Introduction to Data Science: Assignment 1

Jisoo Lim (100516011) & Alyssa Wright (100517836)

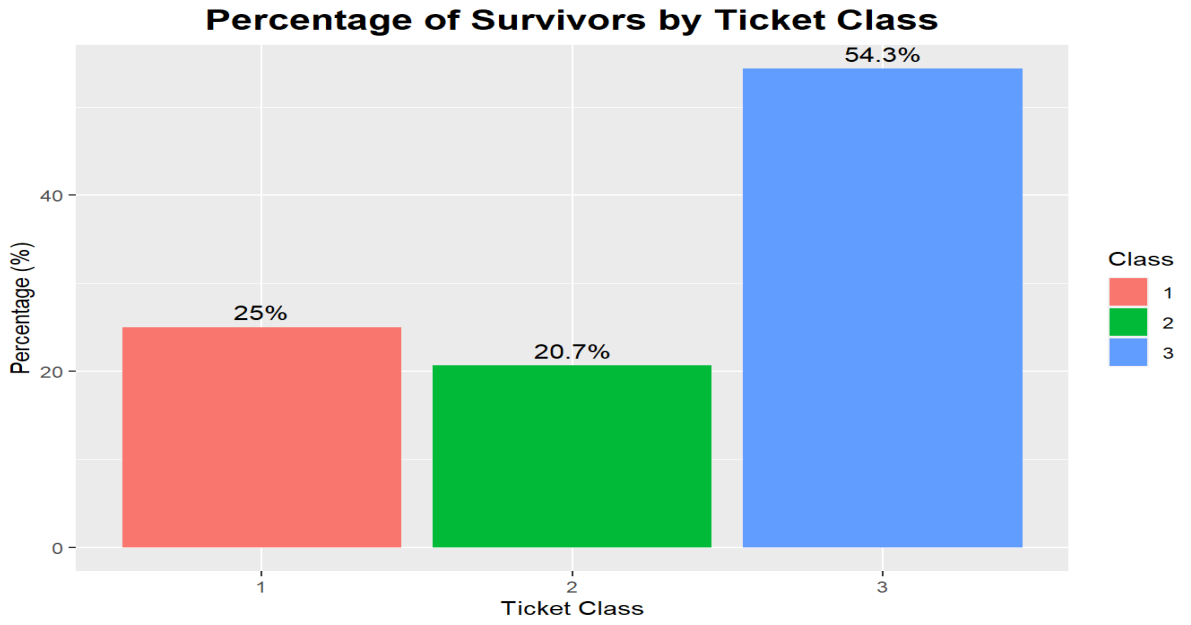
Exploratory Data Analysis

1. Did passengers from a particular class have a higher likelihood of survival compared to others? How did survival rates vary by class?
2. Were women more likely to survive the Titanic disaster compared to men? How significant was the gender-based difference in survival rates?
3. Did age play a role in survival? Were children more likely to survive, or did age not significantly impact survival rates?
4. Did the survival rate vary for passengers of different ages within each ticket class?
5. Were passengers who had family members (siblings and spouses) with them more likely to survive compared to those traveling alone?
6. Were passengers who had family members (siblings and spouses) with them more likely to survive compared to those traveling alone?
7. Were there any interesting interactions or correlations between the variables that were not evident from individual analysis, such as the combined effect of ticket class, gender, and fare on survival?
8. Did the location from which the passengers embarked have an influence on their survival?

Introduction

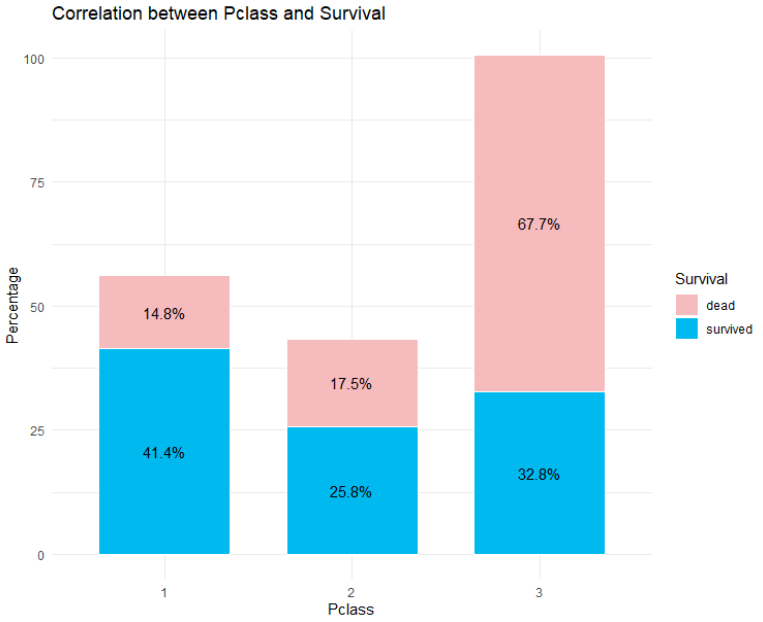
When we think of the last scene of escape from the movie “Titanic”, most passengers on lifeboats were first-class passengers, and there were many young children and female passengers. Was that true in reality? That's why we explored data focusing on survival.

Question 1 Did passengers from a particular class have a higher likelihood of survival compared to others? How did survival rates vary by class?

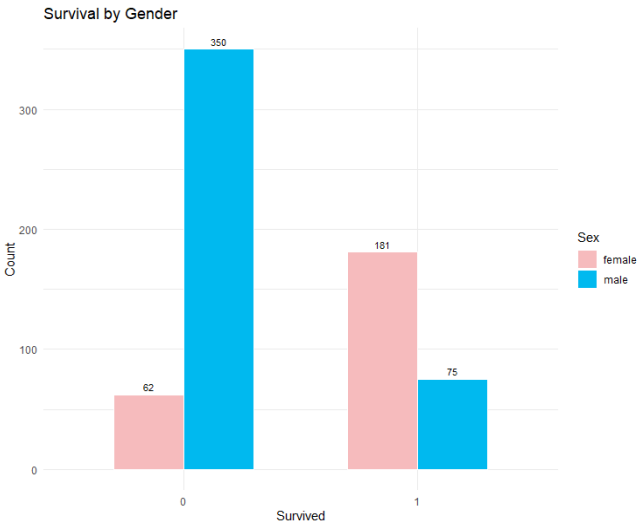


We first analyzed the percentage of passengers that survived within each ticket class. Class 1 tended to be the wealthier passengers, American and British elite, who most likely paid more for their fares. These wealthier passengers typically resided at the top of the boat where nicer amenities and rooms were located. The Class 2 passengers were leisure tourists, academics, members of the clergy, and middle-class English, Scottish and American families. The Class 3 passengers were immigrants hoping to begin new lives in the United States and Canada to save enough money for their families back in their home countries.

We calculated the frequency of survivors within each class using frequency tables and solved for their percentages, and the bar chart above shows these percentages. We can see that Class 3 passengers had more survivors than the other two classes out of all of the surviving passengers. The graph shows that among the surviving passengers, the proportion of passengers on first-class tickets is high. On the other hand, most passengers who did not survive are passengers on third-class tickets.



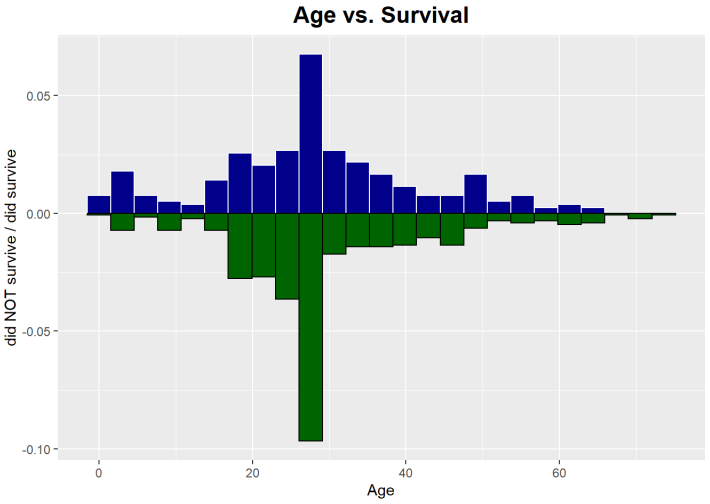
Question 2 Were women more likely to survive the Titanic disaster compared to men? How significant was the gender-based difference in survival rates?



