**Name: Pratik Patel**

**Class: CS 483 Big Data Analytics Capstone**

**Professor: Dr. Elena Braynova**

**Course Project**

**Overview**

The project I chose to do is based on one of the horrific tragedy that took place in the year 1912. The Titanic sank and 1502 people died in this heartbreaking incident. I am going to use the dataset that is available to me and do the analysis of the people who are likely to survive in this kind of similar incident using R.

**Dataset**

I am using two datasets for this projection. The Test dataset and Train dataset. The descriptions of the variables are as following.

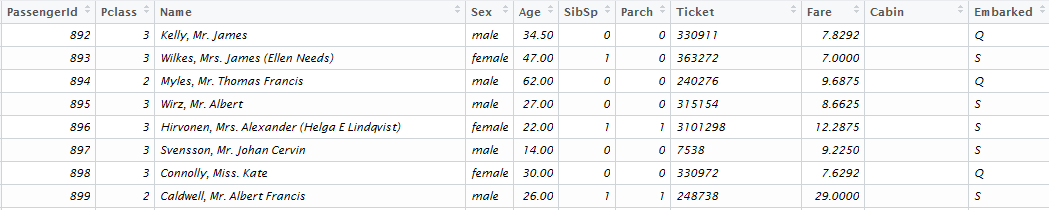
Survival: (0 = no, 1 = yes)

Pclass: (1 = 1st class, 2 = 2nd class, 3 = 3rd class)

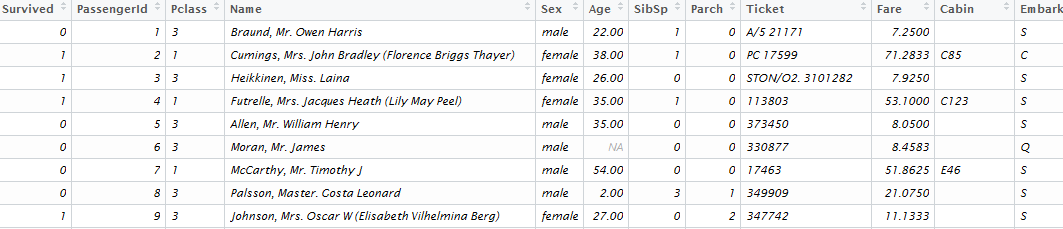
SibSp: Number of siblings/spouses aboard

Embarked: (C = Cherbourg, Q = Queenstown, S = Southampton)

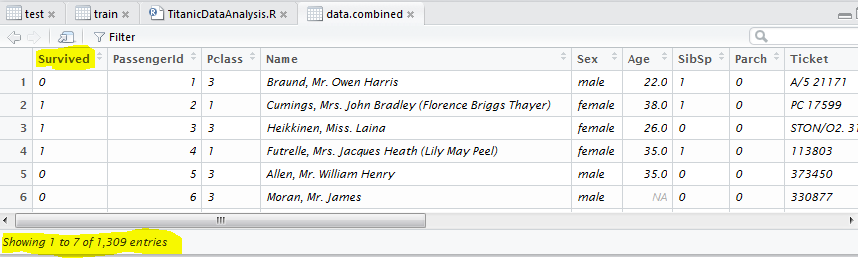
Test.csv: Test dataset has 418 observations in them with 11 different variables.



Train.csv: Train dataset consist 891 observations with 12 variables.

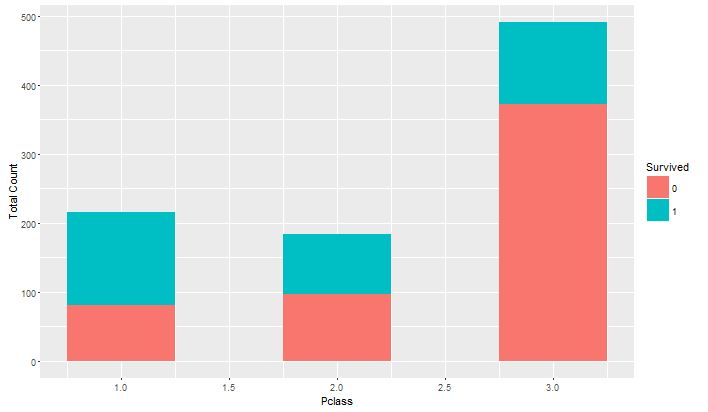


Using these both datasets I created a new test. Survived dataset with survived variable added into it. After this I combined both of the data sets and created data combined. The following is the example of the data.combined dataset which holds 1309 observations with 12 variables. I also converted Pclass from int to factor and also converted survived from Chr to factor.

