Capstone Project

Flight cost prediction

Notes-1

K.Vineet *PATNAIK*

Table of contents

1. Introduction

The problem statement is to predict the cost of the flight at various time periods for different locations for different times of the day for different airlines.For this to analyse furtherly a dataset is given with different details such as Flight source,destination,route details, stops in between,arrival time etc. along with some additional information .

The dataset consists of 4 months of data and the respective prices of each flight in it.The main need to understand the data is to determine the fair price of the flight .Some of the objectives are :

1. Identify Important variables .
2. Establish a relationship b/w time of journey and flight prices .

2. Data Report

The data was taken from the month of March to July.some of the misprinted data was inspected in the excel sheet and some mistakes were corrected.Also Delhi and New delhi are considered as delhi

Air Asia	Air India	GoAir
319	1752	194
IndiGo	Jet Airways	Jet Airways Business
2053	3849	6
Multiple carriers	Multiple carriers	Premium economy
1196	13	SpiceJet
Trujet	Vistara	Vistara Premium economy
1	479	3

There are different variables in the dataset where duration is a continuous variable and remaining are all categorical variables .Basically much to understand how the data is behaving, so will be continued in the further analysis .

3. Exploratory data analysis

```
any(is.na(dataset))
sum(is.na(dataset))
data = na.omit(dataset)
sum(is.na(data))
```

So all the NA values will be removed and a new variable ‘data’ is created .

Variable transformation is done for arrival and destination and Hour and minutes are separated from time .Some unwanted variables are seperated like ‘Additional data’ and ‘number of stops’ and New Delhi was converted to Delhi .All of these changes were directly made in the excel sheet

To do Univariate and bivariate analysis we need to transform the variables for correlation etc.

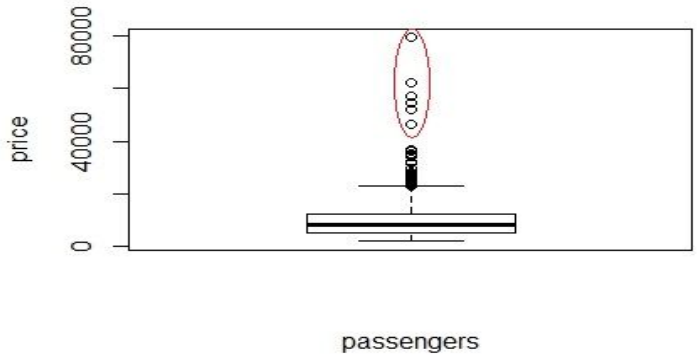
```
data$Airline = as.numeric(data$Airline)
data$Date_of_Journey = as.numeric(data$Date_of_Journey)
data$Source = as.numeric(data$Source)
data$Destination = as.numeric(data$Destination)
data$Route = as.numeric(data$Route)
data$Dep_Time = as.numeric(data$Dep_Time)
data$Arrival_Time = as.numeric(data$Arrival_Time)
data$Duration= as.numeric(data$Duration)
data$hourD = as.numeric(data$hourD)
data$hourA = as.numeric(data$hourA)
data$minuteA = as.numeric(data$minuteA)
data$minuteD = as.numeric(data$minuteD)
```

Missing outliers treatment

Some of the high prices here does not help us in predicting the accurate value of the prices but stay as outliers,so they can be removed

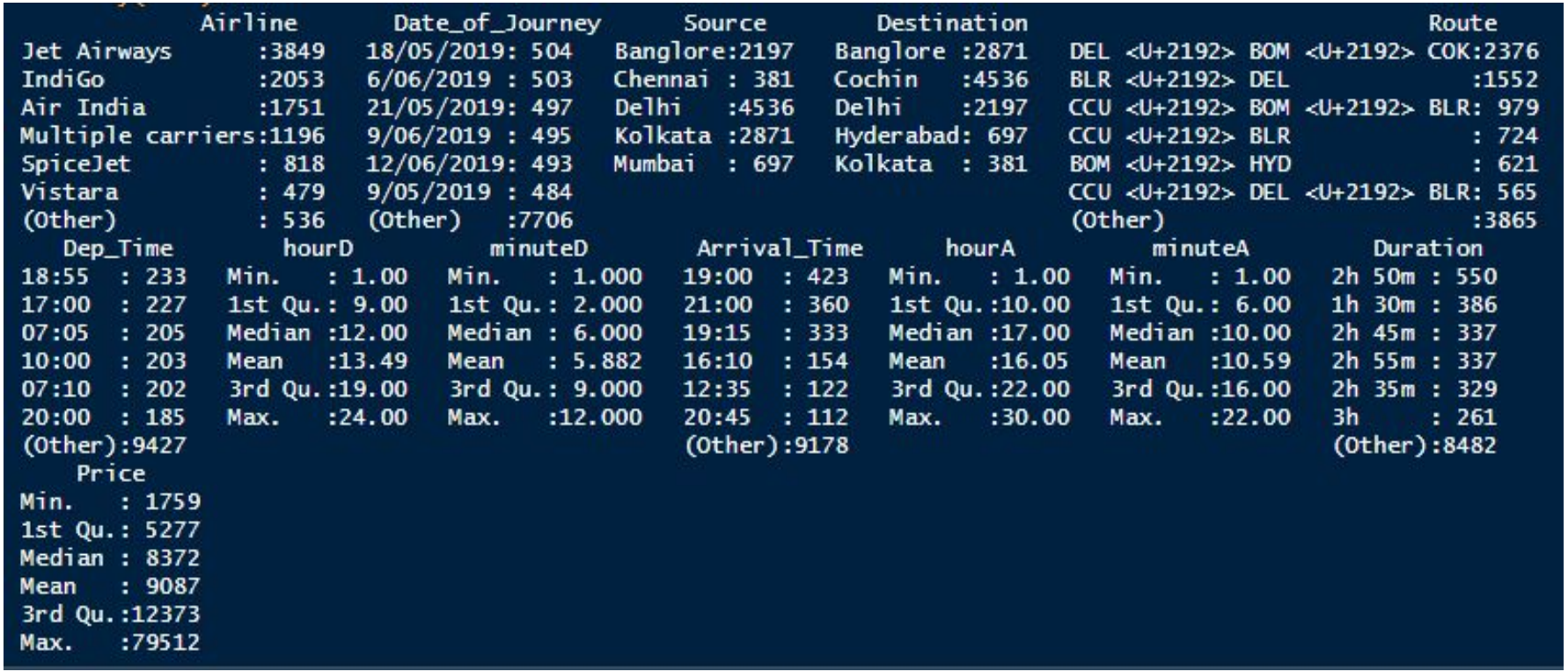
```
boxplot(Price,ylab = "price",xlab = "passengers")
quantile(Price,probs = seq(0,1,0.05),
na.rm = FALSE)

data[data$Price < quantile(data$Price, 0.95), ]
data[data$Price > quantile(data$Price, 0.05), ]
```



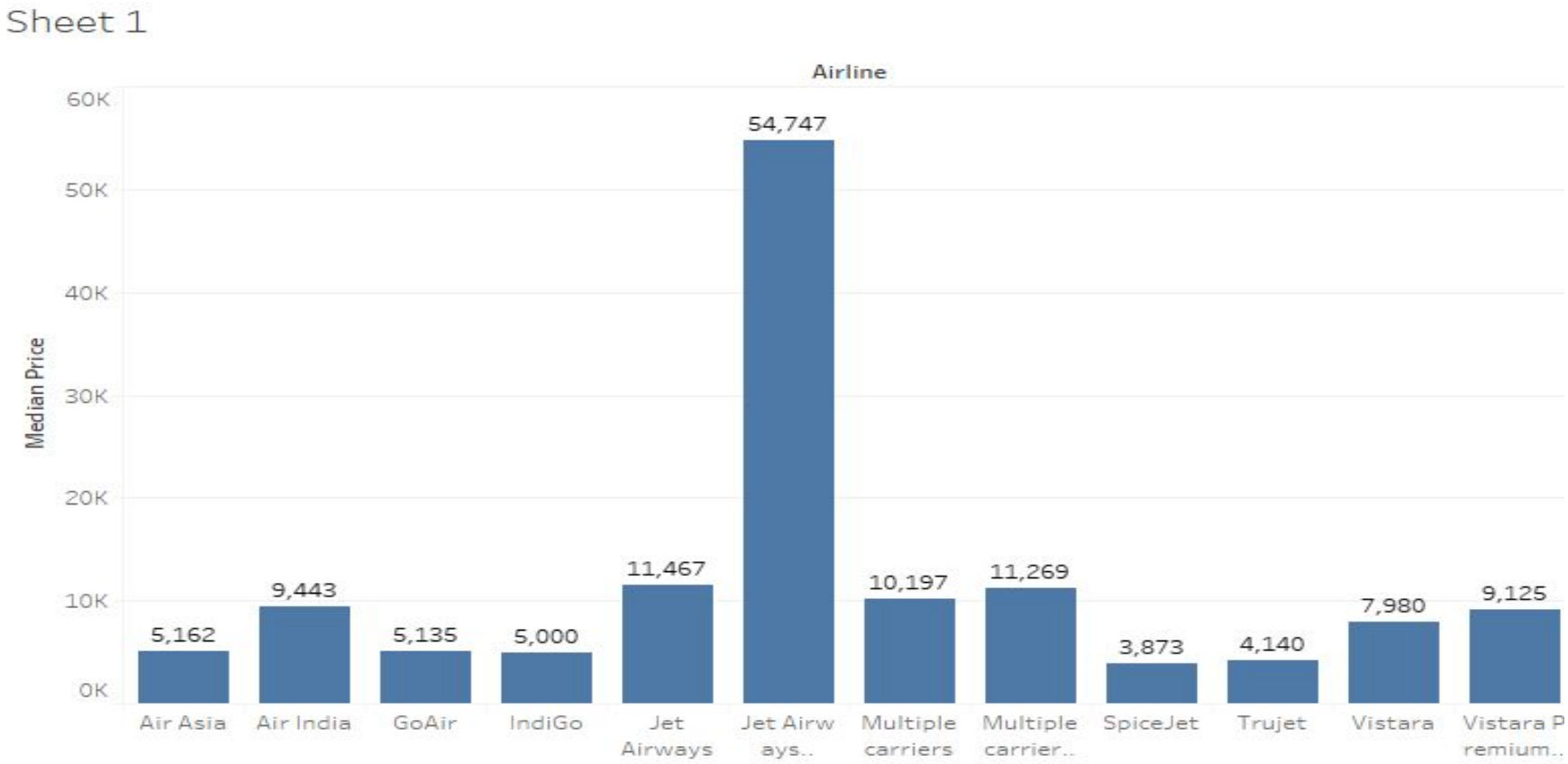
We contain the first and last 5 percentile of the data so that it becomes a balanced data and dealing with it is now better .

4.Univariate & Bivariate Analysis

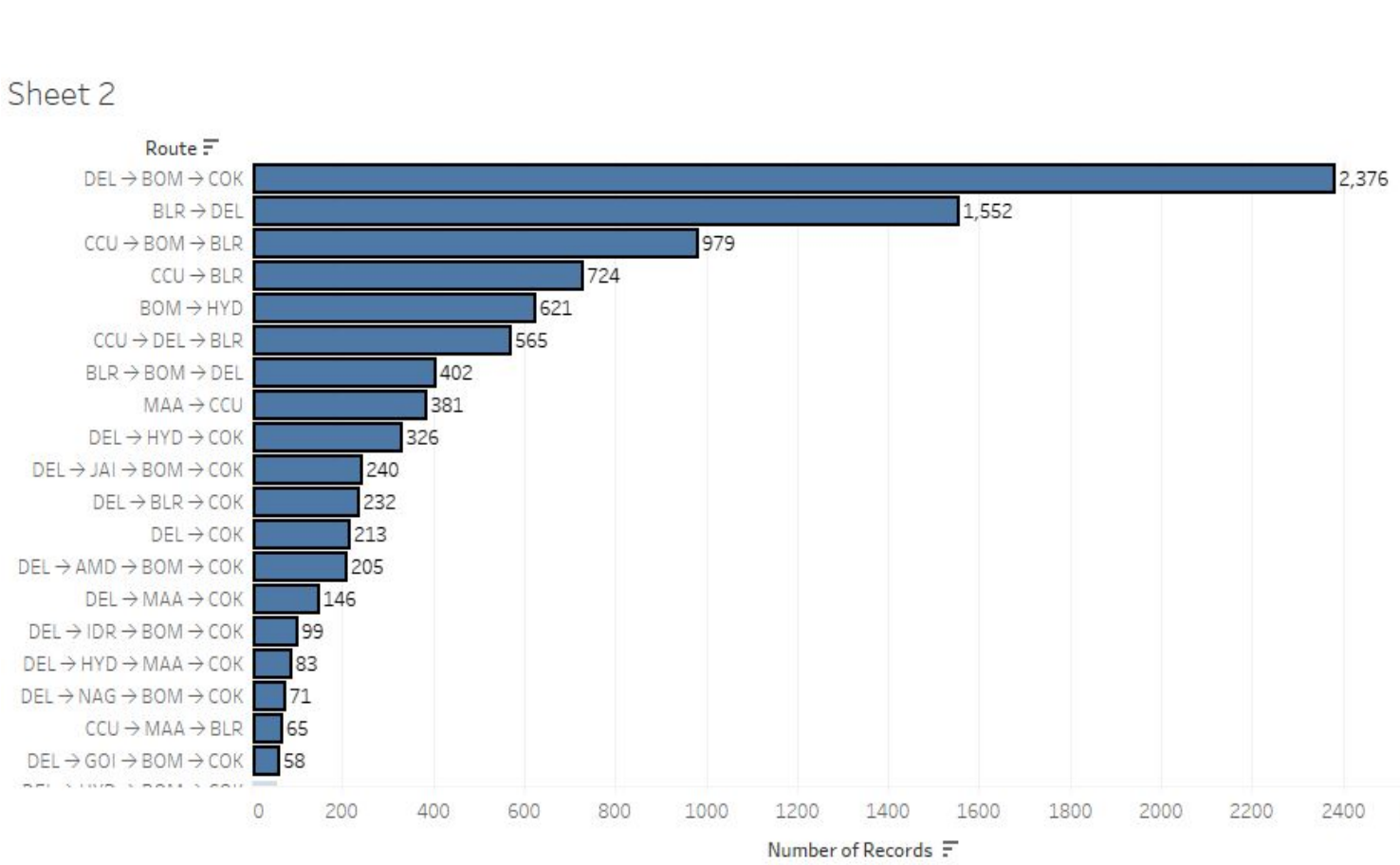


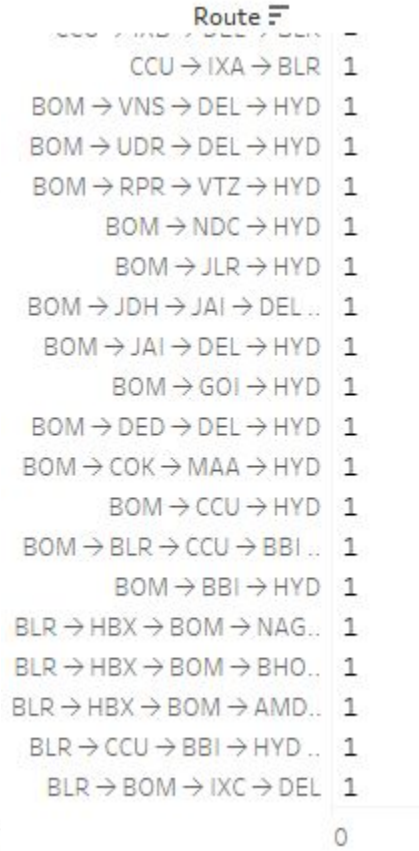
```
                Airline       Date_of_Journey      Source       Destination                                 Route
Jet Airways         :3849   18/05/2019: 504   Banglore:2197   Banglore :2871   DEL <U+2192> BOM <U+2192> COK:2376
IndiGo              :2053   6/06/2019 : 503   Chennai : 381   Cochin   :4536   BLR <U+2192> DEL             :1552
Air India           :1751   21/05/2019: 497   Delhi   :4536   Delhi    :2197   CCU <U+2192> BOM <U+2192> BLR: 979
Multiple carriers:1196   9/06/2019 : 495   Kolkata :2871   Hyderabad: 697   CCU <U+2192> BLR             : 724
SpiceJet            : 818   12/06/2019: 493   Mumbai  : 697   Kolkata  : 381   BOM <U+2192> HYD             : 621
Vistara             : 479   9/05/2019 : 484                                    CCU <U+2192> DEL <U+2192> BLR: 565
(Other)             : 536   (Other)   :7706                                    (Other)                      :3865
   Dep_Time        hourD          minuteD        Arrival_Time     hourA          minuteA        Duration
18:55  : 233   Min.   : 1.00   Min.   : 1.000   19:00  : 423   Min.   : 1.00   Min.   : 1.00   2h 50m : 550
17:00  : 227   1st Qu.: 9.00   1st Qu.: 2.000   21:00  : 360   1st Qu.:10.00   1st Qu.: 6.00   1h 30m : 386
07:05  : 205   Median :12.00   Median : 6.000   19:15  : 333   Median :17.00   Median :10.00   2h 45m : 337
10:00  : 203   Mean   :13.49   Mean   : 5.882   16:10  : 154   Mean   :16.05   Mean   :10.59   2h 55m : 337
07:10  : 202   3rd Qu.:19.00   3rd Qu.: 9.000   12:35  : 122   3rd Qu.:22.00   3rd Qu.:16.00   2h 35m : 329
20:00  : 185   Max.   :24.00   Max.   :12.000   20:45  : 112   Max.   :30.00   Max.   :22.00   3h     : 261
(Other):9427                                    (Other):9178                                   (Other):8482
    Price
Min.   : 1759
1st Qu.: 5277
Median : 8372
Mean   : 9087
3rd Qu.:12373
Max.   :79512
```

Some of the analysis was done in Tableau shows which airlines passengers travelled most



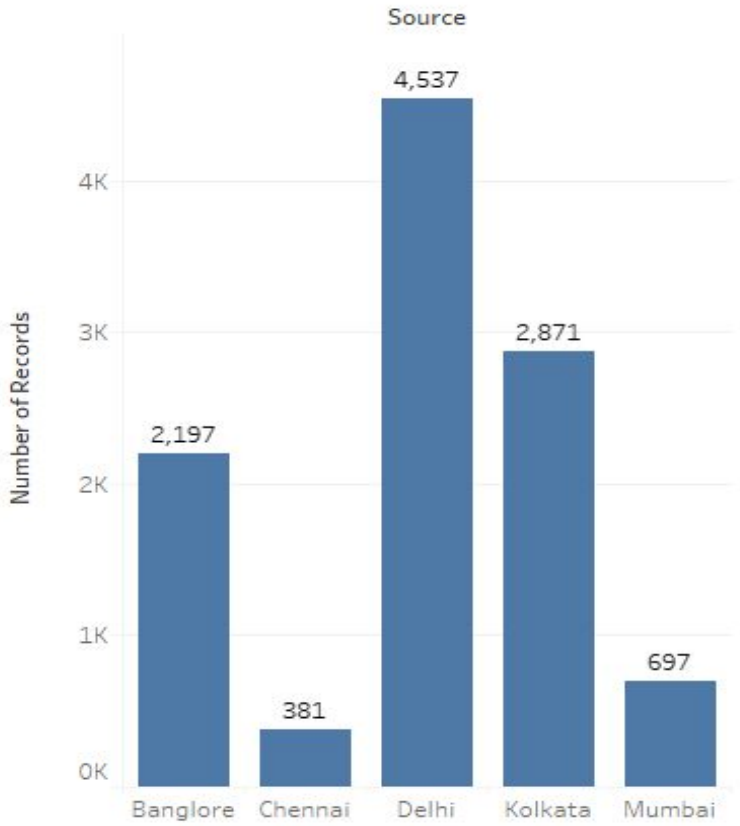
