In [52]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
```

In [53]:
```
table1 = pd.read_csv(r"C:\Users\Joshwa\Downloads\test.csv")
table1
```

Out[53]:

| | PassengerId | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 892 | 3 | Kelly, Mr. James | male | 34.50 | 0 | 0 | 330911 | 7.83 | NaN | Q |
| **1** | 893 | 3 | Wilkes, Mrs. James (Ellen Needs) | female | 47.00 | 1 | 0 | 363272 | 7.00 | NaN | S |
| **2** | 894 | 2 | Myles, Mr. Thomas Francis | male | 62.00 | 0 | 0 | 240276 | 9.69 | NaN | Q |
| **3** | 895 | 3 | Wirz, Mr. Albert | male | 27.00 | 0 | 0 | 315154 | 8.66 | NaN | S |
| **4** | 896 | 3 | Hirvonen, Mrs. Alexander (Helga E Lindqvist) | female | 22.00 | 1 | 1 | 3101298 | 12.29 | NaN | S |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **413** | 1305 | 3 | Spector, Mr. Woolf | male | NaN | 0 | 0 | A.5. 3236 | 8.05 | NaN | S |
| **414** | 1306 | 1 | Oliva y Ocana, Dona. Fermina | female | 39.00 | 0 | 0 | PC 17758 | 108.90 | C105 | C |
| **415** | 1307 | 3 | Saether, Mr. Simon Sivertsen | male | 38.50 | 0 | 0 | SOTON/O.Q. 3101262 | 7.25 | NaN | S |
| **416** | 1308 | 3 | Ware, Mr. Frederick | male | NaN | 0 | 0 | 359309 | 8.05 | NaN | S |
| **417** | 1309 | 3 | Peter, Master. Michael J | male | NaN | 1 | 1 | 2668 | 22.36 | NaN | C |

418 rows × 11 columns

In [54]:
```
table2 = pd.read_csv(r"C:\Users\Joshwa\Downloads\gender_submission.csv")
table2
```

Out[54]:

| | PassengerId | Survived |
|---|---|---|
| **0** | 892 | 0 |
| **1** | 893 | 1 |
| **2** | 894 | 0 |
| **3** | 895 | 0 |
| **4** | 896 | 1 |
| **...** | ... | ... |
| **413** | 1305 | 0 |
| **414** | 1306 | 1 |
| **415** | 1307 | 0 |
| **416** | 1308 | 0 |
| **417** | 1309 | 0 |

418 rows × 2 columns

In [55]:
```
table3 = pd.read_csv(r"C:\Users\Joshwa\Downloads\train.csv")
table3
```

Out[55]:

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 0 | 3 | Braund, Mr. Owen Harris | male | 22.00 | 1 | 0 | A/5 21171 | 7.25 | NaN | S |
| **1** | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th... | female | 38.00 | 1 | 0 | PC 17599 | 71.28 | C85 | C |
| **2** | 3 | 1 | 3 | Heikkinen, Miss. Laina | female | 26.00 | 0 | 0 | STON/O2. 3101282 | 7.92 | NaN | S |
| **3** | 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35.00 | 1 | 0 | 113803 | 53.10 | C123 | S |
| **4** | 5 | 0 | 3 | Allen, Mr. William Henry | male | 35.00 | 0 | 0 | 373450 | 8.05 | NaN | S |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **886** | 887 | 0 | 2 | Montvila, Rev. Juozas | male | 27.00 | 0 | 0 | 211536 | 13.00 | NaN | S |
| **887** | 888 | 1 | 1 | Graham, Miss. Margaret Edith | female | 19.00 | 0 | 0 | 112053 | 30.00 | B42 | S |
| **888** | 889 | 0 | 3 | Johnston, Miss. Catherine Helen "Carrie" | female | NaN | 1 | 2 | W./C. 6607 | 23.45 | NaN | S |
| **889** | 890 | 1 | 1 | Behr, Mr. Karl Howell | male | 26.00 | 0 | 0 | 111369 | 30.00 | C148 | C |
| **890** | 891 | 0 | 3 | Dooley, Mr. Patrick | male | 32.00 | 0 | 0 | 370376 | 7.75 | NaN | Q |

891 rows × 12 columns

In [56]:
```
table_1 = pd.merge(table1, table2, on='PassengerId', how='left')
table_1
```

Out[56]:

| | PassengerId | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked | Survived |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 892 | 3 | Kelly, Mr. James | male | 34.50 | 0 | 0 | 330911 | 7.83 | NaN | Q | 0 |
| **1** | 893 | 3 | Wilkes, Mrs. James (Ellen Needs) | female | 47.00 | 1 | 0 | 363272 | 7.00 | NaN | S | 1 |
| **2** | 894 | 2 | Myles, Mr. Thomas Francis | male | 62.00 | 0 | 0 | 240276 | 9.69 | NaN | Q | 0 |
| **3** | 895 | 3 | Wirz, Mr. Albert | male | 27.00 | 0 | 0 | 315154 | 8.66 | NaN | S | 0 |
| **4** | 896 | 3 | Hirvonen, Mrs. Alexander (Helga E Lindqvist) | female | 22.00 | 1 | 1 | 3101298 | 12.29 | NaN | S | 1 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **413** | 1305 | 3 | Spector, Mr. Woolf | male | NaN | 0 | 0 | A.5. 3236 | 8.05 | NaN | S | 0 |
| **414** | 1306 | 1 | Oliva y Ocana, Dona. Fermina | female | 39.00 | 0 | 0 | PC 17758 | 108.90 | C105 | C | 1 |
| **415** | 1307 | 3 | Saether, Mr. Simon Sivertsen | male | 38.50 | 0 | 0 | SOTON/O.Q. 3101262 | 7.25 | NaN | S | 0 |
| **416** | 1308 | 3 | Ware, Mr. Frederick | male | NaN | 0 | 0 | 359309 | 8.05 | NaN | S | 0 |
| **417** | 1309 | 3 | Peter, Master. Michael J | male | NaN | 1 | 1 | 2668 | 22.36 | NaN | C | 0 |

418 rows × 12 columns

In [110…

