Task_5_Exploratory Data Analysis (EDA)

April 28, 2025

[4]:

```
# Import Required Libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Set Seaborn style
sns.set_style('whitegrid')
%matplotlib inline
```

[5]:

```
# Load Titanic training dataset
train_df = pd.read_csv('train.csv')  # Make sure train.csv is in the same folder
```

[6]:

```
# Basic Info
train_df.info()
```

```
<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
```

[7]:
```
# Statistical Summary
train_df.describe()
```

[7]:
```
       PassengerId    Survived      Pclass         Age       SibSp  \
count   891.000000  891.000000  891.000000  714.000000  891.000000
mean    446.000000    0.383838    2.308642   29.699118    0.523008
std     257.353842    0.486592    0.836071   14.526497    1.102743
min       1.000000    0.000000    1.000000    0.420000    0.000000
25%     223.500000    0.000000    2.000000   20.125000    0.000000
50%     446.000000    0.000000    3.000000   28.000000    0.000000
75%     668.500000    1.000000    3.000000   38.000000    1.000000
max     891.000000    1.000000    3.000000   80.000000    8.000000

            Parch        Fare
count  891.000000  891.000000
mean     0.381594   32.204208
std      0.806057   49.693429
min      0.000000    0.000000
25%      0.000000    7.910400
50%      0.000000   14.454200
75%      0.000000   31.000000
max      6.000000  512.329200
```

[8]:
```
train_df.isnull().sum
```

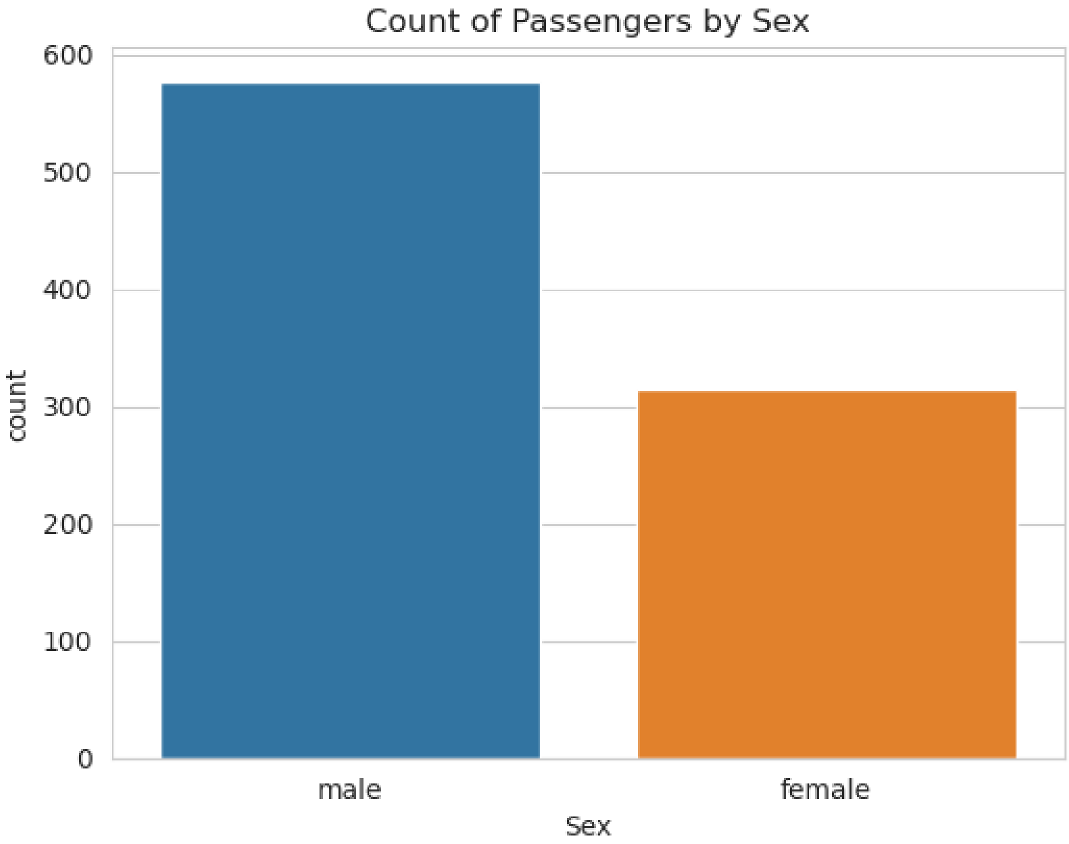
[8]:
```
<bound method NDFrame._add_numeric_operations.<locals>.sum of      PassengerId
Survived  Pclass   Name   Sex   Age  SibSp  Parch  Ticket  \
0           False     False   False  False  False  False  False  False    False
1           False     False   False  False  False  False  False  False    False
2           False     False   False  False  False  False  False  False    False
3           False     False   False  False  False  False  False  False    False
4           False     False   False  False  False  False  False  False    False
..            …         …     …     …     …     …     …     …       …
886         False     False   False  False  False  False  False  False    False
887         False     False   False  False  False  False  False  False    False
888         False     False   False  False  False   True  False  False    False
889         False     False   False  False  False  False  False  False    False
890         False     False   False  False  False  False  False  False    False

      Fare  Cabin  Embarked
0    False   True     False
1    False  False     False
2    False   True     False
3    False  False     False
4    False   True     False
..     …     …         …
886  False   True     False
```

