Airline Satisfaction

Final Report

Syracuse University

IST 687: Introduction to Data Science

Spring 2021

Instructor: Gary Krudys

Team Members: Elizabeth Axworthy, Jessica Calhoun, Emily Murdoch

TABLE OF CONTENTS

EXECUTIVE SUMMARY

1. INTRODUCTION
2. BUSINESS QUESTIONS
3. DATA AQUISITION
4. DESCRIPTIVE ANALYSIS
5. DATA MODELING T ECHNIQUES
6. RESULTS
   1. Actionable insights
   2. Interpretation
   3. Summary
7. APPENDIX

**INTRODUCTION**

Gaining and maintaining customer loyalty is a competitive field for airlines, as seen by the abundance of incentives like credit card rewards and frequent flyer programs. It’s easy to understand why; customers purchase your services; you make a profit. As such, finding ways to increase loyalty is an important question for airlines. One way to do this is to optimize customer satisfaction in the hopes that happy customers are frequent customers, or at least customers that come to you instead of others. In this study, we will be acting as a group of consultants for the airline Cheapseats. We will take a collection of customer satisfaction responses recorded by an airport and attempt to find the factors that lead towards greater customer satisfaction. We will then use our findings to build a profile for the average Cheapseats customer and make general recommendations that any airline could apply to boost their own numbers.

**BUSINESS QUESTIONS**

To help with our overall aim of optimizing customer satisfaction, we developed the following eight business questions to help focus our analysis. We then linked the questions to the appropriate variables. Each member of the team took on three questions, with the names noted in the table below. The definitions of the various attributes can be found in the *Data Acquistion* section.

|  |  |  |
| --- | --- | --- |
| **Question #1:** How cancellations impact satisfaction, do certain cities have more cancellations? [Lizzy] | **Attributes:** Satisfaction, Flight Cancelled, Origin City, Destination City | **Reasoning:** Delays are a common problem at airlines, and something that can cause a lot of dissatisfaction. By nailing down details, we can better work out how to mitigate this problem with the greatest effect. |
| **Question #2:** Which origin city has the highest satisfaction? Which destination city has the lowest satisfaction?[Emily] | **Attributes:** Satisfaction, Origin City, Destination City | **Reasoning:** With this question we are looking to see if certain cities provide a higher or lower satisfaction and experience for travelers utilizing those airlines. |
| **Question #3:** Does Class have an impact on satisfaction? [Emily] | **Attributes:** Satisfaction, Class, Airline Name, Airline Code | **Reasoning:** One would assume that a better class on a flight means a better flight experience. We are looking to see if this is the case or if traveler’s satisfaction level is not impacted by the class, as well as if there’s a difference between satisfaction of classes amongst airlines. |
| **Question #4:** Relationship between loyalty and satisfaction? [Emily] | **Attributes:** Airline Name, Airline Code, Satisfaction, Airline Status, Percent of Flights with Other Airlines, Number of Other Loyalty Cards | **Reasoning:** Understanding why someone chose a specific airline is important to know, so we are looking to see if the satisfaction levels of a traveler translates to airline loyalty and repeat business. |
| **Question #5**: What is the most common type of travel customers are doing through our airline? [Jessica] | **Attributes:** Airline Code, Airline Name, Class, Type of Travel | **Reasoning:**  This question helps us show the types of travelers that are most common with our airline and can also be used as a comparison versus other airlines and the type of passenger their customers are. |
| **Question #6:** How does Flight Time effect satisfaction? [Lizzy] | **Attributes:** Satisfaction,Flight time in minutes, Flight Distance. | **Reasoning:** If there is a pattern between longer fights and less satisfaction, we could offer things like in-flight rewards for customer mollification. |
| **Question #7:** How in-airport behavior impacts a customer satisfaction? (Shopping Amounts, Eating and Drinking Amount) [Lizzy] | **Attributes:** Satisfaction, Shopping Amount, Eating and Drinking. | **Reasoning:** If there is a strong correlation between in-airport behavior and overall satisfaction, we might be able to utilize that with things such as coupons to restaurants, or by providing more complementary services. |
| **Question #8**: How departure delays impact satisfaction, does the origin state or the destination state impact customer satisfaction? Is there a relationship between the type of travel (business, personal, mileage) and the origin or destination state? [Jessica] | **Attributes:** Satisfaction, Origin State, Destination State, Departure Delay in Minutes, Type of Travel | **Reasoning:** One of the most common attributes that impact customer satisfaction is delays and cancellations. If we can pinpoint departure states, delay times, etc., we might better understand locations that need to be improved or discarded. |

**DATA AQUISTION**

*Data Selection*

For our project, we used the airport satisfaction survey data provided by the course. This data was originally provided as an excel workbook containing 129889 rows of 28 attributes, with the output variable being overall satisfaction in the responding customer’s airline experience, and the input variables including factors such as starting point, destination, flight length in distance and time, class of flight, and some demographic information about the customer.

*Data Structure – Attributes and Meanings*

**Attributes Name:**

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

*Data Cleaning*

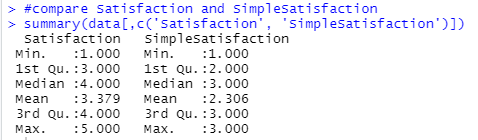
An initial analysis of the data revealed a few abnormalities that had to be cleaned up. The first was that in the Satisfaction column, 3 rows had values in an odd format which made the intended value difficult to understand. As there were relatively few of these, they were dropped. The nest change was that 6 rows had a number above 100% for Percentage of Flights Taken with Other Airlines, which seemed like an error. Again, these were dropped. The column ‘Orgin City’ was renamed to ‘Origin City’. Many rows had blanks in the ‘Departure Delay in Minute’ or ‘Arrival Delay in Minutes’ columns, which would have become NA’s once they were loaded into R. These were filled with 0’s to prevent problems. 337 uncancelled flights had no flight time in minutes and due to not being able to accurately estimate what the actual time might be, these were dropped. As these were all found during the preliminary look at the data, these changes were all made in Excel and saved to a new workbook to maintain the original data in case it was needed.

*Data Transformation*

For some parts of the analysis, a secondary satisfaction column was used. This column was generated by taking the existing satisfaction responses which used the 1-5 scale, and translating them into a 1-3 scale. The translation was as follows, with the old score being on the left and the new score being on the right: [1,2]-1, (2,4)-2, [4,5]-3. In our new column we thus have 1 signifying a low score, 2 a medium score, and 3 a high score. This new column is named SimpleSatisfaction. Three sub dataframes were created based on the SimpleSatisfaction rating, these being highSatisfaction, mediumSatisfaction, and lowSatisfaction. Other transformations were also used when investigating different questions, and will be described under the appropriate heading for that question in the following section.

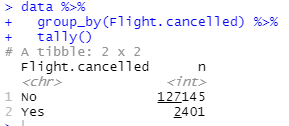
**DESCRIPTIVE ANALYSIS**

*General*



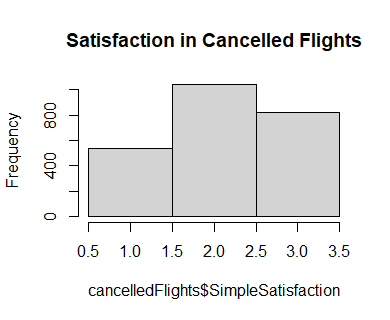
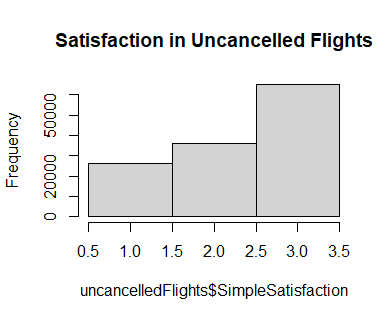
*Question 1: Cancellations*

First, we get a count of how many flights are canceled vs. Uncancelled:

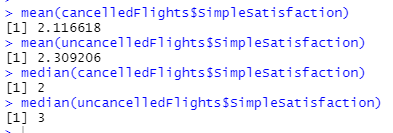


As we would hope, the majority of flights are uncancelled. However, about 1.89% of the flights in our data are cancelled. This is a large enough portion of flights to be worth investigating.

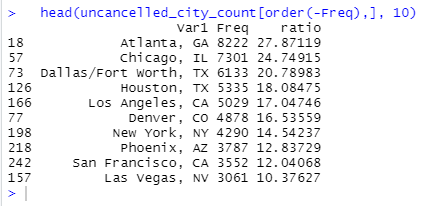
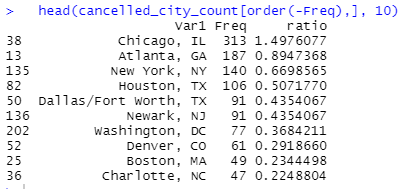
For quick comparison, two smaller dataframes were extracted from the larger ‘data’ dataframe to separate out cancelled flights vs non cancelled flights. Using SimpleSatisfaction, we take a look at the frequency of different satisfaction values across cancelled and uncancelled flights and compare their distributions.



We can see a left shift in the mean in the distribution of the cancelled flights. This is confirmed by taking a look at the mean and median values of SimpleSatisfaction.



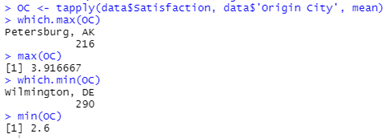
Our next step is to take a look at the origin cities for the most cancelled flights and compare against the uncancelled flights, which will be our control group so that a city won’t be over represented as having a lot of cancellations purely because it has a lot of flights:



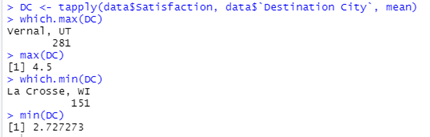
The unique values found in top 10 cities represented in our cancelled flights data frame are Newark NJ, Washington DC, Boston MA, and Charlotte NC.

*Question 2: Origin and Destination*

We wanted to know if there was a particular city that went above and beyond and provided excellent satisfaction or if there was a city that was lacking. To gain this insight, we created a data frame with the Satisfaction level and Origin City first. The statistical analysis on this was a tapply in order to pull each Origin City with the mean satisfaction level for each. From this data, we wanted to know the highest and lowest levels. A simple which.max and which.min told us the cities and max and min told us the value of that city.



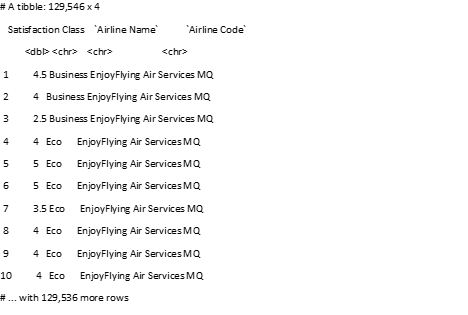
From this one can see that Petersburg, Arkansas had the highest average satisfaction levels with a 3.97 compared to the lowest, Wilmington, DE with 2.6 level. This information does not give a lot of insight, but it shows that the average satisfaction levels are not incredibly high or low, given the range from 1 to 5. Next, we looked at the Destination Cities average satisfaction.



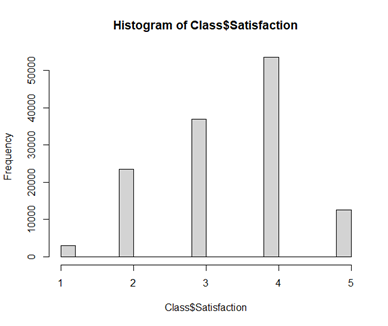
The highest average of the Destination Cities was Vernal, UT with 4.5 satisfaction. The lowest average was La Crosse, WI with a 2.73 satisfaction. Comparing these cities to the Origin Cities, the highest and lowest are not the same. Therefore, just because one city has a high average for one instance does not mean it has the same levels of satisfaction for another instance. In conclusion, much more insight into the specific airport cities would be needed to find makes them have a high or low average satisfaction level. From these simple statistics, one can see however, that no airport in one city is far exceeding or underachieving compared to the others.

*Question 3: Class*

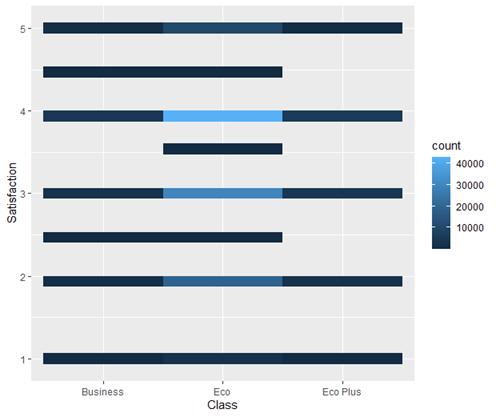
The next question we investigated was if the class that a person is flying had an impact on satisfaction or not. We created a data frame with the pertinent columns of data, including Satisfaction, Class, Airline Name and Airline Code, as shown below.



Next, we wanted to get a good idea of the spread and distribution of the levels of satisfaction. A histogram was able to show this, with the most occurring frequencies of satisfaction falling at a level of 4, followed by a 3.

