School of Information Studies
SYRACUSE UNIVERSITY

AIRLINE CUSTOMER SATISFACTION ANALYSIS

Ryan French, Rebecca LoSurdo, Tanushree Shetty, Abhishek Singh, Zefeng Zhang

IST 687
Prof Jeff Saltz

Introduction

The purpose of this project was to utilize customer flight data to determine drivers that influence the satisfaction level of airline clientele with the aim of increasing overall customer satisfaction. The provided flight data came in the form of customer surveys that included each customer's rating of their experience along with various other variables pertaining to the customer's demographics, behavior, history and specific flight information. Through analysis and data modelling, we were able to identify the variables that were most consistently correlated with customer satisfaction so that these areas could be a source of focus during the implementation of new policies aimed at improving the overall customer experience.

Business Questions

1. What variables have the most significant impact on the satisfaction of airline customers?
2. Which of these variables are difficult or impossible to manipulate (ex: what airport a customer flies out of)?
3. Which of these variables are potentially manipulatable to further increase this satisfaction (ex: how quickly a customer accumulates miles) and what possible recommendations can be made from these insights?
4. What topics could be useful if researched further?

Understanding the Data

To understand the data we split all the available variables into five groups, grouping logically by placing similar variables in a group. We then explored the variables individually and how they affect our target variable i.e. Satisfaction. Our five groups are as follows:

1. **Airline Analysis**: Airline Status, % of Flight with other Airlines, Airline Code, Airline Name, No. of other Loyalty Cards, Class, Type of Travel

2. **Flight Analysis**: Day of Month, Flight date, Flight time in minutes, Flight Distance

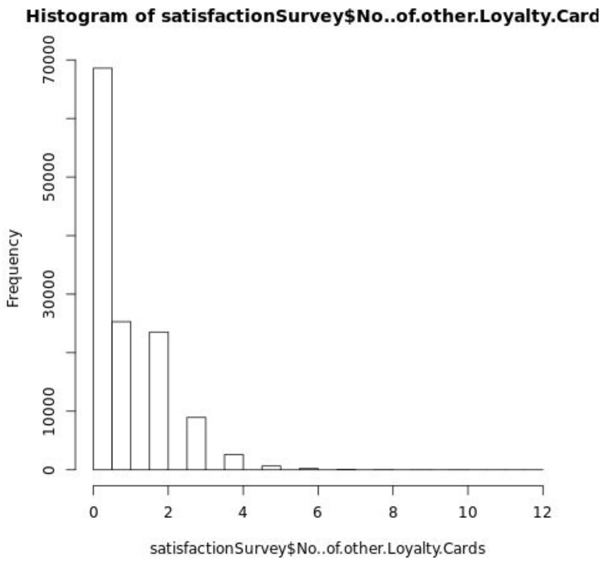
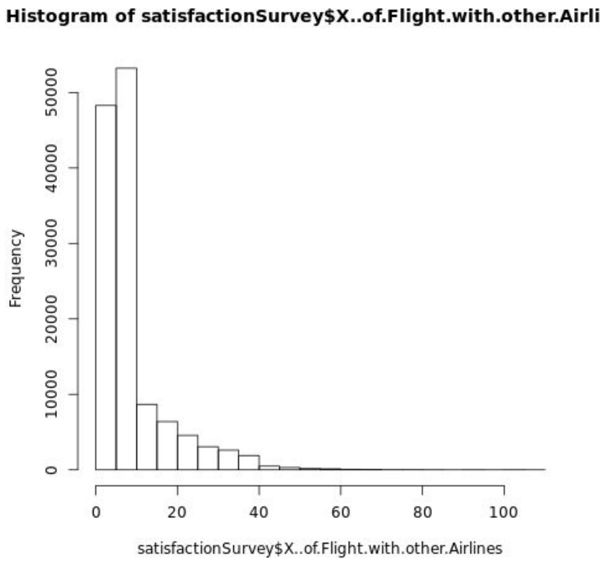
3. **Delay Analysis**: Scheduled Departure Hour, Departure Delay in Minutes, Arrival Delay in Minutes, Flight canceled, Arrival Delay greater 5 Mins

4. **Geographic Analysis**: Origin State, Destination State

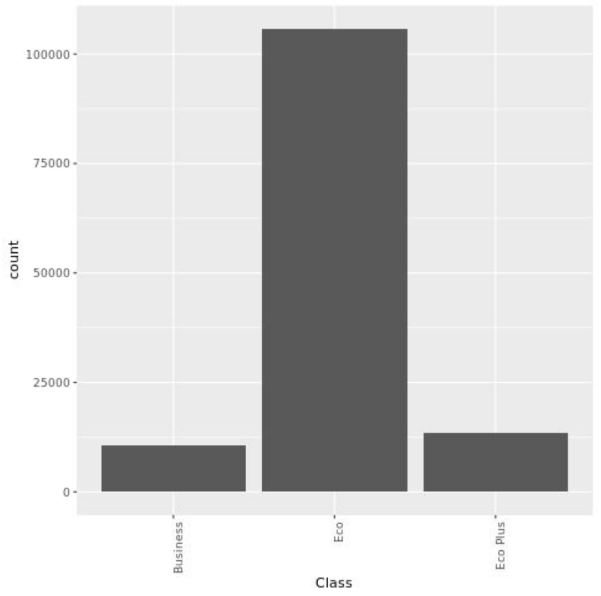
5. **Passenger Demographics**: Age, Gender, Year of First Flight, No of Flights p.a., Shopping Amount at Airport, Price Sensitivity, Eating and Drinking at Airport

1. Airline Analysis:

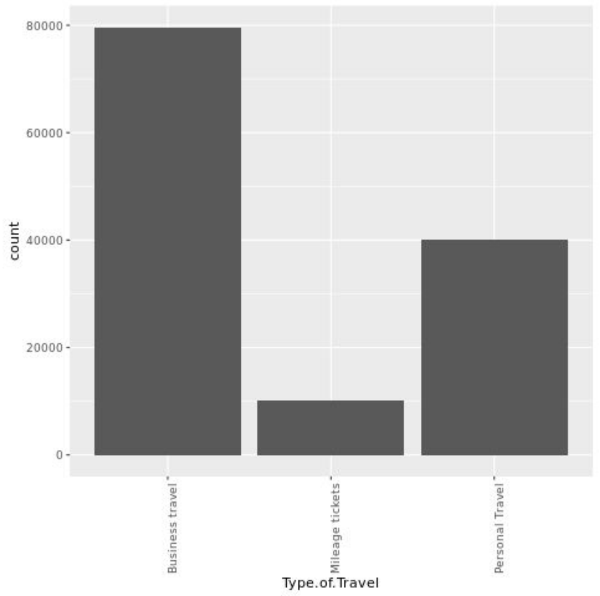
When initially working to understand the data, we utilized histogram and bar graph visualizations in order to better understand the demographics of the customers and the areas with the most potential for growth.

	
Frequency vs Number of Loyalty Cards	Frequency vs Number of Flights with other Airlines

From these two figures, it can be deduced that the majority of the customers interviewed are either relatively new flyers or infrequent flyers, the majority having flown less than ten times with other airlines and having a relatively few number of other loyalty cards. This is useful in framing the mindset of the customers and their flying experience.

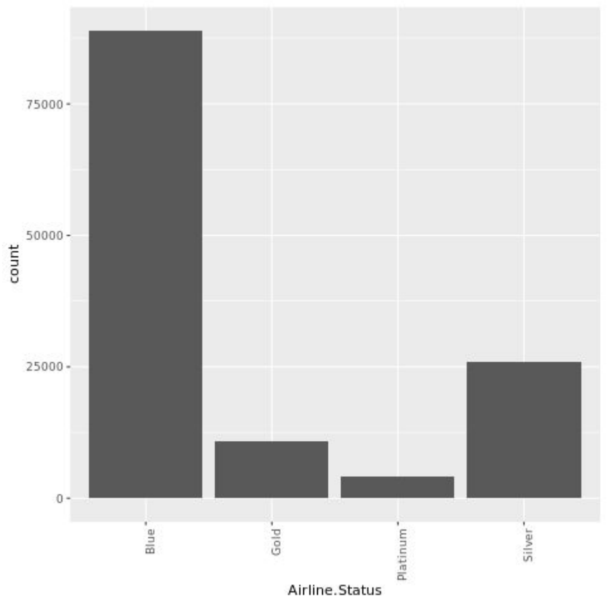


Count vs Travel Class



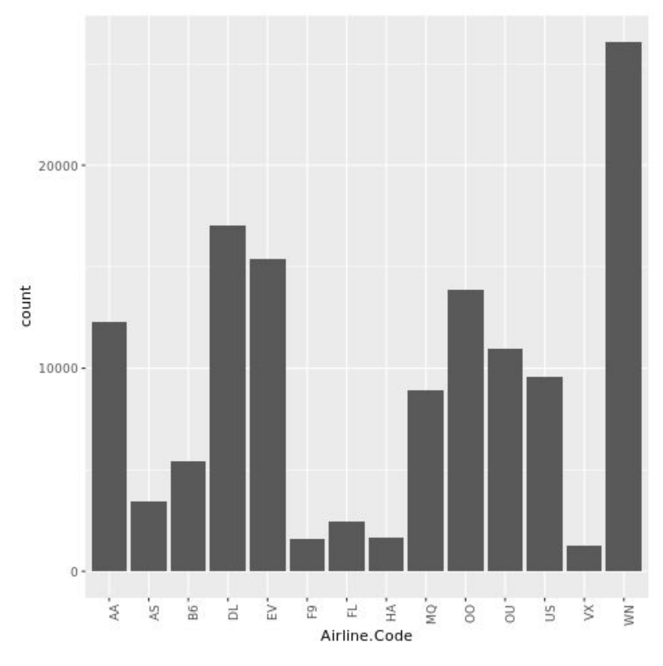
Count vs Travel Type

From these figures we can deduce that the vast majority of customers are traveling for the sake of business but in economy seating allowing us to further understand the largest passenger demographic; those flying economy for business.



Count vs Airline Status

Airline Status provides an additional facet to this demographic as again we see a skew in the data towards the lowest level of membership, seeming to confirm that passengers are likely primarily new to flying or fly infrequently.



Count vs Airline Code

We additionally looked at the number of respondents for each Airline Code, however this proved to be largely arbitrary information as this overall analysis is not tied to a single airline nor is it possible in some cases to change the airline with which customers are flying.

In this way, preliminary analysis of airline passenger attributes helped us to both determine potentially relevant variables such as status, class, and travel type as well as helping to frame the mindset of customers through their number of loyalty cards and flights with other airlines. Finally, the airline code variable seemed unlikely to be useful due to its rather immobile nature and the scope of this project extending beyond any singular airline.

2. Flight Analysis:

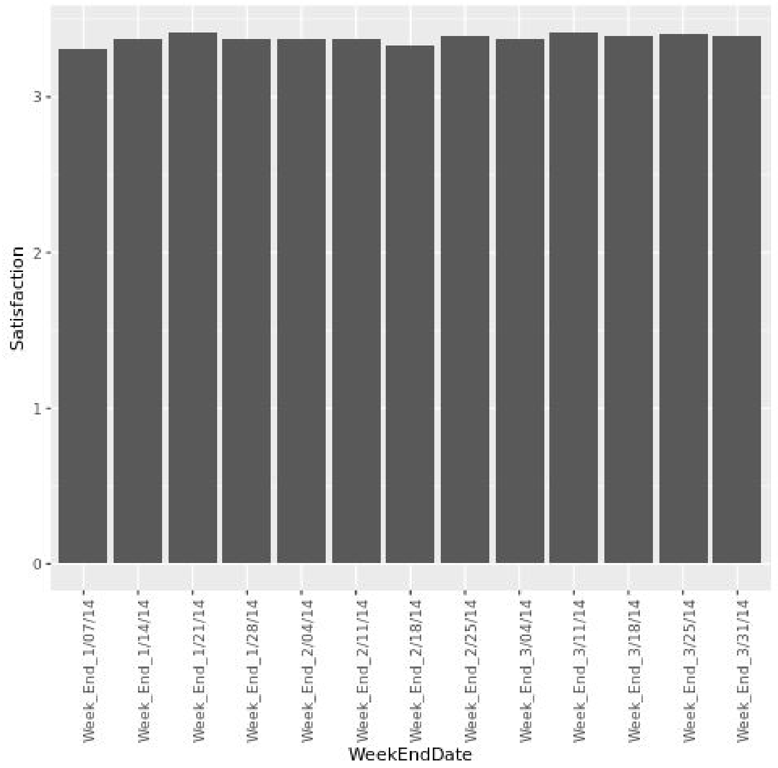
Performing exploratory analysis on the portion of our dataset that pertained to flight information included examining the relationships between Satisfaction and each of the following variables: Day of the Month, Flight Date, Flight Time (in minutes) and Flight Distance. We began by comparing average Satisfaction ratings across each variable while normalizing for the fact that some flight dates, times and distances are more frequent than others. If any particular instances of the variables were shown to positively or negatively affect Satisfaction, we would then investigate these instances further (frequency, central tendency, etc.)

As there were only 31 possible unique values for the Day of the Month, we were able to create a bar graph to examine the relationship between each day of the month and overall satisfaction. To account for the fact that more flights may occur during certain times of the month, we first

normalized the data by grouping it by day, reporting the average Satisfaction per group, and coercing the results into a dataframe to prepare them for processing.
Using the ggplot2 package, we then plotted the results to provide a visualization that would allow us to observe any obvious trends relating to the days. Upon viewing the results, we concluded that there did not seem to be any correlation between Satisfaction and Day of the Month. This preliminary hypothesis would be further validated with modelling techniques later in our process.

Next, we performed some exploratory analysis on the flight dates themselves. As we had three months of data which translated to approximately 90 unique dates, it was necessary to group these into related buckets before trying to create a bar plot that would be of any use. To accomplish this, we created a function that put Flight Date into buckets that were each one week in length (with the exception of the final bucket which was shorted by the months end). This function also required some formatting work to account for the specific date-time format found in the dataset. In order to catch any values that did not fall into our pre-defined categories, we created a default value (“Error”) that would be returned in the event that a particular date didn’t get assigned to a date range.

We were then able to store the results of the function in a new column in the dataset so that we could find the average Satisfaction rating for each week, essentially normalizing for heavy travel periods and allowing us to see if there were any trends in Satisfaction changes based solely on the calendar.
We used a similar plotting method to that of Day of the Month but as our group names were elongated, we adjusted the labels on the x-axis for ease of reading. Similar to Day of the Month, we did not view any correlation between Satisfaction and Flight Date during this preliminary analysis.



Next, we analyzed the Flight Time to see if we could find any correlation between time in the air and Satisfaction. In order to have a reasonable number of groupings to work with, we created a function that put Flight Time into buckets of 100 minute intervals. As we did with the categorization of Flight Date, we first created a default group labelled “Error” so that if any values did not fall into our preset categories, we would be able to identify them.

Once we created the function, we were able to pass the Flight Time data into it and store the results in a variable for validation. In validating, we found that there were some entries that the function assigned the “Error” label to meaning that they did not fall into any of the predefined minute intervals. We confirmed that this was because there are NA entries for Flight Time in some of the surveys which we would later address in the data cleaning phase of the project.

In the meantime, we continued to prepare the data for preliminary exploration by creating a new column within the original dataframe that defines our new groups:

```
Flight.minute.bins <-
defineMinuteBuckets(satisfactionSurvey$Flight.time.in.minutes)
satisfactionSurvey$Flight.minute.bins <- Flight.minute.bins[,1]
```

In order to normalize for the fact that some Flight Time values occur more often than others, we looked at the average Satisfaction for each group rather than the sum. We then prepared these grouped averages for plotting by coercing them into a dataframe and removing the “Error” grouping. This was important because the “NA” values would not be able to be averaged which would prevent us from plotting the rest of the data. For the sake of our current work, we disregarded this issue although we would revisit it later. From there, we plotted the results. We saw no obvious trends or interesting correlations between any of the groups of Flight Time and Satisfaction; they all looked to have relatively the same average Satisfaction rating.

