Titanic Dataset - Exploratory Data Analysis

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

sns.set_style("whitegrid")
```

In [10]:
```
import pandas as pd

data = pd.read_csv("titanic.csv")
data.head()
```

Out[10]:

	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 [11]:
```
data.info()
```

```
<class 'pandas.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    str    
 4   Sex          891 non-null    str    
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64  
 7   Parch        891 non-null    int64  
 8   Ticket       891 non-null    str    
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    str    
 11  Embarked     889 non-null    str    
dtypes: float64(2), int64(5), str(5)
memory usage: 83.7 KB
```

In [12]:
```
data.describe()
```

Out[12]:

	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

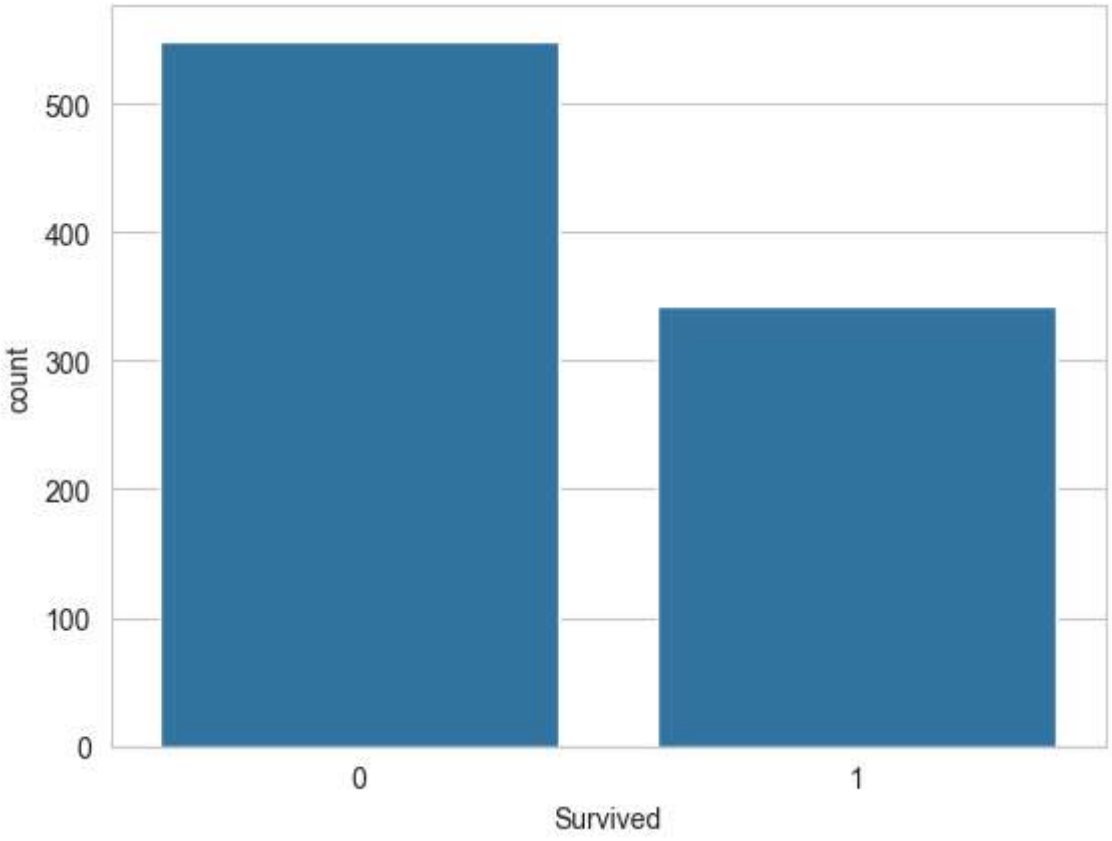
In [13]:
```
data.isnull().sum()
```

Out[13]:
```
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 [14]:
```
data['Survived'].value_counts()
data['Sex'].value_counts()
data['Pclass'].value_counts()
```