```
temp_pop = table_1.pop('Survived')
```

In [58]:

```
table_1.insert(1, 'Survived', temp_pop)
table_1
```

Out[58]:

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 892 | 0 | 3 | Kelly, Mr. James | male | 34.50 | 0 | 0 | 330911 | 7.83 | NaN | Q |
| **1** | 893 | 1 | 3 | Wilkes, Mrs. James (Ellen Needs) | female | 47.00 | 1 | 0 | 363272 | 7.00 | NaN | S |
| **2** | 894 | 0 | 2 | Myles, Mr. Thomas Francis | male | 62.00 | 0 | 0 | 240276 | 9.69 | NaN | Q |
| **3** | 895 | 0 | 3 | Wirz, Mr. Albert | male | 27.00 | 0 | 0 | 315154 | 8.66 | NaN | S |
| **4** | 896 | 1 | 3 | Hirvonen, Mrs. Alexander (Helga E Lindqvist) | female | 22.00 | 1 | 1 | 3101298 | 12.29 | NaN | S |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **413** | 1305 | 0 | 3 | Spector, Mr. Woolf | male | NaN | 0 | 0 | A.5. 3236 | 8.05 | NaN | S |
| **414** | 1306 | 1 | 1 | Oliva y Ocana, Dona. Fermina | female | 39.00 | 0 | 0 | PC 17758 | 108.90 | C105 | C |
| **415** | 1307 | 0 | 3 | Saether, Mr. Simon Sivertsen | male | 38.50 | 0 | 0 | SOTON/O.Q. 3101262 | 7.25 | NaN | S |
| **416** | 1308 | 0 | 3 | Ware, Mr. Frederick | male | NaN | 0 | 0 | 359309 | 8.05 | NaN | S |
| **417** | 1309 | 0 | 3 | Peter, Master. Michael J | male | NaN | 1 | 1 | 2668 | 22.36 | NaN | C |

418 rows × 12 columns

In [59]:
```
df = pd.concat([table3, table_1], ignore_index = True)
df
```

Out[59]:

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 0 | 3 | Braund, Mr. Owen Harris | male | 22.00 | 1 | 0 | A/5 21171 | 7.25 | NaN | S |
| **1** | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th... | female | 38.00 | 1 | 0 | PC 17599 | 71.28 | C85 | C |
| **2** | 3 | 1 | 3 | Heikkinen, Miss. Laina | female | 26.00 | 0 | 0 | STON/O2. 3101282 | 7.92 | NaN | S |
| **3** | 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35.00 | 1 | 0 | 113803 | 53.10 | C123 | S |
| **4** | 5 | 0 | 3 | Allen, Mr. William Henry | male | 35.00 | 0 | 0 | 373450 | 8.05 | NaN | S |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **1304** | 1305 | 0 | 3 | Spector, Mr. Woolf | male | NaN | 0 | 0 | A.5. 3236 | 8.05 | NaN | S |
| **1305** | 1306 | 1 | 1 | Oliva y Ocana, Dona. Fermina | female | 39.00 | 0 | 0 | PC 17758 | 108.90 | C105 | C |
| **1306** | 1307 | 0 | 3 | Saether, Mr. Simon Sivertsen | male | 38.50 | 0 | 0 | SOTON/O.Q. 3101262 | 7.25 | NaN | S |
| **1307** | 1308 | 0 | 3 | Ware, Mr. Frederick | male | NaN | 0 | 0 | 359309 | 8.05 | NaN | S |
| **1308** | 1309 | 0 | 3 | Peter, Master. Michael J | male | NaN | 1 | 1 | 2668 | 22.36 | NaN | C |

1309 rows × 12 columns

In [60]:
```
pd.set_option('display.float_format', lambda x: '%.2f' % x)
```

In [61]:
```
df.describe()
```

Out[61]:

| | PassengerId | Survived | Pclass | Age | SibSp | Parch | Fare |
|---|---|---|---|---|---|---|---|
| **count** | 1309.00 | 1309.00 | 1309.00 | 1046.00 | 1309.00 | 1309.00 | 1308.00 |
| **mean** | 655.00 | 0.38 | 2.29 | 29.88 | 0.50 | 0.39 | 33.30 |
| **std** | 378.02 | 0.48 | 0.84 | 14.41 | 1.04 | 0.87 | 51.76 |
| **min** | 1.00 | 0.00 | 1.00 | 0.17 | 0.00 | 0.00 | 0.00 |
| **25%** | 328.00 | 0.00 | 2.00 | 21.00 | 0.00 | 0.00 | 7.90 |
| **50%** | 655.00 | 0.00 | 3.00 | 28.00 | 0.00 | 0.00 | 14.45 |
| **75%** | 982.00 | 1.00 | 3.00 | 39.00 | 1.00 | 0.00 | 31.27 |
| **max** | 1309.00 | 1.00 | 3.00 | 80.00 | 8.00 | 9.00 | 512.33 |

In [108…

