Titanic Survival Machine learning Model

# Executive Summary:

The task is to model the survival of the passengers who were aboard on Titanic. I used Random Forest model and achieved an accuracy of around 83% on the test data set .The most important features were Sex, Travel Class and Age. 75% females survived and 65% people in first class survived. These were determined by visualising data using Mosaic and bar plot. Decision Trees, Random Forest parameter importance, confirmed this finding.

# Process:

## Introduction:

Titanic was a disaster, which struck on night of 14th April1912. Some people survived on that night. The challenge here is that given the data build a machine-learning model which will predict whether a person will survive or die based on the data provided for that person. Thus, the question I am looking to answer is that can I build a machine-learning model to predict whether a person will survive or not.

## Background:

The challenge was posted on the Machine Learning competition site Kaggle [2]. Kaggle team provided the data and explanation of each of the field is provided as well on the Kaggle page. Thus, I got the data and I did not need to look for it and capture it from various sites. There are 2 data sets one is train.csv and another one is test.csv. Please see [3] for more data and the explanation of the each variable. Some of these are obvious and I will explain it if it is not obvious. Below are 2 which I think are not obvious.

SibSp Number of Siblings/Spouses Aboard

Parch Number of Parents/Children Aboard

## Data Cleansing & Analysis:

This is one of the most important parts of the building machine learning models. This provides basic information about the data, types and structure of data and data distribution. This is useful in finding missing data and also whether we could convert some of the data from continuous to categorical variables. Also, the data is quite clean otherwise except few missing values.

### Missing Data:

There are some of missing rows in Cabin, Ticket, Embarked and Age. I will impute the missing data for Age (using median age). I could have used the median age based on Travel Class and Sex and Port Of embarkment to further find the proper median age. Also, I replaced 2 blank values in Embarked with “S”. I will not impute the data for remaining 2 columns, as I will not be using this in the Model or these are not useful columns from modelling point of view.

I did some basic analysis using the table command in R and here are few observations by looking at the data. This is just observation based on the data analysis and we will see how our other methods (data viz and Importance of parameters) supports or rejects these.

I have decided that I will be creating the age and fare column as categorical variables. Age will have a bracket size of 10 whereas for fare it will be 50. This, I will do when I will do modelling i.e. in TTN\_Model.rmd.

1. Females had more chances of survival than Males.
2. First Class passengers had more chance of survival than Second and 3rd class.
3. People in age bracket 20-30 had the highest chances of being dead.
4. Females with parent and child or siblings or spouses had less chances of survival whereas for man it was other way around.

Code for this is Data\_Analysis.rmd. Full output of this can be found in document Data\_Analysis.docx.

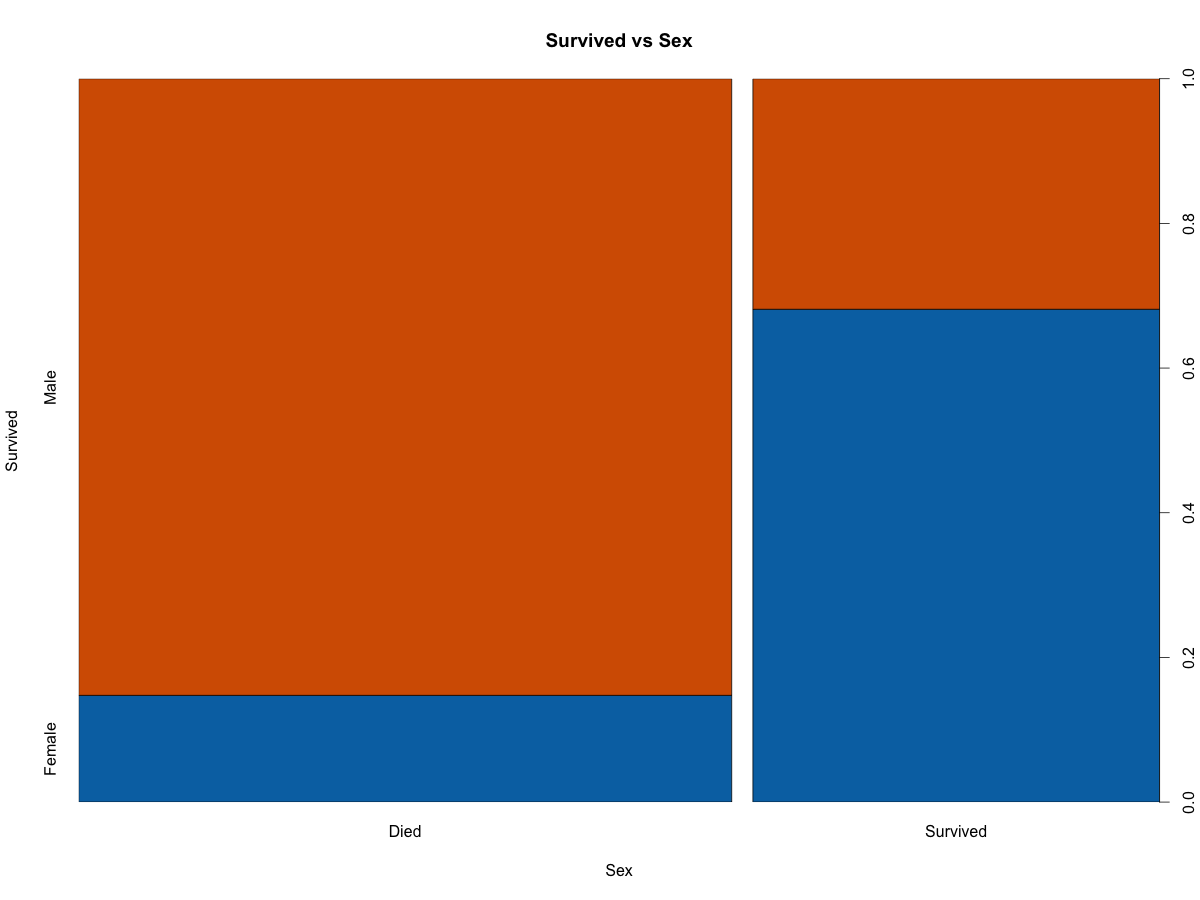
## Data Visualization:

Data Analysis above showed some nice patterns. But visualization, using graphs are a much better way to look at pattern and present these patterns. Thus, I have run the various data graphs and here I will be including only the graphs which are most important from the Machine Leaning modelling point of view for this given task. I will explain these graphs in 1-2 lines like what information these graphs are providing. These graphs use the % rather than absolute count.

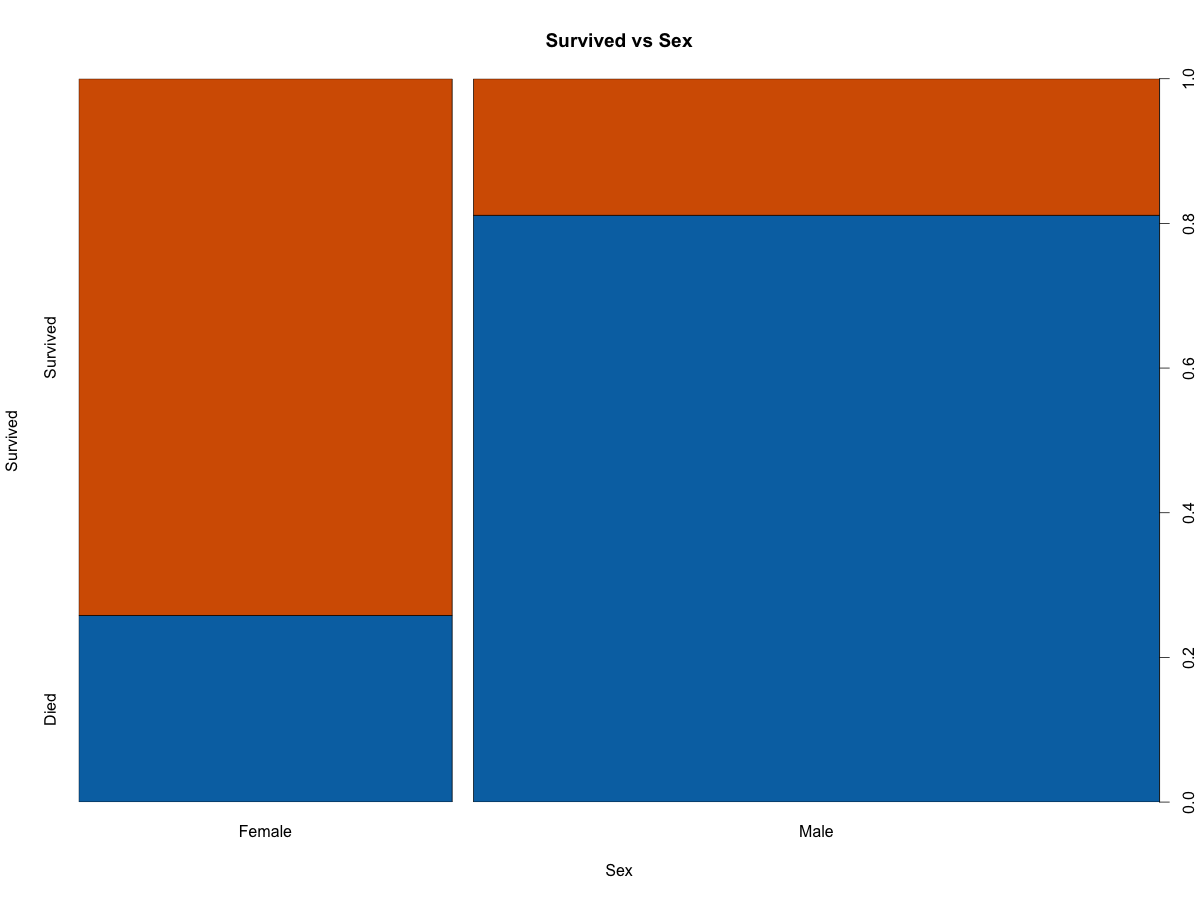
All of my data is categorical rather than continuous (except age and fare). Also, it is a classification problem and thus the bar chart and Mosaic plot would be much better idea. I was especially interested in Mosaic plots as these are quite good when we combine more than 2 or 3 variables. I have selected only graphs, which were showing some of the best patterns.