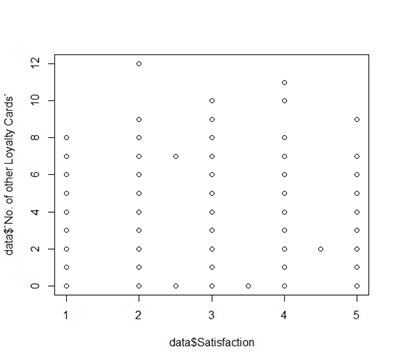


Then, the next distribution we wanted to look at was between Class and Satisfaction. We utilized ggplot for this and created a two-dimensional bin plot. From this, one can see the three different class levels and the counts in each satisfaction level. The highest count is found at a four level of satisfaction (aligning with the previous histogram), and in the Eco class. We can also see that in the Eco class, it has a higher count for the two, three, and five levels as well. This gives us the preliminary idea that the Eco class tends to impact satisfaction levels, or that the Eco class just has more overall counts than Business and Eco Plus. We will have to look at the regression analysis next to see if there is a linear relationship.

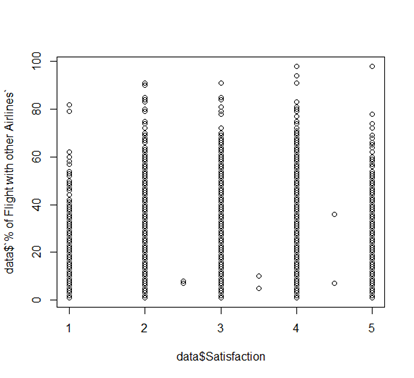


*Question 4: Loyalty and Satisfaction*

The fourth question we had was if there was a relationship between loyalty and satisfaction. For this we were looking at the variables of Satisfaction, No. of other Loyalty Cards, and % of Flight with other Airlines. First, we plotted the level of satisfaction to the No. of other Loyalty Cards. This showed us if there were people with certain numbers of loyalty cards at each level of satisfaction. For example, at the level of two, there were people with 0-9 loyalty cards and 12 loyalty cards.

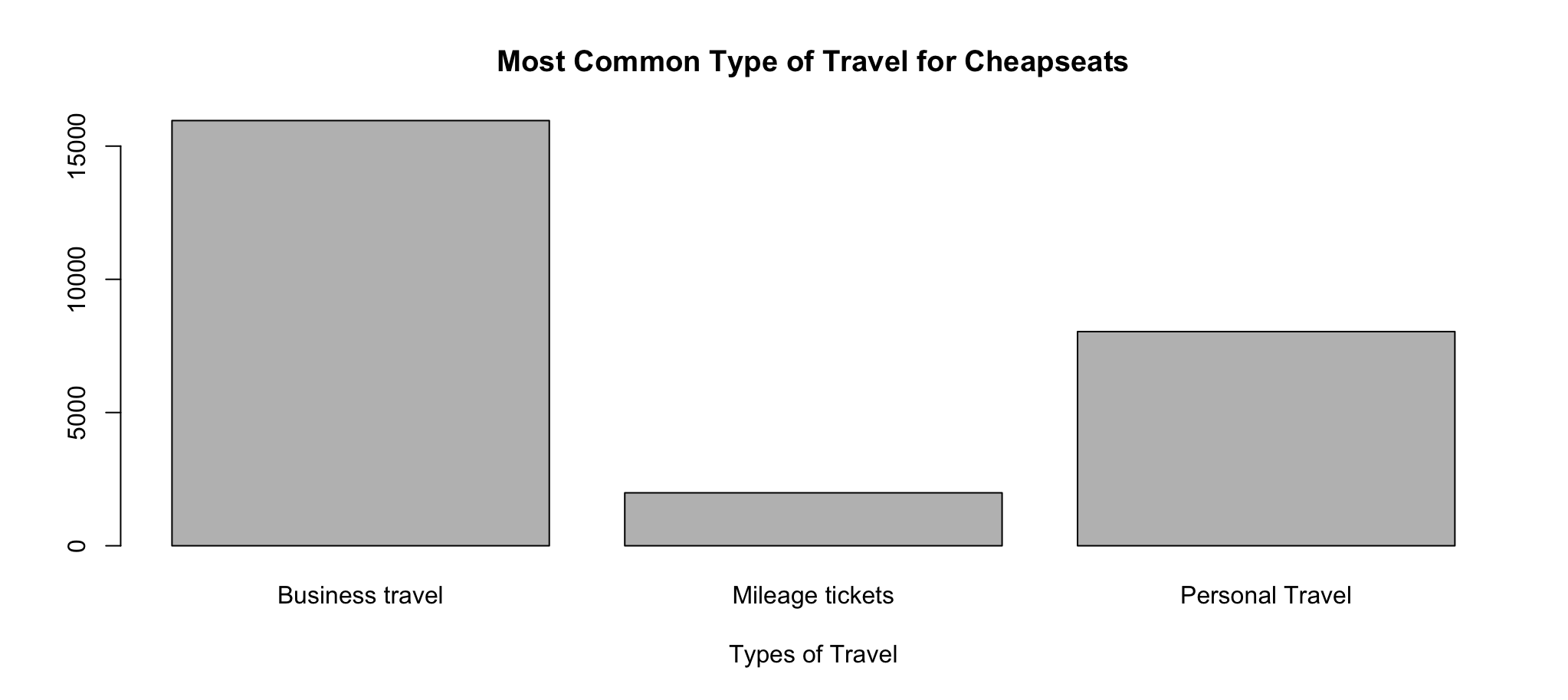


The next plot was between Satisfaction and % of Flight with other Airlines. Again, it shows the plotted frequency of percentages at each level of satisfaction. Both plots do not give us much insight other than seeing frequencies and all of which are similar at each level of satisfaction. Therefore, a linear regression model will give us a better indication of relationship among the variables.

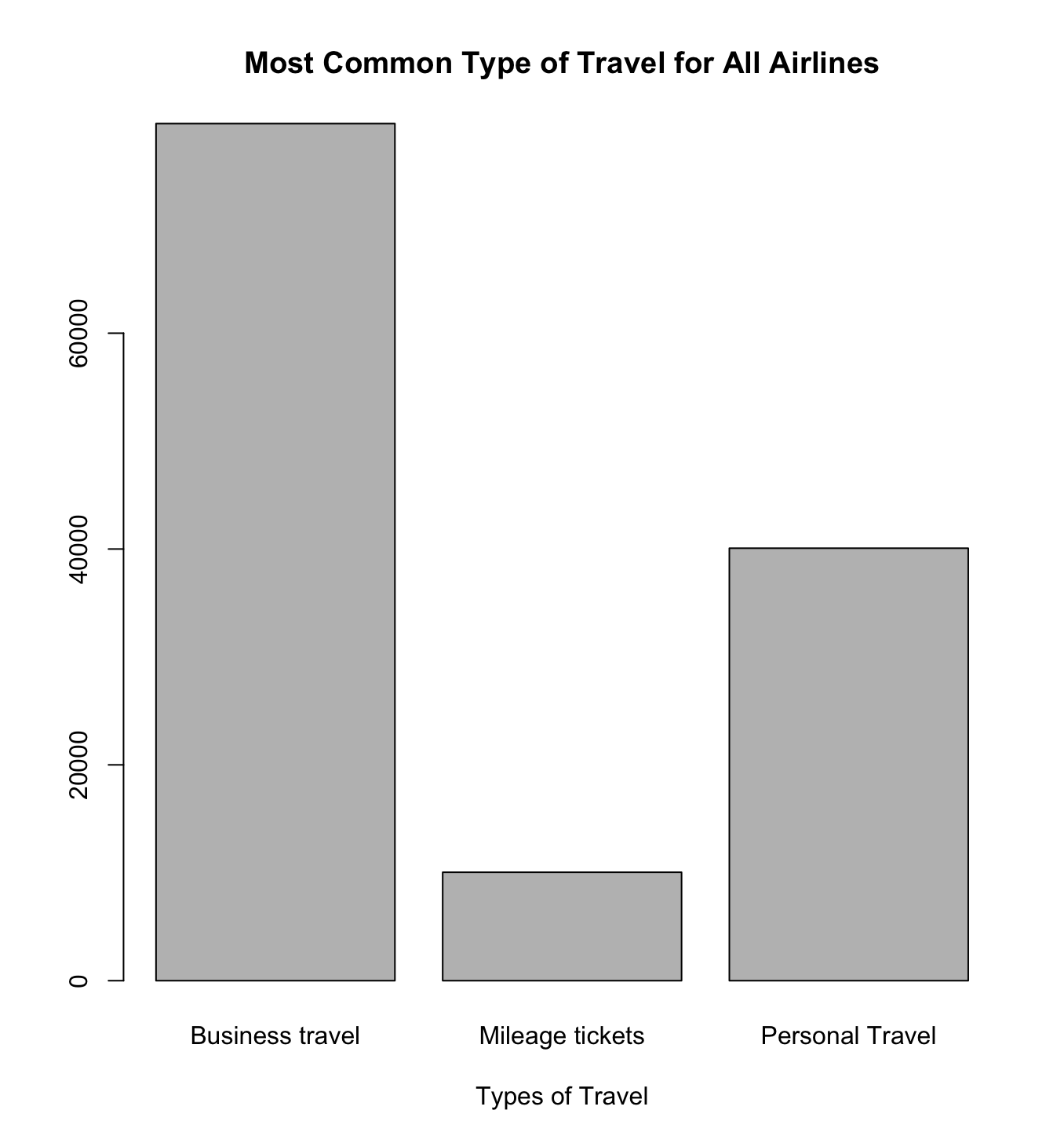


*Question 5: Most Frequent Type of Travel*

We wanted to understand the most common type if travel for our Cheapseats Airlines customers. Understanding our core customer base is an important element in ensuring that we are providing the best service possible to our customers. For this we looked at only responses from Cheapseats Airlines flights. We filtered down that data and removed all other airline flight information. Then we used a bar chart to provide a visual representation of the type of travel our customers are doing.



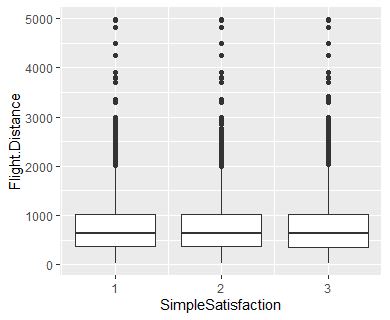
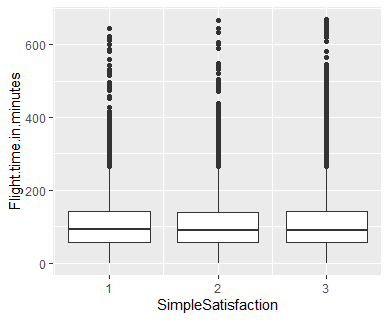
As we can see from the data, our customers are primarily doing Business Travel, with a smaller portion doing Personal Travel and finally a very small portion of our customers are using Mileage Tickets to fly with Cheapseats Airlines.



We can see that our individual airline’s distribution of Business, Personal, and Mileage travel closely mirrors the overall distribution of travel types for all airlines.

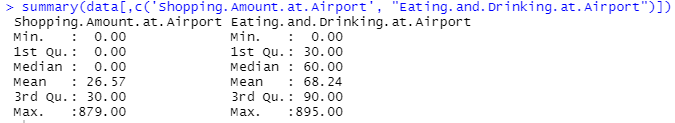
*Question 6: Flight Time and Satisfaction*

A boxplot comparison of SimpleSatisfaction and Flight.time or Flight.Distance reveals no significant effect on satisfaction caused by a longer flight. As there is not a lot that could be done to reduce flight length, this is the better outcome.