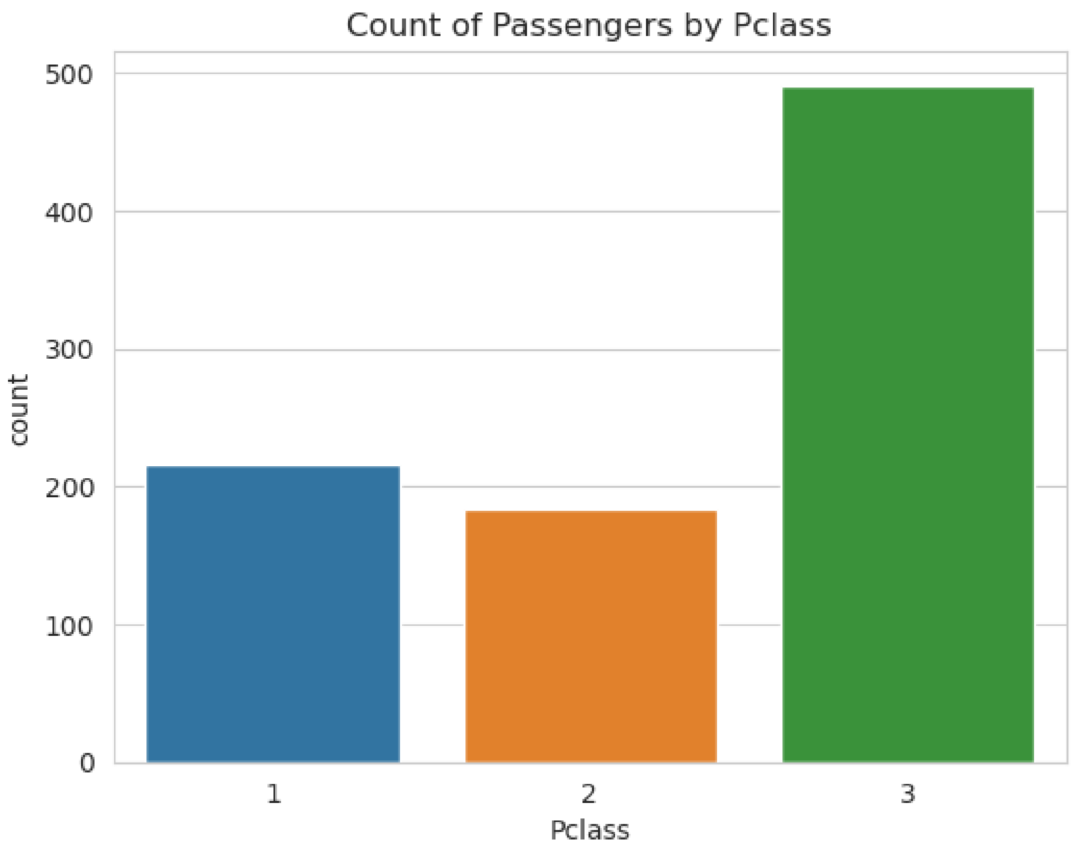
```
887  False  False     False
888  False   True     False
889  False  False     False
890  False   True     False

[891 rows x 12 columns]>
```

