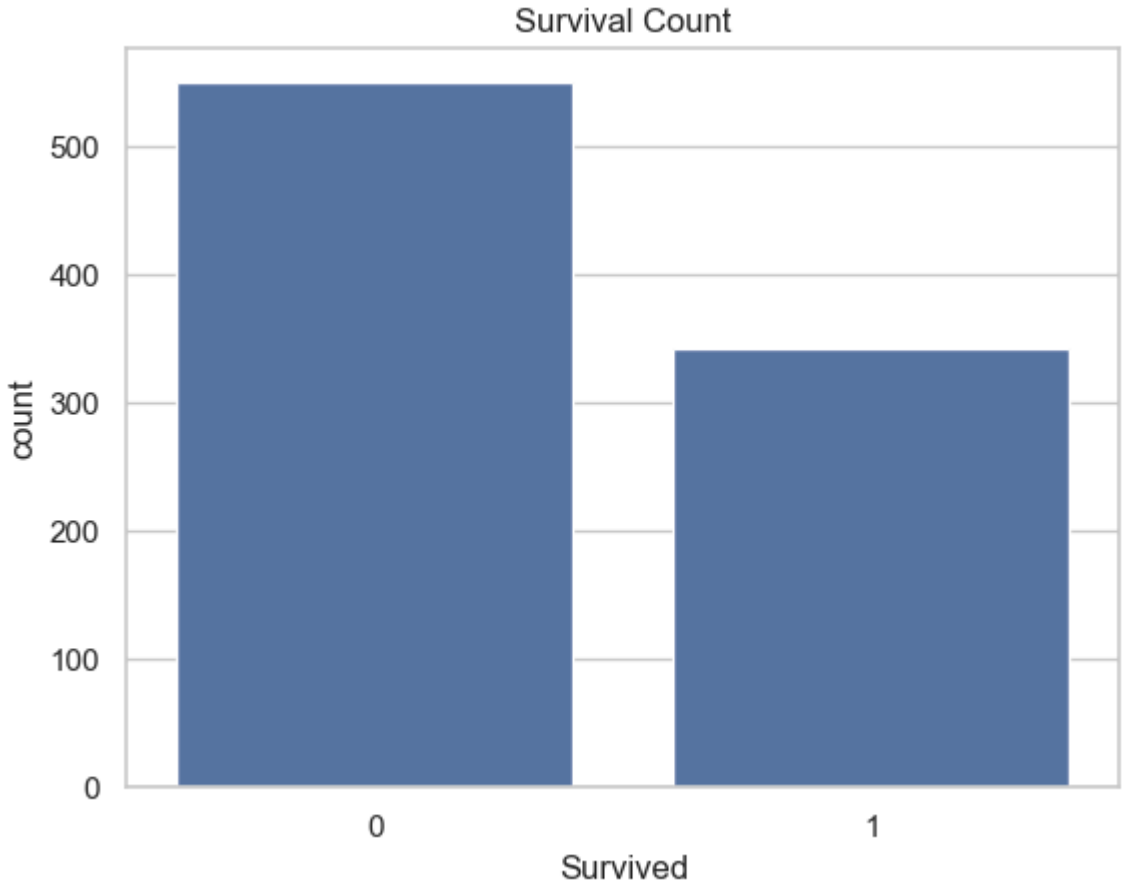
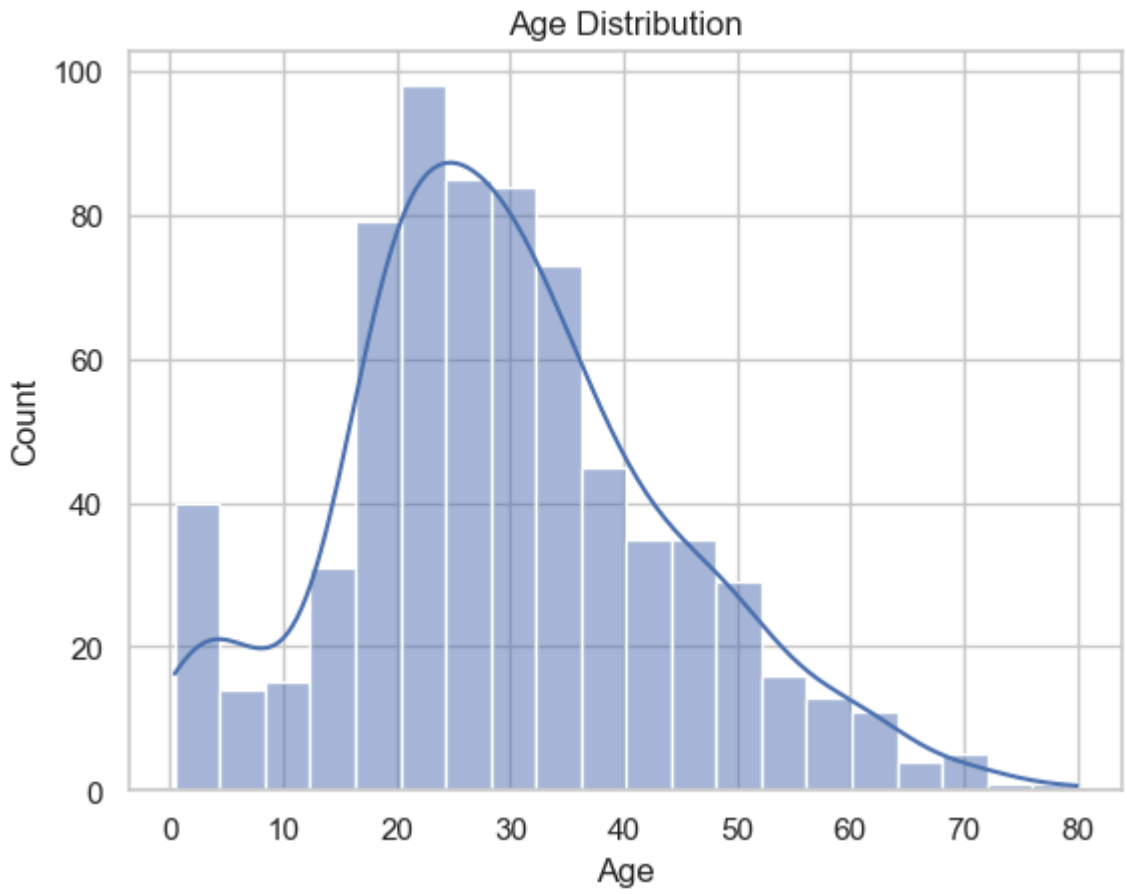
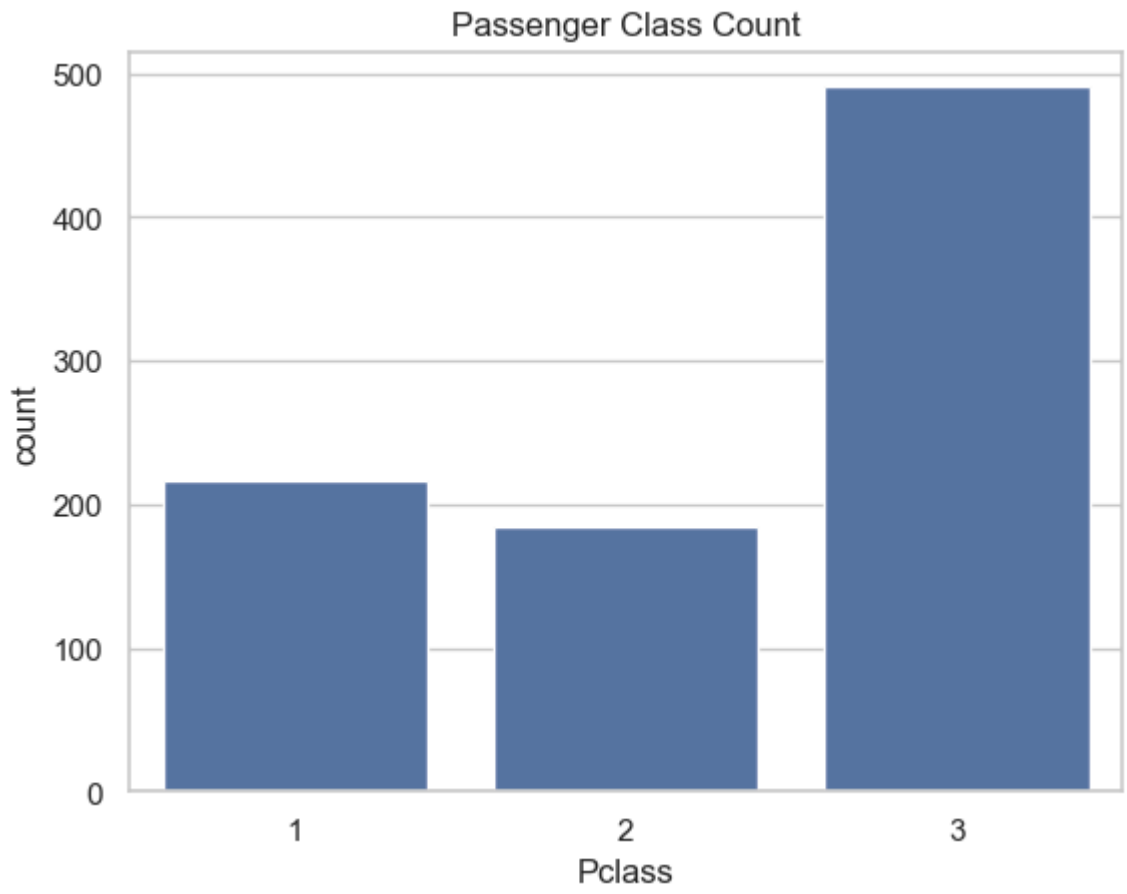
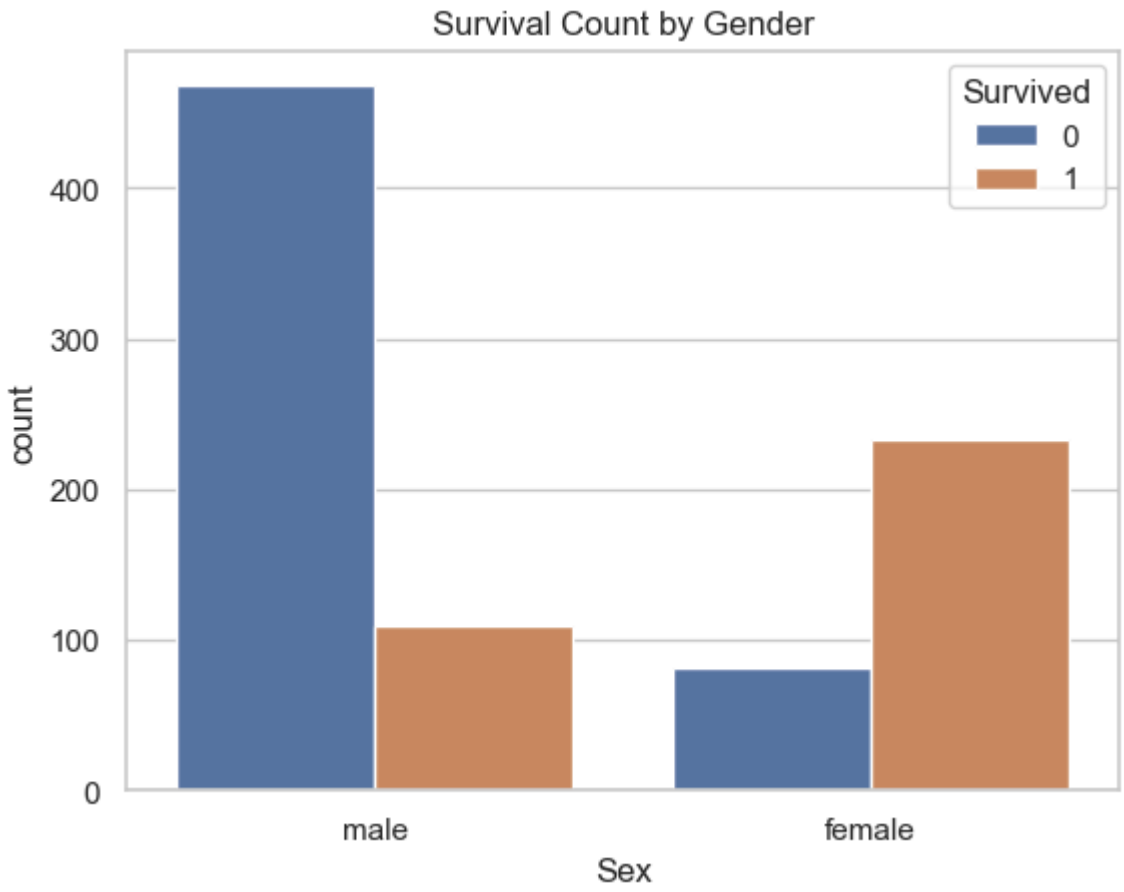
In [4]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
sns.set(style="whitegrid")
df = pd.read_csv("train.csv")
df.head()
df.info()
df.describe()
df.isnull().sum()
sns.countplot(x=df["Survived"])
plt.title("Survival Count")