### Sex vs Survival:

Graph shows that if you were a female than chances of survival were very high as compared to the males. Graph shows that of the people who survived 70% were females and 30 % males. And, of people died the males were 85% whereas females were just 15%.

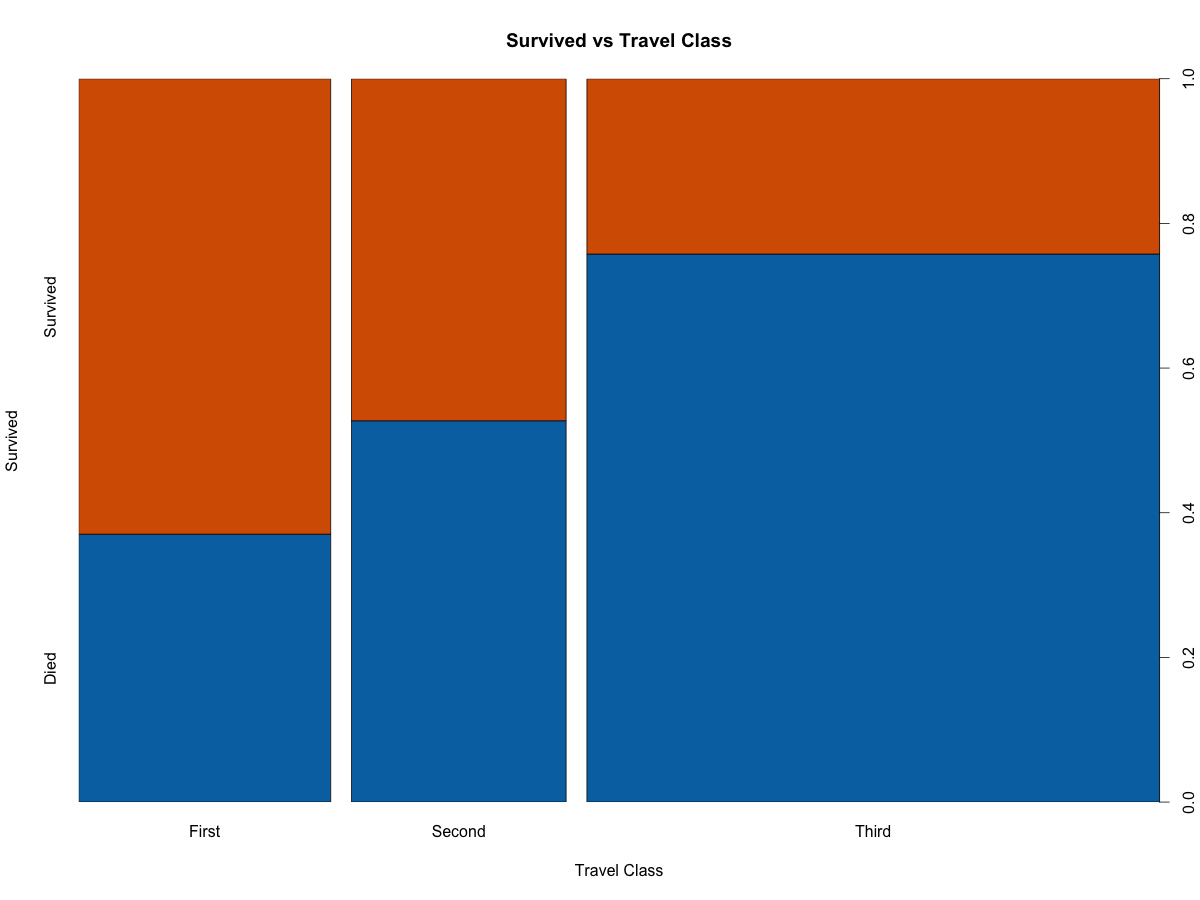


Below graph is another presentation of the above graph and it also shows the same pattern. Graph shows that survival chance for female is around 75% and for male it is just under 20%.

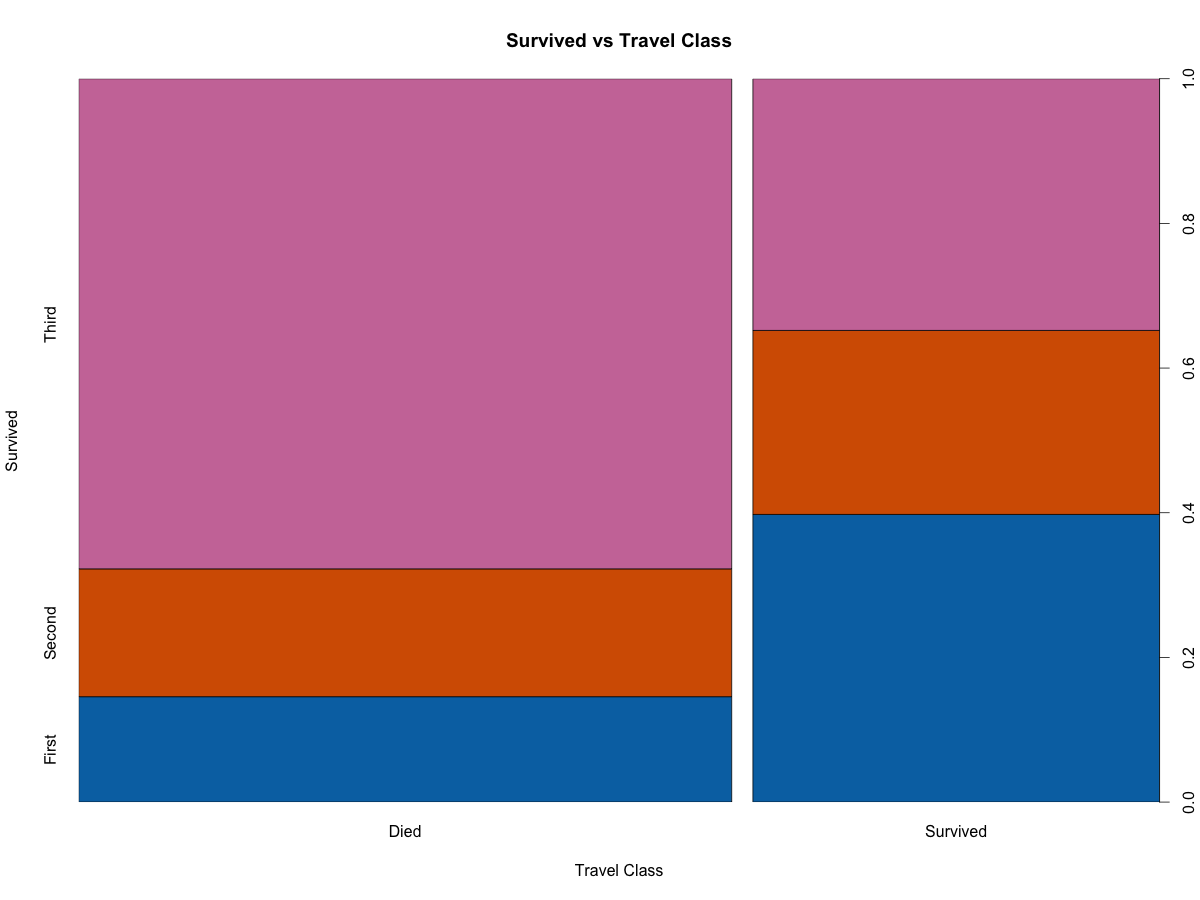


### Travel Class Vs Survival:

Travel class is divided into 3 parts First, Second and Third and these present the social status of the people. This again shows that the people in First class had much better chances of survival as compared to the people in 2nd and 3rd class. Only 22% people in 3rd class survived whereas for 1st class it is around 65%.

