Table of Contents

[**Introduction to Dataset**](#_mhrro3tcuga4) **2**

[**Objective**](#_y137x1y7mn9) **3**

[**Process**](#_xdip8swhux9c) **3**

[**Business Questions**](#_w61pggh6fuyw) **4**

[**Initial Analysis**](#_93ktlo2ilr7) **5**

[Summary of the Dataset](#_t5fslz5vr5at) 5

[Structure of the Dataset](#_tnk44p5tu5a6) 7

[Cleaning the Dataset](#_e2lq45hfedl2) 8

[I. Demographics](#_qlfd3pt0mxy0) 8

[Age](#_arik3nopdsxl) 8

[Histogram of Age Filled with Satisfaction](#_e2q61p2dqfwu) 8

[Gender](#_8nn943tzb54d) 9

[Pie Chart of Gender](#_v1dqm0b83zus) 9

[Histogram of Gender Filled with Satisfaction](#_ngtmds69r2nh) 10

[Age and Gender](#_dpty6m7b5dfq) 11

[II. Travel preferences](#_jspgyre1jvph) 13

[Airlines](#_3z2mfaa9tv04) 13

[Histogram of Airline Companies Filled with Satisfaction](#_m7a3zg15qoi3) 14

[Customer Status](#_pdi6ro7ky6qq) 14

[Plot of Airline Status Filled with Satisfaction](#_i5d46f97e9jk) 15

[Customer Class](#_6n83la9q67u3) 15

[Class Filled with Satisfaction](#_a50s9s807epp) 16

[Type of Travel](#_kkb54zkz0ut4) 16

[Type of Travel Filled with Satisfaction](#_64n7c6kgumdb) 17

[Pie Chart for Type of Travel](#_wajs82ep9c6o) 18

[Price Sensitivity](#_l49vowi1p81p) 19

[Histogram of Price Sensitivity Filled with Satisfaction](#_y7eriigyve8p) 19

[Flights Per Year](#_lpmdgdq9iz3u) 20

[Histogram of Flights Per Year Filled with Satisfaction](#_arzz7kb2eo9y) 20

[Airport Expenditure](#_hveoql8n35e5) 21

[Spending Money at the Airport by Rank and Type of Travel](#_1470nhyd7d2q) 21

[States](#_str3cnjqkq5e) 22

[Origin States with Satisfaction Rates](#_o7n5id436nvd) 22

[Destination States with Satisfaction Rates](#_k26uudvo6nuj) 23

[Lollipop Chart for States vs. Mean Satisfactions](#_ang5ihq5gld3) 24

[III. Customer Satisfaction vs. Loyalty](#_gd156kjvrwxo) 25

[**Data Subset - I**](#_9pldunp5tpkw) **26**

[**Data Modeling - I**](#_oisrtehfspve) **26**

[Linear Model](#_xb8gr5v2m1cv) 26

[Association Rules](#_76qxlmkej91m) 28

[Support Vector Machine](#_5ypi83qub7ek) 29

[**Data Subset - II**](#_cc88atatj61r) **30**

[**Data modeling - II**](#_v70tc37868op) **34**

[Linear model](#_i3sntt7908fe) 34

[Further Analysis](#_l9fc1ypoq4g5) **37**

[Linear Model](#_861fzvvy2uas) 37

[Association Rules](#_viu7cm3ih7k1) 42

[**Actionable Insights**](#_fyrnb0j47yrb) **43**

[**Recommendations**](#_122dcd4la47r) **44**

[**Appendix 1**](#_mumfledennsy) **45**

[**Appendix 2 (Code)**](#_tgv3tn8i037e) **47**

[Descriptive Analysis](#_nax1sf982hc4) 47

[Visualization Code](#_kjjvwsl3059q) 53

[Linear Modeling](#_h1p6eexddxdb) 60

# Introduction to Dataset

This dataset contained information on the customer satisfaction of 194,833 airline customers based on 29 different variables across 14 partner airlines. (Refer to Appendix one for details regarding each of the variables) The key variable in this dataset is the customer satisfaction that the partner airlines are trying to maximize. Each row in the dataset represents one flight segment by one airline customer. The dataset captures characteristics of the flight as well as characteristics of the customer. The dataset also contains a simple survey-based rating of each customer’s satisfaction with the flight segment.

For example, this dataset contains flights to 295 different cities, within the fifty states, Puerto Rico and U.S. Pacific Territories. These flights were from one of fourteen partner companies. Each customer belonged to one of four airline status: blue, silver, gold, or platinum.Also each customer was able to fly in one of three classes: eco, eco plus, or business. The customers had three types of travel: personal, mileage and business. The dataset also contained information pertaining to airline routes, origin and destinations. Additionally, there was information regarding departure delay and arrival delay times.

The team will use the R software to this dataset and derive actionable insights from analysis before presenting its recommendations to the management teams at the partner airlines to improve the satisfaction of their customers. This report will elaborate the steps involved in the analysis of this dataset, key takeaways or actionable insights from this and a few recommendations that the team believes would be useful to the partner airlines companies.

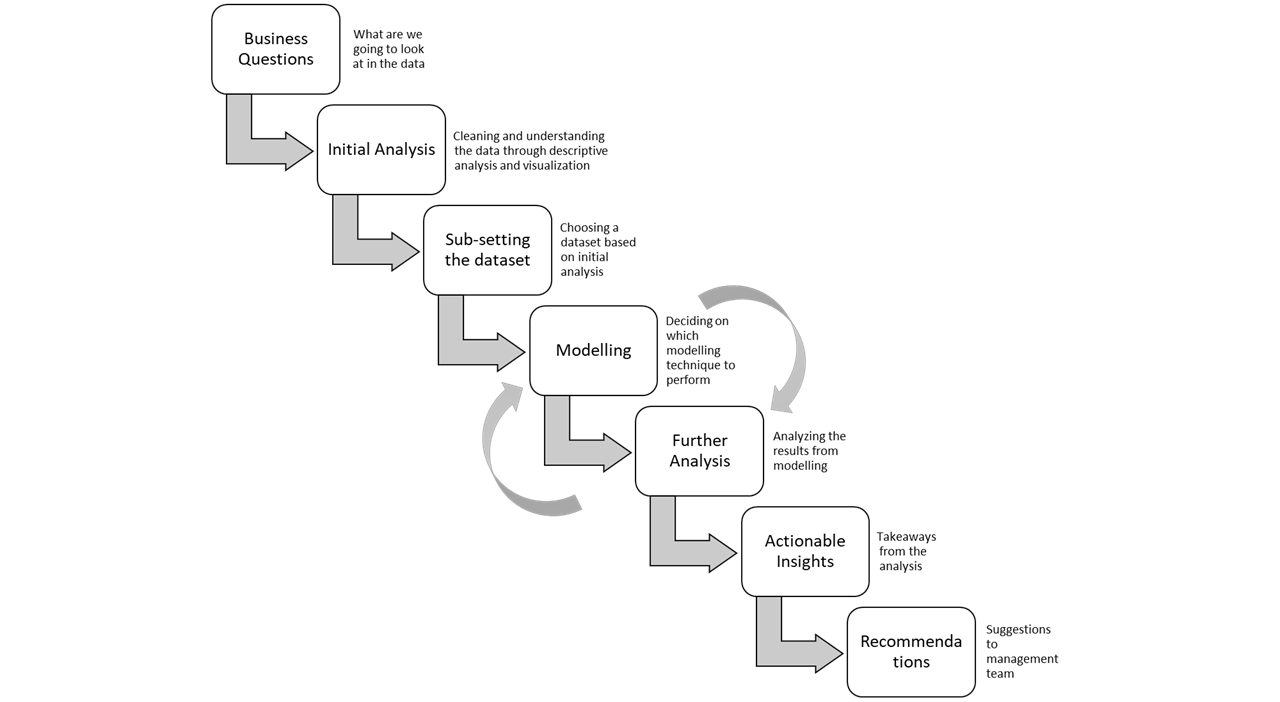
# 

# Objective

The objective of this analysis is to identifying the customers who are not satisfied and increase their satisfaction. In other words, this analysis should be able to provide information about what factors affect the satisfaction of the customers. Thus, the overall objective is to ensure how to maximize the overall satisfaction of the customers especially the unsatisfied ones.

# Process

In order to achieve this objective the team followed the process shown below. The team first defined the business questions before cleaning the dataset and performing initial analysis on the data. During the initial analysis descriptive analysis and visualization techniques were used. Based on this analysis, the data was then sub-setted and data modeling was performed. If the data model did not work, the team revisited the sub-setting part and used a new subset of the data to perform data modeling. Further analysis was then performed on the data model results to get actionable insights on the basis of which the team presented its recommendations to the management team of the partner airline companies.



# 

# Business Questions

The business questions were divided into three broad categories - customers, airlines and states. For the customers, we look at questions that help us understand why customers are satisfied or not. While looking at questions pertaining to airlines, we look at factors affecting the airlines such as delays, origin city, departure city, and partner airlines to see their effect on customer satisfaction. For states, the team looked at how departure delay and arrival delay are distributed across states and what impact does it have on customer satisfaction. The team identified the following business questions to be considered to analyze the data.

For Customers:

(i) Which status customers are least satisfied?

(ii) Which type of travel customers are satisfied? Which are not?

(iii) What is satisfaction according to Customer Status and which is least satisfied?

(iv) Which class of customers are more satisfied and which are the least?

(v) Does age affect customer satisfaction? Which age group is highest/least satisfied?

(vi) Does gender impact customer satisfaction? What are the demographics of the dataset? Which gender is least satisfied?

(vii) What is the overall customer satisfaction rate according to the airline status?

(viii) Analyze the price sensitivity with Customer Satisfaction.

(ix) Analyze number of flights per year with Customer Satisfaction.

(x) Analyze spending money at the airport by type of travel.

For Airlines:

(i) What is the customer satisfaction for partner airlines and which airlines has the lowest/highest Customer Satisfaction?

(ii) Does delay time affect Customer Satisfaction?

(iii) Does Loyalty to an airline affect the Customer Satisfaction?

(iv) Which airline has the highest number of Customers?

(v) Compare the significant factors for Cheapseats, Going North and West Airlines.

For States:

(i) Which state has the least outgoing/incoming flights satisfaction rate?

(ii) Does departure delay affect the satisfaction of a customer for a particular state?

(iii) Which state has the least satisfaction rate in Blue airline status and Eco class?

(iv) Which state has the highest satisfaction rate in Blue airline status and Eco class?

(v) Which state has the highest customer satisfaction rate overall?

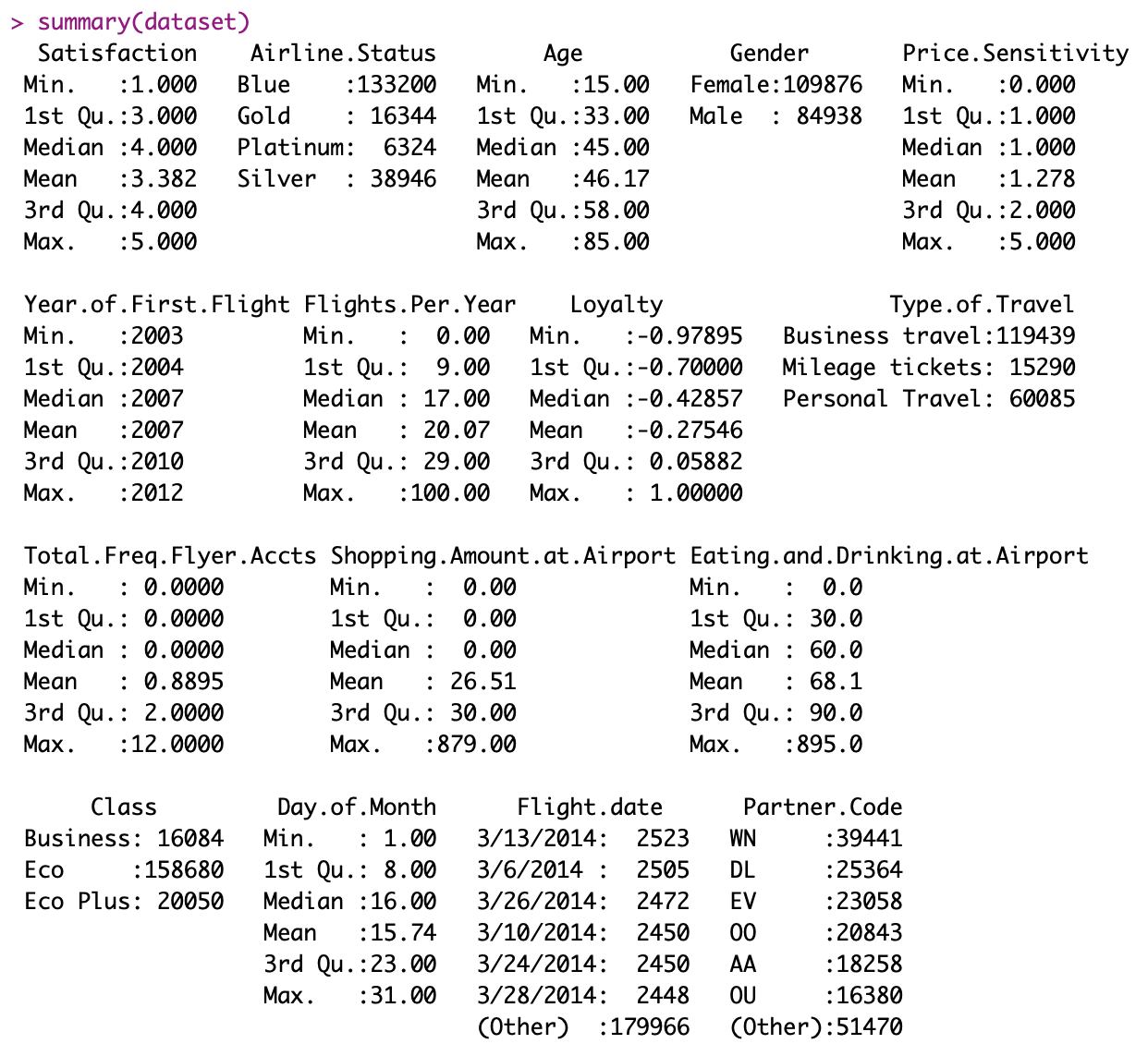
(vi) Find out which state has the least satisfaction rate when it is an origin state. Analyze this in Blue and Eco class.

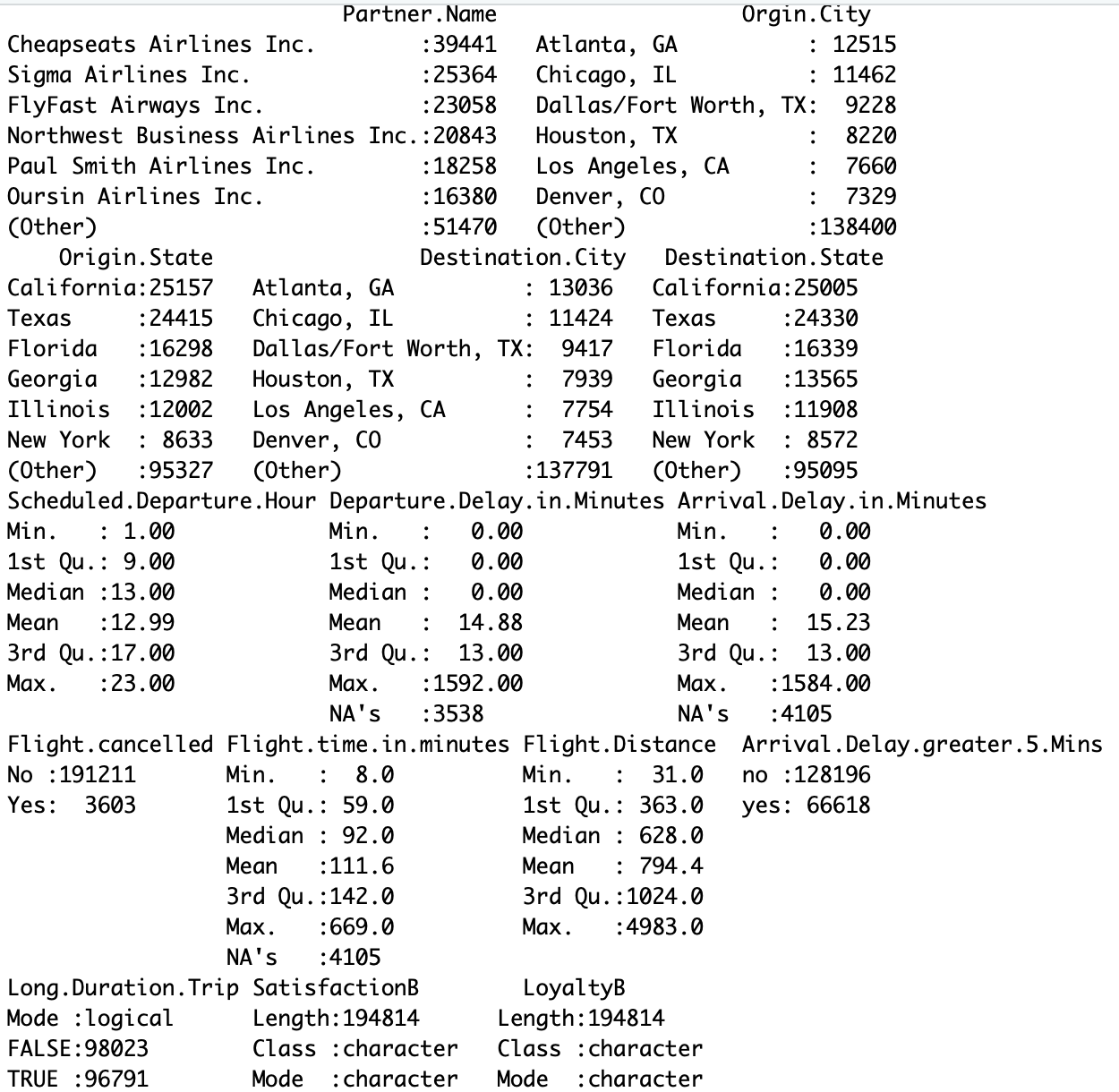
(vii) Find out which state has the least satisfaction rate when it is a destination state. Analyze this in Blue and Eco class.

# Initial Analysis

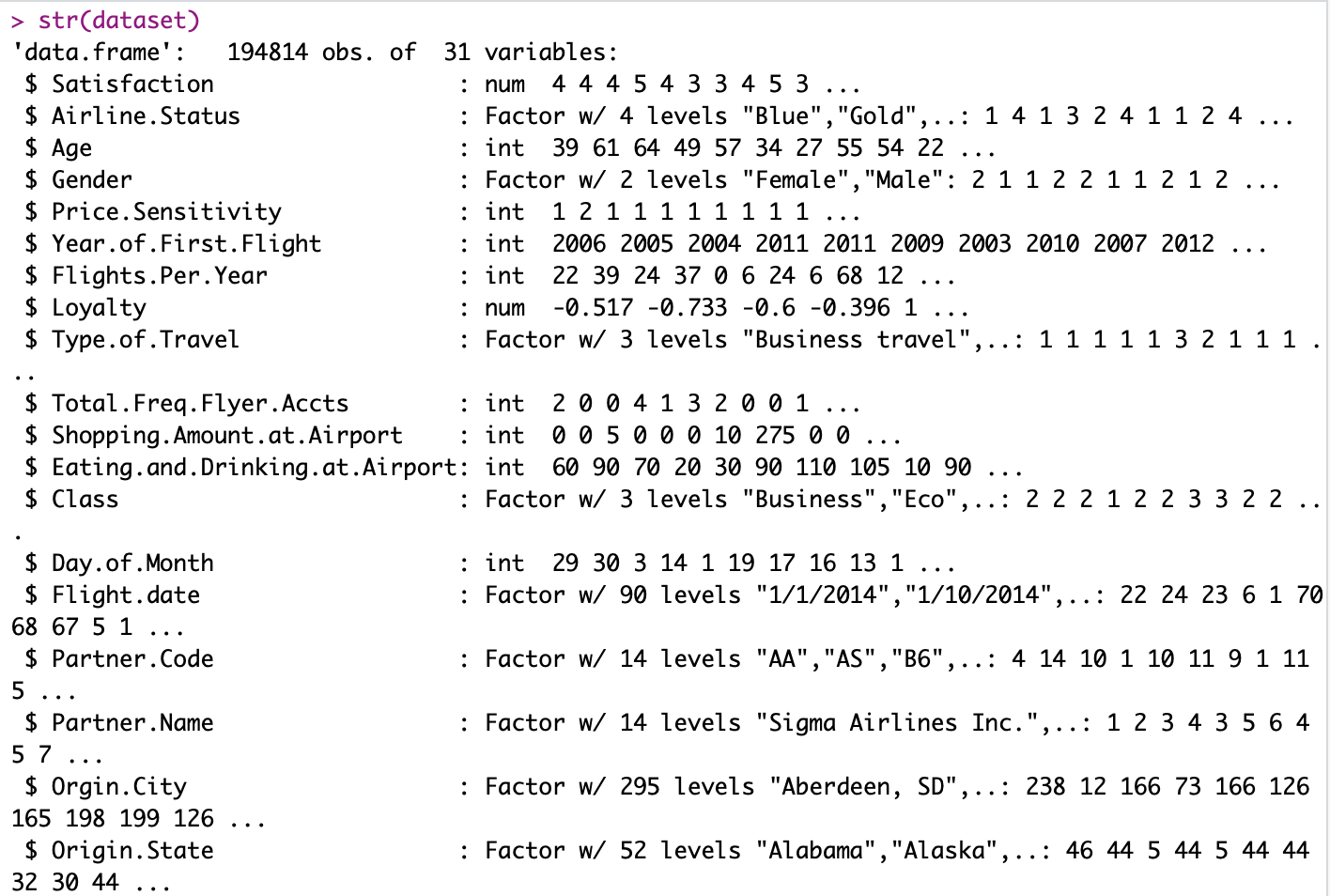
Before conducting the initial analysis, the summary and structure commands were run to get an idea about the descriptive statistics as well as the data structure of the dataset.

