IST 687

INTRODUCTION ………………………………………………………………………………….. 1 ABOUT OUR PROJECT ……………………………………………………………….... 2 BUSINESS QUESTIONS ………………………………………………………………………… 2 DATA REQUISITION AND PREPROCESSING ……………………………………………….

DATA REQUISITION …………………………………………………………………….. 3

DATA PREPROCESSING ………………………………………………………………. 3

APPLICATION OF MODELING TECHNIQUES………………………

MODELING 1 - Association Rule Mining …………………………………………….. 31 MODELING 2 - Linear Modeling ………………………………………………………. 34 VALIDATION …………………………………………………………………………………….. Support Vector Machines …………………………………………………………….... 38

USE OF DESCRIPTIVE STATISTICS AND DATA VISUALIZATION ……………………… 5

GGMAP ……………………………………………………………………………………. 5

GGPPLOT2 ……………………………………………………………………………… 13

Correlation Matrix(Preparation) …………………………………………………… 29

ACTIONABLE INSIGHTS ……………………………………………………………………… 42

**INTRODUCTION**

With the advent of new technology, flights have become the most reliable and time efficient mode of transport for customers especially in today's time when people don’t have a lot of time to spare and need to get to the destination as fast as possible. As more people have joined the workforce and more jobs require the employees to travel for their work, flights are the best mode of transportation. And with time air travel has also become more affordable and hence the demand has also increased for flights. This has led to more airlines coming into effect which has given rise to increasing competition in this industry. The customers now have a wide range of options in terms of airlines to choose from and would prefer to travel by that airlines more frequently that provides higher satisfaction compared to the others. So, customer satisfaction is an important factor which leads to the growth of an airline per say.

There are many ways in which feedback can be collected from the customers in order to determine and then work out ways to improve the customer satisfaction. However, the best way to do so is to conduct a customer survey. This report focuses on the customer survey collected from a large number of people for various airlines. We are working on a dataset consisting of 14 airlines, 29 different variables or attributes that affect the customer satisfaction and has 194833 observations.

We have used a number of key performance indicators like flight path, flight delays, the airline type etc in order to analyse ou survey. This helped us to understand the dataset better and as a result we have used different data analysis models and different data visualization tools to analyze the customer feedback from the dataset.

The business questions that were created for the dataset were answered by using different analysis models that was useful to find the correlation between different attributes and the customer satisfaction. This was helpful in order to understand what are the main factors that influence customer satisfaction and hence find better solutions to increase the customer satisfaction.

We have used Linear Modeling and Association Rule Mining to analyse our dataset and then in order to validate our results we have used Support Vector Machine model.

**BUSINESS QUESTIONS**

The customer feedback helps us gain better insights into what aspects of the Airlines services particularly need more attention so that the customer satisfaction can be increased which help us to develop the actionable insights. As mentioned before as the competition in the airline industry is increasing rapidly so each airline should focus more on providing better customer satisfaction so as to increase customer base and loyalty to generate greater customer recommendation and in turn higher revenues

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

1)What are the main factors that affect customer satisfaction of our dataset?

2)How do individual attributes affect the customer satisfaction?

3)Which attributes are positively correlated with the customer satisfaction?

4)Which attributes are negatively correlated with the customer satisfaction?

5)Which factors contribute towards increasing the probability of the customer providing positive feedback and recommending the airlines.

6)What are the services that provide greater customer satisfaction?

7) Which factors should be focused on that can be improved upon to provide better customer satisfaction and hence lead to rise in revenue?

**DATA PREPROCESSING**

For data analysis to be conducted on any dataset the first step should always be cleaning the data which helps us to analyse the data better. So, the first step is basically loading the dataset, cleaning and munging it.

A. DATA LOADING

The dataset was provided to us by our professors for our project. We downloaded the csv file and then loaded it in order to analyse it in detail. The data consisted of approximately 194833 different observations based on 29 variables.

Code for Data Loading:

**flightData <- read.csv("spring19survey.csv")**

**dataset <-flightData**

**View(dataset)**

**str(dataset)**

**summary(dataset)**

**B. DATA PREPROCESSING**

There were many attributes in the dataset that needed pre processing before analysis. For example there were a lot of NA values in the dataset especially in the columns like flight time in minutes, arrival delay etc. So we replaced all the NA values by the mean value of the whole column. We also trimmed the white spaces in all the columns. We deleted the partner code, partner name columns as they are not relevant to us. Moreover, in cases where the flight is cancelled the satisfaction is bound to be less so we have considered only those cases where the flight has not been cancelled.

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)) #4113**

**ncol(dataset) #29**

**nrow(dataset) #194833**

**#Taking only the data whose flights are not cancelled**

**Airdata <- dataset[dataset$Flight.cancelled=="No",]**

**nrow(Airdata)#191230 ->>>>>> 3603 flights cancelled**

**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)) #510**

**sum(is.na(Airdata$Arrival.Delay.in.Minutes)) #0**

**sum(is.na(Airdata$Flight.time.in.minutes)) #502**

**sum(is.na(Airdata$Departure.Delay.in.Minutes)) #0**

**is.null(Airdata)**

**#View(Airdata)**

**nrow(Airdata)**

**nrow(dataset)**

**str(Airdata)**

**h1 <- ggplot(Sati\_flight\_1,aes(x=Age))**

**h1 <- h1 + geom\_histogram(aes(fill=Satisfaction),position = "dodge")**

**h1 <- h1 + ggtitle("Satisfaction versus Age")**

**h1**

**#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**

**# complete.cases: Return a logical vector indicating which cases are complete, i.e., have no missing values.**

**sum(!complete.cases(Airdata)) # reduced from 510 to zero**

**is.null(Airdata)**

**#View(Airdata)**

**nrow(Airdata)**

**nrow(dataset)**

**str(Airdata)**

**#removing unwanted columns from the Survey dataset**

**#16 Partner code**

**#17 Partner name**

**Airdata <- Airdata[-c(16,17)]**

**str(Airdata) # 27 variables**

**summary(Airdata)**

**View(Airdata)**

**USE OF DESCRIPTIVE STATISTICS AND DATA VISUALIZATION**

This section focuses on finding the correlation of the different factors with each other using ggmap in order to map some of our finding on the map and ggplot2 is used to plot histograms, barplots etc. to find the relationship between two distinct attributes.

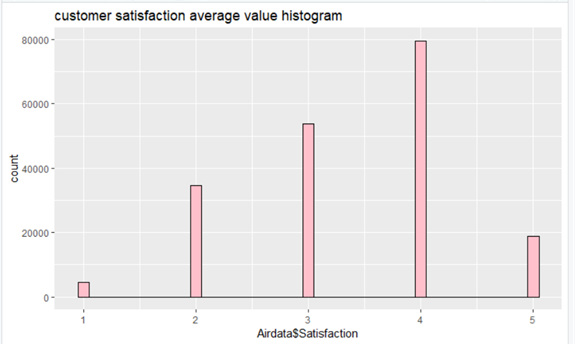
1)This graph shows the number of people or count for each customer rating value. We can see that maximum customers have provided rating 4 and least number of customers have provided rating 1. But there are more customers who have provided a rating lower than 4 ( other than 4 and 5)

g <- ggplot(Airdata, aes(x=Airdata$Satisfaction))

g <- g + geom\_histogram(binwidth=0.1, color="black", fill = "pink",col.bg="blue", col.grid="blue") + scale\_fill\_brewer()

g <- g + ggtitle("customer satisfaction average value histogram")

g



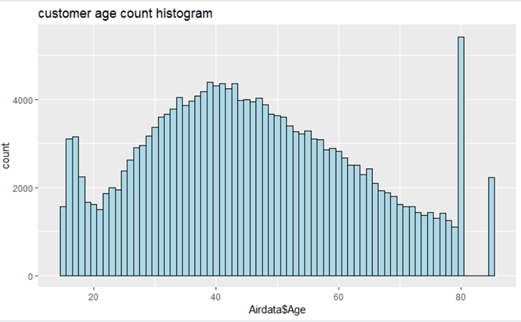
2) This shows us the count for the number of people for different ages that have provided the customer satisfaction rating. We can infer that maximum people are in the age bracket of 20-60.

g1 <- ggplot(Airdata, aes(x=Airdata$Age))

g1 <- g1 + geom\_histogram(binwidth=1, color="black", fill = "light blue",col.bg="blue", col.grid="blue") + scale\_fill\_brewer()

g1 <- g1 + ggtitle("customer age count histogram")

g1



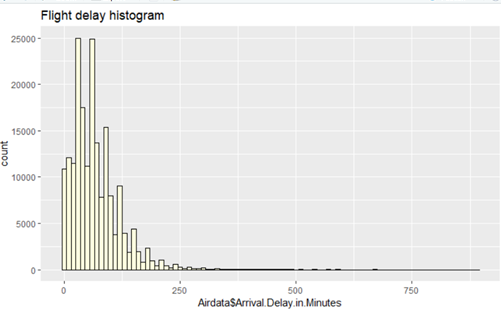
3) This graph shows us that more number of flights have been delayed for less than 120 minutes than greater than that.