Finally, we examined any potential effect that Flight Distance might have on Satisfaction. Similar to the above steps, we began by creating a function that would group the Flight Distances into 500 mile intervals, ensuring that any Flight Distances outside of our ranges would show as “Error” for further investigation. We then ran our data through the function and stored the results in a variable for validation before creating a new column in the original dataframe that defined the new categories of Flight Distance.

```
Flight.distance.bins <-
defineMileBuckets(satisfactionSurvey$Flight.Distance)
satisfactionSurvey$Flight.distance.bins <-
Flight.distance.bins[,1]
```

To normalize for more common Flight Distances, we took the average Satisfaction for each interval and coerced it into a dataframe for plotting, similar to the procedures for Day of the Month, Flight Date and Flight Time. In plotting the results, we did not see any cases where a particular grouping of distances impacted the average Satisfaction in any meaningful way.
As Day of the Month, Flight Date, Flight Time and Flight Distance did not have any direct impact on Satisfaction, we concluded our analysis without further need to investigate frequencies of the instances of any of these variables.

3. Delay analysis:

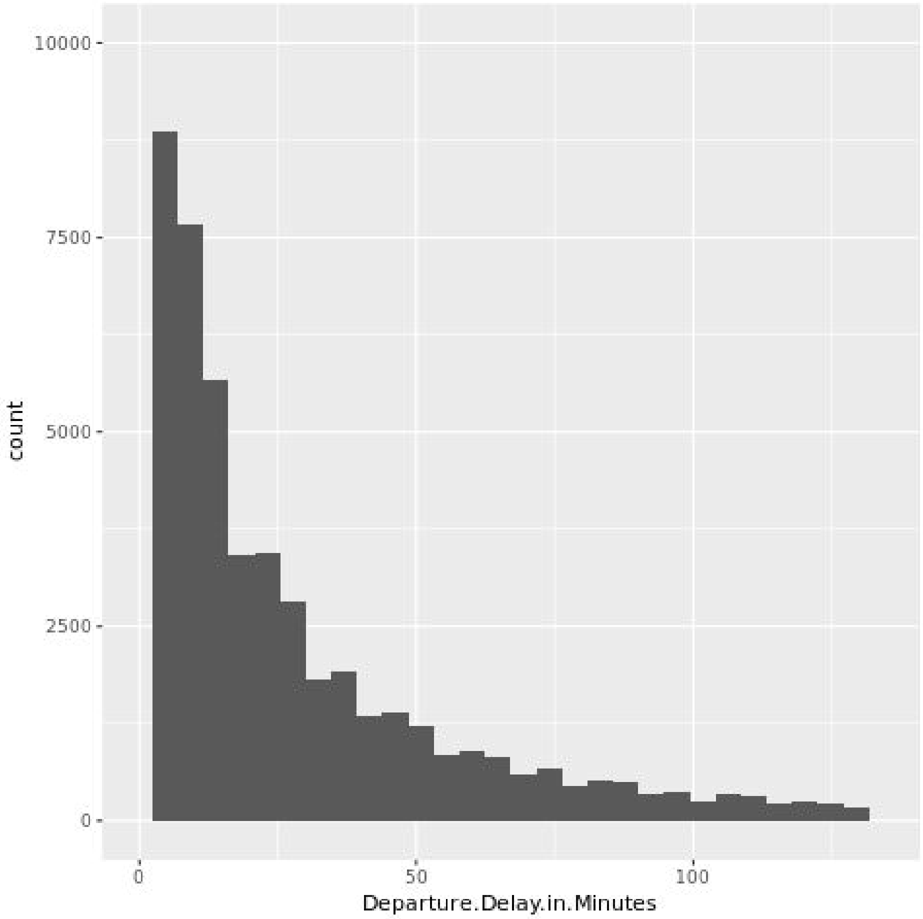
Here we discuss our analysis of the variables related to flight delay which are Scheduled Departure Hour, Departure Delay in Minutes, Arrival Delay in Minutes, Flight canceled, Arrival Delay Greater Than 5 Mins.
To start we performed univariate analysis to study the distributions of variables. First we explored the departure hour variable which contains the specific time for flight departure.
After seeing that it ranged from 1 AM to 23 PM we further investigated how this variable was distributed. We saw that most of our data had departure times between 5 AM and 8 PM and the general shape of a normal distribution curve.
Next we examined the Departure Delay in Minutes variable which described the delay for each flights in minutes. We noticed that there were approximately two thousand NA’s in this variable

and it seemed to be skewed towards the third and fourth quartiles. There was also a huge difference between the third quartile and the maximum value hinting at the presence of outliers.

```
ddplot <- ggplot(aes(Departure.Delay.in.Minutes), data =
subset(satisfactionSurvey,!is.na(satisfactionSurvey$Departure.Delay
.in.Minutes))) +
  geom_histogram()
```

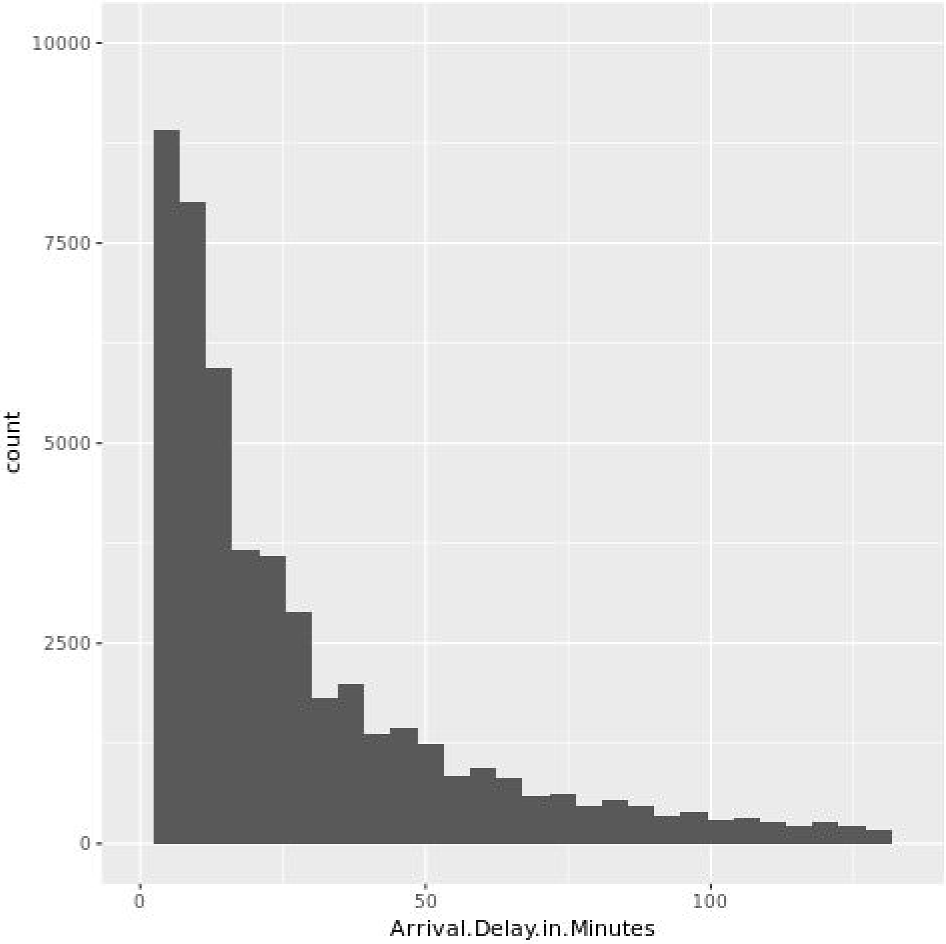
The plot generated by the code above confirmed our theory, as the data was indeed skewed towards the right of our distribution, hence why some of the flights had a delay of more than 500 minutes. We created another plot which showed a slice from the 98th percentile to 100th and also fixed our Y axis accordingly.

```
ddplot + scale_x_continuous(limits =
c(0,quantile(delay.df$Departure.Delay.in.Minutes,probs = 0.98,na.rm
= TRUE))) + scale_y_continuous(limits = c(0,10000))
```



This gave a clearer picture showing the distribution of delay variable. Most flights were delayed by less than an hour with very few flights being delayed over an hour. Further analysis showed that there were 19 flights with a delay of more than 12 hours and one with delay of more than a day which drives the histogram towards the right; these flights being ones that could potentially get rescheduled for the next day. Solutions to our problem with NA values would be dealt with later.

Next we investigated a similar variable; Arrival Delay in Minutes. The initial look showed us that this variable had a similar distribution as the previous one as it was skewed to the right and contained many NA values.
The above plot showed a very similar distribution as Departure Delay. To take a closer look we created the following plot.



Thus we can deduce similar conclusions as hypothesized about this variable.

Next we looked at the Flights Cancelled variable. We saw that Flights Cancelled was a categorical variable taking values 'YES' and 'NO'. We looked at the proportions for delayed flights as follows:

```
table(delay.df$Flight.cancelled)
➞No    Yes
127488   2401
```

There were around 2000 flights that had been cancelled. However, we had seen a similar number before during our initial analysis. We suspected that these NA values may be related to the ones we had previously encountered in the Arrival Delay variable. To inspect this, we performed the following analysis:

```
unique(delay.df[which(delay.df$Flight.cancelled=='Yes'),'Arrival.De
lay.in.Minutes'])
➞NA
sum(delay.df[which(delay.df$Flight.cancelled=='Yes'),'Arrival.Delay
.in.Minutes'],na.rm = TRUE)
➞0
```

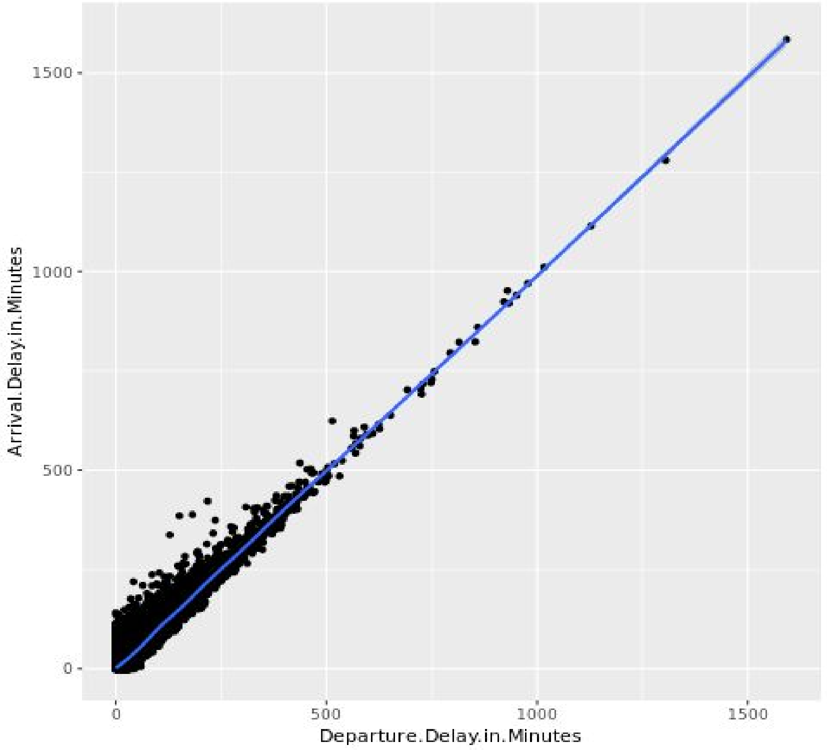
The first command showed us that the only delay values for cancelled flights were NA’s while the second command confirmed this. We saw there are only NA values which makes logical sense as for a cancelled flight, delay variable didn’t make sense. Further analysis showed that 98% of our data was for flights that were on time. We planned to subset our data using delay while conducting regression analysis.

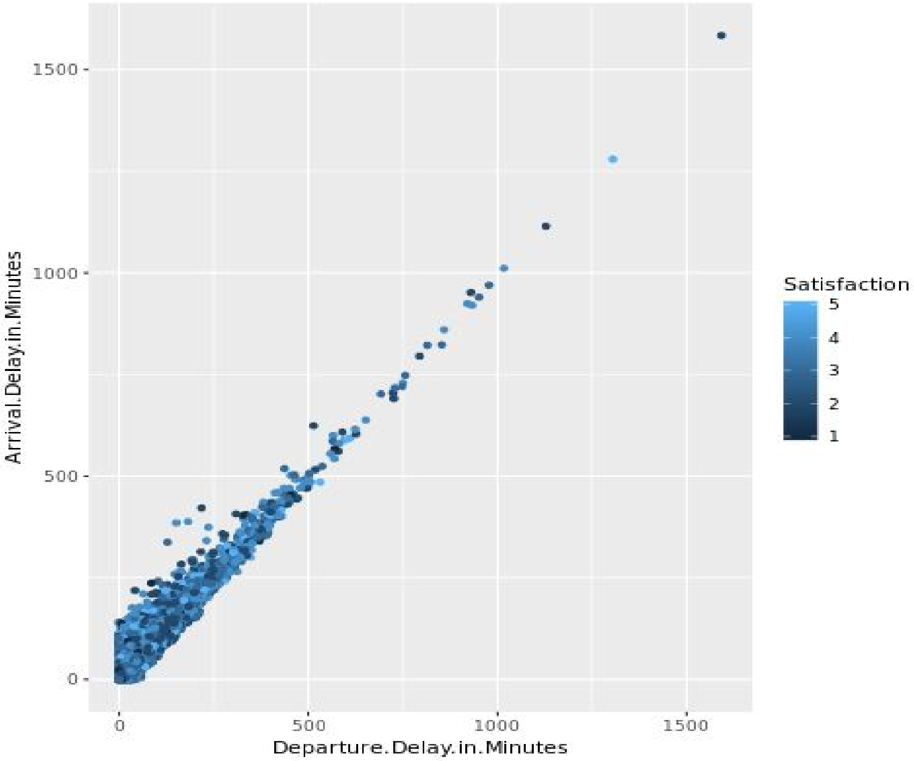
Other variables like Arrival Delay Greater Than 5 Minutes were categorical, indicating if the delay is greater than 5 minutes. This didn’t seem to be particularly useful as we already had the delay value for flights. 65% of flights in our data had a delay of less than 5 mins as observed earlier.

Next we looked at the bivariate relationships in our data. First we considered two similar variables: Arrival Delay and Departure Delay. As previously observed from our analysis, both these delay variables had a similar distribution so we further examined how these variables are related by creating a scatter plot.
We saw a strong linear relationship between the two as expected as, if a flight's departure is delayed, its arrival time is also affected. This linear relationship can be highlighted as follows:

Checking the correlation coefficient we obtained a value of 0.965 which confirmed that these are highly correlated variables and both of them need not be included in our model to avoid redundancy of data.

Further we looked at how delay affects our satisfaction ratings by coloring the previous plot using the Satisfaction variable:





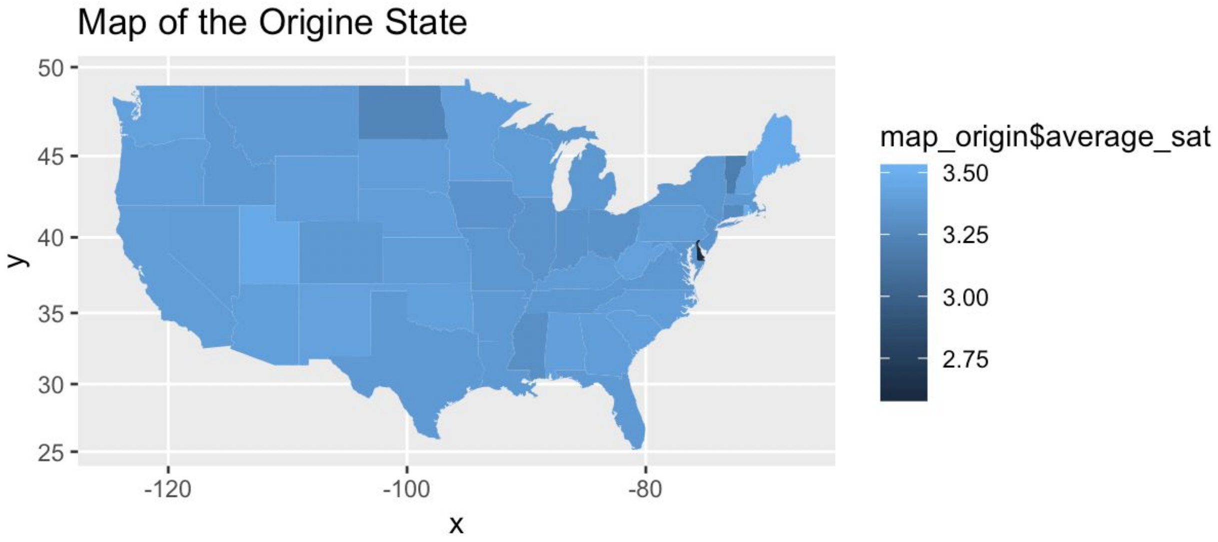
However a clear trend could not be observed here, contrary to the expectation that higher delay time may result in low satisfaction value. This hinted towards the fact that customer satisfaction was not solely dependent on delay, but also on other factors. For example, a customer may experience a higher delay but may be still satisfied due to other reasons.

Next we created a model which tries to predict customer satisfaction using only delay variables. We considered only flights which weren't cancelled for our model. As expected it didn’t perform well, as Adjusted R square was only 0.0075 which was evident from the last plot itself. We also implemented a different model with only one variable i.e. the departure delay as its predictor variable, dropping arrival delay as they were highly correlated. We also saw that the scheduled hour did not play a big role in predicting satisfaction. As expected our Adjusted R square did not drop very much, hence this variable may be considered as a key variable with respect to delay analysis.

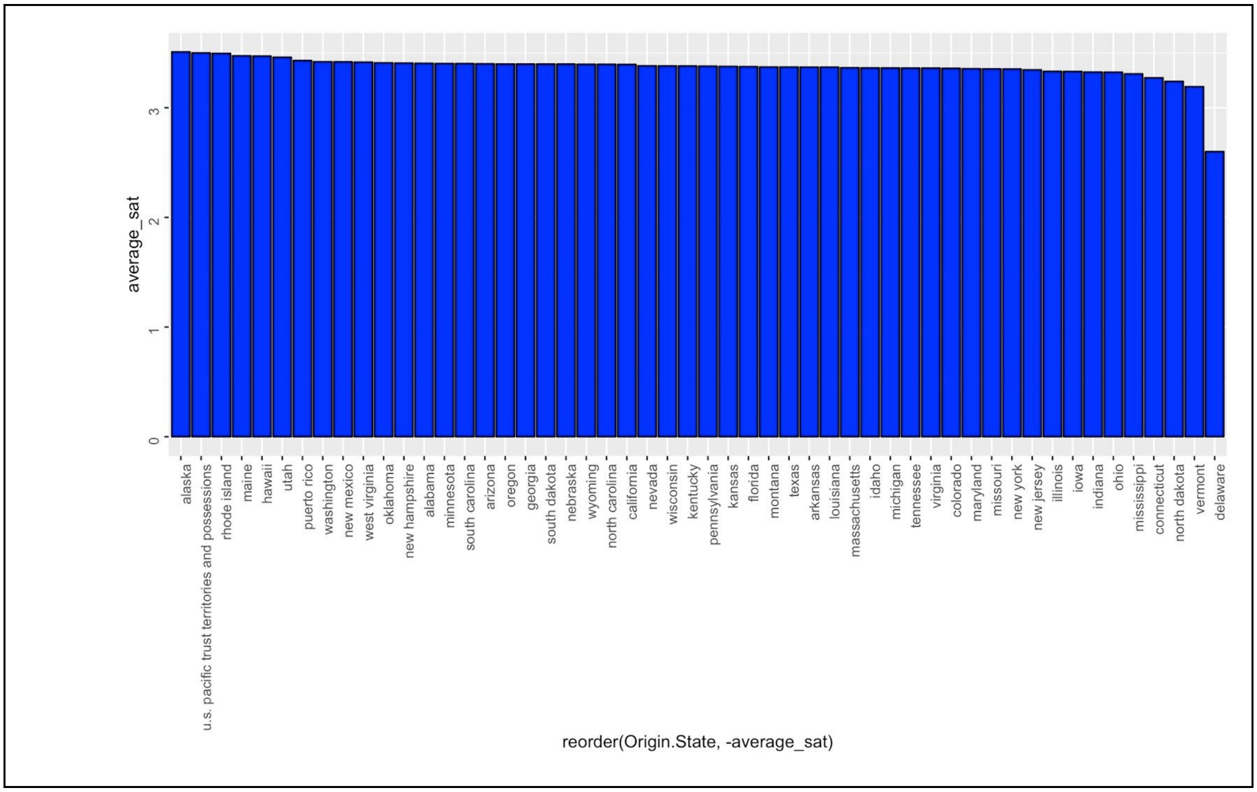
4. Geographical analysis:

The team also created color coded maps based on the average satisfaction level of the Origin State and the Destination State.
When then created a color coded map based on the origin States:

```
oStateMap <- ggplot(map_origin, aes(map_id
=map_origin$Origin.State))
oStateMap <- oStateMap+geom_map(map = us,
aes(fill=map_origin$average_sat))
oStateMap <- oStateMap+expand_limits(x = us$long, y = us$lat)
oStateMap <- oStateMap+coord_map() +ggtitle ("Map of the Origine
State")
oStateMap
```

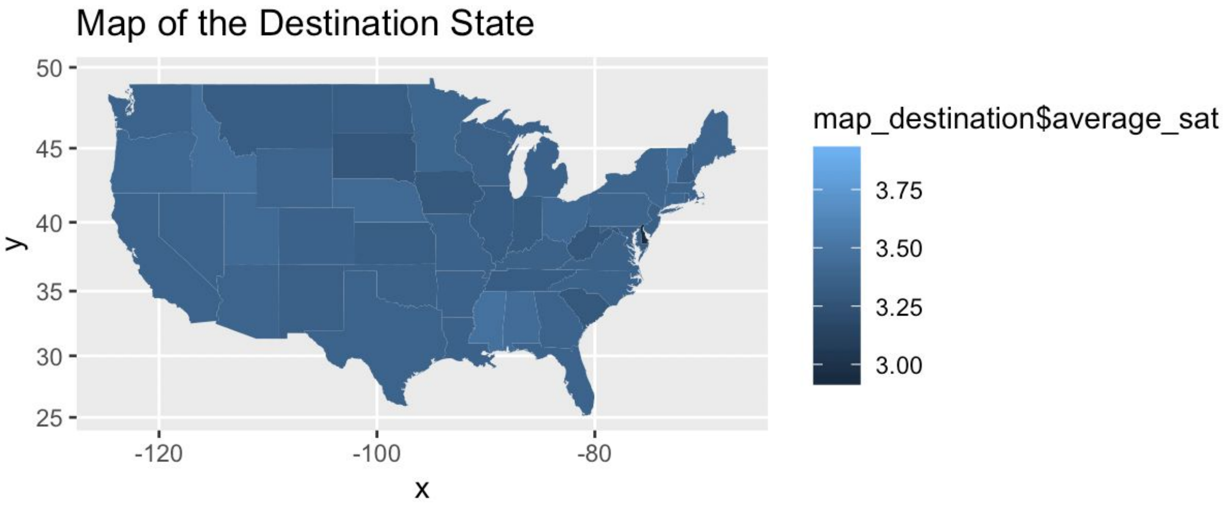


The geographic map did not clearly show the average satisfaction level by state, so we decided to create a bar chart to get a better understanding of the overall satisfaction level among different origin states:

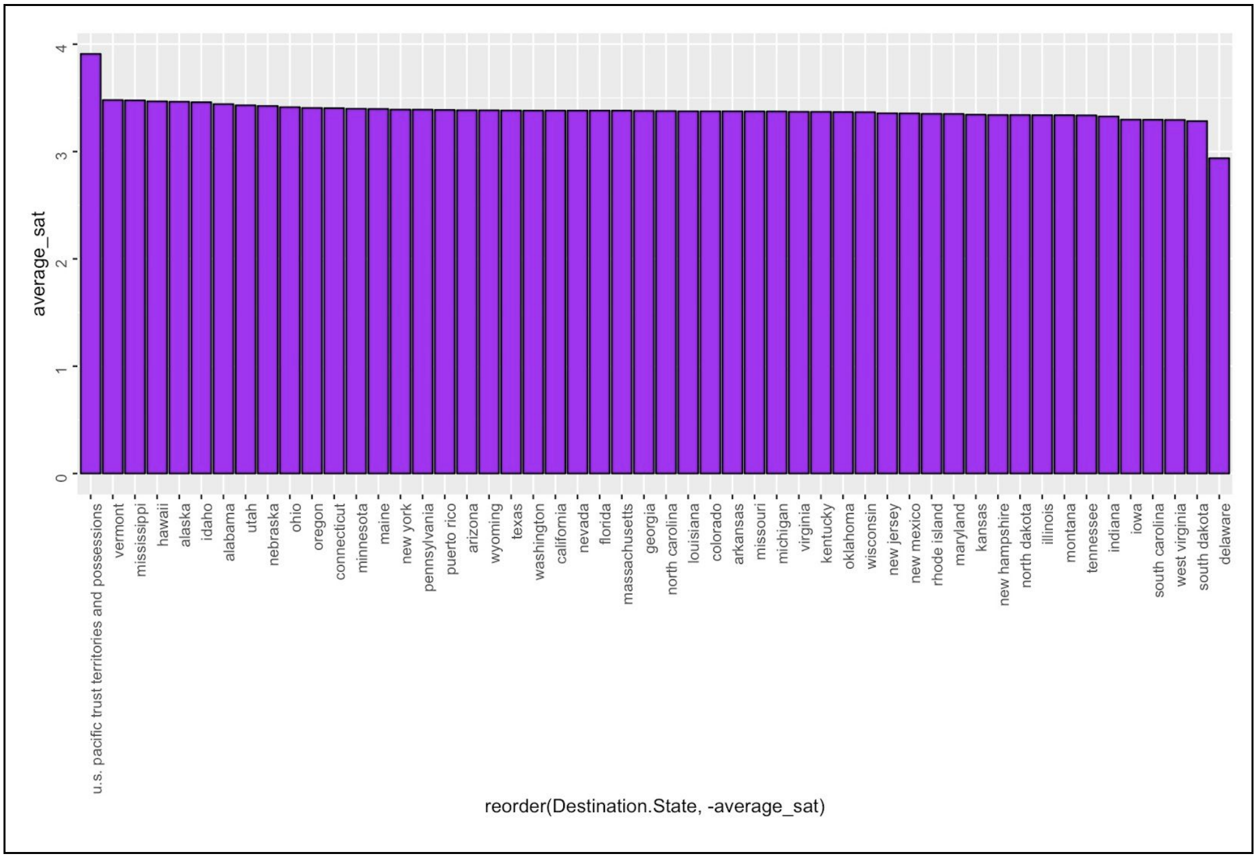


The bar chart show that the satisfaction level among most origin states is similar, ranging from 3.2 to 3.7. However, Delaware's satisfaction is only about 2.5, the lowest State in the U.S.
We then created a color coded map based on the destination state:

```
oStateMap <- ggplot(map_destination, aes(map_id
=map_destination$Destination.State))
oStateMap <- oStateMap+geom_map(map = us,
aes(fill=map_destination$average_sat))
oStateMap <- oStateMap+expand_limits(x = us$long, y = us$lat)
oStateMap <- oStateMap+coord_map() +ggtitle ("Map of the Origine
State")
oStateMap
```



The geographic map did not clearly show the average satisfaction level by state, so we created another bar chart to get a better view of the overall satisfaction level among different destination states:



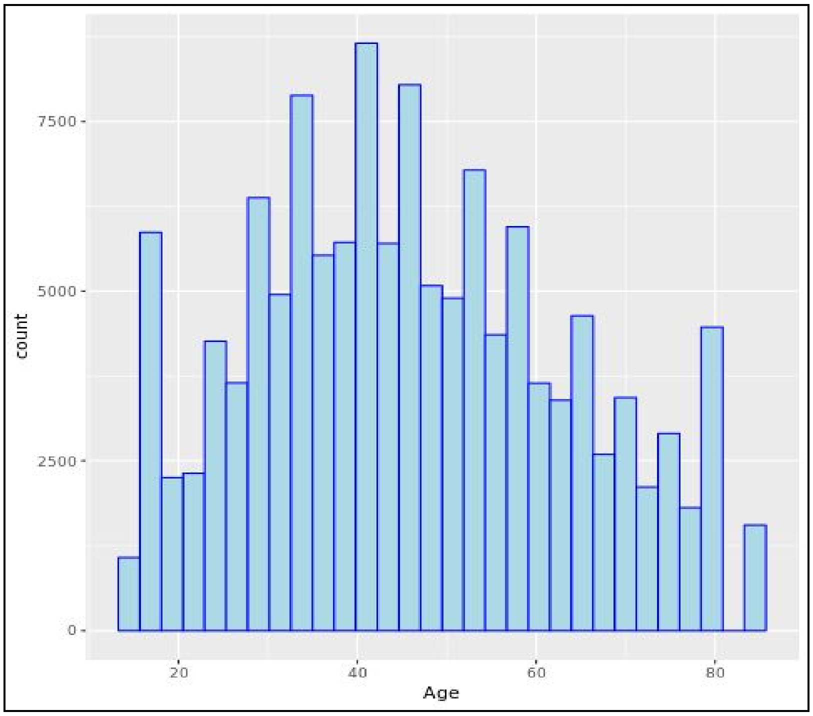
The bar chart shows that U.S pacific trust territories and possessions along with Delaware rank the highest and the lowest respectively. The rest of the states share similar satisfaction levels, ranging from 3.7 to 4. However, Delaware only averages about 2.5, the lowest State in the U.S.

We found that Delaware ranks the lowest as both the destination state and origin state, and that its overall satisfaction level is significantly lower than the rest of the States. Measures should be taken to boost the overall satisfaction level there.

5. Passenger Demographics:

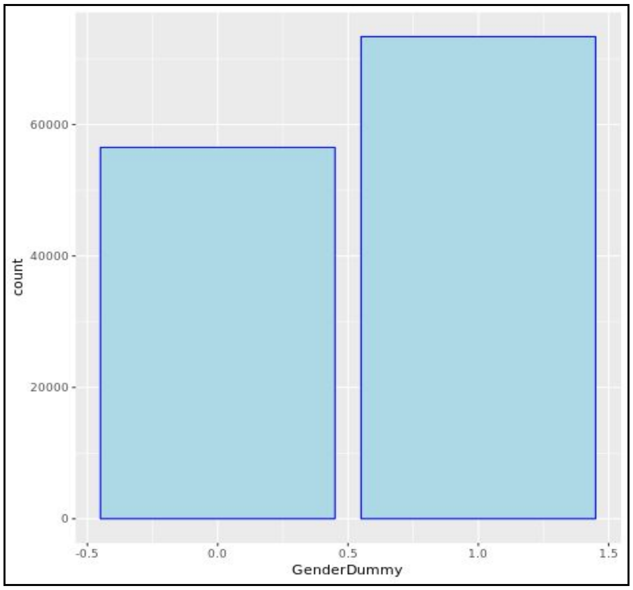
We analysed the variables that define the passenger demographics i.e. Age, Gender, No. of Flights, Shopping Amount at Airport, Price Sensitivity, Eating and Drinking at Airport.

