Exploratory Data Analysis Report: Titanic Dataset

Data Overview: The Titanic dataset contains information about passengers aboard the Titanic, including demographic details, travel information, and survival status. The key variables include:

- **Survived:** Survival status (0 = No, 1 = Yes)
- **Pclass:** Passenger class (1st, 2nd, 3rd)
- **Name:** Passenger name
- **Sex:** Passenger sex (male, female)
- **Age:** Passenger age
- **SibSp:** Number of siblings/spouses aboard
- **Parch:** Number of parents/children aboard
- **Ticket:** Ticket number
- **Fare:** Passenger fare
- **Cabin:** Cabin number
- **Embarked:** Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

Key Findings:

1. **Survival Rate:**
    - The overall survival rate is approximately 38%.
    - Females had a significantly higher survival rate than males.
    - Passengers in higher classes (1st and 2nd) had a greater chance of survival compared to those in 3rd class.
2. **Age and Fare:**
    - The age distribution is slightly right-skewed, with most passengers being between 20 and 40 years old.

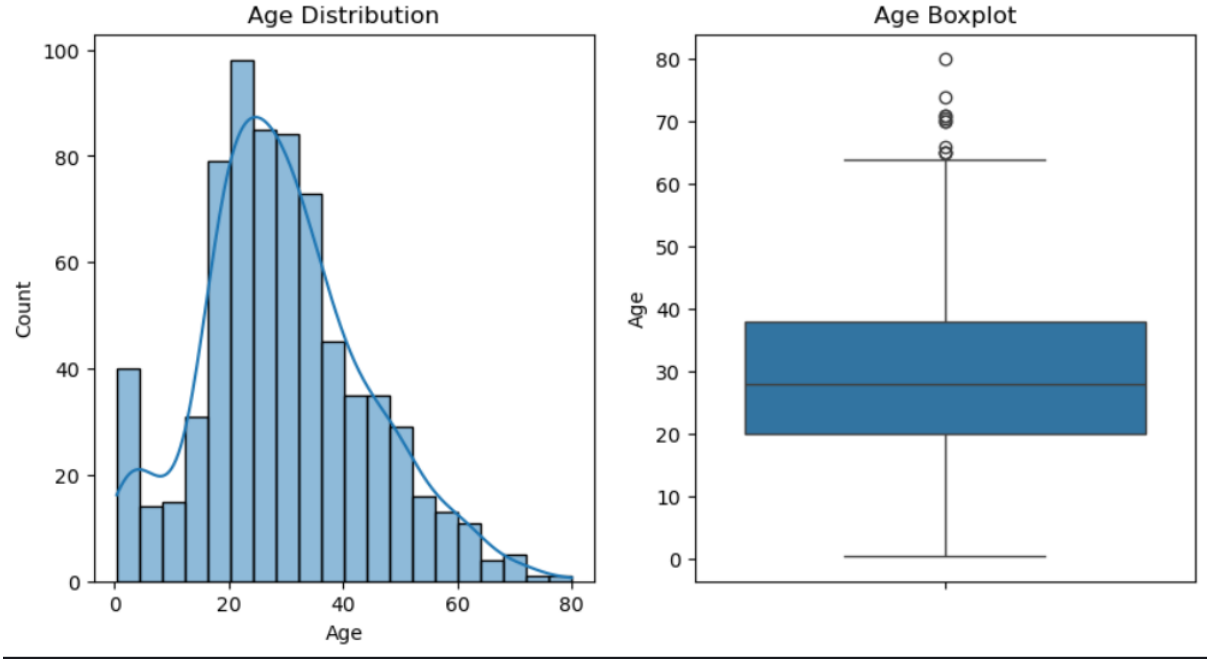
   - Passenger fares are heavily right-skewed, indicating that most passengers paid lower fares, with a few paying significantly higher fares.
   - There is no strong linear correlation between age and fare.