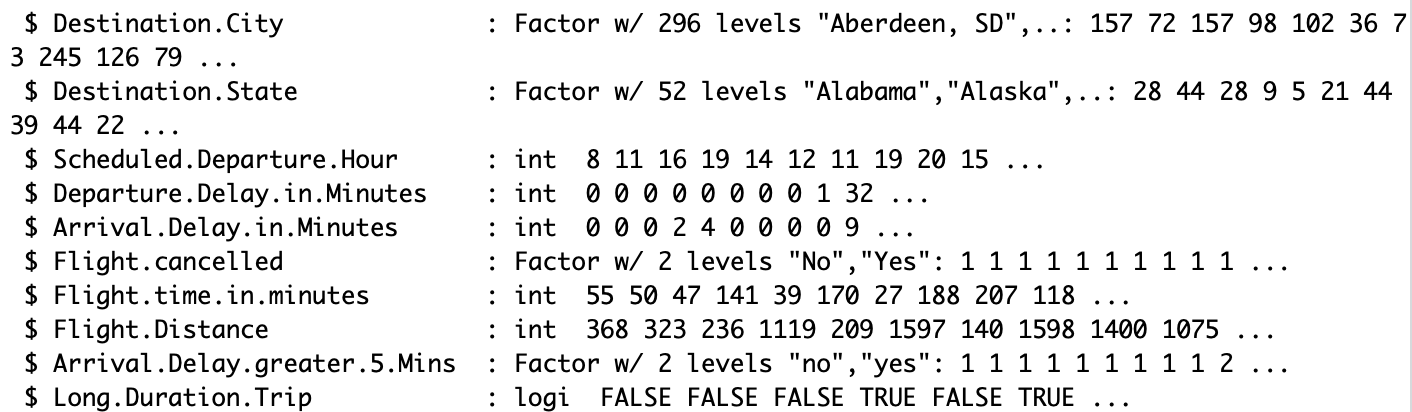
## Summary of the Dataset





## Structure of the Dataset





## 

## Cleaning the Dataset

Prior to performing the descriptive analysis the team cleaned the data using the complete cases command. The following is just one example. This was carried out on other columns in a similar fashion.

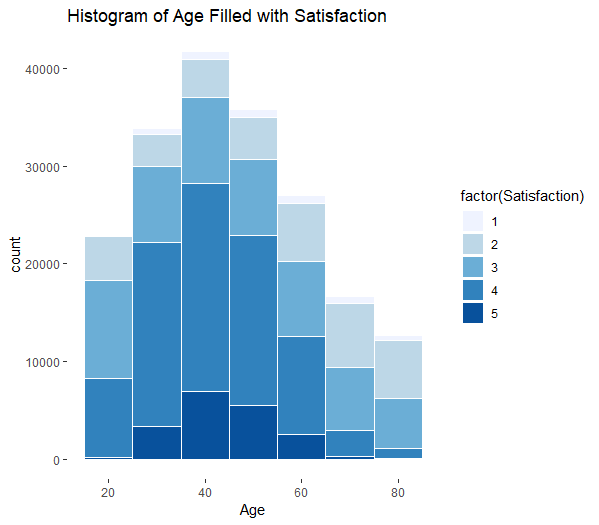
# Remove NAs and Clean the Data

dataset <- dataset[complete.cases(dataset$Partner.Name),]

### I. Demographics

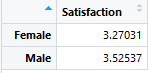
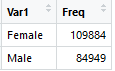
#### Age

##### Histogram of Age Filled with Satisfaction



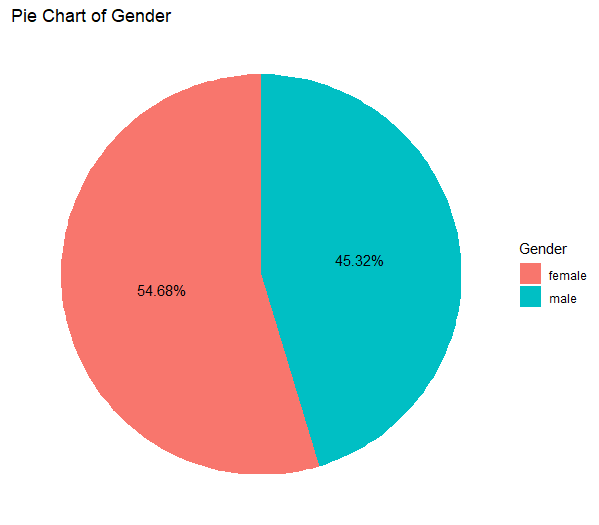
This histogram shows the number of customers in each level of satisfaction based on age. The plot shows that the happiest (most satisfied) customers are between the ages of 30-60 years old. We would need to improve the satisfaction for those under the age of 30 and those over the age of 60.

#### Gender



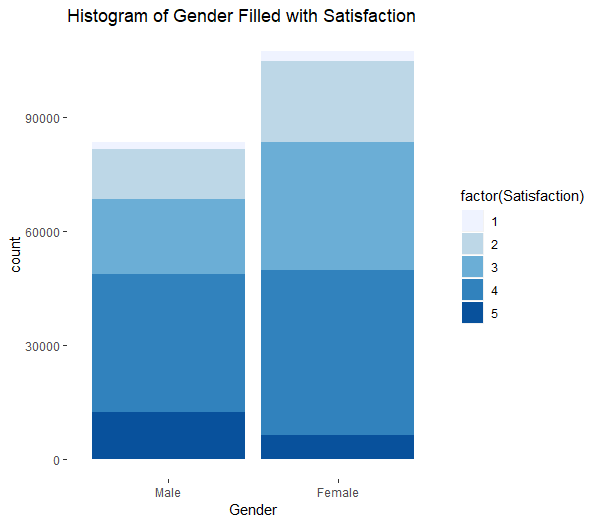
We found that women made up the majority of this dataset, and women had a lower satisfaction than men.

##### Pie Chart of Gender



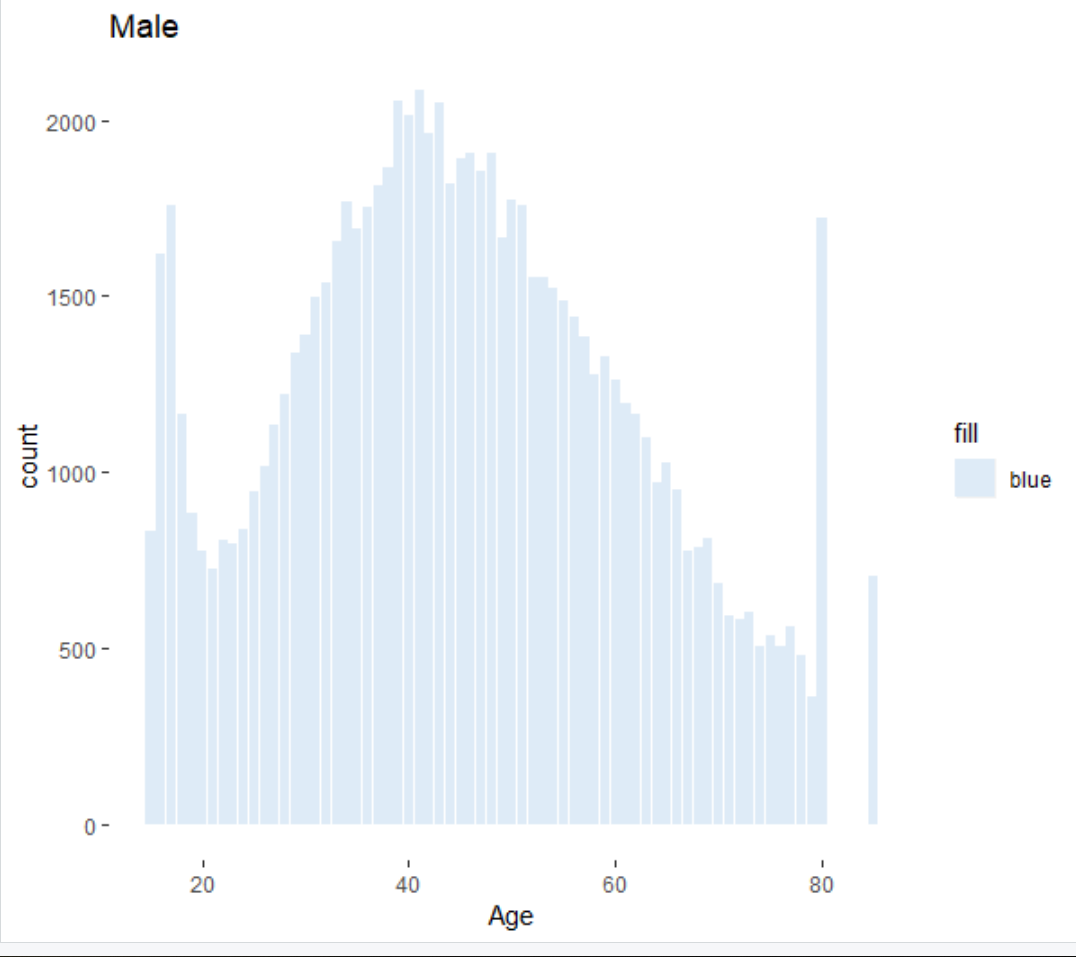
This pie chart shows the breakdown of customers in this dataset based on gender. 54.68% of the customers were women and therefore, they made up the majority of the data.

##### Histogram of Gender Filled with Satisfaction

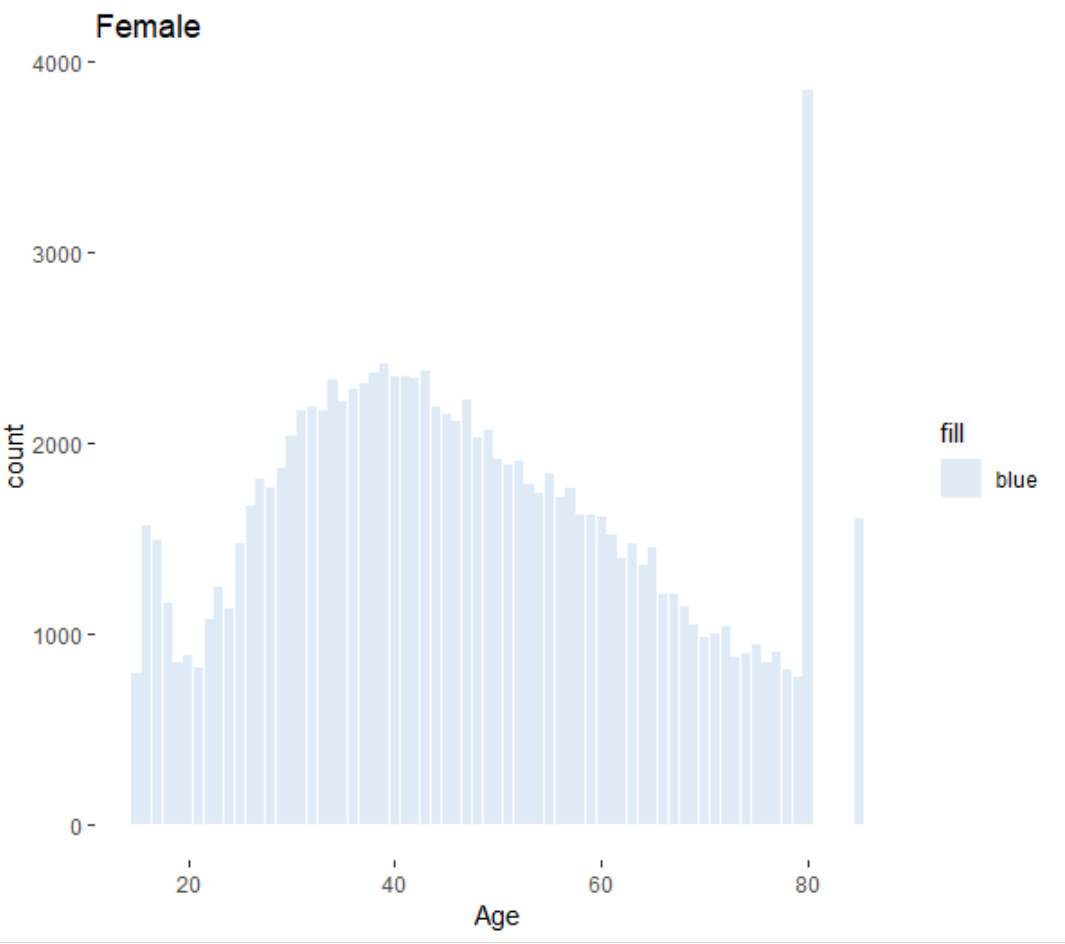


This histogram shows the number of customers in each level of satisfaction based on gender. This plot also shows that the majority of customers were women. And although women are the majority, they are also particularly unhappy (not satisfied).

#### Age and Gender

****

The above chart shows the total number of Male gender customers in the dataset according to their age. It is observed that the dataset forms a nice bell curve, with majority customers ranging between 35 to 50.

****

The above chart shows the total number of Female gender customers in the dataset according to their age. It is observed that the dataset forms a nice bell curve.

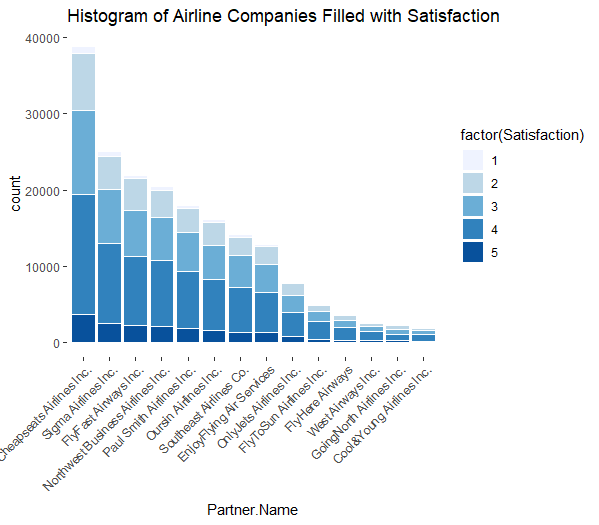
### II. Travel preferences

#### Airlines

#### 

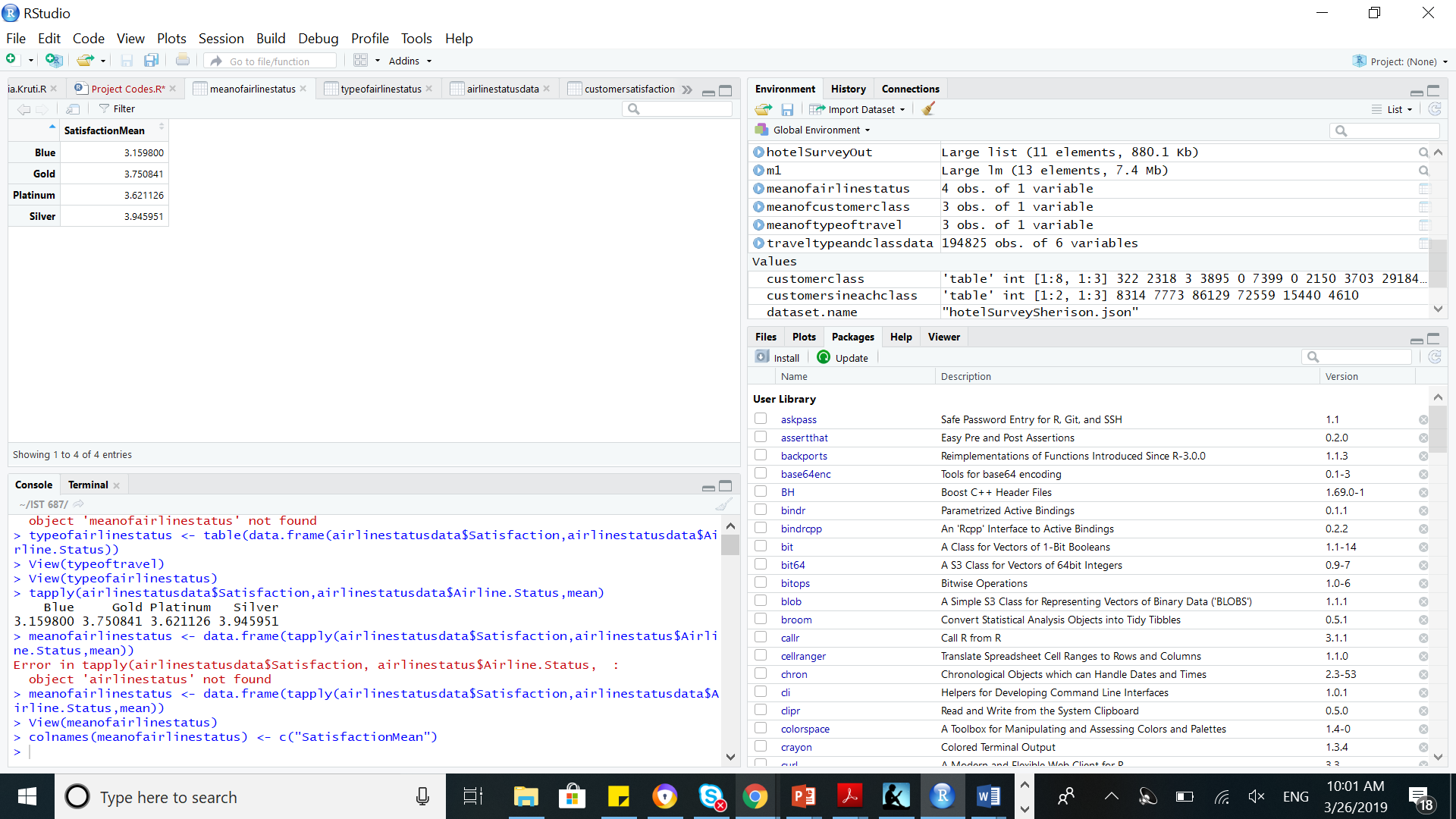
After looking at satisfaction based on the airlines, we found that West Airways had the highest customer satisfaction while GoingNorth Airlines had the lowest. However, we also looked at the volume of customers each company had and saw that Cheapseats Airlines had the most customers while GoingNorth Airlines and West Airways had the lowest two customer totals in comparison to the other partner companies. Therefore, we decided to look into Cheapseats since it had the largest total of customers and the third lowest customer satisfaction. We will also look into GoingNorth and West as comparison for Cheapseats.

##### Histogram of Airline Companies Filled with Satisfaction



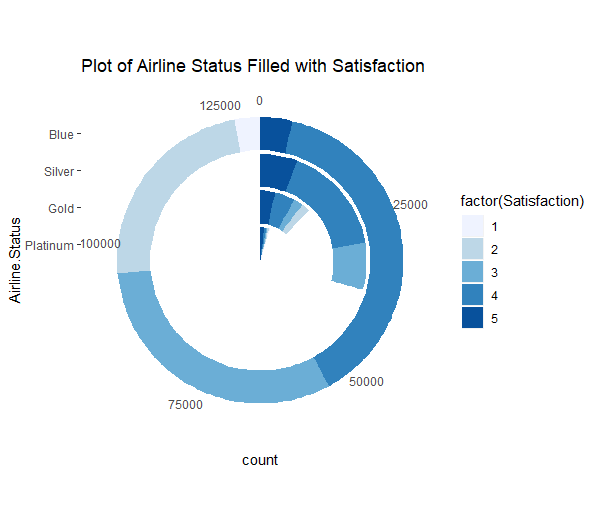
This plot shows the breakdown of blue status customers and their satisfaction level based on the airline that they flew. Here we can see that Cheapseats has the majority of blue status customers and Cheapseats also had the lowest overall customer satisfaction compared to all of the other partner airlines.

#### Customer Status



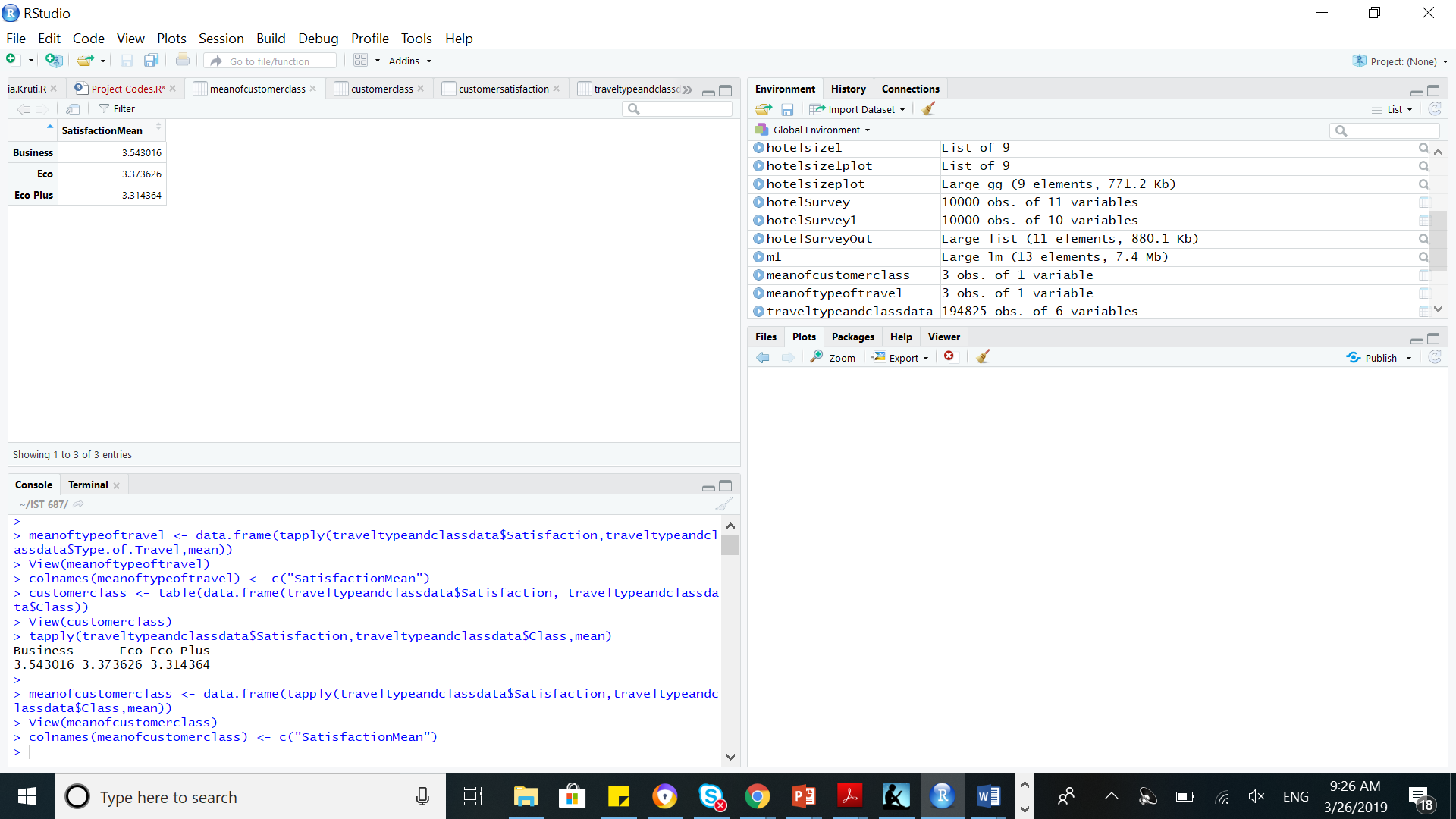
For customer status, blue status had the lowest customer satisfaction rate while silver had the highest for this dataset. However, we recognized that blue status contained the majority of the customers in the total dataset but also had the lowest satisfaction rate. It was one of the deciding factors when forming the subset that we would focus on to improve customer satisfaction.

