Titanic Dataset - Exploratory Data Analysis

Step 1: Import necessary libraries

We import the essential libraries needed for EDA:

- **pandas** for data handling
- **matplotlib.pyplot** and **seaborn** for data visualization
- `%matplotlib inline` to display plots inside the notebook

In [17]:
```
# Importing libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# To display plots inside the notebook
%matplotlib inline
```

In [18]:
```
df.describe()
```

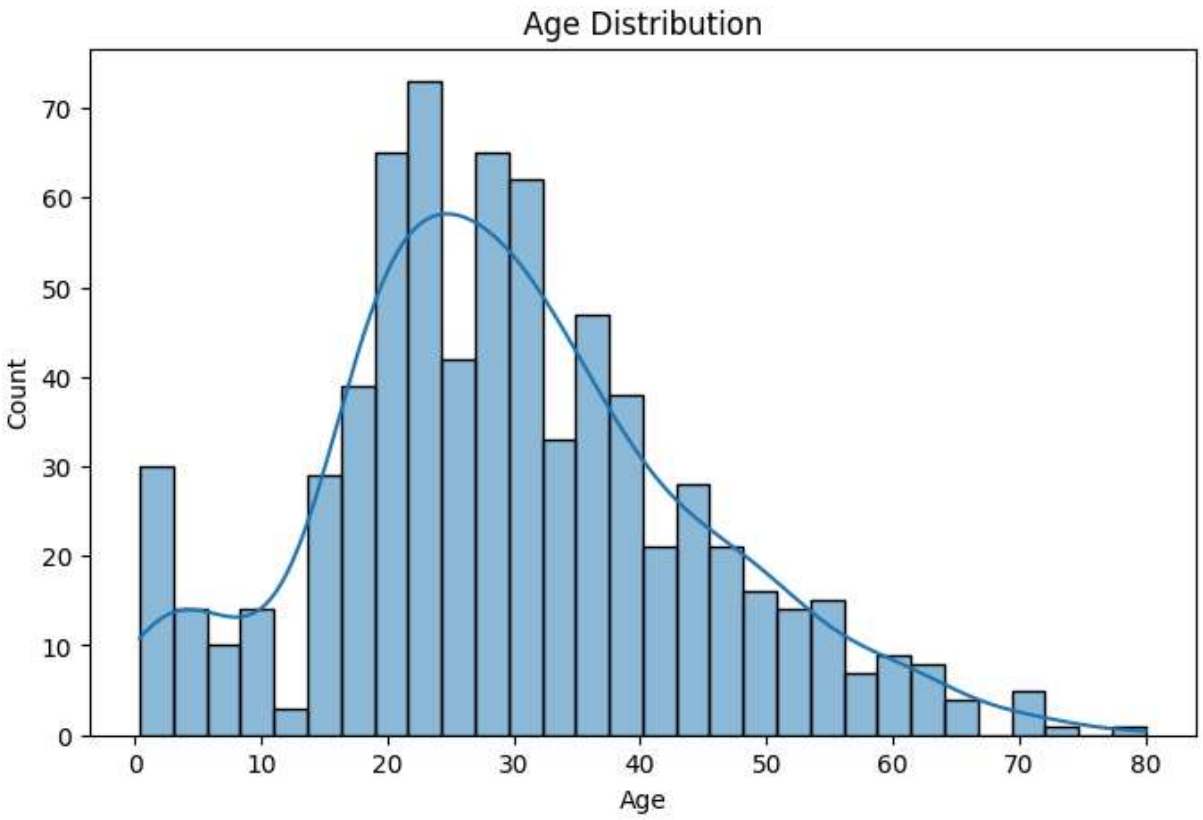
Out[18]:

	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 [6]:
```
# Value counts for categorical columns
print("Sex Value Counts:\n", df['Sex'].value_counts())
print("\nPclass Value Counts:\n", df['Pclass'].value_counts())
print("\nEmbarked Value Counts:\n", df['Embarked'].value_counts())
```

