Customer Satisfaction In The Airline Industry

SouthEast Airlines

Data Analysis Report

- Hope Puzzanghera
- Shwet Jain
- Osama Junaid
- Pat Carlin

CONTENTS

Chapter 1 - INTRODUCTION

Introduction:

In the last 25 years, the aviation industry has been growing rapidly. In addition to its technological developments, the growing of airline industry is due to its role as supporting the world trade, international investment, and tourism activities. Because of these roles, it is often said that the aviation industry is the center of globalization for other industries. The growing of the airline industry provides opportunities as well as challenges to the business entities in this industry. The opportunities arise due to the increasing demand for the airline services. While the challenges arise not only because of the high level of competition between the airlines, but also due to growing consumer demands for better service. Growing of this industry has led to the governments around the world to deregulate the industry by enabling companies to form private airlines. Further, the stringent rules on safety flight is currently applied in almost all countries. These conditions cause all players in the airline industry require to continuously innovate in terms of both services and technology used to deliver services and better safety to the consumers. This report provides analysis of SouthEast Airlines on different factors that influence consumer satisfaction which can then be used to drive customers satisfaction and thus provide a competitive edge.

Background

The Dataset contains about 129889 responses (rows) from airline customers survey throughout 3 months, and contains data from 14 airlines. The Dataset has 28 columns, which consist of data obtained from surveys submitted by its airline customers. The columns broadly focus on several categories, including customer's gender, age, number of flights, shopping amount at the airport, type of travel, etc.

Objective

We are working as a consulting company for Southeast Airlines. The focus of our survey analysis project is to provide some useful insights to Southeast Airlines so that they can improve customer satisfaction which will drive business and in turn profits.

Context:

The Airlines data is a dataset collected from information about the customers taking various airline flights and giving their satisfaction ratings about the overall experience they had with the flight.

Scope :

The study will benefit the airline company 'South Asia' in evaluating their performance. It could have a competitive advantage if they could brand themselves appropriately. This study will help Airline 'South Asia' and airport authority in developing an effective service. The results of the study will be significant to the airline 'South Asia' in which the majority passengers belong to and also it enables other airlines to offer better service. It will serve as a guide as how passengers prefer airline and the satisfaction level of passengers will guide the airline 'SouthAsia" for their improvement.

Chapter 2 - BUSINESS QUESTIONS ADDRESSED

Every research or analysis needs to have a business reason.

The two key reasons are:

Improved customer service such as those below. There can be more than one reason :

- Increasing take up of services among key groups to achieve targets
- Making it easier to access services
- Giving a better service
- Giving a service targeted to individual needs
- Giving access to a broader range of services

Improving efficiency by one or more of:

- Increasing take up among key groups to increase income
- Increasing early take up and reducing more expensive interventions later
- Improving processes to streamline services and reduce costs (including one touch contact)
- Switching customers to more cost efficient channels

1) Does Gender play an important role wrt. Satisfaction?
2) What are the important attributes that drive Satisfaction ?
3) Does Origin City and Destination City affect Satisfaction?
4) Does No.of Other Flights taken by the customer affect Satisfaction?
5) Discuss the correlation between Distance Covered and Arrival Delay?
6) Does the relationship between the Origin City and the Destination City together have something to do with Customer Satisfaction.
7) Does the relationship between Price Sensitivity, Type Of Travel affect the corresponding Satisfaction?
8) Relationship between the Airline Status and Satisfaction.

Chapter 3 - DATA EXPLORATION

The process of amending or removing data in a database that is incorrect, incomplete, improperly formatted, or duplicated.

Data Acquisition

1) Approximately 1,30,000 survey responses
2) 25 fields in the Survey
3) Some entries in the data-set are blank (NA)

File Used: "**Satisfaction Survey.csv**"

Pre-Cleansing

Incorrect, Incomplete, Improperly formatted and duplicated

Class	Day.of.Month	Flight.date	Airline.Code	Airline.Name	Orgin.City	Origin.State	Destination.City	Destination.State
Business	18	3/18/14	MQ	EnjoyFlying Air Services	Madison, WI	Wisconsin	Dallas/Fort Worth, TX	Texas
Business	11	1/11/14	MQ	EnjoyFlying Air Services	Madison, WI	Wisconsin	Dallas/Fort Worth, TX	Texas
Business	25	1/25/14	MQ	EnjoyFlying Air Services	Milwaukee, WI	Wisconsin	Dallas/Fort Worth, TX	Texas
Eco	20	2/20/14	MQ	EnjoyFlying Air Services	Madison, WI	Wisconsin	Dallas/Fort Worth, TX	Texas
Eco	25	2/25/14	MQ	EnjoyFlying Air Services	Milwaukee, WI	Wisconsin	Dallas/Fort Worth, TX	Texas
Eco	16	1/16/14	MQ	EnjoyFlying Air Services	Madison, WI	Wisconsin	Dallas/Fort Worth, TX	Texas
Eco	6	3/6/14	MQ	EnjoyFlying Air Services	Madison, WI	Wisconsin	Dallas/Fort Worth, TX	Texas
Eco	5	2/5/14	MQ	EnjoyFlying Air Services	Madison, WI	Wisconsin	Dallas/Fort Worth, TX	Texas
Eco	21	1/21/14	MQ	EnjoyFlying Air Services	Milwaukee, WI	Wisconsin	Dallas/Fort Worth, TX	Texas
Eco	19	1/19/14	MQ	EnjoyFlying Air Services	Madison, WI	Wisconsin	Dallas/Fort Worth, TX	Texas
Eco	19	3/19/14	MQ	EnjoyFlying Air Services	Milwaukee, WI	Wisconsin	Dallas/Fort Worth, TX	Texas
Eco	4	2/4/14	MQ	EnjoyFlying Air Services	Milwaukee, WI	Wisconsin	Dallas/Fort Worth, TX	Texas

Also, handling the NA' s in the data-set.

Cleansing, Transformation & Munging

1. Dealing with Column Names:

```
colnames(ProjectData)[6]<- "YearOfFirstFlight"
colnames(ProjectData)[8]<- "FlightsWithOtherAirlines"
colnames(ProjectData)[7]<- "FlightsPerYear"
colnames(ProjectData)[10]<- "NoOfOtherLoyaltyCards"
colnames(ProjectData)[11]<- "ShoppingAtAirport"
colnames(ProjectData)[12]<- "EatingAndDrinkingAtAirport"
colnames(ProjectData)[14]<- "DayOfMonth"
colnames(ProjectData)[15]<- "FlightDate"
colnames(ProjectData)[25]<- "FlightsCancelled"
colnames(ProjectData)[26]<- "FlightTimeInMinutes"
colnames(ProjectData)[28]<- "ArrivalDelayGreaterThan5mins"
colnames(ProjectData)[23]<- "DepartureDelayInMinutes"
colnames(ProjectData)[24]<- "ArrivalDelayInMinutes"
colnames(ProjectData)[18]<- "OriginCity"
```

2. Dealing with ' . ' s in the Column Names:

```
names(ProjectData) <- gsub("\\.","", names(ProjectData))
```

3. Dealing with NA's: NA's will be replaced by means of their respective columns

```
SouthData$ArrivalDelayInMinutes[is.na(SouthData$ArrivalDelayInMinutes)] <- round(mean(SouthData$ArrivalDelayInMinutes, na.rm = TRUE))
SouthData$FlightTimeInMinutes[is.na(SouthData$FlightTimeInMinutes)] <- round(mean(SouthData$FlightTimeInMinutes, na.rm = TRUE))
SouthData$DepartureDelayInMinutes[is.na(SouthData$DepartureDelayInMinutes)] <- round(mean(SouthData$DepartureDelayInMinutes, na.rm = TRUE))
```

4. Dealing with City,State in the Same Column:

```
ProjectData$DestinationCity<- gsub("(.*),.*", "\\1",ProjectData$DestinationCity)
ProjectData$OriginCity<- gsub("(.*),.*", "\\1",ProjectData$OriginCity)
```

5. Extracting Data (SouthEast Airlines)

```
SouthData <- ProjectData[which(ProjectData$AirlineName == "Southeast Airlines Co. " ),] #For Seperate DataFrame
str(SouthData)
```

Data Set After Performing Cleansing, Transformation & Munging Operations

Class	DayOfMonth	FlightDate	AirlineCode	AirlineName	OriginCity	OriginState	DestinationCity	DestinationState
Business	13	2/13/14	US	Southeast Airlines Co.	Milwaukee	Wisconsin	Phoenix	Arizona
Business	25	3/25/14	US	Southeast Airlines Co.	Milwaukee	Wisconsin	Phoenix	Arizona
Business	8	3/8/14	US	Southeast Airlines Co.	Milwaukee	Wisconsin	Phoenix	Arizona
Eco	13	2/13/14	US	Southeast Airlines Co.	Milwaukee	Wisconsin	Phoenix	Arizona
Eco	26	1/26/14	US	Southeast Airlines Co.	Milwaukee	Wisconsin	Phoenix	Arizona
Eco	8	3/8/14	US	Southeast Airlines Co.	Milwaukee	Wisconsin	Phoenix	Arizona
Eco	17	3/17/14	US	Southeast Airlines Co.	Milwaukee	Wisconsin	Phoenix	Arizona
Eco	2	2/2/14	US	Southeast Airlines Co.	Milwaukee	Wisconsin	Phoenix	Arizona
Eco	4	2/4/14	US	Southeast Airlines Co.	Milwaukee	Wisconsin	Phoenix	Arizona
Eco	14	2/14/14	US	Southeast Airlines Co.	Milwaukee	Wisconsin	Phoenix	Arizona
Eco	21	3/21/14	US	Southeast Airlines Co.	Milwaukee	Wisconsin	Phoenix	Arizona
Eco	19	1/19/14	US	Southeast Airlines Co.	Milwaukee	Wisconsin	Phoenix	Arizona
Eco	11	3/11/14	US	Southeast Airlines Co.	Milwaukee	Wisconsin	Phoenix	Arizona
Eco	27	2/27/14	US	Southeast Airlines Co.	Milwaukee	Wisconsin	Phoenix	Arizona
Eco	19	3/19/14	US	Southeast Airlines Co.	Milwaukee	Wisconsin	Phoenix	Arizona
Eco	30	1/30/14	US	Southeast Airlines Co.	Milwaukee	Wisconsin	Phoenix	Arizona
Eco	2	3/2/14	US	Southeast Airlines Co.	Milwaukee	Wisconsin	Phoenix	Arizona
Eco Plus	19	3/19/14	US	Southeast Airlines Co.	Milwaukee	Wisconsin	Phoenix	Arizona
Eco	16	2/16/14	US	Southeast Airlines Co.	Seattle	Washington	Philadelphia	Pennsylvania
Eco	3	2/3/14	US	Southeast Airlines Co.	Seattle	Washington	Philadelphia	Pennsylvania
Eco	14	3/14/14	US	Southeast Airlines Co.	Seattle	gton	Philadelphia	Pennsylvania
Eco	4	1/4/14	US	Southeast Airlines Co.	Seattle	Washington	Philadelphia	Pennsylvania
Eco	6	1/6/14	US	Southeast Airlines Co.	Seattle	Washington	Philadelphia	Pennsylvania

Seattle

Chapter 4 - DESCRIPTIVE STATISTICS AND VISUALIZATION

Descriptive Statistics

- Used to describe the basic features of the data in a study.
- Provide simple summaries about the sample and the measures.

Screenshots: Using the summary(), range(), and sd() functions we get

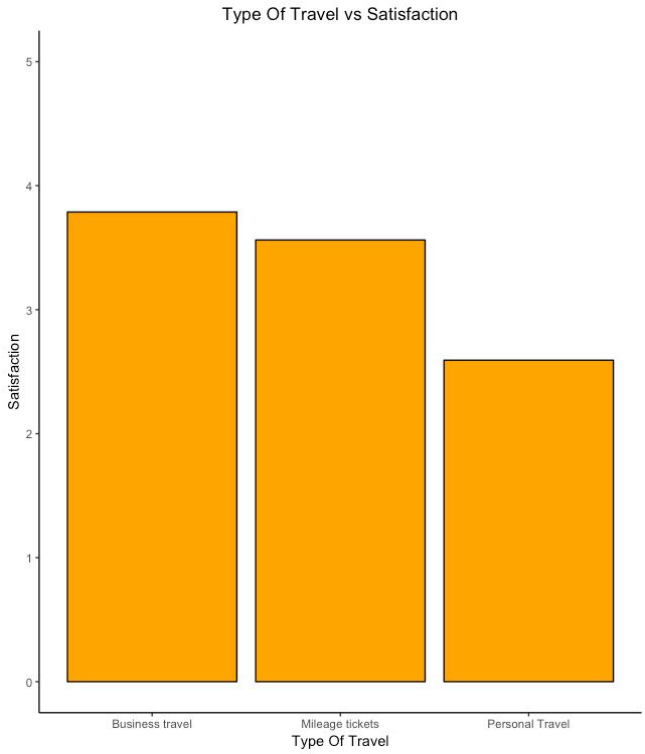
```
> summary(ProjectData)
  Satisfaction    AirlineStatus        Age          Gender      PriceSensitivity YearOfFirstFlight
 4      :53758   Blue    :88910   Min.   :15.0   Female:73374   Min.   :0.000    Min.   :2003
 3      :36984   Gold    :10837   1st Qu.:33.0   Male  :56515   1st Qu.:1.000    1st Qu.:2004
 2      :23587   Platinum: 4172   Median :45.0                  Median :1.000    Median :2007
 5      :12552   Silver  :25970   Mean   :46.2                  Mean   :1.276    Mean   :2007
 1      : 2999                    3rd Qu.:59.0                  3rd Qu.:2.000    3rd Qu.:2010
 2.5    :    2                    Max.   :85.0                  Max.   :5.000    Max.   :2012
 (Other):    7
 FlightsPerYear   FlightsWithOtherAirlines          TypeofTravel   NoOfOtherLoyaltyCards ShoppingAtAirport
 Min.   :  0.00   Min.   :  1.000          Business travel:79630   Min.   : 0.0000       Min.   :  0.00
 1st Qu.:  9.00   1st Qu.:  4.000          Mileage tickets:10070   1st Qu.: 0.0000       1st Qu.:  0.00
 Median : 17.00   Median :  7.000          Personal Travel:40189   Median : 0.0000       Median :  0.00
 Mean   : 20.08   Mean   :  9.314                                  Mean   : 0.8838       Mean   : 26.55
 3rd Qu.: 29.00   3rd Qu.: 10.000                                  3rd Qu.: 2.0000       3rd Qu.: 30.00
 Max.   :100.00   Max.   :110.000                                  Max.   :12.0000       Max.   :879.00

 EatingAndDrinkingAtAirport      Class          DayOfMonth      FlightDate      AirlineCode
 Min.   :  0.00             Business: 10548   Min.   : 1.00   3/13/14:  1641   WN     :26058
 1st Qu.: 30.00             Eco     :105735   1st Qu.: 8.00   3/10/14:  1640   DL     :17037
 Median : 60.00             Eco Plus: 13606   Median :16.00   3/21/14:  1638   EV     :15407
 Mean   : 68.24                               Mean   :15.72   3/26/14:  1628   OO     :13840
 3rd Qu.: 90.00                               3rd Qu.:23.00   3/27/14:  1622   AA     :12248
 Max.   :895.00                               Max.   :31.00   3/24/14:  1619   OU     :10968
                                                              (Other):120101   (Other):34331
                          AirlineName     OriginCity          OriginState    DestinationCity
 Cheapseats Airlines Inc.         :26058   Length:129889      California:16751   Length:129889
 Sigma Airlines Inc.              :17037   Class :character   Texas     :16346   Class :character
 FlyFast Airways Inc.             :15407   Mode  :character   Florida   :10894   Mode  :character
 Northwest Business Airlines Inc. :13840                      Georgia   : 8751
 Paul Smith Airlines Inc.         :12248                      Illinois  : 7989
```

This helped us get a basic overview of the entire data-set

Visualization

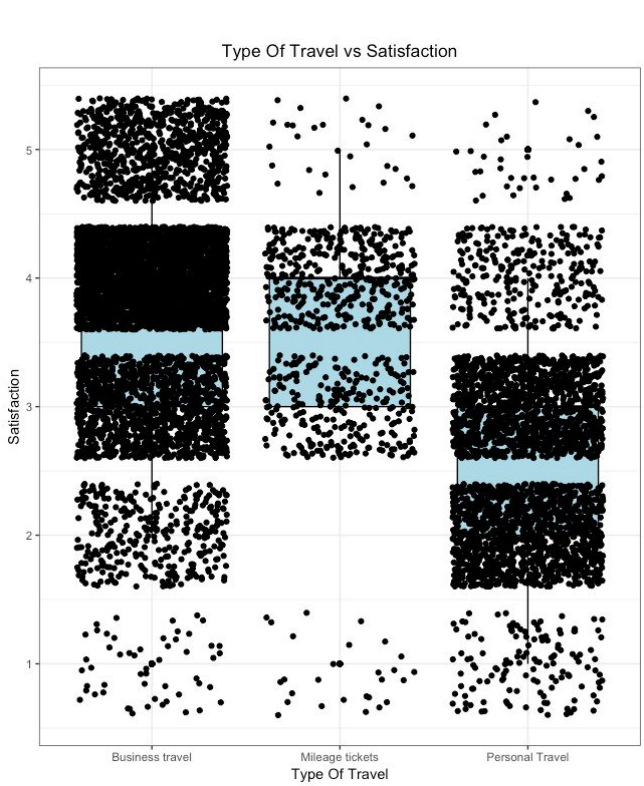
Visualization is one of the most important tasks in data analysis. Being able to visually represent an issue and a solution within a company will allow the technical analyst to touch all employees of a company. Aimed to appeal to everyone from upper management to the most technical employee at Southeast Airlines Co., well developed and thought provoking insights into the happenings and issues that have arisen. Our goal is to visualize our interpretation of the business questions previously discussed with sound graphical analysis.

