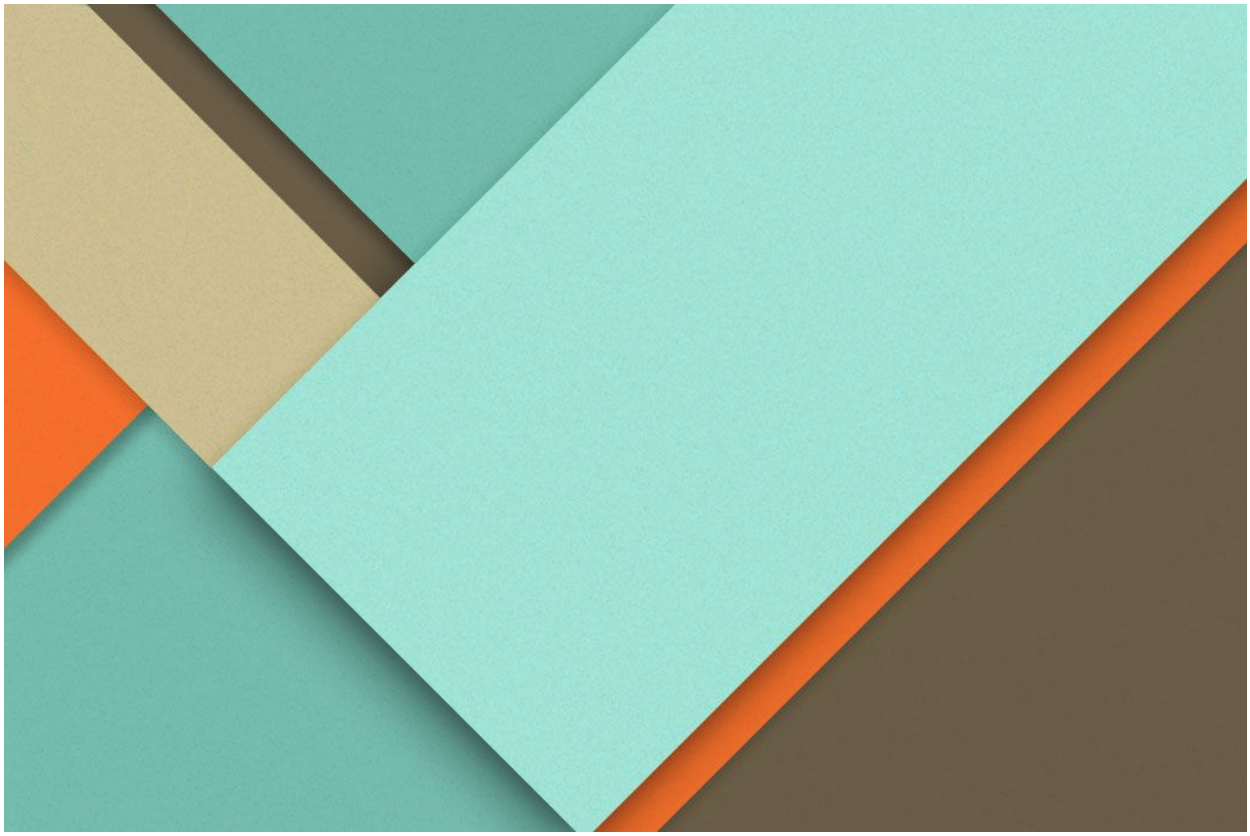


Customer Statisfaction Analysis of Southeast Airlines

Group A

Mahesh Kumar Badam Venkata

Ankit Mohite

Aman Awana

Yashila Upendra Kumar

Sripad Amol Laddha

Yucan Dai

Data Overview

There are two primary problems Southeast Airlines faced. For one thing, the loyalty program which was considered the best practice of the industry does not work well in today's market. For another, the most commonly used indicator, customer churn, is lagging. It cannot avoid having the customers leave and prevent the loss. So we want to figure out some more efficient and leading indicators to help us make predictions before we lost customers.The survey dataset we analysed gave us a lot of information. It contained thousands of observations of flight segment data. Each row represents one flight segment for a specific customer. Each column represents an attribute of that particular flight segment. After we clean the dataset, there are 27 attributes in our dataset, and we divide them into 3 categories, airline attributes, customer attributes, and location attributes. We can check the table below to better understand each attribute.

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

Our objective was to provide suggestions for best practices to reduce customer churn for Southeast Airlines.

Business Questions

- How important is age and gender as a factor to determine customer's airline rating and travel frequency ?
- How to use a customer's Likelihood to recommend an airline as a factor to correlate LTR to other factors such as age,gender, origin city,destination city ,comments etc..
- How to use the comments and feedbacks by customers to determine if their responses are positive or negative in order to understand if the customer is likely to use the airline again.
- How can we properly analyze the diverse set of customer base that the airline possesses.
- Finding out the source states having less likelihood to recommend.

Data Cleaning

We read data from the json file first and then start data cleanse.

Firstly, we use trimws function to remove leading and trailing whitespace of each column. Then we rename each column to make them more understandable by using colnames function. Furthermore, we separate the city name and state name in both destination and departure place by using separate function. After that, we found that in column "Comments", there is a lot of NAs. To solve this problem, we convert the argument to character type. Using is.na function to find missing values and assign "No comments" to these values.

Next, we leverage the which(is.na) to identify the location of NAs in all columns and arrange 0 to the NAs.

As for now, we have done the data cleaning part. Then we add several new attributes to the data frame to better process the data. They are Arrival Delay Greater Than 5 Minutes, Mean Travel Time, Mean Departure Delay Time and if it's Long Duration Flight.

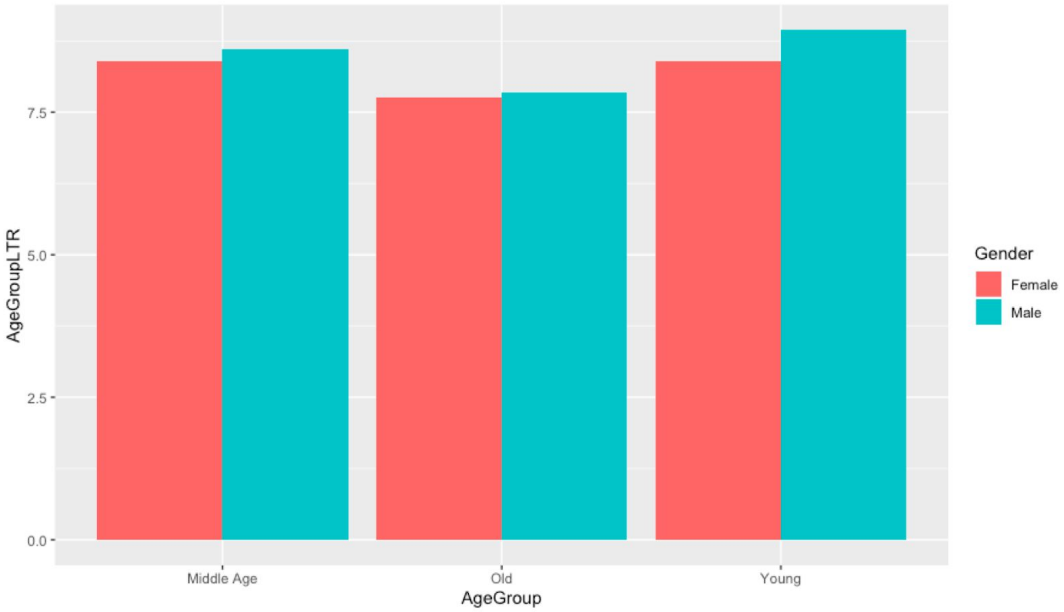
Data Visualization

In order to present clear and intuitive results to our audience, we then create several plots.

The first plot is Agegroup LTR.

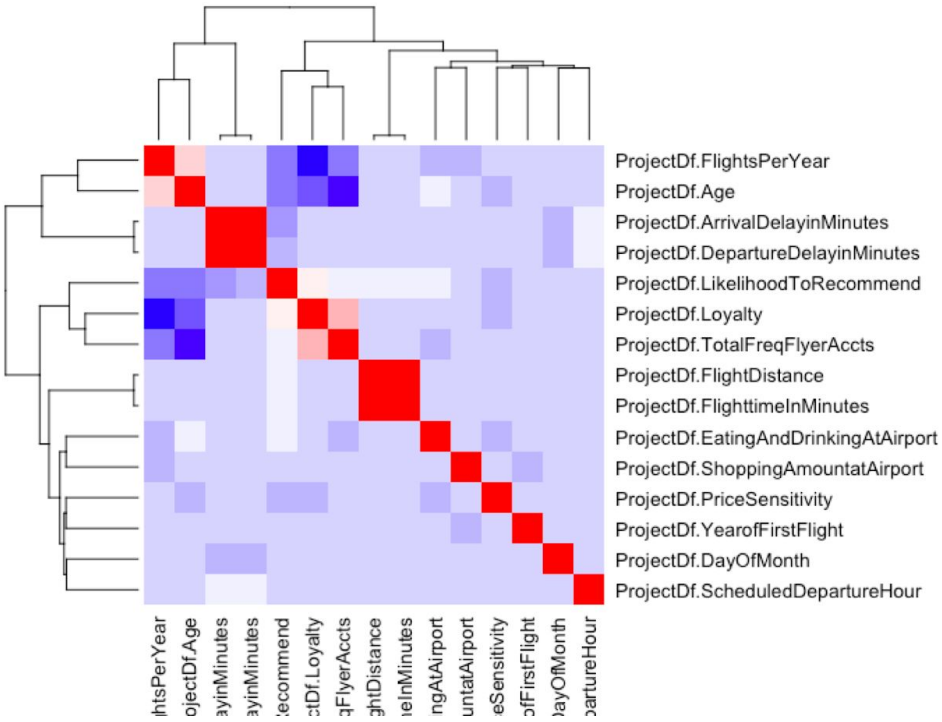
This plot demonstrates the likelihood to recommend for different age groups as well as the gender. From the graph, we can see that gender and age groups both will affect likelihood to recommend Southeast Airlines.

Firstly, we divide people into three age groups, young, middle, and old. In each age group, we calculate the likelihood of male and female respectively. Compared to other age groups, young age group tend to have higher likelihood to recommend, while old age group is less likely to recommend. Furthermore, male customers are more likely to recommend our client company to others than female customers.

