Mohammed Shamlan

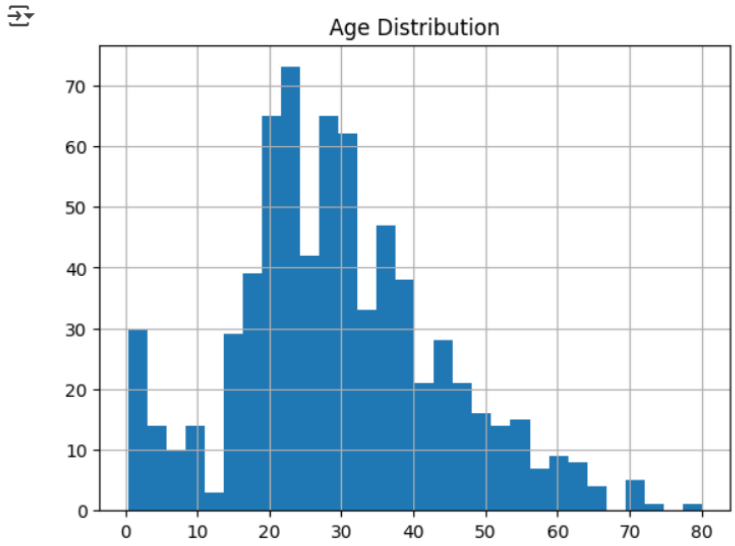
Task 5: Exploratory Data Analysis (EDA)

Topic:
Observations for visuals of dataset & summary of findings

Observations for visuals:

1) Histograms for numerical features:

```
[20] data['Age'].hist(bins=30)
     plt.title('Age Distribution')
     plt.show()
```



The distribution appears to be right-skewed. This means the tail on the right side of the distribution is longer or fatter than the tail on the left side. In simpler terms, there are more individuals in the younger age groups compared to the older age groups in this dataset.

The peak of the distribution (the mode) seems to be around the 20-30 age range. This suggests that this age group has the highest frequency of observations in the dataset.

There's a noticeable number of observations in the younger age groups, particularly between 0 and 30. The bars in this range are generally taller than those in the older age ranges.

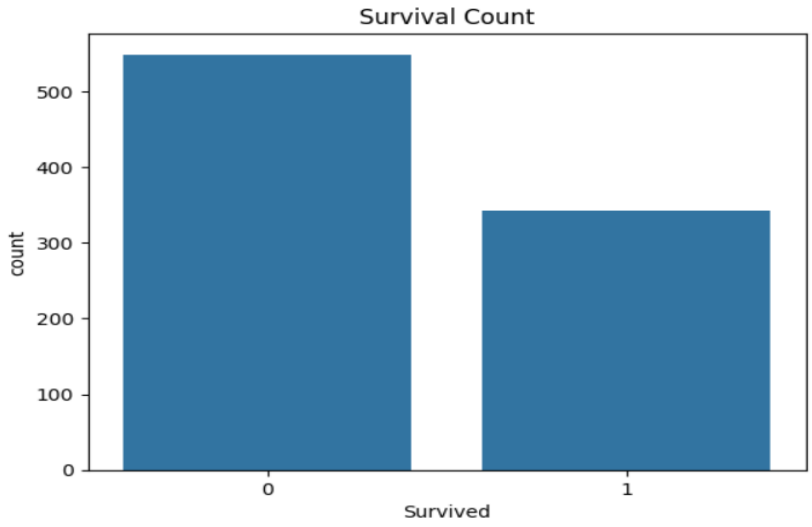
The frequency of observations decreases as age increases beyond the 30-year mark. The bars become progressively shorter, indicating fewer individuals in the older age categories.

There are relatively few observations in the very young (0-10) and very old (70+) age groups. The bars at the extreme ends of the x-axis are quite short.

The histogram uses 30 bins, which provides a reasonable level of detail in visualizing the distribution.

