Data Analysis for Titanic Survivors

19085436d Songi Kim
19085595d Yeseo Park
19091827d Junyoung Jung

Table of Content

Motivation: The reason why we should understand the Titanic data (from the perspectives of history, social science, and so forth).

Our team initially chose this topic for the fascination, but the importance of analyzing this historic event appeared afterward. Genuinely, the data from people aboard the Titanic offer numerical evidence that illustrates the people's particular social behavior in that historic event. Since it is difficult to infer whichever intentions people had, to save a particular group of boarding members, analyzing the data helps the process of assuming why few of the majority survived, for example being young, adolescent children or women, or people in upper-class suites. It nonetheless provides clues of how the people of the 17th century tended to behave amongst other people generally, how men were protective of women and why children were valued more. The social ambiance which differs from the ambiance today is attained through understanding the Titanic's data.

Description: Describe the methods (or tools) you used to analyze the features and why they can be helpful. Also, give a brief introduction of Naïve Bayes models (in your own words).

Our team used R language and Rstudio IDE to analyze the given data. Since there were limitations to predict data with the local library, we imported outer function from e1071 and ggplot2, for convenience. This outer library provides a parameter tuning for prediction and splitting data.

Utilizing outer libraries to manipulate data is conveniently such that it reduces the number of codes that require translation to the equation to codes. While converting the massive amount of parameters along with variables and data that are not consistent in data type, the local library only provides basic tools making users formulate all the equations. But, with the external library, it offers pre-made functions

<**Naïve Bayes models>**

Naive Bayes is based on the Bayes theorem assuming that it is conditionally independent. It is roughly calculated with the number of given factors. The equation is shown below. Detailed factors of K are s1,s2,s3, and s4. As factor events occur independently, events that happen simultaneously are the multiplication of each probability.

$$P(C_1|K) = \frac{P(K|C1)P(C1)}{P(K)} = P(C_1|s_1, s_2, s_3, s_4)$$

$$P(s1, s2, s3, s4) = P(s1) \cdot P(s2) \cdot P(s3) \cdot P(s4)$$

The likelihood of the probability that c1 will come out is when each feature is given.

$$\prod_{p,m} P(s_p|C_1) = [P(s_1|C_1) \cdot P(s_2|C_1) \cdot P(s_3|C_1) \dots P(s_p|C_1)]$$

Implementation: How you implement the Naïve Bayes models (describing both variables and functions used in the R scripts)

In this project, we have implemented the Naive Bayes model to calculate the probability based on factors including age, gender, Pclass, fare, embarked place, and parch. First, we designated each factors to "as.factor" like "titanic.full$Age <- as.factor(titanic.full$Age)." Based on the above equation, C1 is survived. As we used the library e1071, it brings naive Bayes directly, "model <- naiveBayes(Survived~., train)." It receives the data, that has been manipulated into the adequate sufficiency, the functions accept the manipulated data and trains based on the algorithm. Then, with the trained dataset, the program compares the existing biases to that of data from "Test.csv" file. This ultimately outputs the accuracy rate of the result. Once the result is created, we made a csv file, which predicts the given data on our specific orders.

```
46  combi_titanic$Pclass <- as.factor(combi_titanic$Pclass)
47  combi_titanic$Embarked <- as.factor(combi_titanic$Embarked)
48  combi_titanic$Age <- as.factor(combi_titanic$Age)
49  combi_titanic$Fare <- as.factor(combi_titanic$Fare)
50  combi_titanic$Title <- as.factor(combi_titanic$Title)
51  combi_titanic$IsMale <- as.factor(combi_titanic$IsMale)
52
```

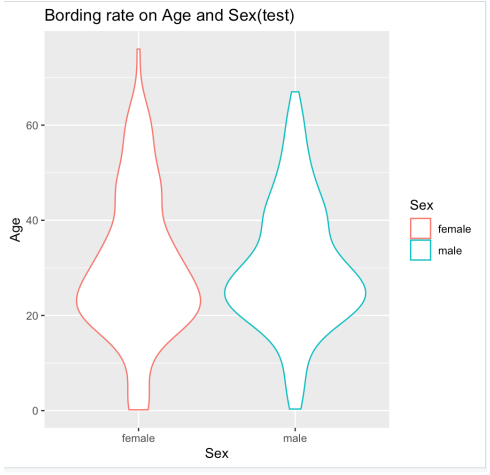
Data: Present the key statistics in the data you are working on (both train.csv and test.csv), such as the average value, range, distributions, etc.