##### Plot of Airline Status Filled with Satisfaction

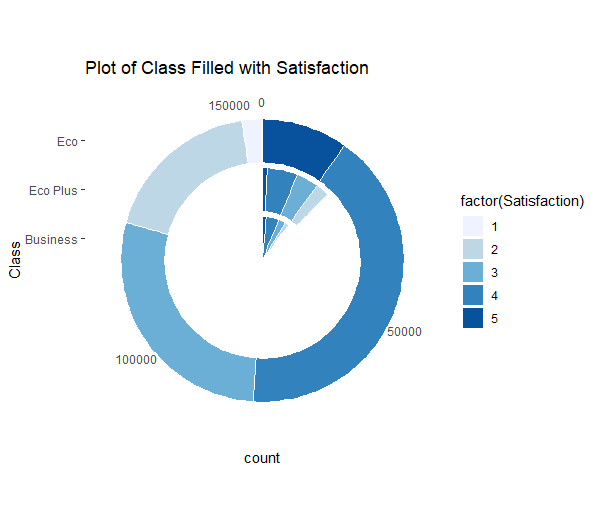


This plot shows the number of customers in each level of satisfaction based on airline status. Blue status contained the most customers and had more customers in Satisfaction levels 4 and 5 than the other status’ have combined. This shows that we should focus on improving the satisfaction of Blue status customers since they make up the majority.

#### Customer Class

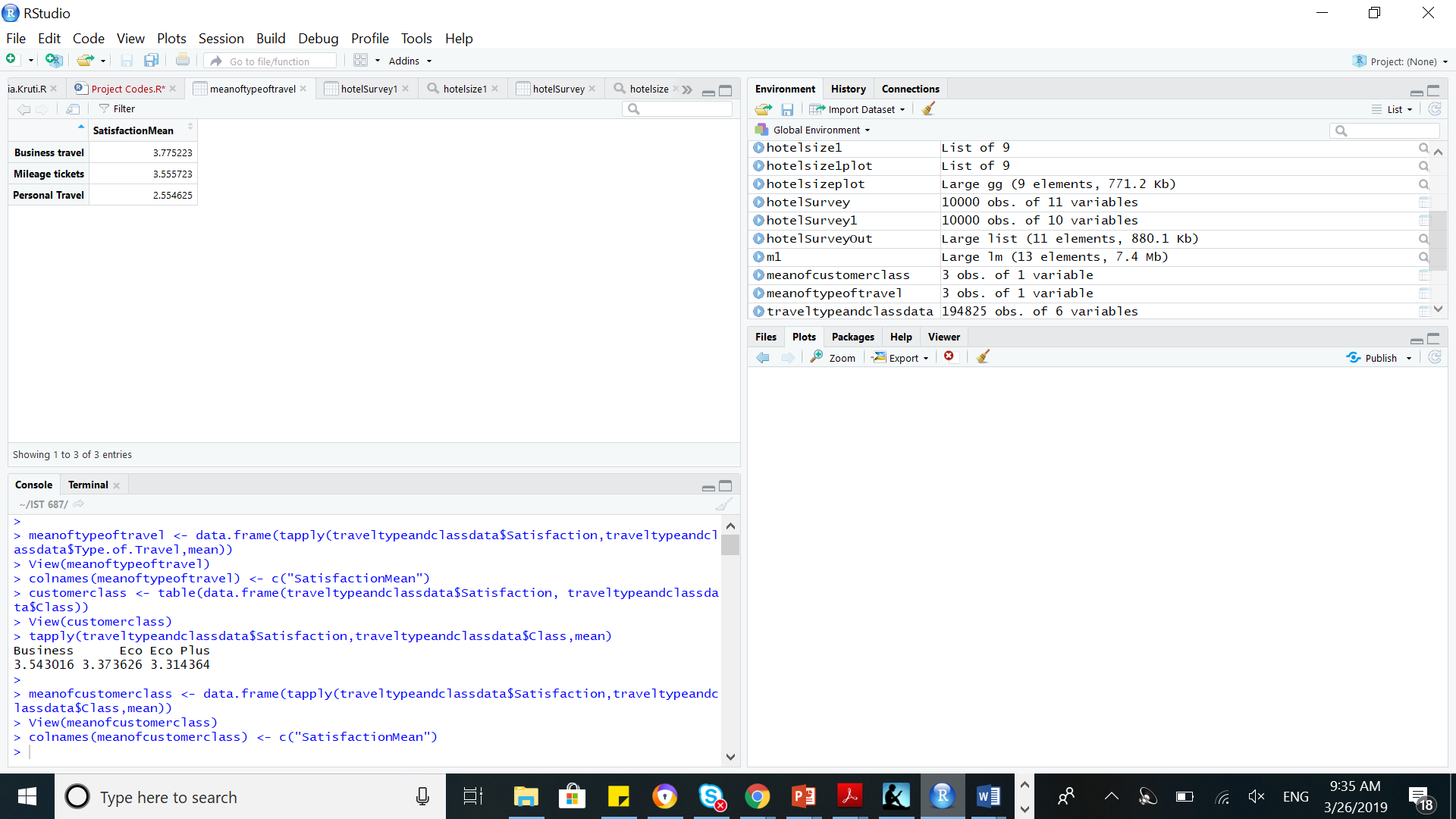
Customer class had a difference in satisfaction for each of the three classes. Business class had the highest satisfaction rate of the three classes while eco plus had the lowest. However, we noticed that a majority of the customers flew eco class and decided to further investigate into this class for further analysis.

##### Class Filled with Satisfaction



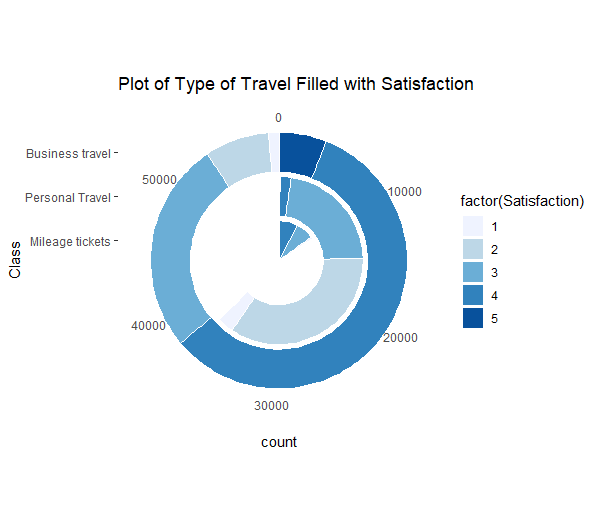
This plot shows the number of customers in each satisfaction level based on class. Eco class made up the majority for this dataset. Due to the large count of customers who flew in eco class, we decided to make this a subset and look into increasing overall customer satisfaction.

#### Type of Travel



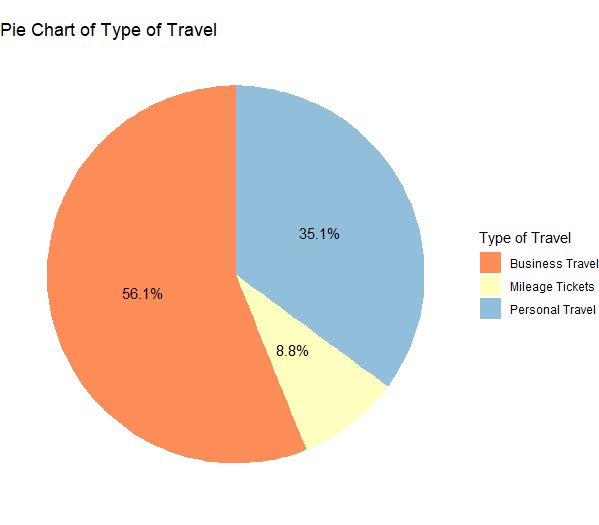
For our subset of blue eco customers, type of travel played a role in overall customer satisfaction. Business travel had the highest satisfaction rate whereas personal travel had the lowest.

##### Type of Travel Filled with Satisfaction



These plots show the statistics for the Blue Eco customers that we decided to subset from the data. The first pie chart shows the percentages of customers flying based on type of travel. Although business travel was the majority of the customers, personal travel made up over a third of the dataset. The second plot shows the number of customers in each level of satisfaction based on type of travel. Business travel was the majority and were the most satisfied. However, personal travel were dissatisfied.

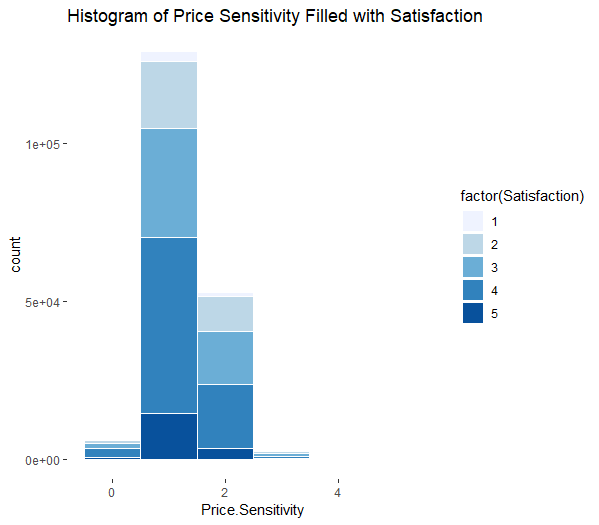
##### Pie Chart for Type of Travel



In the above Visualization, the data that is used to generate the pie chart has customers belonging to Blue airline status and Eco class. It is evident from the chart that most customers from this category are Business travelers with more than half in this sub-category. Then, the second highest type of travelers are Personal travelers and lastly there is a small percentage of Mileage tickets.

#### Price Sensitivity

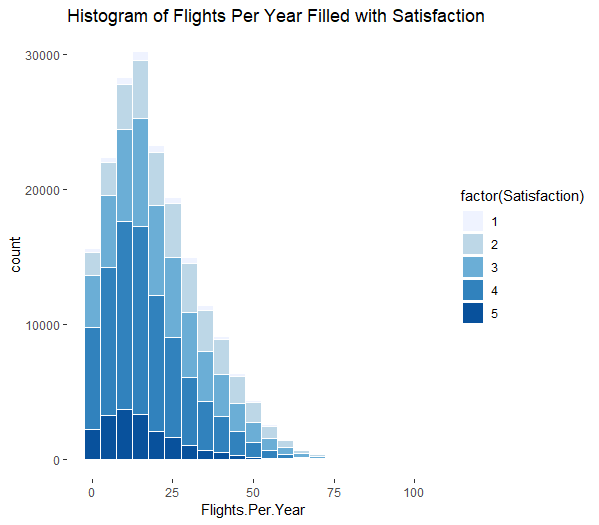
##### Histogram of Price Sensitivity Filled with Satisfaction



This histogram shows customer satisfaction based on the price sensitivity. Most customers rated price sensitivity as either 1 or 2, therefore it was an insufficient way to conclude satisfaction.

#### Flights Per Year

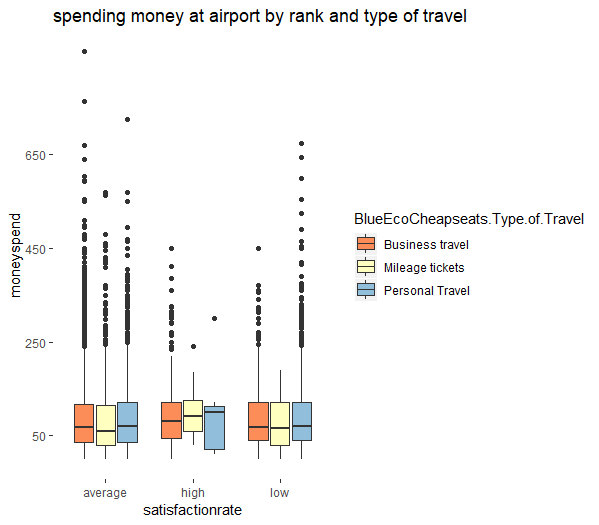
##### Histogram of Flights Per Year Filled with Satisfaction



This histogram shows the number of customers in each satisfaction level based on the number of flights they take per year. People who took 5-15 flight.

#### Airport Expenditure

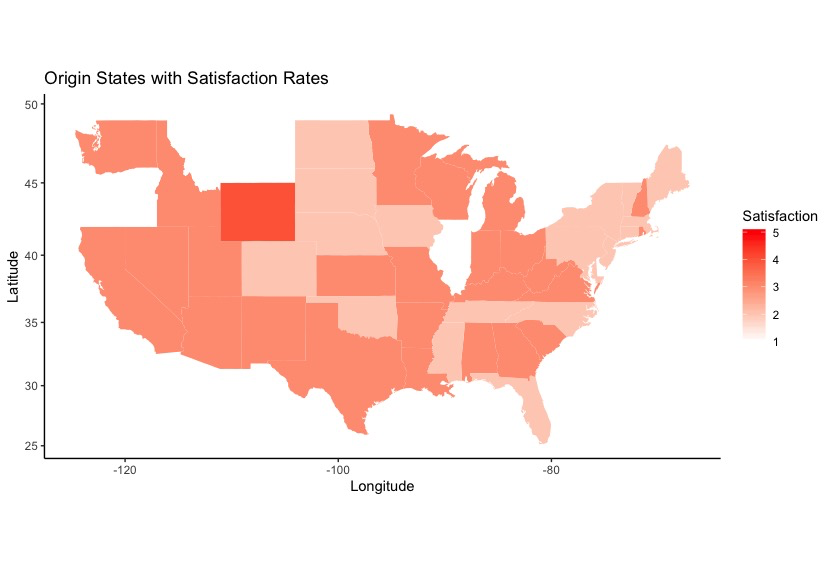
##### Spending Money at the Airport by Rank and Type of Travel



This plot shows different satisfaction rates based on type of travel and money spent at the airport. Low satisfaction is defined as a satisfaction rate less than 3, average is defined as satisfaction rate equal to 3, and high satisfaction rate is defined as satisfaction rate greater than 3. We found that the satisfaction of personal travel was high when they spent the least amount at the airport.

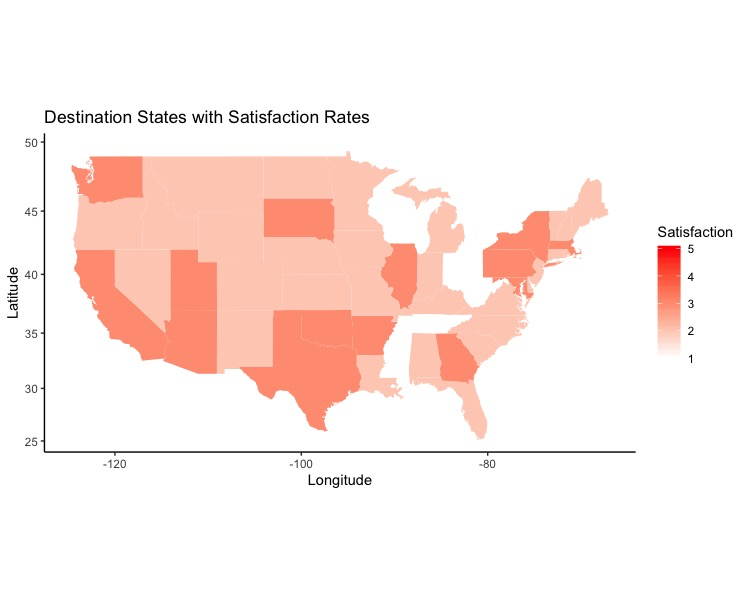
#### States

##### Origin States with Satisfaction Rates

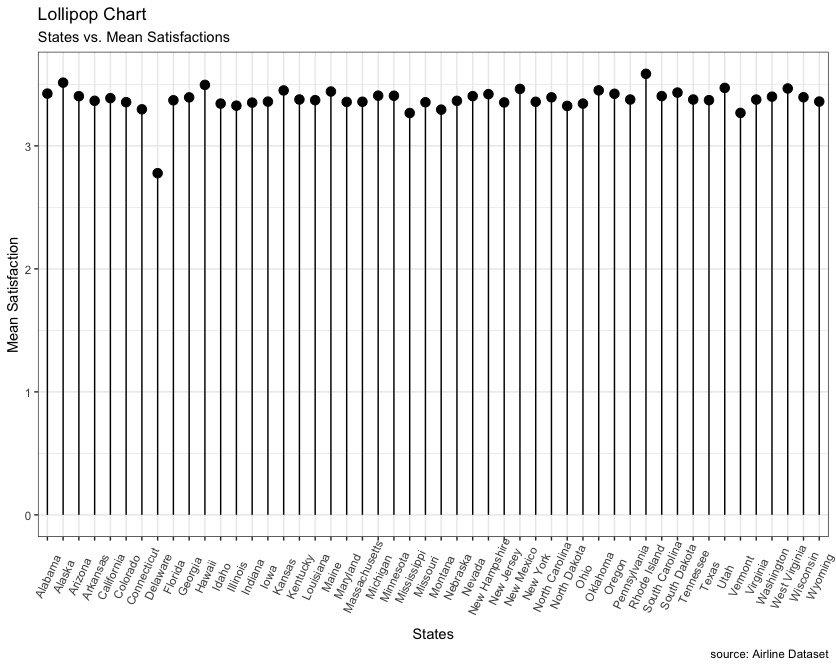


Data: The Data used to create this visualization was a subset of the original data set. We chose Blue customers, since Blue customers had the least satisfaction of all the statuses. Then, we chose Eco class since that had the least satisfaction. Further we chose Personal travelers since they had the least satisfaction. Without any surprise, the plot shows that the satisfaction is low, especially in Delaware, where it is the least. It is not evident from the map, but when we create a table with all the states and their mean satisfaction rates, Delaware has the least. For this plot, the Origin States were taken and then in the following plot, it was compared to Destination States.

##### Destination States with Satisfaction Rates



##### Lollipop Chart for States vs. Mean Satisfactions



From the above graph, it is evident that Delaware has the lowest mean satisfaction of all the states and Rhode Island has the highest satisfaction. Upon further analysis, we calculate the mean Arrival Delay in Minutes for both Rhode Island and Delaware. But to our surprise, we also found out that Delaware had a total of 9 entries for Origin State and 20 entries for Destination State. That meant that there were only 29 entries for Delaware to conduct analysis, and therefore it was no surprise if Delaware’s ratings were skewed. But just for the analysis part, here were some observations: Firstly, all the flights coming in and going out of Delaware are through Going North partner airlines, which also has the lowest satisfaction overall. Then we also calculated the mean Arrival Delay in Minutes for both the States. Delaware had a mean delay of 28 minutes and on the other hand, Rhode Island had a mean delay of just 13 minutes (13.39 to be precise). But before making any other conclusions, that Departure Delay, let’s look at Mississippi, which is the second least satisfactory state to travel to/from. Our finding for this state was, that it had a mean Departure Delay of 12.34 minutes, which is less than that of Rhode Island. So there are definitely other factors which are influencing Satisfaction.

### 

### III. Customer Satisfaction vs. Loyalty

To analyse Customer Satisfaction against Loyalty, we first created buckets to organize the Customer satisfaction in Low, Medium and High Buckets. Similarly we also arranged the Loyalty into two buckets unlike satisfaction, into Low and High.

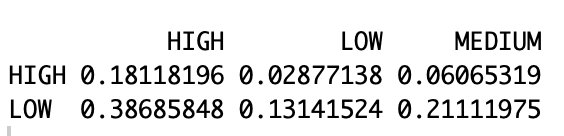
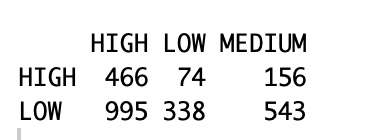
The buckets were created in the following manner:

For Customer Satisfaction: 1-2 (Low), 3 (Medium), 4-5 (High)

For Loyalty: Anything below 0 (Low), 0 and anything above 0 (High)

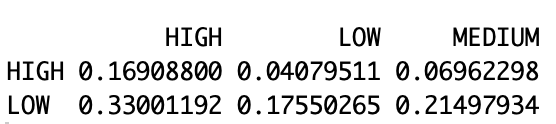
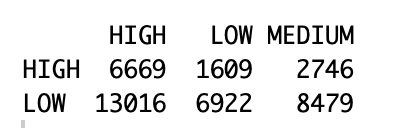
Based on this, we looked at different partner airlines so as to find out whether Loyalty impacts Customer Satisfactions. The following are some observations about different airlines.

West Airways

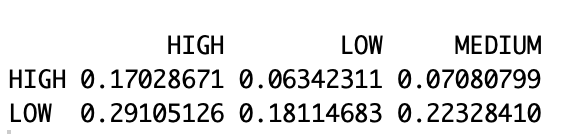
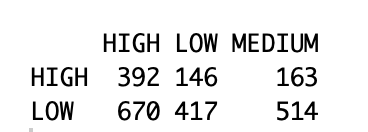


The first airlines that we looked at were West Airways. The reason behind selecting these airlines first was that West Airways has the highest mean satisfaction rate of all the partner airlines in the dataset. In the left diagram above, we can see the number of customers in every category. To reiterate, Loyalty has buckets Low and High, and Customer Satisfaction has buckets Low, Medium and High. As seen from the right figure, majority of the customers (about 38.6%) from the total customers, even though having low customer loyalty, show a high satisfaction rate.

To further our analysis, we looked at Cheapseats airlines as they had the highest number of travelers.



Upon some further analyses we observe a similar trend in Cheapseats Airlines as well. Majority of Customers (about 33%) from the total customers, even though having low customer loyalty, showed a high satisfaction rate. To sum it up, we looked a the airlines which had the lowest satisfaction rate, Going North Airlines.