plt.show()
sns.histplot(df["Age"], kde=True)
plt.title("Age Distribution")
plt.show()
sns.countplot(x=df["Pclass"])
plt.title("Passenger Class Count")
plt.show()
sns.countplot(x="Sex", hue="Survived", data=df)
plt.title("Survival Count by Gender")
plt.show()
sns.countplot(x="Pclass", hue="Survived", data=df)
plt.title("Survival by Passenger Class")
plt.show()
sns.boxplot(x="Survived", y="Age", data=df)
plt.title("Age vs Survival")
plt.show()
sns.pairplot(df[["Survived","Age","Fare","Pclass"]])
plt.show()
plt.figure(figsize=(10,6))
numeric_df = df.corr(numeric_only=True)
sns.heatmap(numeric_df, annot=True, cmap="coolwarm")
plt.title("Correlation Heatmap")
plt.show()
df["Age"] = df["Age"].fillna(df["Age"].median())
df["Embarked"] = df["Embarked"].fillna(df["Embarked"].mode()[0])
print("""
KEY INSIGHTS:
1. Females had a much higher survival rate than males.
2. Passengers in 1st class survived more compared to 2nd & 3rd.
3. Younger passengers had slightly higher survival.
4. Fare is positively correlated with survival.
5. Age had missing values which were filled using median.
6. No strong correlation between most numeric features.
