Exploratory Data Analysis on the Titanic Dataset

1. Introduction

- **Objective of the analysis:**

Extract insights using visual and statistical exploration.

- **Dataset source :** [Titanic - Machine Learning from Disaster | Kaggle]

Basic Info:

- Rows: 891 passengers
- Columns: 12 attributes + target (Survived)

OUTPUT:

```
# 3. Basic Info & Summary
print(df.info()) print(df.describe())
print(df.isnull().sum())
print(df['Survived'].value_couts())
```

```
<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
       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
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
Survived
0    549
1    342
Name: count, dtype: int64
```

2.Missing Values:

- **Age: ~177 missing**
- **Cabin: many missing**
- **Embarked: 2 missing**

3. Data Cleaning

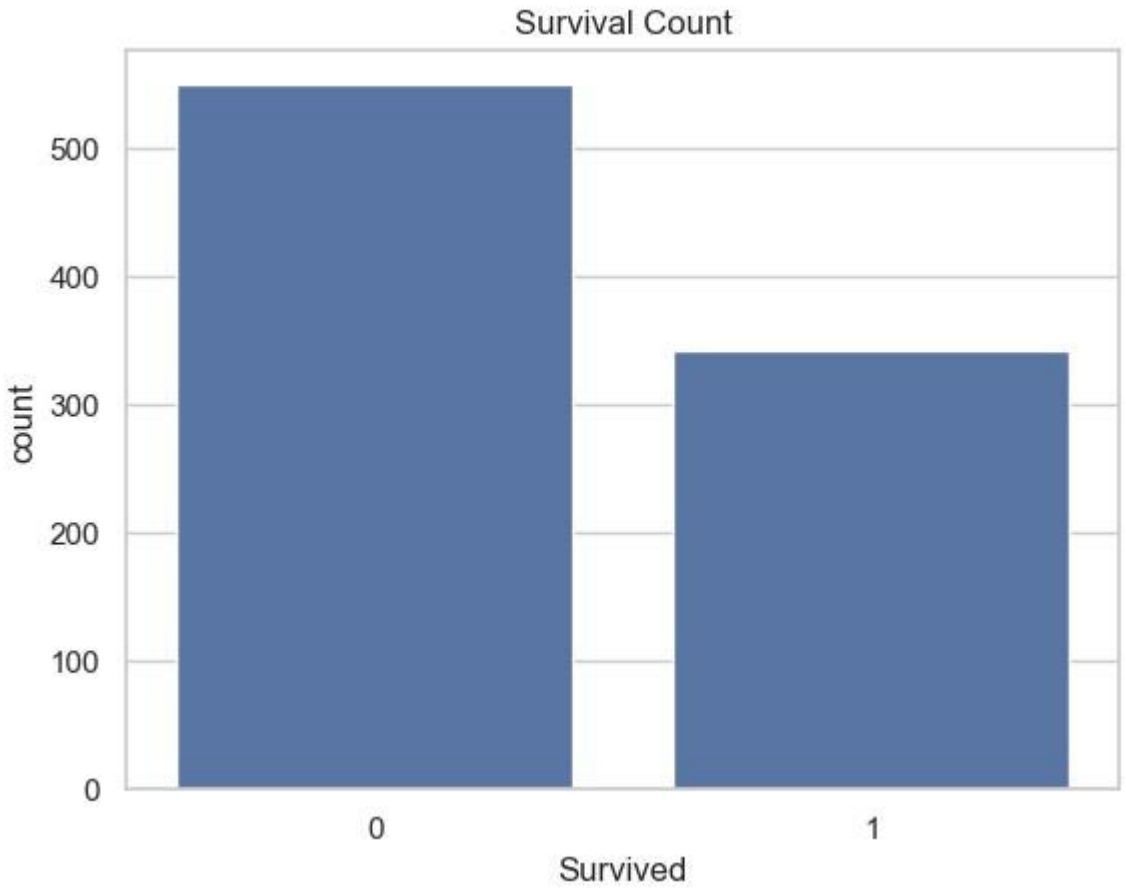
- **Dropped irrelevant columns: Ticket, Cabin, and Name.**
- **Imputed missing Age values with median.**
- **Filled missing Embarked with mode**

```
plt.title("Fare Distribution") plt.show()
```

