

```
In [3]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

sns.set(style="whitegrid")
```

```
In [4]: df = pd.read_csv("train.csv")
df.head()
```

```
Out[4]:
```

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

```
In [5]: df.shape
```

```
Out[5]: (891, 12)
```

```
In [6]: df.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
```

In [7]: `df.describe()`

Out[7]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

In [8]: `df.isnull().sum()`

Out[8]:

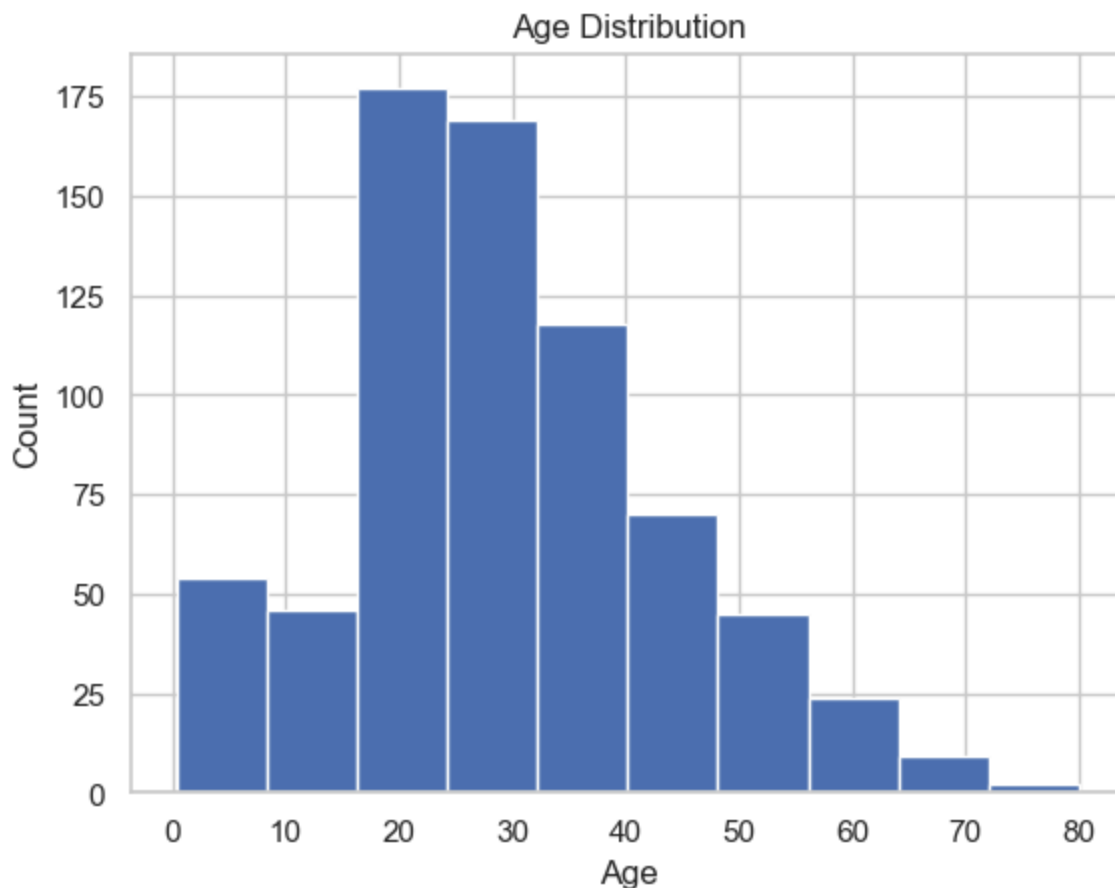
```
PassengerId    0
Survived        0
Pclass         0
Name           0
Sex            0
Age           177
SibSp          0
Parch          0
Ticket         0
Fare           0
Cabin        687
Embarked        2
dtype: int64
```

```
In [19]: df['Sex'].value_counts()  
df['Pclass'].value_counts()  
df['Survived'].value_counts()  
df['Embarked'].value_counts()
```

