Project Tittle – Titanic Survival Analysis

Group Information:

1. Eshika Biswas -22-46243-1

2.Md. Mahmudul Hasan -22-46256-1

3.Md. Shafin Ahamed -22-46274-1

4.Maeed Ahammed -22-46280-1

1. Introduction

The Titanic dataset is a well-known dataset that contains demographic and travel information for passengers aboard the RMS Titanic. It includes variables such as passenger class (Pclass), sex, age, family relationships (SibSp, Parch), fare paid (Fare), port of embarkation (Embarked) and survival status (Survived).

The primary questions explored in this analysis were:

- What passenger characteristics were most associated with survival?
- How did demographic variables such as age and gender influence survival chances?
- Were there notable patterns in survival based on ticket class, fare paid or family size?

2. Data Cleaning Process

The dataset was loaded directly from an online repository using:

url <- "https://raw.githubusercontent.com/Shafin06/Data-Science/61570c317223692a79b9186057c964b6dfbef40e/titanic.csv"

titanic <- read.csv(url, stringsAsFactors = FALSE)

Cleaning steps included:

1. **Removing Duplicates:**
   - Used distinct() to ensure each passenger record was unique.
   - Rationale: Duplicate entries can bias survival rates and distort descriptive statistics.
2. **Handling Missing Values:**
   - **Age:** Replaced missing ages with the median value (28 years).
     Rationale: Median is less sensitive to extreme values than mean.

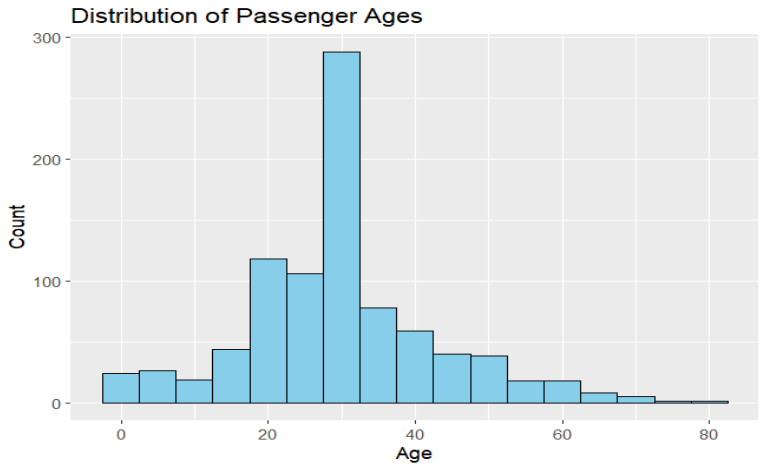
- **Embarked:** Replaced missing values with the most common port (S for Southampton).
  Rationale: Ensures no passengers are excluded from categorical analysis.

3. **Feature Engineering:**
   - Created a FamilySize variable (SibSp + Parch + 1).
     Rationale: Family size can influence survival patterns.
4. **Variable Selection:**
   - Selected only relevant variables for analysis:
     Survived, Pclass, Sex, Age, SibSp, Parch, Fare, Embarked, FamilySize.
     Rationale: Simplifies analysis and focuses on key predictors.
5. **Handling Outliers:**
   - For continuous variables (Age and Fare) extreme values were detected using the Interquartile Range (IQR) method.
   - Any values below Q1 − 1.5 × IQR or above Q3 + 1.5 × IQR were considered outliers.
   - Instead of removing these passengers, their values were capped at the nearest valid boundary to reduce skewness.
   - Rationale: Outliers can distort averages plots, and survival pattern analysis. Capping preserved overall data structure while limiting the impact of extreme cases.

3. Key Findings & Visualizations

3.1 Age Distribution of Passengers

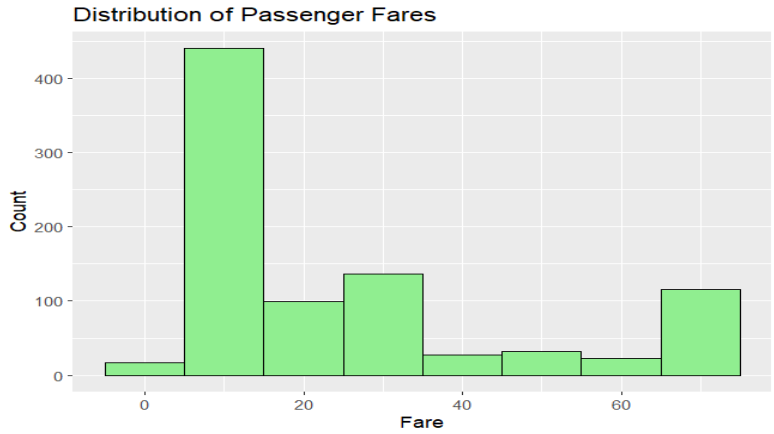
Figure 1: The histogram shows that most passengers were between 20 and 40 years old with a smaller number of children and elderly passengers.



Insight: Younger passengers (especially children) had a higher survival rate which aligns with the “women and children first” evacuation protocol.

