Airline Analysis

IST687: Introduction to Data Science Group 1

Group Members:
Aditya Kini
Praneshwar Srinivasan
Yujing Yuan
Ushma Desai

TABLE OF CONTENT

1.INTRODUCTION

The airplane easily one of the most influential inventions of the 20th century, if not all time. Air travel really made this world small, changing travel times from months to hours. Southeast Airlines is one of the top four airlines in the United States. Southeast Airlines needed to lower their customer churn (sometimes referred to as customer attrition).

Additionally, customer churn is actually a lagging indicator, meaning the loss has already occurred. As such, it was a measurement of the damage inflicted. The real goal is to reduce churn by getting ahead of the loss (of the customer) by identifying some leading indicators, or metrics, that might help keep a customer.

Southeast often surveyed their customers, and in fact, possessed thousands of recently completed customer surveys. Southeast has been using the surveys to calculate NPS. The concept of NPS is that customers who are promoters are good customers to keep. Customers who are detractors are really problematic in that they may actively tell their social connections not to use the product or service.

1.1 PROJECT OVERVIEW

The survey dataset contained thousands of observations of flight segment data collected by Southeast Airlines. Each row represents one flight segment, by one airline (either southeast or one of its partner airlines), for a specific customer. Each column represents an attribute of that particular flight segment.

Each row captures 26 characteristics of the flight (ex. day of month, date, airline, origin and destination city, if the flight was delayed), the customer (ex. age, gender, price sensitivity, the person's frequent flyer status). The row also contains a simple survey-based rating of each customer's likelihood to recommend the airline that they just flew as well as a field for open-ended text comments.

In our analysis, we have identified and reviewed the key characteristics like flight cancellations, flight delays, partner name etc. This gave us an insight on the feedback provided by the customers and helped us analyze it using different data analysis models. We found out various trends and relations between variables of the data set and likelihood to recommend using the analysis models which thus helped us answer business questions formulated for the data set. After answering the business questions we could work out on better insights and solutions for increasing the customer satisfaction and airline services.

2. BUSINESS QUESTIONS

Following business questions which will be considered in the project:

Q1. Which customers are happy and which customers are unhappy based on the NPS?

Q2. What makes a successful flight experience?

Q3. How to make customers happy in the future?

3. DATA MUNGING

CLEANING

We cleaned the data before performing any Descriptive analysis or Modeling techniques by first updating the Column names according to naming convention and changing all the ‘.’ To ‘_’. Also, we have removed the last column as the Likelihood_to_recommend column appropriately tells us how much satisfied the customer is which corresponds to a Positive comment (For Promoters) and a negative comment (For Detractors).Thereby we didn’t need the ‘freeText’ column and thus removed it.

Code Snippet

```
colnames(cleaned_data) <- colnames(cleaned_data) %>% str_replace_all("\\.","_")

cleaned_data <- cleaned_data[,-which(colnames(cleaned_data)=="freeText")]
```

Next, We cleaned the NA values detected in the ‘Arrival Delay’ and ‘Departure Delay’ columns by substituting the NA with the median of the data as the median is a proper measure of central tendency.

Code Snippet

```
cleaned_data$Departure_Delay_in_Minutes[which(is.na(cleaned_data$Departure_Delay_in_Minutes))] = median(cleaned_data$Departure_Delay_in_Minutes,na.rm = TRUE)

cleaned_data$Arrival_Delay_in_Minutes[which(is.na(cleaned_data$Arrival_Delay_in_Minutes))] = median(cleaned_data$Arrival_Delay_in_Minutes,na.rm = TRUE)
```

Finally, we needed to clear the NA values in the Flight time Column. We found that the flight time is also dependent on the flight distance and therefore can’t just be substituted by mean or median, as this could result in 2 flights of different distances (i.e. 200 and 2000 miles) ending up with the same flight time.

Therefore to clean the NA’s, we first ordered the data frame according to the ‘Flight_distance’ column and then using the ‘na_interpolation’ function to interpolate the NA’s between flights of almost the same distance.

Code Snippet

```
cleaned_data <- cleaned_data %>% arrange(Flight_Distance)

cleaned_data$Flight_time_in_minutes <-
na_interpolation(cleaned_data$Flight_time_in_minutes)
```

TRANSFORMATION

After the cleaning process was completed, we transformed a few existing columns as well as added new columns of our own to aid in the analysis and prediction process.

First, we transformed the age column into 4 categories ‘1’ – Between Ages 15 and 32, ‘2’ – Between Ages 33 and 44, ‘3’ – Between Ages 45 and 58 and ‘4’ – Age 59 and above.

We found these intervals by getting the summary of the age column wherein

1 – Between Min value and 1st Quartile

2 – Between 1st Quartile and Median

3 – Between Median and 3rd Quartile

4 – Between 3rd Quartile and Max Value (or in this case just greater than 3rd Quartile)

Code Snippet

```
summary(transform_data$Age)

for (var in 1:dim(transform_data)[1]) {

  if ((transform_data$Age[var]>=15) & (transform_data$Age[var]<33))

            { transform_data$t_age[var] <- 1

  } else if ((transform_data$Age[var]>=33) & (transform_data$Age[var]<45))

            { transform_data$t_age[var] <- 2

  } else if ((transform_data$Age[var]>=45) & (transform_data$Age[var]<59))

            { transform_data$t_age[var] <- 3
```

```
 } else

          transform_data$t_age[var] <- 4

}
```

Next, we created a new variable Arrival_delay_greater_5 which took a value of 1 if the flight arrival delay was greater than 5 and 0 otherwise.

Code Snippet

```
summary(transform_data$Arrival_Delay_in_Minutes)

for (var in 1:dim(transform_data)[1]) {

  if(transform_data$Arrival_Delay_in_Minutes[var] > 5) {

    transform_data$Arrival_Delay_Greater_5[var] = 1 }

  else

    transform_data$Arrival_Delay_Greater_5[var] = 0

}
```

Similarly, we create another variable ‘Long_duration_flight’ which is Boolean and takes values TRUE when flight duration is greater than 92 and FALSE when flight duration is less than 92. We arrived at the value 92 as this was the median of the flight time column.

Code Snippet

```
summary(transform_data$Flight_time_in_minutes)

for (var in 1:dim(transform_data)[1]) {

  if(transform_data$Flight_time_in_minutes[var] > 92) {

    transform_data$Long_duration_flight[var] = TRUE }

  else

    transform_data$Long_duration_flight[var] = FALSE}
```

Finally, we create the LoyaltyIndex variable which takes in 4 values – ‘1’ When loyalty between -1 and -0.5, ‘2’ – When loyalty between -0.5 and 0, ‘3’ – When loyalty between 0 and 0.5, ‘4’ – When loyalty between 0.5 and 1

The intervals denote the following type of people

1 – Disloyal Customers

2 – Likely Disloyal Customers

3 – Likely Loyal Customers

4 – Loyal Customers

Code Snippet

```
summary(transform_data$Loyalty)

for (var in 1:dim(transform_data)[1]) {

  if((transform_data$Loyalty[var] >= -1) && (transform_data$Loyalty[var] < -0.5)) {

            transform_data$LoyalityIndex[var] = 1 }

  else if((transform_data$Loyalty[var] >= -0.5) && (transform_data$Loyalty[var] < 0)) {

    transform_data$LoyalityIndex[var] = 2 }

  else if((transform_data$Loyalty[var] >= 0) && (transform_data$Loyalty[var] < 0.5)) {

    transform_data$LoyalityIndex[var] = 3 }

  else

    transform_data$LoyalityIndex[var] = 4

}
```

We have also created other variables to help us during Visualizations and these variables have been included in the visualizations part for better understanding of their use.

The variables are as follows

- Arrival_group
- Departure_group
- Age_Category
- Passenger_Category
- Passenger_Sensitivity
- Delay_Group

4. USE OF DESCRIPTIVE STATISTICS AND DATA VISUALIZATION

4.1 Customer Attributes

Now moving into the next part we will be considering various variables for creating plots for the purpose of predicting and answering the business questions in order to help to increase the number of happy customers by reducing the churn.First of all we need to install few libraries and then we use the data to generate plots..

Code:

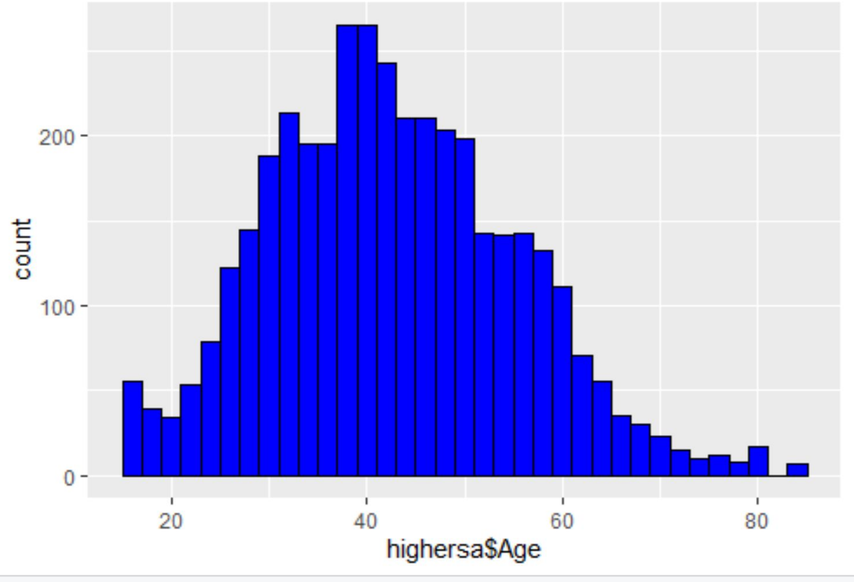
```
#Installing library
install.packages(RCurl)
library(RCurl)

install.packages(jsonlite)
library(jsonlite)

install.packages(ggplot2)
library(ggplot2)

install.packages(tidyverse)
library(tidyverse)
```

Firstly, the likelihood to recommend is divided by two categories. The likelihood over 8 is seen as having higher possibility to recommend and the likelihood under 7 is seen as having lower possibility to recommend.

```
lowersa<-  jsonfile_data[which(airlines$Likelihood.to.recommend<7),]
highersa<- jsonfile_data[which(airlines$Likelihood.to.recommend>8),]
```

Code:

```
highersa%>%
  ggplot()+
  aes(x=highersa$Age)+
geom_histogram(binwidth = 2, color="black", fill="blue")
```

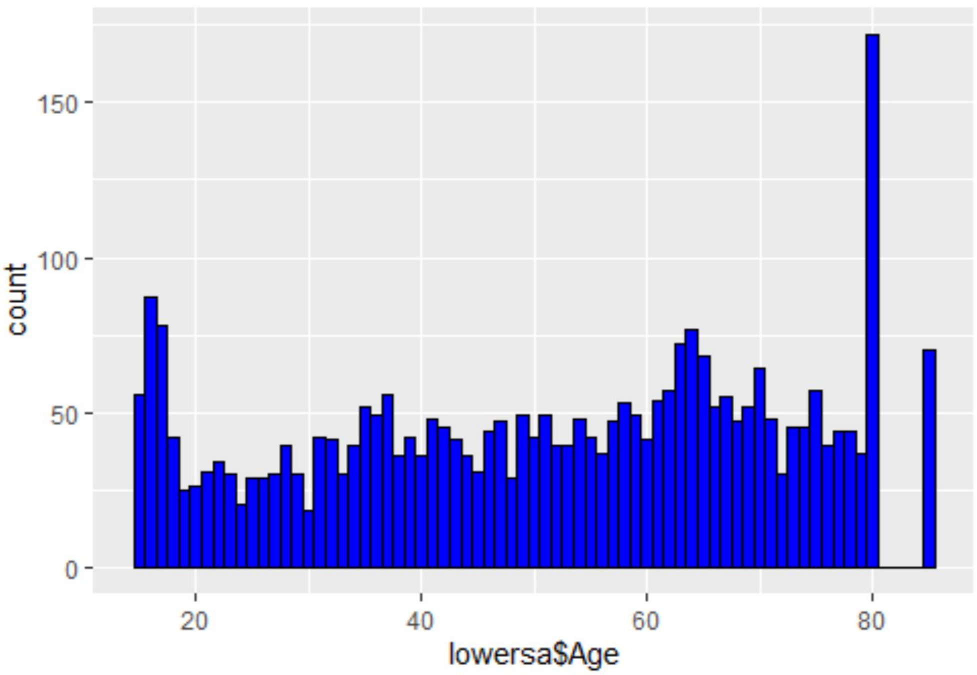


From this chart, The bar chart that reflects which age range of customers have higher possibility to recommend, and it shows that customers' age from 30 to 50 has the higher possibilities to be satisfied. In addition, we also generated a bar plot for lower likelihood to recommend.

Code:

```
lowersa%>%
  ggplot()+
  aes(x=lowersa$Age)+
  geom_histogram(binwidth = 1, color="black", fill="blue")
```

Visualization:



The bar chart shows that the customers' age under 20 and over 80, and from 60 to 70 have higher possibilities of being unsatisfied.

NPS and Customer Attributes

Now we have a general understanding that different age groups have different possibilities to recommend. However, we should further investigate other customer attributes that might influence the NPS.

Hence a few attributes which are related to customers like Age,Gender and class. Along with this attributes we will calculate NPS which is Net Promoter Score. So we can say that NPS asks customers to respond, on a scale of 1 -10, to one simple question: “How likely is it that you will recommend our airline to a friend or colleague?”. If respondents score less than 7, they’re detractors. If they scored above an 8, they’re promoters. In the middle range (a score of 7 or 8), then they’re “passive”. In a given group, subtracting the percent of respondents who are detractors from the percent of respondents who were promoters provides the overall NPS score

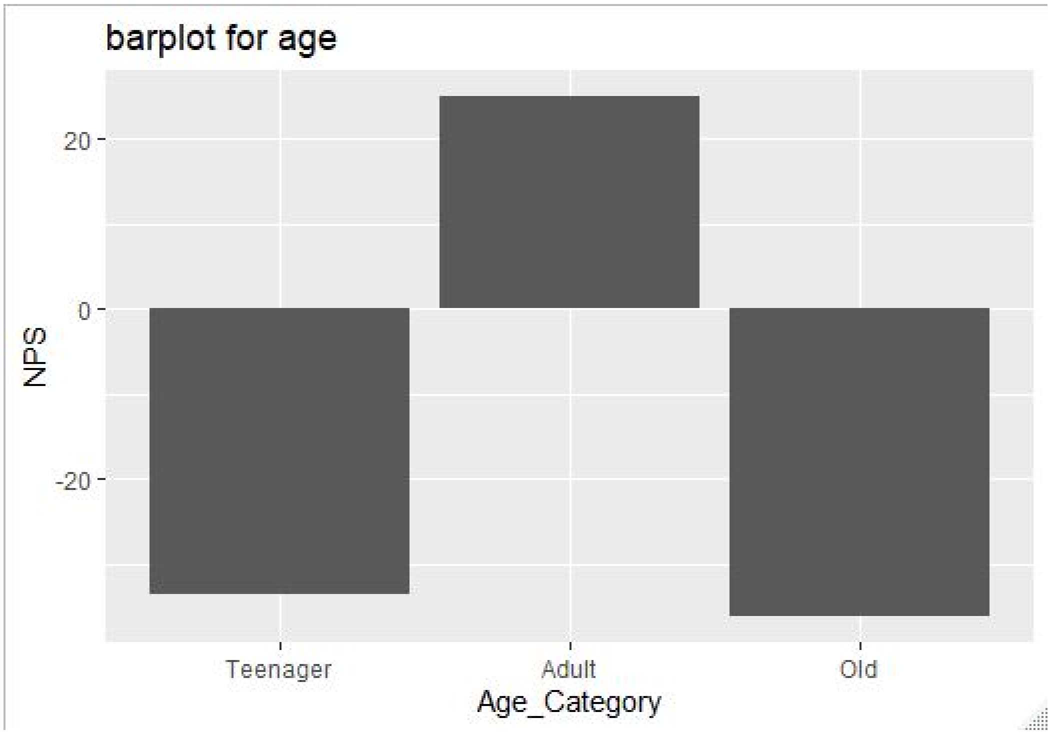
Now first we are considering generating a plot for NPS and Age.Before this we need to group the data. In the code below we notice Age Category and Passenger Category attributes which we were created by us. In Age Category we have grouped the customers into 4 categories based on age attribute. The customers who are between 0 to 10 are “Children” , the customers who are between 10 to 18 are “Teenagers” , the customers who are between 18 to 59 are “Adult” and customers who are between 59 to 100 are “Old”. Similarly to Age Category we have grouped the customers in 3 categories based on Likelihood to Recommend attributes.Hence the customers who have rated between 0 to 6 are “detractors”, the customers who rated between 6 to 8 are “passive” and the customers who rated between 8 to 10 are “promoters”.

Code:

1. **NPS vs Age**

```
age_detractor <- jsonfile_data %>%
  filter(Passenger_Category == "detractor") %>%
  group_by(Age_Category) %>%
  summarise(detractor_freq_age = n())

#We generated a table considering Passenger Category as Detractors and group the data by Age Category attribute

age_promoter <- jsonfile_data %>%
  filter(Passenger_Category == "promoter") %>%
  group_by(Age_Category) %>%
  summarise(promoter_freq_age = n())

#We generated a table considering Passenger Category as Promoters and group the data by Age Category attribute

age_total <- jsonfile_data %>%
  group_by(Age_Category) %>%
  summarise(total_freq_age = n())
```

```
#We generated a table considering all the customers and group the data by Age Category
attribute

Age_merged= merge(age_detractor,age_promoter, by="Age_Category")
Age_merged= merge(Age_merged,age_total,by ="Age_Category")

#From the above two lines of code we have merged all the three tables :
age_detractor,age_promoter and age_total into a single table Age_merged based on Age
Category attribute

Age_merged$Age_detractor_percent =
(Age_merged$detractor_freq_age/Age_merged$total_freq_age)*100

#Above we are calculating the percentage of detractors

Age_merged$Age_promoter_percent =
(Age_merged$promoter_freq_age/Age_merged$total_freq_age)*100

#Above we are calculating the percentage of promoters

Age_merged$NPS = Age_merged$Age_promoter_percent -
Age_merged$Age_detractor_percent

#Above we are calculating the NPS

#Hence now we have Age_merged as our final table and based on this table we can generate a
plot for NPS vs Age.

ggplot(Age_merged) +
  aes(x=Age_Category,y=NPS)+
  geom_col()+
  ggtitle("barplot for age")
```

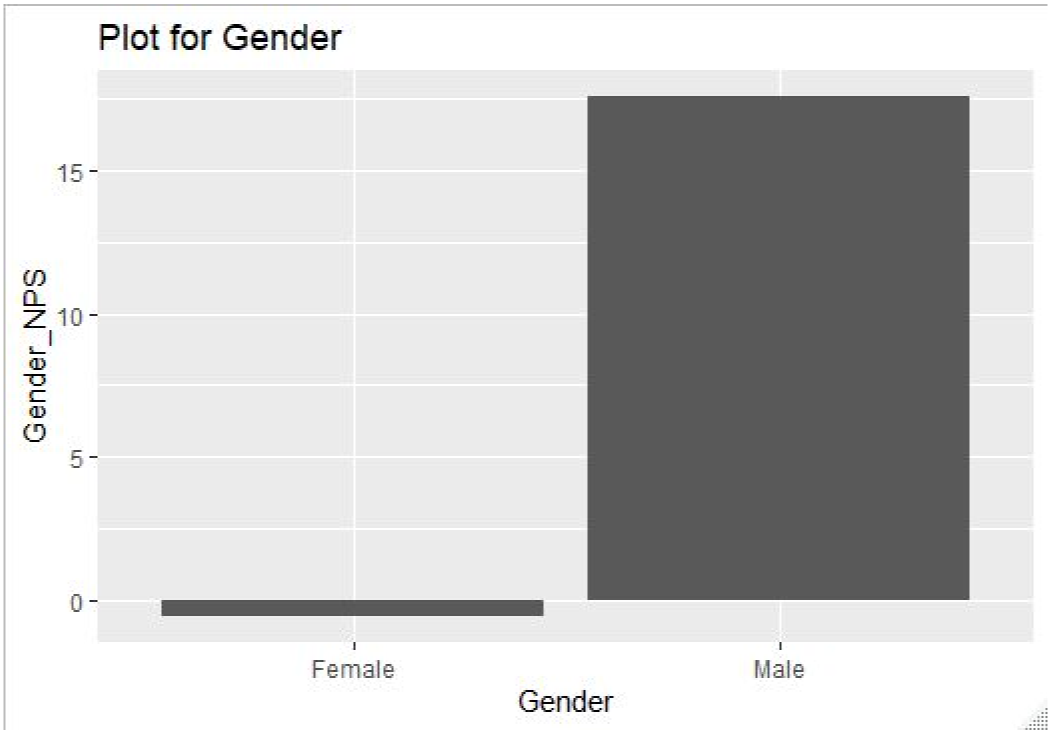


From the plot we can observe that the adults have the best positive NPS among all the age group category and they are Promoters.Since teenagers and old age groups have negative NPS are detractors.

2. **NPS vs Gender**

```
Gender_detractor <- jsonfile_data %>%
  filter(Passenger_Category == "detractor") %>%
  group_by(Gender) %>%
  summarise(detractor_freq_gender = n())

#We generated a table considering Passenger Category as Detractors and group the data by Gender attribute

Gender_promoter <- jsonfile_data %>%
  filter(Passenger_Category == "promoter") %>%
  group_by(Gender) %>%
  summarise(promoter_freq_gender = n())

#We generated a table considering Passenger Category as Promoters and group the data by Gender attribute

Gender_total <- jsonfile_data%>%
  group_by(Gender) %>%
  summarise(total_freq_gender = n())

#We generated a table considering all the customers and group the data by Gender attribute

Gender_merged = merge(Gender_detractor,Gender_promoter,by="Gender")
Gender_merged = merge(Gender_total,Gender_merged,by="Gender")

#From the above two lines of code we have merged all the three tables : Gender_detractor,Gender_promoter and Gender_totalinto a single table Gender_merged based on Gender attribute

Gender_merged$Gender_detractor_percent=(Gender_merged$detractor_freq_gender/Gender_merged$total_freq_gender)*100

#Above we are calculating the percentage of detractors

Gender_merged$Gender_promoter_percent=(Gender_merged$promoter_freq_gender/Gender_merged$total_freq_gender)*100

#Above we are calculating the percentage of promoters

Gender_merged$Gender_NPS=Gender_merged$Gender_promoter_percent - Gender_merged$Gender_detractor_percent

#Above we are calculating the NPS
```

```
ggplot(Gender_merged)+
  aes(x=Gender,y=Gender_NPS)+
  geom_col()+
  ggtitle("Plot for Gender")
```



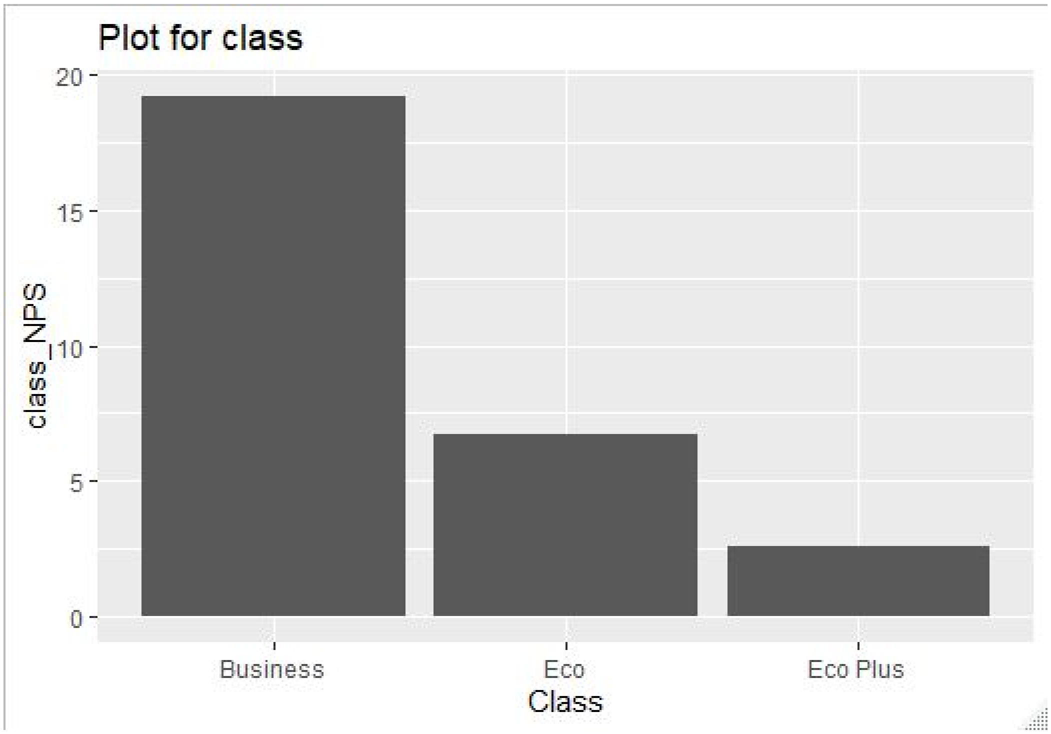
From the plot we can observe that both male and female have positive NPS but in comparison Male gender has a higher NPS than female. Hence we can say that male are promoters and female are detractors.

3. NPS vs Class

```
  Class_detractor <- jsonfile_data %>%
  filter(Passenger_Category == "detractor") %>%
  group_by(Class) %>%
  summarise(detractor_freq_class = n())
#We generated a table considering Passenger Category as Detractors and group the data by Class attribute
```

```
Class_promoter <- jsonfile_data %>%
  filter(Passenger_Category == "promoter") %>%
  group_by(Class) %>%
  summarise(promoter_freq_class = n())

#We generated a table considering Passenger Category as Promoters and group the data by
Class attribute

Class_total <- jsonfile_data %>%
  group_by(Class) %>%
  summarise(total_freq_class = n())

#We generated a table for total customers and group the data by Gender attribute

Class_merged = merge(Class_promoter,Class_detractor,by="Class")
Class_merged = merge(Class_total,Class_merged,by="Class")

#In the above two lines we merged all the three tables Class_promoter,Class_detractor and
Class_total based on Class attribute.

Class_merged$detractor_class_percent =
(Class_merged$detractor_freq_class/Class_merged$total_freq_class)*100

#Above we are calculating the percentage of detractors

Class_merged$promoter_class_percent =
(Class_merged$promoter_freq_class/Class_merged$total_freq_class)*100

#Above we are calculating the percentage of promoters

Class_merged$class_NPS = Class_merged$promoter_class_percent -
Class_merged$detractor_class_percent
#Above we are calculating the NPS
```

```
ggplot(Class_merged)+
  aes(x=Class,y=class_NPS)+
  geom_col()+
  ggtitle("Plot for class")
```



From the plot we can observe that in all the three class the Business class has the highest NPS followed Eco and Eco plus. Hence we can say that Business class are promoters and the least NPS is for Eco plus hence they are detractors.

.

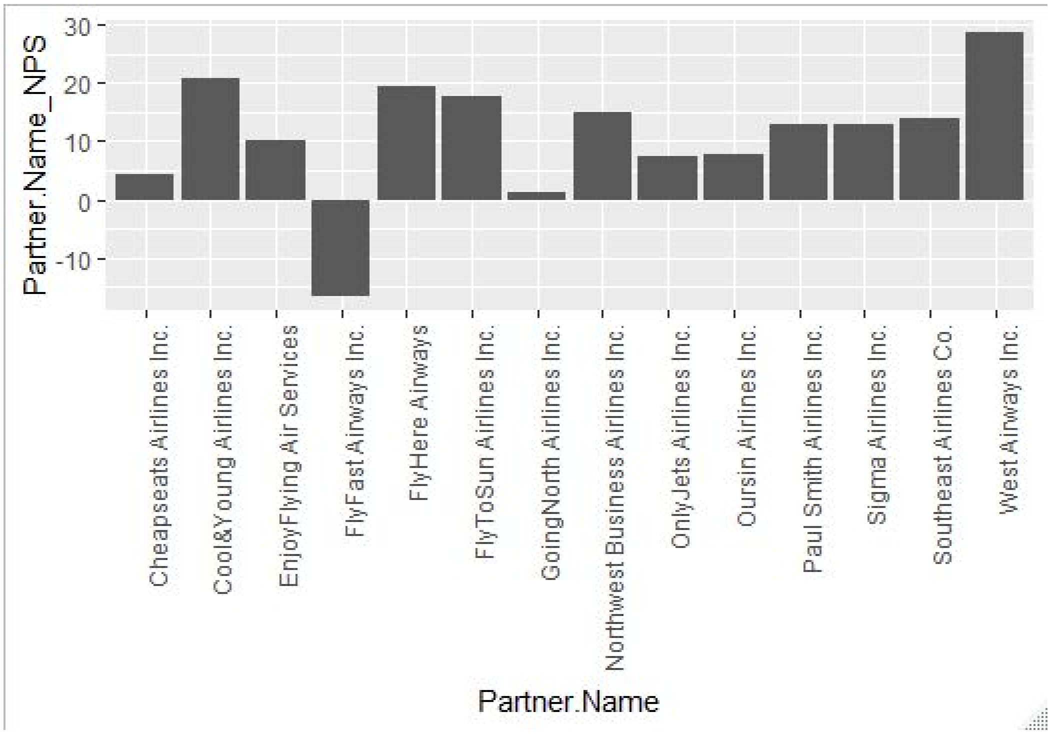
4.2 Flight Attributes

As we used various attributes above for the customer attribute we have similarly we used various attributes to build plots.Further we are calculating NPS and then generated a plot against Partner Airlines to understand which airlines have a positive impact and customers experienced a good service.Below is the code :

1.NPS vs Partner Airlines

```
  Partner.Name_detractor <- jsonfile_data %>%
  filter(Passenger_Category == "detractor") %>%
  group_by(Partner.Name) %>%
  summarise(detractor_freq_Partner.Name = n())

#We generated a table considering Passenger Category as Detractors and group the data by Partner name attribute

Partner.Name_promoter <- jsonfile_data %>%
  filter(Passenger_Category == "promoter") %>%
  group_by(Partner.Name) %>%
  summarise(promoter_freq_Partner.Name = n())

#We generated a table considering Passenger Category as Promoters and group the data by Partner name attribute


Partner.Name_total <- jsonfile_data %>%
  group_by(Partner.Name) %>%
  summarise(total_freq_Partner.Name = n())

#We generated a table for total customers and group the data by Partner Name attribute
```

```
Partner.Name_merged =
merge(Partner.Name_promoter,Partner.Name_detractor,by="Partner.Name")
Partner.Name_merged =
merge(Partner.Name_total,Partner.Name_merged,by="Partner.Name")

#From the above two lines of code we have merged all the three tables :
Partner.Name_promoter,Partner.Name_detractor and Partner.Name_total into a single table
Partner.Name_merged based on Partner Name attribute

Partner.Name_merged$detractor_Partner.Name_percent =
(Partner.Name_merged$detractor_freq_Partner.Name/Partner.Name_merged$total_freq_Partn
er.Name)*100

#Above we are calculating the percentage of detractors

Partner.Name_merged$promoter_Partner.Name_percent =
(Partner.Name_merged$promoter_freq_Partner.Name/Partner.Name_merged$total_freq_Partn
er.Name)*100
#Above we are calculating the percentage of promoters



Partner.Name_merged$Partner.Name_NPS =
Partner.Name_merged$promoter_Partner.Name_percent -
Partner.Name_merged$detractor_Partner.Name_percent

#Above line of code is calculating the NPS

ggplot(Partner.Name_merged)+
  aes(x=Partner.Name,y=Partner.Name_NPS)+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  geom_col()
```



From the plot we can observe that for the airlines partners only FlyFast airlines has a negative NPS and rest all the airlines partners have a positive NPS. Hence we can conclude from this plot that customers travelling by FlyFast are unhappy and are detractors. Rest of the airlines customers are happy and are promoters.

2. Frequency of departure delay

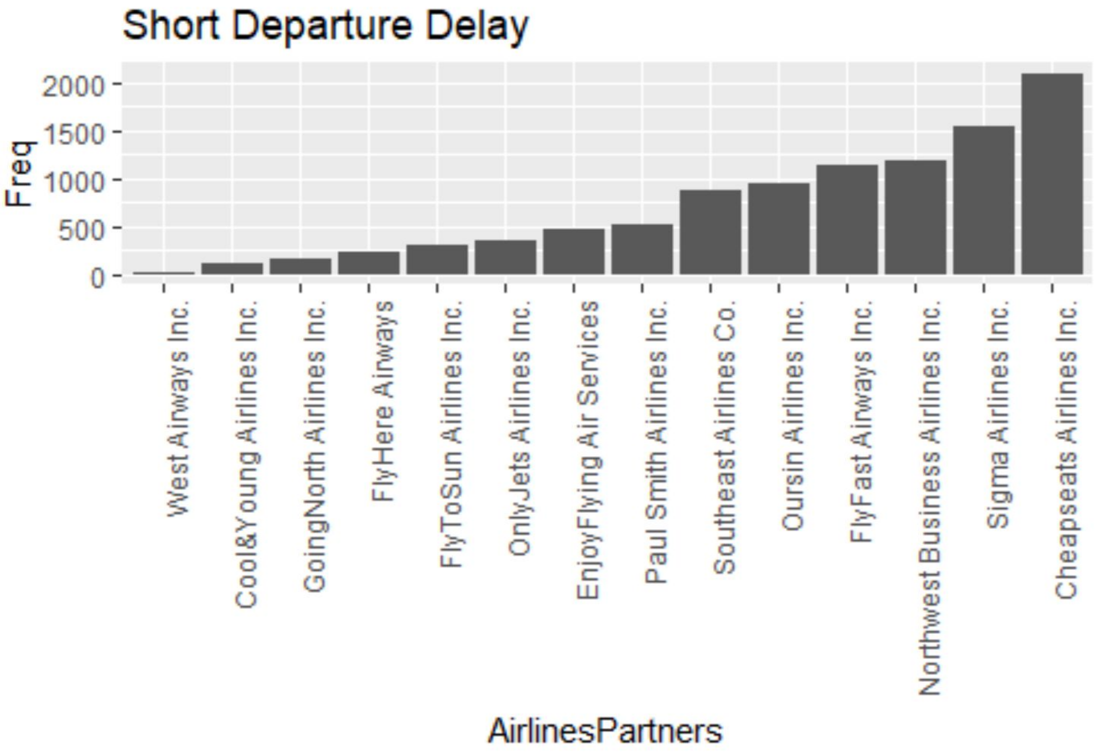
It is clear that the longer the flight is delayed, the less the customers will be satisfied. So we generated some bar charts that reflect different airlines partners and their delay time frequency. Firstly, we categorized the flight departure delay and arrival delay as short delay, medium delay and long delay. The short delay time is from 0 to 100 minutes, medium delay time is from 100 to 400 minutes and long delay time is from 400 to 730 minutes. All the NA values are replaced by the average delay time.

Code:

```
airlines$Departure.Delay.in.Minutes[is.na(airlines$Departure.Delay.in.Minutes)]<-
mean(airlines$Departure.Delay.in.Minutes,na.rm=TRUE)
airlines$Arrival.Delay.in.Minutes[is.na(airlines$Arrival.Delay.in.Minutes)]<-
mean(airlines$Departure.Delay.in.Minutes,na.rm=TRUE)

#Categorize the arrival delay
airlines <- airlines %>%
  mutate(Arrival_group = cut(Arrival_Delay_in_Minutes,
                             breaks = c(0, 100, 400, 730),
                             labels = c("low","medium","high")))
airlines$Arrival_group[is.na(airlines$Arrival_group)]<- c("low")
#Categorize departure delay
airlines <- airlines %>%
  mutate(Departure_group = cut(Departure_Delay_in_Minutes,
                               breaks = c(0, 100, 400, 730),
                               labels = c("low","medium","high")))
airlines$Departure_group[is.na(airlines$Departure_group)]<- c("low")
```

<u>Code:</u>

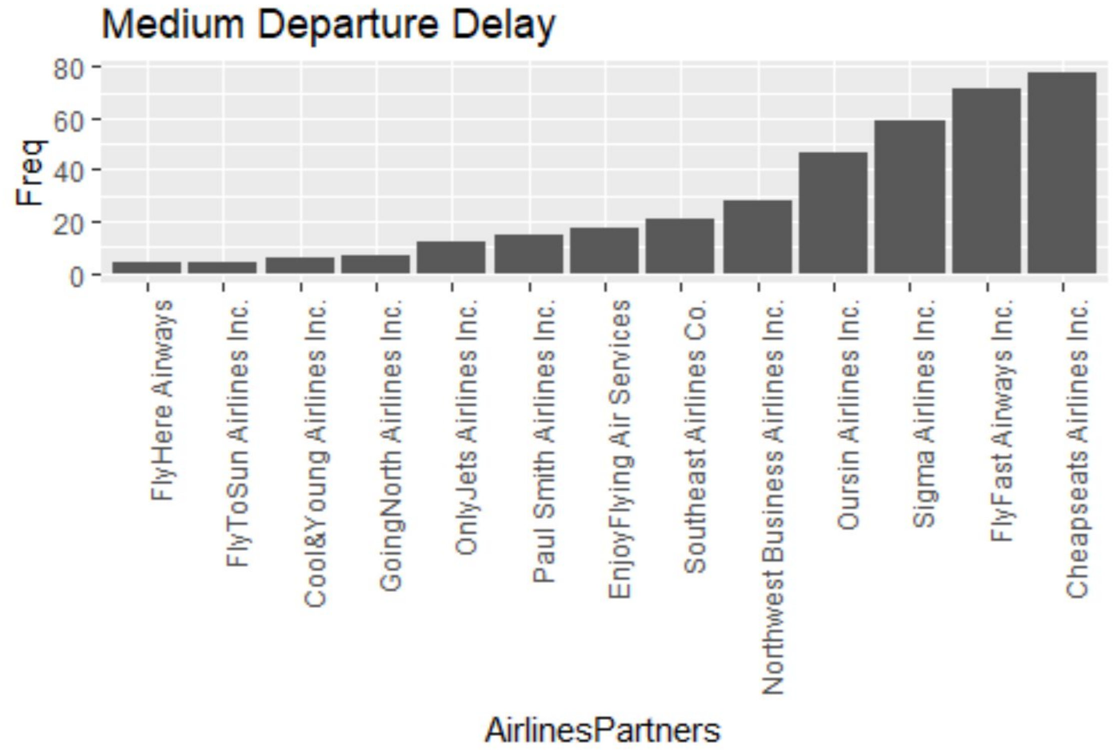
```
departure_low<- airlines%>%
  filter(airlines$Departure_group=="low")%>%
  group_by(Partner_Name)
view(departure_low)
delay_low<- as.data.frame(table(departure_low$Partner_Name))
view(delay_low)
colnames(delay_low)[which(names(delay_low) == "Var1")] <- "AirlinesPartners"

ggplot(delay_low)+
  aes(x=reorder(AirlinesPartners,Freq),y=Freq)+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  ggtitle("Short Departure Delay")+
  labs(x="AirlinesPartners")+
  geom_col()
```



From this bar plot, we can see that Cheapseats Airlines company has the most flights with short departure delay. In addition, Sigma also has relatively many flights that not delayed for too long.

Code:

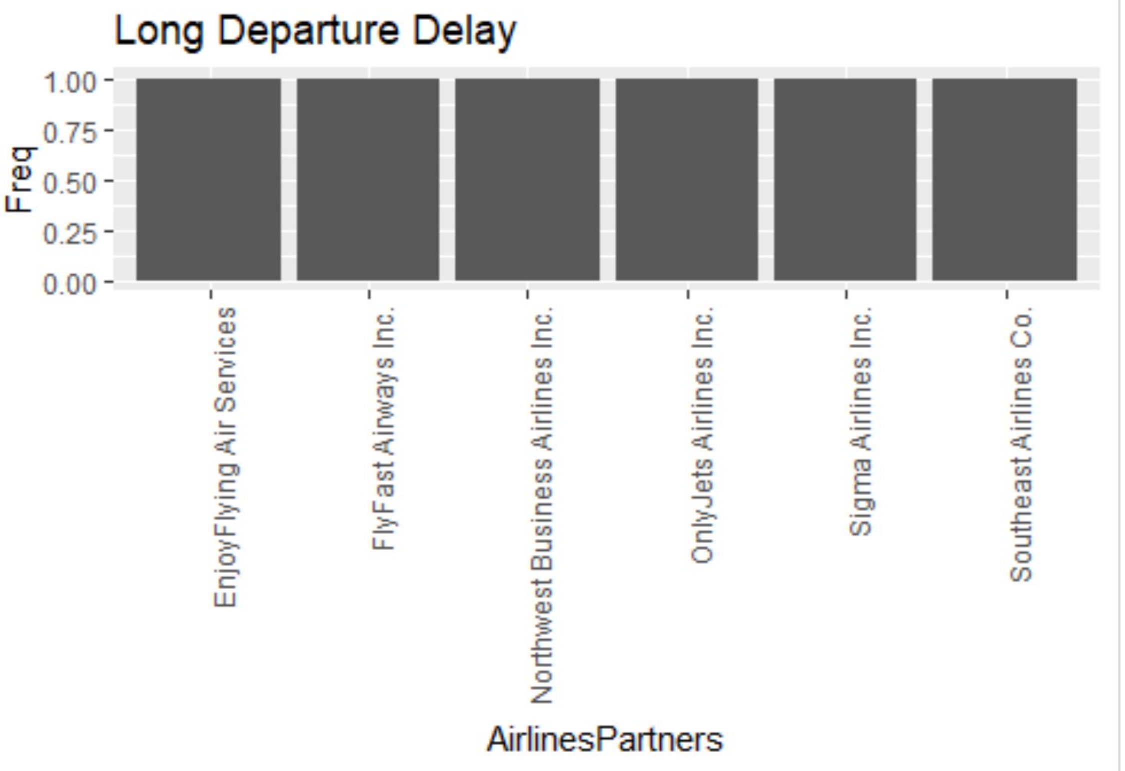
```
departure_medium<- airlines%>%
  filter(airlines$Departure_group=="medium")%>%
  group_by(Partner_Name)
delay_medium<- as.data.frame(table(departure_medium$Partner_Name))
view(delay_medium)
colnames(delay_medium)[which(names(delay_medium) == "Var1")] <- "AirlinesPartners"
ggplot(delay_medium)+
  aes(x=reorder(AirlinesPartners,Freq),y=Freq)+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  ggtitle("Medium Departure Delay")+
  labs(x="AirlinesPartners")+
  geom_col()
```



In this bar plot, we can see that Cheapseats still has the most medium time delay, which means the delay time for this airline partner is mainly short but with some exceptions. Furthermore, we found that Flyfast has relatively high possibility to have longer delay based on it's proportion of short delay and medium delay. Sigma's flights also have many medium delay time.

Code:

```
departure_high<- airlines%>%
  filter(airlines$Departure_group=="high")%>%
  group_by(Partner.Name)
delay_high<- as.data.frame(table(departure_high$Partner.Name))
view(delay_high)
colnames(delay_high)[which(names(delay_high) == "Var1")] <- "AirlinesPartners"
ggplot(delay_high)+
  aes(x=AirlinesPartners,y=Freq)+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  ggtitle("Long Departure Delay")+
  geom_col()
```

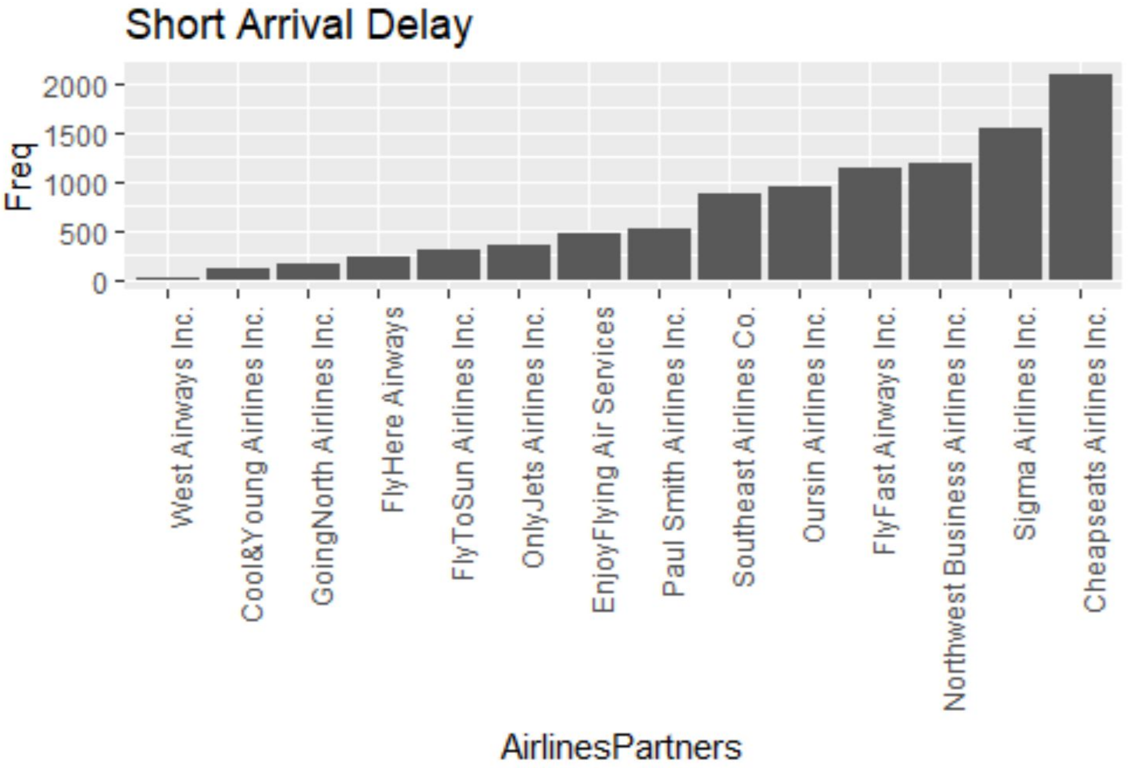


This chart shows that long departure delay is really rare, it may be due to some external condition that force the flight to delay. Overall, the lone departure delay appears only once for some airlines partners, but we still have to pay attention on it.

2. Frequency of arrival delay

Code:

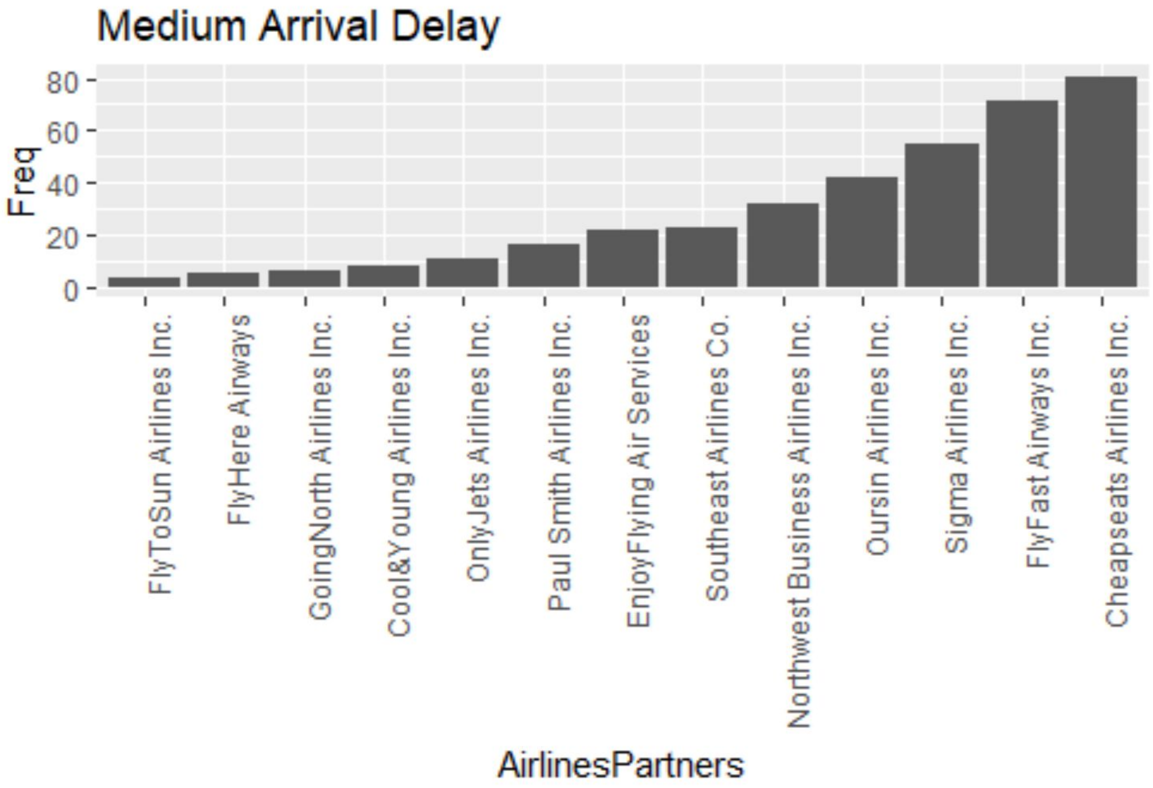
```
arrival_low<- airlines%>%
  filter(airlines$Arrival_group=="low")%>%
  group_by(Partner.Name)
view(arrival_low)
delay_low1<- as.data.frame(table(arrival_low$Partner.Name))
view(delay_low1)
colnames(delay_low1)[which(names(delay_low1) == "Var1")] <- "AirlinesPartners"
ggplot(delay_low1)+
  aes(x=reorder(AirlinesPartners,Freq),y=Freq)+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  labs(x="AirlinesPartners")+
  ggtitle("Short Arrival Delay")+
  geom_col()
```



For short arrival delay, Cheapseats has the shortest delays which is not surprising since it has the most short departure delays. At the same time, Sigma has the second most short delays.

Code:

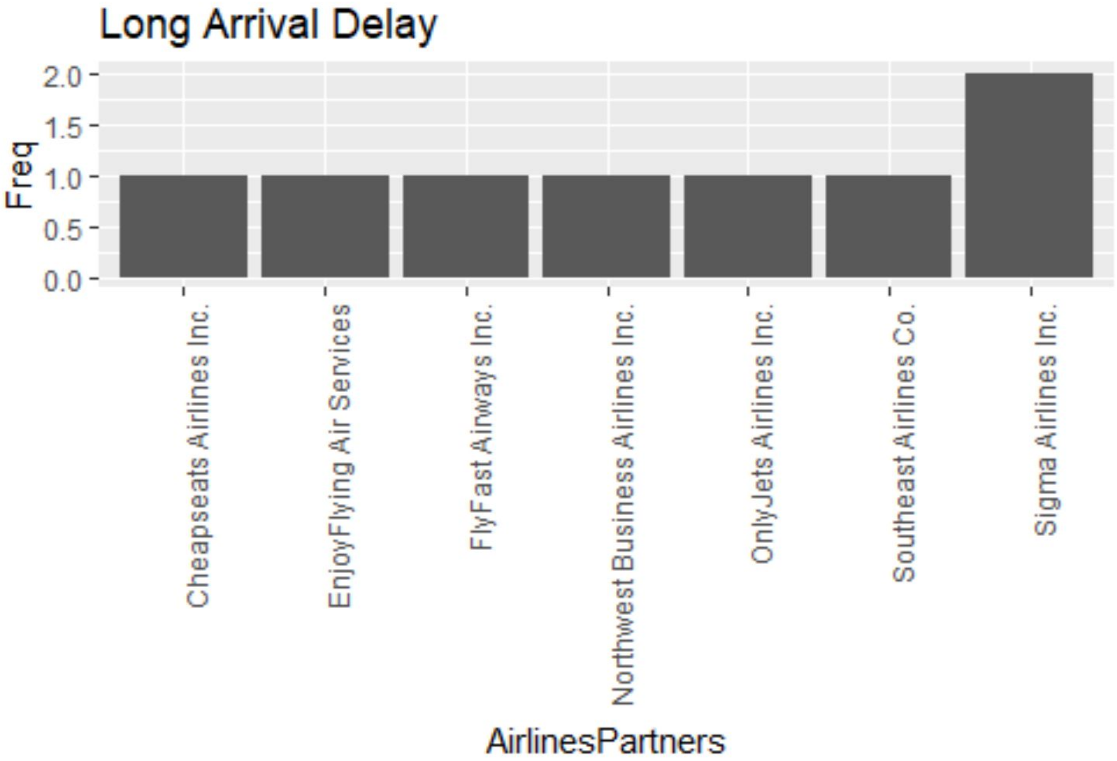
```
arrival_medium<- airlines%>%
  filter(airlines$Arrival_group=="medium")%>%
  group_by(Partner.Name)
view(arrival_medium)
delay_medium1<- as.data.frame(table(arrival_medium$Partner.Name))
view(delay_medium1)
colnames(delay_medium1)[which(names(delay_medium1) == "Var1")] <- "AirlinesPartners"
ggplot(delay_medium1)+
  aes(x=reorder(AirlinesPartners,Freq),y=Freq)+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  ggtitle("Medium Arrival Delay")+
  labs(x="AirlinesPartners")+
  geom_col()
```



In this barchart, we can see that the frequency of medium arrival delay is almost the same as the departure delay.

Code:

```
arrival_high<- airlines%>%
  filter(airlines$Arrival_group=="high")%>%
  group_by(Partner.Name)
view(arrival_high)
delay_high1<- as.data.frame(table(arrival_high$Partner.Name))
view(delay_high1)
colnames(delay_high1)[which(names(delay_high1) == "Var1")] <- "AirlinesPartners"
ggplot(delay_high1)+
  aes(x=reorder(AirlinesPartners,Freq),y=Freq)+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  ggtitle("Long Arrival Delay")+
  labs(x="AirlinesPartners")+
  geom_col()
```



In this chart, although Sigma has two very long arrival delay, it is not frequent compared to the total number of flights. The frequency of long arrival delay of other airlines company remains once.

Generally, most of the flight delay is focused on short time delay. Cheapseats airlines has the most frequent delays and Sigma is the second. The medium departure delay is not frequent, but it still unsatisfying customers, and in this part Cheapseat still has the most frequent medium delay. However, the second is Flyfast and the third is Sigma. The difference is not big which means Sigma is doing slightly better than Flyfast in preventing medium time delay. On the other hand, the frequency of arrival time is almost the same as the frequency of departure time. It means that flights are on time if they depart on time.

5. USE OF MODELING TECHNIQUES

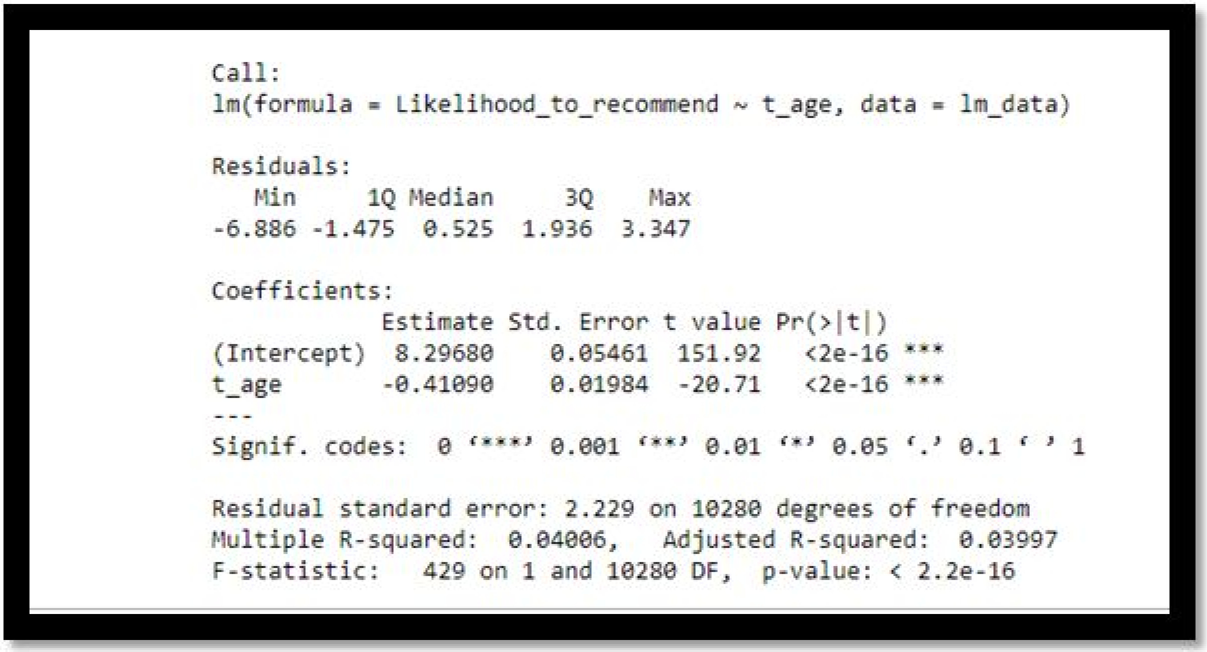
5.1 Linear Modeling

To establish a causal relationship between the different variables and Customer Satisfaction, and to find out how much each variable affects the Likelihood to recommend, we have used bivariate and multivariate linear regression Models to try and predict patterns.

First, we ran bivariate linear regression models on all variables that seemed to have a relation with Customer satisfaction. Using this we also wanted to find out if the variables are significant.

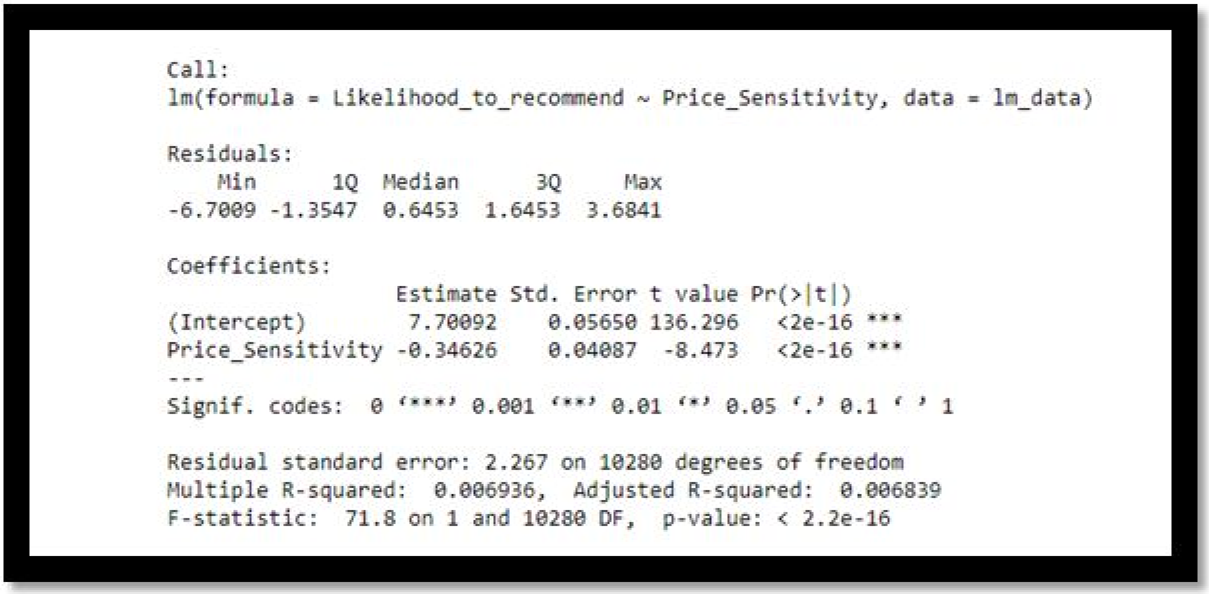
We Got the following results

Regression Model 1 - Age



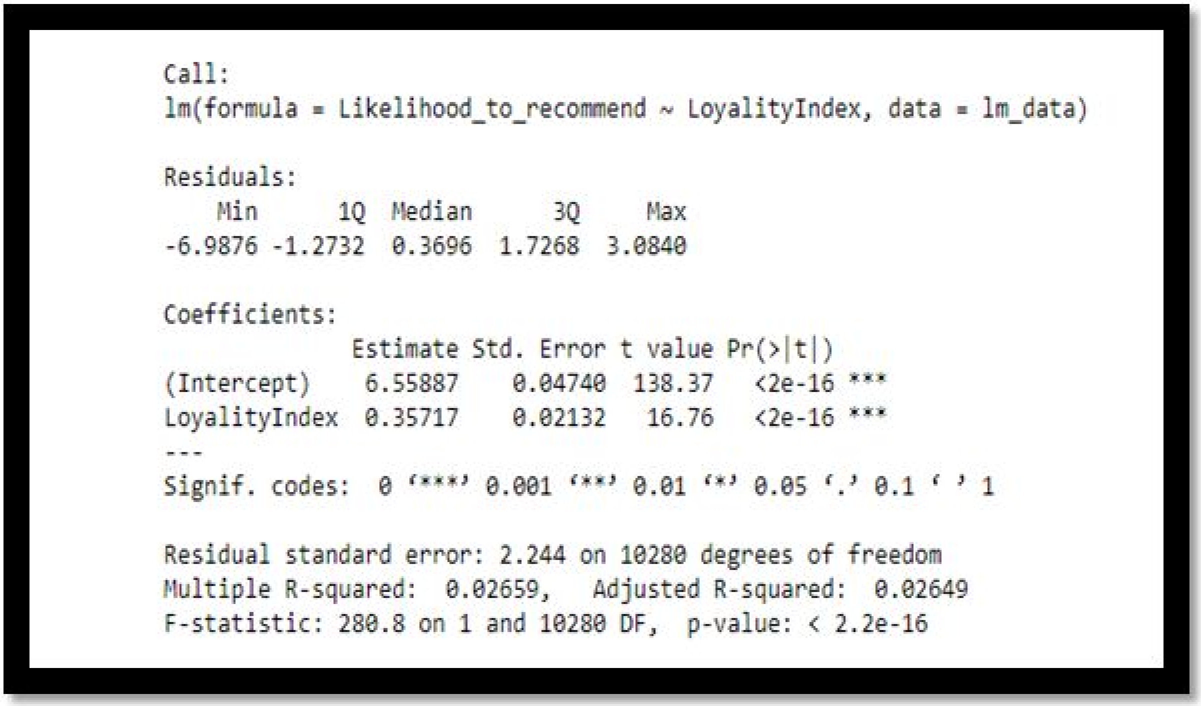
```
Call:
lm(formula = Likelihood_to_recommend ~ t_age, data = lm_data)

Residuals:
   Min     1Q Median     3Q    Max 
-6.886 -1.475  0.525  1.936  3.347 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.29680    0.05461  151.92   <2e-16 ***
t_age       -0.41090    0.01984  -20.71   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.229 on 10280 degrees of freedom
Multiple R-squared:  0.04006,   Adjusted R-squared:  0.03997 
F-statistic:   429 on 1 and 10280 DF,  p-value: < 2.2e-16
```

Regression Model 2 – Price Sensitivity



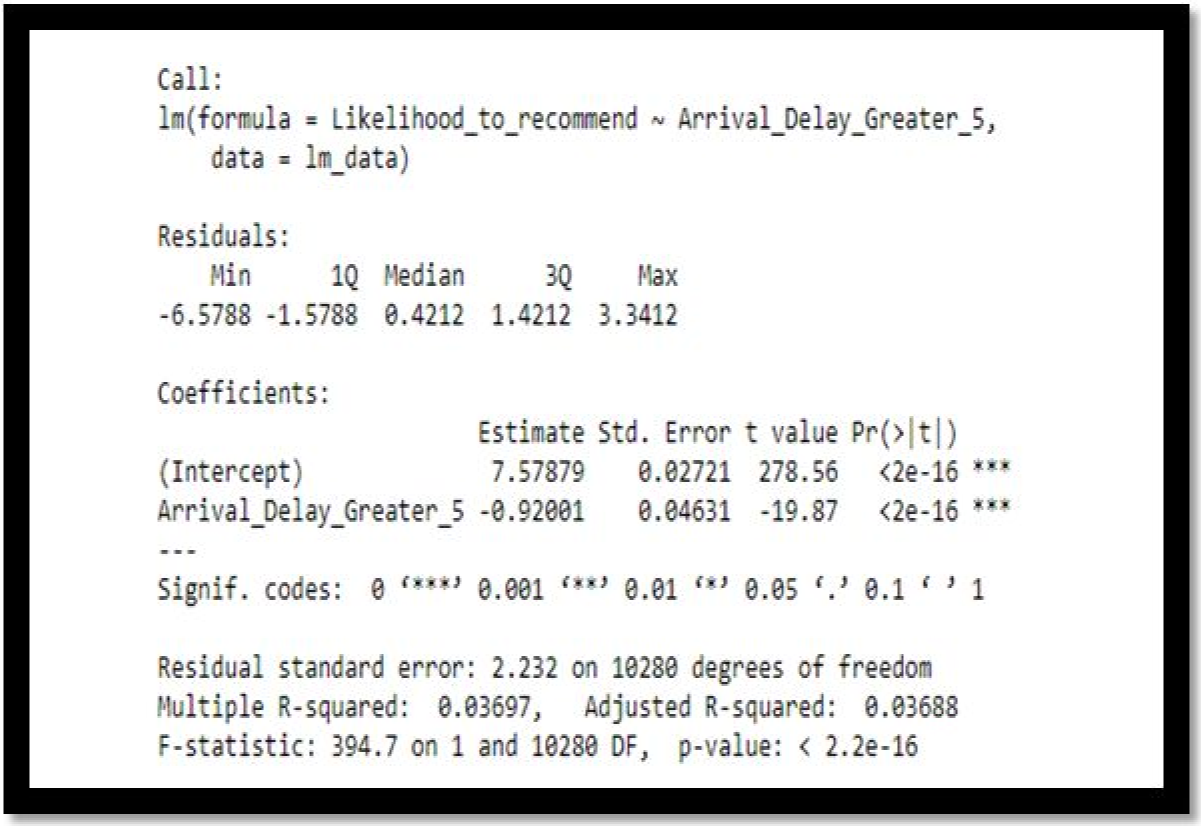
```
Call:
lm(formula = Likelihood_to_recommend ~ Price_Sensitivity, data = lm_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.7009 -1.3547  0.6453  1.6453  3.6841 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        7.70092    0.05650 136.296   <2e-16 ***
Price_Sensitivity -0.34626    0.04087  -8.473   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.267 on 10280 degrees of freedom
Multiple R-squared:  0.006936,	Adjusted R-squared:  0.006839 
F-statistic:  71.8 on 1 and 10280 DF,  p-value: < 2.2e-16
```