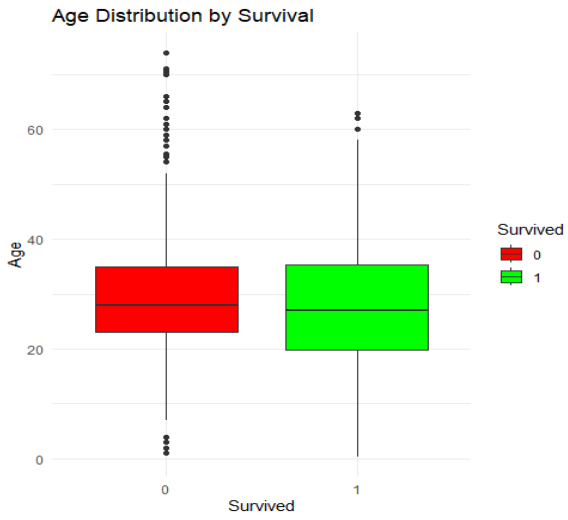
There were a total of 668 individuals. Among them, there were 425 males and 243 females. When expressed as percentages, the distribution is approximately 64% for males and 36% for females. There were 350 male deaths and 62 female deaths. The death rate is approximately 52% for males and 9% for females. The male death rate is approximately 5.65 times higher than the female death rate. On the other hand, examining the survival rates, we find that the survival rate is approximately 48% for males and 91% for females. The female survival rate is about 2 times higher than the male survival rate.

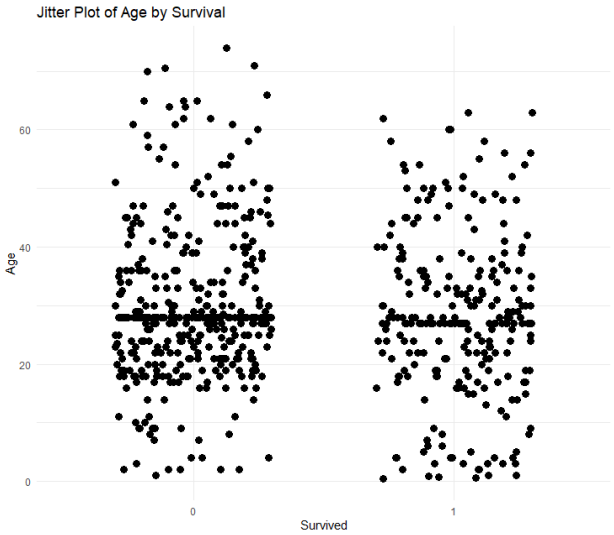
Question 3 Did age play a role in survival? Were children more likely to survive, or did age not significantly impact survival rates?

One would think that during an emergency situation such as the Titanic that children and older people would have priority. We tested this theory using a frequency table and histogram. The top dark blue bars were passengers that survived, and the bottom green bars were passengers who did not survive. This histogram demonstrates that the theory is partially correct: there were a higher number of adolescent passengers who survived than did not survive, however, elderly passengers had more of a 1 to 1 ratio with the survival rate. To continue, answering this question, we can see that middle aged passengers had higher death



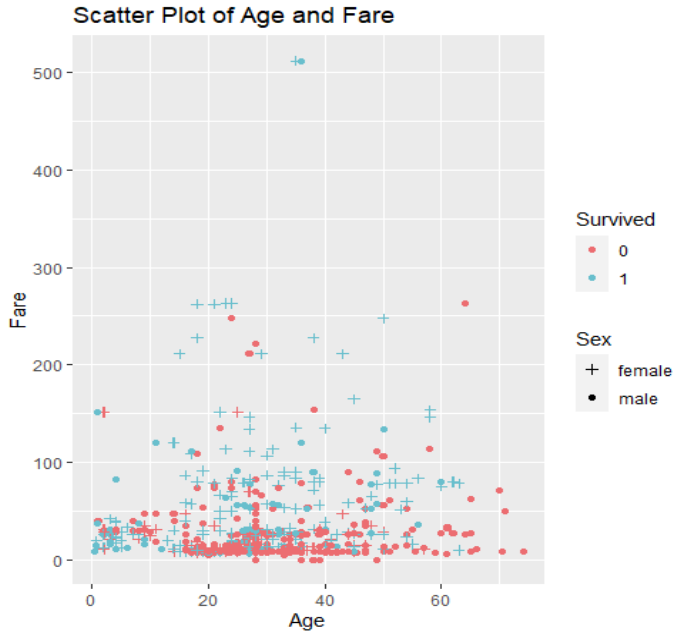
rates than children, but there were less youths on the ship, as well. Considering all of this, we think it is still valid to say that being older negatively affected the survival of a passenger.





It seems that the age is relatively low, but there is no significant difference. However, when we checked the data with a jitter plot, it revealed that the relatively lower age group exhibited a higher survival rate, while, conversely, passengers in the higher age group showed a lower survival rate.

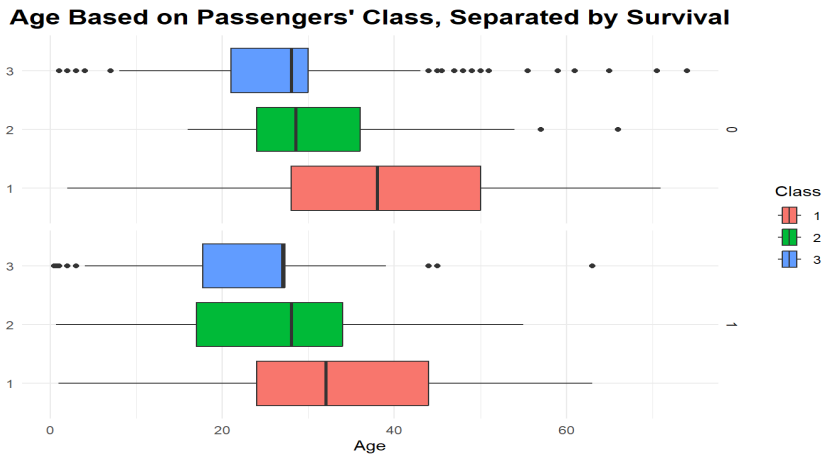
Question 4 What is the effect of a passenger's age and sex on his/her survival?



According to this scatter plot, it becomes possible to grasp a more comprehensive overview of the data at a glance. Looking to the left side of the graph, it is noticeable that even with lower fares, the survival rate is higher for children. Most of the adult deaths were passengers with lower-class tickets/lower fares, particularly indicating a lower survival rate for male passengers. Towards the upper part of the graph, as age increases, passengers with expensive tickets show a higher survival rate, and female passengers show a relatively higher survival rate compared to males.

Question 5 What is the effect of a passenger's age and class on his/her survival? Did the survival rate vary for passengers of different ages within each ticket class?

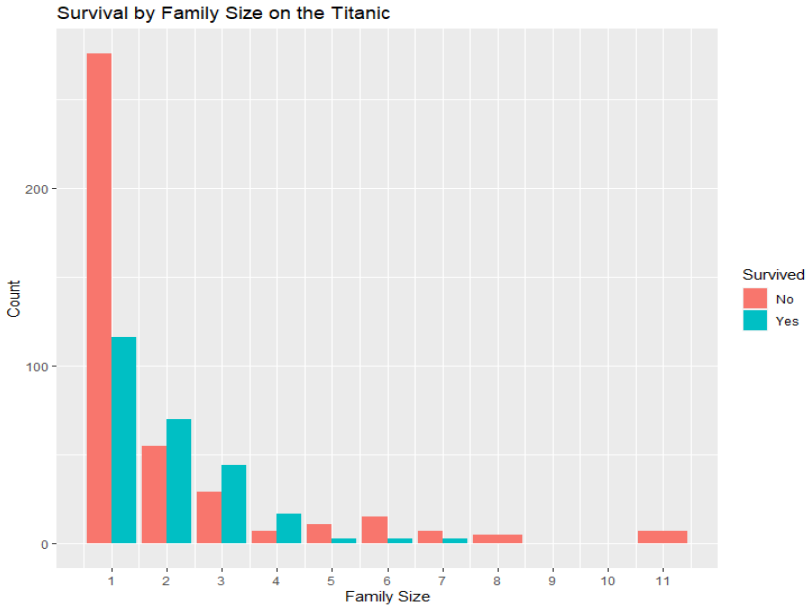
In the top portion of the box plots labeled 0, also known as not surviving, we see that the interquartile range for Class 3 was between the ages 20 to 30 years old, Class 2 was 25 to 35 years old, and Class 1 was 30 to 50 years old. For those who survived in the lower section, we can see all of the box plots have been shifted left suggesting that regardless of class, that the age ranges were younger in comparison to the groups who did die from the incident. Even more, when we take a look at Class 2 passengers who survived specifically, we noticed