4. Univariate Analysis

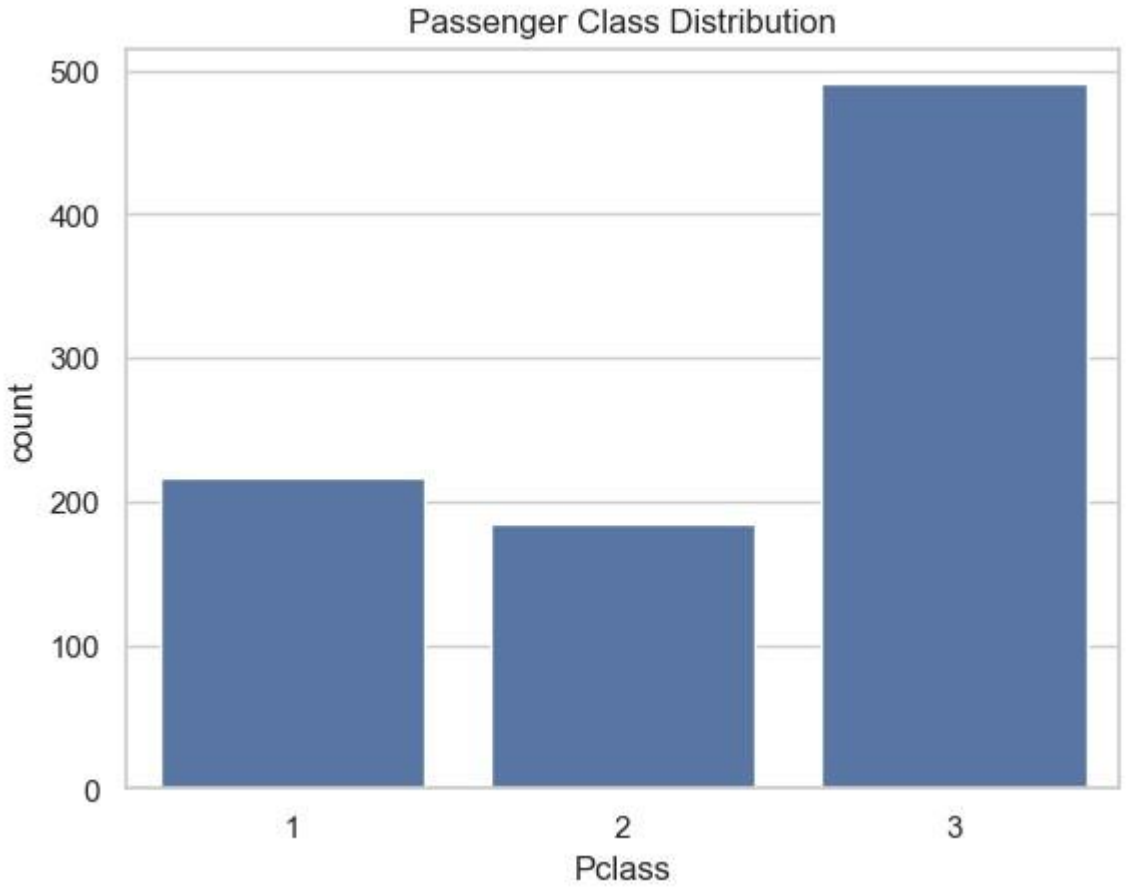
➤ Survival Count

- 0: Did not survive – 549
- 1: Survived – 342



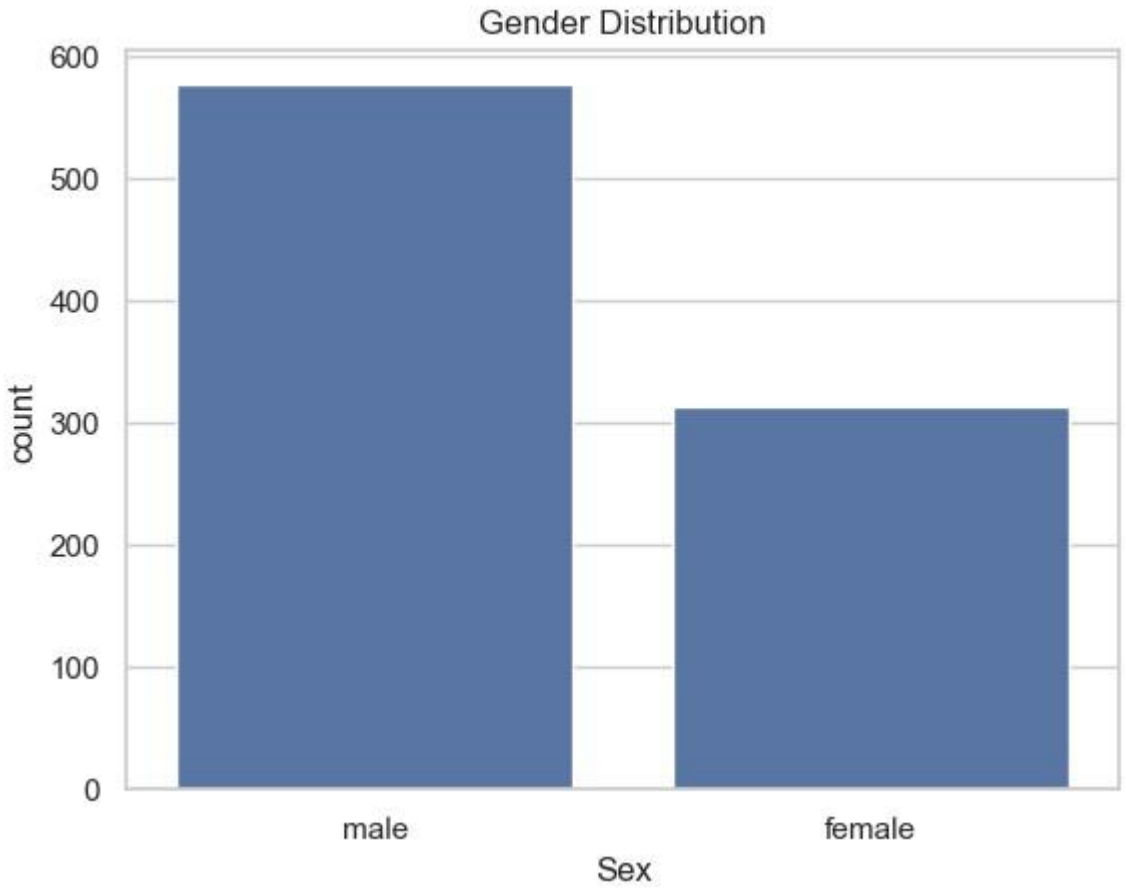
➤ passenger class distribution