First we created a histogram of the Age attribute:



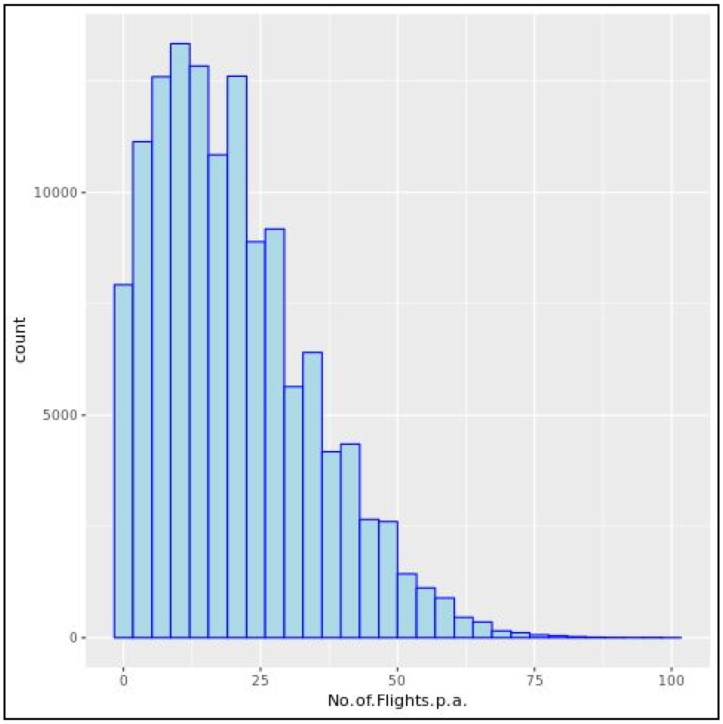
The Histogram shows that People in the age group of 35 to 55 travel more frequently, peaking at 40-45. This helps a great deal in knowing which age group to focus on in terms of primary demographics.

Next we created a barplot of the Gender attribute. We created a dummy variable for Male and Female using as.numeric, where Female=1, Male=0



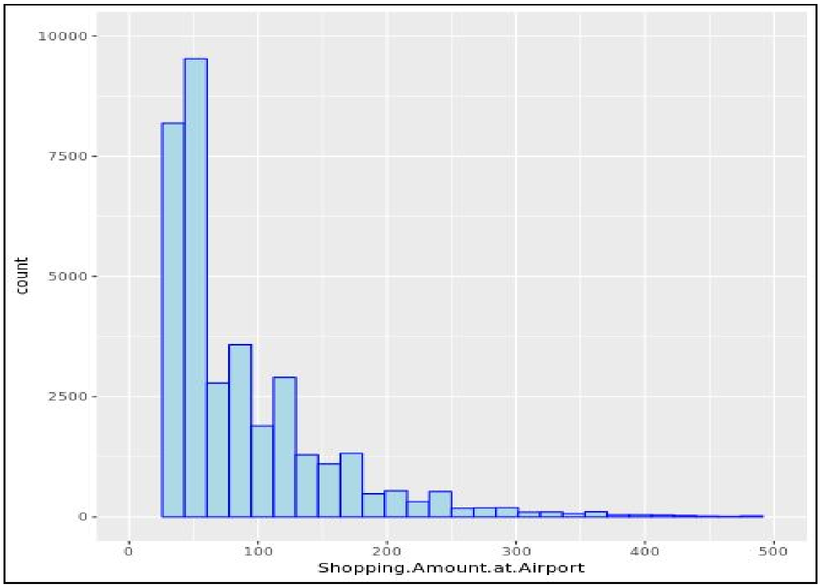
The barplot clearly shows that the number of female passengers is higher than males.

Next we created a plot to analyze No of Flights taken by a person on average.



On average, customers appear to have not taken very many flights. The graph is a left skewed graph which indicates mainly less than 30 flights.

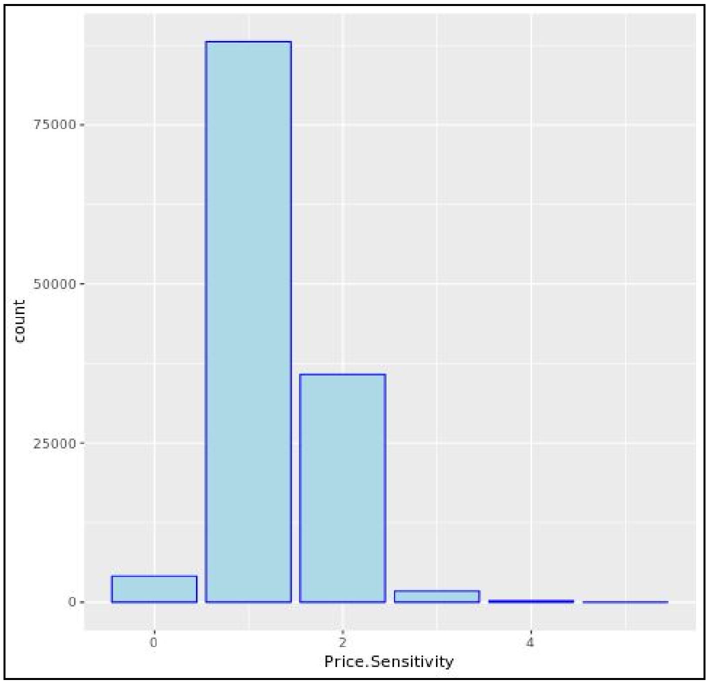
We also utilized Shopping Amount at Airport to examine the amount spent on average by customers at the airports:



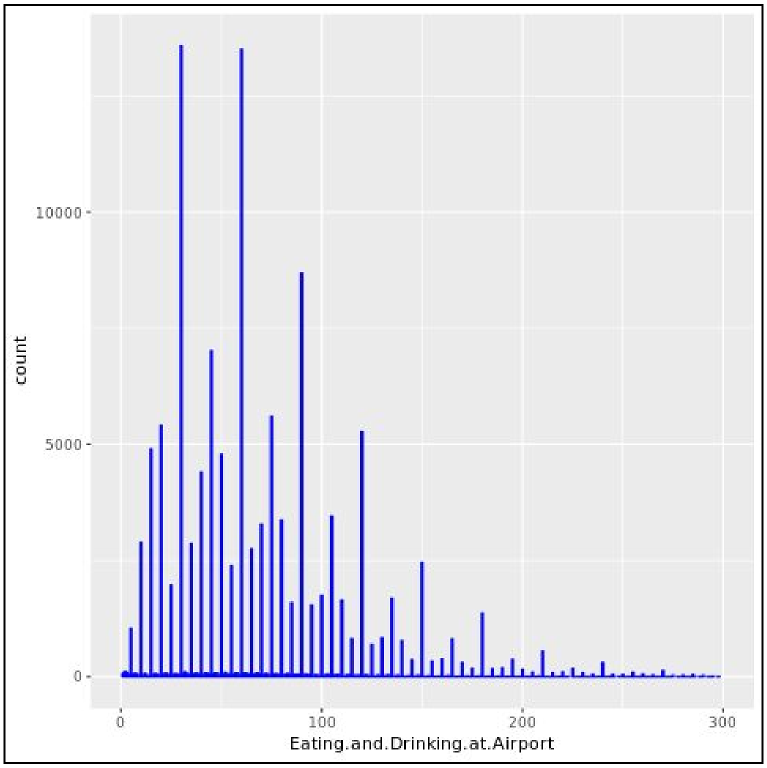
This data is right skewed and shows that on an average, people spent between 0 to 100 dollars at the airport. The summary shows that more than 50% of passengers do not spend any money shopping at the airport. Very few passengers spend upwards of $350, indicating that these cases are generally outliers.

Price Sensitivity has a range from 0 to 5 which is the amount which the ticket price affects whether a customer follows through with their transaction.

Under the assumption that a value greater than 2 indicates that passengers find the products at the airport expensive:



The maximum grading is one, which means the passengers do not find airports too expensive. Finally, we looked at the Eating and Drinking attribute:



People have eaten at the airport quite a few times(30-90 times) but this plot or the data does not give a precise data. Each of these attributes under passenger demographics are quite important to estimate the customer satisfaction since they give in depth details of our customers base.

Data Preparation

Data Acquisition:

In order to determine the variables that have the greatest effect on overall customer satisfaction, we were given three months of survey data that included each customer's self-reported overall Satisfaction rate for their flight experience as well as 27 other data elements including demographic data (Age, Gender), flight information (Location, Delays), flight history (Past Flight Quantities, Airline Status), and behavior (Price Sensitivity, Airport Expenditure) among other things.

We received this data in a .csv format within the MIDST environment and read it into our R modules by simply using the flow feature to identify the .csv as an Input Port with the variable satisfactionSurvey. From there we were able to use R code to do the exploratory analysis that allowed us to determine what steps we would take to clean and transform the data.

Data Transformation:

Since the Satisfaction variable was the dependent variable that we were interested in measuring, it was important that it was formatted correctly and that its values were measurable. First, we explored the unique entries in the Satisfaction column to discover any anomalies or unexpected values that were not equal to a number between 0 and 5.

```
unique(satisfactionSurvey$Satisfaction)
➞
"4.5"      "4"        "2.5"      "5"       "3.5"       "2"
"3"        "1"        "4.00.5"   "4.00.2.00"
```

Our exploration showed that there were some Satisfaction ratings of 4.00.5 and 4.00.2.00 in the data which fell outside of our expected format and are non-numeric (unable to be used in numerical analysis). Assuming that these were most likely errors, we decided to determine what percentage of the data was affected to help us decide the best way to address them.

The Satisfaction numbers that appear to be errors only make up 0.002% of the data available which is statistically insignificant. As they are unable to be analyzed as entered, because we cannot be sure of what numbers they were intended to be, and because they won't significantly influence overall trends, our group opted to remove them completely from the data set and create a new variable before moving forward.

Next, we examined the NULL values in Departure Delays, Arrival Delays and Flight Times that were found during the preliminary analysis of these variables. Because that analysis found that these instances made up approximately 1.5% of the data, we opted not to remove them entirely but instead to assign reasonably estimated values where the data was missing. This would allow us to estimate our models for these variables and keep the data from the other variables intact for these (approximately) 2400 surveys.

First, we created a dataframe that explored the relationship between the Flight Cancelled variable and missing Delay and Flight Time data to determine if the missing fields were legitimate records (as a cancelled flight most likely would not have a delay time or a flight time).

```
                               measures blankCounts
1                     Cancelled flights        2401
2                      Departure Delays        2345
3 Departure Delays for Cancelled flights        2345
4                        Arrival Delays        2738
5   Arrival Delays for Cancelled Flights        2401
6                           Flight Time        2738
7      Flight Time for Cancelled Flights        2401
```

From our analysis, we found that all of the blanks for Departure Delays are for flights that were cancelled along with 88% (2401 out of 2738) of the blanks for Arrival Delays and Flight Time. Since a cancelled flight has no recorded time in the air and cannot land late by any amount of time, we can reasonably set these instances to 0 for Arrival Delay and Flight Time. There were less blank Departure Delays than total flights cancelled, presumably because a flight can be delayed initially and also ultimately cancelled (after sitting on the tarmac). Still, it is reasonable to set the remaining blank Departure Delays to 0 where the flights were cancelled as an unreported delay on a flight that didn't fly most likely did not exist in the first place.

To clean the data in accordance with the above assumptions, we created new columns that were based upon the data in Departure Delays, Arrival Delays and Flight Time but that set these values to zero where they were blank and where Flight Cancelled was "Yes.”

Next, we needed to decide what to do with the remaining 338 instances where flights were not cancelled but Arrival Delay and Flight Time were still missing. Since we knew that these flights took place and that this is most likely data that was actually missing, so we decided to attempt to estimate the missing values using other known variables. To see if our other data could be used to accurately estimate the missing values, we used linear models to determine the correlation between the missing variables and those related to them.

For example, Flight Time should be related to Flight Distance with some slight variances due to things like head and tail winds, flight direction, etc. We used a linear model to see if these variables made a difference or if Flight Distance remained a reasonable predictor of Flight Time. At 0.9542, the R-Squared in our linear model was extremely close to 1 and at 2.2e-16, the p-value was very far below 0.05 which together showed that Flight Distance is reasonably good at predicting Flight Time and that the results are statistically significant.

Similarly, Departure Delay should be reasonably related to Arrival Delay with the only discrepancies between the two being attributed to mid-flight gains or losses in travel time. To test how these mid-flight changes affected the correlation of these two variables, we used another linear model. In this linear model, an R-Squared of 0.9316 and a p-value of 2.2e-16 together showed a strong relationship between Departure Delay and Arrival Delay and determined that we could reasonably estimate Arrival Delay data from Departure Delay data.

Using these linear models, we replaced the NULL values for Flight Time and Arrival Delay with estimated values based on Flight Distance and Departure Delay, respectively.

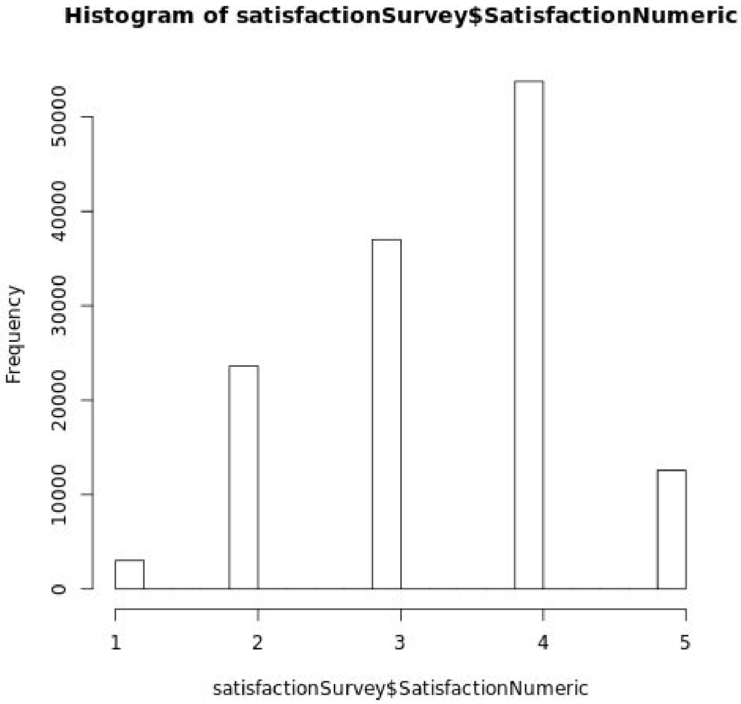
After cleaning our data for errors and NULL values, we began to prepare it for future analysis. First, we coerced the Satisfaction ratings into a new (numeric) column so that it could be better used in future models and analysis.

```
satisfactionSurvey$SatisfactionNumeric <-
as.numeric(as.character(satisfactionSurvey$Satisfaction))
```

Our final data munging step was to create binary Satisfaction categories of “High” and “Low” to aid in distinguishing between these two groups in our future analysis. First, we determined the threshold that made the most sense within the distribution of Satisfaction numbers in our data by examining their central tendency and dispersion.

```
summary(satisfactionSurvey$SatisfactionNumeric)
➞
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   3.000   4.000   3.379   4.000   5.000

hist(satisfactionSurvey$SatisfactionNumeric)
➞
```



It seemed that the Satisfaction numbers fell into two groups distributed around 3.5, so we made that our threshold for defining "High" and "Low" Satisfaction. We know from previous analysis that there is an insignificant quantity of 3.5 ratings so we grouped them in the high group for the sake of simplicity when creating our new Binary Satisfaction column.
In following the above steps of analysis and transformation, we were able to modify our dataframe (satisfactionSurvey) to be used in further analysis.

