20250410_01

April 10, 2025

1 Titanic Dataset: Mini Analysis Report

In this mini project, we analyze a simplified Titanic dataset to explore potential relationships between fare, passenger class and survival status. Visualization techniques are applied to illustrate data distributions and group differences.

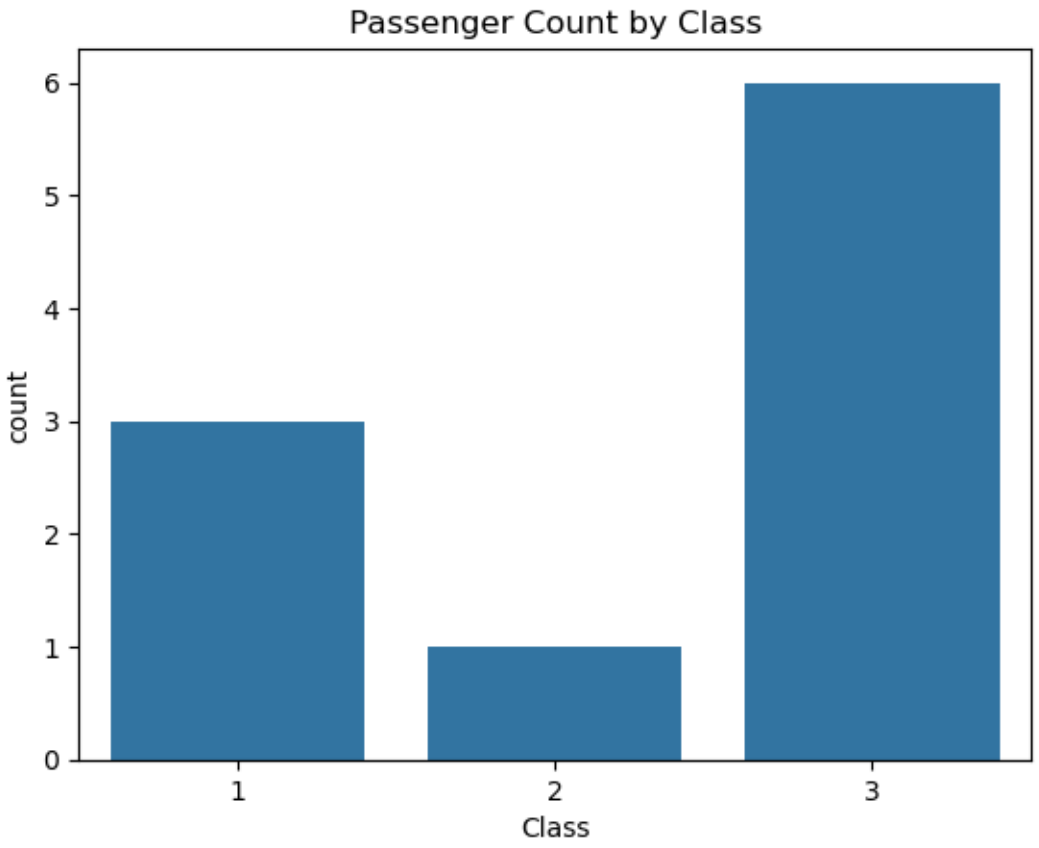
[5]:
```
import pandas as pd

df = pd.read_csv('cleaned_titanic.csv')
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10 entries, 0 to 9
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   ID        10 non-null     int64
 1   Survived  10 non-null     int64
 2   Class     10 non-null     int64
 3   Name      10 non-null     object
 4   Gender    10 non-null     object
 5   Age       10 non-null     float64
 6   Fare      10 non-null     float64
dtypes: float64(2), int64(3), object(2)
memory usage: 692.0+ bytes
```

The dataset contains 10 records with no missing values.