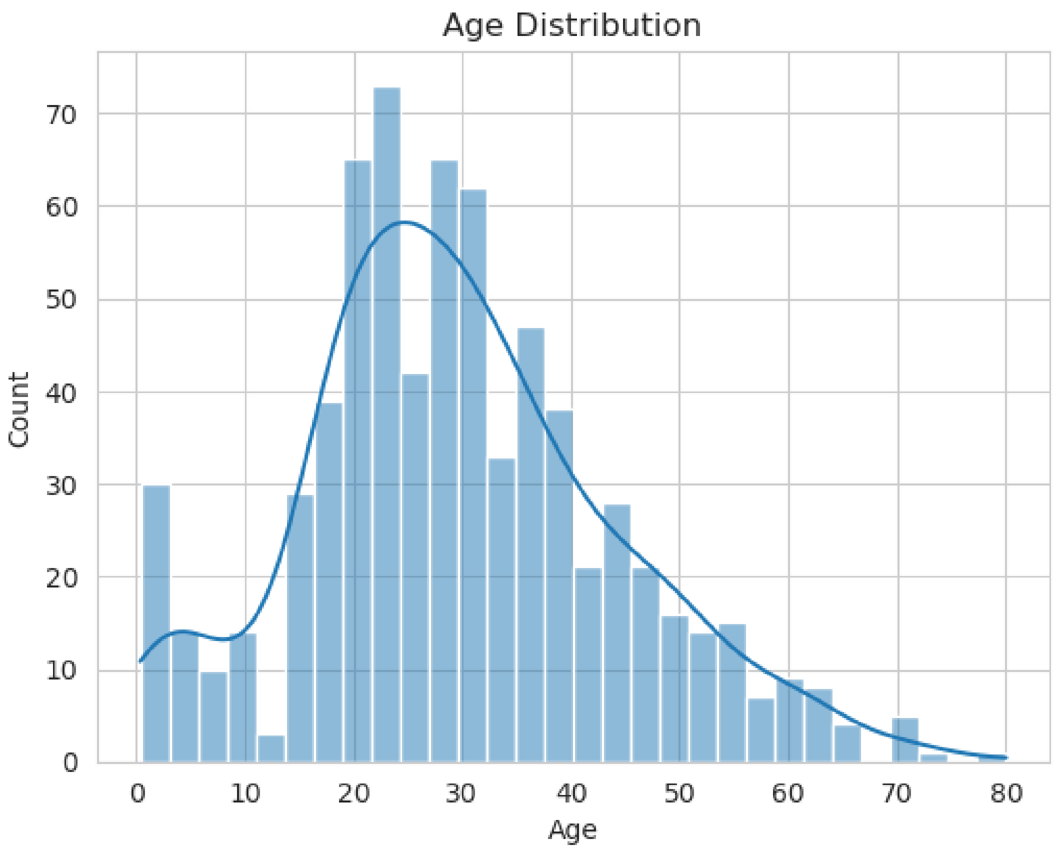
[9]:
```
# Categorical Columns
sns.countplot(data=train_df, x='Sex')
plt.title('Count of Passengers by Sex')
plt.show()
```



[11]:
```
sns.countplot(data=train_df,x='Pclass')
plt.title('Count of Passengers by Pclass')
plt.show()
```

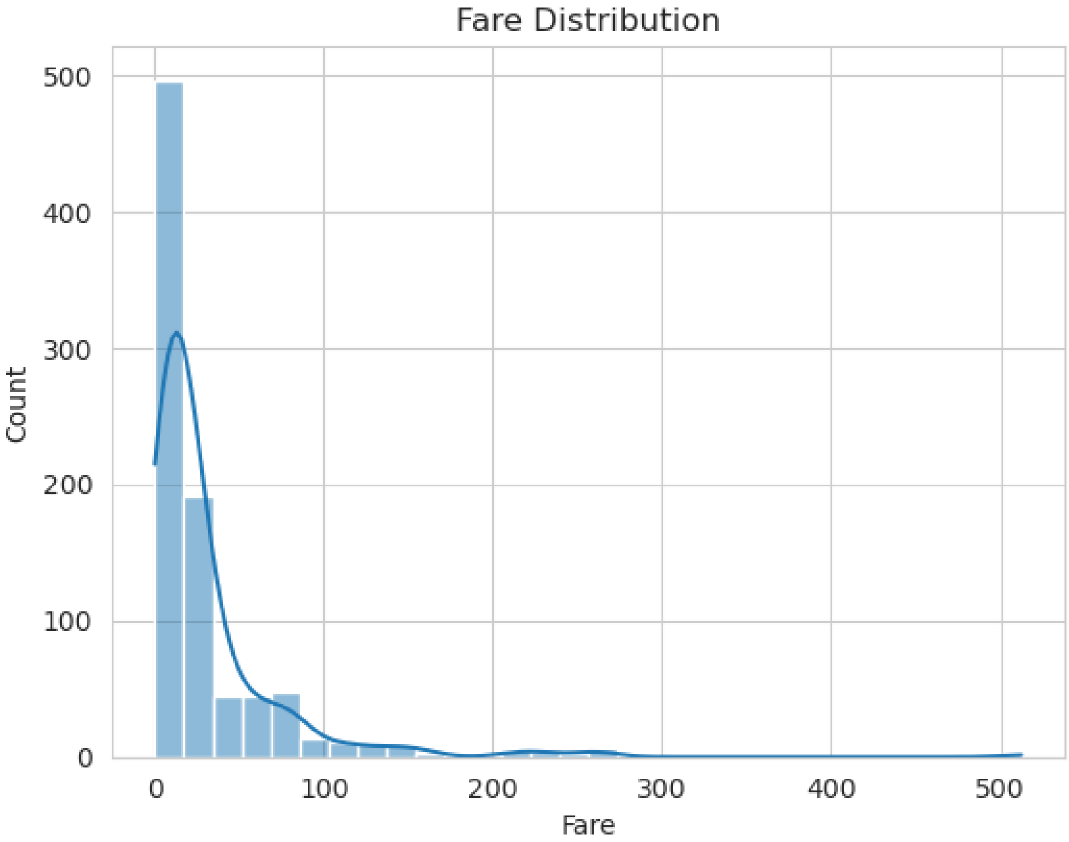


[12]:
```
sns.histplot(train_df['Age'].dropna(), kde=True, bins=30)
plt.title('Age Distribution')
plt.show()
```

