# COMP 1433 Final Project Report

## Motivation

The Titanic incident refers to a giant and luxury cruise ship: RMS Titanic which under the jurisdiction of the British White Star Shipping Company in 1912, sank on April 15 after a collision with a huge iceberg. Ghandour and Abdalla (2017) stated that 1,502 of the 2,224 crew and passengers were killed, the incident was known as the worst peacetime shipwreck in modern history afterward.

In our project, we focus on related data list provided by Kaggle Titanic competition, which includes 10 attributes: the passenger's name, Pclass, passengerId, gender, SibSp, Parch, Ticket, Fare, cabin, Embarked of passengers. SibSp represents the number of siblings / spouses who board together, and Parch represents or the number of children or parents who board together, Ticket is the ticket number, Fare is the boarding fee, while Pclass is divided into 3 levels (first/second/third class), Cabin is the specific cabin room number. Embarked indicates the ports of embarkation (Q for Queenstown, S for Southampton, C for Cherbourg).

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | PassengerId | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
| 2 | 892 | 3 | Kelly, Mr. James | male | 34.5 | 0 | 0 | 330911 | 7.8292 | | Q |
| 3 | 893 | 3 | Wilkes, Mrs. James (Ellen Needs) | female | 47 | 1 | 0 | 363272 | 7 | | S |
| 4 | 894 | 2 | Myles, Mr. Thomas Francis | male | 62 | 0 | 0 | 240276 | 9.6875 | | Q |
| 5 | 895 | 3 | Wirz, Mr. Albert | male | 27 | 0 | 0 | 315154 | 8.6625 | | S |
| 6 | 896 | 3 | Hirvonen, Mrs. Alexander (Helga E Lindqvist) | female | 22 | 1 | 1 | 3101298 | 12.2875 | | S |
| 7 | 897 | 3 | Svensson, Mr. Johan Cervin | male | 14 | 0 | 0 | 7538 | 9.225 | | S |

In reality, identity of a passenger attached with a name represents his/her social status. A study found that celebrities or the rich who are in a high status probably will possess considerable wealth to purchase a high-priced ferry ticket, and will more likely to survive(Kakde & Agrawal, 2014) since plenty of lifeboats were close to the first-class cabin(Frey, Savage & Torgler, 2009). Worse still, most of the steamship lifeboats are stored in first and second class cabins. And poor people and workers cannot afford high fare are concentrated in third-class cabins, which are closer to cracks that leaked after hitting icebergs, thus they are less likely to survive from seepage of cold water from the crack formed by the crash. Fare, identity, and cabin are key elements revealing the huge gap between rich and poor, which indicated that class differences profoundly affect survival rates in Titanic. In the last hour of sinking, under the forcible command of captain and deputy, women and children of each cabin can be rescued first via getting into the lifeboat (Ghandour & Abdalla, 2017). Resulting in survivors were concentrated in the group of children and women. Hence, age and gender are two key factors affecting the survival rate of passengers as well.

In preliminary survival rate analysis of Titanic data, which pointed to passenger information table, (train.csv) it seems that compared with Gender, Identity, and Fare, the SibSp, Parch, Embarked didn't directly have a significant impact on the survival rate of each individual on board. But these data are interrelated, they will form different data set, each passenger or crew is mapped to one or even a few data sets, so in the process of deduction from the analysis on overall survival rate to individuals', analyzing a single data feature isn't adequate to accurately

predict the survival rate of a specific passenger, whose survival probability may be affected by other factors concurrently. Thus the preliminary text-form analysis is vague, before constructing a data model to predict the survival rate, we are required to accurately explore the specific gap of survival rate revealed by the same type of data, and judge whether each data type has a strong or weak correlation of another data.

Furthermore, analysis of the Titanic data can give latter shipbuilders and designers insight to reasonably adjust cruise ship cabin settings, verify the number of crew loads ship can bear. The survival rate curve in the context of various data obtained through analysis and processing can allow insurance companies to efficiently estimate the survival rate and risk of a passenger, and promote related insurance products for both passengers and shipping companies, protecting the rights and interests of passengers, especially for spouses and persons who own children or parents. Additionally, it is helpful for historians and sociology experts to regard the Titanic data as a model to indirectly analyze the moral level and the gap between the rich and the poor in the early 20th century of the western world. The group of crew and passengers (2,224) was used as a dramatic sample to assess whether they comply with gentility and chivalry - the moral principle of giving up priority to vulnerable groups. It will assist shipping companies to learn from the disaster and come up with a set of exercise plans with the highest survival rate and the lowest loss as well as compensation to avoid more shipwrecks.

