Airlines Satisfaction Analysis Report

IST-687 (M009)

Priyank Jethva

Ashutosh Jha

Yaksh Suresh Shobhavat

Sharvil Turbadkar

Nathan Wendel

Contents

I. Introduction:

We have created this project to help our client, Southeast Airlines, in reducing their customer churn. Southeast Airlines is one of the top-four airlines in the United States and has many regional airline partners throughout the country. But similar to other airlines, it has trouble keeping its customers from switching to a different airline provider. This is represented by a term called “customer churn”, which can also be called “customer attrition”. In order to reduce customer churn, Southeast Airlines had implemented a loyalty program for frequent flyers. This loyalty program was intended to keep customers with Southeast Airlines because the more that they flew with them, the more “miles” that they would acquire. These miles would allow the customer to rise through the different levels of service that the airline provided, while also allowing them the opportunity at free or discounted flights. Many other airlines were also using similar programs, as it was the “accepted industry best practice”.

However, there were problems with this. For one, it was difficult to gauge the effect of the loyalty program on customer churn. The effects of these loyalty programs were becoming in question, as customers seemed to be valuing them less. They also seemed to cause issues for the airlines, as well, as it was suggested that frequent flyer points were behind almost $12 billion of “loyalty debt” from these programs. Moreover, it was difficult to track customer churn, in general, because it is a “lagging indicator”. This refers to the fact that when customers switch to a different airline, the customer churn has already occurred. Therefore, it is very difficult to track in real-time, and is ultimately more of a measurement of the damage that was done by a customer switching airlines. As a result, it was necessary for us to conduct an analysis on what actually impacts customer churn so that we can suggest aspects that the airline should emphasize in order to maintain their customers.

This was done by analyzing the Net Promoter Score data provided to us. Net Promoter Score, or NPS for short, is a survey that asks customers to respond to the following question on a scale of 1-10: “How likely is it that you will recommend our airline to a friend or colleague?”. A score above an 8 categorizes a customer as a “promoter”, meaning that they are a good asset to the airline, and will likely help to increase their business. A score between 7-8 categorizes a customer as “passive”, meaning that the customer is not necessarily an asset or a detriment to the

company. Finally, a score below 7 categorizes the customer as a "detractor", meaning that they are likely to switch to a different airline. Moreover, the overall NPS score can be calculated by subtracting the percentage of detractors in a study by the percentage of promoters. Although this may seem like a simple process, it also gives the individuals analyzing the results the opportunity to examine certain attributes of the customers within the study. Although the score itself is valuable information, it is also important to know what kind of customers are the most, or least, satisfied with the airline, and ultimately which customers are more likely to contribute to customer churn.

In our study, the following attributes were provided to us:

1. ***Likelihood to Recommend –*** *rated on a scale of 1 to 10, which shows how likely the customer is to recommend the airline to their friends (10 is very likely, and 1 is not very likely).*
2. ***Airline Flyer Status*** *– each customer has a different type of airline status, which are platinum, gold, silver, and blue (based on level of travel with the airline)*
3. ***Age*** *– the specific customer's age. Ranging from 15 to 85 years old.*
4. ***Gender*** *– male or female.*
5. ***Price Sensitivity*** *– the grade to which the price affects to customers purchasing. The price sensitivity has a range from 0 to 5.*
6. ***Year of First Flight*** *– this attribute shows the first flight of each single customer. The range of year of the first flight for each customer has been started in 2003 until 2012.*
7. ***Flights Per Year*** *– The number of flights that each customer has taken in the most recent 12 months. The range starting from 0 to 100.*
8. ***Loyalty*** *– An index of loyalty ranging from -1 to 1 that reflects the proportion of flights taken on other airlines versus flights taken on this airline. A higher index means more loyalty.*
9. ***Type of Travel*** *– One of business travel, mileage tickets, or personal travel (ex. vacation)*
10. ***Total Frequent Flyer Accounts*** *– How many frequent flyer accounts the customer has.*
11. ***Shopping Amount at Airport*** *– The spending on non-food & services at the airport (in $)*
12. ***Eating and Drinking at Airport*** *– The spending on food/drink at the airport (in $).*
13. ***Class*** *– three different kinds of service level (business, economy plus, and economy).*
14. ***Day of Month*** *–the traveling day of each costumer (ranges from 1 to 31).*
15. ***Flight date*** *– the passenger's flight date of travel.*

16. ***Partner Code*** *– This airline works with wholly- and partially-owned subsidiary companies to deliver regional flights. For example, AA, AS, B6, and DL.*
17. ***Partner Name*** *– These are the full names of the partner airline companies.*
18. ***Origin City*** *– the place where passenger departed from. For example, Boston MA.*
19. ***Origin State*** *– the place where passenger departed from. For example, Texas.*
20. ***Destination City*** *– the place to which passenger travels to. For example, Boston MA.*
21. ***Destination State*** *– the place to which passenger travels to. For example, Texas.*
22. ***Scheduled Departure Hour*** *– the specific time at which the plane was scheduled to depart.*
23. ***Departure Delay in Minutes*** *– How long the flight's departure was delayed, when compared to schedule.*
24. ***Arrival Delay in Minutes*** *– How long the arrival was delayed.*
25. ***Flight Cancelled*** *– occurs when the airline does not operate the flight.*
26. ***Flight time in minutes*** *–the length of time, in minutes, to reach the destination.*
27. ***Flight Distance*** *– the distance between the departure and arrival destination.*
28. ***Comment*** *– a free form text field of the passenger comment, with respect to the flight.*

It should be noted that the "Comment" attribute was removed in our analysis, as it does not contribute any statistical data to the study. Along with this, we were also able to create and add a number of other attributes to the study that we believed would help us in understanding customer churn. The attributes that were added are the following:

1. ***Customer Type-*** *Whether the customer is a promoter or a detractor*
2. ***Loyalty Buckets-*** *Bucketing customers depending upon their loyalty (-1 to -0.27 = Unloyal, -0.28 to 1 = Loyal)*
3. ***Age Buckets-*** *Bucketing customers by age groups (Less than 23 = Young, 23 to 61 = Adult, Greater than 61 = Old)*

The remaining 6 attributes were those that changed previous non-numeric variables to numeric. For example, this included changing the Flight Cancellation attribute, which was originally a"yes" or "no" variable, to one that instead gave a response of 0 or 1.

II. Business Questions Addressed:

- Which are the factors which tend to influence the satisfaction of the customers towards the airlines, and which is the most effective one?

- Is there any relationship between these factors?

- Are the variables positively or negatively related, in case a relationship exists between them?

- Is there any solution which could help them stand out in achieving higher customer satisfaction in correspondence to Southeast Airlines?

- How does gender, airline status, airline class, age, price sensitivity, travel type, and flight cancellation status vary according to the current satisfaction variable? Is it possible for us to identify a trend or a feature for people who are more likely to give a higher satisfaction score and also a lower satisfaction score?

- Can a method be suggested to deal with these problems in order to gain higher customer satisfaction?

III. Data Acquisition, Cleansing, Transformation and Munging:

The dataset for this project contained a total of 10,282 data points with 27 original attributes (after one attribute was removed) and 9 added attributes. We started the data munging process by altering the names of the dataset columns so that the dots can be removed from column names, deleted values with decimals, and then re-numbered the rows. Then we converted the columns to characters and numeric to ease the application of functions on each of them. The date format of flight date column was inconsistent, so we changed the format of this column to a standard date format.

Our main focus for us was to increase customer satisfaction and to find out how customer satisfaction is affected. Here, we took into consideration only the data of Southeast Airlines, and we researched on all the factors which could possibly improve the satisfaction of all the customers.

We replaced the NA values for arrival delay in minutes and departure delay in minutes. In case, both the arrival delay in minutes and delay in minutes were NA then they were replaced by zero. This was done by making a linear model between arrival delay in minutes and departure delay in minutes as they have a high correlation of 95.9. The NA values for likelihood to recommend was replaced by the mean. The NA values for Flight time in minutes were replaced by linear regression with flight distance as they have high correlation of 95.4. The NPS score is -0.11 which means that there are more promoters than detractors.

Code:

```
# Converting to data frame fro JSON format
df<-jsonlite::fromJSON(vector)

View(df)
# Calculating NA values across all attributes
sapply(df,function(x)sum(is.na(x)))

# The following attributes have NA values

# Likelihood to recommend
# Flight Time in Minutes
# Freetext
# Departure.Delay.in.Minutes
# Arrival.Delay.in.Minutes

#6531
head(df$Likelihood.to.recommend)

which(is.na(df$Likelihood.to.recommend))
# 6531 index value is na

table(df$Class)
str(df$Class)


# Replacing NA values

################# Departure.Delay.in.Minutes and Arrival.Delay.in.Minutes #########
View(df$Departure.Delay.in.Minutes)
rm(list = ls())

#Replacing  na values by MODE which is zero as there are many cases where there are no delays

cor(df$Arrival.Delay.in.Minutes,df$Departure.Delay.in.Minutes,use = "pairwise.complete.obs")
# High Correlation of 95.9% is there

lm(df$Arrival.Delay.in.Minutes~df$Departure.Delay.in.Minutes,data = df)
lm(df$Departure.Delay.in.Minutes~df$Arrival.Delay.in.Minutes,data = df)
table(is.na(df$Departure.Delay.in.Minutes))
```

```r
# Temporaray data frame
df1<-data.frame(df)

for(i in 1:nrow(df1))
{
  if((is.na(df1$Arrival.Delay.in.Minutes[i])=="TRUE")&&(is.na(df1$Departure.Delay.in.Minutes[i]))=="TRUE")
    # If both arrival and departure times are not mentioned
  {
    df1$Arrival.Delay.in.Minutes[i]=0
    df1$Departure.Delay.in.Minutes[i]=0
  }
  if(is.na(df1$Arrival.Delay.in.Minutes[i])=="TRUE")
    # If Arrival time is NA
  {
    df1$Arrival.Delay.in.Minutes[i]=0.8066+0.9881 *df1$Departure.Delay.in.Minutes[i]
  }
  if(is.na(df1$Departure.Delay.in.Minutes[i])=="TRUE")
    # If Departure Time is NA
  {
    df1$Departure.Delay.in.Minutes[i]=0.4808+0.9314*df1$Arrival.Delay.in.Minutes[i]
  }
}

# To check whether all na values are removed

sum(is.na(df1$Arrival.Delay.in.Minutes))
sum(is.na(df1$Departure.Delay.in.Minutes))


# df$Departure.Delay.in.Minutes[which(is.na(df$Departure.Delay.in.Minutes))]<- 0
# df$Arrival.Delay.in.Minutes[which(is.na(df$Arrival.Delay.in.Minutes))]<-0
mean(df1$Departure.Delay.in.Minutes)
# Mean Departure Delay is 15.22 minutes
mean(df1$Arrival.Delay.in.Minutes)
# Mean Arrival Delay is 15.84 minutes
```