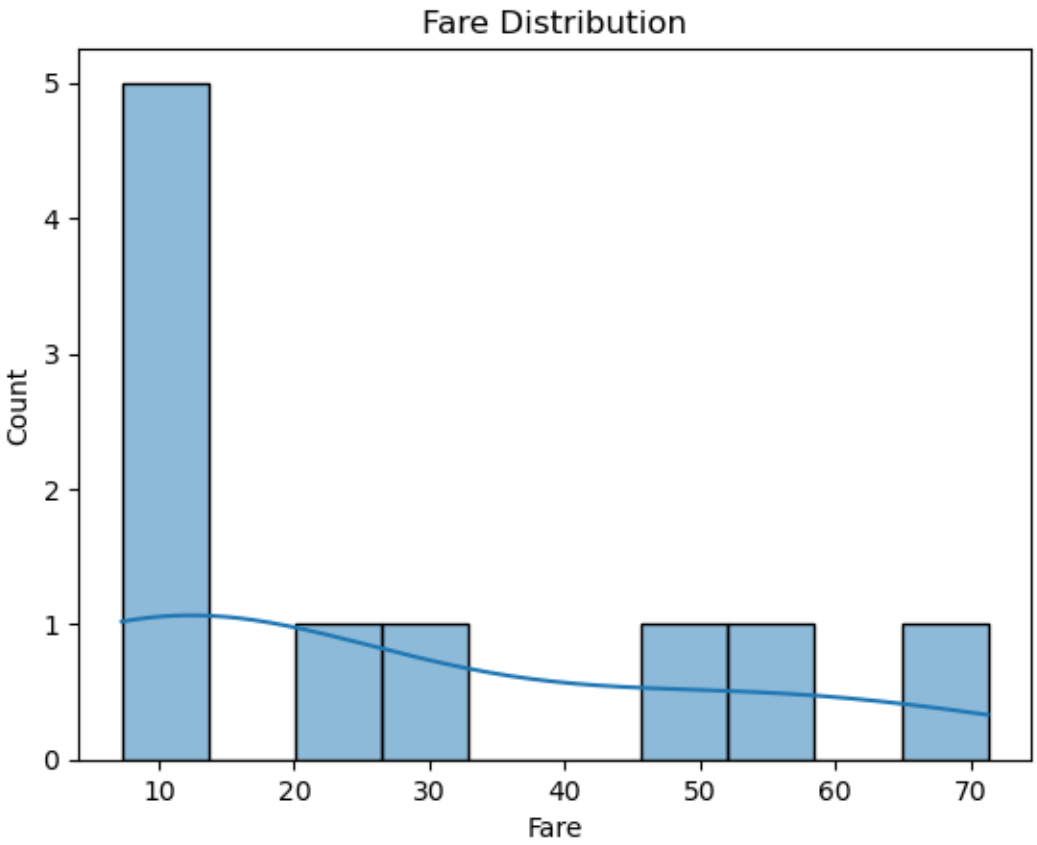
[7]:
```
import seaborn as sns
import matplotlib.pyplot as plt

sns.countplot(x = 'Class', data = df)
plt.title('Passenger Count by Class')
plt.show
```

[7]:
```
<function matplotlib.pyplot.show(close=None, block=None)>
```



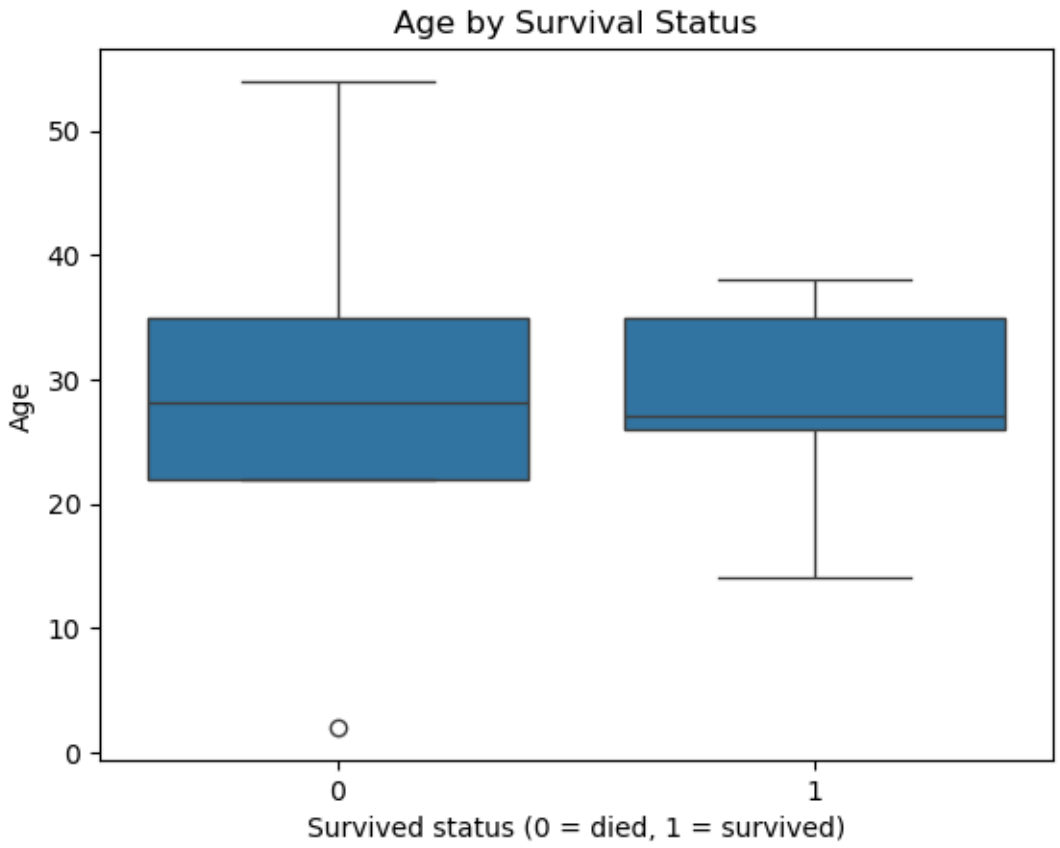
Most passengers were in class 3, suggesting tickets with lower price were sold most commonly.