Pclass 3 has highest distribution



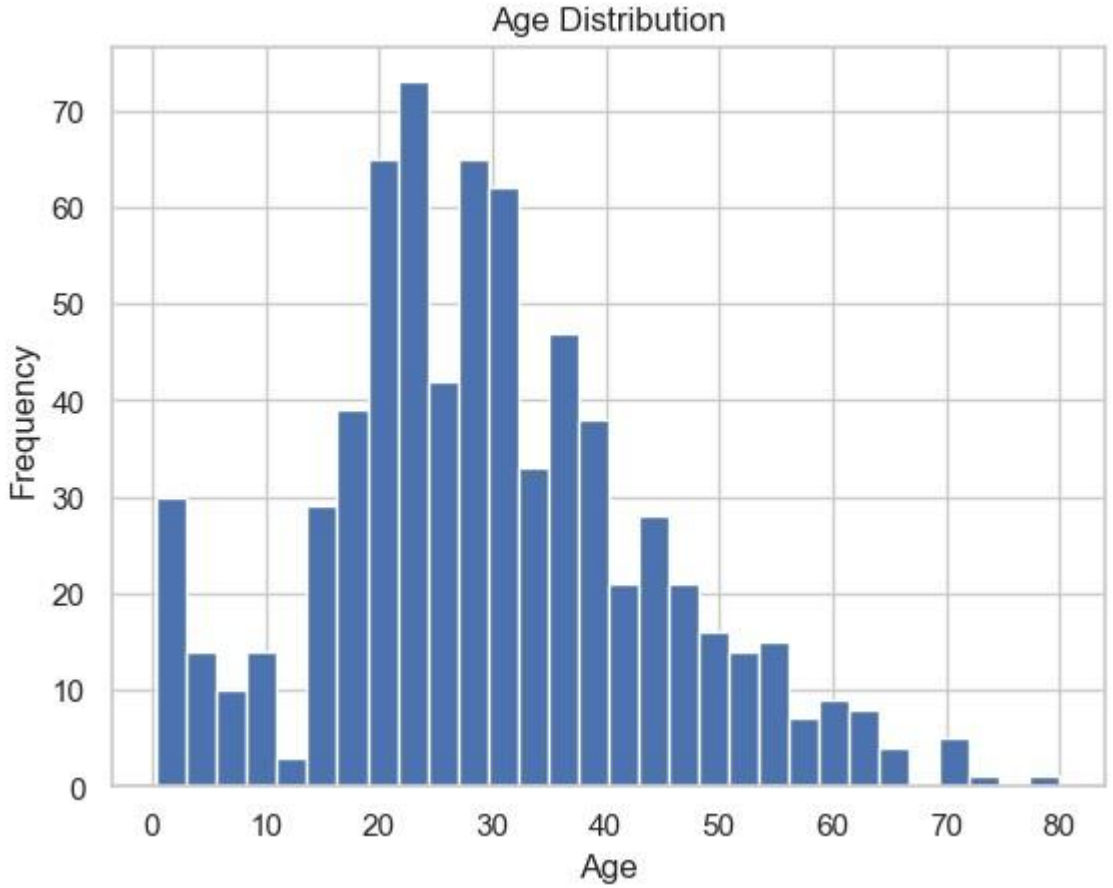
➤ Gender Distribution

- Male: 577
- Female: 314



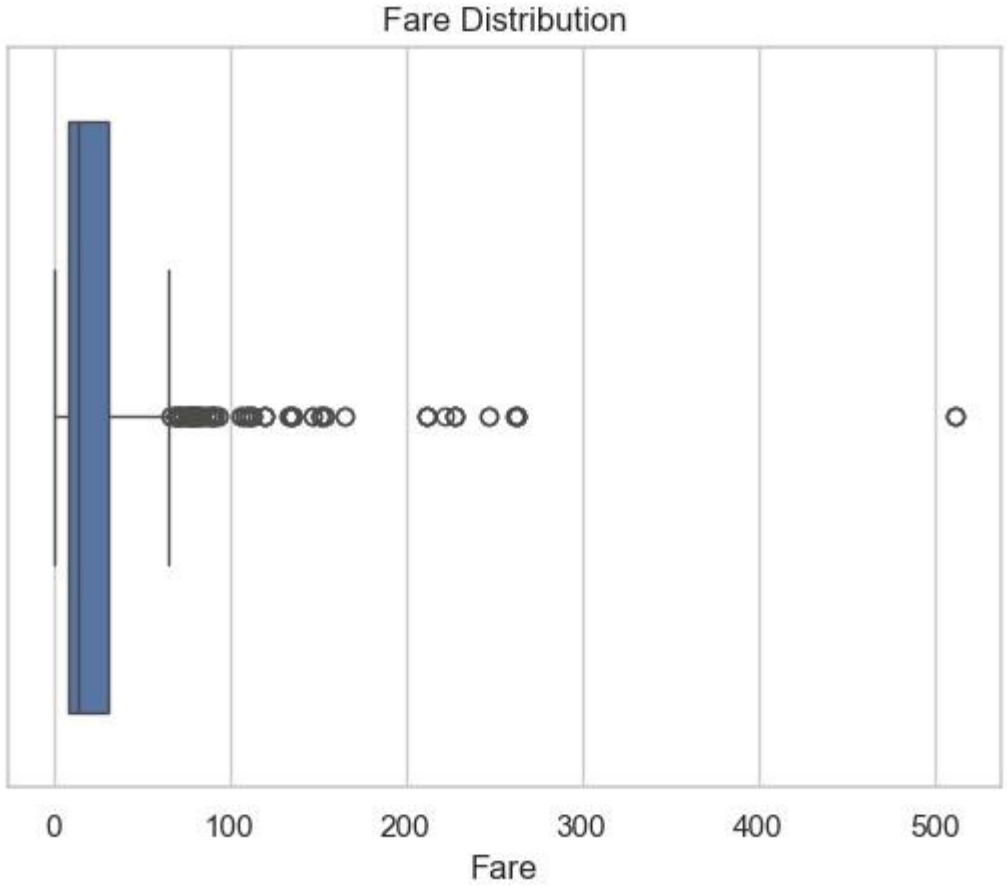
➤ Age Distribution

Most passengers were between 20–30 years old.


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

➤ Fare Distribution

Highly skewed, with some high-paying outliers.



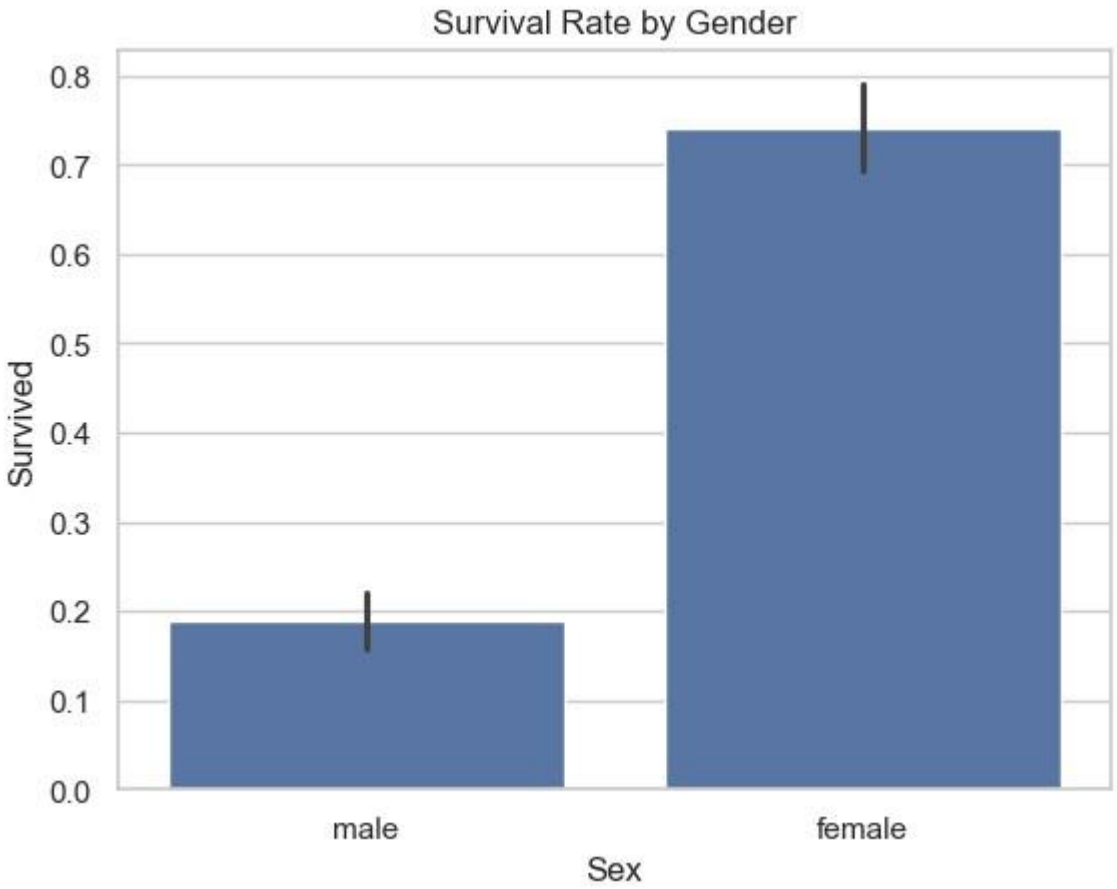
In [13]:
```
# Survival by Sex
sns.barplot(data=df, x='Sex', y='Survived')
plt.title("Survival Rate by Gender") plt.show()

# Survival by Class
sns.barplot(data=df, x='Pclass', y='Survived')
plt.title("Survival Rate by Passenger Class") plt.show()

# Age vs Survival
sns.violinplot(data=df, x='Survived', y='Age')
```

```
plt.title("Age vs Survival") plt.show()
```

5.Bivariate / Multivariate Analysis

➤ Survival by Gender

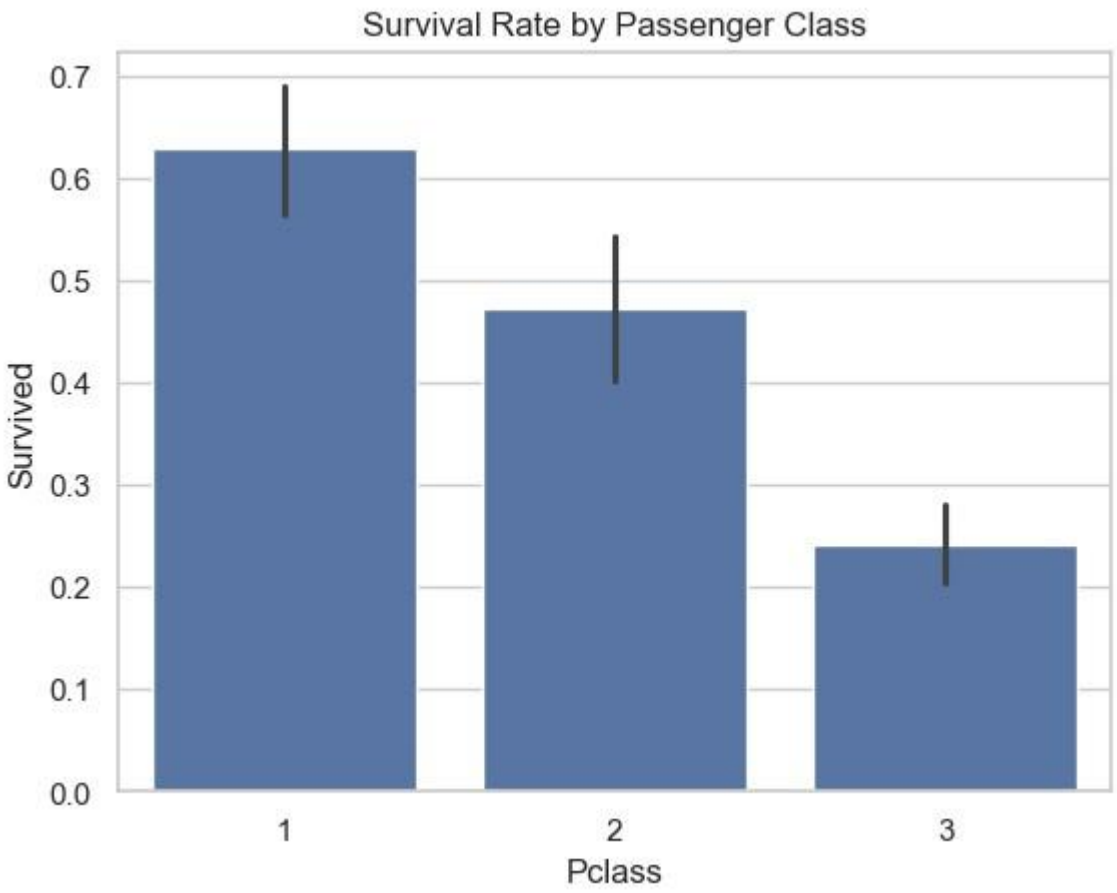
Female survival rate is much higher than males