Maximum and minimum via routes taken



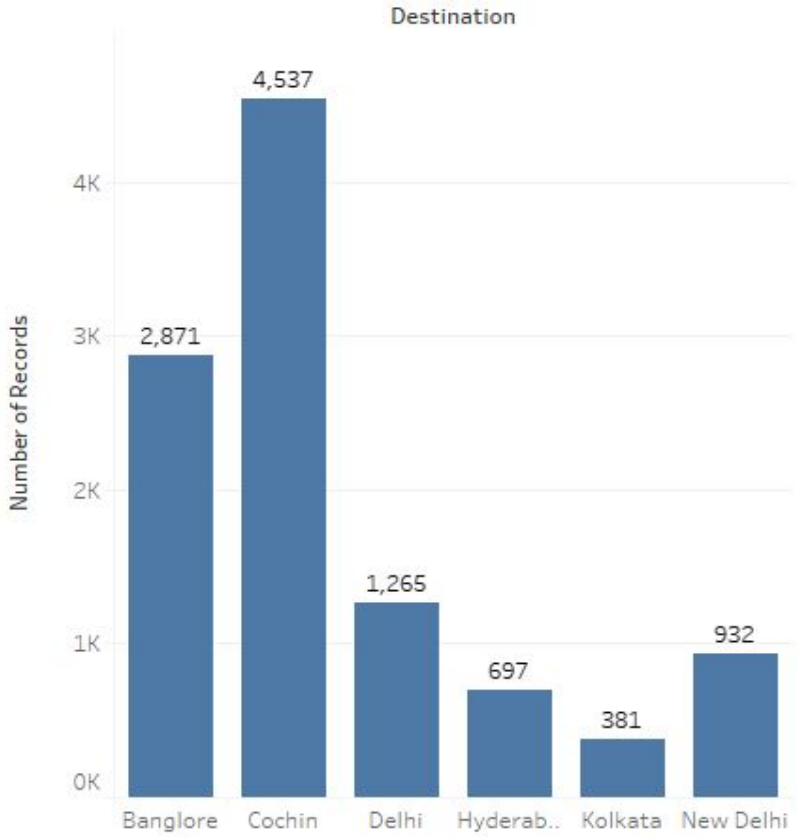


Number of passengers from source and to destination

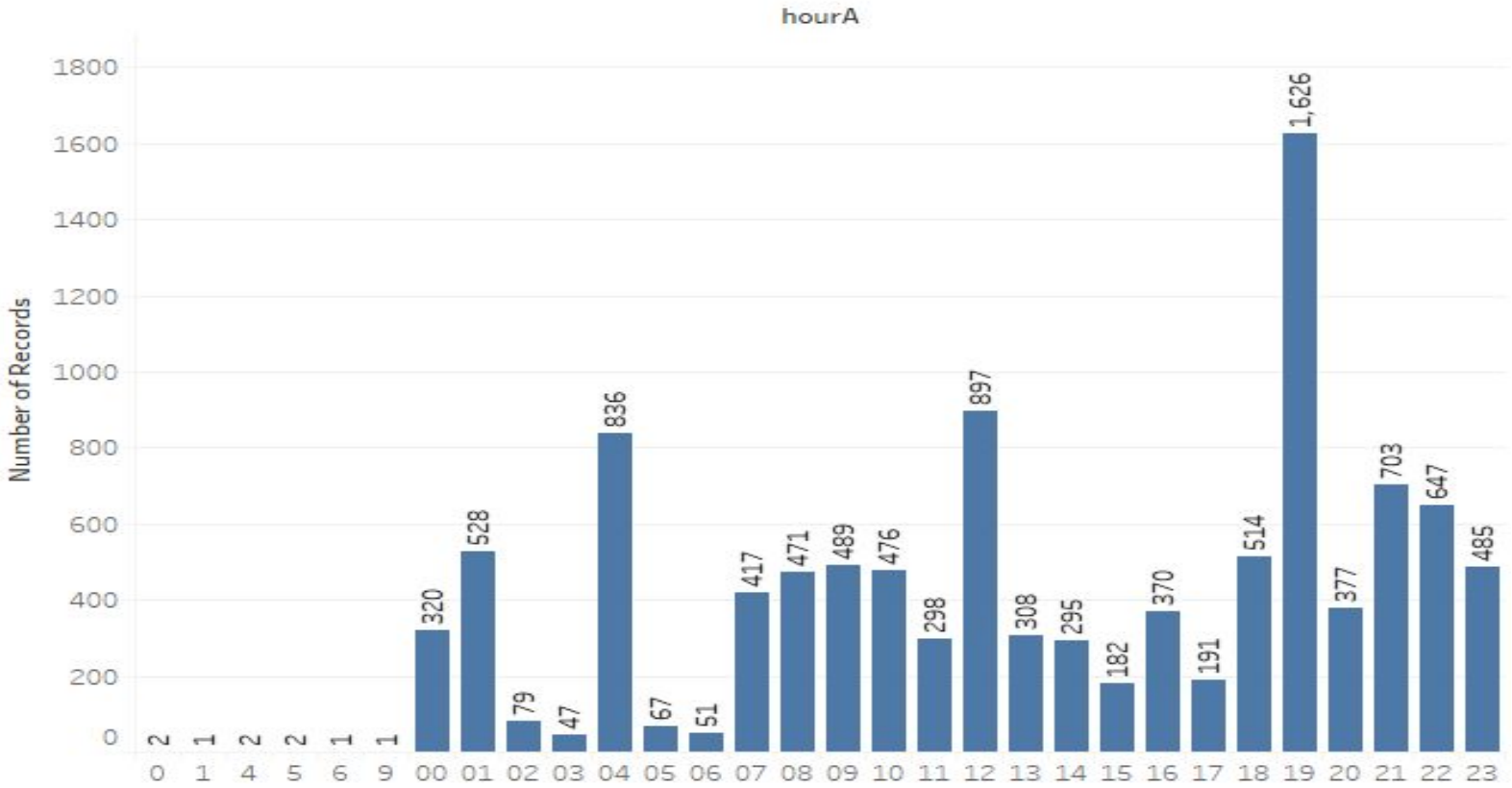
Sheet 4



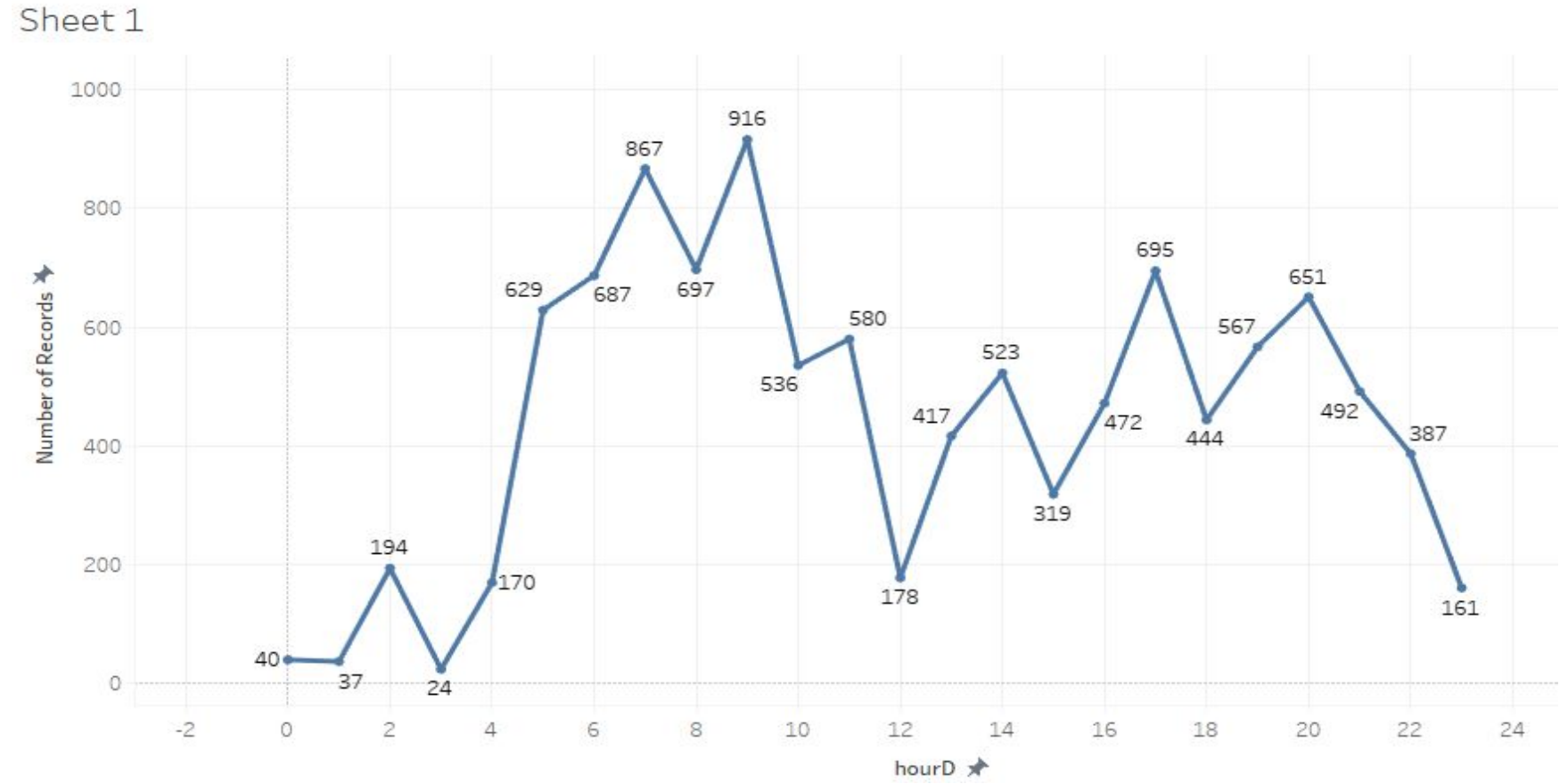
Sheet 5



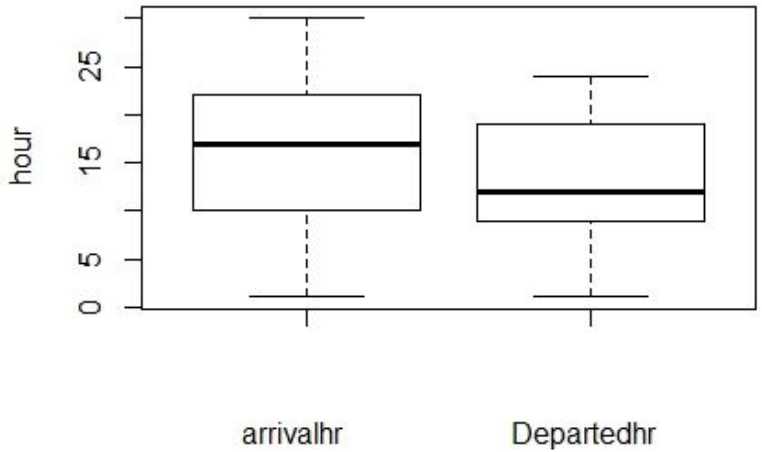
Times of hour at which passengers arrived

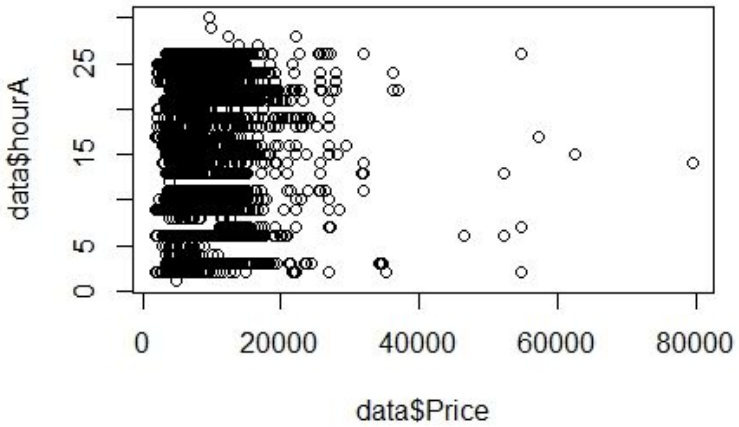


Times of hour at which passengers arrived



boxplot(data$hourA,data$hourD,ylab = "hour",xlab = "arrivalhr Departedhr")





Correlation explains the relationship between two variables whether positive, negative or 0 .

```
> cor(data)
                  Airline Date_of_Journey       Source Destination       Route     Dep_Time         hourD     minuteD
Airline        1.00000000     0.022390145 -0.013397047  0.07053230  0.02521435 -0.039507727 -0.035268951 -0.05992238
Date_of_Journey 0.02239014     1.000000000  0.167554764 -0.08857713  0.27512067 -0.007776155 -0.005080167 -0.05339896
Source         -0.01339705     0.167554764  1.000000000 -0.43422668  0.40341175  0.055193758  0.059046984 -0.05699796
Destination     0.07053230    -0.088577135 -0.434226685  1.00000000 -0.23169883 -0.079475573 -0.087777203  0.09672222
Route           0.02521435     0.275120674  0.403411749 -0.23169883  1.00000000 -0.082013353 -0.075002160 -0.06807113
Dep_Time       -0.03950773    -0.007776155  0.055193758 -0.07947557 -0.08201335  1.000000000  0.997451976  0.02898226
hourD          -0.03526895    -0.005080167  0.059046984 -0.08777720 -0.07500216  0.997451976  1.000000000 -0.02474545
minuteD        -0.05992238    -0.053398961 -0.056997961  0.09672222 -0.06807113  0.028982258 -0.024745451  1.00000000
Arrival_Time   -0.00993505    -0.009641640  0.023982730 -0.05925580  0.01090540  0.001860146  0.007453559  0.03990173
hourA          -0.00665482    -0.009363038  0.023040771 -0.05582718  0.01311232 -0.002947469  0.002467771  0.04108863
minuteA        -0.08837701    -0.060515825  0.003579417 -0.01861967 -0.12943295  0.212546173  0.212625337 -0.01673390
Duration        0.02788577    -0.001136741 -0.192008850  0.02582566 -0.06207829  0.041604668  0.041507543  0.03087595
Price          -0.03956478    -0.036906543  0.015999249 -0.26216428  0.16414933  0.002931165  0.006799237 -0.02445781
               Arrival_Time        hourA      minuteA     Duration        Price
Airline        -0.009935052 -0.006654820 -0.088377011  0.027885773 -0.039564779
Date_of_Journey -0.009641640 -0.009363038 -0.060515825 -0.001136741 -0.036906543
Source          0.023982730  0.023040771  0.003579417 -0.192008850  0.015999249
Destination    -0.059255798 -0.055827178 -0.018619674  0.025825657 -0.262164283
Route           0.010905395  0.013112318 -0.129432949 -0.062078287  0.164149327
Dep_Time        0.001860146 -0.002947469  0.212546173  0.041604668  0.002931165
hourD           0.007453559  0.002467771  0.212625337  0.041507543  0.006799237
minuteD         0.039901732  0.041088632 -0.016733901  0.030875949 -0.024457812
Arrival_Time    1.000000000  0.989738028 -0.193739735  0.033103379  0.021040371
hourA           0.989738028  1.000000000 -0.220435683  0.027652329  0.024578580
minuteA        -0.193739735 -0.220435683  1.000000000 -0.022889941  0.005083145
Duration        0.033103379  0.027652329 -0.022889941  1.000000000 -0.144280285
Price           0.021040371  0.024578580  0.005083145 -0.144280285  1.000000000
>
```

5.Insights from EDA

Some of the values of the variable ‘price’ have some extremities on both ends, so the top and bottom 5 percentile of the variable is removed to make it balanced so that when further analysis is done the result won’t be biased .

Looking at the Correlation matrix ,we can observe that there is very less correlation and also negative relation between variables .there is high correlation only between created variables from Dep_time & Arr_time.

There are no insights using clustering