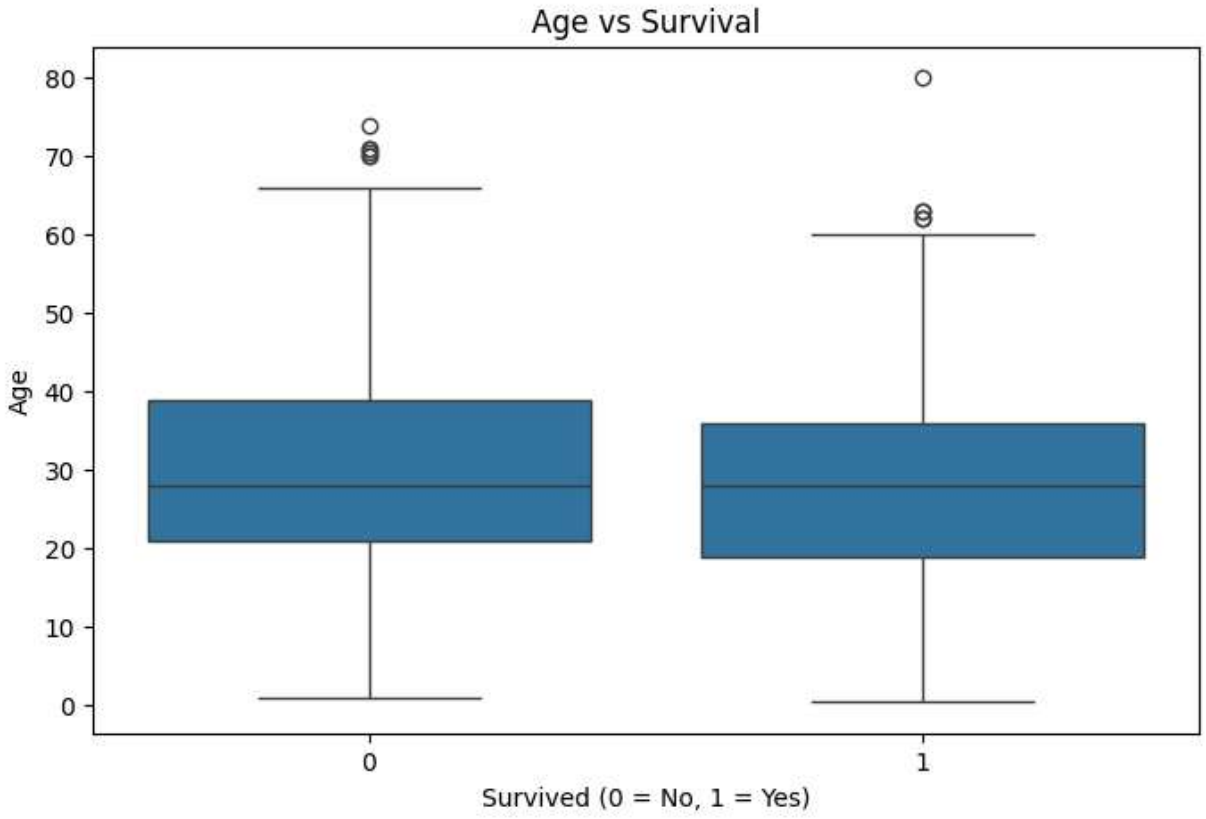
```
Sex Value Counts:
 Sex
male      577
female    314
Name: count, dtype: int64

Pclass Value Counts:
 Pclass
3    491
1    216
2    184
Name: count, dtype: int64

Embarked Value Counts:
 Embarked
S    644
C    168
Q     77
Name: count, dtype: int64
```

In [15]:
```
# Plotting a histogram for Age distribution
plt.figure(figsize=(8,5))
sns.histplot(df['Age'], kde=True, bins=30)
plt.title('Age Distribution')
plt.xlabel('Age')
plt.ylabel('Count')
plt.show()
```

