OSLOMET

OSLO METROPOLITAN UNIVERSITY
STORBYUNIVERSITETET

Statistical Learning
ACIT4510

Final Assignment

Titanic's survival rate analysis

Introduction

The Titanic incident, a tragic maritime disaster, occurred on April 15, 1912, when a British ocean liner sank in the North Atlantic Ocean. On its maiden voyage from Southampton, England, to New York City, the Titanic collided with an iceberg causing at least five of its sixteen compartments to rupture[2]. This resulted in many thousands of deaths and didn't care whether you were rich or poor, young or old, man or female, in the end everyone fell victim to its demise.
The purpose of this analysis is to explore the survival rate on different key variables based on the dataset[1] . These key-variables as we will go in depth later on this report are Survival status, gender, ticket class, age and fare. Combining these key variables we will observe that the outcome for the survival rate varies and this will help us come to a conclusion of how it really was back there on the Titanic.

Methods and Data

- **Survival status**
  The survival status here is described as “Survived” and the values are 1 and 0. 1 represents survived and 0 represents not survived. This variable is crucial and very relevant to our problem statement which is figuring out the survival rate, as it shows us which passenger survived or not.
- **Gende**r
  Gender in the dataset is described with the binary 0 and 1. 0 is for the female part and 1 for the male counterpart. In our analysis, gender plays a big role for distinguishing the survivability of the passengers based on the different policies that were implied when the crisis was ongoing.
- **Pclass**
  Pclass also known as Passenger class, is categorized into 3 different classes, First class, with number 1, being the high class, second class, with number 2 being the middle class and third class, with number 3 being the lower class. These categorizations are created based on the price that the passengers paid for their fare.

- **Age**
  Age is represented with decimals and is also a very important variable for our analysis. In the age column, I have divided according to the different ages, age groups. The reason age is a key variable is because it helps further explain the survival rate of these different age groups. Based on the facts for the event of the Titanic, older people and people in their late 20s were victims to the demise of the Titanic, while younger age groups were among the ones that had survived the incident due to the policies that mandated them to use the lifeboats first.
- **Fare**
  Fare similar to passenger class, describes the total amount of money each passenger spent for their fare. With fare, we will be able to create groups such as “Rich” and “Poor” as they were used later on to give additional insights on the survival rate.

As for our statistical methods I used chi-squared tests where we tested the null hypothesis that survival is independent of these categorical variables against the alternative hypothesis and that there is an association between them. Moreover I have used different graphs such as bar-graphs, Kaplan-meier survival curves and scatterplots to investigate potential and discover potential relationships between our key-variables. Finally I have performed a logistic regression model that helped me check the validity of the model recognizing the survival rate based on the key variables.

Results

Overall Survival Rate

Fig 1 shows a simplified visual comparison between the proportion of Titanic passengers who survived and those who did not survive.

- The blue bar represents the proportion of passengers who survived the Titanic disaster.
- The orange bar represents the proportion of passengers who did not survive.

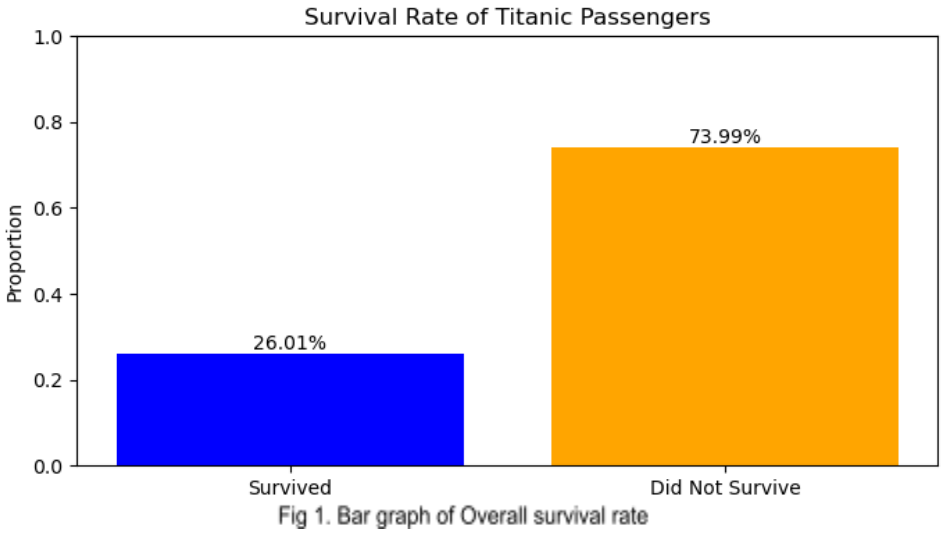


Fig 1. Bar graph of Overall survival rate

- The height of each bar indicates the relative proportion of each group, with the y-axis showing the scale from 0 to 1, which corresponds to 0% to 100%.

