Flight_Capstone

Nitin Yadav

29/02/2020

Introduction

Problem Statement

Airline industries are in continuous tussle to get more and more customers and in turn are working on very thin margins. The price of flight tickets are very unpredictable considering the dynamic nature of business and governing the law of demand and supply. At times we have noted that for a particular city or destination when we search for flight price , the price keeps getting dynamically updated depending on the search criteria, seat availability, date and time of travel etc. Hence it becomes very important for the Airline industry to have a right price prediction mechanism which is backed up by data and helps the industry to take a data driven decision.

Need of Study Project

This is a problem of machine learning where we have been given 2 data sets i.e Train and Test set. Train data consist of 10683 records Test data consist of 2671 records

Understanding business/social opportunity

This is a MAchine learning problem based on supervised learning. Here we train the algorithm using the Train dataset. In supervised machine learnng we know the Target variable and we try to identify the kep predictors on which the response variable (Y) is dependent. Based on the trained machine model, we then try to predict the target using a Test set. This is very crucial in Airline industry as price of a flight is very important parameter for a customer to take the travel decision and a right price point will be beneficial for both customer and the Airline company. Hence the better the MAchine learning model, better would be the accuracy and hence minimum error.

Data Report

Data collection in terms of time, frequency and methdology

If we see the data collected , we notice that data provided comprises of 4 months data starting from MArch till June 2019 and the data is provided for Weekday, weekends and for

24 hour time period across all days. There can be various methodologies to collect data i.e through APIs as the direct hstorical data for airline flights is not available, howwver different travel websites provided data in various fields which has to be cleaned first to get data in desired format

visual inspection of Data (Rows, Columns, Descriptive Stats)

The Train dataset comprises of 10683 rows and 10 columns. Test set consist of 2671 row items and 10 variables. We observe the following in dataset

1. Price is dependent variable, all other variable are independent or predictors

2. Except Price which is numeric, all other variables are in “Char” format which needs to be converted to categorical or right class

3. Date of Journey column needs to be separated into “Date”, “Month” and “Year” columns and convert to Date format

4. Route Info has starting city as “Source” and end city as “Destination”. We need to do feature engineering to create 2 columns i.e for Source and Destination using the separate function and see if this matches with the existing source and destination information provided.

5. Departure time and arrival time have to be converted to time format and the duration has to be put in either “Total hours” or “Total mins”. We have taken “Total Mins”

6. Total stops have to be converted to factor category

7. Jet Airways and Indigo have the maximum number of flights followed by Air India

8. Delhi, Kolkata and Bangalore have the maximum flights starting from them as Source City

9. Cochin, Bangalore and Delhi have the maximum flights reaching there as Destination City
10. There are 3491 non-stop flights and 5625 flights with 1 stop.

11. There is a huge variation in the price , minimum is 1750 and maximum goes upto 79500. There are possibility of outliers in the PRice column.

12. Dates within 1st to 10th of month have highest number of flights and maximum flight are in month of May-June (Possibility of Summer Holidays)

Understanding of Attributes (variable info)

1. Convert Date of Journey in Date, month and Year columns
2. Convert the required variables to Factor or Date formats
3. Separate the Duration column in Hour and minutes to calculate Total Minutes

4. Convert the Departure time in two brackets i.e day time (9am-9pm) and night time (9pm-9am)
5. Get the weekday information and create a separate column for the day of week from date of journey field

Exploratory Data Analysis

Univariate Analysis

1. Price and Total Duration are numeric categories and all other columns are either categorical or date class
2. Boxplot and histogram of Price shows the presence of outliers
3. Skewness is a mesure of symmetry , positive skewness for price (1.85) means the mean is more than median of the entries and hence it is right skewed
4. Kurtosis define the tail shape of data distribution , in this we have excess kurtosis (13.5) which is towards positive hence it indicates Fat tailed distribution or leptokurtic
5. Day of Travel shows that maximum number of flights are on Monday , Wednesday and Thursday 6 Departure Time and arrival time shows that maximum number of flights arrive and depart around 7 pm in evening
6. Minimum of duration (in mins) is 75 mins and maximum is 2860.
7. Total count of flights is highest during Daytime, on wednesday as day of week and flight with 1 stop

Bivariate Analysis

1. Average flight price on Sunday and Friday are highest and on Monday are lowest
2. Price of Daytime flight is more than night time
3. Jet Airways, Air India and Indigo have highest number of flights in May june month which is maximum or peak season from flights perspective due to summer season
4. Delhi, Kolkata and Bangalore are the popular choice as Source for boarding the flights
5. Cochin, Bangalore are the popular choice as Destination
6. Average Flight price per week is high in the months of May and June compared to MArch April
7. Jet Airways command the highest price among the Airline categories as evident from box plot
8. Average flight price is high during the first 15 days of month compared to the month end days unless there is some specific festive occasion
9. Delhi and Kolkata commands the highest median price among the other source cities
10. Delhi and Kolkata has highest number of flights as source city and also the count of 1 stops is high for these cities
11. Flight price and Flight duration in mins have a positive correlation of 0.56, means as duration increases flight price increases.

Unwanted variable Removal, outlier treatment and Missing value Treatment

1. We remove unwanted variable like Route information and Additional info from our dataset as they are not contributing to the model and we have already extracted source and destination information from Route information.

2. For outlier treatment, we notice that outlier present in price , we take maximum value of price as 22500, and drop the data points above that point. By doing this we have eliminated around 322 entires of flight price having value higher than 22500

3. For NA values , we notice that there is 1 NA present in Total Stops column, hence we take complete cases and drop the single entry. After doing this transformation the final row count is 10361 and 16 rows

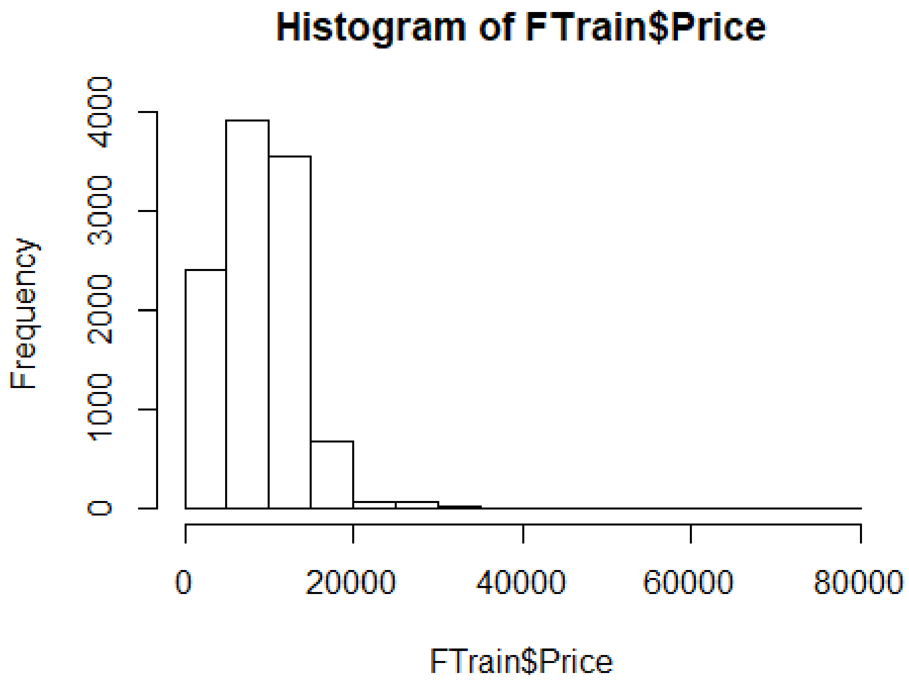
Addition of New Variables

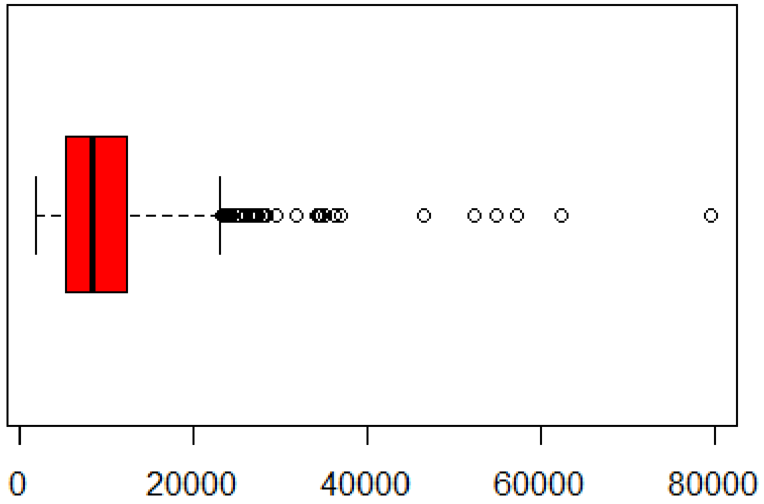
This step we have already covered as part of earlier description provided.

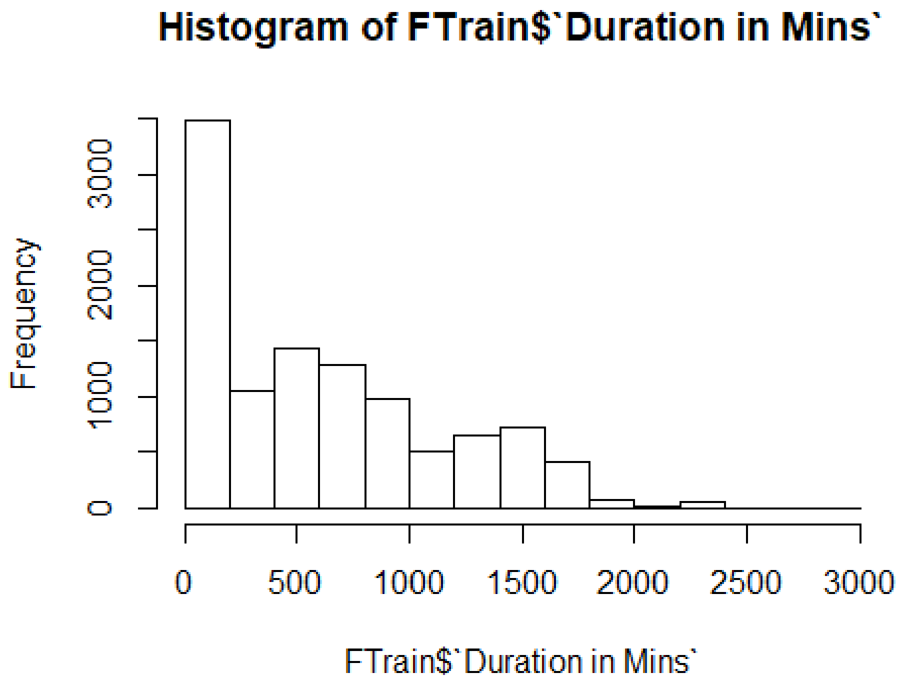
Insights from EDA

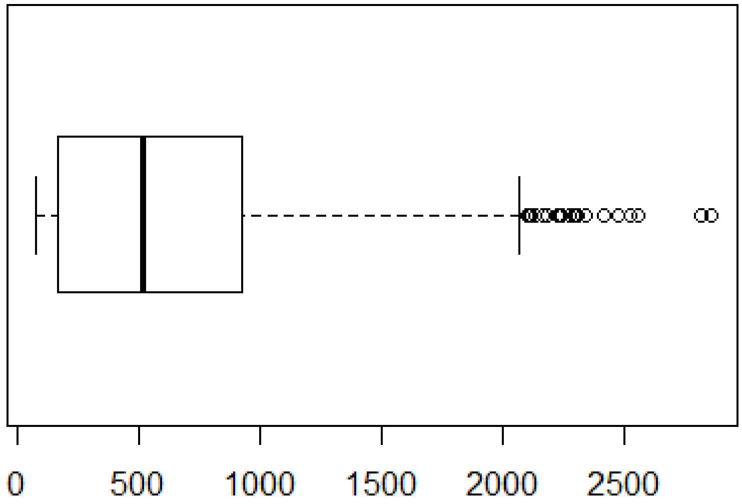
We have already covered insights from EDA. For data imbalance, it make more sense when the classification is binary (0 or 1) but in our case the response variable (Price) is numeric so data imbalance would not play much role here. also the imbalance due to oultier entries is around 3% and very minimal.

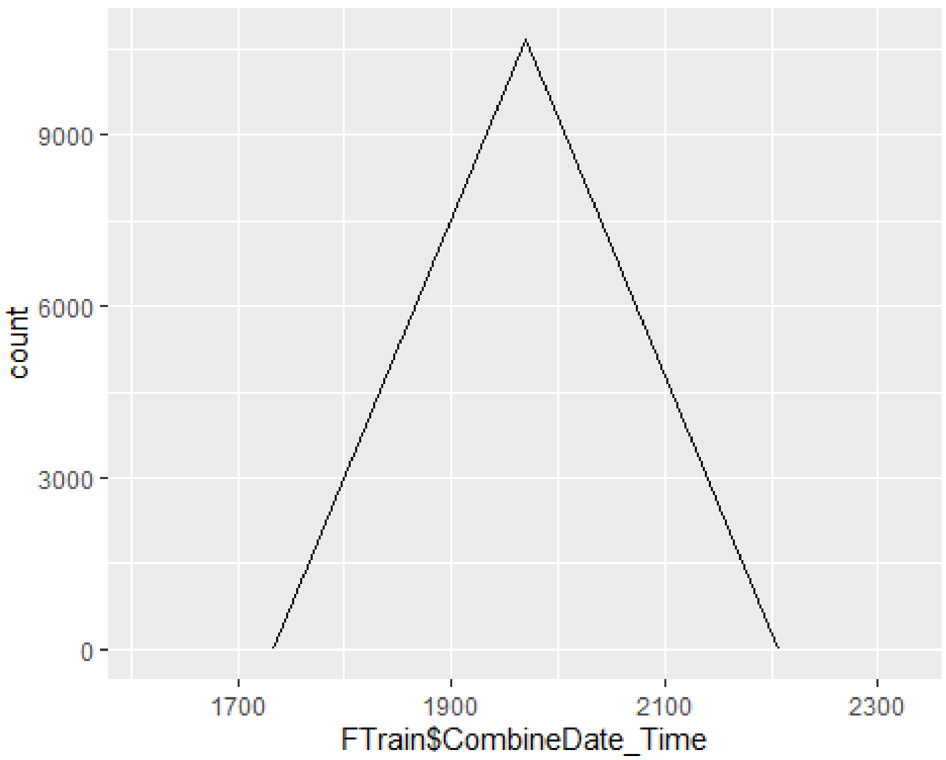
Also techniques like clustering and PCA would have played role where we didnt have target column and we are trying to predict the target, but in our case we have been given the price information and we need to use the same to predict the test data once the model gets sufficient learning and tuning from train data.

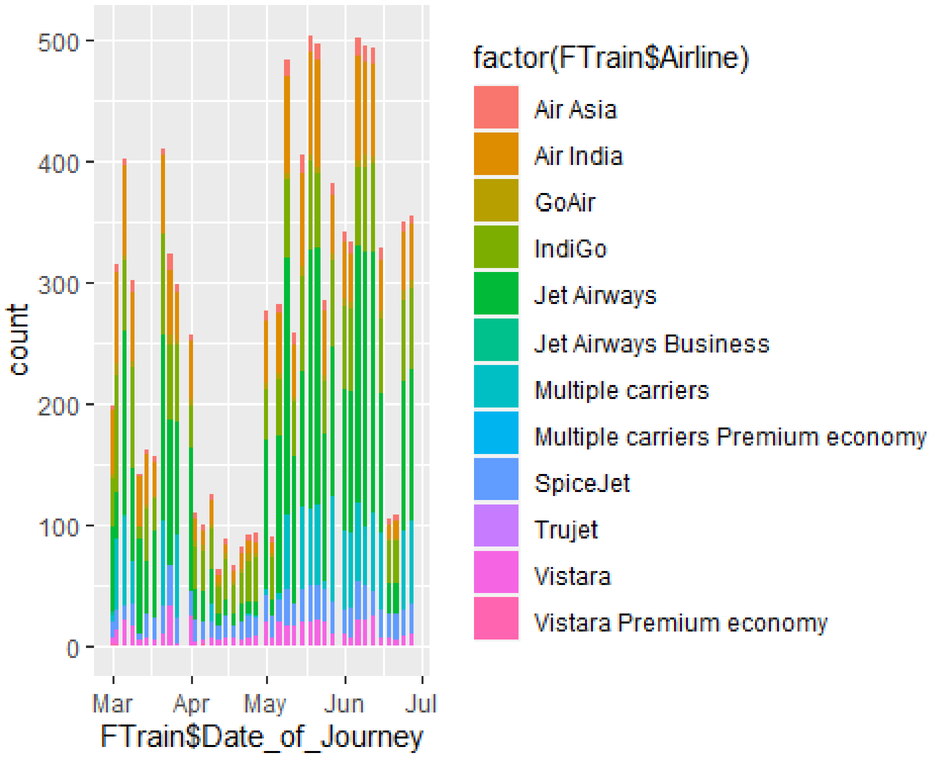

Histogram of FTrain$Price
Frequency
0
1000
2000
3000
4000
0
20000
40000
60000
80000
FTrain$Price

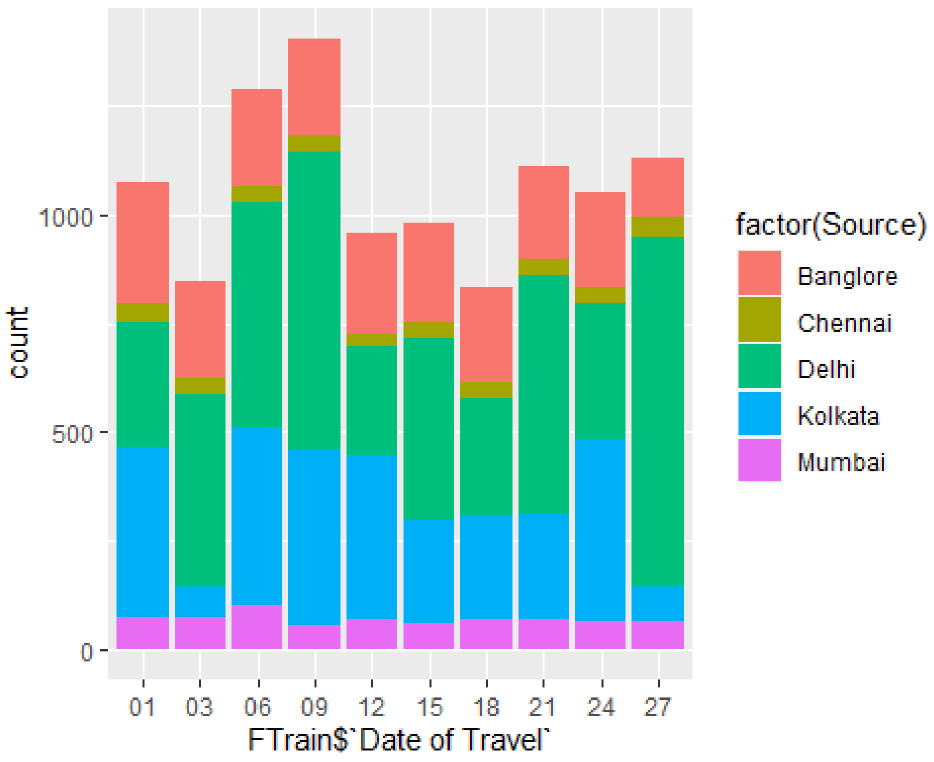

0
20000
40000
60000
80000

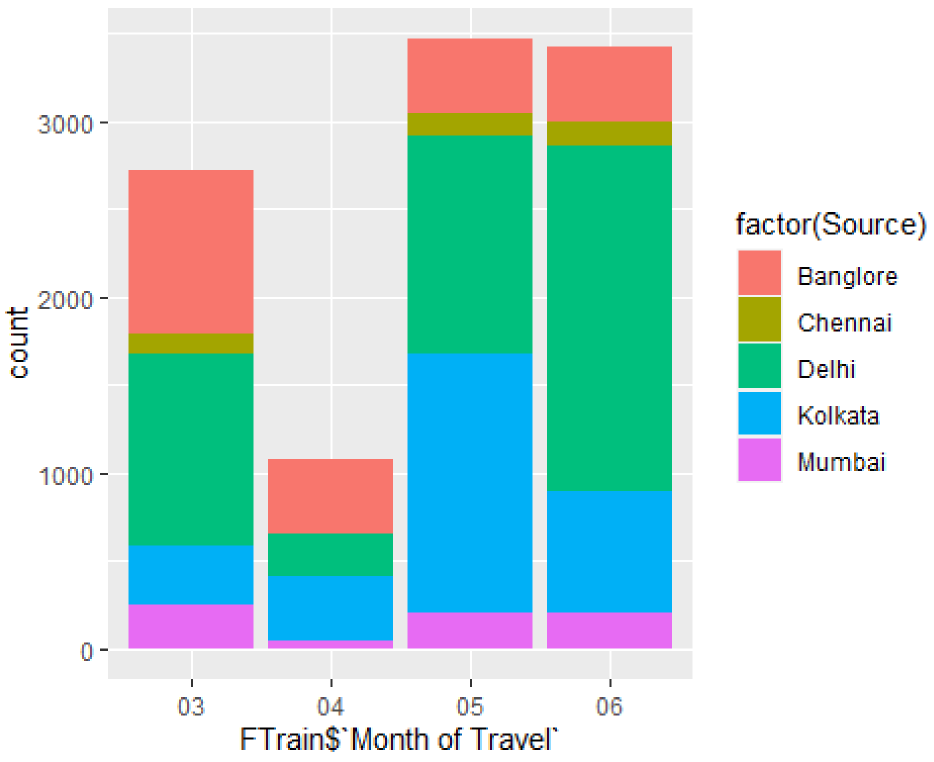

Histogram of FTrain$`Duration in Mins`
Frequency
0
1000
2000
3000
0
500
1000
1500
2000
2500
3000
FTrain$`Duration in Mins`

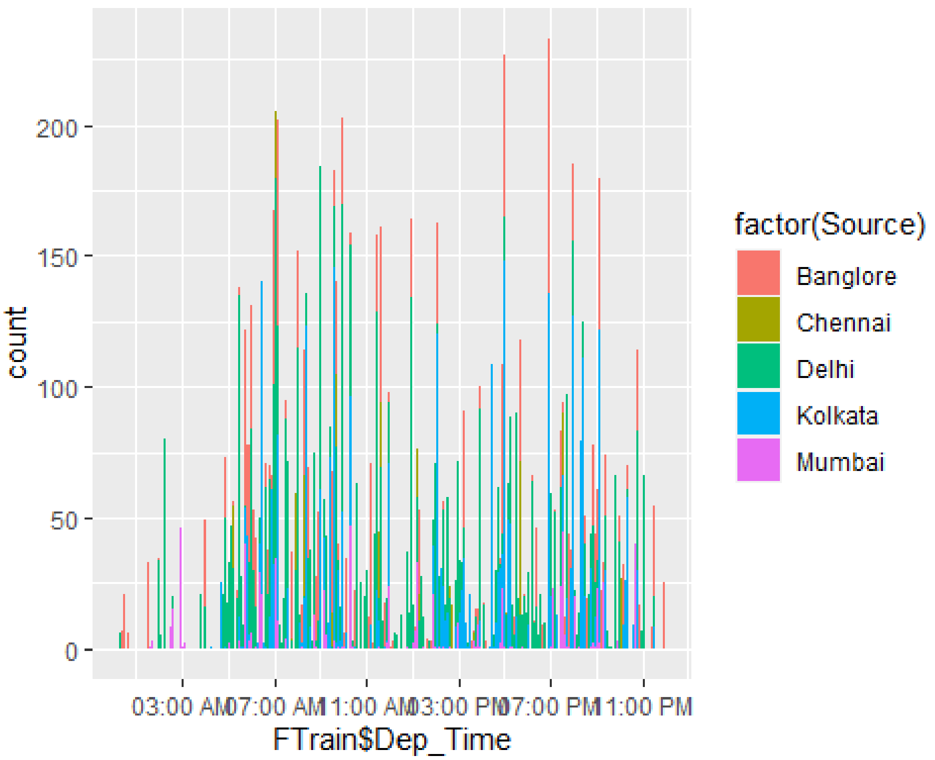

0
500
1000
1500
2000
2500

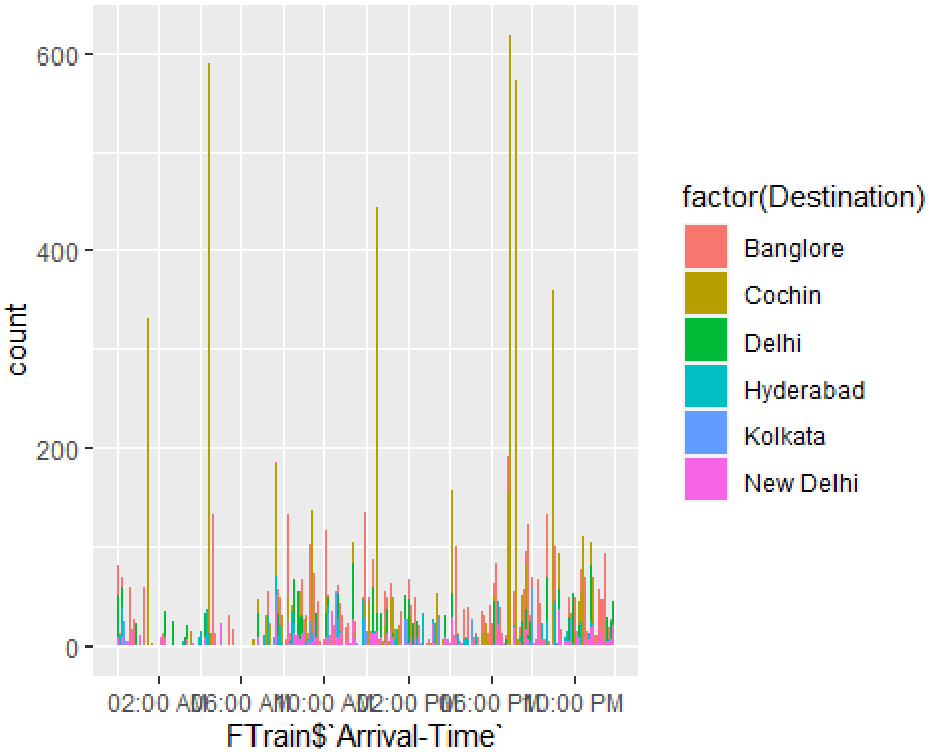

count
9000
6000
3000
0
1700
1900
2100
2300
FTrain$CombineDate_Time

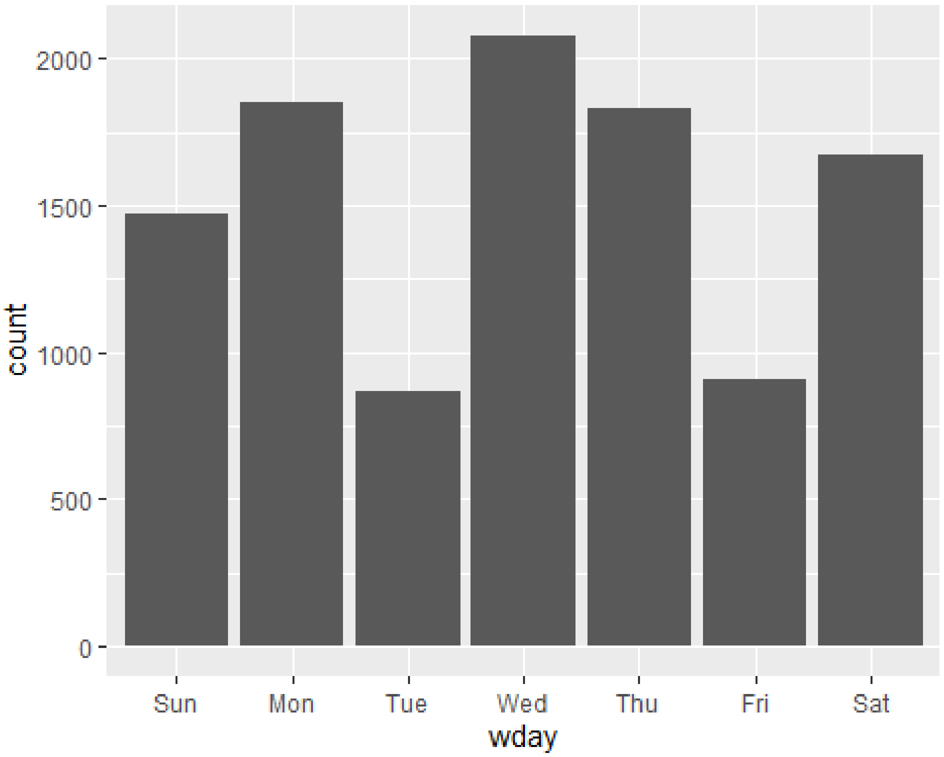

count
500
400
300
200
100
0
Mar
Apr
May
Jun
Jul
FTrain$Date_of_Journey
factor(FTrain$Airline)
Air Asia
Air India
GoAir
IndiGo
Jet Airways
Jet Airways Business
Multiple carriers
Multiple carriers Premium economy
SpiceJet
Trujet
Vistara
Vistara Premium economy

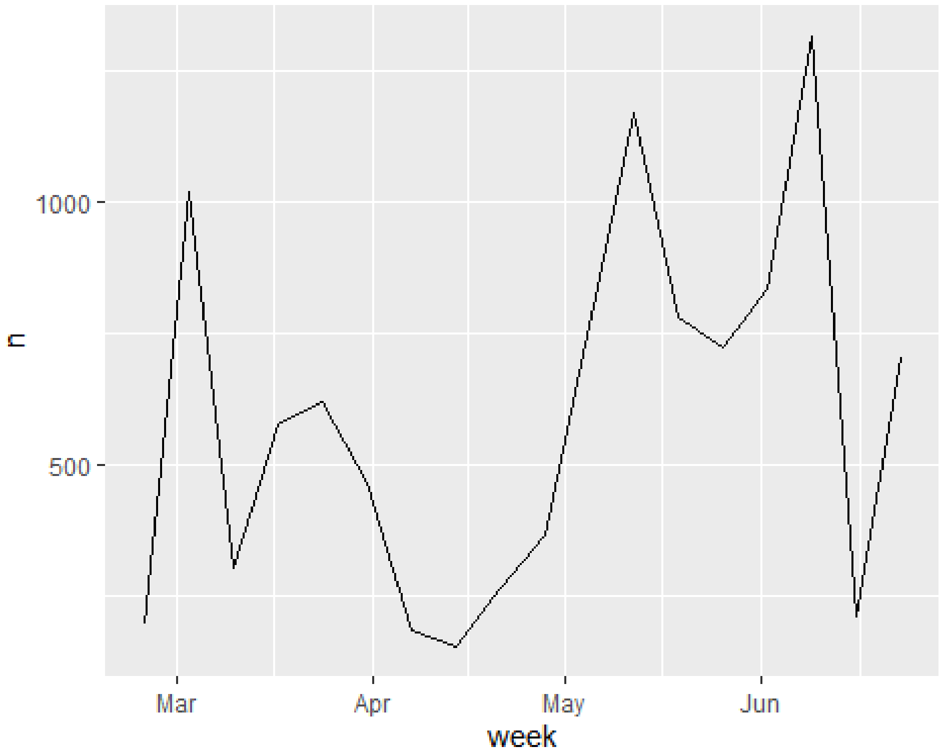

count
0
500
1000
01
03
06
09
12
15
18
21
24
27
FTrain$`Date of Travel`
factor(Source)
Banglore
Chennai
Delhi
Kolkata
Mumbai

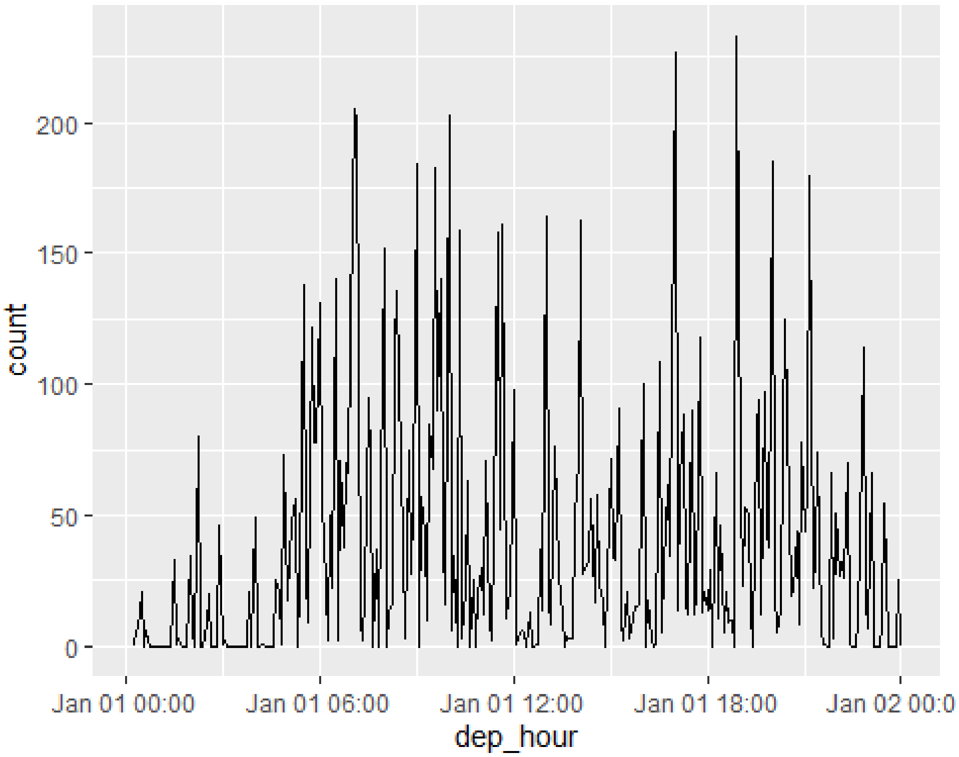

count
0
1000
2000
3000
03
04
05
06
FTrain$`Month of Travel`
factor(Source)
Banglore
Chennai
Delhi
Kolkata
Mumbai

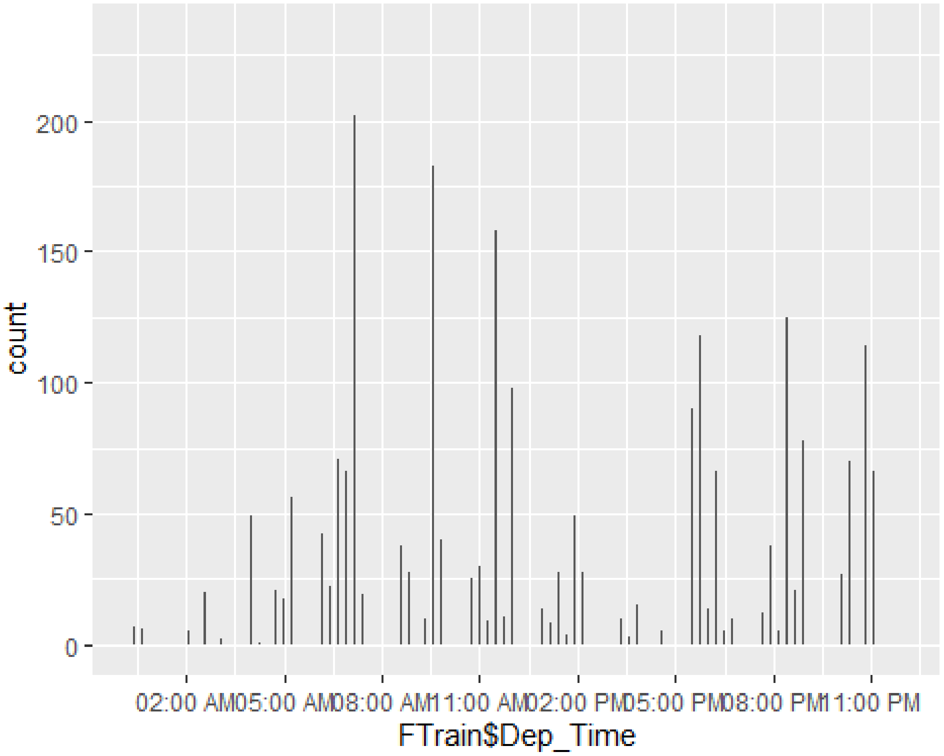

count
200
150
100
50
0
03:00 AM
07:00 AM
11:00 AM
03:00 PM
07:00 PM
11:00 PM
FTrain$Dep_Time
factor(Source)
Banglore
Chennai
Delhi
Kolkata
Mumbai

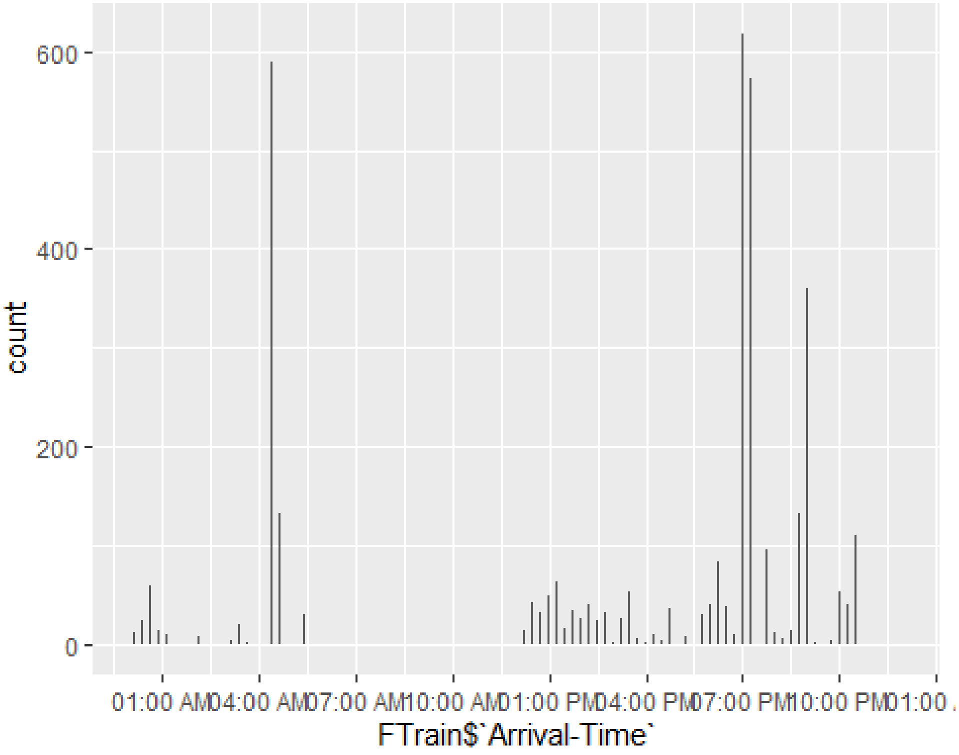

count
600
400
200
0
02:00 AM
06:00 AM
10:00 AM
02:00 PM
06:00 PM
10:00 PM
FTrain$`Arrival-Time`
factor(Destination)
Banglore
Cochin
Delhi
Hyderabad
Kolkata
New Delhi

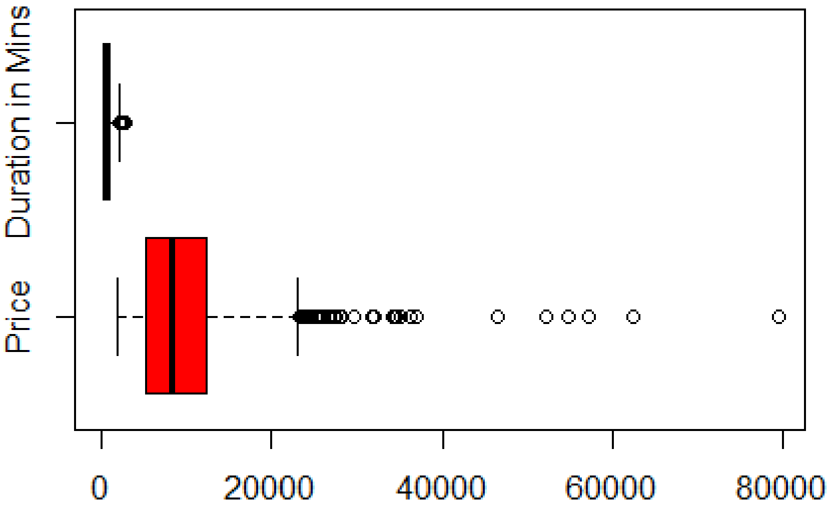

2000
1500
1000
500
0
count
Sun
Mon
Tue
Wed
Thu
Fri
Sat
wday


1000
500
n
Mar
Apr
May
Jun
week

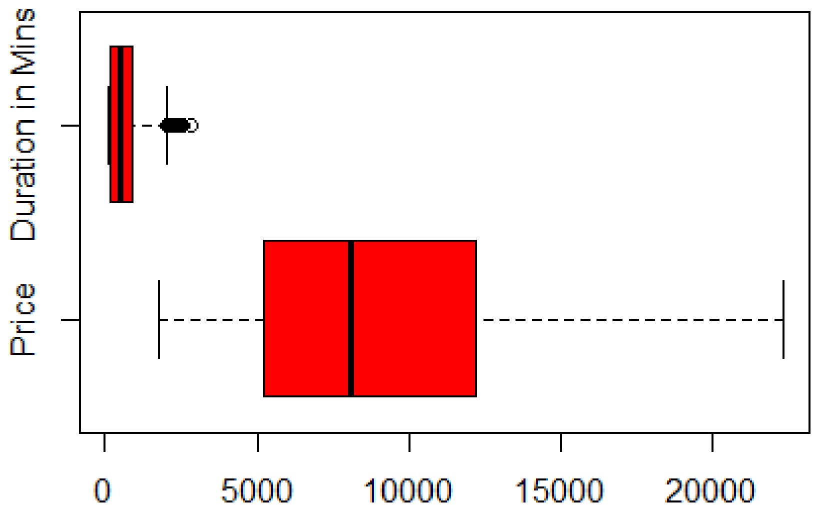

count
0
50
100
150
200
Jan 01 00:00
Jan 01 06:00
Jan 01 12:00
Jan 01 18:00
Jan 02 00:0
dep_hour

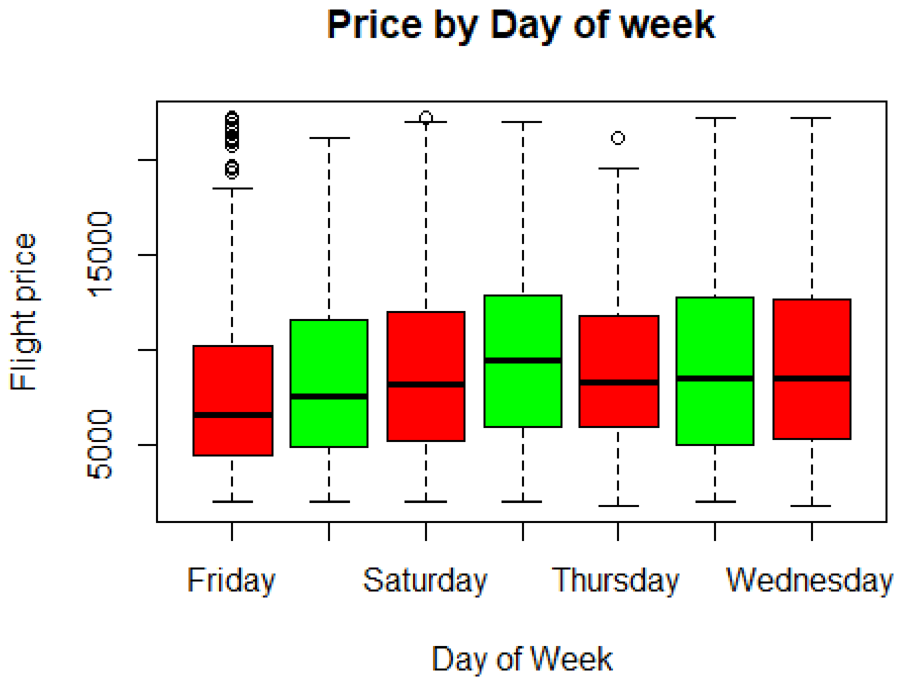

count
0
50
100
150
200
02:00 AM
05:00 AM
08:00 AM
11:00 AM
02:00 PM
05:00 PM
08:00 PM
11:00 PM
FTrain$Dep_Time

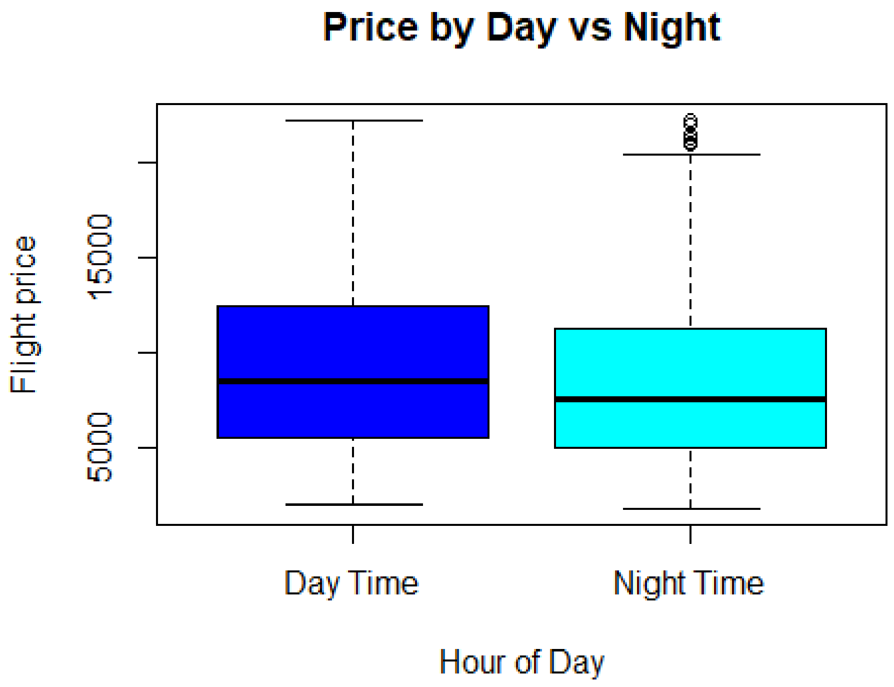


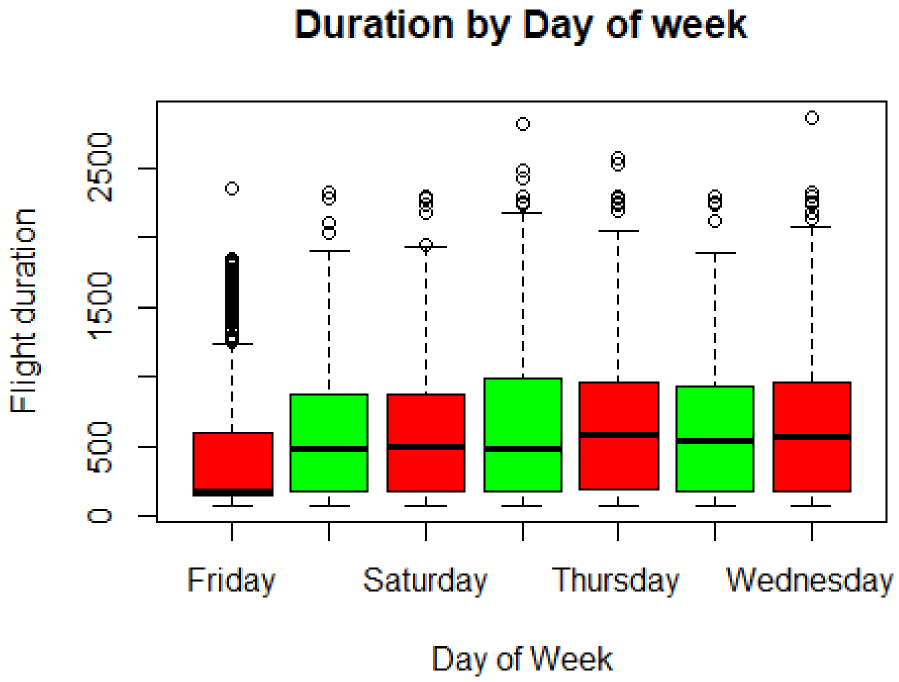


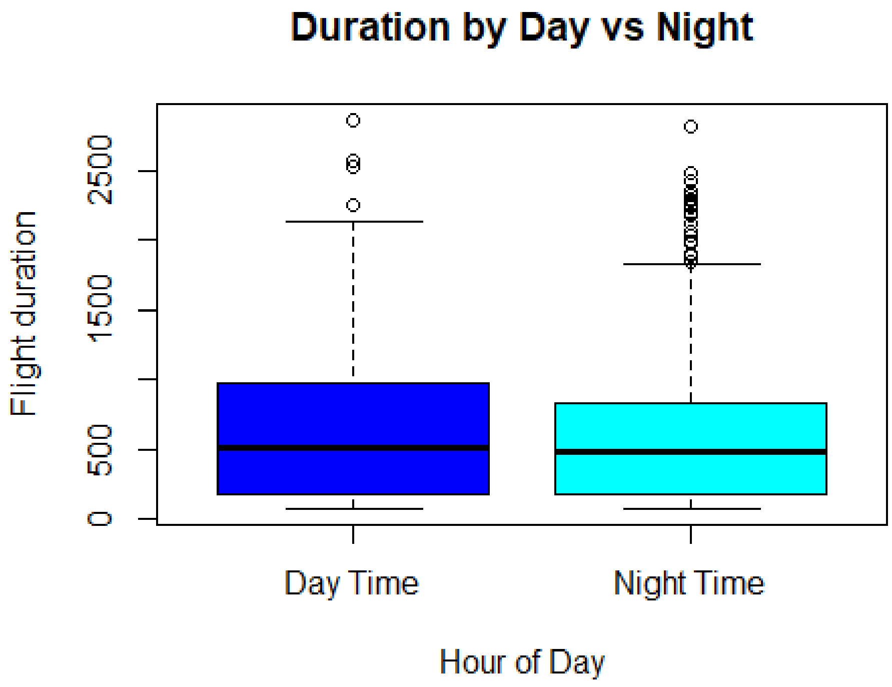
```
# Through boxplot we find that approx 3% of entries in the total dataset are
outliers, so dropping those
```

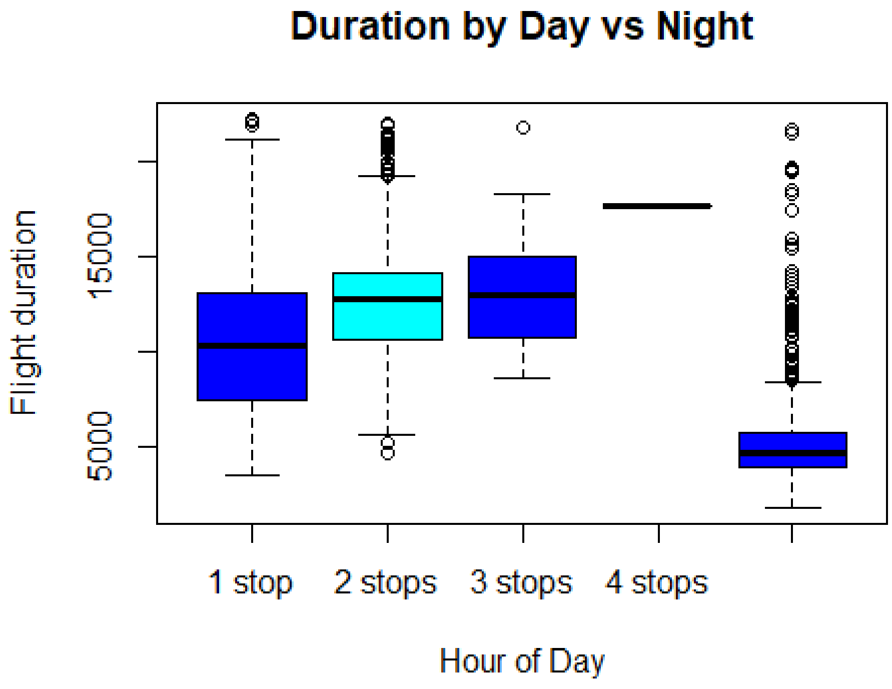

Duration in Mins
Price
0
5000
10000
15000
20000


Price by Day of week
Flight price
5000
15000
Friday
Saturday
Thursday
Wednesday
Day of Week


Price by Day vs Night
Flight price
5000
15000
Day Time
Night Time
Hour of Day


Duration by Day of week
Flight duration
0
500
1500
2500
Friday
Saturday
Thursday
Wednesday
Day of Week


Duration by Day vs Night
Flight duration
0
500
1500
2500
Day Time
Night Time
Hour of Day


Duration by Day vs Night
Flight duration
5000
15000
1 stop
2 stops
3 stops
4 stops
Hour of Day

Independent variables that are singificant

Based on the data transformation and feature engineering we have done above , we can say that except the columns “Route” and “Additional Info”, all other columns are significant in the model building. The same will get validated once we start building the model using Multiple linear regression, Decision Tree, Random Forest, Gradient Boost etc.

Relationship between time of journey and Flight prices

Response to this section we have covered earlier. Flight prices are costlier during the day time and specific during the evening time. Also flight price on weekend are costlier compared to weekdays. Flight price in the morning hours 8-9 am and in evening 4-6 pm are higher compared to other time.

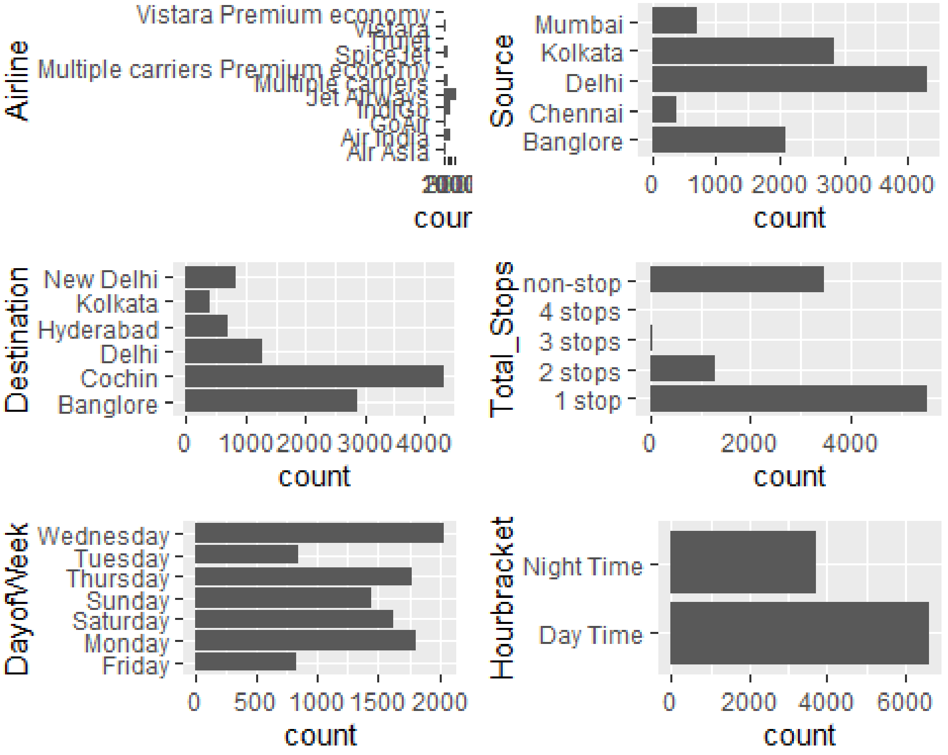
Hypothesis Testing

Flight Prices on Weekdays are cheaper than flight prices on weekends.

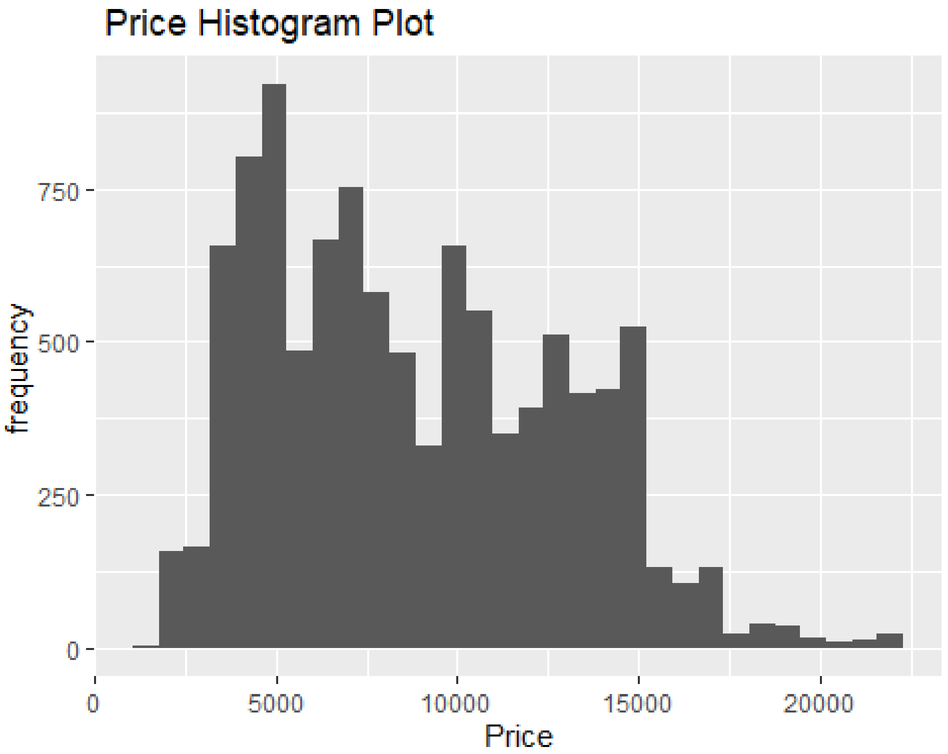
We did anova testing on the linear model built using Price and “DayofWeek” and found the P value very small and hence null hypothesis is rejected and we can say that flight price on weekends are costlier compared to weekdays

Flight Prices during peak hours (9 AM till 9 PM) are costlier than flights at other times.

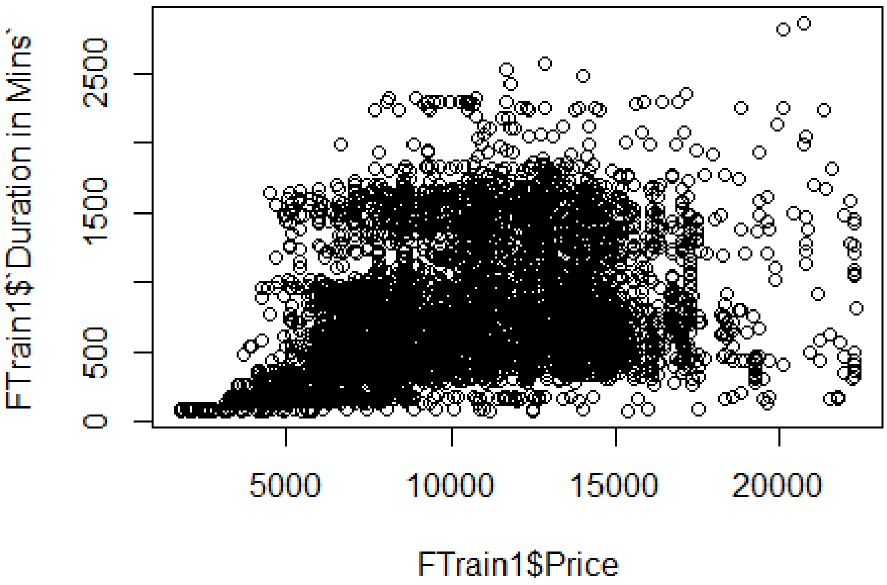
We did a 2-tail t test for the same and found that P value is very small and less than 0.05 , hence null hypothesis is rejected and hence Flight price during peak hours 9am-9pm are higher than non peak hours ie 9pm-9am.

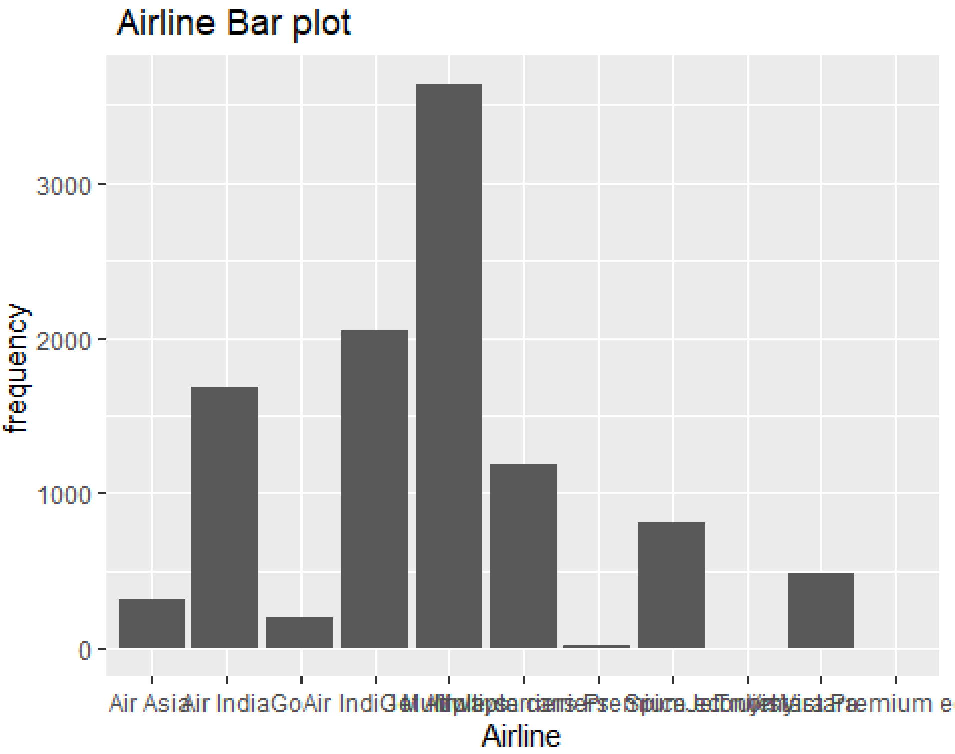


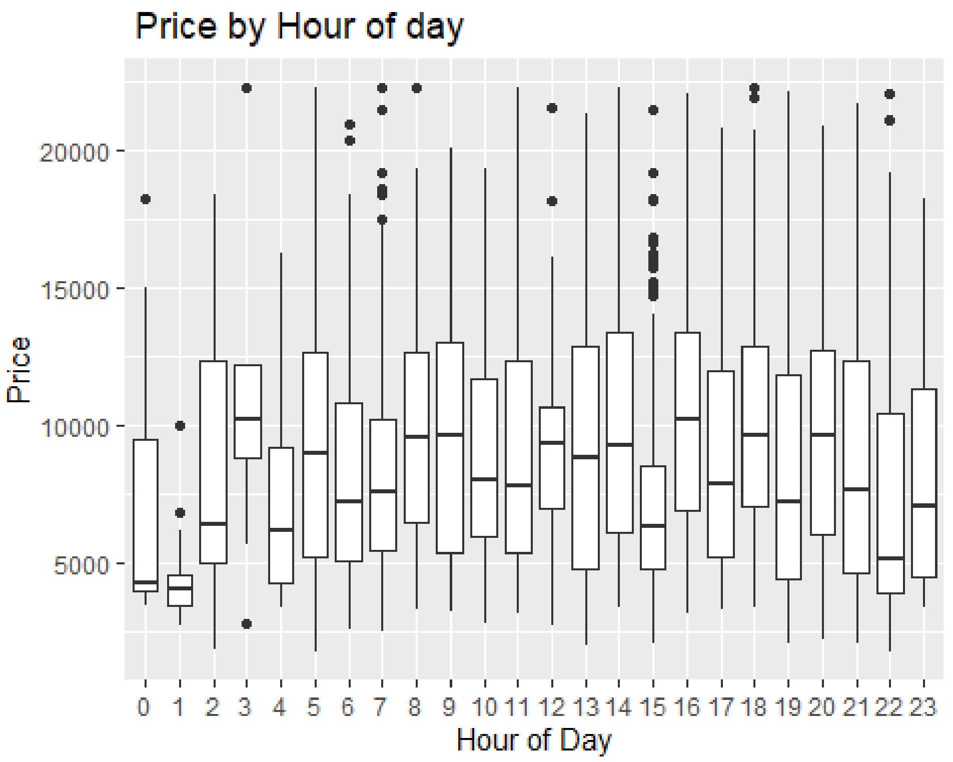
Airline
Vistara Premium economy
Vistara
Trujet
SpiceJet
Multiple carriers Premium economy
Multiple carriers
Jet Airways
IndiGo
GoAir
Air India
Air Asia
cour
Source
Mumbai
Kolkata
Delhi
Chennai
Banglore
0
1000
2000
3000
4000
count
Destination
New Delhi
Kolkata
Hyderabad
Delhi
Cochin
Banglore
0
1000
2000
3000
4000
count
Total_Stops
non-stop
4 stops
3 stops
2 stops
1 stop
0
2000
4000
count
DayofWeek
Wednesday
Tuesday
Thursday
Sunday
Saturday
Monday
Friday
0
500
1000
1500
2000
count
Hourbracket
Night Time
Day Time
0
2000
4000
6000
count

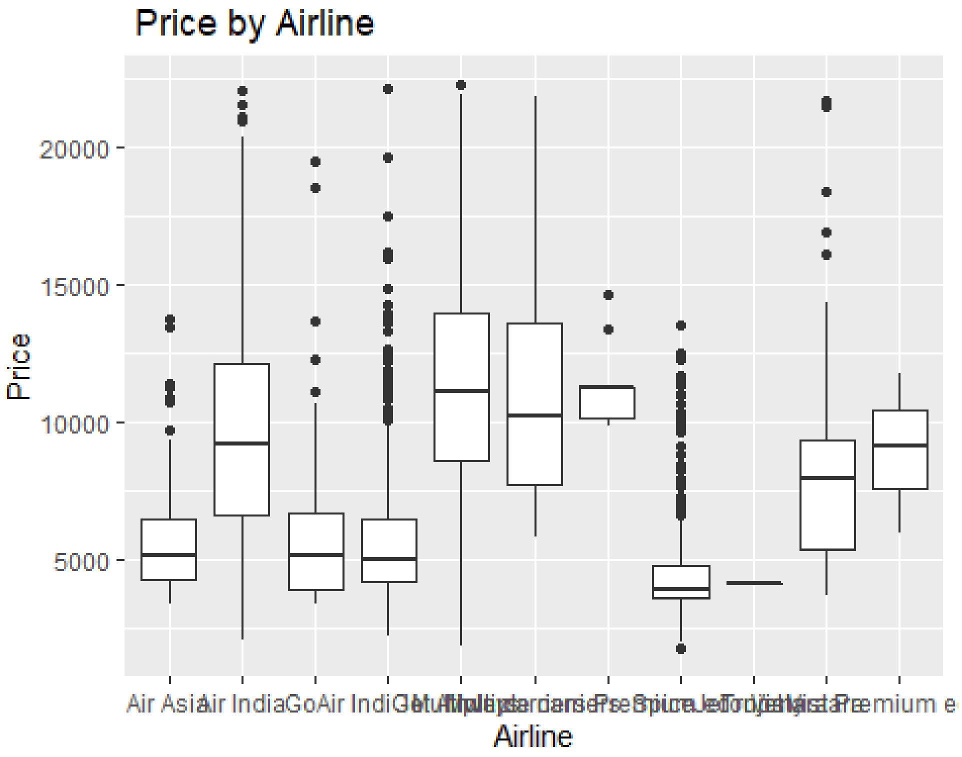


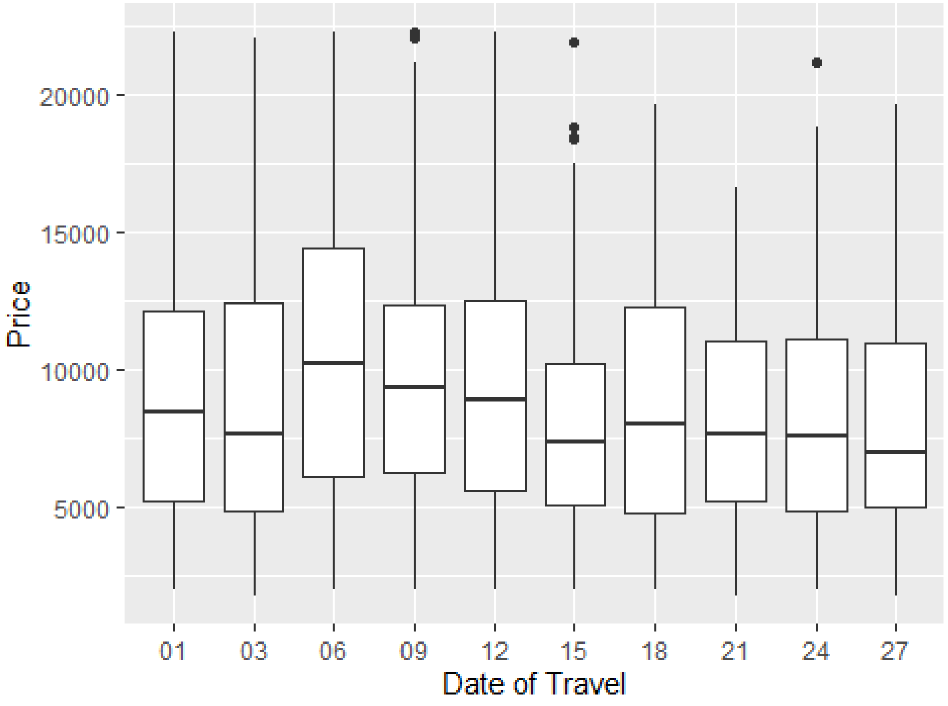
Price Histogram Plot
frequency
750
500
250
0
0
5000
10000
15000
20000
Price


FTrain1$`Duration in Mins`
0
500
1500
2500
5000
10000
15000
20000
FTrain1$Price

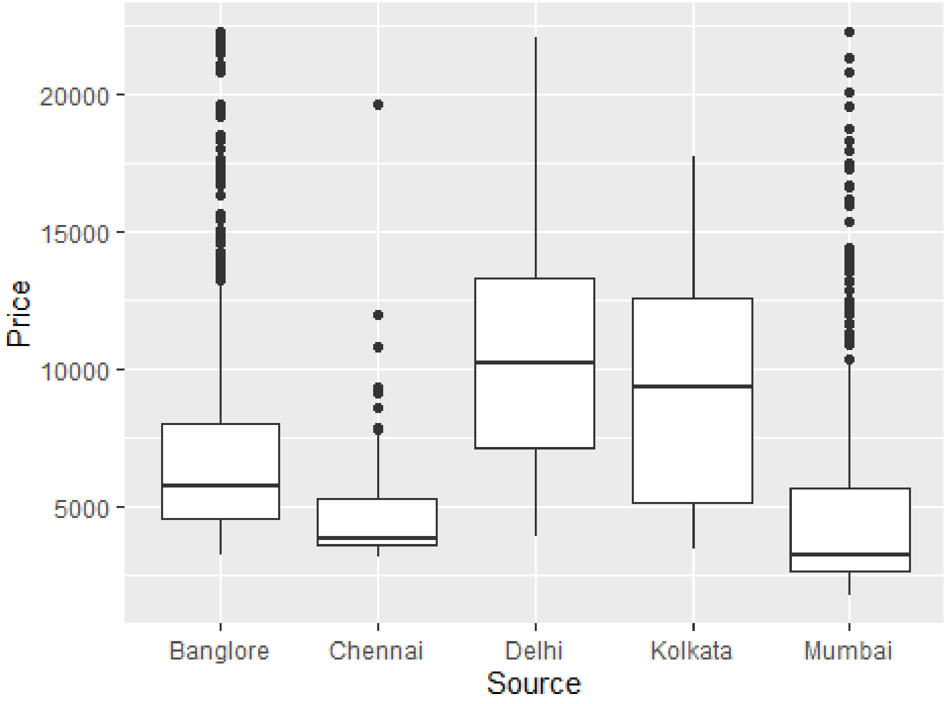

Airline Bar plot
frequency
0
1000
2000
3000
Air Asia
Air India
GoAir
Premium e
Airline


Price by Hour of day
Price
20000
15000
10000
5000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Hour of Day

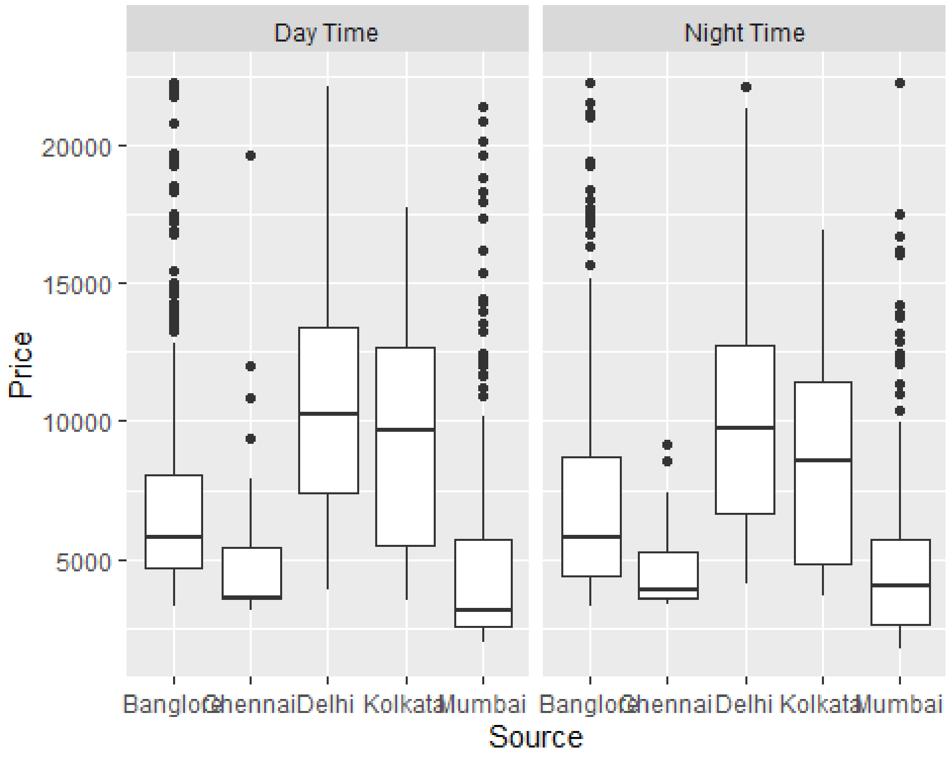

Price by Airline
Price
20000
15000
10000
5000
Air Asia
Air India
GoAir
IndiGo
Premium e
Airline

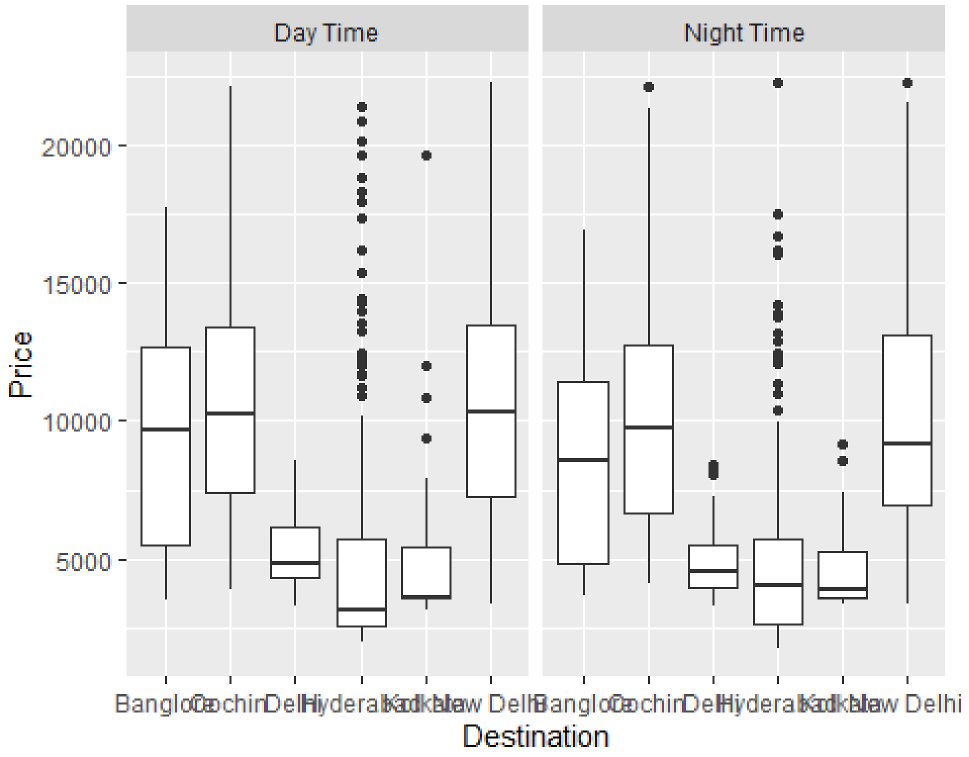


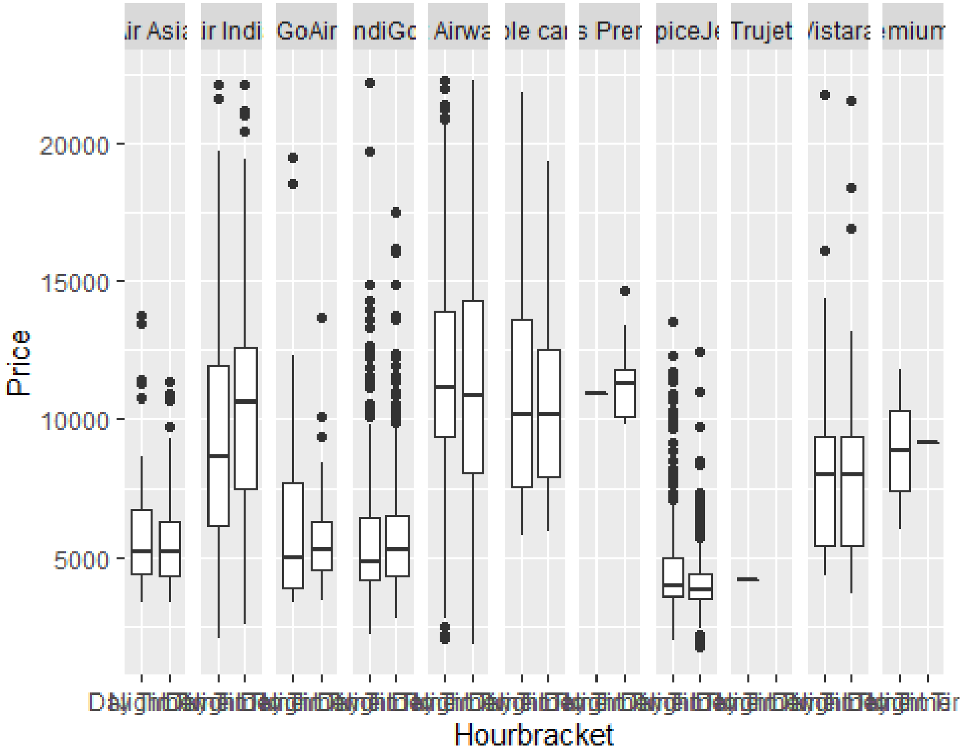
Price by Date of Travel

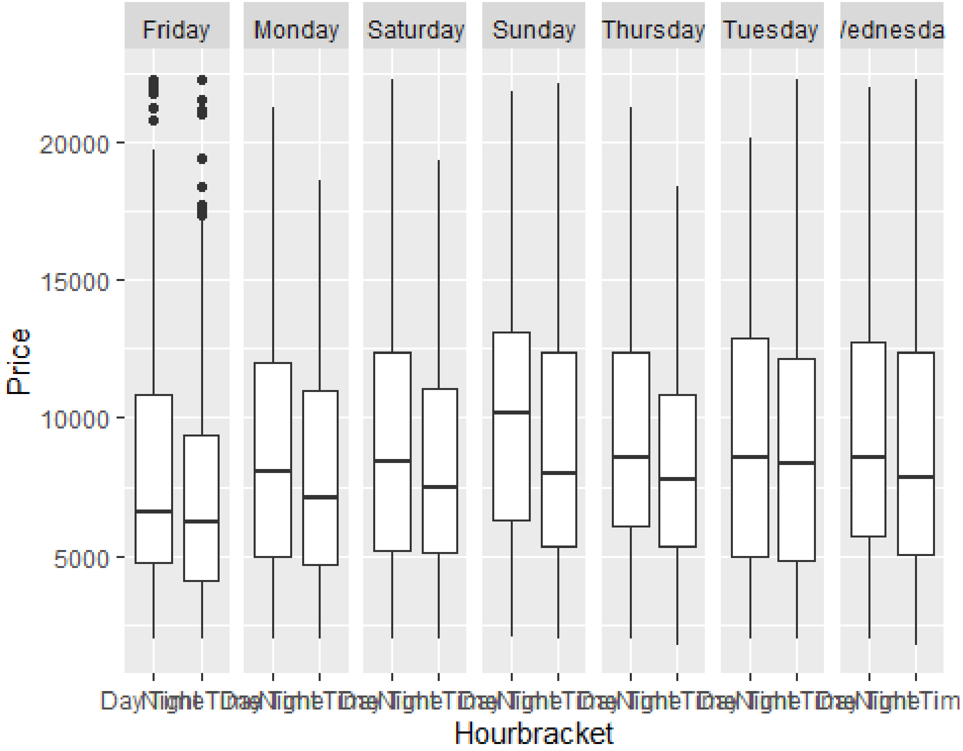


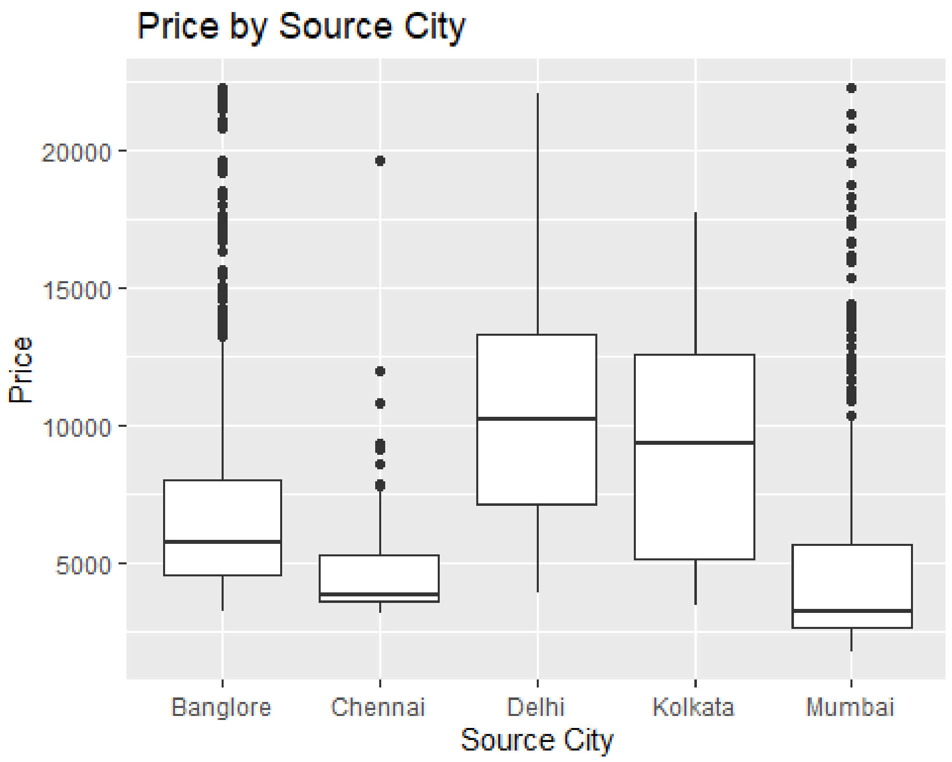
Price by source

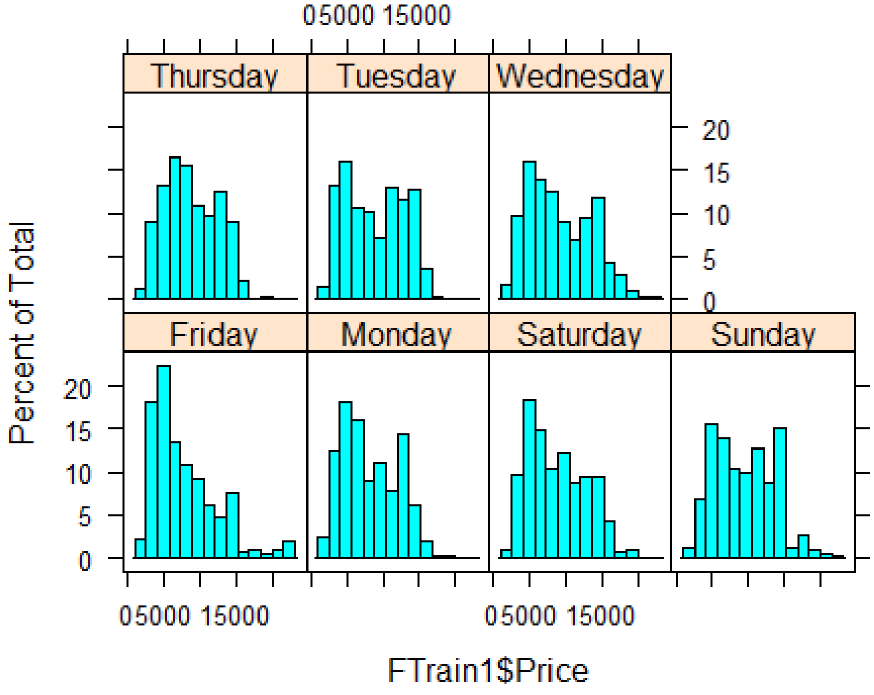

Day Time
Night Time
Price
20000
15000
10000
5000
Banglore
Chennai
Delhi
Kolkata
Mumbai
Banglore
Chennai
Delhi
Kolkata
Mumbai
Source

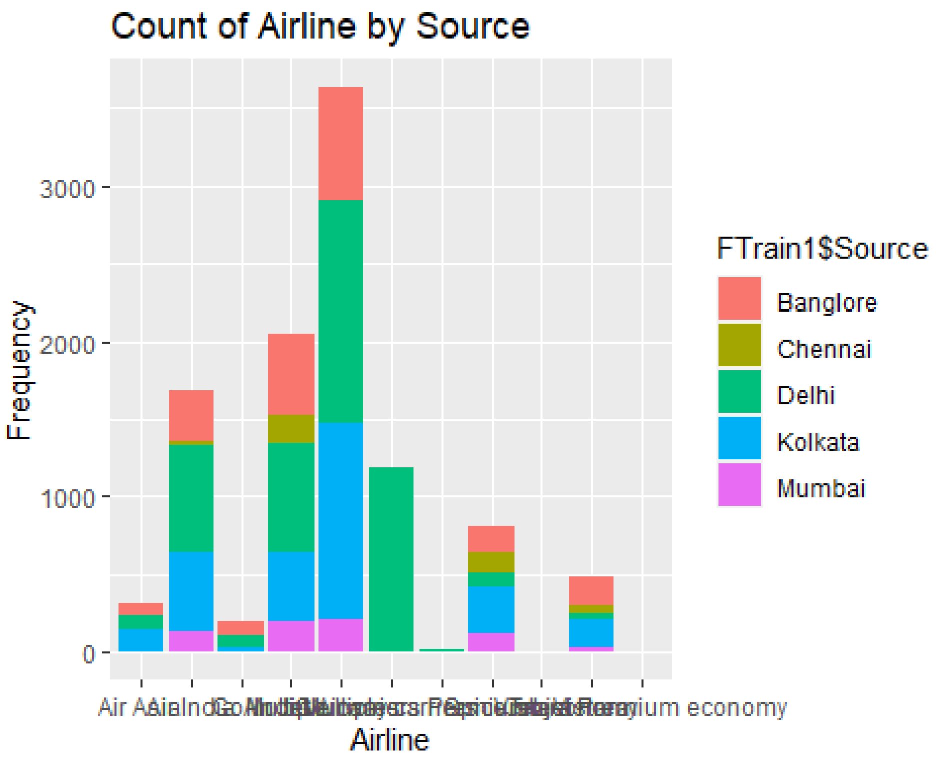

Day Time
Night Time
Price
20000
15000
10000
5000
Banglore
Cochin
Delhi
Hyderabad
Kolkata
New Delhi
Banglore
Cochin
Delhi
Hyderabad
Kolkata
New Delhi
Destination

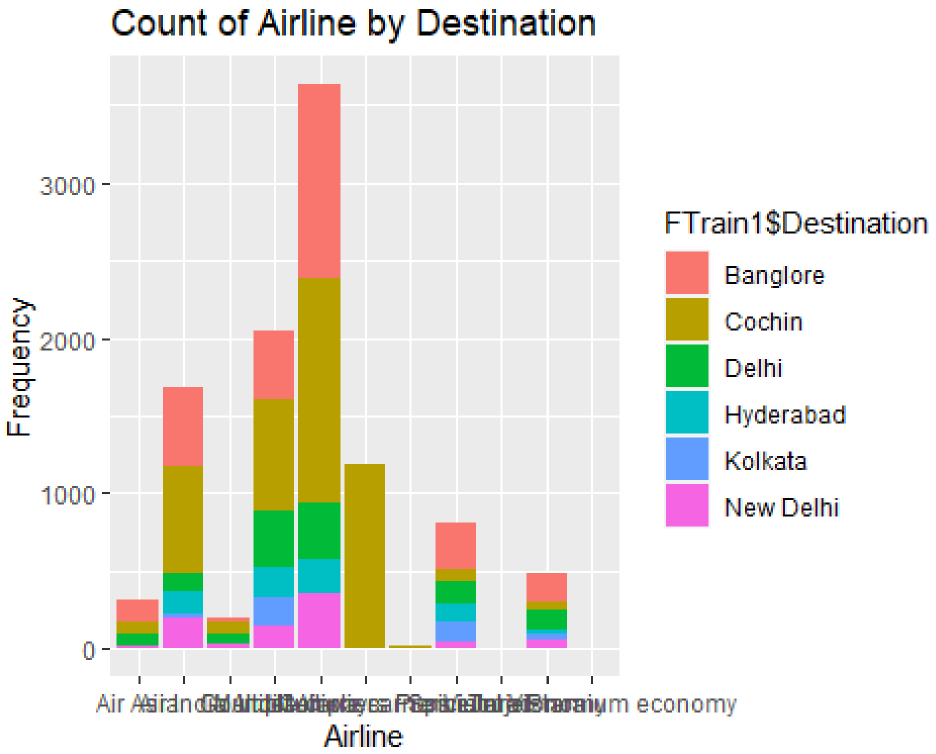

ir Asia
ir Indi
GoAir
ndiGo
Airwa
ble car
s Prer
piceJe
Trujet
Vistara
emium
20000
15000
10000
5000
Price
Day
NightTime
Hourbracket

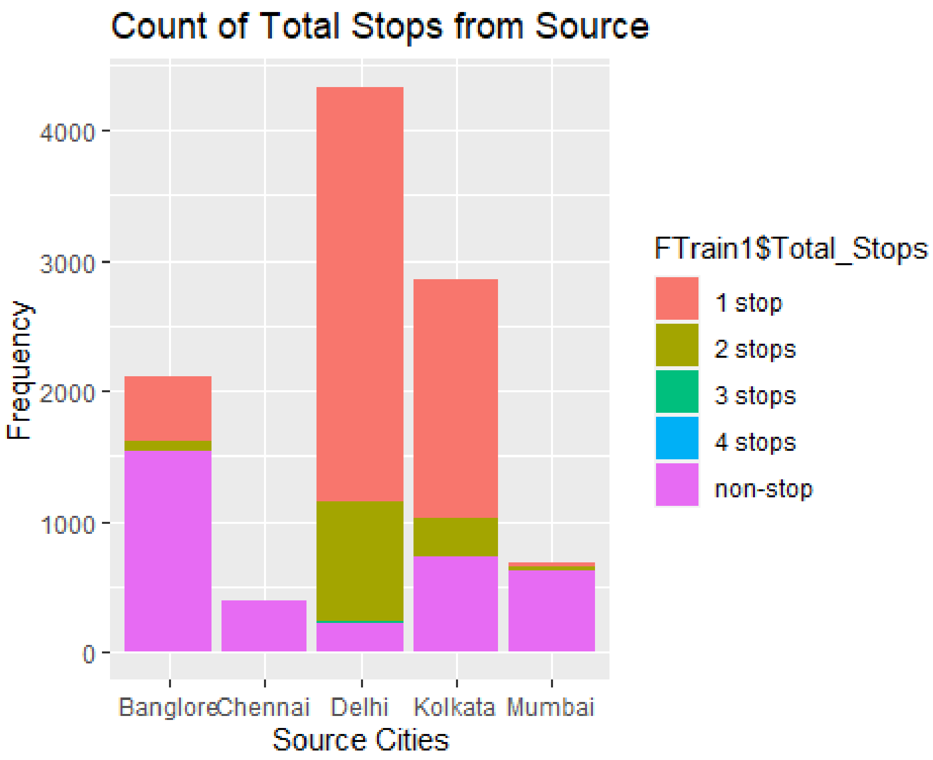

Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Vednesda
20000
15000
10000
5000
Price
Day
NightTime
Hourbracket

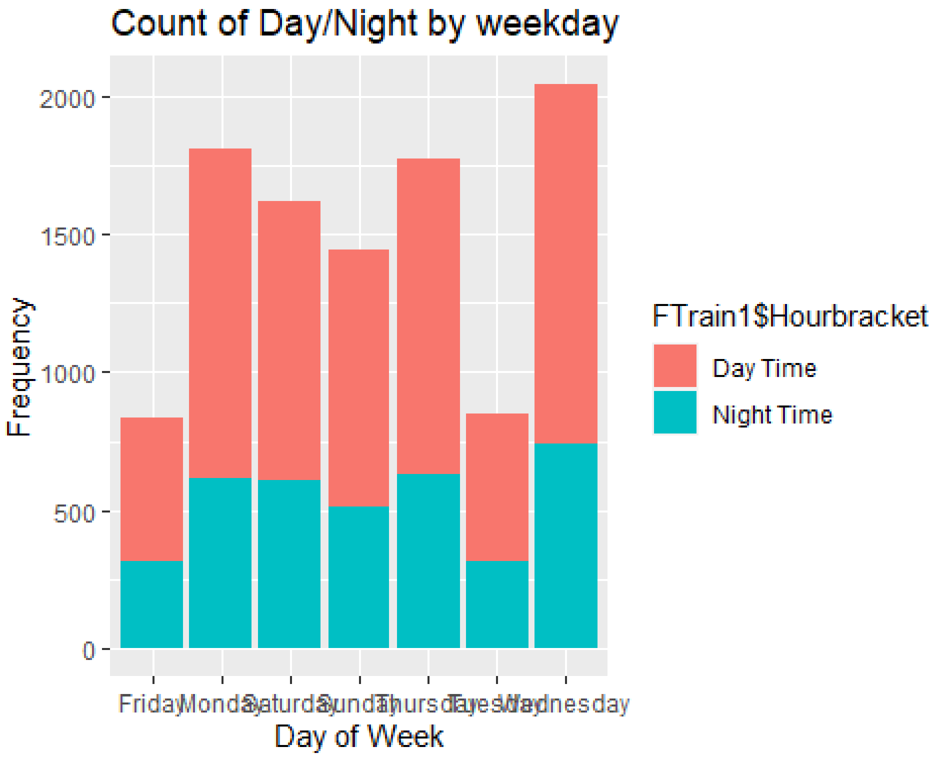

Price by Source City
Price
20000
15000
10000
5000
Banglore
Chennai
Delhi
Kolkata
Mumbai
Source City

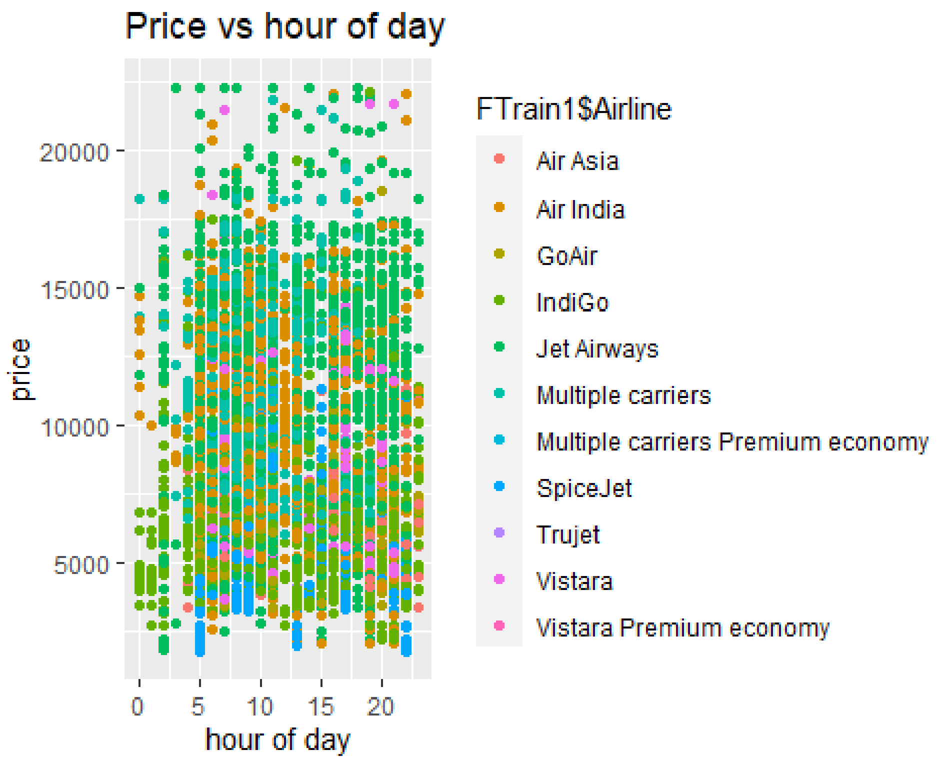

05000 15000
Thursday
Tuesday
Wednesday
20
15
10
5
0
Friday
Monday
Saturday
Sunday
20
15
10
5
0
Percent of Total
05000 15000
05000 15000
FTrain1$Price

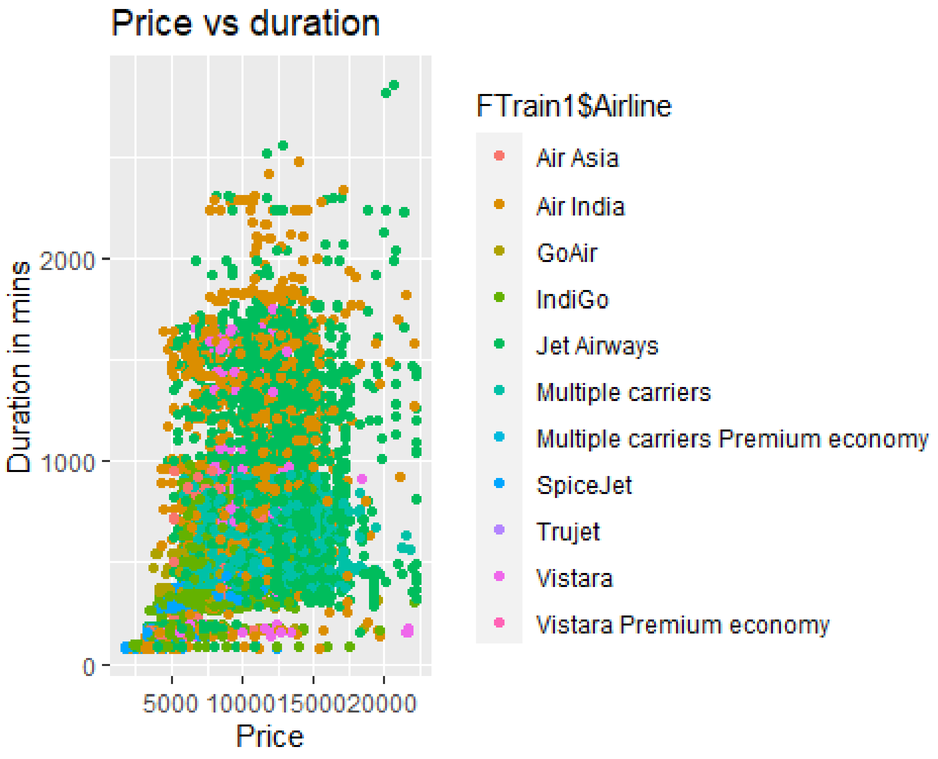

Count of Airline by Source
Frequency
0
1000
2000
3000
Airline
Air A
ium economy
FTrain1$Source
Banglore
Chennai
Delhi
Kolkata
Mumbai

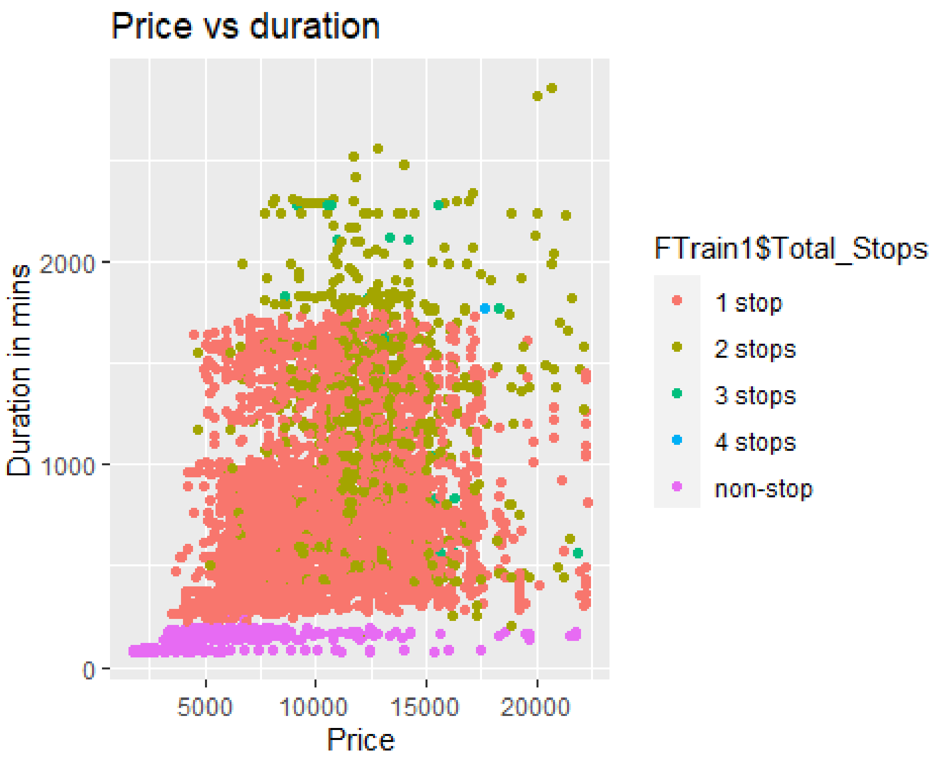

Count of Airline by Destination
Frequency
0
1000
2000
3000
Airline
Air A
um economy
FTrain1$Destination
Banglore
Cochin
Delhi
Hyderabad
Kolkata
New Delhi

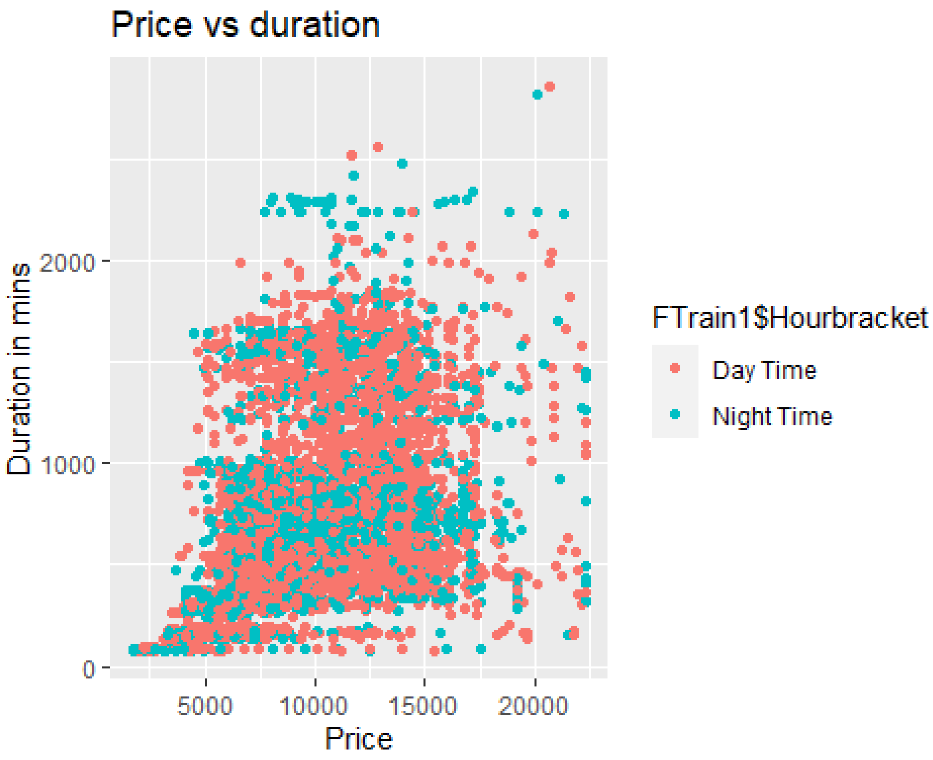

Count of Total Stops from Source
Frequency
4000
3000
2000
1000
0
Banglore
Chennai
Delhi
Kolkata
Mumbai
Source Cities
FTrain1$Total_Stops
1 stop
2 stops
3 stops
4 stops
non-stop

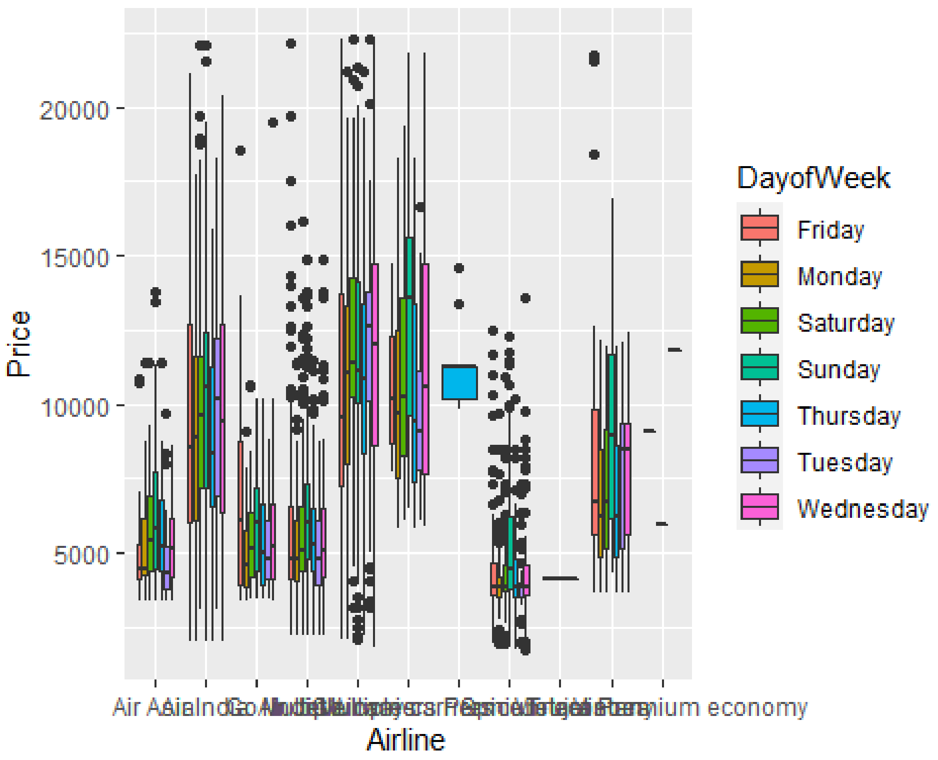

Count of Day/Night by weekday
Frequency
2000
1500
1000
500
0
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday
Day of Week
FTrain1$Hourbracket
Day Time
Night Time

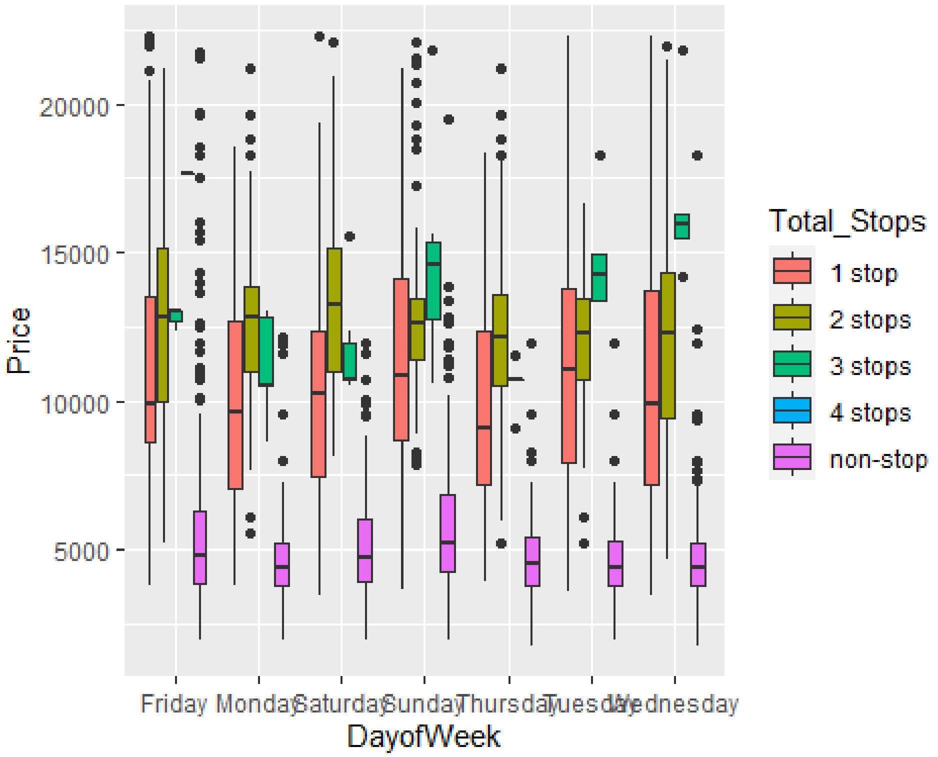

Price vs hour of day
price
20000
15000
10000
5000
0
5
10
15
20
hour of day
FTrain1$Airline
Air Asia
Air India
GoAir
IndiGo
Jet Airways
Multiple carriers
Multiple carriers Premium economy
SpiceJet
Trujet
Vistara
Vistara Premium economy

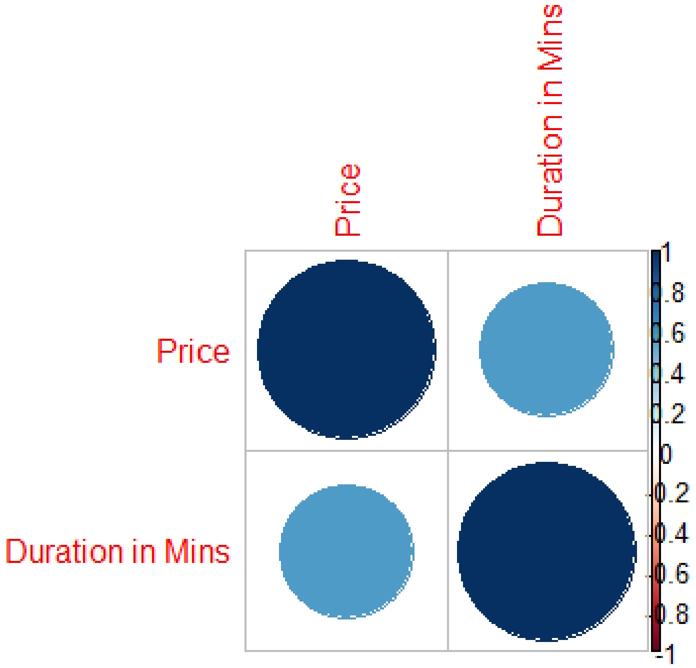

Price vs duration
Duration in mins
2000
1000
0
5000
10000
15000
20000
Price
FTrain1$Airline
Air Asia
Air India
GoAir
IndiGo
Jet Airways
Multiple carriers
Multiple carriers Premium economy
SpiceJet
Trujet
Vistara
Vistara Premium economy


Price vs duration
Duration in mins
0
1000
2000
5000
10000
15000
20000
Price
FTrain1$Total_Stops
1 stop
2 stops
3 stops
4 stops
non-stop


Price vs duration
Duration in mins
0
1000
2000
5000
10000
15000
20000
Price
FTrain1$Hourbracket
Day Time
Night Time


Price
20000
15000
10000
5000
Air Asia
Air India
GoAir
IndiGo
Premium economy
Airline
DayofWeek
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday


Price
20000
15000
10000
5000
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday
DayofWeek
Total_Stops
1 stop
2 stops
3 stops
4 stops
non-stop



Multiple Linear Regression

Model Intrepretation

For the model building , we take encoded data set set where all the categorical variables have been encoded with dummy variables as 0 or 1 . The intercept consider the effect of variables that we have not considered. We can intrepret the model such as For each one unit change of independent variable the price increase or decrease by the amount represented as slope keeping other variable constant

R square interpretation - 63% of the variation in the flight price is explained by the independent variables used in the model

We reject null hypothesis as t-stat has extremely low p-values which is p<0.05 and there is significant evidence that regression model exist

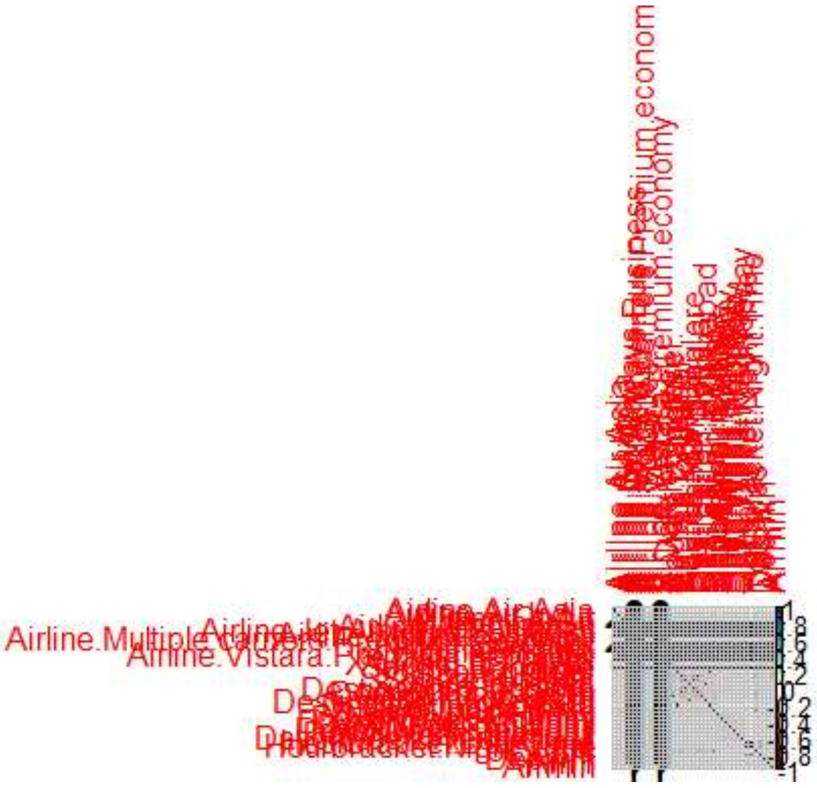
Adjusted R square shows the effect of adding more variables in the model. For eg current model is explained 63% by the independent variables which means remaining 40% of model is unexplained or by residuals.so we inflate the error component by multiplier which is division of Total degree of freedom by residual degree of freedom

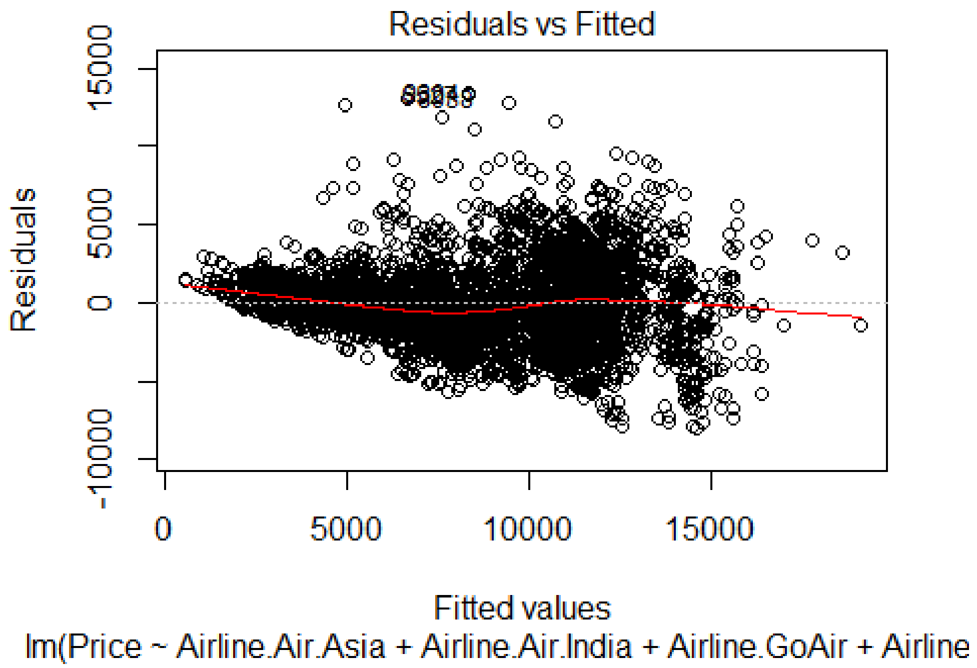
Based on the model and P-stat we keep only the significant variables in the final model building and again compute the R2 and root mean square error

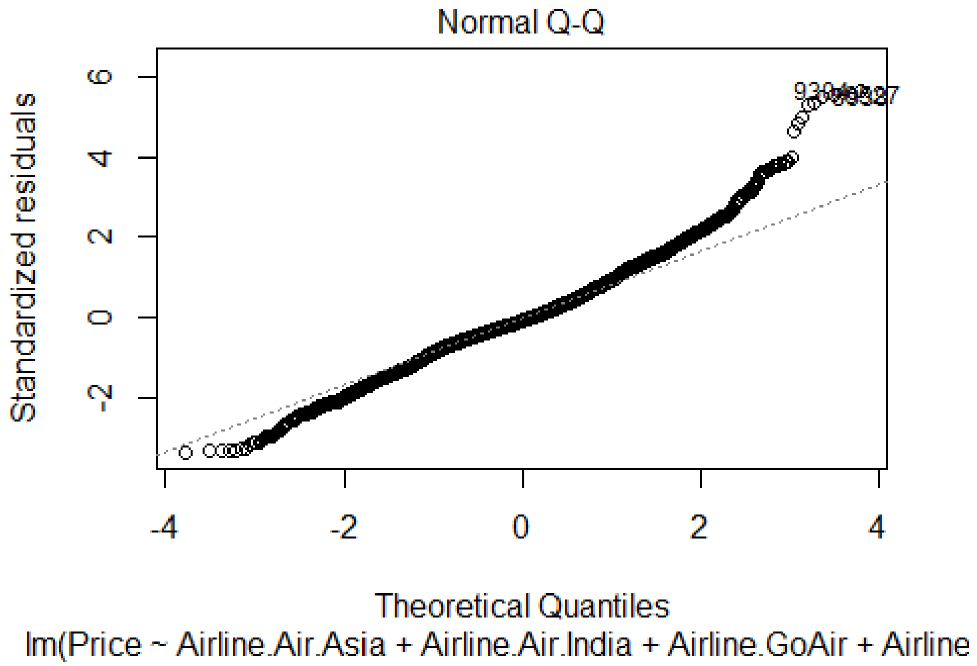
We also check for correlation between the models using VIF

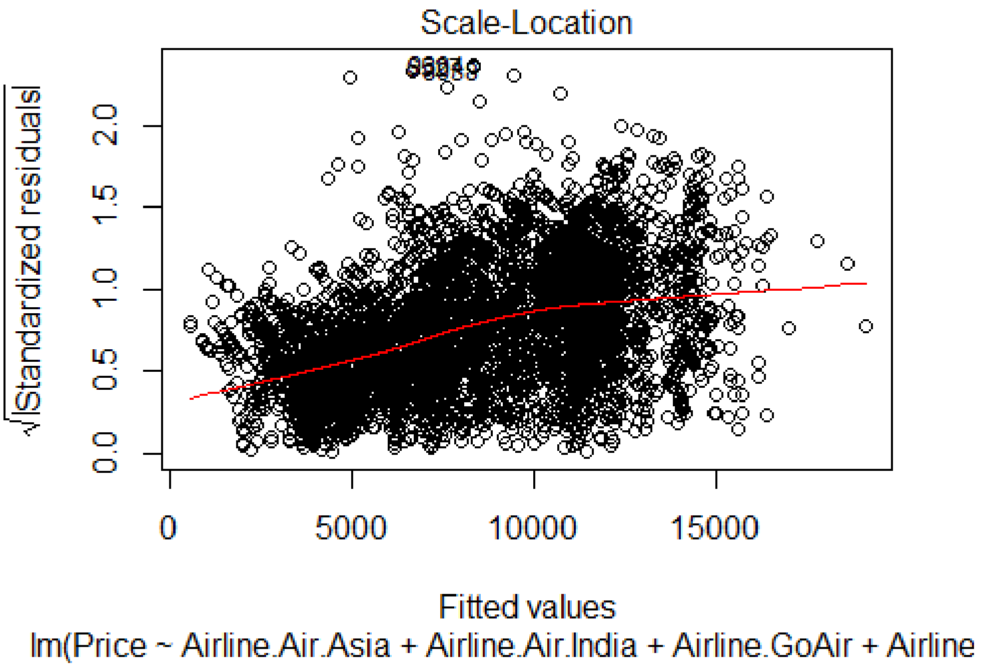
RMSE - 2629.421

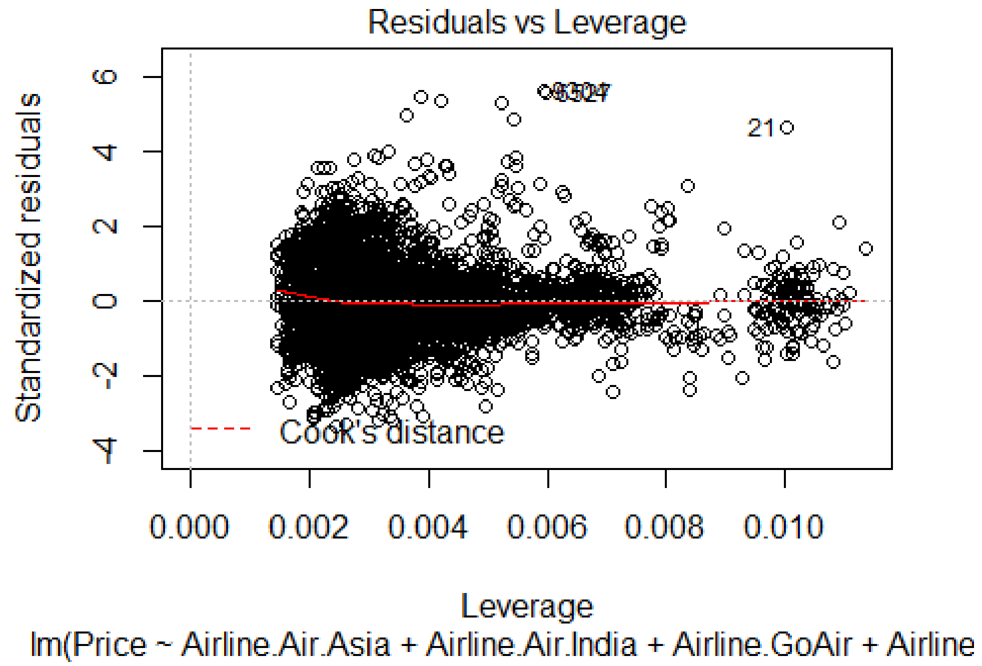
R2 - .633




Residuals vs Fitted
Residuals
15000
5000
0
-10000
0
5000
10000
15000
Fitted values
lm(Price ~ Airline.Air.Asia + Airline.Air.India + Airline.GoAir + Airline


Normal Q-Q
Standardized residuals
6
4
2
0
-2
-4
-2
0
2
4
Theoretical Quantiles
lm(Price ~ Airline.Air.Asia + Airline.Air.India + Airline.GoAir + Airline


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0
5000
10000
15000
Fitted values
lm(Price ~ Airline.Air.Asia + Airline.Air.India + Airline.GoAir + Airline


Residuals vs Leverage
Standardized residuals
-4
-2
0
2
4
6
21
Cook's distance
0.000
0.002
0.004
0.006
0.008
0.010
Leverage
lm(Price ~ Airline.Air.Asia + Airline.Air.India + Airline.GoAir + Airline

Ridge Regression

Regularization -

Normal regression works by selecting the coefficients that minimize the loss function, however if the coefficient are large there may be chance of overfitting the training data and will not generalize well on unseen test data .To overcome this we will do regularization that will impact the large coefficients

Ridge Regression - Loss function is minimized by adding a penalty parameter equivalent to square of magnitude of coefficients. The model will be tuned thru the hyperparameter lambda. This model will generalize well on test data as it will be less sensitive to extreme variance.

RMSE - 2579 R2 - .634

Lasso Regression

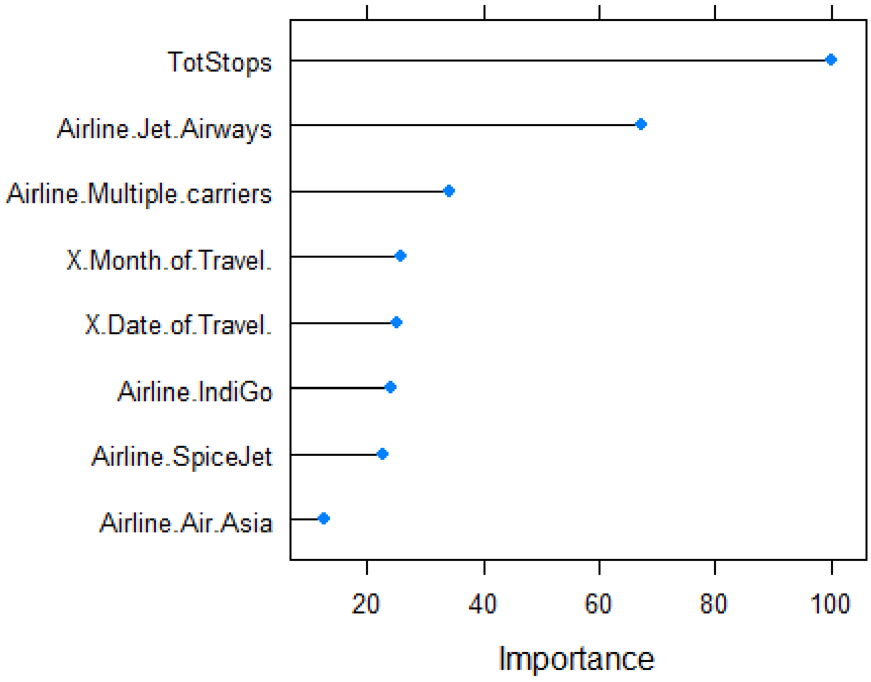
Lasso Regression - in this loss function is modified to minimize the complexity of model by limiting sum of absolute values of model coefficients

RMSE - 2626 R2 - .633

Elastic Net Regression

Elstic Net - It uses the properties of both ridge and lasso . It works by penalizing the model using both optimum alpha and lambda values. We use caret package to find optimal values and using the hyper tuning parametrs

RMSE - 2627 R2 - .633



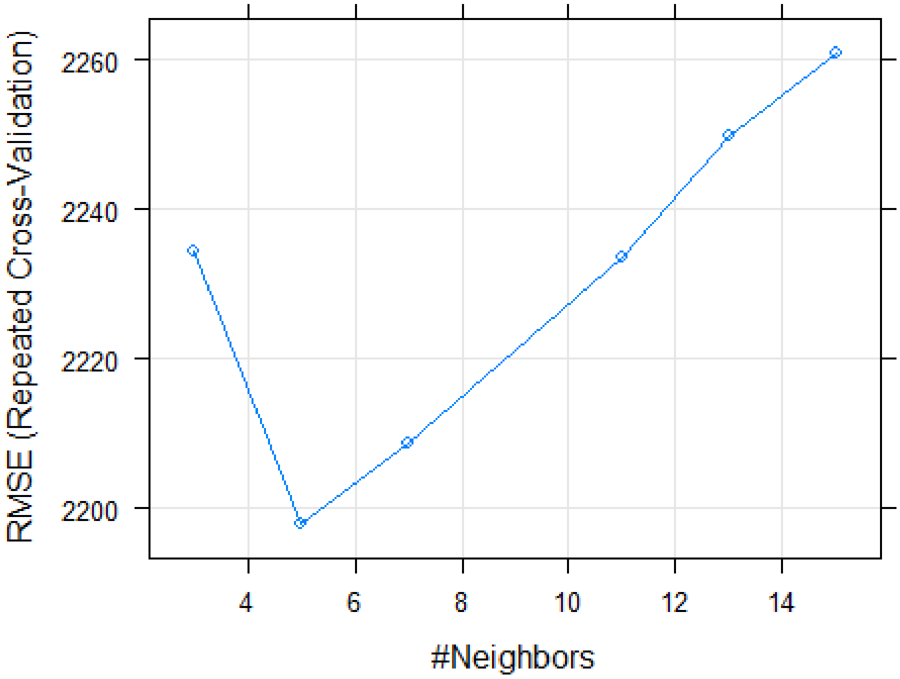
KNN Regressor

KNN - used caret package , to find the optimal number of nearest neighbors to predict the value. We used cross validation method where 9 samples were used for train and one is used for test

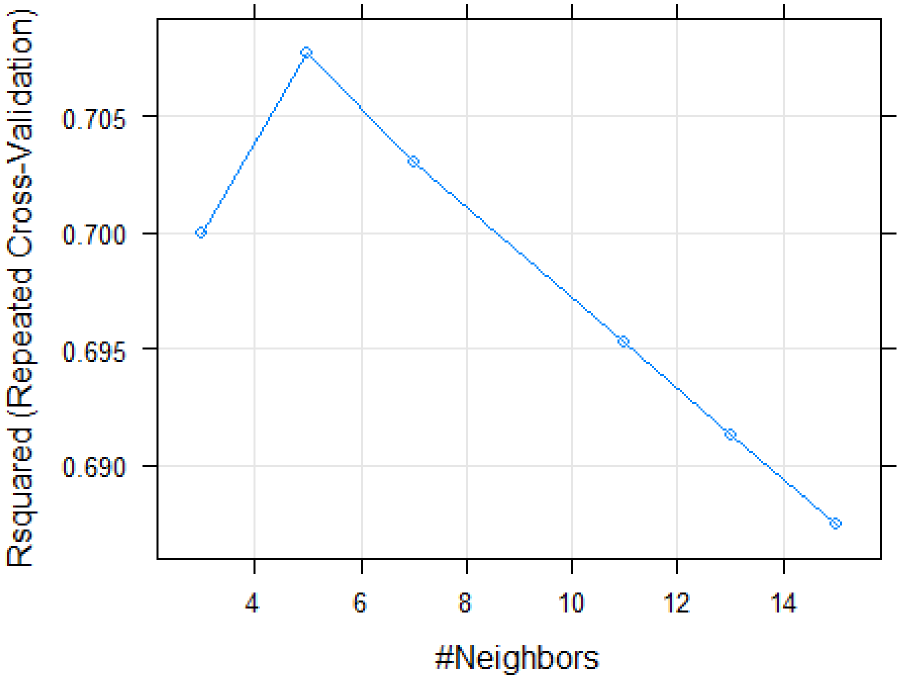
Minimum RMSE is achieved when K is 5 , we use both default method of RMSE and Rsquared method, and find out that Rsquare gives a slight better RMSE reduction

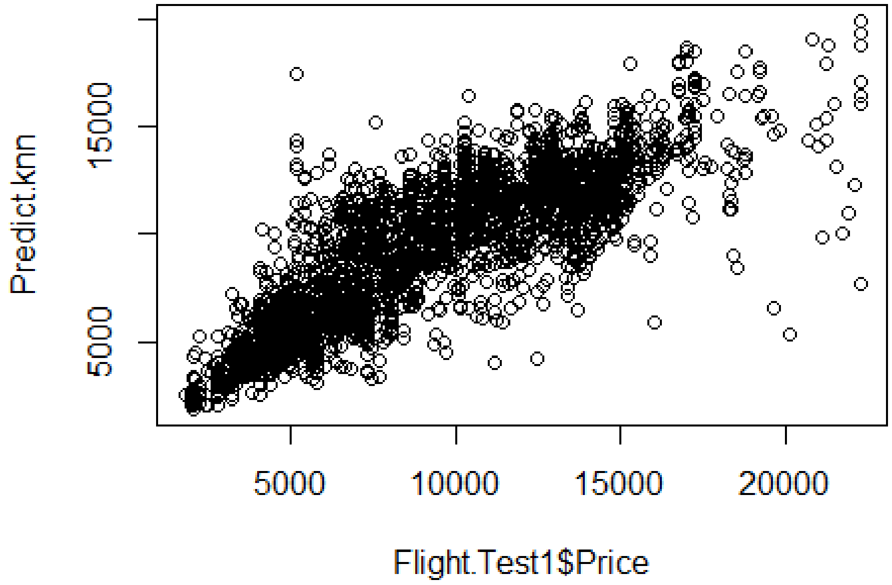
RMSE - 2116 R2 - .73

```
plot(fit.knn)
```



```
plot(fit.knn1)
```





Decision Trees

Decision Tree - In this method the we split the tree starting from root node based on the condition that each split will have minimum impurity or less Gini index . Impure nodes will have more Gini index and have maximum variance .Pure node will have zero Gini index and impure as maximum Gini of 0.5 The split should be such that when we come from root node to child node the Gini gain is maximum for that particular split compared to other split for different predictors and Gini impurity reduces for that node compared to root node.

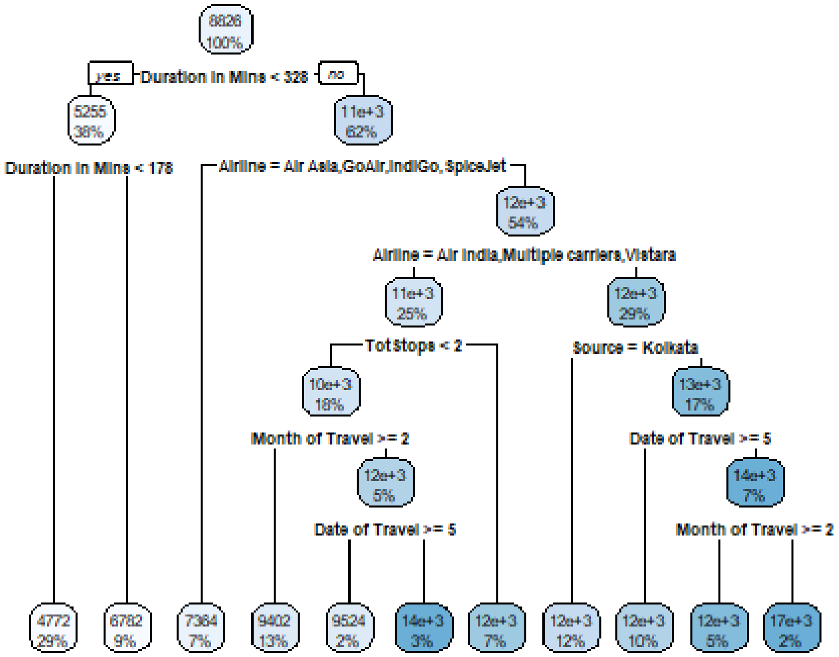
Decision Trees are powerful but have tendency to over fitting the data.We prune the tree to avoid overfitting based on complexity Parameter which says that error decrease should be more than Alpha ,if it is less than we stop adding the branches . Alpha is threshold and we generally take around 0.015

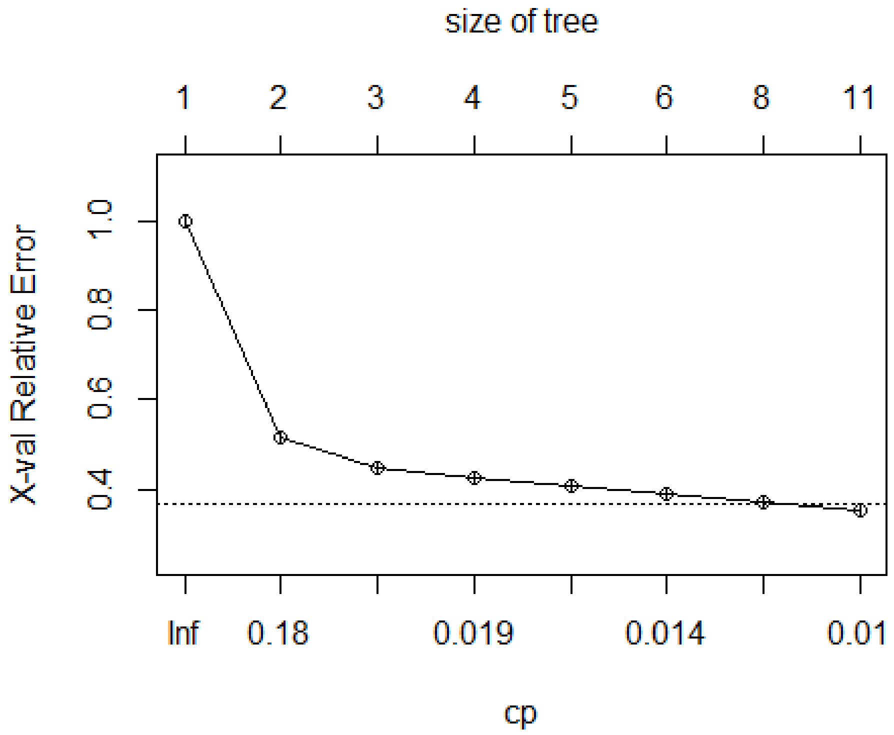
Common problem with Decision Trees are -

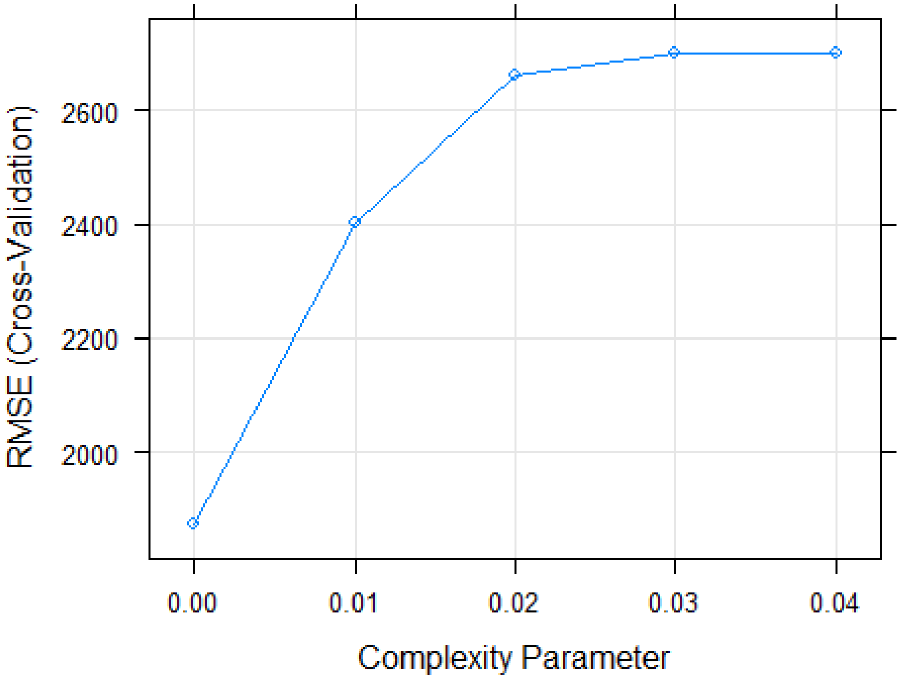
Too many branches - chances of overfitting (capture of noise) Too less branch - under fitting (Misclassification error)

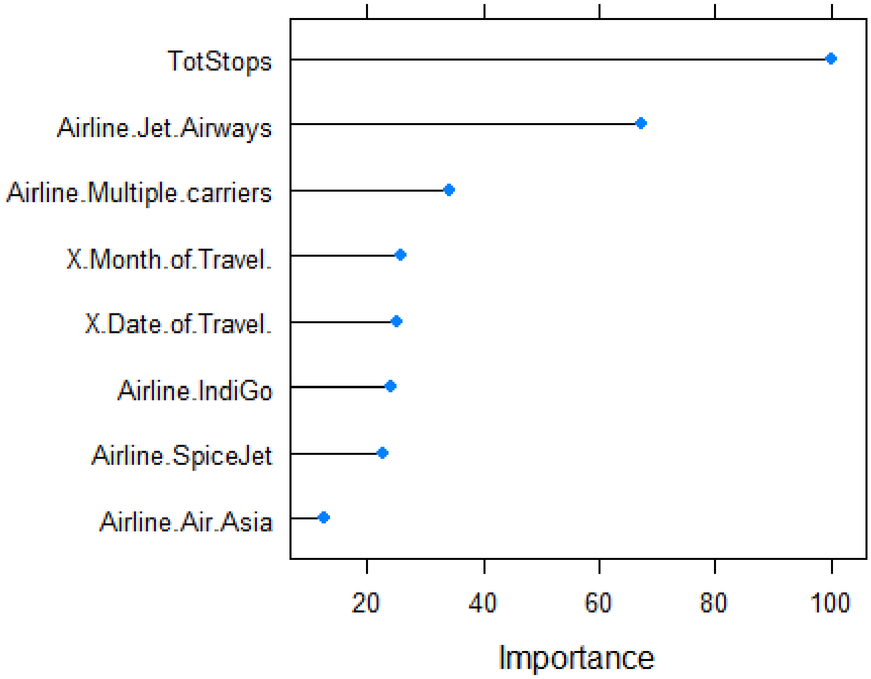
Optimum number of branches are arrived based on pruning.CP is error reduction per node . Choose the node where cross validation error is minimum i.e xerror.

RMSE - 1885.7 R2 -.787


8826
100%
yes
Duration in Mins < 328
no
5255
38%
11e+3
62%
Duration in Mins < 178
Airline = Air Asia,GoAir,IndiGo,SpiceJet
12e+3
54%
Airline = Air India,Multiple carriers,Vistara
11e+3
25%
12e+3
29%
Tot Stops < 2
Source = Kolkata
10e+3
18%
13e+3
17%
Month of Travel >= 2
Date of Travel >= 5
12e+3
5%
14e+3
7%
Date of Travel >= 5
Month of Travel >= 2
4772
29%
6782
9%
7364
7%
9402
13%
9524
2%
14e+3
3%
12e+3
7%
12e+3
12%
12e+3
10%
12e+3
5%
17e+3
2%


size of tree
1
2
3
4
5
6
8
11
X-val Relative Error
1.0
0.8
0.6
0.4
Inf
0.18
0.019
0.014
0.01
cp


RMSE (Cross-Validation)
2000
2200
2400
2600
0.00
0.01
0.02
0.03
0.04
Complexity Parameter


TotStops
Airline.Jet.Airways
Airline.Multiple.carriers
X.Month.of.Travel.
X.Date.of.Travel.
Airline.IndiGo
Airline.SpiceJet
Airline.Air.Asia
20
40
60
80
100
Importance

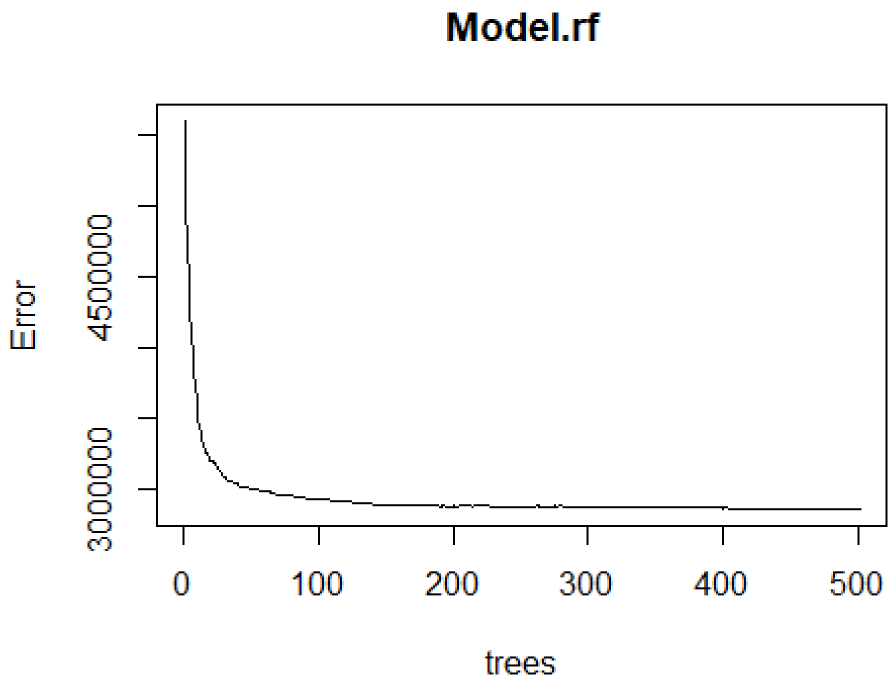
Random Forest

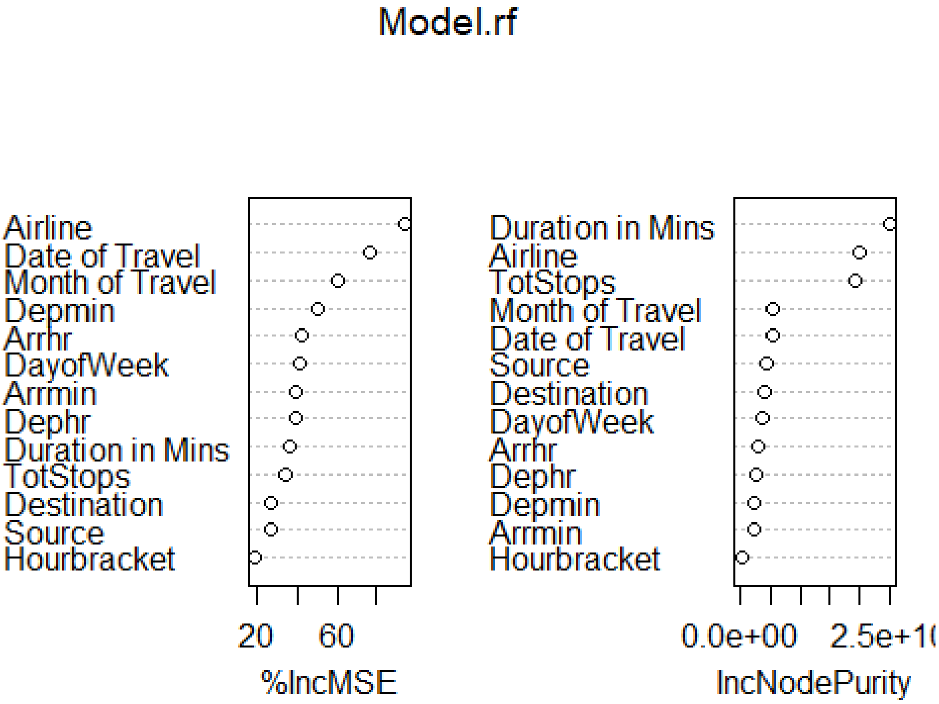
Random forest -this is an ensemble model technique in which we combine multiple models (Decision Tree in this case) to improve the predictive power of the model.

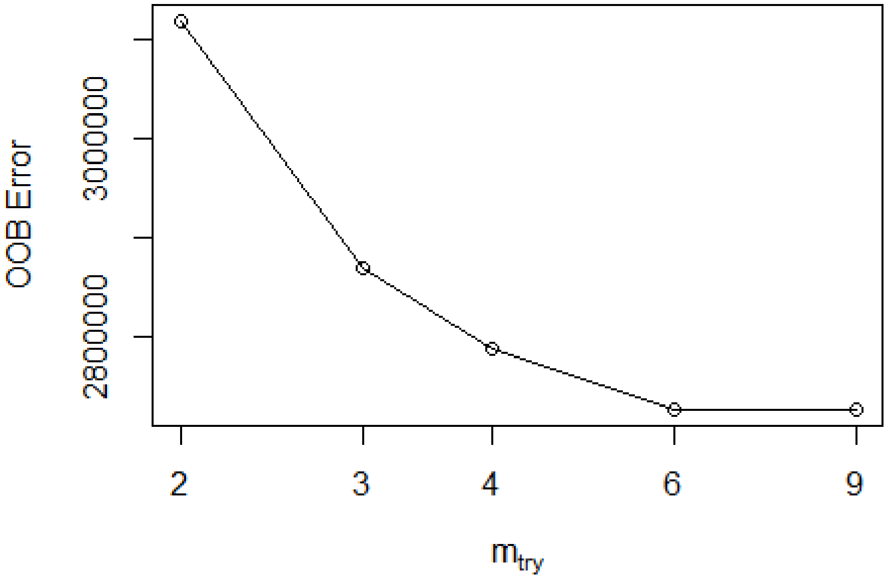
We use bootstrap aggregating (which is generating new training subset by sampling the data subset over and over again with replacement) Through this we ensure that the each tree built within the forest is diverse and at the same time share communality that they have been built from the same subset of data.

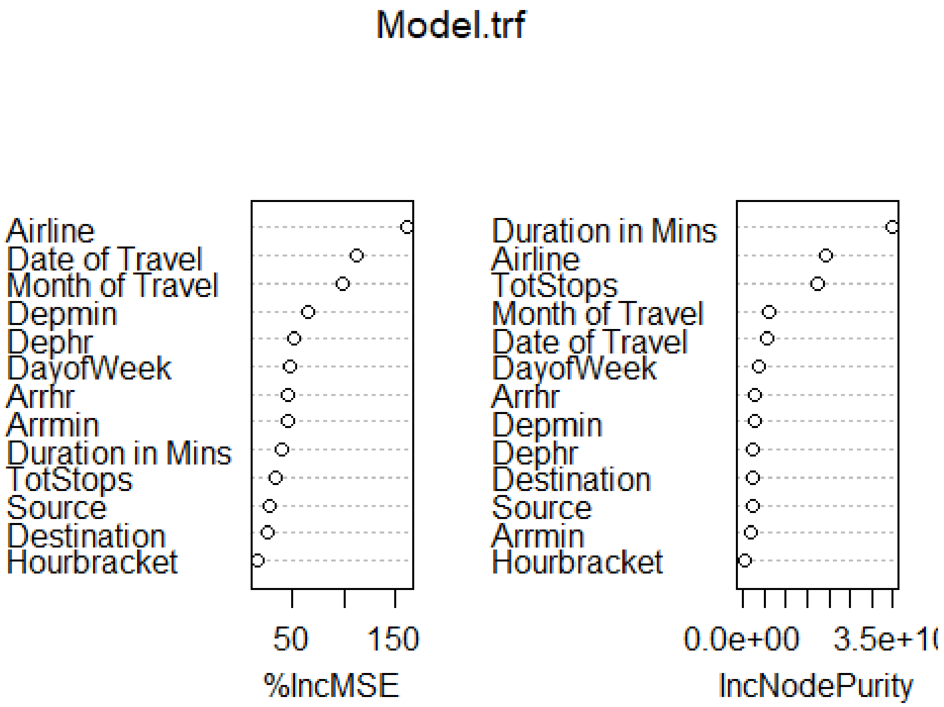
We have to be careful with value of m “mtry in mode”, if m is large variables become too correlated , if m is small, then predictive power of model decreases. Optimal choice of m plays vital role in random forest model.

RMSE - 1654.5 R2 - .836




Model.rf
Airline
Date of Travel
Month of Travel
Depmin
Arrhr
DayofWeek
Arrmin
Dephr
Duration in Mins
TotStops
Destination
Source
Hourbracket
20
60
%IncMSE
Duration in Mins
Airline
TotStops
Month of Travel
Date of Travel
Source
Destination
DayofWeek
Arrhr
Dephr
Depmin
Arrmin
Hourbracket
0.0e+00
2.5e+1(
IncNodePurity


OOB Error
3000000
2800000
2
3
4
6
9
m_{try}



Boosting method

Gradient Boosting

Boosting - we will do boosting to sequentially train the weak learners.Difference in bagging and boosting is that bagging is parallel and boosting is sequential

Gradient boosting method - It builds on each model, trying to fit the next model based on the residuals of previous model. We will use several tuning parameters to arrive at optimal model performance

RMSE - 2388.8 R2 - .66

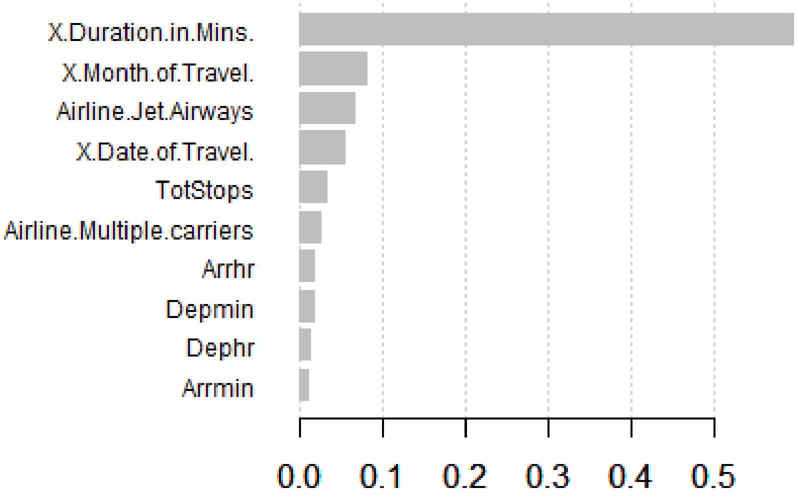
Extreme gradient Boosting

Extreme gradient boosting - specialized implementation of gradient boosting decision trees designed for performance . Types are gradient boosting ,stochastic and regularized boosting. Some of advantages of using XGboost are

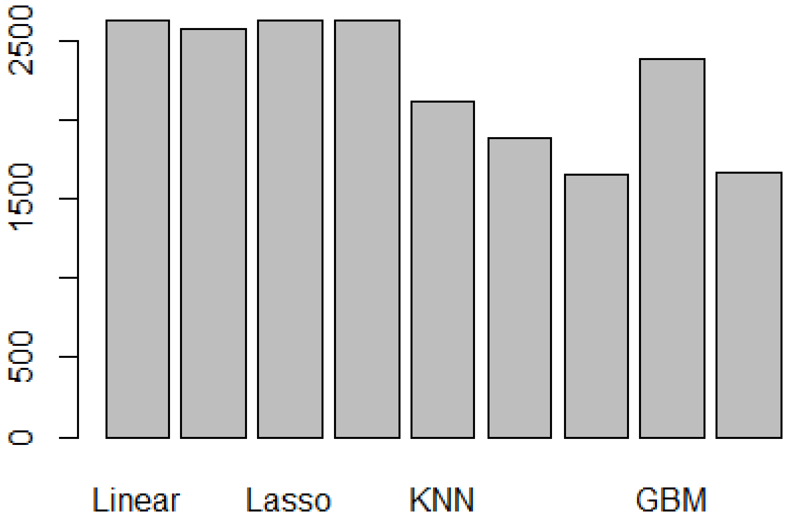
Parallel computing - it is enabled with parallel processing Regularization - it is used to avoid overfitting in linear and tree based models *Enabled cross validation - it is enabled with internal CV and not needed any additional package* Missing values - model can handle missing values *Tree pruning - it grows the tree upto a max depth and then prune backward until improvement in loss function.* Bias and Variance - Unlike bagging model like Random forest which takes care of overfitting (high variance) but still retains some bias, Boosting method can handle both bias and variance

We will use several hyper parameters to tune the model . These values can be further ioptimized to get a most optimum XGBoost model. In my current case i have take learning rate as 0.045, depth as 6 and number of rounds as 650

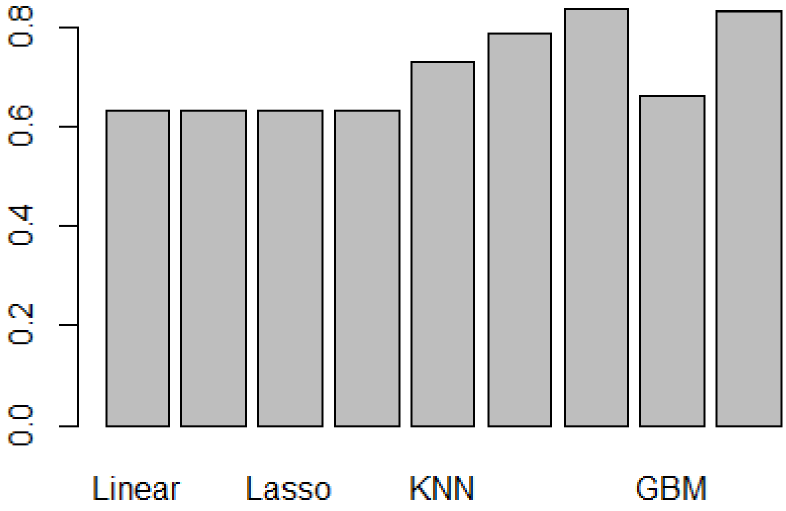
RMSE - 1677 R2 - .832



Final model building for predicting the Airfare prices for Test dataset.



```r
barplot.default(FModel$Rsquare , names.arg = Model)
```


Linear
Lasso
KNN
GBM