

```
In [33]: import pandas as pd      # For data handling
import numpy as np      # For numerical operations
import matplotlib.pyplot as plt # For plotting graphs
import seaborn as sns    # For advanced visualizations

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

```
In [35]: df = pd.read_csv("C:/Users/shiba/Downloads/titanic/train.csv")
```

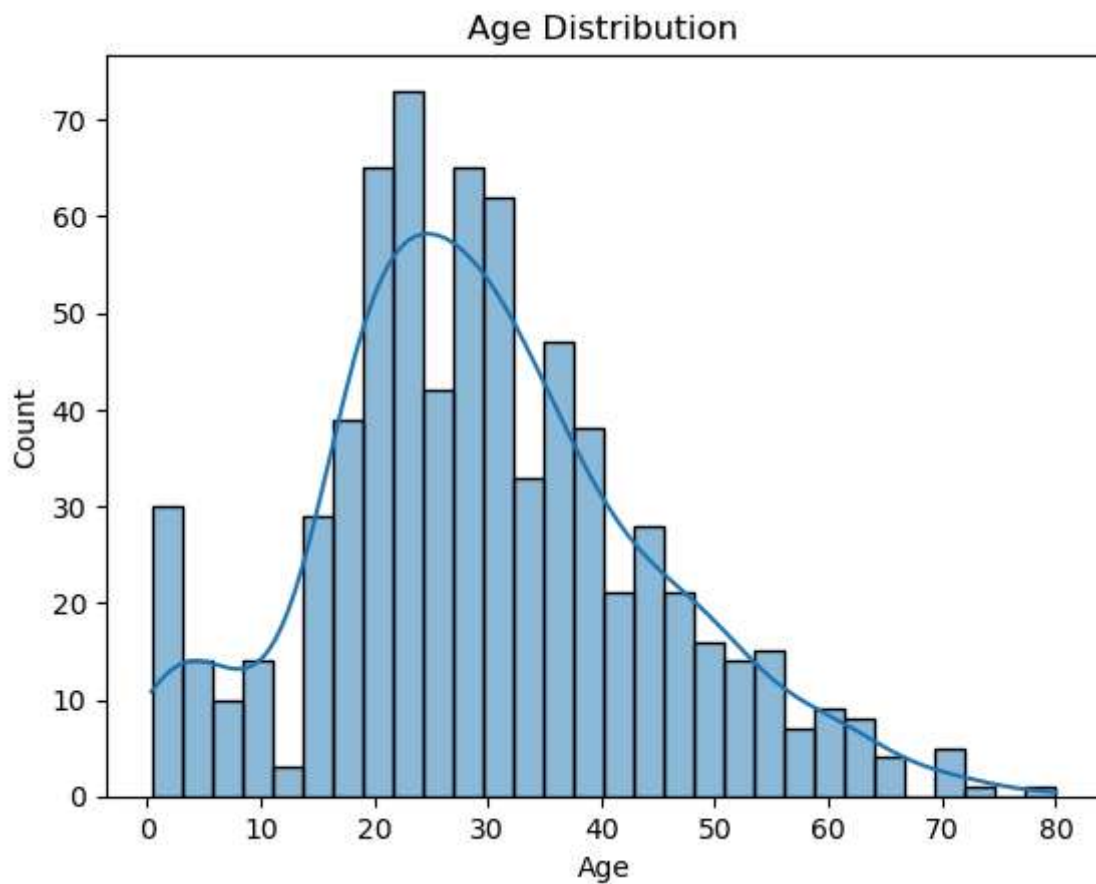
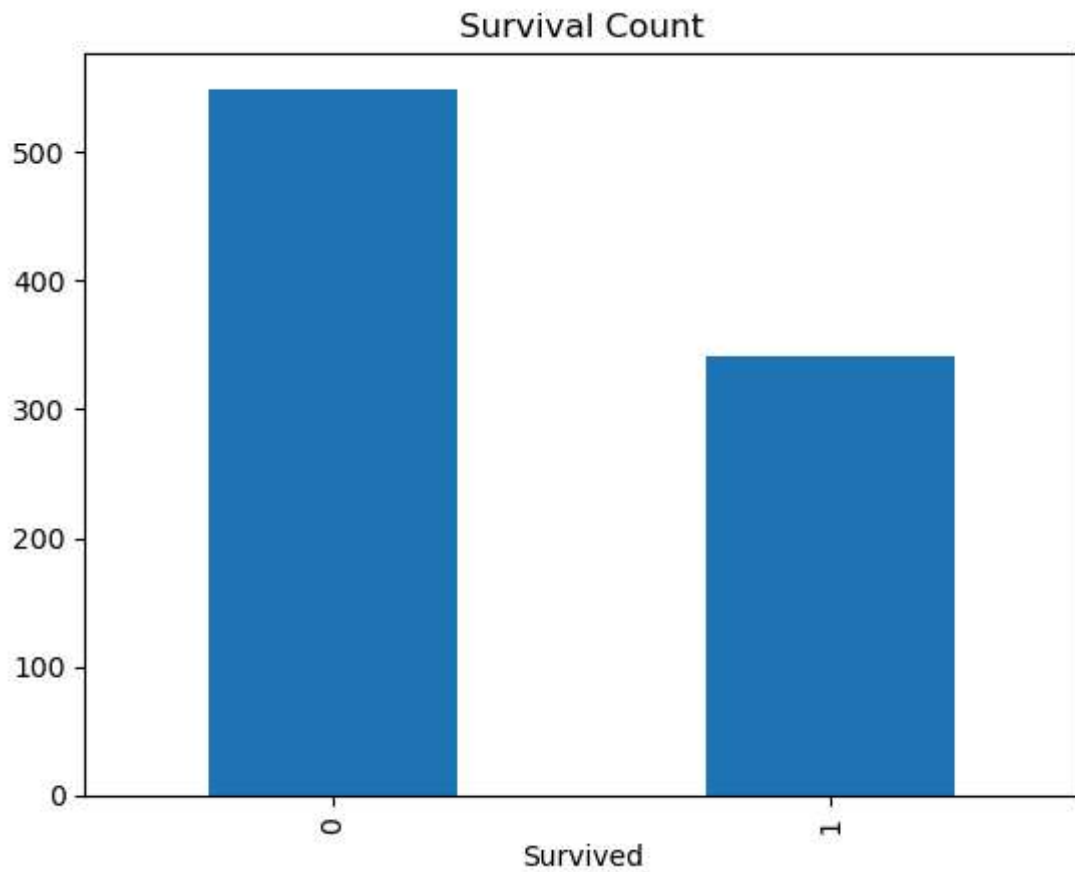
```
In [37]: df.head()           # First 5 rows
df.info()           # Data types and missing values
df.describe()       # Statistical summary
df.isnull().sum()   # Total missing values per column
df.columns          # List of all columns
```

