Titanic_EDA

June 11, 2025

[11]:
```
import pandas as pd
df = pd.read_csv("titanic.csv")
```

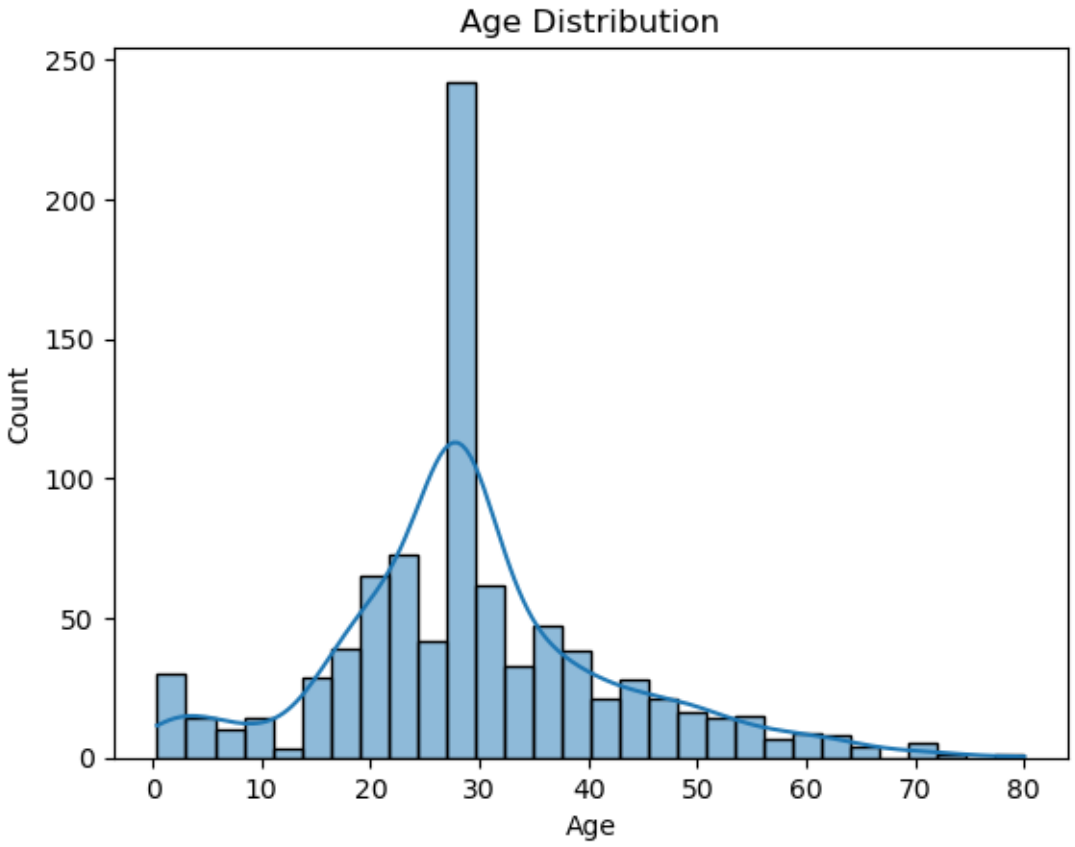
[15]:
```
df.shape
df.info()

df.describe(include='all')

df.isnull().sum()

df['Sex'].value_counts()
df['Embarked'].value_counts()
```