Regression Model 3 – Loyalty Index



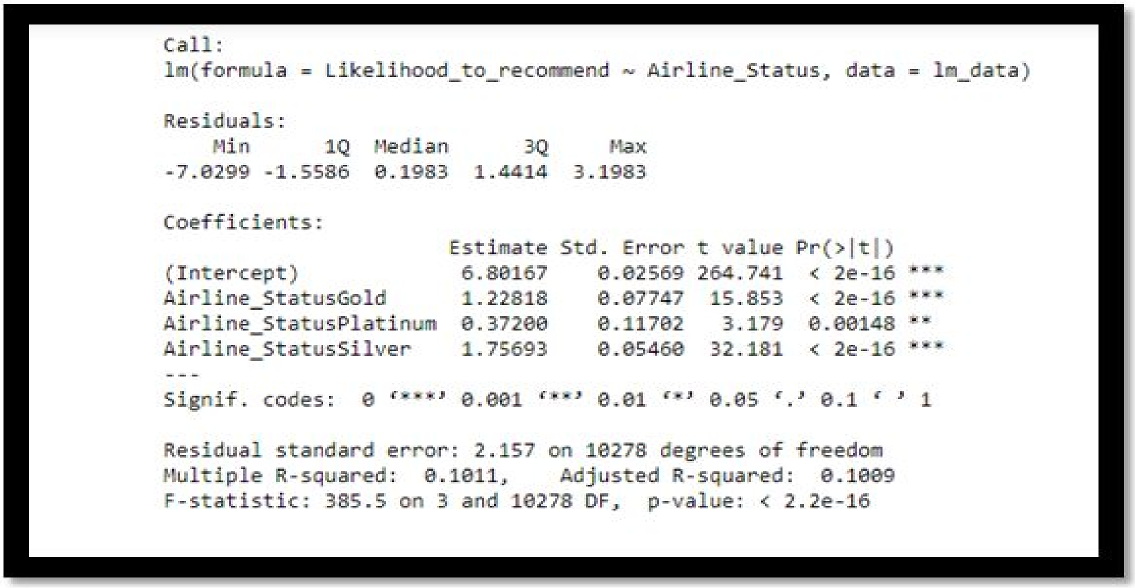
```
Call:
lm(formula = Likelihood_to_recommend ~ LoyalityIndex, data = lm_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.9876 -1.2732  0.3696  1.7268  3.0840 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    6.55887    0.04740  138.37   <2e-16 ***
LoyalityIndex  0.35717    0.02132   16.76   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.244 on 10280 degrees of freedom
Multiple R-squared:  0.02659,	Adjusted R-squared:  0.02649 
F-statistic: 280.8 on 1 and 10280 DF,  p-value: < 2.2e-16
```

Regression Model 4 – Arrival_delay_Greater_5



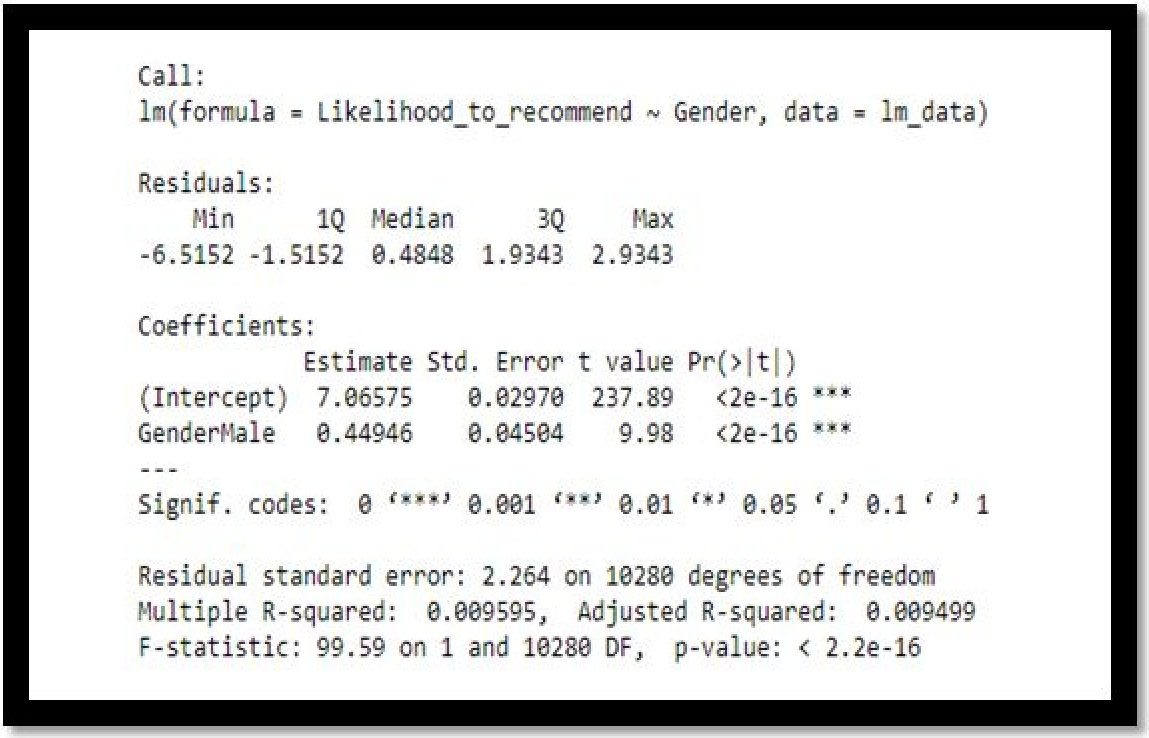
```
Call:
lm(formula = Likelihood_to_recommend ~ Arrival_Delay_Greater_5,
    data = lm_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.5788 -1.5788  0.4212  1.4212  3.3412 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)              7.57879    0.02721  278.56   <2e-16 ***
Arrival_Delay_Greater_5 -0.92001    0.04631  -19.87   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.232 on 10280 degrees of freedom
Multiple R-squared:  0.03697,	Adjusted R-squared:  0.03688 
F-statistic: 394.7 on 1 and 10280 DF,  p-value: < 2.2e-16
```

Regression Model 5 – Airline Status



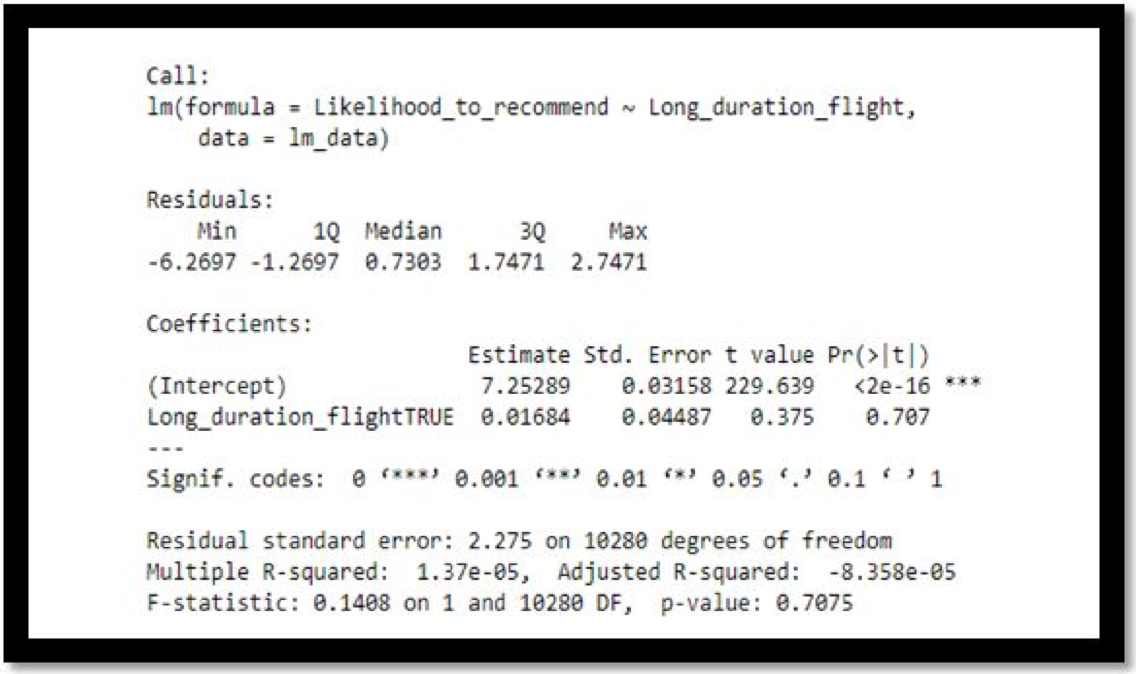
```
Call:
lm(formula = Likelihood_to_recommend ~ Airline_Status, data = lm_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.0299 -1.5586  0.1983  1.4414  3.1983 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)            6.80167    0.02569 264.741  < 2e-16 ***
Airline_StatusGold     1.22818    0.07747  15.853  < 2e-16 ***
Airline_StatusPlatinum 0.37200    0.11702   3.179  0.00148 ** 
Airline_StatusSilver   1.75693    0.05460  32.181  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.157 on 10278 degrees of freedom
Multiple R-squared:  0.1011,	Adjusted R-squared:  0.1009 
F-statistic: 385.5 on 3 and 10278 DF,  p-value: < 2.2e-16
```

Regression Model 6 – Gender



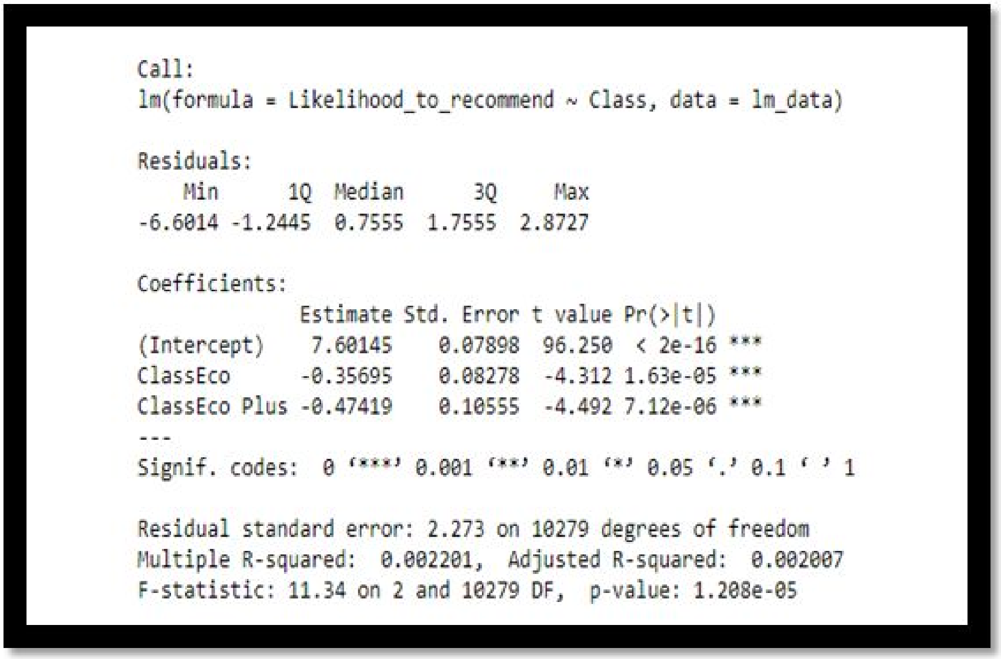
```
Call:
lm(formula = Likelihood_to_recommend ~ Gender, data = lm_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.5152 -1.5152  0.4848  1.9343  2.9343 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.06575    0.02970  237.89   <2e-16 ***
GenderMale   0.44946    0.04504    9.98   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.264 on 10280 degrees of freedom
Multiple R-squared:  0.009595,	Adjusted R-squared:  0.009499 
F-statistic: 99.59 on 1 and 10280 DF,  p-value: < 2.2e-16
```

Regression Model 7 – Long_duration_flight



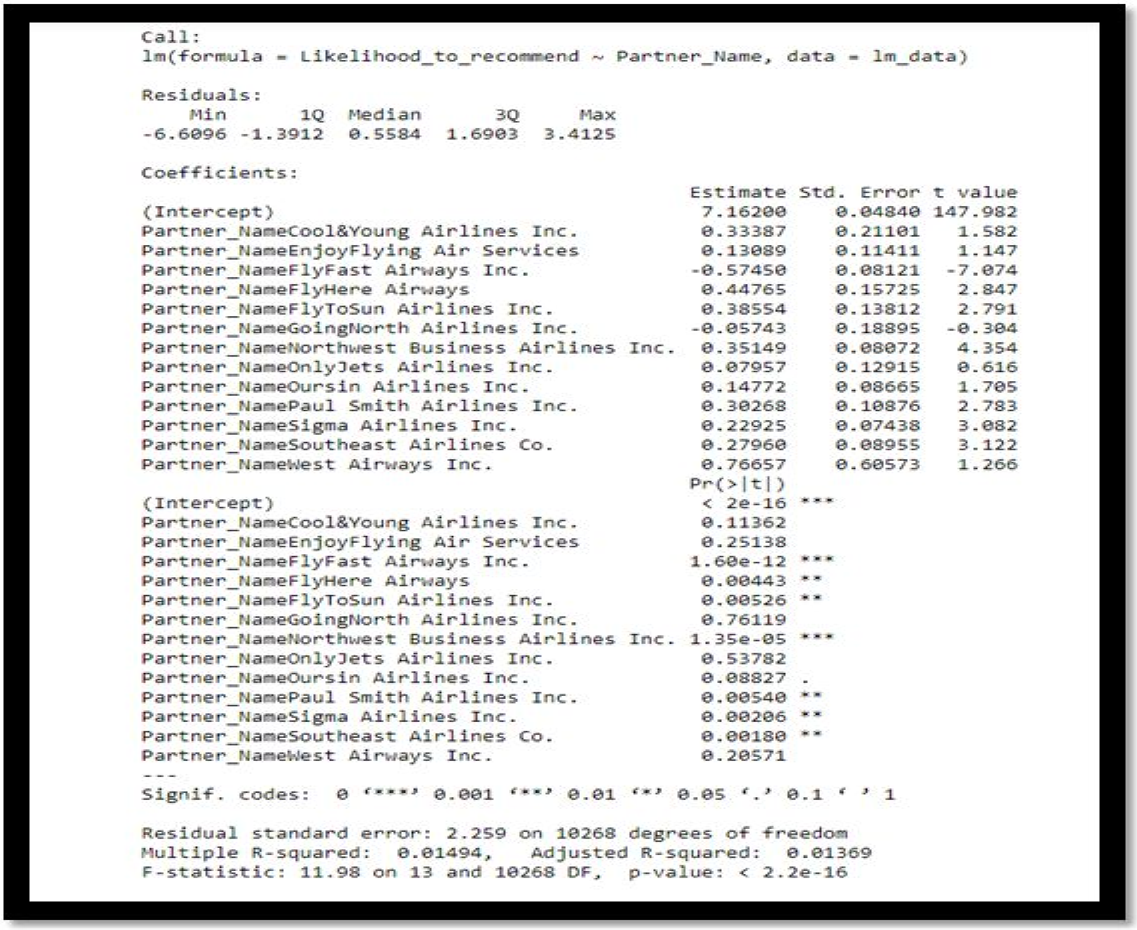
```
Call:
lm(formula = Likelihood_to_recommend ~ Long_duration_flight,
    data = lm_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.2697 -1.2697  0.7303  1.7471  2.7471 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)               7.25289    0.03158 229.639   <2e-16 ***
Long_duration_flightTRUE  0.01684    0.04487   0.375    0.707    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.275 on 10280 degrees of freedom
Multiple R-squared:  1.37e-05,  Adjusted R-squared:  -8.358e-05 
F-statistic: 0.1408 on 1 and 10280 DF,  p-value: 0.7075
```

Regression Model 8 – Class



```
Call:
lm(formula = Likelihood_to_recommend ~ Class, data = lm_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.6014 -1.2445  0.7555  1.7555  2.8727 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    7.60145    0.07898  96.250  < 2e-16 ***
ClassEco      -0.35695    0.08278  -4.312 1.63e-05 ***
ClassEco Plus -0.47419    0.10555  -4.492 7.12e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.273 on 10279 degrees of freedom
Multiple R-squared:  0.002201,  Adjusted R-squared:  0.002007 
F-statistic: 11.34 on 2 and 10279 DF,  p-value: 1.208e-05
```

Regression Model 9 – Partner Airlines

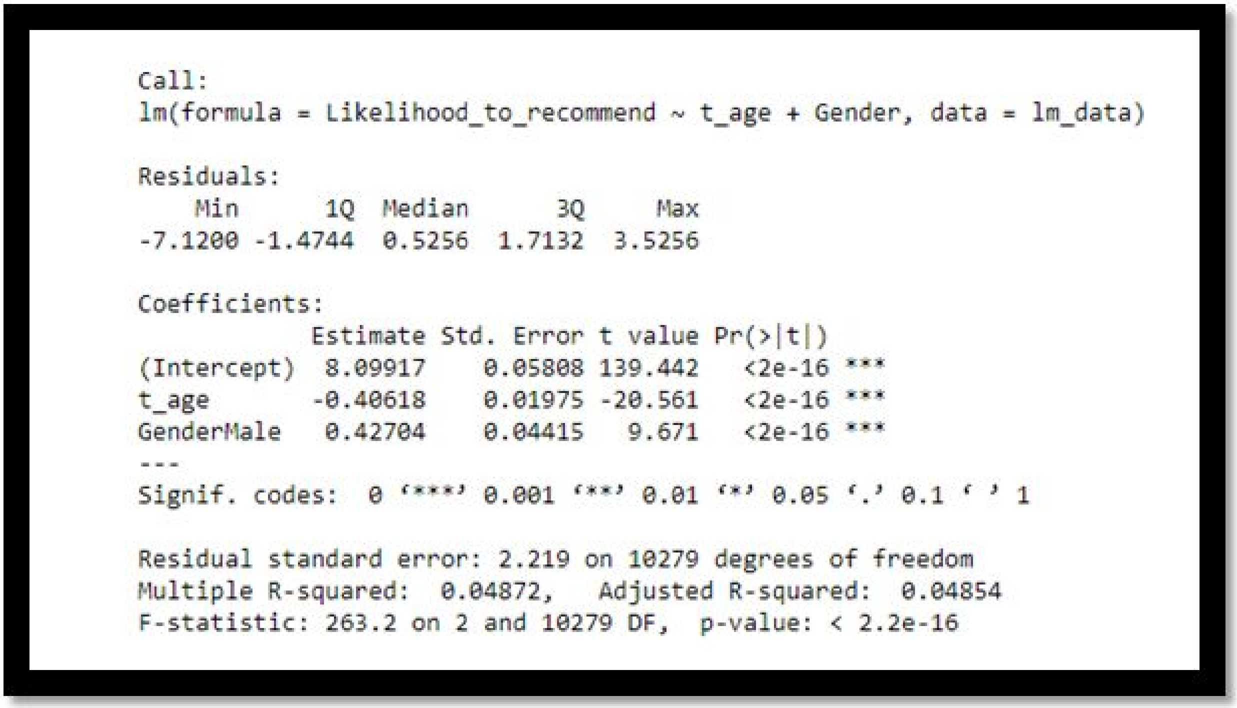


```
Call:
lm(formula = Likelihood_to_recommend ~ Partner_Name, data = lm_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.6096 -1.3912  0.5584  1.6903  3.4125 

Coefficients:
                                          Estimate Std. Error t value
(Intercept)                                7.16200    0.04840 147.982
Partner_NameCool&Young Airlines Inc.       0.33387    0.21101   1.582
Partner_NameEnjoyFlying Air Services       0.13089    0.11411   1.147
Partner_NameFlyFast Airways Inc.          -0.57450    0.08121  -7.074
Partner_NameFlyHere Airways                0.44765    0.15725   2.847
Partner_NameFlyToSun Airlines Inc.         0.38554    0.13812   2.791
Partner_NameGoingNorth Airlines Inc.      -0.05743    0.18895  -0.304
Partner_NameNorthwest Business Airlines Inc.  0.35149    0.08072   4.354
Partner_NameOnlyJets Airlines Inc.         0.07957    0.12915   0.616
Partner_NameOursin Airlines Inc.           0.14772    0.08665   1.705
Partner_NamePaul Smith Airlines Inc.       0.30268    0.10876   2.783
Partner_NameSigma Airlines Inc.            0.22925    0.07438   3.082
Partner_NameSoutheast Airlines Co.         0.27960    0.08955   3.122
Partner_NameWest Airways Inc.              0.76657    0.60573   1.266
                                          Pr(>|t|)    
(Intercept)                                < 2e-16 ***
Partner_NameCool&Young Airlines Inc.       0.11362    
Partner_NameEnjoyFlying Air Services       0.25138    
Partner_NameFlyFast Airways Inc.          1.60e-12 ***
Partner_NameFlyHere Airways                0.00443 ** 
Partner_NameFlyToSun Airlines Inc.         0.00526 ** 
Partner_NameGoingNorth Airlines Inc.       0.76119    
Partner_NameNorthwest Business Airlines Inc. 1.35e-05 ***
Partner_NameOnlyJets Airlines Inc.         0.53782    
Partner_NameOursin Airlines Inc.           0.08827 .  
Partner_NamePaul Smith Airlines Inc.       0.00540 ** 
Partner_NameSigma Airlines Inc.            0.00206 ** 
Partner_NameSoutheast Airlines Co.         0.00180 ** 
Partner_NameWest Airways Inc.              0.20571    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.259 on 10268 degrees of freedom
Multiple R-squared:  0.01494,	Adjusted R-squared:  0.01369 
F-statistic: 11.98 on 13 and 10268 DF,  p-value: < 2.2e-16
```

From the results we found that Age, Price Sensitivity, Loyalty Index, Gender, Class, Airline Status we more significant than the other variables, we started out by building our model using a Step-wise Approach.

Continuing on from the regression result for the ‘Age’ variable, we next added Gender to check if this improved our R².

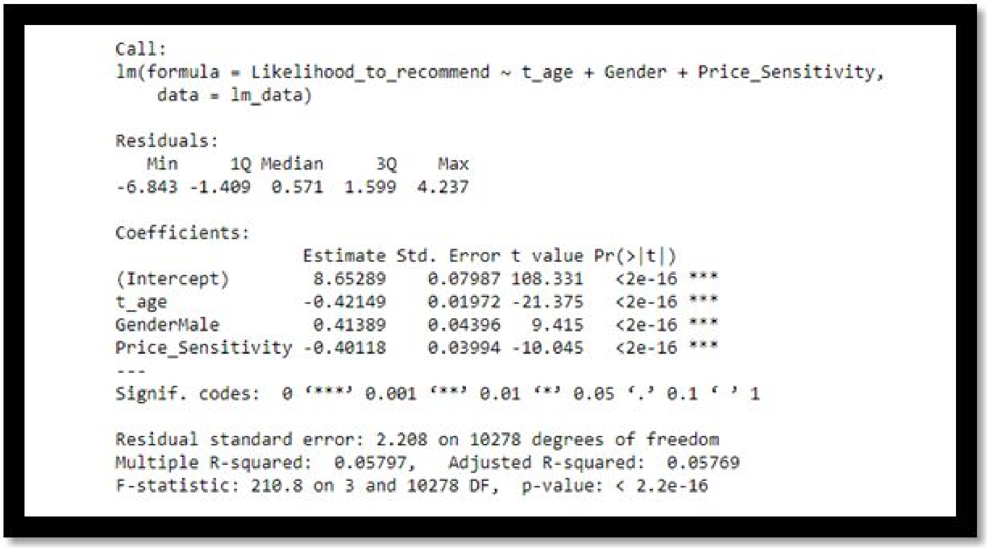
Regression Model – Age + Gender



```
Call:
lm(formula = Likelihood_to_recommend ~ t_age + Gender, data = lm_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.1200 -1.4744  0.5256  1.7132  3.5256 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.09917    0.05808 139.442   <2e-16 ***
t_age       -0.40618    0.01975 -20.561   <2e-16 ***
GenderMale   0.42704    0.04415   9.671   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.219 on 10279 degrees of freedom
Multiple R-squared:  0.04872,	Adjusted R-squared:  0.04854 
F-statistic: 263.2 on 2 and 10279 DF,  p-value: < 2.2e-16
```

Here the Adjusted R² has increased from ‘0.03997’ to ‘0.04854’. Thereby we include Gender in the Linear Model. Also, from the results we can see that a Male is inclined to be more satisfied than a female.