## Description

In this question, we prefer to use the R language to solve the problems. As we all know, the R language is a programming language that is widely used among mathematics. It has tremendous useful built-in functions that we can use for Titanic observation. By using the “NaiveBayes” function or “cforest” function in the random forest, we can get different types of prediction. As mentioned in the implementation part, the “NaiveBayes” function is less sensitive to missing data, aims to connect survived with other classification. By using such a function, we can build a model that can forest results. As to make it more visualized, we can use “ggplot2” function in our analysis. It will display the relationship between Pclass and survival rate, gender, and survival rate as well, and some other facts that will affect the survival rate. In the histogram, you will clearly view the results. It can also generate a broken line part. For this program, we choose different types of graphs in different situations. In the ggplot function, we may also add titles and put different databases to x and set y as the survived result.

Naive Bayes models mean to judge whether the thing is C according to plenty of other factors like $F_1$, $F_2$, ……， $F_n$. There is a probability that if it meets with the specification, it will be C. The obvious advantage of applying it is for convenience, which doesn’t need numerous databases. Besides, Naive Bayes models have high accuracy on the database. At the same time, its classification accuracy does not rely on the dependencies. (Domingos & Pazanni 1997).

Naive Bayes models mean to judge whether the thing is C according to a great deal of other factors like $F_1$, $F_2$, ……， $F_n$. There is a probability that if it meets with the specification,

it will be C. The obvious advantage of applying it is for convenience, which doesn't need numerous databases. Besides, Naive Bayes models have high accuracy on the database. At the same time, its classification accuracy does not rely on the dependencies. (Domingos & Pazanni 1997).

## Implementation

The idea of naive Bayes:
For the given item to be classified, from probability of each category under the condition of the occurrence of this item, the largest one is considered to be the category item belongs to.

The formal definition of naive Bayesian classification is as follows:

- There is an item to be classified $x = \{a_1, a_2, \ldots, a_m\}$, and each a is a characteristic attribute of x
- Category set $C = \{y_1, y_2, \ldots, y_n\}$
- Calculate $p(y_1|x), p(y_2|x), \ldots, p(y_n|x)$
- If $p(y_i|x)=\max\{p(y_1|x), p(y_2|x), \ldots, p(y_n|x)\}$, then x belongs to i

From the above definition, we can know which type X belongs to as long as we calculate the conditional probabilities in step 3.

$$p(yi|x)=p(x|yi)p(yi)/p(x)$$

```
library(e1071)                  #naivebayes function
model <- naiveBayes(Survived ~ Pclass + Sex + Age2 + SibSp + Parch + Embarked + Fare +
Title + FamilySize, newtrain ,na.action = na.pass)
prediction <- predict(model,newtest)
result <- data.frame(newtest$PassengerId,prediction[1])
names(result) <- c("PassengerId","Survived")
write.csv(result,file = "D:/ZXY.csv",row.names = FALSE)
```

Implementing NaiveBayes model has used the "naiveBayes" function in "e1071" with variables(Pclass, Sex, Age2, SibSp, Parch, Embarked, Fare, Title, FamilySize).
In the variables, the missing values in Fare, Embarked, Pclass and Age are filled.

## Data

We tried to calculate the following statistics for the initial analysis:

| Data Names | Average | Range | Correlation coefficient (Pearson, Kendall, Spearman) |
|---|---|---|---|
| **PassengerID** | - | - | - |
| **Survived** | - | (0, 1) | - |
| **Pclass** | - | (1, 2, 3) | -0.36, -0.34, -0.36 |

| | | | |
|---|---|---|---|
| **Name** | - | - | - |
| **Sex** | - | (male, female) | -0.54, -0.54, -0.54 |
| **Age** | 29.88114 | (0.17, 80) | -0.08, -0.04, -0.05 |
| **SibSp** | 0.4988541 | (0, 8) | -0.02, 0.07, 0.07 |
| **Parch** | 0.3850267 | (0, 9) | 0.09, 0.15, 0.16 |
| **Ticket** | - | - | - |
| **Fare** | 33.29548 | (0, 512.3292) | 0.27, 0.28, 0.34 |
| **Cabin** | - | - | 0.27, 0.30, 0.32 |
| **Embarked** | - | - | -0.19, -0.17, -0.18 |

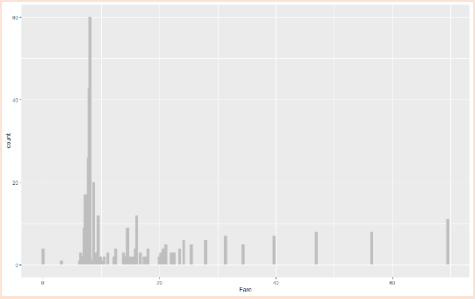
The following are the steps we take to process the data.
(The data and graphs in Data part bellow got from R script Format 2)
Process the data in R preparing for prediction.
Check the missing data

```
PassengerId    Survived      Pclass        Name         Sex         Age
          0         418           0           0           0         263
      SibSp       Parch      Ticket        Fare       Cabin    Embarked
          0           0           0           1        1014           2
```

We filled the missing data from small to large (for "Fare" item we only supplement the missing one).

**1.Fill Fare missing one**

Check the information of the passenger.

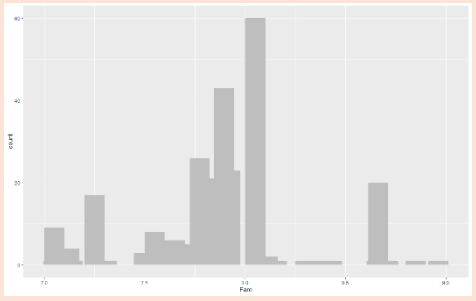


```
> combination[combination$PassengerId==1044,]
     PassengerId Survived Pclass                Name  Sex  Age SibSp Parch Ticket Fare Cabin Embarked
1044        1044     <NA>      3 Storey, Mr. Thomas male 60.5     0     0   3701    1  <NA>        S
```

We got that the man was in Pclass 3 and Embarked at S. Therefore, we picked out all passenger in Pclass3 and Embarked

```
> Pclass3_EnbarS<-Pclass3_EnbarS[which(Pclass3_EnbarS$Fare<9 & Pclass3_EnbarS$Fare>7),]
> Pclass3_EnbarS<-data.frame(Fare=Pclass3_EnbarS)
> pic_Pclass3_EnbarS2<- ggplot(Pclass3_EnbarS,aes(x=Fare)) + geom_bar(stat="count",width=0.1, fill='gray')
```

at S.It is obvious that the most frequent value of Fare is between 7 and 9. We narrowed this picture.

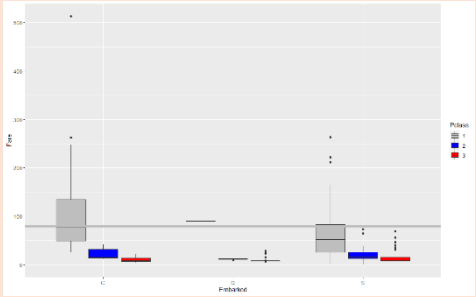


We can ensure the modal number is very close to 8.0, so we assign the missing Fare as 8.0. So 8.0 is used for fill this missing Fare.

**2.Fill Embarked missing two**

```
> print(combination[combination$PassengerId==830,])
     PassengerId Survived Pclass                                                 Name    Sex Age SibSp Parch Ticket Fare Cabin
830          830        1      1 Stone, Mrs. George Nelson (Martha Evelyn) female  62     0     0 113572   80   B28
    Embarked
830     <NA>
```

```
> print(combination[combination$PassengerId==62,])
    PassengerId Survived Pclass                Name    Sex Age SibSp Parch Ticket Fare Cabin Embarked
62           62        1      1 Icard, Miss. Amelie female  38     0     0 113572   80   B28     <NA>
```

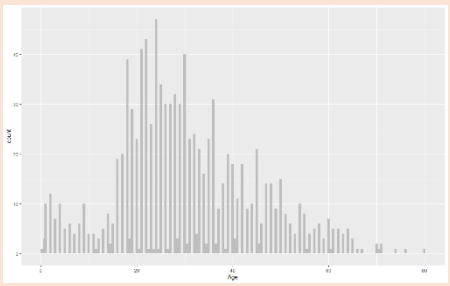


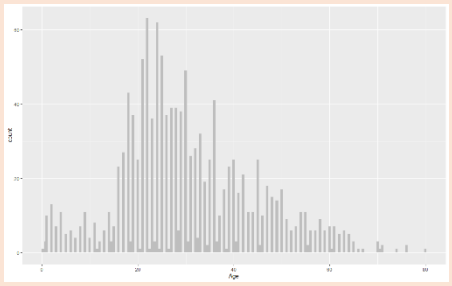
The embarked row has two missing values. After getting the two row's data, we find that Pclass are both 1 and Fare are both 80. We tried a variety of data and distribution maps, finally using boxplot and three variables to find the most likely case. From this graph, we can get that we can get that most Pclass 1 people trend to embarked at C, so we fill them as C.

**3.Fill Age missing 263**

```
> pic_Pclass_Fare<-ggplot(Pclass_Fare, mapping=aes(x=Embarked, y=Fare, fill=Pclass)) +geom_boxplot() +
+     geom_hline(aes(yintercept=80), colour='gray', lwd=2)+scale_fill_manual(values=c('gray','blue','red'))
> pic_Pclass_Fare
> Have_age<-combination[which(!is.na(combination$Age)),]
> no_need <- c('PassengerId','Ticket','Cabin','Survived')
> result<- mice(combination[, !names(combination) %in% no_need], method="rf")
```

Lack of age than the Fare and Embarked tend to be Missing Completely at Random, because not only from the lack of a lot of people, we almost can't find the rule. But also, in fact, Age can cause deficits for a variety of reasons. Therefore, we used random forest to make the age distribution of the missing person consistent with the age distribution of the non-missing person. Get people have age info.





On the left is the known age distribution, and on the right is the predicted age distribution. The two are very close, indicating that the predicted results are acceptable

**4. As for the missing Cabin**

There are 1014 missing Cabin, which accounts 1014/1309=77%, so we didn't think about it.

## Results and Observations

(The data and graphs in this part bellow got from R script Format 1)

```
> names(train)      #names
 [1] "PassengerId" "Survived"    "Pclass"      "Name"        "Sex"
 [6] "Age"         "SibSp"       "Parch"       "Ticket"      "Fare"
[11] "Cabin"       "Embarked"
> sapply(data,function(x) sum(is.na(x)))    #the missing data
PassengerId    Survived      Pclass        Name         Sex         Age
          0         418           0           0           0         263
      SibSp       Parch      Ticket        Fare       Cabin    Embarked
          0           0           0           1        1014           2
> str(data)
tibble [1,309 x 12] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ PassengerId: num [1:1309] 1 2 3 4 5 6 7 8 9 10 ...
 $ Survived   : num [1:1309] 0 1 1 1 0 0 0 0 1 1 ...
 $ Pclass     : num [1:1309] 3 1 3 1 3 3 1 3 3 2 ...
 $ Name       : chr [1:1309] "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley
 (Florence Briggs Thayer)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath
 (Lily May Peel)" ...
 $ Sex        : chr [1:1309] "male" "female" "female" "female" ...
 $ Age        : num [1:1309] 22 38 26 35 35 NA 54 2 27 14 ...
 $ SibSp      : num [1:1309] 1 1 0 1 0 0 0 3 0 1 ...
 $ Parch      : num [1:1309] 0 0 0 0 0 0 0 1 2 0 ...
 $ Ticket     : chr [1:1309] "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
 $ Fare       : num [1:1309] 7.25 71.28 7.92 53.1 8.05 ...
 $ Cabin      : chr [1:1309] NA "C85" NA "C123" ...
 $ Embarked   : chr [1:1309] "S" "C" "S" "S" ...
```

The names of the variables

Using str(data) to see the data types

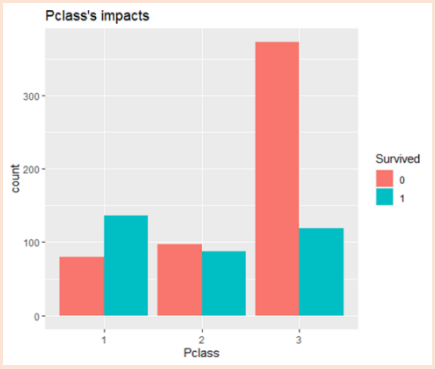
Then convert the Survived, Pclass, Sex, Embarked data to factor type.

Looking for the missing data and fill them.

We have described how we fill the missing values in different lines.

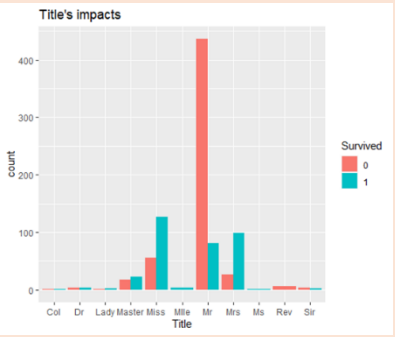
Then analyze the data and use "ggplot2" to draw graphs.

**1. Impacts of Pclass**




In the Pclass graph, people who did not survive are mainly from the third-class cabin and survived people are relatively more from the first-class cabin. Also, only the number of first-class reveals more survived than dead.

**2. Impacts of Title**



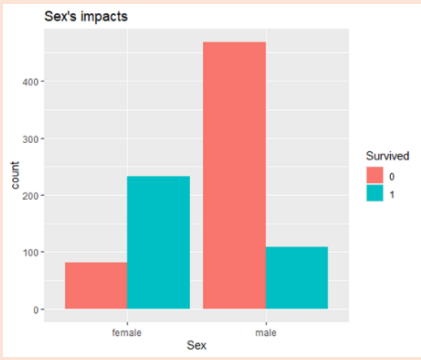

In the Title graph, people whose title is “Mr”, which only the male used, have extremely high counting of dead. “Miss” and “Mrs”, which are used to describe female, have higher survived number than dead. It reveals that female is more likely to survive than a man.

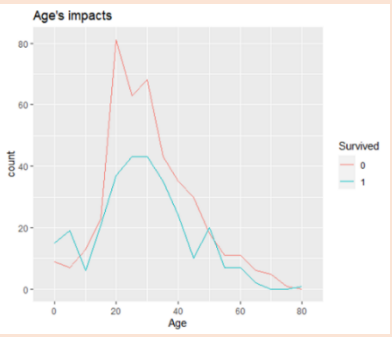
```
> table(data$Title)

    Capt      Col      Don     Dona       Dr 
       1        4        1        1        8 
Jonkheer    Lady    Major   Master     Miss 
       1        1        2       61      260 
    Mlle      Mme       Mr      Mrs       Ms 
       2        1      757      197        2 
     Rev      Sir the Countess                
       8        1        1                  
```

### 3. Impacts of Sex

In the graph of Sex, the female has more survived number than dead, male has more dead number than survived.



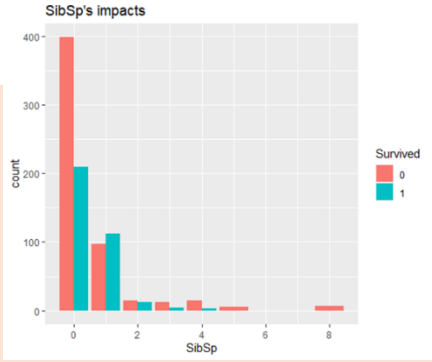

### 4. Impacts of Age

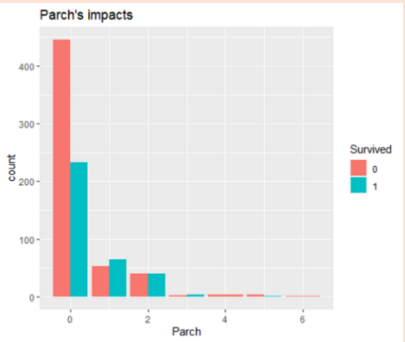
In the graph of Age, people who are under 10 years old are more likely to survive than people of other ages. It is suggested that children have a considerably high rate to survival. For future predictions, we divided the ages into the following four groups, and covert the format to factor.

```
library(rpart)
Age.model <- rpart(Age~Pclass+Sex+SibSp+Parch+Fare+Embarked+Title+FamilySize,data = data[!is.na(data$Age),]
data$Age[is.na(data$Age)] <- predict(Age.model,data[is.na(data$Age),])
data$Age2 <- "15-"
data$Age2[data$Age >= 15 & data$Age < 30 ] <- "15-30"
data$Age2[data$Age >= 30 & data$Age < 45 ] <- "30-45"
data$Age2[data$Age >= 45 & data$Age < 60 ] <- "45-60"
data$Age2[data$Age >= 60 ] <- "60+"
```

### 5. Impacts of SibSp

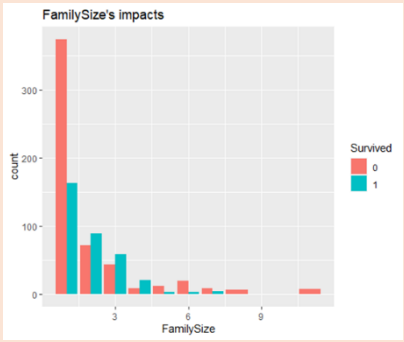
In the Siblings/Spouse graph, people have no sibling or spouse are more likely to fail to survive. People have one sibling or spouse have more survived number than dead. For two or more siblings/spouse, the number of survived are all less than dead.



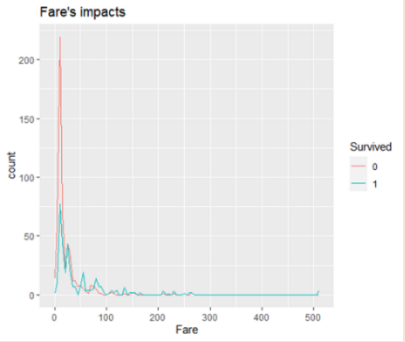

### 6. Impacts of Parch

In the Parents/Children graph, for people who came with no parent or children, they can rarely survive. For people who have one to three parents/children, the number of survived is larger than dead. For people with four or more parents/children, they are less likely to survive.

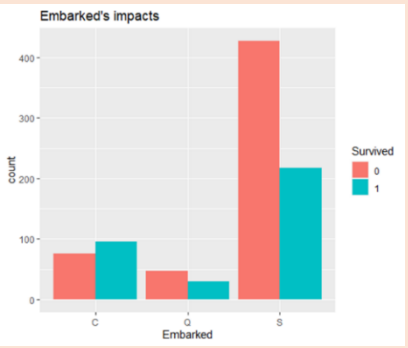



### 7. Impacts of FamilySize

In comparison, we have a graph of FamilySize, which shows that people who came alone or came with four or more family members are probably cannot survive. People came with one to three family members have more opportunities to survive.