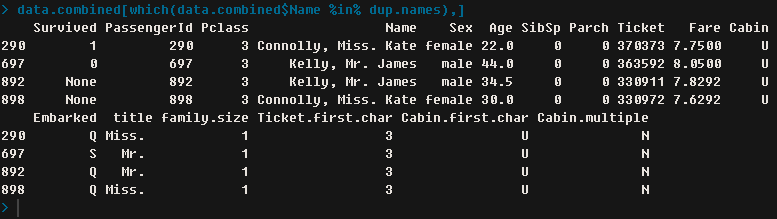


**Hypothesis**

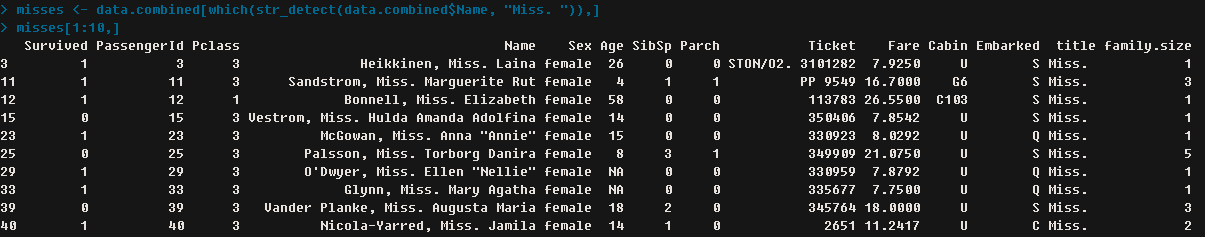
**1)** The first ggplot will show that the passengers in the first class have easy access to the life boats. So the general setup in the ships is the higher amount you pay for your tickets the higher you will be on the ship deck. That will keep you closer to the life boats.



In the next step I found whether my dataset holds duplicate data entries or not. My results showed only two names which are duplicated. Now I followed this lead to look into combined dataset with these names associated.

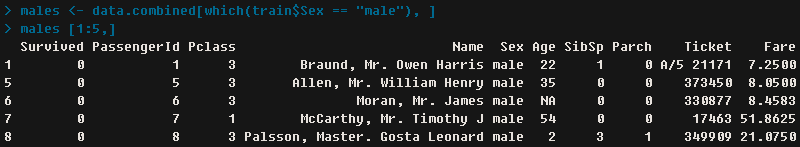


Another thing I focused on is the titles of these names. The titles like Mr. Mrs. And miss are there with every single name. They could be there for a reason, so I checked their predictive powers.



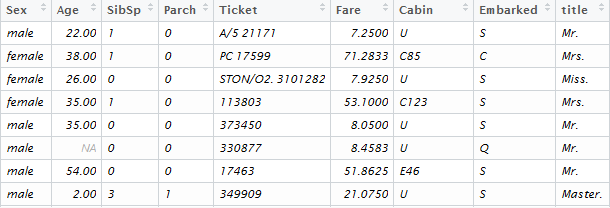
We can already see some interesting number analysis. If we looked at the survived, out of the 10 records 7 of them survived. Other interesting stats are from Pclass. Out of the 10 number of records 8 of them are from 3rd class. If we look at the Parch field it suggests the number of parents or children aboard. Only two entries show that they had parents or children onboard. Looking at the age we can assume that they have parents onboard with them.

**2)** Second hypothesis consist name title correlation with the sex of the passengers. The following screenshot will show you the male passenger’s data.

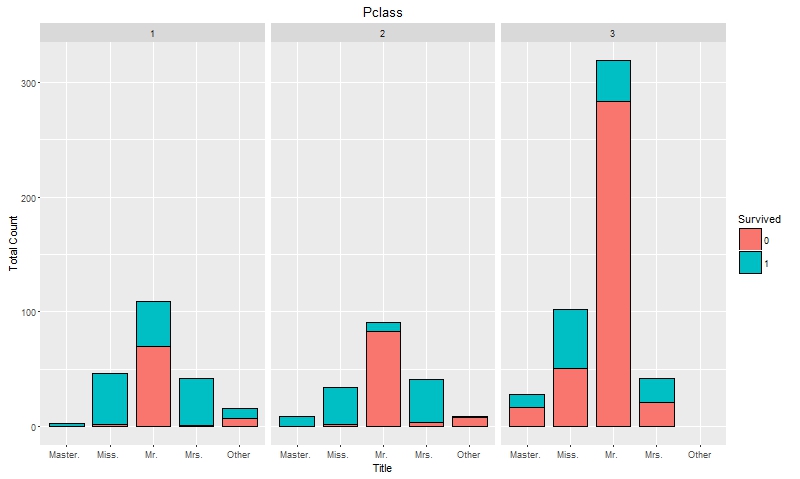


If we look at the numbers on the previous screen we can see the pattern that continues. The number of passengers that survived was not as high from 3rd class compared to the 1st and 2nd class.