This is consistent with historical accounts, which report that over half of the passengers and crew on the Titanic did not survive after it struck an iceberg and sank.

Breakdown of Key Variables:

In this breakdown we are going to analyze and create helpful graphs that will explain in detail how the survival rate of the passengers on the Titanic were based on the key variables which are: Gender, Age, Fare, PClass.

Kaplan-Meier survival curves

Survival probability Gender vs age:

For this representation I have decided to use Kaplan-Meier Estimator. Although Kaplan-Meier estimators use 'time' which refers to the duration until an event occurs, I have decided to proxy time with age.

Fig 2 shows curves for two groups, differentiated by gender: males (in blue) and females (in orange).

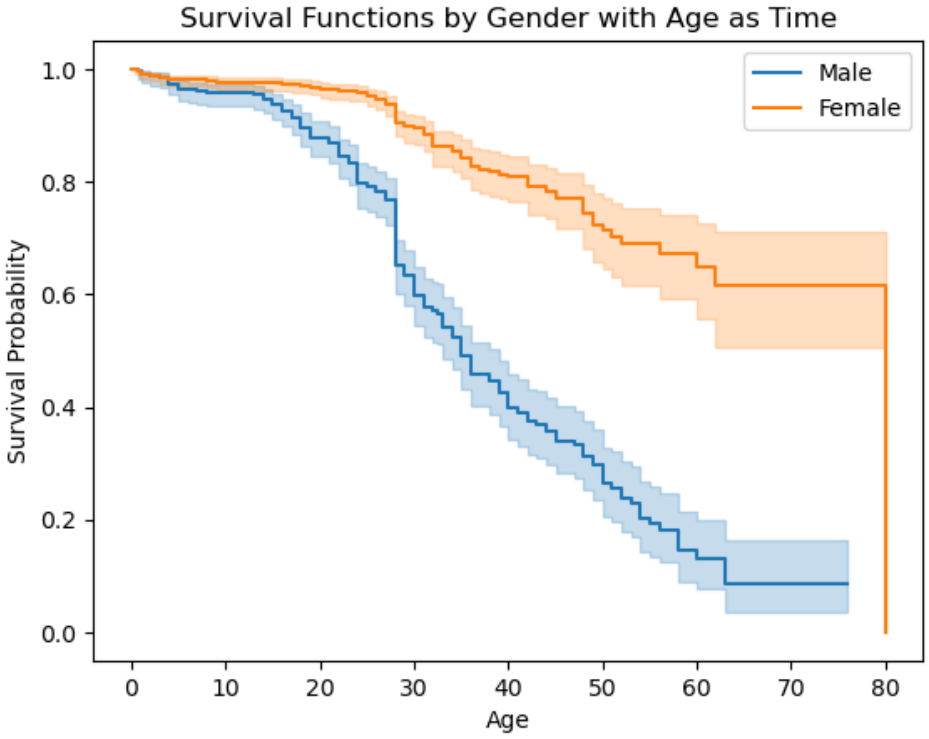


Fig 2. Kaplan meier curves for genders and age

As the graph suggests there is a higher survival probability for females as there is for males. Throughout the entire age range, the survival probability for females is consistently higher than that for males. This is indicative of a significant difference in survival likelihood between genders, with females more likely to survive.

Both males and females survival probability decreases with age. However there is a convergence of survival probabilities at older ages (around 70+), suggesting older ages (around 70+), suggesting that the survival advantage for females diminishes as age increases.

As for the male passengers, the curve takes a steeper decline, particularly from the start between the late 20s, indicating that age had a more pronounced impact on survival for males than for females in that age range. The shaded areas around the curves represent confidence intervals, providing a sense of the uncertainty around the survival probability estimates. The wider the shaded area, the more uncertainty there is in the estimate. As the graph suggests that the confidence intervals widen for older ages, which is expected due to smaller sample sizes for higher age groups.

Survival probability rich vs poor

We use Kaplan Meier curves once more, to figure out this time the difference between rich and poor. In *Fig 3* we are able to distinguish the survival probabilities between the rich and poor based also on their ages.

The y-axis represents the survival probability, ranging from 0 to 1 (0% to 100%). The x-axis represents age, which, although not a typical time-to-event metric used in survival analysis, is used here as a proxy to indicate the point at which the observation was made.

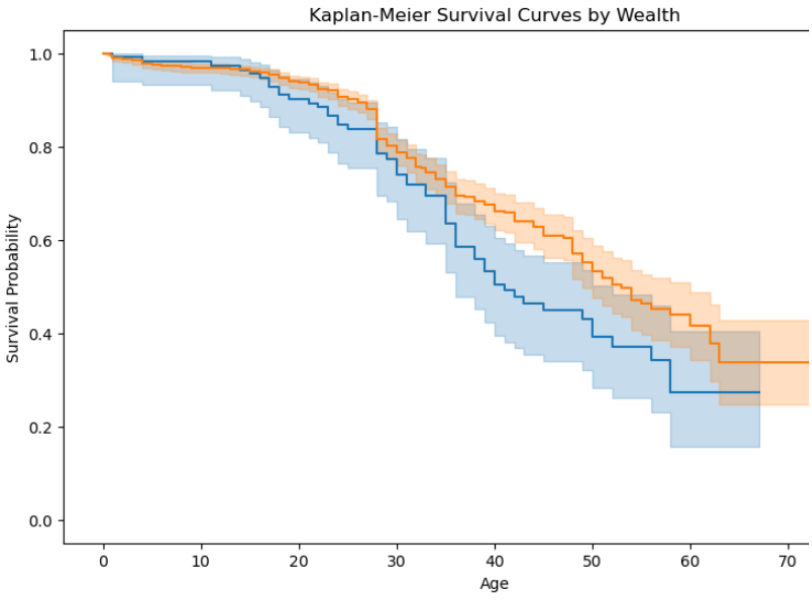


Fig 3. Kaplan meier curves for rich(Light orange) vs poor(Light blue) and age

We create the groups ‘Rich’ and ‘Poor’ by categorizing the Fare column in our dataset.The **'Rich'** *(Light Orange*)group corresponds to passengers with higher ticket price, while the **'Poor'***(Light blue)* group corresponds to passengers with lower ticket price.

As we can observe from the graph, initially, both **'Rich'** and **'Poor'** groups start with a survival probability close to 1, indicating that at younger ages, the survival probability is high for all. However, as age increases, the survival probability decreases for both groups. On the other side, the survival probability for the 'Rich' group is consistently higher than for the 'Poor' group across all ages.The gap between the curves suggests that wealthier passengers had a better chance of survival, which could be due to various factors, such as priority access to lifeboats.

Bar graphs:

Now we will move on to using Bar graphs for visualizing the survival rate based on these factors:

- Sex
- Age groups
- Pclass (High class, middle class, lower class)

Bar graph of Survival Rate by Gender:

The *Fig 4.* compares the survival rates between male and female passengers. The survival rate for males is significantly lower at approximately 12.93% compared to females, which is around 49.78%. This substantial difference highlights the impact of the 'women and children first' policy that was largely followed during the evacuation of the ship.

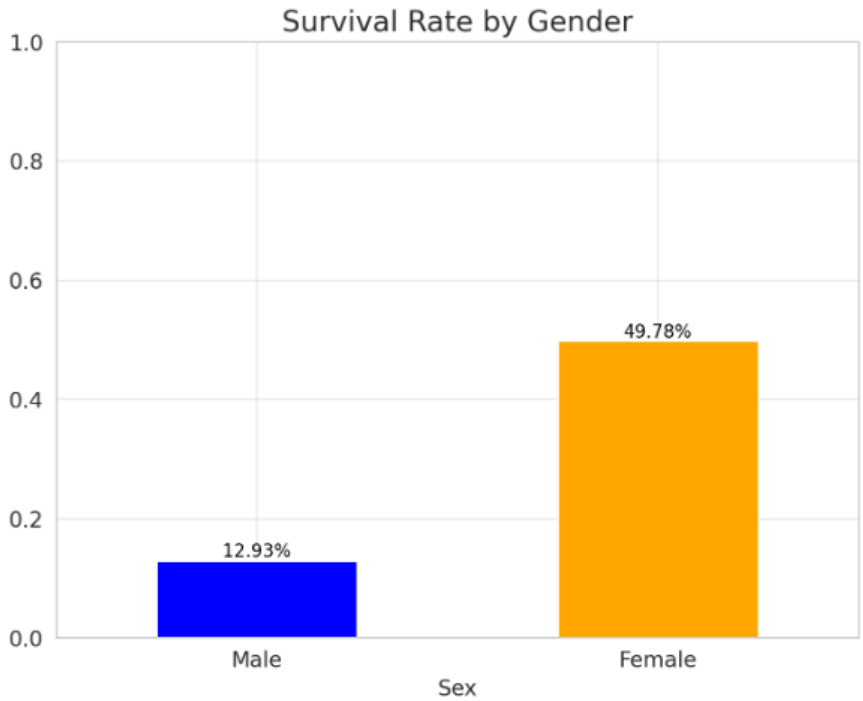


Fig 4. Bar graph with the survival rate based on gender

Bar graph of Survival Rate by age groups

Fig 5, I have taken every age in the dataset and created based on them different age groups:

- Child
- Teen
- Adult
- Senior

The survival rates among different age groups are depicted, with children having the highest survival rate at approximately 42.55%. The survival rate decreases with age, with teens at about 30.30%, adults at 24.58%, and seniors at 12.50%. This gradient suggests that age was a factor in survival, with younger passengers being more likely to survive, which is in line with historical accounts prioritizing the evacuation of children.

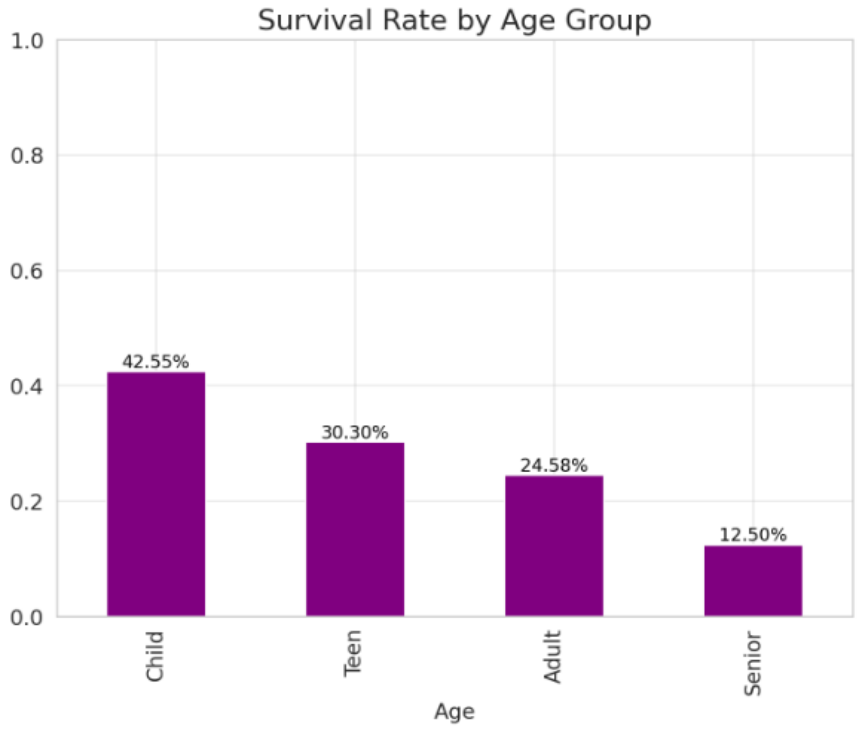


Fig 5. Bar graph with the survival rate based on ages

Bar graph of Survival Rate by ticket class

Fig 6 illustrates the survival rates for the three passenger classes on the Titanic. First-class passengers had the highest survival rate at 41.74%, followed by second-class passengers at 31.41%, and third-class passengers had the lowest at 16.78%. The graph indicates a socio-economic divide in survival chances, with wealthier passengers having had better access to life-saving resources.

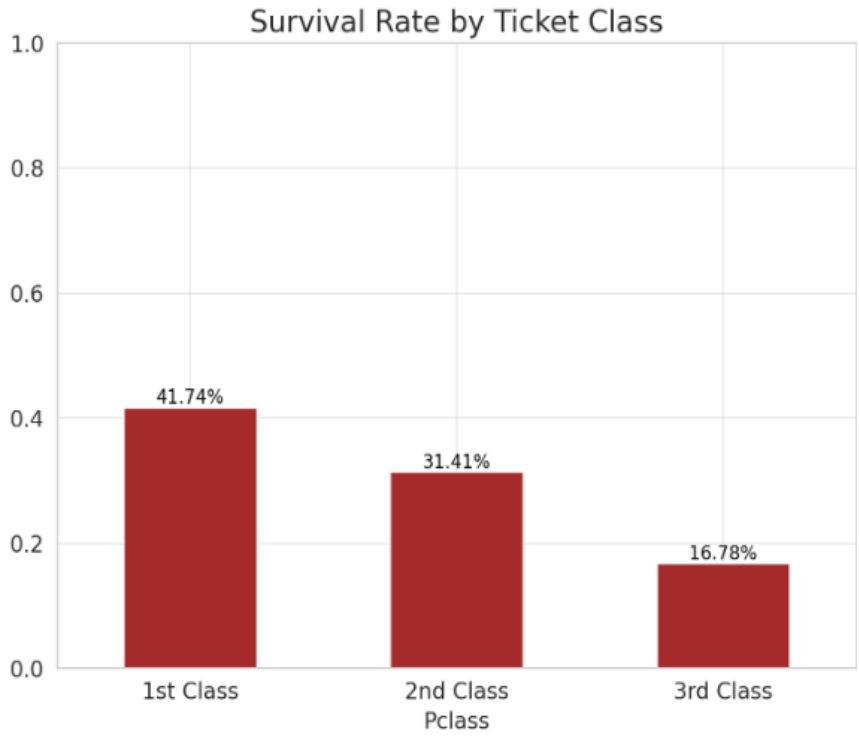


Fig 6. Bar graph with the survival rate based on Passenger class

Scatterplots:

Now we move to the different scatterplots that their purpose are to further explain in more detail the survival rate of every passenger based on the following parameters:

- Age and gender
- Fare and gender
- Age and Fare

Survival scatter plot by age and gender

Fig 7. Scatterplot by age and gender. More males died than females

The x-axis represents the age of the passengers. The y-axis has two discrete values, 'Did Not Survive' (0) and 'Survived' (1), with some jitter (random noise) added to the points to avoid overlapping and make individual data points more distinguishable. Blue points represent male passengers, and red points represent female passengers. The distribution of blue and red points across the 'Did Not Survive' and 'Survived' categories gives a visual indication of the survival pattern for different ages and genders. As we can observe from *Fig 7,* we can see that the majority of male passengers ended up not surviving the incident than females. On the other side, females were scattered much more on the survived

area which makes total sense how women were prioritized to use the lifeboats than men. We can also observe a chunk of dots gathered in both conditions (survived, not survived) between the ages of early 20s to late 20s.This is also due to the fact that many passengers were around those ages.

Survival scatter plot by gender and fare

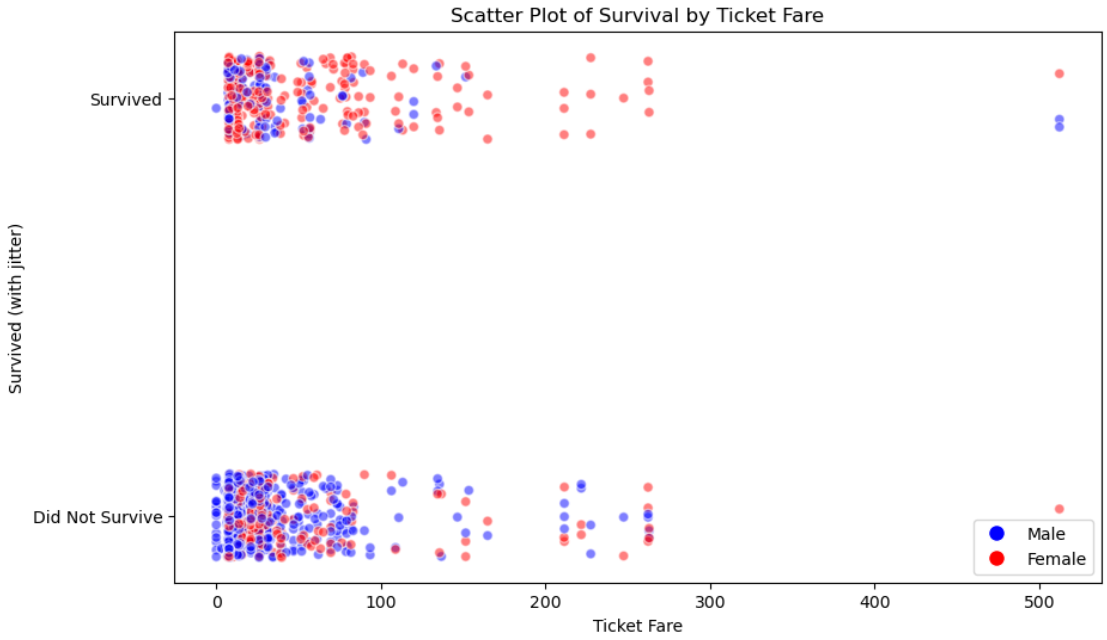


Fig 8. Scatterplot by fare and gender.

Fig 8 illustrates the relationship between ticket fare and survival on the Titanic, with the survival status shown on the y-axis (0 for 'Did Not Survive' and 1 for 'Survived') and ticket fare on the x-axis. The points are color-coded by gender, with blue representing male passengers and red representing female passengers.

There is a higher concentration of points toward the lower fare range, indicating that most passengers bought cheaper tickets. The spread of points becomes sparser as the fare increases. As we move on, the two distinct horizontal bands represent those who did not survive (at y=0) and those who survived (at y=1). There is a higher concentration of red points (female passengers) in the 'Survived' band, suggesting that females had a higher survival rate than males regardless of how much they had paid for their tickets, which is consistent with

historical records and our previous analysis. Those who had 0 as Ticket fare, were most likely the staff of the ship who based on the graph were men which as we can remember from the past, men worked primarily on boats than women.The dataset does not explain the different occupations of the passengers but if i were to guess, i would suggest that some of the dots around that area may had been the crew of the boat. Moving on to the different fare prices, passengers who paid higher fares, particularly over $100, appear to have a higher survival rate. This observation aligns with the historical context, where wealthier passengers, likely in higher classes, had better access to lifeboats. Moreover it's noticeable that there are some passengers with very high fares (extending towards $500), and these points seem to be associated with survival, reinforcing the idea that higher-paying passengers had better chances of survival.

The survival rate for females (red points) is visibly higher across all fare ranges, which further supports the "women and children first" policy during the ship's evacuation. There is a notable cluster of blue points (male passengers) at the lower fare end who did not survive, suggesting that male passengers in lower classes were less likely to survive.

Survival scatter plot by age-groups vs ticket fare:

Fig 9. Scatterplot by fare and age groups

Fig 9 visualizes the survival of passengers on the Titanic as a function of their ticket fare and age group. Each dot represents a passenger, with the position along the y-axis indicating survival status (Survived or Did Not Survive) and the position along the x-axis indicating the fare they paid. The colors represent different age groups.

The y-axis is binary, with the top representing those who survived and the bottom those who did not. The x-axis shows the ticket fare from 0 to 300. As we can observe on the scatter plot, there's a wide spread of fares, with a clustering of points at the lower end of the fare scale, indicating that a large number of passengers bought cheaper tickets. Some passengers who paid higher fares, which are scattered across the right side of the plot, seem to have higher survival rates. This suggests that passengers with more expensive tickets had better survival chances, which could be correlated with higher class and better access to lifeboats.

The colors indicate the passenger's age group. It appears that all age groups are represented across the fare spectrum. However there are key differences in different age groups:

- The 'Child' group (blue) has a spread throughout the fare range, with a good number of survivors, possibly due to the prioritization of children for lifeboat seats.
- The 'Senior' group (purple) seems less represented, which might be due to fewer passengers in that age group, and shows a mix of survival outcomes.
- The ‘Adult’ group (red) seems to have not survived the incident.

Chi squared tests for categorical variables

In this section I ran the chi-squared tests. Here I tested the null hypothesis that survival is independent of these categorical variables against the alternative hypothesis that there is an association between them. Here are the results that I have figured out by implementing these tests:

```
Out[28]: (1.9617694244098499e-47, 0.00034815696365977044, 2.0631614019402776e-17)
```

Fig 10. Chi squared value

Case 1: Gender and Survival: The p-value is approximately **1.96 × 1 0 $^{-47}$**. This value, which is significantly lower than the typical alpha level of 0.05. This indicates that there is a statistically significant association between gender and survival on the Titanic.

Case 2: Age Group and Survival: The p-value is approximately **0.000348** which also is well below the alpha level of 0.05, suggesting a statistically significant association between age group and survival.

Case 3: Ticket Class and Survival: The p-value is approximately **2.06 × 1 0 $^{-17}$**, which is again much lower than 0.05, indicating a statistically significant association between ticket class and survival.

In all of the three cases, the null hypothesis that survival is independent of these categorical variables were rejected. There is strong evidence to suggest that gender, age group, and ticket class were all associated with the likelihood of survival on the Titanic.

Logistic regression

For the logistic regression model, I considered 'Survived' as a dependent variable and 'Sex', 'Age', and 'Pclass' as independent variables. 'Sex' is encoded as a binary variable (1 for male, 0 for female) if it's not already, and use the age bins and 'Pclass' as categorical variables in the model.

The model's coefficients for the predictors are as follows:

Sex (Gender): The coefficient is approximately ***1.66***, suggesting that being female (coded as '0') is associated with higher odds of survival compared to being male (coded as '1').

Pclass (Ticket Class): The coefficient is approximately ***-0.75***, indicating that higher classes (lower numerical 'Pclass' values) are associated with higher odds of survival.

Age: The coefficient is approximately ***-0.025***, which implies that as age increases, the likelihood of survival decreases slightly. The intercept of the model is about ***0.47***.

Based on the classification report we can say that for non-survivors ('0'), the precision is **0.77**, and for survivors ('1'), it's **0.70**. This indicates that the model is slightly better at predicting non-survivors than survivors. The model has a recall of **0.94** for non-survivors and 0.32 for survivors, suggesting that it is quite good at identifying non-survivors but not as effective at identifying survivors. The *F1*-score is a balance between precision and recall, being **0.85** for non-survivors and **0.44** for survivors, indicating that the model predicts non-survivors with higher accuracy. The overall accuracy of the model is **0.76**, meaning it correctly predicts the survival status **76%** of the time on the test set. Based on these

values that our model generated brings it to the conclusion that it performs well, particularly in predicting non-survivors.

To further ensure the accuracy level of the model, I used Receiver Operating Characteristic. The ROC curve on *Fig 11* illustrates the true positive rate (sensitivity) against the false positive rate (1 - specificity) at various threshold settings. The area under the curve (AUC) is a measure of the model's ability to distinguish between the classes. An AUC of 1.0 represents a perfect model, while an AUC of 0.5 represents a model that is no better than random guessing. That being said, the **AUC** for this model is around **0.76**, which suggests that the model has a good ability to distinguish between survivors and non-survivors. However, there is still room for improvement, as the curve is not at the top left corner.

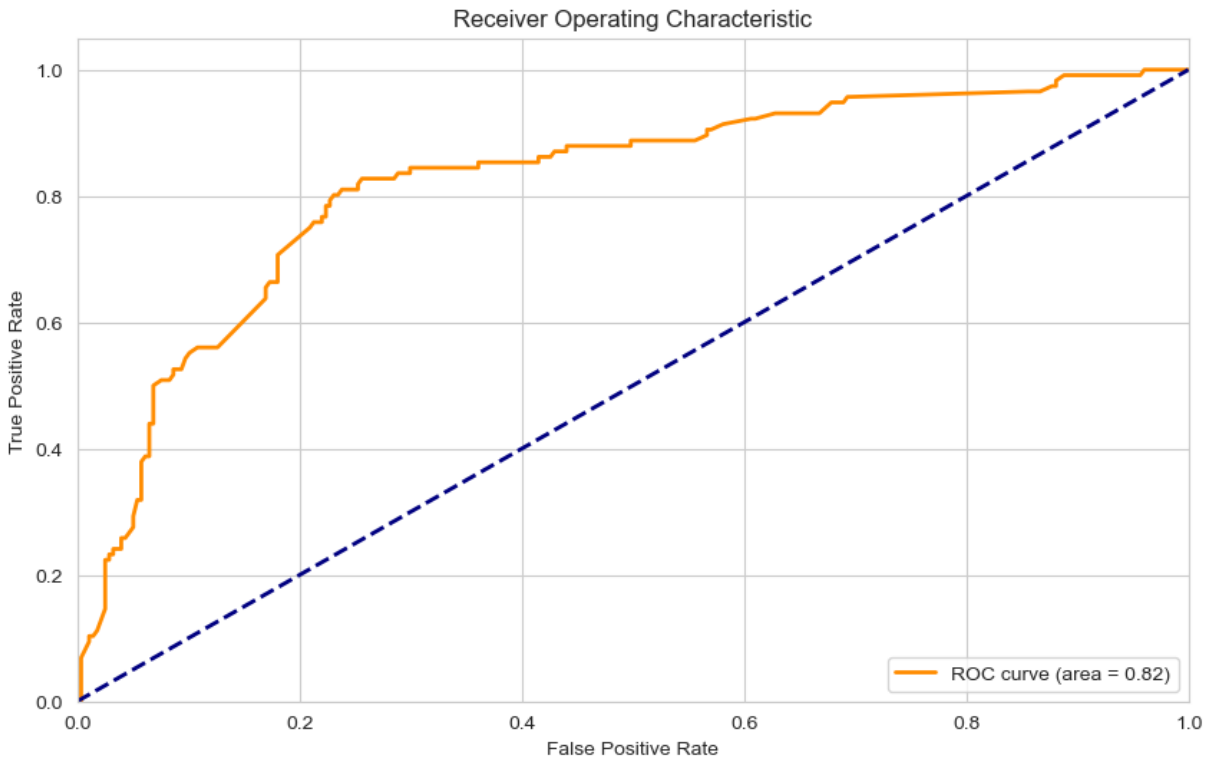
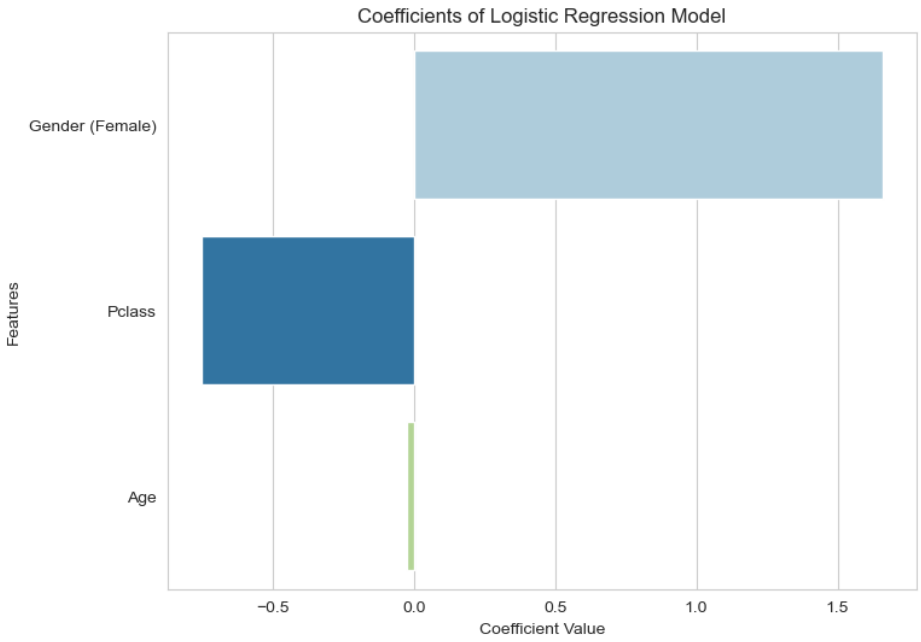


Fig 11. ROC graph indicating the accuracy of the model

Discussions

Interpretation of Results

With our logistic regression model, we end up with the fact that being female significantly increased the likelihood of survival. As we can see from the graph, the bar extends to the right, signifying a positive relationship with the outcome (survival). This also aligns with historical accounts that women were given priority during lifeboat boarding. Moreover in accordance with the Ticket class, the negative coefficient for ticket class indicates that as the class number increases, meaning going from 1 to 3 that is from high class to lower class, the likelihood of survival decreases.The bar extends to the left, indicating a negative relationship. Since first-class tickets, this suggests that passengers in higher classes had better survival odds. However the age predictor had a small negative coefficient, implying that as age increases, the probability of survival slightly decreases.The very short bar, close to zero, suggests a weaker relationship compared to gender or class.This suggests that younger passengers had slightly better chances of survival, which could be due to prioritization of children for lifeboats or greater physical resilience.



Gender-Based Survival Patterns

One of the most striking observations is the significant gender-based disparity in survival rates. The Kaplan-Meier curves and bar graphs show that females had a higher probability of survival than males. This aligns with the historical narrative

of the 'women and children first' policy during lifeboat allocation. Logistic regression further substantiates this, with gender emerging as a strong predictor of survival. The chi-squared test confirms the statistical significance of this association, rejecting the null hypothesis of independence between gender and survival.

Age as a Factor in Survival

Age also played a crucial role in survival chances. While younger passengers, particularly children, had higher survival rates, the likelihood of survival diminished with age. This pattern is evident in both Kaplan-Meier curves and bar graphs categorizing survival rates by age groups. The logistic regression model further quantifies this trend, showing a negative correlation between age and survival likelihood. The convergence of survival probabilities at older ages suggests a diminishing age-related advantage, possibly due to the physical challenges of evacuation (handicap people or injured old people).

Socio-Economic Divide and Survival

Based on all of the analysis we get to see that a clear socio-economic divide in survival probabilities. Passengers in higher classes had significantly better survival rates, as shown in the bar graphs and scatter plots. This trend might be attributable to better access to life-saving resources like lifeboats and preferential treatment during the evacuation process. The chi-squared test results further reinforce the statistical significance of the association between ticket class and survival.

Fare and Survival Correlation

With the help of the scatter plot analysis, it helped me to reveal an interesting correlation between fare and survival. Higher fares, possibly indicative of higher socio-economic status, correlated with better survival chances. This observation is consistent across gender and age groups, suggesting that wealth was a critical factor in survival, independent of other variables.

Conclusion

In the analysis of the Titanic dataset, we observed that survival during the tragic event was profoundly influenced by a combination of gender, age, and socio-economic status. Women had a notably higher survival rate compared to men, reflecting the historical 'women and children first' policy that was likely in effect during the lifeboat allocations. Additionally, age emerged as a significant factor, with younger passengers, particularly children, showing higher survival rates, suggesting a prioritization of the young and vulnerable during rescue efforts.

The socio-economic status, as indicated by passenger class and fare, also played a crucial role in survival chances. Passengers in higher classes, who typically paid higher fares, had substantially better survival rates. This trend points to a socio-economic divide where wealth and status may have granted better access to life-saving resources like lifeboats.

Moreover, the logistic regression model reinforced these observations, identifying gender, age, and socio-economic status as key predictors of survival. This comprehensive analysis not only aligns with historical narratives but also provides a statistical backdrop to understand the dynamics of human behavior and decision-making in crisis situations. It highlights the need for equitable treatment and access to resources in emergency responses, offering invaluable lessons for managing disasters and crises.

References

[1] Heptapod. (2017). Titanic Dataset [Data set]. Kaggle. https://www.kaggle.com/datasets/heptapod/titanic

[2] "Titanic Timeline and Facts." Britannica, https://www.britannica.com/story/titanic-timeline-and-facts Accessed [15 November 2023].