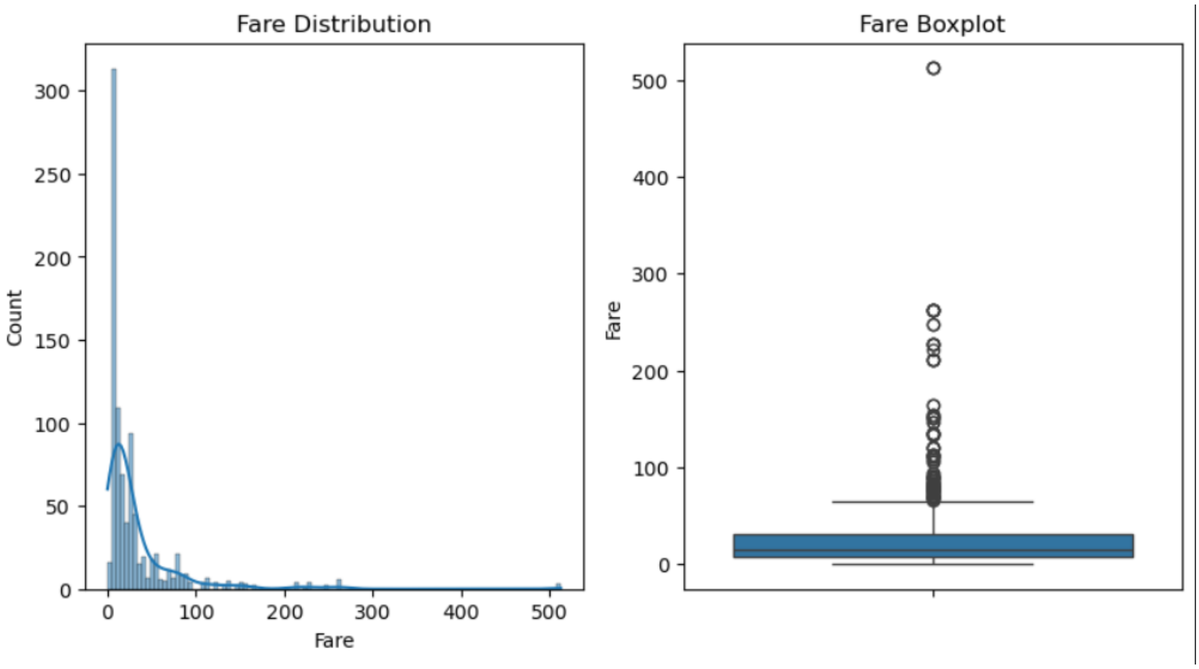
3. **Passenger Class:**
   - Most passengers were in the 3rd class.
   - Survival rates decrease as passenger class decreases.
4. **Embarked:**
   - Most passengers embarked from Southampton (S).
   - The port of embarkation appears to have some correlation with survival, though less pronounced than sex or passenger class.
5. **Family Size:**
   - Most passengers traveled without siblings/spouses or parents/children.
   - Passengers with a small family size had a higher survival rate.

