# Lab 1 Report

Jiayue Meng

*Abstract*— **This report illustrates the relationship between the survival rate of passengers on the Titanic and factors like gender, age, and social status. Through box plots, bar graphs, heatmaps, and histograms, the results of the analysis are visualized.**

## I. INTRODUCTION

In the report, I expect to see that women and children have a higher survival rate since passengers might follow the "women and children first" principle. I also expect that passengers who paid more have a higher survival rate since they probably have priorities and more opportunities to receive help. The survival situation, survival rates, and correlations between numerical variables are analyzed in this report. Passengers are divided into different groups according to their class, title, and gender. It is found that females and children have higher survival rates than males do. Thus, a reasonable inference – "women and children first" was applied – can be made. Also, it's found that the higher the ticket fare, the higher the survival rate. It might be because richer passengers in better classes have more chances to board lifeboats, which matched my expectation.

## II. DATA

In the project, we have 891 data in the titanicdata.xslx. There are 12 attributes in the data: PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, and Embarked. Besides these attributes, I created NotAlone according to the columns SibSp and Parch. If the passenger does not have any siblings/spouses or parents/children, then NotAlone would be 0, otherwise, it would be 1. I dropped all of the passengers from the dataset who have no age information. There are 177 passengers with no age information, and it is about 0.19865 of the entire dataset. After dropping the missing values, there are 714 data left. I used this 714 data for the later analysis.

## III. RESULTS

### A. Lab Results

To analyze the data better, I classified the data into three groups according to the class of each passenger. I found that there are 186 passengers in class 1, 173 passengers in class 2, and 355 passengers in class 3. Class 3 is the biggest passenger class. After analyzing the mean and median of Age and Fare of each class, I found that passengers in class 1 have the highest mean and median for both Age and Fare, which implies that they might be a group of people who are richer and older than other passengers and probably have certain social status. Passengers in class 3 have the lowest mean and median for Fare and Age. I think these statistics are reasonable because people who can afford the fare for class 1 are typically richer and older than passengers in other classes due to wealth accumulation.

I also created two datasets – titanicdfsurvived and titanicdfnotsurvived – using the Survived feature. By counting the observations in each dataset, we knew that 290 people survived and 424 people did not survive. If we group the survivals in class, we will know that 122 people survived in class 1, 83 people survived in class 2, and 85 people survived in class 3. Similarly, we knew that 64 people did not survive in class 1, 90 people did not survive in class 2, and 270 people did not survive in class 3.

By calculating the correlation between all numerical values possible, I got the following correlation matrix

| | Survived | Age | SibSp | Parch | Fare | NotAlone |
|---|---|---|---|---|---|---|
| Survived | 1.000000 | −0.077221 | −0.017358 | 0.093317 | 0.268189 | 0.196140 |
| Age | −0.077221 | 1.000000 | −0.308247 | −0.189119 | 0.096067 | −0.198270 |
| SibSp | −0.017358 | −0.308247 | 1.000000 | 0.383820 | 0.138329 | 0.629818 |
| Parch | 0.093317 | −0.189119 | 0.383820 | 1.000000 | 0.205119 | 0.577524 |
| Fare | 0.268189 | 0.096067 | 0.138329 | 0.205119 | 1.000000 | 0.260136 |
| NotAlone | 0.196140 | −0.198270 | 0.629818 | 0.577524 | 0.260136 | 1.000000 |

We can see from the matrix that SibSp and NotAlone are relatively close related comparing to other variables because their correlation is the highest among all correlations. This is reasonable because NotAlone is created according to SibSp. Also, if the passenger is with siblings/spouse then he is not alone. Correlations between all numerical values are below 0.6, which means that they all have weak correlations, except the correlation between SibSp and NotAlone. The correlation between Parch and NotAlone is about 0.5775, which is very close to 0.6. It's the second-highest correlation. It's because the passenger who is with parents/children is considered as not alone when created the column NotAlone. There are no strongly correlated values because none of the correlations exceeds 0.8.

The standard deviation for Age is 14.5265. The standard deviation for Fare is 52.9189. The interquartile range for the Age is 17.875, and the interquartile range for the Fare is 25.325. From the box plot of Age, we can see that the data is right-skewed, and the majority of the observations are slightly below the mean. From the box plot of Fare, we can see that the data is right-skewed, and the majority of the observations are below the mean. There is a very extreme outlier, indicating that a passenger paid about 500 dollars for the ticket, which is much higher than the fare of other passengers. Also, there are more outliers in Fare than that in Age. For both box plots, the means are higher than the median.

By calculating the conditional probability that a person survives according to their sex and passenger class, I found

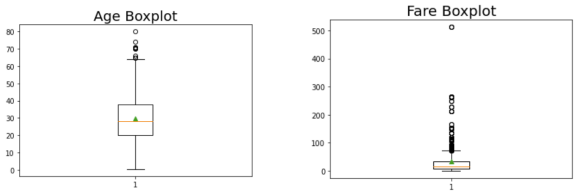


Fig. 1. (a) Age Boxplot (b) Fare Boxplot

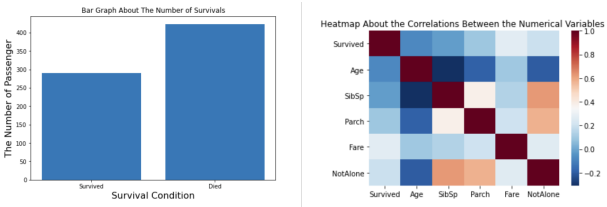


Fig. 2. (a) Bar Graph (b) Heatmap

that the survival rate for

- Female in class 1 = 9647058823529412
- Female in class 2 = 0.918918918918919
- Female in class 3 = 0.46078431372549017
- Male in class 1 = 0.39603960396039606
- Male in class 2 = 0.15151515151515152
- Male in class 3 = 0.15019762845849802

