In [18]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Settings for prettier plots
sns.set(style="whitegrid")
%matplotlib inline
```

In [15]:
```
# Load Titanic dataset
file_names =[r'C:\Users\raham\Downloads\titanic\gender_submission.csv',r'C:\User
dfs = [pd.read_csv(file) for file in file_names]
```

In [16]:
```
dfs
```

Out[16]:
```
[     PassengerId  Survived
0            892         0
1            893         1
2            894         0
3            895         0
4            896         1
..           ...       ...
413         1305         0
414         1306         1
415         1307         0
416         1308         0
417         1309         0

[418 rows x 2 columns],
     PassengerId  Pclass                                          Name  \
0            892       3                              Kelly, Mr. James
1            893       3              Wilkes, Mrs. James (Ellen Needs)
2            894       2                     Myles, Mr. Thomas Francis
3            895       3                              Wirz, Mr. Albert
4            896       3  Hirvonen, Mrs. Alexander (Helga E Lindqvist)
..           ...     ...                                           ...
413         1305       3                           Spector, Mr. Woolf
414         1306       1                 Oliva y Ocana, Dona. Fermina
415         1307       3                 Saether, Mr. Simon Sivertsen
416         1308       3                          Ware, Mr. Frederick
417         1309       3                     Peter, Master. Michael J

        Sex   Age  SibSp  Parch              Ticket      Fare Cabin Embarked
0      male  34.5      0      0              330911    7.8292   NaN        Q
1    female  47.0      1      0              363272    7.0000   NaN        S
2      male  62.0      0      0              240276    9.6875   NaN        Q
3      male  27.0      0      0              315154    8.6625   NaN        S
4    female  22.0      1      1             3101298   12.2875   NaN        S
..      ...   ...    ...    ...                 ...       ...   ...      ...
413    male   NaN      0      0           A.5. 3236    8.0500   NaN        S
414  female  39.0      0      0            PC 17758  108.9000  C105        C
415    male  38.5      0      0  SOTON/O.Q. 3101262    7.2500   NaN        S
416    male   NaN      0      0              359309    8.0500   NaN        S
417    male   NaN      1      1                2668   22.3583   NaN        C

[418 rows x 11 columns],
     PassengerId  Survived  Pclass  \
0              1         0       3
1              2         1       1
2              3         1       3
3              4         1       1
4              5         0       3
..           ...       ...     ...
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         0       3

                                                 Name     Sex   Age  SibSp  \
0                               Braund, Mr. Owen Harris    male  22.0      1
1    Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1
2                                Heikkinen, Miss. Laina  female  26.0      0
3         Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                              Allen, Mr. William Henry    male  35.0      0
```

```
..                                                 ...     ...   ...    ...
886                            Montvila, Rev. Juozas    male  27.0      0
887                     Graham, Miss. Margaret Edith  female  19.0      0
888         Johnston, Miss. Catherine Helen "Carrie"  female   NaN      1
889                            Behr, Mr. Karl Howell    male  26.0      0
890                              Dooley, Mr. Patrick    male  32.0      0

     Parch            Ticket     Fare Cabin Embarked
0        0         A/5 21171   7.2500   NaN        S
1        0          PC 17599  71.2833   C85        C
2        0  STON/O2. 3101282   7.9250   NaN        S
3        0            113803  53.1000  C123        S
4        0            373450   8.0500   NaN        S
..     ...               ...      ...   ...      ...
886      0            211536  13.0000   NaN        S
887      0            112053  30.0000   B42        S
888      2        W./C. 6607  23.4500   NaN        S
889      0            111369  30.0000  C148        C
890      0            370376   7.7500   NaN        Q

[891 rows x 12 columns]]
```

In [21]:
```
# merge all files into one file
df = pd.concat(dfs,ignore_index= True)
```

In [47]:
```
df.tail()
```

Out[47]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fa
1722	887	No	2nd class	Montvila, Rev. Juozas	male	27.0	0.0	0.0	211536	13.
1723	888	Yes	1st class	Graham, Miss. Margaret Edith	female	19.0	0.0	0.0	112053	30.
1724	889	No	3rd class	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1.0	2.0	W./C. 6607	23.