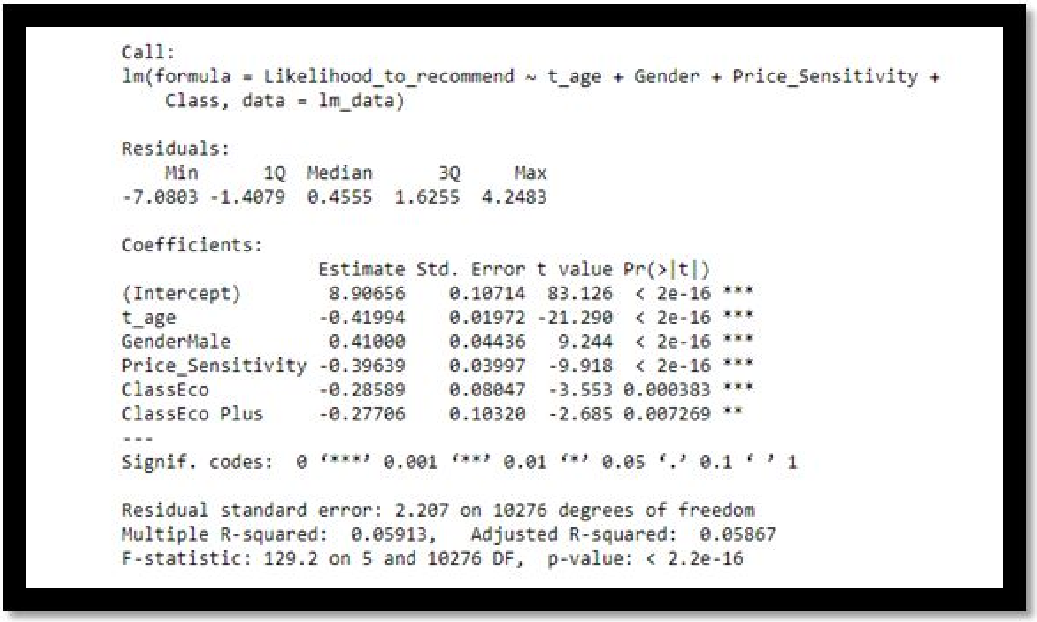
Regression Model – Age + Gender + Price Sensitivity



```
Call:
lm(formula = Likelihood_to_recommend ~ t_age + Gender + Price_Sensitivity,
    data = lm_data)

Residuals:
   Min     1Q Median     3Q    Max 
-6.843 -1.409  0.571  1.599  4.237 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        8.65289    0.07987 108.331   <2e-16 ***
t_age             -0.42149    0.01972 -21.375   <2e-16 ***
GenderMale         0.41389    0.04396   9.415   <2e-16 ***
Price_Sensitivity -0.40118    0.03994 -10.045   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.208 on 10278 degrees of freedom
Multiple R-squared:  0.05797,	Adjusted R-squared:  0.05769 
F-statistic: 210.8 on 3 and 10278 DF,  p-value: < 2.2e-16
```

Here the Adjusted R² has increased from ‘0.04854’ to ‘0.05769’. Thereby we include Price Sensitivity in the Linear Model. Also, from the results we can see that lower the price sensitivity higher customer satisfaction.

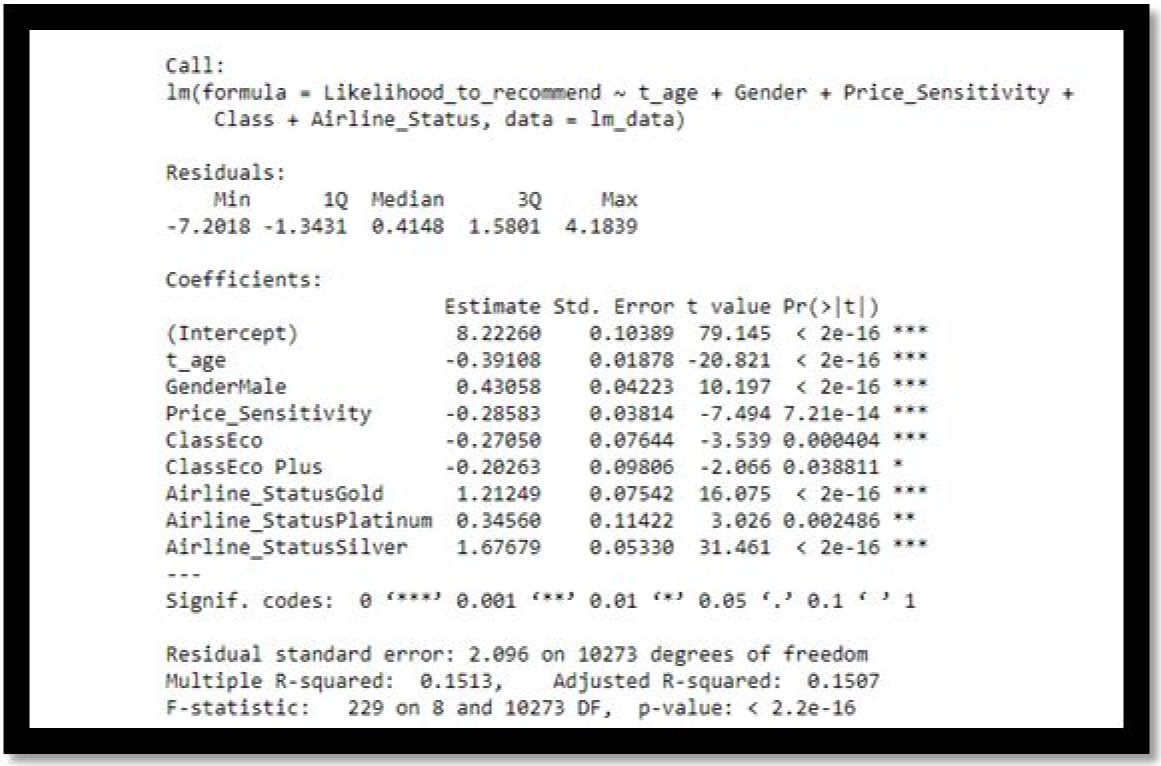
<u>Regression Model – Age + Gender + Price Sensitivity + Class</u>



```
Call:
lm(formula = Likelihood_to_recommend ~ t_age + Gender + Price_Sensitivity + 
    Class, data = lm_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.0803 -1.4079  0.4555  1.6255  4.2483 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        8.90656    0.10714  83.126  < 2e-16 ***
t_age             -0.41994    0.01972 -21.290  < 2e-16 ***
GenderMale         0.41000    0.04436   9.244  < 2e-16 ***
Price_Sensitivity -0.39639    0.03997  -9.918  < 2e-16 ***
ClassEco          -0.28589    0.08047  -3.553 0.000383 ***
ClassEco Plus     -0.27706    0.10320  -2.685 0.007269 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.207 on 10276 degrees of freedom
Multiple R-squared:  0.05913,	Adjusted R-squared:  0.05867 
F-statistic: 129.2 on 5 and 10276 DF,  p-value: < 2.2e-16
```

Here the Adjusted R² has increased from ‘0.05769’ to ‘0.05867’. Thereby we include Class in the Linear Model. Also, from the results we can see that Business Class has a higher satisfaction than Eco or Eco Plus.

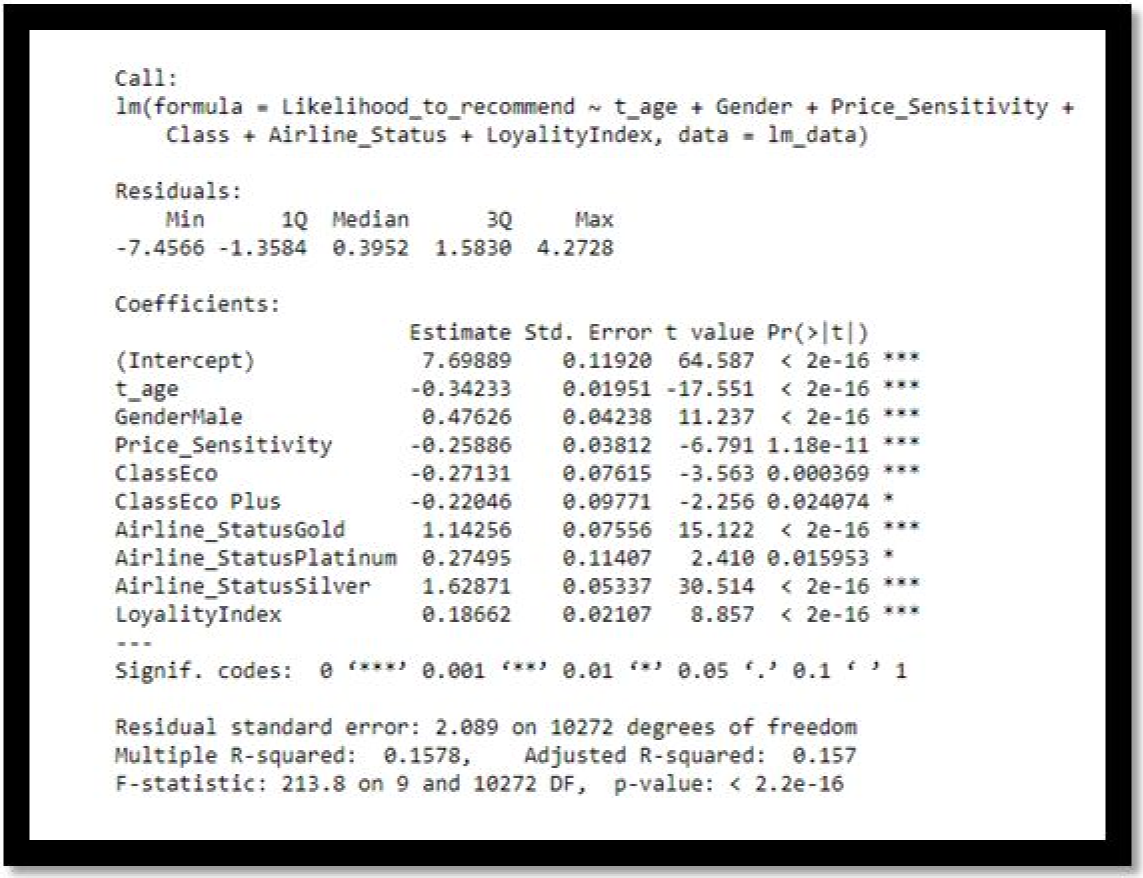
Regression Model – Age + Gender + Price Sensitivity + Class + Airline Status



```
Call:
lm(formula = Likelihood_to_recommend ~ t_age + Gender + Price_Sensitivity +
    Class + Airline_Status, data = lm_data)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2018 -1.3431  0.4148  1.5801  4.1839

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              8.22260    0.10389  79.145  < 2e-16 ***
t_age                   -0.39108    0.01878 -20.821  < 2e-16 ***
GenderMale               0.43058    0.04223  10.197  < 2e-16 ***
Price_Sensitivity       -0.28583    0.03814  -7.494 7.21e-14 ***
ClassEco                -0.27050    0.07644  -3.539 0.000404 ***
ClassEco Plus           -0.20263    0.09806  -2.066 0.038811 *
Airline_StatusGold       1.21249    0.07542  16.075  < 2e-16 ***
Airline_StatusPlatinum   0.34560    0.11422   3.026 0.002486 **
Airline_StatusSilver     1.67679    0.05330  31.461  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.096 on 10273 degrees of freedom
Multiple R-squared:  0.1513,    Adjusted R-squared:  0.1507
F-statistic:   229 on 8 and 10273 DF,  p-value: < 2.2e-16
```

Here the Adjusted R² has increased from ‘0.05867’ to ‘0.1507’. Thereby we include Airline Status in the Linear Model. Also, from the results we can see that Gold and Silver Status have a higher satisfaction than Platinum and Blue.

Regression Model – Age + Gender + Price Sensitivity + Class + Airline Status + Loyalty Index



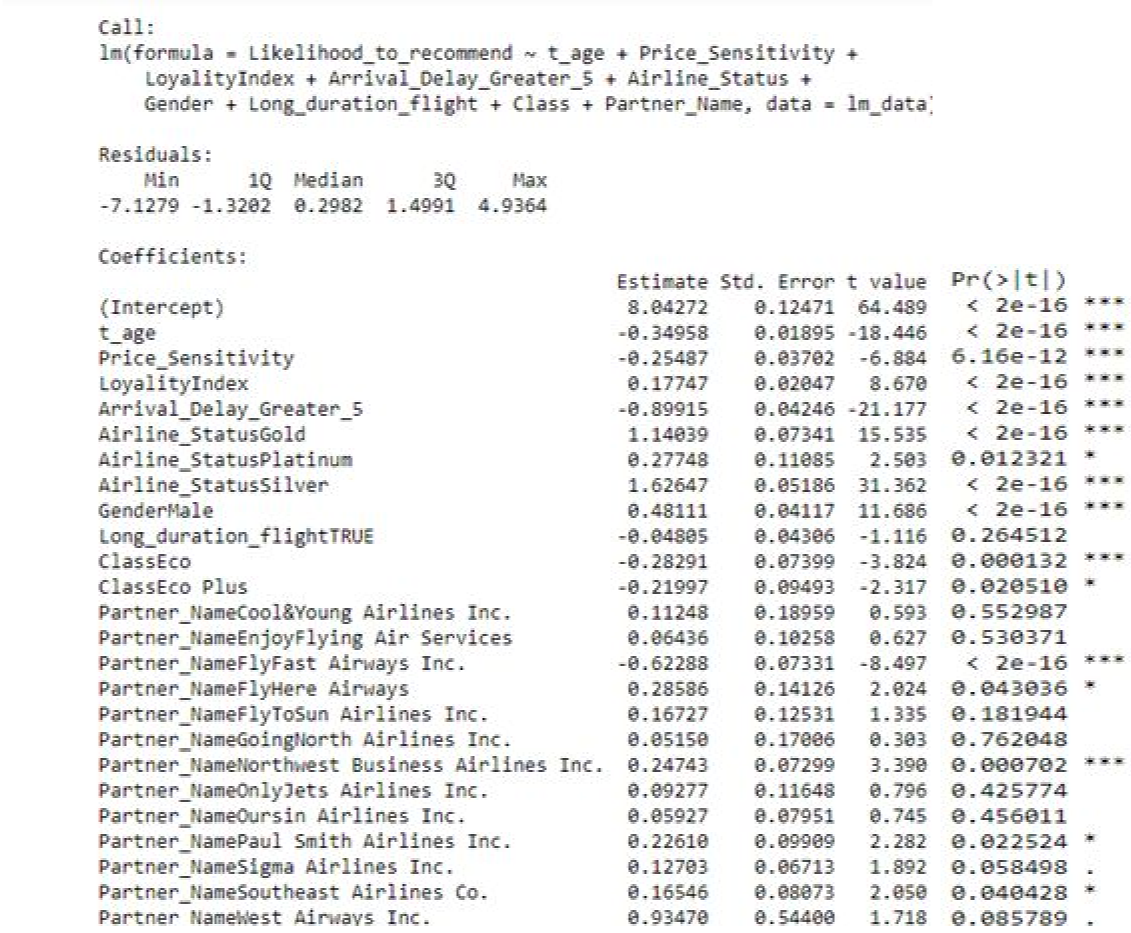
```
Call:
lm(formula = Likelihood_to_recommend ~ t_age + Gender + Price_Sensitivity +
    Class + Airline_Status + LoyalityIndex, data = lm_data)

Residuals:
    Min      1Q  Median      3Q     Max
-7.4566 -1.3584  0.3952  1.5830  4.2728

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             7.69889    0.11920  64.587  < 2e-16 ***
t_age                  -0.34233    0.01951 -17.551  < 2e-16 ***
GenderMale              0.47626    0.04238  11.237  < 2e-16 ***
Price_Sensitivity      -0.25886    0.03812  -6.791 1.18e-11 ***
ClassEco               -0.27131    0.07615  -3.563 0.000369 ***
ClassEco Plus          -0.22046    0.09771  -2.256 0.024074 *
Airline_StatusGold      1.14256    0.07556  15.122  < 2e-16 ***
Airline_StatusPlatinum  0.27495    0.11407   2.410 0.015953 *
Airline_StatusSilver    1.62871    0.05337  30.514  < 2e-16 ***
LoyalityIndex           0.18662    0.02107   8.857  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.089 on 10272 degrees of freedom
Multiple R-squared:  0.1578,    Adjusted R-squared:  0.157
F-statistic: 213.8 on 9 and 10272 DF,  p-value: < 2.2e-16
```

Here the Adjusted R² has increased from ‘0.05867’ to ‘0.1570’. Thereby we include Loyalty Index in the Linear Model. Also, from the results we can see that the higher Loyalty, more satisfaction.

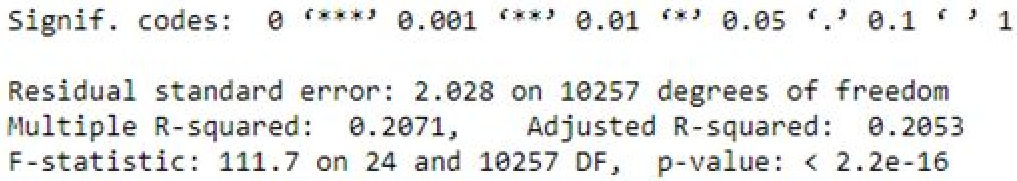
We also tried adding the Long Duration Flight and Partner Airlines and found that it improved our Adjusted R² from ‘0.1570’ to ‘0.2053’. Thereby we decided to add the 2 variables to our Linear Model as well.

We found that customers in short flights were predicted to be more satisfied then customers in long flights. Also, we have found that ‘West Airways’ customers are likely to be satisfied more than other partner airlines customers.

This is our final regression model as well. We have been able to predict 20% of the change in the Likelihood to recommend by the variables shown above.



```
Call:
lm(formula = Likelihood_to_recommend ~ t_age + Price_Sensitivity +
    LoyalityIndex + Arrival_Delay_Greater_5 + Airline_Status +
    Gender + Long_duration_flight + Class + Partner_Name, data = lm_data)

Residuals:
    Min      1Q  Median      3Q     Max
-7.1279 -1.3202  0.2982  1.4991  4.9364

Coefficients:
                                           Estimate Std. Error t value  Pr(>|t|)
(Intercept)                                 8.04272    0.12471  64.489  < 2e-16 ***
t_age                                      -0.34958    0.01895 -18.446  < 2e-16 ***
Price_Sensitivity                          -0.25487    0.03702  -6.884 6.16e-12 ***
LoyalityIndex                               0.17747    0.02047   8.670  < 2e-16 ***
Arrival_Delay_Greater_5                    -0.89915    0.04246 -21.177  < 2e-16 ***
Airline_StatusGold                          1.14039    0.07341  15.535  < 2e-16 ***
Airline_StatusPlatinum                      0.27748    0.11085   2.503 0.012321 *
Airline_StatusSilver                        1.62647    0.05186  31.362  < 2e-16 ***
GenderMale                                  0.48111    0.04117  11.686  < 2e-16 ***
Long_duration_flightTRUE                   -0.04805    0.04306  -1.116 0.264512
ClassEco                                   -0.28291    0.07399  -3.824 0.000132 ***
ClassEco Plus                              -0.21997    0.09493  -2.317 0.020510 *
Partner_NameCool&Young Airlines Inc.        0.11248    0.18959   0.593 0.552987
Partner_NameEnjoyFlying Air Services        0.06436    0.10258   0.627 0.530371
Partner_NameFlyFast Airways Inc.           -0.62288    0.07331  -8.497  < 2e-16 ***
Partner_NameFlyHere Airways                 0.28586    0.14126   2.024 0.043036 *
Partner_NameFlyToSun Airlines Inc.          0.16727    0.12531   1.335 0.181944
Partner_NameGoingNorth Airlines Inc.        0.05150    0.17006   0.303 0.762048
Partner_NameNorthwest Business Airlines Inc. 0.24743    0.07299   3.390 0.000702 ***
Partner_NameOnlyJets Airlines Inc.          0.09277    0.11648   0.796 0.425774
Partner_NameOursin Airlines Inc.            0.05927    0.07951   0.745 0.456011
Partner_NamePaul Smith Airlines Inc.        0.22610    0.09909   2.282 0.022524 *
Partner_NameSigma Airlines Inc.             0.12703    0.06713   1.892 0.058498 .
Partner_NameSoutheast Airlines Co.          0.16546    0.08073   2.050 0.040428 *
Partner NameWest Airways Inc.               0.93470    0.54400   1.718 0.085789 .
```



```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.028 on 10257 degrees of freedom
Multiple R-squared:  0.2071,    Adjusted R-squared:  0.2053
F-statistic: 111.7 on 24 and 10257 DF,  p-value: < 2.2e-16
```

The analysis of the regression tells us the following

- With increasing age, satisfaction decreases
- When price sensitivity is high, satisfaction decrease, which makes sense intuitively
- Satisfaction of our loyal customers are comparatively higher, which is good
- If Arrival Delay is greater than 5 minutes, then satisfaction is expected to fall.
- A Business class customer is a more satisfied than an Eco or Eco Plus customer, which is what we need
- A male on average is more satisfied than a female.
- An Airline Gold and Silver Status customer is a more satisfied than a Platinum Status customer> This is not good and requires more attention towards it.
- Customers Travelling through West Airways have comparatively higher satisfaction than other Partner airlines flight. Also, there are some airlines which are not statistically significant to the regression equation.

Therefore, a Loyal Male Business Class Gold Member has a higher chance of rating a higher Likelihood to Recommend.