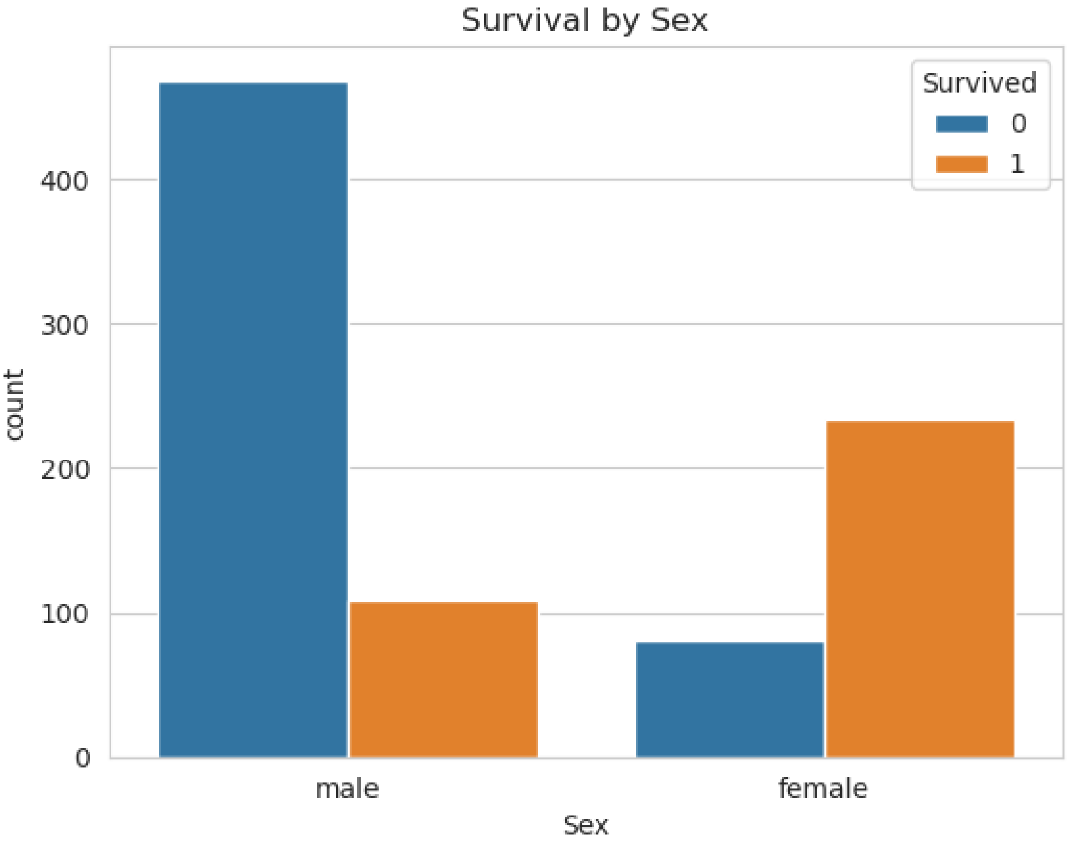


[13]:

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

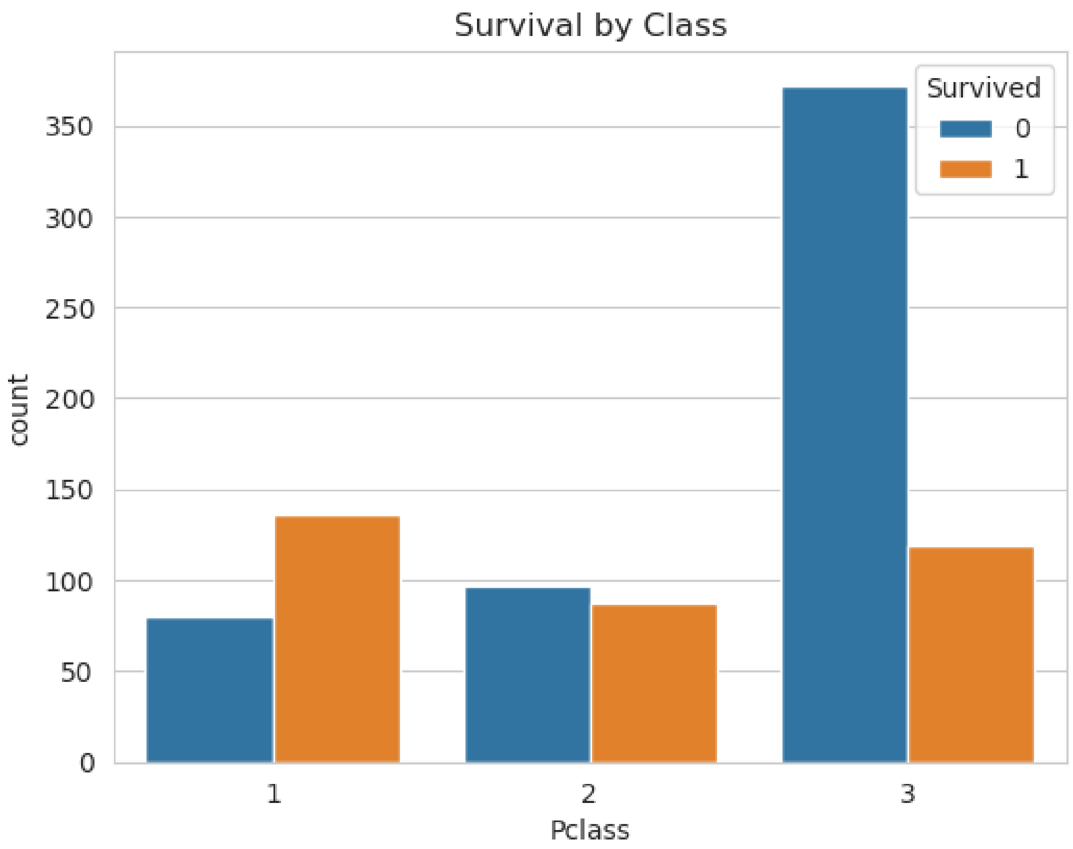


[14]:
```
# Sex vs Survived
sns.countplot(data=train_df, x='Sex', hue='Survived')
plt.title('Survival by Sex')
plt.show()
```



