Exploratory Data Analysis Report: Titanic Dataset

Author : Suhani Pancholi

Date : April 14, 2025

Objective :

This exploratory data analysis aims to uncover key patterns and insights in the Titanic dataset. The goal is to identify factors that influenced passenger survival using statistical methods and data visualization techniques.

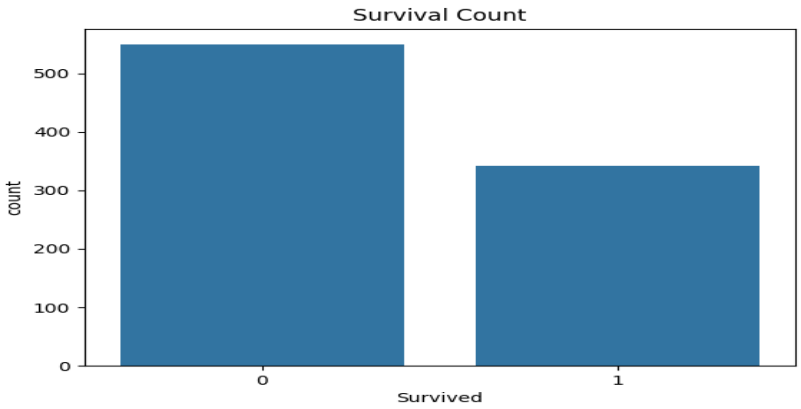
1. Dataset Overview

- **Number of records: 891**
- **Number of columns: 12**
- **Features: PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked**
- **Missing Values: Found in Age, Embarked, and Cabin**
- **Handling: Filled Age with median, Embarked with mode, Dropped Cabin**

2. Observations for Each Visual

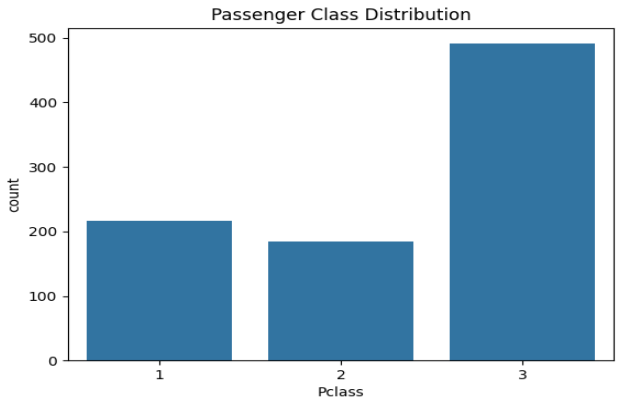
(i) Univariate Analysis

- **Survival Count**
  → Fewer people survived (1) than did not (0).



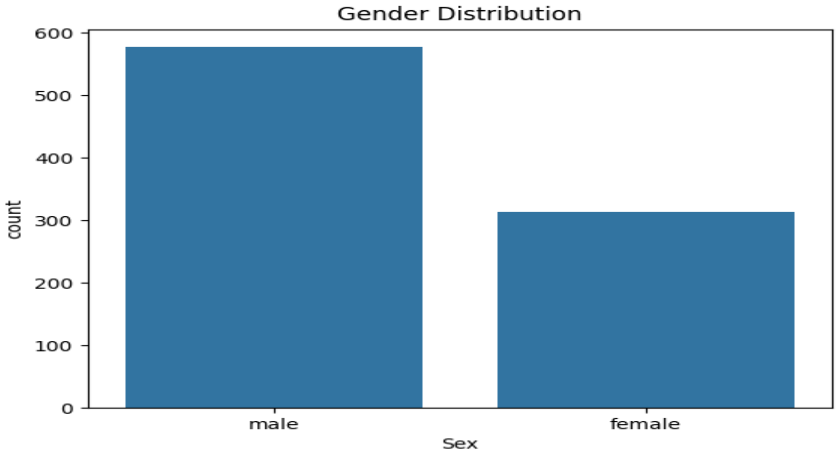
- **Pclass Distribution**

  → Most passengers were in 3rd class.



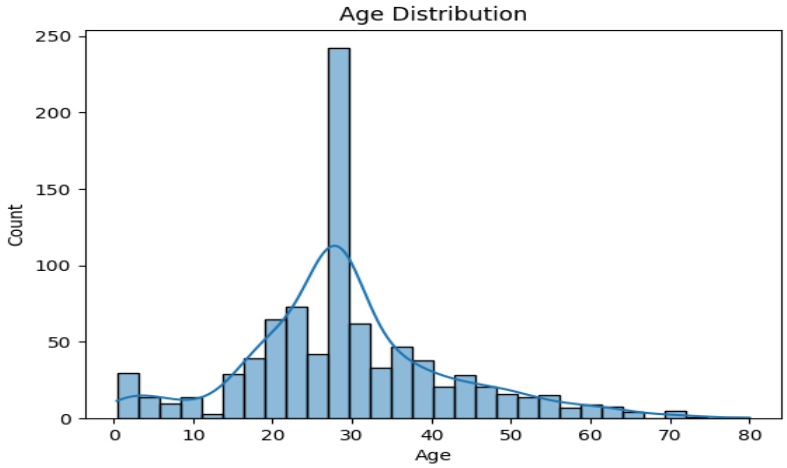
- **Gender Distribution**

  → More males than females were onboard.



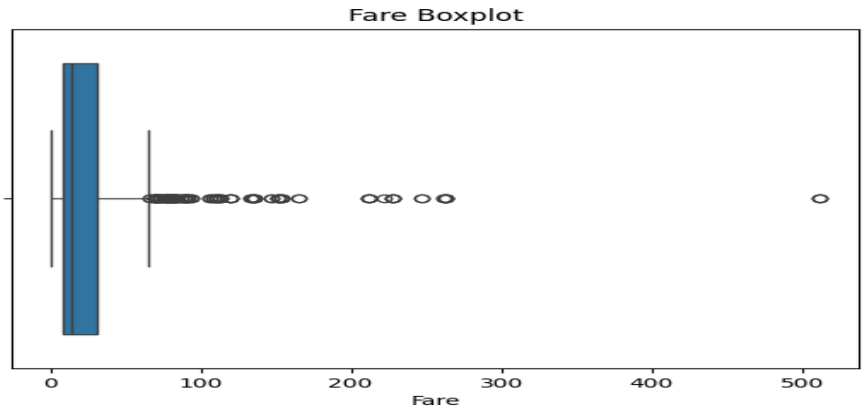
- **Age Distribution**

  → Age is right-skewed; most passengers were between 20 and 40 years old.



- **Fare Boxplot**

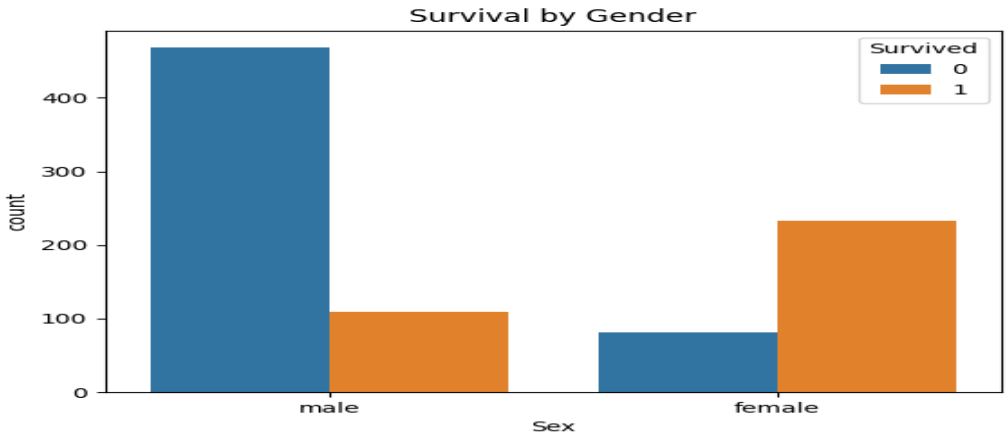
  → Most fares were under 100, but there are extreme outliers with fares above 500.



(ii) Bivariate Analysis

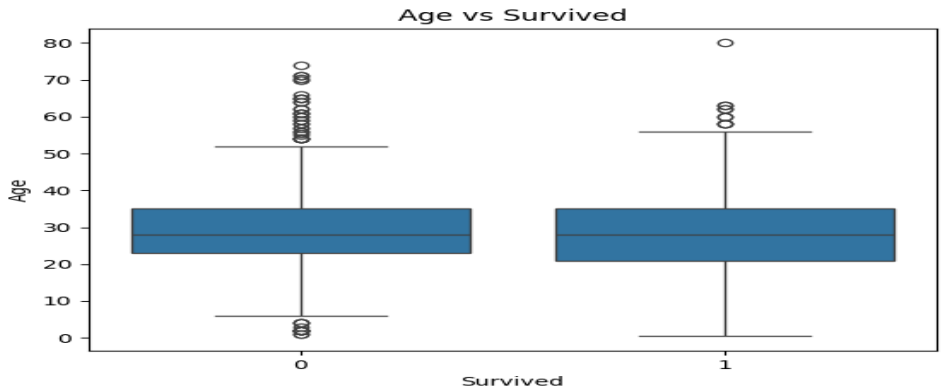
- **Survival by Gender**

  → Higher survival rate among females compared to males.



- **Age vs Survival (Boxplot)**

  → Median age of survivors was slightly lower. Young children had better survival chances.



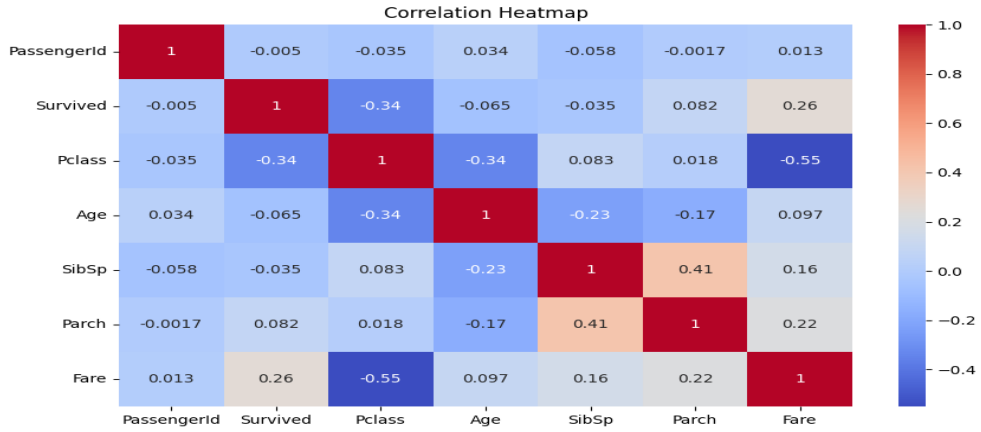
- **Survival by Passenger Class**

  → 1st class passengers had the highest survival rate, while 3rd class had the lowest.

- **Heatmap of Correlations**

  → Strongest positive correlation: Fare and Pclass (inverse).

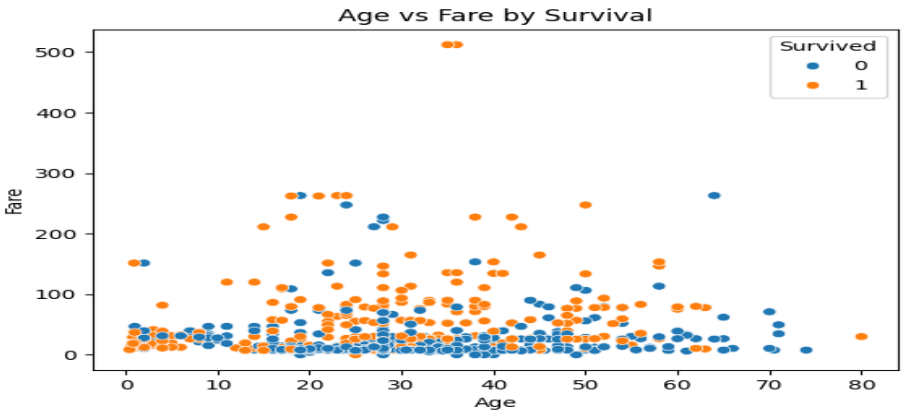
  → Sex, Pclass, and Fare show noticeable correlation with survival.



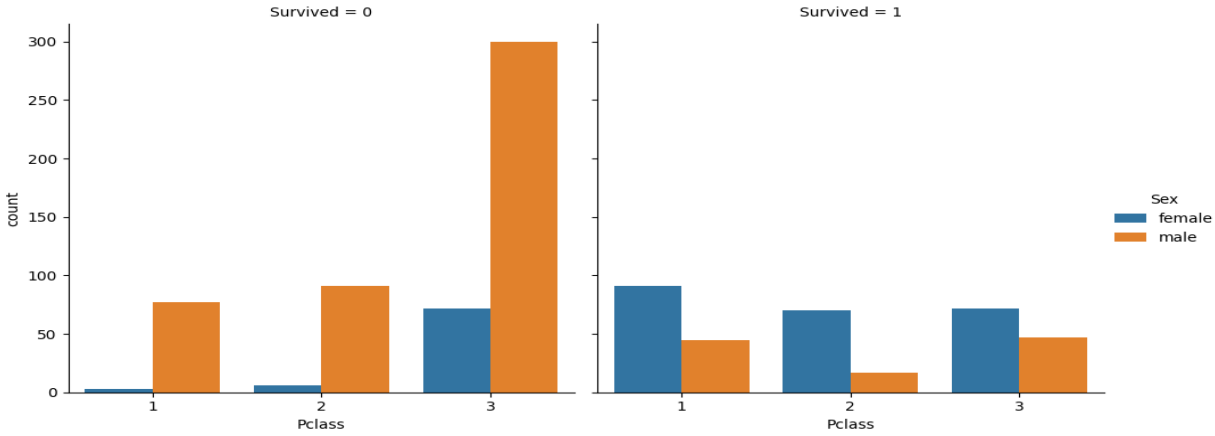
(iii) Multivariate & Scatterplot

- **Age vs Fare by Survival (Scatterplot)**

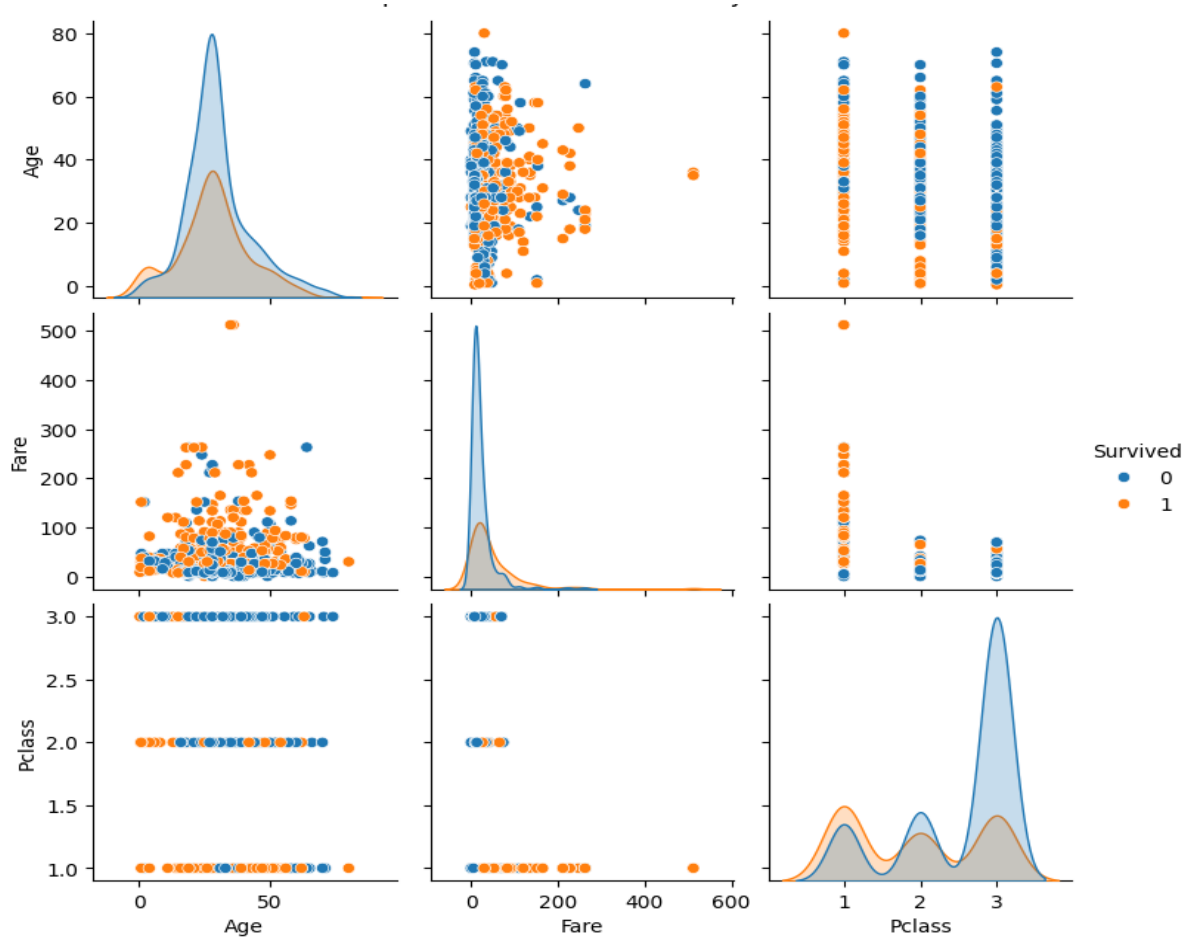
  → Survivors tended to pay higher fares and were slightly younger overall.



- **Survival by Class and Gender (Catplot)**
  → Women in 1st and 2nd class had the highest survival rate.
  → Men in 3rd class had the lowest.



- **Pairplot**
  → Showed positive relationship between Fare and Pclass.
  → Clearly separates some survival trends when considering multiple features together.



3. Summary of Findings

- **Sex is a strong predictor: Females had much higher survival rates.**
- **Pclass is crucial: 1st class passengers were most likely to survive, especially women.**
- **Fare shows correlation with survival — higher fare, higher survival likelihood.**
- **Age has a moderate effect — young children had better chances.**
- **The Cabin column was dropped due to excessive missing values.**
- **Embarked had minimal effect but was cleaned using mode imputation.**
- **Data shows a clear social-class and gender-based bias in survival outcomes.**

4. Conclusion

The EDA indicates that survival on the Titanic was heavily influenced by gender, socio-economic status (Pclass), and fare amount. Females and higher-class passengers exhibited significantly higher survival rates, emphasizing strong feature importance for Sex, Pclass, and Fare in predictive modeling.