**School of Information Studies**

**SYRACUSE UNIVERSITY**

AIRLINES ANALYSIS

FINAL PROJECT REPORT

M004 GROUP 4

Juilee Salunkhe

Po Kai Huang

Raj Desai

Rishabh Upadhye

Xioben Yin

12/10/2019

Table of Contents

[1. FINAL PROJECT SUMMARY 3](#_Toc15842091)

[1.1 BUSINESS QUESTIONS 3](#_Toc15842092)

[1.2 EXPLORING AND DATA CLEANING 4](#_Toc15842092)

[2. DATA MODELLING 12](#_Toc15842091)

[2.1 LINEAR MODELLING 12](#_Toc15842094)

[2.2 ASSOCIATION RULES 28](#_Toc15842095)

[2.3 DECISION TREE 35](#_Toc15842097)

[3. ACTIONABLE INSIGHTS 6](#_Toc15842098)

Project summary

Southeast Airlines needed to lower their customer churn, to lower the same the airlines tried to have a robust loyalty program for the frequent flyers. The customers could bank their points and use it for their next flight with the airlines. However, the customers were valuing the program less, so this is the reason why the airline does not want to rely on this program to keep the customer churn less.

The real goal is to reduce the churn by getting ahead of loss of customers by identifying some indicators or metrics that will help us identify when a customer was about to stop flying in the Southeast Airlines. We have a dataset of customers who the airlines had recently surveyed. Using the 32 characteristics of the data we provided some business solutions which will help the airlines to lower the customer churn and will help to select the regional airlines the company has to do business in future.

BUSINESS QUESTIONS

1. Customers who are likely to recommend Southeast Airlines or Partner Airlines.
2. Ways to convert a Detractor to Promoter.
3. All attributes that contribute to one becoming a detractor. Targeting those attributes.
4. Airline/s which are not meeting customer’s expectation.
5. Concentration of Customers as per the Region/city/location that is not happy with the airlines.
6. Understand and promote Customers class-wise.
7. Actionable insights to increase the customer satisfaction overall.

EXPLORING & [DATA CLEANING](https://word-edit.officeapps.live.com/we/wordeditorframe.aspx?ui=en%2DUS&rs=en%2DUS&wopisrc=https%3A%2F%2Fsumailsyr-my.sharepoint.com%2Fpersonal%2Frmupadhy_syr_edu%2F_vti_bin%2Fwopi.ashx%2Ffiles%2Ff6858e2a5cae4b83b55b2892796de412&wdenableroaming=1&mscc=1&wdodb=1&hid=2643209F-500D-A000-A1A4-7F6D652C96B9&wdorigin=Sharing&jsapi=1&newsession=1&corrid=74747185-6ef3-42eb-9eb3-f4fc8c1a61b5&usid=74747185-6ef3-42eb-9eb3-f4fc8c1a61b5&instantedit=1&wopicomplete=1&wdredirectionreason=Unified_SingleFlush#_Toc15842093)

The survey dataset contained thousands of observations of flight segment data collected by Southeast Airlines. To begin with, we had an airline data consisting of total 32 variables (Each column represents an attribute of that flight segment) with 10282 entries. Further delving showed us that there are few variables that can be used to demographically represent the data of the flights (Destination city, state, Origin city, state ,latitudes and longitudes of destination and Origins),few others can be termed as Personal Customer’s data (Age, Gender, shopping at airports, eating & drinking at airports, Class, Flights per year, Type of travel, Year of the first flight) and some provide us the information on flight’s time, departure delay, arrival delay, Date ,etc. There are also comments, suggestions provided by the customers under the free text column.

We performed basic descriptive statistical analysis on the numeric data (e.g. the minimum Age of passengers is 15 whereas maximum is 85 and the dataset is of the airline that operates in 198 cities). We have studied the dataset column-wise and understood few business rules that can be seen in the data. For e.g. the price sensitivity is scaled from 0 to 4, Type of travel has three inputs (Mileage tickets, Business travel and Personal travel) and so on.

1. Data Munging

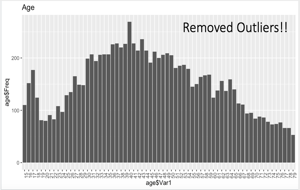
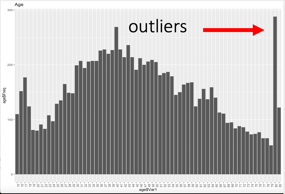
Data cleaning is necessary before we can explore and create models. After understanding the dataset, we started the data munging process by altering the names of the dataset columns and replaced NAs or missing values.

A screenshot of a cell phone

Description automatically generated



Then we converted the columns to factors and numeric to ease the application of functions on each of them. Also, we removed outliers from the data set by dividing the entries of 80-85 age equally to all the age groups. Eliminated unwanted data.

1. Net Promoter Score

Net Promoter is both a loyalty metric and a discipline for using customer feedback to fuel profitable growth. The Net Promoter Score, or NPS® is calculated by placing a company's customers into three categories: Promoters, Passives, and Detractors. Customers are asked one key loyalty question: How likely is it that you would recommend [Company X or Brand X] to a friend or colleague? Customers respond on a 0-to-10 point rating scale, with 0 being not at all likely, and 10 being extremely likely to recommend. Responses are categorized as follows:

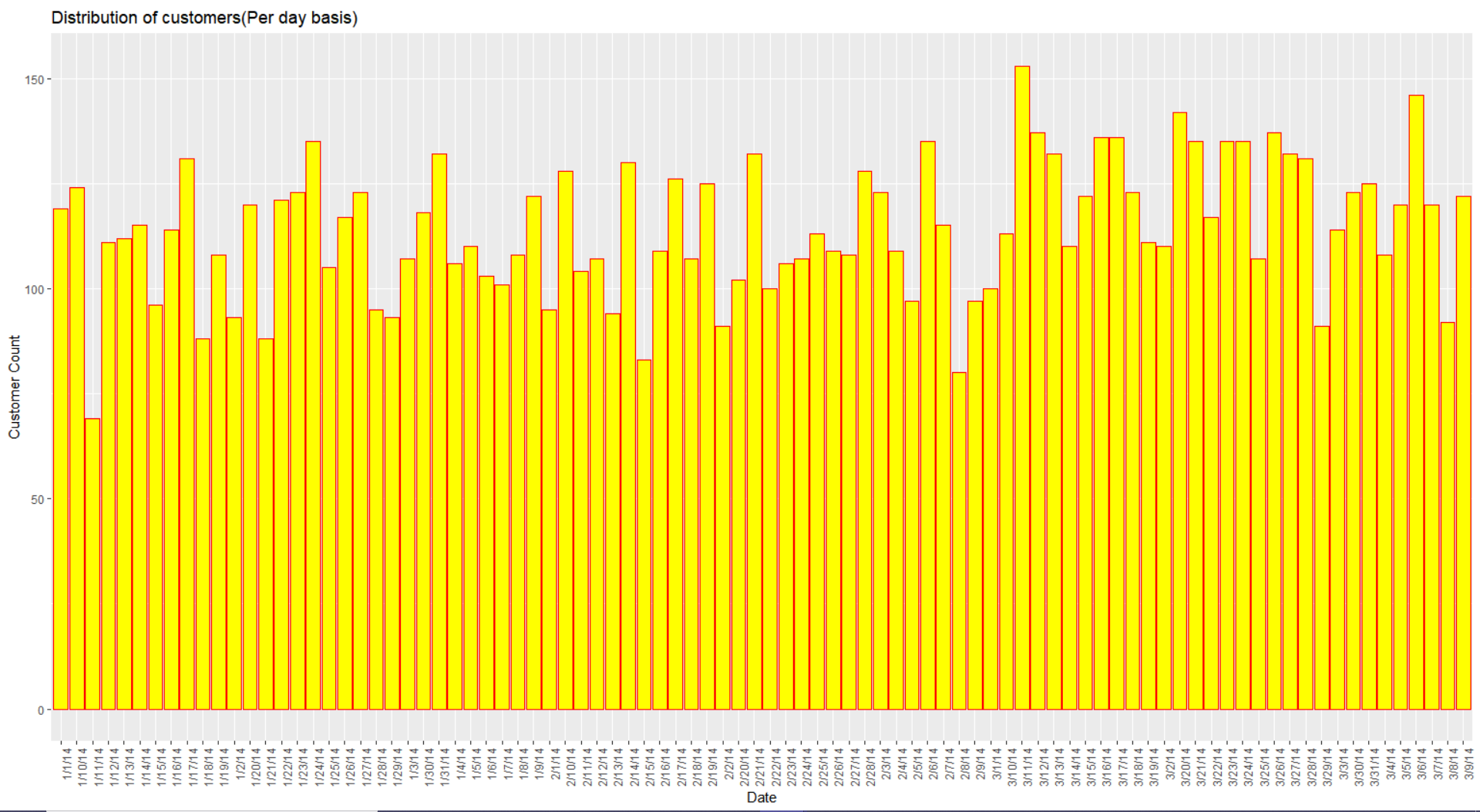
* Promoters (score 9-10) are loyal enthusiasts who will keep buying and refer others, fueling growth.
* Passives (score 7-8) are satisfied but unenthusiastic customers who are vulnerable to competitive offerings.
* Detractors(score0-6) are unhappy customers who can damage your brand and impede growth through negative word-of-mouth.

To calculate a company's Net Promoter Score, take the percentage of customers who are Promoters and subtract the percentage who are Detractors. The resulting NPS can provide insight on competitive position among companies in a given industry. By understanding attributes of the customer experience that influence the recommend score and NPS, companies can make improvements to product and service design and delivery to support profitable growth.

For calculating further insights about the dataset, we calculated the NPS value for each and every partner airlines. For implementing this, we created a function to which we passed the airlines data frame so the function automatically calculates the value of NPS and stores into an array. Now, array is converted into the data frame so that we can perform various plots on this new data frame for visualizing the values. What we found that West Airlines had the highest NPS value and the northwest business airlines had the lowest NPS value.

1. Narrowing down the data set (Analysis using NPS)

We focused on narrowing the data to the airlines with lowest median value of customer satisfaction. So, we narrowed down the dataset to the bottom 2 airlines having the least NPS score. We were planning to do this so that we can understand why these airlines are having a greater number of detractors. This would help us to convert these detractors in promoters.