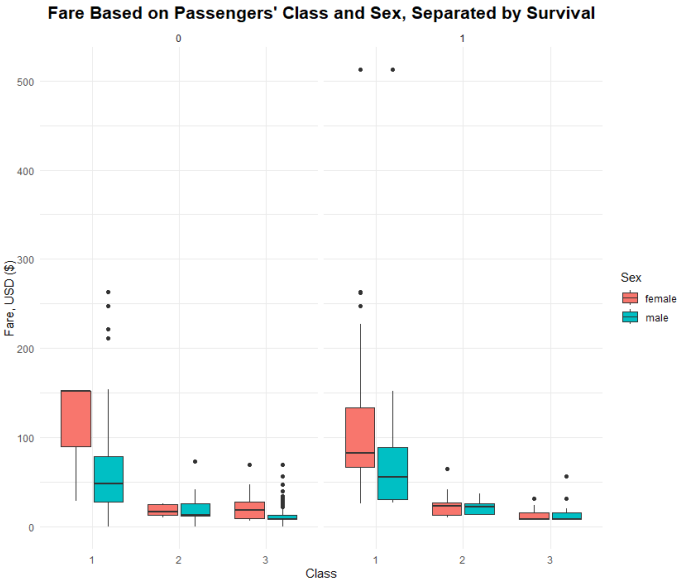
that the interquartile range was slightly larger, approximately having ages 15 to 35 years old passengers.

Question 6 Were passengers who had family members (siblings and spouses) with them more likely to survive compared to those traveling alone?

The visualization of passenger survival reveals distinct patterns based on the family size. Individuals on board exhibit a significantly lower survival rate. Majority of the passengers who survived were those who had 1 to 2 family members on board of the Titanic. However, for groups ranging from 5 to 11 family members or traveled alone, the survival rate, once again, showed a decline. This trend suggests that being aboard the Titanic with a moderate-sized family may have been advantageous for survival. Nevertheless, as the group size exceeds a certain threshold, the chances of survival appear to decline. This trend shows the impact of family size on survival outcomes aboard the Titanic.