```
df.head(10)
```

Out[108...

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Embarked |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 0 | 3 | Braund, Mr. Owen Harris | male | 22.00 | 1 | 0 | A/5 21171 | 7.25 | S |
| **1** | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th... | female | 38.00 | 1 | 0 | PC 17599 | 71.28 | C |
| **2** | 3 | 1 | 3 | Heikkinen, Miss. Laina | female | 26.00 | 0 | 0 | STON/O2. 3101282 | 7.92 | S |
| **3** | 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35.00 | 1 | 0 | 113803 | 53.10 | S |
| **4** | 5 | 0 | 3 | Allen, Mr. William Henry | male | 35.00 | 0 | 0 | 373450 | 8.05 | S |
| **5** | 6 | 0 | 3 | Moran, Mr. James | male | 28.00 | 0 | 0 | 330877 | 8.46 | Q |
| **6** | 7 | 0 | 1 | McCarthy, Mr. Timothy J | male | 54.00 | 0 | 0 | 17463 | 51.86 | S |
| **7** | 8 | 0 | 3 | Palsson, Master. Gosta Leonard | male | 2.00 | 3 | 1 | 349909 | 21.07 | S |
| **8** | 9 | 1 | 3 | Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg) | female | 27.00 | 0 | 2 | 347742 | 11.13 | S |
| **9** | 10 | 1 | 2 | Nasser, Mrs. Nicholas (Adele Achem) | female | 14.00 | 1 | 0 | 237736 | 30.07 | C |