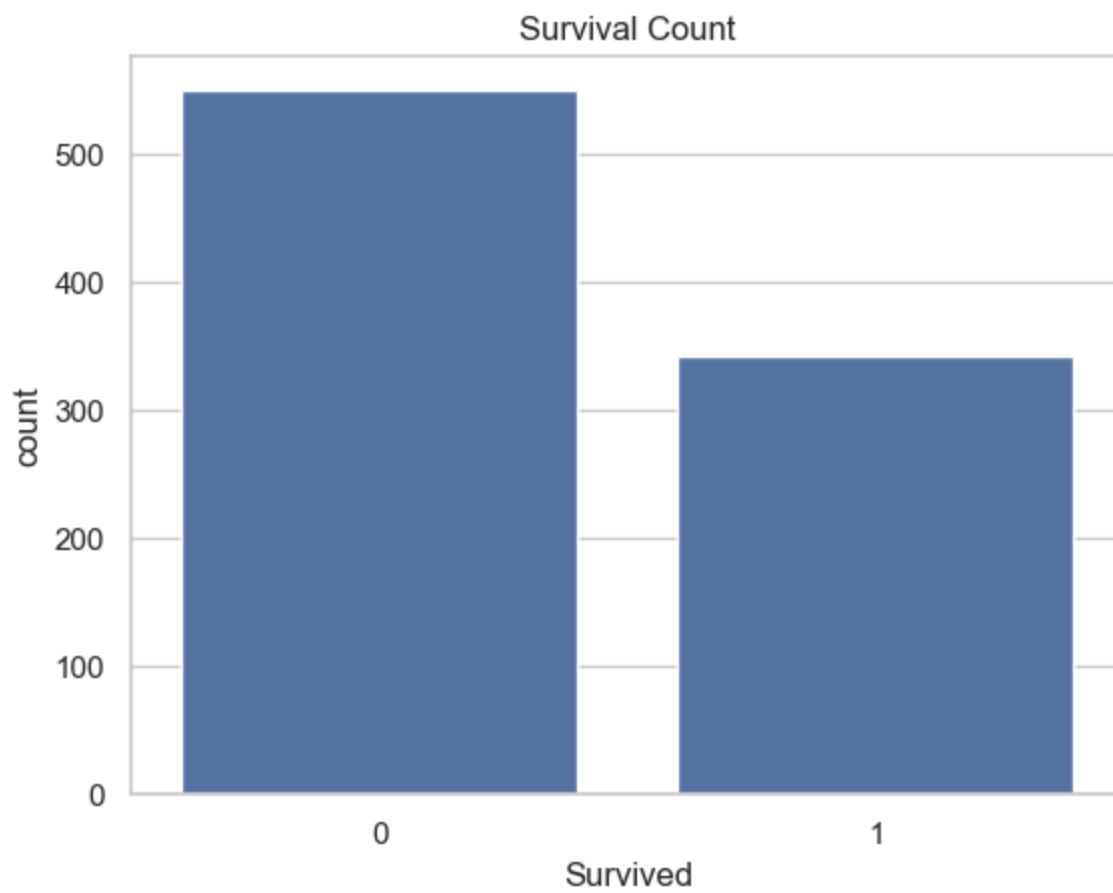
```
Out[19]: Embarked  
S      644  
C      168  
Q       77  
Name: count, dtype: int64
```

- 1.Age and Cabin columns have missing values
- 2.Dataset contains 891 rows and 12 columns
- 3.Fare has high variation