""")
```

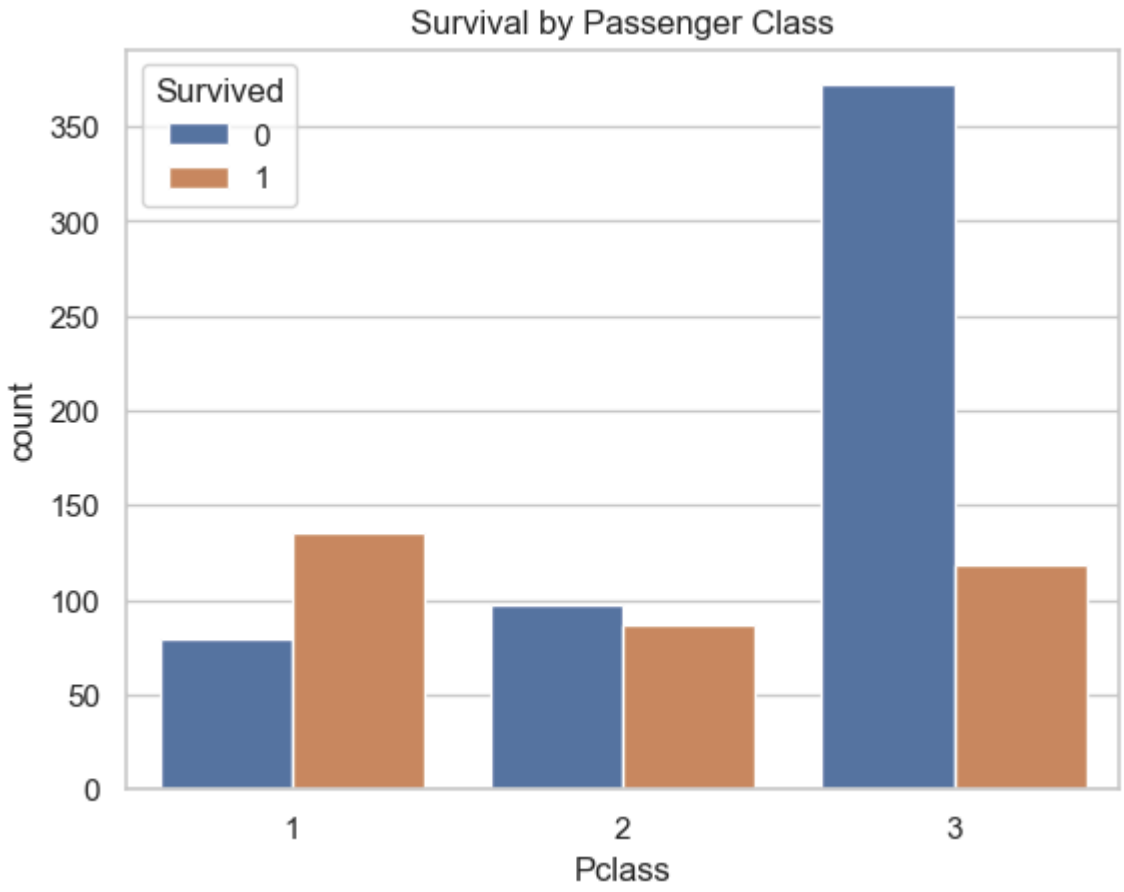
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null    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
```

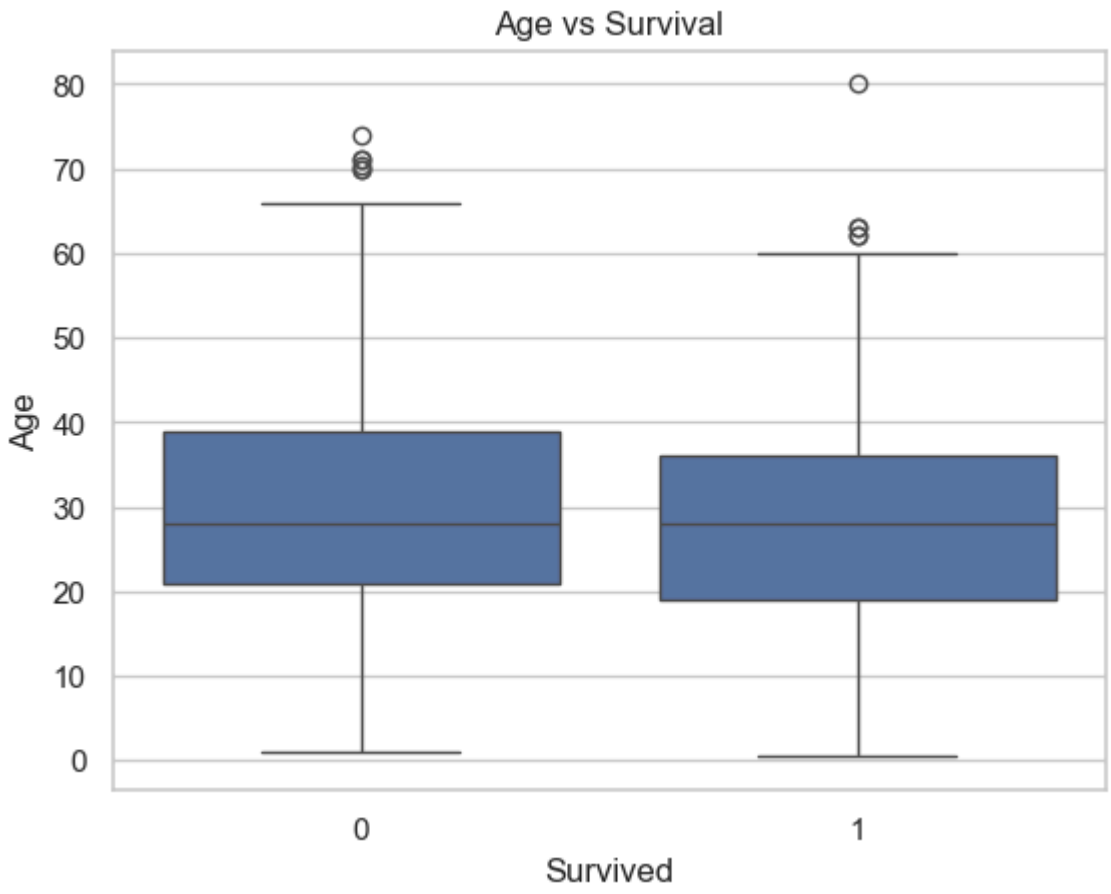


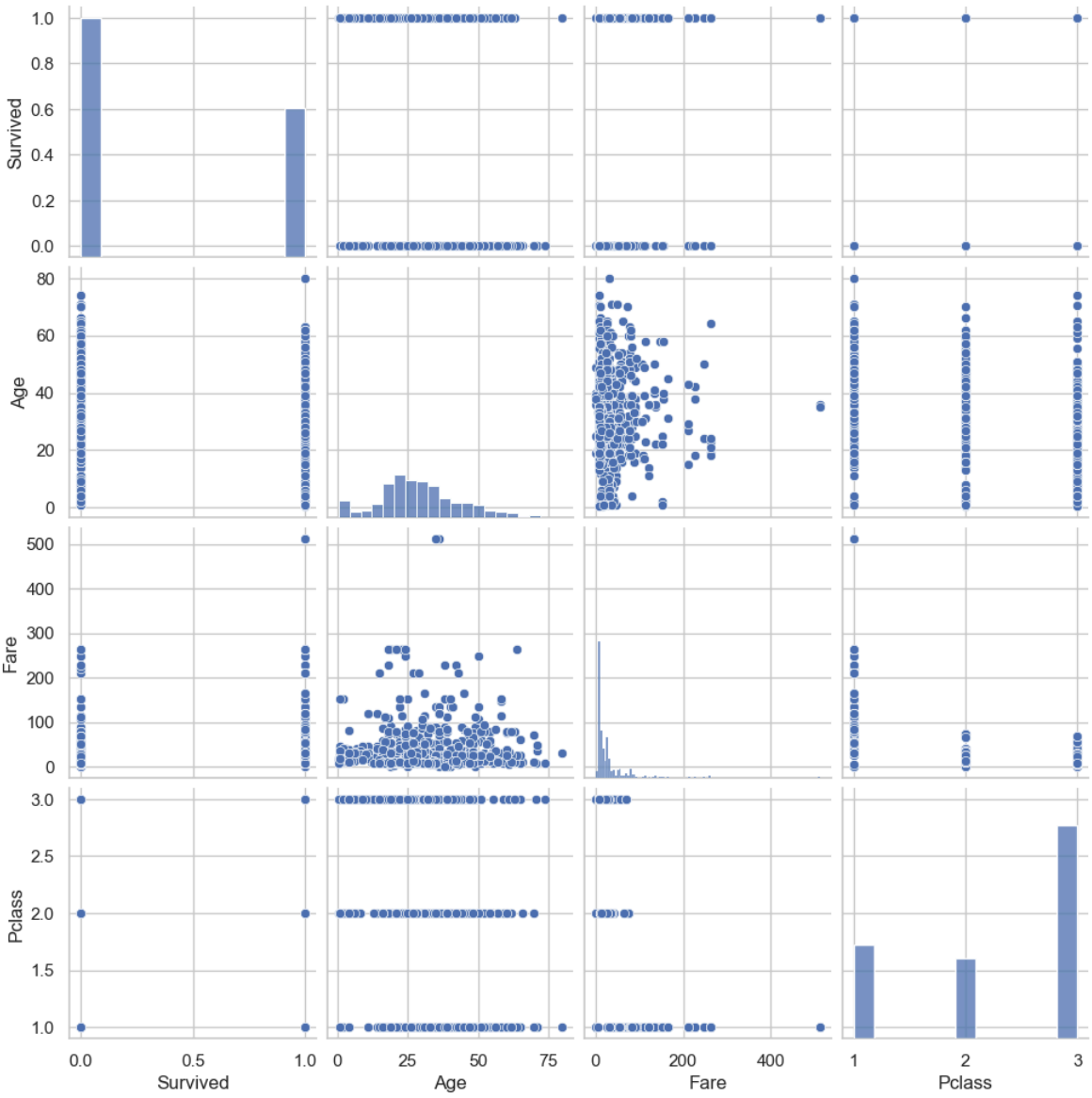

Age Distribution
Count
Age
0
20
40
60
80
100
0
10
20
30
40
50
60
70
80

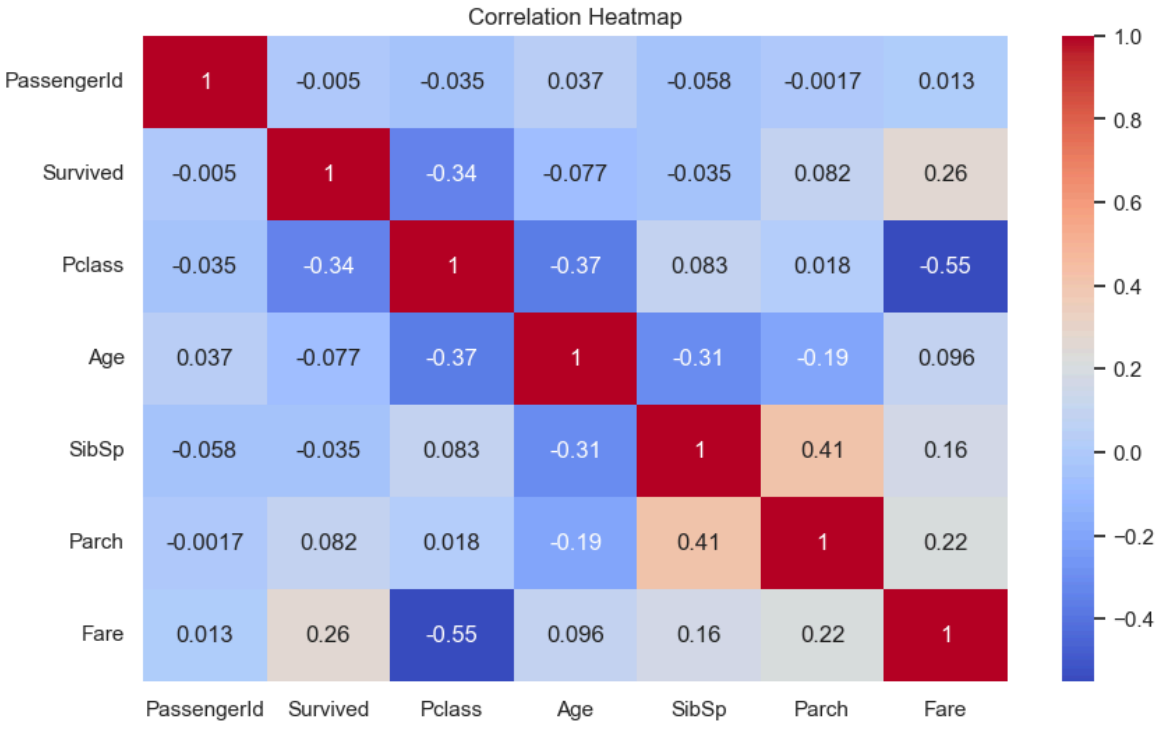

Passenger Class Count
count
Pclass
0
100
200
300
400
500
1
2
3


Survival Count by Gender
Survived
0
1
count
400
300
200
100
0
male
female
Sex


Survival by Passenger Class
Survived
0
1
count
350
300
250
200
150
100
50
0
1
2
3
Pclass


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


Survived
Age
Fare
Pclass
0.0
0.5
1.0
0
25
50
75
200
400
1
2
3


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

KEY INSIGHTS:
1. Females had a much higher survival rate than males.
2. Passengers in 1st class survived more compared to 2nd & 3rd.
3. Younger passengers had slightly higher survival.
4. Fare is positively correlated with survival.
5. Age had missing values which were filled using median.
6. No strong correlation between most numeric features.