Modeling & Visualizations

In this section we discuss the models created by us with an aim to explain our dependent variable Satisfaction with other explanatory variables in our data. We started by creating a simple linear model using linear regression for modelling, followed by mining our data by applying association rule mining. We discuss our results from these approaches and try to validate the same using other modelling approaches. We also built a logistics regression model predicting the binary satisfaction (high/low) and a SVM and a Decision tree model to validate our results and draw insights from our models.

1. Linear Modelling

We began our modelling process by creating a linear model that took each of the independent variables provided in the dataset and used them to attempt to predict the outcome of the dependent variable, overall customer Satisfaction. By examining the relative strength of each correlation, we were able to begin to hone in on which variables had a greater impact on customer Satisfaction, confirming our preliminary analysis and moving us one step closer to identifying actionable insights.

First, we created our linear model formula including all of the dependent variables and the numberized version of our satisfaction variable. We stored the summary of this model in another variable for ease of analysis.

```
originalModel <- summary(lm(formula = SatisfactionNumeric ~.,
      data = satisfactionSurvey))
```

From this original model, we were able to determine that the following variables were statistically insignificant or at least not of primary importance: Origin.City, Origin.State, Destination.City, Destination.State, Flight.date, Airline.Name, X..of.Flight.with.other.Airlines, No..of.other.Loyalty.Cards, Eating.and.Drinking.at.Airport, Day.of.Month , Airline.Code, Departure.Delay.in.Minutes.0, Arrival.Delay.in.Minutes.0.Est, Flight.time.in.minutes.0.Est, and Flight.Distance. While we still wanted to validate these findings with other modelling techniques, it was encouraging that these results matched our preliminary analyses.

Next, we re-ran the linear model with the statistically significant variables and stored it in a new variable to aid in future analysis. Each of the independent variables in this model have a significance code of “ *** ” because their p-values are below 0.001. Thus, we were able to use linear modelling to narrow down our list of potentially correlated variables, verify our exploratory analysis, and prepare for further validation using other modelling techniques.

```
finalModel <- summary(lm(formula = SatisfactionNumeric ~ 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 + Flight.cancelled +
Arrival.Delay.greater.5.Mins, data = satisfactionSurvey))
Residuals:
➞
    Min      1Q  Median      3Q     Max
-3.1441 -0.4148  0.0786  0.4716  2.8461

Coefficients:
                                  Estimate Std. Error  t value Pr(>|t|)
(Intercept)                    -5.603e+00  1.348e+00   -4.156 3.24e-05
***
Airline.StatusGold              4.368e-01  7.404e-03   58.999  < 2e-16
***
Airline.StatusPlatinum          2.536e-01  1.151e-02   22.039  < 2e-16
***
Airline.StatusSilver            6.215e-01  5.150e-03  120.673  < 2e-16
***
Age                            -2.311e-03  1.257e-04  -18.376  < 2e-16
***
GenderMale                      1.291e-01  4.146e-03   31.138  < 2e-16
***
Price.Sensitivity              -3.912e-02  3.704e-03  -10.563  < 2e-16
***
Year.of.First.Flight            4.712e-03  6.719e-04    7.013 2.35e-12
***
No.of.Flights.p.a.             -3.195e-03  1.507e-04  -21.204  < 2e-16
***
Type.of.TravelMileage tickets  -1.414e-01  7.690e-03  -18.383  < 2e-16
***
Type.of.TravelPersonal Travel  -1.069e+00  4.914e-03 -217.510  < 2e-16
***
Shopping.Amount.at.Airport      1.656e-04  3.789e-05    4.372 1.23e-05
***
```

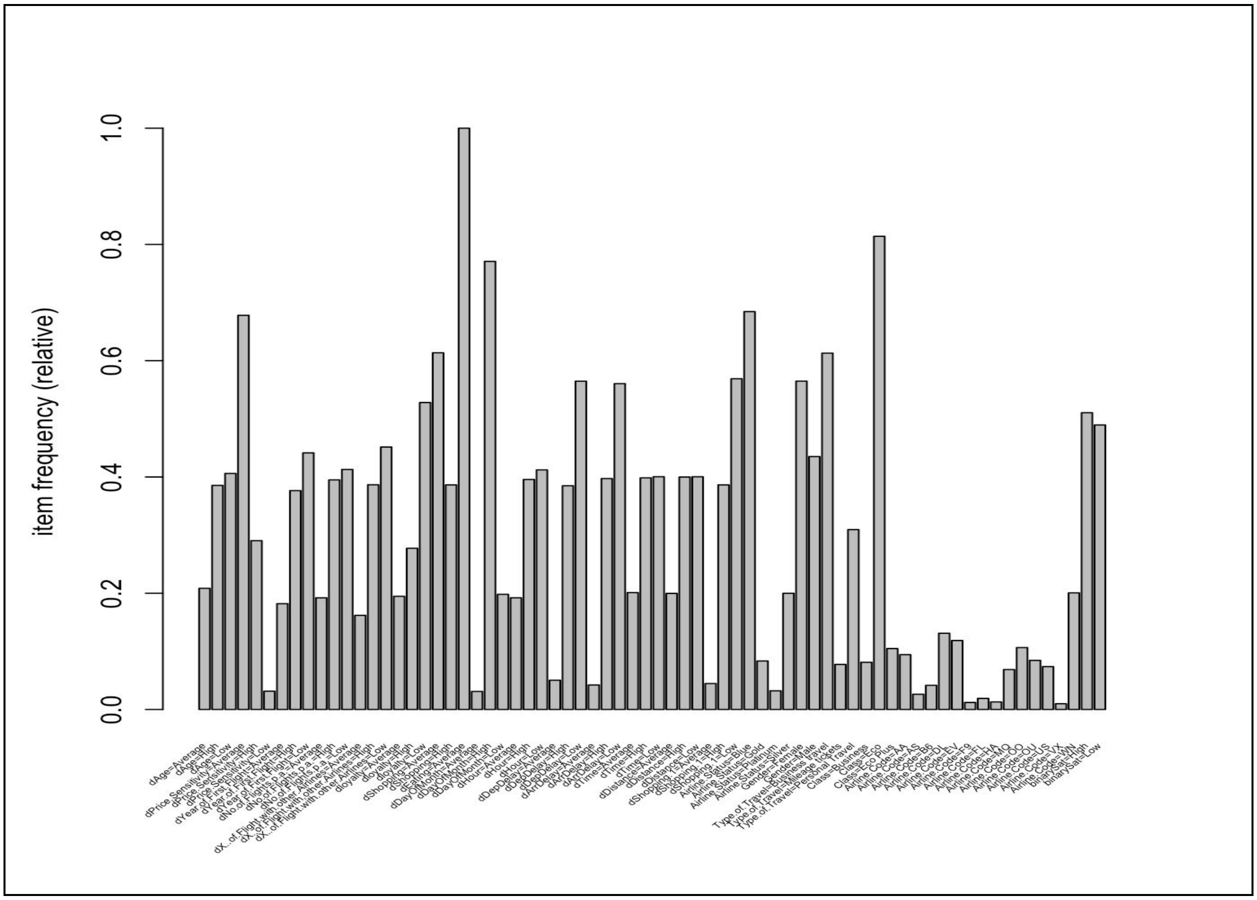
```
ClassEco                           -7.796e-02  7.350e-03  -10.607  < 2e-16
***
ClassEco Plus                      -7.069e-02  9.408e-03   -7.514 5.77e-14
***
Scheduled.Departure.Hour            3.754e-03  4.347e-04    8.636  < 2e-16
***
Flight.cancelledYes                -1.502e-01  1.491e-02  -10.073  < 2e-16
***
Arrival.Delay.greater.5.Minsyes -3.408e-01  4.255e-03  -80.095  < 2e-16
***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.719 on 129869 degrees of freedom
Multiple R-squared:  0.4445,    Adjusted R-squared:  0.4445
F-statistic:  6496 on 16 and 129869 DF,  p-value: < 2.2e-16
```

2. Association Rules Mining

Discretization of Data:
The team discretized 18 variables including Departure Delay in Minutes, Arrival Delay in Minutes, Eating and Drinking in the Airport, Satisfaction level and Number of Flights. The variables were categorized into “low”, “median” and “high” levels. We also factorized several variables and then combined them with the discretized columns producing a new csv file called discretized data (129886 obs. of 21 variables) for association rules mining. We created a function to discretize the assigned variables (using the cleaned and/or estimated ones where applicable) and stored them in new variables. We followed the same steps for other variables as well and created a new dataframe with this discretized data.

Next we coerced the discretized satisfaction data frame into a sparse transactions matrix.
We used the inspect( ), itemFrequency( ), and itemFrequencyPlot( ) commands to explore the contents of the new dataframe.

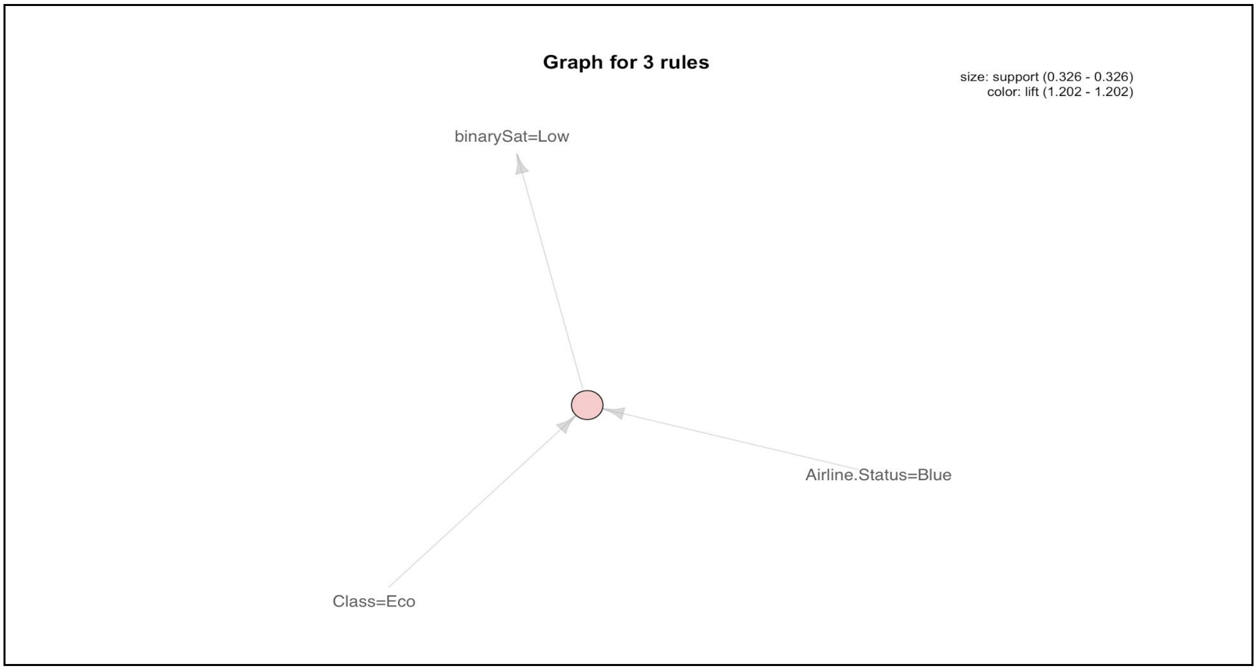


Then we used arules to discover patterns. Running the apriori command allowed us to try and predict unhappy customers.
The following code allowed us to rank the resulting rules:

```
rules.new <- rulesLow[order(-quality(rulesLow)$lift),]
inspect(head(rules.new,5))
```

```
  lhs
rhs              support   confidence lift     count
[1] {Airline.Status=Blue,Class=Eco}                          =>
{binarySat=Low} 0.3263631 0.5885456  1.202439 42390
[2] {dEating=Average,Airline.Status=Blue,Class=Eco}         =>
{binarySat=Low} 0.3263631 0.5885456  1.202439 42390
[3] {dDayOfMonth=High,Airline.Status=Blue}                   =>
{binarySat=Low} 0.3081086 0.5840570  1.193268 40019
[4] {dEating=Average,dDayOfMonth=High,Airline.Status=Blue} =>
{binarySat=Low} 0.3081086 0.5840570  1.193268 40019
[5] {Airline.Status=Blue}                                    =>
{binarySat=Low} 0.3991500 0.5831131  1.191340 51844
```

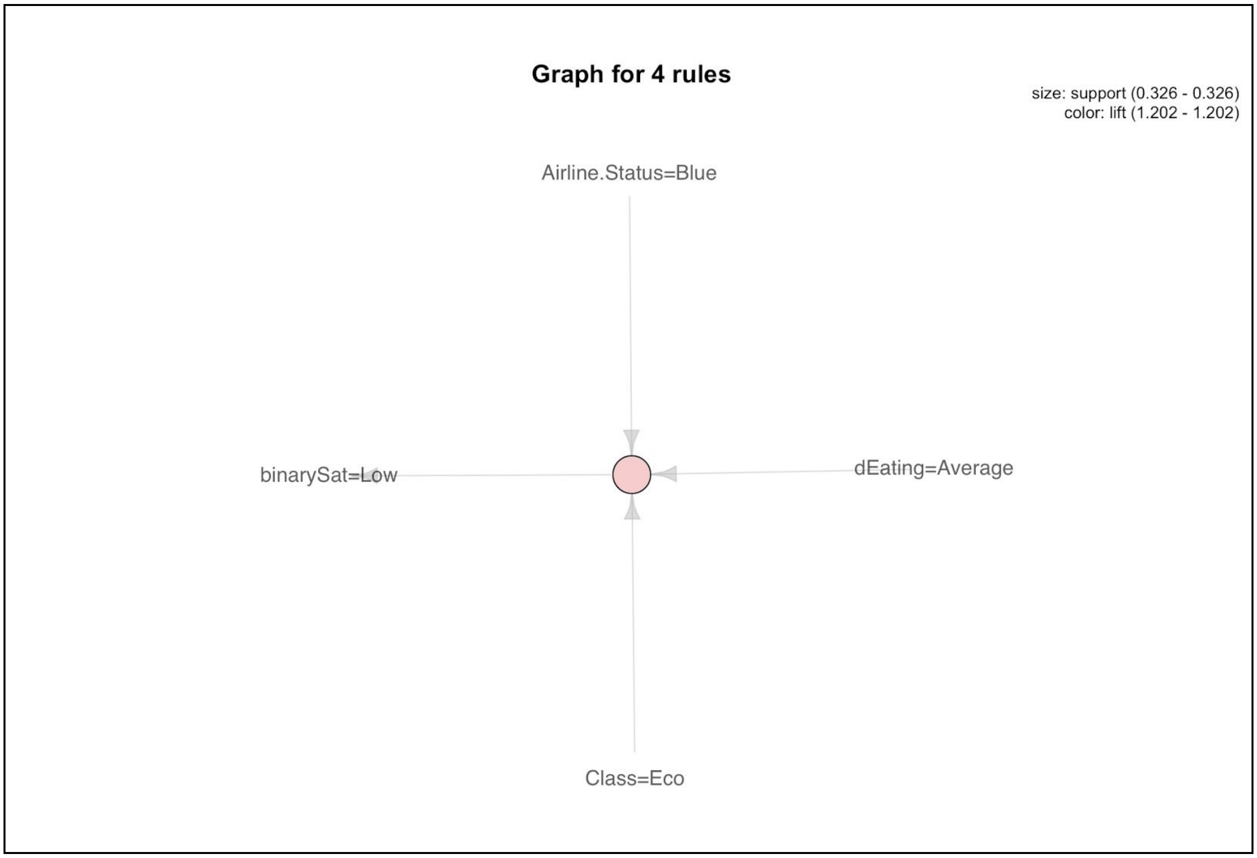
We found that the top five rules contained the attribute rules Airline Status=Blue, Class=Eco, Eating and Drinking =Average, Day Of Month=High. In other words, they are strong rules for low customer service satisfaction level.