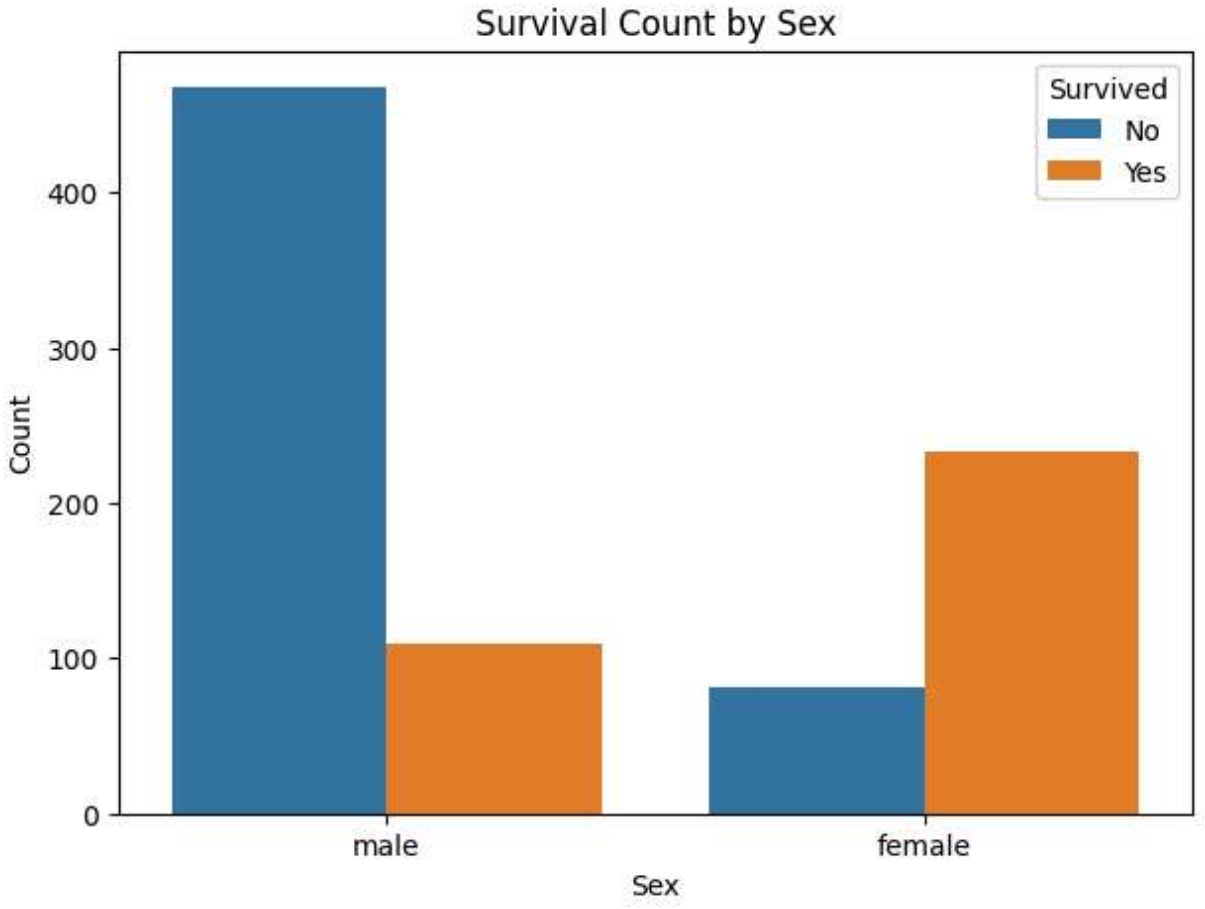


In [8]:
```
# Boxplot: Age vs Survived
plt.figure(figsize=(8,5))
sns.boxplot(x='Survived', y='Age', data=df)
```