CODE:

```
# Regression Model 1 - Age

RM1_age <- lm(formula = Likelihood_to_recommend ~ t_age,data = lm_data)

summary(RM1_age)



# Regression Model 2 - Price Sensitivity

RM2_PS <- lm(formula = Likelihood_to_recommend ~ Price_Sensitivity,data = lm_data)

summary(RM2_PS)



# Regression Model 3 - Loyality Index
```

```
RM3_Loyal <- lm(formula = Likelihood_to_recommend ~ LoyalityIndex,data = lm_data)

summary(RM3_Loyal)


# Regression Model 4 - Arrival Delay Greater than 5

RM4_ArrivalDelay <- lm(formula = Likelihood_to_recommend ~ Arrival_Delay_Greater_5,data = lm_data)

summary(RM4_ArrivalDelay)


# Regression Model 5 - Airline Status

RM5_status <- lm(formula = Likelihood_to_recommend ~ Airline_Status,data = lm_data)

summary(RM5_status)


# Regression Model 6 - Gender

RM6_Gender <- lm(formula = Likelihood_to_recommend ~ Gender,data = lm_data)

summary(RM6_Gender)


# Regression Model 7 - Long Flight Duration

RM7_flightduration <- lm(formula = Likelihood_to_recommend ~ Long_duration_flight,data = lm_data)

summary(RM7_flightduration)
```

```
# Regression Model 8 - class

RM8_class <- lm(formula = Likelihood_to_recommend ~ Class,data = lm_data)

summary(RM8_class)


# Regression Model 9 - Partnered Airline

RM9_Partner_Airline <- lm(formula = Likelihood_to_recommend ~ Partner_Name,data =
lm_data)

summary(RM9_Partner_Airline)

# Regression Model A - Age + Gender

RM_A <- lm(formula=Likelihood_to_recommend ~ t_age + Gender,data = lm_data)

summary(RM_A)


# Regression Model B - Age + Gender + Price Sensitivity

RM_B <- lm(formula=Likelihood_to_recommend ~ t_age + Gender + Price_Sensitivity,data =
lm_data)

summary(RM_B)


# Regression Model C - Age + Gender + Price Sensitivity + Class

RM_C <- lm(formula=Likelihood_to_recommend ~ t_age + Gender + Price_Sensitivity +
Class,data = lm_data)

summary(RM_C)
```

```
# Regression Model D - Age + Gender + Price Sensitivity + Class + Airline Status

RM_D <- lm(formula=Likelihood_to_recommend ~ t_age + Gender + Price_Sensitivity + Class +
Airline_Status,data = lm_data)

summary(RM_D)


# Regression Model E - Age + Gender + Price Sensitivity + Class + Airline Status +
Loyality Index

RM_E <- lm(formula=Likelihood_to_recommend ~ t_age + Gender + Price_Sensitivity + Class +
Airline_Status + LoyalityIndex,data = lm_data)

summary(RM_E)


# OVR 1

RM_ovr1 <- lm(formula = Likelihood_to_recommend ~ t_age + Price_Sensitivity + LoyalityIndex
+ Arrival_Delay_Greater_5 + Airline_Status + Gender + Long_duration_flight + Class +
Partner_Name,data  = lm_data)

summary(RM_ovr1)
```

5.2 Association Rule Modeling

We used Association rules model to answer the business questions:
Q1. Who are happy and unhappy customers?
Q2. What makes a successful flight experience?
Q3. How to make customers happy?

The reason for choosing Association Rules since this modeling technique is one of the best one for analyzing and predicting customer behaviour. They play an important role the customer analytics,product catalog,market based analytics and store layout. In this method each set of elements are taken into consideration i.e. finding associations among items often among the large databases. An association rule has two parts: an antecedent (if) and a consequent (then). An antecedent is an item found within the data. A consequent is an item found in combination with the antecedent.

A number of different measures are calculated which helps us in relating and finding insights in relation to one another.Support is an indication of how frequently the items appear in the data. Confidence indicates the number of times the if-then statements are found true. A third metric, called lift, can be used to compare confidence with expected confidence.In this method we generate rules with the help of Apriori algorithm which is a classical algorithm in data mining. It is used for mining frequent itemsets and relevant association rules. It is devised to operate on a database containing a lot of transactions.
Also we can make use of these measures and based on all these measures we will be able to filter large amounts of rules generated to a smaller number for better understanding and developing actionable insights.

First we used all the categorical variables to generate the model around them.Following attributes were used: Destination City, Origin City, Airline Status, Gender, Type of Travel, Class, Partner Name, Origin State, Destination State, Flight cancelled, Age Category, Passenger Category.

Below is the code along with the plots which helped us in understanding the data and building our analysis for improving the customer experience quality with Southeast airlines.

Code:

#Association Rule on Entire set

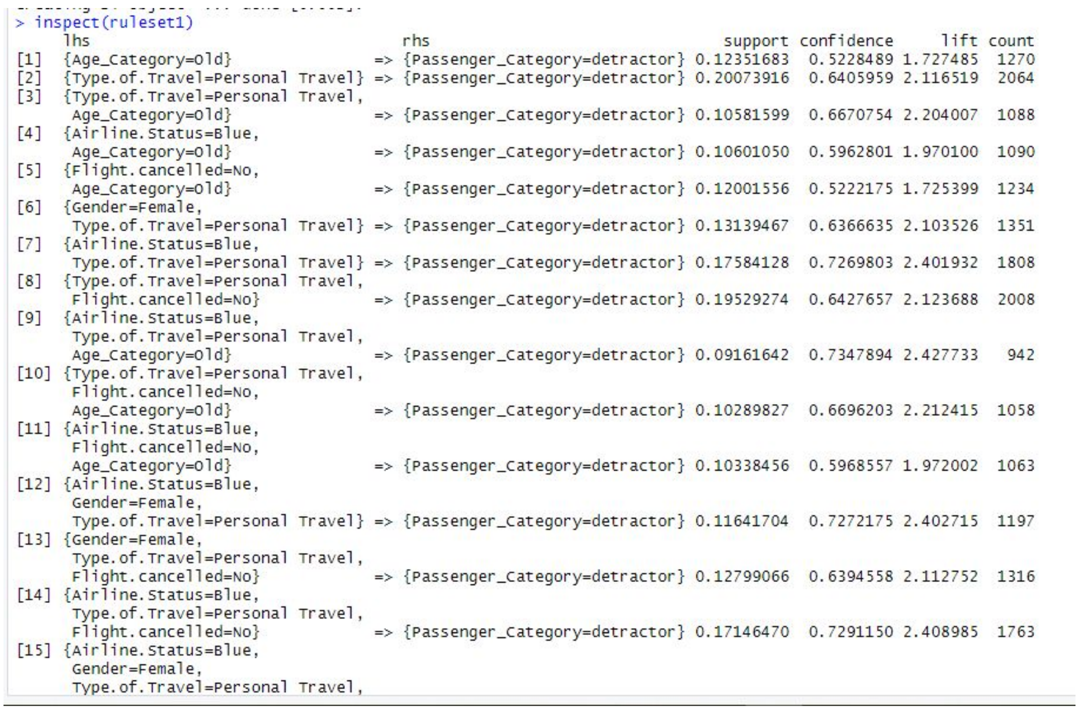
```
#Installing library
install.packages(arules)
library(arules)

install.packages("arulesViz")
library(arulesViz)


jsonfile_data <- jsonfile_data %>%
  mutate(Arrival_group = cut(Arrival_Delay_in_Minutes,
                        breaks = c(-1,100,400,730),
                        labels = c("Low","Medium","High")))

jsonfile_data <- jsonfile_data %>%
  mutate(Delay_group = cut(Departure_Delay_in_Minutes,
                       breaks = c(-1,100,400,751),
                       labels = c("Low","Medium","High")))



discrete_dataset<- jsonfile_data[c('Destination_City', 'Origin_City', 'Airline_Status', 'Gender',
'Type_of_Travel', 'Class', 'Partner_Name', 'Origin_State', 'Destination_State', 'Flight_cancelled',
'Age_Category', 'Passenger_Category')]

str(discrete_dataset)

discrete_dataset$Destination_City= as.factor(discrete_dataset$Destination_City)
discrete_dataset$Origin_City = as.factor(discrete_dataset$Origin_City)
discrete_dataset$Airline_Status=as.factor(discrete_dataset$Airline_Status)
discrete_dataset$Gender=as.factor(discrete_dataset$Gender)
discrete_dataset$Type_of_Travel=as.factor(discrete_dataset$Type_of_Travel)
discrete_dataset$Class=as.vector.factor(discrete_dataset$Class)
discrete_dataset$Partner_Name=as.factor(discrete_dataset$Partner_Name)
discrete_dataset$Origin_State=as.factor(discrete_dataset$Origin_State)
discrete_dataset$Destination_State=as.factor(discrete_dataset$Destination_State)
discrete_dataset$Flight_cancelled=as.factor(discrete_dataset$Flight_cancelled)


DATA_X <- as(discrete_dataset,"transactions")
DATA_X #Transaction Matrix

inspect(DATA_X) #Obtained 10282 rules
```

```
summary(DATA_X)
```

```
#Detractor code for entire dataset
```

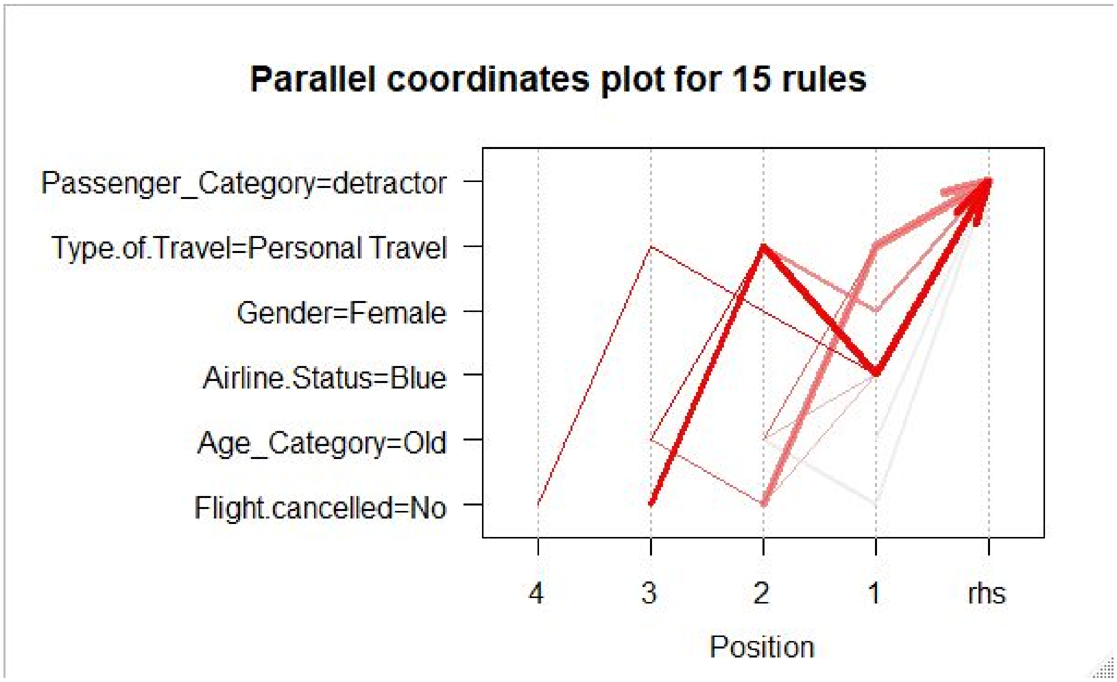
```
ruleset1 <- apriori(DATA_X,
                    parameter=list(support=0.09,confidence=0.5),
                    appearance = list(default="lhs", rhs=("Passenger_Category=detractor")))
```

```
inspect(ruleset1)
```



```
> inspect(ruleset1)
     lhs                                   rhs                            support confidence     lift count
[1]  {Age_Category=Old}                 => {Passenger_Category=detractor} 0.12351683  0.5228489 1.727485  1270
[2]  {Type.of.Travel=Personal Travel}   => {Passenger_Category=detractor} 0.20073916  0.6405959 2.116519  2064
[3]  {Type.of.Travel=Personal Travel,
      Age_Category=Old}                 => {Passenger_Category=detractor} 0.10581599  0.6670754 2.204007  1088
[4]  {Airline.Status=Blue,
      Age_Category=Old}                 => {Passenger_Category=detractor} 0.10601050  0.5962801 1.970100  1090
[5]  {Flight.cancelled=No,
      Age_Category=Old}                 => {Passenger_Category=detractor} 0.12001556  0.5222175 1.725399  1234
[6]  {Gender=Female,
      Type.of.Travel=Personal Travel}   => {Passenger_Category=detractor} 0.13139467  0.6366635 2.103526  1351
[7]  {Airline.Status=Blue,
      Type.of.Travel=Personal Travel}   => {Passenger_Category=detractor} 0.17584128  0.7269803 2.401932  1808
[8]  {Type.of.Travel=Personal Travel,
      Flight.cancelled=No}              => {Passenger_Category=detractor} 0.19529274  0.6427657 2.123688  2008
[9]  {Airline.Status=Blue,
      Type.of.Travel=Personal Travel,
      Age_Category=Old}                 => {Passenger_Category=detractor} 0.09161642  0.7347894 2.427733   942
[10] {Type.of.Travel=Personal Travel,
      Flight.cancelled=No,
      Age_Category=Old}                 => {Passenger_Category=detractor} 0.10289827  0.6696203 2.212415  1058
[11] {Airline.Status=Blue,
      Flight.cancelled=No,
      Age_Category=Old}                 => {Passenger_Category=detractor} 0.10338456  0.5968557 1.972002  1063
[12] {Airline.Status=Blue,
      Gender=Female,
      Type.of.Travel=Personal Travel}   => {Passenger_Category=detractor} 0.11641704  0.7272175 2.402715  1197
[13] {Gender=Female,
      Type.of.Travel=Personal Travel,
      Flight.cancelled=No}              => {Passenger_Category=detractor} 0.12799066  0.6394558 2.112752  1316
[14] {Airline.Status=Blue,
      Type.of.Travel=Personal Travel,
      Flight.cancelled=No}              => {Passenger_Category=detractor} 0.17146470  0.7291150 2.408985  1763
[15] {Airline.Status=Blue,
      Gender=Female,
      Type.of.Travel=Personal Travel,
```

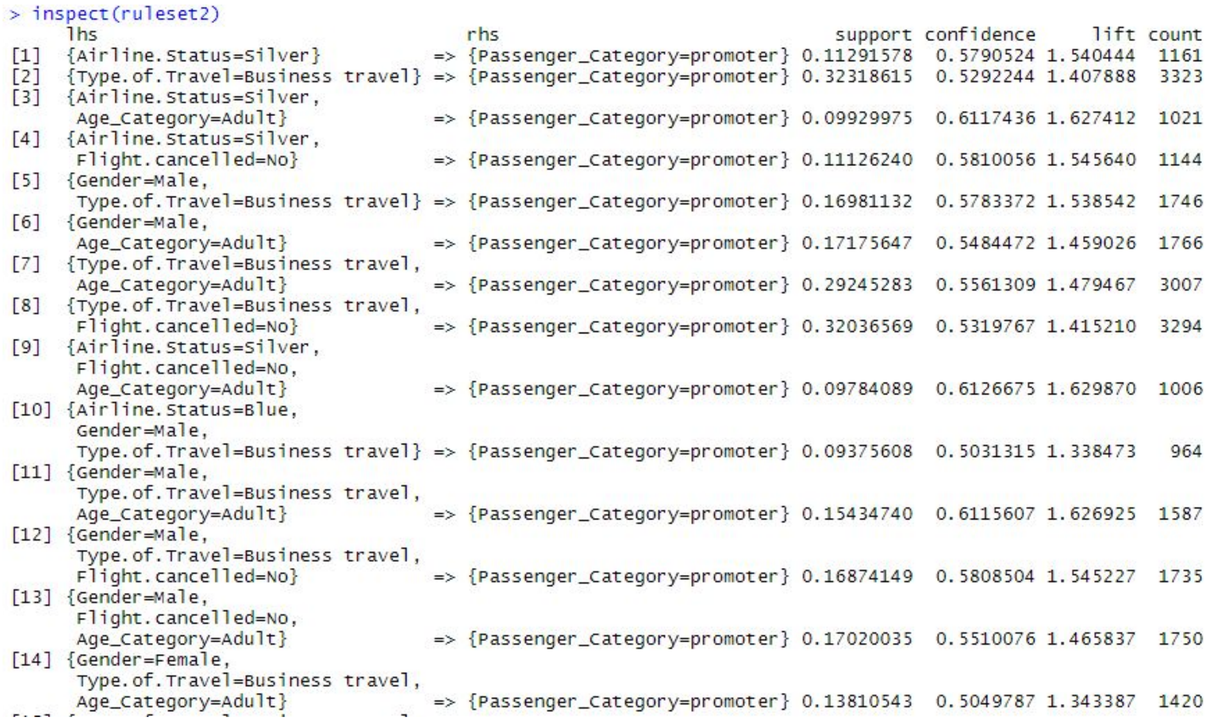
```
subrules1 = ruleset1[quality(ruleset1)$confidence>0.4]
#Plot for subrules1
plot(subrules1,method = "paracoord")
```



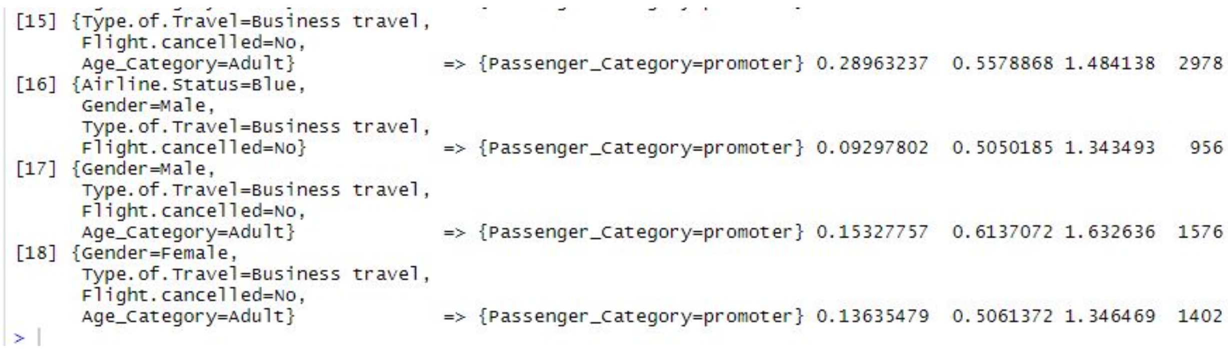
From the above plot we can observe various lines but derving from the most solid line we can say that the customer who has an Airline status as Blue and travelling as a personal type are detractors.

#Promoter code for entire dataset

```
ruleset2 <- apriori(DATA_X,
                parameter=list(support=0.09,confidence=0.5),
                appearance = list(default="lhs", rhs=("Passenger_Category=promoter")))

inspect(ruleset2)
```

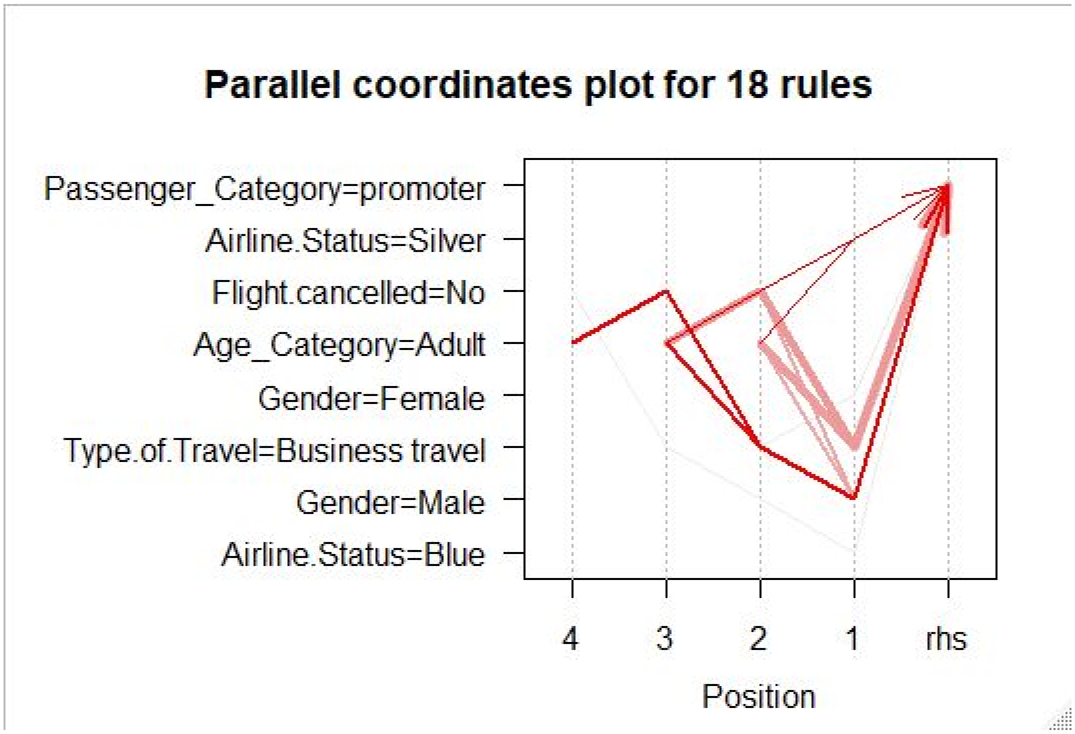


```
> inspect(ruleset2)
     lhs                                rhs                             support confidence     lift count
[1]  {Airline.Status=Silver}         => {Passenger_Category=promoter} 0.11291578  0.5790524 1.540444  1161
[2]  {Type.of.Travel=Business travel} => {Passenger_Category=promoter} 0.32318615  0.5292244 1.407888  3323
[3]  {Airline.Status=Silver,
      Age_Category=Adult}            => {Passenger_Category=promoter} 0.09929975  0.6117436 1.627412  1021
[4]  {Airline.Status=Silver,
      Flight.cancelled=No}           => {Passenger_Category=promoter} 0.11126240  0.5810056 1.545640  1144
[5]  {Gender=Male,
      Type.of.Travel=Business travel} => {Passenger_Category=promoter} 0.16981132  0.5783372 1.538542  1746
[6]  {Gender=Male,
      Age_Category=Adult}            => {Passenger_Category=promoter} 0.17175647  0.5484472 1.459026  1766
[7]  {Type.of.Travel=Business travel,
      Age_Category=Adult}            => {Passenger_Category=promoter} 0.29245283  0.5561309 1.479467  3007
[8]  {Type.of.Travel=Business travel,
      Flight.cancelled=No}           => {Passenger_Category=promoter} 0.32036569  0.5319767 1.415210  3294
[9]  {Airline.Status=Silver,
      Flight.cancelled=No,
      Age_Category=Adult}            => {Passenger_Category=promoter} 0.09784089  0.6126675 1.629870  1006
[10] {Airline.Status=Blue,
      Gender=Male,
      Type.of.Travel=Business travel} => {Passenger_Category=promoter} 0.09375608  0.5031315 1.338473   964
[11] {Gender=Male,
      Type.of.Travel=Business travel,
      Age_Category=Adult}            => {Passenger_Category=promoter} 0.15434740  0.6115607 1.626925  1587
[12] {Gender=Male,
      Type.of.Travel=Business travel,
      Flight.cancelled=No}           => {Passenger_Category=promoter} 0.16874149  0.5808504 1.545227  1735
[13] {Gender=Male,
      Flight.cancelled=No,
      Age_Category=Adult}            => {Passenger_Category=promoter} 0.17020035  0.5510076 1.465837  1750
[14] {Gender=Female,
      Type.of.Travel=Business travel,
      Age_Category=Adult}            => {Passenger_Category=promoter} 0.13810543  0.5049787 1.343387  1420
```



```
[15] {Type.of.Travel=Business travel,
      Flight.cancelled=No,
      Age_Category=Adult}            => {Passenger_Category=promoter} 0.28963237  0.5578868 1.484138  2978
[16] {Airline.Status=Blue,
      Gender=Male,
      Type.of.Travel=Business travel,
      Flight.cancelled=No}           => {Passenger_Category=promoter} 0.09297802  0.5050185 1.343493   956
[17] {Gender=Male,
      Type.of.Travel=Business travel,
      Flight.cancelled=No,
      Age_Category=Adult}            => {Passenger_Category=promoter} 0.15327757  0.6137072 1.632636  1576
[18] {Gender=Female,
      Type.of.Travel=Business travel,
      Flight.cancelled=No,
      Age_Category=Adult}            => {Passenger_Category=promoter} 0.13635479  0.5061372 1.346469  1402
> |
```

subrules2 = ruleset2[quality(ruleset2)$confidence>0.4]
#Plot for subrules2
plot(subrules2,method = "paracoord")