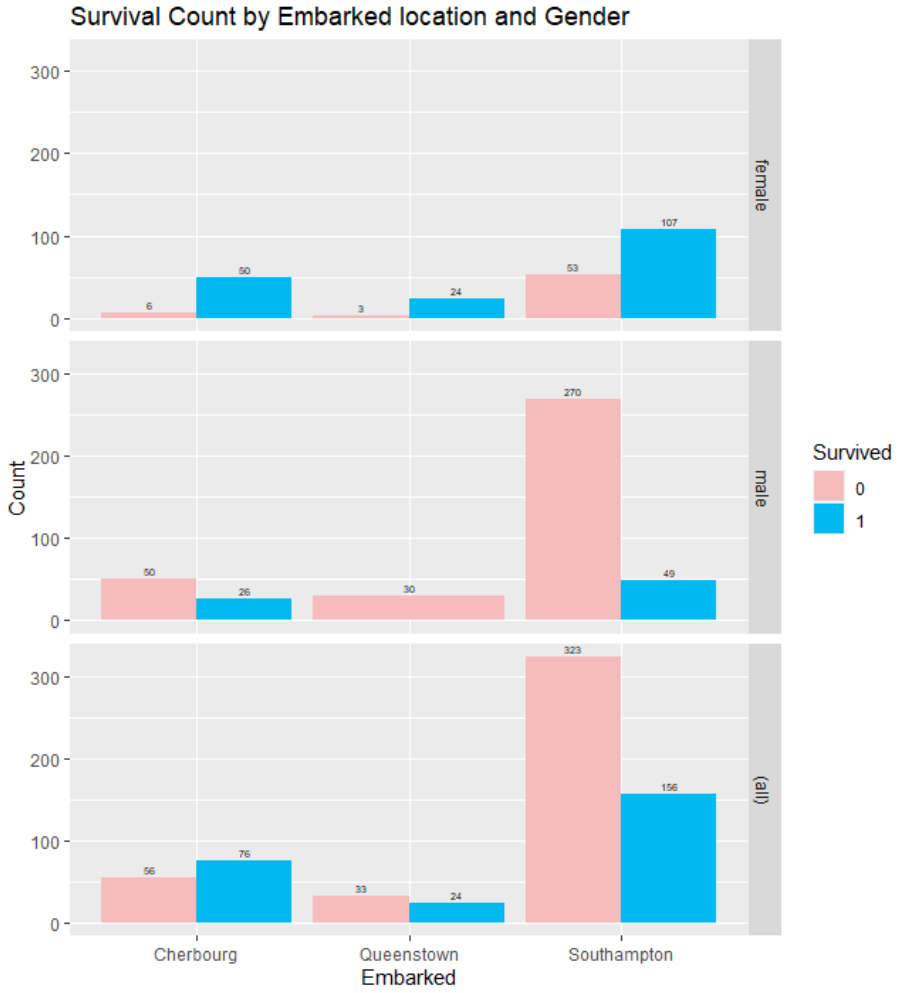


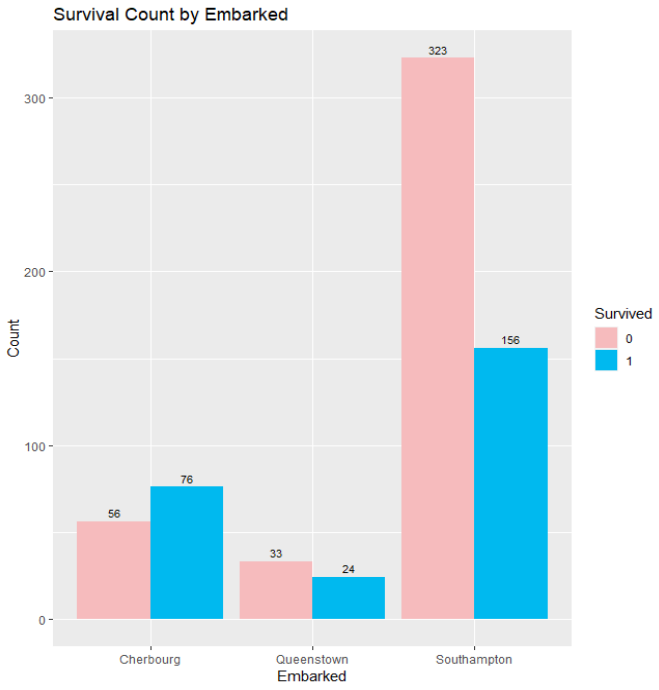
Question 7 Were there any interesting interactions or correlations between the variables that were not evident from individual analysis, such as the combined effect of ticket class, gender, and fare on survival?



In these box plots we see that the first class female passengers were most likely to purchase more expensive tickets, but there was a wider range of those types of passengers. Despite this increase in pay, the fares that each passenger paid did not seem to influence a person's death because the average fare price of those who died or survived in any of the classes are fairly similar to those who survived.

Question 8 Did the location from which the passengers embarked have an influence on their survival?





Analyzing this bar graph from each of the locations that the Titanic embarked from, it appears that the majority of the deaths of passengers embarked from Southampton while the least amount of deaths came from Queenstown. Both Cherbourg and Queenstown also have a more even ratio of those that survived or did not, unlike Southampton.

Considering all of the graphs, we can conclude that a wealthy passenger between the ages 20 to 30 years old that embarked from Southampton alone or with a considerably large family (more than 3 family members) was more likely to not survive the Titanic than any other demographic.