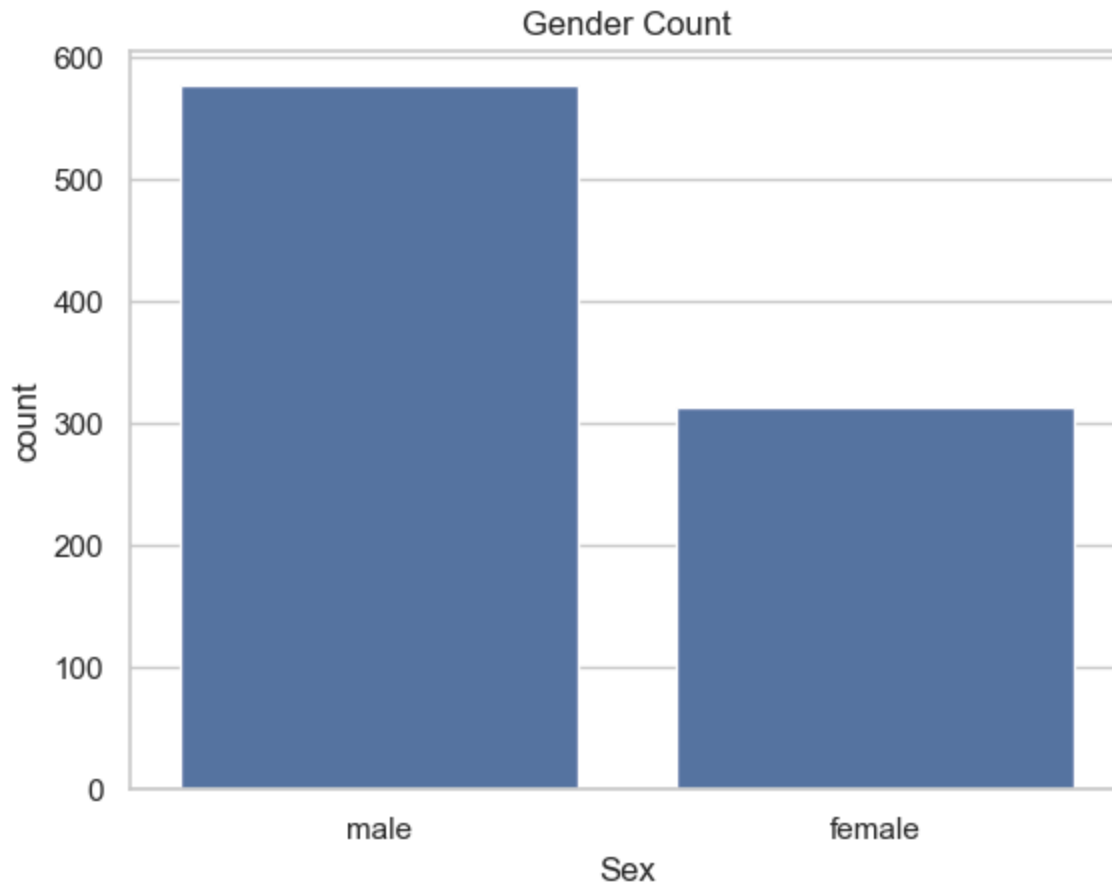
```
In [9]: plt.hist(df["Age"].dropna())  
plt.title("Age Distribution")  
plt.xlabel("Age")  
plt.ylabel("Count")  
plt.show()
```



```
In [10]: sns.countplot(x="Survived", data=df)  
plt.title("Survival Count")  
plt.show()
```