➤ **Survival by passenger class**

1st class: highest survival

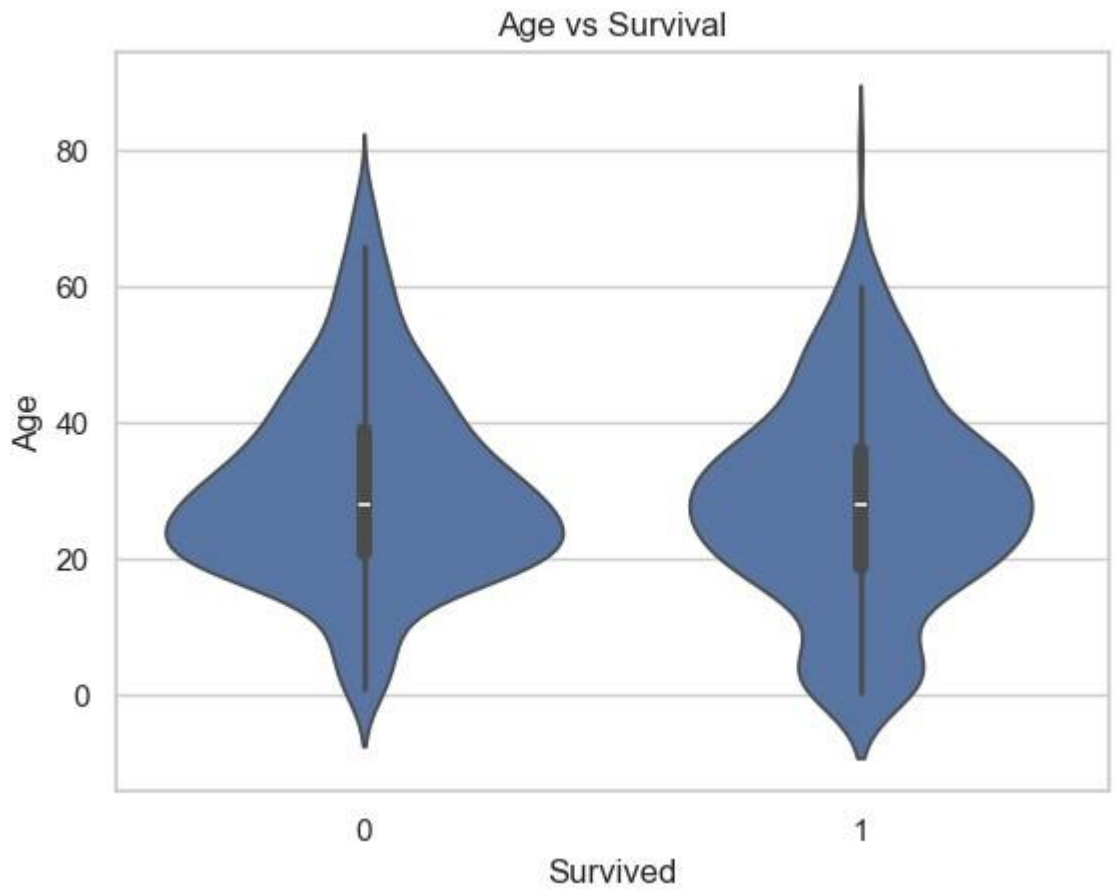
3rd class: lowest survival



➤ **age vs survival**

Children (0–10 years) appear more prominently on the survival side:

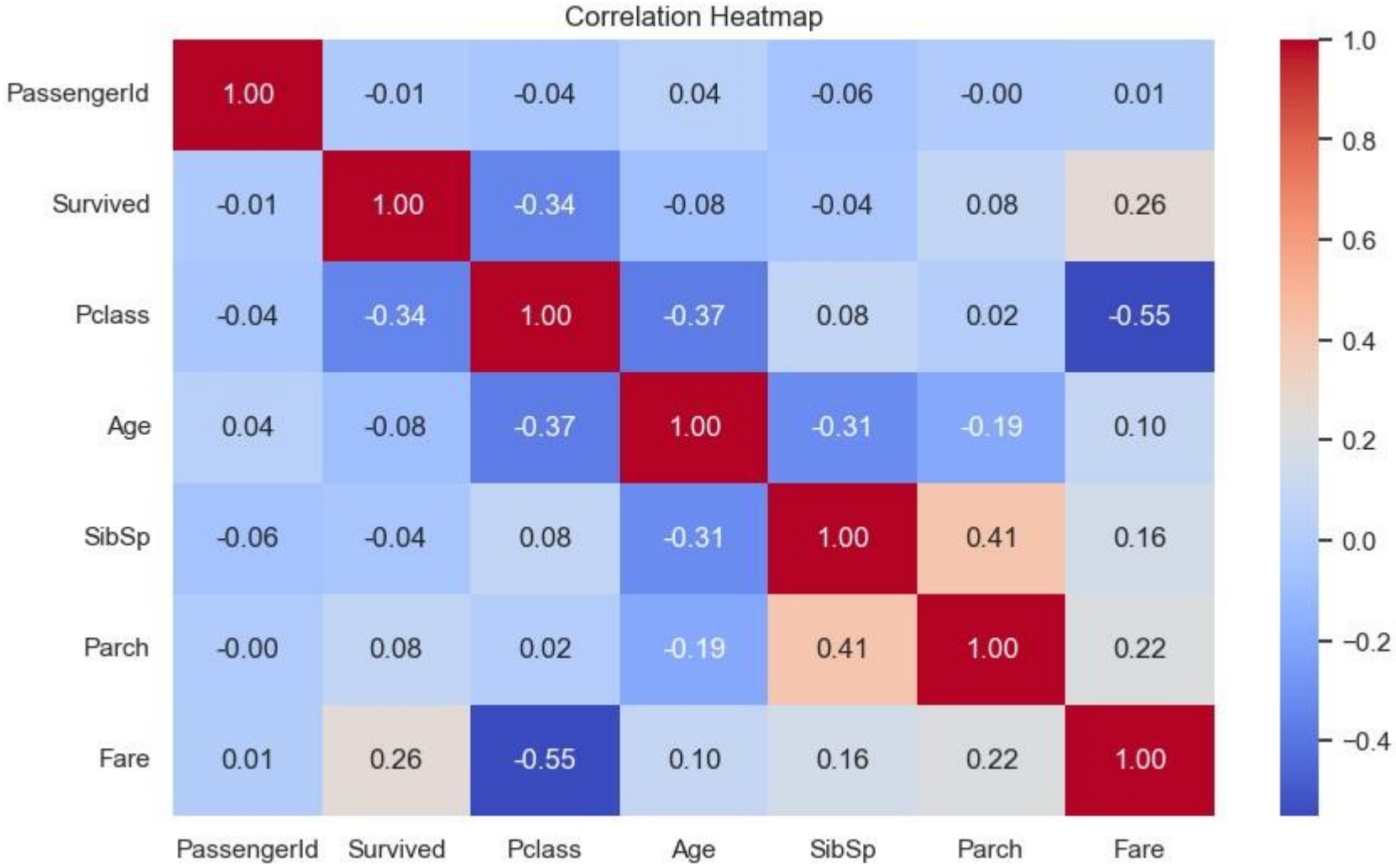
- There is a wider base on the left side (Survived = 1) for young children.