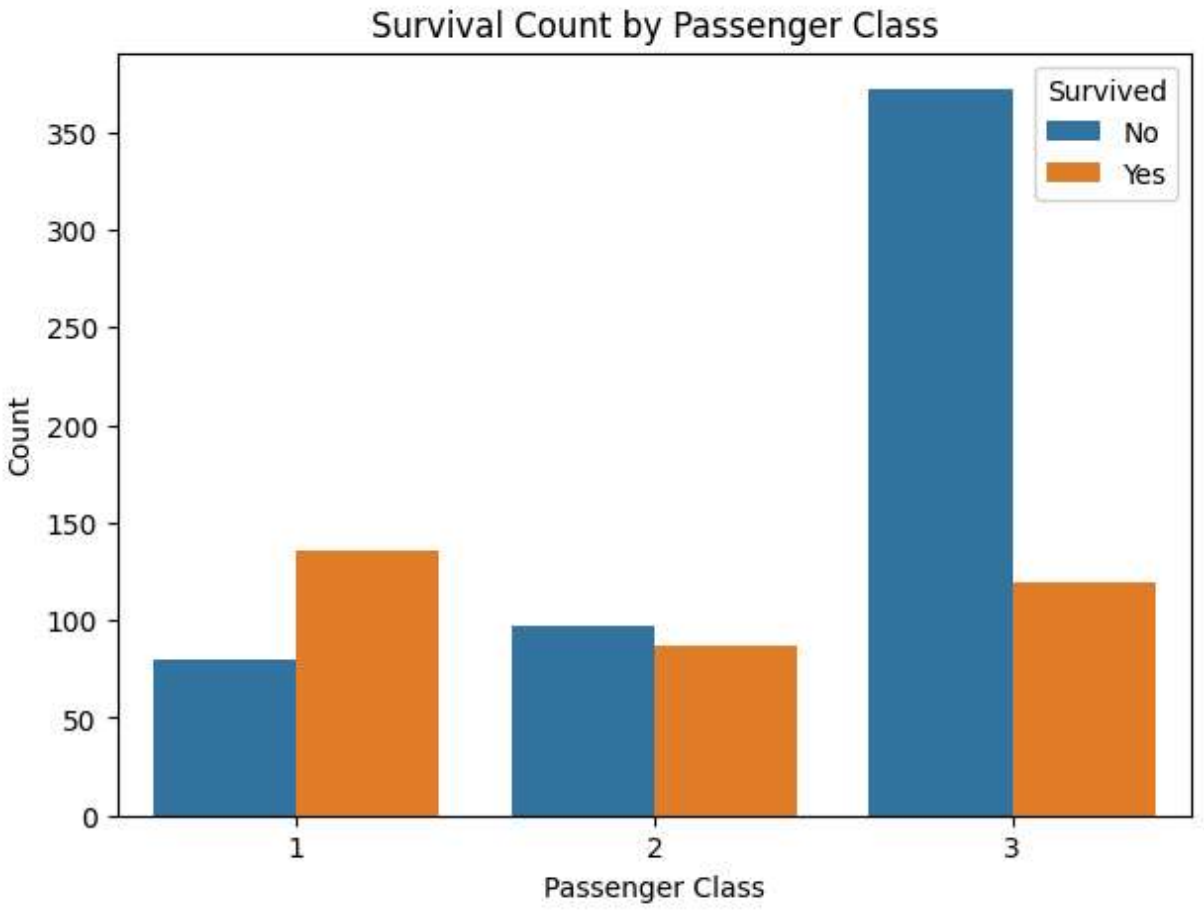
```
plt.title('Age vs Survival')
plt.xlabel('Survived (0 = No, 1 = Yes)')
plt.ylabel('Age')
plt.show()
```



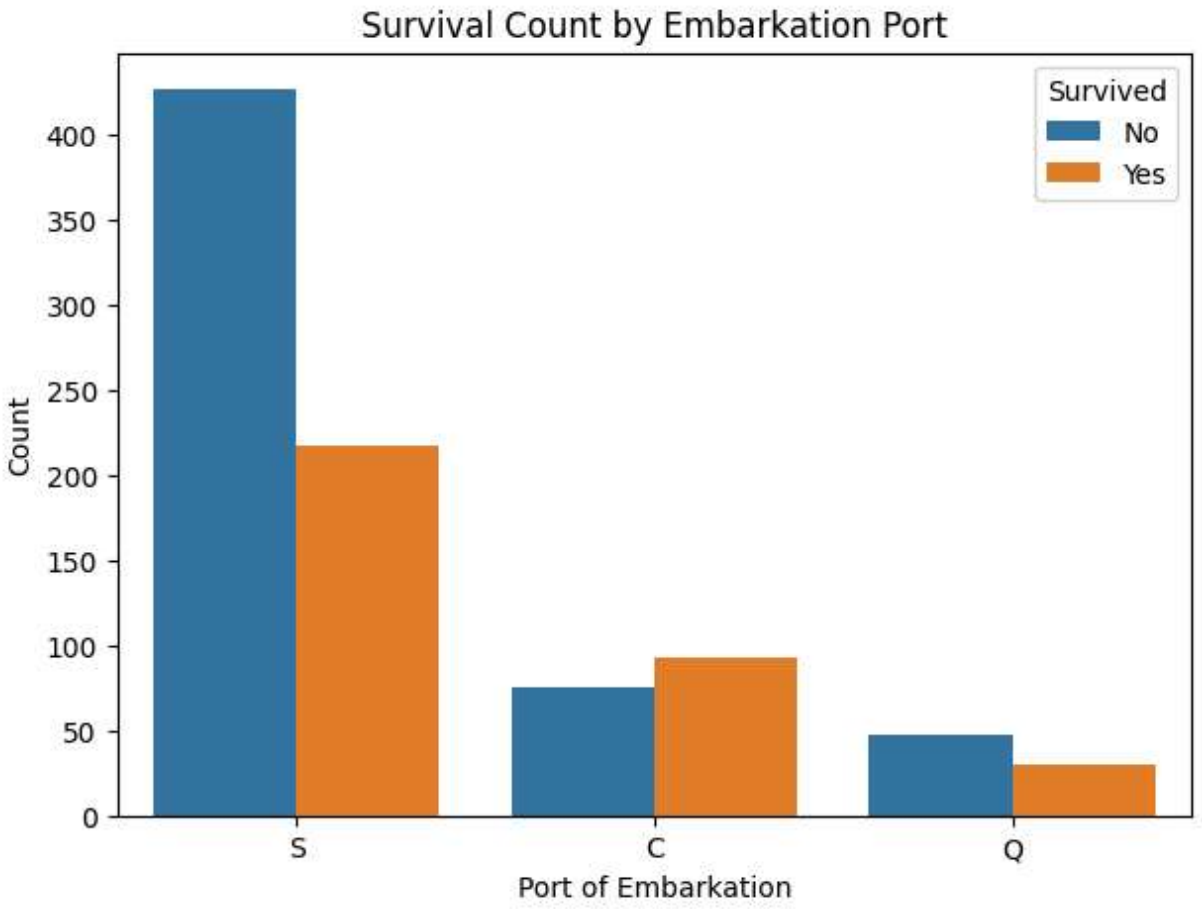
In [9]:
```
# Countplot: Sex vs Survival
plt.figure(figsize=(7,5))
sns.countplot(x='Sex', hue='Survived', data=df)
plt.title('Survival Count by Sex')
plt.xlabel('Sex')
plt.ylabel('Count')
plt.legend(title='Survived', labels=['No', 'Yes'])
plt.show()
```



