**Titanic: Machine Learning from Disaster**

**1. Problem Statement**

**Background**

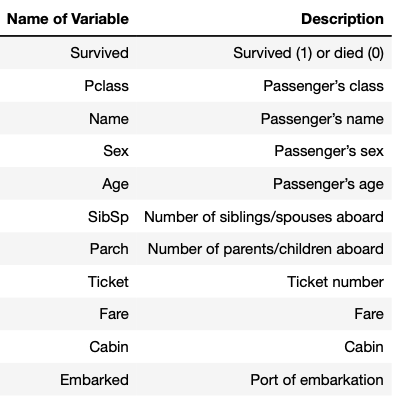
The Titanic sank after its collision with an iceberg on April 15, 1912. The tragedy led to the deaths of 1,502 out of 2,224 passengers and crew, calling for better safety regulations for ships. The large number of casualties was partly caused by the insufficient number of lifeboats for the passengers and crew. Besides luck, it seemed that certain groups of people such as women, children, and the upper-class were more likely to survive than others.

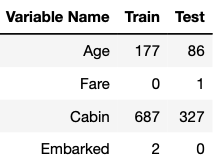
**Goal**

This report aims to predict the survival of the passengers provided in the Test set, therefore 0 for \*Died\* and 1 for \*Survived\* based on the X independent variables.

**2. Data Description**

There are 2 datasets namely \*training set (train.csv)\* and \*test set (test.csv)\*. The training set and test set each contains 891 obeservations (12 variables) and 418 observations (11 variables)respectively. The training set contains one more variable Survived which depicts the binary predictions, therefore, 1 for \*Survived\* and 0 for \*Died\*.  Description of the variables are as follows:

  
  
We note that there are various missing data in the Train and Test datasets. In particular, Age and Cabin has large number of missing data. In Section .3.2, we will be imputing these missing data



**3.1 Data Exploration and Feature Engineering**

We will first form insights through data exploration to create new relevant features and impute missing data. We will then train our models based on selected features in the Train dataset. The model with the highest performance will be improved and used to predict the Survived indicator in the Test dataset.

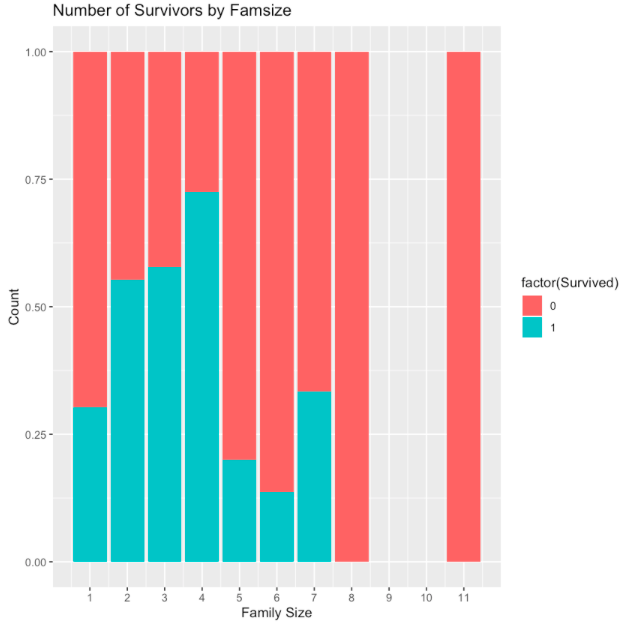
This section will cover data exploration, feature engineering, missing data imputation and modelling.

**3.1 Data Exploration and Feature Engineering**

In this section, we will be exploring the data to create new and relevant features, which will be used to build our models in Section 3.3. To do so, we will combine the training and test sets to have clearer picture of the population for data exploration and cleansing.

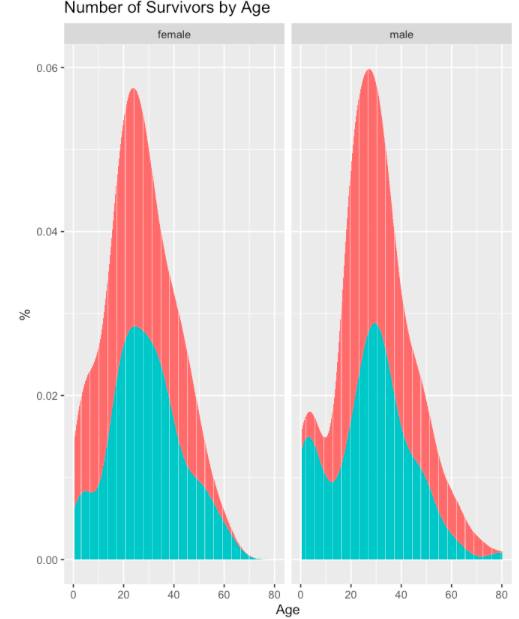
**3.1.1 How does family size affect survival?**

We will create a new family size variable which is the sum of one's siblings, spouses, parents,children and the passenger himself/herself. From the chart, we can see that singles and bigger family with a size of 4 or more have a lower chance of survival. Smaller families with a size of 2-4 are more likely to survive. We will therefore discretize our family size variable to 3 categories namely, single, small and large.



**3.1.2a How does age affects survival?**

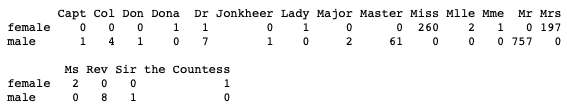
Age may seem to be a strong predictor of survivability. From the chart, it is apparent that older women passengers and younger child passengers (in particular male) had a higher chance of survival. From the density histogram chart below, we can see that survivial rate is lower for adult passengers and higher for senior passengers. It is therefore meaningful to perform discretisation to create a new AgeGroup variable. As there is a significant number of missing Age values, we will impute these missing Age values and perform discretization in section 3.2.2.



**3.1.2b How does one's title relate to age?**

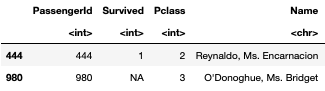
As there are no missing data in Name, Name may be a fairly robust predictor for Age. The structure of Names is "surname, title. given name (maiden name)". Interestingly, a maiden name will be present in parentheses to indicate that the passenger is a married woman.

We can see that some of the titles are underrepresented while are abundant and noted that Titles with small number of occurrences has limited predictive power. We will feature engineer the titles to ensure the predictive power for our imputation model for the missing Age data.

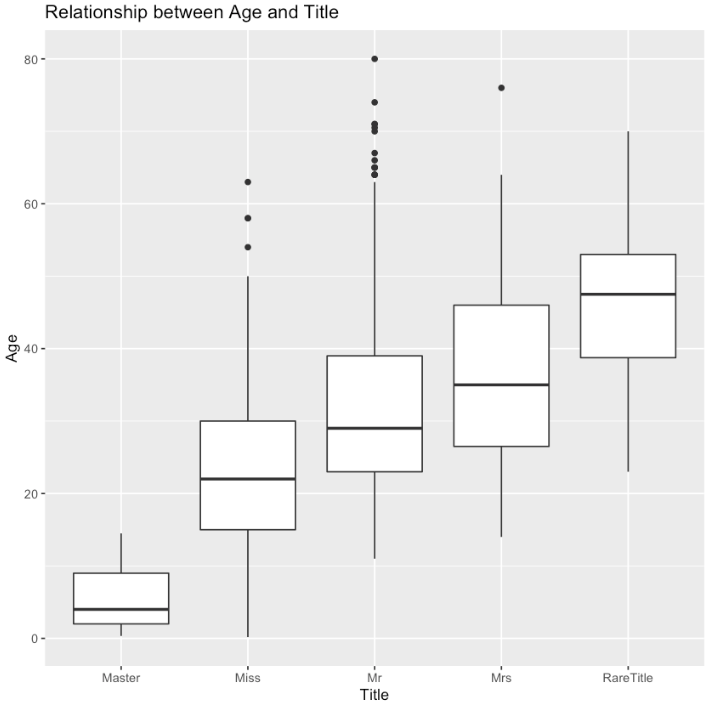


Interestingly, maiden name presented in parentheses indicates that the passenger is a married woman.

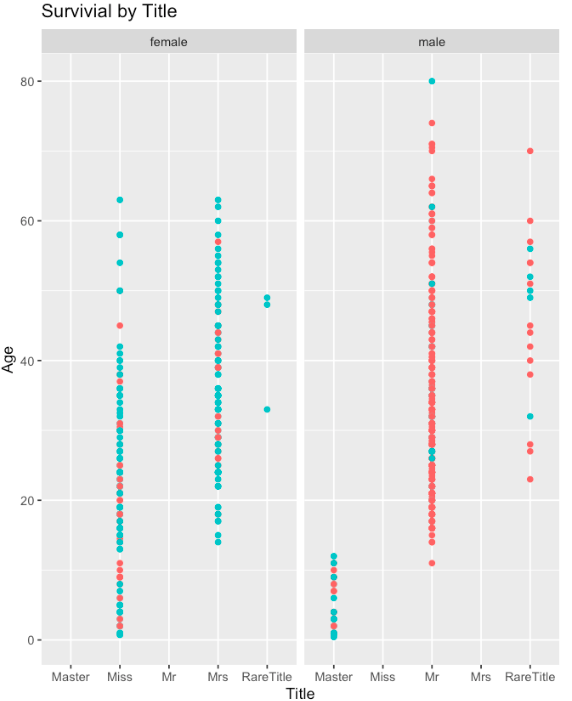
As the 2 instances of “Ms” do not have maiden name in parentheses, we will assign it together with “Miss”. We have also assigned the French designations Madamoiselle (“Mlle.”) and Madame (“Mme.”) to their corresponding English designations of “Miss” and “Mrs” respectively.



We will categorise the rest of the Title that had few occurences namely "Capt", "Col", "Don", "Dona", "Dr","Jonkheer", "Lady", "Major","Rev","Sir", "the Countess" into \*RareTitle\* as these titles indicate an adult age due to honorifics and professional terms used. We can see that passengers with titles 'Master' and 'Miss' are of younger age. On the other hand, passengers with'RareTitle' are of higher age. It therefore shows that Title is a fairly robust predictor for Age.

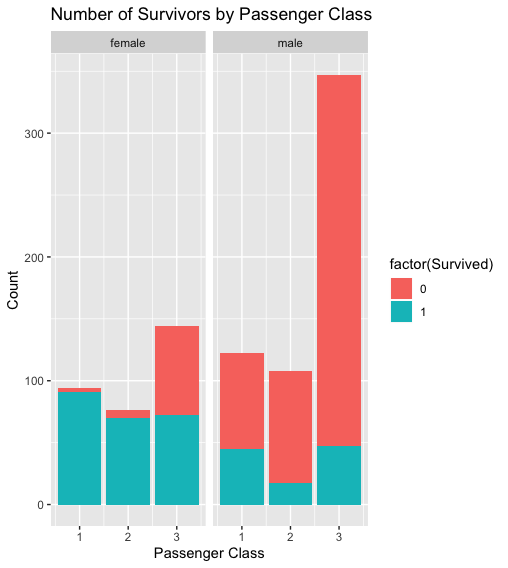


We can also see that all the younger male passengers who survived are with the 'Master' title and all the women passengers with 'RareTitle' all survived.

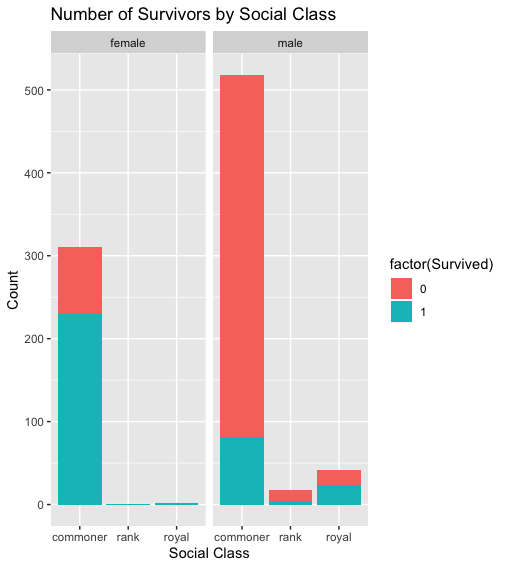


3.1.3 **How does passenger class and social class affect survival?**

Generally, passengers from Class 1 have a higher chance of survival as compared to the other classes, for both men and women. Among those who are from passenger class 1, passengers with high social status may have a higher rate of survival. We will discretize the Names variable to depict the different social class namely Commoner, Royal and Rank.

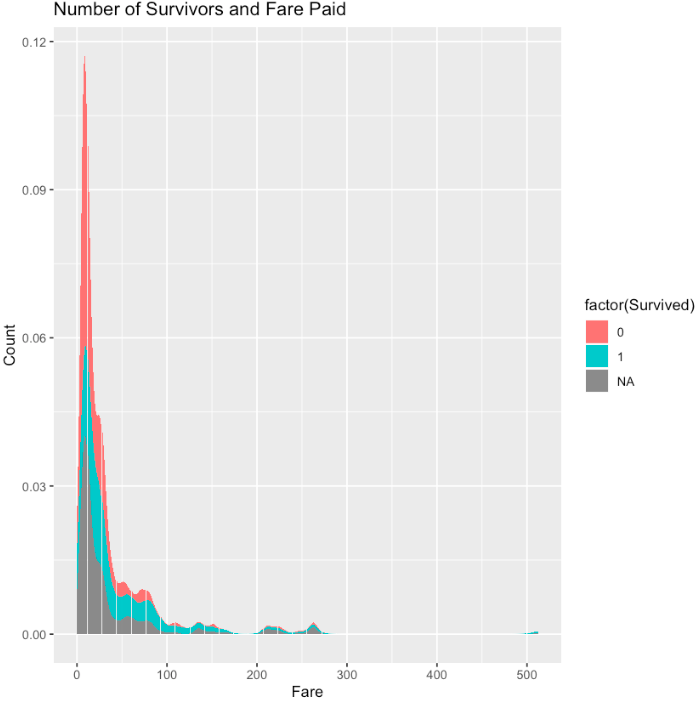


We can see that passengers from a royal social status have higher rate of survival than other social classes. In particular, women from the social class of *rank* and *royal* in passenger Class 1 all survived. On the other hand, men of social class *rank* has a lower survival rate as they are likely to be the heroes who helped others to escape.



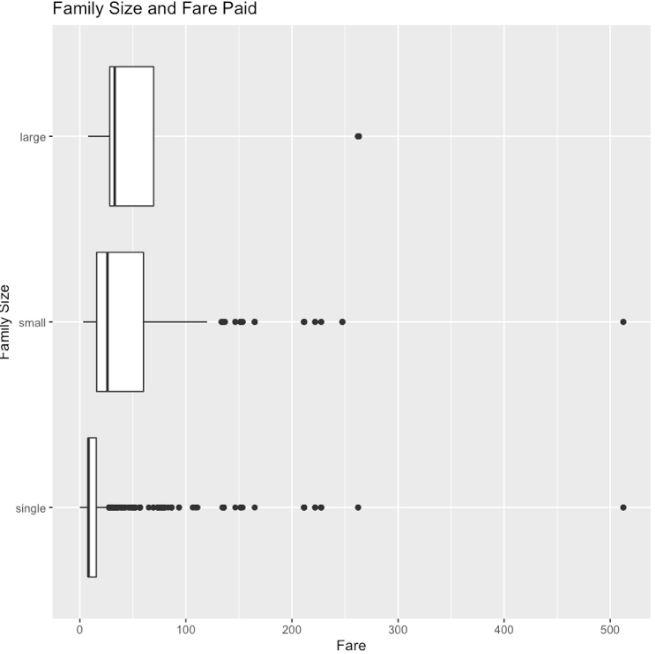
**3.1.4a How does passenger fare affects survival?**

We can see that passengers who paid a lower fare have a lower chance of survival. It is interesting to note that all the male passengers who paid 0 fare in the train dataset all died. All of them embarked from Southampton and were all singles! These men were mostly likely the crew who sacrificed themselves and helped the other passengers to escape.



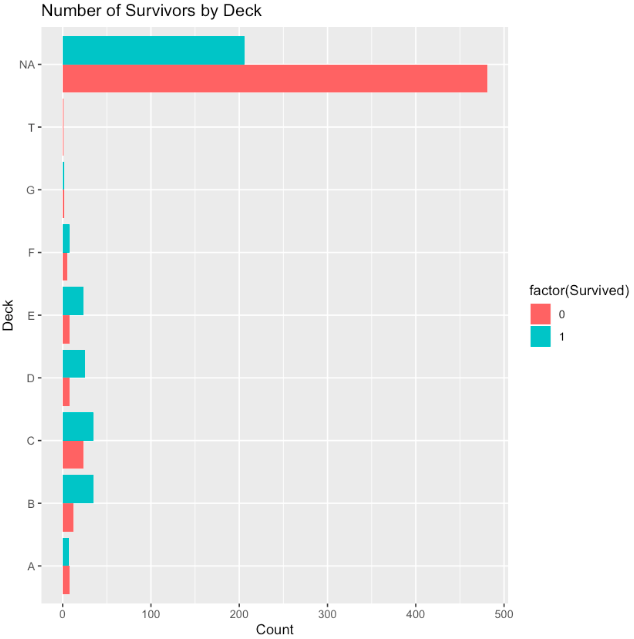
**3.1.4b How does family size relates to passenger fare?**

We can see bigger family, has higher median fare. It is likely that that passenger fare is the total amount paid for the entire family. We also note that that the fare of singles is more extreme as can be seen from the outliers.In this regard, we will create a new feature "PerPassengerFare".



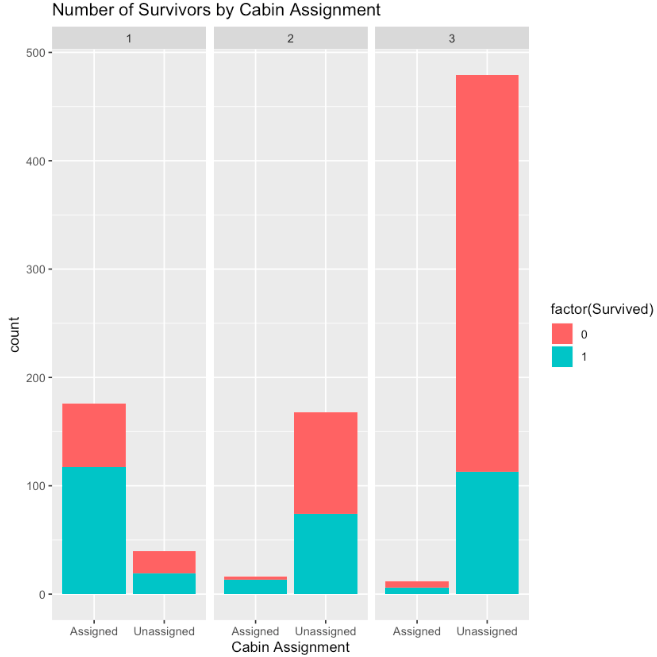
**3.1.5a How does cabin affects survival?**

The Cabin feature could be a strong predictor for survivability. Passengers who are located in cabins that are nearer to exit and/or access to lifeboats will have higher chance of survival. However, there are many missing Cabin values. The first letter of the Cabin feature denotes the deck. The relationship between the Deck and survival is not clear. However, it seems that survival rate is relatively high for passengers where Cabin value is not missing. The missing Cabin values may not be missing data but an indication that these passengers were not assigned a Cabin!



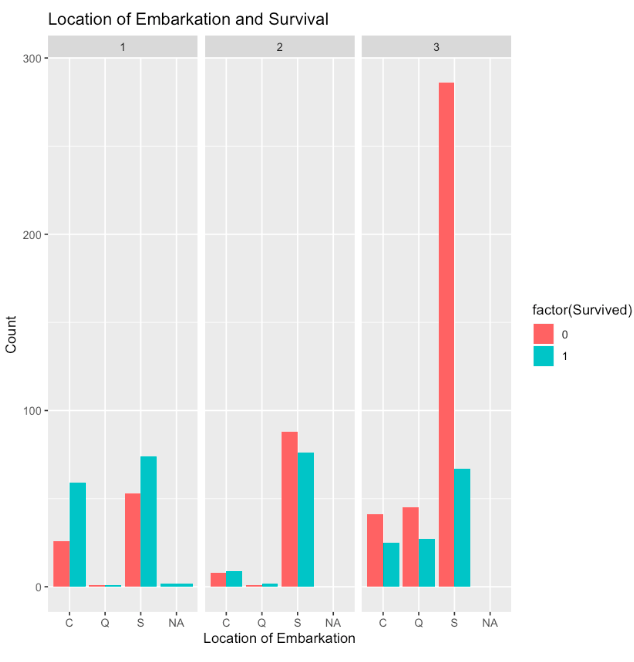
**3.1.5b How does the assignment of cabin affects survival?**

We can see that passengers who were being assigned a cabin have a higher survival rate than those who were not assigned a cabin for all passenger classes. Being assigned a cabin\ beforehand may denote one’s social status and wealth. As we created a new binary CabinAssignment feature, which could be seen as a strong predictor, there is no need to fill in the missing Cabin values. We will use this new binary CabinAssignmnet feature as one of the independent variables to build our models.



**3.1.6 How does location of embarkation affects survival?**

Passengers embarked Titanic from 3 different location namely Cherbourg (=C), Queenstown (=Q) and Southampton (=S). Near 7 in 10 passengers boarded the Titanic from Southanpton. A compared to S and Q, passengers who embarked from C has higher survival rate than passengers. Furthermore, a passenger who embarked from Cherbourg has higher survival rate if he was from passenger class 1 as compared to passengers who were from the other passenger classes whom he/she had embarked the ship with.



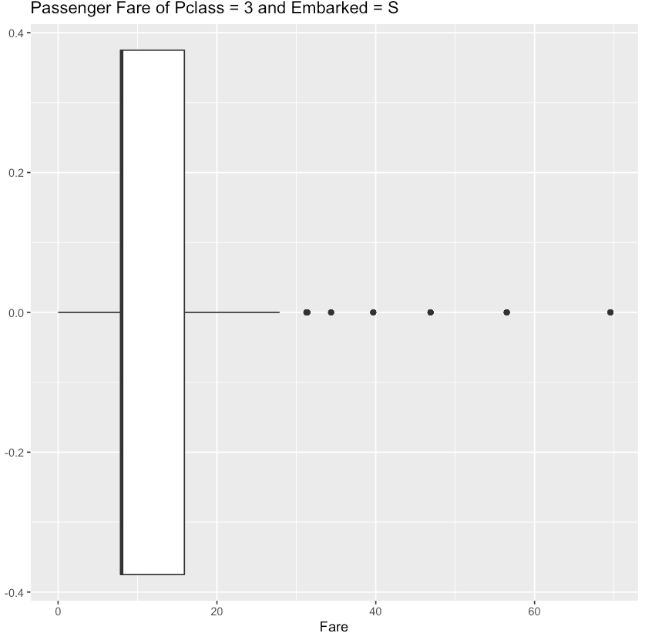
**3.2 Missing Data Imputation**

**3.2.1 Imputation of Missing Embarked Values**

We note that passenger with ID 62 and 830 have missing Embarked values. We will impute these missing values by S as it is where most of the passengers embarked as see n previously.

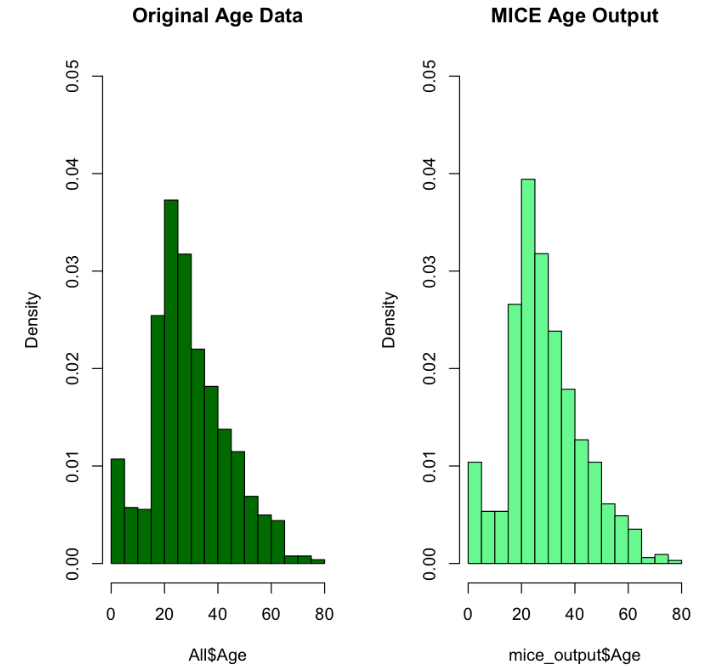
**3.2.1 Imputation of Missing Fare**

The missing fare is from a Pclass 3 passenger who boarded the Titanic from Southampton. We will look at the fares of the other passengers who were also from Pclass 3 whom had boarded the Titanic from Southampton. As there seems to be a few passengers who paid higher prices (outliers), we will impute the missing fare with median instead of mean fare of the Pclass 3 passengers who boarded the Titanic from S.



**3.2.3 Predictive Imputation of Missing Age Values**

In section 3.1.2b, we saw that Title may be a robust predictor of the missing Age values. We will perform the \_mice\_ imputation that will exclude the less than useful variables. As the Age distribution of the MICE output is in line with the original data, we will replace the original Age with our MICE output.



**Discretization of Age**

In section 3.1.2, we note that survival rate differs for different age range. Now that we have all the imputed missing Age values, we will perform discretization to create a new AgeGroup variable and to ensure that each group is not overrepresented. We will categorize the Age variable into a new AgeGroup variable consisting 5 categories as follows:

- young children: less than 12 years old

- children and young adults: more than or equal to 12 and less than 21 year old

- adults: more than or equal to 21 and less than 30 year old

- middle aged adults: more than or equal to 30 and less than 50 year old

- older adults: more than or equal to 50 years old

The frequency of each age group is as shown below. If we do not adults into different category, the adults class will be overrepresented and may result in imbalance issue.



**3.3 Feature Selections and Importance**

**3.4 Building Machine Learning Models**

As the independent variable (Survived) is missing, we are not able to use the test set to validate the performance of our model. We will therefore use a 10 fold cross validation repeated 10 times to validate the performance of our models. We will compare among all the models that we trained and choose the one with the best performance for submission of the results to Kaggle.

**3.4.1 Logistic Regression**

Logistic regression is a predictive modelling technique for binary categorical variable. We will first run a logistic regression on our train dataset and do a repeated 10-fold cross validation based on our train dataset. The results show that…

**3.4.2 Random Forest**

**3.4.3 Neural Network**

**3.4.4 Naiye Bayes**

**3.4.5 Decision Tree**

**3.4.6 K Nearest Neighbour**

**4.1 Model Comparison**

**4.2 Kaggle Submission Score(s)**

**5.1 Conclusion**

**5.2 Future Directions for Improvement**