Flight_Capstone

Nitin Yadav

29/02/2020

1. Introduction

a. 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.

b. 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

c. 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.

2. Data Report

a. 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

b. 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)

c. 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

3. Exploratory Data Analysis

a. 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

b. 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.

c. 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

d. Addition of New Variables

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

4. 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.

a. Independent variables that are significant

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.

b. 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.

c. Hypothesis Testing

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

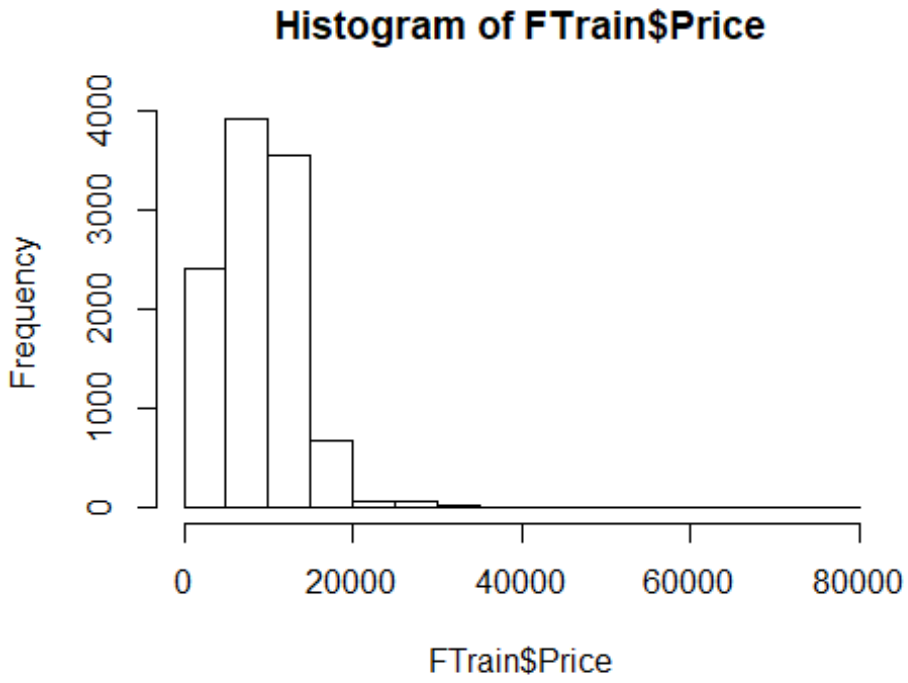
We did anova testing on the linear model built using Price and “DayofWeek” and found the P value very small and hence null hypothesis is rejected and we can say that flight price on weekends are costlier compared to weekdays

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

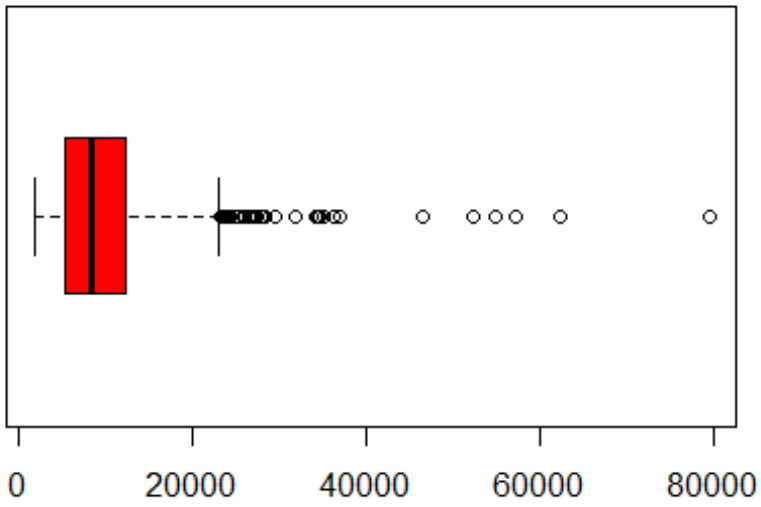
We did a 2-tail t test for the same and found that P value is very small and less than 0.05 , hence null hypothesis is rejected and hence Flight price during peak hours 9am-9pm are higher than non peak hours ie 9pm-9am.

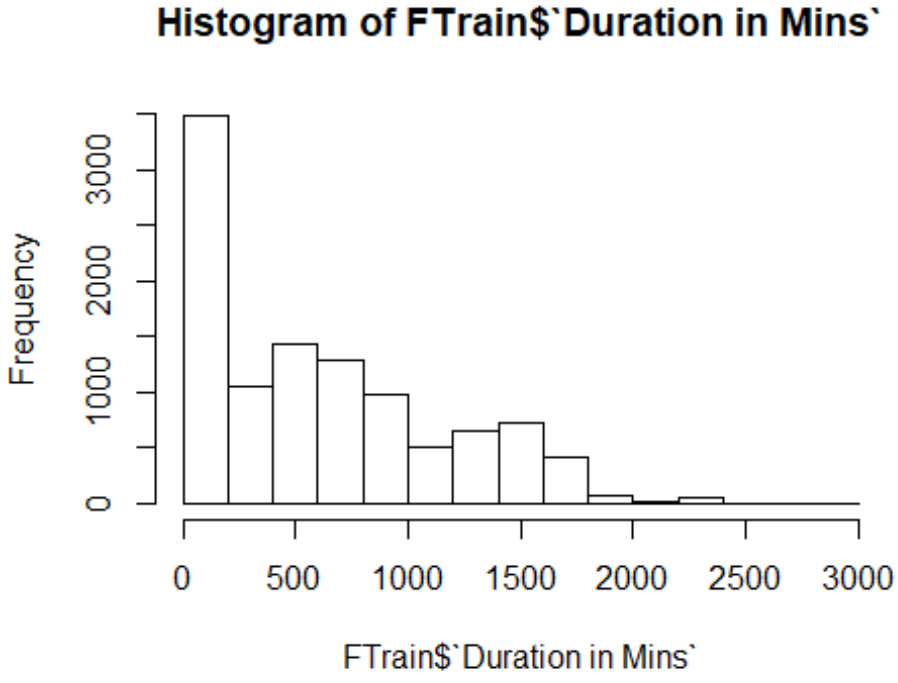
Appendix

(Attached below are the exploratory data analysis. These are not covered as part of report but provided for reference and additional insights.)

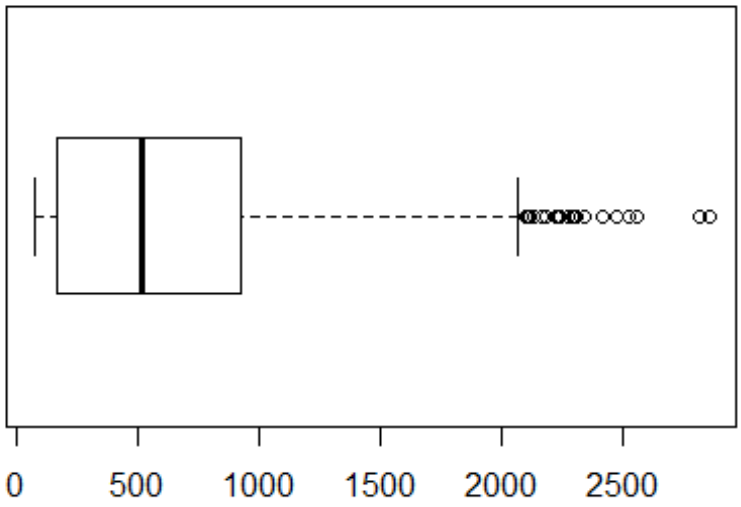


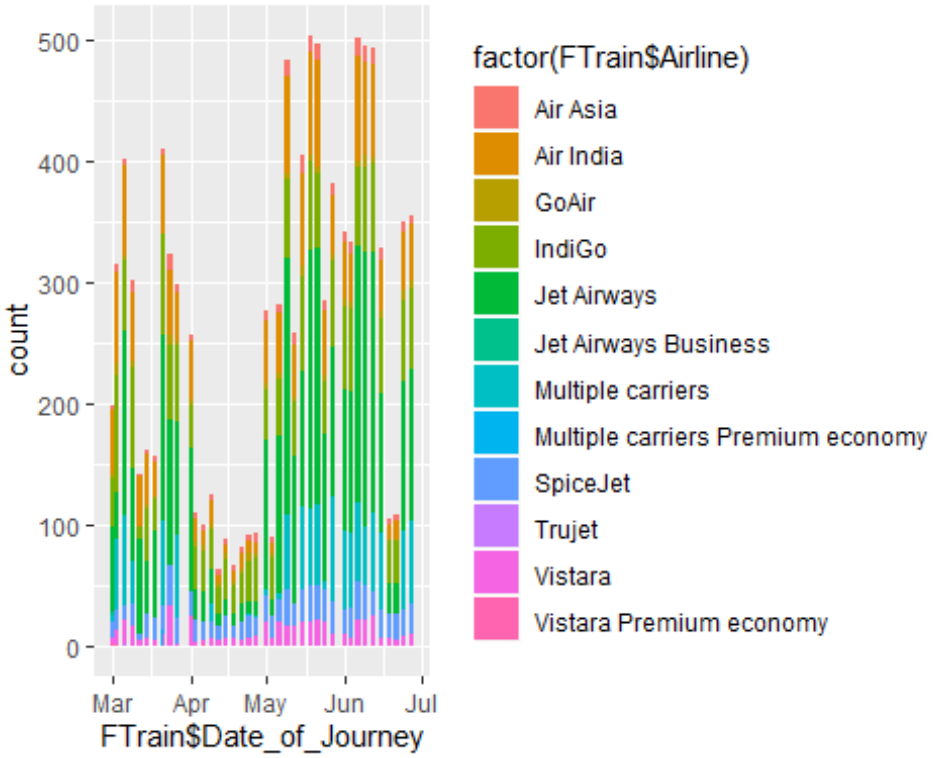
Boxplot for Price

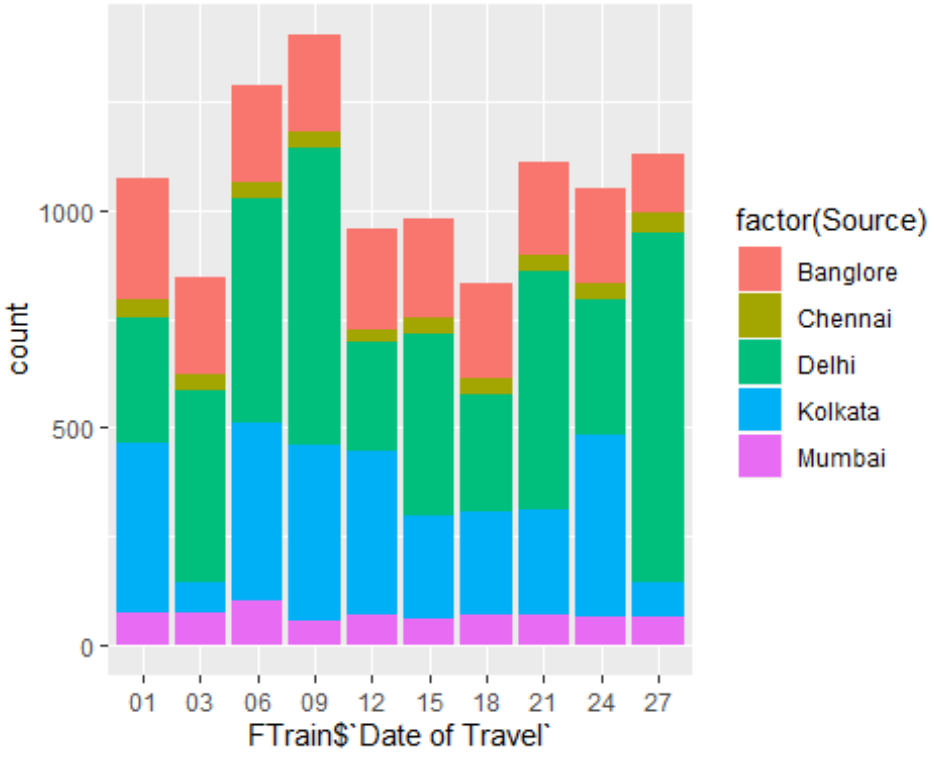


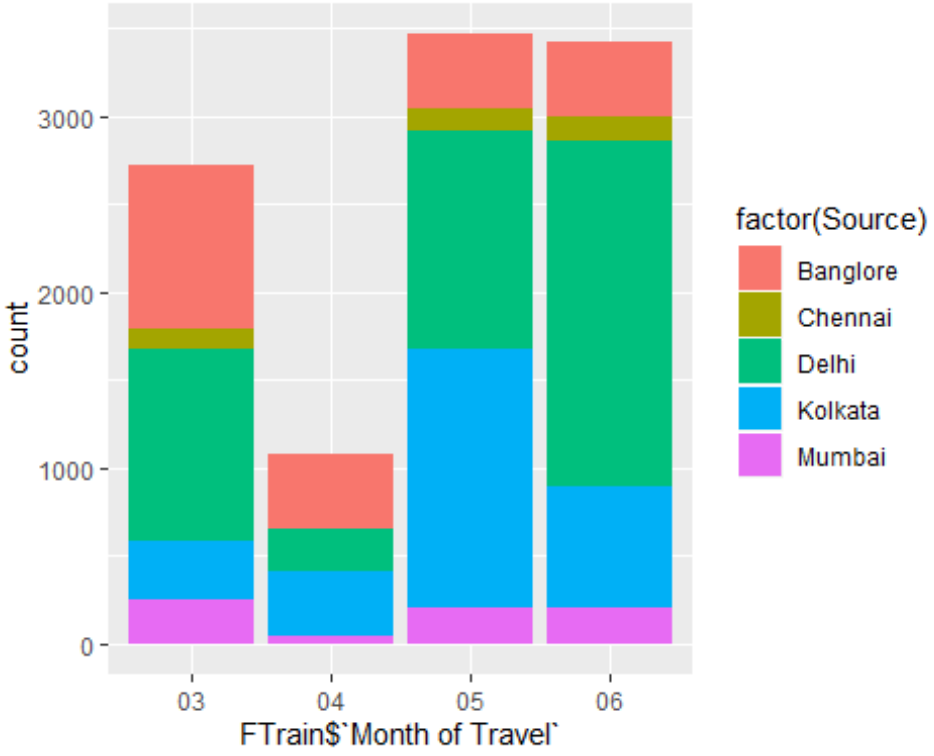


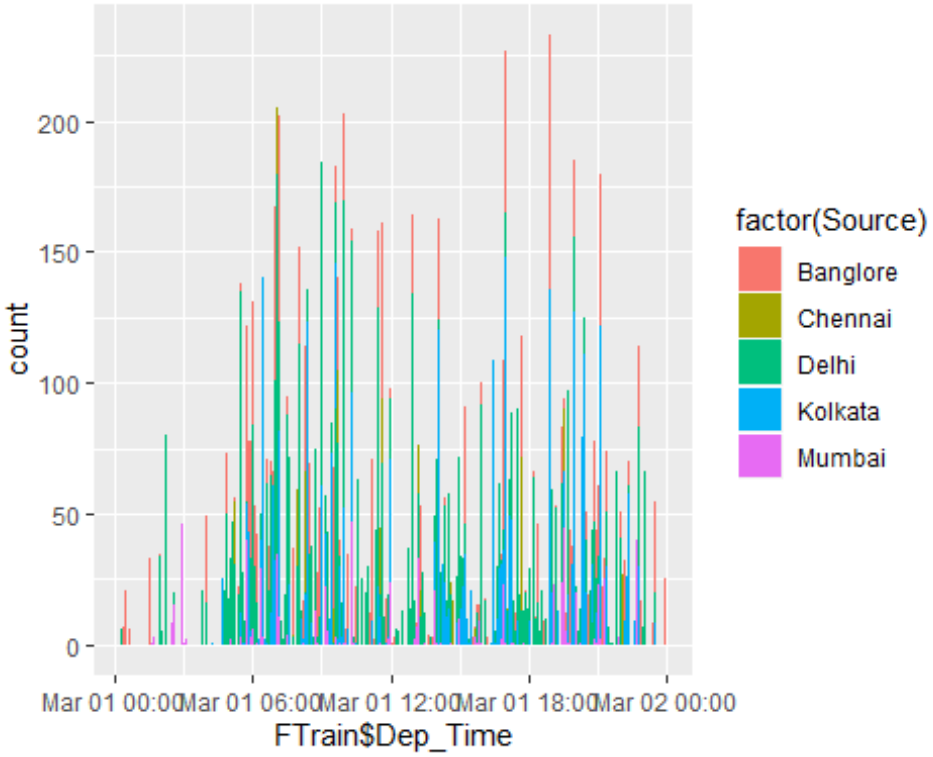
Boxplot for Duration

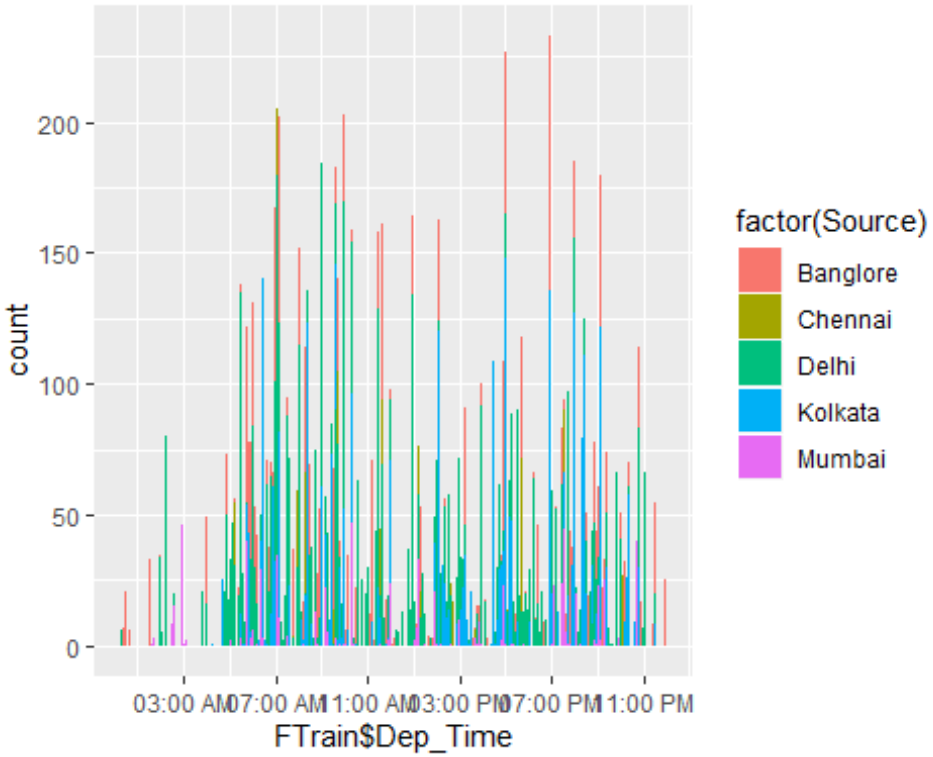

0
500
1000
1500
2000
2500

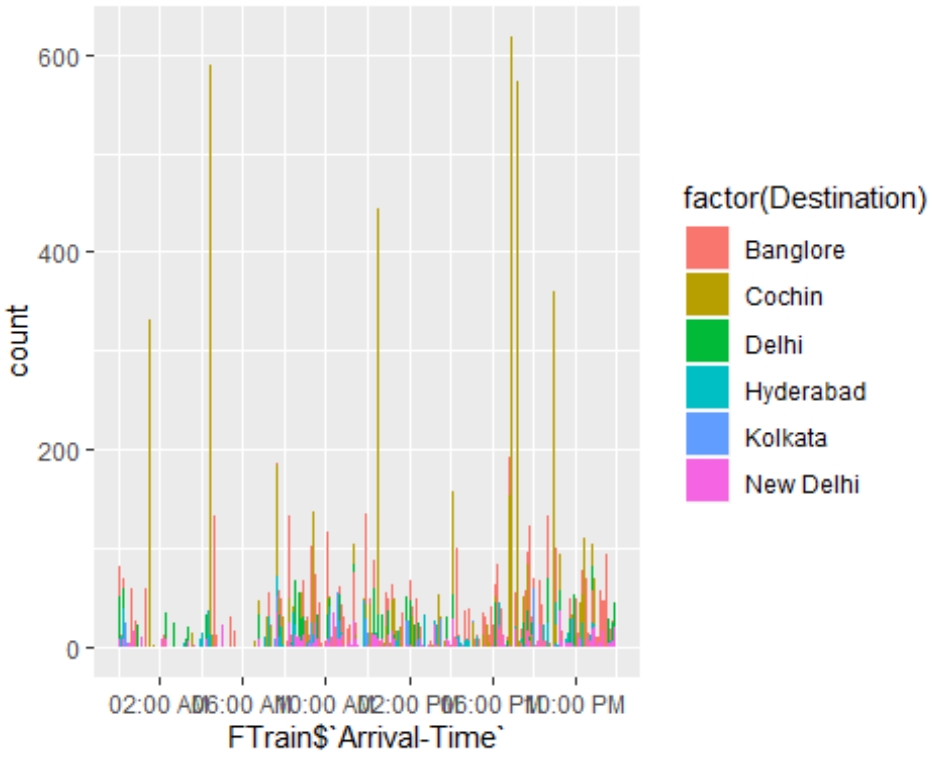

count
500
400
300
200
100
0
Mar
Apr
May
Jun
Jul
FTrain$Date_of_Journey
factor(FTrain$Airline)
Air Asia
Air India
GoAir
IndiGo
Jet Airways
Jet Airways Business
Multiple carriers
Multiple carriers Premium economy
SpiceJet
Trujet
Vistara
Vistara Premium economy

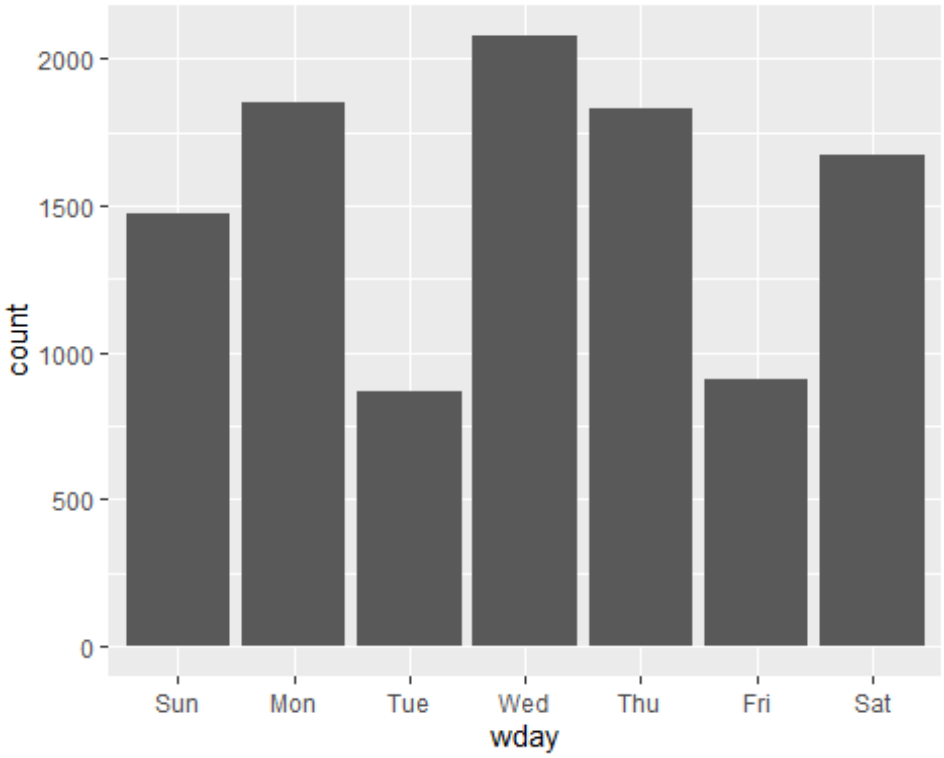

count
1000
500
0
01
03
06
09
12
15
18
21
24
27
FTrain$`Date of Travel`
factor(Source)
Banglore
Chennai
Delhi
Kolkata
Mumbai

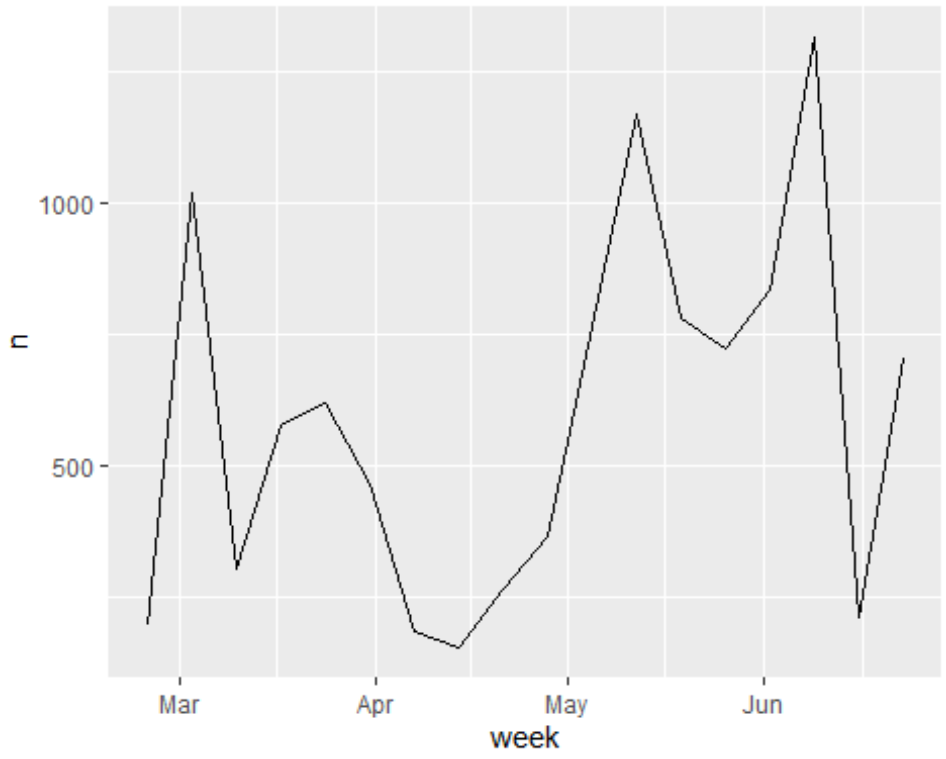

count
3000
2000
1000
0
03
04
05
06
FTrain$`Month of Travel`
factor(Source)
Banglore
Chennai
Delhi
Kolkata
Mumbai

