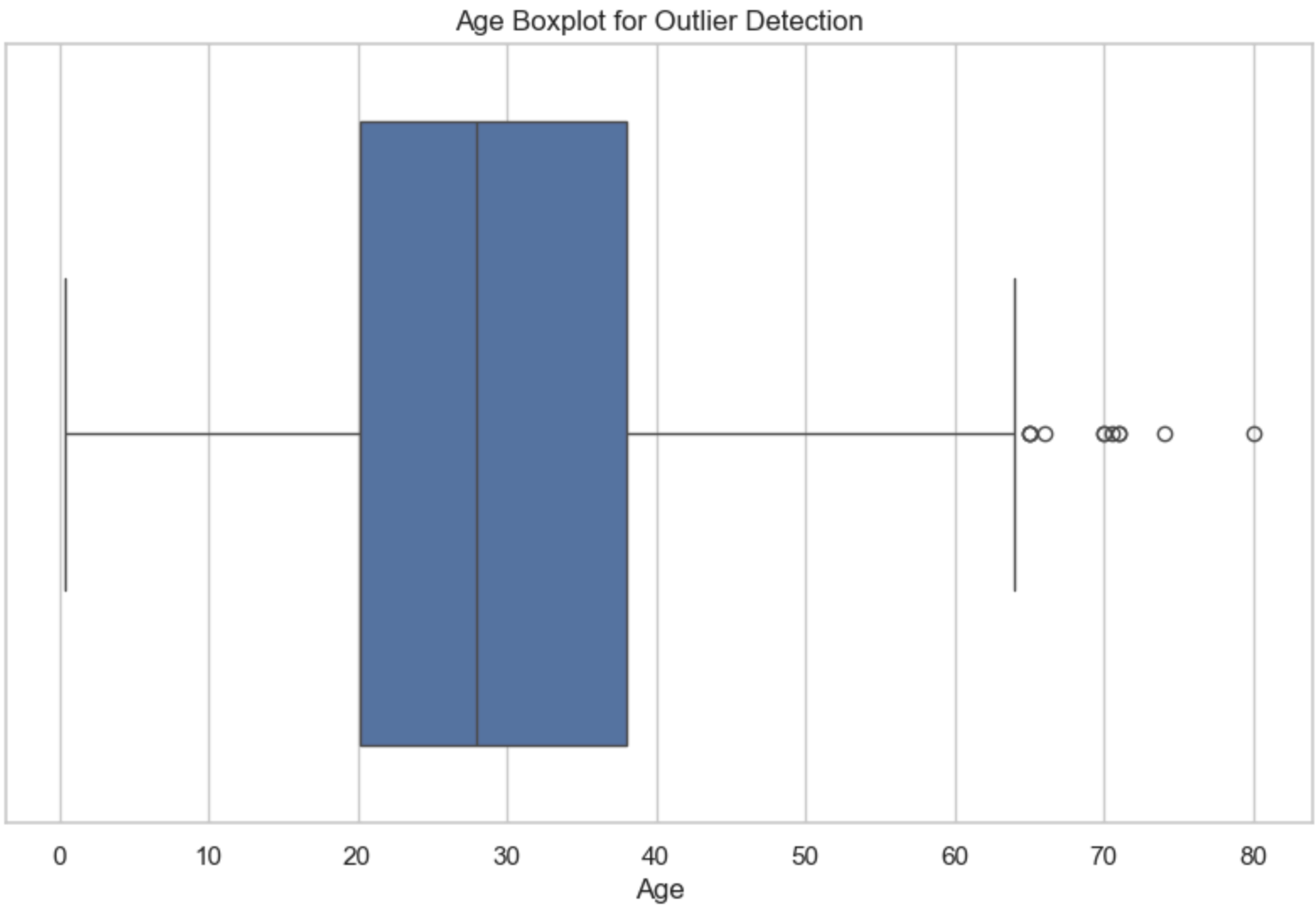

Age vs Fare by Survival Status
Survived
0
1
Fare
500
400
300
200
100
0
0
10
20
30
40
50
60
70
80
Age