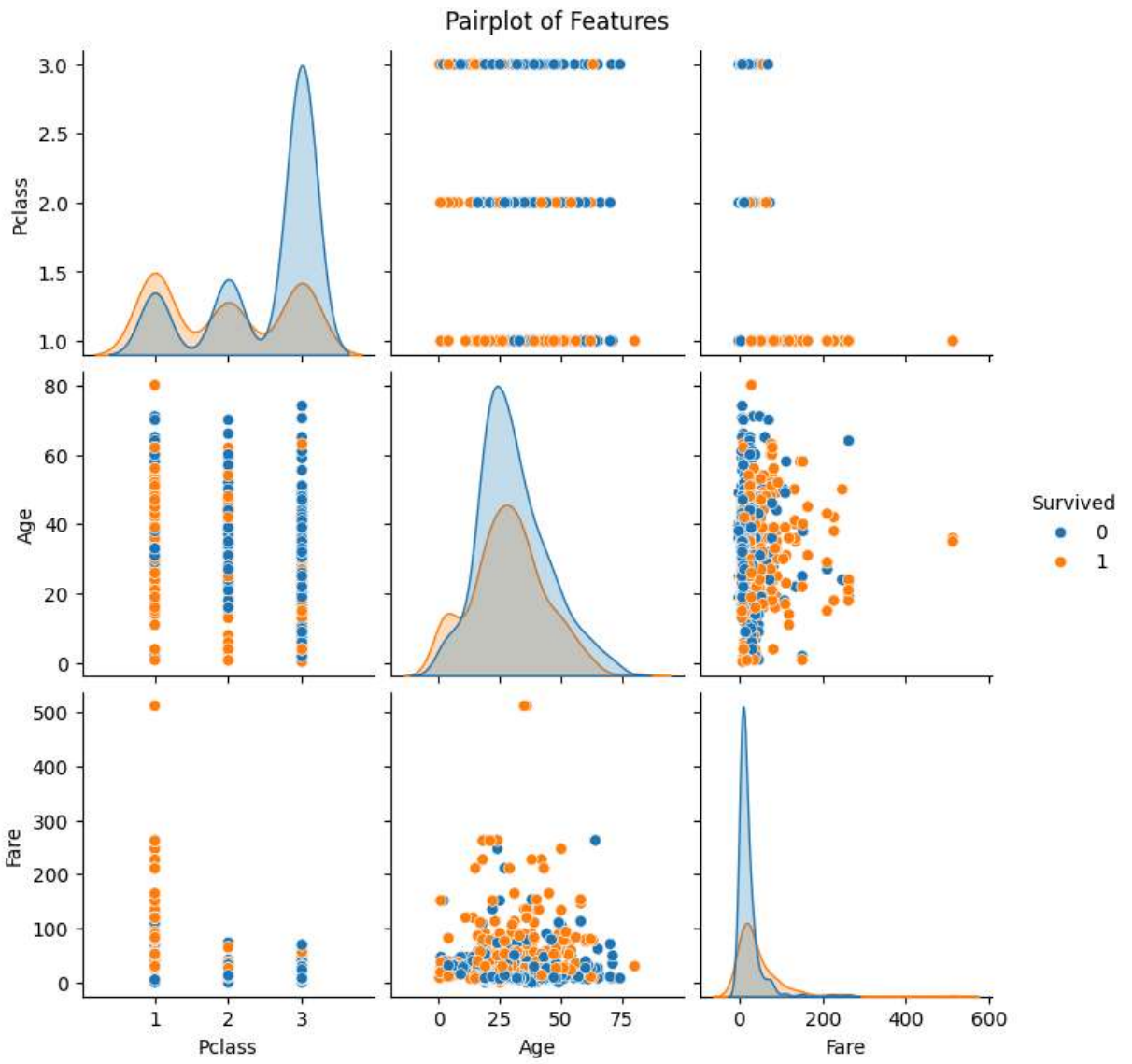
In [7]:
```
# Countplot: Pclass vs Survival
plt.figure(figsize=(7,5))
sns.countplot(x='Pclass', hue='Survived', data=df)
plt.title('Survival Count by Passenger Class')
plt.xlabel('Passenger Class')
plt.ylabel('Count')
plt.legend(title='Survived', labels=['No', 'Yes'])
plt.show()
```