Looking at the titles in the names of the passenger it makes us wonder if that has any patterns in them that we are missing. So in order to figure that out I added the title column in data.combined. The following screenshot will show you that.

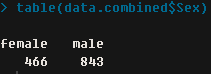


Now let’s combine the title and survived label to create ggplot and see if that gives us any insight.

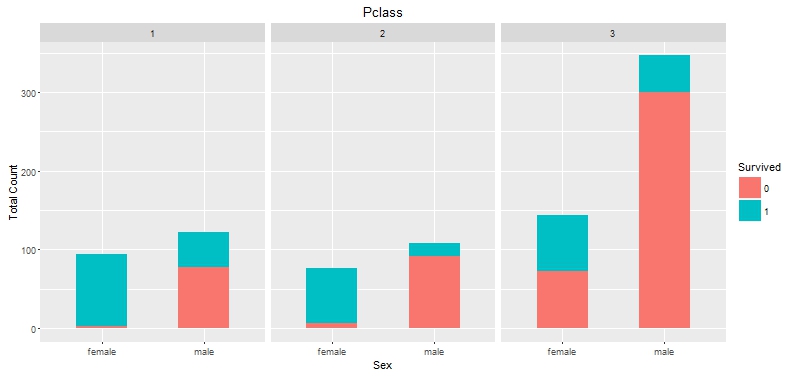


Looking at the chart above we can clearly see that for Mr. in the 3rd class has really less chance to survive. If you are Miss or Mrs. in the 3rd classes your survival chances are 50/50. But even looking at the whole plot we can see that the chances for Mr. to survive are less compare to Mrs., miss or master. This could be related to the old saying of women and children first.

Now we knew looking at the distribution of the females to males that females are more likely to survive compare to males. As the numbers suggested the following.



The following plot will show us three way relationships between sex, Pclass and survival.



The interesting point we can take out of this graph is that males in second class did not have better chances to survive compare to 3rd class proportionally. But overall female has more chances to survive compare to male. This plot also suggests that our previous plot that holds title values is correct and we are on the right track of our analysis.

**3)** Let’s take our focus to age variable and missing values. The following screenshot will show you the summary of age of the entire dataset.

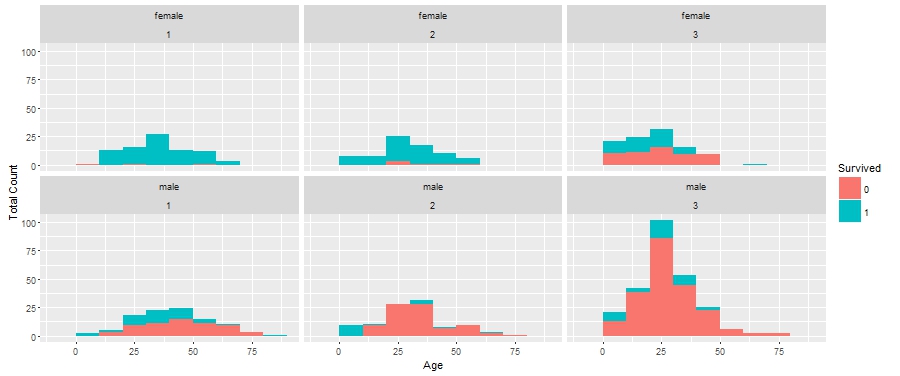
**P:\STUDY\Data Analysis\final project\age summary.PNG**