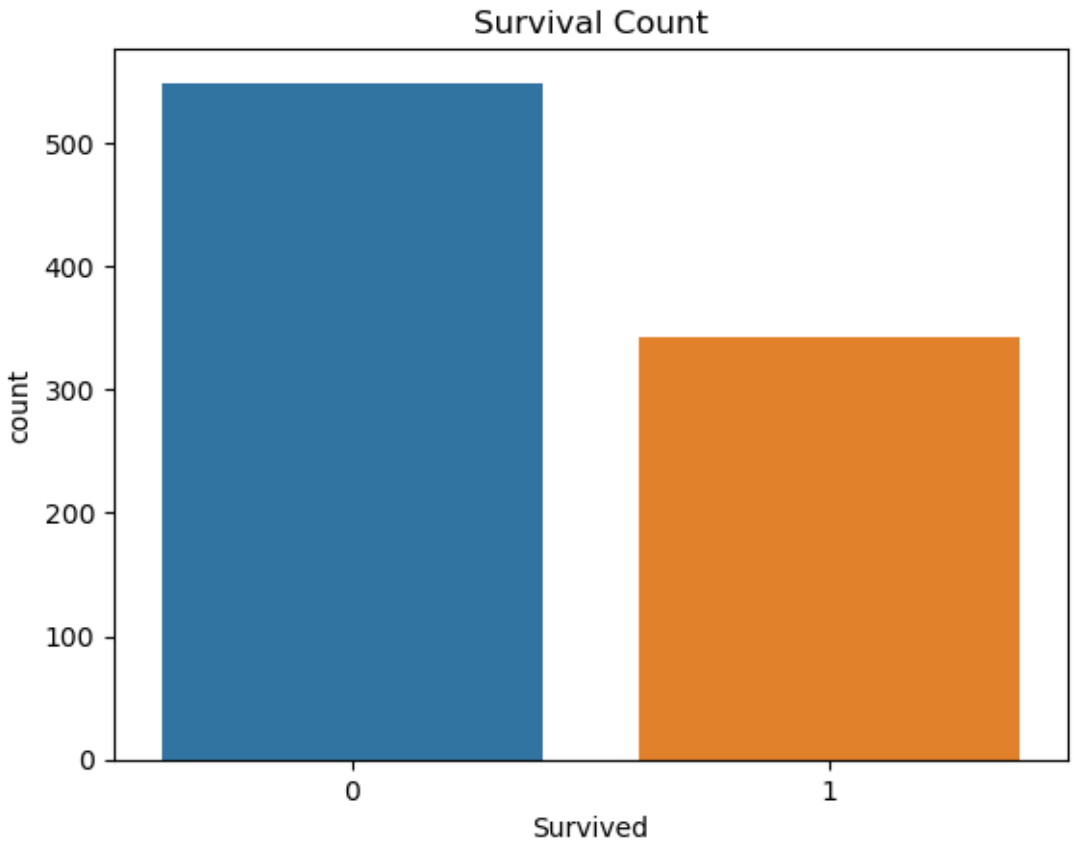
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null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

