In [10]:
```
# Titanic Dataset EDA Notebook

# 1. Import Libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Settings for better plots
sns.set(style="whitegrid")
plt.rcParams['figure.figsize'] = (10,6)
```

In [11]:
```
# 2. Load Dataset
train_df = pd.read_csv('train.csv')
test_df = pd.read_csv('test.csv')
```

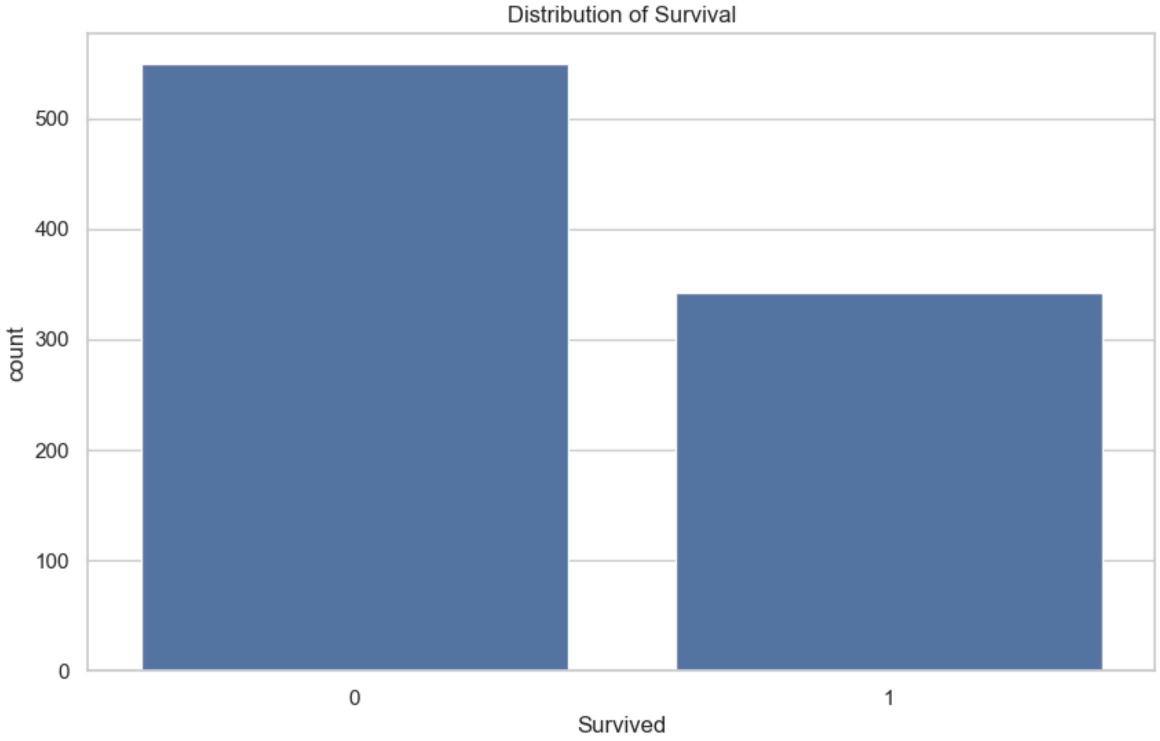
In [18]:
```
# 3. Data Overview
print("\n--- Data Info ---")
print(train_df.info())

print("\n--- Data Description ---")
print(train_df.describe())

print("\n--- Missing Values ---")
print(train_df.isnull().sum())