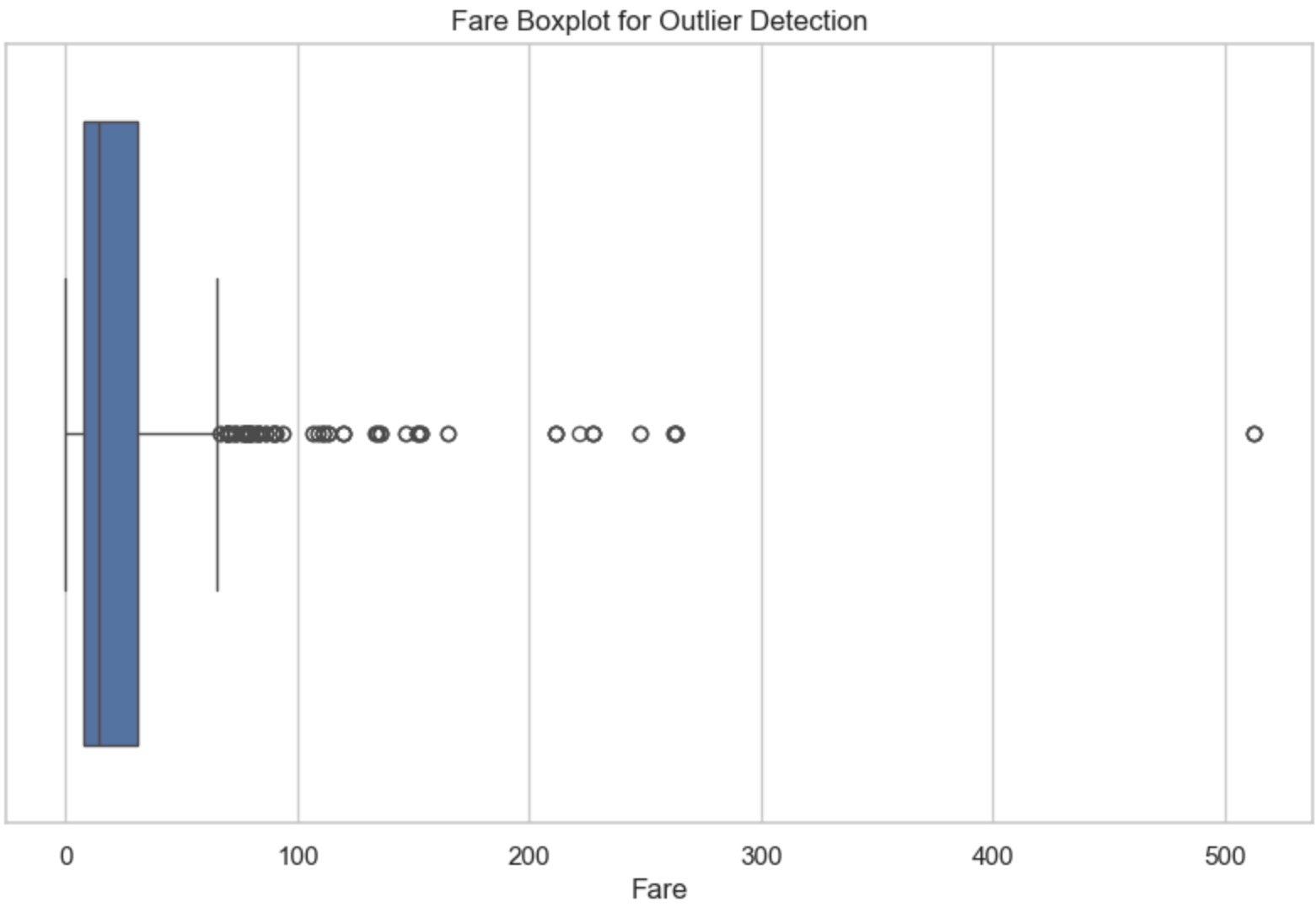



```
C:\Users\ASIM ALI\AppData\Local\Temp\ipykernel_9864\17829154.py:123: FutureWarning:

The `ci` parameter is deprecated. Use `errorbar=None` for the same effect.

  sns.barplot(x='FamilySize', y='Survived', data=df, ci=None)
```