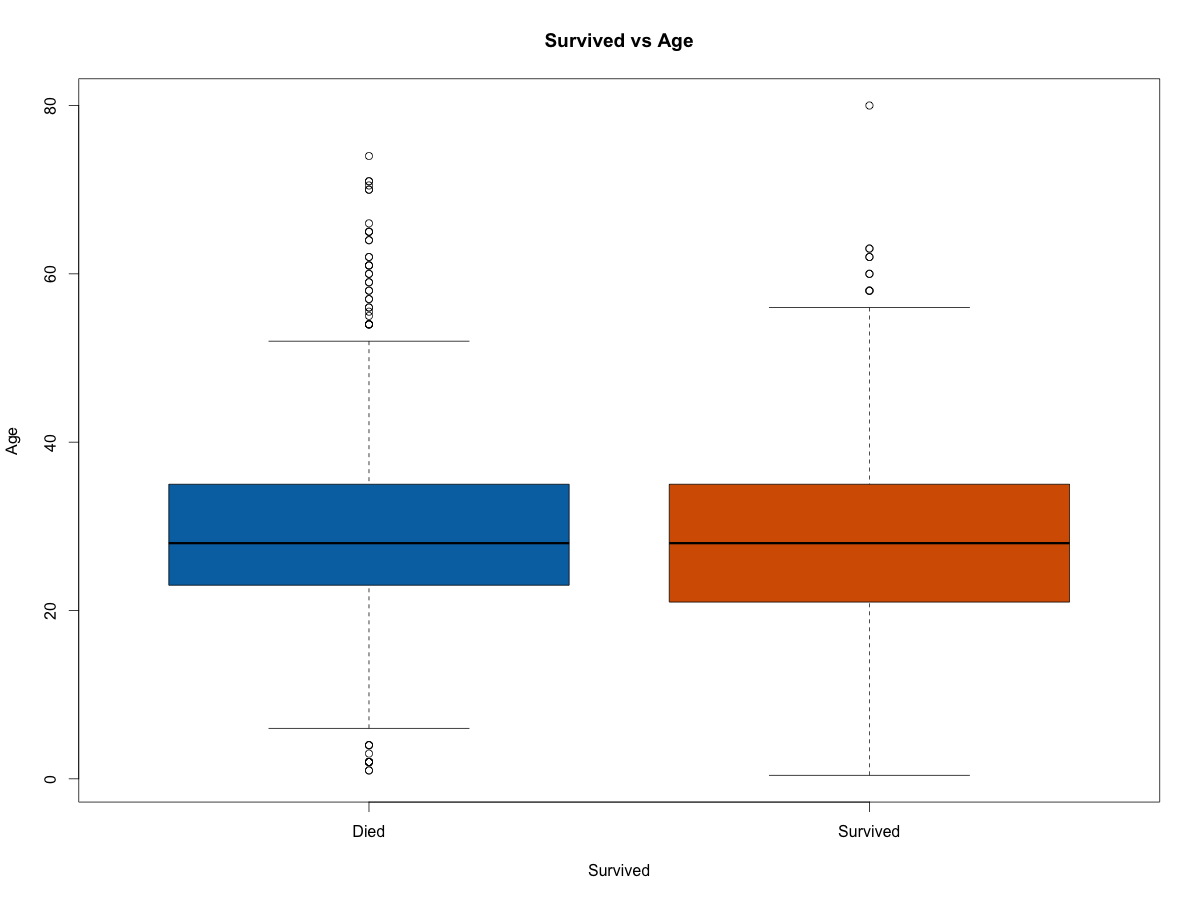


This is another view of the above graph. It shows similar information but in a bit different way. People of 3rd class were worst sufferers. Of the people survived 40% were from First class,25% from 2nd class and remaining 35% were from 3rd class.



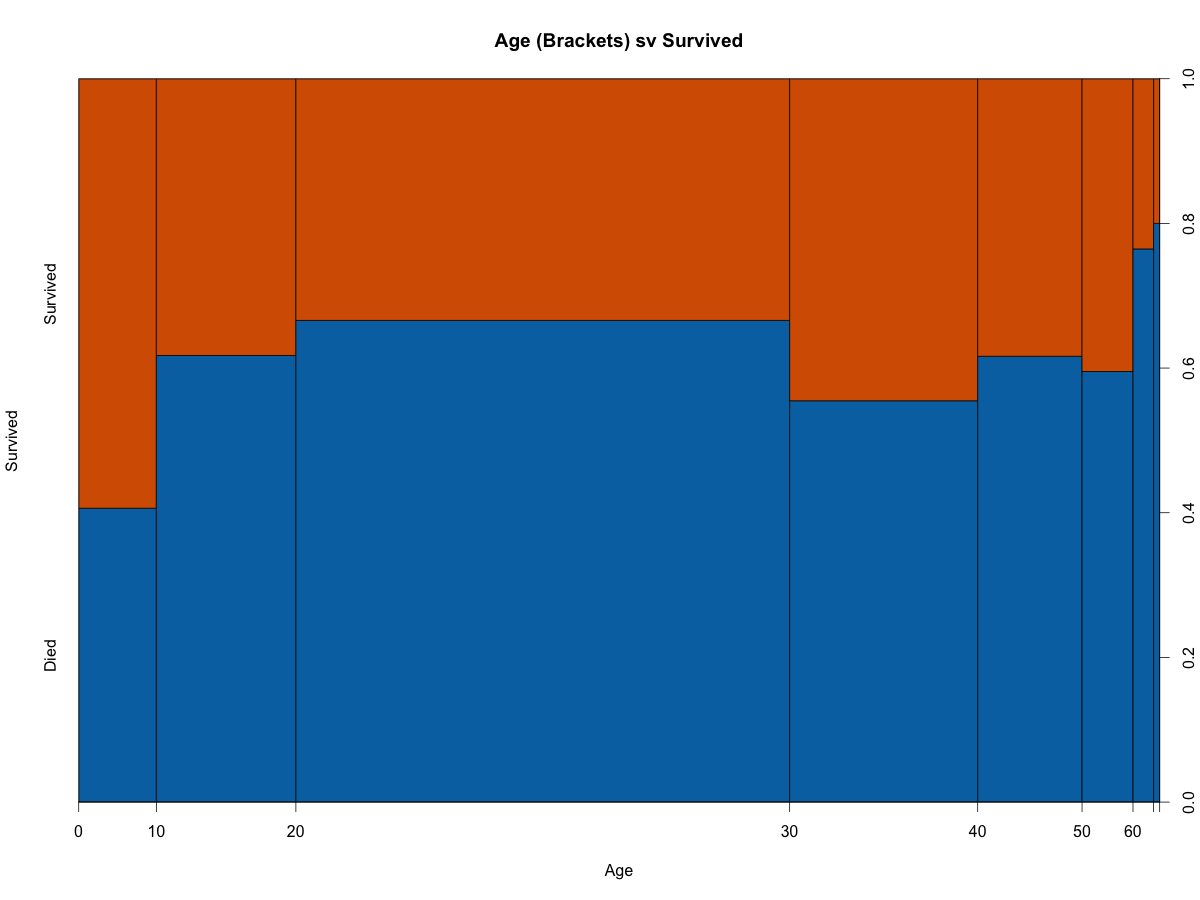
### Age Vs Survival:

Boxplot:



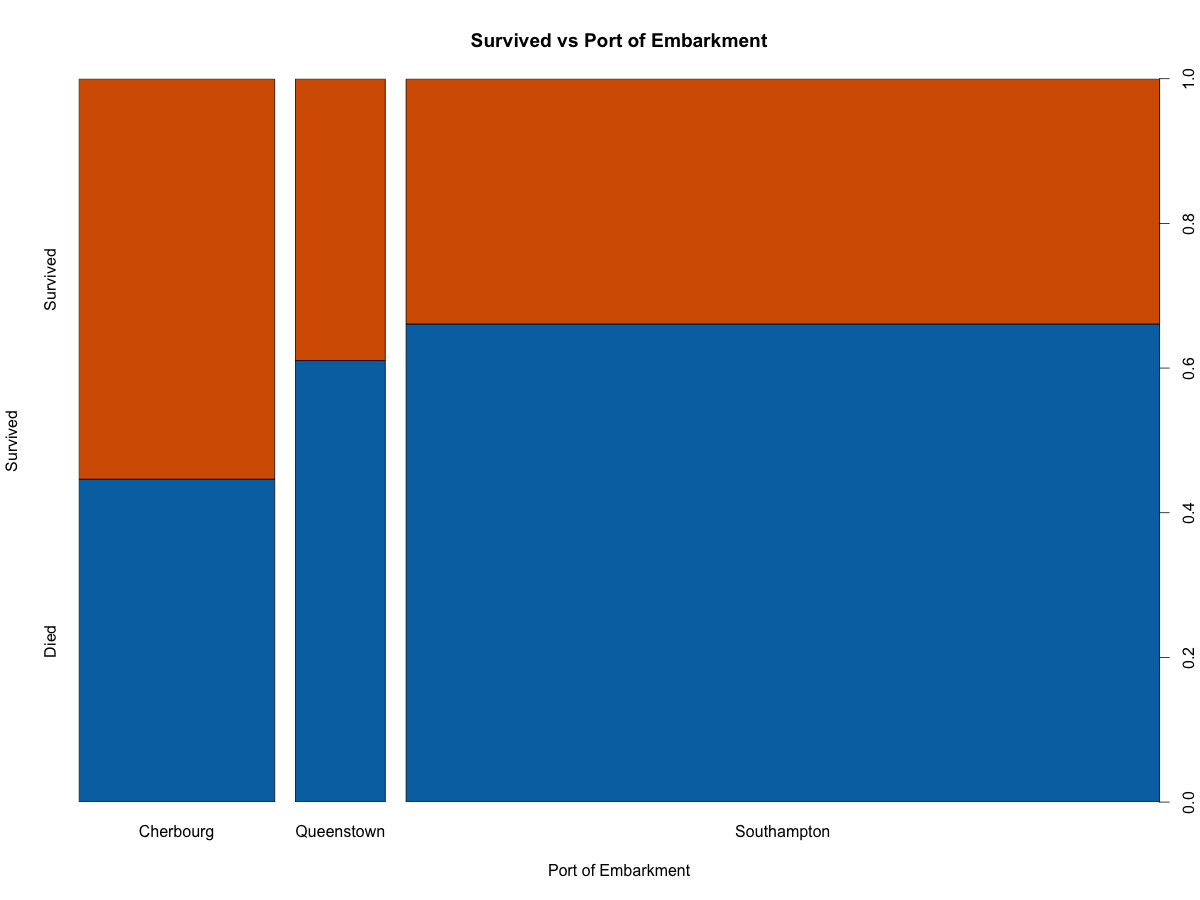
Boxplot is a nice way to see if there are some values of continuous variable above or below which we can distinguish 2 classes. There isn’t much here except that we could say that the young people had much better chance of survival, There are 2 things which are outlier, For people survived we could see that there are some who were more than 60 years. On the other hand we could see that there were kid who died.

Below graph shows the similar information but here age data is in brackets of 10 and width of bracket shows the number of people. Here we could see that people in range of 20-30 had very less chances along with people above 60 years.



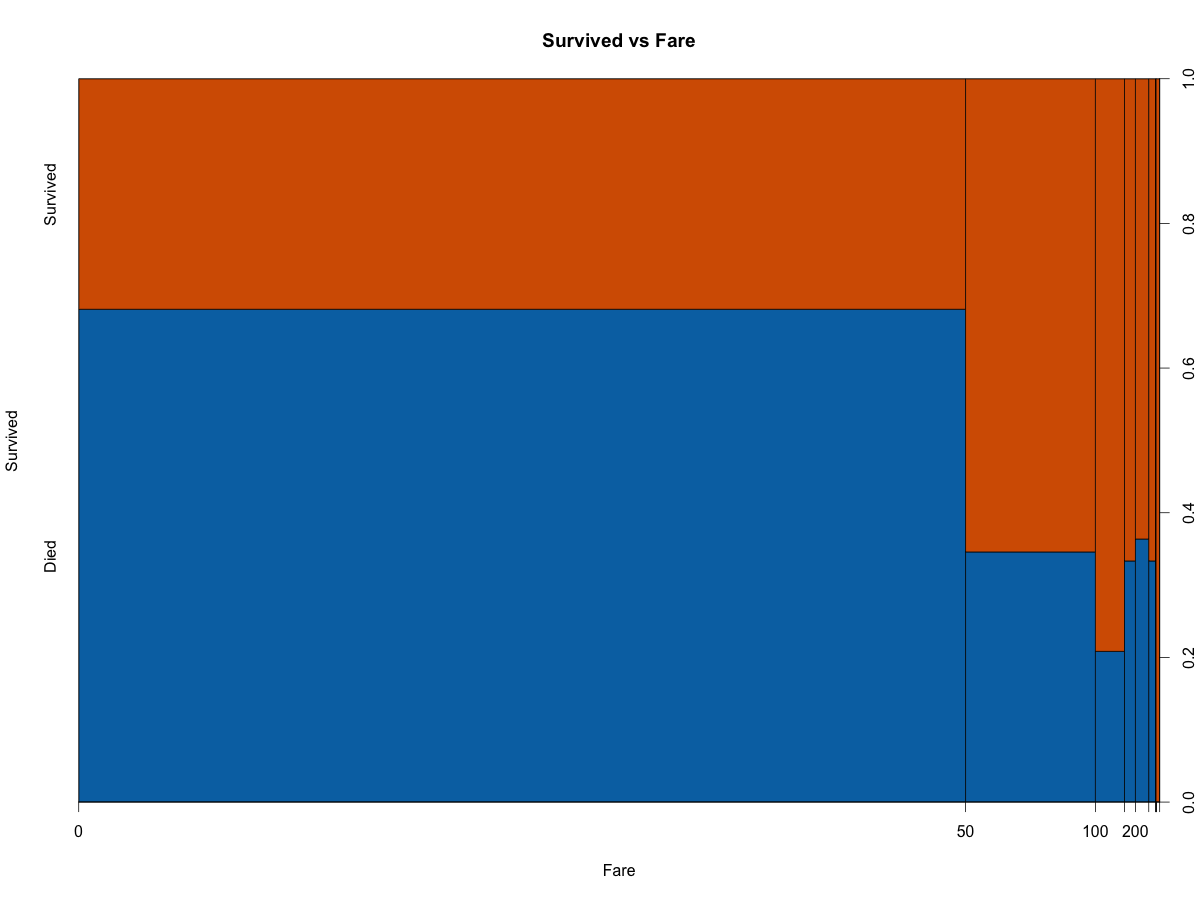
### Port of Embarkment Vs Survival:

This graph shows that people from Cherbourg had much better chance of survival followed by Queenstown. People from Southampton had only 30% chance of survival.



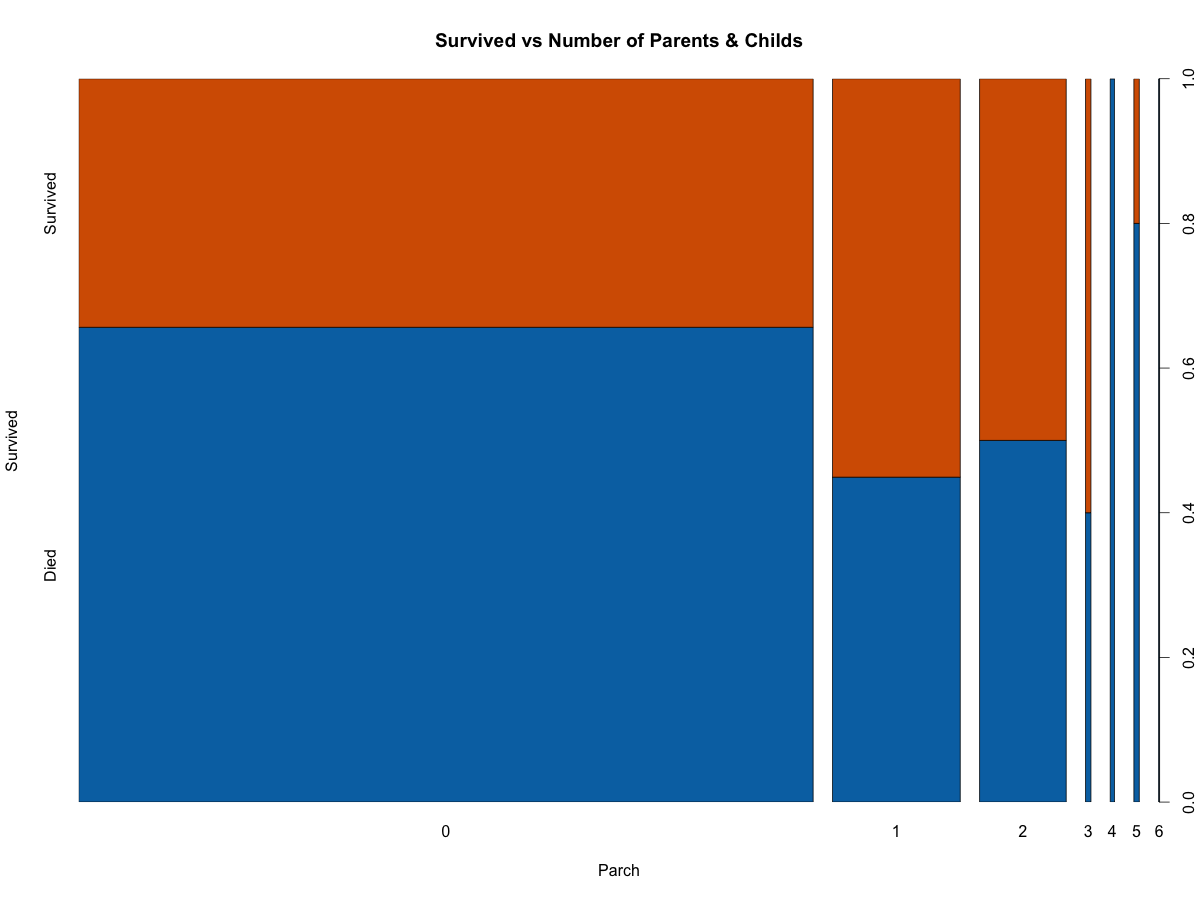
### Fare Vs Survival:

The fare is indirectly related to the Travel Class and thus it should show similar trend. Here the fare is divided into buckets of size 50(fare value). Again if you paid more money as fare chances of the survival were much better.