1725	890	Yes	1st class	Behr, Mr. Karl Howell	male	26.0	0.0	0.0	111369	30.
1726	891	No	3rd class	Dooley, Mr. Patrick	male	32.0	0.0	0.0	370376	7.

In [30]:
```
# Data information
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1727 entries, 0 to 1726
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  1727 non-null   int64  
 1   Survived     1309 non-null   float64
 2   Pclass       1309 non-null   float64
 3   Name         1309 non-null   object 
 4   Sex          1309 non-null   object 
 5   Age          1046 non-null   float64
 6   SibSp        1309 non-null   float64
 7   Parch        1309 non-null   float64
 8   Ticket       1309 non-null   object 
 9   Fare         1308 non-null   float64
 10  Cabin        295 non-null    object 
 11  Embarked     1307 non-null   object 
dtypes: float64(6), int64(1), object(5)
memory usage: 162.0+ KB
```

In [31]:
```
# Statistical summary
df.describe()
```

Out[31]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	1
count	1727.000000	1309.000000	1309.000000	1046.000000	1309.000000	1309.000000	
mean	762.828025	0.377387	2.294882	29.881138	0.498854	0.385027	
std	385.032264	0.484918	0.837836	14.413493	1.041658	0.865560	
min	1.000000	0.000000	1.000000	0.170000	0.000000	0.000000	
25%	432.500000	0.000000	2.000000	21.000000	0.000000	0.000000	
50%	864.000000	0.000000	3.000000	28.000000	0.000000	0.000000	
75%	1093.500000	1.000000	3.000000	39.000000	1.000000	0.000000	
max	1309.000000	1.000000	3.000000	80.000000	8.000000	9.000000	

In [32]:
```
# Checking missing values
df.isnull().sum()
```

Out[32]:
```
PassengerId       0
Survived        418
Pclass          418
Name            418
Sex             418
Age             681
SibSp           418
Parch           418
Ticket          418