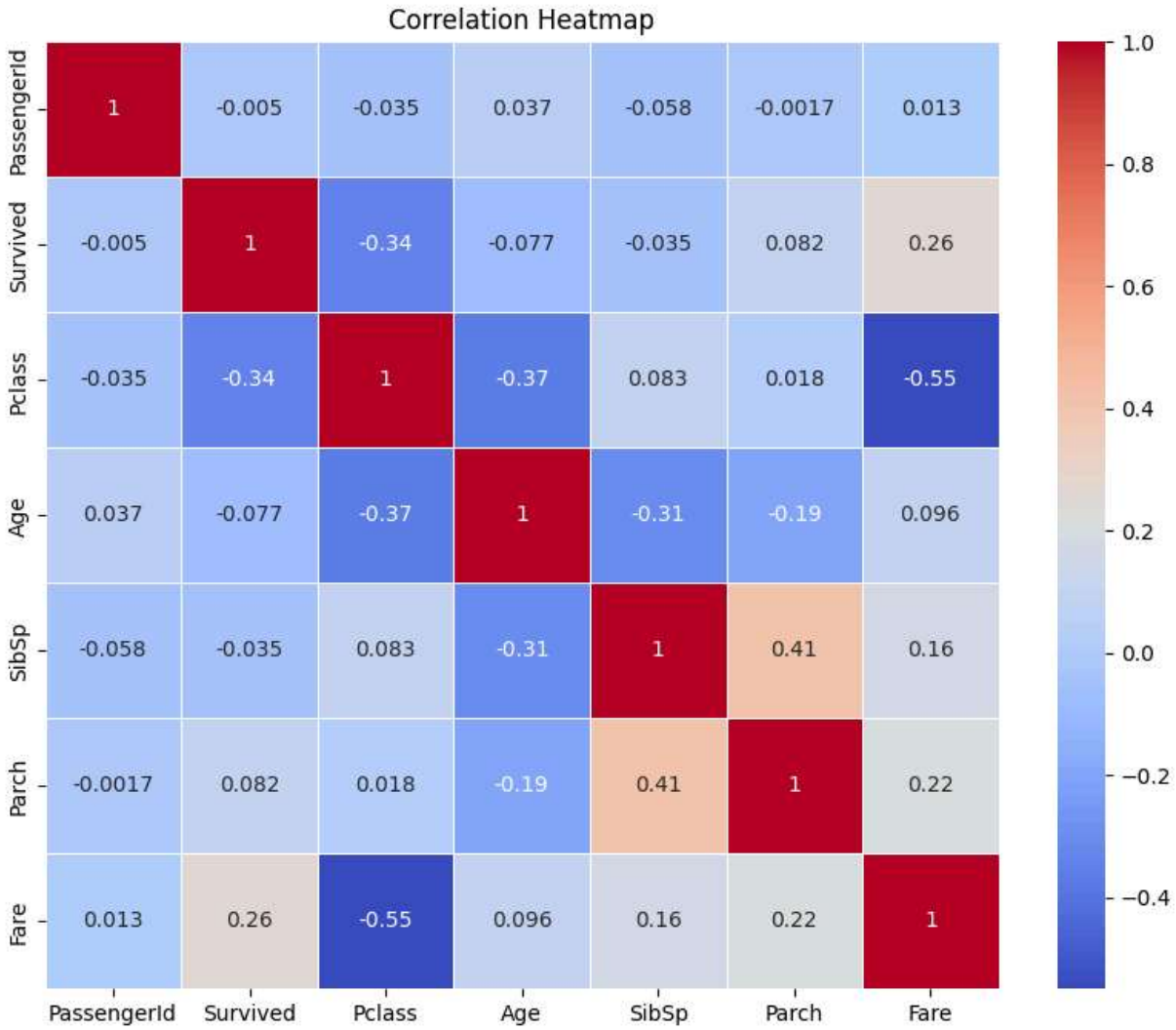
In [6]:
```
import pandas as pd

url = 'https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.cs
df = pd.read_csv(url)

# ab plot kar sakte ho
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(7,5))
sns.countplot(x='Embarked', hue='Survived', data=df)
plt.title('Survival Count by Embarkation Port')
plt.xlabel('Port of Embarkation')
plt.ylabel('Count')
plt.legend(title='Survived', labels=['No', 'Yes'])
plt.show()
```