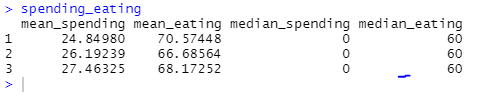
*Question 7: In Airport Behavior*

We first took a look at the range of values for Shopping.Amount and Eating.and.Drinking

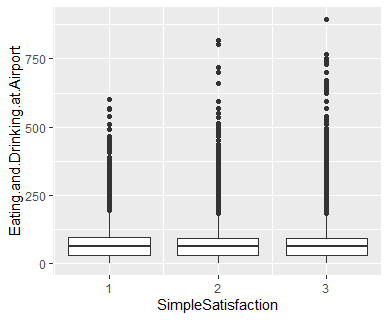
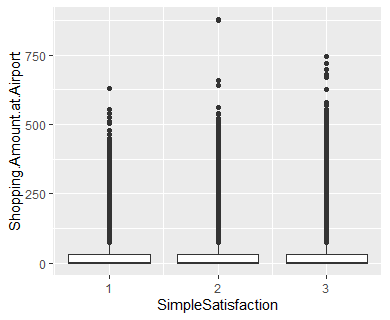


Over 50% of customers spend nothing, and of those that do spend money, 50% spend less than $30. Most customers do spend money on food and drinks at the airport, with the majority of spending being between $30 and $60

Next we’ll look at the mean and median of Spending and Eating grouped by Simple Satisfaction:



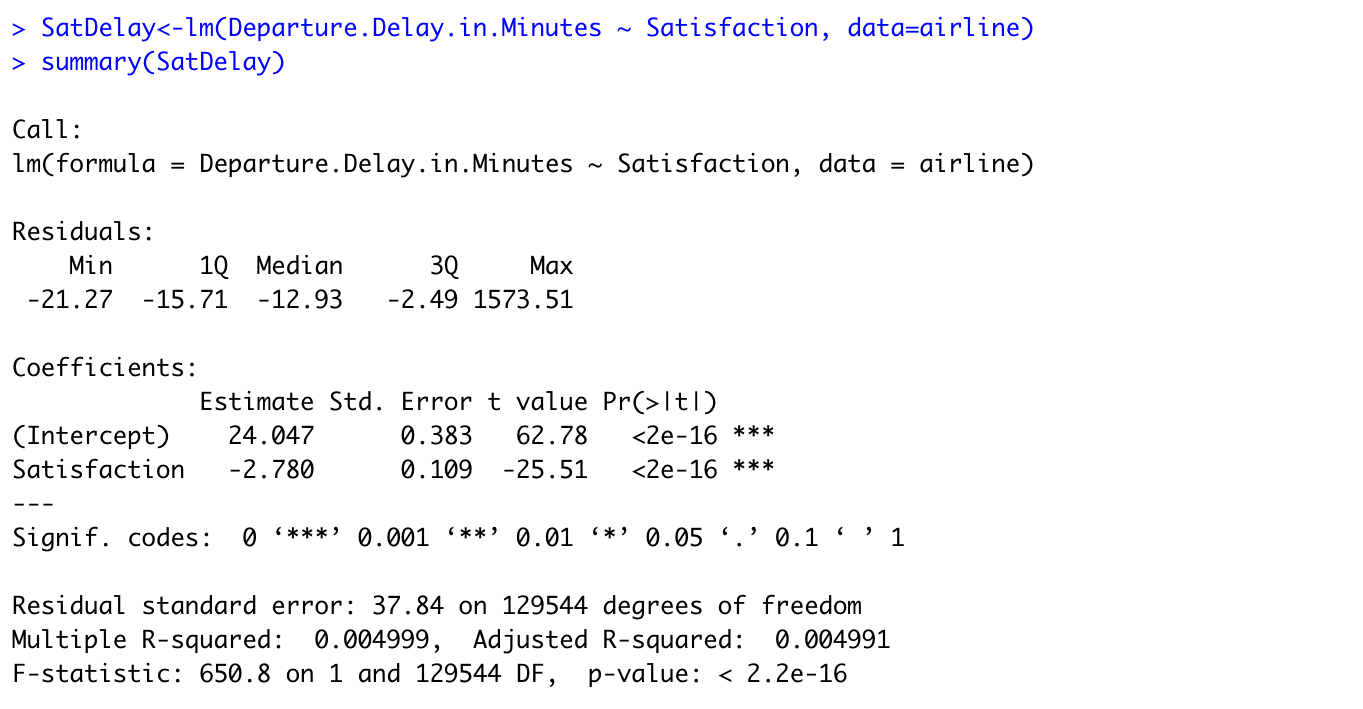
The differences between Spending and Eating between the different satisfaction groups is minimal in the means, and non-existent in the medians. This suggests that these areas aren’t worth further looking into if the goal is to increase satisfaction, but we’ll do one further study to confirm this before abandoning this avenue.



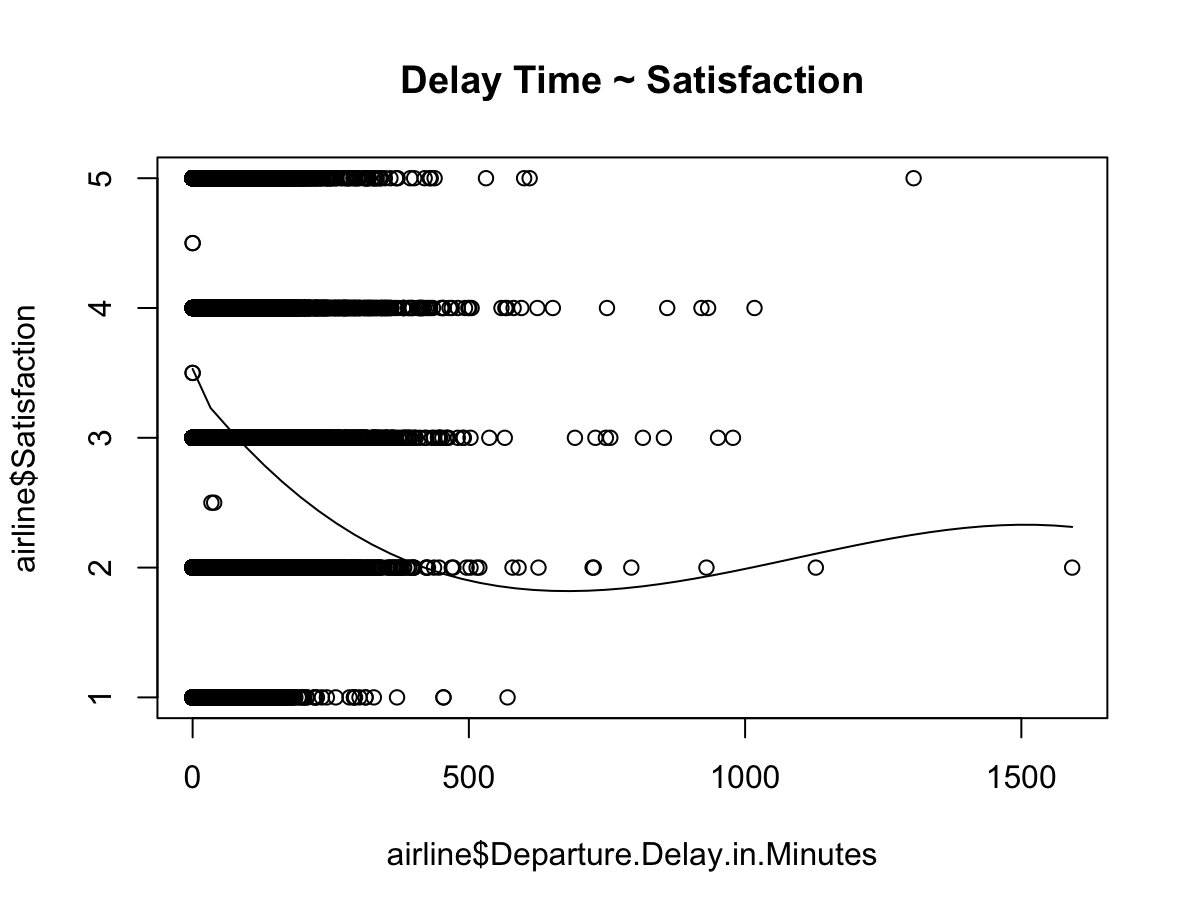
Boxplots for Spending and Eating, grouped by Simple Satisfaction show identical ranges outside of the outliers. As there is no significant difference is satisfaction from different spending amounts, no further analysis will be done. We can use the takeaway that almost everyone spend money on food though and incorporate that into our recommendations in the results portion of our analysis.

*Question 8: Satisfaction & Delay, Satisfaction & Origin/Destination State, and Type of Travel & Origin/Destination State*

We ran a linear regression with the variables of Satisfaction and Delay in Minutes. We used the model to determine if satisfaction and delay in minutes had a low p-value or a high p-value.

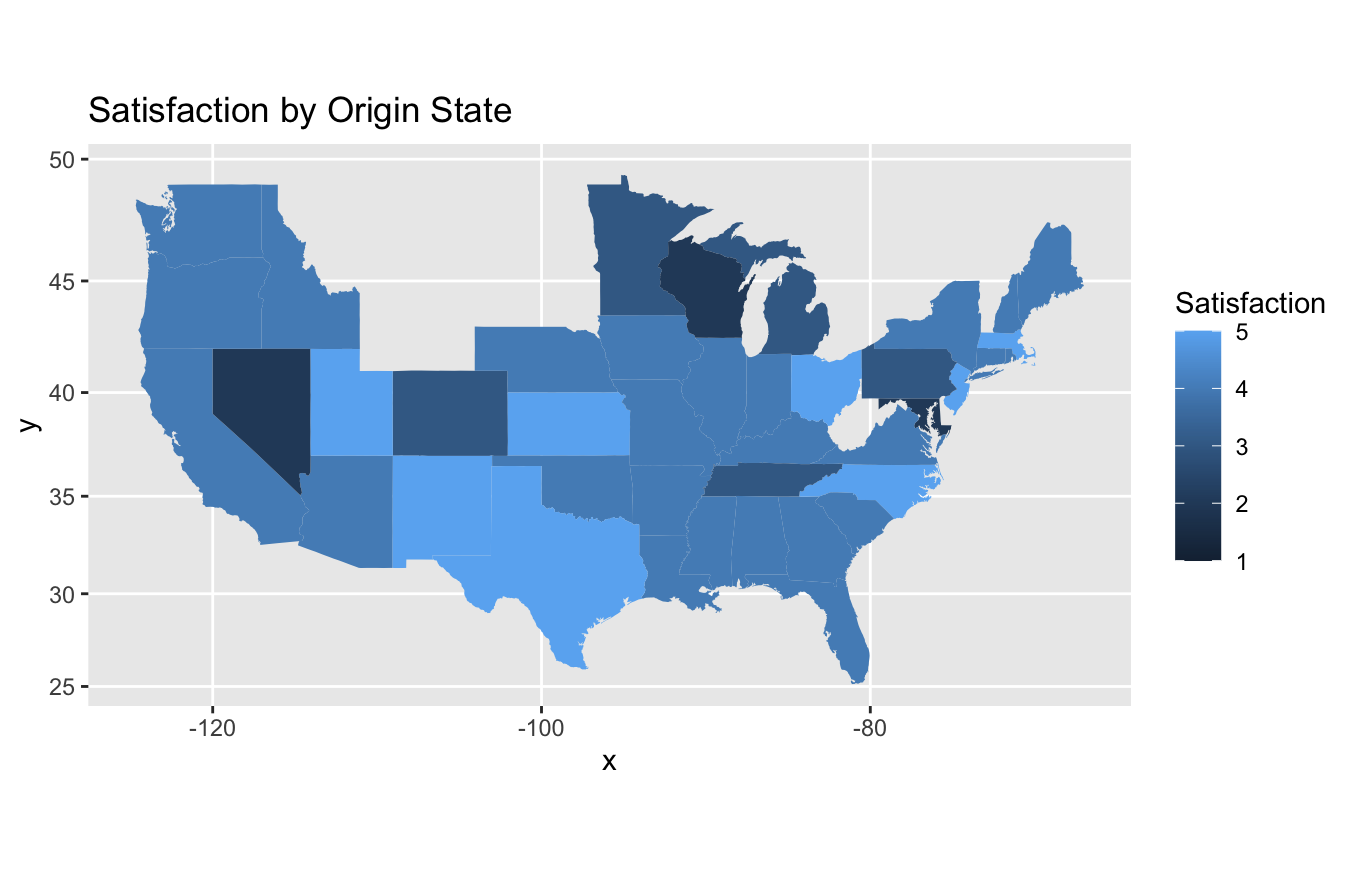


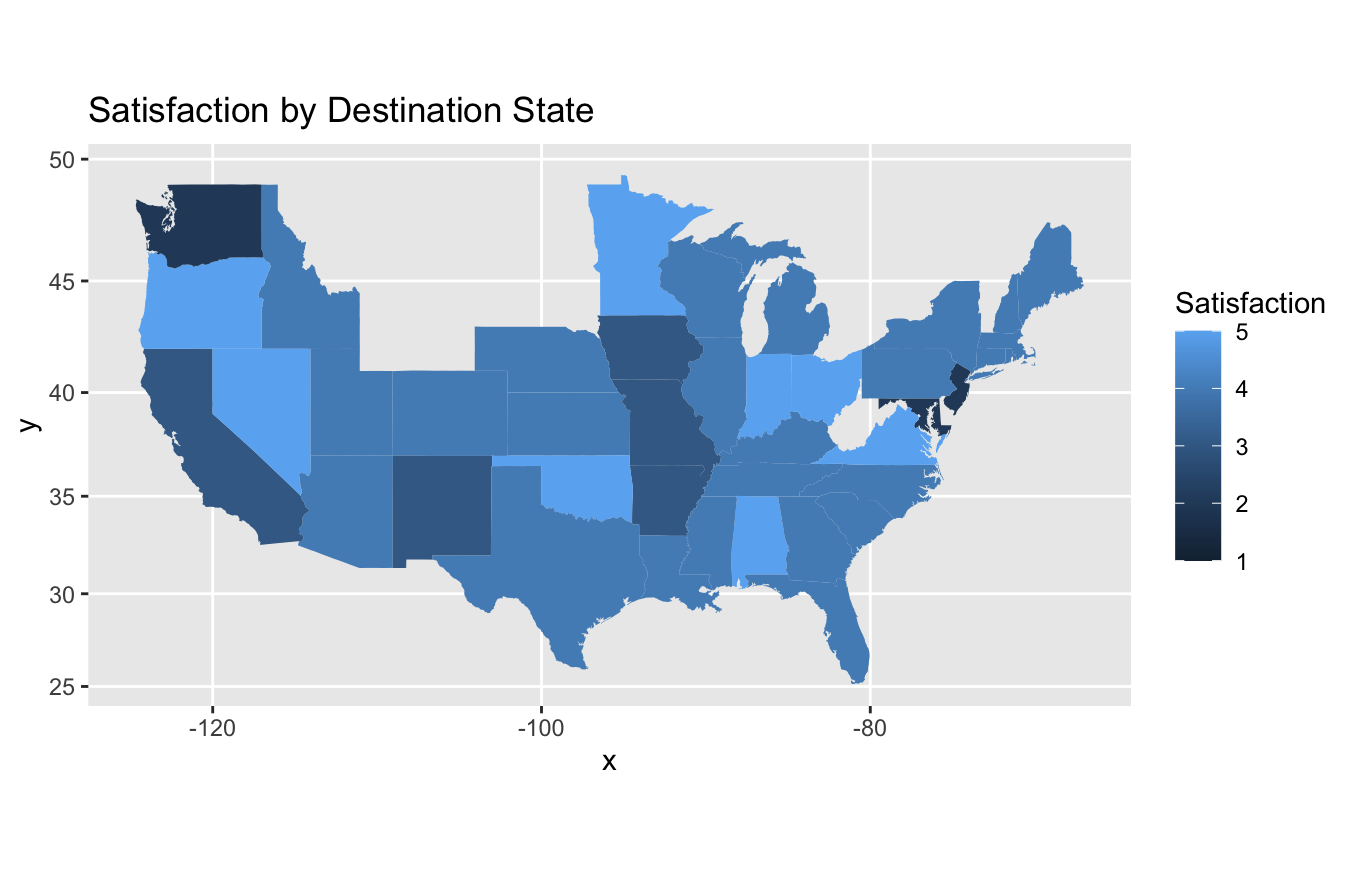
The linear regression clearly shows that our p-value is very low for these two data points. We then used the data points to run a scatter plot to better understand the relationship between Departure Delay in Minutes and customer Satisfaction.