Fare            419
Cabin          1432
Embarked        420
dtype: int64
```

In [44]:
```
# Value counts of important columns
print(df['Survived'].value_counts())
```

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

```
Survived
No     815
Yes    494
Name: count, dtype: int64
Pclass
3rd class    709
1st class    323
2nd class    277
Name: count, dtype: int64
Sex
male      843
female    466
Name: count, dtype: int64
```

In [39]:
```
# Observation: Majority of passengers are in 3rd class.
# Observation: Majority of passengers are died than survived.
# Observation: Majority of passengers are Male than Females.
```

In [41]:
```
df['Survived'] = df['Survived'].replace({0: 'No', 1: 'Yes'})
```

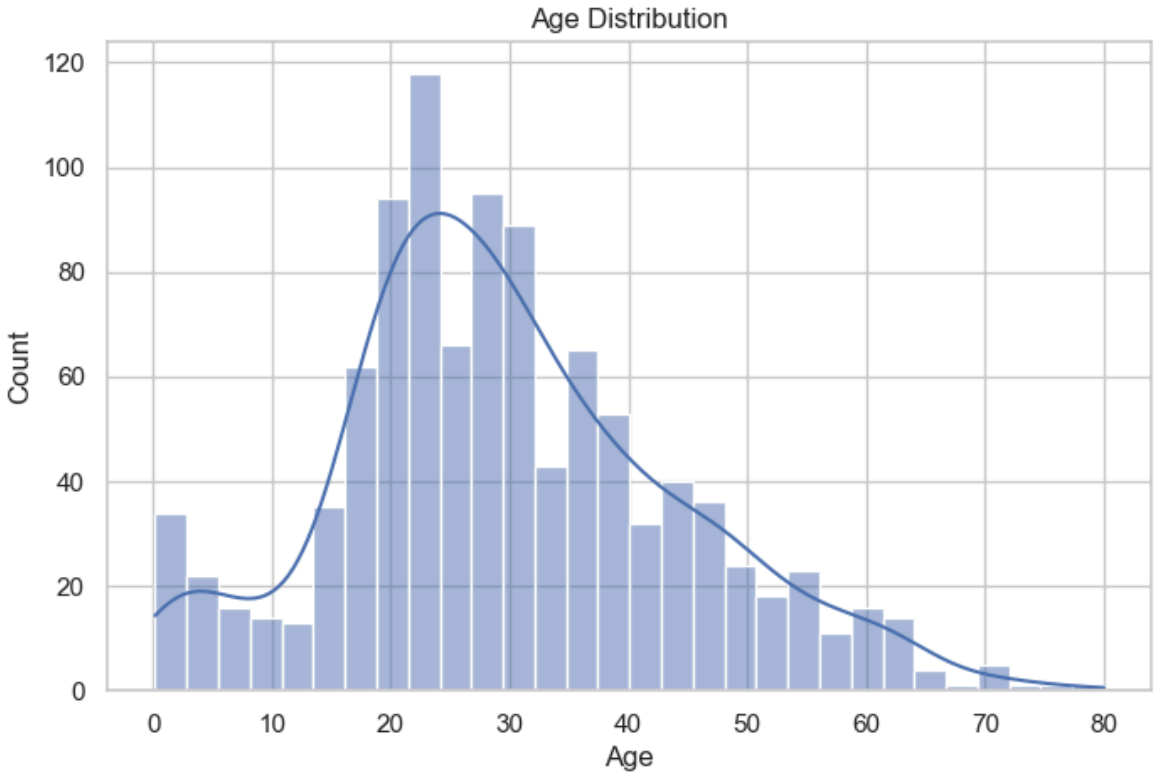
In [43]:
```
df['Pclass'] = df['Pclass'].replace({1: '1st class', 2: '2nd class', 3: '3rd cla
```

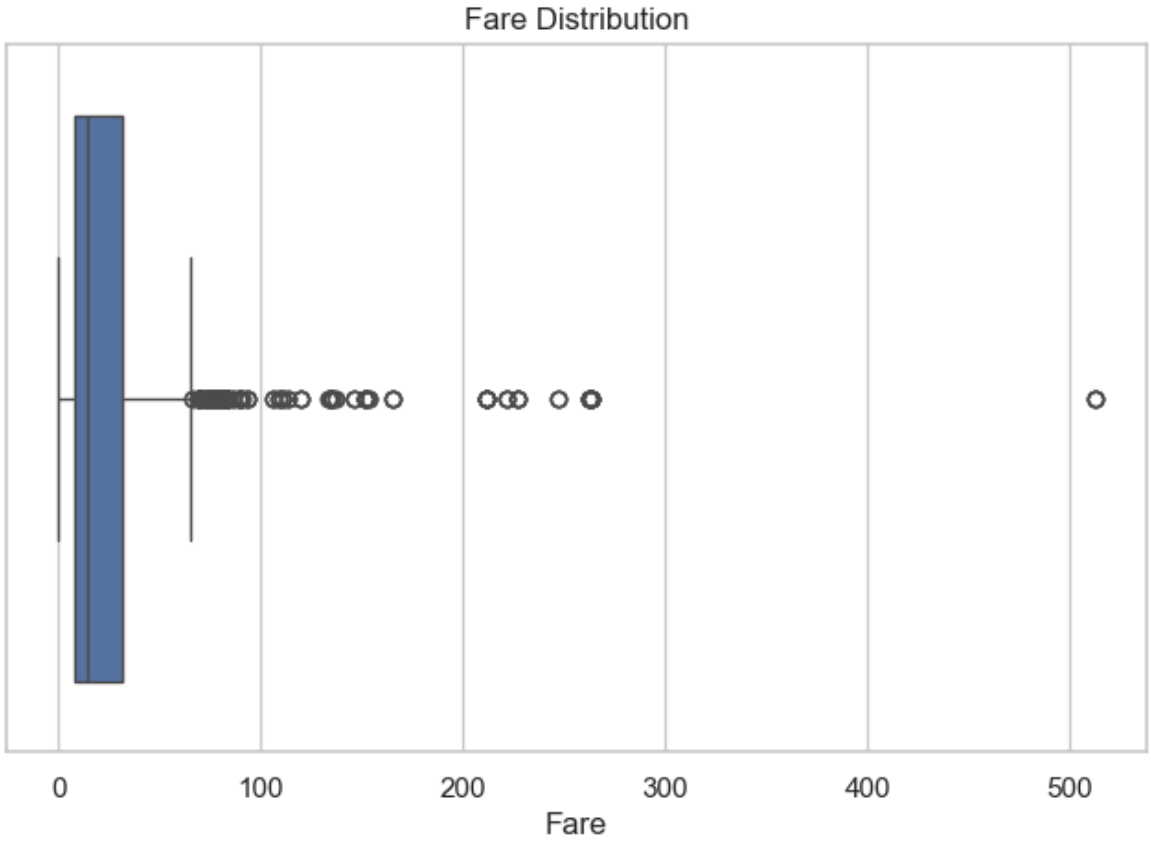