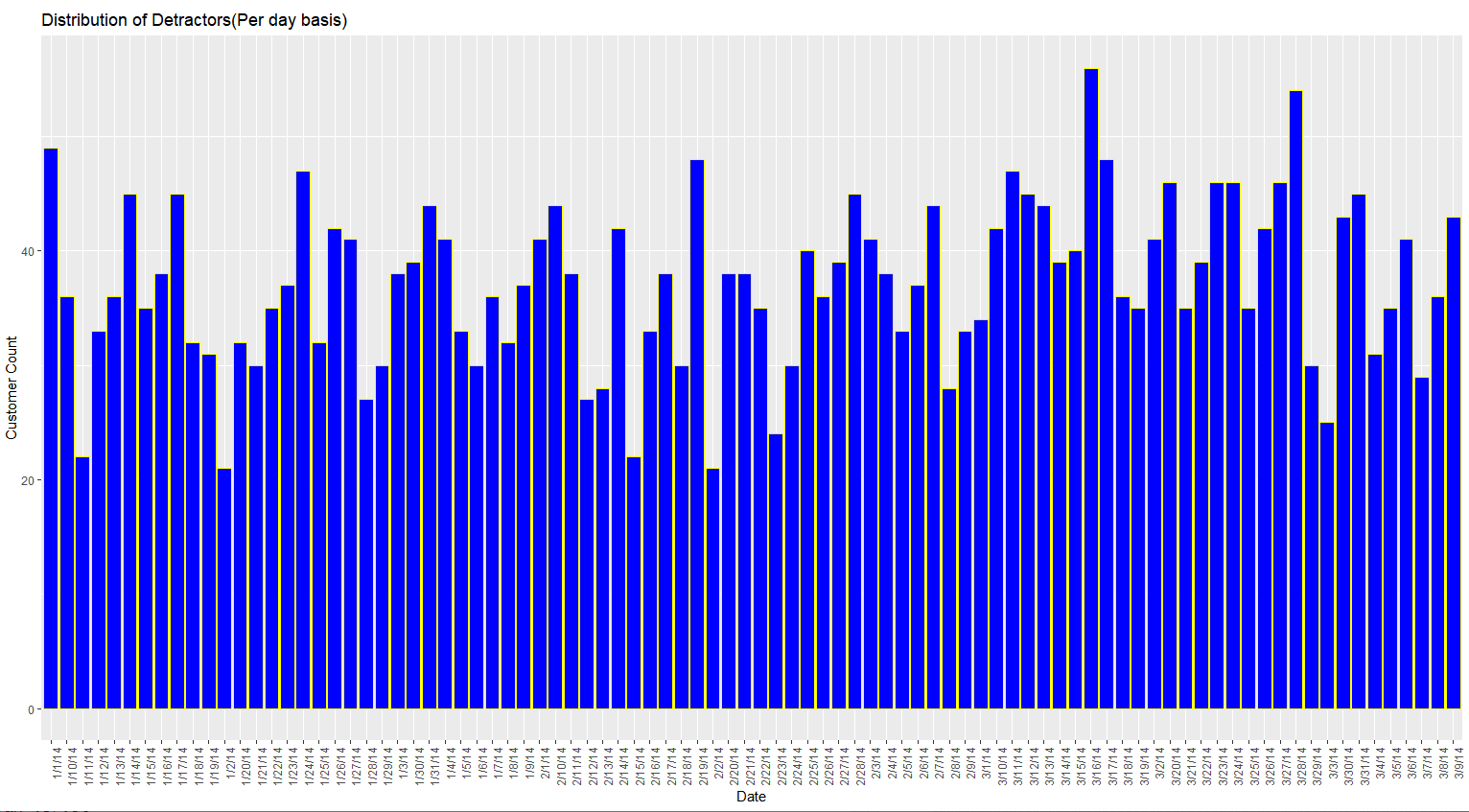
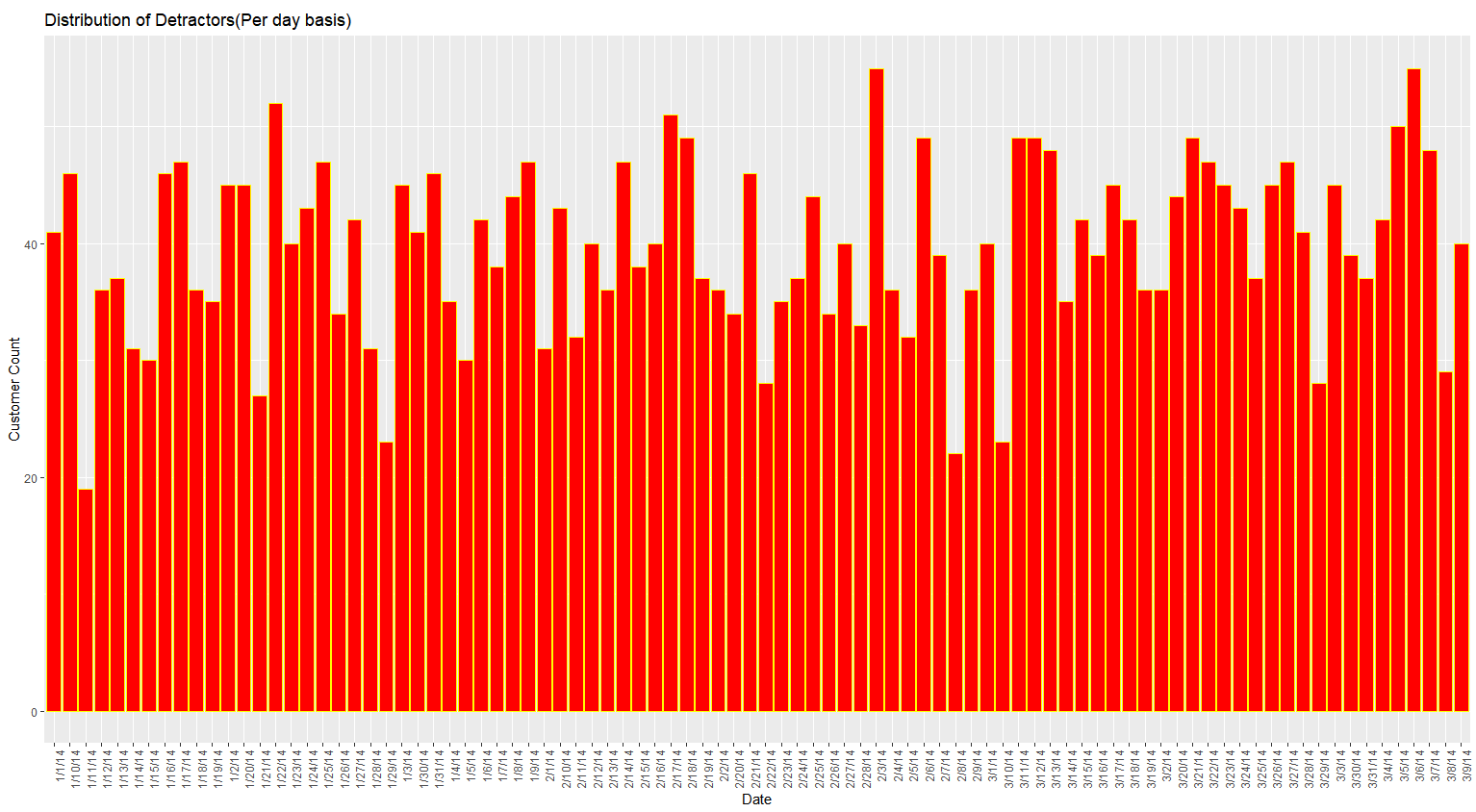


Average Customer per day: 114

Average Customer per day: **114**

[NARROWING DOWN THE DATA SET](https://word-edit.officeapps.live.com/we/wordeditorframe.aspx?ui=en%2DUS&rs=en%2DUS&wopisrc=https%3A%2F%2Fsumailsyr-my.sharepoint.com%2Fpersonal%2Frmupadhy_syr_edu%2F_vti_bin%2Fwopi.ashx%2Ffiles%2Ff6858e2a5cae4b83b55b2892796de412&wdenableroaming=1&mscc=1&wdodb=1&hid=2643209F-500D-A000-A1A4-7F6D652C96B9&wdorigin=Sharing&jsapi=1&newsession=1&corrid=74747185-6ef3-42eb-9eb3-f4fc8c1a61b5&usid=74747185-6ef3-42eb-9eb3-f4fc8c1a61b5&instantedit=1&wopicomplete=1&wdredirectionreason=Unified_SingleFlush#_Toc15842094)

We focused on narrowing the data to the airlines with lowest median value of customer satisfaction.



Distribution of Promoters (Per day basis)

Average Detractors per day: **40**

Average Passive per day: **38**

Average Promoters per day: **37**

Distribution of Detractors (Per day basis)

Average Detractors per day: **40**

Divide the Passives:

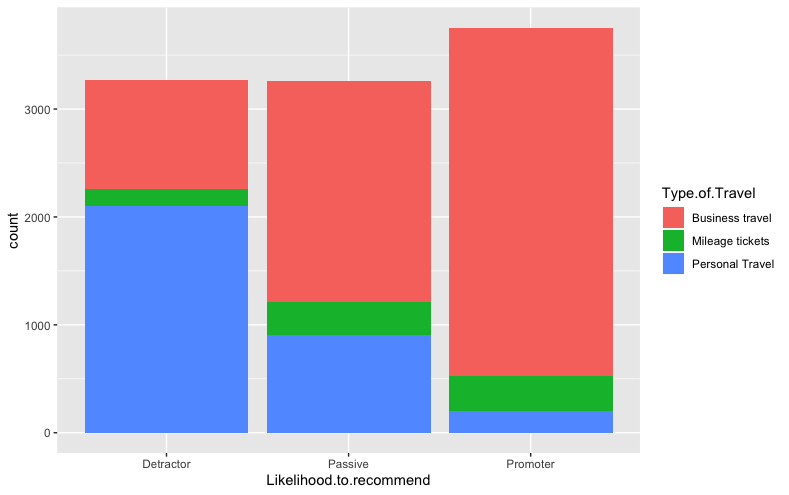
Avg. Promoters: 37+19 = **56**

Avg. Detractors: 40+19 = **59**

Percentage of Daily Detractors:

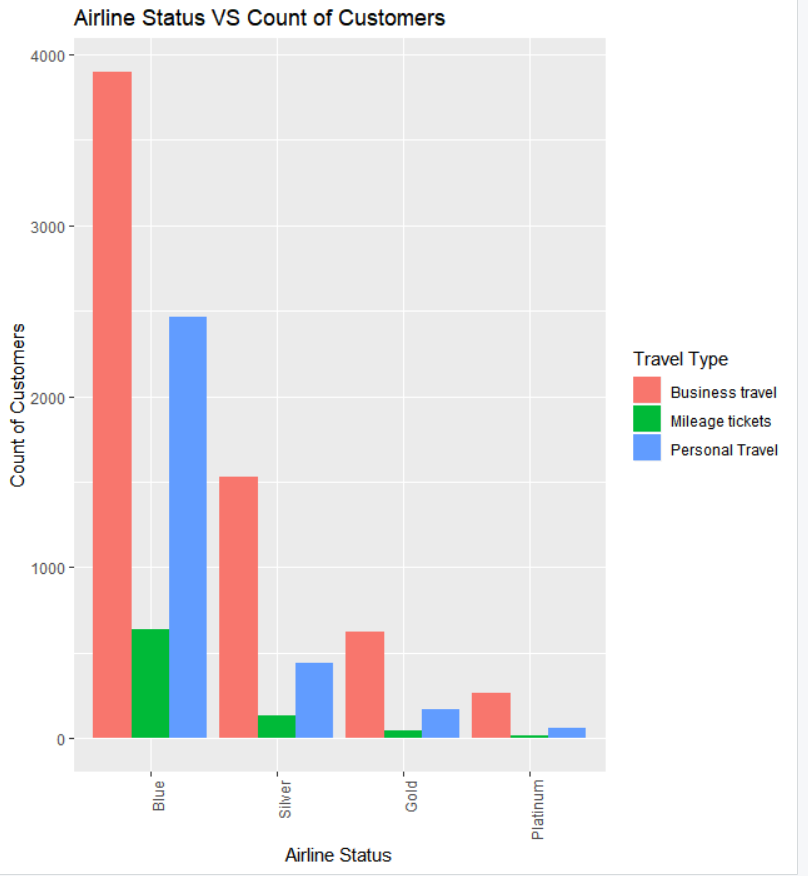
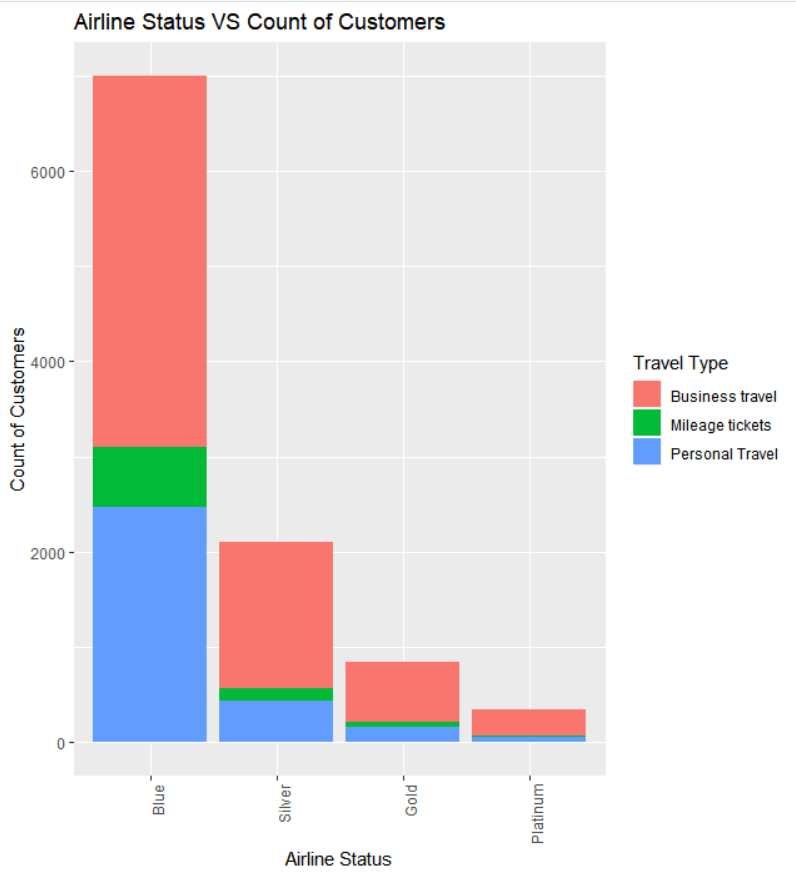
**59∕115 \* 100 = 51.30%**

On daily basis we have a 51.30% of customers who happen to fall under the detractor category. We need to develop BUSINESS SOLUTIONS to minimize this percentage.