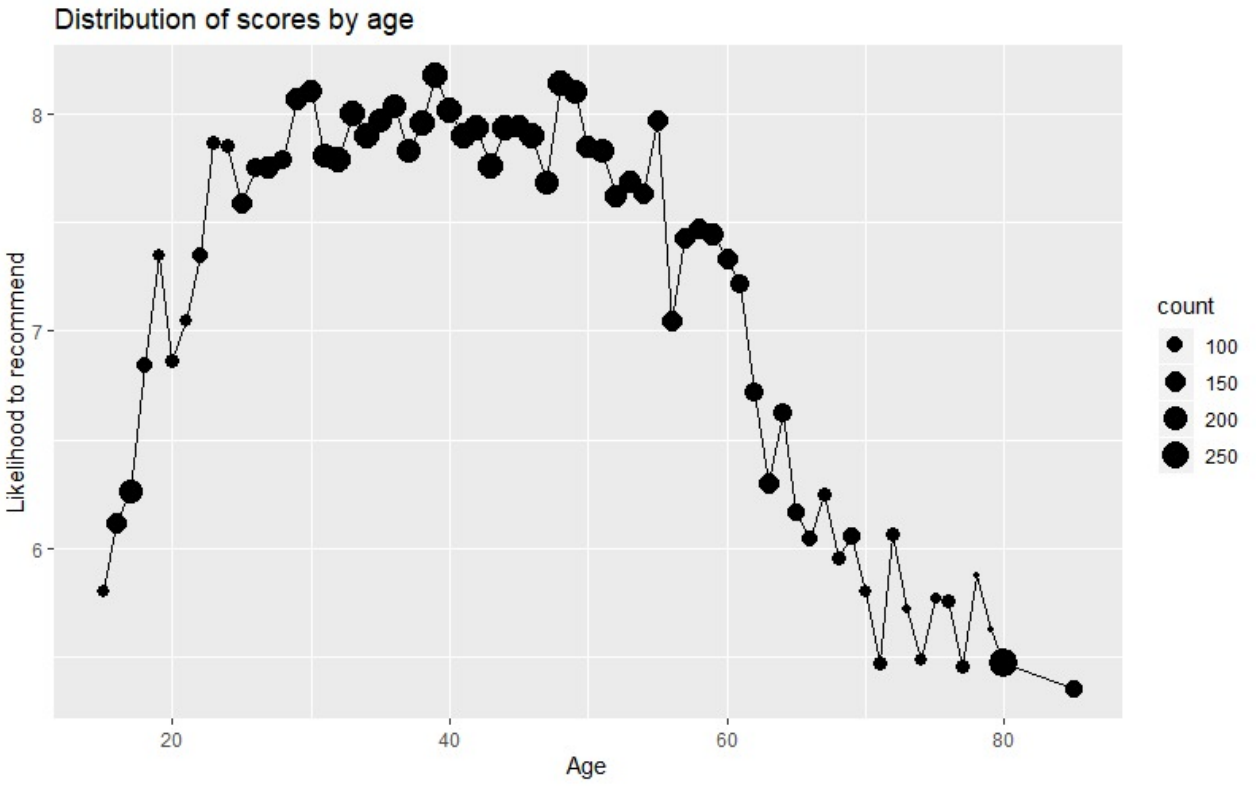
```r
######### Flight Time in Minutes##########
table(is.na(df$Flight.time.in.minutes))
which(is.na(df$Flight.time.in.minutes))
```

```r
for(i in 1:nrow(df1))
{
  if((is.na(df1$Arrival.Delay.in.Minutes[i])=="TRUE")&&(is.na(df1$Departure.Delay.in.Minutes[i]))=="TRUE")
    # If both arrival and departure times are not mentioned
  {
    df1$Arrival.Delay.in.Minutes[i]=0
    df1$Departure.Delay.in.Minutes[i]=0
  }
  if(is.na(df1$Arrival.Delay.in.Minutes[i])=="TRUE")
    # If Arrival time is NA
  {
    df1$Arrival.Delay.in.Minutes[i]=0.8066+0.9881 *df1$Departure.Delay.in.Minutes[i]
  }
  if(is.na(df1$Departure.Delay.in.Minutes[i])=="TRUE")
    # If Departure Time is NA
  {
    df1$Departure.Delay.in.Minutes[i]=0.4808+0.9314*df1$Arrival.Delay.in.Minutes[i]
  }
}
```

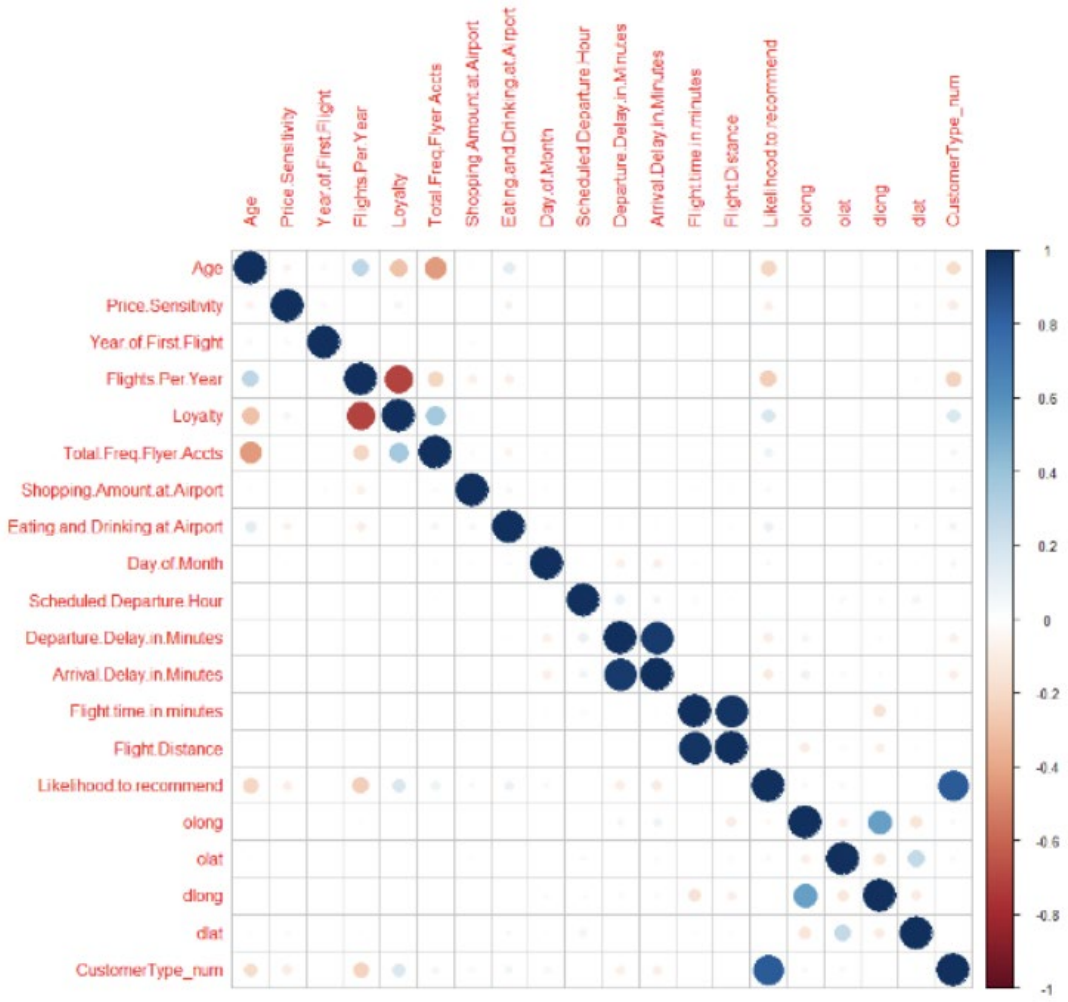
```r
for(i in 1:nrow(df1))
{
  if(is.na(df1$Flight.time.in.minutes[i])=="TRUE")
  {
    df1$Flight.time.in.minutes[i]=17.6625+0.1176*df1$Flight.Distance[i]
  }
}
```

