IST 687 – Applied Data Science

Lab Section M005 | Group 1

FLIGHT FEEDBACK SURVEY ANALYSIS

Recommendations to Increase Customer Satisfaction

Submitted by:
Adarsh Patil | Shashikant Dhuppe | Vidisha Badhe | Preksha Agrawal | Shravya Kamath

Table of Contents

1. Description

In this research we have been provided with an aggregate dataset of different airlines of a parent airway. We have dataset from a satisfaction survey with 129,889 rows of records and it represents answers to questions collected from a vast set of customers flying within United States using various airlines. We are focusing on a specific airline, Cheapseats Airlines and have done an extensive research on this airline's customer satisfaction survey. The project revolves around analyzing the data collected from customer evaluations and feedback, booking information, and overall customer satisfaction, in order to provide recommendations to improve satisfaction rates. The recommendations simply answer business questions to increase customer satisfaction.

2. Project Scope

2.1 Objectives

Scope of the project is based on the extent of the data set which is two-fold. Firstly, the origin and destination geographical area of flights, which in this case is limited within the United States. Secondly, time frame for which data is collected and made available, which is January 2014 through March 2014 in our case.

Objective of the project is to provide suggestions for best practices for Cheapseats airlines with the main purpose of increasing customer satisfaction. Various data analysis models have been used to find trends between various parameters of the data set with the aim of providing accurate results for solving the business questions that would arise eventually.

2.2 Assumptions

A. The customer survey data is given by a parent airways company. The airlines mentioned in the survey cater to different airlines of the same airways.

B. Hierarchy of prices of flight tickets from high to low:
   a. Airline Status → Platinum
   b. Airline Status → Gold
   c. Airline Status → Silver
   d. Airline Status → Blue

C. Range of Price sensitivity varies from 0 to 5 in which 0 means customers are not so sensitive about price variations of flight tickets. 5 means customers are sensitive to price variations of flight tickets.

D. Every record in the data sheet is of one customer travelling in a particular airline

3. Deliverables

The following deliverables have been identified based on the purpose of the project:

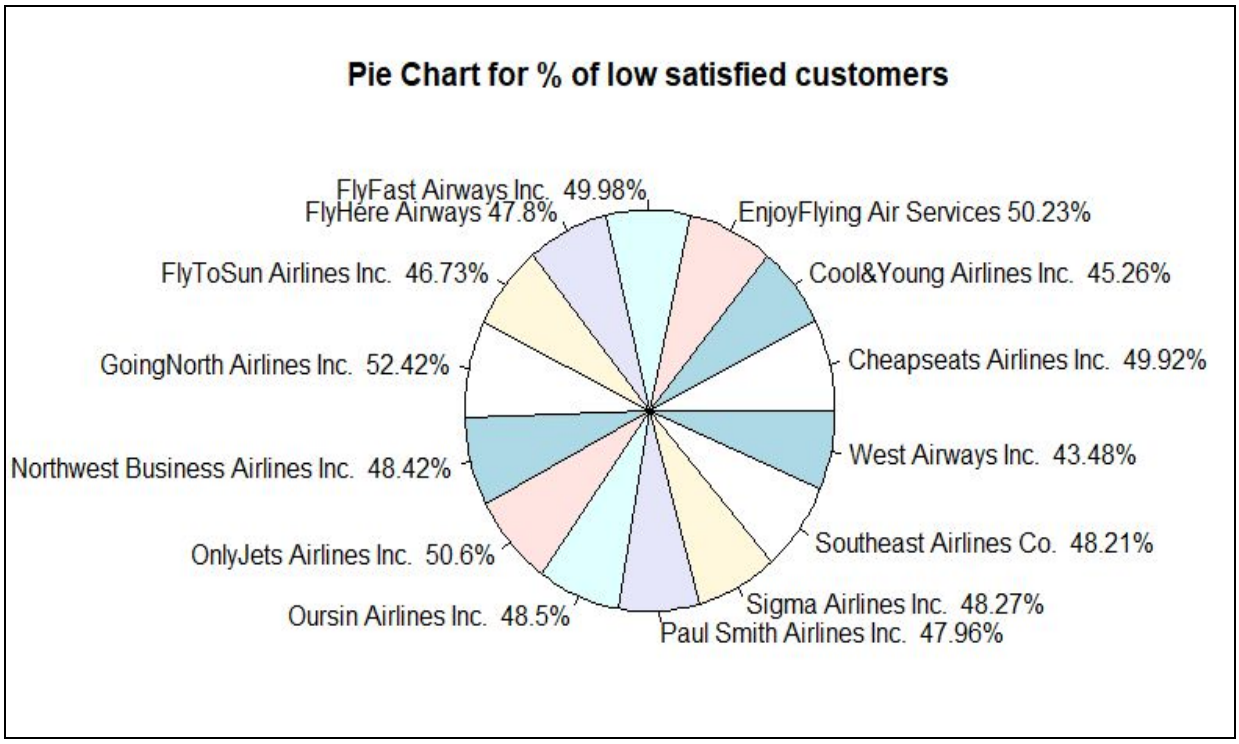
A. Performing data munging on the entire data set to eliminate redundant & undefined values.
B. Applying filters on the data set to retrieve a new set of data specific to one particular airline which has the highest number of low satisfied customers.
C. Performing descriptive statistics on the retrieved new set of data to obtain visuals which tell the basic insights & relationships between the various attributes.
D. Identification of the models which help in determining the attributes which affect satisfaction rate of the customers. In this project one such model is used: Linear Modeling
E. Using association rules mining technique which gives us the specific attributes & values which are responsible for the low or high satisfaction of customers.
F. Using a predictive technique which would predict the future values of customer satisfaction in order to draw inferences to perform modifications in airline services.
G. Providing recommendations based on the insights which would increase the overall customer satisfaction of the entire data set.

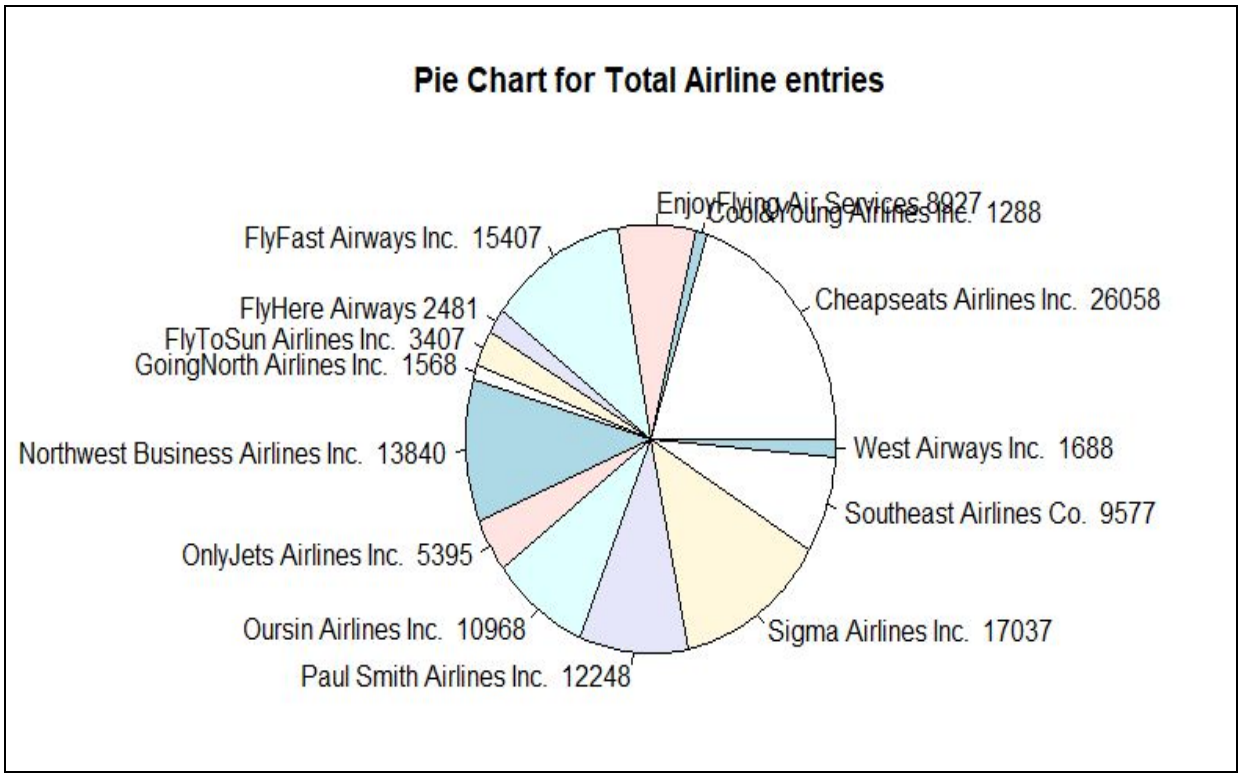
4. Business Questions addressed

A. Analyze the diverse set of customer base that the airline possesses
B. Understand the factors affecting the promoters of the airlines
C. Analyze the factors affecting the detractors of the airlines
D. Determine the go-to market strategies to improve net promoter score of airlines
E. Find out states to start implementing solutions which can increase net promoter score

5. Why Cheapseat airlines?

As per the below pie chart showing percentage of low satisfaction, Cheapseats is among the top 3. Also, 20% of the total data of the parent airway caters to Cheapseats airlines, hence, improving the satisfaction of Cheapseats airlines increases the overall satisfaction of the parent airway.


Pie Chart for % of low satisfied customers
FlyFast Airways Inc. 49.98%
FlyHere Airways 47.8%
EnjoyFlying Air Services 50.23%
FlyToSun Airlines Inc. 46.73%
Cool&Young Airlines Inc. 45.26%
GoingNorth Airlines Inc. 52.42%
Cheapseats Airlines Inc. 49.92%
Northwest Business Airlines Inc. 48.42%
West Airways Inc. 43.48%
OnlyJets Airlines Inc. 50.6%
Southeast Airlines Co. 48.21%
Oursin Airlines Inc. 48.5%
Sigma Airlines Inc. 48.27%
Paul Smith Airlines Inc. 47.96%


Pie Chart for Total Airline entries
EnjoyFlying Air Services 8927
Cool&Young Airlines Inc. 1288
FlyFast Airways Inc. 15407
Cheapseats Airlines Inc. 26058
FlyHere Airways 2481
FlyToSun Airlines Inc. 3407
GoingNorth Airlines Inc. 1568
Northwest Business Airlines Inc. 13840
West Airways Inc. 1688
Southeast Airlines Co. 9577
OnlyJets Airlines Inc. 5395
Oursin Airlines Inc. 10968
Sigma Airlines Inc. 17037
Paul Smith Airlines Inc. 12248

6. Data Requisition

Data was made available to us by the course instructors. The downloaded data was for the time period of January-March 2014. Before any kind of data munging, this data set consisted of approximately 129,889 rows and 28 variables. This data was extensively studied to determine usable variables. After initial analysis, the data set was forwarded to the preprocessing phase of the project for data munging.

Following is the code for data requisition:

```
5
6  data<-read.csv("Satisfaction Survey.csv", stringsAsFactors = FALSE)
7
```

Output obtained for the above code consists of data sets for the months from January 2014 through March 2014.

7. Data Preprocessing

The data sets contained numerous NA values, which were best dealt with by using dirty values. Omitting these rows would deprive us of valuable data insights that could be gained from these rows and using calculated values such as mean would misguide us in our observations. All columns of data have been kept, while a subset of the entire dataset has been picked out for the analysis. The chosen subset of data was that of Cheapseat Airlines and it had 26,058 rows.

In summary, the data preprocessing phase provided us with a subset of useable data for the project, consisting of only the rows that needed to be worked with and eliminating Null/NA values. The code for data preprocessing is given below.

Following is the code for data preprocessing:

```
apply(dataCleaned,2, function(x) any(is.na(x)))

# Now we come to know that only three columns contain NA values
# Those are 1. Departure.Delay.in.Minutes 2. Arrival.Delay.in.Minutes 3.Flight.time.in.minutes
# Rest columns are free from NA

# d<-dataCleaned

dataCleaned[is.na(dataCleaned)]<-9999 # Dirty value

apply(dataCleaned,2, function(x) any(is.na(x)))
```

```
# Replacing .(dots) by _(underscore) for easy analysis

name_column <- colnames(dataCleaned)
real_names <- gsub("\\.", "_", name_column)
colnames(dataCleaned) <- real_names
```

Output of the snippet consist of clean data set for the months from January 2014 through March 2014. These data sets consists of only the required variables and rows, and tackle the issue of NA values.

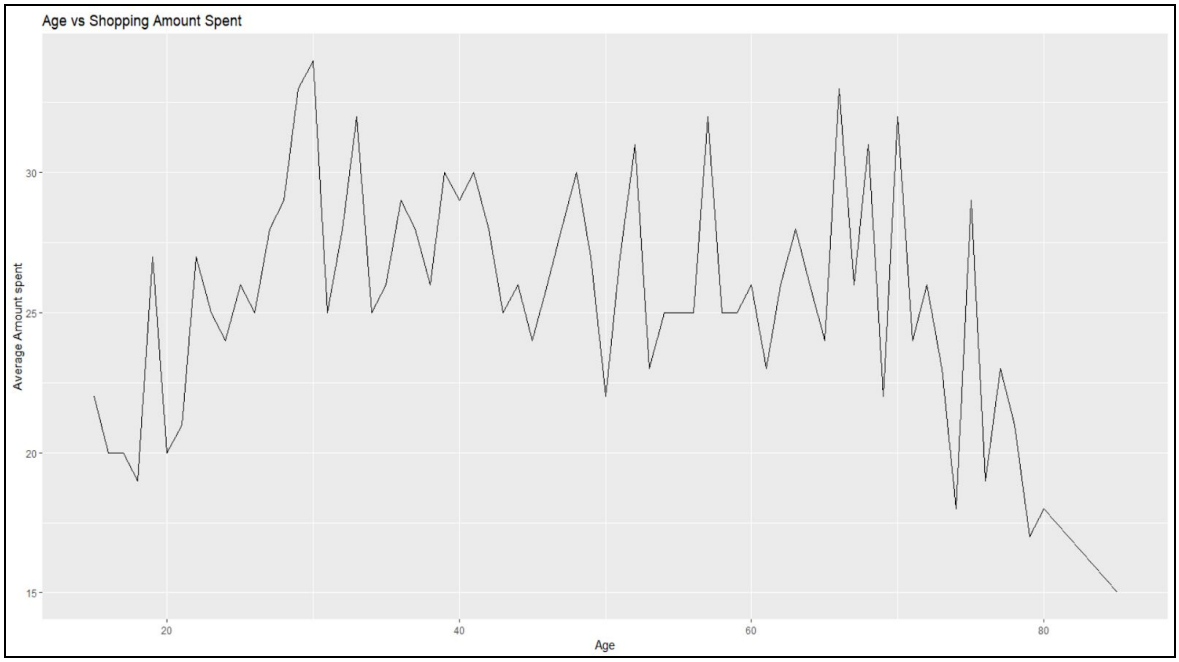
8. Descriptive Statistics

In this section we have explored the data set which we obtained after data munging i.e. the data set for Cheapseats airlines using the visualization techniques in R. Some of the visualization techniques are ggplot, maps.