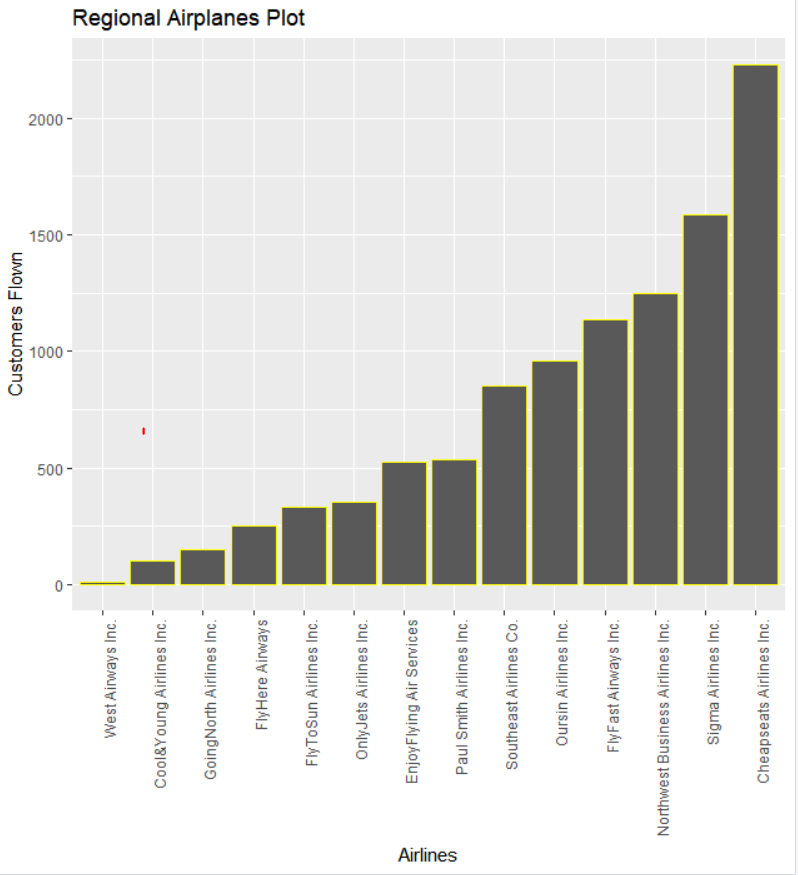
NPS VS Count of Customers

Here ,we plot a bar graph between NPS and the count of customers. What we observe is that most of the Promoters are in Business Class and most number of detractors are in Personal Travel.

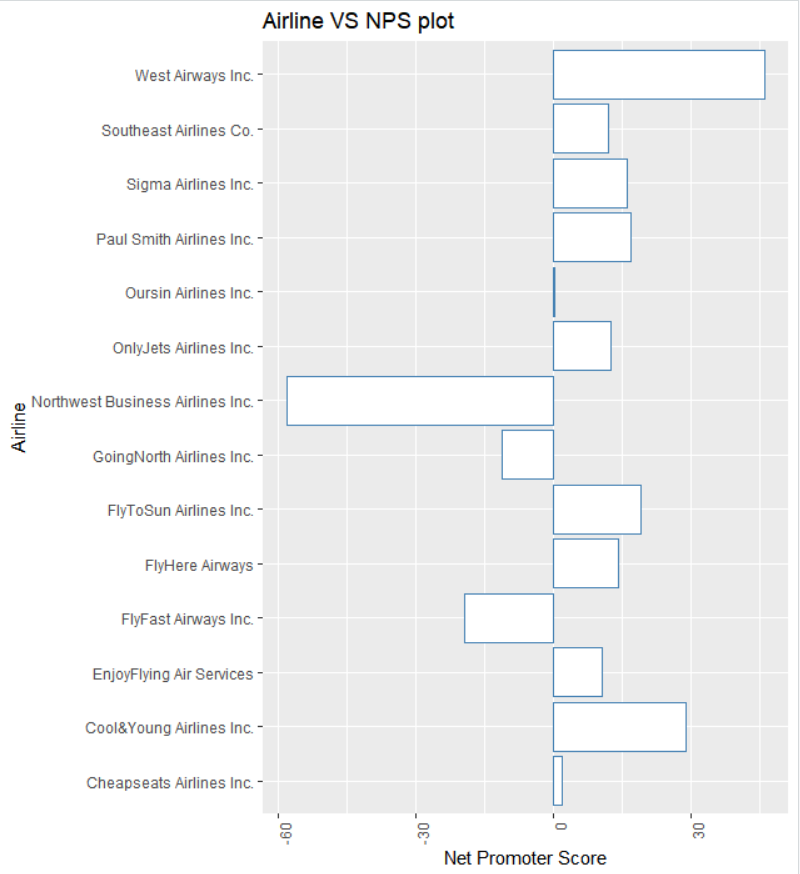
Airline Status VS Count of Customers

Here we plot a graph between the airline status and count of customers so we can find that most of the business travel customers fall into Blue airlines status. With respect to mileage travel and personal travel, business travel customers are more in each of the categories: Blue, Silver, Gold and Platinum.



Regional Airplanes Plot

Here we can see with each airline, how many customers have flown. What we observe is that most customers have flown with Cheapest Airlines Inc followed by Sigma Airlines Inc. The least number of customers flew from West Airways and Cool & Young Airlines Inc.



Airline VS NPS Plot

Here, we plot airline wise all the NPS scores and observe the following:

Best performing partner airlines:

1.West Airways

2.Cool & Young Airlines

3.Fly to Sun Airlines

Worst performing partner airlines:

1.Northwest Business Airline

2.Going North Airline

3.Fly Fast Airways

Percentage detractor per top three airline:

1.Cheap seats Airlines 32.70%

2.Sigma Airlines 26.83%

3.Northeast Business Airline 58.73%

[DATA MODELLING](https://word-edit.officeapps.live.com/we/wordeditorframe.aspx?ui=en%2DUS&rs=en%2DUS&wopisrc=https%3A%2F%2Fsumailsyr-my.sharepoint.com%2Fpersonal%2Frmupadhy_syr_edu%2F_vti_bin%2Fwopi.ashx%2Ffiles%2Ff6858e2a5cae4b83b55b2892796de412&wdenableroaming=1&mscc=1&wdodb=1&hid=2643209F-500D-A000-A1A4-7F6D652C96B9&wdorigin=Sharing&jsapi=1&newsession=1&corrid=74747185-6ef3-42eb-9eb3-f4fc8c1a61b5&usid=74747185-6ef3-42eb-9eb3-f4fc8c1a61b5&instantedit=1&wopicomplete=1&wdredirectionreason=Unified_SingleFlush#_Toc15842091)

1. Linear Models

**Age**

First of all, we check whether there are errors in the age column, which is less than 0. And also, if their first flight age is less than 0, then the data is error. So, we create 2 new columns to check the data set.

# Check whether all data in age is correct

New\_flight$First\_flight\_age <- New\_flight$Age - (2012 - New\_flight$First\_flight\_year)

# Create a new column that shows the first flight age of all passengers

auz <- New\_flight %>%

filter(First\_flight\_age <= 0)

auzn <- nrow(auz)

auzn

# Check if there's any customer's age is under 0

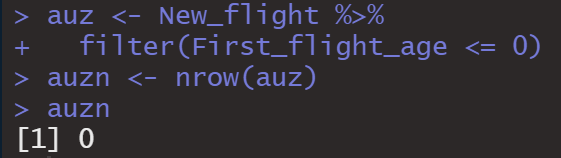
ffa <- New\_flight %>%

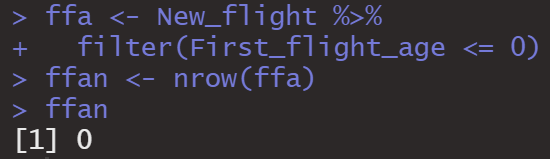
filter(First\_flight\_age <= 0)

ffan <- nrow(ffa)

ffan

# Check if there's any First\_flight\_age is under 0, which means a wrong data





The result shows that both auzn and ffan equal to 0, which means that there's no logical mistake with age data.

Our next step is group customers into different groups by their age, because age may cause different reasons for traveling. We create 3 different groups, which is “Low”means age under 35, “Medium” means age between 35 and 65, and “High” means age above 65.

# Group customers into different age range

for(i in 1:n)

if (New\_flight$Age[i] <= 35) {

New\_flight$Age\_Group[i]="Low"

} else if (New\_flight$Age[i] > 35 && New\_flight$Age[i] <= 65) {

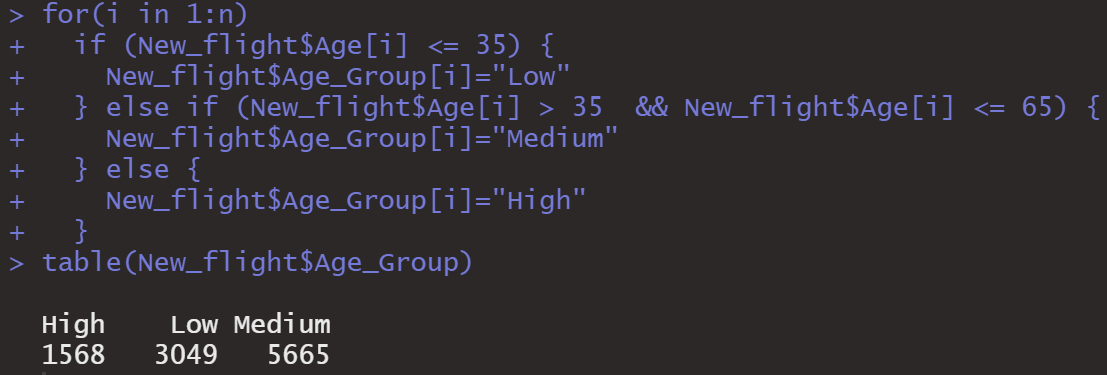
New\_flight$Age\_Group[i]="Medium"

} else {

New\_flight$Age\_Group[i]="High"

}

table(New\_flight$Age\_Group)



From the results above we can see that there are 1568 people who are older than 65, 3049 who are below 35 years old and 5665 people who are between 35 and 65 years old.

Then we separate the data frame by those three age groups we just created.

LAC <- New\_flight %>%

filter(Age\_Group == "Low")

MAC <- New\_flight %>%

filter(Age\_Group == "Medium")

HAC <- New\_flight %>%

filter(Age\_Group == "High")

Now since the new data frames are created, we can run the linear models. What we did is examine 3 groups with numeric columns using backward linear model method.