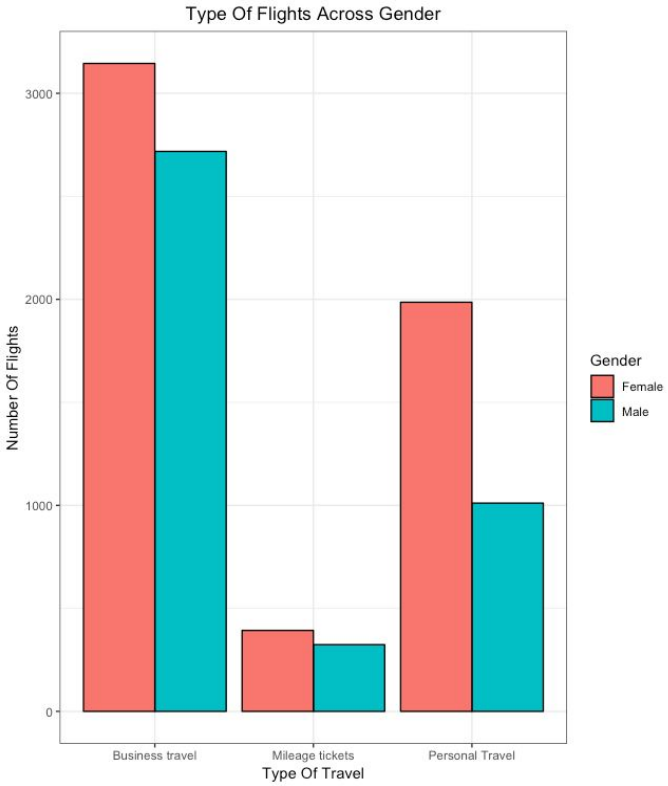
1. *Viewing the Overall Satisfaction within separate Flights and Ages [Graph 1.1]*



This visual representation graphs the different types of travel compared to the averaged overall customer satisfaction. Notice that Business travel has the highest satisfaction with around 3.8, second is the mileage tickets at 3.5, and last would be the personal travel at well below 3. This brings up an apparent question when we decide on where to go next with our analysis; why, out of all the types of travel, is the personal (leisure) travel have the lowest satisfaction?

2. *Finding the frequencies of the Types of Travel [Graph 1.2:1.3]*

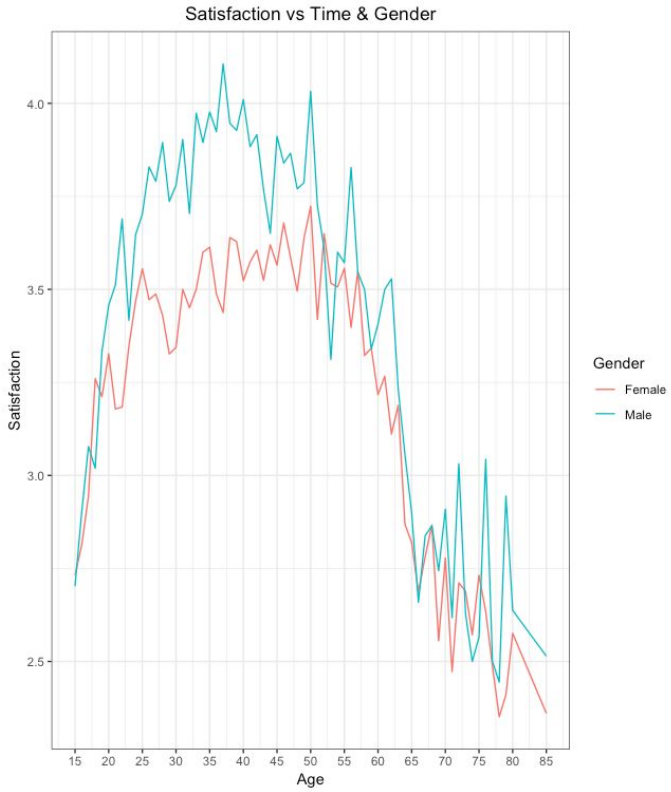


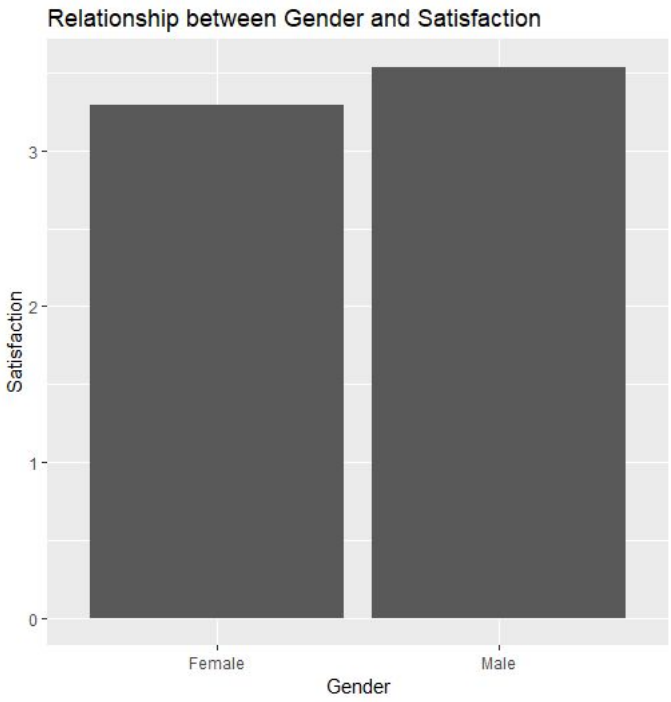


Furthering into our analysis of the overall customer satisfaction within the different types of travel leads us to the boxplot in 'Type of Travel vs Satisfaction' boxplot, by utilizing the geom_jitter() function within our plot, we can see how the frequencies within each type are dispersed. Notice that within 'Business travel', having the highest frequency, tends to be skewed above the 3.5 median. Having a majority of the frequencies above the median in this case will ultimately lead us to a higher mean. transition to the travel type with the lowest mean in graph 1.1, 'Personal Travel' had a fairy even distribution around the median of 2.5, giving us validation of why we see the low average. With the least amount of visual frequencies 'Mileage Tickets' showed a fairly true distribution, around the median of 3.5.

Graph 1.3 is a visual comparing the frequencies of each Type of Travel, with gender taken into account for a deeper understanding of the data. We see that Graph 1.3 is clearly a better visual of how the frequencies of each travel type are allocated, again validating the reasoning behind Graph 1.1. But what is the reasoning for the 'Personal Travel', which has the second highest frequency, have the lowest average satisfaction? Notice the gender frequency for females within 'Personal travel' is double the frequency of males.

3. Gender effects on Overall Satisfaction [Graph 1.4:1.5]



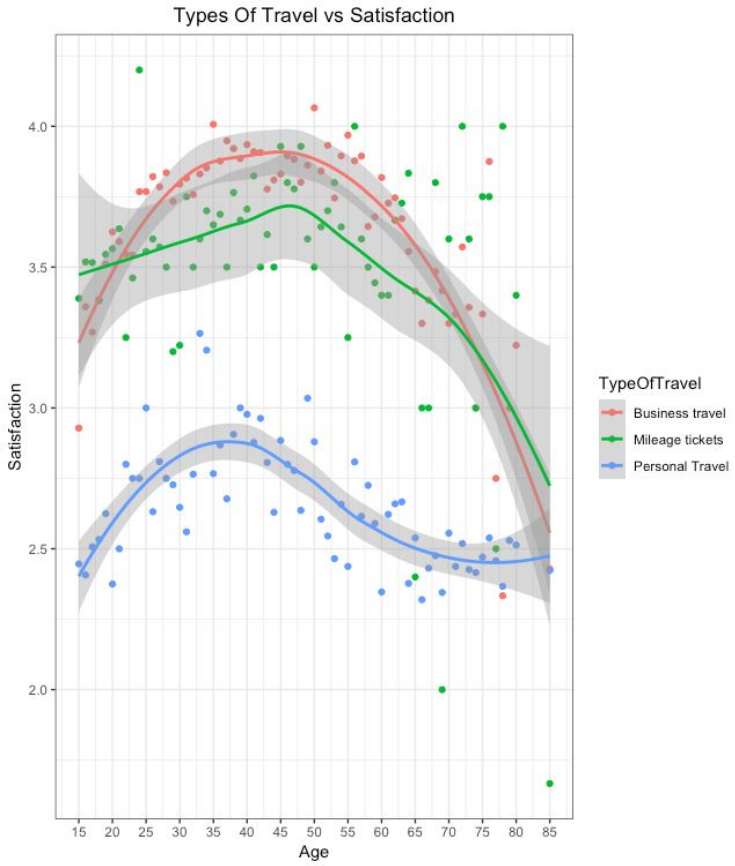


While discussing the impacts of types of travel on overall satisfaction, we notices that there was an unequal dispersion within the frequencies of male and female in personal travel. To take a step back and understand the reasoning behind the disparity, let's take a look at overall averaged satisfactions of males and females across the age groups given.

Two interesting values arise from what is visible within Graph 1.4. Up until the age of around 21, men and women seem to give back the same average satisfaction. From there up until about age 55 do the two lines meet again. Why would that be, we know that women on average have a higher frequency across all categories of travel type. Would it be a causal effect of traveling more? Or an exogenous variable acting on the subset?

Another interesting point to take away from Graph 1.4 would be the downward trend from age 60-70. We notice that the satisfaction drops lower than at any other point in the graph within this group. Does retirement play a role in the the overall flight satisfaction. If does the type of travel change with the age?

4. *Age Impacts across Satisfaction and Type of Travel [Graph 1.6]*

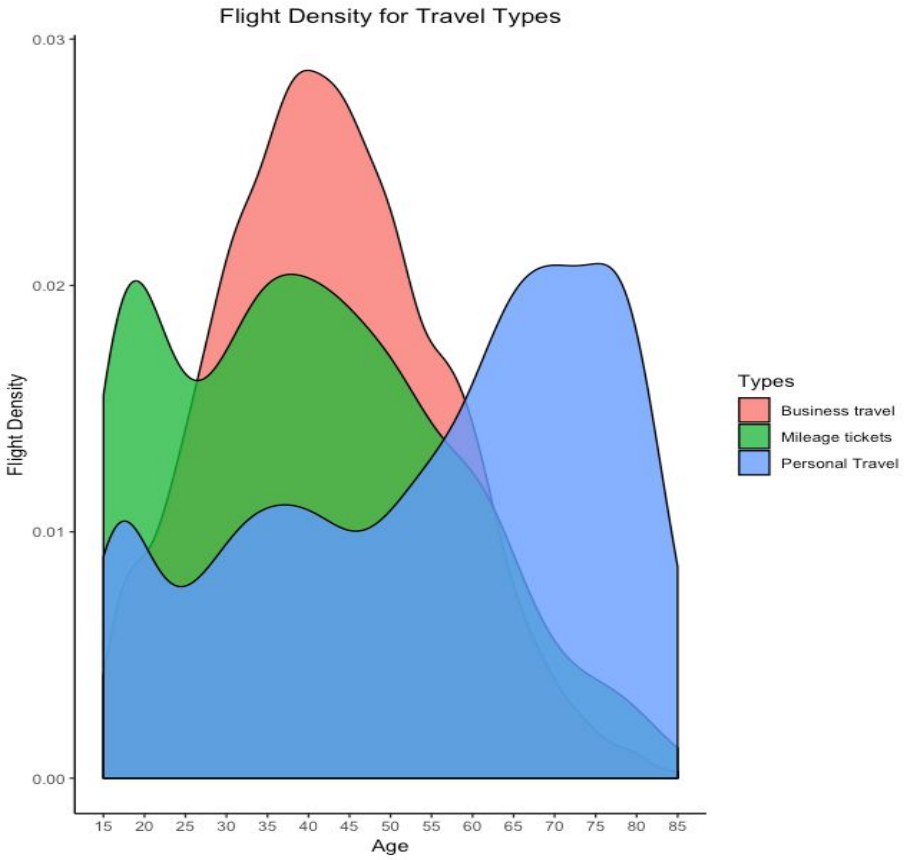


What we are attempting to view within this graph is how each type of travel changes across surveyed ages. For 'Types of Travel vs Satisfaction' we see a scatter plot, coupled with a smoothed average across ages. Visually we notice that as age increases, the overall satisfaction of the lines increases across all types of travel. But once we get passed 50 years of age, Business travel and Mileage tickets both fall exponentially. But the average satisfaction of Personal Travel seems to remain the, around 2.5 at age 85, while Business Travel and Mileage Tickets drop a full point of satisfaction across the same span.

But the smoothed line of Personal travel switched concavity at around 60 years of age. Just as we noticed before, something happens around age of 55-60 that spikes our attention. Graph 1.6 shows that something within Personal Travel will change as age increases.

Giving our insight into the data set, you need to understand the mindset of each flight, business travel is usually stipend while earning points towards other flights. On a business trip, being more of a duty than leisurely activity, amenities are not as big a factor as they would be on a personal travel where someone is paying for their flight.

5. Visualizing the Density of Type Of Travel across Ages [Graph 1.7]



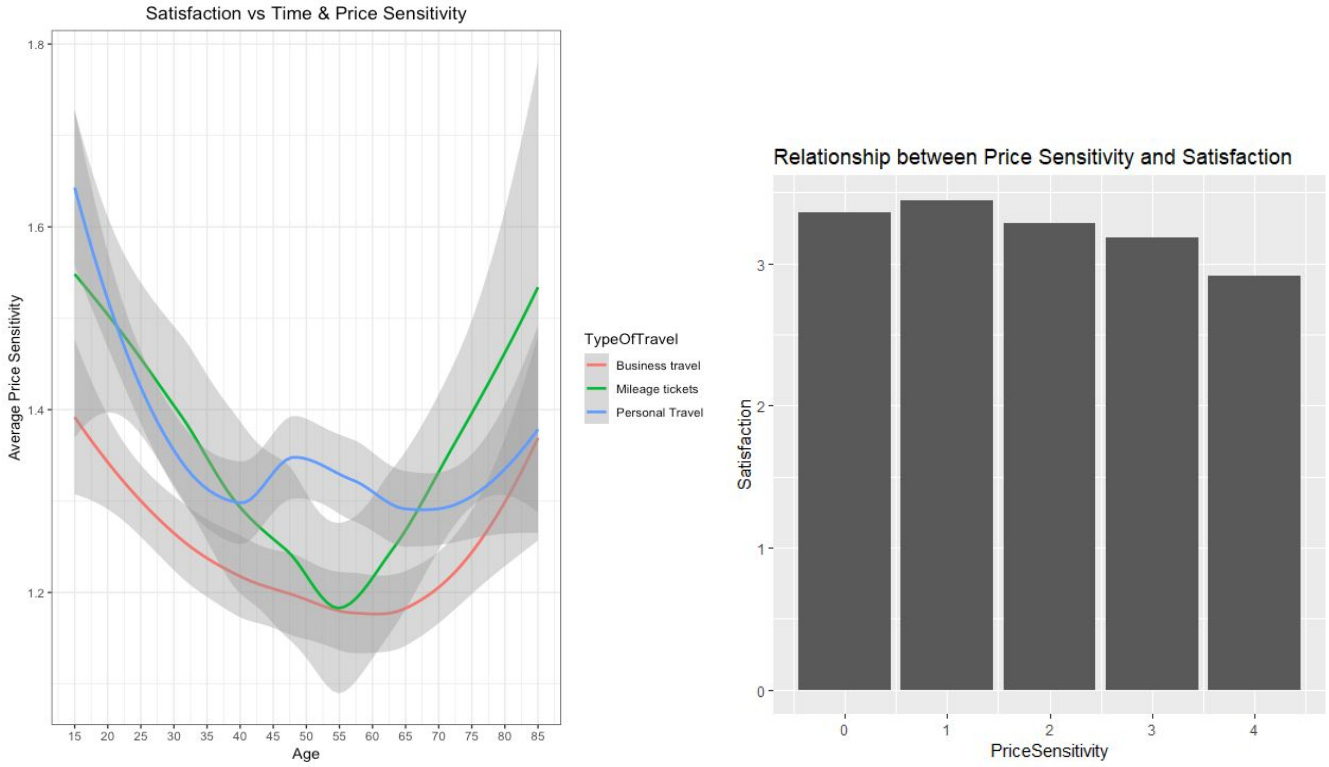
We are left wondering about how age impacted the specific types of travel, we needed to visualize how each type of travel frequency changes with age. Graph 1.7 depicts each type density, across ages 15 through 85.

We can notice the business travel has a parabolic path with a maximum value around the age of 45. Inferring that business travel is paid for by the employees company we will not see any fluctuation within the frequency of business travel. While travelling for business, employees usually earn mileage points which can be put towards mileage tickets.

We see that mileage tickets has an initial peak in between age 15 to 25. How I would translate this initial peak, would be that mileage tickets earned from a parents business travel are being used for their children and family. Families that use mileage tickets earned from business travel for their vacations (personal travel) would explain the spike in personal travel between age 15 to 25.

Based of of the visual in the "Flight Density for Travel Types" we can notice that at age 45 people decrease both Business travel and Mileage tickets, while Personal travel skyrockets until the age of 80 then begins to fall. As people begin to move out of the workforce they do not receive the same benefits from Business travel, therefore the type of travel turns into personal travel as one nears retirement.

6. Price Sensitivity Graph against Age [Graph 1.8:1.9]

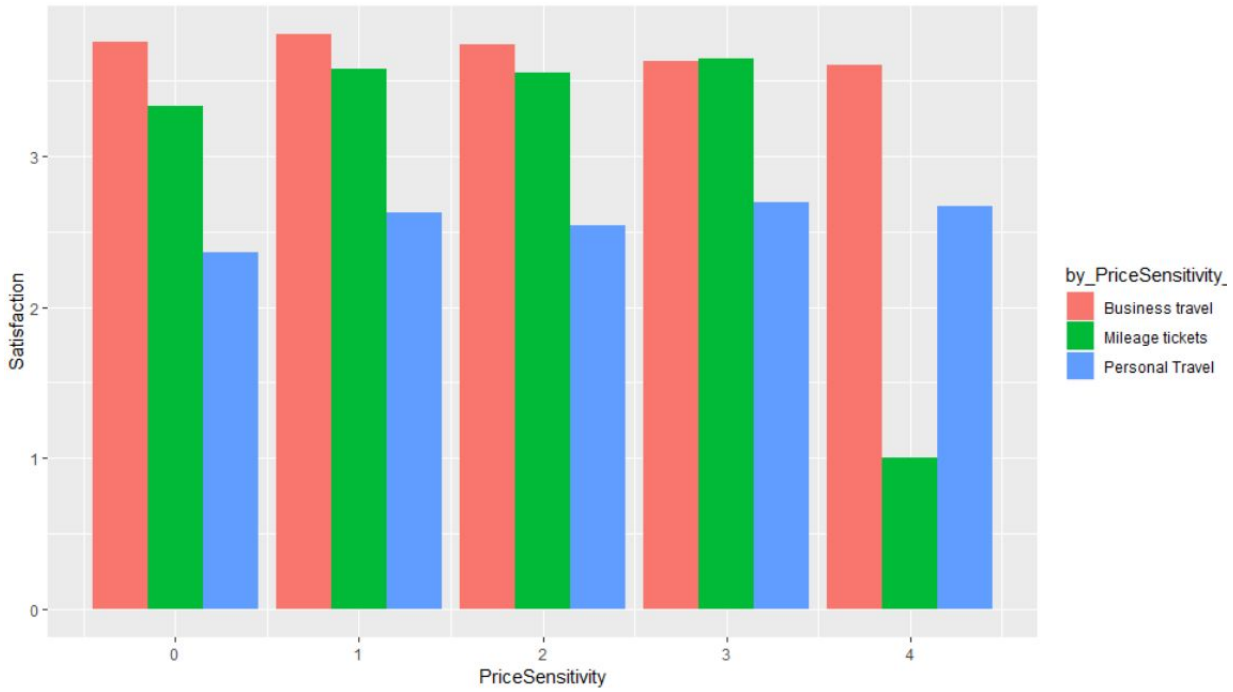


What we noticed in Graph 1.8 was how frequency of types of travel changed throughout the age scale given in the survey. Graphing the averaged price sensitivity across ages will show us how and why people decide on making their decisions for travel. Again the age 55 becomes an interesting value to look at when discussing the graph because it is the critical value of all the types of travel. At age 55, Business travel and mileage tickets reach their lowest price sensitivity. As viewed in graph 1.8, from age 55 to 65 is when we see the frequency of personal travel surpass that of the business travel and mileage tickets.