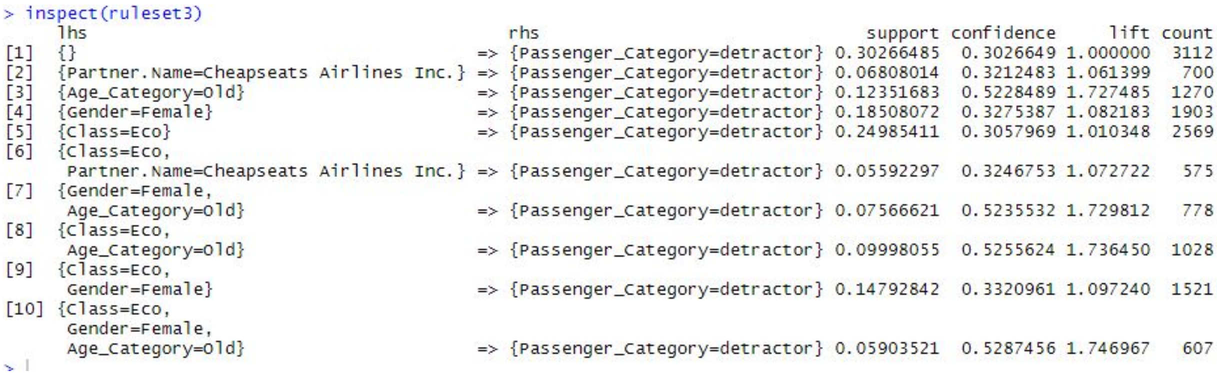
From the above plot we can observe that Male adult travelling for business type and when the flights are not cancelled are major Promoters.

#Association Rule and Bar plot verify
#1
#Detractor
#Class,Gender,Age Category,Partner Name,Passenger Category

```
dataset1<-
jsonfile_data[c('Class','Gender','Age_Category','Partner_Name','Passenger_Category')]
dataset1

dataset1$Gender=as.factor(dataset1$Gender)
dataset1$Class=as.factor(dataset1$Class)
dataset1$Partner_Name=as.factor(dataset1$Partner_Name)

str(dataset1)
DATA_X1 <- as(dataset1,"transactions")
DATA_X1 #Transaction Matrix


ruleset3 <- apriori(DATA_X1,
```

parameter=list(support=0.07,confidence=0.2),
appearance = list(default="lhs", rhs=("Passenger_Category=detractor")))

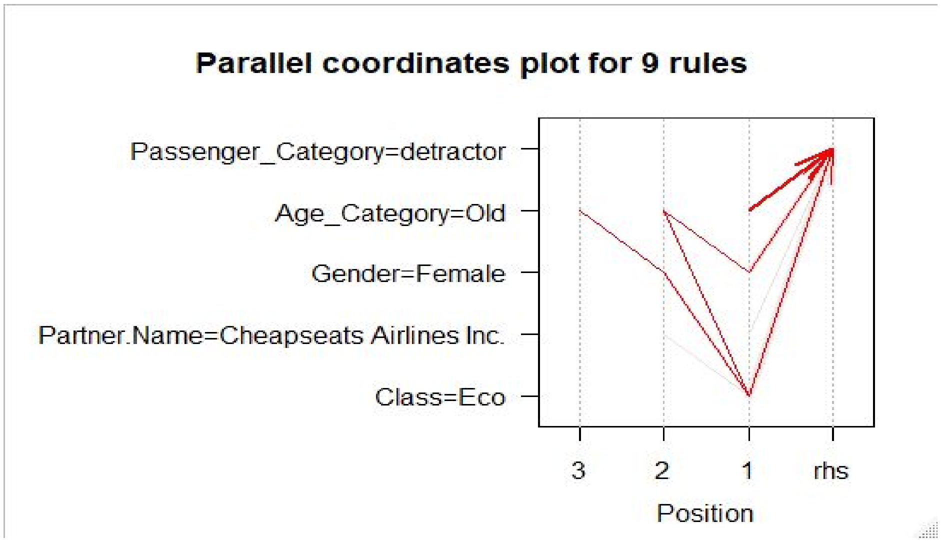
inspect(ruleset3)



```
> inspect(ruleset3)
     lhs                                     rhs                              support confidence     lift count
[1]  {}                                   => {Passenger_Category=detractor} 0.30266485  0.3026649 1.000000  3112
[2]  {Partner.Name=Cheapseats Airlines Inc.} => {Passenger_Category=detractor} 0.06808014  0.3212483 1.061399   700
[3]  {Age_Category=Old}                   => {Passenger_Category=detractor} 0.12351683  0.5228489 1.727485  1270
[4]  {Gender=Female}                      => {Passenger_Category=detractor} 0.18508072  0.3275387 1.082183  1903
[5]  {Class=Eco}                          => {Passenger_Category=detractor} 0.24985411  0.3057969 1.010348  2569
[6]  {Class=Eco,
      Partner.Name=Cheapseats Airlines Inc.} => {Passenger_Category=detractor} 0.05592297  0.3246753 1.072722   575
[7]  {Gender=Female,
      Age_Category=Old}                   => {Passenger_Category=detractor} 0.07566621  0.5235532 1.729812   778
[8]  {Class=Eco,
      Age_Category=Old}                   => {Passenger_Category=detractor} 0.09998055  0.5255624 1.736450  1028
[9]  {Class=Eco,
      Gender=Female}                      => {Passenger_Category=detractor} 0.14792842  0.3320961 1.097240  1521
[10] {Class=Eco,
      Gender=Female,
      Age_Category=Old}                   => {Passenger_Category=detractor} 0.05903521  0.5287456 1.746967   607
> 
```

subrules3 = ruleset3[quality(ruleset3)$confidence>0.4]

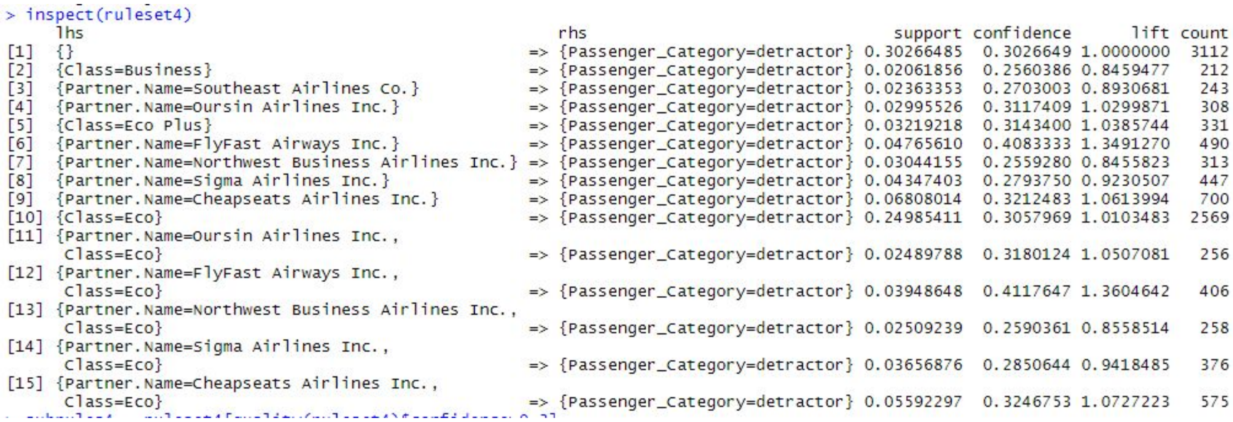
#Plot for subrules3
plot(subrules3,method = "paracoord")



From the above plot we get to know that the Old female travelling through Eco class are detractors.

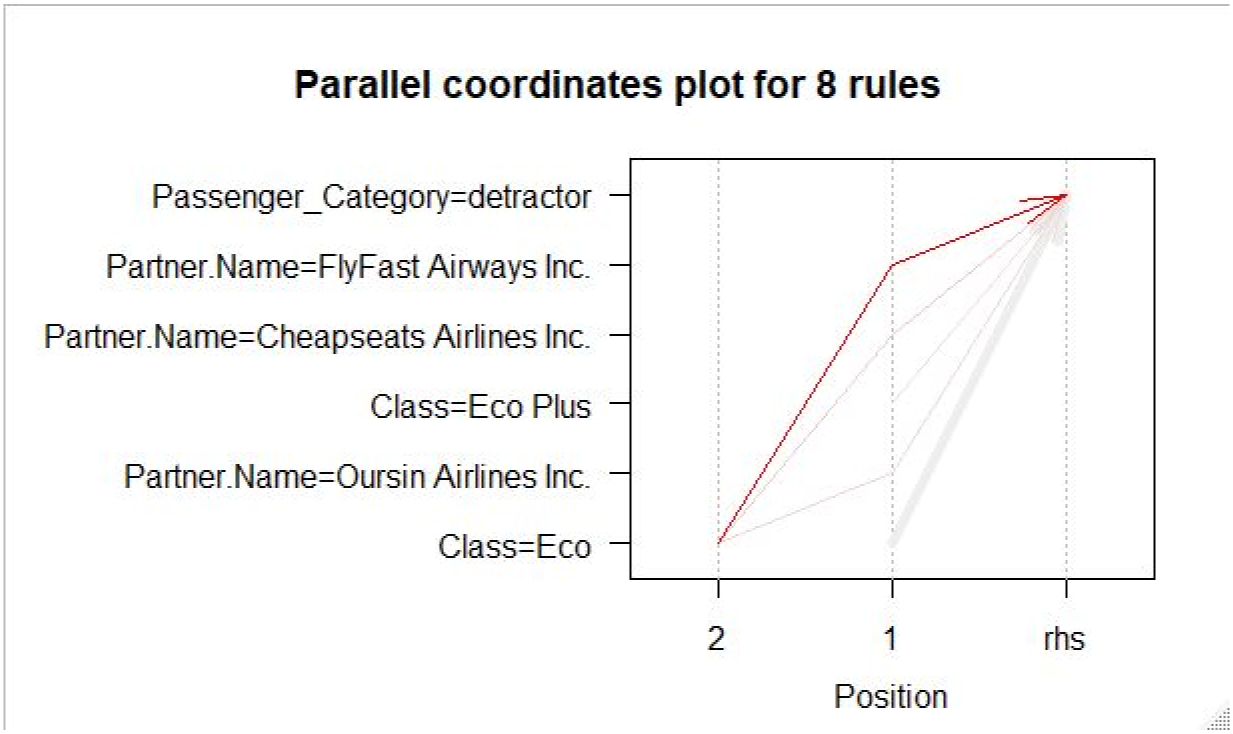
#2
#Detractor
#Partner Name,Class, Passenger Category

```
dataset2<- jsonfile_data[c('Partner_Name','Class', 'Passenger_Category')]
dataset2

str(dataset2)

dataset2$Partner_Name=as.factor(dataset2$Partner_Name)
dataset2$Class=as.factor(dataset2$Class)

DATA_X3 <- as(dataset2,"transactions")
DATA_X3 #Transaction Matrix


ruleset4 <- apriori(DATA_X3,
                 parameter=list(support=0.02,confidence=0.25),
                 appearance = list(default="lhs", rhs=("Passenger_Category=detractor")))

inspect(ruleset4)
```



```
> inspect(ruleset4)
     lhs                                           rhs                            support confidence      lift count
[1]  {}                                         => {Passenger_Category=detractor} 0.30266485  0.3026649 1.0000000  3112
[2]  {Class=Business}                           => {Passenger_Category=detractor} 0.02061856  0.2560386 0.8459477   212
[3]  {Partner.Name=Southeast Airlines Co.}      => {Passenger_Category=detractor} 0.02363353  0.2703003 0.8930681   243
[4]  {Partner.Name=Oursin Airlines Inc.}        => {Passenger_Category=detractor} 0.02995526  0.3117409 1.0299871   308
[5]  {Class=Eco Plus}                           => {Passenger_Category=detractor} 0.03219218  0.3143400 1.0385744   331
[6]  {Partner.Name=FlyFast Airways Inc.}        => {Passenger_Category=detractor} 0.04765610  0.4083333 1.3491270   490
[7]  {Partner.Name=Northwest Business Airlines Inc.} => {Passenger_Category=detractor} 0.03044155  0.2559280 0.8455823   313
[8]  {Partner.Name=Sigma Airlines Inc.}         => {Passenger_Category=detractor} 0.04347403  0.2793750 0.9230507   447
[9]  {Partner.Name=Cheapseats Airlines Inc.}    => {Passenger_Category=detractor} 0.06808014  0.3212483 1.0613994   700
[10] {Class=Eco}                                => {Passenger_Category=detractor} 0.24985411  0.3057969 1.0103483  2569
[11] {Partner.Name=Oursin Airlines Inc.,
      Class=Eco}                                => {Passenger_Category=detractor} 0.02489788  0.3180124 1.0507081   256
[12] {Partner.Name=FlyFast Airways Inc.,
      Class=Eco}                                => {Passenger_Category=detractor} 0.03948648  0.4117647 1.3604642   406
[13] {Partner.Name=Northwest Business Airlines Inc.,
      Class=Eco}                                => {Passenger_Category=detractor} 0.02509239  0.2590361 0.8558514   258
[14] {Partner.Name=Sigma Airlines Inc.,
      Class=Eco}                                => {Passenger_Category=detractor} 0.03656876  0.2850644 0.9418485   376
[15] {Partner.Name=Cheapseats Airlines Inc.,
      Class=Eco}                                => {Passenger_Category=detractor} 0.05592297  0.3246753 1.0727223   575
```

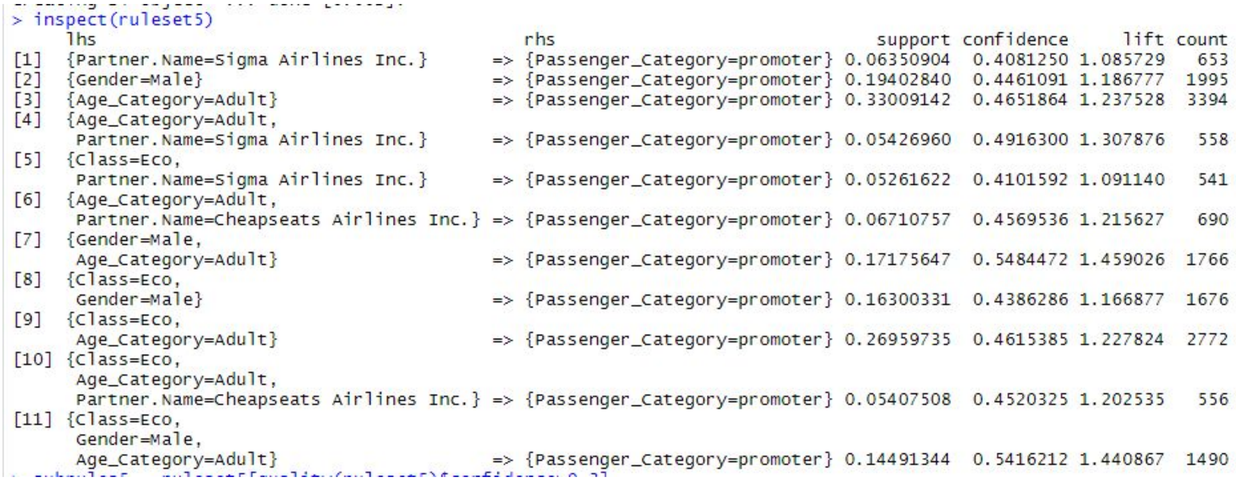
```
subrules4 = ruleset4[quality(ruleset4)$confidence>0.3]

#Plot for subrules4
plot(subrules4,method = "paracoord")
```



From the above plot we can observe that the customers travelling by Eco class with FlyFast Airways are the major detractors.

```
#3
#promoter
#Class,Gender,Age Category,Partner Name,Passenger Category

dataset3<-
jsonfile_data[c('Class','Gender','Age_Category','Partner_Name','Passenger_Category')]
dataset3

dataset3$Gender=as.factor(dataset3$Gender)
dataset3$Class=as.factor(dataset3$Class)
dataset3$Partner_Name=as.factor(dataset3$Partner_Name)

str(dataset3)
DATA_X3 <- as(dataset1,"transactions")
DATA_X3 #Transaction Matrix


ruleset5 <- apriori(DATA_X3,
                  parameter=list(support=0.07,confidence=0.2),
                appearance = list(default="lhs", rhs=("Passenger_Category=promoter")))
```

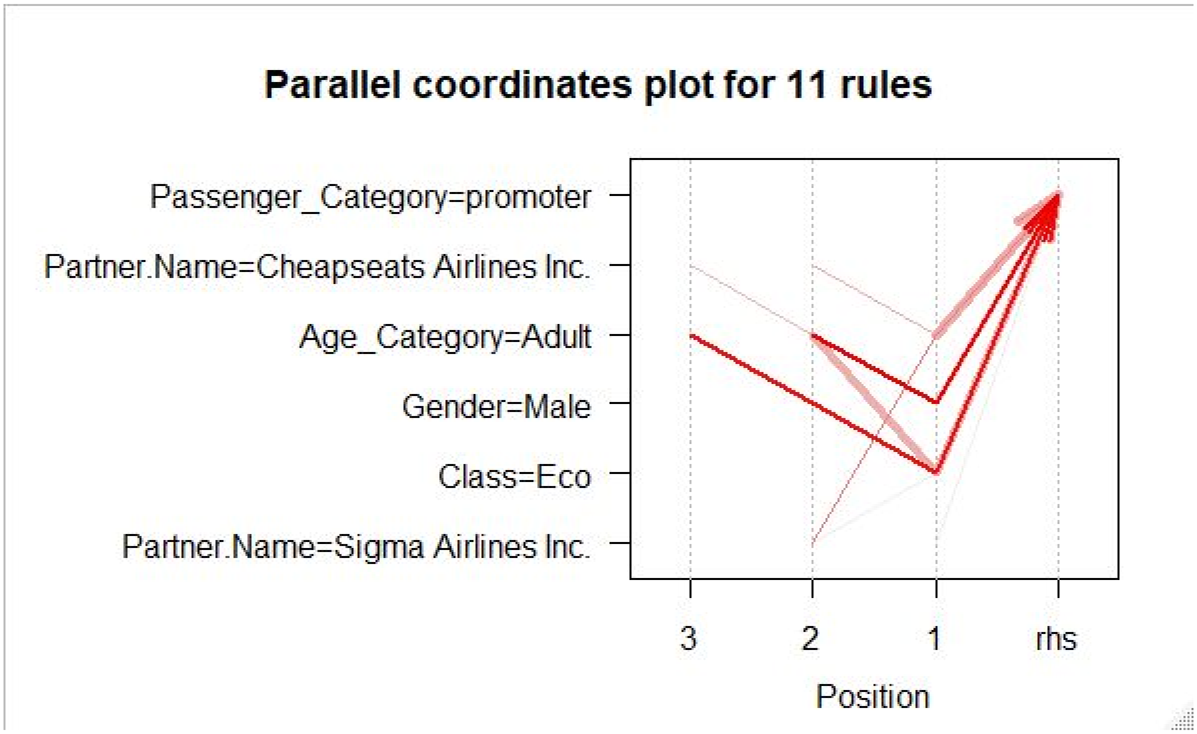
```
inspect(ruleset5)
```



```
> inspect(ruleset5)
     lhs                                        rhs                           support confidence     lift count
[1]  {Partner.Name=Sigma Airlines Inc.}      => {Passenger_Category=promoter} 0.06350904  0.4081250 1.085729   653
[2]  {Gender=Male}                           => {Passenger_Category=promoter} 0.19402840  0.4461091 1.186777  1995
[3]  {Age_Category=Adult}                    => {Passenger_Category=promoter} 0.33009142  0.4651864 1.237528  3394
[4]  {Age_Category=Adult,
      Partner.Name=Sigma Airlines Inc.}      => {Passenger_Category=promoter} 0.05426960  0.4916300 1.307876   558
[5]  {Class=Eco,
      Partner.Name=Sigma Airlines Inc.}      => {Passenger_Category=promoter} 0.05261622  0.4101592 1.091140   541
[6]  {Age_Category=Adult,
      Partner.Name=Cheapseats Airlines Inc.} => {Passenger_Category=promoter} 0.06710757  0.4569536 1.215627   690
[7]  {Gender=Male,
      Age_Category=Adult}                    => {Passenger_Category=promoter} 0.17175647  0.5484472 1.459026  1766
[8]  {Class=Eco,
      Gender=Male}                           => {Passenger_Category=promoter} 0.16300331  0.4386286 1.166877  1676
[9]  {Class=Eco,
      Age_Category=Adult}                    => {Passenger_Category=promoter} 0.26959735  0.4615385 1.227824  2772
[10] {Class=Eco,
      Age_Category=Adult,
      Partner.Name=Cheapseats Airlines Inc.} => {Passenger_Category=promoter} 0.05407508  0.4520325 1.202535   556
[11] {Class=Eco,
      Gender=Male,
      Age_Category=Adult}                    => {Passenger_Category=promoter} 0.14491344  0.5416212 1.440867  1490
```

```
subrules5 = ruleset5[quality(ruleset5)$confidence>0.4]
```

```
#Plot for subrules5
plot(subrules5,method = "paracoord")
```

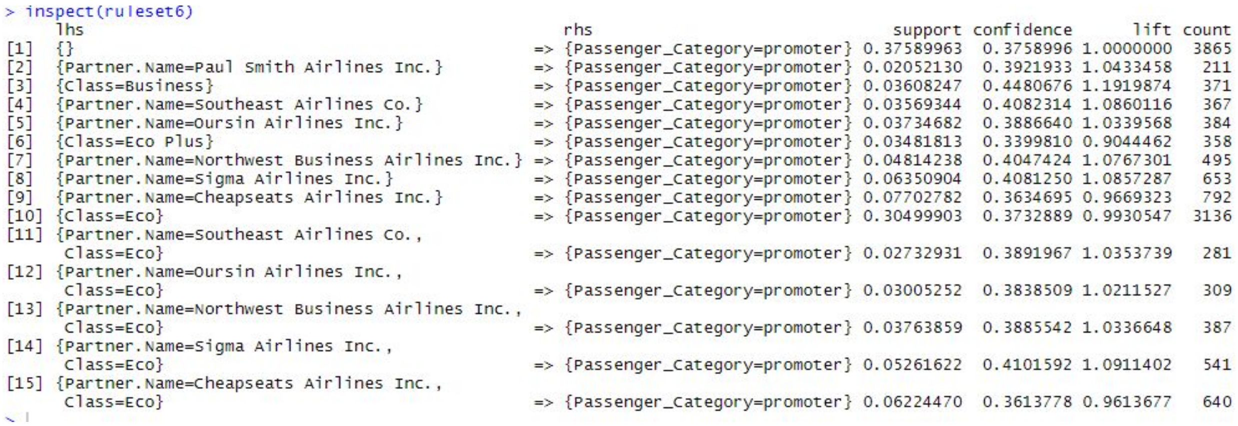


From the above plot we can observe that adult male customers travelling by Eco class contribute to be promoters.