Following are the statistics obtained -

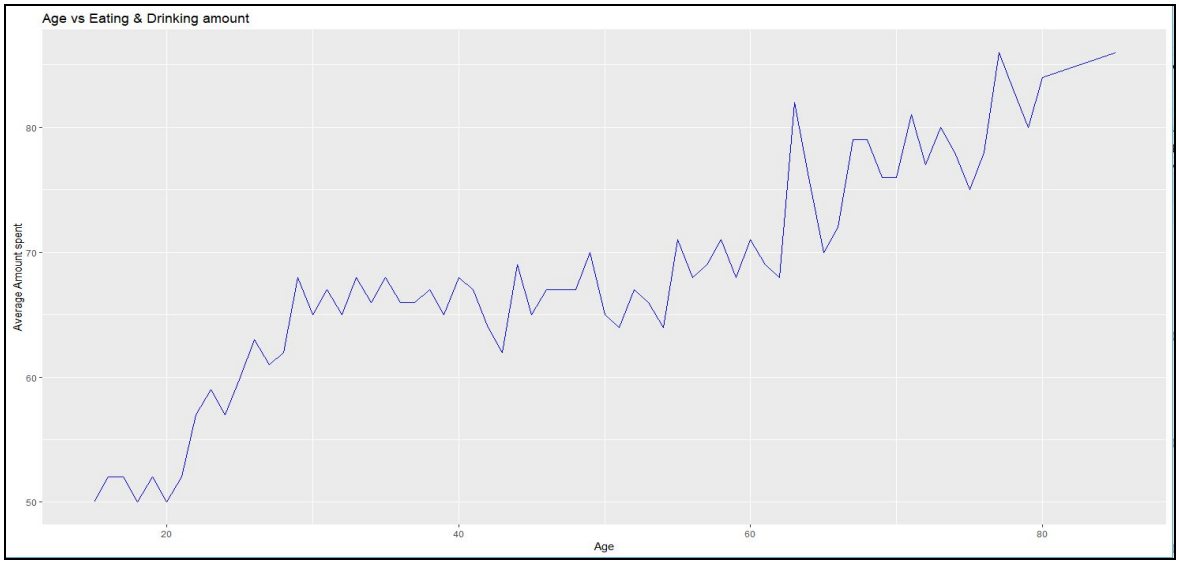
A. Line graph of Age VS Shopping Amount

```
ggplot(AgeVsShop, aes(x = AgeVsShop$Age, y = AgeVsShop$Shopping_Amount_at_Airport)) +
  geom_line(stat = "identity", color = "Black") +
  ggtitle("Age vs Shopping Amount Spent") +
  scale_x_continuous(name="Age") +
  scale_y_continuous(name = "Average Amount spent")
```



B. Line graph of Age VS Eating and Drinking Amount

```
ggplot(AvgVsE_D, aes(x = AvgVsE_D$Age , y = AvgVsE_D$Eating_and_Drinking_at_Airport )) +
  geom_line(stat = "identity", color = "Blue") +
  ggtitle("Age vs Eating & Drinking amount") +
  scale_x_continuous(name="Age") +
  scale_y_continuous(name = "Average Amount spent")
```

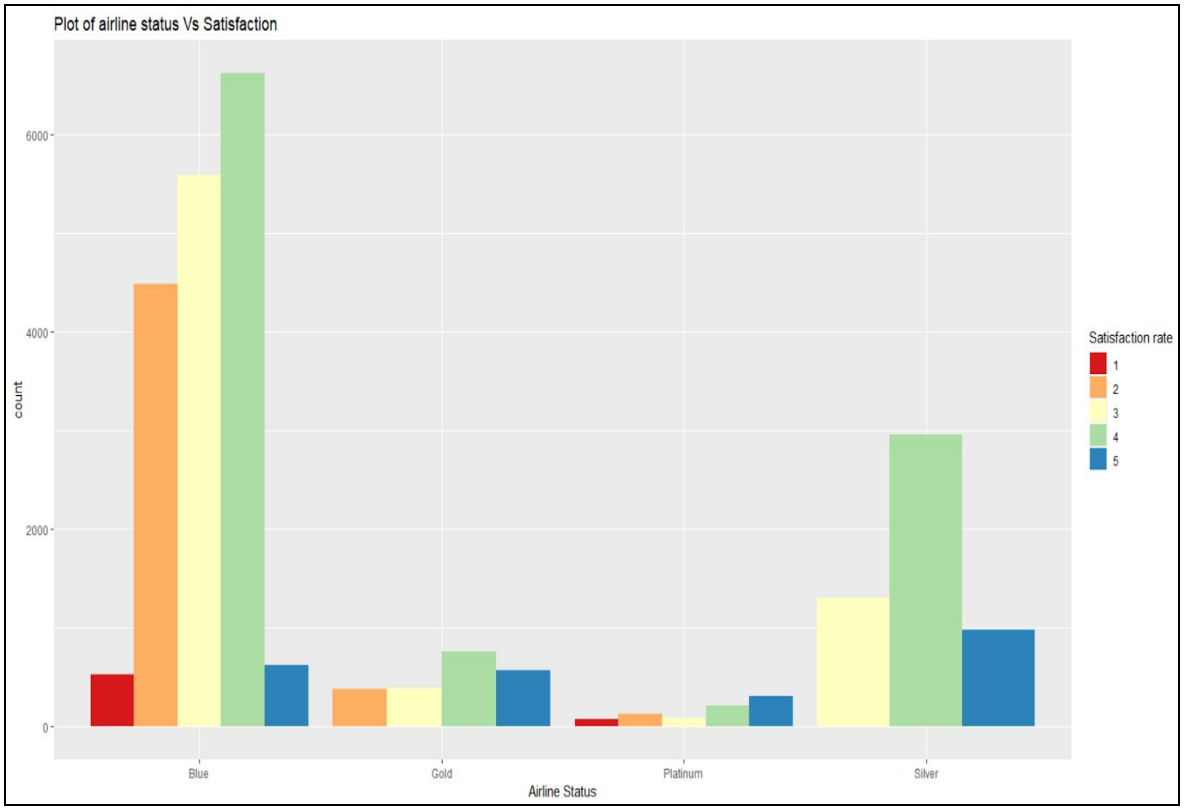


***Insight**:*

As the age increases, the amount of eating & drinking products purchased also increases.

C. Bar plot of Airline Status Vs Satisfaction

```
statusVsSatisfaction <- table(dataCleaned$Satisfaction, dataCleaned$Airline_Status)
statusVsSatisfaction_df <- data.frame(statusVsSatisfaction)
colnames(statusVsSatisfaction_df) <- c("Satisfaction", "Status", "Frequency")
statusVsSatisfaction_plot <- ggplot(statusVsSatisfaction_df, aes(y = jitter(statusVsSatisfaction_df$Frequency,20), x = statusVsSatisfaction_df$Status))
statusVsSatisfaction_plot <- statusVsSatisfaction_plot + geom_point(aes(color = statusVsSatisfaction_df$Satisfaction, size = 1)) +
  guides(color = guide_legend(title = "Satisfaction Rate"))
statusVsSatisfaction_plot <- statusVsSatisfaction_plot + scale_x_discrete("Airline Status") + scale_y_continuous("Frequency of each Status") +
  ggtitle("Plot of Airline Status Vs Satisfaction")
statusVsSatisfaction_plot
```

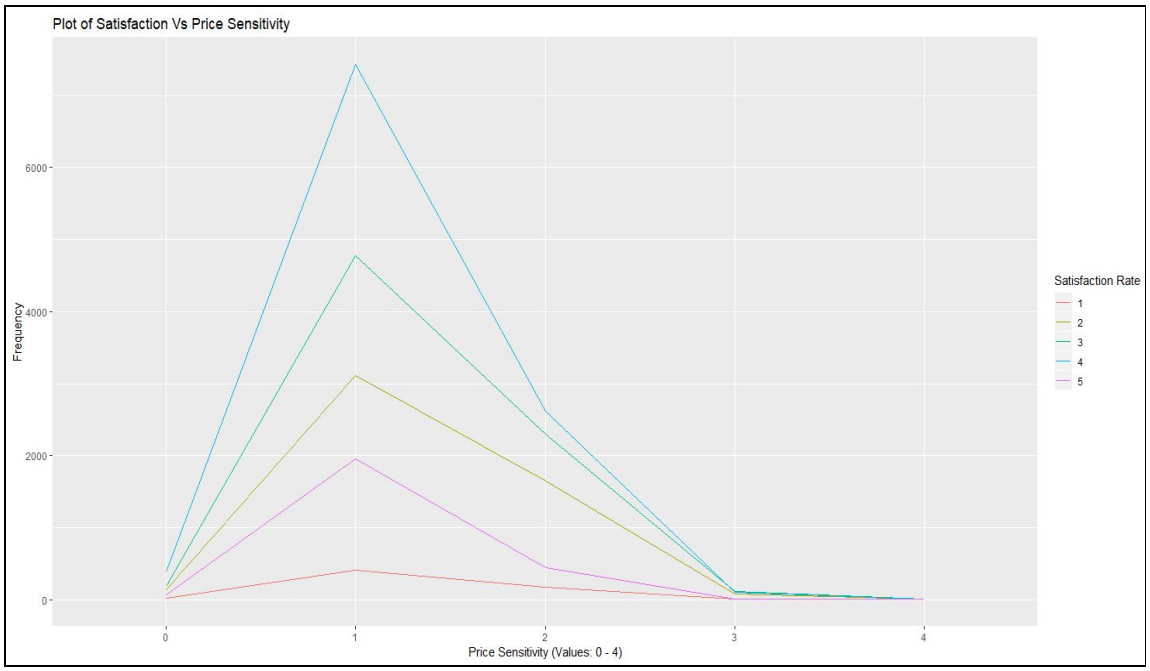


Insight:

The total number of low satisfied customers(with satisfaction rate 1,2 & 3) are more in Blue status of the airline.

D. Line Plot of Price Sensitivity Vs Satisfaction

```
satisfactionVsSensitivity <- table(dataCleaned$Satisfaction, dataCleaned$Price_Sensitivity)
satisfactionVsSensitivity_df <- data.frame(satisfactionVsSensitivity)
colnames(satisfactionVsSensitivity_df) <- c("Satisfaction", "Sensitivity", "Frequency")
satisfactionVsSensitivity_plot <- ggplot(satisfactionVsSensitivity_df, aes(x = Sensitivity, y = Frequency, group = Satisfaction))
satisfactionVsSensitivity_plot <- satisfactionVsSensitivity_plot + geom_line(aes(color = satisfactionVsSensitivity_df$Satisfaction)) +
  guides(color = guide_legend(title = "Satisfaction Rate"))
satisfactionVsSensitivity_plot <- satisfactionVsSensitivity_plot + scale_x_discrete("Price Sensitivity (Values: 0 - 4)") +
  scale_y_continuous("Frequency") + ggtitle("Plot of Satisfaction Vs Price Sensitivity")
satisfactionVsSensitivity_plot
```

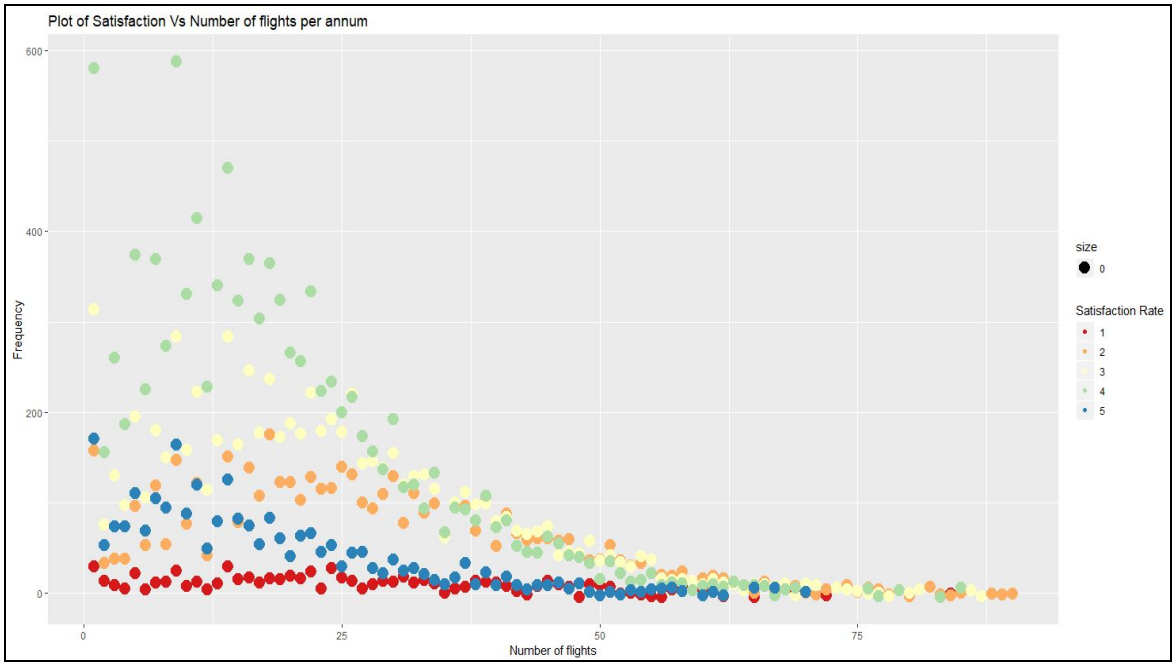


Insight:

The satisfaction rate does not depend on the price of tickets.

E. Scatter plot of Number of flights per annum Vs Satisfaction

```
noofflightsVsSatisfaction <- table(dataCleaned$Satisfaction, dataCleaned$No_of_Flights_p_a_)
noofflightsVsSatisfaction_df <- data.frame(noofflightsVsSatisfaction)
colnames(noofflightsVsSatisfaction_df) <- c("Satisfaction", "NumberOfFlights", "Frequency")
noofflightsVsSatisfaction_df$Satisfaction <- as.factor(noofflightsVsSatisfaction_df$Satisfaction)
noofflightsVsSatisfaction_df$NumberOfFlights <- as.numeric(noofflightsVsSatisfaction_df$NumberOfFlights)
noofflightsVsSatisfaction_df <- noofflightsVsSatisfaction_df[noofflightsVsSatisfaction_df$Frequency > 0,]
noofflightsVsSatisfaction_plot <- ggplot(noofflightsVsSatisfaction_df, aes(x = NumberOfFlights, y = jitter(Frequency, 30)))
noofflightsVsSatisfaction_plot <- noofflightsVsSatisfaction_plot + geom_point(aes(color = noofflightsVsSatisfaction_df$Satisfaction, size = 0)) +
  guides(color = guide_legend(title = "Satisfaction Rate")) + scale_color_brewer(palette = "Spectral")
noofflightsVsSatisfaction_plot <- noofflightsVsSatisfaction_plot + scale_x_continuous("Number of flights") + scale_y_continuous("Frequency") +
  ggtitle("Plot of Satisfaction Vs Number of flights per annum")
noofflightsVsSatisfaction_plot
```



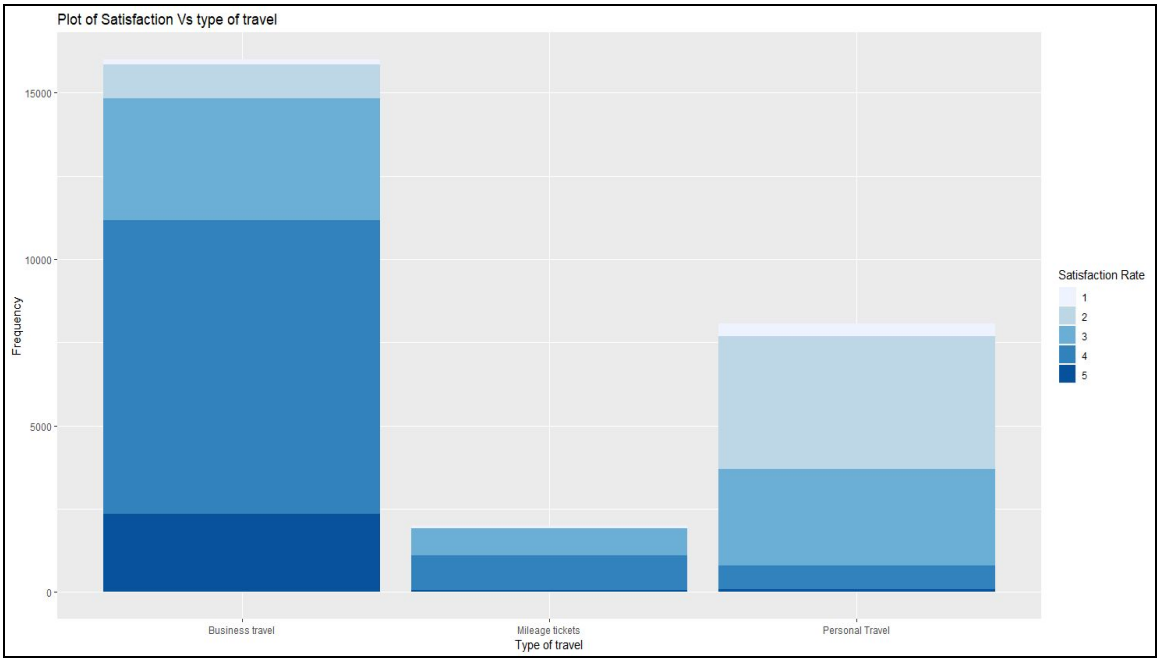
Insight:

As the number of flights taken per year increases, the total number of customer satisfaction decreases.