Out[14]:
```
Pclass
3    491
1    216
2    184
Name: count, dtype: int64
```

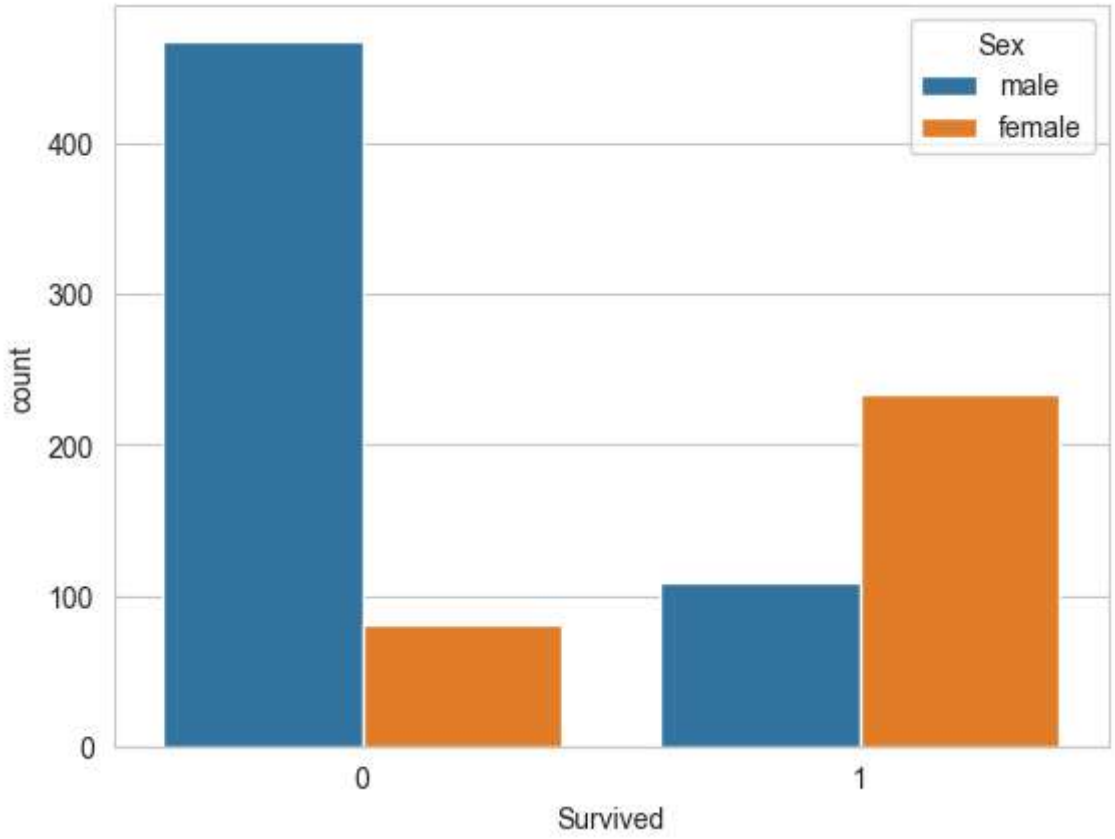
In [15]:
```
sns.countplot(x='Survived', data=data)
plt.show()
```



Observation:

Most passengers did not survive.