IV. Descriptive Statistics and Visualization:

- **Inferences Made**

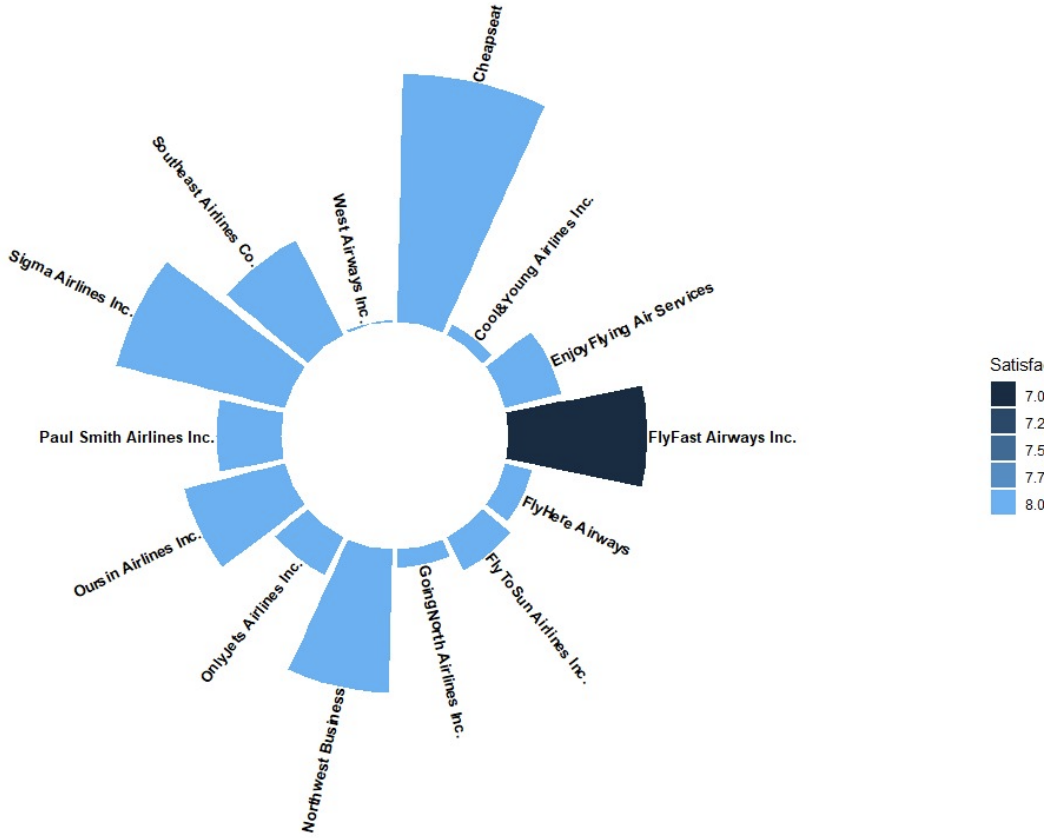


As we can see by the distribution of scores by age, adults are more likely to travel and recommend Southeast Airlines.

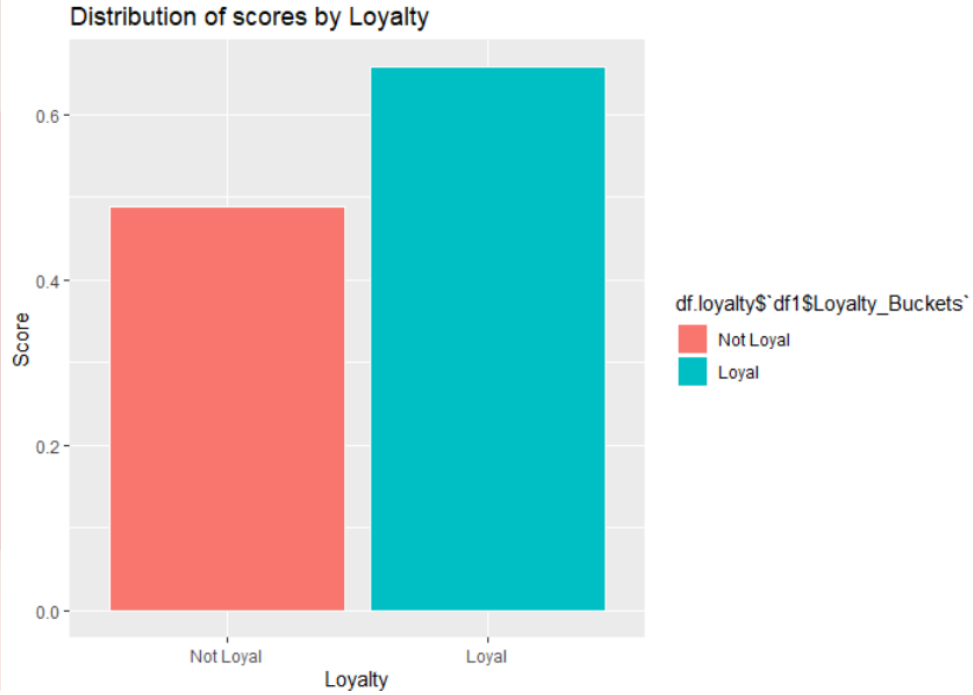


A correlation matrix is used to investigate dependencies between multiple variables at the same time. This results in a table of correlation coefficients between each variable. It measures the percentage of fluctuation in one variable that can be explained by another variable. A correlation of 1 means the variables move in perfect unison, a correlation of -1 means the variables move in the complete opposite direction, and a correlation of 0 means there is no relationship at all between the two variables.

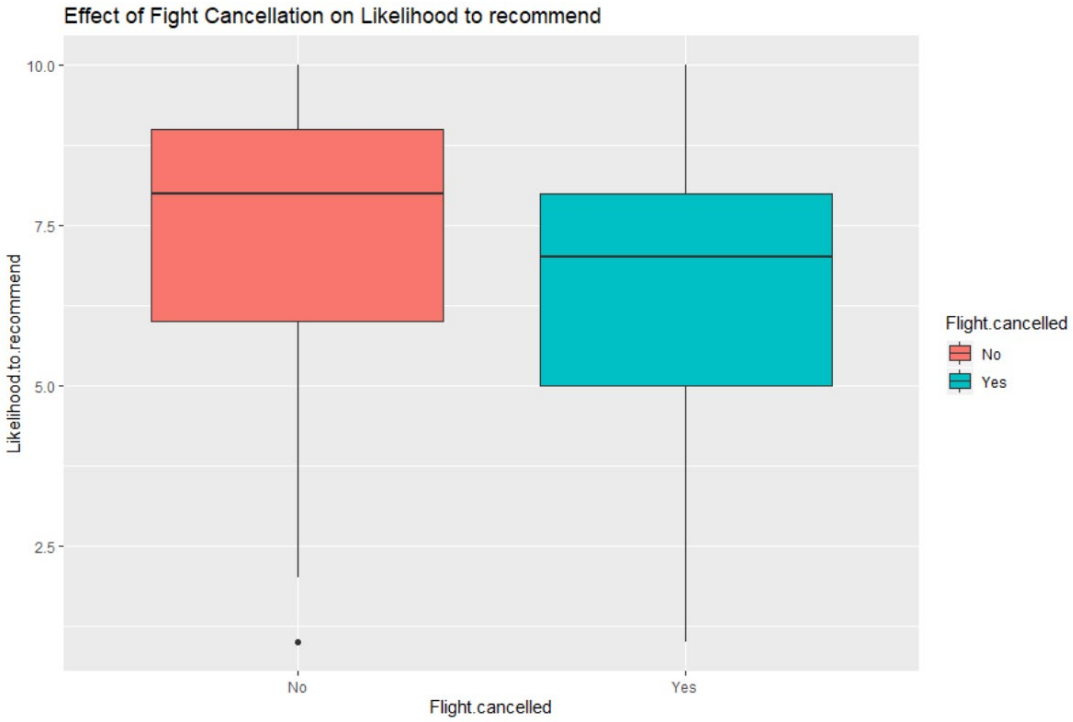
The correlation matrix here captures the correlation between 2 non numeric variables. It quatifies the value between -1 and +1 where -1 being very weakly associated and +1 being most strongly associated.



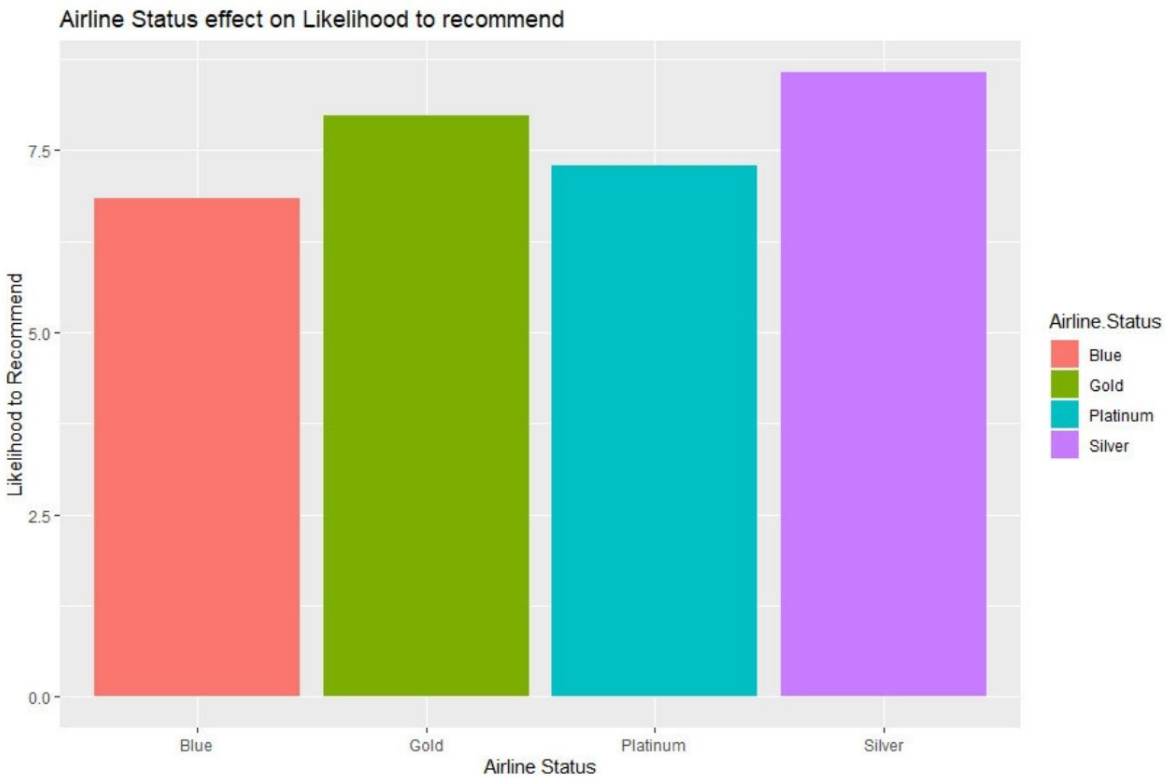
From the circular boxplot above, we can conclude that Cheapset airline has most number of customers and the Flyfast Airways has customers with most promoters



