IST687 – FINAL PROJECT

CUSTOMER SATISFACTION ANALYSIS PROJECT

Ananth Adisesh
ISCHOOL | SYRACUSE UNIVERSITY

Data Overview:

Southeast Airlines has two major issues. First, in today's market, the loyalty system perceived to be the industry's best practice is not successful. The other metric is the customer's turnover, which is the most commonly used. This cannot keep customers from quitting and from losing. We would also like to figure out a few more effective and leading indicators, which will allow us to forecast before we lose our customers. It included thousands of flight segment data observations. For a specific customer, each row represents one flight section. Each column defines a specific flight segment characteristic. We cleanse the dataset by splitting 27 attributes into three types, airline attributes, attributes of the customer, and location attributes in our dataset. In order to understand each attribute better, we should test the following table.

Airline Attributes	Customer Attributes	Location Attributes
Day.of.Month	Age,	Destination.State,
Partner.Name	Gender, Price.Sensitive,	Origin.City,
Flight.cancelled	Flights.Per.Year,	Destination.City,
Flight.time.in.minutes	Type of Travel,	Origin.State,
Flight.time.in.minutes	Shopping.Amount.at.Airport	Dlat
Flight.Distance	Eating.and.Drinking.at.Airport,	Dlong,
Flight.Month	Year.Length,	Olong,
Scheduled.Departure.Hour	Airline.Status, Free Text,	Olat
Arrival.Delay.in.Minutes	Class	

Objective:

The objective is to provide better business solutions to reduce customer churn for Southeast Airlines.

Business Questions:

- Are ‘Age’ and ‘Gender’ important factors in determining customer travel frequency and airline rating?
- Can customers’ likelihood to recommend an airline, be used as a factor to correlate LTR to other factors such as Destination City, Comments, and Origin City etc.?
- How can the diverse set of customer base of the airline be analysed for gains?
- Determining the key states where the customer is less likely to recommend.
- Using the comments/feedback given by the customers, how to analyse if the customer will recommend the airline to another person?

Data Cleaning:

The first step is to read the data file (given on JSON format) and import it to R Studio to start the data cleaning.

Second, we can use trimws to erase the leading and trailing column whitespace. Then, with colnames, we rename each column to make it more comprehensible. We can also distinguish the name of the city and of the state by using a distinguish feature both in the destination and in the departure. After that, I found that there are several NA’s in the "Comments" section. We can convert the argument into the character form in order to solve this problem. Using is.na function, missing values are found and "No Comments" for those values are added.

The next thing to do is to decide where the NA’s are in all columns and assign ‘0’ to the NAs. I did the data cleaning part of the details now. We can then add various new attributes to the data structure to enhance data processing. They are more arrival delay than five minutes, average travel time, average time for late departure and if it is a long trip.

Data Visualization using descriptive statistics and Histograms:

In order to interpret the data using statistical or easily recognizable visual statistics, descriptive statistics are used to summarize the quantitative values in R to draw clear conclusions about the distribution of the results.

The data frame 'df' used in our project has 10 numeric columns, which can be statistically analysed and visualised. These columns can be grouped into two sets of attributes: The Customer attributes and the Airline attributes.

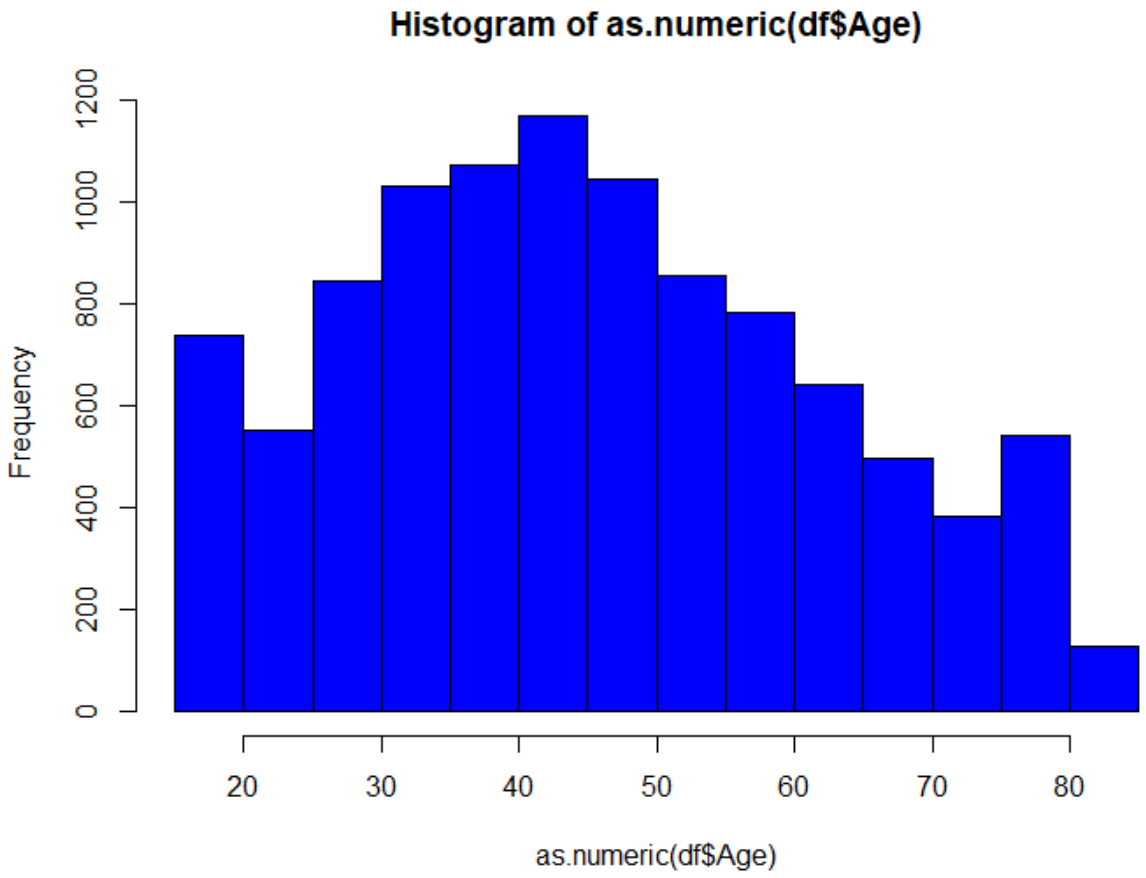
1. Customer Attributes:

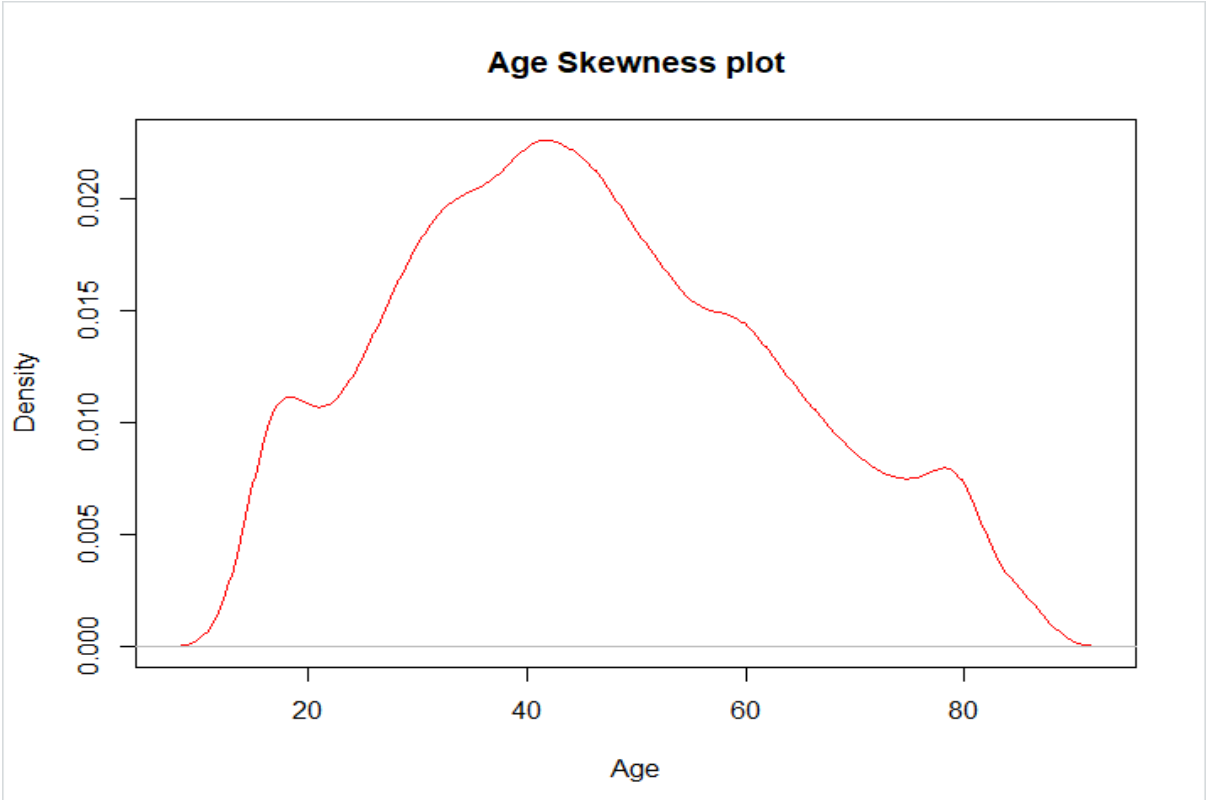
a. Age: Code Snippet:

```
#Descriptive statistics, histogram and plot - 1
desc(df$Age)
summary(df$Age)
as.numeric(df$Age)
hist(as.numeric(df$Age), col = "blue")
mean(df$Age,na.rm = TRUE)
df$Age[is.na(df$Age)] <- mean(df$Age,na.rm = TRUE)
plot(density(as.numeric(df$Age)),main = "Age Skewness plot", xlab = "Age",col = "red")
```

Descriptive statistics :

Mean	46.32
Median	45
Minimum	15
Maximum	85
Range	70d
Trimmed	45.8
N (total number of variables)	10282
1st quartile	33.0
3rd quartile	59.00
Standard deviation	17.37
Mean Absolute Deviation	19.27
Standard error	0.17
Skew	0.24
kurtosis	-0.73


Histogram of as.numeric(df$Age)
Frequency
0
200
400
600
800
1000
1200
20
30
40
50
60
70
80
as.numeric(df$Age)

