exploratory data

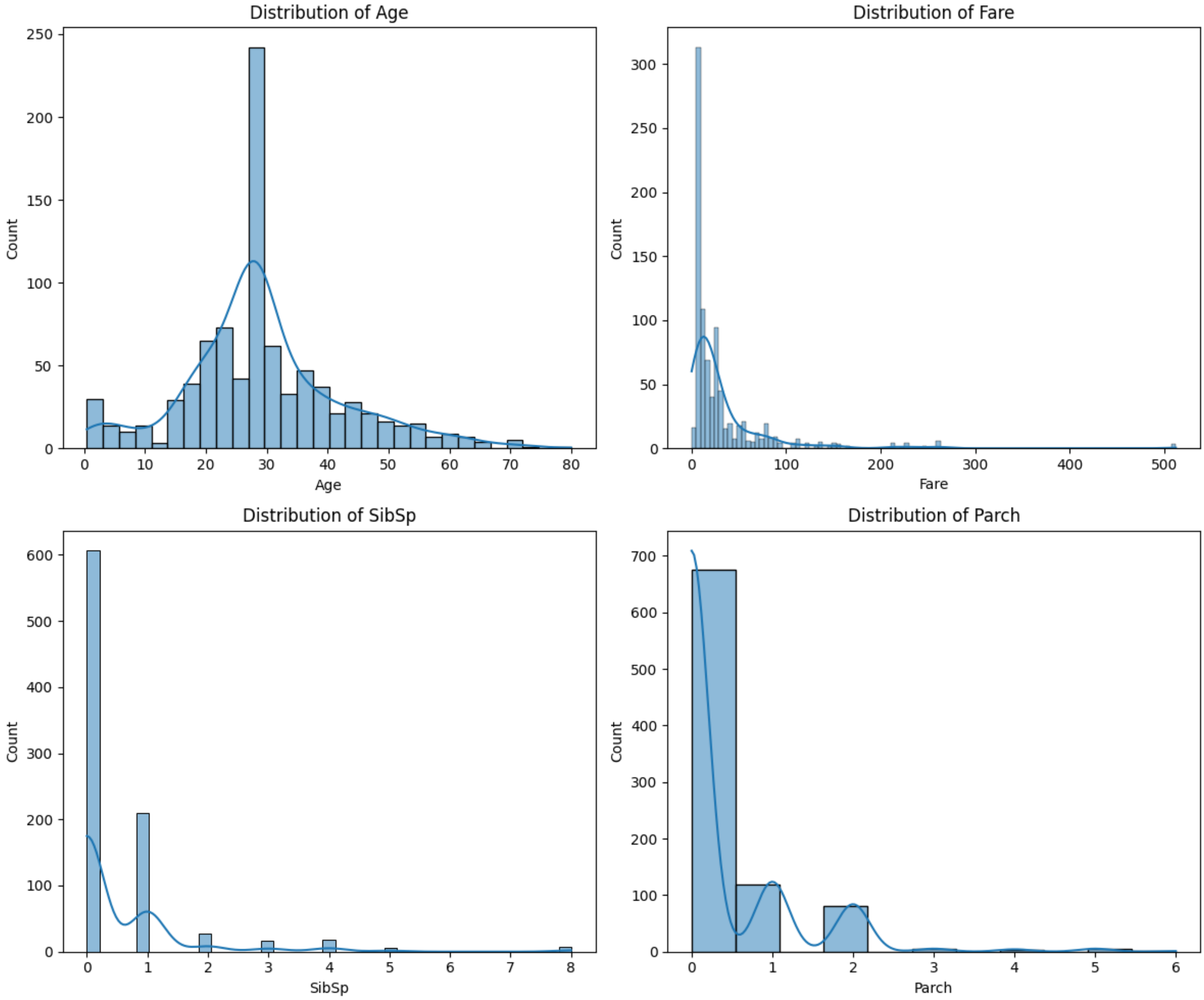
In [4]:

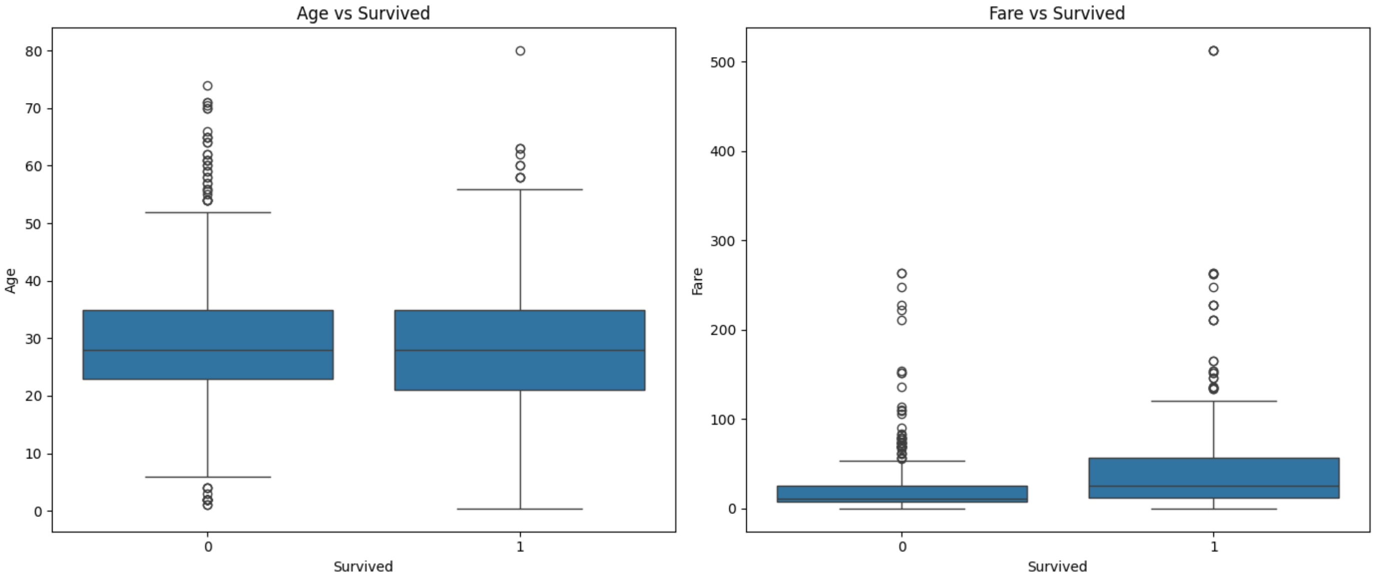
```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns


df = pd.read_csv("C:/Users/achsa/OneDrive/Desktop/project/train.csv")


df.info()


print(df.describe())


print(df['Sex'].value_counts())
print(df['Embarked'].value_counts())
print(df['Pclass'].value_counts())


df['Age'].fillna(df['Age'].median(), inplace=True)
df.dropna(subset=['Embarked'], inplace=True)

numeric_features = ['Age', 'Fare', 'SibSp', 'Parch']
plt.figure(figsize=(12, 10))
for i, col in enumerate(numeric_features):
    plt.subplot(2, 2, i + 1)
    sns.histplot(df[col], kde=True)
    plt.title(f'Distribution of {col}')
plt.tight_layout()
plt.show()


plt.figure(figsize=(14, 6))
plt.subplot(1, 2, 1)
sns.boxplot(x='Survived', y='Age', data=df)
plt.title('Age vs Survived')

plt.subplot(1, 2, 2)
sns.boxplot(x='Survived', y='Fare', data=df)
plt.title('Fare vs Survived')
plt.tight_layout()
plt.show()


plt.figure(figsize=(12, 6))
sns.countplot(x='Pclass', hue='Survived', data=df)
plt.title('Survival by Passenger Class')
plt.show()

sns.countplot(x='Sex', hue='Survived', data=df)
plt.title('Survival by Sex')
plt.show()
plt.figure(figsize=(10, 8))
sns.heatmap(df.corr(numeric_only=True), annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()


sns.pairplot(df[['Survived', 'Age', 'Fare', 'SibSp', 'Parch']], hue='Survived')
plt.suptitle('Pairplot of Features by Survival', y=1.02)
plt.show()



"""
# Observations:

- Age: Majority of passengers were between 20-40 years old.
- Fare: Most paid under $100; a few paid over $500.
- Sex: Females had significantly higher survival rates than males.
- Pclass: 1st class had the highest survival rate, followed by 2nd and 3rd.
- Heatmap: Survival is positively correlated with Fare and being female; negatively with
- SibSp/Parch: Having 1-2 relatives onboard slightly increased survival.
- Pairplot: Shows distinct clusters based on survival, especially along Age and Fare.

# Summary of Findings:

- Survival was not evenly distributed — women, younger passengers, and 1st-class travele
- Fare and class were strong indicators of survival.
- There is a need to handle missing values like `Cabin` for deeper analysis.

"""
```

