Titanic Lab

Dang Pham and Kairuo Yan

Abstract— **More than half of passengers died in the Titanic Sinking at the beginning of the 20th century. In this research, we would like to explore how demographic, social, and other factors affect passengers' survival rates in the Titanic. We conducted the statistical analysis, and we calculated the survival rate for different passenger categories. Our research results are close to our initial assumption, and we found that females and people in the higher class have higher survival rates. However, there are causal inferences problems that exist in the research. Thus, the findings only applied to people in the RMS Titanic rather than the larger population.**

I. INTRODUCTION

Titanic was a passenger liner from Britain, and it sank in the North Atlantic Ocean after colliding with an iceberg in 1912. In this lab, we look at how the demographic, social, and other factors could affect passengers' survival rates in Titanic. In our expectations, people in first class and women and children are more likely to survive. To find the answer, we took five significant steps. In the beginning, we did some data cleaning by eliminating missing values. Then we categorize data into different groups to gain exogenous variables that we are interested in exploring, such as NotAlone, Pclass, Sex, and Age. We performed several statistical tests, including univariate analysis and bivariate analysis. Additionally, we calculated the survival rate by category. In the end, we made graphs that help us to visually understand the existing relationships between exogenous variables and survival rate. As a result, we observed three significant findings. First, females have higher survival rates compared to males. Second, people in the higher passenger class are more likely to survive. Third, people who paid more for the ticket are more likely to survive compared to people who paid less.

II. DATA

For this assignment, we will be using the Titanic Dataset. This dataset provides information on the passengers of the famous cruise ship Titanic. There are 892 rows and 12 columns (Passenger ID, Survived, Passenger Class, Name, Sex, Age, number of Siblings/Spouses, number of Parents/Children, Ticket number, Ticket's fare, Cabin, and Embarked). After doing some data cleaning, we realized that there are 177 rows with missing 'Age' values and this is accounted for 24.79% of the dataset.

III. RESULTS

While setting up our data set, we faced many missing values in the 'Age' column. 177 out of 892 rows were missing, this is about 24.8% of the data set missing. However, the missing data did not affect our project.

Exploring the data set shows us that there are 186 people in the first class, 173 in the second class, and 355 in the third class, which is also the biggest passenger class on the Titanic. We also look at the mean and median age and fare of each passenger class. The first class has the highest mean age and fare, and median age and fare, while the third class has the lowest age and fare, both mean and median.

The Titanic incident was tragic. There were 424 people did not make it alive, while 290 people got lucky enough and make it back to safety.

The Titanic data set has some numeric variables, so we calculate their correlation with each other to see which values correlated. We found out that there is no set of variables has strong correlation. There are only medium correlation between 'NotAlone' and 'SibSp', and 'NotAlone' and 'Parch', they are 0.6298 and 0.5775 respectively. The rest of the correlation table is weak correlation.

The standard deviation for Age is 14.516321150817317, and the standard deviation for Fare is 52.881858444051744. With the median age is 28, and the 1st quartile is 20.25 and the 3rd quartile is 38 (IQR = 17.75). This means that 25% of passengers are younger than 20.25, 25% of passengers between 20.25 and 28, 25% between 28 and 38, and 25% are more than 38 years old. The shape of the data is leaning toward younger passengers with a long right tail. With the mean of 29.699, this means that the majority of observations are below the mean.

We calculated the survival rate for females and males in three different passenger classes. We found that females in the first class have the highest survival rate among females, which is 96.47%, and females in the third class have the lowest survival rate, which is 46.08%. Males have the highest survival rate in the first class and the lowest survival rate in the third class, 39.6% and 15.02%.

We used for-loop functions to subtract a list of unique titles. We also add a new variable Title into our dataset. We calculated the survival rate for people with different titles. As a result, we found that people with Miss titles, Mrs and Mr, have the highest survival rate among all people in Titanic, which are 14.71%, 11.90%, and 9.38%. However, people with titles Don, Rev, Capt, and Jokheer have the lowest survival rate, and they all died.

IV. VISUALIZATIONS

A. *Figure I*

The bar graph shows that people survived in Titanic and people died in Titanic, 290 and 424. More people died than people who survived.

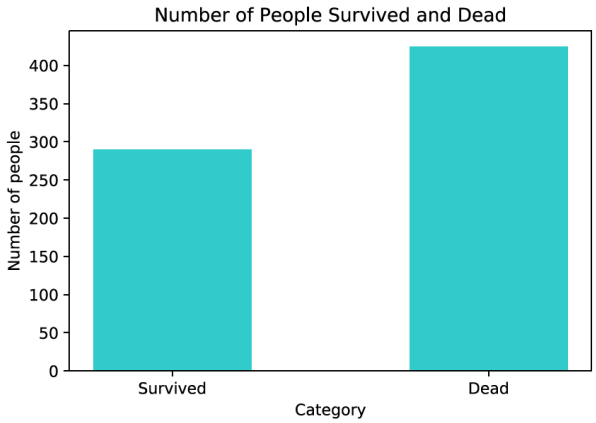

Fig. 1. Number of people survived and dead on the Titanic

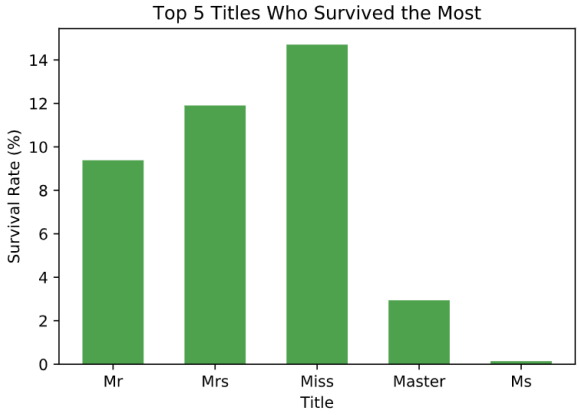

Fig. 3. Top five titles who most likely to survive the Titanic

B. *Figure II*

The heat-map is showing the survival rate in different sex and passenger classes. The higher the level they are, the higher the survival chance for both females and males. Females have a higher survival rate compared to males in general. The survival rate for females in the first class is more than 90%, and females in the third class are around 45%. Whereas the survival rate for males in the first class is about 40%, and for males in the third class, the survival rate is less than 20%.