In [63]:
```
df.isnull().sum()
```

Out[63]:
```
PassengerId       0
Survived          0
Pclass            0
Name              0
Sex               0
Age             263
SibSp             0
Parch             0
Ticket            0
Fare              1
Cabin          1014
Embarked          2
dtype: int64
```

In [ ]:
```
df['Age'].fillna(df['Age'].median())
df['Embarked'].fillna(df['Embarked'].mode()[0])
df['Fare'].fillna(df['Fare'].median())
df.drop(columns=['Cabin'])
```

In [65]:
```
df.isnull().sum()
```

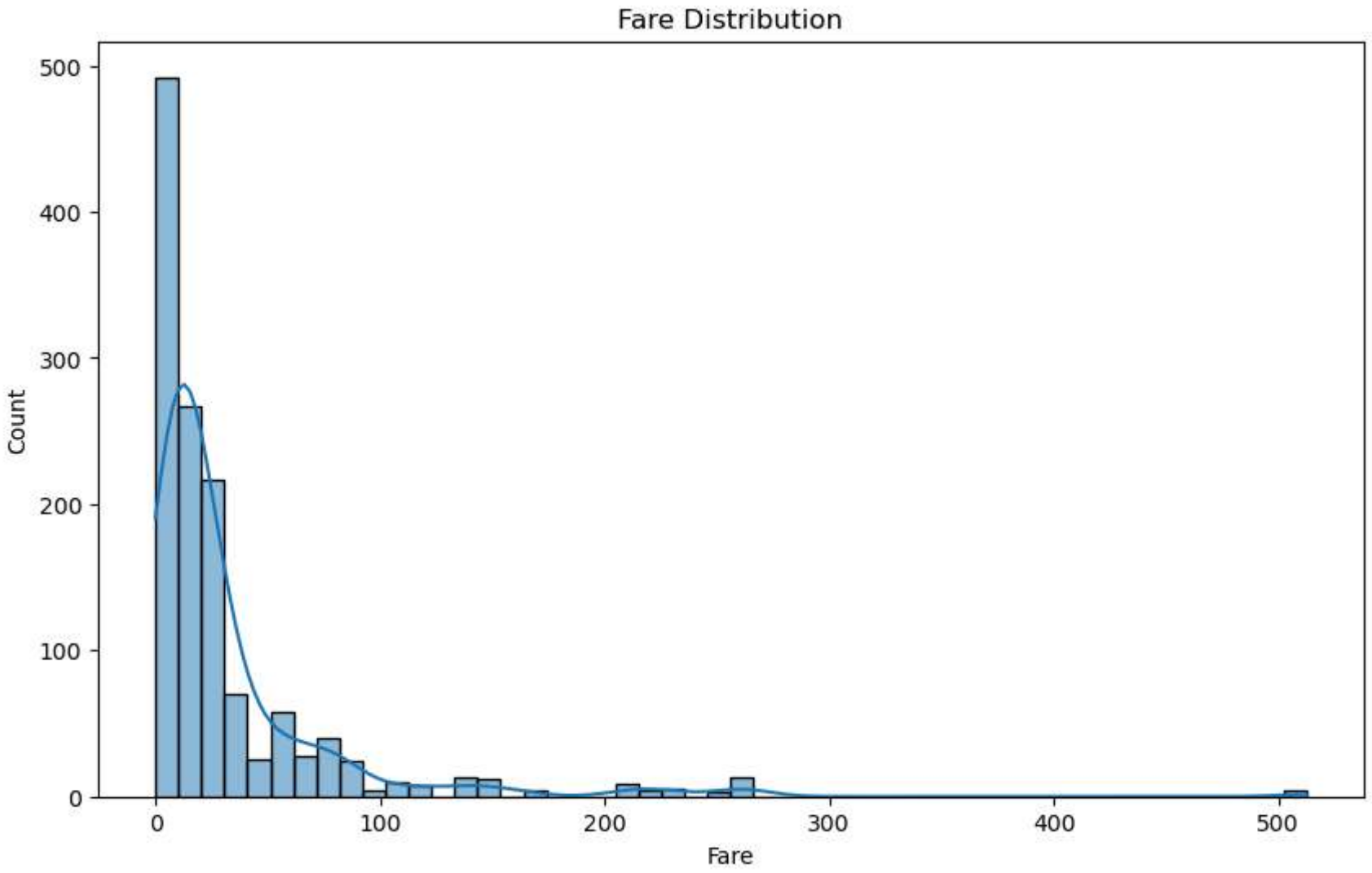
Out[65]:
```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Embarked       0
dtype: int64
```

In [66]:
```
df.nunique()
```

Out[66]:
```
PassengerId    1309
Survived          2
Pclass            3
Name           1307
Sex               2
Age              98
SibSp             7
Parch             8
Ticket          929
Fare            281
Embarked          3
dtype: int64
```

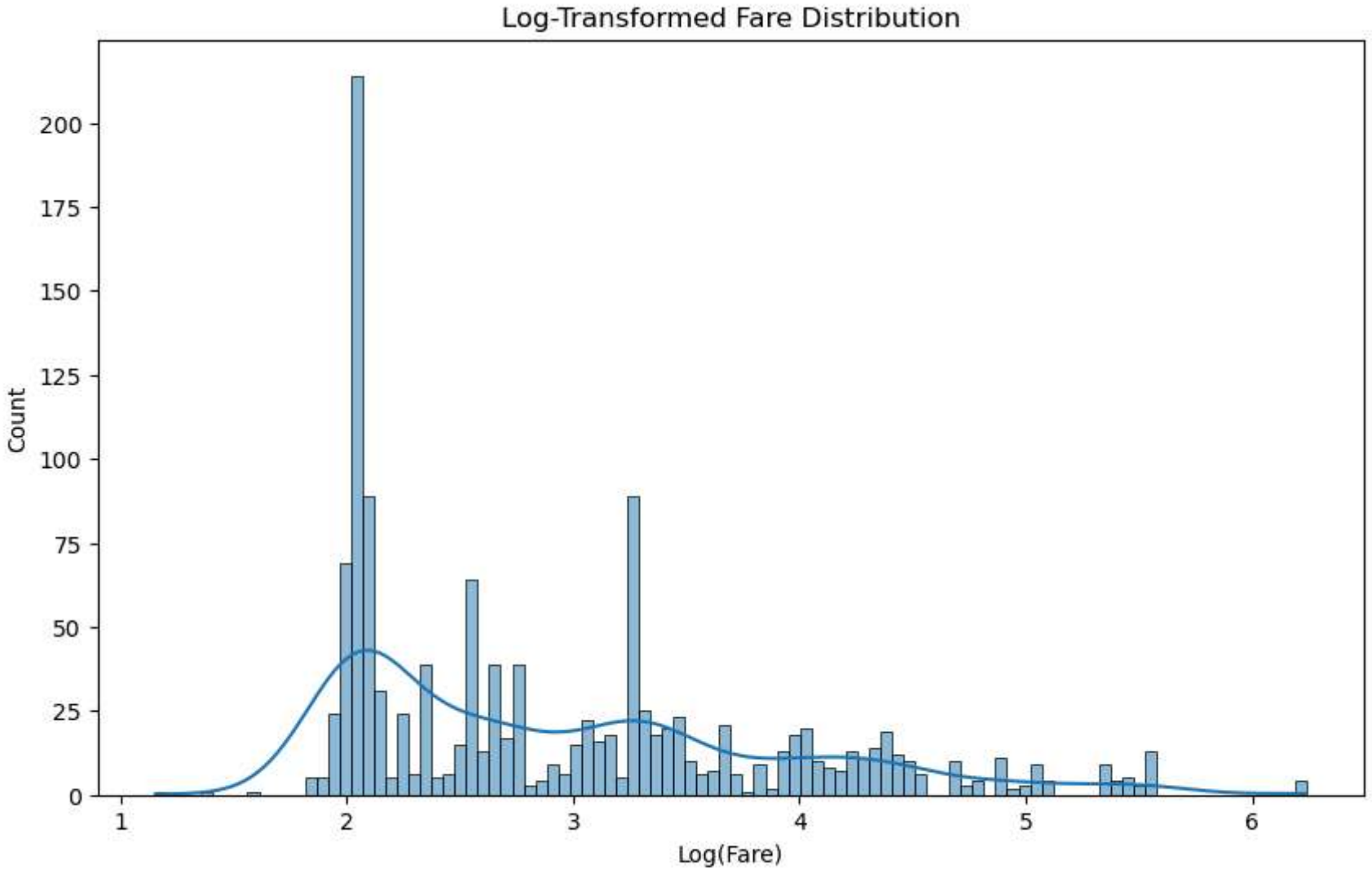
In [67]:
```
plt.figure(figsize=(10, 6))
sns.histplot(df['Fare'], kde=True, bins=50)
plt.title('Fare Distribution')
plt.xlabel('Fare')
```