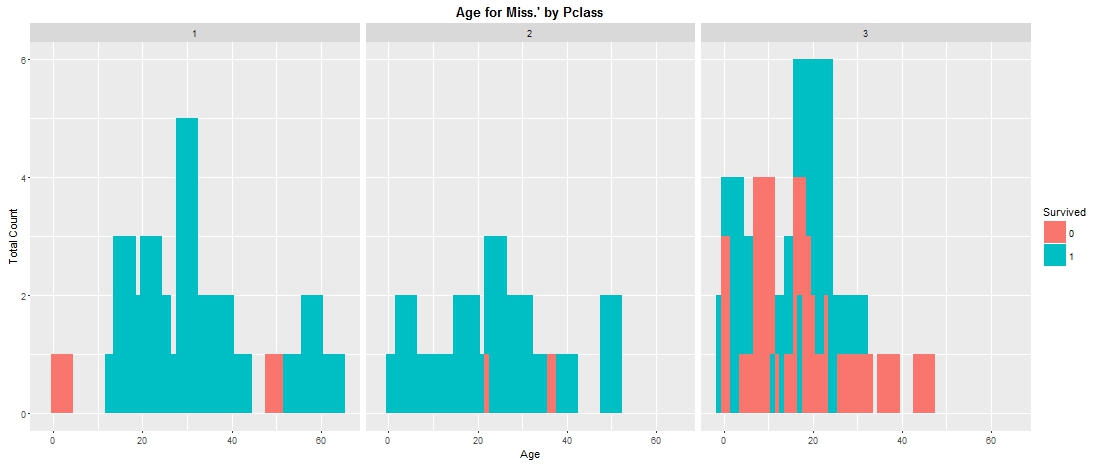
Now looking at the total numbers of variable which are 1309 in data.combined, 263 are missing values in the age variable. It almost closes to 1:5 ratios. We can go further and see how many missing values we have in our train dataset.

P:\STUDY\Data Analysis\final project\train missing value summary.PNG

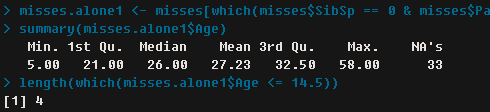
So we can see with the screenshot above that not only we have so many missing values but more than half of them are in the training data. The missing values are more skewed towards our training data. We have to make sure we find some other variable to find it as proxy to the age.

The following plot will show the survival rates broken down by the sex, passenger class and age.

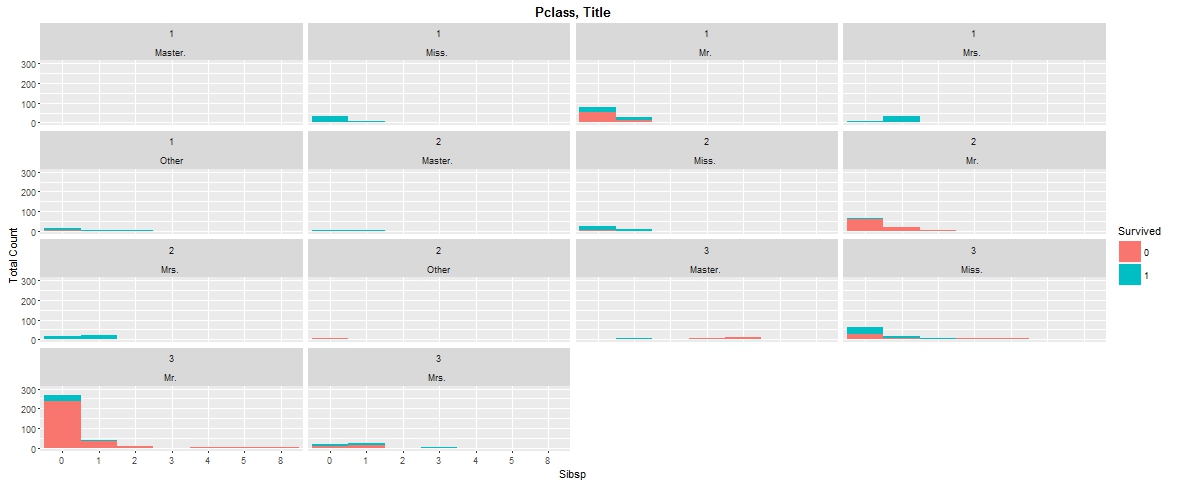


The plot shows the breakdown of female and male by class. It also shows the survival rates throughout the different age group. So we can clearly see that in the first class hardly any female did not survive. Another interesting stat we can see is if you are between the ages of 1 to 22 female in second class you are most likely to survive. For the female in the 3rd class only have 50/50 chances to survive. On the other hand the survival chance for male in 3rd class is bad especially there were bad chances to survive as older the age is for the male. That is the case even in 1st and 2nd class for male. As age increased the chances are slimmer for them to be survived. Let’s take our investigation further with female that has title miss in to their name. We can see one very young female did not survive and one female with the age around 50 did not survived. In the second class the story is similar. Where in the third class the chances looked mix. The chances are more of at 50/50. And we can see as you grow older the chances are lesser for women to survive with the title miss in the 3rd class. The following plot does present us with this information. 

One aspect we have to figure out with this is to if you have title Miss, is the person a child? This question needs to be solved in order to get accurate analysis of our data. Because we can see looking at this plot there are many miss in 3rd class compare to others. The following screenshot will show the Miss which are under the age of 14.5. There are only 4 female children under the age of 14.5.

