

IST 687: Introduction to Data Science:

Customer Intelligence on Airline Data for South-East Airlines

                              Akshita Chandiramani

         Eashani Deorukhkar

Jay Kachhadia

          Rohan Mahajan

Yesaswi Avula

**Professor: Jeffrey Saltz**

**TA: Ivan Shamshurin**

Contents

[Introduction](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.gjdgxs) 3

[Objective](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.30j0zll) 3

[Background](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.1fob9te) 3

[Context](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.3znysh7) 3

[Scope 4](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.2et92p0)

[Business Questions 5](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.tyjcwt)

[Initial business questions 5](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.3dy6vkm)

[Final Business questions 6](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.1t3h5sf)

[Data Analysis 6](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.4d34og8)

[Data Acquisition 6](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.2s8eyo1)

[Data Cleansing 7](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.17dp8vu)

[Data Transformation 8](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.3rdcrjn)

[Data Munging](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.26in1rg)   8

[Descriptive statistics & Visualizations 9](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.lnxbz9)

[BQ 1: How is the overall satisfaction of Southeast customers compared to other airlines?](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.3whwml4)             10

[BQ 2: Does flight distance, flight time, arrival or departure delay affect customer satisfaction?](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.2bn6wsx) 11

[BQ3: How does age affect customer satisfaction? Which age group gives lower ratings?](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.qsh70q)              13

[BQ 4: Is there any relationship between gender and customer satisfaction?](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.3as4poj) 15

[Are females more likely to give](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.3as4poj) lower rating to the airline[?](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.3as4poj)

[BQ 5: Does Year of first flight and No. of flights p.a affect customer satisfactio](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.49x2ik5)n?

[BQ 6: Is it possible to associate airline status, type of travel and class with customer rating?](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.2p2csry) 16

[BQ 7: How much does Shopping, eating and drinking at the airport affect customer satisfaction?](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.147n2zr) 18

B[Q 8:  Does date of travel (day, week, month) affect customer satisfaction?](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.3o7alnk) 19

BQ 9: Do particular Origins and Destinations cause low customer satisfaction? 21

BQ 10: Do cancelled flights cause customers to give a low customer rating? 22

BQ11: Does Price Sensitivity affect Customer Satisfaction?                          22

[Use of modeling techniques & Visualizations](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.35nkun2) 23

[Linear Models](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.1ksv4uv) 23

Logistic Regression 23

[Naïve Bayes](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.2xcytpi) 24

Association rule mining 26

[Actionable Insights / Overall interpretation of results](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.1ci93xb) 29

[Conclusion 3](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.23ckvvd)0

**Introduction**

Objective

We are working as a consulting company for Southeast Airlines. The focus of our survey analysis project is to provide some notions and directions to Southeast Airlines so that they can improve its profit and consumer satisfaction. Concrete analysis of the airlines' data containing 14 major airline companies in the United States including Southeast Airlines helps us generate some useful insights into customers mindsets.

Background

The full 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.

Our goal here is to understand what factors influence a customer’s satisfaction after taking a particular airline’s flight. There can be different relationships between customers and all these 28 variables. Some of them make sense whereas some of them won’t make sense. Also, Various recommendations could be provided to our client Southeast airlines based on these insights we get from data analysis and model creation.

There are many business questions to be answered for improving Southeast airlines’ service. So, they could better satisfy their customers and increase their profits. Also, It’s essential to understand the market in which you are playing your game. So, Understanding your competitor’s strategy gives you actionable insights on where you should stress on to improve your business.

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. Getting actionable insights from this dataset helps them improve the services, win more customers and increase revenues. The dataset that we got was dirty and required much cleaning due to inappropriate values and missing values in multiple columns which needs to be handled thoughtfully.

Southeast Airlines for understanding their customers and their experiences. We think that answering these questions will help Southeast executives know more about the ways to improve Southeast airline Services.

Models were also created to make predictions on customer satisfaction so that we could better understand the critical features and whether or not all the models depict the same findings or not and why.

Scope

By playing with the features in the dataset doing some feature engineering, we can get insights on how we can improve Southeast services to improve customer satisfaction, attract new customers and encourage existing customers to avail Southeast airline services more frequently.

We have 29 features out of which some were not useful actually. here’s the detailed description of all the features.

1. **Satisfaction** – it is rated from 1 to 5, that how satisfied is the customer?
   1. 5 means higher satisfied, and 1 is lowest level of satisfaction.
2. **Airline Status** – each customer has a different type of airline status or package, which are platinum, gold, silver, and blue.
3. **Age** – the specific customer’s age. That is starting from 15 to 85 years old.
4. **Gender** – male or female.
5. **Price Sensitivity** – the grade to which the price affects to customers purchasing. The price sensitivity has a range from 0 to 5.
6. **Year of First Flight** – this attributes shows the first flight of each single customer. The range of year of the first flight for each customer has been started in 2003 until 2012.
7. **No of Flights p. a.** – this could be the number of flights that each customer has taken. The range starting from 0 to 100.
8. **Percent of Flight with other Airlines** – if we were Southeast Airline, we would like to know how many time that customer fly with other Airlines.
9. **Type of Travel** – is provide three traveling purpose for each consumer, which are business travel, mileage tickets that based on loyalty card, and personal travel like to see the family or in vacation
10. **No. Of other Loyalty Cards** – it is kind of membership card of each customer, that for retail establishment to gain a benefits such as, discounts.
11. **Shopping Amount at Airport** – showing the costumer’s result of how many products have been purchased. The range of shopping amount is from 0 to 875.
12. **Eating and Drinking at Airport** – it is the quantity eating and drinking per each consumer at the airport. The masseur of how often for eating and drinking, which is 0 to 895.
13. **Class** – it consisted of three different kinds of service level such as, business, and economy plus, economy. Moreover, customers have optional to choose their seat.
14. **Day of Month** – it means the traveling day of each costumer. In this attribute, shows total of 31 days of the month.
15. **Flight date** – all of these data are abbreviate the passenger’s flight date travel, which were since 2014 and only in January, February, and March.
16. **Airline Code** – basically, it is unique two or three digits that mean what is the specific type of airline. There are several codes that consumers have been going with. For example, AA, AS, B6, and DL.
17. **Airline Name** – There are several airlines company names such as, West Airways, Southeast Airlines Co, and FlyToSun Airlines Inc. This attribute provide what airline name that passenger have been used.
18. **Origin City** – refers to actual city that customers have departed from. For example, Yuma AZ, Waco TX, and Toledo HO.
19. **Origin State** – same thing as origin city such as, what state that customers have departed from? A good example, Texas, Ohio, Alaska, and Utah.
20. **Destination City** – the place to which passenger travels to. For example, Akron HO, Alpena MI, Austin TX, and Boston MA.
21. **Destination State** – also, it is the same thing as origin city, such as, to what state passenger travel to? Some example of destination states, Alaska, Kentucky, Iowa, and Florida.
22. **Scheduled Departure Hour** – the specific time at which passengers are scheduled to depart. In this data in scheduled departure hour is starting at 1 am until 23 pm.
23. **Departure Delay in Minutes** – which are minutes of departure delayed for each passenger, when compared to schedule. In this data the rage are starting from 0 until 1128 minutes.
24. **Arrival Delay in Minutes** – how many minutes of arrival delayed of each passenger. Rang of delayed minutes in this data are starting from 0 until 1115 minutes.
25. **Flight Cancelled** – occurs when the airline dose not operates the flight at all, and that is for a certain reason.
26. **Flight time in minutes** – indicate to period time to the destination.
27. **Flight Distance** – the extent of space between two places. Also, that means how many minutes are passenger traveling between two different places. Rang in this data starting from 31 until 4983 minutes.
28. **Arrival Delay greater 5 Minutes** – It means the delay of arrival airline time, which is more than 5 minutes per each passenger in the data.