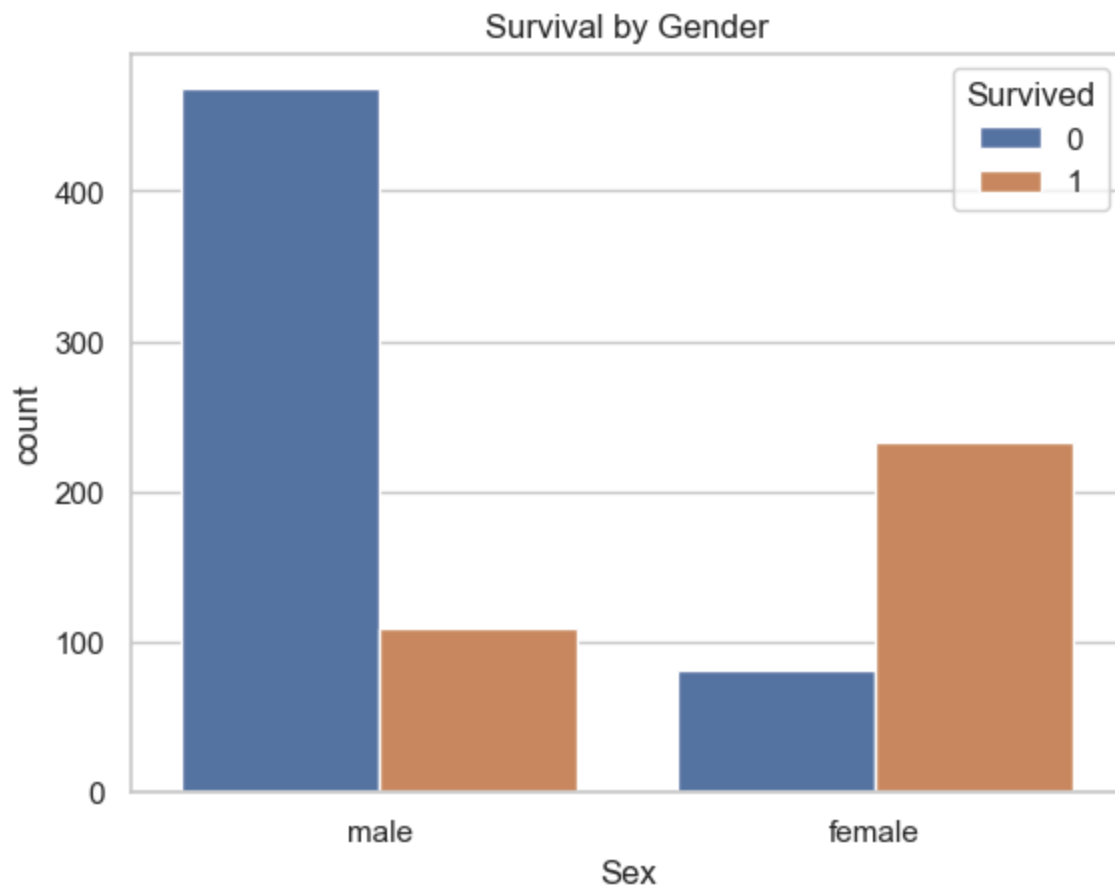


```
In [11]: sns.countplot(x="Sex", data=df)
plt.title("Gender Count")
plt.show()
```