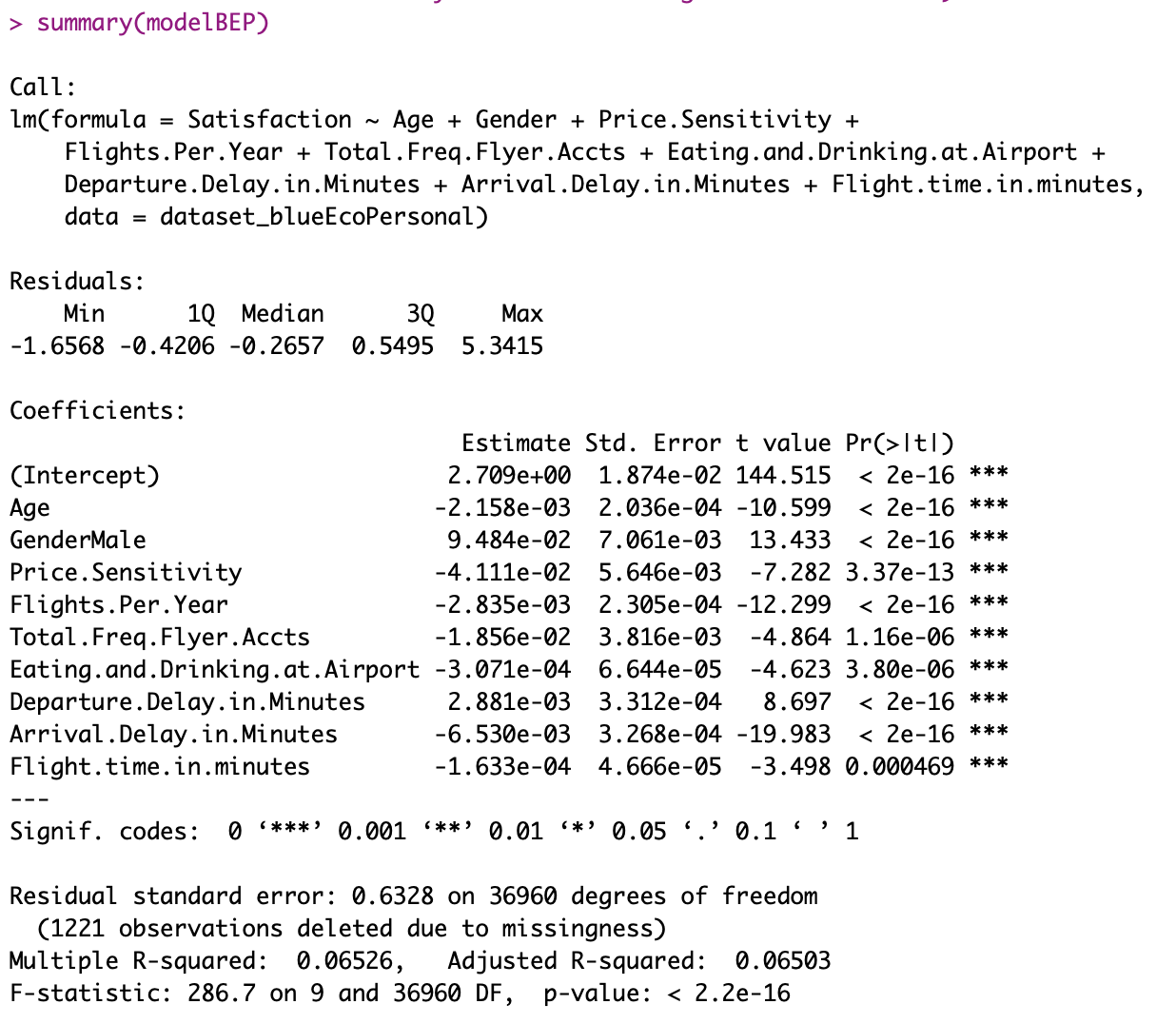
A similar trend is observed here as well, with majority of customers (about 29%) from the total customers, even though having low customer loyalty, show a high satisfaction rate.

# Data Subset - I

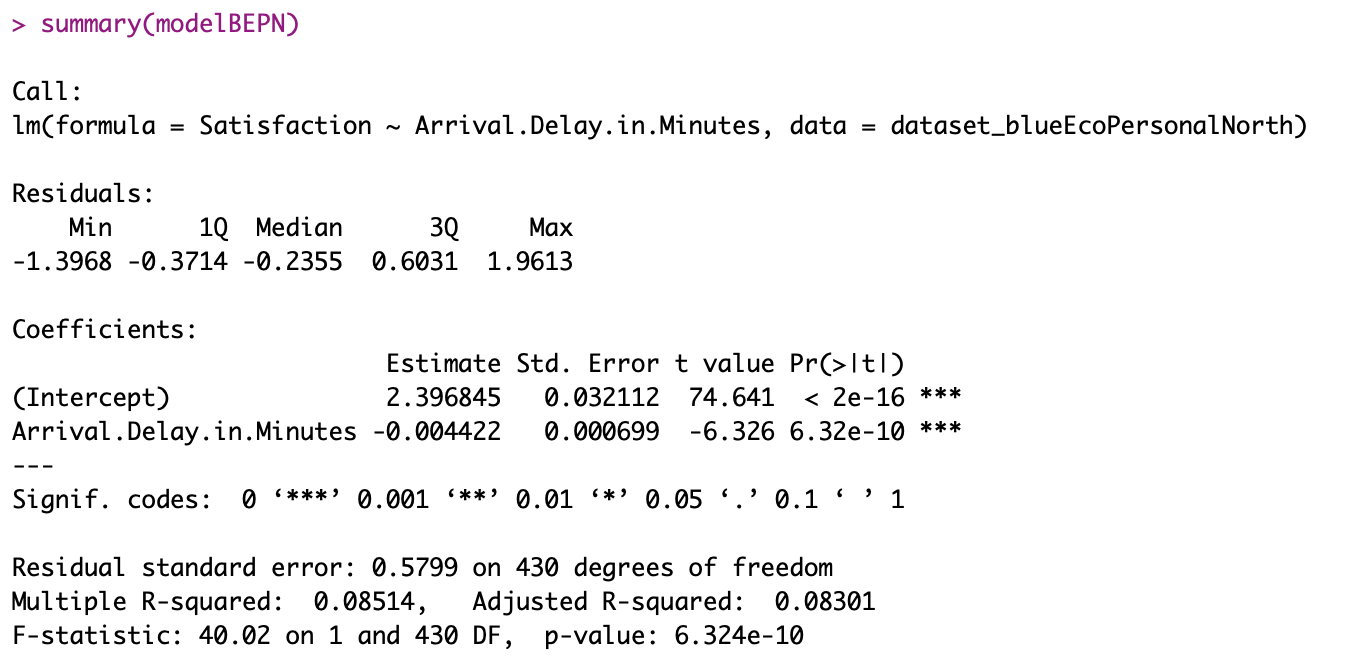
Based on the initial descriptive analysis the team, subset the data into customers belonging to the blue status who travel on economic class for personal travel. Blue status held the majority of the customers in this dataset, therefore it was the direction that we decided to pursue when it came to improving satisfaction. Eco class also held the majority of the customers in this dataset. This drove us to subset our data and focus on the majority group in order to improve customer satisfaction. Therefore, the subset we decided to focus on was Blue status Eco class customers who had Personal type of travel.

# Data Modeling - I

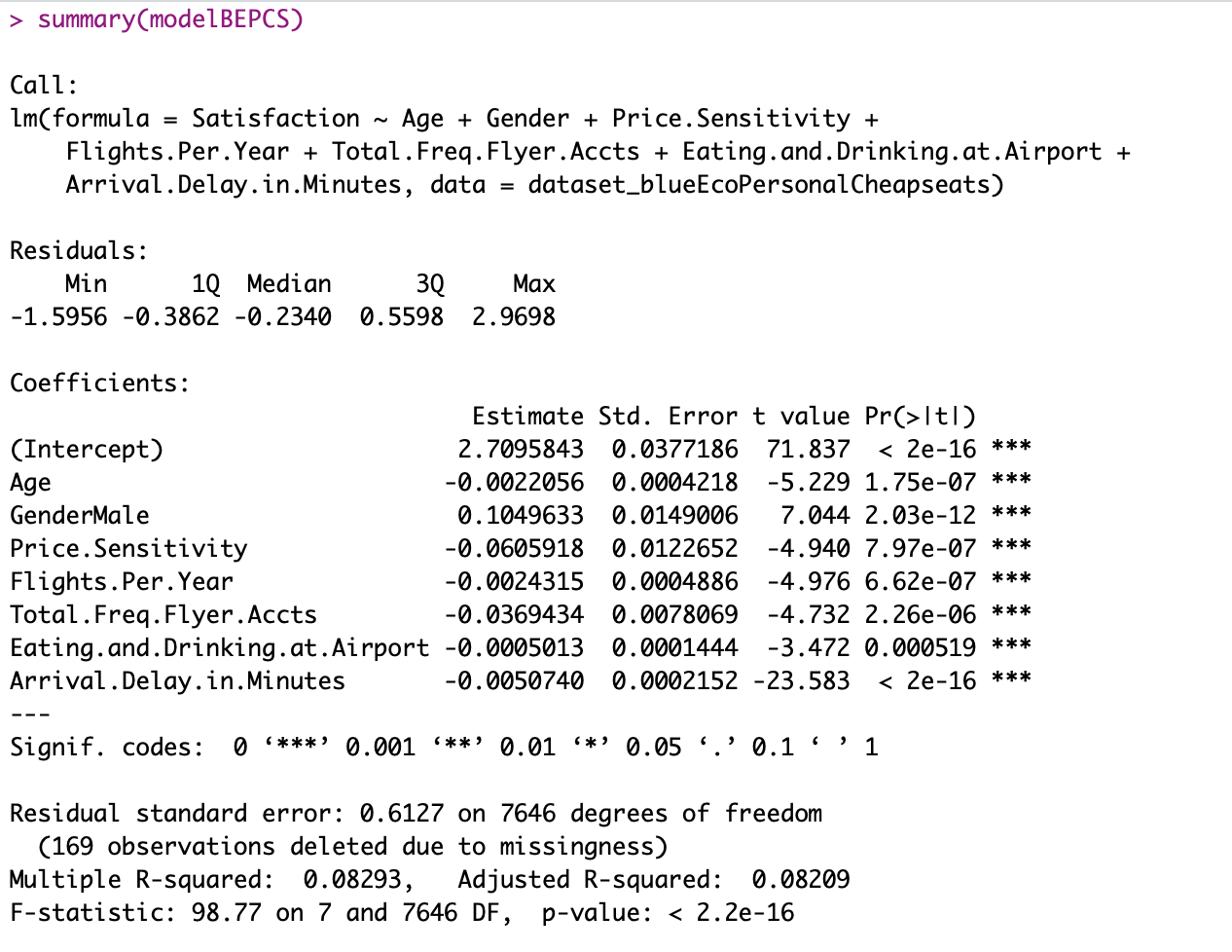
## Linear Model



As seen above, the first thing that we notice in the model is the poor R-Squared value. This model gives an R-Squared of just 6% which is very poor. This led us to focus more on a particular airline, to explore whether the R-Squared increased or not. Therefore, since Going North Airlines had the least satisfaction rate, we created a subset of the above data which had customers having partner airlines only Going North. The following were the results of the model.



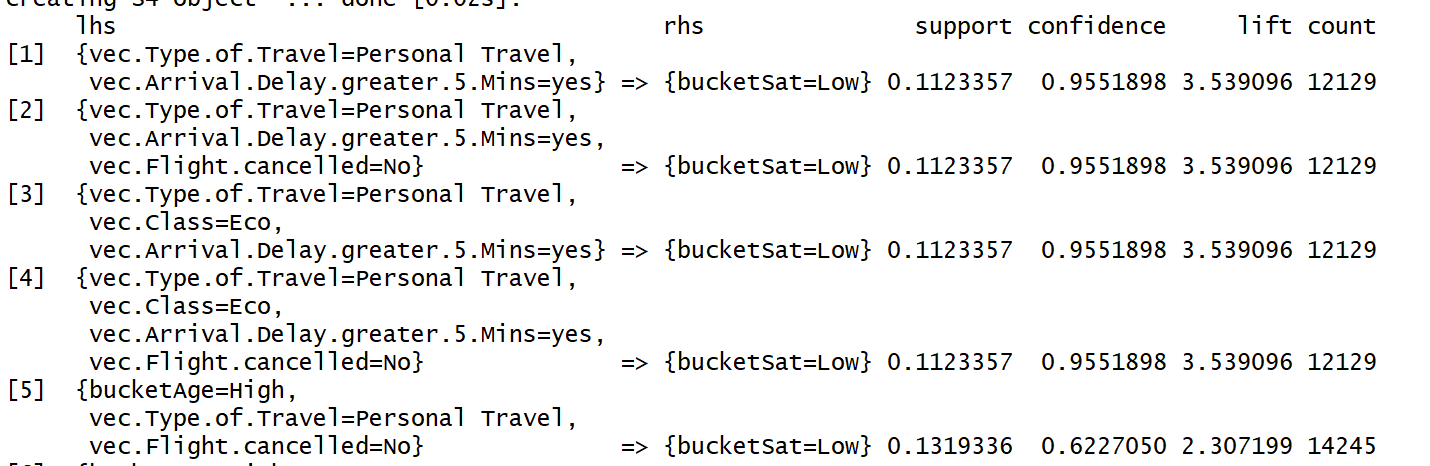
The first thing that we notice in this model is that we have a slightly improved R-Squared, but it is still poor. And also, the only significant factor that we found was of Arrival Delay Minutes in Going North Airlines compared to other significant coefficient observed in the first model. This led us to thinking, why not run the model on Cheapseats airlines which has the most number of travelers.



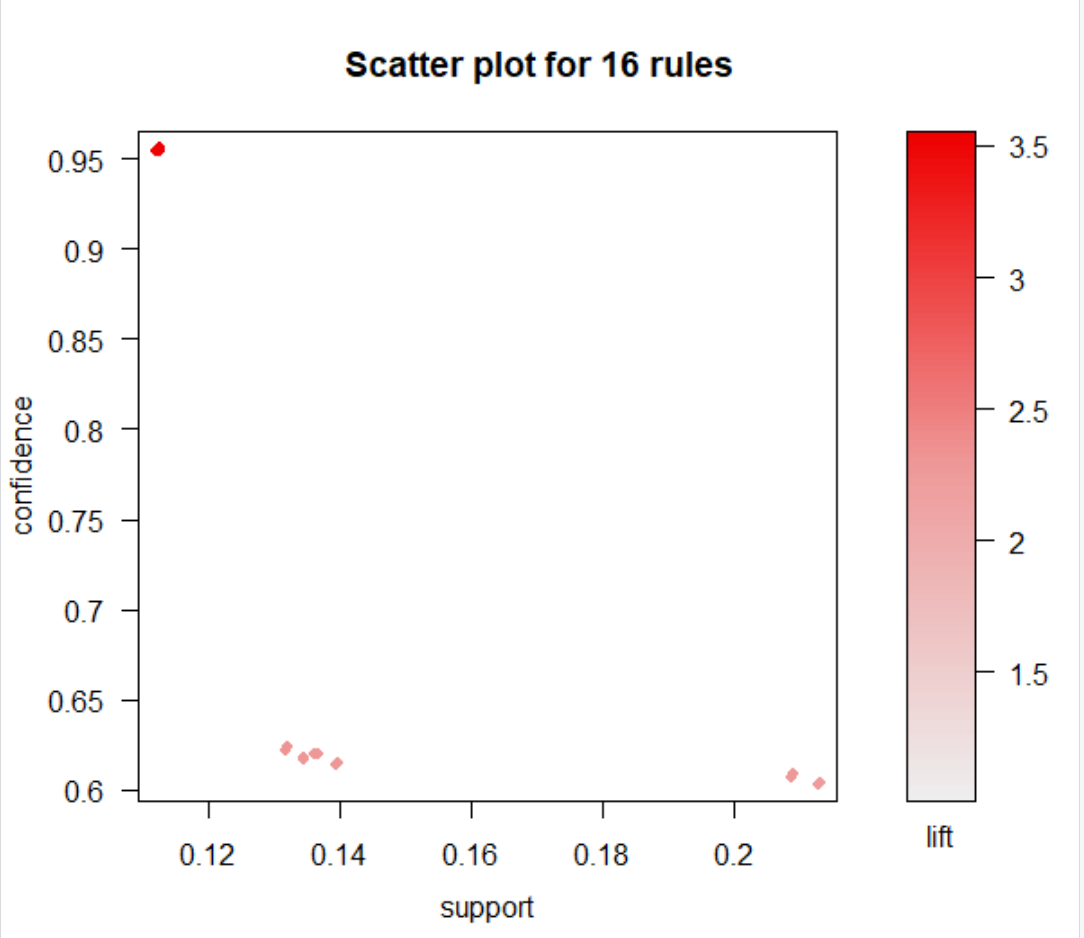
In the above model, it is evident that the model is not doing well since it has an R-Squared of just 8%. This led us to thinking that do we need to back track our dataset?

## Association Rules

We also implemented association rules to the subset of Eco Class customers with airline status of Blue and pick up the first 5 rules having highest confidence.



With rules above, we have about 95.5% of confidence to say that low satisfaction happens in association with personal traveller and arrival delay greater than 5 mins.

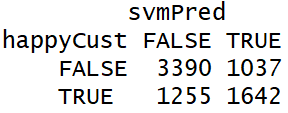


## Support Vector Machine

SVMs are built to try to make prediction on customer satisfaction based on different attributes. For this analysis we define happy customer as a customer with satisfaction greater than or equal to 3.

The first SVM: happyCust ~ Age + Gender + Price Sensitivity

Try to predict customer satisfaction base on age, gender, price sensitivity.

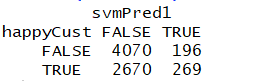


Error Rate: 31.3%

The error rate is acceptable because customer satisfaction is kind of human behavior. Hence, we can conclude that age, gender and price sensitivity are important to blue status customers who travel in eco class.

The second SVM: happyCust ~ Departure Delay in Minutes + Arrival Delay in Minutes

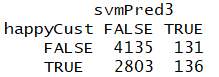
Try to predict customer satisfaction base on departure delay in minutes and arrival delay in minutes



Error Rate: 39.7%

The third SVM: happyCust ~ Flight.time.in.minutes + Flight.Distance

Try to predict customer satisfaction base on flight time in minutes and flight distance.



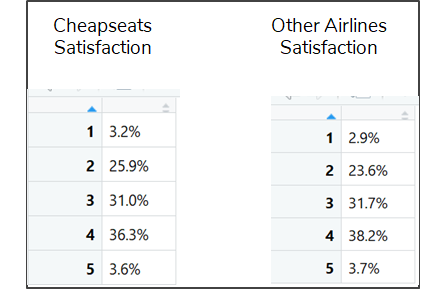
Error Rate: 40.7%

However, the error rate is high for second and third SVM.

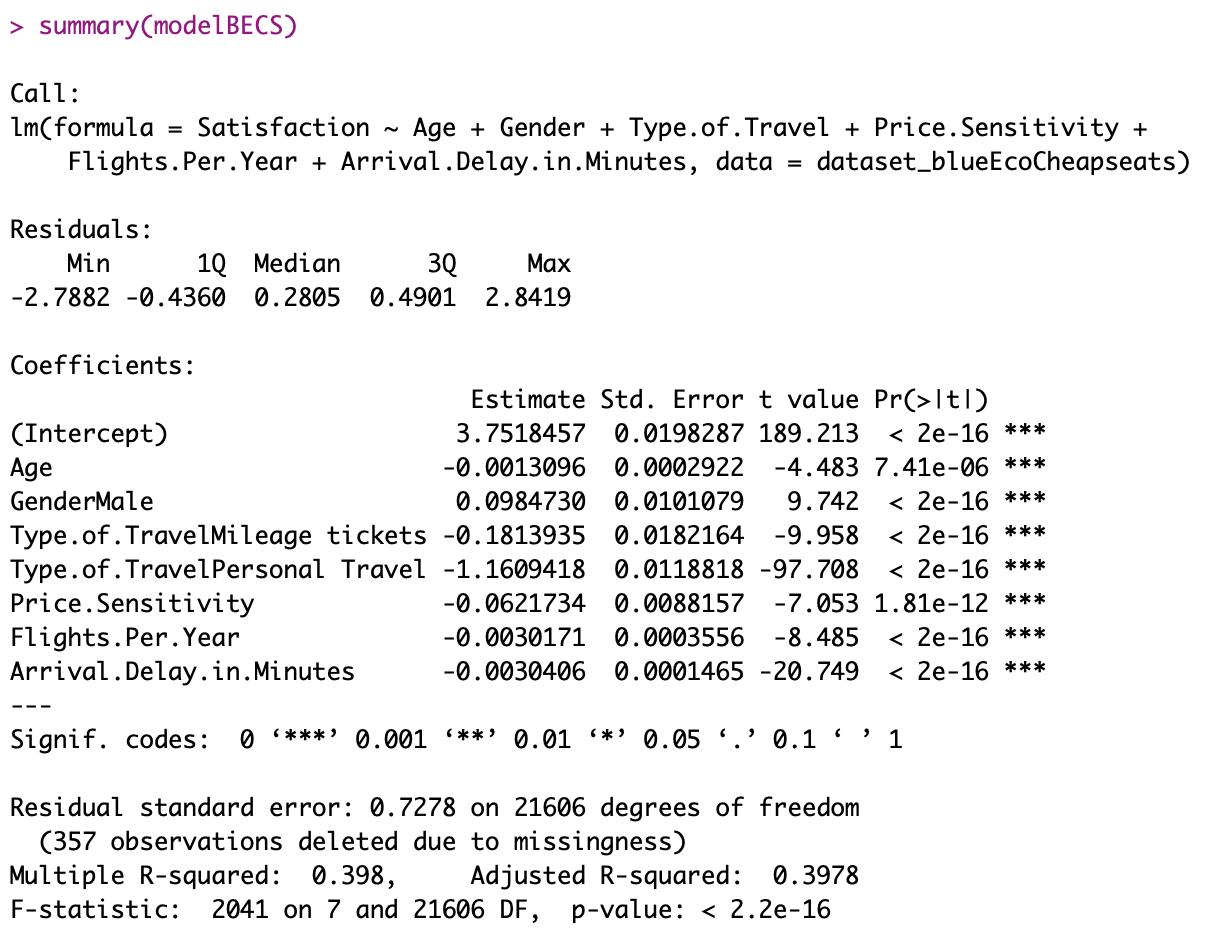
# 

# Data Subset - II

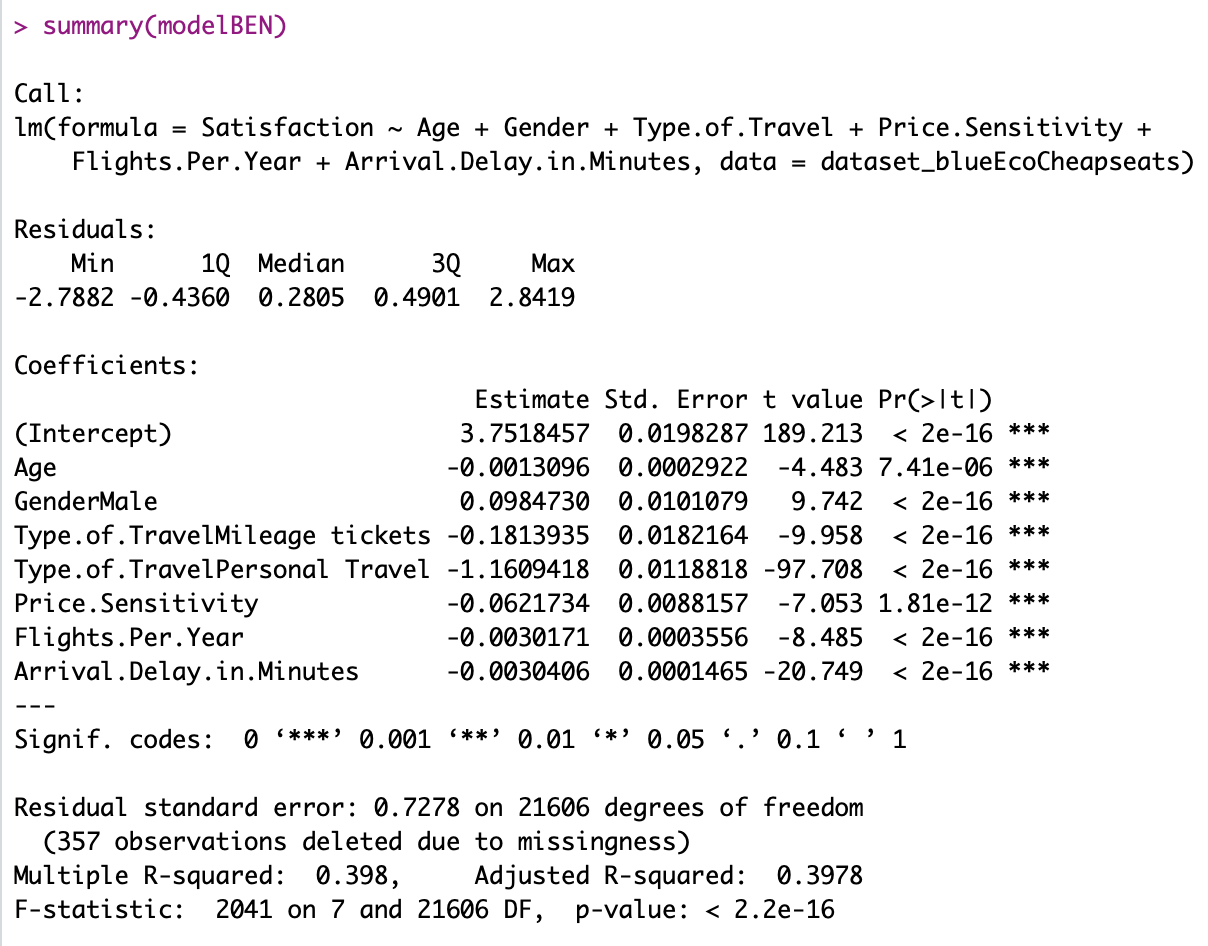
Further analysis on the first subset of data did not show significant trends and did not yield satisfactory results. Hence, the team subset the data further. The team looked at the satisfaction rates across the partner airlines. As noted earlier, Cheapseat Airline had the highest number of customers. Upon further analysis, it was found that Cheapseat Airline even had higher proportion of customers with satisfaction 3 and below as compared to the other partner airlines. Therefore, the team will focus on the Cheapseat Airline.



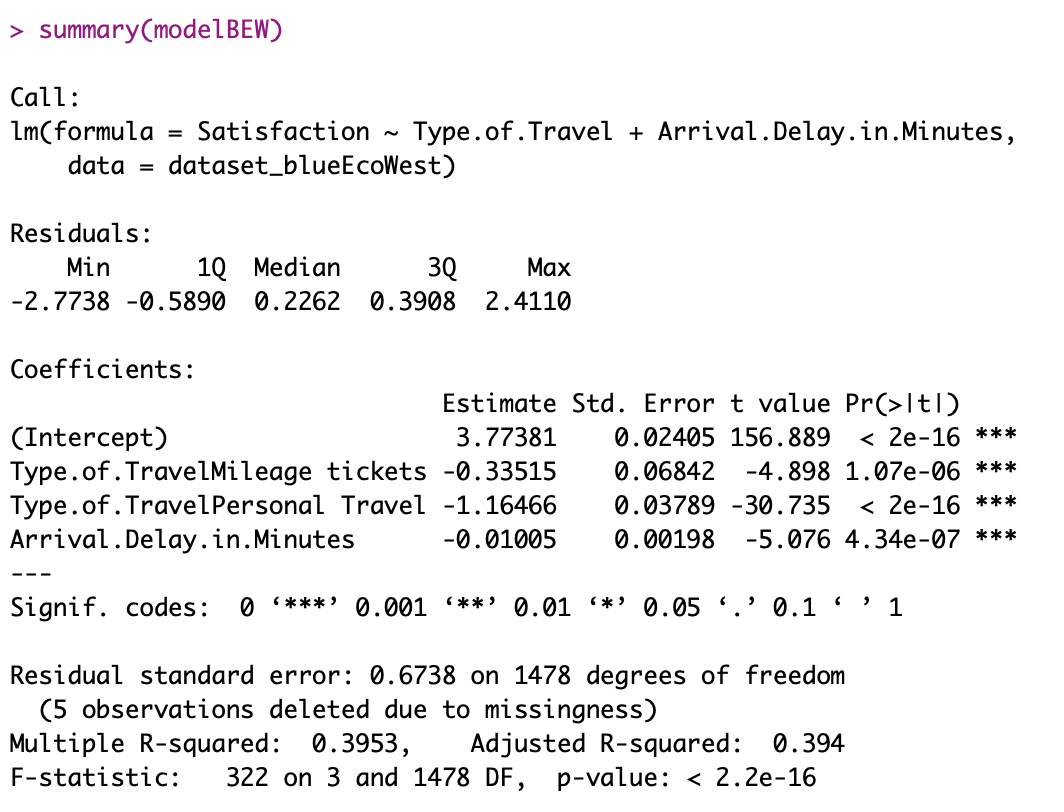
The data set for further analysis is customers belonging to the blue status who travel on economic class flying Cheapseats, Going North and West Airlines. The code for the condensed data is in Appendix 2. We chose these three airlines as Cheapseats had the most number of customers, Going North had the least satisfaction and West had the highest satisfaction.



The first thing that we observe in the above figure is the improved R-Squared value. Once we did not include type of travel, the performance of the model increased five folds. This is understandable as Type of Travel is a highly significant coefficient. We further decided to study the models for Going North Airlines.



In this model, we observe that the R-Squared is better than the previous Going North model. Another observation is that the number of significant factors has increased compared to just Arrival Delay in Minutes for the previous model. We also start observing a pattern in the coefficients which are highly significant, which could be a takeaway for Going North. Since these variables work for Cheapseats in increasing the satisfaction, it will also work for Going North if they concentrate on increasing the quality of service in these areas.



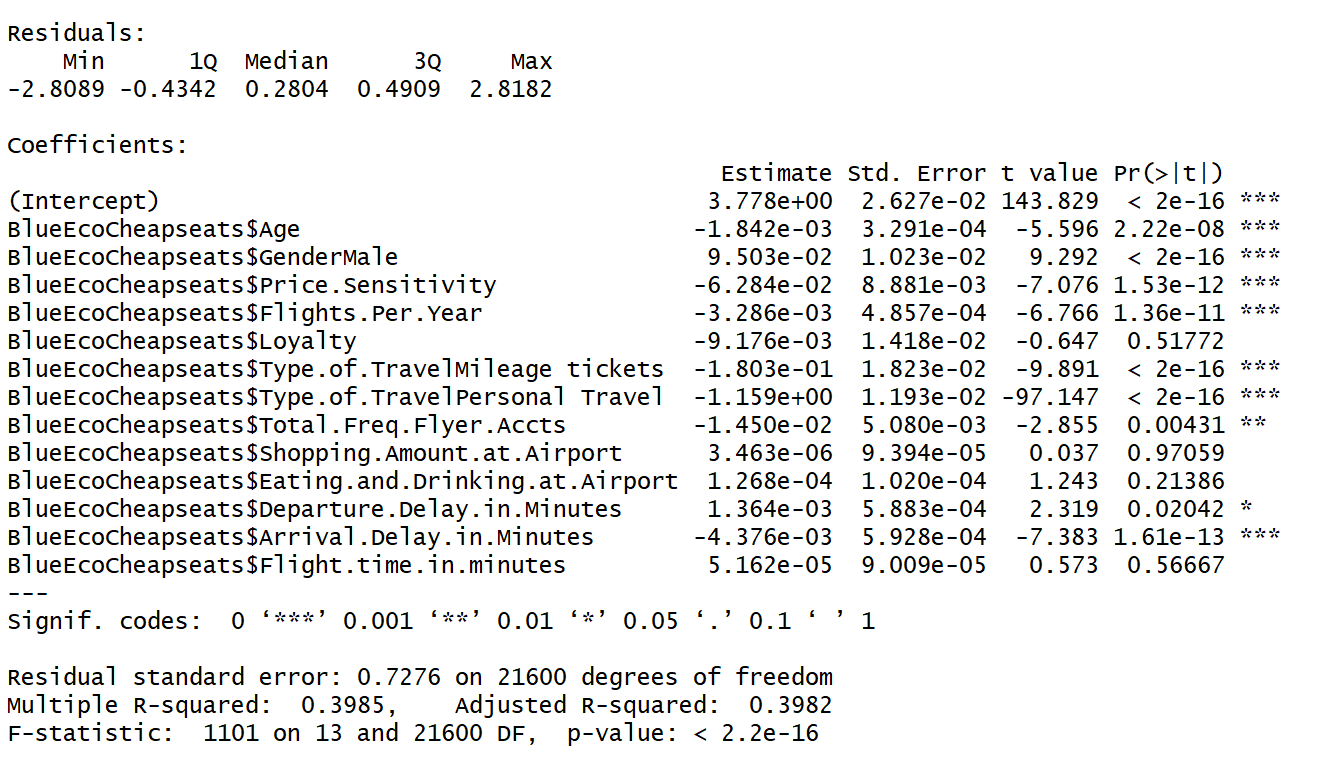
We made this model just for the sake of comparison. This model is only for the West Airlines data, which has the most satisfied customers. Here we observe that the highly significant variables are different than the ones that we observed in the previous models. For this airlines, to keep their customers satisfied, it should focus their efforts of Mileage, Personal tickets as well as minimizing the arrival delay in minutes as it affects the overall customer satisfaction.

# Data modeling - II

## Linear model

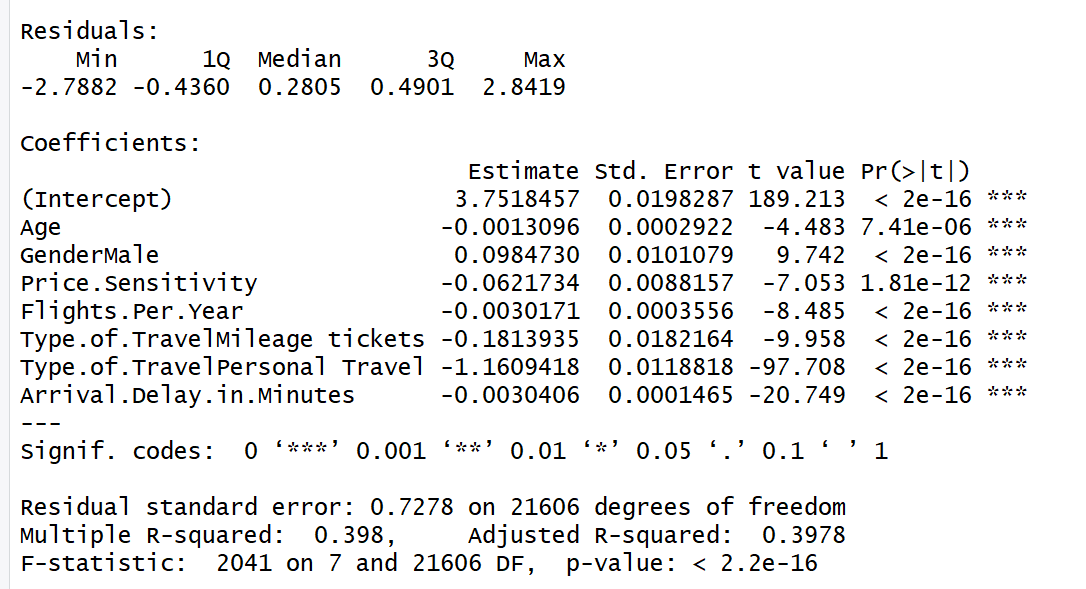
Different linear regressions are built after we focused to Eco customers from Cheapseats Airline Inc. with airline status of blue. We try to find out the attributes that will play important role in customer overall tendency and relationships among these attributes. In this part, we used the same data that narrows down to Eco customer from Cheapseats Airline Inc. with airline status of blue.

The first linear model we generated with attributes Age, Gender, Price Sensitivity, Flights per year, Loyalty, Type of travel, Total Freq Flyer Accts, Shopping Amount at Airport, Eating and Drinking at Airport, Departure Delay in Minutes, Arrival Delay in Minutes, Flight time in minutes.

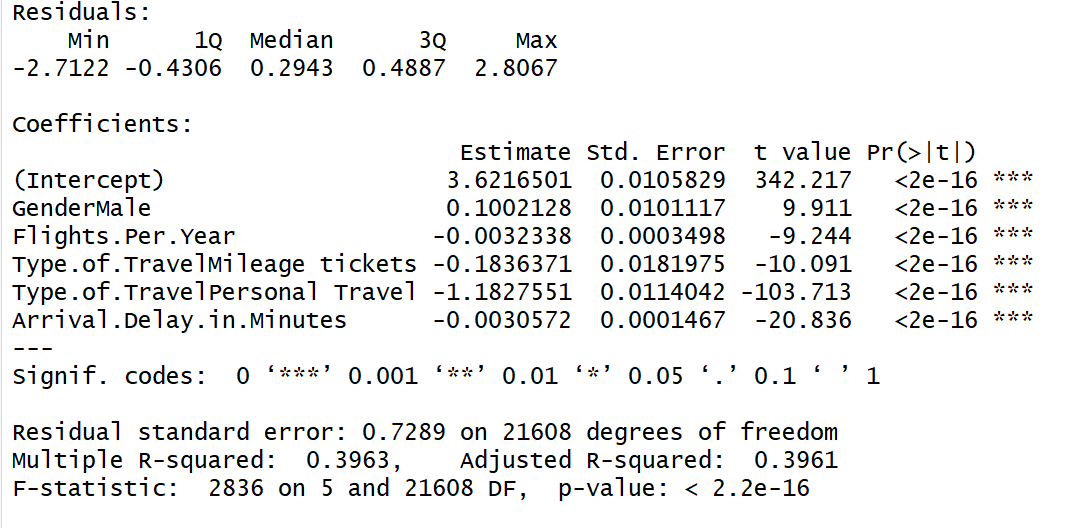


From the results we can see that R-squared is about 40% which means in 40% confidence we can make description in relationship between customer overall satisfaction and all the attributes cite above. We can also notice that loyalty, shopping amount at airport, eating and drinking at airport and flight time in minutes is not significant which means they may not be able to descript customer satisfaction in this linear model.

In next linear regression we will delete those insignificant attributes and try to find the key factors that can describe customer overall satisfaction in linear.

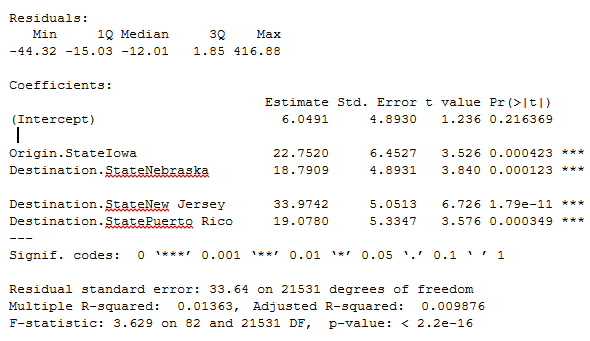


Although we can find that all attributes above are significant we still want to scale down the attribute. We will not include Age and Price Sensitivity in next regression because they have the higher possibility of being wrong because of higher P-values.



From this linear model we can see that gender do have impact on customer satisfaction that Male will have 0.1 higher satisfaction. Type of travel are also important that personal travel will have 1.18 lower satisfaction. And every 1 time increased in flights per year and every 1 minutes increased in arrival delay will cause 0.003 points decreasing in customer overall satisfaction.

In next linear model, we will try to find the relationship between arrival delay in minutes and states.



Although the R-squared is small but we can find significant state that related to arrival delay in minutes. They are Iowa as origin state and Nebraska, New Jersey and Puerto Rico as destination states.

So we recommend that to improve the situation of airport in states Nebraska, New Jersey and Puerto Rico to reduce the arrival delay in order to have better overall customer satisfaction.

Airlines departure from Iowa should improve the situation of arrival delay.

# Further Analysis

## Linear Model

A close up of a map

Description generated with high confidence

The mileage tickets have highest rate of decline and personal travels have the least overall satisfaction but the least rate of decline.

We recommend that Cheapseats should try to slightly reduce the mileage of tickets by changing mileage tickets policy. May offer more mileage tickets to customer having lower flights per year and reduce the number of mileage tickets to customer having high flights per year but offer better benefits.

A close up of a map

Description generated with high confidence

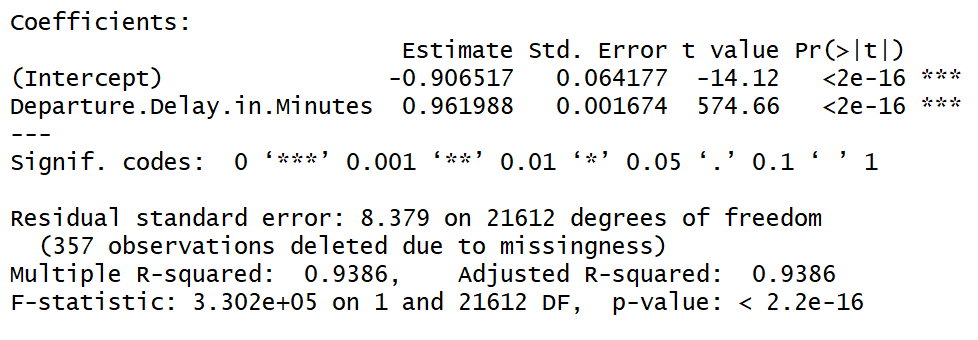
This plot shows the Linear Regression of Satisfaction vs. Arrival Delay in Minutes in Cheapseats Airline Blue customers in Eco Class.

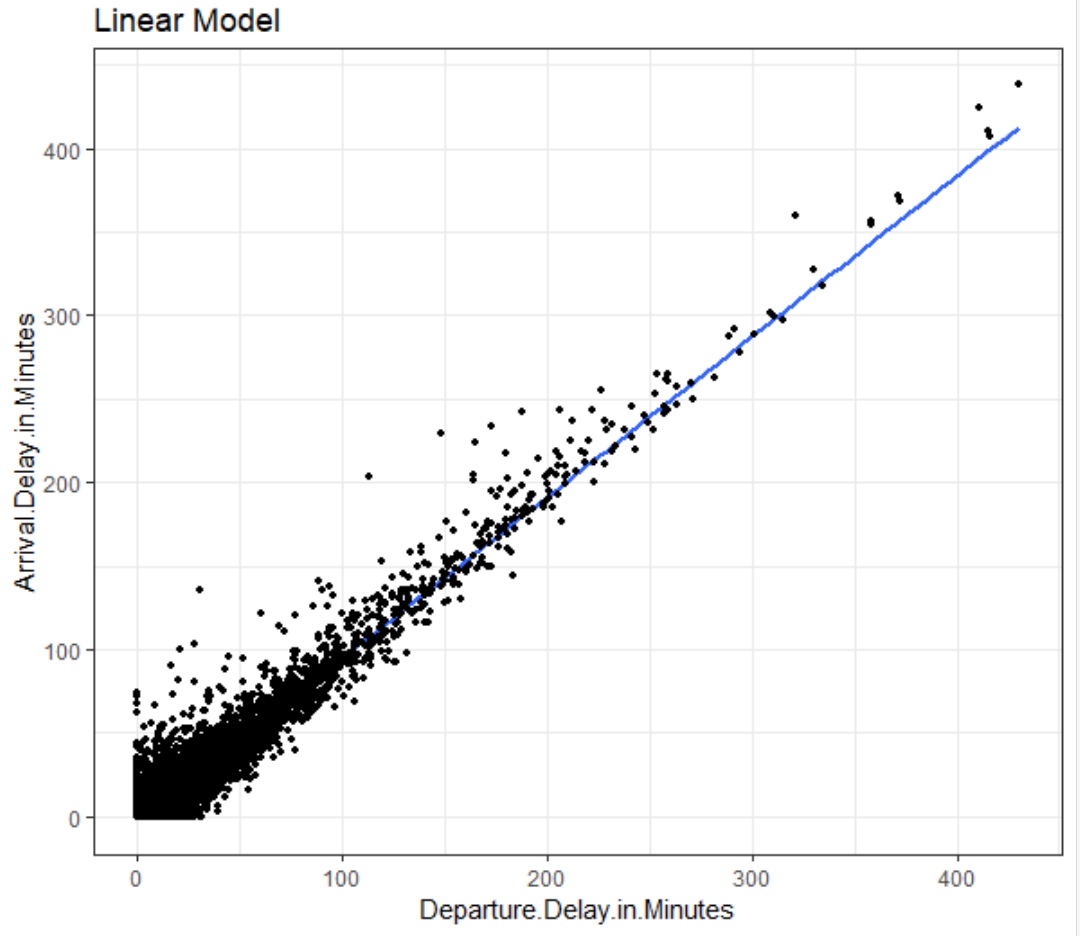
It shows clearly that in overall, satisfaction trends to decline when arrival delay in minutes increasing. Compared among types of travel we can see that personal travel will have more rate of satisfaction decline. It may be because the arrival delay will destroy their travel plan, for example, some of personal travels might have bought interline tickets.