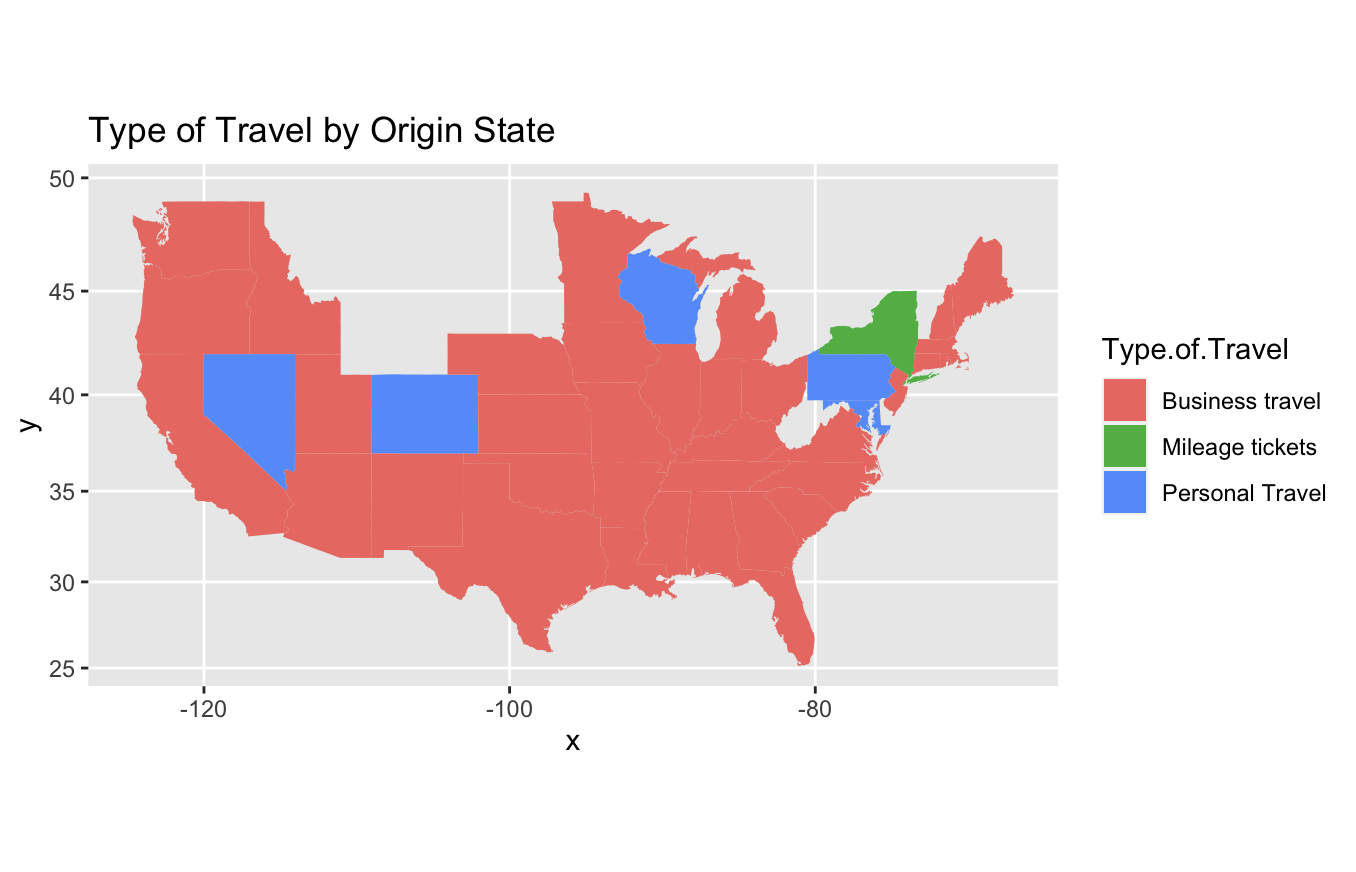
As we can see from the scatter plot, the customer satisfaction immediately drops from our Mean satisfaction of 3.357157, ending in a satisfaction rating of just around a satisfaction level of ~2 at 1500 minutes of delay. The moment there is any type of delay, the satisfaction level drops, which is an important data point when reviewing ways to improve customer satisfaction.

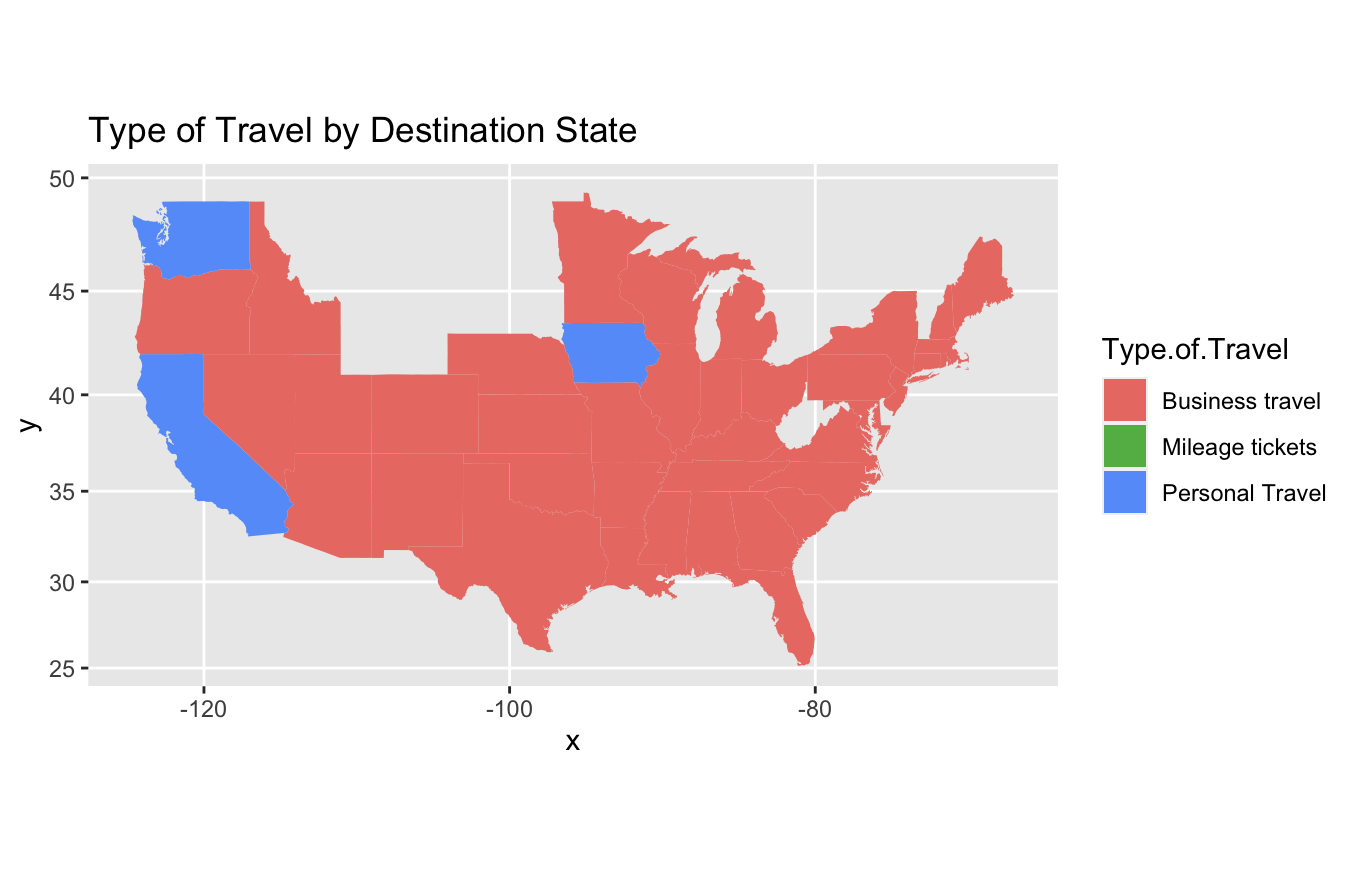
Next, we wanted to see how the state of origin or the destination state for Cheapseats customers had any impact on a customer’s overall satisfaction. Using ggplot, we are able to see which states our customers are most satisfied when they depart, and which states customers are most satisfied when they arrive. This helps us better understand where our hubs should be located, and which destinations are ideal for satisfied customers.





We also wanted to see if there was any focus on the type of travel that our customers were doing (business, personal, mileage) and the state of origin or the state of destination. We were hoping these maps might provide more insight into where our customers are and where to focus our efforts for business travelers, however, the data is too widespread to yield any tangible results.

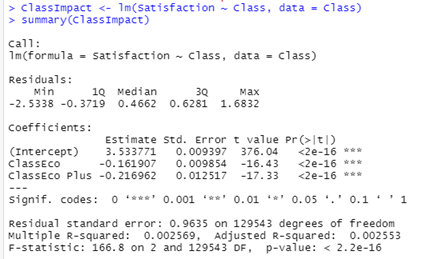




**DATA MODELING TECHNIQUES**

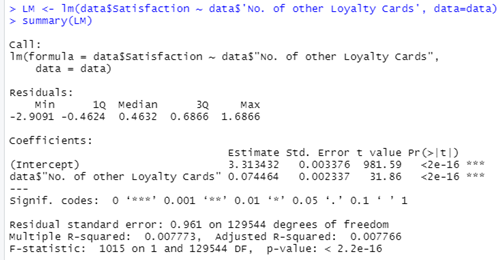
*Model 1: Class*

A linear regression model was ran with the variables of Satisfaction and Class. According to the results, the intercept, Eco, and Eco Plus are all significant with very low p-values for each. This lets us look at the r^2 value of 0.0025. A value that indicates that 0.25% of the class level determines the satisfaction levels. This is a very low percentage. While class may play some role in influencing satisfaction level, it does not play a significant role in impacting it.



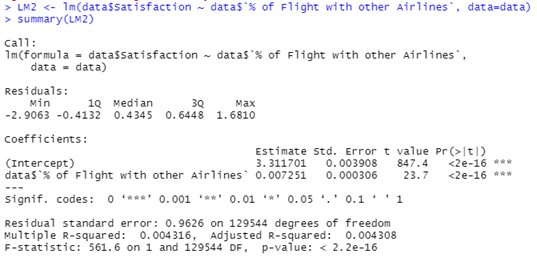
*Model 2: Loyalty and Satisfaction*

The first linear regression between Satisfaction and No. of other Loyalty Cards is shown below.



The intercept and the No. of other Loyalty Cards are both statistically significant with very low p-values. The r^2 value is 0.0078, which indicates that 0.78% of the No. of other Loyalty Cards determines the level of satisfaction. This is a very low percentage and therefore, cannot be a main influence of the level of satisfaction.