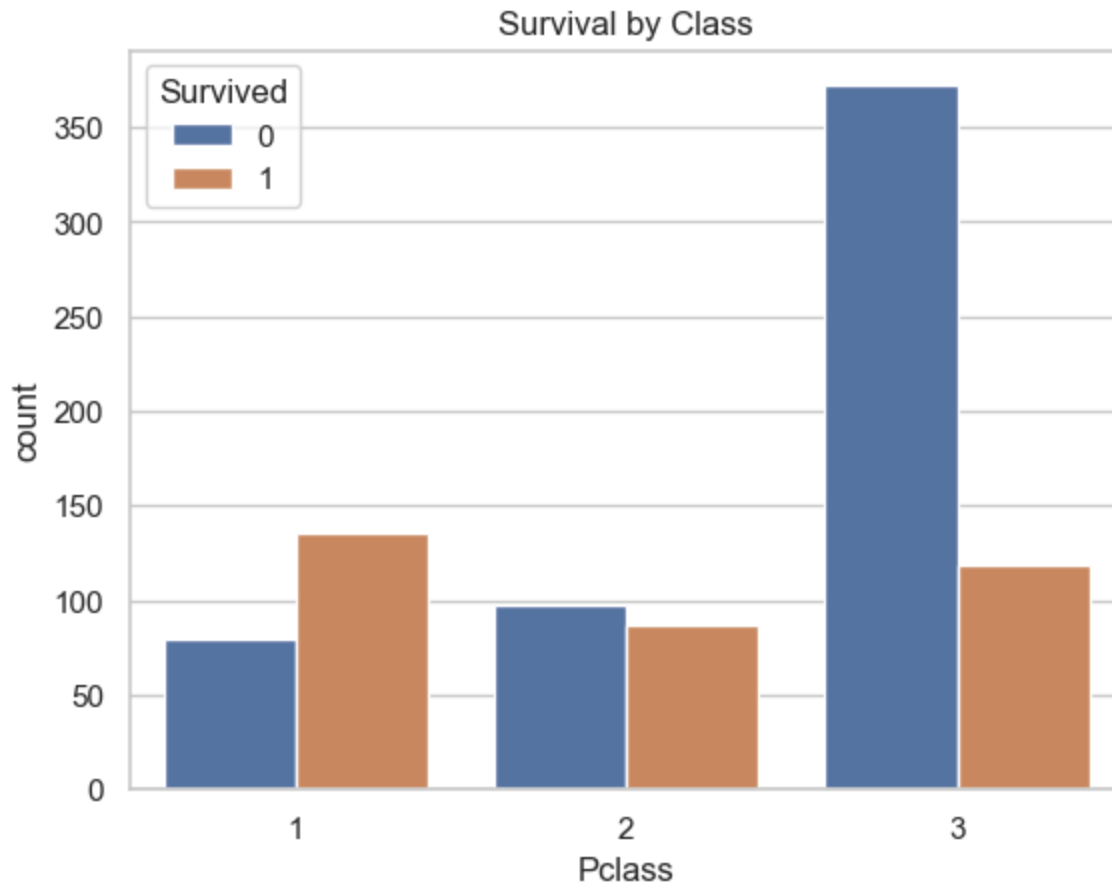


1. Most passengers did not survive 2. More males than females were on board

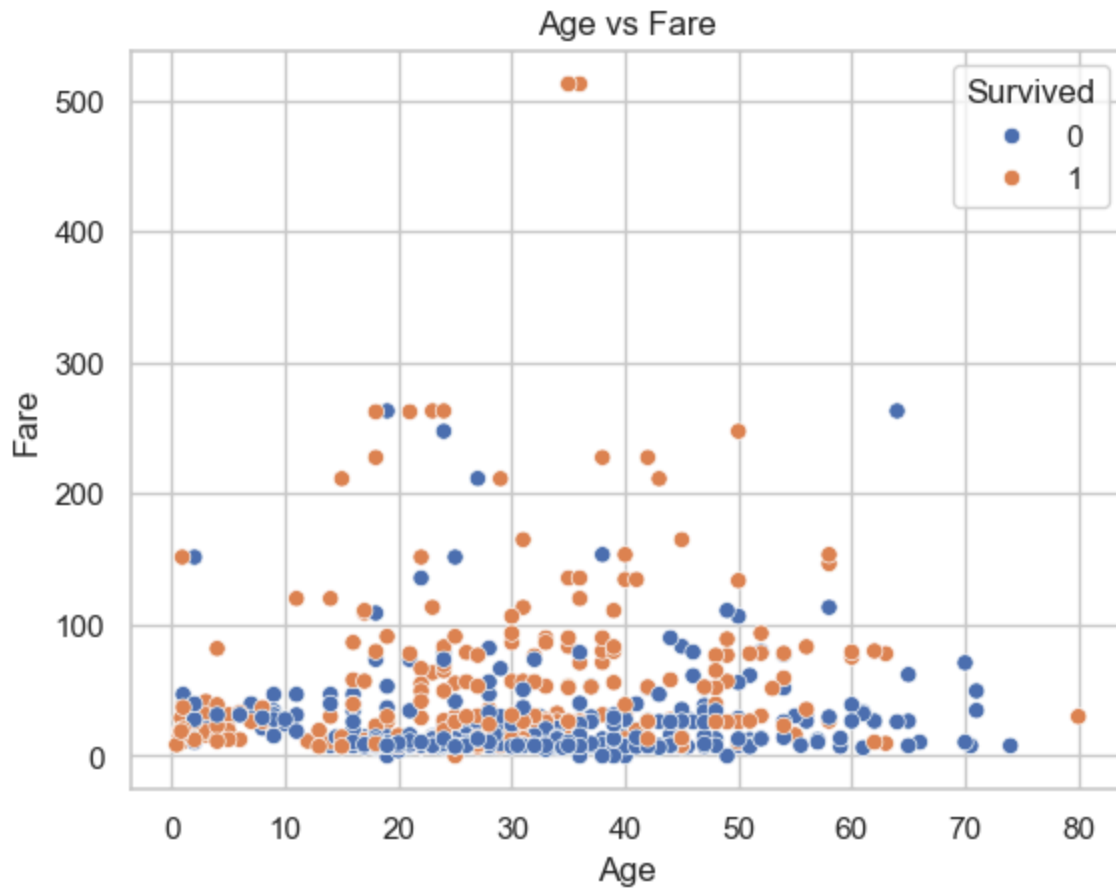
```
In [12]: sns.countplot(x="Sex", hue="Survived", data=df)
plt.title("Survival by Gender")
plt.show()
```



```
In [13]: sns.countplot(x="Pclass", hue="Survived", data=df)
plt.title("Survival by Class")
plt.show()
```