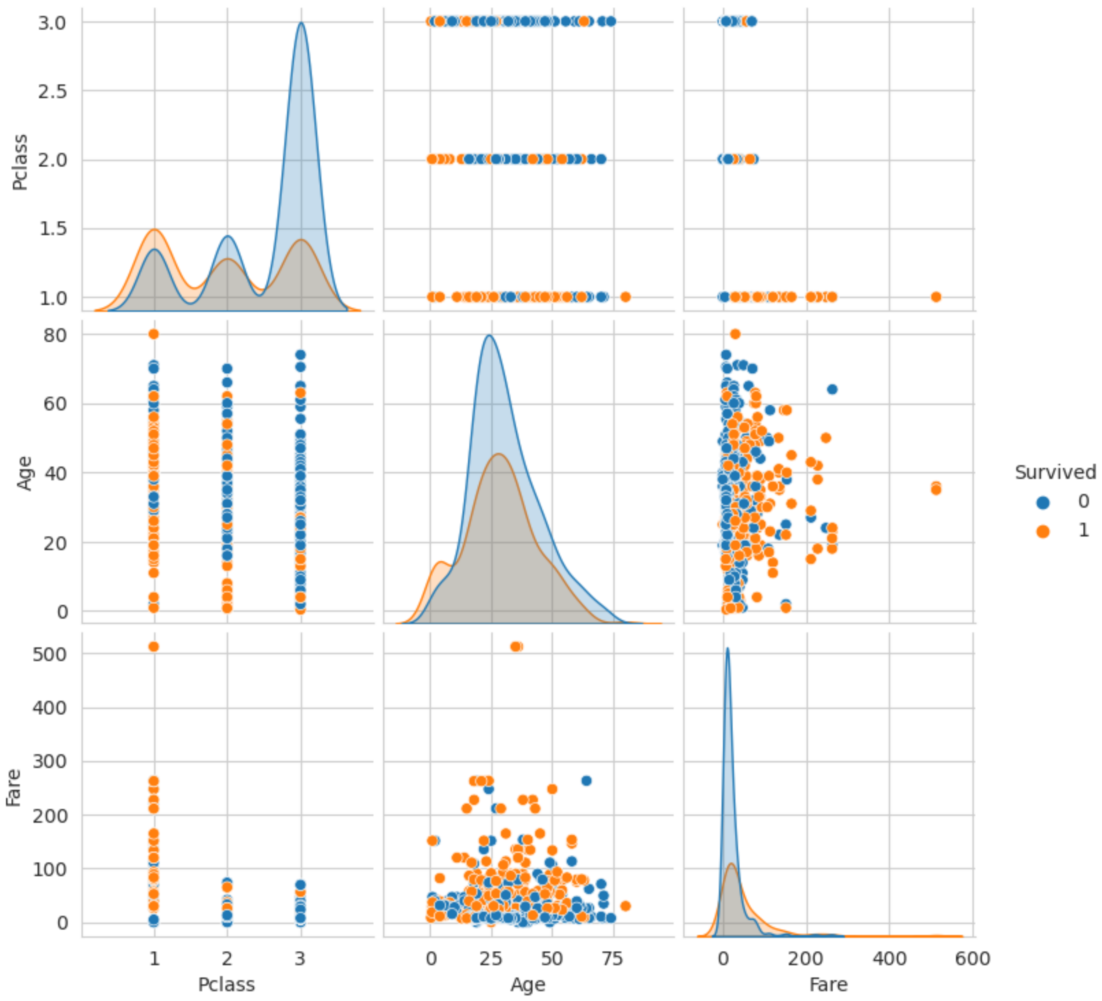
[16]:

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



[15]:
```
# Pairplot
sns.pairplot(train_df[['Survived', 'Pclass', 'Age', 'Fare']], hue='Survived')
plt.show()
```