F. Type of travel Vs Satisfaction & Age

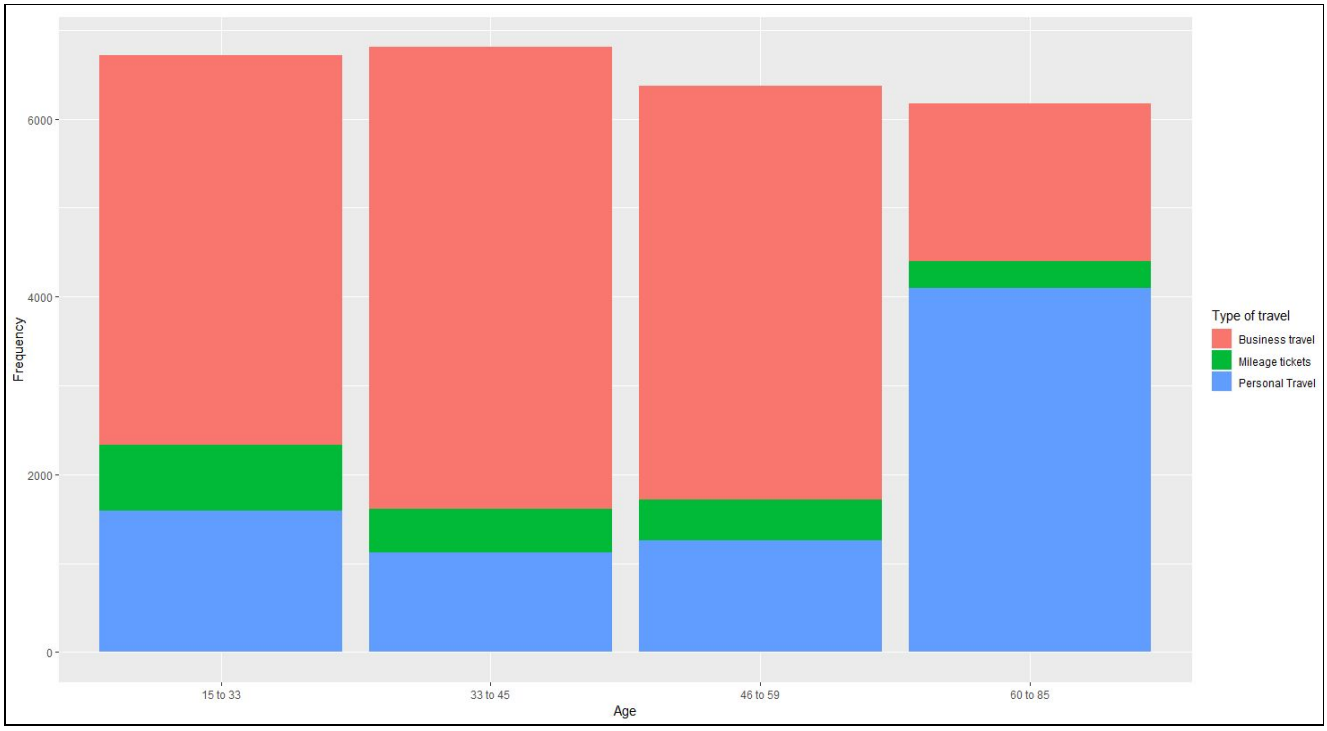
a.Type of travel Vs Satisfaction

```
traveltypeVsSatisfaction <- table(dataCleaned$Satisfaction, dataCleaned$Type_of_Travel)
traveltypeVsSatisfaction_df <- data.frame(traveltypeVsSatisfaction)
noofflightsVsSatisfaction_df$Satisfaction <- as.numeric(noofflightsVsSatisfaction_df$Satisfaction)
colnames(traveltypeVsSatisfaction_df) <- c("Satisfaction", "Type", "Frequency")
traveltypeVsSatisfaction_plot <- ggplot(traveltypeVsSatisfaction_df, aes(x = traveltypeVsSatisfaction_df$Type, y = jitter(Frequency, 30)))
traveltypeVsSatisfaction_plot <- traveltypeVsSatisfaction_plot + geom_bar(stat = "identity", aes(fill = traveltypeVsSatisfaction_df$Satisfaction)) +
  scale_fill_brewer(palette = "Blues") + guides(fill = guide_legend(title = "Satisfaction Rate"))
traveltypeVsSatisfaction_plot <- traveltypeVsSatisfaction_plot + scale_x_discrete("Type of travel") + scale_y_continuous("Frequency") +
  ggtitle("Plot of Satisfaction Vs type of travel")
traveltypeVsSatisfaction_plot
```



b.Type of travel Vs Age

```
dataPlot<-dataCleaned
head(ageTravel_df)
ggplot(data=)
q<-quantile(dataPlot$Age, c(0.25, 0.5, 0.75))
vec<-dataPlot$Age
vBuckets <- replicate(length(vec), "Average")
vBuckets[vec <= q[1]] <- "15 to 33"
vBuckets[vec > q[1] & vec<=q[2]] <- "33 to 45"
vBuckets[vec> q[2] & vec<=q[3]] <- "46 to 59"
vBuckets[vec>q[3]] <- "60 to 85"
dataPlot$Age<-as.factor(vBuckets)
str(dataPlot$Age)
ageTravel<-table(dataPlot$Age,dataPlot$Type_of_Travel)
ageTravel_df<-data.frame(ageTravel)
colnames(ageTravel_df)<-c("Age","Type_of_Travel","Frequency")
ggplot(ageTravel_df,aes(x=Age,y=Frequency))+
  geom_bar(stat = "identity",aes(fill = ageTravel_df$Type_of_Travel))+guides(fill = guide_legend(title = "Type of travel"))
  scale_fill_brewer(palette = "Spectral")
```



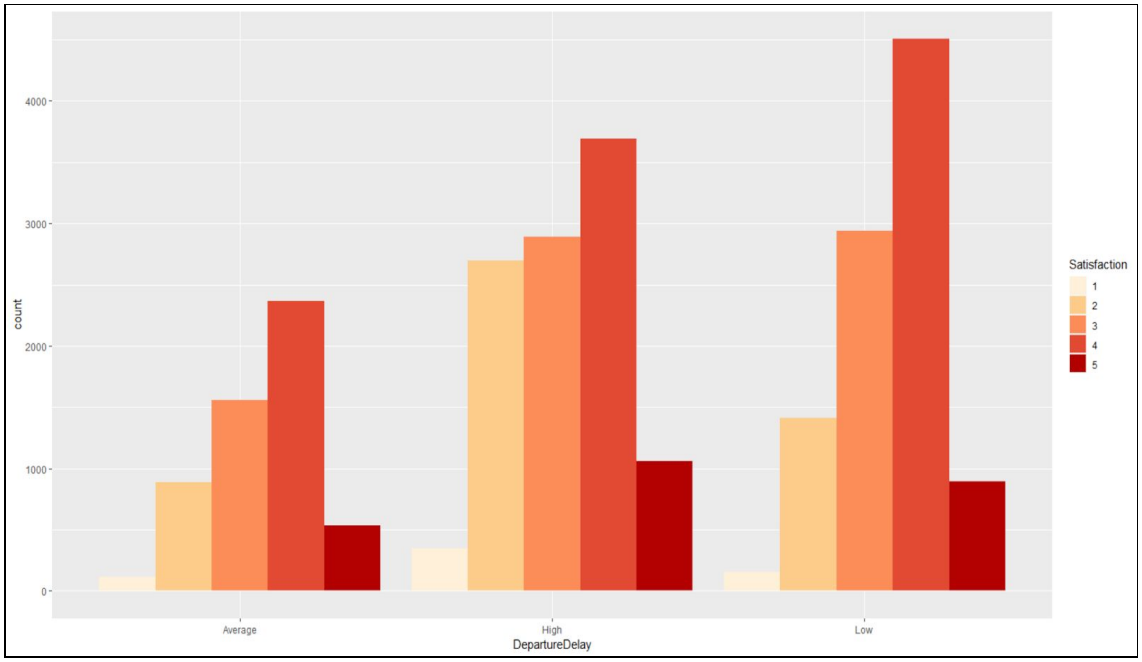
Insight:

Customers travelling for personal purposes are less satisfied and it mostly includes people who fall in the age group bracket of 60-85.

G. Barplot for Departure delay Vs Customer satisfaction

```
depdelay_df <- data.frame(dataCleaned$Satisfaction, dataCleaned$Departure_Delay_in_Minutes)
colnames(depdelay_df) <- c("Satisfaction", "DepartureDelay")
depdelay_func <- function(vec){
  q <- quantile(vec, c(0.4, 0.6))
  vBuckets <- replicate(length(vec), "Average")
  vBuckets[vec >= q[2]] <- "High"
  vBuckets[vec < q[1]] <- "Low"
  return(vBuckets)
}
str(depdelay_df)
depdelay_df$Satisfaction <- as.factor(depdelay_df$Satisfaction)
depdelay_df$DepartureDelay <- depdelay_func(depdelay_df$DepartureDelay)

ggplot(depdelay_df, aes(x =DepartureDelay, fill =Satisfaction)) + geom_bar(position = "dodge") +
  scale_fill_brewer(palette = "OrRd")
```

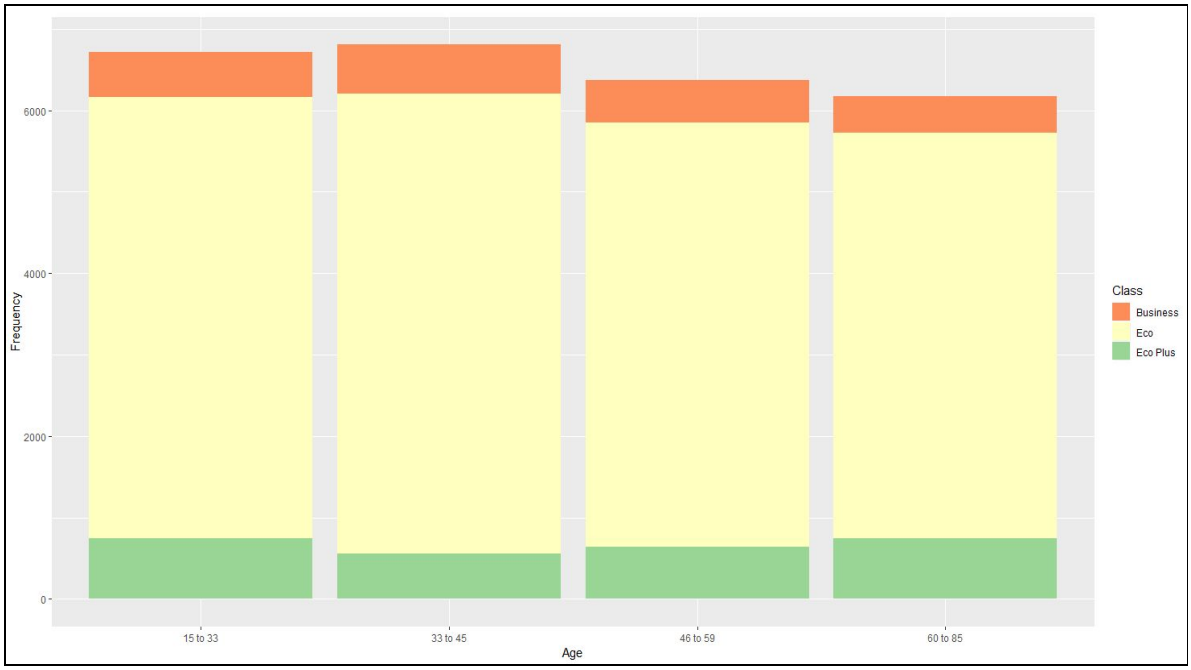


Insight:

If the departure delay is high, the rate of satisfaction is high and departure delay is low, the rate of satisfaction is low.

H. Bar plot of Age Vs Class

```
ageClass<-table(dataPlot$Age,dataPlot$Class)
ageClass_df<-data.frame(ageClass)
colnames(ageClass_df)<-c("Age","Class","Frequency")
ggplot(ageClass_df,aes(x=Age,y=Frequency))+
  geom_bar(stat = "identity",aes(fill = ageClass_df$Class))+
  scale_fill_brewer(palette = "Spectral") +
  guides(fill=guide_legend(title="Class"))
```



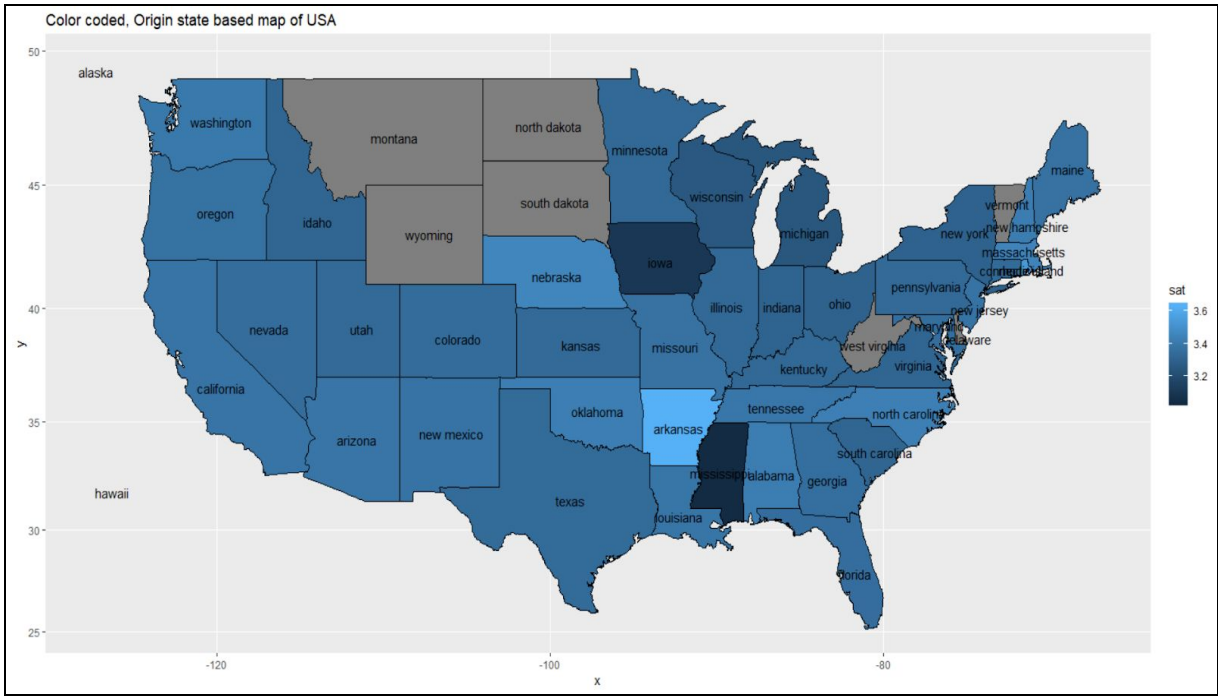
Insight:

Majority of the people are in the age group of 33 to 45 & maximum of them travel by Economy class.

I. Color coded map for Satisfaction Vs Origin State & Destination state

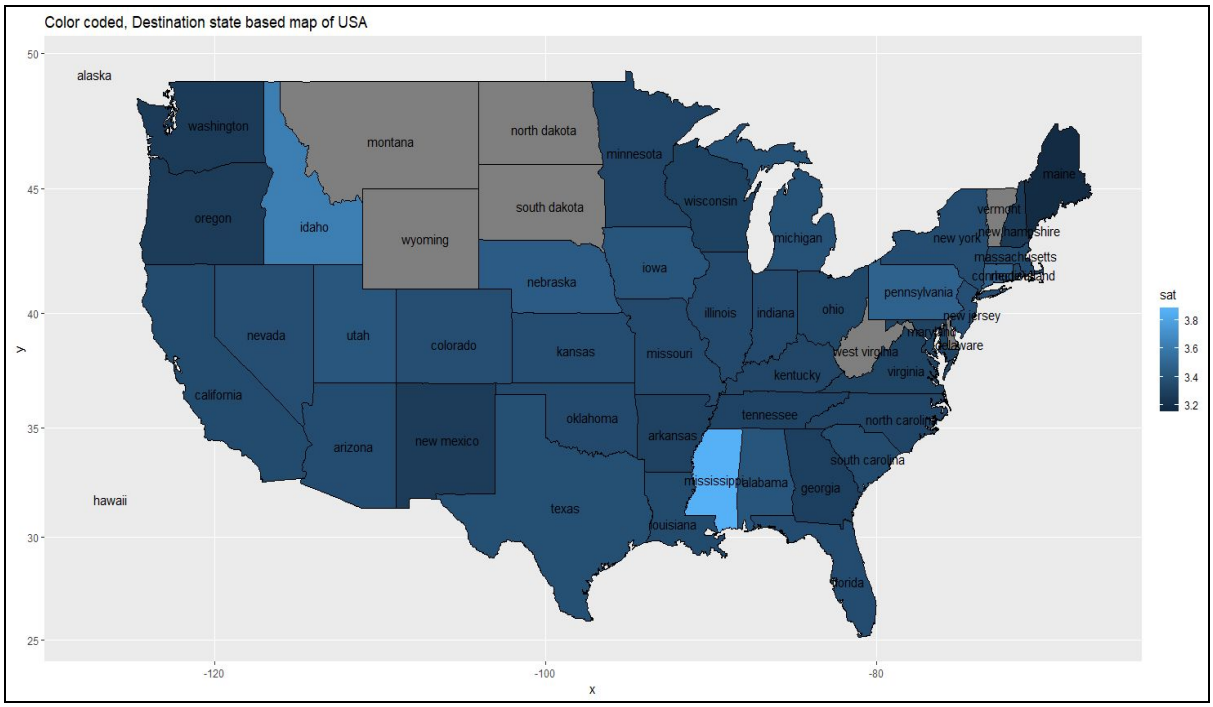
a. Map for Origin State Vs Satisfaction

```
stateSat<-sqldf("Select avg(Satisfaction) as sat,Origin_State FROM dataCleaned GROUP BY Origin_State")
stateSat
stateSatDes<-sqldf("Select avg(Satisfaction) as sat,Destination_State FROM dataCleaned GROUP BY Destination_State")
stateSatDes
stateSat$stateName<-tolower(stateSat$Origin_State)
stateSatDes$stateName<-tolower(stateSatDes$Destination_State)
USAStateA<-data.frame(state.name,state.center,state.area)
USAStateA$stateName<-tolower(USAStateA$state.name)
library(ggplot2)
merged_df<-merge(USAStateA,stateSat,all.x=TRUE)
merged_df
merged_dfDes<-merge(USAStateA,stateSatDes,all.x=TRUE)
merged_dfDes
ggplot(merged_df , aes(map_id=stateName)) +
  geom_map(map=us,aes(fill=sat), color="black") +
  expand_limits(x=us$long,y=us$lat) +
  coord_map() +
  ggtitle("Color coded, Origin state based map of USA")+
  geom_text(aes(x=merged_df$x,y=merged_df$y,label=merged_df$stateName))
```



