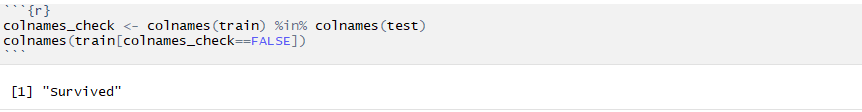
**Research Question:**

Who were the people most likely to survive in such an accident, where the survival was just a matter of luck and what sorts of people were likely to survive?

**1. Introduction:**

Initially the data is loaded into the R-Studio and then the packages that are required which are then imported for the project. Once the packages are imported, data is loaded and then divided into test and train datasets where we use train data for training the model and the test data to check the accuracy of the model built.

The training set has 891 observations and 12 variables and the testing set has 418 observations and 11 variables. The training set has 1 extra variable. We could see that in a very small dataset like this, but if its larger we want two compare them. As we can see we are missing the Survived in the test set.



**2. Data Analysis:**

Analysis will be performed by considering age, class, gender, family and some other important features of the passengers. This document is more about training and evaluating the model. To do that, a dataset that has the complete passenger information is considered i.e. the train dataset. I will also try feature engineering where we will consider passenger name because we can break it down into additional meaningful variables which can feed predictions or be used in the creation of additional new variables.

Before feature engineering let’s look into missing values. For studying the complete data set lets join test and train data set. The below Fig1 is the combine data of train and test datasets and named as full. Here test data had 418 obs. and it didn’t have survival data. Whereas there are lot of other missing values like Cabin and Age variable too.

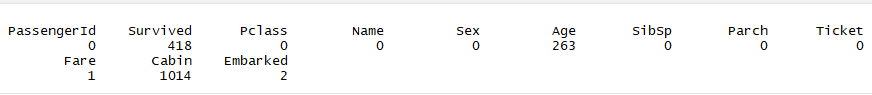


Fig1: Missing value count per variables used in full

Below Fig2 displays all existing combinations of missing (red) and observed (lightseagreen) values. For example, this plot reveals that if variable Cabin are missing, they are also missing Age and Fare variables. However, when Embarked variable is missing, Cabin, Age and Fare variables have data. We are not concerned with Survival’s missing observations since they are mostly from test dataset.

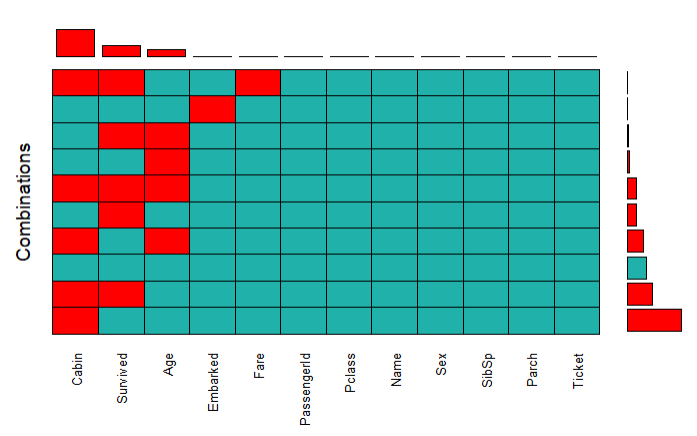


Fig2: Aggr function used to find the missing values

The data can be missing for many reasons. Generally, there are 3 steps required in dealing with missing data:

* Find the missing data
* Investigate the cause of missing data &
* Replace the missing values with sensible data or delete the missing data rows.

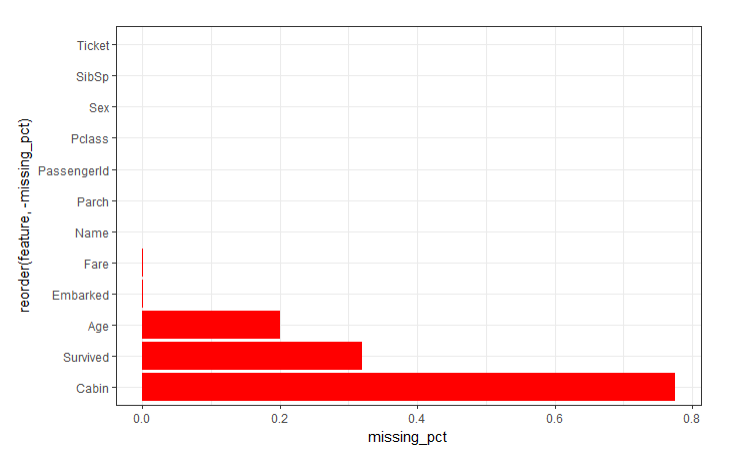
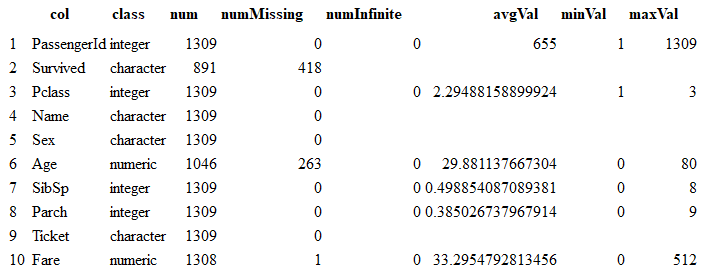
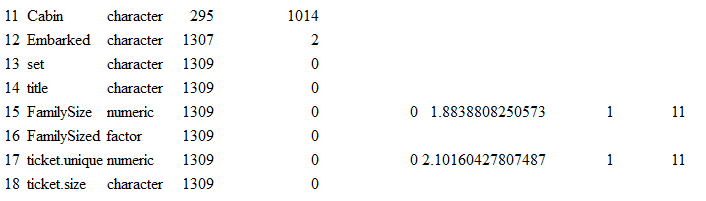


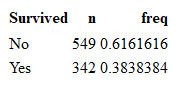
Fig3: Missing values in percentage

The below table shows useful data quality function for missing values





The independent variable, Survived, is labeled as a Bernoulli trial where a passenger or crew member surviving is encoded with the value of 1. Among observations in the train set, approximately 38% of passengers and crew survived. The below figure shows for train dataset(891 rows)

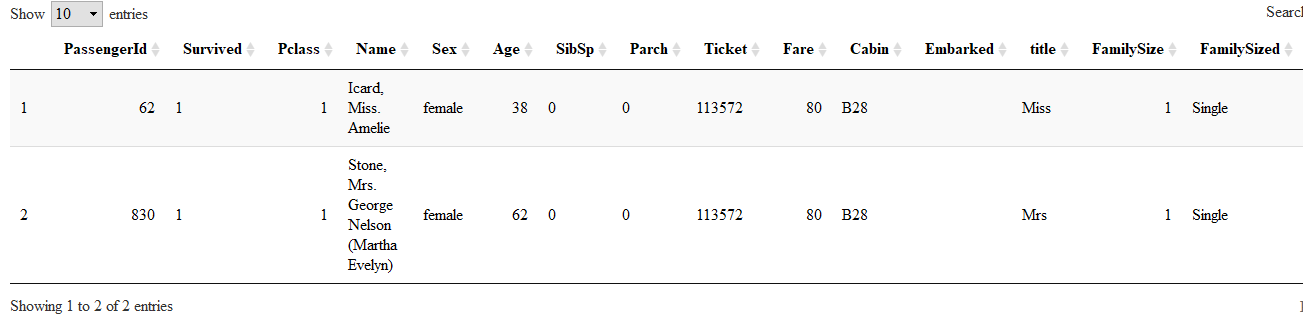
****

**3. Feature Engineering:**

Missing data can’t be used in modeling, it has to be patched up with data. So now we can focus on preparing the data so that it can be used for study, such as exploratory data analysis and modeling fitting.

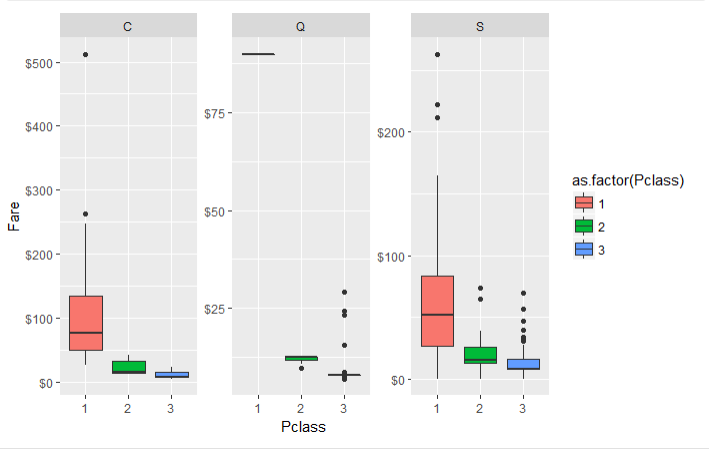
Firstly, lets us look at missing values of Embarked variables. We know from our missing data plot [Missing values](http://rstudio-pubs-static.s3.amazonaws.com/241333_e3e176ecc7a44094a6bba75899a88f72.html#anchor) that there are no other variables missing when combined with Embarked variables.

The below Fig4 that there are two observations with missing data from Embarked variable. We also know that there are three Port of Embarkation -> C = Cherbourg; Q = Queenstown; S = Southampton. The table above shows that both tickets were from Pclass 1; Fare $80; Ticket no 113572 & Cabin B28.

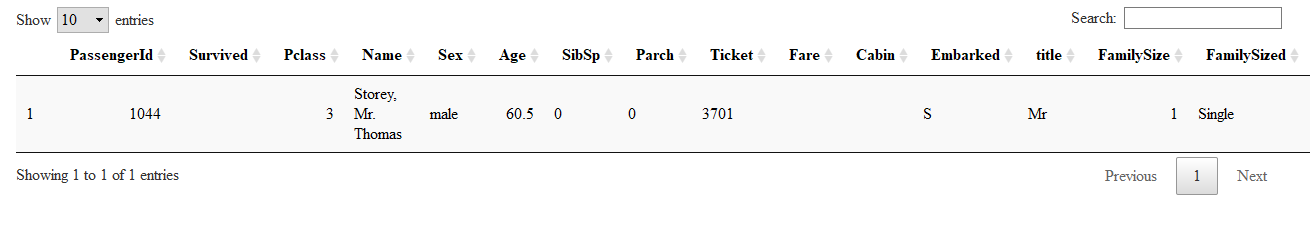


**Fig4: Missing Embarked**

We can see from the box plot that Pclass 1 median fare is approx $80 when embarked from C which is Cherbourg. Hence now I think it is sensible to assign the missing embarked with “C” for those two tickets whose Embarked values were missing.

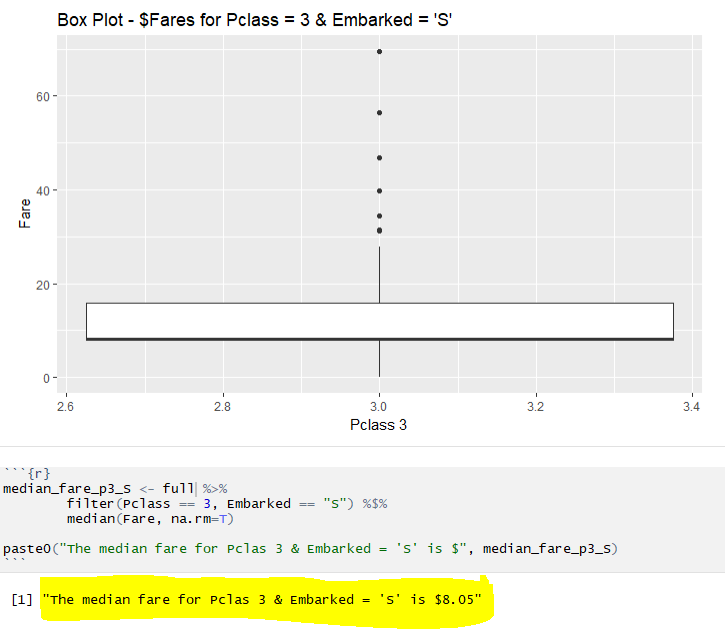


Now the PassengerId 1044 has Fare value missing as shown below in Fig5



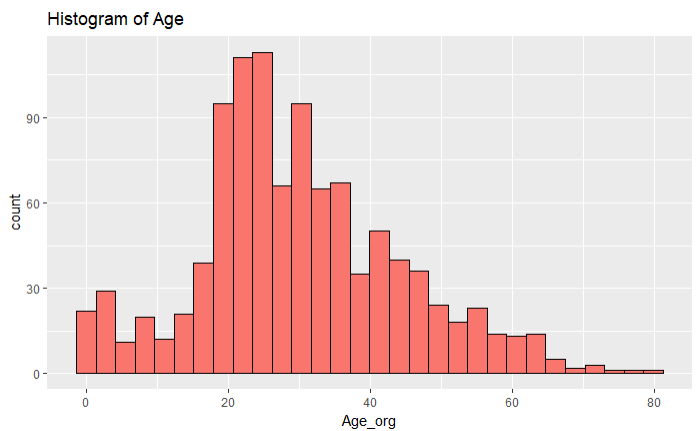
**Fig5: Missing Fare**

The density plot in below Fig6 shows the counts and the area underneath the polygon is one. This plot tells us approx. $8.05 accounts to around 15 percent of the fare in Pclass 3 and Embarked from S. I think it is sensible to replace NA with $8.05

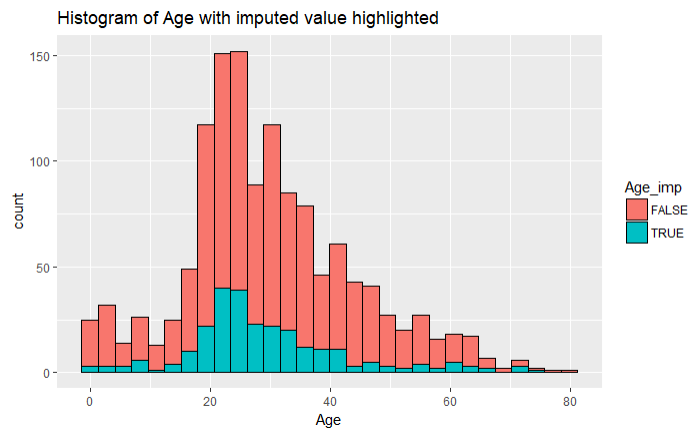


**Fig6: median fare for Pclass 3 and Embarked S**

We’ll use MICE imputation method to impute missing Age variables. We’ll use Pclass, SibSp, Parch, Fare, Embarked, Title, FamilySize variables as predictors to impute missing values in Age.



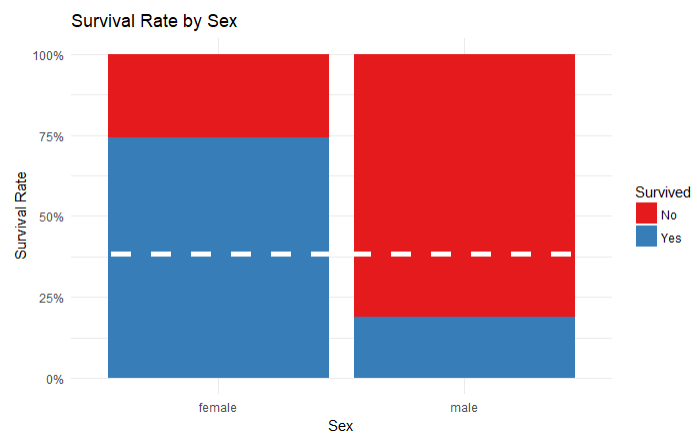
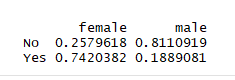
Now fill the missing values for age. The below shape of histogram has not changed significantly from original values and after imputing missing values. I think its sensible to use the imputed values.



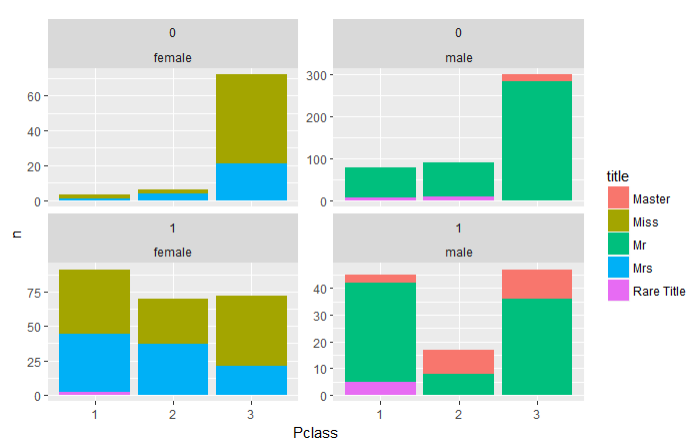
**4. Relationship to Survival Rate**

* 1. **Sex vs Survival:**

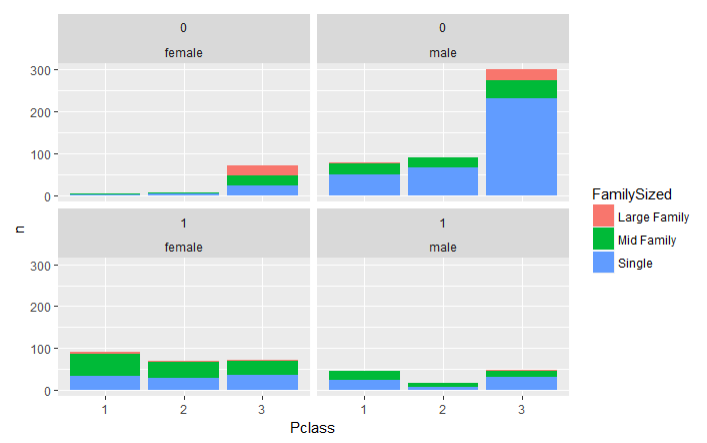
74 % of all the females in train data survived whereas only ~18% out of all male population survived. That is huge differences!!!

The below plot clearly shows that there are lot more female survivors then men. In all 3 classes female survivors outnumber male survivors. If you were female you had more chances of survival in all Pclass. You’ll have a good chance of survival if you were single or in small size family.

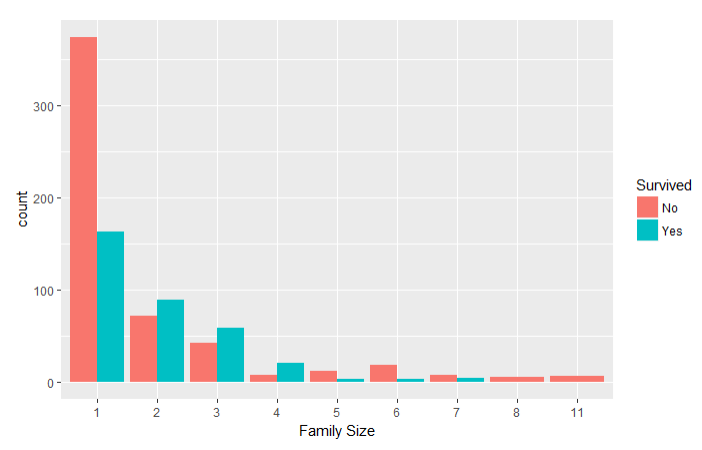


**Fig: Relation: Title | Pclass | Survival**

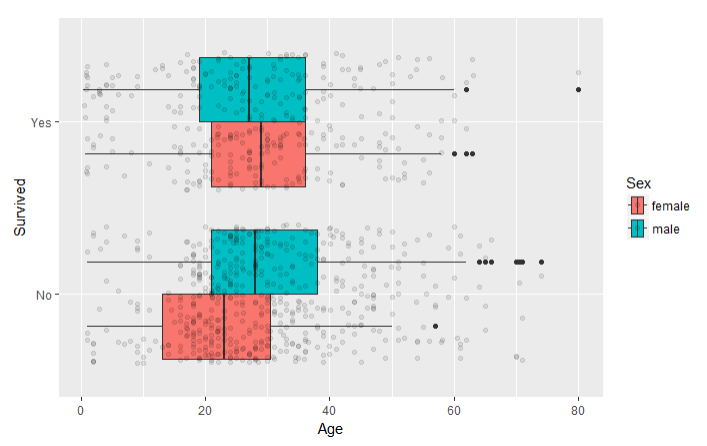


**Fig: Relation: Family Sized | Pclass | Survival**

And this confirm visually that small family have survived better than large family

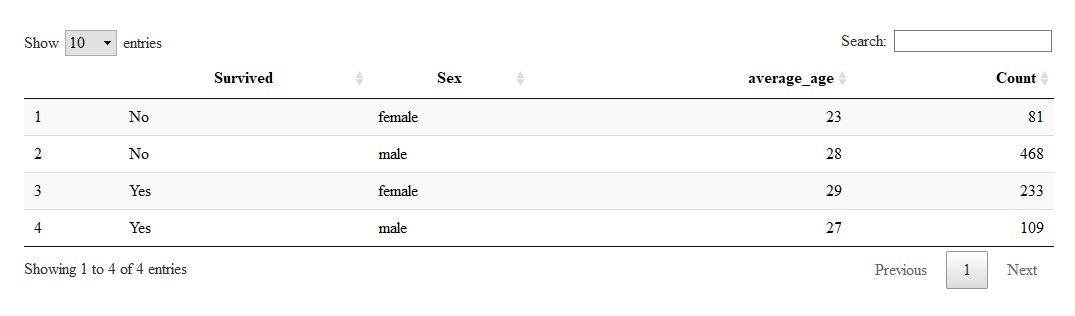


* 1. **Age vs Survival Boxplot**



**Fig: BoxPlot for Age Vs Survival**

The below data table (train data) shows that those who did not survive, on average females were much younger than males.

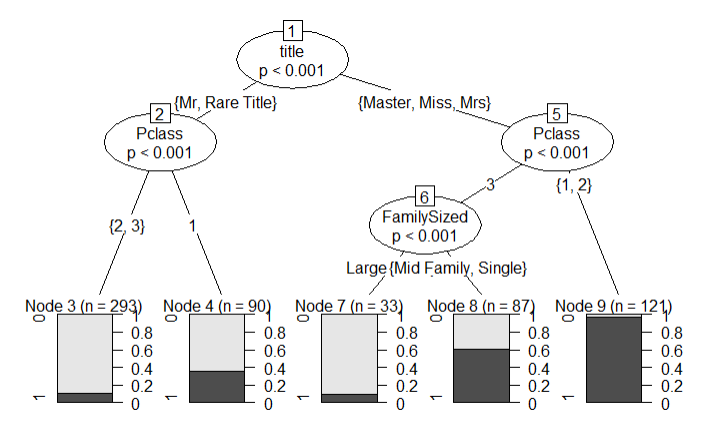


**5. Prediction:**

1. **Using Conditional inference tree (ctree)**

I feel conditional inference trees can be used as base learner for Random Forests. We’ll split the training set (891 rows)- train into trainset & testset using 70/30 ratio. And will build the ctree (conditional inference tree) model using:

Survived ~ Pclass + Sex + SibSp + Parch + Fare + Embarked + Title + FamilySize + Age + FamilySized



The above tree diagram has confirmed what we already knew from our data exploration. The tree diagram claims that the most important variable is Title and if you’re female and travel in Pclass 1 & 2 you have much chances of survival. If you are male the worst-case scenario would be to travel in Pclass 2 & 3. Since we had split the train data set into trainset and testset we know the survival status. Going forward we can use this information to check our model prediction power.

Does your model match up with what your research question is asking?

Yes, I always thought title, age, Pclass, Sex are the important variables to help match my research question. I was wondering why it didn’t use Sex in above model. I still digging more on both the model (ctree and decision tree).

1. **Using Decision tree**

The decisions that have been found go a lot deeper. Decisions have been found for the port of embarkation one that I didn’t even look at. The decision is made on the current node which appear to be the best at the time but are unable to change their minds as they grow new nodes.