**Business Questions**

Questions that drive the success of any business are called business questions. There are some questions which are very essential to understand in order to improve the service, attract more customers and increase customer satisfaction. Our Client Southeast airlines is interested in the same type of questions which can help them increase their profits and gain more customers.

We here also try to compare all the features of the dataset for Southeast airline and other airlines inorder to figure out where they are lacking and can improve upon. Trends shown by each of the feature in association with the customer satisfaction is taken into account to create a list of initial business questions.

Initial business questions

1. How much are the customers satisfied with Airline services in the USA?
2. In which areas Southeast airlines lag with respect to the other airlines?
3. What are the factors which really affect the customer satisfaction?
4. Are there any difference in how females rate airline services of flights than males?
5. Does Age have any relationship with customer satisfaction?
6. Can we model customer satisfaction using some of features in the dataset?
7. Does type of travel affects customer satisfaction?

After performing analysis on airlines data, We found that there are many other questions which can be answered for southeast airlines and in general for all airline services. Our team was also interested in knowing whether any actionable insights can be drawn from other features which are either customer centric or airline centric. So, Following are the business questions we came up with after doing the exploratory data analysis and modeling.

Final Business questions

Our new business questions were:

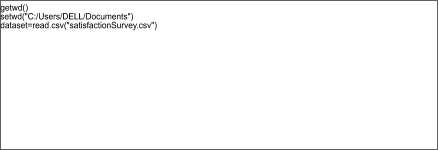
1. [How is the overall satisfaction of Southeast customers compared to other airlines?](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.3whwml4)
2. [Does flight distance, flight time, arrival or departure delay affect customer satisfaction?](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.2bn6wsx)
3. [How does age affect customer satisfaction? Which age group gives lower ratings?](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.qsh70q)
4. [Is there any relationship between gender and customer satisfaction?](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.3as4poj)
5. [Does Year of first flight and No. of flights p.a affect customer satisfaction](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.49x2ik5)?
6. [Is it possible to associate airline status, type of travel and class with customer rating](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.2p2csry)?
7. [How much does Shopping, eating and drinking at the airport affect customer satisfaction?](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.147n2zr)
8. [Does date of travel (day, week, month) affect customer satisfaction?](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.3o7alnk)
9. Do particular Origins and Destinations cause low customer satisfaction?
10. Do cancelled flights cause customers to give a low customer rating?
11. Does Price Sensitivity affect Customer Satisfaction?

**Data Analysis**

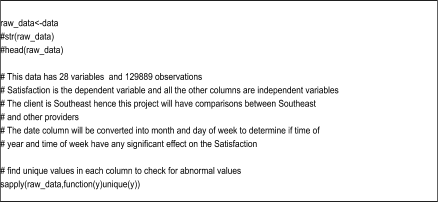
Data Acquisition

Our professor provided us with a dataset that contains customer feedback surveys obtained from the guests that have travelled through southeast airlines which is our client. This data set a csv file that contains surveys from the year 2014.

We downloaded the file using the following code:

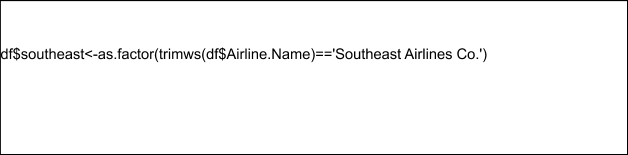


We checked the information of the given dataset using the following code.



Data Cleansing

For the cleaning process, our team decided to work with all the months but subset the dataset based on the columns that can help us to answer our business questions. We separated the southeast airlines as it is our client. We separated it from the other airlines for the sake of the results.



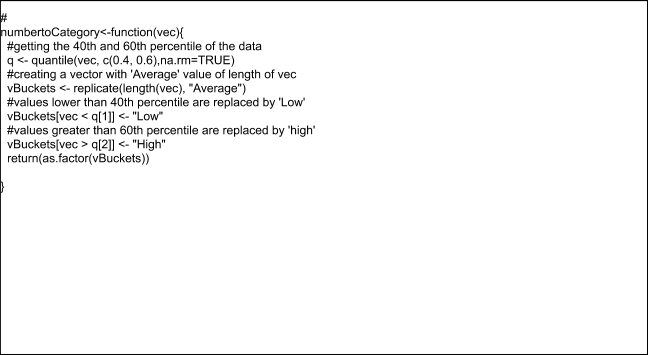
After selecting the columns, we decided to use the method na.omit to delete the missing values wherever needed.

We decided not to fulfil blanks with an average value because we considered that these assumptions can affect our final results. Also, the majority of the columns presented categorical data, so it was difficult to get an average or middle value.

Data Transformation

As it is mentioned in the previous section, many of the columns that we were working with contain categorical data. In order to work with linear models, it was necessary to transform the categorical values into numerical values. We transformed these values into numeric values using the following code:

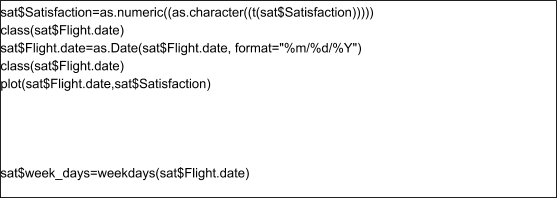


Data Munging****

Data munging is the process of programmatically transforming data into a format that makes it easier to work with.

After doing some experiments with the dataset, we noticed that our results were not what our team was expecting. We had some talks with our Professor, and he recommended us that we remove the variables which we felt that are not important or change them into usable methods. The date variable is split into weekdays and the client southeast is separated too.

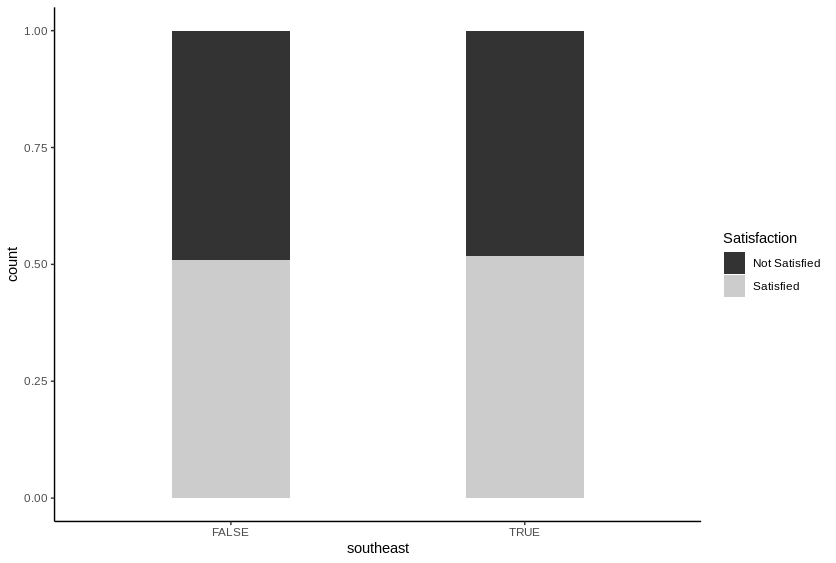
Different functions with other dataframes and vectors are used for all this. We used this to plot the graph as a part of our visualisation. The date is changed in the                                 following manner. The southeast is added as mentioned above.



**Descriptive Statistics & Visualizations**

After cleaning the dataset, our dataset contains 129889 observations and 28 columns. We have customer survey data of our client Southeast as well as other airlines. We converted the satisfaction into two categories namely “Satisfied” and “Not Satisfied”. All other variables where it was needed were converted into bins of categories for proper visualization.