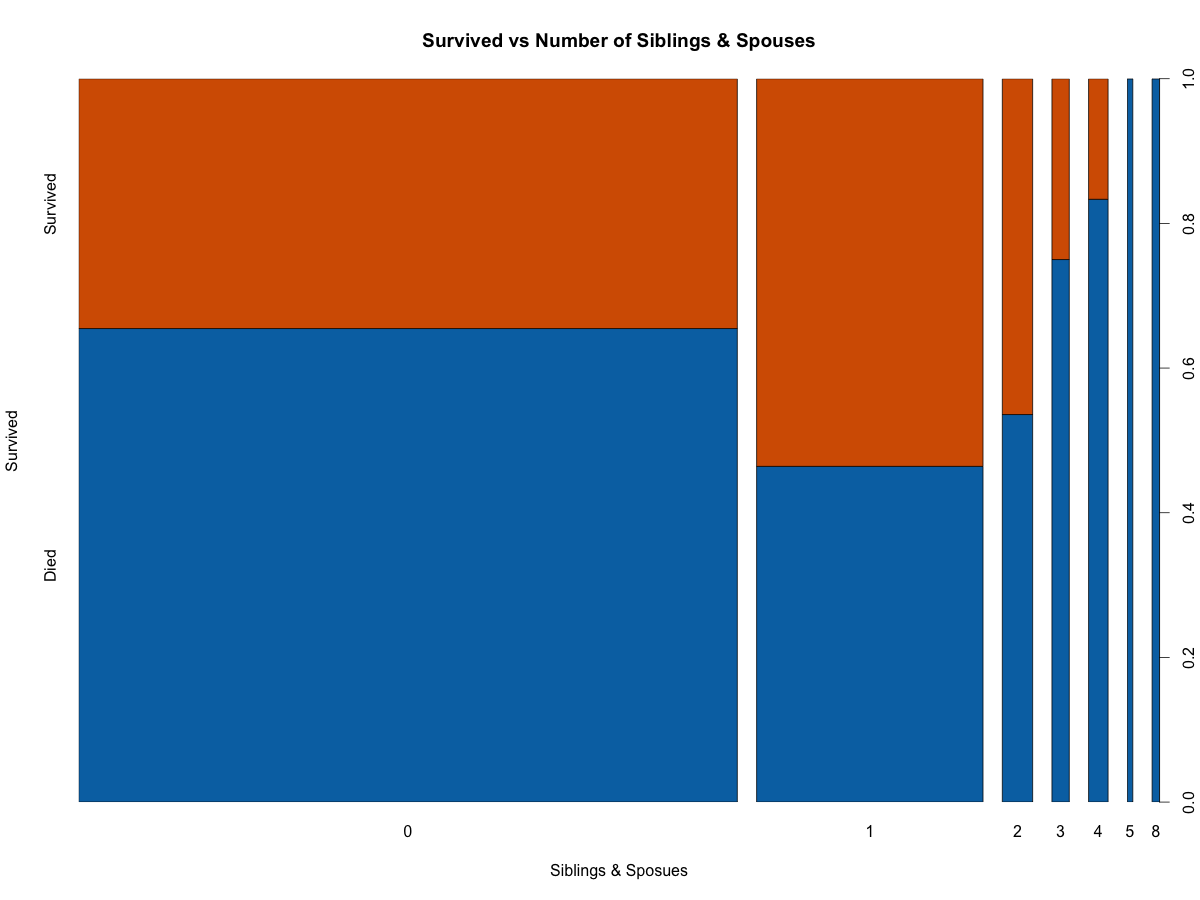
### Parent & Child Vs Survival:

Here it shows that if you had 1 or 2 number of Parent and Childs then chances of survival were much higher.



Siblings & Spouses vs Survival:

Here it shows that if you had 1 or 2 number of Spouse and Siblings then chances of survival were much higher.

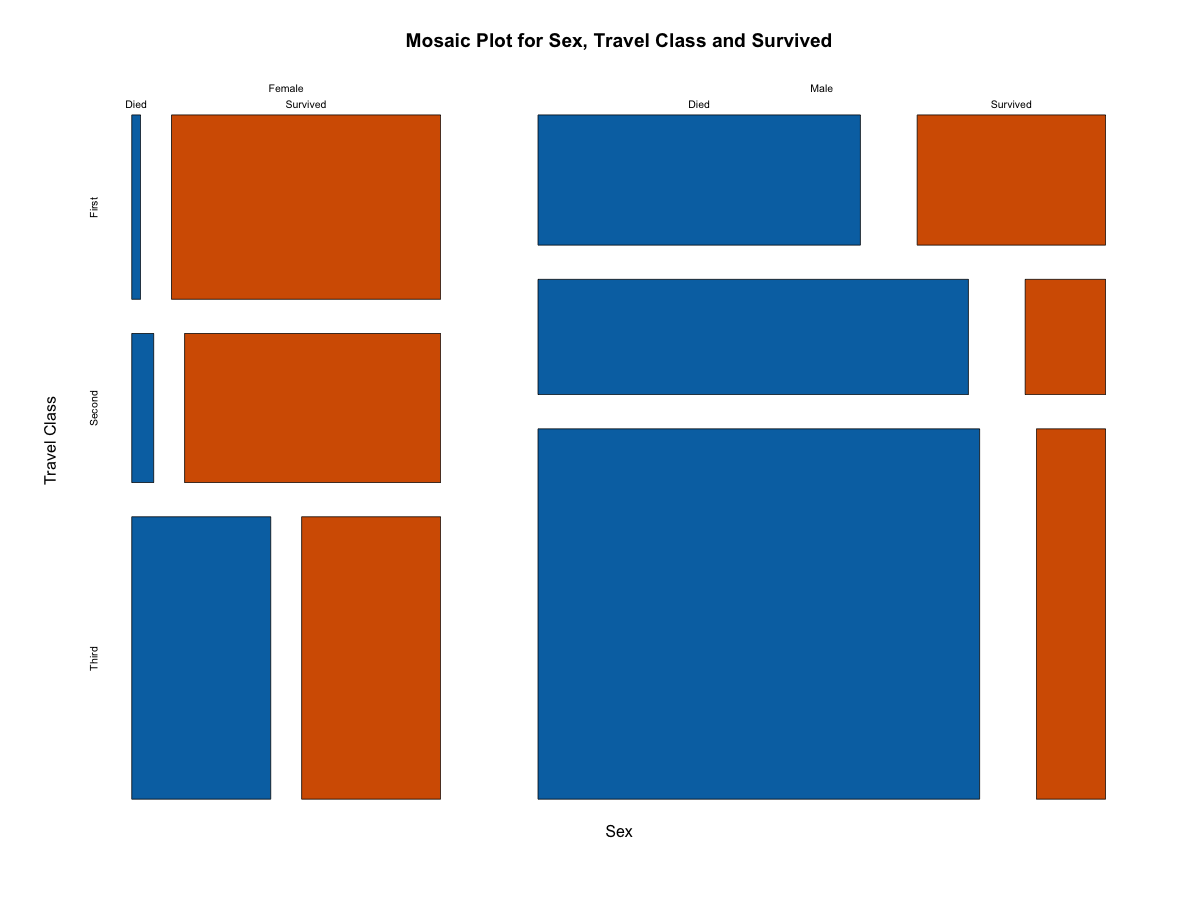


These graphs explain a lot of the data but it is quite possible that one of these fields is very important and that is actually causing other features to be quite important. E.g. Sex might be best feature and then say for Port of Embarkment where high survival rate maybe because in that Port of Embarkment most of person were Females. Also, we might see some pattern based on the feature interaction. Thus, we could see that we can use the interaction terms especially if we will be using the Logistic Regression (although I will not be using this model here).

