# Introduction:

Titanic was a disaster, which struck on night of 14th April1912 from England to New York. There were lost of people who survived and the challenge is to find out based on the data given in trai.csv to build a Machine Learning Model to predict whether a person would hav esurvied or died during the Titanic Crash. The data is provided.

# Background:

Kaggle team provided the data and explanation of each of the field is provided as well on the kaggle page. Please see below for more details on the data. Please see [2] 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.

# Data Analysis:

This is one of the most important part of the Building Machine learning models. This provides basic information about the data and data distribution as well as it provides whether there are some data missing.

Missing Data:

There are some of missing rows in Cabin, Ticket, Embarked and Age. I will impute the missing data for Age (median age) and Embarked(with S). I will not impute the data for remaining 2 columns, as I will not be using this in the Model.

I did some basic analysis using the table command in R and here are few observations by looking at the data.

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. After you run this the output of this can be found at Data\_Analysis.docx.

# Data Visualization:

Data Analysis above showed some nice patterns. Graphs are a nice way to look ate the pattern. Thus, I have run the various data graphs and here I will be including only the graphs which are most important from the Machine Leanrng modelliung point of view. I will specify the Graph and will try to explain in a line or 2 what information is being presented by the graph. PClass, Sex and Age were most important factors and Below are the graphs for these columns for survival and non survival.

1. Sex vs Survival:
2. Travel Class Vs Survival
3. Port of Embarkment Vs Survival
4. Fare Vs Survival
5. Parent & Child Vs Survival
6. Siblings & Spouses vs Survival

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 Embarkemnet where highs survival rate maybe 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 Regressiom (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.

Code for this can be found in Data\_Viz.rmd and the output for the full code is in Data\_Viz.docx. This document has only the graphs which are part of this document. I have not included the code where I created few graphs which were not important.

Finally, Here are some of D3 graphs. These graphs use the absolute count. Please look at the graph 4. It shows that the Females in First class had very high chances of Survival. These can be found at [6] <https://www.doc.gold.ac.uk/~dsing001/TitanCharts.html>

# Model:

Modelling is final step where we use training data and test data. Training data is around 70% and our model will be trained on this data and finally It will be tested on the test\_data and we will fine tuen the model based on how will be the performamce. I concentrated the data vizualation in this case, thus I did not spend too much time on fine tuning the models.

I wanted to use Decision Tree and Random Forest for my models. Thus, I ran the decision tree model to see how will it perform and most importantly which are some of the bestfeatures as decision tree is a method which can provide which are most important featires. I will check this with my findings in Data Vizualiation section.

Below is a generated Tree.

As you could see that the Sex is most important feature and followed by Pclass and Age and then Port of Embarkment.

Finaly, for the model. I wanted to use Random Forest which is an ensemble of various decision Trees. I have used ensemble of 1000 tress. There is slight diff how trees are generated as these are random.

Random forest provides which are most important features and below is the graph generated from the same. It shows similar ot what we got in Decision Tree an dwhat we got from Data Vizuaalizaion part.

Results: I got around 80% accuracy on this data. Below are results.

# 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 varibales) or by using some sort of Machine Learning Algorithm to find out which are important features based on the data set. Then you can valise only say top 10 features and see how’s things looks like. The vizaulation makes it easier to present and it is very important part of the final model.

## Further work:

Names have title and the title can be further used as a feature. This is called feature engineering and I have not done this. Also, there are other ways to add more feature or remove features and then run the Machine Leanring algorthim. I could have tried these things further but these are more Machien Leanring rather than

# References:

1. <http://en.wikipedia.org/wiki/Sinking_of_the_RMS_Titanic>
2. http://www.kaggle.com/c/titanic-gettingStarted
3. http://en.wikipedia.org/wiki/Decision\_tree\_learning
4. <http://en.wikipedia.org/wiki/Random_forest>
5. <http://en.wikipedia.org/wiki/Feature_engineering>
6. <https://www.doc.gold.ac.uk/~dsing001/TitanCharts.html>