The graph show that Class=Eco and Airline.Status=Blue lead to a low satisfaction level among customers.

The following code allowed us to visualize the second lowest satisfaction rule:

```
rules.new.2<-rules.new[2]
plot(rules.new.2,method="graph",main="Graph for 4 rules")
```

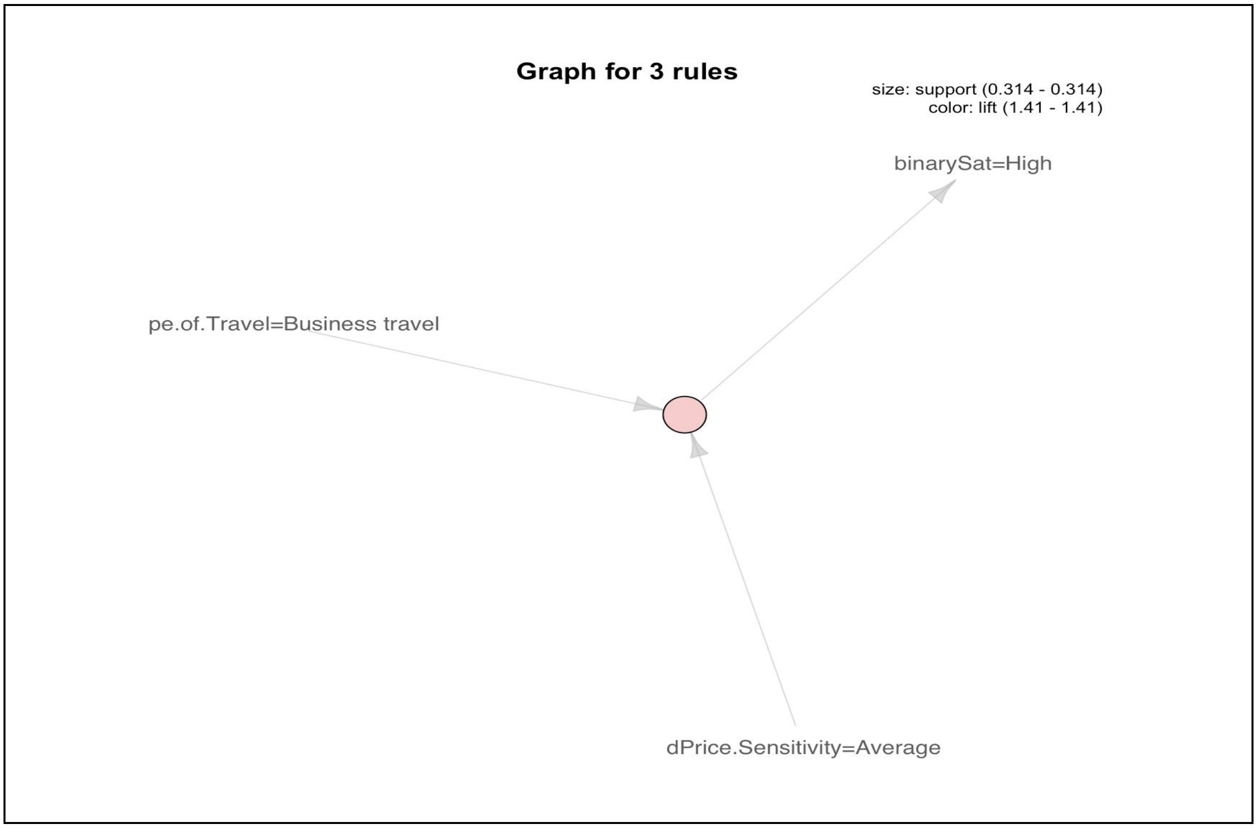


This graph showed that Class=Eco, Eating and Drinking =Average, and Airline Status=Blue lead to a low satisfaction level among customers.

The team also ran the apriori command allowed us to predict happy customers.

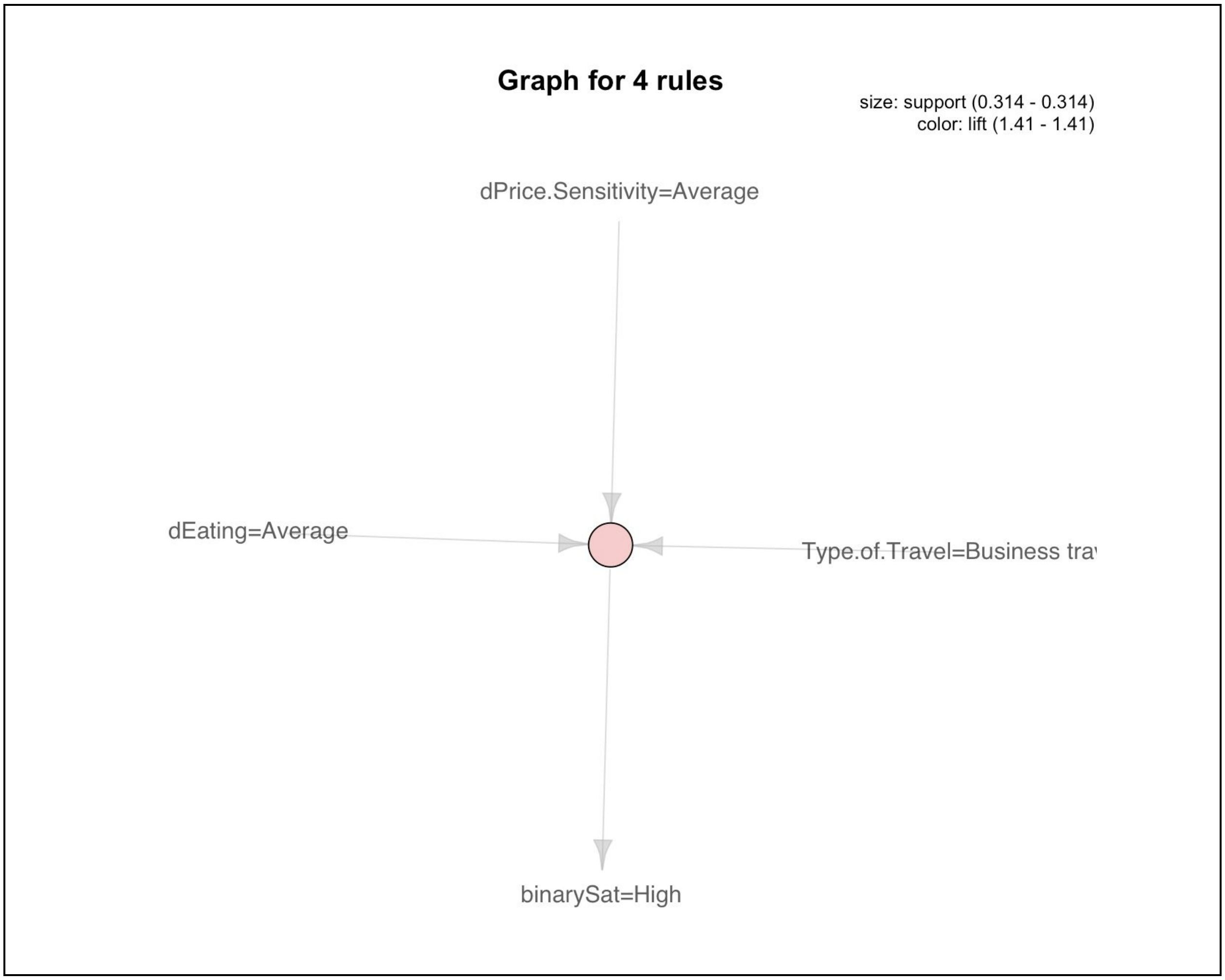
```
  lhs                                    rhs               support confidence
lift count
[1] {dPrice.Sensitivity=Average,
     Type.of.Travel=Business travel} => {binarySat=High} 0.3144142
0.7199929 1.410258 40838
[2] {dPrice.Sensitivity=Average,
     dEating=Average,
     Type.of.Travel=Business travel} => {binarySat=High} 0.3144142
0.7199929 1.410258 40838
[3] {Type.of.Travel=Business travel} => {binarySat=High} 0.4346427
0.7089806 1.388688 56454
[4] {dEating=Average,
     Type.of.Travel=Business travel} => {binarySat=High} 0.4346427
0.7089806 1.388688 56454
[5] {dDayOfMonth=High,
     Type.of.Travel=Business travel} => {binarySat=High} 0.3347243
0.7071568 1.385115 43476
```

We found that the top five rules contained the attribute rules Type of Travel = Business, Eating and Drinking =Average, Day of Month=High and Price Sensitivity=Average. In other words, they are strong rules for high customer service satisfaction level.
The highest satisfaction rule:


Graph for 3 rules
size: support (0.314 - 0.314)
color: lift (1.41 - 1.41)
binarySat=High
pe.of.Travel=Business travel
dPrice.Sensitivity=Average

This graph showed that Type of Travel = Business and Price Sensitivity=Average lead to a high satisfaction level among customers.

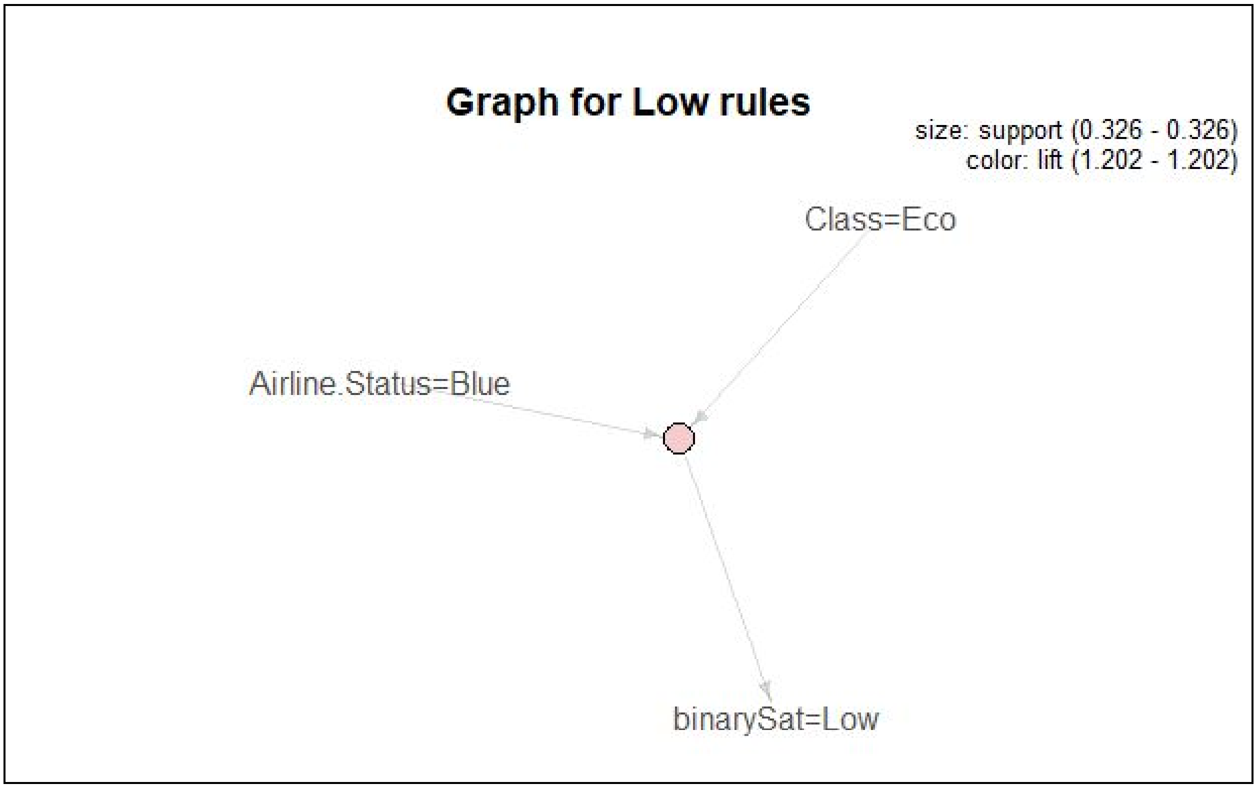
The second lowest satisfaction rule:


Graph for 4 rules
size: support (0.314 - 0.314)
color: lift (1.41 - 1.41)
dPrice.Sensitivity=Average
dEating=Average
Type.of.Travel=Business tra
binarySat=High

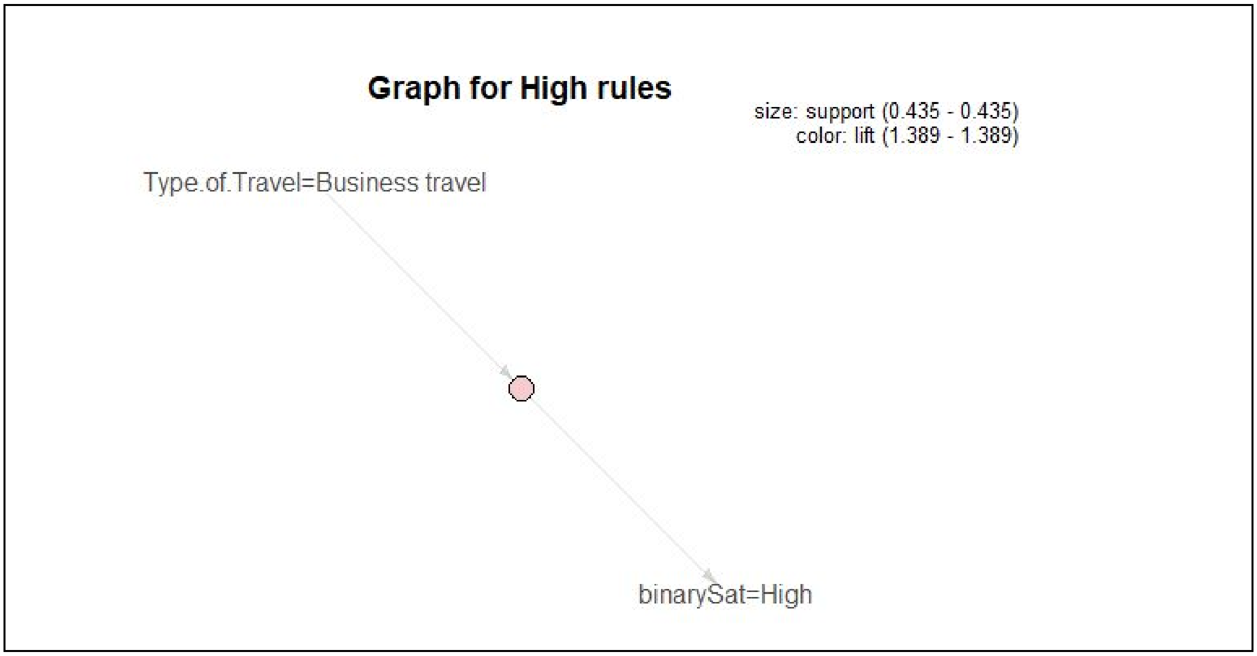
This graph showed that Type of Travel = Business, Eating and Drinking =Average and Price Sensitivity=Average lead to a high satisfaction level among customers.