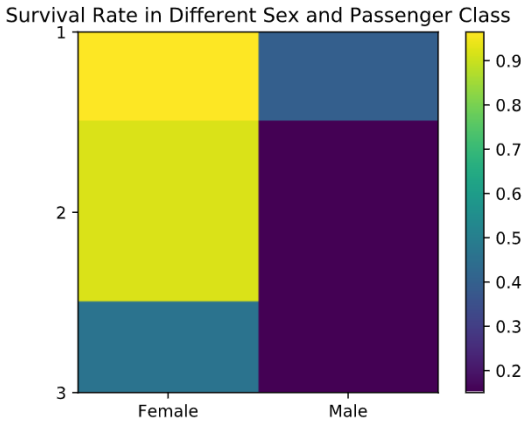

Fig. 2. Survival rate of different sex and passenger class on the Titanic

C. *Figure III*

The bar chart is illustrating the survival rates of different name titles of passengers. The top five titles that have the highest survival rate on the Titanic are "Mr", "Miss", "Mrs", "Master", and "Ms". From the graph, we can see that "Miss" and "Mrs" have the highest survival rate on the Titanic (14.71% and 11.9% respectively). "Mr" stands at the third place with 9.38%, and finally, 'Master" and "Ms" at the fourth and fifth place with 2.94% and 0.14% survival rate.

D. *Figure IV*

From the bar chart, we can see that among survivors, there are more likely he or she is from a lower fare range. About 17.5 percent of survivors paid less than 25 for the Titanic trip. 20-50 and 50-100 group had their survival rate of 9.38% and 8.68% respectively. While the group of passenger paid above 100 for the Titanic trip had lower survival rate, this can be explained by looking that number of people actually paid more than 100 to be on the trip. It is understandable that more people paid less than 25 for the trip so there will be more who survived.

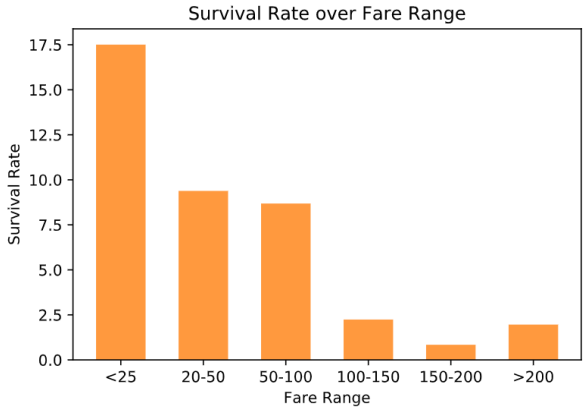

Fig. 4. Relationship of Survival rate versus Fare

V. CAUSAL INFERENCE

A. *What are some of the potential causes that helped passengers survive the accident?*

Four potential causes helped passengers survive. They are passenger class, sex, whether people are alone or not, and fare.

We found that passengers in passenger class1 are more likely to survive for passenger class, while passengers in passenger class 3 are more likely to die. We also found females are more likely to survive, while males are less likely to survive. In addition to that, people who are not alone are

more likely to survive than those alone. We found that people who paid a higher fare are more likely to survive than those who spent less for the fare.

B. Do you think "women and children first" was applied in this disaster?

In this disaster, we do believe that "women and children come first".

We found that the survival rate of female is much higher compares to male. There were 197 females alive and 64 reported dead, while there were only 93 male alive and 360 were dead. This means that about 75.5% of women on board got back to safety, and only 20.5% male could. This shows that being a women during a disaster can boost your chance of survive.

Also, being a child can boost your chance of survival. Our data shows that almost 60% of children on board came back alive, three times higher than the survival rate of male.

C. 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?

From our correlation analysis, there are no potential relationships that are statistically satisfactory. All of my correlations are weak correlations, with only two exception to be medium correlations ('NotAlone' - 'SibSp': 0.6289, and 'NotAlone' - 'Parch': 0.5775). We think that with a larger data set, there will be more room for correlations to happens.

D. Do you see any typical causal inference problems (selection bias, simultaneity, omitted variable bias) in this analysis?

There are selection bias and omitted variable bias problems in this analysis. Due to selection bias, we could not get a clear causal inference from the data set regarding our initial expectation that women and children in the first class have the highest survival rate. Because there are many more males than females, there are many more adults than children; and there are more people in the third class than first and second class. If we would like to explore whether women and children in first class are more likely to survive, we do not have enough randomized samples. Another example is the people with different titles. Most people's titles are Mr, Mrs or Miss. Only very few people have titles like Rev, Capt, or Jonkheer. The survival rate and people with different titles are biased and not a causal inference. Thus, the Titanic sample is not representative of the population.

VI. CONCLUSIONS

Women, children, and people in higher passenger classes have a bigger chance of survival than males and people in lower passenger classes. Some other potential causes also helped passengers gain a higher survival rate. However, there is no strong statistical correlation between factors and survival rates. We need a larger dataset to find statistically satisfied causal relationships and reduce data bias.