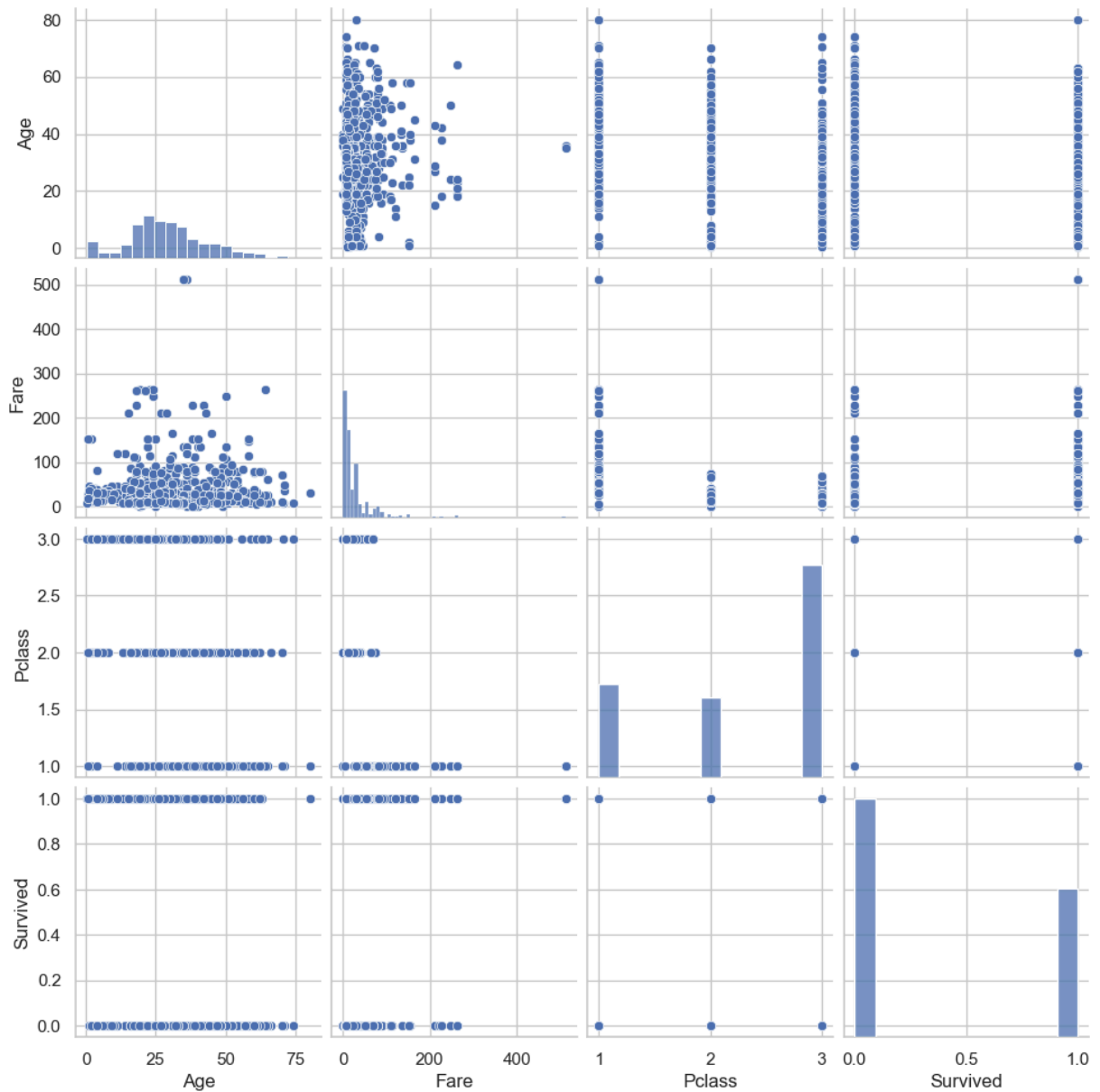
```
In [14]: sns.scatterplot(x="Age", y="Fare", hue="Survived", data=df)
plt.title("Age vs Fare")
plt.show()
```



