School of Information Studies
SYRACUSE UNIVERSITY

IST 687 – Applied Data Science

Lab Section M004 | Group 2

Data Analysis for Airline Satisfaction Survey

Recommendations to Improve Customer Satisfaction

=]

Submitted by:

Madhu Chheda | Madhavi Kadam | Ruifeng Chen | Shuming Song | Yash Kapadia

TABLE OF CONTENT

INTRODUCTION

As a mean transportation in our life, airplane could save our time during the trip. The flights are not blocked by mountains, rivers, deserts and oceans. With the development of economy, there is a growing concern over the feedback from the customers about its satisfaction. It will affect the customer's choice about which flight company which they are going to pick in the fierce competition.

At this time, a survey could directly collect the real feedback from customer, to evaluating their flight, measuring their loyalty, figure out their preference. So in this report, we are going to analyse the satisfaction survey about their flight experience.

A. PROJECT APPROACH

Satisfaction Survey is a data set containing data of 14 different airlines with 129889 different observations based on 28 variables. All the variables are related to the satisfaction feedback of the customers.

In our analysis of the Satisfaction Survey, we have identified and reviewed the key performance indicators like flight cancellations, flight delays, flight path etc. This gave us an insight on the feedback provided by the customers and helped us analysis it using different data analysis models.

We created correlational trends between various variables of the data set and customer satisfaction using the analysis models and this helped us answer the business questions formulated for the data set. On answering the business questions we could infer better insights and solutions for increasing the customer satisfaction and airline services.

Moreover, we use system modeling in Math and Correlation Matrix to prepare our data before we do the modeling part. In this report, we use the Association Rule Mining and Linear Modeling to analysis our data, and another two models, which is Support Vector Machine and Random Forest Model to validate our result.

BUSINESS QUESTIONS

Feedback from every customer provides information that lead to formulation of actionable insights. Analysing this feedback would help airlines improve their business practices and services, which in turn would result in higher revenues.

Following are the business questions that have been identified and answered through project:

1. Defining Aspect:

- What are the characteristics of customers with low satisfaction and high satisfaction?
- What attributes influence the likelihood of customers recommending the airline services and providing it a good feedback?

2. Analysis Aspect:

- How the possible attributes affect the customer's satisfaction individually? Is it negatively or positively relevant to the customers' satisfaction?
- How can detractors be converted into promoters based on satisfaction rate?

3. Recommendation Aspect:

- What are the facilities and services provided by airlines that need to be focused on to increase overall satisfaction rate for airlines?
- How can the airline company predict the customer's satisfaction systematically?
- Standing in the perspective of airline company, what corrective actions can be done to increase the customer's satisfaction?

DATA REQUISITION AND PREPROCESSING

The first step of our project is to clear up the data that we got. Doing the basic analysis and cleaning the data could help us find the questions and directions for our project. Facing to this new data set, some of the information may not helpful for our analysis, so in this part, we will load the data, clean it and munging it.

A. DATA REQUISITION

Data was made available to us by course instructors. Initially, we downloaded the data and it consisted of approximately 129889 different observations based on 28 variables. This data was extensively studied to determine usable variables. After initial analysis, the data set was forwarded to the pre-processing phase of the project for data munging.

Following is the code for Data Requisition:

```
dataset <- read.csv(file = "Satisfaction Survey.csv",na.strings = "",sep =",")
View(dataset)
str(dataset)
summary(dataset)
```

B. DATA PREPROCESSING

Before processing the data further we had to clean the data. As many columns like departure delay in minutes, arrival delay in minutes, flight time in minutes etc. contained a lot of NA values. In data cleaning we omitted the NA values by replacing it with the mean of the entire column, which could have the least effect on the following analysing process such as linear modeling. Only the columns relevant to the analysis of the data were kept and the remaining were eliminated. Columns such as Origin State and Destination state can be found out by the Origin city and Destination city columns. Thus, eliminating the State columns makes sense. Also, some columns such as Flight distance can be calculated using other attributes so we have eliminated it. The flight date is also not relevant for our analysis purposes. Hence, in the data pre-processing part the attributes which were not important for the analysis have been eliminated and the dataset is narrowed to limited number of columns ~ 21 which can be used for the analysis and visualization process. The white spaces in the columns have been trimmed so that it is easier to use those columns in the models for further analysis. Besides, since it is made as an agreement in the group discussion that the customer whose flight is cancelled are likely to give an emotional and biased comment about their satisfaction, we decided to focus our research area in the customer who successfully take the flight.

In summary, the data pre-processing phase provides a subset of usable data for the project, consisting of only the columns that need to be worked with and eliminating Null/NA values.
Data munging was also performed for transforming and mapping raw data into usable format for the project, enabling easy data analytics on it.

Following is the code for Data Cleaning:

```
#converting all integer columns to numeric
dataset$Age <- as.numeric(dataset$Age)
dataset$Flight.time.in.minutes <- as.numeric(dataset$Flight.time.in.minutes)
dataset$Day.of.Month <- as.numeric(dataset$Day.of.Month)
dataset$Flight.Distance <- as.numeric(dataset$Flight.Distance)
dataset[5:8] <- lapply(dataset[5:8],as.numeric)
dataset[10:12] <- lapply(dataset[10:12],as.numeric)
dataset[22:24] <- lapply(dataset[10:12],as.numeric)

#checking for any rows which are not complete
sum(!complete.cases(dataset))

ncol(dataset)
nrow(dataset)

#Taking only the data whose flights are not cancelled
Airdata <- dataset[dataset$Flight.cancelled=="No",]
nrow(Airdata)
str(Airdata)

#Airdata <- data.frame(lapply(Airdata, trimws), stringsAsFactors = FALSE)
#str(Airdata)

#checking for any rows which are not complete for the new dataset
sum(!complete.cases(Airdata))
sum(is.na(Airdata$Arrival.Delay.in.Minutes))
sum(is.na(Airdata$Flight.time.in.minutes))
sum(is.na(Airdata$Departure.Delay.in.Minutes))

is.null(Airdata)
View(Airdata)
nrow(Airdata)
nrow(dataset)
str(Airdata)

#converting all integer columns to numeric
Airdata$Age <- as.numeric(Airdata$Age)
Airdata$Flight.time.in.minutes <- as.numeric(Airdata$Flight.time.in.minutes)
Airdata$Day.of.Month <- as.numeric(Airdata$Day.of.Month)
Airdata$Flight.Distance <- as.numeric(Airdata$Flight.Distance)
Airdata[5:8] <- lapply(Airdata[5:8],as.numeric)
Airdata[10:12] <- lapply(Airdata[10:12],as.numeric)
Airdata[22:24] <- lapply(Airdata[10:12],as.numeric)
View(Airdata)

#replacing na values with mean
for(i in 1:ncol(Airdata)){
  Airdata[is.na(Airdata[,i]), i] <- mean(Airdata[,i], na.rm = TRUE)
}
```

```
#Validating that the data is cleaned
sum(!complete.cases(Airdata))

is.null(Airdata)
View(Airdata)
nrow(Airdata)
nrow(dataset)
str(Airdata)

#removing unwanted columns from the Survey dataset
Airdata <- Airdata[-c(6,14,19,21,25,27,28)]
str(Airdata)
summary(Airdata)
View(Airdata)

#Converting the Satisfaction Column to numeric - Run this line only when required
Airdata$Satisfaction <- as.numeric(as.character(Airdata$Satisfaction))
str(Airdata)

#trimming the Airline name column
Airdata$Airline.Name <- trimws(Airdata$Airline.Name)
Airdata$Airline.Name <- as.factor(Airdata$Airline.Name)
```

USE OF DESCRIPTIVE STATISTICS AND DATA VISUALIZATION

In this part, we are trying to find the relationship between the factors by using GGMAP, and draw the airline information on the map. We also used histogram and density graphs in GGPLOT2 to compare the relationship between two factors that we picked.

A. GGMAP

```
load("R-Project.RData")
Sati_flight_1 <- na.omit(Sati_flight)
install.packages("ggmap")
install.packages("geosphere")
install.packages("jpeg")
install.packages("grid")
install.packages("plyr")
library(ggmap)
library(geosphere)
library(jpeg)
library(grid)
library(plyr)
library(pacman)
p_load(tidyverse, data.table, geosphere, grid, jpeg, plyr,dbplyr)

#Download the map
download.file("https://www.nasa.gov/specials/blackmarble/2016/globalmaps/BlackMarble_2016_01deg.jpg",destfile = "BlackMarble_2016_01deg.jpg", mode = "wb")

#render the map
earth <- readJPEG("BlackMarble_2016_01deg.jpg", native = TRUE)
earth <- rasterGrob(earth, interpolate = TRUE)

# Get the point of airport
Allcity <- factor(c(Sati_flight_1$Orgin.City,Sati_flight_1$Destination.City))
Allevel <- levels(Allcity)

latlon_Org <- geocode(source = "dsk",Allevel)
View(latlon_Org)

latlon_Org <- data.frame("City Name"=Allevel,"Lat"=latlon_Org[1],"Long"=latlon_Org[2])

match_Org <- match(Sati_flight_1$Orgin.City,latlon_Org$City.Name)
match_Des <- match(Sati_flight_1$Destination.City,latlon_Org$City.Name)

Flightpath <- data.frame(Sati_flight_1$Satisfaction,Sati_flight_1$Airline.Name,
                         Sati_flight_1$Orgin.City,"Orgin_lat"=latlon_Org$lat[match_Org],"Orgin_lon"=latlon_Org$lon[match_Org],
                         Sati_flight_1$Destination.City,"Dest_lat"=latlon_Org$lat[match_Des],"Dest_lon"=latlon_Org$lon[match_Des])

Flightpath_all <-
as.matrix(data.frame(Flightpath$Orgin.lat,Flightpath$Orgin.lon,Flightpath$Dest.lat,Flightpath$Dest.lon))
```

```
Flightpath_all <- t(Flightpath_all)
Flightpath_all <- matrix(as.vector(Flightpath_all),ncol = 2,byrow = T)
group <-
as.vector(t(as.matrix(data.frame(1:length(Flightpath$Dest.lat),1:length(Flightpath$Dest.lat)
))))

Flightpath_all <- cbind(group,data.frame(Flightpath_all))
colnames(Flightpath_all) <- c("long","lat","id")
str(Flightpath_all)

flightpath_split <- split(Flightpath, Flightpath$Sati_flight_1.Airline.Name)

flights_all <- lapply(flightpath_split, function(x) gcIntermediate(x[, c("Orgin_lon",
"Orgin_lat")], x[, c("Dest_lon", "Dest_lat")], n=100, breakAtDateLine = FALSE,
addStartEnd = TRUE, sp = TRUE))

# convert it into dataframe
flightpath_fortified <- lapply(flights_all, function(x) ldply(x@lines, fortify))

# Unsplit lists
flightpath_fortified <- do.call("rbind", flightpath_fortified)

# Add and clean column with airline names
flightpath_fortified$name <- rownames(flightpath_fortified)
flightpath_fortified$name <- gsub("\\..*", "", flightpath_fortified$name)

levels(factor(flightpath_fortified$name))

# Extract first and last observations for plotting source and destination points (i.e.,
airports)
flightpath_points <- flightpath_fortified %>%
  group_by(group) %>%
  filter(row_number() == 1 | row_number() == n())
#learn it from the stackoverflow
#Calculate intermediate points between each two locations

Fmap <- ggplot() +
       annotation_custom(earth, xmin = -180, xmax = 180, ymin = -90, ymax = 90) +
       geom_path(aes(long, lat, group = id, color = name), alpha = 0.0, size = 0.0, data =
flightpath_fortified) +
       geom_path(aes(long, lat, group = id), alpha = 0.2, size = 0.3, color = "#f9ba00", data
= flightpath_fortified[flightpath_fortified$name=="West Airways Inc",]) +
       geom_path(aes(long, lat, group = id), alpha = 0.2, size = 0.3, color = "#ff0000", data
= flightpath_fortified[flightpath_fortified$name == "Southeast Airlines Co", ]) +
       geom_path(aes(long, lat, group = id), alpha = 0.2, size = 0.3, color = "#075aaa", data
= flightpath_fortified[flightpath_fortified$name == "Northwest Business Airlines Inc", ]) +
       geom_point(data = flightpath_points, aes(long, lat), alpha = 0.8, size = 0.1, colour =
"white") +
      theme(panel.background = element_rect(fill = "#05050f", colour =
"#05050f"),panel.grid.major = element_blank(),panel.grid.minor =
element_blank(),axis.title = element_blank(),axis.text = element_blank(),axis.ticks.length =
unit(0, "cm"),legend.position = "none") +
        annotate("text", x = -150, y = 13, hjust = 0, size = 14,label = paste("Northwest
Business Airlines Inc"), color = "#075aaa") +
```

```
        annotate("text", x = -150, y = 10, hjust = 0, size = 14,label = paste("West Airways
Inc"), color = "#f9ba00") +
        annotate("text", x = -150, y = 7, hjust = 0, size = 14,label = paste("Southeast Airlines
Co"), color = "#ff0000") +
        annotate("text", x = -150, y = 4, hjust = 0, size = 8,
          label = paste("Flight routes"), color = "white") +
  annotate("text", x = -150, y = 2, hjust = 0, size = 7,
          label = paste("Ruifeng Chen || NASA.gov || IST687 Final Project"), color = "white",
alpha = 0.5) +
         coord_equal()
Fmap

# Divide the flight by Satisfaction
Flightpath$Sati_flight_1.Satisfaction <- ceiling(Flightpath$Sati_flight_1.Satisfaction)
hist(Flightpath$Sati_flight_1.Satisfaction)
index <-
sample(c(TRUE,FALSE),length(Flightpath$Sati_flight_1.Satisfaction),replace=T,prob=c(0.
2,0.8))
Flightpath_sample <- Flightpath[index,]

Sampleit <- function(df,n){
     a <- n/nrow(df)
     index01 <- sample(c(T,F),nrow(df),replace = T, prob = c(a,1-a))
     df1 <- df[index01,]
     return(df1)
}

num <- nrow(Flightpath_sample[Flightpath_sample$Sati_flight_1.Satisfaction=="1",])
Flightpath1_sample <-
rbind(Flightpath_sample[Flightpath_sample$Sati_flight_1.Satisfaction=="1",],
                    Sampleit(Flightpath_sample[Flightpath_sample$Sati_flight_1.Satisfaction
=="2",],num),
                      Sampleit(Flightpath_sample[Flightpath_sample$Sati_flight_1.Satisfactio
n=="3",],num),
                      Sampleit(Flightpath_sample[Flightpath_sample$Sati_flight_1.Satisfactio
n=="4",],num),
                      Sampleit(Flightpath_sample[Flightpath_sample$Sati_flight_1.Satisfactio
n=="5",],num) )

flightpath1_split <- split(Flightpath1_sample,
Flightpath1_sample$Sati_flight_1.Satisfaction)
flights1_all <- lapply(flightpath1_split, function(x) gcIntermediate(x[, c("Orgin_lon",
"Orgin_lat")], x[, c("Dest_lon", "Dest_lat")], n=50, breakAtDateLine = FALSE, addStartEnd
= TRUE, sp = TRUE))

flightpath1_fortified <- lapply(flights1_all, function(x) ldply(x@lines, fortify))
flightpath1_fortified <- do.call("rbind", flightpath1_fortified)

flightpath1_fortified$satisfaction <- rownames(flightpath1_fortified)
flightpath1_fortified$satisfaction <- gsub("\\..*", "", flightpath1_fortified$satisfaction)

flightpath1_points <- flightpath1_fortified %>%
  group_by(group) %>%
  filter(row_number() == 1 | row_number() == n())
```

```
#example of one plot about satisfaction: satisfaction==5
ggplot() +
	annotation_custom(earth, xmin = -180, xmax = 180, ymin = -90, ymax = 90) +
	geom_path(aes(long, lat, group = id, color = satisfaction), alpha = 0, size = 0, data = flightpath1_fortified) +
	geom_path(aes(long, lat, group = id), alpha = 0.2, size = 0.3, color = "#dedcee", data = flightpath1_fortified[flightpath1_fortified$satisfaction=="5",]) +
	geom_point(data = flightpath1_points, aes(long, lat), alpha = 0.8, size = 0.1, colour = "white") +
	theme(panel.background = element_rect(fill = "#05050f", colour = "#05050f"),panel.grid.major = element_blank(),panel.grid.minor = element_blank(),axis.title = element_blank(),axis.text = element_blank(),axis.ticks.length = unit(0, "cm"),legend.position = "none") +
	annotate("text", x = -150, y = 7, hjust = 0, size = 14,label = paste("Satisfaction=5"), color = "#2bde73") +
	annotate("text", x = -150, y = 4, hjust = 0, size = 8,
	  label = paste("Flight routes vs Satisfaction"), color = "white") +
	annotate("text", x = -150, y = 2, hjust = 0, size = 7,
	  label = paste("Ruifeng Chen || NASA.gov || IST687 Final Project"), color = "white", alpha = 0.5) +
	coord_equal()


#Airpot vs satisfaction
Airport_Origin <-
data.frame(tapply(Flightpath$Sati_flight_1.Satisfaction,Flightpath$Sati_flight_1.Orgin.City, mean))
colnames(Airport_Origin) <- "SatisAsOri"
Airport_Origin$Origin_city <- rownames(Airport_Origin)
#View(Airport_Origin)

Airport_Dest <-
data.frame(tapply(Flightpath$Sati_flight_1.Satisfaction,Flightpath$Sati_flight_1.Destination.City,mean))
colnames(Airport_Dest) <- "SatisAsDest"
Airport_Dest$Dest_City <- rownames(Airport_Dest)
#View(Airport_Dest)
Airport_Dest <- Airport_Dest[-211,]
# the 211st city only appears as an destination city.

Airport_Popularity02 <- data.frame(table(Flightpath$Sati_flight_1.Destination.City))
Airport_Popularity01 <- data.frame(table(Flightpath$Sati_flight_1.Orgin.City))
Airport_Popularity02 <- Airport_Popularity02[-211,]

match01_Org <- match(Airport_Origin$Origin_city,latlon_Org$City.Name)
match01_Des <- match(Airport_Origin$Origin_city,Airport_Dest$Dest_City)
Airport_Origin[209:212,]
Airport_Dest[209:212,]
Airport_Popularity02[209:212,]

Airport_Sat <- data.frame("City"=Airport_Origin$Origin_city,
```

```
      "long"=latlon_Org$lon[match01_Org], "lat"=latlon_Org$lat[match01_Org],
      "SatisfactionAsOrigin"=Airport_Origin$SatisAsOri,
      "SatisfactionAsDest"=Airport_Dest$SatisAsDest[match01_Des])
View(Airport_Sat)

Airport_Sat$RankAsOrigin <- rank(Airport_Sat$SatisfactionAsOrigin-MinAsO)
Airport_Sat$RankAsDest <- rank(Airport_Sat$SatisfactionAsDest-MinAsD)
Airport_Sat$Comparison <- Airport_Sat$RankAsOrigin>Airport_Sat$RankAsDest
Airport_Sat$Popularity <- (Airport_Popularity01$Freq+Airport_Popularity02$Freq)

#hist(Airport_Sat$SatisfactionAsOrigin)

ggplot() +
  annotation_custom(earth, xmin = -180, xmax = 180, ymin = -90, ymax = 90) +
       geom_path(aes(long, lat, group = id, color = satisfaction), alpha = 0, size = 0, data =
flightpath1_fortified) +
       geom_point(data = Airport_Sat[Airport_Sat$Comparison,], aes(long, lat,
size=Popularity), alpha = 0.8,color="#ED5485") +
       geom_point(data = Airport_Sat[!Airport_Sat$Comparison,], aes(long, lat,
size=Popularity), alpha = 0.8,color="#f9ba00") +
       theme(panel.background = element_rect(fill = "#05050f", colour =
"#05050f"),panel.grid.major = element_blank(),panel.grid.minor =
element_blank(),axis.title = element_blank(),axis.text = element_blank(),axis.ticks.length =
unit(0, "cm"),legend.position = "none") +
       annotate("text", x = -150, y = 4, hjust = 0, size = 7,label = paste("Airport in
Destination City"), color = "#dedcee") +
       annotate("text", x = -150, y = 2, hjust = 0, size = 4,
         label = paste("Flight Airport vs Satisfaction"), color = "white") +
       annotate("text", x = -150, y = 1, hjust = 0, size = 3.5,
         label = paste("Ruifeng Chen || NASA.gov || IST687 Final Project"), color = "white",
alpha = 0.5) +
       coord_equal()

MaxQ <- quantile(Airport_Sat$Popularity,probs = 0.95)
ggplot(data=Airport_Sat[Airport_Sat$Popularity
<   MaxQ,],aes(x=SatisfactionAsOrigin,y=SatisfactionAsDest))+
       geom_point(aes(size=Popularity,color=Popularity),alpha=0.7)+
       geom_abline(color="red")+geom_abline(slope = 1,color="blue") +coord_equal()
```

We got the following maps after running the code. The first map with pink points shows which airport has better satisfaction on arrival in the United States. The second map with orange points tell us the airport which has the better satisfaction on departure. The third map tells us the difference both in the popularity of the airport (size of the point) as well as whether the customer leaving the airport have higher satisfaction (pink) or the customer arriving at the airport have higher satisfaction (orange). The fourth chart presents the routines of flight where the customer with satisfaction equalling to 4 takes. The fifth chart displays the flight routines of three different airline companies.

However, the difference between airports doesn't possess a significant difference. Therefore, we didn't insist on the exploration of the relationship between satisfaction with geographical information.

Fig1. Plot of Flight Airport vs Satisfaction in the Destination City

Fig 2. Plot of Flight Airport vs Satisfaction in the Original City

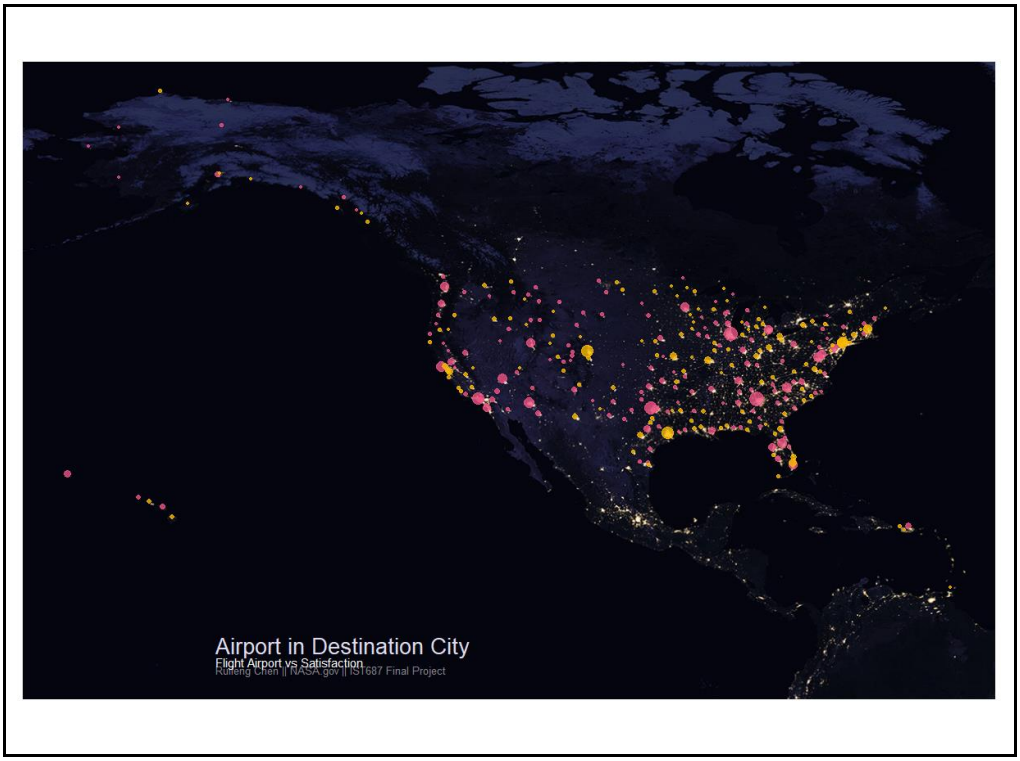


Fig 3. Plot of Satisfaction of Origin city vs Satisfaction of Destination City
(Pink means the city have higher Satisfaction as origin city, while the orange means the city have higher Satisfaction as destination city)

Fig 4. Plot of Flight Routes vs Satisfaction for Satisfaction = 5

Fig.5 Plot of Flight Routes vs three airline companies

B. GGPPLOT2

```
install.packages("ggplot2")
library(ggplot2)

Sati_flight_1 <- na.omit(Sati_flight)
Sati_flight_1$Satisfaction <- factor(ceiling(Sati_flight_1$Satisfaction))

Sati_flight_1 <- Airdata


# Negative factors:
# Satisfaction and Age
Sati_Age <- ggplot(Sati_flight_1,aes(x=Age))
Sati_Age <- Sati_Age + geom_histogram(aes(fill=Satisfaction),position = "dodge")
Sati_Age <- Sati_Age + ggtitle("Satisfaction versus Age")
Sati_Age
```

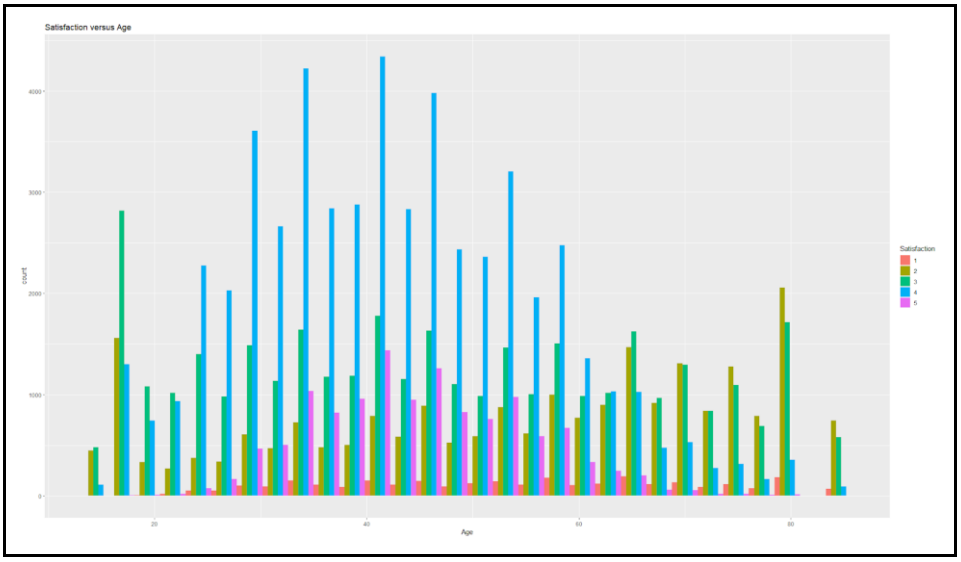


Fig 6. Plot of Satisfaction vs Age

```
Sati_Age_1 <- ggplot(Sati_flight_1,aes(x=Age)) +
geom_density(aes(fill=Satisfaction),position = "fill")
Sati_Age_1
```

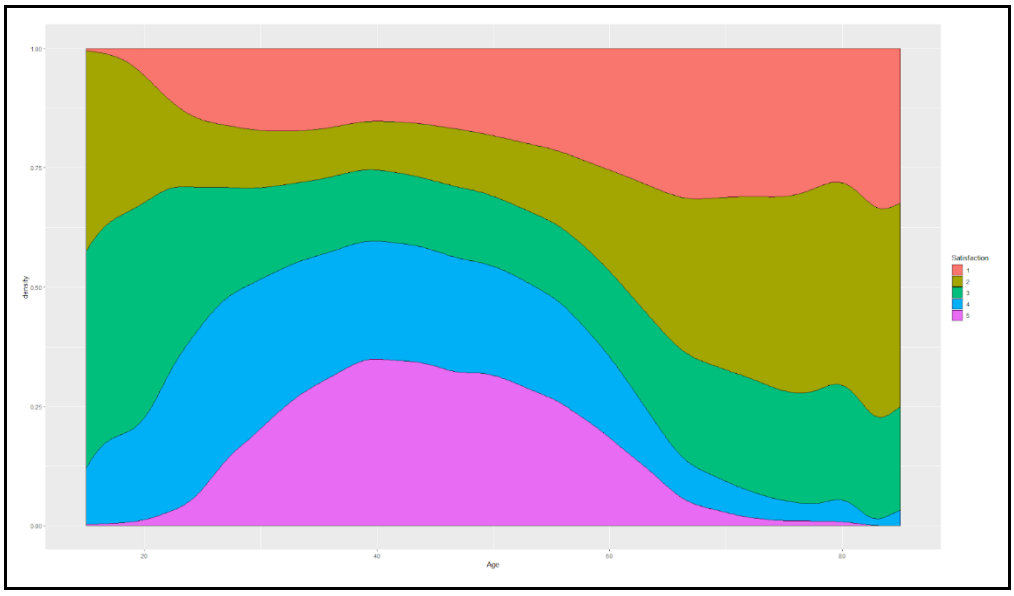


Fig 7. Heat Map of Age

From the graphs of satisfaction and age, we can see that the age around 40 will provide a high score on satisfaction. And the satisfaction decreases as the age increases.

```
# Satisfaction and No of Flights p. a.
Sati_NoFl <- ggplot(Sati_flight_1,aes(x=No.of.Flights.p.a.))
```

```
Sati_NoFl <- Sati_NoFl + geom_histogram(aes(fill=Satisfaction),position = "dodge")
Sati_NoFl <- Sati_NoFl + ggtitle("Satisfaction versus No of Flights")
Sati_NoFl
```

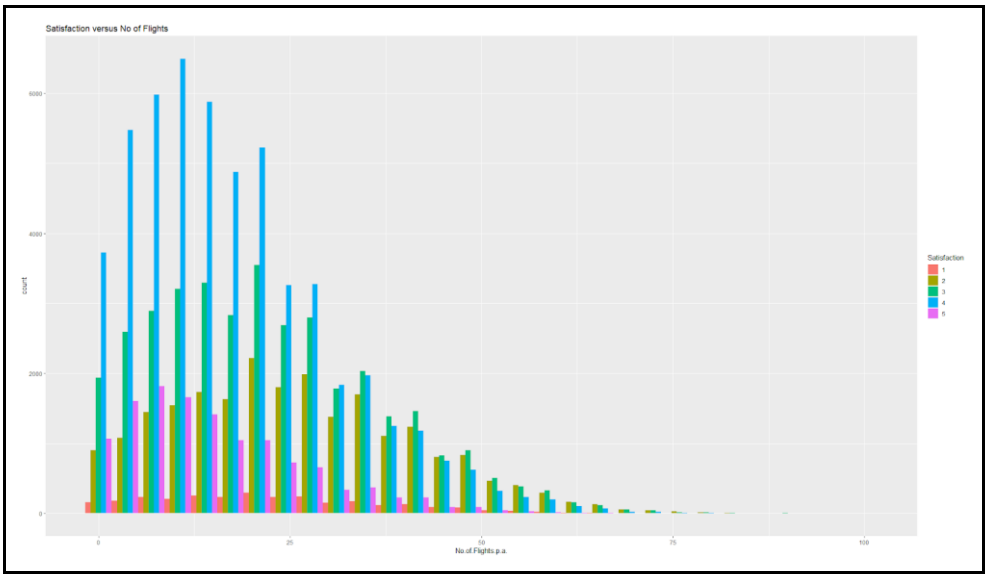


Fig 8. Plot of Satisfaction vs Number of Flights

```
Sati_NoFl_1 <- ggplot(Sati_flight_1,aes(x=No.of.Flights.p.a.)) +
geom_density(aes(fill=Satisfaction),position = "fill")
Sati_NoFl_1
```

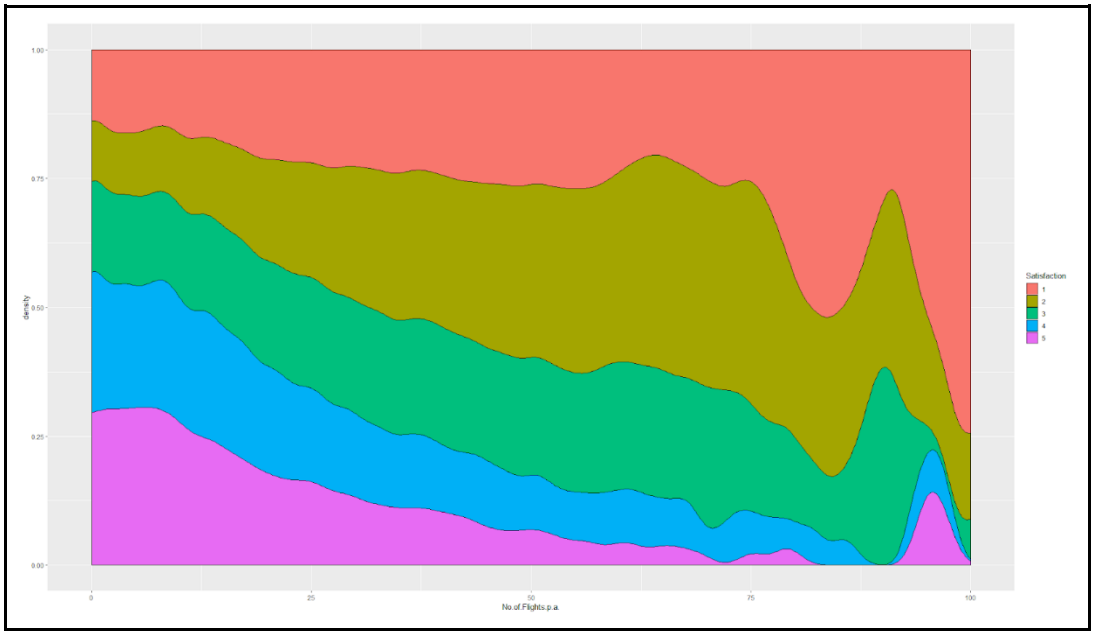


Fig 9. Heat Map of Number of Flights

No of flights p.a means the number of flights that each member has been taken per annum. From these two graphs, we can see that better feedback appeared when the number is less than 30.

```
# Satisfaction and Departure delay
Dpdy96 <- quantile(Sati_flight_1$Departure.Delay.in.Minutes,0.96)Sati_Dpdy <- Sati_Dpdy +
geom_histogram(aes(fill=Satisfaction),position = "dodge")
Sati_Dpdy <- Sati_Dpdy + ggtitle("Satisfaction versus Departure Delay")
Sati_Dpdy
Sati_Dpdy <-
ggplot(Sati_flight_1[Sati_flight_1$Departure.Delay.in.Minutes<Dpdy96,],aes(x=Departure.Delay.in.Minutes))
```

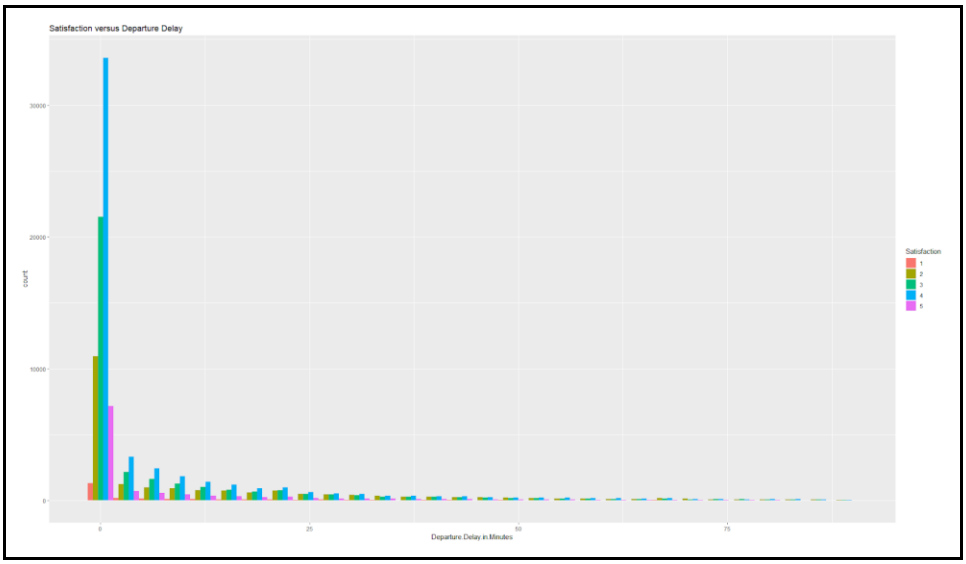


Fig 10. Plot of Satisfaction vs Departure Delay

```
Sati_Dpdy_1 <-
ggplot(Sati_flight_1[Sati_flight_1$Departure.Delay.in.Minutes<Dpdy96,],aes(x=Departure.Delay.in.Minutes))
+ geom_density(aes(fill=Satisfaction),position = "fill")
Sati_Dpdy_1
```

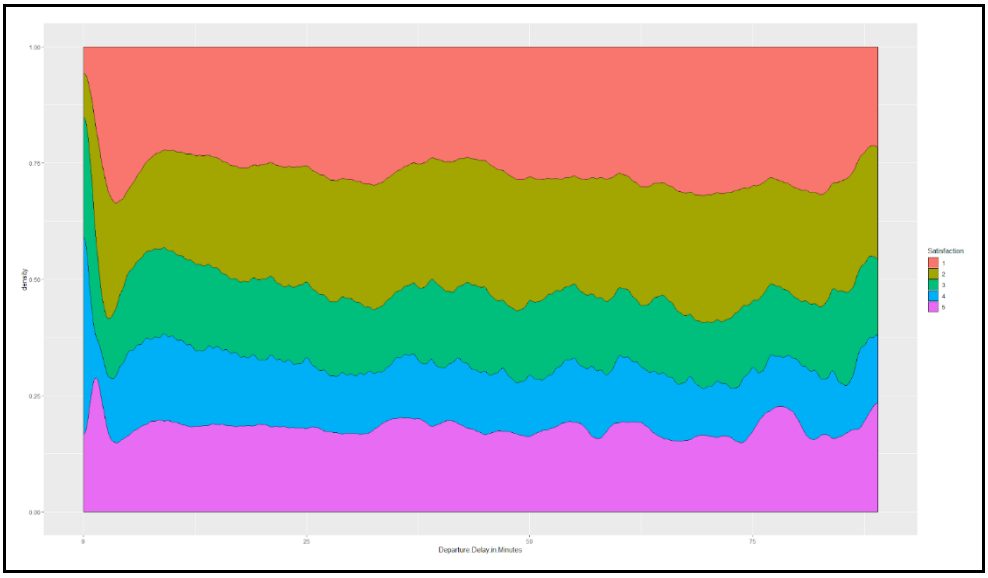


Fig 11. Heat Map of Departure Delay in Minutes

These graphs show that most airplanes usually depart on time and the satisfaction rate is almost stable after the delay time is more than 25 minutes. This means that the satisfaction level is relatively high only when departure delay time is less than 30 minutes.

```
# Satisfaction and Arrival delay
Ardy96 <- quantile(Sati_flight_1$Arrival.Delay.in.Minutes,0.96)
Sati_Ardy <-
ggplot(Sati_flight_1[Sati_flight_1$Arrival.Delay.in.Minutes<Ardy96,],aes(x=Arrival.Delay.in.Minutes))
Sati_Ardy <- Sati_Ardy + geom_histogram(aes(fill=Satisfaction),position = "dodge")
Sati_Ardy <- Sati_Ardy + ggtitle("Satisfaction versus Arrival Delay")
Sati_Ardy
```

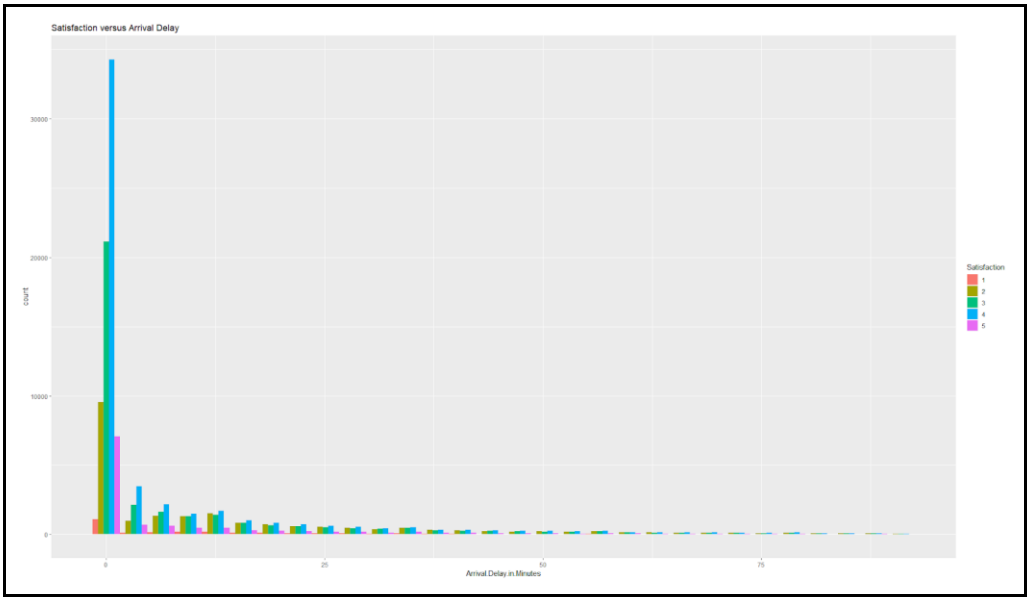


Fig 12. Plot of Satisfaction vs Arrival Delay

```
Sati_Ardy_1 <-
ggplot(Sati_flight_1[Sati_flight_1$Arrival.Delay.in.Minutes<Ardy96,],aes(x=Arrival.Delay.in.Minutes))
+ geom_density(aes(fill=Satisfaction),position = "fill")
Sati_Ardy_1
```

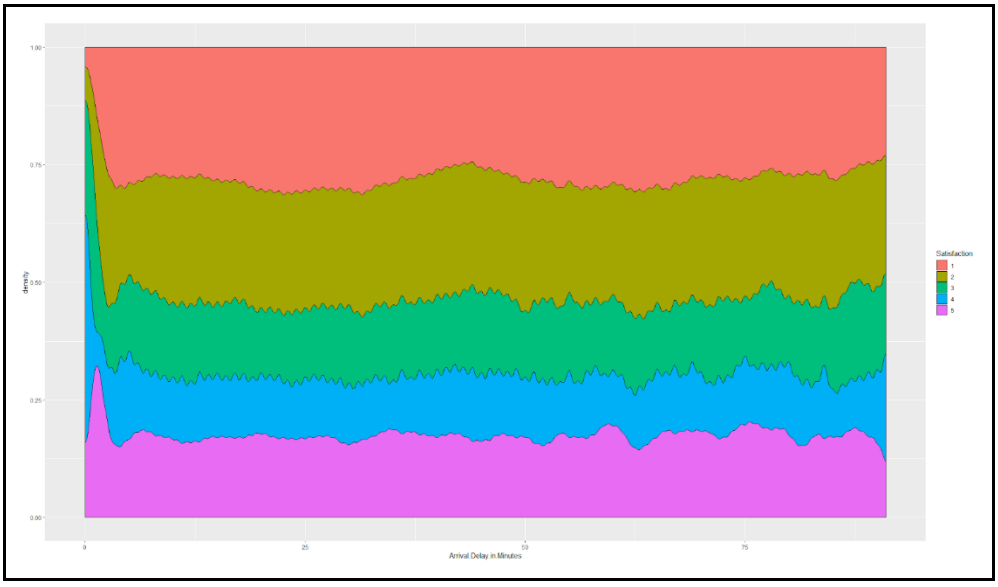


Fig 13. Heat Map of Arrival Delay in Minutes

Same is the case in departure condition, these graphs show that most airplanes arrive on time. And satisfaction keeps almost stable after delay time is more than 25 minutes. This means that the satisfaction level is relatively high only when arrival delay time is less than 30 minutes.

```
# Satisfaction and Price Sensitivity
Sati_Prsy <- ggplot(Sati_flight_1,aes(x=Price.Sensitivity))
Sati_Prsy <- Sati_Prsy + geom_histogram(aes(fill=Satisfaction),position = "dodge")
Sati_Prsy <- Sati_Prsy+ ggtitle("Satisfaction versus Price Sensitivity")
Sati_Prsy
```

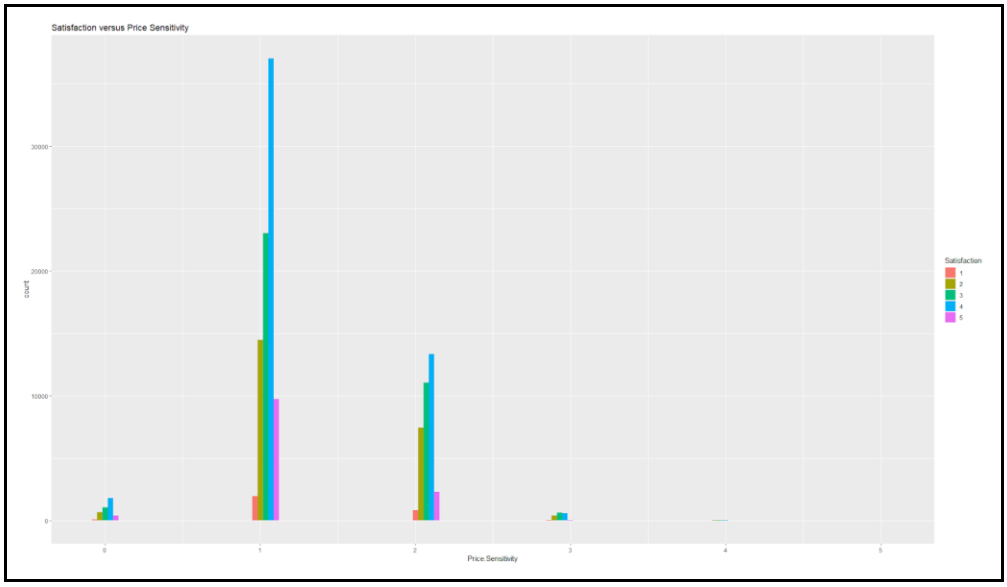


Fig 14. Plot of Satisfaction vs Price Sensitivity

```
Sati_Prsy_1 <- ggplot(Sati_flight_1,aes(x=Price.Sensitivity)) +
geom_density(aes(fill=Satisfaction),position = "fill")
Sati_Prsy_1
```

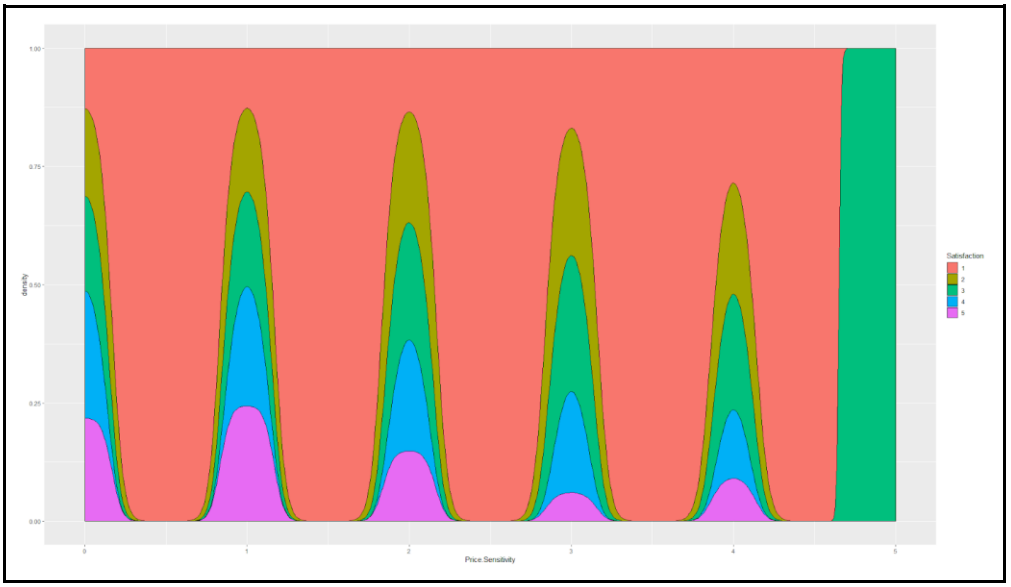


Fig 15. Heat Map of Price Sensitivity

The price sensitivity graphs look different as compared with others and the obvious reason for this is that the density graph is not continuous. This is because we only read data in integer positions. And the tendency is when price sensitivity is higher, the satisfaction will be lower.

 # Positive factors:
 # Satisfaction and Airline Status

```
Sati_As <- ggplot(Airdata_new1,aes(x=Airline.Status))
Sati_As <- Sati_As + geom_histogram(aes(fill=factor(ceiling(Satisfaction))),position =
"dodge")
Sati_As <- Sati_As + ggtitle("Satisfaction versus Airline Status")
Sati_As
```

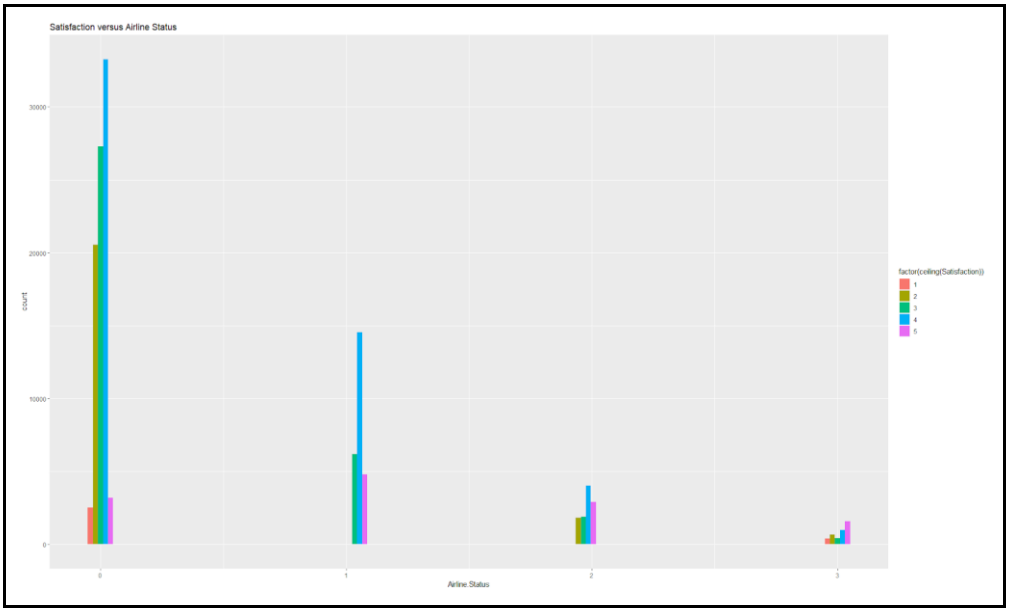


Fig 16. Plot of Satisfaction vs Arrival Status

```
Sati_As_1 <- ggplot(Airdata_new1,aes(x=Airline.Status)) +
geom_density(aes(fill=factor(ceiling(Satisfaction))),position = "fill")
Sati_As_1
```

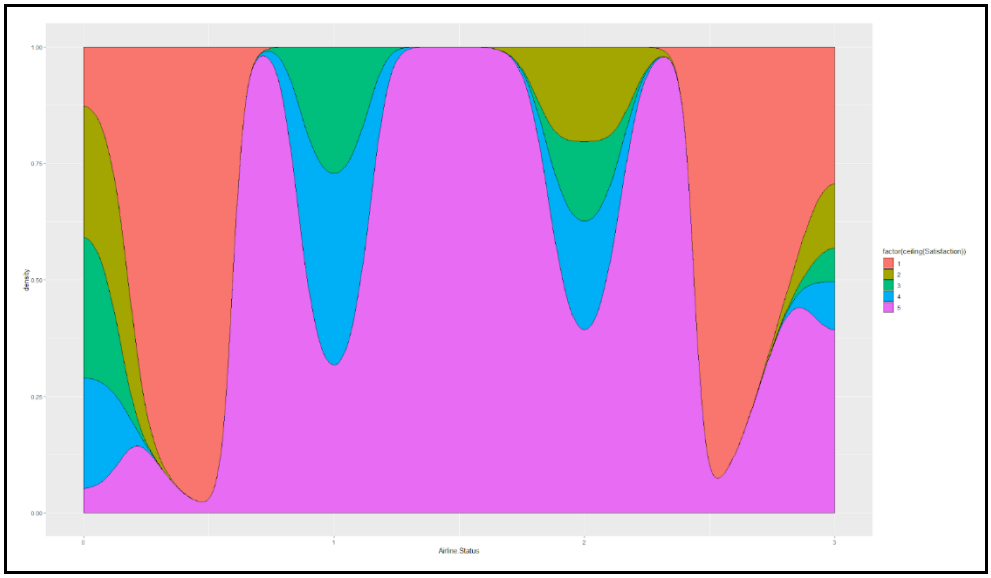


Fig 17. Heat Map of Arrival Status

The graphs of satisfaction vs. airline status tell us that there is no linear relationship between these two factors. But we know that we get a higher satisfaction from silver customer and the lowest or highest membership will give lower satisfaction as compared to others.

```
# Satisfaction and No. Of other Loyalty Cards
maxlc <- quantile(Sati_flight_1$No..of.other.Loyalty.Cards,probs = 0.95)
Sati_Lc <-
ggplot(Sati_flight_1[Sati_flight_1$No..of.other.Loyalty.Cards<maxlc,],aes(x=No..of.other.Loyalty.Cards))
Sati_Lc <- Sati_Lc + geom_histogram(aes(fill=Satisfaction),position = "dodge")
Sati_Lc <- Sati_Lc + ggtitle("Satisfaction versus Loyalty Cards")
Sati_Lc
```

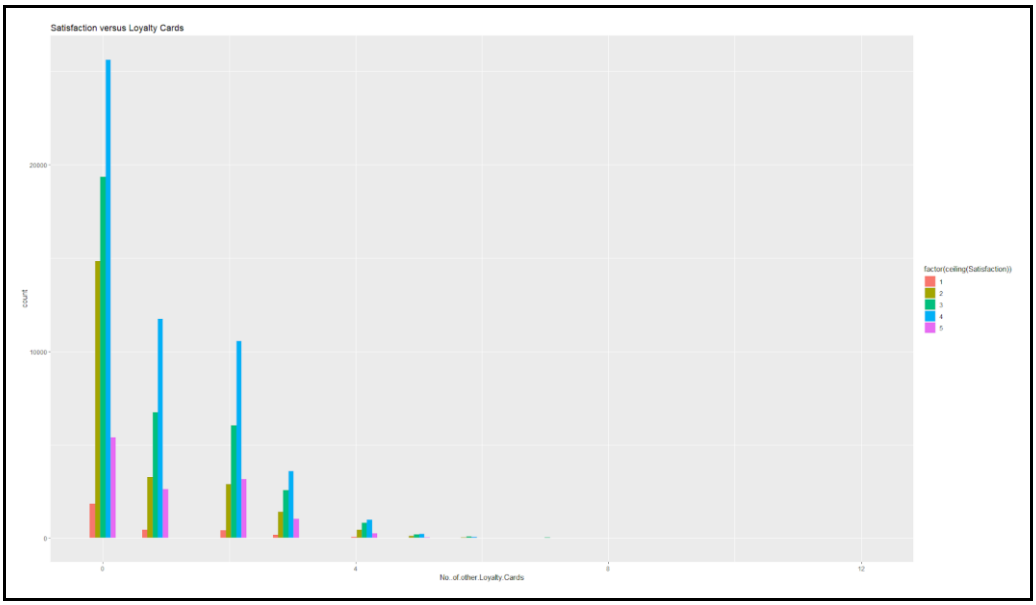


Fig 18. Plot of Satisfaction vs Number of Other Loyalty Cards

```
Sati_Lc_1 <-
ggplot(Sati_flight_1[Sati_flight_1$No..of.other.Loyalty.Cards<maxlc,],aes(x=No..of.other.Loyalty.Cards))
+ geom_density(aes(fill=Satisfaction),position = "fill")
Sati_Lc_1
```

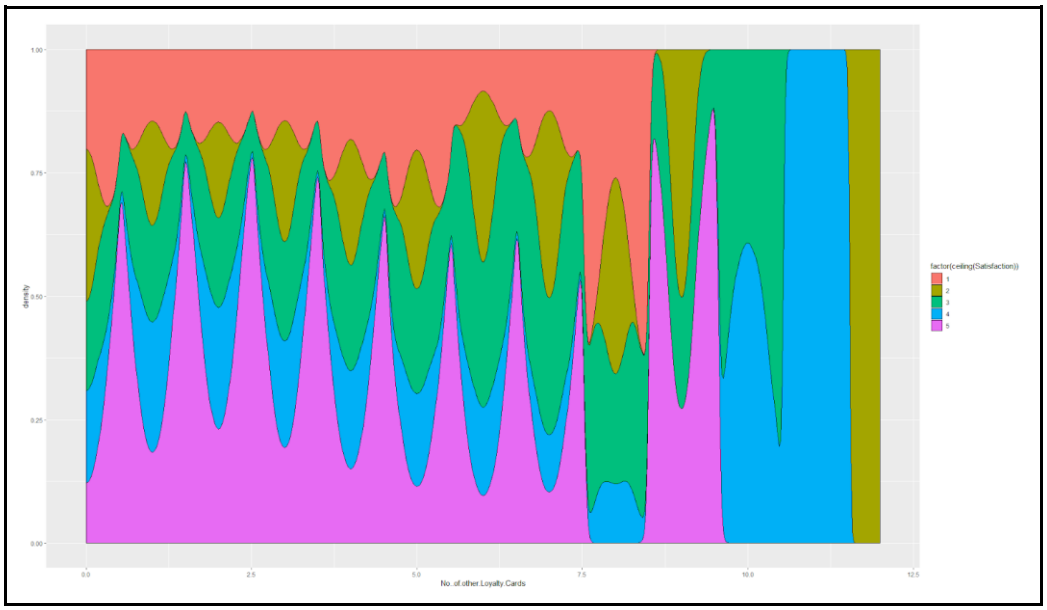


Fig 19. Heat Map of Number of Other Loyalty Cards

Same as price sensitive graph, the density graph of “number of cards” is not continuous because of the integer position data. And from the graph, we could know that most people do not hold any loyalty cards. But for the people who hold loyalty cards, whenever the number of cards is higher, the satisfaction of the customer is lower. Also, there is abnormal high satisfaction when people have 8-9 cards, we think this is inadequate data resulting in outliners.

```
# Satisfaction and Percent of Flight with other Airlines
Fa96 <- quantile(Sati_flight_1$X..of.Flight.with.other.Airlines,0.96)
Sati_Fa <-
ggplot(Airdata_new1[Airdata_new1$X..of.Flight.with.other.Airlines<Fa96,],aes(x=X..of.Flight.with.other.Airlines))
Sati_Fa <- Sati_Fa + geom_histogram(aes(fill=factor(ceiling(Satisfaction))),position = "dodge")
Sati_Fa <- Sati_Fa + ggtitle("Satisfaction versus Flight with other airlines")
Sati_Fa
```

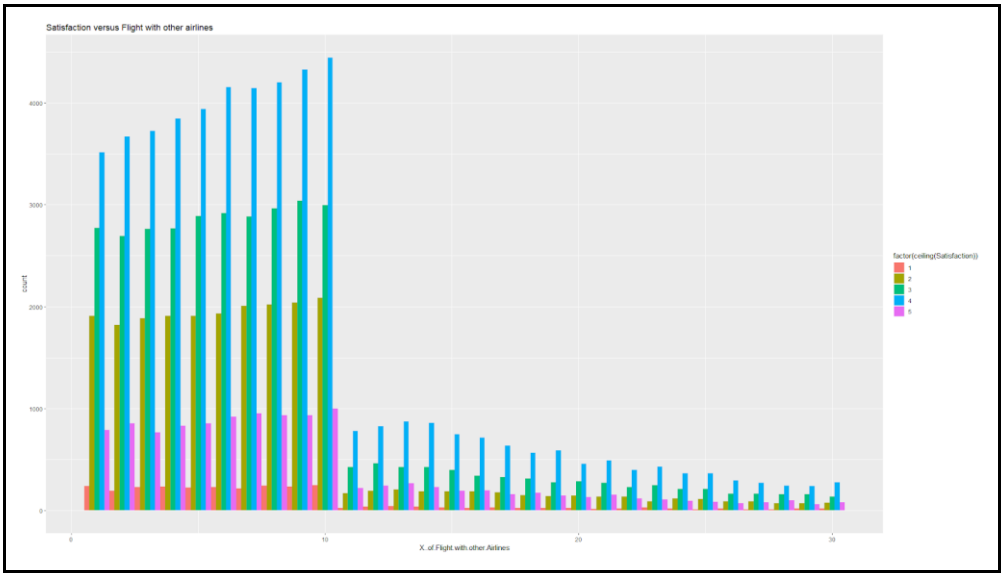


Fig 20. Plot of Satisfaction vs Flight with Other Airlines

```
Sati_Fa_1 <-
ggplot(Airdata_new1[Airdata_new1$X..of.Flight.with.other.Airlines<Fa96,],aes(x=X..of.Flight.with.other.Airlines))
+ geom_density(aes(fill=factor(ceiling(Satisfaction))),position = "fill")
Sati_Fa_1
```

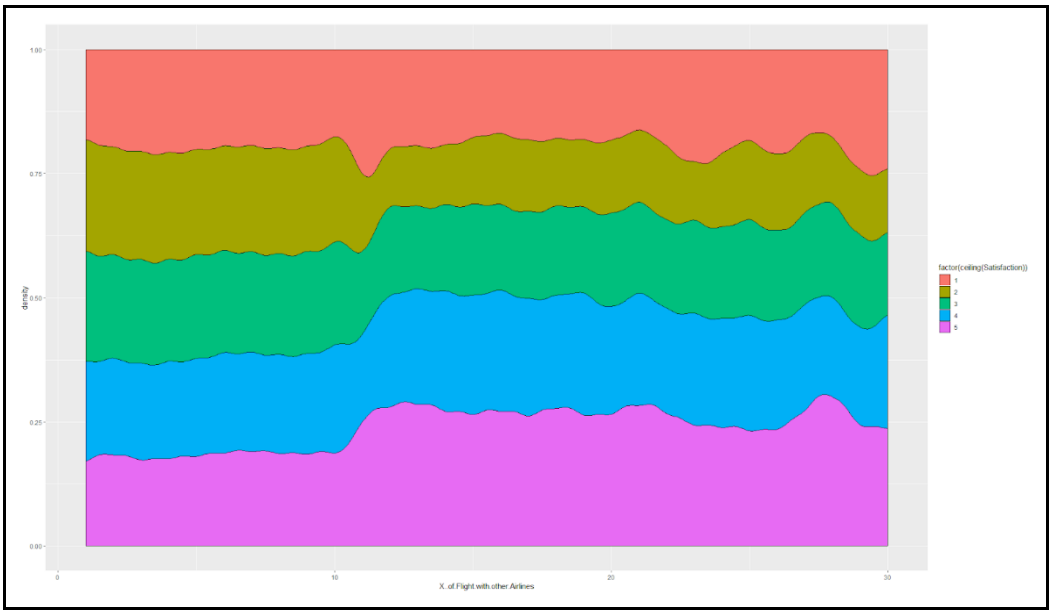


Fig 21. Heat Map of Flight with Other Airline

According to this graph, we could easily make out that most people have lower than 10 flights and higher the number of other flights, keeps the satisfaction stable. Also, there is a significant change in improvement when flight taken are greater than 10.

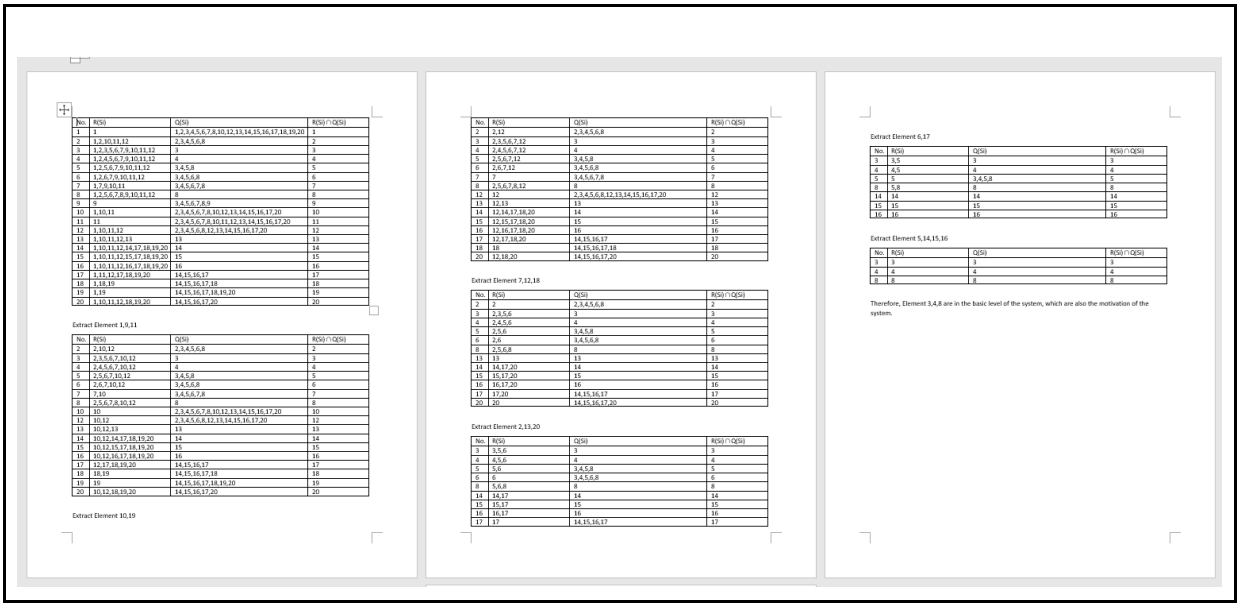
USE OF MODELING TECHNIQUES

A. MODELING 0 – Preparation

a. System Modeling-Reachability Matrix

Reachability matrix is a mathematical tool widely used by system modeling problems. Within a system which has not been understood, all the relationships between the factors in the system can be rather messy and discontinuous. Therefore, this is how the reachability matrix could help us: If we can figure out the one-to-one relationship between any pairs of variables within the system, we can build a reachability matrix after a series of calculation. By kicking out some of the factor according to certain principles, remaining factors can be divided into several levels and, after finishing the link, we can turn it into a hierarchical system with clear and easily understood relationships within the system.

After several calculation, the table of reachability matrix is presented as below:



No.	R(Si)	Q(Si)	R(Si) ∩ Q(Si)
1	1	1,2,3,4,5,6,7,8,10,12,13,14,15,16,17,18,19,20	1
2	1,2,10,11,12	2,3,4,5,6,8	2
3	1,2,3,5,6,7,9,10,11,12	3	3
4	1,2,4,5,6,7,9,10,11,12	4	4
5	1,2,5,6,7,9,10,11,12	3,4,5,8	5
6	1,2,4,7,9,10,11,12	3,4,5,6,8	6
7	1,7,9,10,11	3,4,5,6,7,8	7
8	1,2,5,6,7,8,9,10,11,12	8	8
9	9	3,4,5,6,7,8,9	9
10	1,10,11	2,3,4,5,6,7,8,10,12,13,14,15,16,17,20	10
11	11	2,3,4,5,6,7,8,10,11,12,13,14,15,16,17,20	11
12	1,10,11,12	2,3,4,5,6,8,12,13,14,15,16,17,20	12
13	1,10,11,12,13	13	13
14	1,10,11,12,14,17,18,19,20	14	14
15	1,10,11,12,15,17,18,19,20	15	15
16	1,10,11,12,16,17,18,19,20	16	16
17	1,11,12,17,18,19,20	14,15,16,17	17
18	1,18,19	14,15,16,17,18	18
19	1,19	14,15,16,17,18,19,20	19
20	1,10,11,12,18,19,20	14,15,16,17,20	20

Extract Element 1,9,11

No.	R(Si)	Q(Si)	R(Si) ∩ Q(Si)
2	2,10,12	2,3,4,5,6,8	2
3	2,3,5,6,7,10,12	3	3
4	2,4,5,6,7,10,12	4	4
5	2,5,6,7,10,12	3,4,5,8	5
6	2,6,7,10,12	3,4,5,6,8	6
7	7,10	3,4,5,6,7,8	7
8	2,5,6,7,8,10,12	8	8
10	10	2,3,4,5,6,7,8,10,12,13,14,15,16,17,20	10
12	10,12	2,3,4,5,6,8,12,13,14,15,16,17,20	12
13	10,12,13	13	13
14	10,12,14,17,18,19,20	14	14
15	10,12,15,17,18,19,20	15	15
16	10,12,16,17,18,19,20	16	16
17	12,17,18,19,20	14,15,16,17	17
18	18,19	14,15,16,17,18	18
19	19	14,15,16,17,18,19,20	19
20	10,12,18,19,20	14,15,16,17,20	20

Extract Element 10,19

No.	R(Si)	Q(Si)	R(Si) ∩ Q(Si)
2	2,12	2,3,4,5,6,8	2
3	2,3,5,6,7,12	3	3
4	2,4,5,6,7,12	4	4
5	2,5,6,7,12	3,4,5,8	5
6	2,6,7,12	3,4,5,6,8	6
7	7	3,4,5,6,7,8	7
8	2,5,6,7,8,12	8	8
12	12	2,3,4,5,6,8,12,13,14,15,16,17,20	12
13	12,13	13	13
14	12,14,17,18,20	14	14
15	12,15,17,18,20	15	15
16	12,16,17,18,20	16	16
17	12,17,18,20	14,15,16,17	17
18	18	14,15,16,17,18	18
20	12,18,20	14,15,16,17,20	20

Extract Element 7,12,18

No.	R(Si)	Q(Si)	R(Si) ∩ Q(Si)
2	2	2,3,4,5,6,8	2
3	2,3,5,6	3	3
4	2,4,5,6	4	4
5	2,5,6	3,4,5,8	5
6	2,6	3,4,5,6,8	6
8	2,5,6,8	8	8
13	13	13	13
14	14,17,20	14	14
15	15,17,20	15	15
16	16,17,20	16	16
17	17,20	14,15,16,17	17
20	20	14,15,16,17,20	20

Extract Element 2,13,20

No.	R(Si)	Q(Si)	R(Si) ∩ Q(Si)
3	3,5,6	3	3
4	4,5,6	4	4
5	5,6	3,4,5,8	5
6	6	3,4,5,6,8	6
8	5,6,8	8	8
14	14,17	14	14
15	15,17	15	15
16	16,17	16	16
17	17	14,15,16,17	17

Extract Element 6,17

No.	R(Si)	Q(Si)	R(Si) ∩ Q(Si)
3	3,5	3	3
4	4,5	4	4
5	5	3,4,5,8	5
8	5,8	8	8
14	14	14	14
15	15	15	15
16	16	16	16

Extract Element 5,14,15,16

No.	R(Si)	Q(Si)	R(Si) ∩ Q(Si)
3	3	3	3
4	4	4	4
8	8	8	8

Therefore, Element 3,4,8 are in the basic level of the system, which are also the motivation of the system.

Fig 22. Process of System Modeling

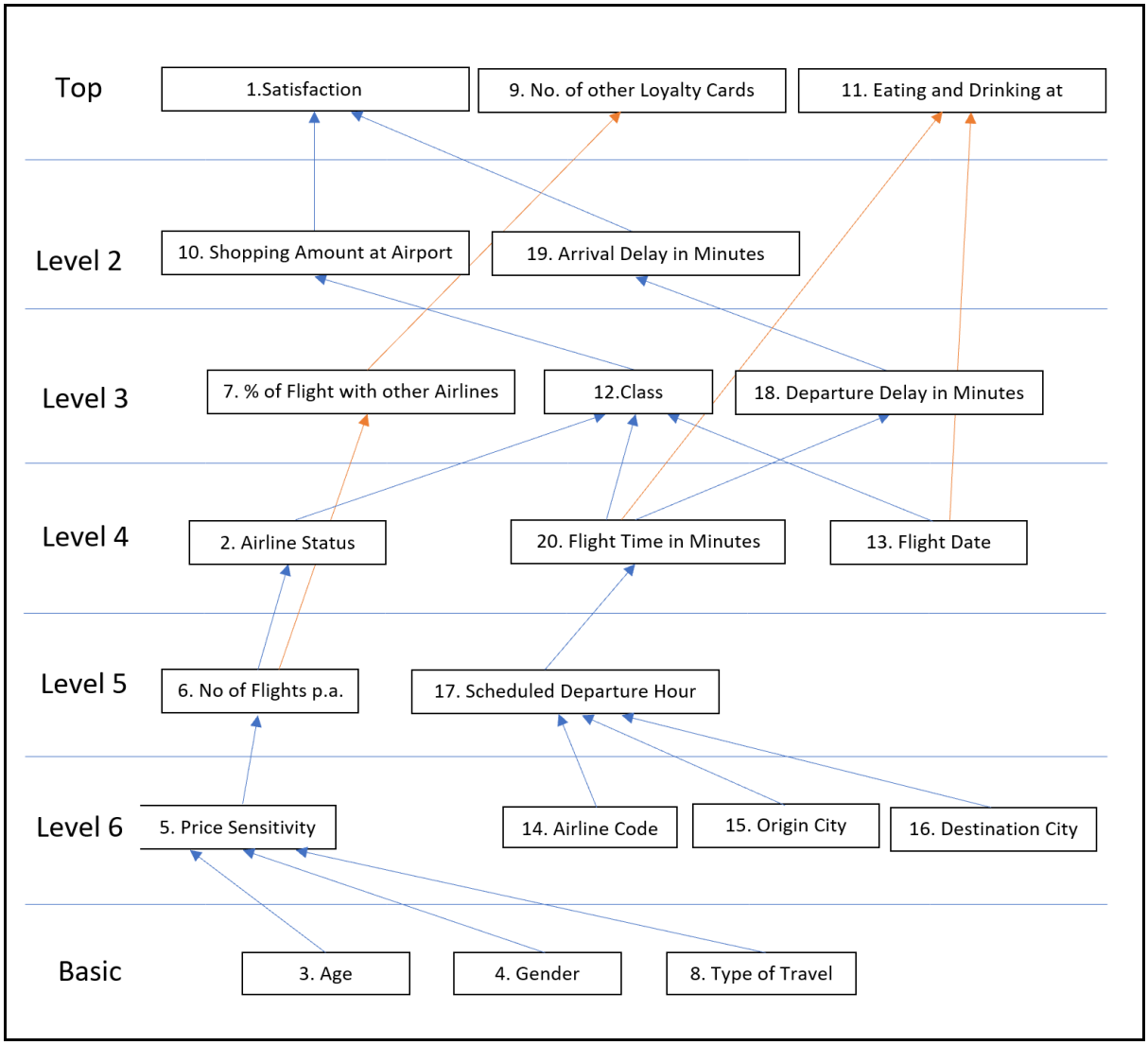


Fig 23. System Modeling

With the reachability matrix, we are able to figure out how the whole “system” is like. However, we still decide to give up the result of it. It is because for the process of building the reachability matrix, we use the logical reasoning to cut off some of the seemly-unreasonable relationship, and after we get the final result of the chart shown above, it didn’t give a good performance during the test of linear modeling. Since there was much work to be done manually to redo the reachability matrix, and we found another matrix with the similar effect on our research. Therefore, we abandoned this result in the end.

This failure taught all of us a good lesson. When we are working with the data, we became aware if it allows us to bring subjective ideas to the data intendedly or unintendedly. Data is honest and intolerant to any “bias”. Therefore, next time when we are dealing with data, we should keep asking ourselves in mind: “Am I treating the data honestly and objectively?”

b. Co-Relation Matrix (Preparation)

A correlation matrix is a table where the correlation coefficients between two variables is presented at the crossing point of the certain row and column. Each cell in the table suggests how relevant it is between two variables. A correlation matrix is often applied as a way to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analysis.
In our research process, there are more than 20 attributes in the data. Some of them would have a strong relationship with the customer's satisfaction while others may have nothing to do the change of satisfaction. Therefore, the correlation matrix can be the first step for us to briefly examine the correlation between satisfaction and other attributes, helping us to select the most relevant attributes for further analysis and study.

```
# Draw correlation matrix and visualize it
Airdata_new1 <- Airdata
Airdata_new1$Gender <- 0
Airdata_new1$Gender[Airdata$Gender == "Male"] <- 1

Airdata_new1$Airline.Status <- as.character(Airdata_new1$Airline.Status)
Airdata_new1$Airline.Status <- 0
Airdata_new1$Airline.Status[Airdata$Airline.Status == "Silver"] <- 1
Airdata_new1$Airline.Status[Airdata$Airline.Status == "Gold"] <- 2
Airdata_new1$Airline.Status[Airdata$Airline.Status == "Platinum"] <- 3

Airdata_new1$Class <- as.character(Airdata_new1$Class)
Airdata_new1$Class <- 0
Airdata_new1$Class[Airdata$Class == "Eco Plus"] <- 1
Airdata_new1$Class[Airdata$Class == "business"] <- 2
Airdata_new1 <- Airdata_new1[,c(-8,-13,-14,-15,-16,-17)]
View(Airdata_new1)
cor_Sati <- cor(Airdata_new1)
cor_Sati

#visualize it
install.packages("corrplot")
library(corrplot)
corrplot(cor_Sati, type = "upper", order = "hclust", tl.col = "black", tl.srt = 45, tl.cex = 0.5)
```

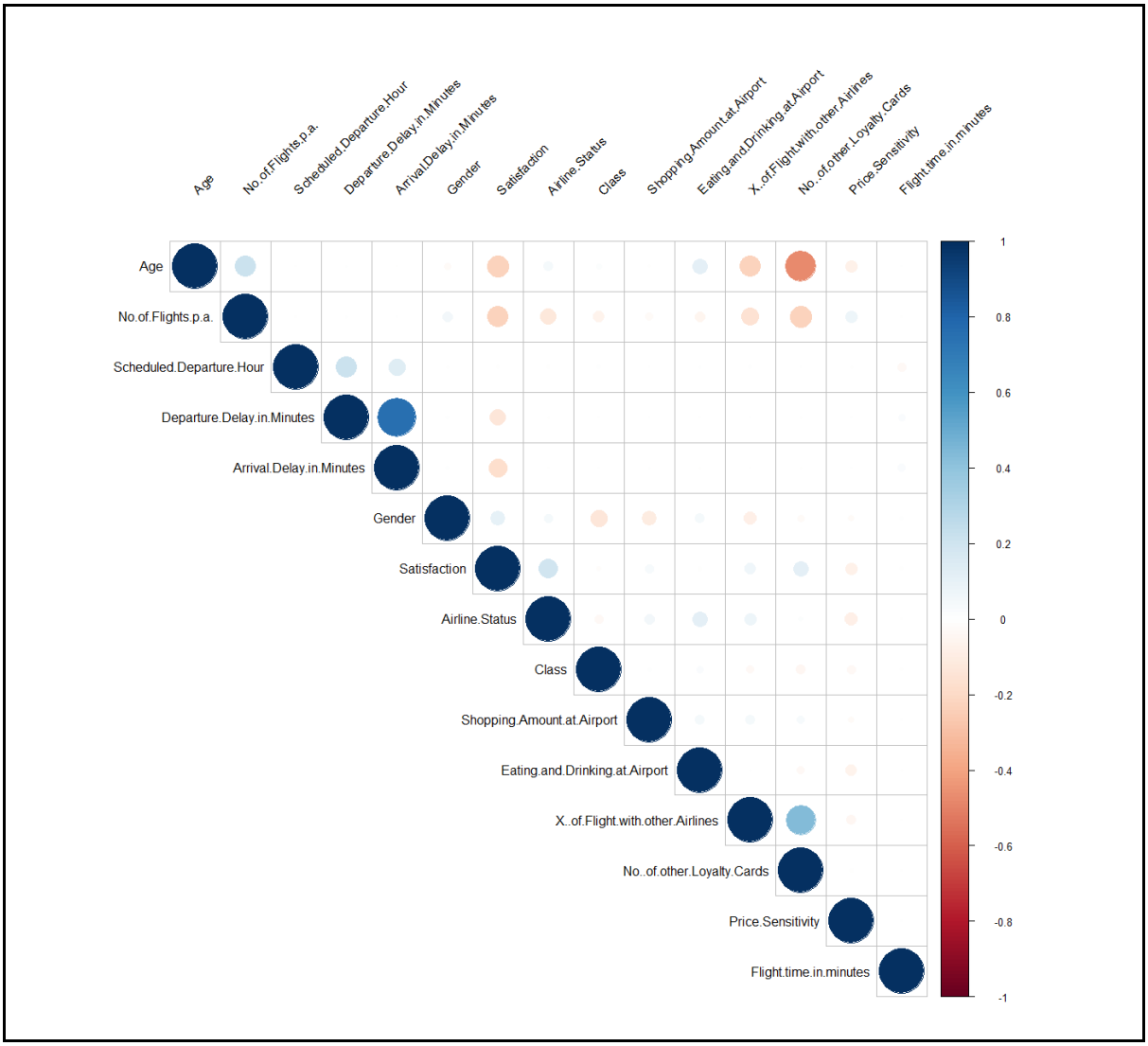


Fig 24. Co- Relation Matrix

```
# Seems not obivious at all
# Normalize it before cor
newSati_2 <- newSati[,c(-8,-13,-14,-15,-16,-17)]
newSati_2$Gender <- newSati_1$Gender
newSati_2$Airline.Status <- newSati_1$Airline.Status
newSati_2$Class <- newSati_1$Class
for( i in 1:15){
  newSati_2[,i] <- as.numeric(newSati_2[,i])
}

new.cor_Sati <- cor(newSati_2)
corrplot(new.cor_Sati, type = "upper", order = "hclust", tl.col = "black", tl.srt = 45, tl.cex = 1)
hist(Airdata$Satisfaction)

install.packages("psych")
library(psych)
pairs.panels(Airdata_new1)
```

What we can see from the heat map, the blue circle suggests the positive impact while the red one tells the opposite meaning. Besides, the size of the circle informs how strongly relevant the two attributes (row and column) are with each other. There are 9 attributes keeping a relationship with the satisfaction, no matter how weak it is, in total. And among all these attributes, 5 of them play negative roles in affecting the customer's satisfaction and the rest 4 of them positively encourage the customer's satisfaction.

Although these results are told by the calculation, they are meaningful and reasonable when combined with daily life's experience. For example, the departure delay in minutes has a negative effect on satisfaction. Obviously, waiting on the airport or on the airplane is directly challenging the people's patience. It is not hard to understand that the longer they wait, the lower satisfaction they provide. Besides this one, percent of flying with other airline was also a big negative attribute to satisfaction. Once clients are more experienced in taking flights and have tried different airline's services, they can't help critically compare the service with "the best service" in their memory, which greatly affects the satisfaction. This data could also reflect customer loyalty, and the impression they have for this airplane.

On the contrary, there are also some interesting points left for our further analysis. For example, number of other loyalty cards has a slight positive effect on satisfaction. Besides, whether the client is a blue card holder, golden holder, or others would affect their satisfaction as well, the higher their membership is, the more likely they are to be satisfied by the experience of taking flight. Why does it happen? How can it make sense? Following analysis are waiting to explore more secrets hid behind these interesting data and relationships.

B. MODELING 1 - Association Rule Mining

We decided to use Association Rule Mining as our first method for discovering interesting relations between the factors in the satisfaction survey data set. This modeling intended to identify strong rules discovered in the data set using some measures of interestingness. It also generates new rules as it analyses more data.

```
install.packages("RJSONIO")
library(RJSONIO)
load("R-Project.RData")
load("Airdata.R")
hist(Sati_flight$Satisfaction,breaks = 10)
nrow(Sati_flight[Sati_flight$Satisfaction<1.5,])
Age_Sat <- lm(formula = Price.Sensitivity ~ .,data = Sati_flight)
summary(Age_Sat)


#Using Association rule modeling to refine to the existing model
#convert useful column into factors
levelit <- function(v){
  new.v <- v
  level4 <- quantile(v[v>0],probs = c(0.3,0.7))
  new.v[v<level4[1]] <- "Low"
```

```
  new.v[level4[1]<=v & v<level4[2]] <- "Average"
  new.v[level4[2]<=v & v<level4[3]] <- "High"
  new.v <- factor(new.v)
  return(new.v)
}
Airdata_new <- Airdata
Airdata_new$Satisfaction <- "Average"
Airdata_new$Satisfaction[Airdata$Satisfaction>3.5] <- "High"
Airdata_new$Satisfaction[Airdata$Satisfaction<3] <- "Low"

newSati <- na.omit(Sati_flight)
newSati$Satisfaction <- ceiling(newSati$Satisfaction)

for (i in c(3,5,6,7,9,10,11,18,19,20,21)){
  Airdata_new[,i] <- levelit(Airdata_new[,i])
}

for (i in 1:21){
  Airdata_new[,i] <- factor(Airdata_new[,i])
}

table(Airdata_new$Satisfaction)

# Start association rules modeling to see the character of customers with diverse satisfaction
install.packages("arules")
library(arules)

newSurvey <- as(Airdata_new,"transactions")
inspect(head(newSurvey,10))
itemFrequency(newSurvey)
itemFrequencyPlot(newSurvey,paramter = list(confidence = 0.9))

result_Sati <- apriori(newSurvey)#,parameter = list(confidence=0.9)

result_Sati_3 <- apriori(newSurvey,parameter = list(confidence=0.6),appearance =
list(default="lhs",rhs="Satisfaction=High"))
summary(result_Sati_3)
result_Sati.new_3 <- result_Sati_3[order(-quality(result_Sati_3)$lift),]
inspect(head(result_Sati.new_3,15))


# Only the top 5 rules are presented for page saving.
#    lhs                          rhs                 support confidence     lift count
#[1]  {Airline.Status=Silver,
      Price.Sensitivity=High,
      Type.of.Travel=Business travel} => {Satisfaction=High} 0.1170020  0.8353027
1.625929 14916
#[2]  {Airline.Status=Silver,
      Type.of.Travel=Business travel} => {Satisfaction=High} 0.1205710  0.8349267
1.625197 15371
#[3]  {Age=Average,
      Gender=Male,
```

```
    Price.Sensitivity=High,
    Type.of.Travel=Business travel} => {Satisfaction=High} 0.1140683  0.7992745
1.555799 14542
#[4]  {Age=Average,
    Gender=Male,
    Type.of.Travel=Business travel} => {Satisfaction=High} 0.1182021  0.7987808
1.554838 15069
#[5]  {Age=Average,
    Gender=Male,
    Type.of.Travel=Business travel,
    Class=Eco}                  => {Satisfaction=High} 0.1002863  0.7941980 1.545918
12785

result_Sati_2 <- apriori(newSurvey,appearance = list(rhs="Satisfaction=Average"))
summary(result_Sati_2)
result_Sati.new_2 <- result_Sati_2[order(-quality(result_Sati_2)$lift),]
inspect(result_Sati.new_2)


#    lhs                           rhs                    support confidence     lift count
#[1]  {Price.Sensitivity=High,
    Type.of.Travel=Personal Travel} => {Satisfaction=Average} 0.1134565  0.3805615
1.349876 14464
#[2]  {Type.of.Travel=Personal Travel} => {Satisfaction=Average} 0.1162960  0.3795796
1.346393 14826
#[3]  {Airline.Status=Blue,
    Gender=Female,
    Price.Sensitivity=High,
    Class=Eco}                  => {Satisfaction=Average} 0.1030082  0.3497390 1.240546
13132
#[4]  {Airline.Status=Blue,
    Gender=Female,
    Class=Eco}                  => {Satisfaction=Average} 0.1056752  0.3485280 1.236251
13472
#[5]  {Airline.Status=Blue,
    Gender=Female,
    Price.Sensitivity=High}        => {Satisfaction=Average} 0.1315292  0.3472139
1.231590 16768

result_Sati_1 <- apriori(newSurvey,parameter = list(confidence=0.3),appearance =
list(default="lhs",rhs="Satisfaction=Low"))
summary(result_Sati_1)
result_Sati.new_1 <- result_Sati_1[order(-quality(result_Sati_1)$lift),]
inspect(result_Sati.new_1)

#     lhs                           rhs                 support confidence     lift count
#[1]  {Airline.Status=Blue,
    Type.of.Travel=Personal Travel} => {Satisfaction=Low} 0.1446131  0.6096359
2.983472 18436
#[2]  {Airline.Status=Blue,
    Price.Sensitivity=High,
    Type.of.Travel=Personal Travel} => {Satisfaction=Low} 0.1403145  0.6092021
2.981348 17888
#[3]  {Airline.Status=Blue,
```

```
    Type.of.Travel=Personal Travel,
    Class=Eco}                  => {Satisfaction=Low} 0.1181786  0.6090718 2.980711
15066
#[4]  {Airline.Status=Blue,
    Price.Sensitivity=High,
    Type.of.Travel=Personal Travel,
    Class=Eco}                  => {Satisfaction=Low} 0.1148606  0.6088059 2.979410
14643
#[5]  {Airline.Status=Blue,
    Type.of.Travel=Personal Travel,
    Shopping.Amount.at.Airport=Low} => {Satisfaction=Low} 0.1022395  0.6051911
2.961719 13034


#final plot for unsatisfied customers
library(arulesViz)
plot(result_Sati.new_1, method = "paracoord")

#final plot for satisfied customers
result_Sati.new_3.top <- head(result_Sati.new_3,15)
plot(result_Sati.new_3.top, method = "paracoord")
```

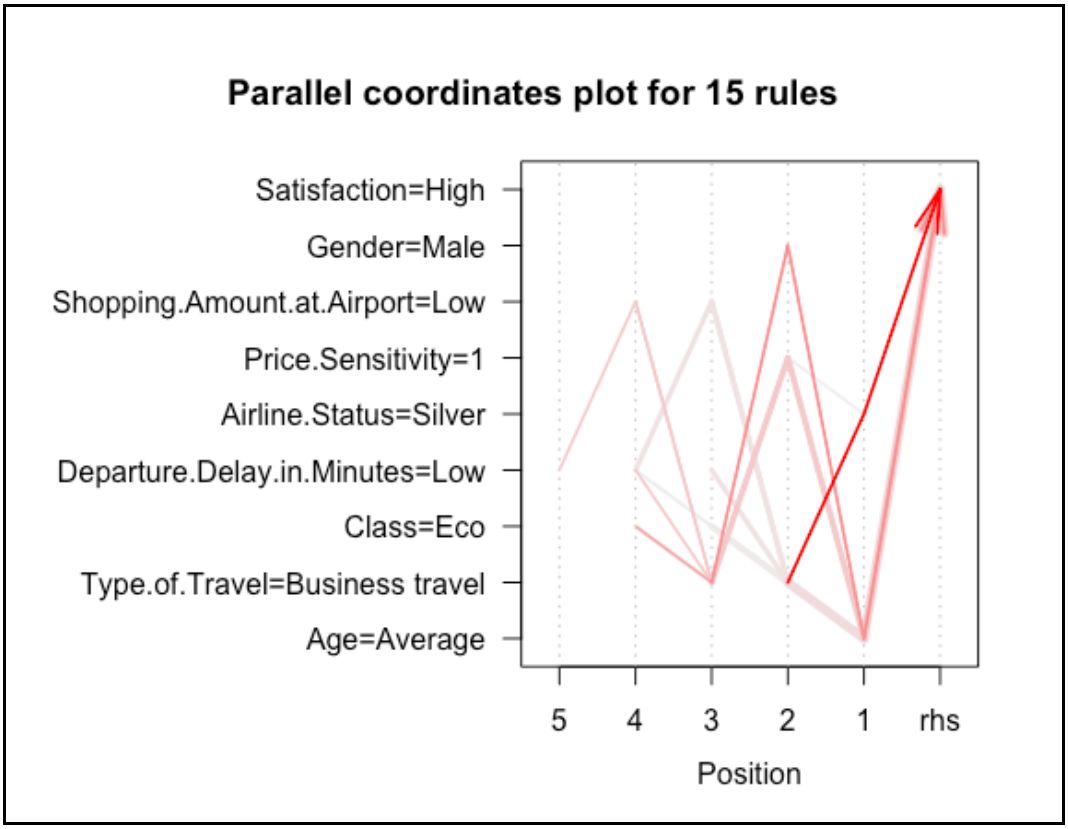


Fig 25. Rule Plot for Satisfied Customers

The above plot is the visualization of association rules created using arulesViz package. This plot is the parallel coordinate plot of the LHS and RHS of the association

rules generated. Here, the top 15 rules are chosen according to the highest lift values. We can see that the Business travellers and Silver status travellers have the highest satisfaction. Also, other factors like reducing the departure delay and shopping amount can make the customers happier.

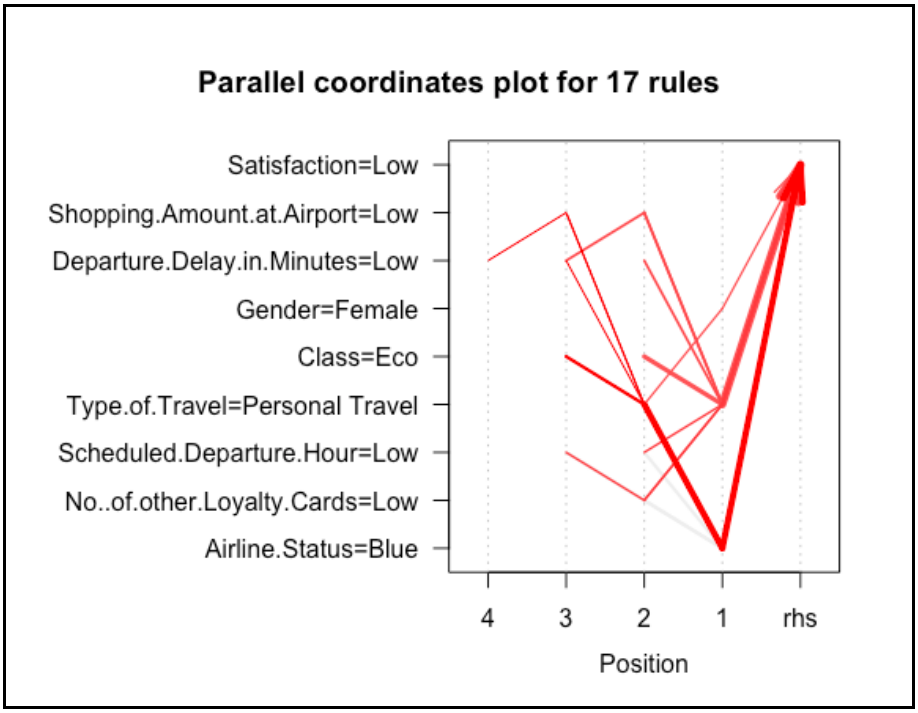


Fig 26. Rule Plot for Unsatisfied Customers

From the above plot, it can be found that Personal travellers and Blue status customers along with Economy class passengers have the lowest Satisfaction. It can also be observed the females who travel for vacation are not satisfied customers.

C. MODELING 2 - Linear Modeling

During our study towards the customer's satisfaction, linear regression is another tool frequently mentioned in our group meetings and online discussions. Linear model tries to explain the relationship between different variables as linear-relevant. By using the linear model, not only could linear modeling be used to deal with many complex information and find a more understandable relationship, but also it is able to tell the difference between attribute's impact on the satisfaction. So, we used linear modeling as another reference to find out the best attribute.

However, even we can apply the result from former analysis to the linear regression, finding the best attribute manually is still a hard work. Therefore, we used the stepAIC() function, which would examine the AIC value of the linear modeling when taking out a single attribute step by step. Then when the Rstudio find the model with the lowest AIC value, it will output it as the model with the best quality, as well as the attributes with the best quality.

```
# Using groupby to normalize the Airdata
Airdata_zscore <- Airdata
for(i in c(1,3,5,6,7,9,10,11,18,19,20,21)){
```

```
Airdata_zscore[,i] <- scale(Airdata_zscore[,i],center = T,scale=T)
}

Air.model4 <- lm(formula = Satisfaction ~., data = Airdata)
library(MASS)
stepAIC(Air.model4)
summary(Air.model4)

#Call: lm(formula = Satisfaction ~ Airline.Status + Age + Gender + Price.Sensitivity +
#No.of.Flights.p.a. + Type.of.Travel + No..of.other.Loyalty.Cards +
#  Shopping.Amount.at.Airport + Class + Flight.date + Airline.Code +
#  Flight.time.in.minutes, data = Airdata)


Air.model6 <-lm(formula = Satisfaction ~ Airline.Status + Age + Gender + Price.Sensitivity
+  No.of.Flights.p.a. + Type.of.Travel + No..of.other.Loyalty.Cards +
              Shopping.Amount.at.Airport + Class + Flight.date + Airline.Code +
              Age*No.of.Flights.p.a. +Age*No..of.other.Loyalty.Cards +
              Flight.time.in.minutes, data = Airdata_zscore)
summary(Air.model6)


lmit <- function(textname){
n<-Airdata_zscore[,-c(14,15)]
n1 <-n[Airdata_zscore$Airline.Name==textname, ]
model <-lm(formula = Satisfaction ~ Airline.Status + Age + Gender + Price.Sensitivity +
               No.of.Flights.p.a. + Type.of.Travel + No..of.other.Loyalty.Cards +
               Shopping.Amount.at.Airport + Class +
               Age*No.of.Flights.p.a. +Age*No..of.other.Loyalty.Cards +
                Flight.time.in.minutes, data =n1)
summary(model)
lmresult <- lmit(levels(Airdata$Airline.Name)[1])
dflm <- data.frame(lmresult$coefficients)
namelist <- rownames(dflm)
esti_result <- dflm$Estimate
name <- levels(Airdata$Airline.Name)
extractesti <- function(textname){
      lmrslt <- lmit(textname)
      ndflm <- data.frame(lmrslt$coefficients)
      nmtch <- match(namelist,rownames(ndflm))
      esti_rslt <- ndflm$Estimate
      return(esti_rslt[nmtch])
```

```
}


#The dataframe of linear modeling's result
finallmresult <- data.frame(extractesti("Cheapseats Airlines Inc."),extractesti("Cool&Young
Airlines Inc."),
       extractesti("EnjoyFlying Air Services"),extractesti("FlyFast Airways Inc."),
       extractesti("FlyHere Airways"),extractesti("FlyToSun Airlines Inc."),
       extractesti("GoingNorth Airlines Inc."),extractesti("Northwest Business Airlines Inc."),
       extractesti("OnlyJets Airlines Inc."),extractesti("Oursin Airlines Inc."),
       extractesti("Paul Smith Airlines Inc."),extractesti("Sigma Airlines Inc."),
       extractesti("Southeast Airlines Co."),extractesti("West Airways Inc."))
rownames(finallmresult) <- namelist
colnames(finallmresult) <- name

# Visualize the linear modeling's result
finallmresult01 <- data.frame(t(as.matrix(finallmresult)))
summary(finallmresult01)
minrslt <- min(unlist(lapply(finallmresult01,min)))
maxrslt <- max(unlist(lapply(finallmresult01,max)))
fnlmrslt01 <- finallmresult01

fnlmrslt01 <-
data.frame(matrix(rank(as.matrix(finallmresult01))/238,nrow(finallmresult01),byrow = F))
colnames(fnlmrslt01) <- colnames(finallmresult01)
rownames(fnlmrslt01) <- rownames(finallmresult01)

fnlmrslt01$AirlineName <- name
fnlmrsltmelt <- reshape2::melt(fnlmrslt01)

#A function to make the ordinary plot to be a radar plot
coord_radar <- function (theta = "x", start = 0, direction = 1)
{
  theta <- match.arg(theta, c("x", "y"))
  r <- if (theta == "x")
    "y"
  else "x"
  ggproto("CordRadar", CoordPolar, theta = theta, r = r, start = start,
          direction = sign(direction),
          is_linear = function(coord) TRUE)
}
# learn it from github

#Compacted radar chart
ggplot(fnlmrsltmelt, aes(x = variable, y = value)) +
      geom_polygon(aes(group = AirlineName, color =AirlineName), fill = NA, size = 2,
show.legend = FALSE) +
      geom_line(aes(group = AirlineName, color = AirlineName), size = 2) +
      theme(strip.text.x = element_text(size = rel(0.8)),
            axis.text.x = element_text(size = rel(0.8)),
            axis.ticks.y = element_blank(),
            axis.text.y = element_blank()) +
      xlab("") + ylab("") +
```

```
      guides(color = guide_legend(ncol=2)) +
      coord_radar()

#Seperate radar chart
ggplot(fnlmrsltmelt, aes(x = variable, y = value)) +
      geom_polygon(aes(group = AirlineName, color = AirlineName), fill = NA, size = 2) +
      facet_wrap(~ AirlineName) +
      theme(axis.ticks.x = element_blank(),
            axis.text.x = element_blank(),
            axis.ticks.y = element_blank(),
            axis.text.y = element_blank()) +
      xlab("") + ylab("") +
      guides(color = "none") +
      coord_radar()


# Making the NPS dataframe
makenps <- function(textname){
  df <- Airdata1[Airdata1$Airline.Name==textname,]
  df.NPS <- nps(df$Satisfaction*2,breaks = list(0:1.5, 2:2.5, 3.5:5))
  return(df.NPS)
}

name <- levels(Airdata$Airline.Name)
Airlinelist.NPS <- unlist(lapply(name, makenps))
names(Airlinelist.NPS) <- name
Airlinelist.NPS <- (Airlinelist.NPS+1)*0.5
View(Airlinelist.NPS)
AirlineNPS <- data.frame("AirlineName"=names(Airlinelist.NPS),"NPS"=Airlinelist.NPS)

#Visualize the NPS value into barplot
ggplot(AirlineNPS,aes(x=AirlineName,y=NPS))+geom_bar(stat =
"identity",aes(fill=AirlineName),alpha=0.6)+
      theme(axis.text.x =element_text(angle = 90, hjust = 1))+
      ggtitle("NPS of different Airlines")
```

According to the result of this code, we can make this form on the spreadsheet:

	Cheapseats Airlines Inc.	Cool&Young Airlines Inc.	EnjoyFlying Air Services	FlyFast Airways Inc.	FlyHere Airways	FlyToSun Airlines Inc.	GoingNorth Airlines Inc.
(Intercept)	0.176358404	0.159128248	0.205976306	0.210242202	0.262330047	0.299209475	0.161105516
Airline.StatusGold	0.449154226	0.525505994	0.415630452	0.44952629	0.408891348	0.33584846	0.390345676
Airline.StatusPlatinum	0.296071857	0.722774196	0.183790273	0.295084274	0.280273908	0.373219805	0.265606114
Airline.StatusSilver	0.658393799	0.616430592	0.6326766	0.61217836	0.599958661	0.573455745	0.711863185
Age	-0.017686019	-0.076436071	-0.022058214	-0.004563199	-0.044566875	-0.007235912	0.017886836
GenderMale	0.127908243	0.115274404	0.180742402	0.136112202	0.118558286	0.087425732	0.112711933
Price.Sensitivity	-0.032307532	0.002683826	-0.017111689	-0.015136683	-0.024822325	-0.044482128	-0.012089634
No.of.Flights.p.a.	-0.039739689	-0.059479243	-0.054109494	-0.051649382	-0.079352879	-0.048991695	-0.013451672
Type.of.TravelMileage tickets	-0.143719825	-0.159884489	-0.173480187	-0.091662449	-0.121042179	-0.083367041	-0.21544149
Type.of.TravelPersonal Travel	-1.097366134	-1.057185315	-1.080053705	-1.115605839	-1.067029531	-1.033779961	-1.145596893
No..of.other.Loyalty.Cards	0.028562104	-0.020573553	0.017061666	0.032887117	-0.015854546	0.038510545	0.067847787
Shopping.Amount.at.Airport	0.004885841	-0.027508874	0.020797355	0.009249678	0.016816069	0.022221328	0.026869
ClassEco	-0.063086965	0.010906096	-0.114771597	-0.105611348	-0.138371833	-0.105061882	-0.066169613
ClassEco Plus	-0.056685945	0.032826179	-0.118167333	-0.078954884	-0.12415658	-0.105202555	-0.002117803
Flight.time.in.minutes	-0.006387244	-0.008551447	-0.022079227	-0.03947518	-0.01854352	-0.003254152	-0.021032786
Age:No.of.Flights.p.a.	0.004217461	-0.010167733	-0.014425517	-0.010901632	-0.008610444	-0.000227343	0.00935961
Age:No..of.other.Loyalty.Cards	0.069578572	0.008555993	0.056895791	0.072741935	0.034440844	0.078018266	0.1413979
Multiple R-squared	0.4265	0.4293	0.4203	0.4349	0.4055	0.3967	0.4352
Adjusted R-squared	0.4262	0.4221	0.4192	0.4343	0.4015	0.3938	0.4293
NPS Result(Adjusted)	***0.2996076***	***0.3205128***	***0.3017542***	***0.3013699***	***0.3125514***	***0.3142899***	***0.2887324***

	Northwest Business Airl	OnlyJets Airlines Inc.	Oursin Airlines Inc.	Paul Smith Airlines Inc.	Sigma Airlines Inc.	Southeast Airlines Co.	West Airways Inc.
(Intercept)	0.237679632	0.176180699	0.281784328	0.220457336	0.195030865	0.229488664	0.290682569
Airline.StatusGold	0.443023123	0.442642642	0.444452458	0.469862037	0.463531416	0.418084875	0.394454371
Airline.StatusPlatinum	0.202069508	0.325153137	0.216520594	0.147475508	0.271858853	0.299643134	0.397643285
Airline.StatusSilver	0.625185212	0.639259503	0.64184345	0.633641703	0.605649663	0.595807803	0.573760276
Age	-0.011966793	-0.056255049	-0.011181207	0.006641757	-0.015420917	-0.019410158	0.000861779
GenderMale	0.136847419	0.128094295	0.114340613	0.132668101	0.129165092	0.13284865	0.117566076
Price.Sensitivity	-0.021979217	-0.036356737	-0.022992295	-0.016865907	-0.013580027	-0.004346563	-0.023953805
No.of.Flights.p.a.	-0.035015073	-0.035288659	-0.056489939	-0.045802409	-0.042044608	-0.039707591	-0.054125673
Type.of.TravelMileage tickets	-0.154741655	-0.149942248	-0.183507266	-0.112368354	-0.122278854	-0.132027236	-0.227825397
Type.of.TravelPersonal Travel	-1.106081039	-1.095893995	-1.095128338	-1.063417322	-1.079066117	-1.026112893	-1.00749299
No..of.other.Loyalty.Cards	0.033088838	0.011013969	0.041291609	0.042842185	0.038065372	0.048140506	0.032645112
Shopping.Amount.at.Airport	0.012848461	0.01392093	-0.005951006	0.005501984	0.007078318	0.001018065	0.005872465
ClassEco	-0.089830379	-0.044450628	-0.124836798	-0.067002175	-0.031096172	-0.060950457	-0.034393301
ClassEco Plus	-0.101671516	-0.042986489	-0.07440193	-0.03078697	-0.030563824	-0.04936992	-0.036328339
Flight.time.in.minutes	-0.0192896	-0.02027147	-0.005250325	-0.007406999	-0.000923405	-0.011488342	0.00224111
Age:No.of.Flights.p.a.	0.002373036	-0.003709323	-0.00758889	-0.015511945	0.002484803	-0.0153405	-0.037078696
Age:No..of.other.Loyalty.Cards	0.072595271	0.075047147	0.076066734	0.076882477	0.079510012	0.09595522	0.07279571
Multiple R-squared	0.4257	0.4323	0.4261	0.412	0.4196	0.4116	0.414
Adjusted R-squared	0.425	0.4306	0.4252	0.4112	0.4191	0.4106	0.4084
NPS Result(Adjusted)	***0.3105821***	***0.2986533***	***0.3088657***	***0.3092541***	***0.3097206***	***0.3091053***	***0.3320498***

Fig 27. Important Attributes from Linear Modeling

Fig 28. Visualization of Separate Impact of Every Attribute on the Airlines

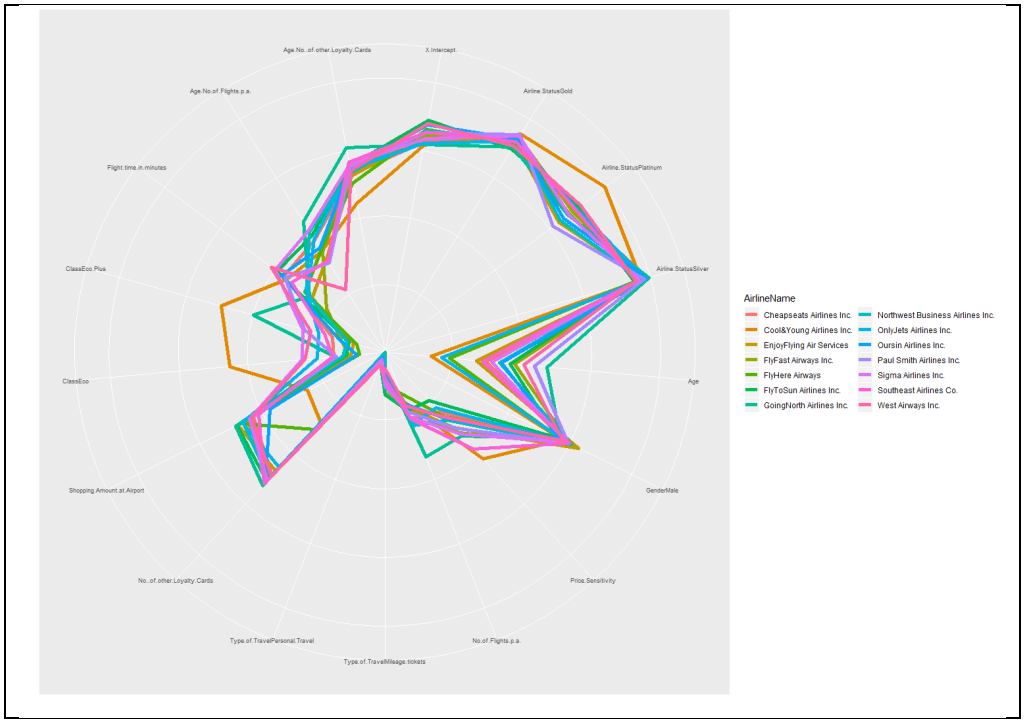


Fig 29. Visualization of Combined impact of Every Attribute on the Airlines

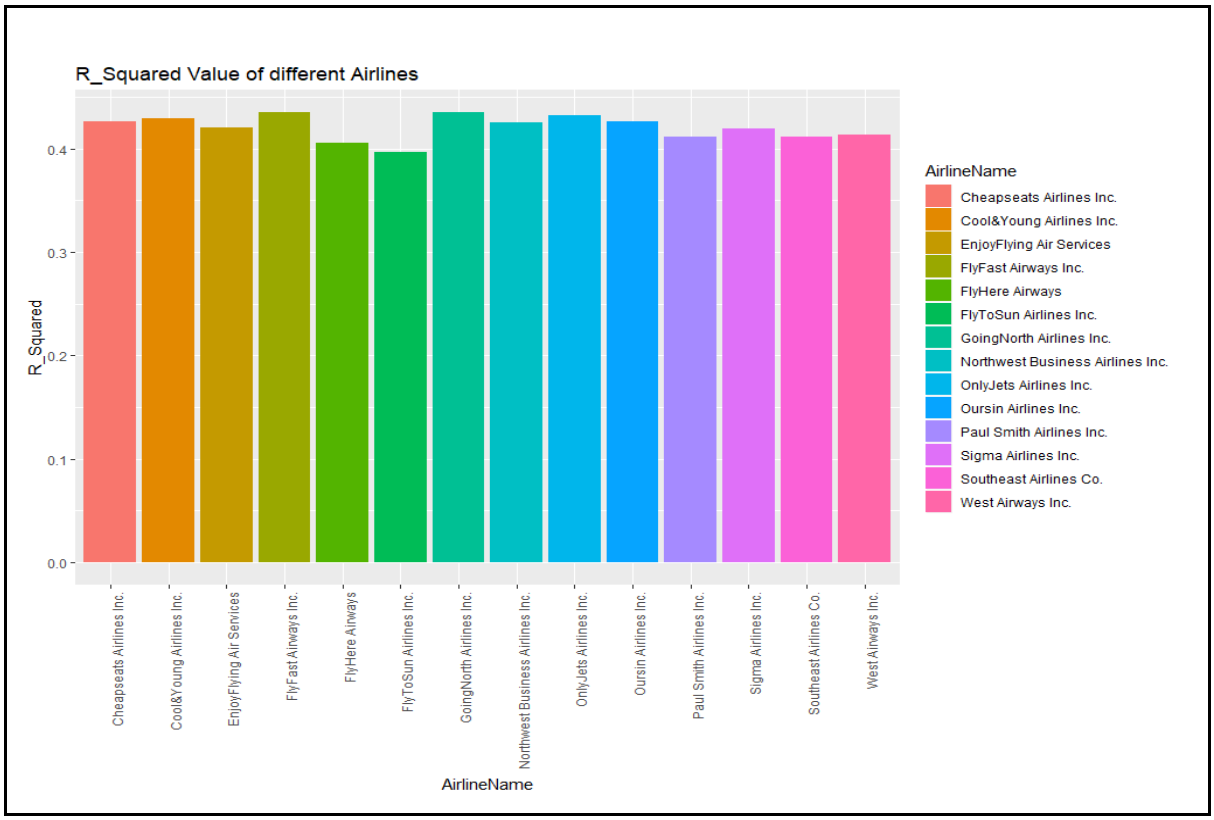


Fig.30. R_Squared Values of Different Airlines

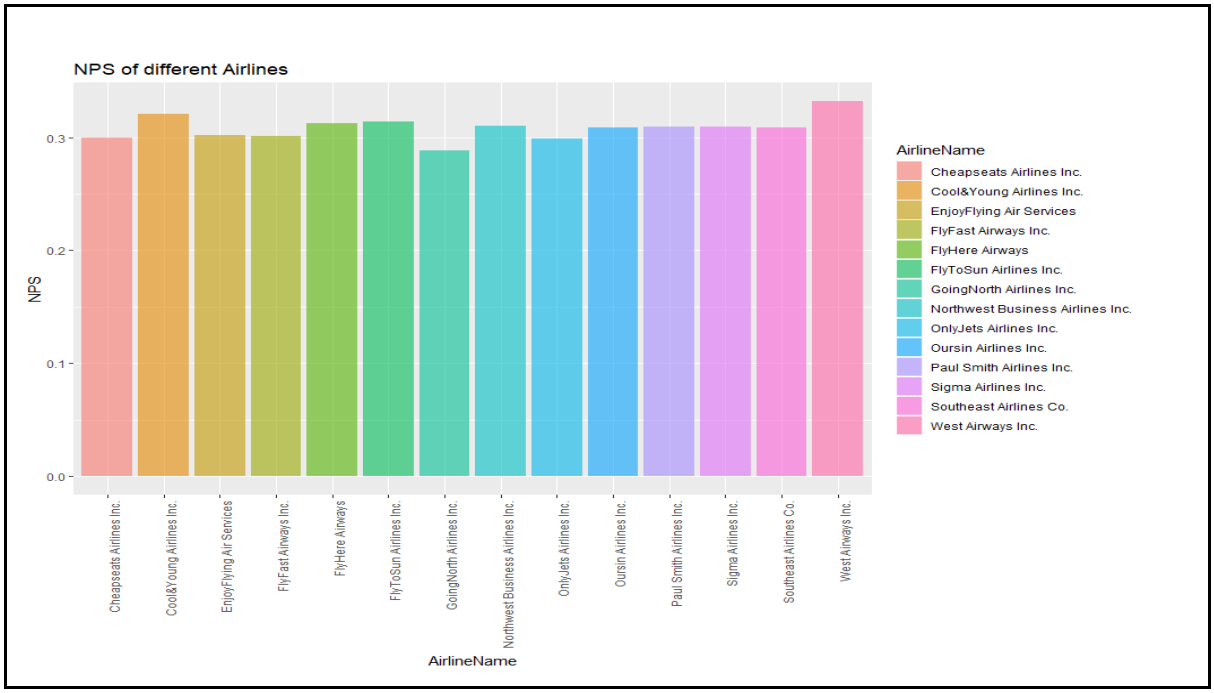


Fig 31. NPS of Different Airlines

The linear modeling part provides 14 models of different airline companies. It can be easily found out on the figure that the impact of a attribute varies in models of different airlines. For example, the shopping amount at airport would encourage the customer's satisfaction in some airline companies such as Cheapest Seat and EnjoyFlying Airline, while it may play a negative role in the model of other airlines such as Cool&Young and Oursin Airline. However, even though these model are different from each other, the R-squared value of them remains approximately same. Therefore,the Net Promoter Score, which is a function to predict recommendable the object will be among their users, is introduced and interestingly, the result of NPS is also similar. Both the airline with the highest NPS and the lowest NPS possess 9 positive factors and most of them are same.

VALIDATION

A. Support Vector Machines

After finding out the attributes affecting the customer's satisfaction more significantly than other attributes in the data, It is still too early to tell that those data are qualified enough for tell the customer's satisfaction accurately. The quality of the data in these attributes should be tested before we can make sure they are absolutely "good factors". Therefore, the SVM model was introduced to analyze the quality of the data in several attributes.

The Support Vector Machine (SVM) is a kind of model for classifying the data according to the traindata prepared in advance. The way we are considering to use SVM in our project is that we train the algorithm on the satisfaction data and then we test it on a brand-new set of data. Once the training works well, the algorithm will 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. In other words, it is the quality of data that decides the quality of the SVM model. Therefore, training and testing the SVM model can be a great way for us to examine the quality of selected attributes, as well as the results and business recommendations based on them.

```
#Using the SVM model to predict the data
install.packages("kernlab")
library(kernlab)

trainindex <- sample(c(1,2,3), nrow(Airdata),replace= T,prob = c(0.15,0.45,0.4))
traindata <- Airdata[trainindex==1,]
testdata <- Airdata[trainindex==2,]

svmOutput <- ksvm(Satisfaction ~ Airline.Status + Age + Gender + Price.Sensitivity +
                  No.of.Flights.p.a. + Type.of.Travel + No..of.other.Loyalty.Cards +
                  Shopping.Amount.at.Airport + Class + Airline.Code +
                  Departure.Delay.in.Minutes+Arrival.Delay.in.Minutes,
data=traindata,kernel="rbfdot", kpar="automatic",C=40,cross=4, prob.model=TRUE)
svmOutput


#Support Vector Machine object of class "ksvm"

#SV type: eps-svr  (regression)
# parameter : epsilon = 0.1  cost C = 40

#Gaussian Radial Basis kernel function.
# Hyperparameter : sigma =  0.0852578788033577

#Number of Support Vectors : 10645

#Objective Function Value : -90086.5
#Training error : 0.403878
#Cross validation error : 0.619004
#Laplace distr. width : 2.023007

svmresult <- predict(svmOutput,testdata,type="votes")
View(svmresult)
compactable <- (testdata[,1]-svmresult)<0.8&(testdata[,1]-svmresult)>-0.7
result <- table(compactable)
result
accuraryratio <- result[2]/(sum(result))
accuraryratio
#0.690532
```

Since we directly applied the SVM to predict the exact score of customer's satisfaction. The result of prediction is numeric. Therefore, we decided to convert the number into boolean types ("True" and "False") by setting a standard. We are assuming if the difference between the predicted score and satisfaction of test dataset is with a range of (-0.7,0.8), (the absolute number of the range is 1.5, 30% of the total range of satisfaction), we regard the predicted result to be "Accurate" ("True"), otherwise it's "Not Accurate" ("False"). In this way we finally translated the predicted score into a table of "True" or "False". Therefore, the accuracy ratio of the SVM model is 69%, which would, in some aspects, guarantee the quality of the selected attributes used in the SVM model.

B. Random Forest Model

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees. Here, we have used the random forest model to build trees which can validate our results which were found out using other models such as the Linear Model or the Apriori algorithm. The variables used in the random forest model were from the correlation matrix which were affecting the Satisfaction column. Also, this model tries to validate the results of the Association rule mining model.

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

install.packages("devtools")
library(devtools)
devtools::install_github('araastat/reprtree')
library(reprtree)

#building the random forest model
rf <- randomForest(Satisfaction ~ Airline.Status + Type.of.Travel, data=Airdata )
importance(rf)
print(rf)
plot(rf)

#plotting the tree for Airline Status and Type of Travel
reprtree:::plot.getTree(rf)

#plotting the tree for Age and Gender
rf1 <- randomForest(Satisfaction ~ Age + Gender, data=Airdata )
reprtree:::plot.getTree(rf1)
```

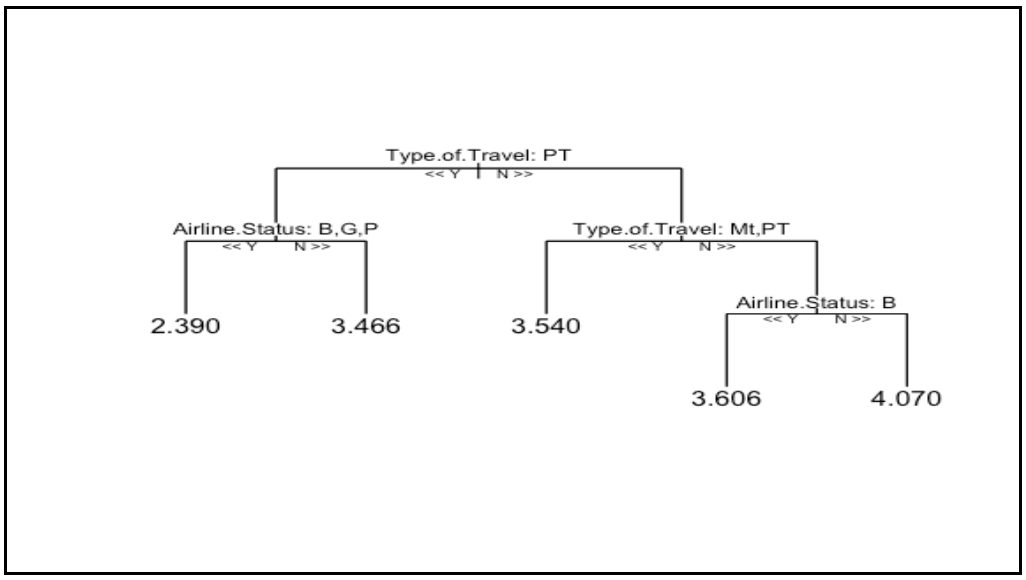


Fig 32. Random Forest Tree predicting Satisfaction
using Type of Travel and Airline Status

The above figure shows the predicted Satisfaction values on the leaf nodes of the Random forest tree. Here, we can make rules and validate our results that we found out with association rule mining technique. When the Type of travel is Business Travel and the Airline Status is Silver and not Blue, Gold or Platinum the Satisfaction of the customers is relatively high (~4). Similarly, Personal Travellers who might be going on a vacation with an Airline Status - blue have a relatively low satisfaction (~2.3).

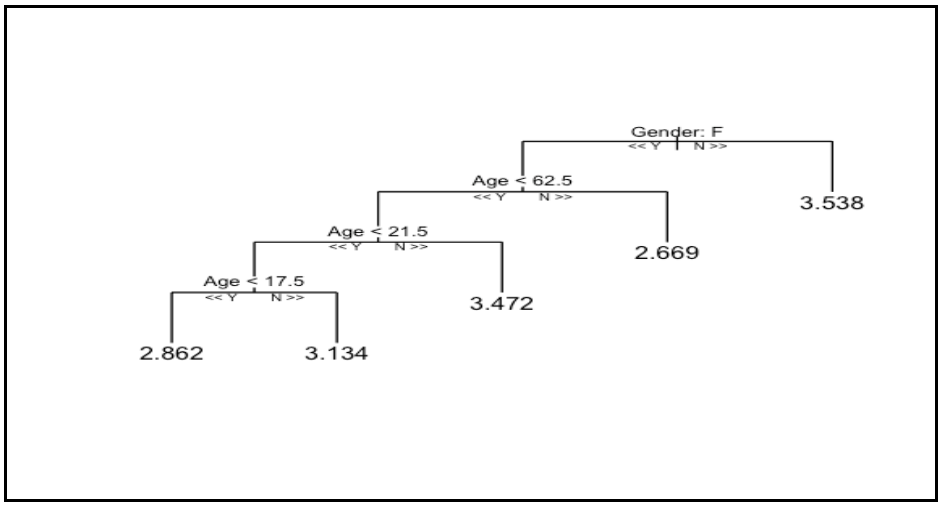


Fig 33. Random Forest Tree predicting Satisfaction using Gender and Age

The above plot justifies that the females have relatively lower satisfaction. Also, the male senior citizens from the age group 60 ~85 are relatively unsatisfied customers.

ACTIONABLE INSIGHTS

It can be observed from the above modeling techniques that the factors that affect the satisfaction rate can be classified in two categories:

A. Factors that gave us high satisfaction rate:

1. Price Sensitivity: The customers are satisfied when the prices are low.
2. Class of Travel: The customers who are traveling by business are highly satisfied.
3. Departure in Minutes: When the departure in minutes is low that is when the customers are satisfied and give good ratings.
4. Age: The age approximately in the average range are the ones who are satisfied with the travel.

B. Factors that give us the low satisfaction rate:

1. Class of Travel: The quality of services for the economy travelers should be improved to increase the satisfaction rate.
2. Type of Travel: From the above plots it can be observed that personal travelers are not satisfied.
3. Airline Status: The travellers traveling by Blue airline package are less satisfied as compared to another traveller package.
4. Gender: Female are less satisfied with the services provided at the airline.

C. Insights:

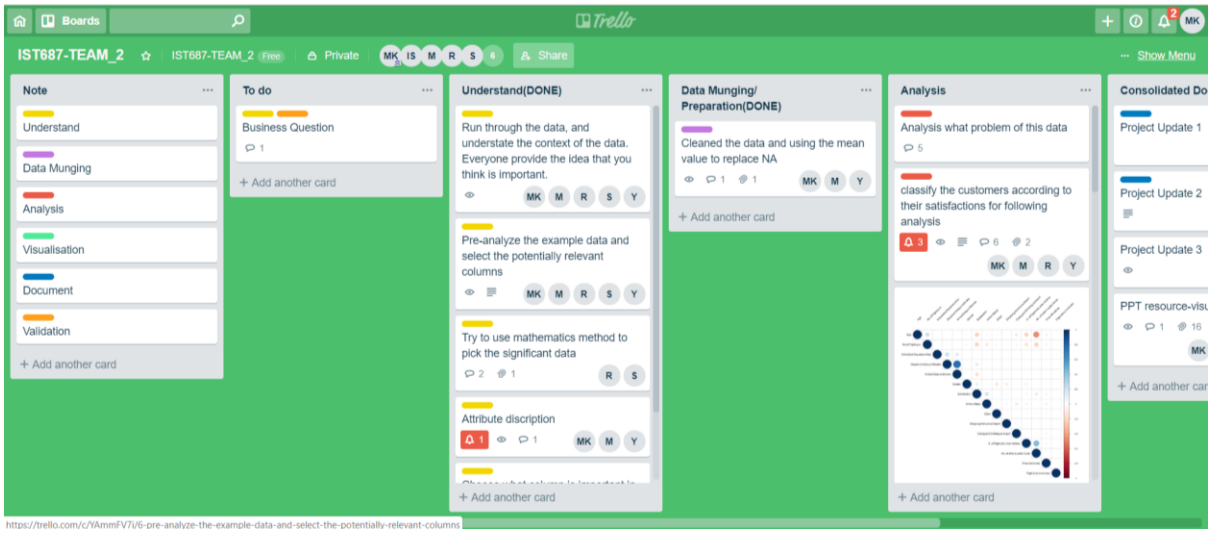
When we finished the analysis, we can get the following information and insights:

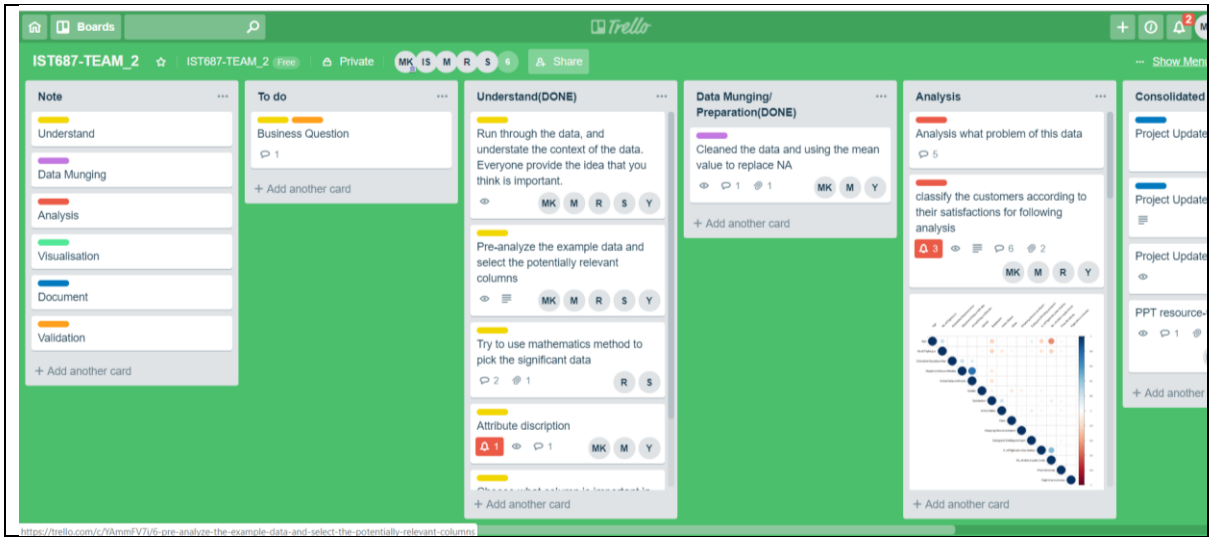
1. The airline status has a positive relationship with the satisfaction for all the airline companies. Thus, **people with Blue airline status should also be given some benefits that would help us increase their satisfaction**.
2. For the number of flights taken annually, we found that the larger number will lead to lower satisfaction. We think this is not caused by the service quality, however too many times of taking flights will make customers feel bored (not fresh anymore) for the airplane service. So, the airplane company may consider more idea to attract customers' interest, like **making a new introduction video.**
3. The personal passengers and mileage ticket owners always had low satisfaction. It may be due to the fact that they are in different class, we find the personal passenger has a high ratio to choose Eco & Eco Plus Class. Thus, **airlines should try to bridge the gap between services provided to personal passengers and others.**
4. The class of Eco & Eco Plus are usually negatively relevant on the satisfaction, except for Cool & Young, we suggest that the **airplane company emphasize on the overall experience of Eco and Eco Plus customers to increase their feedback rating positively.**
5. Usually, the shopping amount reflects the satisfaction of customer on airports directly. While we cannot find any relevant information to determine the most

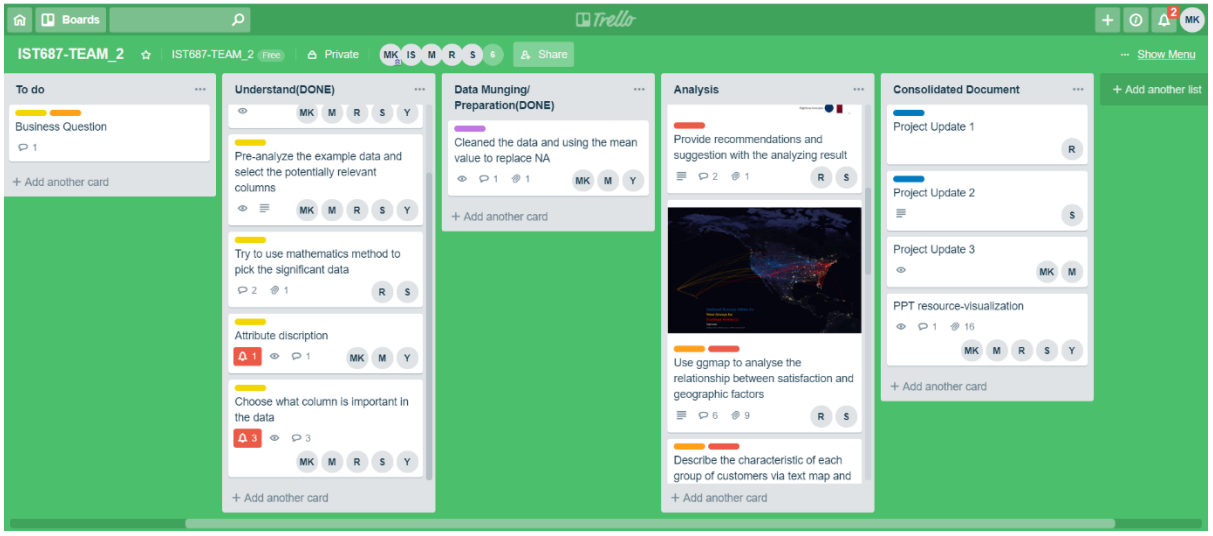
popular airport or airline so, it works only as a crucial factor in exploring the new flights or new airports, or managing the existing flights.

6. What we got from the flight time in minutes is that customer is more likely to be upset for long-distance travel. Hence, **how to improve the user's experience of long-time travel** should be an important factor to consider for airline companies. Including some nice entertainment packages or services to help passengers pass their time would help increase the satisfaction of the customers.
7. The number of loyalty cards can work as a detector for the service of every airline company. Obviously, bigger numbers of loyalty cards is better than smaller numbers of loyalty cards.
8. The relationship between airline service and ages show a negative impact on satisfaction. Thus, **including some better healthcare services and adding facilities that could help aged people to travel easily from one place to other will help in increasing the satisfaction.** Also, **more wheelchairs can be made available and personnel to help with luggage can bring relief to senior citizens.**
9. From gender, in general, men are more likely to be satisfied than women, which reflects the airline may lack the facilities needed for women. For this point, **providing more facilities to take care of women should be considered by airline company, such as special services for pregnant women, young child and infant**, etc.
10. Customer with higher price sensitivity are more likely to be unhappy. Therefore, in order to delight them, **small discounts or bonus on ticket can be a good trick to get their better feedback**.

Trello Board




Boards
Trello
IST687-TEAM_2
IST687-TEAM_2 Free
Private
Share
Show Menu
Note
Understand
Data Munging
Analysis
Visualisation
Document
Validation
+ Add another card
To do
Business Question
+ Add another card
Understand(DONE)
Run through the data, and understate the context of the data. Everyone provide the idea that you think is important.
Pre-analyze the example data and select the potentially relevant columns
Try to use mathematics method to pick the significant data
Attribute discription
+ Add another card
Data Munging/ Preparation(DONE)
Cleaned the data and using the mean value to replace NA
+ Add another card
Analysis
Analysis what problem of this data
classify the customers according to their satisfactions for following analysis
+ Add another card
Consolidated
Project Update
Project Update
Project Update
PPT resource-
+ Add another
https://trello.com/c/YAmmFV7i/6-pre-analyze-the-example-data-and-select-the-potentially-relevant-columns


Boards
Trello
IST687-TEAM_2
IST687-TEAM_2 Free
Private
Share
Show Menu
To do
Business Question
+ Add another card
Understand(DONE)
Pre-analyze the example data and select the potentially relevant columns
Try to use mathematics method to pick the significant data
Attribute discription
Choose what column is important in the data
+ Add another card
Data Munging/ Preparation(DONE)
Cleaned the data and using the mean value to replace NA
+ Add another card
Analysis
Provide recommendations and suggestion with the analyzing result
Use ggmap to analyse the relationship between satisfaction and geographic factors
Describe the characteristic of each group of customers via text map and
+ Add another card
Consolidated Document
Project Update 1
Project Update 2
Project Update 3
PPT resource-visualization
+ Add another card
+ Add another list


Boards
Trello
IST687-TEAM_2
IST687-TEAM_2 Free
Private
Share
Show Menu
Understand(DONE)
Pre-analyze the example data and select the potentially relevant columns
Try to use mathematics method to pick the significant data
Attribute discription
Choose what column is important in the data
+ Add another card
Data Munging/ Preparation(DONE)
Cleaned the data and using the mean value to replace NA
+ Add another card
Analysis
Describe the characteristic of each group of customers via text map and ggplot2
Analysis the relevant variables from the data using ggplot2
+ Add another card
Consolidated Document
Project Update 1
Project Update 2
Project Update 3
PPT resource-visualization
+ Add another card
+ Add another list


Boards
Trello
IST687-TEAM_2
IST687-TEAM_2 Free
Private
Share
Show Menu
Understand(DONE)
Pre-analyze the example data and select the potentially relevant columns
Try to use mathematics method to pick the significant data
Attribute discription
Choose what column is important in the data
Add another card
Data Munging/ Preparation(DONE)
Cleaned the data and using the mean value to replace NA
Add another card
Analysis
Use the linear modeling to analyze different the airline brands and explain the factors
Parallel coordinates plot for 15 rules
Add another card
Consolidated Document
Project Update 1
Project Update 2
Project Update 3
PPT resource-visualization
Add another card
Add another list


Boards
Trello
IST687-TEAM_2
IST687-TEAM_2 Free
Private
Share
Show Menu
Understand(DONE)
Pre-analyze the example data and select the potentially relevant columns
Try to use mathematics method to pick the significant data
Attribute discription
Choose what column is important in the data
Add another card
Data Munging/ Preparation(DONE)
Cleaned the data and using the mean value to replace NA
Add another card
Analysis
Association rules Visualizations
Random Forest Validation Analysis
Add another card
Consolidated Document
Project Update 1
Project Update 2
Project Update 3
PPT resource-visualization
Add another card
Add another list