```
<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
```

```
Out[37]: Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
               'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
              dtype='object')
```

```
In [39]: # Categorical Features
df['Survived'].value_counts().plot(kind='bar', title='Survival Count')
plt.show()

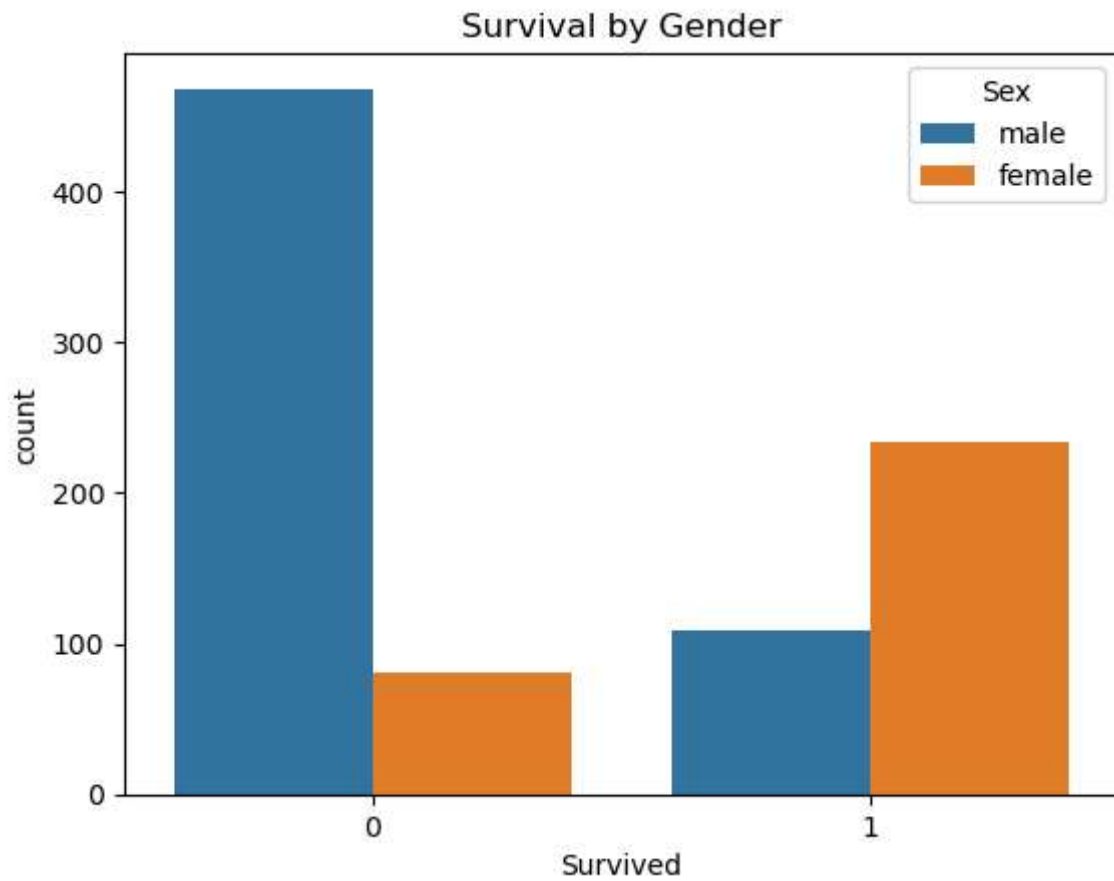
# Numerical Feature - Age
sns.histplot(df['Age'].dropna(), kde=True, bins=30)
plt.title('Age Distribution')
plt.show()
```

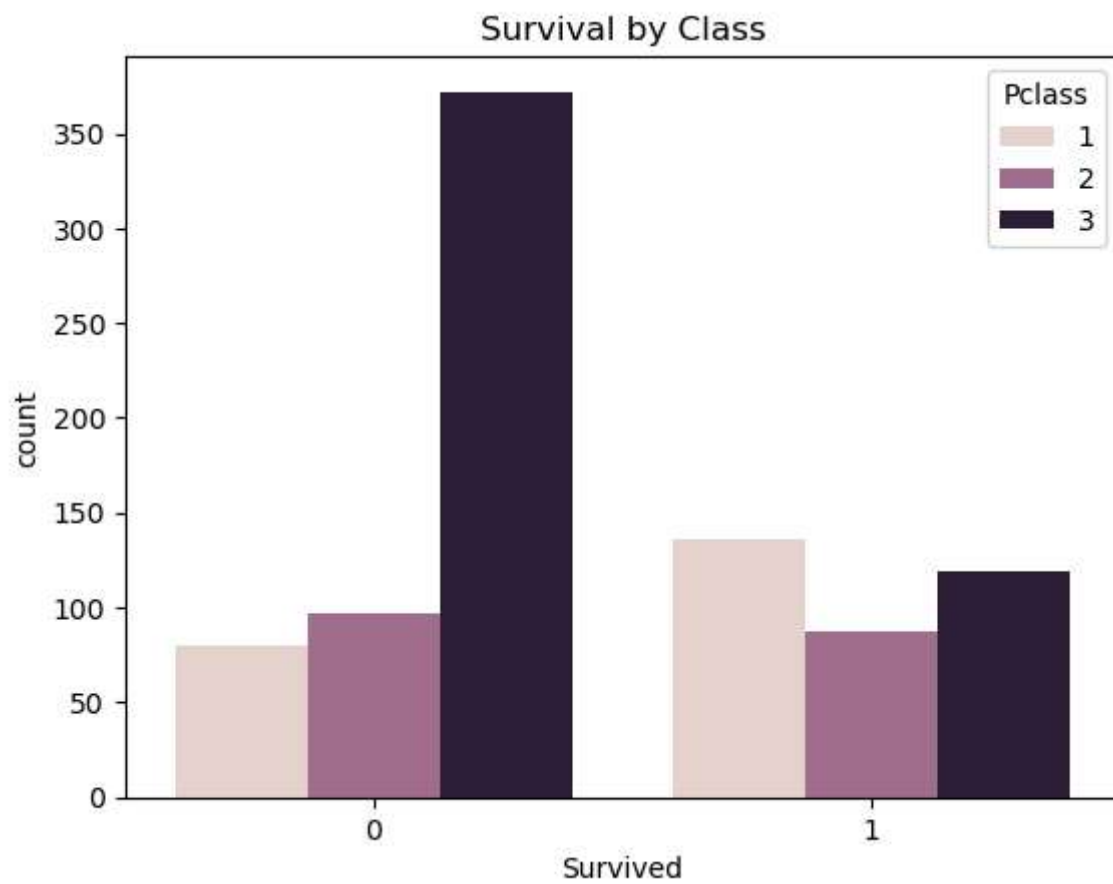