1. Females had higher survival rates 2. 1st class passengers survived more 3. Higher fare → better survival

```
In [18]: sns.pairplot(df[['Age', 'Fare', 'Pclass', 'Survived']])
```

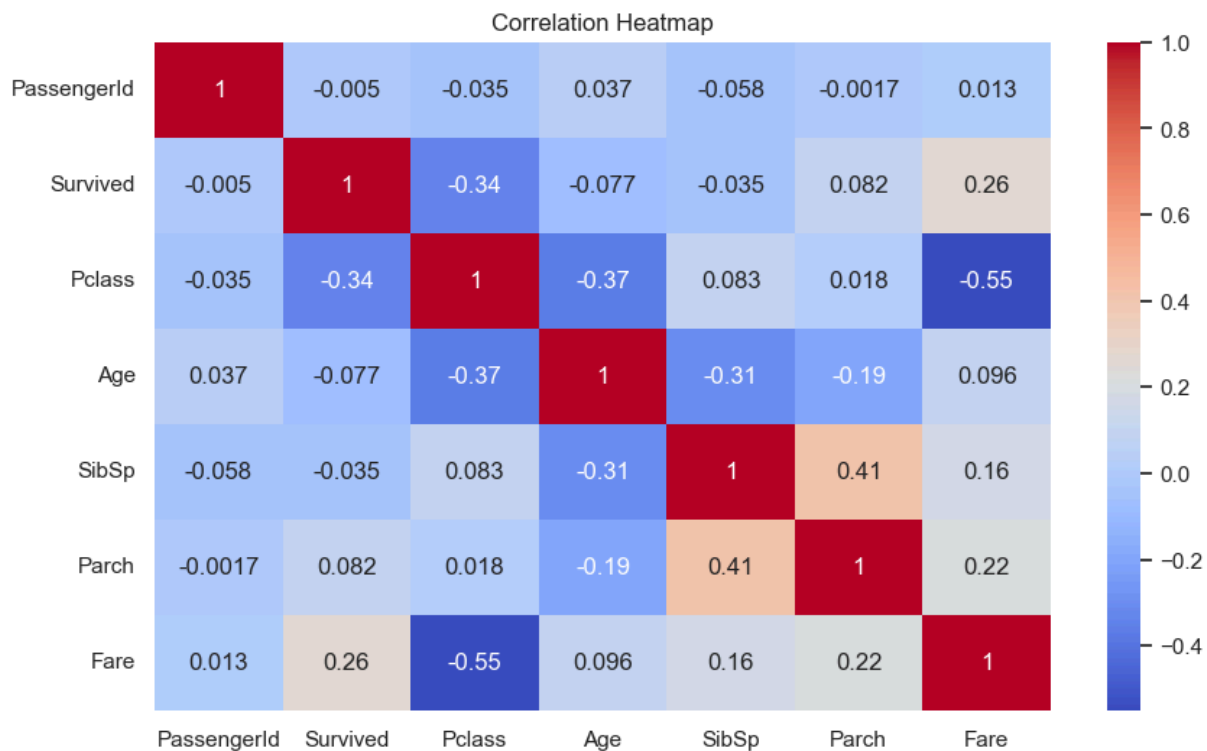
```
Out[18]: <seaborn.axisgrid.PairGrid at 0x169829f1160>
```



