Titanic: Disaster and Insights

Steven Voo Su Yee

May 2019

Table of Contents

[1.0 Introduction 2](#_Toc8746159)

[1.1 Data Overview 2](#_Toc8746160)

[1.2 Impute Missing Data 5](#_Toc8746161)

[2.0 Feature Engineering 6](#_Toc8746162)

[2.0.1 Social Titles of Passengers 6](#_Toc8746163)

[2.0.2 Family Size aboard 7](#_Toc8746164)

[2.0.3 Fare Bin 7](#_Toc8746165)

[2.1 Impute Missing Age 7](#_Toc8746166)

[2.1.1 Age Group 8](#_Toc8746167)

[3.0 Exploratory Data Analysis 9](#_Toc8746168)

[3.1 Relationships between Survival Rate and other Variables 10](#_Toc8746169)

[3.1.1 Sex 10](#_Toc8746170)

[3.1.2 Pclass 10](#_Toc8746171)

[3.1.3 AgeGroup 11](#_Toc8746172)

[3.1.4 Title 11](#_Toc8746173)

[3.1.5 Family 12](#_Toc8746174)

[3.1.6 IsAlone 13](#_Toc8746175)

[3.1.7 Embarked 13](#_Toc8746176)

[3.1.8 Fare 14](#_Toc8746177)

[3.2 Relationships between Survival Rate and other Variables with Interaction 15](#_Toc8746178)

[3.2.1 Sex and Pclass 15](#_Toc8746179)

[3.2.2 AgeGroup and Pclass 15](#_Toc8746180)

[3.2.3 Sex and AgeGroup 16](#_Toc8746181)

[3.2.4 Title and Pclass 16](#_Toc8746182)

[3.2.5 IsAlone and Embarked 17](#_Toc8746183)

[3.2.6 Sex, Pclass, and IsAlone 17](#_Toc8746184)

[3.2.7 Sex, AgeGroup, and Pclass 18](#_Toc8746185)

[3.2.8 Sex, AgeGroup, Pclass, and IsAlone 19](#_Toc8746186)

[3.3 Correlation 20](#_Toc8746187)

[4.0 Predictive Modeling 20](#_Toc8746188)

[4.1 Logistic Regression 21](#_Toc8746189)

[4.2 k-Nearest Neighbour 22](#_Toc8746190)

[4.3 Decision Tree 23](#_Toc8746191)

[4.4 Random Forest 24](#_Toc8746192)

[5.0 Conclusion 25](#_Toc8746193)

# 1.0 Introduction

RMS Titanic was a British passenger liner that sank in the North Atlantic Ocean in the early morning hours of 15 April 1912, after it collided with an iceberg during its maiden voyage from UK to New York City. There were an estimated 2224 passengers and crew aboard the ship, and more than 1500 died, making it one of the deadliest commercial peacetime maritime disasters in modern history. Below are the journey map of Titanic:

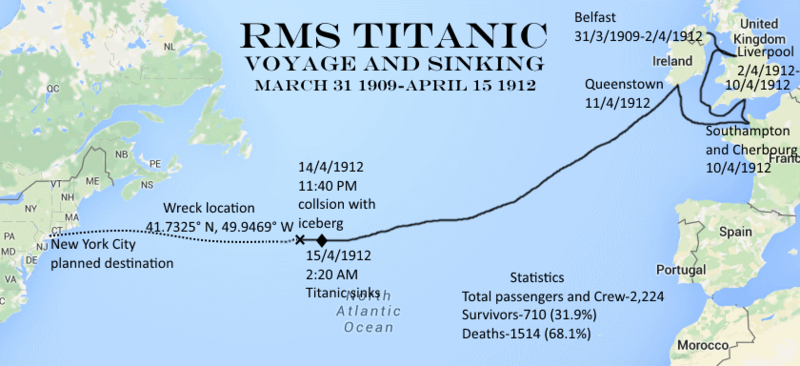


Fig 1: Titanic Journey Map

## 1.1 Data Overview

There are 891 observations and 12 variables in this dataset. This dataset consist only 40% of the overall 2254 passengers and crew on-board of Titanic.

## 'data.frame': 891 obs. of 12 variables:  
## $ PassengerId: int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Survived : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...  
## $ Pclass : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...  
## $ Name : chr "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...  
## $ Sex : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...  
## $ Age : num 22 38 26 35 35 NA 54 2 27 14 ...  
## $ SibSp : int 1 1 0 1 0 0 0 3 0 1 ...  
## $ Parch : int 0 0 0 0 0 0 0 1 2 0 ...  
## $ Ticket : chr "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...  
## $ Fare : num 7.25 71.28 7.92 53.1 8.05 ...  
## $ Cabin : chr NA "C85" NA "C123" ...  
## $ Embarked : Factor w/ 3 levels "C","Q","S": 3 1 3 3 3 2 3 3 3 1 ...

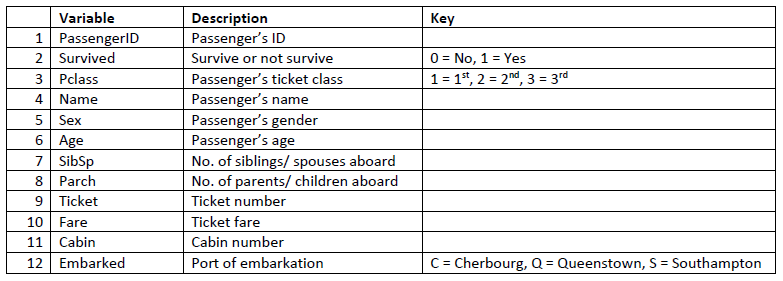


Fig 2: Variables Description

Above are the descriptions of these variables. There are some extra explanations for these variables:

* 1st class ticket means the fare is the most expensive amongst the 3 classes.
* The age is actually fractional if less than 1 and if it is estimated, it is in the form of xx.5
* Siblings also include step-siblings.
* Children also include step-children.

