Project1 Write-up

Yue Pan

1063965

03/16/2020

Write-up

***Background***: The sinking of the Titanic is one of the most infamous shipwrecks in history and Titanic is one of the most famous American movies. On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew. While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others. So I’d like to find what kind of people are more likely to be survived in the accident.

***Hypothesis***: Buying 1stst ticket class, older women passengers are more likely to survive.

***Data***: My project data is a csv table called train.csv which is from Kaggle, and it is like the ground truth because it recorded all 891 passengers conditions in 1912 Titanic which can be found now. There are also some biases about the data, because 1912 is far from now, there were some data was lost and some NA records. Therefore, it’s likely to cause some errors in the conclusion.

***Analysis Steps***: The first step was to import the train.csv to the python, and then found the survival ratio in the know data(by the pie chart which is easier to see the actual ratio). Then, there were 3 main parts in this project which is about to analyze the age, sex and fare to determine which factor was more likely to survive in Titanic Accident.

The first part is using histogram and boxplot to see which age group is more likely to survive in the data. The second part was using the histogram to analyze men or women had more possibility to survive. The third part was using the histogram to see with the age increasing how the numbers of survival people will changed. Then combining age and ticket class together, and Age and sex together to see the results.

***Summary:***

1) The conclusion is same with the hypothesis that buying 1st ticket class, older women passengers are more likely to survive. The results is expected but there were a lot of observations had been lost, so there might be some bias in the conclusion.

2) Challenge Part: Analyzing the data was pretty challenging and sometimes it was difficult to draw some beautiful charts because the raw data should be changed little or more. Moreover, I was confused with Python and R, so there will be some syntax errors.

***Reflection:*** This was the first time for me to use Python to analyze the data, so I only had limited knowledge to handle the problems. If I can do it again, I will use a long programming code to replace few short programming codes, so it will be easy for people to read the code.