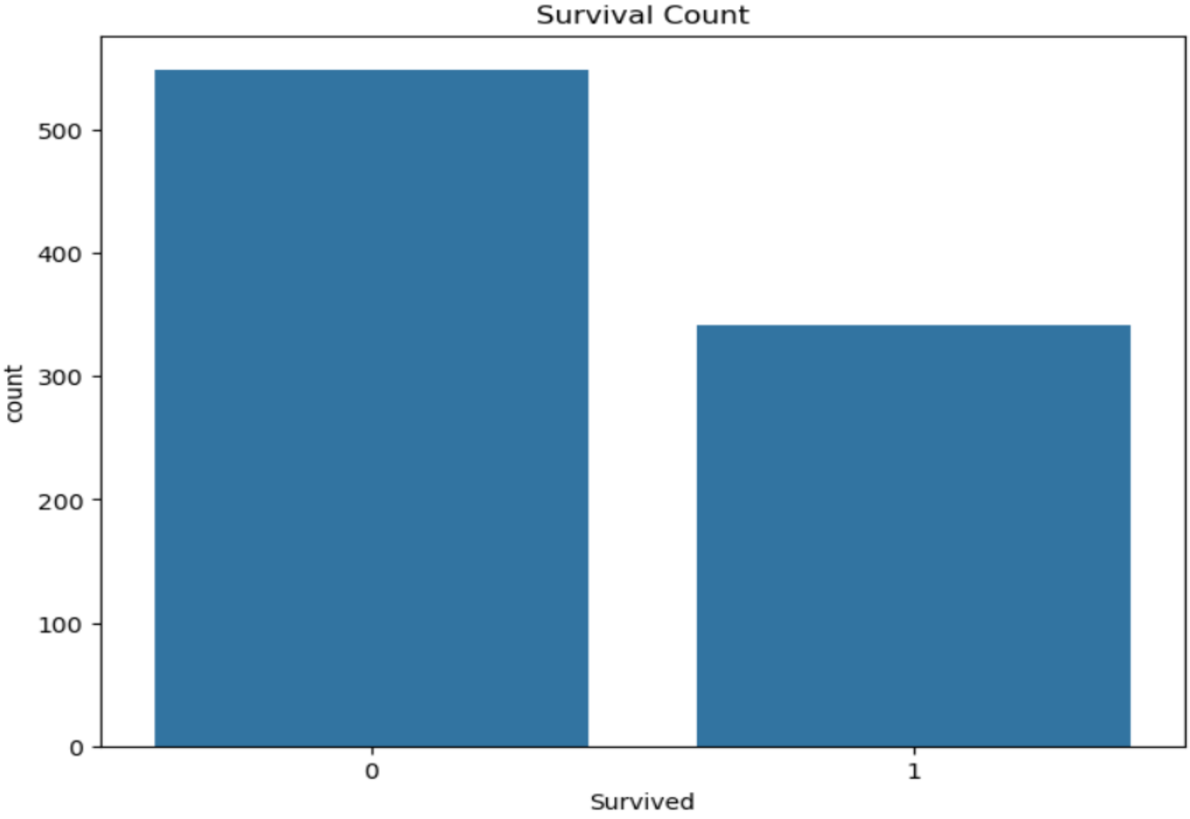
Visualizations:

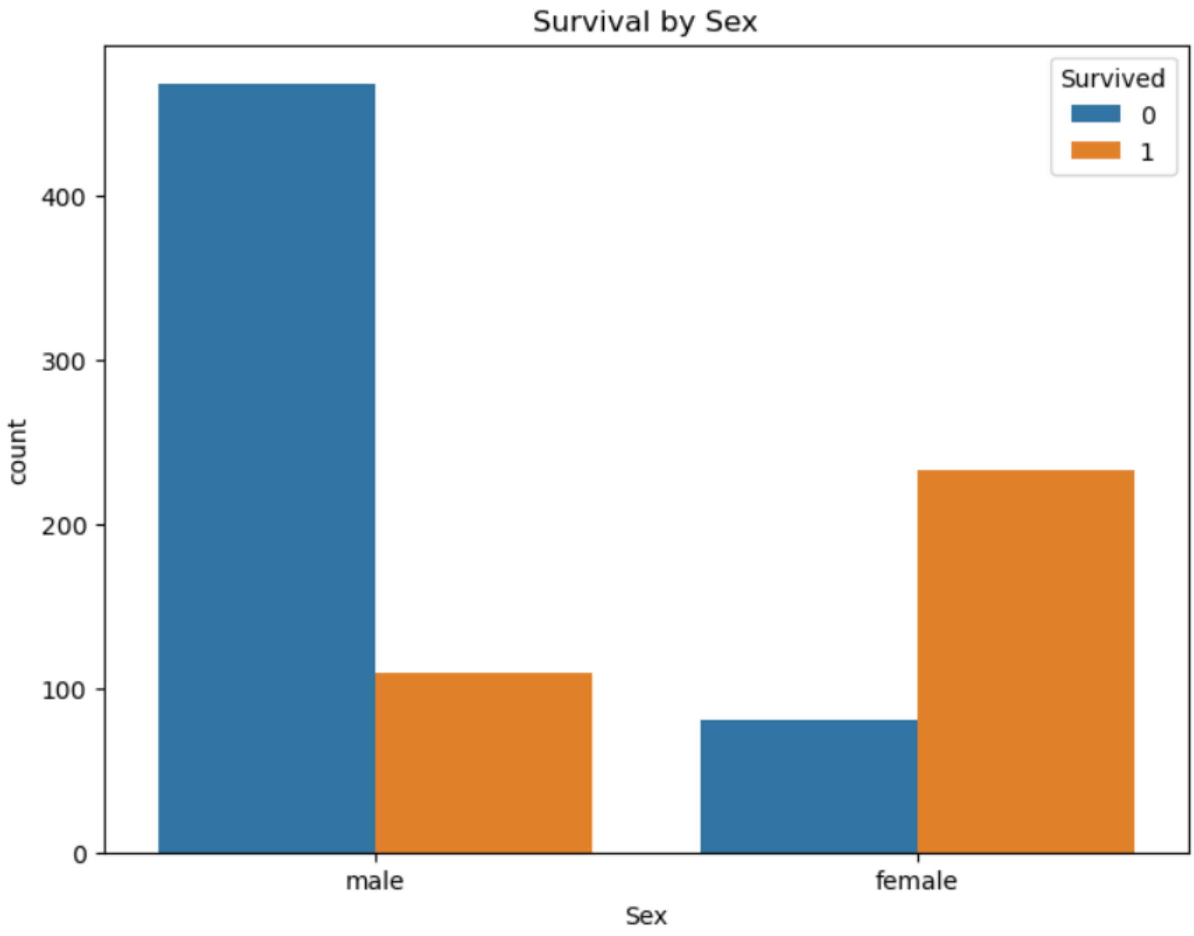
- **Histograms and Boxplots:** Used to visualize the distribution of numerical variables (Age, Fare) and identify potential outliers.
- **Count Plots:** Used to visualize the frequency of categories in categorical variables (Survived, Sex, Pclass, Embarked) and their relationship with survival.
- **Scatterplots:** Used to explore the relationship between two numerical variables (Age vs. Fare).

