FLIGHT PRICE PREDICTION

OBJECTIVE

Our objective is to use the CRISP-DM Methodology, to understand the requirement, explore the data and get an in-depth analysis of the data. Besides this, we tend to find ways to determine or predict the airfare with other components (such as which airline, what duration, etc.) for the purpose of decision making.

DESCRIPTIVE AND VISUAL ANALYTICS

DATA UNDERSTANDING

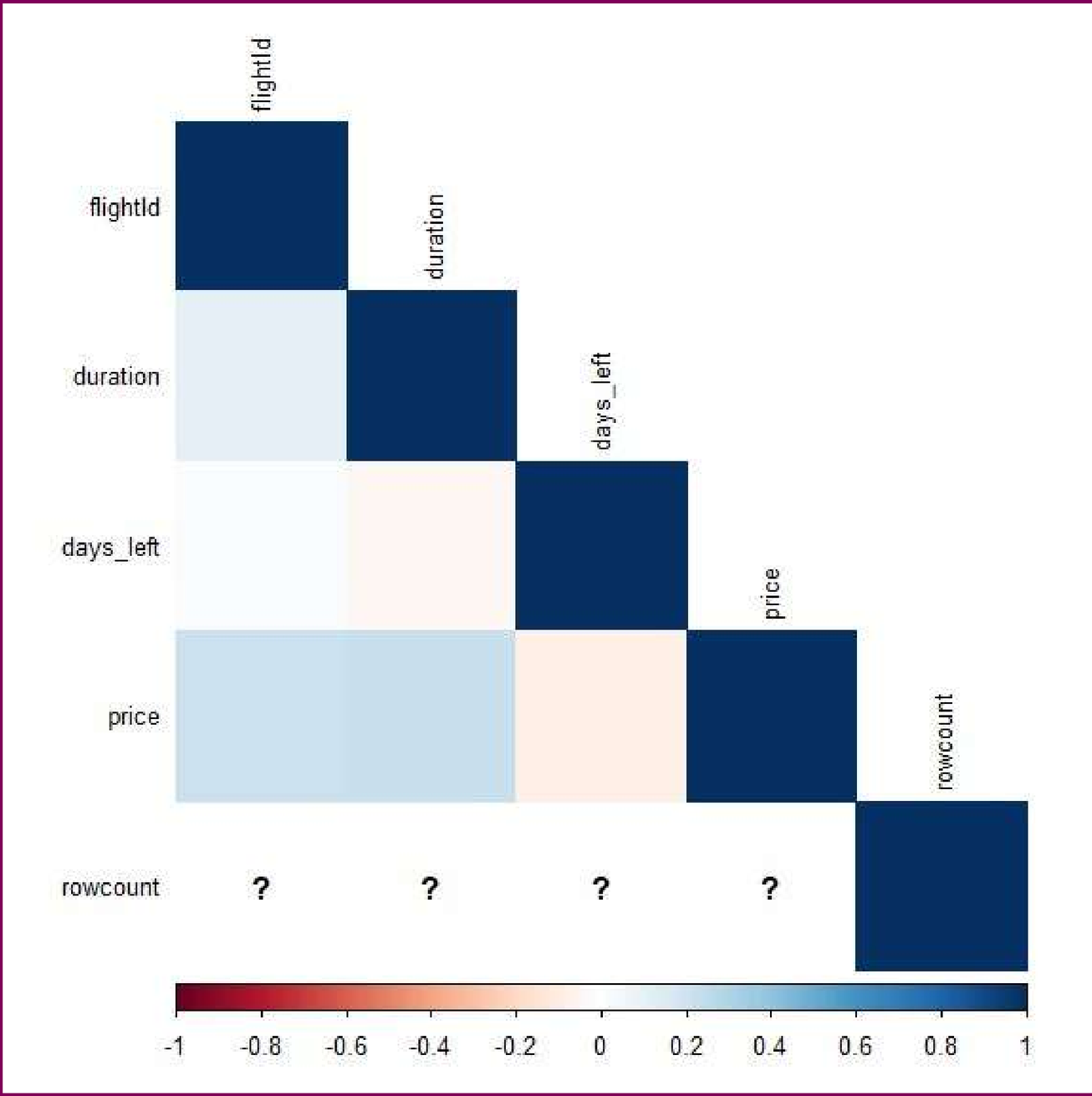
- We used functions like dim, glimpse and describe to understand the data better.