The missing values in each variable are shown next:

## PassengerId Survived Pclass Name Sex Age   
## 0 0 0 0 0 177   
## SibSp Parch Ticket Fare Cabin Embarked   
## 0 0 0 0 687 2

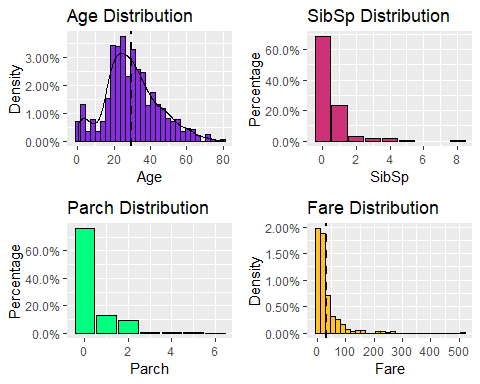
The variable **‘Age’** and **‘Cabin’** are missing a substantial amount of values. **‘Embarked’** has 2 missing values.

**Summary of the data:**

## PassengerId Survived Pclass Name Sex   
## Min. : 1.0 0:549 1:216 Length:891 female:314   
## 1st Qu.:223.5 1:342 2:184 Class :character male :577   
## Median :446.0 3:491 Mode :character   
## Mean :446.0   
## 3rd Qu.:668.5   
## Max. :891.0   
##   
## Age SibSp Parch Ticket   
## Min. : 0.42 Min. :0.000 Min. :0.0000 Length:891   
## 1st Qu.:20.12 1st Qu.:0.000 1st Qu.:0.0000 Class :character   
## Median :28.00 Median :0.000 Median :0.0000 Mode :character   
## Mean :29.70 Mean :0.523 Mean :0.3816   
## 3rd Qu.:38.00 3rd Qu.:1.000 3rd Qu.:0.0000   
## Max. :80.00 Max. :8.000 Max. :6.0000   
## NA's :177   
## Fare Cabin Embarked   
## Min. : 0.00 Length:891 C :168   
## 1st Qu.: 7.91 Class :character Q : 77   
## Median : 14.45 Mode :character S :644   
## Mean : 32.20 NA's: 2   
## 3rd Qu.: 31.00   
## Max. :512.33   
##

* There are 342 out of 891 people (38.38%) survived in this sample of data, which is slighly more than the reported 31.9% survival rate.
* Majority of the passengers were travelling with 3rd class ticket (55.11%), whereas 20.65% with 2nd class ticket and 24.24% with 1st class ticket.
* Majority of the passengers were male (577 out of 891).
* Majority of the passengers embarked from Southampton port (72.28%), followed by Cherbourg (18.86%) and Queenstown (8.64%).

As for numerical variables such as **Age**, **SibSp**, **Parch**, and **Fare**, histograms will be used to visualise the distributions of these variables.



* The distribution of age is positively skewed, majority of the passengers were between 20 to 40 years old.
* Majority of the passengers did not aboard along siblings, spouses, parents, and children.
* Interestingly, majority of the fare price is 0, probably those are crew or a lot of passengers were invited to aboard Titanic for its maiden voyage for free.

## 1.2 Impute Missing Data

The variables **‘Age’**, **‘Cabin’**, and **‘Embarked’** have missing values. **‘Cabin’** will be dropped from the data since it has 687 missing values. As for the **‘Age’**, it has 177 missing values and we will deal with that later. In this section, **‘Embarked’** which only has 2 missing values will be imputed. Below shows which observations have the missing **‘Embarked’** value:

## PassengerId Survived Pclass Name  
## 62 62 1 1 Icard, Miss. Amelie  
## 830 830 1 1 Stone, Mrs. George Nelson (Martha Evelyn)  
## Sex Age SibSp Parch Ticket Fare Cabin Embarked  
## 62 female 38 0 0 113572 80 B28 <NA>  
## 830 female 62 0 0 113572 80 B28 <NA>

We can see that both of the passengers are female, traveled with 1st class ticket and both paid 80 for the fares. The fare might be able to tell which ports they were embarked on.

## Embarked Pclass MedianFare  
## <fct> <fct> <dbl>  
## 1 C 1 78.3   
## 2 C 2 24   
## 3 C 3 7.90  
## 4 Q 1 90   
## 5 Q 2 12.4   
## 6 Q 3 7.75  
## 7 S 1 52   
## 8 S 2 13.5   
## 9 S 3 8.05

Median fare value is used since the distribution of fare is irregular. Fare of 80 is very close to the median fare value of 1st class ticket embarked from Cherbourg. We will impute the 2 missing value with Cherbourg.

## PassengerId Survived Pclass Name  
## 62 62 1 1 Icard, Miss. Amelie  
## 830 830 1 1 Stone, Mrs. George Nelson (Martha Evelyn)  
## Sex Age SibSp Parch Ticket Fare Cabin Embarked  
## 62 female 38 0 0 113572 80 B28 C  
## 830 female 62 0 0 113572 80 B28 C

The missing values in **‘Age’** will be dealt in later section.

# 2.0 Feature Engineering

Feature enginnering is an approach to create additional relevant features (variables) from the existing raw data so that more insights can be exploited and then the predictive power of the predictive models can be increased.

### 2.0.1 Social Titles of Passengers

Notice that the **Name** variable does not only contain the first, middle and last names of the passengers, it also contains the social titles of the passengers (Mr., Mrs., Miss., Col., etc.). These titles related to the passengers’ gender, age, marital status, and contains military ranks, clergy, Royal and other Nobles. These titles might be helpful in representing the social status of the passengers and it may help to predict if a person with Noble status is more likely to survive or not. Below are all the titles that can be extracted in this dataset:

##   
## Capt Col Don Dr Jonkheer Lady Major Master Miss Mlle Mme Mr Mrs  
## female 0 0 0 1 0 1 0 0 182 2 1 0 125  
## male 1 2 1 6 1 0 2 40 0 0 0 517 0  
##   
## Ms Rev Sir the Countess  
## female 1 0 0 1  
## male 0 6 1 0

There are various social titles and we want to group them into few representative categories. First off, Master is a title used to address young boys who are not old enough to be addressed as Mister (Mr.) and since there are 40 of them, we will maintain **Master** as a group itself. **Mr** is also a group itself. Next, it is hard to classify equivalent young ladies since there isn’t general title for them. So we can classify the women into married and unmarried. Miss, Ms, and Mlle (Mademoiselle in French) are to be classified as **Miss** (unmarried women) whereas Mrs and Mme (Madame in French) are to be classified as **Mrs** (married women). Lastly, the military, clergy, Royal, and other Noble titles are to be grouped together as **Noble**.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Master | Miss | Mr | Mrs | Noble |
| female | 0 | 185 | 0 | 126 | 3 |
| male | 40 | 0 | 517 | 0 | 20 |

### 2.0.2 Family Size aboard

There are **SibSp** and **Parch** indicate how many siblings, spouses, parents and children a passenger traveled aboard with. We want to create another feature which combine both of these variables to indicate the total family members included the passenger traveled aboard the Titanic, called **FamSize**.

df$FamSize <- df$SibSp + df$Parch + 1

With that feature, we can also create another feature to indicate whether the passenger traveled alone, called **IsAlone** where IsAlone = 1 if FamSize = 1 and IsAlone = 0 if FamSize > 1.

df$IsAlone <- if\_else(df$FamSize > 1, 0, 1)  
df$IsAlone <- as.factor(df$IsAlone)

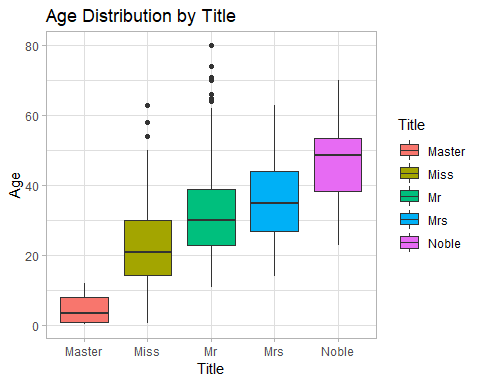
### 2.0.3 Fare Bin

The Fare is a continuous variable, so we want to classify the Fare into few bins (groups). qcut function is used so that the Fare is grouped by its quartile ranges.

df$FareBin <- qcut(df$Fare, cuts = 4)  
df$FareBin <- as.factor(df$FareBin)

## 2.1 Impute Missing Age

There are 177 missing values in Age variable. Just by replacing with the mean/ median age might not be the best solution because the age might be differ by categories of passengers. The social title might provides clues to impute the missing age.



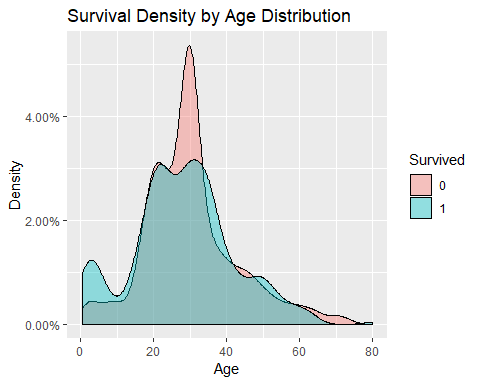
## # A tibble: 5 x 2  
## Title MedianAge  
## <fct> <dbl>  
## 1 Master 3.5  
## 2 Miss 21   
## 3 Mr 30   
## 4 Mrs 35   
## 5 Noble 48.5

From the boxplot above, there is a distinction in distribution of age by different titles. Median age of each title will be imputed to the missing values according to the passengers’ titles.

df$Age[df$Title == "Master" & is.na(df$Age)] <- 3.5  
df$Age[df$Title == "Miss" & is.na(df$Age)] <- 21  
df$Age[df$Title == "Mr" & is.na(df$Age)] <- 30  
df$Age[df$Title == "Mrs" & is.na(df$Age)] <- 35  
df$Age[df$Title == "Noble" & is.na(df$Age)] <- 48.5

### 2.1.1 Age Group

Now that the missing ages are imputed, we want to look at the age distribution by survival status to glean any insights.



Those younger than aorund 16 had a high survival rate which provides good information. The age band between 24 and 80 shows survival is unfavourable. It is not very informative as most passengers did not survive to begin with. We will classify the passengers in this new feature **AgeGroup** by those who were 16 years and below and those who were not.

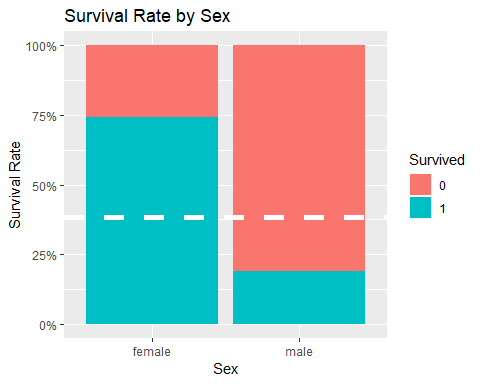
df$AgeGroup <- if\_else(df$Age <= 16, "<=16", ">16")  
df$AgeGroup <- as.factor(df$AgeGroup)

# 3.0 Exploratory Data Analysis

With data processing and cleaning done, We now want to explore the data more deeply with the help of visualisation tools.

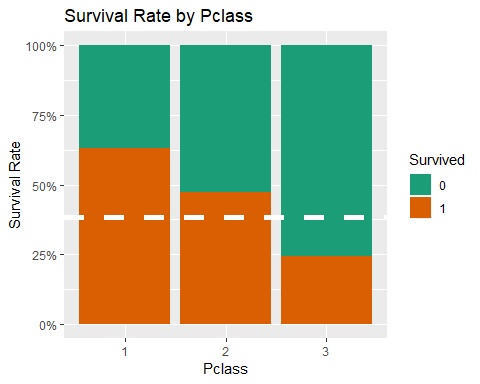
## 3.1 Relationships between Survival Rate and other Variables

### 3.1.1 Sex



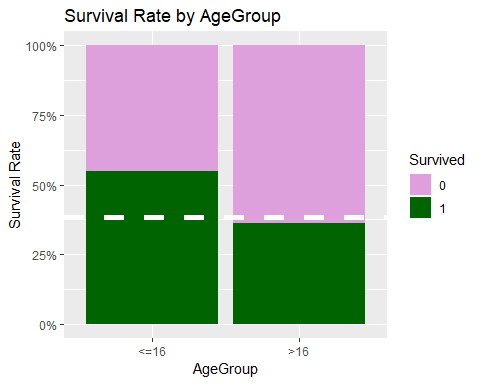
74.2% of females survived the disaster which is far above the mean survival rate of 38.4%, whereas only 18.9% of males survived which is far below the mean survival rate. This shows females survived at a higher rate than the males.

### 3.1.2 Pclass



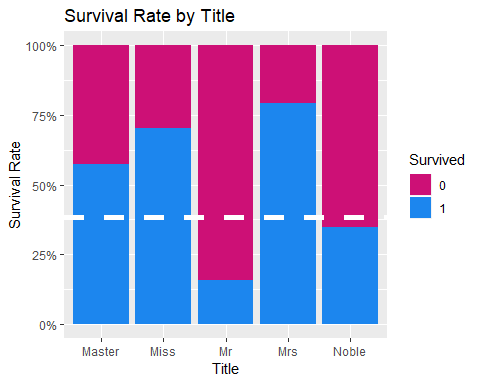
63% of passengers traveled with 1st class ticket survived, 47.3% of passengers traveled with 2nd class ticket survived, and only 24.2% of passengers traveled with 3rd class ticket survived, compared to the mean survival rate of 38.4%.

### 3.1.3 AgeGroup



Children and teens aged 16 and below survived at 54.8%. Passengers who were above 16 years old survived at 36.2% which is closed to the mean survival rate.

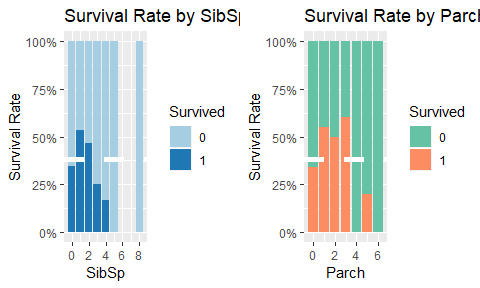
### 3.1.4 Title

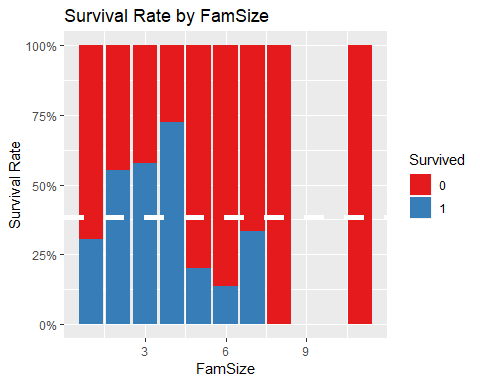


The survival rate of Miss (70.3%) and Mrs (79.4%) are the highest while the survival rate of Mr (15.7%) is the lowest, which in line with the finding that females have a very much higher survival rate than males. The survival rate of Master (57.5%) is also in line with the finding that children survived at higher rate. Passengers with Noble title have a survival rate of 34.8%.

### 3.1.5 Family

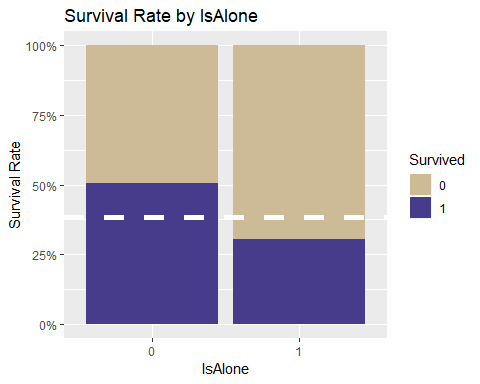
In this section, we will look at **SibSp**, **Parch**, and **FamSize**.





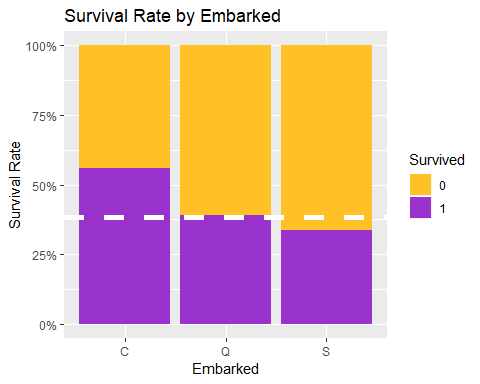
There is no clear takeaway from how family size could affect the survival rate.

### 3.1.6 IsAlone



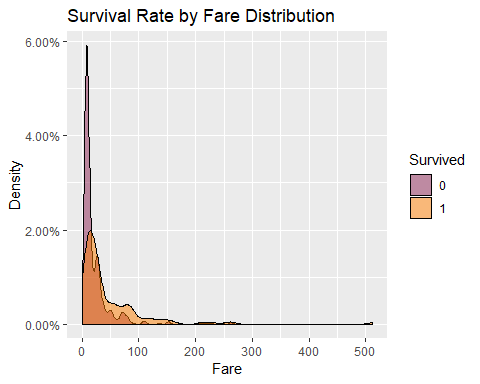
Passengers who were traveling alone had a lower survival rate (30.4%) than those traveling with family members (50.6%).

### 3.1.7 Embarked

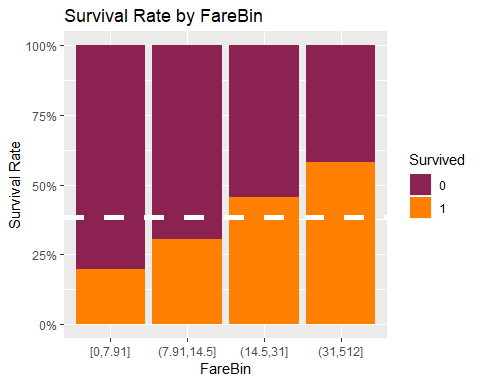


Those embarked on Cherbourg had the highest survival rate (55.9%), followed by those emabrked on Queenstown (39%) and Southampton (33.7%).

### 3.1.8 Fare



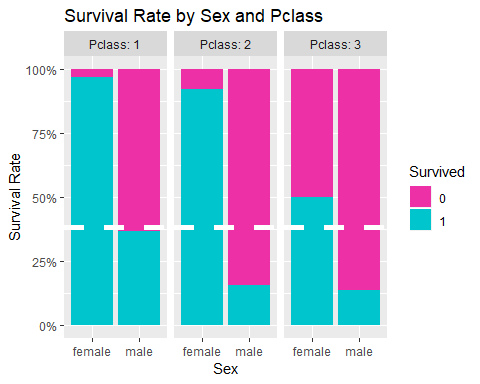
The survival rate increased as the fare is getting higher. We will explore this by looking at the breakdown of fare groups (FareBin).



Those who bought the fare for 31 and above survived at the highest rate (58.1%) followed by those bought the fare for 14.5 - 31 (45.5%). Those who bought the fare below 7.91 survived at the rate of 19.7%.

## 3.2 Relationships between Survival Rate and other Variables with Interaction

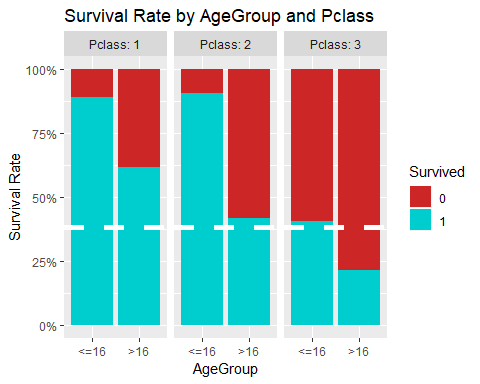
### 3.2.1 Sex and Pclass



Female passengers with 1st class tickets almost all survived (96.8%), the high survival rate also hold for female passengers with 2nd class tickets (92.1%), whereas the female passengers with 3rd class tickets had a survival rate of 50%.

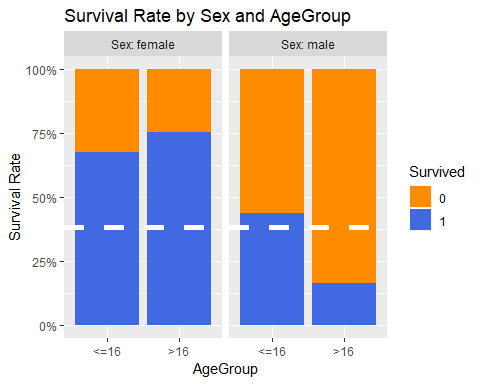
The same pattern hold for male passengers where 1st class survival rate > 2nd class survival rate > 3rd class survival rate, but with much lower survival rate compared to female passengers.

### 3.2.2 AgeGroup and Pclass



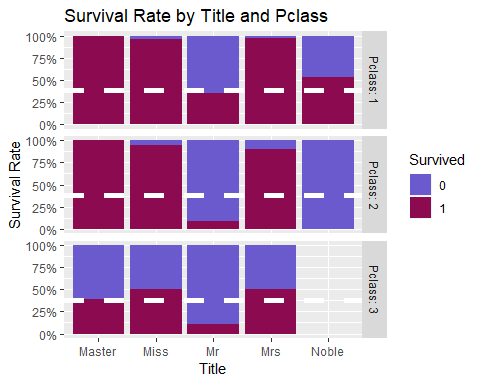
Those aged 16 years old and below traveled with 1st and 2nd class tickets had a very high survival rate (88.9% and 90.5% respectively). Those aged above 16 years old traveled with 1st class ticket also survived at a high rate (61.8%) which further indicate that the survival odds of 1st class passengers are high.

### 3.2.3 Sex and AgeGroup



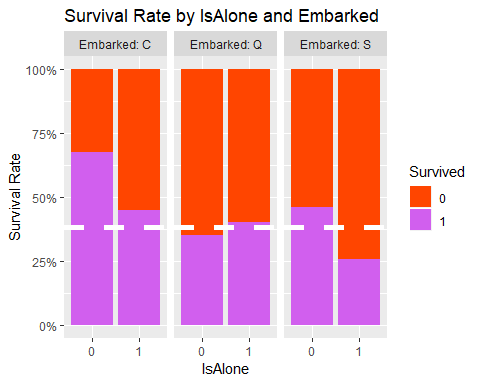
Interestingy, among the females, those who were aged above 16 years old survived at higher rate than those who were 16 years old and below. Whereas among the males, those who were 16 years old and below survived at much higher rate those those who were aged 16 years old and above.

### 3.2.4 Title and Pclass



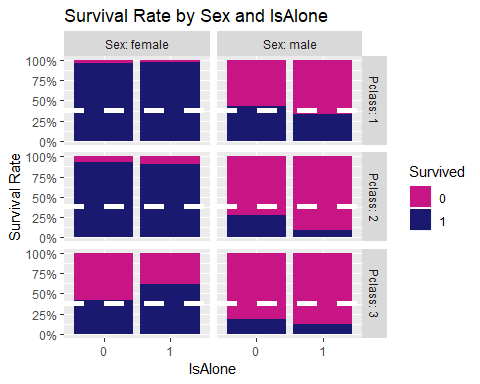
Masters traveled in 1st and 2nd class all survived (100%) whereas Masters traveled in 3rd class had survival rate of 39.3%. As established before, Miss and Mrs traveled at 3rd class had survival rate of 50%. Nobles traveled only in 1st class and 2nd class, those in 1st class had a survival rate of 53.3% whereas those in 2nd class did not survive. Mr had the lowest survival rate amongst all the Title groups, although the odds to survive is slightly higher if they traveled in 1st class.

### 3.2.5 IsAlone and Embarked



Passengers who were traveling alone and embarked from Queenstown had a higher survival rate (40.35%) compared to those embarked from Queenstown but traveling with family members (35%).

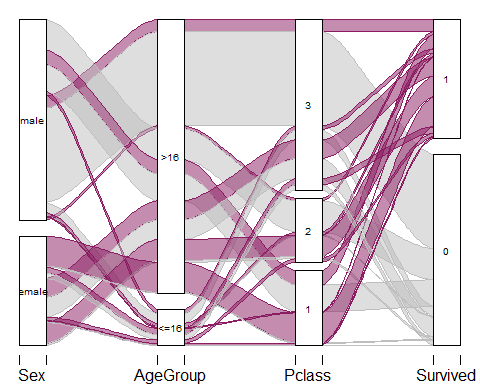
### 3.2.6 Sex, Pclass, and IsAlone



Interestingly, female passengers traveled alone with 3rd class ticket actually had a high survival rate (61.67%). Generally, male passengers traveled alone had slightly lower survival rate in 1st class and 3rd class.

### 3.2.7 Sex, AgeGroup, and Pclass

The following flow diagram is called Alluvial diagram and it is plot using alluvial package. The **maroon** coloured bands represent those who survived, whereas the **grey** coloured bands represent those who did not survive. The larger the band width, the larger the amount of passengers represented in the band.

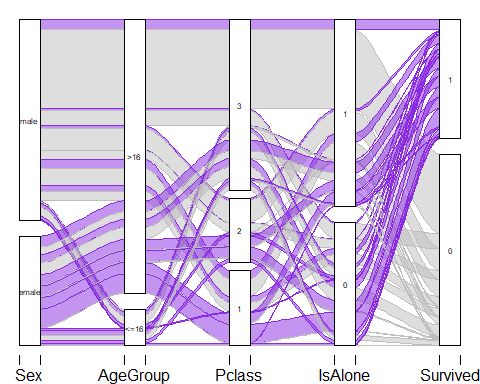


The Alluvial diagram helps summarise what we found in Sex and Pclass, AgeGroup and Pclass, and Sex and AgeGroup:

* Female passengers of any age aboard with 1st class ticket most likely will survive.
* The overall survival rate of 3rd class is 24.2%, but if you are female of any age, you increase your survival rate to 50%.
* Male passengers of 16 years old and below aboard with 1st and 2nd class ticket most likely will survive, but for other conditions, male passengers hardly survive.

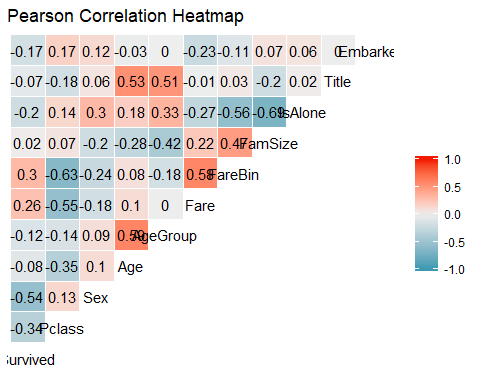
### 3.2.8 Sex, AgeGroup, Pclass, and IsAlone

The following flow diagram is called Alluvial diagram and it is plot using alluvial package. The **violet** coloured bands represent those who survived, whereas the **grey** coloured bands represent those who did not survive. The larger the band width, the larger the amount of passengers represented in the band.



Male passengers of aged 16 and below actually did not travel alone. Female passengers of any age and of any ticket class had higher survival odds even if they were traveling alone.

## 3.3 Correlation



1. Obviously, the new features are highly correlated with the variables they are derived from, such as **Age** and **AgeGroup**, **Fare** and **FareBin**, **FamSize** and **IsAlone**.
2. **Survived** correlates most with **Sex**, followed by **Pclass**, **Fare**, and **IsAlone**.
3. **Pclass** is highly correlated with **Fare** negatively, which suggests that as the fare increased, the more it is under Pclass = 1.
4. **Pclass** is correlated with **Age** negatively, which suggests that the older passengers are richer.
5. **Age** is highly correlated with **Title**.
6. **FareBin** is highly correlated with **IsAlone**, which suggests that the higher the fare bands, the more it is under IsAlone = 0. This makes sense as people would pay to get better fare during family trip.

# 4.0 Predictive Modeling

With data processing, cleaning, and exploration done, the next step will be to predict the survival rate of Titanic disaster. Go through each tabs for different machine learning (ML) models. **5**-fold cross validation will be used in each models to help fine tuning hyperparameters and to determine the models’ performances. Cross-validation accuracy is also a good proxy to estimate the test accuracy of the model. Logistic Regression model is served as baseline model.

The dependent variables selected are *Pclass*, *Sex*, *Embarked*, *Title*, and *IsAlone*. *AgeGroup* is not selected because there were many missing data in *Age*. We do not want to introduce unnecessary noises to the model despite Ages are imputed reasonably. Besides, *Title* contains enough relevant information regarding the age groups. *FareBin* is also not selected because *Pclass* already contains much of the information regarding *FareBin* as suggested in the correlation heatmap.

## 4.1 Logistic Regression

## Generalized Linear Model   
##   
## 891 samples  
## 5 predictor  
## 2 classes: '0', '1'   
##   
## No pre-processing  
## Resampling: Cross-Validated (5 fold)   
## Summary of sample sizes: 712, 713, 713, 713, 713   
## Resampling results:  
##   
## Accuracy Kappa   
## 0.8102756 0.5956287

##   
## Call:  
## NULL  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.5298 -0.6559 -0.3648 0.6235 2.5448   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 18.0663 509.3457 0.035 0.971705   
## Pclass2 -0.9229 0.2731 -3.380 0.000726 \*\*\*  
## Pclass3 -2.1721 0.2478 -8.766 < 2e-16 \*\*\*  
## Sexmale -15.4123 509.3455 -0.030 0.975860   
## EmbarkedQ -0.2667 0.3815 -0.699 0.484559   
## EmbarkedS -0.6963 0.2380 -2.926 0.003433 \*\*   
## TitleMiss -15.4295 509.3457 -0.030 0.975834   
## TitleMr -2.9836 0.4160 -7.172 7.37e-13 \*\*\*  
## TitleMrs -14.9933 509.3457 -0.029 0.976516   
## TitleNoble -3.3816 0.6860 -4.929 8.25e-07 \*\*\*  
## IsAlone1 0.5216 0.2211 2.359 0.018315 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1186.66 on 890 degrees of freedom  
## Residual deviance: 763.77 on 880 degrees of freedom  
## AIC: 785.77  
##   
## Number of Fisher Scoring iterations: 13

The Logistic Regression model has a cross-validation accuracy of 81%. The model estimation result shows that:

* 2nd class passengers are 0.397 (exp(-0.9229)) times as likely to survive as 1st class passengers; whereas 3rd class passengers are only 0.114 (exp(-2.1721)) times as likely to survive as 1st class passengers. The odds of 2nd and 3rd class passengers to survive are very low compared to 1st class passengers.
* Male passengers are almost guaranteed perish (exp(-15.4123)) as compared to females. But the coefficient of this dummy variable is statistically insignificant.
* Passengers embarked from Queenstown are 0.766 (exp(-0.2667)) times as likely to survive as those embarked from Cherbourg; whereas passengers embarked from Southampton are 0.498 (exp(-0.6963)) times as likely to survive as those embarked from Cherbourg.
* Mr are 0.05 (exp(-2.9836)) times as likely to survive as Masters; whereas Nobles are 0.034 (exp(-3.3816)) times as likely to survive as Masters. The odds of Mr and Nobles to survive are very low compared to Masters. The coefficients for Miss and Mrs are statistically insignificant.
* The odds of survive are 1.685 (exp(0.5216)) times higher for passengers travel alone as compared to those who are not traveling alone. This is probably due to the fact that female passengers traveled alone in 3rd class and passengers traveled alone from Queenstown had higher odds of survival.

## 4.2 k-Nearest Neighbour

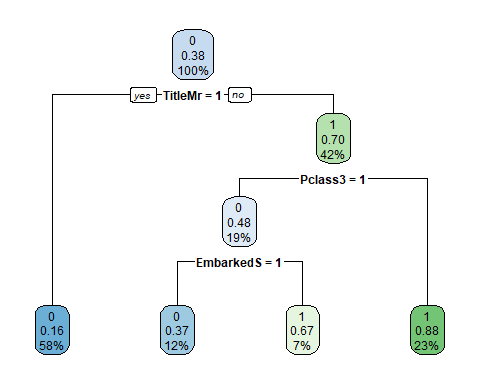
## k-Nearest Neighbors   
##   
## 891 samples  
## 5 predictor  
## 2 classes: '0', '1'   
##   
## No pre-processing  
## Resampling: Cross-Validated (5 fold)   
## Summary of sample sizes: 712, 713, 713, 712, 714   
## Resampling results across tuning parameters:  
##   
## k Accuracy Kappa   
## 5 0.8170774 0.5900708  
## 7 0.8058729 0.5643255  
## 9 0.8069964 0.5643497  
##   
## Accuracy was used to select the optimal model using the largest value.  
## The final value used for the model was k = 5.

The optimised k-Nearest Neighbour model is the 5-Nearest Neighbour model with cross-validation accuracy of 81.7%.

## 4.3 Decision Tree

## CART   
##   
## 891 samples  
## 5 predictor  
## 2 classes: '0', '1'   
##   
## No pre-processing  
## Resampling: Cross-Validated (5 fold)   
## Summary of sample sizes: 712, 713, 713, 713, 713   
## Resampling results across tuning parameters:  
##   
## cp Accuracy Kappa   
## 0.02046784 0.8159689 0.5913317  
## 0.04093567 0.7901324 0.5488081  
## 0.43274854 0.7092461 0.3148347  
##   
## Accuracy was used to select the optimal model using the largest value.  
## The final value used for the model was cp = 0.02046784.

The optimal Decision Tree model is the simplest possible model where the complexity parameter is about 0.0205. The cross-validation accuracy of tuned Decision Tree model is about 81.6%.

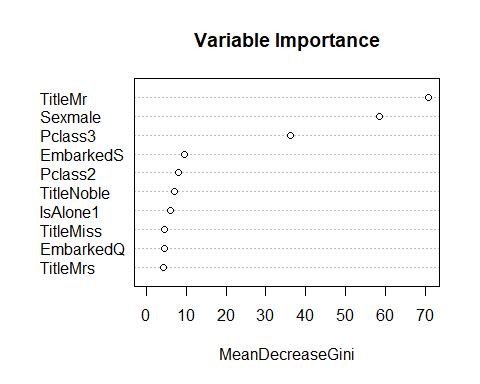


Title being Mr. is the most important feature for the model as it appears at the top internal node. If a passenger is Mr., he will not survive. If a passenger is not a Mr., then the next important feature is the Pclass. If the passenger is not a Mr. and is not traveling in 3rd class, he/ she will survives. But if that passenger is traveling in 3rd class and embarked from Southampton, he/ she will not survive.

## 4.4 Random Forest

## Random Forest   
##   
## 891 samples  
## 5 predictor  
## 2 classes: '0', '1'   
##   
## No pre-processing  
## Resampling: Cross-Validated (5 fold)   
## Summary of sample sizes: 712, 713, 713, 712, 714   
## Resampling results across tuning parameters:  
##   
## mtry Accuracy Kappa   
## 2 0.8137318 0.5905738  
## 6 0.8260726 0.6108300  
## 10 0.8215845 0.6029080  
##   
## Accuracy was used to select the optimal model using the largest value.  
## The final value used for the model was mtry = 6.

The optimised Random Forest model is the model with parameter mtry = 6. This means that at each splits during the trees building process, 6 variables are randomly sampled as split candidates out of the 10 variables. The cross-validation accuracy of tuned Random Forest model is about 82.6%.



For the Random Forest model, the top 3 most important features are TitleMr, Sex, and Pclass3. This shows that males, particularly adult males (Mr.) are very important characteristics in predicting whether a passenger will survives. 3rd class ticket is also a very important feature in predicting whether a passenger will survives. These are in accordance with the findings in EDA where we know that Mr and Males had a very low survival rate, and 3rd class passengers also had a very low survival rate.

# 5.0 Conclusion

There are 2 features that are always given higher priority in fitting the models. These features are probably the most important predictors, they are **TitleMr** and **Pclass3**. As pointed out in EDA, Mr had the lowest survival rate (15.7%) amongst all the Title groups, whereas 3rd class passengers had the lowest survival rate (24.2%) amongst all the ticket classes. Hence, it is no surprise that when passengers fall into these categories, their odds of survival are very low.

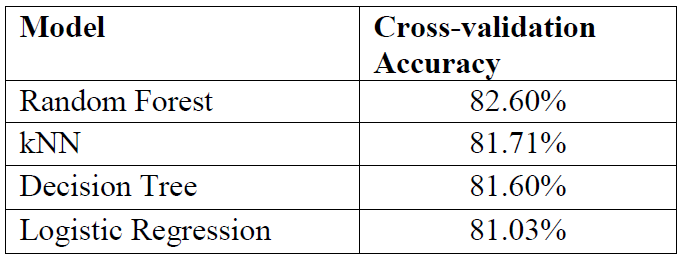


Fig 3: Models Comparison

Random Forest, k-NN, and Decision Tree models perform better than Logistic Regression model. This is probably becasue these models can capture more patterns than Logistic Regression model (without non-linear extension) which can only capture linear patterns. Of the 2224 passengers on-boarded Titanic, 891 observations are used in the above analyses. Random Forest model is the best performing model based on cross-validation accuracy. The cross-validation accuracy provides a realistic estimation of test accuracy, viz, how well the chosen model performs when exposed to previously unseen data (the remaining 1333 observations). However, the true test accuracy of the model and how well the model generalizes can only be assessed by exposing it to other unseen observations.

Lastly, there are 3 features that are not being exploited here due to complexity. One of them is the **family name** of the passengers, the family names might help to explore in detail the relationship amongst the passengers and might even provide insights into whether they were all stayed in the same cabin and whether the younger family members survived the most. The other unexploited features are **Cabin** and **Ticket code**. The ticket codes might provide deeper insights into the Pclass and probably associated with cabins. The cabins might provide a large clue as to which cabins are closer to the upper deck and hence, whether the passengers stayed in these cabins made it to the upper deck faster and got to escape the ship faster. These are the features that can be exploited and might even help to increase the predictive power of the models.