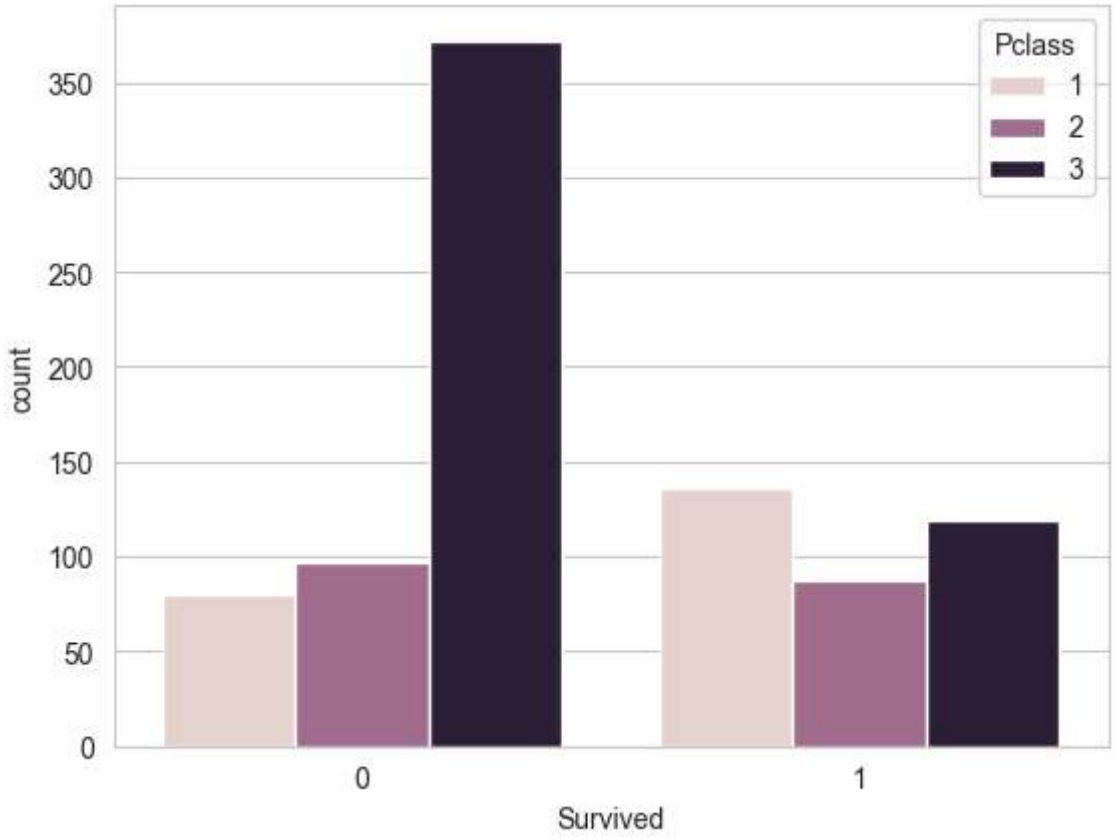
In [16]:
```
sns.countplot(x='Survived', hue='Sex', data=data)
plt.show()
```



Observation:

Females had much higher survival rate than males.

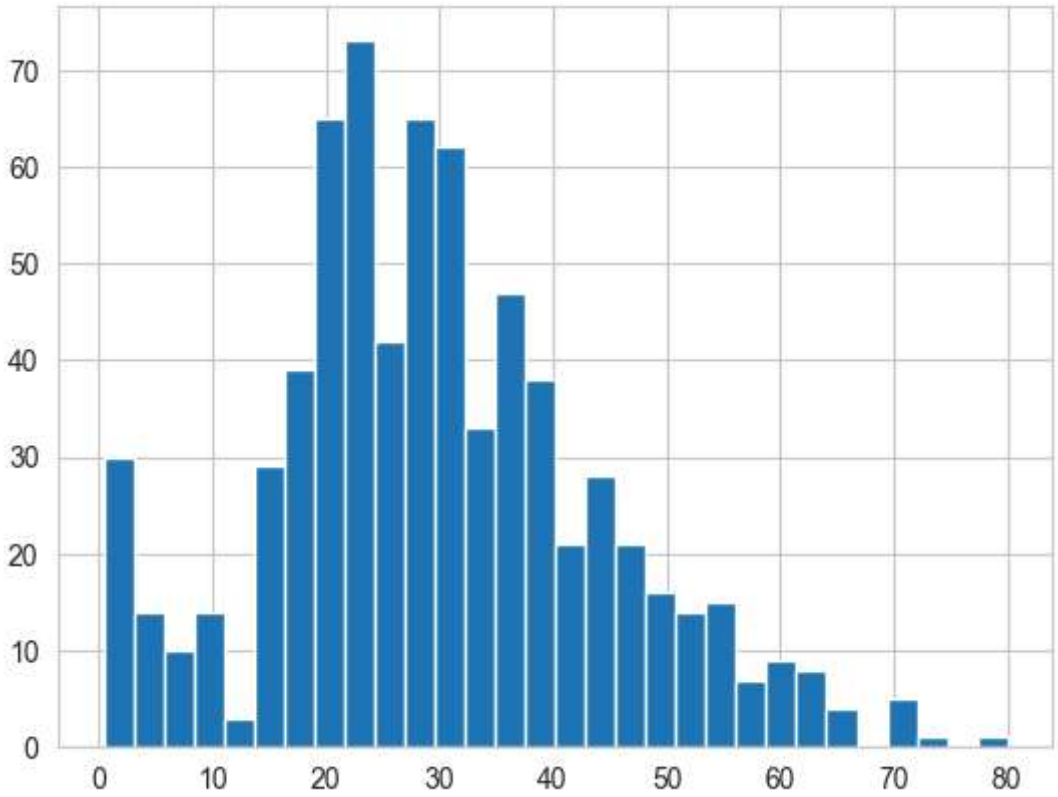
In [17]:
```
sns.countplot(x='Survived', hue='Pclass', data=data)
plt.show()
```