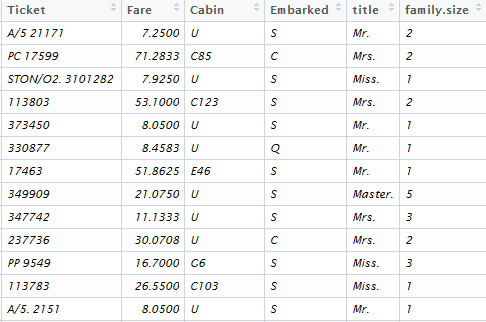


**4)** Now let’s predict survival rates using Sibsp, Pclass and title.

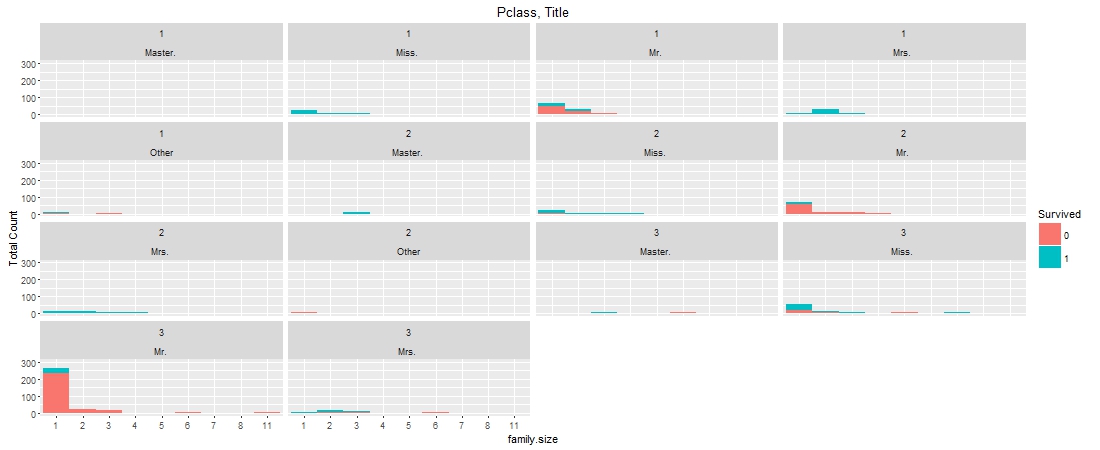
****

The similarities are still there in this plot above which shows if you are child in the 3rd class and travelling with fewer siblings you are more likely to survive.

Now we can create family size feature and use that to see if you have more chances to survive if you are alone or with the family.

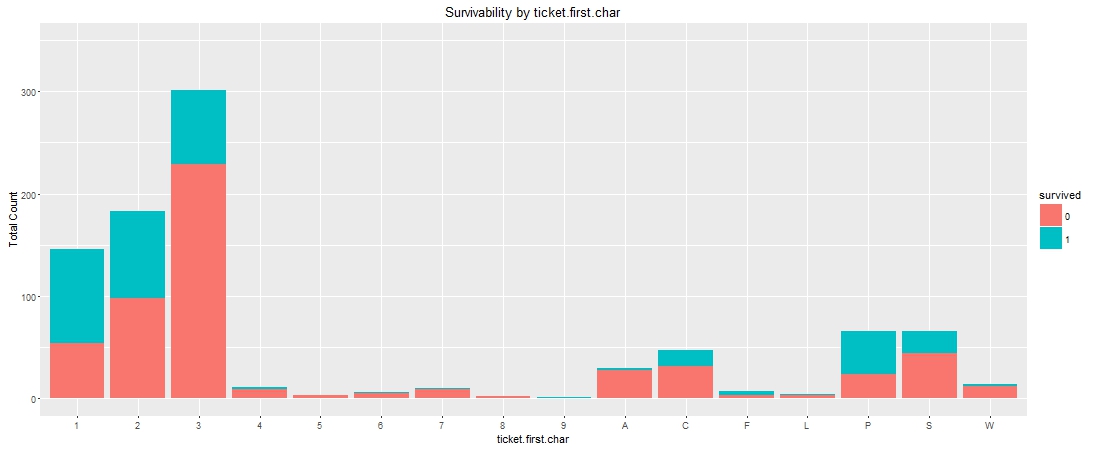


As you can see in the screen above, we now have family size which allows us to see how bigger the family is and how better the survival rates they have.

