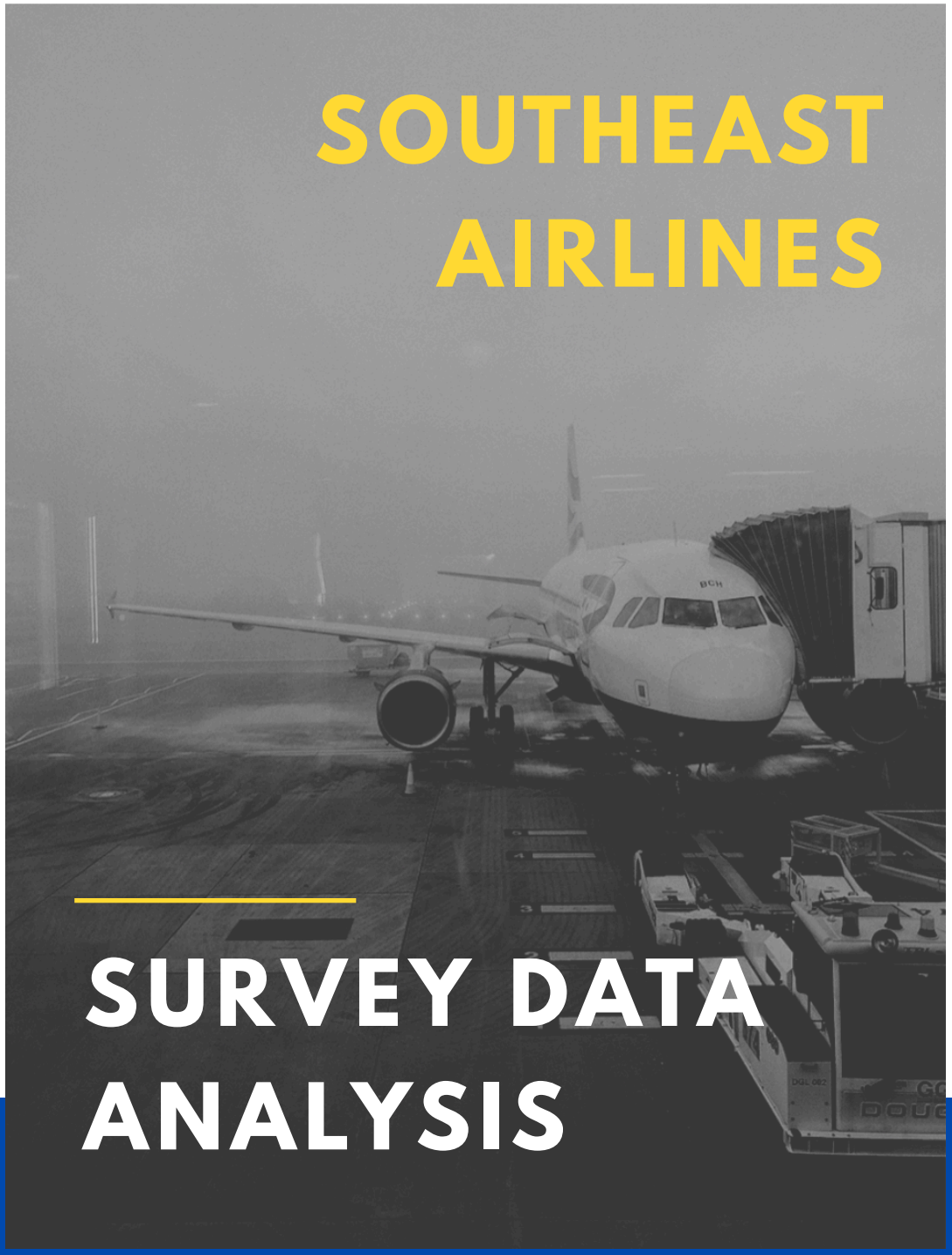


SOUTHEAST AIRLINES

SURVEY DATA ANALYSIS

DECEMBER 3RD, 2020 | IST 687

ZEYANG ZHOU
MICHELLE KE KINCAID
RAYANNA HARDUARSINGH
DIVYA BHUWANSINGH DAMAHE

CONTENTS

INTRODUCTION

As data analysts, we were presented with the challenge of improving Southeast Airline's customer churn. It was our duty to explore, analyze, and interpret the results of a vast survey taken by customers who traveled with Southeast Airlines. After receiving 88,000 survey responses from travelers who rated their overall satisfaction on a scale from 1-10, our main goal was to identify key patterns within this data and derive the best recommendation possible for Southeast Airlines to improve their customer satisfaction. We thoroughly analyzed key factors as well as the travelers to predict which factors have an effect on the satisfaction score.

This proposal consists of our deep analysis on the survey through several visualizations and modeling techniques that contributed to our insights and recommendations. These results can aid in the process of how Southeast Airlines can improve satisfaction with their current and future travelers.

BUSINESS QUESTIONS

After exploring the dataset and its attributes, we thought it was important to ask ourselves questions to keep in mind when doing our analysis. The following are important areas of interests, but were certainly were not limited to:

1. How can we improve NPS score?
2. How does NPS score vary by geographic location?
3. How does NPS score vary by age?
    1. Can we offer senior citizens (60-85) with a special treatment to ease their travel and help improve the NPS score?
4. Depending on the data, who is a better partner?
    1. Does Southeast Airlines need to adjust the number of their partners?
5. What factors contribute to low scores?
    1. Arrival/Departure Delay, Travel Type, etc.
6. What services might contribute to a customer being satisfied?
    1. Shopping, Eating, etc.

DATA ACQUISITION, CLEANSING, TRANSFORMATION,

Loading the Data

To load the dataset into RStudio, we installed the “jsonlite” package to convert the JSON format survey file into a readable file. We then transformed the survey file into a data frame called “airData”.

```
'data.frame':   88100 obs. of  32 variables:
 $ Destination.City             : chr  "Aberdeen, SD" "Aberdeen, SD" "Aberdeen, SD" "Aberdeen, SD" ...
 $ Origin.City                  : chr  "Minneapolis, MN" "Minneapolis, MN" "Minneapolis, MN" "Minneapolis, MN" ...
 $ Airline.Status               : chr  "Silver" "Gold" "Silver" "Blue" ...
 $ Age                          : int  52 27 52 39 63 43 52 50 64 78 ...
 $ Gender                       : chr  "Female" "Female" "Female" "Male" ...
 $ Price.Sensitivity            : int  1 1 1 1 2 1 1 2 0 1 ...
 $ Year.of.First.Flight         : int  2010 2010 2010 2006 2012 2006 2009 2010 2008 2007 ...
 $ Flights.Per.Year             : int  5 11 5 6 48 39 13 19 54 2 ...
 $ Loyalty                      : num  0.444 -0.1 0.444 0.2 -0.778 ...
 $ Type.of.Travel               : chr  "Business travel" "Business travel" "Business travel" "Business travel" ...
 $ Total.Freq.Flyer.Accts       : int  2 0 2 3 0 3 0 0 0 0 ...
 $ Shopping.Amount.at.Airport   : int  65 0 65 0 0 0 0 0 0 0 ...
 $ Eating.and.Drinking.at.Airport: int  40 160 40 75 30 40 50 30 15 45 ...
 $ Class                        : chr  "Eco Plus" "Eco" "Eco Plus" "Eco" ...
 $ Day.of.Month                 : int  30 16 30 11 7 27 3 31 28 6 ...
 $ Flight.date                  : chr  "3/30/14" "2/16/14" "3/30/14" "3/11/14" ...
 $ Partner.Code                 : chr  "OO" "OO" "OO" "OO" ...
 $ Partner.Name                 : chr  "Northwest Business Airlines Inc." "Northwest Business Airlines Inc." "Northwest Business Airlines Inc." "Northwest Business Airlines Inc."
 ...
 $ Origin.State                 : chr  "Minnesota" "Minnesota" "Minnesota" "Minnesota" ...
 $ Destination.State            : chr  "South Dakota" "South Dakota" "South Dakota" "South Dakota" ...
 $ Scheduled.Departure.Hour     : int  13 13 13 13 13 21 13 13 13 21 ...
 $ Departure.Delay.in.Minutes   : int  0 0 0 0 0 0 83 10 0 0 ...
 $ Arrival.Delay.in.Minutes     : int  0 0 0 0 0 0 0 79 2 0 0 ...
 $ Flight.cancelled             : chr  "No" "No" "No" "No" ...
 $ Flight.time.in.minutes       : int  45 44 45 44 44 48 49 48 46 47 ...
 $ Flight.Distance              : int  257 257 257 257 257 257 257 257 257 257 ...
 $ Likelihood.to.recommend      : int  9 9 10 4 4 4 4 10 7 4 ...
 $ olong                        : num  -93.3 -93.3 -93.3 -93.3 -93.3 ...
 $ olat                         : num  45 45 45 45 45 ...
 $ dlong                        : num  -98.5 -98.5 -98.5 -98.5 -98.5 ...
 $ dlat                         : num  45.5 45.5 45.5 45.5 45.5 ...
 $ freeText                     : chr  NA NA NA NA ...
```

Cleaning the Data

It was important we make sure our data was organized in order for our analysis to be accurate and amenable. We checked to see if there were any missing data fields throughout the data set. We did keep in mind that if a field were not applicable, it did not necessarily mean the data was blank, but because the value was in fact meant to be 0. We found that there were four missing values in the Likelihood to Recommend column, so we eliminated those blank values.

Transforming the data set to create:

- Arrival Delay column to be greater than 5 Minutes
- Departure Delay column to be greater than 5 Minutes
- Arrival Delay in Minutes from NA to 0
- Shopping transformed to yes or no
- Changing City, State to lowercase
- Changing certain variable to factors during certain processes

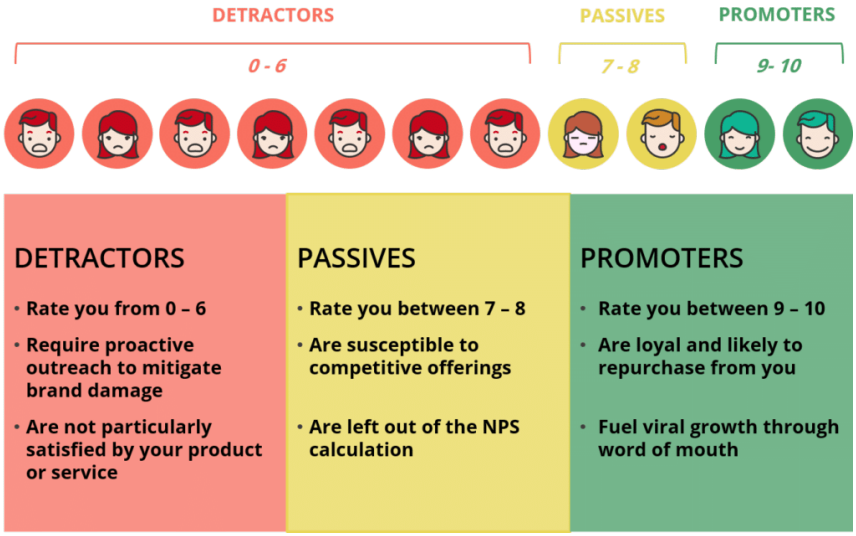
Calculating Net Promoter Score

NPS:

NPS Formula: %Promoters = % Detractors (Natural Language)

```
airdata.pd<-airData
airdata.m<- airData
airdata.pd$Likelihood.to.recommend[which(airData$Likelihood.to.recommend<7)]<-"detractors"
airdata.pd$Likelihood.to.recommend[which(airData$Likelihood.to.recommend>8)]<-"promoters"
airdata.pd$Likelihood.to.recommend[which(airData$Likelihood.to.recommend==7)]<-"passive"
airdata.pd$Likelihood.to.recommend[which(airData$Likelihood.to.recommend==8)]<-"passive"
table(airdata.pd$Likelihood.to.recommend)
NPS<-prop.table(table(airdata.pd$Likelihood.to.recommend))
NPS<-100*(NPS[3]-NPS[1])#Net Promotor Score
NPS#8.90%
```

NPS Score: 8.9%



NPS Function:

```
#NPS:-100%-100%,>50%: GOOD
NPS<- function(a){
  s<-length(a$Likelihood.to.recommend[which(a$Likelihood.to.recommend>8)])/length(a$Likelihood.to.recommend)-
    length(a$Likelihood.to.recommend[which(a$Likelihood.to.recommend<7)])/length(a$Likelihood.to.recommend)
  return(s)
}
```

We packaged it and it could be easily used by just inputting the name of dataset which you would like to calculate its NPS.

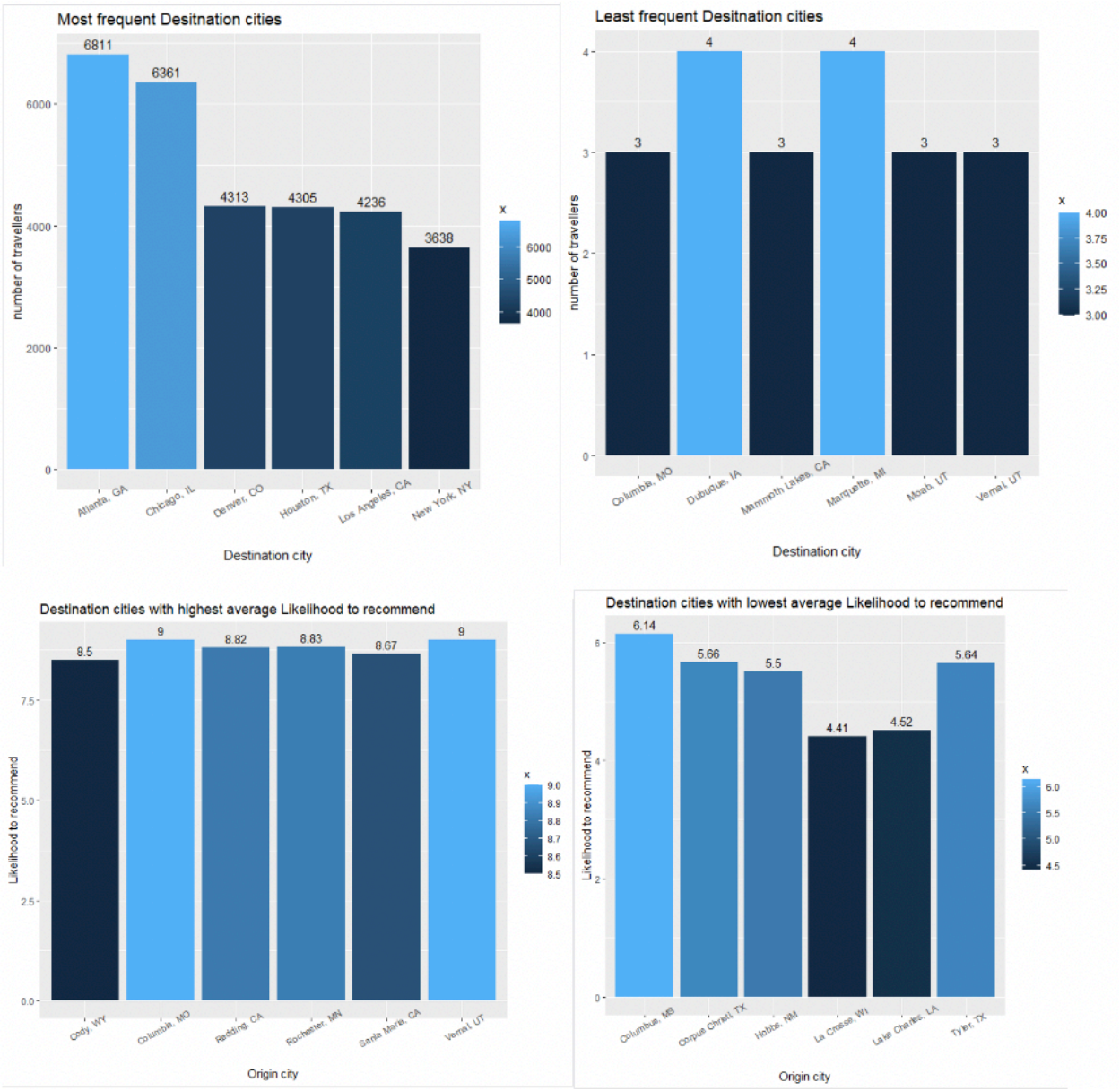
DESCRIPTIVE STATISTICS & VISUALIZATIONS

We first ran a basic summary of the data set to explore the descriptive statistics among the 32 variables to get an idea of what we were working with. We were able to get an overview of the highest, lowest, as well as a range of values.

```
Destination.City   Origin.City        Airline.Status          Age           Gender
Length:88100       Length:88100       Length:88100       Min.   :15.00   Length:88100
Class :character   Class :character   Class :character   1st Qu.:33.00   Class :character
Mode  :character   Mode  :character   Mode  :character   Median :45.00   Mode  :character
                                                         Mean   :46.22
                                                         3rd Qu.:59.00
                                                         Max.   :85.00

Price.Sensitivity Year.of.First.Flight Flights.Per.Year    Loyalty         Type.of.Travel
Min.   :0.000     Min.   :2003         Min.   : 0.00    Min.   :-0.97619   Length:88100
1st Qu.:1.000     1st Qu.:2004         1st Qu.: 9.00    1st Qu.:-0.70000   Class :character
Median :1.000     Median :2007         Median :17.00    Median :-0.42857   Mode  :character
Mean   :1.277     Mean   :2007         Mean   :20.04    Mean   :-0.27419
3rd Qu.:2.000     3rd Qu.:2010         3rd Qu.:29.00    3rd Qu.: 0.05882
Max.   :4.000     Max.   :2012         Max.   :98.00    Max.   : 1.00000

Total.Freq.Flyer.Accts Shopping.Amount.at.Airport Eating.and.Drinking.at.Airport    Class
Min.   : 0.0000        Min.   :  0.00             Min.   :  0.00                  Length:88100
1st Qu.: 0.0000        1st Qu.:  0.00             1st Qu.: 30.00                  Class :character
Median : 0.0000        Median :  0.00             Median : 60.00                  Mode  :character
Mean   : 0.8899        Mean   : 26.62             Mean   : 67.99
3rd Qu.: 2.0000        3rd Qu.: 30.00             3rd Qu.: 90.00
Max.   :12.0000        Max.   :745.00             Max.   :895.00

 Day.of.Month   Flight.date        Partner.Code       Partner.Name       Origin.State
Min.   : 1.00   Length:88100       Length:88100       Length:88100       Length:88100
1st Qu.: 8.00   Class :character   Class :character   Class :character   Class :character
Median :16.00   Mode  :character   Mode  :character   Mode  :character   Mode  :character
Mean   :15.69
3rd Qu.:23.00
Max.   :31.00

Destination.State  Scheduled.Departure.Hour Departure.Delay.in.Minutes Arrival.Delay.in.Minutes
Length:88100       Min.   : 1.00           Min.   :  0.00             Min.   :  0.00
Class :character   1st Qu.: 9.00           1st Qu.:  0.00             1st Qu.:  0.00
Mode  :character   Median :13.00           Median :  0.00             Median :  0.00
                   Mean   :13.02           Mean   : 15.04             Mean   : 15.38
                   3rd Qu.:17.00           3rd Qu.: 13.00             3rd Qu.: 13.00
                   Max.   :23.00           Max.   :978.00             Max.   :970.00
                                           NA's   :1607               NA's   :1838
Flight.cancelled   Flight.time.in.minutes Flight.Distance  Likelihood.to.recommend     olong
Length:88100       Min.   : 13.0          Min.   :  67.0   9      :19883           Min.   :-165.39
Class :character   1st Qu.: 61.0          1st Qu.: 373.0   8      :14904           1st Qu.:-111.93
Mode  :character   Median : 92.0          Median : 628.0   10     :13966           Median : -90.14
                   Mean   :113.1          Mean   : 807.7   7      :13333           Mean   : -95.33
                   3rd Qu.:143.0          3rd Qu.:1024.0   4      : 7789           3rd Qu.: -81.64
                   Max.   :443.0          Max.   :3414.0   (Other):18221           Max.   : -66.12
                   NA's   :1838                            NA's   :    4
     olat            dlong              dlat          freeText
Min.   :18.02   Min.   :-165.39   Min.   :18.02   Length:88100
1st Qu.:33.56   1st Qu.:-111.93   1st Qu.:33.82   Class :character
Median :37.67   Median : -90.14   Median :37.67   Mode  :character
Mean   :37.08   Mean   : -95.33   Mean   :37.08
3rd Qu.:40.72   3rd Qu.: -81.64   3rd Qu.:40.72
Max.   :71.29   Max.   : -66.12   Max.   :71.29
```

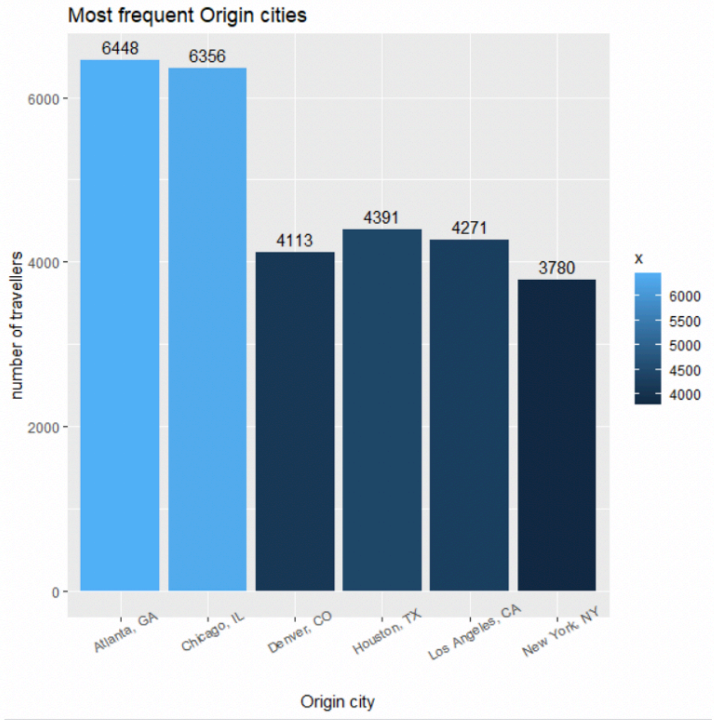
However, visualizations are at the utmost importance when translating this generic data into something that the human eye can easily read. We focused these variables into extensive visualizations to get an idea of which areas we can deeply explore more.

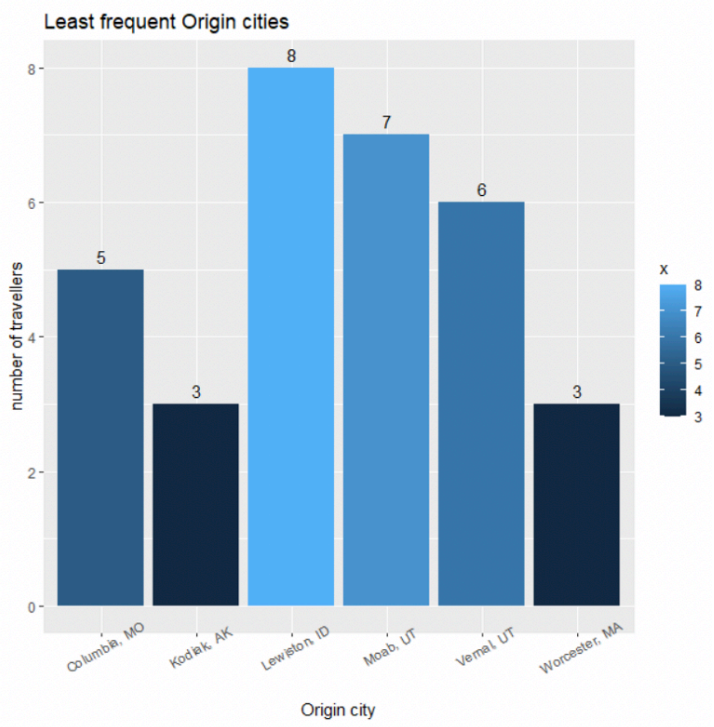
Destination Cities

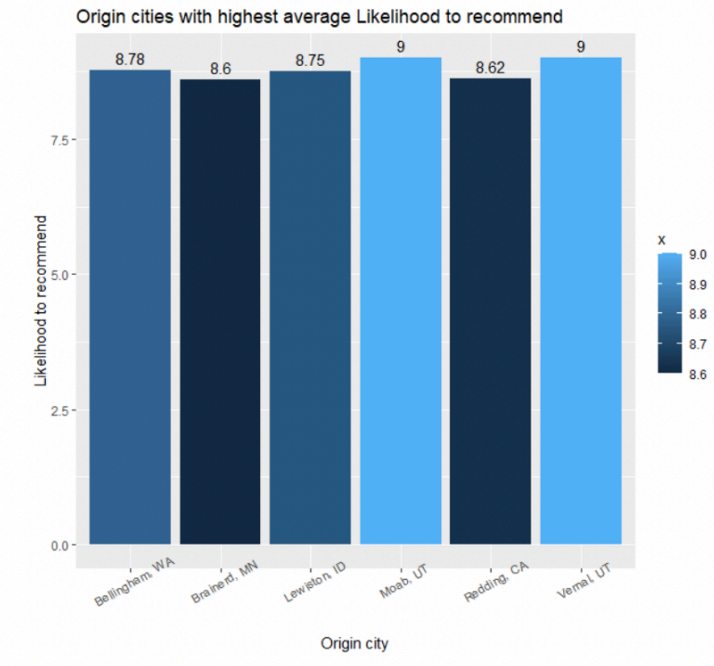


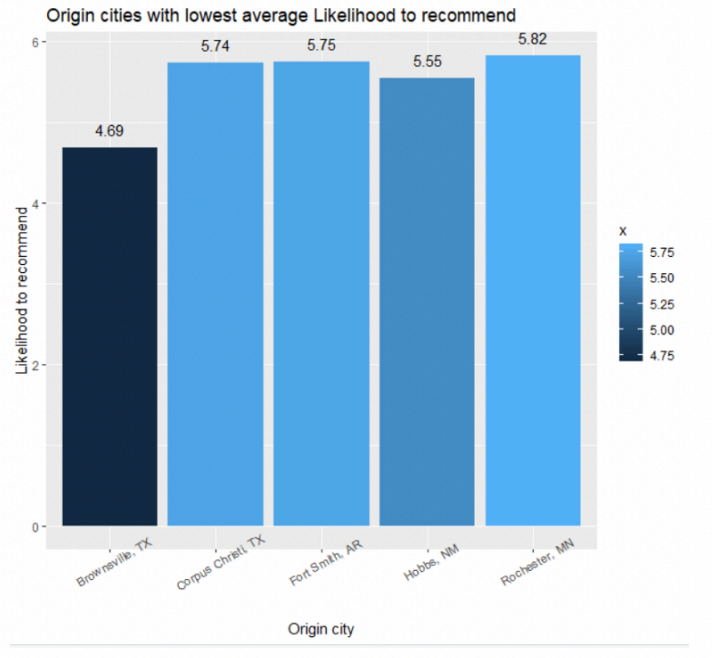
The states of Missouri, Utah, California were among the highest areas where the likelihood to recommend was highest, rating from an average of 8-9. States like Texas, Wisconsin, and Louisiana were among the lowest scores, so it might be possible to keep cities as a factor in mind as to customer churn.

Origin Cities



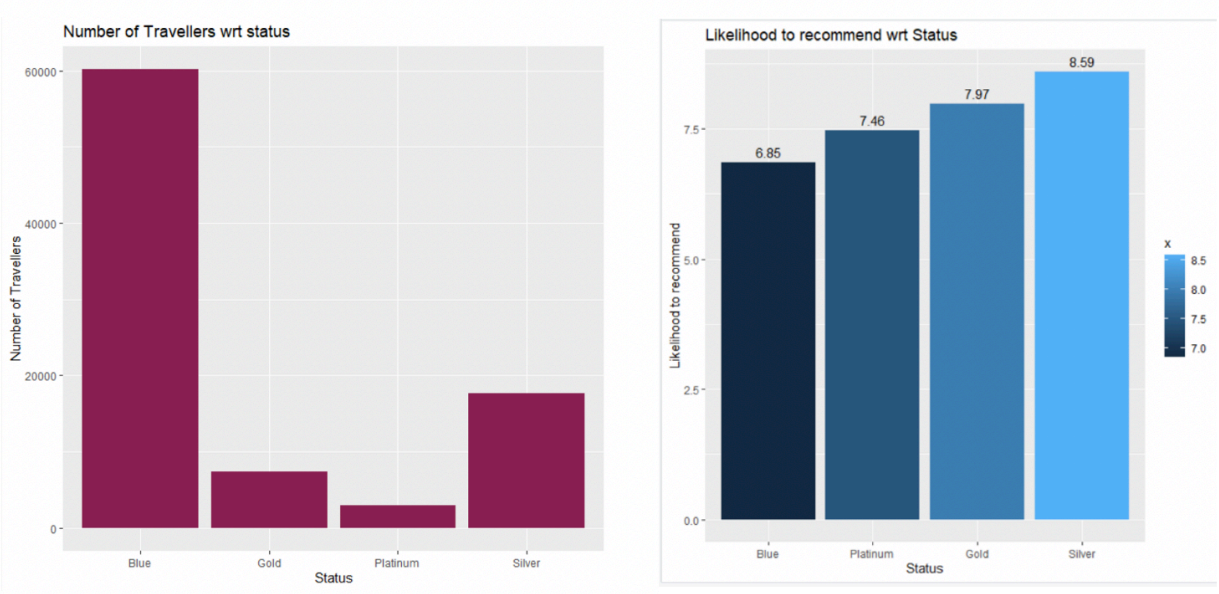






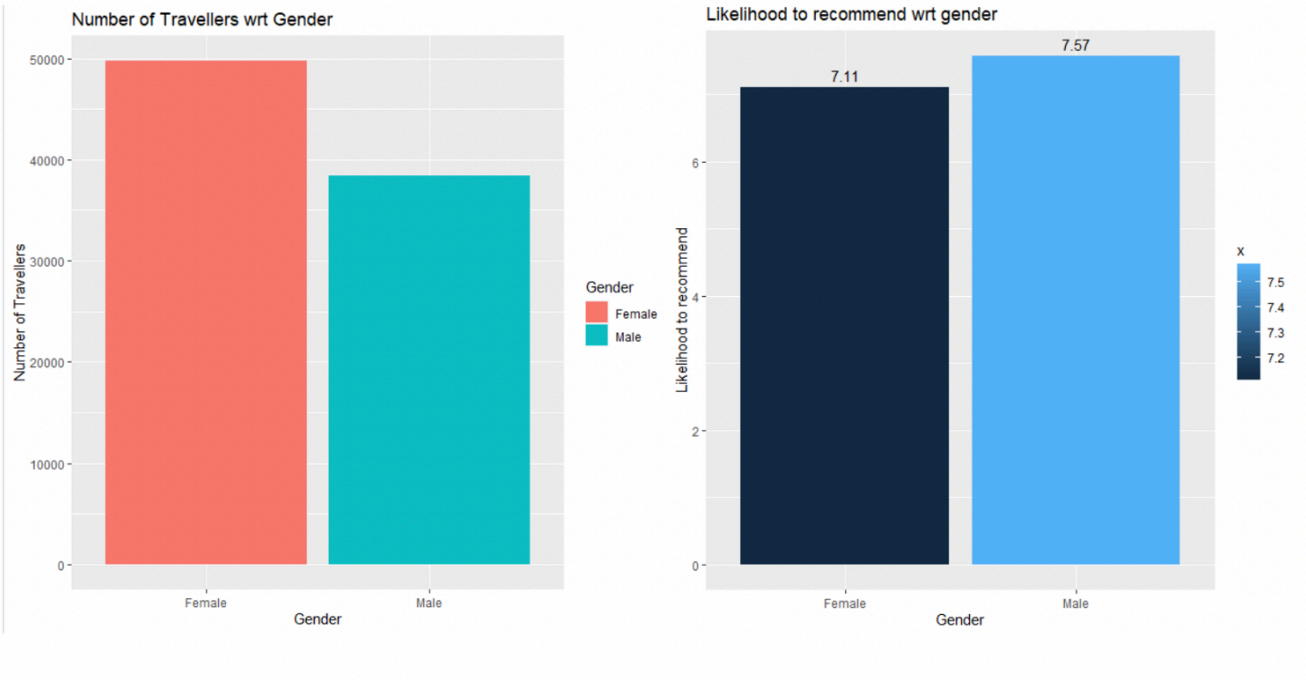
The origin cities seemed to match up closely with the scores in relation to destination cities, so again, cities may have a factor into a traveler's satisfaction score.

Status



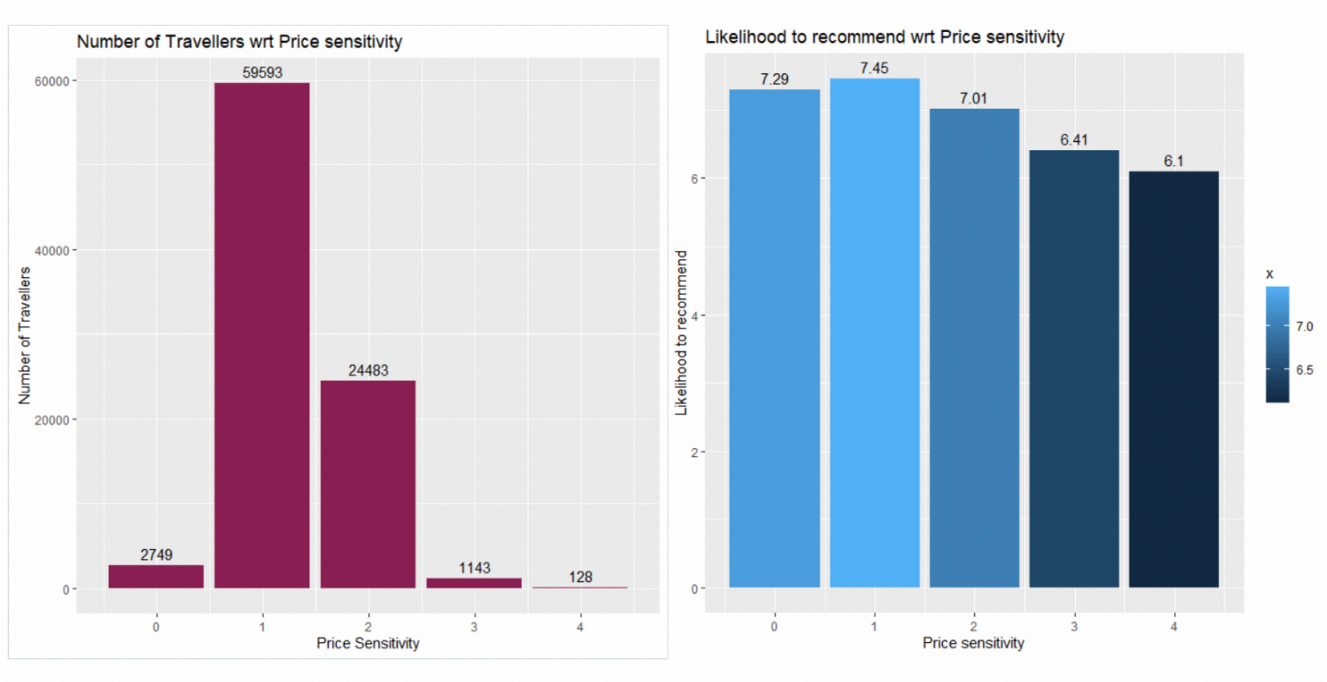
The number of travelers with status: blue is high(~60,000) but the average likelihood to recommend for travelers with status status:blue is lower than others.

Gender



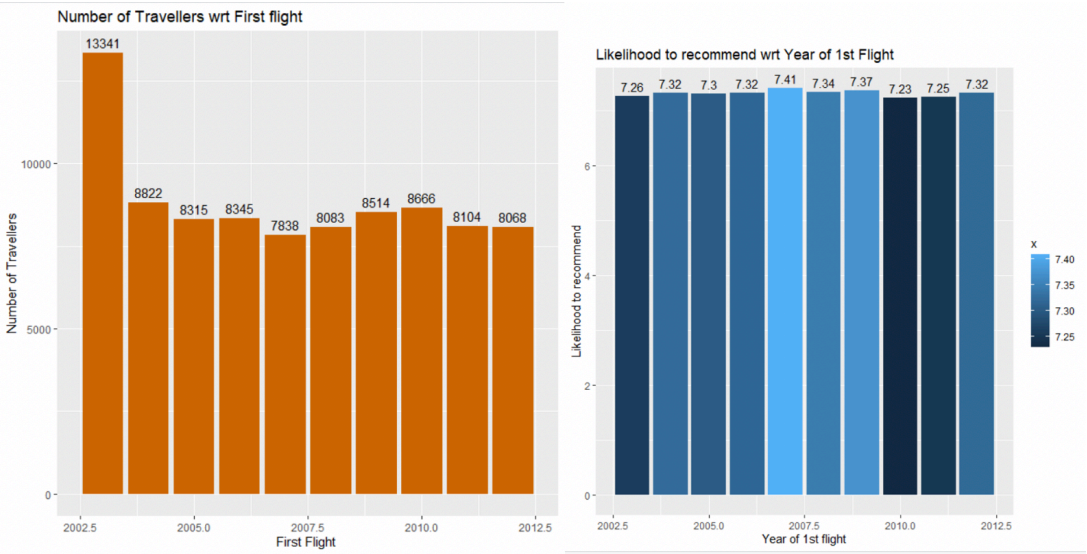
Female travelers are more but the mean likelihood to recommend is lower than that of male travelers.

Price Sensitivity



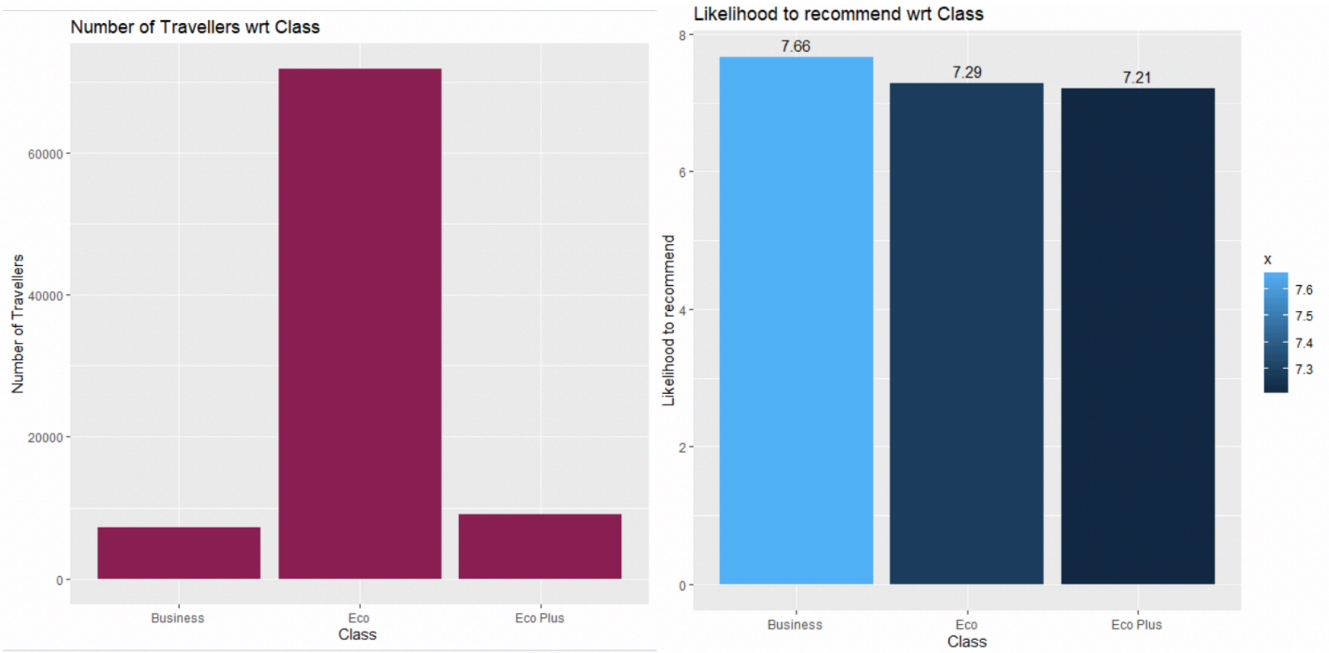
Price definitely seems to factor into the traveler's satisfaction score as it the score seems to be higher with a low price.

First-time Travelers



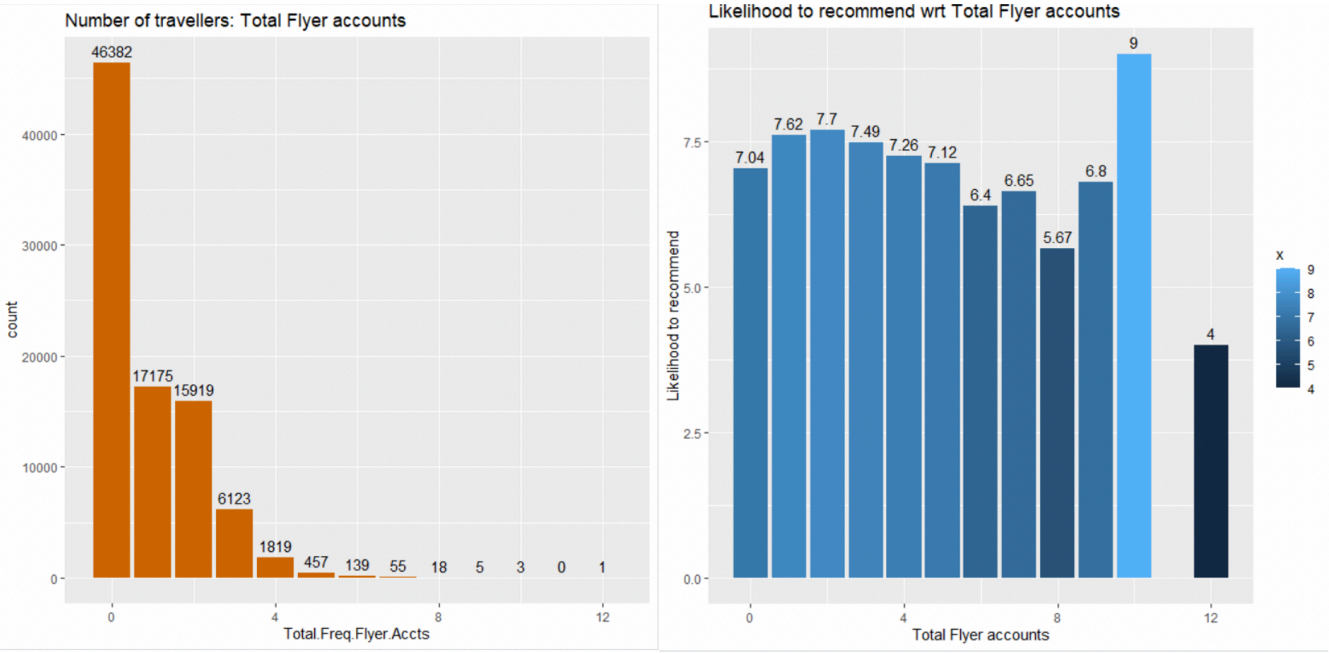
Travelers who fly for the first time have a score in the range of 7, which is okay, but can also mean that it's possible to look into how we can improve their first flight experience.

Travel Class



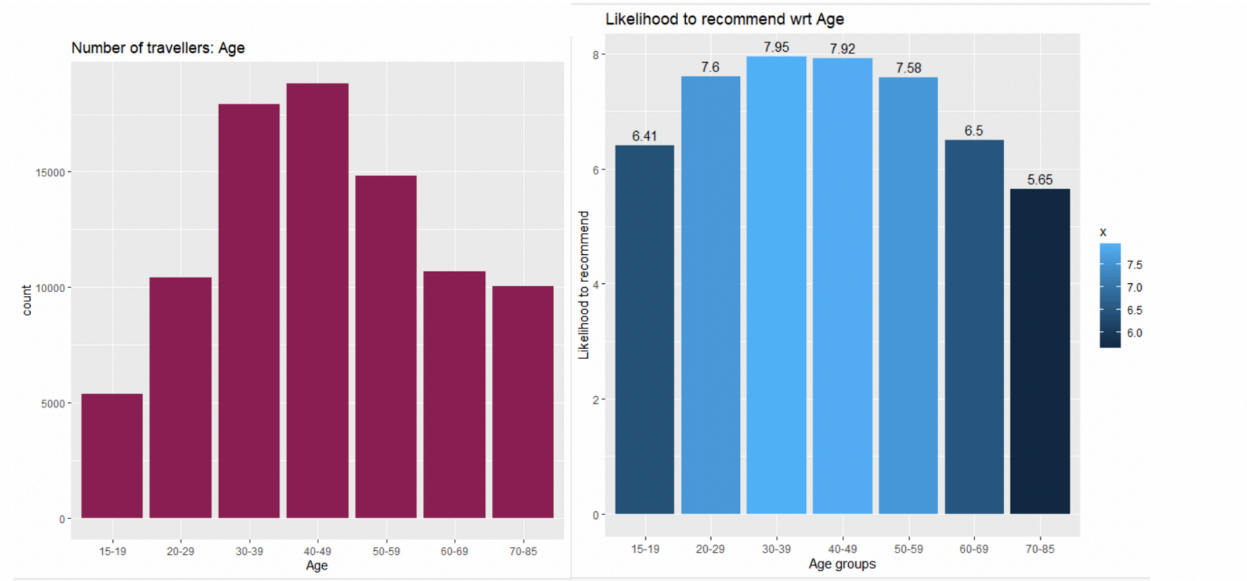
Business, Economy, or Economy Plus Class does not seem to be much of a factor into the recommendation score as the value were in the same range.

Total Flyer Accounts



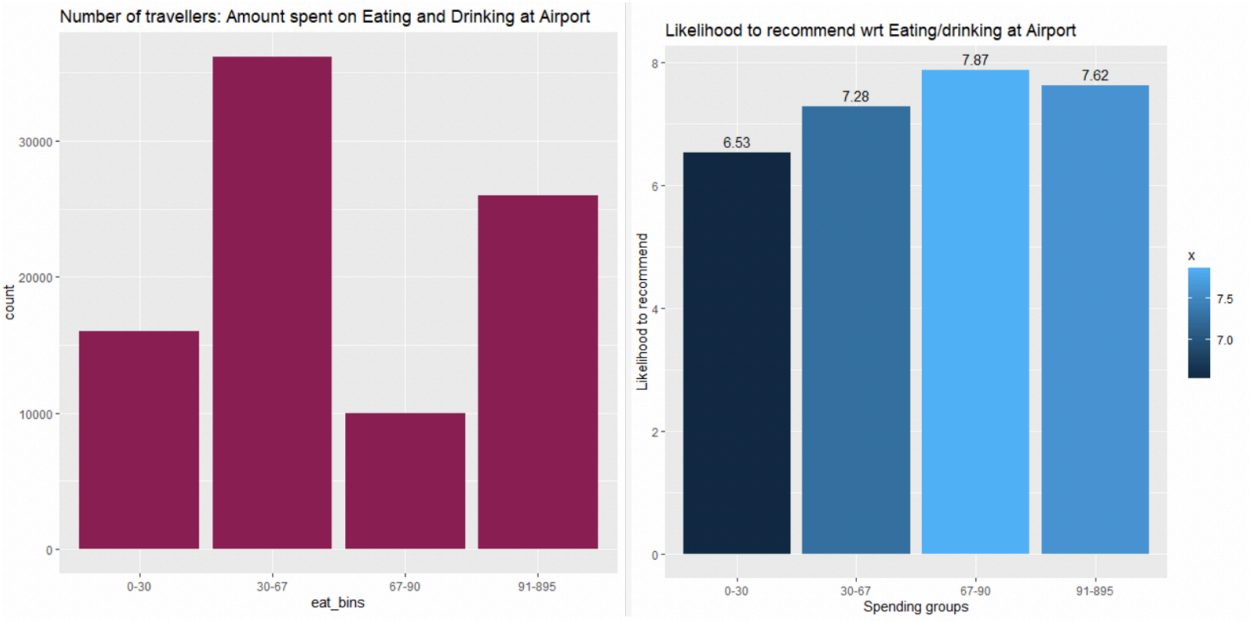
Travelers with higher flyer accounts seemed to rate the highest.

Age Groups



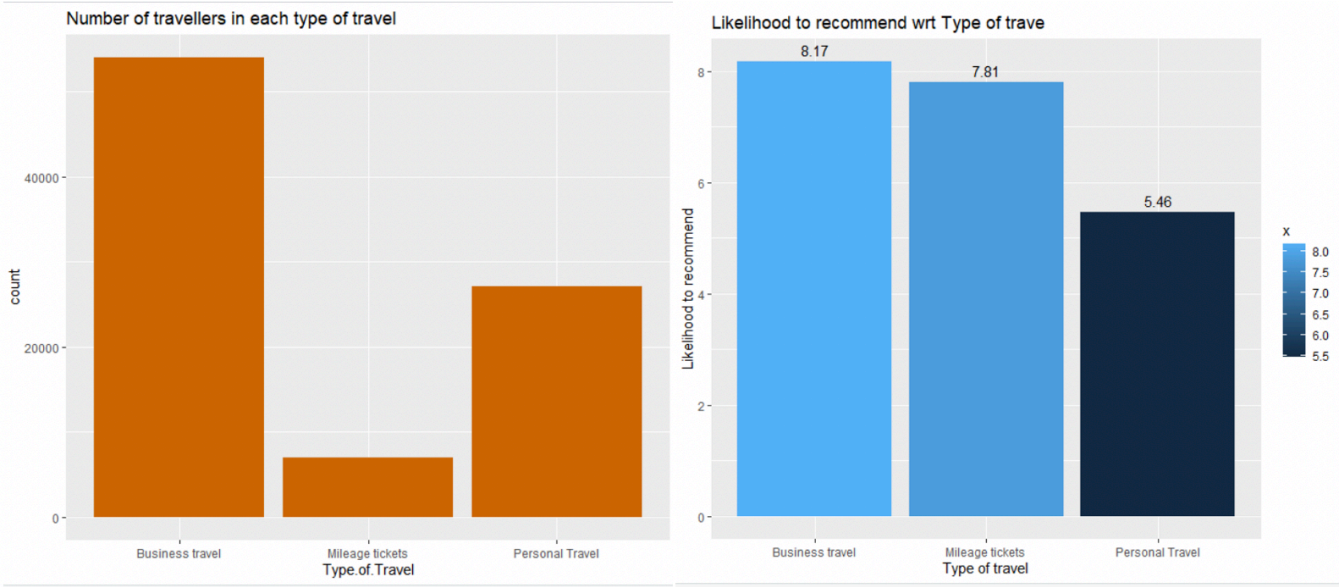
Age groups- 15-19 and 60-85 tend to give lower ratings.

Shopping and Eating



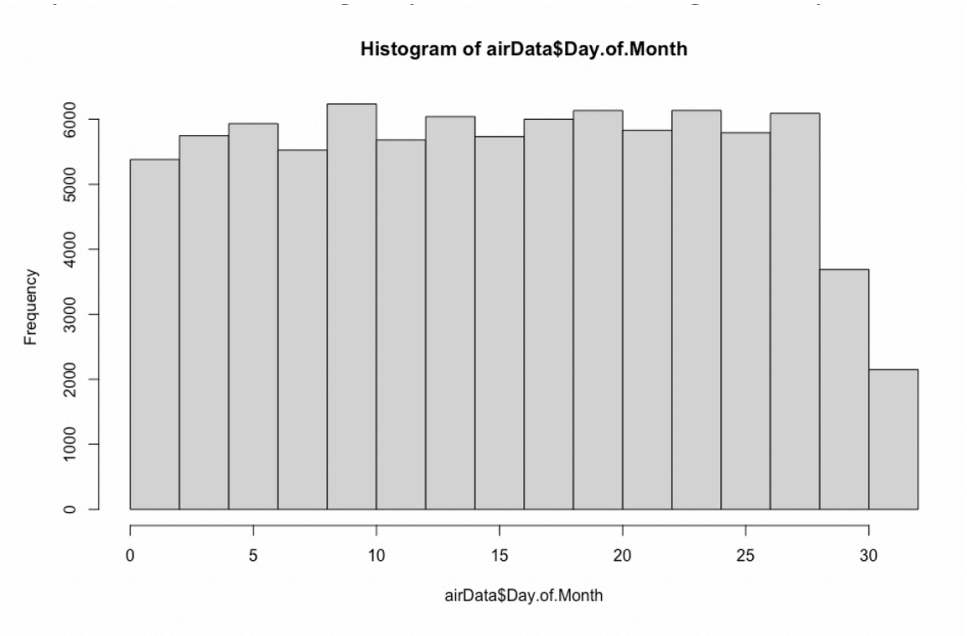
The majority of shoppers and eaters seemed to vote higher on their recommendation score so this can be a possible area for improvement of customer satisfaction.

Type of Travel



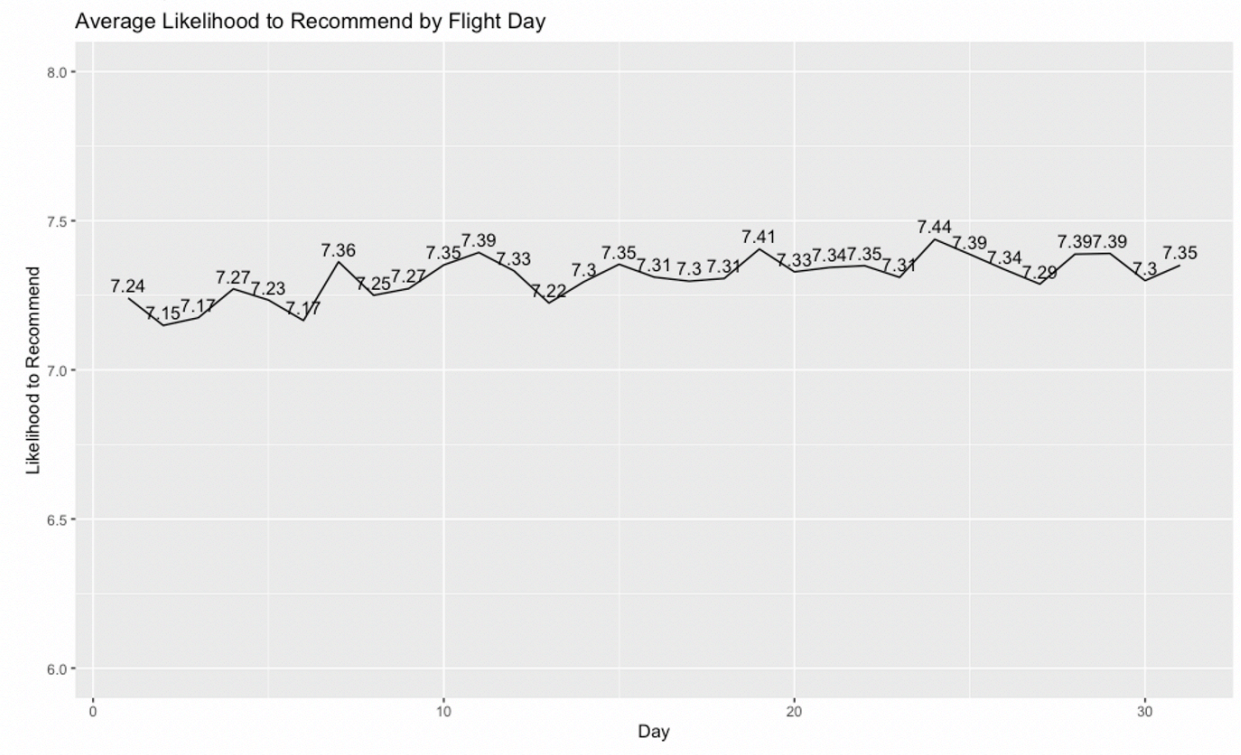
Customers who traveled for business and those who used mileage tickets rated the highest compared to customers who traveled for personal reasons.

Day of Month(the traveling day of each customer (ranges from 1 to 31)



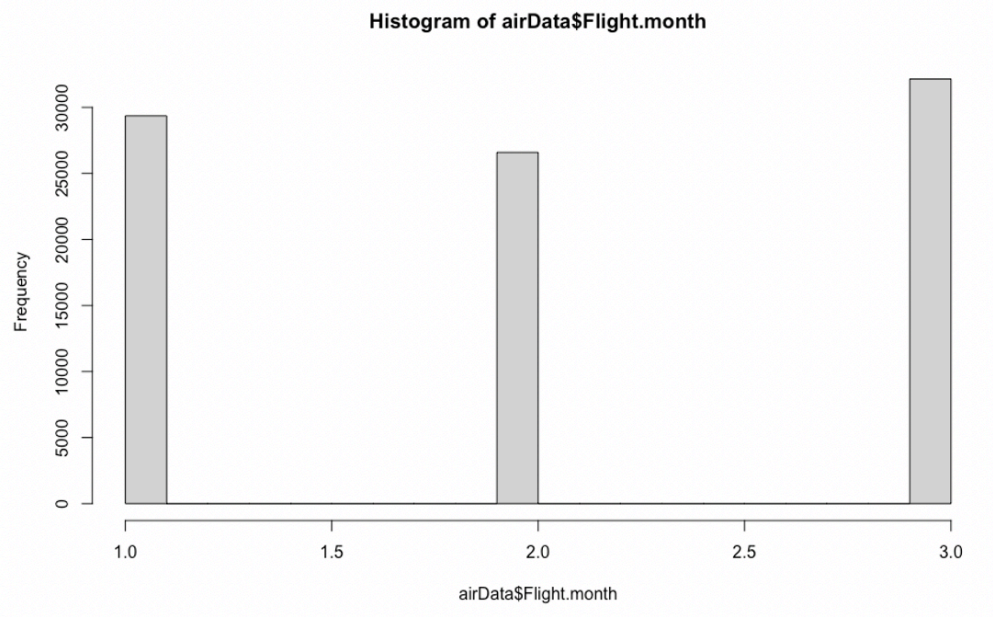
There is little variation in frequency for flights taken on different days of the month. The least frequent day is the 30th. The average day of the month that flights took place was the 15th.

Flight Delays



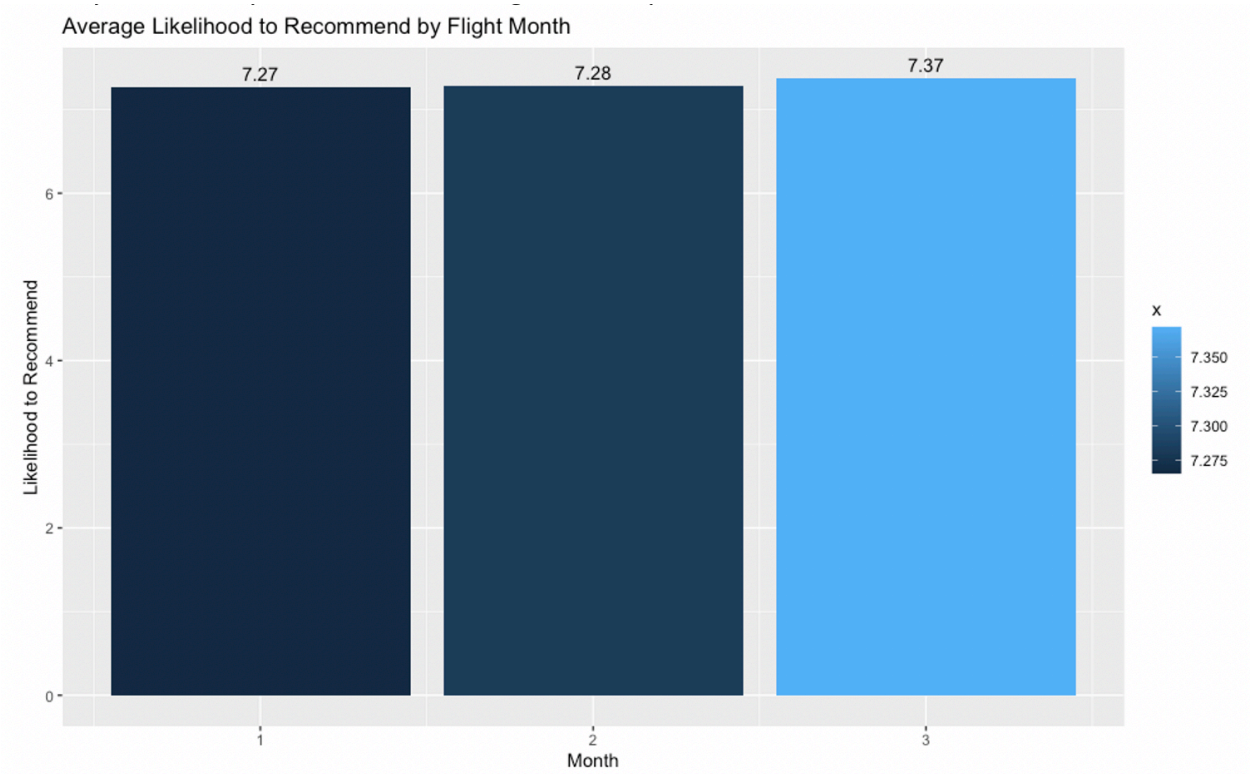
This shows the average likelihood to recommend for specific days of the month. These are aggregate scores for the day number in January, February, and March (the months in this data set). Over the days in the month, the minimum likelihood to recommend score is 7.15 and the maximum is 7.44. Overall, there is very little variation across all days. To this end, targeting specific days may not be a priority area for reducing customer churn due to the lack of variance.