2) plots for categorical features

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



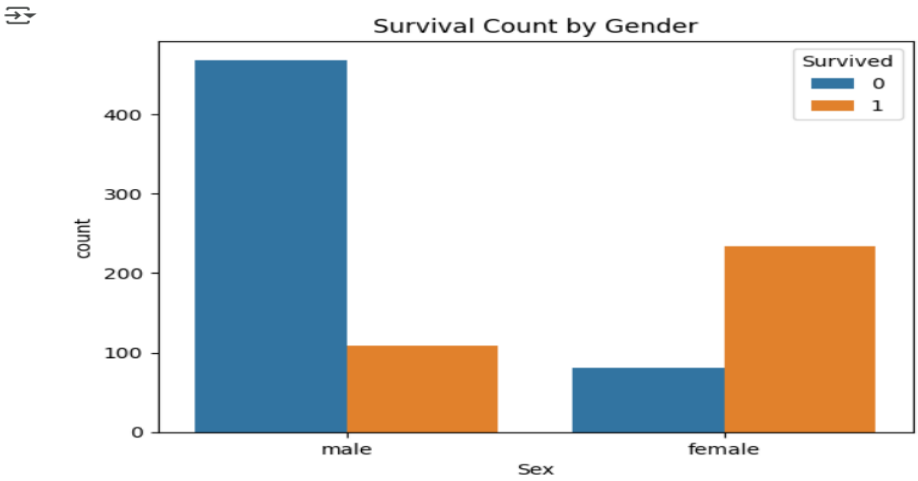
Two Categories: The plot clearly shows two distinct categories for the 'Survived' variable, labeled as '0' and '1'. It's common in such contexts for '0' to represent individuals who did not survive and '1' to represent those who did.

Higher Count of Non-Survivors: The bar corresponding to '0' (did not survive) is significantly taller than the bar corresponding to '1' (survived). This indicates that there were more individuals in the dataset who did not survive compared to those who did.

3) Relationships between two features:

- **Survival by gender:**

```
[24] sns.countplot(x='Sex', hue='Survived', data=data)
     plt.title('Survival Count by Gender')
     plt.show()
```



Two Sex Categories: The x-axis clearly shows two categories: 'male' and 'female'.

Survival Hue: The plot uses different colors (hue) to distinguish between those who did not survive (represented by '0', in blue) and those who survived (represented by '1', in orange).

Survival Count for Males:

- The number of males who did not survive (blue bar) is significantly higher, appearing to be around 450.
- The number of males who survived (orange bar) is considerably lower, around 110.

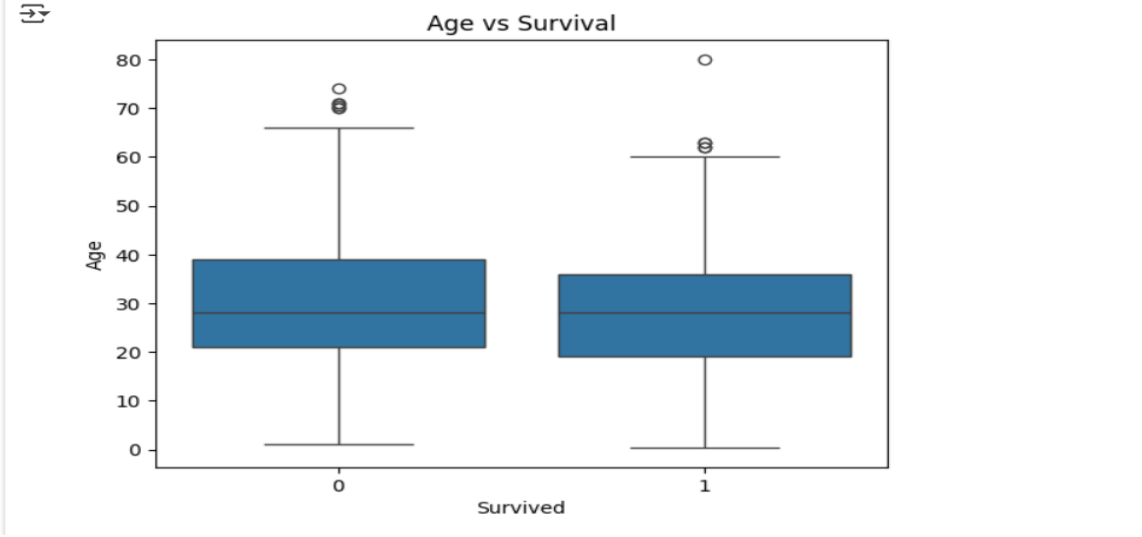
Survival Count for Females:

- The number of females who did not survive (blue bar) is lower than the number of males who did not survive, appearing to be around 80.
- The number of females who survived (orange bar) is notably higher than the number of males who survived, around 230.

Gender Disparity in Survival: There's a clear disparity in survival rates between genders in this dataset. While more males did not survive compared to females, a much higher proportion of females survived compared to males.

- **Age vs Survival:**

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



Median Age: Survivors had a slightly lower median age (around 28) compared to non-survivors (around 30).

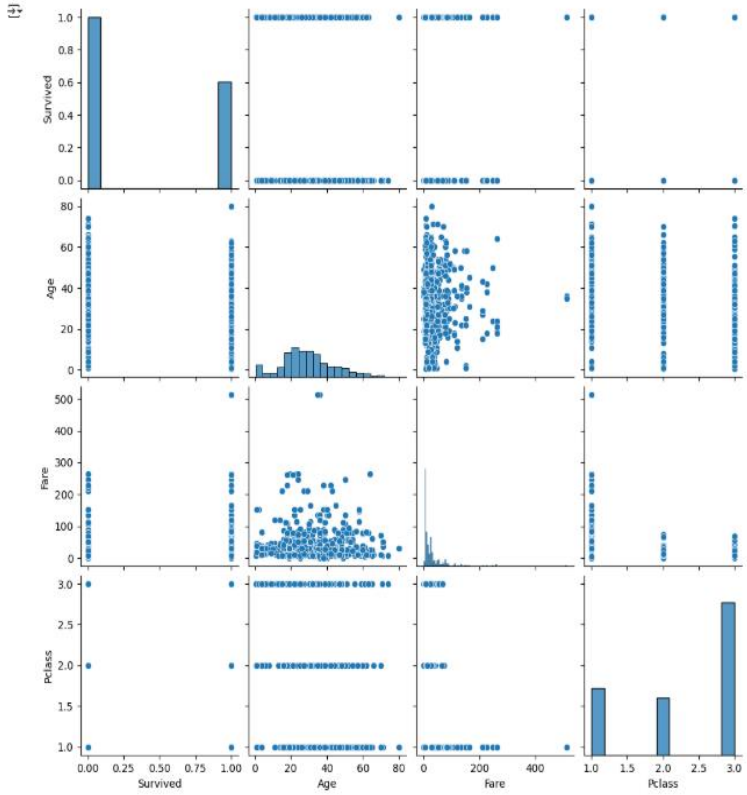
Age Range: The central age ranges (IQR) were similar for both groups (roughly 20-36 for survivors and 22-39 for non-survivors).

Younger Survivors: Survivors included some younger individuals (whiskers extended to a younger age).

Outliers: Older outliers existed in both groups.

4) Pairplots

```
[ ] sns.pairplot(data[['Survived', 'Age', 'Fare', 'Pclass']])
    plt.show()
```



Survival: More deaths than survivals. Higher survival with higher fare and better class (P class 1). Possibly slightly lower survival with increasing age.