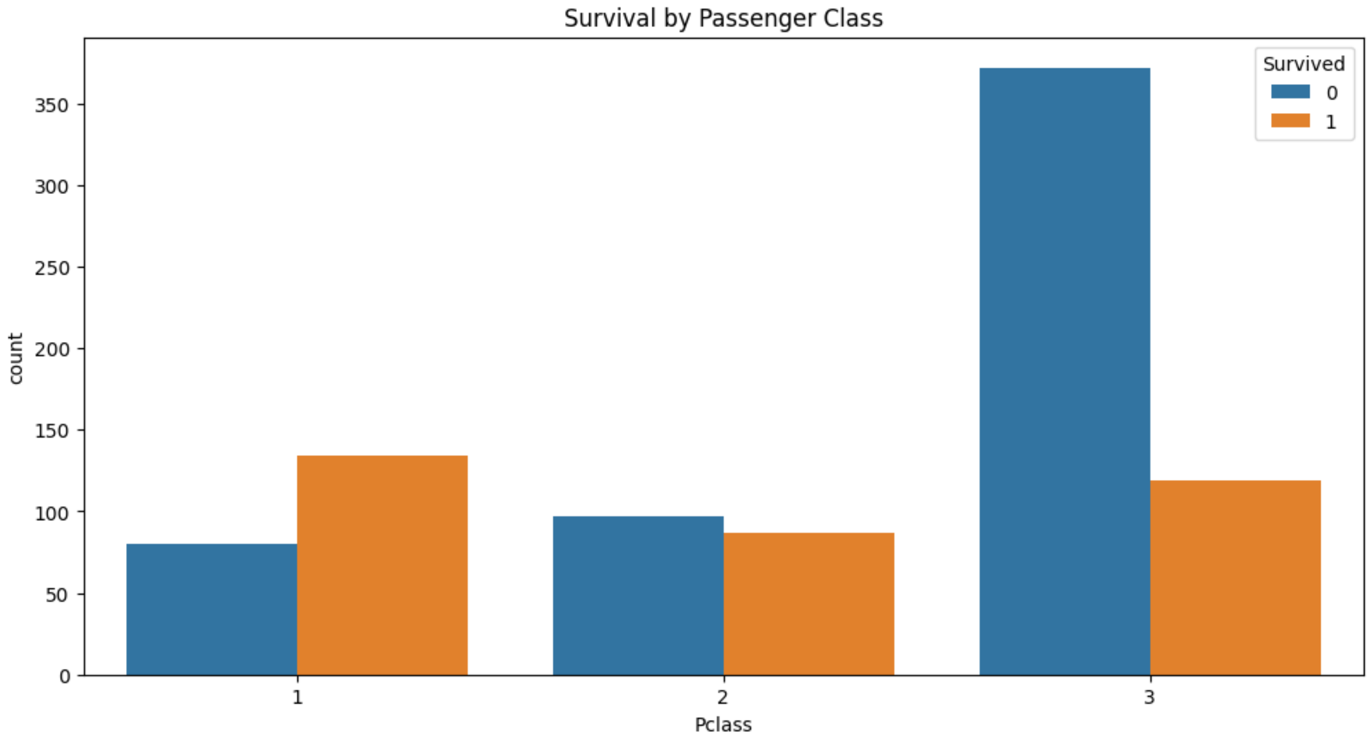
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null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