[15]:
```
Embarked
S    644
C    168
Q     77
Name: count, dtype: int64
```

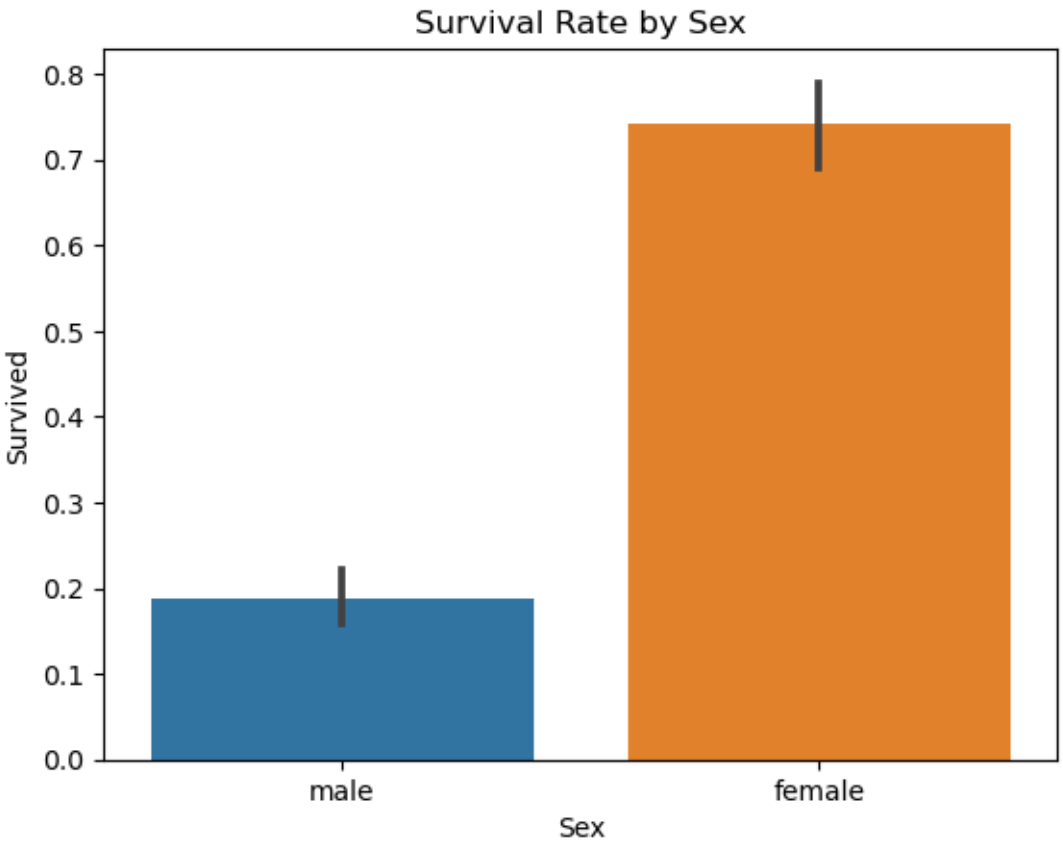
[17]:
```
df['Age'].fillna(df['Age'].median(), inplace=True)
df['Embarked'].fillna(df['Embarked'].mode()[0], inplace=True)
df.drop('Cabin', axis=1, inplace=True)

df.drop_duplicates(inplace=True)
```

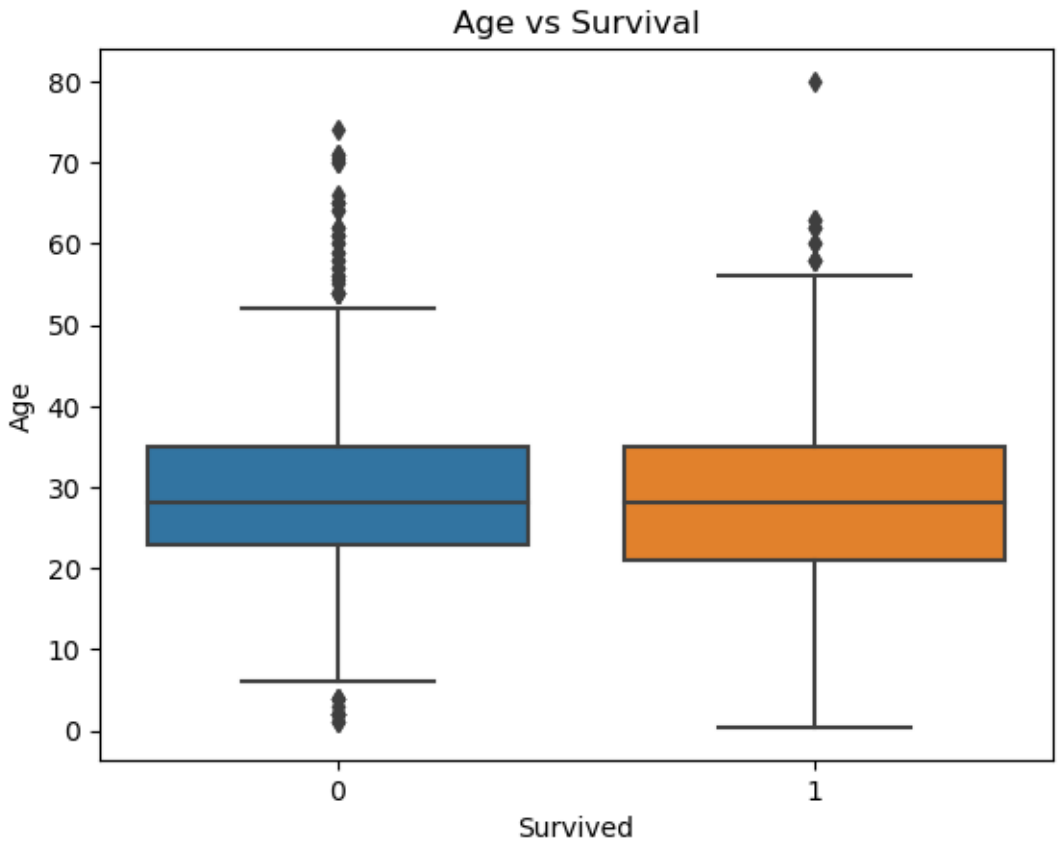
[19]:
```
import matplotlib.pyplot as plt
import seaborn as sns

sns.histplot(df['Age'], kde=True)
plt.title('Age Distribution')
plt.show()

sns.countplot(x='Survived', data=df)
plt.title('Survival Count')
plt.show()
```