[**BQ 1: How is the overall satisfaction of Southeast customers compared to other airlines?**](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.3whwml4)

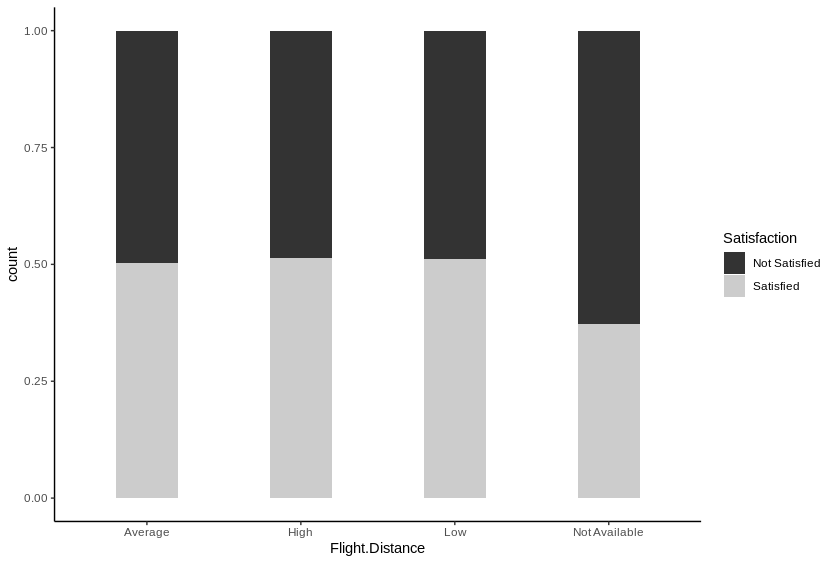


Southeast and other airlines have almost same satisfaction rate. If we see closely, Southeast

slightly more satisfied customers.

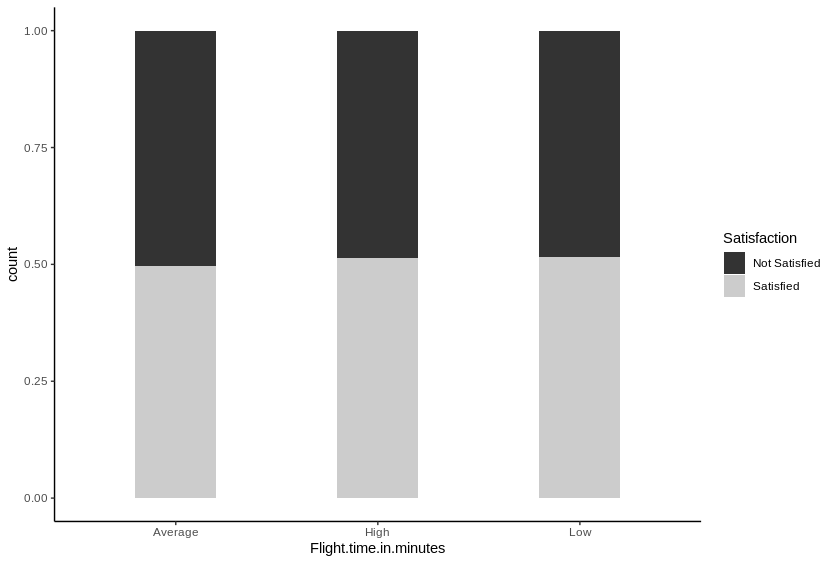
[**BQ 2: Does flight distance, flight time, departure delay affect customer satisfaction?**](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.2bn6wsx)

**Satisfaction vs. Flight distance.**



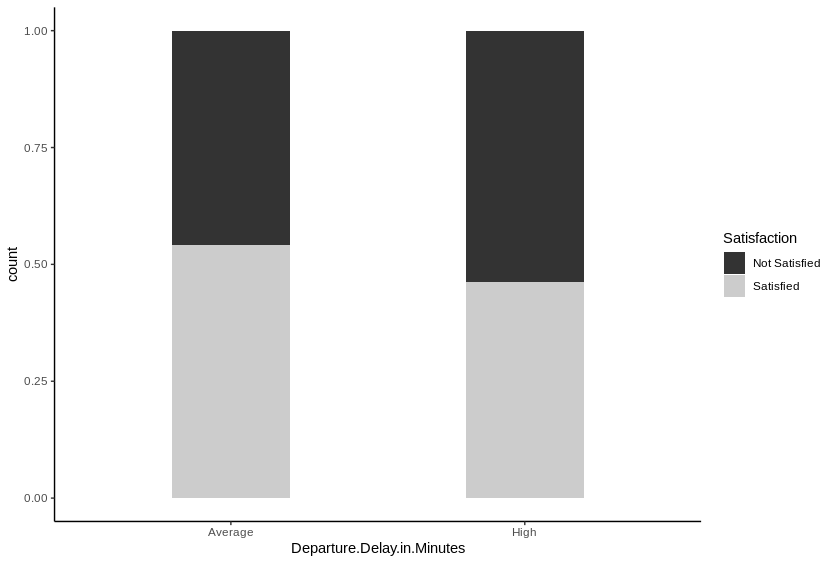
Satisfaction remains almost constant as flight distance increases. But for the cancelled flights which had flight distance not available have lower satisfaction.

**Satisfaction vs. Flight time.**



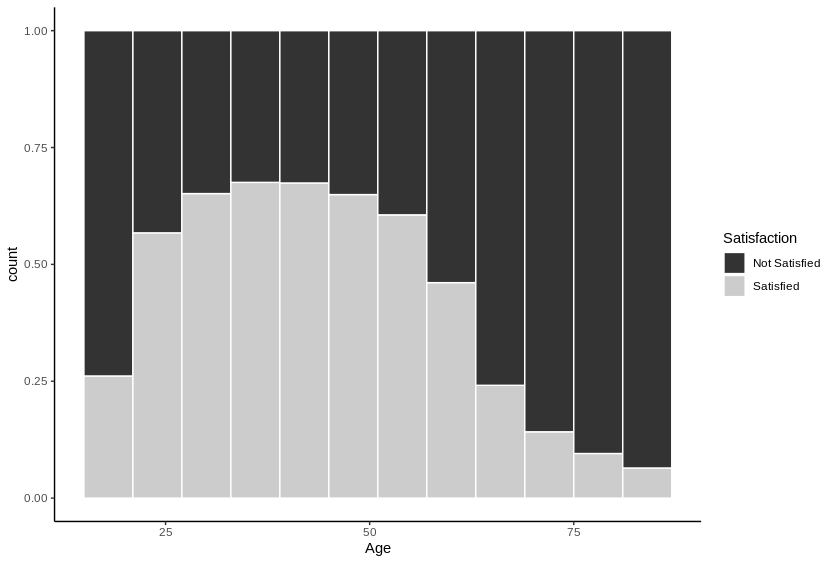
Satisfaction remains almost constant as Flight time increases.

**Satisfaction vs. Departure Delay for Southeast and other airlines.**



As departure delay increases, the satisfaction ratings tend to reduce.