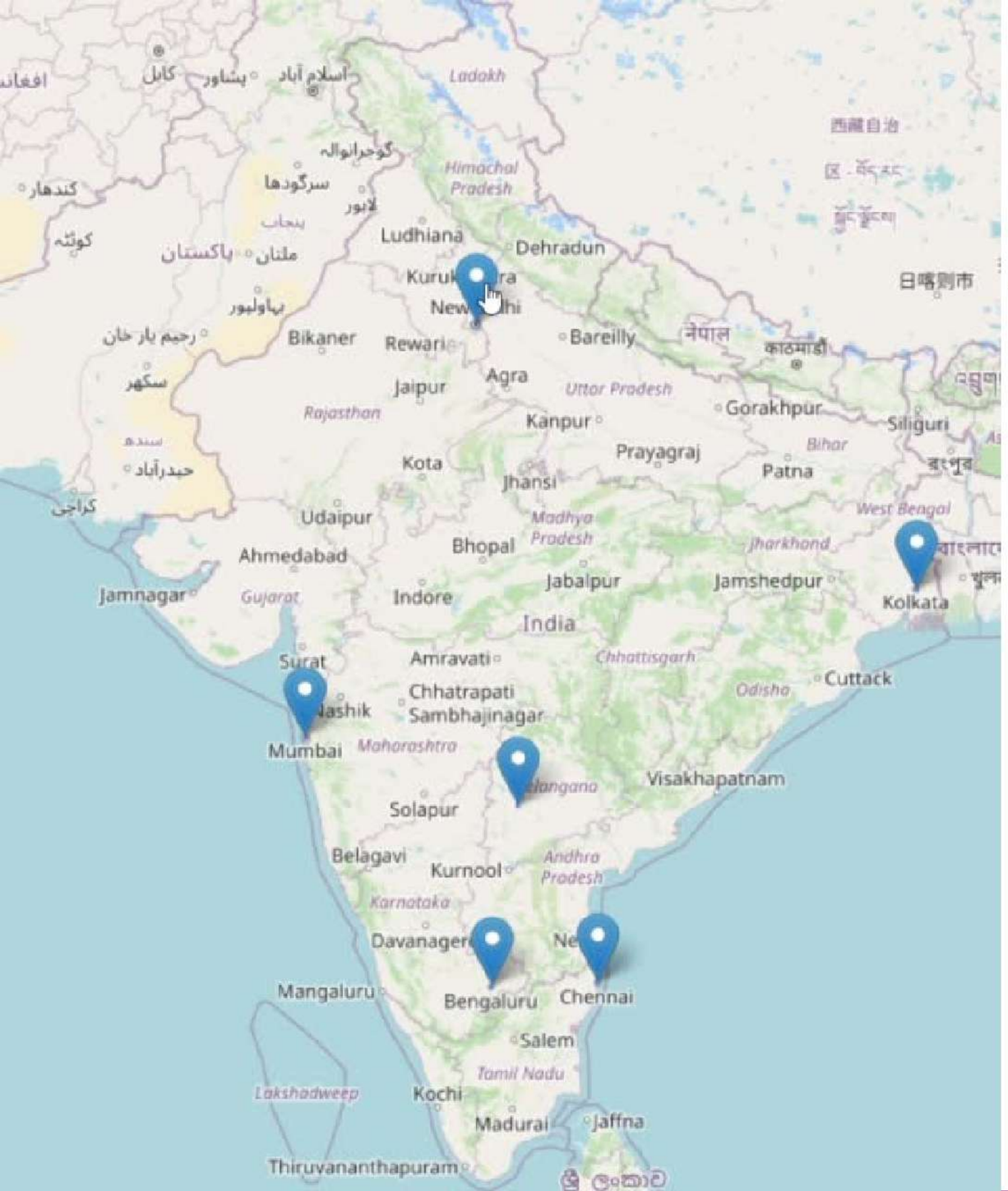
- We used summary function to derive the descriptive statistics of the data and then created a function to automatically detect the numerical columns, as an when they are added and they obtain the descriptive statistics of each attribute using for loop.

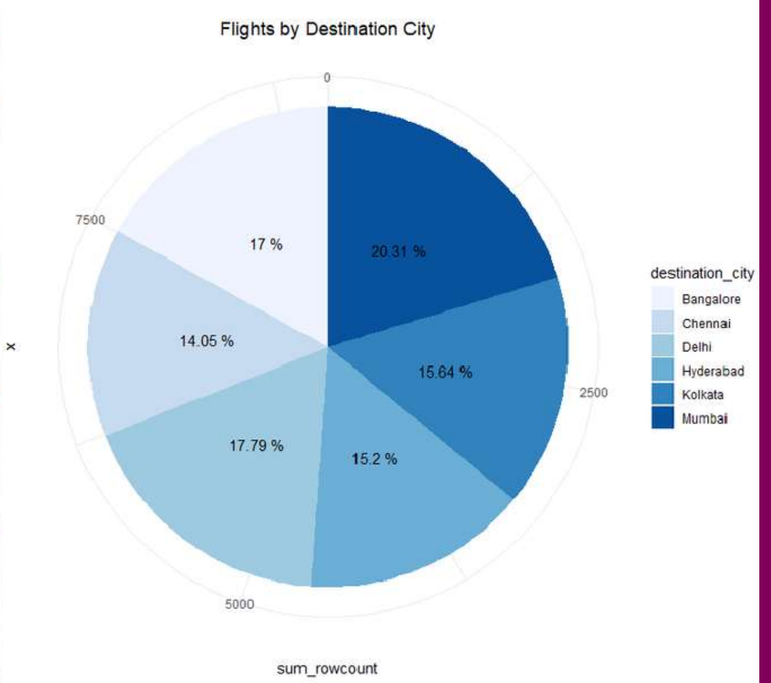
CORRELATION

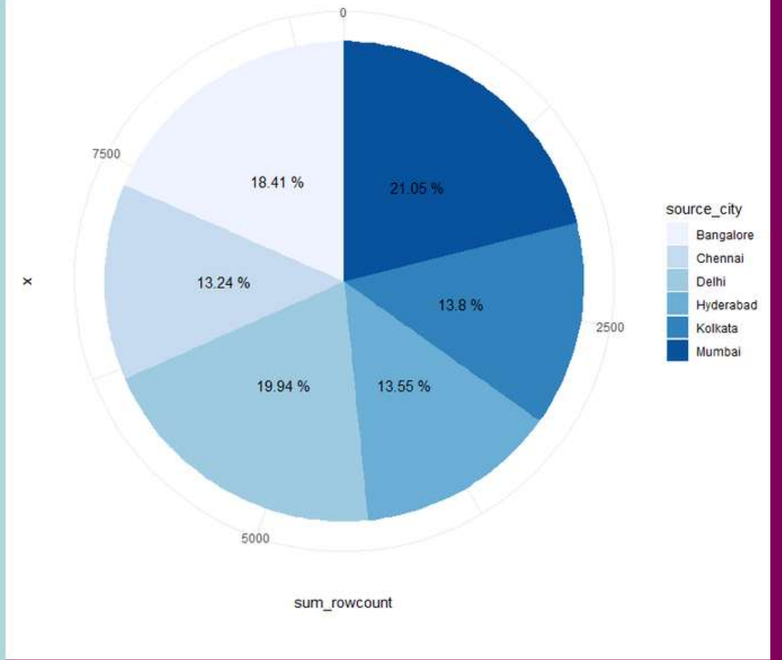
- The heat map gives us an understanding that price is slightly positively correlated to duration and slightly negatively correlated to days_left

- Hence, as duration increases price increases and vice-versa. Also, as number of days_left decreases, price increases and vice-versa