If we take a closer look at the personal travel. we notice that from age 15 to 40 the price sensitivity drops .3 of a sensitivity point, but from age 65 to 85 it remains unaffected. We can see this as after retirement people remain unaffected by the price of the tickets, unlike earlier 'leisure' travel.

This Visualization helps us understand the relationship between the Price Sensitivity and it's respective mean Satisfaction. As we can see from the plot, Customer Satisfaction decreases as the Price Sensitivity increases.

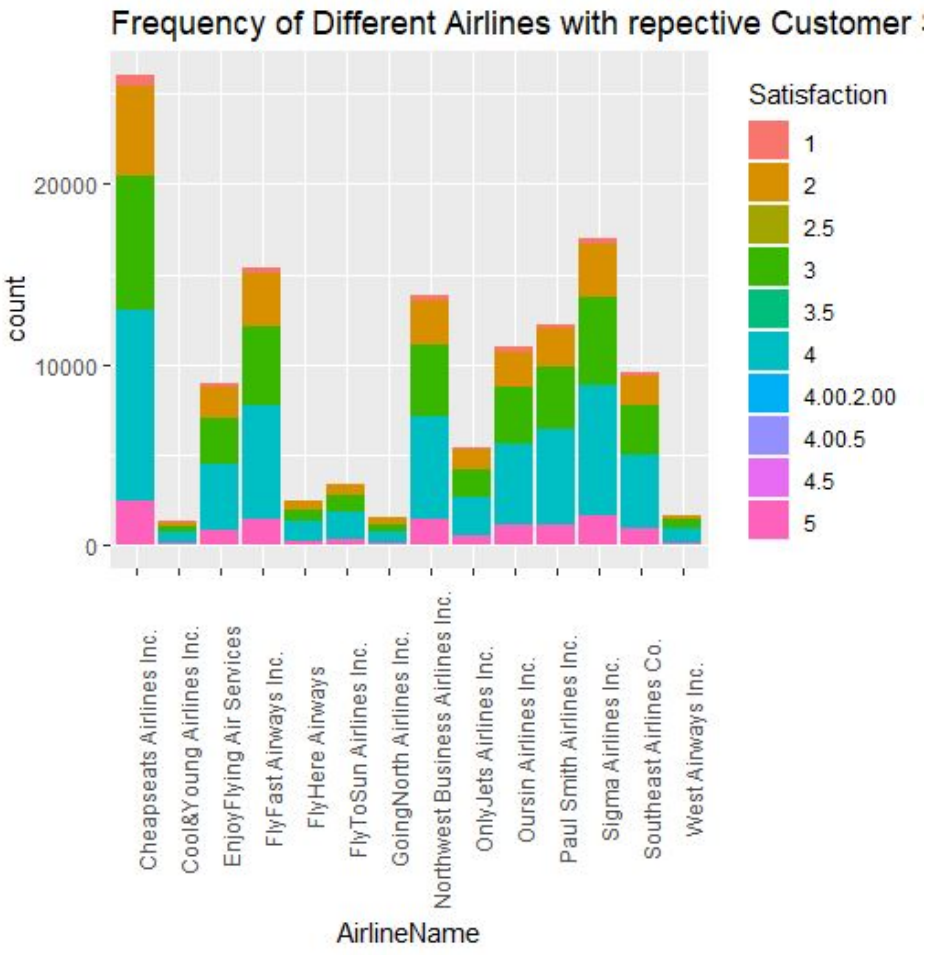
7. *Satisfaction and Price Sensitivity of Types of Travel [Graph 2.0]*



This Visualization describes the relationship between PriceSensitivity, TypeOfTravel and corresponding Satisfaction.

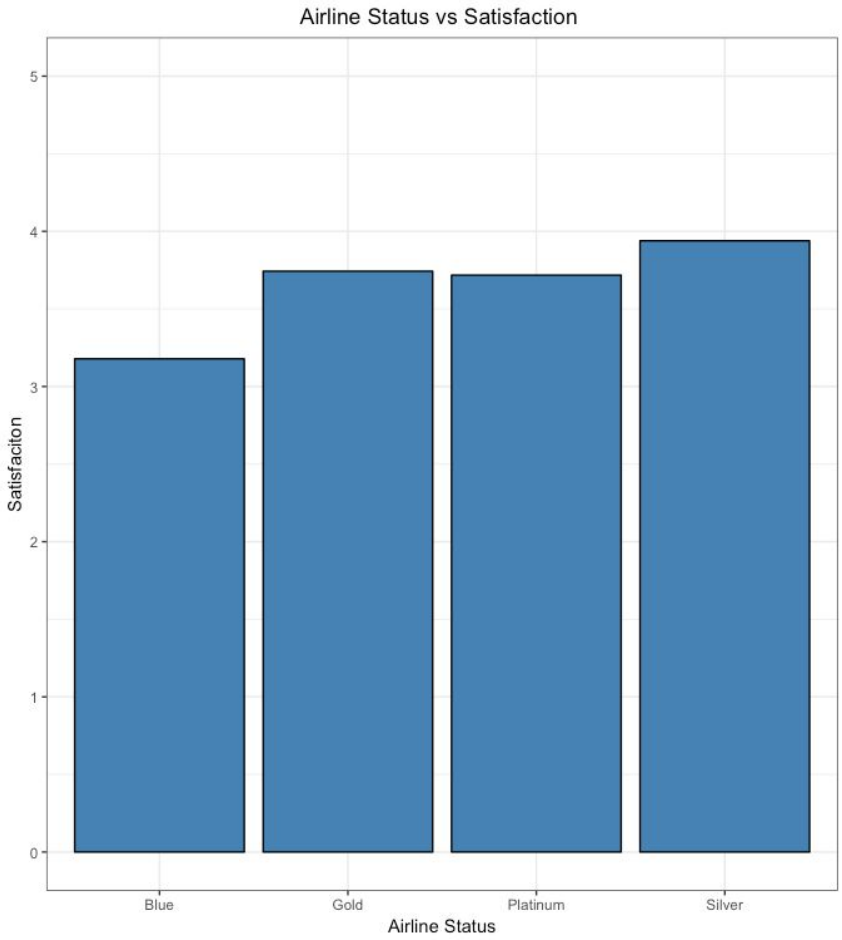
As we can see from the Plot, Price Sensitivity doesn't really affect the Business Travel but in someway or another does affect the other Types of Travel.

8. *Satisfaction Overview of all Airlines [Graph 2.1]*



This Visualization helps us understand how Customer Satisfaction is distributed over all the other Airlines in the data-set along with the frequency.

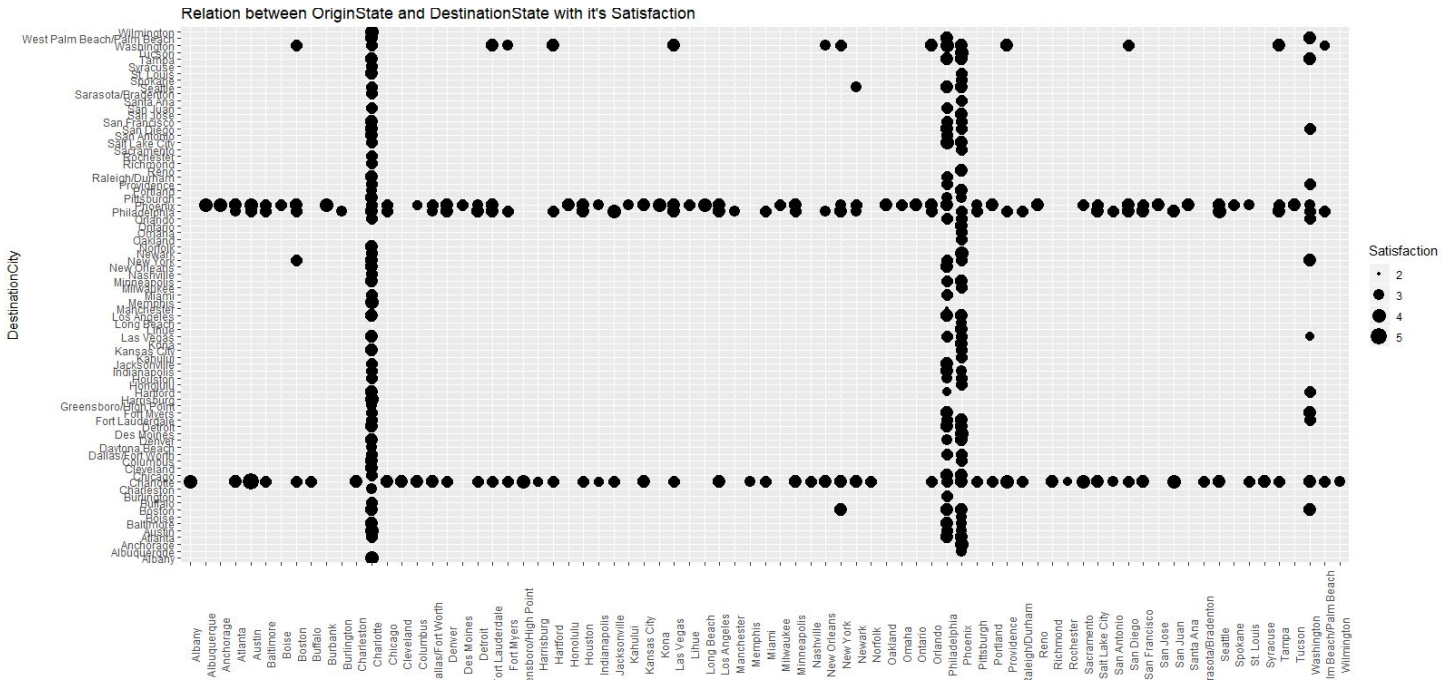
9. Airline Status and Averaged Satisfaction [Graph 2.2]



This Visualization helps us understand the relationship between the Airline Status and Satisfaction.

As we can see from the plot, Airline_Status=Blue has a lower customer Satisfaction when compared to other Airline_Status.

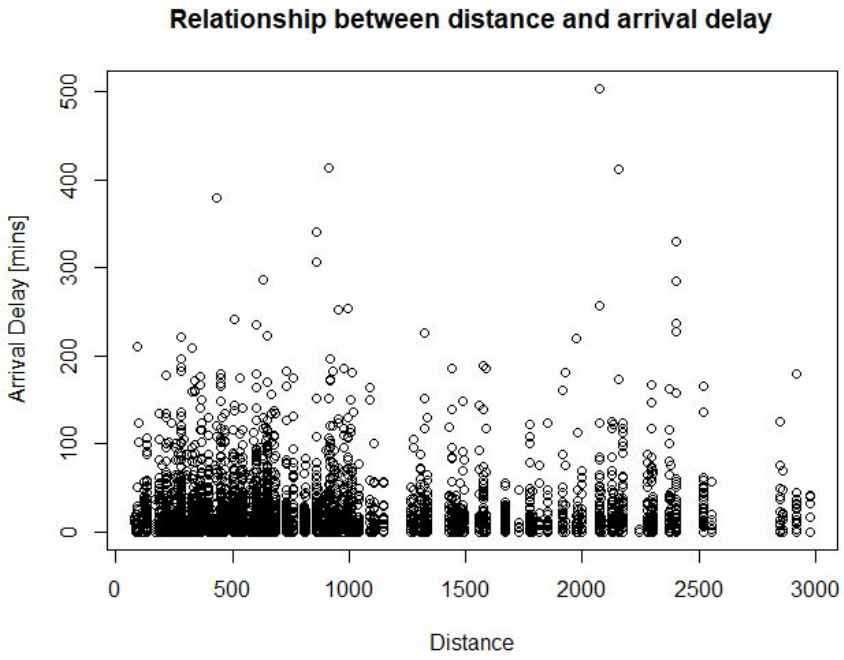
10. Origin and Destination State with Satisfaction [Graph 2.3]



This visualization shows the relationship between the Origin City and the Destination City along with it's mean Satisfaction.

This Visualization could be helpful for quick action. For example, as we can see "Washington" has a lower mean Satisfaction which is an Origin City. We could dive deep into the problem and necessary actions could be taken to solve the problem.

11. Distance against Arrival Delay [Graph 2.4]

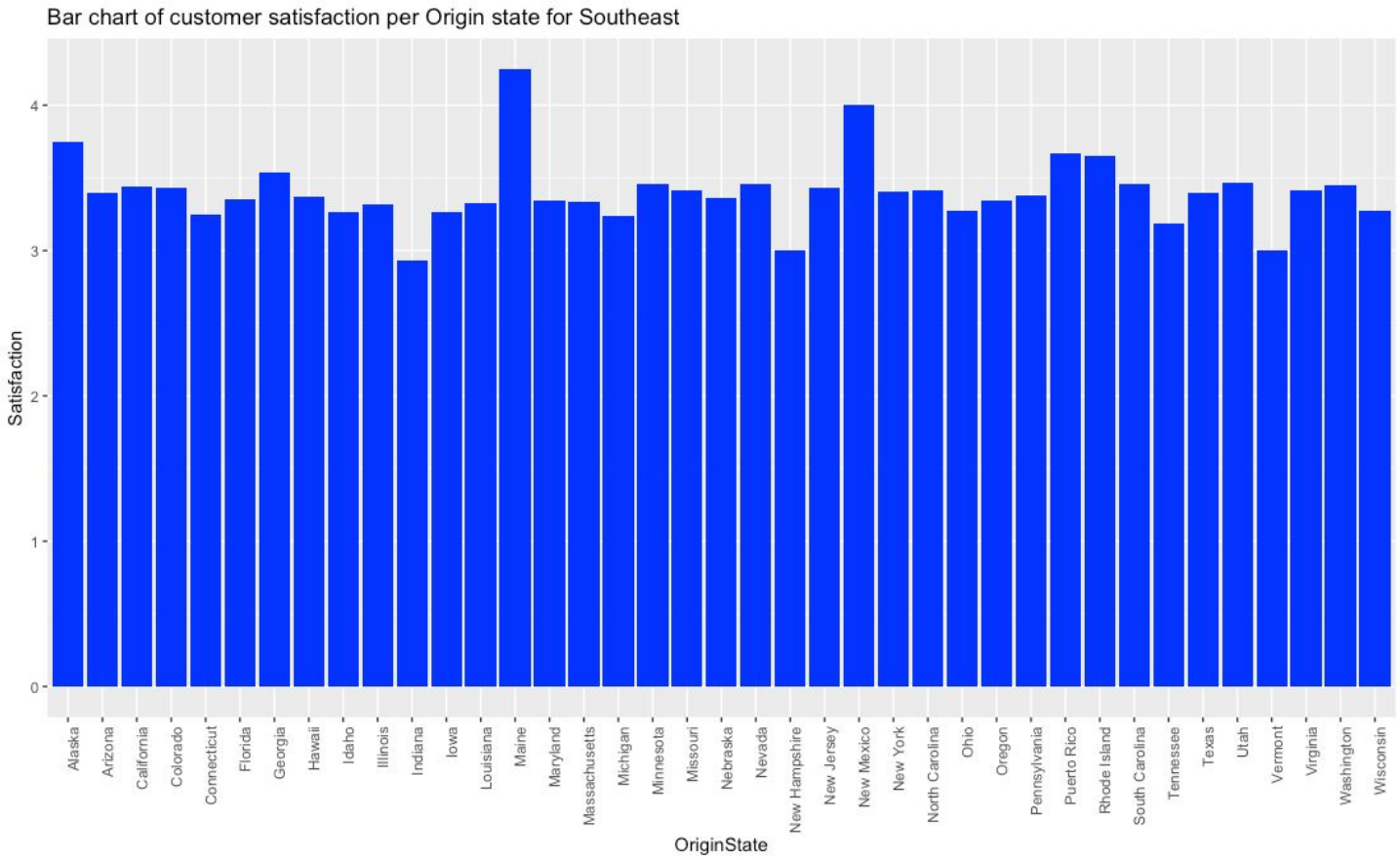


This visualization was created to find out the correlation between the Distance Travelled and the Arrival Delay.

```
Call:
lm(formula = ArrivalDelayInMinutes ~ FlightDistance, data = SouthData)

Residuals:
   Min     1Q Median     3Q    Max
-13.00 -10.37  -9.86  -0.93 491.07

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.468257   0.461029  20.537  < 2e-16 ***
FlightDistance 0.001187   0.000417   2.846  0.00444 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26.68 on 9575 degrees of freedom
Multiple R-squared:  0.000845,  Adjusted R-squared:  0.0007406
F-statistic: 8.097 on 1 and 9575 DF,  p-value: 0.004442
```

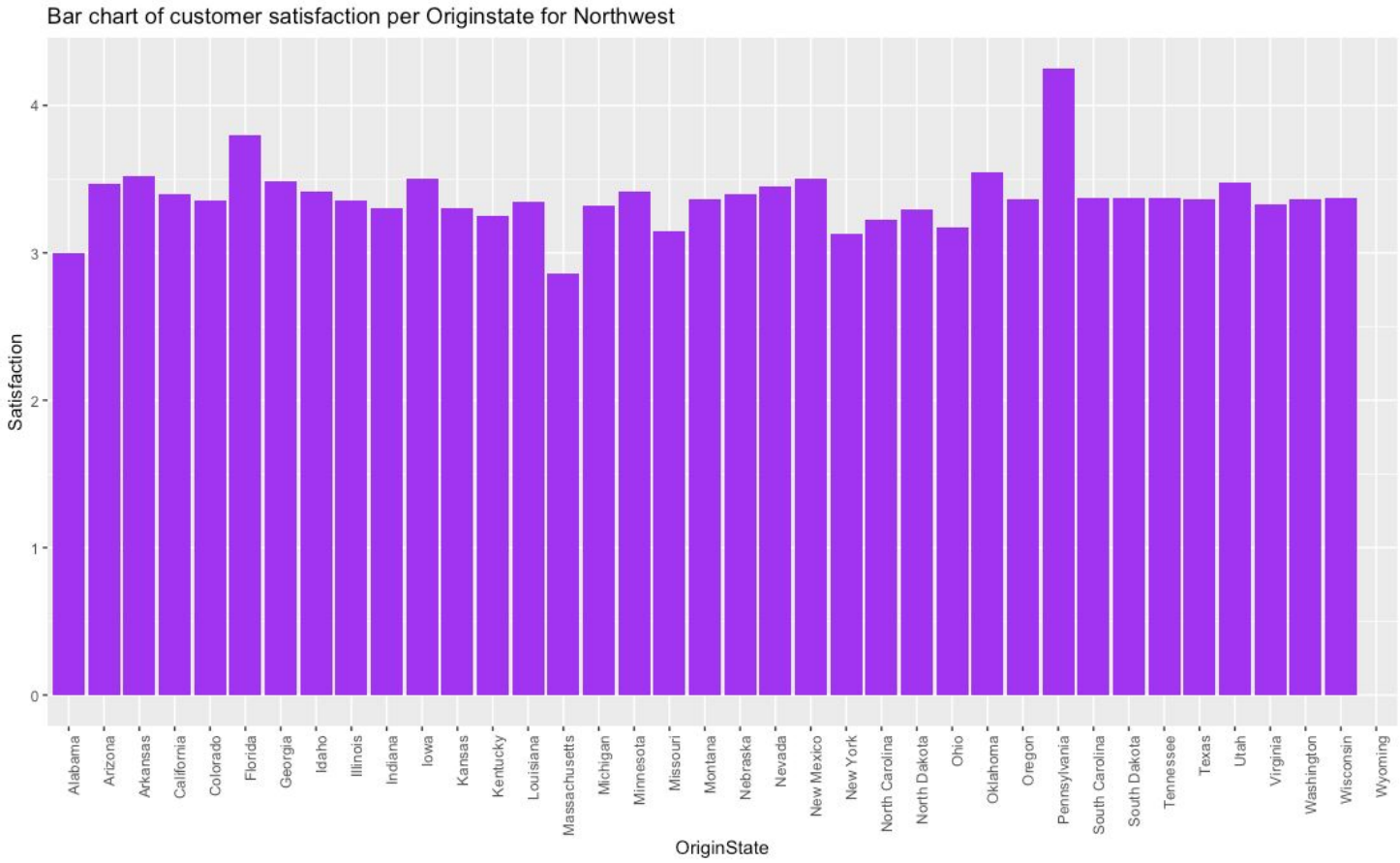
We carried out linear modelling tests for the same. And there was no significant relationship between the two variables.

12. Customer Satisfaction against Origin State [Graph 2.5]



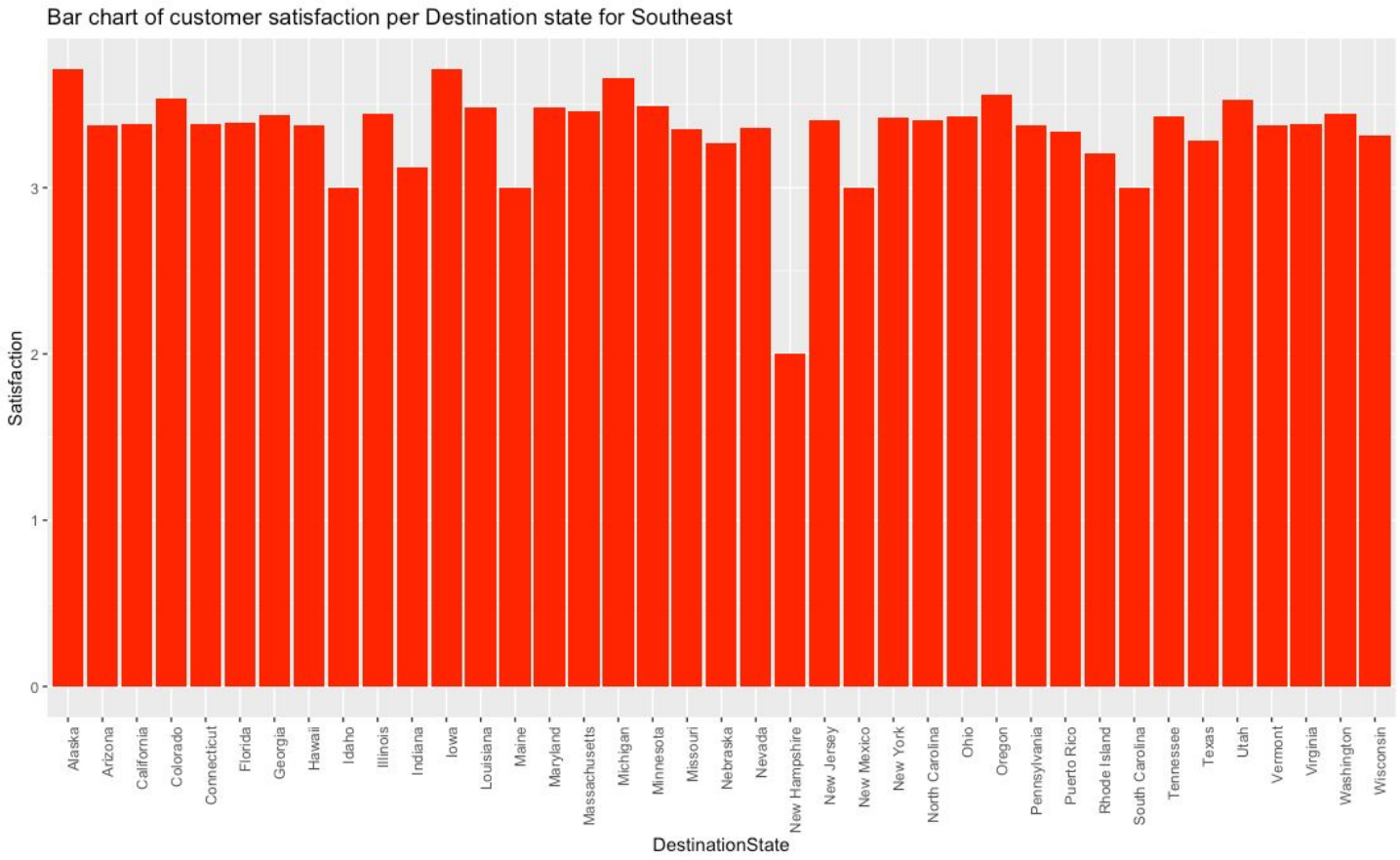
As you can see above, I constructed a Bar chart for the overall customer satisfaction per state for Origin of state. You can see that our Airline, Southeast Airlines Co flies out of 39 states instead of 50 or 51, which means that it doesn't fly out from the rest of the missing states which can be a problem because they are not maximizing profit. If you pay attention to the chart, Maine had the highest overall customer satisfaction (4.250000) and Indiana had the lowest (2.933333) and Vermont and New hampshire being the second lowest tied at (3.0000).

13. Customer Satisfaction against Origin State for Northwest Business Airlines Inc. [Graph 2.6]



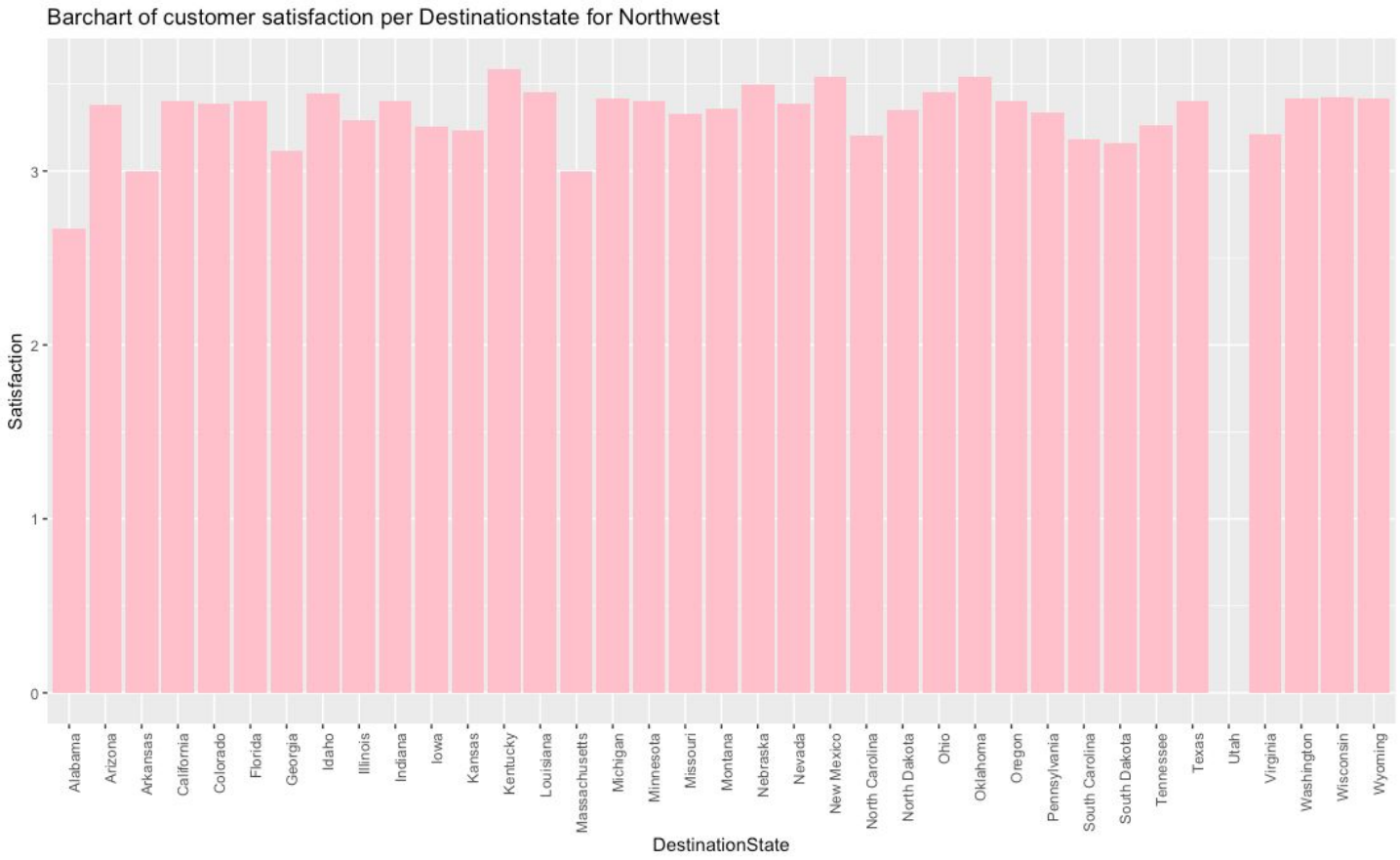
Constructed a different barchart for *Northwest Business Airlines Inc.* to compare it to *Southeast Airlines Co.* You can notice the difference and see that this airline only flies out of 38 states which is less than our airline. on average, there's a higher satisfaction rate compared to Southeast. Pennsylvania has the highest overall satisfaction rating.

14. Customer Satisfaction against Destination State [Graph 2.7]



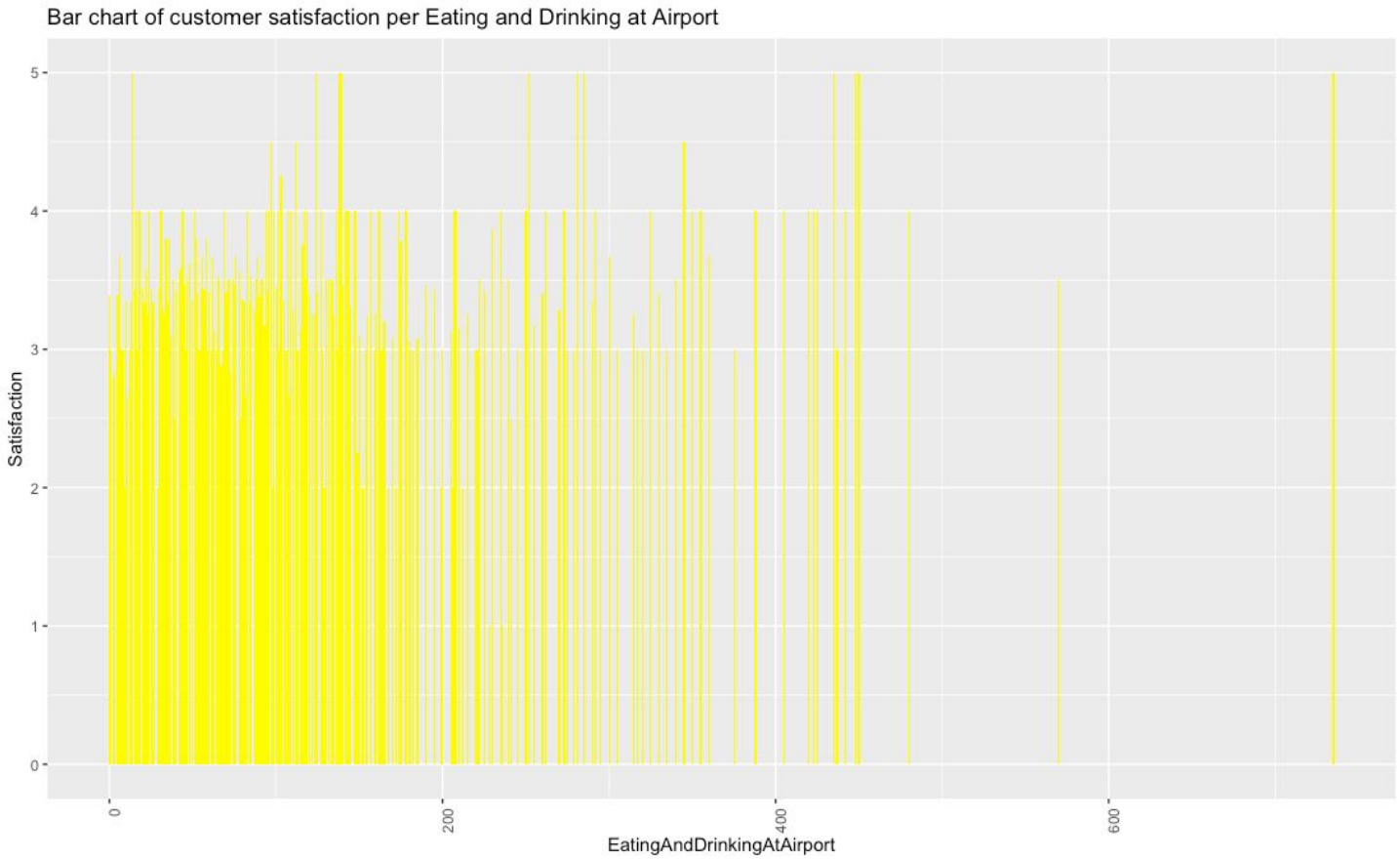
I constructed another Bar chart for the overall customer satisfaction per state for Destination of state. You can see again that Southeast Airlines Co. only flies to 39 states instead of 50 or 51, which means that it doesn't fly to the rest of the missing states which can be another problem because they are not maximizing profit for all the State. If you pay attention to the chart, Alaska has the highest overall customer satisfaction (3.714286) and New hampshire has the lowest (2.000000). South Carolina, New Mexico, Maine, and Idaho are the second lowest tied at (3.0000). From these 2 bar charts you can see that People like flying out of Maine but don't like flying to Maine.

15. Customer Satisfaction against Destination State for Northwest Business Inc. [Graph 2.8]



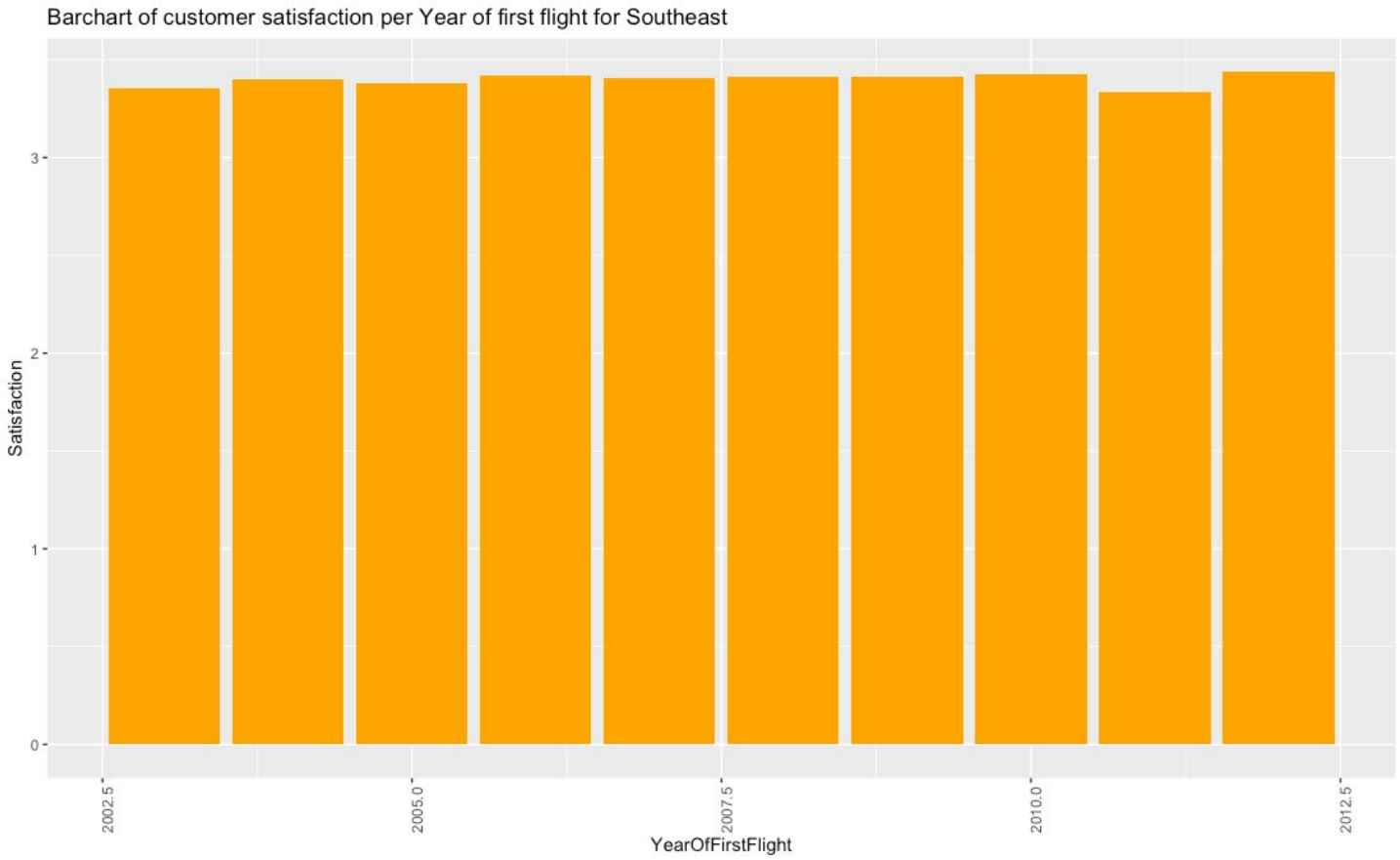
Overall customer satisfaction per state for Destination of state for Northwest Business Airlines Inc. you can see that this airlines only flies to 37 states which less than Southeast Airlines. For Utah there's no Satisfaction. Alabama has the lowest satisfaction rate.