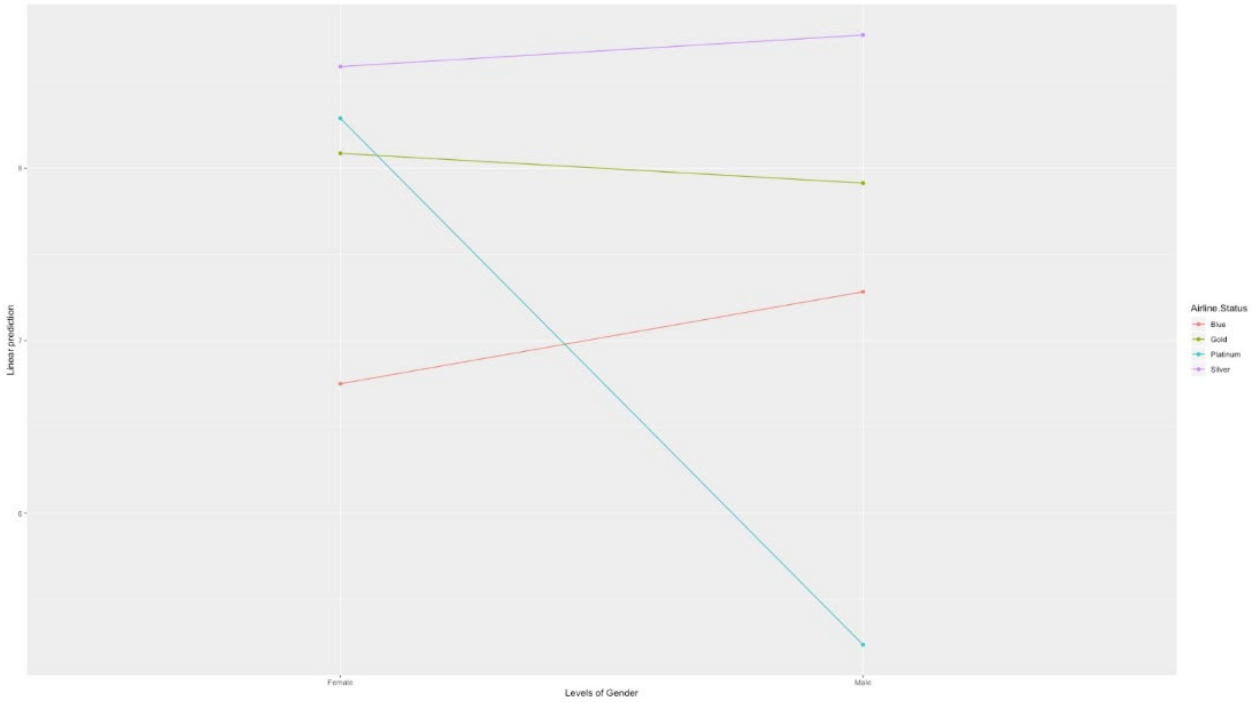
As would make sense, we can see that a loyal customer is more likely to recommend the airline than a non-loyal customer.



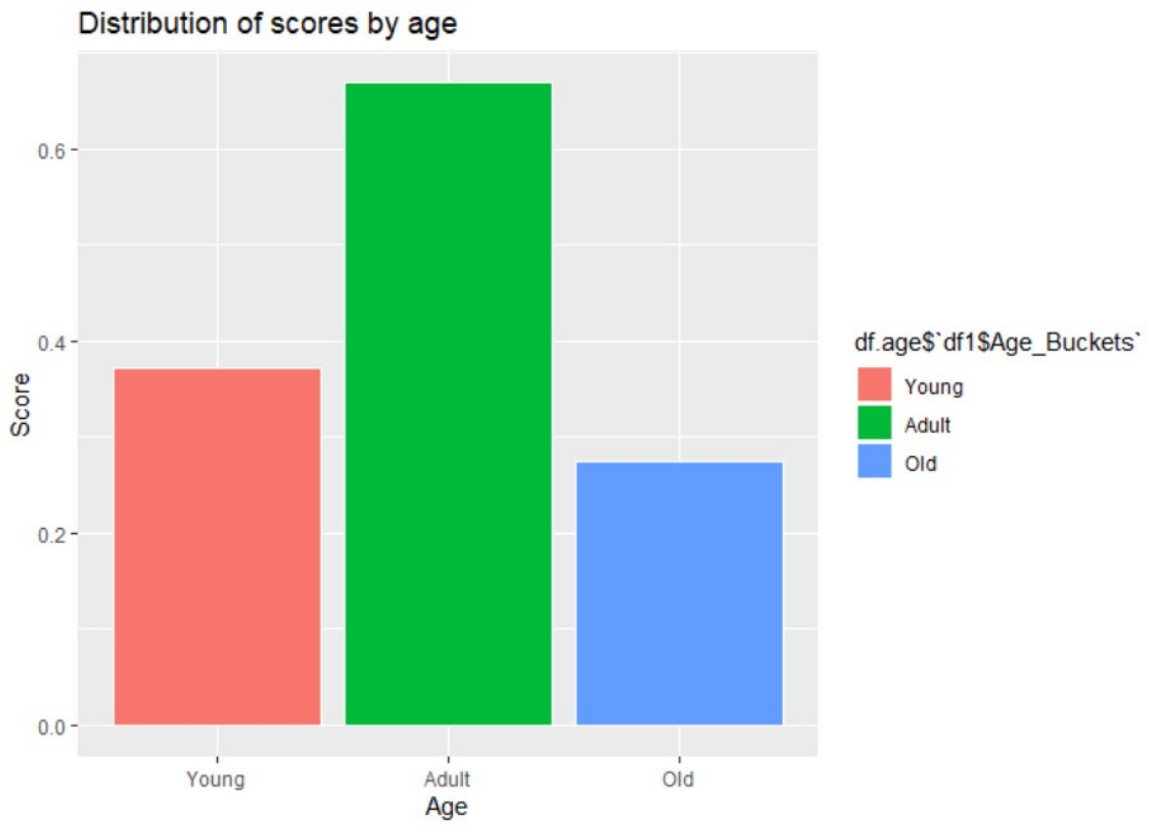
Similarly, it makes sense for a customer to be dissatisfied if their flight was cancelled, and would be less likely to recommend the airline. This can be seen in the boxplot above.



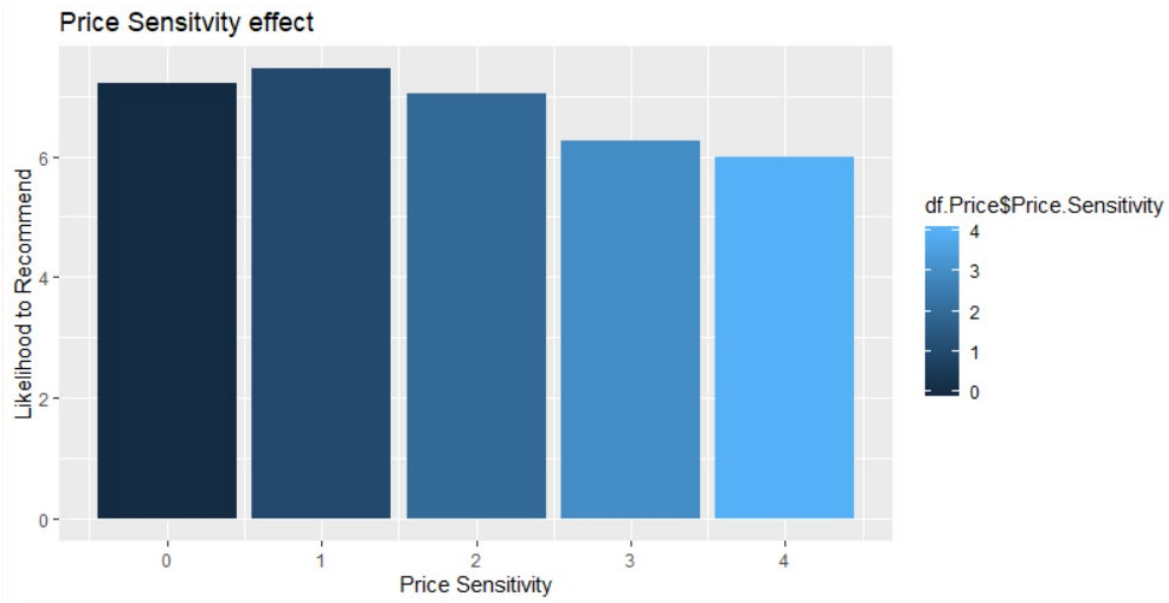
When examining the different airline status options, we see that those customers with the Silver status are the most likely of the customers to recommend Southeast Airlines.



This figure gives us a closer look at the previously discussed airline status, as we are able to see that there is a significant difference between male Silver status customers and female Silver status customers. Although our initial reaction to the previous figure was that Silver status customers are the most likely to recommend the airline, we see that this essentially only applies for female Silver customers. In fact, we can see that male Silver customers are the least likely to recommend the airline. We must keep this in mind throughout the study.



When observing the different age groups, we see that adults are by far the most likely group to recommend the airline.



In this figure, we can see that a cheaper flight ticket correlates with a higher likeliness to recommend the airline.