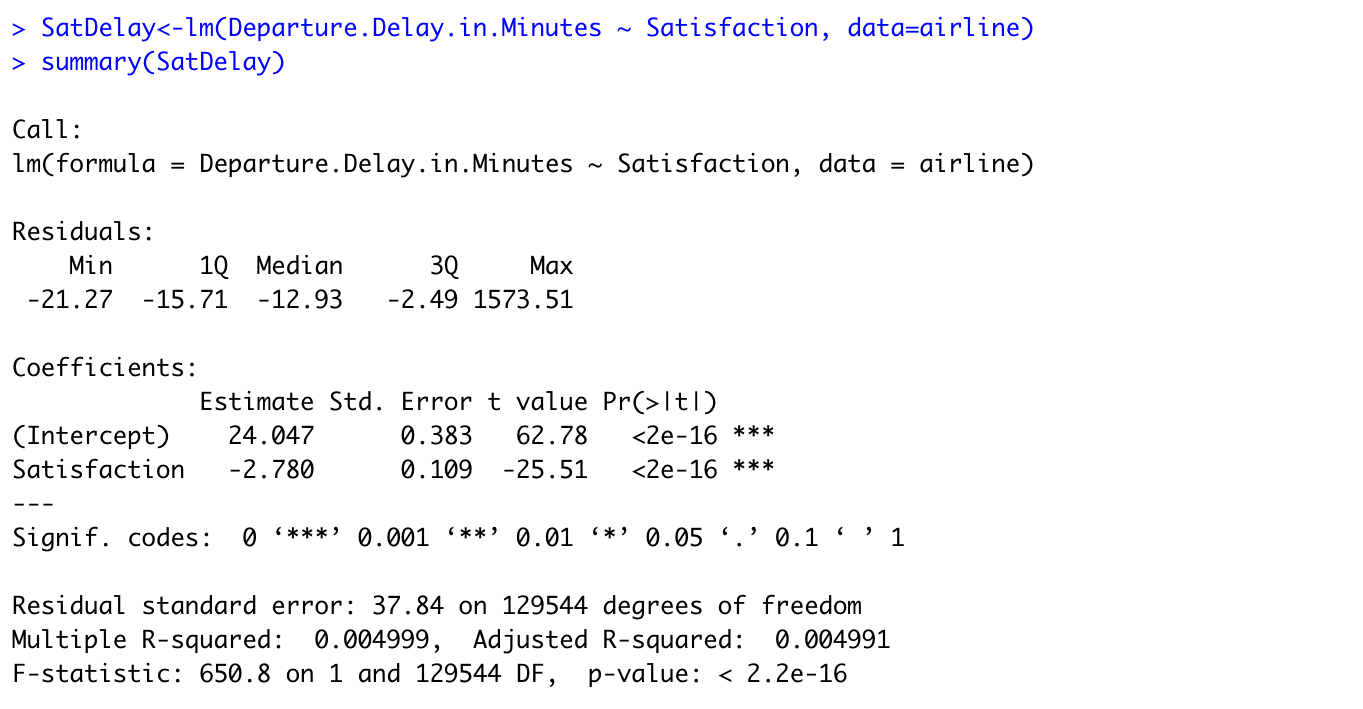
Lastly, we took the linear regression of satisfaction and % of Flight with other Airlines.



Both the % of Flight with other Airlines and the intercept are statistically significant with p-values below 0.05. Therefore, we look at the r^2 value which is 0.0043. This means that 0.43% of the % of Flight with other Airlines determines the level of satisfaction. This is also a very low percentage and shows that it only contributes but is not a main influence on the level of satisfaction.

*Model 3: Departure Delay and Satisfaction*

We ran a linear regression with the variables of Satisfaction and Delay in Minutes. According to the result is a very low p-values for each. This lets us look at the r^2 value of 0.004999. A value that indicates that 0.50% of the departure delays determines the satisfaction levels. This is a very low percentage. While departure delays play some role in influencing satisfaction level, it does not play a significant role in impacting it.



*Bonus Model – All Available Variables in Rcmdr*

I experimented with running a linear regression in Rcmdr with all the variables it would allow, and Satisfaction being the output. As we can see, the overall P-value of our equation is low, so the model is reliable, but the present variables only explain 10.4% of variation. I attempted to improve the model by removing the insignificant variables.

Coefficients:

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

(Intercept) -9.45170098 1.71309757 -5.517 3.45e-08 \*\*\*

Age -0.01122515 0.00016863 -66.566 < 2e-16 \*\*\*

Arrival.Delay.in.Minutes -0.00453490 0.00025868 -17.531 < 2e-16 \*\*\*

Day.of.Month 0.00081745 0.00029417 2.779 0.005456 \*\*

Departure.Delay.in.Minutes 0.00261927 0.00026170 10.009 < 2e-16 \*\*\*

Eating.and.Drinking.at.Airport -0.00001368 0.00004949 -0.276 0.782177

Flight.Distance -0.00005328 0.00001492 -3.570 0.000357 \*\*\*

Flight.time.in.minutes 0.00049769 0.00012168 4.090 4.31e-05 \*\*\*

No..of.other.Loyalty.Cards -0.03680237 0.00269149 -13.674 < 2e-16 \*\*\*

No.of.Flights.p.a. -0.01325611 0.00018606 -71.245 < 2e-16 \*\*\*

Price.Sensitivity -0.17728958 0.00467923 -37.889 < 2e-16 \*\*\*

Scheduled.Departure.Hour 0.00041230 0.00055383 0.744 0.456613

Shopping.Amount.at.Airport 0.00001572 0.00004798 0.328 0.743202

X..of.Flight.with.other.Airlines -0.00047381 0.00032452 -1.460 0.144287

Year.of.First.Flight 0.00691452 0.00085367 8.100 5.55e-16 \*\*\*

---

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

Residual standard error: 0.9132 on 129531 degrees of freedom

Multiple R-squared: 0.104, Adjusted R-squared: 0.1039

F-statistic: 1074 on 14 and 129531 DF, p-value: < 2.2e-16

I also tested for colinearity and found four of the variable had levels above 10, making them unsuitable to be used together.

Age Arrival.Delay.in.Minutes Day.of.Month

1.325076 15.371350 1.007352

Departure.Delay.in.Minutes Eating.and.Drinking.at.Airport 15.310214 1.037463

Flight.Distance Flight.time.in.minutes No..of.other.Loyalty.Cards 12.116623 12.149540 1.468130

No.of.Flights.p.a. Price.Sensitivity

1.112595 1.015275

Scheduled.Departure.Hour Shopping.Amount.at.Airport

1.018352 1.008492

X..of.Flight.with.other.Airlines Year.of.First.Flight

1.249750 1.003193

The second regression has no insignificant variable, but the R-value remains unchanged. The colinearity is still the same so we’ll start cutting those variables out. Flight time and flight distance are obviously related so we’ll choose one to throw out and one to keep. Flight time had the larger coefficient and smaller p-value so we’ll keep that.

Coefficients:

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

(Intercept) -9.40409931 1.71226486 -5.492 3.98e-08 \*\*\*

Age -0.01122142 0.00016710 -67.152 < 2e-16 \*\*\*

Arrival.Delay.in.Minutes -0.00455115 0.00025792 -17.646 < 2e-16 \*\*\*

Day.of.Month 0.00081916 0.00029416 2.785 0.005358 \*\*

Departure.Delay.in.Minutes 0.00264064 0.00026036 10.142 < 2e-16 \*\*\*

Flight.Distance -0.00005328 0.00001492 -3.570 0.000357 \*\*\*

Flight.time.in.minutes 0.00049726 0.00012166 4.087 4.37e-05 \*\*\*

No..of.other.Loyalty.Cards -0.03822817 0.00250498 -15.261 < 2e-16 \*\*\*

No.of.Flights.p.a. -0.01322905 0.00018339 -72.137 < 2e-16 \*\*\*

Price.Sensitivity -0.17684150 0.00466302 -37.924 < 2e-16 \*\*\*

Year.of.First.Flight 0.00689098 0.00085328 8.076 6.75e-16 \*\*\*

---

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

Residual standard error: 0.9132 on 129535 degrees of freedom

Multiple R-squared: 0.104, Adjusted R-squared: 0.1039

F-statistic: 1503 on 10 and 129535 DF, p-value: < 2.2e-16

R-Value has lowed for 0.1%, as has p-value of Flight time and Day of Month.

Coefficients:

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

(Intercept) -9.41674276 1.71233884 -5.499 3.82e-08 \*\*\*

Age -0.01122701 0.00016710 -67.186 < 2e-16 \*\*\*

Arrival.Delay.in.Minutes -0.00434461 0.00025136 -17.285 < 2e-16 \*\*\*

Day.of.Month 0.00086109 0.00029394 2.929 0.0034 \*\*

Departure.Delay.in.Minutes 0.00244634 0.00025462 9.608 < 2e-16 \*\*\*

Flight.time.in.minutes 0.00008121 0.00003496 2.323 0.0202 \*

No..of.other.Loyalty.Cards -0.03827925 0.00250506 -15.281 < 2e-16 \*\*\*

No.of.Flights.p.a. -0.01323460 0.00018339 -72.167 < 2e-16 \*\*\*

Price.Sensitivity -0.17692735 0.00466317 -37.941 < 2e-16 \*\*\*

Year.of.First.Flight 0.00689871 0.00085331 8.085 6.29e-16 \*\*\*

---

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

Residual standard error: 0.9133 on 129536 degrees of freedom

Multiple R-squared: 0.1039, Adjusted R-squared: 0.1038

F-statistic: 1669 on 9 and 129536 DF, p-value: < 2.2e-16

At this point the only remaining problem is with Arrival Delay and Departure Delay. For the same reasoning used previously, we’ll drop departure delay.

Age Arrival.Delay.in.Minutes Day.of.Month 1.301078 14.512776 1.005710

Departure.Delay.in.Minutes Flight.time.in.minutes 14.492451 1.002948

No..of.other.Loyalty.Cards No.of.Flights.p.a. Price.Sensitivity

1.271681 1.080759 1.008237

Year.of.First.Flight

1.002266

Ultimately, we were not able to improve our model by eliminating insignificant variables and collinearity. The p-value indicates that the model still might have some use though.

Coefficients:

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

(Intercept) -9.41674276 1.71233884 -5.499 3.82e-08 \*\*\*

Age -0.01122701 0.00016710 -67.186 < 2e-16 \*\*\*

Arrival.Delay.in.Minutes -0.00434461 0.00025136 -17.285 < 2e-16 \*\*\*

Day.of.Month 0.00086109 0.00029394 2.929 0.0034 \*\*

Departure.Delay.in.Minutes 0.00244634 0.00025462 9.608 < 2e-16 \*\*\*

Flight.time.in.minutes 0.00008121 0.00003496 2.323 0.0202 \*

No..of.other.Loyalty.Cards -0.03827925 0.00250506 -15.281 < 2e-16 \*\*\*

No.of.Flights.p.a. -0.01323460 0.00018339 -72.167 < 2e-16 \*\*\*

Price.Sensitivity -0.17692735 0.00466317 -37.941 < 2e-16 \*\*\*

Year.of.First.Flight 0.00689871 0.00085331 8.085 6.29e-16 \*\*\*

---

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

Residual standard error: 0.9133 on 129536 degrees of freedom

Multiple R-squared: 0.1039, Adjusted R-squared: 0.1038

F-statistic: 1669 on 9 and 129536 DF, p-value: < 2.2e-16

After running a RESET test, we found that part of why we’re not getting very useful data might be that this is non-linear, making the usefullness of the model at all questionable.

data: Satisfaction ~ Age + Arrival.Delay.in.Minutes + Day.of.Month + Flight.time.in.minutes + No..of.other.Loyalty.Cards + No.of.Flights.p.a. + Price.Sensitivity + Year.of.First.Flight