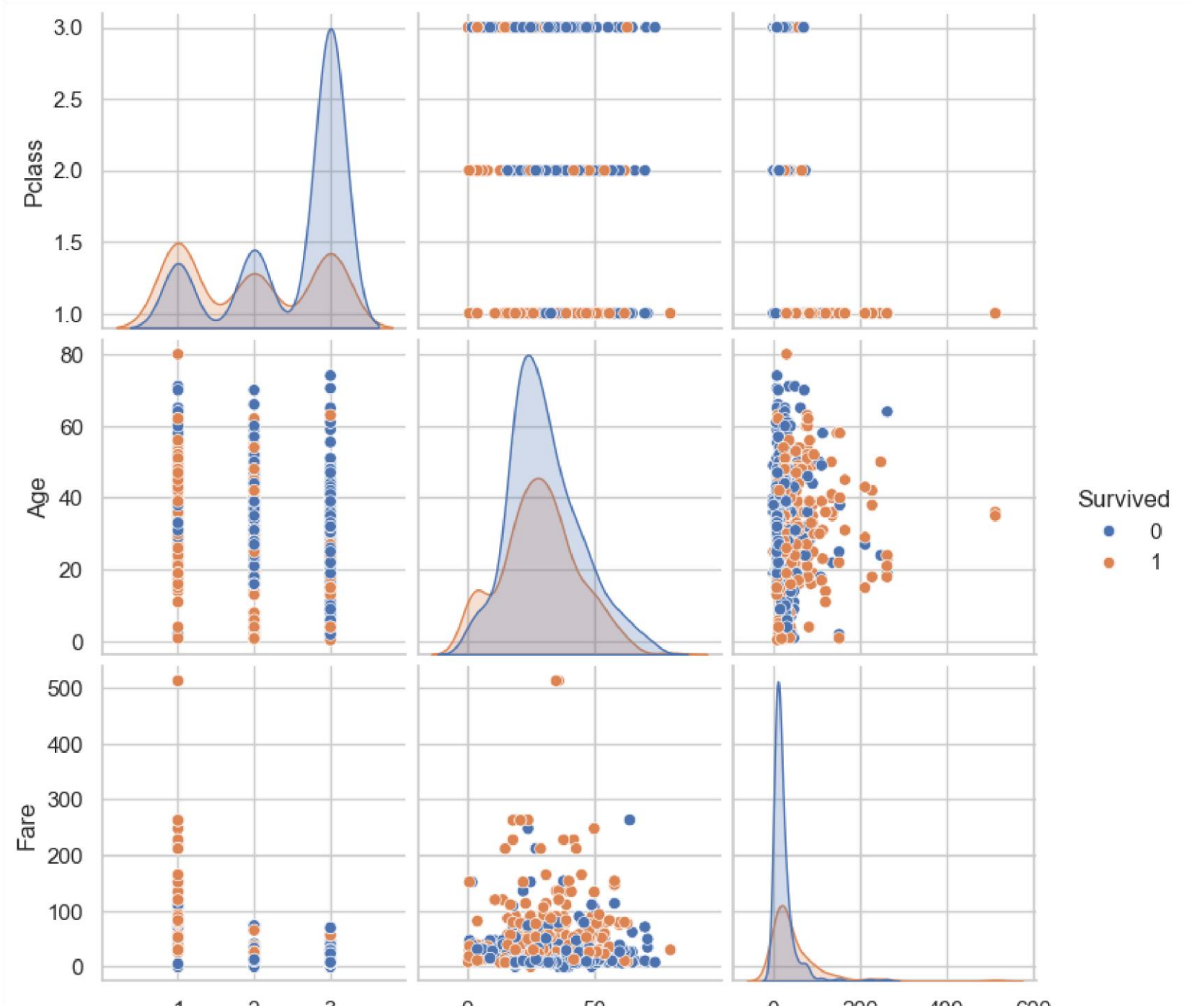

Age vs Survival
80
60
40
20
0
Age
0
1
Survived

In [15]:
```
# Heatmap of correlations plt.figure(figsize=(10, 6))
sns.heatmap(df.corr(numeric_only=True), annot=True, cmap='coolwarm', fmt='.2f')
plt.title("Correlation Heatmap") plt.show()

# Pairplot (selected features)
sns.pairplot(df[['Survived', 'Pclass', 'Age', 'Fare']], hue='Survived') plt.show()
```

➤ Correlation Heatmap

Pclass, Sex, Fare, and Age are most correlated with survival.


Correlation Heatmap
PassengerId
Survived
Pclass
Age
SibSp
Parch
Fare
1.00 -0.01 -0.04 0.04 -0.06 -0.00 0.01
-0.01 1.00 -0.34 -0.08 -0.04 0.08 0.26
-0.04 -0.34 1.00 -0.37 0.08 0.02 -0.55
0.04 -0.08 -0.37 1.00 -0.31 -0.19 0.10
-0.06 -0.04 0.08 -0.31 1.00 0.41 0.16
-0.00 0.08 0.02 -0.19 0.41 1.00 0.22
0.01 0.26 -0.55 0.10 0.16 0.22 1.00
PassengerId
Survived
Pclass
Age
SibSp
Parch
Fare
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


Pclass
Age
Fare
3.0
2.5
2.0
1.5
1.0
80
60
40
20
0
500
400
300
200
100
0
Survived
0
1

6. Key Observations

- 🧍 **Gender matters**: Female passengers had a survival rate of over 70%, while only ~20% of males survived.
- 💼 **Class advantage**: First-class passengers were more likely to survive.
- 👶 **Age factor**: Younger passengers had slightly higher chances.
- 💰 **Fare correlation**: Passengers who paid higher fares had better survival odds.
- 🚢 **Embarkation point** also showed minor differences in survival rates.

7. Conclusion & Summary of Findings

The Titanic dataset offers key insights into what contributed to survival. Our EDA reveals that sex, class, and age significantly influenced the chances of survival. Women and children in first class had the highest survival rates, confirming the historical notion of **“women and children first.”**

These findings can serve as a foundation for further machine learning modeling or deeper socio-economic analysis.