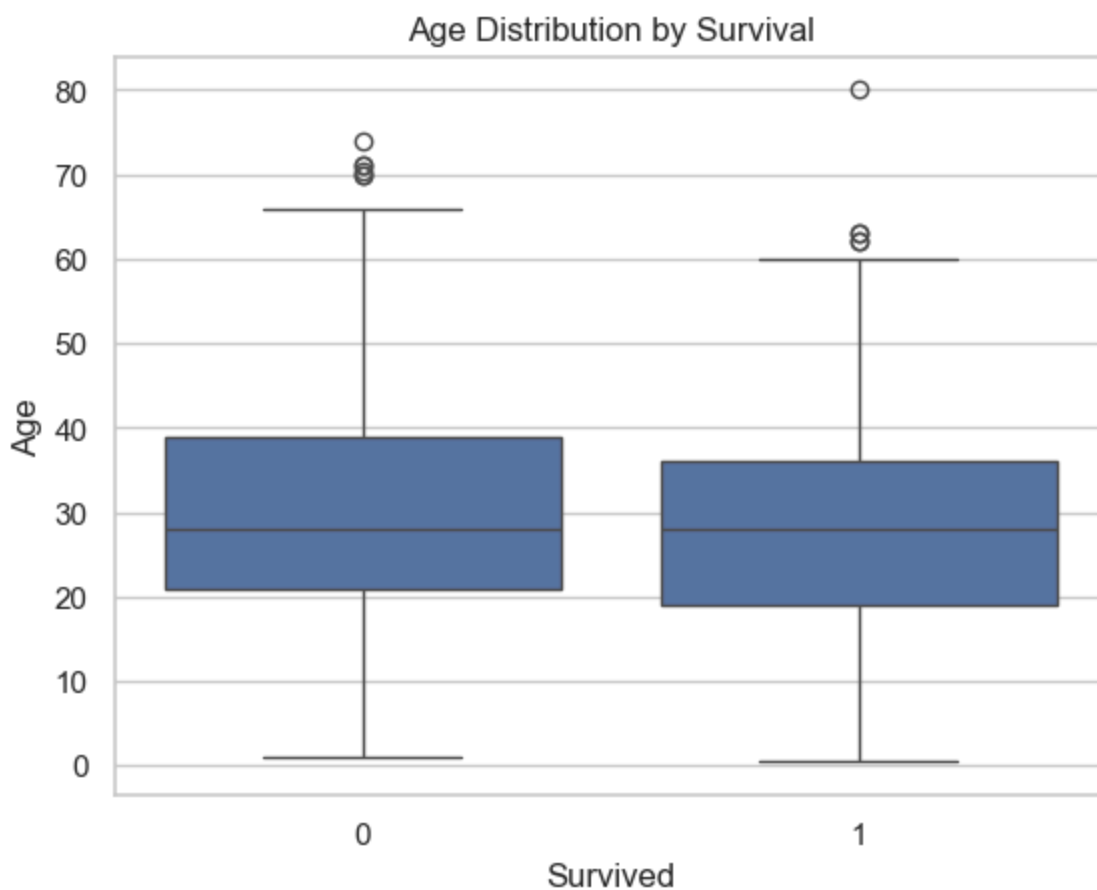
```
In [16]: plt.figure(figsize=(10,6))

numeric_df = df.select_dtypes(include=['int64', 'float64']) # selects only numeri

sns.heatmap(numeric_df.corr(), annot=True, cmap="coolwarm")
plt.title("Correlation Heatmap")
plt.show()
```



```
In [17]: sns.boxplot(x='Survived', y='Age', data=df)
plt.title("Age Distribution by Survival")
plt.show()
```



Summary of Findings

1. Females had a significantly higher survival rate than males.
2. Passengers in 1st class survived more than those in 2nd or 3rd class.
3. Higher ticket fare is strongly linked with increased survival chances.
4. The Age column contains some missing values, and the Cabin column contains many missing values.
5. Majority of passengers were in 3rd class and most of them did not survive.
6. Younger passengers had slightly better survival outcomes.
7. The correlation heatmap shows that Pclass and Fare have a noticeable relationship and both affect survival.