IST 687 - Introduction to Data Science

Lab Section M004 | Group 1

Customer Survey Analysis of Airlines in the United States

Contents:

Introduction

Airline companies all around the world are in the constant process of improving their services to attract more customers to choose their airline over other airlines available in order to increase their revenue. With the introduction of latest technological methods, it is now possible to make this process easier. This is possible due to the ability to apply data analysis techniques on the feedback surveys given by the customers of their respective airlines.

We are similarly going to apply data analysis techniques on sample data of airlines given to us to predict the causes of high and low satisfaction of customers and ways to implement methods in order to attract more customers and keep them loyal to a specific airline.

Scope

The scope of this project is to analyze the customer survey data of different airlines in the United States and understand the key factors affecting the low satisfaction of customers. Further, we need to understand which major factors should be focused upon in order to help the Airlines take business decisions based on the actionable insights identified and increase its revenue.

In the project, the given data will go through data cleaning. After the process of data cleaning, different analysis models will be applied on the cleaned data. After this process, different kinds of visualizations will be created to identify the major factors responsible for the high and low satisfaction of customers.

Project Goals

The following are the project goals:

1. To analyze satisfaction of customers flying within the United States
2. To generate actionable insights from our analysis
3. To identify drivers to improve customer satisfaction

Business Questions

While doing the project, we were successful in analyzing the data i.e. the factors which are responsible for the different responses in the customer survey data of different airlines. We succeeded in doing so by carefully cleaning the data of all the airlines and then applying various models on the cleaned data of the airline where we determined the most satisfied customers and the least satisfied customers after extrapolating the most significant attributes from the data.

Further we were able to gain actionable insights from the significant factors which are responsible for the low customer satisfaction of the airlines thus having all the answers to the business questions addressed.

The following are the business questions that have been answered after doing this project:

1. Which airlines have most satisfied/ happy customers?
2. Which airlines have least satisfied/ unhappy customers?
3. Which factors influence the customer satisfaction?

Data Acquisition

The data of customer surveys of different Airlines in the United States was provided to us by the Professor. The data provided to us contained 129889 survey responses and 28 attributes. The data contained customer ratings ranging from 1 to 5 for fourteen airlines. The dataset also contained a lot of interesting attributes such as type of travel, Airline Status, Arrival_time_greater_than_5, Flight Cancelled, Origin City, Destination City, Age & Gender of the customer, Class and many more such attributes.

Data Cleaning

There were 7000 rows where variables Departure delay in Minute, Arrival Delay in Minutes and Flight time in minutes contained NA's where Flight cancelled corresponded to yes. Hence, we converted them to zero instead of removing them. Further, the column customer satisfaction contained some garbage values which we decided to remove in order to clean the dataset. Further the data type of customer satisfaction was in the format of factors which we then converted to numeric.

Thus, after data cleaning, the remaining total number of survey responses left were 129549.

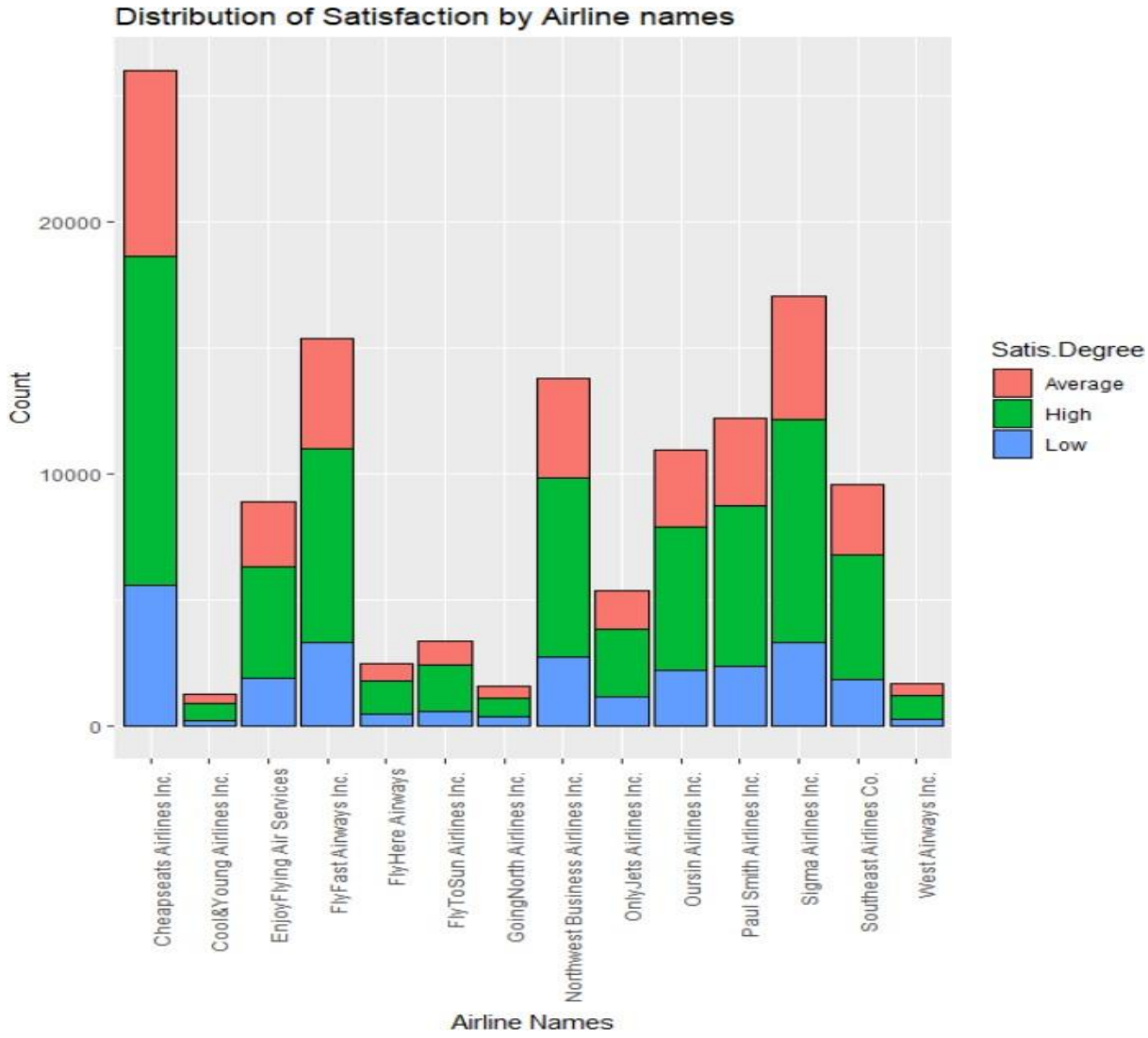
```
# Data cleaning

# Filling in some missing values
df$Departure.Delay.in.Minutes[which(is.na(df$Departure.Delay.in.Minutes) & (df$Flight.cancelled == 'Yes'))] <- 0
df$Arrival.Delay.in.Minutes[which(is.na(df$Arrival.Delay.in.Minutes) & (df$Flight.cancelled == 'Yes'))] <- 0
df$Flight.time.in.minutes[which(is.na(df$Flight.time.in.minutes) & (df$Flight.cancelled == 'Yes'))] <- 0

# Numericalize satisfaction
df$Satisfaction <- as.numeric(as.character(df$Satisfaction))

# Dumping all the rows with missing values
ndf <- na.omit(df)
```

Descriptive Statistics

Descriptive statistics are used to describe the basic features of the data. They provide summaries about the sample and the measures. Together with graphical analysis, they form the basis of every quantitative analysis of data.

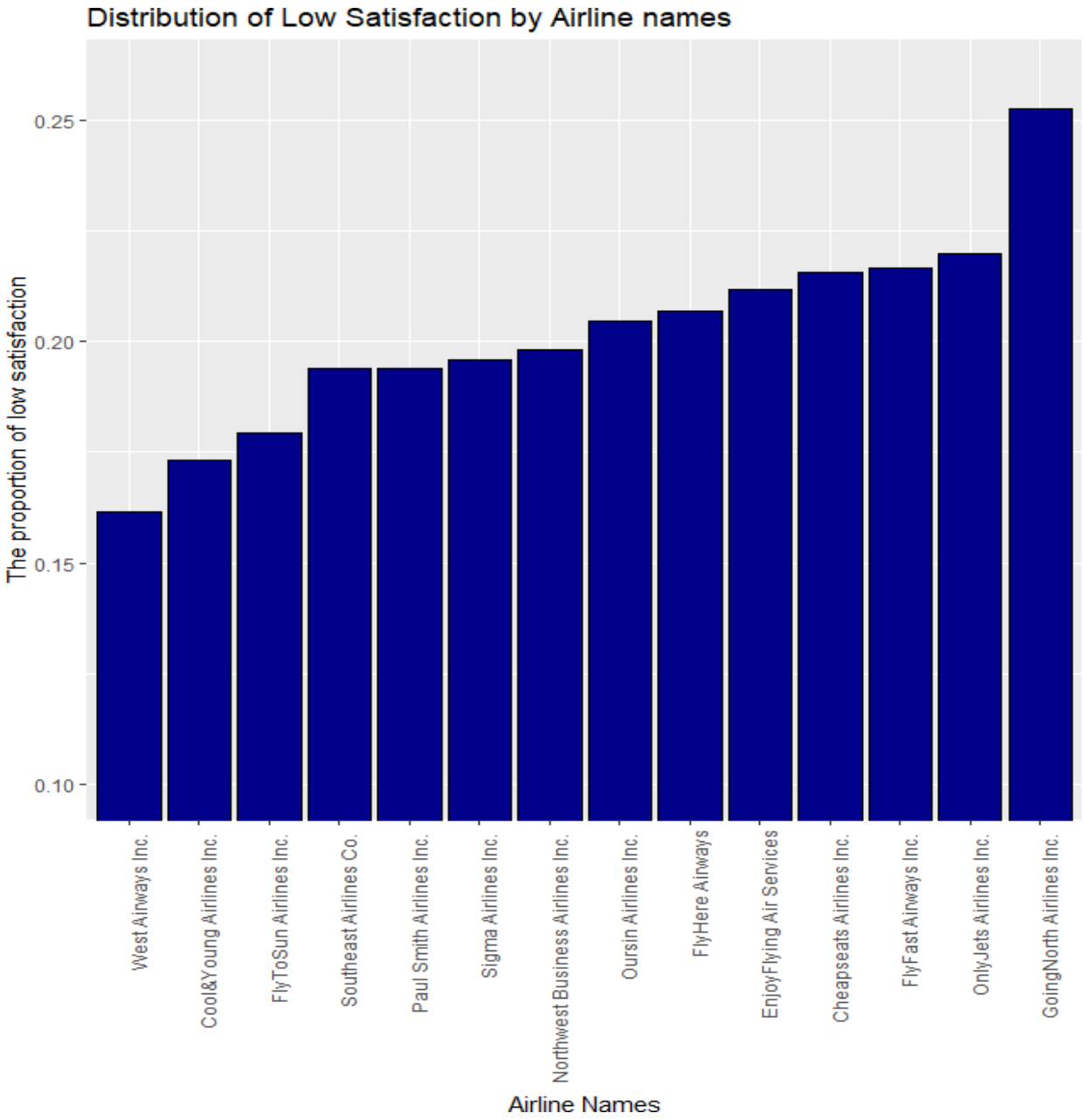
```
#package
library(ggplot2)
library(MASS)

#read data
df <- read.csv("C:/Users/Spencer/Desktop/IST-687/Satisfaction Survey.csv")

#clean missing value
df$Departure.Delay.in.Minutes[which(is.na(df$Departure.Delay.in.Minutes) & (df$Flight.cancelled == 'Yes'))] <- 0
df$Arrival.Delay.in.Minutes[which(is.na(df$Arrival.Delay.in.Minutes) & (df$Flight.cancelled == 'Yes'))] <- 0
df$Flight.time.in.minutes[which(is.na(df$Flight.time.in.minutes) & (df$Flight.cancelled == 'Yes'))] <- 0

#change the data type of 'Satisfaction' to numeric
df$Satisfaction <- as.numeric(as.character(df$Satisfaction))

#omit the missing value
ndf <- na.omit(df)

#get the names of airlines
airline.name <- c(levels(ndf$Airline.Name))

#Insert a new column descriping the degree of satisfaction
ndf$degree <- NA
ndf$Satisfaction <- as.numeric(as.character(ndf$Satisfaction))
ndf$degree[which(ndf$Satisfaction >=4)]<- "High"
ndf$degree[which(ndf$Satisfaction ==3)] <- "Average"
ndf$degree[which(ndf$Satisfaction ==3.5)] <- "Average"
ndf$degree[which(ndf$Satisfaction <3)] <- "Low"

ndf <- na.omit(ndf)

#Build a table counting the number of each degree of satisfaction grouped by airline names
Freq <- as.data.frame(table(ndf$Airline.Name,ndf$degree))
colnames(Freq) <- c("Name","Satis.Degree","Count")
Freq

#Draw a bar chart for all airline companies showing the distribution of satisfaction
Satis_Bar <- ggplot(Freq,aes(x=Name,y=Count,fill=Satis.Degree)) +
  geom_bar(stat="identity",color="black") +
  theme(axis.text.x = element_text(angle=90,hjust = 1)) +
  ggtitle("Distribution of Satisfaction by Airline names")+
  labs(x = "Airline Names", y = "Count")
Satis_Bar
```

```
#Build a propotion table of satisfaction grouped by company
P1 <- prop.table(table(ndf$Airline.Name,ndf$degree),1)
P1
Prop <- data.frame(matrix(data = P1, nrow= length(airline.name), ncol = 3, byrow = F))
Prop
colnames(Prop) <- c('Average','High','Low')
row.names(Prop) <- c(airline.name)
Prop

#perfect group
HighBar <- ggplot(Prop,aes(x=reorder(airline.name,High),y=High)) + geom_col(color="black",fill=" dark green") +
  theme(axis.text.x = element_text(angle=90,hjust = 1))+
  ggtitle("Distribution of High Satisfaction by Airline names")+
  labs(x = "Airline Names", y = "The porpotion of high satisfaction")+
  coord_cartesian(ylim = c(0.4,0.6))
HighBar
#The west Ailrlines has the highest porpotion of high satisfaction
#This company is the most satisfacted company

#soso group
AverageBar <- ggplot(Prop,aes(x=reorder(airline.name,Average),y=Average)) + geom_col(color="black",fill="dark red") +
  theme(axis.text.x = element_text(angle=90,hjust = 1))+
  ggtitle("Distribution of Average Satisfaction by Airline names")+
  labs(x = "Airline Names", y = "The porpotion of average satisfaction")+
  coord_cartesian(ylim = c(0.25,0.3))
AverageBar

#unhappy group
LowBar <- ggplot(Prop,aes(x=reorder(airline.name,Low),y=Low)) + geom_col(color="black",fill="dark blue") +
  theme(axis.text.x = element_text(angle=90,hjust = 1))+
  ggtitle("Distribution of Low Satisfaction by Airline names")+
  labs(x = "Airline Names", y = "The porpotion of low satisfaction")+
  coord_cartesian(ylim = c(0.1,0.26))
LowBar
#The GoingNorth Airlines Inc. has the highest porpotion of low satisfaction.
#So this company is the least satisfacted company.
```



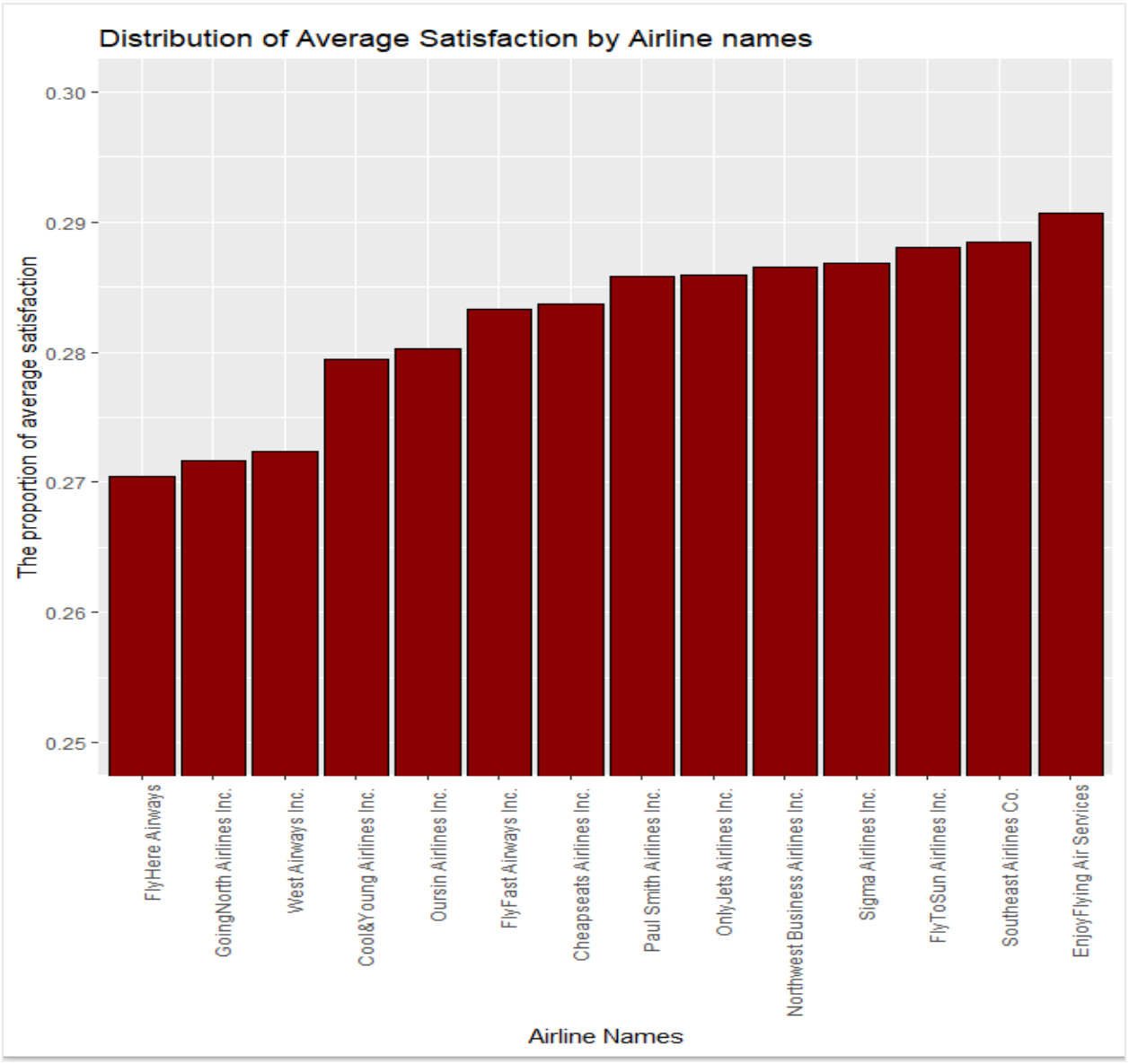
The above visualization gives us the overall count of customers on the basis of degree of satisfaction for all airlines i.e. “Average”, “High” and “Low”.

Based on the count of the customer satisfaction of all the airlines, we cannot conclude the airline with the highest or lowest percentage of customers as this is a count of all the reviews and not the proportion of the customer survey count of each airline. So to find out the proportion we used the prop.table() function.



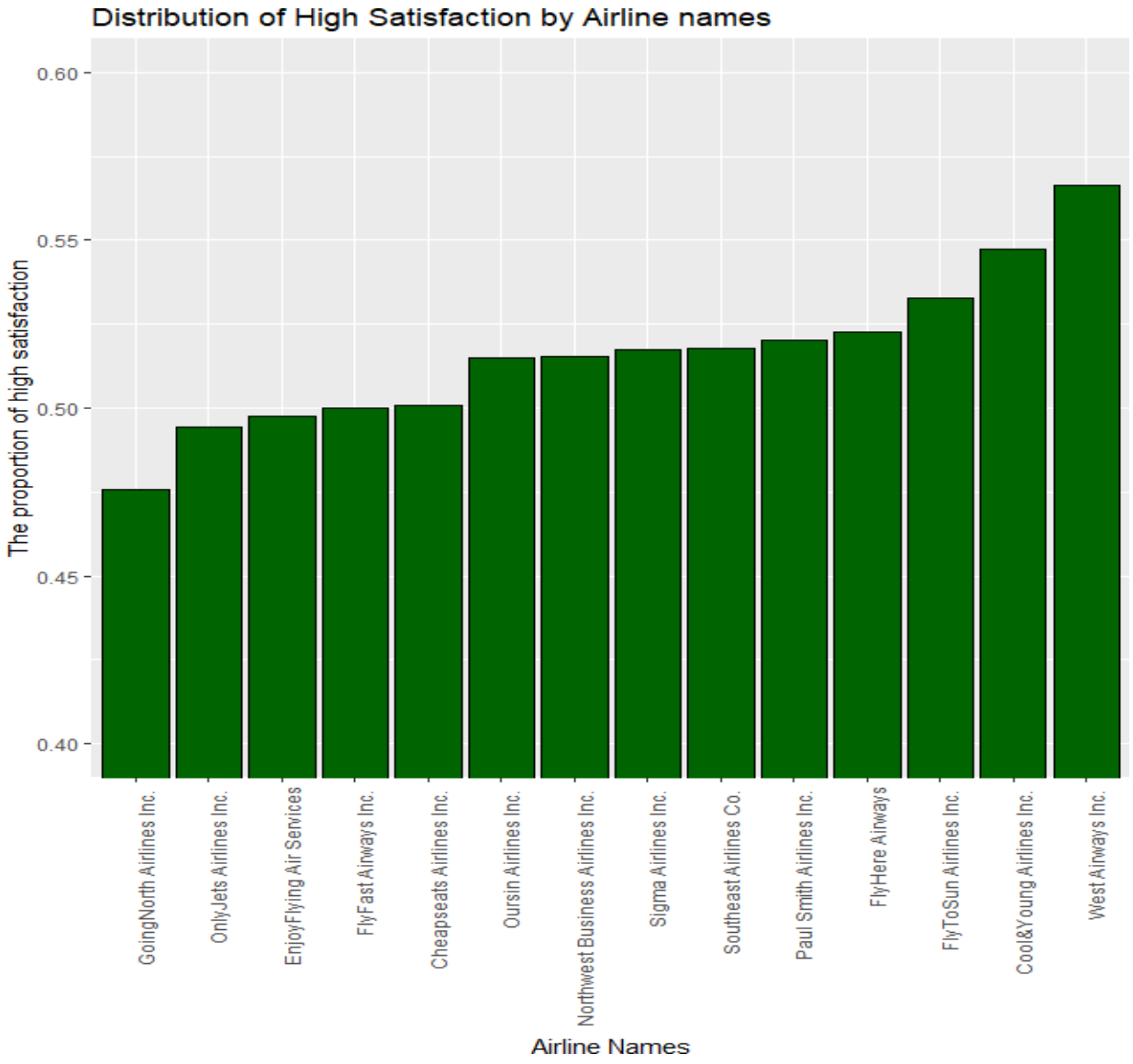
The above visualization gives us the distribution of the proportion of low satisfaction of customers by their respective airline. Here, we can get an estimate of the number of customers who are not satisfied by their respective airline in comparison to others.

According to the visualization, GoingNorth Airlines has the highest proportion of customers who are not satisfied with their airline while West Airways has the lowest proportion of customers who are not satisfied with their airline.



The above visualization gives us the distribution of the proportion of average satisfaction of customers by their airline count. Here, we can get an estimate of the number of customers who are not completely satisfied by their respective airline in comparison to others. Southeast Airlines and EnjoyFlying Airlines has the highest

proportion of customers who were not completely satisfied by their respective airlines.



The above visualization gives us the distribution of proportion of high satisfaction of customers by their airline count. Here, we can get an estimate of the number of customers who are completely satisfied by their respective airline in comparison to others.

According to the visualization, GoingNorth Airlines has the lowest proportion of customers who are completely satisfied with their airline while West Airways has the highest proportion of customers who are completely satisfied with their airline.

Modelling Techniques and Visualizations

Stepwise Linear Regression

Stepwise Regression is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure. In each step, a variable is considered for addition to(forward stepwise linear regression) or subtraction from(backward stepwise linear regression) the set of explanatory variables based on some prespecified criterion(mean R-squared value should increase).

We have used the stepAIC function for performing the stepwise linear regression and in that function we have used the forward stepwise linear regression. We get a set of significant attributes once the stepAIC function is executed.

Initially we used the stepwise linear regression for all the airlines present in the dataset. In doing so we found out that there are 3 attributes that are significant in all the outputs of those stepwise models i.e. type of travel, Airline status and arrival_delay_greater_5_mins.

In the next step we tried to draw comparisons of these 3 attributes for the two airlines that had the most satisfied and the least satisfied customers i.e. West Airways and the GoingNorth Airlines respectively.

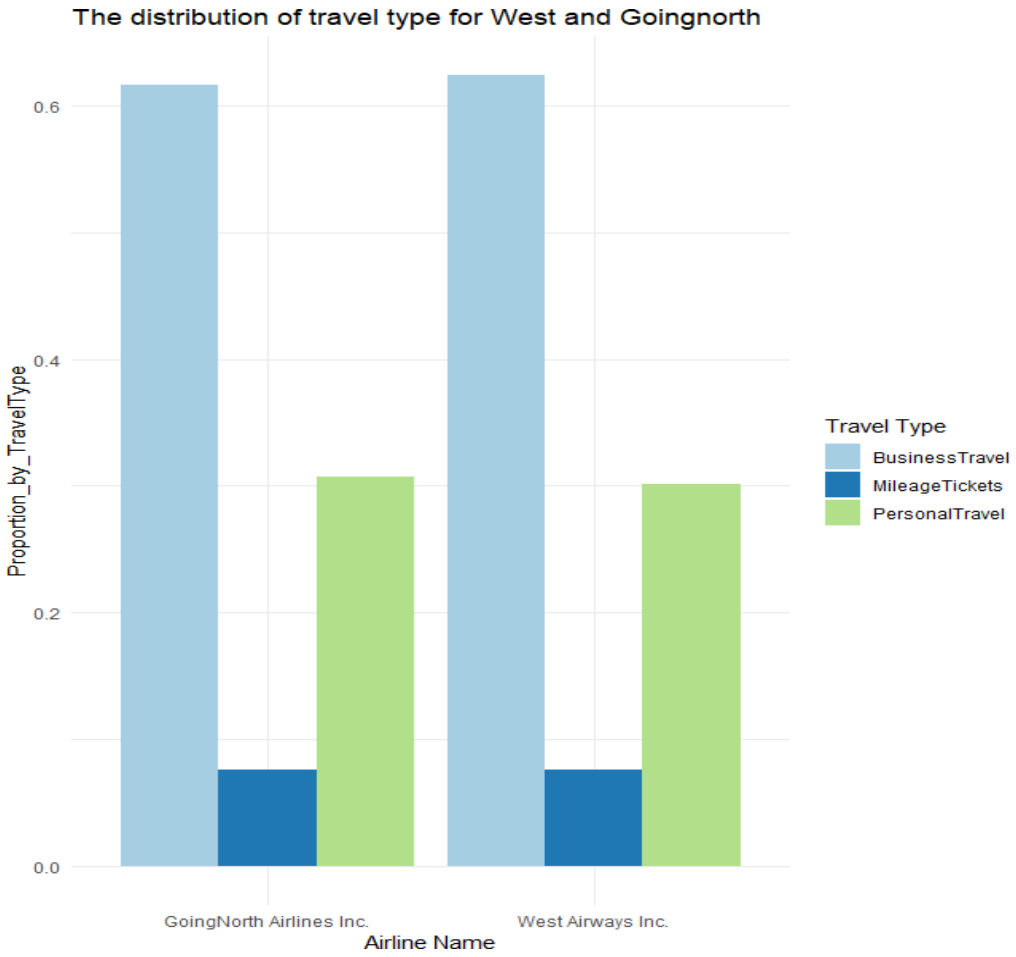
```
#After analyzing the plot, we want to what makes some airline companies having more high satisfaction than others
#So we will build linear models for each airline companies
#subset the data by company
which(colnames(df) == "Airline.Code")
nndf <- df[,-16]
nndf <- na.omit(nndf)
```

```
Cheapseats <- data.frame(subset(nndf, Airline.Name == "Cheapseats Airlines Inc. ")) #Cheapseats Airlines Inc.
CoolYoung <- subset(nndf, Airline.Name == "Cool&Young Airlines Inc. ") #Cool&Young Airlines Inc.
EnjoyFlying <- subset(nndf, Airline.Name == "EnjoyFlying Air Services") #EnjoyFlying Air Services
FlyFast <- subset(nndf, Airline.Name == "FlyFast Airways Inc. ") #"FlyFast Airways Inc. "
FlyHere <- subset(nndf, Airline.Name =="FlyHere Airways") #"FlyHere Airways"
FlyToSun <- subset(nndf, Airline.Name =="FlyToSun Airlines Inc. ") #"FlyToSun Airlines Inc. "
GoingNorth <- subset(nndf, Airline.Name == "GoingNorth Airlines Inc. ") #"GoingNorth Airlines Inc. "
Northwest <- subset(nndf, Airline.Name =="Northwest Business Airlines Inc. ") #"Northwest Business Airlines Inc. "
OnlyJets <- subset(nndf, Airline.Name == "OnlyJets Airlines Inc. ") #"OnlyJets Airlines Inc. "
Oursin <- subset(nndf, Airline.Name == "Oursin Airlines Inc. ") #"Oursin Airlines Inc. "
PaulSmith <- subset(nndf, Airline.Name == "Paul Smith Airlines Inc. ") #"Paul Smith Airlines Inc. "
Sigma <- subset(nndf, Airline.Name == "Sigma Airlines Inc. " ) #"Sigma Airlines Inc. "
Southeast <- subset(nndf, Airline.Name == "Southeast Airlines Co. ") #"Southeast Airlines Co. "
West <- subset(nndf, Airline.Name == "West Airways Inc. ") #"West Airways Inc. "

#use stepwise theory to find the important variables for each company and then build the final model

#build function for stepwise theory
which(colnames(nndf) == "Airline.Name") #16
stepwise <- function(myVec){
  DeleteName <- myVec[,-16]
  FullModel <- lm(Satisfaction~.,data = DeleteName)
  StepModel <- stepAIC(FullModel,direction = "forward", trace = F)
  Result <- which(summary(StepModel)$coef[,4] <= 0.001)
  return(Result)
}

#linear model for each company
#Cheapseats Airlines Inc.
stepwise(Cheapseats)
lm_Cheapseats <- lm(Satisfaction ~ Airline.Status+Age+Gender+Price.Sensitivity+
                      No.of.Flights.p.a.+Type.of.Travel+Class+
                      Flight.date+Scheduled.Departure.Hour+Arrival.Delay.greater.5.Mins,
                    data = Cheapseats)
try <- plot_summs(lm_Cheapseats, scale = TRUE, plot.distributions = TRUE, inner_ci_level = .9)

#Cool&Young Airlines Inc.
stepwise(CoolYoung)
lm_CoolYoung <- lm(Satisfaction ~ Airline.Status+Type.of.Travel+
                     Arrival.Delay.greater.5.Mins, data = CoolYoung)
```

```
#EnjoyFlying Air Services
stepwise(EnjoyFlying)
lm_EnjoyFlying <- lm(Satisfaction ~ Airline.Status+Age+Gender+
                       No.of.Flights.p.a.+Type.of.Travel+Class+
                       Arrival.Delay.greater.5.Mins, data = EnjoyFlying)

#"FlyFast Airways Inc. "
stepwise(FlyFast)
lm_FlyFast <- lm(Satisfaction ~ Airline.Status+Age+Gender+No.of.Flights.p.a.+
                   Type.of.Travel+Class+Flight.cancelled+Arrival.Delay.greater.5.Mins,
                 data = FlyFast)

#"FlyHere Airways"
stepwise(FlyHere)
lm_FlyHere <- lm(Satisfaction ~ Airline.Status+Age+No.of.Flights.p.a.+Type.of.Travel
                 +Arrival.Delay.greater.5.Mins, data = FlyHere)

#"FlyToSun Airlines Inc. "
stepwise(FlyToSun)
lm_FlyToSun <- lm(Satisfaction ~ Airline.Status+Gender+Type.of.Travel+
                    Arrival.Delay.greater.5.Mins, data = FlyToSun)

#"GoingNorth Airlines Inc. "
stepwise(GoingNorth)
lm_GoingNorth <- lm(Satisfaction ~ Airline.Status+Type.of.Travel+
                      Arrival.Delay.greater.5.Mins, data = GoingNorth)
#"Northwest Business Airlines Inc. "
stepwise(Northwest)
lm_Northwest <- lm(Satisfaction ~ Airline.Status+Age+Gender+Price.Sensitivity+
                     No.of.Flights.p.a.+Type.of.Travel+Class+Arrival.Delay.greater.5.Mins,
                   data = Northwest)

#"OnlyJets Airlines Inc. "
stepwise(OnlyJets)
lm_OnlyJets <- lm(Satisfaction ~ Airline.Status+Age+Gender+No.of.Flights.p.a.+
                    Type.of.Travel+Arrival.Delay.greater.5.Mins, data = OnlyJets)

#"Oursin Airlines Inc. "
stepwise(Oursin)
lm_Oursin <- lm(Satisfaction ~ Airline.Status+Age+Gender+No.of.Flights.p.a.+
                  Type.of.Travel+Class+Arrival.Delay.greater.5.Mins,
                data = Oursin)
```

```
#"Paul Smith Airlines Inc. "
stepwise(PaulSmith)
lm_PaulSmith <- lm(Satisfaction ~ Airline.Status+Gender+Year.of.First.Flight+
                     No.of.Flights.p.a.+Type.of.Travel+Scheduled.Departure.Hour+
                     Arrival.Delay.greater.5.Mins, data = PaulSmith)

#"Sigma Airlines Inc. "
stepwise(Sigma)
lm_Sigma <- lm(Satisfaction ~ Airline.Status+Age+Gender+Price.Sensitivity+
                 No.of.Flights.p.a.+Type.of.Travel+Arrival.Delay.greater.5.Mins,
                 data = Sigma)

#"Southeast Airlines Co. "
stepwise(Southeast)
lm_Southeast <- lm(Satisfaction ~ Airline.Status+Age+Gender+No.of.Flights.p.a.+
                     Type.of.Travel+Flight.cancelled+Arrival.Delay.greater.5.Mins,
                   data = Southeast)

#"West Airways Inc. --- no flight cancelled
which(colnames(West)=="Flight.cancelled") #24
WestDeleted <- West[,-24]
stepwise(WestDeleted)
lm_West <- lm(Satisfaction ~ Airline.Status+Gender+Type.of.Travel+
                Arrival.Delay.greater.5.Mins, data = WestDeleted)

#All the models have 'Airline.Status', 'Type.of.Travel', 'Arrival.Delay.greater.5.Mins'
#Compare those three attributes of the most satisfacted and the least satisfacted company

#Relation between 'Airline.Status' and 'Satisfaction
prop.status <- function(vec){
  co.status <- prop.table(table(vec$Airline.Status))
  m <- matrix(co.status,ncol = 1,nrow = 4)
  return(m)
}
w <- prop.status(West)
n <- prop.status(GoingNorth)
wn <- c(w,n)
s.wn <- rep(c("Blue","Gold","Platinum","Silver"),times = 2)
co.wn <- rep(c("West Airways Inc.","GoingNorth Airlines Inc."),each = 4)
co.wn
status <- data.frame(co.wn,s.wn,wn)
colnames(status) <- c("AirlineName","AirlineStatus","StatusPorpotionByAirline")
status
```

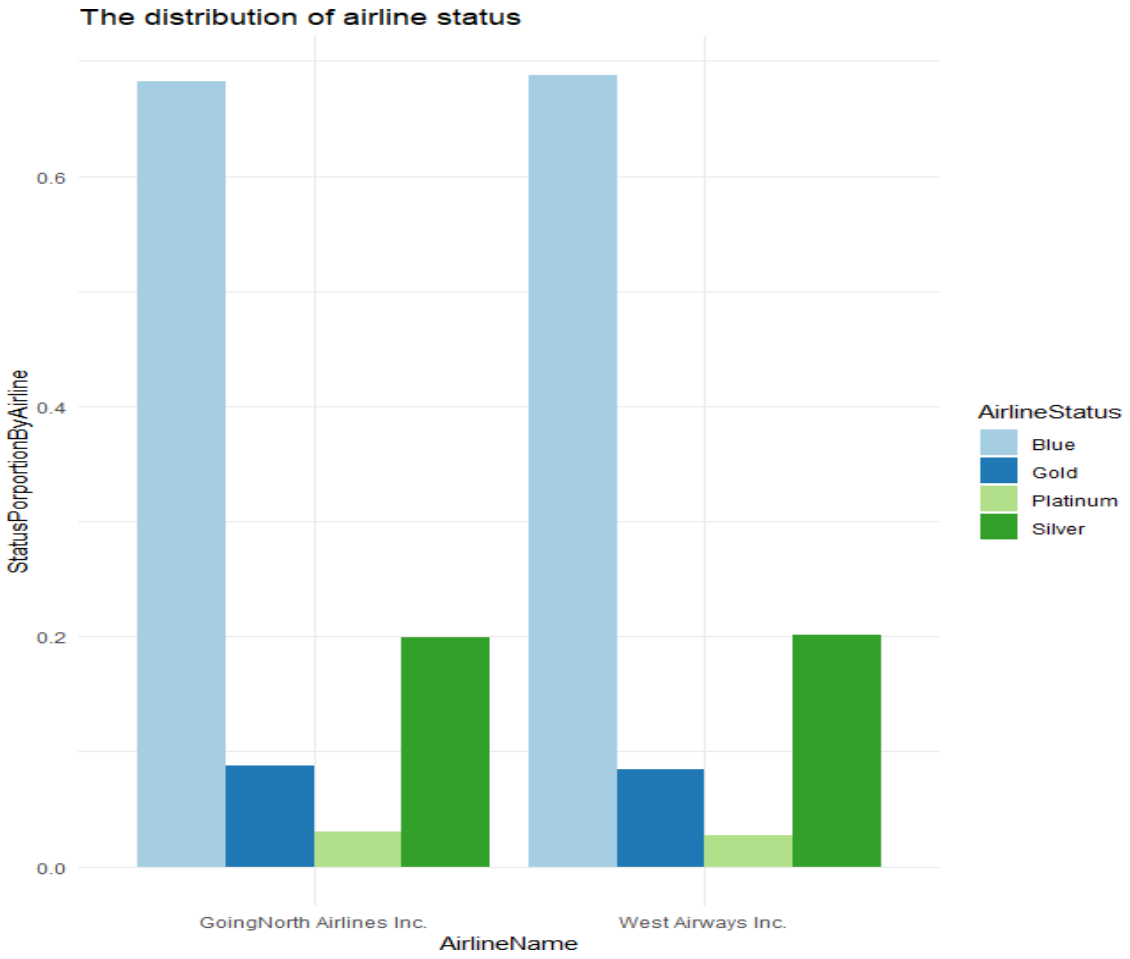
```
bar.status <- ggplot(status, aes(x=AirlineName, y=StatusPorpotionByAirline,fill=AirlineStatus)) +
  geom_bar(stat="identity", position=position_dodge())+
  scale_fill_brewer(palette="Paired")+
  ggtitle("The distribution of airline status")+
  theme_minimal()
bar.status

#Relation between 'Type.of.Travel' and 'Satisfaction'
prop.type <- function(vec){
  co.type <- prop.table(table(vec$Type.of.Travel))
  m <- matrix(co.type,ncol = 1,nrow = 3)
  return(m)
}
wt <- prop.type(West)
nt <- prop.type(GoingNorth)
wnt <- c(wt,nt)
t.wn <- rep(c("BusinessTravel", "MileageTickets", "PersonalTravel"),times = 2 )
co.t <- rep(c("West Airways Inc.","GoingNorth Airlines Inc."),each = 3)
type <- data.frame(co.t,t.wn,wnt)
bar.type <- ggplot(type, aes(x=co.t, y=wnt,fill=t.wn)) +
  geom_bar(stat="identity", position=position_dodge())+
  scale_fill_brewer(palette="Paired")+
  ggtitle("The distribution of travel type for West and Goingnorth")+
  labs(x="Airline Name", y="Porpotion",fill="Travel Type")+
  theme_minimal()
bar.type

#Relation between  'Arrival.Delay.greater.5.Mins' and 'Satisfaction'
prop.delay <- function(vec){
  co.type <- prop.table(table(vec$Arrival.Delay.greater.5.Mins))
  m <- matrix(co.type,ncol = 1,nrow = 2)
  return(m)
}
delay.w <- prop.delay(West)
delay.n <- prop.delay(GoingNorth)
prop.table(table(GoingNorth$Arrival.Delay.greater.5.Mins))
delay.wn <- c(delay.w,delay.n)
DelayGreater5Mins <- rep(c("NO","YES"),times=2)
co.d <- rep(c("West Airways Inc.","GoingNorth Airlines Inc."),each = 2)
delay5min <- data.frame(co.d,DelayGreater5Mins,delay.wn)
bar.delay5min <- ggplot(delay5min, aes(x=co.d, y=delay.wn,fill=DelayGreater5Mins)) +
  geom_bar(stat="identity", position=position_dodge())+
  scale_fill_brewer(palette="Paired")+
  ggtitle("The distribution of arrival delay greater than 5 mins")+
  labs(x="Airline Name", y="Porpotion",fill="Is delayed")+
  theme_minimal()
bar.delay5min
```



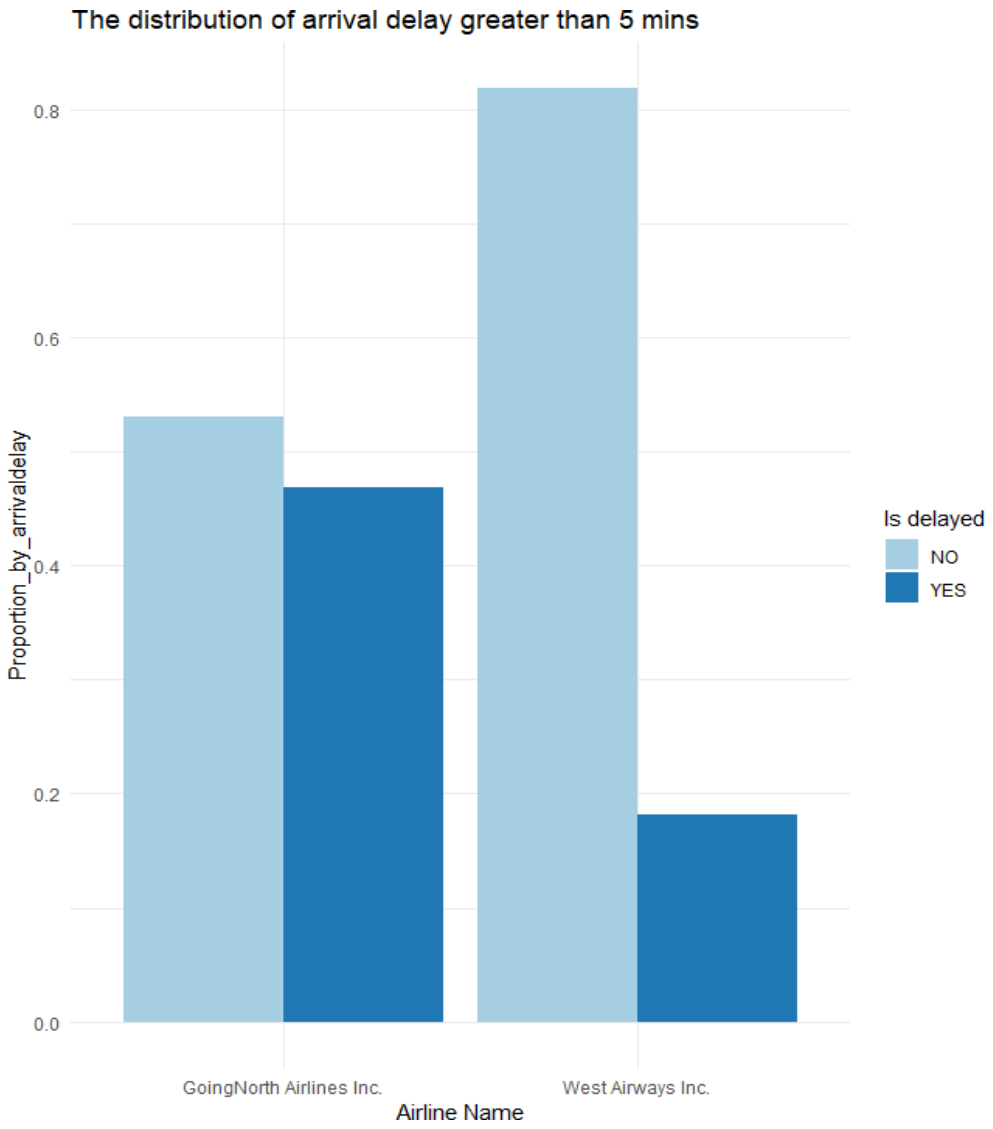
Here, we have implemented a linear model for every airline. We analyzed that Business Travel, Mileage tickets and Personal Travel are the most significant attributes that would affect the customer satisfaction.

Hence, we have chosen the airlines with the highest number of satisfied customers i.e. West Airways and airline with lowest number of satisfied customers i.e. GoingNorth Airlines.

The above visualization shows the comparison of distribution of travel type consisting of three categories which are Business Travel, Mileage Tickets and Personal travel of West Airways and GoingNorth Airlines



The above visualization shows the distribution of Airline Status consisting of four categories which are Blue, Gold, Platinum and Silver for GoingNorth Airlines and West Airways. Concluding from the above visualization, we cannot accurately depict which airline status has led to more or less customer satisfaction as there is not a significant difference to establish the right predictor.



Here we compare GoingNorth Airlines and West Airways with respect to arrival delay greater than 5 mins. The above visualization shows the distribution of arrival delay of GoingNorth Airlines and West Airways. Here we can infer that that arrival delay greater than 5mins is the most significant attribute. We can infer from the comparison that GoingNorth Airlines has a large proportion of flights where the arrival delay time was more than 5 mins. West Airways has a significantly lower proportion of flights which were delayed.

Association Rule Mining

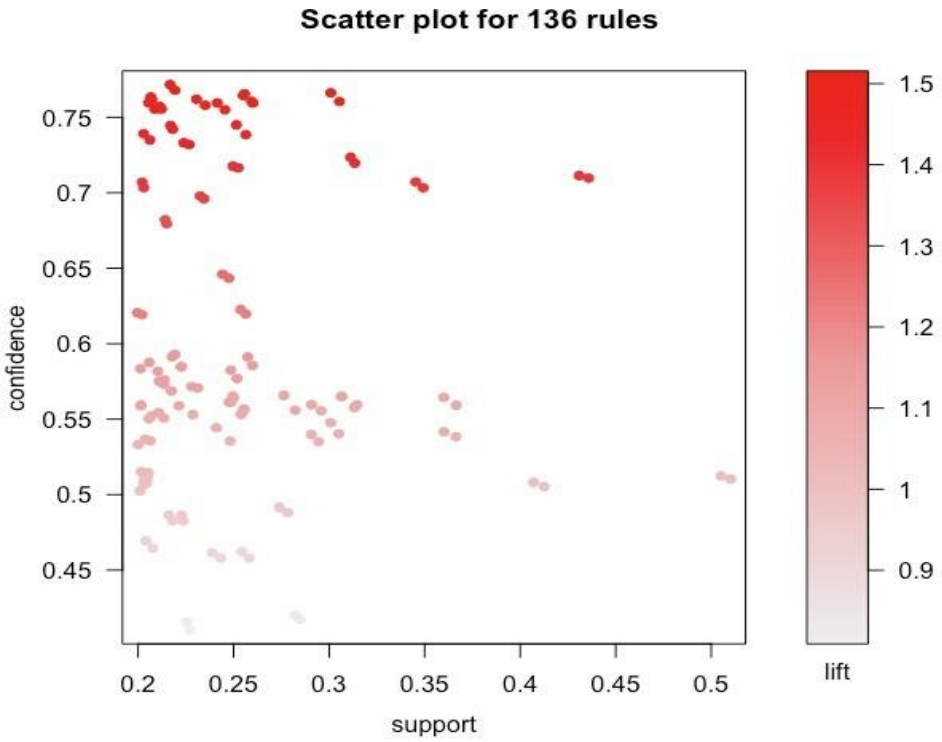
Association rule Mining is a rule-based machine learning method for discovering interesting relations between variables in large databases.

In this modeling technique, we have made two functions for bucketing that are CreateBucketSat() and CreateBuckets(). CreateBucketSat() function is used to assign Happy, Average and Low values to customer satisfaction column depending upon their values while CreateBuckets() function is used to assign high, average and low values to other numeric columns based on quantiles. Then, we apply bucketing functions to satisfaction and other independent attributes. Further, we make a dataframe for these bucketed attributes and use it for executing apriori function. Apriori function gives us the rulesets having attributes affecting customer satifaction.

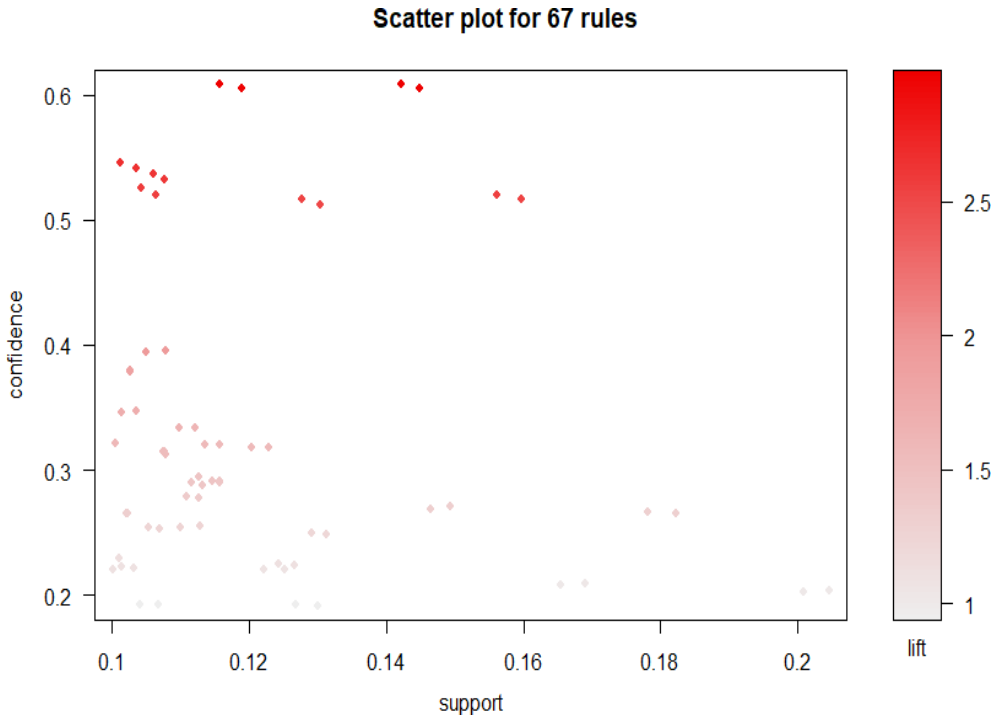
```
#Installing packages
install.packages("ggplot2")
install.packages("arules")
install.packages("arulesviz")

# Loading packages
library('ggplot2')
library('arules')
library('arulesViz')

# Loading data
# Set the working directory where your dataset is saved
setwd("D:/Syracuse University/Fall'18/IST 687/Project")
df <- read.csv("Satisfaction Survey.csv")

# Data cleaning

# Filling in some missing values
df$Departure.Delay.in.Minutes[which(is.na(df$Departure.Delay.in.Minutes) & (df$Flight.cancelled == 'Yes'))] <- 0
df$Arrival.Delay.in.Minutes[which(is.na(df$Arrival.Delay.in.Minutes) & (df$Flight.cancelled == 'Yes'))] <- 0
df$Flight.time.in.minutes[which(is.na(df$Flight.time.in.minutes) & (df$Flight.cancelled == 'Yes'))] <- 0

# Numericalize satisfaction
df$Satisfaction <- as.numeric(as.character(df$Satisfaction))

# Dumping all the rows with missing values
ndf <- na.omit(df)

# Bucketing

# Satisfaction bucketing
createBucketSat <- function(vec){
  vBuckets <- replicate(length(vec), "Average")
  vBuckets[vec >= 4] <- "Happy"
  vBuckets[vec < 3] <- "Unhappy"
  return(vBuckets)
}

# Bucketing numeric numbers
createBuckets <- function(vec){
  q <- quantile(vec, c(0.4, 0.6), na.rm = "TRUE")
  vBuckets <- replicate(length(vec), "Average")
  vBuckets[vec <= q[1]] <- "Low"
  vBuckets[vec > q[2]] <- "High"
  return(vBuckets)
}
```

```
ndf$Satisfaction <- createBucketSat(ndf$Satisfaction)
ndf$Age <- createBuckets(ndf$Age)
ndf$Price.Sensitivity <- as.factor(ndf$Price.Sensitivity)
ndf$No.of.Flights.p.a. <- createBuckets(ndf$No.of.Flights.p.a.)
ndf$X..of.Flight.with.other.Airlines <- createBuckets(ndf$X..of.Flight.with.other.Airlines)
ndf$No..of.other.Loyalty.Cards <- createBuckets(ndf$No..of.other.Loyalty.Cards)
ndf$Shopping.Amount.at.Airport <- createBuckets(ndf$Shopping.Amount.at.Airport)
ndf$Eating.and.Drinking.at.Airport <- createBuckets(ndf$Eating.and.Drinking.at.Airport)
ndf$Departure.Delay.in.Minutes <- createBuckets(ndf$Departure.Delay.in.Minutes)
ndf$Arrival.Delay.in.Minutes <- createBuckets(ndf$Arrival.Delay.in.Minutes)
ndf$Flight.time.in.minutes <- createBuckets(ndf$Flight.time.in.minutes)
ndf$Flight.Distance <- createBuckets(ndf$Flight.Distance)
ndf$Scheduled.Departure.Hour <- as.factor(ndf$Scheduled.Departure.Hour)
ndf$Day.of.Month <- as.factor(ndf$Day.of.Month)
ndf$Year.of.First.Flight <- as.factor(ndf$Year.of.First.Flight)

# Build dataframe for a-rule
aprioriDF <- data.frame(ndf$Satisfaction,
                        ndf$Airline.Name,
                        ndf$Age,
                        ndf$Price.Sensitivity,
                        ndf$No.of.Flights.p.a.,
                        ndf$X..of.Flight.with.other.Airlines,
                        ndf$No..of.other.Loyalty.Cards,
                        ndf$Shopping.Amount.at.Airport,
                        ndf$Eating.and.Drinking.at.Airport,
                        ndf$Departure.Delay.in.Minutes,
                        ndf$Arrival.Delay.in.Minutes,
                        ndf$Flight.time.in.minutes,
                        ndf$Airline.Status,
                        ndf$Gender,
                        ndf$Type.of.Travel,
                        ndf$Class,
                        ndf$Flight.cancelled,
                        ndf$Arrival.Delay.greater.5.Mins,
                        ndf$Flight.time.in.minutes,
                        ndf$Flight.Distance,
                        ndf$Orgin.City,
                        ndf$Origin.State,
                        ndf$Destination.City,
                        ndf$Destination.State,
                        ndf$Scheduled.Departure.Hour,
                        ndf$Day.of.Month,
                        ndf$Year.of.First.Flight)
```

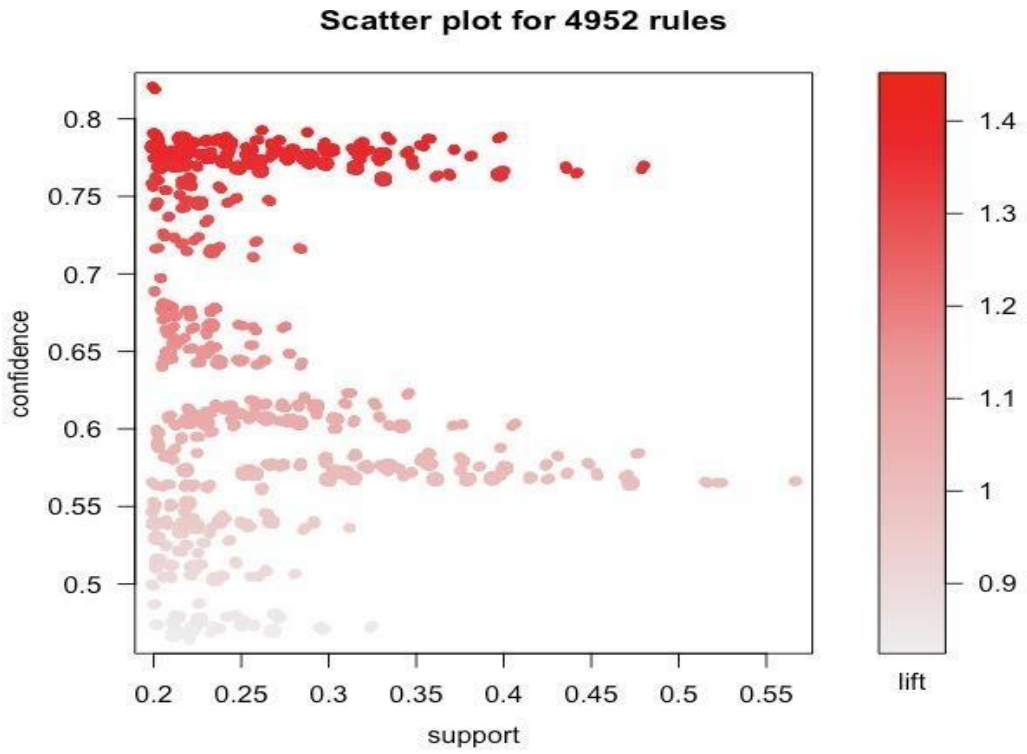
```
# Analyze for the whole dataset

ruleset <- apriori(aprioriDF, parameter = list(support = 0.2, confidence = 0.2), appearance = list(default = 'lhs', rhs = "ndf.Satisfaction=Happy"))
plot(ruleset)

ruleset <- apriori(aprioriDF, parameter = list(support = 0.1, confidence = 0.1), appearance = list(default = 'lhs', rhs = "ndf.Satisfaction=Unhappy"))
plot(ruleset)

# Seperating the dataset into different airline
# With analysis

# Happy one
#"West Airways Inc. "
West <- subset(aprioriDF, aprioriDF$ndf.Airline.Name == "West Airways Inc. ", select = -c(ndf.Airline.Name))

ruleset <- apriori(West, parameter = list(support = 0.2, confidence = 0.2), appearance = list(default = 'lhs', rhs = "ndf.Satisfaction=Happy"))
plot(ruleset)

# Unhappy one
#"GoingNorth Airlines Inc. "
GoingNorth <- subset(aprioriDF, aprioriDF$ndf.Airline.Name == "GoingNorth Airlines Inc. ", select = -c(ndf.Airline.Name))

ruleset <- apriori(GoingNorth, parameter = list(support = 0.1, confidence = 0.1), appearance = list(default = 'lhs', rhs = "ndf.Satisfaction=Unhappy"))
plot(ruleset)
```



The above visualization is done for the entire dataset considering the customer satisfaction greater than equal to 4 as the RHS. Our main aim here is to find the ruleset which are the significant factors predicting high customer satisfaction. The significant attributes which we found are Type of travel-Business, Flight Cancelled-No and Arrival delay is greater than 5 mins-No. These attributes give the highest lift value of 1.49 which will make the customers happy of each airline.

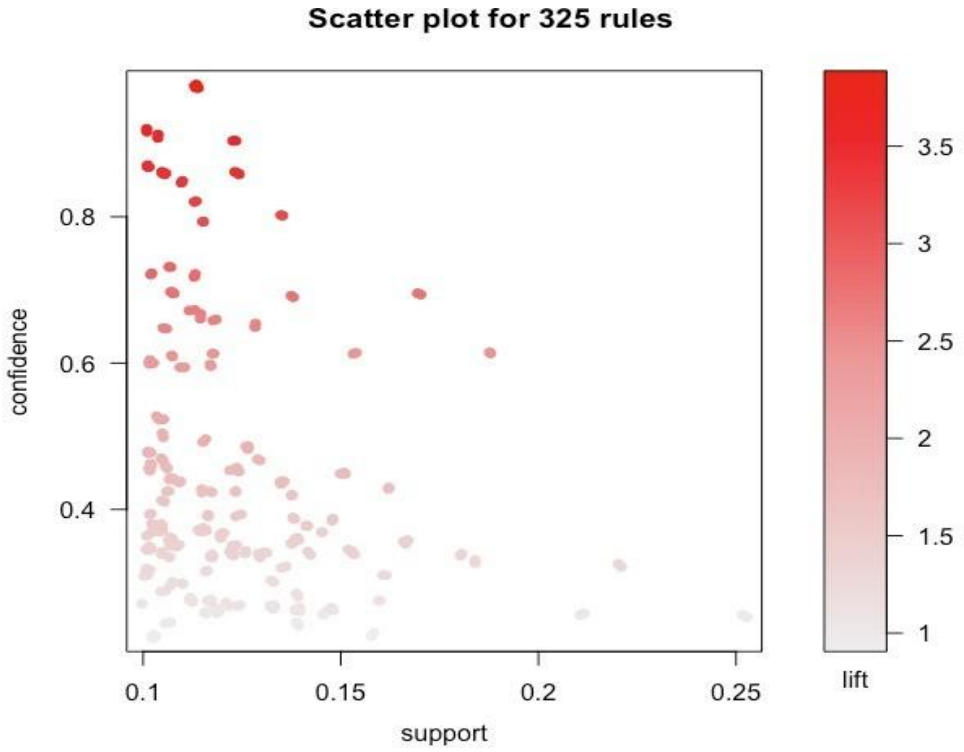


We were keen in finding the attributes which were the major cause of unhappiness in customers of all airlines. Hence we considered the entire data where the customer satisfaction corresponds to less than 3 for all airlines. Our main aim here is to find the ruleset which are the significant factors predicting low customer satisfaction. The significant attributes which we found for making customers unhappy are type of travel-Personal. This attribute give the highest lift value of 2.53 which will make the customers unhappy of each airline.



As we found GoingNorth Airlines with the highest proportion of unhappy customers and WestAirlines with the highest proportion of happy customers, we decided to consider these two airlines out of all the airlines and get the significant attributes which determine the cause of happiness and unhappiness of customers.

The above visualization is done for the customer survey of West Airlines. Hence we considered the data of West Airways where the customer satisfaction corresponds to greater than equal to 4 for all airlines. Our main aim here is to find the ruleset which are the significant factors predicting high customer satisfaction. The significant attributes which we found for making customers happy are type of travel-Business and Flight Cancelled-No. These attributes give the highest lift value of 1.35 which will make the customers happy of West Airlines.



The above visualization is done for the customer survey of GoingNorth Airlines. Hence we considered the data of GoingNorth Airways where the customer satisfaction corresponds to less than 3 for GoingNorth airlines. Our main aim here is to find the ruleset which are the significant factors predicting low customer satisfaction. The significant attributes which we found for making customers unhappy are Airline Status-Blue,Type of travel-Personal and Flight Cancelled-No. These attributes give the highest lift value of 2.75 which will make the customers unhappy of Goingnorth Airlines.

Decision Tree

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes.

Before using decision tree we have made two functions for bucketing that are CreateBucketSat() and CreateBuckets(). CreateBucketSat() function is used to assign Happy, Average and Low values to customer satisfaction column depending upon their values while CreateBuckets() function is used to assign high, average and low values to other numeric columns based on quantiles. Then, we apply bucketing functions to satisfaction and other independent attributes. We have used two third of data as train data while one third of data as test data. We built the decision tree based on the three attributes which we got as significant from stepwise linear regression model(type of travel, airline status and arrival_delay_greater_5_mins). The decision tree was built using the rpart.plot function.

```
#=======================================================================================================

# Part I
#
# Preperation

# Loading packages
library('rpart')
library('rpart.plot')
library('ggplot2')
library('caret')
library('e1071')

# Loading data
setwd('/Users/a111/Documents/data/')
df <- read.csv("Satisfaction Survey.csv")


# Data cleaning

# Filling in some missing values
df$Departure.Delay.in.Minutes[which(is.na(df$Departure.Delay.in.Minutes) & (df$Flight.cancelled == 'Yes'))] <- 0
df$Arrival.Delay.in.Minutes[which(is.na(df$Arrival.Delay.in.Minutes) & (df$Flight.cancelled == 'Yes'))] <- 0
df$Flight.time.in.minutes[which(is.na(df$Flight.time.in.minutes) & (df$Flight.cancelled == 'Yes'))] <- 0

# Numericalize satisfaction
df$Satisfaction <- as.numeric(as.character(df$Satisfaction))

# Dumping all the rows with missing values
ndf <- na.omit(df)

#=======================================================================================================
```

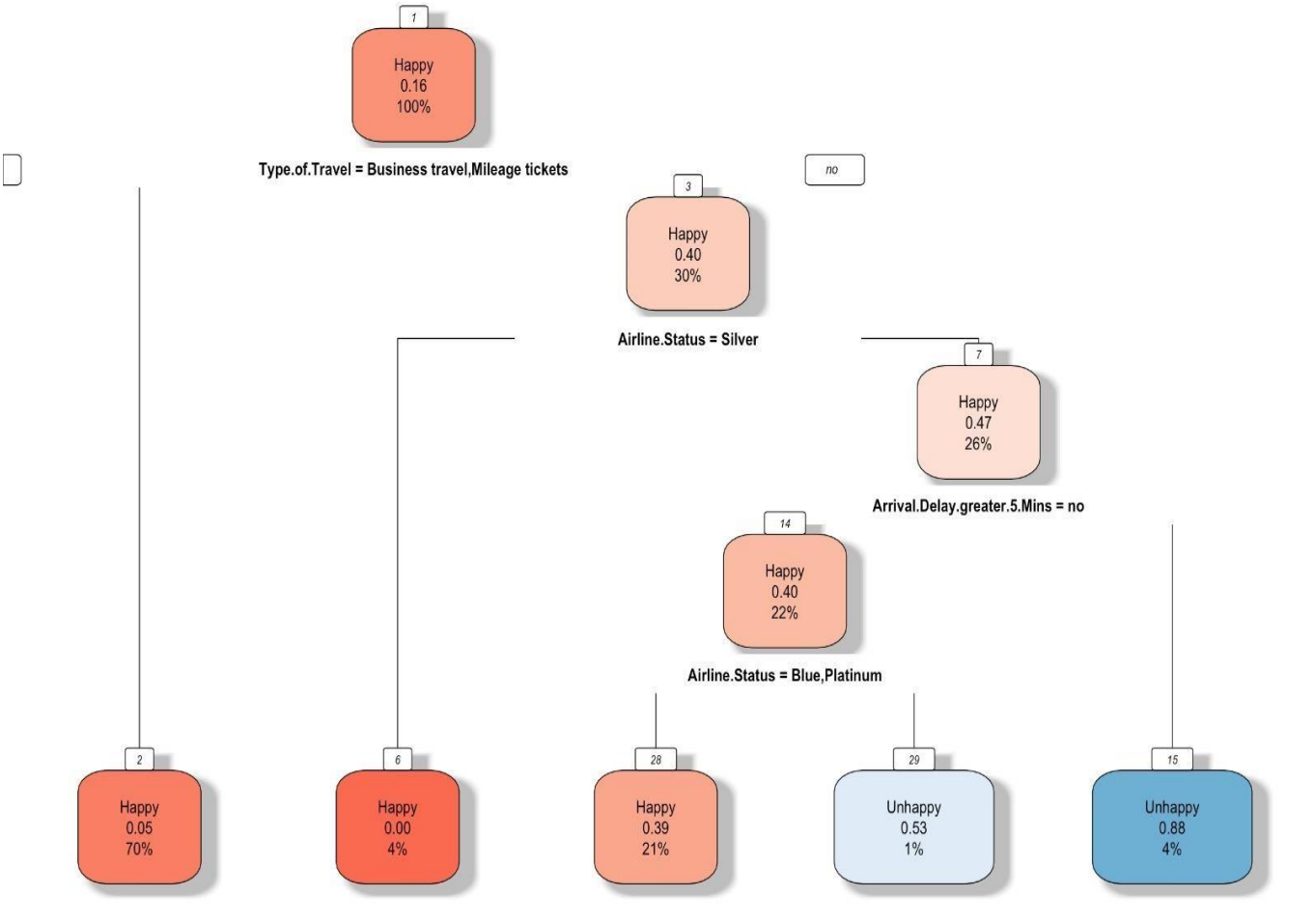
```
# Part II
#
# Data manipulating

# Bucketing

# Satisfaction bucketing
createBucketSat <- function(vec){
  vBuckets <- replicate(length(vec), "Unhappy")
  vBuckets[vec >= 3] <- "Happy"
  vBuckets[vec < 3] <- "Unhappy"
  return(vBuckets)
}

# Bucketing numeric numbers
createBuckets <- function(vec){
  q <- quantile(vec, c(0.4, 0.6), na.rm = "TRUE")
  vBuckets <- replicate(length(vec), "Average")
  vBuckets[vec <= q[1]] <- "Low"
  vBuckets[vec > q[2]] <- "High"
  return(vBuckets)
}

ndf$Satisfaction <- createBucketSat(ndf$Satisfaction)

Treedf <-  data.frame(
  ndf$Satisfaction,
  ndf$Airline.Status,
  ndf$Age,
  ndf$Gender,
  ndf$No.of.Flights.p.a.,
  ndf$X..of.Flight.with.other.Airlines,
  ndf$Type.of.Travel,
  ndf$Shopping.Amount.at.Airport,
  ndf$Eating.and.Drinking.at.Airport,
  ndf$Class,
  ndf$Airline.Name
)

colnames(Treedf) <- c('Sat', 'Status', 'Age', 'Gender', 'NOF','XOF', 'Type', 'Shop', 'Eat', 'Class', 'Airline')

# Generate random numbers of the size of the whole dataset which later be used as indexes
randIndex <- sample(1:dim(Treedf)[1])
head(randIndex,10)
```

```
# Cutting point that seperates training data and testing data
cutPoint2_3 <- floor(2 * dim(Treedf)[1]/3)

# Generating training dataset
trainData <- Treedf[randIndex[1 : cutPoint2_3], ]

# Generating testing dataset
testData <- Treedf[randIndex[(cutPoint2_3+1) : dim(Treedf)[1]], ]

tree <- rpart(Sat ~ ., data=trainData, cp=.0009)
# rpart.plot(tree, box.palette="RdBu", shadow.col="gray", nn=TRUE)

treepred <- predict(tree, newdata = testData)
pred <- replicate(length(treepred)/2, 'Unhappy')
for(i in 1:(length(treepred)/2)){
  if(treepred[i, 1] > treepred[i, 2]){
    pred[i] <- 'Happy'
  }else if(treepred[i, 1] <= treepred[i, 2]){
    pred[i] <- 'Unhappy'
  }
}
pred <- as.factor(pred)
confusionMatrix(pred, testData$Sat)



# Generate random numbers of the size of the whole dataset which later be used as indexes
randIndex <- sample(1:dim(ndf)[1])

# Cutting point that seperates training data and testing data
cutPoint2_3 <- floor(2 * dim(ndf)[1]/3)

# Generating training dataset
ndftrainData <- ndf[randIndex[1 : cutPoint2_3], ]

# Generating testing dataset
ndftestData <- ndf[randIndex[(cutPoint2_3+1) : dim(ndf)[1]], ]

"
```

```
#=============================================================================================

# Business solution tree 1

tree1 <- rpart(Satisfaction ~ Airline.Status + Type.of.Travel + Arrival.Delay.greater.5.Mins, data = ndftrainData, cp = .001)
rpart.plot(tree1, box.palette="RdBu", shadow.col="gray", nn=TRUE)

treepred1 <- predict(tree1, newdata = ndftestData)
pred1 <- replicate(length(treepred1)/2, 'Unhappy')
for(i in 1:(length(treepred1)/2)){
  if(treepred1[i, 1] > treepred1[i, 2]){
    pred1[i] <- 'Happy'
  }else if(treepred1[i, 1] <= treepred1[i, 2]){
    pred1[i] <- 'Unhappy'
  }
}

pred1 <- as.factor(pred1)
confusionMatrix(pred1, as.factor(ndftestData$Satisfaction))

#=============================================================================================

# Business solution tree 2

tree2 <- rpart(Satisfaction ~ Class + Type.of.Travel + Flight.cancelled, data = ndftrainData, cp = .0006)
rpart.plot(tree2, box.palette="RdBu", shadow.col="gray", nn=TRUE)

treepred2 <- predict(tree2, newdata = ndftestData)
pred2 <- replicate(length(treepred2)/2, 'Unhappy')
for(i in 1:(length(treepred2)/2)){
  if(treepred2[i, 1] > treepred2[i, 2]){
    pred2[i] <- 'Happy'
  }else if(treepred2[i, 1] <= treepred2[i, 2]){
    pred2[i] <- 'Unhappy'
  }
}

pred2 <- as.factor(pred2)
confusionMatrix(pred2, as.factor(ndftestData$Satisfaction))
```

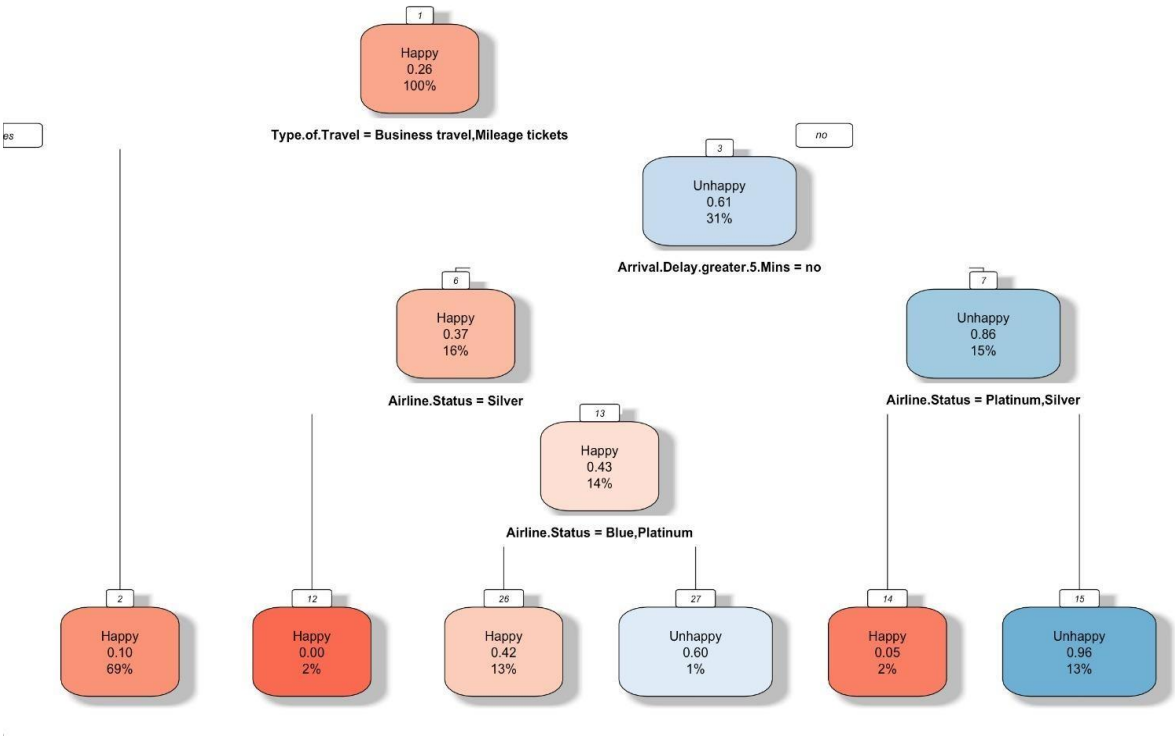
```
# happy airline

West1 <- subset(ndf, Airline.Name == "West Airways Inc. ")
randIndex <- sample(1:dim(West1)[1])
cutPoint2_3 <- floor(2 * dim(West1)[1]/3)
ndftrainData <- West1[randIndex[1 : cutPoint2_3], ]
ndftestData <- West1[randIndex[(cutPoint2_3+1) : dim(West1)[1]], ]

tree3 <- rpart(Satisfaction ~ Type.of.Travel + Flight.cancelled, data = ndftrainData, cp = .001)

tree3 <- rpart(Satisfaction ~ Airline.Status + Type.of.Travel + Arrival.Delay.greater.5.Mins, data = ndftrainData, cp = .001)

rpart.plot(tree3, box.palette="RdBu", shadow.col="gray", nn=TRUE)



treepred3 <- predict(tree3, newdata = ndftestData)
pred3 <- replicate(length(treepred3)/2, 'Unhappy')
for(i in 1:(length(treepred3)/2)){
  if(treepred3[i, 1] > treepred3[i, 2]){
    pred3[i] <- 'Happy'
  }else if(treepred3[i, 1] <= treepred3[i, 2]){
    pred3[i] <- 'Unhappy'
  }
}

pred3 <- as.factor(pred3)
confusionMatrix(pred3, as.factor(ndftestData$Satisfaction))

# Unhappy airline

GoingNorth1 <- subset(ndf, Airline.Name == "GoingNorth Airlines Inc. ")
randIndex <- sample(1:dim(GoingNorth1)[1])
cutPoint2_3 <- floor(2 * dim(GoingNorth1)[1]/3)
ndftrainData <- GoingNorth1[randIndex[1 : cutPoint2_3], ]
ndftestData <- GoingNorth1[randIndex[(cutPoint2_3+1) : dim(GoingNorth1)[1]], ]

tree4 <- rpart(Satisfaction ~ Airline.Status + Type.of.Travel + Arrival.Delay.greater.5.Mins, data = ndftrainData, cp = .0006)

tree4 <- rpart(Satisfaction ~ Airline.Status + Type.of.Travel, data = ndftrainData, cp = .0006)

rpart.plot(tree4, box.palette="RdBu", shadow.col="gray", nn=TRUE)
```

```
treepred4 <- predict(tree4, newdata = ndftestData)
pred4 <- replicate(length(treepred4)/2, 'Unhappy')
for(i in 1:(length(treepred4)/2)){
  if(treepred4[i, 1] > treepred4[i, 2]){
    pred4[i] <- 'Happy'
  }else if(treepred3[i, 1] <= treepred3[i, 2]){
    pred4[i] <- 'Unhappy'
  }
}

pred4 <- as.factor(pred4)
confusionMatrix(pred4, as.factor(ndftestData$Satisfaction))
```



We have used this model of decision tree to predict the customer satisfaction by using the significant attributes of Airline Status, Travel type and Arrival delay greater than 5 mins which we got from the stepwise linear regression. We have made the decision trees for GoingNorth Airlines and West Airways.

The above decision tree showcases what combination of attributes makes the customer happy for West Airways. If type of travel is Business or Mileage tickets, then customers would be highly satisfied. If type of travel is Personal and Airline Status is silver, customers will be happy. If type of travel-Personal, airline status is blue or platinum and arrival is not greater than five mins than customer will be happy.



The above decision tree showcases what combination of attributes makes the customer unhappy for GoingNorth Airlines. If type of travel is not Business or Mileage tickets i.e. Personal, there is a delay of greater than 5 mins and the airline status is either blue or gold then the customer would be unhappy. Similarly a

Personal Traveller experiencing a delay of greater than 5 mins is expected to be unhappy for a majority of times.

Support Vector Machine

Support Vector Machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.

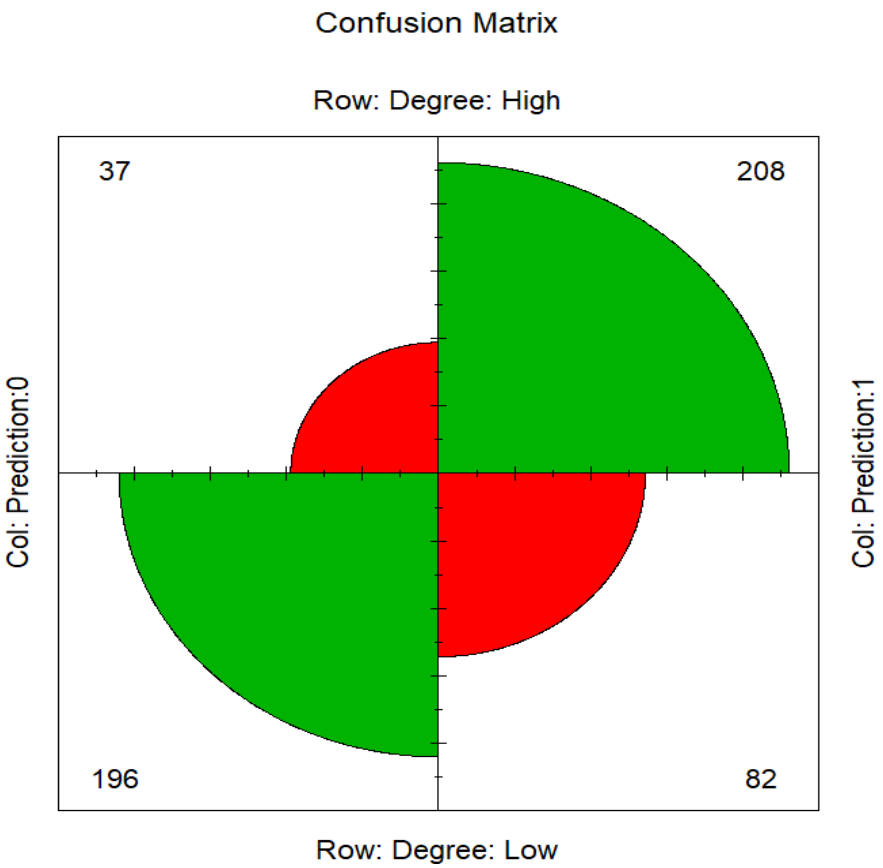
Before using SVM we have assumed customers who have given rating greater than 4 to be happy while customers who have given rating less than four to be unhappy. We have used two third of data as train data while one third of data as test data. We built the SVM model based on the three attributes which we got as significant from stepwise linear regression model(type of travel, airline status and arrival_delay_greater_5_mins). While using the ksvm function we made the value of cost as 5 and considered the cross validation value to be 3.

After the SVM model was built we predicted our results using the test data and got an accuracy of 77% for West Airways while 74% for GoingNorth Airlines.

```
# loading packages
library(ggplot2)
library(MASS)
library(kernlab)
library(RJSONIO)

#read data
# Set the working directory where your dataset is saved
setwd("D:/Syracuse University/Fall'18/IST 687/Project")
df <- read.csv("Satisfaction Survey.csv")
View(df)
#clean missing value
df$Departure.Delay.in.Minutes[which(is.na(df$Departure.Delay.in.Minutes) & (df$Flight.cancelled == 'Yes'))] <- 0
df$Arrival.Delay.in.Minutes[which(is.na(df$Arrival.Delay.in.Minutes) & (df$Flight.cancelled == 'Yes'))] <- 0
df$Flight.time.in.minutes[which(is.na(df$Flight.time.in.minutes) & (df$Flight.cancelled == 'Yes'))] <- 0

#change the data type of 'Satisfaction' to numeric
df$Satisfaction <- as.numeric(as.character(df$Satisfaction))

#omit the missing value
ndf <- na.omit(df)
View(df)
#get the names of airlines
airline.name <- c(levels(ndf$Airline.Name))
airline.name
#Insert a new column describing the degree of satisfaction
ndf$degree <- NA
ndf$Satisfaction <- as.numeric(as.character(ndf$Satisfaction))
ndf$degree[which(ndf$Satisfaction >=4)]<- "High"
ndf$degree[which(ndf$Satisfaction <4)] <- "Low"

ndf <- na.omit(ndf)
View(ndf)

GoingNorth <- subset(ndf, Airline.Name == "GoingNorth Airlines Inc. ") #"GoingNorth Airlines Inc. "
West <- subset(ndf, Airline.Name == "West Airways Inc. ") #"West Airways Inc. "

randIndex <- sample(1:dim(GoingNorth)[1])
summary(randIndex)
```

```
# Creating a breakpoint of 2/3rd and 1/3rd part for GoingNorth
cutPoint2_3_gn <- floor(2 * dim(GoingNorth)[1]/3)
cutPoint2_3_gn
# Creating traindata with 2/3rd of GoingNorth data
trainData_gn <- GoingNorth[randIndex[1:cutPoint2_3_gn],]
# Creating testdata with 1/3rd of GoingNorth data
testData_gn <- GoingNorth[randIndex[(cutPoint2_3_gn+1):dim(GoingNorth)[1]],]
testData_gn
#check dimensions of the data frame GoingNorth, trainData_gn and testData_gn
dim(GoingNorth)
dim(trainData_gn)
View(trainData_gn)
dim(testData_gn)

#GoingNorth Airlines

svmOutput_gn <- ksvm(degree ~ Airline.Status+Type.of.Travel+Arrival.Delay.greater.5.Mins, data = trainData_gn,kernel ="rbfdot",kpar="automatic",C=5,cross=3,prob.model=TRUE)
print(svmOutput_gn)
svmPred_gn <- predict(svmOutput_gn, testData_gn, type ="votes")
svmPred_gn
str(svmPred_gn)
head(svmPred_gn)
# Creating a composite table based on CompTable_gn and svmPred_gn
compTable_gn<-data.frame(testData_gn$degree,svmPred_gn[1,])
View(compTable_gn)
# Creating a confusion matrix
conMatrix_gn<-table(compTable_gn)
conMatrix_gn

ctable <- matrix(c(68, 185, 202, 68), nrow = 2, byrow = TRUE)
ctable

colnames(ctable) <- c("Prediction:0","Prediction:1")
row.names(ctable) <- c("Degree: High", "Degree: Low")

fourfoldplot(ctable, color = c("#ff0000", "#00b300"),
             conf.level = 0, margin = 1, main = "Confusion Matrix")


errorSum_gn<-conMatrix_gn[1,1]+conMatrix_gn[2,2]
errorSum_gn
# Creating percentage of error rate
errorRate_gn<-errorSum_gn/sum(conMatrix_gn)*100
errorRate_gn
```

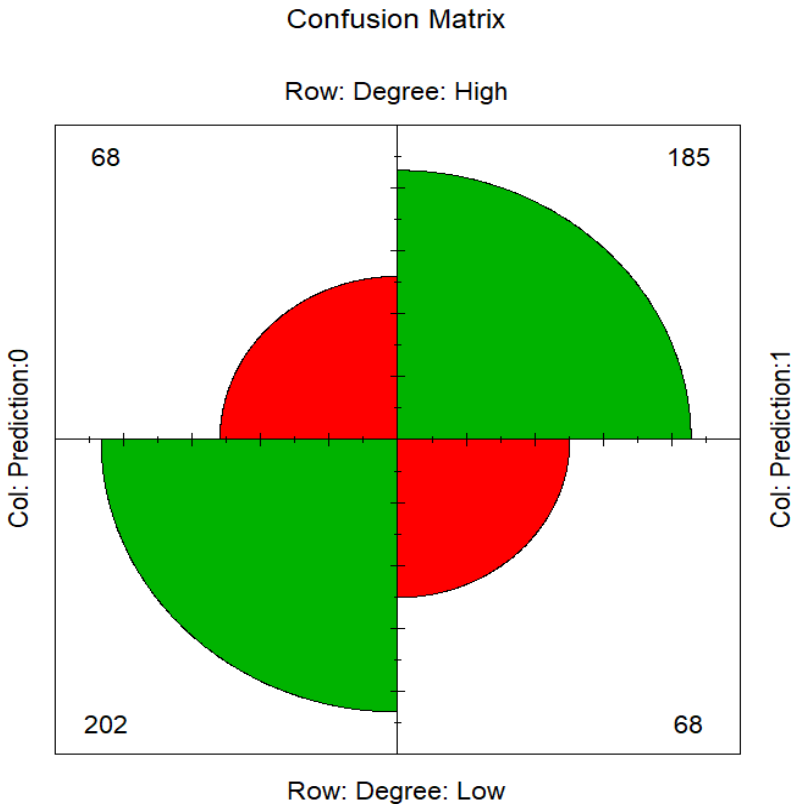
```
#West Airways:

svmOutput_gn1 <- ksvm(degree ~ Airline.Status+Gender+Type.of.Travel+Arrival.Delay.greater.5.Mins, data = trainData_gn,kernel ="rbfdot",kpar="automatic",C=5,cross=3,prob.model=TRUE)

print(svmOutput_gn1)
svmPred_gn1 <- predict(svmOutput_gn1, testData_gn, type ="votes")
svmPred_gn1
str(svmPred_gn1)
head(svmPred_gn1)
# Creating a composite table based on CompTable_gn and svmPred_gn
compTable_gn1<-data.frame(testData_gn$degree,svmPred_gn1[1,])
str(compTable_gn1)
# Creating a confusion matrix
conMatrix_gn1<-table(compTable_gn1)
conMatrix_gn1

ctable <- matrix(c(37, 208, 196, 82), nrow = 2, byrow = TRUE)
ctable

colnames(ctable) <- c("Prediction:0","Prediction:1")
row.names(ctable) <- c("Degree: High", "Degree: Low")

fourfoldplot(ctable, color = c("#ff0000", "#00b300"),
             conf.level = 0, margin = 1, main = "Confusion Matrix")

# Creating a dataframe containing sum of errors
errorSum_gn1<-conMatrix_gn1[1,1]+conMatrix_gn1[2,2]
errorSum_gn1
# Creating percentage of error rate
errorRate_gn1<-errorSum_gn1/sum(conMatrix_gn1)*100
errorRate_gn1
```



Here we are using KSVM model for predicting customer satisfaction for both the airlines GoingNorth Airlines and West Airways.

We have visualized the confusion matrix using the fourfold plot.

The fourfold plot of West Airways gives us 77% accuracy rate i.e (208+196/37+82+208+196) for the three attributes of Airlines Status, Type of travel and arrival delay greater than 5 minutes.



The fourfold plot of Going North Airlines gives us 74% accuracy rate i.e (185+202/68+68+202+185) for the three attributes of Airlines Status, Type of travel and arrival delay greater than 5 minutes.

Things that did not work

In stepAIC function, we did not use the backward stepwise linear regression model since it was taking a large amount of execution time. Therefore, we decided to not go ahead as this method was not feasible.

Results

Stepwise Linear model gave us 3 significant attributes i.e. Type of Travel, Airline Status and Arrival_Greater_Than_5_Mins and through comparisons we found out that Customers are experiencing more delays in GoingNorth Airlines as compared to West Airlines.

The Decision tree models gave us an accuracy of around 87 percent which shows that the significant attributes obtained from stepwise Linear model are indeed significant.

Finally, the SVM model gave us an accuracy of around 77 percent which makes it even more clear that GoingNorth Airlines should work on the above mentioned significant attributes.

Actionable Insights

From our analysis, we found out that GoingNorth Airlines had their majority of flights delayed which made the customers rate the airlines less. From this we can say that GoingNorth Airlines should work more on landing their flights on time.

In addition to this, the type of travel of majority of customers for GoingNorth Airlines was Personal from which we can deduce that personal travelers are the ones who aren't happy with GoingNorth Airlines. In order to improve this grey area, GoingNorth Airlines should try to find out ways in which they can make the Personal Travelers happy i.e. by providing discounted rates, by upgrading their status or by rewarding them with miles.

Finally, the Airline status of majority of unhappy customers of GoingNorth Airlines was Blue which makes us think that there is something wrong with the plan that is provided by the GoingNorth Airlines for the customers with Airline status as Blue. Thus, GoingNorth Airlines should try to do something different for Blue customers or try to incorporate some offers of other tiers like gold, silver, platinum so that the customers having status blue stay happy.

Validation

From our analysis we found out three significant attributes i.e. type of travel, Airline Status and Arrival_greater_than_5_mins. We can say with a high rate of assurance that these three attributes are indeed important and must be looked at by GoingNorth Airlines because we got an accuracy rate of around 87 percent using decision tree model and an accuracy rate of 77 percent using the Support Vector Machine Model. These accuracy rates are high enough to validate our analysis.

Trello Board Update