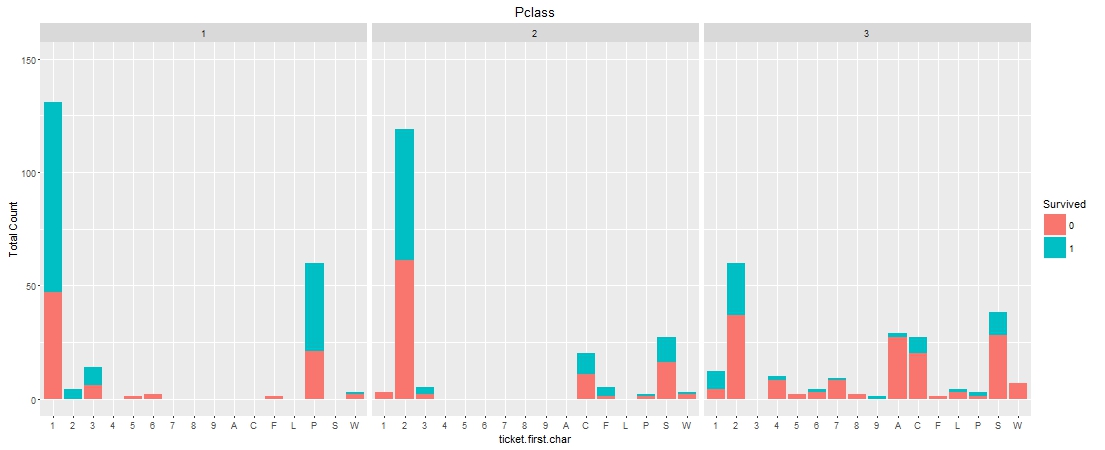


Looking at the plot we can notice the bigger the family size the less chance you have for survival. As we know that is for sure for Mr. in the 2nd and 3rd class.

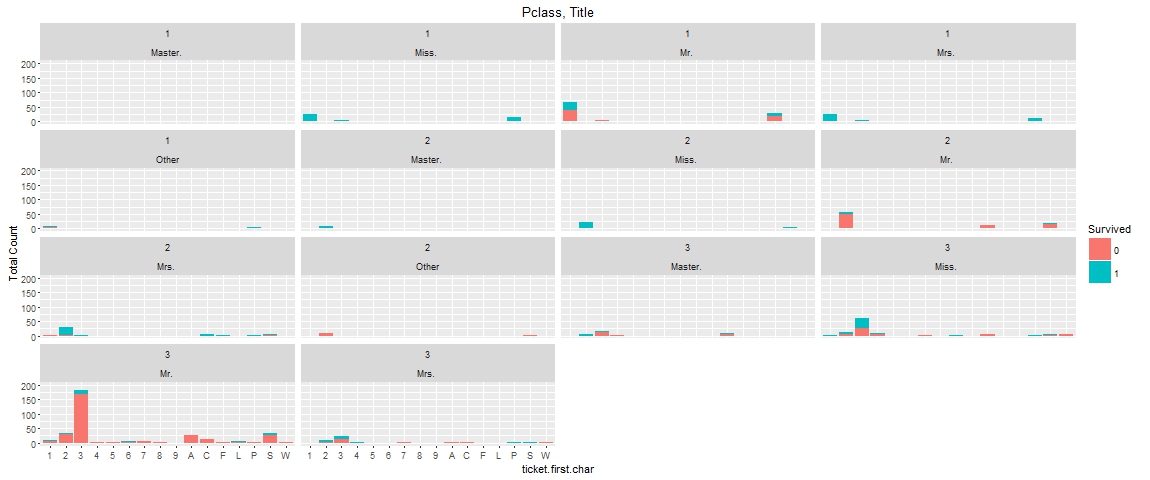
Now after collecting all these values lets plot the factored ticket variable and survived variable. The screen below shows the survival rates according to ticket numbers.

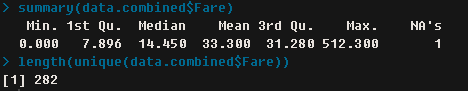


We may not get that much of an idea but the information of tickets does look interesting in itself. So I tried more drilling down approach for this variable.