We recommend that Cheapseats Airline should try to reduce time delayed when arrival for Cheapseats have an overall higher mean Arrival Delay in Minutes. (16.04 for Cheapseats and 15.23 for all airlines).

Also, after arrival delay happened, try to make action to comfort personal travel associated with time delayed so that the rate of reduce in satisfaction influenced by arrival delay would be reduced.

Linear model of Arrival Delay in Minutes vs. Departure Daly in Minutes





A close up of a map

Description generated with high confidence

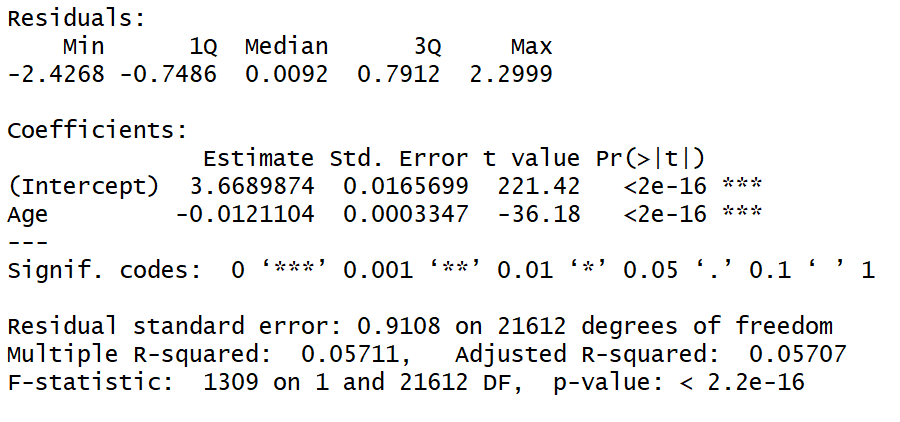
We can see from the plot that there are slightly different between satisfaction vs. flight per year grouped by gender. Male customers have over all higher satisfaction.

A close up of a map

Description generated with very high confidence

Male customers have overall higher satisfaction than female

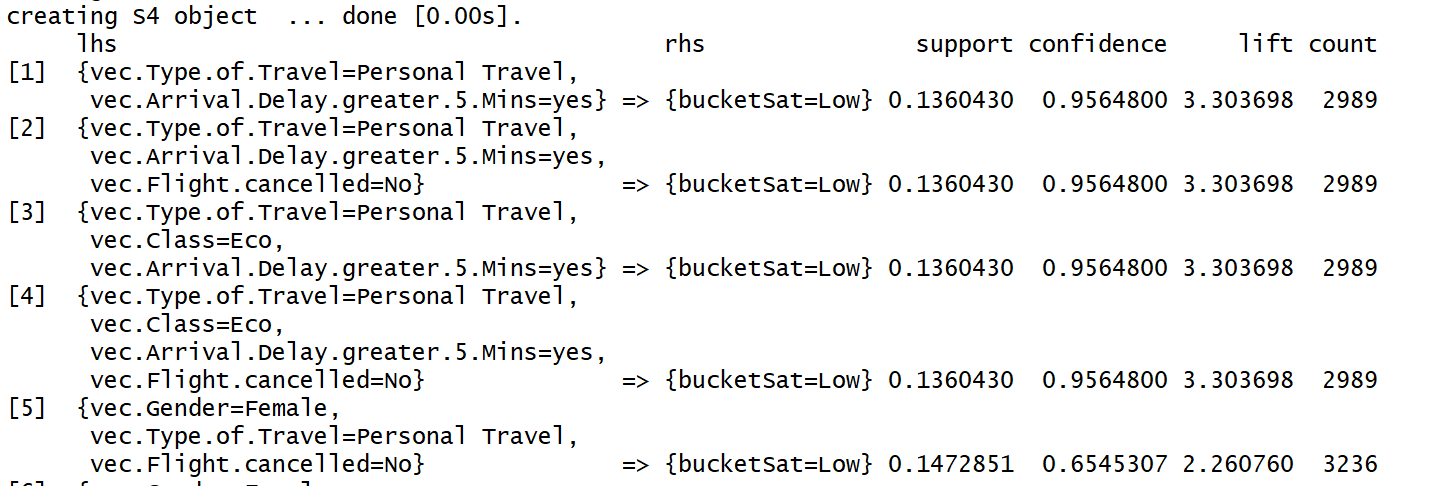
We built linear regression in customer overall satisfaction vs. age.



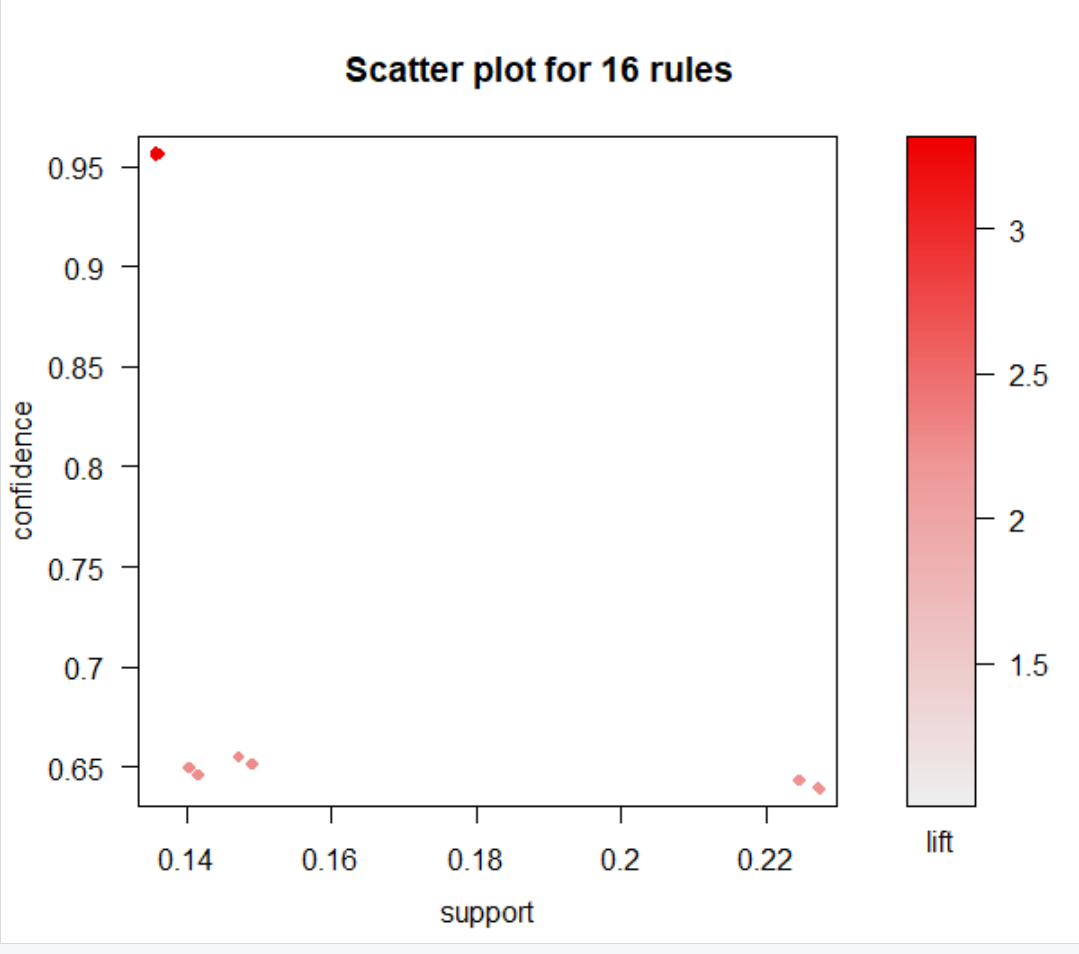
In the linear model we can see that with age goes up, the overall customer satisfaction trends to go down.

## Association Rules

We implemented the association rule to Cheapseats Airline Eco Class customer with airline status of Blue and gathered top 5 rules having highest confidence.



With rules above, we will have about 95.6% of confidence to say that low satisfaction happens in association with personal traveller and arrival delay greater than 5 mins.



# 

# Actionable Insights

Through our analysis above, we have gained some valuable insights that the executives of the partner airlines can look into and take actions to increase the satisfaction levels of their customers. The team will base its recommendations on these actionable insights. Some of our main takeaways are as follows:

1. The analysis for Cheapseats, Going North and West Airways indicated that higher the arrival delay, lower the customer satisfaction. Therefore, arrival delay is a key determinant of customer satisfaction.
2. Another factor that affects customer satisfaction is the number of flights per year. Customer satisfaction also falls with increase in the number of flights customers take per year.
3. In general, customers who were personal travelers seemed to be least satisfied when compared with the business and mileage travelers.
4. Moreover, our analysis indicated that most older people (>60 years) tend to be personal travelers.
5. Female customers seem to be less satisfied than male customers and their satisfaction rates fall faster than males across the two most significant factors that affect customer satisfaction.

# 

# Recommendations

Based on our actionable insights from our analysis, the team would like to present the following recommendations to the management teams at the various partner airlines.

1. For arrival delays, the airlines can train their pilots and crew on how to reduce the time in air to arrive on time. The airlines should further look into the routes from Iowa to New Jersey, Nebraska and Puerto Rico. These routes had the highest arrival delays. The airlines could also provide customers who have faced delays due to airline fault some compensation either in the form of a free meal or a shopping coupon.
2. The airlines should provide consistent service throughout the year, record customer data and provide frequent fliers some sort of incentives such as selecting seat for free or more loyalty points since the customer satisfaction decreases with increase in number of flights per year.
3. Since personal travelers were the least satisfied, the airlines could provide some offers or incentives to them. The airlines should aim to provide these customers with the mileage ticket options or upgrade them.
4. Most of the customers belonging to the personal travel category are older people who are greater than 60 years. The airlines company should not only offer them senior citizen discounts, but provide them comfort during their journey through comfortable and accessible seats. Further, these seats should be assigned depending on the duration of the flight and size of the airplane. The airline companies can also train their staff to be friendly and patient with them and provide assistance. They can also provide them complimentary meal on the flight.
5. Our analysis indicated that overall females are less satisfied than men on an average. However, the data we have is not sufficient to do further analysis as this would require additional psychographic and behavioural data. For this purpose, the airlines could conduct a survey with its female customers to analyze this issue.

Our overall recommendation is that the airlines need to reassess their services offered to all customers, improve them and maintain the quality of the services offered to all customers. This is crucial for the airlines to consider and implement since the customer satisfaction falls with increase in number of flights per year. This is a major concern as this could result in losing some loyal and high satisfaction customers as well. The airlines companies should take constant surveys from their customers after every flight and collect more data, perhaps qualitative data, to understand their behavior and improve its services accordingly. Therefore, continuous feedback and analysis is important if the airlines companies want their customers to be satisfied.

# 

# Appendix 1

**Attributes Name:**

1. **Satisfaction** – it is rated from 1 to 5, that how satisfied is the customer? 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. **Flights Per Year** – The number of flights that each customer has taken in the most recent 12 months. The range starting from 0 to 100.

8. **Loyalty** – An index of loyalty ranging from -1 to 1 that reflects the proportion of flights taken on other airlines versus flights taken on this airline. A higher index means more loyalty.

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. **Total Frequent Flyer Accounts** – How many frequent flyer accounts the customer has.

11. **Shopping Amount at Airport** – The spending in dollars on non-food/drink goods and services at the airport(s) where the customer was before, between, or after flights.

12.  **Eating and Drinking at Airport** – The spending in dollars on food/drink goods and services at the airport(s) where the customer was before, between, or after flights.

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 customer. 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. **Partner Code** – This airline works with wholly- and partially-owned subsidiary companies to deliver regional flights. For example, AA, AS, B6, and DL.

17. **Partner Name** – These are the full names of the subsidiary airline companies. Pseudonyms have been substituted in place of the real names.

18. Origin City – refers to actual city that customers have departed from. For example, Yuma AZ, Waco TX, and Toledo OH.

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.

29. **Long Duration Trip** – A Boolean variable that divides flight segments into two types: FALSE means a shorter duration segment (including average delays), TRUE means a longer duration segment**.**

# Appendix 2 (Code)

### Descriptive Analysis

#Reading the data

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

View(customersatisfaction)

#Structure of data

str(customersatisfaction)

#Subset of data looking at airlines status and partners

airlinestatusdata <- customersatisfaction [,c(1:4,8,9,10,13,16,17)]

View(airlinestatusdata)

airlinestatusdata <- airlinestatusdata[complete.cases(airlinestatusdata),]

#Airline Status and Satisfaction

typeofairlinestatus <- table(data.frame(airlinestatusdata$Satisfaction,airlinestatusdata$Airline.Status))

View(typeofairlinestatus)

tapply(airlinestatusdata$Satisfaction,airlinestatusdata$Airline.Status,mean)

meanofairlinestatus <- data.frame(tapply(airlinestatusdata$Satisfaction,airlinestatusdata$Airline.Status,mean))

View(meanofairlinestatus)

colnames(meanofairlinestatus) <- c("SatisfactionMean")

#Airline Status and Loyalty

typeofairlinestatusl <- table(data.frame(airlinestatusdata$Loyalty,airlinestatusdata$Airline.Status))

View(typeofairlinestatusl)

tapply(airlinestatusdata$Loyalty,airlinestatusdata$Airline.Status,mean)

meanofairlinestatusl <- data.frame(tapply(airlinestatusdata$Loyalty,airlinestatusdata$Airline.Status,mean))

View(meanofairlinestatusl)

colnames(meanofairlinestatusl) <- c("LoyaltyMean")

#Airline Status and gender breakdown

airlinestatusandgender <- table(data.frame(airlinestatusdata$Gender, airlinestatusdata$Airline.Status))

View(airlinestatusandgender)

tapply(airlinestatusdata$Gender, airlinestatusdata$Airline.Status)

#No. of customers for each airline statys

airlinesstatuscount <- table(data.frame( airlinestatusdata$Airline.Status))

View(airlinesstatuscount)

#Partner Airlines and Satisfaction

partnerairlines <- table(data.frame(airlinestatusdata$Satisfaction,airlinestatusdata$Airline.Status))

View(partnerairlines)

tapply(airlinestatusdata$Satisfaction,airlinestatusdata$Partner.Name,mean)

meanofpartnerairlines <- data.frame(tapply(airlinestatusdata$Satisfaction,airlinestatusdata$Partner.Name,mean))

View(meanofpartnerairlines)

colnames(meanofpartnerairlines) <- c("SatisfactionMean")

#Partner Airlines and Loyalty

partnerairlinesl <- table(data.frame(airlinestatusdata$Loyalty,airlinestatusdata$Airline.Status))

View(partnerairlinels)

tapply(airlinestatusdata$Loyalty,airlinestatusdata$Partner.Name,mean)

meanofpartnerairlinesl <- data.frame(tapply(airlinestatusdata$Loyalty,airlinestatusdata$Partner.Name,mean))

View(meanofpartnerairlinesl)

colnames(meanofpartnerairlinesl) <- c("LoyaltyMean")

#Partner airlines and gender breakdown

partnerairlinesandgender <- table(data.frame(airlinestatusdata$Gender, airlinestatusdata$Partner.Name))

View(partnerairlinesandgender)

tapply(airlinestatusdata$Gender, airlinestatusdata$Airline.Status)

#No. of customers for each airline statys

partnerairlinescount <- table(data.frame( airlinestatusdata$Partner.Name))

View(partnerairlinescount)

#Airline status and class

airlinestatusandclass <- table(data.frame(airlinestatusdata$Class,airlinestatusdata$Airline.Status))

View(airlinestatusandclass)

#Airline status and type of travel

airlinestatusandtype <- table(data.frame(airlinestatusdata$Type.of.Travel,airlinestatusdata$Airline.Status))

View(airlinestatusandtype)

#selecting on blue customers

#https://www.statmethods.net/management/subset.html

bluestatus <- airlinestatusdata[which(airlinestatusdata$Airline.Status == 'Blue'),]

View(bluestatus)

#blue customers class

bluecustomerclasscount <- table(data.frame( bluestatus$Class))

View(bluecustomerclasscount)

#blue customers class satisfaction

tapply(bluestatus$Satisfaction,bluestatus$Class,mean)

bluemeanclass <- data.frame(tapply(bluestatus$Satisfaction,bluestatus$Class,mean))

View(bluemeanclass)

colnames(bluemeanclass) <- c("SatisfactionMean")

#blue customers class loyalty

tapply(bluestatus$Loyalty,bluestatus$Class,mean)

bluemeanclassl <- data.frame(tapply(bluestatus$Loyalty,bluestatus$Class,mean))

View(bluemeanclassl)

colnames(bluemeanclassl) <- c("LoyaltyMean")

#blue customers type of travel

bluecustomertypecount <- table(data.frame( bluestatus$Type.of.Travel))

View(bluecustomertypecount)

#blue customers type of travel satisfaction

tapply(bluestatus$Satisfaction,bluestatus$Type.of.Travel,mean)

bluemeantype <- data.frame(tapply(bluestatus$Satisfaction,bluestatus$Type.of.Travel,mean))

View(bluemeantype)

colnames(bluemeantype) <- c("SatisfactionMean")

#blue customers type of travel loyalty

tapply(bluestatus$Loyalty,bluestatus$Type.of.Travel,mean)

bluemeantypel <- data.frame(tapply(bluestatus$Loyalty,bluestatus$Type.of.Travel,mean))

View(bluemeantypel)

colnames(bluemeantypel) <- c("LoyaltyMean")

#blue status class and type of travel

blueclassandtype <- table(data.frame(bluestatus$Type.of.Travel, bluestatus$Class))

View(blueclassandtype)

tapply(bluestatus$Type.of.Travel, bluestatus$Class)

#Travel Type and Class Data subset

traveltypeandclassdata <- customersatisfaction[,c(1,3,4,8,9,13)]

View(traveltypeandclassdata)

traveltypeandclassdata <- traveltypeandclassdata[complete.cases(traveltypeandclassdata),]

#Type of travel and customer satisfaction

typeoftravel <- table(data.frame(traveltypeandclassdata$Satisfaction,traveltypeandclassdata$Type.of.Travel))

View(typeoftravel)

tapply(traveltypeandclassdata$Satisfaction,traveltypeandclassdata$Type.of.Travel,mean)

meanoftypeoftravel <- data.frame(tapply(traveltypeandclassdata$Satisfaction,traveltypeandclassdata$Type.of.Travel,mean))

View(meanoftypeoftravel)

colnames(meanoftypeoftravel) <- c("SatisfactionMean")

#Type of travel and loyalty

typeoftravell <- table(data.frame(traveltypeandclassdata$Loyalty,traveltypeandclassdata$Type.of.Travel))

View(typeoftravell)

tapply(traveltypeandclassdata$Loyalty,traveltypeandclassdata$Type.of.Travel,mean)

meanoftypeoftravell <- data.frame(tapply(traveltypeandclassdata$Loyalty,traveltypeandclassdata$Type.of.Travel,mean))

View(meanoftypeoftravell)

colnames(meanoftypeoftravell) <- c("LoyaltyMean")

#Customer class and satisfaction

customerclass <- table(data.frame(traveltypeandclassdata$Satisfaction, traveltypeandclassdata$Class))

View(customerclass)

tapply(traveltypeandclassdata$Satisfaction,traveltypeandclassdata$Class,mean)

meanofcustomerclass <- data.frame(tapply(traveltypeandclassdata$Satisfaction,traveltypeandclassdata$Class,mean))

View(meanofcustomerclass)

colnames(meanofcustomerclass) <- c("SatisfactionMean")

#Customer class and Loyalty

customerclassl <- table(data.frame(traveltypeandclassdata$Loyalty, traveltypeandclassdata$Class))

View(customerclassl)

tapply(traveltypeandclassdata$Loyalty,traveltypeandclassdata$Class,mean)

meanofcustomerclassl <- data.frame(tapply(traveltypeandclassdata$Loyalty,traveltypeandclassdata$Class,mean))

View(meanofcustomerclassl)

colnames(meanofcustomerclassl) <- c("LoyaltyMean")

#Customer class and gender breakdown

customerclassandgender <- table(data.frame(traveltypeandclassdata$Gender, traveltypeandclassdata$Class))

View(customerclassandgender)

tapply(traveltypeandclassdata$Gender,traveltypeandclassdata$Class)

#No. of customers in each class

customerclasscount <- table(data.frame( traveltypeandclassdata$Class))

View(customerclasscount)

#Customer type of travel and gender breakdown

customertraveltypeandgender <- table(data.frame(traveltypeandclassdata$Gender, traveltypeandclassdata$Type.of.Travel))

View(customertraveltypeandgender)

tapply(traveltypeandclassdata$Gender,traveltypeandclassdata$Type.of.Travel)

#No. of customers of each travel type

customertraveltypeandgendercount <- table(data.frame( traveltypeandclassdata$Type.of.Travel))

View(customertraveltypeandgendercount)

### Visualization Code

Plot of Airline Status Filled with Satisfaction

install.packages("arules")

install.packages("arulesViz")

install.packages("scales")

library(arules)

library(arulesViz)

library(ggplot2)

library(gcookbook)

library(RColorBrewer)

library(dplyr)

library(scales)

#Clean NA and non-integer in satisfaction

GroupByStatus<-group\_by(survey, survey$Airline.Status)

#Clean NA and non-integer in satisfaction

GroupByStatus<-GroupByStatus[-which(GroupByStatus$Satisfaction==2.5),]

GroupByStatus<-GroupByStatus[-which(GroupByStatus$Satisfaction==3.5),]

GroupByStatus<-GroupByStatus[-which(GroupByStatus$Satisfaction==4.5),]

GroupByStatus<-na.omit(GroupByStatus)

View(GroupByStatus)

mytheme <- theme(legend.position = 'right',

panel.background = element\_rect(fill = "white")

)

#bar plot of airline status

AirlineStatus<-GroupByStatus

AirlineStatus$Airline.Status<-factor(AirlineStatus$Airline.Status, levels = c("Platinum","Gold","Silver","Blue"))

ggplot(AirlineStatus,aes(x=Airline.Status)) + geom\_bar(aes(fill=factor(Satisfaction)))+

scale\_fill\_brewer(palette="Blues") + coord\_polar(theta = "y")+mytheme+ggtitle("Plot of Airline Status Filled with Satisfaction")

View(GroupByStatus)

**# Histogram of Age**

Hist\_Age <- ggplot(GroupByStatus, aes(x = Age,fill=factor(Satisfaction))) + geom\_histogram(binwidth = 10, color = "white") +

scale\_fill\_brewer(palette="Blues")+ mytheme #+ coord\_polar(theta = "x")