```
In [41]: # Survival by Gender  
sns.countplot(x='Survived', hue='Sex', data=df)
```

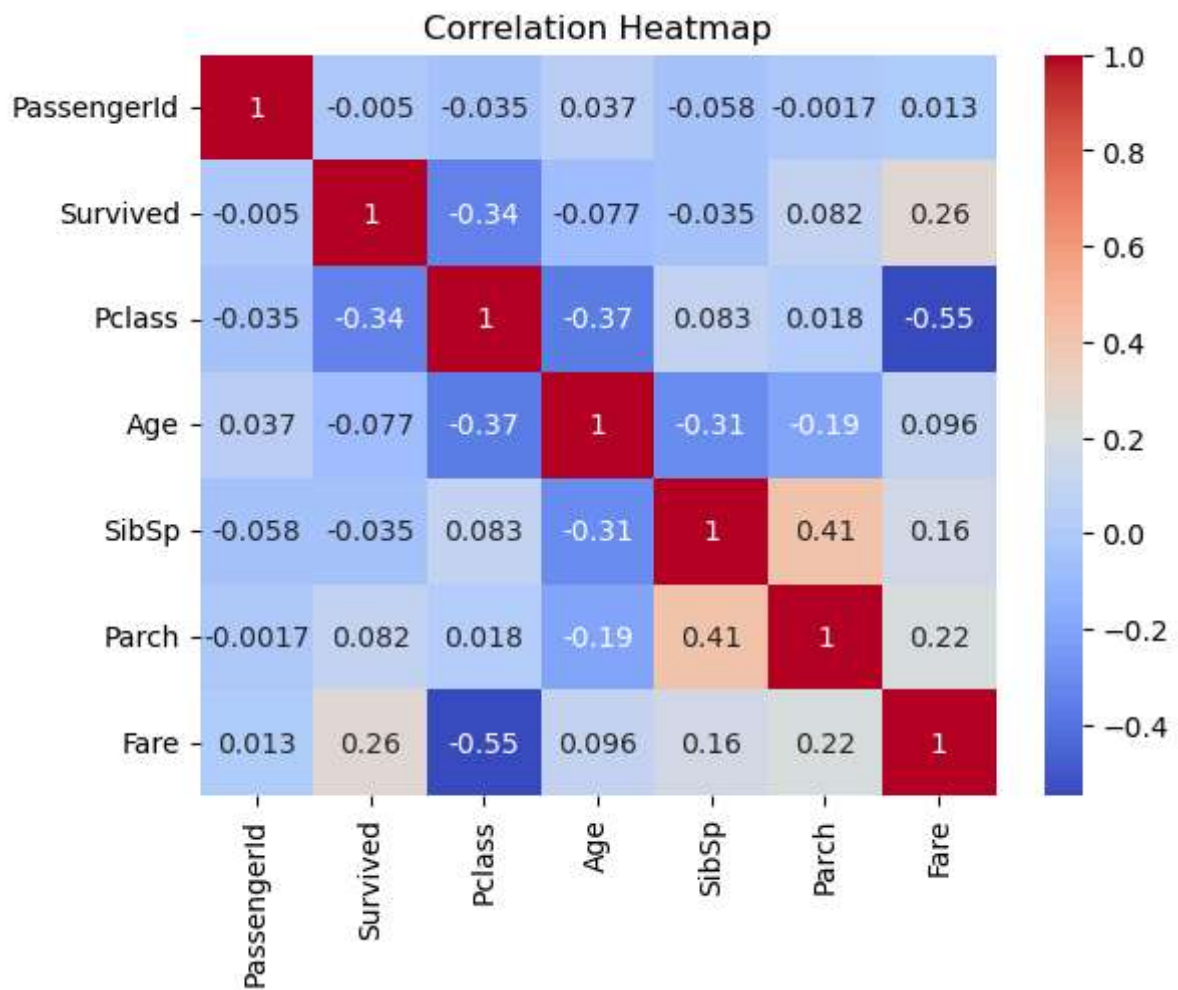
```
plt.title('Survival by Gender')
plt.show()