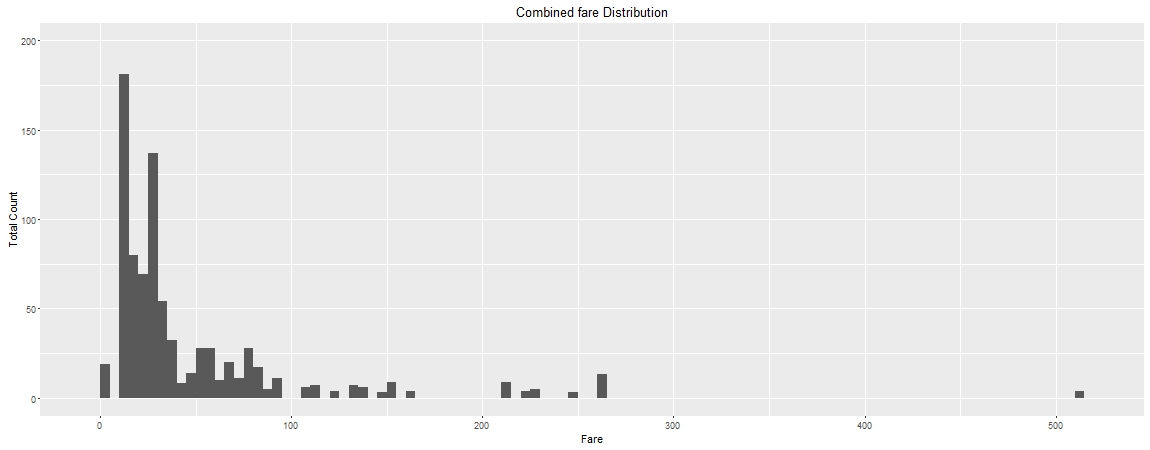
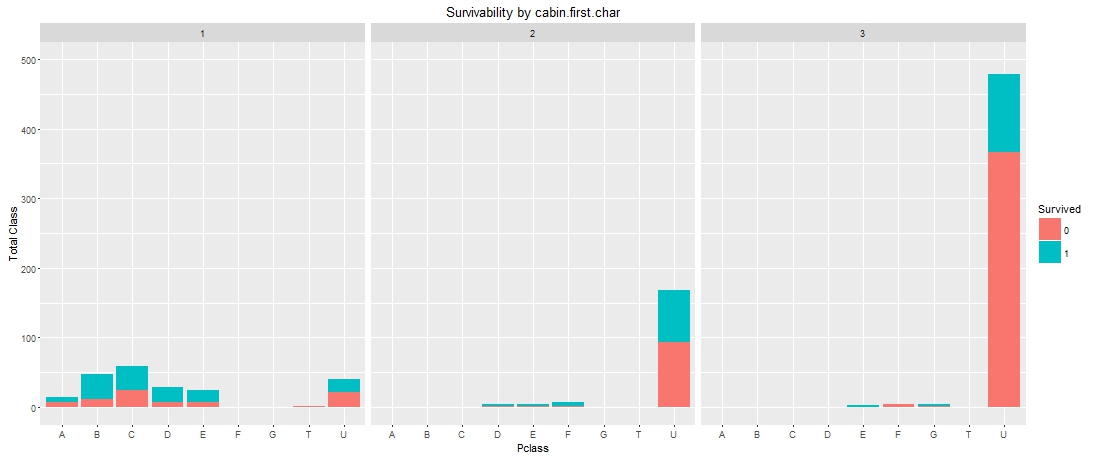
Most of the tickets which start with number 1 were being sold to the 1st class passengers. Tickets that starts with number 2 are mostly sold to 2nd class and 3rd class passengers. Let’s look at the third plot for the tickets analysis. Looking at the third plot there isn’t much information that could be used in order to make our analysis.

 **5)** We can give our focus to fare variable. Since its numeric values, let’s look at the summary of the fare variable.

****

Looking at the summary some of the passengers didn’t pay any fare at all. The first quartile paid 7.896 British pounds. So almost close to 50% of people paid less than 14.450 British pounds. If we look at the max value its 512.300 British pounds.

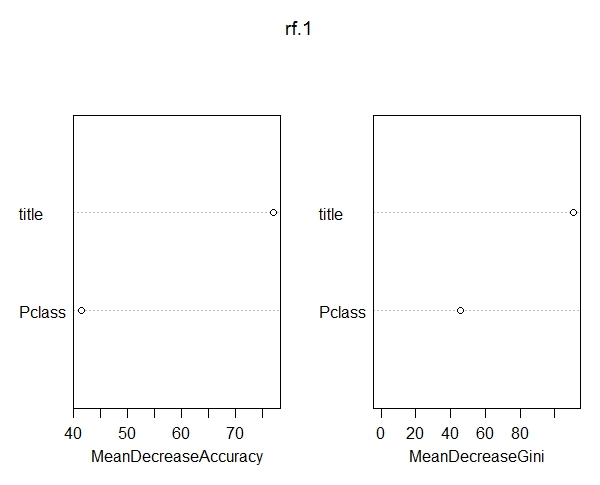
The unique values are 282. Because it carries so many unique values we don’t want to look at this. Let’s look at the fare prediction power by the plot. The following plot will show us all the analysis related to the fare variable. We have one outlier which is the 512.300 pounds fare been paid by a passenger. Looking at the overall graph its expediential decline distribution with the passengers that didn’t pay much to get on the Titanic.