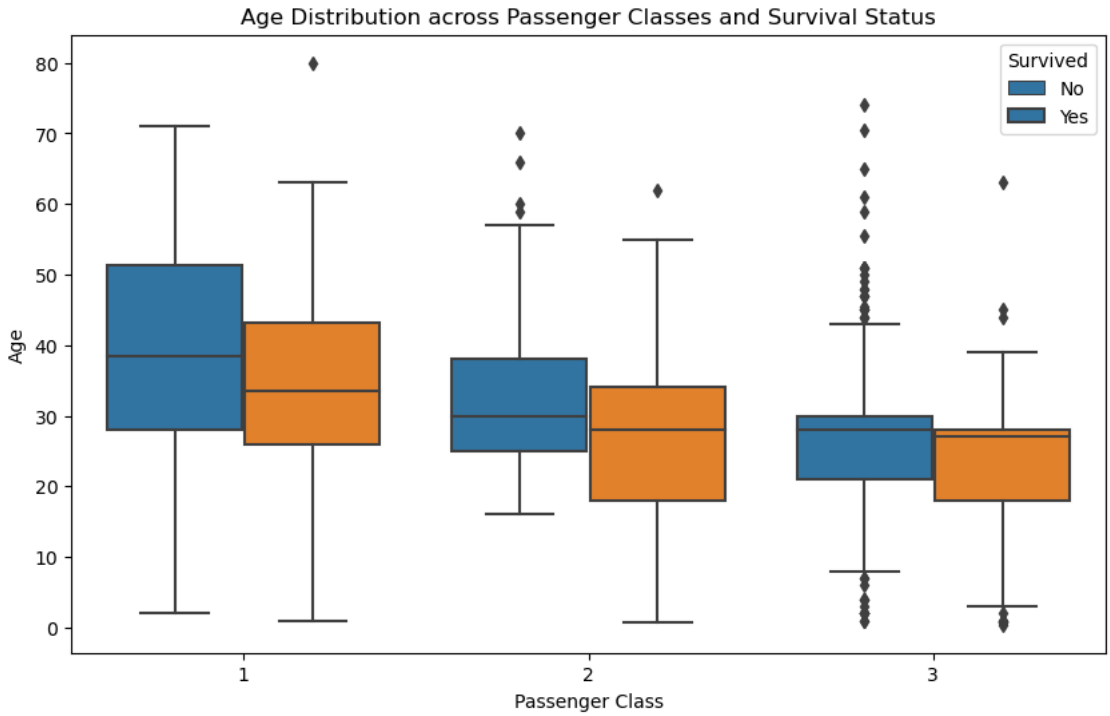


[21]:
```
sns.barplot(x='Sex', y='Survived', data=df)
plt.title('Survival Rate by Sex')
plt.show()

sns.boxplot(x='Survived', y='Age', data=df)
plt.title('Age vs Survival')
plt.show()
```