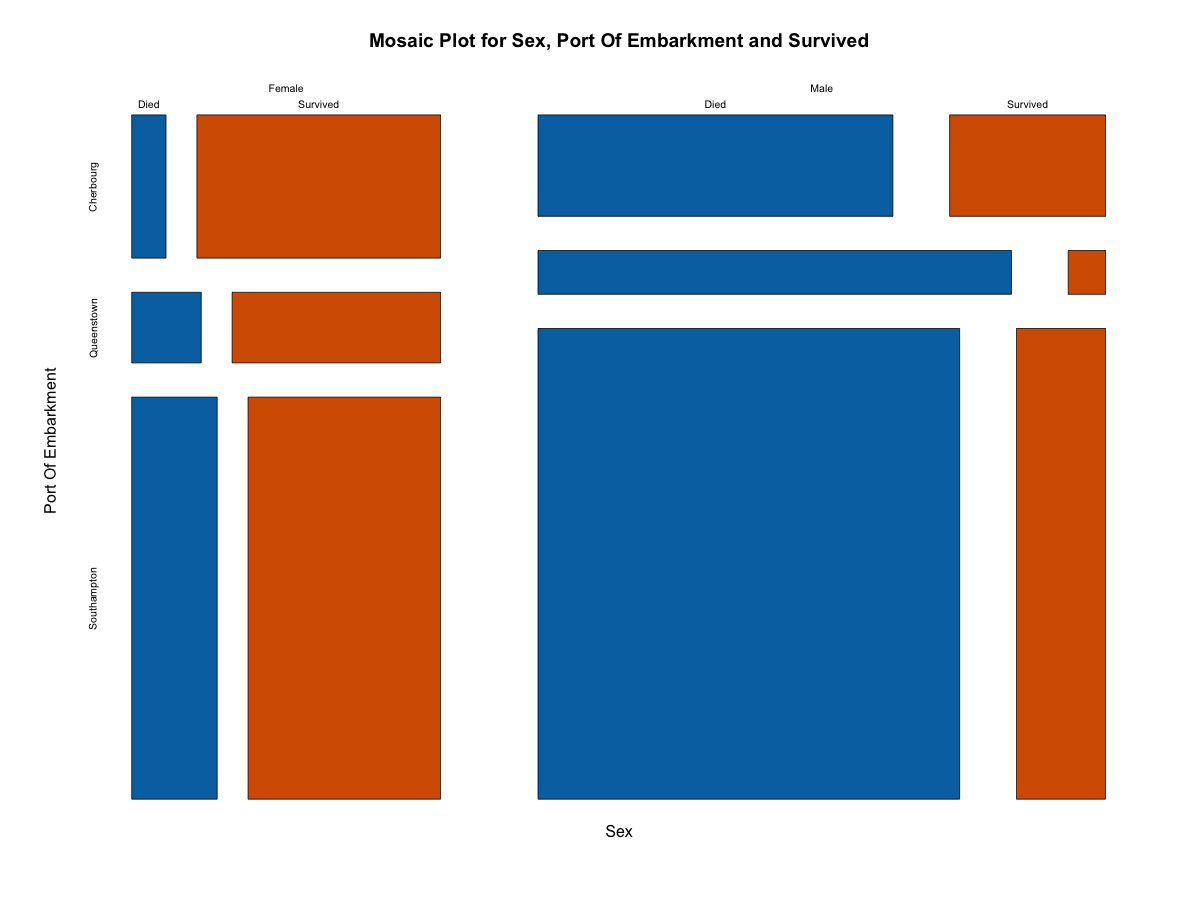
Thus, I wanted to lot some of the Mosaic graphs. I could have used the ggplot2 and plotted these graphs side by side. However, I personally think that in this case Mosaic plots will be much better.

1. Here I wanted to see how Travel Class and Sex data interact with each other. E.g. does Females of First class have more chances of survival than Females of Third Class?

In fact it seems that the Females of First class had much better survival rate than females in Third class. Males also follow the same pattern. Interesting here is that Females of Third class has much better survival rate than Males of the First class and thus it seems that Sex has more importance than Travel class.



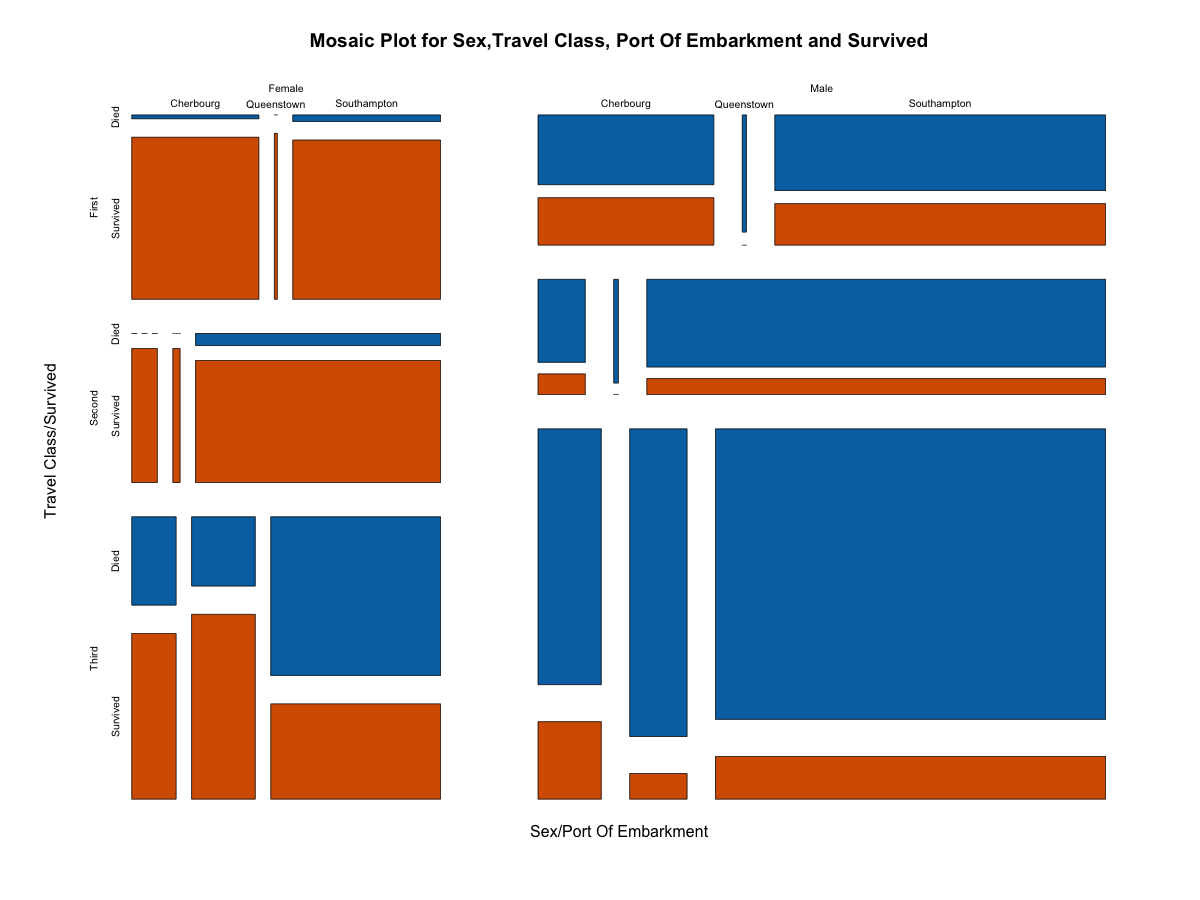
2. Below graph shows that the Port of Embarkment is an important one as well e.g. Males from Cherbourg has much better survival rate than Southampton.



1. Here, I used the Travel Class,sex and Port of Emabrakment.

There are some interesting facts here. E.g. Males from Queenstown had very less chance of survival irrespective of class. In fact only survived looks from 3rd class. There are just 2 people in First and Second class combined. Also, for males port of embarkment did not matter in the first and second class. However, there is a difference for 3rd class.

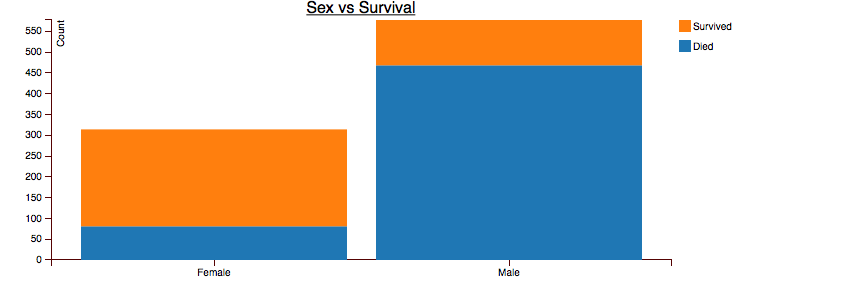
For females however, the port of embarkment had an impact. It was manily in 2nd and 3rd class.