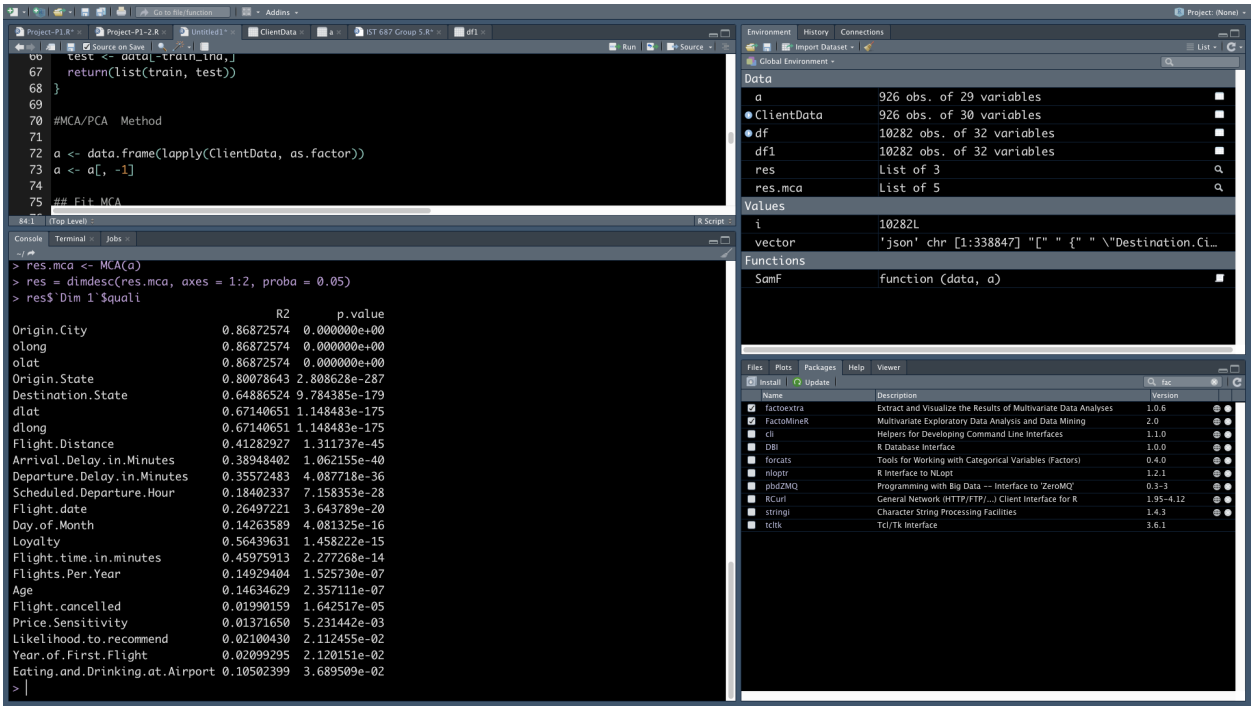
VI. Use of Modeling Techniques and Visualizations

A. Multiple Correspondence Analysis

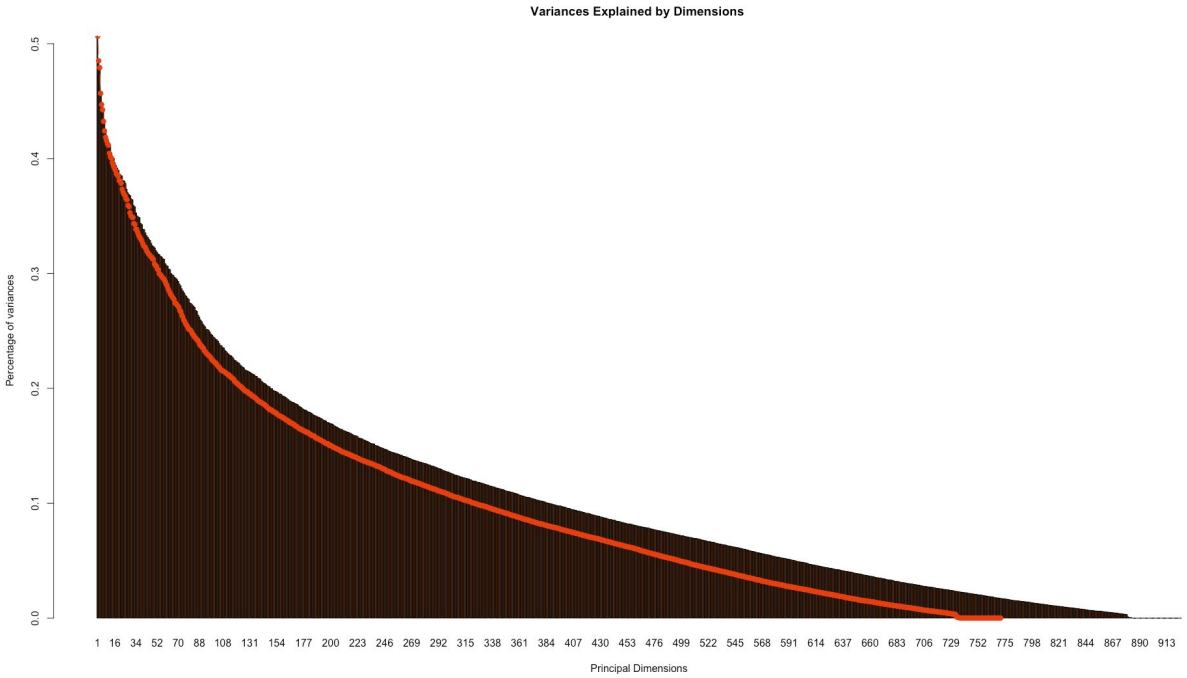
Multiple correspondence analysis (MCA) is an extension of the simple correspondence analysis for summarizing and visualizing a data table containing more than two *categorical variables*. It can also be seen as a generalization of principal component analysis when the variables to be analyzed are categorical instead of quantitative (Abdi and Williams 2010).

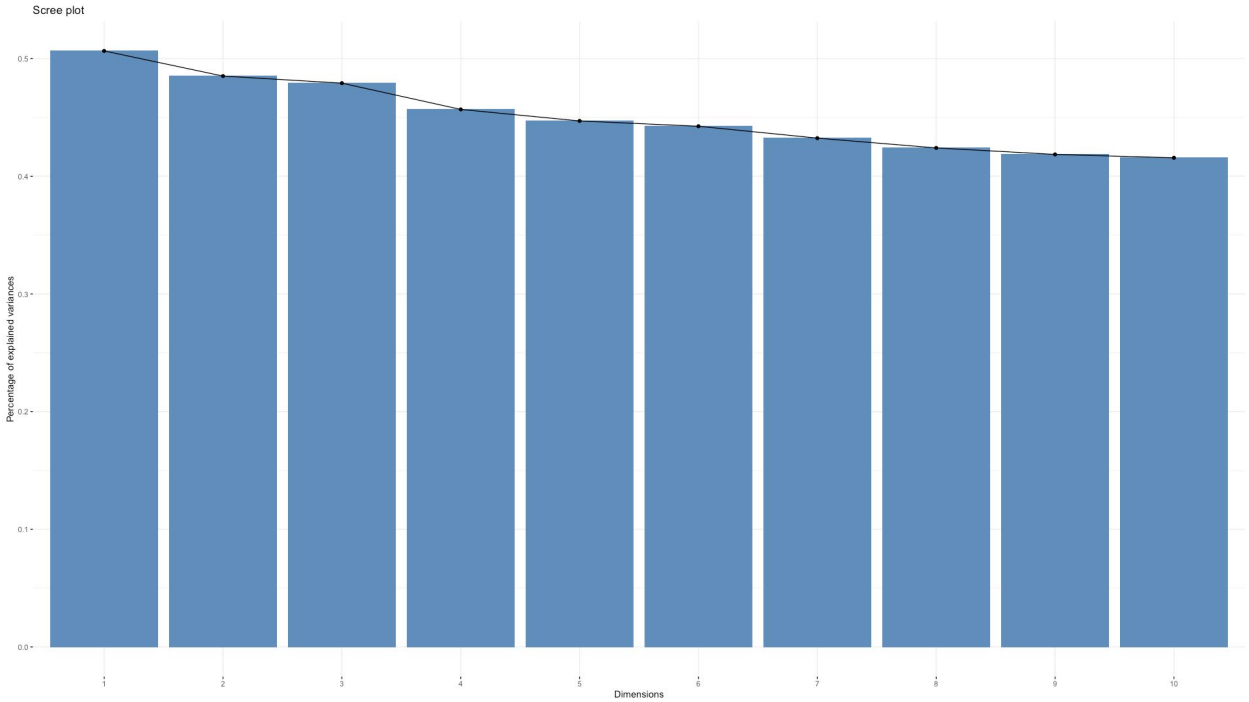
MCA is generally used to analyse a data set from survey. The goal is to identify:

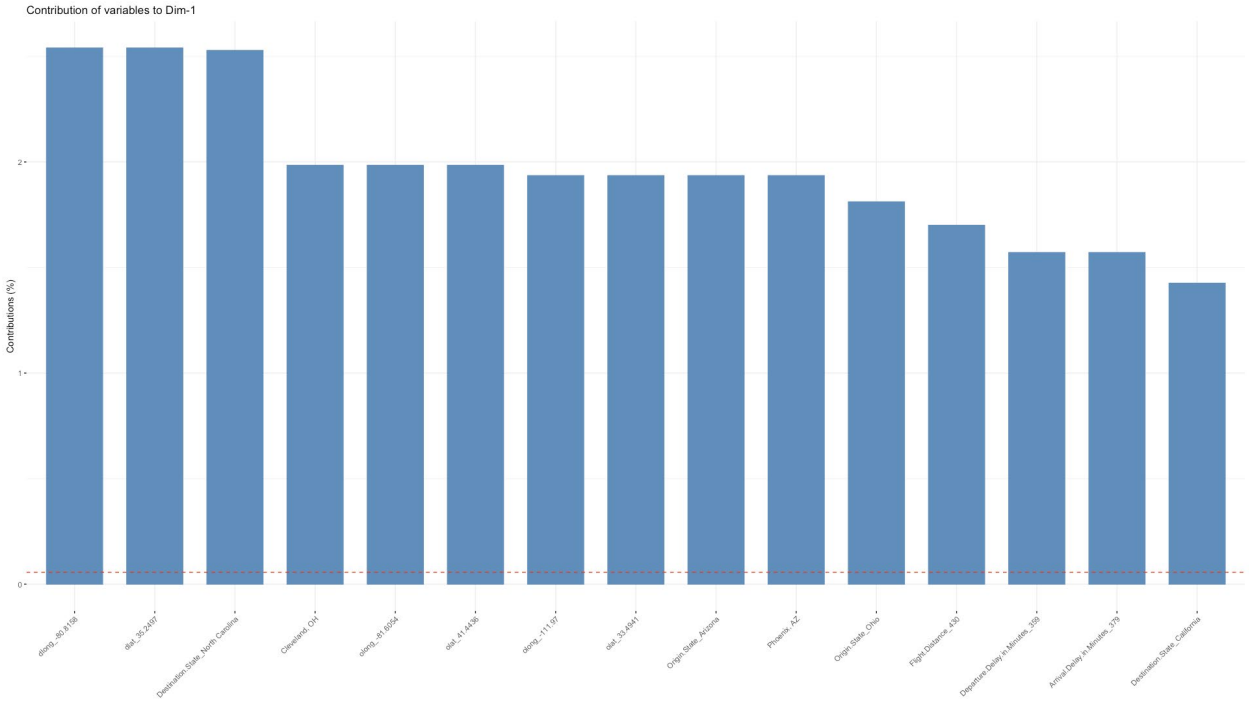
- A group of individuals with similar profile in their answers to the questions
- The associations between variable categories

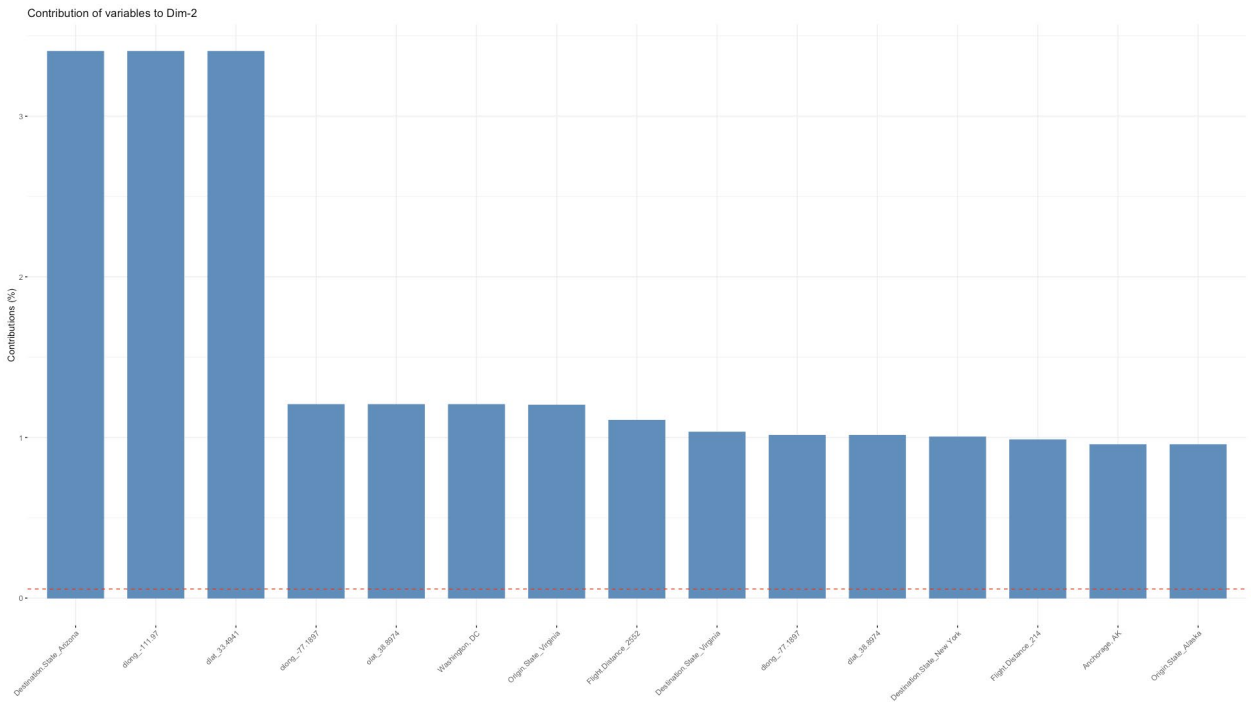


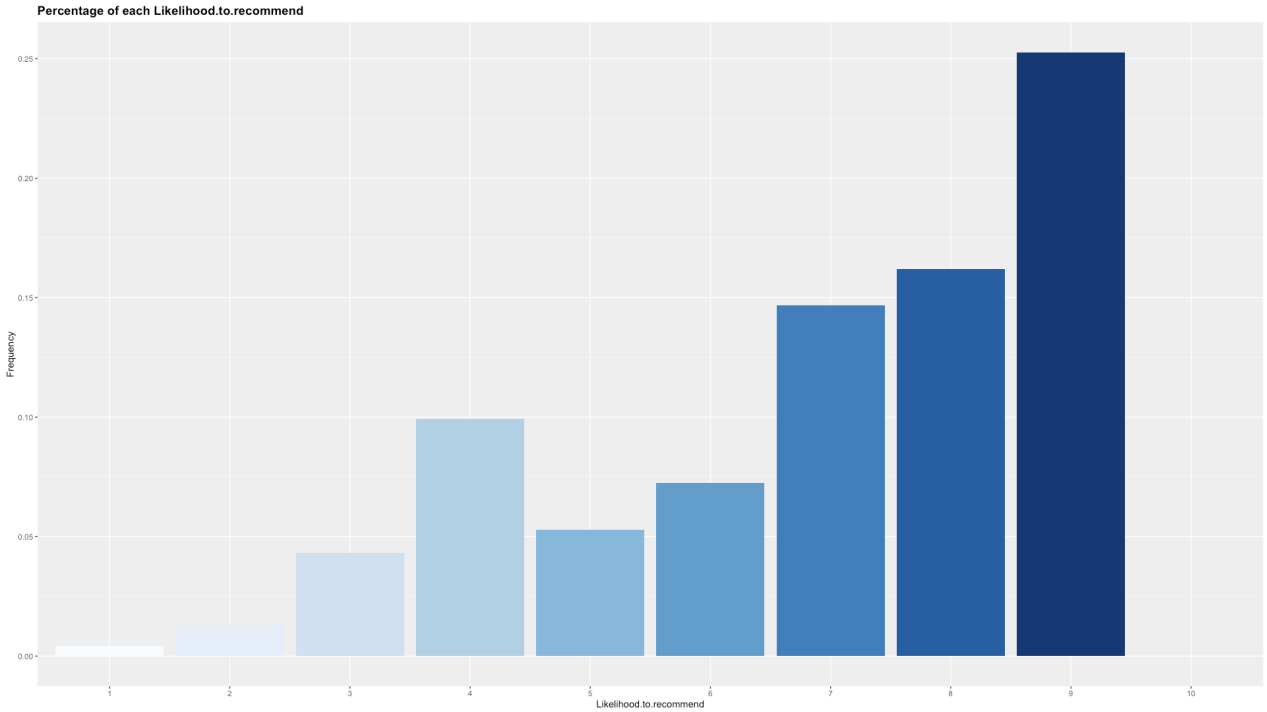
```
Categorical variables (eta2)
                          Dim.1 Dim.2 Dim.3
Origin.City             | 0.869 0.823 0.863 |
Airline.Status          | 0.001 0.002 0.014 |
Age                     | 0.146 0.173 0.169 |
Gender                  | 0.000 0.003 0.011 |
Price.Sensitivity       | 0.014 0.001 0.101 |
Year.of.First.Flight    | 0.021 0.018 0.006 |
Flights.Per.Year        | 0.149 0.131 0.132 |
Loyalty                 | 0.564 0.518 0.682 |
Type.of.Travel          | 0.000 0.003 0.020 |
Total.Freq.Flyer.Accts | 0.006 0.027 0.028 |
>
```




Scree plot
Percentage of explained variances
0.0
0.1
0.2
0.3
0.4
0.5
1
2
3
4
5
6
7
8
9
10
Dimensions


Contribution of variables to Dim-1
Contributions (%)
0
1
2
dlong_-80.8198
dlat_35.2497
Destination.State_North Carolina
Cleveland, OH
dlong_-81.6954
dlat_41.4436
dlong_-111.97
dlat_33.4941
Origin.State_Arizona
Phoenix, AZ
Origin.State_Ohio
Flight.Distance_430
Departure.Delay.in.Minutes_359
Arrival.Delay.in.Minutes_379
Destination.State_California





Upon successfully running the MCA model, we observe that the first 3 dimensions have the most explaining power which contribute to the entire matrix as seen in the tables above. Despite the fact

that only a tiny percentage of variances are explained, we can still learn some general information from this model.

We performed linear modelling on the entire dataset by adding one variable at a time and kept on increasing the R squared value. Finally, we got 10 out of 28 columns which were significant and impacting customer satisfaction column positively as well as negatively as seen in tables above.

- Origin City
- Age
- Gender
- Airline Status
- Price Sensitivity
- Year of First Flight
- Type of Travel
- Flights per travel
- Loyalty
- Total frequent flyer accounts

The above factors are considered to be independent variables and customer satisfaction will be the dependent variable.