```
plt.ylabel('Count')
plt.show()
```



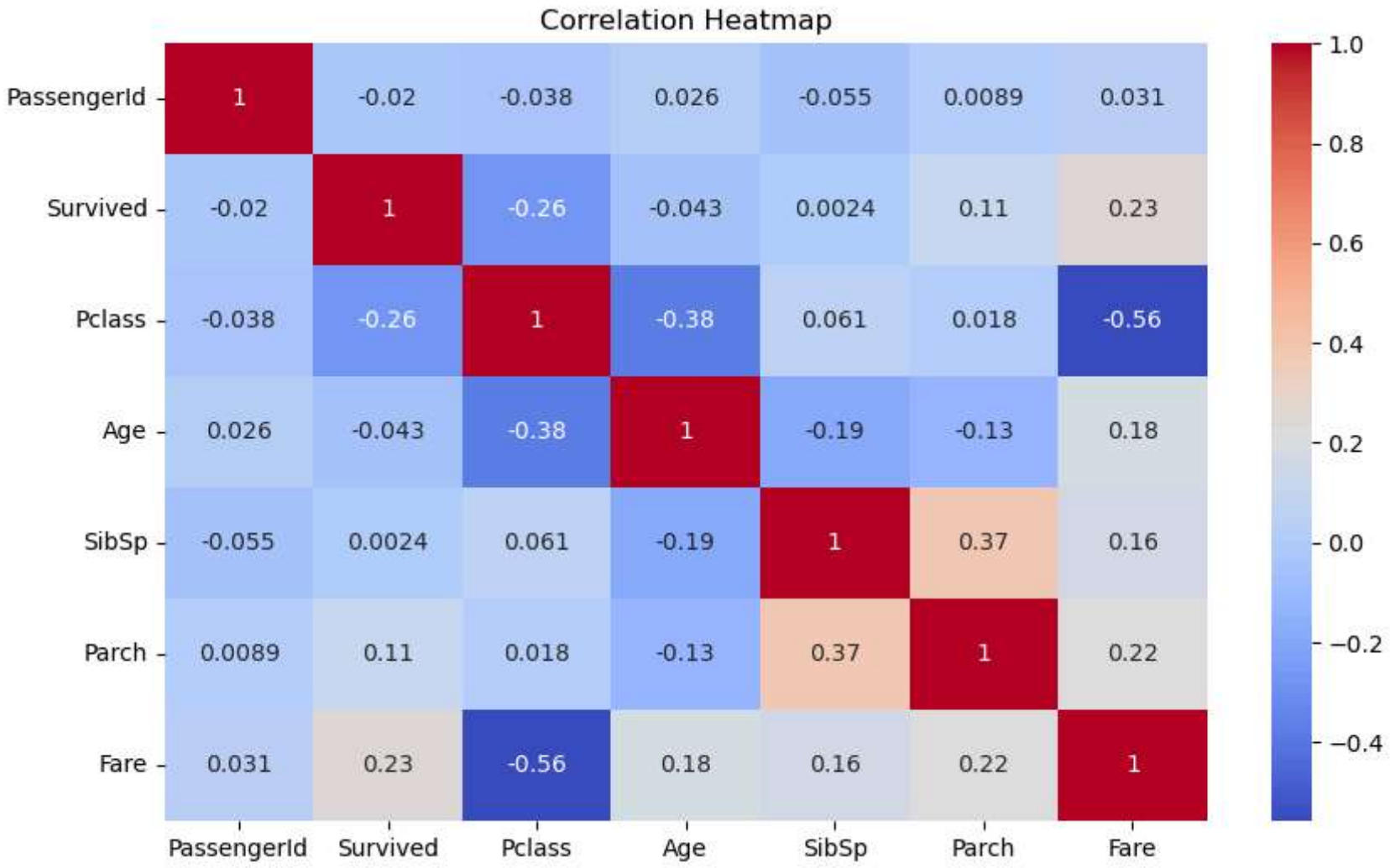

In [68]:
```
df_log = df[df['Fare'] > 0].copy()
df_log['LogFare'] = np.log(df_log['Fare'])