```

```
--- Data Info ---
<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

--- Data Description ---
       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

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

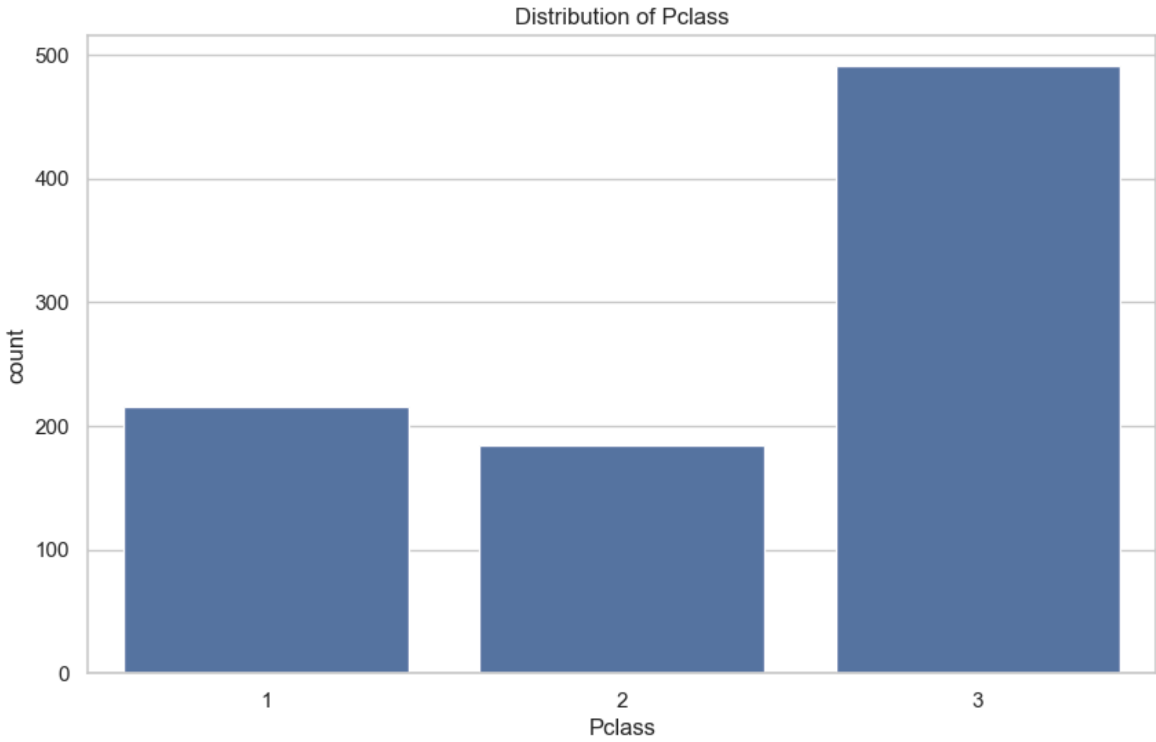
In [12]:
```
# 4. Univariate Analysis
print("\n--- Target Variable: Survival ---")
```

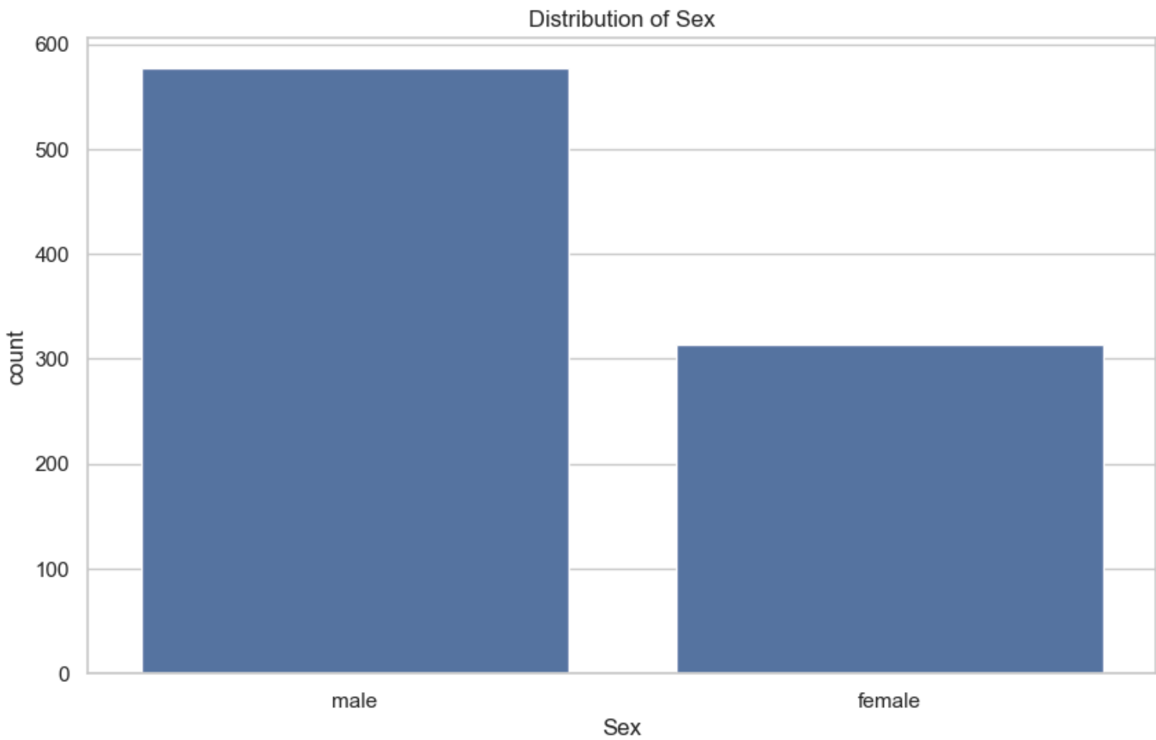
```
sns.countplot(x='Survived', data=train_df)
plt.title('Distribution of Survival')
plt.show()
# Observation:
# - Around 38% of passengers survived, while 62% did not survive.