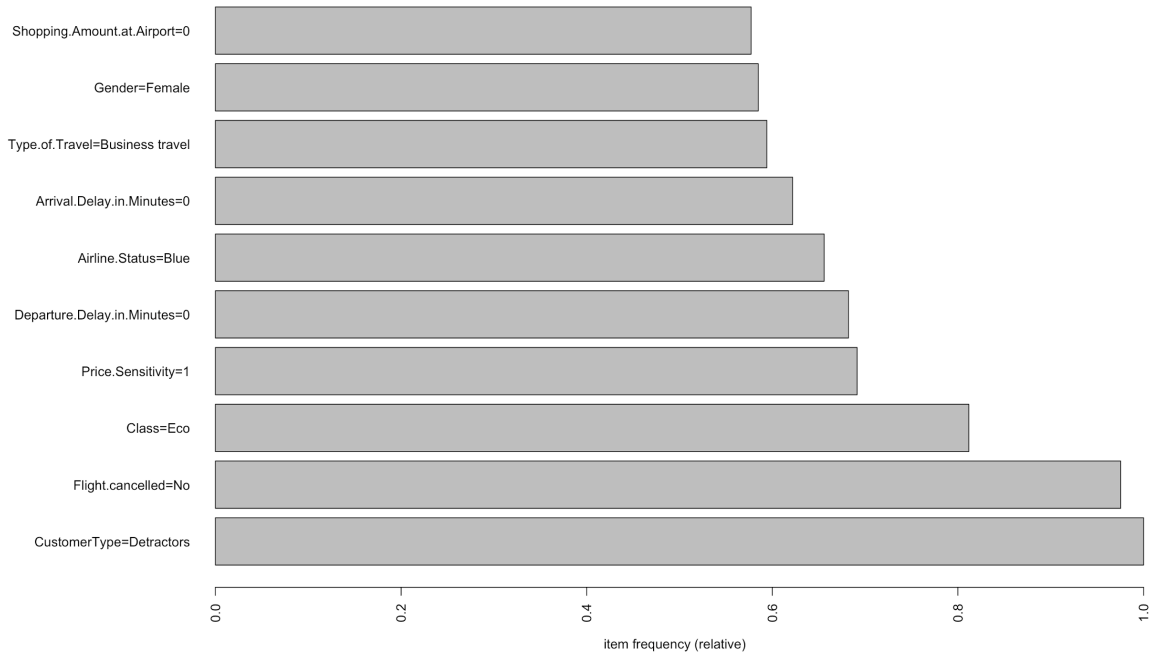
Code:

```
#MCA/PCA  Method
#Remove SATISFACTION

a <- data.frame(lapply(ClientData, as.factor))
a <- a[, -1]

## Fit MCA

library(FactoMineR)
library(factoextra)


res.mca <- MCA(a)
res = dimdesc(res.mca, axes = 1:2, proba = 0.05)
res$`Dim 1`$quali

#
###
ind <- get_mca_ind(res.mca)
var <- get_mca_var(res.mca)
# Coordinates
head(var$coord)
# Cos2: quality on the factore map
head(var$cos2)
# Contributions to the principal components
head(var$contrib)

summary.MCA(res.mca)

## eigvalue
eig.val <- res.mca$eig
barplot(
  eig.val[, 2],
  names.arg = 1:nrow(eig.val),
  main = "Variances Explained by Dimensions ",
  xlab = "Principal Dimensions",
  ylab = "Percentage of variances",
  col = "saddlebrown"
)
# Add connected line segments to the plot
lines(
  x = 1:nrow(eig.val),
```

```
  eig.val[, 2],
  type = "b",
  pch = 19,
  col = "red"
)


fviz_eig(res.mca)
fviz_mca_biplot(res.mca)
fviz_mca_ind(res.mca)
fviz_mca_var(res.mca)
fviz_mca_var(res.mca, choice = "mca.cor",
             repel = TRUE, # Avoid text overlapping (slow)
             ggtheme = theme_minimal())
#these two functions have a loading error they do not load and generates an error
fviz_mca_var(res.mca,
             repel = TRUE, # Avoid text overlapping (slow)
             ggtheme = theme_minimal())
fviz_mca_var(res.mca, col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE, # Avoid text overlapping
             ggtheme = theme_minimal())

# Contributions of rows to dimension 1
fviz_contrib(res.mca, choice = "var", axes = 1, top = 15)
# Contributions of rows to dimension 2
fviz_contrib(res.mca, choice = "var", axes = 2, top = 15)
```

B. Association Rules

Association Rule Mining is a common technique used to find associations between many variables. It is based on the notion that if you buy a certain group of items, you are more or less likely to buy another group of items.



For each item we select their frequency by taking 10 attributes into account like shopping amount, gender, type of travel, arrival delay, status of the airlines, class of the airline in which the passenger is travelling, whether the flight was cancelled or not and the type of the customer whether he is a detractor or not.

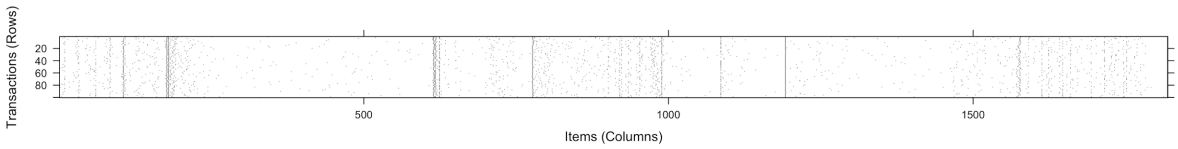
Here association rule mining is done to get the frequency of the items.

Here, we have taken shopping amount is shopping amount at the airport at customers as 0. Gender type is selected as female, the type of travel is business, delay in arrival in minutes is 0,

status of the airline is blue, price sensitivity it is 0 and the class of the airline is economy. The status of the flight is cancelled in this case and the customers are the detractors.

Item frequency is from 0.0 to 1 and it is plotted on the x axis while attributes are posted on the y axis.

The transaction of the items are shown on the next plot.



Code:

```
159 ##########################**********association rule minig********################
160
161 library(arules)
162 library(arulesViz)
163
164 ARData <- ClientData
165
166 ARData$CustomerType<-(cut(ARData$Likelihood.to.recommend,breaks = c(0.5,floor(mea
167 vBuckets <- replicate(length(ARData$CustomerType), "Detractors")
168 vBuckets[ARData$CustomerType >3] <- "Y"
169 ARData$CustomerType <- vBuckets
170
171 for (i in 1:ncol(ARData)) {
172   ARData[, i] <- factor(ARData[, i])
```

```r
}
train <- SamF(ARData, 0.7)[[1]]
test <- SamF(ARData, 0.7)[[2]]
ARDataX <- as(train, "transactions")
inspect(ARDataX[1:5])
summary(ARDataX)
itemFrequencyPlot(ARDataX, topN = 10, horiz = T)
image(sample(ARDataX, 100))

RuleDF <-apriori(ARDataX ,parameter = list(support = 0.1, confidence = 0.8),appea
RuleDF <-apriori(train ,parameter = list(support = 0.2, confidence = 0.8),appeara

inspect(RuleDF [1:20])

ordered_rules <- sort(RuleDF, by = "lift")
inspect(ordered_rules[1:20])
```

Output:

```
Source
Console  Terminal  Jobs
[1]  {Airline.Status=Blue,
      Type.of.Travel=Personal Travel,
      Shopping.Amount.at.Airport=0,
      Flight.cancelled=No}            => {CustomerType=Detractors} 0.1450617  0.8785047 2.070076    94
[2]  {Airline.Status=Blue,
      Type.of.Travel=Personal Travel,
      Total.Freq.Flyer.Accts=0,
      Shopping.Amount.at.Airport=0}   => {CustomerType=Detractors} 0.1095679  0.8765432 2.065455    71
[3]  {Airline.Status=Blue,
      Gender=Female,
      Type.of.Travel=Personal Travel,
      Class=Eco}                      => {CustomerType=Detractors} 0.1095679  0.8765432 2.065455    71
[4]  {Airline.Status=Blue,
      Type.of.Travel=Personal Travel,
      Shopping.Amount.at.Airport=0}   => {CustomerType=Detractors} 0.1496914  0.8738739 2.059165    97
[5]  {Airline.Status=Blue,
      Type.of.Travel=Personal Travel,
      Shopping.Amount.at.Airport=0,
      Class=Eco,
      Flight.cancelled=No}            => {CustomerType=Detractors} 0.1280864  0.8736842 2.058718    83
[6]  {Airline.Status=Blue,
      Type.of.Travel=Personal Travel,
      Total.Freq.Flyer.Accts=0,
      Shopping.Amount.at.Airport=0,
      Flight.cancelled=No}            => {CustomerType=Detractors} 0.1049383  0.8717949 2.054266    68
[7]  {Airline.Status=Blue,
      Gender=Female,
      Type.of.Travel=Personal Travel,
      Class=Eco,
      Flight.cancelled=No}            => {CustomerType=Detractors} 0.1049383  0.8717949 2.054266    68
[8]  {Airline.Status=Blue,
```

```
     lhs                                    rhs                          support confidence     lift count
[1]  {Airline.Status=Silver}             => {CustomerType=Promoters}  0.1651235  0.8699187 1.511280   107
[2]  {Airline.Status=Silver,
      Gender=Female}                     => {CustomerType=Promoters}  0.1003086  0.8441558 1.466523    65
[3]  {Airline.Status=Silver,
      Type.of.Travel=Business travel}    => {CustomerType=Promoters}  0.1280864  0.9120879 1.584539    83
[4]  {Airline.Status=Silver,
      Arrival.Delay.in.Minutes=0}        => {CustomerType=Promoters}  0.1095679  0.9220779 1.601894    71
[5]  {Airline.Status=Silver,
      Departure.Delay.in.Minutes=0}      => {CustomerType=Promoters}  0.1172840  0.9156627 1.590749    76
[6]  {Airline.Status=Silver,
      Price.Sensitivity=1}               => {CustomerType=Promoters}  0.1280864  0.8736842 1.517821    83
[7]  {Airline.Status=Silver,
      Class=Eco}                         => {CustomerType=Promoters}  0.1435185  0.8532110 1.482254    93
[8]  {Airline.Status=Silver,
      Likelihood.to.recommend=Y}         => {CustomerType=Promoters}  0.1651235  0.8699187 1.511280   107
[9]  {Airline.Status=Silver,
      Flight.cancelled=No}               => {CustomerType=Promoters}  0.1604938  0.8739496 1.518282   104
[10] {Type.of.Travel=Business travel,
      Total.Freq.Flyer.Accts=1}          => {CustomerType=Promoters}  0.1296296  0.8316832 1.444854    84
[11] {Gender=Female
```

C. Random Forest

Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object.

