In [27]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

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

titanic_df.head()
```

Out[27]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	I
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0

In [11]:
```
titanic_df.isnull().sum()
```

Out[11]:
```
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 [12]:
```
# Data Cleaning

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

```
titanic_df['Embarked'].fillna(titanic_df['Embarked'].mode()[0], inplace=True)
titanic_df.drop(columns=['Cabin'], inplace=True)
titanic_df.drop(columns=['Ticket'], inplace=True)
```

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

Out[13]:
```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Fare           0
Embarked       0
dtype: int64
```

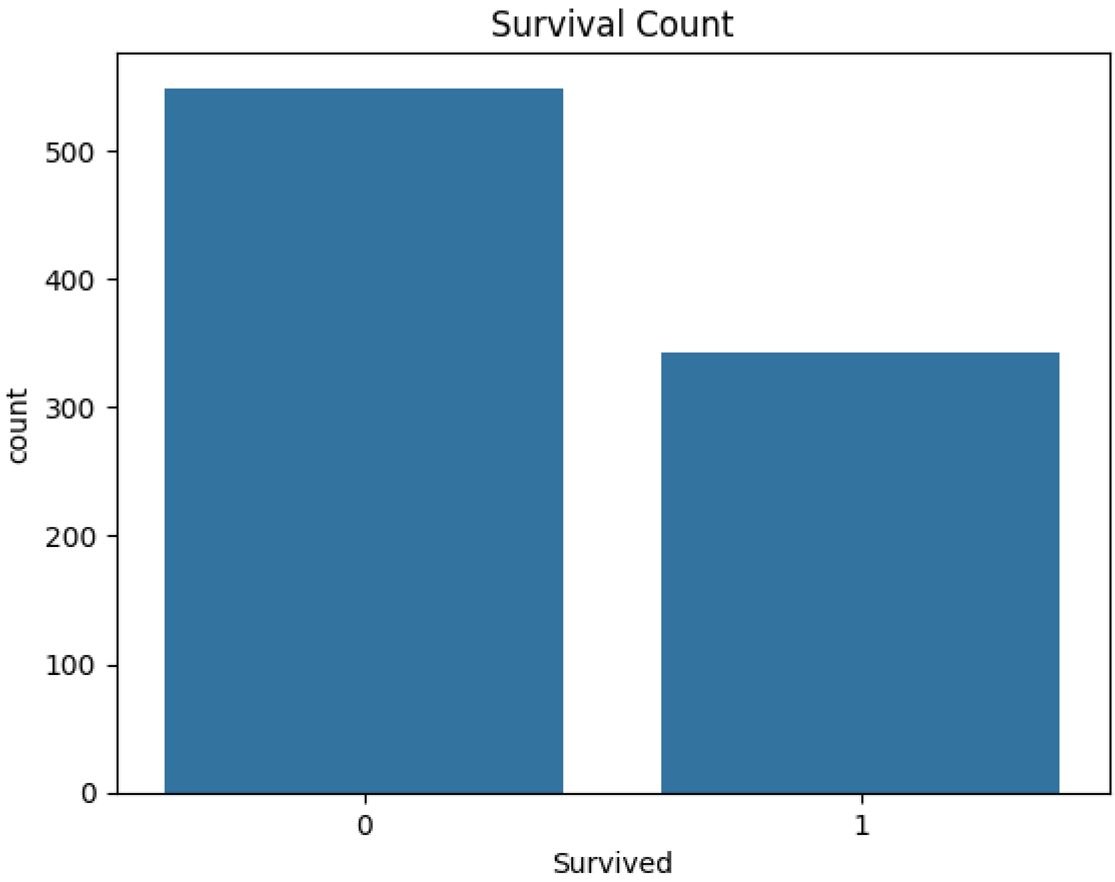
In [14]:
```
#EDA

titanic_df.describe()
```

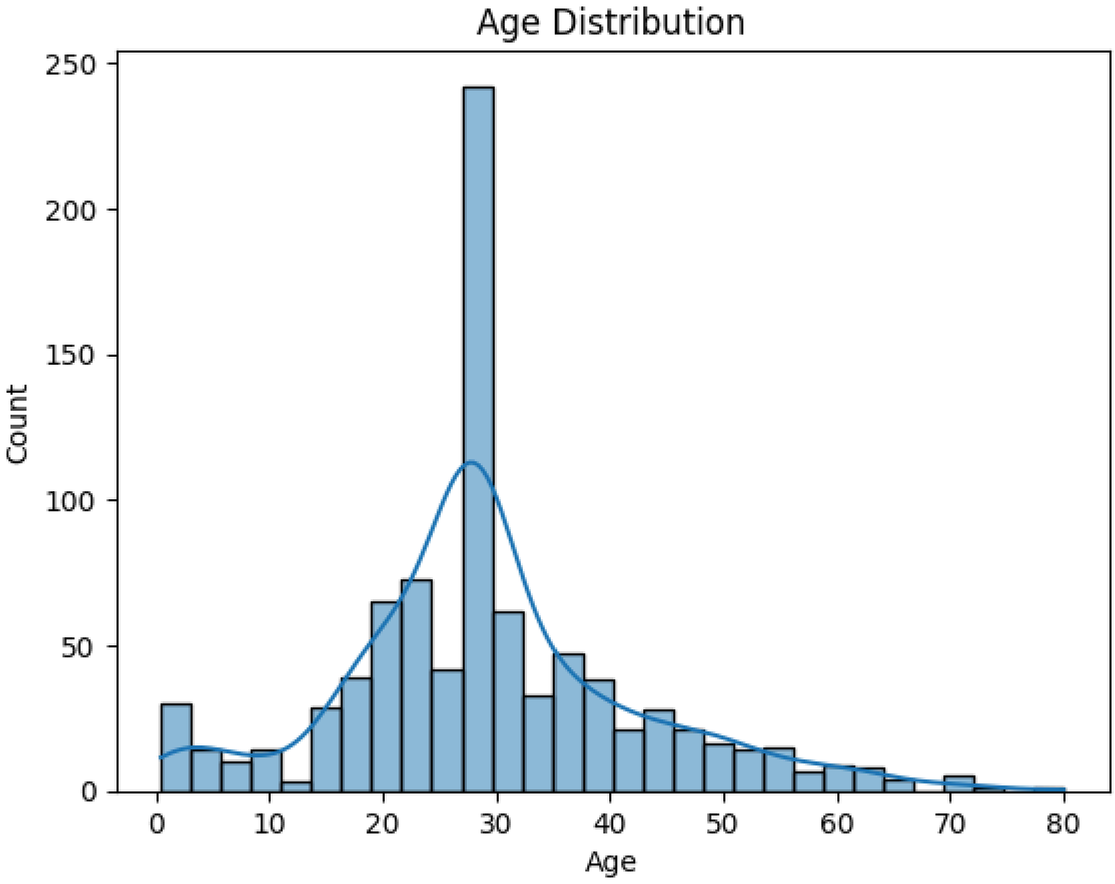
Out[14]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	
count	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000	891.000
mean	446.000000	0.383838	2.308642	29.361582	0.523008	0.381594	32.204
std	257.353842	0.486592	0.836071	13.019697	1.102743	0.806057	49.693