print("\n--- Categorical Features ---")
categorical_cols = ['Pclass', 'Sex', 'Embarked']
for col in categorical_cols:
    sns.countplot(x=col, data=train_df)
    plt.title(f'Distribution of {col}')
    plt.show()
    # Observation example:
    # - For Pclass: Most passengers belonged to 3rd class.
    # - For Sex: More males were onboard than females.
    # - For Embarked: Most passengers embarked from Southampton.

print("\n--- Numerical Features ---")
train_df[['Age', 'Fare']].hist(bins=30, figsize=(12,6))
plt.suptitle('Histograms of Age and Fare')
plt.show()
# Observation:
# - Most passengers were aged between 20-40.
# - Most fares were low, with a few very expensive tickets.

print("\n--- Boxplots to Detect Outliers ---")
for col in ['Age', 'Fare']:
    sns.boxplot(x=train_df[col])
    plt.title(f'Boxplot of {col}')
    plt.show()
    # Observation:
    # - Fare has several high-value outliers.
    # - Age distribution is relatively normal but with few extreme ages.
```

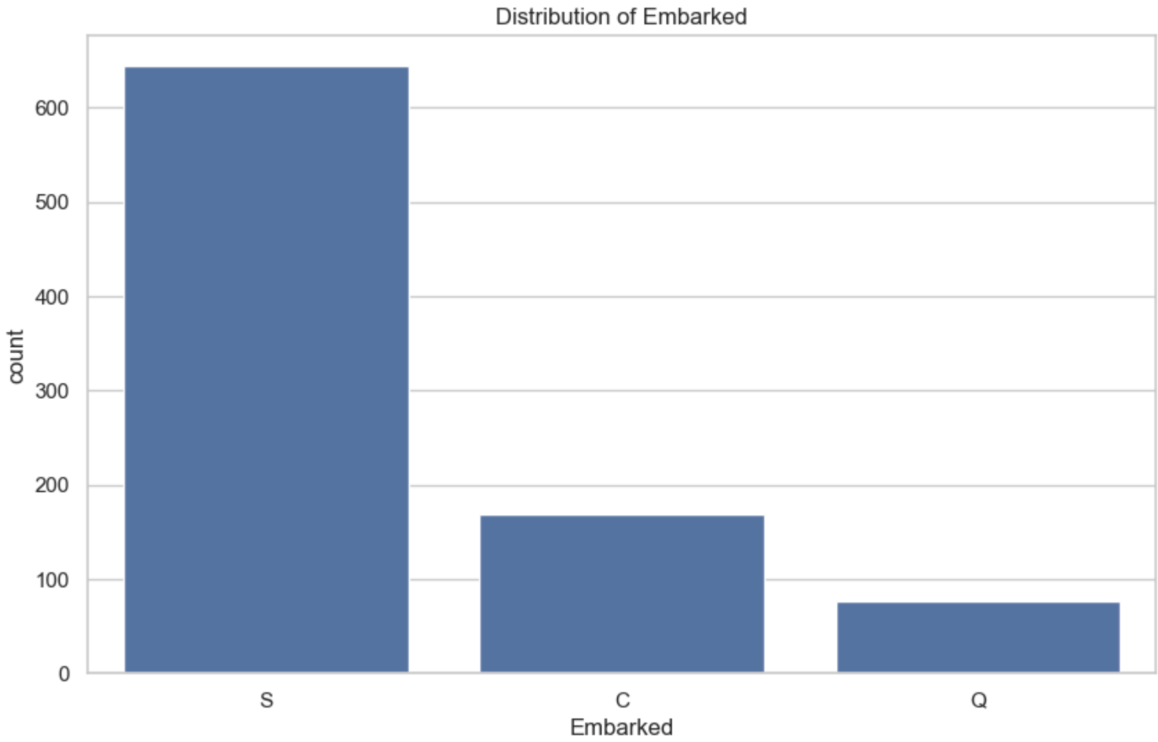
```
--- Target Variable: Survival ---
```



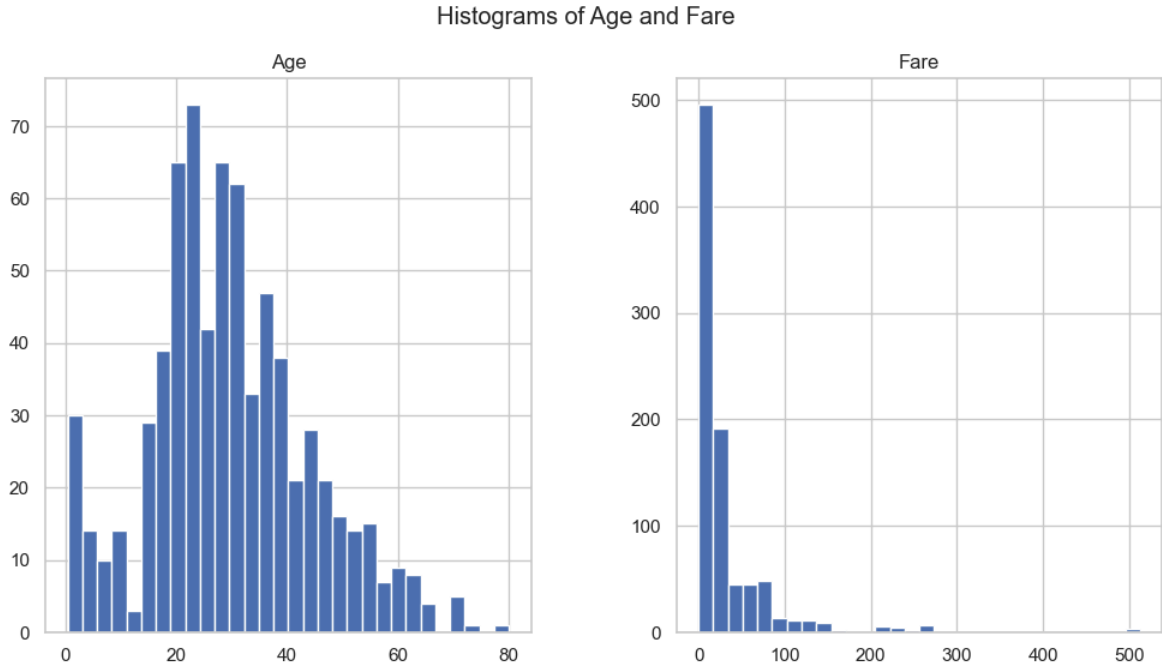