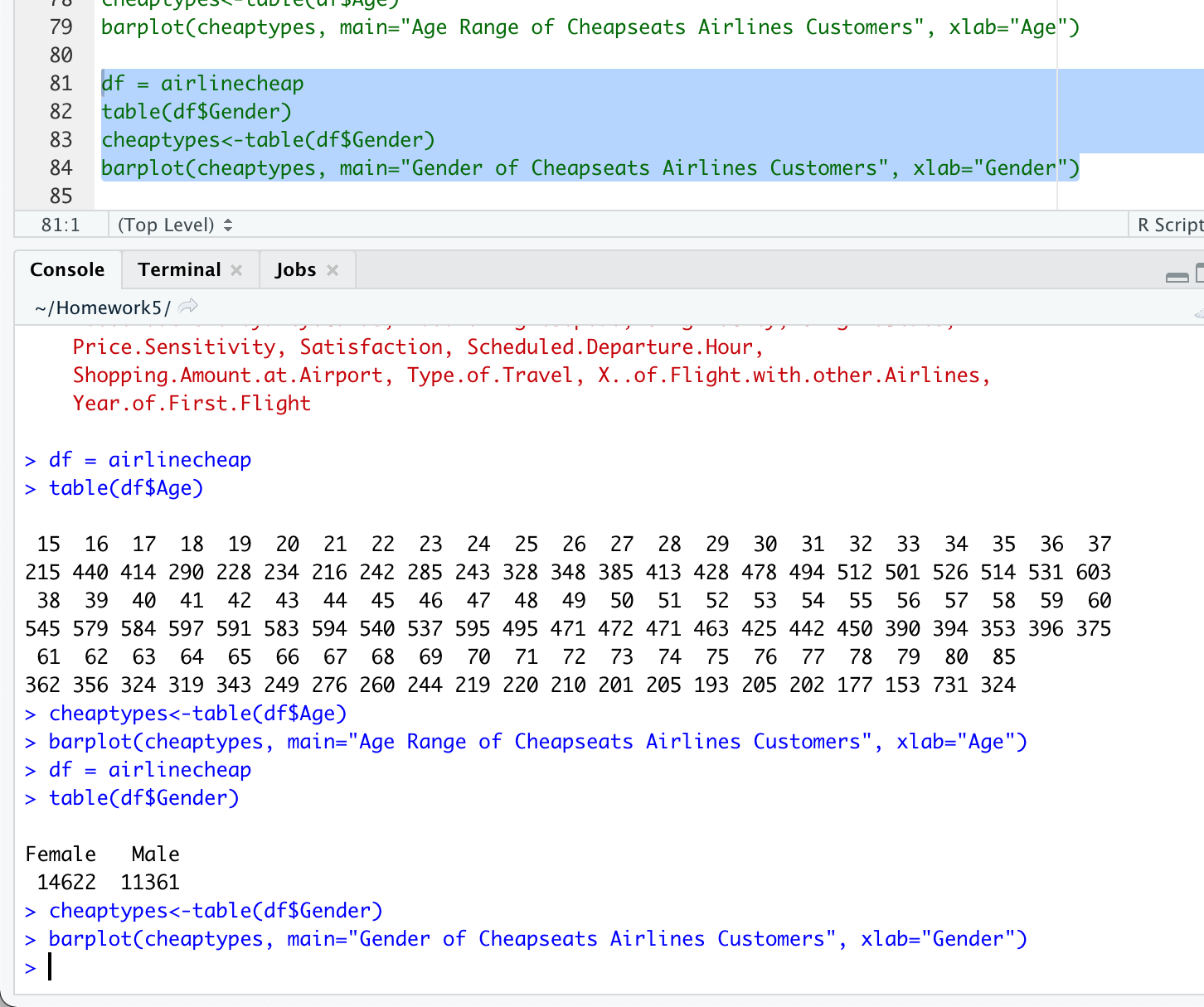
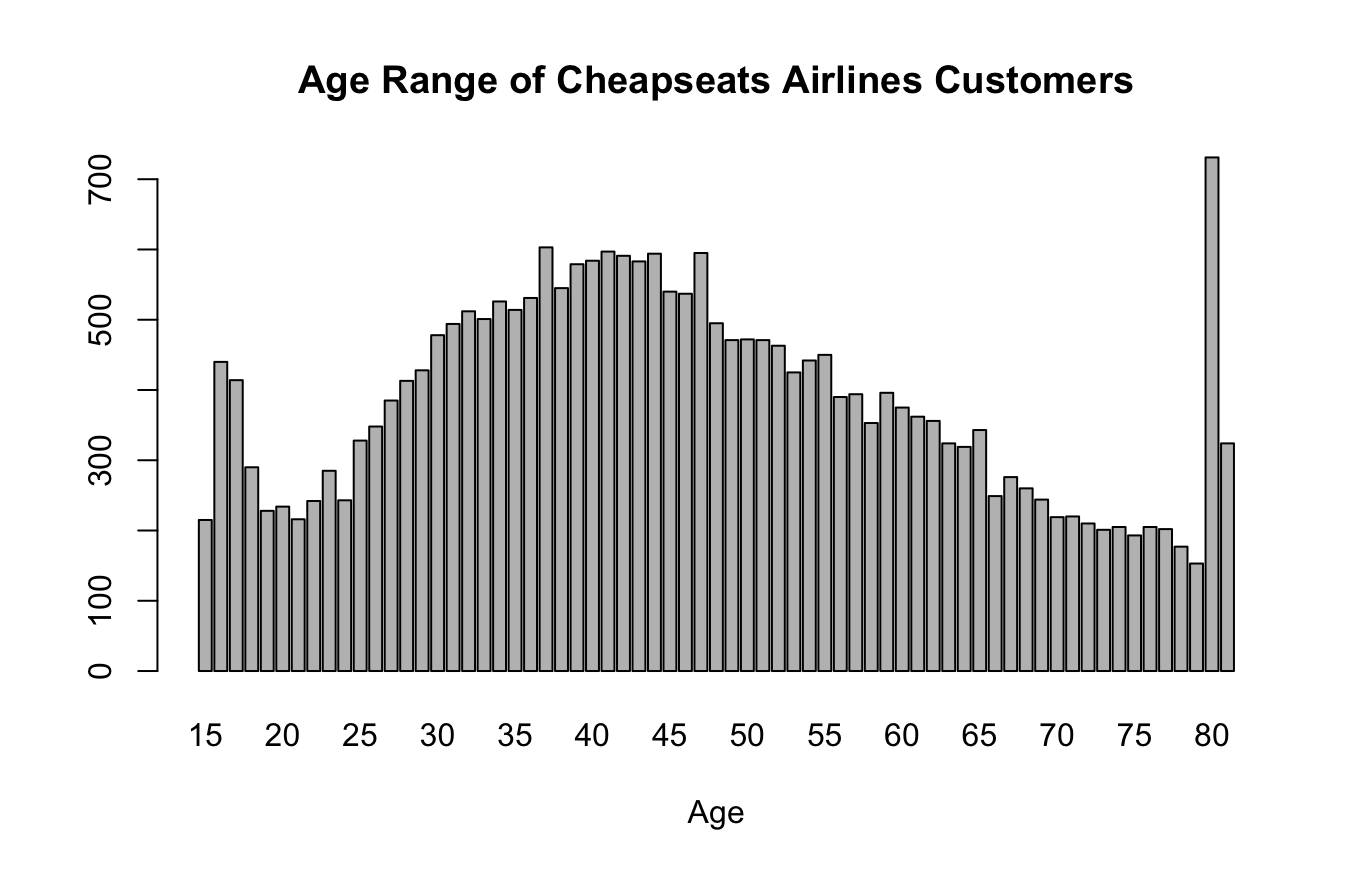
RESET = 958.43, df1 = 16, df2 = 129521, p-value < 2.2e-16

**RESULTS**

*Profile of Cheapseats Customers*

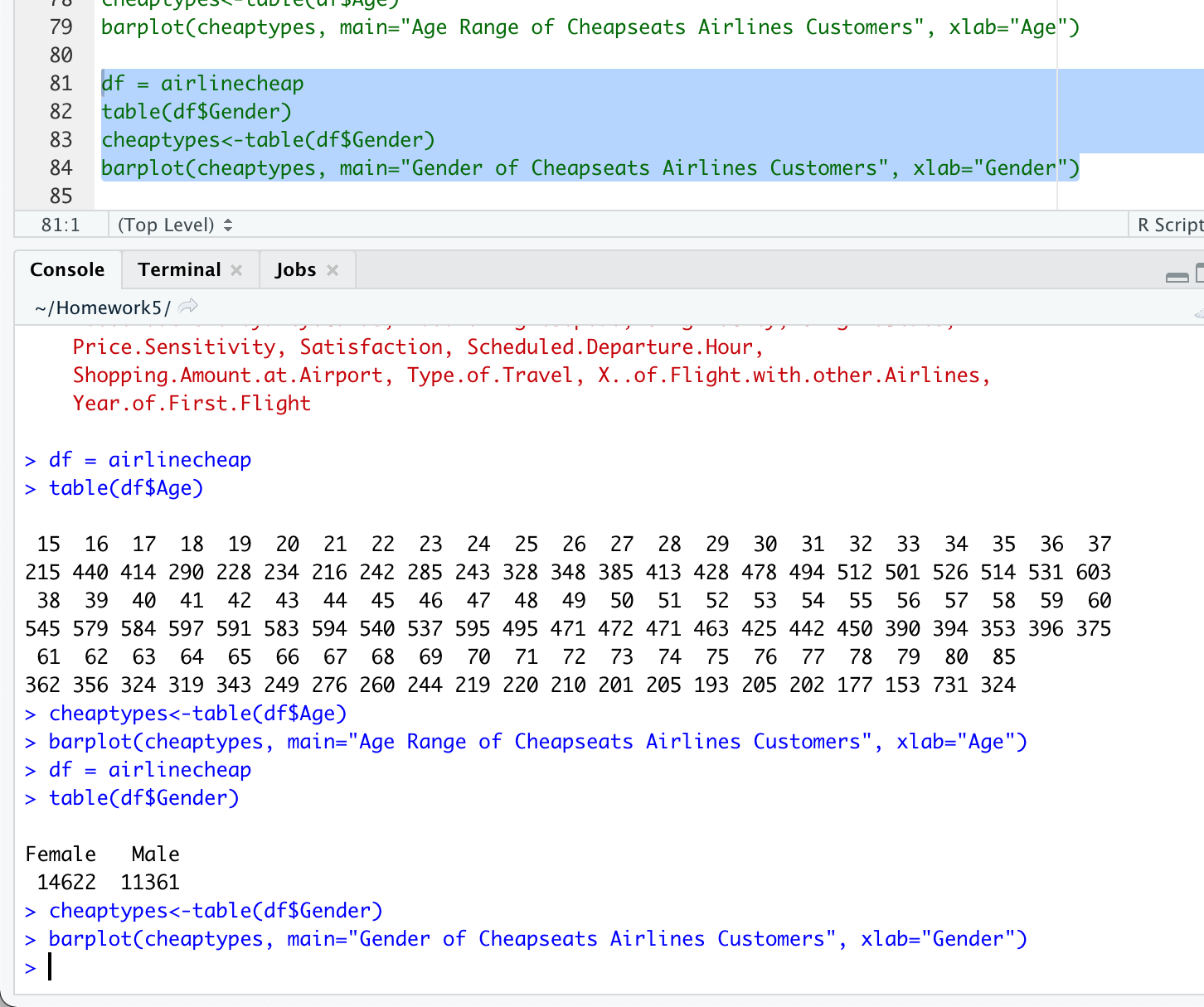
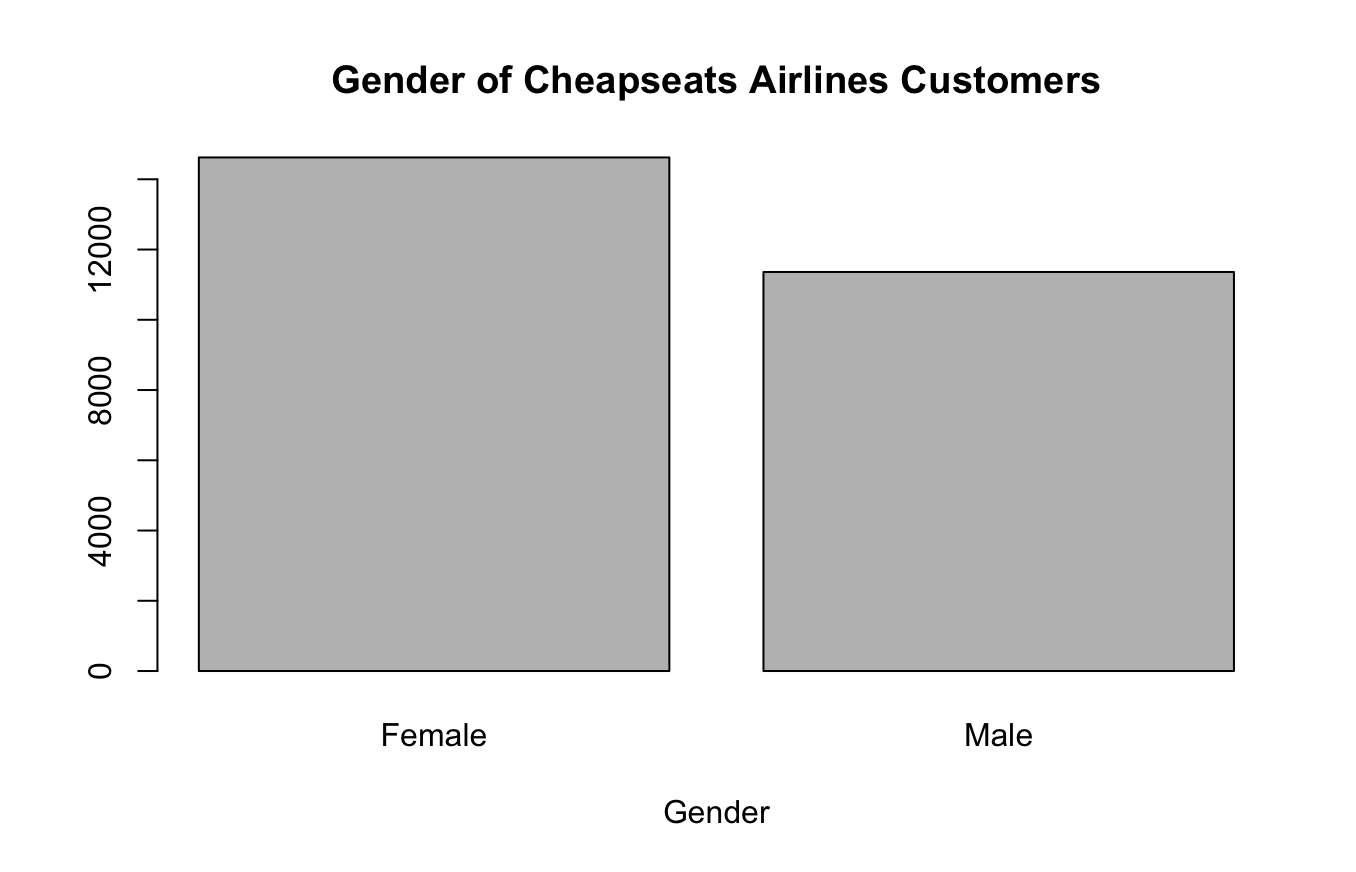
Age Range

We can see from the data of Cheapseats Airlines Customers, that the age ranges fall along a fairly normal distribution of ages with the peak age of customers being in the 40-45 year old range. We can also see a few outliers in the 80 year old customers and the 16-19 age range.



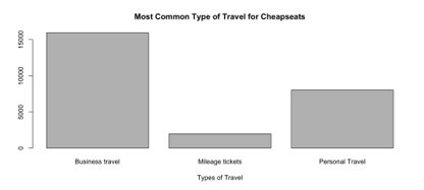
Gender

We can see from the data of Cheapseats Airlines Customers, that the distribution between genders is relatively close between male and female customers, with female customers accounting for 56% of our customers and male customers accounting for 44% of our customers.



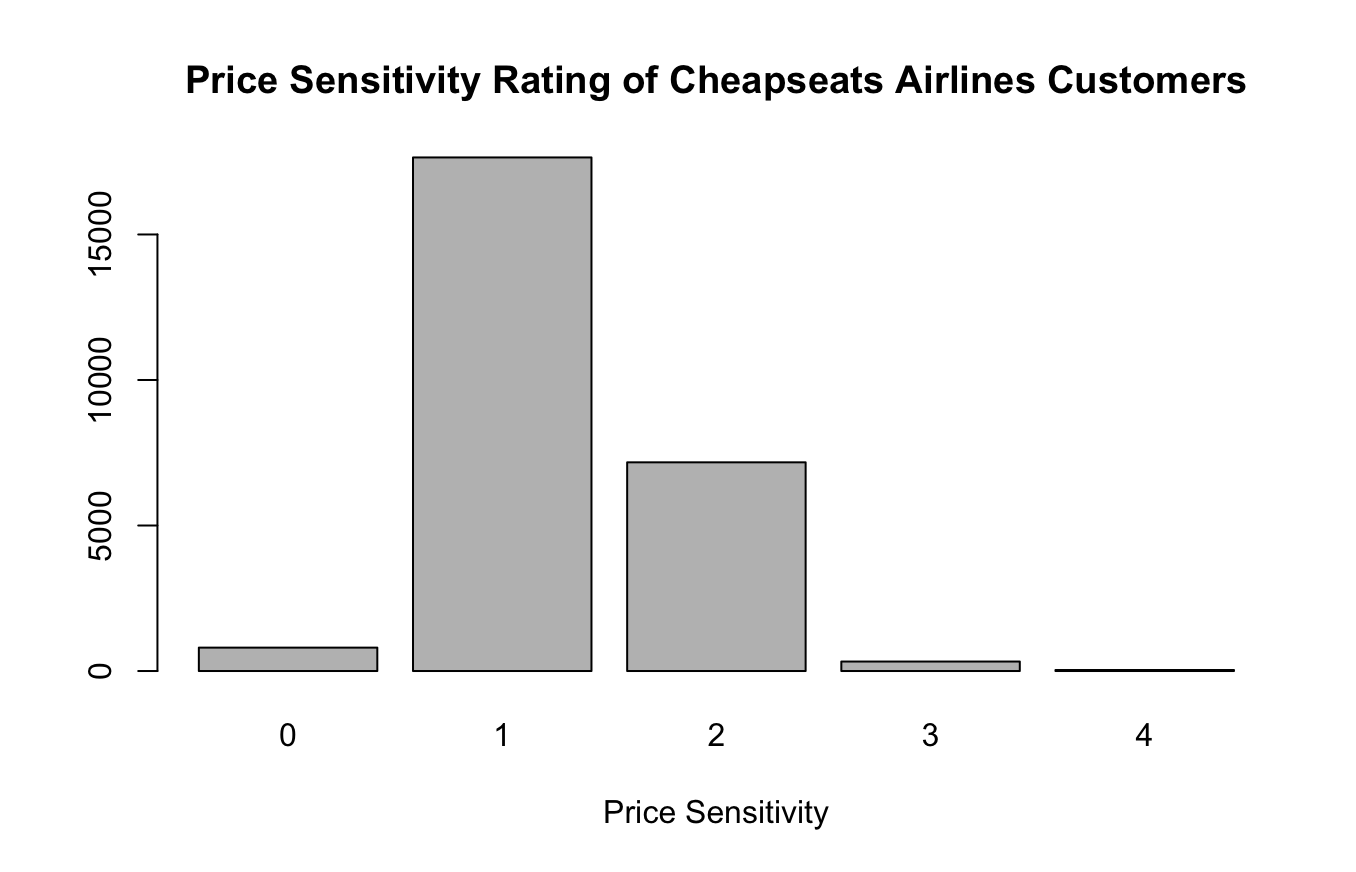
Type of Travel

We can see from the data of Cheapseats Airlines Customers, that the vast majority of our customers are flying with us for business travel. The smallest percentage of our customers are using mileage tickets to fly with Cheapseats Airlines.



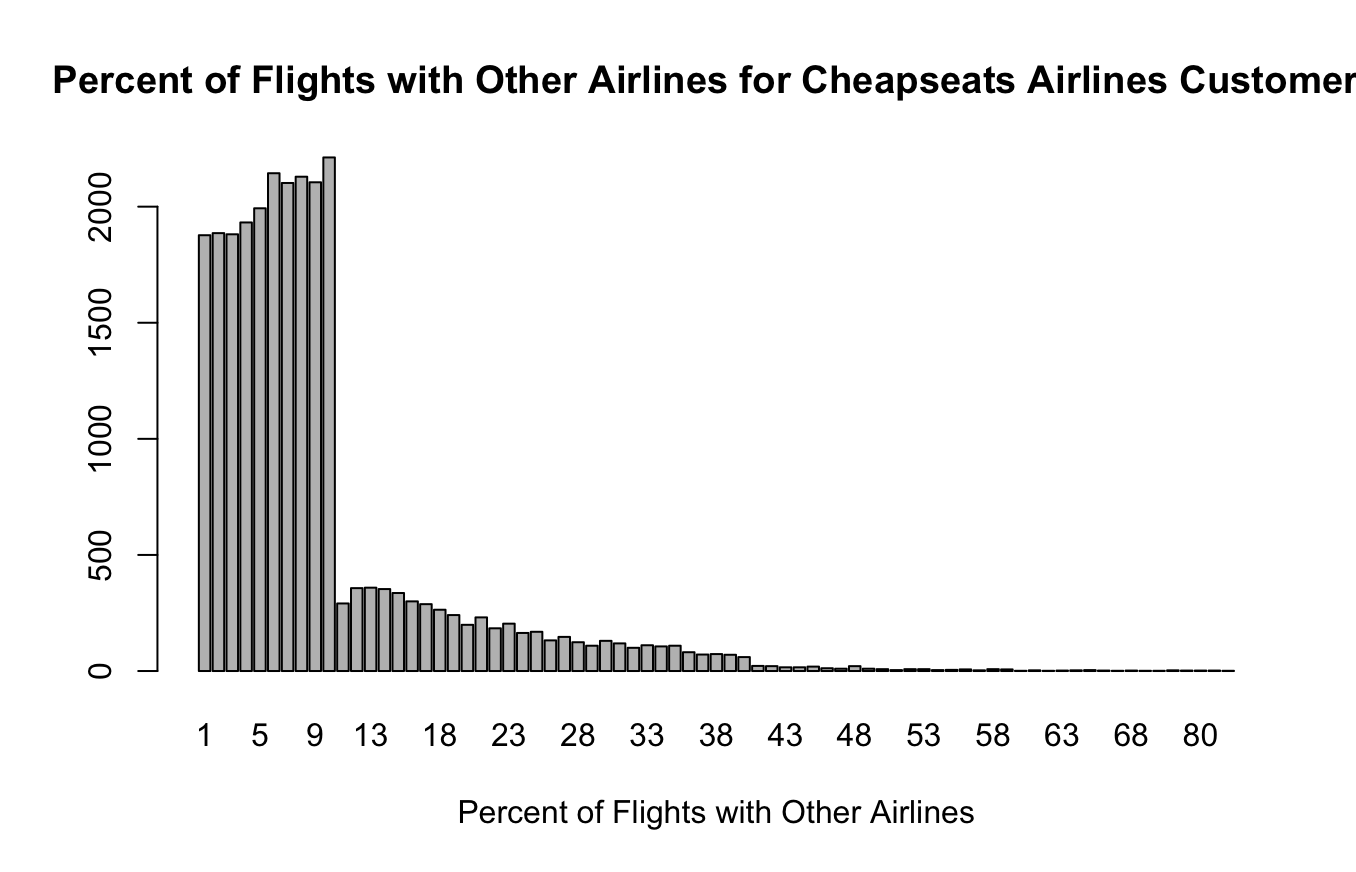
Price Sensitivity Range

We can see from the data of Cheapseats Airlines Customers, our customers are very price sensitive when booking our flights.



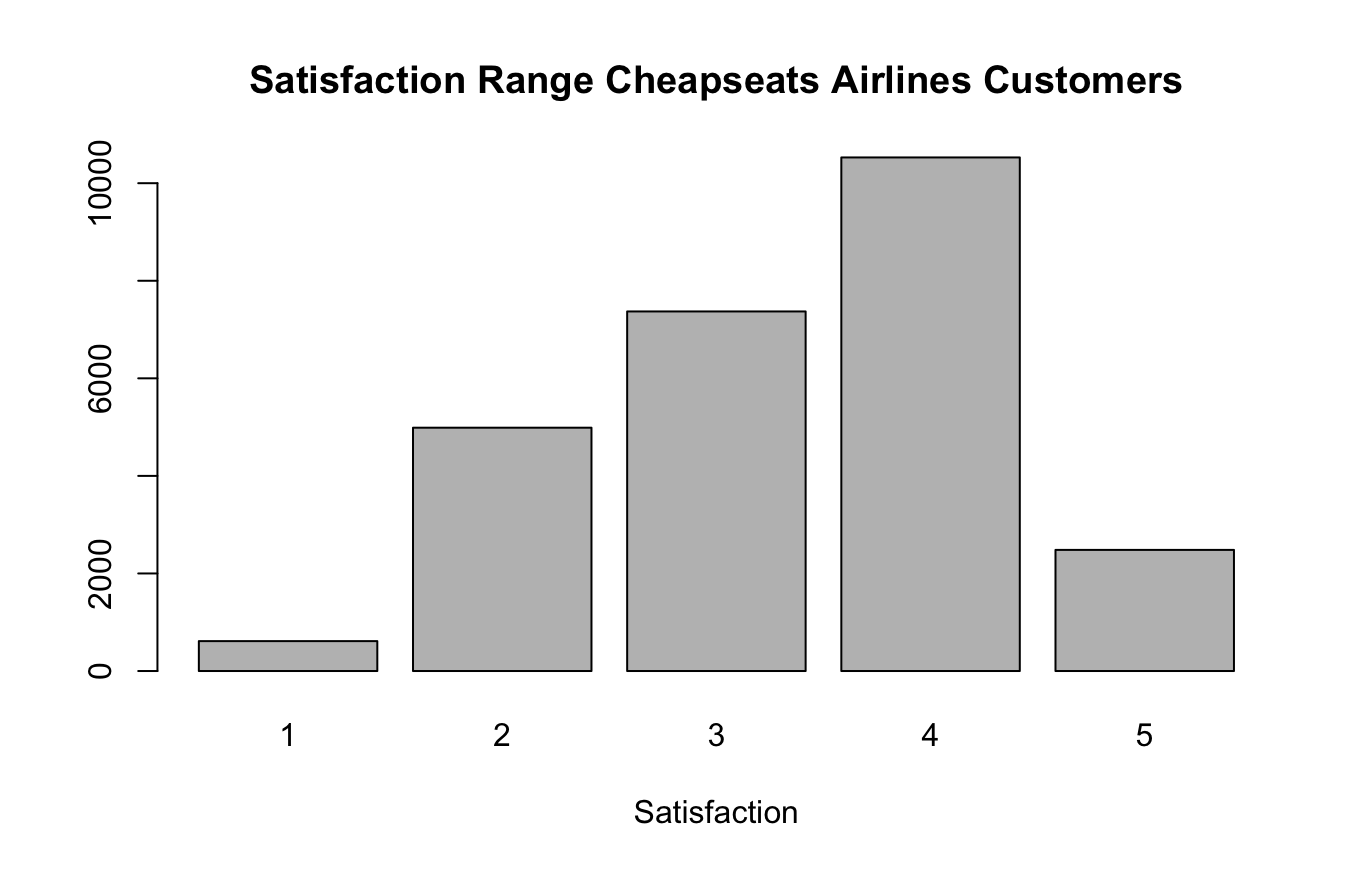
Percent of Flights with Other Airlines

We can see from the data of Cheapseats Airlines Customers that our customers not flying with our competitors, with the vast majority of our customers only choosing other airlines between 1 and 10% of the time.



Overall Satisfaction

We can see from the data of Cheapseats Airlines Customers, our customers are rating their satisfaction as a 4. A very small number of our customers are giving us a low rating of a 1 or a 2. Conversely, our customers are also not giving us a rating of 5, which means there are improvements that can be made to increase customer satisfaction. Overall, our mean customer satisfaction rating was 3.357157.



*Recommendations*

1. The delay of flights is one cause of lower satisfaction ratings. Therefore, our recommendation would either be to hire new personnel to address delayed flight issues or build in the flight delay times into the flight schedule. If there is a possibility of making up the delayed time in the air, then arrival times could be pushed back to accommodate delays.
2. As business travelers in their mid-forties are the relative target demographic, advertising and marketing campaigns should target this demographic. Our data suggests that Cheapseats Airlines customers are relatively loyal to our brand, we recommend new ways of rewarding loyal customers. This could include regular upgrades to our Business or Economy Plus class, which would also familiarize our customers with a higher class and this in turn may encourage them to book Business class or Economy Plus in the future.
3. Many of our customers are not utilizing the Business class or Economy Plus class of travel. This means that we need to review the price point of tickets for those classes. Also, as mentioned above, rewarding our loyal customers with a higher class of travel may incentivize them to book Business class or Economy Plus in the future.
4. For the cities that occurred more frequently in the cancelled flights than they did in uncancelled flights, perhaps invest in technology that could predict whether a cancelation will be needed and reroute flights to avoid that airport or make other changes that could prevent the cancellation.
5. As a customer is incredibly likely to spend money on food, offering coupons for in airport dining, or offering more inflight complementary meals could influence customer’s decision to fly with a certain airline.
6. Increase incentives for loyalty cards and rewards to help customers utilize them more.

APPENDIX

*Code*

library("readxl")

library('ggplot2')

library('dplyr')

#change this so that you're set to the directory where your datafile is

setwd('./Intro\_To\_Datascience')

data <- read.csv('Survey\_Cleaned.csv')

#check for any NA's in the data

sum(is.na(data))

#Section 1 – Lizzy

#take a preliminary look at our satisfaction summary

summary(data[,c('Satisfaction')])

#create SimpleSatisfaction col

data$SimpleSatisfaction <- data$Satisfaction

data$SimpleSatisfaction <- ifelse(between( data$SimpleSatisfaction, 0, 2), 1, data$SimpleSatisfaction)

data$SimpleSatisfaction <- ifelse(between( data$SimpleSatisfaction, 2.1, 3.9), 2, data$SimpleSatisfaction)

data$SimpleSatisfaction <- ifelse(between( data$SimpleSatisfaction, 4, 5), 3, data$SimpleSatisfaction)

#compare Satisfaction and SimpleSatisfaction

summary(data[,c('Satisfaction', 'SimpleSatisfaction')])

#seperate High, Medium, and Low Satisfaction

highSatisfaction <- data[data$SimpleSatisfaction==3,]

mediumSatisfaction <- data[data$SimpleSatisfaction==2,]

lowSatisfaction <- data[data$SimpleSatisfaction==1,]

#Make sure no flights missing

nrow(data) == nrow(highSatisfaction) + nrow(mediumSatisfaction) + nrow(lowSatisfaction)

#QUESTION 1: Cancelled Flights

#get counts of cancelled vs uncancelled flights

data %>%

group\_by(Flight.cancelled) %>%

tally()

#seperate cancelled and uncancelled flights

cancelledFlights <- data[data$Flight.cancelled=='Yes',]

uncancelledFlights <- data[data$Flight.cancelled=='No',]

#Make sure no flights have been lost

length(data$Satisfaction) == length(cancelledFlights$Satisfaction) + length(uncancalledFlights$Satisfaction)

#generate means

mean(cancelledFlights$SimpleSatisfaction)

mean(uncancelledFlights$SimpleSatisfaction)

#generate medians

median(cancelledFlights$SimpleSatisfaction)

median(uncancelledFlights$SimpleSatisfaction)

#histograms

hist(uncancelledFlights$SimpleSatisfaction,

main = 'Satisfaction in Uncancelled Flights', breaks = seq(from=0.5, to=3.5, by=1))

hist(cancelledFlights$SimpleSatisfaction,

main = 'Satisfaction in Cancelled Flights', breaks = seq(from=0.5, to=3.5, by=1))

#get count of cities with most cancelled flights

cancelled\_city\_count <- as.data.frame(table(cancelledFlights$Origin.City))

cancelled\_city\_count$ratio = cancelled\_city\_count$Freq/nrow(cancelled\_city\_count)

attach(cancelled\_city\_count)

head(cancelled\_city\_count[order(-Freq),], 10)

#get count of cities with uncancelled flights

uncancelled\_city\_count <- as.data.frame(table(uncancelledFlights$Origin.City))

uncancelled\_city\_count$ratio = uncancelled\_city\_count$Freq/nrow(uncancelled\_city\_count)

attach(uncancelled\_city\_count)

head(uncancelled\_city\_count[order(-Freq),], 10)

#QUESTION 6: Flight Time

#summary of flight time and flight distance

summary(data[c('Flight.time.in.minutes', 'Flight.Distance')])

#scatter plots

chart <- ggplot(data, aes(x=Flight.time.in.minutes, y=SimpleSatisfaction)) + geom\_point()

chart

chart <- ggplot(data, aes(x=Flight.Distance, y=SimpleSatisfaction)) + geom\_point()

chart

#boxplots

ggplot(data, aes(x=SimpleSatisfaction, group=SimpleSatisfaction, y=Flight.time.in.minutes)) + geom\_boxplot()

ggplot(data, aes(x=SimpleSatisfaction, group=SimpleSatisfaction, y=Flight.Distance)) + geom\_boxplot()

#QUESTION 7: In-Airport Behavior

#summary of eating and spending

summary(data[,c('Shopping.Amount.at.Airport', "Eating.and.Drinking.at.Airport")])

#comparative dataframe

mean\_spending <- tapply(data$Shopping.Amount.at.Airport, data$SimpleSatisfaction, mean)

mean\_eating <- tapply(data$Eating.and.Drinking.at.Airport, data$SimpleSatisfaction, mean)

median\_spending <- tapply(data$Shopping.Amount.at.Airport, data$SimpleSatisfaction, median)

median\_eating <- tapply(data$Eating.and.Drinking.at.Airport, data$SimpleSatisfaction, median)

spending\_eating <- data.frame(mean\_spending, mean\_eating, median\_spending, median\_eating)

spending\_eating

#boxplots

ggplot(data, aes(x=SimpleSatisfaction, group=SimpleSatisfaction, y=Shopping.Amount.at.Airport)) + geom\_boxplot()

ggplot(data, aes(x=SimpleSatisfaction, group=SimpleSatisfaction, y=Eating.and.Drinking.at.Airport)) + geom\_boxplot()

#Section 2 – Emily

OC <- tapply(data$Satisfaction, data$'Origin City', mean)

which.max(OC)

max(OC)

which.min(OC)

min(OC)

DC <- tapply(data$Satisfaction, data$`Destination City`, mean)

which.max(DC)

max(DC)

which.min(DC)

min(DC)

Class <- data[ , c("Satisfaction", "Class", "Airline Name", "Airline Code")]

Class

hist(Class$Satisfaction)

ggplot(data=Class, aes(x = Class, y = Satisfaction)) + geom\_bin2d()

ClassImpact <- lm(Satisfaction ~ Class, data = Class)

summary(ClassImpact)

plot(data$Satisfaction, data$`No. of other Loyalty Cards`)

plot(data$Satisfaction, data$`% of Flight with other Airlines`)

LM <- lm(data$Satisfaction ~ data$'No. of other Loyalty Cards', data=data)

summary(LM)

LM2 <- lm(data$Satisfaction ~ data$`% of Flight with other Airlines`, data=data)

summary(LM2)

#Section 3 - Jessica

#profile of cheapseats customers

#created new csv file to ensure we are only working with cheapseats airline customers.

fname <- file.choose()

airlinecheap <- read.csv(file=fname

, header=TRUE

, stringsAsFactors = F)

str(airlinecheap)

attach(airlinecheap)

#profile of age

df = airlinecheap

table(df$Age)

cheaptypes<-table(df$Age)

barplot(cheaptypes, main="Age Range of Cheapseats Airlines Customers", xlab="Age")

#profile of gender

df = airlinecheap

table(df$Gender)

cheaptypes<-table(df$Gender)

barplot(cheaptypes, main="Gender of Cheapseats Airlines Customers", xlab="Gender")

#profile for type of traveldf = airlinecheap

table(df$Type.of.Travel)

cheaptypes<-table(df$Type.of.Travel)

barplot(cheaptypes, main="Most Common Type of Travel for Cheapseats", xlab="Types of Travel")

#All Airlines Type of Travel

df = airline #used the full airlines data set, rather than just the cheapseats customers

table(df$Type.of.Travel)

cheaptypes<-table(df$Type.of.Travel)

barplot(cheaptypes, main="Most Common Type of Travel for All Airlines", xlab="Types of Travel")

#profile for percent of flights with other airlines

df = airlinecheap

table(df$X..of.Flight.with.other.Airlines)

cheaptypes<-table(df$X..of.Flight.with.other.Airlines)

barplot(cheaptypes, main="Percent of Flights with Other Airlines for Cheapseats Airlines Customers", xlab="Percent of Flights with Other Airlines")

#profile of satisfaction

df = airlinecheap

table(df$Satisfaction)

cheaptypes<-table(df$Satisfaction)

barplot(cheaptypes, main="Satisfaction Range Cheapseats Airlines Customers", xlab="Satisfaction")

#profile of price sensitivity

df = airlinecheap

table(df$Price.Sensitivity)

cheaptypes<-table(df$Price.Sensitivity)

barplot(cheaptypes, main="Price Sensitivity Rating of Cheapseats Airlines Customers", xlab="Price Sensitivity")

#business question #5

str(airlinecheap)

attach(airlinecheap)

df = airlinecheap

table(df$Type.of.Travel)

cheaptypes<-table(df$Type.of.Travel)

barplot(cheaptypes, main="Most Common Type of Travel for Cheapseats", xlab="Types of Travel")

df = airline

alltypes<-table(df$Type.of.Travel)

barplot(alltypes, main="Most Common Type of Travel for All Airlines", xlab="Types of Travel")

#business questions #7

#linear model – Satisfaction versus Departure Delay in Minutes

SatDelay<-lm(Departure.Delay.in.Minutes ~ Satisfaction, data=airline)

summary(SatDelay)

Call:

lm(formula = Departure.Delay.in.Minutes ~ Satisfaction, data = airline)

Residuals:

Min 1Q Median 3Q Max

-21.27 -15.71 -12.93 -2.49 1573.51

Coefficients:

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

(Intercept) 24.047 0.383 62.78 <2e-16 \*\*\*

Satisfaction -2.780 0.109 -25.51 <2e-16 \*\*\*

---

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

Residual standard error: 37.84 on 129544 degrees of freedom

Multiple R-squared: 0.004999, Adjusted R-squared: 0.004991

F-statistic: 650.8 on 1 and 129544 DF, p-value: < 2.2e-16

scatter.smooth(x=airline$Departure.Delay.in.Minutes, y=airline$Satisfaction, main="Delay Time ~ Satisfaction")

#satisfaction by origin state

dfStates<- airlinecheap

dfStates$Origin.State <-tolower(dfStates$Origin.State)

map.popColor <- ggplot(dfStates, aes(map\_id=Origin.State))

map.popColor <- map.popColor + geom\_map(map=us,aes(fill=Satisfaction))

map.popColor <-map.popColor + expand\_limits(x=us$long, y=us$lat)

map.popColor <- map.popColor + coord\_map() + ggtitle("Satisfaction by Origin State")

map.popColor

#satisfaction by destination state

dfStates<- airlinecheap

dfStates$Destination.State <-tolower(dfStates$Destination.State)

map.popColor <- ggplot(dfStates, aes(map\_id=Destination.State))

map.popColor <- map.popColor + geom\_map(map=us,aes(fill=Satisfaction))

map.popColor <-map.popColor + expand\_limits(x=us$long, y=us$lat)

map.popColor <- map.popColor + coord\_map() + ggtitle("Satisfaction by Destination State")

map.popColor

#type of travel by origin state

dfStates<- airlinecheap

dfStates$Origin.State <-tolower(dfStates$Origin.State)

map.popColor <- ggplot(dfStates, aes(map\_id=Origin.State))

map.popColor <- map.popColor + geom\_map(map=us,aes(fill=Type.of.Travel))

map.popColor <-map.popColor + expand\_limits(x=us$long, y=us$lat)

map.popColor <- map.popColor + coord\_map() + ggtitle("Type of Travel by Origin State")

map.popColor

#type of travel by destination state

dfStates<- airlinecheap

dfStates$Destination.State <-tolower(dfStates$Destination.State)

map.popColor <- ggplot(dfStates, aes(map\_id=Destination.State))

map.popColor <- map.popColor + geom\_map(map=us,aes(fill=Type.of.Travel))

map.popColor <-map.popColor + expand\_limits(x=us$long, y=us$lat)

map.popColor <- map.popColor + coord\_map() + ggtitle("Type of Travel by Destination State")

map.popColor