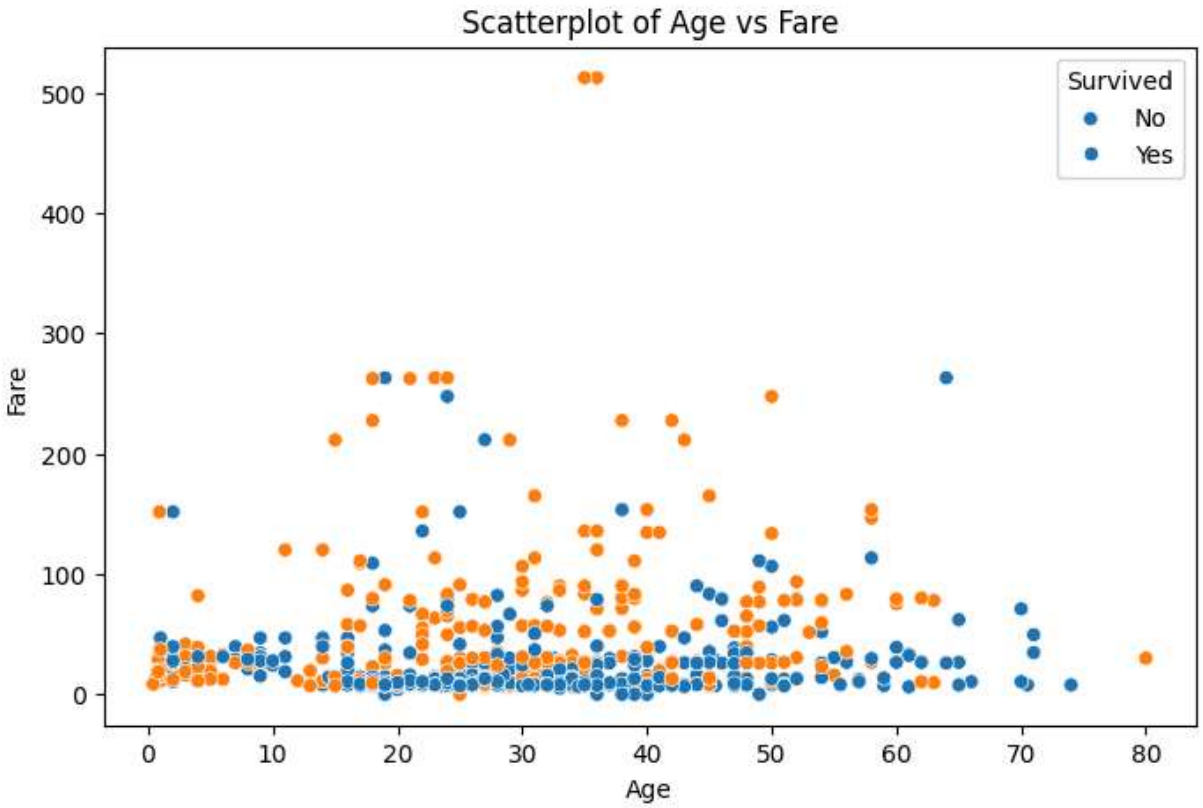
In [8]:
```
# Pairplot of selected features
sns.pairplot(df[['Survived', 'Pclass', 'Sex', 'Age', 'Fare']], hue='Survived', diag
plt.suptitle('Pairplot of Features', y=1.02)
plt.show()
```