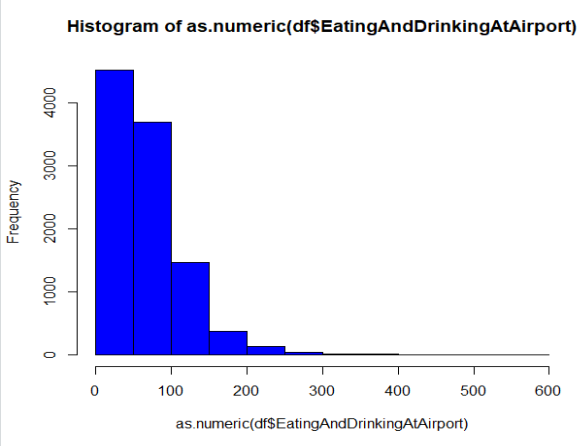

Age Skewness plot
Density
0.000
0.005
0.010
0.015
0.020
20
40
60
80
Age

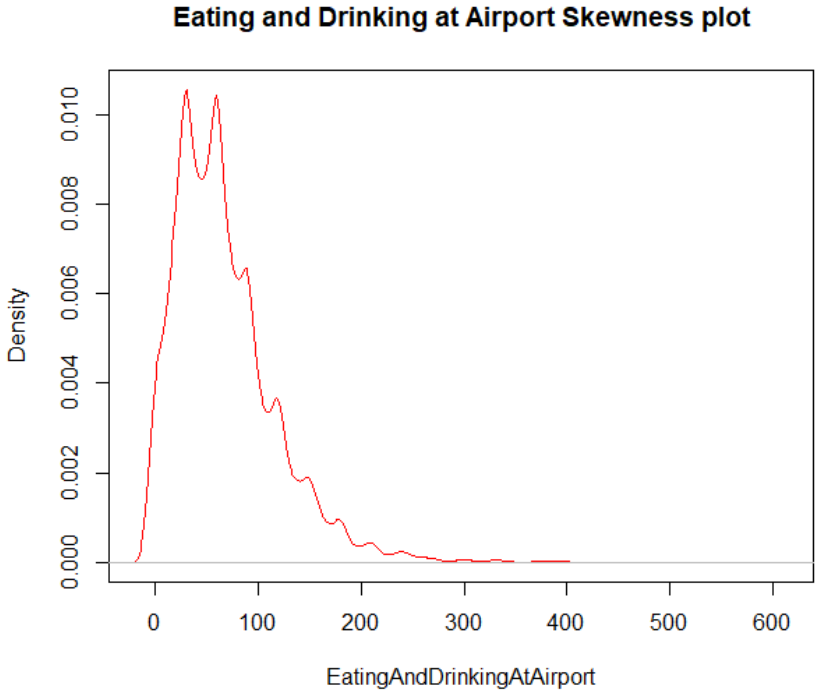
b. Money Spent on Eating and Drinking at the airport: Code snippet:

```
#Descriptive statistics, histogram and plot - 2
desc(df$EatingAndDrinkingAtAirport)
summary(df$EatingAndDrinkingAtAirport)
as.numeric(df$EatingAndDrinkingAtAirport)
hist(as.numeric(df$EatingAndDrinkingAtAirport), col = "blue")
mean(df$EatingAndDrinkingAtAirport,na.rm = TRUE)
df$EatingAndDrinkingAtAirport[is.na(df$EatingAndDrinkingAtAirport)] <- mean(df$EatingAndDrinkingAtAirport,na.rm = TRUE)
plot(density(as.numeric(df$EatingAndDrinkingAtAirport)),main = "Eating and Drinking at Airport Skewness plot", xlab = "EatingAndDrinkingAtAirport",col = "red")
```

Descriptive Statistics :

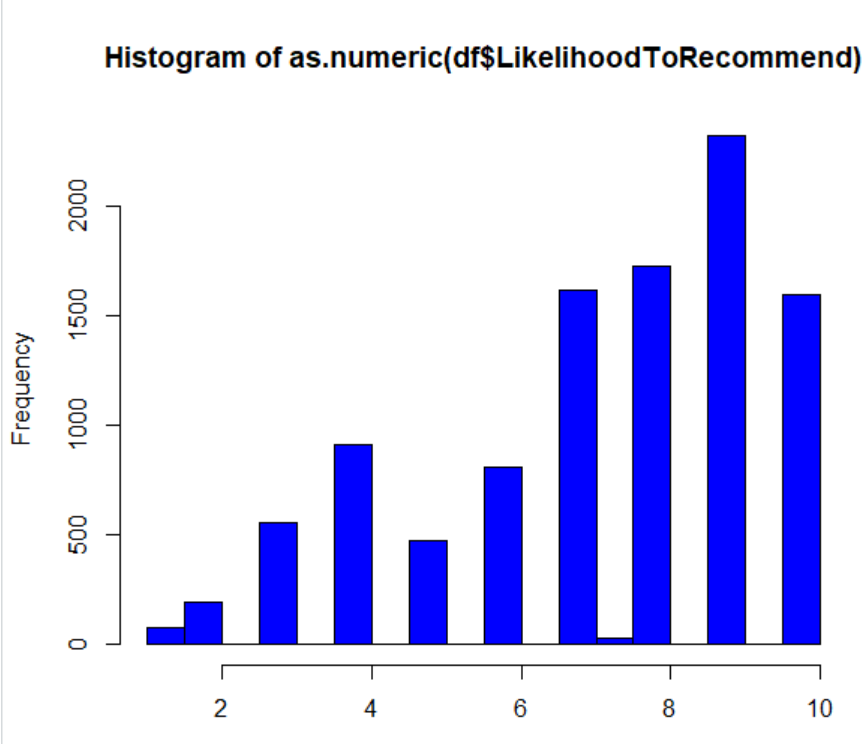
Mean	68.02
Median	60.00
Minimum	0.00
Maximum	805.00
Range	805
Trimmed	61.56
N (total number of variables)	10282
1st quartile	30.00
3 rd quartile	90.00
Standard deviation	53.58
Mean Absolute Deviation	19.27
Standard error	0.53
Skew	2.4
kurtosis	14.98

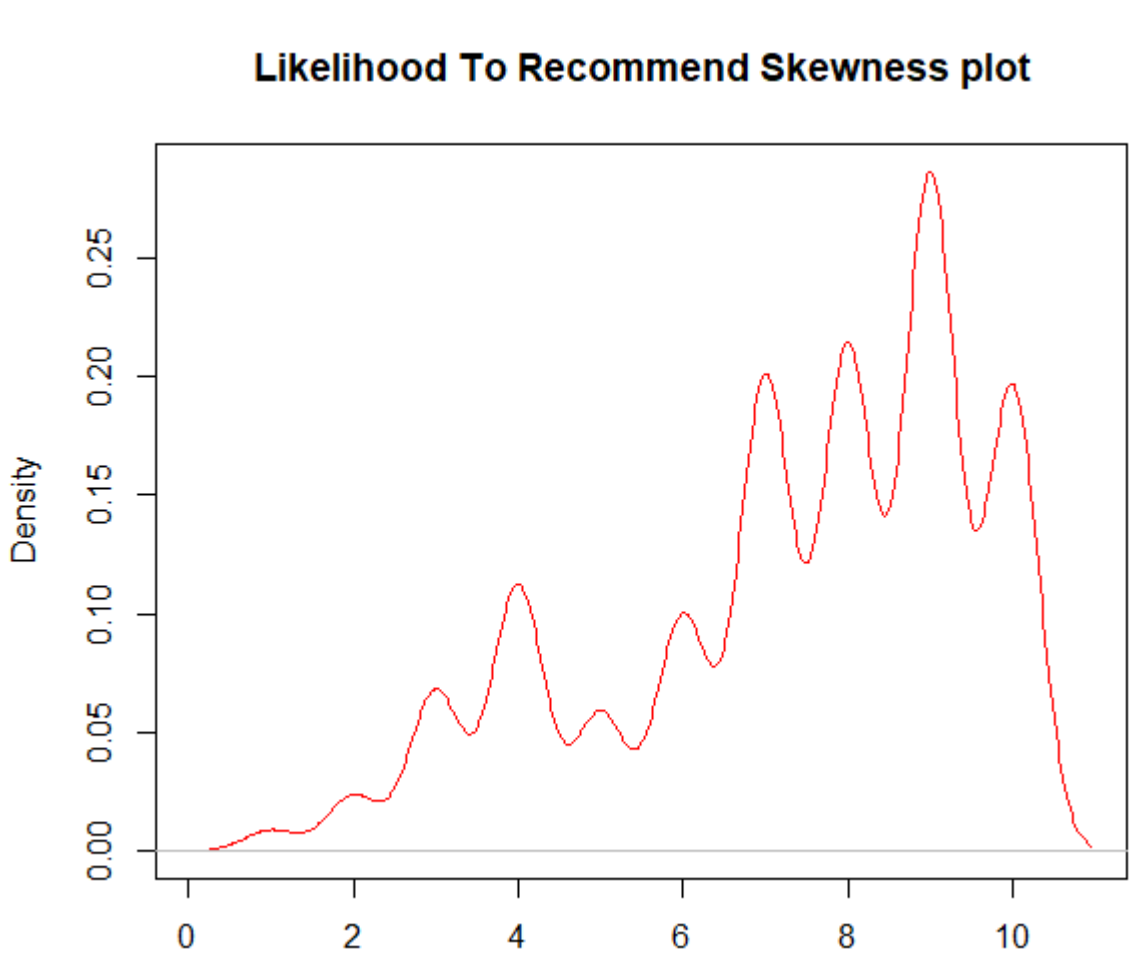




c. Likelihood to recommend by the customer: Code Snippet:

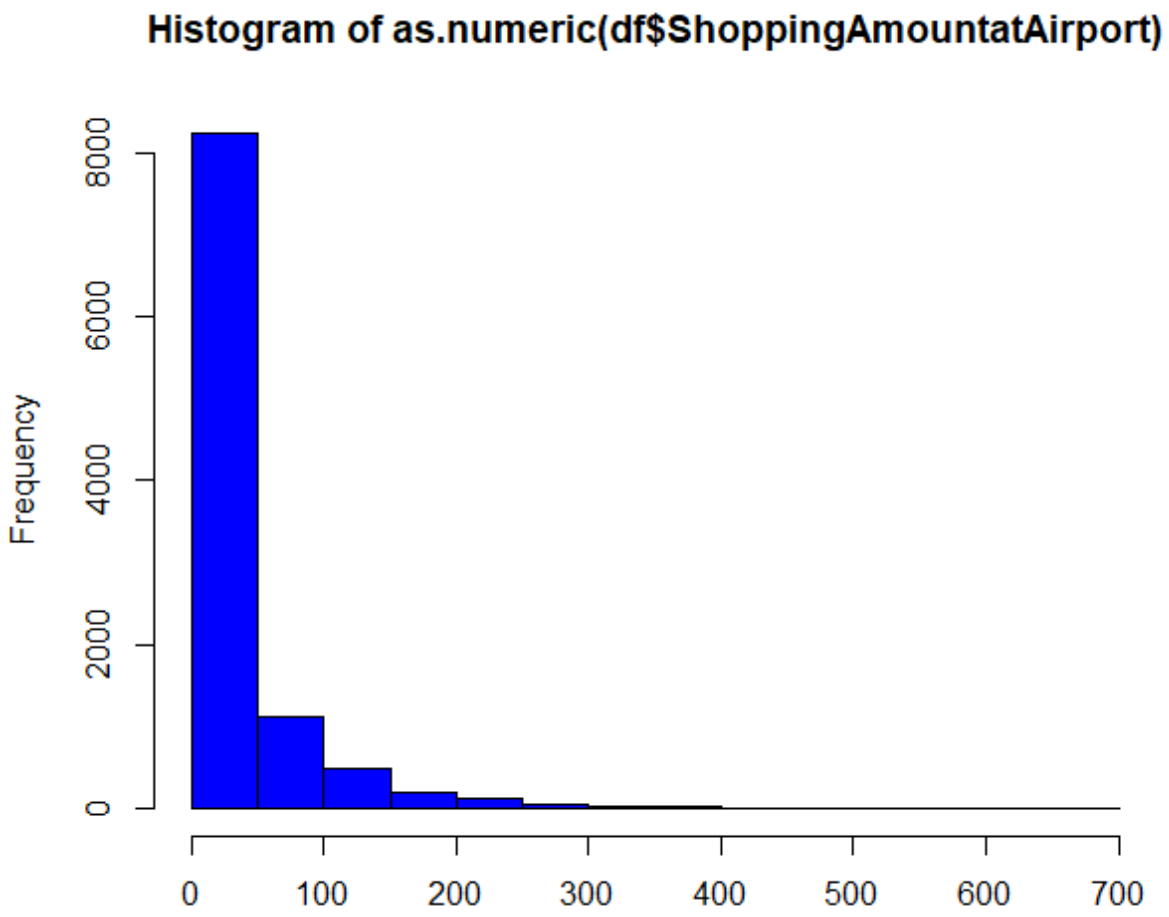
```
#Descriptive statistics, histogram and plot - 3
desc(df$LikelihoodToRecommend)
summary(df$LikelihoodToRecommend)
as.numeric(df$LikelihoodToRecommend)
hist(as.numeric(df$LikelihoodToRecommend), col = "blue")
mean(df$LikelihoodToRecommend,na.rm = TRUE)
df$LikelihoodToRecommend[is.na(df$LikelihoodToRecommend)] <- mean(df$LikelihoodToRecommend,na.rm = TRUE)
plot(density(as.numeric(df$LikelihoodToRecommend)),main = "Likelihood To Recommend Skewness plot", xlab = "LikelihoodToRecommend",col = "red")
```