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

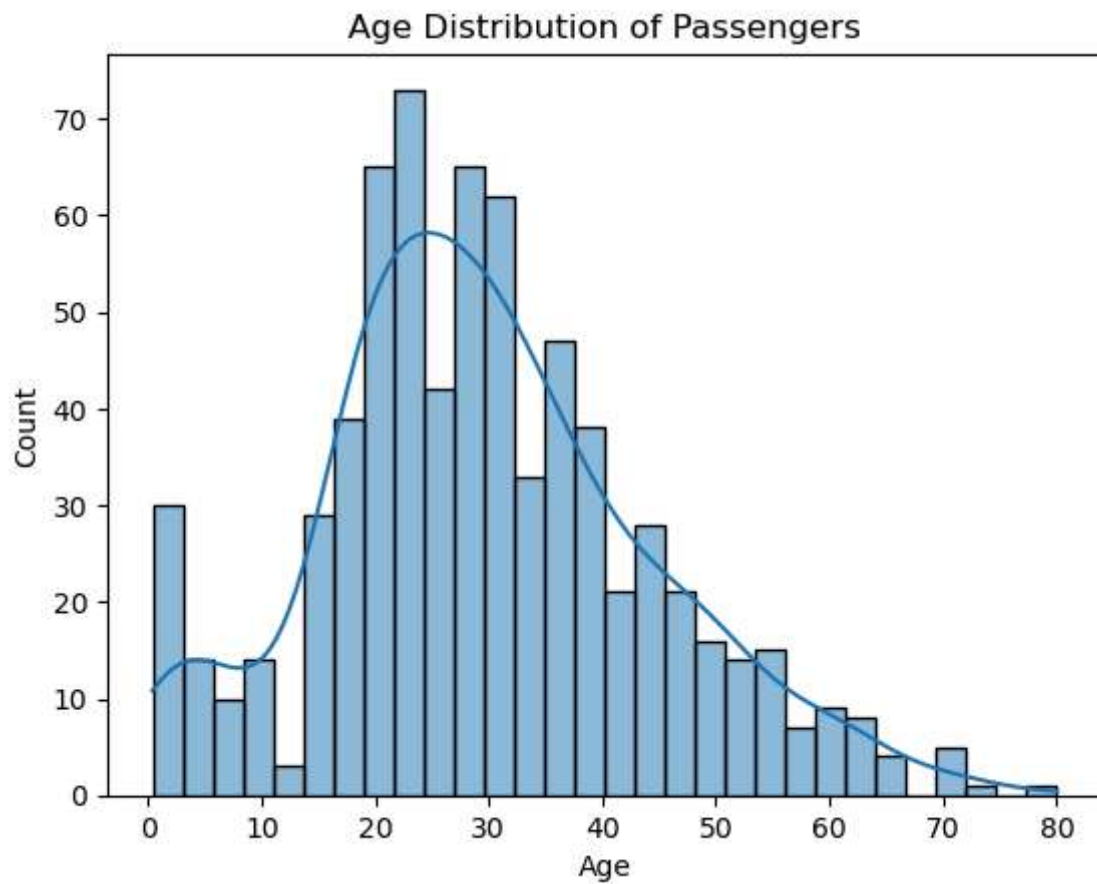




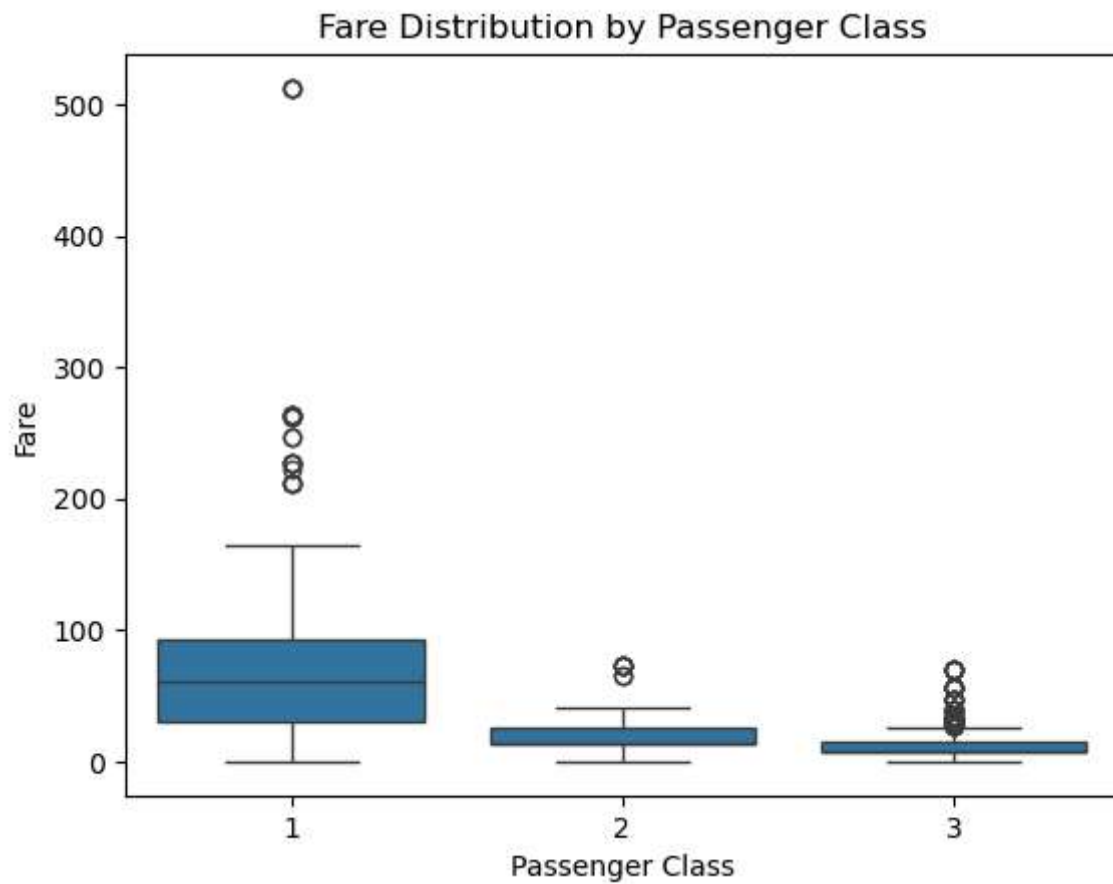
```
In [45]: sns.heatmap(df.corr(numeric_only=True), annot=True, cmap='coolwarm')  
plt.title('Correlation Heatmap')  
plt.show()
```



