Task5 Observations:

Dataset Overview

The dataset consists of three CSV files related to Titanic passenger survival prediction:

- train.csv: Training data including survival status
- test.csv: Test data without survival status
- gender_submission.csv: Sample submission for predictions

Missing Values

- Age has 177 missing entries, which were filled with the **median** age.
- Cabin has 687 missing entries, considered too sparse to fill, so dropped or treated as a categorical feature.
- Embarked had 2 missing values, which were filled with the **mode** (most frequent value).

Key Insights from Data Exploration:

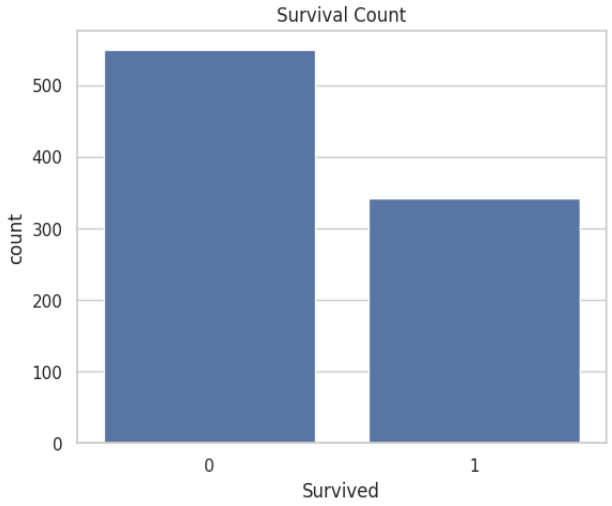
1. **Survival Rate**:
    - Around **38%** of the passengers survived, meaning only about one in three passengers made it through the disaster.
2. **Gender Impact**:
    - There was a clear **gender disparity** in survival. **Females** had a significantly higher chance of survival compared to males, which may reflect societal norms at the time or the "women and children first" policy.
3. **Class Impact**:
    - The survival rate was also strongly influenced by **class**. Passengers in **Pclass 1** (First Class) had the highest survival rates, while those in **Pclass 3** (Third Class) had the lowest, possibly due to proximity to lifeboats or the location of their cabins.
4. **Age Distribution**:
    - The **average age** of passengers was around **29 years**. A large proportion of passengers were quite young, with children and young adults making up a significant portion of the onboard population.
5. **Fare Distribution**:

- The **Fare** feature is highly **right-skewed**, meaning most passengers paid a relatively low fare, but a small number of passengers paid very high fares (likely the first-class passengers).

6. **Embarked Locations**:
    - Most passengers boarded at **Southampton (S)**, followed by **Cherbourg (C)**, and a smaller number boarded at **Queenstown (Q)**.

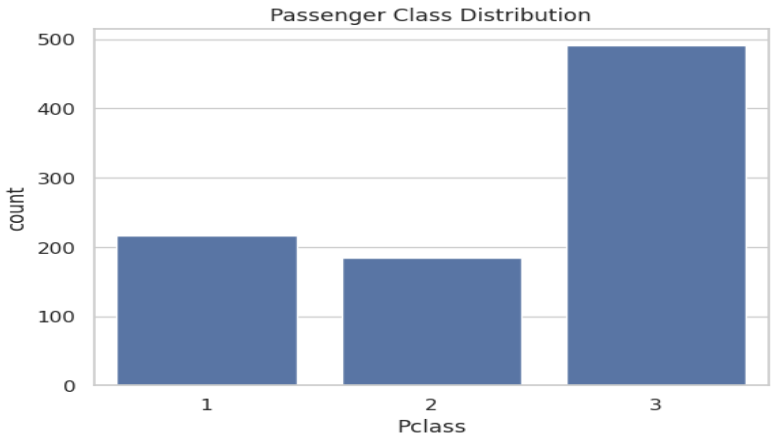
Plots

1. **Survival count:**



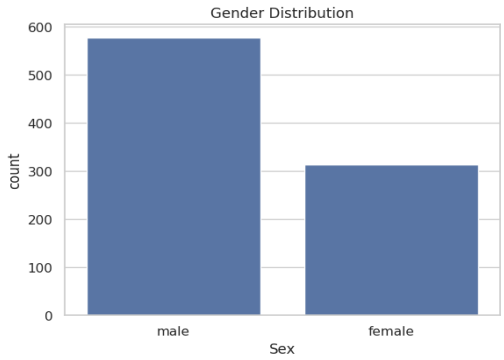
It is showing more people didn't survived appx.200 more than survived.

2. **Passenger Class Distribution**:



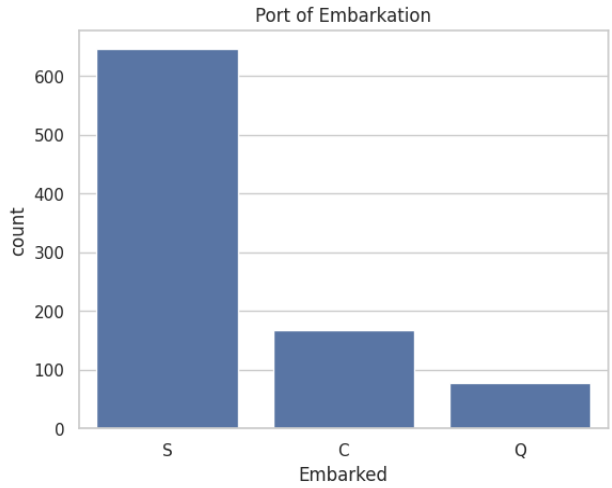
It is showing more people are in 3rd class.