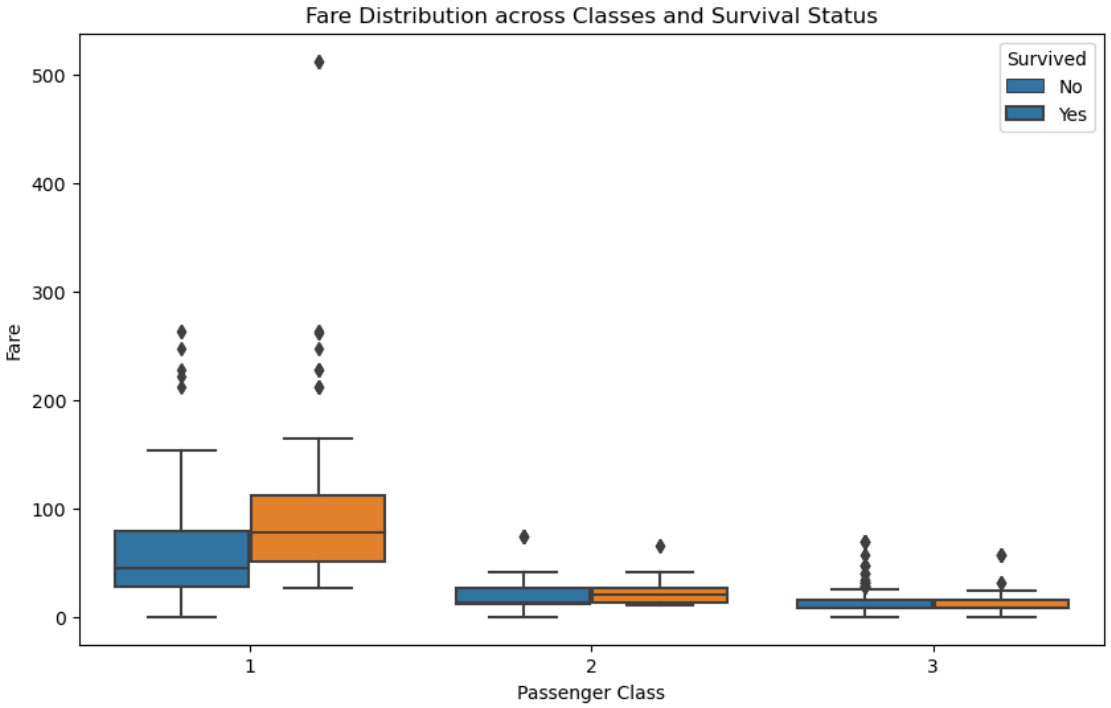

Survival Rate by Sex
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Survived
male
female
Sex



[31]:
```
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(10, 6))
sns.boxplot(x='Pclass', y='Age', hue='Survived', data=df)
plt.title('Age Distribution across Passenger Classes and Survival Status')
plt.xlabel('Passenger Class')
plt.ylabel('Age')
plt.legend(title='Survived', labels=['No', 'Yes'])
plt.show()
```



[33]:
```
plt.figure(figsize=(10, 6))
sns.boxplot(x='Pclass', y='Fare', hue='Survived', data=df)
plt.title('Fare Distribution across Classes and Survival Status')
plt.xlabel('Passenger Class')
plt.ylabel('Fare')
plt.legend(title='Survived', labels=['No', 'Yes'])
plt.show()
```



[ ]:
```
## Summary of Insights:
- Females had a significantly higher survival rate than males.
- Passengers in 1st class had higher chances of survival.
- Younger passengers had slightly higher survival chances.
- Fare and class were positively correlated.
- Age was not strongly correlated with survival but showed some patterns.
```