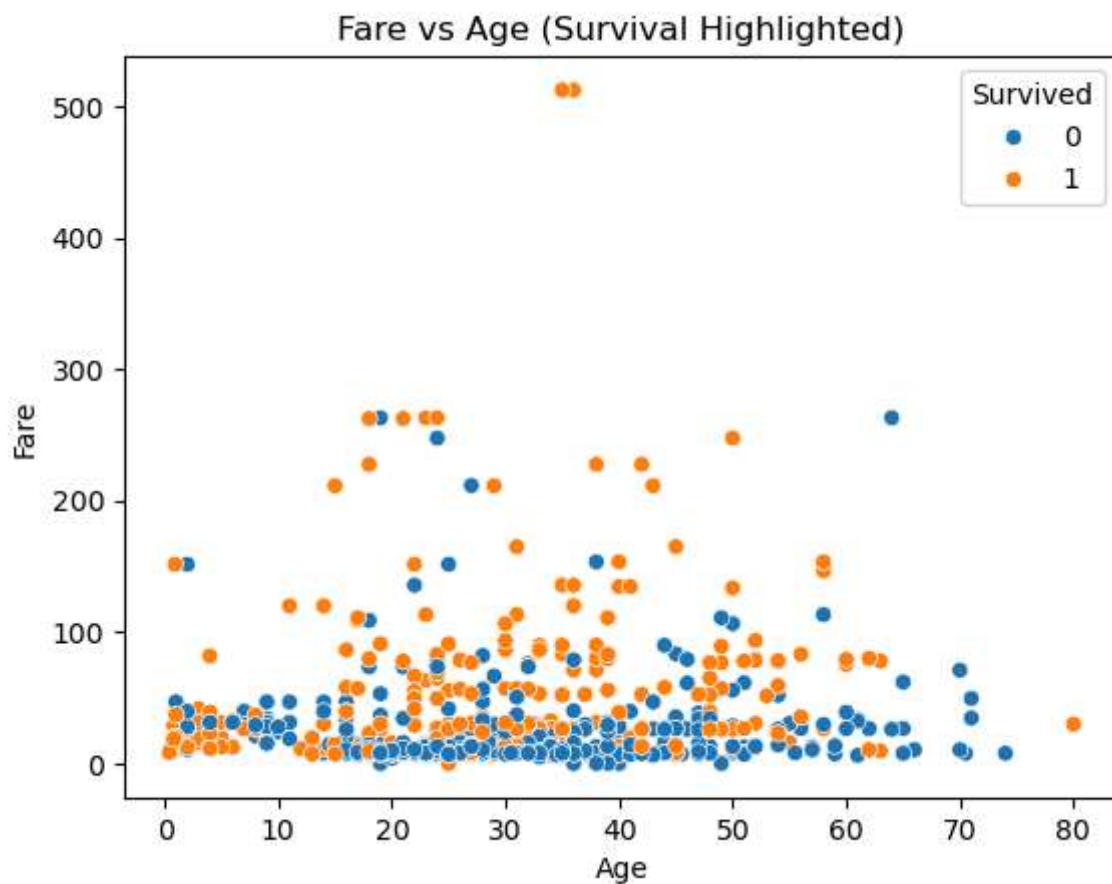
```
In [49]: sns.histplot(df['Age'].dropna(), kde=True, bins=30)
plt.title('Age Distribution of Passengers')
plt.xlabel('Age')
plt.ylabel('Count')
plt.show()
```



```
In [51]: sns.boxplot(x='Pclass', y='Fare', data=df)
plt.title('Fare Distribution by Passenger Class')
plt.xlabel('Passenger Class')
plt.ylabel('Fare')
plt.show()
```



```
In [53]: sns.scatterplot(x='Age', y='Fare', hue='Survived', data=df)
plt.title('Fare vs Age (Survival Highlighted)')
plt.xlabel('Age')
plt.ylabel('Fare')
plt.show()
```



## Summary of EDA

- Females had higher survival rates.
- Passengers in 1st class had better survival rates.
- Younger passengers tended to survive more.
- There is strong correlation between Fare and Pclass.

In [ ]: