Who can survive on Titanic?

Linfeng Zhou, Tianyi Gu, Xiaoge Wu, Yi Zhang

November 13, 2015

1 Introduction

The sinking of Titanic is one of the most horrific events in the 20th century. Even though there are numerous people died on that ship, some of particular group were survived in this tragedy. Our goal is to find which attributes may effect surviving rate in this event and predict whether a given passenger can survive or not in the sinking of the Titanic.

2 Getting and Wrangling Data

The data available for this problem is from Kaggle competition Titanic: Machine Learning from Disaster. The source of the dataset is available from their data page[1].There are two main datasets for this problem. One is our training dataset, another is testing dataset.

The training dataset has 891 passengers, and the testing dataset consists of data for 418 samples. To be more precise, each observation from both datasets contains 11 variables to provide detail of this passenger(variable name and meanings are in the following table). Of course, the training dataset also provides the label whether this observation was dead or survived after Titanic tragedy.

Table 1: Meanings of Each Variable

Variable Name	Data Structure	Detail
PassengerId	Int	PassengerId Number
Survived	Int	Survival(0 = No; 1 = Yes)

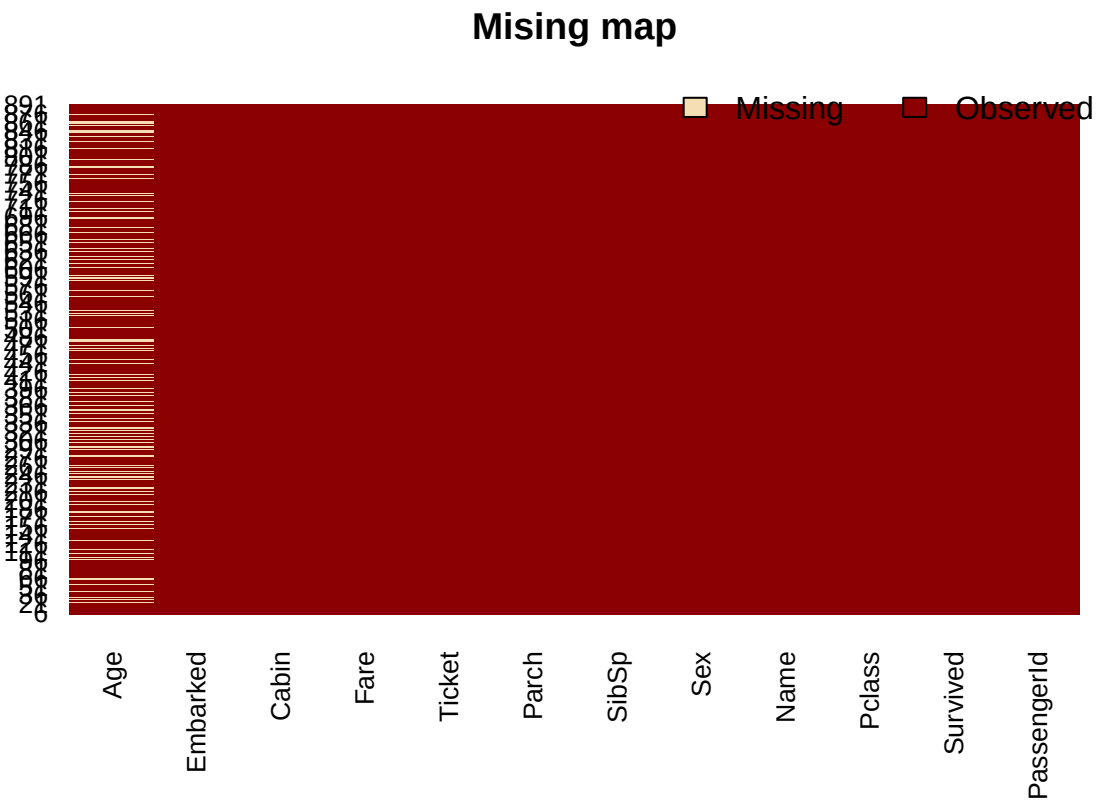
[1] https://www.kaggle.com/c/titanic/data

Variable Name	Data Structure	Detail
PClass	Int	Passenger Class(1 = 1st; 2 = 2nd; 3 = 3rd)
Name	Character	Name
Sex	Character	Sex
Age	Int	Age
Sibsp	Int	Number of Siblings/Spouses Aboard
Parch	Int	Number of Parents/Children Aboard
Ticket	Character	Ticket Number
Fare	Number	Passenger Fare
Cabin	Character	Cabin
Embarked	Character	Port of Embarkation

In order to map data into a convenient format for training model, data mugging is inevitable. As the Table 1 showed, types of Survived and Pclass variables are both integer right now. However, the table also provided the essential information that both variables are categorical data. Therefore, the first step of data cleaning is converting those variables into categorical data(It is called factor data in R).

There are several passenger samples with missing value in some of the variables. Those missing data might distort inference about the population. In fact, some machine learning techniques cannot handle missing value. Therefore, filling unknown data should be done before generating any statistics or training model. Using missmap fuction from package Amelia[2], the figure showed that age variable has lots of unknown data.

[2]James Honaker, Gary King, Matthew Blackwell (2011). Amelia II: A Program for Missing Data. Journal of Statistical Software, 45(7), 1-47.

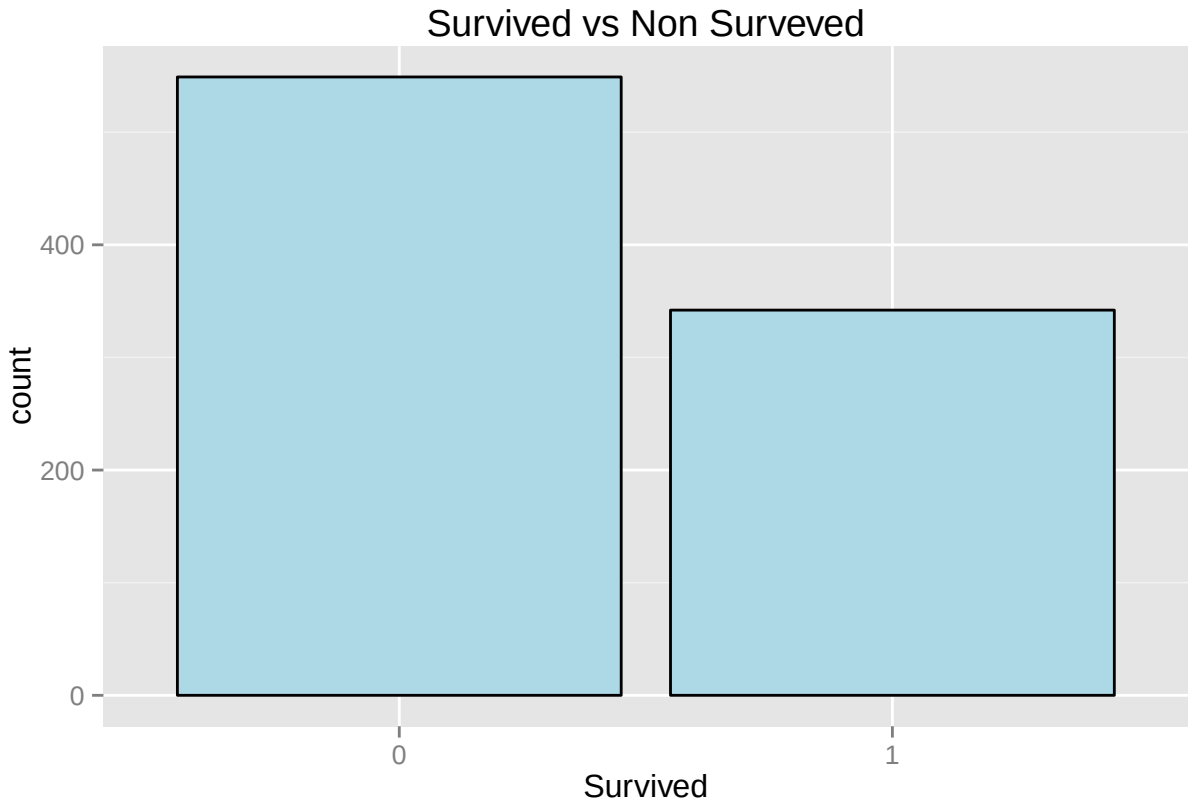


One approach is to use similarity between observations to fill those missing values. This main idea came from a classification method called k nearest neighbor which find the most similarity observations of given sample. In this process, we select the nearest observations and use his or her age as filling number[3].

3 Exploratory Data Analysis

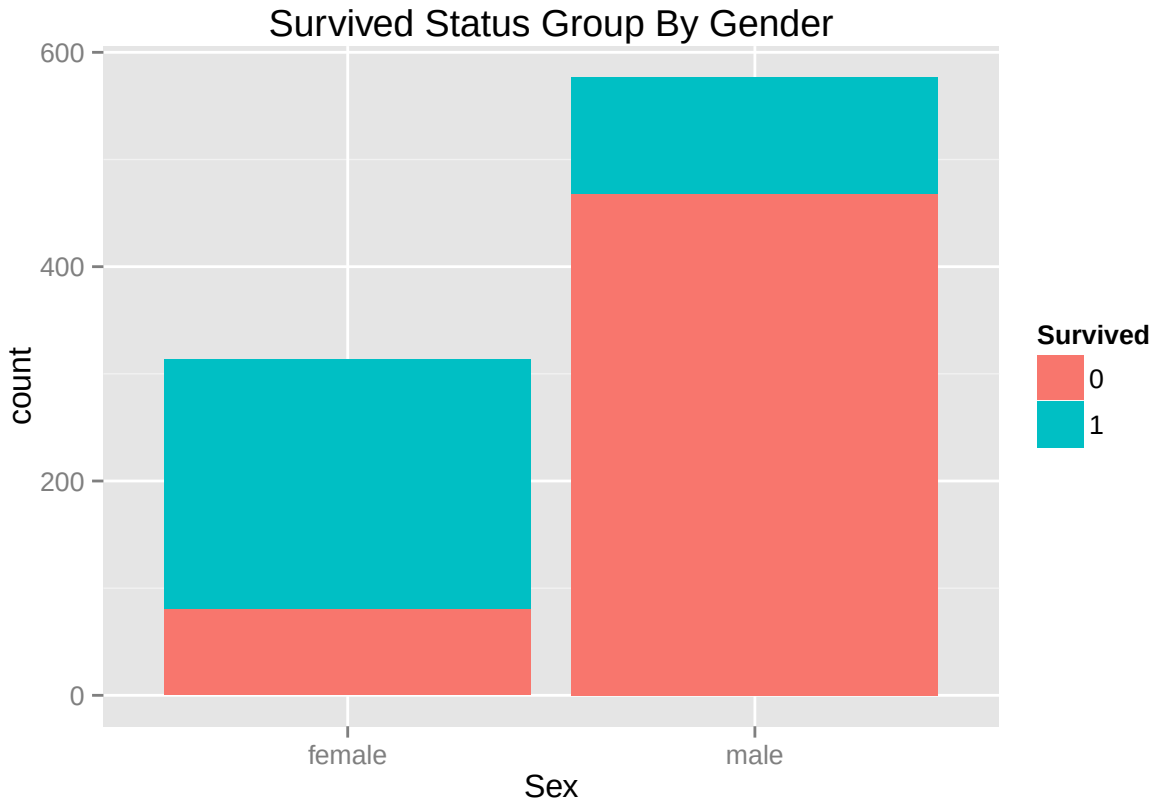
First, we can use a bar plot to generate descriptive statistic for passenger survival:

[3]Torgo, L. (2010). Data Mining with R,learning with case studies Chapman and Hall/CRC.

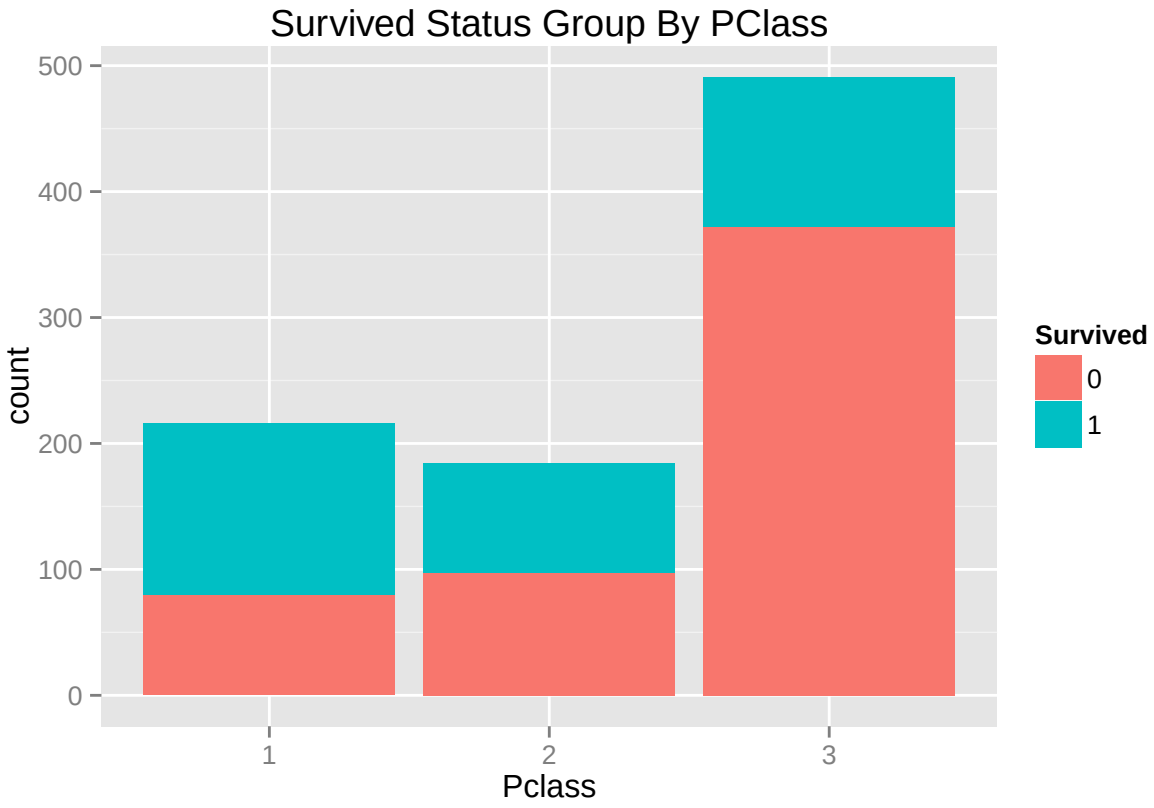


As 0 stands for perished and 1 stands for survived, we are able to see that approximately 570 passengers (63%) perished and 340 passengers (37%) survived in our train data.

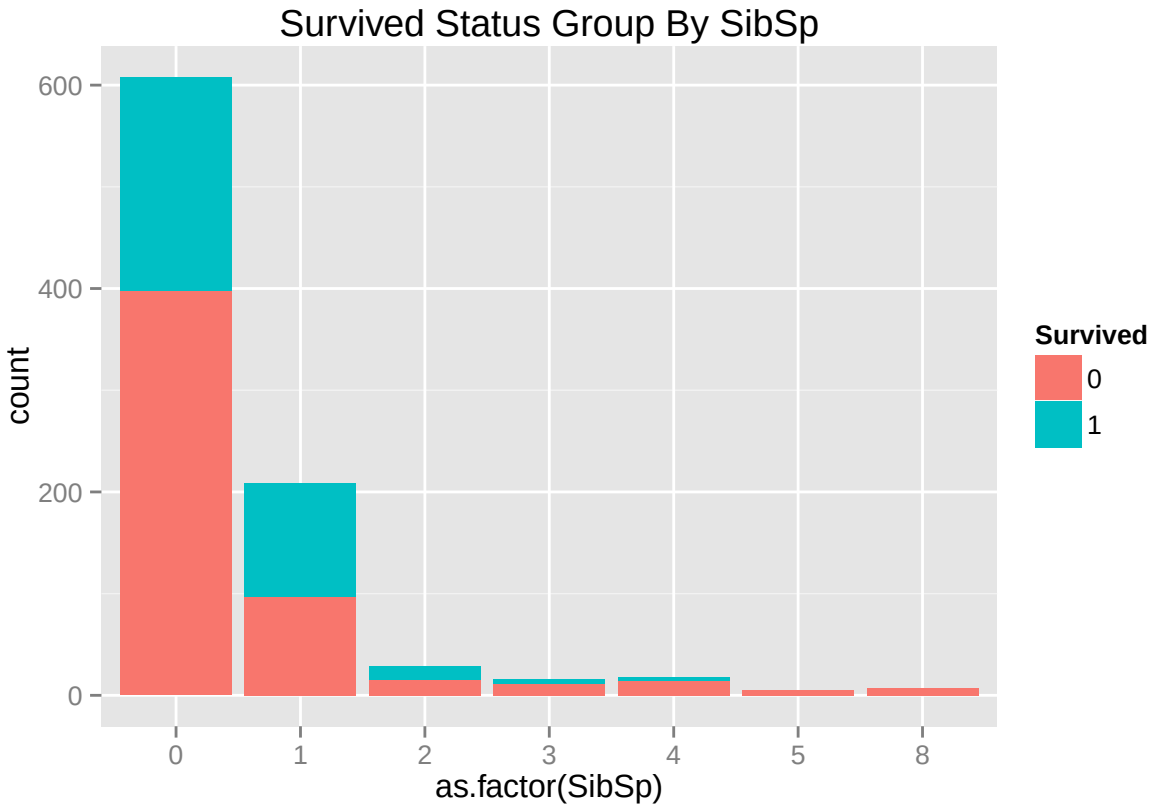
We use bar plots to see survived status by each group, including gender, Pclass, SibSp, Parch, and Embarked.



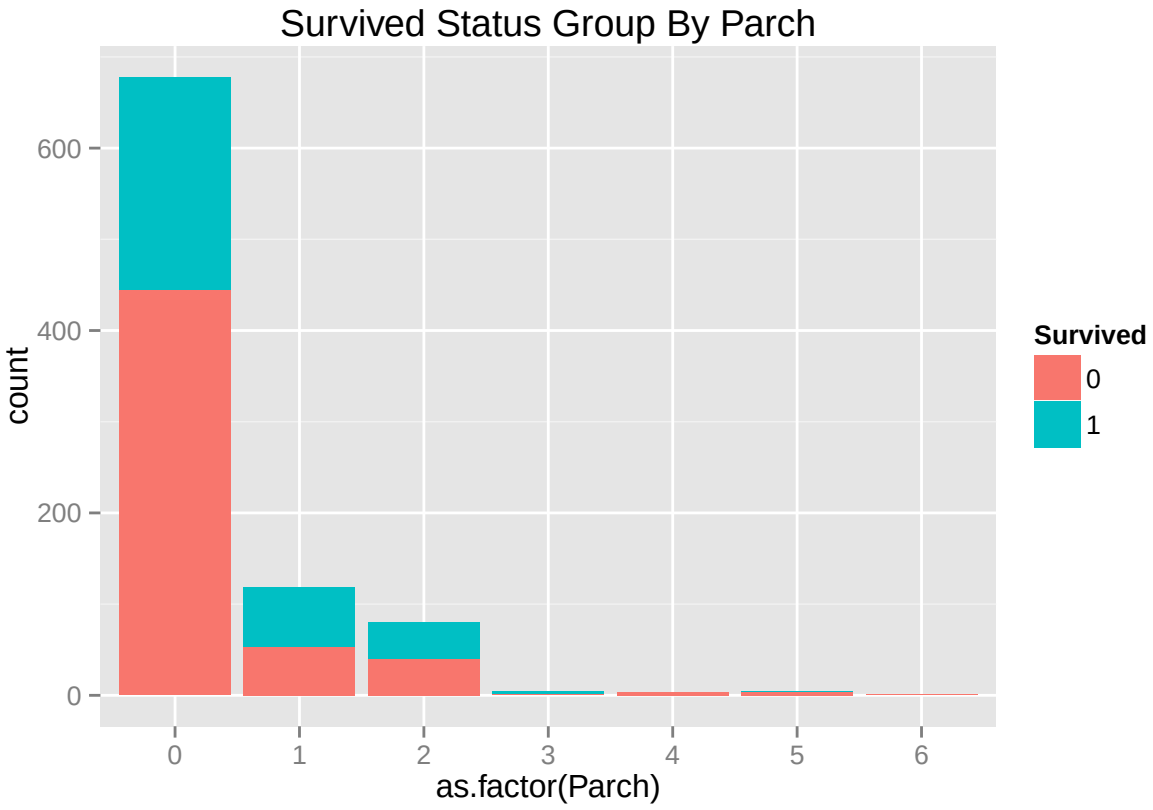
Based on the bar plot above, we have approximately 325 female and 585 male passengers on board. We can see that about 75% female survived while only 19% male survived. The probability for a female to survive is significantly higher than a male on board in our train data.



Tatanic has three passenger classes, in which the 1st and 2nd class have 200 people respectively while the 3rd class has almost 500 passengers. The first class has an approximate survival rate of 64%, the second class has a rate of 48%, and the third class has a rate of 25%. It clearly shows that the higher the passenger class, the higher probability that a passenger would survive.

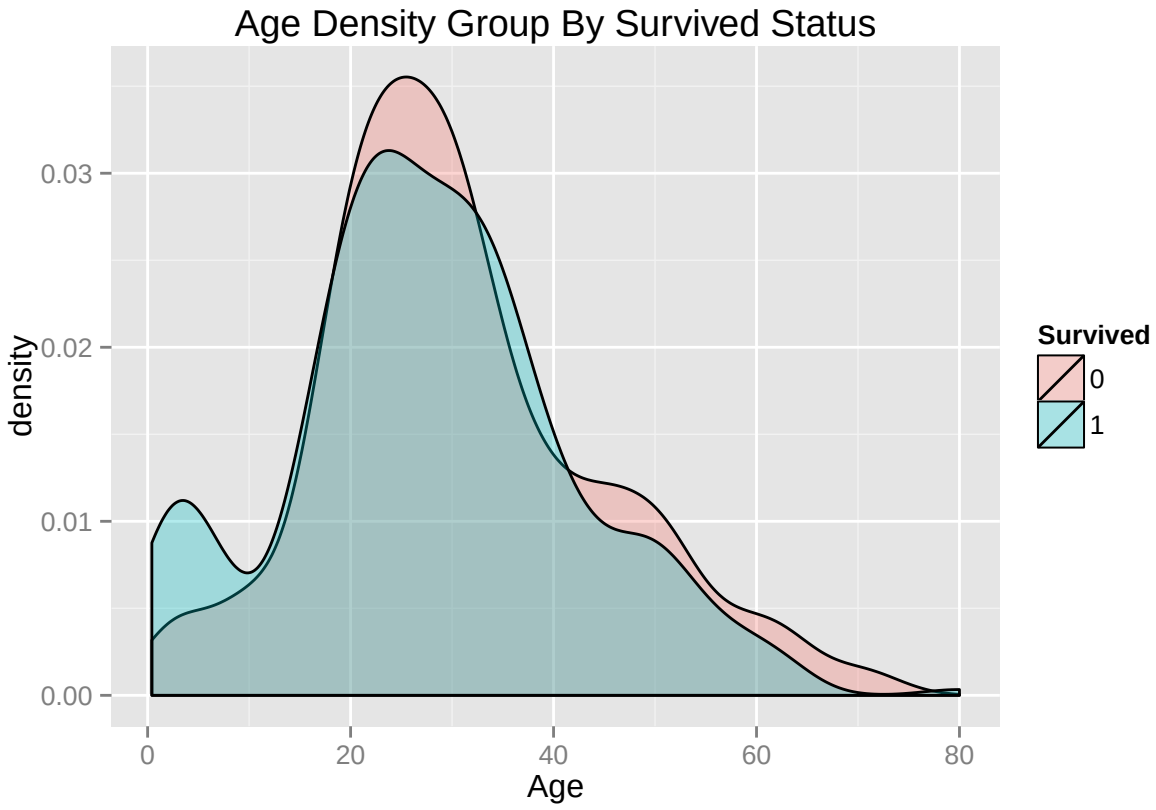


Based on the bar plot of survived status group by SibSp, we can see that more than 60% passengers came on board without siblings or spouses. However, we are not able to figure out a clear relationship that how the number of SibSp would impact the probability of survival.

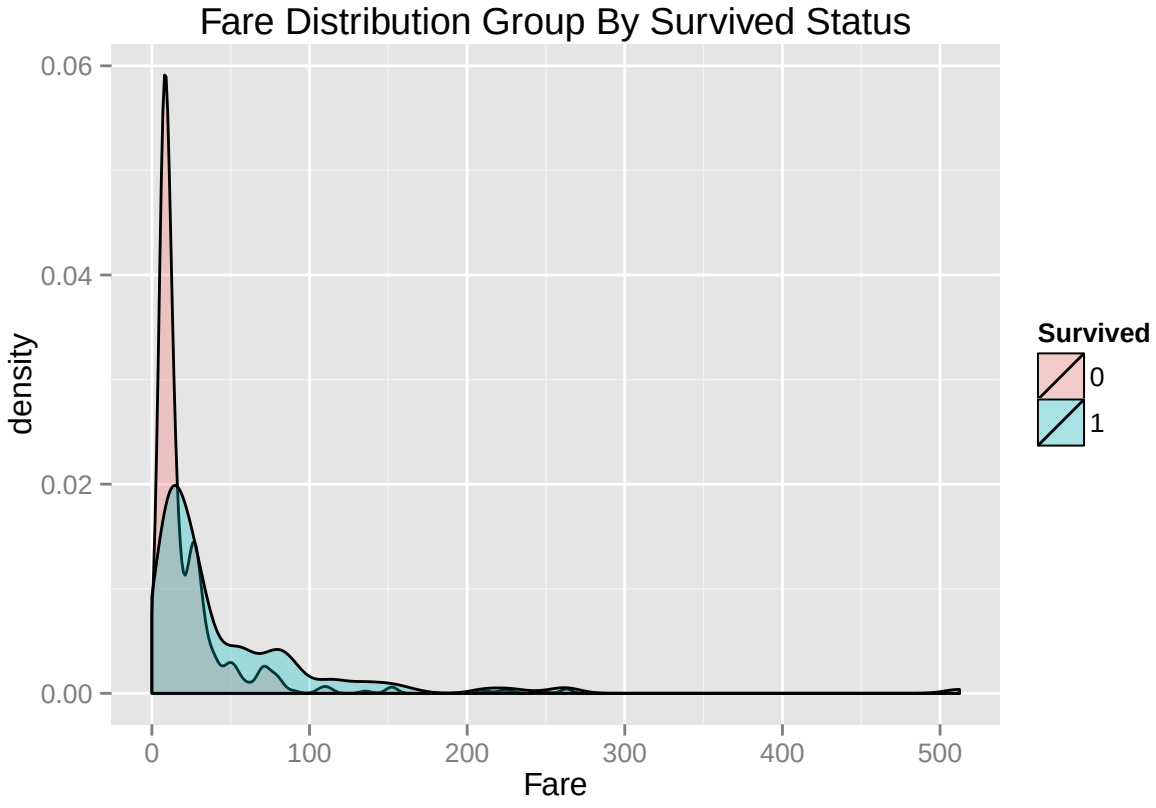


We get the distribution of the passenger parch and find out that about 75% passengers did not bring any children or parents abroad. However, we cannot see a clear trend that the number of parents or children would impact a passenger's likelihood of survival.

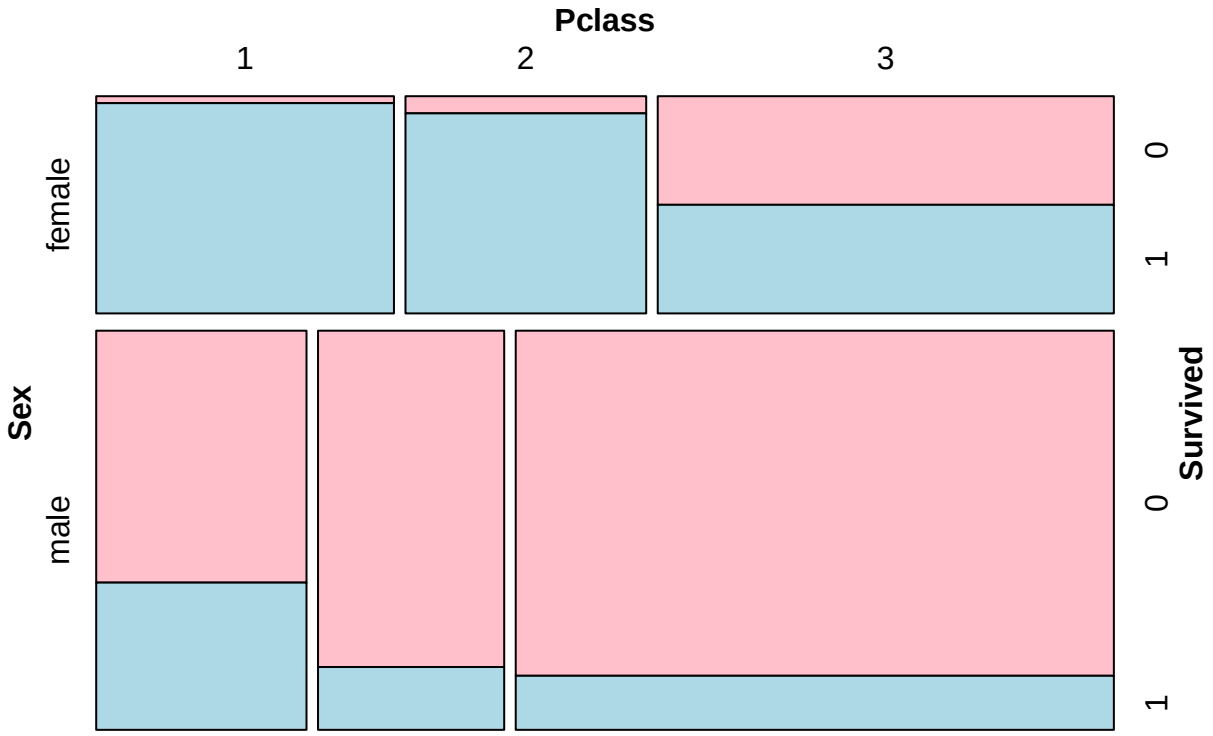
Tatanic has three different ports of embarkation, in which c stands for Cherbourg, Q stands for Queenstown, and S stands for Southampton. Passengers embarked from Port C has an approximately 56% survival rate, Port Q has a 38% and Port S has a 33% survival rate.



We can see from the age density that the majority of passengers are from 20 to 40 years old. The overall impact of age density on survival status is not significantly shown in the graph. However, when the age below 10, the survival rate are relatively high. We can conclude that children has a higher probability of survival than adults, but other than that, age is not a significant factor on survival status.



The fare distribution shows us a clear picture that higher fare gives a passenger higher probability of survival. Based on the graph, for passengers whose fare belows to $20, the likelihood of survival is significantly lower than being perished. However, for passengers whose fare are from $30 to $200, the likelihood of survival is higher than being perished.



The graph above illustrates the relationship between sex, pclass and survived. It clearly shows that more passengers survived in higher class, which indicates the passenger who were in higher class had higher probability surviving in the sinking despite of sex. To consider sex, female were more likely survived than male. Female who were in first class had the most probability surviving in the sinking and male who were in third class had the least probability surviving in the sinking.

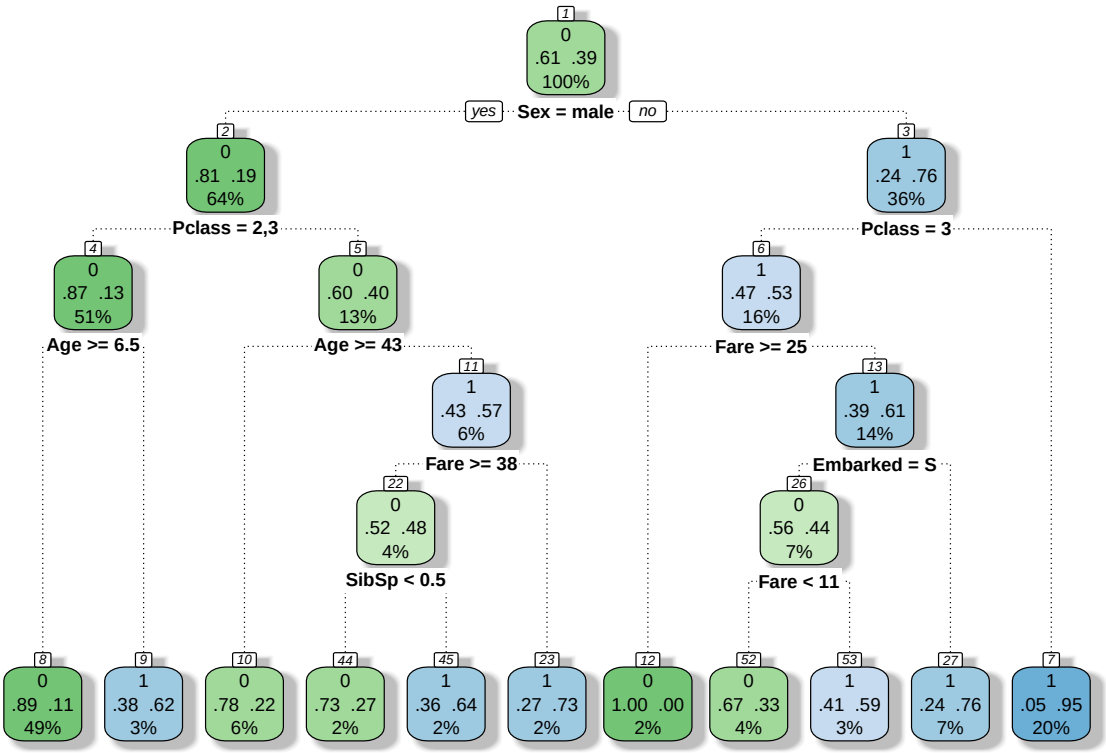
4 Training Predict Model

We separate the training dataset so that 70% of the data are used for training, and 30% are used for validation. Our first basic prediction model is based on decision tree, selecting features include Sex, PClass, Age, Sibsp, Parch, Embarked and Fare. We use rpart[4] to train our model and rattle[5] to vasualize the decision tree result.

```
## Warning: Failed to load RGtk2 dynamic library, attempting to install it.
```

[4]Terry Therneau, Beth Atkinson and Brian Ripley (2015). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-9. http://CRAN.R-project.org/package=rpart

[5]Williams, G. J. (2011), Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery.



In evaluation part, firstly we generate confusion matrix. A confusion matrix can provide information about how many observations are correctly predicted and how many passengers are wrongly classified.

Table 2: Confusion Matrix for Decision Tree

	Prediction == 0	Prediction == 1
True Mark == 0	145	27
False Mark == 1	32	64

We use four measurements (accuracy, precision, recall and F1 score) to approach the performance of our models. The definitions of the four measures are described below [ˆ7]. Accuracy is defined by the number of items categorized correctly divided by the total number of items. Precision is what fraction of the items the classifier flags as being in the class actually are in the class. Recall is what fraction of the things that are in the class are detected by the classifier. In short, precision is a measure of confirmation while recall is a measure of utility. F1 score is a useful combination of precision and recall, so that F1 has a small value

if either precision or recall is very small. The evaluation result of model 1 will be stated at the end of this section.

Table 3: Classifier Performance Measures

Measure	Formula
Accuracy	$(TP + TN)/(TP + FP + TN + FN)$
Precision	$TP/(TP + FP)$
Recall	$TP/(TP + FN)$
F1	$2 * precision * recall/(precision + recall)$

Basic decision tree model may perform very weak at the beginning, embedding learning algorithm(such as adaboost, bagging, random forest) can avoid this problem. In this section, we use random forest to train our data again, and generate the evaluation results for both decision tree and random forest.

Here is the confusion matrix for random forest.

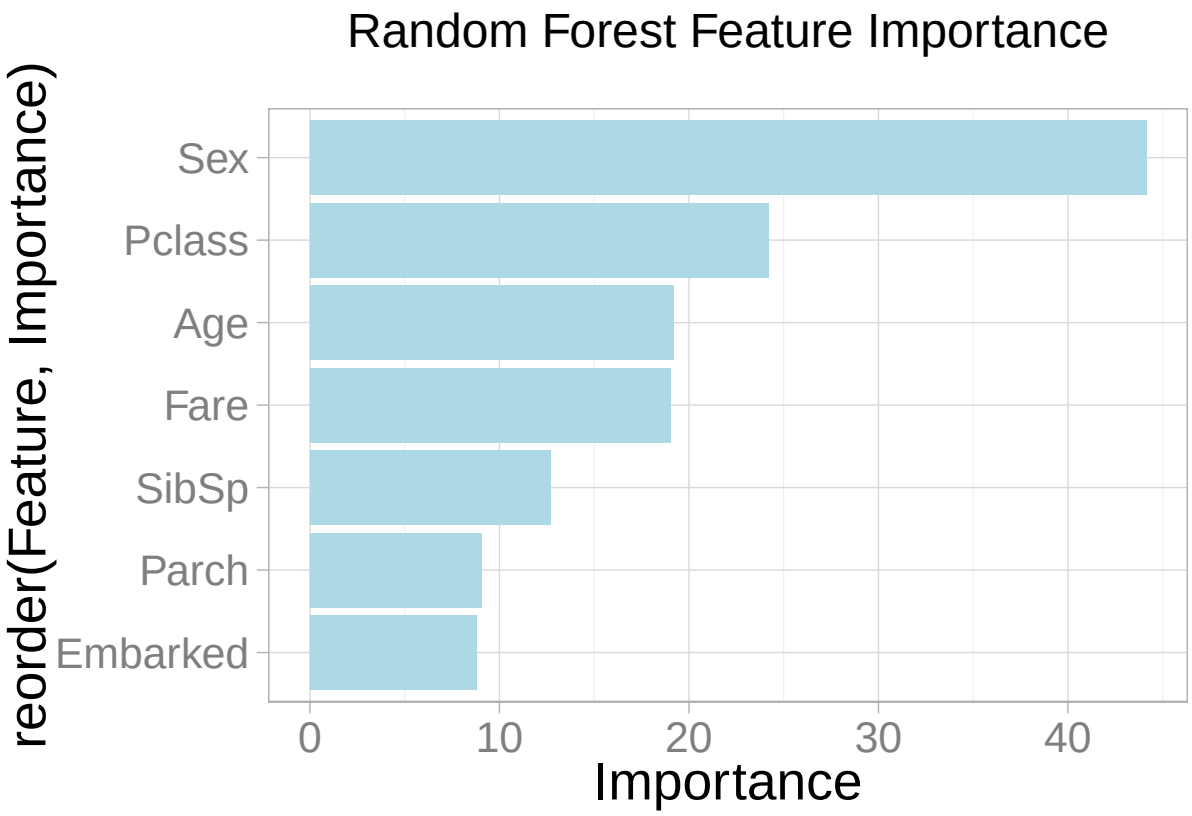
Table 4: Confusion Matrix for Random Forest

	Prediction == 0	Prediction == 1
True Mark == 0	162	10
False Mark == 1	28	68

Table 5: Evaluation of Prediction Models

	Decision Tree	Random Forest
Accuracy	0.77	0.85
Precision	0.70	0.87
Recall	0.66	0.70

	Decision Tree	Random Forest
F1	0.68	0.78



5 Improved Model

By comparing the two density plots of age and fare that were generated above we could conclude that there is no significant difference in age distributions for survival and non-survival; however, there is observable difference in fare distributions for difference survival status. According to the description of plots above, it can also be seen that there is no significant effect of Sibsp, Parch, Embarket on the result of survival. Hence we decide to use sex, Pclass and fare as the only variables in our model 2.

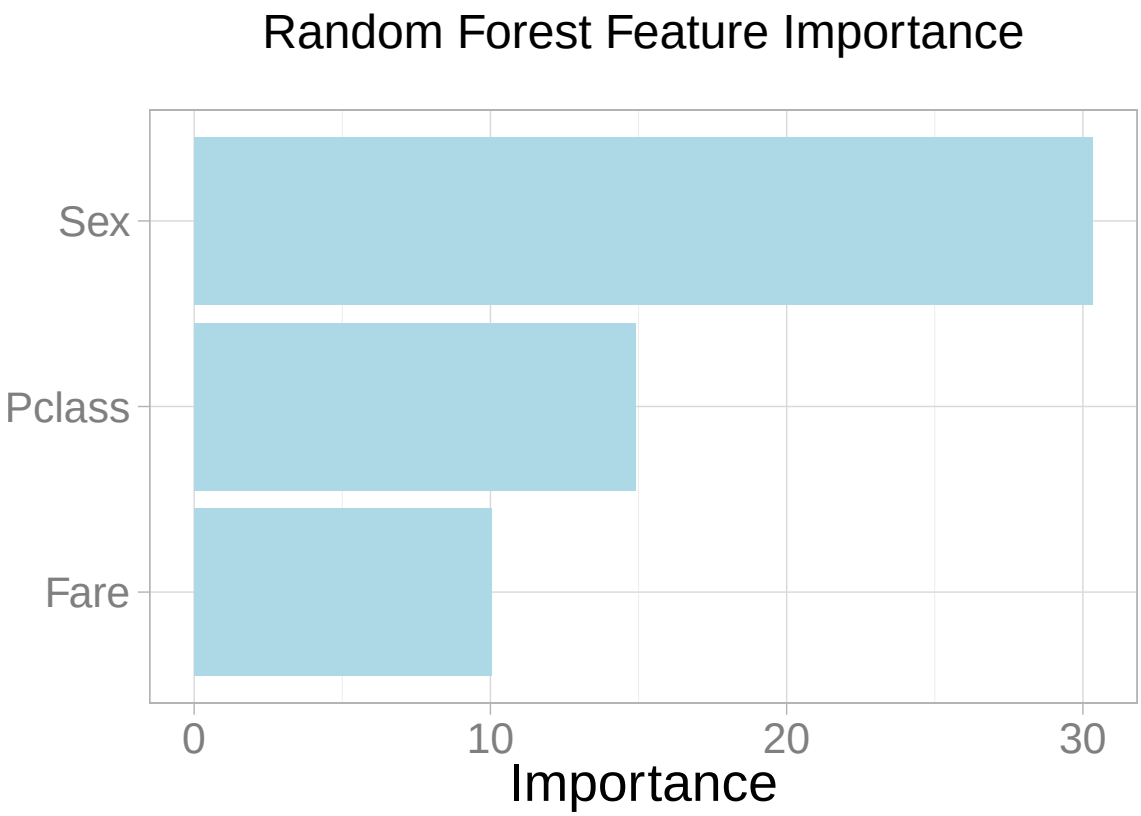
Table 6: Confusion Matrix for Random Forest

	Prediction == 0	Prediction == 1
True Mark == 0	153	19
False Mark == 1	36	60

We calculate evaluation scores of model 2 and compare them to model 1. The result indicates that the second model is not fit traning dataset well compared to model1. However, it gain a better score than model 1 at kaggle test dataset(Which we don't know the survived status of those data). In this situation, we can conclude that the first model have some irrelevant features which need to be transformed or removed.

Table 7: Evaluation of Prediction Models

	Rondom Forest 1	Random Forest 2
Accuracy	0.85	0.79
Precision	0.87	0.76
Recall	0.70	0.63
F1	0.78	0.68
Kaggle Scores	0.75	0.79



6 Discussion

Based on our topic, we can draw the conclusion that gender was the most significant factor which determined whether the passenger survived or not. From the analysis above, it is obviously that female had higher probability survived in the sinking. In the meantime, passengers who were in higher class were more likely survived than those were in lower class. There are some points we want to mention:

- Firstly, adding all or many variables as factors to the model is not always a good idea. In fact, sometimes it will create an over-fitting phenomenon on model and make the accuracy of prediction even worse.
- Secondly, some variables may have affect to survived rate such as titles of name and cabin. We may discuss effects of those features at further study.
- Last but not least, all of our models are unify model, which didn't consider the difference between gender or Pclass. It is possible that survived features are different between male or female.

7 Contribute

- Introduction (Linfeng Zhou)
- Getting and Wrangling Data (Xiaoge Wu)
- Exploratory Data Analysis(Tianyi Gu, Xiaoge Wu)
- Training Predict Model(Linfeng Zhou, Yi Zhang)
- Improved Model(Yizhang, Tianyi Gu)
- Discussion(Linfeng Zhou, Tianyi Gu)