Project1 Proposal

Yue Pan

1063965

02/26/2020

Proposal

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

The first part is using histogram and boxplot to see which age group are more likely to survive in the unknown data. The second part is using the histogram to analyze men or women have more possibility to survive. The third part is 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.

***Potential Roadblocks***: This is the first time for me to use python, so there’re a lot of basic questions for me to analyze the data and make some beautiful charts. Moreover, when I try to analyze them, I am sometimes confused with Python and R, so there are a lot of syntax errors in my code and I have to fix it.