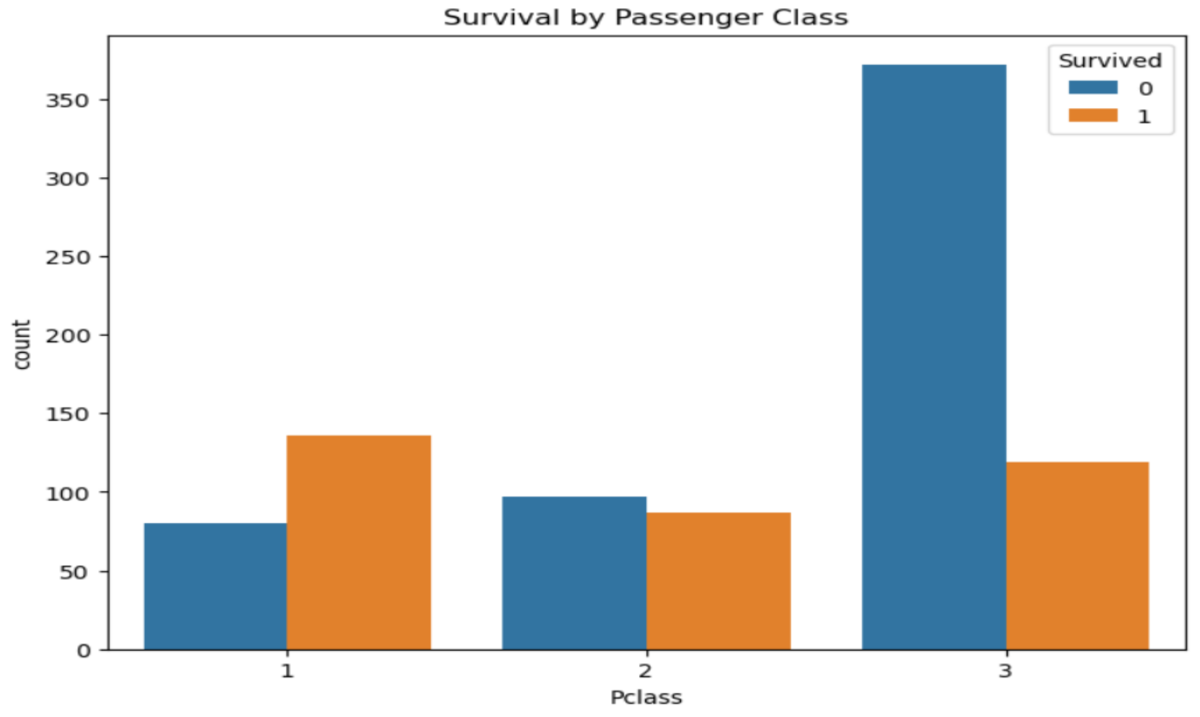
- **Crosstabs and Bar Plots:** Used to visualize the relationship between categorical variables (Sex vs. Survived, Pclass vs. Survived).
- **Pairplot:** Used to visualize pairwise relationships between multiple numerical variables.
- **Heatmap:** Used to display the correlation matrix and visualize correlations between numerical variables.

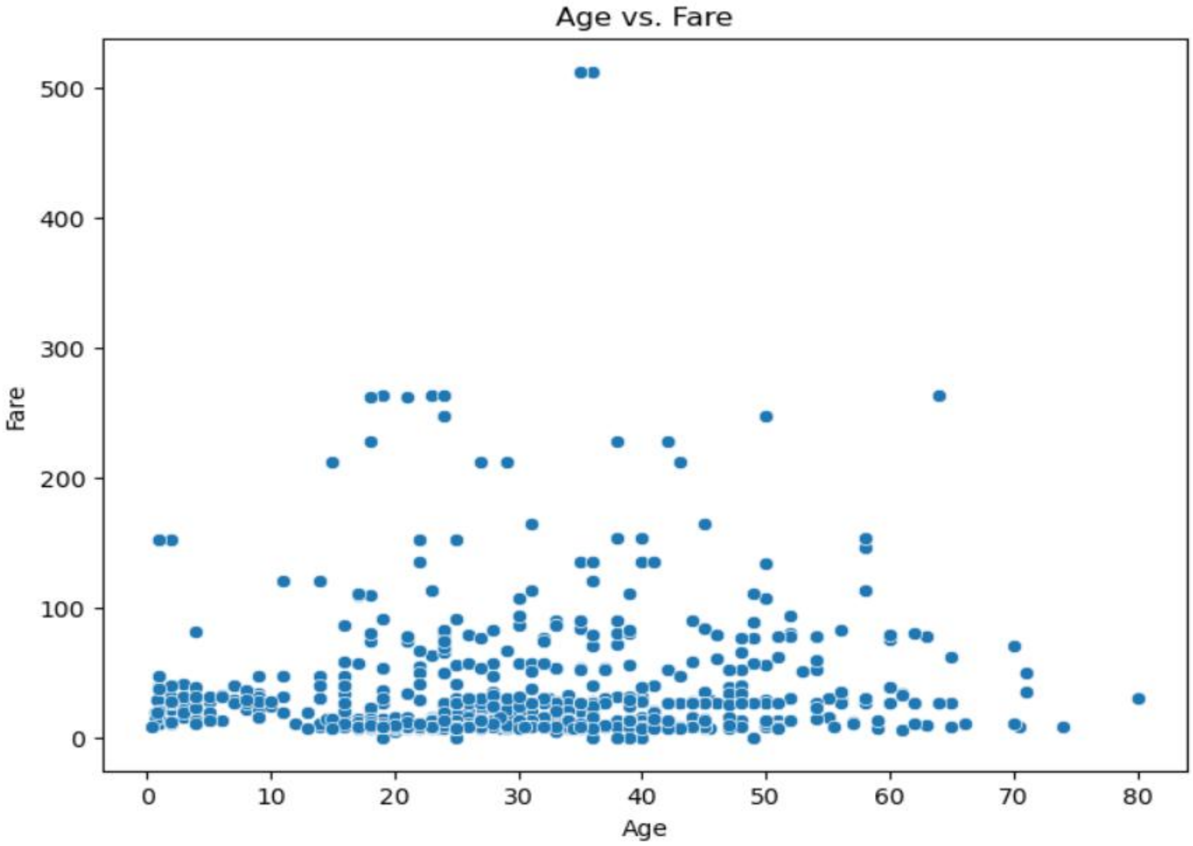


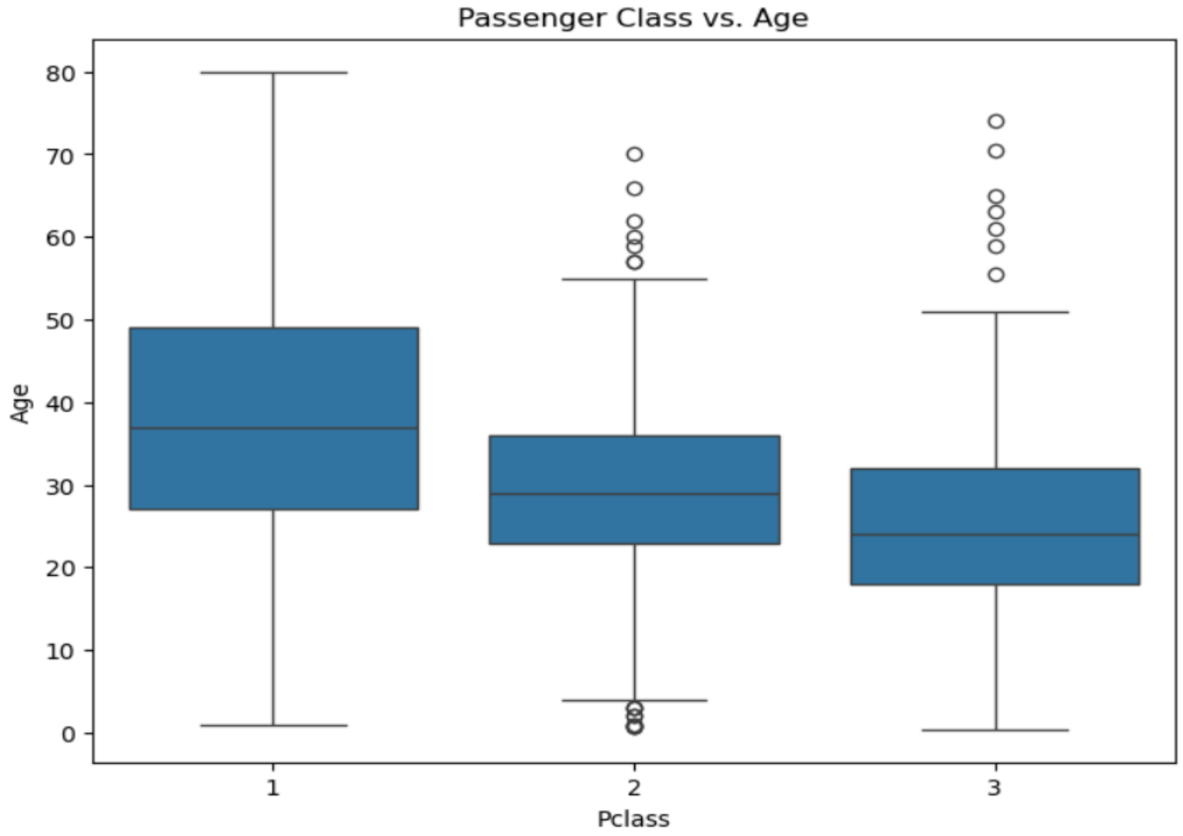

Fare Distribution
Count
300
250
200
150
100
50
0
0
100
200
300
400
500
Fare
Fare Boxplot
Fare
500
400
300
200
100
0

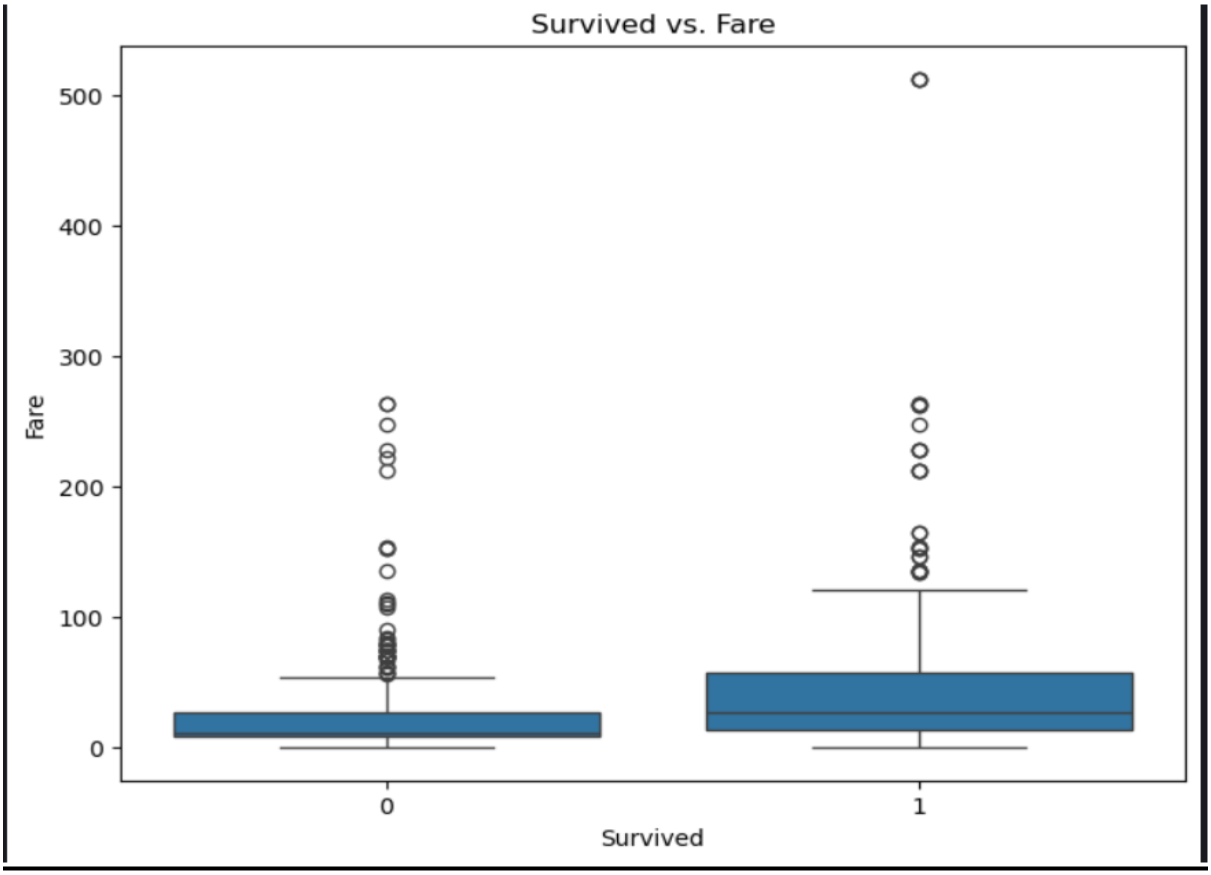

Survival Count
count
500
400
300
200
100
0
0
1
Survived

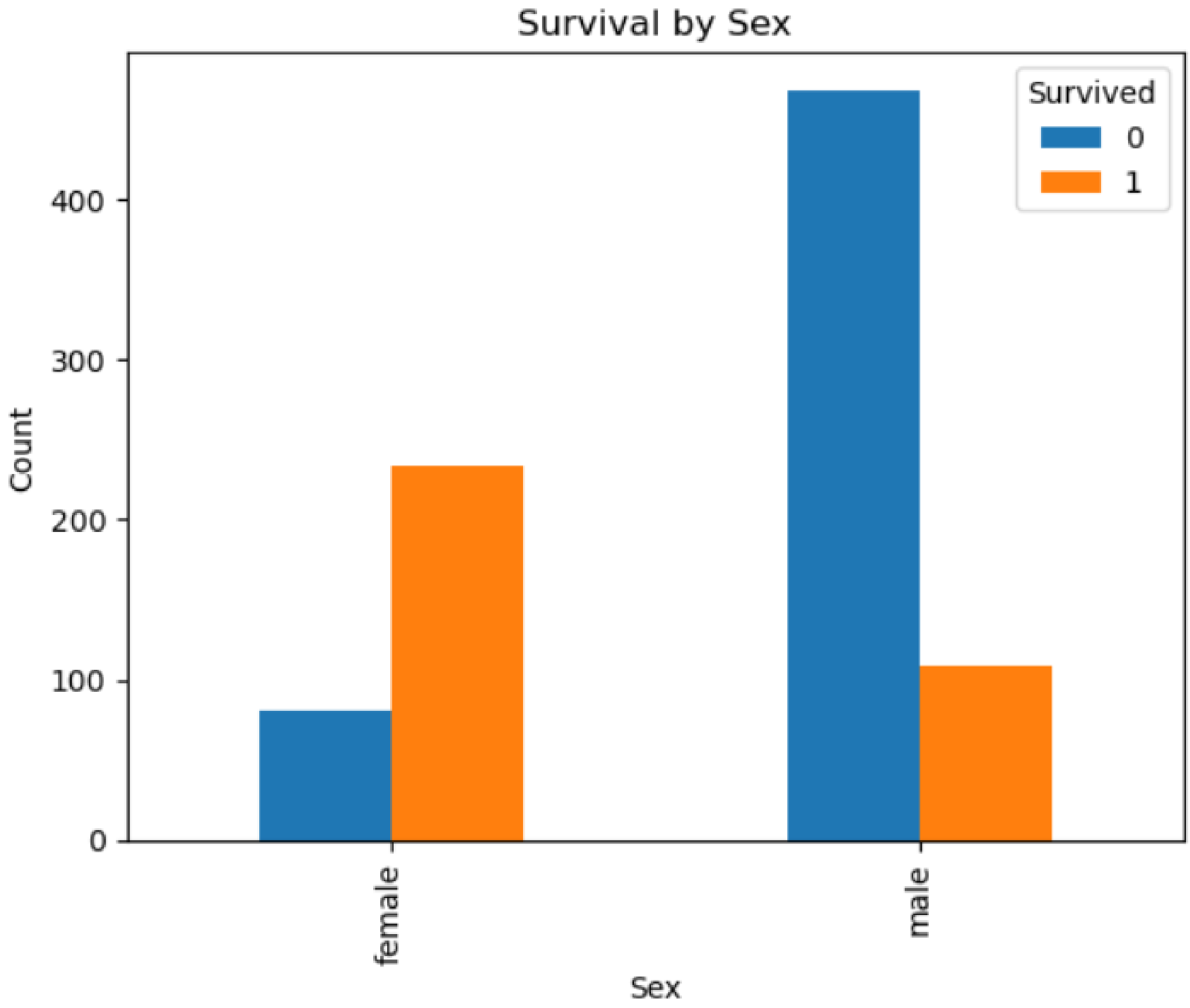

Survival by Sex
Survived
0
1
400
300
200
100
0
count
male
female
Sex


Survival by Passenger Class
Survived
0
1
350
300
250
200
150
100
50
0
count
1
2
3
Pclass


Age vs. Fare
500
400
300
200
100
0
Fare
0
10
20
30
40
50
60
70
80
Age


Passenger Class vs. Age
80
70
60
50
40
30
20
10
0
Age
1
2
3
Pclass


Survived vs. Fare
500
400
300
200
100
0
Fare
0
1
Survived



Summary: The EDA reveals that survival on the Titanic was influenced by several factors, including sex, passenger class, and potentially family size and port of embarkation. Females and passengers in higher classes had a higher likelihood of survival. Age and fare did not appear to be strong predictors of survival.