Observation:

1st class passengers survived more compared to 3rd class.

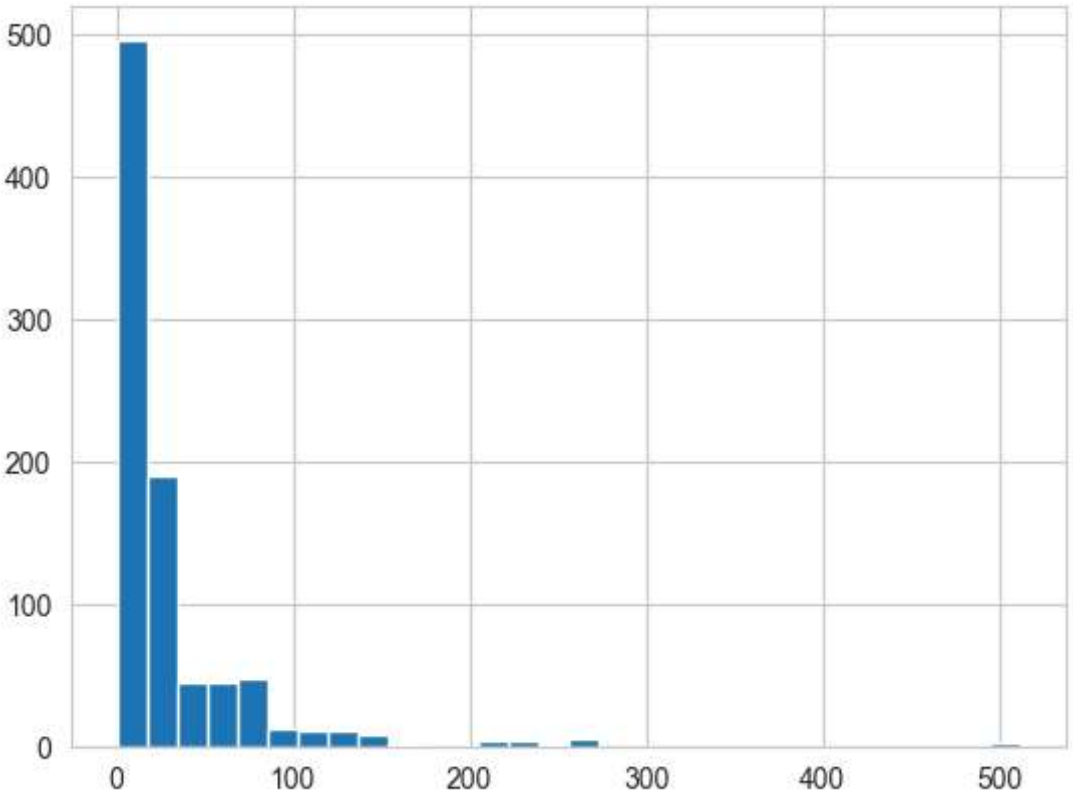
In [18]:
```
data['Age'].hist(bins=30)
plt.show()
```



Observation:

Most passengers were between 20–40 years old.