# Low age customers

LAC\_LR1 <- lm(Likelihood\_to\_recommend ~ Age + First\_flight\_age + Price\_sensitivity + loyalty + `F\_time(mins)` + F\_Distance + `Scheduled\_Departure(Hours)` + `Departure\_delay(mins)` + `Arrival\_delay(mins)` + F\_per\_year + `Freq\_flyer\_accts(total)` + Shopping\_at\_airport + Eat\_drink\_at\_airport, data = LAC)

summary(LAC\_LR1)

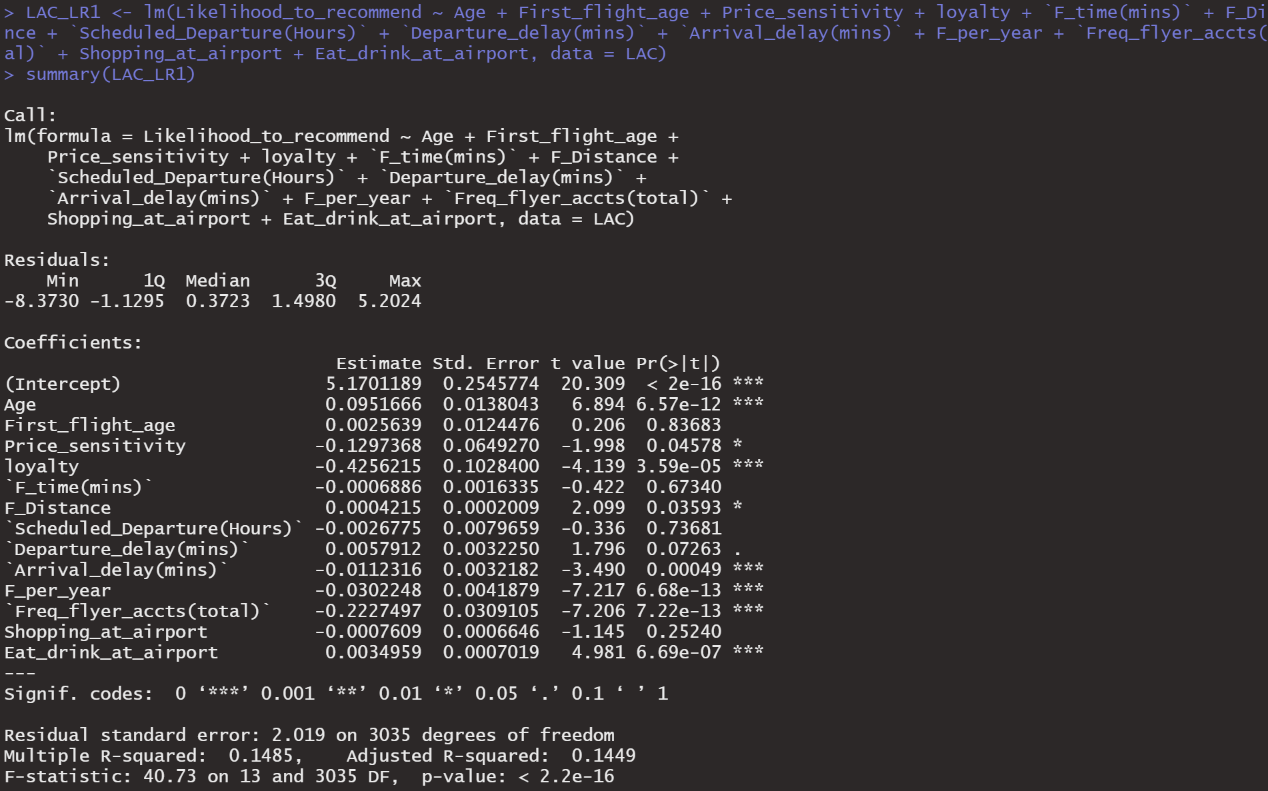
# Filter out columns not important

LAC\_LR2 <- lm(Likelihood\_to\_recommend ~ Age + Price\_sensitivity + loyalty + F\_Distance + `Arrival\_delay(mins)` + F\_per\_year + `Freq\_flyer\_accts(total)` + Eat\_drink\_at\_airport, data = LAC)

summary(LAC\_LR2)

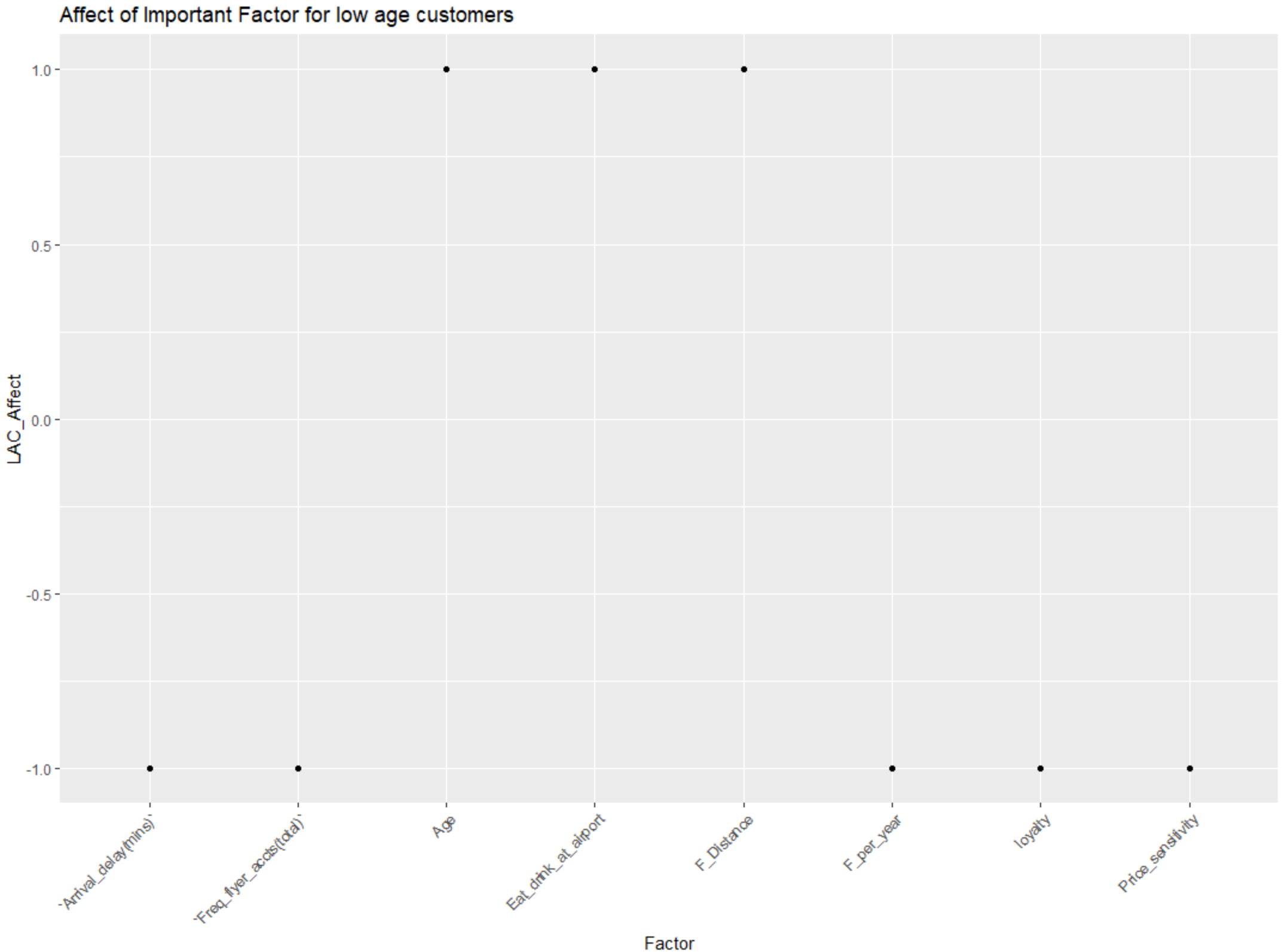
LAC\_Number\_of\_Important\_Fac <- length(LAC\_LR2$coefficients) - 1

LAC\_Number\_of\_Important\_Fac



From the results we can see that there are 8 variables whose p-value is less than 0.05, which means that they are statistically significant and they are factors that matters much to affect the likelihood of recommendation.

Since we got these factors, we would like to figure out whether these significant variables affects positively or negatively to likelihood to recommend. Because they are correlated to each other, we cannot judge from their coefficients.



LAC\_Affect <- sign(coefficients(LAC\_LR2))

# Delete the first row intercept coefficient

LAC\_Affect <- LAC\_Affect[-1]

Factor <- names(LAC\_Affect)

# Create a new data frame using two above list

LAC\_IF <- data.frame(Factor, LAC\_Affect)

# Create a plot so we can easily inspect the sign of affect each factor poses

ggplot(LAC\_IF) +

aes(x = Factor, y = LAC\_Affect) +

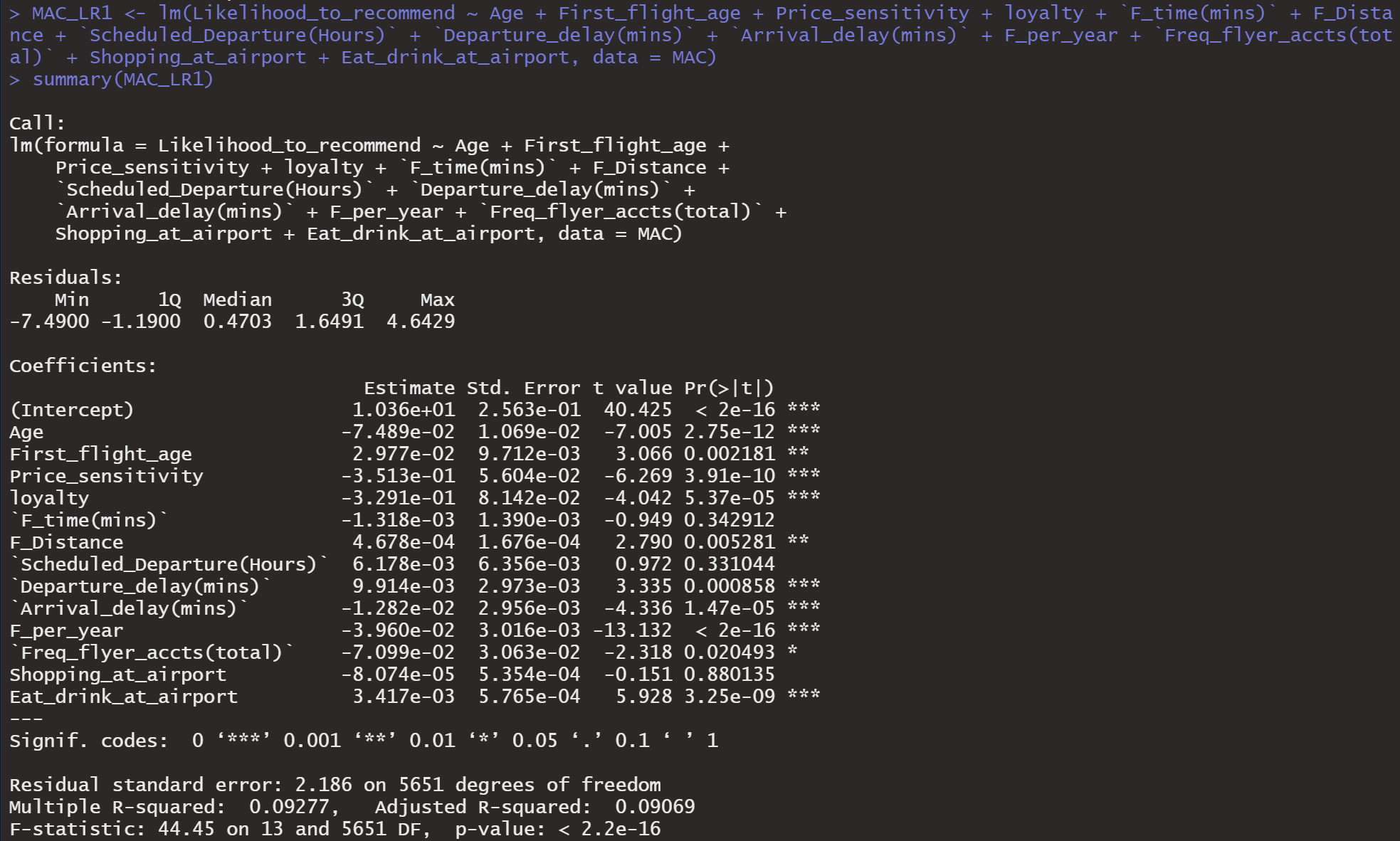
geom\_point() +

theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +

ggtitle("Affect of Important Factor for low age customers")

We can conclude from the plot that there are totally three factors that positively affects the likelihood, which are age, flight distance and eat and drink at the airport. So the first conclusion we can drive from this model is that for people who are younger than 35, if they are going on shorter flight distance and spend more money on the eat and drink at the airport, then they tend to be more satisfied and more likely to recommend to others.

We did the same thing for other two groups, the codes and plot are as follows:



# Medium age customers

MAC\_LR1 <- lm(Likelihood\_to\_recommend ~ Age + First\_flight\_age + Price\_sensitivity + loyalty + `F\_time(mins)` + F\_Distance + `Scheduled\_Departure(Hours)` + `Departure\_delay(mins)` + `Arrival\_delay(mins)` + F\_per\_year + `Freq\_flyer\_accts(total)` + Shopping\_at\_airport + Eat\_drink\_at\_airport, data = MAC)

summary(MAC\_LR1)

# Filter out columns not important

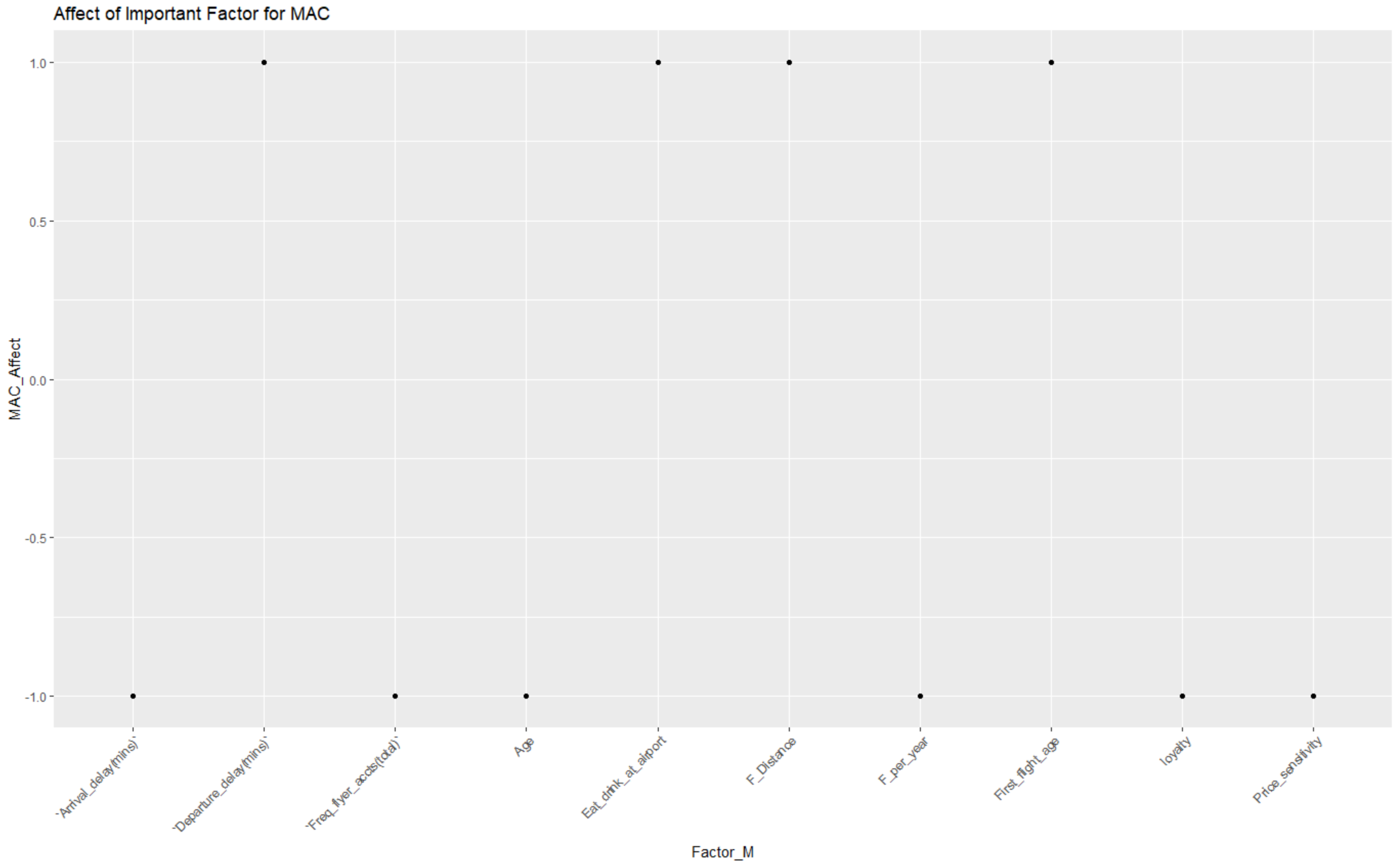
MAC\_LR2 <- lm(Likelihood\_to\_recommend ~ Age + First\_flight\_age + Price\_sensitivity + loyalty + F\_Distance + `Departure\_delay(mins)` + `Arrival\_delay(mins)` + F\_per\_year + `Freq\_flyer\_accts(total)` + Eat\_drink\_at\_airport, data = MAC)

summary(MAC\_LR2)

MAC\_Number\_of\_Important\_Fac <- length(MAC\_LR2$coefficients) - 1

MAC\_Number\_of\_Important\_Fac

The result shows that for middle age group, there are 10 factors that affect significantly to the likelihood to recommend.



# Medium age customers

MAC\_LR1 <- lm(Likelihood\_to\_recommend ~ Age + First\_flight\_age + Price\_sensitivity + loyalty + `F\_time(mins)` + F\_Distance + `Scheduled\_Departure(Hours)` + `Departure\_delay(mins)` + `Arrival\_delay(mins)` + F\_per\_year + `Freq\_flyer\_accts(total)` + Shopping\_at\_airport + Eat\_drink\_at\_airport, data = MAC)

summary(MAC\_LR1)

# Filter out columns not important

MAC\_LR2 <- lm(Likelihood\_to\_recommend ~ Age + First\_flight\_age + Price\_sensitivity + loyalty + F\_Distance + `Departure\_delay(mins)` + `Arrival\_delay(mins)` + F\_per\_year + `Freq\_flyer\_accts(total)` + Eat\_drink\_at\_airport, data = MAC)

summary(MAC\_LR2)

MAC\_Number\_of\_Important\_Fac <- length(MAC\_LR2$coefficients) - 1

MAC\_Number\_of\_Important\_Fac

# Find out whether these important factors affect positively or negatively to likelihood to recommend

MAC\_Affect <- sign(coefficients(MAC\_LR2))

# Delete the first row intercept coefficient

MAC\_Affect <- MAC\_Affect[-1]

Factor\_M <- names(MAC\_Affect)

# Create a new data frame using two above list

MAC\_IF <- data.frame(Factor\_M, MAC\_Affect)

# Create a plot so we can easily inspect the sign of affect each factor poses

ggplot(MAC\_IF) +

aes(x = Factor\_M, y = MAC\_Affect) +

geom\_point() +

theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +

ggtitle("Affect of Important Factor for MAC")

For middle age people, there are four factors that positively affect the likelihood, which are first flight age, flight distance, departure delay, and eat and drink at the airport.

Follows are the same kind of analysis for the elder:

# High age customers

HAC\_LR1 <- lm(Likelihood\_to\_recommend ~ Age + First\_flight\_age + Price\_sensitivity + loyalty + `F\_time(mins)` + F\_Distance + `Scheduled\_Departure(Hours)` + `Departure\_delay(mins)` + `Arrival\_delay(mins)` + F\_per\_year + `Freq\_flyer\_accts(total)` + Shopping\_at\_airport + Eat\_drink\_at\_airport, data = HAC)

summary(HAC\_LR1)

# Filter out columns not important

HAC\_LR2 <- lm(Likelihood\_to\_recommend ~ Age + First\_flight\_age + `Scheduled\_Departure(Hours)` + `Arrival\_delay(mins)` + F\_per\_year, data = HAC)

summary(HAC\_LR2)

#Filter out columns not important

HAC\_LR3 <- lm(Likelihood\_to\_recommend ~ Age + `Scheduled\_Departure(Hours)` + `Arrival\_delay(mins)` + F\_per\_year, data = HAC)

summary(HAC\_LR3)

HAC\_Number\_of\_Important\_Fac <- length(HAC\_LR3$coefficients) - 1

HAC\_Number\_of\_Important\_Fac

# Find out whether these important factors affect positively or negatively to likelihood to recommend

HAC\_Affect <- sign(coefficients(HAC\_LR3))

# Delete the first row intercept coefficient

HAC\_Affect <- HAC\_Affect[-1]

Factor\_H <- names(HAC\_Affect)

# Create a new data frame using two above list

HAC\_IF <- data.frame(Factor\_H, HAC\_Affect)

# Create a plot so we can easily inspect the sign of affect each factor poses

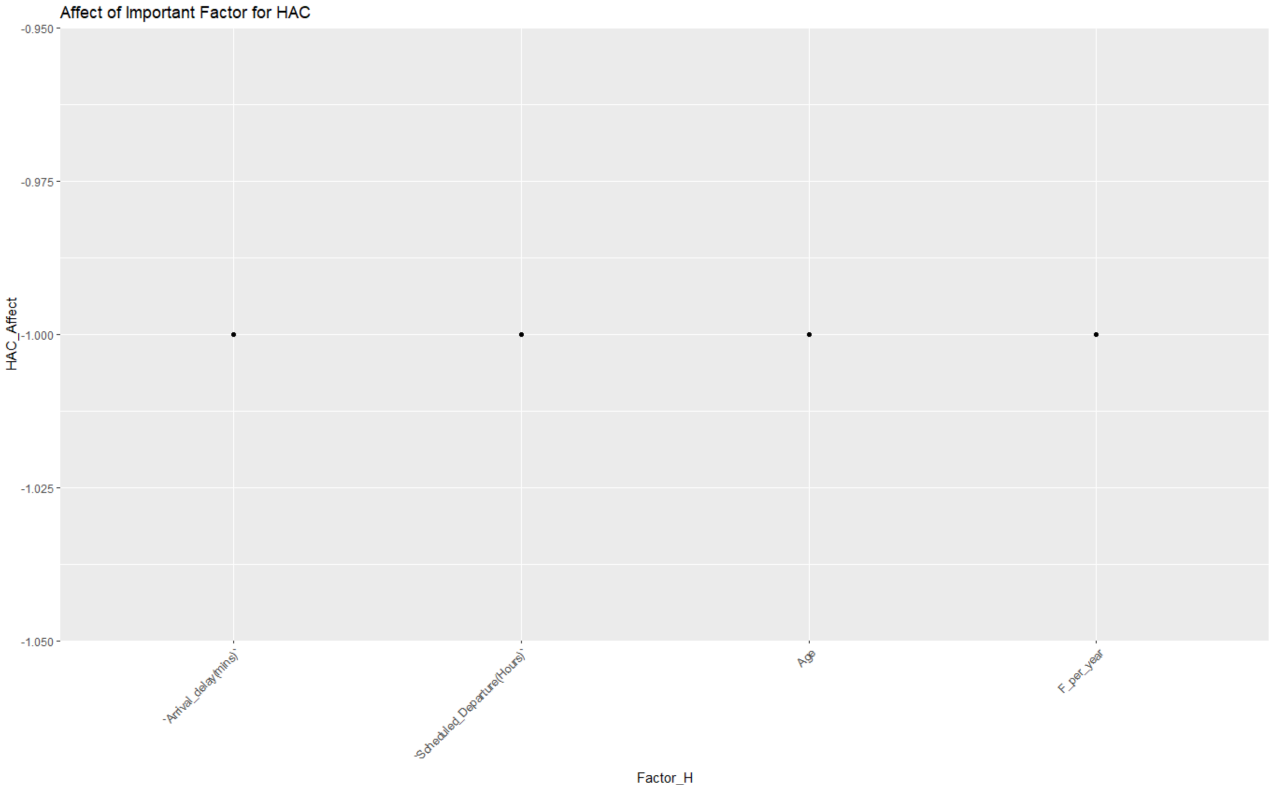
ggplot(HAC\_IF) +

aes(x = Factor\_H, y = HAC\_Affect) +

geom\_point() +

theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +

ggtitle("Affect of Important Factor for HAC")



We found out an interesting thing that older people are less likely to recommend flight to others.

Owing that different age groups have different number of customers, multiply the coefficient of linear models and sum the weighted values up using data frame

Factors <- c("Age", "First\_flight\_age", "Price\_sensitivity", "loyalty", "F\_time(mins)", "F\_Distance", "Scheduled\_Departure(Hours)", "Departure\_delay(mins)", "Arrival\_delay(mins)", "F\_per\_year", "Freq\_flyer\_accts(total)", "Shopping\_at\_airport", "Eat\_drink\_at\_airport")

Weighted\_value <- c(0,0,0,0,0,0,0,0,0,0,0,0,0)

FWV <- data.frame(Factors, Weighted\_value)

# Input weighted value of LAC into FWV

LAC\_A <- coefficients(LAC\_LR2)

LAC\_A <- LAC\_A[-1]

LAC\_WA <- table(New\_flight$Age\_Group)[2] \* LAC\_A

FWV[1,2] <- FWV[1,2] + LAC\_WA[1]

FWV[3,2] <- FWV[3,2] + LAC\_WA[2]

FWV[4,2] <- FWV[4,2] + LAC\_WA[3]

FWV[6,2] <- FWV[6,2] + LAC\_WA[4]

FWV[9,2] <- FWV[9,2] + LAC\_WA[5]

FWV[10,2] <- FWV[10,2] + LAC\_WA[6]

FWV[11,2] <- FWV[11,2] + LAC\_WA[7]

FWV[13,2] <- FWV[13,2] + LAC\_WA[8]

# Input weighted value of MAC into FWV

MAC\_A <- coefficients(MAC\_LR2)

MAC\_A <- MAC\_A[-1]

MAC\_WA <- table(New\_flight$Age\_Group)[3] \* MAC\_A

FWV[1,2] <- FWV[1,2] + MAC\_WA[1]

FWV[2,2] <- FWV[2,2] + MAC\_WA[2]

FWV[3,2] <- FWV[3,2] + MAC\_WA[3]

FWV[4,2] <- FWV[4,2] + MAC\_WA[4]

FWV[6,2] <- FWV[6,2] + MAC\_WA[5]

FWV[8,2] <- FWV[8,2] + MAC\_WA[6]

FWV[9,2] <- FWV[9,2] + MAC\_WA[7]

FWV[10,2] <- FWV[10,2] + MAC\_WA[8]

# Sort the data and output the top 5 weighted value of important factors based on coefficient and number of customers

FWV$Weighted\_value <- round(FWV$Weighted\_value, 3)

FWV <- FWV[order(abs(FWV$Weighted\_value), decreasing = TRUE),]

Strongest\_Overall\_Factors\_Affecting\_Age <- head(FWV, 5)

Strongest\_Overall\_Factors\_Affecting\_Age

rownames(Strongest\_Overall\_Factors\_Affecting\_Age) <- Strongest\_Overall\_Factors\_Affecting\_Age$Factors

Strongest\_Overall\_Factors\_Affecting\_Age <- Strongest\_Overall\_Factors\_Affecting\_Age[,-1]

barplot(Strongest\_Overall\_Factors\_Affecting\_Age)

FWV[11,2] <- FWV[11,2] + MAC\_WA[9]

FWV[13,2] <- FWV[13,2] + MAC\_WA[10]

# Input weighted value of HAC into FWV

HAC\_A <- coefficients(HAC\_LR3)

HAC\_A <- HAC\_A[-1]

HAC\_WA <- table(New\_flight$Age\_Group)[1] \* HAC\_A

FWV[1,2] <- FWV[1,2] + HAC\_WA[1]

FWV[7,2] <- FWV[7,2] + HAC\_WA[2]

FWV[9,2] <- FWV[9,2] + HAC\_WA[3]

FWV[10,2] <- FWV[10,2] + HAC\_WA[4]

rownames(Strongest\_Overall\_Factors\_Affecting\_Age) <- Strongest\_Overall\_Factors\_Affecting\_Age$Factors

Strongest\_Overall\_Factors\_Affecting\_Age <- Strongest\_Overall\_Factors\_Affecting\_Age[,-1]

barplot(Strongest\_Overall\_Factors\_Affecting\_Age)

FWV[11,2] <- FWV[11,2] + MAC\_WA[9]

FWV[13,2] <- FWV[13,2] + MAC\_WA[10]

# Input weighted value of HAC into FWV

HAC\_A <- coefficients(HAC\_LR3)

HAC\_A <- HAC\_A[-1]

HAC\_WA <- table(New\_flight$Age\_Group)[1] \* HAC\_A

FWV[1,2] <- FWV[1,2] + HAC\_WA[1]

FWV[7,2] <- FWV[7,2] + HAC\_WA[2]

FWV[9,2] <- FWV[9,2] + HAC\_WA[3]

FWV[10,2] <- FWV[10,2] + HAC\_WA[4]

By inspecting the top 5 important factors, we can tell that these strongest factors are all negative values. We conclude that In order to increase the likelihood to recommend value, airline can keep these factors low.

Gender

First, let’s get to know about the distribution of customers based on gender

table(New\_flight$Gender)

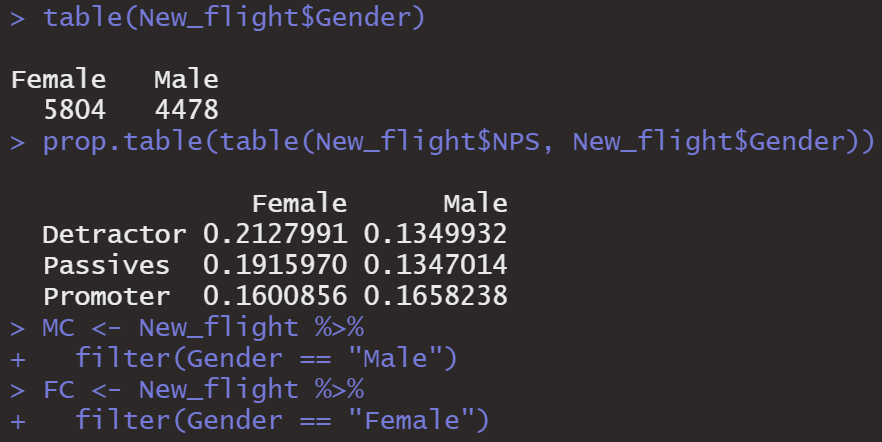
prop.table(table(New\_flight$NPS, New\_flight$Gender))

MC <- New\_flight %>%

filter(Gender == "Male")

FC <- New\_flight %>%

filter(Gender == "Female")



Then we examine gender by using backward linear model method as well.

# Male Customers

MC\_LR1 <- lm(Likelihood\_to\_recommend ~ Age + First\_flight\_age + Price\_sensitivity + loyalty + `F\_time(mins)` + F\_Distance + `Scheduled\_Departure(Hours)` + `Departure\_delay(mins)` + `Arrival\_delay(mins)` + F\_per\_year + `Freq\_flyer\_accts(total)` + Shopping\_at\_airport + Eat\_drink\_at\_airport, data = MC)

summary(MC\_LR1)

MC\_LR2 <- lm(Likelihood\_to\_recommend ~ Age + Price\_sensitivity + F\_Distance + `Arrival\_delay(mins)` + F\_per\_year + Shopping\_at\_airport + Eat\_drink\_at\_airport, data = MAC)

summary(MC\_LR2)

MC\_LR3 <- lm(Likelihood\_to\_recommend ~ Age + Price\_sensitivity + F\_Distance + `Arrival\_delay(mins)` + F\_per\_year + Eat\_drink\_at\_airport, data = MAC)

summary(MC\_LR3)

MC\_Number\_of\_Important\_Fac <- length(MC\_LR3$coefficients) - 1

MC\_Number\_of\_Important\_Fac

There are 6 important factors that enables us to predict likelihood of recommendation for male customers. We then find out whether these important factors affect positively or negatively to likelihood to recommend. And create a plot

of these factors.

MC\_Affect <- sign(coefficients(MC\_LR3))

# Delete the first row intercept coefficient

MC\_Affect <- MC\_Affect[-1]

Factor\_MC <- names(MC\_Affect)

# Create a new data frame using two above list

MC\_IF <- data.frame(Factor\_MC, MC\_Affect)

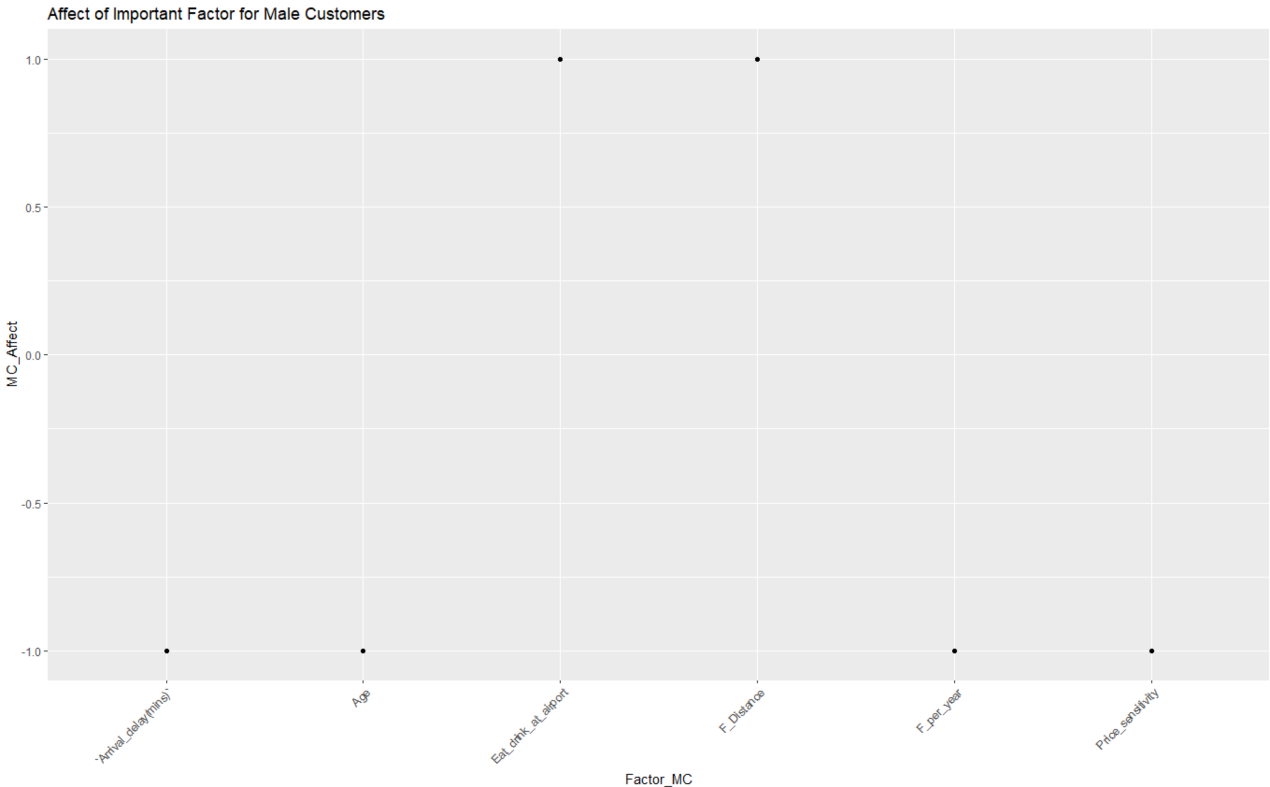
ggplot(MC\_IF) +

aes(x = Factor\_MC, y = MC\_Affect) +

geom\_point() +

theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +

ggtitle("Affect of Important Factor for Male Customers")



We can conclude that for male customers, eat and drink at the airport and flight distance drive positive effect on the likelihood to recommend.

Then we did the similar model for female customers.

FC\_LR1 <- lm(Likelihood\_to\_recommend ~ Age + First\_flight\_age + Price\_sensitivity + loyalty + `F\_time(mins)` + F\_Distance + `Scheduled\_Departure(Hours)` + `Departure\_delay(mins)` + `Arrival\_delay(mins)` + F\_per\_year + `Freq\_flyer\_accts(total)` + Shopping\_at\_airport + Eat\_drink\_at\_airport, data = FC)

summary(FC\_LR1)

FC\_LR2 <- lm(Likelihood\_to\_recommend ~ Age + Price\_sensitivity + F\_Distance + `Departure\_delay(mins)` + `Arrival\_delay(mins)` + F\_per\_year + `Freq\_flyer\_accts(total)` + Eat\_drink\_at\_airport, data = FC)

summary(FC\_LR2)

FC\_Number\_of\_Important\_Fac <- length(FC\_LR2$coefficients) - 1

FC\_Number\_of\_Important\_Fac

There are 8 important factors that enables us to predict likelihood of recommend for female customers. Then we draw a plot to show how these factors affect the likelihood.

FC\_Affect <- sign(coefficients(FC\_LR2))

# Delete the first row intercept coefficient

FC\_Affect <- FC\_Affect[-1]

Factor\_FC <- names(FC\_Affect)

# Create a new data frame using two above list

FC\_IF <- data.frame(Factor\_FC, FC\_Affect)

# Create a plot so we can easily inspect the sign of affect each factor poses

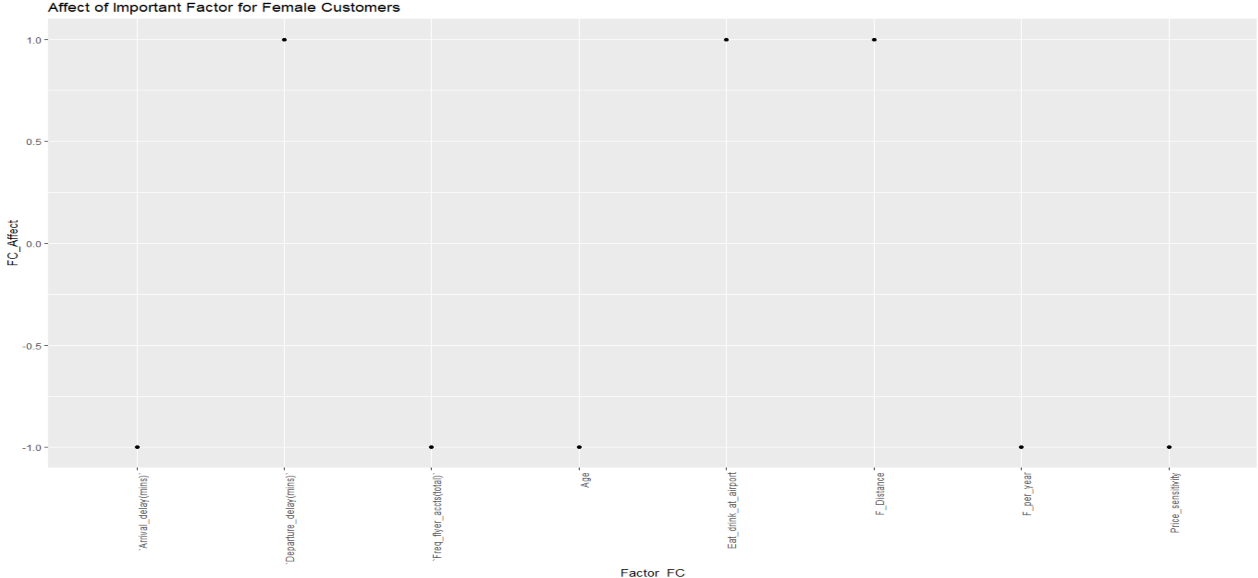
ggplot(FC\_IF) +

aes(x = Factor\_FC, y = FC\_Affect) +

geom\_point() +

theme(axis.text.x = element\_text(angle = 90, hjust = 1)) +

ggtitle("Affect of Important Factor for Female Customers")



We conclude that for female customers, eat and drink at the airport, flight distance and departure delay time are those that positively affects the likelihood.

Because different gender has different number of customers, we multiply the coefficient of linear models and sum the weighted values up using data frame

Factors <- c("Age", "First\_flight\_age", "Price\_sensitivity", "loyalty", "F\_time(mins)", "F\_Distance", "Scheduled\_Departure(Hours)", "Departure\_delay(mins)", "Arrival\_delay(mins)", "F\_per\_year", "Freq\_flyer\_accts(total)", "Shopping\_at\_airport", "Eat\_drink\_at\_airport")

Weighted\_value\_gender <- c(0,0,0,0,0,0,0,0,0,0,0,0,0)

GF <- data.frame(Factors, Weighted\_value\_gender)

# Input weighted value of male customers into GF

MCA <- coefficients(MC\_LR3)

MCA <- MCA[-1]

MCAA <- table(New\_flight$Gender)[2] \* MCA

GF[1,2] <- GF[1,2] + MCAA[1]

GF[3,2] <- GF[3,2] + MCAA[2]

GF[6,2] <- GF[6,2] + MCAA[3]

GF[9,2] <- GF[9,2] + MCAA[4]

GF[10,2] <- GF[10,2] + MCAA[5]

GF[13,2] <- GF[13,2] + MCAA[6]

# Input weighted value of male customers into GF

FCA <- coefficients(FC\_LR2)

FCA <- FCA[-1]

FCAA <- table(New\_flight$Gender)[1] \* FCA

GF[1,2] <- GF[1,2] + FCAA[1]

GF[3,2] <- GF[3,2] + FCAA[2]

GF[6,2] <- GF[6,2] + FCAA[3]

GF[8,2] <- GF[8,2] + FCAA[4]

GF[9,2] <- GF[9,2] + FCAA[5]

GF[10,2] <- GF[10,2] + FCAA[6]

GF[11,2] <- GF[11,2] + FCAA[7]

GF[13,2] <- GF[13,2] + FCAA[8]

# Sort the data and output the top 5 weighted value of important factors based on coefficient and number of customers

GF$Weighted\_value\_gender <- round(GF$Weighted\_value\_gender, 3)

GF <- GF[order(abs(GF$Weighted\_value\_gender), decreasing = TRUE),]

Strongest\_Overall\_Factors\_Affecting\_Gender <- head(GF, 5)

barplot(Strongest\_Overall\_Factors\_Affecting\_Gender)

Strongest\_Overall\_Factors\_Affecting\_Gender

# By inspecting the top 5 important factors, we can tell that these strongest factors are all negative values.

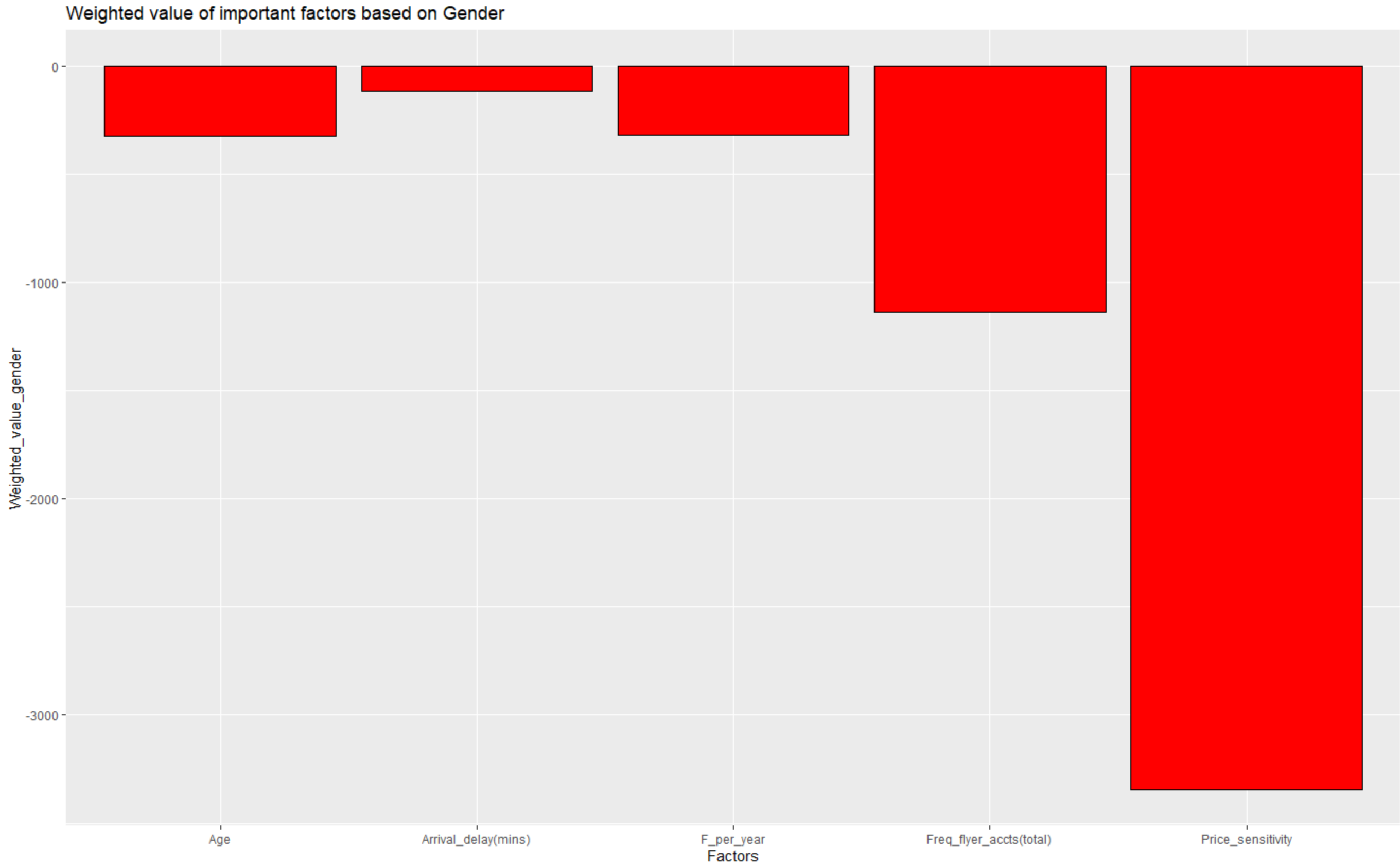
# In order to increase the likelihood to recommend value, airline can keep these factors low.

myPlot2 <- ggplot(Strongest\_Overall\_Factors\_Affecting\_Gender, aes(x = Factors, y = Weighted\_value\_gender))

myPlot2 <- myPlot2 + geom\_col(color = "black", fill = "red")

myPlot2 <- myPlot2 + ggtitle("Weighted value of important factors based on Gender")

myPlot2



myPlot2 <- ggplot(Strongest\_Overall\_Factors\_Affecting\_Gender, aes(x = Factors, y = Weighted\_value\_gender))

myPlot2 <- myPlot2 + geom\_col(color = "black", fill = "red")

myPlot2 <- myPlot2 + ggtitle("Weighted value of important factors based on Gender")

myPlot2

By inspecting the top 5 important factors, we can tell that these strongest factors are all negative values. In order to increase the likelihood to recommend value, airline can keep these factors low.

1. Association Rules

We used association rule mining which is a technique that we used to find associations between the NPS values and the other variables.

library(arules) #Invoke Library

library(arulesViz) #Invoke Library

df$ShoppingAirportCategorical <- ifelse(df$Shopping\_at\_airport>0,'Yes','No')

#Converting numerical variable to categorical variable

df$EatingAirportCategorical <- ifelse(df$Eat\_drink\_at\_airport>0,'Yes','No')

#Converting numerical variable to categorical variable

df$LoyaltyCategorical <- ifelse(df$loyalty>0,'Loyal','Not Loyal')

#Converting numerical variable to categorical variable

df$DelayFlightCategorical <- ifelse(df$`Departure\_delay(mins)`>0,'Delayed','No Delay')

#Converting numerical variable to categorical variable

df$ArrivalDelayCategorical <- ifelse(df$`Arrival\_delay(mins)`>0,'Arr Delayed', 'Arr No Delay')

#Converting numerical variable to categorical variable

df$PriceSensitivityCategorical2 <- ifelse(df$Price\_sensitivity<2,'Not Sensitive','Sensitive')

#Converting numerical variable to categorical variable

n <-nrow(df)

#Store number of rows in the data set

for(i in 1:n) #Converting numerical variable to categorical variable

if (df$Price\_sensitivity[i]>=0 && df$Price\_sensitivity[i]<=2) {

df$PriceSensitivityCategorical[i]="Not Sensitive"

} else if (df$Price\_sensitivity[i]>2 && df$Price\_sensitivity[i]<=3) {

df$PriceSensitivityCategorical[i]="Less Sensitive"

} else {

df$PriceSensitivityCategorical[i]="Sensitive"

}

View(df) #Display data frame

For implementing Apriori algorithm we had to convert the numerical variables into ordinal categorical variables. We’ve done the same in the code mentioned above.

Then we created a new dataframe named ‘Dfcat’, which has only categorical variables stored in it. And created a transactional matrix.

dfcat<-df[,c('Partner\_Name','Gender','Type\_of\_travel','Airline\_status','Class\_of\_travel','ShoppingAirportCategorical','EatingAirportCategorical','PriceSensitivityCategorical','PriceSensitivityCategorical2','LoyaltyCategorical','NPS')]

#Creating a new data frame consisting of categorical variables

View(dfcat)

#Display the Data set

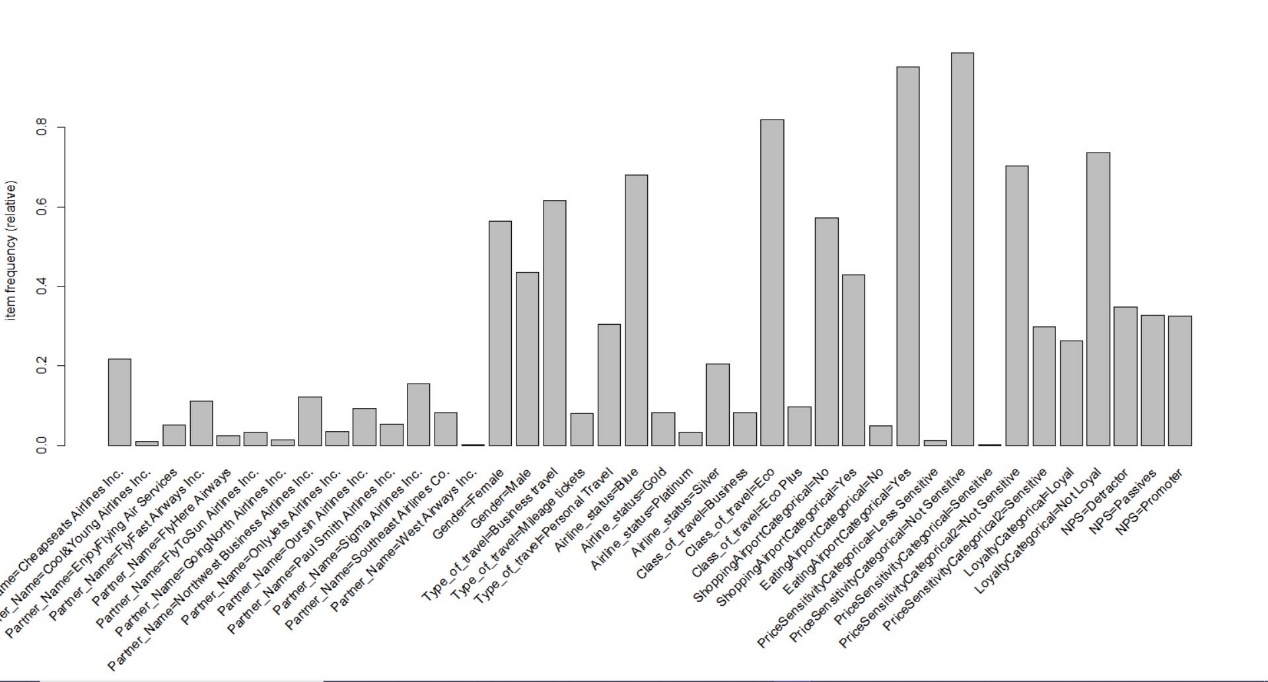
dfcatX <- as(dfcat,"transactions") #Creating a transactions matrix for Apriori algorithm

inspect(dfcatX)

#Inspect the Transactional Matrix

itemFrequencyPlot(dfcatX)

#Plots the item frequency

Item Frequency Plot

.A screenshot of a cell phone

Description automatically generated

#Promoter Analysis

rulesetPromoter <- apriori(dfcatX,

parameter=list(support=0.015,confidence=0.60),

appearance = list(default="lhs", rhs=("NPS=Promoter"))) #Applying apriori algorithm by keeping Promoter on RHS

summary(rulesetPromoter) #Summary of Ruleset

inspectDT(rulesetPromoter) #Interactive ruleset interface

Promoter association rules

The rule set in the figure above helps us infer the variables associated with NPS=” Promoter”.

The variables are:

1. Partner Name is Sigma Airlines
2. Gender is Male
3. Type of Travel is Business Travel
4. Price sensitivity is Not Price Sensitive
5. Eats food at the airport

#Detractor Analysis

rulesetDetractor <- apriori(dfcatX,

parameter=list(support=0.015,confidence=0.60),

appearance = list(default="lhs", rhs=("NPS=Detractor"))) #Applying apriori algorithm by keeping Detractor on RHS

summary(rulesetDetractor) #Summary of RuleSet

inspectDT(rulesetDetractor) #Interactive ruleset interface

A screenshot of a cell phone

Description automatically generated

Detractor Association Rules

The rule set in the figure above helps us infer the variables associated with NPS=”Detractor”.

The variables are:

1. Partner Name is Northwest Business Airline
2. Gender is Female
3. Type of Travel is Personal
4. Class of Travel is Eco
5. Not Price Sensitive
6. Loyalty Category is not Loyal

A picture containing window, drawing

Description automatically generated

As we can see from the above diagram that most of the detractors are from Northwest Business Airlines, so we analyzed the detractors of the airlines to get a clearer reason behind customer being a detractor.

#North West Business Airlines Apriori

dfBA <- dfcat[dfcat$Partner\_Name == 'Northwest Business Airlines Inc.',]

#Creating a data frame consisting only values of Northwest Business Airline Inc.

View(dfBA) #Display the dataframe

dfBAX <- as(dfBA,"transactions") #Creating a transactional matrix

rulesetNBADetractor <- apriori(dfBAX,

parameter=list(support=0.015,confidence=0.60),

appearance = list(default="lhs", rhs=("NPS=Detractor"))) #Applying apriori algorithm by keeping Detractor on RHS

inspectDT(rulesetNBADetractor) #Interactive ruleset interface

A screenshot of a cell phone

Description automatically generated

Detractor Association Rule of Northwest Business Airline

The ruleset in the figure above helps us infer that in Northwest Business Airline the variables associated with detractors are:

1. Type of travel is Personal Travel
2. Airline Status is Blue
3. Class of Travel is EcoPlus
4. Loyalty category is Not Loyal

If we combine both the analysis of detractors of whole data set and detractors of NorthWest Airlines, we see that **Personal**, **Blue** and **Not loyal** categories are major contributors behind being detractors.

To dig in more, we created a data frame of blue airline status and personal type of travel and applied Apriori algorithm to get the reason behind a customer becoming a detractor and also analyzed the promoters in this data set which will help us get the reasons for a person in this category being a promoter.

#Blue and Personal VS Promoter and detractors

View(dfcat)

blueAS <- dfcat %>%

filter(Airline\_status=="Blue")

#Creating a data frame of only blue Airline Status

blueAS <- blueAS %>%

filter(Type\_of\_travel== "Personal Travel")

#Creating a data frame of Blue airline status and Personal Travel

View(blueAS)

dfblueAS <- as(blueAS,"transactions") #Creating a transactional matrix

rulesetDStatB <- apriori(dfblueAS,

parameter=list(support=0.015,confidence=0.60),

appearance = list(default="lhs", rhs=("NPS=Detractor")))

#Applying apriori algorithm by keeping Detractor on RHS

summary(rulesetDStatB) #Summary of the ruleset

inspectDT(rulesetDStatB) #Interactive ruleset interface

A screenshot of a cell phone

Description automatically generated

Detractor Association Rule of Personal and Blue Categories

Here we can again observe that the reason of being a detractor is **Northwest Business Airlines, Airline Status is blue**. One interesting observation we get over here is that a **Loyal customer** is also a detractor for this data set.

rulesetPStatB <- apriori(dfblueAS,

parameter=list(support=0.001,confidence=0.1),

appearance = list(default="lhs", rhs=("NPS=Promoter")))

#Applying apriori algorithm by keeping Promoter on RHS

summary(rulesetPStatB) #Summary of the ruleset

inspectDT(rulesetPStatB) #Interactive ruleset interface

A screenshot of a cell phone

Description automatically generated

Promoter Association Rule of Personal and Blue Categories

For a very low value of support and confidence we get the Promoter association for this data set. We can observe that the promoters in this category fly in the **FlyHere Airways,** travel in the **Business Class.**

1. Decision Tree

colSums(is.na(NumMatrix))

install.packages("rpart")

install.packages("rpart.plot")

library(rpart)

library(rpart.plot)

tree = rpart(formula=LTR ~ F\_time + Price\_Sensitivity + F\_per\_year + Loyalty + Departure\_delay +Age, data=trainData)

We tried multiple values for creating a decision tree and finally it gave as an optimized result.

tree = rpart(formula=LTR ~ F\_time + Price\_SensitivitY + Loyalty, data=trainData)

We tried using 3 values and creating a decision tree from these values but we didn't get an appropriate output to interpret so we dropped the idea.

prp(tree)

A close up of a map

Description automatically generated

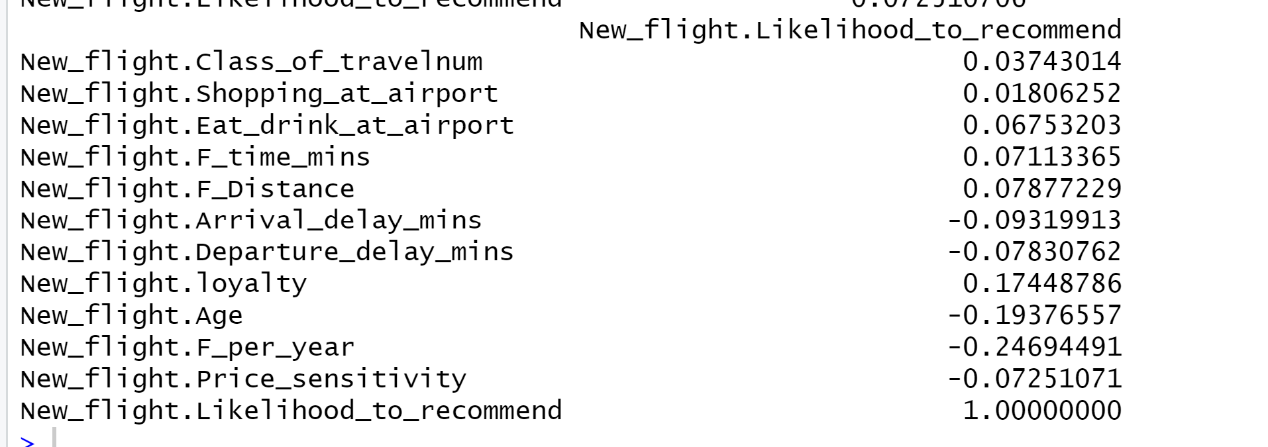
Decision Tree

So, we created a decision tree using the rpart() function which gives us an optimized result of all the variables we have provided to the model. This decision tree can be interpreted in the following way:

If age>=63, then we parse towards the left hand side of the tree i.e. towards the “Yes” side, further it checks that the departure delay is greater than equal to 5.4 minutes or not. Again if it is, then the member is likely to give a recommendation score of 4 otherwise, the member would give a score of 7.

Now, what if age>=63 is not satisfied I.e. it parses towards the right side of the tree towards “No” and it further checks if the age<23 then the customer is likely to recommend a score of 7 and if it is greater than 23 i.e. (the age is between 23 and 63 then the customer is likely to give a recommendation score of 9.

1. Correlation Matrix



Here, we calculate the correlation matrix w.r.t Likelihood to recommend to observe which variables are associated with the likelihood to recommend and accordingly, we use all these values for implementing the models.

5. Ordinal Logistic Regression

install.packages("MASS")

library(MASS)#we need this library to perform the ordinal logistic regression

m <- polr(as.factor(NPSnum) ~ Price\_sensitivity+Class\_of\_travelnum, data = New\_flight, Hess=TRUE)

#It calculates the ordinal logistic regression for which we get all the p-values

#significant so we can interpret that they are good linear predictors.

summary(m)

(ctable <- coef(summary(m)))

p <- pnorm(abs(ctable[,"t value"]), lower.tail = FALSE) \* 2 #this helps to

#find the p-value

(ctable <- cbind(ctable, "p value" = p)) #this adds another column p value

#to the original result

A screenshot of a cell phone

Description automatically generated

Output of Ordinal Logistic Regression

Since NPS has 3 different values(0,1,2) and in logistic regression, the dependent variable can only have values between 0 and 1 so we had to use the ordinal logistic regression in this case. What we found out is that the variable Price sensitivity and Class of travel both have p-values less than 0.05 which confirms from our previous models that they are good predictors of the model.

ACTIONABLE INSIGHTS

1. Business travellers who are male customers are more likely to recommend Southeast Airlines. Advertisements and marketing strategies can be developed focusing the same category. As Southeast airlines have already served them better and a word of mouth from them to personal travellers will be great boost for the detractors becoming a promoter.
2. Try to promote blue class to silver class as the silver class customers are more likely to be a promoter. We can provide some of the silver class services to the blue class customers for a limited time so that we can make the blue class customers a promoter.
3. Northwest Business Airline Travellers has been failing in comparison to other partners as most of the customers didn’t have a pleasant experience with them. So, the contract with them should not be extended and should be transferred to the airlines with positive NPS score like the West Airways, Cool & Young Airlines and Fly to Sun Airlines.
4. Age based price discrimination may help decrease the number of detractors for 18-24 and 63+ years age groups.
5. A limited time of access to the lounge of the airline should be given to the personal type of travellers as discounts to them won’t help as they are not sensitive to the pricing.
6. A ‘first impression’ price discount will also help decrease detractors who are Blue class flyers.
7. Improve service for customers in California state as the number of detractors over there are more.
8. Eco Plus seems to be an effective offering to increase the likelihood to recommend should be promoted more in the advertisements.