g2 <- ggplot(Airdata, aes(x=Airdata$Arrival.Delay.in.Minutes))

g2 <- g2 + geom\_histogram(binwidth=10, color="black", fill = "light yellow") + scale\_fill\_brewer()

g2 <- g2 + ggtitle("Flight delay histogram")

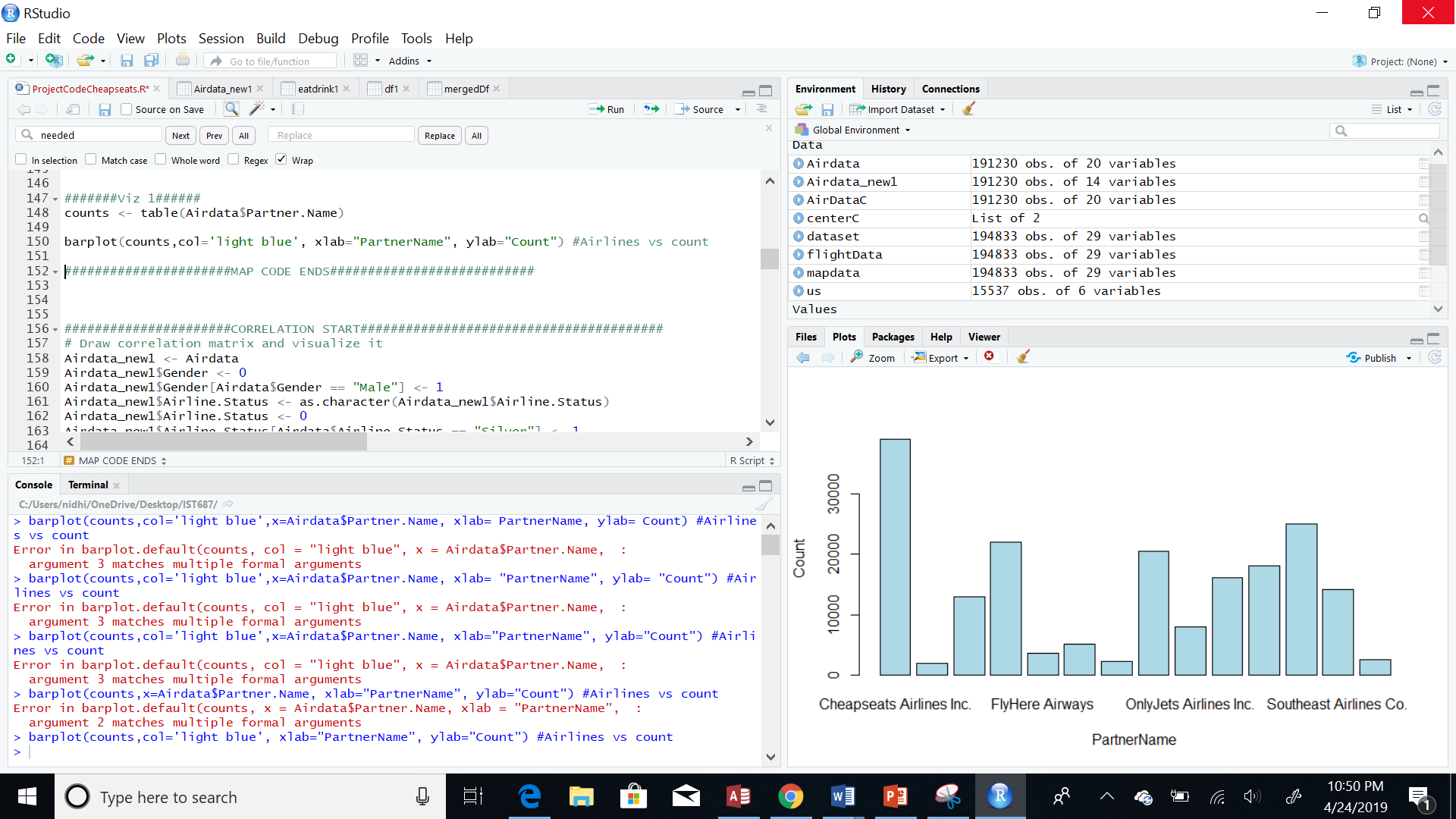
g2



4) This graph shows us the number of customers for each airline and one main point to notice is cheapseat airlines has the highest number of customer count and hence the largest subset of customer satisfaction data.

counts <- table(Airdata$Partner.Name)

barplot(counts,col='light blue', xlab="PartnerName", ylab="Count") #Airlines vs count



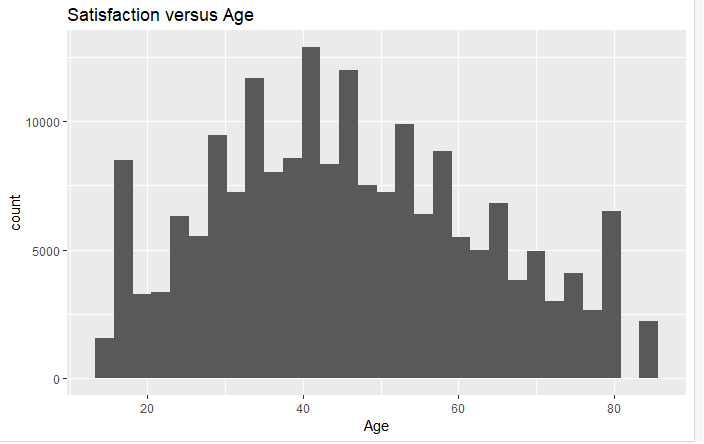
5) The following graph shows that maximum people fall in the age bracket 20- 60. This helped us identify the age bracket that we need to focus on in order to improve the customer satisfaction.

h1 <- ggplot(Airdata,aes(x=Age))

h1 <- h1 + geom\_histogram(aes(fill=Satisfaction),position = "dodge")

h1 <- h1 + ggtitle("Satisfaction versus Age")

h1



6) The Age vs shop graph shows us that people in the age bracket of 20-60 have spent more money on shopping at the airport than the others.

ggplot(AgeVsShop, aes(x = AgeVsShop$Age, y = AgeVsShop$Shopping\_Amount\_at\_Airport)) + geom\_line(stat = "identity", color = "Black") +

ggtitle("Age vs Shopping Amount Spent at Airport") +

scale\_x\_continuous(name="Age") +

scale\_y\_continuous(name = "Average Amount spent")



7) Plotting line graph of Age vs Eating and Drinking Amount gives us an idea that people above the age of 60 are spending more on eating and drinking at the airport. Hence something can be provided by the airlines to customers so that this amount can be reduced and hence increase customer satisfaction.

AvgVsEatDrink<-sqldf("Select Age,count(age), avg(Eating\_and\_Drinking\_at\_Airport) AS Eating\_and\_Drinking\_at\_Airport FROM df\_Filter1 GROUP BY Age;")

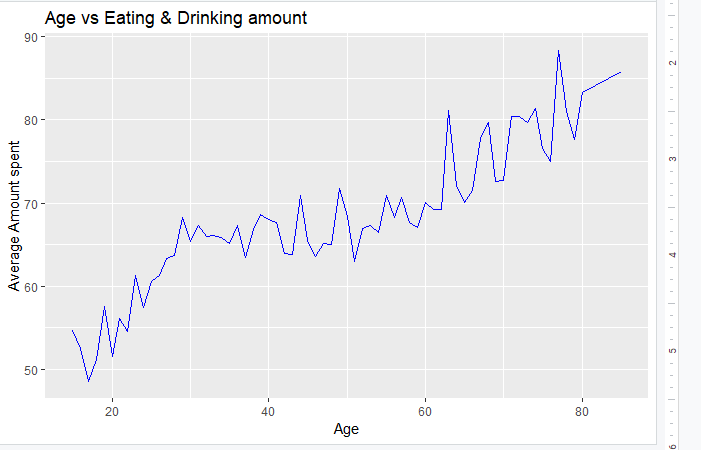
#str(AvgVsE\_D)

ggplot(AgeVsEatDrink, aes(x = AgeVsEatDrink$Age, y = AgeVsEatDrink$Eating\_and\_Drinking\_at\_Airport)) + geom\_line(stat = "identity", color = "Black") +

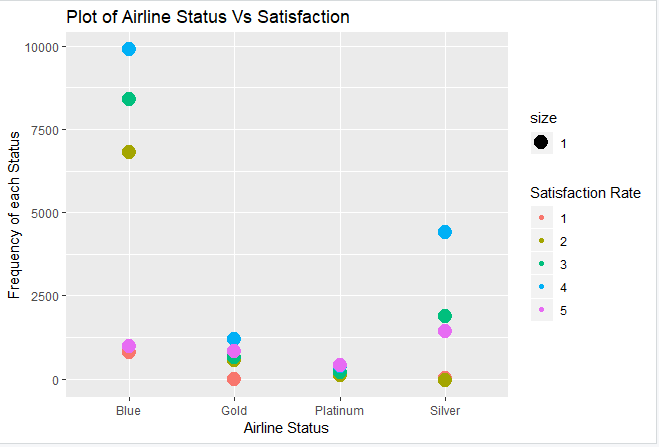
ggtitle("Age vs Eating & Drinking amount") +

scale\_x\_continuous(name="Age") +

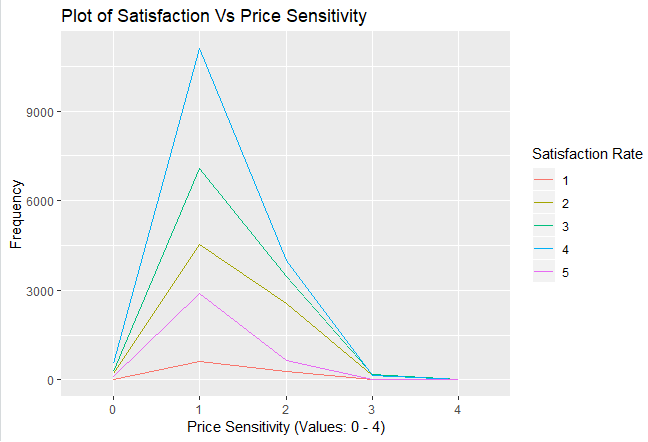
scale\_y\_continuous(name = "Average Amount spent")



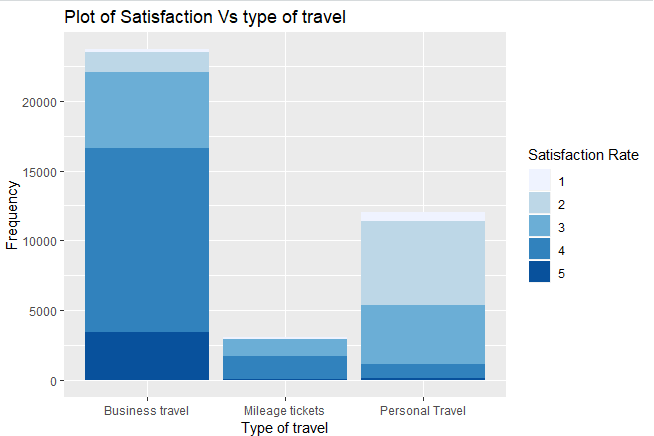
8) The graph of Satisfaction vs airline status shows us that maximum people with the airline status Blue have a customer satisfaction of 2 which is low so this is an important factor and the airline status Blue should be focused upon to improve customer satisfaction.



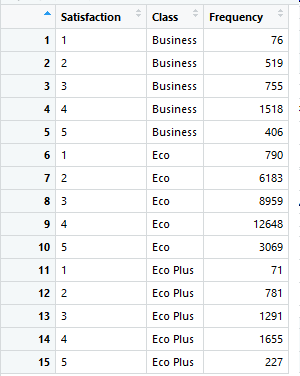
9)The graph of Price Sensitivity vs satisfaction shows us that the customers who have rated the airlines 2 or 3 are affected by the price sensitivity than extreme satisfaction values like 5 and 1.



10) The graph Type of travel vs satisfaction shows us that people on personal travel have rated the airlines lower (1,2,3) compared to those who are on business travel.

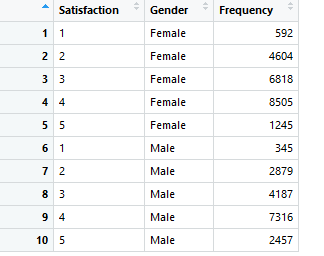


11) This Class vs Satisfaction shows us that more people who travel by economy class have given lower satisfaction ratings than those who travel Business or Eco plus class. So this is an important factor that should be considered. The services provided to economy class customers should be focused upon to increase customer satisfaction ratings.



12) This Gender vs Satisfaction shows us that more lower ratings that is 1,2,3 have been given by females than males.

Gender vs Satisfaction



**USE OF MODELING TECHNIQUES**

**MODELING 1** - Linear Modeling

Linear model is one of the models that we learnt in class which gives us the relationship between variables and also the amount of impact each attribute has on the customer satisfaction. It helps us find significant variables that affect customer satisfaction.

# Linear Modeling on Cust Sat - 1,2,3

model\_lm0 <- lm(formula = Satisfaction ~ Age + Price.Sensitivity +

Flights.Per.Year + Loyalty + Total.Freq.Flyer.Accts +

Shopping.Amount.at.Airport + Eating.and.Drinking.at.Airport

+ Scheduled.Departure.Hour + Departure.Delay.in.Minutes +

Flight.time.in.minutes+ Arrival.Delay.in.Minutes+ Airline.Status+Price.Sensitivity+ Type.of.Travel+Class +Flight.date+Orgin.City + Destination.City+ Long.Duration.Trip +Gender,

data = df1)

summary(model\_lm0)

**#Adjusted R-squared: 0.4292**

**#Significant variables are::: Age,Total.Freq.Flyer.Accts, Airline.Status, Type.of.Travel ,Class**

model\_lm1 <- lm(formula = Satisfaction ~ Age + Price.Sensitivity +

Flights.Per.Year + Flight.time.in.minutes + Airline.Status +

Type.of.Travel+Class+Flight.date,

data = df1)

summary(model\_lm1)

**#Adjusted R-squared: .4201**

**# attributes with 3 stars of significance::: Age,Price.Sensitivity,Flights.Per.Year, Loyalty, Total.Freq.Flyer.Accts, Shopping.Amount.at.Airport**

model\_lm2a <- lm(formula = Satisfaction ~ Age,

data = df1)

summary(model\_lm2a) #Adjusted R-squared: 0.02819

model\_lm2b<- lm(formula = Satisfaction ~ Price.Sensitivity,

data = df1)

summary(model\_lm2b) # Adjusted R-squared: 9.733e-05

model\_lm2c <- lm(formula = Satisfaction ~ Flights.Per.Year,

data = df1)

summary(model\_lm2c) #Adjusted R-squared: 0.009052

model\_lm2d <- lm(formula = Satisfaction ~ Flight.time.in.minutes,

data = df1)

summary(model\_lm2d) #Adjusted R-squared: 0.003476

model\_lm2e <- lm(formula = Satisfaction ~ Airline.Status,

data = df1)

summary(model\_lm2e) #Adjusted R-squared: 0.001806

model\_lm2f <- lm(formula = Satisfaction ~ Type.of.Travel,

data = df1)

summary(model\_lm2f) #Adjusted R-squared: 0.0001945

model\_lm2g <- lm(formula = Satisfaction ~ Flight.date,

data = df1)

summary(model\_lm2g) #Adjusted R-squared: 0.08568

model\_lm2h <- lm(formula = Satisfaction ~ Class,

data = df1)

summary(model\_lm2h)

**#Adjusted R-squared: 0.0008464**

**########################SUMMARY FROM LM: Type.of.Travel, Airline.Status, Age**

**MODELING 2** - Association Rule Mining

This is also one of the models covered in class and it helps us to analyse the relationship between the different attributes and it also helps in identification of rules which can then be sorted in order of strength or importance.

######################Association rules start########################################

createBucket1 <- function(vec){

vBuckets <- replicate(length(vec), "Average")

vBuckets[vec > 3] <- "High"

vBuckets[vec < 3] <- "Low"

return(vBuckets)

}

happyCust <- createBucket1(df1$Satisfaction)

#View(happyCust)

price\_sen <- createBucket1(df1$Price.Sensitivity)

#View(price\_sen)

freq\_flyer <- createBucket1(df1$Total.Freq.Flyer.Accts)

View(freq\_flyer )

dep\_hour <- createBucket1(df1$Scheduled.Departure.Hour)

#View(dep\_hour )

createBucket2<- function(vec){

vBuckets <- replicate(length(vec), "Average")

vBuckets[vec > 60] <- "High"

vBuckets[vec < 20] <- "Low"

return(vBuckets)

}

flight\_perYear <- createBucket2(df1$Flights.Per.Year)

#View(flight\_perYear )

age <- createBucket2(df1$Age)

#View(age)

shopping<- createBucket2(df1$Shopping.Amount.at.Airport)

#View(shopping)

eat\_drink <- createBucket1(df1$Eating.and.Drinking.at.Airport)

#View(eat\_drink )

dep\_delay <- createBucket1(df1$Departure.Delay.in.Minutes)

#View(dep\_delay )

flight\_time <- createBucket1(df1$Flight.time.in.minutes)

#View(flight\_time )

arrival\_delay <- createBucket1(df1$Arrival.Delay.in.Minutes)

#View(arrival\_delay )

createBucket3 <- function(vec){

q <- quantile(vec, c(0.4, 0.6)) # consideriong quantiles of 40% and 60%

vBuckets <- replicate(length(vec), "Average") #values greater than 40% and less than 60% quantile are average

vBuckets[vec <= q[1]] <- "Low" # values with less than or equal to 40%quantile are marked low

vBuckets[vec > q[2]] <- "High" #values greater than 60%quantile are marked high

return(vBuckets)

}

loyalty <- createBucket3(df1$Loyalty)

#View(loyalty )

ruleDF <- data.frame(happyCust, loyalty, age, flight\_perYear, shopping, eat\_drink, price\_sen, freq\_flyer, dep\_hour, arrival\_delay, flight\_time,dep\_delay )

#View(ruleDF)

df2Arule1 <- as(ruleDF, "transactions")

itemFrequency(df2Arule1 )

itemFrequencyPlot(df2Arule1)

#View(df2Arule1)

ruleSet1 <- apriori(df2Arule1,

parameter = list(support =0.1, confidence=.29),

appearance = list(default="lhs", rhs=("happyCust=Low"))

)

ruleSet1 #2rules

inspect(ruleSet1)

ruleSet1\_sorted <-ruleSet1[order(-quality(ruleSet1)$lift),]

inspect(head(ruleSet1\_sorted,15))

Important Rules

#############arules cheapseats#####################

> inspect(head(ruleSet1\_sorted,15))

lhs rhs support confidence lift count

[1] {flight\_perYear=Average,

eat\_drink=High,

freq\_flyer=Low} => {happyCust=Low} 0.1142292 0.2940516 1.360181 4449

[2] {flight\_perYear=Average,

freq\_flyer=Low,

arrival\_delay=High} => {happyCust=Low} 0.1142292 0.2940516 1.360181 4449

[3] {flight\_perYear=Average,

eat\_drink=High,

dep\_hour=Low} => {happyCust=Low} 0.1142292 0.2940516 1.360181 4449

[4] {flight\_perYear=Average,

dep\_hour=Low,

arrival\_delay=High} => {happyCust=Low} 0.1142292 0.2940516 1.360181 4449

[5] {flight\_perYear=Average,

eat\_drink=High,

freq\_flyer=Low,

dep\_hour=Low} => {happyCust=Low} 0.1142292 0.2940516 1.360181 4449

[6] {flight\_perYear=Average,

freq\_flyer=Low,

dep\_hour=Low,

arrival\_delay=High} => {happyCust=Low} 0.1142292 0.2940516 1.360181 4449

[7] {flight\_perYear=Average,

eat\_drink=High,

freq\_flyer=Low,

arrival\_delay=High} => {happyCust=Low} 0.1142292 0.2940516 1.360181 4449

[8] {flight\_perYear=Average,

eat\_drink=High,

freq\_flyer=Low,

flight\_time=High} => {happyCust=Low} 0.1142292 0.2940516 1.360181 4449

[9] {flight\_perYear=Average,

freq\_flyer=Low,

arrival\_delay=High,

flight\_time=High} => {happyCust=Low} 0.1142292 0.2940516 1.360181 4449

[10] {flight\_perYear=Average,

eat\_drink=High,

dep\_hour=Low,

arrival\_delay=High} => {happyCust=Low} 0.1142292 0.2940516 1.360181 4449

[11] {flight\_perYear=Average,

eat\_drink=High,

dep\_hour=Low,

flight\_time=High} => {happyCust=Low} 0.1142292 0.2940516 1.360181 4449

[12] {flight\_perYear=Average,

dep\_hour=Low,

arrival\_delay=High,

flight\_time=High} => {happyCust=Low} 0.1142292 0.2940516 1.360181 4449

[13] {flight\_perYear=Average,

eat\_drink=High,

freq\_flyer=Low,

dep\_hour=Low,

arrival\_delay=High} => {happyCust=Low} 0.1142292 0.2940516 1.360181 4449

[14] {flight\_perYear=Average,

eat\_drink=High,

freq\_flyer=Low,

dep\_hour=Low,

flight\_time=High} => {happyCust=Low} 0.1142292 0.2940516 1.360181 4449

[15] {flight\_perYear=Average,

freq\_flyer=Low,

dep\_hour=Low,

arrival\_delay=High,

flight\_time=High} => {happyCust=Low} 0.1142292 0.2940516 1.360181 4449

Customers who spent more in eat drink -- have rated a s low (actionable insight) : maybe when they check in they can be given a small meal at the counter so that they can munch on while waiting….

**VALIDATION**

**Support Vector Machines**

The SVm is used to analyze the data content or the quality of the data present in the variables. It is used for classifying the data based on after training dataset.It tells us which attributes are more important.

######################SVM Starts########################################

#Using the SVM model to predict the data

install.packages("kernlab")

library(kernlab)

View(df1)

str(df1)

trainindex <- sample(c(1,2,3), nrow(df1),replace= T,prob = c(0.15,0.45,0.4))

traindata <- df1[trainindex==1,] #17659 obs

testdata <- df1[trainindex==2,] # 5801 obs

svmOutput <- ksvm(Satisfaction ~ Airline.Status + Age + Gender + Price.Sensitivity +

Flights.Per.Year + Type.of.Travel+Loyalty+ Total.Freq.Flyer.Accts +

Shopping.Amount.at.Airport+ Eating.and.Drinking.at.Airport+Class+

+ Flight.date + Partner.Code +Partner.Name+ Orgin.City +Destination.City

+Scheduled.Departure.Hour+ Departure.Delay.in.Minutes

+ Arrival.Delay.in.Minutes+Flight.time.in.minutes+ Long.Duration.Trip ,

data=traindata,kernel="rbfdot", kpar="automatic",C=40,cross=4, prob.model=TRUE)

**svmOutput**

**#OBSERVATION::**

**#Number of Support Vectors : 4987**

**#Training error : 0.084237 ~ 8%**

**#Cross validation error : 0.689985 ~ 68%**

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

#View(svmresult)

str(svmresult) #Observation -- num [1:17659, 1]

head(svmresult)

happyPred <- svmresult[,1]

#View(happyPred)

happyPred[happyPred>.8] <- 1

happyPred[happyPred<.8] <- 0

#View(testdata)

str(testdata)

ctable <- data.frame(testdata$Satisfaction, happyPred)

table(ctable)

**# Observation**

**# happyPred**

**#testdata.Satisfaction 0 1**

**# 1 0 447**

**# 2 12 3440**

**# 3 3 4984**

**# 4 0 7122**

**# 5 0 1651**

#11. Calculate an error rate based on what you see in the confusion matrix. See pages 243-244 for more information.

errorrate <- ((table(ctable)[1,2] + table(ctable)[2,1])/(table(ctable)[1,1]+ table(ctable)[1,2] + table(ctable)[2,1] +table(ctable)[2,2]))\*100

**errorrate #We find that error rate is around 11.72% which is good since it means more than only 10% of our predictions are wrong**

**# SVM model for variables in lm and arules::**

svmOutput2<- ksvm(Satisfaction ~ Age + Price.Sensitivity +

Flights.Per.Year + Type.of.Travel+Loyalty+ Total.Freq.Flyer.Accts +

Eating.and.Drinking.at.Airport+Class+

+ Flight.date+Scheduled.Departure.Hour

+ Arrival.Delay.in.Minutes ,

data=traindata,kernel="rbfdot", kpar="automatic",C=40,cross=4, prob.model=TRUE)

**svmOutput2**

**#Number of Support Vectors : 5059**

**#Training error : 0.161649**

**#Cross validation error : 0.790994**

svmresult2 <- predict(svmOutput2,testdata,type="votes")

#View(svmresult2)

str(svmresult2) #Observation -- num [1:17659, 1]

head(svmresult2)

happyPred2 <- svmresult2[,1]

#View(happyPred2)

happyPred2[happyPred2>.8] <- 1

happyPred2[happyPred2<.8] <- 0

#View(testdata)

str(testdata)

ctable2 <- data.frame(testdata$Satisfaction, happyPred2)

table(ctable2)

**# happyPred2**

**#testdata.Satisfaction 0 1**

**# 1 0 447**

**# 2 9 3443**

**# 3 9 4978**

**# 4 0 7122**

**# 5 0 1651**

errorrate2 <- ((table(ctable2)[1,2] + table(ctable2)[2,1])/(table(ctable2)[1,1]+ table(ctable2)[1,2] + table(ctable2)[2,1] +table(ctable2)[2,2]))\*100

**errorrate2 #We find that error rate is around 11.69% which is good since it means more than only 10% of our predictions are wrong**

**TRELLO SCREENSHOTS**

We have used Trello the entire semester in order to keep track of the tasks completed by the group, which helped us to know what we have completed and what needs to be done which helped us to remain on track and complete the project tasks in time.