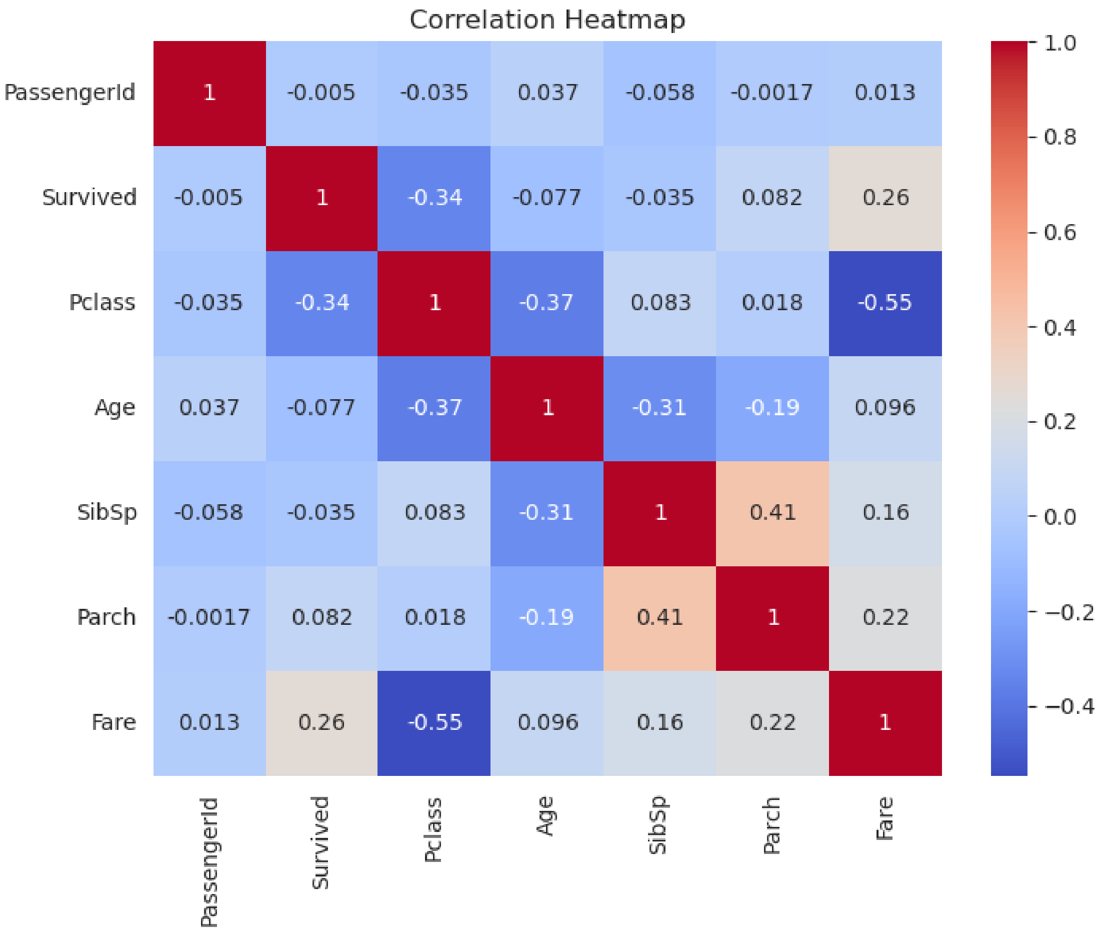
```
/opt/conda/envs/anaconda-panel-2023.05-py310/lib/python3.11/site-
packages/seaborn/axisgrid.py:118: UserWarning: The figure layout has changed to
tight
  self._figure.tight_layout(*args, **kwargs)
```



```
[17]: # Correlation Heatmap

      numeric_df = train_df.select_dtypes(include=['float64', 'int64'])

      # Now plot heatmap
      plt.figure(figsize=(8,6))
      sns.heatmap(numeric_df.corr(), annot=True, cmap='coolwarm')
      plt.title('Correlation Heatmap')
      plt.show()
```



1 Summary of Findings

- **Gender Impact**: Females had a significantly higher survival rate than males.
- **Class Impact**: 1st Class passengers had better survival chances.
- **Fare**: Higher fare-paying passengers survived more often.
- **Age**: Younger passengers (especially children) had better chances of survival.
- **Embarked Port**: Most passengers boarded from Southampton.
- **Missing Data**: 'Cabin' and 'Age' columns have missing valuesediction!