#4
#promoter
#Partner Name,Class, Passenger Category

```
dataset4<- jsonfile_data[c('Partner_Name','Class', 'Passenger_Category')]
dataset4

str(dataset4)

dataset4$Partner_Name=as.factor(dataset4$Partner_Name)
dataset4$Class=as.factor(dataset4$Class)

DATA_X4 <- as(dataset4,"transactions")
DATA_X4 #Transaction Matrix


ruleset6 <- apriori(DATA_X4,
                parameter=list(support=0.02,confidence=0.25),
                appearance = list(default="lhs", rhs=("Passenger_Category=promoter")))
```

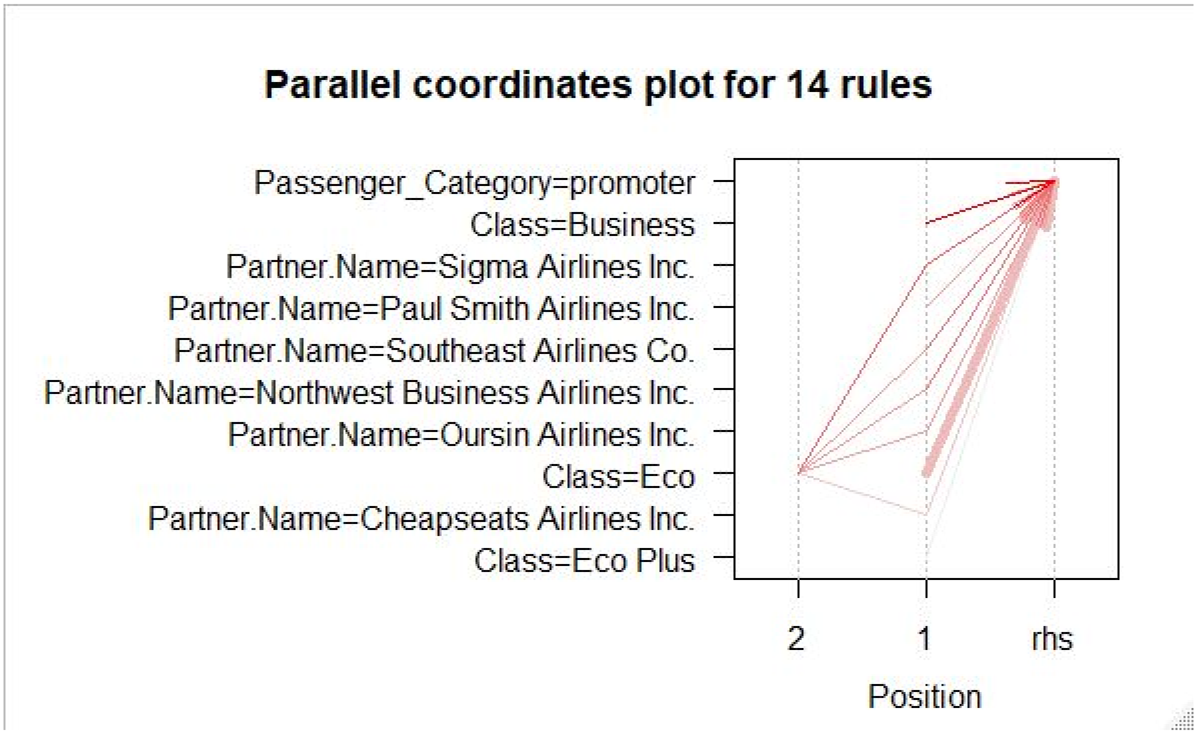
inspect(ruleset6)



```
> inspect(ruleset6)
     lhs                                                rhs                               support confidence      lift count
[1]  {}                                              => {Passenger_Category=promoter} 0.37589963  0.3758996 1.0000000  3865
[2]  {Partner.Name=Paul Smith Airlines Inc.}         => {Passenger_Category=promoter} 0.02052130  0.3921933 1.0433458   211
[3]  {Class=Business}                                => {Passenger_Category=promoter} 0.03608247  0.4480676 1.1919874   371
[4]  {Partner.Name=Southeast Airlines Co.}           => {Passenger_Category=promoter} 0.03569344  0.4082314 1.0860116   367
[5]  {Partner.Name=Oursin Airlines Inc.}             => {Passenger_Category=promoter} 0.03734682  0.3886640 1.0339568   384
[6]  {Class=Eco Plus}                                => {Passenger_Category=promoter} 0.03481813  0.3399810 0.9044462   358
[7]  {Partner.Name=Northwest Business Airlines Inc.} => {Passenger_Category=promoter} 0.04814238  0.4047424 1.0767301   495
[8]  {Partner.Name=Sigma Airlines Inc.}              => {Passenger_Category=promoter} 0.06350904  0.4081250 1.0857287   653
[9]  {Partner.Name=Cheapseats Airlines Inc.}         => {Passenger_Category=promoter} 0.07702782  0.3634695 0.9669323   792
[10] {Class=Eco}                                     => {Passenger_Category=promoter} 0.30499903  0.3732889 0.9930547  3136
[11] {Partner.Name=Southeast Airlines Co.,
      Class=Eco}                                     => {Passenger_Category=promoter} 0.02732931  0.3891967 1.0353739   281
[12] {Partner.Name=Oursin Airlines Inc.,
      Class=Eco}                                     => {Passenger_Category=promoter} 0.03005252  0.3838509 1.0211527   309
[13] {Partner.Name=Northwest Business Airlines Inc.,
      Class=Eco}                                     => {Passenger_Category=promoter} 0.03763859  0.3885542 1.0336648   387
[14] {Partner.Name=Sigma Airlines Inc.,
      Class=Eco}                                     => {Passenger_Category=promoter} 0.05261622  0.4101592 1.0911402   541
[15] {Partner.Name=Cheapseats Airlines Inc.,
      Class=Eco}                                     => {Passenger_Category=promoter} 0.06224470  0.3613778 0.9613677   640
>
```

subrules6 = ruleset6[quality(ruleset6)$confidence>0.3]

#Plot for subrules6
plot(subrules6,method = "paracoord")



From the above plot we can observe that the customer travelling by Business class are promoter and also after that the customer travelling by Eco class from Sigma airlines are also good promoters.

In summary, we ran the association rule analysis on the entire dataset (as we can see in the above plots of subrules 1 and subrules 2), and also on small datasets (as we can see in above plots for subrules 3,subrules 4,subrules 5 and subrules 6) in order to address our business questions of: 1) Who are the happy or unhappy customers , 2)What makes a successful flight experience.

Our conclusions are listed below:

1. For larger dataset we found the following results:

- We came up with an overall conclusion that customers who are travelling personally with airline status as blue in general and when the flights are not cancelled are most likely to have been into detractors category and Southeast airlines officials must look into it to improve their experience while travelling so that the next flight experience will be better and will turn as happy customers.

- We also found that apart from the above points which is for detractors we were able to find that the male adults customers who are travelling for business trips and when their flights are not cancelled are happy customers.

2. For smaller dataset we found the following results:

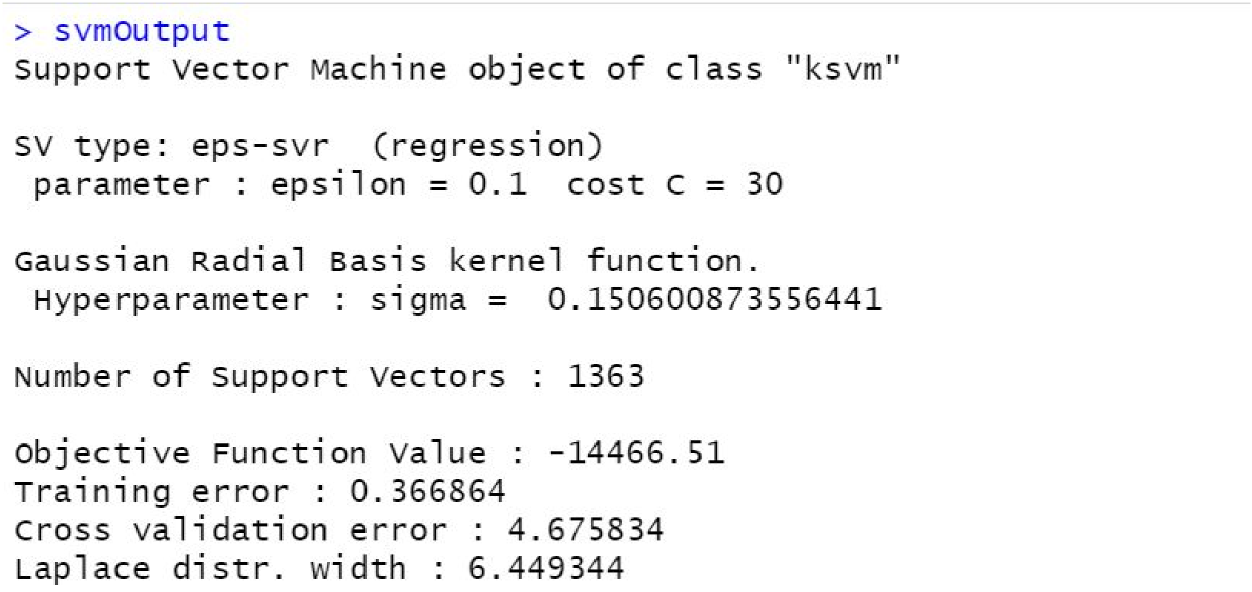
- Female old customers travelling by Eco class are unhappy customers. Also we can conclude that specifically travelling by FlyFast Airlines are potentially the unhappy customers and can be grouped into detractors passenger category.
- Business class customers and adults male customers travelling by Eco class through Sigma Airlines Inc. have provided a positive feedback. Hence they had a good flight experience which makes them happy. Such customers can be grouped into promoters passenger category.
- Hence the services provided for promoters (from the above point) must be looked into and similar services must be provided to customers who are into detractors category.

5.3 Support Vector Machine Modeling

SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. By using the above two modeling techniques, we found out the attributes affecting the customer's likelihood to recommend more significantly than other attributes in the data. The quality of the data in these attributes should be tested before we can make sure they are termed good attributes. Therefore, the SVM model was introduced to analyze the quality of the data in several attributes. The reason support vector machines are considered a supervised learning technique is that we "train" the algorithm on an initial set of data and then we test it out on a brand-new set of data. If the training we accomplished worked well, then the algorithm should be able to predict the right outcome most of the time in the test data.

CODE:

```
library(kernlab)

trainindex <- sample(c(1,2,3), nrow(transform_data),replace= T,prob = c(0.15,0.45,0.4))

traindata <- transform_data[trainindex==1,]

testdata <- transform_data[trainindex==2,]

svmOutput <- ksvm(Likelihood_to_recommend ~ Airline_Status + Age +  Price_Sensitivity +

          Flight_Distance +  Type_of_Travel +  Flight_cancelled +

          Shopping_Amount_at_Airport + Class +

          Departure_Delay_in_Minutes + Arrival_Delay_in_Minutes,

         data=traindata,kernel="rbfdot", kpar="automatic",C=20,cross=10,
prob.model=TRUE)

svmOutput
```



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

SV type: eps-svr  (regression)
 parameter : epsilon = 0.1  cost C = 30

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.150600873556441

Number of Support Vectors : 1363

Objective Function Value : -14466.51
Training error : 0.366864
Cross validation error : 4.675834
Laplace distr. width : 6.449344
```

```
#View(testdata)

svmresult <- predict(svmOutput,testdata,type="votes")

#View(svmresult)

compactable <- (testdata[,27]- svmresult)<10&(testdata[,27]-svmresult) > 0

result <- table(compactable)

result

accuraryratio <- result[2]/(sum(result))

accuraryratio

#accuracy ratio: 0.5214379
```

The accuracy ratio of the SVM model is 52%, which would, in some aspects, guarantee the quality of the selected attributes used in the SVM model. predict the right outcome most of the time in our data set. It seems that the SVM model is testing the model rather than the quality of data. However, since we can manage the attribute imported in the SVM, the best quality of the model also suggests that whether the data in these attributes could be classified with high quality.

6. ACTIONABLE INSIGHTS

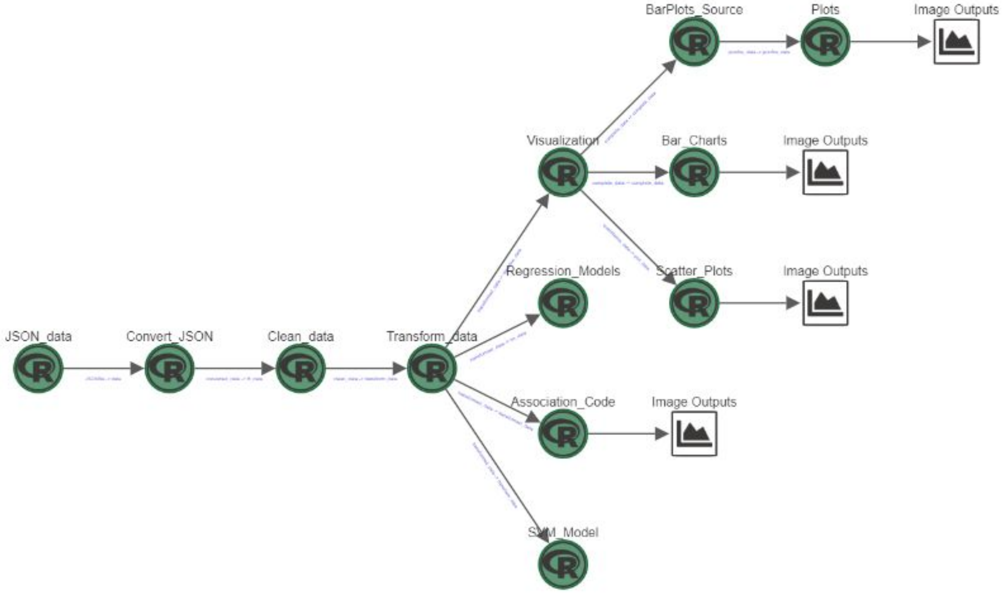
Here are some actionable insights and suggestions that Southeast Airlines can take.

1. Provide more services and lower price for Blue members to prevent them from becoming detractors.
2. Services provided by Sigma Airlines for economy class must be implemented by other airlines partners.
3. There are some targeted customers that have higher NPS and they are more likely to be promoters. For example, customers travelling by business Class and male adult by economy class. We think that targeted advertisements on mobile phone or website can be implemented to keep promoters and encourage them to travel frequently.
4. Team should focus more on old female travelling by economy class through FlyFast Airlines.
5. Services provided to personal travel customers travelling through Blue status should be improved which will make them a happy customer.

7. MIDST

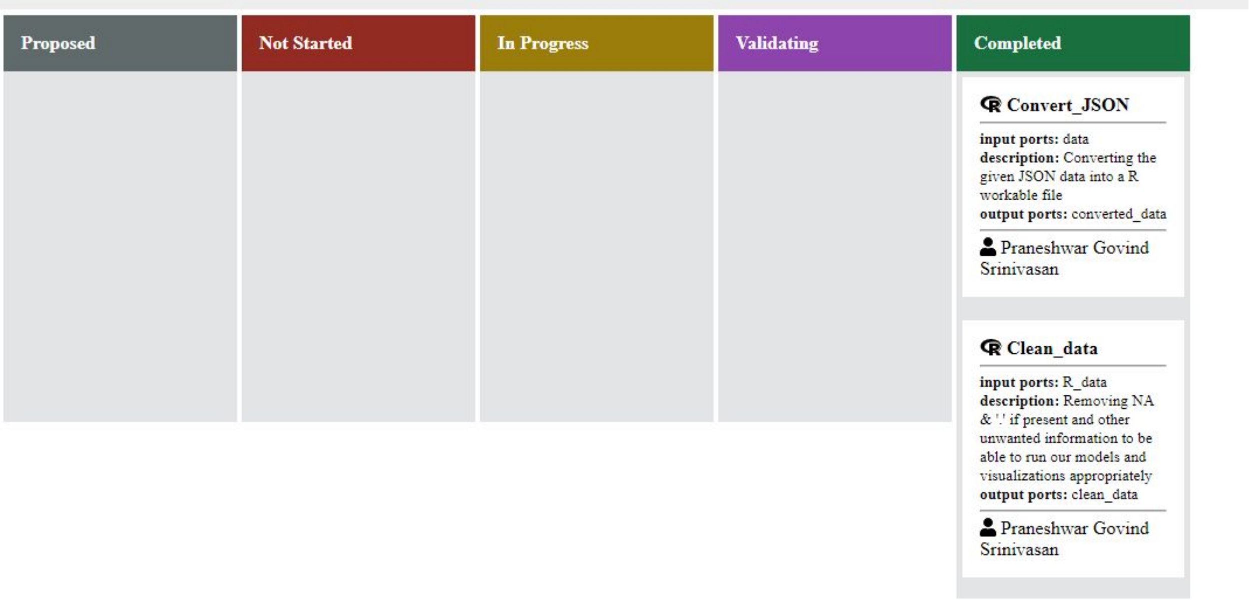
For working and collaboration we made use of MIDST. During the tenure of the project as a group we divided up the tasks and decided who will work on which part and updates were done in the MIDST.Major task which were completed are cleaning of data and exploring ahead followed by building models using Linear Regression Model,Association Rules and Support Vector Machine. Then we analysed each of the results and provided our analysis for the problem

NETWORK:



.

KANBAN BOARD:



Proposed

Not Started

In Progress

Validating

Completed

Convert_JSON

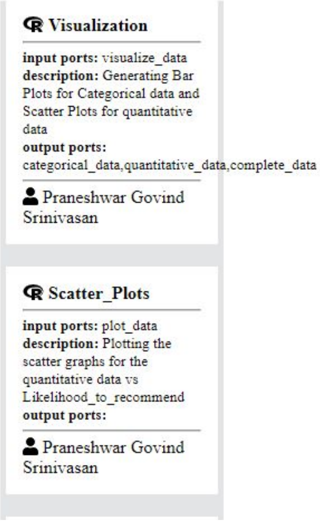
input ports: data
description: Converting the given JSON data into a R workable file
output ports: converted_data

Praneshwar Govind Srinivasan

Clean_data

input ports: R_data
description: Removing NA & '.' if present and other unwanted information to be able to run our models and visualizations appropriately
output ports: clean_data

Praneshwar Govind Srinivasan



Visualization

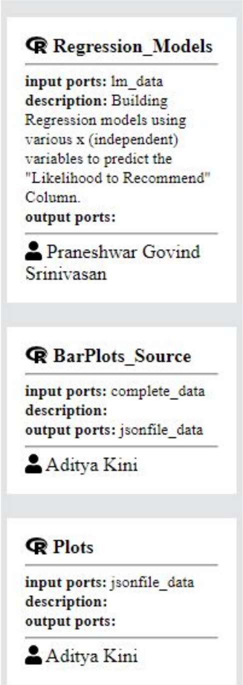
input ports: visualize_data
description: Generating Bar Plots for Categorical data and Scatter Plots for quantitative data
output ports: categorical_data,quantitative_data,complete_data

Praneshwar Govind Srinivasan

Scatter_Plots

input ports: plot_data
description: Plotting the scatter graphs for the quantitative data vs Likelihood_to_recommend
output ports:

Praneshwar Govind Srinivasan



Regression_Models

input ports: lm_data
description: Building Regression models using various x (independent) variables to predict the "Likelihood to Recommend" Column.
output ports:

Praneshwar Govind Srinivasan

BarPlots_Source

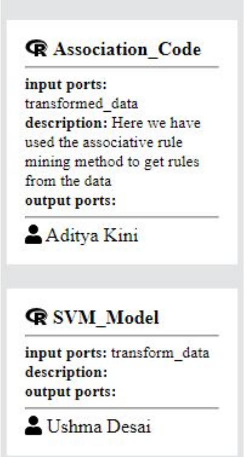
input ports: complete_data
description:
output ports: jsonfile_data

Aditya Kini

Plots

input ports: jsonfile_data
description:
output ports:

Aditya Kini



Association_Code

input ports: transformed_data
description: Here we have used the associative rule mining method to get rules from the data
output ports:

Aditya Kini

SVM_Model

input ports: transform_data
description:
output ports:

Ushma Desai