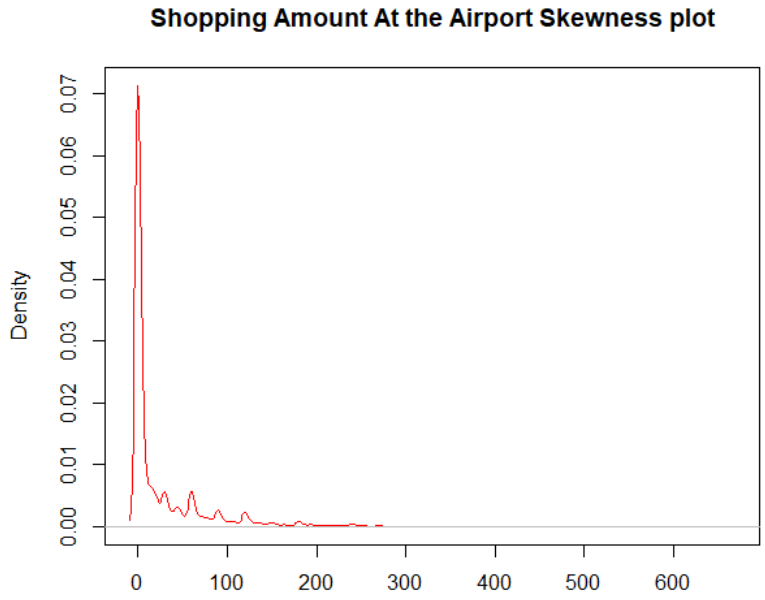

Histogram of as.numeric(df$LikelihoodToRecommend)
Frequency
0
500
1000
1500
2000
2
4
6
8
10


Likelihood To Recommend Skewness plot
Density
0.00
0.05
0.10
0.15
0.20
0.25
0
2
4
6
8
10

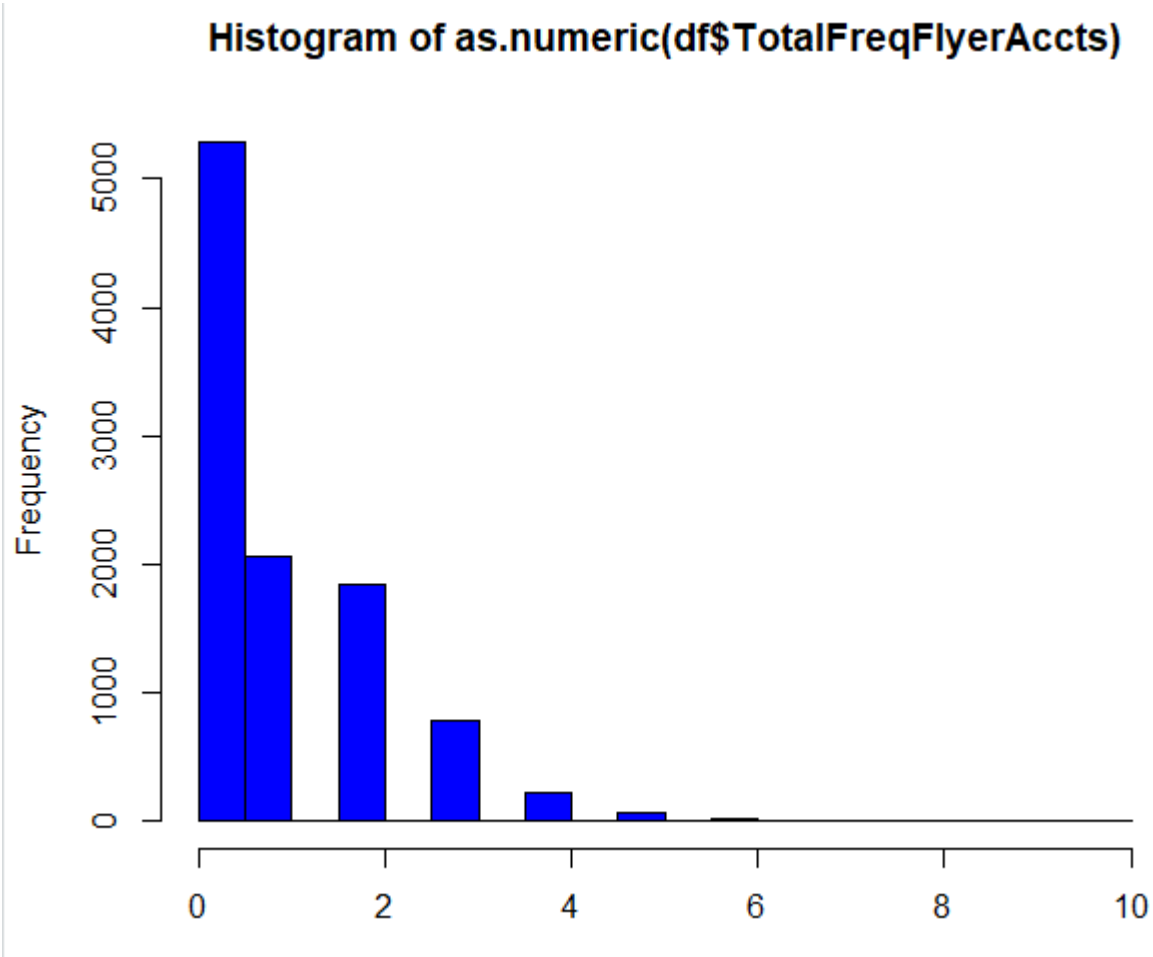
d. Shopping amount at the airport by the customer: Code snippet:

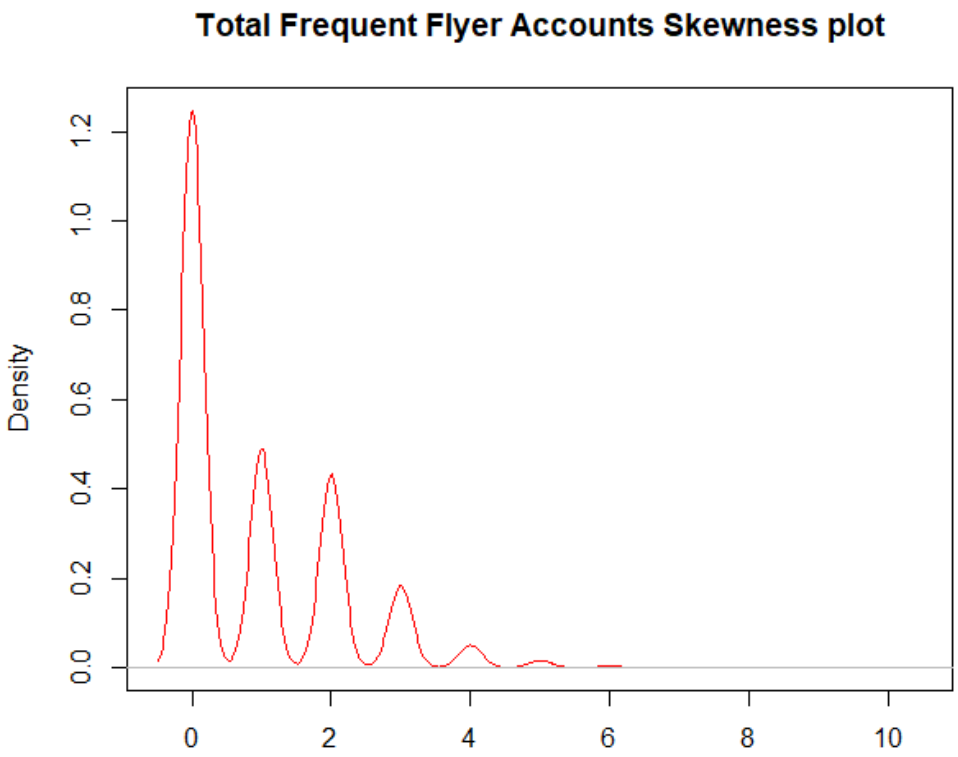
```
#Descriptive statistics, histogram and plot - 4
desc(df$ShoppingAmountatAirport)
summary(df$ShoppingAmountatAirport)
as.numeric(df$ShoppingAmountatAirport)
hist(as.numeric(df$ShoppingAmountatAirport), col = "blue")
mean(df$ShoppingAmountatAirport,na.rm = TRUE)
df$ShoppingAmountatAirport[is.na(df$ShoppingAmountatAirport)] <- mean(df$ShoppingAmountatAirport,na.rm = TRUE)
plot(density(as.numeric(df$ShoppingAmountatAirport)),main = "Shopping Amount At the Airport Skewness plot", xlab = "ShoppingAmountAtAirport",col = "red")
```