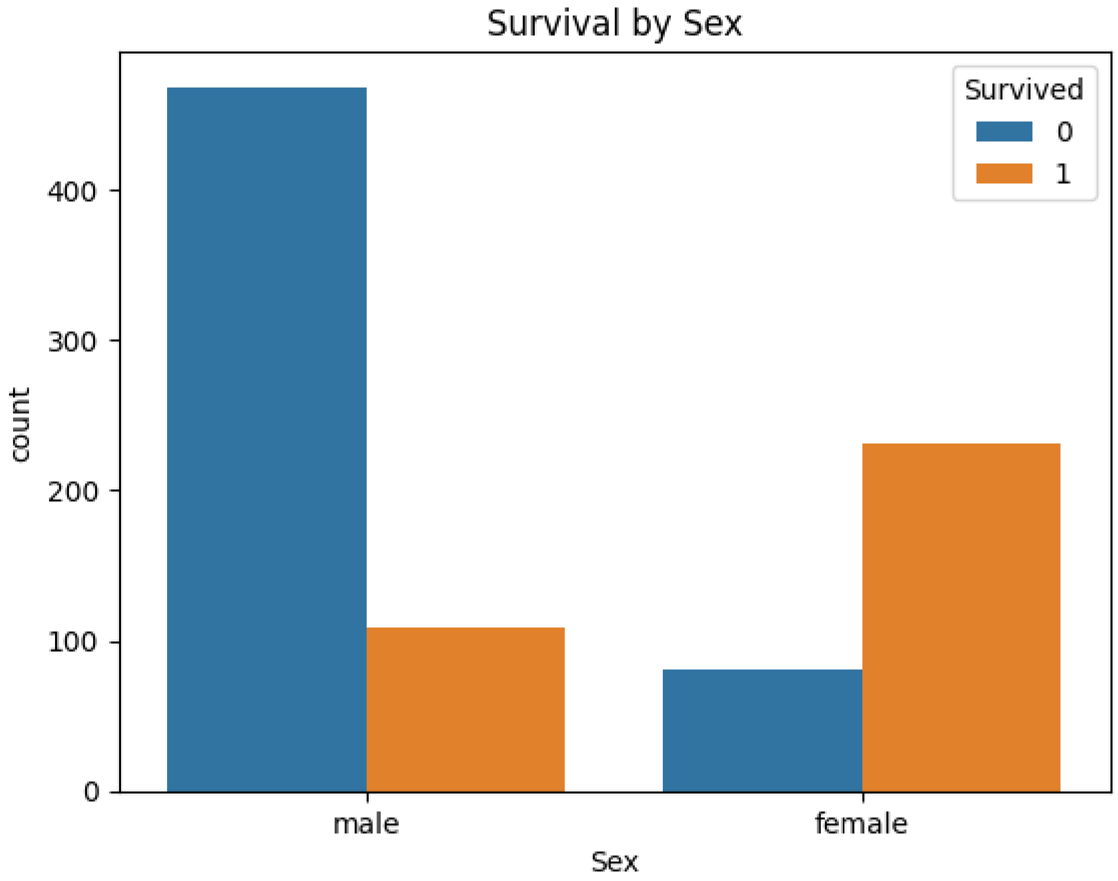
```
            Parch        Fare
count  891.000000  891.000000
mean     0.381594   32.204208
std      0.806057   49.693429
min      0.000000    0.000000
25%      0.000000    7.910400
50%      0.000000   14.454200
75%      0.000000   31.000000
max      6.000000  512.329200
Sex
male      577
female    314
Name: count, dtype: int64
Embarked
S    644
C    168
Q     77
Name: count, dtype: int64
Pclass
3    491
1    216
2    184
Name: count, dtype: int64
```

