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, to be able to observe the data at hand, the existing observations for different predictors will be formatted into data structures that can be handled by R in a meaningful way.

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.
- Titles have been extracted from the name variable and factored into Mr, Mrs, Miss, Master, Noble and Soldier levels.
- Cabin predictor is transformed into factor type.
- Embarked predictor is transformed into factor type.
- Ticket predictor is factored into L, F, P, C, A, S, W based on the letter they contain and into the number of digits if there is no letter in the ticket name. 3, 7, A, F, L, W types are combined into "Other" level as there are not so many observations in those factored levels.
- 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:

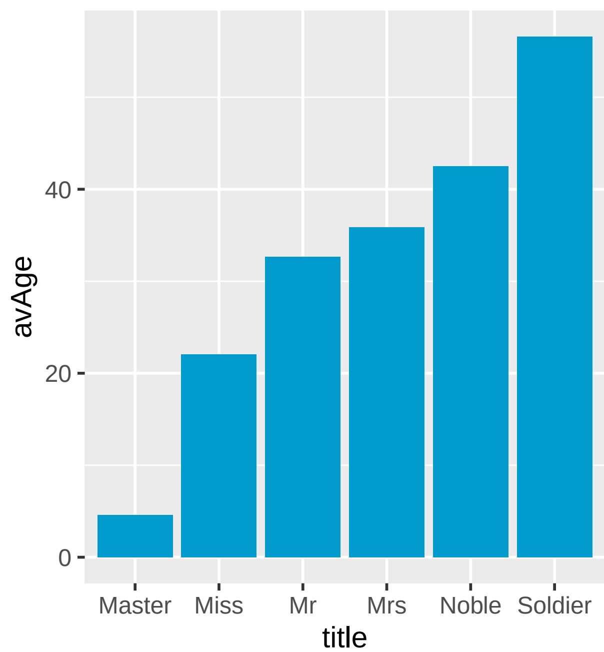
```
##   PassengerId   Pclass       Age          Ticket        Fare
##  Min.   :   1   1:323   Min.   : 0.17   4    :144   Min.   :  0.000
##  1st Qu.: 328   2:277   1st Qu.:21.00   5    :193   1st Qu.:  7.896
##  Median : 655   3:709   Median :28.00   6    :596   Median : 14.454
##  Mean   : 655           Mean   :29.88   C    : 88   Mean   : 33.295
##  3rd Qu.: 982           3rd Qu.:39.00   Other: 99   3rd Qu.: 31.275
##  Max.   :1309           Max.   :80.00   P    :132   Max.   :512.329
##                         NA's   :263     S    : 57   NA's   :1
```

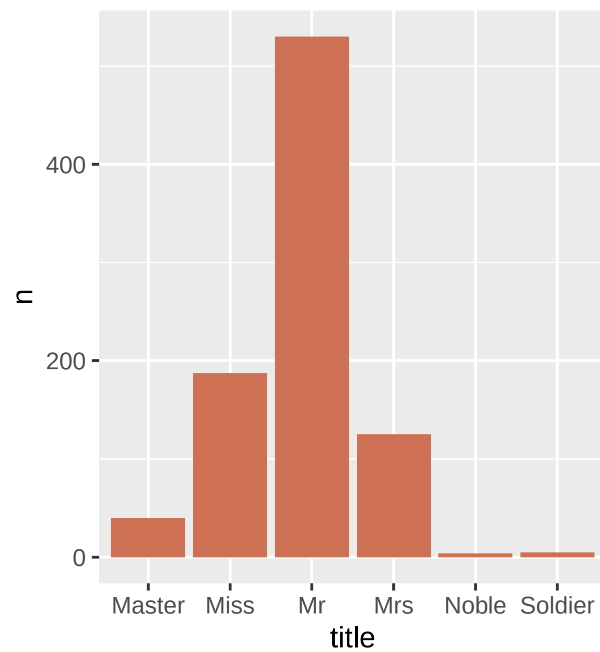
```
##      Cabin      Embarked       title          fSize     isMale
##  C      :  94   C   :270   Master : 61   large    : 82   0:466
##  B      :  65   Q   :123   Miss   :266   singleton:790   1:843
##  D      :  46   S   :914   Mr     :773   small    :437
##  E      :  41   NA's:  2   Mrs    :197
##  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 will be dropped but the rest of the predictors will be imputed. However, before getting to that part, it will be best to look at the data summary below showing the structure of each predictor:

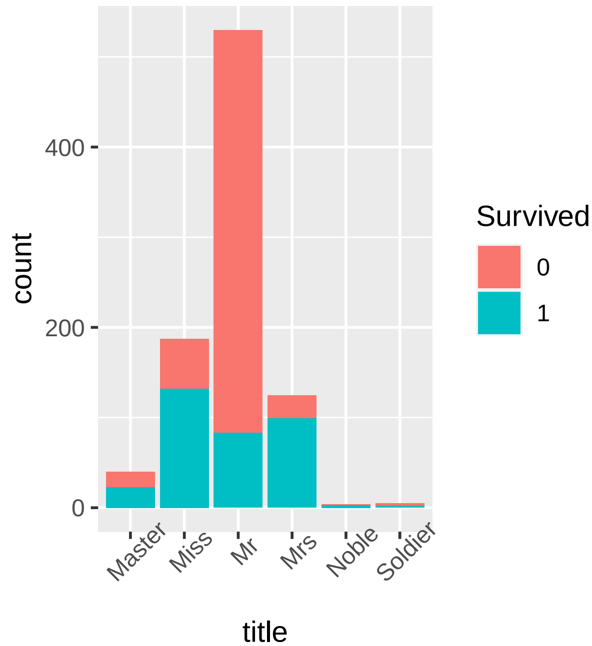
```
## 'data.frame':    1309 obs. of  10 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Pclass     : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ Ticket     : Factor w/ 7 levels "4","5","6","C",..: 5 6 7 3 3 3 2 3 3 3 ...
##  $ 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 ...
```

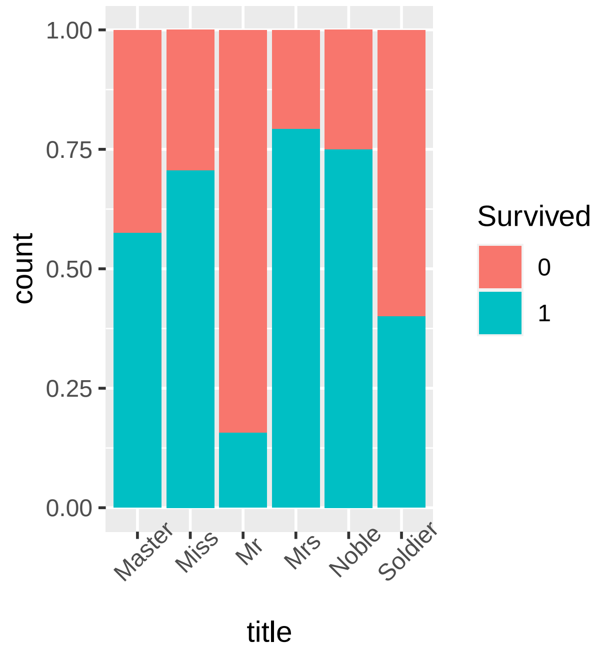
Another step to take before getting into data imputation is to look at the available data visually. However, to keep the test set information seperate, the exploratory visual checks will be performed only on training set. As a starting point to that, the first plot on below left shows the average age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



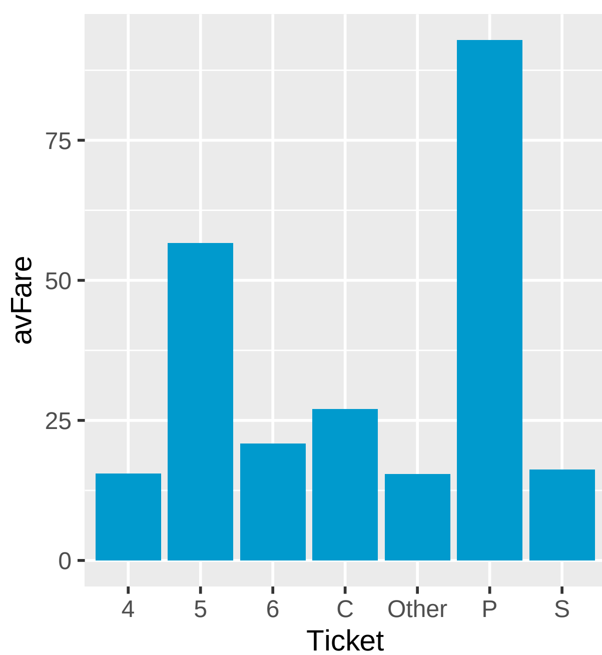


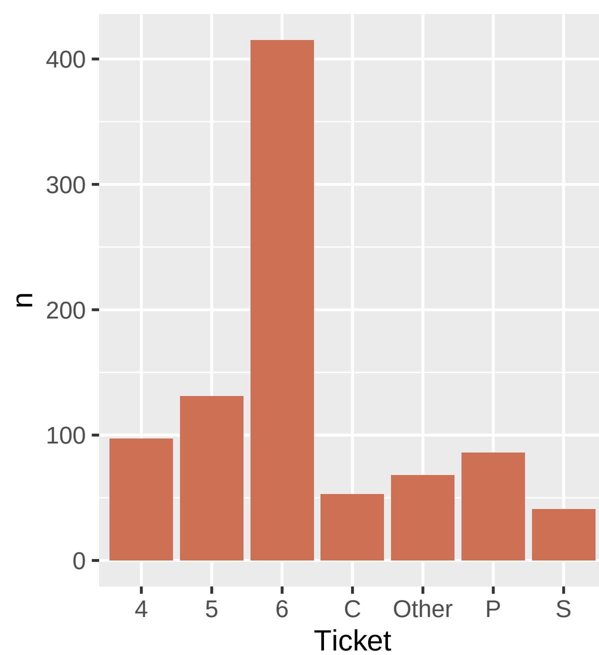
The average age for the title make sense with Master having the lowest average age and Soldier representing seniors that have been in military at some point. It can also be seen that most of the passengers are middle aged men and there are few of nobles and soldiers. The next two graphs below show the 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:



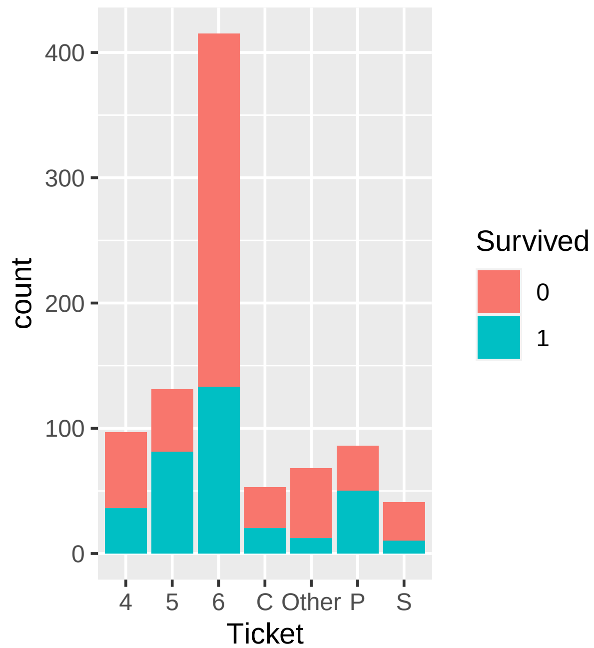


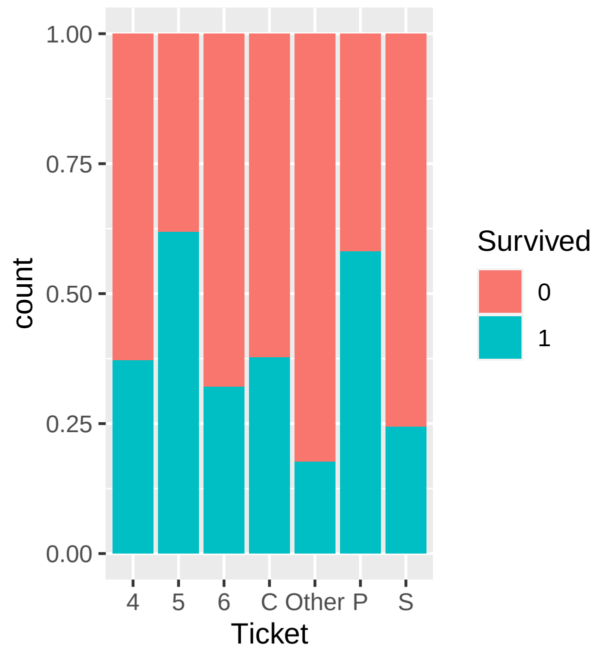
It can be seen that middle aged men have lower survival rates in comparison to the other titles on the ship. Young passengers and female passengers seem to have a higher survival rate. Next graph below shows the average fare per ticket type whereas the one on the right shows the number of observations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



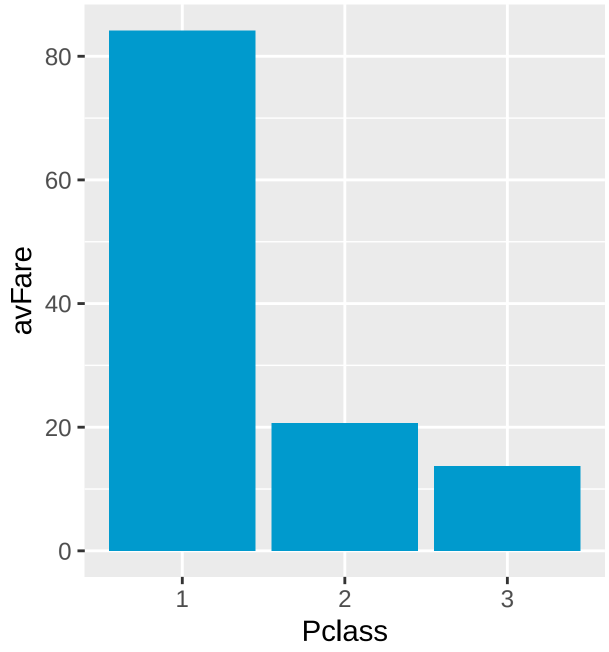


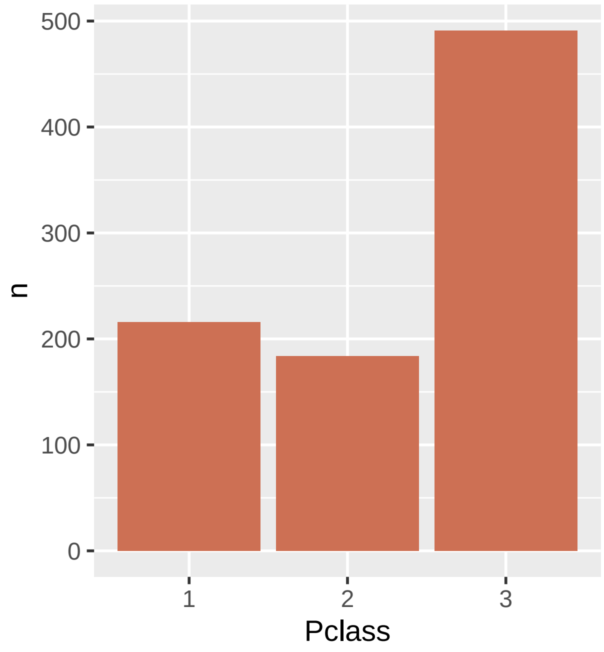
It can be seen that the average ticket prices are similar except for type 5 and type P, both of which are much more expensive compared to other ticket types. It can also be seen that ticket type 6 is the most sold one compared to the other types which are similar in amount. The next two graphs below 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:



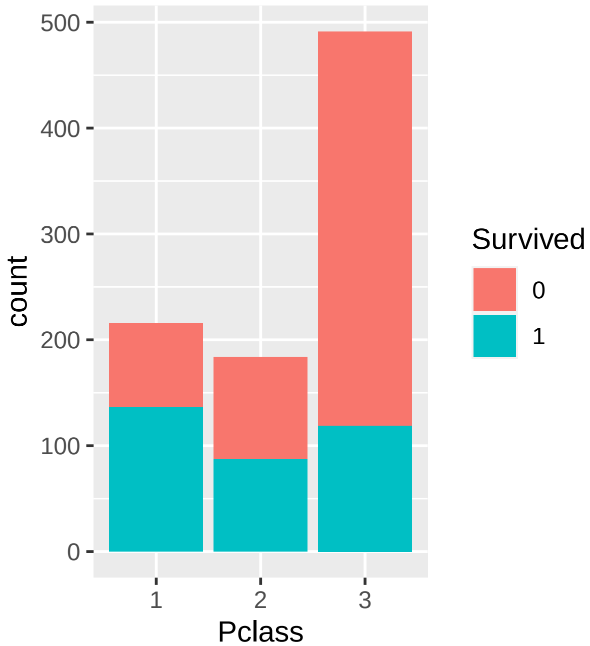


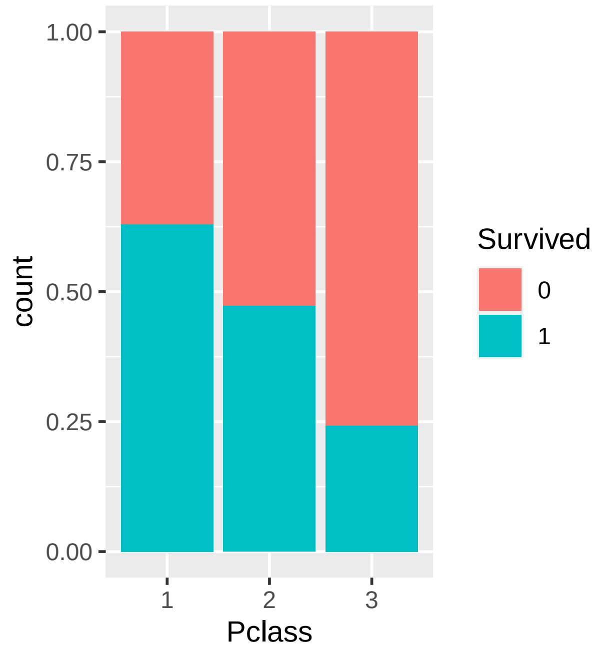
It is a bit hard to make separate conclusions for each ticket type but an obvious one is that the passengers having expensive ticket types have higher survival rates. Next graph below shows the average fare per passenger class whereas the one on the right shows the number of observations for each passenger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



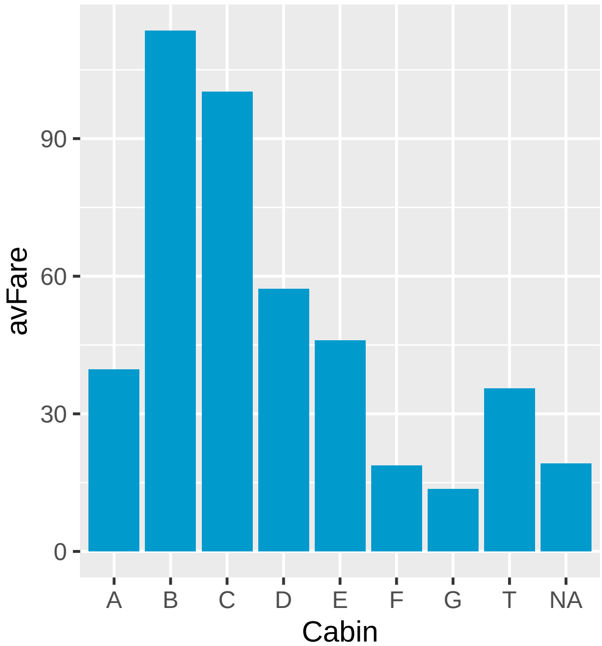


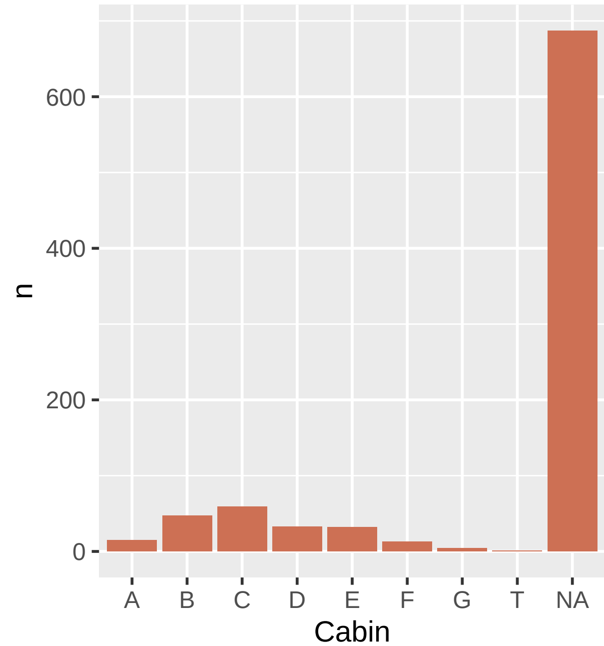
As expected, class 1 passengers pay the highest price compared to other passenger classes. It can also be observed that Class 3 passengers have the highest number. The interesting observation is that class 2 passengers are less in amount in comparison to class 1 passengers. The next two graphs below 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:



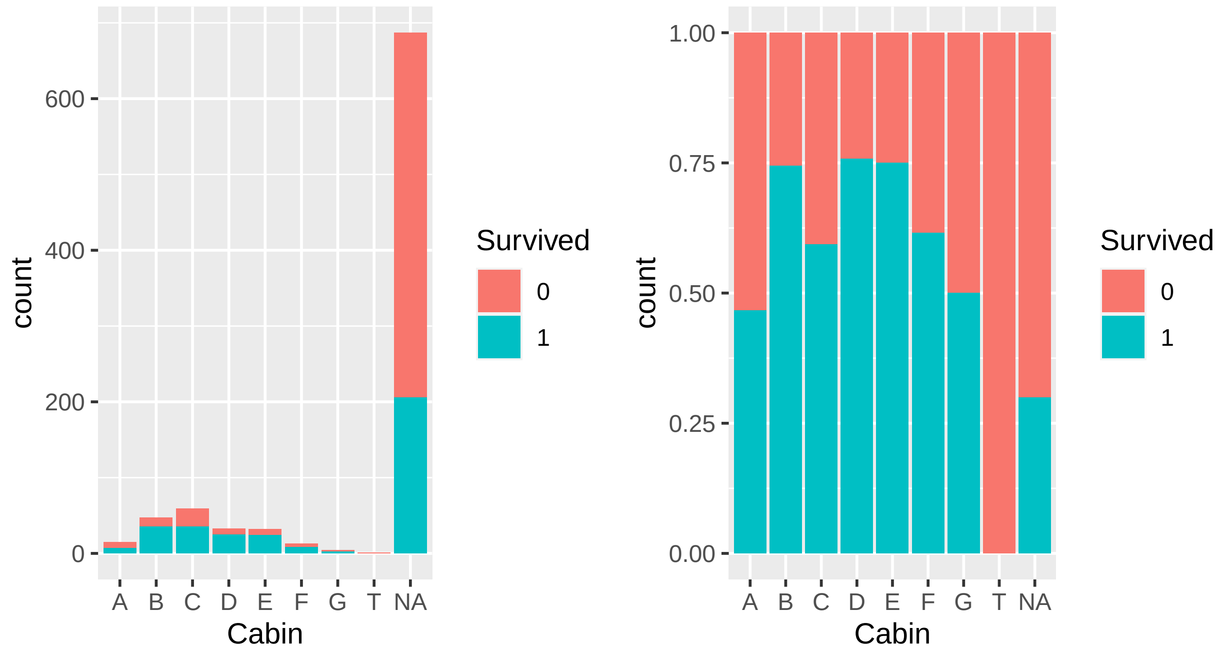


The survival rate is proportional to the passenger class with passengers that pay the highest price survive the most proportionally. Next we will check the cabin predictor in the same manner.

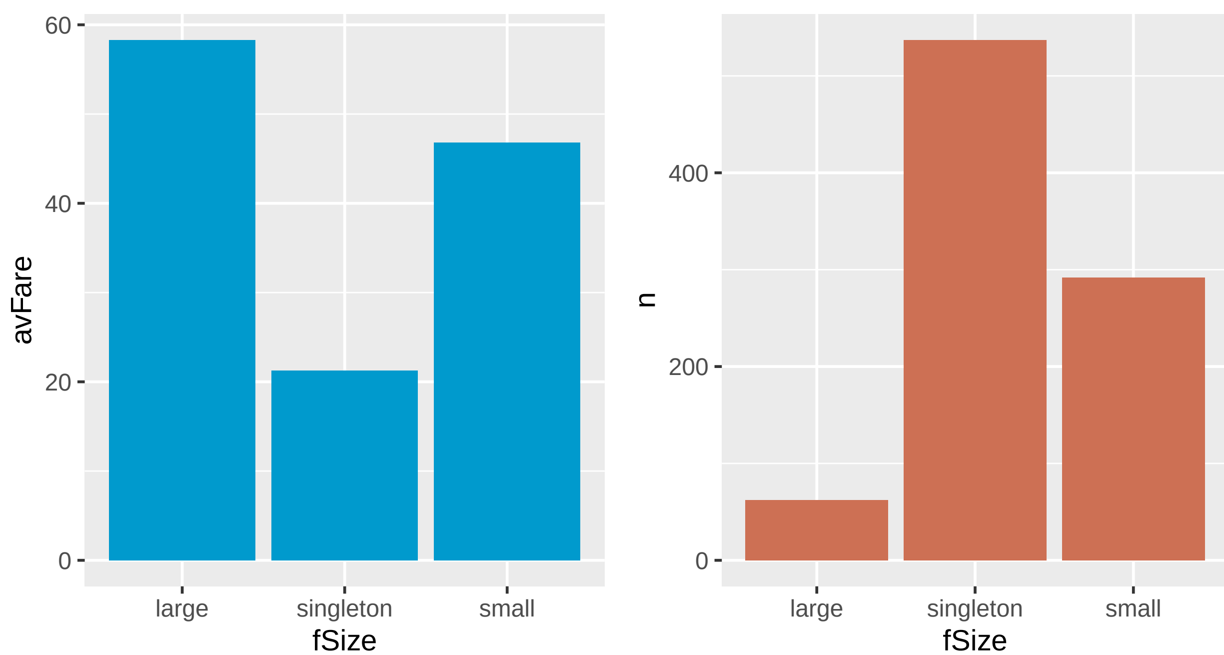




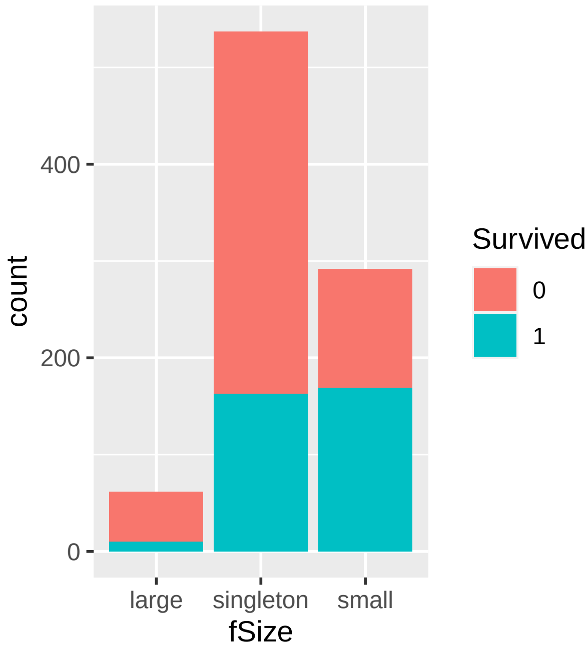
As mentioned before, the number of missing values for the cabin predictor is very high. This means either the cabin predictor will be dropped from analysis, which will result in loss of information, or a logic will be applied for imputation of the missing values. The following graphs show the relationship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

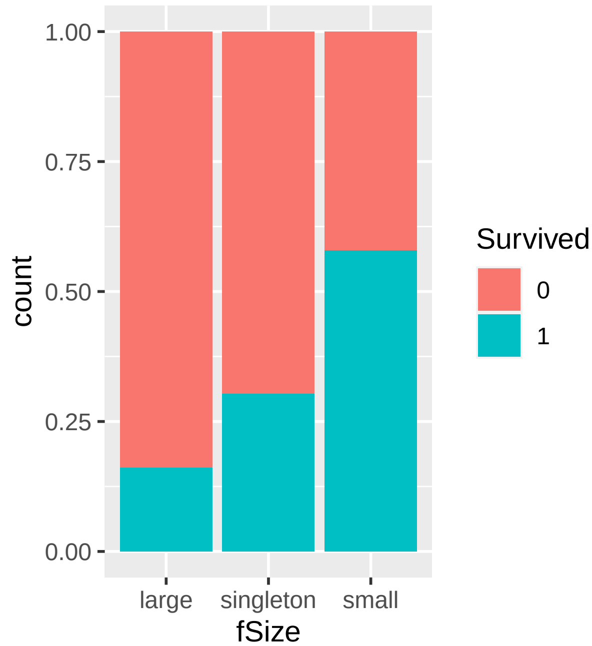


There is not a particular pattern that can be noticed immediately but again cabins with passengers that have paid higher ticket fees tend to have higher survival rate. Next we will check the family size predictor in similar fashion to other predictors above:



The next two graphs show the family size relationship to survival rate:





We will impute the missing age observations by replacing them with median of each title group that the individual belongs to.

Individuals that are upper class passengers, are members of small families(2 to 4 family members), embarked from C and are below 60 years old tend to have higher survival rates.

Cabin variable will be dropped as it doesn't seem possible to impute this predictor with lots of missing observations. However, it might still makes sense to impute age predictor.

The imputed dataset will be used for analysis.

3- Prior Predictive Checking

We will perform a prior predictive check with 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.

We have used weakly informative robust prior of student_t(3,0,2.5) for both population-level and group-level parameters. We also used lkj(2) for the correlation matrix. It can be seen that the spread for prior predictive samples have much higher variance than the actual response variable for the training set. It confirms that the chosen prior can be used for posterior predictive check along with the likelihood.

4 - Model Fitting and Algorithm Diagnostics

5 - Posterior Predictive Checking

6 - Additional Models and Model Improvements

7 - Model Comparison

8 - Prediction Submission