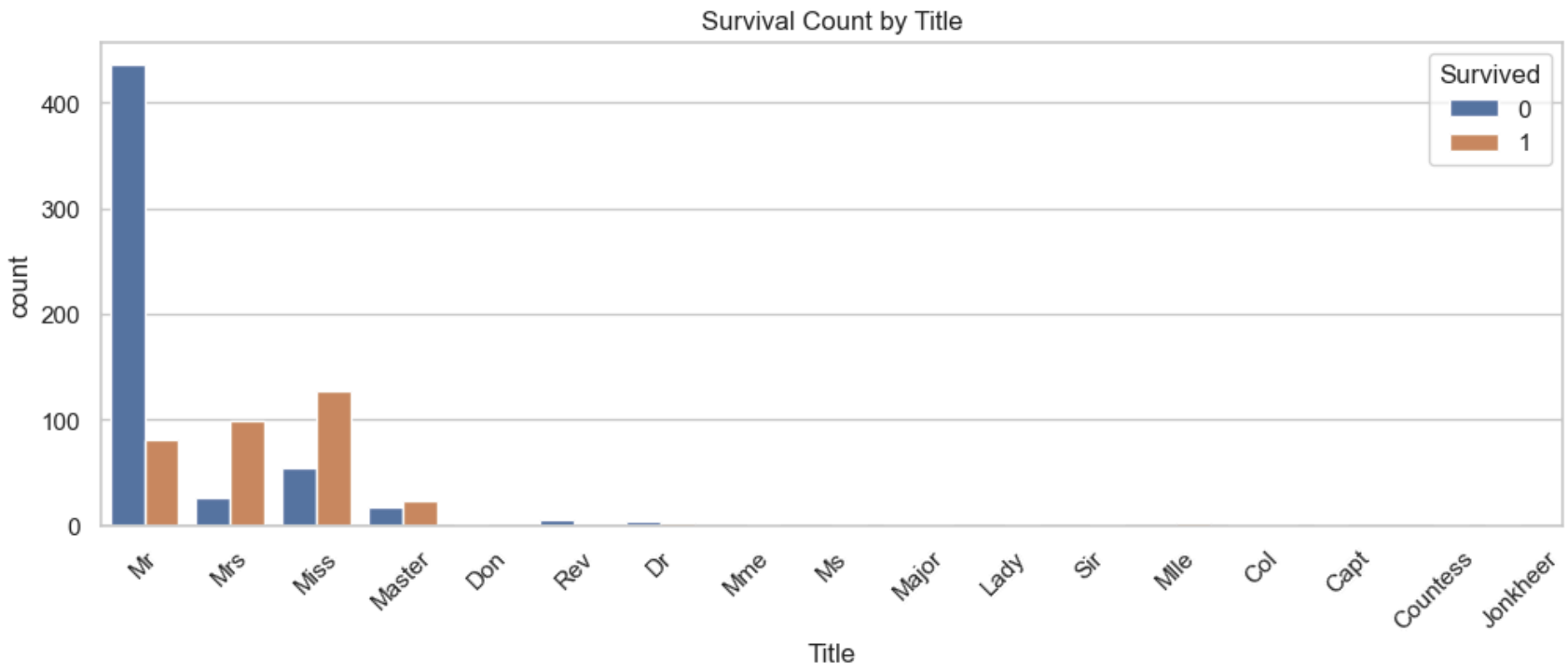

Survival Rate by Family Size
Survival Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
1
2
3
4
5
6
7
8
11
FamilySize


Correlation Matrix
PassengerId 1 -0.005 -0.035 0.037 -0.058 -0.0017 0.013 -0.04
Survived -0.005 1 -0.34 -0.077 -0.035 0.082 0.26 0.017
Pclass -0.035 -0.34 1 -0.37 0.083 0.018 -0.55 0.066
Age 0.037 -0.077 -0.37 1 -0.31 -0.19 0.096 -0.3
SibSp -0.058 -0.035 0.083 -0.31 1 0.41 0.16 0.89
Parch -0.0017 0.082 0.018 -0.19 0.41 1 0.22 0.78
Fare 0.013 0.26 -0.55 0.096 0.16 0.22 1 0.22
FamilySize -0.04 0.017 0.066 -0.3 0.89 0.78 0.22 1
PassengerId Survived Pclass Age SibSp Parch Fare FamilySize
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


Age Boxplot for Outlier Detection
0
10
20
30
40
50
60
70
80
Age


Fare Boxplot for Outlier Detection
0
100
200
300
400
500
Fare

```
=== Hypothesis Testing ===
Chi-square test for Pclass vs Survival: p-value = 0.0000

Female survival rate: 74.20%
Male survival rate: 18.89%

T-test for age difference: p-value = 0.0391

=== Title Analysis ===
Sex       female  male
Title
Capt           0     1
Col            0     2
Countess       1     0
Don            0     1
Dr             1     6
Jonkheer       0     1
Lady           1     0
Major          0     2
Master         0    40
Miss         182     0
Mlle           2     0
Mme            1     0
Mr             0   517
Mrs          125     0
Ms             1     0
Rev            0     6
Sir            0     1
```


Survival Count by Title
Survived
0
1
count
400
300
200
100
0
Mr
Mrs
Miss
Master
Don
Rev
Dr
Mme
Ms
Major
Lady
Sir
Mlle
Col
Capt
Countess
Jonkheer
Title