Bayesian Data Analysis Report on Titanic Dataset

Alp Gunsever

13/01/2021

1 - Introduction

This is a data analysis report intended to test bayesian data analysis skills using the titanic dataset from Kaggle. A bayesian data analysis framework will be followed to perform a full analysis on the titanic dataset. The analysis will be finalized by coming up with a final model and making a prediction with it in the Kaggle competition. All the results and observations will be shared and explained as much as possible for the whole process.

2- Exploratory Data Analysis

It will be beneficial to look into the raw data available before getting into any analysis work. However, it must be stated that any comments on this section will be based on point averages and should be taken as initial observation not causal relationships. In order to do so, the existing observations for different predictors will be first formatted into data structures that can be handled by R in a meaningful way.

Before getting into any data cleaning or modification, let's have a look at some of the rows of the raw data:

```
##   PassengerId Pclass                                                Name    Sex
## 1           1      3                             Braund, Mr. Owen Harris   male
## 2           2      1 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female
## 3           3      3                              Heikkinen, Miss. Laina female
## 4           4      1        Futrelle, Mrs. Jacques Heath (Lily May Peel) female
## 5           5      3                            Allen, Mr. William Henry   male
## 6           6      3                                    Moran, Mr. James   male
##   Age SibSp Parch           Ticket    Fare Cabin Embarked
## 1  22     1     0        A/5 21171  7.2500              S
## 2  38     1     0         PC 17599 71.2833   C85        C
## 3  26     0     0 STON/O2. 3101282  7.9250              S
## 4  35     1     0           113803 53.1000  C123        S
## 5  35     0     0           373450  8.0500              S
## 6  NA     0     0           330877  8.4583              Q
```

The following adjustments have been implemented to the training and test data sets together after they are combined into a single dataset:

- Passenger class predictor is transformed into factor type with classes "1", "2" and "3" set as separate levels.
- Titles have been extracted from the name variable and factored into "Mr", "Mrs", "Miss", "Master", "Noble" and "Soldier" levels based on the implications of various titles.
- Cabin predictor is transformed into factor type according to the letter in the cabin name.
- Embarked predictor is transformed into factor type with the first letters of the ports representing different levels as "C", "Q" and "S".
- Ticket predictor is factored into the number of digits of the number part of the ticket.

- Family size predictor is created based on number of siblings and spouses including self. Family type predictor is created based on the family size predictor. If the total number of family members are equal to 1, then family type is assigned to "Singleton". If family size is between 1 and 4, then family type is assigned to "small". If family size is greater than 4, then family type is assigned to "large". Family type is factored into these 3 levels.
- Sex predictor is changed to isMale and factored as a binary predictor with 1 indicating a male passenger and 0 a female passenger.

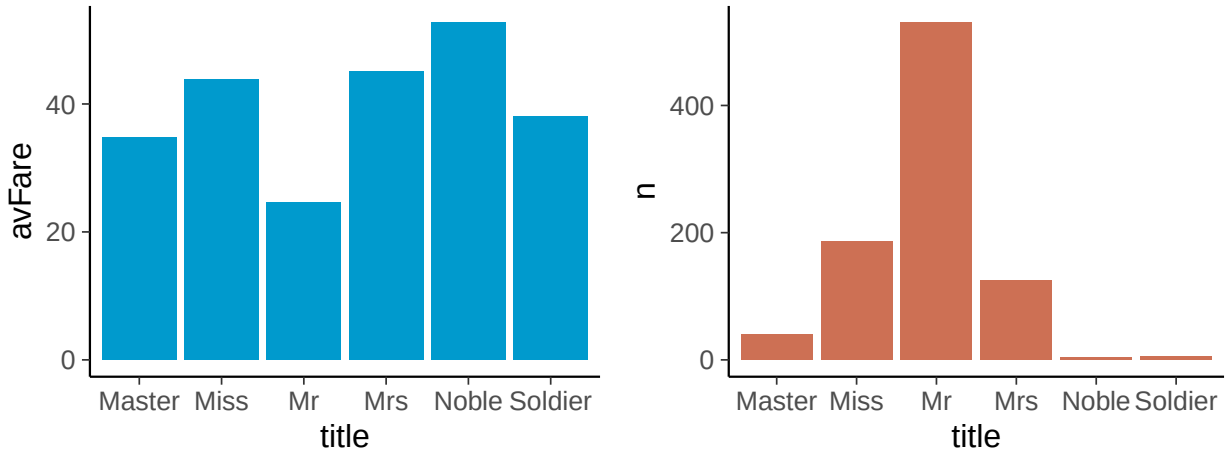
The data summary for training and test sets combined can be seen below:

```
##   PassengerId   Pclass       Age        Ticket       Fare
##  Min.   :   1   1:323   Min.   : 0.17   2:  3   Min.   :  0.000
##  1st Qu.: 328   2:277   1st Qu.:21.00   3: 14   1st Qu.:  7.896
##  Median : 655   3:709   Median :28.00   4:249   Median : 14.454
##  Mean   : 655           Mean   :29.88   5:377   Mean   : 33.295
##  3rd Qu.: 982           3rd Qu.:39.00   6:620   3rd Qu.: 31.275
##  Max.   :1309           Max.   :80.00   7: 46   Max.   :512.329
##                         NA's   :263             NA's   :1
##      Cabin      Embarked       title          fSize     isMale
##  C      :  94   C   :270   Master : 61   large    : 82   0:466
##  B      :  65   Q   :123   Miss   :266   singleton:790   1:843
##  D      :  46   S   :914   Mr     :773   small    :437
##  E      :  41   NA's:  2   Mrs    :197
##  A      :  22              Noble  :  5
##  (Other):  27              Soldier:  7
##  NA's   :1014
```

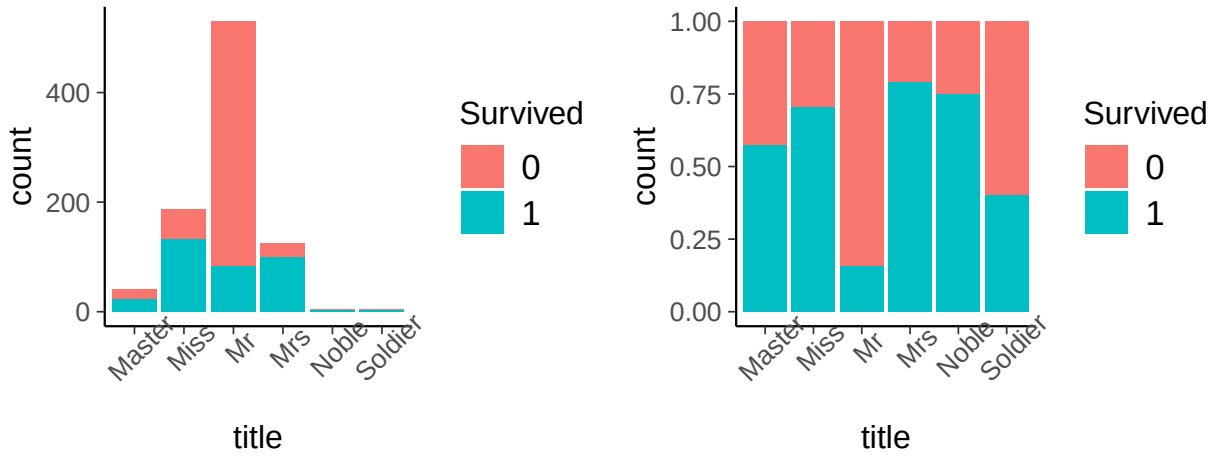
Age predictor has 263 missing observations, fare has 1, cabin has 1014 and embarked has 2 out of a total of 1309 observations. It seems still possible to replace the missing observations even in Age predictor but imputing the cabin predictor can bring a lot of noise to the data. So the cabin predictor might be dropped but the rest of the predictors will be definitely imputed. However, before getting to that part, it will be best to look at the data summary below showing the structure of each predictor:

```
## 'data.frame':    1309 obs. of  10 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Pclass     : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ Ticket     : Factor w/ 6 levels "2","3","4","5",..: 4 4 6 5 5 5 4 5 5 5 ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : Factor w/ 8 levels "A","B","C","D",..: NA 3 NA 3 NA NA 5 NA NA NA ...
##  $ Embarked   : Factor w/ 3 levels "C","Q","S": 3 1 3 3 3 2 3 3 3 1 ...
##  $ title      : Factor w/ 6 levels "Master","Miss",..: 3 4 2 4 3 3 3 1 4 4 ...
##  $ fSize      : Factor w/ 3 levels "large","singleton",..: 3 3 2 3 2 2 2 1 3 3 ...
##  $ isMale     : Factor w/ 2 levels "0","1": 2 1 1 1 2 2 2 2 1 1 ...
```

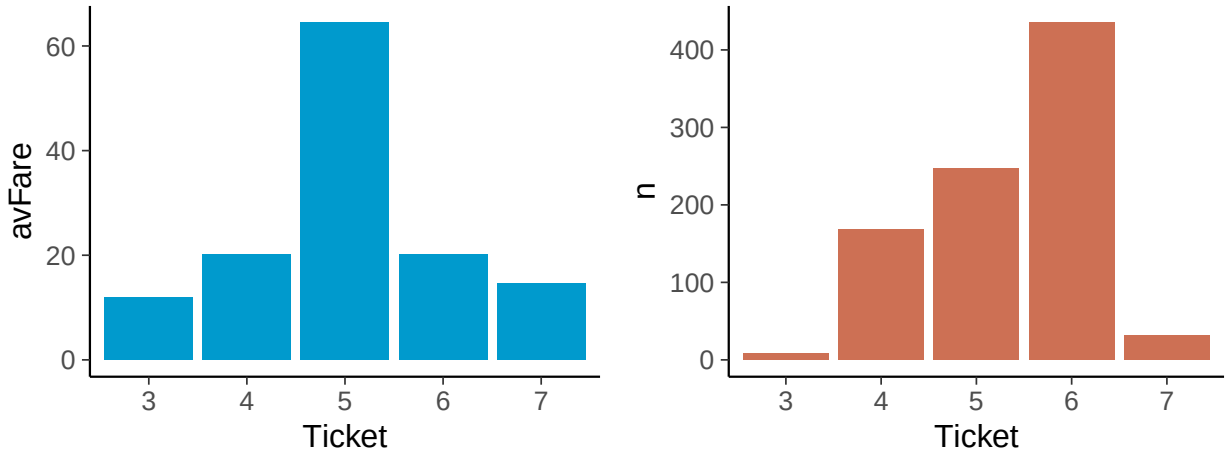
Another step to take before getting into data imputation is to look at the available data visually. However, to keep the test set information seperate, the exploratory visual checks will be performed only on training set. As a starting point to that, the first plot on below left shows the average fee paid for the ticket of each title. The one on the right shows the number of observations for each title:



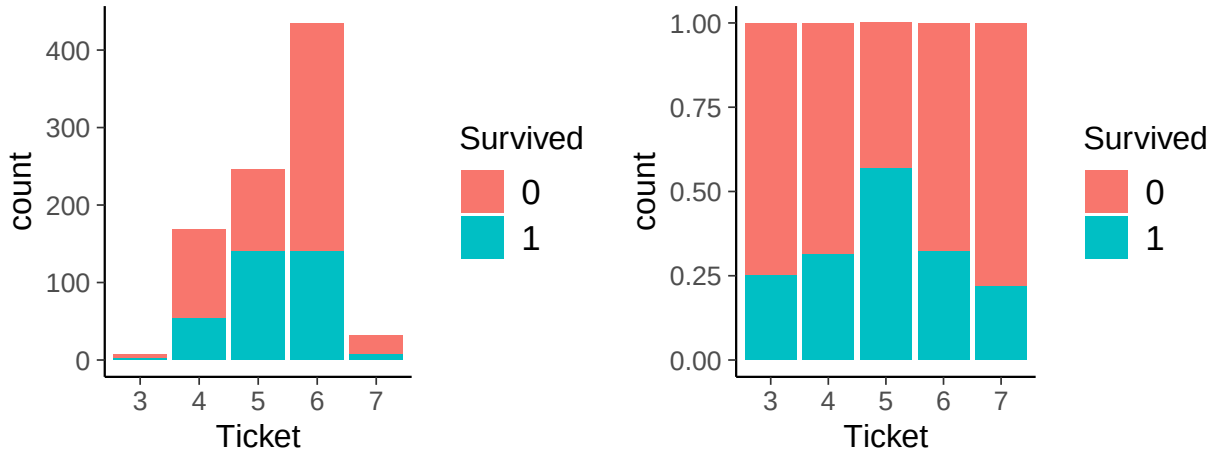
The average fee for the title make sense with Noble title having the highest average fee and Mr title having the lowest most probably because of the contribution of single male passengers that have entered the ship with a cheap ticket. The next two graphs below show the relationship between these titles and the survival rate. The one on the left shows the total number of counts represented by coloured bars with green representing the number of survived passengers and red representing the ones that haven't survived the accident. The normalized version showing the proportion of survived and not survived passengers for each title is shown on the below right graph:



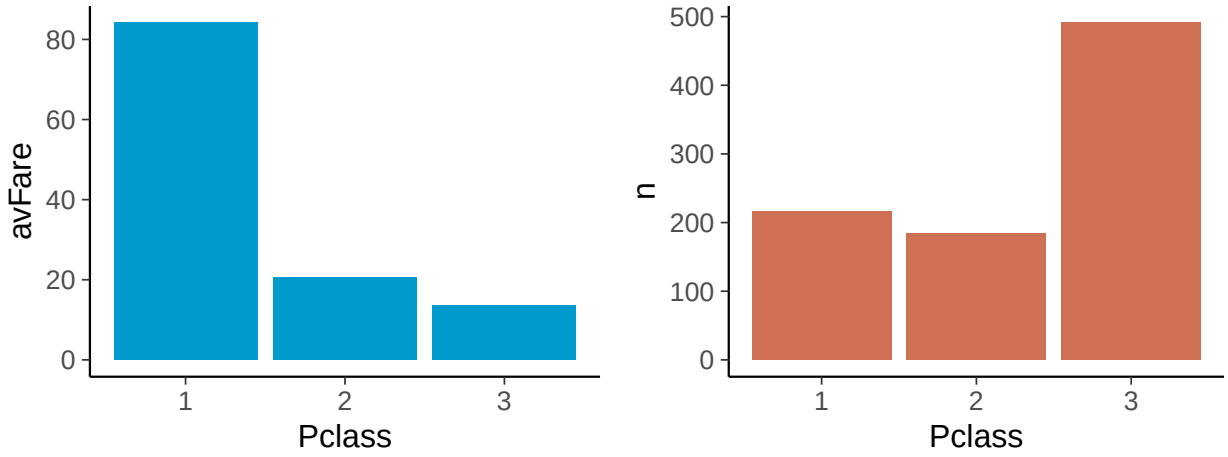
It can be seen that men with cheaper tickets have lower survival rates in comparison to the other titles on the ship. Young passengers, female passengers and nobles seem to have a higher survival rate. Next graph below shows the average fare per ticket type whereas the one on the right shows the number of observations for each ticket type:



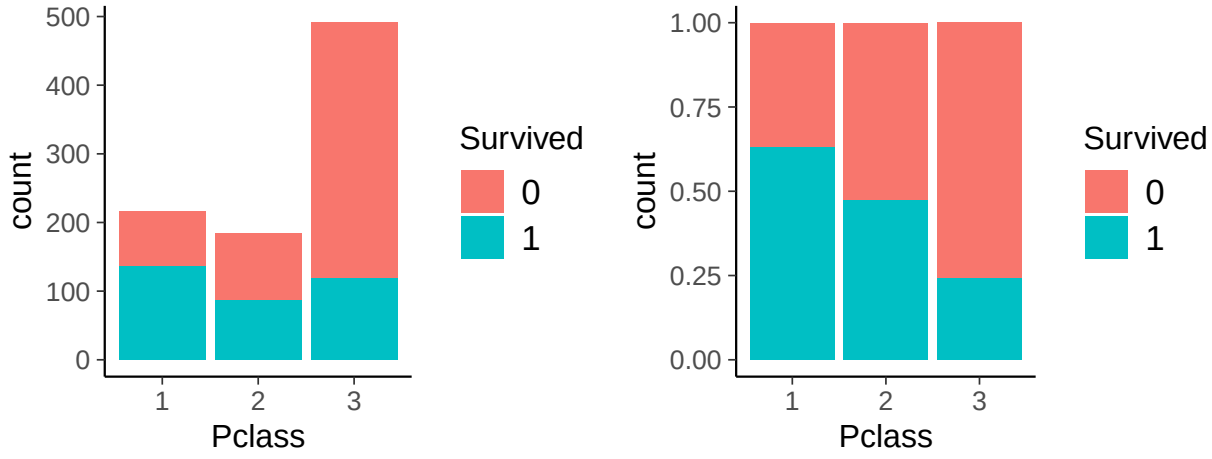
It can be seen that the average ticket prices are similar between types 3 and 7 and between types 4 and 6. Ticket type 5 is highest compared to other ticket types. On the other hand, types 3 and 7 are owned by lowest number of passengers on the ship whereas type 6 is owned most of the passengers. As mentioned before, the visual exploratory checks are performed only on the training set. However in the summary table above, there was also a ticket type 2 which is coming from the test set. It will be hard to make a prediction about this ticket type if we can't model it on the training set. There are 3 passengers that have ticket type 2, 14 passengers with ticket type 3 and 46 passengers with ticket type 7. Based on the assumption that the average fee paid for these ticket types are similar, it might make sense to combine these three ticket types before modeling. The next two graphs below show the relationship between these ticket types and the survival rate. The one on the left shows the total number of counts represented by coloured bars with green representing the number of survived passengers and red representing the ones that haven't survived the accident. The normalized version showing the proportion of survived and not survived passengers for each ticket type is shown on the below right graph:



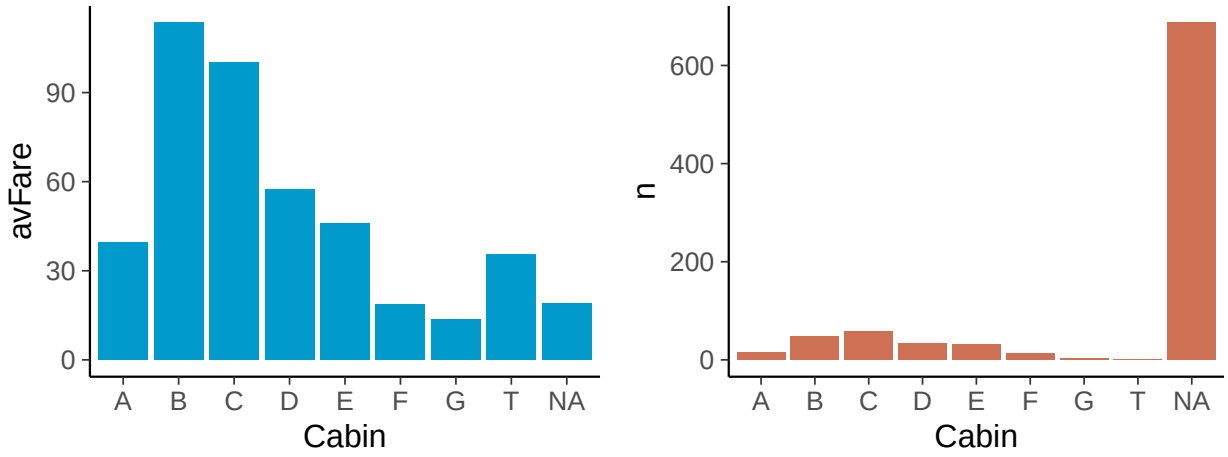
It is a bit hard to make separate conclusions for each ticket type but an obvious one is that the passengers having expensive ticket types have higher survival rates. Next graph below shows the average fare per passenger class whereas the one on the right shows the number of observations for each passenger class:



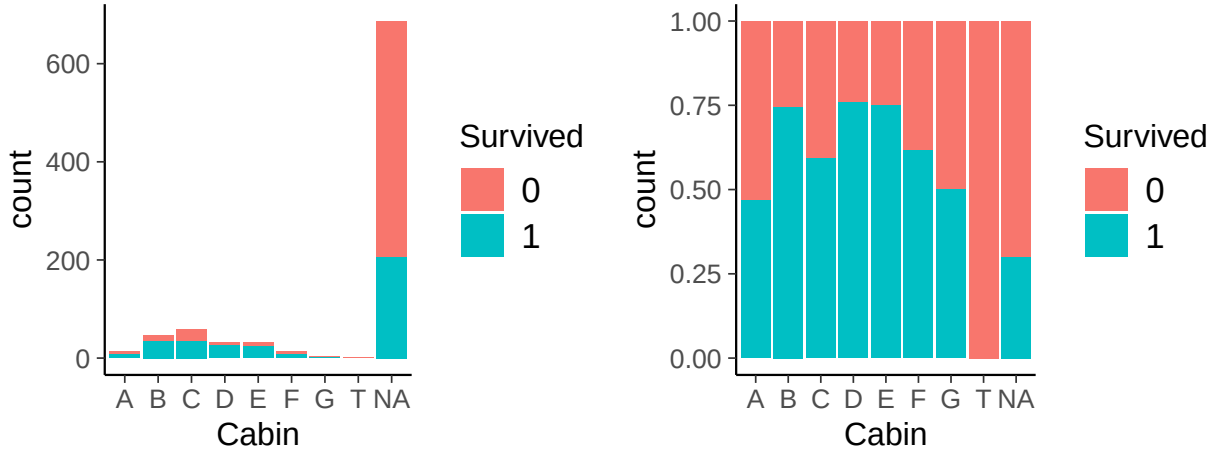
As expected, class 1 passengers pay the highest price compared to other passenger classes. It can also be observed that Class 3 passengers have the highest number. The interesting observation is that class 2 passengers are less in amount in comparison to class 1 passengers. The next two graphs below show the relationship between these passenger classes and the survival rate. The one on the left shows the total number of counts represented by coloured bars with green representing the number of survived passengers and red representing the ones that haven't survived the accident. The normalized version showing the proportion of survived and not survived passengers for each passenger class is shown on the below right graph:



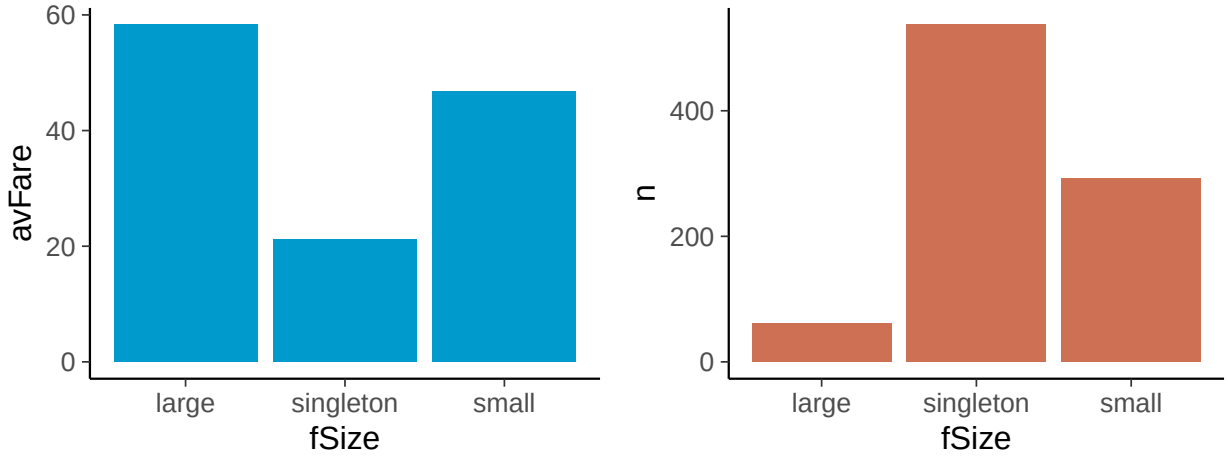
The survival rate is proportional to the passenger class with passengers that pay the highest price survive the most proportionally. Next we will check the cabin predictor in the same manner.



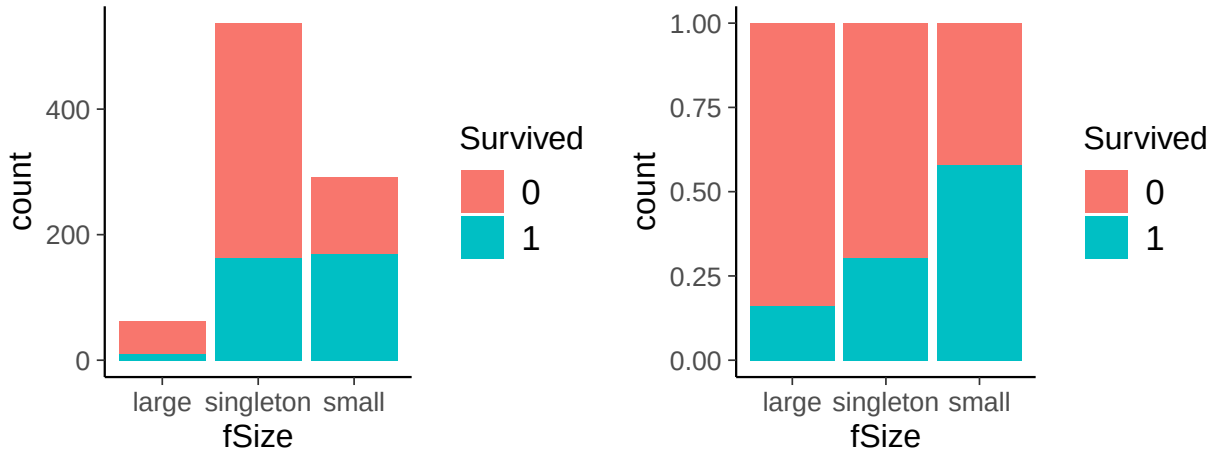
As mentioned before, the number of missing values for the cabin predictor is very high. This means either the cabin predictor will be dropped from analysis, which will result in loss of information, or a logic will be applied for imputation of the missing values. The following graphs show the relationship of the cabin predictor to survival rate:



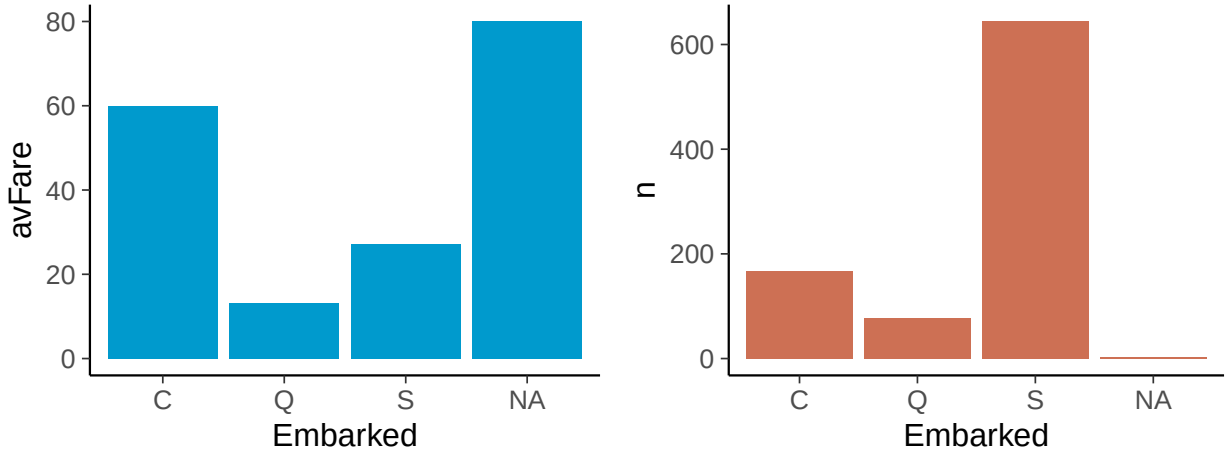
There is not a particular pattern that can be noticed immediately but again cabins with passengers that have paid higher ticket fees tend to have higher survival rate. Next we will check the family size predictor in similar fashion to other predictors above:



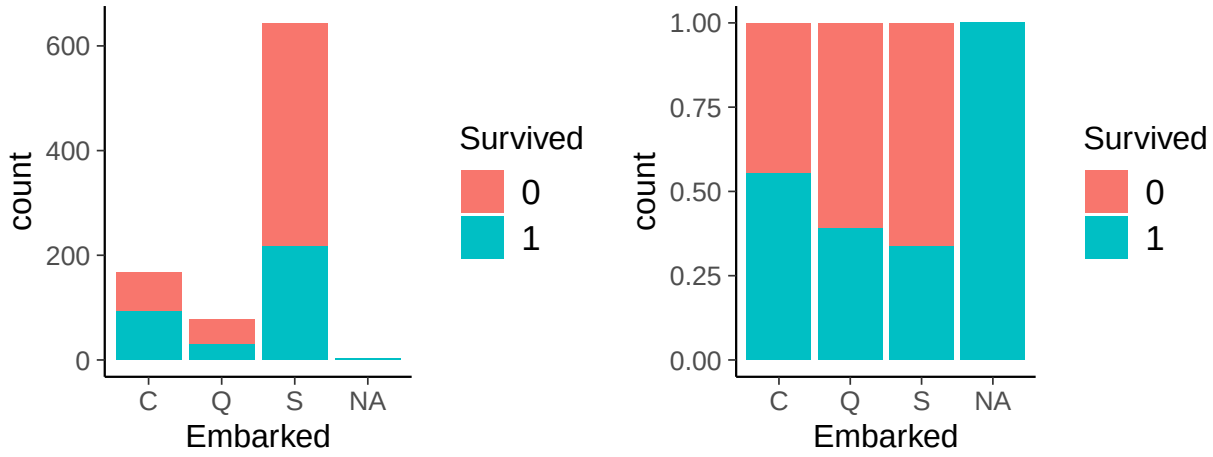
Large families have higher average fare but less in total number. The highest number of individuals are single ones whereas small families also make up a big part of the total number of passengers. The next two graphs show the family size relationship to survival rate:



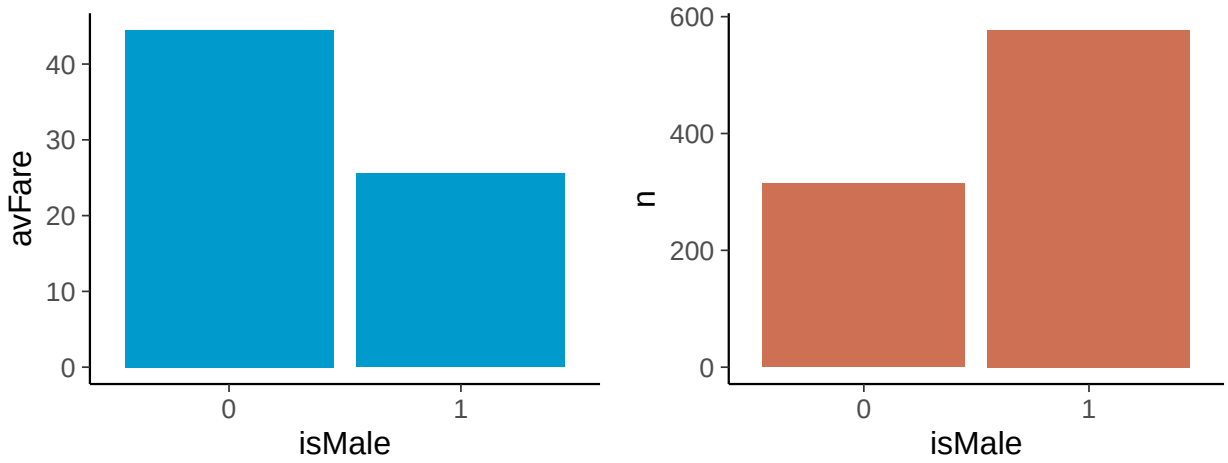
It can be seen that large families and single passengers don't have high survival rate whereas small families have a higher survival rate. Still, the total number of survived single passengers and small families are very close to each other. The next predictor that is going to be observed is the embarkation port:



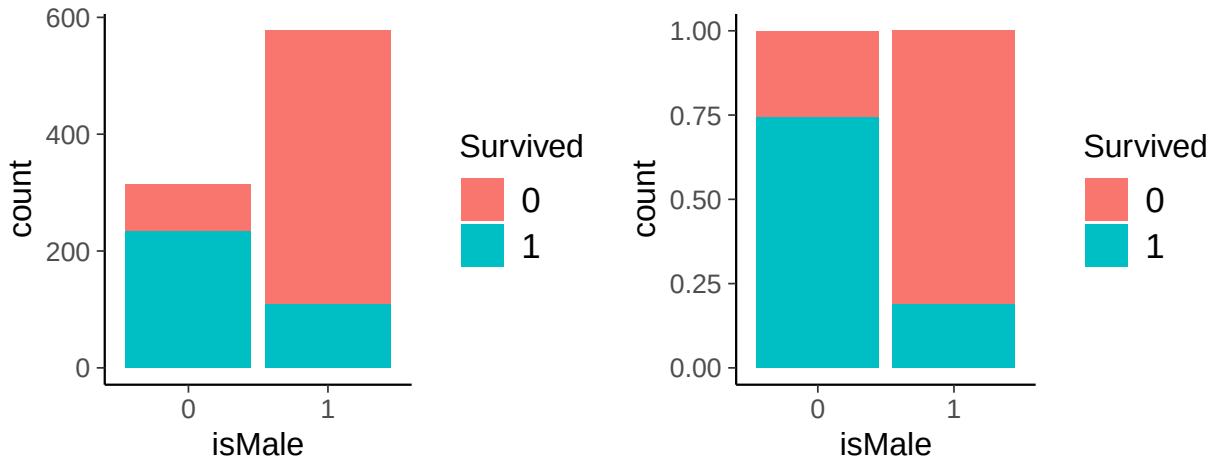
The figure on the upper left shows the average fare for passengers that embarked from different ports. The 2 missing values have the highest rate. The passengers that embarked from port S are the highest in number. Next, the relationship to survival rate will be checked:



It can again be seen that passengers that have paid the highest rate for their tickets in average embarked from a similar port and have a higher survival rate. We will check the sex variable below in similar fashion although it doesn't totally make sense to check the ticket fee for each sex. We will do it for the sake of keeping the same format but will add more specific plots for the predictors:



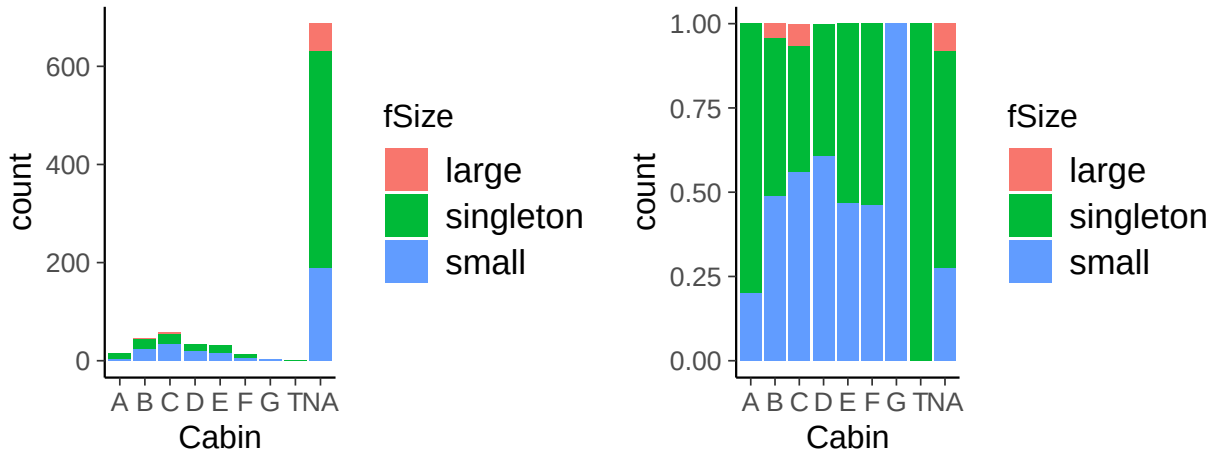
Interestingly, it looks like female passengers have paid more in average for their tickets. There are more male passengers on board which might show the indication that the male passengers are part of a lower passenger class. We will check the relationship to the survival rate now:



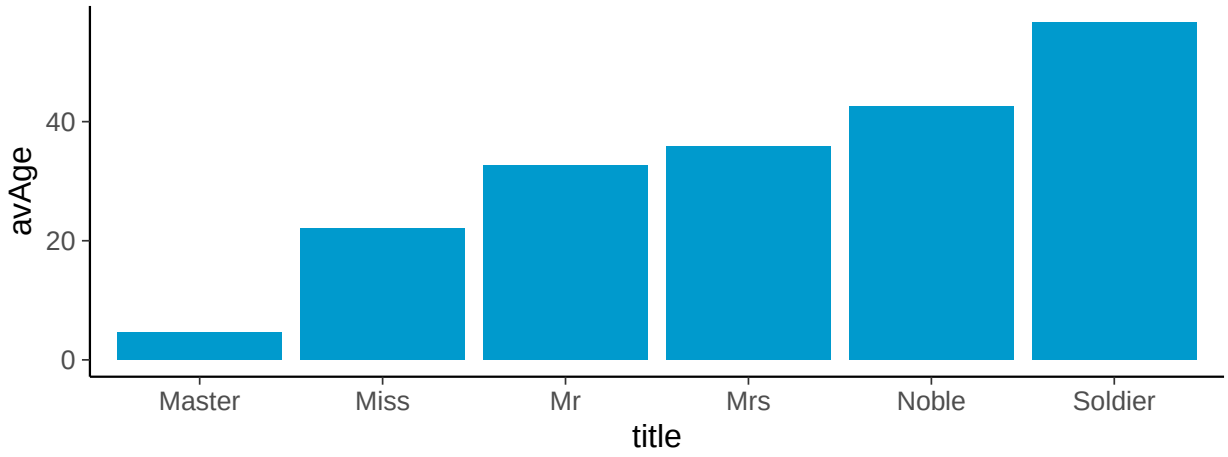
It can be easily noticed that the survival rate is high among female passengers. This predictor is highly correlated to the suvival rate.

We've checked all the categorical variables in relation to the average fee and survival rate. However, it doesn't actually make total sense to check all of the categorical variables in relation to average ticket fee such as cabin type or titles. The following graphs will try to add more meaningful visual checks for the available data.

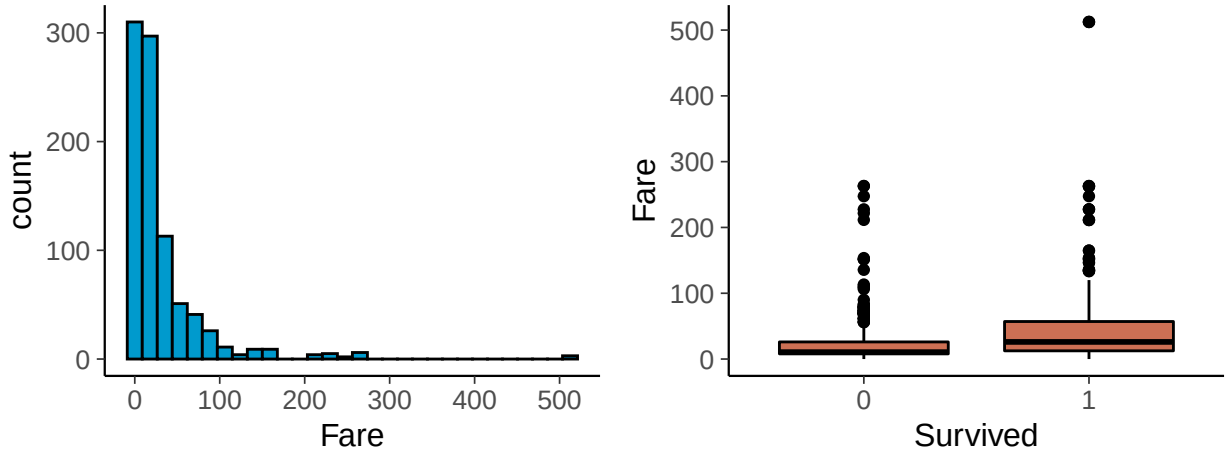
First we will check the relationship between the cabin and the family size:



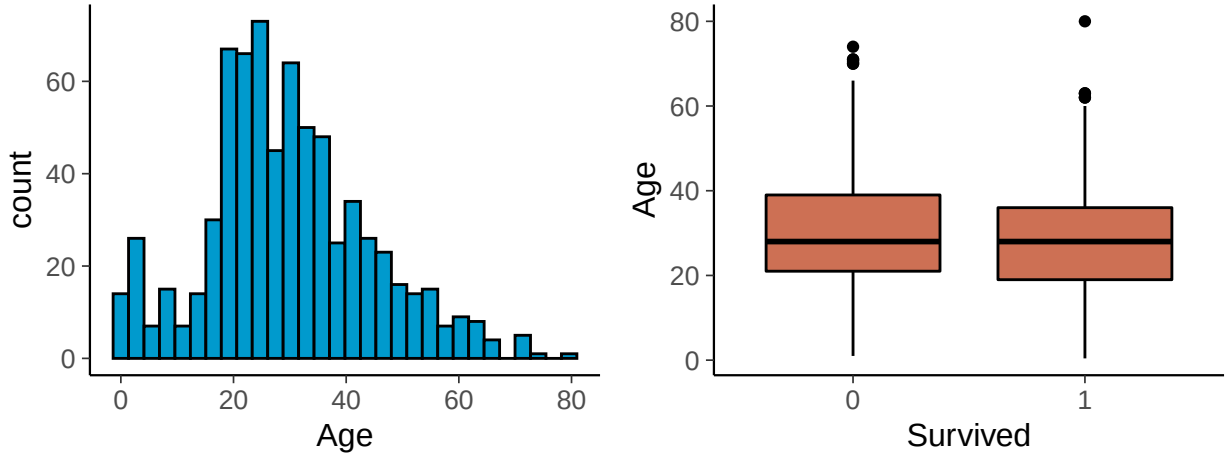
The number of missing values are so high that the graph on the left above doesn't give much information. Data imputation is necessary to be able to get something meaningful out of the above plot. We will get back to this plot after the imputation is finalized in the next section. We can now check the average age of each title to be able to assess if it makes sense to impute the missing age observations based on the title:



It can be seen that there is correlation between the average age and the title so it makes sense to use title to impute the missing age values. However, before that let's have a look into the continous predictor fare for tickets.



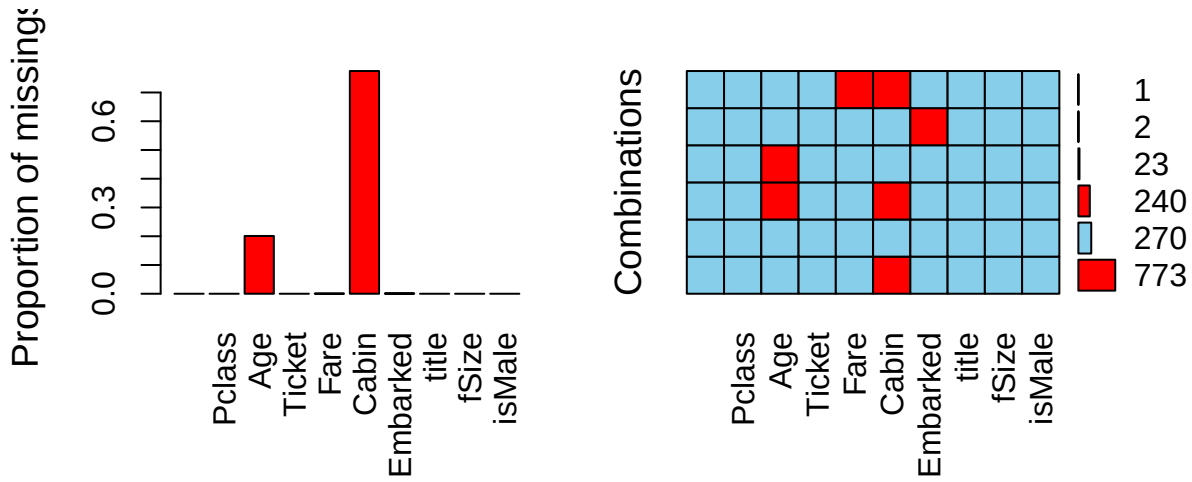
The fare predictor has a skewed distribution so a log transformation or normalization might be necessary. Age predictor has a lot of missing observations so we will look at the available data for now in the same manner as the fare predictor:



We can see a slight implication on the above right figure of younger passengers having a higher survival rate in general. In the next sub-section, we will try to impute the missing observations, modify the data to its final form before modelling and show some final visual exploratory checks.

2.1 Data Imputation

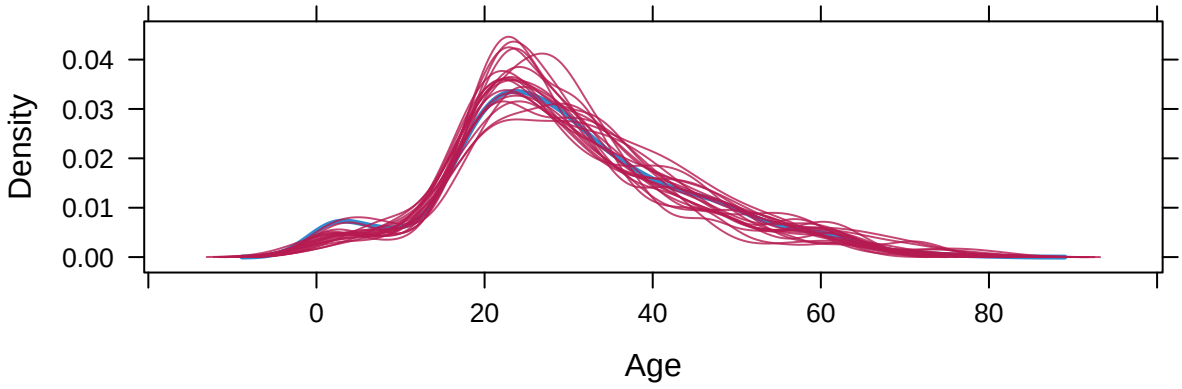
In the summary table above for a combined dataset of both training and test datasets, we have observed 263 missing observations for Age predictor, 1 for Fare predictor, 1014 for Cabin predictor and 2 for Embarked predictor. We will try to impute all the missing observations although an evaluation needs to be made for the Cabin predictor if the imputed dataset makes sense. First let's have a look at a graphical representation of the missing observations:



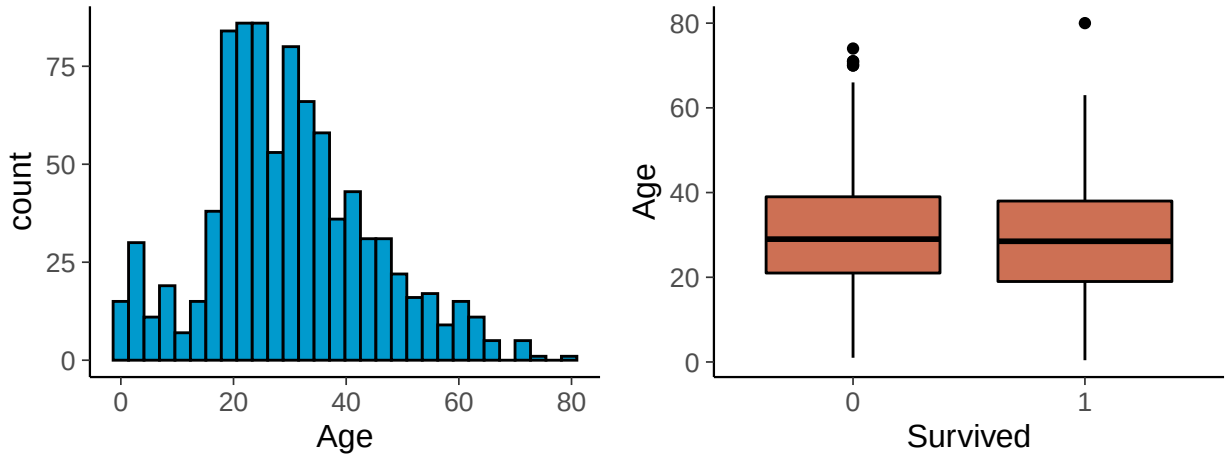
The plot on the left shows the percentage of the missing data per predictor. It can be seen as mentioned before as well that Age and Cabin predictors have the highest number of missing observations. The plot on the right shows the missing observations as a combination. Each row can be read as a specific combination of observations with columns representing the predictors. For example the first row represents the single observation where both Fare and Cabin predictors are missing at the same time. The second row represents the two observations where missing values are for the Embarked predictor. The same goes for all the combinations of missing observations with the fifth row representing 270 observations without any missing values. The methodology for imputation of each predictor will be explained in the following paragraphs. The imputation process will be carried out for both training and test sets combined.

We will start imputing the missing observations in the Age predictor. We wil replace the missing observations using mice package which is implementing multiple imputation methods for missing observations. The details of the algorithm will not be discussed here but they can be accessed via the package official document. The Age predictor will be imputed by using title and family size predictors as they are assessed to be the most suitable ones for predicting the missing Age observations. The method implemented will be weighted predictive mean matching and 20 multiple imputations will be performed.

The plot below shows the density plot for the Age predictor with blue curve representing the observed data and the red curves representing the imputed data for each imputation:



It can be seen from the above plot that the imputed data is very close to the existing data. Now we can have a look at the exploratory graphs for the Age predictor after imputation for all the missing observations is completed. The visualization will represent only the training set values as done before:



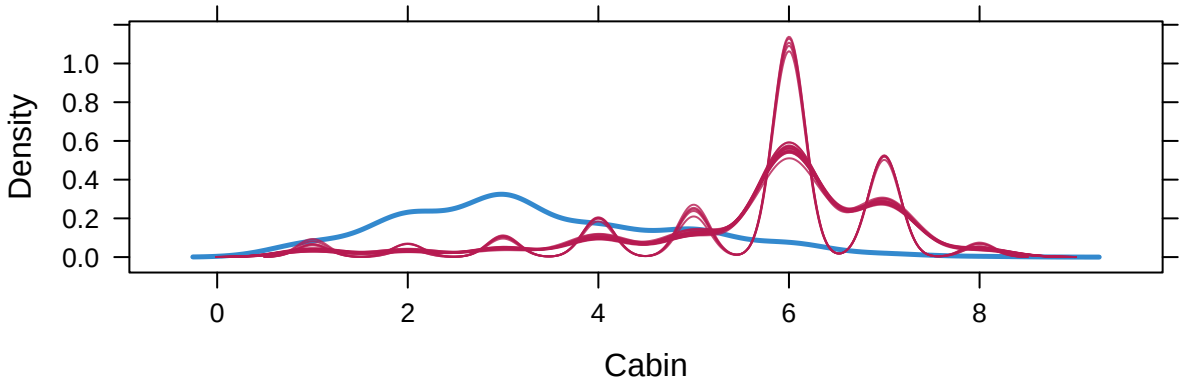
It can be observed above that the number of missing observations are mostly around 30 years of age. Nevertheless, the general shape is very similar to the raw data.

Next we will look at the imputation of 2 missing observations from Embarked predictor. As the number of missing observations is very low compared to the total number of observations and as the missing observations have both survived the accident, they will be imputed as part of the factor level, which is "C".

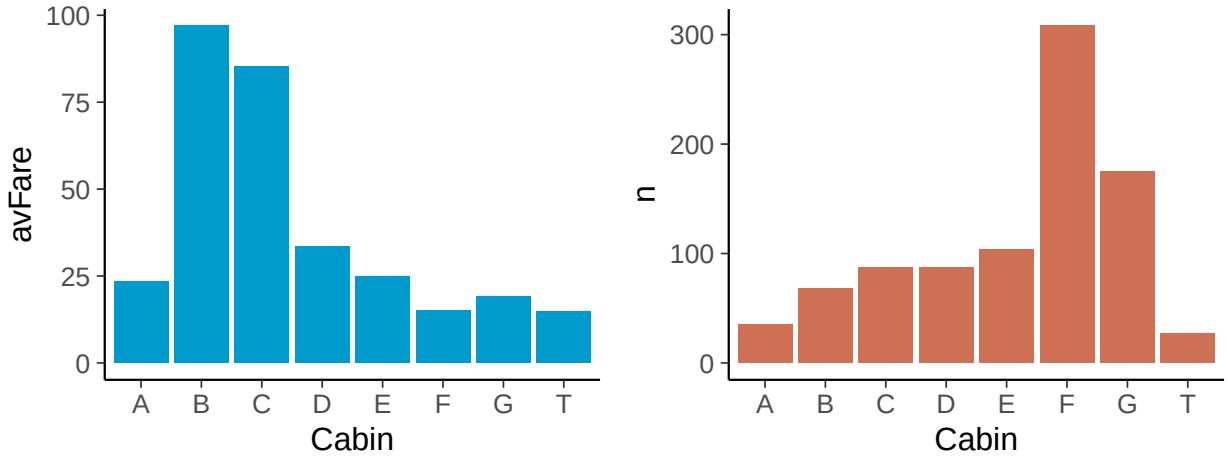
The other missing observation is from Fare predictor and it will be imputed by replacing the missing observation with the median value of the existing observations. Again, this imputation isn't expected to have a huge effect on the analysis.

The final predictor to be imputed is Cabin. It has the most amount of missing observations so it is expected to add noise to the analysis. Again mice package will be used for imputation using all the available and previously imputed data with the method of polytomous logistic regression for prediction of missing values.

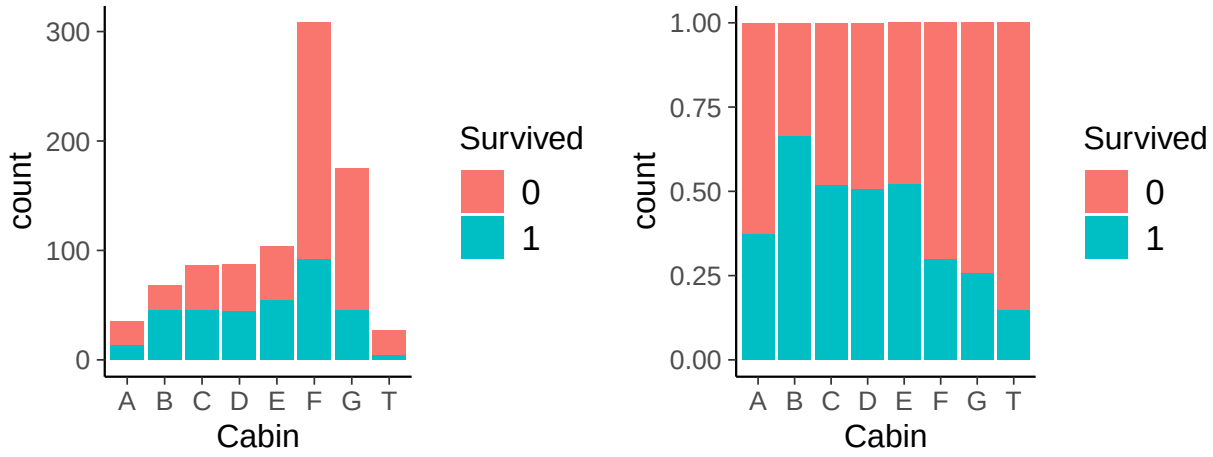
Again the density plot for the observed data and the imputed data can be seen in the below graph:



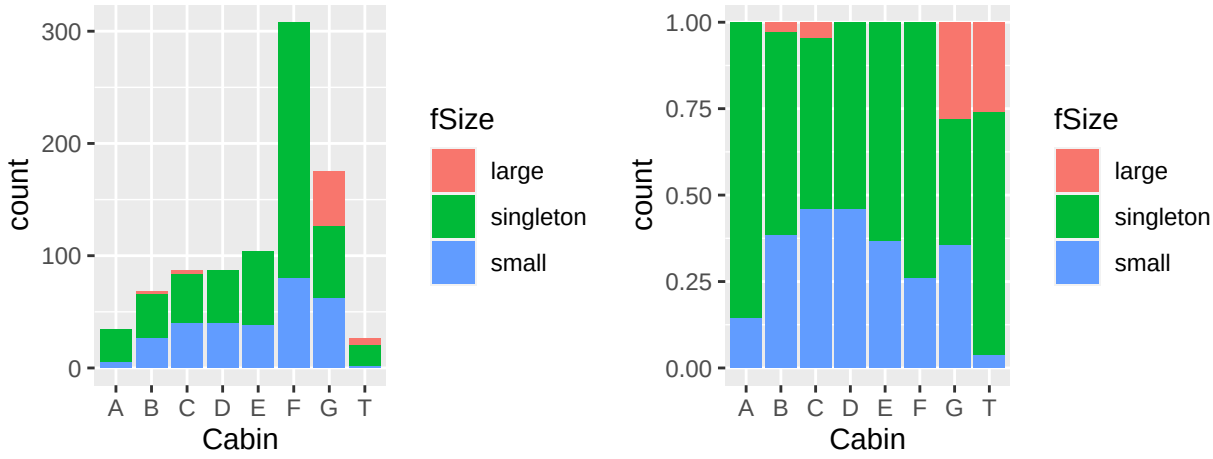
The x-axis numbers actually represent the levels of the Cabin predictor starting from "A". The imputed data shows a different pattern than the available data which can be expected due to high number of missing observations in comparison to the existing ones. It can be seen in the above plot that the imputed values are gathered around level "6" which represents the cabin type "F". Now it makes sense to look at the exploratory graphs one more time:



It can be seen from the graph above that cabin types "B" and "C" have the passengers with the highest average fare whereas the number of passengers are highest in cabin types "F" and "G". The following graphs show the relationship of the cabin predictor to survival rate:



It can be observed that the correlation between the passengers that reside in cabins with highest ticket fee tend to have higher survival rates. We can also duplicate the plot where the relationship between cabin and family size can be seen after the imputations are completed:



As expected, cabins for larger families tend to have lower survival rate whereas the ones for small families tend to have higher survival rate. Next, we will look into final transformations of predictors before we start the modelling.

2.2 Data Transformation

In this section, we will look into transforming the imputed data to its final shape. We will start with the title predictor. It was observed that "Noble" and "Soldier" titles had few observations compared to other factor levels and therefore could end up in unbalanced groups if hierarchical modeling is going to be applied. The number of observations for each title group can be seen below. The number of individuals are presented for both training and test data sets combined:

```
##  Master    Miss      Mr     Mrs   Noble Soldier
##      61     266     773     197       5       7
```

The "Noble" title group will be combined with the "Master" title group and will be named "Other". The "Soldier" title group will be combined with the "Mr" title group. These decisions are made based on the similarity of the relationship to the survival rate.

The title factor levels have changed as below after the transformation:

```
##  Miss    Mr   Mrs Other
##   266   780   197    66
```

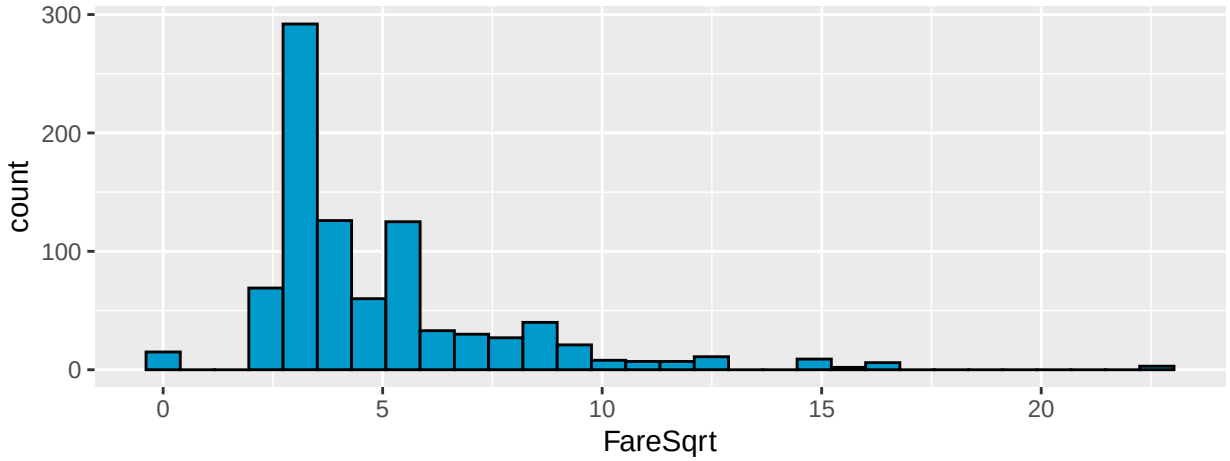
Next we we will look into the ticket predictor. As mentioned in the initial part of the exploratory analysis, the test set had a ticket type "2" that the training set didn't have. The number of type "2" tickets are very low like in types "3" and "7" with similar fares. We can have a look at the number of observations for each level for both training and test datasets below:

```
##   2   3   4   5   6   7
##   3  14 249 377 620  46
```

The transformation here would be to combine types "2", "3" and "7" as "Other" ticket type so that the modeling can be performed with the training set and proper predictions can be applied to the test set. The final number of observations for each ticket type can be seen below after the transformation is applied:

```
##     4     5     6 Other
##   249   377   620    63
```

The final transformation will be applied to the Fare predictor. As has been observed, this predictor has a skewed distribution and it will be beneficial to use the transformed version. The squareroot will be used for transforming the positive skew of Fare predictor. The distribution for the transformed version of this predictor for only training set can be seen below:



It can be observed that the transformation wasn't enough to get rid of the whole skewness. The summary of the training dataset can be seen below:

```
##   PassengerId    Survived Pclass       Age          Ticket        Cabin
##  Min.   :  1.0   0:549    1:216   Min.   : 0.42   4    :169   F      :308
##  1st Qu.:223.5   1:342    2:184   1st Qu.:21.00   5    :247   G      :175
##  Median :446.0            3:491   Median :29.00   6    :435   E      :104
##  Mean   :446.0                    Mean   :29.99   Other: 40   C      : 87
##  3rd Qu.:668.5                    3rd Qu.:39.00               D      : 87
##  Max.   :891.0                    Max.   :80.00               B      : 68
##                                                               (Other): 62
##  Embarked   title           fSize     isMale     FareSqrt
##  C:170    Miss :187   large    : 62   0:314   Min.   : 0.000
##  Q: 77    Mr   :535   singleton:537   1:577   1st Qu.: 2.813
##  S:644    Mrs  :125   small    :292           Median : 3.802
##           Other: 44                           Mean   : 4.851
##                                               3rd Qu.: 5.568
##                                               Max.   :22.635
##
```

Both the training and the test datasets have been nomalized to be used for analysis and prediction. The summary of the normalized version of the training set can be seen below:

```
##     Pclass2           Pclass3            Age              Ticket5
##  Min.   :-0.5099   Min.   :-1.1073   Min.   :-2.04553   Min.   :-0.619
##  1st Qu.:-0.5099   1st Qu.:-1.1073   1st Qu.:-0.62195   1st Qu.:-0.619
##  Median :-0.5099   Median : 0.9021   Median :-0.06857   Median :-0.619
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.000
##  3rd Qu.:-0.5099   3rd Qu.: 0.9021   3rd Qu.: 0.62317   3rd Qu.: 1.614
##  Max.   : 1.9591   Max.   : 0.9021   Max.   : 3.45926   Max.   : 1.614
##     Ticket6          TicketOther          CabinB            CabinC
##  Min.   :-0.9762   Min.   :-0.2167   Min.   :-0.2873   Min.   :-0.3288
##  1st Qu.:-0.9762   1st Qu.:-0.2167   1st Qu.:-0.2873   1st Qu.:-0.3288
##  Median :-0.9762   Median :-0.2167   Median :-0.2873   Median :-0.3288
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 1.0233   3rd Qu.:-0.2167   3rd Qu.:-0.2873   3rd Qu.:-0.3288
##  Max.   : 1.0233   Max.   : 4.6099   Max.   : 3.4770   Max.   : 3.0383
##      CabinD            CabinE            CabinF            CabinG
##  Min.   :-0.3288   Min.   :-0.3633   Min.   :-0.7264   Min.   :-0.4941
##  1st Qu.:-0.3288   1st Qu.:-0.3633   1st Qu.:-0.7264   1st Qu.:-0.4941
##  Median :-0.3288   Median :-0.3633   Median :-0.7264   Median :-0.4941
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.:-0.3288   3rd Qu.:-0.3633   3rd Qu.: 1.3750   3rd Qu.:-0.4941
##  Max.   : 3.0383   Max.   : 2.7493   Max.   : 1.3750   Max.   : 2.0216
##      CabinT          EmbarkedQ         EmbarkedS          titleMr
##  Min.   :-0.1767   Min.   :-0.3074   Min.   :-1.614   Min.   :-1.2252
##  1st Qu.:-0.1767   1st Qu.:-0.3074   1st Qu.:-1.614   1st Qu.:-1.2252
##  Median :-0.1767   Median :-0.3074   Median : 0.619   Median : 0.8153
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.000   Mean   : 0.0000
##  3rd Qu.:-0.1767   3rd Qu.:-0.3074   3rd Qu.: 0.619   3rd Qu.: 0.8153
##  Max.   : 5.6537   Max.   : 3.2495   Max.   : 0.619   Max.   : 0.8153
##     titleMrs          titleOther      fSizesingleton      fSizesmall
##  Min.   :-0.4037   Min.   :-0.2278   Min.   :-1.2310   Min.   :-0.6978
##  1st Qu.:-0.4037   1st Qu.:-0.2278   1st Qu.:-1.2310   1st Qu.:-0.6978
##  Median :-0.4037   Median :-0.2278   Median : 0.8115   Median :-0.6978
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.:-0.4037   3rd Qu.:-0.2278   3rd Qu.: 0.8115   3rd Qu.: 1.4315
##  Max.   : 2.4741   Max.   : 4.3850   Max.   : 0.8115   Max.   : 1.4315
##     isMale1           FareSqrt
##  Min.   :-1.3548   Min.   :-1.6466
##  1st Qu.:-1.3548   1st Qu.:-0.6920
##  Median : 0.7373   Median :-0.3562
##  Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.7373   3rd Qu.: 0.2432
##  Max.   : 0.7373   Max.   : 6.0362
```

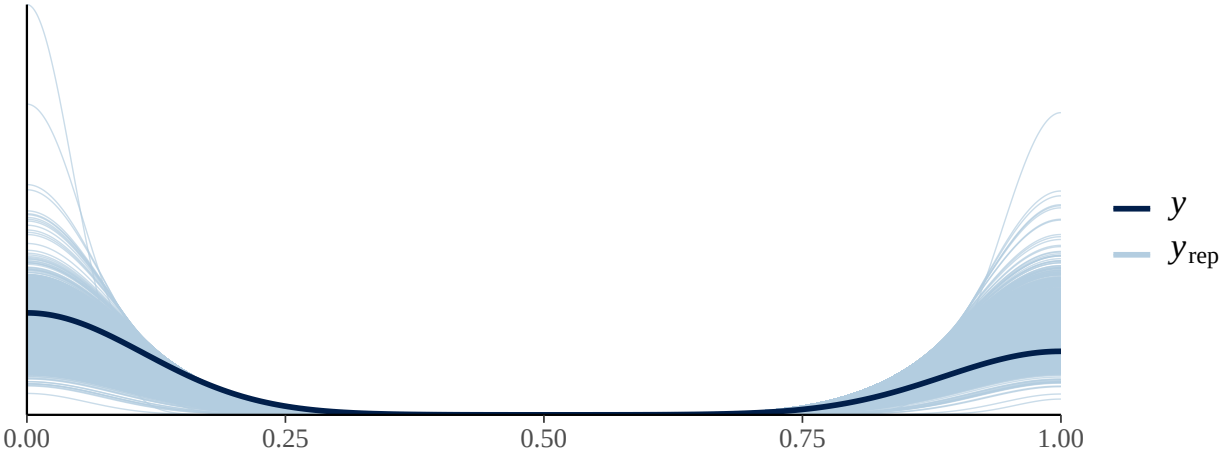
Now, both the training and test datasets are ready for prior predictive checking.

3- Prior Predictive Checking

In this section, we will perform a prior predictive check by only using the priors and no data.The reason for making prior predictive analysis is to make a sanity check on the priors without using the likelihood.

In our analysis, we will look into different models where we will need to move some of the predictors as a grouping variable and the rest for estimating the coefficients. Such an example can be a hierarchical

model where we might need to use the title predictor for grouping the rest of the variables according to their title group. In any case, we will use weakly informative robust prior of student_t(7,0,2.5) for both population-level means with an lkj(2) distribution for the correlation matrix of the group level effects using a multivariate normal distribution. The estimated parameter coefficients will be used in a generalized linear model to estimate the bernoulli logit input. Below we will demonstrate the prior predictive distribution for the non-hierarchical model including all the predictors. It can also be named as a fully pooled model consisting of only population level effects:



In the above figure, the solid black line is the actual distribution of the survival rate in the training set represented by "y" and the blue lines are the simulations of the survival rate using parameters picked from a robust student-t prior distribution represented by "y-rep". A bernoulli model with a logit link function has been used for modelling. By looking at the figure, it can be concluded that the chosen prior is suitable for analysis and prediction using the same model.

4 - Model Fitting and Algorithm Diagnostics

We will start model fitting with 8 different models which are decided on a partially random logic. The descriptions of the models are below:

- **Model 1:** Only intercept model with a single population level effect using a student-t(7,0,2.5) prior.
- **Model 2:** All predictors model with population effects only. It's the same model used for prior predictive checking using a student-t(7,0,2.5) prior.
- **Model 3:** Again all predictors model with population effects only but this time using a horseshoe prior model with a coefficient of 1 for making it weakly informative.
- **Model 4:** Only intercept model with both population level and group level effects using a student-t(7,0,2.5) prior. The grouping variable is title.
- **Model 5:** All predictors model with population and group level effects using a student-t(7,0,2.5) prior and a lkj(2) for the covariance matrix of the parameters variation. The grouping variable is title.
- **Model 6:** All predictors model with population and group level effects using a student-t(7,0,2.5) prior and a lkj(2) for the covariance matrix of the parameters variation. The grouping variable is pClass.
- **Model 7:** All predictors model with population and group level effects using a student-t(7,0,2.5) prior and a lkj(2) for the covariance matrix of the parameters variation. The grouping variable is Cabin.
- **Model 8:** All predictors model with population and group level effects using a student-t(7,0,2.5) prior. The grouping variable is Cabin. The normal model will be used for estimating the parameters of the generalized linear model instead of multivariate linear model.

5 - Posterior Predictive Checking

6 - Additional Models and Model Improvements

7 - Model Comparison

8 - Prediction Submission