<train.csv>

In the train.csv, Initially, 314 women and 577 males were on board. It was from 0.42 to 80, and the average age was 29.7. The range of the fare was huge, from 0 to 512.33. The median was 14.45, and the mean was 32.20.

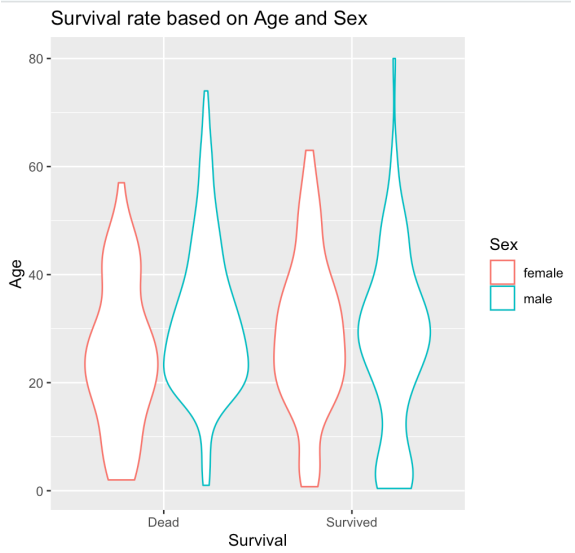
<test.csv>

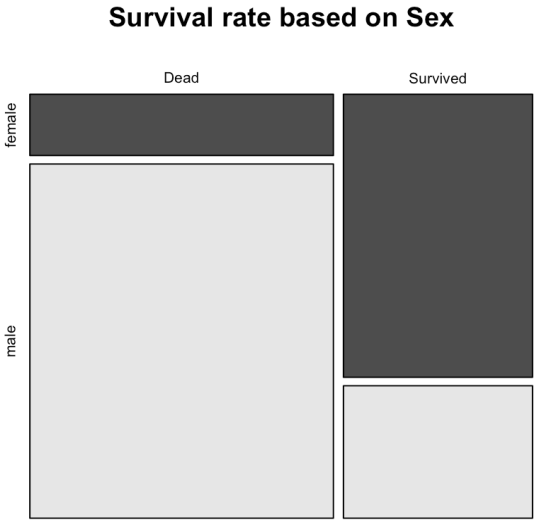
Initially, 152 women and 266 males were on board in the test. It was from 0.17 to 76, and the average age was 30.27. The graph below shows the age distribution by gender for people on the test.csv. Both groups had the most significant number of 20s. With the oldest on the women's side, the age range of women came out more diverse than that of men. The scope of the fare was enormous, from 0 to 512.329. The median was 14.454, but because of the outlier mean was 35.627 while 3rd quartile was 31.5.



Results and observation: Show your analysis results (in figures, tables, or numbers) and list the observations drawn from the results. Here please also provide the screenshot of your team profile and the scores you got from Kaggle.

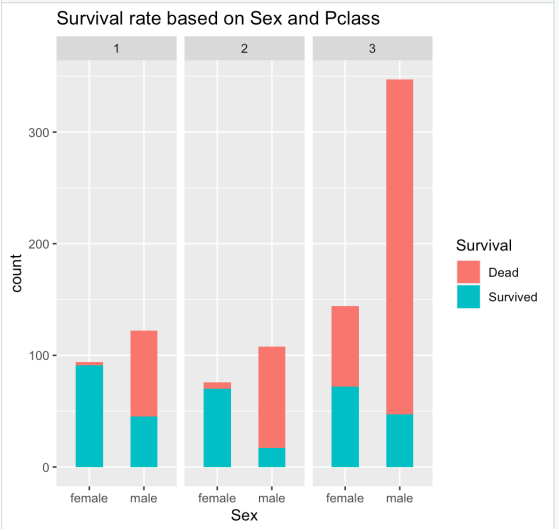
Before analyzing the test, we observed trends in train.csv. Several graphs and images are below.

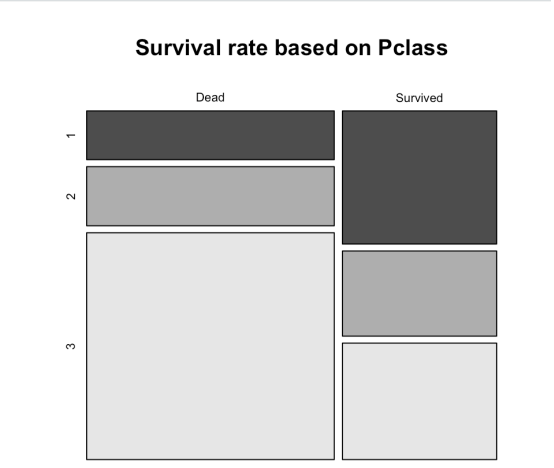




Each table shows the violin and mosaicplot graphs of the result. Both graphs represent the comparison of the survival rate of passengers to gender. From the distribution of the graph on the left, we can see that both men and women had the most in their 20s. The right graph describes, the overall ratio of deaths and survivors for men and women, and the death rate of men was much higher than women.

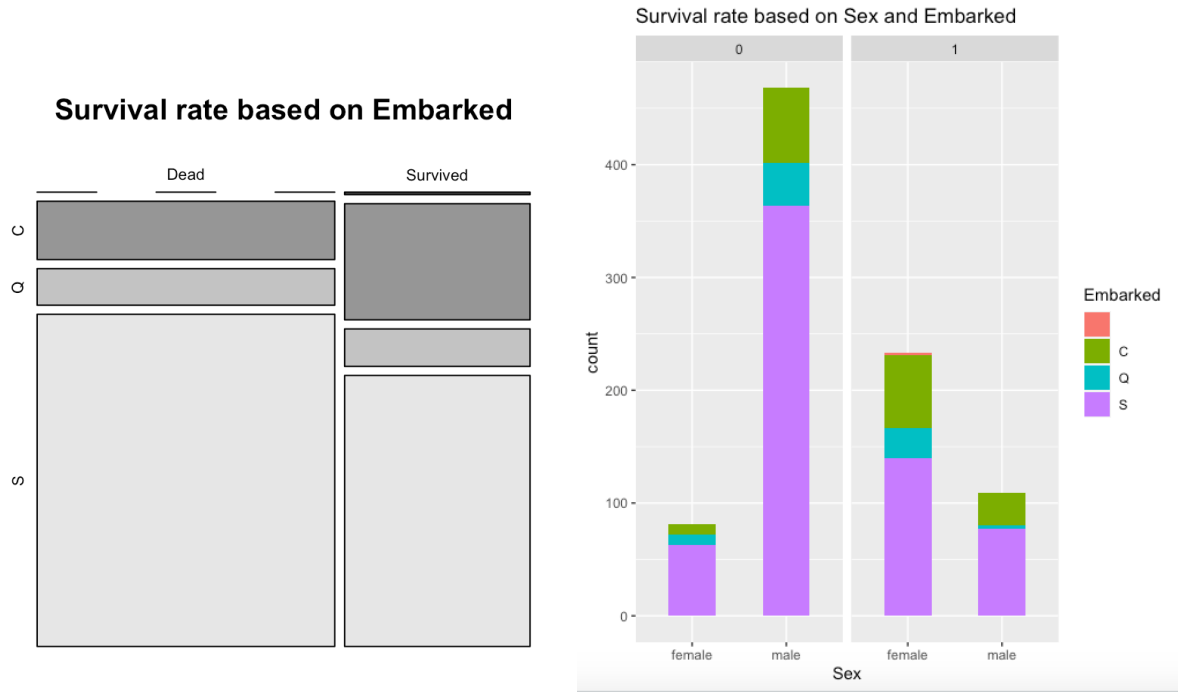
The following graphs specifically divided not only gender but also various factors.





These graphs express the data to the category of "P class." The left bar graph has numerical data while the right one shows the ratio. Regardless of gender, most of the people are third

class. The majority of female passengers in first class and second class survived, however, the survival rate of a second and third-class male passenger was relatively low. It can be deduced that the rate of survival increases proportionally to the higher class.



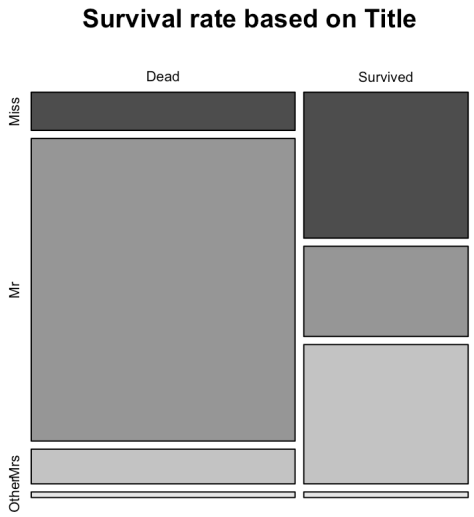
The above two graphs indicated the survival rate, according to the embarked place. The graph on the left was divided for the entire people by embarking place, and the right was divided in detail, according to gender. The left graph illustrates that S port has a higher rate of death compared to Q and C ports and the reason can be assumed through the right graph. We can observe from the right graph that S port has a considerably higher rate of male passengers compared to the other two ports, which is related to the survival rate of S port.

Using the output trend from the train.csv data, our team applied such to the test.csv data. The following data corresponds to the observation made on the research. Similar trends appeared in the test, as analyzed by Train. The graph below shows that the tests did not cover the train.

```
> table(Sex, Survived_p)
        Survived_p
Sex       Dead Survived
  Female    9      143
  Male    244       22
```

```
> table(test$Pclass, Survived_p)
   Survived_p
    Dead Survived
1   38        69
2   61        32
3  154        64
```

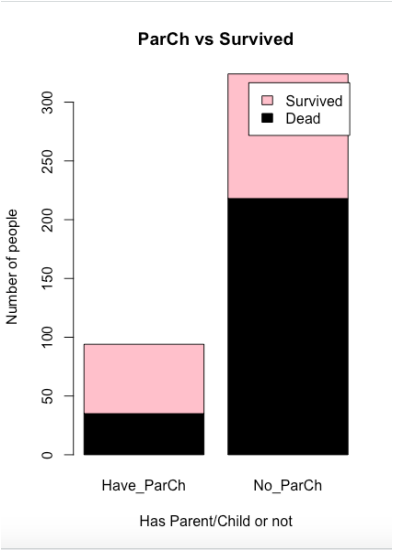
```
> table(test$Embarked, Survived_p)
   Survived_p
    Dead Survived
       0        0
  C   48       54
  Q   22       24
  S  183       87
```




```
combi_titanic$Title <- gsub("^.*, (.*?)\\..*$", "\\1", combi_titanic$Name)
table(combi_titanic$Sex, combi_titanic$Title)
combi_titanic$Title[combi_titanic$Title == 'Mlle' | combi_titanic$Title=='Ms'| combi_titanic$Title == 'Lady'] <- 'Miss'
combi_titanic$Title[combi_titanic$Title == 'Major'| combi_titanic$Title == 'Sir'
                    |combi_titanic$Title == 'Master'|combi_titanic$Title == 'Rev'] <- 'Mr'
combi_titanic$Title[combi_titanic$Title == 'Mme'] <- 'Mrs'
combi_titanic$Title[combi_titanic$Title == 'Capt' |combi_titanic$Title == 'Col' |combi_titanic$Title == 'Don'
                    |combi_titanic$Title == 'Dona'|combi_titanic$Title == 'Dr' |combi_titanic$Title == 'Jonkheer'
                    | combi_titanic$Title =='the Countess'] <- 'Other'

mosaicplot(table(ifelse(train$Survived==1,"Survived","Dead"),train$Title),color=TRUE,main="Survival rate based on Title")
```

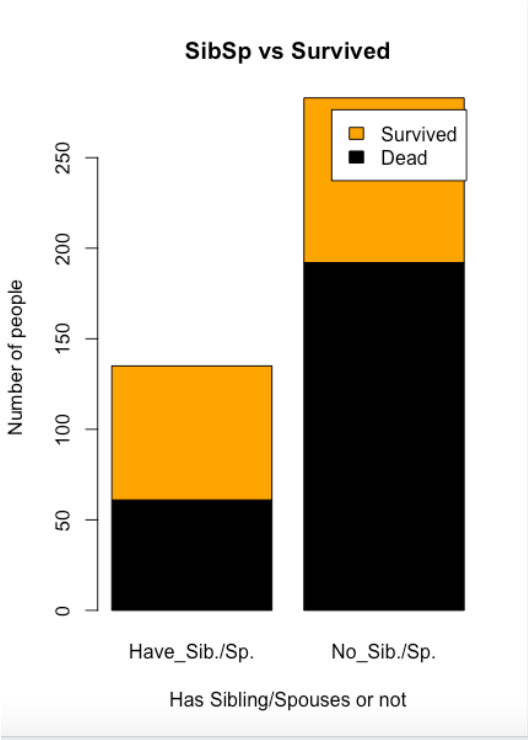
This graph shows the based on the title of their name. To begin with, we separate title to four big categories: Miss, Mr., Mrs, and others. Accounted for the most significant portion of deaths was Mr, followed by Miss and Mrs. On the contrary, Miss and Mrs were the most alive among the survivors, and the number Mr survivors were small. The survival rate comparison for women and men in the preceding train.csv can be compared with the title here.



	Survived_Par	
Parch	Dead	Survived
Have_ParCh	35	59
No_ParCh	218	106

In terms of the number of survivors, people without parents or children (parch) survived 47 people more than those who have a parch. However, in terms of the proportions, the

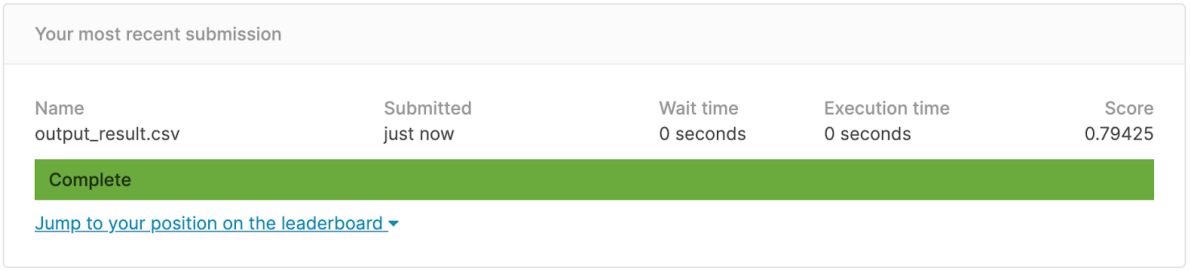
probability of survival is higher with a parch. If there is a parch, it lives a 62.8% chance, while 32.7% without.




```
> table(SibSp,Survived_Sp)
                Survived_Sp
SibSp            Dead Survived
  Have_Sib./Sp.    61       74
  No_Sib./Sp.     192       91
```

The number of survivors in each category was similar, 74 people who have siblings or spouses(SibSp) and 91 without Sibsp . Similar to the parch, the Sibsp group had a higher chance of survival in terms of proportion.
Therefore, the Parch and Sibsp graphs explained that chances of survival increase with family than singles because children and women were escaped first.

<Summary>
The survival rate is influenced by a person's gender, age, Pclass, and ticket fare. So, the data are used to allude that generally healthy full-grown male adults sacrifice themselves for the lives of women, children, and the elderly in times of disaster.



Discussion: Your thoughts and opinions after analyzing the data (can be from different perspectives of history, social science, etc.).

After analyzing the data, we observed that as age decreases, the person is more likely to survive. Also, if the person's gender female, the survival rate goes up. As we delve into the results and debated through, we approached at an interesting conclusion. In our perspective of the particular behavior of men saving women and children before others, it could be deduced that the western women's social position in that historical period was much higher than that of eastern women. In 1912, when the Titanic sank, most of the eastern countries possessed a relatively conservative mindset. Taking Korea, for example, in the early 20th century, Korea was still in a dynasty where there were social classes and Kings. Amongst many prejudices, the majority of people received, gender inequality was a serious issue. In an extreme case, women did not have free will to make a free opinion on any words from men. For this much social pressure, if there was a similar situation in Korea, where a ship full of people of all class, age, and gender sank, it is obvious that men would have a much higher survival rate. Moreover, the concrete eastern ideal of "Confucian ideal" indicates that all younger people must respect the elderly since all previous generations have raised and developed what younger people nourish now. Thus the elderly must be a priority over children. Thus, another difference to the observed conclusion from our team's experiment to the eastern culture is that the survival rate of the elderly would be much higher than children. The survival rate would be the direct opposition to each other. The fascinating idea was aroused due to the immense amount of care was given to women and children despite the dominant male society in that period.

Individual Contribution

o Songi Kim: Took over the comprehensive overview in macroscopical perspective. Focused mainly on implementing the concept of Naïve Bayes model into codes. Also contributed in studying the necessary and appropriate external libraries to minimize tedious works.
o Yeseo Park: Focused on studying the Naïve Bayes model. Manipulated data structures to most accurately predict the data. Similarly, contributed in making the code more efficient and reducing the hard works. Finally, designed the graphs to more concisely express all the output at one view.
o Junyoung Jung: Composed all the data and analyzed the output both numerically and qualitatively. Mainly coordinated the codes to work accurately. Supported other members into understanding the Naïve Bayes model while suggesting multiple possible ways of proceeding with the experiment. Suggested the correlation of the output to the social ambience, and related that to the historical background of eastern countries. Organized the works into documentation.