[9]:
```
sns.histplot(df['Fare'], bins = 10, kde = True)
plt.title('Fare Distribution')
plt.show()
```



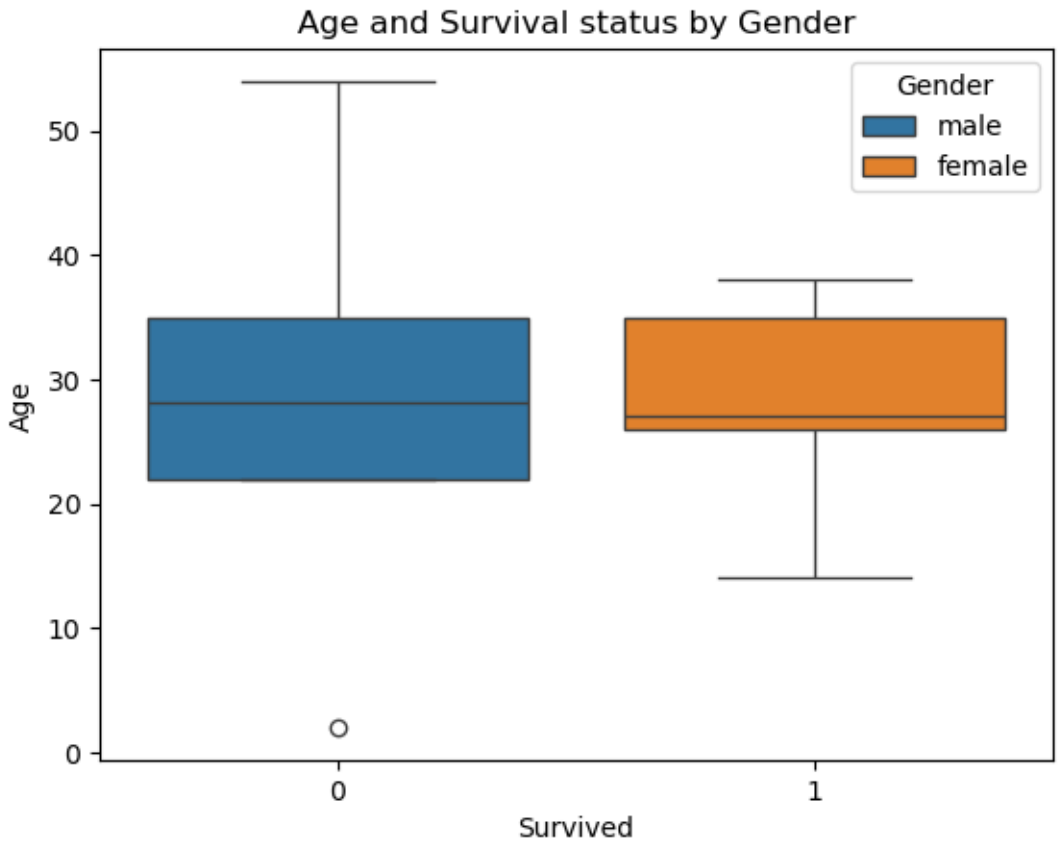
Fare distribution is right-skewed, with most tickets priced around 10 dollars. A few tickets were sold in higher price, which create visible outliers.

[21]:
```
sns.boxplot(x = 'Survived', y = 'Age', data = df)
plt.xlabel('Survived status (0 = died, 1 = survived)')
plt.title('Age by Survival Status')
plt.show()
```



The median ages of survivors and non-survivors are similar. However, the most extreme ages (both youngest and oldest) belong to non-survivors, indicating greater age deviance in said group.

[23]:
```
sns.boxplot(x = 'Survived', y = 'Age', hue = 'Gender', data = df)
plt.title('Age and Survival status by Gender')
plt.show()
```



The chart shows that all survivors were female and all non-survivors were male in this sample.

1.0.1 Conclusion:

- Most Passengers in the dataset were from class 3 and the fare distribution is uneven, with the majority paying lower prices.
- There is no clear linear relationship between age and survival status, though non-survivors exhibit greater variance in age.
- Gender appears to be the strongest factor in survival. All female passengers survived, while all male passengers did not, in this sample.

[ ]: