Cheapseats Airlines

Prepared for [Cheapseats Airlines]

Created by [Red Panda Consultants]

IST687_M002_Group#2

Daniel Trevino

Chiau Yin Yang

Soham Adhikari

Maithili Shah

Ankita Singh

RedPanda Consultants specialize in providing services to aid businesses at all sizes to understand their data and provide insightful analysis. RedPanda consultants additionally adheres to the parent company data request and does not retain company data post the analysis and reported outcome, unless otherwise directed by the business owner.

1. Introduction

Cheapseats Airlines maintains a set of survey data related to how customers feel, and multiple variables associated to their domestic flights. Cheapseats Airlines stood out as one of the larger airlines with 26k observations reported. Due to the law of large numbers, we believe that the dataset provided on which Cheapseats analysis is performed is statistically sufficient for RedPanda consultants to perform their analysis.

2. Understanding the Data

The initial data set contained multiple airlines data and values that proved a hindrance when performing analysis. The dataset contained various data types, including integer, factor, and character that often confused RStudio. In order to generate a dataset that is most usable, we generated the following set of assumptions.

(1) Null values were few in number compared to the overall data, therefore the team extracted all survey data observations that contained null values. In addition, most variables that were used while analyzing are with no null values.
(2) The data proved initially to be to large and thus the team agreed to only focus on Cheapseats Airlines as the total sample population of Cheapseats closely models the overall data set.
(3) Where values fell within the metadata, team members used the integer values within their analysis, this provides a more complete dataset.
(4) For the purpose of this proposal, any reference to West Airlines, being a subsidiary of Cheapseats, is used as a reference point.
(5) To effectively create an increase in the overall satisfaction, a high satisfaction score is benchmarked. The benchmark for the satisfaction score, for purposes of the assessment, will be greater than or equal to 4. Thus, customers with a survey rating of 4 or higher are classified as giving a high rating. Ratings of 3 are average and a rating of 1 or 2 is below average.

The survey data was initially analyzed as a whole data set. This generated a good understanding of the baseline for all the airlines. After a baseline was understood, the team focused on one Airline to perform the analysis over again. This was decided upon as it is easier to help one airline than the conglomerate. After the decision was agreed upon, the survey data was cleaned and a subset of data was created just for Cheapseat Airlines. The following report covers the steps team Red Panda followed from understanding the data to providing valid and usable recommendations for Cheapseat Airlines.

3. Scope & Business Questions

Scope:

(1) Improve overall customer satisfaction
(2) Generate actionable insight for Cheapseats Airlines.

Problem Statement:

Given a dataset of over 130,000 survey responses, what actionable insights can be understood to help Cheapseats Airlines to improve their customer satisfaction

Purpose:

Team Red Panda will analyze a dataset consisting of 130,000 responses within the course of 8 weeks in order to provide actionable insight for Cheapseat Airlines to improve overall customer satisfaction.

Business Questions:

1. **Business Rule # 1 -** All satisfied customers will rate the airlines with a rating of 4 or higher (Average - Very High) while all unsatisfied customers will give a rating less than or equal to 3.
2. **Business Rule # 2 -** Incentives shall be distributed in the event of delayed flights.
3. **Business Rule # 3 -** Ensure senior citizen passengers are properly cared for while flying our airline.
4. **Business Rule # 4 -** Fill as many seats per each flight as possible, and this includes complimentary upgrades when seats are not filled and available.
5. **Business Rule # 5 -** Any flight more than 5 min delayed upon departure or arrival is considered "delayed".
6. **Business Rule # 6 -** Every customer should sign up for the travel type status program (blue, silver, gold, platinum).

Data Questions:

1. Why CheapSeats Airlines?
2. Is age a factor in lower ratings or higher ratings within the data set?
3. Does a person's travel type affect on the rating they will give during the survey?
4. How many flights are there per months / day? Do delayed flights have a significant effect on customer ratings?
5. Is there a gender bias in customer satisfaction ratings?
6. What class is most often traveled and does the class have an effect on the customer rating?
7. Does the airport location of the airport have an impact overall customer satisfaction? Which cities were traveled to most on any given day and is there a pattern of lower ratings based on the most

visited city?

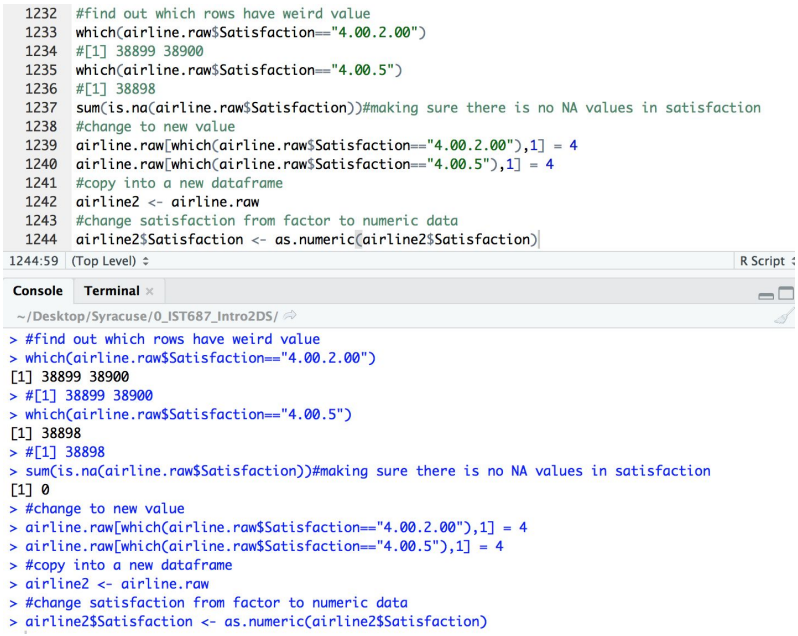
8. What is the distribution of satisfaction ratings per rating?
9. Does the number of times a person flys impact their overall satisfaction rating?
10. Which travel type gives the highest satisfaction rating and how much does travelType affect the overall satisfaction score?

4. Data Munging

How did we transform the data from initial receipt to what we have with Cheapseats?

Upon original receipt of the dataset, which contained 129889 observations and 28 variables, the team began to perform some minor data cleaning. A primary issue with the Satisfaction variable stood out. There were values not conducive for analysis which included a float within a float. The index numbers 38898, 38899, 38900, were found to have values, respectively 4.00.2.00 and 4.00.5. These values were required to be cleaned. For the purpose of the survey, the values were reset to 4, 2, and 4. This decision was made via two factors, the satisfaction scores in the data set surrounding the index values were equal, and changing 3 values within the entire dataset has a minimal effect on the overall dataset (00.2% ~ 0%).

After cleaning the erroneous numbers from the satisfaction variable, the dataset owner did not provide any guidance on why they were listed as "factors." The satisfaction variable was therefore changed to a classification of integer with the as.numeric() function. Looking at the dataset also revealed that there were scores that were floats. These score were left in the overall dataset and found to belong to West Airlines. They were left in the dataset as their number fell between 1 and 5, which matches the metadata requirements. After the satisfaction variable was cleaned, a new dataset was distributed throughout the team.



From the new survey dataset, multiple issues arrived as we dealt with each variable. For more information on these variables, see section 5B on Descriptive Statistics. Some of the more generic issues,

were transforming variables between factors, characters, or integers. Several of the variables also contained NULL values for which the team agreed to drop (na.rm = TRUE) the values during analysis as their total number did not affect the overall analysis. To ensure the data was properly cleaned and everyone was using the same initial dataset, the team utilized the Syracuse Google service. A Google Drives was established and used to exchange dataset and store reports and documentation. As data was transformed and distributed, the Google Drive was the primary resource for all large files.

To read the data into R, the team used two primary techniques. The first technique was via the library(readxl) and the second was via the library(rjsonio). Once the dataset was read into R, the library(dplyr) package was extremely useful for scraping whitespace, filtering, and understanding the data. Issues encountered while munging through the data with dplyr consisted of using the entire dataset to perform analysis with prediction models and how to filter results to perform descriptive statistics. The %>% (piping) technique proved useful for filtering specific sets of information and manipulating the dataset into smaller datasets. An example of piping the Cheapseats Airline into a new dataframe is seen below:

```
cheapseats <- survey %>%
  select("satisfaction", "airlineStatus", "age", "gender", "priceSensitivity", "firstYearFlight", "numberOfFlights", "percentOfOtherFlight", "travelType", "loyaltyCardsNum",
         "shoppingAtAirport", "foodAtAirport", "class", "day", "date", "airlineCode", "airlineName", "originCity", "originState", "destCity", "destState", "scheduledDepHr",
         "scheduledDepDelay", "arriveDelayMinute", "flightCancelled", "flightTimeMinute", "flightDistance", "arriveDelay>5") %>%
  filter(airlineName == "Cheapseats Airlines Inc.")
```

```
 satisfaction    airlineStatus        age           gender      priceSensitivity firstYearFlight numberOfFlights percentOfOtherFlight           travelType
Min.   :1.000   Blue     :17864   Min.   :15.00   Female:14666   Min.   :0.000    Min.   :2003    Min.   : 0.00   Min.   :  1.000      Business travel:15996
1st Qu.:3.000   Gold     : 2123   1st Qu.:33.00   Male  :11392   1st Qu.:1.000    1st Qu.:2004    1st Qu.: 9.00   1st Qu.:  4.000      Mileage tickets: 1993
Median :4.000   Platinum:  823   Median :45.00                  Median :1.000    Median :2007    Median :17.00   Median :  7.000      Personal Travel: 8069
Mean   :3.357   Silver   : 5248   Mean   :46.18                  Mean   :1.274    Mean   :2007    Mean   :20.02   Mean   :  9.352
3rd Qu.:4.000                     3rd Qu.:59.00                  3rd Qu.:2.000    3rd Qu.:2010    3rd Qu.:29.00   3rd Qu.: 10.000
Max.   :5.000                     Max.   :85.00                  Max.   :4.000    Max.   :2012    Max.   :93.00   Max.   :110.000

 loyaltyCardsNum  shoppingAtAirport foodAtAirport        class            day              date       airlineCode                  airlineName           originCity
Min.   :0.0000   Min.   :  0.00    Min.   :  0.00   Business: 2111   Min.   : 1.00   3/19/2014:  349   WN:26058    Cheapseats Airlines Inc.:26058   Chicago, IL  : 1665
1st Qu.:0.0000   1st Qu.:  0.00    1st Qu.: 30.00   Eco     :21261   1st Qu.: 9.00   3/20/2014:  347                                             Las Vegas, NV: 1637
Median :0.0000   Median :  0.00    Median : 60.00   Eco Plus: 2686   Median :16.00   3/13/2014:  345                                             Baltimore, MD: 1383
Mean   :0.8907   Mean   : 26.45    Mean   : 67.97                    Mean   :15.83   3/17/2014:  340                                             Phoenix, AZ  : 1342
3rd Qu.:2.0000   3rd Qu.: 30.00    3rd Qu.: 90.00                    3rd Qu.:23.00   3/21/2014:  337                                             Denver, CO   : 1227
Max.   :8.0000   Max.   :745.00    Max.   :765.00                    Max.   :31.00   2/28/2014:  336                                             Houston, TX  : 1202
                                                                                     (Other)  :24004                                             (Other)      :17602
    originState            destCity         destState     scheduledDepHr  scheduledDepDelay arriveDelayMinute flightCancelled flightTimeMinute flightDistance
California: 4687   Las Vegas, NV: 1686   California: 4616   Min.   : 5.00   0      : 9910     0      :12077     No :25742       NA     :  389    Min.   : 148.0
Texas     : 3476   Chicago, IL  : 1632   Texas     : 3393   1st Qu.: 9.00   1      : 1008     1      :  625     Yes:  316       53     :  369    1st Qu.: 365.0
Florida   : 2416   Phoenix, AZ  : 1326   Florida   : 2441   Median :13.00   2      :  746     2      :  596                     58     :  340    Median : 587.0
Nevada    : 1802   Baltimore, MD: 1323   Nevada    : 1834   Mean   :13.05   3      :  629     3      :  554                     62     :  340    Mean   : 709.5
Illinois  : 1665   Denver, CO   : 1312   Illinois  : 1632   3rd Qu.:17.00   4      :  573     4      :  546                     55     :  339    3rd Qu.: 957.0
Arizona   : 1470   Houston, TX  : 1162   Arizona   : 1442   Max.   :22.00   6      :  538     6      :  482                     63     :  338    Max.   :2329.0
(Other)   :10542   (Other)      :17617   (Other)   :10700                   (Other):12654     (Other):11178                     (Other):23943
arriveDelay.5
no :15263
yes:10795
```

The final munging issue is using the whole survey dataset for running statistical models. Performing Association Rules (arules) and Kernel Support Vector Machines (ksvm) on the cleaned original dataset proved time consuming. Arules ran for several hours while ksvm lasted almost 18 hours on a robust computer. This munging around lead the team to two conclusions, (1) a smaller data set is required and (2) using a smaller dataset with a mean close the original data set is a good sample of the population. In choosing a smaller subset of sample data, the team collectively decided to choose the largest Airline to support. The following table and graph show the distribution of airlines to the survey satisfaction score.

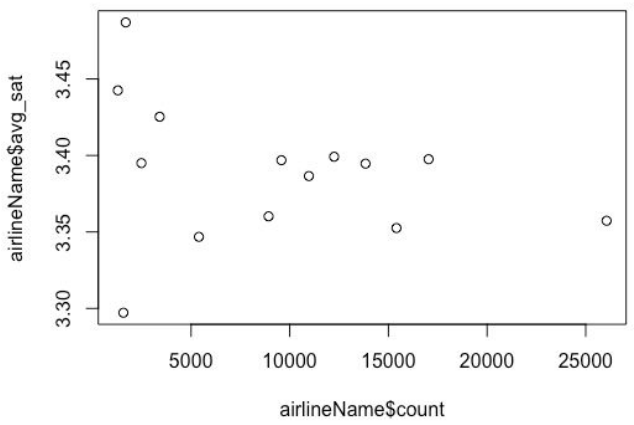
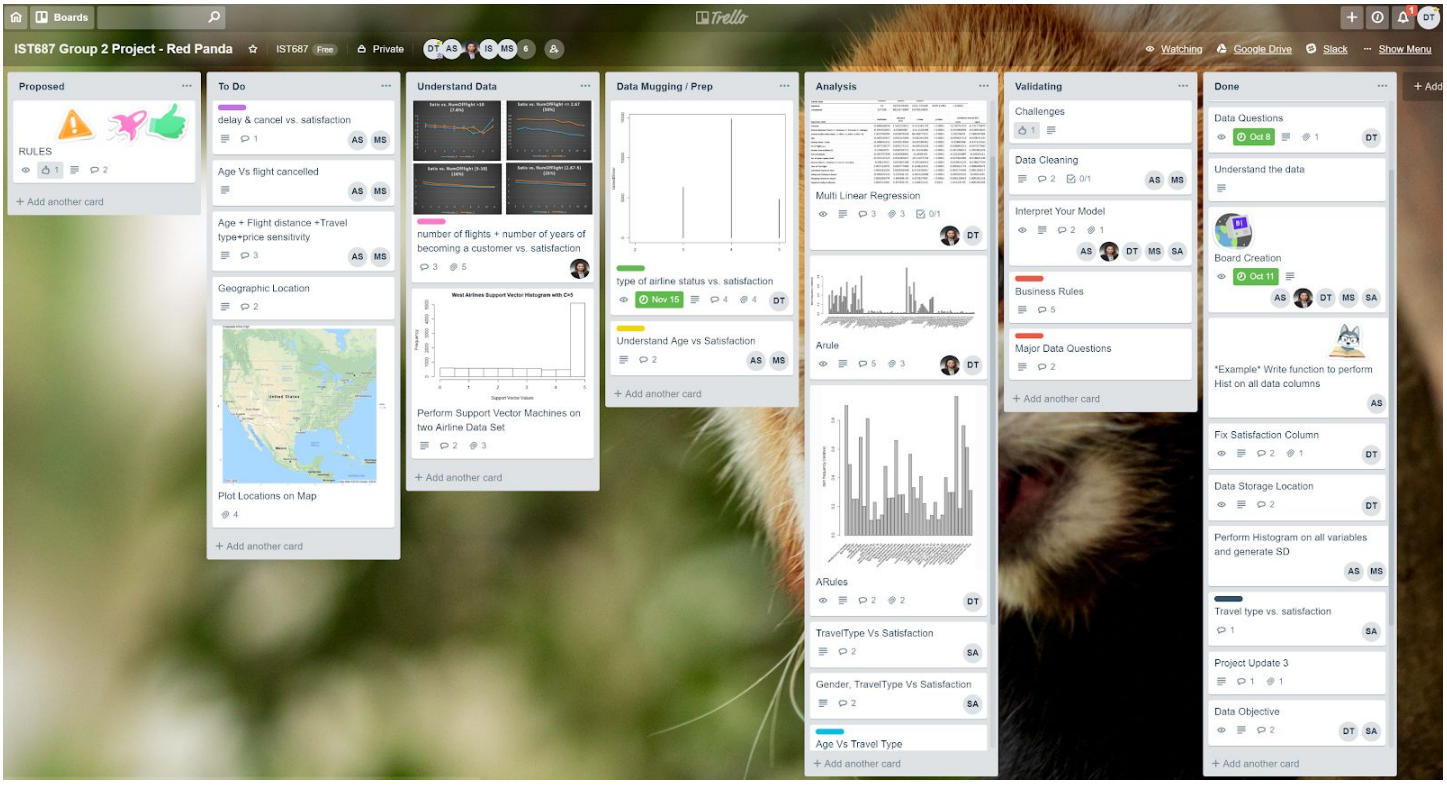


Figure 1: Airline Count vs Satisfaction

	airlineName	count	avg_sat	percent
2	GoingNorth Airlines Inc.	1568	3.297194	1.2071846
6	OnlyJets Airlines Inc.	5395	3.346803	4.1535465
12	FlyFast Airways Inc.	15407	3.352567	11.8616665
14	Cheapseats Airlines Inc.	26058	3.357318	20.0617450
7	EnjoyFlying Air Services	8927	3.360199	6.8727914
9	Oursin Airlines Inc.	10968	3.386534	8.4441331
11	Northwest Business Airlines Inc.	13840	3.394653	10.6552518
4	FlyHere Airways	2481	3.395002	1.9100925
8	Southeast Airlines Co.	9577	3.396888	7.3732187
13	Sigma Airlines Inc.	17037	3.397547	13.1165842
10	Paul Smith Airlines Inc.	12248	3.399167	9.4295899
5	FlyToSun Airlines Inc.	3407	3.425301	2.6230089
1	Cool&Young Airlines Inc.	1288	3.442547	0.9916159
3	West Airways Inc.	1688	3.486967	1.2995712

Figure 2: Airline Count vs Satisfaction

Once again, after pulling the sample dataset (Cheapseats Airlines Inc.) the data cleaning process started over again. To assist with managing the data and processes for cleaning, analysis, and assigning roles, Trello is used. The following is a snapshot of the Trello board team RedPanda used while working on the project. Updating and documenting analysis is a systemic problem across all disciplines of work, such is the same with this management tool where team members labously updated the Trello boards.



5. Analysis

Ethics Statement: Team Red Panda shall not disclose to any third party any details regarding the Client's business, including, without limitation any information regarding any of the Client's customer information, business plans, or price points (the Confidential Information), (ii) make copies of any Confidential Information or any content based on the concepts contained within the Confidential Information for personal use or for distribution unless requested to do so by the Client, or (iii) use Confidential Information other than solely for the benefit of the Client.

To best answer and demonstrate the value of Team Red Panda's analysis, understanding the geography of Cheapseats Airlines is useful. Maps were created to help visualize the data locations represented within the data set. Likewise, attempts were made to plot the lines of travel between the origin and destination cities. Furthermore, to provide a thorough understanding the data, the following two levels will both highlight the work performed and provide analysis on the return of the data. The first level will describe several of the independent variables and how they interact by themselves and with the dependent variable Satisfaction. The second level will cover three models used in R to help predict which variables have the most significance and weight against Satisfaction.

Geography and Mapping:

The initial Data set, as provided, is not very friendly for creating map data. The longitude and latitude were not contained within the data set and therefore are required for plotting points. Additionally, the city names of the origin and destination locations were not clean. Several of the cities where conjoined to provide a geographic regional airport. For example, "Dallas / Arlington, TX". The first step in turning the data into something usable is to clean the city names. To do this, a text replacement function is used, which is part of the base R package. The function used is:

```
gsub(pattern, replacement, x, ignore.case = FALSE, perl = FALSE, fixed = FALSE, useBytes = FALSE)
```

The technique performed by Team Red Panda is to just create a seperate gsub for each multi city observation. The following is an example of the gsub() pattern used. This pattern replaces the multicity with a single city and places back into the original dataset.

```
cheapseats$originCity <- gsub("West Palm Beach/Palm Beach","Palm Beach",cheapseats$originCity)
```

After all the multicity observations are replaced with one variable, as seen below, the next step of finding the longitude and latitude using the geocode() function is used. As seen in the below table, Chicago, IL has the most origination flights and Richmond, VA has the least origination flights.

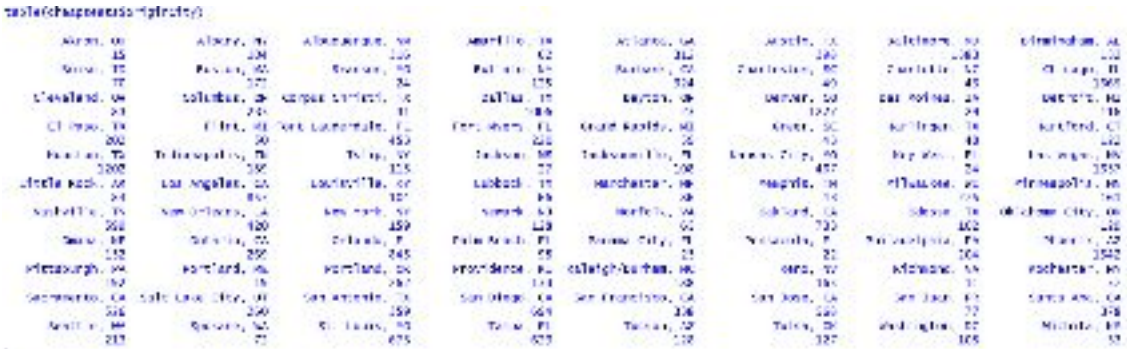


Figure 3: Origin Cities by Count

To plot the city locations, the geocode() function, which is part of the ggmap package is used. The geocode function is designed to first query Google Maps for the longitude and latitude, however, Google has blocked this functionality and request usage of a Google Maps API. This was initially a problem, thus to mitigate the issue, the "dsk" datascience tool kit is used as a parameter within the geocode function. Next, the cheapseats survey data set is passed into the geocode function to return the longitude and latitude for each city. This was initially performed with an evil For{} loop as seen below:

```
#loop through originCity
for(i in 1:nrow(cheapseats))
{
  # Print("Working...")
  result <- geocode(cheapseats$originCity[i], output = "latlona", source = "dsk")
  cheapseats$lon[i] <- as.numeric(result[1])
  cheapseats$lat[i] <- as.numeric(result[2])
  cheapseats$geoAddress[i] <- as.character(result[3])
}
```

This however, was later discovered to be mitigated with the lapply fuction.

```
cheapseatsLatLon <- lapply(cheapseats$originCity, geocode)
```

After the latitude and longitude are implemented, the map was then plotted with geom_point from the ggplot2 package. The following maps are then produced using the ggmap package and the ggplot2 package. The first map pulls the map from google maps, then plots the points on the map.

Figure 4: Cheapseats Location Map 1

The second map plots the points with qmplot on a watercolor map. The negative with qmplot is the map selection is not very wide, however it selects a map based on the parameters passed to it. Unlike get_map, you have to select what map you want to plot on.

Figure 5: Cheapseats Location Map 2

For both maps, you can see where cheapseats airlines flies. The visualization of geographic locations is useful for Team Red Panda as the map provides insightful location analysis on the the magnitude of flights per each geographic location. This data will be discovered further in the descriptive statistics section below. In order to get a good idea of how each city relates to one another, the following map is created. The map shows all cities that have a one or more connecting flights as both the origin or destination To define the map further, cities with the most connection flights have a larger point. This map is used to assist the team with identifying hubs or major Cheapseat locations that act as hubs.

Figure 6: Cheapseats Location Map 3

5B. 2 Levels of Understanding

The levels of understanding will walk through each team members processes as well as provide insightful analysis on each of the variables and models used. The end result of the levels is to assist the team in answering the data questions and providing predictive analysis to answer the problem statement listed at the beginning of this report.

5B-1. Level 1: Descriptive Statistics

The descriptive statistics walks through each of the variables the team selected. The selections were based on a process of elimination. Upon receipt of the data set, after being cleaned, the team selected which variables to develop more insightful evidence. The variable selection process centered around performing descriptive statistics. The variables were broken down by team members for the whole dataset. However, given the time and magnitude of the data, only variables analysis was performed on will follow in the report.

A. Dependent Variable:

a. Variable: Satisfaction

The satisfaction is the most important variable in the survey data. This variable is dependent upon all the customers other ratings and scores. The following table shows that Cheapseats Airline has a median satisfaction score of 4.0 but a mean(average) score of 3.357. This means there are a larger number of higher ratings but the scores for lower ratings were so low that they pulled down the average satisfaction score. Understanding the group that is pulling the satisfaction score down is an important finding within the survey.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   3.000   4.000   3.357   4.000   5.000
```

Figure 7: Cheapseats Summary Data

To visualize the distribution of scores, the following chart is used. The chart shows there is a large number of scores with rating of 4 and a low rating of scores with 1 and 2. This histogram also illustrates the summary values above.

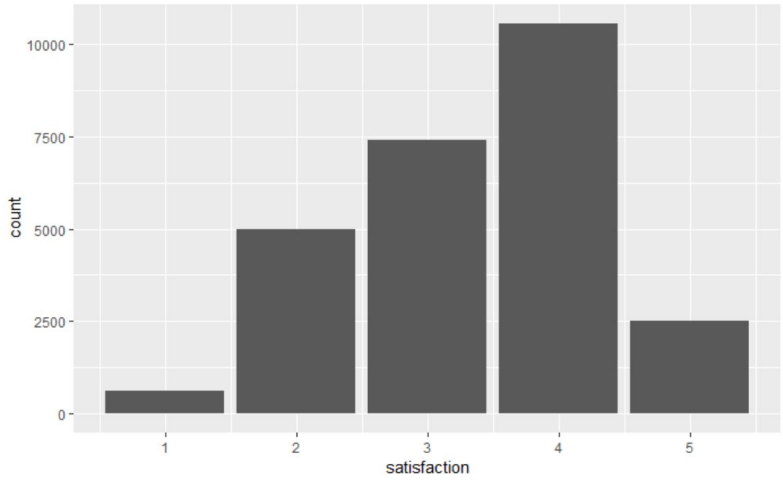


Figure 8: Histogram of Cheapseats Satisfaction

Given how much the Satisfaction score relies on independent variables, multiple independent variables will be analyzed via the descriptive statistics technique and against the satisfaction score. Not all independent variables are assessed due to the time constraint of processing each variable. Choosing the independent variables and splitting them up across the team gives the largest breath of focus for discovering interesting insights. Thus, every team member is given two independent variables, but only the first variables were analyzed with the descriptive technique. The following independent variables are therefore assessed: Airline Status, Age, Travel type, Number of Flights & First Year of Flight, Origin City & Destination City.

B. Independent Variables

a. Variable: Airline Status [Blue, Silver, Gold, Platinum]

The variable airlineStatus is classified as a “factor” in the original dataset. This classification

proved difficult to use initially. In order to get the most out the descriptive statistics process for the variable, the vector needed to be changed into "character." To do this, the as.character() is used on the airline status to convert the factor into characters. However, after performing several hours of analysis, the discovery was made to use the tapply() and ggplot2 package to plot the tables and graphs. The conversion of factors to characters is not entirely required when using tapply() or ggplot().

To understand the airline status of customers, a count of each status is provided in the bar chart. To generate further understanding of how Airline Status is related to the dependent variable, the following table is made to display the relationship.

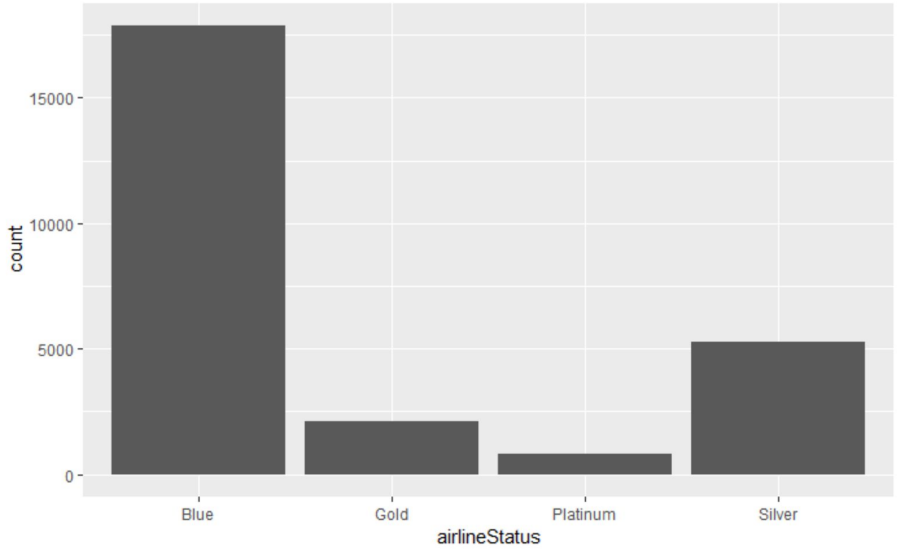


Figure 9: Bar chart of Airline Status.

There are a large number of Blue customers within Cheapseats Airlines. The number of Blue customers makes up 68.6% of the total population. Silver is 20.13%, Gold is 8.14%, and Platinum status is 3.16%. Given the count of Blue customers (17864), their satisfaction scores will have more weight on the mean and median values listed above in the satisfaction dependent variable. For further analysis, satisfaction vs airline status descriptive statistics is performed.

The following tables give a distribution of each airline status vs satisfaction.

Blue

	Var1	Freq
1	1	536
2	2	4484
3	3	5595
4	4	6627
5	5	622

Silver

	Var1	Freq
1	3	1312
2	4	2957
3	5	979

Gold

	Var1	Freq
1	2	383
2	3	399
3	4	766
4	5	575

Platinum

	Var1	Freq
1	1	80
2	2	129
3	3	90
4	4	211
5	5	313

Figure 10: Airlines Status vs Satisfaction

Assessing the tables above, most customers give a rating of 4 or higher. Given this, we are able to validate our assumption number 5 listed in section 2 of this report, that happy customers give a rating of 4 or higher. An average rating is a rating of 3 and ratings of 1 or 2 are below average. Ratings of 2 or

below make up 23% of the overall survey data for Cheapseats Airlines. The majority of these ratings are from Blue status customers (roughly 83.5%). Given the quantity of the ratings, targeting Blue customers with a rating of 2 or below will assist in enhancing the overall customer satisfaction score. Further descriptive statistics will be performed on variables within the data set for more analysis and insight.

b. Variable: Age

The variable, Age, is an easy one, there is no data cleaning work needs to be done, no NAs and is integer. However, to be analyze further, we grouped them into a period of 10 years, so we could investigate further with characteristics of each group of people. We also created a bar chart using ggplot() and fill with average satisfaction ratings. In the below chart, it is obvious that the majority of the customer are in their mid age (30-59), which accounts for 57% of the population. Besides, younger (10-19) and elderly (60 and up) customers have average satisfaction rating of below 3.

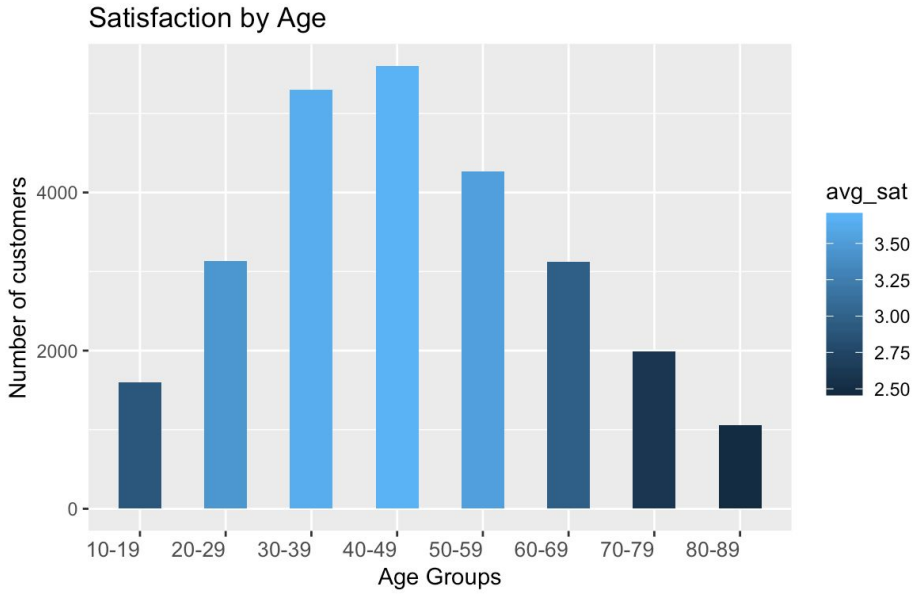


Figure 11: Age Group vs Satisfaction rating

c. Variable: Travel Type (Business, Mileage tickets and Personal)

The variable travelType is classified as a "character" in the original dataset. This classification was convenient to use as in order to get the most out of the descriptive statistics process for the variable, the vector is supposed to be of the type "character". After analyzing the data frame based on travel type ggplot2 tapply(), dlpyr() packages were used to clean the data frame, plot the graphs and tables and to continue further analysis.

First, to analyse and observe the traveltype closely for more understanding, the first step is to prepare a histogram that provides a count of each travel type. For further understanding and analysis, the following histogram is created to display the relationship.

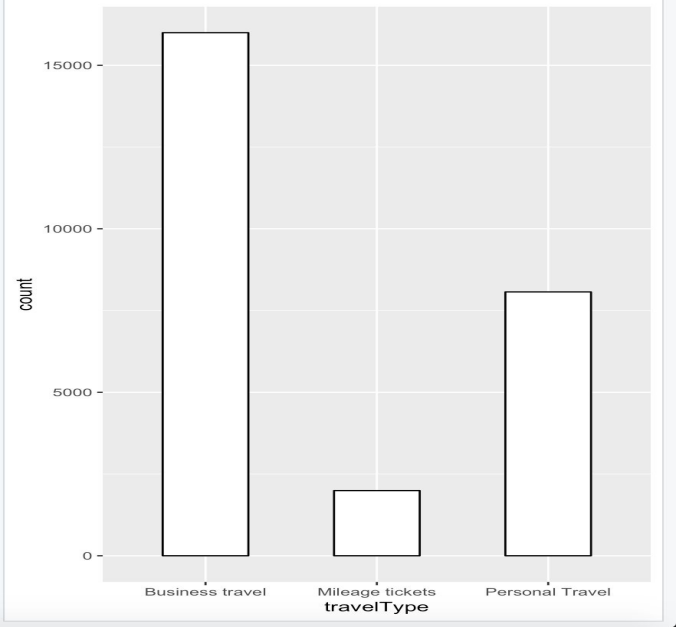


Figure 12: Bar chart of Travel Type

After analysing the bar chart and histogram above, we conclude that the number of business type travelers in Cheapseats is the largest of all the three different types. The number of Business Travelers makes up to 61.3% of the total flyers population. Mileage is 7.6% and Personal is about 31% of the total population. Given the count of Business customers (15996), their satisfaction scores will have more weight on the mean and median values in the satisfaction dependent variable. For further analysis, satisfaction vs travelType descriptive statistics is performed.

The following bar charts provide a distribution of each travel type vs satisfaction

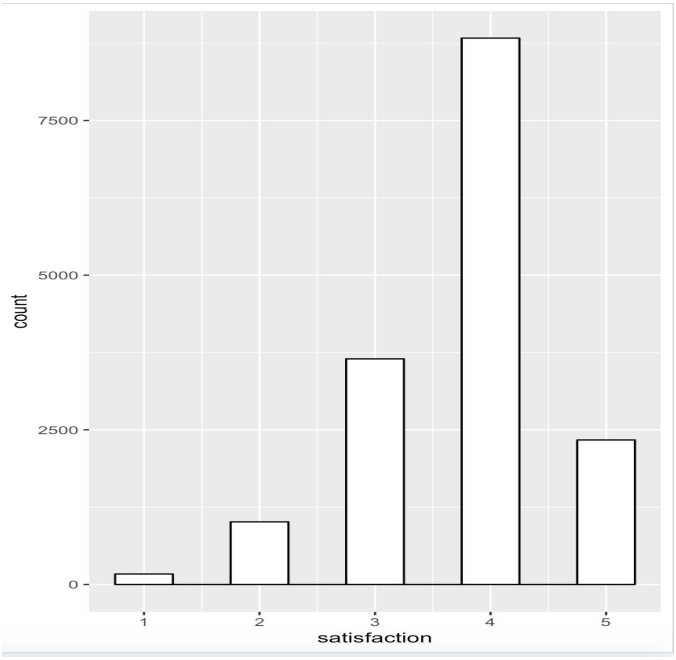


Figure 13: Bar chart of Business type traveler vs Satisfaction

Later in the analysis we categorize high ratings and low ratings. According to the business rules, any satisfaction score that is 4 and above is considered as high rating and any score that is lower than 3 is

considered as low rating. Average is 3 and we can assume that the average ratings would not give us a significant insight of the dependencies on the satisfaction score. In the above chart, we observe that out of the 15996 business type flyers about 12000 customers have given a high rating. Hence, we can make an assumption here that the service offered in the business travel by Cheapseats satisfies customers' needs. Further analysis on the other travel types are done to find out the number of high ratings and low ratings.

According to the explanation provided above, the high average and low ratings have been grouped in the following table.

	Var1	Freq
1	Average	7396
2	High	13050
3	Low	5612

Figure 14: Average, high and low customer satisfaction

Next, this is the bar chart for Mileage Tickets customer and their Satisfaction ratings.

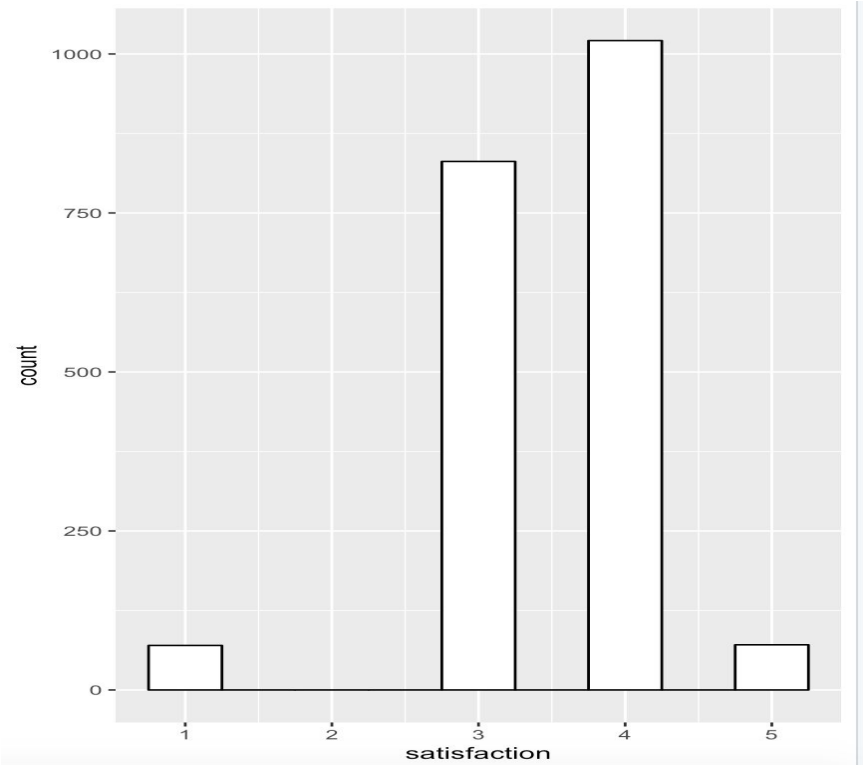


Figure 15: Mileage Tickets travel vs Satisfaction

Based on the information we get from the above bar chart, we observe that out of the total number of Mileage tickets travelers i.e. 1993, about 1090 (i.e. almost 50% of the total number of mileage travelers) have given a high rating (4 and 5) and a very small fraction of people have given a low rating. Hence, evaluating this observation we can assume that the services provided by Mileage are decent. But in order to find out if there are other parameters involved in the satisfaction rating.

Finally, we can see the bar chart of personal travel and Satisfaction.

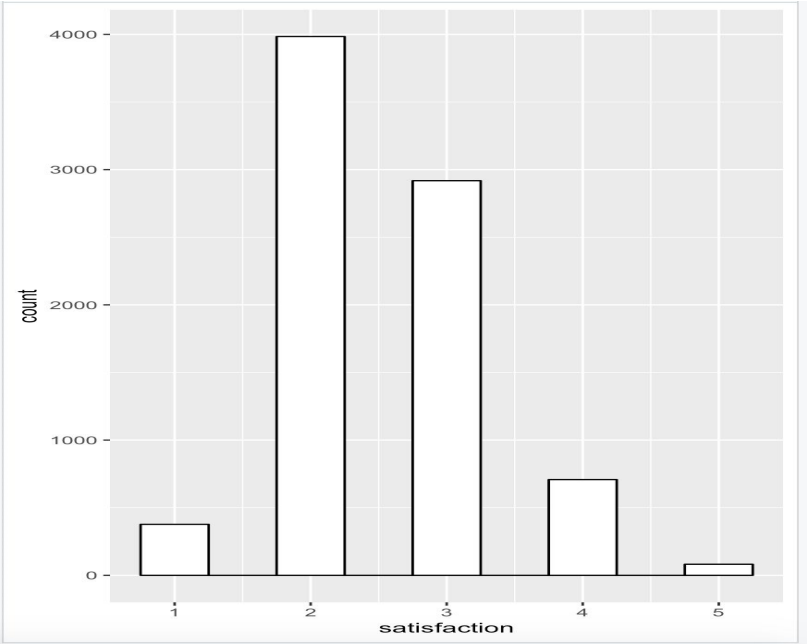


Figure 16: Personal Travel type vs Satisfaction

From the above bar chart, we observe that in the Personal travel type, out of the total number of 8069 flyers about 4200 flyers have given a low rating (below 3) i.e. a little more than 50% of the personal travel type travelers. This observation helps us assume that the travel type plays a significant role in impacting the customer satisfaction. Hence, in the further analysis these values are going to assist in enhancing the overall customer satisfaction rating.

Additionally, from the travel type category, we may assess that the personal travel type weighs in the most on lowering the overall satisfaction score as almost 50% of the personal travel types give a lower rating. The insightful information gained from this observation is that Cheapseats tends to cater more towards the business traveler than the personal traveler.

d. Variable: Number of Flights & First Year of Flight

We would like to look at these two variables together as they are tightly related to each other. The data cleaning job is also easy for these two variables as they are both integers and no NAs. To generate insightful analysis, we created two new columns to calculate (1) the number of years being our customers (2) the average number of flights per year. Below are the comparison between 2 airlines - Cheapseats and West, and you can see that more customers of Cheapseats are long-term customers (left), and West has more newer customers.

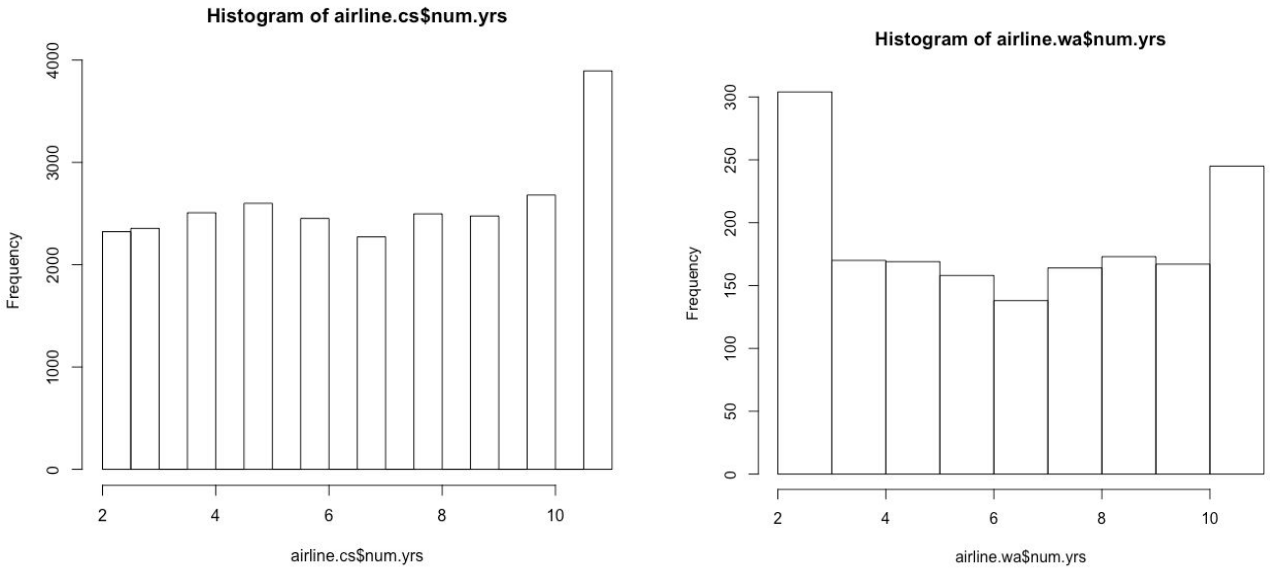


Figure 17: Number of year of being a customer histogram Cheapseats (left) vs. West Airline (right)

Next, we further standardized by dividing the number of flights by the number of years, and here you can see that in Cheapseats, most customers take about less than 4 flights a year, and about 39% take over 10 flights a year, and 5.7% of customer take more than 20%. Also, you can see that over 60% of the customers do not fly more than 10% of other airlines.

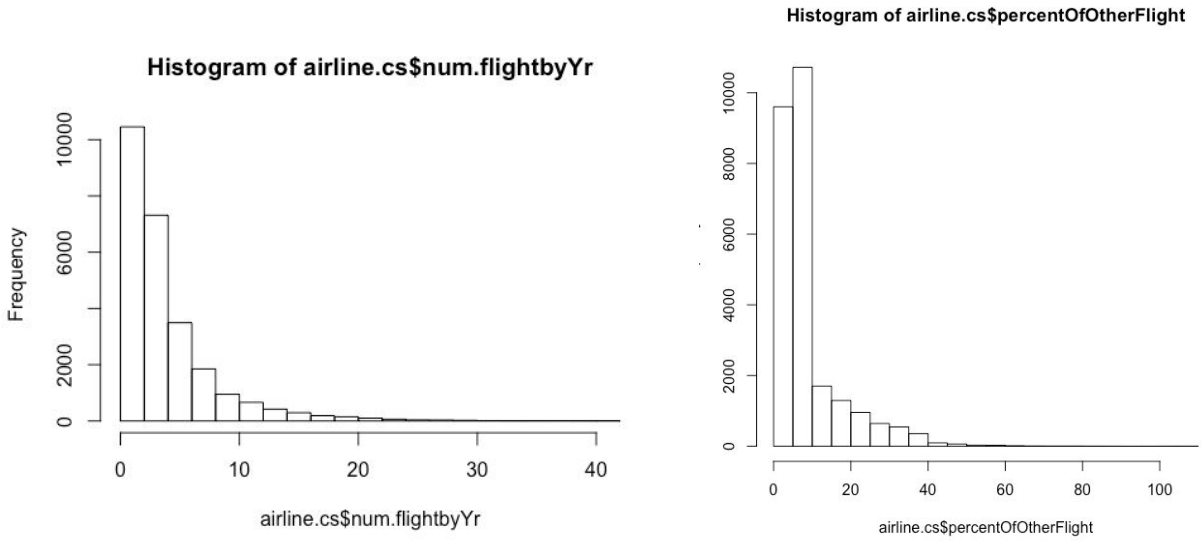


Figure 18: Average number of flights taken per year(left); percentage of taking other airlines (right)

So, we roughly divided those customers into 4 groups to see if the satisfaction ratings changes when customer take more flights. Here is how we group them: group#1 - average number of flights per year is fewer than median (2.67 flights), group#2 - between 2.67 to 5 flights, group#3 - between 5 and 10, and lastly, group#4 - over 10 flights. As you can see below, customers, who took less than 5 flights a year, still rated mostly at 4, however, for the customers who flew more than 5 flights a year, you can see the histogram shifts slightly to rating 2 and 3, especially, the most frequent flyers rated mostly 3, which is less than median rating.

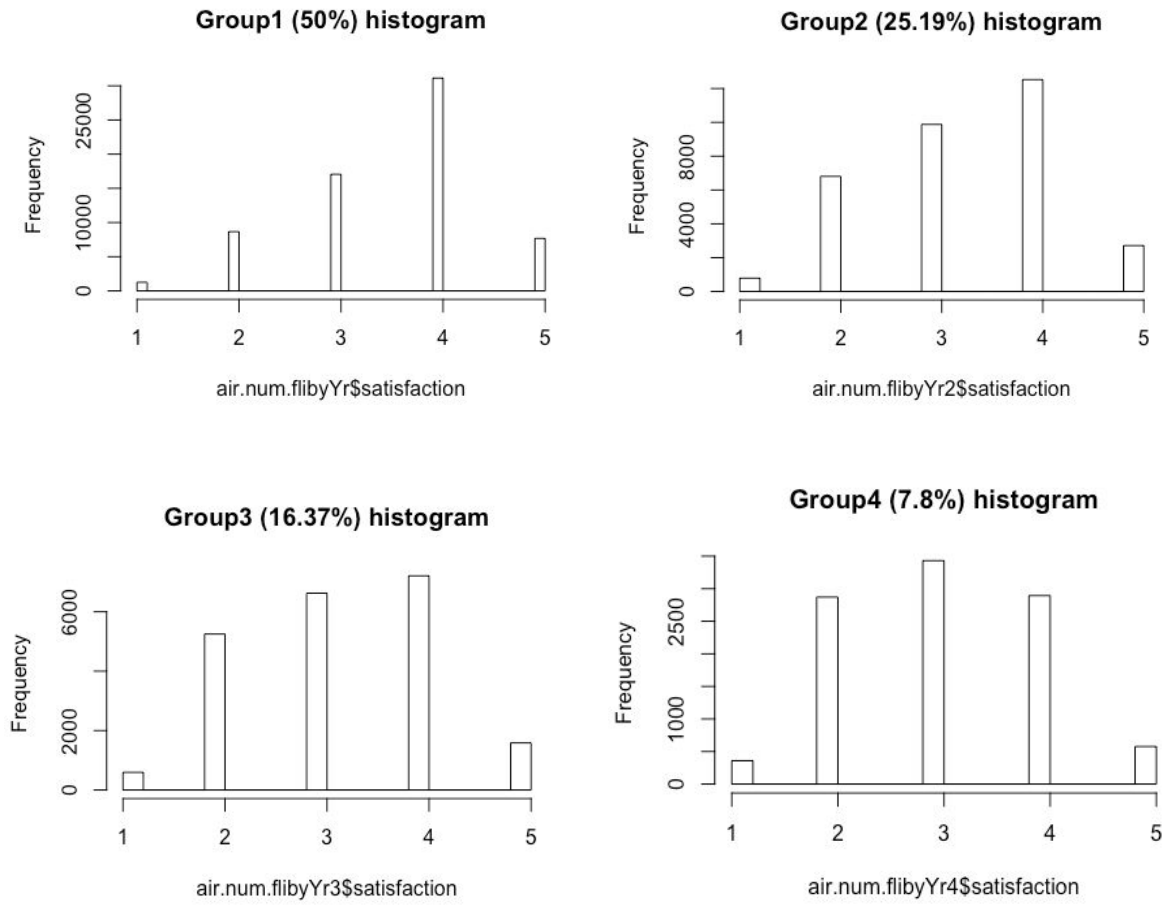


Figure 19: histogram changes of satisfaction ratings for different average number of flights per year

e. Variable: Origin City & Destination City

The next two variables were assessed via are the Origin and Destination City. The team is curious to know if the geographic location has an effect on the satisfaction score. The two bar charts show the difficulty in using the origin or destination city as a factor in understanding the overall satisfaction score. Within the Origin bar chart, there are a few cities that have a higher rating, but their significance is low on the overall population. So given the origin and destination location, their overall effect or impact on the satisfaction score is low. Thus, we conclude that geographic location has little to no impact on the customer satisfaction rating.

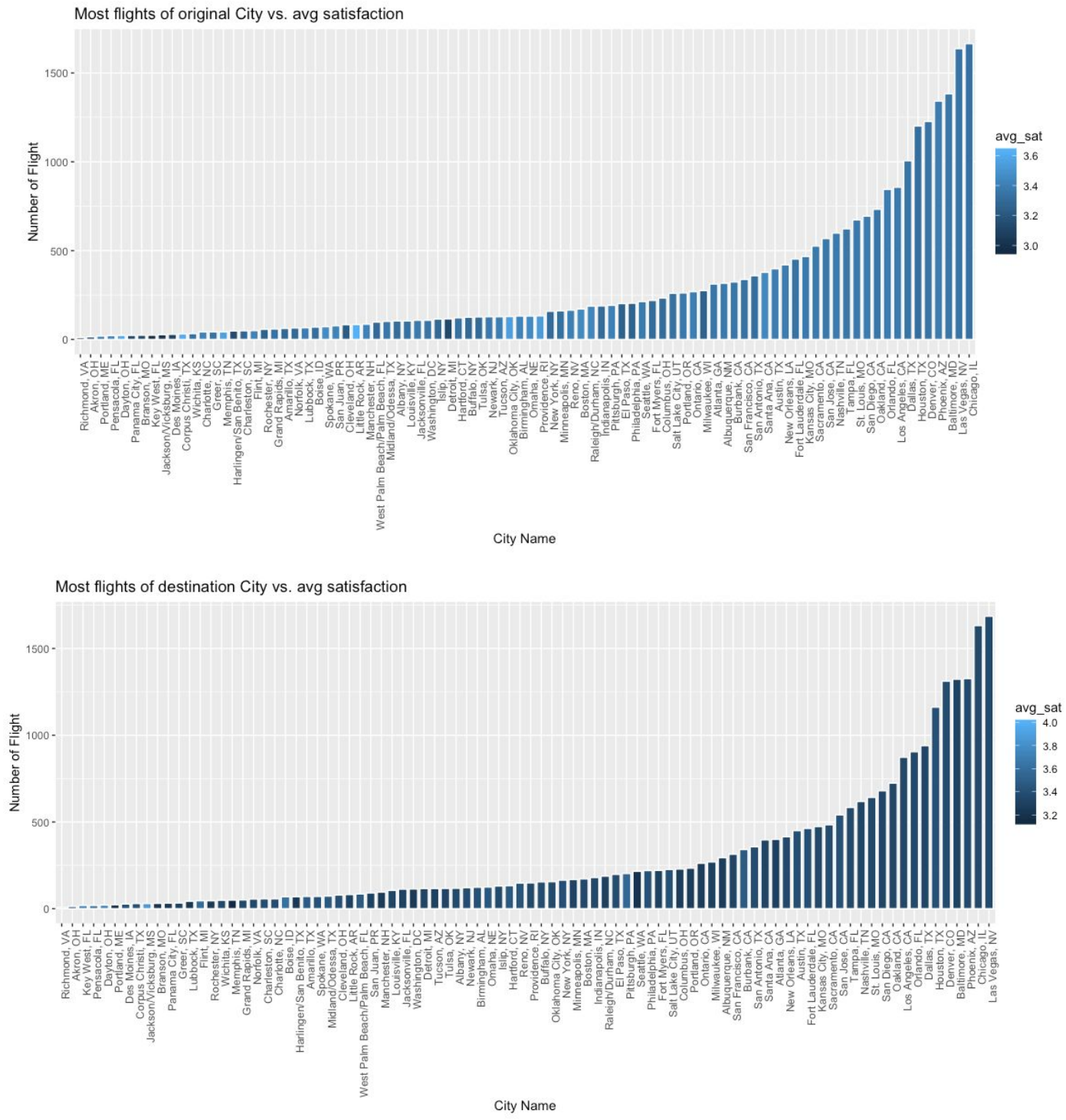


Figure 20: bar chart of original city (upper) and destination city (bottom)

g. Variable : Delay & Cancellation

Here, we investigated two variables - arriveDelay.5 and flightCancelled, both of them are in good form and no NAs. There are 58.57% flights not delayed over 5 minutes, and approximately 1.2% cancelled flights. The average customer ratings for flights were delayed is 3.16 while no delay is about 3.49, which is about 10.22% difference. The average customer ratings for flights that were cancelled is 3.13, which is 7.27% lower than overall average ratings. You can also see from the histogram below that there are more people gave lower ratings (< = 3) when the flight is cancelled.

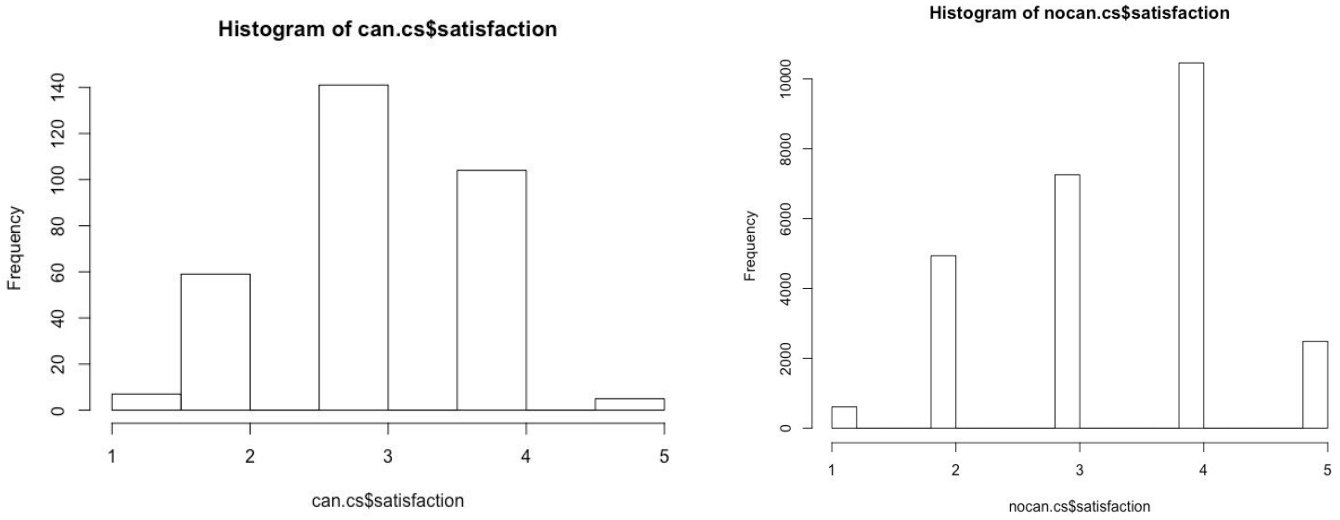


Figure 21: histogram of satisfaction ratings for cancelled (right) and non-cancelled (left) flights

5B-2. Level 2: Statistical Models

A) Linear Regression

Simple linear Regression model was performed on various individual variables of the dataset for CheapSeats Airlines in order to predict the level of impact it has on customer satisfaction. Below are results from the models that performed the best within the Cheapseat Airlines dataset and some that did not. The process ultimately similarly modeled a stepwise linear regression process, where on the variables with significance were added to the model. Several of the linear models are below.

1. **Age:** The linear regression model depicts that age has R squared value of 0.0488 i.e. having around 4.8% impact on customer satisfaction.

```
> summary(regression)

Call:
lm(formula = satisfaction ~ age, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6819 -0.6695  0.2685  0.6404  2.0618

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.9297791  0.0167129  235.13   <2e-16 ***
age         -0.0123950  0.0003388  -36.59   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9486 on 26056 degrees of freedom
Multiple R-squared:  0.04887,   Adjusted R-squared:  0.04883
F-statistic:  1339 on 1 and 26056 DF,  p-value: < 2.2e-16
```

Figure 22: Lm() Satisfaction vs Age

2. **Airline Status**: Airline status has an R-squared value of .1237, which means it has statistical significance for all the levels of airline status. Below in the excerpt from RStudio, it is observed that airlines status has a significant effect on the satisfaction score. So far what the individual linear regressions depict is that age and airline status are a good predictor of what a person's

satisfaction rating will be. The R-squared value is interesting as for Cheapseats, airline status has a higher R-squared value than the entire survey dataset (#Multiple R-squared: 0.1184 <= full survey)

```
Call:
lm(formula = satisfaction ~ airlineStatus, data = cheapseats)

Residuals:
     Min       1Q   Median       3Q      Max
-2.66586 -0.93655  0.06345  0.87041  1.87041

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.129590   0.006813  459.38   <2e-16 ***
airlineStatusGold   0.592501   0.020904   28.34   <2e-16 ***
airlineStatusPlatinum 0.536266   0.032463   16.52   <2e-16 ***
airlineStatusSilver   0.806957   0.014297   56.44   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9106 on 26054 degrees of freedom
Multiple R-squared:  0.1237,    Adjusted R-squared:  0.1236
F-statistic:  1225 on 3 and 26054 DF,  p-value: < 2.2e-16
```

Figure 23: Lm() Satisfaction vs AirlineStatus

3. **Loyalty Card Number**: In contrast, the number of loyalty cards a customer has is statistically significant, however with a R-squared value of .007 or .734%, the number of loyalty cards a customer has does not significantly influence their satisfaction score. For our model, we will leave the number of loyalty cards out as the effect is really small.

```
Call:
lm(formula = satisfaction ~ loyaltyCardsNum, data = cheapseats)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8077 -0.4391  0.3397  0.7084  1.7084

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.291650   0.007624  431.77   <2e-16 ***
loyaltyCardsNum 0.073729   0.005276   13.97   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.969 on 26056 degrees of freedom
Multiple R-squared:  0.007439,  Adjusted R-squared:  0.007401
F-statistic: 195.3 on 1 and 26056 DF,  p-value: < 2.2e-16
```

Figure 24: Lm() Satisfaction vs Loyalty Card Number

4. **Origin City**: To assess the effect the geography has on the customer's satisfaction score, the linear regression is performed on the Origin City of flight. This is to test if the there is a significant effect or other geography impact on the overall satisfaction score. The following model shows that the origin city of flight has little to no effect on the overall customer satisfaction score. The significance is less than .3%. Thus for the model, we will exclude geographic locations in the multiple linear regression model.

```
originCityWashington, DC                 0.218519   0.268054   0.815    0.415
originCityWest Palm Beach/Palm Beach, FL  0.162585   0.269717   0.603    0.547
originCityWichita, KS                    0.200000   0.302932   0.660    0.509
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9728 on 25970 degrees of freedom
Multiple R-squared:  0.002967,  Adjusted R-squared:  -0.0003728
F-statistic: 0.8884 on 87 and 25970 DF,  p-value: 0.7622
```

Figure 25: Lm() Satisfaction vs Origin City

5. **Arrival Delay in Minutes**: With a R-squared score of 4.3%, we assessed that the arrival delay

variable is significant enough in identifying what a customers satisfaction score will be. Not fully shown in the figure below, but among the values that arrival delay, there are several minutes that display the most significance. The times are 10-30 minutes of delay. These values are interesting and while it cannot be proven with the data, the likelihood of those times being significant is possibly due to missing or making layovers at major airports.

```
arriveDelayMinute30  -0.3787288  0.0810759  -4.671 3.01e-06 ***
arriveDelayMinute300  1.5007038  0.9571914   1.568 0.116935
arriveDelayMinute301 -0.4992962  0.9571914  -0.522 0.601935
arriveDelayMinute302  0.5007038  0.6768645   0.740 0.459464
arriveDelayMinute304 -1.4992962  0.9571914  -1.566 0.117279
arriveDelayMinute305  0.5007038  0.6768645   0.740 0.459464
arriveDelayMinute31  -0.1535018  0.0929404  -1.652 0.098625 .
 [ reached getOption("max.print") -- omitted 111 rows ]
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9572 on 25747 degrees of freedom
Multiple R-squared:  0.04309,   Adjusted R-squared:  0.03157
F-statistic:  3.74 on 310 and 25747 DF,  p-value: < 2.2e-16
```

Figure 26: Lm() Satisfaction vs Delay Arrival Time(min)

6. **Date**: looking at the linear models and their ability to inspect issues as per a customer rating, the Date variable is run through the linear regression model. Inspecting the date variable returns a .36% reliability that date affects the satisfaction score. Therefore we will not include this in the multiple linear regression model. One interesting finding does appear within the date model however. Figure # depicts that on 19 February 2014, US Homeland Security published a warning about a possible shoe bombing threat. This date was the displayed as the most significant within the date linear model:

```
date2/19/2014   0.249300    0.081692    3.052   0.00228 **
```

http://articles.latimes.com/2014/feb/19/news/la-nn-shoebomb-threat--20140219

```
date3/28/2014  0.157979   0.080416   1.965  0.04948 *
date3/29/2014  0.095398   0.085627   1.114  0.26524
date3/3/2014   0.082862   0.083559   0.992  0.32138
date3/30/2014  0.101453   0.083420   1.216  0.22393
date3/31/2014  0.098236   0.081006   1.213  0.22526
date3/4/2014   0.139889   0.083772   1.670  0.09495 .
date3/5/2014   0.084069   0.083630   1.005  0.31479
date3/6/2014   0.108526   0.082176   1.321  0.18663
date3/7/2014   0.158788   0.081931   1.938  0.05263 .
date3/8/2014   0.123666   0.084434   1.465  0.14303
date3/9/2014   0.112550   0.082554   1.363  0.17278
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9725 on 25968 degrees of freedom
Multiple R-squared:  0.003607,  Adjusted R-squared:  0.0001922
F-statistic: 1.056 on 89 and 25968 DF,  p-value: 0.3382
```

Figure 27: Lm() Satisfaction vs date

Homeland Security warns airlines of possible shoe bomb threat

February 19, 2014 | By Brian Bennett

Email Share G+ Tweet Recommend 0

WASHINGTON -- The Department of Homeland Security warned airlines Wednesday to watch for explosives hidden in the shoes of passengers flying into the United States from overseas, officials said.

The alert was based on new intelligence indicating that a shoe bomb may be used to blow up a U.S.-bound jetliner, said two law enforcement officials who described the bulletin on the condition of anonymity.

Officials said the threat was not specific to a particular airline, flight, country or time. It was not related to the Winter Olympics underway in Sochi, Russia.

The alert was issued "out of an abundance of caution," said a homeland security official.

Airport screeners at international airports were instructed to step up scrutiny of passengers boarding flights for the United States.

Screeners will increase uses of swabs that can detect traces of explosive powder on shoes, bags and hands. They also are likely to pull aside more passengers for pat-downs and full-body screening, officials said.

TSA agents work at a checkpoint for pre-cleared passengers at Hartsfield-Jackson... (Kent D. Johnson / Associated...)

Figure 28: LA Times Article

7. **Class** : It has a R- Squared value of 0.002196 ,i.e 0.2196% and it has very minimum impact on customer satisfaction, further analysed it needs to be combined with other factors to have a significant impact on satisfaction.

```
Call:
lm(formula = satisfaction ~ class, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5055 -0.3497  0.4945  0.6503  1.6988

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.50545    0.02115 165.775  < 2e-16 ***
classEco      -0.15575    0.02217  -7.025 2.20e-12 ***
classEco Plus -0.20426    0.02826  -7.228 5.04e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9716 on 26055 degrees of freedom
Multiple R-squared:  0.002272,  Adjusted R-squared:  0.002196
F-statistic: 29.67 on 2 and 26055 DF,  p-value: 1.344e-13
```

Figure 29: Lm() Satisfaction vs class

8. **Travel Type**: It has a R-squared value of 0.3361 , i.e 33.61% and it is a factor which has a significant effect on travel type, the travel type has a major effect on the customer satisfaction , it matters whether a trip is for personal or a mileage type ticket. Customers who have a mileage type ticket will prefer to travel more frequently and will be more satisfied.

```
Call:
lm(formula = satisfaction ~ travelType, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7598 -0.5209  0.2402  0.4791  2.4791

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                3.759815   0.006266   600.0   <2e-16 ***
travelTypeMileage tickets -0.246518   0.018826   -13.1   <2e-16 ***
travelTypePersonal Travel -1.238933   0.010822  -114.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7925 on 26055 degrees of freedom
Multiple R-squared:  0.3361,    Adjusted R-squared:  0.3361
F-statistic:  6596 on 2 and 26055 DF,  p-value: < 2.2e-16
```

Figure 30: Lm() Satisfaction vs Travel Type

9. **Gender** : The R-squared value is 0.01884 i.e 1.88% and it is a factor which doesn't have significant impact on satisfaction. We haven't considered this parameter as gender doesn't affect the satisfaction.

```
Call:
lm(formula = satisfaction ~ gender, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5090 -0.5090  0.4910  0.7605  1.7605

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.239534   0.007955  407.21   <2e-16 ***
genderMale  0.269420   0.012032   22.39   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9634 on 26056 degrees of freedom
Multiple R-squared:  0.01888,   Adjusted R-squared:  0.01884
F-statistic: 501.4 on 1 and 26056 DF,  p-value: < 2.2e-16
```

Figure 31: Lm() Satisfaction vs Gender

10. **Price Sensitivity** : The R-squared value is 0.0093 i.e 0.93% and it is a factor which doesn't have significant impact on satisfaction. We haven't considered this parameter as Price sensitivity doesn't affect the satisfaction.

```
Call:
lm(formula = satisfaction ~ priceSensitivity, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5783 -0.4048  0.4217  0.5952  2.1155

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.57830    0.01531  233.67   <2e-16 ***
priceSensitivity -0.17346    0.01106  -15.68   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9681 on 26056 degrees of freedom
Multiple R-squared:  0.009351,  Adjusted R-squared:  0.009313
F-statistic:   246 on 1 and 26056 DF,  p-value: < 2.2e-16
```

Figure 32: Lm() Satisfaction vs PriceSensitivity

11. **Number of flights** : The R-squared value is 0.0552 i.e 5.52% and it is a factor which we have considered for the model.

```
Call:
lm(formula = satisfaction ~ numberOfFlights, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6780 -0.5819  0.3220  0.6263  2.4270

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.6780118  0.0100902  364.51   <2e-16 ***
numberOfFlights -0.0160153  0.0004103  -39.03   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9454 on 26056 degrees of freedom
Multiple R-squared:  0.05524,   Adjusted R-squared:  0.0552
F-statistic:  1523 on 1 and 26056 DF,  p-value: < 2.2e-16
```

Figure 33: Lm() Satisfaction vs Number of Flights

12. **Shopping At Airport & Food At Airport**: The R-squared value is 0.0001322 i.e 0.1322% and it is a factor which we have considered for the model.

```
Call:
lm(formula = satisfaction ~ shoppingAtAirport + foodAtAirport,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4169 -0.3714  0.5608  0.6498  1.6569

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.3431194  0.0102710 325.491   <2e-16 ***
shoppingAtAirport 0.0002343  0.0001146   2.044    0.041 *
foodAtAirport     0.0001177  0.0001165   1.011    0.312
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9726 on 26055 degrees of freedom
Multiple R-squared:  0.000209,  Adjusted R-squared:  0.0001322
F-statistic: 2.723 on 2 and 26055 DF,  p-value: 0.06569
```

Figure 34: Lm() Satisfaction vs Shopping At Airport

13. **Scheduled Departure Delay** : The R-squared value is 0.00515 i.e 0.51% and it is a factor which we have considered for the model.

```
Call:
lm(formula = satisfaction ~ scheduledDepDelay, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3964 -0.3964  0.6036  0.6177  2.4081

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.3963785  0.0068170  498.22   <2e-16 ***
scheduledDepDelay -0.0020061  0.0001731  -11.59   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9716 on 25740 degrees of freedom
  (316 observations deleted due to missingness)
Multiple R-squared:  0.005189,  Adjusted R-squared:  0.00515
F-statistic: 134.3 on 1 and 25740 DF,  p-value: < 2.2e-16
```

Figure 35: Lm() Satisfaction vs Scheduled Departure Delay

Linear Regression Conclusion

Independently, each linear regression performed above shows whether the independent variable is significant enough to include in a model for predictability. In the next section, several of these single linear models will be added to perform a stepwise regression and assess which variables are the best at predicting the satisfaction score of a customer.

B) Multiple Linear Regression

Multiple Linear Regression is a model that analyzes multiple independent variables in a dataset against a dependent variable. The model analyzes which variables have the best fitted linear line or logarithmic line and is used to assist decision makers in predicting the likelihood the independent variables will influence the dependent variable. For this model, Multiple Linear Regression models were

performed on the Cheapseats Airlines Inc dataset. Below are results from the models that performed the best within the Cheapseat Airlines dataset.

The individual variables had a low impact on customer satisfaction. Thus, we decided to perform multiple linear regression model by combining the variables to see the whether there were any significant changes. so far, the individual variables had quite a low impact on customer satisfaction but when they were combined with other variables against customer satisfaction, the R-squared values were increasing showing the increased effect the combination of various independent variables have on customer satisfaction. Below are the results from the models that had significant impact on customer satisfaction.

1. **Age & Airline Status**: It has a R-Squared value of 0.1733, i.e. 17.33 % impact on customer satisfaction. This model is kept due to the significant increase in predictability age and airline status have on satisfaction.

```
Call:
lm(formula = satisfaction ~ age + airlineStatus, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.88427 -0.70397  0.03345  0.75748  2.29603

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.7059252  0.0159871  231.81   <2e-16 ***
age                -0.0125244  0.0003163  -39.60   <2e-16 ***
airlineStatusGold   0.6209737  0.0203145   30.57   <2e-16 ***
airlineStatusPlatinum  0.5791231  0.0315473   18.36   <2e-16 ***
airlineStatusSilver    0.7991767  0.0138868   57.55   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8844 on 26053 degrees of freedom
Multiple R-squared:  0.1734,    Adjusted R-squared:  0.1733
F-statistic:  1366 on 4 and 26053 DF,  p-value: < 2.2e-16
```

Figure 36: Age + airlineStatus

2. **Airline Status & Travel type**: It has a R-Squared value of 0.4139, i.e. 41.39 % impact on customer satisfaction. This model with increases several percentage points with these two independent variables. Therefore, we assess that travel type is significant in predicting the satisfaction score of a customer.

```
> data <- read.csv("cheapseats.csv",stringsAsFactors = FALSE)
> View(data)
> data <- data[,-1]
> View(data)
> regression <- lm(formula = satisfaction ~  airlineStatus + travelType, data = data)
> summary(regression)

Call:
lm(formula = satisfaction ~ airlineStatus + travelType, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.86164 -0.39082 -0.04214  0.44999  2.60918

Coefficients:
                              Estimate Std. Error  t value Pr(>|t|)
(Intercept)                   3.550006   0.006966  509.640   <2e-16 ***
airlineStatusGold             0.443114   0.017163   25.818   <2e-16 ***
airlineStatusPlatinum         0.311637   0.026635   11.700   <2e-16 ***
airlineStatusSilver           0.661880   0.011777   56.199   <2e-16 ***
travelTypeMileage tickets    -0.169750   0.017746   -9.565   <2e-16 ***
travelTypePersonal Travel    -1.159190   0.010275 -112.813   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7446 on 26052 degrees of freedom
Multiple R-squared:  0.414,     Adjusted R-squared:  0.4139
F-statistic:  3681 on 5 and 26052 DF,  p-value: < 2.2e-16

> plot(regression)
Hit <Return> to see next plot:
```

Figure 37: airlineStatus + travelType

3. **Age & Travel type**: It has a R-Squared value of 0.337, i.e. 33.7 % impact on customer satisfaction. Once again, the combination of travel type increases the base Adjusted R-Square Score and therefore Age & travel type are good predictors of a customers satisfaction score.

```
lm(formula = satisfaction ~ age + travelType, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7926 -0.5505  0.2261  0.4402  2.5075

Coefficients:
                            Estimate Std. Error  t value Pr(>|t|)
(Intercept)                3.8395714  0.0142099  270.204  < 2e-16 ***
age                       -0.0018776  0.0003003   -6.253  4.1e-10 ***
travelTypeMileage tickets -0.2496869  0.0188189  -13.268  < 2e-16 ***
travelTypePersonal Travel -1.2156748  0.0114355 -106.307  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7919 on 26054 degrees of freedom
Multiple R-squared:  0.3371,    Adjusted R-squared:  0.337
F-statistic:  4416 on 3 and 26054 DF,  p-value: < 2.2e-16
```

Figure 38: Age + travelType

4. **Age & airline status & travel type**: It has a R-Squared value of 0.4157, i.e. 41.57 % impact on customer satisfaction. The Adjusted R-square value keeps rising with the addition of the three variables. Further evaluation is needed to identify which other variables increase the adjusted R-Square score.

```
lm(formula = satisfaction ~ age + airlineStatus + travelType,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8991 -0.4437 -0.0263  0.4949  2.6055

Coefficients:
                            Estimate Std. Error  t value Pr(>|t|)
(Intercept)                3.6579773  0.0136880  267.240   <2e-16 ***
age                       -0.0025910  0.0002829   -9.158   <2e-16 ***
airlineStatusGold          0.4529551  0.0171695   26.381   <2e-16 ***
airlineStatusPlatinum      0.3265990  0.0266427   12.258   <2e-16 ***
airlineStatusSilver        0.6641800  0.0117613   56.471   <2e-16 ***
travelTypeMileage tickets -0.1732173  0.0177221   -9.774   <2e-16 ***
travelTypePersonal Travel -1.1261807  0.0108737 -103.569   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7434 on 26051 degrees of freedom
Multiple R-squared:  0.4159,    Adjusted R-squared:  0.4157
F-statistic:  3091 on 6 and 26051 DF,  p-value: < 2.2e-16
```

Figure 39: Age + airlineStatus + travelType

5. **Age & class**: It has a R-Squared value of 0.0507 , i.e 5.07%, it needs to be combined with other variables so that it can have a significant impact and we can increase the adjusted R-square score.

```
Call:
lm(formula = satisfaction ~ age + class, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6729 -0.6717  0.1785  0.6361  2.1070

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.0687252  0.0257551 157.977  < 2e-16 ***
age          -0.0123601  0.0003385 -36.517  < 2e-16 ***
classEco     -0.1486512  0.0216256  -6.874 6.39e-12 ***
classEco Plus -0.1869450  0.0275671  -6.781 1.22e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9476 on 26054 degrees of freedom
Multiple R-squared:  0.05085,   Adjusted R-squared:  0.05074
F-statistic: 465.3 on 3 and 26054 DF,  p-value: < 2.2e-16
```

Figure 40: Age + class

6. **Age & airline status & travel type & class**: It has a R-Squared value of 0.4161, i.e. 41.61 % impact on customer satisfaction.

```
Call:
lm(formula = satisfaction ~ age + airlineStatus + travelType +
    class, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.94777 -0.44608 -0.02537  0.49652  2.60826

Coefficients:
                            Estimate Std. Error  t value Pr(>|t|)
(Intercept)                3.7193345  0.0206293  180.294  < 2e-16 ***
age                       -0.0025877  0.0002829   -9.148  < 2e-16 ***
airlineStatusGold          0.4520296  0.0171699   26.327  < 2e-16 ***
airlineStatusPlatinum      0.3241818  0.0266484   12.165  < 2e-16 ***
airlineStatusSilver        0.6635639  0.0117621   56.416  < 2e-16 ***
travelTypeMileage tickets -0.1722329  0.0177190   -9.720  < 2e-16 ***
travelTypePersonal Travel -1.1246757  0.0108782 -103.388  < 2e-16 ***
classEco                  -0.0657698  0.0169771   -3.874 0.000107 ***
classEco Plus             -0.0786928  0.0216478   -3.635 0.000278 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7432 on 26049 degrees of freedom
Multiple R-squared:  0.4163,    Adjusted R-squared:  0.4161
F-statistic:  2322 on 8 and 26049 DF,  p-value: < 2.2e-16
```

Figure 41: Age + airlineStatus + travelType + class

7. **Age vs airline status vs travel type vs class vs number of flights**: It has a R-Squared value of 0.4172, i.e. 41.72 % impact on customer satisfaction.

```
lm(formula = satisfaction ~ age + airlineStatus + travelType +
    class + numberOfFlights, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.96687 -0.44665 -0.02876  0.50243  2.67237

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                3.7494532  0.0210498 178.123  < 2e-16 ***
age                       -0.0022184  0.0002874  -7.717 1.23e-14 ***
airlineStatusGold          0.4407470  0.0172287  25.582  < 2e-16 ***
airlineStatusPlatinum      0.3130205  0.0266708  11.736  < 2e-16 ***
airlineStatusSilver        0.6548963  0.0118155  55.427  < 2e-16 ***
travelTypeMileage tickets -0.1615824  0.0177670  -9.095  < 2e-16 ***
travelTypePersonal Travel -1.1082526  0.0111158 -99.701  < 2e-16 ***
classEco                  -0.0661560  0.0169614  -3.900 9.63e-05 ***
classEco Plus             -0.0877850  0.0216662  -4.052 5.10e-05 ***
numberOfFlights           -0.0024375  0.0003463  -7.038 2.00e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7425 on 26048 degrees of freedom
Multiple R-squared:  0.4174,    Adjusted R-squared:  0.4172
F-statistic:  2073 on 9 and 26048 DF,  p-value: < 2.2e-16
```

Figure 42: Age + airlineStatus + travelType + class + numberOfFlights

8. **Age & airline status & travel type & class & number of flights & arrival delay minutes**: It has a R-Squared value of 0.4262, i.e. 42.62 % impact on customer satisfaction.

```
lm(formula = satisfaction ~ age + airlineStatus + travelType +
    class + numberOfFlights + arriveDelayMinute, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.95951 -0.43581 -0.03002  0.49803  2.77327

Coefficients:
                            Estimate Std. Error  t value Pr(>|t|)
(Intercept)                3.7917989  0.0211931  178.917  < 2e-16 ***
age                       -0.0022063  0.0002884   -7.650 2.07e-14 ***
airlineStatusGold          0.4477426  0.0172559   25.947  < 2e-16 ***
airlineStatusPlatinum      0.3309475  0.0267234   12.384  < 2e-16 ***
airlineStatusSilver        0.6536387  0.0118202   55.299  < 2e-16 ***
travelTypeMileage tickets -0.1665325  0.0177983   -9.357  < 2e-16 ***
travelTypePersonal Travel -1.1173652  0.0111561 -100.157  < 2e-16 ***
classEco                  -0.0687672  0.0169315   -4.061 4.89e-05 ***
classEco Plus             -0.0887706  0.0216557   -4.099 4.16e-05 ***
numberOfFlights           -0.0024337  0.0003467   -7.019 2.29e-12 ***
arriveDelayMinute         -0.0023675  0.0001315  -17.998  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7381 on 25658 degrees of freedom
  (389 observations deleted due to missingness)
Multiple R-squared:  0.4264,    Adjusted R-squared:  0.4262
F-statistic:  1907 on 10 and 25658 DF,  p-value: < 2.2e-16
```

Figure 43: Age + airlineStatus + travelType + class + numberOfFlights + arriveDelayMinute

C) Association Rules

Association rule learning is a machine learning technique where it looks for interesting relationships among variables in the dataset ("transaction") by evaluating the "support", "confidence" and "lift" value in the rules. In the association rule we ran, we included 20 variables, here are the list of variables we included in the model (in order): satisfaction, average number of flight per year, number of years of being customer (2014 minus the first year of customer), foodAtAirport, shoppingAtAirport, numberOfFlights, loyaltyCardsNum, percentOfOtherFlight, flight Distance, airlineStatus, gender, travelType, class, flightCancelled, arriveDelay.5, priceSensitivity, age (by groups), date (by month), originCity, destCity

```
> str(ruleDF.cs)
'data.frame':   26058 obs. of  20 variables:
 $ happyCust      : Factor w/ 3 levels "Average","High",..: 3 2 2 2 1 1 2 1 1 1 ...
 $ avg.fl         : Factor w/ 3 levels "Average","High",..: 3 3 3 1 3 3 1 1 2 3 ...
 $ yr.cs          : Factor w/ 3 levels "Average","High",..: 2 2 2 1 2 2 2 2 3 2 ...
 $ food.cs        : Factor w/ 3 levels "Average","High",..: 1 2 1 2 1 1 2 2 3 1 ...
 $ shopping.cs    : Factor w/ 3 levels "Average","High",..: 3 1 2 2 3 3 3 2 2 2 ...
 $ numOfFl.cs     : Factor w/ 3 levels "Average","High",..: 3 3 3 1 1 3 2 1 2 3 ...
 $ loyCarNum.cs   : Factor w/ 3 levels "Average","High",..: 2 3 3 2 3 2 3 1 1 1 ...
 $ percOtherFl.cs : Factor w/ 3 levels "Average","High",..: 2 3 3 2 3 3 1 2 2 1 ...
 $ flDis.cs       : Factor w/ 3 levels "Average","High",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ airlineStatus  : Factor w/ 4 levels "Blue","Gold",..: 1 4 1 4 4 1 4 4 1 1 ...
 $ gender         : Factor w/ 2 levels "Female","Male": 1 1 1 1 2 1 1 1 1 1 ...
 $ travelType     : Factor w/ 3 levels "Business travel",..: 3 1 1 1 1 1 2 3 3 3 ...
 $ class          : Factor w/ 3 levels "Business","Eco",..: 1 2 2 2 2 2 2 2 2 2 ...
 $ flightCancelled: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ arriveDelay.5  : Factor w/ 2 levels "no","yes": 2 2 1 1 2 2 1 2 1 1 ...
 $ pri.sen        : Factor w/ 5 levels "0","1","2","3",..: 2 2 3 3 2 3 3 2 3 2 ...
 $ AgeGroup       : Factor w/ 11 levels "0-9","10-19",..: 4 6 3 5 6 3 8 6 5 5 ...
 $ month          : Factor w/ 3 levels "FEB","JAN","MAR": 3 2 2 3 2 2 1 1 1 3 ...
 $ originCity     : Factor w/ 88 levels "Akron, OH","Albany, NY",..: 48 48 48 48 48 48 48 48 48 48 ...
 $ destCity       : Factor w/ 88 levels "Akron, OH","Albany, NY",..: 80 80 80 80 80 80 80 80 80 80 ...
```

Figure 44: the structure of the dataset of transactions we ran in arule

For the satisfaction ratings, we defined "high" for ratings that are larger or equal to 4, "low" for ratings that are lower or equal to 3, and the rest being "average". Since we are focusing on studying high customer ratings and low customer ratings, we managed to make the equal proportion of high (50.08%) and low (49.92%) groups. As for the rest of the variables, we defined "low" if it is lower than percentile 35th, "high" if it is over 60th, and the rest will be "average".

```
> airlineX.cs1 <- as(ruleDF.cs1,"transactions")
> itemFrequency(airlineX.cs) #this tells you the frequencies for each items
      happyCust=High        happyCust=Low       avg.fl=Average
         0.5008058945         0.4991941055         0.3012126794
          avg.fl=High           avg.fl=Low          yr.cs=Average
         0.3480313148         0.3507560058         0.2771125950
           yr.cs=High            yr.cs=Low       food.cs=Average
         0.3473405480         0.3755468570         0.2606876967
         food.cs=High          food.cs=Low shopping.cs=Average
         0.3382454525         0.4010668509         0.0839281603
     shopping.cs=High      shopping.cs=Low     numOfFl.cs=Average
         0.3459206386         0.5701512012         0.2662138307
```

Figure 45: quick glimpse of item frequency matrix

Next, in the following 2 figures, you can see the frequencies for each item, and since we included quite some variables, it is hard to tell which item is related to which. So, I eliminated both original city and destination city to get a clearer view.

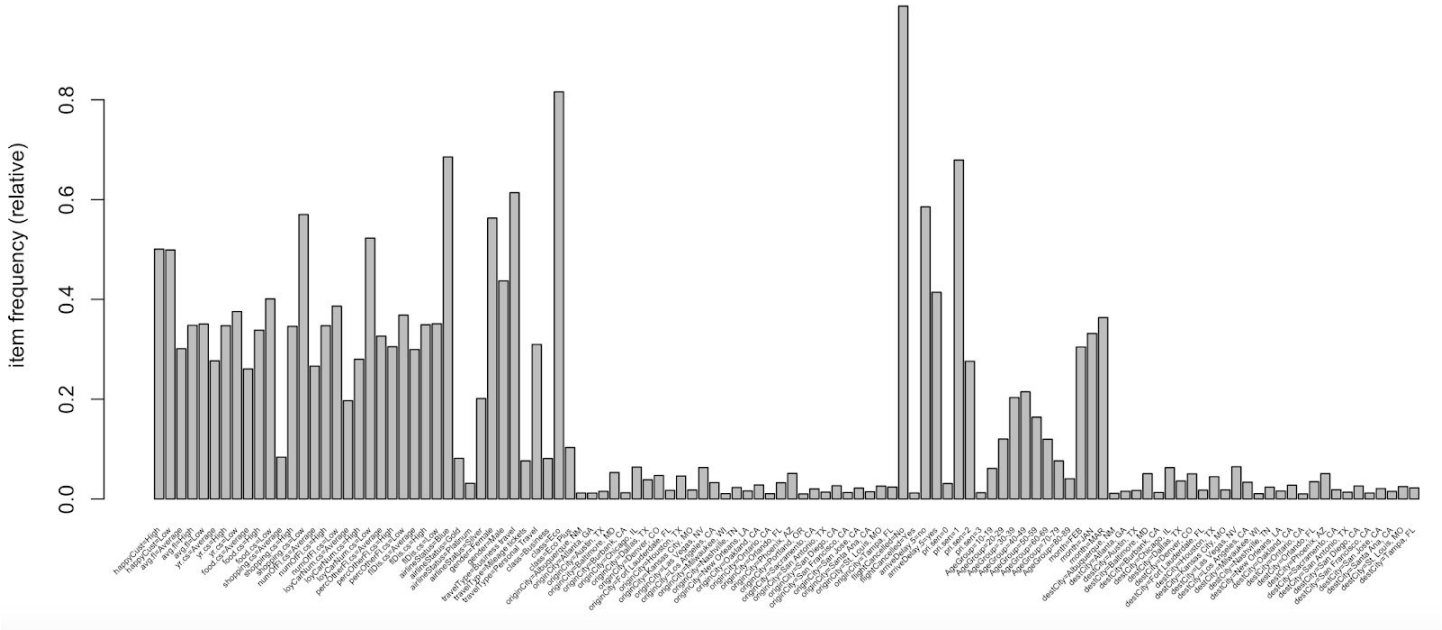


Figure 46: Item Frequency matrix with City

Figure 47: Item Frequency matrix without city

Next, we started apriori with support=0.1 and confidence = 0.1 and it generates 15731 rules. That is too many of rules for us to look at, so we narrowed it down by increasing support and confidence to 0.2, and the results of 1724 rules are as follows. The bright red dots will be the ones that we need to look into further because those are high in lift value. After trying a variety of combination of support and confidence, we decide to stick to 0.2 for both and dive into the model (the right figure).

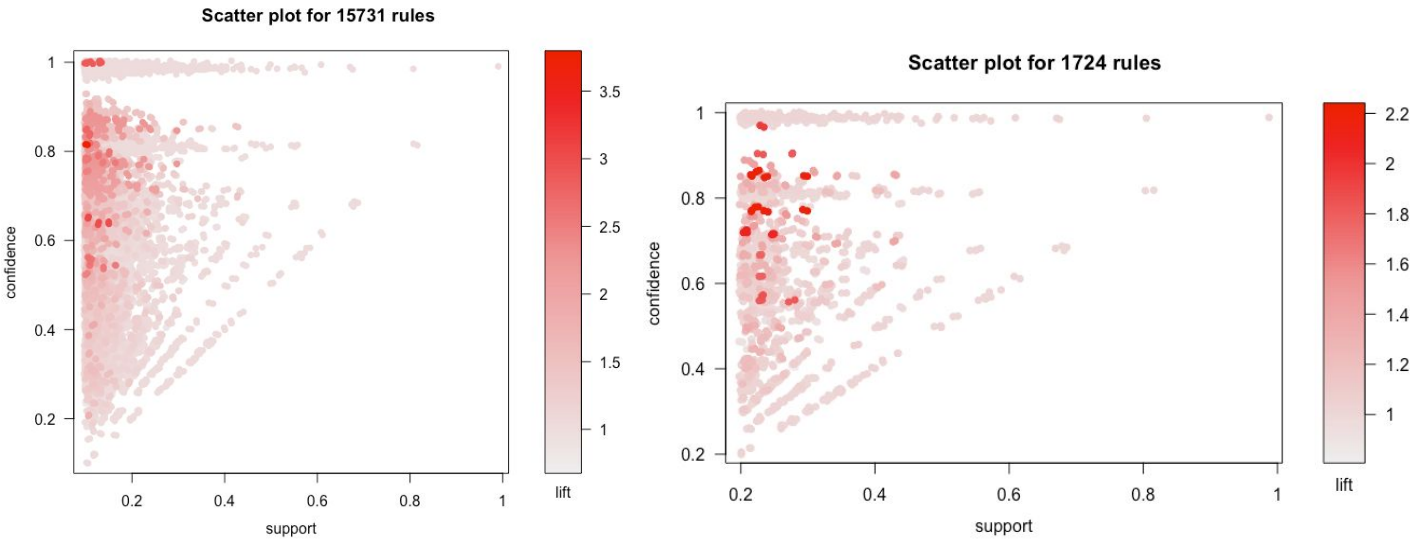


Figure 48: scatter plot for different models we generated - (left) support and confident = 0.1; (right) = 0.2

We can quickly take a look at the red dots by sorting the rules by lift - in the one with original city and destination city, interestingly, we saw original city = Milwaukee, WI, and destination city = seattle,

WA in the list. And we found that both cities are in relatively lower in average satisfaction rating compared to other cities.

```
> inspect(head(airlinerule.cs1, n = 5, by ="lift"))
     lhs                                                                rhs                                support    confidence lift     count
[1] {travelType=Business travel,flightCancelled=No,pri.sen=1} => {happyCust=High}                       0.3087344 0.7130196  1.423744  8045
[2] {travelType=Business travel,pri.sen=1}                    => {happyCust=High}                       0.3108834 0.7114878  1.420686  8101
[3] {happyCust=High,flightCancelled=No,pri.sen=1}             => {travelType=Business travel}           0.3087344 0.8634754  1.406629  8045
[4] {happyCust=High,pri.sen=1}                                => {travelType=Business travel}           0.3108834 0.8625426  1.405110  8101
[5] {travelType=Business travel,flightCancelled=No}           => {happyCust=High}                       0.4256658 0.6996342  1.397017 11092
```

Figure 49: quick glimpse of the first few rules of the red dots (without city)

```
    items                                transactionID
[1] {happyCust=Low,
     avg.fl=Low,
     yr.cs=High,
     food.cs=Average,
     shopping.cs=Low,
     numOfFl.cs=Low,
     loyCarNum.cs=High,
     percOtherFl.cs=High,
     flDis.cs=High,
     airlineStatus=Blue,
     gender=Female,
     travelType=Personal Travel,
     class=Business,
     originCity=Milwaukee, WI,
     flightCancelled=No,
     arriveDelay.5=yes,
     pri.sen=1,
     AgeGroup=30-39,
     month=MAR,
     destCity=Seattle, WA}                1

[2] {happyCust=High,
     avg.fl=Low,
     yr.cs=High,
     food.cs=High,
     shopping.cs=Average,
     numOfFl.cs=Low,
     loyCarNum.cs=Low,
     percOtherFl.cs=Low,
     flDis.cs=High,
     airlineStatus=Silver,
     gender=Female,
     travelType=Business travel,
     class=Eco,
     originCity=Milwaukee, WI,
     flightCancelled=No,
     arriveDelay.5=yes,
     pri.sen=1,
     AgeGroup=50-59,
     month=JAN,
     destCity=Seattle, WA}                2

[3] {happyCust=High,
     avg.fl=Low,
     yr.cs=High,
     food.cs=Average,
     shopping.cs=High,
     numOfFl.cs=Low,
     loyCarNum.cs=Low,
     percOtherFl.cs=Low,
     flDis.cs=High,
     airlineStatus=Blue,
     gender=Female,
     travelType=Business travel,
     class=Eco,
     originCity=Milwaukee, WI,
     flightCancelled=No,
     arriveDelay.5=no,
     pri.sen=2,
     AgeGroup=20-29,
     month=JAN,
     destCity=Seattle, WA}                3
```

Figure 50: quick glimpse of the first few rules of the red dots (with city)

Except for the scatter plot, we also ran grouped matrix to see how our model performs in general. We collect them with high customer ratings here, you can see some of them are closely related to one another. Obviously, since the variables are related, you will see some of them lead to obvious results.

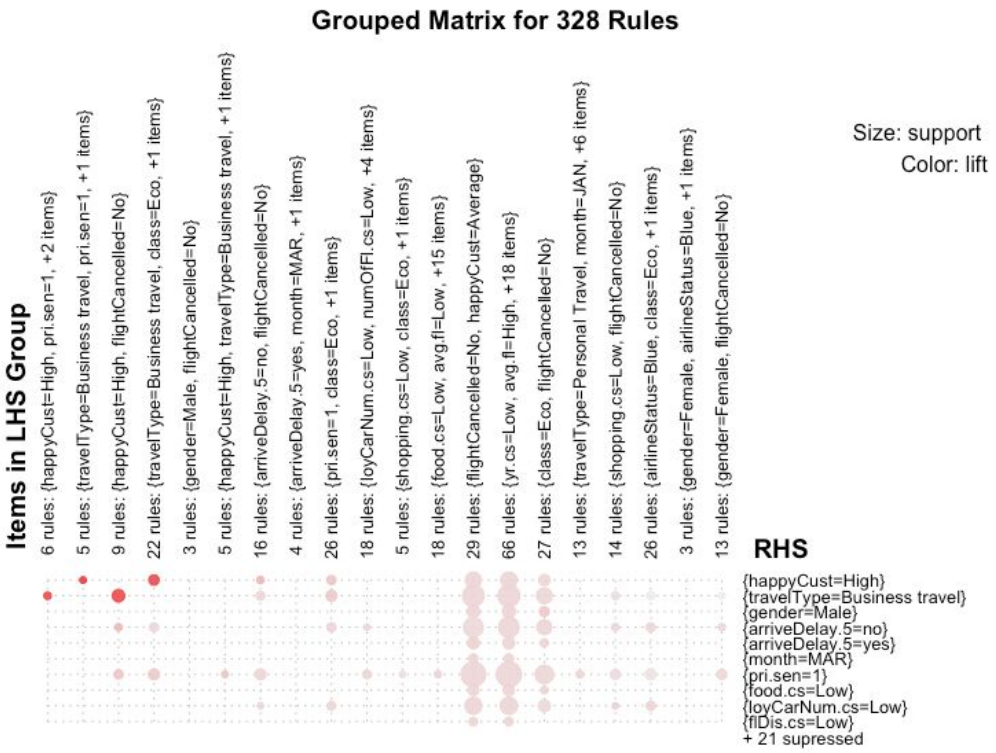


Figure 51: general grouped matrix for the model

All these rules above merely give a general idea of how each option connect to one another. However, our goal is to increase overall customer satisfaction ratings, in order to do so, we further focused on what makes high customer ratings and low customer ratings. We further separated our model into two groups - one is high customer ratings - HappyCust = high, and low customer ratings, HappyCust = low. We analyzed them by setting the rhs to customer rating high or low. We generated grouped matrix by limiting lift is larger than quartile 3. We started with happy customer group and found high customer ratings are mostly related to business travel.

We tried to focus on why might lead to higher customer ratings, so we tried various combination of support and confidence and finally landed at support=0.15, confidence=0.15, which generates about 180 rules for us. Below you can find a paracoord graph that shows most related items that can cause high customer ratings.

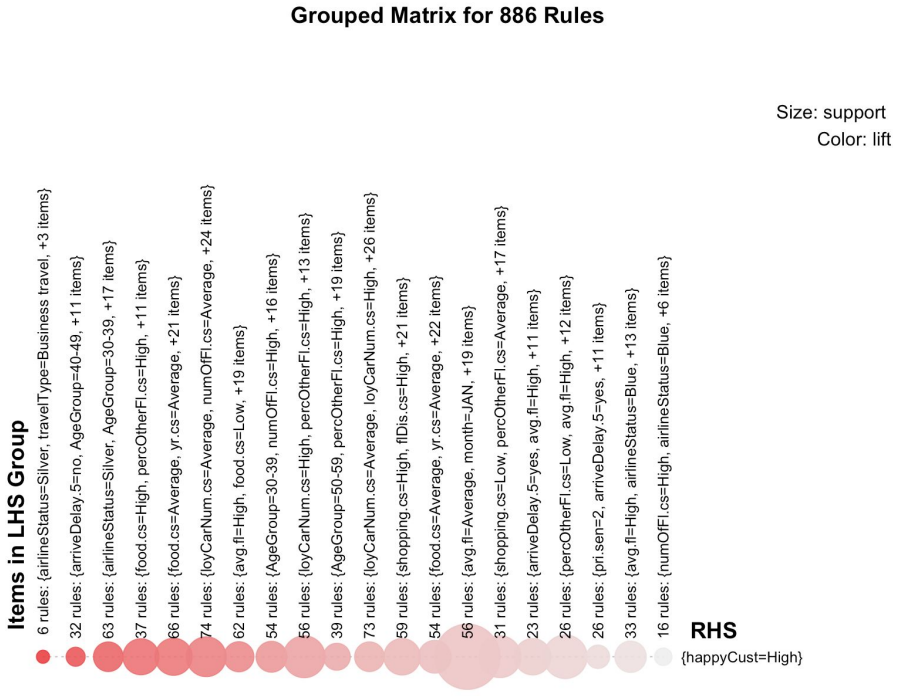


Figure 52: grouped matrix for rhs = high customer ratings

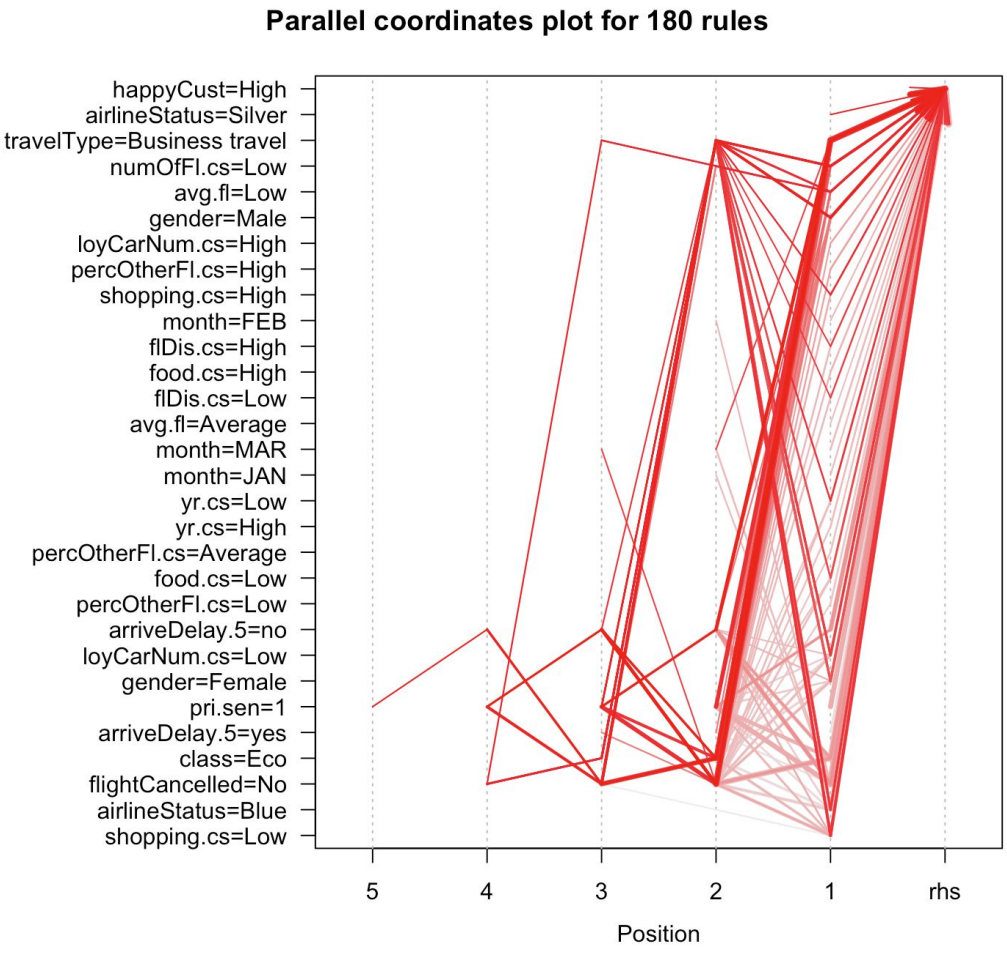


Figure 53: Paracoord graph showing what might cause high customer ratings

Again, we further investigated by limiting the lift that is above quartile 3, which generated only 45 rules left. Here are a few common findings that could cause high customer ratings: (1) males generally give higher ratings than females (2) business travel (3) no delay over 5 minutes and no cancel (4) age group between 30 to 49 (5) customers' number of flights is low (6) class is eco (7) price sensitivity level =1 (8) airline status is silver (9) membership card is high and spending on shopping is also high.

```
> inspect(head(top.lift, n=15))
     lhs                                                                                                  rhs              support   confidence lift     count
[1]  {travelType=Business travel,flightCancelled=No,arriveDelay.5=no,pri.sen=1}                           => {happyCust=High} 0.1923402 0.7671820  1.531895 5012
[2]  {travelType=Business travel,class=Eco,flightCancelled=No,arriveDelay.5=no,pri.sen=1}                 => {happyCust=High} 0.1546166 0.7639363  1.525414 4029
[3]  {travelType=Business travel,arriveDelay.5=no,pri.sen=1}                                             => {happyCust=High} 0.1944892 0.7637131  1.524968 5068
[4]  {travelType=Business travel,flightCancelled=No,arriveDelay.5=no}                                    => {happyCust=High} 0.2674035 0.7617798  1.521108 6968
[5]  {travelType=Business travel,class=Eco,arriveDelay.5=no,pri.sen=1}                                   => {happyCust=High} 0.1566122 0.7606710  1.518894 4081
[6]  {travelType=Business travel,class=Eco,flightCancelled=No,arriveDelay.5=no}                          => {happyCust=High} 0.2150203 0.7592141  1.515985 5603
[7]  {travelType=Business travel,arriveDelay.5=no}                                                       => {happyCust=High} 0.2703201 0.7583163  1.514192 7044
[8]  {travelType=Business travel,class=Eco,arriveDelay.5=no}                                             => {happyCust=High} 0.2177450 0.7558279  1.509223 5674
[9]  {airlineStatus=Silver}                                                                              => {happyCust=High} 0.1510477 0.7500000  1.497586 3936
[10] {gender=Male,travelType=Business travel,flightCancelled=No,pri.sen=1}                               => {happyCust=High} 0.1601811 0.7499102  1.497407 4174
[11] {gender=Male,travelType=Business travel,pri.sen=1}                                                  => {happyCust=High} 0.1609103 0.7474153  1.492425 4193
[12] {numOfFl.cs=Low,travelType=Business travel,flightCancelled=No,pri.sen=1}                            => {happyCust=High} 0.1612557 0.7416167  1.480847 4202
[13] {numOfFl.cs=Low,travelType=Business travel,pri.sen=1}                                               => {happyCust=High} 0.1623686 0.7407213  1.479059 4231
[14] {gender=Male,travelType=Business travel,flightCancelled=No}                                         => {happyCust=High} 0.2183590 0.7340988  1.465835 5690
[15] {gender=Male,travelType=Business travel}                                                            => {happyCust=High} 0.2195103 0.7322069  1.462057 5720
```

Figure 54: screenshot of result of first 15 items in the model of 45 rules

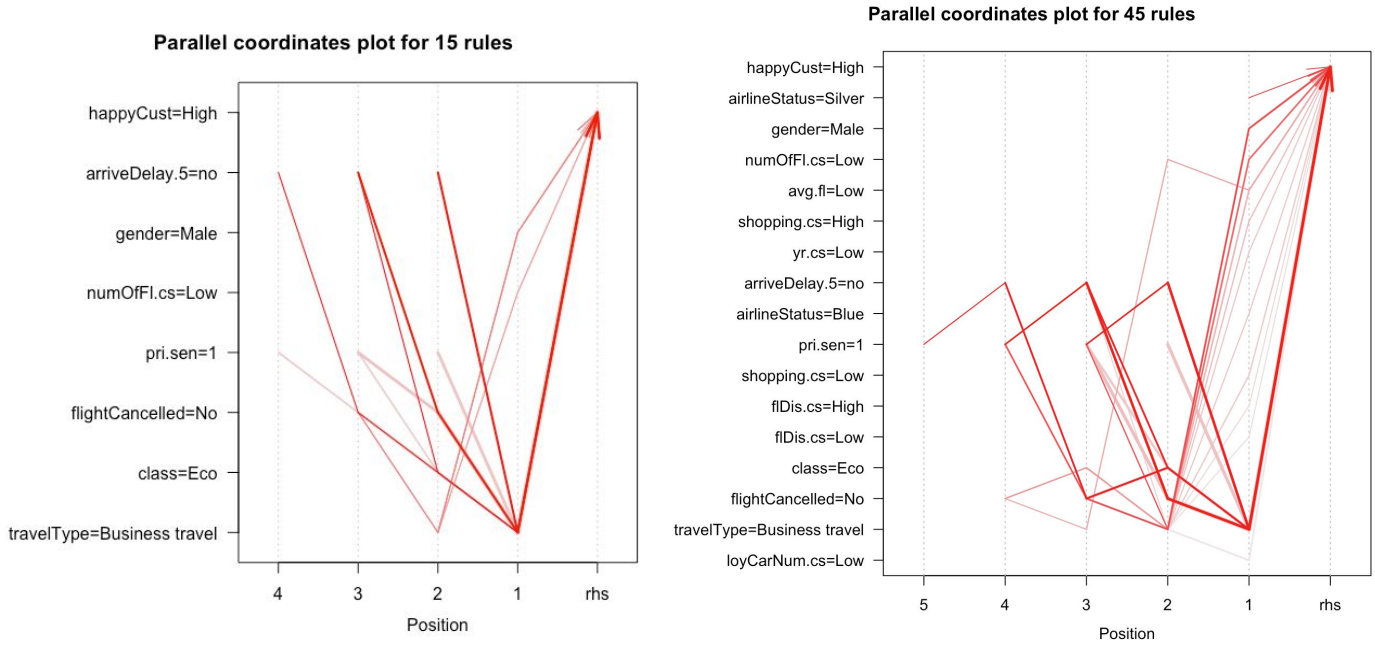


Figure 55: Paracoord graph for first 15 items and the 45 rules

On the other hand, we followed similar pattern on low customer ratings, and here is the Paracoord graph of low customer ratings to quickly get an idea of what the model is.

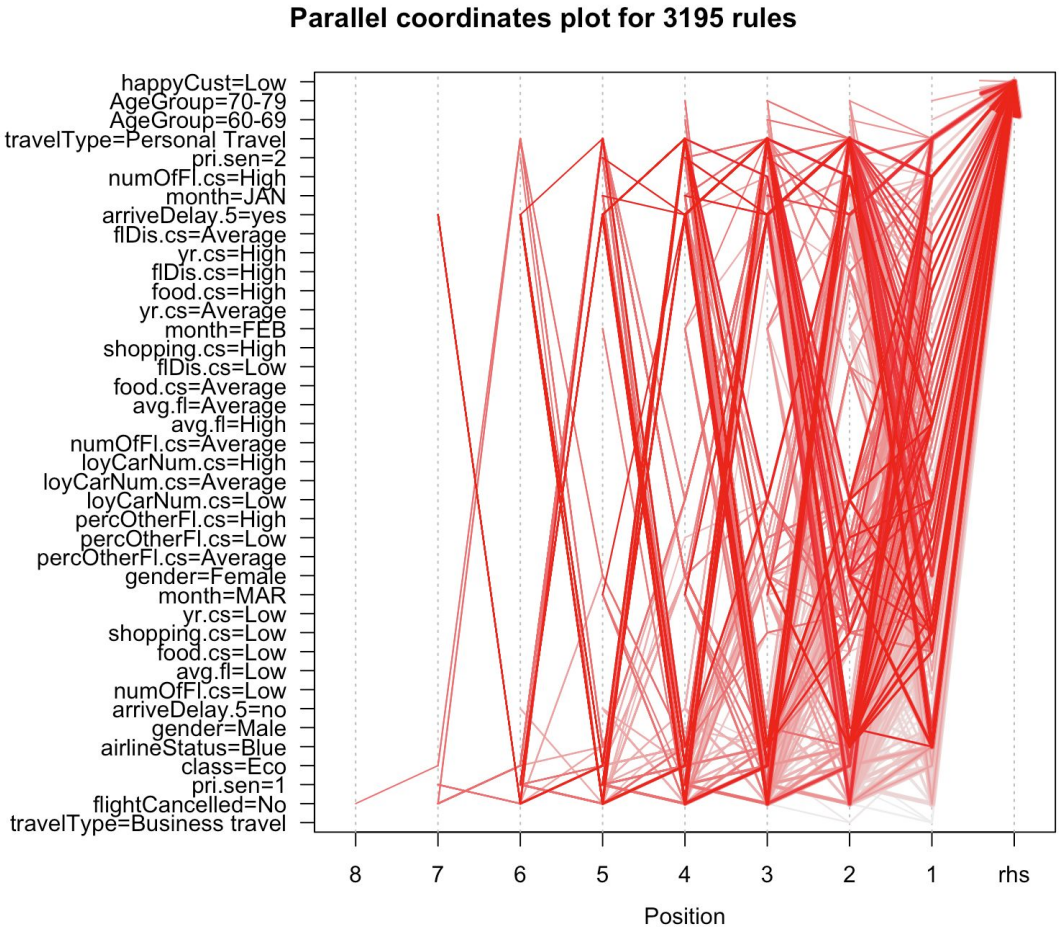


Figure 56: Paracoord graph of low customer ratings

We followed the similar steps to narrowed down our options by adjusting the support and confidence, and limiting the lift to be above third quartile 3 (Q3) to extract the top 15 rules (as follows). Finally, we landed at support=0.12, confidence=0.12 and set rhs to low customer ratings. We narrowed it down by decreasing the number of rules from 374 to 175, and to 44 rules. Here are the findings that are associated to lower customer satisfaction ratings: (1) personal travel (2) with delays over 5 minutes, so usually not cancelled (3) have fewer loyal cards and spent less at shopping (4) airline status is blue (5) age is between 60-79 (6) price sensitivity level is 2 (7) number of flights is high (8) females generally give lower ratings than males.

	lhs		rhs	support	confidence	lift	count
[1]	{loyCarNum.cs=Low,airlineStatus=Blue,travelType=Personal Travel}	=>	{happyCust=Low}	0.1520838	0.9794859	1.962134	3963
[2]	{airlineStatus=Blue,gender=Female,travelType=Personal Travel}	=>	{happyCust=Low}	0.1572262	0.9792065	1.961575	4097
[3]	{airlineStatus=Blue,gender=Female,travelType=Personal Travel,flightCancelled=No}	=>	{happyCust=Low}	0.1542329	0.9788115	1.960783	4019
[4]	{airlineStatus=Blue,travelType=Personal Travel,class=Eco}	=>	{happyCust=Low}	0.1921099	0.9688407	1.940810	5006
[5]	{airlineStatus=Blue,travelType=Personal Travel,class=Eco,flightCancelled=No}	=>	{happyCust=Low}	0.1885026	0.9684543	1.940035	4912
[6]	{airlineStatus=Blue,travelType=Personal Travel}	=>	{happyCust=Low}	0.2324046	0.9683403	1.939807	6056
[7]	{airlineStatus=Blue,travelType=Personal Travel,flightCancelled=No}	=>	{happyCust=Low}	0.2280298	0.9679101	1.938945	5942
[8]	{numOfFl.cs=High,travelType=Personal Travel,flightCancelled=No}	=>	{happyCust=Low}	0.1500499	0.9408085	1.884655	3910
[9]	{numOfFl.cs=High,travelType=Personal Travel}	=>	{happyCust=Low}	0.1524292	0.9405636	1.884164	3972
[10]	{loyCarNum.cs=Low,travelType=Personal Travel,class=Eco}	=>	{happyCust=Low}	0.1504720	0.9174076	1.837777	3921
[11]	{loyCarNum.cs=Low,travelType=Personal Travel}	=>	{happyCust=Low}	0.1830916	0.9171473	1.837256	4771
[12]	{loyCarNum.cs=Low,travelType=Personal Travel,flightCancelled=No}	=>	{happyCust=Low}	0.1795226	0.9168953	1.836751	4678
[13]	{shopping.cs=Low,travelType=Personal Travel,flightCancelled=No}	=>	{happyCust=Low}	0.1689692	0.9150042	1.832963	4403
[14]	{shopping.cs=Low,travelType=Personal Travel}	=>	{happyCust=Low}	0.1721928	0.9147808	1.832515	4487
[15]	{gender=Female,travelType=Personal Travel}	=>	{happyCust=Low}	0.1877734	0.9127029	1.828353	4893

Figure 57: screenshot of first 15 rules that cause low customer ratings

Here you can see some other visualizations that show the criteria that may cause low customer ratings. As you can see most of them were mentioned above, and they are pretty consistent.

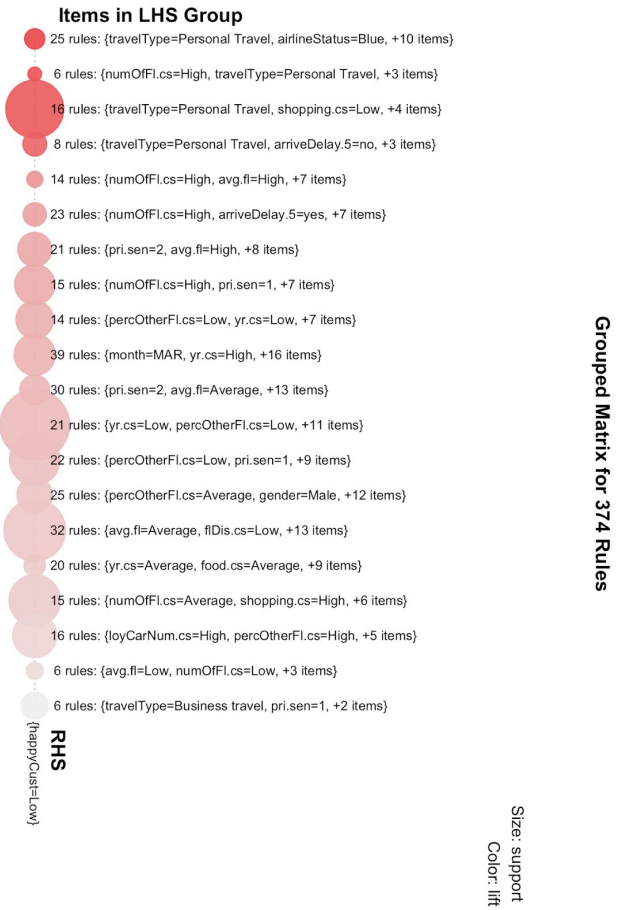


Figure 58: Grouped matrix for low customer ratings (374 rules)

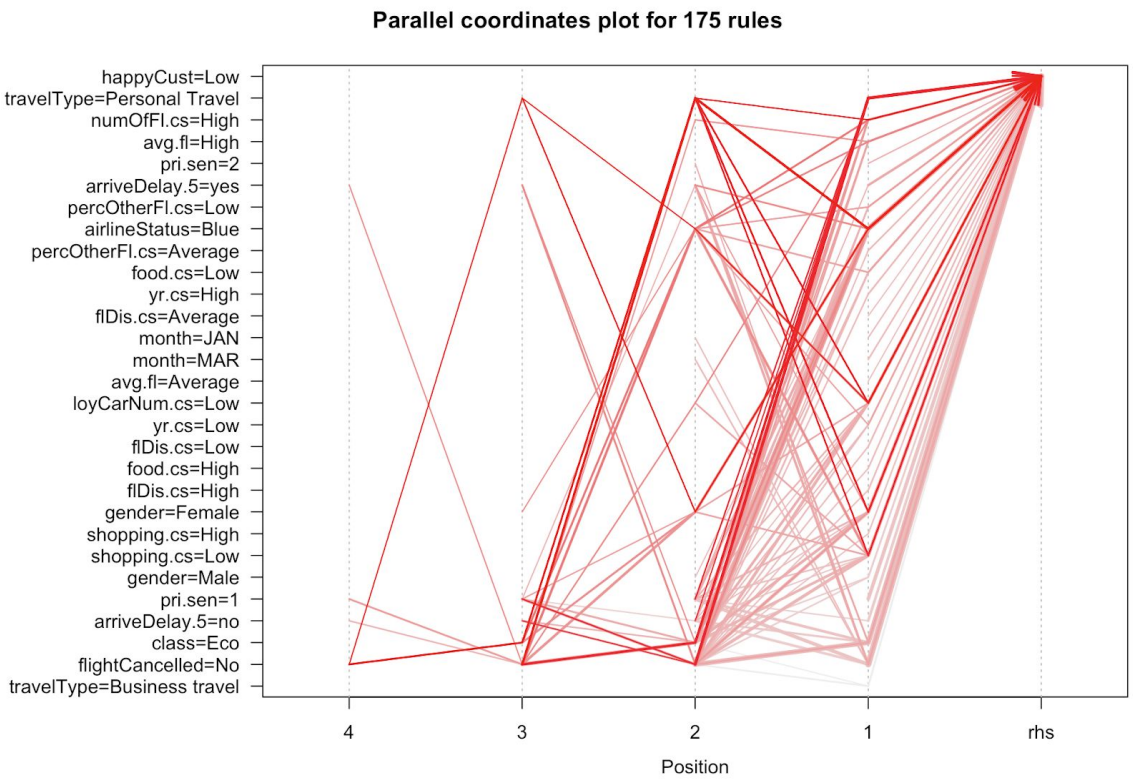


Figure 59: Paracoord graph of low customer ratings (175 rules)

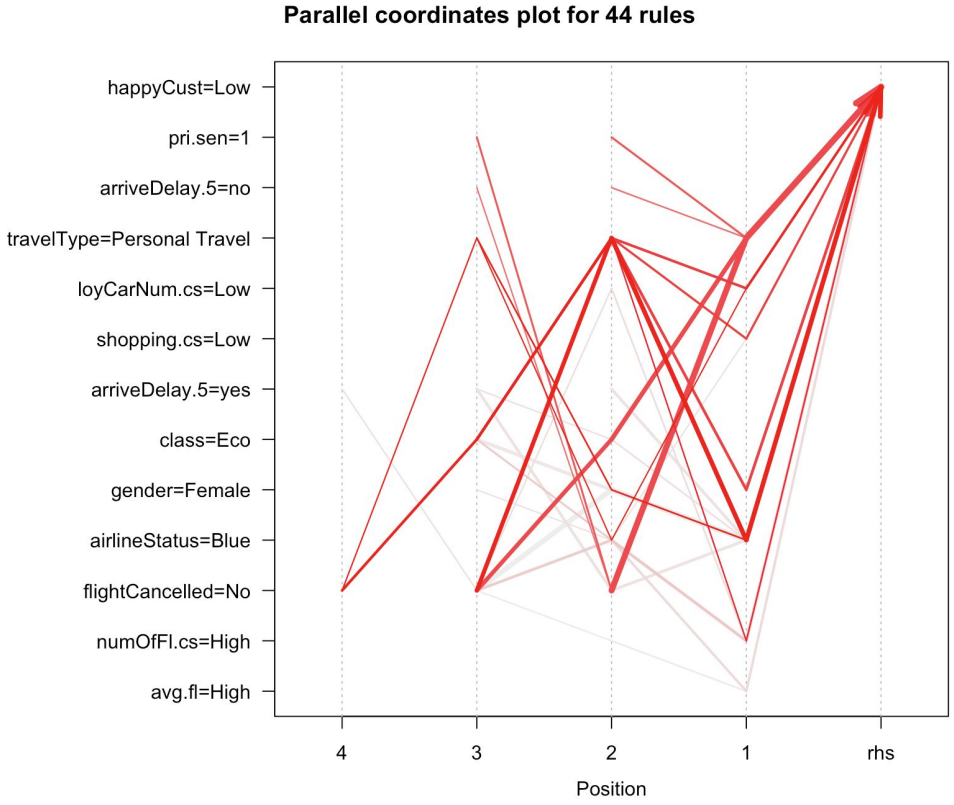


Figure 60: Paracoord graph of low customer ratings (44 rules)

Association Rule Conclusion

Therefore, based on the previous findings, we can conclude that (1) customers generally give higher ratings during business travel rather than personal travel (2) airline status with silver rated higher than ones with airline status at blue (3) no cancellation or delays can generally cause higher ratings but it is not always the case (4) The number of memberships (loyal cards) for other shops is lower, leading to lower the spendings on the shopping, might cause lower satisfaction ratings (5) those who gave lower customer ratings tend to be more sensitive on price changes (higher on price sensitivity variable), in other words, they are less willing to pay more for air tickets than people who gave high ratings (6) the frequent a customer flys, the lower ratings he/she gives

Based on above conclusion, I found a few interesting insights:

1. Cheapseats does not do a good job at making the frequent flyers happy. What can we do to improve the experiences of a loyal customer?
2. Given the other loyalty cards are granted by the frequency a customer flys, those who gave lower ratings are more sensitive to price changes but are not given enough benefits of being loyal customers, i.e. higher airline status or more other loyalty cards. Should the granted benefits be reconsidered by the frequency a person flys?
3. Assuming the airline status is leveled by the number of miles a customer flys, so although the more a customer flys, the less benefits he/she may get. Is it possible to negotiate with our business partners to provide more benefits to those frequent flyers? i.e. discounts on other shops etc.

```
> mean(silver.cs$num.flightbyYr)
[1] 3.428731
There were 32 warnings (use warnin
> mean(blue.cs$num.flightbyYr)
[1] 4.157977
```

D) Kernel Support Vector Machines (KSVM)

Using West Airlines as a reference, as it contains the least data points, the following variables were run against customer satisfaction. For the purpose of performing this model, customer satisfaction is broken into happy customers and unhappy customers. A happy customer is anyone who has given a rating of 4 or higher, which makes up about 56.5% of the total population. The following variables were run through the model against happy and unhappy customers.

1.airlineStatus	2. age	3. travelType	4. arriveDelay	5. numberFlights	6. originCity
7. destCity	8. date	9. day	10. foodAirport	11. class	

The output of the model is as follows:

Number of Support Vectors : 715 Training error : 0.072 Cross validation error : 0.221333	westtestData.happyCustomer 0 1 FALSE 152 104 TRUE 35 272

Insight:

The model gives an error rating of: 152+272 / 563 = **75.3%**. This error rating is high and is concluded that the model is not very useful for predicting a happy customer. This rating tells two stories, (1) the data set is not large enough and (2) we need to evaluate the model more closely see if we can decrease the error for better prediction accuracy. The West Airlines data set was used to run one more KSVM model for reference, with higher cost, but the cross validation error rate remained around 22%. We determined that using the West Airlines as a reference for model prediction is not good due to the small size of sample data. For the remaining data models within KSVM, we will only use Cheapseats Airlines.

Happy Customers

Several KSVM models were performed on the Cheapseats Airline dataset. The output of the models follows in the table below.

CheapOutput	CheapOutput2	CheapOutput3
Number of Support Vectors : 10096 Training error : 0.078172 Cross validation error : 0.220758	Number of Support Vectors : 10144 Training error : 0.101312 Cross validation error : 0.23653	Number of Support Vectors : 8702 Training error : 0.175627 Cross validation error : 0.231176

Assessing the three models listed above, the first model will be used for further prediction of happy customers. This model is selected because the training error is the smallest and it has a roughly 78% rate of predicting happy customers. The graphic below is a histogram of CheapOutput and shows a large portion of the data has cost = 5 or near 5. We will look as several values along the left side of the histogram, focusing on the range < 0.1. There are approx 111 index values that show a happy customer. We will evaluate 25 of them.

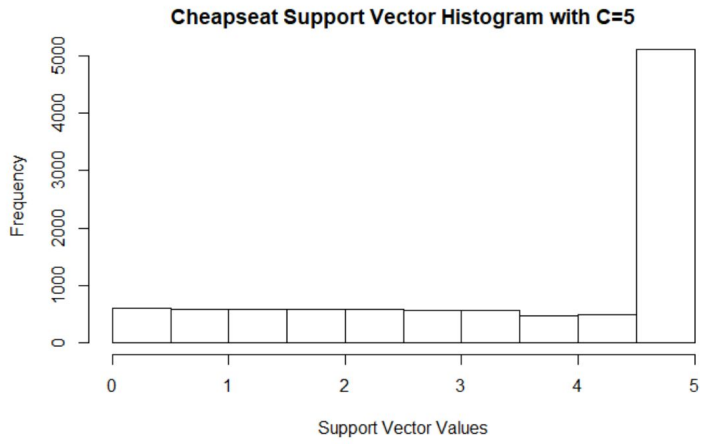


Figure 61: KSVM Histogram of Happy Customers

happyCustomer	0	1
FALSE	3026	1297
TRUE	578	3785

Ratio of .216, or 21.6% error, giving an accuracy rating of 78.4%

Figure 62: Happy Customer Error Ratio

The following data represents the first prediction of a happy customer. This customer was delayed, yet still gave a happy rating. The indicator that stands out is they were flying a business travel type. We will evaluate more customers to see what overlaps.

Customer 1

satisfaction	airlineStatus	age	gender	priceSensitivity	firstYearFlight	numberOfFlights	percentOfOtherFlight	travelType	loyaltyCardsNum	shoppingAtAirport	foodAtAirport	class	day	date
4	Gold	61	Female	1	2011	13	7	Business travel	1	280	45	Eco	11	3/11/2014

airlineName	originCity	originState	destCity	destState	scheduledDepHr	scheduledDepDelay	arriveDelayMinute	flightCancelled	flightTimeMinute	flightDistance	arriveDelay>5	happyCustomer
Cheapseats Airlines Inc.	Denver, CO	Colorado	Tulsa, OK	Oklahoma	17	24	20	No	67	541	yes	TRUE

Customer 2

satisfaction	airlineStatus	age	gender	priceSensitivity	firstYearFlight	numberOfFlights	percentOfOtherFlight	travelType	loyaltyCardsNum	shoppingAtAirport	foodAtAirport	class	day	date
4	Blue	44	Male	1	2003	7	5	Business travel	0	0	30	Eco	21	1/21/2014

originCity	originState	destCity	destState	scheduledDepHr	scheduledDepDelay	arriveDelayMinute	flightCancelled	flightTimeMinute	flightDistance	arriveDelay>5	happyCustomer
Baltimore, MD	Maryland	Los Angeles, CA	California	10	24	39	No	309	2329	yes	TRUE

Similarities between customer 1 and customer 2 show they both were traveling business class and had a price sensitivity of 1, and flew economy class. More customers will be evaluated and the results will be listed. 25 index numbers of happyCustomer were selected and 21 of the 25 returned a happy customer. Evaluating the similarities between the happy customers, the following describe what predicts a happy customer:

Airline Status = Blue, Age = 22 - 61, Price Sensitivity = 1, number of flights = 10 - 20, travel type = business travel, class = eco, and flight is NOT canceled, even though delayed.

Of the 4 unhappy customers falsely predicted to be happy, one customer was travel person and had a class of business. There are no other indications of why the customer rated the flight with a satisfaction score = 1. Next, we will evaluate unHappy customers to see if we can predict what makes a customer give a rating of 3 or lower.

Unhappy Customers

3 KSVM Models were run for unhappy customers where the satisfaction score is less than or equal to 3. Of the three models, the UnHappyCustomer2 data will be analyzed. Analyzing the the ration for predicition failure, UnHappyCustomer2 gives an 80.4% prediction success rate.

UnHappyCustomer	UnHappyCustomer1	UnHappyCustomer2
Number of Support Vectors : 9987 Training error : 0.1035 Cross validation error : 0.233076	Number of Support Vectors : 10081 Training error : 0.101888 Cross validation error : 0.234458	Number of Support Vectors : 8392 Training error : 0.184895 Cross validation error : 0.227895

unHappyCustomer2

	0	1
FALSE	4025	338
TRUE	1367	2956

Ratio of .196, or 19.6% error, giving an accuracy rating of 80.4%

Figure 63: unHappy Customer Error Ratio

Using the ratio, we will evaluate how what the model predicted as unhappy customers. Using the same cost = 5, we focus on the range of predicted values with score < 0.1. The return of values is 98 predictions of not happy customers. Analyzing the returns on the 98, the following are the values the model predicted are most to make an unhappy customer:

Airline Status = Blue, Flight Class = Eco, Travel Type = Personal Travel, Average Number of Previous flight = 20 or more, Age = 44 years or older, gender = female, Price Sensitivity = 2, Flight Time < 120 min.

Happy vs UnHappy Customers

Analyzing both happy and unhappy customers, the primary factors that stand out between the two that make a customer unhappy or give a satisfaction rating lower than 3 are:

(1) Travel Type = Personal
(2) Have a history of more than 20 flights
(3) Show a price Sensitivity of 2

6. Data Question Answers

1. **Data Question # 1 - Why CheapSeats Airlines?**

Cheapseats has the most observations (about 20% of the whole dataset) and has relatively lower average satisfaction score compared to other airlines (about 3.7% lower than the highest rating in this survey). The Cheapseats' average satifaction rating is lower than first quartile among all airlines (see below). As our goal is to increase customer ratings, analyzing the group with low ratings can help generate more insights. In addition, based on our descriptive analysis, the characteristics are very similar to the whole dataset, so we can confidently say that Cheapseats is representative of the overall airline performance.

```
> summary(airlineName$avg_sat)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.297   3.358   3.395   3.388   3.399   3.487
```

Figure 64: Satisfaction scores analysis among airlines

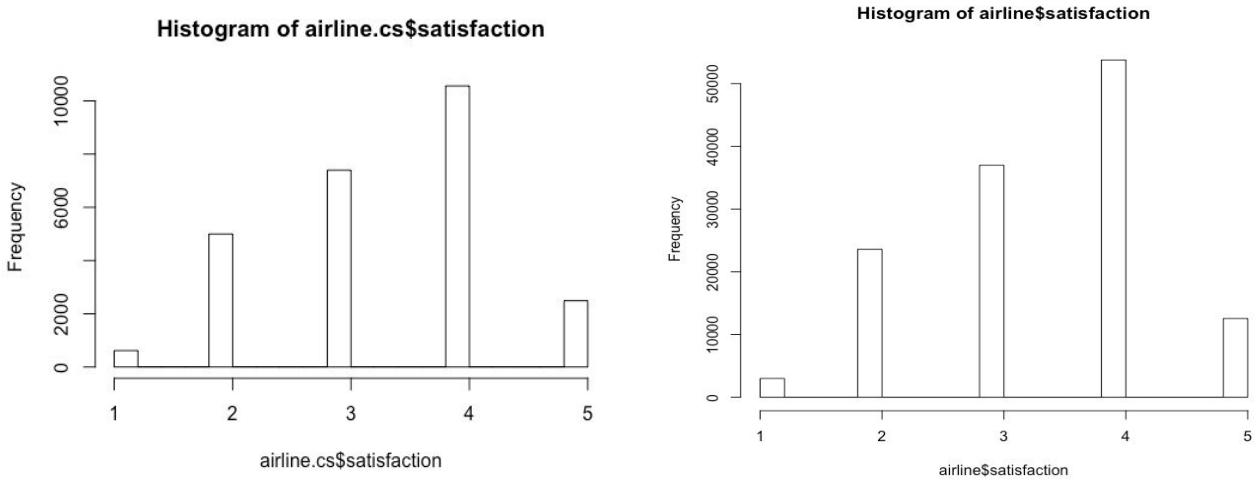


Figure 65: Histogram of satisfaction of Cheapseat airline (left) vs. overall satisfaction (right)

2. Data Question # 2 - Is age a factor in lower ratings or higher ratings within the data set?

Age is definitely a factor in predicting the satisfaction score. The models clearly show there multiple factors related to age. To illustrate the factor, the following to bar charts illustrate the distribution of ages to their satisfaction. To understand the discrepancy between ages, we break up the ages into two groups. Group one is classified as ages 10-19 and 60 - older ad group two as 20 - 59. As seen in the bar charts below, group one has significantly lower scores. The breakup of Group one makes up almost 26% of the overall population. Thus, we conclude that yes, ages is a factor in the rating of satisfaction for Cheapseats Airlines.

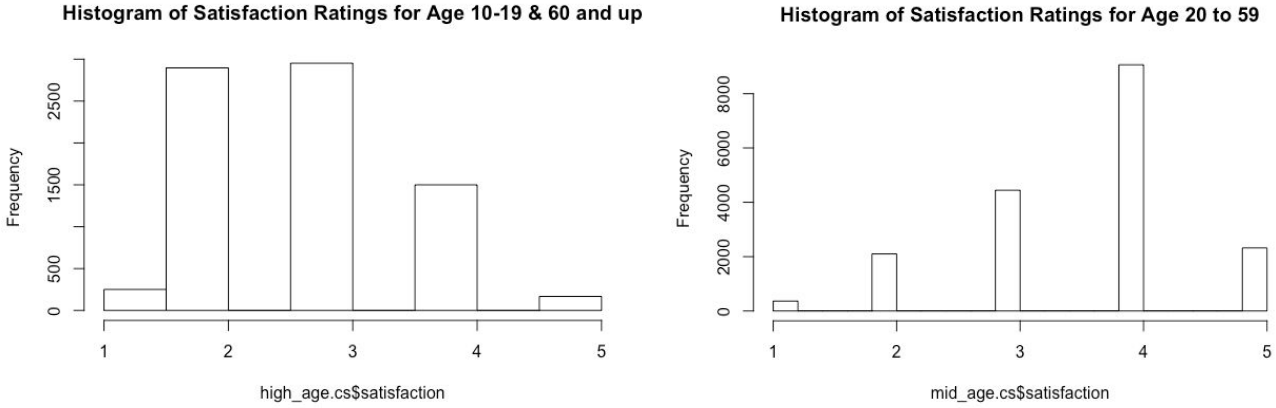


Figure 67: Histogram of satisfaction of customer age 10-19 and 60 and up(left) vs. 20-69 (right)

3. Data Question # 3 - Does a persons travel type an effect on the rating they will give during the survey?

Travel type stood out in two of the three models. The majority of travelers on Cheapseats were of the travel type blue. After running the models and performing the analysis, we later discovered that the travel type in conjunction with other factors directly affects the rating a customer gives. Interpreting the scores of airline status with age groups, it is seen that Blue customers generally give the lowest score.

Adding the factor of age into the description, the table on the right shows that Blue customers give the lowest average satisfaction rating. Thus, we conclude, yes, the travel type does affect the rating a rating of satisfaction.

airlineStatus	count	avg_sat	
10-19	Blue	1590	2.903145
10-19	Platinum	2	2.5
10-19	Silver	2	3.5
20-29	Blue	2348	3.321124
20-29	Gold	125	3.84
20-29	Platinum	23	4.043478
20-29	Silver	634	3.865931
30-39	Blue	3094	3.379767
30-39	Gold	566	3.95583
30-39	Platinum	204	3.990196
30-39	Silver	1431	4.013277
40-49	Blue	3439	3.443734
40-49	Gold	520	4.011538
40-49	Platinum	226	3.933628
40-49	Silver	1414	4.08628
50-59	Blue	2798	3.310936
50-59	Gold	438	3.899543
50-59	Platinum	174	3.931034
50-59	Silver	857	3.989498
60-69	Blue	2197	2.788803
60-69	Gold	287	3.198606
60-69	Platinum	121	3
60-69	Silver	517	3.675048
70-79	Blue	1540	2.454545
70-79	Gold	139	2.546763
70-79	Platinum	53	2.396226
70-79	Silver	262	3.492366
80-89	Blue	858	2.356643
80-89	Gold	48	2.4375
80-89	Platinum	20	2.1
80-89	Silver	131	3.40458

4. **Data Question # 4 - How many flights are there per months / day and do delayed flights per month have a significant effect on customer ratings?**

In the regression model, the R-square for date is not significant. However, from the descriptive analysis, it is interesting to see that February has more flights that are above the overall average ratings than below, whereas flights in March and January have more flights that are below average ratings. You can see the average number of flights per day by month below. Also, the month itself should not affect the average satisfaction ratings as much as delays do.

month	count	avg_sat	avg_flight	above avg	below avg
JAN	8641	3.346719	278.7419355	0.4516129032	0.5483870968
FEB	7937	3.373567	283.4642857	0.6785714286	0.3214285714
MAR	9480	3.353376	305.8064516	0.4838709677	0.5161290323

	arriveDelay.5 <fct>	month <chr>	count <int>	avg_sat <dbl>	Percent <dbl>
1	no	FEB	4815	3.50	18.5
2	no	JAN	4820	3.48	18.5
3	no	MAR	5628	3.49	21.6
4	yes	FEB	3122	3.18	12.0
5	yes	JAN	3821	3.18	14.7
6	yes	MAR	3852	3.15	14.8

5. **Data Question # 4 - Is there a gender bias in customer satisfaction ratings?**

Yes, according to the descriptive analysis and association rule model, women usually give out lower ratings than men. Also, there are more women (56.3%) than men (43.7%) in this survey data. Also, in the regression model, it is only less than 2% in the R-square, so it might not be as important.

	gender <fct>	count <int>	avg_sat <dbl>	Percent <dbl>
1	Female	14666	3.24	56.3
2	Male	11392	3.51	43.7

6. **Data Question # 6 - What class is most often traveled and does the class have an effect on the customer rating?**

In the regression model, the R-squared is around 0.002196, which means it has very minimum impact on customer satisfaction. It has to combine with other variables to weight more significantly. Also, based on our descriptive analysis, class Eco accounted for 81.6% of the observations, yet the average satisfaction rating is just similar to overall ratings. So, we could assume that customers do not give out lower ratings because it is the lowest level of the seat, however, those who took business class did give

out higher ratings.

	class	count	avg_sat	Percent
	<fct>	<int>	<dbl>	<dbl>
1	Business	2111	3.51	8.10
2	Eco	21261	3.35	81.6
3	Eco Plus	2686	3.30	10.3

7. **Data Question # 7 - Does the airport location of the airport have an impact overall customer satisfaction? Which cities were traveled to most on any given day and is there a pattern of lower ratings based on the most visited city?**

Based on the linear regression model we did above, the airport location (city) does not have significant effects on the satisfaction ratings statistically. The adjusted R-square is within 1% changes when added the original cities or destination cities.

Chicago IL, Las Vegas NV, Baltimore MD, Phoenix AZ, and Denver CO are the top 5 most visited original and destination cities, and their average ratings are relatively lower than overall average (3.357).

8. **Data Question # 8 - What is the distribution of satisfaction ratings per rating?**

This question should be able to be answered by the histogram in the descriptive analysis section, however, you can also see that those who rated 4 and above actually account for about 50% of the customers, and there are around 47% of people rated 2 and 3.

	satisfaction	count	Percent
	<int>	<int>	<dbl>
1	1	616	2.36
2	2	4996	19.2
3	3	7396	28.4
4	4	10561	40.5
5	5	2489	9.55

9. **Data Question # 9 - Does the number of times a person flys impact their overall satisfaction rating?**

Yes, as mentioned in the variable descriptive analysis section, you can see that customers, who took less than 5 flights a year, still rated mostly at 4, however, for the customers who flew more than 5 flights a year, you can see the histogram gradually shifts to rating 2 and 3, especially, the most frequent flyers rated mostly 3, which is below average rating. In our association rule, we can also see that the more a customer flys, the less happy a customer is. It is not hard to speculate that the more a customer flys, the more uncertainty events a customer could encounter. Therefore, customers who flys often are more likely to give out lower ratings given the higher chance of getting a bad experience.

10. **Data Question # 10 - Which travel type gives the highest satisfaction rating and how much does travelType affect the overall satisfaction score?**

Business travel type has the highest ratings (at about 3.76), whereas personal travel group yields about 2.52 as average satisfaction ratings. This is also be reaffirmed in the descriptive analysis and the association rule model. Furthermore, we may assess that the personal travel type weighs in the most on decreasing the overall satisfaction score as almost 50% of the personal travel types give a lower rating.

	travelType	count	avg_sat
	<fct>	<int>	<dbl>
1	Business travel	15996	3.76
2	Mileage tickets	1993	3.51
3	Personal Travel	8069	2.52

7. Overall Model Conclusions

Based on the analysis and the data above, Team Red Panda recommends the following in order to increase the customer satisfaction.

(1) Descriptive Statistics modeling shows that customers age, airlines status, travel type, and number of flights typically shows as most significant.

(2) Linear Regression model shows the following are significant and directly attributed to the satisfaction score: Age, Airline Status, Travel Type, Class, Number of Flights, and Arrival Delay Minute.

(3) Arules model shows lower customer satisfaction ratings: (a) personal travel (b) with delays over 5 minutes, so usually not cancelled (c) have fewer loyal cards and spent less at shopping (d) airline status is blue (e) age is between 60-79 (f) price sensitivity level is 2 (g) number of flights is high (h) females generally give lower ratings than males

(4) Arules model shows high customer ratings: (a) males generally give higher ratings than females (b) business travel (c) no delay over 5 minutes and no cancel (d) age group between 30 to 49 (e) customers' number of flights is low (f) class is eco (g) price sensitivity level =1 (h) airline status is silver (i) membership card is high and spending on shopping is also high.

- The number of memberships (loyal cards) for other shops is lower, leading to lower the spendings on the shopping, might cause lower satisfaction ratings
- Those who gave lower customer ratings tend to be more sensitive on price changes (higher on price sensitivity variable), in other words, they are less willing to pay more for air tickets than people who gave high ratings
- Inference:
  - (i) Cheapseats does not do a good job at making the frequent flyers happy. What can we do to improve the experiences of a loyal customer?
  - (ii) Given the other loyalty cards are granted by the frequency a customer flys, those who gave lower ratings are more sensitive to price changes but are not given enough benefits of being loyal customers, i.e. higher airline status or more other loyalty cards. Should the granted benefits be reconsidered by the frequency a person flys?

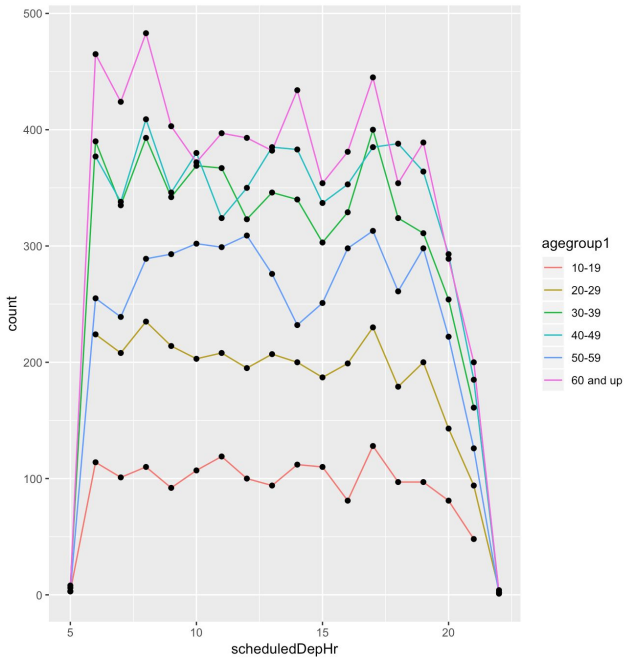
(5) Kernel Support Vector Machine shows that happy customers are typically Airline Status = Blue, Age = 22 - 61, Price Sensitivity = 1, number of flights = 10 - 20, travel type = business travel, class = eco, and flight is NOT canceled, even though delayed.

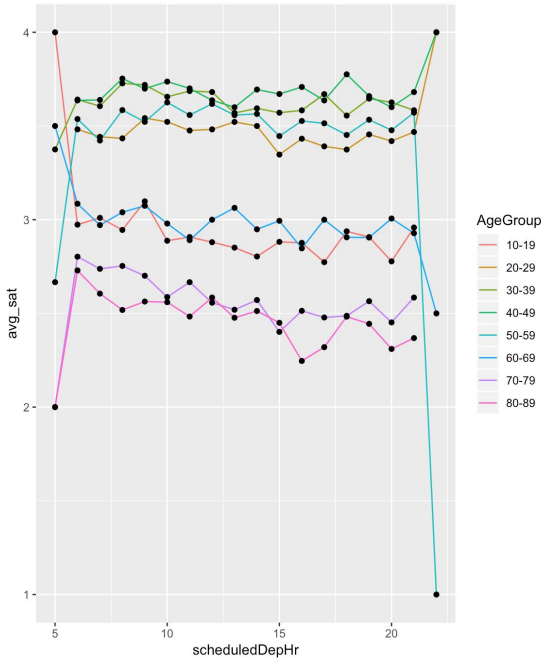
(6) Kernel Support Vector Machines shows that unhappy customers typically have: Airline Status = Blue, Flight Class = Eco, Travel Type = Personal Travel, Average Number of Previous flight = 20 or more, Age = 44 years or older, gender = female, Price Sensitivity = 2, Flight Time < 120 min.

8. Cheapseats Airlines Inc. Recommendations

(1) Age is a factor within Cheapseats. Customers who fall within the Age range of 10-19 or 60 and older give the lowest satisfaction rating. Cheapseats airlines should invest more to take care of younger and older customers.

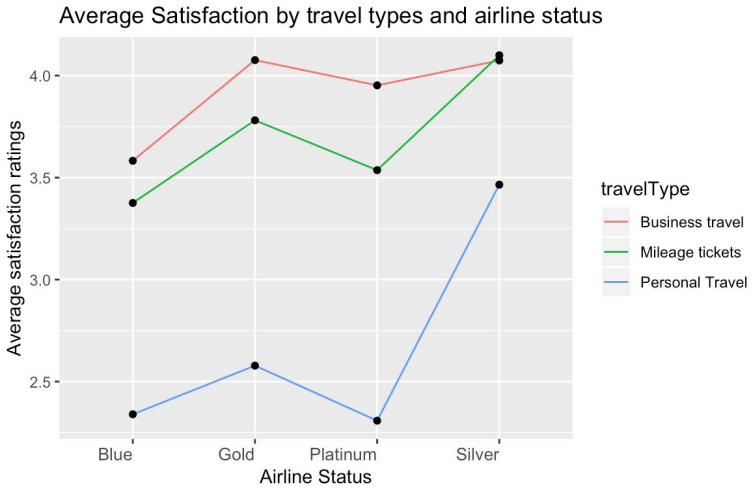
(a) We could find out what time throughout the day does people age 10-19 or 60 and above take flights often, and invest a bit more support during that time.





(b) provide quick questionnaire to ask at the time they check-in so we can assist better.

(2) Frequent flyers traveling on personal travel tend to give a lower rating. Cheapseats Airlines, per the data, appears to cater to only Business travel customers. The Airline should understand the needs of travelers flying via a personal travel type further.



(3) The models show that delayed departures directly affect the satisfaction of flyers which have a short flight time or shorter flight distance. Historically, shorter flight times are related to regional

flights to major airline hubs. The map at the beginning of the report shows more flights into and out of Houston, Dallas, San Francisco, and Los Angeles and Chicago. Recommend Cheapseat Airlines Inc focus on regional flights into and out of major hubs to reduce the delay of flights.

9. Validation

Team Red Panda assesses with a high level of confidence that our techniques in cleaning, manipulating, and analyzing the data are correct. While working with the survey data, the team often discovered about 3 weeks into the analysis errors in our calculations. Upon discovery, the team would discuss during the team meetings and collectively come to a conclusion on the errors. The factor / integer / character conversion proved to be challenging when working with the initial dataset. Issues with NULL values were less apparent in the Cheapseats Dataset as well. After coming to conclusion on our models and providing the recommendations above, the team began to find additional validation within the data by performing more descriptive statistics. These statistics focused on evaluating if our conclusions are correct. By using multiple models, our results are surprisingly similar across all the models. Our interpretations have data to back our findings and conclusions. Thus, by validating our validations, we conclude that our assessment above is correct and in full terms accurate.

10. Appendix - Please see our R-code in the attached file.

11. Indemnification

The Client agrees to indemnify, defend, and protect the Consultant from and against all lawsuits and costs of every kind pertaining to the Client's business including reasonable legal fees due to any act or failure to act by the Client based upon the Consulting Services.

12. No Modification Unless in Writing

No modification of this Agreement shall be valid unless in writing and agreed upon by both Parties.

13. Applicable Law

This Consulting Agreement and the interpretation of its terms shall be governed by and construed in accordance with the laws of the State of [State] and subject to the exclusive jurisdiction of the federal and state courts located in [County], [State].

IN WITNESS WHEREOF, each of the Parties has executed this Consulting Agreement, both Parties by its duly authorized officer, as of the day and year set forth below.

[Company]

_____	_____	_____
[First name]	[Last name]	[Title]

[Client]

_____	_____	_____
[First name]	[Last name]	[Title]