Analyzing Titanic Survival Rates

Machine Learning Fall 2024
Hiva Abolhadizadeh

Abstract

This report analyzes the Titanic dataset to identify factors influencing survival. Key findings show that women (82.5%) had a significantly higher survival rate than men (12.9%). First-class passengers (57.3%) were more likely to survive than those in lower classes, highlighting the role of socio-economic status. Smaller family sizes (1-3 members) had better survival outcomes, while children aged 0-12 had the highest survival rate (55.3%). These insights provide a clearer understanding of survival patterns aboard the Titanic.

Key Word: Titanic, Survival Rate, Gender, Passenger Class, Family Size, Age Group, Data Analysis

Introduction

The sinking of the RMS Titanic on April 15, 1912, stands among the most tragic maritime disasters in history, with over 1,500 human lives lost. Being considered an engineering and luxury marvel of its time, passengers came from all sections of society. The factors that influenced the survival rate during this catastrophe provide lots of insight into the behavioral patterns of humans in crises.

This report analyzes the Titanic dataset, which includes information about passengers' demographics, ticket classes, and survival status. We will find important patterns and relationships that may exist within these variables to determine which ones contribute most significantly to a person's survival. Key factors such as gender, passenger class, family size, and age are explored to understand their impact on survival rates.

This report aims to pinpoint the disparities in survival through the use of statistical analysis and visualizations, showcasing different groups of passengers and providing a deeper understanding of socio-economic and demographic influences during the disaster. In the end, this is an analysis that not only remembers lives lost but also brings light to the nuances involved in human decision-making when emergencies take place.

Explanation of the Dataset

The data used here is from, specifically the Titanic competition page from Kaggale. This data gives a very big picture of the passengers on the RMS Titanic, which sank in April 1912 on its maiden voyage. It finds broad applications in data science, especially predictive modeling and data analysis.

The dataset consists of several key features that contribute to understanding the factors influencing survival rates:

- **PassengerId**: A unique identifier assigned to each passenger.
- **Pclass**: The class of travel for the passenger, where 1 = 1st class, 2 = 2nd class, 3 = 3rd class. This feature reflects the socio-economic status of passengers.
- **Name**: Passenger name, may imply additional information to be derived from the title of analysis.
- **Sex**: The gender of the passenger, is crucial for examining survival trends based on gender.
- **Age**: The age of the passenger in years, allowing analysis of how age impacts survival chances.
- **SibSp**: The number of siblings or spouses aboard the Titanic, is useful for understanding family dynamics during the evacuation.
- **Parch**: The number of parents or children aboard the Titanic, similar to *SibSp* in understanding family relationships.
- **Ticket**: The ticket number for the passenger, which may provide context but is not directly useful for survival analysis.
- **Fare**: The fare paid by the passenger, indicating economic status.
- **Cabin**: The cabin number where the passenger stayed, though many entries are missing.
- **Embarked**: The port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton), which may influence survival rates based on location.

The main target variable in this dataset is represented by the **Survived** column, which contains binary values showing whether a passenger survived-1 or did not survive-0.

The data set had missing values recorded for several of its key features: there were 263 missing values in the *Age* column, 1 missing value in *Fare*, 2 missing values in *Embarked*, and a whopping 1,014 missing values in *Cabin*. First, an attempt was made to fill in the missing *Age* values using relationships based on the *SibSp* and *Parch* columns; however, the identification of strong connections proved challenging. Hence, the missing values in *Age* were filled based on the mean age for each gender group. Since the missing data in the *Cabin* column was too extensive to be useful, this column was dropped from the analysis along with *Name* and *Ticket*, which were deemed irrelevant for predicting survival. By cleaning up the missing value and removing the irrelevant features, the dataset was now ready to go. This allowed the key variables, which were most likely to impact the passenger's survival chances, such as gender, passenger class, and age. The final dataset used for analysis consists of 1,306 samples.

The reason the Titanic dataset is so important in data analysis is that this dataset is expected to provide insight into human behavior at times of crisis and how different factors affect survival. It acts as a basic source to practice data analysis and predictive modeling as machine learning techniques. This dataset will be enlightening for socio-economic and demographic influences that play a pivotal role in enhancing our understanding of the complexities surrounding survival in emergencies.

Data Analysis

Key patterns and insights into the survival of the passengers from the Titanic disaster were unveiled through the analysis of this dataset. The variables of gender, passenger class, age, and family size are all important contributors to the varied survival rates among these passengers.

Overall Passenger Characteristics

The age range of the passengers is approximately 30 years, with the youngest being an infant and the oldest being 80 years. Most of the passengers traveled alone, as depicted by the low average number of siblings/spouses and parents/children on board. Equally, most passengers were traveling in second and third class, as evidenced by an average fare suggesting that many paid relatively cheap prices.

Overall Survival Rate

The overall survival rate for all passengers amounts to (37.67%). For this reason, it is a baseline for a more detailed analysis of factors influencing survival. The consecutive sections present the survival rates with respect to gender, class, age, and family size. Figure 1 shows a pie chart presenting the proportion of survivors versus non-survivors.

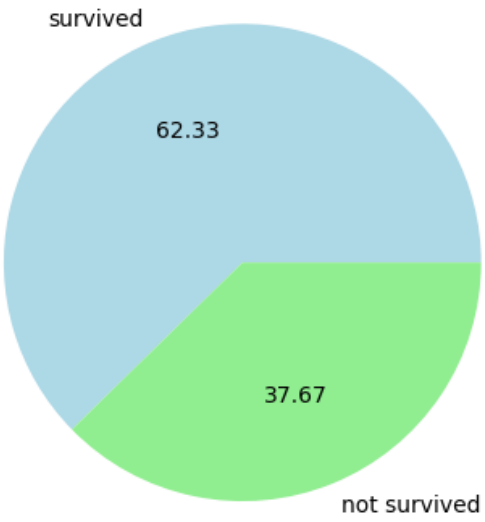

Figure 1: Overall Survival Rate of Passengers

Survival Rate by Gender

Figure 2 shows that gender played an important part in determining the ratio of whether a person survived. The female passengers had a huge advantage in staying alive, with a huge survival rate of 82.54%. This is because the societal norms favored women and children during evacuation.

On the other hand, only 12.95% of male passengers survived due to the low priority in evacuation and also during the chaos.

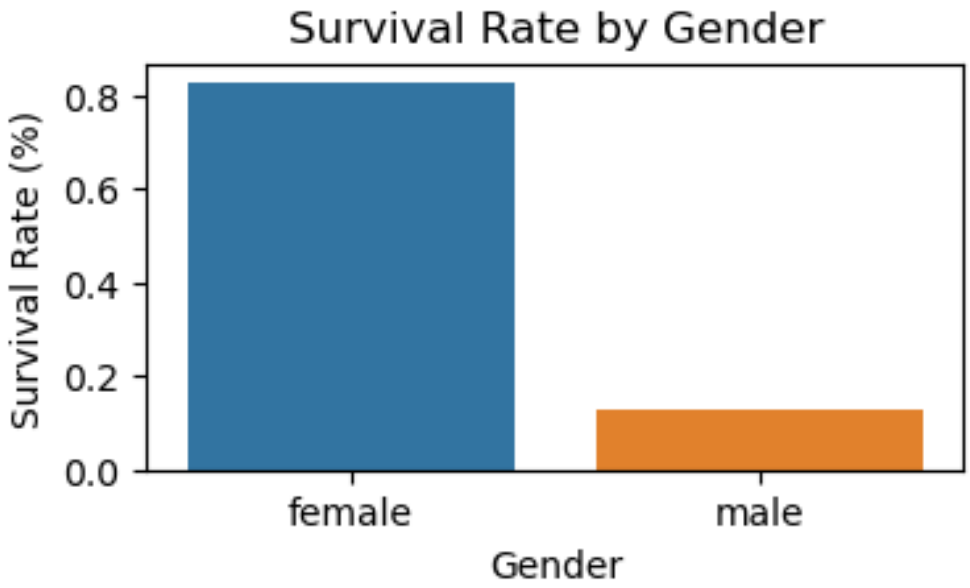

Figure 2: Survival Rate by Gender

Survival Rate by Passenger Class

Passenger class played a huge role in survival, as evident from Figure 3. First-class passengers were the most likely to survive at 57.32%, perhaps because they were more provided with lifeboats and had a higher social standing. Second-class passengers survived at a rate of 42.24%, with the third-class passengers having the lowest survival rate at 26.98%. This disparity suggests that the third-class passengers, located in lower parts of the ship, may have faced greater challenges in accessing the lifeboats, as depicted in Figure 4, which displays the location of each passenger class on the Titanic.

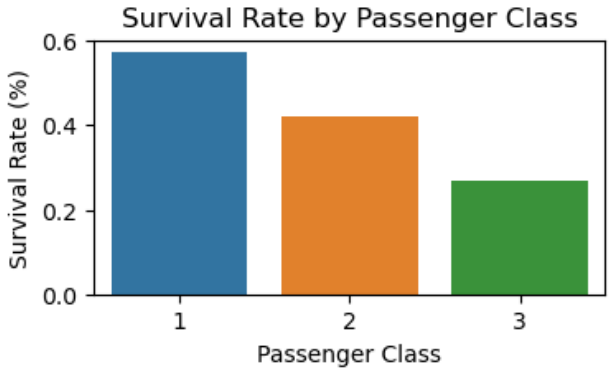

Figure 3: Survival Rate by Passenger Class

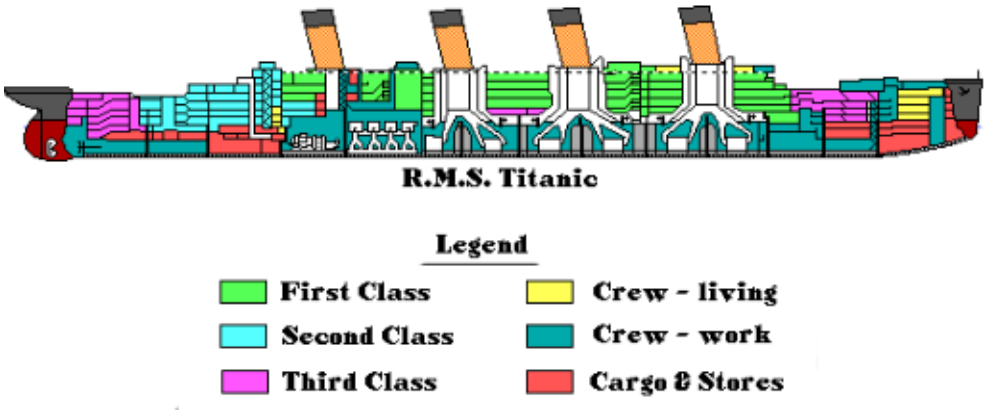

Figure 4: Location of Passenger Classes on the Titanic

Survival Rate by Age Group

As shown in Figure 5, age was another critical factor in survival. Children aged 0-12 had the highest survival rate at 55.32%, reflecting the prioritization of children in the evacuation. In contrast, older adults aged 61 and above had the lowest survival rate at 24.00%. Young adults (ages 13-30) had relatively good survival rates, while middle-aged adults (31-50) experienced a noticeable decline. The vulnerability of both children and elderly passengers is magnified in these findings as it further justifies the relevance of age to survival outcomes during emergencies on the Titanic.

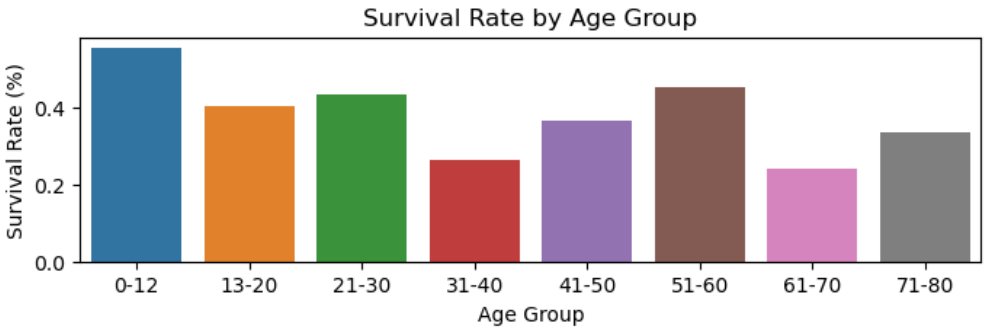

Figure 5: Survival Rate by Age Group

Survival Rate by Family Size

Family size also influenced survival, as illustrated in Figure 6. Passengers traveling alone (Family Size 0) had a survival rate of 29.10%, indicating that being alone did not provide a significant advantage during evacuation. However, passengers with one family member had a survival rate of 53.19%, suggesting that having at least one companion might have improved coordination during the evacuation process.

Medium-sized families (Family Sizes 2 and 3) had even higher survival rates of 55.97% and 72.09%, respectively, likely benefiting from mutual support. In contrast, larger families (Family Sizes 4 to 7) had lower survival rates, with challenges likely arising from the difficulties of keeping

the entire group together during the evacuation. For example, Family Size 4 had a survival rate of 22.73%, and Family Size 5 had a survival rate of only 20.00%.

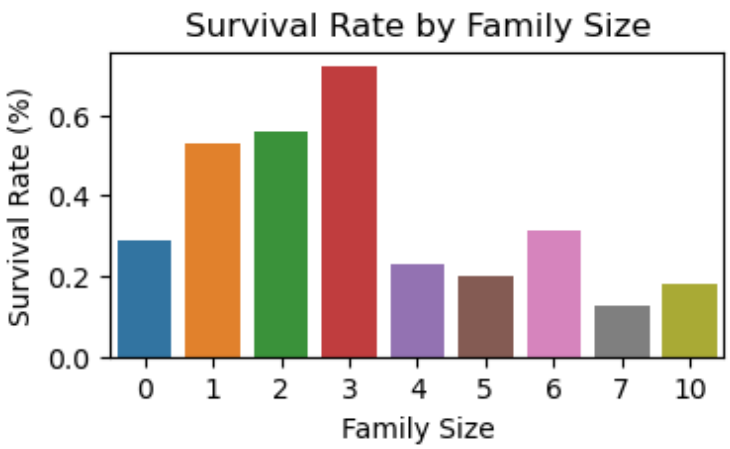

Figure 6: Survival Rate by Family Size

Data Analysis

The following section represents the main visualizations that represent an overview of passenger survival on the Titanic.

Data Distribution in Different Groups

Figure 7 gives the distribution of the data which shows that survivors are mainly concentrated in the age bracket between 20-40 years. On the other hand, non-survivors equally have a concentration in the lower group of ages, although there are noticeably more aged non-survivors. This indicates that though younger adults had a better survival rate, the chances of surviving among older passengers were small, especially for those above 50 years.

Histograms of Age Distribution

The second visualization, presented in Figure 8, reveals that most survivors are concentrated between the ages of 20 and 40, particularly between 20-30 years. Additionally, fewer passengers aged over 60 survived. This highlights that younger passengers had a significantly higher survival rate, especially those in their 20s and 30s.

Box Plot of Age Distribution by Class

Thirdly, the age distribution according to class is shown in more detail in Figure 9.
- First Class (1): Survivors (in green) show a wider spread in terms of age, containing many elderly passengers, whereas the non-survivors shown in blue have a smaller scale.
- Second Class (2): A similar trend is observed, with survivors generally being younger than those in 1st class.
- Third Class (3): The plot indicates that survivors in the 3rd class are predominantly younger, with non-survivors being more concentrated in the younger age range. This suggests that younger passengers in 3rd class had a higher survival rate.

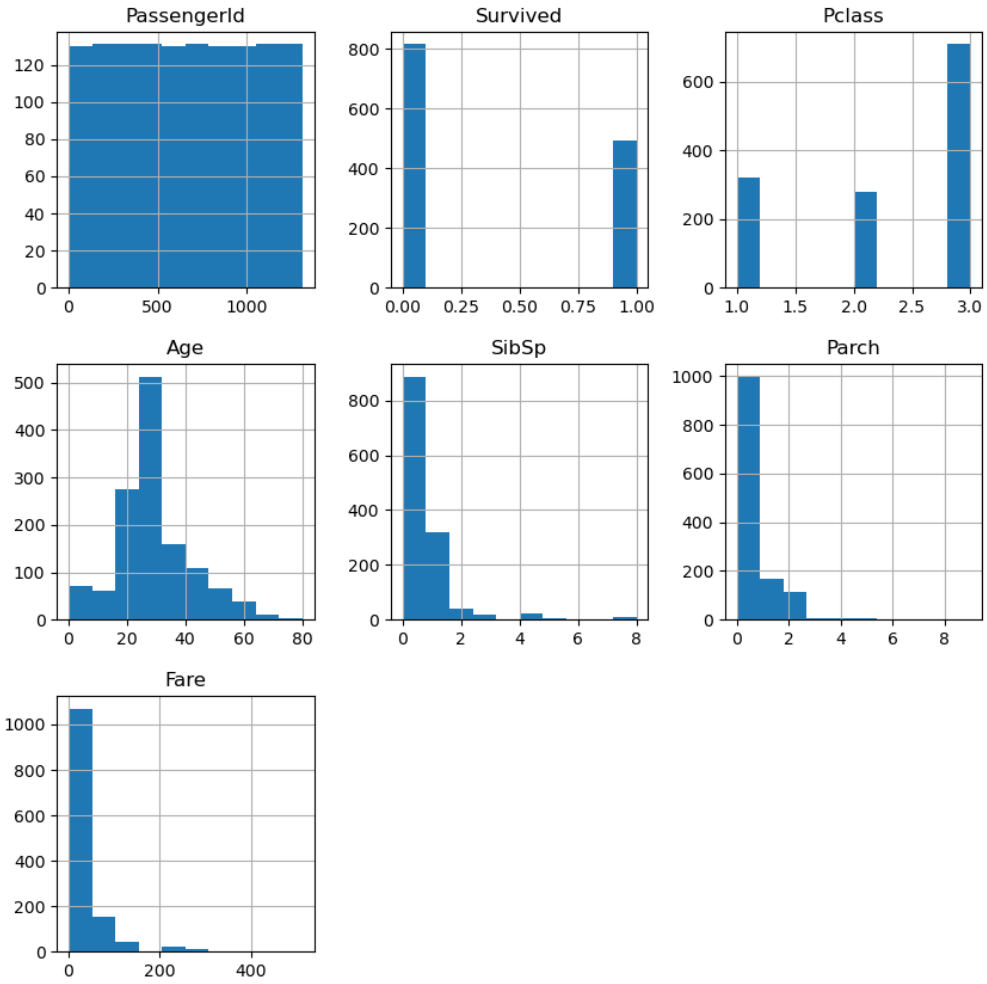

Figure 7: Data Distribution by Age Group

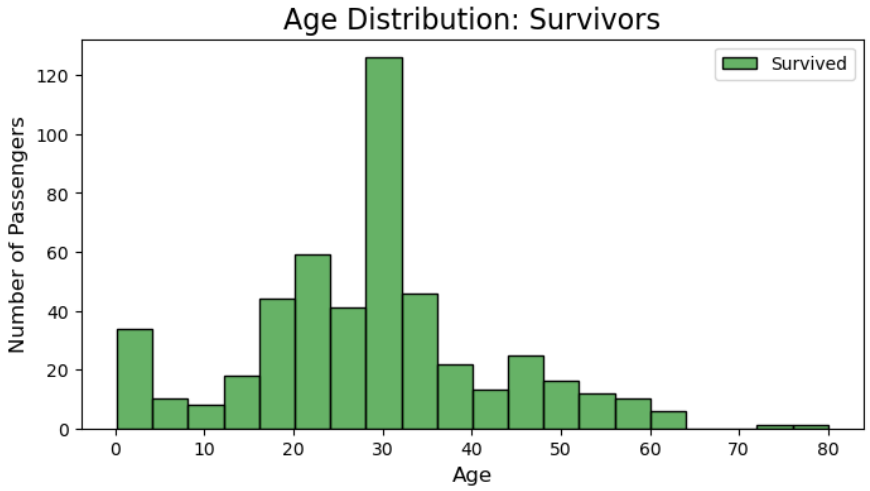

Figure 8: Histograms of Age Distribution

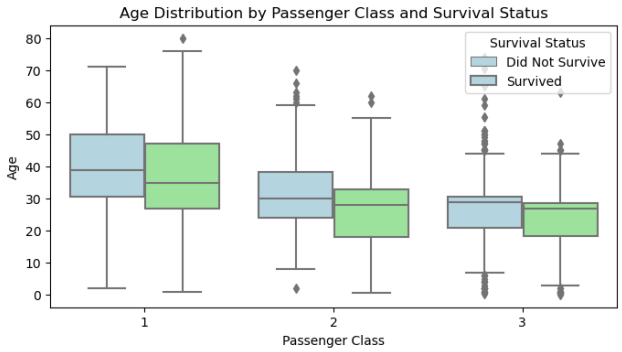

Figure 9: Box Plot of Age Distribution by Class

Understanding the Correlation in Data

As shown in Figure 10, it is clear that:

-Pclass and Survival: A moderate negative correlation of -0.26 indicates that passengers in lower classes (3rd class) were less likely to survive.

- Fare and Survival: A positive correlation of 0.23 infers that the higher the fare paid - probably in a high class - the better the chance of survival.

- Age and Pclass: A correlation of -0.37 suggests that the older the age, the greater the class.

- Other Factors: The analysis also suggests that the SibSp and Parch features are weakly positively correlated with survival; this may mean that passengers who had family members on board were slightly more likely to survive than not.

These correlations confirm that class and fare significantly impacted survival rates, with wealthier passengers more likely to survive the Titanic disaster.

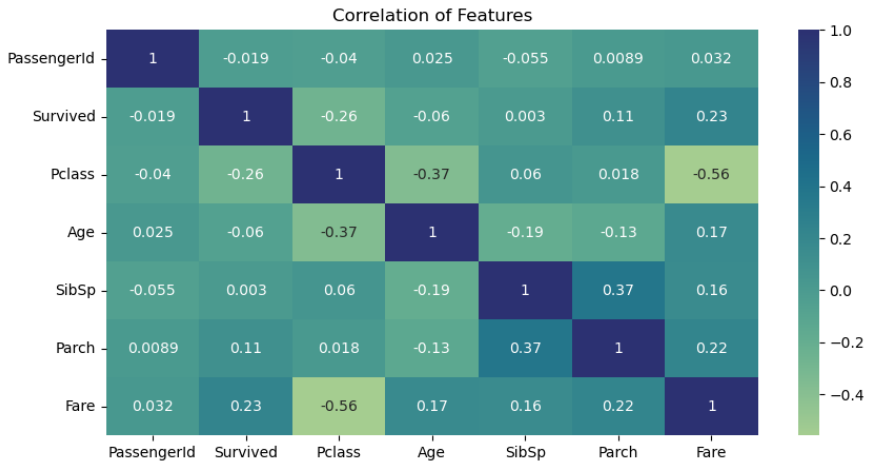

Figure 10: Correlation Heatmap of Features

Conclusion

This critical factor listed below had the most telling impact on the survival rate of the passengers of the Titanic Dataset:

1. Role of Gender: The results showed that gender played the most important role, with 82.54% of the female passengers surviving, compared to just 12.94% among their male counterparts. This difference shows, rather clearly, the efficiency of the "women and children first" policy, usually followed in emergencies, which favored the women's evacuation.

2. Influence of Pclass: The important influence that passenger class had on the outcome is uncovered in the results: the survival rate among first-class passengers was the highest at 57.32%, followed by second-class passengers, while in third-class passengers, this rate had the worst value at only 26.98%. These data suggest that a passenger of higher socio-economic status might have been better placed in terms of access to lifeboats and other evacuation resources.

3. Age Impact: Age also significantly influenced survival, while children under 12 demonstrated the highest survival rate of 55.32%. On the other hand, higher age groups, specifically those within the age brackets of 31-40 with 26.40% and 61-70 with 24%, demonstrated decreased survival rates. The trend clearly shows the high vulnerability of older passengers in emergencies.

4. Impact of Family Size: The analysis revealed that family size affected survival rates, with individuals from smaller families (1 to 3 members) experiencing better survival outcomes, particularly families of three, which had a survival rate of 72.09%. Conversely, larger families (5 or more) faced significant challenges, resulting in lower survival rates. This suggests that smaller family units were better equipped to navigate the evacuation process.

In short, gender, passenger class, age, and family size were all important determinants of the outcome of the Titanic. The results indicate that women, children, first-class passengers, and medium-sized families had a greater survival rate, whereas men, older people, third-class passengers, and those with large-sized families were at greater risk. This analysis not only provides insight into the dynamics of survival in the Titanic disaster but also points out the larger implications of socio-economic and demographic factors in crisis situations.

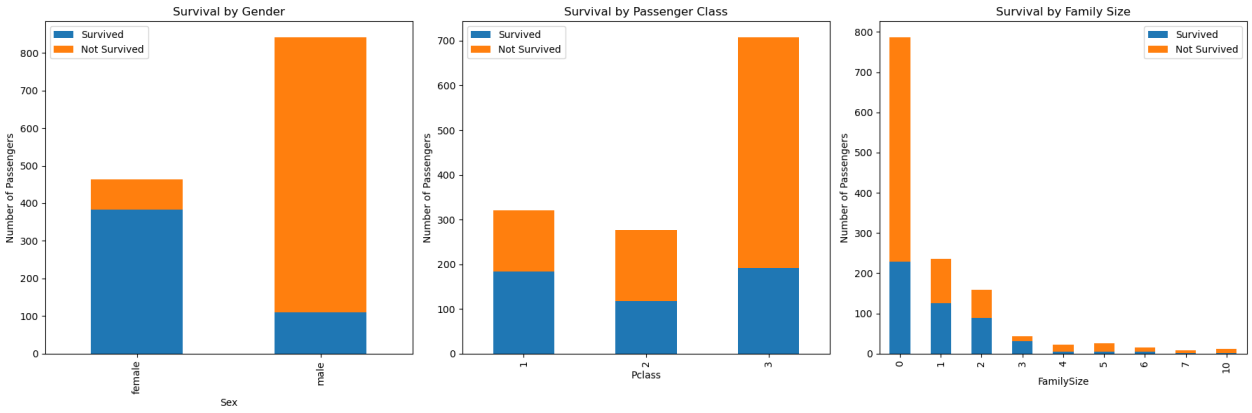

Figure 11: Summary of Survival Rates by Key Factors

References

[1] Kaggle, *Titanic Dataset*, 2024. Available: https://www.kaggle.com/competitions/titanic/data. [Accessed: 2024-10-01].

[2] History.com Editors, *Titanic*, 2010. Available: https://www.history.com/topics/early-20th-century-us/titanic. [Accessed: 2024-10-01].

[3] Wikipedia Contributors, *RMS Titanic*, 2024. Available: https://en.wikipedia.org/wiki/RMS_Titanic. [Accessed: 2024-10-01].

[4] Titanic II, *Titanic II Cabins*, 2024. Available: https://titanicll.wordpress.com/titanic-ii-cabins/. [Accessed: 2024-10-01].