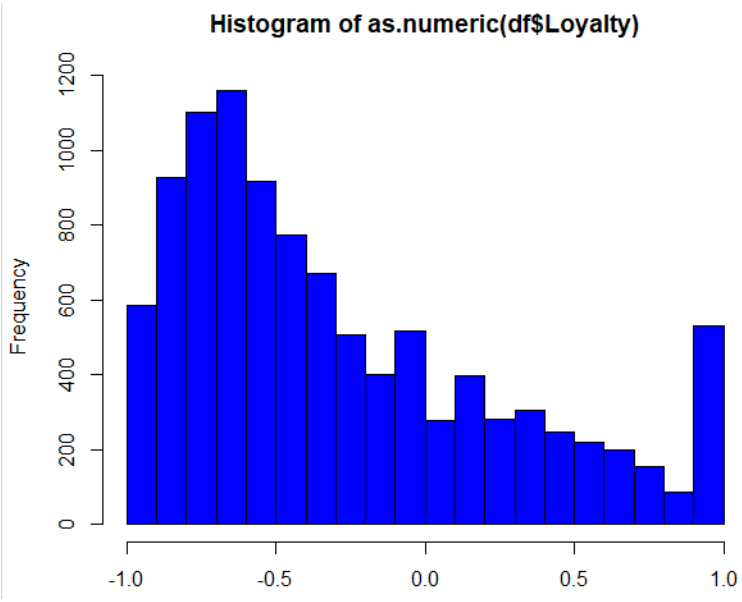


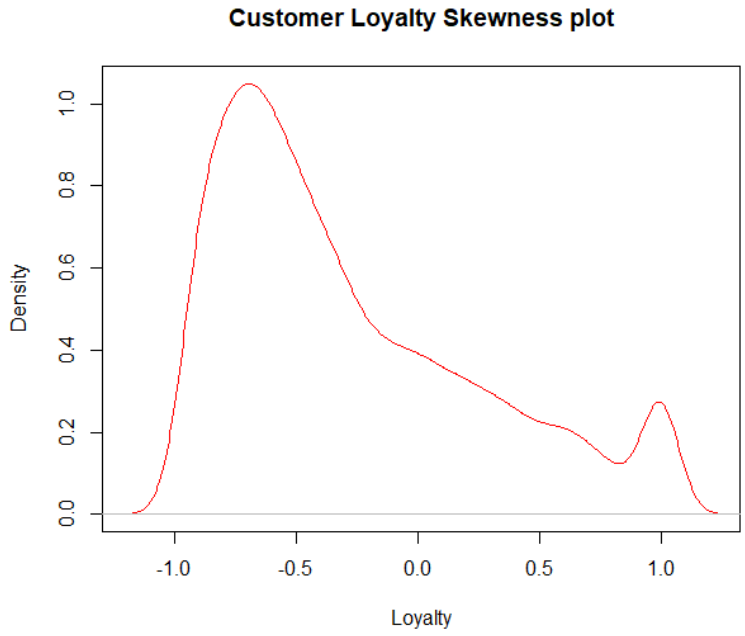
e. **Total number of Frequent Flier Accounts of the customers:**





f. Loyalty of customers:



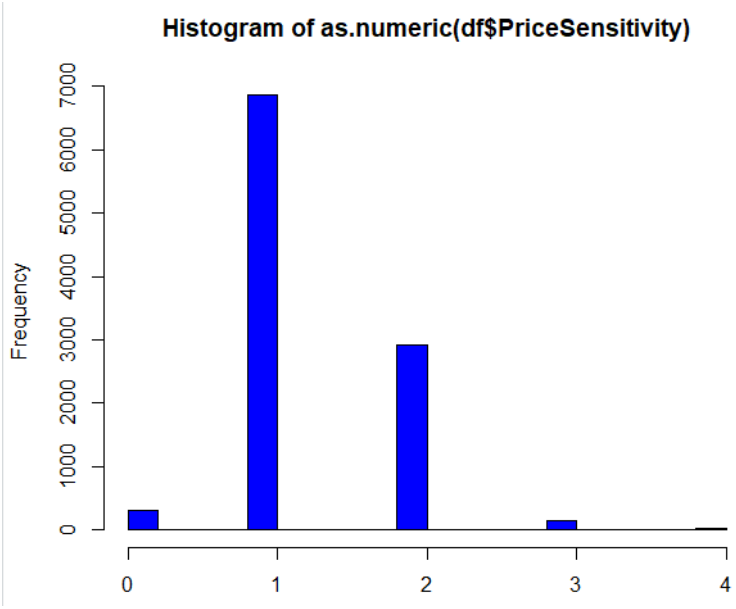


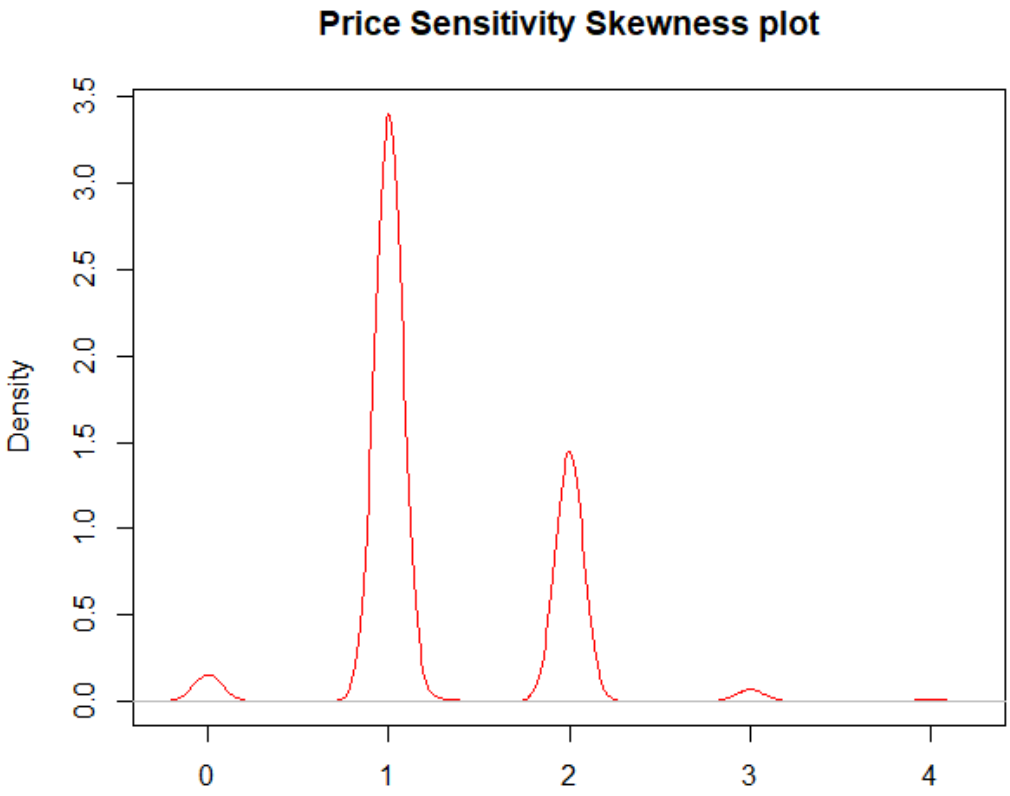
2. Airline Attributes:

a. Price Sensitivity of the tickets:

Descriptive Statistics :

Mean	1.28
Median	1.000
Minimum	0.00
Maximum	4.00
Range	4
Trimmed	1.24
N (total number of variables)	10282
1st quartile	1.00
3rd quartile	2.00
Standard deviation	0.55
Mean Absolute Deviation	0
Standard error	0.01
Skew	0.71
kurtosis	0.86

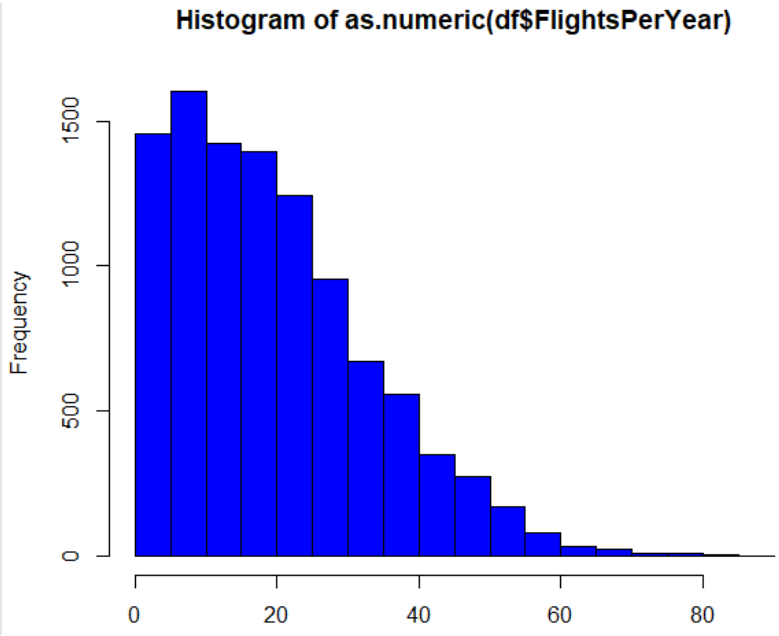

Histogram of as.numeric(df$PriceSensitivity)
Frequency
0
1000
2000
3000
4000
5000
6000
7000
0
1
2
3
4

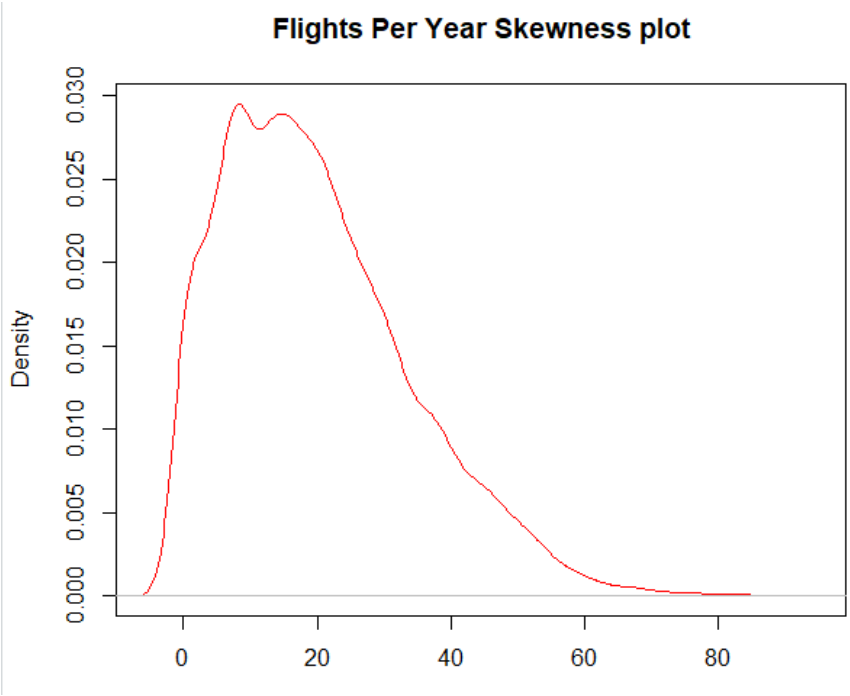

Price Sensitivity Skewness plot
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0
1
2
3
4

b. Number of Flights Per Year:

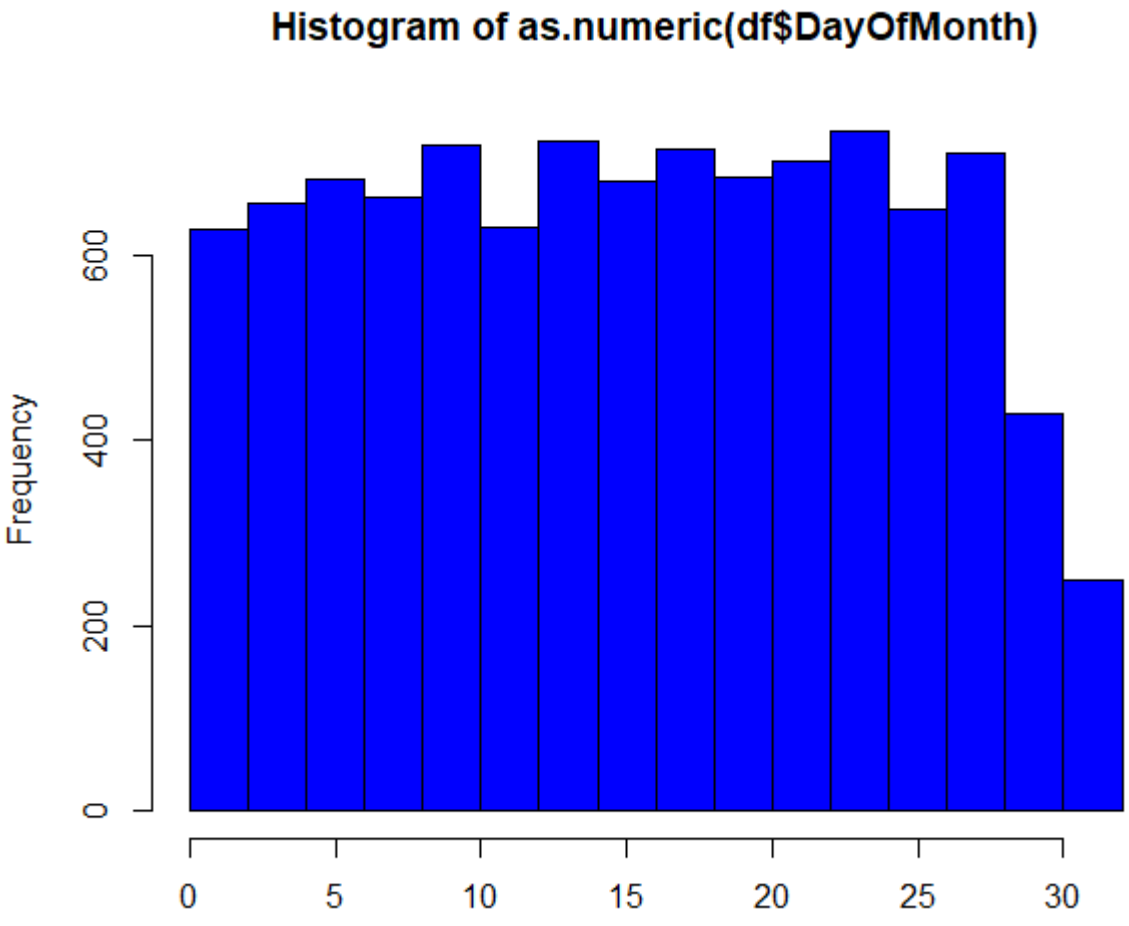
Descriptive Statistics:

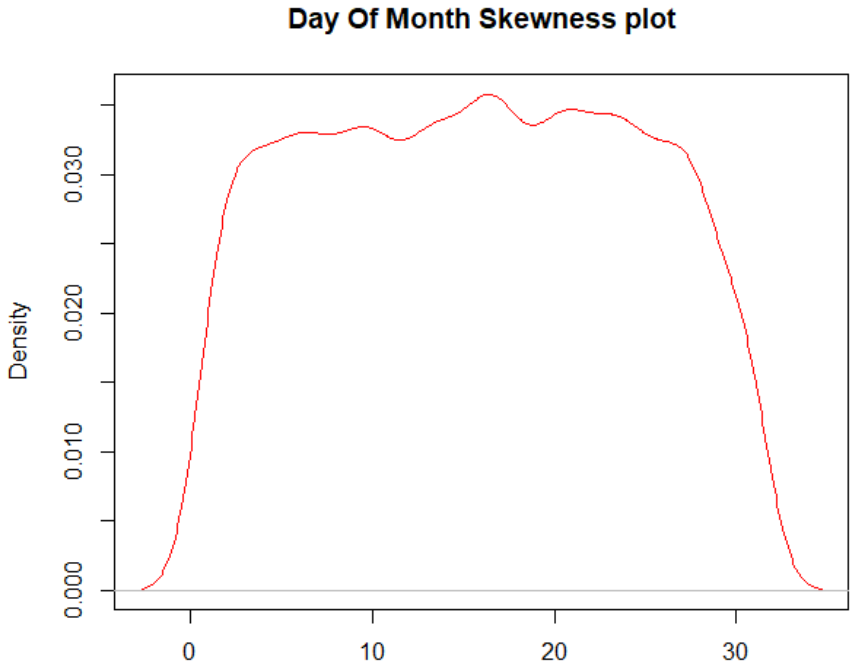
Mean	20.12
Median	18.00
Minimum	0.00
Maximum	89.00
Range	89
Trimmed	18.88
N (total number of variables)	10282
1st quartile	9.00
3rd quartile	29.00
Standard deviation	0.55
Mean Absolute Deviation	14.83
Standard error	0.01
Skew	0.71
Kurtosis	0.86



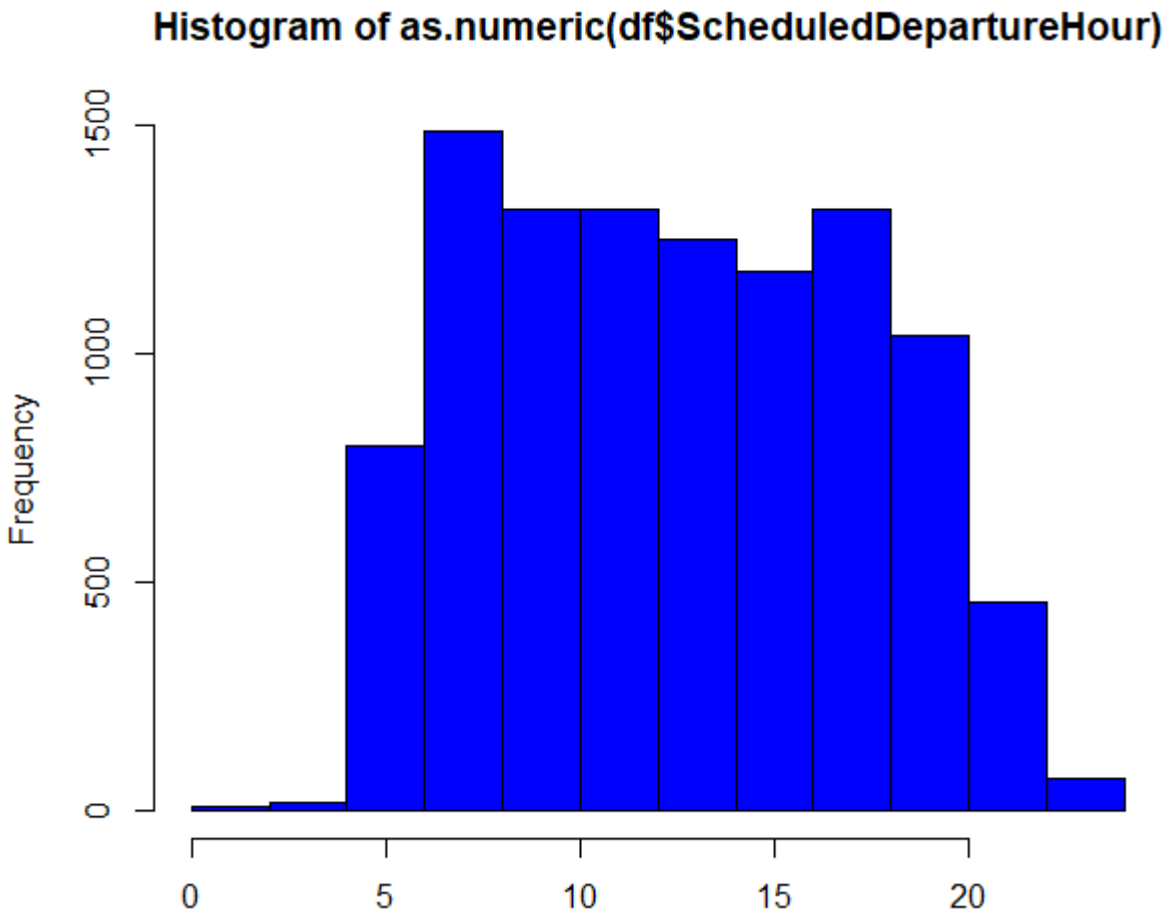


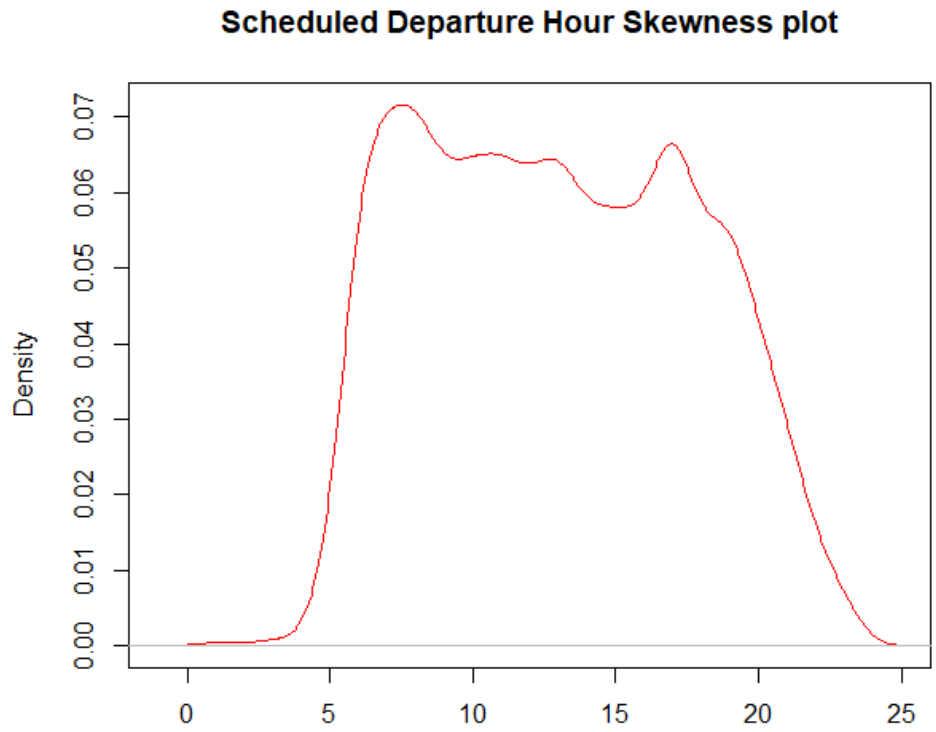
c. Days of the month the airlines run:





d. The scheduled departure hour of the flight:

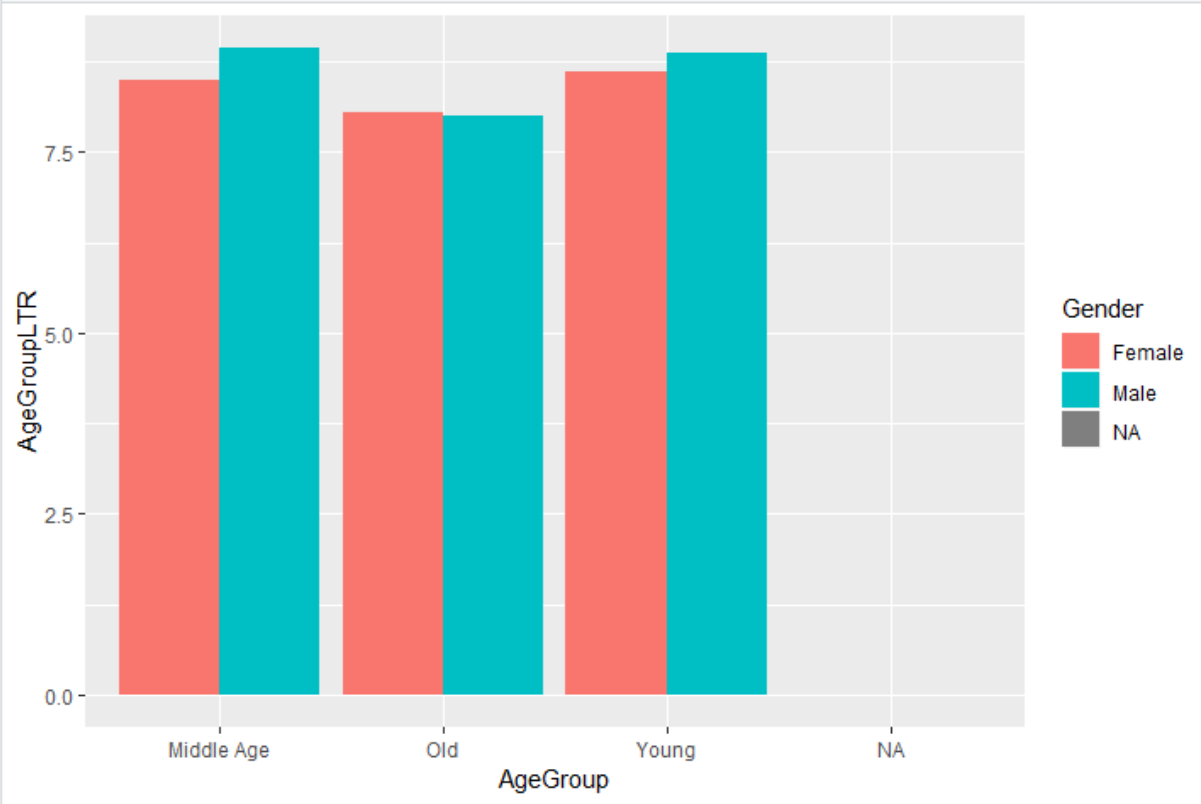




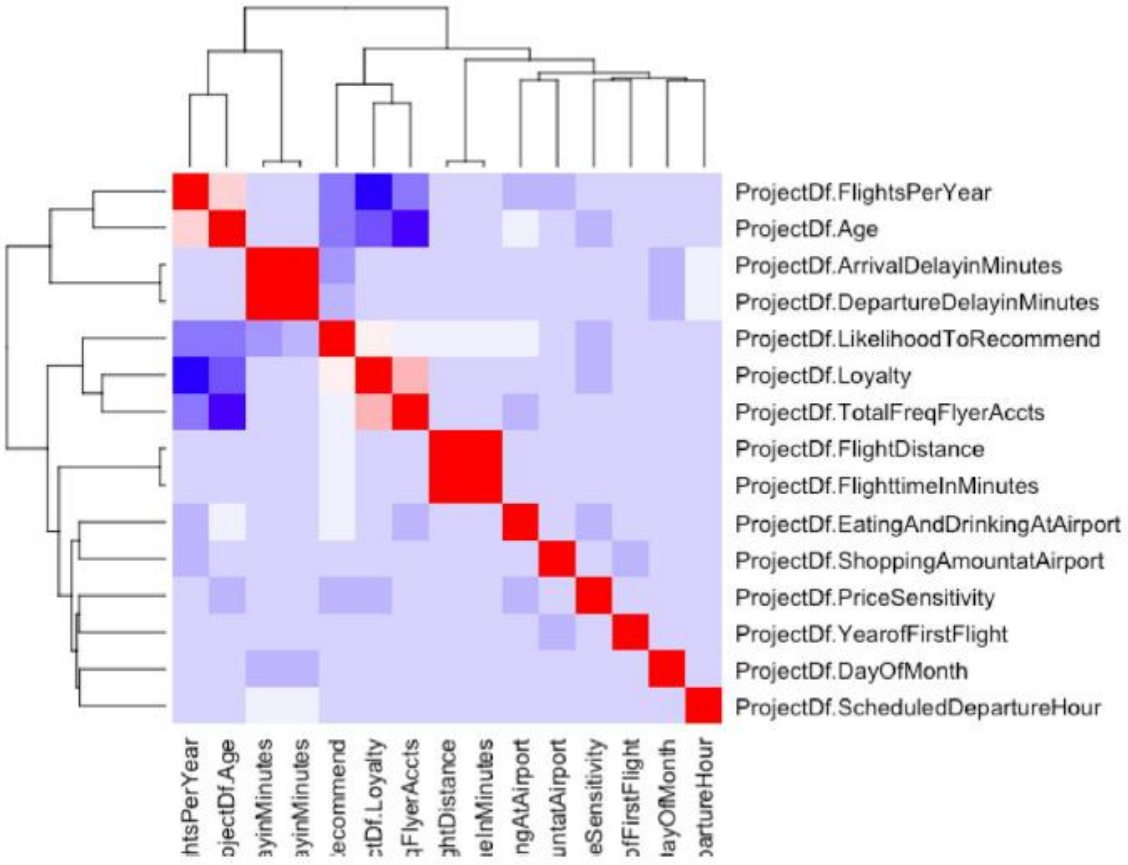
Data Visualization using Maps and Plots:

Plots are used to help understand the data given. The first plot is 'Age Group' created using 'Age' column.

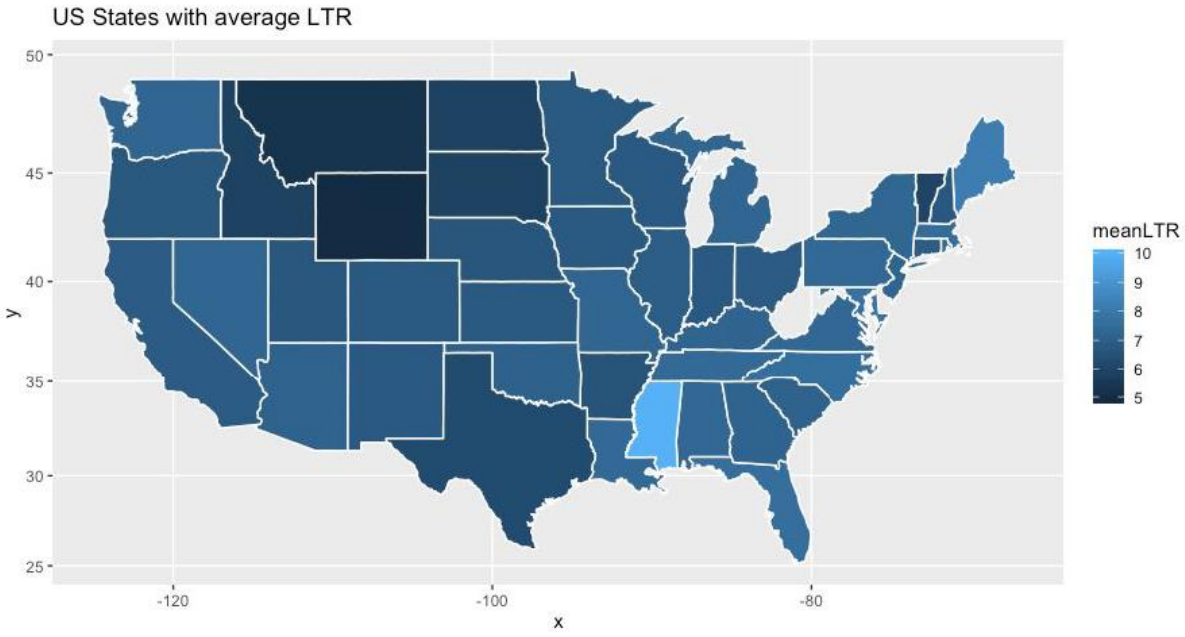
This plot helps us understand how 'Age' and 'Gender' affects the likelihood to recommend. Firstly, we divide people into young, mid and old in three age groups. We can measure the probability of both men and women in each age group. The younger age group is more likely to recommend compared to other age groups, while the old age group is less likely to suggest. In addition, male clients would be more likely than female clients to recommend our customer company to others would.



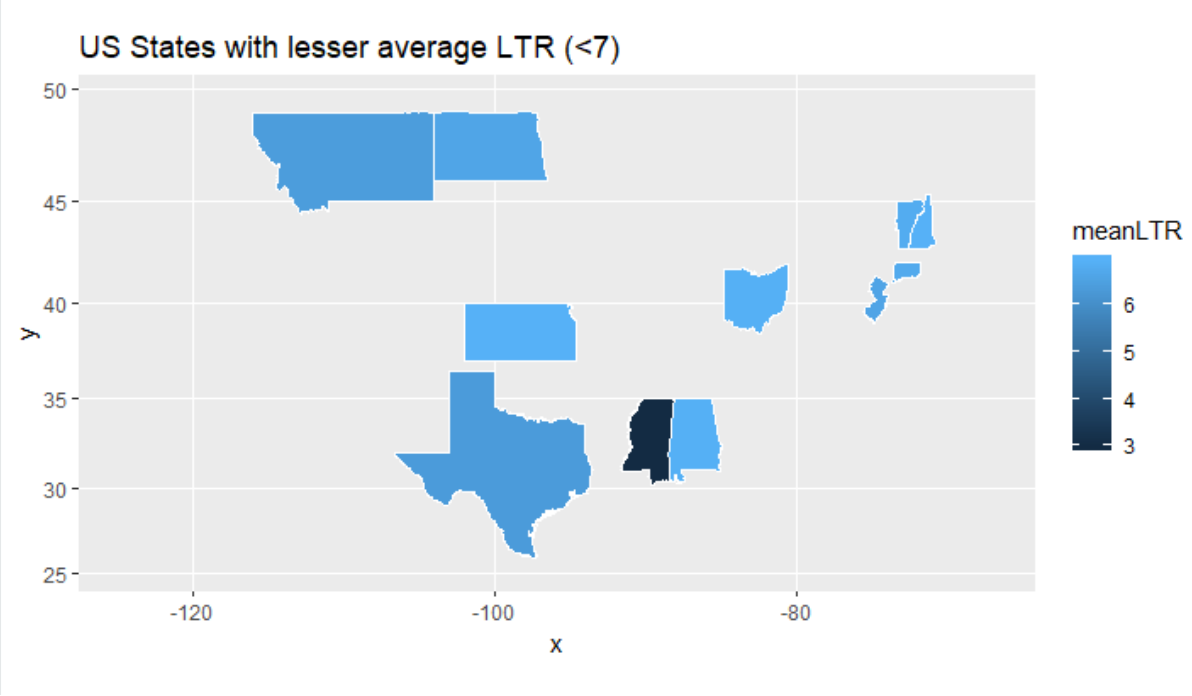
The second plot is a Heat Map. The correlation between each attribute is given. In heat maps, the correlation is more for brighter colours. Blue colour is for positive correlation white is for negative correlation. From this graph, we can see that loyalty and the number of flights that each customer has taken have stronger positive correlation, along with 'Age' and 'Frequent Flyer Accounts' the customer has.



The next one is the Average Likelihood to Recommend for each state:

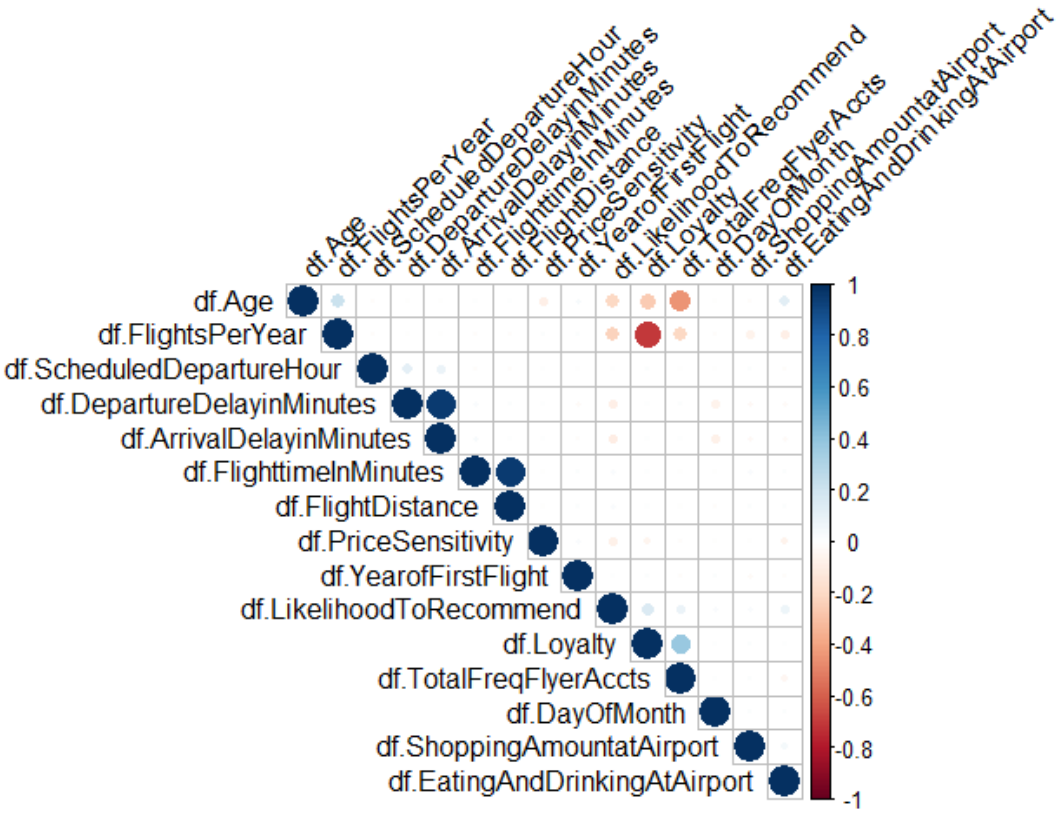


The next plot is the average Likelihood to Recommend < 7:



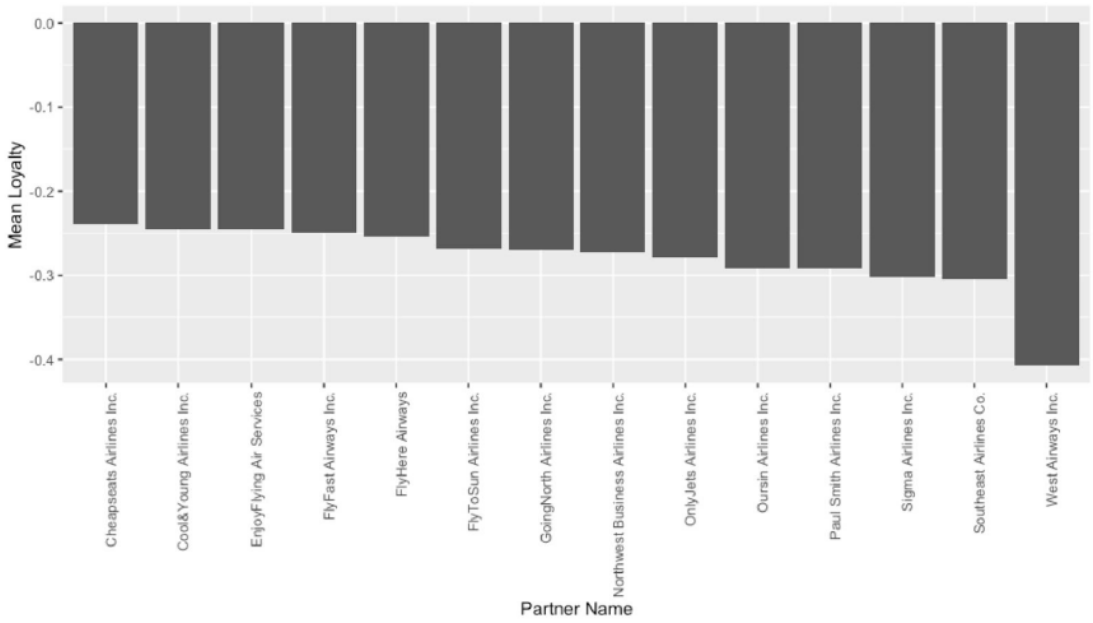
Correlation Matrix:

Also in a correlation matrix, I want to present this correlation. It not only shows the direction to which our attributes correlate, but also shows the magnitude. It is therefore more obvious to show our results. While blue is positive, red is negative. Blue is positive. The darker the colour, the stronger it is. We see that the link between time to reach the destination in minutes and distance of the departure to the arrival destination is strong among such correlations; the correlation is also strong between time delay of departure in minutes and delay of arrival in minutes, both approximately more than 0.8. However, every customer's association between loyalty and the number of flights is the lowest, approaching -1.



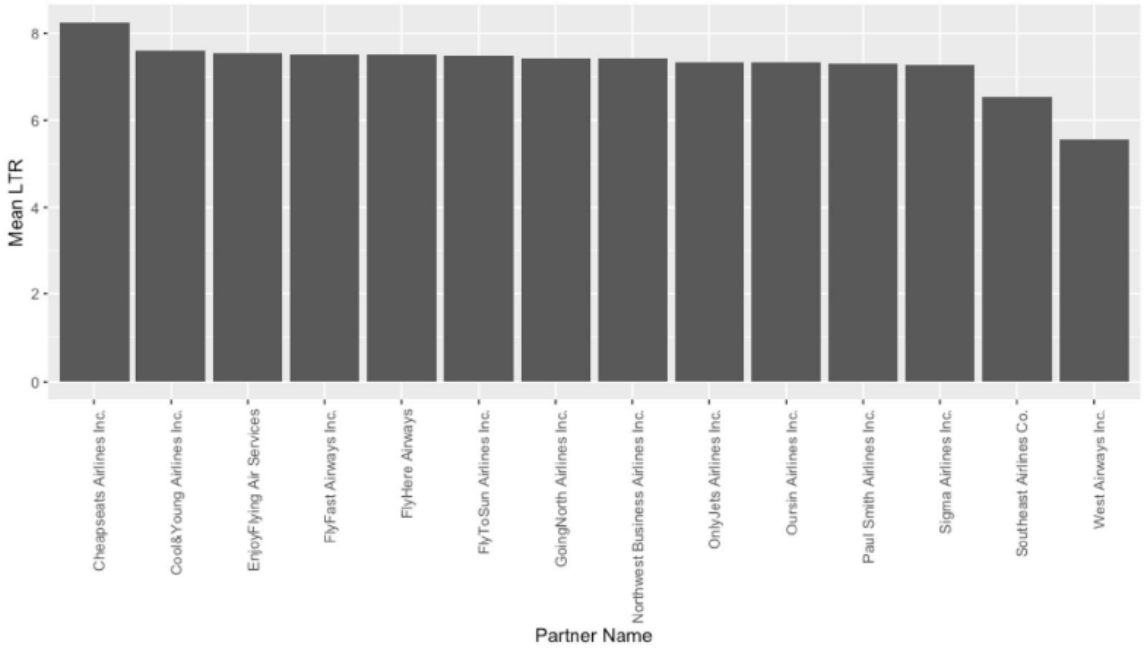
Loyalty Plot:

For a joint carrier, the loyalty plot reveals how many flights other aviation firms relative to those operated by that airline take. The graph reveals that the index is negative for all partner airlines, which means that customer satisfaction is not so high for these airline firms. West Airways has the least loyalty to customers among these firms, while Cheapseats has a fairly high loyalty. Generally, the loyalty of our customers does not vary a lot from those airlines.



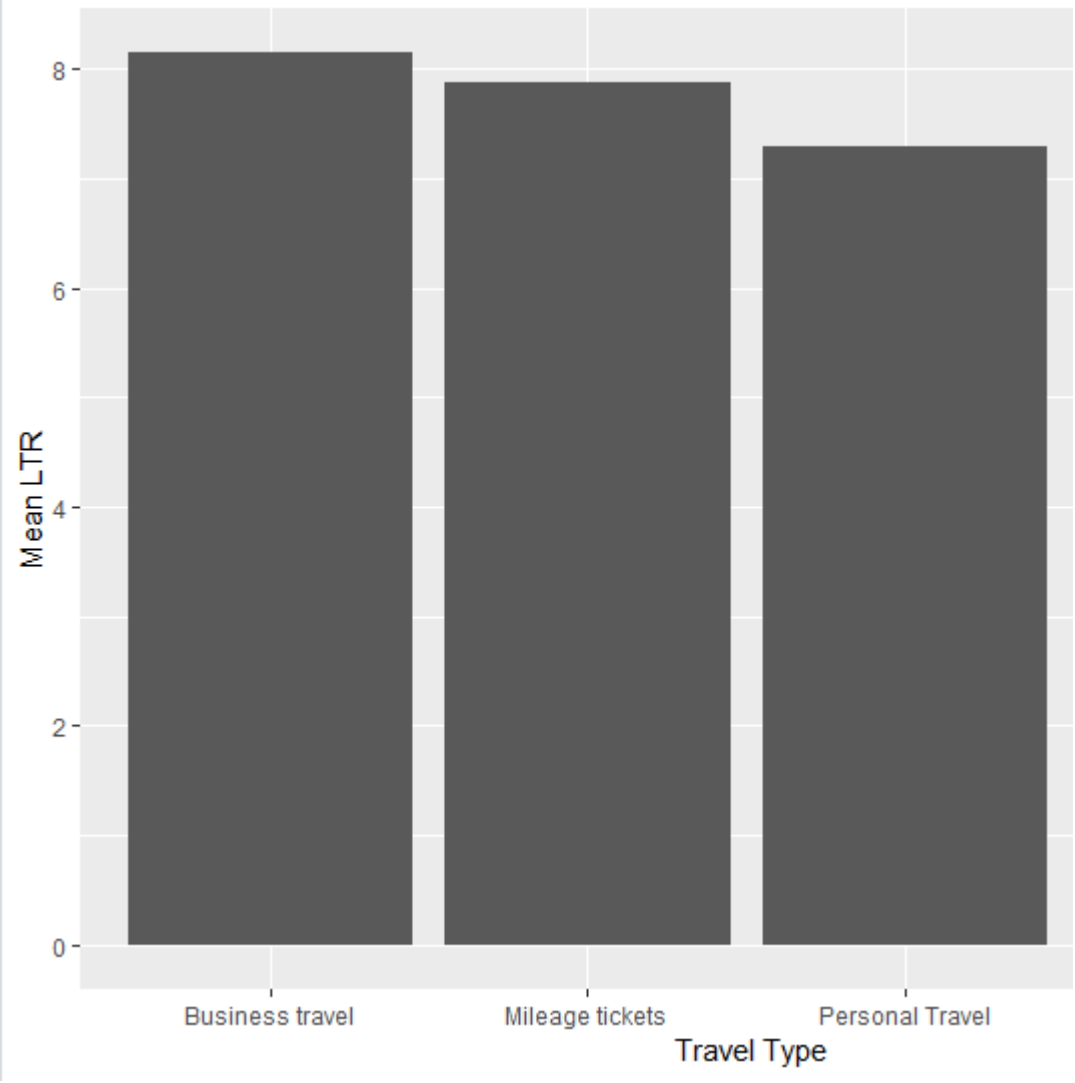
Mean LTR Partner:

This graph displays a customer's average chance that a certain affiliate airline would recommend to another. They assess the willingness to prescribe by counting 1 to 10, 1 of which is less likely to suggest and 10 are highly recommended. The chart indicates that consumers are more likely to prefer airlines that are Cheapseats but that they are less likely to recommend West Airways. This discovery supports our former loyalty story. As the consumer satisfaction of West Airways is low, it is not uncommon for customers to request it less frequently. In comparison, the chance score for Southeast airlines, which is about 7, is relatively low. Apart from the bottom 2 and the top airline, the ratings of the other are almost the same.



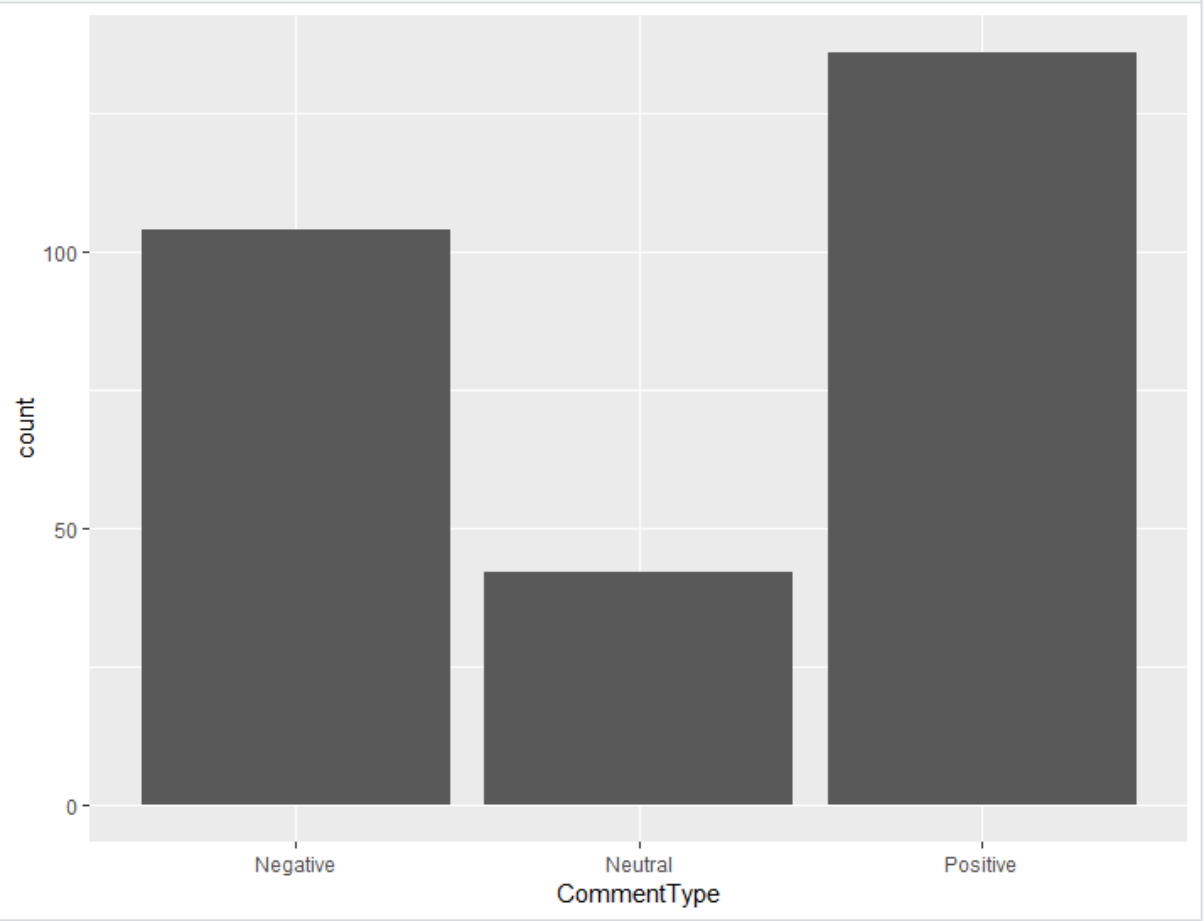
Travel LTR:

Present the possibility to indicate a particular form of travel to airlines from another point of view. The chance varies between the three main forms of flight. The most likely advice for business travel is around 8. Business travel. Yet with the ranking a little more than 5, personal travel is more likely to recommend. Mileage tickets are in the middle and corporate travel scores are very similar to them. The reasons for this may be that the costs should be reimbursed on work visits and that consumers should spend extra for a quality service. Consequently, they had a better travel experience than passengers did on their personal trip.



Text Mining:

I evaluated each customer's feedback and graded as optimistic, supportive and negative reaction to his or her commentary. This graph shows that the majority of words are positive, a limited number of words are neutral, and the remaining words are negative. Moreover, the general mood of the comments can be rated as positive.

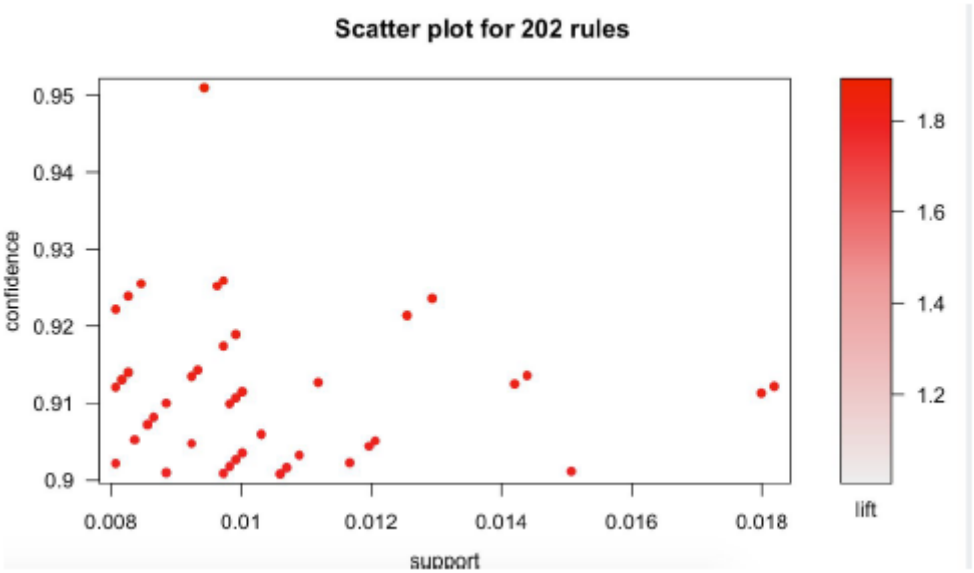


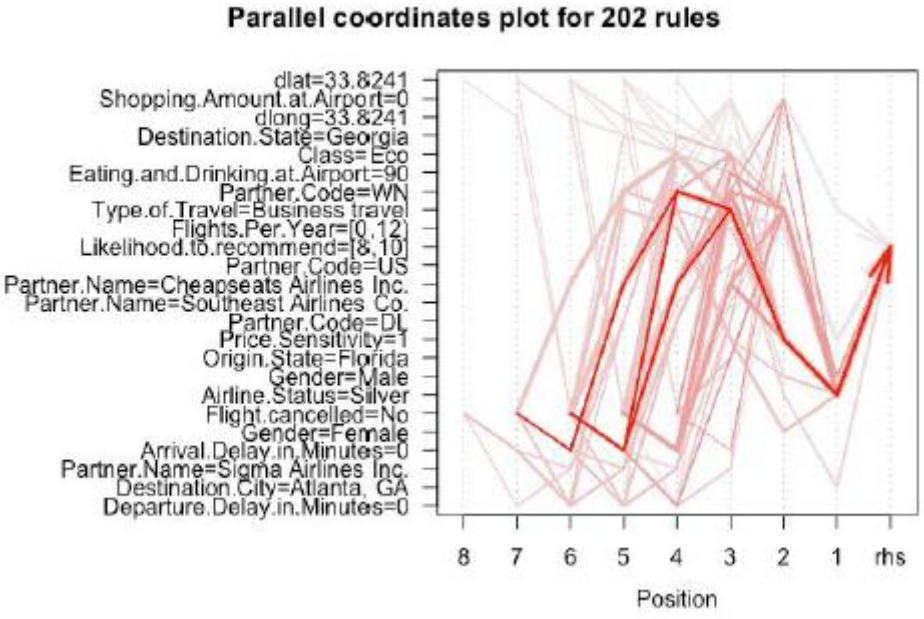
Modelling:

Association Rules Mining:

The method used to find relationships in the dataset between attributes. Help and trust power the pattern. For seeking the right connection, we used our help as 0.008 and 0.9. I used categorical imperatives and found that the column variables that may contribute to a developer are:

- Flight cancelled = No
- Arrival Delay = 0
- Partner Name = Cheapseats Airlines Inc.
- Type of Travel = Business travel
- Price sensitivity = 1 and
- Airline Status = Silver





Support Vector Machines:

SVM modelling deals with a model based on multiple regression approaches and SVM finds the right one to match and is trained automatically. We split the dataset into training and evaluation data sets in which we train the first and the second.

I divided the Likelihood to recommend into two types:

- Promoter: Value greater than seven (P>7).

- Detractor: Value less than seven (P<7).

```
> svmOutput <- ksvm(Emotion ~.,data = df1, kernel = "rbfdot", kpar = "automatic", C=5, cross = 3, prob.model = TRUE)
> svmOutput
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 5

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.486868516943466

Number of Support Vectors : 305

Objective Function Value : -169.9709
Training error : 0
Cross validation error : 0.001073
Probability model included.
```

```
> table1

FALSE  TRUE
 1534  1894
> Result <- (table1[1]/(table1[1]+table1[2]))*100
> Result
   FALSE
44.74912
```

Since the accuracy is less than 50%, this model cannot be used for analysis.

Linear Modelling:

- **Univariate Modelling:** Each data attribute was mapped to 'Likelihood to Recommend' and we find out which attributes are statistically important 'Likelihood to Recommend' predictors.

  We also checked whether the resemblance of the suggestion relies on statistically significant predictors. Instead, by evaluating R-squared quantities, we measured how much it depends on them. The size of the spectrum of the R-squared value is between 0 and 1, which means that the Probability to Suggest attribute does not depend on the predictor, even if the R-squared value is 1. 'Likelihood to Recommend' is mostly dependant on type of travel and status of flight.
- **Multivariate Modelling:** I mapped the 'Likelihood To Recommend' with multiple attributes to find the significant predictors.

Actionable Insights:

- Improvement of in-flight services especially for Female and Older people as they had experienced more or less discomfort during their journey.
- There are approximately 25 having less than average likelihood to recommend of 7, which shows the areas of improvement to be made use of.
- Text mining and Descriptive Statistics prove that the loyalty is coming down and the two partners responsible for it can be fired.
- The Association rules mining shows us the attributes of customer where the airlines can continue to be focused to sustain those customers and reduce customer churn.
- The status of flights if cancelled can be communicated in a better way so as it increase the rating given by customers.
- The airlines must focus more on the business passengers for better rating. Marketing and sales have to make sure of that.
- The Silver airline status passengers play a major role in promoters of the company. So special offers and discounts can make them loyal to Southeast airlines.
- The 'Age Group' plot shows that the passengers between the ages of 30-60 have a higher likelihood to recommend and must be targeted to increase sales.