In [46]:
```
df['Embarked'] = df['Embarked'].replace({'C': 'Cherbourg', 'Q': 'Queenstown', 'S
```

In [48]:
```
# Histogram for Age
plt.figure(figsize=(8,5))
sns.histplot(df['Age'].dropna(), bins=30, kde=True)
plt.title('Age Distribution')

# Boxplot for Fare
plt.figure(figsize=(8,5))
sns.boxplot(x='Fare', data=df)
plt.title('Fare Distribution')
```

Out[48]:
```
Text(0.5, 1.0, 'Fare Distribution')
```

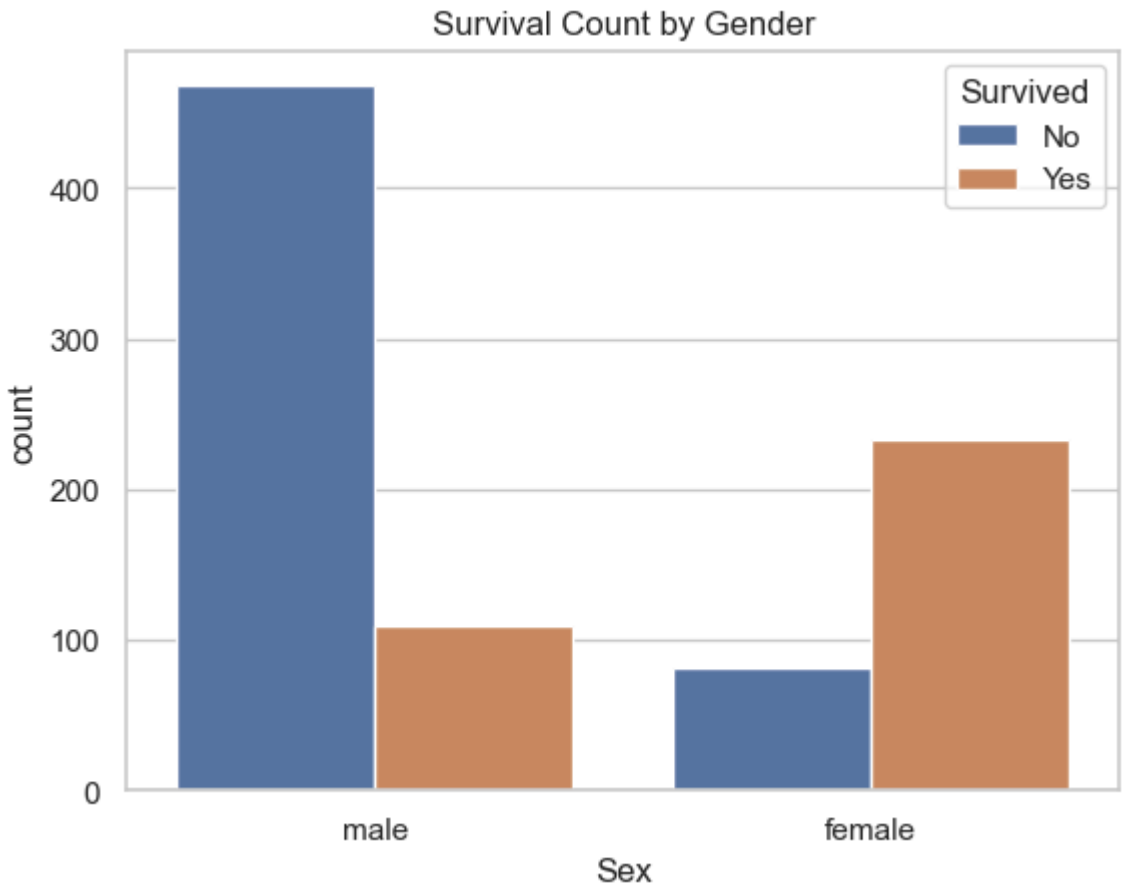




In [49]:
```
# Age Distribution: The majority of Titanic passengers were young adults between
# Fare Distribution: Most Titanic passengers paid low fares (under 50), but a fe
```

In [50]:
```
# Survival based on Gender
sns.countplot(x='Sex', hue='Survived', data=df)
plt.title('Survival Count by Gender')
```

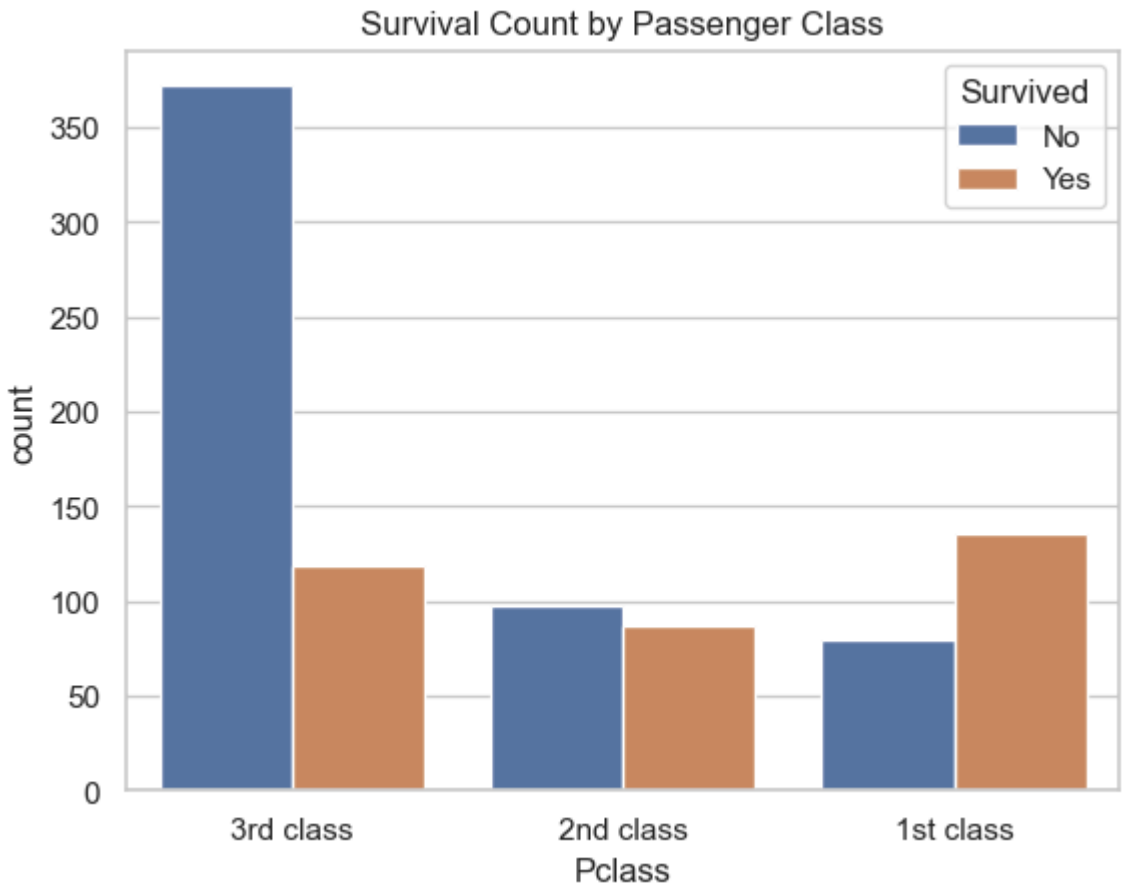
Out[50]:
```
Text(0.5, 1.0, 'Survival Count by Gender')
```



In [51]:
```
# Being female increased the chance of survival on the Titanic.
# Most males did not survive.
```

In [52]:
```
# Survival based on Passenger Class
sns.countplot(x='Pclass', hue='Survived', data=df)
plt.title('Survival Count by Passenger Class')
```

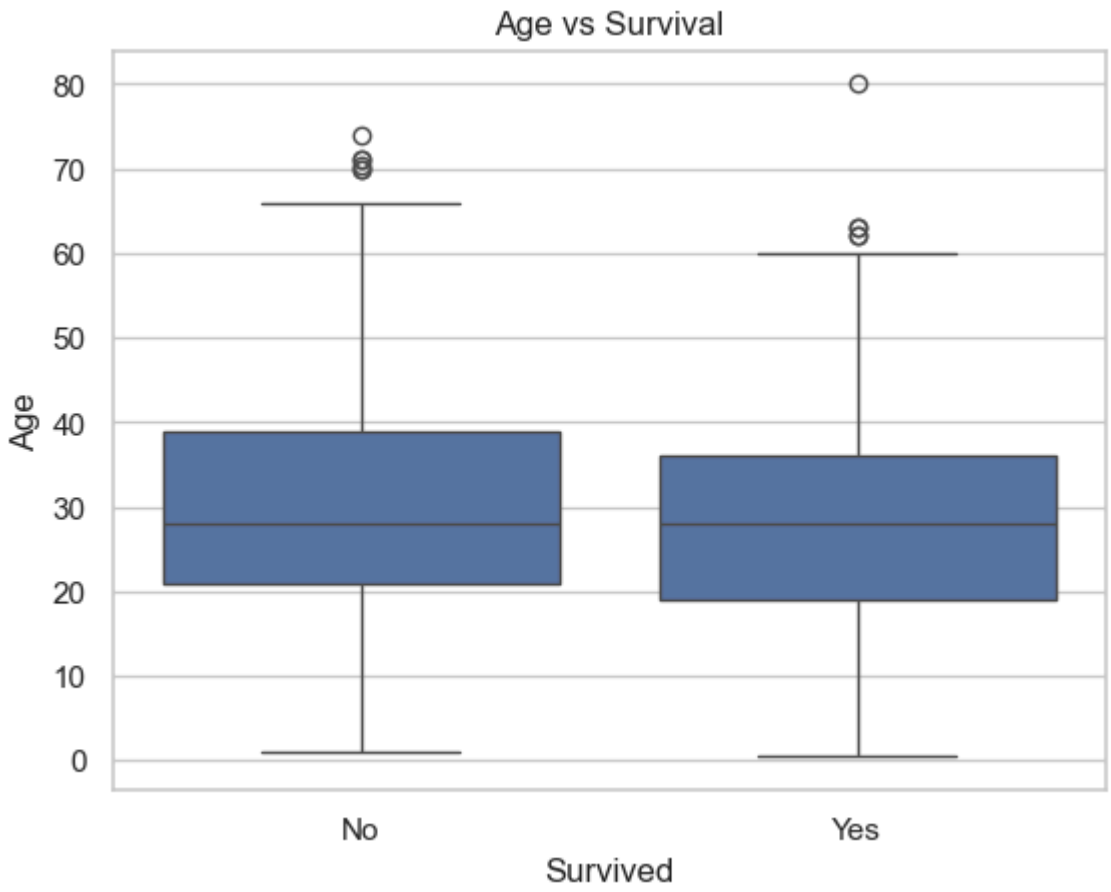
Out[52]:
```
Text(0.5, 1.0, 'Survival Count by Passenger Class')
```



In [53]:
```
# Higher class = Higher chance of survival on the Titanic.
# 3rd class passengers were most affected with more deaths.
```

In [54]:
```
# Boxplot of Age by Survival
sns.boxplot(x='Survived', y='Age', data=df)
plt.title('Age vs Survival')
```

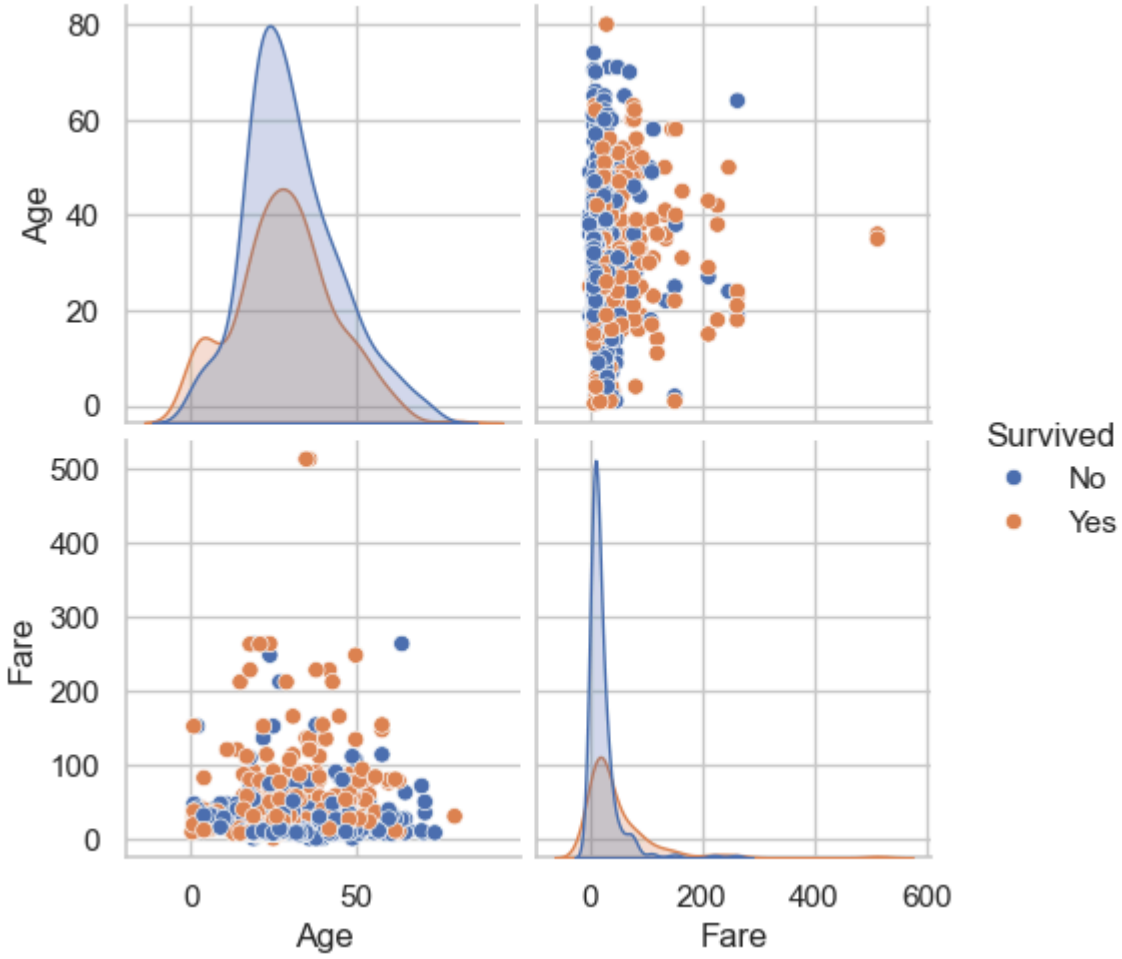
Out[54]:
```
Text(0.5, 1.0, 'Age vs Survival')
```



In [55]:
```
# Younger passengers had a slightly better chance of survival.
# Older passengers were more likely to not survive.
```

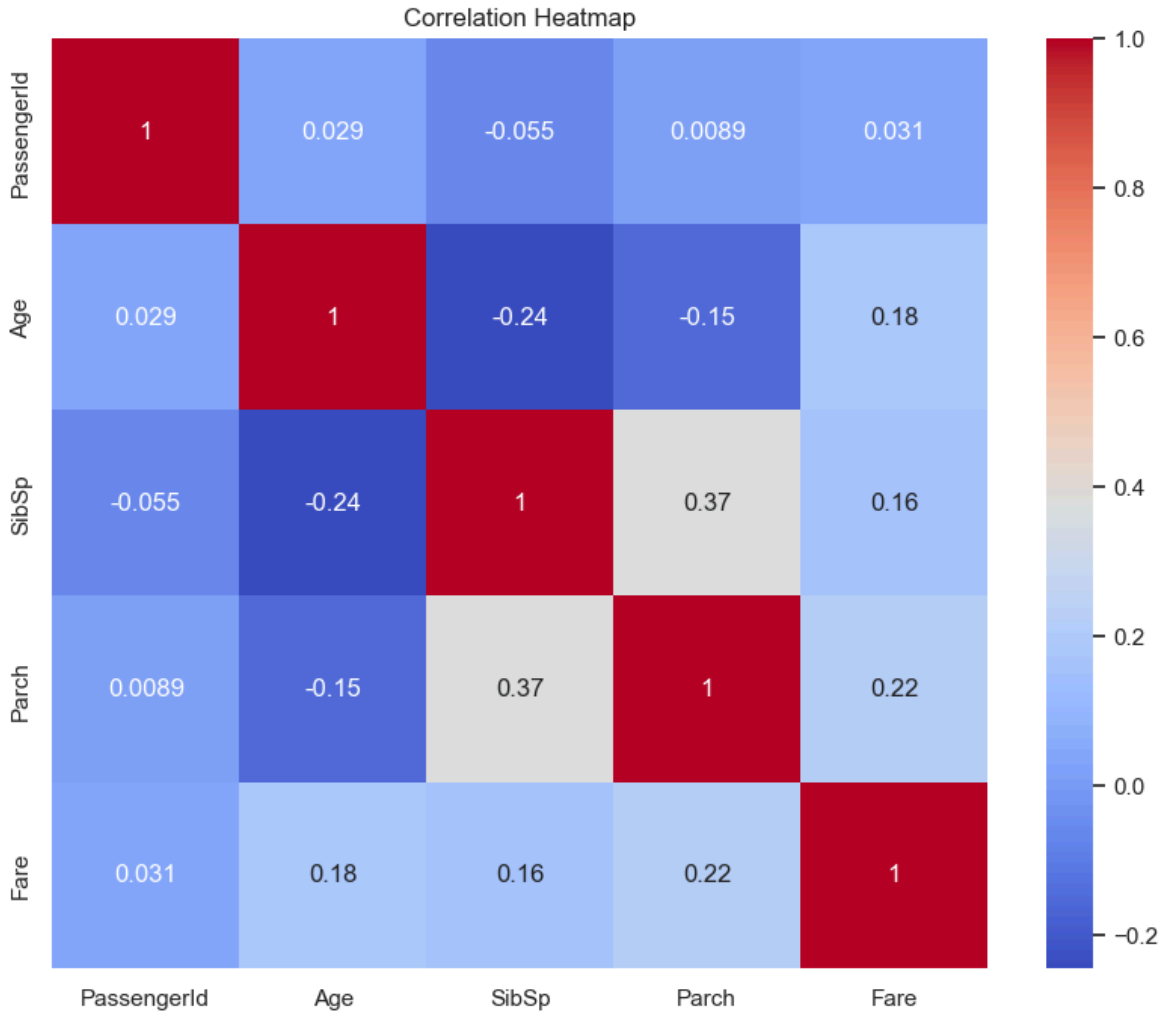
In [58]:
```
selected_cols = ['Survived', 'Pclass', 'Sex', 'Age', 'Fare']
sns.pairplot(df[selected_cols], hue='Survived')
```

Out[58]: <seaborn.axisgrid.PairGrid at 0x20c3bbe5130>



In [59]:
```
# Higher the fare paid, better the chance of survival.
# Age has some effect, but fare seems more influential in survival chances.
```

In [61]:
```
plt.figure(figsize=(10,8))
sns.heatmap(df.select_dtypes(include=['float64', 'int64']).corr(), annot=True, c
plt.title('Correlation Heatmap')
plt.show()
```



In [62]:
```
# All correlation values are weak (none are near 1 or -1).
# Therefore, no strong linear relationships are observed among these features.
```

In [63]:
```
# Final Summary of Findings
# - Female passengers had a much higher survival rate than male passengers.
# - Passengers in 1st class had the highest survival rates.
# - Younger passengers tended to survive slightly more compared to older passeng
# - Higher fares were paid by passengers who had a better chance of survival.
# - The dataset had missing values in 'Age', 'Cabin', and 'Embarked' columns.
# - 'Pclass' (passenger class) and 'Fare' showed some strong relationship with s
# - 'Sex' and 'Pclass' are important features to predict survival.
# - Family members traveling together (SibSp and Parch) have some positive relat
```

In [ ]:
```

```