**Project Report**

**On**

**Airline Satisfaction**

**-Prepared By:**

**Chaitali Nautiyal**

**James Eakins**

**Rupesh Ghosh**

**Sanket Vichare**

TABLE OF CONTENTS

1. PROBLEM STATMENT -------------------------------------------------- 3
2. DATA CLEANING--------------------------------------------------------- 4
3. DATA ANALYSIS---------------------------------------------------------- 5
4. DATA MODELING ------------------------------------------------------ 18
5. DATA VISUALIZATIONS ----------------------------------------------- 23
6. DATA VALIDATION ---------------------------------------------------- 37
7. RECOMMENDATION -------------------------------------------------- 58

Problem Statement

In recent years, the airline industry changed dramatically and became an extremely competitive industry. Because operational excellence became a best practice in the airline industry, airline companies seek to understand customer needs in order to drive and sustain competitive advantage over others. By analyzing customer survey data from a variety of airlines, this project aims to not only provide Southeast Airlines the knowledge and tools needed to understand drivers of customer satisfaction, but also information on how the airline can stand at the forefront of customer service.

Objectives

The goals of this project are to identify the key factors that contribute towards overall customer satisfaction in the airline industry, build a model specifically for Southeast Airlines to predict customer satisfaction, and provide Southeast Airlines with actionable insights for future business operations.

Business Questions

* Which factors impact customer satisfaction the most?
* Does travel source and destination affect the customer satisfaction?
* Which airline has the best and the worst average customer satisfaction rating?
* What impact has time on the customer satisfaction level and what role the flight cancellation plays in context?
* What are the cities that have the most delays and what you can recommend to change it?
* Does the age of the customer affect the customer rating?

Data cleaning

The original dataset was received in CSV format and contained a total of 194835 records. After loading the data file into R, we cleaned the data. Furthermore, we have removed spaces as well as dots from the column names in order to ease the process of modeling building later on. Moreover, we deleted records of flights that were canceled for purposes of our delay time analysis.

**Code:**

#Reading the Data

dataset <- read.csv("spring19survey.csv")

#Cleaning Data

cleanedDataset <- dataset

View(cleanedDataset)

which(is.na(dataset$Departure.Delay.in.Minutes))

cleanedDataset$Departure.Delay.in.Minutes <- ifelse(is.na(cleanedDataset$Departure.Delay.in.Minutes) & cleanedDataset$Flight.cancelled == "Yes", 9999, cleanedDataset$Departure.Delay.in.Minutes)

cleanedDataset$Arrival.Delay.in.Minutes <- ifelse(is.na(cleanedDataset$Arrival.Delay.in.Minutes) & cleanedDataset$Flight.cancelled == "Yes", 9999, cleanedDataset$Arrival.Delay.in.Minutes)

cleanedDataset$Flight.time.in.minutes <- ifelse(is.na(cleanedDataset$Flight.time.in.minutes) & cleanedDataset$Flight.cancelled == "Yes", 9999, cleanedDataset$Flight.time.in.minutes)

cleanedDataset <- na.omit(cleanedDataset)

str(dataset)

summary(cleanedDataset)

str(cleanedDataset)

which(is.na(cleanedDataset$Departure.Delay.in.Minutes)) #Checking Na's

cleanedDataset$Satisfaction <- as.numeric(as.character(cleanedDataset$Satisfaction)) #Changing to numeric

View(cleanedDataset)

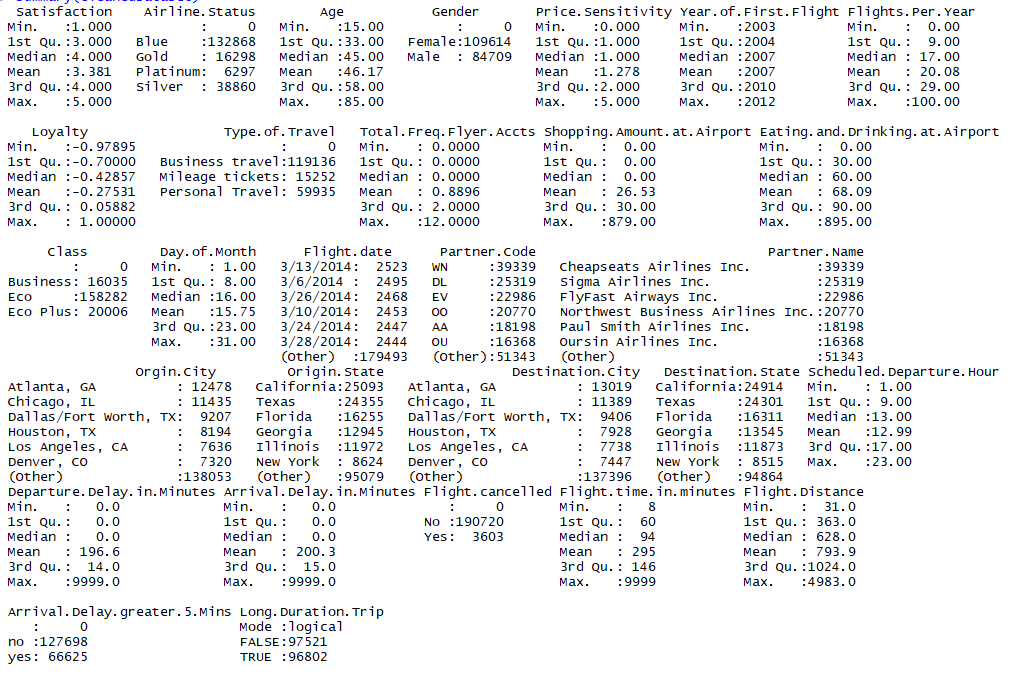
Data Analysis

**1. Descriptive Statistics:**

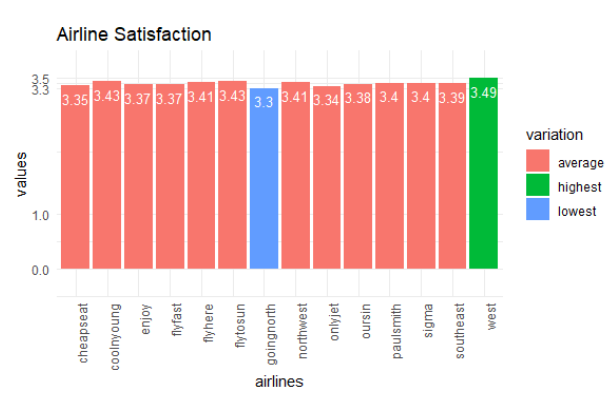
**Summary of the CleanedDataset**

In this you can see, the entire summary which includes the mean, median, mode, minimum value, and maximum values and quantiles.

You can find all the essential columns which are in the dataset and their vital insights



The below bar chart depicts the overall average customer satisfaction for each airline and helped us made the decision to select the airlines with the best and the worst average customer satisfaction rating,



1. **Customer Information**

**Gender:**

The chart below shows the comparison of gender satisfaction between West Airways and GoingNorth and found out that

* **For Males:**

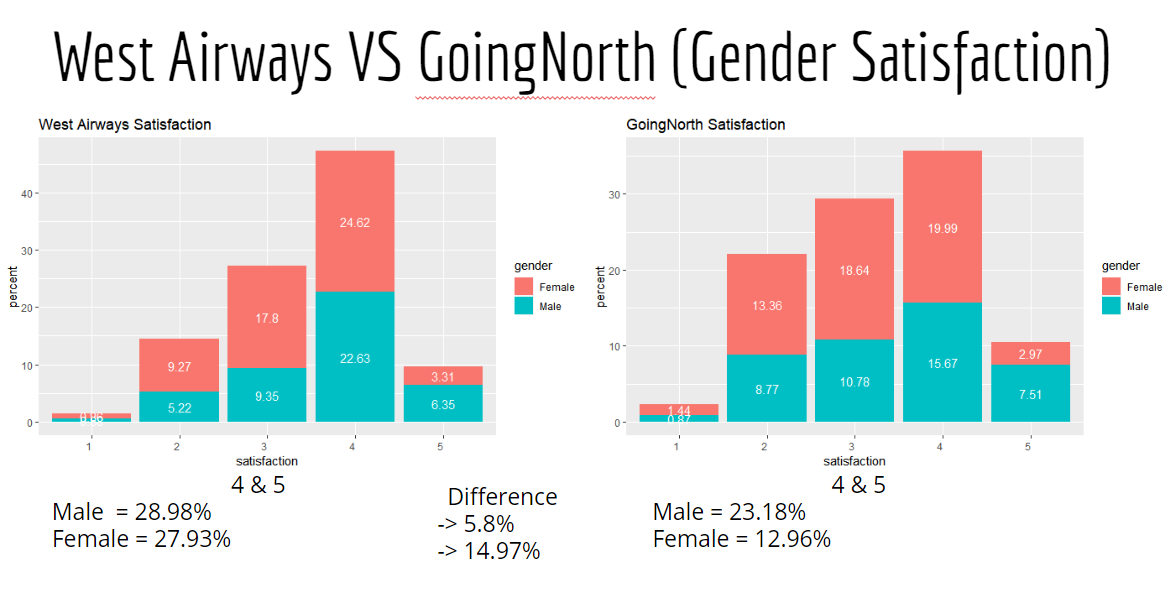
West Airways the total percentile of ratings given as 4 and 5 adds up to be 28.98% and for GoingNorth it is 23.18% giving us a difference of 5.8%.

* **For Females:**

West Airways the total percentile of ratings given as 4 and 5 adds up to be 27.93% and for GoingNorth it is12.96% giving us a difference of 14.97%.

This provides us a brief idea that both males and females are not happy with GoingNorth Airlines due to the certain reason which we will look further. Therefore, affecting the average satisfaction level.

So, in conclusion, targeting the female audience is really vital to improve airline satisfaction rating.



**AirlineStatus**

The next factor that we took into consideration was AirlineStatus.

After performing statistical analysis from the customers perspective we found out that

* **For Blue**

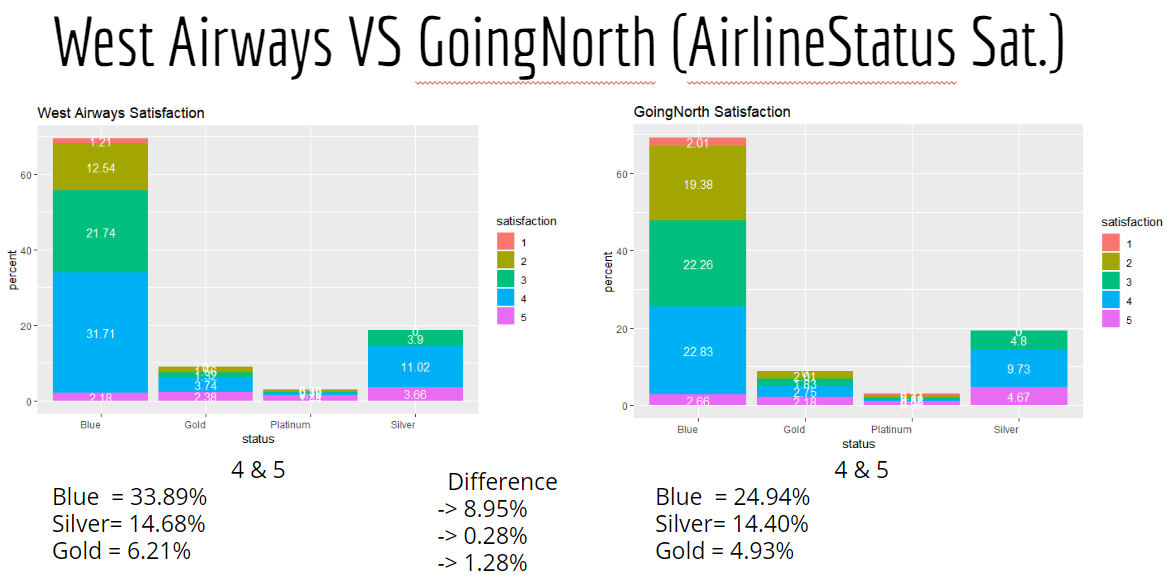
The customers who gave higher satisfaction rating to WestAirways in Blue AirlineStatus is 33.89% while for GoingNorth is 24.94% and the difference between both comes out to be 8.95%.

* **For Silver**

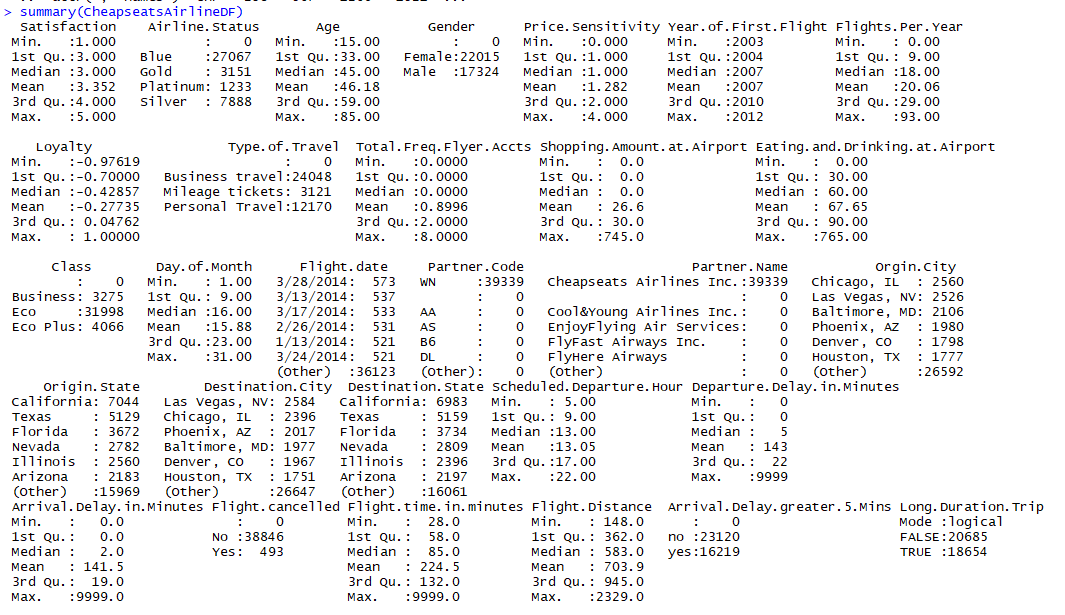
The customers who gave higher satisfaction rating to WestAirways in Silver AirlineStatus is 14.68% while for GoingNorth is 14.40% and the difference between both comes out to be 0.28%.