16. Satisfaction against Eating and Drinking at the Airport [Graph 2.9]



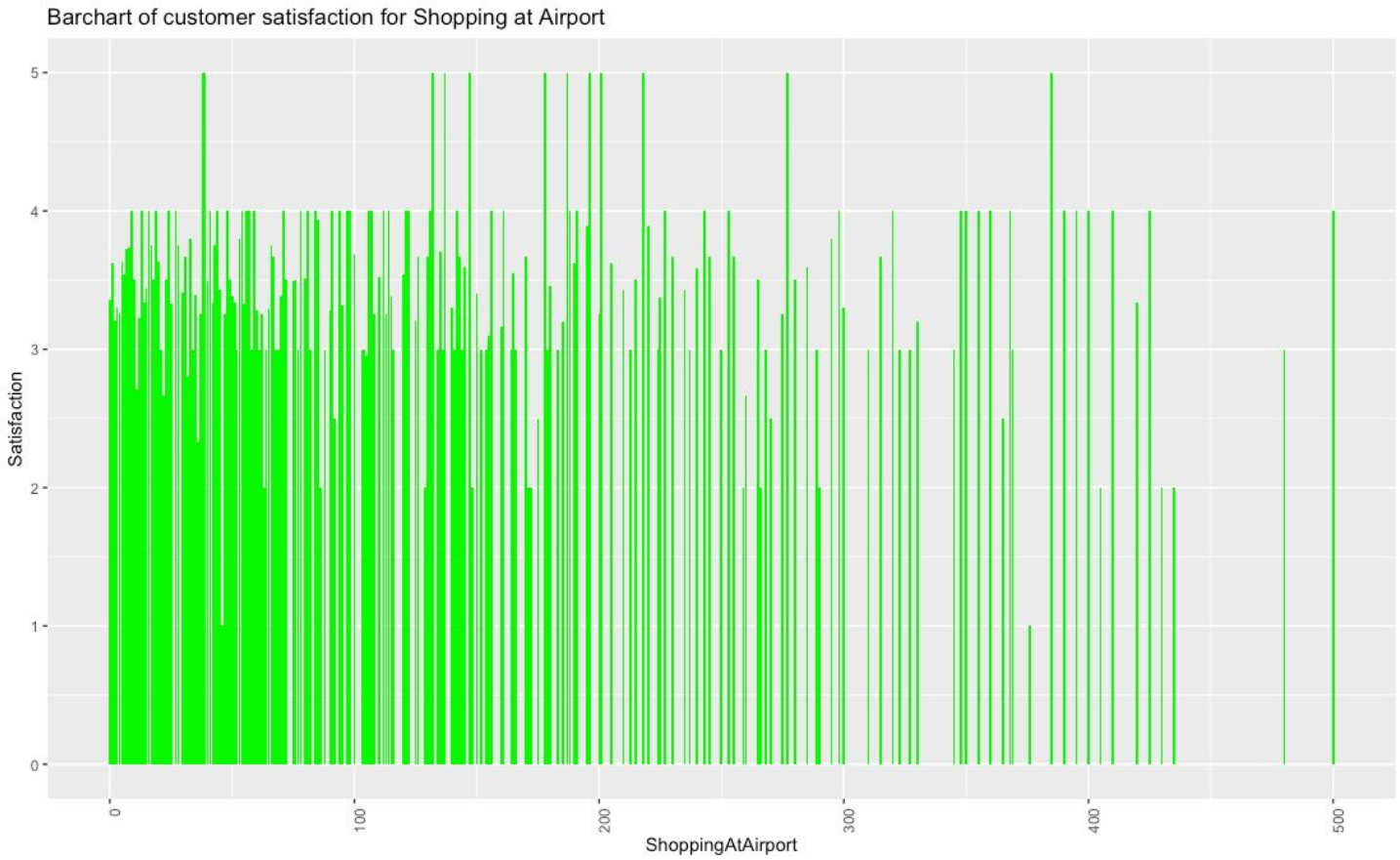
This barchart shows the overall customer satisfaction of eating and drinking at a Airport. The X-axis represents the amount of money people spent. There's a outlier at 735 with a satisfaction of (5.000000) which can mean that someone with a lot of money spent a lot and enjoyed the food at the Airport which shows great customer service. You can say that the data is right skewed which means the mean is greater than the median.

17. Satisfaction and Year of First Flight [Graph 3.0]



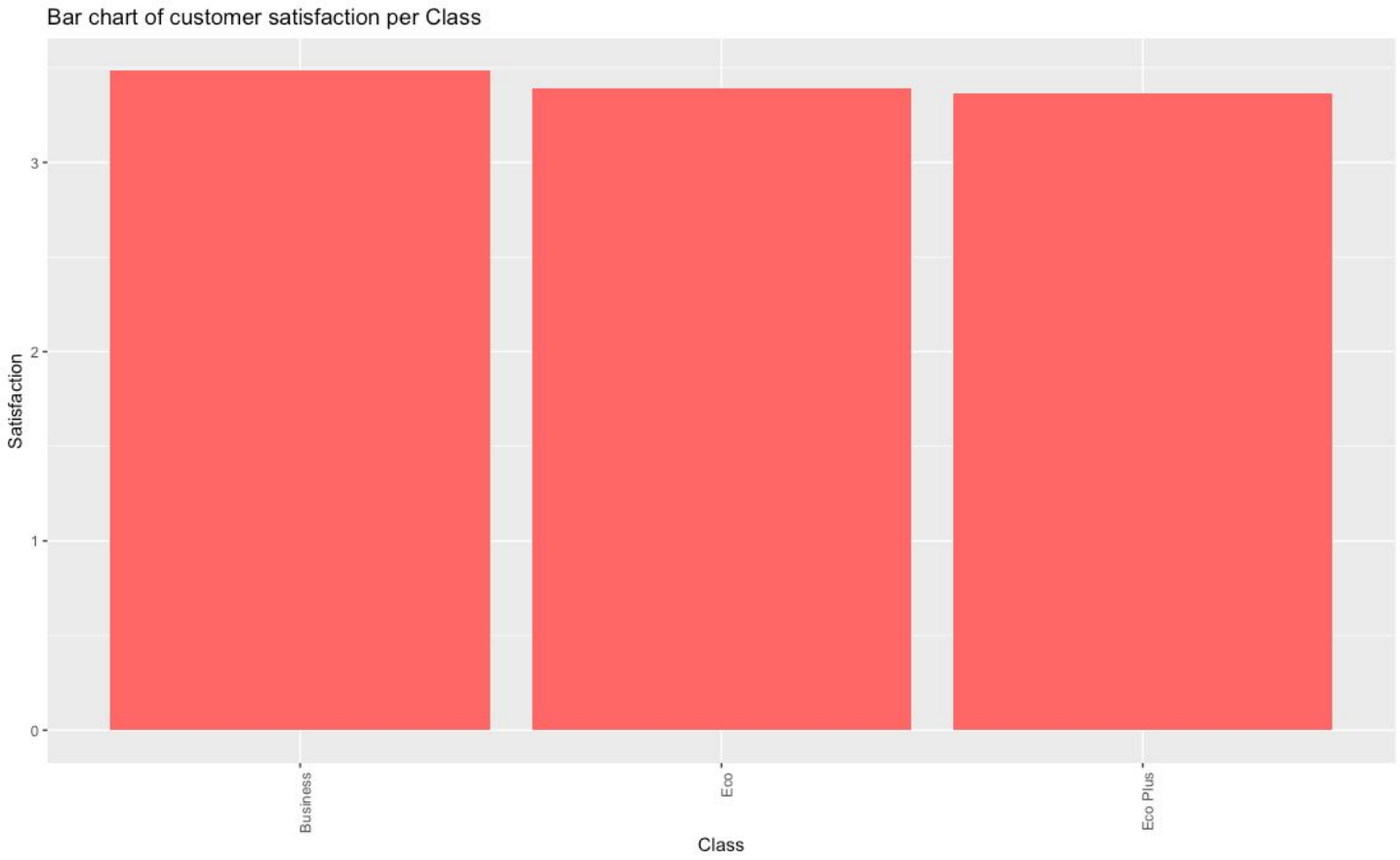
This Barchart was constructed to find out the overall customer Satisfaction per Year of first flight for Southeast. It proved that the most current "first flight" would have the highest satisfaction because it is new and understands what the customers want due to past research.

18. Average Satisfaction Against Shopping at the Airport [Graph 3.1]



Overall customer satisfaction for shopping at the Airport. The X-axis represents the money spent. A lot of people were in the range of $0 - $150. This shows that there are people that enjoy shopping at the Airport and are willing to spend money.

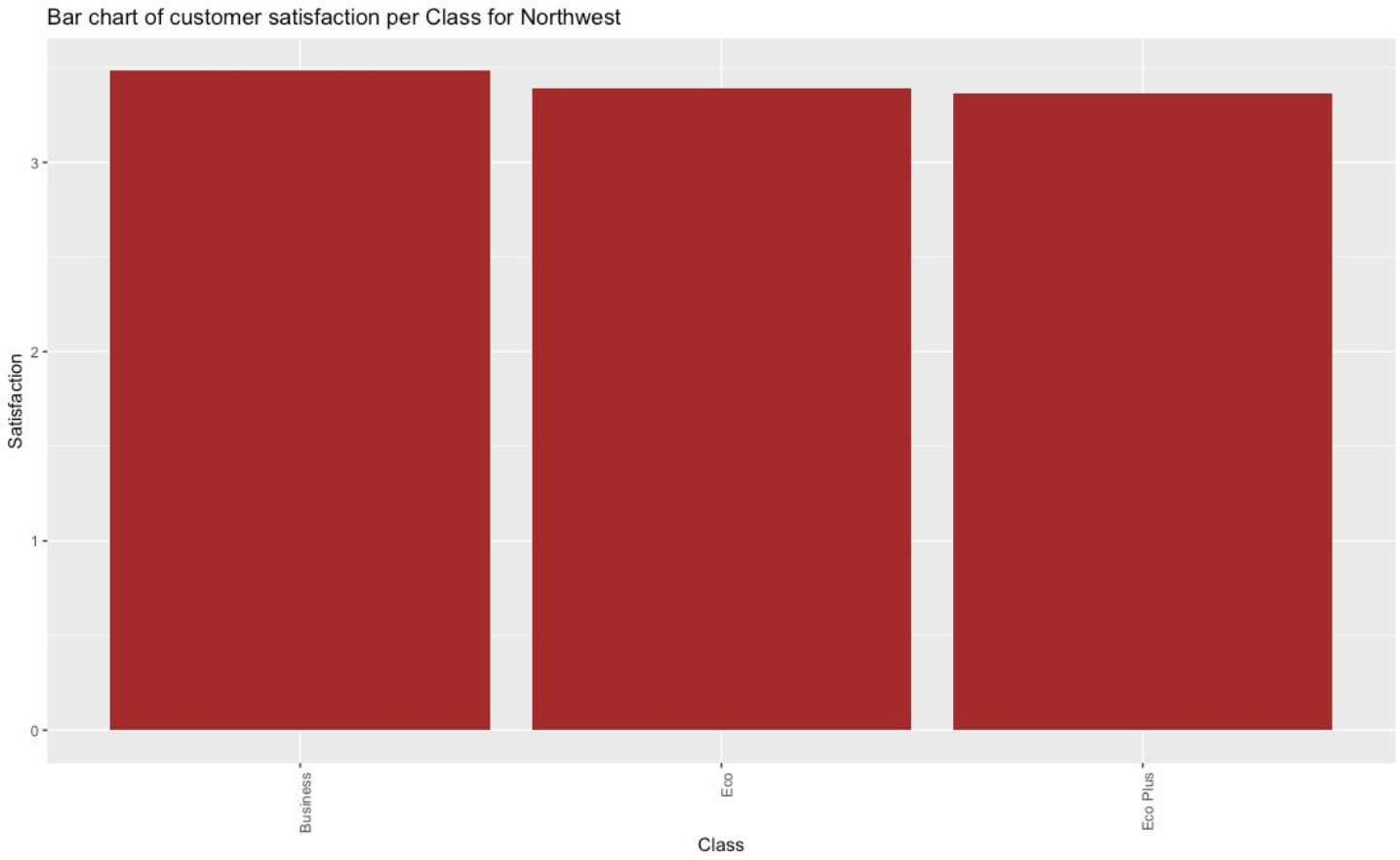
19. Customer Satisfaction per Class in Southeast Airlines [Graph 3.2]



This visual representation graphs the different types of classes, Business, Economy and Economy Plus. Business has the highest Satisfaction maybe due to the fact that you get what you pay for. The difference between the classes isn't great.

	Class	Satisfaction
1.	Business	3.483146
2.	Eco	3.392472
3.	Eco Plus	3.361895

20. Customer Satisfaction per Class in Northeast Airlines [Graph 3.3]



This visual representation graphs the different types of classes for Northwest Airlines; Business, Economy and Economy Plus. It looks very similar to Southeast's classes

	Class	Satisfaction
1.	Business	3.580292
2.	Eco	3.310452
3.	Eco Plus	3.279054

Chapter 5 - USE OF MODELLING TECHNIQUES

- ❏ Apriori Algorithm Model
- ❏ Linear Model
- ❏ Support Vector Machine Model

1) Apriori Algorithm (Association Rules)

Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases.

It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database.

Code Snippet:

Made functions and passed variables in order to later convert them in factors (High, Low, Average)

```
giveQuant <- function(y)
{
  q<- quantile(y,c(0.4,0.6))
  Buckets<-replicate(length(y),"Average")
  Buckets[y<=q[1]] <- "Low"
  Buckets[y>q[2]] <- "High"
  return(Buckets)
}

giveLevel<-function(x)
{
  BucketsClean <- replicate(length(x),"Average")
  BucketsClean[x>=4] <- "High"
  BucketsClean[x<=3.3] <- "Low"
  return (BucketsClean)
}

SouthData <- SouthData[,-14:-17]
SouthData$Age<- giveQuant(SouthData$Age)
SouthData$PriceSensitivity<-giveQuant(SouthData$PriceSensitivity)
SouthData$YearOfFirstFlight<-giveQuant(SouthData$YearOfFirstFlight)
SouthData$FlightsPerYear<-giveQuant(SouthData$FlightsPerYear)
SouthData$FlightsWithOtherAirlines<-giveQuant(SouthData$FlightsWithOtherAirlines)
SouthData$NoOfOtherLoyaltyCards<-giveQuant(SouthData$NoOfOtherLoyaltyCards)
SouthData$ShoppingAtAirport<-giveQuant(SouthData$ShoppingAtAirport)
SouthData$EatingAndDrinkingAtAirport<-giveQuant(SouthData$EatingAndDrinkingAtAirport)
SouthData$FlightDistance<-giveQuant(SouthData$FlightDistance)
SouthData$ArrivalDelayInMinutes<-giveQuant(SouthData$ArrivalDelayInMinutes)
SouthData$FlightTimeInMinutes<-giveQuant(SouthData$FlightTimeInMinutes)
SouthData$DepartureDelayInMinutes<-giveQuant(SouthData$DepartureDelayInMinutes)
SouthData$ScheduledDepartureHour<-giveQuant(SouthData$ScheduledDepartureHour)

SouthData$Satisfaction<- as.numeric(as.character(SouthData$Satisfaction))

SouthData$Satisfaction <- giveLevel(SouthData$Satisfaction)
```

Install necessary packages for Apriori and convert them to factors.

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

install.packages("arulesViz")
library(arulesViz)


str(SouthData)
SouthData$Satisfaction <- as.factor(SouthData$Satisfaction)
SouthData$Age<-as.factor(SouthData$Age)

SouthData$PriceSensitivity <- as.factor(SouthData$PriceSensitivity)
SouthData$YearOfFirstFlight<- as.factor(SouthData$YearOfFirstFlight)
SouthData$FlightsPerYear<-as.factor(SouthData$FlightsPerYear)
SouthData$FlightsWithOtherAirlines<- as.factor(SouthData$FlightsWithOtherAirlines)
SouthData$NoOfOtherLoyaltyCards <- as.factor(SouthData$NoOfOtherLoyaltyCards)
SouthData$ShoppingAtAirport<-as.factor(SouthData$ShoppingAtAirport)
SouthData$EatingAndDrinkingAtAirport<-as.factor(SouthData$EatingAndDrinkingAtAirport)
SouthData$DepartureDelayInMinutes<-as.factor(SouthData$DepartureDelayInMinutes)
SouthData$ArrivalDelayInMinutes<-as.factor(SouthData$ArrivalDelayInMinutes)
SouthData$FlightTimeInMinutes<- as.factor(SouthData$FlightTimeInMinutes)
SouthData$FlightDistance<-as.factor(SouthData$FlightDistance)
SouthData$ScheduledDepartureHour<-as.factor(SouthData$ScheduledDepartureHour)
SouthData$OriginCity<-as.factor(SouthData$OriginCity)
SouthData$DestinationCity<-as.factor(SouthData$DestinationCity)
```

Run the code in order to achieve the Association Rules for Satisfaction=High on RHS.

```
121 ruleset <- apriori(SouthData, list(support = 0.35,confidence = 0.40))
122 inspect(ruleset)
123 ruleSub<- subset(ruleset, subset = rhs %in% "Satisfaction=High")
124 inspect(ruleSub)
125
128:1  (Top Level)
Console  Terminal
~/
```

```
     lhs                                          rhs                      support confidence      lift count
[1]  {}                                        => {Satisfaction=High} 0.5179075  0.5179075 1.0000000  4960
[2]  {TypeofTravel=Business travel}            => {Satisfaction=High} 0.4405346  0.7195975 1.3894325  4219
[3]  {DepartureDelayInMinutes=Low}             => {Satisfaction=High} 0.3586718  0.5448057 1.0519363  3435
[4]  {ArrivalDelayGreaterThan5mins=no}         => {Satisfaction=High} 0.3950089  0.5594499 1.0802120  3783
[5]  {PriceSensitivity=Low}                    => {Satisfaction=High} 0.3848804  0.5423779 1.0472486  3686
[6]  {Class=Eco}                               => {Satisfaction=High} 0.4190247  0.5155447 0.9954378  4013
[7]  {FlightsCancelled=No}                     => {Satisfaction=High} 0.5146706  0.5218634 1.0076383  4929
[8]  {TypeofTravel=Business travel,
      Class=Eco}                               => {Satisfaction=High} 0.3561658  0.7188620 1.3880123  3411
[9]  {TypeofTravel=Business travel,
      FlightsCancelled=No}                     => {Satisfaction=High} 0.4381330  0.7222031 1.3944635  4196
[10] {DepartureDelayInMinutes=Low,
      FlightsCancelled=No}                     => {Satisfaction=High} 0.3585674  0.5448199 1.0519638  3434
[11] {FlightsCancelled=No,
      ArrivalDelayGreaterThan5mins=no}         => {Satisfaction=High} 0.3917720  0.5659125 1.0926904  3752
[12] {PriceSensitivity=Low,
      FlightsCancelled=No}                     => {Satisfaction=High} 0.3823744  0.5464856 1.0551800  3662
[13] {Class=Eco,
      FlightsCancelled=No}                     => {Satisfaction=High} 0.4161011  0.5198956 1.0038388  3985
[14] {TypeofTravel=Business travel,
      Class=Eco,
      FlightsCancelled=No}                     => {Satisfaction=High} 0.3539731  0.7215837 1.3932675  3390
```

Run the code in order to achieve the Association Rules for Satisfaction=Low on RHS.