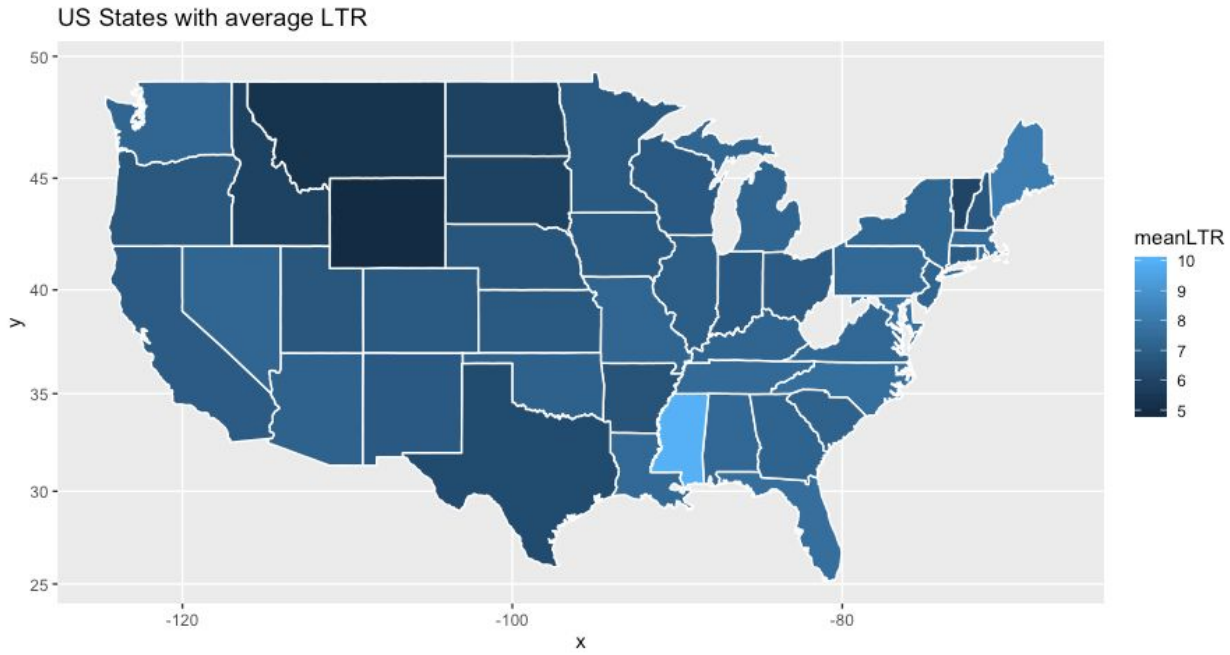


The second plot is Heat map.

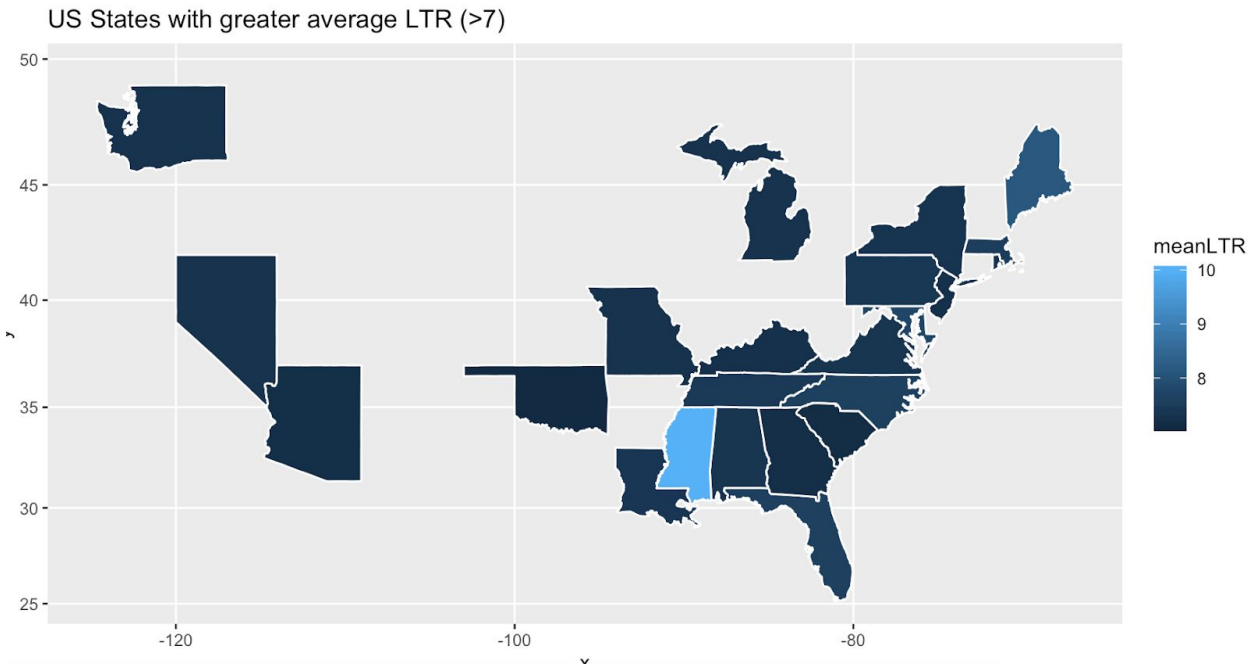
We present the correlation between each attribute in a heat map. In heat map, the brighter the color the stronger the correlation. And blue refers to positive correlation while shades of white refers to negative correlation. From this graph, we can reach the following conclusion, loyalty and the number of flights that each customer has taken have stronger positive correlation, so does age and frequent flyer accounts the customer has.



The third plot is Average Likelihood to recommend for each state:



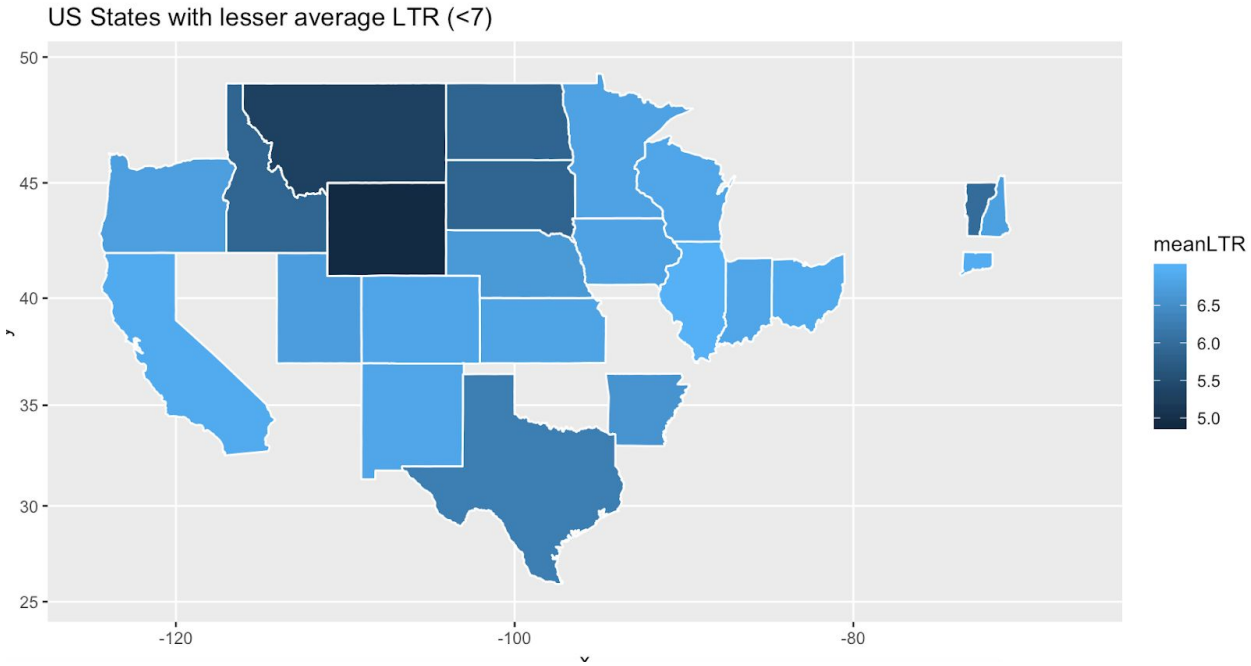
The average LTR plot of average greater than 7:



ALASKA	MISSOURI	SOUTH CAROLINA	KENTUCKY	MASSACHUSETTS
NEVADA	MISSISSIPPI	NORTH CAROLINA	MICHIGAN	MARYLAND
ARIZONA	ALABAMA	VIRGINIA	MAINE	RHODE ISLAND
OKLAHOMA	GEORGIA	PENNSYLVANIA	NEW YORK	DELAWARE
LOUISIANA	FLORIDA	TENNESSEE	NEW JERSEY	

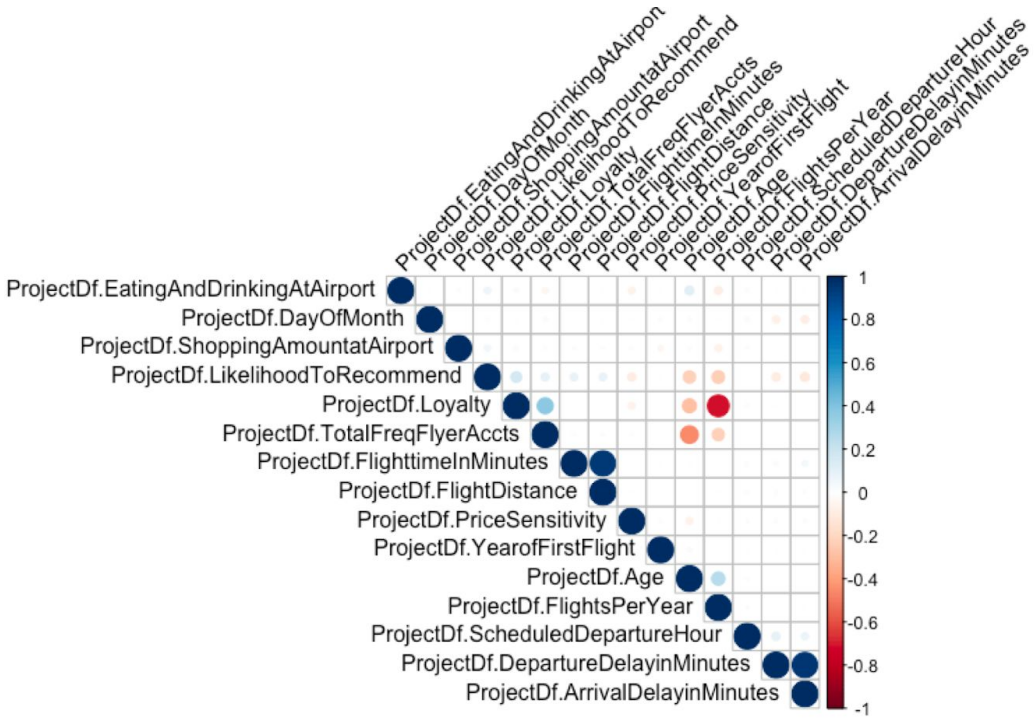
The average LTR plot of average less than 7:

CALIFORNIA	NORTH DAKOTA	MINNESOTA	IOWA	INDIANA
OREGON	SOUTH DAKOTA	NEBRASKA	COLORADO	OHIO
MONTANA	UTAH	KANSAS	MINNESOTA	VERMONT
IDAHO	NEW MEXICO	WYOMING	WISCONSIN	NORTH HAMPSHIRE
MONTANA	COLORADO	ARKANSAS	ILLINOIS	CONNECTICUT



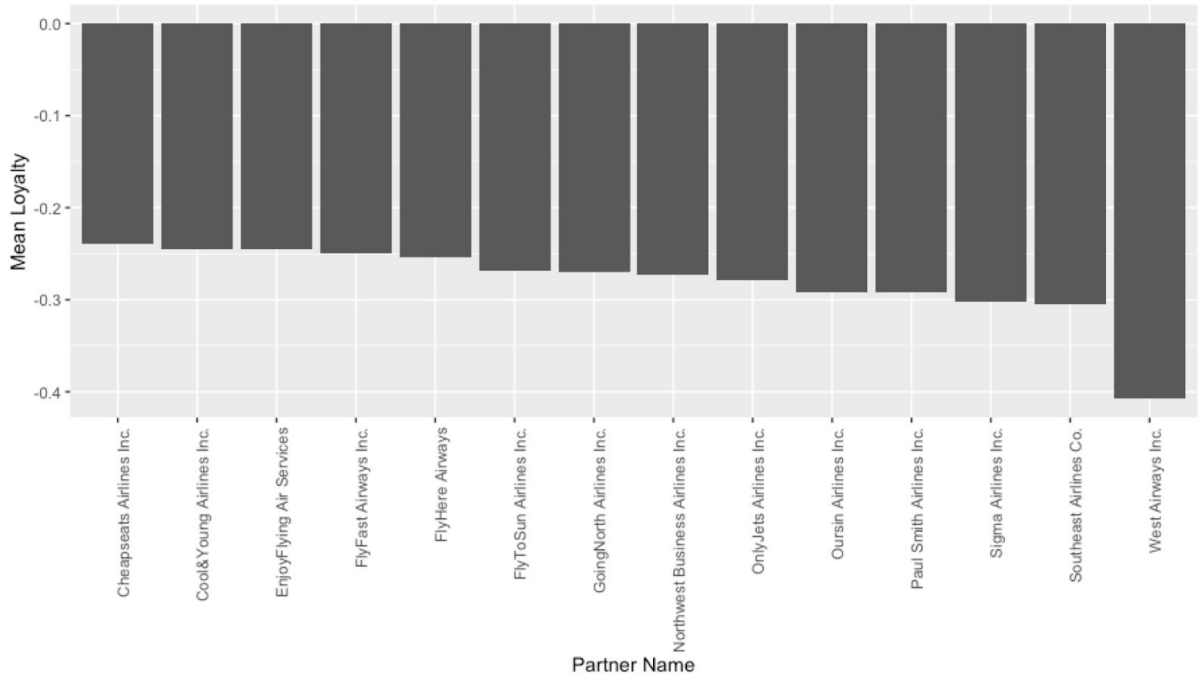
Correlation Matrix:

We present this correlation also in a Correlation Matrix. It not only displays the direction of the correlation between our attributes, but also shows the magnitude. So it's more clearly to exhibit our outcome. Blue represents positive correlation while red represents negative correlation. The darker the color is, the stronger the correlation is. We can see that among those correlations, the correlation between the length of time, in minutes, to reach the destination and the distance between the departure and arrival destination is strong, so is the correlation between Departure Delay in Minutes and arrival delay in minutes, both of them are approximately more than 0.8. However, the correlation between loyalty and the number of flights that each customer has taken is the weakest, approaching -1, which fit with our common sense.



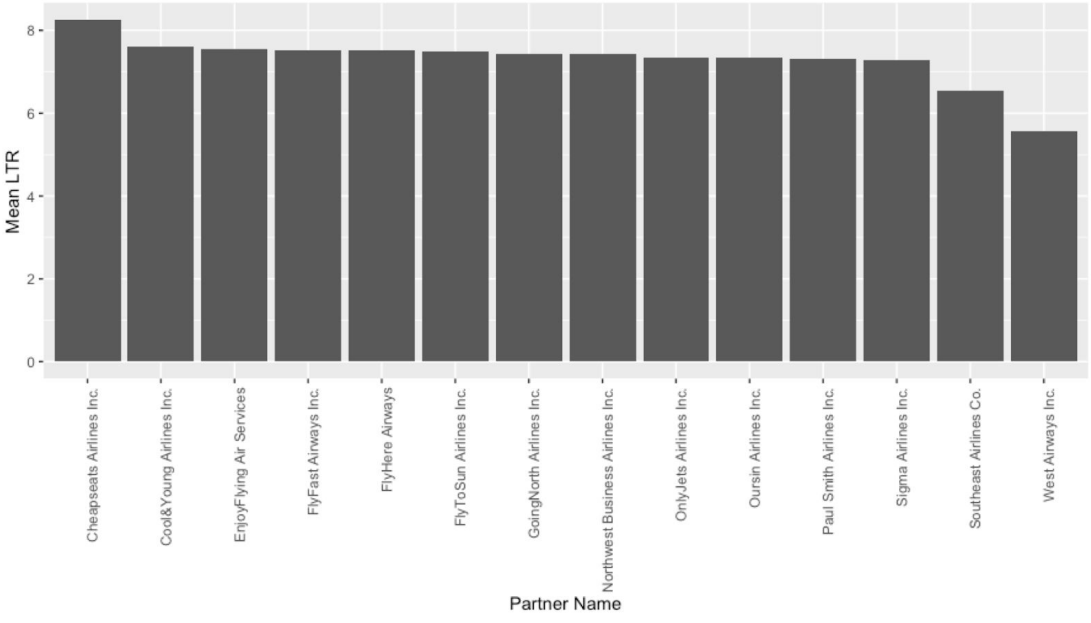
Loyalty plot:

Loyalty plot shows us for each partner company, what the proportion of flights taken on other airlines versus flights taken on this airline. It can be seen from the graph, the index for all partner airlines is negative, which means customer loyalty for these airlines is not so strong. Among those companies, West Airways has the weakest customer loyalty, while Cheapseats airlines has a relatively stronger customer loyalty. All in all, customer loyalty does not have much difference among those partner airlines.



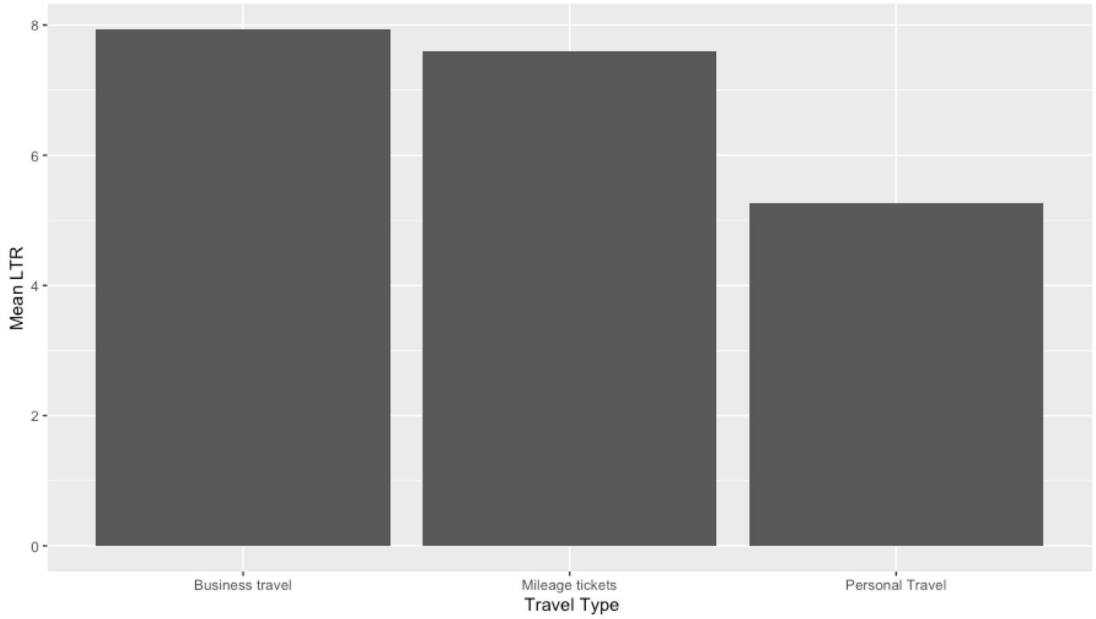
Mean LTR Partner:

This graph displays the average likelihood for a customer to recommend a certain partner airline to others. We measure the will to recommend by numbering it from 1 to 10, of which 1 represents less likely to recommend while 10 means strongly recommend. From the graph, we can conclude that customers are more likely to recommend Cheapseats airlines but less likely to recommend West Airways. This conclusion confirms our previous one from the loyalty plot. Since West Airways has the weakest customer loyalty, it is not surprising that customers are less likely to recommend it. On top of that, the likelihood score is also relatively low for Southeast airlines, which is rated around 7. Besides the top 1 and bottom 2 airlines, the ratings for other airlines are almost the same.



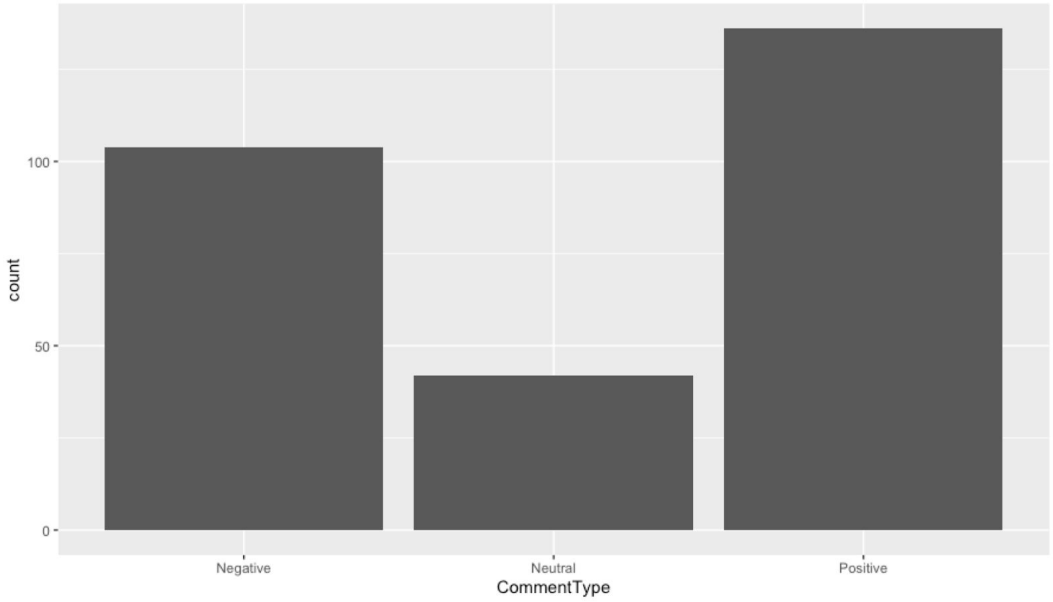
Travel LTR:

Let's see the likelihood to recommend the airlines from a different perspective, travel type. This graph demonstrates the likelihood varies among the three travel types. Business travel has the highest likelihood to recommend, the rating of which is around 8. However, personal travel has the lowest likelihood to recommend with the rating a little bit more than 5. Mileage tickets is in the middle, and the rating is very close to business travel. The reasons behind this maybe in business travel, the expense can be reimbursed, so customers are willing to pay more to receive a better service. Therefore, they have a better flight experience than customers in personal travel, so they are more likely to recommend.



Text mining:

We analyzed the comments from each customer, and categorized the attitude for their comments, which are positive, neutral and negative. It can be seen from this graph that the majority of the words are positive, a small proportion of words can be considered neutral, the rest are negative words. So we can classify the overall attitude of the comments as positive.



Data Visualization

Descriptive statistics are used to summarize the quantitative values in R in order to analyse the data through mathematical statistics or visual easy-to-understand plots so that we can draw meaningful inferences about the data distribution.

The data frame ProjectDf 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, the Airline attributes .

Customer Attributes:

The descriptive statistics for Customer attributes can be seen as follows:

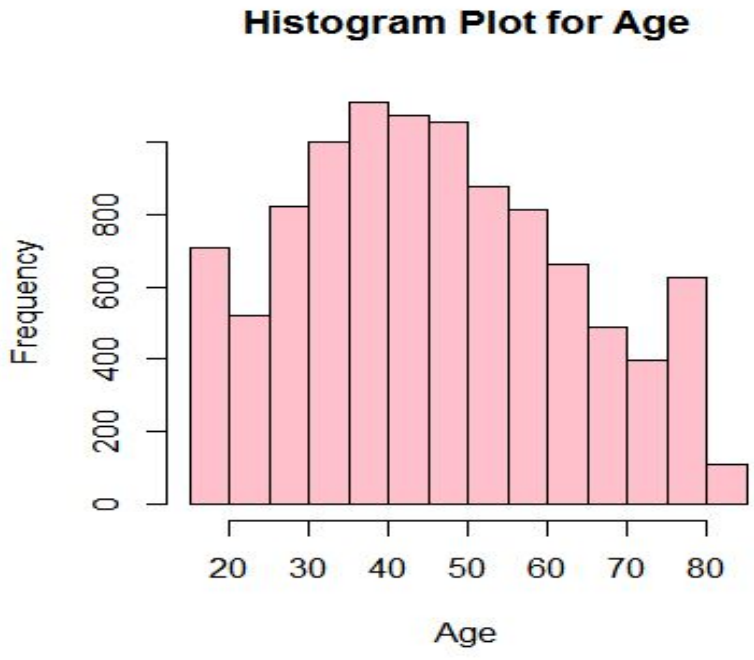
1) Age :

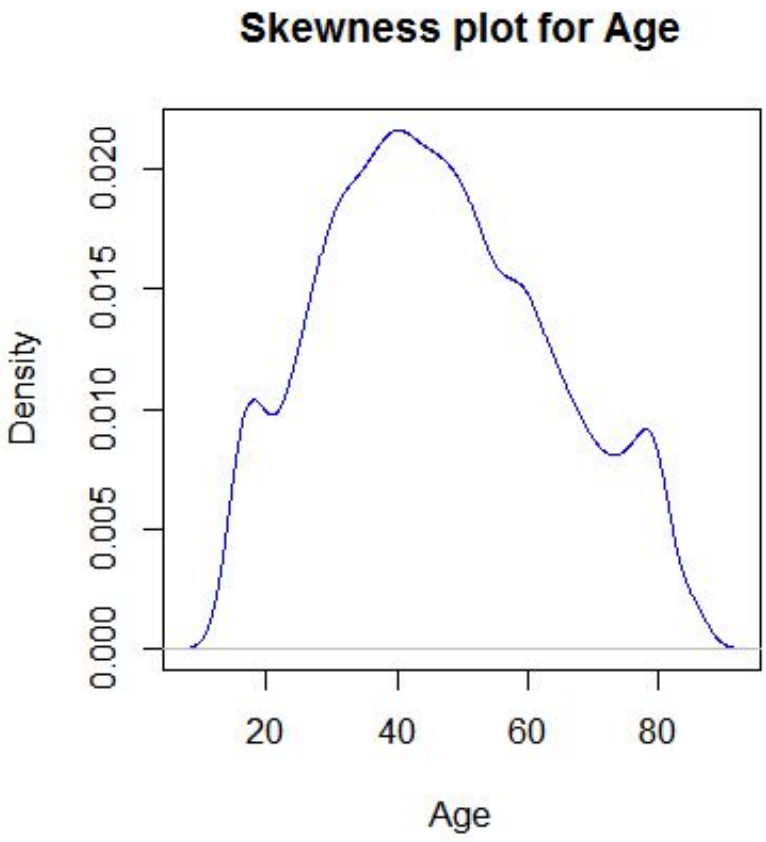
Code Snippet:

```
install.packages("pastecs") #installing package pastec
library(pastecs)
install.packages("psych") #installing package psych
library(psych)
describe(ProjectDf$Age)  #use of describe function to measure the
        statistics of the Age column data frame.
summary(ProjectDf$Age)   #use of summary function to measure the
        statistics of Age column in the data frame
hist(ProjectDf$Age ,col="pink")      # use of hist function to show the
        histogram distribution of Age column in the data frame.
plot(density(ProjectDf$Age),main ="Skewness plot for
        Age",xlab="Age",col="blue") #plotting the density distribution of
        Age coulmn in the dataframe to show the measure of skewness.
```

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





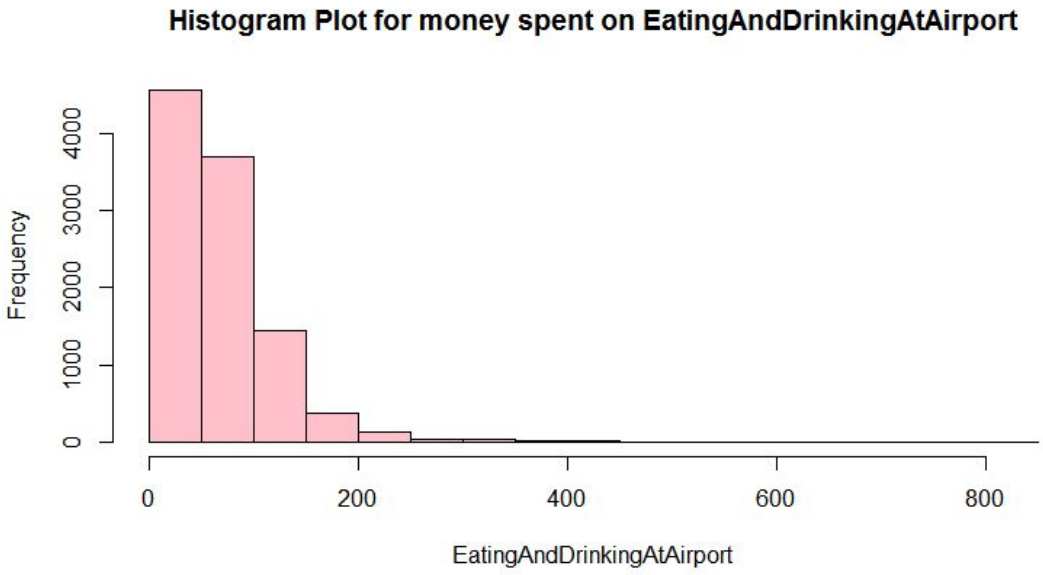
2) Money spent on Eating and Drinking at the airport.

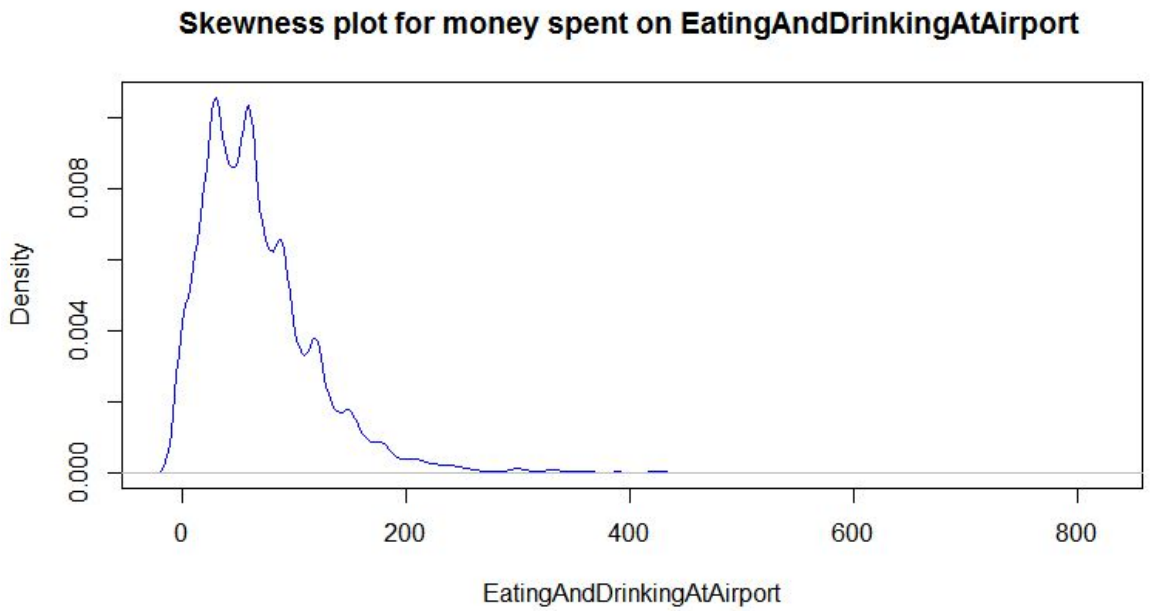
Code snippet:

```
install.packages("pastecs") #installing package pastec
library(pastecs)
install.packages("psych") #installing package psych
library(psych)
describe(ProjectDf$EatingAndDrinkingAtAirport)  #use of describe function to
measure the statistics of the Age column data frame.
summary(ProjectDf$EatingAndDrinkingAtAirport)   #use of summary function to
measure the statistics of Age column in the data frame
hist(ProjectDf$EatingAndDrinkingAtAirport ,main =" Histogram Plot for money
spent on EatingAndDrinkingAtAirport ",xlab =
"EatingAndDrinkingAtAirport",col="pink") # use of hist function to show the
histogram distribution of Age column in the data frame.
plot(density(ProjectDf$EatingAndDrinkingAtAirport),main ="Skewness plot for
money spent on EatingAndDrinkingAtAirport
",xlab="EatingAndDrinkingAtAirport",col="blue") #plotting the density distribution
of Age coulmn in the dataframe to show the measure of skewness.
```

Descriptive Statistics :

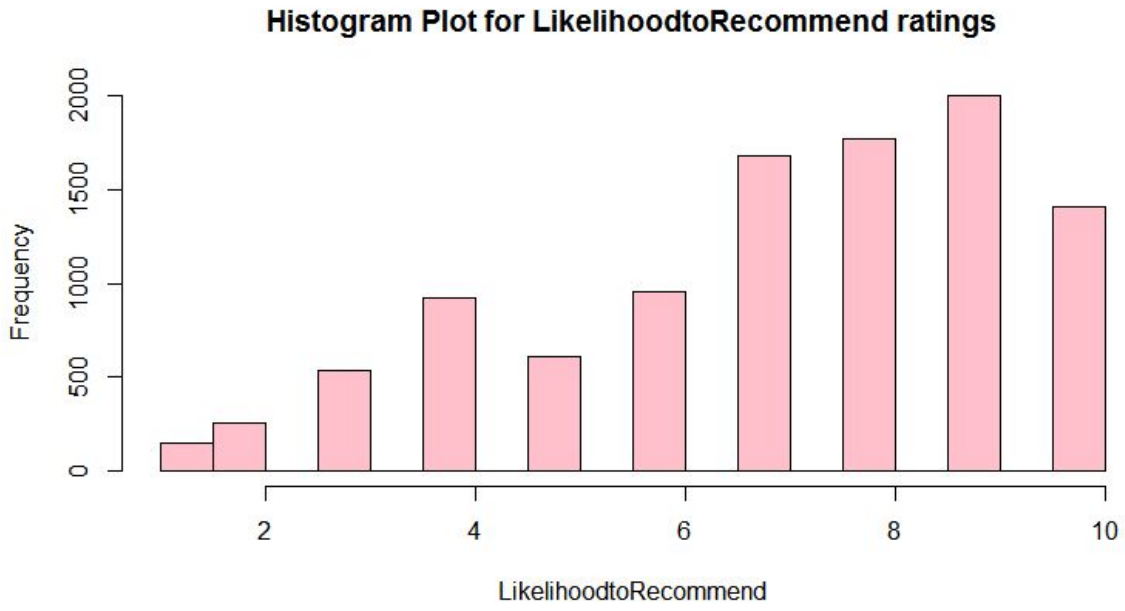
Mean	68.02
Median	60.00
Minimum	0.00
Maximum	805.00
Range	805
Trimmed	61.56
N (total number of variables)	10282
1st quartile	30.00
3 rd quartile	90.00
Standard deviation	53.58
Mean Absolute Deviation	19.27
Standard error	0.53
Skew	2.4
kurtosis	14.98

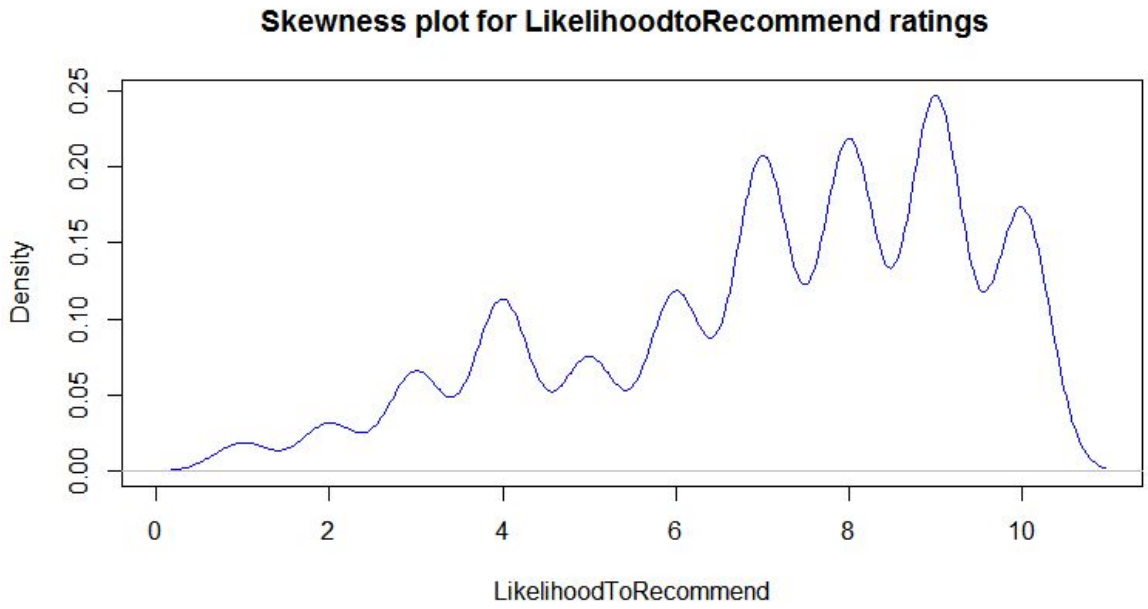




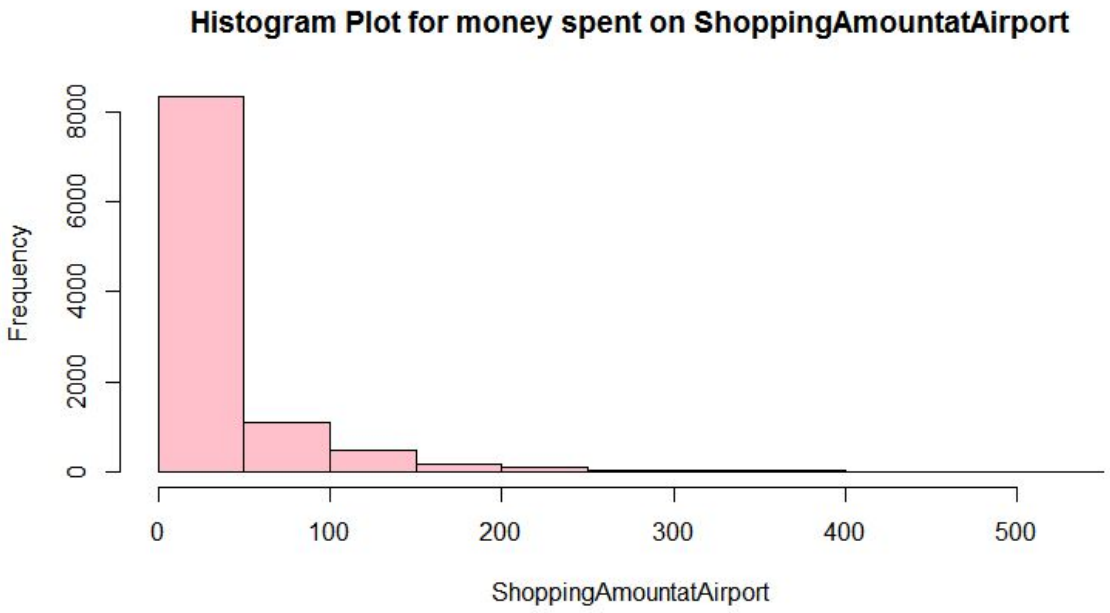
Similarly the Plots for other Customer Attributes can be seen as follows:

3)Likelihood to Recommend ratings given by the Customer:

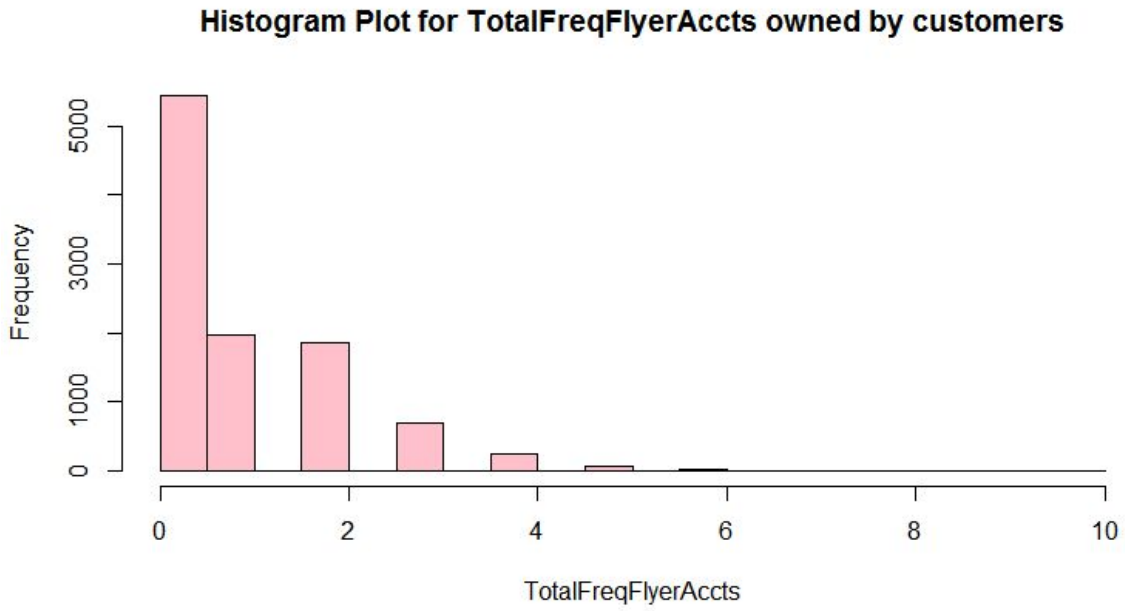


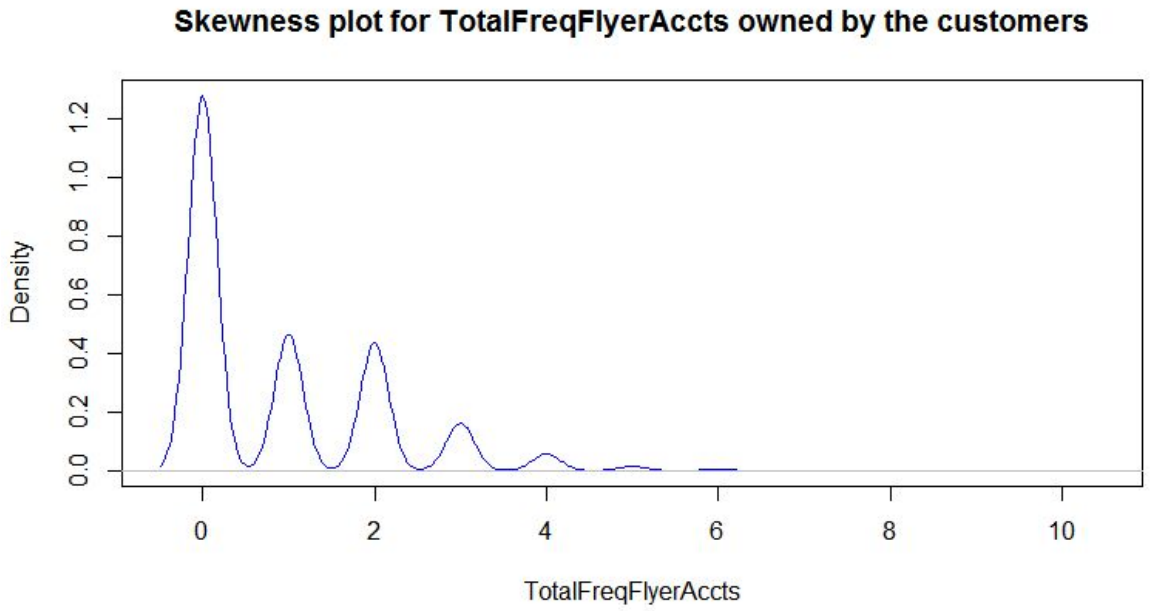


4) Money spent on Shopping at the airport by Customers:

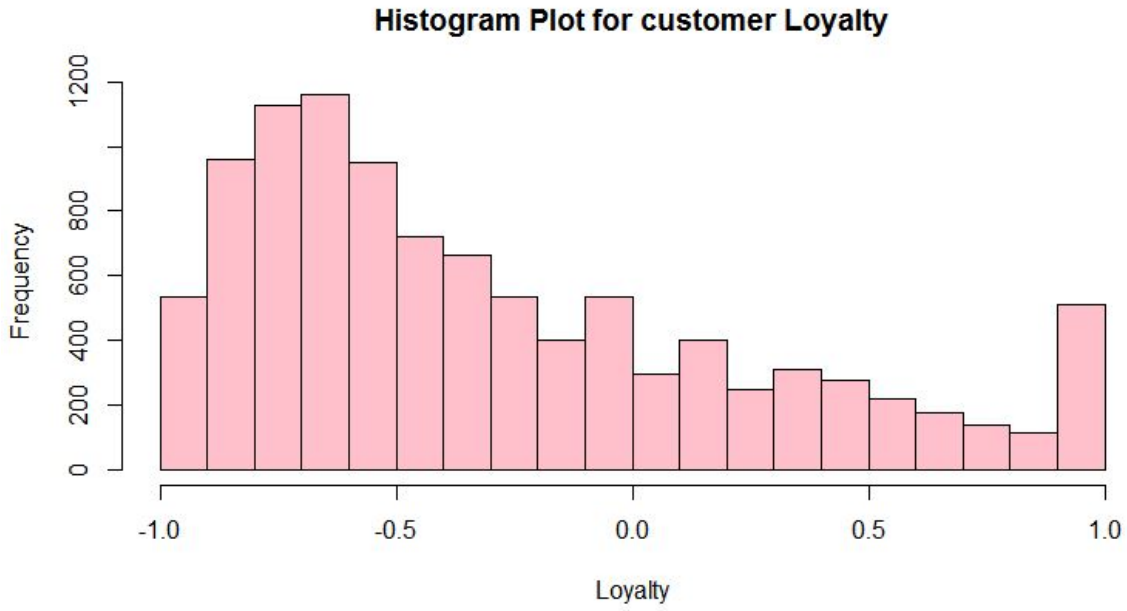


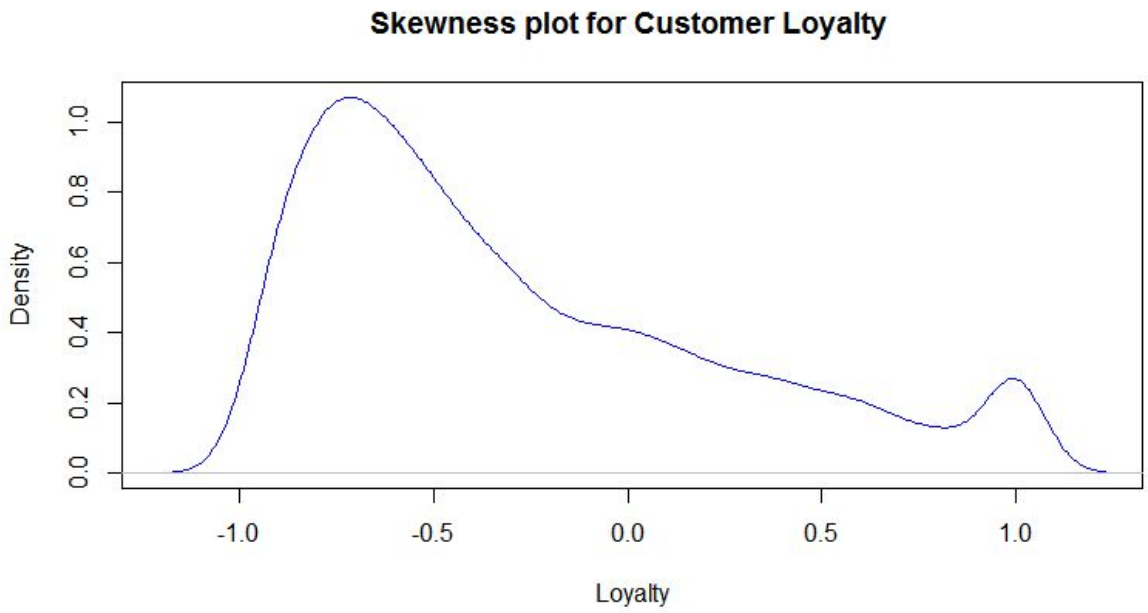
5) Total number of Frequent flier accounts owned by customers:





6)Customer Loyalty:





Airline Attributes:

The descriptive analysis for Airline attributes can be seen as follows:

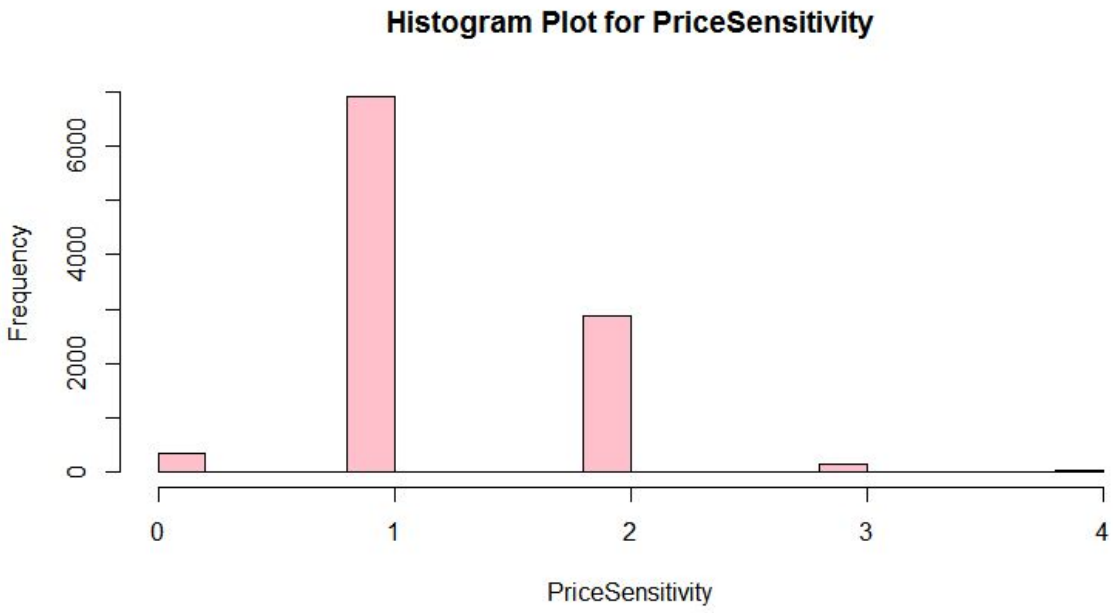
1) Price Sensitivity of the airlines:

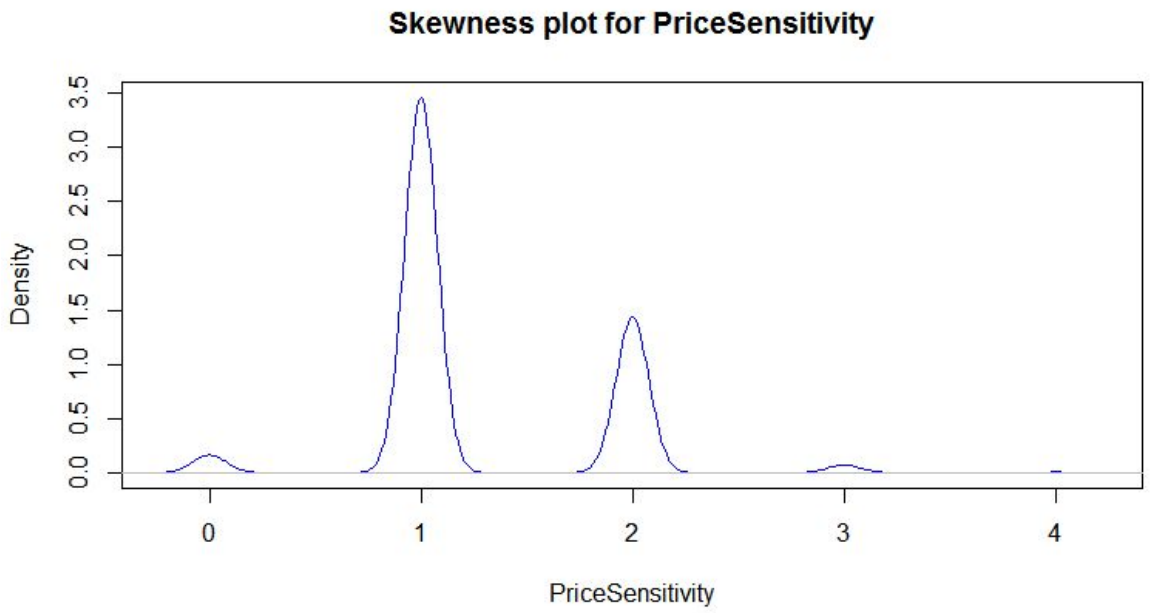
Code snippet:

```
install.packages("pastecs") #installing package pastec
library(pastecs)
install.packages("psych") #installing package psych
library(psych)
describe(ProjectDf$PriceSensitivity)  #use of describe function to measure the statistics of the PriceSensitivity  column data frame.
summary(ProjectDf$PriceSensitivity)   #use of summary function to measure the statistics of PriceSensitivity column in the data frame

hist(ProjectDf$PriceSensitivity ,main =" Histogram Plot for PriceSensitivity",xlab = "PriceSensitivity",col="pink") # use of hist function to show the histogram distribution of PriceSensitivity column in the data frame.

plot(density(ProjectDf$PriceSensitivitys),main ="Skewness plot for PriceSensitivity",xlab="PriceSensitivity",col="blue") #plotting the density distribution of PriceSensitivity coulmn in the dataframe to show the measure of skewness.
```

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





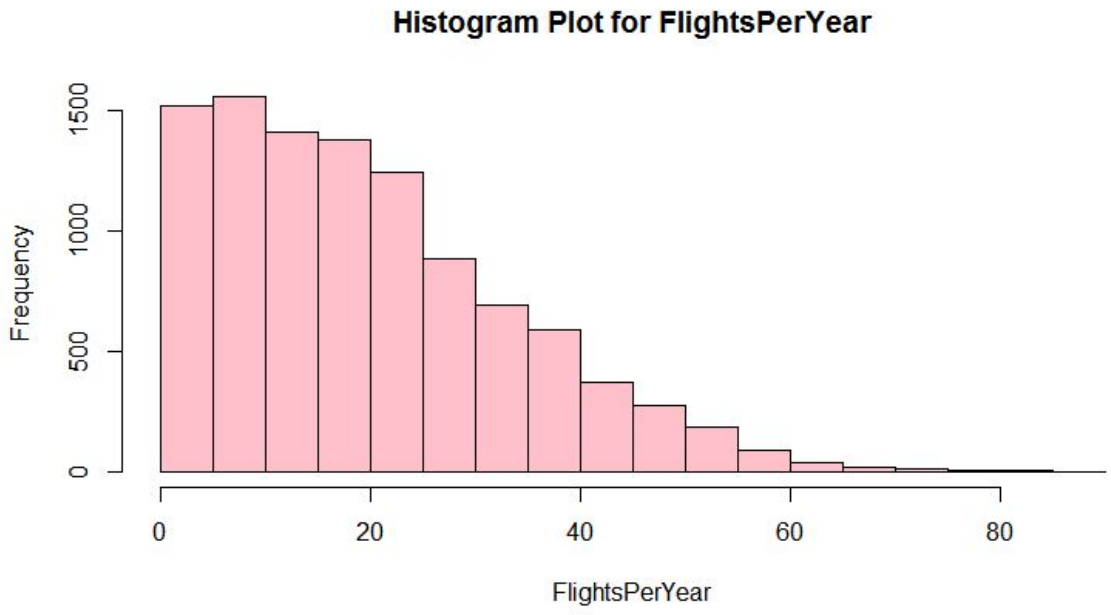
2) Flights run per year for airlines:

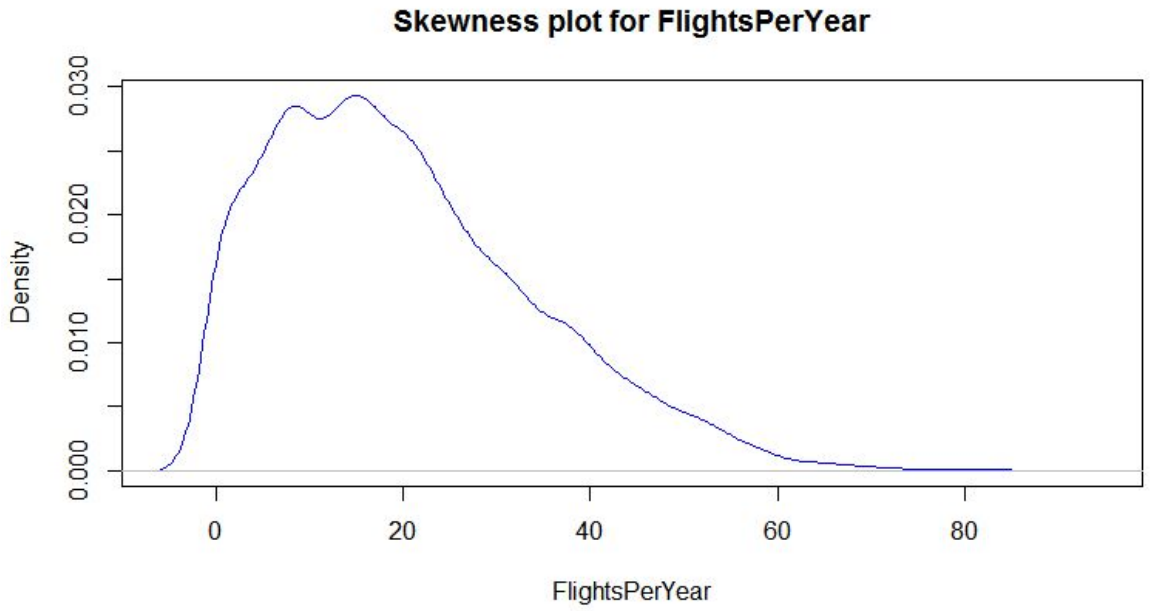
Code Snippet:

```
install.packages("pastecs") #installing package pastec
library(pastecs)
install.packages("psych") #installing package psych
library(psych)
describe(ProjectDf$FlightsPerYear)  #use of describe function to measure the statistics of the
FlightsPerYear  column data frame.
summary(ProjectDf$FlightsPerYear)   #use of summary function to measure the statistics of
FlightsPerYear column in the data frame
hist(ProjectDf$FlightsPerYear ,main =" Histogram Plot for FlightsPerYear",xlab =
"FlightsPerYear",col="pink") # use of hist function to show the histogram distribution of
FlightsPerYear column in the data frame.
plot(density(ProjectDf$FlighttimeInMinutes),main ="Skewness plot for
FlightsPerYear",xlab="FlightsPerYear",col="blue") #plotting the density distribution of FlightsPerYear
coulmn in the dataframe to show the measure of skewness.
```

Descriptive Statistics:

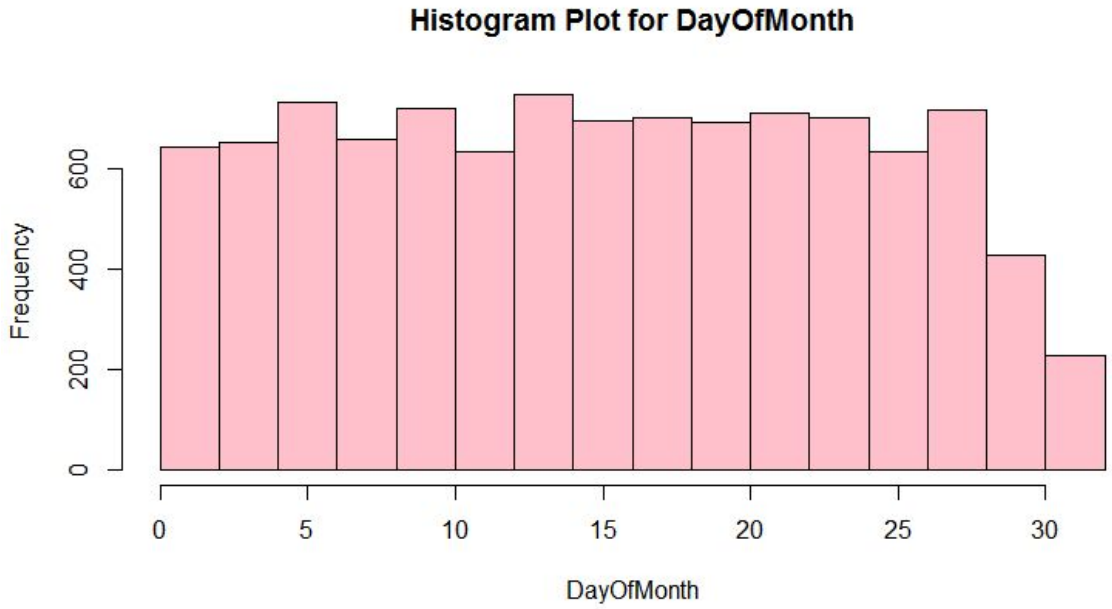
Mean	20.12
Median	18.00
Minimum	0.00
Maximum	89.00
Range	89
Trimmed	18.88
N (total number of variables)	10282
1st quartile	9.00
3rd quartile	29.00
Standard deviation	0.55
Mean Absolute Deviation	14.83
Standard error	0.01
Skew	0.71
Kurtosis	0.86

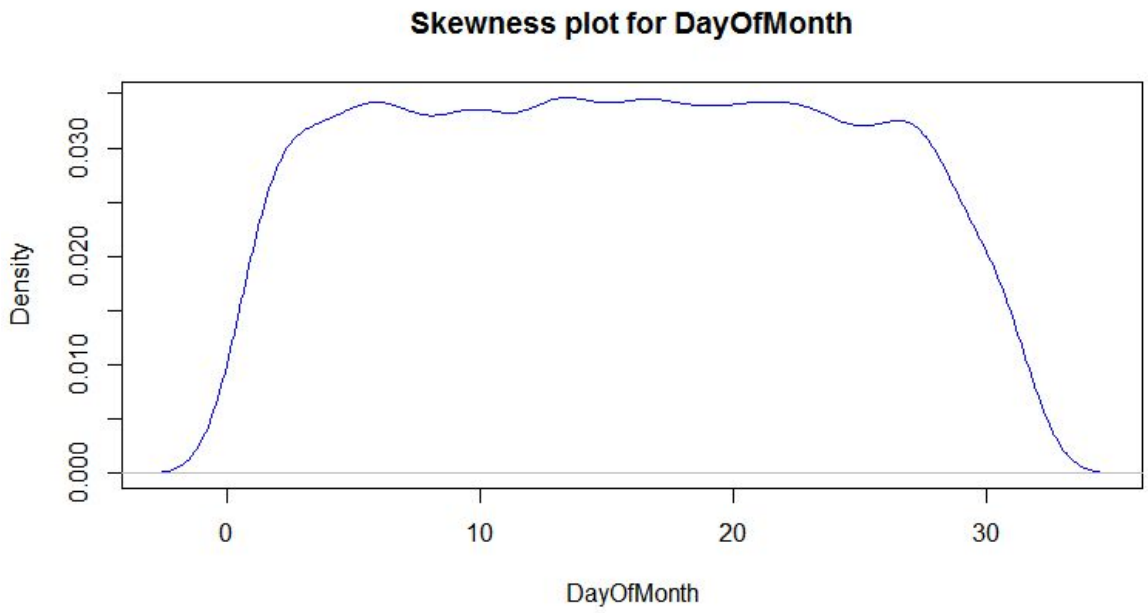




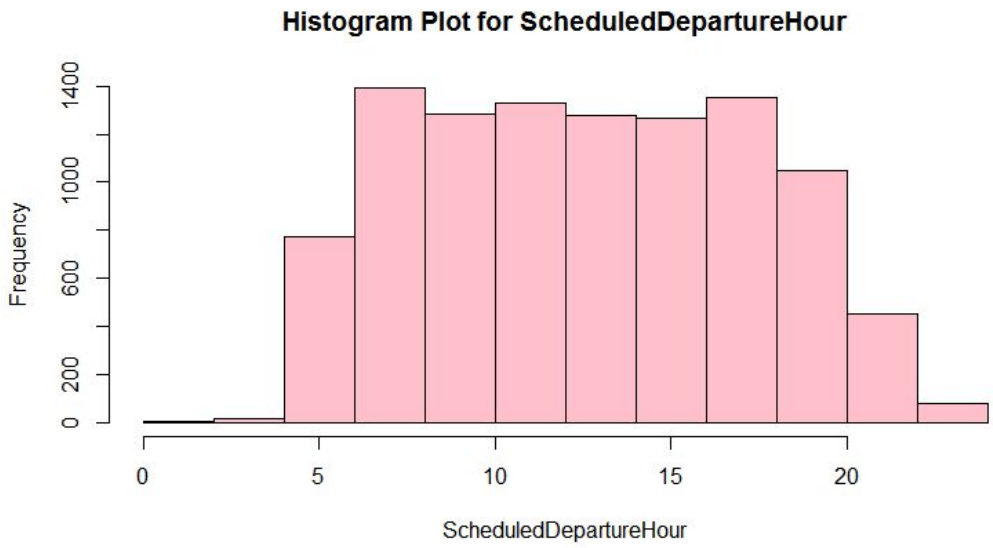
Similarly the Plots for other Airline Attributes can be seen as follows:

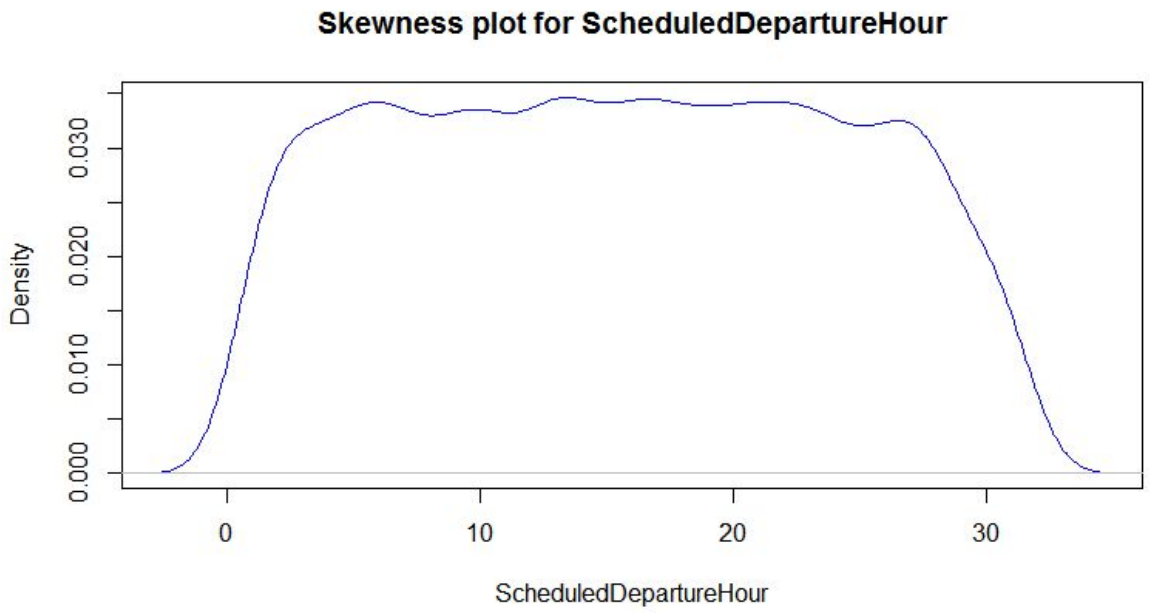
3) Day of the Month the airlines run :





4) Scheduled Departure Hour for the airlines:





Modelling

In-order to distinguish significant attributes from insignificant ones we have used following modelling techniques.

Association Rule Mining:

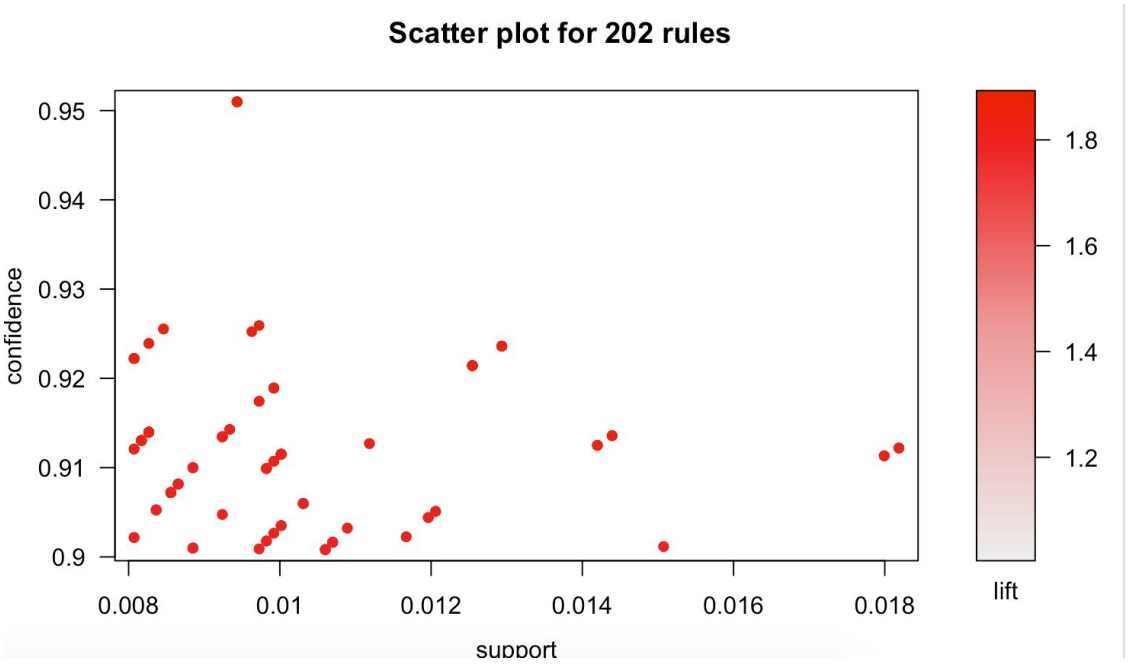
The technique used to find relationships between attributes in the dataset. The model is controlled by **support** and **confidence**. We used our support as 0.008 and 0.9 for finding the best association. We used apriori and found out that the factors of column that can lead to a **promoter** are:

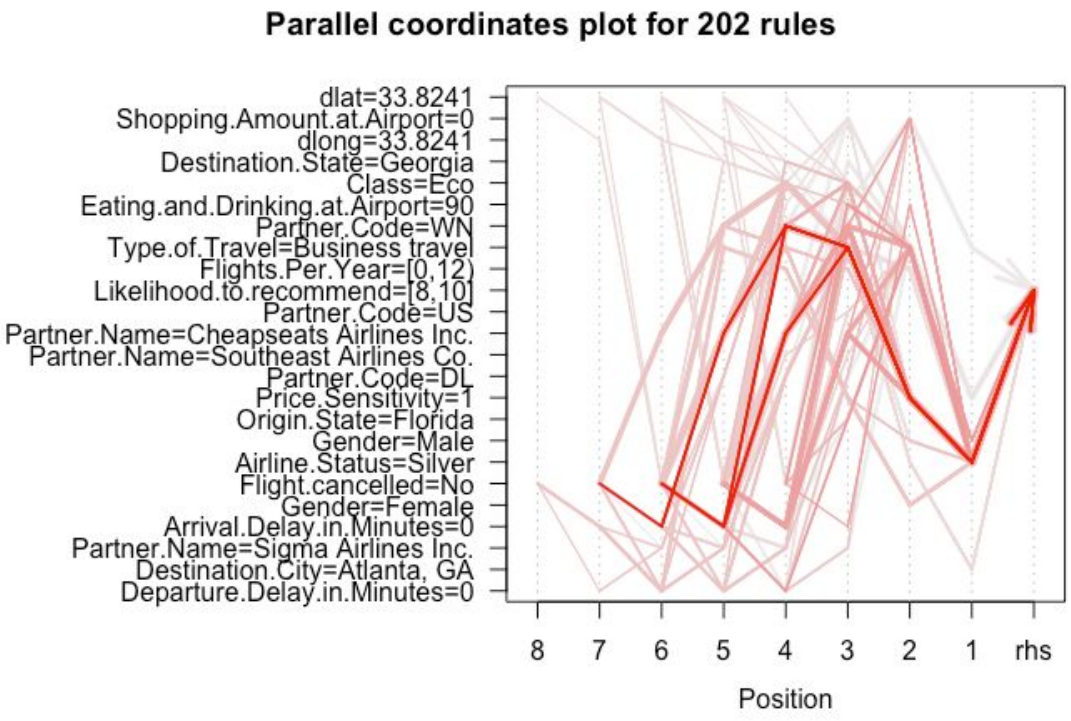
1. Flight cancelled = No
2. Arrival Delay = 0
3. Partner Name = Cheapseats Airlines Inc.
4. Type of Travel = Business travel
5. Price sensitivity = 1 and
6. Airline Status = Silver

Below is the code snippet.

```
> ruleset <- apriori(df,parameter = list(supp=0.008, conf=0.9),appearance = list(rhs=c("Likelihood.t
o.recommend=[8,10]"),default="lhs"),control = list(verbose=F))
> ruleset<-sort(ruleset,decreasing = TRUE,by="lift")
> summary(ruleset)
set of 202 rules

rule length distribution (lhs + rhs):sizes
 4  5  6  7  8  9
 4 27 65 68 32  6

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.000   6.000   7.000   6.569   7.000   9.000

summary of quality measures:
    support           confidence          lift           count
 Min.   :0.008072   Min.   :0.9008   Min.   :1.789   Min.   : 83.0
 1st Qu.:0.008559   1st Qu.:0.9053   1st Qu.:1.798   1st Qu.: 88.0
 Median :0.009337   Median :0.9100   Median :1.807   Median : 96.0
 Mean   :0.009944   Mean   :0.9114   Mean   :1.810   Mean   :102.2
 3rd Qu.:0.010309   3rd Qu.:0.9140   3rd Qu.:1.815   3rd Qu.:106.0
 Max.   :0.018187   Max.   :0.9510   Max.   :1.889   Max.   :187.0

mining info:
 data ntransactions support confidence
   df          10282   0.008        0.9
```

```
plot(ruleset,jitter=0)
 inspect(ruleset)
rules<- subset(ruleset,subset= rhs %in% "Likelihood.to.recommend=[8,10]")
rules
plot(rules, method="paracoord", control=list(reorder=TRUE))
```





SVM Modeling:

SVM modeling is used to train a model based on various regression techniques and svm automatically finds the best one to fit and gets trained. We have divided the dataset into training and testing dataset where we train with former and test with the later one.

For this, we divided the Likelihood to recommend column data into two categories.

1. With a value greater than 7 as “Promoter”.
2. With a value less than 7 as “Detracter”

Below is the code snippet we used to get the SVM output:

```
ProjectDf$Emotion <- "Detractor"
ProjectDf$Emotion[which(ProjectDf$LikelihoodToRecommend > 7)] <- "Promoter"

newdf <- data.frame(ProjectDf$Gender, ProjectDf$DepartureDelayinMinutes, ProjectDf$ArrivalDelayin
newdf$ltr <- as.integer(newdf$ltr)
newdf$Emotion <- "Detractor"
newdf$Emotion[which(newdf$ltr > 7)] <- "Promoter"

randIndex <- sample(1:dim(newdf)[1])
twothird <- floor(2*dim(newdf)[1]/3)
training_ds <- newdf[randIndex[1:twothird],]
testing_ds <- newdf[randIndex[(twothird+1):dim(newdf)[1]],]

svmOutput <- ksvm(Emotion ~.,data = newdf, kernel = "rbfdot", kpar = "automatic", C=5, cross = 3,
svmOutput
svmPred <- predict(svmOutput, testing_ds)
comptable <- table(svmPred == testing_ds[,7])
Accuracy <- (comptable[1]/(comptable[1]+comptable[2]))*100
Accuracy
```

Training results:

```
> svmOutput
Support Vector Machine object of class "ksvm"

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

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.204008169037089

Number of Support Vectors : 821

Objective Function Value : -139.0646
Training error : 0
Cross validation error : 0
Probability model included.
```

Prediction results with testing data:

```
> comptable <- table(svmPred == testing_ds[,7])

> comptable

FALSE  TRUE
 1691  1737

> Accuracy <- (comptable[1]/(comptable[1]+comptable[2]))*100
> Accuracy
   FALSE
49.32905
```

The results are a bit unconvincing since the accuracy is less than 50%. Hence we concluded that svm is not fit for our prediction of other data.

Linear Modeling:

We performed Linear Modeling in 2 phases:

1. Univariate modeling
2. Multivariate Modeling

Univariate Modeling:

Each attribute of the given data was mapped with Likelihood to Recommend attribute and we found out which attributes are statistically significant predictors of Likelihood to Recommend.

We checked whether the Likelihood to Recommend is dependent on statistically significant predictors or not. If they were then we assessed how much they are dependent on them by analyzing the R-squared values. R squared value's range is between 0 and 1, 0 means Likelihood to Recommend attribute doesn't depend on the predictor whereas if R squared value is 1 then Likelihood to Recommend attribute is completely dependent on predictor.

Type of Travel and Airline Status these are the attributes on which Likelihood to Recommend depends the most.

Multivariate Modeling:

In this Modeling we did mapped a cluster of attributes together with Likelihood to Recommend attribute and we found out whether attributes collectively are statistically significant predictors of Likelihood to Recommend.

But the results that we got weren't in an acceptable range as a result we couldn't derive keen insights out of the multivariate modeling.

Actionable Insights

1. Improvement of in-flight services especially for Female and Older people as they had experienced more or less discomfort during their journey.
2. There are approximately 25 having less than average likelihood to recommend of 7 which gives the regions of improvement to work on.
3. Text mining and Descriptive Statistics prove that the loyalty is coming down and the below two partners responsible for it can be cut loose.
4. The flights being cancelled can play a major role in customers giving a good rating so the flight cancellation decision can be communicated in a more pleasing way to the customer.
5. The Association rules mining shows us the attributes of customer where the airlines can continue being focused to sustain those customers and reduce customer churn.

6. The more focused the airlines focused on business travel passengers the better the rating. So the marketing and sales can make sure that they aren't losing those customers.
7. The Silver airline status passengers play a substantial role in promoters of the company. So special offers and discounts can make them loyal to Southeast airlines.
8. From the age group plot, we can understand that middle aged(30-60) passengers are being at the edge of promoters which we need to focus because the persistent retaining efforts on these customers can help the airlines to sustain in the market.
9. Overall, the average likelihood to recommend is at the edge of promoter(7.07) so, it is high time for the company to make efforts on the above categorical customer by cutting loose under-performing partners and deploying new sales and marketing strategies.