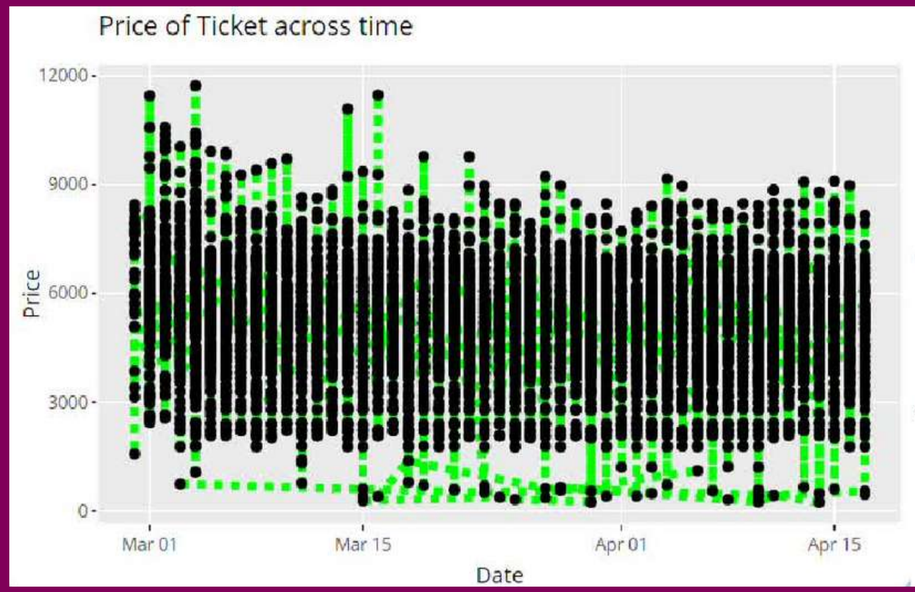


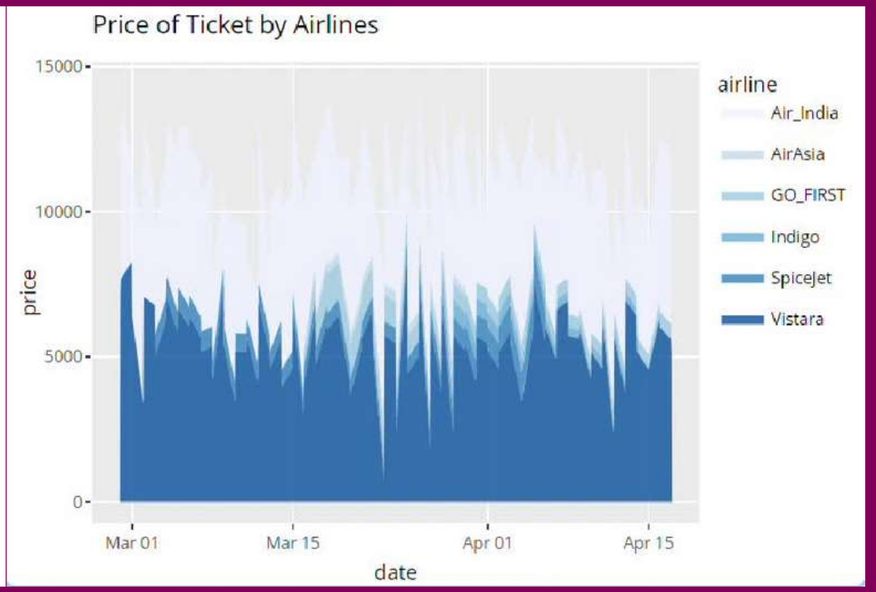






AIRPORT OVERVIEW





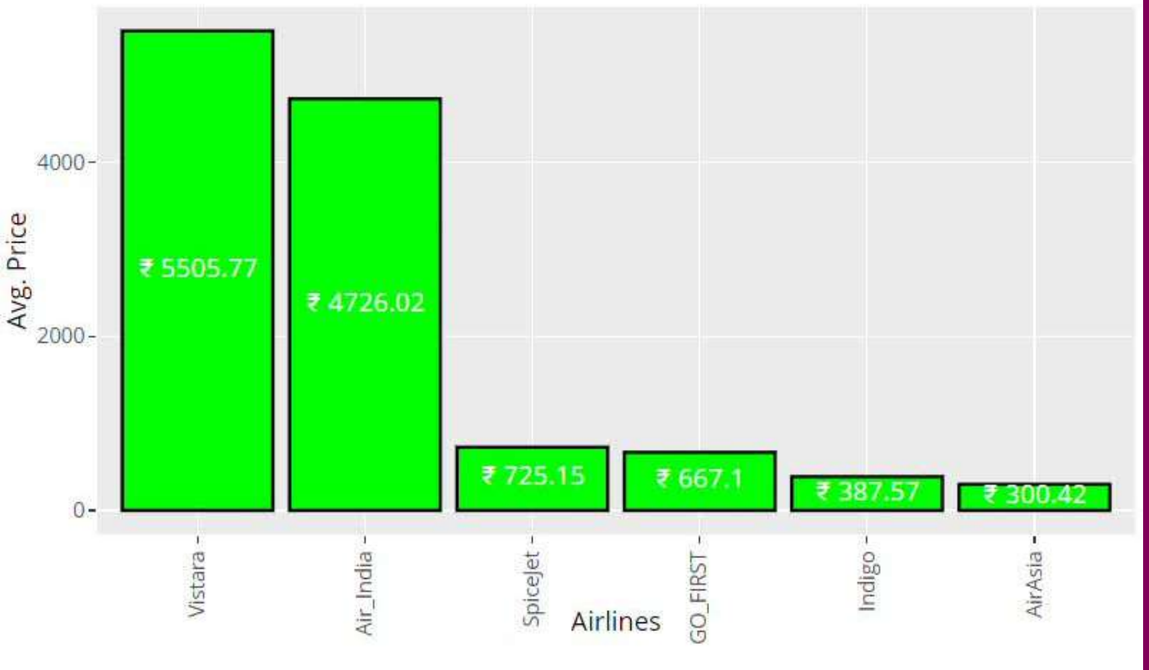
PRICING OVERVIEW