We can see that females in class 1 survived the most. Almost all females in class 1 survived. Males in class 3 survived the least. Only about 15 percent of males in class 3 survived. Overall, females survived more in percentage than males did. Even females in class 3, which is the class with the lowest survival rate, had a higher survival rate than that of the males in the first class. For both males and females, people in class 1 survived the most, and people in class 3 survived the least. I also found that the possibility that a child who is in class 3 is 10 years old or younger survives is 0.4318, which is higher than that of males in all classes, but lower than that of all females in all classes.

The mean fare is 87.96158225806447 for class 1, 21.47155606936416 for class 2, 13.229435211267623 for class 3. The mean fare for class 1 is significantly higher than that of other classes. It might be because of the many outliers that we see in the fare box plot.

From the titanicdata.xslx we know that each passenger has a title of social status/nobility. There are 17 unique titles in the dataset, they are Capt, Col, Don, Dr, Jonkheer, Lady, Major, Master, Miss, Mlle, Mme, Mr, Mrs, Ms, Rev, Sir, and the Countess. When grouping the passengers by titles, the highest survival rate is 1.0, which means all passengers who have the title survived. The lowest survival rate is 0.0, which means all passengers who have the title died. 4 titles have 0 survival rate, 3 titles have 0.5 survival rate, and 6 titles have 1 survival rate. Females have a higher survival rate. For example, all Lady, Mlle, Mme, Ms, and the Countess survived. Also, Miss and Mrs have a survival rate higher than 0.7. On the other hand, most males have a survival rate equal to or lower than 0.5. It seems that the passengers on the Titanic followed the 'women first'.

By drawing a bar graph about the number of survivors in this disaster, I found that more passengers died than survived.

For the heatmap, the darker the blue is, the weaker the correlation is. From the heatmap, we can clearly see that almost all numerical variables have weak correlations except SibSp and NotAlone. Survived, Age, and Fare have weak correlations with all other numerical variables. SibSp and Age have the weakest correlation.

According to Q9, Lady, Mlle, Mme, Ms, Sir, and the Countess all have survival rates of 1.0, which are equally high. So, there are 6 highest survival rates. I drew a bar graph to show the survival rate of the top 5 titles who survived the most. I also drew a bar graph showing the top 5 titles that have the most survivals. They are Dr, Master, Miss, Mr, and Mrs.

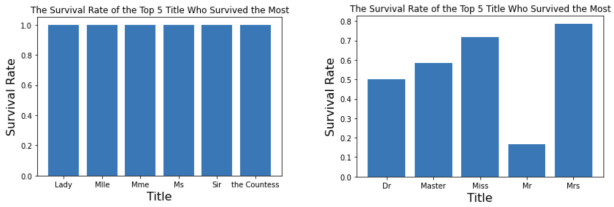


Fig. 3. (a) Top 5 Titles That Have the Highest Survival Rate (b) Top 5 Titles That Have the Most survivals

We can see that even though the title Mr has many survivors, its survival rate is very low. This indicates that there are a lot of passengers with the title Mr, but only a small portion of them survived. This graph also illustrates that females survived more in the disaster. Title Miss and Mrs not only have high survival rates but also have a large number of survivals.

From the histogram about the relationship between ticket fare and survival count, we know that passengers with lower ticket fare survived more. Maybe it's because there are more passengers in the range of low ticket fare. However, we can see that as the ticket fare increases, the survival rate increases. Passengers with high ticket fares have a higher survival rate. It is reasonable to infer that passengers with higher ticket fares are in a better class like class 1, so they have priority to board the lifeboats.

### *B. Casual Inference*

What are some of the potential causes that helped passengers survive the accident?-Do you think "women and children first" was applied in this disaster?-Do you think the correlation analysis that you did to uncover some potential relationships is statistically satisfactory? Do you see any potential room for improvement?

I think one potential cause that helped passengers survived the accident is the ticket fare. I can infer that passengers who paid higher ticket fares lived in a higher class, vice

versa. For example, rooms for class 1 are probably located in the upper part of Titanic, while rooms for class 3 are probably located in the bottom of Titanic. When the accident happened, passengers in class 1 have higher chances to board the lifeboats first and survived.

I think "women and children first" was applied in this disaster because women and children have higher survival rates than men. Even the children in class 3 have a survival rate higher than that of males in all classes. For women, we can see from both bar graphs and survival rates that they have higher survival rates.

The correlation analysis that I did to uncover some potential relationships is not very statistically satisfactory. Excluding the column I created NotAlone, I didn't find any meaningful strongly or medially correlated variables, but I found that all variables are weakly correlated.

There are some places that can be improved in the analysis. For the top 5 titles that have the highest survival rate, I think the results are not very meaningful because there are some titles like the Countess only represents one passenger. Thus, the high survival rate for this title doesn't include much information. We can set restrictions to both the titles and the number of passengers when analyzing the data to improve the results.