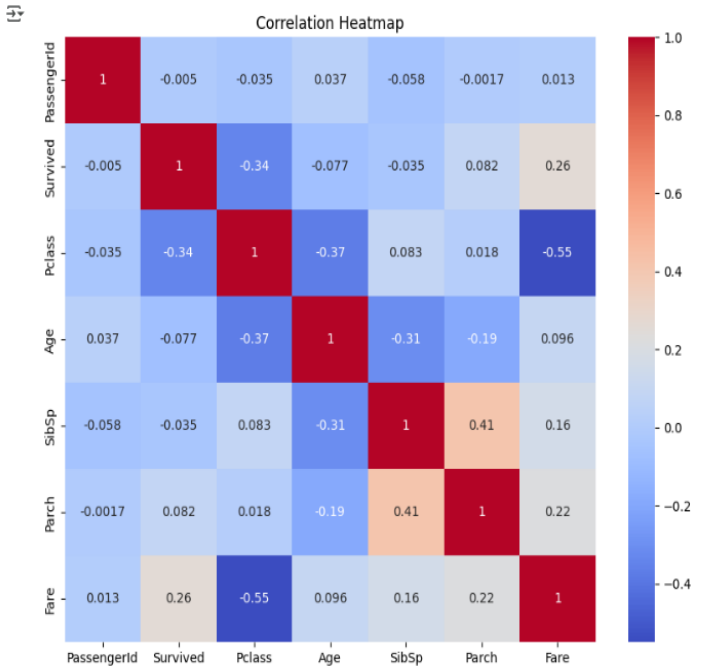
Age: Right-skewed (more younger people). Older passengers tended to be in higher P class. No clear link with fare.

Fare: Heavily right-skewed (mostly lower fares). Higher fare strongly linked to better P class. No clear link with age.

P class: Most were in P class 3. Better P class linked to higher fare. Older passengers tended to be in better P class. Survival better in better P class.

5) Heatmap for correlations

```
plt.figure(figsize=(10,8))
numeric_data = data.select_dtypes(include=['int64', 'float64'])  # Select only numeric columns
sns.heatmap(numeric_data.corr(), annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```



Survival: Better for higher class (lower P class) and higher fare. Little link to age.

P class: Higher class means higher fare. Weakly linked to older age.

Fare: Higher fare for higher class. Weakly linked to having more parents/children.

Age: Older passengers tended to have fewer siblings/spouses or parents/children.

SibSp & Parch: Moderately positively correlated (more family members together).

Summary of Findings

1. General Dataset Overview

- The dataset contains **891 rows** and **12 columns**.
- Several columns like Age, Cabin, and Embarked have **missing values**.
- Cabin has too many missing values (~77% missing).

2. Target Variable Analysis (Survived)

- Around **38%** of the passengers survived (Survived = 1).
- Around **62%** did not survive (Survived = 0).

3. Gender vs Survival

- **Females** had a significantly higher survival rate than males.
    - **Survival rate for females**: ~74%.
    - **Survival rate for males**: ~19%.
- Gender played a major role in survival chances.

4. Passenger Class (P class) vs Survival

- **1st Class** passengers had the highest survival rate (~63%).
- **3rd Class** passengers had the lowest survival rate (~24%).
- Higher-class passengers were prioritized for rescue.

5. Age Distribution

- The majority of passengers were between **20-40 years old**.
- **Children (<10 years old)** had a better survival rate than young adults.
- **Older passengers (>60)** had low survival chances.

6. Fare vs Survival

- Higher fare prices were associated with higher survival rates.
- Most of the high-paying passengers were from 1st class

7. Embarked Port Analysis

- Most passengers boarded from **Southampton (S)**.
- Passengers who boarded at **Cherbourg (C)** had a better survival rate.

8. Missing Data

- Age had missing values (~20%).
- Cabin had very high missing values (~77%); might be dropped or further analysed.
- Embarked had very few missing values (just 2).

9. Correlations (from Heatmap)

- Fare and P class have a moderate negative correlation (-0.55).
- Survived is positively correlated with Fare and P class (1st class passengers paid more and survived more).
- No extremely strong multicollinearity between variables.