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000
25%	223.500000	0.000000	2.000000	22.000000	0.000000	0.000000	7.910
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454
75%	668.500000	1.000000	3.000000	35.000000	1.000000	0.000000	31.000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329

In [15]:
```
# Survival rate
sns.countplot(x='Survived', data=titanic_df)
plt.title('Survival Count')
plt.show()
```

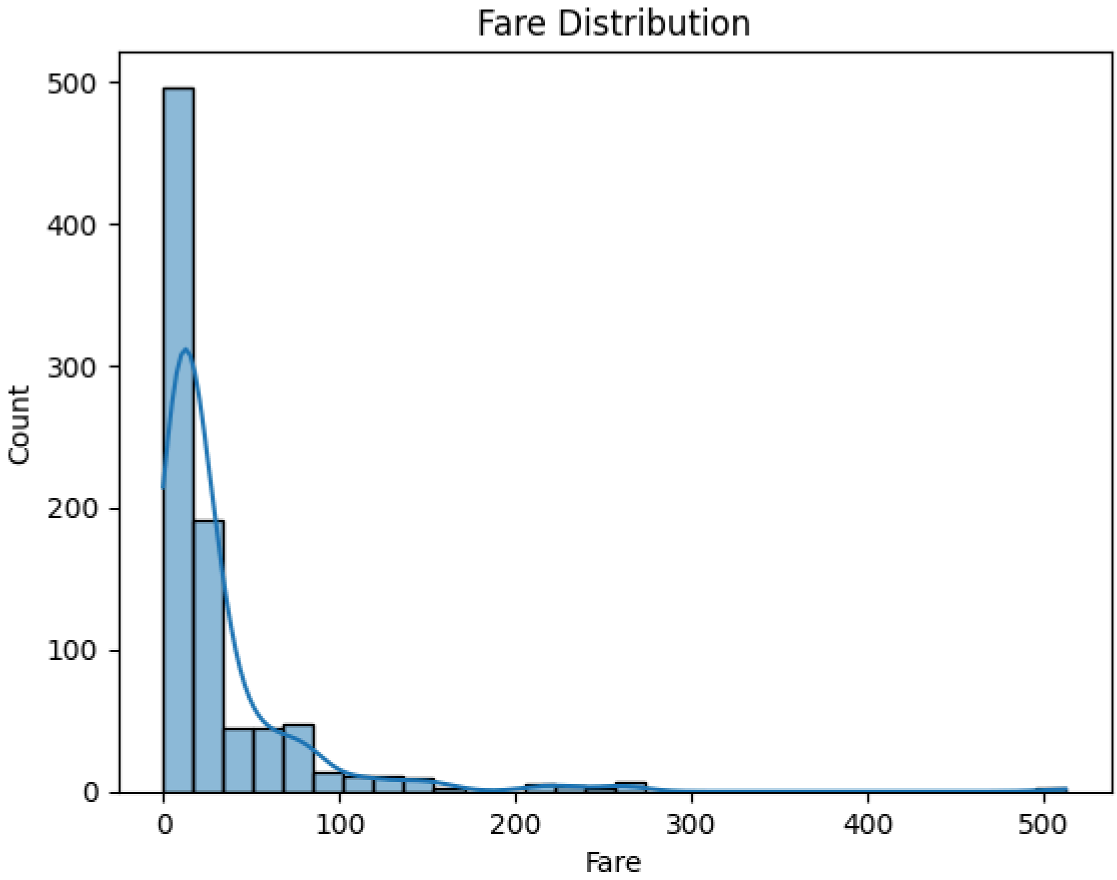


```
In [16]: # Distribution of Age
         sns.histplot(titanic_df['Age'], bins=30, kde=True)
         plt.title('Age Distribution')
         plt.show()
```



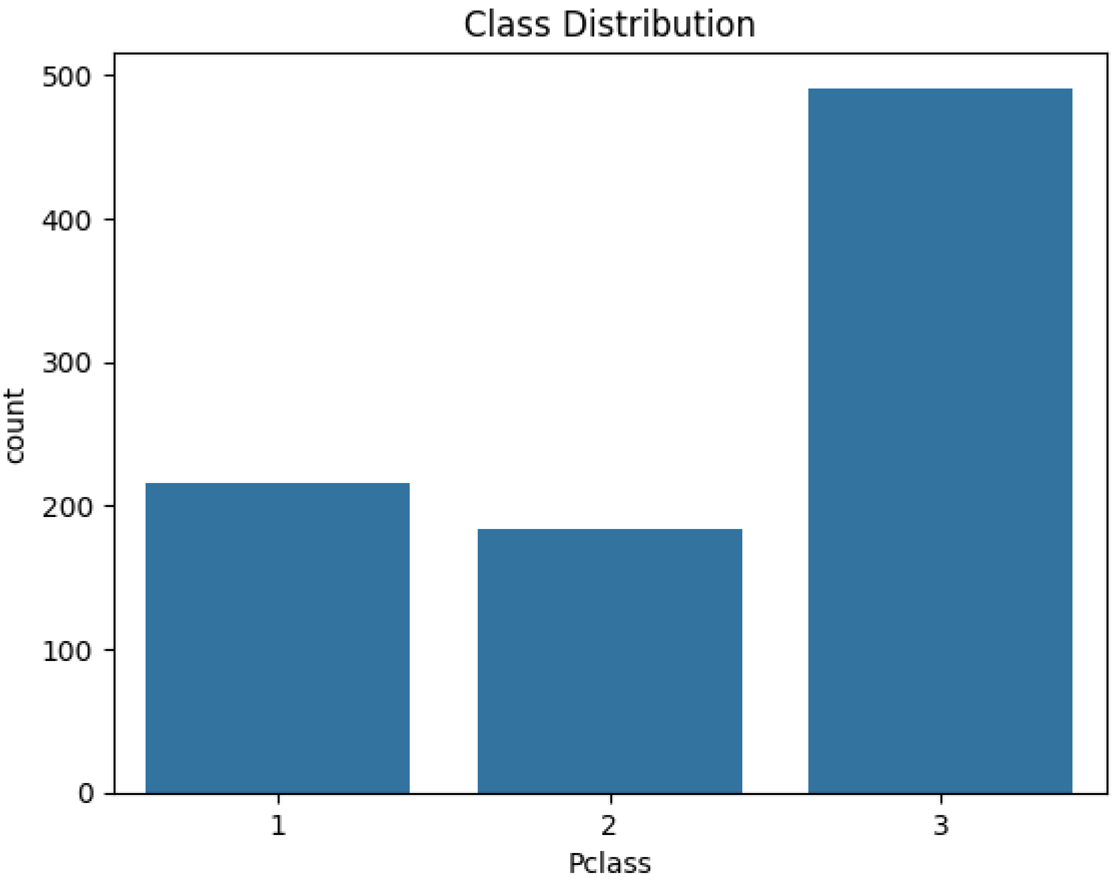
In [17]:

```
# Distribution of Fare
sns.histplot(titanic_df['Fare'], bins=30, kde=True)
plt.title('Fare Distribution')
plt.show()
```



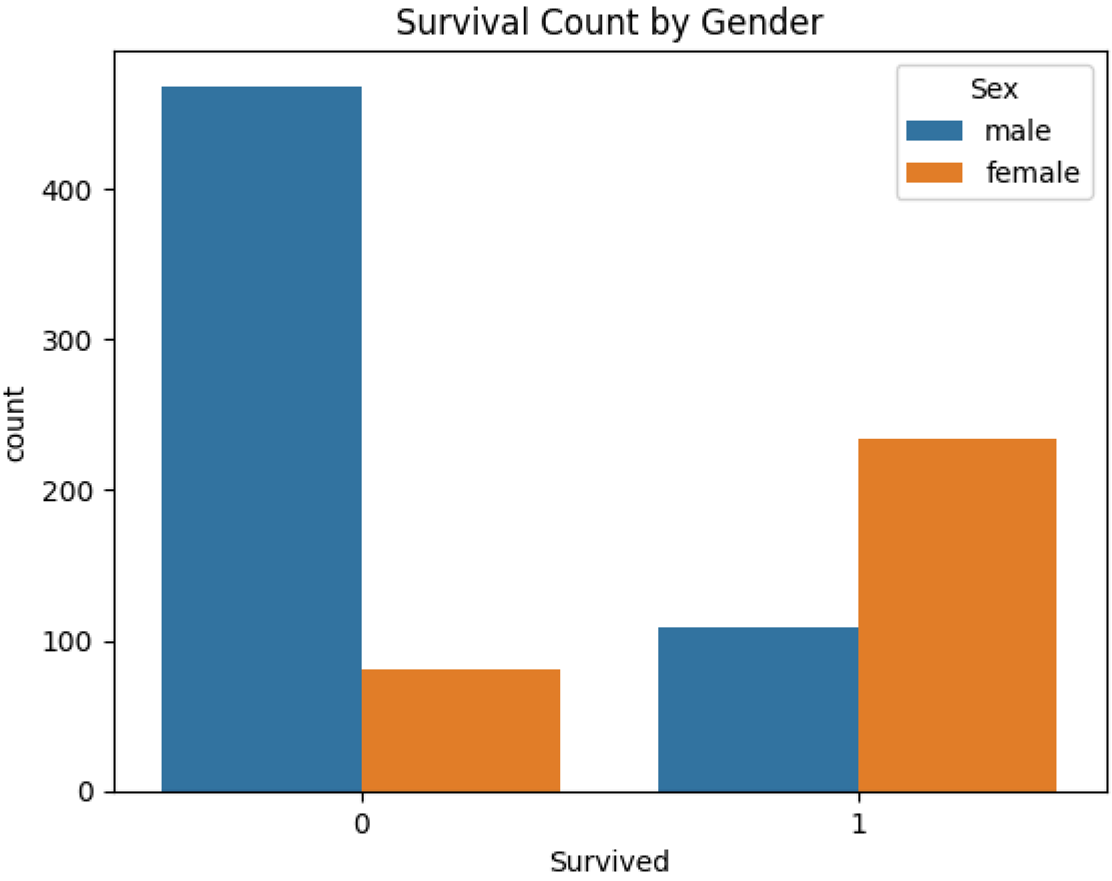