### 8. Impacts of Fare



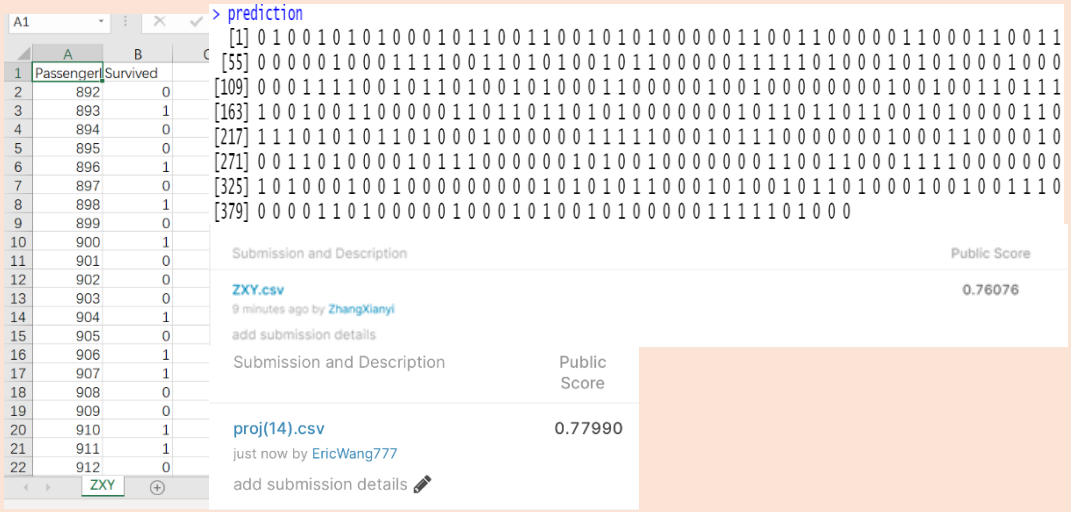

In the Fare graph, people with the lower fare, less than $50, are highly possible not to survive. To compare these two graphs, it proves that people paid less on fare are likely the people who purchased a third-class cabin and they are easier to fail to survive. (Fare in R script format 2 in distinguish to 2 groups, one is less than or equal to 80 and another is more than 80)

### 9. Impacts of Embarked

In the Embarked graph, Embarked = C has the highest rate of survival, Embarked = S has a high rate of death.

Finally, using the "NaiveBayes" function to predict. We got the result as follow:

The scores are **0.76076** on Kaggle on R script format 1 (Left) and **0.77990** on format 2 (Right).




## Discussions

Based on the survival rate of Titanic, our reflection stands on perspectives of Culture and Social science. From a cultural perspective, different sizes of families own various survival rates, it's probably due to tradition and attic faith that family members should stay and go through a hard period together. Moreover, the impact of Pclass reveals that passengers in the first-class cabin have a higher possibility to survive numerically. Such an observative revelation could be proved by the impacts of Fare as well. In our opinion, such a phenomenon discloses survival rate is positively correlated with status and individual wealth.

On point of view of social science, "Saving female and children first" is enforced until the last sinking moment. For human beings, the continuation of life is a primitive instinct. Kids are vulnerable but regard as the future of a country as well. Adults are willing to offer their survival chance to the kids, they can raise the sympathy of other people, and then they will give the survival chance to them. Additionally, young children lack the living ability and need breastfeeding, which means women are needed to leave over to look after them. Hence, a group of children and females are saved first for race continues.

Partially missing data like Age and Cabin in csv files make data processing complex to deal with, however, implementation of Naïve Bayes can balance error of unbalanced data set, maintain a degree of accuracy in Titanic data under the circumstance of losing part of features. However, we encounter some inevitable issues when applying it, such as error rate in classification decisions, observation effect is not optimal due to sampling attributes are interrelated (Given that assumption of the independence of sample attributes). Our estimate model will be optimized If Kaggle provides intact and open-sourced data sets. And still, Ticket owns data integrity and how Ticket difference affect one's survival rate is worth to be further explored and discussed, due to such issue is simply ignored by the ordinary observer but has research value. One of the most disadvantages of the analysis is that we have ignored the effect of the cabin. Although it affects the rate of the dead, we still ignore it because of its large amount of missing.

## Teamwork.

| Student Name | Student ID | Work Allocation |
|---|---|---|
| **YANG Jun Peng** | 19107412D | Write the motivation and part of discussions. |
| **WANG Yi Feng** | 19087103D | Write the description and part of discussions. |
| **PAN Ze Wen** | 19083967D | Write the data part and analyze data, assist to write the second draft of R script. |
| **WANG Xiang Zhi** | 19082878D | Write the second draft of R script(Format 2), supply data and graphs, data part. |
| **ZHANG Xian Yi** | 19078513D | Write the first draft of R script, give the data and graphs, and write the data, implementation and results part. |
| **ZHANG Zi Xuan** | 19068277D | Find references. Write part of results & observation and discussion of report. |
| | | |

(In the order of the parts of the report that each person is responsible for.)

## Reference


Dina, A.M.G., & May, A.M.A. (2017). THE TRAGEDY OF TITANIC: A LOGISTIC REGRESSION ANALYSIS. INTERNATIONAL JOURNAL OF ADVANCED RESEARCH, 5(6),1454-1465.DOI: 10.21474/IJAR01/4558

Kakde, Y. & Agrawal, S. (2018). Predicting Survival on Titanic by Applying Exploratory Data Analytics and Machine Learning Techniques. International Journal of Computer Applications. 179. 32-38. 10.5120/ijca2018917094.

Domingos, P., and Pazzani, M. (1997). Beyond independence: Conditions for the optimality of the simple Bayesian classifier. Machine Learning 29:103–130.