b.Map for Destination state Vs Satisfaction

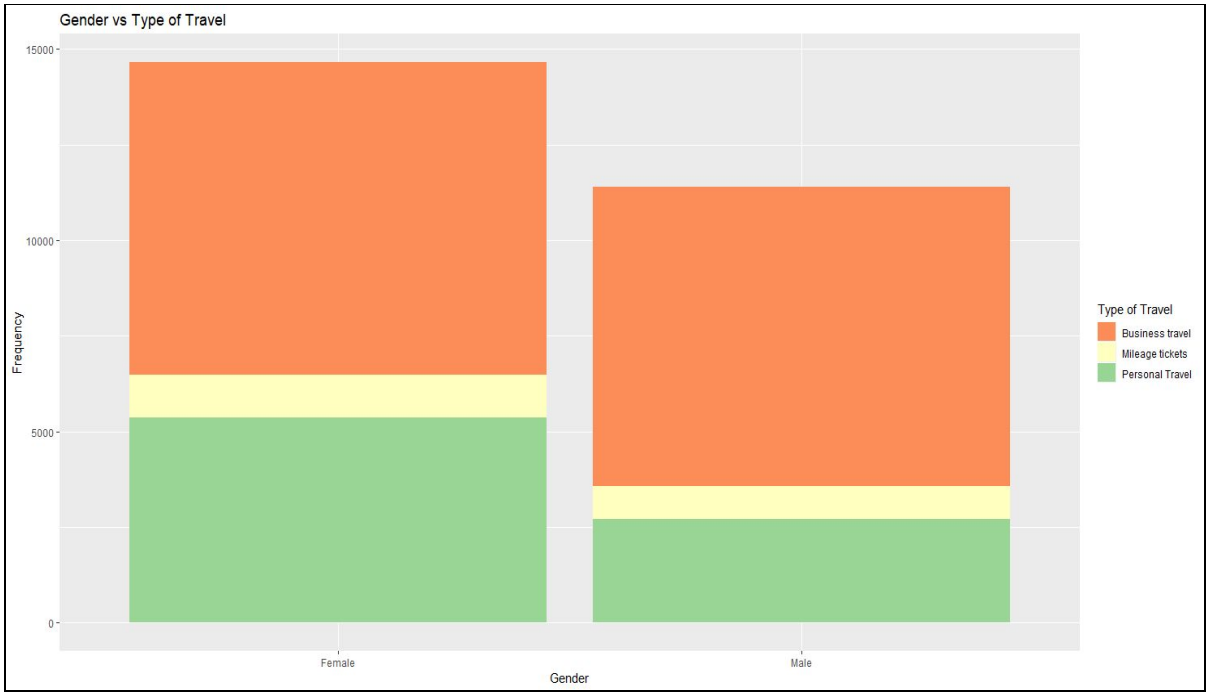
```
ggplot(merged_dfDes , aes(map_id=stateName )) +
  geom_map(map=us,aes(fill=sat), color="black") +
  expand_limits(x=us$long,y=us$lat) +
  coord_map() +
  ggtitle("Color coded, Destination state based map of USA")+
  geom_text(aes(x=merged_df$x,y=merged_df$y,label=merged_df$stateName))
```



J. Plots of Gender with Type of travel, Class, Eating & Drinking Amount, Shopping Amount

a. Gender Vs Type of Travel

```
GenderTypeOfTravel <-table(dataPlot$Gender, dataPlot$Type_of_Travel)
GenderTypeOfTravel_df<-data.frame(GenderTypeOfTravel)
colnames(GenderTypeOfTravel_df)<-c("Gender","Type_of_Travel","Frequency")
ggplot(GenderTypeOfTravel_df,aes(x= Gender,y=Frequency))+
  geom_bar(stat = "identity",aes(fill = GenderTypeOfTravel_df$Type_of_Travel))+
  scale_fill_brewer(palette = "Spectral") +
  guides(fill=guide_legend(title="Type of Travel"))+
  ggtitle("Gender vs Type of Travel")
```

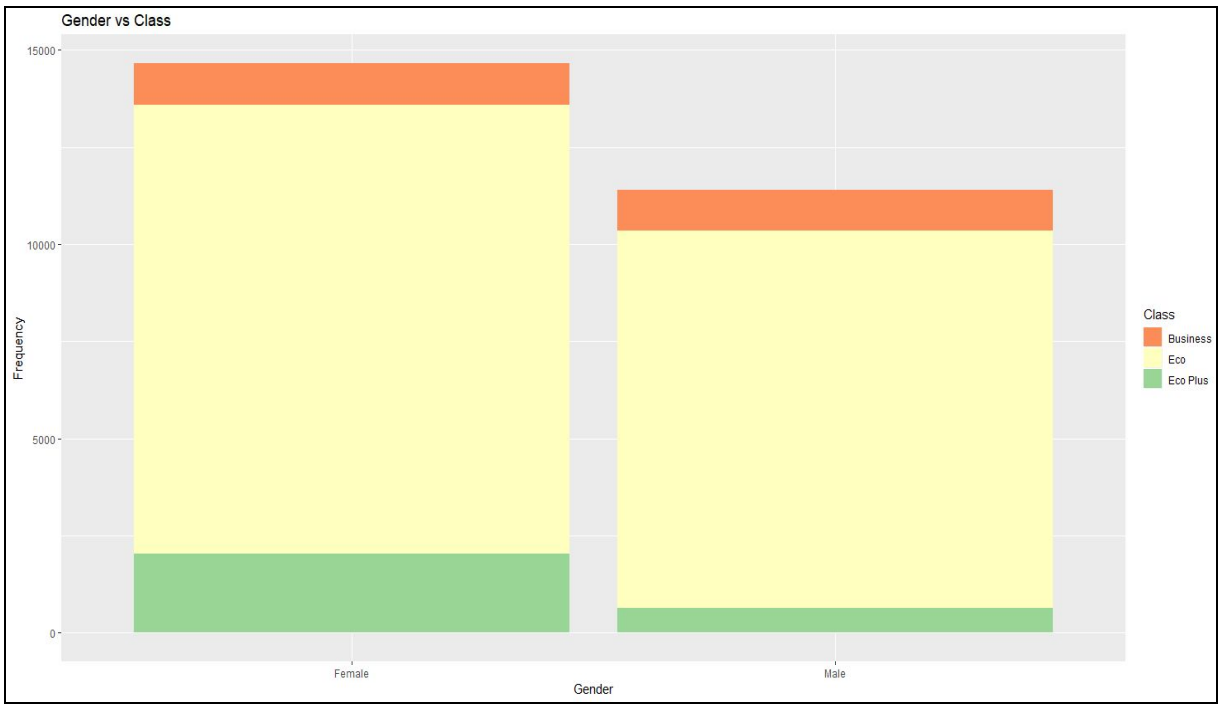


Insight:

Firstly, there are more female passengers travelling and most of them are travelling for personal purposes.

b. Gender Vs Class

```
GenderClass<-table(dataPlot$Gender,dataPlot$Class)
GenderClass_df<-data.frame(GenderClass)
colnames(GenderClass_df)<-c("Gender","Class","Frequency")
ggplot(GenderClass_df ,aes(x= Gender,y=Frequency))+
  geom_bar(stat = "identity",aes(fill = GenderClass_df$Class ))+
  scale_fill_brewer(palette = "Spectral") +
  guides(fill=guide_legend(title="Class"))+
  ggtitle("Gender vs Class")
```



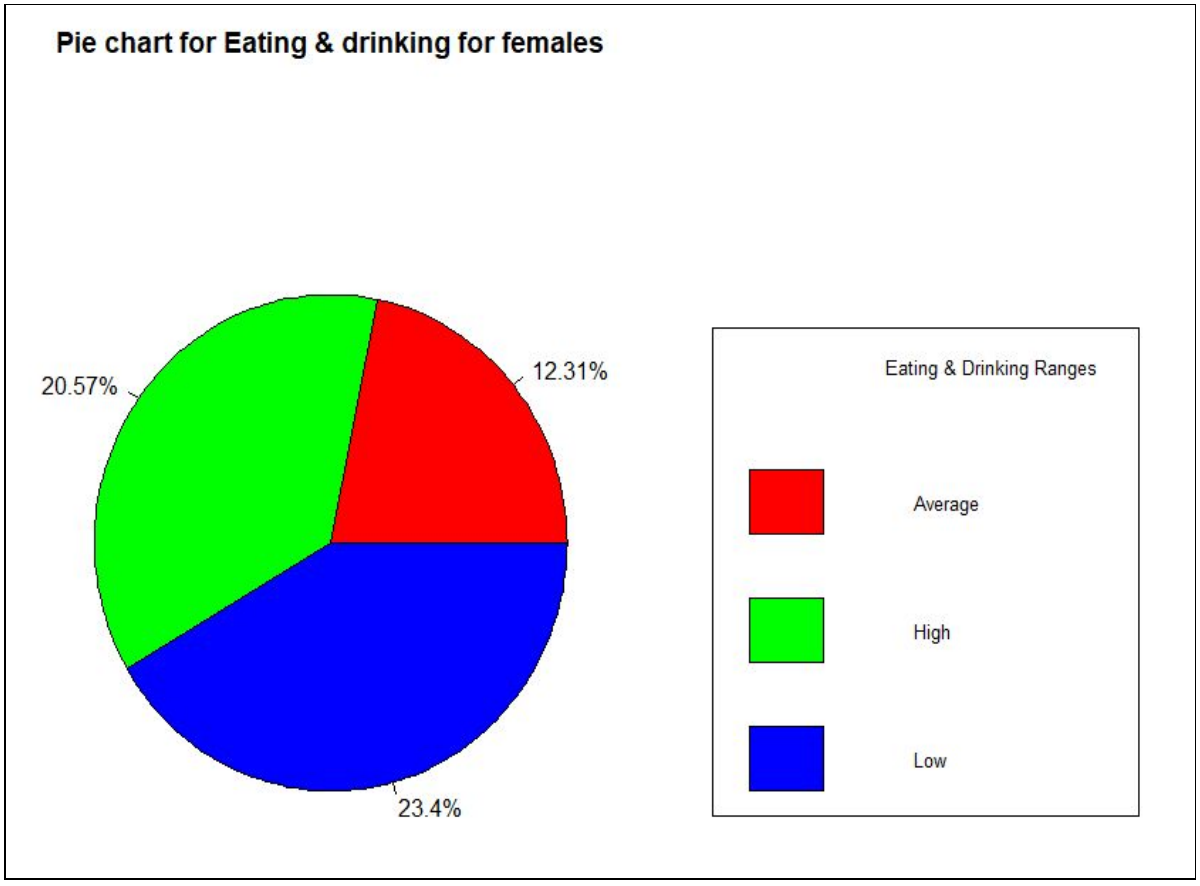
Insight:

Maximum number of people are travelling by Economy class followed by Economy Plus & then Business.

c. Gender Vs Amount of Eating & Drinking products

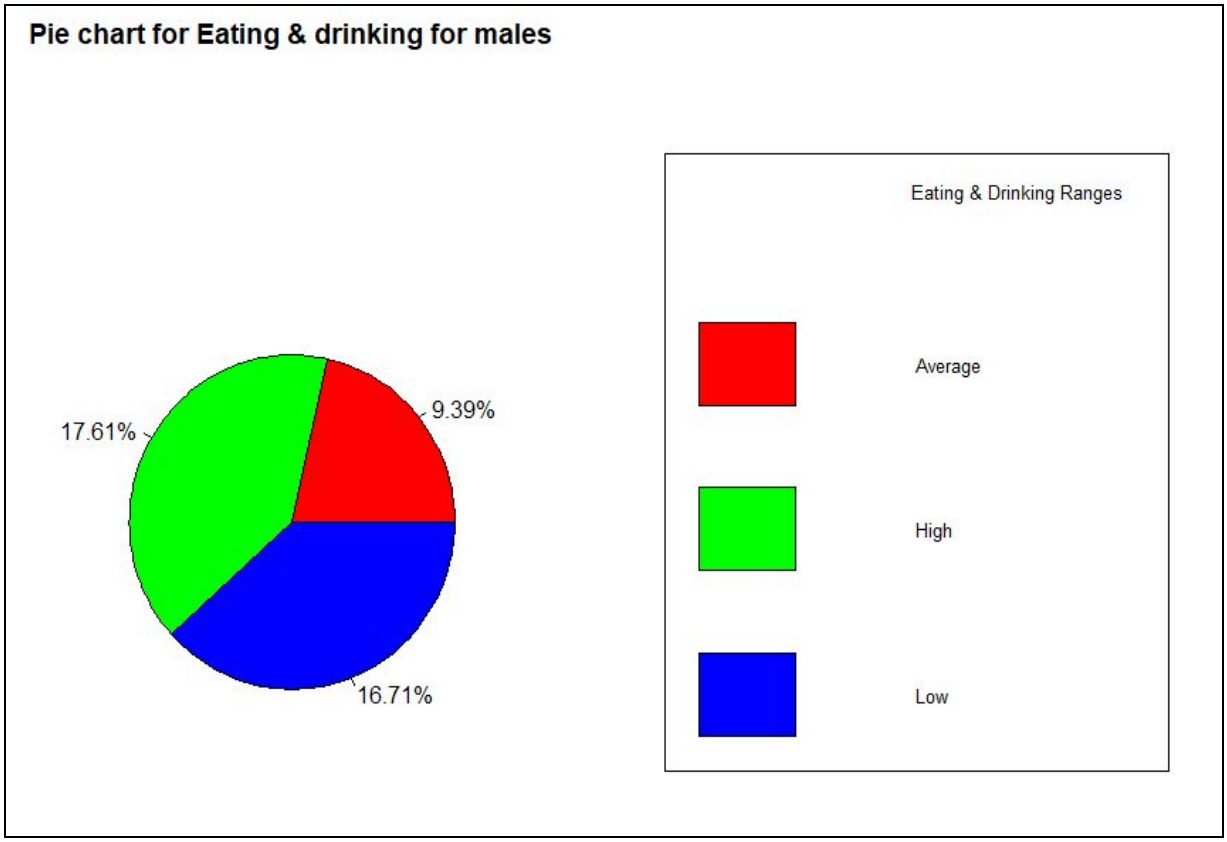
i. For Females:

```
total <- sum(GenderEatDrink_df$Frequency)
GenderEatDrink_df$Gender <- as.character(GenderEatDrink_df$Gender)
GenderEatDrink_df$Eating_and_Drinking <- as.character(GenderEatDrink_df$Eating_and_Drinking)
femaleCount_df <- GenderEatDrink_df[GenderEatDrink_df$Gender == "Female",]
femaleCount <- sum(femaleCount_df$Frequency)
femaleCount_df$TotalCount <- total
femaleCount_df$Percentage <- (femaleCount_df$Frequency/femaleCount_df$TotalCount)*100
femaleCount_df$Percentage <- round(femaleCount_df$Percentage, 2)
slicesGenderEatDrink <- femaleCount_df$Percentage
lblsGenderEatDrink <- femaleCount_df$Percentage
lblsGenderEatDrink <- paste(lblsGenderEatDrink,"%", sep = "")
pie(slicesGenderEatDrink, labels = lblsGenderEatDrink, col = rainbow(length(lblsGenderEatDrink)), main = "Pie chart for Eating & drinking for females")
legend("bottomleft", c("Average", "High", "Low"),cex = 0.8, fill = rainbow(length(lblsGenderEatDrink)), title = "Eating & Drinking Ranges")
```



ii. For Males:

```
total <- sum(GenderEatDrink_df$Frequency)
GenderEatDrink_df$Gender <- as.character(GenderEatDrink_df$Gender)
GenderEatDrink_df$Eating_and_Drinking <- as.character(GenderEatDrink_df$Eating_and_Drinking)
maleCount_df <- GenderEatDrink_df[GenderEatDrink_df$Gender == "Male",]
maleCount <- sum(femaleCount_df$Frequency)
maleCount_df$TotalCount <- total
maleCount_df$Percentage <- (maleCount_df$Frequency/maleCount_df$TotalCount)*100
maleCount_df$Percentage <- round(maleCount_df$Percentage, 2)
slicesGenderEatDrink <- maleCount_df$Percentage
lblsGenderEatDrink <- maleCount_df$Percentage
lblsGenderEatDrink <- paste(lblsGenderEatDrink,"%", sep = "")
pie(slicesGenderEatDrink, labels = lblsGenderEatDrink, col = rainbow(length(lblsGenderEatDrink)), main = "Pie chart for Eating & drinking for males")
legend("bottomleft", c("Average", "High", "Low"),cex = 0.8, fill = rainbow(length(lblsGenderEatDrink)), title = "Eating & Drinking Ranges")
```



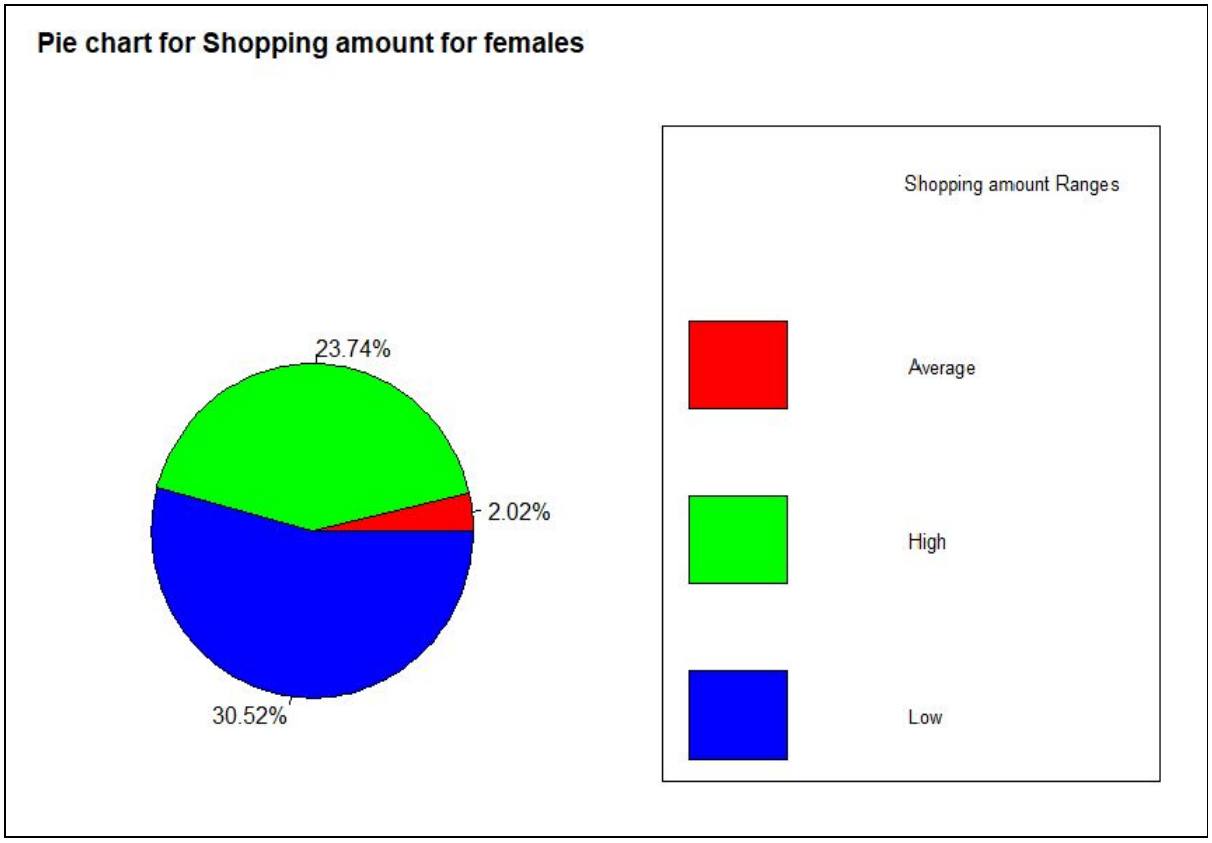
Insight:

Females spend more on eating & drinking compared to males.

d. Gender Vs Shopping Amount

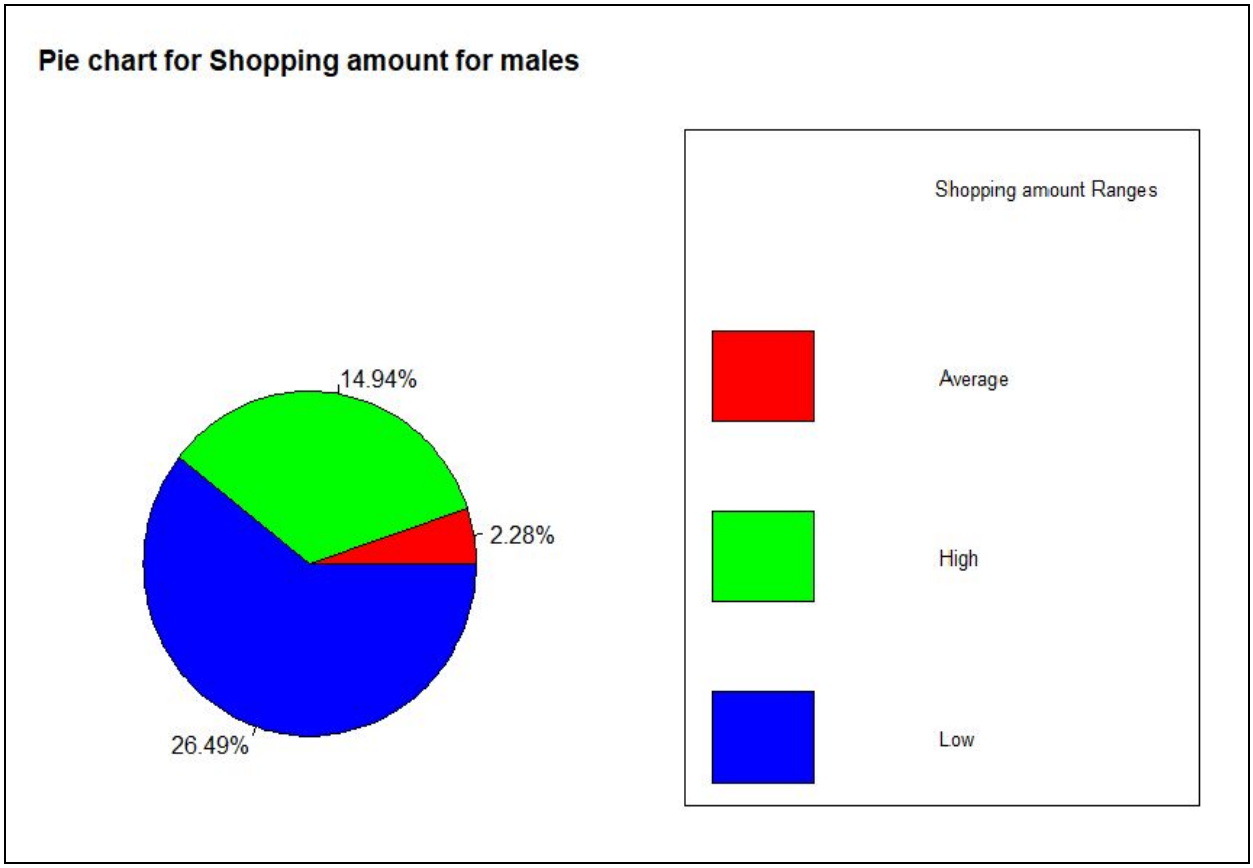
i.For Females

```
Gender <- dataCleaned$Gender
ShoppingAmount <- map(dataCleaned$Shopping_Amount_at_Airport)
GenderShop_df <- data.frame(table(Gender, ShoppingAmount))
colnames(GenderShop_df) <- c("Gender", "ShoppingAmount","Frequency")
total <- sum(GenderShop_df$Frequency)
GenderShop_df$Gender <- as.character(GenderShop_df$Gender)
femaleCount_df <- GenderShop_df[GenderShop_df$Gender == "Female",]
femaleCount <- sum(femaleCount_df$Frequency)
femaleCount_df$TotalCount <- total
femaleCount_df$Percentage <- (femaleCount_df$Frequency/femaleCount_df$TotalCount)*100
femaleCount_df$Percentage <- round(femaleCount_df$Percentage, 2)
slicesGenderShop <- femaleCount_df$Percentage
lblsGenderShop <- femaleCount_df$Percentage
lblsGenderShop <- paste(lblsGenderShop,"%", sep = "")
pie(slicesGenderShop, labels = lblsGenderShop, col = rainbow(length(lblsGenderShop)), main = "Pie chart for Shopping amount for females")
legend("bottomleft", c("Average", "High", "Low"),cex = 0.8, fill = rainbow(length(lblsGenderShop)), title = "Shopping amount Ranges")
```



ii.For males

```
maleCount_df <- GenderShop_df[GenderShop_df$Gender == "Male",]
maleCount <- sum(maleCount_df$Frequency)
maleCount_df$TotalCount <- total
maleCount_df$Percentage <- (maleCount_df$Frequency/maleCount_df$TotalCount)*100
maleCount_df$Percentage <- round(maleCount_df$Percentage, 2)
slicesGenderShop <- maleCount_df$Percentage
lblsGenderShop <- maleCount_df$Percentage
lblsGenderShop <- paste(lblsGenderShop,"%", sep = "")
pie(slicesGenderShop, labels = lblsGenderShop, col = rainbow(length(lblsGenderShop)), main = "Pie chart for Shopping amount for males")
legend("bottomleft", c("Average", "High", "Low"),cex = 0.8, fill = rainbow(length(lblsGenderShop)), title = "Shopping amount Ranges")
```



Insight:

Females spend more amount on shopping compared to males.

9. Modelling Techniques and Visualizations

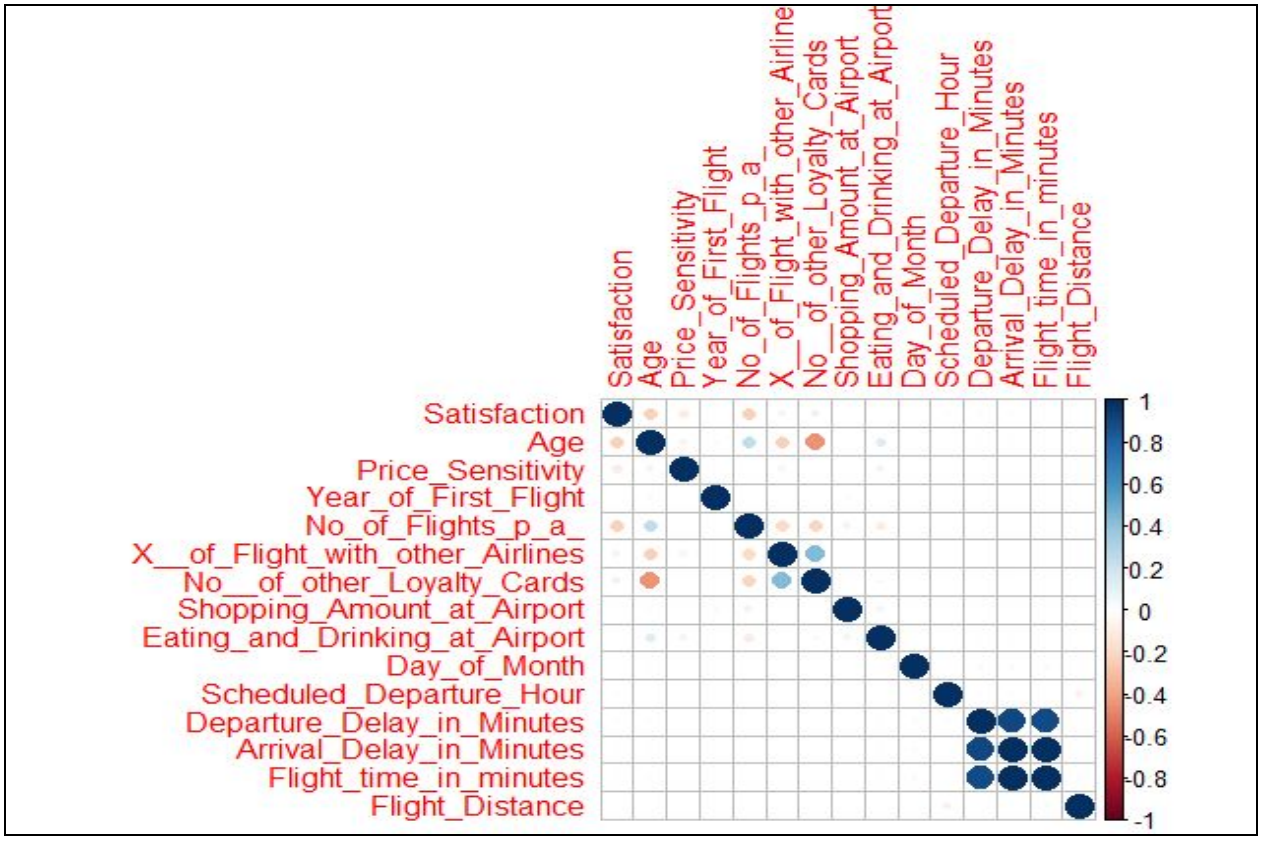
9.1 Linear modelling

In order to determine the factors affecting the satisfaction of Cheapseats customers, linear modelling was performed on all attributes of the dataset.

```
options(scipen = 100)
View(dataCleaned)
allattributemodel<-lm(formula = Satisfaction~., data = dataCleaned)
summary(allattributemodel)
```

On performing linear modelling on the all the attributes, 7 attributes were determined which are most statistically significant. They are:

1. Airline.Status
2. Age
3. Gender
4. Price.Sensitivity
5. No.of.Flights.p.a.
6. Type.of.Travel
7. Class
8. Year of first flight



On performing the linear modelling on 8 attributes, R square value of 0.422 was achieved.

In order to analyse the factors affecting number of flights per annum, linear modelling was performed. 3 new attributes, Number of other loyalty cards, shopping amount, eating and drinking at airport, came up as statistically significant as shown below:

```
87  m4 <- lm(No.of.Flights.p.a. ~ .
88           , data=dataCleaned)
89  summary(m4)
90
```

92:1 (Top Level)

Console B:/Syracuse/Data Science/Project details/

Age	0.1010118	0.0055677	18.143	< 0.0000000000000002	***
GenderMale	2.5096209	0.1686911	14.877	< 0.0000000000000002	***
Price.Sensitivity	-0.5997573	0.1504899	-3.985	0.00006756073536	***
Year.of.First.Flight	0.0513464	0.0271168	1.894	0.058299	.
X..of.Flight.with.other.Airlines	-0.1367771	0.0103395	-13.229	< 0.0000000000000002	***
Type.of.Travel2	-6.2730851	0.2359973	-26.581	< 0.0000000000000002	***
Type.of.Travel3	-1.8783338	0.3586668	-5.237	0.00000016449521	***
No..of.other.Loyalty.Cards	-1.3400414	0.0852694	-15.715	< 0.0000000000000002	***
Shopping.Amount.at.Airport	-0.0137115	0.0015338	-8.940	< 0.0000000000000002	***
Eating.and.Drinking.at.Airport	-0.0316512	0.0015789	-20.046	< 0.0000000000000002	***
ClassEco	-0.1117944	0.2954521	-0.378	0.705148	
ClassEco Plus	-3.3406735	0.3793506	-8.806	< 0.0000000000000002	***

Interesting insight that can be seen here is that the number of loyalty cards, shopping amount at airport and eating and drinking at airport affect frequent travellers. One other interesting attribute that was seen is Percentage of flight with other airlines.

On using linear model to see the factors affecting the percentage of flight with other airlines, interesting insight was found that, it is affected by Number of loyalty cards.

```
m3 <- lm(X..of.Flight.with.other.Airlines ~ .
         , data=dataCleaned)
summary(m3)
#Insight: New attribute that affects %of flight with other airlines is No of other loyalty cards
```

To understand the attributes affecting the attribute, Number of loyalty cards linear modelling was done as below:

```
96  m6 <- lm(formula = No..of.other.Loyalty.Cards ~ ., data=dataCleaned)
97  summary(m6)
99:1  (Top Level)

Console  B:/Syracuse/Data Science/Project details/
Satisfaction5                  0.06633662  0.04481497   1.480              0.138823
Airline.Status1               -0.05428193  0.01597377  -3.398              0.000679 ***
Airline.Status2               -0.12238247  0.02239471  -5.465        0.0000000468 ***
Airline.Status3               -0.05480990  0.03463878  -1.582              0.113587
Age                           -0.02365021  0.00037919 -62.370 < 0.0000000000000002 ***
GenderMale                     0.00279462  0.01229738   0.227              0.820229
Price.Sensitivity             -0.04282484  0.01092400  -3.920        0.0000886840 ***
Year.of.First.Flight           0.00415466  0.00196835   2.111              0.034805 *
No.of.Flights.p.a.            -0.00706087  0.00044930 -15.715 < 0.0000000000000002 ***
X..of.Flight.with.other.Airlines  0.04404112  0.00070149  62.782 < 0.0000000000000002 ***
Type.of.Travel2               -0.08935986  0.01735437  -5.149        0.0000002636 ***
Type.of.Travel3               -0.00068030  0.02604903  -0.026              0.979165
Shopping.Amount.at.Airport     0.00002935  0.00011151   0.263              0.792365
Eating.and.Drinking.at.Airport -0.00023788  0.00011549  -2.060              0.039430 *
ClassEco                       0.00253287  0.02144659   0.118              0.905988
ClassEco Plus                 -0.09478330  0.02757160  -3.438              0.000588 ***
```

No new insight was found on this other than the fact that male customers are not so concerned about loyalty cards.

Linear modelling was performed on the attribute, Type of Travel to understand if any major attributes are contributing to it.

```
92  m5 <- lm(Type.of.Travel ~ .
93           , data=dataCleaned)
94  summary(m5)
96:1  (Top Level)

Console  B:/Syracuse/Data Science/Project details/
Coefficients: (82 not defined because of singularities)
                                 Estimate  Std. Error t value             Pr(>|t|)
(Intercept)                   -3.45068345  2.89644461  -1.191             0.233527
Satisfaction2                 -0.27928838  0.01996538 -13.989 < 0.0000000000000002 ***
Satisfaction3                  0.23823884  0.01977795  12.046 < 0.0000000000000002 ***
Satisfaction4                  0.53097928  0.01978147  26.842 < 0.0000000000000002 ***
Satisfaction5                  0.49015627  0.02165221  22.638 < 0.0000000000000002 ***
Airline.Status1               -0.14651829  0.00785738 -18.647 < 0.0000000000000002 ***
Airline.Status2               -0.01118335  0.01109645  -1.008             0.313545
Airline.Status3                0.05922897  0.01714458   3.455             0.000552 ***
Age                           -0.00575628  0.00019734 -29.170 < 0.0000000000000002 ***
GenderMale                     0.06064083  0.00606670   9.996 < 0.0000000000000002 ***
Price.Sensitivity             -0.01965836  0.00539696  -3.642             0.000271 ***
Year.of.First.Flight           0.00288920  0.00097461   2.964             0.003035 **
No.of.Flights.p.a.            -0.00276054  0.00022022 -12.535 < 0.0000000000000002 ***
X..of.Flight.with.other.Airlines -0.00211088  0.00037272  -5.663       0.000000015 ***
No..of.other.Loyalty.Cards    -0.00498103  0.00307849  -1.618             0.105673
Shopping.Amount.at.Airport    -0.00003874  0.00005523  -0.701             0.483058
Eating.and.Drinking.at.Airport -0.00023045  0.00005716  -4.032         0.000055469 ***
```

Factors affecting the airline.status based on results of linear model:

```
 99  m7 <- lm(formula = Airline.Status ~ ., data=dataCleaned)
100  summary(m7)
102:1  (Top Level)

Console  B:/Syracuse/Data Science/Project details/
Age                                  0.003836224  0.000309341  12.400 < 0.0000000000000002 ***
GenderMale                          -0.000001645  0.009364867   0.000              0.99986
Price.Sensitivity                   -0.051135110  0.008329650  -6.139         0.000000000843 ***
Year.of.First.Flight                 0.003752572  0.001501481   2.499              0.01245 *
No.of.Flights.p.a.                  -0.004527583  0.000343137 -13.195 < 0.0000000000000002 ***
X..of.Flight.with.other.Airlines     0.007474422  0.000572071  13.066 < 0.0000000000000002 ***
Type.of.Travel2                     -0.002658065  0.013144956  -0.202              0.83975
Type.of.Travel3                     -0.098512377  0.019760994  -4.985         0.000000622998 ***
No..of.other.Loyalty.Cards          -0.024549069  0.004741218  -5.178         0.000000226200 ***
Shopping.Amount.at.Airport           0.000191204  0.000085049   2.248              0.02457 *
Eating.and.Drinking.at.Airport       0.001184538  0.000087758  13.498 < 0.0000000000000002 ***
ClassEco                            -0.023397640  0.016359718  -1.430              0.15267
ClassEco Plus                       -0.109276804  0.021027733  -5.197         0.000000204293 ***
```

Interesting insight from this is that airline status is affected by Number of loyalty cards

Factors affecting the age of customers based on results of linear model:

```
102  m8 <- lm(formula = Age ~ ., data=dataCleaned)
103  summary(m8)
105:1  (Top Level)

Console  B:/Syracuse/Data Science/Project details/
Airline.Status1                      1.7610761  0.2440640   7.216 0.000000000000531632 ***
Airline.Status2                      3.1697034  0.3420680   9.266 < 0.0000000000000002 ***
Airline.Status3                      4.3068394  0.5290102   8.141 0.000000000000000408 ***
GenderMale                          -1.5804595  0.1877826  -8.416 < 0.0000000000000002 ***
Price.Sensitivity                   -2.6998124  0.1662429 -16.240 < 0.0000000000000002 ***
Year.of.First.Flight                 0.1899008  0.0300775   6.314 0.000000000276821666 ***
No.of.Flights.p.a.                   0.1244472  0.0068594  18.143 < 0.0000000000000002 ***
X..of.Flight.with.other.Airlines    -0.1154771  0.0114928 -10.048 < 0.0000000000000002 ***
Type.of.Travel2                     -8.2803981  0.2604600 -31.791 < 0.0000000000000002 ***
Type.of.Travel3                     -9.3541687  0.3940449 -23.739 < 0.0000000000000002 ***
No..of.other.Loyalty.Cards          -5.5297777  0.0886602 -62.370 < 0.0000000000000002 ***
Shopping.Amount.at.Airport          -0.0041512  0.0017048  -2.435              0.014900 *
Eating.and.Drinking.at.Airport       0.0320407  0.0017548  18.259 < 0.0000000000000002 ***
```

Interesting attributes that can be see here are Number of loyalty cards and eating and drinking at airport.

Factors affecting the price sensitivity based on results of linear model:

```
105  m9 <- lm(formula = Price.Sensitivity ~ ., data=dataCleaned)
106  summary(m9)
107
108:1  (Top Level)

Console  B:/Syracuse/Data Science/Project details/

GenderMale                          -0.01019469  0.00699862  -1.457              0.145219
Year.of.First.Flight                 0.00493183  0.00111994   4.404  0.00001068604289664 ***
No.of.Flights.p.a.                  -0.00102365  0.00025685  -3.985  0.00006756073535713 ***
X..of.Flight.with.other.Airlines    -0.00433943  0.00042775 -10.145 < 0.0000000000000002 ***
Type.of.Travel2                     -0.10612472  0.00986002 -10.763 < 0.0000000000000002 ***
Type.of.Travel3                     -0.00555837  0.01482543  -0.375              0.707722
No..of.other.Loyalty.Cards          -0.01387171  0.00353847  -3.920  0.00008868398895371 ***
Shopping.Amount.at.Airport          -0.00004535  0.00006346  -0.715              0.474883
Eating.and.Drinking.at.Airport      -0.00051005  0.00006566  -7.768  0.00000000000000825 ***
ClassEco                            -0.00050640  0.01220606  -0.041              0.966907
ClassEco Plus                       -0.07448997  0.01568877  -4.748  0.00000206559304053 ***
```

Interesting attributes that should be seen are number of other loyalty cards, eating and drinking at airport and percentage of flight with other airlines.

Therefore, it is important to note the attributes which do not directly affect the customer satisfaction attribute of the data set. Hence, the attributes which affect the factors affecting the customer satisfaction are:

1. Number of Loyalty cards
2. Eating and Drinking at airport
3. Shopping Amount
4. Percentage of flight with other Airlines

9.2 Association Rules Mining

Association Rules Mining is usually used to determine the effect of set of columns at Left Hand Side(LHS) towards Right Hand Side(RHS) and hence find the relation.

To do that, we need to make all the variables categorical so that it can map and we can see the relation. We will be mapping all the numerical values to the appropriate label using the quantile function. We have defined the quantile function which will take column as its input and will return the categorical variable.

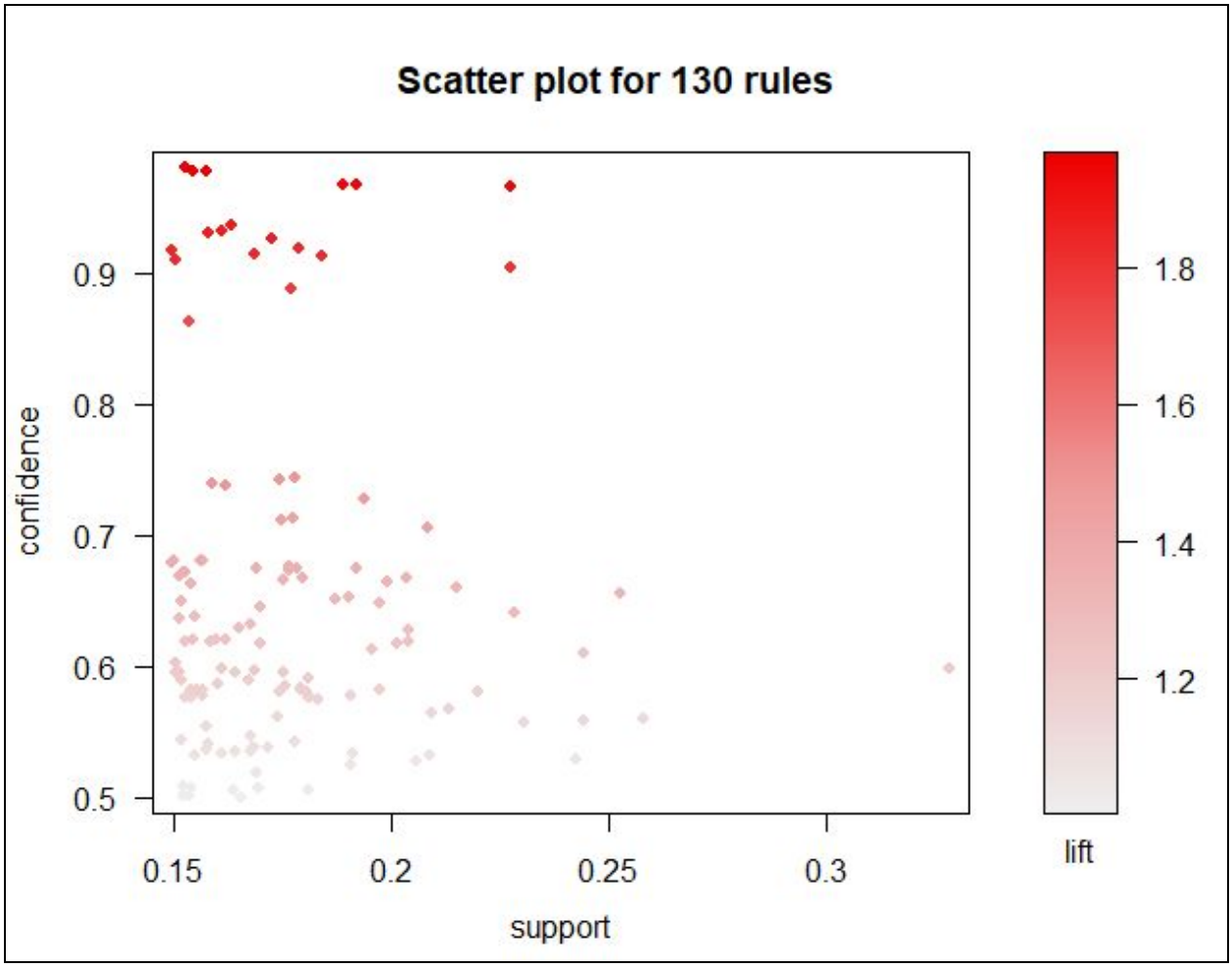
```
560
561 map<-function(vec){
562   q <- quantile(vec, c(0.4, 0.6))
563   vBuckets <- replicate(length(vec), "Average")
564   vBuckets[vec <= q[1]] <- "Low"
565   vBuckets[vec > q[2]] <- "High"
566   return(vBuckets)
567 }
```

We will be then passing all the columns into the function and converting all the columns. Then the entire database will be converted into transactions.

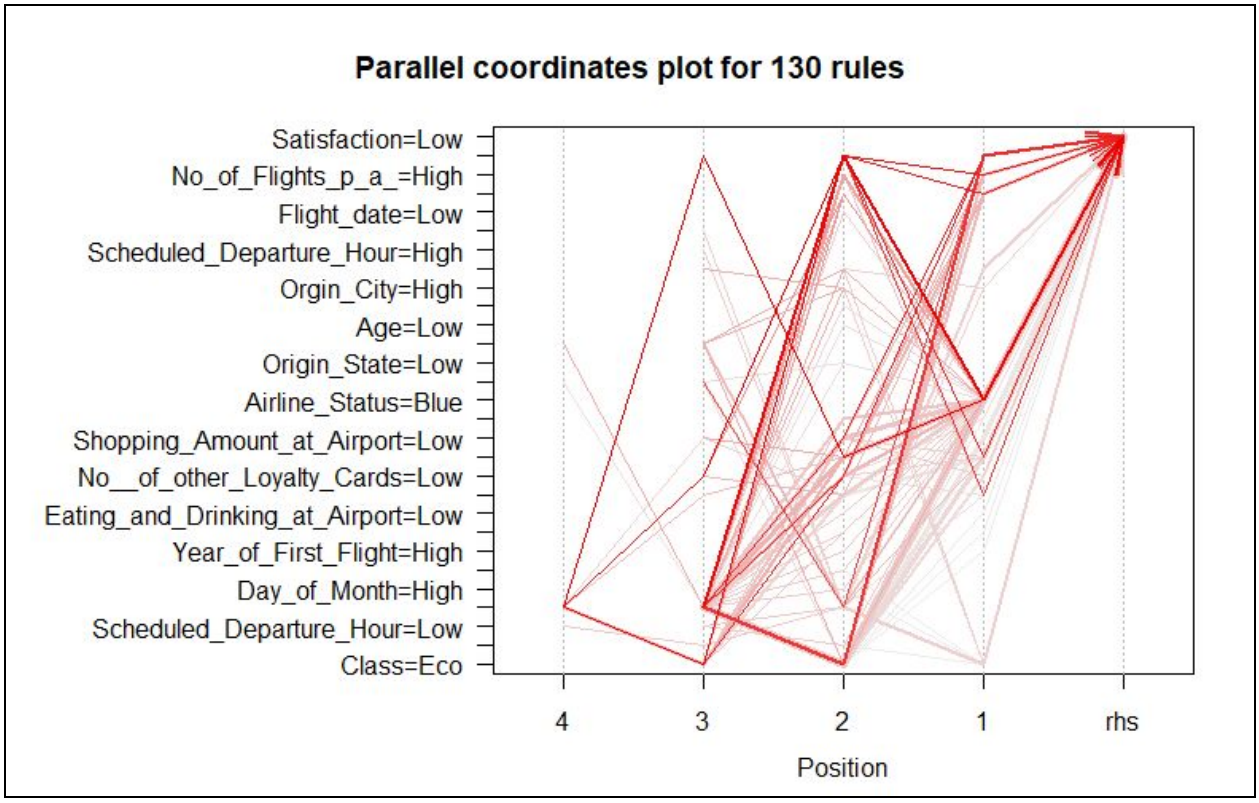
After obtaining the transactions, we will be applying Apriori algorithm to determine the factors which contribute to the level of Satisfaction.

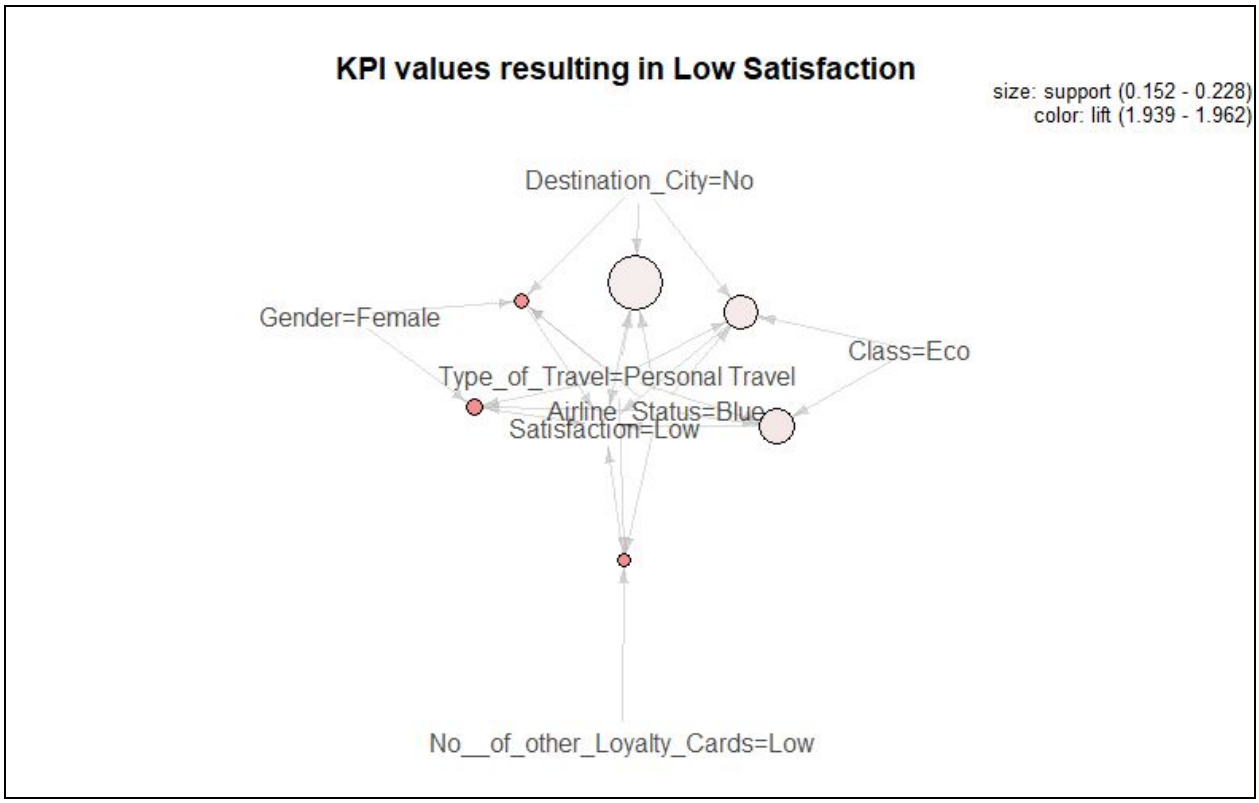
For detractors, we will be using the following rule:

```
639
640 # For customers having low satisfaction
641
642 rulesLow<-apriori(surveyDataX ,
643                   parameter = list(support=0.15,confidence=0.5,minlen=4,maxtime=0,maxlen=5),
644                   appearance = list(rhs=c("Satisfaction=Low")))
645 # We got 130 rules
646
647 inspect(rulesLow) # seeing the results affecting the satisfaction as low
648
649 plot(rulesLow) # Ploting the results
650
```



By using the above plot, we will be creating good rules by evaluating the value of lift. The above graph says, the optimum lift value should be greater than 1.8.


Parallel coordinates plot for 130 rules
Satisfaction=Low
No_of_Flights_p_a_=High
Flight_date=Low
Scheduled_Departure_Hour=High
Orgin_City=High
Age=Low
Origin_State=Low
Airline_Status=Blue
Shopping_Amount_at_Airport=Low
No__of_other_Loyalty_Cards=Low
Eating_and_Drinking_at_Airport=Low
Year_of_First_Flight=High
Day_of_Month=High
Scheduled_Departure_Hour=Low
Class=Eco
4
3
2
1
rhs
Position


KPI values resulting in Low Satisfaction
size: support (0.152 - 0.228)
color: lift (1.939 - 1.962)
Destination_City=No
Gender=Female
Class=Eco
Type_of_Travel=Personal Travel
Airline_Status=Blue
Satisfaction=Low
No__of_other_Loyalty_Cards=Low

These are the rules, we identified which contribute the most to detractors'

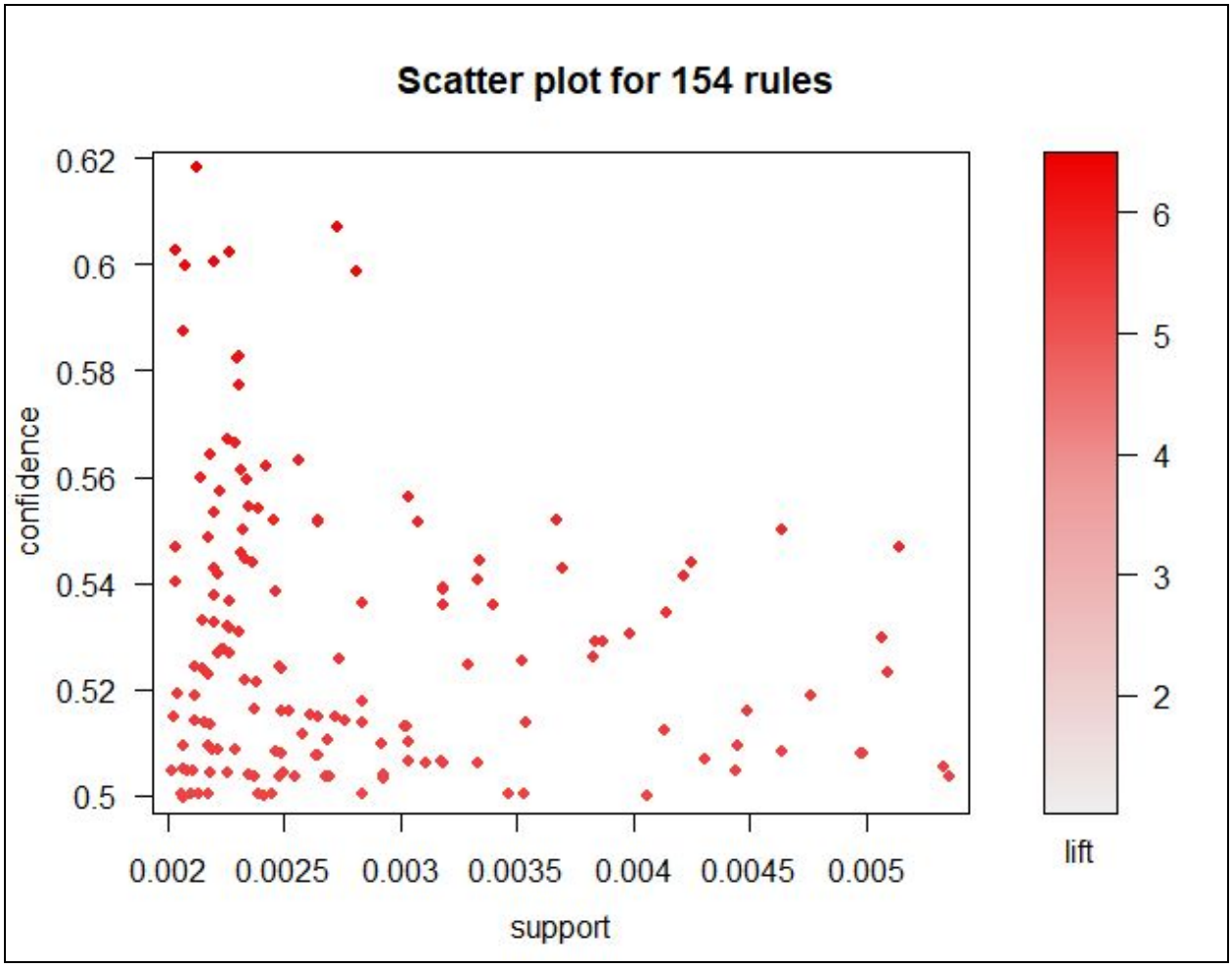
```
# [1] {Airline_Status=Blue, Type_of_Travel=Personal Travel,
No__of_other_Loyalty_Cards=Low}
# [2] {Airline_Status=Blue, Gender=Female, Type_of_Travel=Personal Travel}
# [3] {Airline_Status=Blue, Type_of_Travel=Personal Travel, Class=Eco}
# [4] {Airline_Status=Blue, Type_of_Travel=Personal Travel, Destination_City=No}
# [5] {Airline_Status=Blue, Gender=Female, Type_of_Travel=Personal Travel,
Destination_City=No}
# [6] {Airline_Status=Blue, Type_of_Travel=Personal Travel, Class=Eco,
Destination_City=No}
```

From the above rules we can determine the particular values of the attributes which lead to low satisfaction of the customers.

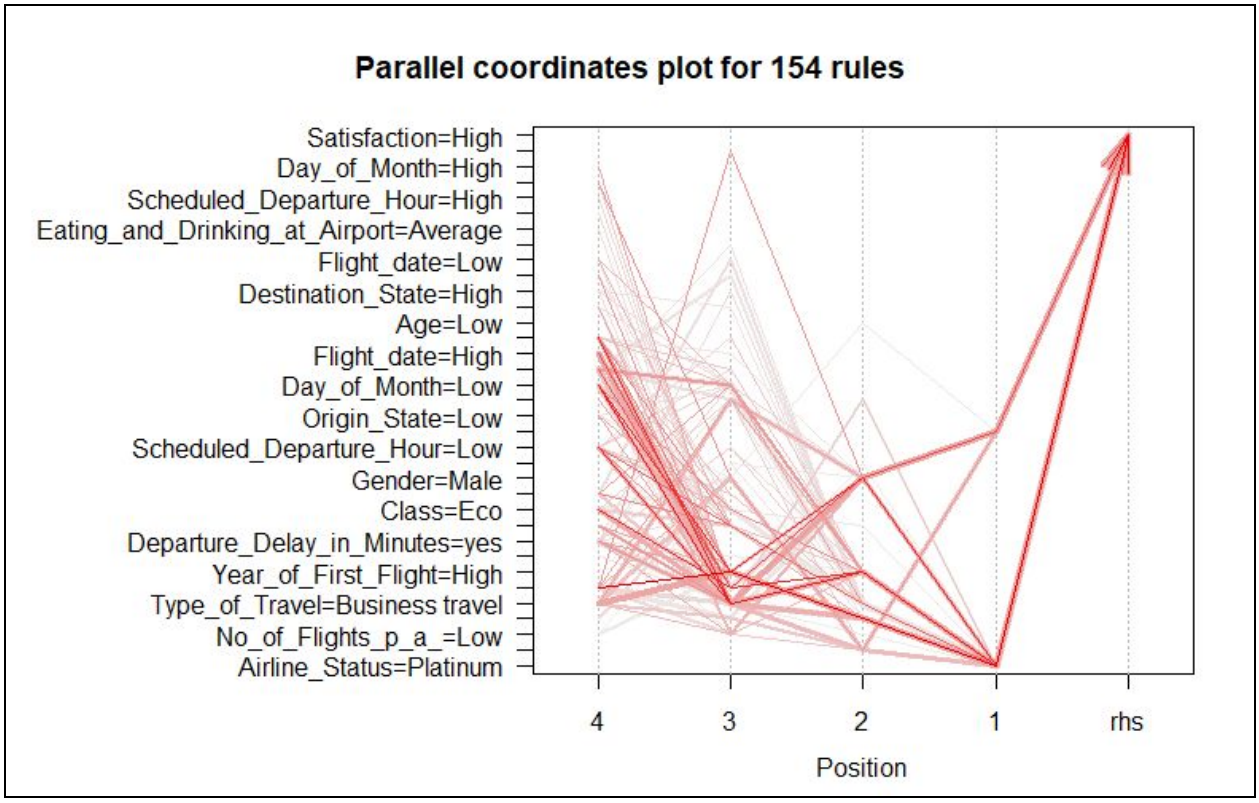
Similarly, the entire process was repeated for promoters' by changing the values of support and confidence.

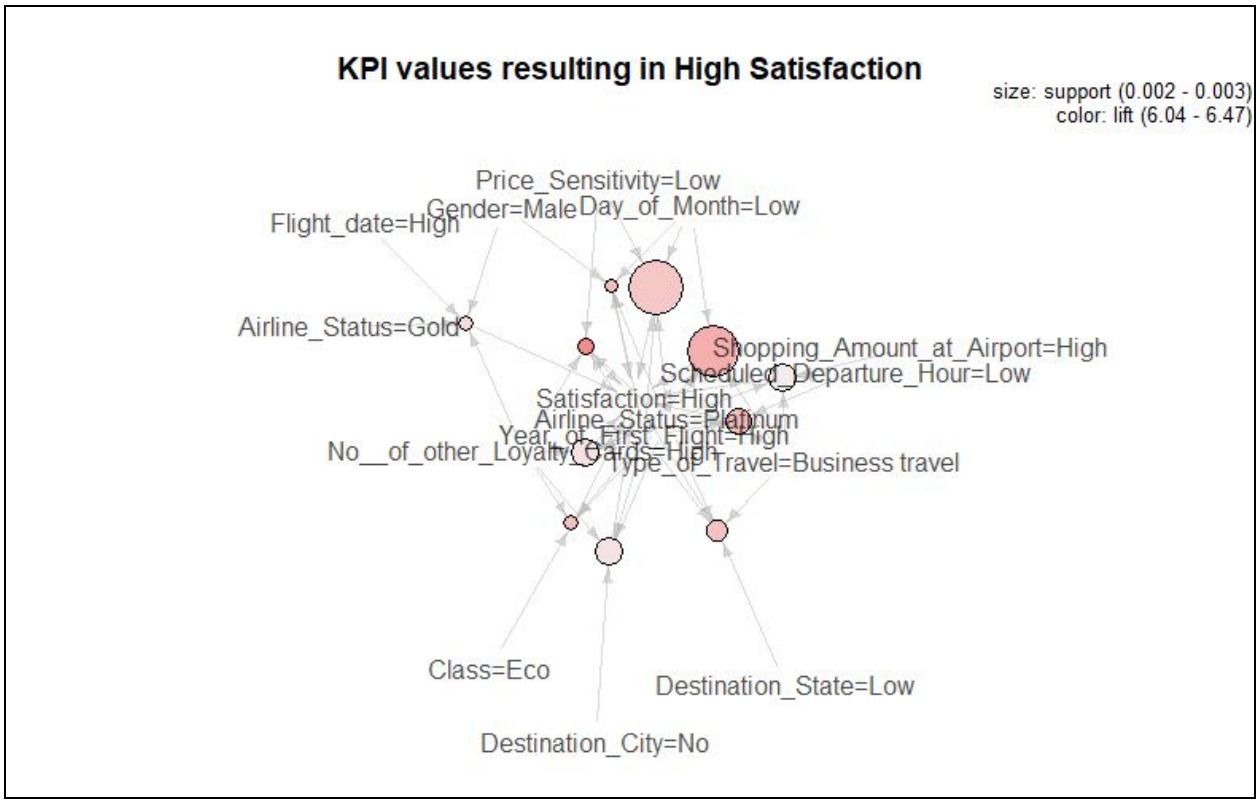
```
670
671 rulesHigh<-apriori(surveyDataX ,
672                     parameter = list(support=0.002,confidence=0.5,minlen=4,maxtime=0,maxlen=5),
673                     appearance = list(rhs=c("Satisfaction=High")))
674 # 154 rules
675
676 inspect(rulesHigh)
677
678 plot(rulesHigh)
679
```

The scatter plot for the available rules will be,



Here, the value of the lift is significantly higher than the previous one, because here the confidence is quite less.


Parallel coordinates plot for 154 rules
Satisfaction=High
Day_of_Month=High
Scheduled_Departure_Hour=High
Eating_and_Drinking_at_Airport=Average
Flight_date=Low
Destination_State=High
Age=Low
Flight_date=High
Day_of_Month=Low
Origin_State=Low
Scheduled_Departure_Hour=Low
Gender=Male
Class=Eco
Departure_Delay_in_Minutes=yes
Year_of_First_Flight=High
Type_of_Travel=Business travel
No_of_Flights_p_a_=Low
Airline_Status=Platinum
4
3
2
1
rhs
Position


KPI values resulting in High Satisfaction
size: support (0.002 - 0.003)
color: lift (6.04 - 6.47)
Price_Sensitivity=Low
Gender=Male
Day_of_Month=Low
Flight_date=High
Airline_Status=Gold
Shopping_Amount_at_Airport=High
Scheduled_Departure_Hour=Low
Satisfaction=High
Airline_Status=Platinum
Year_of_First_Flight=High
No__of_other_Loyalty_Cards=High
Type_of_Travel=Business travel
Class=Eco
Destination_State=Low
Destination_City=No

So, by evaluating the results and restricting it by the value of lift, we generate good rules for promoters'

```
# [1] {Airline_Status=Platinum, Price_Sensitivity=Low, Year_of_First_Flight=High,
No__of_other_Loyalty_Cards=High}

# [2] {Airline_Status=Platinum, Year_of_First_Flight=High, Type_of_Travel=Business travel,
Scheduled_Departure_Hour=Low}

# [3] {Airline_Status=Platinum, Gender=Male, Year_of_First_Flight=High,
Day_of_Month=Low}

# [4] {Airline_Status=Platinum, Year_of_First_Flight=High, Type_of_Travel=Business travel,
Day_of_Month=Low}
```

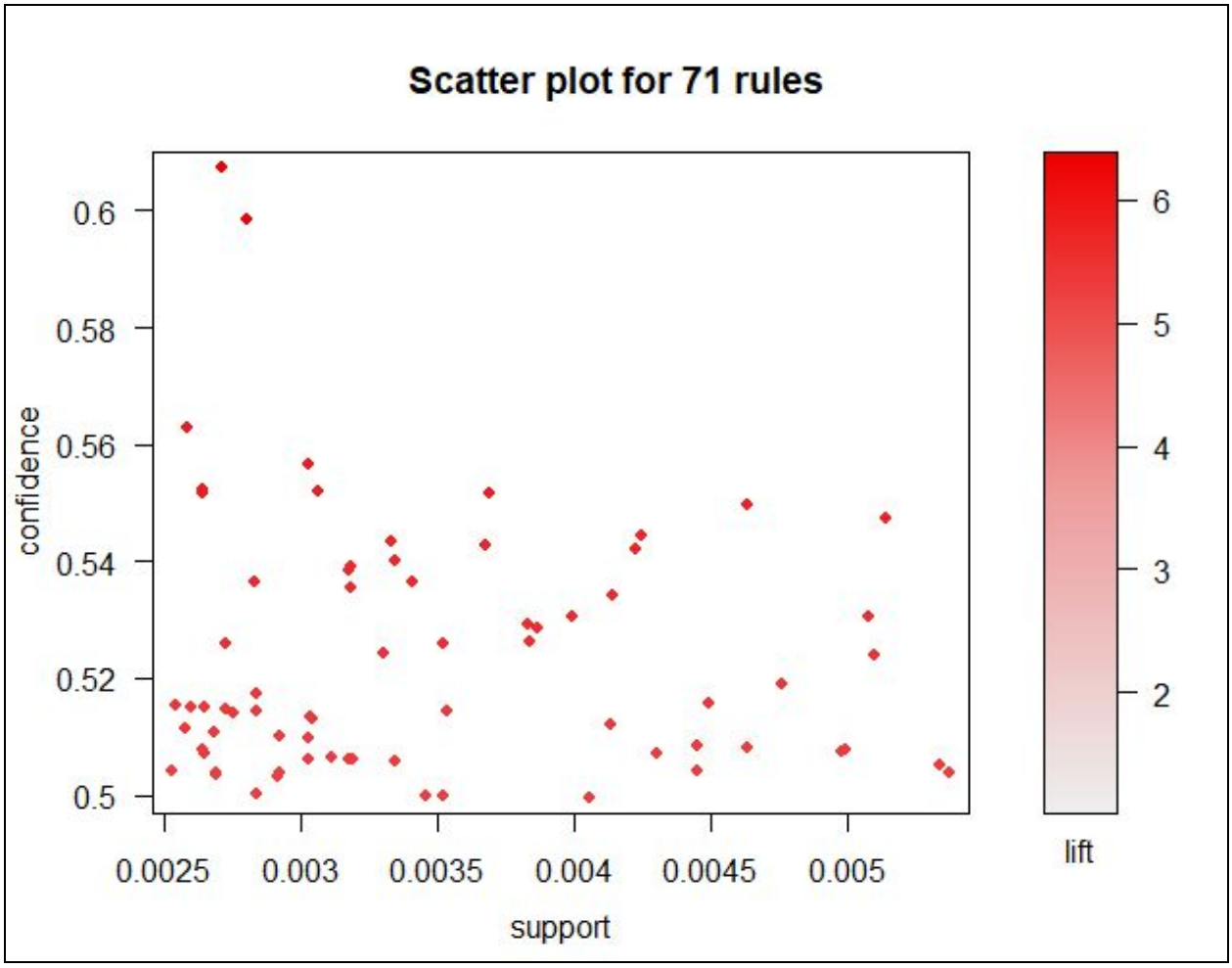
From the above rules we can determine the particular values of the attributes which lead to high satisfaction of the customers.

This, we have done for the entire Cheapseats Airlines Inc. We will be applying these rules to the smaller subset of data which we got from applying Linear modelling.

These are the significant columns, we got:

1. Airline Status
2. Age
3. Gender
4. Type of Travel
5. Price Sensitivity
6. Number of Flights per annum
7. Class
8. Year of first flight

This further confirmed our results and strengthen our analysis.


Scatter plot for 71 rules
confidence
0.6
0.58
0.56
0.54
0.52
0.5
0.0025
0.003
0.0035
0.004
0.0045
0.005
support
lift
6
5
4
3
2

9.3. Support Vector Machines

We have used kSVM to evaluate the results from the Linear Modeling and Association Rules Mining. In the SVM, we divide the entire data into two parts.

The first part called as Training Data, will have 2/3rd of random rows from the entire dataset and the second part called as Testing Data will have remaining 1/3rd of data.

The Training Data, as the name suggests, will be used as training data frame to make the model learn the entire algorithm. While executing the model, we will be giving an input as permutations and combinations of the columns got from the two modelling techniques.

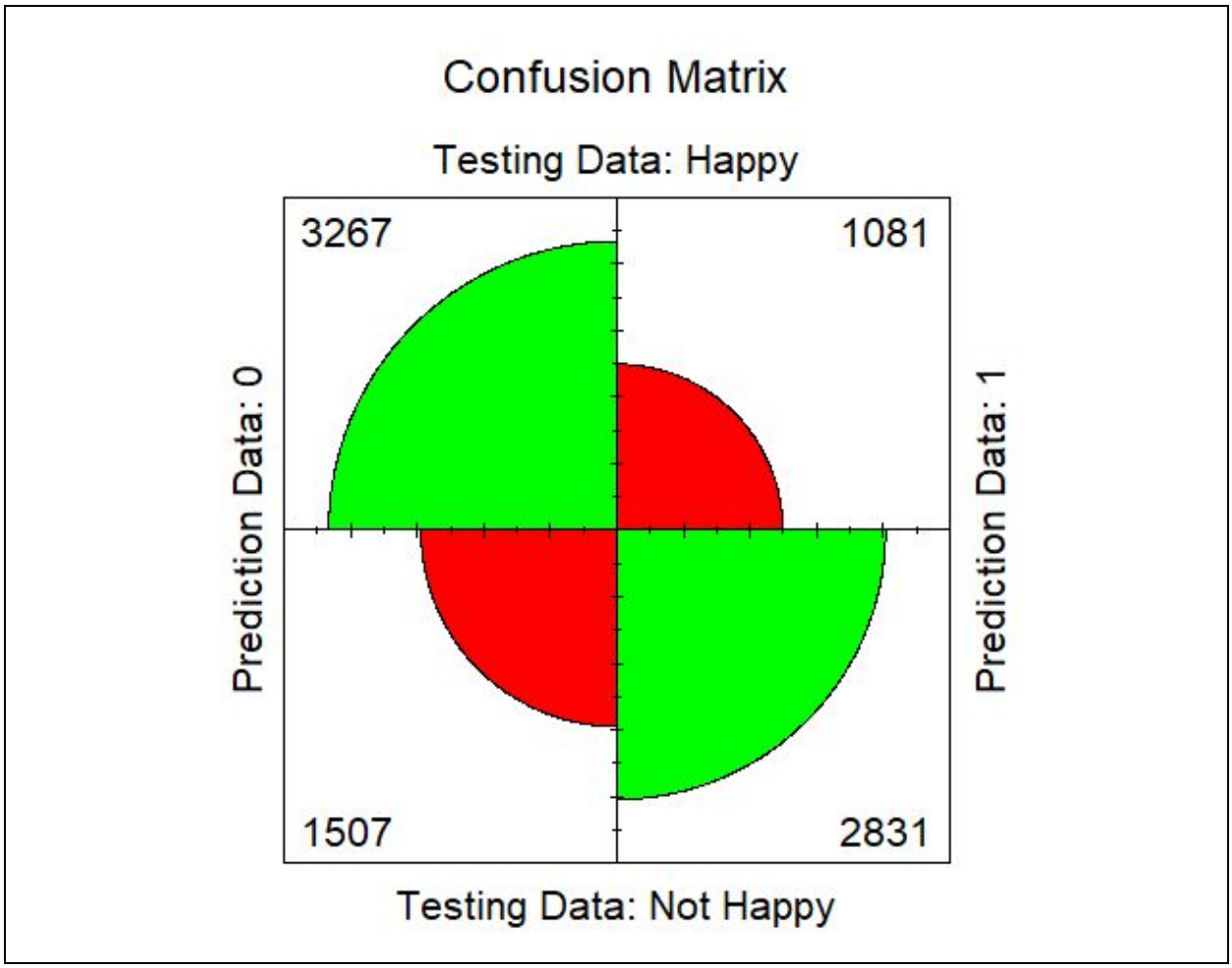
The Testing Data will be then used for observing how our model fared against the prediction. Finally, we will calculate the error rate to check the accuracy of our results.

```
745
746 # kSVM
747
748 Survey_dataSVM<-dataCleaned
749
750 dummy<-ifelse(as.numeric(Survey_dataSVM$Satisfaction) > 3, "Happy", "Not Happy")
751
752 Survey_dataSVM$HappyCust <-dummy # Creating a new column and insrting the above generated value.
753
754 dim(Survey_dataSVM)
755
756 randIndex<-sample(1:dim(Survey_dataSVM)[1]) # Creating a dataframe of random indices
757
758 cutPoint2_3<-floor(2*dim(Survey_dataSVM)[1]/3) # Creating a breakpoint of 2/3rd and 1/3rd part
759
760 trainData<-Survey_dataSVM[randIndex[1:cutPoint2_3],] # Creating traindata with 2/3rd
761
762 testData <-Survey_dataSVM[randIndex[(cutPoint2_3+1):dim(Survey_dataSVM)[1]],] # Creating testdata with 1/3rd
763
764 dim(trainData) # Checking the dimension
765
766 dim(testData) # Checking the dimension
767
768 install.packages("kernlab") # installing the package: kernlab
769 library(kernlab) # Including the library: kernlab
770
771 svmOutput<-ksvm(HappyCust~Airline_Status+Age+Price_Sensitivity+No_of_Flights_p_a_+Class, data = trainData,
772                 kernel= "rbfdot", kpar = "automatic", C = 5, cross = 3, prob.model = TRUE)
773 svmOutput
774
```

After getting the output, it will be compared against testing data.

```
774
775 svmPred <- predict(svmOutput, testData, type = "votes") # Making a Prediction variable based on number of votes
776
777 str(svmPred)
778 head(svmPred[2,])
779
780 compTable<-data.frame(testData$HappyCust,svmPred[2,]) # Creating a composite table based on HappyCustomer and svmPredicti
781 colnames(compTable)<-c("Testing Data","Prediction Data")
782
783
784 conMatrix<-table(compTable) # Creating a confusion matrix
785 conMatrix # Displaying the result onto console
786
787 errorSum<-conMatrix[1,2]+conMatrix[2,1] # Creating a dataframe containing sum of errors
788 errorRate<-errorSum/sum(conMatrix)*100 # Creating percentage of error rate
789 errorRate
790
791 # Drawing Confusion Matrix
792
793 fourfoldplot(conMatrix, color = c("Red", "Green"),
794              conf.level = 0, margin = 1, main = "Confusion Matrix")
795
```

We can also look upon fourfold graph to evaluate our results.



This gives us a different perspective and hence we can evaluate our results in a more visual manner.

We also have carried out different combinations and this was our analysis.

1. Airline_Status, Price_Sensitivity and Class has error rate 37.37048
2. Airline_Status and Class has error rate 37.07
3. Airline_Status and Age has error rate 31.19
4. Type of travel and age has error rate 22.93
5. Age and Gender has error rate 32.0053
6. Type of travel and class 24.91
7. Type of travel, class and airline status has error rate 24.68
8. Type of travel, age and class has error rate 23.73
9. Type of travel and airline status has error rate 24.83
10. Airline_Status,Age,Price_Sensitivity,No_of_Flights_p_a_ and Class has error rate of 29.97

10. Actionable Insights

1. Provide more convenience and luxury to higher age group of customers such as
    - Early on-boarding
    - Better in-flight services
    - Seats with extra legroom
2. Provide more offers to travel for personal reasons.
    - Example collaboration with hotels to provide more offers
    - Discounts during holiday season
    - Travel and tourism packages
3. Provide better customer services for people travelling in economy class and economy plus class by giving them one ticket upgrade to business class every year. Provide referral scheme to customers travelling in economy class so that they get discounts/miles for every referee they get.
4. As the data suggests, customers are not price sensitive. Hence, providing better but paid facilities to people travelling with Blue status to increase net promoter score.
5. Provide more loyalty cards for female customers based on frequency of flights.
6. Frequent travellers have low satisfaction and high percentage of flight with other airlines. Hence, providing more perks on loyalty cards such as travel miles, discount while shopping and eating at the restaurants at the airport can encourage the customers to choose Cheapseats over other airlines.
7. As the data suggests, business travelers are highly satisfied, hence giving free baggage insurance and corporate benefits to retain such customers
8. Start implementing these solutions in the below 5 states as it caters to 50% of Cheapseats customer base:
    - California
    - Texas
    - Florida
    - Nevada
    - Illinois

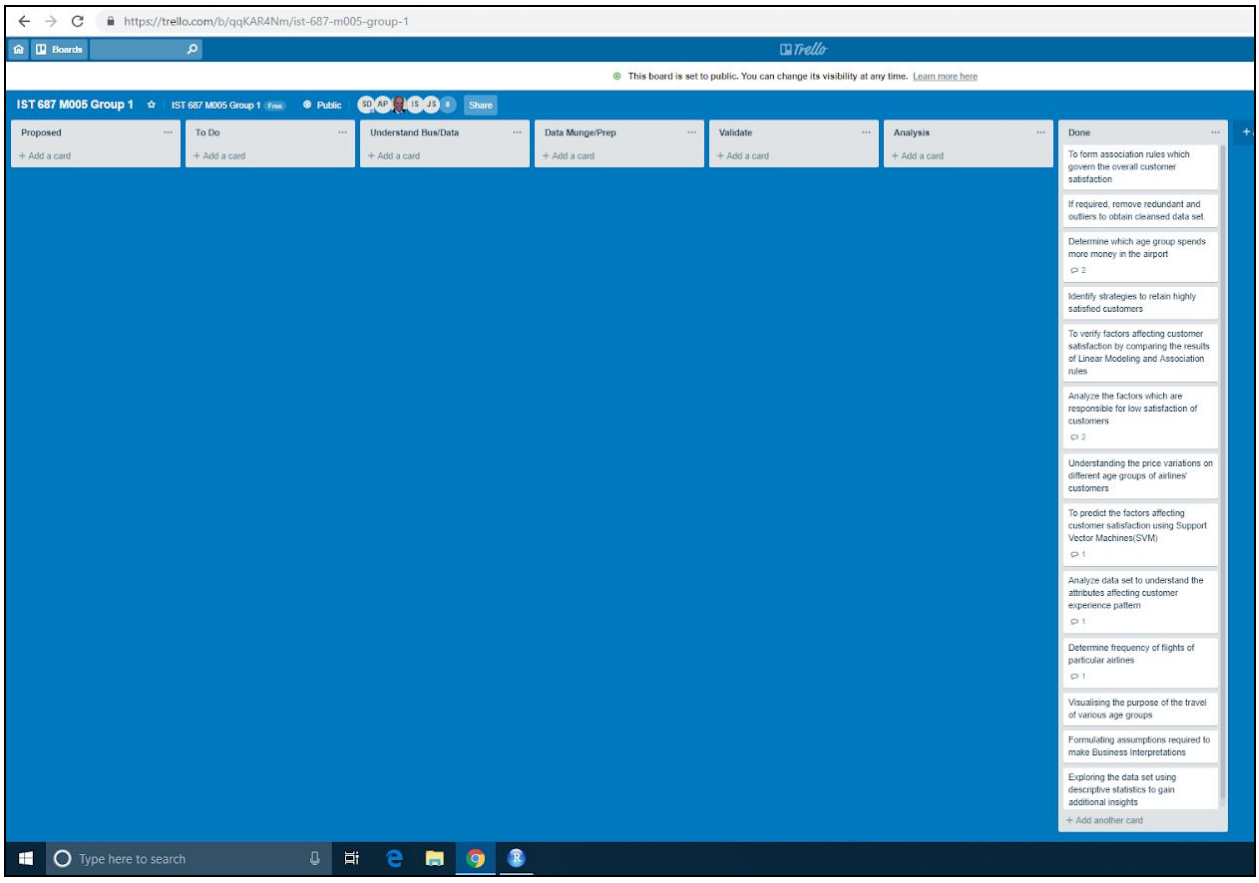
Appendix 1: R Code

Following is the github link of the project:

https://github.com/shashi-dhuppe/Data-Science-IST-687/blob/master/Data%20Science%20Project

Appendix 2: Trello Board

The tasks were discussed in team meetings and individual tasks were assigned to all. The individuals who were assigned with the tasks moved the tasks across the Trello board based on the results obtained.
Results and analysis were discussed during the team meetings.



Please find below the link the trello board:

https://trello.com/b/qqKAR4Nm/ist-687-m005-group-1https://trello.com/b/qqKAR4Nm/ist-687-m005-group-1

http://prezi.com/w-swjkjn0hmx/?utm_campaign=share&utm_medium=copy

https://prezi.com/w-swjkjn0hmx/edit/#0