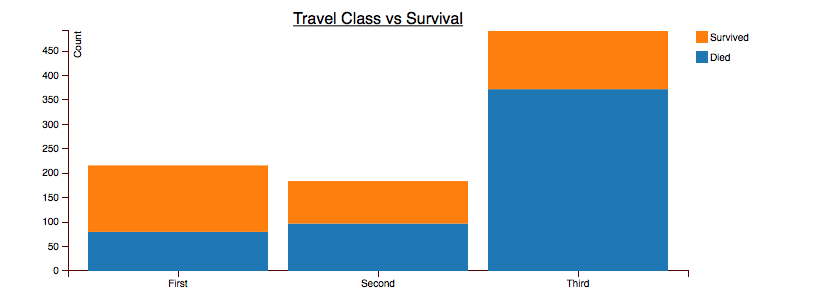
Code for this can be found in Data\_Viz.rmd and the output for the full code is in Data\_Viz.docx. You will find some other graph in this document. I did not include all of them here.

Finally, Here are some of D3 graphs. These are same graphs as we had seen above but these are using D3 and uses absolute count. First 3 are obvious but look at the graph 4. It shows that the Females in First class had very high chances of Survival. And worst are Males from Third class.

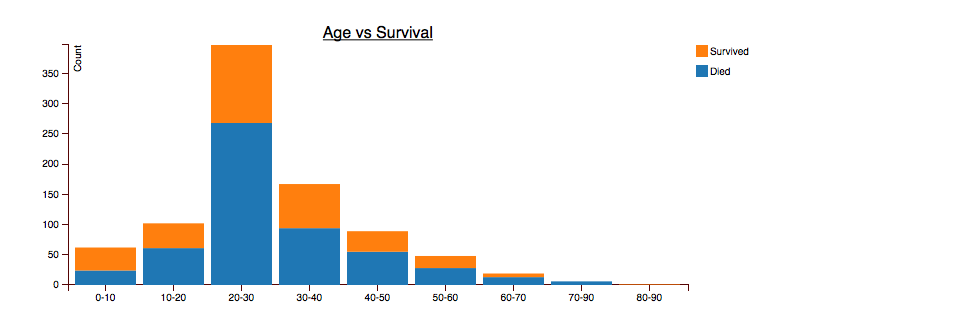
1.



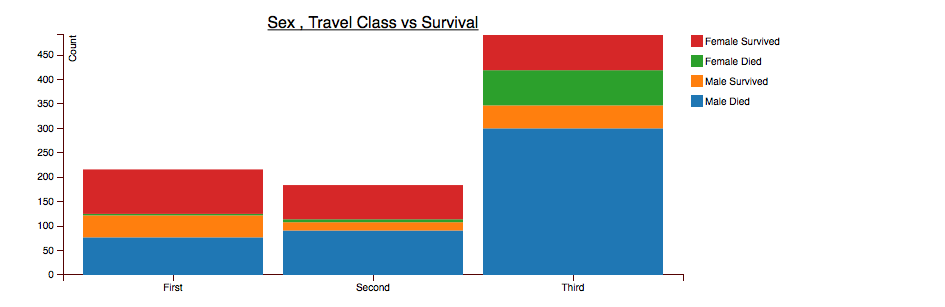
2.



3.



4.



These can be found at [6].

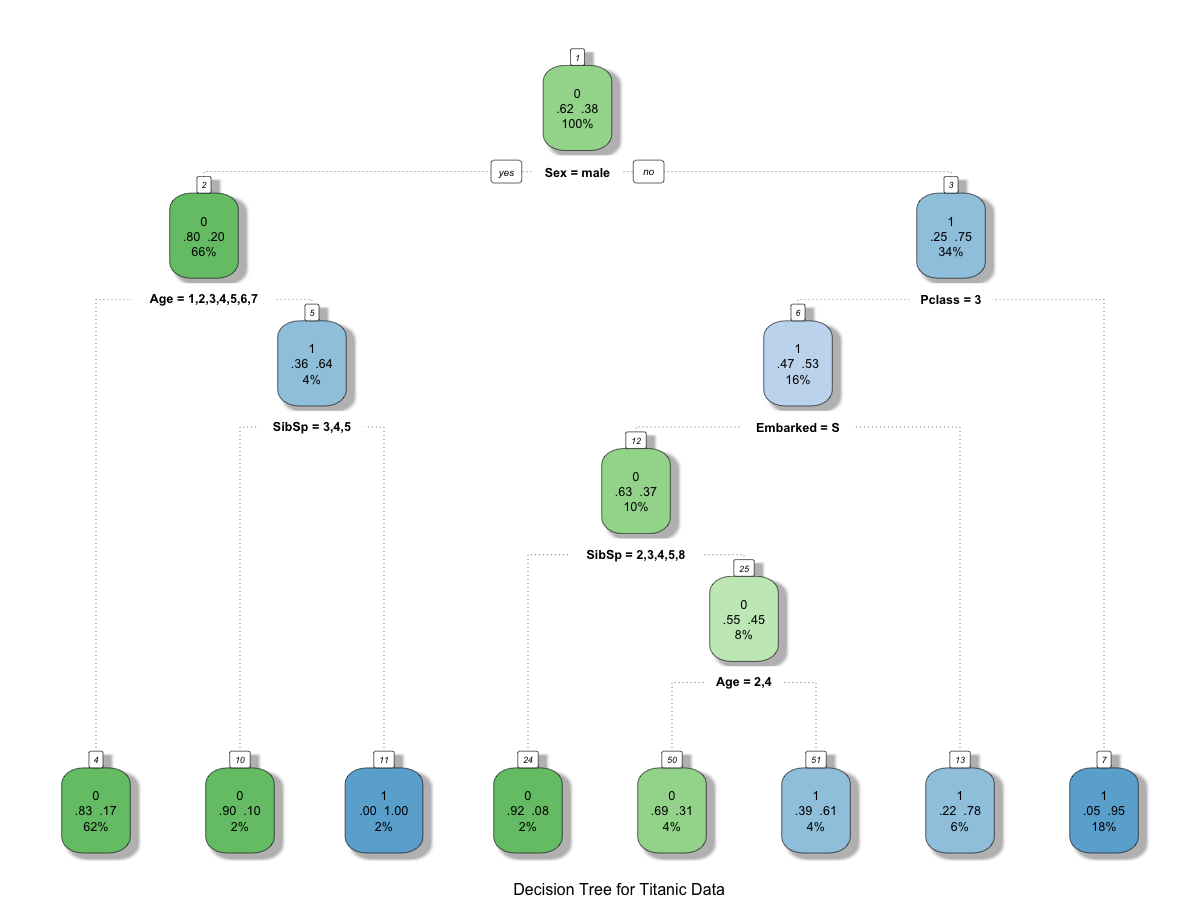
<http://www.doc.gold.ac.uk/~dsing001/TitanCharts.html>

## Model:

Modelling is final step where we use training data and test data. I have divided the data into training and test set. Training set has around 80% of actual training data. I concentrated the data vizualization in this case, thus I did not spend too much time on fine-tuning the model and thus achieve the maximum accuracy.