Hist\_Age <- Hist\_Age+ggtitle("Histogram of Age Filled with Satisfaction")

Hist\_Age

**# Histogram of Partner Name with Satisfaction**

PartnerName<-GroupByStatus

PartnerName$Partner.Name<-factor(PartnerName$Partner.Name,levels = c("Cheapseats Airlines Inc.","Sigma Airlines Inc.","FlyFast Airways Inc.","Northwest Business Airlines Inc.","Paul Smith Airlines Inc.","Oursin Airlines Inc.","Southeast Airlines Co.","EnjoyFlying Air Services","OnlyJets Airlines Inc.","FlyToSun Airlines Inc.","FlyHere Airways","West Airways Inc.","GoingNorth Airlines Inc.","Cool&Young Airlines Inc."))

ggplot(PartnerName, aes(x = Partner.Name,fill=factor(Satisfaction))) + geom\_bar( color = "white") +

scale\_fill\_brewer(palette="Blues") + mytheme + theme(axis.text.x = element\_text(angle = 45,hjust = 1))+ggtitle ("Histogram of Airline Companies Filled with Satisfaction")#+ coord\_polar(theta = "y")

**# Pie Chart of Gender**

T1<-data.frame(table(BlueEco$Gender))

T1[1,2]

T1[2,2]

length(which(BlueEco$Gender == "Female"))

length(which(BlueEco$Gender == "Male"))

prop <- c(T1[1,2],T1[2,2])

prop<-prop/sum(prop)

prop

gender <- c("male","female")

value<-c(prop[2],prop[1])

df<-data.frame(gender,value)

Pie\_gender <- ggplot(df,aes(x="", y=value, fill=gender)) +

geom\_col() +

geom\_text(aes(label = percent(value)), position = position\_stack(vjust = 0.5)) +

coord\_polar("y") +

theme\_void() +

labs(title = "Piechart of Gender",

fill = "Gender")+

ggtitle("Pie Chart of Gender")

Pie\_gender

**# Histogram of Gender with Satisfaction**

Gender<-GroupByStatus

Gender$Gender<-factor(Gender$Gender, levels = c("Male","Female"))

ggplot(Gender,aes(x=Gender)) + geom\_bar(aes(fill=factor(Satisfaction)))+

scale\_fill\_brewer(palette="Blues") + mytheme + ggtitle ("Histogram of Gender Filled with Satisfaction")#+ coord\_polar(theta = "x")

**# Histogram of Classes Filled with Satisfaction**

Class<-GroupByStatus

Class$Class<-factor(Class$Class, levels = c("Business","Eco Plus","Eco"))

ggplot(Class, aes(x = Class,fill=factor(Satisfaction))) + geom\_bar( color = "white") +

scale\_fill\_brewer(palette="Blues") + mytheme + coord\_polar(theta = "y")+ggtitle ("Plot of Class Filled with Satisfaction")

**# Plot of Price Sensitivity**

ggplot(GroupByStatus, aes(x = Price.Sensitivity,fill=factor(Satisfaction))) + geom\_histogram(binwidth = 1, color = "white") +

scale\_fill\_brewer(palette="Blues") + mytheme + ggtitle ("Histogram of Price Sensitivity Filled with Satisfaction")#+ coord\_polar(theta = "x")

**# Plot of Flights Per Year with Satisfaction**

ggplot(GroupByStatus, aes(x = Flights.Per.Year,fill=factor(Satisfaction))) + geom\_histogram(binwidth = 5, color = "white") +

scale\_fill\_brewer(palette="Blues") + mytheme + ggtitle ("Histogram of Flights Per Year Filled with Satisfaction")#+ coord\_polar(theta = "x")

**# Visualization for Origin and Destination States**

# Looking at States

tapply(dataset$Satisfaction, dataset$Origin.State, mean)

library(ggmap)

library(maps)

library(tidyverse)

us <- map\_data("state")

dataset\_blueEcoPersonal$Origin.StateL <- tolower(dataset\_blueEcoPersonal$Origin.State)

us$region

######### Origin States ############

ggplot(dataset\_blueEcoPersonal, aes(map\_id = Origin.StateL)) +

geom\_map(map = us, aes(fill = dataset\_blueEcoPersonal$Satisfaction)) +

expand\_limits(x = us$long, y = us$lat) + coord\_map() +

scale\_fill\_continuous(name = "Satisfaction", low = "white", high = "red") +

xlab("Longitude") + ylab("Latitude") + ggtitle("Origin States with Satisfaction Rates")

########## Destination States #######

dataset\_blueEcoPersonal$Destination.StateL <- tolower(dataset\_blueEcoPersonal$Destination.State)

ggplot(dataset\_blueEcoPersonal, aes(map\_id = Destination.StateL)) +

geom\_map(map = us, aes(fill = dataset\_blueEcoPersonal$Satisfaction)) +

expand\_limits(x = us$long, y = us$lat) + coord\_map() +

scale\_fill\_continuous(name = "Satisfaction", low = "white", high = "red") +

xlab("Longitude") + ylab("Latitude") + ggtitle("Destination States with Satisfaction Rates")

**# States vs. Mean Satisfactions**

# Plot the satisfaction according to States

ggplot(dataset\_StateS, aes(x = dataset\_StateS$States, fill = factor(dataset\_StateS$MeanSatisfaction))) +

geom\_bar(color = "white") +

theme(axis.text = element\_text(angle = 45, hjust = 1))

# Plot

library(ggplot2)

theme\_set(theme\_bw())

dataset\_StateS <- dataset\_StateS[order(-dataset\_StateS$MeanSatisfaction),]

rownames(dataset\_StateS) <- NULL

ggplot(dataset\_StateS, aes(x=dataset\_StateS$States, y=dataset\_StateS$MeanSatisfaction)) +

geom\_point(size=3) +

geom\_segment(aes(x=dataset\_StateS$States,

xend=dataset\_StateS$States,

y=0,

yend=dataset\_StateS$MeanSatisfaction)) +

labs(title="Lollipop Chart",

subtitle="States vs. Mean Satisfactions",

caption="source: Airline Dataset") +

theme(axis.text.x = element\_text(angle=65, vjust=0.6)) +

xlab("States") + ylab("Mean Satisfaction")

**# Pie Chart of Type of Travel**

T1<-data.frame(table(BlueEco$Type.of.Travel))

T1

T1[1,2]

T1[2,2]

T1[3,2]

prop <- c(T1[1,2],T1[2,2],T1[3,2])

prop<-prop/sum(prop)

prop

class <- c("Business Travel","Mileage Tickets","Personal Travel")

value<-c(prop[1],prop[2],prop[3])

df<-data.frame(class,value)

Pie\_gender <- ggplot(df,aes(x="", y=value, fill=class)) +

geom\_col() +

geom\_text(aes(label = percent(value)), position = position\_stack(vjust = 0.5)) +

coord\_polar("y") +

theme\_void() +

labs(title = "Piechart of Class",

fill = "Type of Travel")+

ggtitle("Pie Chart of Type of Travel")+scale\_fill\_brewer(palette="RdYlBu")

Pie\_gender

**# Type of Travel Filled with Satisfaction**

Class<-BlueEco

Class$Class<-factor(Class$Type.of.Travel, levels = c("Mileage tickets","Personal Travel","Business travel"))

ggplot(Class, aes(x = Class,fill=factor(Satisfaction))) + geom\_bar( color = "white") +

scale\_fill\_brewer(palette="Blues") + mytheme + coord\_polar(theta = "y")+ggtitle ("Plot of Type of Travel Filled with Satisfaction")

**# Airline Companies Filled with Satisfaction**

PartnerName<-GroupByStatus

PartnerName$Partner.Name<-factor(PartnerName$Partner.Name,levels = c("Cheapseats Airlines Inc.","Sigma Airlines Inc.","FlyFast Airways Inc.","Northwest Business Airlines Inc.","Paul Smith Airlines Inc.","Oursin Airlines Inc.","Southeast Airlines Co.","EnjoyFlying Air Services","OnlyJets Airlines Inc.","FlyToSun Airlines Inc.","FlyHere Airways","West Airways Inc.","GoingNorth Airlines Inc.","Cool&Young Airlines Inc."))

ggplot(PartnerName, aes(x = Partner.Name,fill=factor(Satisfaction))) + geom\_bar( color = "white") +

scale\_fill\_brewer(palette="Blues") + mytheme + theme(axis.text.x = element\_text(angle = 45,hjust = 1))+ggtitle ("Histogram of Airline Companies Filled with Satisfaction")#+ coord\_polar(theta = "y")

**# Spending Money at the Airport by Rank and Type of Travel**

# To discover the amount of money spent at airport between different type of travel

satisfactionrate <- function(vec){

q <- quantile(vec, c(2, 3), na.rm = TRUE)

agerate <- replicate(length(vec), "average")

agerate[vec <= q[1]] <- "low"

agerate[vec > q[2]] <- "high"

return(agerate)

}

satisfactionrate <- satisfactionrate(BlueEcoCheapseats$Satisfaction)

moneyspend <- BlueEcoCheapseats$Shopping.Amount.at.Airport+BlueEcoCheapseats$Eating.and.Drinking.at.Airport

shoppingBytravel <- data.frame(moneyspend,satisfactionrate,BlueEcoCheapseats$Type.of.Travel)

ggplot(shoppingBytravel,aes(x=satisfactionrate,y=moneyspend,fill=BlueEcoCheapseats.Type.of.Travel))+

geom\_boxplot()+scale\_fill\_brewer(palette="RdYlBu")+mytheme +scale\_x\_discrete(breaks=c("low","average","high"),

labels=c("low","average","high"))+

scale\_y\_continuous(breaks=c(50,250,450,650))+labs(title = "spending money at airport by rank and type of travel")

Shopping <- data.frame(moneyspend,satisfactionrate,BlueEcoCheapseats$Gender)

str(Shopping)

ggplot(Shopping,aes(x=satisfactionrate,y=moneyspend,fill=BlueEcoCheapseats.Gender))+

geom\_boxplot()+scale\_x\_discrete(breaks=c("low","average","high"),

labels=c("low","average","high"))+

scale\_y\_continuous(breaks=c(50,250,450,650))+mytheme+labs(title = "spending money at airport by satisfaction and gender")

### 

### Linear Modeling

# Code for Linear Modeling

lm4<-lm(data=BlueEcoCheapseats,BlueEcoCheapseats$Satisfaction~BlueEcoCheapseats$Age+BlueEcoCheapseats$Gender+BlueEcoCheapseats$Price.Sensitivity+BlueEcoCheapseats$Flights.Per.Year+BlueEcoCheapseats$Loyalty+BlueEcoCheapseats$Type.of.Travel+BlueEcoCheapseats$Total.Freq.Flyer.Accts+BlueEcoCheapseats$Shopping.Amount.at.Airport+BlueEcoCheapseats$Eating.and.Drinking.at.Airport+BlueEcoCheapseats$Departure.Delay.in.Minutes+BlueEcoCheapseats$Arrival.Delay.in.Minutes+BlueEcoCheapseats$Flight.time.in.minutes)

summary(lm4)

Residuals:

Min 1Q Median 3Q Max

-2.8089 -0.4342 0.2804 0.4909 2.8182

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.778e+00 2.627e-02 143.829 < 2e-16 \*\*\*

BlueEcoCheapseats$Age -1.842e-03 3.291e-04 -5.596 2.22e-08 \*\*\*

BlueEcoCheapseats$GenderMale 9.503e-02 1.023e-02 9.292 < 2e-16 \*\*\*

BlueEcoCheapseats$Price.Sensitivity -6.284e-02 8.881e-03 -7.076 1.53e-12 \*\*\*

BlueEcoCheapseats$Flights.Per.Year -3.286e-03 4.857e-04 -6.766 1.36e-11 \*\*\*

BlueEcoCheapseats$Loyalty -9.176e-03 1.418e-02 -0.647 0.51772

BlueEcoCheapseats$Type.of.TravelMileage tickets -1.803e-01 1.823e-02 -9.891 < 2e-16 \*\*\*

BlueEcoCheapseats$Type.of.TravelPersonal Travel -1.159e+00 1.193e-02 -97.147 < 2e-16 \*\*\*

BlueEcoCheapseats$Total.Freq.Flyer.Accts -1.450e-02 5.080e-03 -2.855 0.00431 \*\*

BlueEcoCheapseats$Shopping.Amount.at.Airport 3.463e-06 9.394e-05 0.037 0.97059

BlueEcoCheapseats$Eating.and.Drinking.at.Airport 1.268e-04 1.020e-04 1.243 0.21386

BlueEcoCheapseats$Departure.Delay.in.Minutes 1.364e-03 5.883e-04 2.319 0.02042 \*

BlueEcoCheapseats$Arrival.Delay.in.Minutes -4.376e-03 5.928e-04 -7.383 1.61e-13 \*\*\*

BlueEcoCheapseats$Flight.time.in.minutes 5.162e-05 9.009e-05 0.573 0.56667

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7276 on 21600 degrees of freedom

Multiple R-squared: 0.3985, Adjusted R-squared: 0.3982

F-statistic: 1101 on 13 and 21600 DF, p-value: < 2.2e-16

lm5<-lm(data=BlueEcoCheapseats,BlueEcoCheapseats$Satisfaction~BlueEcoCheapseats$Age+BlueEcoCheapseats$Gender+BlueEcoCheapseats$Price.Sensitivity+BlueEcoCheapseats$Flights.Per.Year+BlueEcoCheapseats$Type.of.Travel+BlueEcoCheapseats$Total.Freq.Flyer.Accts+BlueEcoCheapseats$Arrival.Delay.in.Minutes)

summary(lm5)

Residuals:

Min 1Q Median 3Q Max

-2.8112 -0.4355 0.2793 0.4931 2.8303

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.7916081 0.0235043 161.316 < 2e-16 \*\*\*

BlueEcoCheapseats$Age -0.0017672 0.0003262 -5.417 6.13e-08 \*\*\*

BlueEcoCheapseats$GenderMale 0.0962178 0.0101312 9.497 < 2e-16 \*\*\*

BlueEcoCheapseats$Price.Sensitivity -0.0632163 0.0088201 -7.167 7.89e-13 \*\*\*

BlueEcoCheapseats$Flights.Per.Year -0.0031692 0.0003588 -8.834 < 2e-16 \*\*\*

BlueEcoCheapseats$Type.of.TravelMileage tickets -0.1797310 0.0182203 -9.864 < 2e-16 \*\*\*

BlueEcoCheapseats$Type.of.TravelPersonal Travel -1.1582736 0.0119095 -97.256 < 2e-16 \*\*\*

BlueEcoCheapseats$Total.Freq.Flyer.Accts -0.0154124 0.0048943 -3.149 0.00164 \*\*

BlueEcoCheapseats$Arrival.Delay.in.Minutes -0.0030438 0.0001465 -20.775 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7277 on 21605 degrees of freedom

Multiple R-squared: 0.3983, Adjusted R-squared: 0.3981

F-statistic: 1788 on 8 and 21605 DF, p-value: < 2.2e-16

lm6<-lm(data=BlueEcoCheapseats,BlueEcoCheapseats$Satisfaction~BlueEcoCheapseats$Age+BlueEcoCheapseats$Gender+BlueEcoCheapseats$Price.Sensitivity+BlueEcoCheapseats$Flights.Per.Year+BlueEcoCheapseats$Type.of.Travel+BlueEcoCheapseats$Arrival.Delay.in.Minutes)

summary(lm6)

Residuals:

Min 1Q Median 3Q Max

-2.7882 -0.4360 0.2805 0.4901 2.8419

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.7518457 0.0198287 189.213 < 2e-16 \*\*\*

BlueEcoCheapseats$Age -0.0013096 0.0002922 -4.483 7.41e-06 \*\*\*

BlueEcoCheapseats$GenderMale 0.0984730 0.0101079 9.742 < 2e-16 \*\*\*

BlueEcoCheapseats$Price.Sensitivity -0.0621734 0.0088157 -7.053 1.81e-12 \*\*\*

BlueEcoCheapseats$Flights.Per.Year -0.0030171 0.0003556 -8.485 < 2e-16 \*\*\*

BlueEcoCheapseats$Type.of.TravelMileage tickets -0.1813935 0.0182164 -9.958 < 2e-16 \*\*\*

BlueEcoCheapseats$Type.of.TravelPersonal Travel -1.1609418 0.0118818 -97.708 < 2e-16 \*\*\*

BlueEcoCheapseats$Arrival.Delay.in.Minutes -0.0030406 0.0001465 -20.749 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7278 on 21606 degrees of freedom

Multiple R-squared: 0.398, Adjusted R-squared: 0.3978

F-statistic: 2041 on 7 and 21606 DF, p-value: < 2.2e-16

lm7<-lm(data=BlueEcoCheapseats,Satisfaction~Gender+Flights.Per.Year+Type.of.Travel+Arrival.Delay.in.Minutes)

summary(lm7) #Final Linear Model

Residuals:

Min 1Q Median 3Q Max

-2.7122 -0.4306 0.2943 0.4887 2.8067

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.6216501 0.0105829 342.217 <2e-16 \*\*\*

GenderMale 0.1002128 0.0101117 9.911 <2e-16 \*\*\*

Flights.Per.Year -0.0032338 0.0003498 -9.244 <2e-16 \*\*\*

Type.of.TravelMileage tickets -0.1836371 0.0181975 -10.091 <2e-16 \*\*\*

Type.of.TravelPersonal Travel -1.1827551 0.0114042 -103.713 <2e-16 \*\*\*

Arrival.Delay.in.Minutes -0.0030572 0.0001467 -20.836 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7289 on 21608 degrees of freedom

Multiple R-squared: 0.3963, Adjusted R-squared: 0.3961

F-statistic: 2836 on 5 and 21608 DF, p-value: < 2.2e-16

Code

# Code for Linear Modeling and Visualizations

lm1<-lm(data = BlueEco,BlueEco$Satisfaction~BlueEco$Age+BlueEco$Price.Sensitivity+BlueEco$Flights.Per.Year+BlueEco$Loyalty+BlueEco$Shopping.Amount.at.Airport+BlueEco$Eating.and.Drinking.at.Airport+BlueEco$Departure.Delay.in.Minutes+BlueEco$Arrival.Delay.in.Minutes+BlueEco$Flight.time.in.minutes)

summary(lm1)

lm2<-lm(data = BlueEco,BlueEco$Satisfaction~BlueEco$Age+BlueEco$Price.Sensitivity+BlueEco$Flights.Per.Year+BlueEco$Loyalty+BlueEco$Shopping.Amount.at.Airport+BlueEco$Eating.and.Drinking.at.Airport+BlueEco$Departure.Delay.in.Minutes+BlueEco$Arrival.Delay.in.Minutes)

summary(lm2)

lm3<-lm(data = BlueEco,BlueEco$Satisfaction~BlueEco$Age+BlueEco$Price.Sensitivity+BlueEco$Flights.Per.Year+BlueEco$Loyalty+BlueEco$Eating.and.Drinking.at.Airport+BlueEco$Departure.Delay.in.Minutes+BlueEco$Arrival.Delay.in.Minutes)

summary(lm3)

PN<-table(BlueEco$Partner.Name)

PN<-data.frame(PN)

PN[1,1]

BlueEcoCheapseats<-subset(BlueEco,BlueEco$Partner.Name==PN[1,1])

ggplot(BlueEcoCheapseats, aes(x=factor(1), fill=Gender))+

geom\_bar(width = 1)+mytheme+ coord\_polar(theta = "y")

ggplot(BlueEcoCheapseats, aes(x=factor(1), fill=Type.of.Travel))+

geom\_bar(width = 1)+mytheme+ coord\_polar(theta = "y")+scale\_fill\_brewer(palette = "RdYlBu")

CheapType<-BlueEcoCheapseats

ggplot(CheapType, aes(x = Type.of.Travel,fill=factor(Satisfaction))) + geom\_bar( color = "white") +

scale\_fill\_brewer(palette="Blues") + mytheme #+ coord\_polar(theta = "y")

lm4<-lm(data=BlueEcoCheapseats,BlueEcoCheapseats$Satisfaction~BlueEcoCheapseats$Age+

BlueEcoCheapseats$Age+

BlueEcoCheapseats$Gender+

BlueEcoCheapseats$Price.Sensitivity+

BlueEcoCheapseats$Flights.Per.Year+

BlueEcoCheapseats$Loyalty+

BlueEcoCheapseats$Type.of.Travel+

BlueEcoCheapseats$Total.Freq.Flyer.Accts+

BlueEcoCheapseats$Shopping.Amount.at.Airport+

BlueEcoCheapseats$Eating.and.Drinking.at.Airport+

BlueEcoCheapseats$Departure.Delay.in.Minutes+

BlueEcoCheapseats$Arrival.Delay.in.Minutes+

BlueEcoCheapseats$Flight.time.in.minutes)

summary(lm4)

lm5<-lm(data=BlueEcoCheapseats,BlueEcoCheapseats$Satisfaction~BlueEcoCheapseats$Age+BlueEcoCheapseats$Age+BlueEcoCheapseats$Gender+BlueEcoCheapseats$Price.Sensitivity+BlueEcoCheapseats$Flights.Per.Year+BlueEcoCheapseats$Type.of.Travel+BlueEcoCheapseats$Total.Freq.Flyer.Accts+BlueEcoCheapseats$Arrival.Delay.in.Minutes)

summary(lm5)

lm6<-lm(data=BlueEcoCheapseats,Satisfaction~Age+

Gender+

Price.Sensitivity+

Flights.Per.Year+Type.of.Travel+

Arrival.Delay.in.Minutes)

summary(lm6)

lm7<-lm(data=BlueEcoCheapseats,Satisfaction~Gender+

Flights.Per.Year+

Type.of.Travel+

Arrival.Delay.in.Minutes)

summary(lm7)

lm8<-lm(data=BlueEcoCheapseats,Satisfaction~Gender+

Flights.Per.Year+Type.of.Travel+

Arrival.Delay.in.Minutes+

Origin.State+

Destination.State)

summary(lm8)

lm9<-lm(data=BlueEcoCheapseats,Arrival.Delay.in.Minutes~Origin.State+

Destination.State)

summary(lm9)

#Origin.StateIowa 22.7520 6.4527 3.526 0.000423 \*\*\*

# Destination.StateCalifornia 10.0665 3.8300 2.628 0.008586 \*\*

# Destination.StateColorado 10.8690 3.9136 2.777 0.005487 \*\*

# Destination.StateNebraska 18.7909 4.8931 3.840 0.000123 \*\*\*

# Destination.StateNew Jersey 33.9742 5.0513 6.726 1.79e-11 \*\*\*

# Destination.StatePuerto Rico 19.0780 5.3347 3.576 0.000349 \*\*\*

mean(BlueEcoCheapseats$Arrival.Delay.in.Minutes)

mean(GroupByStatus$Arrival.Delay.in.Minutes)