[**BQ3: How does age affect customer satisfaction? Which age group gives lower ratings?**](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.qsh70q)

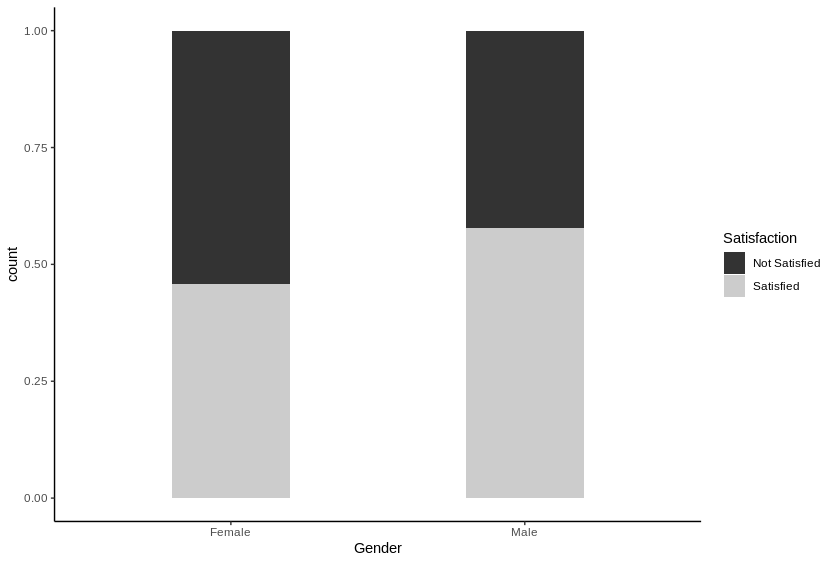


Age affects satisfaction rating. As customer’s age increases from 20 to 50, the satisfaction rating increases, after which it starts decreasing till the age of 80.

[**BQ 4: Is there any relationship between gender and customer satisfaction?**](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.3as4poj)

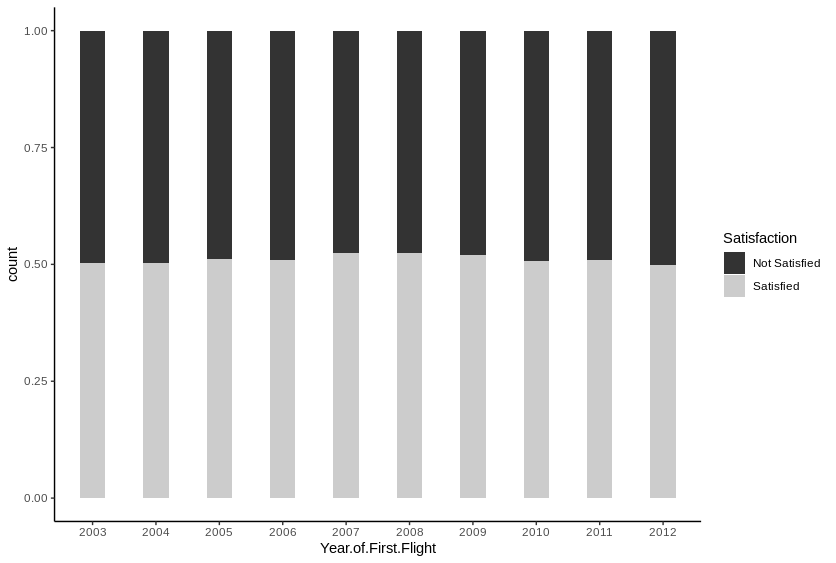
[**Are females more likely to give**](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.3as4poj) **lower rating to the airline**[**?**](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.3as4poj)

**Satisfaction vs. Gender.**

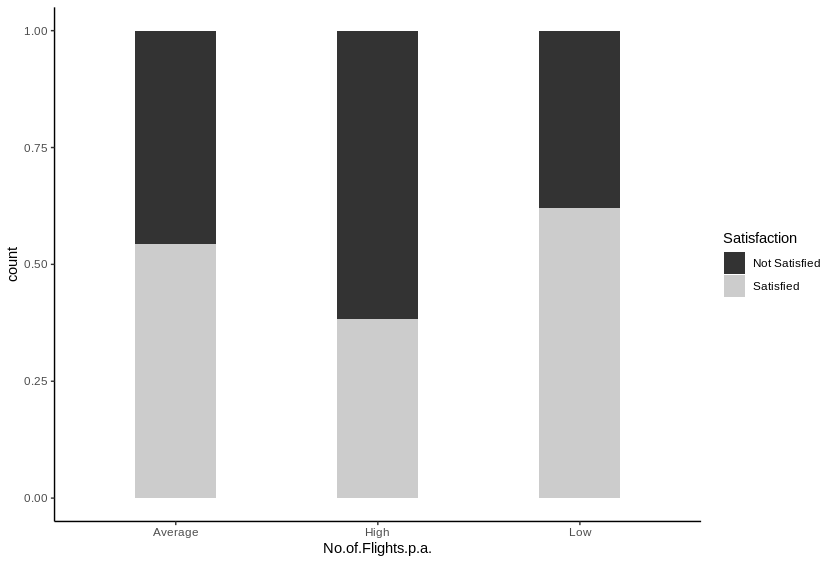


On an average, females give lesser satisfaction ratings compared to males.

[**BQ 5: Does Year of first flight and No. of flights p.a affect customer satisfaction?**](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.49x2ik5)



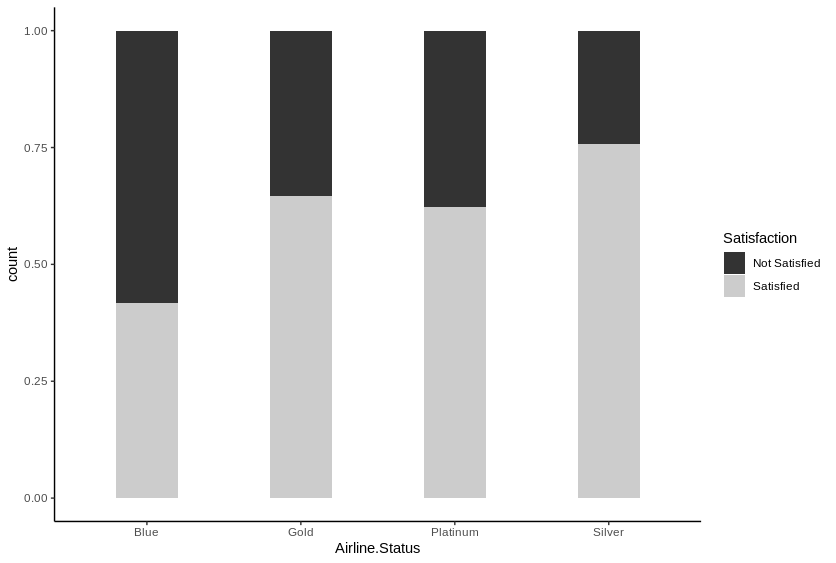
Year of flights does not affect the satisfaction ratings by a significant amount.



For southeast airlines, satisfaction tends to increase as the no. of flights p.a. decreases for customers.

[**BQ 6: Is it possible to associate airline status, type of travel and class with customer rating?**](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.2p2csry)

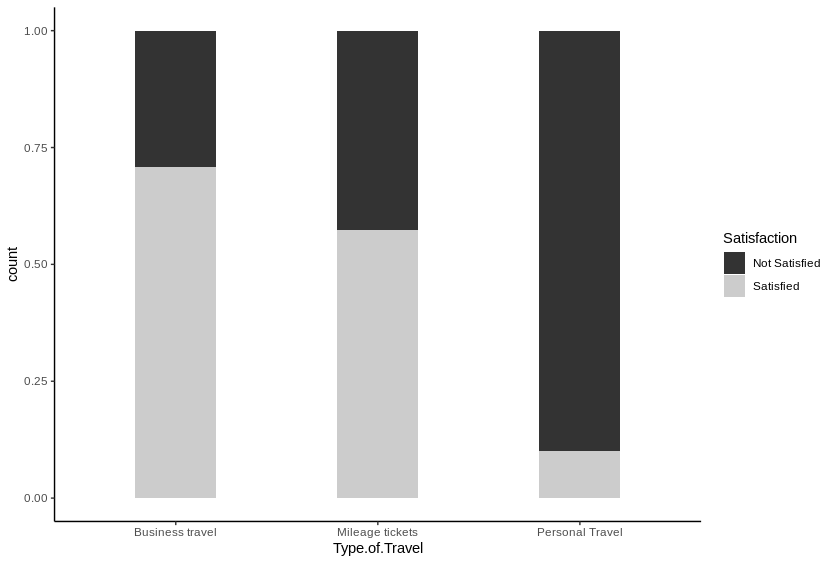
**Satisfaction vs. Airline Status.**



Customer’s travelling by blue status have the lowest satisfaction rating, followed by Platinum.