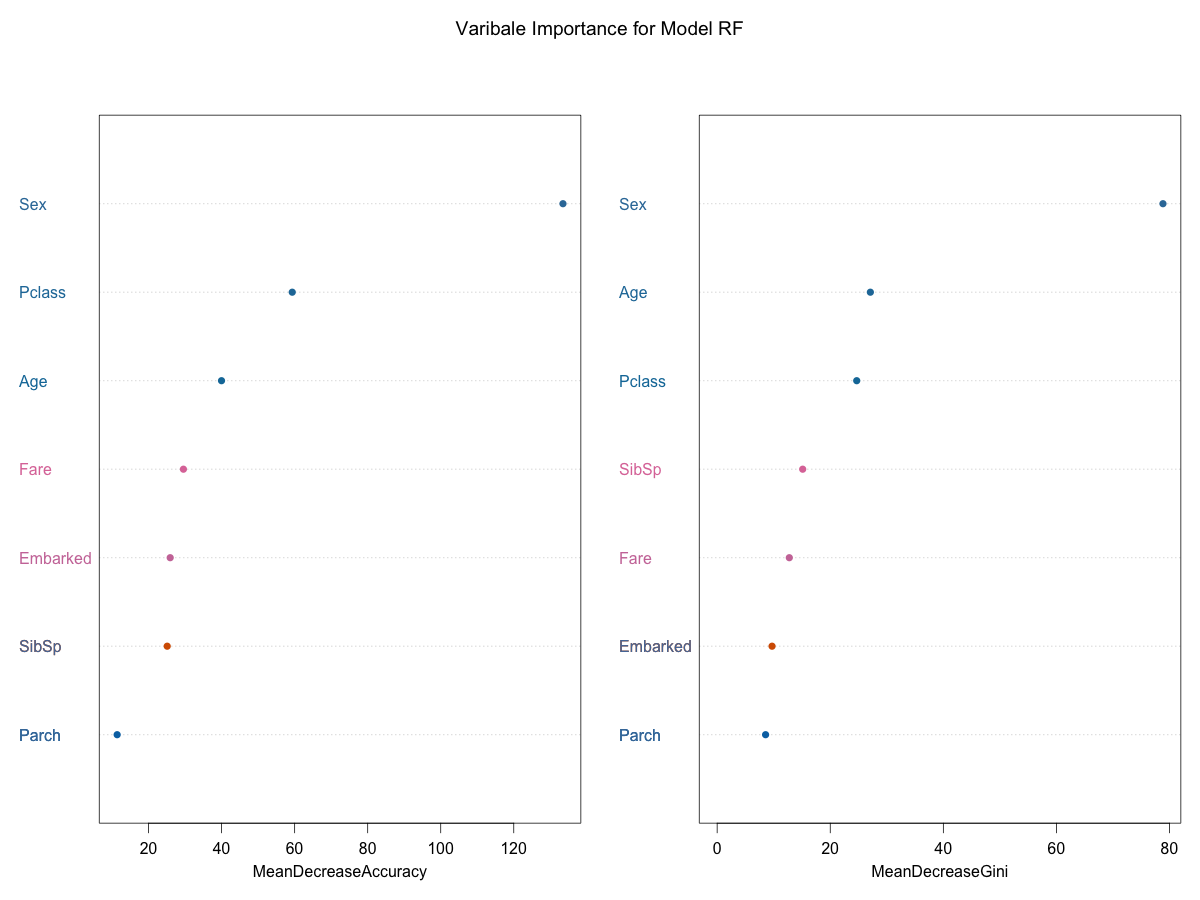
Finally, for the model I wanted to use Random Forest [5], which is an ensemble of various, decision Trees [4]. I have used ensemble of 1000 tress. There is slight diff how trees are generated, as these are random. However, I will generate a decision tree using decision tree model to show you how the decision tree looks.

Below is a generated Tree. It shows that Sex is most important features followed by Travel Class and Age and then port of Embarkment and Number of Siblings and Spouses.

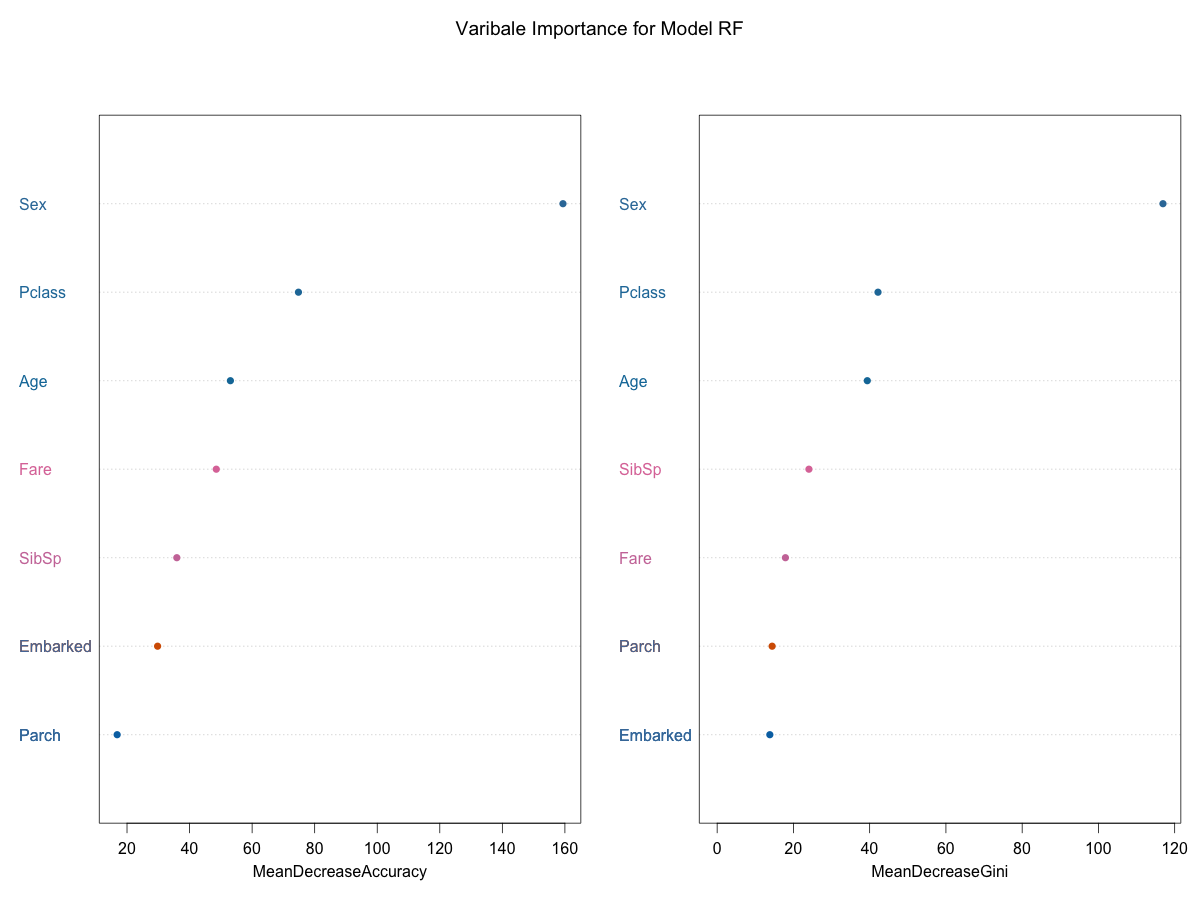


Random forest generates the tree but uses a bit of randomization and thus the trees might not be exactly same but they will look similar and there will be 1000 of them, which are more or less similar. In Random Forest, we can specify the importance of parameter which can provides us most important features.

Below is a graph, which shows parameter importance when only 80% of the trn data was used for the training. It indicates the same thing but Fare is higher than Embarked and SibSp.



Below is a graph, which shows parameter importance when full data set of training\_data data was used for the training.



### Results:

Accuracy on 20% validation data was around 83%. Full results can be seen in the output of the TTN\_Model.rmd output TTN\_Model.docx.

## Conclusion:

Feature selection and finding pattern are very important tasks and these takes can be accomplished either by using vizualition (for smaller data set and with limited number of variables) or by using some sort of machine learning algorithm to find out which are important features based on the data set. Then you can validated only say top 10 features and see how’s things looks like. The vizualition makes it easier to present and it is very important part of the final model.

I personally feel that as a data scientists R and python are much better vizualition tools and D3 can be used to represent the final output as in D3 you should know what you want to plot.

### Further work/Shortcomings:

I have not explored all the possible modelling techniques to achieve high accuracy. I will try to fine-tune the random forest parameters. Also, I have ignored the column names. Names have title and the title can be further used as a feature. This is called feature engineering [6] and I have not done this. Also, there are other ways to add more features or remove features and then run the machine-learning algorithm. I could have tried these things further but these techniques are more machine learning modelling rather than related to task at hand, which is more from visualization point of view.

In D3 graphs, I have used stacked bar chart [9] for showing interaction between Travel class and Sex w.r.t. Survived. I could have done a Merimekko chart [8]. It could have been much clearer like graphs generated in R.

**Challenges:**

I have not used much of Mosaic plots before this data analysis and I wanted to interpret these graphs accurately. Thus, it was bit difficult at first to interpret but after analysing the data both using graph and actual data I was able to grasp how to interpret these and slowly became good at it.

### Learning:

I have learned how to create the Stacked Bar Chart using D3 as well/Although, most important take away from here is that I am able to generate and interpret the Mosaic Plot especially when there are lots of variables involved (like 3 or 4). These graphs are nice way to see the pattern. I used to use ggplot2 and used to generate the graphs between 2 variables for each value in 3rd variable and then place them to next to each other and interpret them.

Also, I learnt how to generate a pretty decision tree. The standard package generates a simple (ugly) tree, which is hard to interpret sometimes.

## Code:

Code can be found at github at following location.

<https://github.com/darshanmeel/ttn>

## References:

1. <http://en.wikipedia.org/wiki/Sinking_of_the_RMS_Titanic>
2. <http://www.kaggle.com/c/titanic-gettingStarted>
3. <http://www.kaggle.com/c/titanic-gettingStarted/data>
4. <http://en.wikipedia.org/wiki/Decision_tree_learning>
5. <http://en.wikipedia.org/wiki/Random_forest>
6. <http://en.wikipedia.org/wiki/Feature_engineering>
7. <https://www.doc.gold.ac.uk/~dsing001/TitanCharts.html>
8. <http://bl.ocks.org/mbostock/1005090>
9. <http://bost.ocks.org/mike/bar/>
10. <http://www.d3noob.org/>