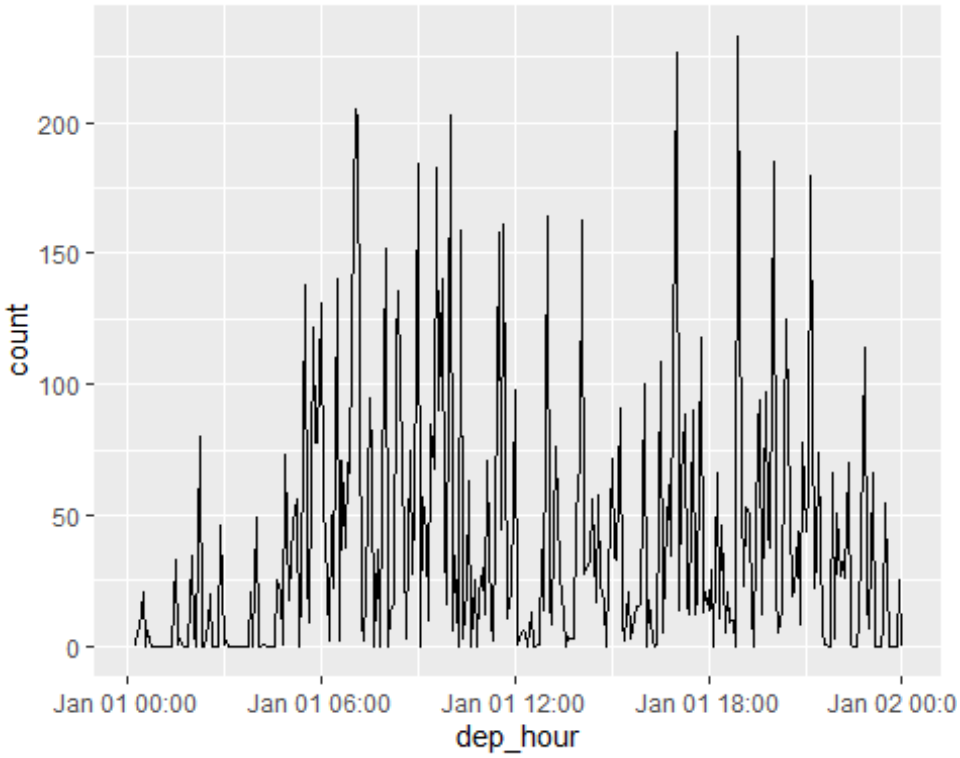

200
150
100
50
0
count
Mar 01 00:00
Mar 01 06:00
Mar 01 12:00
Mar 01 18:00
Mar 02 00:00
FTrain$Dep_Time
factor(Source)
Banglore
Chennai
Delhi
Kolkata
Mumbai

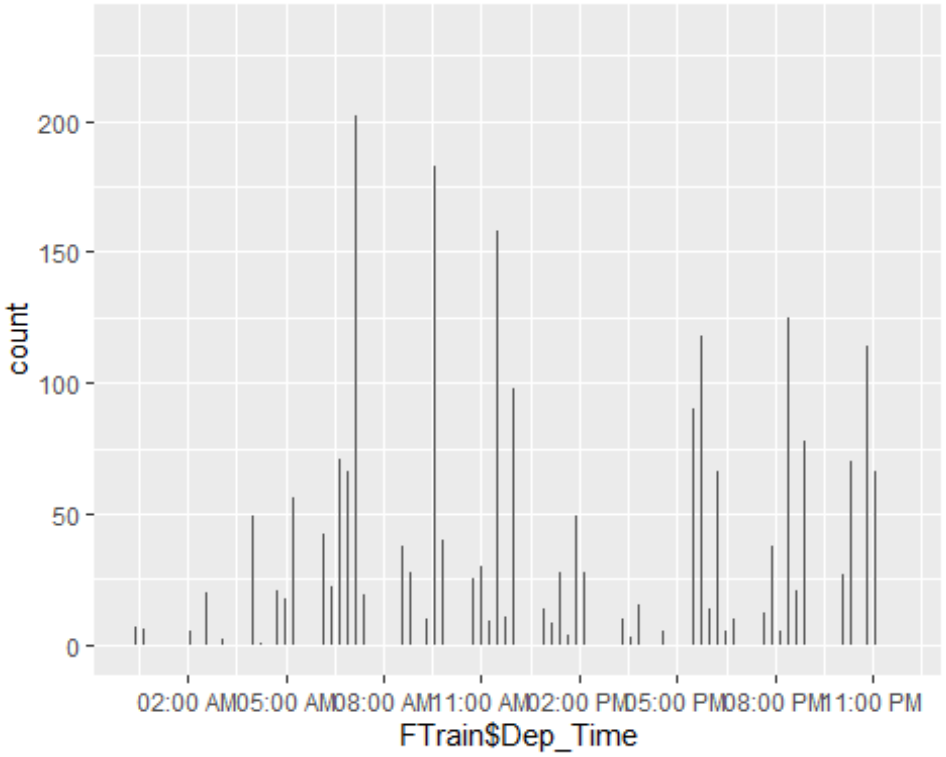

200
150
100
50
0
count
03:00 AM
07:00 AM
11:00 AM
03:00 PM
07:00 PM
11:00 PM
FTrain$Dep_Time
factor(Source)
Banglore
Chennai
Delhi
Kolkata
Mumbai

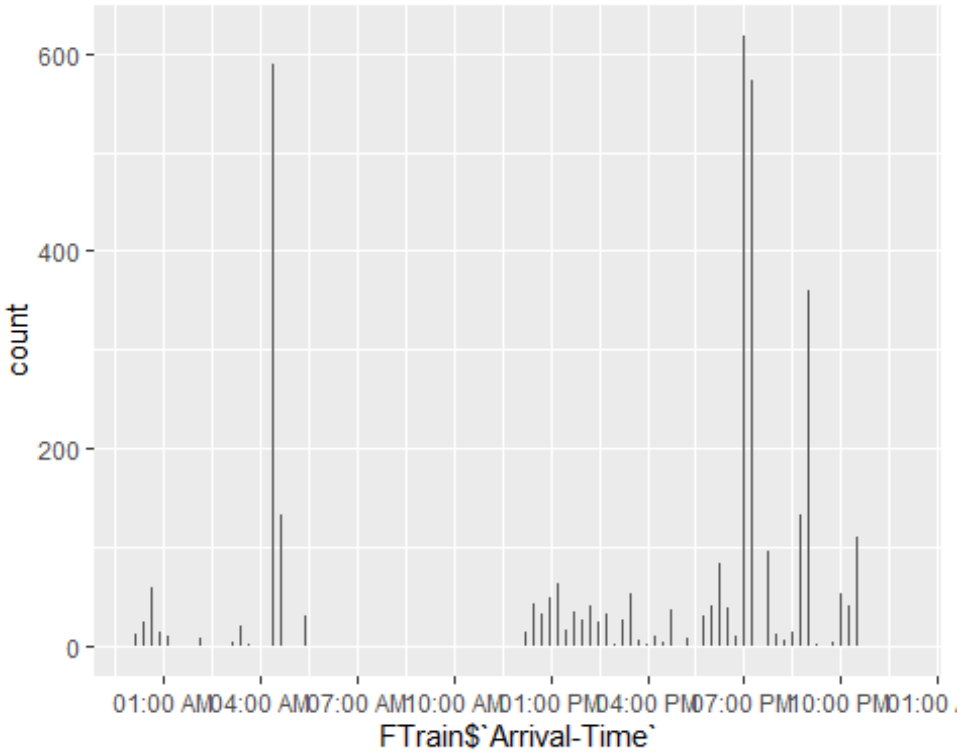

600
400
200
0
count
02:00 AM
06:00 AM
10:00 AM
02:00 PM
06:00 PM
10:00 PM
FTrain$`Arrival-Time`
factor(Destination)
Banglore
Cochin
Delhi
Hyderabad
Kolkata
New Delhi


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


1000
500
n
Mar
Apr
May
Jun
week


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



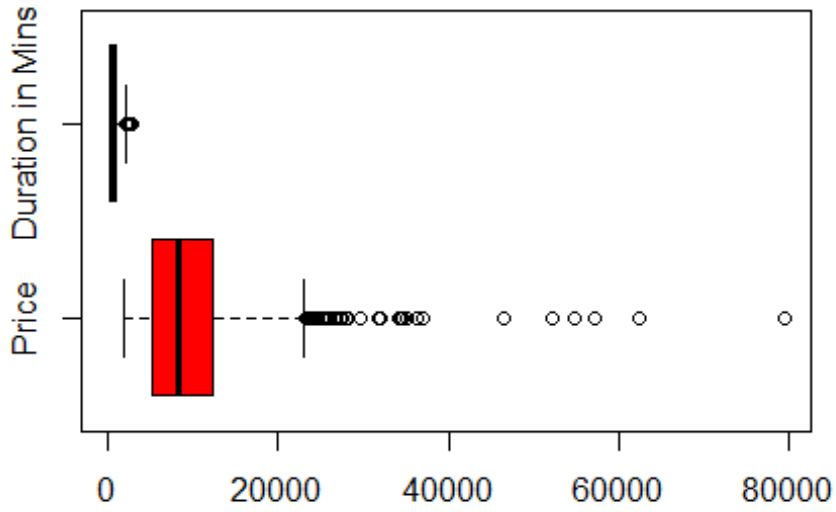


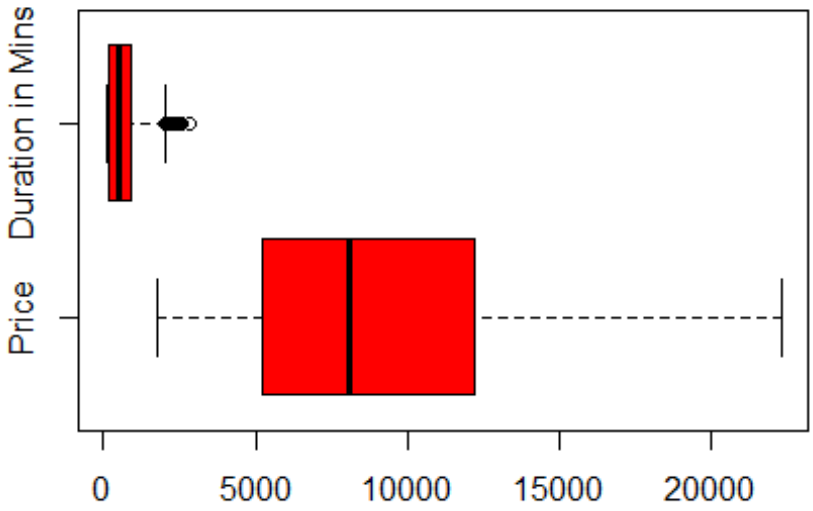

```
# Measurement of skewness and kurtosis
```

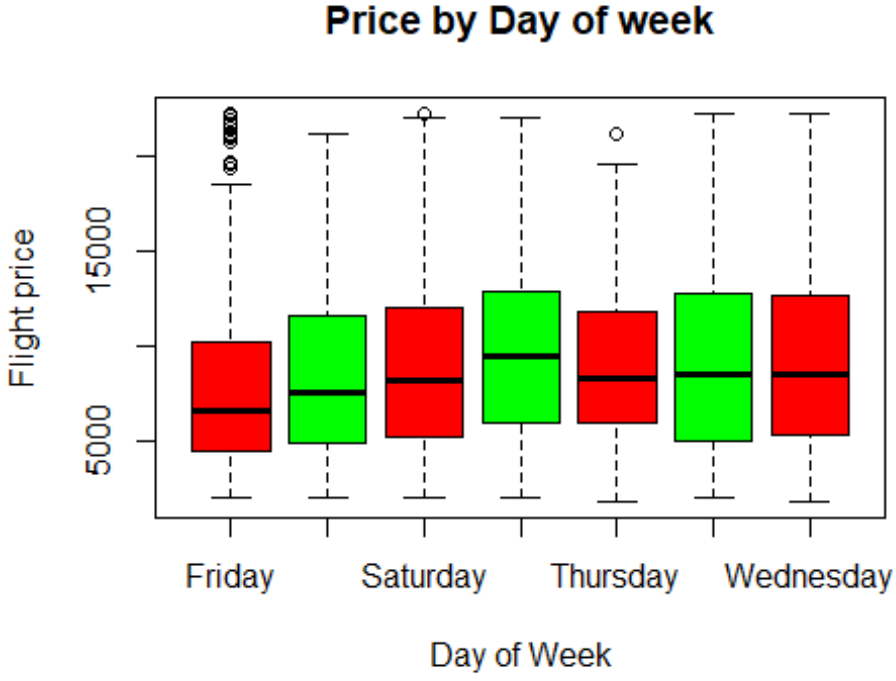
```
# Skewness is a mesure of symmetry , positive skewness in this case means the mean is more than median of the entries and hence it is right skewed

# Kurtosis define the tail shape of data distribution , in this we have excess kurtosis which is towards positive hence it indicates Fat tailed distribution or leptokurtic
```

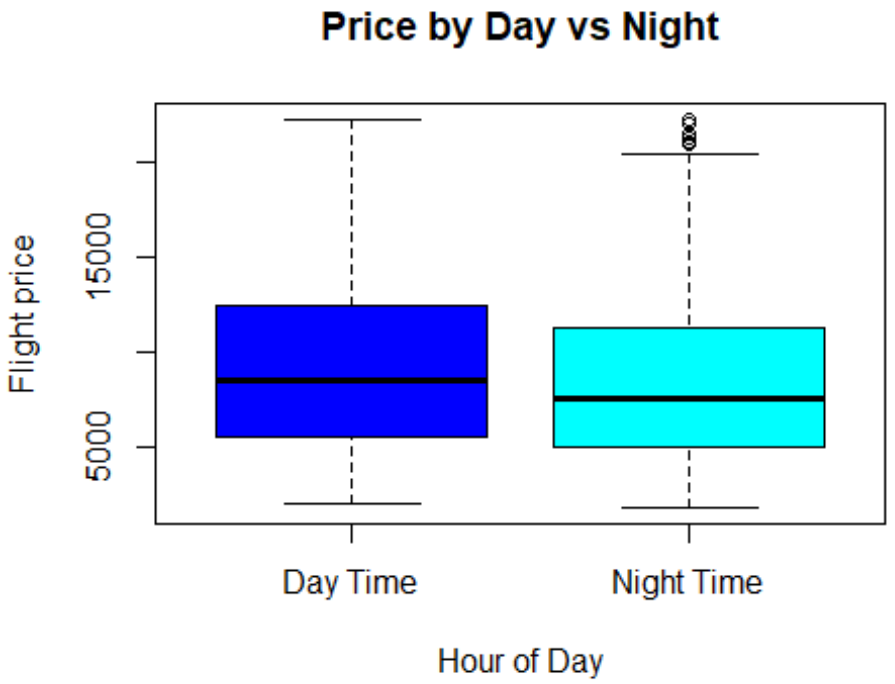
```
##   DayofWeek    Price
## 1    Friday 9709.828
## 2    Monday 8401.684
## 3  Saturday 8875.708
## 4    Sunday 9537.369
## 5  Thursday 8834.626
## 6   Tuesday 8945.932
## 7 Wednesday 9237.084
```

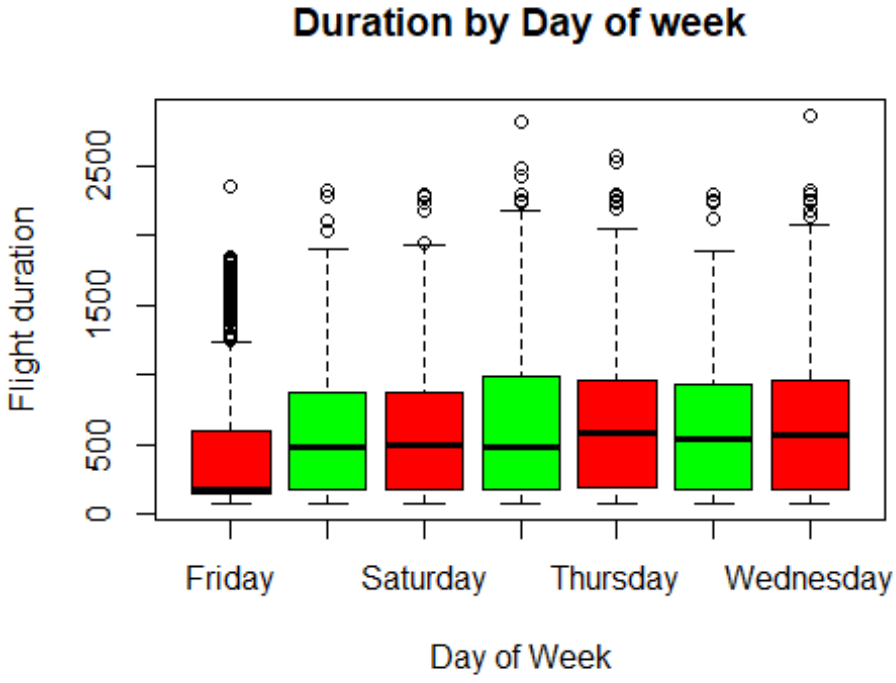
```
##   Hourbracket    Price
## 1    Day Time 9199.099
## 2  Night Time 8723.786
```

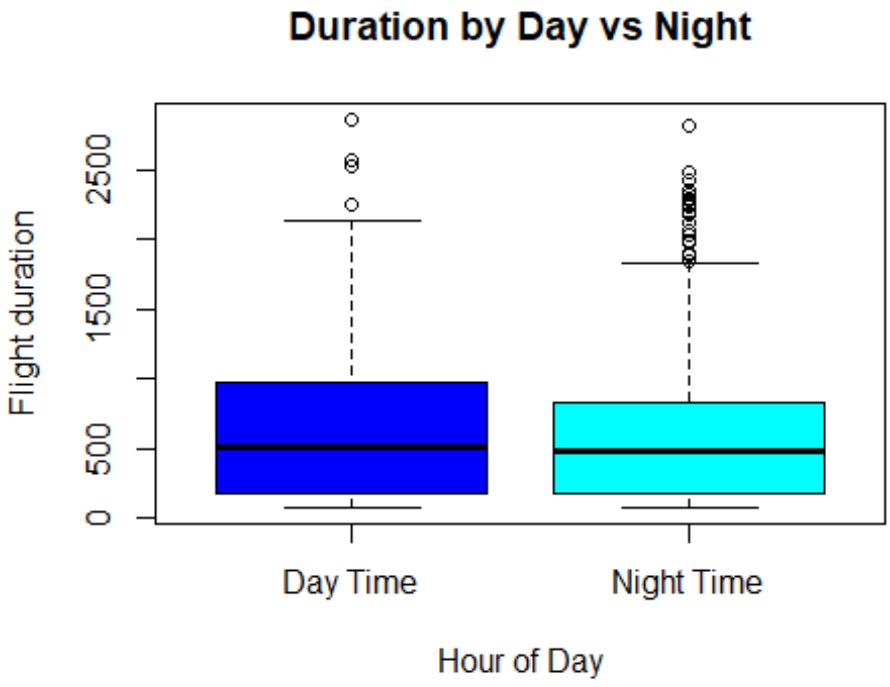


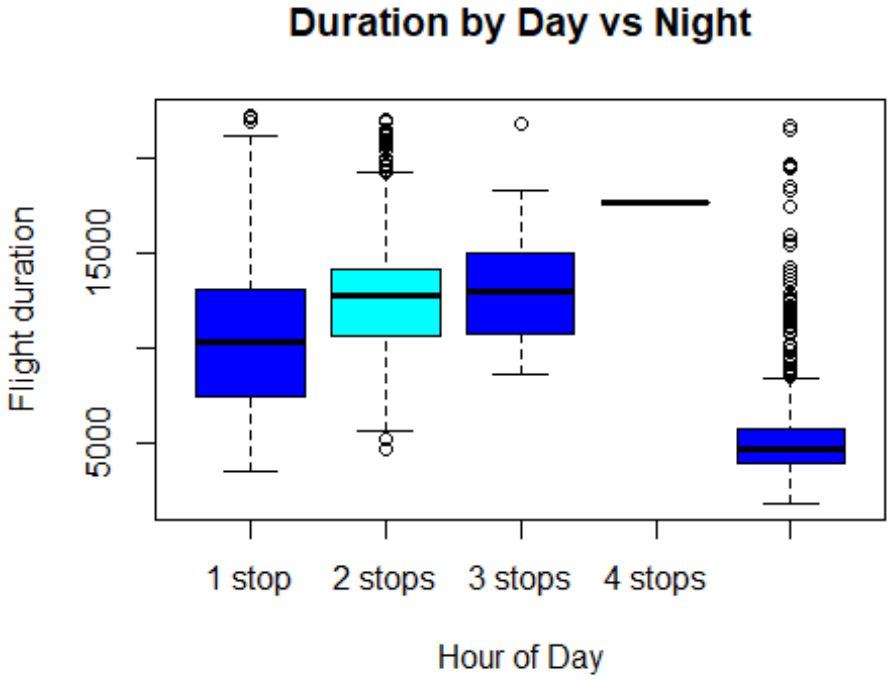

Duration in Mins
Price
0
5000
10000
15000
20000

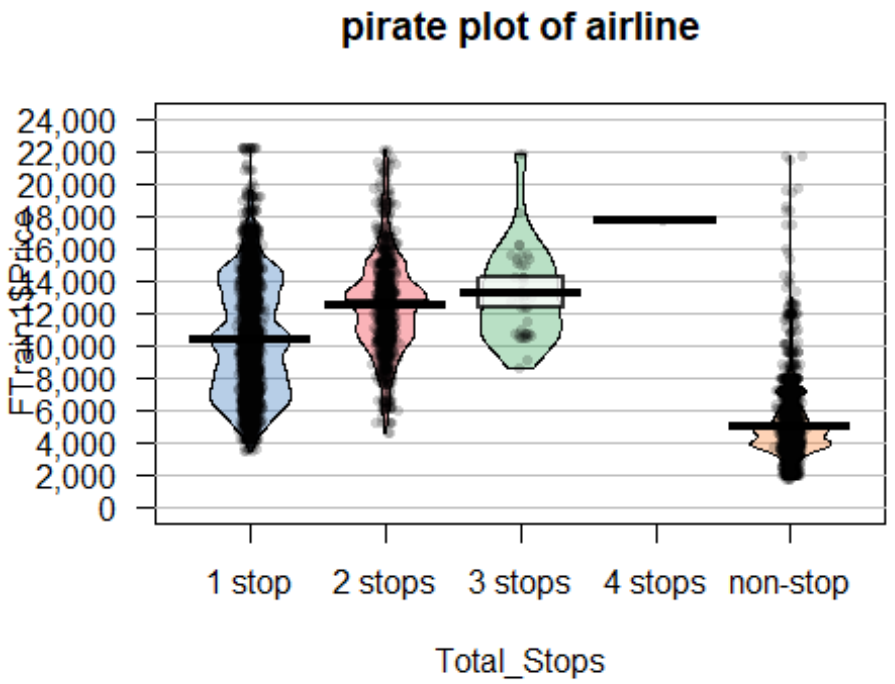

Price by Day of week
Flight price
5000
15000
Friday
Saturday
Thursday
Wednesday
Day of Week

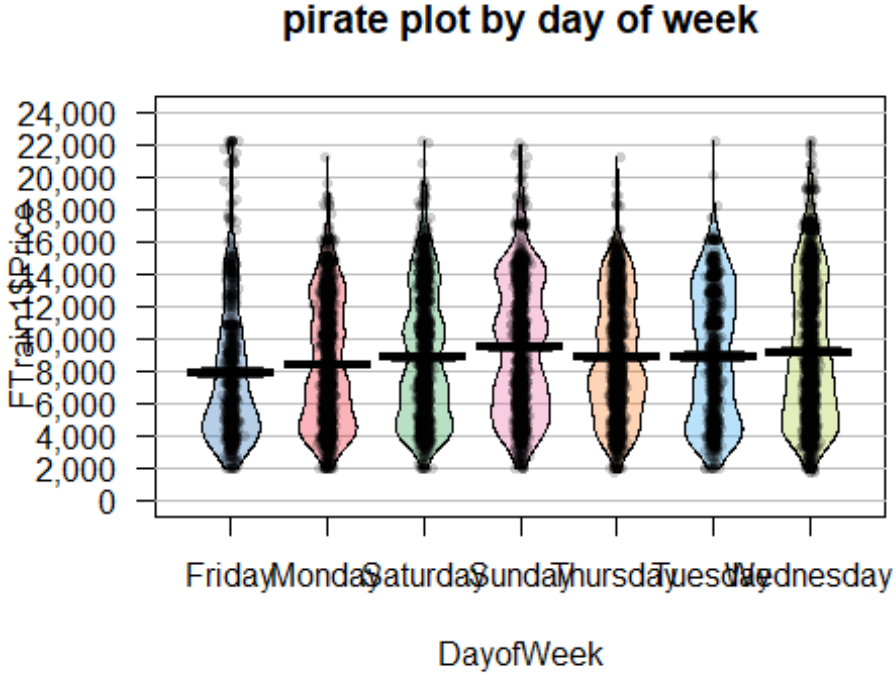

Price by Day vs Night
Flight price
5000
15000
Day Time
Night Time
Hour of Day


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

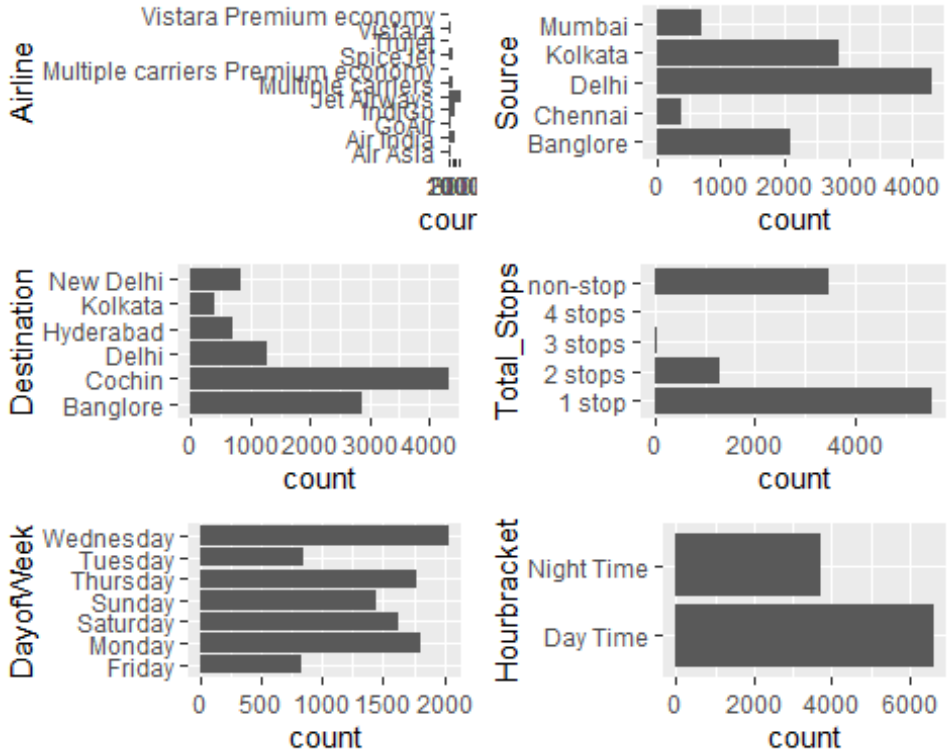

Duration by Day vs Night
Flight duration
5000
15000
1 stop
2 stops
3 stops
4 stops
Hour of Day

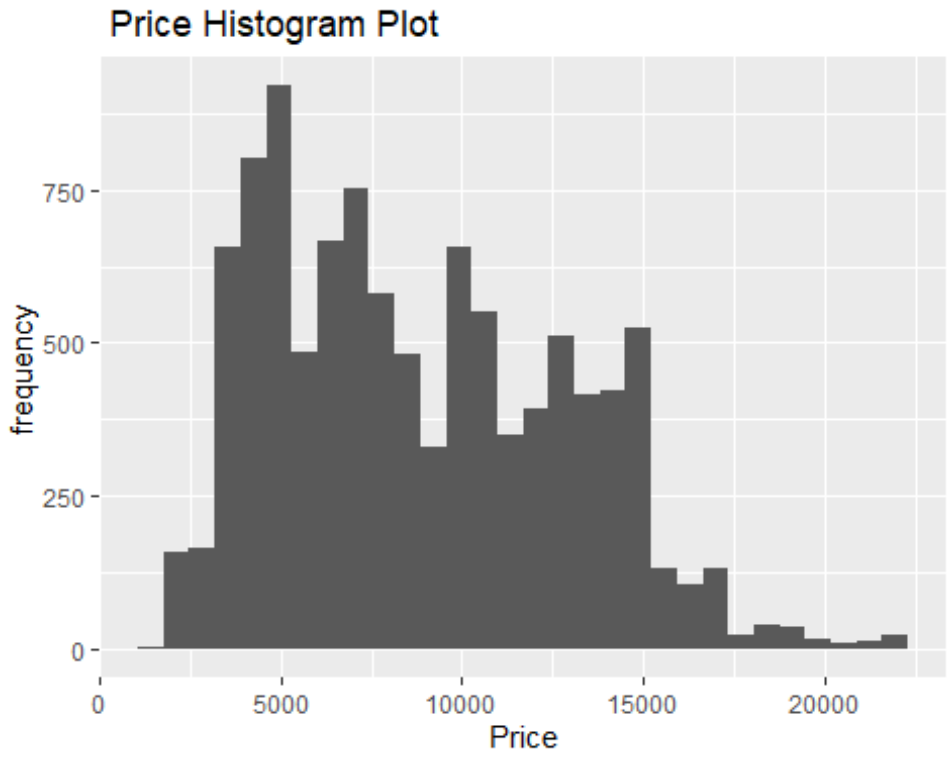


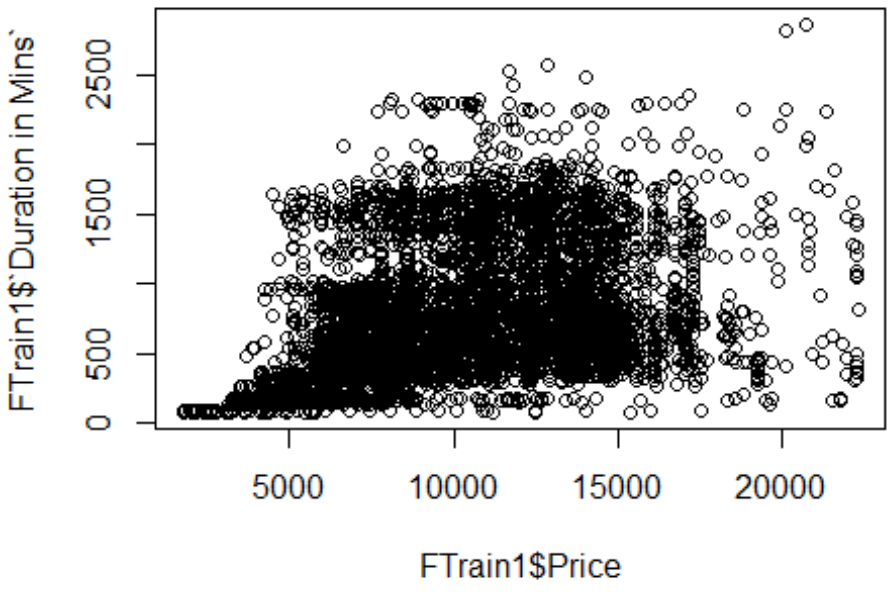


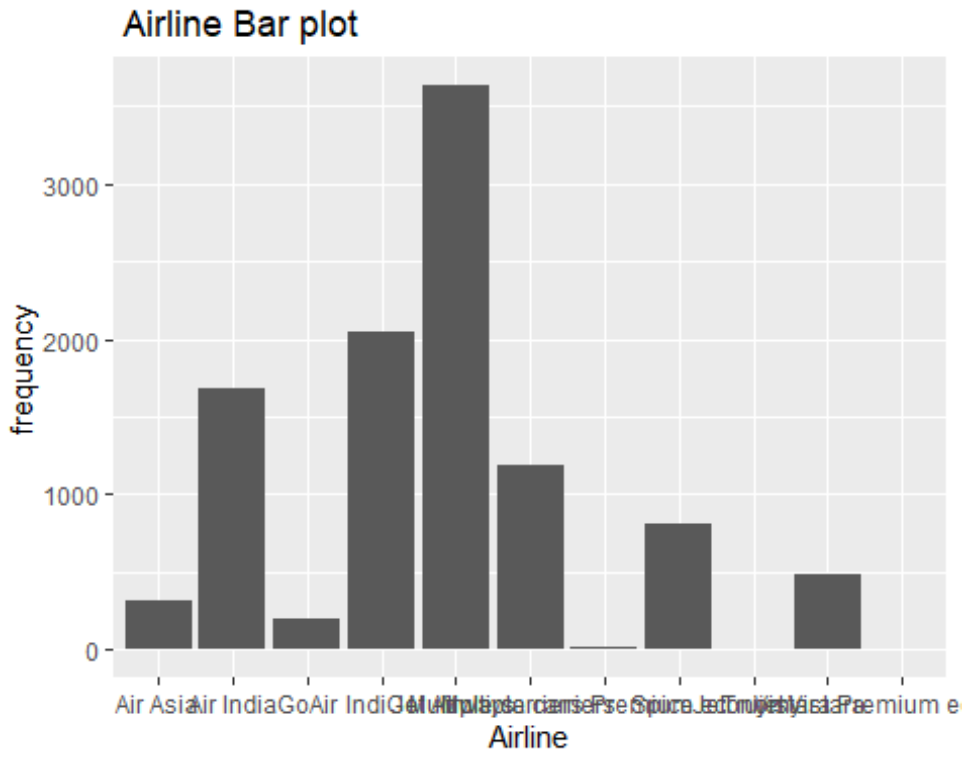

```
# Hypothesis testing

# when we do hypothesis testing for the flight price to see if there is a sig
```

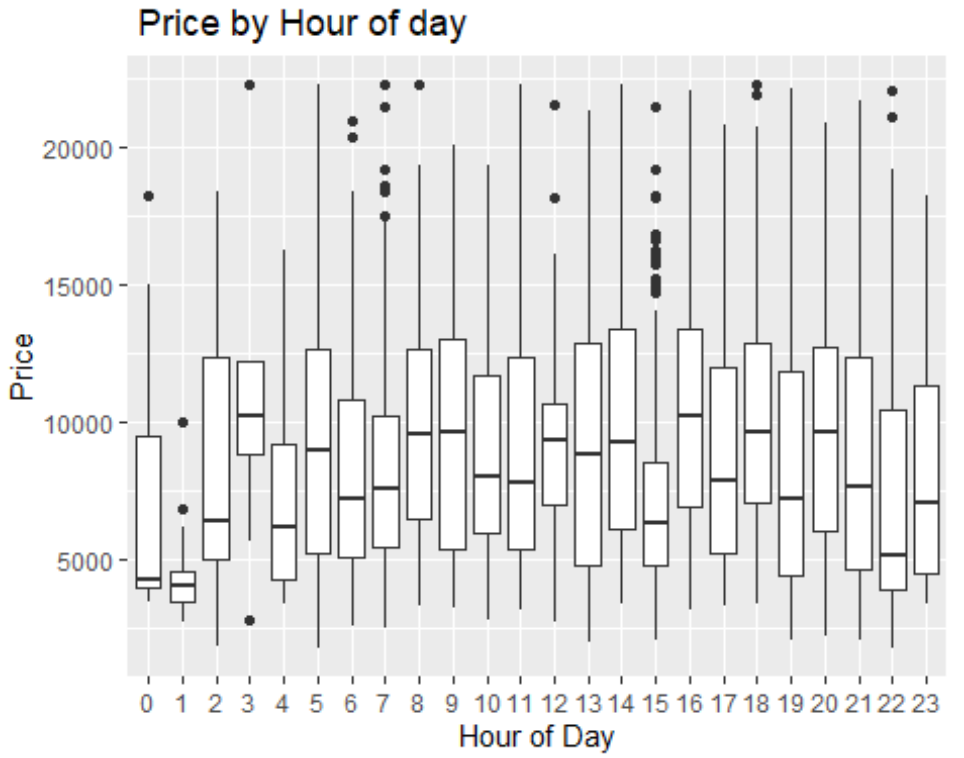
```
nificant different of flight price in day time versus night time, we found
# P value of less than 0.05 which means we reject the null hypothesis and acc
ept alternative that there is significant price difference in flight during d
ay time versus night time
```

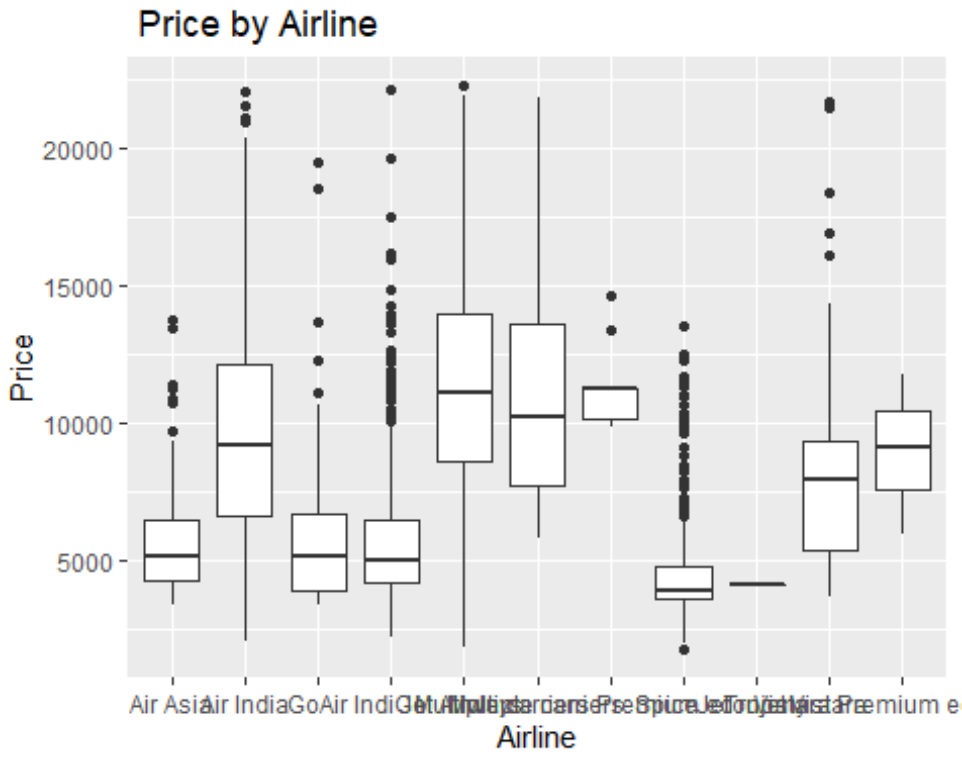


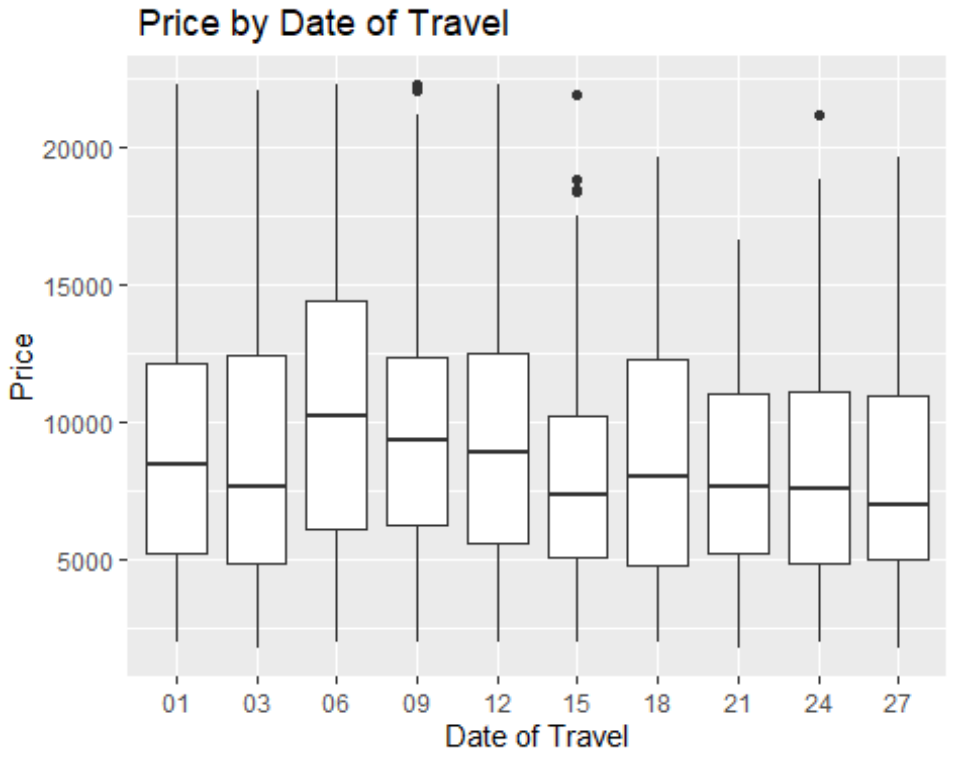

Price Histogram Plot
frequency
750
500
250
0
0
5000
10000
15000
20000
Price

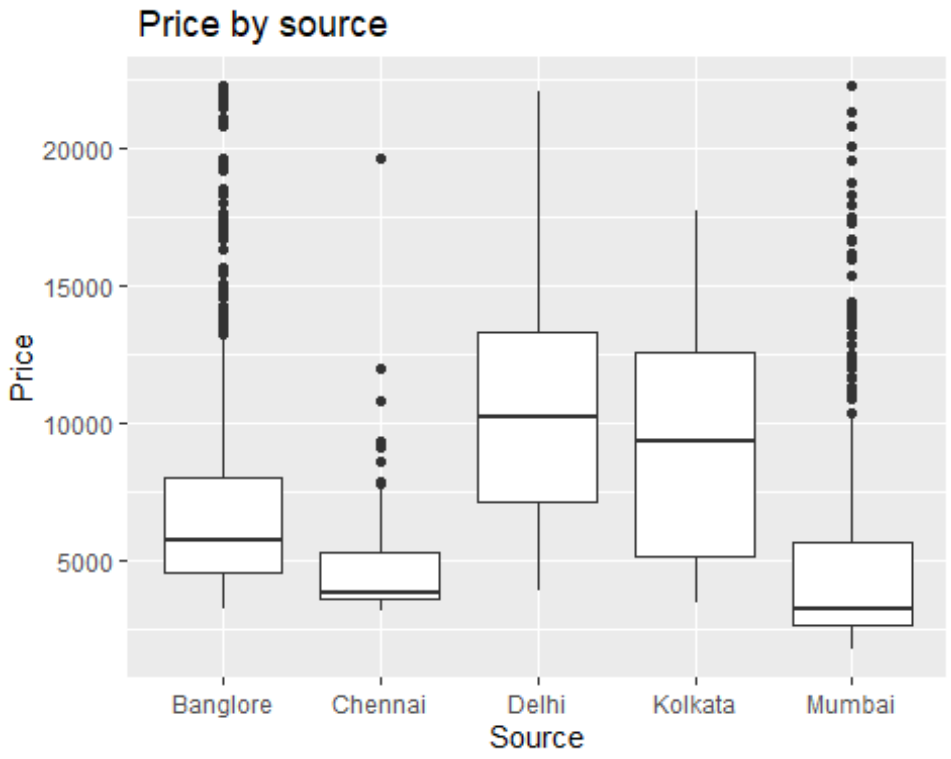

FTrain1$`Duration in Mins`
0
500
1500
2500
5000
10000
15000
20000
FTrain1$Price

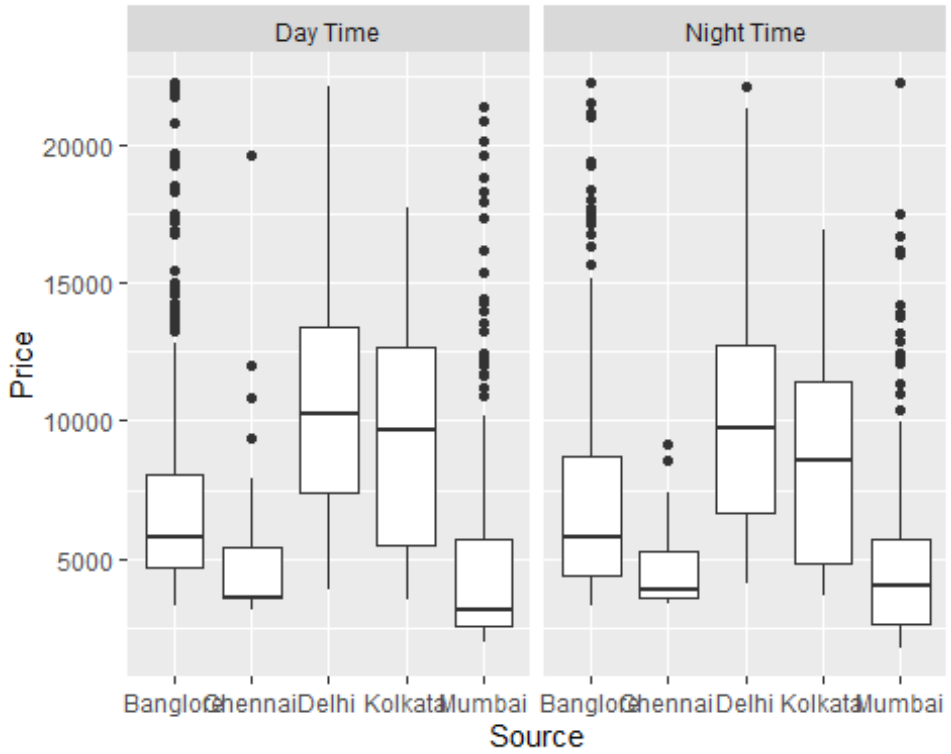

Airline Bar plot
frequency
3000
2000
1000
0
Air Asia
Air India
GoAir
Airline

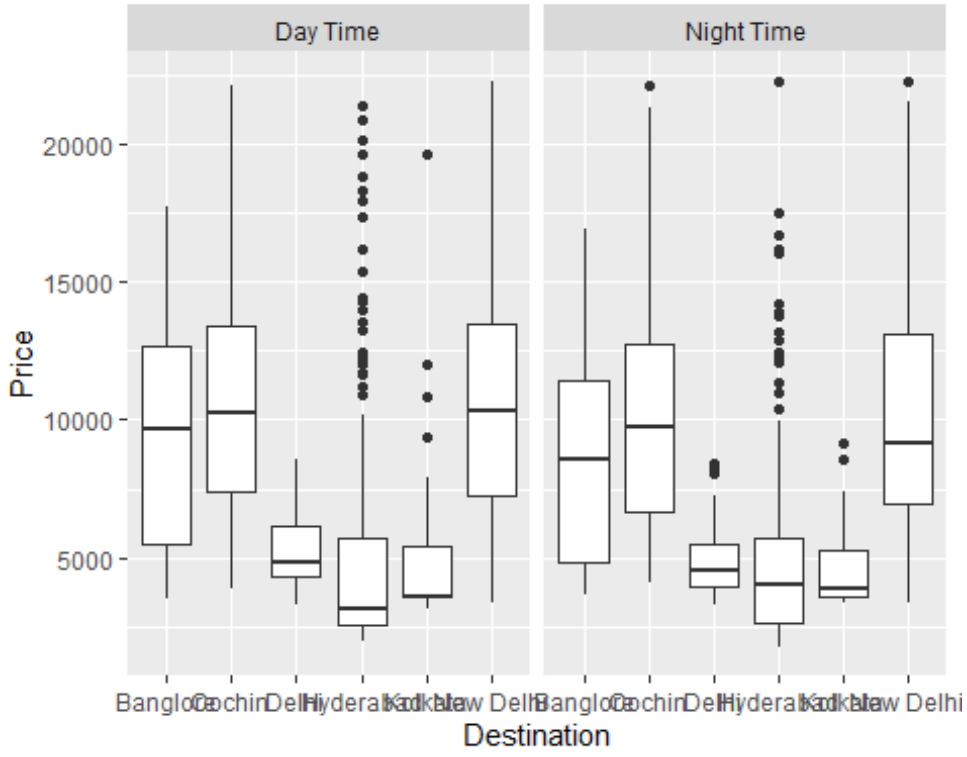

Price by Hour of day
Price
20000
15000
10000
5000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Hour of Day

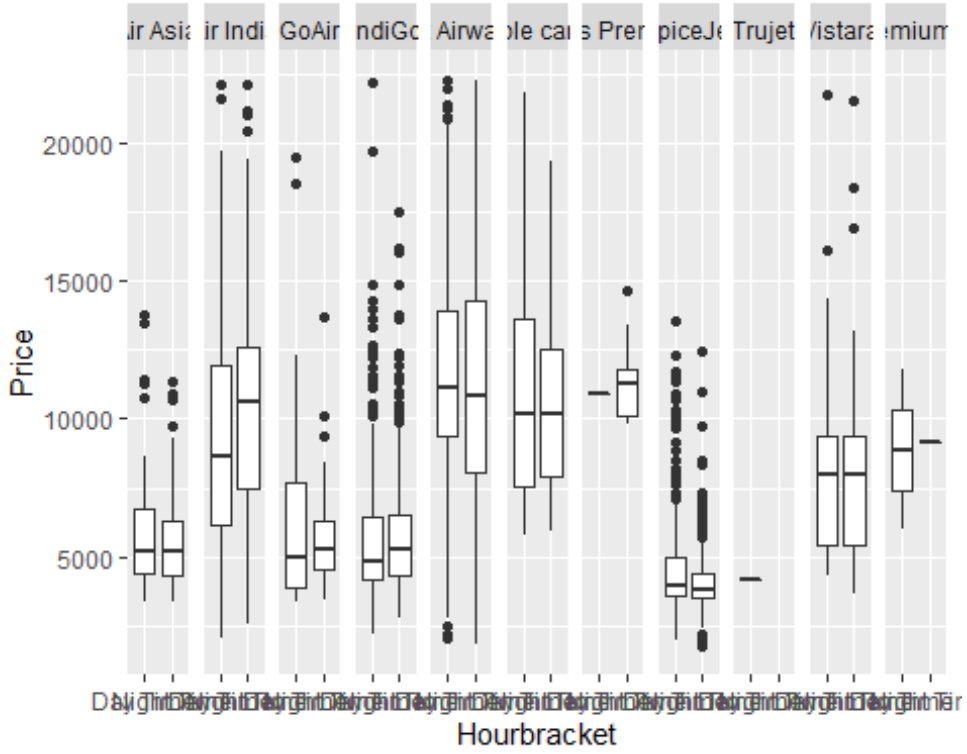

Price by Airline
20000
15000
10000
5000
Price
Air Asia
Air India
GoAir
Airline

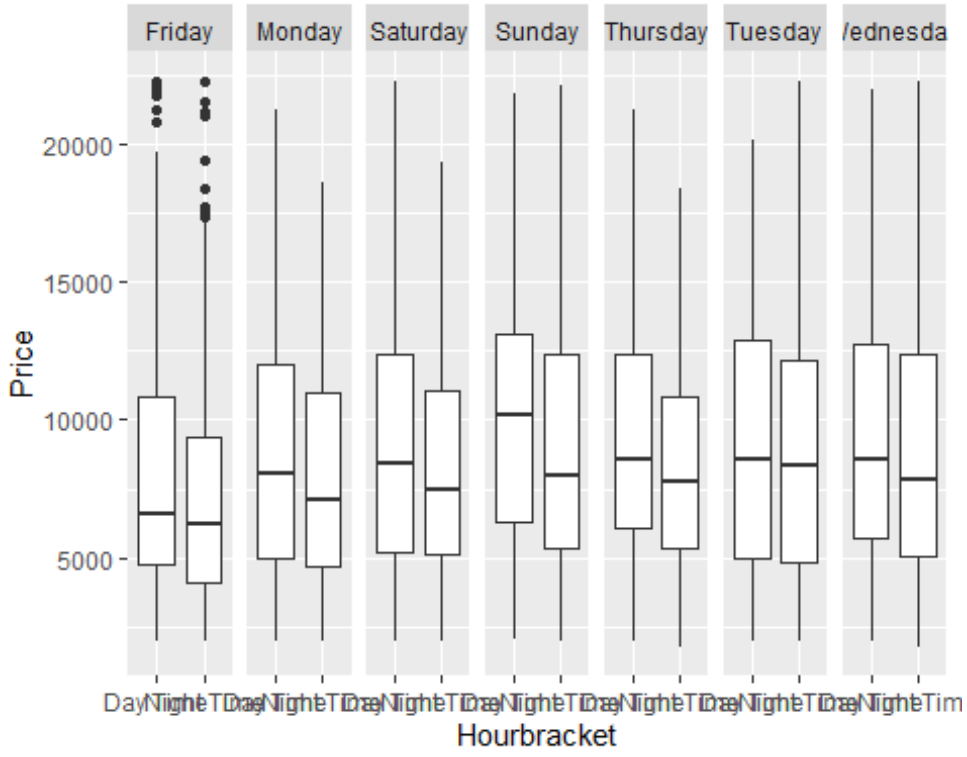

Price by Date of Travel
20000
15000
10000
5000
Price
01
03
06
09
12
15
18
21
24
27
Date of Travel

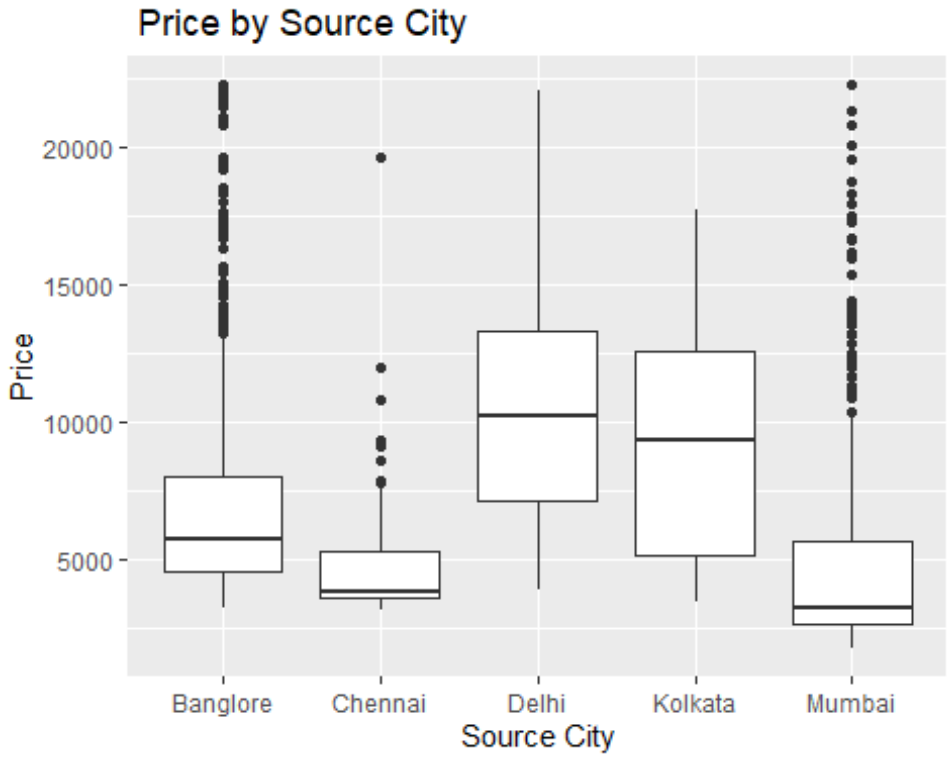

Price by source
20000
15000
10000
5000
Price
Banglore
Chennai
Delhi
Kolkata
Mumbai
Source

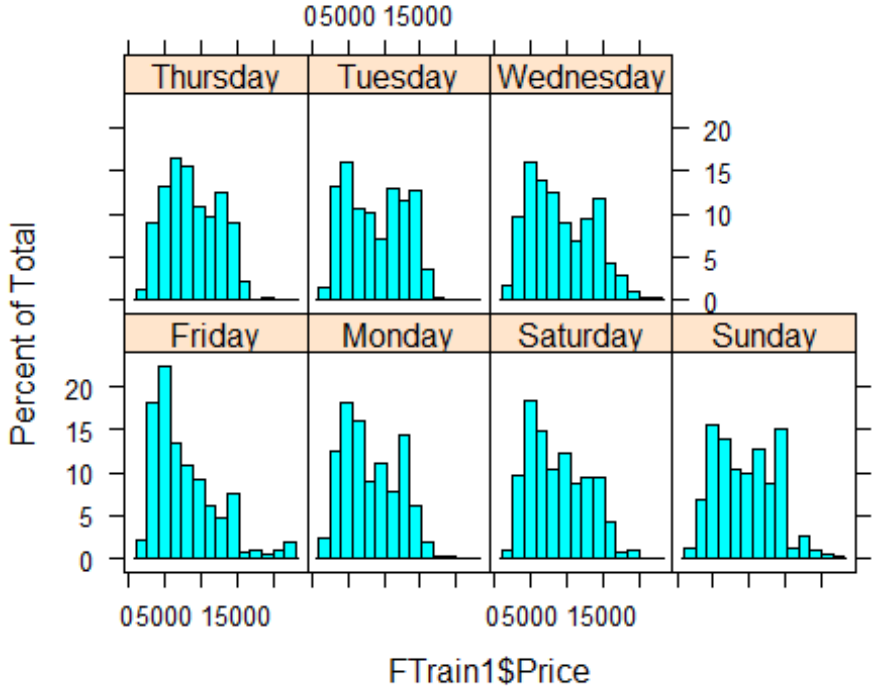

Day Time
Night Time
20000
15000
10000
5000
Price
Banglore
Chennai
Delhi
Kolkata
Mumbai
Source

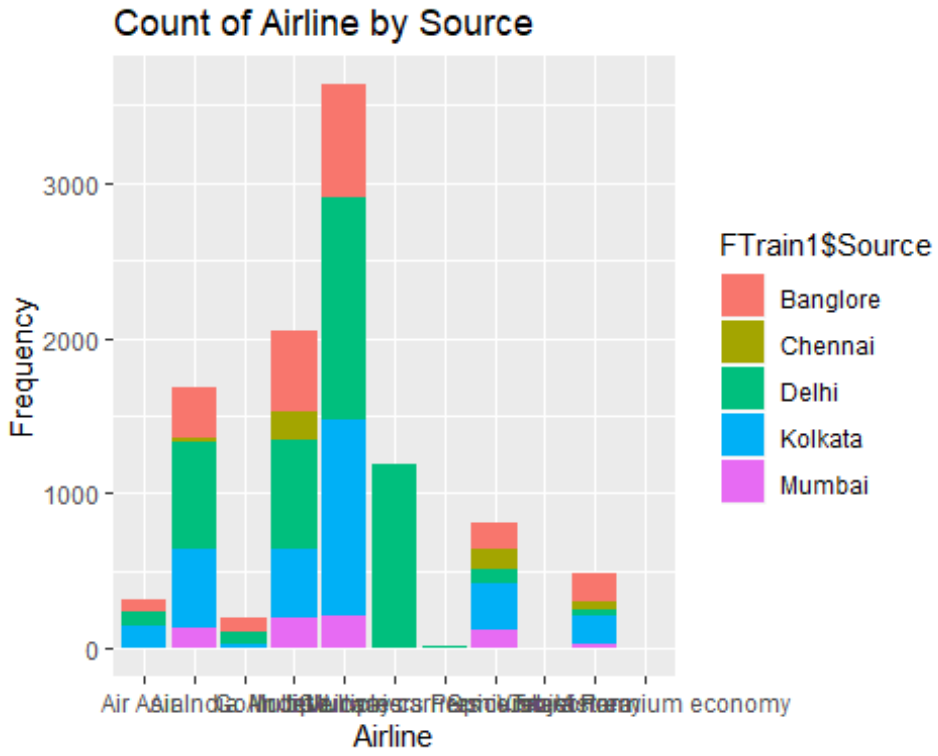

Day Time
Night Time
20000
15000
10000
5000
Price
Destination

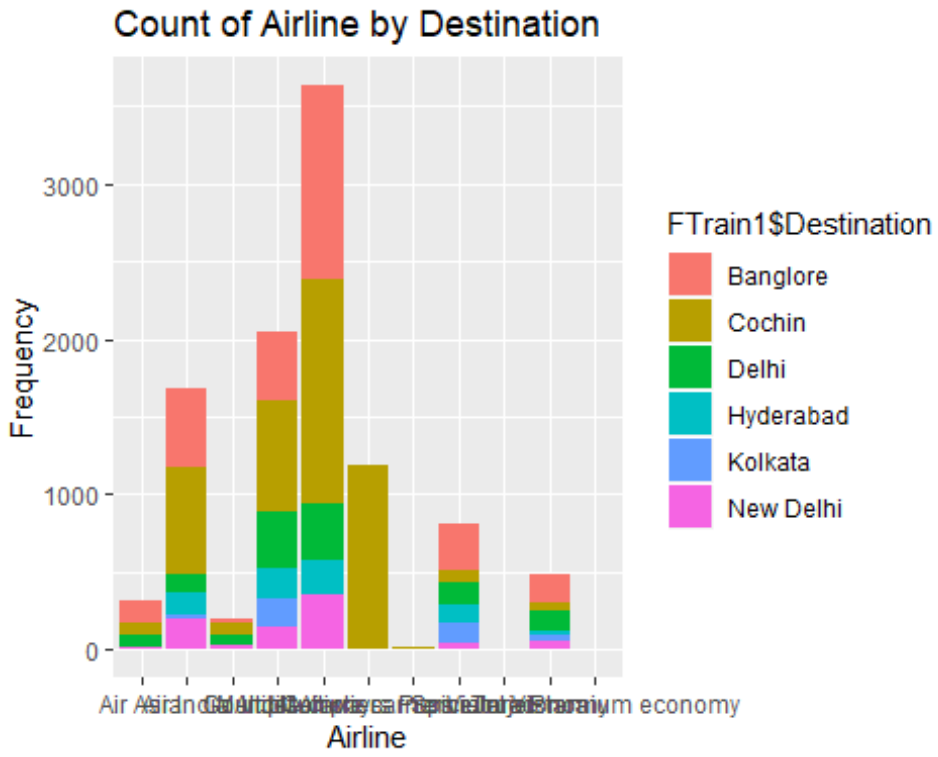

20000
15000
10000
5000
Price
Hourbracket

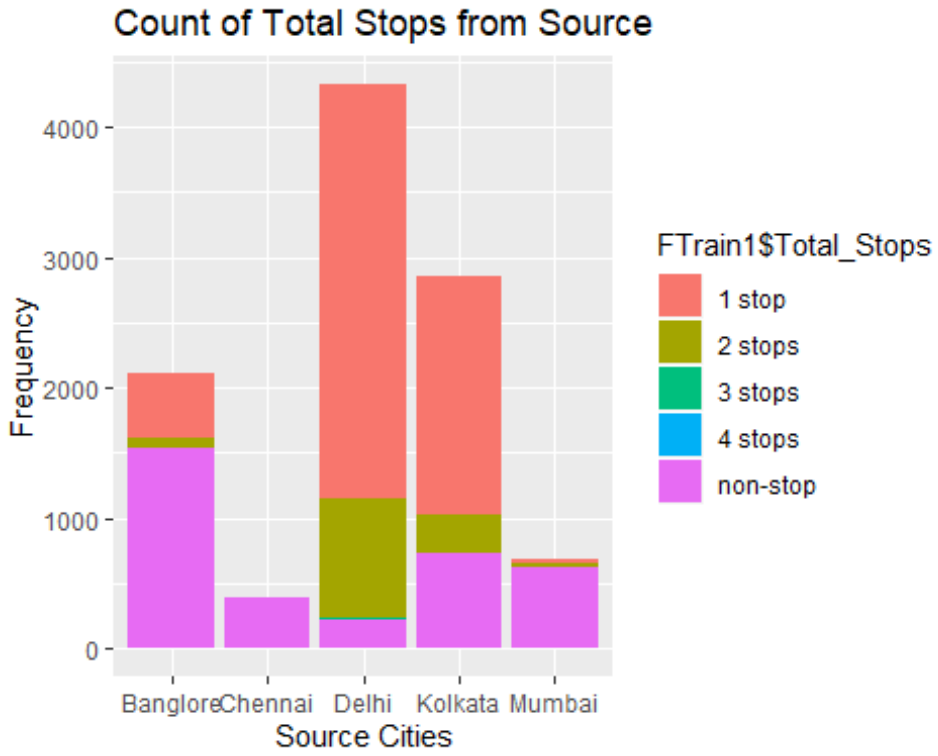

Friday
Monday
Saturday
Sunday
Thursday
Tuesday
/ednesda
20000
15000
10000
5000
Price
Hourbracket

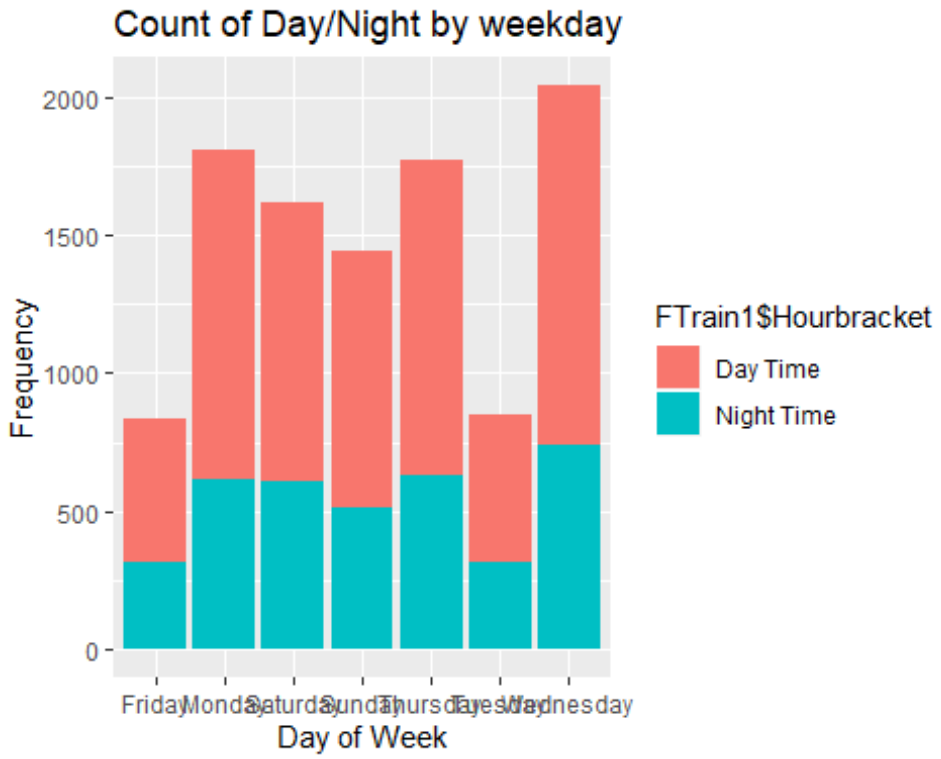

Price by Source City
20000
15000
10000
5000
Price
Banglore
Chennai
Delhi
Kolkata
Mumbai
Source City

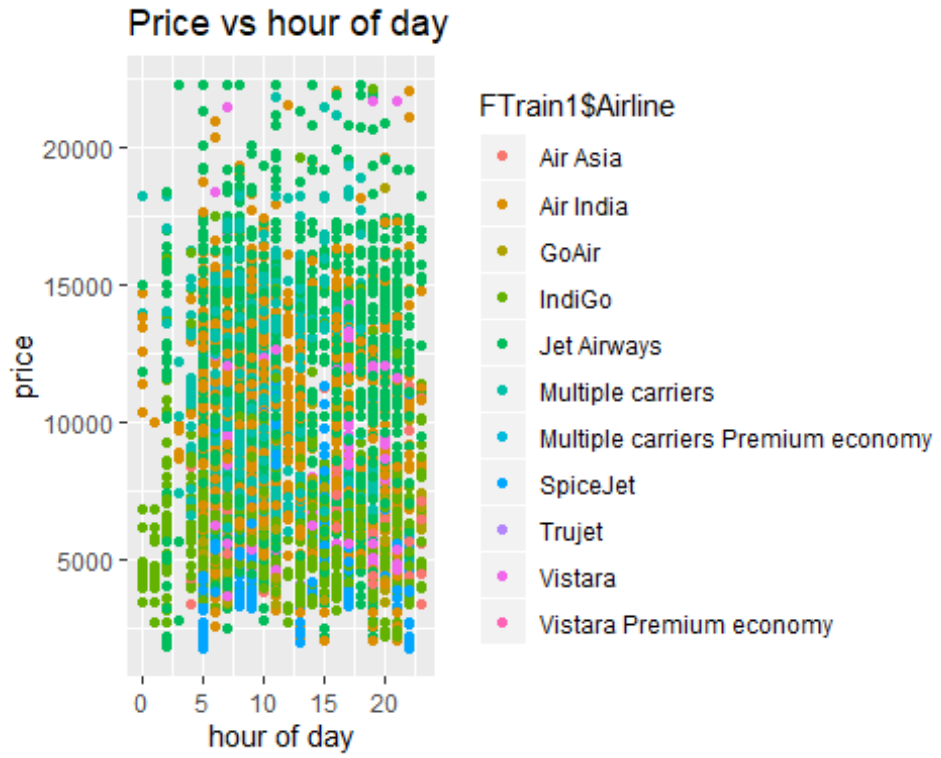

05000 15000
Thursday
Tuesday
Wednesday
Friday
Monday
Saturday
Sunday
Percent of Total
0
5
10
15
20
05000 15000
05000 15000
FTrain1$Price

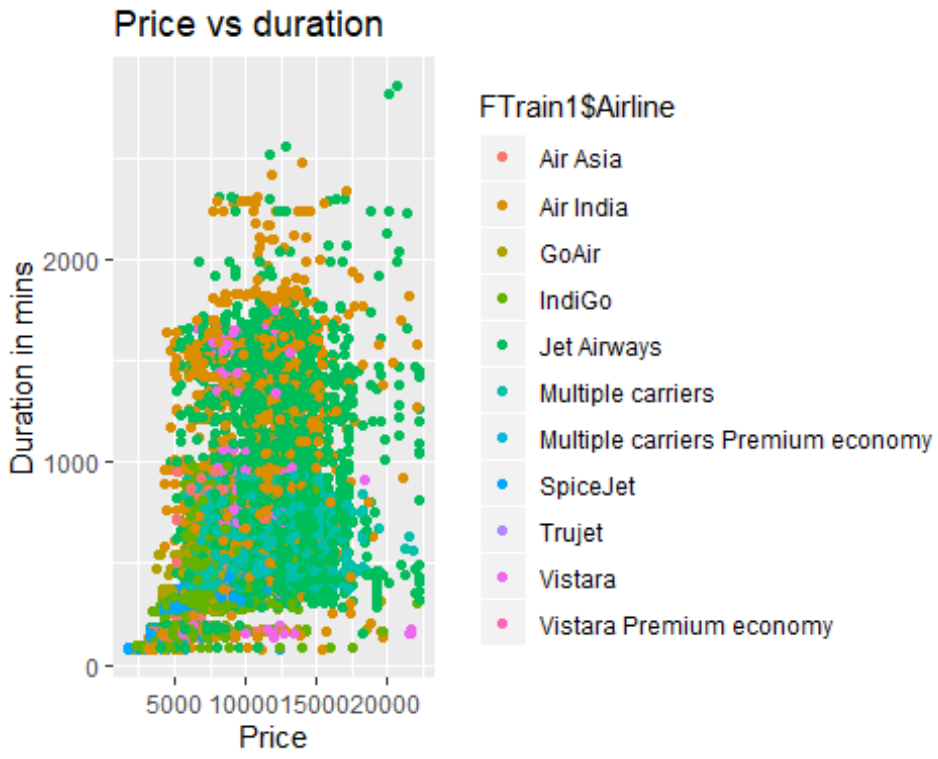

Count of Airline by Source
Frequency
0
1000
2000
3000
FTrain1$Source
Banglore
Chennai
Delhi
Kolkata
Mumbai
Airline

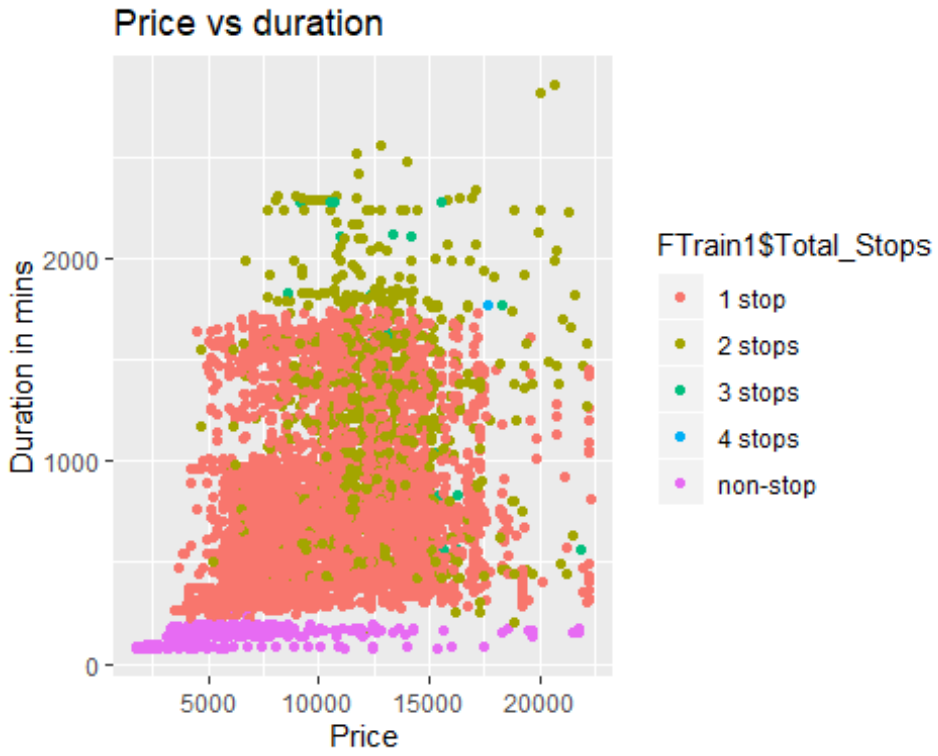

Count of Airline by Destination
Frequency
3000
2000
1000
0
Airline
FTrain1$Destination
Banglore
Cochin
Delhi
Hyderabad
Kolkata
New Delhi

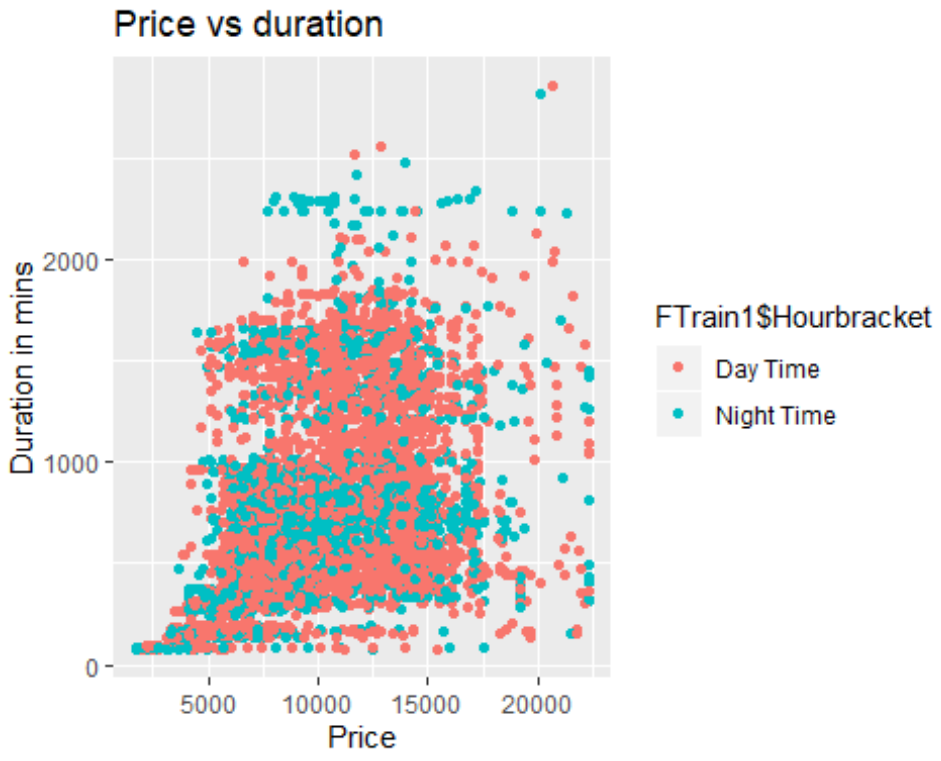

Count of Total Stops from Source
Frequency
4000
3000
2000
1000
0
Banglore
Chennai
Delhi
Kolkata
Mumbai
Source Cities
FTrain1$Total_Stops
1 stop
2 stops
3 stops
4 stops
non-stop

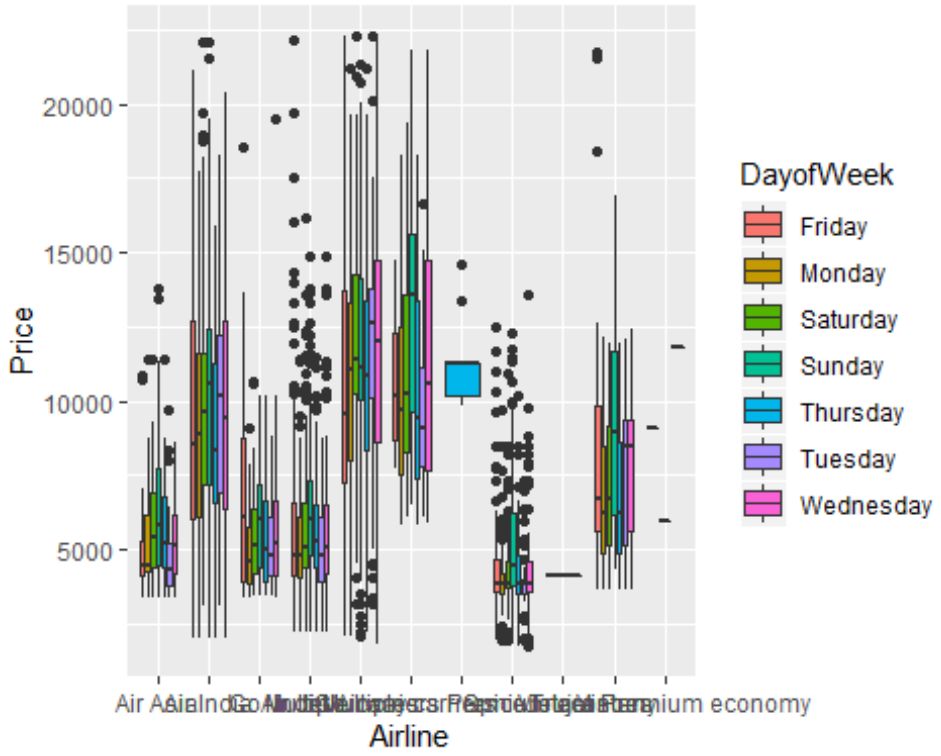

Count of Day/Night by weekday
Frequency
2000
1500
1000
500
0
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday
Day of Week
FTrain1$Hourbracket
Day Time
Night Time

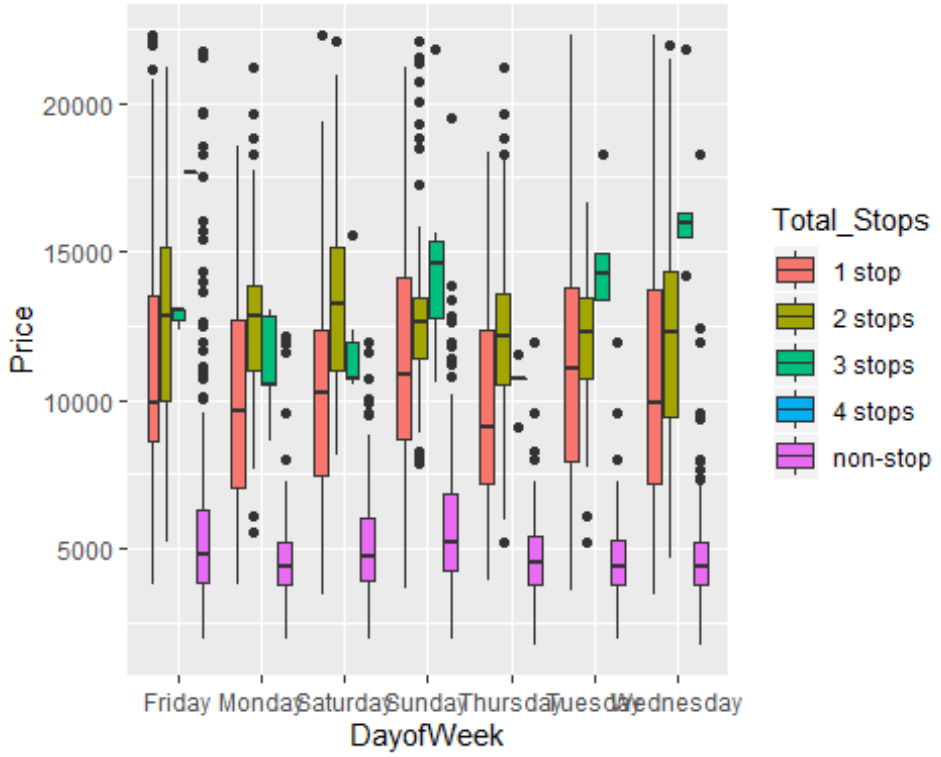

Price vs hour of day
price
20000
15000
10000
5000
0
5
10
15
20
hour of day
FTrain1$Airline
Air Asia
Air India
GoAir
IndiGo
Jet Airways
Multiple carriers
Multiple carriers Premium economy
SpiceJet
Trujet
Vistara
Vistara Premium economy

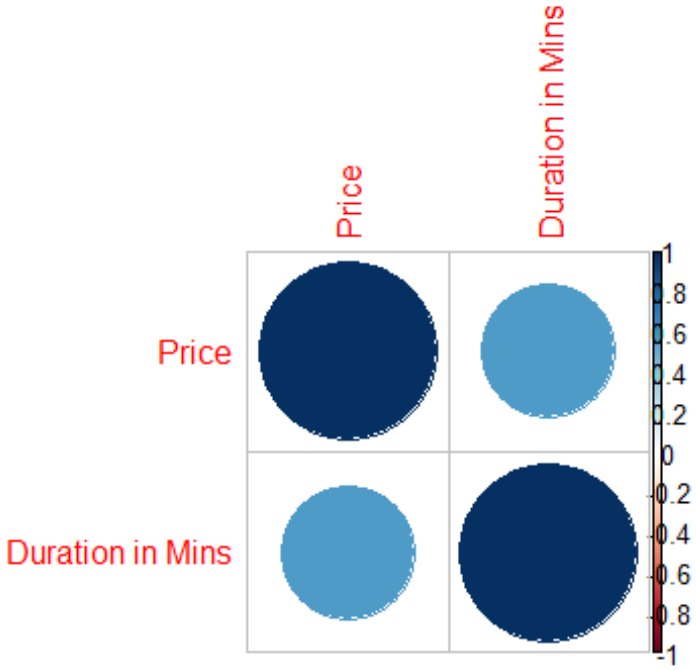

Price vs duration
Duration in mins
2000
1000
0
5000 10000 15000 20000
Price
FTrain1$Airline
Air Asia
Air India
GoAir
IndiGo
Jet Airways
Multiple carriers
Multiple carriers Premium economy
SpiceJet
Trujet
Vistara
Vistara Premium economy


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


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


Price
20000
15000
10000
5000
Air
Airline
DayofWeek
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday






```
##                      Price Duration in Mins
## Price            1.0000000        0.5684513
## Duration in Mins 0.5684513        1.0000000
```