 We can also look at the other variable which is cabin. As data analyst we cannot underestimate any variables in the dataset. 

As we can see in the plot that most of the first letters belongs to the 1st class passengers where the lower cabin letters belong to the 2nd and 3rd classes. The interesting visualization is if your cabin number starts from E or F in the 2nd class your chances of survival are higher. On the other hand in the 3rd class your chances of survivals are rare.

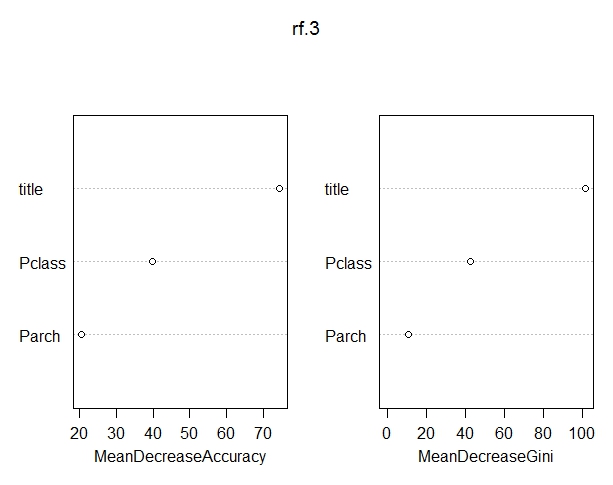
After all these analysis our data.combined consist 17 variables instead of 12. We added title, family size, cabin and multiple cabin variables to our dataset in order to do our analysis.

Now let’s focus on building a model from our dataset. We are using randomForest model to see our data through. In the following plot you can see we used title and Pclass.

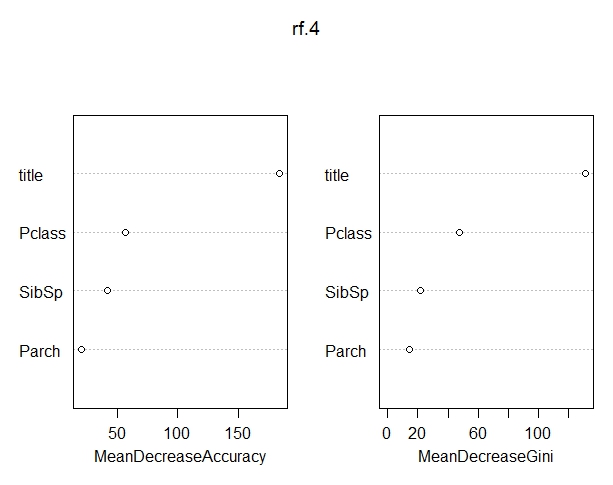


With the RandomForest further to the right is your variable the more important that variable becomes. In the RandomForest above our Title variable is far more important than Pclass variable. With using two variables we got error rate of 21.44%

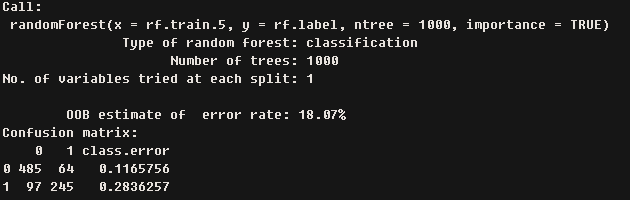
Now, let’s create the RandomForest using the title, sibsp and Pclass. Now if we look at our error rate using three variables together it decreased by 2%. The error rate it gave with three variables is 19.75%. Two percent in the error rate is really good improvement in the randomForest. It increased the accuracy rate for the people who survived.



Now let’s try another variable to make our model stronger and let’s see if we can decrease our error rate. We are going to use Title, Pclass, sibsp and Parch for the RandomForest below.



This looks even better from our previous model. The error rate decreases even more. We are getting the error rate of 18.86%. We can see looking at this model that the number of family matters. Also the women and children come first. We also can agree that wealthier people are tend to survive more than others.

Smaller families tend to survive then the larger families. The randomForest model I created using Pclass, title and family size gives us very good numbers. 

The model suggests the error rate of 18.07% using the family size variable with the title and Pclass. That is the best model that I created which will definitely work for sure. So using all these models we can get the result of 81.93% accuracy using our model.

**Work Cited:**

https://www.kaggle.com/c/titanic