```
--- Categorical Features ---
```


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

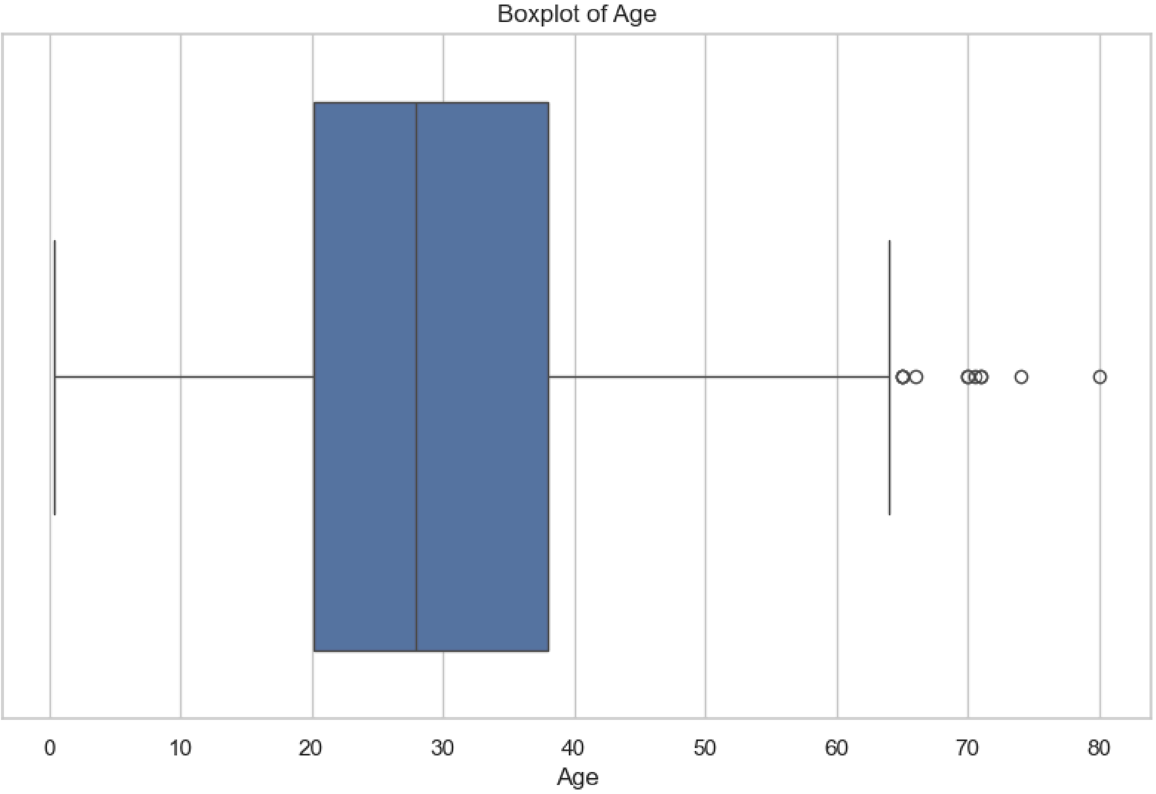

Distribution of Sex
count
0
100
200
300
400
500
600
male
female
Sex

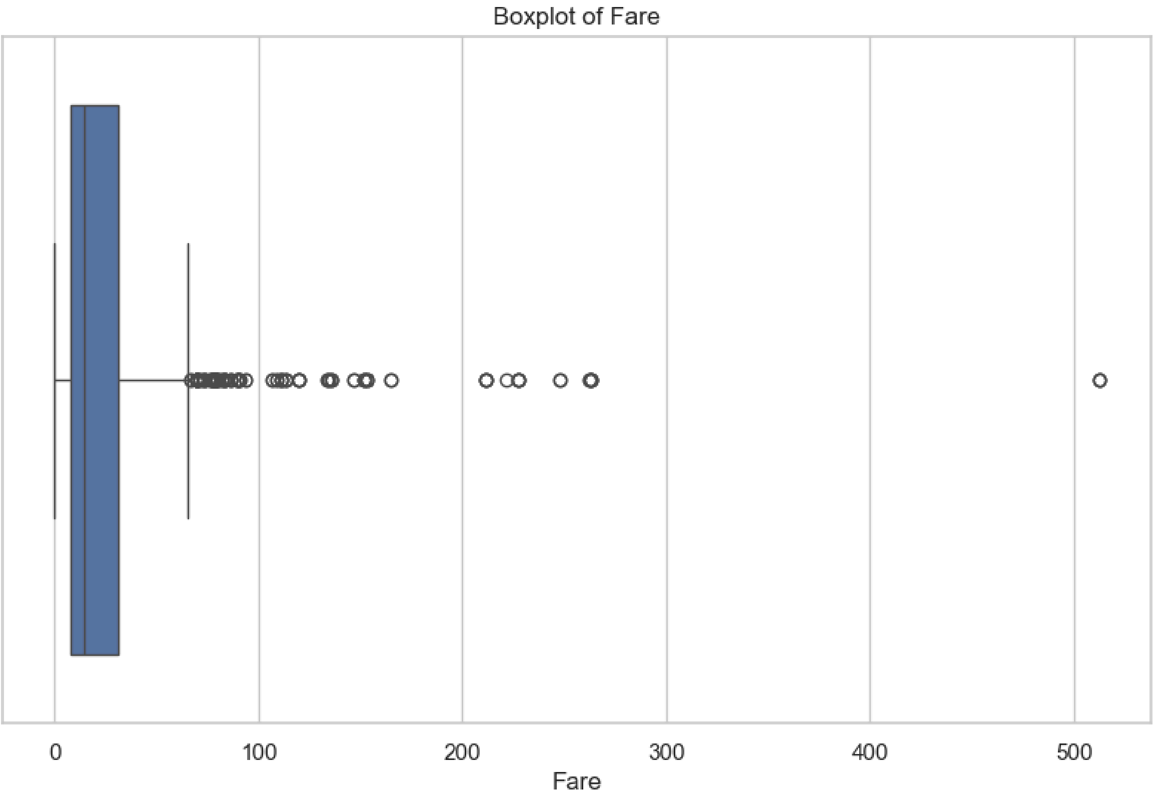