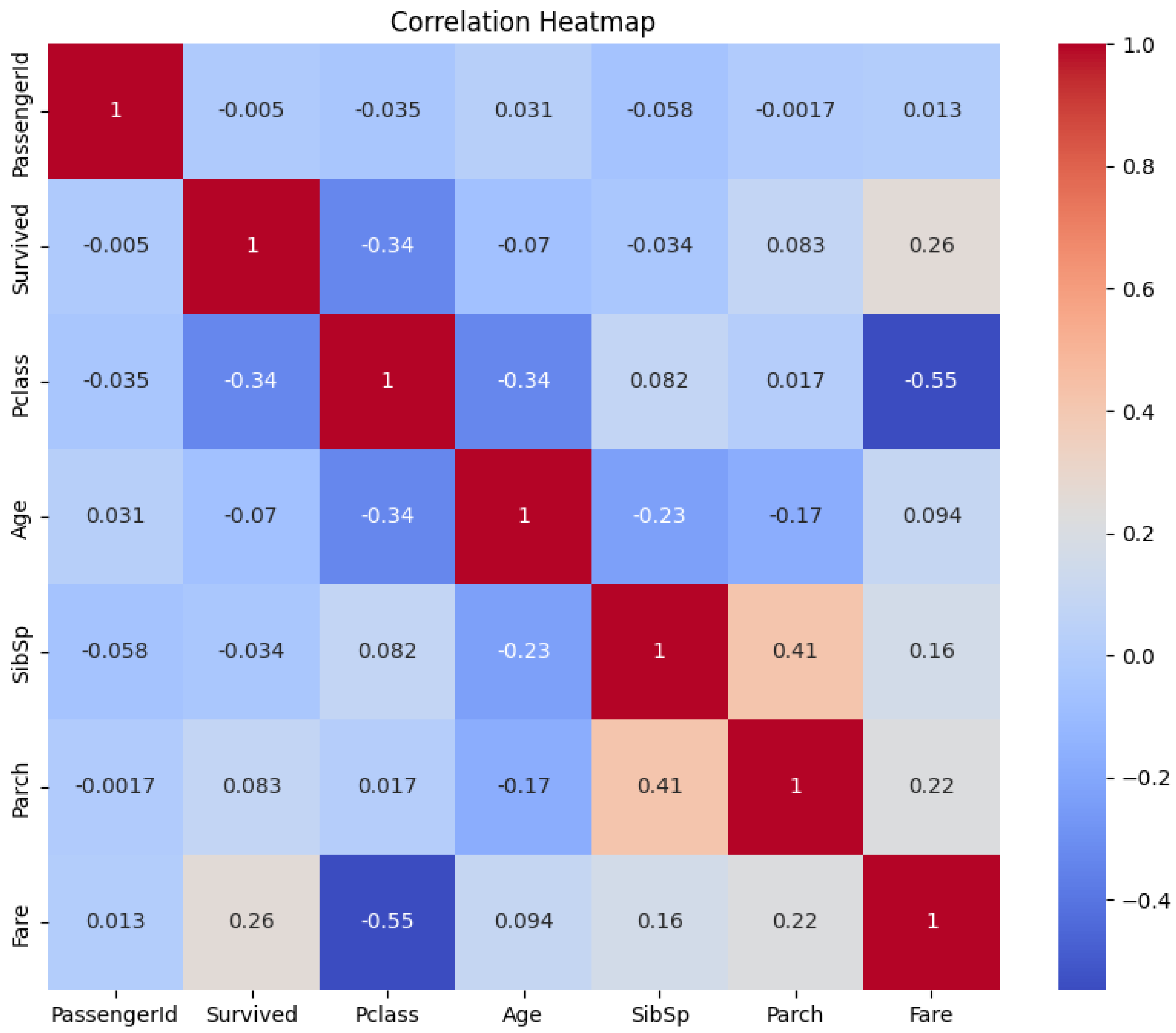
```
C:\Users\achsa\AppData\Local\Temp\ipykernel_21284\621609974.py:20: FutureWarning: A valu
e is trying to be set on a copy of a DataFrame or Series through chained assignment usin
g an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work because the
intermediate object on which we are setting values always behaves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({co
l: value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform the ope
ration inplace on the original object.


  df['Age'].fillna(df['Age'].median(), inplace=True)
```