3. **Gender distribution:**



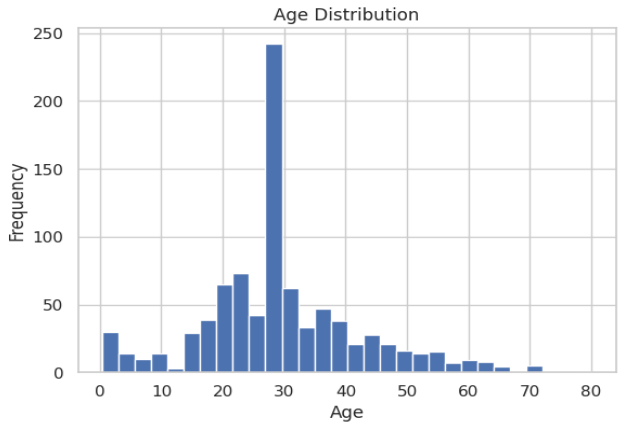
It is showing more males are there in the ship.

4. **Port of Embarkation:**



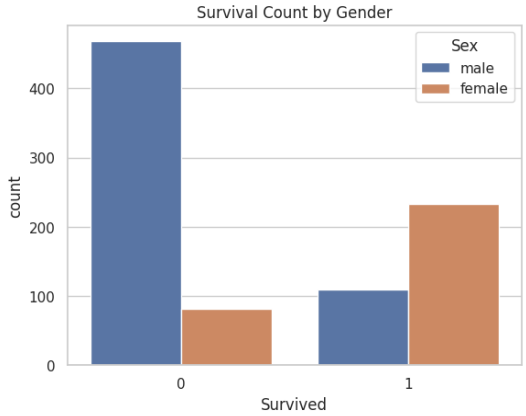
It is showing majority of passengers boarded at **Southampton** and least from **Queenstown.**

5. **Age Distribution:**



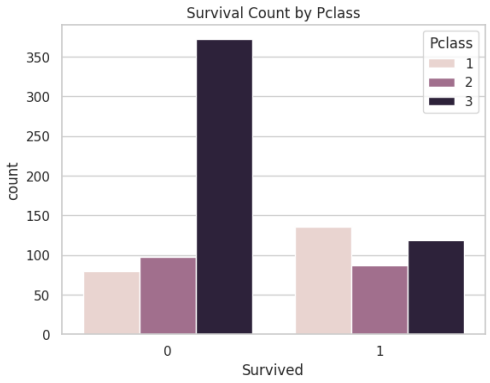
It is showing more people are around the age **30** and looks like a right skewed distribution as more people on the right side of peak.

6. Survival Count by Gender:



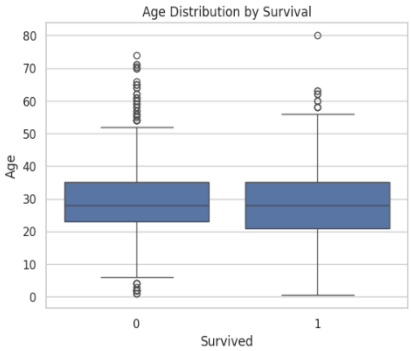
It is showing more females had been survived than males.

7. Survival Count by Passenger class:



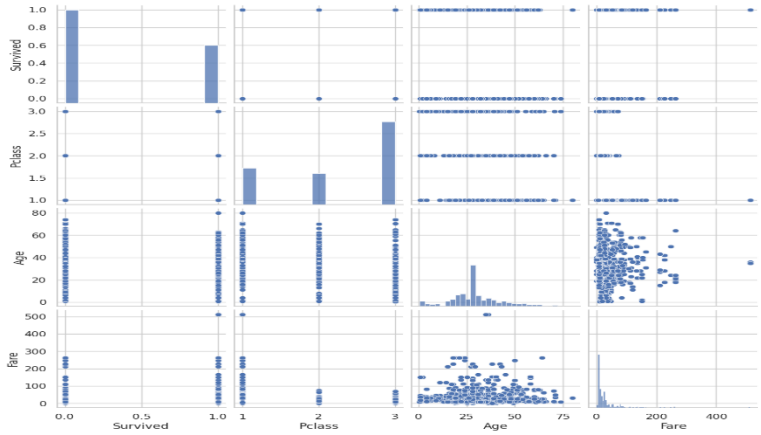
It is showing that more people in the 3rd class had been dead and 1st class people had been survived well.

8. Age Distribution by Survival:



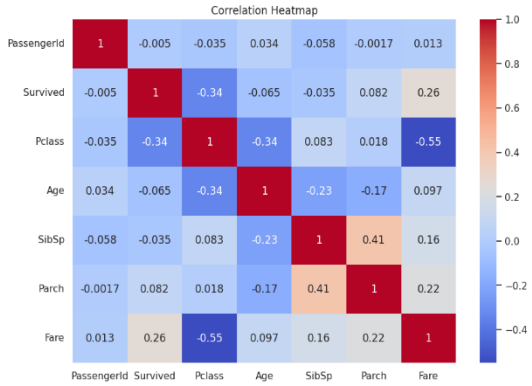
It is showing more survived people are less age than who died.

9. **Pair plot:**



It is showing that passengers in **1st class** and those who paid higher **fares** had better survival chances. **Fare** is heavily skewed with a few high-paying outliers.

10. **Correlation Heatmap:**



Heatmap is showing a **positive correlation between Fare and survival**, and a **negative correlation between Pclass and survival**. No other strong correlations observed.

11. **Fare Distribution with Outliers:**

The boxplot indicates that most passengers paid low fares, with a few exceptionally high-paying outliers.