```
--- Numerical Features ---
```




```
--- Boxplots to Detect Outliers ---
```

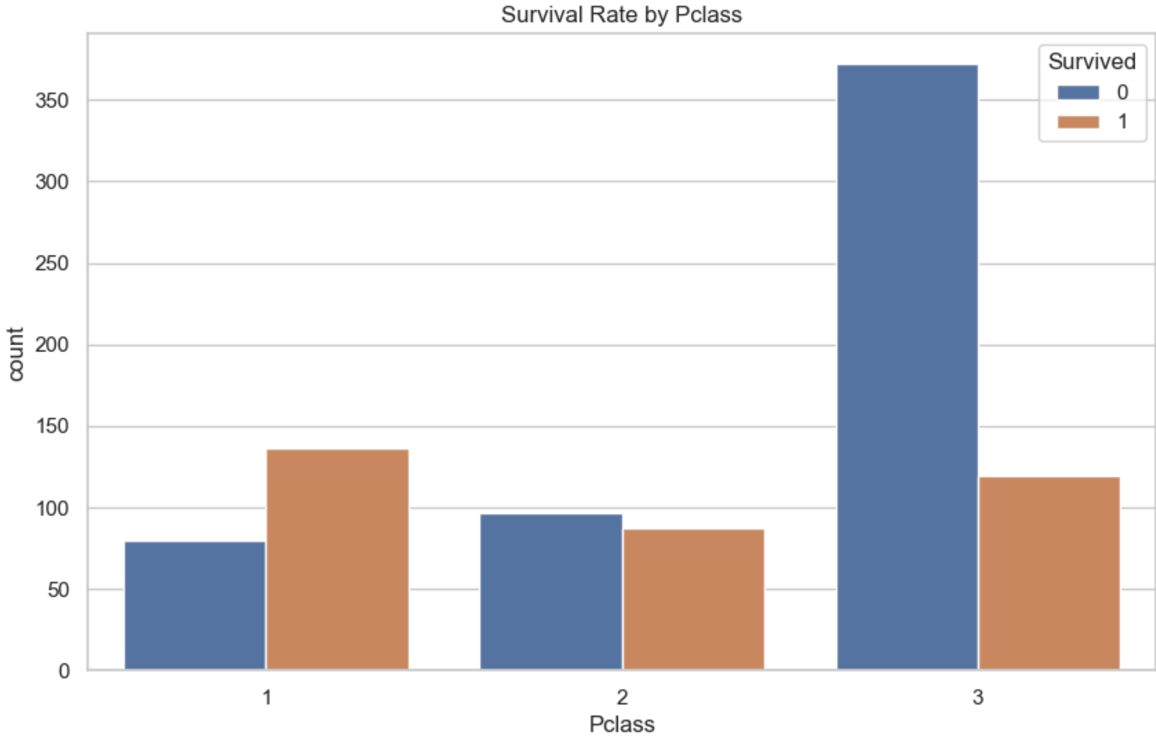


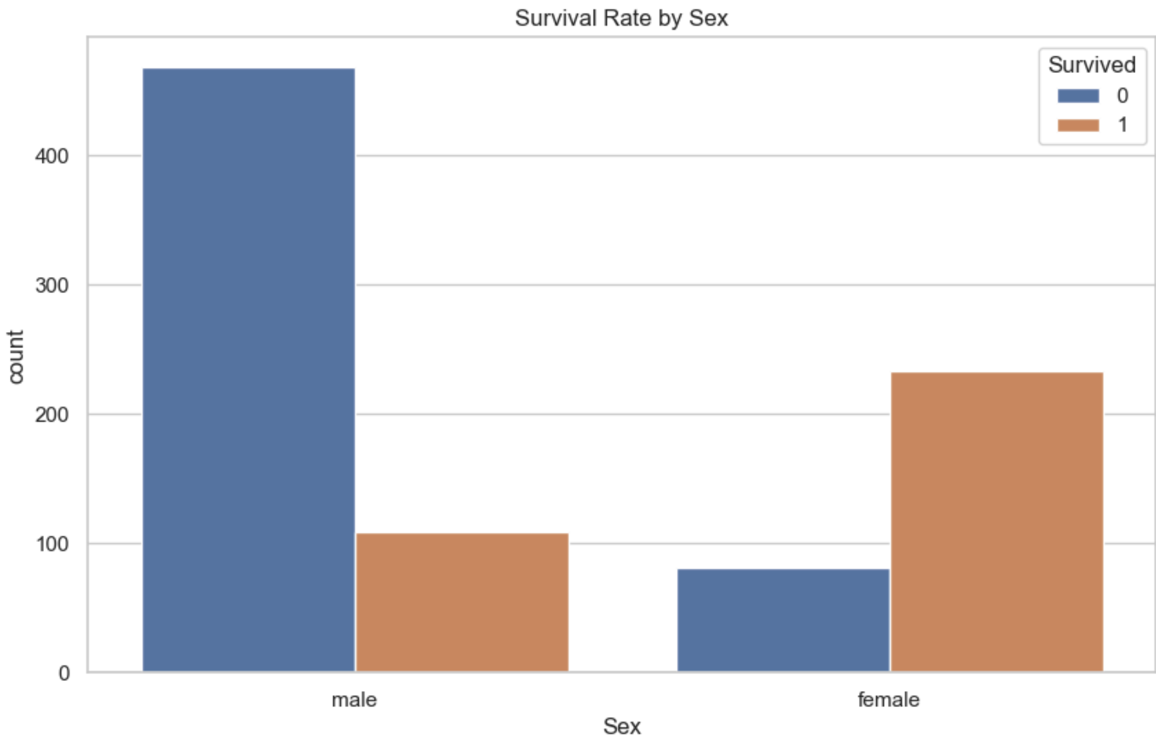


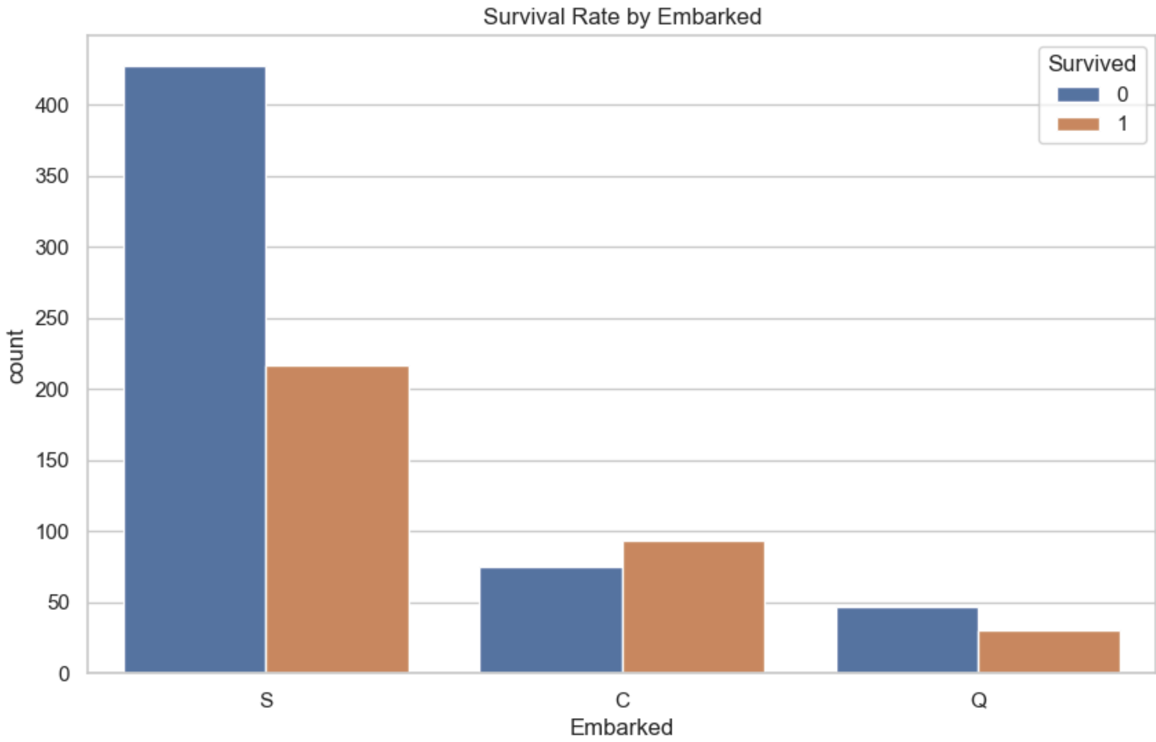
In [13]:
```
# 5. Bivariate Analysis
print("\n--- Survival vs Categorical Features ---")
for col in categorical_cols:
    sns.countplot(x=col, hue='Survived', data=train_df)
    plt.title(f'Survival Rate by {col}')
    plt.show()
    # Observation:
    # - Females had a much higher survival rate than males.
    # - 1st class passengers had better survival chances.

print("\n--- Survival vs Numerical Features ---")
```

```
sns.histplot(data=train_df, x='Age', hue='Survived', multiple='stack')
plt.title('Age Distribution by Survival')
plt.show()
# Observation:
# - Young children had higher survival rates.

sns.histplot(data=train_df, x='Fare', hue='Survived', multiple='stack')
plt.title('Fare Distribution by Survival')
plt.show()
# Observation:
# - Higher fare-paying passengers had higher survival rates.
```

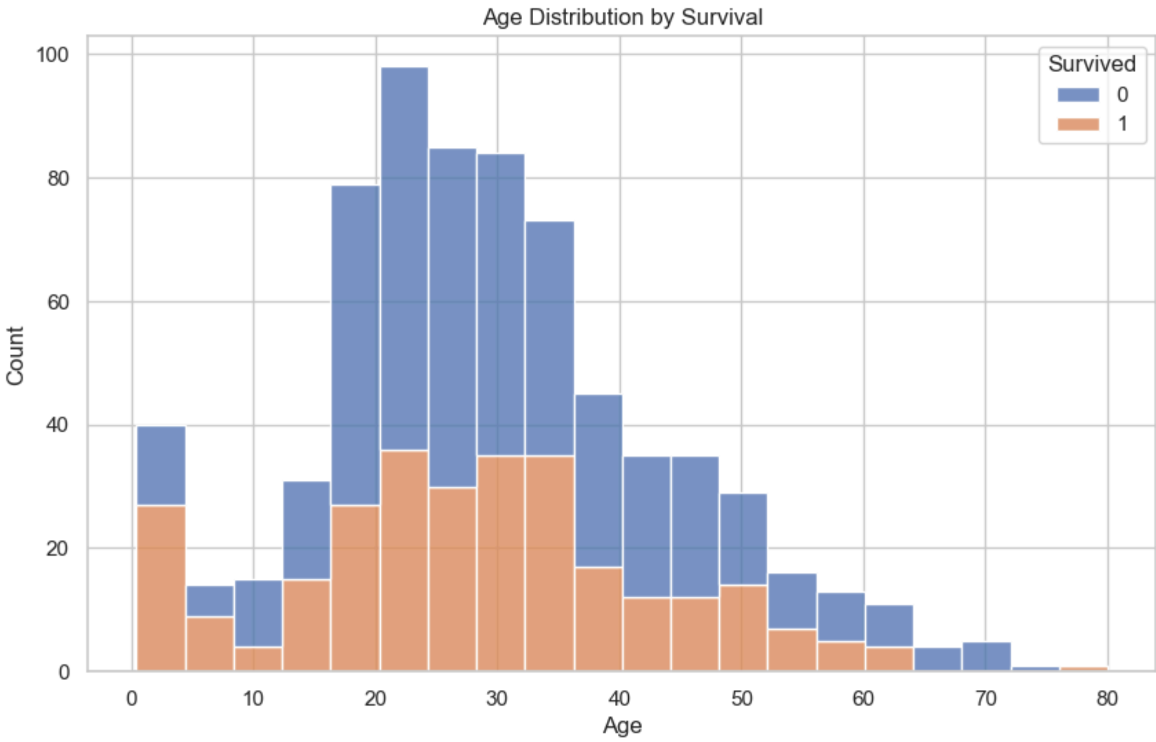
```
--- Survival vs Categorical Features ---
```

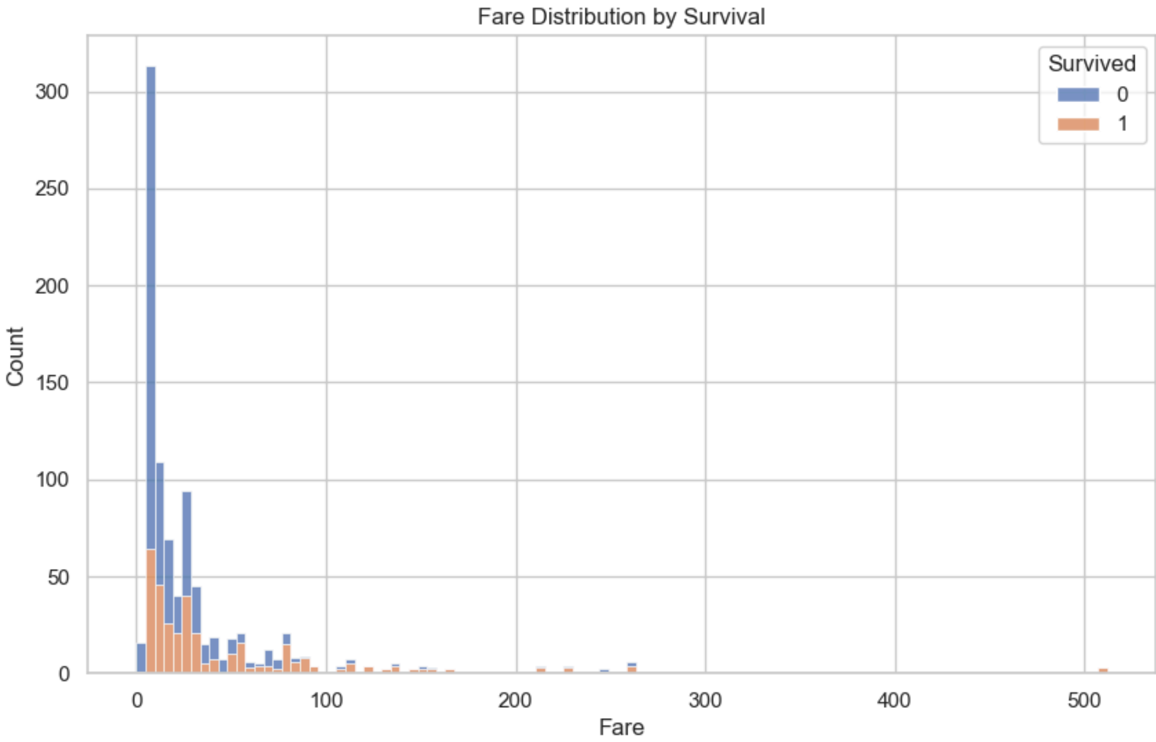






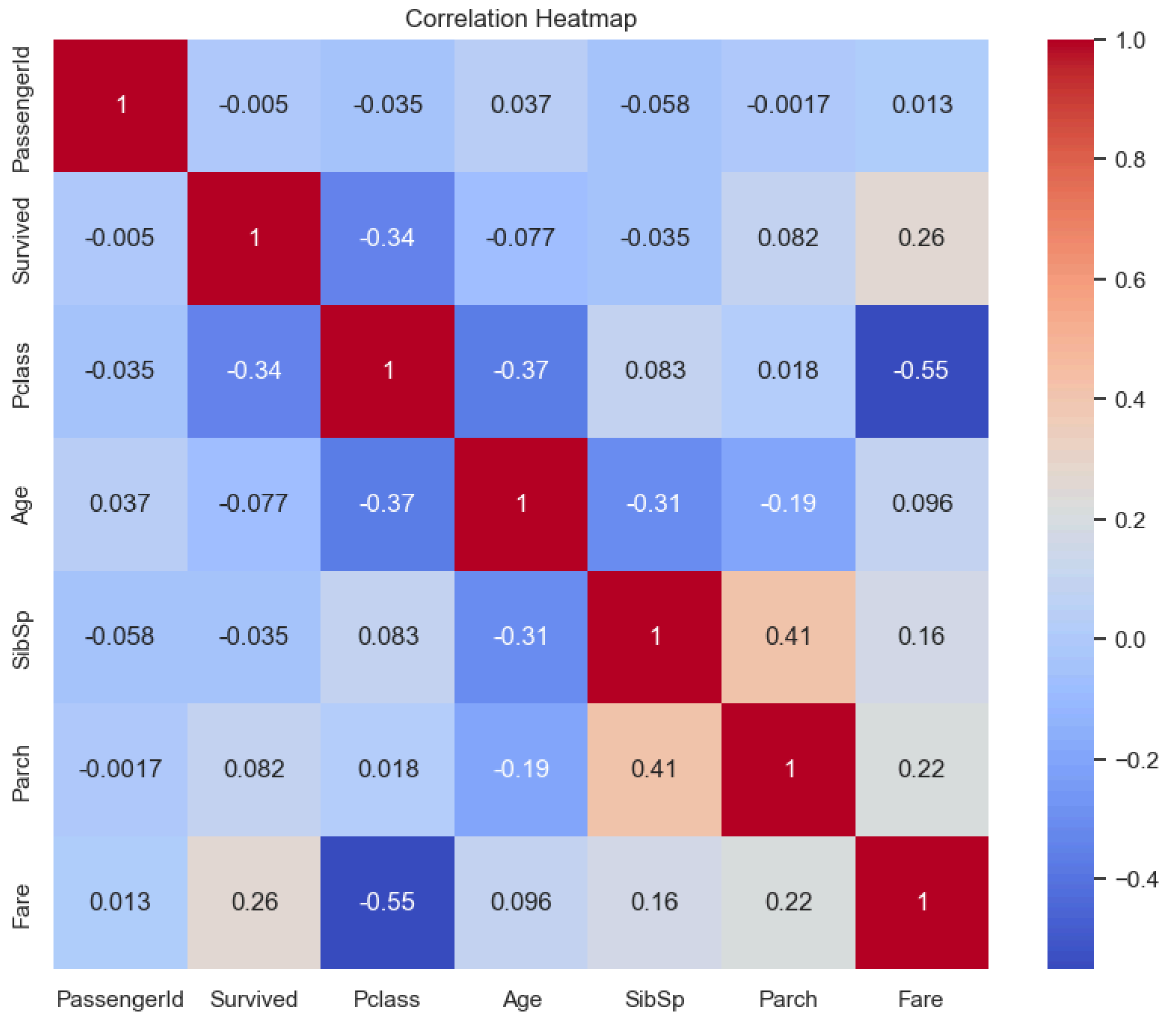