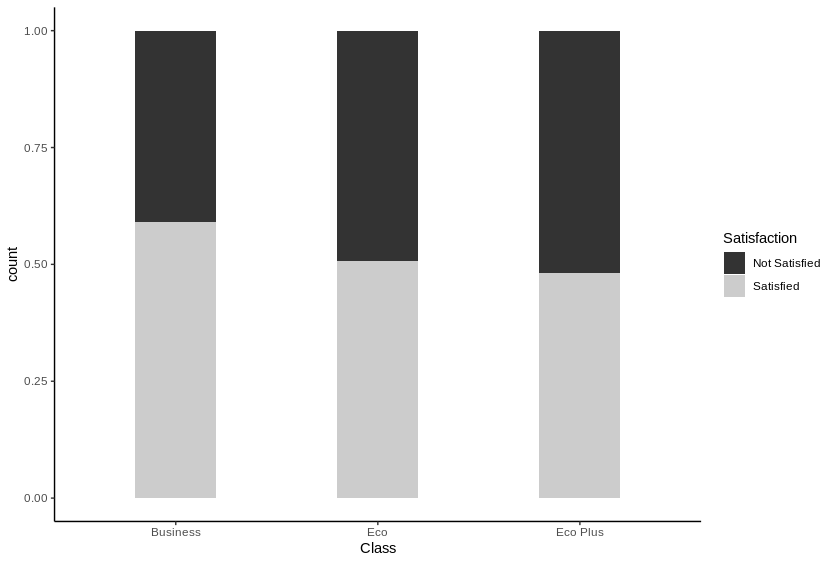
The customer ratings are higher for Gold and maximum for Silver. This holds true for both southeast and other airlines

**Satisfaction vs. Type of travel for Southeast and other airlines.**



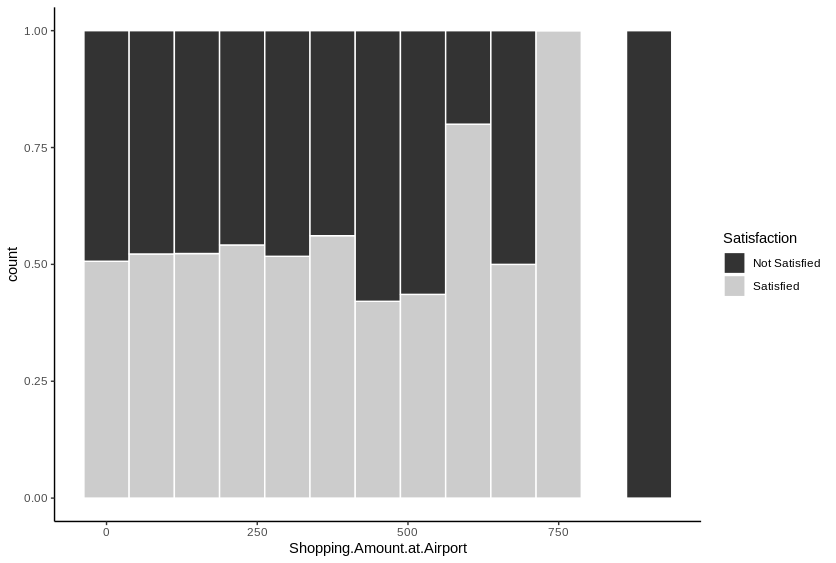
When customer’s travel for personal reasons, their satisfaction is the lowest (around 2.5)

Whereas, it is higher when the tickets are Mileage tickets and even higher for Business travel.



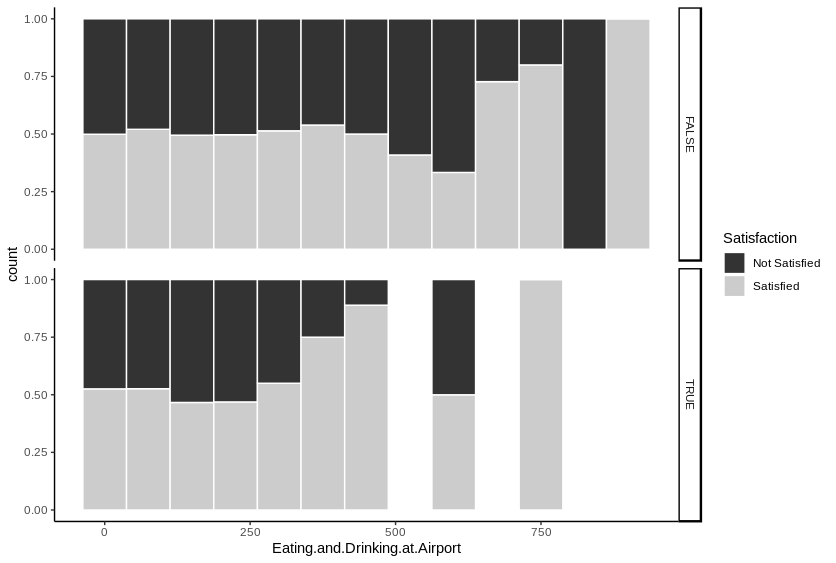
The satisfaction ratings are slightly higher for Business class compared to Eco and Eco plus.

[**BQ 7: How much does Shopping, eating and drinking at the airport affect customer satisfaction?**](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.147n2zr)



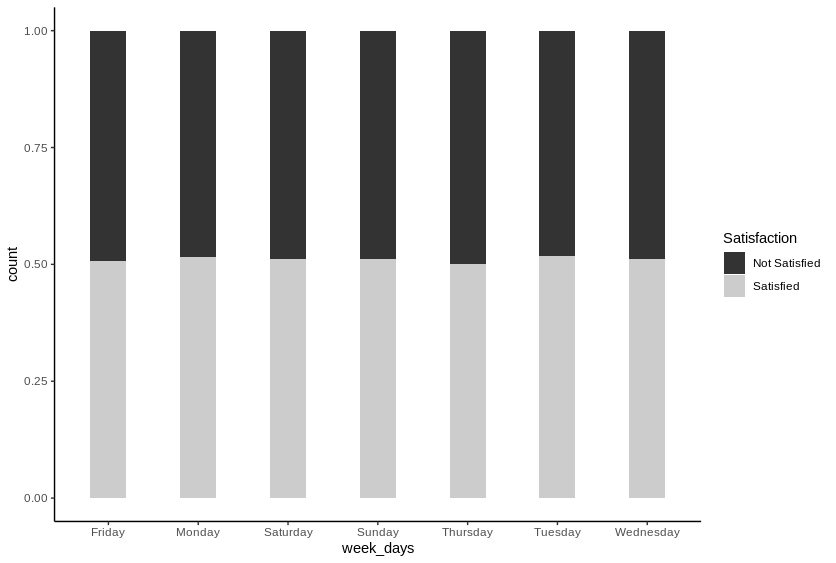
Customers who shop between 600-800 give a high customer satisfaction rating for other airlines

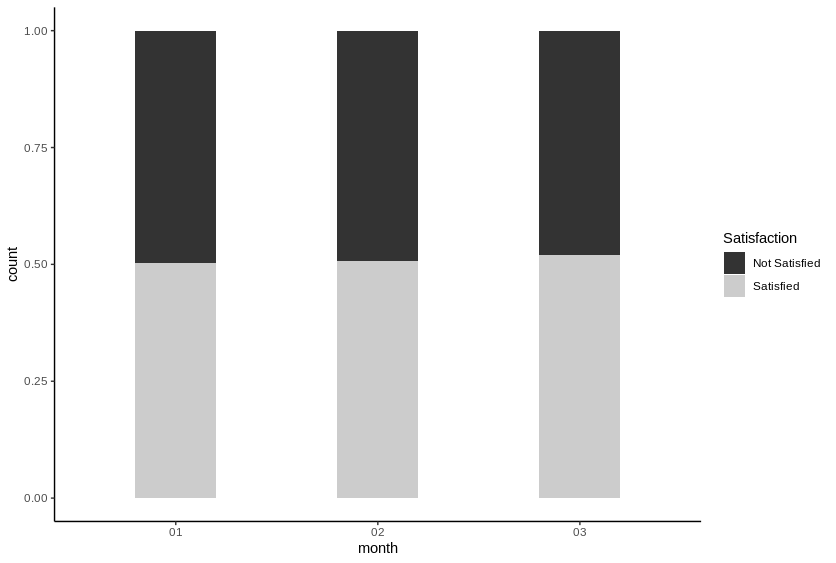
whereas those who shop for 500-600 give a high customer rating for southeast airlines.