The independent variables in random forest are:

- Airline Status
- Age
- Class
- Gender
- Price Sensitivity
- Year of first flight
- Loyalty
- Type of travel
- Eating and Drinking at Airport
- Departure Delay in mins
- Arrival Delay in mins
- Likelihood to recommend

We build a Random Forest model with the remaining variables using all of our client data and predicted satisfaction based on the Random Forest model by using ntree=100. The number of variables available for splitting at each tree node - mtry is 3.

The mean of squared residuals is 2.318 and we found out the variance to be 53.7%.

Code:

```
#install.packages("randomForest")
library(randomForest)

Random_Forest<- df1[c(3,4,5,6,7,9,10,13,14,22,23,27)]
str(Random_Forest)
Random_Forest$Gender<- as.factor(Random_Forest$Gender)
Random_Forest$Age<- as.numeric(Random_Forest$Age)
Random_Forest$Loyalty<- as.numeric(Random_Forest$Loyalty)
Random_Forest$Price.Sensitivity<- as.numeric(Random_Forest$Price.Sensitivity)
Random_Forest$Departure.Delay.in.Minutes<- as.numeric(Random_Forest$Departure.Delay.in.Minutes)
Random_Forest$Year.of.First.Flight<- as.numeric(Random_Forest$Year.of.First.Flight)
Random_Forest$Eating.and.Drinking.at.Airport<- as.numeric(Random_Forest$Eating.and.Drinking.at.Airport)
Random_Forest$Class<- as.factor(Random_Forest$Class)


randomforest_classifier = randomForest(Likelihood.to.recommend ~ ., data=Random_Forest, ntree=100, mtry=3, importance=TRUE)
randomforest_classifier
```

Output:

```
Call:
 randomForest(formula = Likelihood.to.recommend ~ ., data = df2,      ntree = 100, mtry = 3, importance = TRUE)
               Type of random forest: regression
                     Number of trees: 100
No. of variables tried at each split: 3

          Mean of squared residuals: 2.318063
                    % Var explained: 53.7
```

D. Support Vector Machine

Support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.

The independent variables in SVM are:

- Airline Status
- Flight Cancelled
- Class
- Age Buckets
- Price Sensitivity
- Loyalty
- Type of travel
- Customer Type

- Departure Delay in mins
- Arrival Delay in mins
- Flights per year

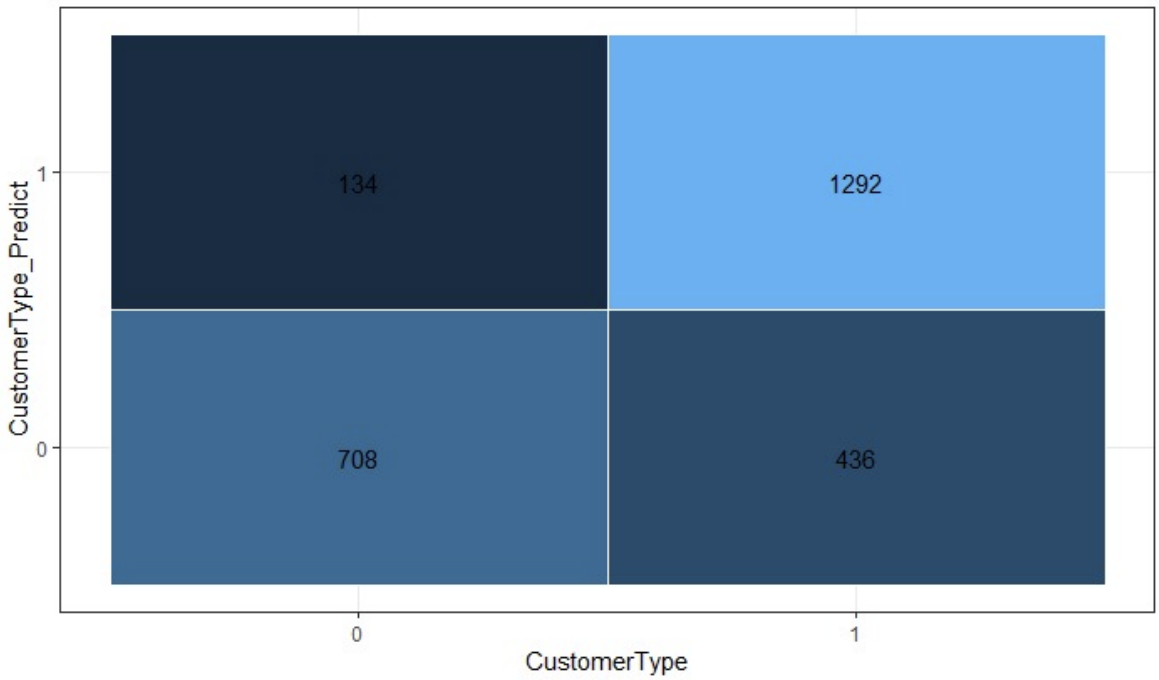
We build an SVM model with the remaining variables using the entire data of our client data and predicted the satisfaction. Here the, C=5 stands for a low cost of constraints parameter which implies that the point of differentiation for deciding the value of dependent variable cut is more generalized. It also implies more bias and lesser variance which is good for any data holistically. However it might give more cross validation errors as it will classify on a more generalized scale. If C is high then it will predict values and generate algorithm which will generate few errors for trained dataset as bias is very low and variance is high, but one which will generate many more errors for test dataset or other models.

We have used a three-fold cross validation model. The number of support vectors is 4022. The training error is found out to be 0.201 and the cross validation error is 0.226. Finally, we were able to achieve an accuracy rate of 77.8%.

```
##SVM Creating a Random Index
rand<-sample_n(df1,10282,replace = F)

View(rand)
# Cutoff point between Training and Test Dataset
cutpoint<-round(dim(rand)[1]*(3/4))
df1$Departure.Delay.in.Minutes
df1$Loyalty

mean(df1$Departure.Delay.in.Minutes)
library()
df1$Airline.Status
View(df1)
# Training and test datasets have been made
trainset<- rand[1:cutpoint,              c("Airline.Status","Flight.cancelled","Price.Sensitivity",
                                           "Arrival.Delay.in.Minutes","Departure.Delay.in.Minutes",
                                           "CustomerType","Loyalty","Age_Buckets",
                                           "Class","Type.of.Travel","Flights.Per.Year")]
testset<-rand[(cutpoint+1):(dim(rand)[1]),c("Airline.Status","Flight.cancelled","Price.Sensitivity",
                                           "Arrival.Delay.in.Minutes","Departure.Delay.in.Minutes",
                                           "CustomerType","Loyalty","Age_Buckets",
                                           "Class","Type.of.Travel","Flights.Per.Year")]
install.packages("kernlab")
library(kernlab)

svmOutput<-ksvm(CustomerType~.,data=trainset,kpar="automatic",kernal = "rbfdot", C = 5, cross = 5,
                prob.model = TRUE)

print(svmOutput)
str(trainset)
str(testset)

svmOutput
```

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

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

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.165815999497183

Number of Support Vectors : 4022

Objective Function Value : -16692.77
Training error : 0.201115
Cross validation error : 0.226141
Probability model included.
>
```

```
# Now Predicting for test dataset
svmpred<-predict(svmOutput,testset,type="votes")
# Creating a confusion Matrix

confusionMatrix<- data.frame(testset$CustomerType,svmpred[2,])
confusionMatrixTable<- table(confusionMatrix)

confusionMatrixTable
errorRate<-((confusionMatrixTable[1,2]+confusionMatrixTable[2,1])/sum(confusionMatrixTable))
errorRate

accuracyRate<-((confusionMatrixTable[1,1]+confusionMatrixTable[2,2])/sum(confusionMatrixTable))
accuracyRate
# Accuracy is 77.8%
```



While running the model successfully, we observed that our model correctly predicted 1292 customers as Promoters and falsely predicted 436 Promoters as detractors. The model also correctly predicted 708 customers as Detractors and falsely predicted 134 Detractors as Promoters.

VI. Actionable Insights

- Older people

Customers aged from 60-85 tend average a lower satisfaction. This could be resolved by providing more personal travel amenities for senior customers and taking additional care. For instance, Southeast airlines could provide wheelchairs at the airport.

- Female

We can observe that all female related components appear to be in dissatisfaction. We suggest that the Southeast airlines should provide more feminine care to improve the satisfaction level. Additionally, they can also provide more help to the mothers who carry their babies. For instance, we can provide a separate area for them.

- Personal Travel

It is observed that the people who are travelling on their own, tend to give a lower satisfaction rating. Therefore, in order to improve the user experience of personal travel customers, we suggest that a survey be conducted based on these customers. Further analysis on the survey results can help Southeast to come up with a specific plan to increase their satisfaction level.

- Blue status

Passengers in the blue status are the most yet they tend to give low satisfaction. One reason for this could be that they do not get the service which they had expected. The Southeast airlines definitely need to change some policies for blue status customers and improve their service towards their clients.

- Economy Plus Class

Out of all the classes, the passengers of the Economy Plus class were the least satisfied, in spite Economy Plus being more expensive and potentially meant to be better than Economy. The airlines needs to investigate their Economy Plus services and management to find out what exactly is making their customers unhappy. A possible reason could be that the customers have paid more for Economy Plus but did not feel that the additional services were worth the extra cost.

VII. Recommendations:

- The Southeast Airlines should target adult passengers between the ages of 23-60.
- They should be targeting business travelers.
- The insights indicate that people get detracted when there is a delay in arrival or departure.
- Female passengers belonging to the Silver Airline status should be targeted.
- Insights indicate that customers get detracted when the flight is cancelled.
- Southeast Airlines should strive to keep their prices less as to not detract the passengers.