```
--- Survival vs Numerical Features ---
```



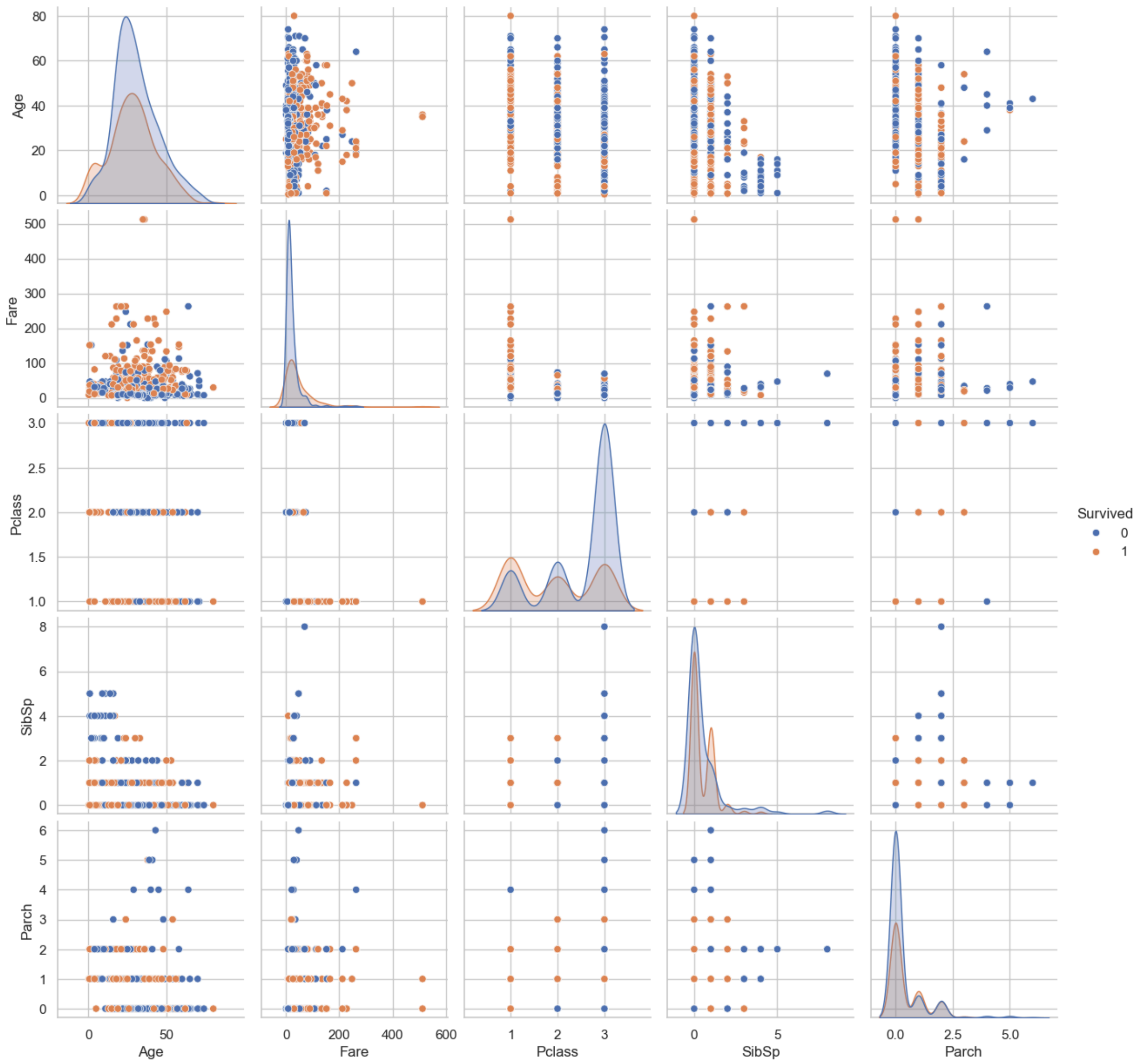


In [14]:
```
# 6. Multivariate Analysis
print("\n--- Correlation Heatmap ---")
plt.figure(figsize=(10,8))
numeric_cols = train_df.select_dtypes(include=['int64', 'float64'])
sns.heatmap(numeric_cols.corr(), annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
# Observation:
# - Strong correlation between Fare and Pclass.
# - Sex and Survival are correlated.

print("\n--- Pairplot of Selected Features ---")
sns.pairplot(train_df, vars=['Age', 'Fare', 'Pclass', 'SibSp', 'Parch'], hue='Su
plt.show()
# Observation:
# - Clear patterns between Fare, Age, and Survival
```

```
--- Correlation Heatmap ---
```




```
--- Pairplot of Selected Features ---
```



In [15]:
```
# 7. Insights and Findings
# Example Insights:
# - Females survived at a much higher rate than males.
# - 1st Class passengers had a higher survival rate.
# - Higher fare was correlated with higher survival.
# - Young children had better survival odds.
```

In [16]:
```
# 8. Conclusion
# - Key factors influencing survival: Sex, Pclass, Fare, Age.
# - Recommend feature engineering and modeling as next steps.
```

In [17]:
```
# Summary of Findings
# - Females and 1st class passengers had the highest survival chances.
# - Higher fare-paying passengers were more likely to survive.
# - Young children (<10 years) had higher survival rates.
# - 3rd class males had the lowest survival rates.
# - Passengers from Cherbourg had better survival rates.
# - Important features: Sex, Pclass, Fare, Age.
```