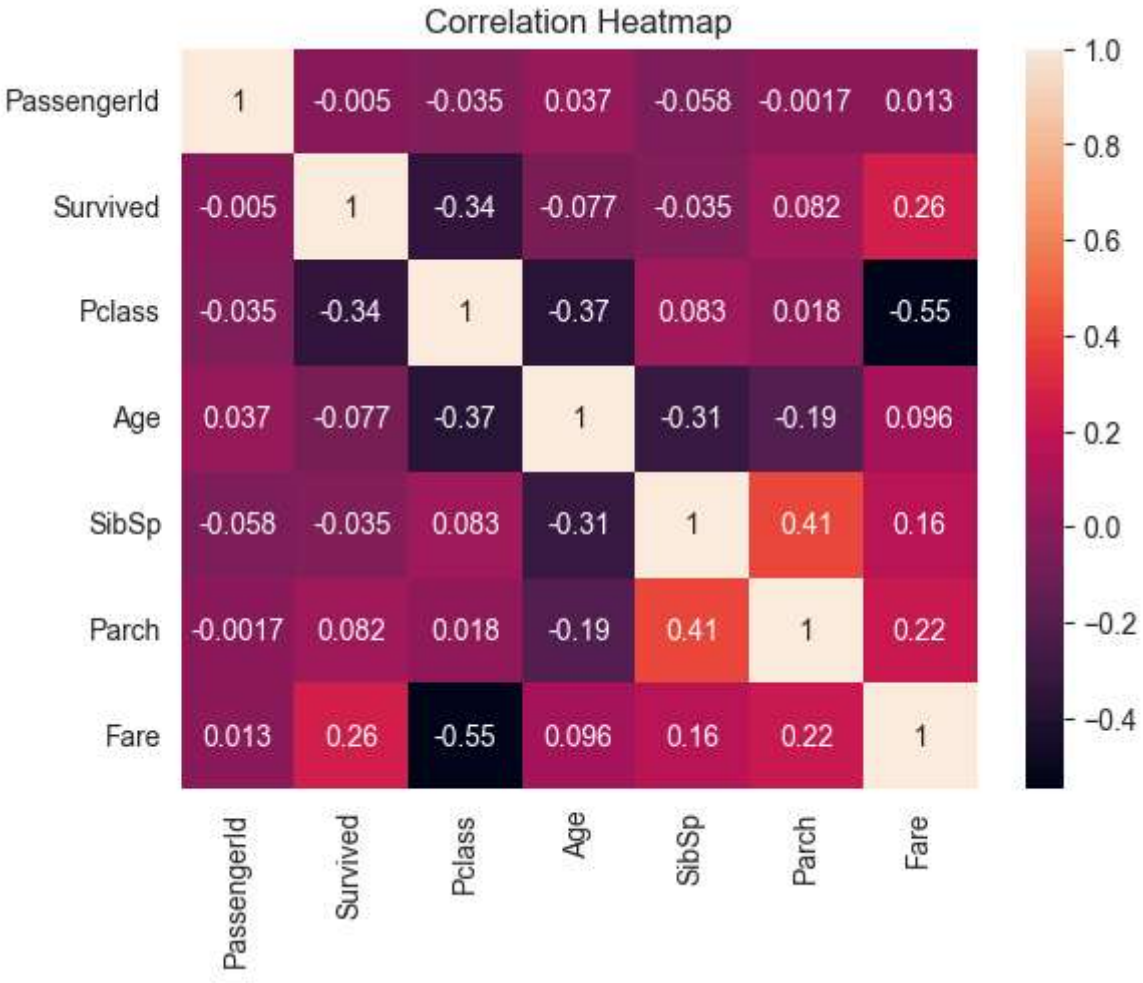
In [19]:
```
data['Fare'].hist(bins=30)
plt.show()
```



Observation:

Most passengers paid lower fare. Few paid very high fare.

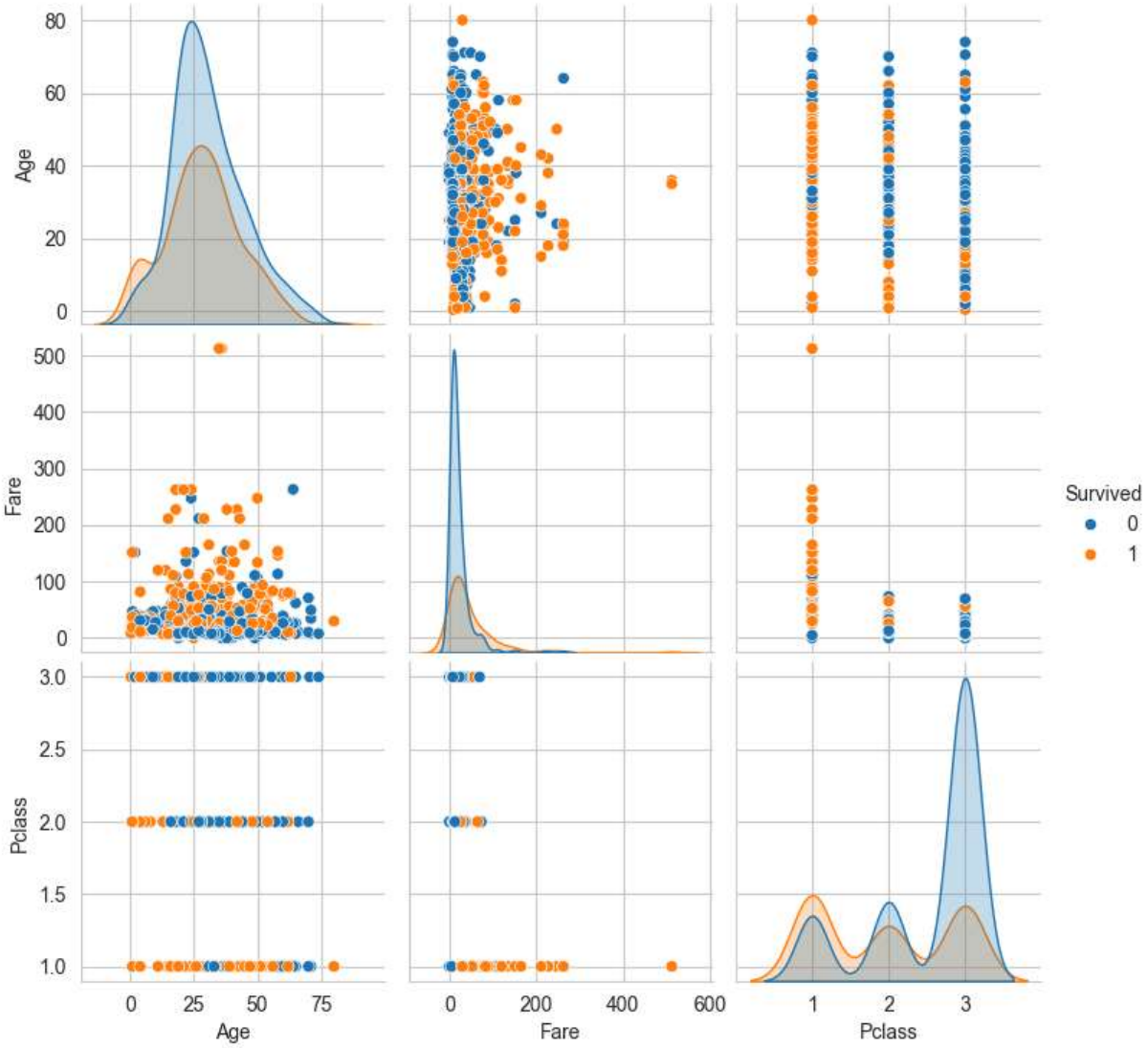
In [24]:
```
corr = data.corr(numeric_only=True)

sns.heatmap(corr, annot=True)

plt.title("Correlation Heatmap")   # Add this line
plt.show()
```



Observation:

Fare positively correlates with survival. Pclass negatively correlates with survival.

In [25]:
```
sns.pairplot(data[['Survived','Age','Fare','Pclass']], hue='Survived')
plt.show()
```



Observation:

Higher fare and 1st class passengers had better survival rate.

Final Summary of Findings

- Females survived more than males.
- 1st class passengers had higher survival rate.
- Passengers who paid higher fare survived more.
- Young passengers had slightly better survival rate.
- Pclass and Fare strongly affect survival.