plt.figure(figsize=(10, 6))
sns.histplot(df_log['LogFare'], kde=True, bins=100)
```

```
plt.title('Log-Transformed Fare Distribution')
plt.xlabel('Log(Fare)')
plt.ylabel('Count')
plt.show()
```




In [69]:
```
corr_matrix = df.select_dtypes(include=['float64', 'int64']).corr()
plt.figure(figsize=(10,6))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
```

```
plt.title('Correlation Heatmap')
plt.show()
```

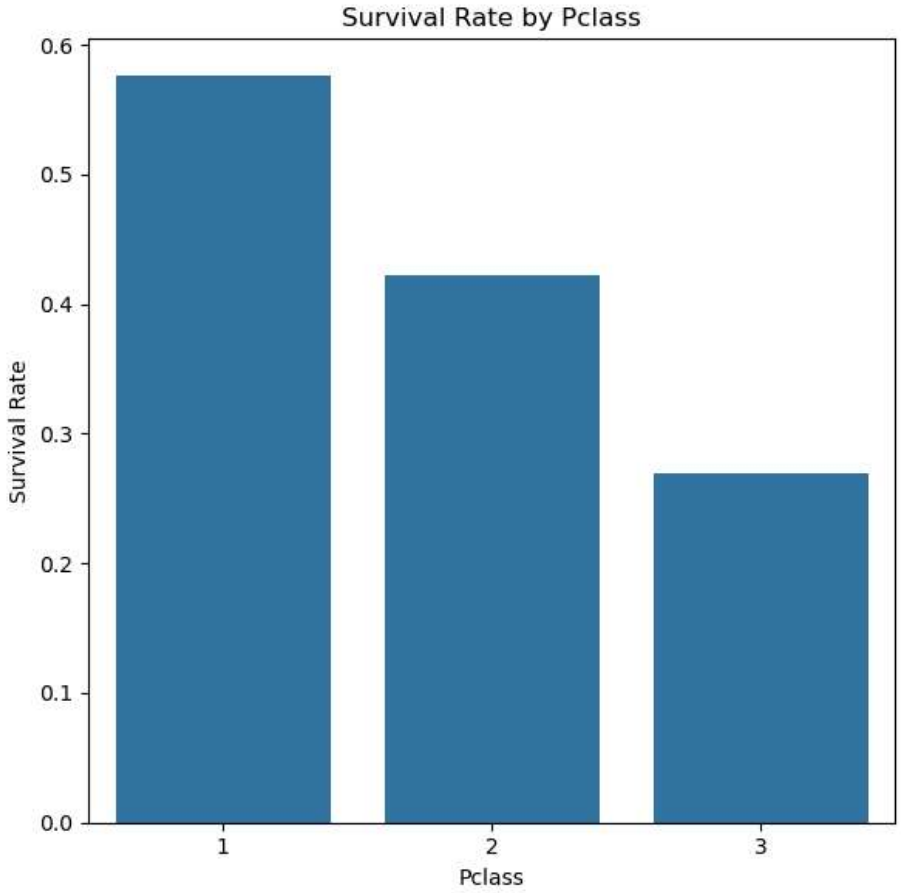


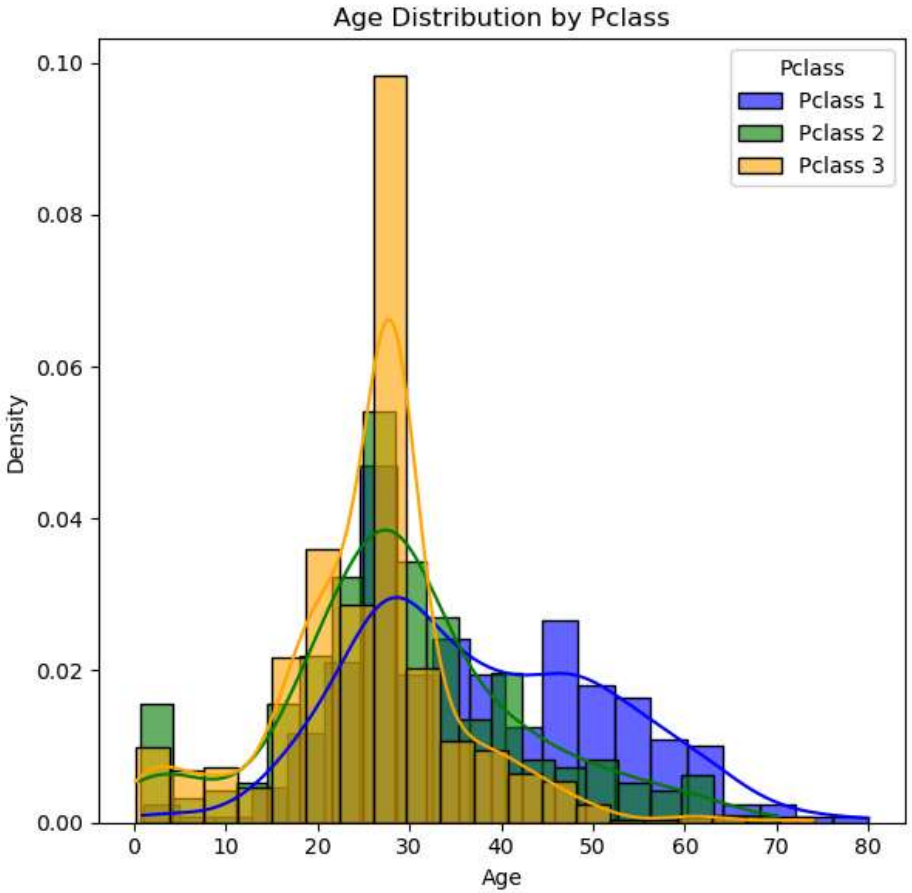

In [70]:
```
# Group by Pclass, and calculate survival rate
survival_by_pclass = df.groupby('Pclass')['Survived'].mean()

# Age range (min, max) by Pclass
age_range_by_pclass = df.groupby('Pclass')['Age'].agg(['min', 'max'])
```

```
# Combine both into a single DataFrame
result = pd.concat([survival_by_pclass, age_range_by_pclass], axis=1)
result.columns = ['Survival Rate', 'Age Min', 'Age Max']

# Display the result
print(result)
```

```
        Survival Rate  Age Min  Age Max
Pclass
1                0.58     0.92     80.00
2                0.42     0.67     70.00
3                0.27     0.17     74.00
```

In [71]:
```
fig, ax = plt.subplots(1, 2, figsize=(12, 6))

# Plot 1: Survival Rate by Pclass (Bar Plot)
sns.barplot(x=survival_by_pclass.index, y=survival_by_pclass.values, ax=ax[0])
ax[0].set_title('Survival Rate by Pclass')
ax[0].set_xlabel('Pclass')
ax[0].set_ylabel('Survival Rate')

# Create histograms for each Pclass
sns.histplot(df[df['Pclass'] == 1]['Age'], kde=True, color='blue', label='Pclass 1', bins=20, stat='density', alpha=0.6)
sns.histplot(df[df['Pclass'] == 2]['Age'], kde=True, color='green', label='Pclass 2', bins=20, stat='density', alpha=0.6)
sns.histplot(df[df['Pclass'] == 3]['Age'], kde=True, color='orange', label='Pclass 3', bins=20, stat='density', alpha=0.6)

# Customize plot
plt.title('Age Distribution by Pclass')
plt.xlabel('Age')
plt.ylabel('Density')
plt.legend(title='Pclass')

# Display the plots
plt.tight_layout()
plt.show()
```




In [72]:
```
pd.crosstab(df['Pclass'], df['Survived'], normalize = 'index')*100
```

Out[72]:

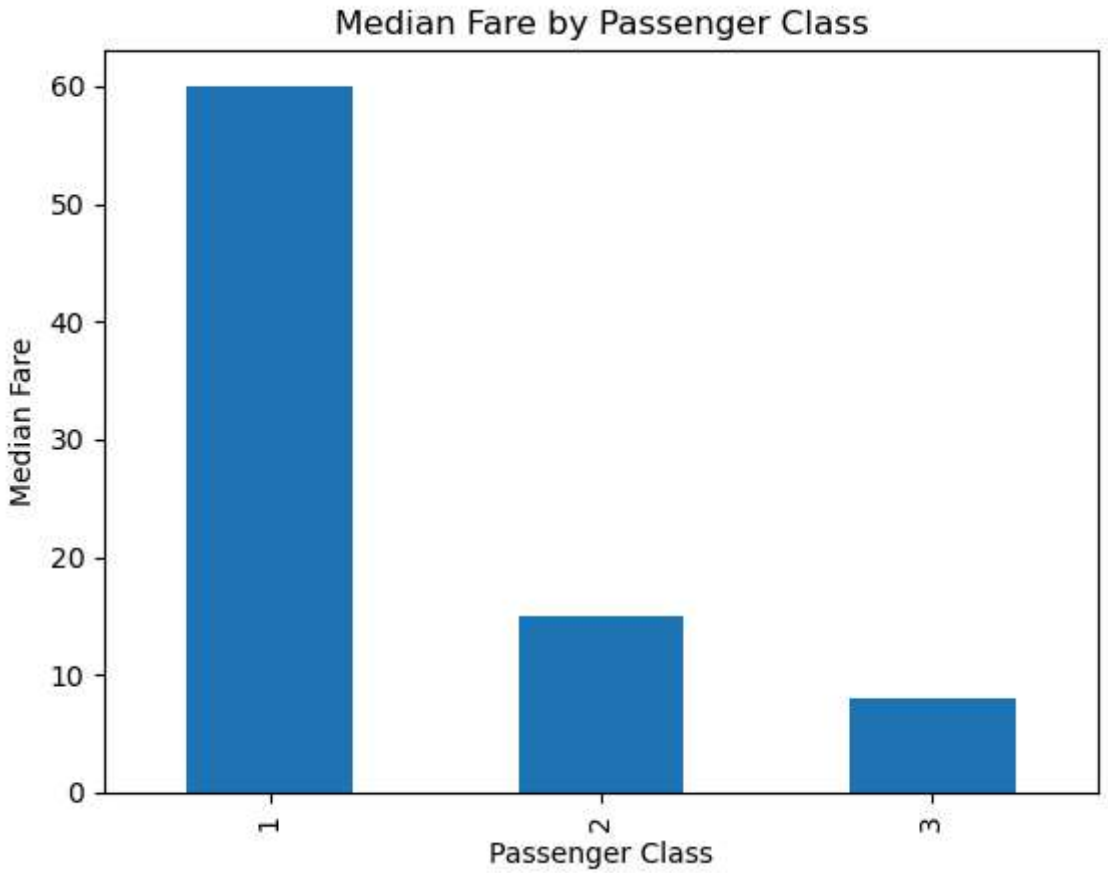
| Pclass | Survived 0 | 1 |
|---|---|---|
| 1 | 42.41 | 57.59 |
| 2 | 57.76 | 42.24 |
| 3 | 73.06 | 26.94 |

In [98]:
```
# Group by Pclass and see Fare statistics
df.groupby('Pclass')['Fare'].describe()
```

Out[98]:

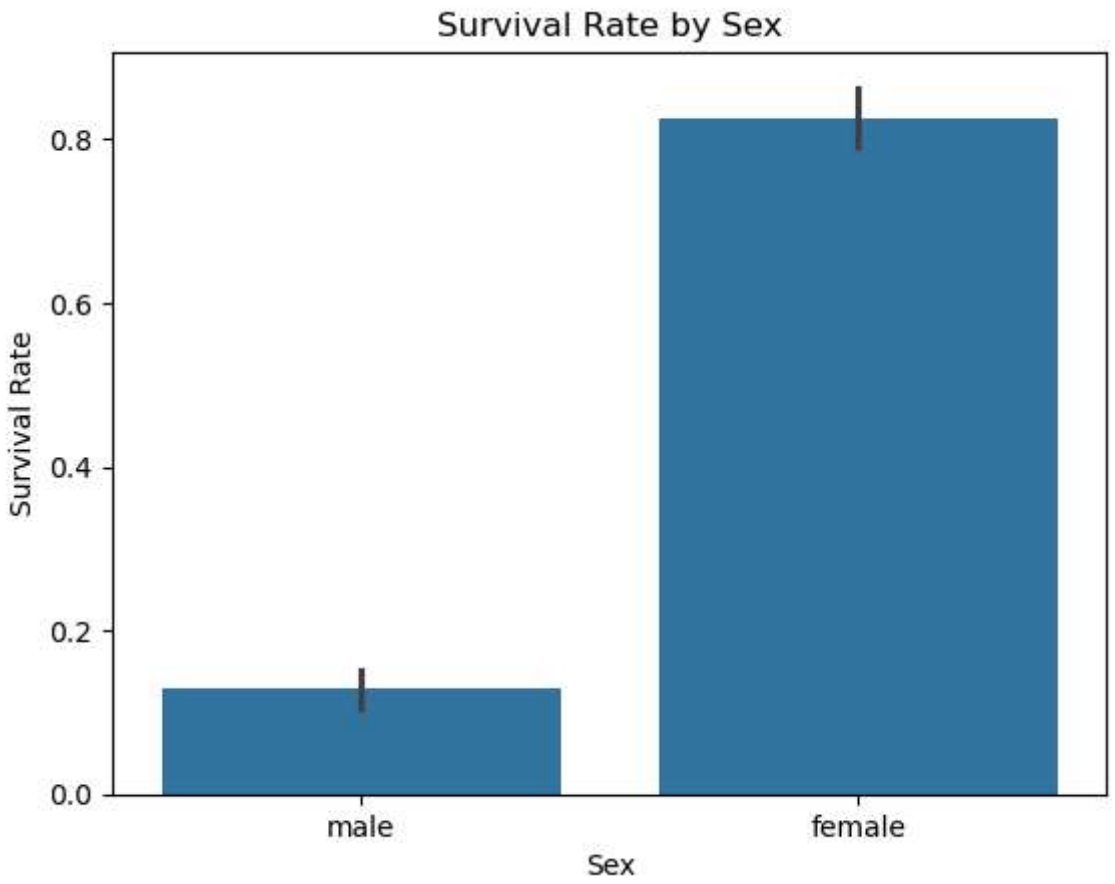
| Pclass | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| 1 | 323.00 | 87.51 | 80.45 | 0.00 | 30.70 | 60.00 | 107.66 | 512.33 |
| 2 | 277.00 | 21.18 | 13.61 | 0.00 | 13.00 | 15.05 | 26.00 | 73.50 |
| 3 | 709.00 | 13.30 | 11.49 | 0.00 | 7.75 | 8.05 | 15.25 | 69.55 |

In [106…
```
# Median fare by class
median_fares = df.groupby('Pclass')['Fare'].median()

# Plotting
median_fares.plot(kind='bar')
plt.title('Median Fare by Passenger Class')
plt.xlabel('Passenger Class')
plt.ylabel('Median Fare')
plt.show()
```



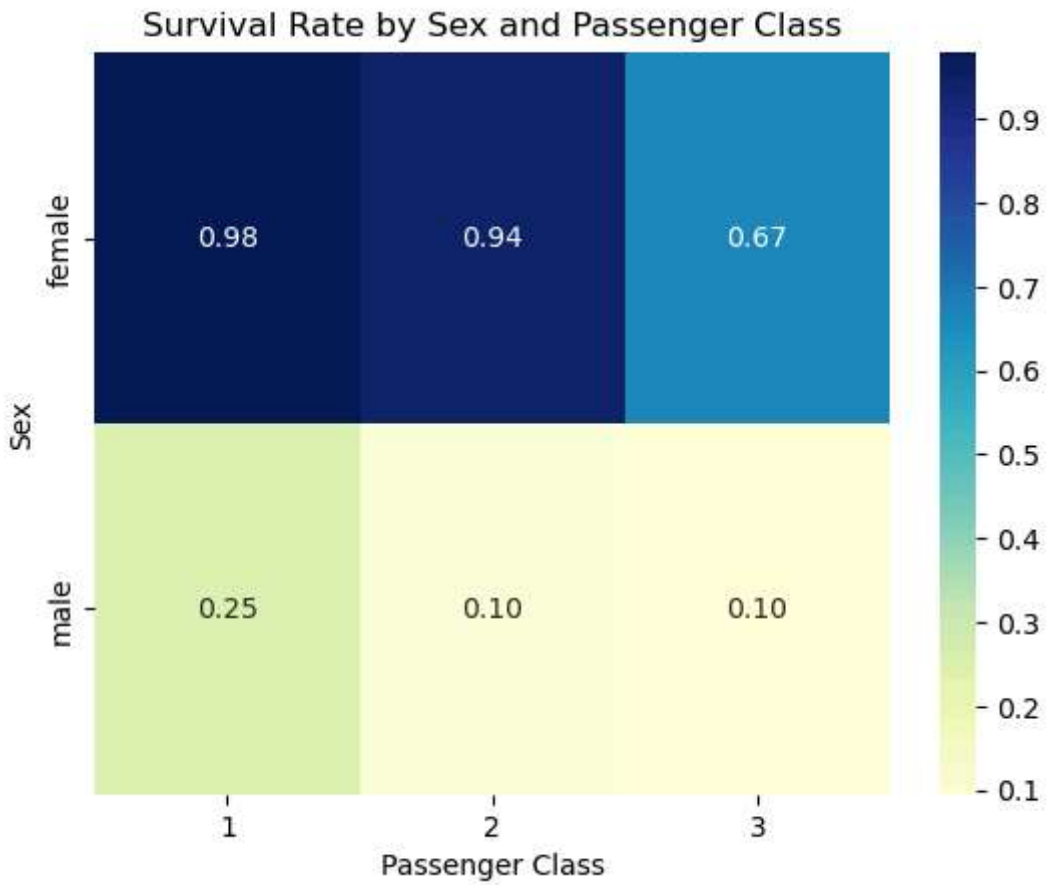

In [112…

```
# Group by Sex and find mean survival
survival_rate_by_sex = df.groupby('Sex')['Survived'].mean()

sns.barplot(x='Sex', y='Survived', data=df)
plt.title('Survival Rate by Sex')
plt.ylabel('Survival Rate')
plt.xlabel('Sex')
plt.show()
```




In [116…

```
# Group by both Sex and Pclass
survival_rate_by_sex_class = df.groupby(['Sex', 'Pclass'])['Survived'].mean().unstack()


sns.heatmap(survival_rate_by_sex_class, annot=True, cmap="YlGnBu", fmt=".2f")
plt.title('Survival Rate by Sex and Passenger Class')
plt.ylabel('Sex')
plt.xlabel('Passenger Class')
plt.show()
```


Survival Rate by Sex and Passenger Class
female
male
Sex
0.98
0.94
0.67
0.25
0.10
0.10
1
2
3
Passenger Class
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1