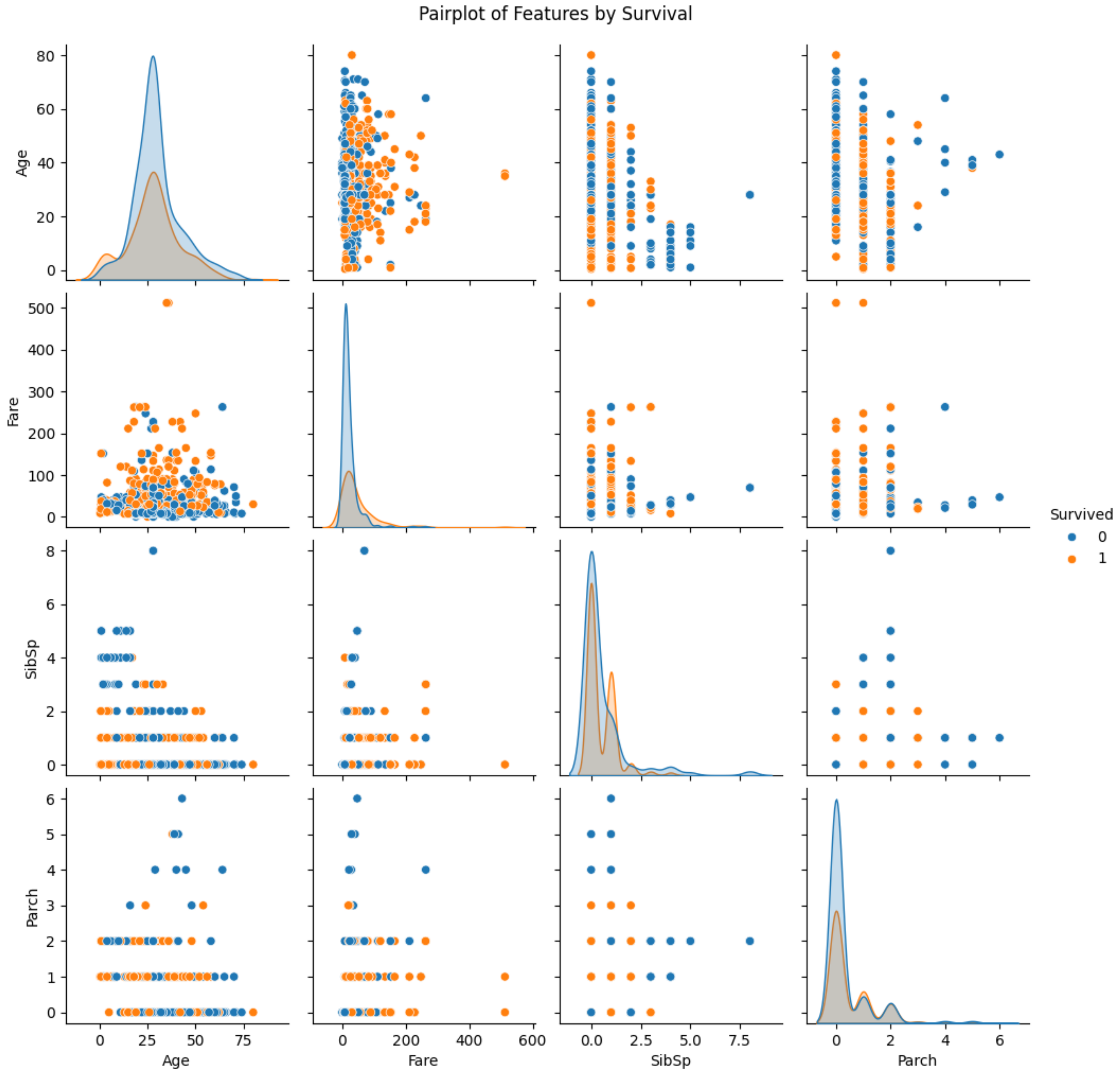

Distribution of Age
Count
250
200
150
100
50
0
0
10
20
30
40
50
60
70
80
Age
Distribution of Fare
Count
300
250
200
150
100
50
0
0
100
200
300
400
500
Fare
Distribution of SibSp
Count
600
500
400
300
200
100
0
0
1
2
3
4
5
6
7
8
SibSp
Distribution of Parch
Count
700
600
500
400
300
200
100
0
0
1
2
3
4
5
6
Parch


Age vs Survived
Age
80
70
60
50
40
30
20
10
0
0
1
Survived
Fare vs Survived
Fare
500
400
300
200
100
0
0
1
Survived


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


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


Correlation Heatmap
PassengerId
Survived
Pclass
Age
SibSp
Parch
Fare
1 -0.005 -0.035 0.031 -0.058 -0.0017 0.013
-0.005 1 -0.34 -0.07 -0.034 0.083 0.26
-0.035 -0.34 1 -0.34 0.082 0.017 -0.55
0.031 -0.07 -0.34 1 -0.23 -0.17 0.094
-0.058 -0.034 0.082 -0.23 1 0.41 0.16
-0.0017 0.083 0.017 -0.17 0.41 1 0.22
0.013 0.26 -0.55 0.094 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



Out[4]:

```
'\n# Observations:\n\n- Age: Majority of passengers were between 20-40 years old.\n- Far
e: Most paid under $100; a few paid over $500.\n- Sex: Females had significantly higher
survival rates than males.\n- Pclass: 1st class had the highest survival rate, followed
by 2nd and 3rd.\n- Heatmap: Survival is positively correlated with Fare and being femal
e; negatively with Pclass.\n- SibSp/Parch: Having 1-2 relatives onboard slightly increas
ed survival.\n- Pairplot: Shows distinct clusters based on survival, especially along Ag
e and Fare.\n\n# Summary of Findings:\n\n- Survival was not evenly distributed — women,
younger passengers, and 1st-class travelers were more likely to survive.\n- Fare and cla
ss were strong indicators of survival.\n- There is a need to handle missing values like
`Cabin` for deeper analysis.\n\n'
```