PRICING

Average price of ticket by Departure Time

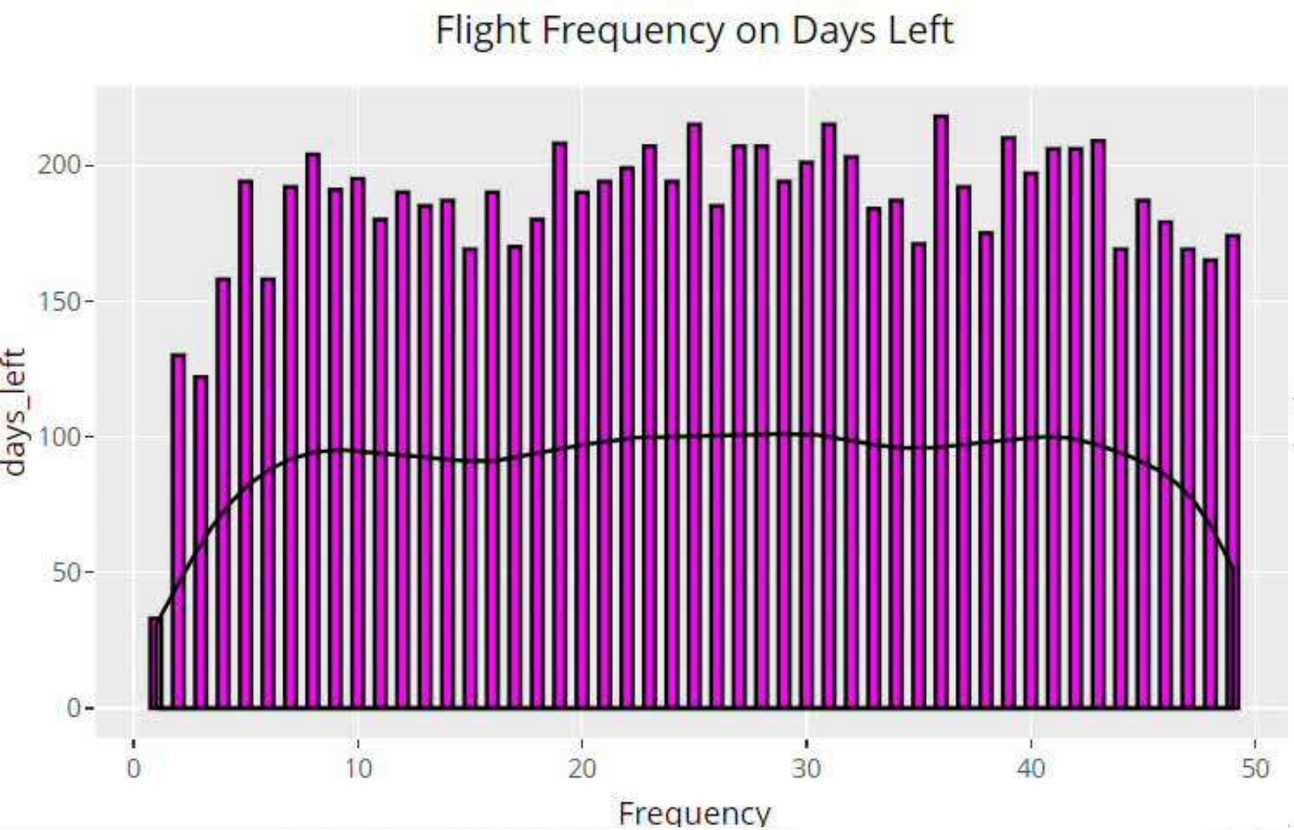
Average price of ticket by Arrival Time

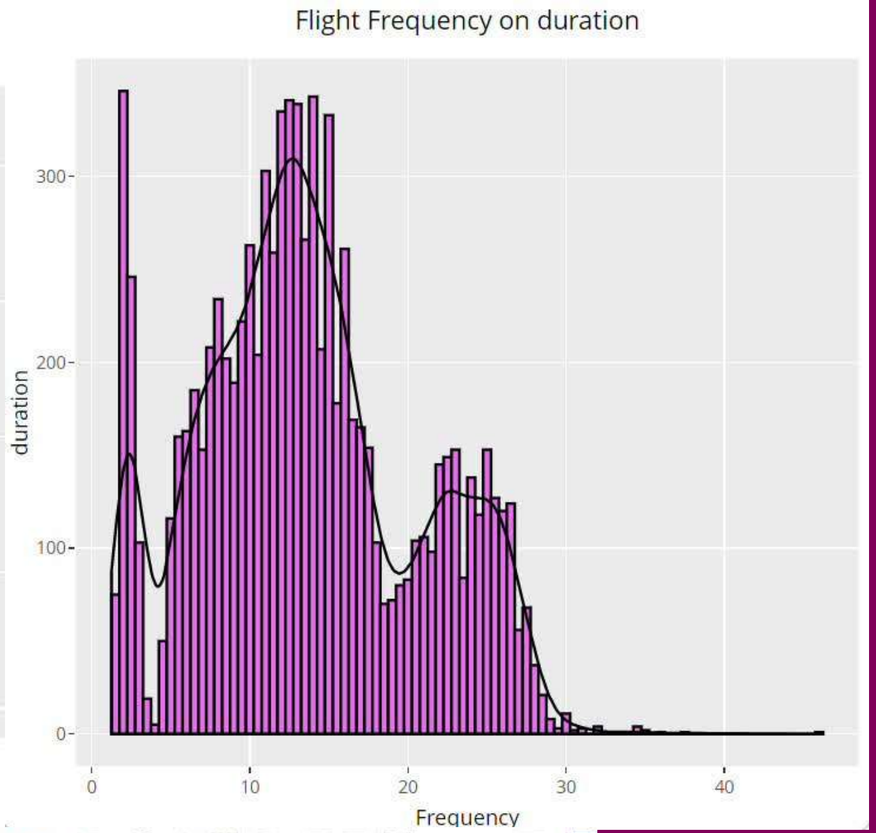
Average price of ticket by Airlines

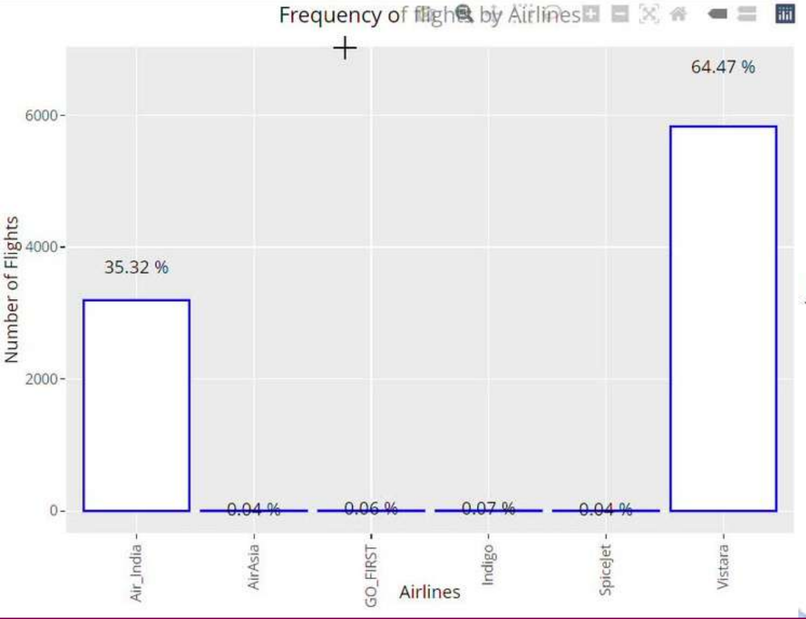


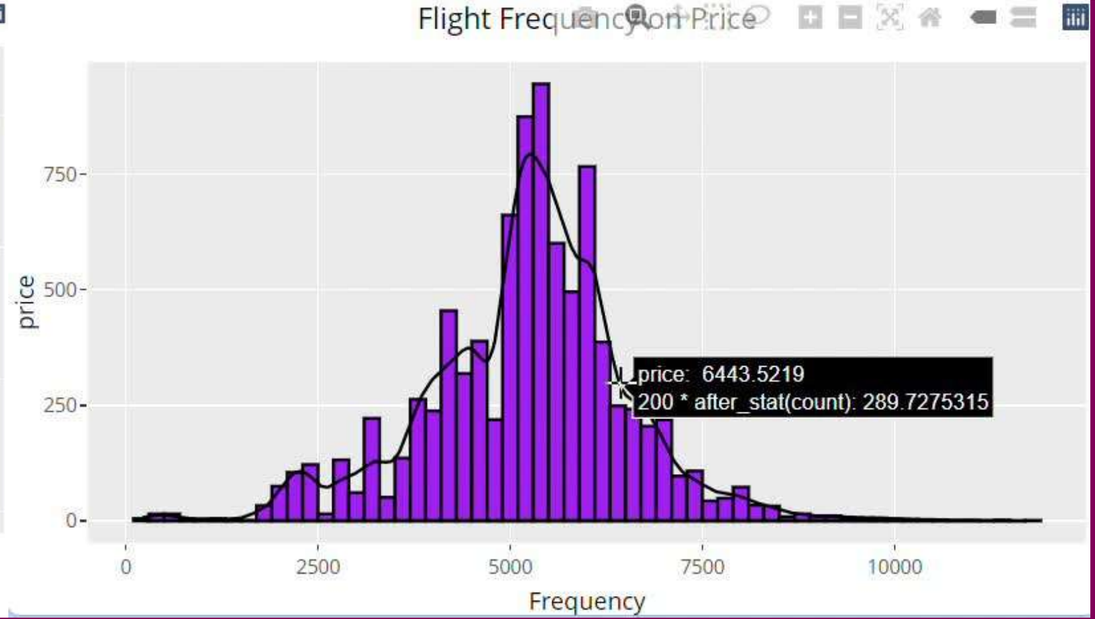
Average price of ticket by Number of Stops

FREQUENCY OVERVIEW



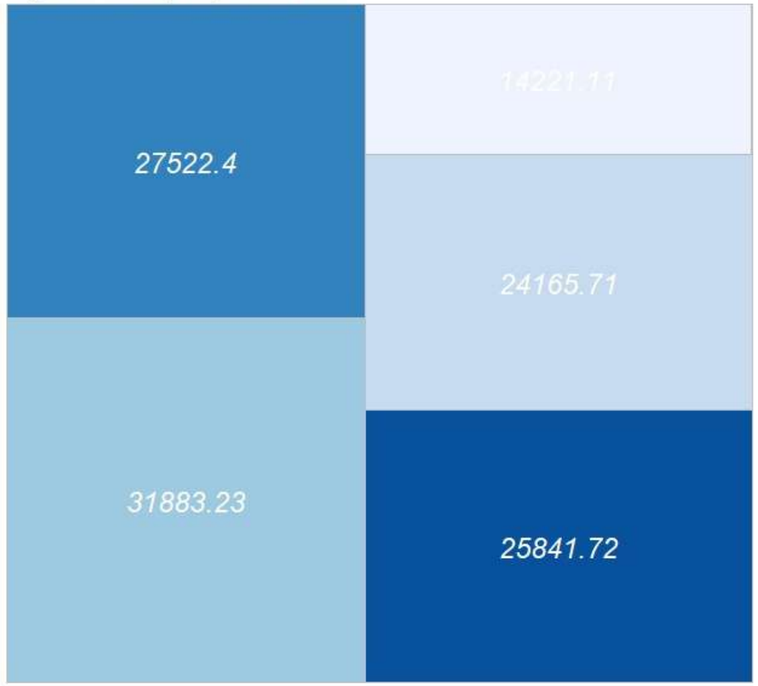






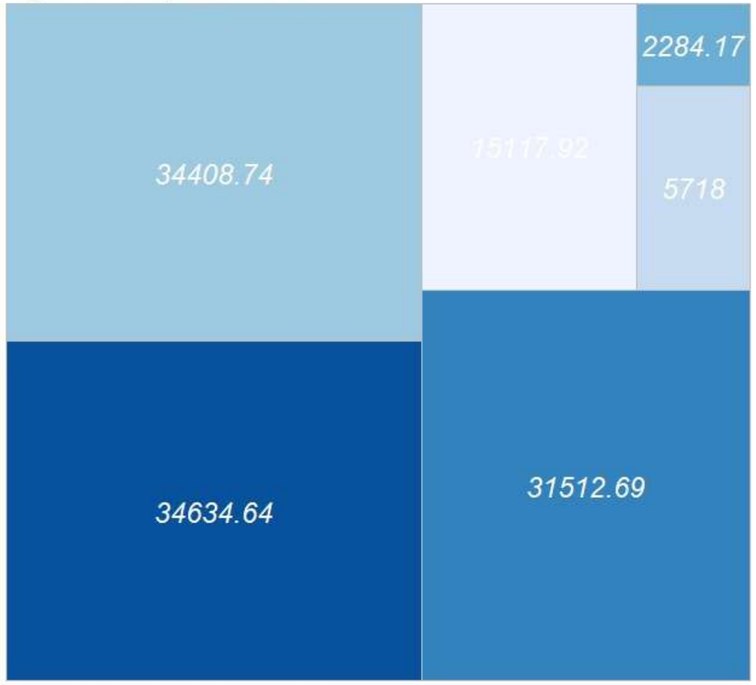
FLIGHT DURATION OVERVIEW

Flight Duration by Departure Time



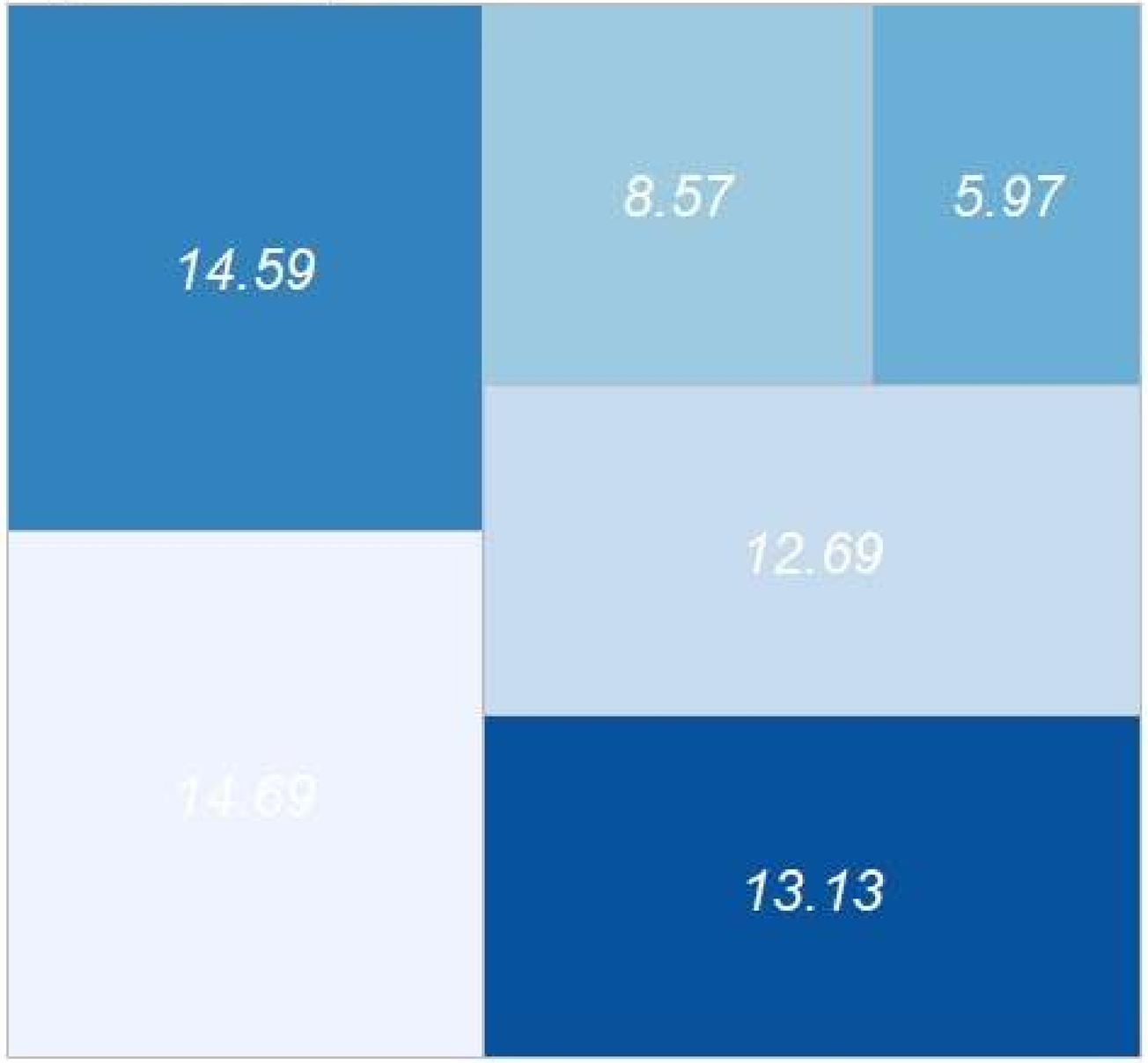
departure_time
- Afternoon
- Early_Morning
- Evening
- Late_Night
- Morning
- Night

Flight Duration by Arrival Time



arrival_time
- Afternoon
- Early_Morning
- Evening
- Late_Night
- Morning
- Night

Flight Duration by Airlines



airline
- Air_India
- AirAsia
- GO_FIRST
- Indigo
- SpiceJet
- Vistara

INFERENTIAL AND PREDICTIVE ANALYTICS

INDEPENDENT T-TEST


```r
# Independent T-test for varied departure time
#H0 (Null Hypothesis): Pre-sunset prices >= Post-sunset time prices
#H1 (Alternative Hypothesis): Pre-sunset prices < Post-sunset time prices
dep_time_results<-t.test(x=df$price[df$departure_time==c("Early_Morning","Morning","Afternoon")],
                         y=df$price[df$departure_time==c("Evening","Night","Late_Night")],
                         alternative = "less")
print(dep_time_results)

# Since, p-value is greater than 0.05, that means we accept the Null Hypothesis.
# Hence, the prices for tickets Pre-sunset (i.e. Early Morning,Morning and Afternoon) is higher or equal to the
# prices of tickets Post-sunset (Evening, Night and Late Night) while the passenger departures.

#-------------------------------------------------------------------------------------------------

# Independent T-test for varied arrival time
#H0 (Null Hypothesis): Pre-sunset prices >= Post-sunset time prices
#H1 (Alternative Hypothesis): Pre-sunset prices < Post-sunset time prices
arr_time_results<-t.test(x=df$price[df$arrival_time==c("Early_Morning","Morning","Afternoon")],
                         y=df$price[df$arrival_time==c("Evening","Night","Late_Night")],
                         alternative = "less")
print(arr_time_results)

# Since, p-value is less than 0.05, that means we reject the Null Hypothesis.
# Hence, the prices for tickets Pre-sunset (i.e. Early Morning,Morning and Afternoon) is lesser to the
# prices of tickets Post-sunset (Evening, Night and Late Night) while the passenger arrives.
```

ONE-SIDED T-TEST

```r
# One-sided T-test using Average Population Mean for Ticket Price
#H0 (Null Hypothesis): mean sample price  = mean population price of ₹4854 (given population mean)
#H1 (Alternative Hypothesis): mean sample price  != mean population price of ₹4854 (given population mean)
results<-t.test(x=df$price,
                alternative = "two.sided",
                mu=4854)

print(results)

# Since, p-value is less than 0.05, that means we reject the Null Hypothesis.
# Hence, there is a significant difference between the sample mean price with population mean price of ₹4854,
# where the sample mean price (₹5219.848) is greater than that of the population mean price (₹4854)
```

ONE-WAY ANOVA TEST

```r
# One way Anova
#H0 (Null Hypothesis): All population means are equal
#H1 (Alternative Hypothesis): Not all of the population means are equal
anova_test_results<-aov(price~airline,data=df) # Comparing price airline wise
summary(anova_test_results)

# Since, p-value is less than 0.05, that means we reject the Null Hypothesis.
# Hence, Price of all Airlines are not significantly same.
```

LINEAR REGRESSION

The Significant Model for Predicting Price would be:
price= 4819.6004 +43.4959*duration -7.5340*days_left

Q. Let's say the budget for airfare is 4500 and we want the duration of flight to be 2 hours, how many days earlier, should we book the ticket?
(We define a function and find the desired answer.)
find_days_left(4500,2)
Hence, You should be booking the ticket 53.97 days before the departure date.

The Significant Model for Predicting Price with the inclusion of number of stops would be:
price= 7338.2588 -15.6204*duration -8.5386*days_left
-4286.4288*stop_zero_Dummy-1459.0417*stop_one_Dummy

LOGISTIC REGRESSION

The predictive model is:

price= 4819.6004 +43.4959*duration -7.5340*days_left

Q. Find the price of the ticket when duration is 2 hours and we have 30 days left for departure.

Based on our model, the Price of the ticket would be ₹4680.571 when the duration of flight is 2 hours and 30 days are left for departure.

Since, the p-value of the predict class is greater than 0.5, we can conclude that the predicted class is loyal because the value is TRUE.

LINEAR DISCRIMINANT ANALYSIS

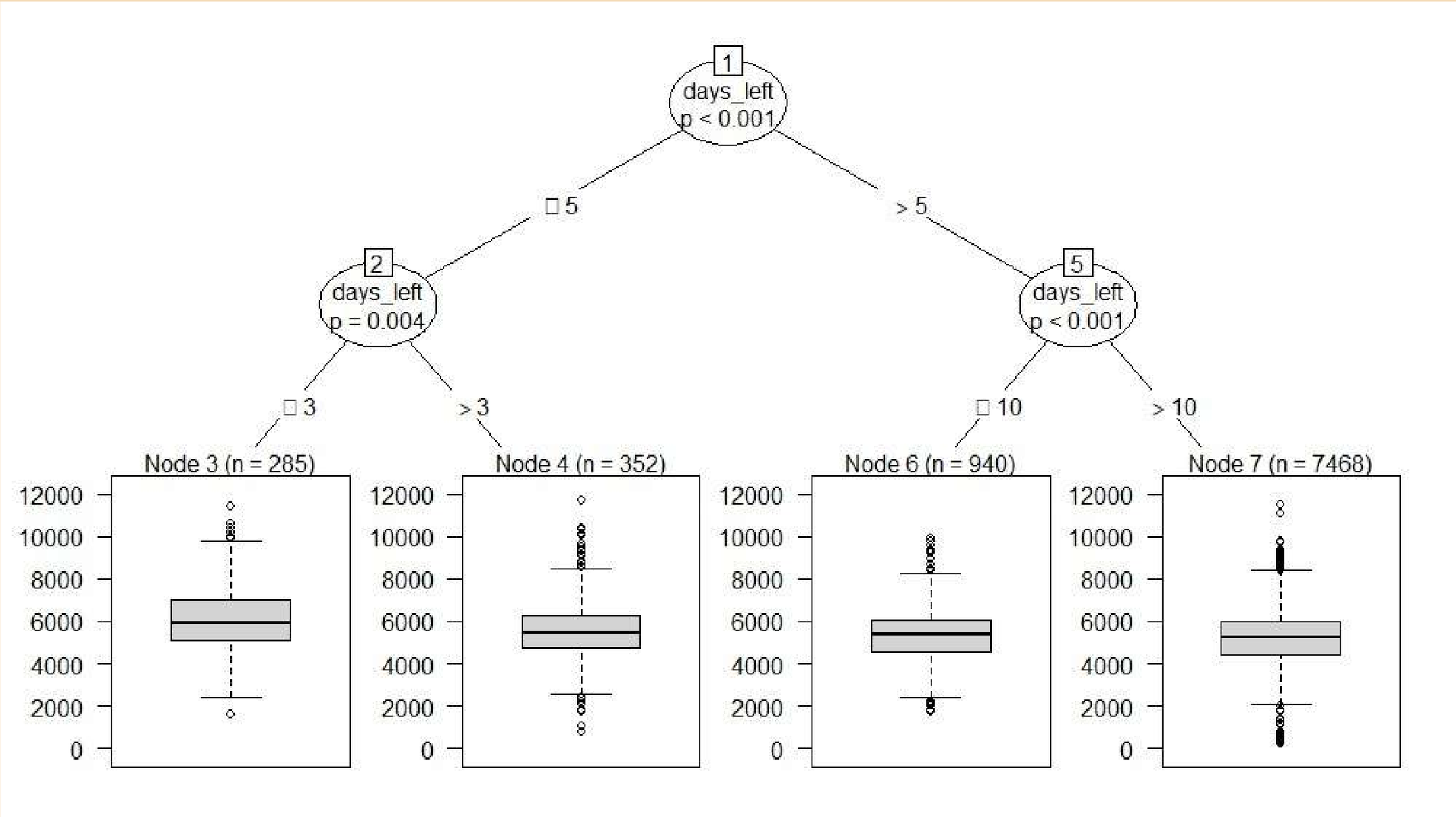
Group means:

airline	price	duration	days_left
Air_India	4726.023	14.693161	25.22848
AirAsia	300.425	12.6875	24.5
GO_FIRST	667.1	8.566	31.2
Indigo	387.5667	5.971667	39.16667
SpiceJet	725.15	14.585	23.75
Vistara	5505.765	13.127038	26.12022

Q. Predicting a Case with Price of the ticket as ₹6000, Duration as 4.25 hours and Days_left as 6 days

Based on our model and case, the predicted airline for the said parameters is Vistara.

DECISION TREE ANALYSIS



CLASSIFICATION

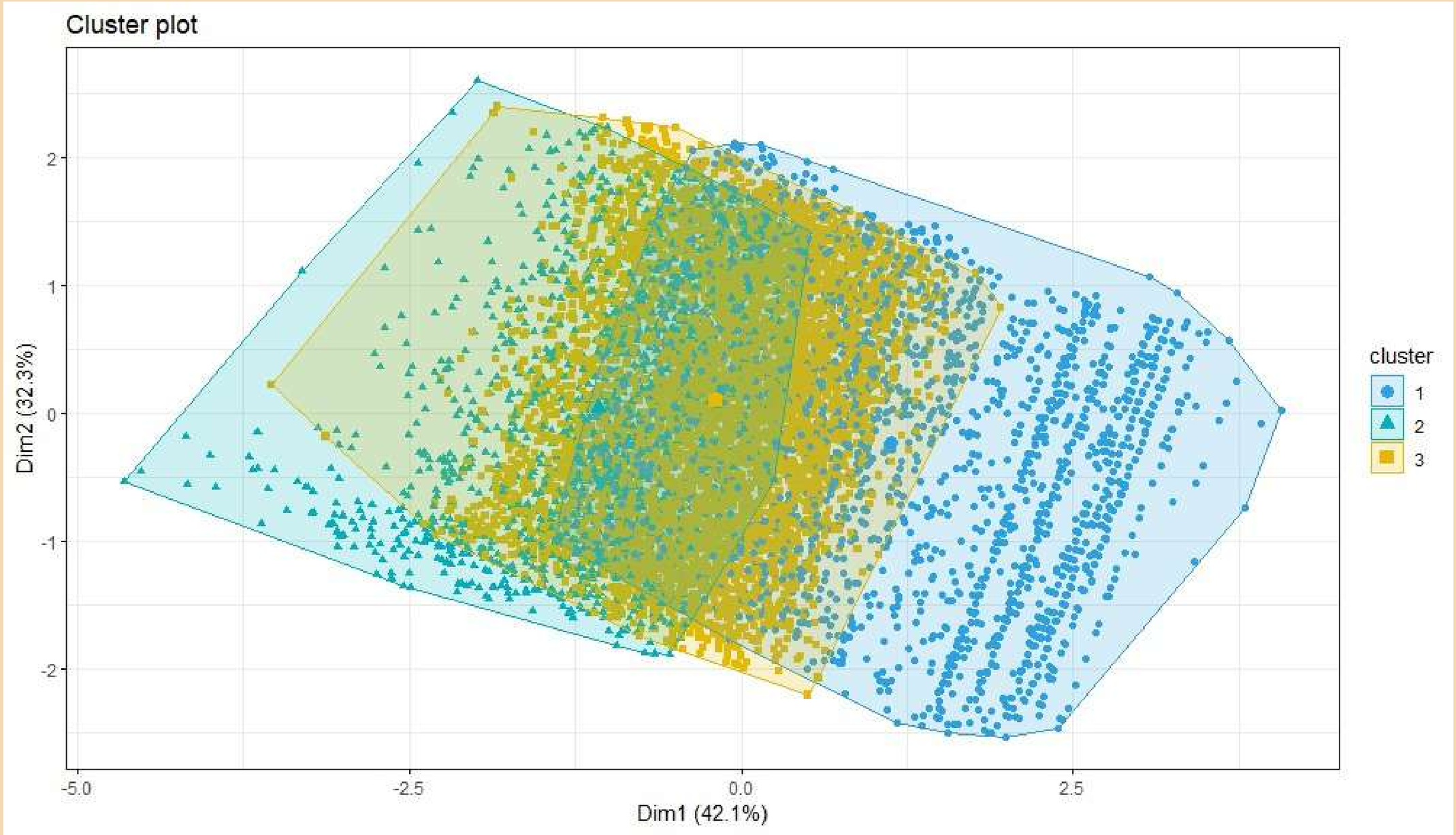
Q. Predicting the source city with price as ₹4500, duration as 3.75 hours and days_left as 18.

Based on our case and model, we get, Bangalore has the highest probability for our said case.

Q. Predicting the destination_city with price as ₹4500, duration as 3.75 hours and days_left as 18.

Based on our case and model, we get, Chennai has the highest probability for our said case.

K-MEANS CLUSTER ANALYSIS




THANK YOU
FOR YOUR
ABOARD!
SOUTHEAST AIRLINES
SA
674103
ROM
ITALY
PCS.
WGT.
41-03
NATIONAL
NO. OF PIECES
(STRAP) TAG
WEIGHT
DATE
ROM
DIA