In [18]:

```
# Class Distribution
sns.countplot(x='Pclass', data=titanic_df)
plt.title('Class Distribution')
plt.show()
```



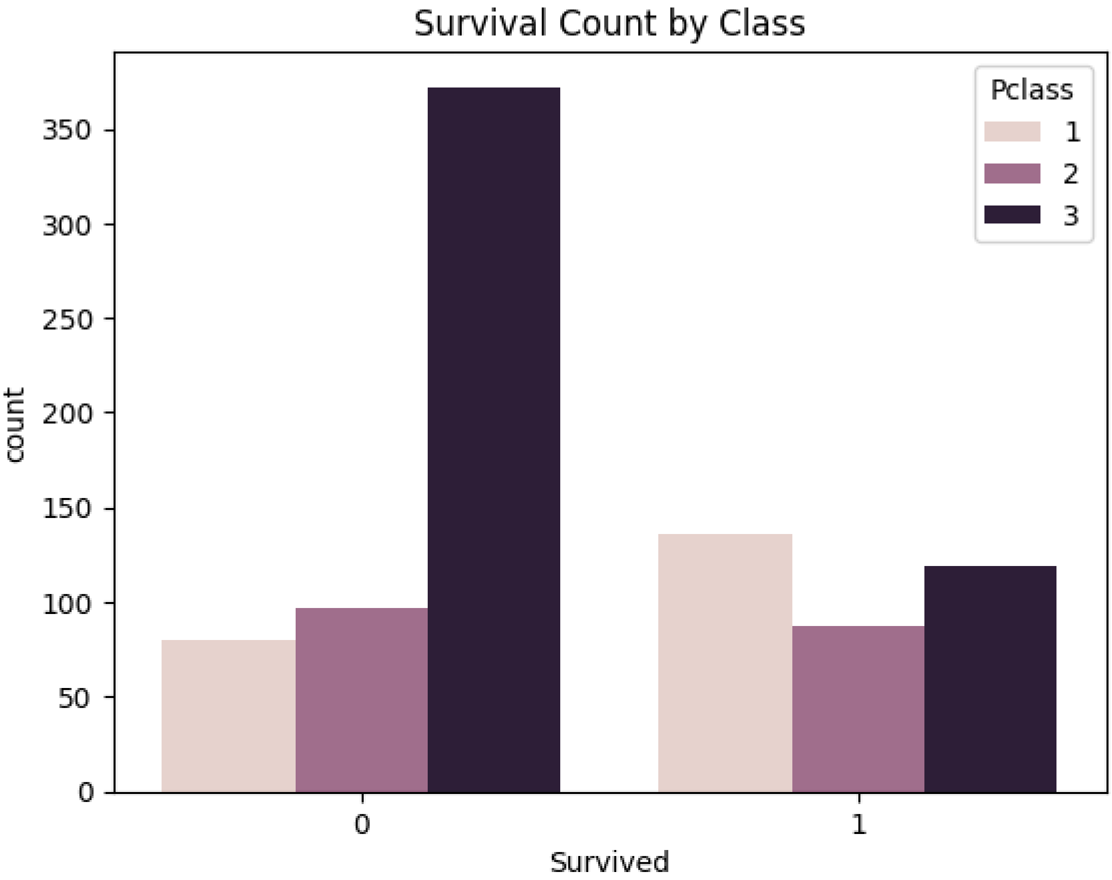
In [19]:

```
# Survival rate by gender
sns.countplot(x='Survived', hue='Sex', data=titanic_df)
plt.title('Survival Count by Gender')
plt.show()
```



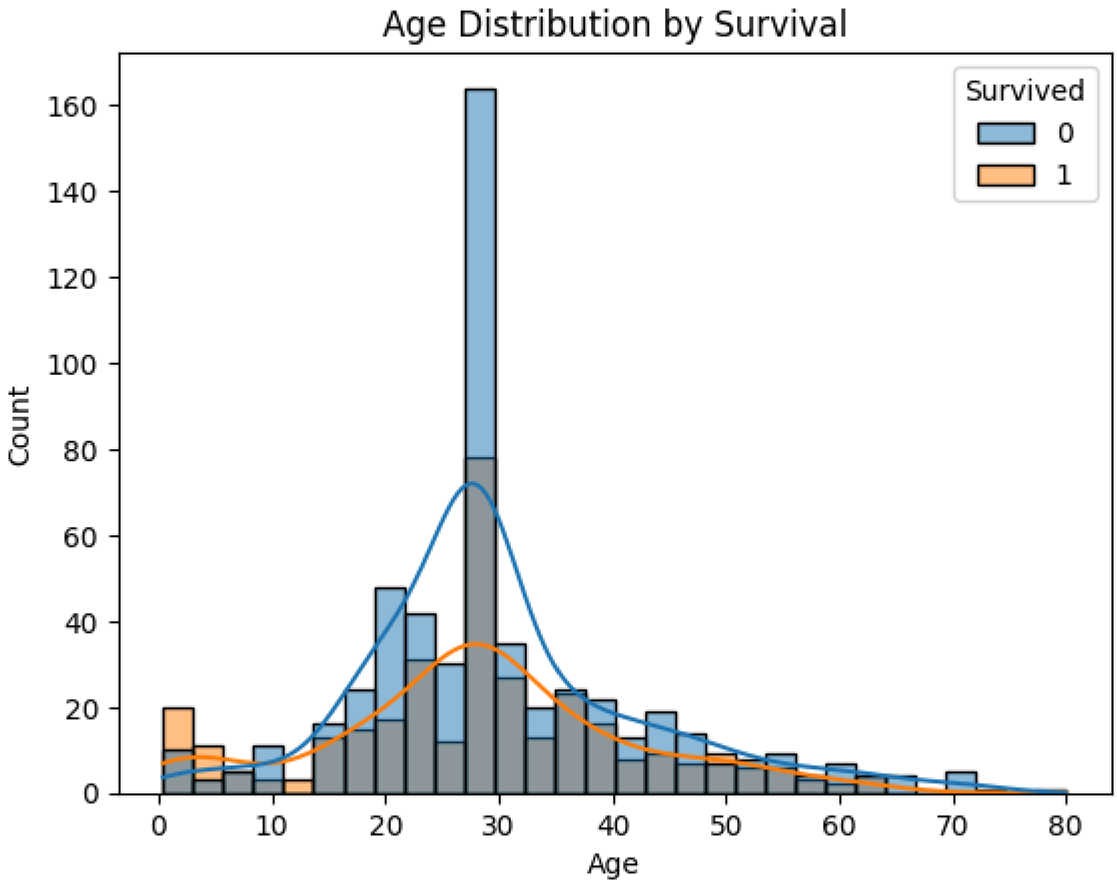
In [20]:

```