It’s almost the same trend for eating and drinking at airport.

**B**[**Q 8:  Does date of travel (day, week, month) affect customer satisfaction?**](https://docs.google.com/document/d/171SSYVGzls6pyAVv3-7Bq1azb59-9y_RPTWHCtwRO2c/edit#heading=h.3o7alnk)





As shown in the graph, there isn’t a difference in the ratings of customer satisfaction. Therefore,

day of the month is not a factor which affects customer satisfaction.

**BQ 9: Do particular Origins and Destinations cause low customer satisfaction?**

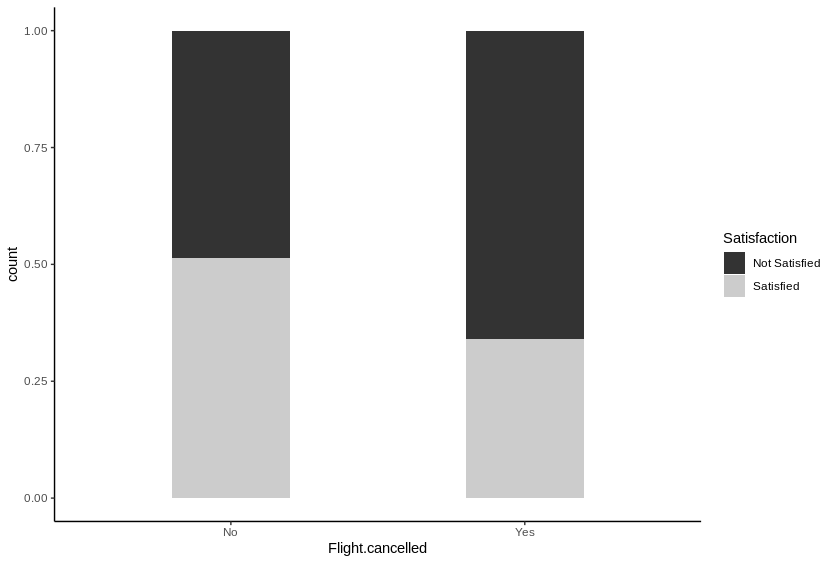


Customers travelling from WA,  NM, SC, OH, NE give lower ratings to south east.

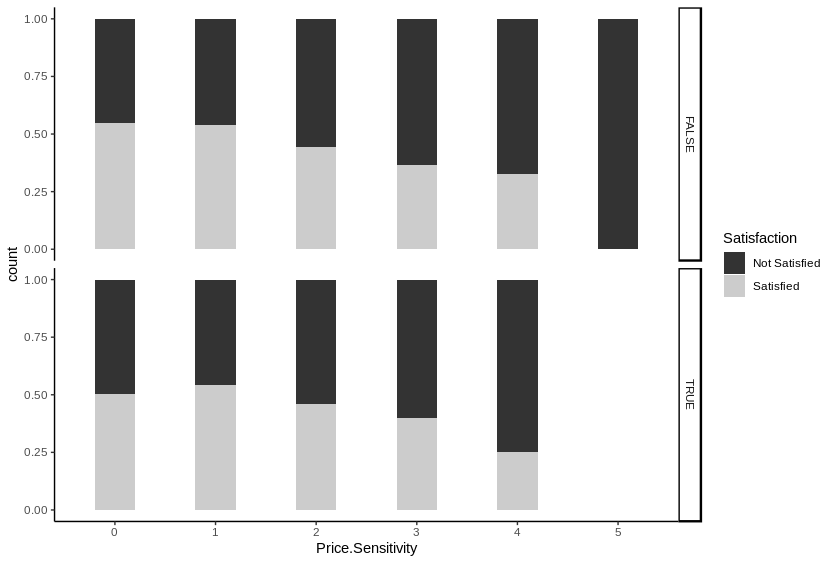


Customers traveling to OR,  UT, NE, IL, PA give lower rating.

**BQ 10: Do cancelled flights cause customers to give a low customer rating?**



Cancelled flights cause customers to give a lower satisfaction rating for both southeast and other airlines.

**BQ11: Does Price Sensitivity affect Customer Satisfaction?**  


Less price sensitive customers are in general more satisfied. For southeast specially, more price sensitive customers tend not to take southeast maybe because southeast have high prices.

**Use of modeling techniques & Visualizations**

Linear Models

summary(lm(Satisfaction ~ Airline.Status + Age + Gender +

           Price.Sensitivity + Year.of.First.Flight + No.of.Flights.p.a. +

              Type.of.Travel +

              Shopping.Amount.at.Airport +

              Class +

             Scheduled.Departure.Hour +

             Departure.Delay.in.Minutes +

             Arrival.Delay.greater.5.Mins, data = clean\_data ))

Residual standard error: 0.7218 on 85438 degrees of freedom

  (1568 observations deleted due to missingness)

Multiple R-squared:  0.4432, Adjusted R-squared:  0.4431

F-statistic:  4250 on 16 and 85438 DF,  p-value: < 2.2e-16

Logistic regression

happyCust <- 1

med <- median(clean\_data$Satisfaction)

happyCust[clean\_data$Satisfaction>=med] <- 3

happyCust[clean\_data$Satisfaction<med] <- 2

happyCust <- as.factor(happyCust)

clean\_data1 <- cbind(clean\_data,happyCust)

clean\_data1 <- clean\_data1[,-c(1)]

clean\_data1$Departure.Delay.in.Minutes[is.na(clean\_data1$Departure.Delay.in.Minutes)] <- mean(clean\_data1$Departure.Delay.in.Minutes,na.rm=T)

clean\_data1$Arrival.Delay.in.Minutes[is.na(clean\_data1$Arrival.Delay.in.Minutes)] <- mean(clean\_data1$Arrival.Delay.in.Minutes,na.rm=T)

clean\_data1$Flight.time.in.minutes[is.na(clean\_data1$Flight.time.in.minutes)] <- mean(clean\_data1$Flight.time.in.minutes,na.rm=T)

# origin city, destination, airline code, flights with other airlines,arrival and departure delay and day of month don't affect happiness significantly(high p values)

model3 <- glm(happyCust~Airline.Status+Age+Gender+Price.Sensitivity+No.of.Flights.p.a.+ Type.of.Travel+Shopping.Amount.at.Airport+Class+Arrival.Delay.greater.5.Mins+Eating.and.Drinking.at.Airport+southeast,family="binomial",data=clean\_data1)

summary(model3)

# # evaluation of logistic regression model

logtraindata <- clean\_data1

logtestdata <- test

happyCust <- 1

med <- median(logtestdata$Satisfaction)

happyCust[logtestdata$Satisfaction>=med] <- 3

happyCust[logtestdata$Satisfaction<med] <- 2

happyCust <- as.factor(happyCust)

logtestdata$Departure.Delay.in.Minutes[is.na(logtestdata$Departure.Delay.in.Minutes)] <- mean(logtestdata$Departure.Delay.in.Minutes,na.rm=T)

logtestdata$Arrival.Delay.in.Minutes[is.na(logtestdata$Arrival.Delay.in.Minutes)] <- mean(logtestdata$Arrival.Delay.in.Minutes,na.rm=T)

logtestdata$Flight.time.in.minutes[is.na(logtestdata$Flight.time.in.minutes)] <- mean(logtestdata$Flight.time.in.minutes,na.rm=T)

logpred <- predict(model3,logtestdata,type="response")

pos\_or\_neg <- ifelse(logpred > 0.5, 3, 2)

happyCusttest <- factor(pos\_or\_neg)

x <- table(happyCust,happyCusttest)

error <- (x[1,1]+x[2,2])/sum(x)

error

# 77% accuracy

Naïve Bayes

    After logistic regression, our team did another analysis using the Naive Bayes classifier. Naive Bayes is an algorithm used on complex datasets. To use Naive Bayes classifier, we loaded the e1071 package, which is the Naive Bayes package. The NA values on the data were removed and the median of the Satisfaciton was considered. After that, we made a separate column to classify whether the customer is happy or not.

Then, we used Naive Bayes to make the prediction on the selected predictors.

We then made a confusion matrix, and took the percentage of the matched values and predicted the accuracy of the prediction, which came up to 76.79%.

library(e1071)

clean\_data <- na.omit(clean\_data)

happyCust <- 1

med <- median(clean\_data$Satisfaction)

happyCust[clean\_data$Satisfaction>=med] <- 3

happyCust[clean\_data$Satisfaction<med] <- 2

happyCust <- as.factor(happyCust)

clean\_data1 <- cbind(clean\_data,happyCust)

clean\_data1 <- clean\_data1[,-c(1)]

nb <- naiveBayes(happyCust ~ Airline.Status + Age + Gender +

           Price.Sensitivity + Year.of.First.Flight + No.of.Flights.p.a. +

              Type.of.Travel +

              Shopping.Amount.at.Airport +

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

             Arrival.Delay.greater.5.Mins, data = clean\_data1)

summary(nb)

nb

------ OUTPUT For Naive Bayes--------------------

      Length Class  Mode

apriori  2 table  numeric

tables  12 -none- list

levels   2 -none- character

call     4 -none- call

Naive Bayes Classifier for Discrete Predictors

Call:

naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:

Y

      2        3

0.485563 0.514437

Conditional probabilities:

  Airline.Status

Y         Blue Gold   Platinum Silver

 2 0.81628756 0.05895719 0.02475283 0.10000242

 3 0.55883091 0.10604851 0.03979101 0.29532958

  Age

Y       [,1] [,2]

 2 49.78390 20.30972

 3 42.73826 12.87123

  Gender

Y      Female     Male

 2 0.6276197 0.3723803

 3 0.5064911 0.4935089

  Price.Sensitivity

Y       [,1]   [,2]

 2 1.326138 0.5700696

 3 1.230259 0.5180028

  Year.of.First.Flight

Y       [,1] [,2]

 2 2007.199 2.996252

 3 2007.207 2.959490

  No.of.Flights.p.a.

Y       [,1] [,2]

 2 23.41321 15.27507

 3 16.77218 12.57571

  Type.of.Travel

Y   Business travel Mileage tickets Personal Travel

 2      0.36524934      0.06715173 0.56759893

 3      0.85489060      0.08528600 0.05982340

  Shopping.Amount.at.Airport

Y       [,1] [,2]

 2 25.68725 52.67634

 3 27.49250 53.46744

  Class

Y     Business        Eco Eco Plus

 2 0.06884382 0.82063864 0.11051754

 3 0.09480025 0.80681740 0.09838235

  Scheduled.Departure.Hour

Y       [,1] [,2]

 2 12.92260 4.634437

 3 13.02437 4.580868

  Departure.Delay.in.Minutes

Y       [,1] [,2]

 2 17.73618 42.04413

 3 12.31951 34.72743

  Arrival.Delay.greater.5.Mins

Y          no yes

 2 0.5776064 0.4223936

 3 0.7161697 0.2838303

predP <- predict(nb, newdata = clean\_data1, type = "class")

matrixSat <- table(predP,clean\_data1$happyCust)

sumMatrix <- matrixSat[1,1] + matrixSat[2,2]

percSat <- sumMatrix/sum(matrixSat)

percSat <- percSat\*100

percSat

-----OUTPUT for Accuracy------

[1] 76.9396

#76.79% Accuracy

Association rule mining

For Association Rules Mining, Firstly the variables were converted to factors which is the foremost step.

data$Flight.date=as.Date(data$Flight.date, format="%m/%d/%Y")

data$week\_days=weekdays(data$Flight.date)

data$month<- strftime(data$Flight.date, "%m")

numbertoCategory<-function(vec){

  #getting the 40th and 60th percentile of the data

  q <- quantile(vec, c(0.4, 0.6),na.rm=TRUE)

  #creating a vector with 'Average' value of length of vec

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

  #values lower than 40th percentile are replaced by 'Low'

  vBuckets[vec < q[1]] <- "Low"

  #values greater than 60th percentile are replaced by 'high'

  vBuckets[vec > q[2]] <- "High"

  vBuckets[vec==999]<-"Not Available"

  return(as.factor(vBuckets))

}

data$Age=numbertoCategory(data$Age)

data$No.of.Flights.p.a.=numbertoCategory(data$No.of.Flights.p.a.)

data$X..of.Flight.with.other.Airlines=numbertoCategory(data$X..of.Flight.with.other.Airlines)

data$Shopping.Amount.at.Airport=numbertoCategory(data$Shopping.Amount.at.Airport)

data$Departure.Delay.in.Minutes=numbertoCategory(data$Departure.Delay.in.Minutes)

data$Arrival.Delay.in.Minutes=numbertoCategory(data$Arrival.Delay.in.Minutes)

data$Flight.time.in.minutes=numbertoCategory(data$Flight.time.in.minutes)

data$Flight.Distance=numbertoCategory(data$Flight.Distance)

data$Arrival.Delay.in.Minutes[is.na(data$Arrival.Delay.in.Minutes)]<-999

data$Departure.Delay.in.minutes[is.na(data$Departure.Delay.in.Minutes)]=999

data$Flight.time.in.Minutes[is.na(data$Arrival.Delay.in.Minutes)]=999

data$Eating.and.Drinking.at.Airport=numbertoCategory(data$Eating.and.Drinking.at.Airport)

data$Arrival.Delay.in.Minutes[data$Flight.cancelled=='Yes']

data$Satisfaction[data$Satisfaction>=4]="Satisfied"

data$Satisfaction[data$Satisfaction<4]="Not Satisfied"

data$Satisfaction=as.factor(data$Satisfaction)

data$Price.Sensitivity=as.factor(data$Price.Sensitivity)

data$No..of.other.Loyalty.Cards=as.factor(data$No..of.other.Loyalty.Cards)

data$Day.of.Month=as.factor(data$Day.of.Month)

data$Scheduled.Departure.Hour=as.factor(data$Scheduled.Departure.Hour)

data$week\_days=as.factor(data$week\_days)

data$month=as.factor(data$month)

data=data[,-15]

facna=addNA(data$Arrival.Delay.in.Minutes)

levels(facna) <- c(levels(data$Arrival.Delay.in.Minutes), 'Not Available')

data$Arrival.Delay.in.Minutes<-facna

facna=addNA(data$Departure.Delay.in.Minutes)

levels(facna) <- c(levels(data$Departure.Delay.in.Minutes), 'Not Available')

data$Departure.Delay.in.Minutes<-facna

facna=addNA(data$Flight.time.in.minutes)

levels(facna) <- c(levels(data$Flight.time.in.minutes), 'Not Available')

data$Arrival.Delay.in.Minutes<-facna

data$southeast=as.factor(data$southeast)

data$Year.of.First.Flight=as.factor(data$Year.of.First.Flight)

data<-data.frame(sapply(data,as.factor))

Now, as all variables are factors, We can apply association rule mining to get the most associated rules in the data set

dataX <- as(data,"transactions")

ruleset <- apriori(dataX, parameter=list(support = 0.3,confidence = 0.3))

#subsetting to get only those rules that result into high overallCustSat

sub<-subset(ruleset, subset = rhs %in% "Satisfaction=Satisfied")

#inspect to see those rules

inspect(sub)

dataX <- as.data.frame(dataX)

# Output

# lhs                                 rhs                    support confidence  lift count

# [1]  {}                           => {Satisfaction=Satisfied} 0.5106351  0.5106351 1.0000000 44437

# [2]  {Shopping.Amount.at.Airport=Average} => {Satisfaction=Satisfied} 0.3031957  0.4951861 0.9697454 26385

# [3]  {Type.of.Travel=Business travel} => {Satisfaction=Satisfied} 0.4348850  0.7098378 1.3901076 37845

# [4]  {Departure.Delay.in.Minutes=Average} => {Satisfaction=Satisfied} 0.3323374  0.5401857 1.0578702 28921

# [5]  {Arrival.Delay.greater.5.Mins=no} => {Satisfaction=Satisfied} 0.3659033  0.5575556 1.0918865 31842

# [6]  {Price.Sensitivity=1}            => {Satisfaction=Satisfied} 0.3652827  0.5386792 1.0549201 31788

# [7]  {Class=Eco}                      => {Satisfaction=Satisfied} 0.4112936  0.5057224 0.9903793 35792

# [8]  {southeast=FALSE}                => {Satisfaction=Satisfied} 0.4720706  0.5094433 0.9976660 41081

# [9]  {Flight.cancelled=No}            => {Satisfaction=Satisfied} 0.5042920  0.5140203 1.0066294 43885

# [10] {Type.of.Travel=Business travel,

# Arrival.Delay.greater.5.Mins=no} => {Satisfaction=Satisfied} 0.3040805  0.7604679 1.4892588 26462

# [11] {Price.Sensitivity=1,

# Type.of.Travel=Business travel} => {Satisfaction=Satisfied} 0.3138596  0.7208118 1.4115985 27313

# [12] {Type.of.Travel=Business travel,

# Class=Eco}                      => {Satisfaction=Satisfied} 0.3489307  0.7045898 1.3798302 30365

# [13] {Type.of.Travel=Business travel,

# southeast=FALSE}                => {Satisfaction=Satisfied} 0.4019512  0.7083494 1.3871928 34979

# [14] {Type.of.Travel=Business travel,

# Flight.cancelled=No}            => {Satisfaction=Satisfied} 0.4306678  0.7122522 1.3948360 37478

# [15] {Departure.Delay.in.Minutes=Average,

# Arrival.Delay.greater.5.Mins=no} => {Satisfaction=Satisfied} 0.3035175  0.5564615 1.0897438 26413

# [16] {Departure.Delay.in.Minutes=Average,

# southeast=FALSE}                => {Satisfaction=Satisfied} 0.3036209  0.5399407 1.0573905 26422

# [17] {Departure.Delay.in.Minutes=Average,

# Flight.cancelled=No}            => {Satisfaction=Satisfied} 0.3260517  0.5465367 1.0703077 28374

# [18] {Arrival.Delay.greater.5.Mins=no,

# southeast=FALSE}                => {Satisfaction=Satisfied} 0.3366351  0.5568862 1.0905757 29295

# [19] {Flight.cancelled=No,

# Arrival.Delay.greater.5.Mins=no} => {Satisfaction=Satisfied} 0.3595601  0.5641599 1.1048200 31290

# [20] {Price.Sensitivity=1,

# southeast=FALSE}                => {Satisfaction=Satisfied} 0.3379911  0.5377148 1.0530314 29413

# [21] {Price.Sensitivity=1,

# Flight.cancelled=No}            => {Satisfaction=Satisfied} 0.3610195  0.5422053 1.0618253 31417

# [22] {Class=Eco,

# southeast=FALSE}                => {Satisfaction=Satisfied} 0.3801064  0.5043916 0.9877730 33078

# [23] {Class=Eco,

# Flight.cancelled=No}            => {Satisfaction=Satisfied} 0.4056629  0.5091145 0.9970221 35302

# [24] {Flight.cancelled=No,

# southeast=FALSE}                => {Satisfaction=Satisfied} 0.4659458  0.5127336 1.0041096 40548

# [25] {Price.Sensitivity=1,

# Type.of.Travel=Business travel,

# Flight.cancelled=No}            => {Satisfaction=Satisfied} 0.3110902  0.7235214 1.4169049 27072

# [26] {Type.of.Travel=Business travel,

# Class=Eco,

# southeast=FALSE}                => {Satisfaction=Satisfied} 0.3223516  0.7025470 1.3758298 28052

# [27] {Type.of.Travel=Business travel,

# Class=Eco,

# Flight.cancelled=No}            => {Satisfaction=Satisfied} 0.3452421  0.7072338 1.3850082 30044

# [28] {Type.of.Travel=Business travel,

# Flight.cancelled=No,

# southeast=FALSE}                => {Satisfaction=Satisfied} 0.3979063  0.7107786 1.3919500 34627

# [29] {Flight.cancelled=No,

# Arrival.Delay.greater.5.Mins=no,

# southeast=FALSE}                => {Satisfaction=Satisfied} 0.3305103  0.5634305 1.1033916 28762

# [30] {Price.Sensitivity=1,

# Flight.cancelled=No,

# southeast=FALSE}                => {Satisfaction=Satisfied} 0.3338773  0.5411421 1.0597431 29055

# [31] {Class=Eco,

# Flight.cancelled=No,

# southeast=FALSE}                => {Satisfaction=Satisfied} 0.3746711  0.5076525 0.9941591 32605

# [32] {Type.of.Travel=Business travel,

# Class=Eco,

# Flight.cancelled=No,

# southeast=FALSE}                => {Satisfaction=Satisfied} 0.3188238  0.7052080 1.3810410 27745

Actionable Insights / Overall interpretation of results

|  |  |
| --- | --- |
| **Finding** | **Recommendation** |
| Customers with cancelled flights tend to give lower ratings | Reduce frequency of cancelled flights |
| Price sensitive customers are unsatisfied | Reduce prices of flights |
| Customers who  travel for personal reasons are unsatisfied | Provide family plans to reduce flight cost |
| Customers  travelling  to certain states  are unsatisfied | Improve facilities for  OR, UT, NE, IL, PA, NC |
| Customers in a higher age range are unsatisfied | Improve facilities for older customers |
| Customers who travel frequently give lower ratings | Provide incentives to frequent travellers |
| Customers who  travel by blue  status flights  are unsatisfied | Improve amenities for blue status flights |
| Customers travelling from certain states are unsatisfied | Improve facilities for NM, WA, NE, SC, OH |

**Conclusion**

The data analyzed for south east airlines provided results that helped determine

customers with low satisfaction and thus insights and recommendations were

provided to help mitigate the situation.

**MIDST Usage/Info**

 We used the MIDST tool as instructed by our professor to divide our code into chunks and assign tasks to each team member. To approach our problem statement, we divided our code into necessary chunks according to each task as explained earlier. Our MIDST link with screenshots is given below:

**MIDST link:** <https://midst.syr.edu/project/team55>

**MIDST Screenshots:**Our network is given below:

