20250411_01

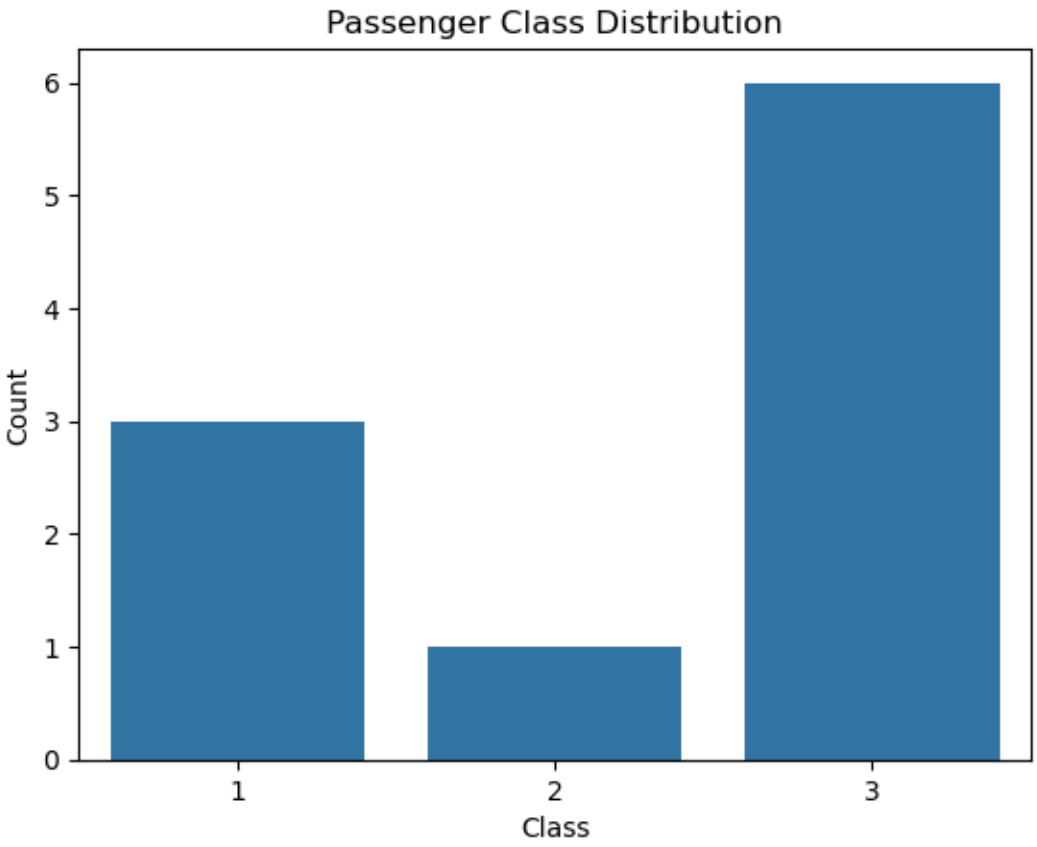
April 11, 2025

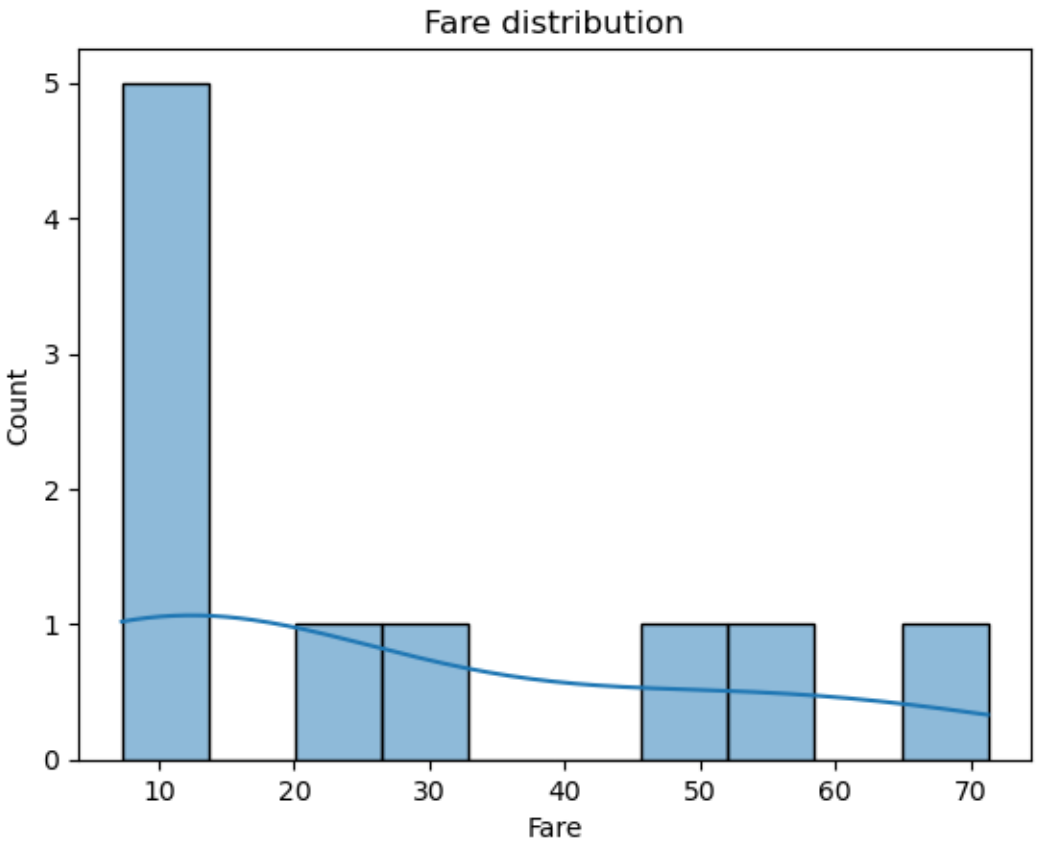
[1]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

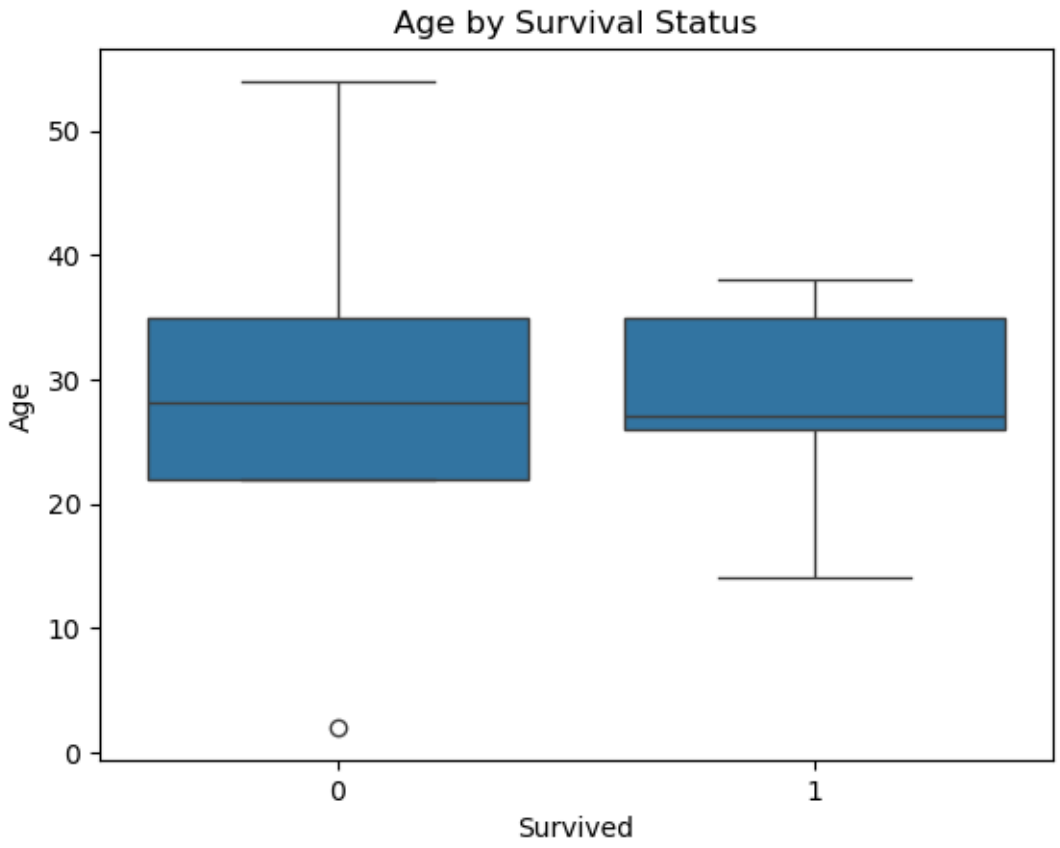
[7]:
```
def clean_titanic_data(filepath):
    df = pd.read_csv(filepath)

    # Use mean value to fill nulls in Age
    mean_age = df['Age'].mean()
    df['Age'] = df['Age'].fillna(mean_age)

    #Change the name of PassengerId to ID
    df.rename(columns = {'PassengerId':'ID'}, inplace = True)
    df.rename(columns = {'Pclass':'Class'}, inplace = True)
    df.rename(columns = {'Sex':'Gender'}, inplace = True)

    return df
```

[5]:
```
def plot_hist(df, column, title = ''):
    sns.histplot(df[column], bins = 10, kde = True)
    plt.title(title)
    plt.xlabel(column)
    plt.ylabel('Count')
    plt.show()

def plot_count(df, column, title = ''):
    sns.countplot(x = column, data = df)
    plt.title(title)
    plt.xlabel(column)
    plt.ylabel('Count')
    plt.show()

def plot_box(df, x, y, title = ''):
    sns.boxplot(x = x, y = y, data = df)
    plt.title(title)
    plt.xlabel(x)
    plt.ylabel(y)
    plt.show()
```

[15]:
```
df = clean_titanic_data('titanic_day2.csv')
plot_count(df, 'Class', 'Passenger Class Distribution')
plot_hist(df, 'Fare', 'Fare distribution')
plot_box(df, 'Survived', 'Age', 'Age by Survival Status')
```




Fare distribution
Count
Fare
0
1
2
3
4
5
10
20
30
40
50
60
70


Age by Survival Status
Age
50
40
30
20
10
0
0
1
Survived

[ ]: