Titanic Data Analysis Report

Introduction

The Titanic dataset from Kaggle gives information about the passengers on the ship, including their demographics, social status, and whether they survived the 1912 disaster. The goal of this analysis is to find patterns in survival, look at how factors like sex, passenger class, age and fare affected outcomes, and prepare the data for possible predictive modeling. The dataset comes in three files: a train dataset with 891 rows and 12 columns, a test dataset with 418 rows and 11 columns, and a gender submission file with sample survival predictions for the test dataset. Together, these provide a solid base for analysis.

Data Overview

The training dataset contains the target variable Survived (0 = did not survive, 1 = survived). Other features include demographic information (Sex, Age), socio-economic indicators (Pclass, Fare, Cabin), family details (SibSp, Parch), identifiers (PassengerId, Name, Ticket), and embarkation port (Embarked). Some features are categorical, such as sex and embarkation, while others like age and fare are numerical. A preview of the data shows that there were both children and adults in different classes, with varying ticket prices, family sizes, and boarding points.

Data Cleaning

Before analysis, the dataset needed cleaning. Several columns had missing values. The Age and Cabin columns had the largest gaps, Embarked was missing two values, and Fare was missing one value in the test set.

For Embarked, the two missing passengers were both first-class women, aged 38 and 62, traveling together with ticket number 113572. Their fare was 80 and their cabin was B28. Looking at passengers with similar fares and class showed that this fare was usually linked to Cherbourg (C). Therefore, their embarkation values were imputed as “C.”

For Age, missing values were replaced with the median age of 28. The median was chosen because it is less affected by extreme values and gives a fair central value for the dataset.

The missing Fare was filled with the median fare based on passenger class and embarkation point, since fares differed by both factors. The Cabin column was too incomplete to fill, so instead a new variable HasCabin was created, showing whether a passenger had a recorded cabin or not. Finally, outliers in fare were handled by applying a log transformation for modeling, while age outliers were kept but trimmed in charts for clarity.

Feature Engineering

New features were created to give more insight. FamilySize was calculated by adding siblings, spouses, parents, children, and the passenger. From this, a feature called IsAlone was created to mark passengers traveling alone. Titles such as Mr, Mrs, Miss, and Master were extracted from names, since these carry information about age, gender, and social role. Age and fare were grouped into categories (AgeGroup and FareGroup) to show clearer patterns. Lastly, the HasCabin variable was added to reflect whether cabin information was available, which often related to class and survival chances.

Findings from the data

a) Looking at Single Variables

When first looking at the data, survival was clearly imbalanced. Only about 38% of the passengers survived, while 62% died. This shows that the disaster claimed more lives than it saved. The passenger list also had more men than women, which becomes important when we compare survival rates later. Most passengers were in third class, meaning they were from lower socio-economic backgrounds. First-class passengers were fewer, but their survival rates turned out to be much higher.

The ports of embarkation also show differences. Southampton (S) was the most common port, followed by Cherbourg (C) and Queenstown (Q). This tells us about the routes people took, and later analysis shows that where passengers boarded had an influence on survival.

The age distribution reveals that the Titanic carried mostly young adults. However, there were also children and elderly passengers. This makes sense, as many families traveled together. The age pattern is important because younger passengers, especially children, were given preference during evacuation.

Fare values show a very uneven distribution. Many tickets were cheap, which is expected since third-class had the most passengers. However, some tickets were extremely expensive, belonging to first-class passengers. This created a “long tail” in the fare data. This difference in ticket prices is another sign of the clear class divide among the passengers.

a) Comparing Each Variable with Survival

When survival is compared to each variable, strong patterns emerge. The most obvious difference is between men and women. Women survived at a much higher rate than men, showing that the “women and children first” rule was followed. In fact, most men died, while many women, even from second or third class, managed to survive.

Passenger class also had a strong effect. First-class passengers had the highest survival rate, followed by second class, while third-class passengers had the lowest. This suggests that wealth and access to better parts of the ship, such as proximity to lifeboats, made a big difference.

Age was another important factor. Children were more likely to survive compared to adults, especially boys with the title “Master.” This again reflects the practice of saving children first. On the other hand, older men had very low survival chances.

Fares were strongly connected to survival as well. Those who paid higher fares, usually in first class, had much better chances of survival. Even within the same class, paying a higher fare often meant a better cabin location, which likely helped during evacuation.

The port of embarkation also made a difference. Passengers who boarded at Cherbourg (C) had higher survival rates than those from Southampton (S) or Queenstown (Q). This may be linked to the fact that many first-class passengers boarded at Cherbourg, giving them an advantage.

Family structure also influenced survival. Passengers traveling in small families of two to four people had the best chances. This may be because they could help each other during the chaos without being too large a group to manage. Passengers traveling alone had lower survival rates, likely because they had no one to assist them. Very large families also had poor outcomes, possibly because they could not all escape together. Having a recorded cabin also helped survival, as it was mostly first- and second-class passengers who had cabins assigned.

b) Looking at Combinations of variables

When variables are combined, the survival patterns become even clearer. The most powerful combination was sex and class. First-class women had the highest survival

rate, while third-class men had the lowest. This shows that both gender and wealth were key factors.

Age and sex also interacted strongly. Women of all ages had better chances of survival than men. Among men, younger boys survived more often than older men, again showing the focus on saving children.

Family size was another strong interaction. Small families survived the most, while those traveling alone or in very large families had much worse outcomes. This highlights the balance between having support but not being overwhelmed in large groups.

Even within the same class, money still mattered. First-class passengers who paid more for their tickets survived more often, which suggests that expensive cabins were in better positions for reaching lifeboats. This proves that survival was not only about class in general but also about how much individuals could afford to pay within that class.

Summary of insights

The most important predictors of survival were sex, passenger class, fare, age (especially with titles), and family structure (family size, being alone, and cabin presence). The findings match historical records: women and children had priority, wealthier passengers had better chances, and third-class passengers faced the worst conditions. Survival was highest for those in small families, while embarkation port and cabin access also played roles.

Preparing the data for modeling

The dataset was prepared by filling missing values, encoding categorical variables, and handling skewed features. Passenger class could be treated as ordered or converted into dummy variables. Fare was log-transformed to reduce skewness. To build predictive models, stratified train-test splits would keep survival ratios balanced. Logistic regression with interaction terms could provide a clear baseline, while tree-based models could capture more complex patterns.

Conclusion

In conclusion, survival on the Titanic was most likely for women in first class, especially those traveling in small families, paying higher fares, and boarding from Cherbourg. The least likely to survive were men in third class, particularly those traveling alone or in large families with cheap fares. These insights, supported by charts and visualizations, help explain the survival patterns and also prepare the dataset for building prediction models.