# Survival rate by class
sns.countplot(x='Survived', hue='Pclass', data=titanic_df)
plt.title('Survival Count by Class')
plt.show()
```

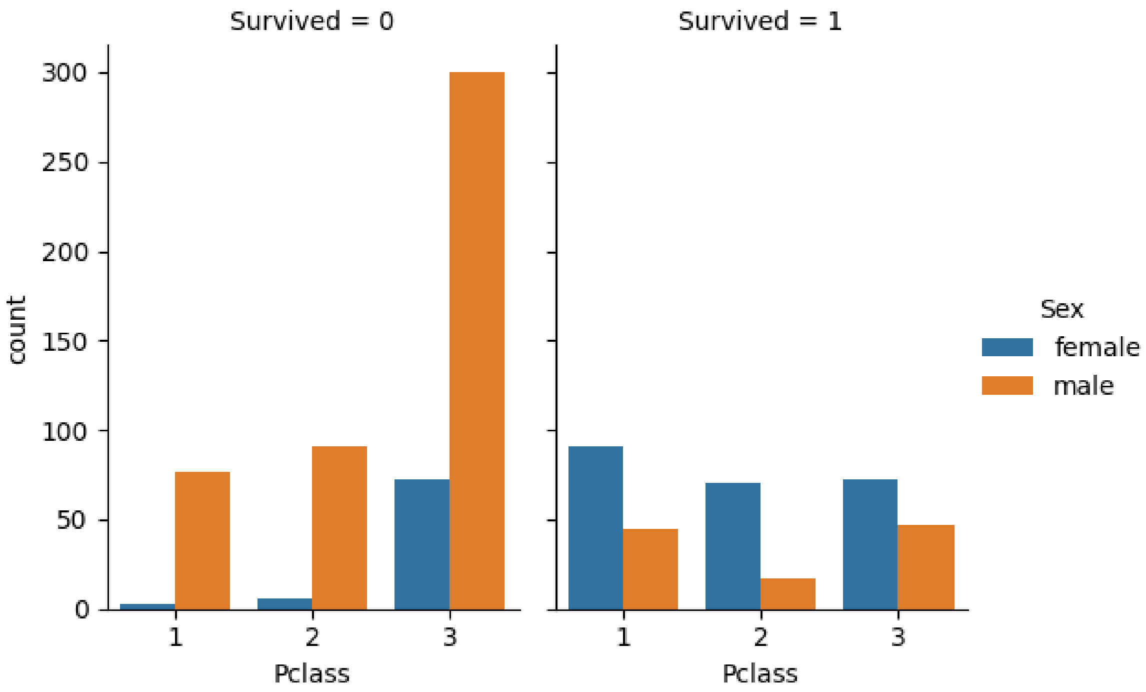


In [21]:

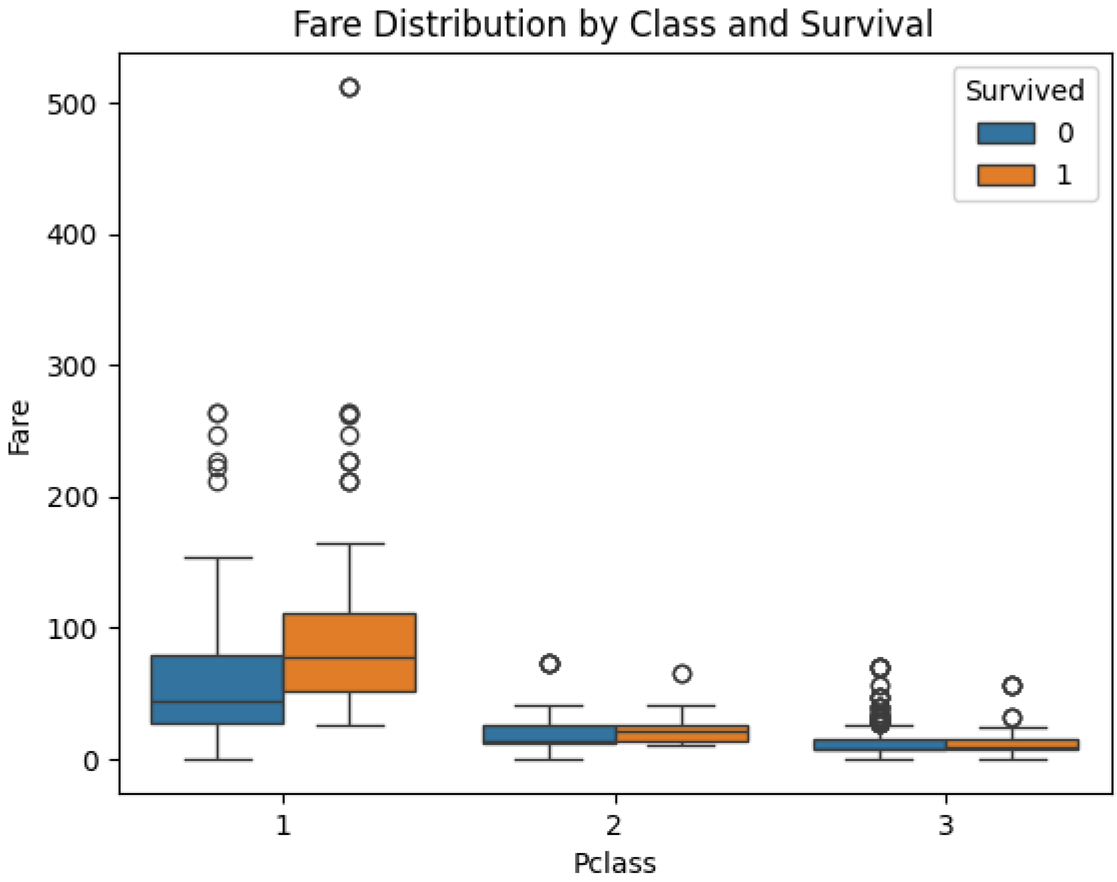
```
# Age distribution by survival
sns.histplot(data=titanic_df, x='Age', hue='Survived', bins=30, kde=True)
plt.title('Age Distribution by Survival')
plt.show()
```



In [22]:
```python
# Survival rate by class and gender
sns.catplot(x='Pclass', hue='Sex', col='Survived', data=titanic_df, kind='count'
plt.show()
```



In [25]:
```python
# Fare distribution by class and survival
sns.boxplot(x='Pclass', y='Fare', hue='Survived', data=titanic_df)
plt.title('Fare Distribution by Class and Survival')
plt.show()
```


Fare Distribution by Class and Survival
Survived
0
1
500
400
300
200
100
0
Fare
1
2
3
Pclass