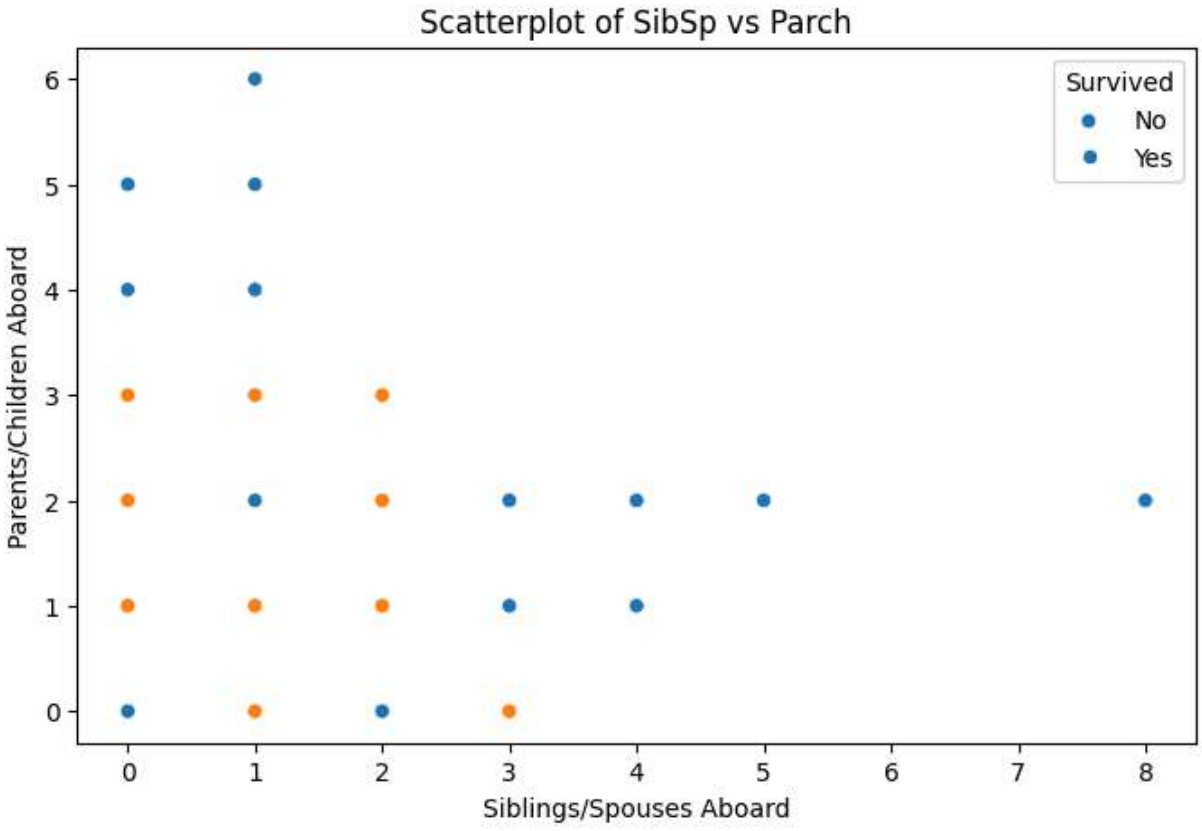
In [10]:
```
# Correlation Heatmap (numeric columns only)
plt.figure(figsize=(10,8))
numeric_df = df.select_dtypes(include=['float64', 'int64'])  # Only numeric columns
sns.heatmap(numeric_df.corr(), annot=True, cmap='coolwarm', linewidths=0.5)
plt.title('Correlation Heatmap')
plt.show()
```



In [11]:
```
plt.figure(figsize=(8,5))
sns.scatterplot(x='Age', y='Fare', hue='Survived', data=df)
plt.title('Scatterplot of Age vs Fare')
plt.xlabel('Age')
plt.ylabel('Fare')
plt.legend(title='Survived', labels=['No', 'Yes'])
plt.show()
```



In [12]:
```
plt.figure(figsize=(8,5))
sns.scatterplot(x='SibSp', y='Parch', hue='Survived', data=df)
plt.title('Scatterplot of SibSp vs Parch')
plt.xlabel('Siblings/Spouses Aboard')
plt.ylabel('Parents/Children Aboard')
plt.legend(title='Survived', labels=['No', 'Yes'])
plt.show()
```



Summary of Insights:

- Female passengers had a much higher survival rate compared to males.
- Passengers from higher classes (Pclass 1) survived more often.
- Ports of embarkation (C, S, Q) show slight survival variations, but C had better survival rates.
- Younger passengers and those who paid higher fares had higher survival probabilities.
- Family members traveling together (small family groups) had slightly better chances to survive.