Further analysis with the significant variables obtained from all our models on an average lead to the following two major rules for High and Low satisfaction:

```
lhs                                 rhs             support   confidence lift
{Airline.Status=Blue,Class=Eco} => {binarySat=Low} 0.3263631 0.588545
1.202439
 count
 42390
```



```
lhs                                   rhs              support   confidence
{Type.of.Travel=Business travel} => {binarySat=High} 0.4346427 0.7089806
Lift      count
1.388688  56454
```


Graph for High rules
size: support (0.435 - 0.435)
color: lift (1.389 - 1.389)
Type.of.Travel=Business travel
binarySat=High

Validation:

In this section we create different models to validate the results of the linear model and association rules.

1. SVM

The importance of the previously determined significant variables by the linear model were further confirmed through the use of an SVM model. Due to the sheer size of the data and the number of attributes involved as well as the cost of compute power, we elected to sample the data and run our SVM validation on a collection of 10,000 of the 129,886 instances available. The seed is set at 123 to provide consistency in replicating these results.

```
set.seed(123) #set seed to ensure consistent results
sampleData <- cleanData[sample(nrow(cleanData),10000,
replace=FALSE),]
```

Of this sample data, ⅔ was selected for training purposes while the file ⅓ was utilized for testing. This training data was then utilized to create an SVM classifier. After iterations of testing, a cost of 500 appeared to create a worthwhile distinction between the satisfied and unsatisfied classes, the seed is set at 123 to provide consistency in replicating this model. Again, variables were chosen based on the variables deemed significant by the linear and association rules models. The code for the model is displayed below:

```
svmOutput <- ksvm(binSat ~.,data=sampleData, kernel="rbfdot",
C=250, cross=10, set.seed=123, prob.model=TRUE)
```

Which runs with the following results:

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

SV type: C-svc  (classification)
parameter : cost C = 250

Gaussian Radial Basis kernel function.
Hyperparameter : sigma =  0.115144917738097

Number of Support Vectors : 4733

Objective Function Value : -794617.7
Training error : 0.1356
```

```
Cross validation error : 0.2417
Probability model included.
```

With a Cross Validation error of ~24%, the model is able to correctly predict satisfied customers about 75% of the time. The fact that this is around double the value of our Training Error (13.56%) makes sense as the parameters generally do not perform as well on other data as the original training data.

With a cost of 250, the plurality of instances were difficult to classify and therefore had a significant impact on the structure of the model. Only around 600 instances were easily classified. The support vector voting ease is depicted in the form of a histogram below:

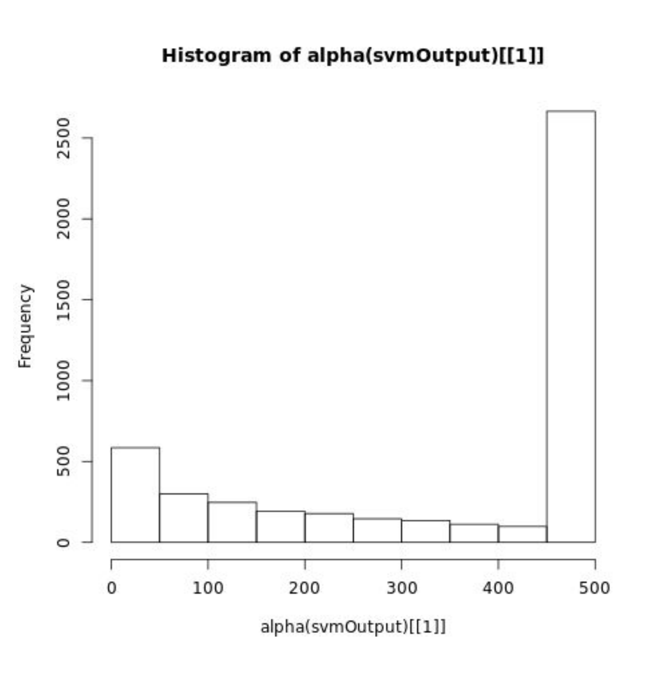


Figure - Frequency vs Ease of Vote (SVM)

Once our SVM model was generated, we were able to further validate our results by running our model on the test data and comparing the votes from the model to the actual satisfaction attributes of the test data. This information was then outputted in the form of a confusion matrix in order to easily identify errors that occured.

The confusion matrix that results from this test data compared to the prediction data is as follows:

```
              svmPred.2...
testData.binSat    0    1
             0  1246  376
```

```
            1    67 1645
```

To calculate the accuracy of our predictions for the testing data, we can simply add up the correctly classified instances (1246 + 1645 = 2891) and the incorrectly classified ones (376 + 67 = 443) and then divide the incorrectly classified total by the correctly classified total (443 / 2891) to receive an overall accuracy of 15.32%.

As a result of this information, we can be relatively confident in the relevance of our significant variables as well as the accuracy of our model. In this way, this model could be utilized in the future to predict customer satisfaction with a reasonable degree of accuracy.

2. Logistic Regression

Logistic regression is another modelling technique we used for our analysis. Here, we converted the Satisfaction column into binary values i.e. HIGH=1, LOW=0. The dependent variable(Y axis) i.e the Satisfaction column, is compared with other independent variables(X axis) through logistic regression to predict variables that affect the satisfaction rate the most. We used the Amelia and caTools library for our R code and then split the data into training data and test data.
The glm() function was used to run the logistic regression on the training data in R:

```
logReg <- glm(binarySat ~. , data = train, family = binomial)
summary(logReg)

→
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4953  -0.5768   0.3770   0.8494   3.1845

Coefficients:
                                  Estimate Std. Error  z value Pr(>|z|)
(Intercept)                      2.690e+00  7.141e-02   37.676  < 2e-16
***
Airline.Status                   5.206e-01  7.982e-03   65.228  < 2e-16
***
Age                             -9.616e-03  6.075e-04  -15.827  < 2e-16
***
Gender                          -3.388e-01  1.760e-02  -19.253  < 2e-16
***
Price.Sensitivity               -1.526e-01  1.588e-02   -9.611  < 2e-16
***
No.of.Flights.p.a.              -1.128e-02  6.580e-04  -17.135  < 2e-16
```

```
***
X..of.Flight.with.other.Airlines  3.408e-03  1.107e-03    3.077  0.00209
**
Type.of.Travel                   -1.379e+00  1.137e-02 -121.276  < 2e-16
***
No..of.other.Loyalty.Cards       -6.163e-02  8.888e-03   -6.934 4.08e-12
***
Class                            -1.264e-01  1.987e-02   -6.358 2.05e-10
***
Day.of.Month                      2.231e-03  9.958e-04    2.240  0.02507
*
Airline.Code                     -4.217e-03  1.936e-03   -2.178  0.02938
*
Scheduled.Departure.Hour          1.073e-02  1.872e-03    5.733 9.88e-09
***
Departure.Delay.in.Minutes        4.419e-03  9.069e-04    4.873 1.10e-06
***
Flight.Distance                   3.757e-04  7.149e-05    5.255 1.48e-07
***
Arrival.Delay.in.Minutes         -1.024e-02  8.979e-04  -11.406  < 2e-16
***
Flight.time.in.minutes           -2.935e-03  5.899e-04   -4.976 6.49e-07
***
Expenditure.at.Airport            1.935e-04  1.141e-04    1.696  0.08981
.
```

Through the logistic regression model, we found that the variables Airline Status, Age, Gender, Price Sensitivity, No. of Flights p.a., Type of Travel and Arrival Delay in minutes are statistically significant variables since they have the lowest p-value i.e. 2e-16

We used our model to predict the satisfaction from the test data and obtained the probability of Satisfaction being HIGH or LOW by setting type=”response”. We used a threshold of 0.5 to convert these probabilities into HIGH and LOW labels.

Next, we used these predictions to construct a confusion matrix:

```
logProbs = predict(logReg, newdata = test, type = "response")
logPred = ifelse(logProbs > 0.5, "High", "Low")
test.binarySat  High   Low
             Low  7078 13583
             High 18811  3020
```

Further we calculated the classification error to be 0.2376

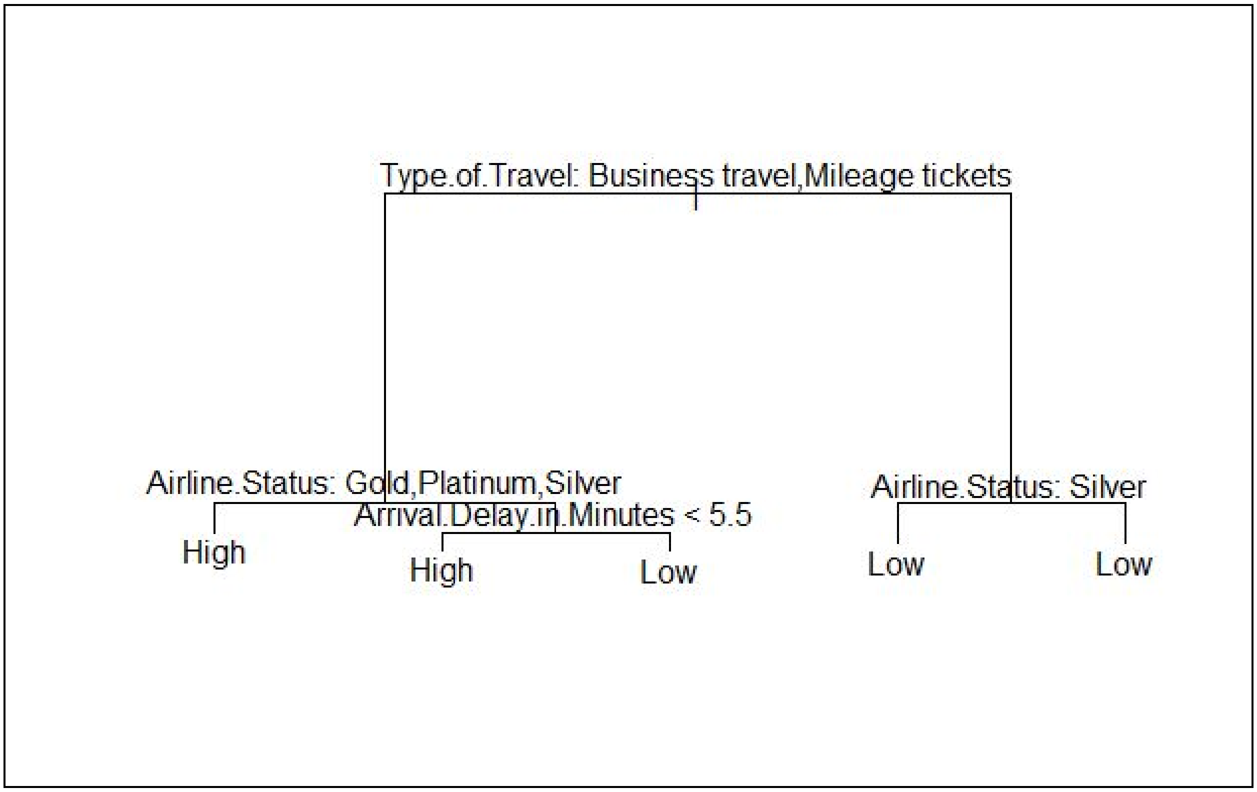
3. Decision Tree

We trained a decision tree model on our data and further analyzed its results. A tree model allowed us to use all the categorical variables in our dataset, since a tree can learn and make decisions using these as well, unlike a linear regression which requires numerical variables only.

We used the tree library to create our model. Our target variable here was binary satisfaction created earlier and we aimed to predict factors yielding higher satisfaction for customers:
The model and the significant variables as per our model are shown below:

```
summary(tree.model)
➞
Classification tree:
tree(formula = binarySat ~ ., data = train)
Variables actually used in tree construction:
[1] "Type.of.Travel"           "Airline.Status"
"Arrival.Delay.in.Minutes"
Number of terminal nodes:  5
Residual mean deviance:  0.9536 = 81040 / 84980
Misclassification error rate: 0.2363 = 20079 / 84986
```

Notable from this output were the variables actually used in tree construction: Type.of.Travel, Airline.Status and Arrival.Delay.in.Minutes. So according to our model these variables were significant in helping us predict the Satisfaction. Also the misclassification error rate for our model is 0.2363 (20079 / 84986) which may be high due to overfitting.
Next we plotted our Decision Tree:

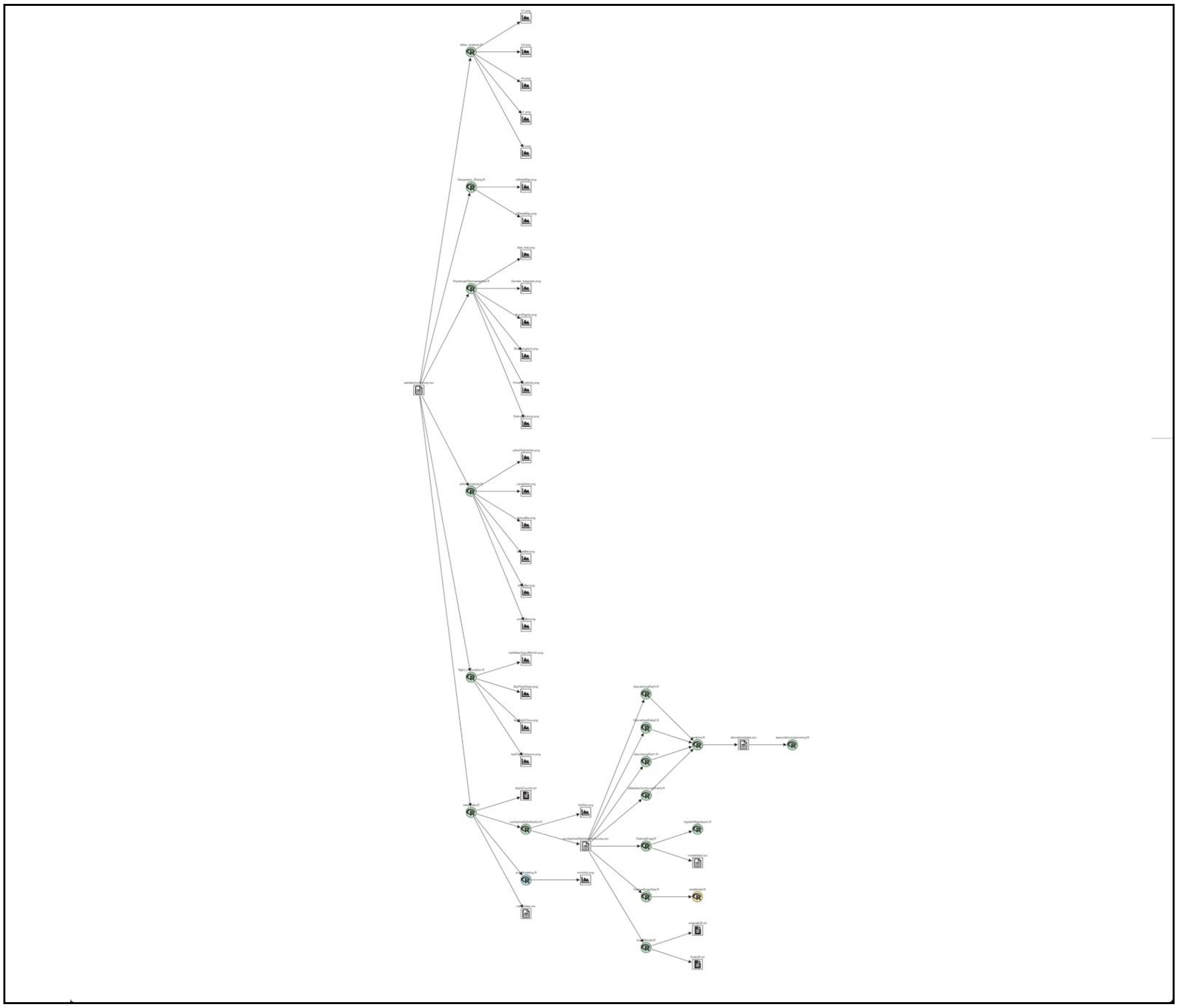

Type.of.Travel: Business travel,Mileage tickets
Airline.Status: Gold,Platinum,Silver
Arrival.Delay.in.Minutes < 5.5
Airline.Status: Silver
High
High
Low
Low
Low

Insights that can be drawn from the tree are that Type of Travel, Airline Status, and Arrival Delay help us determine satisfaction. Business and mileage travelers had higher satisfaction than personal travel. Gold and Platinum airline status were also indications for higher satisfaction and with different results for different levels of delay, more satisfied customers experiencing delays of less than 5 mins.

We made predictions on the test data using our model and constructed the confusion matrix using the target variable and predictions generated by the model as follows:

```
tree.pred = predict(tree.model,test,type="class")
#Confusion matrix:
      prediction
target  High   Low
  High 17265  4619
  Low   5449 15160
```

MIDST:



Utilizing the MIDST platform, we conducted our work as a team while assigning specific tasks to individuals. Our web of MIDST nodes is pictured above.

For the initial exploratory data analysis, we divided the work by splitting the variables into five logical groups and each group was assigned to a team member for this phase. Using the results from this phase we discussed as a group about how we will be cleaning and transforming our data and the same was done in code by a single team member. Further logically discretizing the data for data mining was again split amongst the team with each member discretizing a group of variables, after which a member compiled this work together. Finally, the modeling phase was split amongst members, where each member implemented a different model for the data. To review our work, after each phase the work done by a team member was validated by a different team member, chosen on rotation basis.

Actionable Insights:

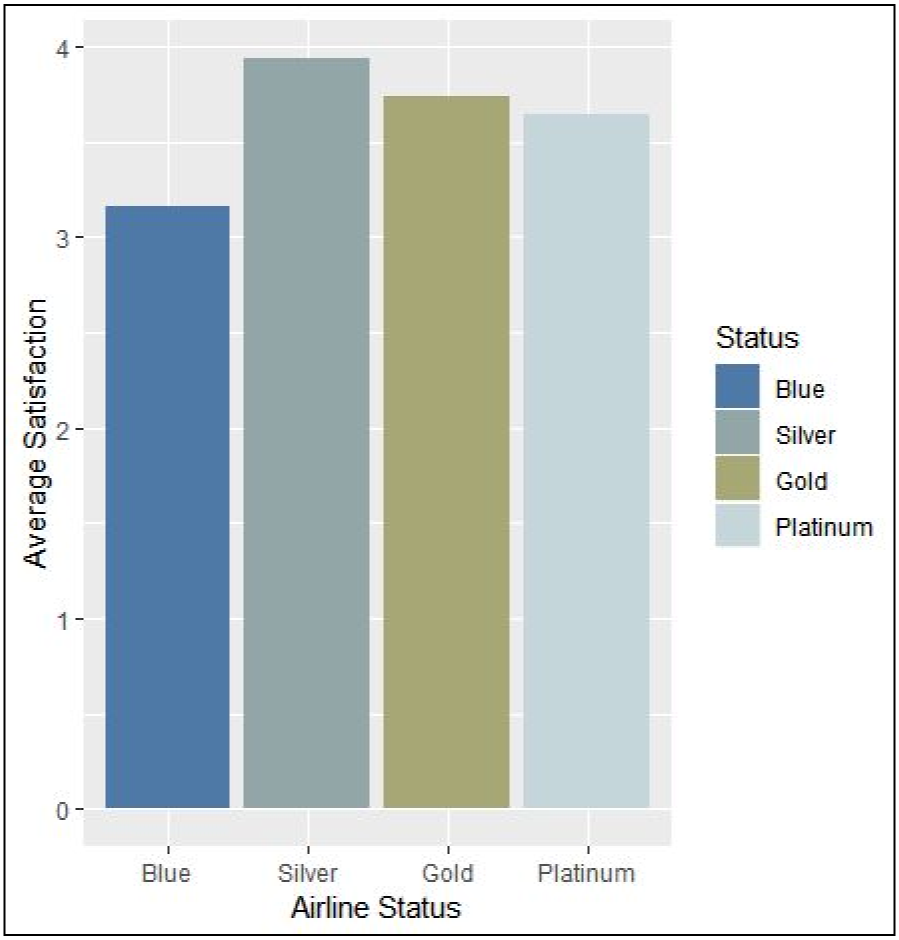
Summary:

Model	**Significant Variables**
Linear	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, Flight.cancelled,Arrival.Delay.greater.5.Mins
Logistic	Airline.Status, Age, Gender, Price.Sensitivity,No.of.Flights.p.a.,Type.of.Travel, Arrival Delay in minutes
Association Rules	
High	Type.of.Travel=Business travel, dPrice.Sensitivity=Average, dEating=Average, dDayOfMonth=High
Low	Airline.Status=Blue,Class=Eco, dEating=Average,dDayOfMonth=High
Decision Tree	Type.of.Travel, Airline.Status and Arrival.Delay.in.Minutes.

Recommendations:

1. Status Upgrade:

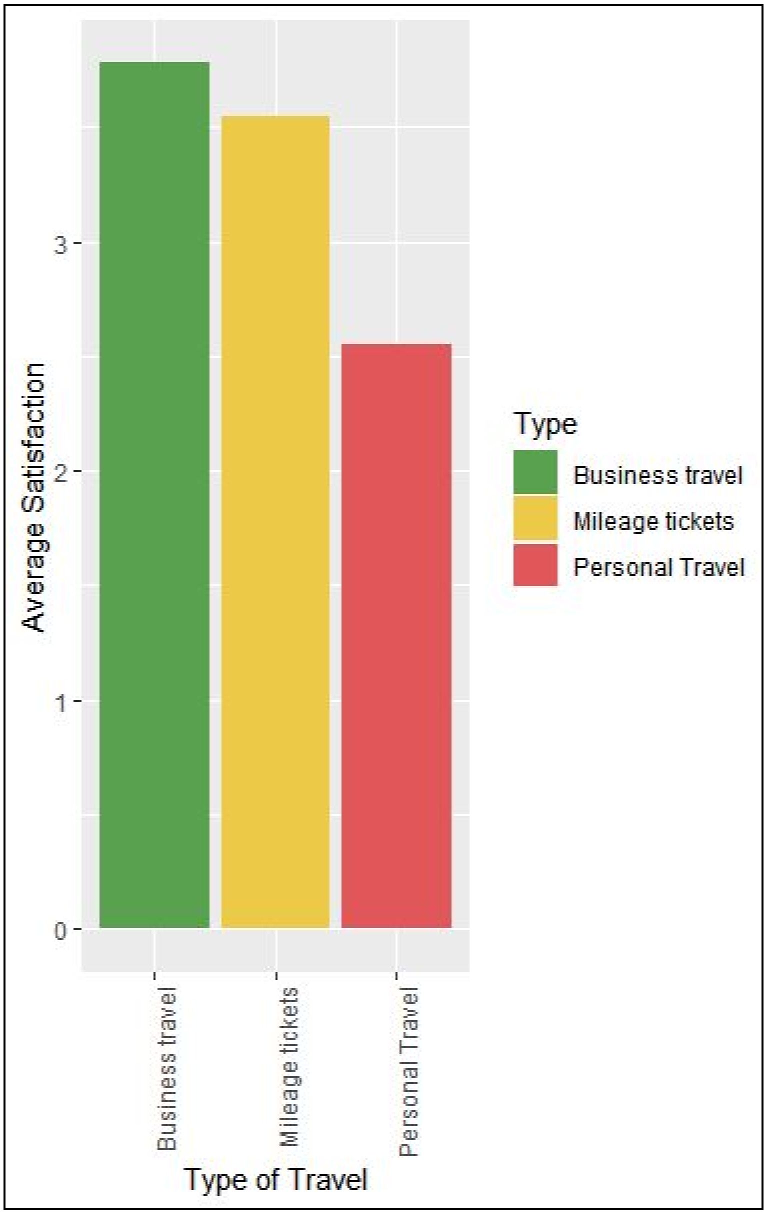
Make it easier for customers to move from Blue to Silver status



Though our Linear model and Logistic Regression, the team found that Airline Status variable had a significant impact on customer satisfaction. Further on applying association rules to the 5 significant variables from our prior models, we found airline status ‘Blue’ leads to low satisfaction. Through the Decision Tree model, we learnt that the Satisfaction rate for Gold, Silver and Platinum members are high. So this led us to conclude that moving customers from Blue to higher status will result in a positive impact on Satisfaction. While satisfaction remains largely the same, the increase from Blue to Silver is significant enough that we recommend making customers more aware of this option and potentially altering the guidelines to make obtaining Silver status easier since it is the status after Silver.

2. Increase mileage-ticket travellers:

Offer a variety of opportunities for customers to acquire miles efficiently so that more flights are taken via earned airline miles



Both our regression models, linear and logistic, showed Type.of.Travel to be a significant variable. Further the decision tree showed business and mileage travelers had higher satisfaction than personal travel. All these results led us to the conclusion that customers are generally happier when flying on mileage tickets, likely because they feel as though they are being rewarded. Although customers are somewhat price sensitive, we believe that it may be worth considering marginally increasing prices and providing extra miles as a result. In this way, customers will not only be flying more on mileage and be generally happier as a result.

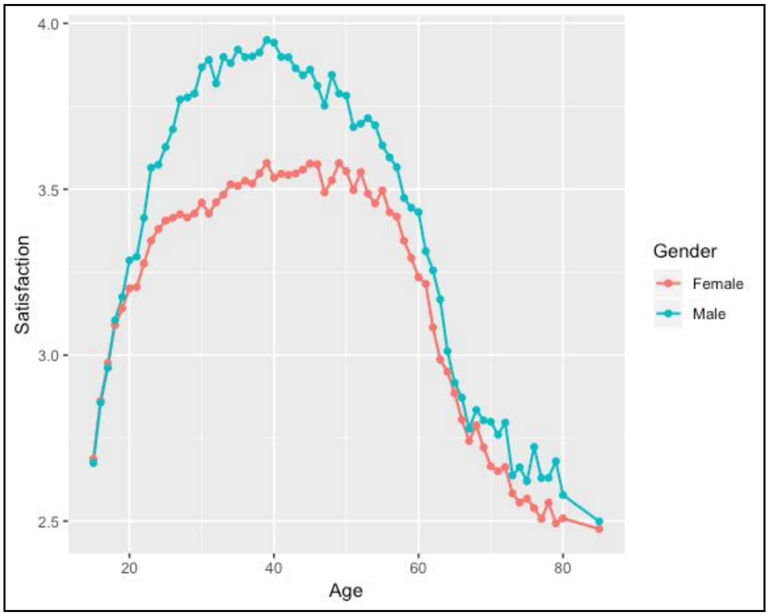
3. Avoid Delay:

Where possible, avoid arrival delays. If they occur, adequately compensate customers

Our logistic and decision tree models show that arrival delay significantly affects customer's satisfaction level. Even though there are multiple uncontrollable factors such as weather conditions relating to arrival delay, but airline companies still have certain control over it. For instance, they can improve the the human factor in air-traffic performance. Using historical data about what causes delays, and develops algorithms that help air-traffic controllers and pilots better manage who's taking off and when, and be able to predict congestion. Beyond that, airline companies can pay for meals, accommodations, future travel discount or other compensation.

Further Research:

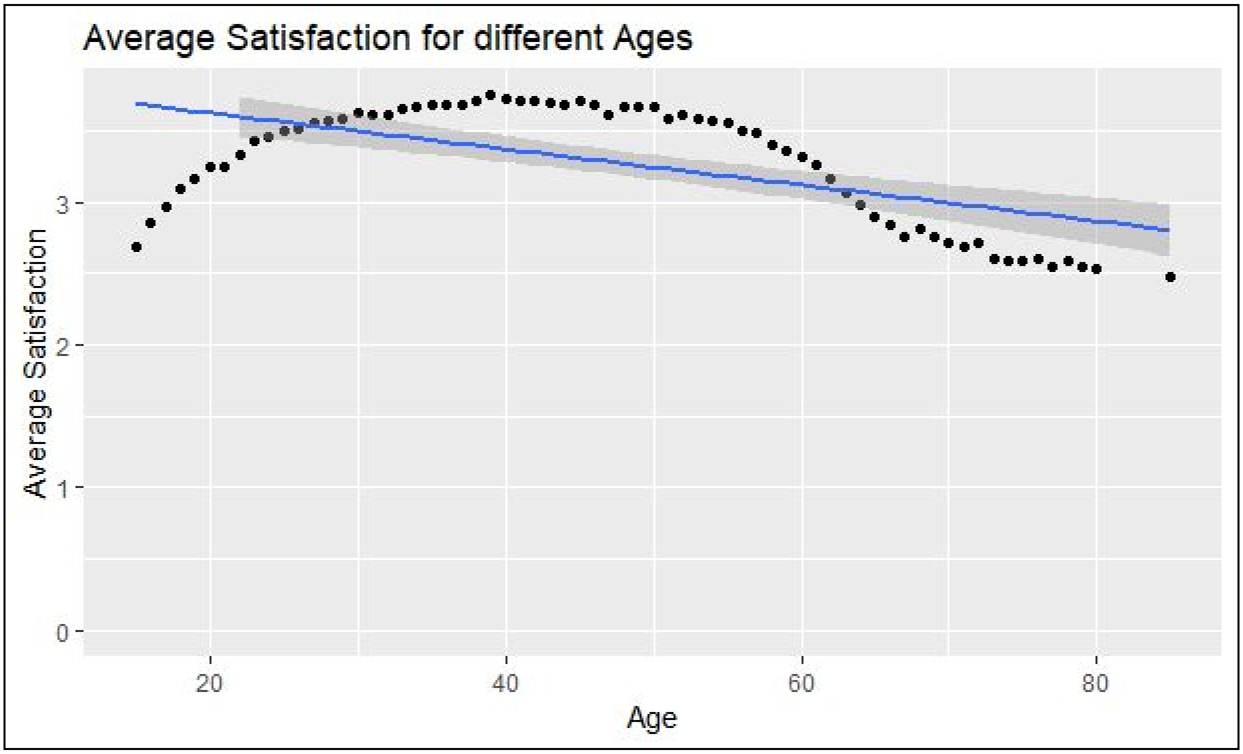
1. Dissatisfaction of women between their teenage and senior years.



Age vs Satisfaction by Gender

Between the ages of 20 and 55, women are significantly less satisfied with their airline experiences. Potential reasons for this are that they are traveling with children or for business but another much more insidious possibility is that they are concerned with their safety. Between the years of 2014 and 2017, FBI investigations into sexual assaults that occured on commercial flights have increased by 66%. The number of investigations has increased from year to year within that span as well from 38 in 2014, to 40 in 2015, to 57 in 2016, to 63 in 2017. While this is an observation, it is anecdotal evidence at best and will require further research to definitively determine any concrete causation.

2. Dissatisfaction of customers in their retirement years and beyond.



Age vs Average Satisfaction

Considering the trend of Satisfaction vs Age there is a clear drop off in satisfaction between the ages of 60 and 65 which continues decline the older customers get. When considering why this might be, it occurred to us that most passengers around this age are likely retired and may require additional accessibility support such as wheelchair access or space on the airplanes that better accommodates them. Surveying this older demographic specifically may provide more insight into what would appease them.