* **For Gold**

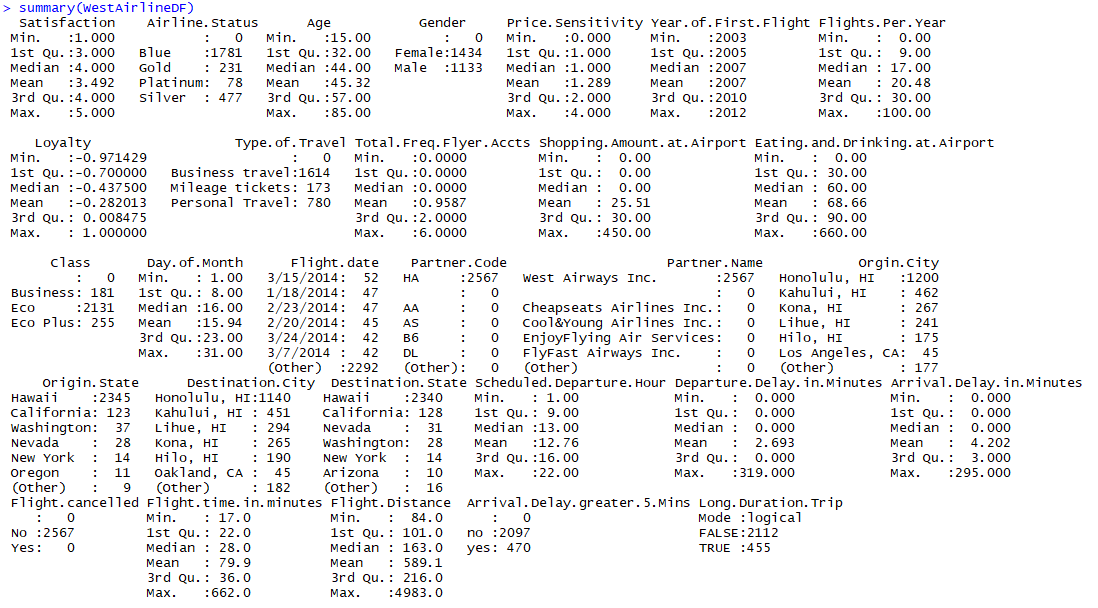
The customers who gave higher satisfaction rating to WestAirways in Gold AirlineStatus is 6.21% while for GoingNorth is 4.93% and the difference between both comes out to be 1.28%.



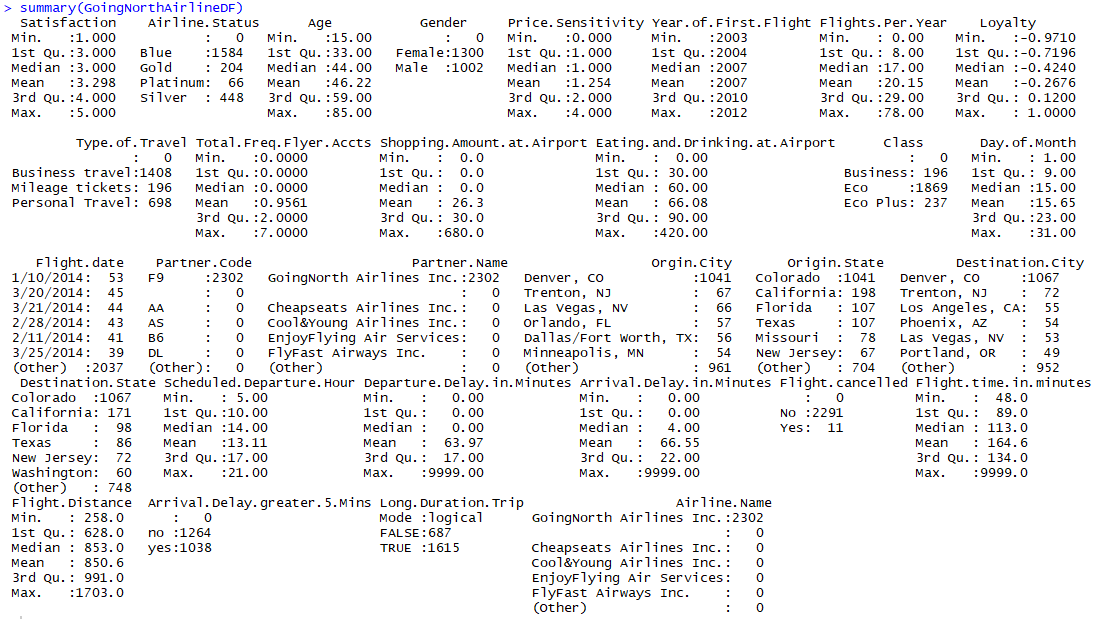
**Summary of Cheapseats Airlines**



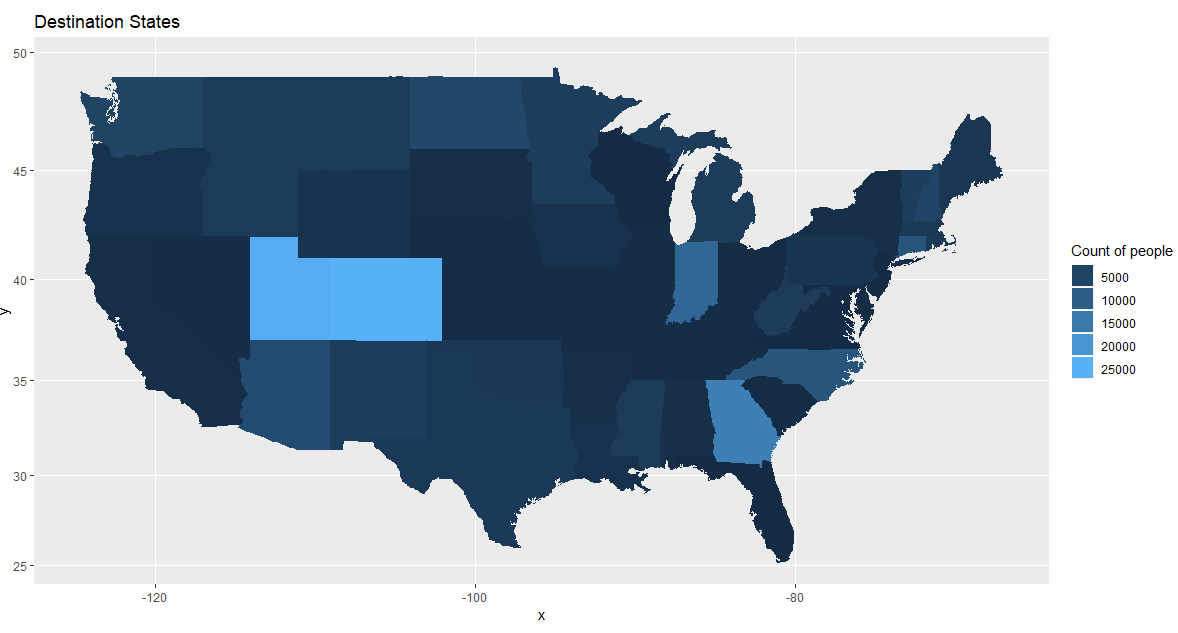
**Summary of WestAirlines**



**Summary of GoingNorth**

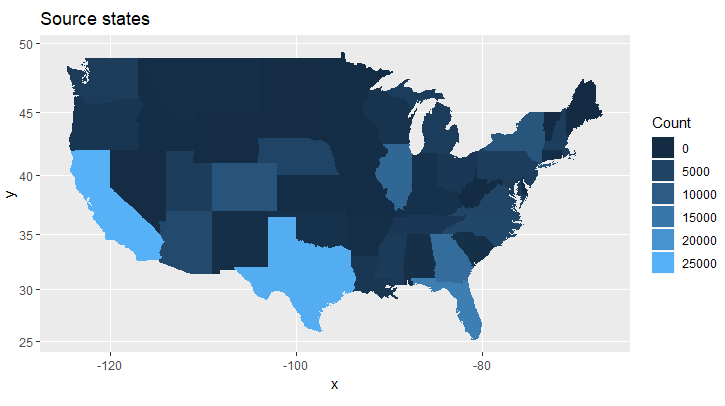


**Destination States**



We tried focusing on the destination states initially to see if we could make some important decisions in accordance with the factor. So we plotted a map, considering the count of the people traveling to the state but unfortunately could not move forward with that idea.

**Source States**

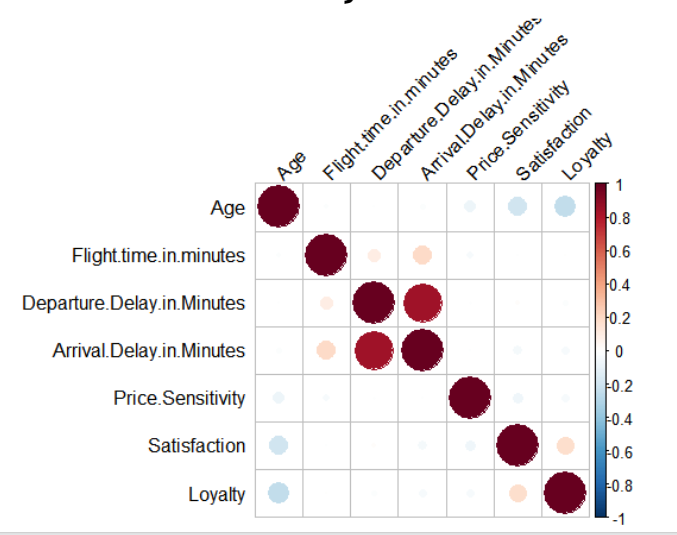


We tried focusing on the source states initially to see if we could make some important decisions in accordance with the factor. So we plotted a map, considering the count of the people from the state but unfortunately could not move forward with that idea.

**Correlation matrix**

A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between the two variables. A correlation matrix is used as a way to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses.

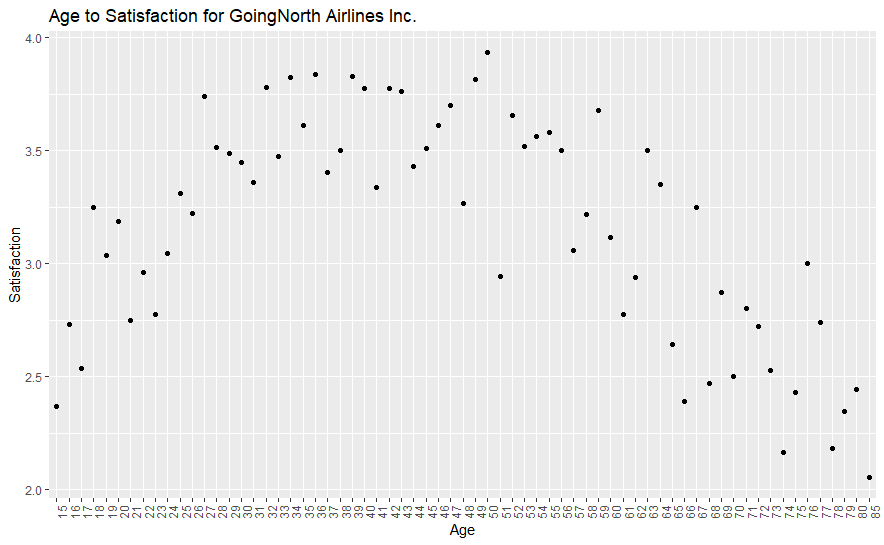
Here we can see the correlation matrix between the columns of the dataset taken from the survey which affects the satisfaction. The sublimation of colors on the extreme right in a bar format gives us the basic idea of the statistical values ranging from -1 to 1 along with the size of the dots in the matrix and the color.



**Analysis of Impact of Age attribute on Customer Satisfaction:**

The next factor that we would consider is the impact of age on customer satisfaction. So here we plot a graph of age against customer satisfaction for GoingNorth Airlines Inc., West Airways, and Cheapseats Airlines Inc.

* **GoingNorth Airlines Inc.**



We notice that lower satisfaction index is given by the people belonging to the extreme end, i.e. higher age group. To quantify the results, we categorize the customers into three groups based on their age:

High: Above 60

Average: 40-60

Low: Below 40

**Average High Low**

1 21 18 14

2 139 235 136

3 222 153 302

4 390 89 343

5 144 18 78

Displaying the above result in percentage for better understanding

**Average High Low**

1 0.91% 0.78% 0.61%

2 6.04% 10.21% 5.91%

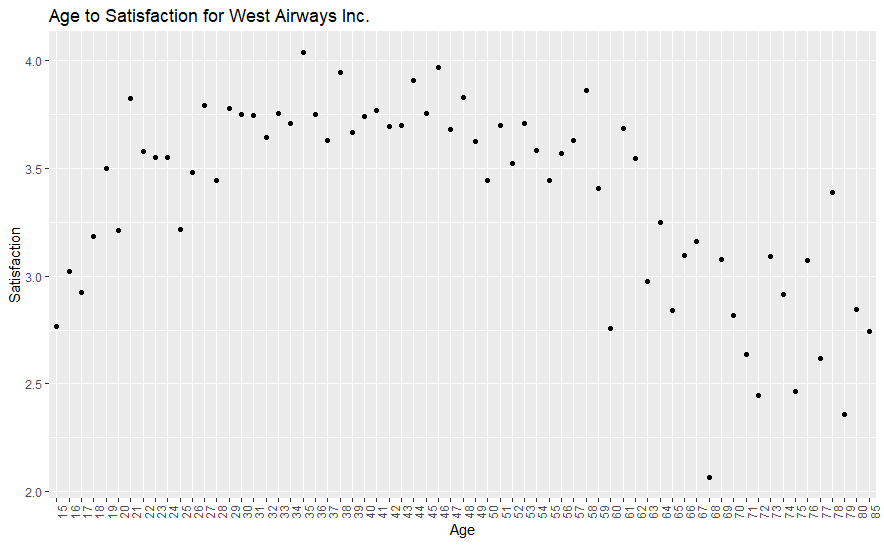
3 9.64% 6.65% 13.12%

4 16.94% 3.87% 14.9%

5 6.26% 0.78% 3.39%

Here, we notice that higher age group is giving lower ratings (1 & 2).

* **West Airways**



We notice that lower satisfaction index is given by the people belonging to the extreme end, i.e. higher age group. To quantify the results, we categorize the customers into three groups based on their age:

High: Above 60

Average: 40-60

Low: Below 40

**Average High Low**

1 17 8 12

2 100 148 127

3 234 235 230

4 495 113 605

5 140 18 90

Displaying the above result in percentage for better understanding

**Average High Low**

1 0.66% 0.31% 0.47%

2 3.89% 5.75% 4.94%

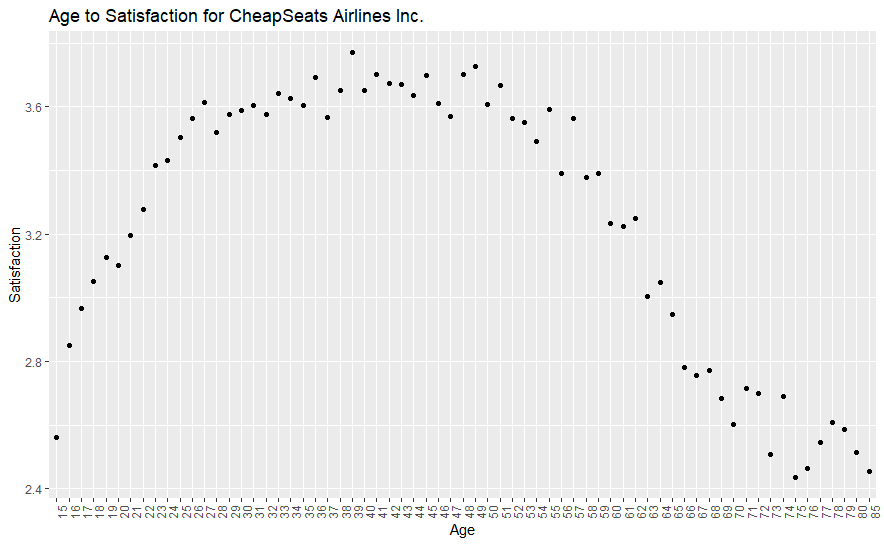
3 9.10% 9.14% 8.94%

4 19.25% 4.39% 23.52%

5 5.44% 0.70% 3.50%

Again, we notice that higher age group is giving lower ratings (1 & 2).

* **CheapSeats Airlines Inc.**



We notice that lower satisfaction index is given by the people belonging to the extreme end, i.e. higher age group. To quantify the results, we categorize the customers into three groups based on their age:

High: Above 60

Average: 40-60

Low: Below 40

**Average High Low**

1 345 362 245

2 2021 3478 2080

3 3473 3206 4546

4 7306 1574 7099

5 2250 207 1249

We convert the above results into percentage for better understanding.

**Average High Low**

1 0.87% 0.92% 0.62%

2 5.12% 8.82% 5.27%

3 8.81% 8.13% 11.53%

4 18.52% 3.99% 18.00%

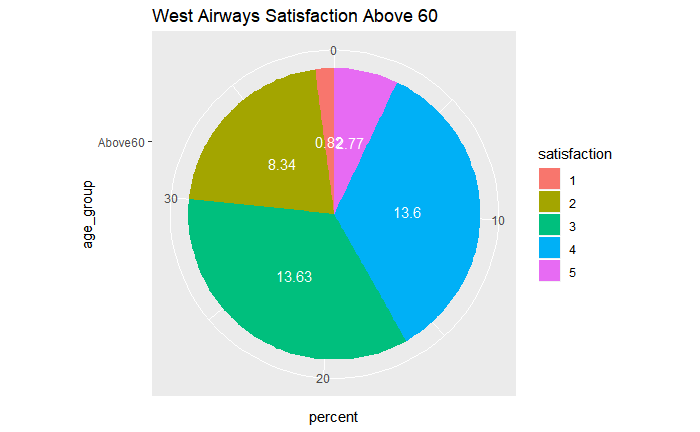
5 5.70% 0.52% 3.17%

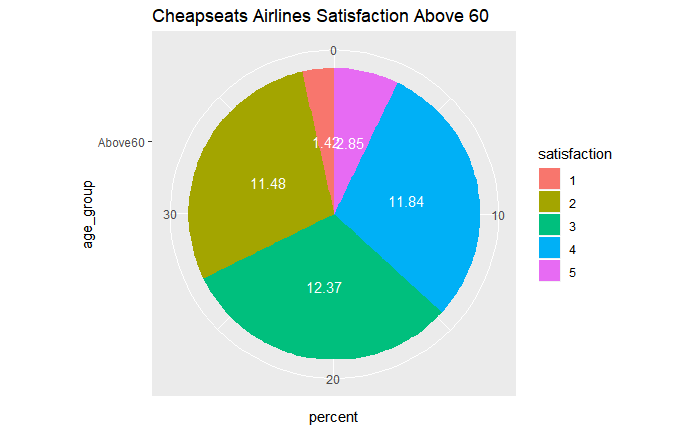
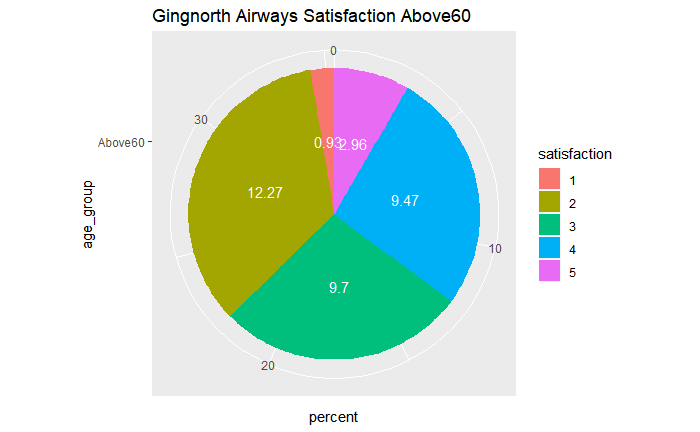
Again, people belonging to higher age group are giving lower ratings.

Based on the analysis of age for all the three airlines, we conclude that higher age is consistently giving lower ratings for GoingNorth Airlines, West Airways and CheapSeats Airlines. Therefore, this could be an area of improvement for the airlines in order to increase their customer satisfaction.

To further analyze and validate our findings, we created visualization for the customers belongs to high age group (Age >60).

* **For West Airways:**



* **CheapSeats Airlines Inc.****GoingNorth Airlines Inc**.

Based on the findings that we have gathered from the analysis; we could provide actionable insights to improve the satisfaction of the customers for these three airlines.

Data Modelling

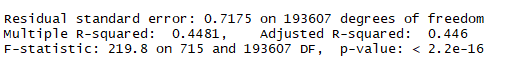
In order to derive evidence-based insights and understand which of the variables in the dataset the key drivers for overall customer satisfaction are, we have utilized three different modeling techniques for our data analysis.

**A) Linear Regression Model**

Linear Regression model is the starting point of our analysis as it allows us to monitor the statistical relationship between a variety of variables that are of interest to us. In detail, by creating linear regression models we can understand how selected variables, called independent or explanatory variables, impact one target variable, called the dependent or response variable. Customer satisfaction is going to be the dependent variable for each of our linear regression models as we seek to understand which factors impact customer satisfaction positively or negatively.

For the first linear regression model, we intentionally included all variables captured by the data to see the statistical significance and impact on the response variable – customer satisfaction. In other words, we utilize our first regression analysis to filter and exclude variables that are of no statistical significance. In order to understand whether a variable has statistical significance, we need to consider its p-value. A low p-value indicates that the null hypothesis of a given statistical model can be rejected. In plain language, an explanatory variable that has a low p-value, usually below .05, is considered to have a meaningful impact on the response variable, because changes in the response variable can be explained by changes in the explanatory variable.

**Output:**



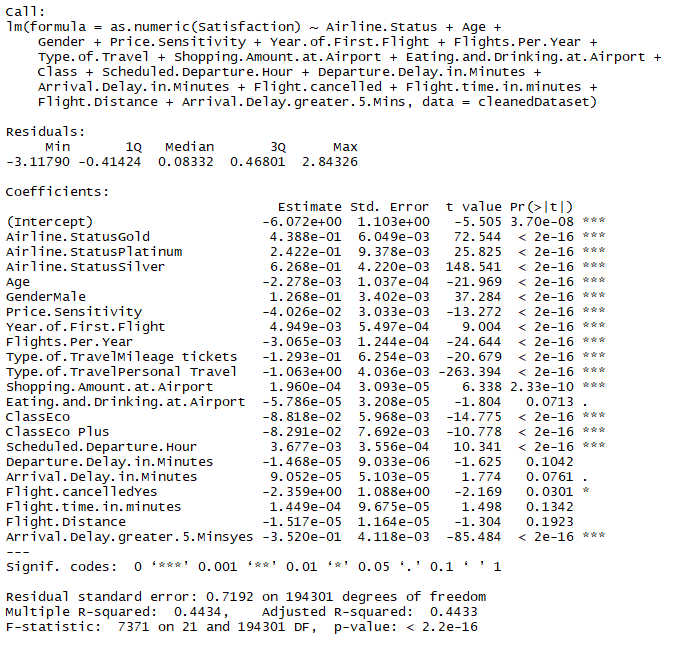
**Explanation:**

Based on the p-values of the regression result below, we notice that the following 17 variables have a statistical impact on customer satisfaction:

Airline.Status, Age, Gender, Price.Sensitivity, Year.of.First.Flight, Flights.Per.Year, Type.of.Travel, Shopping.Amount.at.Airport, Eating.and.Drinking.at.Airport, Class, Scheduled.Departure.Hour, Departure.Delay.in.Minutes, Arrival.Delay.in.Minutes, Flight.cancelled, Flight.time.in.minutes, Flight.Distance, Arrival.Delay.greater.5.Mins,

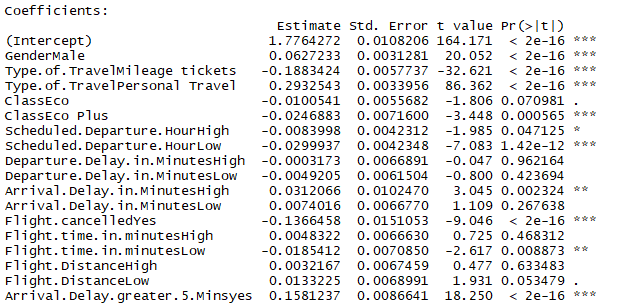
Furthermore, we can use the R-squared measure to understand the quality of our model. For this model, the adjusted R-Squared is 0.4433, which means that 44.33 percent of the variability in customer satisfaction can be explained by the variability of our explanatory variables. To put the obtained R-Squared in perspective, it is fair to say that it is rather low.

**Output:**



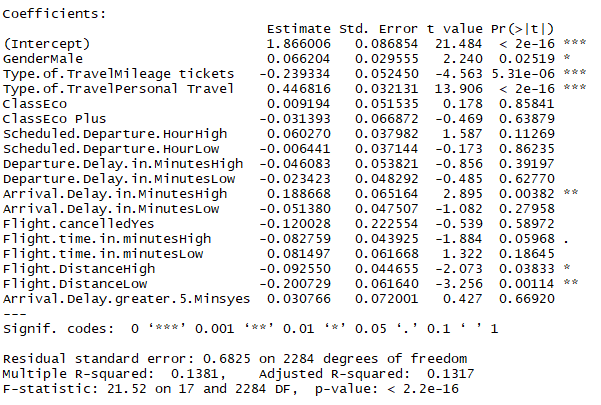
* **Linear model of WestAirways**

**Output:**



* **Linear model of GoingNorth**

**Output:**



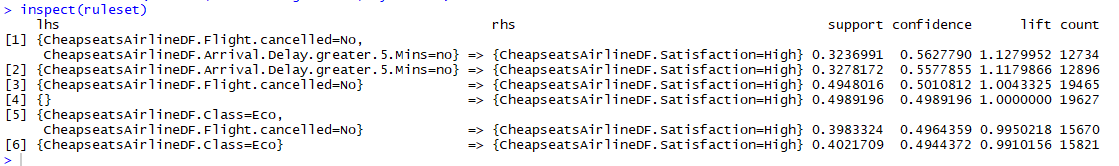
**B) Association Rules Mining**

After determining which the key drivers of customer satisfaction are, we used R’s arules package to perform association rules mining for only customers of Southeast Airline. Association rules mining is a data mining technique that is primarily used to discover and understand the co-occurrence of variables. The goal of this analysis was to find a combination of variables that inevitably lead to high or low customer satisfaction. The target variable in association rule mining model cannot be numerical and, therefore, customer satisfaction had to be converted from numerical to nominal. Purely for the purpose of this analysis, we classified customer satisfaction ratings lower than two to low, equal to three as average, and greater than three as high. Our rule selection was primarily based on the lift parameter, which is a measure of rule importance, to avoid trivial rule generation.

* **For Cheapseats Airline**

Below we found out there are 72 interesting rules we have selected for high customer satisfaction. However, 72 rules would be way too much so we narrowed it down further to 6 rules on the basis of the higher lift value.

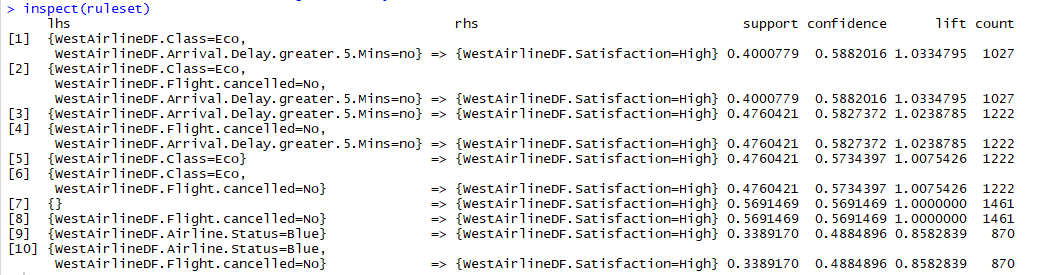
**Output:**



* **For WestAirways**

We found out 10 rules out of 332 rules.

**Output:**



**C) Support Vector Machines**

SVM is a supervised machine-learning algorithm that is used in data mining for classification purposes. SVM is considered supervised because the output dataset is already known. In this context, we are intending to build a classification model with customer satisfaction as the target variable. In other words, we are seeking to predict whether a customer has low, average, or high customer satisfaction based on personal customer information, the flight details, and flight performance. Assuming that the accuracy of the model is satisfying, Any airline can use this SVM classification model as a tool to continuously track customer satisfaction without generating any further surveys in the future. That means an Airline is longer dependent on customer feedback as they are able to generate their own predicted customer satisfaction rating.

* **For Cheapseats airlines**

The below confusion matrix shows us both statistically as well as visually that the SVM classification model was implemented successfully. Testing our SVM model on a test dataset resulted in an accuracy of approximately 88% and with an error rate of 11.66%. Out of the numbers of customers contained in the test data, the SVM model was able to predict 10206 satisfied customers correctly, while incorrectly predicting 103 customers. Moreover, the model correctly classified 1426 customers as dissatisfied while incorrectly classifying 1378 for Cheapseats airlines.

**Output:**



Similarly, we performed SVM on the dataset of West Airways and GoingNorth Airlines giving us some vital results for our validation and analysis.

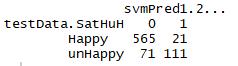
* **For WestAirways**

**Output:**



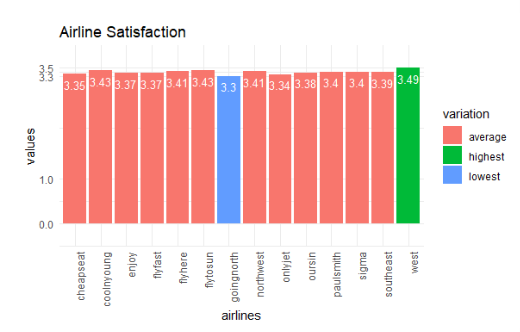
* **For GoingNorth**

**Output:**



Data Visualizations

1. **Airline Satisfaction**

****

**Explanation:**

This visualization shows the average satisfaction of all 14 airlines from the survey data. We used this visualization to determine the highest and lowest average satisfaction airlines.

**Code:**

a<-table(survey$Partner.Name, survey$Satisfaction)

a<- as.data.frame(a)

colnames(a)<-c("Partner.Name", "Satisfaction", "Count")

cheapseat <- a[a$Partner.Name == "Cheapseats Airlines Inc.", ]

cheapseat$percent <- (cheapseat$Count/sum(cheapseat$Count)) \* 100

cheapseat\_rating<- ((1 \* subset(cheapseat, cheapseat$Satisfaction == '1',select= Count))

+(2 \* subset(cheapseat, cheapseat$Satisfaction == '2',select= Count))

+(3 \* subset(cheapseat, cheapseat$Satisfaction == '3',select= Count))

+(4 \* subset(cheapseat, cheapseat$Satisfaction == '4',select= Count))

+(5 \* subset(cheapseat, cheapseat$Satisfaction == '5',select= Count)) ) / sum(cheapseat$Count)

sigma <- a[a$Partner.Name == "Sigma Airlines Inc.", ]

sigma$percent <- (sigma$Count/sum(sigma$Count)) \* 100

sigma\_rating<- ((1 \* subset(sigma, sigma$Satisfaction == '1',select= Count))

+(2 \* subset(sigma, sigma$Satisfaction == '2',select= Count))

+(3 \* subset(sigma, sigma$Satisfaction == '3',select= Count))

+(4 \* subset(sigma, sigma$Satisfaction == '4',select= Count))

+(5 \* subset(sigma, sigma$Satisfaction == '5',select= Count)) ) / sum(sigma$Count)

flyfast <- a[a$Partner.Name == "FlyFast Airways Inc.", ]

flyfast$percent <- (flyfast$Count/sum(flyfast$Count)) \* 100

flyfast\_rating<- ((1 \* subset(flyfast, flyfast$Satisfaction == '1',select= Count))

+(2 \* subset(flyfast, flyfast$Satisfaction == '2',select= Count))

+(3 \* subset(flyfast, flyfast$Satisfaction == '3',select= Count))

+(4 \* subset(flyfast, flyfast$Satisfaction == '4',select= Count))

+(5 \* subset(flyfast, flyfast$Satisfaction == '5',select= Count)) ) / sum(flyfast$Count)

northwest <- a[a$Partner.Name == "Northwest Business Airlines Inc.", ]

northwest$percent <- (northwest$Count/sum(northwest$Count)) \* 100

northwest\_rating<- ((1 \* subset(northwest, northwest$Satisfaction == '1',select= Count))

+(2 \* subset(northwest, northwest$Satisfaction == '2',select= Count))

+(3 \* subset(northwest, northwest$Satisfaction == '3',select= Count))

+(4 \* subset(northwest, northwest$Satisfaction == '4',select= Count))

+(5 \* subset(northwest, northwest$Satisfaction == '5',select= Count)) ) / sum(northwest$Count)

paulsmith <- a[a$Partner.Name == "Paul Smith Airlines Inc.", ]

paulsmith$percent <- (paulsmith$Count/sum(paulsmith$Count)) \* 100

paulsmith\_rating<- ((1 \* subset(paulsmith, paulsmith$Satisfaction == '1',select= Count))

+(2 \* subset(paulsmith, paulsmith$Satisfaction == '2',select= Count))

+(3 \* subset(paulsmith, paulsmith$Satisfaction == '3',select= Count))

+(4 \* subset(paulsmith, paulsmith$Satisfaction == '4',select= Count))

+(5 \* subset(paulsmith, paulsmith$Satisfaction == '5',select= Count)) ) / sum(paulsmith$Count)

oursin <- a[a$Partner.Name == "Oursin Airlines Inc.", ]

oursin$percent <- (oursin$Count/sum(oursin$Count)) \* 100

oursin\_rating<- ((1 \* subset(oursin, oursin$Satisfaction == '1',select= Count))

+(2 \* subset(oursin, oursin$Satisfaction == '2',select= Count))

+(3 \* subset(oursin, oursin$Satisfaction == '3',select= Count))

+(4 \* subset(oursin, oursin$Satisfaction == '4',select= Count))

+(5 \* subset(oursin, oursin$Satisfaction == '5',select= Count)) ) / sum(oursin$Count)

enjoy <- a[a$Partner.Name == "EnjoyFlying Air Services", ]

enjoy$percent <- (enjoy$Count/sum(enjoy$Count)) \* 100

enjoy\_rating<- ((1 \* subset(enjoy, enjoy$Satisfaction == '1',select= Count))

+(2 \* subset(enjoy, enjoy$Satisfaction == '2',select= Count))

+(3 \* subset(enjoy, enjoy$Satisfaction == '3',select= Count))

+(4 \* subset(enjoy, enjoy$Satisfaction == '4',select= Count))

+(5 \* subset(enjoy, enjoy$Satisfaction == '5',select= Count)) ) / sum(enjoy$Count)

southeast <- a[a$Partner.Name == "Southeast Airlines Co.", ]

southeast$percent <- (southeast$Count/sum(southeast$Count)) \* 100

southeast\_rating<- ((1 \* subset(southeast, southeast$Satisfaction == '1',select= Count))

+(2 \* subset(southeast, southeast$Satisfaction == '2',select= Count))

+(3 \* subset(southeast, southeast$Satisfaction == '3',select= Count))

+(4 \* subset(southeast, southeast$Satisfaction == '4',select= Count))

+(5 \* subset(southeast, southeast$Satisfaction == '5',select= Count)) ) / sum(southeast$Count)

onlyjet <- a[a$Partner.Name == "OnlyJets Airlines Inc.", ]

onlyjet$percent <- (onlyjet$Count/sum(onlyjet$Count)) \* 100

onlyjet\_rating<- ((1 \* subset(onlyjet, onlyjet$Satisfaction == '1',select= Count))

+(2 \* subset(onlyjet, onlyjet$Satisfaction == '2',select= Count))

+(3 \* subset(onlyjet, onlyjet$Satisfaction == '3',select= Count))

+(4 \* subset(onlyjet, onlyjet$Satisfaction == '4',select= Count))

+(5 \* subset(onlyjet, onlyjet$Satisfaction == '5',select= Count)) ) / sum(onlyjet$Count)

flytosun <- a[a$Partner.Name == "FlyToSun Airlines Inc.", ]

flytosun$percent <- (flytosun$Count/sum(flytosun$Count)) \* 100

flytosun\_rating<- ((1 \* subset(flytosun, flytosun$Satisfaction == '1',select= Count))

+(2 \* subset(flytosun, flytosun$Satisfaction == '2',select= Count))

+(3 \* subset(flytosun, flytosun$Satisfaction == '3',select= Count))

+(4 \* subset(flytosun, flytosun$Satisfaction == '4',select= Count))

+(5 \* subset(flytosun, flytosun$Satisfaction == '5',select= Count)) ) / sum(flytosun$Count)

goingnorth <- a[a$Partner.Name == "GoingNorth Airlines Inc.", ]

goingnorth$percent <- (goingnorth$Count/sum(goingnorth$Count)) \* 100

goingnorth\_rating<- ((1 \* subset(goingnorth, goingnorth$Satisfaction == '1',select= Count))

+(2 \* subset(goingnorth, goingnorth$Satisfaction == '2',select= Count))

+(3 \* subset(goingnorth, goingnorth$Satisfaction == '3',select= Count))

+(4 \* subset(goingnorth, goingnorth$Satisfaction == '4',select= Count))

+(5 \* subset(goingnorth, goingnorth$Satisfaction == '5',select= Count)) ) / sum(goingnorth$Count)

flyhere <- a[a$Partner.Name == "FlyHere Airways", ]

flyhere$percent <- (flyhere$Count/sum(flyhere$Count)) \* 100

flyhere\_rating<- ((1 \* subset(flyhere, flyhere$Satisfaction == '1',select= Count))

+(2 \* subset(flyhere, flyhere$Satisfaction == '2',select= Count))

+(3 \* subset(flyhere, flyhere$Satisfaction == '3',select= Count))

+(4 \* subset(flyhere, flyhere$Satisfaction == '4',select= Count))

+(5 \* subset(flyhere, flyhere$Satisfaction == '5',select= Count)) ) / sum(flyhere$Count)

coolnyoung <- a[a$Partner.Name == "Cool&Young Airlines Inc.", ]

coolnyoung$percent <- (coolnyoung$Count/sum(coolnyoung$Count)) \* 100

coolnyoung\_rating<- ((1 \* subset(coolnyoung, coolnyoung$Satisfaction == '1',select= Count))

+(2 \* subset(coolnyoung, coolnyoung$Satisfaction == '2',select= Count))

+(3 \* subset(coolnyoung, coolnyoung$Satisfaction == '3',select= Count))

+(4 \* subset(coolnyoung, coolnyoung$Satisfaction == '4',select= Count))

+(5 \* subset(coolnyoung, coolnyoung$Satisfaction == '5',select= Count)) ) / sum(coolnyoung$Count)

west <- a[a$Partner.Name == "West Airways Inc.", ]

west$percent <- (west$Count/sum(west$Count)) \* 100

west\_rating<- ((1 \* subset(west, west$Satisfaction == '1',select= Count))

+(2 \* subset(west, west$Satisfaction == '2',select= Count))

+(3 \* subset(west, west$Satisfaction == '3',select= Count))

+(4 \* subset(west, west$Satisfaction == '4',select= Count))

+(5 \* subset(west, west$Satisfaction == '5',select= Count)) ) / sum(west$Count)

rating1<-data.frame()

rating1 <- rbind(cheapseat\_rating, sigma\_rating, flyfast\_rating, coolnyoung\_rating, enjoy\_rating, goingnorth\_rating, northwest\_rating, onlyjet\_rating, oursin\_rating, paulsmith\_rating, southeast\_rating, west\_rating, flytosun\_rating, flyhere\_rating )

rownames(rating1)<- c("cheapseat", "sigma","flyfast", "coolnyoung", "enjoy","goingnorth","northwest","onlyjet", "oursin", "paulsmith", "southeast","west","flytosun","flyhere")

colnames(rating1) <- c("rating")

airlines <- rep(c("cheapseat", "sigma","flyfast", "coolnyoung", "enjoy","goingnorth","northwest","onlyjet", "oursin", "paulsmith", "southeast","west","flytosun","flyhere"))

ratings<- rating1$rating

values<-ratings

mydata<-data.frame(airlines,values)

is.num <- sapply(mydata, is.numeric)

mydata[is.num] <- lapply(mydata[is.num], round, 2)

mydata$variation[(mydata$values> 3.45)]<- "highest"

mydata$variation[(mydata$values< 3.31)]<- "lowest"

mydata$variation[(mydata$values> 3.31 & mydata$values< 3.45 )]<- "average"

p <-ggplot(mydata, aes(airlines, values, fill=variation))

p<- p + ggtitle("Airline Satisfaction")

p<-p +geom\_bar(stat = "identity")+ geom\_text(aes(label=values), vjust=1.6, color="white", size=3.5)+

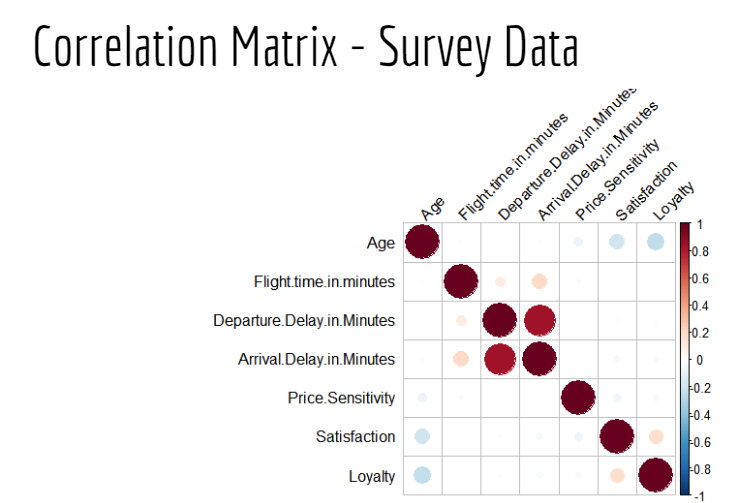
theme\_minimal()

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

p<- p+ scale\_y\_continuous(expand = c(0,0.5), breaks=c(0,1,3.3,3.5))

p

**2. Correlation Matrix on Survey Data**



**Explanation:**

This visualization shows the correlation between 7 attributes that we considered as important numeric factors to the operation of the Airlines. The light blue dot shows the negative correlations between Age to price sensitivity, satisfaction, and loyalty. As the age goes up, the price sensitivity, satisfaction, loyalty go down. This contradicts to the pink dots, which shows the positive correlations between Departure Delay in minutes, Arrival Delay in Minutes and Flight Time in minutes. As Departure Delay in minutes and Arrival Delay in Minutes go up, Flight Time in Minutes will increase.

**Code:**

survey<-read.csv("C:/Users/james/OneDrive-SyracuseUniversity/Desktop/IST687/Homework/spring19survey.csv")

source("http://www.sthda.com/upload/rquery\_cormat.r")

survey\_co<-select(survey, Price.Sensitivity, Age, Satisfaction, Loyalty, Departure.Delay.in.Minutes,Arrival.Delay.in.Minutes,Flight.time.in.minutes )

rquery.cormat(west\_airways,type="full")

**3. Word Cloud**



**Explanation:**

We created this Wordcloud to find out the most commonly used airlines in the entire data in an innovative way rather than just counting the number of occurrences of the airline name ( Partner.Name) in the data set.

Here you can see that Cheapseats is in a bigger font as compared to others as it is the most common occurred word in the data set which made us also take Cheapseats Airlines into consideration even though it does not have the best or the worst average customer rating.

Now, this Wordcloud made us choose Cheapseats airlines as another reason to analyze the airline along with the most commonly used airline in the provided dataset.

**Code:**

install.packages("tm")

library(tm)

install.packages("wordcloud")

library(wordcloud)

createWordCounts<- function(vFtext)

{

words.vec <- VectorSource(vFtext) #create a Corpus, a "Bag of Words"

words.corpus <- Corpus(words.vec)

words.corpus

words.corpus <- tm\_map(words.corpus,content\_transformer(tolower))

words.corpus <- tm\_map(words.corpus, removePunctuation)

words.corpus <- tm\_map(words.corpus, removeNumbers)

words.corpus <- tm\_map(words.corpus, removeWords, stopwords("english"))

words.corpus <- tm\_map(words.corpus, removeWords, c("airlines","inc"))

tdm<- TermDocumentMatrix(words.corpus)

tdm

m<- as.matrix(tdm)# create a matrix

wordCounts <- rowSums(m)

wordCounts<- sort(wordCounts, decreasing=TRUE)

return(wordCounts)

}

wordCounts<- createWordCounts(cleanedDataset$Partner.Name)

View(wordCounts)

genWordCloud <- function(wordCounts)

{

cloudFrame <- data.frame( word= names(wordCounts), frequency = wordCounts)

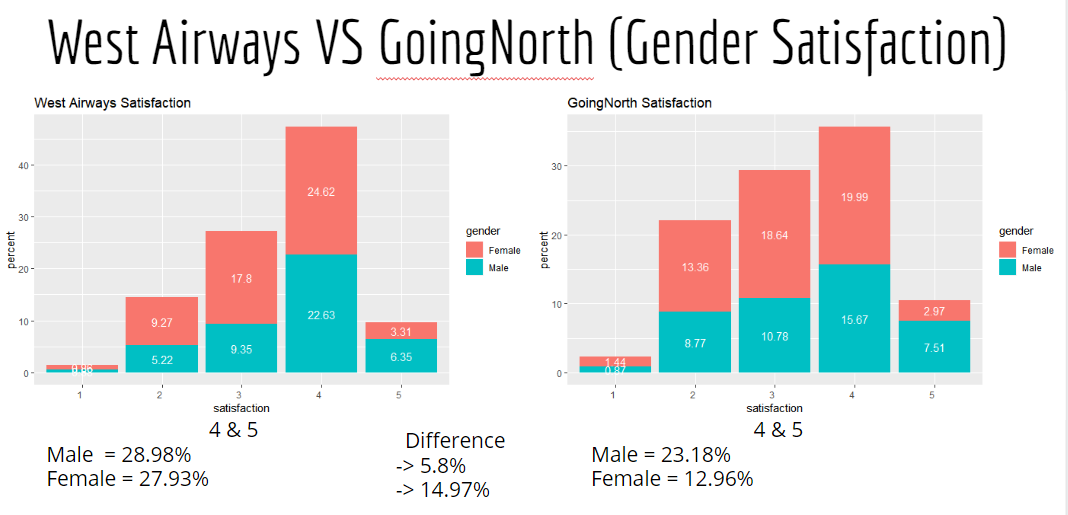
wordcloud(names(wordCounts),wordCounts, min.freq = 2, max.words=30, rot.per=0.35,

colors= brewer.pal(8,"Dark2"))

}

genWordCloud(wordCounts)

**4. Impact of Gender on Customer Satisfaction**



**Explanation:**

These graphs show Gender to Satisfaction of West Airways and GoingNorth Airlines. By showing the satisfaction on the x-axis and percentage on the y-axis, the viewers can easily read the graphs without any problem.

**Code:**

West Airways Satisfaction - Gender Satisfaction

west\_airways<- Survey[survey$Partner.Name =="West Airways Inc.", ]

west\_airways<-na.omit(west\_airways)

west\_gender<-select(west\_airways, Satisfaction, Gender)

west\_gender<-table(west\_gender$Satisfaction,west\_gender$Gender)

west\_gender<-as.data.frame(west\_gender, header=F)

names(west\_gender)<-c("satisfaction","gender","count")

west\_gender$percent<-c(west\_gender$count/2567)\*100

west\_gender$percent <- lapply(west\_gender$percent, round, 2)

p <-ggplot(west\_gender, aes(satisfaction,percent, fill=gender))

p<- p + ggtitle("West Airways Satisfaction")

p<-p +geom\_bar(stat = "identity")

theme\_minimal()

p <- p + theme (axis.text.x = element\_text(angle =0, hjust =1))+ geom\_text(aes(label=percent),position=position\_stack(vjust=0.5), colour="white")

p

GoingNorth Satisfaction - Gender Satisfaction

goingnorth<- survey[survey$Partner.Name =="GoingNorth Airlines Inc.", ]

goingnorth<-na.omit(goingnorth)

goingnorth\_gender<-select(goingnorth, Satisfaction, Gender)

goingnorth\_gender<-table(goingnorth\_gender$Satisfaction,goingnorth\_gender$Gender)

goingnorth\_gender<-as.data.frame(goingnorth\_gender, header=F)

names(goingnorth\_gender)<-c("satisfaction","gender","count")

goingnorth\_gender$percent<-c(goingnorth\_gender$count/2291)\*100

goingnorth\_gender$percent <- lapply(goingnorth\_gender$percent, round, 2)

p <-ggplot(goingnorth\_gender, aes(satisfaction,percent, fill=gender))

p<- p + ggtitle("GoingNorth Satisfaction")

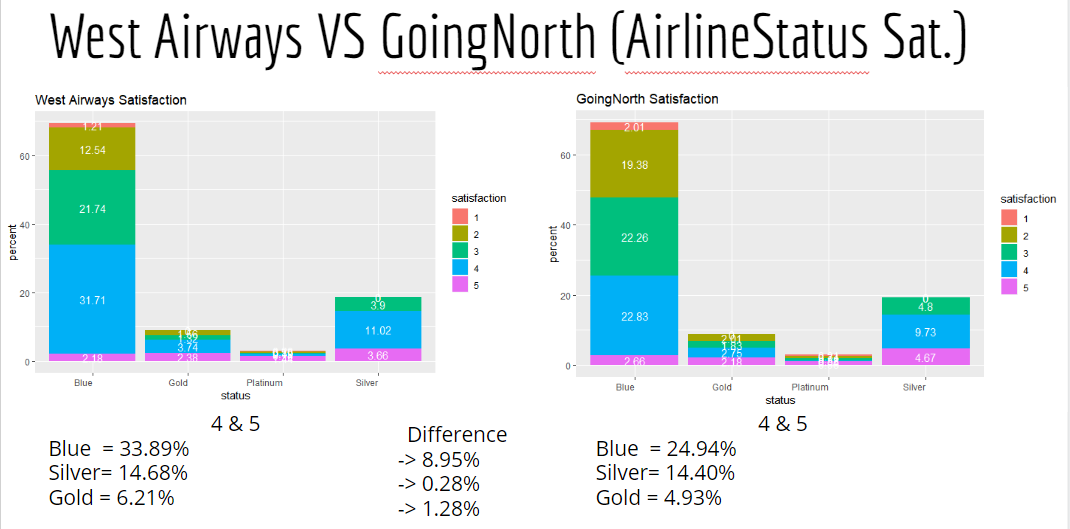
p<-p +geom\_bar(stat = "identity")

theme\_minimal()

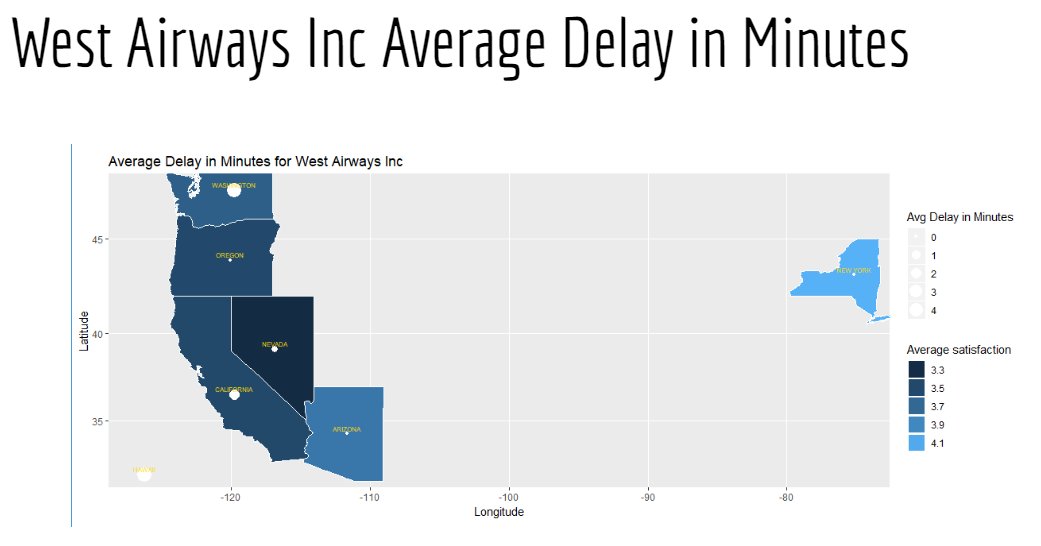
p <- p + theme (axis.text.x = element\_text(angle =0, hjust =1))+ geom\_text(aes(label=percent),position=position\_stack(vjust=0.5), colour="white")

p

**5. Impact of Airline Status on Customer Satisfaction**



**6. Average Delay in Mins**



**Explanation:**

This visualization shows the average delay in minutes for West airways who has a higher satisfaction average rating affecting the satisfaction rate according to each state in the United States.

So we can say that Avg Delay in minutes is directly proportional to Average satisfaction.

**Code:**

#for ggmaps

install.packages("sqldf")

library(sqldf)

install.packages("Rcpp") #Installing package Rcpp

library(Rcpp) #Using the library

install.packages("ggplot2") #Installing package ggplot2

library(ggplot2) #Using the library

install.packages("ggmap") #Installing package ggmap

library(ggmap) #Using the library

install.packages("maps") #Installing package maps

library(maps) #Using the library

install.packages("mapdata") #Installing package mapdata

library(mapdata) #Using the library

install.packages("mapproj") #Installing package mapproj

library(mapproj) #Using the library

View(dfAir1)

summary(dfAir1)

#Removing 9999 Values which were introduced for removing Na's

sqldf(' select count("Arrival.Delay.in.Minutes") from dfAir1 where "Arrival.Delay.in.Minutes" = 9999')

dfAirif<- dfAir1[dfAir1$Arrival.Delay.in.Minutes!=9999,]

dfAir1<- dfAirif

#extracting data form the dataset for destination and average arrival in delay

statesDelay<- sqldf('select "Destination.State" as "stateName", avg("Arrival.Delay.in.Minutes") as "adih",avg("Satisfaction") as "AverageSatisfaction" from dfAir1 group by "Destination.State"')

#taking and merging default system data with our dataset

area <- state.area

latlong<- state.center

stateName<- state.name

mergeDf<- data.frame(stateName,latlong,area)

fds<- merge(mergeDf,statesDelay, by='stateName')

#using lower case for stateName

fds$stateName<-tolower(fds$stateName)

us <- map\_data("state")

#ggmaps

mapplot<-ggplot(fds , aes(map\_id=stateName))

mapplot<- mapplot + geom\_map(map = us,aes(fill=fds$AverageSatisfaction),color="white") + guides(fill=guide\_legend(title="Average satisfaction"))

mapplot<- mapplot + expand\_limits(x= fds$x,y=fds$y)

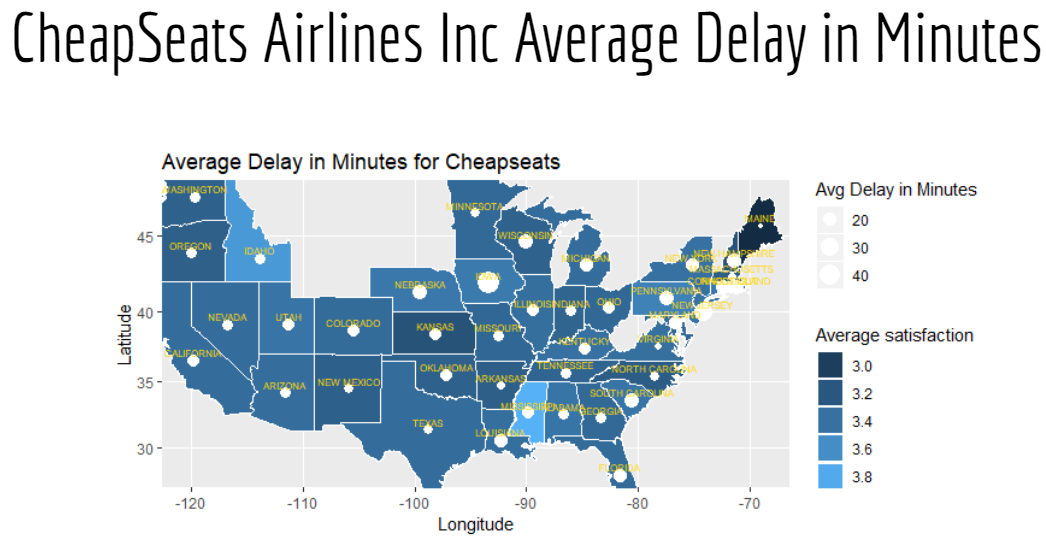
mapplot<- mapplot + geom\_point(data=fds, aes(x=fds$x,y=fds$y,size=fds$adih), color ="white")+ scale\_size(name="Avg Delay in Minutes")

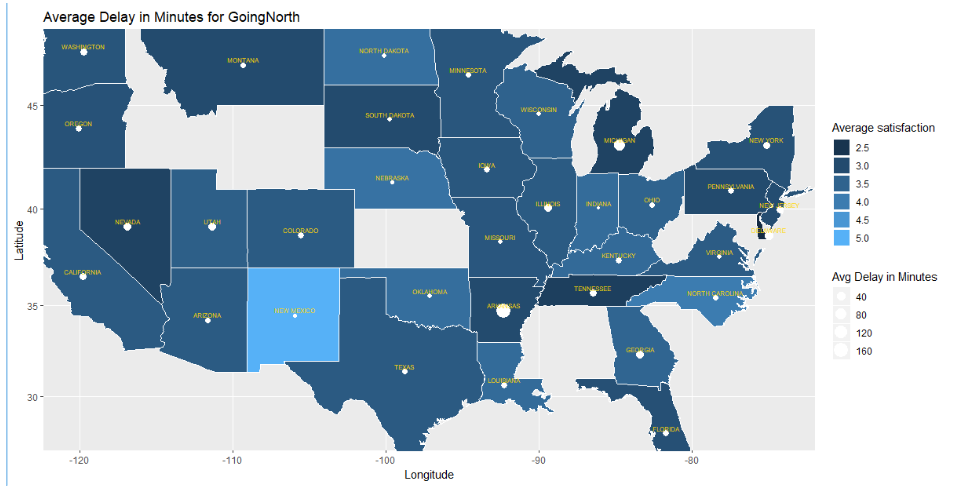
mapplot<- mapplot + geom\_text( data=fds, hjust=0.5, vjust=-0.5, aes(x=x, y=y, label=toupper(stateName)), colour="gold", size=2 )

mapplot<- mapplot + coord\_map() + ggtitle("Average Delay in Minutes for West Airways Inc")+ xlab("Longitude") + ylab("Latitude")

Mapplot

Similarly for Cheapseats and GoingNorth





**Thus, we conclude that the key Drivers for Low Customer Satisfaction**

* Gender : Female
* Airline Status : Blue
* Age : Over 60 years old customers
* Delay in Minutes : More than 5 minutes
* Flight Cancellation : Yes

Data Validation

From the above data analysis, modelling techniques and visualizations, we have identified the attributes that can lead to lower customer satisfaction. To validate the accuracy of the above attributes, we run the following models.

**Association Rules Mining**

We used Association Rules Mining for making recommendations to improve customer satisfaction by grouping some relevant attributes together.

Apriori rules were applied on GoingNorth Airlines,West Airways and Cheapseats Airlines for determining the customer satisfaction based on their Age,Flight Delays,Flight Length and Day of the month flights were being booked.

One thing we found in common for all the flights was that when the customers fell in the low and average age bucket, they were giving higher customer ratings.

Simailarly, the customer satisfaction was high when the departure delay and arrival delay of the flights was less.

This further confirmed our previous findings that recommendations should be provided to increase the customer ratings for higher age groups along with taking measures to reduce the flight delays for the airlines.

**Code:**

AirlineData <- read.csv("spring19survey.csv")

View(AirlineData)

#install the required packages

install.packages("arules")

install.packages("arulesViz")

#run the installed packages

library(arules)

library(arulesViz)

cleanedDataset <- AirlineData

View(cleanedDataset)

which(is.na(dataset$Departure.Delay.in.Minutes))

cleanedDataset$Departure.Delay.in.Minutes <- ifelse(is.na(cleanedDataset$Departure.Delay.in.Minutes) & cleanedDataset$Flight.cancelled == "Yes", 9999, cleanedDataset$Departure.Delay.in.Minutes)

cleanedDataset$Arrival.Delay.in.Minutes <- ifelse(is.na(cleanedDataset$Arrival.Delay.in.Minutes) & cleanedDataset$Flight.cancelled == "Yes", 9999, cleanedDataset$Arrival.Delay.in.Minutes)

cleanedDataset$Flight.time.in.minutes <- ifelse(is.na(cleanedDataset$Flight.time.in.minutes) & cleanedDataset$Flight.cancelled == "Yes", 9999, cleanedDataset$Flight.time.in.minutes)

cleanedDataset <- na.omit(cleanedDataset)

str(cleanedDataset)

summary(cleanedDataset)

str(cleanedDataset)

which(is.na(cleanedDataset$Departure.Delay.in.Minutes)) #Checking Na's

cleanedDataset$Satisfaction <- as.numeric(as.character(cleanedDataset$Satisfaction)) #Changing to numeric

View(cleanedDataset)

#Applying Apriori rules on Cheapseats Airlines

CheapseatsDF <- cleanedDataset[cleanedDataset$Partner.Name=="Cheapseats Airlines Inc.",]

View(CheapseatsDF)

#Creating a bucket for the attributes using quantiles

createBuckets <- function(vec){

q <- quantile(vec, c(0.4, 0.6))

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

vBuckets[vec <= q[1]] <- "Low"

vBuckets[vec > q[2]] <- "High"

return(vBuckets)

}

#Converting the given attributes to buckets by calling the createBuckets() function

age <- createBuckets(CheapseatsDF$Age)

View(age)

departuredelay <- createBuckets(CheapseatsDF$Departure.Delay.in.Minutes)

View(departuredelay)

monthDay <- createBuckets(CheapseatsDF$Day.of.Month)

View(monthDay)

flightLength <- createBuckets(CheapseatsDF$Flight.Distance)

View(flightLength)

arrivalDelay <- createBuckets(CheapseatsDF$Arrival.Delay.in.Minutes)

View(arrivalDelay)

summary(cleanedDataset)

ConvertSat <- function(vec) {

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

vBuckets[vec > 3] <- "High"

vBuckets[vec < 3] <- "Low"

return(vBuckets)

}

happyCust <- ConvertSat(CheapseatsDF$Satisfaction)

View(happyCust)

ruleDF <- data.frame(happyCust,age,departuredelay,arrivalDelay,monthDay,flightLength)

View(ruleDF)

flightSurveyX <- as(ruleDF,"transactions")

View(flightSurveyX)

ruleset <- apriori(flightSurveyX,

parameter = list(support=0.1, confidence=.5),

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

summary(ruleset)

inspect(ruleset)

# lhs rhs support confidence lift count

# [1] {age=Average} => {happyCust=High} 0.1251430 0.6545672 1.311969 4923

# [2] {monthDay=Average} => {happyCust=High} 0.1030275 0.5049838 1.012155 4053

# [3] {monthDay=High} => {happyCust=High} 0.1896591 0.5002011 1.002569 7461

# [4] {flightLength=High} => {happyCust=High} 0.2017082 0.5059296 1.014050 7935

# [5] {age=Low} => {happyCust=High} 0.2270775 0.5547069 1.111816 8933

# [6] {departuredelay=Low} => {happyCust=High} 0.2318564 0.5472819 1.096934 9121

# [7] {arrivalDelay=Low} => {happyCust=High} 0.2654109 0.5648975 1.132241 10441

# [8] {arrivalDelay=Low,monthDay=High} => {happyCust=High} 0.1050103 0.5649617 1.132370 4131

# [9] {arrivalDelay=Low,flightLength=High} => {happyCust=High} 0.1067643 0.5769231 1.156345 4200

# [10] {arrivalDelay=Low,flightLength=Low} => {happyCust=High} 0.1057729 0.5514180 1.105224 4161

# [11] {age=Low,departuredelay=Low} => {happyCust=High} 0.1066372 0.6200118 1.242709 4195

# [12] {age=Low,arrivalDelay=Low} => {happyCust=High} 0.1219146 0.6404059 1.283585 4796

# [13] {arrivalDelay=Low,monthDay=Low} => {happyCust=High} 0.1033326 0.5691683 1.140802 4065

# [14] {departuredelay=Low,arrivalDelay=Low} => {happyCust=High} 0.1973614 0.5608206 1.124070 7764

#Applying Apriori rules on West Airways

WestDF <- cleanedDataset[cleanedDataset$Partner.Name=="West Airways Inc.",]

View(WestDF)

#Creating a bucket for the attributes using quantiles

createBuckets <- function(vec){

q <- quantile(vec, c(0.4, 0.6))

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

vBuckets[vec <= q[1]] <- "Low"

vBuckets[vec > q[2]] <- "High"

return(vBuckets)

}

#Converting the given attributes to buckets by calling the createBuckets() function

age <- createBuckets(WestDF$Age)

View(age)

departuredelay <- createBuckets(WestDF$Departure.Delay.in.Minutes)

View(departuredelay)

monthDay <- createBuckets(WestDF$Day.of.Month)

View(monthDay)

flightLength <- createBuckets(WestDF$Flight.Distance)

View(flightLength)

arrivalDelay <- createBuckets(WestDF$Arrival.Delay.in.Minutes)

View(arrivalDelay)

summary(cleanedDataset)

ConvertSat <- function(vec) {

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

vBuckets[vec > 3] <- "High"

vBuckets[vec < 3] <- "Low"

return(vBuckets)

}

happyCust <- ConvertSat(WestDF$Satisfaction)

View(happyCust)

ruleDF <- data.frame(happyCust,age,departuredelay,arrivalDelay,monthDay,flightLength)

View(ruleDF)

flightSurveyX <- as(ruleDF,"transactions")

View(flightSurveyX)

ruleset <- apriori(flightSurveyX,

parameter = list(support=0.1, confidence=.5),

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

summary(ruleset)

inspect(ruleset)

# lhs rhs support confidence lift count

# [1] {} => {happyCust=High} 0.5691469 0.5691469 1.0000000 1461

# [2] {monthDay=Average} => {happyCust=High} 0.1141410 0.5955285 1.0463529 293

# [3] {age=Average} => {happyCust=High} 0.1347877 0.6878728 1.2086033 346

# [4] {flightLength=Average} => {happyCust=High} 0.1227113 0.5909944 1.0383864 315

# [5] {departuredelay=High} => {happyCust=High} 0.1246591 0.5904059 1.0373525 320

# [6] {flightLength=High} => {happyCust=High} 0.1928321 0.5749129 1.0101310 495

# [7] {arrivalDelay=High} => {happyCust=High} 0.1963381 0.5448649 0.9573361 504

# [8] {monthDay=High} => {happyCust=High} 0.2271134 0.5760870 1.0121939 583

# [9] {age=Low} => {happyCust=High} 0.2707441 0.6550424 1.1509198 695

# [10] {monthDay=Low} => {happyCust=High} 0.2278925 0.5503293 0.9669372 585

# [11] {flightLength=Low} => {happyCust=High} 0.2536034 0.5549872 0.9751213 651

# [12] {arrivalDelay=Low} => {happyCust=High} 0.3728087 0.5828258 1.0240341 957

# [13] {departuredelay=Low} => {happyCust=High} 0.4444877 0.5634568 0.9900025 1141

# [14] {departuredelay=Low,

# flightLength=Average} => {happyCust=High} 0.1036229 0.6004515 1.0550027 266

# [15] {departuredelay=High,

# arrivalDelay=High} => {happyCust=High} 0.1008960 0.5630435 0.9892763 259

# [16] {arrivalDelay=Low,

# flightLength=High} => {happyCust=High} 0.1215427 0.5988484 1.0521860 312

# [17] {departuredelay=Low,

# flightLength=High} => {happyCust=High} 0.1406311 0.5703002 1.0020264 361

# [18] {age=Low,

# monthDay=High} => {happyCust=High} 0.1024542 0.6368039 1.1188744 263

# [19] {monthDay=High,

# flightLength=Low} => {happyCust=High} 0.1044020 0.5751073 1.0104726 268

# [20] {arrivalDelay=Low,

# monthDay=High} => {happyCust=High} 0.1499805 0.5772114 1.0141695 385

# [21] {departuredelay=Low,

# monthDay=High} => {happyCust=High} 0.1803662 0.5571600 0.9789390 463

# [22] {age=Low,

# monthDay=Low} => {happyCust=High} 0.1145306 0.6697039 1.1766802 294

# [23] {age=Low,

# flightLength=Low} => {happyCust=High} 0.1188157 0.6448203 1.1329594 305

# [24] {age=Low,

# arrivalDelay=Low} => {happyCust=High} 0.1916634 0.7028571 1.2349311 492

# [25] {age=Low,

# departuredelay=Low} => {happyCust=High} 0.2212700 0.6528736 1.1471091 568

# [26] {arrivalDelay=Low,

# monthDay=Low} => {happyCust=High} 0.1484223 0.5594714 0.9830000 381

# [27] {departuredelay=Low,

# monthDay=Low} => {happyCust=High} 0.1795871 0.5481570 0.9631204 461

# [28] {arrivalDelay=Low,

# flightLength=Low} => {happyCust=High} 0.1636151 0.5592543 0.9826187 420

# [29] {departuredelay=Low,

# flightLength=Low} => {happyCust=High} 0.2002337 0.5416228 0.9516397 514

# [30] {departuredelay=Low,

# arrivalDelay=Low} => {happyCust=High} 0.3490456 0.5743590 1.0091578 896

# [31] {departuredelay=Low,

# arrivalDelay=Low,

# flightLength=High} => {happyCust=High} 0.1016751 0.5748899 1.0100905 261

# [32] {departuredelay=Low,

# arrivalDelay=Low,

# monthDay=High} => {happyCust=High} 0.1429684 0.5734375 1.0075387 367

# [33] {age=Low,

# departuredelay=Low,

# flightLength=Low} => {happyCust=High} 0.1020647 0.6501241 1.1422782 262

# [34] {age=Low,

# departuredelay=Low,

# arrivalDelay=Low} => {happyCust=High} 0.1776393 0.6898638 1.2121016 456

# [35] {departuredelay=Low,

# arrivalDelay=Low,

# monthDay=Low} => {happyCust=High} 0.1382937 0.5478395 0.9625626 355

# [36] {departuredelay=Low,

# arrivalDelay=Low,

# flightLength=Low} => {happyCust=High} 0.1620569 0.5583893 0.9810987 416

#Applying Apriori rules on goingNorth Airways

goingNorthDF <- cleanedDataset[cleanedDataset$Partner.Name=="GoingNorth Airlines Inc.",]

View(goingNorthDF)

#Creating a bucket for the attributes using quantiles

createBuckets <- function(vec){

q <- quantile(vec, c(0.4, 0.6))

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

vBuckets[vec <= q[1]] <- "Low"

vBuckets[vec > q[2]] <- "High"

return(vBuckets)

}

#Converting the given attributes to buckets by calling the createBuckets() function

age <- createBuckets(goingNorthDF$Age)

View(age)

departuredelay <- createBuckets(goingNorthDF$Departure.Delay.in.Minutes)

View(departuredelay)

monthDay <- createBuckets(goingNorthDF$Day.of.Month)

View(monthDay)

flightLength <- createBuckets(goingNorthDF$Flight.Distance)

View(flightLength)

arrivalDelay <- createBuckets(goingNorthDF$Arrival.Delay.in.Minutes)

View(arrivalDelay)

summary(cleanedDataset)

ConvertSat <- function(vec) {

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

vBuckets[vec > 3] <- "High"

vBuckets[vec < 3] <- "Low"

return(vBuckets)

}

happyCust <- ConvertSat(goingNorthDF$Satisfaction)

View(happyCust)

ruleDF <- data.frame(happyCust,age,departuredelay,arrivalDelay,monthDay,flightLength)

View(ruleDF)

flightSurveyX <- as(ruleDF,"transactions")

View(flightSurveyX)

ruleset <- apriori(flightSurveyX,

parameter = list(support=0.1, confidence=.5),

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

summary(ruleset)

inspect(ruleset)

# lhs rhs support confidence lift count

# [1] {age=Average} => {happyCust=High} 0.1194613 0.6152125 1.333540 275

# [2] {arrivalDelay=Low} => {happyCust=High} 0.2345786 0.5590062 1.211706 540

# [3] {departuredelay=Low} => {happyCust=High} 0.2593397 0.5168831 1.120400 597

# [4] {age=Low,arrivalDelay=Low} => {happyCust=High} 0.1072980 0.6365979 1.379895 247

# [5] {age=Low,departuredelay=Low} => {happyCust=High} 0.1185925 0.5711297 1.237985 273

# [6] {departuredelay=Low,flightLength=Low} => {happyCust=High} 0.1146829 0.5290581 1.146791 264

# [7] {departuredelay=Low,arrivalDelay=Low} => {happyCust=High} 0.1941790 0.5437956 1.178736 447

**Analysis on influence of Age on Satisfaction**

We carried out analysis on the age attribute for GoingNorth Airlines, West Airways and CheapSeats Airlines Inc.

**Code:**

AirlineData <- read.csv("spring19survey.csv")

View(AirlineData)

GoingNorthDF <- AirlineData[AirlineData$Partner.Name=="GoingNorth Airlines Inc.",]

View(GoingNorthDF)

WestAirDF <- AirlineData[AirlineData$Partner.Name=="West Airways Inc.",]

View(WestAirDF)

CheapSeat <- AirlineData[AirlineData$Partner.Name=="Cheapseats Airlines Inc.",]

View(CheapSeat)

ConverToBucket <- function(vec)

{

q <- quantile(vec, c(0.4, 0.6))

vBuckets[[vec >= 40 & vec <=60]] <- "Average"

vBuckets[[vec < 40]] <- "Low"

vBuckets[vec > 60] <- "High"

return(vBuckets)

}

GoingNorthDF$AgeBucket[(GoingNorthDF$Age> 60)]<- "High"

GoingNorthDF$AgeBucket[(GoingNorthDF$Age<40)]<- "Low"

GoingNorthDF$AgeBucket[(GoingNorthDF$Age>= 40 & GoingNorthDF$Age<=60)]<- "Average"

WestAirDF$AgeBucket[(WestAirDF$Age> 60)]<- "High"

WestAirDF$AgeBucket[(WestAirDF$Age<40)]<- "Low"

WestAirDF$AgeBucket[(WestAirDF$Age>= 40 & WestAirDF$Age<=60)]<- "Average"

CheapSeat$AgeBucket[(CheapSeat$Age> 60)]<- "High"

CheapSeat$AgeBucket[(CheapSeat$Age<40)]<- "Low"

CheapSeat$AgeBucket[(CheapSeat$Age>= 40 & CheapSeat$Age<=60)]<- "Average"

View(GoingNorthDF)

View(WestAirDF)

View(CheapSeat)

#GoingNorthDF$AgeBucket <- ConverToBucket(GoingNorthDF$Age)

#WestAirDF$AgeBucket <- ConverToBucket(WestAirDF$Age)

#CheapSet$AgeBucket <- ConverToBucket(CheapSet$Age)

NorthAgeonSat <- table(GoingNorthDF$Satisfaction,GoingNorthDF$AgeBucket)

NorthAgeonSat

# Average High Low

# 1 21 18 14

# 2 139 235 136

# 3 222 153 302

# 4 390 89 343

# 5 144 18 78

prop.table(NorthAgeonSat)

# Average High Low

# 1 0.009122502 0.007819288 0.006081668

# 2 0.060382276 0.102085143 0.059079062

# 3 0.096437880 0.066463944 0.131190269

# 4 0.169417897 0.038662033 0.149000869

# 5 0.062554301 0.007819288 0.033883579

WestAgeonSat <- table(WestAirDF$Satisfaction,WestAirDF$AgeBucket)

WestAgeonSat

# Average High Low

# 1 17 8 12

# 2 100 148 127

# 3 234 235 230

# 4 495 113 605

# 5 140 18 90

prop.table(WestAgeonSat)

# Average High Low

# 1 0.006609642 0.003110420 0.004665630

# 2 0.038880249 0.057542768 0.049377916

# 3 0.090979782 0.091368585 0.089424572

# 4 0.192457232 0.043934681 0.235225505

# 5 0.054432348 0.006998445 0.034992224

CheapseatAgeonSat <- table(CheapSeat$Satisfaction,CheapSeat$AgeBucket)

CheapseatAgeonSat

# Average High Low

# 1 345 362 245

# 2 2021 3478 2080

# 3 3473 3206 4546

# 4 7306 1574 7099

# 5 2250 207 1249

prop.table(CheapseatAgeonSat)

# Average High Low

# 1 0.008747243 0.009178266 0.006211810

# 2 0.051241094 0.088182348 0.052737000

# 3 0.088055577 0.081285971 0.115260769

# 4 0.185238711 0.039907710 0.179990365

# 5 0.057047235 0.005248346 0.031667554

# Lower customer satisfaction (1 & 2) is mostly given by people belonging to higher age (>60)

# This is true for both GoingNorth, WestAirways and CheapSeats

# Thus, recommendations can be suggested to improve ratings for higher age group

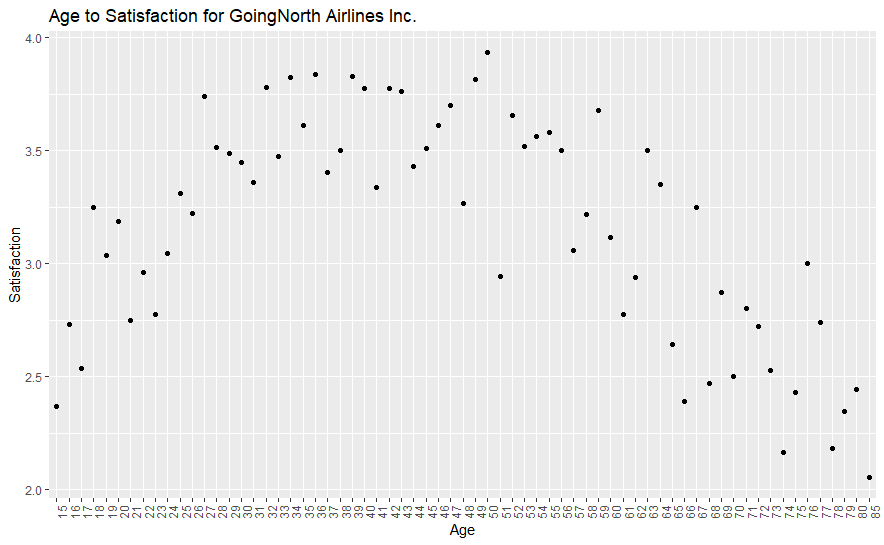
**Plotting Graphs for Age:**

# Creating a scatter plot

install.packages("ggplot2")

library(ggplot2)

**Plotting Age to Satisfaction for Going North**



GoingNorthGraph <-tapply(GoingNorthDF$Satisfaction, GoingNorthDF$Age,mean)

GoingNorthData <-data.frame(customersatisfaction=GoingNorthGraph,age=row.names(GoingNorthGraph))

GoingNorthData

GoingNorthbar <- ggplot(GoingNorthData, aes(x=age,customersatisfaction, y=customersatisfaction))

GoingNorthbar <- GoingNorthbar + geom\_point(color = "Black", fill = "Grey")

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

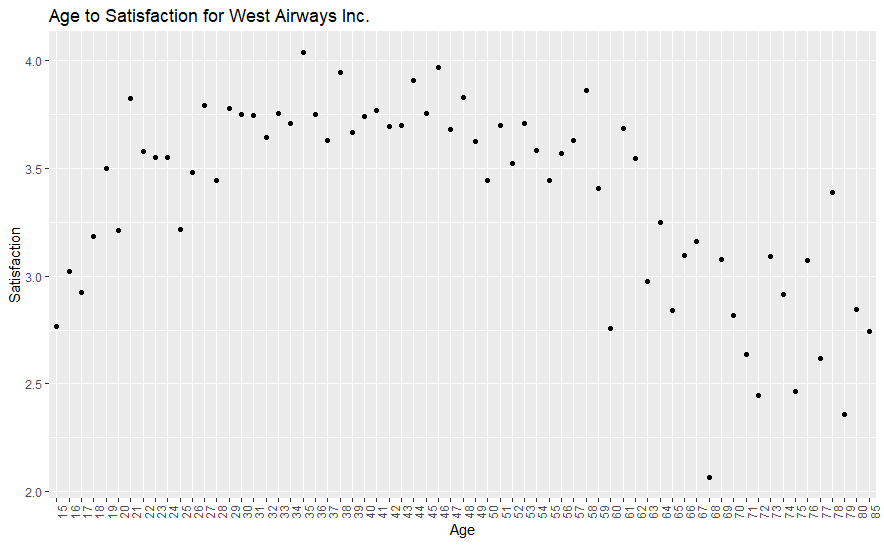
GoingNorthbar <- GoingNorthbar + ggtitle("Age to Satisfaction for GoingNorth Airlines Inc.")

GoingNorthbar <- GoingNorthbar + labs(x = "Age")

GoingNorthbar <- GoingNorthbar + labs(y = "Satisfaction")

GoingNorthbar

**Plotting Age to Satisfaction for West Airways**



WestAirGraph <-tapply(WestAirDF$Satisfaction, WestAirDF$Age, mean)

WestAirData <-data.frame(customersatisfaction=WestAirGraph,age=row.names(WestAirGraph))

WestAirData

WestAirbar <- ggplot(WestAirData, aes(x=age,customersatisfaction, y=customersatisfaction))

WestAirbar <- WestAirbar + geom\_point(color = "Black", fill = "Grey")

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

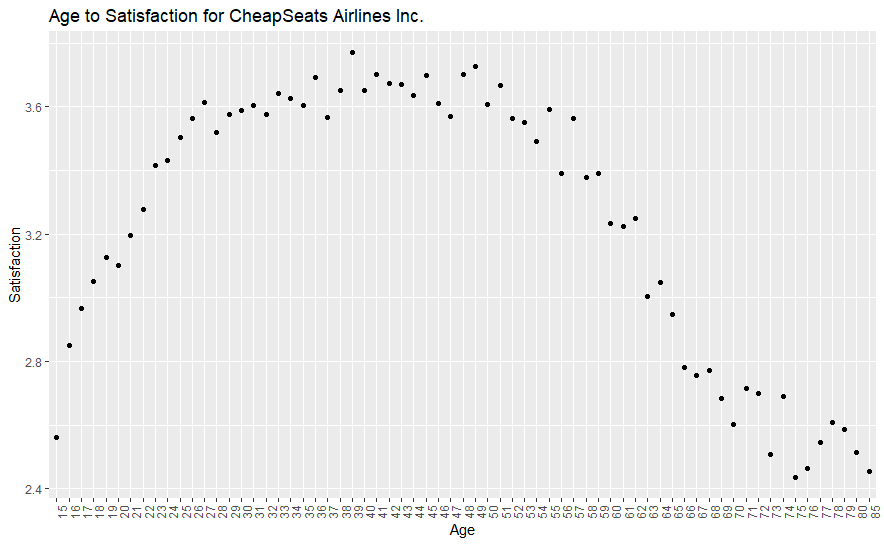
WestAirbar <- WestAirbar + ggtitle("Age to Satisfaction for West Airways Inc.")

WestAirbar <- WestAirbar + labs(x = "Age")

WestAirbar <- WestAirbar + labs(y = "Satisfaction")

WestAirbar

**Plotting Age to Satisfaction for Cheapseat Airlines**



CheapseatGraph <-tapply(CheapSeat$Satisfaction, CheapSeat$Age, mean)

CheapSetData <-data.frame(customersatisfaction=CheapseatGraph,age=row.names(CheapseatGraph))

CheapSetData

cheapSetbar <- ggplot(CheapSetData, aes(x=age,customersatisfaction, y=customersatisfaction))

cheapSetbar <- cheapSetbar + geom\_point(color = "Black", fill = "Grey")

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

cheapSetbar <- cheapSetbar + ggtitle("Age to Satisfaction for CheapSeats Airlines Inc.")

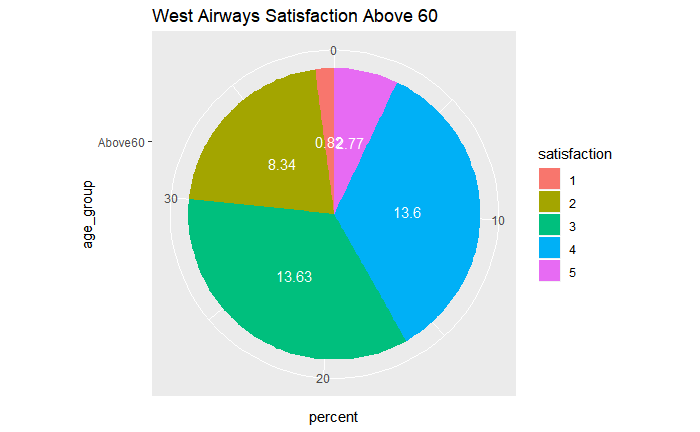
cheapSetbar <- cheapSetbar + labs(x = "Age")

cheapSetbar <- cheapSetbar + labs(y = "Satisfaction")

cheapSetbar

We further perform in depth analysis on satisfaction of customers belonging to high age group (Age>60)

**West Airways**



createBuckets <- function(vec){

  q <- quantile(vec, c(0.4, 0.6))

  vBuckets <- replicate(length(vec), "40-60")

  vBuckets[vec <= q[1]] <- "Below40"

  vBuckets[vec > q[2]] <- "Above60"

  return(vBuckets)

}

createLBuckets <- function(vec){

  q <- quantile(vec, c(0.4, 0.6))

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

  vBuckets[vec <= q[1]] <- "Lower\_than\_40%"

  vBuckets[vec > q[2]] <- "Higher\_than\_60%"

  return(vBuckets)

}

west\_airways<-survey[survey$Partner.Name == "West Airways Inc.", ]

west\_airways<-select(west\_airways, Age, Price.Sensitivity)

age\_bucket <- createBuckets(west\_airways$Age)

west\_airways<- cbind(west\_airways,age\_bucket)

west\_age<-table(west\_airways$Price.Sensitivity, west\_airways$age\_bucket)

west\_age<-as.data.frame(west\_age, header=F)

names(west\_age)<-c("price.sensitivity","age\_group", "count")

west\_age$percent<-c(west\_age$count/2567)\*100

west\_age$percent <- lapply(west\_age$percent, round, 2)

west\_low<- west\_age[west\_age$age\_group == "Below40",]

west\_avg<- west\_age[west\_age$age\_group == "40-60",]

west\_high<- west\_age[west\_age$age\_group == "Above60",]

p <-ggplot(west\_age, aes(age\_group,percent, fill=Loyalty))

p<- p + ggtitle("West Airways Price Sensitivity Below40")

p<-p +geom\_bar(stat = "identity")

  theme\_minimal()

p <- p + theme (axis.text.x = element\_text(angle =0, hjust =1))+ geom\_text(aes(label=percent),position=position\_stack(vjust=0.5), colour="white")

p

west\_airways<-survey[survey$Partner.Name == "West Airways Inc.", ]

west\_airways<-select(west\_airways, Age, Satisfaction)

age\_bucket <- createBuckets(west\_airways$Age)

west\_airways<- cbind(west\_airways,age\_bucket)

west\_age<-table(west\_airways$Satisfaction, west\_airways$age\_bucket)

west\_age<-as.data.frame(west\_age, header=F)

names(west\_age)<-c("satisfaction","age\_group", "count")

west\_age$percent<-c(west\_age$count/2572)\*100

west\_age$percent <- lapply(west\_age$percent, round, 2)

west\_low<- west\_age[west\_age$age\_group == "Below40",]

west\_avg<- west\_age[west\_age$age\_group == "40-60",]

west\_high<- west\_age[west\_age$age\_group == "Above60",]

p <-ggplot(west\_high, aes(age\_group,percent, fill=satisfaction))

p<- p + ggtitle("West Airways Satisfaction Above60")

p<-p +geom\_bar(stat = "identity")

  theme\_minimal()

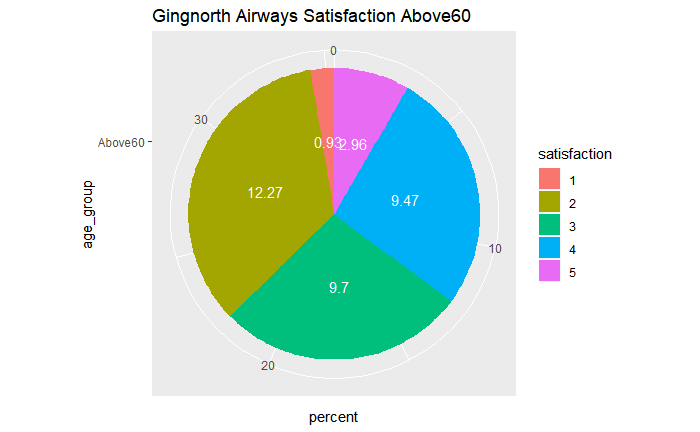
p <- p + theme (axis.text.x = element\_text(angle =0, hjust =1))+ geom\_text(aes(label=percent),position=position\_stack(vjust=0.5), colour="white")

p

pie <- p + coord\_polar("y", start=0)

pie

**GoingNorth Airlines Inc.**



goingnorth\_airways<-survey[survey$Partner.Name == "GoingNorth Airlines Inc.", ]

goingnorth\_airways<-select(goingnorth\_airways, Age, Satisfaction)

age\_bucket <- createBuckets(goingnorth\_airways$Age)

goingnorth\_airways<- cbind(goingnorth\_airways,age\_bucket)

goingnorth\_age<-table(goingnorth\_airways$Satisfaction, goingnorth\_airways$age\_bucket)

goingnorth\_age<-as.data.frame(goingnorth\_age, header=F)

names(goingnorth\_age)<-c("satisfaction","age\_group", "count")

goingnorth\_age$percent<-c(goingnorth\_age$count/2567)\*100

goingnorth\_age$percent <- lapply(goingnorth\_age$percent, round, 2)

goingnorth\_low<- goingnorth\_age[goingnorth\_age$age\_group == "Below40",]

goingnorth\_avg<- goingnorth\_age[goingnorth\_age$age\_group == "40-60",]

goingnorth\_high<- goingnorth\_age[goingnorth\_age$age\_group == "Above60",]

p <-ggplot(goingnorth\_high, aes(age\_group,percent, fill=Loyalty))

p<- p + ggtitle("Gingnorth Airways Loyalty Above60")

p<-p +geom\_bar(stat = "identity")

  theme\_minimal()

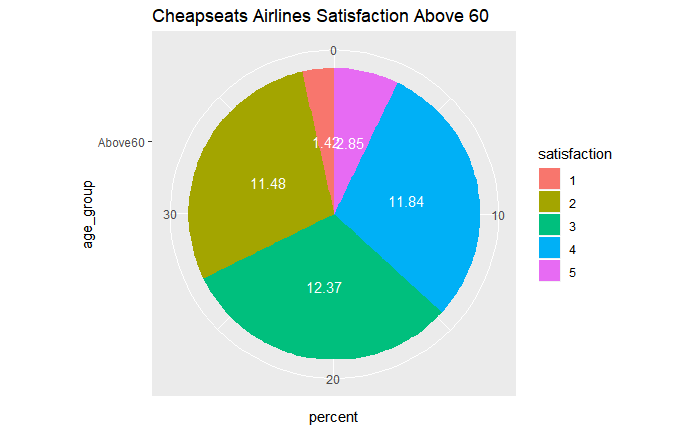
p <- p + theme (axis.text.x = element\_text(angle =0, hjust =1))+ geom\_text(aes(label=percent),position=position\_stack(vjust=0.5), colour="white")

p

pie <- p + coord\_polar("y", start=0)

pie

**For CheapSeat Airlines Inc.**



cheapseat\_airways<-survey[survey$Partner.Name == "Cheapseats Airlines Inc.", ]

cheapseat\_airways<-select(cheapseat\_airways, Age, Satisfaction)

age\_bucket <- createBuckets(cheapseat\_airways$Age)

cheapseat\_airways<- cbind(cheapseat\_airways,age\_bucket)

cheapseat\_age<-table(cheapseat\_airways$Satisfaction, cheapseat\_airways$age\_bucket)

cheapseat\_age<-as.data.frame(cheapseat\_age, header=F)

names(cheapseat\_age)<-c("satisfaction","age\_group", "count")

cheapseat\_age$percent<-c(cheapseat\_age$count/39441)\*100

cheapseat\_age$percent <- lapply(cheapseat\_age$percent, round, 2)

cheapseat\_low<- cheapseat\_age[cheapseat\_age$age\_group == "Below40",]

cheapseat\_avg<- cheapseat\_age[cheapseat\_age$age\_group == "40-60",]

cheapseat\_high<- cheapseat\_age[cheapseat\_age$age\_group == "Above60",]

p <-ggplot(cheapseat\_high, aes(age\_group,percent, fill=satisfaction))

p<- p + ggtitle("Cheapseats Airlines Satisfaction Above 60")

p<-p +geom\_bar(stat = "identity")

  theme\_minimal()

p <- p + theme (axis.text.x = element\_text(angle =0, hjust =1))+ geom\_text(aes(label=percent),position=position\_stack(vjust=0.5), colour="white")

p

pie <- p + coord\_polar("y", start=0)

pie

Response: Business Questions

* Age plays a quite significant role with respect to customer satisfaction. People within the age group of 30 to 65 are the travelers that give the highest customer satisfaction for all the three airlines analyzed. Furthermore, we identified that people that are older than 70 years old are giving the worse ratings.
* Delay in arrival of flights at major cities like Los Angeles, San Jose, Seattle, San Diego, Phoenix, Flint, Norfolk, Rochester, West Palm Beach/Palm Beach, New York, NY has affected the satisfaction of the customers.

Recommendations and actionable insights

1. Aim to make the delay time less than 5 minutes, so that customers feel that their time is valued and in return can provide better ratings.
2. Offer better price and services to Blue Status Customers, as most customers belong to the Blue Status category and are giving lower ratings.
3. Collect Female suggestions and feedback to improve services that target female customers to make them feel more involved.
4. Give special assistance like wheelchair assistance, seats with greater leg space and comfort to the elderly as they are giving lower ratings