Dataset Analysis Report

Date: 6th January 2025

Introduction

The Titanic dataset provides detailed information about passengers aboard the RMS Titanic, which tragically sank in 1912. This report focuses on exploratory data analysis (EDA) and visualization to uncover patterns and relationships within the data, highlighting factors influencing passenger survival. The insights gained contribute to developing a compelling narrative based on the dataset's features.

Part 1: Understanding the Dataset

Dataset Overview

The dataset contains 891 observations and 12 variables, including demographic details, socio-economic status, and survival outcomes.

Data Dictionary

Column Name	Data Type	Description
Passenger Id	Integer	Unique identifier for each passenger
Survived	Factor	Survival status (0 = Did not survive, 1 = Survived)
Pclass	Factor	Passenger class (1 = Upper, 2 = Middle, 3 = Lower)
Name	Character	Full name of the passenger
Sex	Character	Gender of the passenger
Age	Numeric	Age of the passenger
Sib Sp	Integer	Number of siblings/spouses aboard
Parch	Integer	Number of parents/children aboard
Ticket	Character	Ticket number
Fare	Numeric	Fare paid by the passenger
Cabin	Character	Cabin number
Embarked	Character	Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

Part 2: Exploratory Data Analysis (EDA)

Descriptive Statistics and Data Cleaning

- The dataset has missing values in the Age column (177 entries). These were imputed using the median value (28 years) to minimize bias.
- Variables Survived and Pclass were converted to factors for better analysis.
- Outliers in Age were identified but retained to preserve dataset variability.

Univariate Analysis

- The majority of passengers were male (65%) and in the third class (55%).
- Survival rate: 38% survived, with more females surviving (74%) compared to males (19%).

Bivariate Analysis

- Survival rates differed significantly by class: 62% of first-class passengers survived compared to 24% in third class.
- Survival was higher among females across all classes.

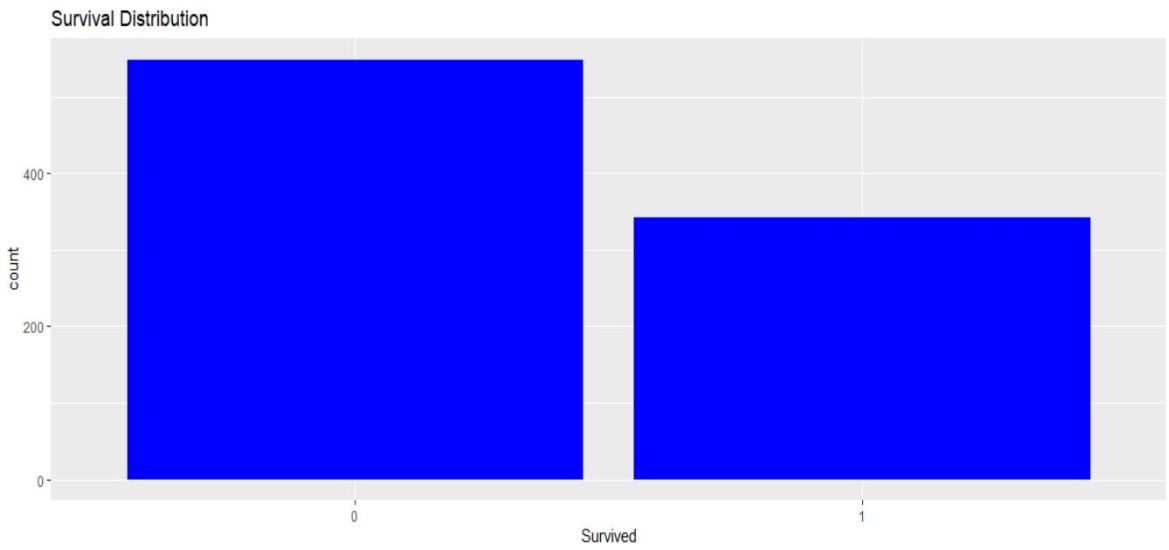
Multivariate Analysis

- Combining Pclass and Sex, it was observed that first-class females had the highest survival rate, while third-class males had the lowest.

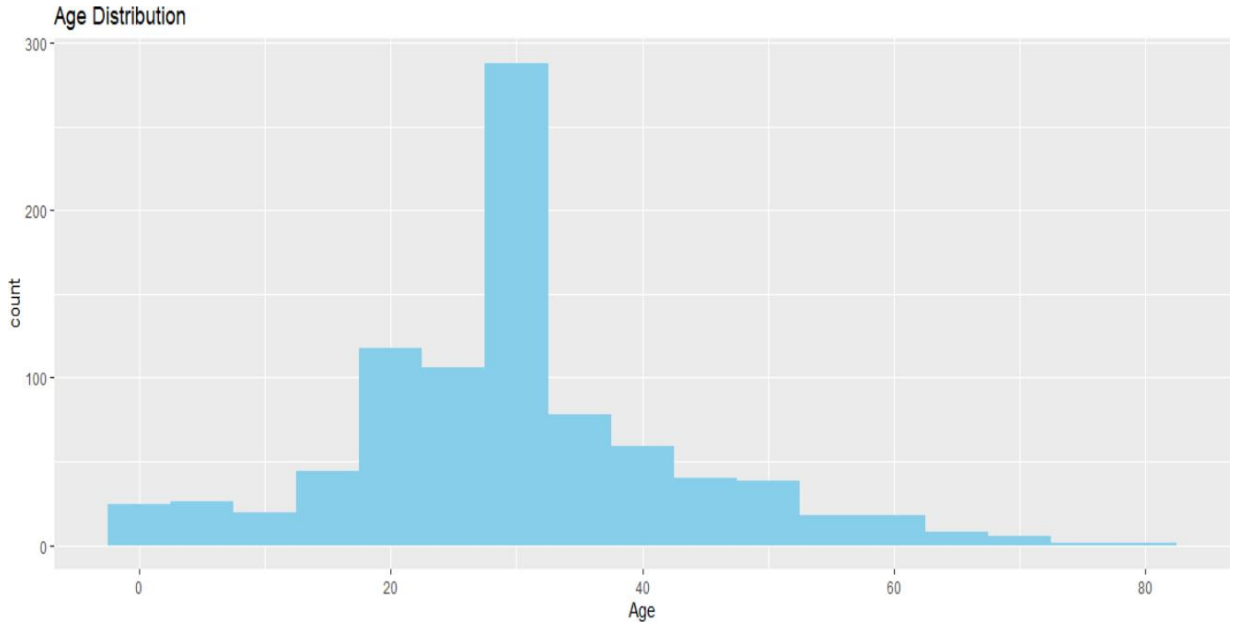
Part 3: Visualization

Key Visualizations

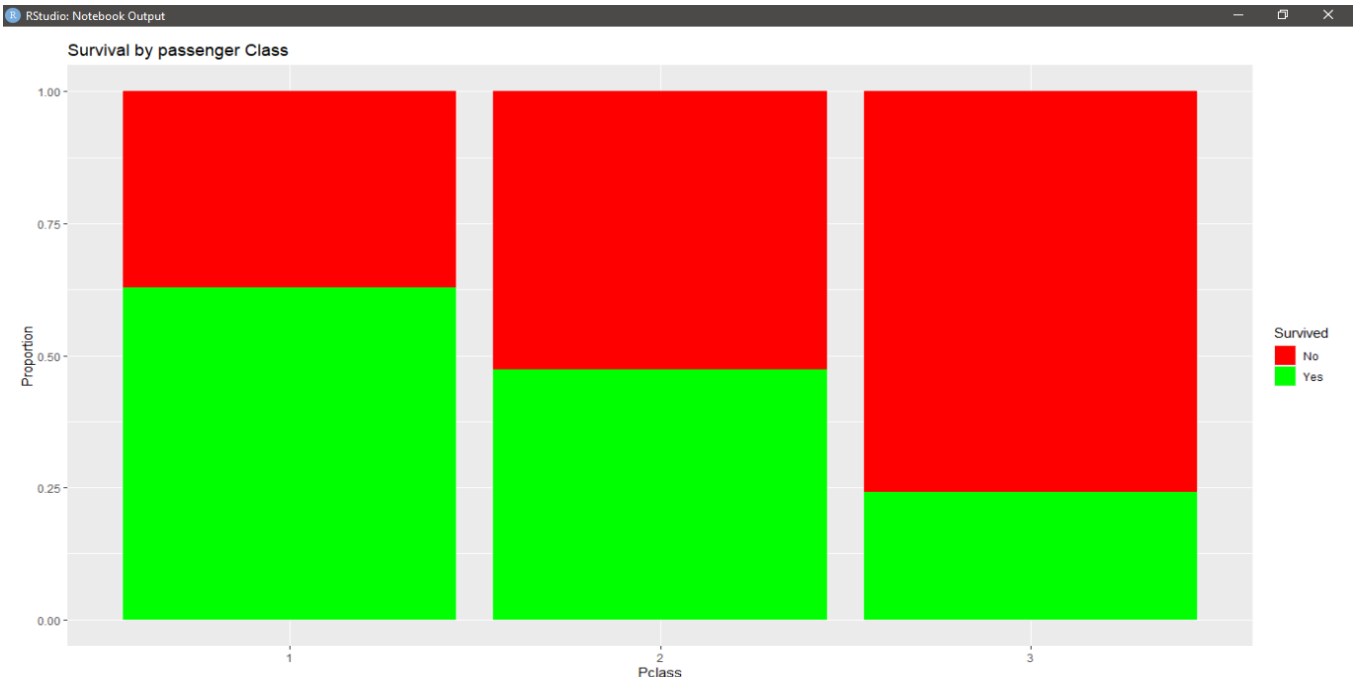
A bar plot showed that first-class passengers had a higher survival rate.



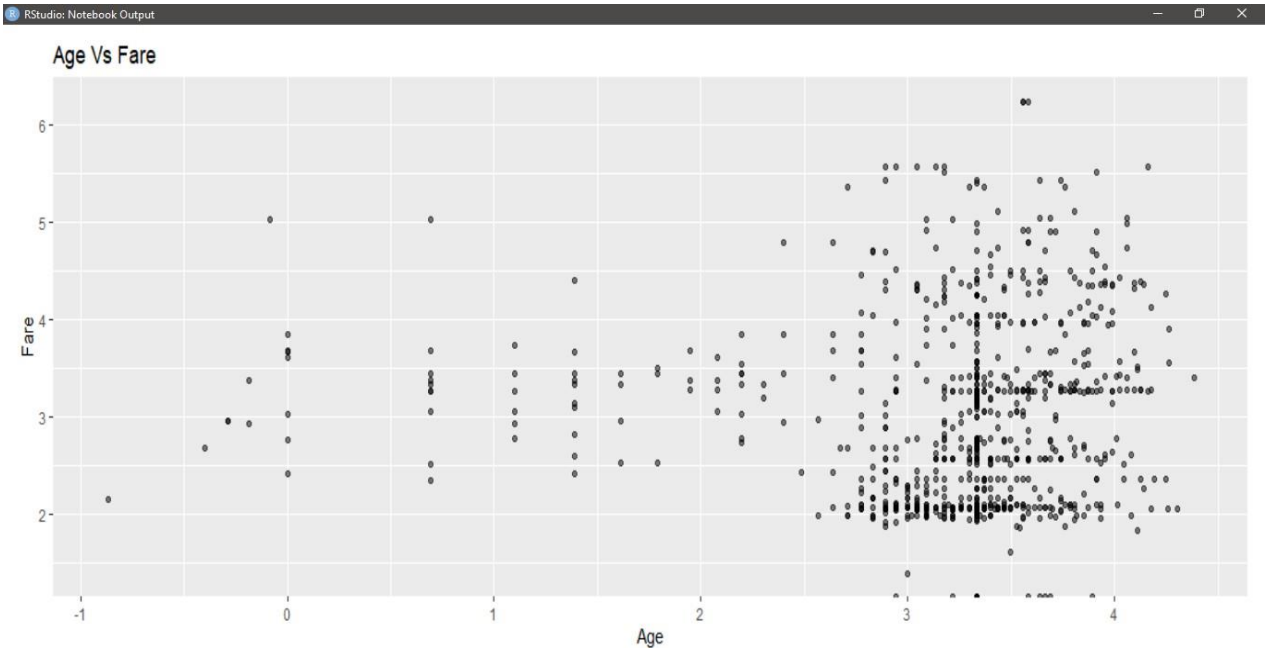
From the bar chart above the number of people who perished in the titanic tragedy was more than the number of the passengers who were on board.



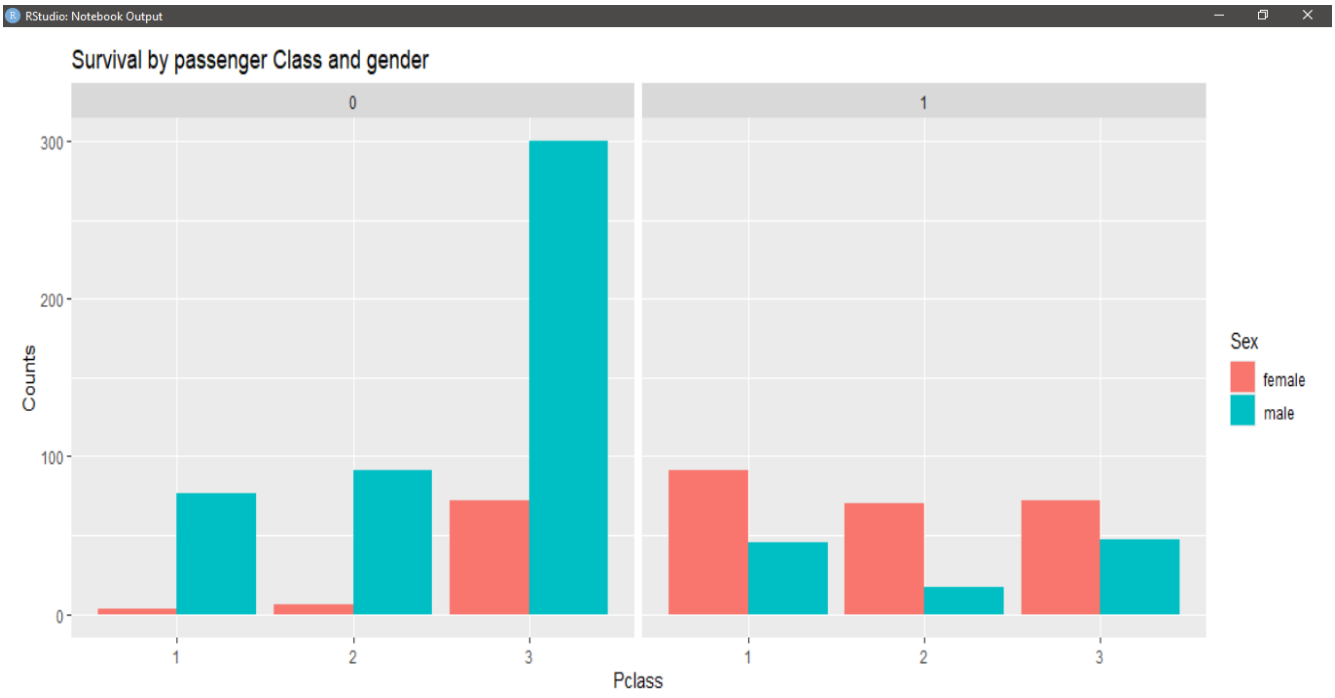
A histogram highlighted that children had better survival rates.



The above bar chart The proportion of passengers who survived during the accident in the first class cabin were more than those who did not survive, Many of the passengers who perished in the tragedy were from the third class which had the highest proportion of fatality.



Survival by Gender and Class: A faceted bar plot indicated that female survival was significantly higher across all classes.



The passenger were travelling in different classes that is first class, second class and Third class. From the graph the number of passengers who were travelling in third class died where most of them were males, the number of Passenger who died travelling in first class cabin were fewer with the female passengers experiencing close to zero deaths. In general in all the three classes' cabin the number of male passengers who died were more than that of female passengers. Many of the female passengers who were travelling in first class cabin survived the tragedy than the males while the number of females who survived the tragedy in second and third class cabin were equal. In general the number of females who survived were more than that of male passengers with male passengers in second class cabin experiencing a minimum number of survival.

Insights from Visualizations

- Passengers in lower socio-economic classes (third class) were disproportionately affected.

- Age and gender significantly influenced survival; women and children had priority access to lifeboats.

Part 4: Storytelling with Data

The analysis reveals stark differences in survival rates based on socio-economic status, gender, and age. Women and children, especially in first-class cabins, had the highest survival chances due to evacuation priorities. Conversely, third-class passengers, particularly males, faced the highest fatalities, potentially due to restricted access to lifeboats. This tragic event underscores the influence of socio-economic disparities on survival. It also highlights the importance of equitable safety measures to prevent future biases in emergency situations.

Conclusion and Recommendations

The Titanic dataset analysis demonstrates how survival was influenced by gender, age, and class. Future policies should ensure equal access to safety for all individuals, regardless of socio-economic status. These insights, supported by visualizations, provide a comprehensive understanding of the factors affecting survival.