3.2 Distribution of Passenger Fares

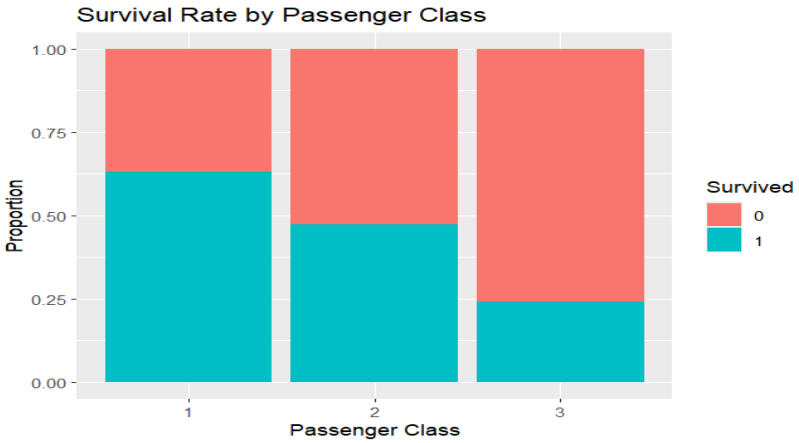
Figure 6: Most passengers paid fares below 100 while a few paid extremely high amounts.



Insight: The fare distribution is right-skewed. Higher fares were linked to first-class passengers who had better survival chances but fare alone did not guarantee survival.

3.3 Survival Rate by Passenger Class

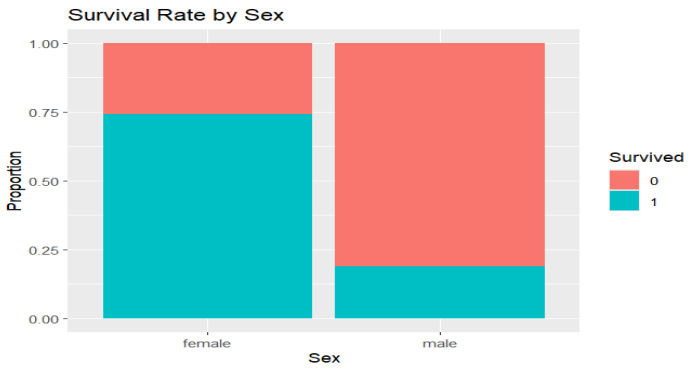
Figure 2: Proportion of survivors by ticket class. Higher classes had higher survival rates.



Insight: First-class passengers were more likely to survive than those in second or third class. Access to lifeboats and cabin location may have contributed.

3.4 Survival Rate by Gender

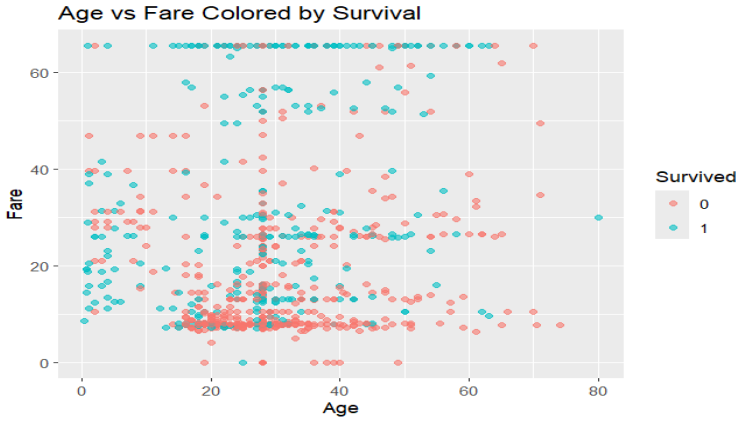
Figure 3: Women had a much higher survival rate than men.



Insight: Female passengers had about a 74% survival rate while male passengers had around 19%. This strongly supports the "women and children first" policy.

3.5 Family Size and Survival

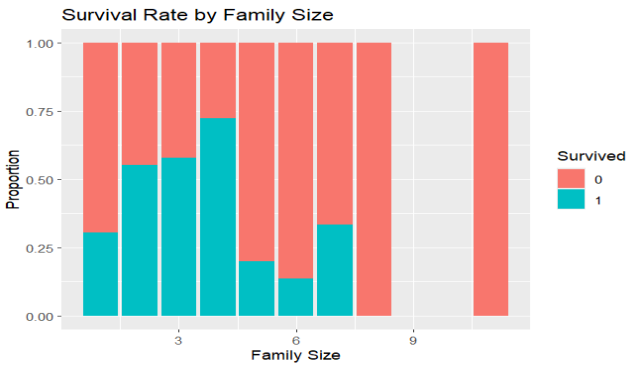
Figure 4: Passengers with small families (2–4 members) had higher survival rates than those traveling alone or with very large families.



Insight: Being alone or in a very large group reduced survival chances. Smaller groups may have been better able to coordinate escape.

3.6 Age vs. Fare

Figure 5: Scatter plot of Age vs. Fare colored by survival.



Insight: Many high-fare passengers survived particularly younger ones. However high fare alone did not guarantee survival.

4. Conclusion

From the analysis several clear patterns emerged:

- **Gender:** Women had a significantly higher survival rate.
- **Passenger Class:** Higher class passengers were more likely to survive.
- **Age:** Children had better survival chances.
- **Family Size:** Small families fared better than solo travellers or very large families.
- **Fare:** Higher fares were associated with higher survival but not universally.

Additionally, by handling outliers in Age and Fare the analysis became more robust. This ensured that extreme ticket prices or rare age values did not disproportionately influence the results. However it is important to note that capping outliers may slightly mask the true variability of the data.

Limitations:

- The dataset does not include all possible survival factors such as lifeboat location or crew assistance.
- Missing age values were imputed which could slightly distort age related findings.
- The dataset is historical and may not reflect modern maritime evacuation procedures.