#plot of price sensitivity

sensitive<-subset(BlueEcoCheapseats,BlueEcoCheapseats$Price.Sensitivity>3)

sensitive

ggplot(Personal, aes(x = Price.Sensitivity,fill=factor(Price.Sensitivity))) + geom\_histogram(binwidth = 1, color = "white") +

scale\_fill\_brewer(palette="Blues") + mytheme + ggtitle ("Histogram of Price Sensitivity Filled with Type of Travel")#+ coord\_polar(theta = "x")

#plot of flight per year

x <- BlueEcoCheapseats$Arrival.Delay.in.Minutes

y <- BlueEcoCheapseats$Satisfaction

group <- BlueEcoCheapseats$Gender

test<-data.frame(x,y,group)

ggplot(test, aes(x, y, shape=group, colour=group, fill=group)) +

geom\_smooth(method="lm") +

geom\_point(size=1) +

theme\_bw() +

xlab("Flights per Year") +

ylab("Satisfaction") +

ggtitle("Linear Model") +

#scale\_fill\_brewer(palette = "RdYlBu")+

expand\_limits(y=0) #+

# scale\_y\_continuous(breaks = 0:10) +

# scale\_x\_continuous(breaks = 1998:2015)

x <- BlueEcoCheapseats$Arrival.Delay.in.Minutes

y <- BlueEcoCheapseats$Satisfaction

group <- BlueEcoCheapseats$Type.of.Travel

test<-data.frame(x,y,group)

ggplot(test, aes(x, y, shape=group, colour=group, fill=group)) +

geom\_smooth(method="lm") +

geom\_point(size=1) +

theme\_bw() +

xlab("Arrival Delay in Minutes") +

ylab("Satisfaction") +

ggtitle("Linear Model") +

#scale\_fill\_brewer(palette = "RdYlBu")+

expand\_limits(y=0) #+

# scale\_y\_continuous(breaks = 0:10) +

# scale\_x\_continuous(breaks = 1998:2015)

x <- BlueEcoCheapseats$Flights.Per.Year

y <- BlueEcoCheapseats$Satisfaction

group <- BlueEcoCheapseats$Gender

test<-data.frame(x,y,group)

ggplot(test, aes(x, y, shape=group, colour=group, fill=group)) +

geom\_smooth(method="lm") +

geom\_point(size=1) +

theme\_bw() +

xlab("Flight per Year") +

ylab("Satisfaction") +

ggtitle("Linear Model") +

#scale\_fill\_brewer(palette = "RdYlBu")+

expand\_limits(y=0) #+

# scale\_y\_continuous(breaks = 0:10) +

# scale\_x\_continuous(breaks = 1998:2015)

x <- BlueEcoCheapseats$Arrival.Delay.in.Minutes

y <- BlueEcoCheapseats$Satisfaction

group <- BlueEcoCheapseats$Gender

test<-data.frame(x,y,group)

ggplot(test, aes(x, y, shape=group, colour=group, fill=group)) +

geom\_smooth(method="lm") +

geom\_point(size=1) +

theme\_bw() +

xlab("Arrival Delay in Minutes") +

ylab("Satisfaction") +

ggtitle("Linear Model") +

#scale\_fill\_brewer(palette = "RdYlBu")+

expand\_limits(y=0) #+

scale\_y\_continuous(breaks = 0:10) +

scale\_x\_continuous(breaks = 1998:2015)

lm<-lm(data=BlueEcoCheapseats,Satisfaction~Age)

summary(lm)

summary(BlueEcoCheapseats)

table(BlueEcoCheapseats$Type.of.Travel)

Personal<-subset(BlueEcoCheapseats,BlueEcoCheapseats$Type.of.Travel=="Personal Travel")

View(Personal)

str(Personal)

personal<-Personal[,c(-2,-9,-13,-16,-17)]

View(personal)

str(personal)

personal<-personal[,-25]

lm7<-lm(data=personal,Satisfaction~Age+Gender+Price.Sensitivity+Flights.Per.Year+Loyalty+Total.Freq.Flyer.Accts+Shopping.Amount.at.Airport+Eating.and.Drinking.at.Airport+Departure.Delay.in.Minutes+Arrival.Delay.in.Minutes+Flight.time.in.minutes)

summary(lm7)

lm8<-lm(data=personal,Satisfaction~Age+Gender+Price.Sensitivity+Flights.Per.Year+Total.Freq.Flyer.Accts+Shopping.Amount.at.Airport+Eating.and.Drinking.at.Airport+Arrival.Delay.in.Minutes)

summary(lm8)

display.brewer.all()

#To discover the amount of money spent at airport between different type of travel

satisfactionrate <- function(vec){

q <- quantile(vec, c(0.25, 0.75), na.rm = TRUE)

agerate <- replicate(length(vec), "average")

agerate[vec <= q[1]] <- "low"

agerate[vec > q[2]] <- "high"

return(agerate)

}

satisfactionrate <- satisfactionrate(BlueEcoCheapseats$Satisfaction)

moneyspend <- BlueEcoCheapseats$Shopping.Amount.at.Airport+BlueEcoCheapseats$Eating.and.Drinking.at.Airport

shoppingBytravel <- data.frame(moneyspend,satisfactionrate,BlueEcoCheapseats$Type.of.Travel)

ggplot(shoppingBytravel,aes(x=satisfactionrate,y=moneyspend,fill=BlueEcoCheapseats.Type.of.Travel))+

geom\_boxplot()+scale\_fill\_brewer(palette="RdYlBu")+mytheme +scale\_x\_discrete(breaks=c("low","average","high"),

labels=c("low","average","high"))+

scale\_y\_continuous(breaks=c(50,250,450,650))+labs(title = "spending money at airport by rank and type of travel")

Shopping <- data.frame(moneyspend,satisfactionrate,BlueEcoCheapseats$Gender)

str(Shopping)

ggplot(Shopping,aes(x=satisfactionrate,y=moneyspend,fill=BlueEcoCheapseats.Gender))+

geom\_boxplot()+scale\_x\_discrete(breaks=c("low","average","high"),

labels=c("low","average","high"))+

scale\_y\_continuous(breaks=c(50,250,450,650))+mytheme+labs(title = "spending money at airport by satisfaction and gender")

str(BlueEcoCheapseats)

sensitive<-subset(BlueEcoCheapseats,BlueEcoCheapseats$Price.Sensitivity>3)

sensitive

ggplot(Personal, aes(x = Price.Sensitivity,fill=factor(Price.Sensitivity))) + geom\_histogram(binwidth = 1, color = "white") +

scale\_fill\_brewer(palette="Blues") + mytheme + ggtitle ("Histogram of Price Sensitivity Filled with Type of Travel")#+ coord\_polar(theta = "x")

#plot of flight per year

x <- Others$Arrival.Delay.in.Minutes

y <- Others$Satisfaction

group <- Others$Gender

test<-data.frame(x,y,group)

ggplot(test, aes(x, y, shape=group, colour=group, fill=group)) +

geom\_smooth(method="lm") +

geom\_point(size=1) +

theme\_bw() +

xlab("Flights per Year") +

ylab("Satisfaction") +

ggtitle("Linear Model") +

#scale\_fill\_brewer(palette = "RdYlBu")+

expand\_limits(y=0) +

scale\_y\_continuous(breaks = 0:10) +

scale\_x\_continuous(breaks = 1998:2015)

x <- Others$Arrival.Delay.in.Minutes

y <- Others$Satisfaction

group <- Others$Type.of.Travel

test<-data.frame(x,y,group)

ggplot(test, aes(x, y, shape=group, colour=group, fill=group)) +

geom\_smooth(method="lm") +

geom\_point(size=1) +

theme\_bw() +

xlab("Arrival Delay in Minutes") +

ylab("Satisfaction") +

ggtitle("Linear Model") +

#scale\_fill\_brewer(palette = "RdYlBu")+

expand\_limits(y=0,x=0) #+

# scale\_y\_continuous(breaks = 0:10) +

# scale\_x\_continuous(breaks = 1998:2015)

x <- Others$Flights.Per.Year

y <- Others$Satisfaction

group <- Others$Gender

test<-data.frame(x,y,group)

ggplot(test, aes(x, y, shape=group, colour=group, fill=group)) +

geom\_smooth(method="lm") +

geom\_point(size=1) +

theme\_bw() +

xlab("Flight per Year") +

ylab("Satisfaction") +

ggtitle("Linear Model") +

#scale\_fill\_brewer(palette = "RdYlBu")+

expand\_limits(y=0) #+

# scale\_y\_continuous(breaks = 0:10) +

# scale\_x\_continuous(breaks = 1998:2015)

x <- Others$Arrival.Delay.in.Minutes

y <- Others$Satisfaction

group <- Others$Gender

test<-data.frame(x,y,group)

ggplot(test, aes(x, y, shape=group, colour=group, fill=group)) +

geom\_smooth(method="lm") +

geom\_point(size=1) +

theme\_bw() +

xlab("Arrival Delay in Minutes") +

ylab("Satisfaction") +

ggtitle("Linear Model") +

#scale\_fill\_brewer(palette = "RdYlBu")+

expand\_limits(y=0) #+

#scale\_y\_continuous(breaks = 0:10) +

#scale\_x\_continuous(breaks = 1998:2015)

lm<-lm(data=BlueEcoCheapseats,Satisfaction~Age)

summary(lm)

x <- BlueEcoCheapseats$Age

y <- BlueEcoCheapseats$Satisfaction

group <- BlueEcoCheapseats$Gender

test<-data.frame(x,y,group)

ggplot(test, aes(x, y, shape=group, colour=group, fill=group)) +

geom\_smooth(method="lm") +

geom\_point(size=1) +

theme\_bw() +

xlab("Age") +

ylab("Satisfaction") +

ggtitle("Linear Model") +

#scale\_fill\_brewer(palette = "RdYlBu")+

expand\_limits(y=0)

#histogram of age

Hist\_Age <- ggplot(BlueEcoCheapseats, aes(x = Age,fill=factor(Type.of.Travel))) + geom\_histogram(binwidth = 10, color = "white") +

scale\_fill\_brewer(palette="Blues")+ mytheme #+ coord\_polar(theta = "x")

Hist\_Age <- Hist\_Age+ggtitle("Histogram of Age Filled with Satisfaction")

Hist\_Age

summary(BlueEcoCheapseats)

table(BlueEcoCheapseats$Type.of.Travel)

Personal<-subset(BlueEcoCheapseats,BlueEcoCheapseats$Type.of.Travel=="Personal Travel")

View(Personal)

str(Personal)

personal<-Personal[,c(-2,-9,-13,-16,-17)]

View(personal)

str(personal)

personal<-personal[,-25]

lm7<-lm(data=personal,Satisfaction~Age+Gender+Price.Sensitivity+Flights.Per.Year+Loyalty+Total.Freq.Flyer.Accts+Shopping.Amount.at.Airport+Eating.and.Drinking.at.Airport+Departure.Delay.in.Minutes+Arrival.Delay.in.Minutes+Flight.time.in.minutes)

summary(lm7)

lm8<-lm(data=personal,Satisfaction~Age+Gender+Price.Sensitivity+Flights.Per.Year+Total.Freq.Flyer.Accts+Shopping.Amount.at.Airport+Eating.and.Drinking.at.Airport+Arrival.Delay.in.Minutes)

summary(lm8)

# Using the subset to run Association Rules Mining

#group age

createage<-function(vec){

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

vBuckets[between(vec,1,30)] <- "young"

vBuckets[between(vec,31,60)] <- "middle aged"

vBuckets[between(vec,60,85)] <- "old"

return(vBuckets)

}

#Flight time class

createFlightTime<-function(vec){

vBuckets <- replicate(length(vec), "late more than one hour")

vBuckets[between(vec,0,15)] <- "late less than 15 min"

vBuckets[between(vec,16,30)] <- "late between 16 and 30 min"

vBuckets[between(vec,31,60)] <- "late between 31 and 60 min"

return(vBuckets)

}

createarriveTime<-function(vec){

vBuckets <- replicate(length(vec), "late more than one hour")

vBuckets[between(vec,0,10)] <- "late less than 10 min"

vBuckets[between(vec,11,30)] <- "late between 11 and 30 min"

vBuckets[between(vec,31,40)] <- "late between 31 and 40 min"

vBuckets[between(vec,41,60)] <- "late between 40 and 60 min"

return(vBuckets)

}

#converting age group

dfbluecheapEco$Agehierachy<-createage(dfbluecheapEco$Age)

#converting Departure late time and Arrive late time

dfbluecheapEco$DepartureDelayMin1<-createFlightTime(dfbluecheapEco$DepartureDelayMin)

dfbluecheapEco$ArrivalDelayMin1<-createarriveTime(dfbluecheapEco$ArrivalDelayMin)

# Coerce the dfbluecheapEco data frame into a sparse transactions matrix using:

ruleDF<-data.frame(dfbluecheapEco$Satisfactionwith3,dfbluecheapEco$Agehierachy,dfbluecheapEco$DepartureDelayMin1,dfbluecheapEco$ArrivalDelayMin1)

dfbluecheapEcoX<-as(ruleDF,'transactions')

# Make sure you create a data frame before creating matrix

# Run the apriori command to try and predict happy customers (as defined by their overall satisfaction being high - above 3).

resultset<-apriori(dfbluecheapEcoX,

parameter = list(support=0.1,confidence=0.5),

appearance = list(default='lhs',rhs=('dfbluecheapEco.Satisfactionwith3=happy')))

summary(resultset)

inspect(resultset)

#lhs rhs support confidence lift count

#[4] {dfbluecheapEco.Agehierachy=middle aged} => {dfbluecheapEco.Satisfactionwith3=happy} 0.4062645 0.7940139 1.1191273 8781

#[5] {dfbluecheapEco.ArrivalDelayMin1=late less than 10 min} => {dfbluecheapEco.Satisfactionwith3=happy} 0.5134635 0.7776065 1.0960018 11098

#[6] {dfbluecheapEco.DepartureDelayMin1=late less than 15 min} => {dfbluecheapEco.Satisfactionwith3=happy} 0.5258166 0.7611680 1.0728325 11365

#[13] {dfbluecheapEco.DepartureDelayMin1=late less than 15 min,

#dfbluecheapEco.ArrivalDelayMin1=late less than 10 min} => {dfbluecheapEco.Satisfactionwith3=happy} 0.4838531 0.7802149 1.0996781 10458

plot(resultset)

Svm on the same Dataset

library(dplyr)

library(kernlab)

GroupByStatus<-group\_by(spring19survey, spring19survey$Airline.Status)

#Clean NA and non-integer in satisfaction

GroupByStatus<-GroupByStatus[-which(GroupByStatus$Satisfaction==2.5),]

GroupByStatus<-GroupByStatus[-which(GroupByStatus$Satisfaction==3.5),]

GroupByStatus<-GroupByStatus[-which(GroupByStatus$Satisfaction==4.5),]

GroupByStatus<-na.omit(GroupByStatus)

Blue<-subset(GroupByStatus,GroupByStatus$Airline.Status=="Blue")

BlueEco<-subset(Blue,Blue$Class=="Eco")

PN<-table(BlueEco$Partner.Name)

PN<-data.frame(PN)

PN[1,1]

BlueEcoCheapseats<-subset(BlueEco,BlueEco$Partner.Name==PN[1,1])

Cheap<-BlueEcoCheapseats

Cheap$happyCust<-Cheap$Satisfaction>3

dim(Cheap)

dim(Cheap)[1]

RandomIndex<-sample(1:dim(Cheap)[1])

length(RandomIndex)

cutpoint<-floor(2\*dim(Cheap)[1]/3)

cutpoint

train<-Cheap[RandomIndex[1:cutpoint],]

View(train)

test<-Cheap[RandomIndex[(cutpoint+1):dim(Cheap)[1]],]

#---happyCust with Age + Gender + Price.Sensitivity

SVM<-ksvm(happyCust ~ Age + Gender + Price.Sensitivity , data=train,

kernel= "rbfdot", kpar = "automatic", C = 5, cross = 3, prob.model = TRUE)

# Linear Modeling

# Blue Customers Only

# Why Blue Customers?

tapply(dataset$Satisfaction, dataset$Airline.Status, mean)

# It seems that Blue Customers have lowest mean Satisfaction

dataset\_blue <- dataset[dataset$Airline.Status == 'Blue',]

# Why Eco Class?

tapply(dataset\_blue$Satisfaction, dataset\_blue$Class, mean)

# It seems that Eco Class customers are least satisfied

dataset\_blueEco <- dataset\_blue[dataset\_blue$Class == 'Eco',]

# Which Type of Travel

tapply(dataset\_blueEco$Satisfaction, dataset\_blueEco$Type.of.Travel, mean)

# It seems Personal Travel has least Satisfaction

dataset\_blueEcoPersonal <- dataset\_blueEco[dataset\_blueEco$Type.of.Travel == 'Personal Travel',]

str(dataset\_blueEcoPersonal)

# Linear Modeling on Blue-Eco-Personal Travel

modelBEP <- lm(data = dataset\_blueEcoPersonal,

Satisfaction ~ Age + Gender +

Price.Sensitivity + Flights.Per.Year +

Loyalty + Total.Freq.Flyer.Accts + Shopping.Amount.at.Airport +

Eating.and.Drinking.at.Airport + Departure.Delay.in.Minutes +

Arrival.Delay.in.Minutes + Flight.time.in.minutes)

summary(modelBEP)

# Keeping only Highly Significant Coefficients

modelBEP <- lm(data = dataset\_blueEcoPersonal,

Satisfaction ~ Age + Gender +

Price.Sensitivity + Flights.Per.Year +

Total.Freq.Flyer.Accts + Eating.and.Drinking.at.Airport + Departure.Delay.in.Minutes +

Arrival.Delay.in.Minutes + Flight.time.in.minutes)

summary(modelBEP)

# Choosing Different Airlines in BlueEcoPersonal

tapply(dataset\_blueEcoPersonal$Satisfaction, dataset\_blueEcoPersonal$Partner.Name, mean)

# Selecting GoingNorth as it has least Satisfaction

dataset\_blueEcoPersonalNorth <- dataset\_blueEcoPersonal[dataset\_blueEcoPersonal$Partner.Name == 'GoingNorth Airlines Inc.',]

# Linear Modeling on GoingNorth

modelBEPN <- lm(data = dataset\_blueEcoPersonalNorth,

Satisfaction ~ Age + Gender +

Price.Sensitivity + Flights.Per.Year +

Loyalty + Total.Freq.Flyer.Accts + Shopping.Amount.at.Airport +

Eating.and.Drinking.at.Airport + Departure.Delay.in.Minutes +

Arrival.Delay.in.Minutes + Flight.time.in.minutes)

summary(modelBEPN)

# Keeping only significant Co-efficients

modelBEPN <- lm(data = dataset\_blueEcoPersonalNorth,

Satisfaction ~ Arrival.Delay.in.Minutes)

summary(modelBEPN)

# Selecting Cheapseats as it has most travelers

dataset\_blueEcoPersonalCheapseats <- dataset\_blueEcoPersonal[dataset\_blueEcoPersonal$Partner.Name == 'Cheapseats Airlines Inc.',]

# Linear Modeling on Cheapseats

modelBEPCS <- lm(data = dataset\_blueEcoPersonalCheapseats,

Satisfaction ~ Age + Gender +

Price.Sensitivity + Flights.Per.Year +

Loyalty + Total.Freq.Flyer.Accts + Shopping.Amount.at.Airport +

Eating.and.Drinking.at.Airport + Departure.Delay.in.Minutes +

Arrival.Delay.in.Minutes + Flight.time.in.minutes)

summary(modelBEPCS)

# Keeping on Significant Variables

modelBEPCS <- lm(data = dataset\_blueEcoPersonalCheapseats,

Satisfaction ~ Age + Gender +

Price.Sensitivity + Flights.Per.Year +

Total.Freq.Flyer.Accts +

Eating.and.Drinking.at.Airport +

Arrival.Delay.in.Minutes)

summary(modelBEPCS)

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

# Selecting Cheapseats as it has most travelers

dataset\_blueEcoCheapseats <- dataset\_blueEco[dataset\_blueEco$Partner.Name == 'Cheapseats Airlines Inc.',]

# Linear Modeling on Cheapseats

modelBECS <- lm(data = dataset\_blueEcoCheapseats,

Satisfaction ~ Age + Gender + Type.of.Travel +

Price.Sensitivity + Flights.Per.Year +

Loyalty + Total.Freq.Flyer.Accts + Shopping.Amount.at.Airport +

Eating.and.Drinking.at.Airport + Departure.Delay.in.Minutes +

Arrival.Delay.in.Minutes + Flight.time.in.minutes)

summary(modelBECS)

# Keeping only Significant Variables

modelBECS <- lm(data = dataset\_blueEcoCheapseats,

Satisfaction ~ Age + Gender + Type.of.Travel +

Price.Sensitivity + Flights.Per.Year +

Arrival.Delay.in.Minutes)

summary(modelBECS)

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

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

# Selecting GoingNorth as it has least satisfaction

dataset\_blueEcoNorth <- dataset\_blueEco[dataset\_blueEco$Partner.Name == 'GoingNorth Airlines Inc.',]

# Linear Modeling on Going North

modelBEN <- lm(data = dataset\_blueEcoCheapseats,

Satisfaction ~ Age + Gender + Type.of.Travel +

Price.Sensitivity + Flights.Per.Year +

Arrival.Delay.in.Minutes)

summary(modelBEN)

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

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

# Selecting West as it has most satisfaction

dataset\_blueEcoWest <- dataset\_blueEco[dataset\_blueEco$Partner.Name == 'West Airways Inc.',]

# Linear Modeling on Cheapseats

modelBEW <- lm(data = dataset\_blueEcoWest,

Satisfaction ~ Age + Gender + Type.of.Travel +

Price.Sensitivity + Flights.Per.Year +

Loyalty + Total.Freq.Flyer.Accts + Shopping.Amount.at.Airport +

Eating.and.Drinking.at.Airport + Departure.Delay.in.Minutes +

Arrival.Delay.in.Minutes + Flight.time.in.minutes)

summary(modelBEW)

# Keeping only Significant Variables

modelBEW <- lm(data = dataset\_blueEcoWest,

Satisfaction ~ Type.of.Travel +

Arrival.Delay.in.Minutes)

summary(modelBEW)

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