```
126 ruleset <- apriori(SouthData, list(support = 0.30,confidence = 0.43))
127 inspect(ruleset)
128 ruleSub<- subset(ruleset, subset = rhs %in% "Satisfaction=Low")
129 inspect(ruleSub)
130
131
```

giveLevel(x)

139:1 (Top Level)

Console Terminal

~/

```
> inspect(ruleSub)
     lhs                                                  rhs                  support   confidence
[1]  {}                                                => {Satisfaction=Low} 0.4820925 0.4820925
[2]  {Gender=Female}                                   => {Satisfaction=Low} 0.3075076 0.5331282
[3]  {AirlineStatus=Blue}                              => {Satisfaction=Low} 0.3916675 0.5746017
[4]  {ArrivalDelayGreaterThan5mins=no}                 => {Satisfaction=Low} 0.3110577 0.4405501
[5]  {PriceSensitivity=Low}                            => {Satisfaction=Low} 0.3247363 0.4576221
[6]  {Class=Eco}                                       => {Satisfaction=Low} 0.3937559 0.4844553
[7]  {FlightsCancelled=No}                             => {Satisfaction=Low} 0.4715464 0.4781366
[8]  {Gender=Female,FlightsCancelled=No}               => {Satisfaction=Low} 0.3010337 0.5292822
[9]  {AirlineStatus=Blue,Class=Eco}                    => {Satisfaction=Low} 0.3178448 0.5768429
[10] {AirlineStatus=Blue,FlightsCancelled=No}          => {Satisfaction=Low} 0.3831054 0.5711395
[11] {FlightsCancelled=No,ArrivalDelayGreaterThan5mins=no} => {Satisfaction=Low} 0.3005116 0.4340875
[12] {PriceSensitivity=Low,FlightsCancelled=No}        => {Satisfaction=Low} 0.3173228 0.4535144
[13] {Class=Eco,FlightsCancelled=No}                   => {Satisfaction=Low} 0.3842539 0.4801044
[14] {AirlineStatus=Blue,Class=Eco,FlightsCancelled=No} => {Satisfaction=Low} 0.3101180 0.5730272
     lift      count
[1]  1.0000000 4617
[2]  1.1058628 2945
[3]  1.1918910 3751
[4]  0.9138290 2979
[5]  0.9492413 3110
[6]  1.0049011 3771
[7]  0.9917942 4516
[8]  1.0978851 2883
[9]  1.1965398 3044
[10] 1.1847093 3669
[11] 0.9004236 2878
[12] 0.9407207 3039
```

Meaningful Insights (Apriori Algorithm):

1) **Airline Status** = BLUE → **Satisfaction** = LOW
2) **Gender** = Female → **Satisfaction** = LOW
3) **Type Of Travel** = Business Travel → **Satisfaction** = HIGH
4) **Price Sensitivity** = LOW → **Satisfaction** = HIGH
5) **Departure Delay in Minutes** = LOW → **Satisfaction** = HIGH
6) **Flights Cancelled** = NO → **Satisfaction** = HIGH

2) Linear Model

- Linear models describe a continuous response variable as a function of one or more predictor variables.
- They can help you understand and predict the behavior of complex systems or analyze experimental, financial, and biological data.
- The variable that we're trying to model or predict is known as the dependent variable, and the variables that we use to make predictions are known as independent variables, or covariates

Code Snippet:

```
### Reading the data into R
ProjectData <- read.csv("Documents/IST 687/Satisfaction Survey.csv", header = TRUE)

#display attributes of ProjectData
str(ProjectData)

### Removing "." in ProjectData
names(ProjectData)
names(ProjectData) <- gsub("\\.","", names(ProjectData))

#Display summary of ProjectData
summary(ProjectData)

#display and rename ProjectData columns
colnames(ProjectData)
colnames(ProjectData)[6]<- "YearOfFirstFlight"
colnames(ProjectData)[8]<- "FlightsWithOtherAirlines"
colnames(ProjectData)[7]<- "FlightsPerYear"
colnames(ProjectData)[10]<- "NoOfOtherLoyaltyCards"
colnames(ProjectData)[11]<- "ShoppingAtAirport"
colnames(ProjectData)[12]<- "EatingAndDrinkingAtAirport"
colnames(ProjectData)[14]<- "DayOfMonth"
colnames(ProjectData)[15]<- "FlightDate"
colnames(ProjectData)[25]<- "FlightsCancelled"
colnames(ProjectData)[26]<- "FlightTimeInMinutes"
colnames(ProjectData)[28]<- "ArrivalDelayGreaterThan5mins"
colnames(ProjectData)[23]<- "DepartureDelayInMinutes"
colnames(ProjectData)[24]<- "ArrivalDelayInMinutes"
colnames(ProjectData)[18]<- "OriginCity"
ProjectData$DestinationCity<- gsub("(.*),.*", "\\1",ProjectData$DestinationCity)
ProjectData$OriginCity<- gsub("(.*),.*", "\\1",ProjectData$OriginCity)
```

```
#Convert categorical data into dummy variables
ProjectData$statusBlue <- ifelse(ProjectData$AirlineStatus == "Blue",1,0)
ProjectData$statusSilver <- ifelse(ProjectData$AirlineStatus == "Silver",1,0)
ProjectData$statusPlatinum<- ifelse(ProjectData$AirlineStatus == "Platinum",1,0)
ProjectData$statusGold <- ifelse(ProjectData$AirlineStatus == "Gold",1,0)
ProjectData$genderMale <- ifelse(ProjectData$Gender == "Male", 1, 0)
ProjectData$businessTravel <- ifelse(ProjectData$TypeofTravel == "Business travel", 1, 0)
ProjectData$mileageTickets <- ifelse(ProjectData$TypeofTravel == "Mileage tickets", 1, 0)
ProjectData$personalTravel <- ifelse(ProjectData$TypeofTravel == "Personal Travel", 1, 0)
ProjectData$businessClass <- ifelse(ProjectData$Class == "Business", 1, 0)
ProjectData$ecoClass <- ifelse(ProjectData$Class == "Eco", 1, 0)
ProjectData$ecoplusClass <- ifelse(ProjectData$Class == "Eco Plus", 1, 0)
ProjectData$flightsCancelled <- ifelse(ProjectData$FlightsCancelled == "Yes", 1, 0)
ProjectData$arrivalDelayGreaterThan5mins <- ifelse(ProjectData$ArrivalDelayGreaterThan5mins == "yes", 1, 0)

#create new data set containing only data for the Southeast Airline
SouthData <- ProjectData[which(ProjectData$AirlineName == "Southeast Airlines Co. " ),] #For Seperate DataFrame
#display attributes of new dataset
str(SouthData)

#calculate number of nulls in the data set
sum(is.na(SouthData$ArrivalDelayInMinutes)) #154
sum(is.na(SouthData$FlightTimeInMinutes)) #154
sum(is.na(SouthData$DepartureDelayInMinutes)) #129

#replace null values with variable mean value
SouthData$ArrivalDelayInMinutes[is.na(SouthData$ArrivalDelayInMinutes)] <- round(mean(SouthData$ArrivalDelayInMinutes, na.rm = TRUE))
SouthData$FlightTimeInMinutes[is.na(SouthData$FlightTimeInMinutes)] <- round(mean(SouthData$FlightTimeInMinutes, na.rm = TRUE))
SouthData$DepartureDelayInMinutes[is.na(SouthData$DepartureDelayInMinutes)] <- round(mean(SouthData$DepartureDelayInMinutes, na.rm = TRUE))

#convert data to numeric variables
SouthData$Satisfaction <- as.numeric(as.character(SouthData$Satisfaction))
SouthData$OriginState <- as.numeric(SouthData$OriginState)
SouthData$DestinationState <- as.numeric(SouthData$DestinationState)


#remove unneeded data columns
SouthData <- SouthData[,-15:-17]

View(SouthData)

#create and examine linear model using all variables
model1 <- lm(formula=Satisfaction ~ Age + genderMale + PriceSensitivity + YearOfFirstFlight + FlightsPerYear + FlightsWithOtherAirlines
             + NoOfOtherLoyaltyCards + ShoppingAtAirport + EatingAndDrinkingAtAirport + DayOfMonth + ScheduledDepartureHour + DepartureDelayInMinutes
             + ArrivalDelayInMinutes + FlightTimeInMinutes + FlightDistance + statusSilver + statusPlatinum + statusGold + mileageTickets + personalTravel
             + businessClass + ecoplusClass + flightsCancelled + arrivalDelayGreaterThan5mins , data= SouthData)
#display summary of model
summary(model1)

#refine model and remove variables that do not contribute to the overall strength of the model
summary(lm(formula=Satisfaction ~ Age + genderMale + PriceSensitivity + YearOfFirstFlight + FlightsPerYear +
             EatingAndDrinkingAtAirport + FlightsWithOtherAirlines + ScheduledDepartureHour +
             statusSilver + statusPlatinum + statusGold + mileageTickets +
             personalTravel + businessClass + flightsCancelled + arrivalDelayGreaterThan5mins, data= SouthData))

#create subset including all the variables in the linear model
lmSouthData <- subset(SouthData, select=c("Satisfaction", "Age", "genderMale", "PriceSensitivity", "YearOfFirstFlight", "FlightsPerYear",
                              "EatingAndDrinkingAtAirport", "FlightsWithOtherAirlines", "ScheduledDepartureHour",
                              "statusSilver", "statusGold", "statusPlatinum", "mileageTickets",
                              "personalTravel", "businessClass", "flightsCancelled", "arrivalDelayGreaterThan5mins"))
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.1160 -0.4218  0.1185  0.4499  2.7409

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -7.681e+00  4.934e+00  -1.557 0.119574
Age                           -2.792e-03  5.141e-04  -5.432 5.71e-08 ***
genderMale                     1.314e-01  1.531e-02   8.582  < 2e-16 ***
PriceSensitivity              -1.562e-02  1.372e-02  -1.138 0.255153
YearOfFirstFlight              5.696e-03  2.459e-03   2.317 0.020526 *
FlightsPerYear                -3.263e-03  5.628e-04  -5.798 6.91e-09 ***
FlightsWithOtherAirlines       5.751e-04  9.649e-04   0.596 0.551130
NoOfOtherLoyaltyCards          3.364e-03  7.807e-03   0.431 0.666587
ShoppingAtAirport              8.522e-05  1.406e-04   0.606 0.544447
EatingAndDrinkingAtAirport    -4.048e-04  1.430e-04  -2.832 0.004641 **
DayOfMonth                     1.431e-03  8.404e-04   1.703 0.088672 .
ScheduledDepartureHour         4.871e-03  1.504e-03   3.239 0.001205 **
DepartureDelayInMinutes        1.147e-03  7.423e-04   1.546 0.122230
ArrivalDelayInMinutes         -8.031e-04  7.629e-04  -1.053 0.292502
FlightTimeInMinutes           -4.056e-04  3.796e-04  -1.069 0.285257
FlightDistance                 5.199e-05  4.476e-05   1.162 0.245449
statusSilver                   6.029e-01  1.896e-02  31.803  < 2e-16 ***
statusPlatinum                 2.976e-01  4.204e-02   7.080 1.55e-12 ***
statusGold                     4.230e-01  2.715e-02  15.580  < 2e-16 ***
mileageTickets                -1.316e-01  2.864e-02  -4.595 4.38e-06 ***
personalTravel                -1.026e+00  1.812e-02 -56.591  < 2e-16 ***
businessClass                  7.973e-02  2.660e-02   2.997 0.002732 **
ecoplusClass                  -3.420e-03  2.454e-02  -0.139 0.889173
flightsCancelled              -2.379e-01  6.344e-02  -3.750 0.000178 ***
arrivalDelayGreaterThan5mins -3.463e-01  2.036e-02 -17.006  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7157 on 9552 degrees of freedom
Multiple R-squared:  0.4365,    Adjusted R-squared:  0.4351
F-statistic: 308.3 on 24 and 9552 DF,  p-value: < 2.2e-16
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.1043 -0.4202  0.1168  0.4477  2.7215

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -7.4625068  4.9288831  -1.514 0.130050
Age                           -0.0028648  0.0004758  -6.021 1.80e-09 ***
genderMale                     0.1305822  0.0151205   8.636  < 2e-16 ***
PriceSensitivity              -0.0154257  0.0136960  -1.126 0.260070
YearOfFirstFlight              0.0055999  0.0024560   2.280 0.022625 *
FlightsPerYear                -0.0032942  0.0005585  -5.898 3.80e-09 ***
EatingAndDrinkingAtAirport    -0.0003979  0.0001428  -2.787 0.005338 **
FlightsWithOtherAirlines       0.0007031  0.0008990   0.782 0.434200
ScheduledDepartureHour         0.0051191  0.0014979   3.418 0.000634 ***
statusSilver                   0.6030942  0.0189369  31.847  < 2e-16 ***
statusPlatinum                 0.2960908  0.0420220   7.046 1.97e-12 ***
statusGold                     0.4247251  0.0271269  15.657  < 2e-16 ***
mileageTickets                -0.1317763  0.0286264  -4.603 4.21e-06 ***
personalTravel                -1.0249720  0.0180831 -56.681  < 2e-16 ***
businessClass                  0.0804198  0.0264634   3.039 0.002381 **
flightsCancelled              -0.2467738  0.0630666  -3.913 9.18e-05 ***
arrivalDelayGreaterThan5mins -0.3486045  0.0161264 -21.617  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7157 on 9560 degrees of freedom
Multiple R-squared:  0.436,     Adjusted R-squared:  0.4351
F-statistic: 461.9 on 16 and 9560 DF,  p-value: < 2.2e-16
```

Meaningful Insights:

Positive Impact on Satisfaction:

- Gender
- Year of First Flight
- Scheduled Departure Hour
- Airline Status - Silver, Gold, Platinum
- Traveling Business Class

Negative Impact on Satisfaction:

- Age
- Number of Flights per Year
- Eating and Drinking at Airport
- Type of Travel - mileage
- Flight Cancelled
- Arrival Delay Greater Than 5 Minutes

3) Support Vector Machines

- A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane.
- In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.
- In two dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.

```
SouthData$Satisfaction <- ifelse(SouthData$Satisfaction > 3, "Satisfied", "Not Satisfied")

#generate randomized index to choose cases for the training and test sets
randIndex <- sample(1:dim(SouthData)[1])

#use summary and length command to check the length of index list is the same as the number of rows in the SouthData dataset
summary(randIndex)
length(randIndex)

#check indices are randomized
head(randIndex)

#calculate cut point that would divide the SouthData data into a two-thirds training set and one-third test set
cutPoint2_3 <- floor(2*dim(SouthData)[1]/3)
#display cut point
cutPoint2_3

#create training data set using cutPoint2_3
trainData <- SouthData[randIndex[1:cutPoint2_3],]

#create test data set
testData <- SouthData[randIndex[(cutPoint2_3+1):dim(SouthData)[1]],]

#use dim() function to demostrate the test and training data sets contain the appropriate number of cases
dim(trainData)
dim(testData)
```

Code Snippet :

```
#add kernlab package to library
library(kernlab)

#build support vector model using training data set
svmOutput <- ksvm(Satisfaction ~ Age + genderMale + PriceSensitivity + YearOfFirstFlight + FlightsPerYear +
                    EatingAndDrinkingAtAirport + FlightsWithOtherAirlines + ScheduledDepartureHour +
                    statusSilver + statusPlatinum + statusGold + mileageTickets +
                    personalTravel + businessClass + flightsCancelled + arrivalDelayGreaterThan5mins, data=trainData, kernel="rbfdot", kpar="automatic",
                  C=5, cross=3, prob.model=TRUE)
svmOutput

#validate model against test data using the predict() function
svmPred <- predict(svmOutput, testData, type = "votes")


#review contents of svmPred using str() command
str(svmPred)

#create a confusion matrix (2x2 table) that compares the second row of svmPred to the contents of SouthData$Satisfaction variable
compTable <-data.frame(testData$Satisfaction, svmPred[2,])
results <- table(compTable)

#calculate an error rate based on the confusion matrix
((results[1,2]+results[2,1])/length(SouthData$Satisfaction))
```

```
> #use summary and length command to check the length of index list is the same as the number of rows in the SouthData dataset
> summary(randIndex)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      1    2395    4789    4789    7183    9577
> length(randIndex)
[1] 9577
>
> #check indices are randomized
> head(randIndex)
[1] 9191  782 3222 1383  979 7267
>
> #calculate cut point that would divide the SouthData data into a two-thirds training set and one-third test set
> cutPoint2_3 <- floor(2*dim(SouthData)[1]/3)
> #display cut point
> cutPoint2_3
[1] 6384
>
> #create training data set using cutPoint2_3
> trainData <- SouthData[randIndex[1:cutPoint2_3],]
>
> #create test data set
> testData <- SouthData[randIndex[(cutPoint2_3+1):dim(SouthData)[1]],]
>
> #use dim() function to demostrate the test and training data sets contain the appropriate number of cases
> dim(trainData)
[1] 6384   17
> dim(testData)
[1] 3193   17
```

```
> svmOutput
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 5

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.045684040639401

Number of Support Vectors : 3072

Objective Function Value : -12479.82
Training error : 0.178728
Cross validation error : 0.207237
Probability model included.
>
> #validate model against test data using the predict() function
> svmPred <- predict(svmOutput, testData, type = "votes")
>
>
> #review contents of svmPred using str() command
> str(svmPred)
 num [1:2, 1:3193] 0 1 1 0 1 0 1 0 1 0 ...
>
> #create a confusion matrix (2x2 table) that compares the second row of svmPred to the contents of SouthData$Satisfaction variable
> compTable <-data.frame(testData$Satisfaction, svmPred[2,])
> results <- table(compTable)
> results
                     svmPred.2...
testData.Satisfaction    0    1
        Not Satisfied 1017  514
        Satisfied      152 1510
>
> #calculate an error rate based on the confusion matrix
> ((results[1,2]+results[2,1])/length(SouthData$Satisfaction))
[1] 0.06954161
```

Meaningful Insights:

1. Satisfaction was reported as greater than 3

2. Not Satisfied Customers, Satisfaction was reported as 3 or less

3. SVM Model Correctly Identified
    a. 1063 NotSatisfied Customers
    b. 1480 Satisfied Customers

4. 80% of the responses were predicted correctly from the svm model

Chapter 6 - ACTIONABLE INSIGHTS / OVERALL INTERPRETATION OF THE RESULTS

OBSERVATION	RECOMMENDATION
As age increases, Personal travel surpasses both Business and Mileage ticket travel	Possible stipend for seniors, as their frequency of travel increase while price sensitivity remains unchanged, essentially allowing seniors to travel more aiding in higher satisfaction
Price Sensitivity for Business travel and Mileage has the least variation throughout age and frequency	To increase profits, increasing the price of business travel would have little no impact on the overall change in frequency. Coupled with an increase benefit
Mileage tickets and Personal travel are correlated from 15 - 25, while satisfaction is rather low	Promote a higher mileage point award system for business travel around age 45, where the highest frequency is, and where we believe people are using mileage tickets for both personal and mileage type of travel outside of work. This will raise the overall satisfaction as price sensitivity is correlated to satisfaction
Customers who travel by BLUE STATUS flights are unsatisfied	Improve amenities for blue status flights and work
Customers who are in the category of Personal travel have the lowest overall satisfaction	Provide family plans to reduce flight cost, this would increase overall satisfaction within the subcategory of Personal travel.
Customers in a higher age range are unsatisfied	Better healthcare facilities and proper care of their needs which will drive their Satisfaction. Discount Offers for Senior citizens.
Women have 1000 more flights in Personal travel, which has the lowest average satisfaction.	Women have 1000 more flights in Personal travel, which has the lowest average satisfaction. By alleviate low satisfaction for this group in Personal travel in any way would increase in the overall satisfaction of that type of travel.

Chapter 7 - CONCLUSION

The data analyzed for Southeast Airlines provided useful insights that helped Southeast Airlines in order to improve customer satisfaction which will in turn gain profits giving a competitive edge over competitors.

Chapter 8 - APPENDIX

```
### Pat Carlin
### IST 687 FinalProject
################### Section 0: Loading and Cleaning the Data
### Insalling Packages
install.packages("ggplot2")
install.packages("dplyr")
install.packages("ggmap")
install.packages("mapdata")
install.packages("wordcloud")
install.packages("tm")
library(mapdata)
library(ggmap)
library(ggplot2)
library(dplyr)
library(wordcloud)
library(tm)
### Reading the data into R
ProjectData <- read.csv("~/IST 687 Project/ProjectFullDataSet.csv")
str(ProjectData)
 ### Cleaning the names and punctuation of the dataset
colnames(ProjectData)
colnames(ProjectData)[6]<- "YearOfFirstFlight"
colnames(ProjectData)[8]<- "FlightsWithOtherAirlines"
colnames(ProjectData)[7]<- "FlightsPerYear"
colnames(ProjectData)[9]<- "TypeOfTravel"
colnames(ProjectData)[10]<- "NoOfOtherLoyaltyCards"
colnames(ProjectData)[11]<- "ShoppingAtAirport"
```

```
colnames(ProjectData)[12]<- "EatingAndDrinkingAtAirport"

colnames(ProjectData)[14]<- "DayOfMonth"

colnames(ProjectData)[15]<- "FlightDate"

colnames(ProjectData)[25]<- "FlightsCancelled"

colnames(ProjectData)[26]<- "FlightTimeInMinutes"

colnames(ProjectData)[28]<- "ArrivalDelayGreaterThan5mins"

colnames(ProjectData)[23]<- "DepartureDelayInMinutes"

colnames(ProjectData)[24]<- "ArrivalDelayInMinutes"

colnames(ProjectData)[18]<- "OriginCity"

ProjectData$DestinationCity<- gsub("(.*),.*", "\\1",ProjectData$DestinationCity)

ProjectData$OriginCity<- gsub("(.*),.*", "\\1",ProjectData$OriginCity)

names(ProjectData) <- gsub("\\.","", names(ProjectData))

names(ProjectData)

#################### Section 1: Focusing on our Southest Airlines Co.

### Isolating the Southeast Airlines Co.

SoutheastData <- ProjectData[which(ProjectData$AirlineName == "Southeast Airlines Co. " ),] #For Seperate DataFrame

View(SoutheastData)

### Transforming columns in SoutheastData to 'Factor' and 'Numeric'

SoutheastData$Satisfaction <- as.numeric(as.character(SoutheastData$Satisfaction))

SoutheastData$Age <- as.numeric(as.character(SoutheastData$Age))

SoutheastData$PriceSensitivity <- as.numeric(as.character(SoutheastData$PriceSensitivity))

SoutheastData$YearOfFirstFlight <- as.numeric(as.character(SoutheastData$YearOfFirstFlight))

SoutheastData$FlightsPerYear <- as.numeric(as.character(SoutheastData$FlightsPerYear))

SoutheastData$FlightsWithOtherAirlines <- as.numeric(as.character(SoutheastData$FlightsWithOtherAirlines))

#################### Section 3: Exploring how the TypeOfTravel effects the Satisfaction

### Plotting the average Satisfaction against TypeOfTravel

TypeOfTravSat <- SoutheastData %>% group_by(TypeOfTravel) %>% summarise(TypeSat = mean(Satisfaction))

TypeOfTravPlot <- ggplot(TypeOfTravSat, aes(x = TypeOfTravel, y = TypeSat)) + geom_bar(stat = "identity", color = "black", fill =
"orange") + ylim(0,5)

TypeOfTravPlot <- TypeOfTravPlot + ggtitle("Type Of Travel vs Satisfaction") + xlab("Type Of Travel") + ylab("Satisfaction")

TypeOfTravPlot <- TypeOfTravPlot + theme_classic() + theme(plot.title = element_text(hjust = 0.5))

TypeOfTravPlot

### Creating a boxplot to visualize frequency of the TypeOfTravel

TypeSatBoxPlot <- ggplot(data = SoutheastData, aes(x = TypeOfTravel, y = Satisfaction)) + geom_boxplot(fill = "lightblue", colour =
"black")

TypeSatBoxPlot <- TypeSatBoxPlot + ggtitle("Type Of Travel vs Satisfaction") + xlab("Type Of Travel") + ylab("Satisfaction") +
theme_bw()
```

```
TypeSatBoxPlot <- TypeSatBoxPlot + theme(plot.title = element_text(hjust = 0.5)) + geom_jitter()

TypeSatBoxPlot


### Creating a smoothed scatter plot to see the effects of TypeOfTravel on Age and average Satisfaction

TypeAgeSatisfaction <- SoutheastData %>% group_by(Age, TypeOfTravel) %>% summarise(TypeAgeSat = mean(Satisfaction))

AgeTypeLinePlot <- ggplot(data = TypeAgeSatisfaction, aes( x = Age, y = TypeAgeSat, group = TypeOfTravel, colour = TypeOfTravel)) + geom_point() + geom_smooth()

AgeTypeLinePlot <- AgeTypeLinePlot + xlab("Age") + ylab("Satisfaction") + ggtitle("Types Of Travel vs Satisfaction") + theme_bw() + theme(plot.title = element_text(hjust = 0.5))

AgeTypeLinePlot <- AgeTypeLinePlot + scale_x_continuous(breaks = seq(15, 85, 5))

AgeTypeLinePlot

### Plotting the AgeOfPlane against TypeOfTravel

AgeTypePlot <- ggplot(data = SoutheastData, aes(x = Age, fill = TypeOfTravel)) + geom_histogram(binwidth = 2)

AgeTypePlot <- AgeTypePlot + xlab("Age") + ylab("Number Of Flights") + ggtitle("Types Of Travel Across Age") + theme_classic() + theme(plot.title = element_text(hjust = 0.5))

AgeTypePlot <- AgeTypePlot + scale_x_continuous(breaks = seq(15, 85, 5))

AgeTypePlot ### Need to find a better visual

### Finding the density plot for the TypeOfTravel and Age

TypeDensityPlot <- ggplot(SoutheastData, aes(Age)) + geom_density(aes(fill = factor(TypeOfTravel)), alpha=0.8)

TypeDensityPlot <- TypeDensityPlot + xlab("Age") + ylab("Flight Density") + ggtitle("Flight Density for Travel Types") + theme_classic() + theme(plot.title = element_text(hjust = 0.5))

TypeDensityPlot <- TypeDensityPlot + scale_x_continuous(breaks = seq(15, 85, 5)) + guides(fill=guide_legend(title="Types"))

TypeDensityPlot

### Seeing the effects of Age on average PriceSensitivity between TypeOfTravel

PriceAgeSat <- SoutheastData %>% group_by(Age, TypeOfTravel) %>% summarise(PriceSat = mean(PriceSensitivity)

PriceAgePlot <- ggplot(data = PriceAgeSat, aes(x = Age, y = PriceSat, group = TypeOfTravel, colour = TypeOfTravel)) + geom_smooth()

PriceAgePlot <- PriceAgePlot + ggtitle("Satisfaction vs Time & Price Sensitivity") + xlab("Age") + ylab("Average Price Sensitivity") + theme_bw()

PriceAgePlot <- PriceAgePlot + theme_bw() + theme(plot.title = element_text(hjust = 0.5))

PriceAgePlot <- PriceAgePlot + scale_x_continuous(breaks = seq(15, 85, 5))

PriceAgePlot

############## Section 4: Exploring how Gender and Age affect Satisfaction within SoutheastData

### Showing variance in Gender and frequency of flight between the TypeOfTravel

TypeGenderPlot <- ggplot(data = SoutheastData, aes(x = TypeOfTravel, fill = Gender)) + geom_bar(colour = "black", stat="count", position = position_dodge(), size = .5)

TypeGenderPlot <- TypeGenderPlot + scale_fill_hue(name="Gender") + xlab("Type Of Travel") + ylab("Number Of Flights") + ggtitle("Type Of Flights Across Gender")

TypeGenderPlot <- TypeGenderPlot + theme_bw() + theme(plot.title = element_text(hjust = 0.5))
```

```
TypeGenderPlot
### Plotting average Satisfaction against Age for Gender
GenderAgeSat <- SoutheastData %>% group_by(Age, Gender) %>% summarise(GenAgeSat = mean(Satisfaction))

GenderAgePlot <- ggplot(data = GenderAgeSat, aes(x = Age, y = GenAgeSat, group = Gender, colour = Gender)) + geom_line()
GenderAgePlot <- GenderAgePlot + ggtitle("Satisfaction vs Time & Gender") + xlab("Age") + ylab("Satisfaction") + theme_classic()
GenderAgePlot <- GenderAgePlot + theme_bw() + theme(plot.title = element_text(hjust = 0.5))
GenderAgePlot <- GenderAgePlot + scale_x_continuous(breaks = seq(15, 85, 5))
GenderAgePlot



##ShwetJain
##IST 687 Final Project
#CleanedData
#Different Plots
### Installing Packages
install.packages("ggplot2")
library(ggplot2)
### Reading the data into R
ProjectData <- read.csv("Satisfaction Survey.csv", header = TRUE)
str(ProjectData)
### Removing "." in ProjectData
names(ProjectData)
names(ProjectData) <- gsub("\\.","", names(ProjectData))
names(ProjectData)
View(ProjectData)



colnames(ProjectData)
colnames(ProjectData)[6]<- "YearOfFirstFlight"
colnames(ProjectData)[8]<- "FlightsWithOtherAirlines"
colnames(ProjectData)[7]<- "FlightsPerYear"
colnames(ProjectData)[10]<- "NoOfOtherLoyaltyCards"
colnames(ProjectData)[11]<- "ShoppingAtAirport"
```

```
colnames(ProjectData)[12]<- "EatingAndDrinkingAtAirport"

colnames(ProjectData)[14]<- "DayOfMonth"

colnames(ProjectData)[15]<- "FlightDate"

colnames(ProjectData)[25]<- "FlightsCancelled"

colnames(ProjectData)[26]<- "FlightTimeInMinutes"

colnames(ProjectData)[28]<- "ArrivalDelayGreaterThan5mins"

colnames(ProjectData)[23]<- "DepartureDelayInMinutes"

colnames(ProjectData)[24]<- "ArrivalDelayInMinutes"

colnames(ProjectData)[18]<- "OriginCity"



ProjectData$DestinationCity<- gsub("(.*),.*", "\\1",ProjectData$DestinationCity)

ProjectData$OriginCity<- gsub("(.*),.*", "\\1",ProjectData$OriginCity)


SouthData <- ProjectData[which(ProjectData$AirlineName == "Southeast Airlines Co. " ),] #For Seperate DataFrame

View(SouthData)

str(SouthData)


sum(is.na(SouthData$ArrivalDelayInMinutes)) #154

sum(is.na(SouthData$FlightTimeInMinutes)) #154

sum(is.na(SouthData$DepartureDelayInMinutes)) #129

#REPLACING THE NA'S

SouthData$ArrivalDelayInMinutes[is.na(SouthData$ArrivalDelayInMinutes)] <- round(mean(SouthData$ArrivalDelayInMinutes, na.rm =
TRUE))

SouthData$FlightTimeInMinutes[is.na(SouthData$FlightTimeInMinutes)] <- round(mean(SouthData$FlightTimeInMinutes, na.rm =
TRUE))

SouthData$DepartureDelayInMinutes[is.na(SouthData$DepartureDelayInMinutes)] <-
round(mean(SouthData$DepartureDelayInMinutes, na.rm = TRUE))


install.packages("ggplot2")

library(ggplot2)


install.packages("dplyr")

library(dplyr)
```

```
SouthData$Satisfaction <- as.numeric(as.character(SouthData$Satisfaction))


str(SouthData)

by_AirlineStatus <- group_by(SouthData,AirlineStatus)

by_AirlineStatus<-summarise(by_day,Satisfaction=mean(Satisfaction))

Plot_AirlineStatus <- ggplot(by_day,aes(x=AirlineStatus,y=Satisfaction)) + geom_col()+ theme(axis.text.x = element_text(angle = 90))+
ggtitle("Relationship between AirlineStatus and Satisfaction")   #Using the theme function and setting it to 90 will flip it and give a clear
representation.




Age_Gender<-ggplot(SouthData, aes(x=Age, fill=Gender, color=Gender)) + geom_histogram(position="identity", alpha=0.5) +
geom_vline(data=mu, aes(xintercept=grp.mean, color=Gender),linetype="dashed")


Airlines_Count <- ggplot(ProjectData, aes(AirlineName)) + geom_bar() +  theme(axis.text.x = element_text(angle = 90)) +
geom_bar(aes(fill = Satisfaction)) + ggtitle("Frequency of Different Airlines with repective Customer Satisfaction")


ProjectData$Satisfaction<- as.factor(ProjectData$Satisfaction)


by_Gender <- group_by(SouthData,Gender)

by_Gender<-summarise(by_Gender,Satisfaction=mean(Satisfaction))

Gender_Plot <-ggplot(by_Gender, aes(x = Gender, y = Satisfaction)) + geom_bar(stat = "identity") + ggtitle("Relationship between
Gender and Satisfaction")



by_PriceSensitivity <- group_by(SouthData,PriceSensitivity)

by_PriceSensitivity <-summarise(by_PriceSensitivity,Satisfaction=mean(Satisfaction))

PriceS_Plot<- ggplot(by_PriceSensitivity, aes(x =PriceSensitivity, y = Satisfaction)) + geom_bar(stat = "identity") + ggtitle("Relationship
between Price Sensitivity and Satisfaction")



by_Class <- group_by(SouthData,Class)

by_Class <-summarise(by_Class,Satisfaction=mean(Satisfaction))

ClassPlot <-ggplot(by_Class, aes(x=Class, y=Satisfaction)) + geom_bar(stat="identity", fill="steelblue") + theme_minimal()
```

```
by_OriginDestination <- group_by(SouthData,OriginCity,DestinationCity)

by_OriginDestination <-summarise(by_OriginDestination,Satisfaction=mean(Satisfaction),count=n())

OriginDestinationPlot<- ggplot(by_OriginDestination, aes(x=OriginCity, y=DestinationCity)) + geom_point(aes(size=Satisfaction)) +
theme(axis.text.x = element_text(angle = 90)) + ggtitle("Relation between OriginState and DestinationState with it's Satisfaction")



by_NoOfOtherLoyaltyCards <- group_by(SouthData,NoOfOtherLoyaltyCards)

by_NoOfOtherLoyaltyCards <-summarise(by_NoOfOtherLoyaltyCards,Satisfaction=mean(Satisfaction))

LoyaltyCardsPlot<- ggplot(by_NoOfOtherLoyaltyCards, aes(x =NoOfOtherLoyaltyCards, y = Satisfaction)) + geom_bar(stat = "identity")



by_PriceSensitivity_TypeOfTravel <- group_by(SouthData,PriceSensitivity,TypeofTravel)

by_PriceSensitivity_TypeOfTravel<-summarise(by_PriceSensitivity_TypeOfTravel,Satisfaction=mean(Satisfaction),count=n())

plot(by_PriceSensitivity_TypeOfTravel)


ggplot(by_PriceSensitivity_TypeOfTravel, aes(fill=by_PriceSensitivity_TypeOfTravel$TypeofTravel,
y=by_PriceSensitivity_TypeOfTravel$Satisfaction, x=by_PriceSensitivity_TypeOfTravel$PriceSensitivity)) + xlab("PriceSensitivity")+
ylab("Satisfaction")+geom_bar(position="dodge", stat="identity")

#Association Model

### Insalling Packages

install.packages("ggplot2")

library(ggplot2)

### Reading the data into R

ProjectData <- read.csv("Satisfaction Survey.csv", header = TRUE)

str(ProjectData)

### Removing "." in ProjectData

names(ProjectData)

names(ProjectData) <- gsub("\\.","", names(ProjectData))

names(ProjectData)

View(ProjectData)

summary(ProjectData)

colnames(ProjectData)

colnames(ProjectData)[6]<- "YearOfFirstFlight"
```

```
colnames(ProjectData)[8]<- "FlightsWithOtherAirlines"

colnames(ProjectData)[7]<- "FlightsPerYear"

colnames(ProjectData)[10]<- "NoOfOtherLoyaltyCards"

colnames(ProjectData)[11]<- "ShoppingAtAirport"

colnames(ProjectData)[12]<- "EatingAndDrinkingAtAirport"

colnames(ProjectData)[14]<- "DayOfMonth"

colnames(ProjectData)[15]<- "FlightDate"

colnames(ProjectData)[25]<- "FlightsCancelled"

colnames(ProjectData)[26]<- "FlightTimeInMinutes"

colnames(ProjectData)[28]<- "ArrivalDelayGreaterThan5mins"

colnames(ProjectData)[23]<- "DepartureDelayInMinutes"

colnames(ProjectData)[24]<- "ArrivalDelayInMinutes"

colnames(ProjectData)[18]<- "OriginCity"

ProjectData$DestinationCity<- gsub("(.*),.*", "\\1",ProjectData$DestinationCity)

ProjectData$OriginCity<- gsub("(.*),.*", "\\1",ProjectData$OriginCity)

SouthData <- ProjectData[which(ProjectData$AirlineName == "Southeast Airlines Co. " ),] #For Seperate DataFrame

View(SouthData)

str(SouthData)

sum(is.na(SouthData$ArrivalDelayInMinutes)) #154

sum(is.na(SouthData$FlightTimeInMinutes)) #154

sum(is.na(SouthData$DepartureDelayInMinutes)) #129

SouthData$ArrivalDelayInMinutes[is.na(SouthData$ArrivalDelayInMinutes)] <- round(mean(SouthData$ArrivalDelayInMinutes, na.rm =
TRUE))

SouthData$FlightTimeInMinutes[is.na(SouthData$FlightTimeInMinutes)] <- round(mean(SouthData$FlightTimeInMinutes, na.rm =
TRUE))

SouthData$DepartureDelayInMinutes[is.na(SouthData$DepartureDelayInMinutes)] <-
round(mean(SouthData$DepartureDelayInMinutes, na.rm = TRUE))

str(SouthData)

giveQuant <- function(y)

{

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

  Buckets<-replicate(length(y),"Average")

  Buckets[y<=q[1]] <- "Low"

  Buckets[y>q[2]] <- "High"

  return(Buckets)
```

```
}



giveLevel<-function(x)
{
  BucketsClean <- replicate(length(x),"Average")
  BucketsClean[x>=4] <- "High"
  BucketsClean[x<=3.3] <- "Low"
  return (BucketsClean)
}

SouthData <- SouthData[,-14:-17]
SouthData$Age<- giveQuant(SouthData$Age)
SouthData$PriceSensitivity<-giveQuant(SouthData$PriceSensitivity)
SouthData$YearOfFirstFlight<-giveQuant(SouthData$YearOfFirstFlight)
SouthData$FlightsPerYear<-giveQuant(SouthData$FlightsPerYear)
SouthData$FlightsWithOtherAirlines<-giveQuant(SouthData$FlightsWithOtherAirlines)
SouthData$NoOfOtherLoyaltyCards<-giveQuant(SouthData$NoOfOtherLoyaltyCards)
SouthData$ShoppingAtAirport<-giveQuant(SouthData$ShoppingAtAirport)
SouthData$EatingAndDrinkingAtAirport<-giveQuant(SouthData$EatingAndDrinkingAtAirport)
SouthData$FlightDistance<-giveQuant(SouthData$FlightDistance)
SouthData$ArrivalDelayInMinutes<-giveQuant(SouthData$ArrivalDelayInMinutes)
SouthData$FlightTimeInMinutes<-giveQuant(SouthData$FlightTimeInMinutes)
SouthData$DepartureDelayInMinutes<-giveQuant(SouthData$DepartureDelayInMinutes)
SouthData$ScheduledDepartureHour<-giveQuant(SouthData$ScheduledDepartureHour)

SouthData$Satisfaction<- as.numeric(as.character(SouthData$Satisfaction))

SouthData$Satisfaction <- giveLevel(SouthData$Satisfaction)

install.packages("arules")
library(arules)

install.packages("arulesViz")
```

```
library(arulesViz)



str(SouthData)
SouthData$Satisfaction <- as.factor(SouthData$Satisfaction)
SouthData$Age<-as.factor(SouthData$Age)

SouthData$PriceSensitivity <- as.factor(SouthData$PriceSensitivity)
SouthData$YearOfFirstFlight<- as.factor(SouthData$YearOfFirstFlight)
SouthData$FlightsPerYear<-as.factor(SouthData$FlightsPerYear)
SouthData$FlightsWithOtherAirlines<- as.factor(SouthData$FlightsWithOtherAirlines)
SouthData$NoOfOtherLoyaltyCards <- as.factor(SouthData$NoOfOtherLoyaltyCards)
SouthData$ShoppingAtAirport<-as.factor(SouthData$ShoppingAtAirport)
SouthData$EatingAndDrinkingAtAirport<-as.factor(SouthData$EatingAndDrinkingAtAirport)
SouthData$DepartureDelayInMinutes<-as.factor(SouthData$DepartureDelayInMinutes)
SouthData$ArrivalDelayInMinutes<-as.factor(SouthData$ArrivalDelayInMinutes)
SouthData$FlightTimeInMinutes<- as.factor(SouthData$FlightTimeInMinutes)
SouthData$FlightDistance<-as.factor(SouthData$FlightDistance)
SouthData$ScheduledDepartureHour<-as.factor(SouthData$ScheduledDepartureHour)
SouthData$OriginCity<-as.factor(SouthData$OriginCity)
SouthData$DestinationCity<-as.factor(SouthData$DestinationCity)

ruleset <- apriori(SouthData, list(support = 0.35,confidence = 0.40))
inspect(ruleset)
ruleSub<- subset(ruleset, subset = rhs %in% "Satisfaction=High")
inspect(ruleSub)

ruleset <- apriori(SouthData, list(support = 0.30,confidence = 0.43))
inspect(ruleset)
ruleSub<- subset(ruleset, subset = rhs %in% "Satisfaction=Low")
inspect(ruleSub)
```

```
#####Osama Junaid ##########
#final project
#install ggplot and ggmap
install.packages("ggplot2")
install.packages("mapdata")
install.packages("ggmap")
install.packages("dplyr")
library("ggplot2")
library(ggplot2)
library(ggmap)
library(maps)
library(mapdata)
library(dplyr)
install.packages("arules")
library(arules)
install.packages("arulesViz")
library(arulesViz)
install.packages("tmap")
install.packages("tmaptools")
install.packages("sf")
install.packages("leaflet")
library("tmap")
library("tmaptools")
library("sf")
library("leaflet")
ProjectData <- read.csv(file.choose(), header = TRUE)
View(ProjectData)
str(ProjectData)
```

```
### Removing "." in ProjectData
names(ProjectData)
names(ProjectData) <- gsub("\\.","", names(ProjectData))
names(ProjectData)
View(ProjectData)
colnames(ProjectData)
colnames(ProjectData)[6]<- "YearOfFirstFlight"
colnames(ProjectData)[8]<- "FlightsWithOtherAirlines"
colnames(ProjectData)[7]<- "FlightsPerYear"
colnames(ProjectData)[10]<- "NoOfOtherLoyaltyCards"
colnames(ProjectData)[11]<- "ShoppingAtAirport"
colnames(ProjectData)[12]<- "EatingAndDrinkingAtAirport"
colnames(ProjectData)[14]<- "DayOfMonth"
colnames(ProjectData)[15]<- "FlightDate"
colnames(ProjectData)[25]<- "FlightsCancelled"
colnames(ProjectData)[26]<- "FlightTimeInMinutes"
colnames(ProjectData)[28]<- "ArrivalDelayGreaterThan5mins"
colnames(ProjectData)[23]<- "DepartureDelayInMinutes"
colnames(ProjectData)[24]<- "ArrivalDelayInMinutes"
colnames(ProjectData)[18]<- "OriginCity"
ProjectData$DestinationCity<- gsub("(.*),.*", "\\1",ProjectData$DestinationCity)
ProjectData$OriginCity<- gsub("(.*),.*", "\\1",ProjectData$OriginCity)
names(ProjectData) <- gsub("\\.", "", names(ProjectData))
names(ProjectData)
View(ProjectData)
############################################################
#New data called Southeast
SoutheastData <- ProjectData[which(ProjectData$AirlineName == "Southeast Airlines Co. " ),] #For Seperate DataFrame
View(SoutheastData)
#############################################################################################################
#Calculating mean of satisfaction and grouping it by Origin state
originstatemean <- SoutheastData %>%
group_by(OriginState) %>%
summarize(Satisfaction = mean(as.numeric(Satisfaction)))
#Plotting graph for satisfaction vs origin state
```

```
SoutheastData$Satisfaction <- as.numeric(as.character(SoutheastData$Satisfaction))
originstatemean<-as.data.frame(originstatemean)
originstatemean
originState <- ggplot(originstatemean, aes(x=OriginState, y=Satisfaction)) + geom_col()
originState <- originState + ggtitle("Bar chart of customer satisfaction per Origin state for Southeast")
originState <- originState + theme(axis.text.x = element_text(angle = 90, hjust = 1))
originState <- originState + geom_bar(stat = "identity", fill ="blue")
originState
#############################################################################################################
#calculating mean of satisfaction and grouping by destination state
destinationstatemean <- SoutheastData %>%
group_by(DestinationState) %>%
summarize(Satisfaction = mean(as.numeric(Satisfaction)))
#plotting graph of satisfaction vs destination state
destinationstatemean<-as.data.frame(destinationstatemean)
destinationstatemean
destinationState <- ggplot(destinationstatemean, aes(x=DestinationState, y=Satisfaction)) + geom_col()
destinationState <- destinationState + ggtitle("Bar chart of customer satisfaction per Destination state for Southeast")
destinationState <- destinationState + theme(axis.text.x = element_text(angle = 90, hjust = 1))
destinationState <- destinationState + geom_bar(stat = "identity", fill ="Red")
destinationState
#############################################################################################################
#
#calculating mean of satisfaction and grouping by Eating and Drinking at Airport
EatandDrinkmean <- SoutheastData %>%
group_by(EatingAndDrinkingAtAirport) %>%
summarize(Satisfaction = mean(as.numeric(Satisfaction)))
#plotting graph of satisfaction vs destination state
EatandDrinkmean<-as.data.frame(EatandDrinkmean)
EatandDrinkmean
EatingandDrinking <- ggplot(EatandDrinkmean, aes(x=EatingAndDrinkingAtAirport, y=Satisfaction)) + geom_col()
EatingandDrinking  <- EatingandDrinking  + ggtitle("Bar chart of customer satisfaction per Eating and Drinking at Airport")
EatingandDrinking  <- EatingandDrinking  + theme(axis.text.x = element_text(angle = 90, hjust = 1))
EatingandDrinking  <- EatingandDrinking + geom_bar(stat = "identity", fill ="yellow")
EatingandDrinking
```

```
#####################################################################################################################
#
#culating mean of satisfaction and grouping by Year of first flight
yearofflightmean <- SoutheastData %>%
group_by(YearOfFirstFlight) %>%
summarize(Satisfaction = mean(as.numeric(Satisfaction)))
#plotting graph of satisfaction vs destination state
yearofflightmean<-as.data.frame(yearofflightmean)
yearofflightmean
YearofFF <- ggplot(yearofflightmean, aes(x=YearOfFirstFlight, y=Satisfaction)) + geom_col()
YearofFF  <- YearofFF  + ggtitle("Barchart of customer satisfaction per Year of first flight for Southeast")
YearofFF <- YearofFF + theme(axis.text.x = element_text(angle = 90, hjust = 1))
YearofFF <- YearofFF + geom_bar(stat = "identity", fill ="orange")
YearofFF
#####################################################################################################################
#
#culating mean of satisfaction and grouping by Shopping at airport
shopatAirportmean <- SoutheastData %>%
group_by(ShoppingAtAirport) %>%
summarize(Satisfaction = mean(as.numeric(Satisfaction)))
#plotting graph of satisfactionstate
shopatAirportmean<-as.data.frame(shopatAirportmean)
shopatAirportmean
shoppingatAirport <- ggplot(shopatAirportmean, aes(x=ShoppingAtAirport, y=Satisfaction)) + geom_col()
shoppingatAirport  <- shoppingatAirport  + ggtitle("Barchart of customer satisfaction for Shopping at Airport")
shoppingatAirport <- shoppingatAirport + theme(axis.text.x = element_text(angle = 90, hjust = 1))
shoppingatAirport <- shoppingatAirport + geom_bar(stat = "identity", fill ="green")
shoppingatAirport
#####################################################################################################################
#
#culating mean of satisfaction and grouping by class
classmean <- SoutheastData %>%
group_by(Class) %>%
summarize(Satisfaction = mean(as.numeric(Satisfaction)))
#plotting graph of satisfaction vs destination state
classmean<-as.data.frame(classmean)
classmean
```

```
classes <- ggplot(classmean, aes(x=Class, y=Satisfaction)) + geom_col()
classes  <- classes  + ggtitle("Bar chart of customer satisfaction per Class")
classes <- classes + theme(axis.text.x = element_text(angle = 90, hjust = 1))
classes <- classes + geom_bar(stat = "identity", fill ="#FF6666")
classes
#########################################################################
#culating mean of satisfaction and grouping by Type of travel
typeoftravelmean <- SoutheastData %>%
group_by(TypeofTravel) %>%
summarize(Satisfaction = mean(as.numeric(Satisfaction)))
#plotting graph of satisfaction vs destination state
typeoftravelmean<-as.data.frame(typeoftravelmean)
typeoftravelmean
typeoftravel <- ggplot(typeoftravelmean, aes(x=TypeofTravel, y=Satisfaction)) + geom_col()
typeoftravel  <- typeoftravel  + ggtitle("Bar chart of customer satisfaction per Type of travel")
typeoftravel <- typeoftravel + theme(axis.text.x = element_text(angle = 90, hjust = 1))
typeoftravel <- typeoftravel + geom_bar(stat = "identity", fill ="green")
typeoftravel
#########################################################################################################################
#
shoppingamountatAirport <- ggplot(SoutheastData, aes(x=shoppingatAirport, y=Satisfaction, colour="purple")) + geom_point()
shoppingamountatAirport
#########################################################################################################################
#created a new airlines
NorthwestData <- ProjectData[which(ProjectData$AirlineName == "Northwest Business Airlines Inc. " ),]
View(NorthwestData)
#type of travel for this airlines
#culating mean of satisfaction and grouping by Type of travel
typeoftravelmeann <- NorthwestData %>%
group_by(TypeofTravel) %>%
summarize(Satisfaction = mean(as.numeric(Satisfaction)))
#plotting graph of satisfaction vs destination state
typeoftravelmeann<-as.data.frame(typeoftravelmean)
typeoftravelmeann
typeoftravell <- ggplot(typeoftravelmeann, aes(x=TypeofTravel, y=Satisfaction)) + geom_col()
typeoftravell  <- typeoftravell  + ggtitle("Bar chart of customer satisfaction per Type of travel for Northwest")
```

```
typeoftravell <- typeoftravell + theme(axis.text.x = element_text(angle = 90, hjust = 1))

typeoftravell <- typeoftravell + geom_bar(stat = "identity", fill ="purple")

typeoftravell

#############################################################################################

originstatemeann <- NorthwestData %>%

group_by(OriginState) %>%

summarize(Satisfaction = mean(as.numeric(Satisfaction)))

#Plotting graph for satisfaction vs origin state

NorthwestData$Satisfaction <- as.numeric(as.character(NorthwestData$Satisfaction))

originstatemeann<-as.data.frame(originstatemeann)

originstatemeann

originStatee <- ggplot(originstatemeann, aes(x=OriginState, y=Satisfaction)) + geom_col()

originStatee <- originStatee + ggtitle("Bar chart of customer satisfaction per Originstate for Northwest")

originStatee <- originStatee + theme(axis.text.x = element_text(angle = 90, hjust = 1))

originStatee <- originStatee + geom_bar(stat = "identity", fill ="Purple")

originStatee

#############################################################################################################
#

EatandDrinkmeann <- NorthwestData %>%

group_by(EatingAndDrinkingAtAirport) %>%

summarize(Satisfaction = mean(as.numeric(Satisfaction)))

#plotting graph of satisfaction vs destination state

EatandDrinkmeann <-as.data.frame(EatandDrinkmeann)

EatandDrinkmeann

EatingandDrinkingg <- ggplot(EatandDrinkmeann, aes(x=EatingAndDrinkingAtAirport, y=Satisfaction)) + geom_col()

EatingandDrinkingg  <- EatingandDrinkingg  + ggtitle("Bar chart of customer satisfaction per Eating and Drinking at Airport for
Northwest")

EatingandDrinkingg  <- EatingandDrinkingg  + theme(axis.text.x = element_text(angle = 90, hjust = 1))

EatingandDrinkingg  <- EatingandDrinkingg + geom_bar(stat = "identity", fill ="green")

EatingandDrinkingg

#############################################################################################################
#

#culating mean of satisfaction and grouping by class

classmeann <- NorthwestData %>%

  group_by(Class) %>%

  summarize(Satisfaction = mean(as.numeric(Satisfaction)))

#plotting graph of satisfaction vs destination state
```

```
classmeann<-as.data.frame(classmeann)
classmeann
classess <- ggplot(classmean, aes(x=Class, y=Satisfaction)) + geom_col()
classess  <- classess  + ggtitle("Bar chart of customer satisfaction per Class for Northwest")
classess <- classess + theme(axis.text.x = element_text(angle = 90, hjust = 1))
classess <- classess + geom_bar(stat = "identity", fill ="brown")
classess
#########################################################################################################################
#culating mean of satisfaction and grouping by Shopping at airport for Northwest
shopatAirportmeann <- NorthwestData %>%
group_by(ShoppingAtAirport) %>%
summarize(Satisfaction = mean(as.numeric(Satisfaction)))
#plotting graph of satisfactionstate
shopatAirportmeann <-as.data.frame(shopatAirportmeann)
shopatAirportmeann
shoppingatAirportt <- ggplot(shopatAirportmeann, aes(x=ShoppingAtAirport, y=Satisfaction)) + geom_col()
shoppingatAirportt  <- shoppingatAirport  + ggtitle("Bar chart of customer satisfaction per Shopping at Airport for Northwest")
shoppingatAirportt <- shoppingatAirportt + theme(axis.text.x = element_text(angle = 90, hjust = 1))
shoppingatAirportt <- shoppingatAirportt + geom_bar(stat = "identity", fill ="red")
shoppingatAirportt
#############################################################################
destinationstatemeann <- NorthwestData %>%
group_by(DestinationState) %>%
summarize(Satisfaction = mean(as.numeric(Satisfaction)))
#plotting graph of satisfaction vs destination state
destinationstatemeann<-as.data.frame(destinationstatemeann)
destinationstatemeann
destinationStatee <- ggplot(destinationstatemeann, aes(x=DestinationState, y=Satisfaction)) + geom_col()
destinationStatee <- destinationStatee + ggtitle("Barchart of customer satisfaction per Destinationstate for Northwest")
destinationStatee <- destinationStatee + theme(axis.text.x = element_text(angle = 90, hjust = 1))
destinationStatee <- destinationStatee + geom_bar(stat = "identity", fill ="pink")
destinationStatee
```

Thank You