Flight Month



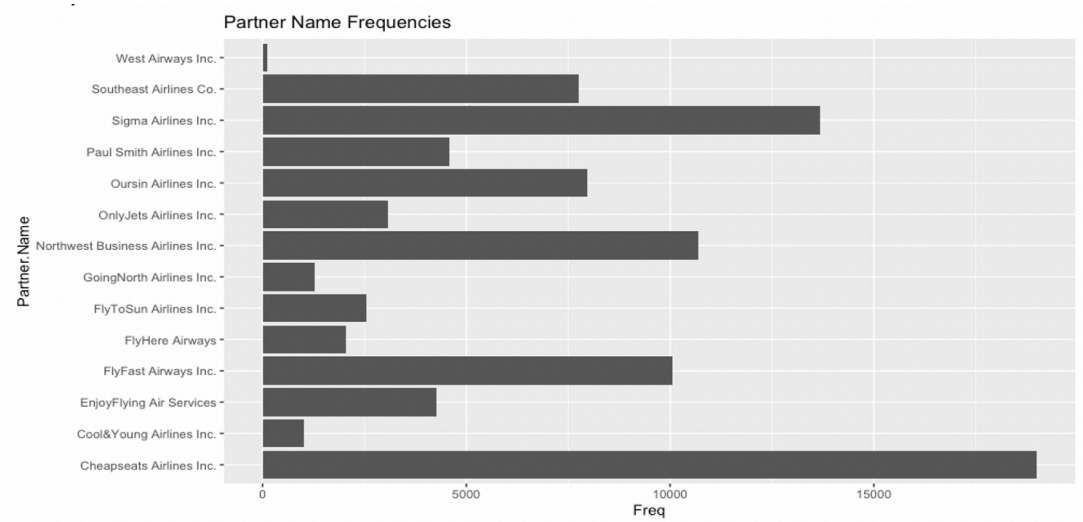
Parsed from Flight Date – the month in which the flight took place (January – March). There is little variation across frequencies of flights for the months included in this survey. They range from January to March . There is a fairly even distribution of flights across these months.

Recommendation by Month



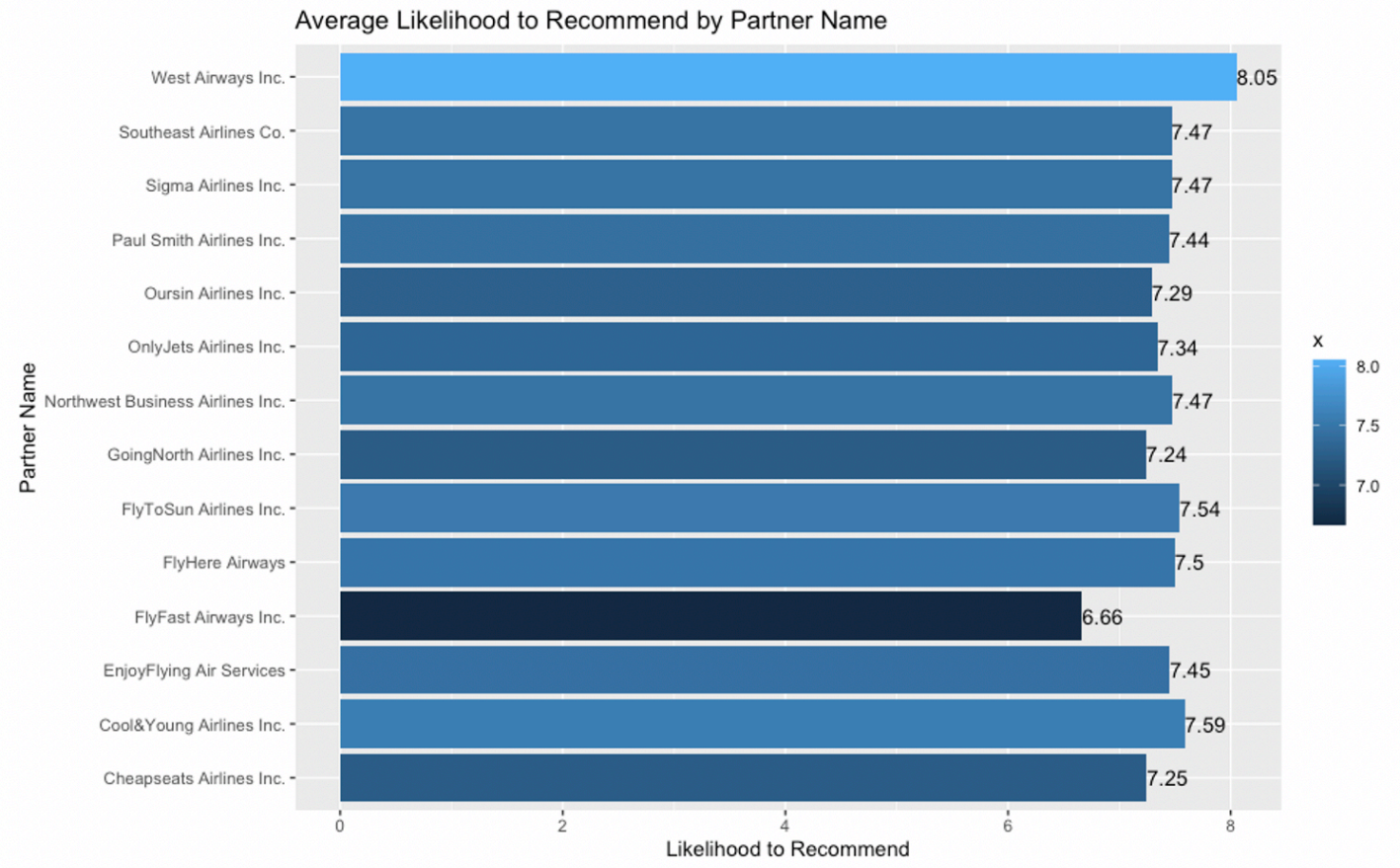
The range of Likelihood to Recommend ratings does not vary greatly across the months. The difference between the lowest and highest rating (7.27 vs 7.37) is only one tenth of a point. To this end, it may be unlikely that these specific months have significant impact on likelihood to recommend.

Partner Name



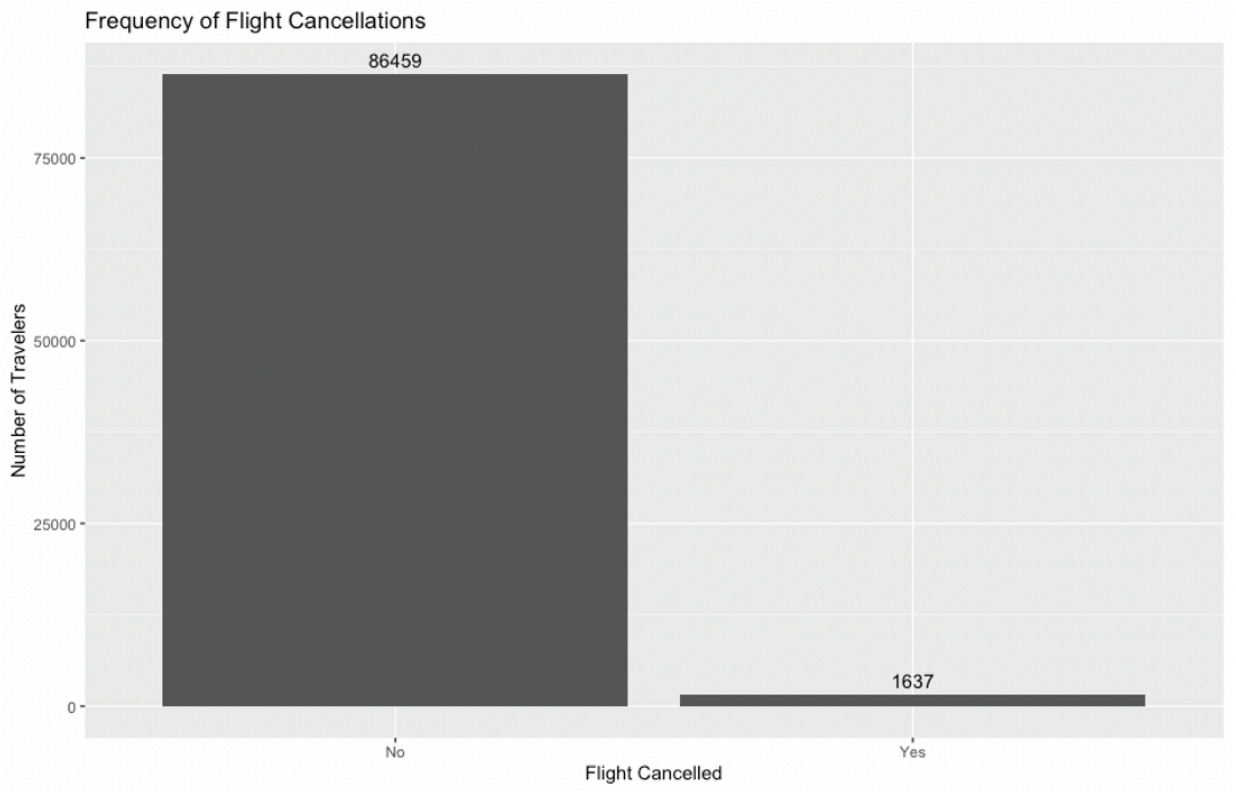
These are the full names of the partner airline companies. There is some variation in the number of flights taken across the different types of partner airlines. The most frequently used partner airline in this sample is Cheapseats Airlines Inc. The least frequent is West Airways Inc.

Recommendation by Partner Name

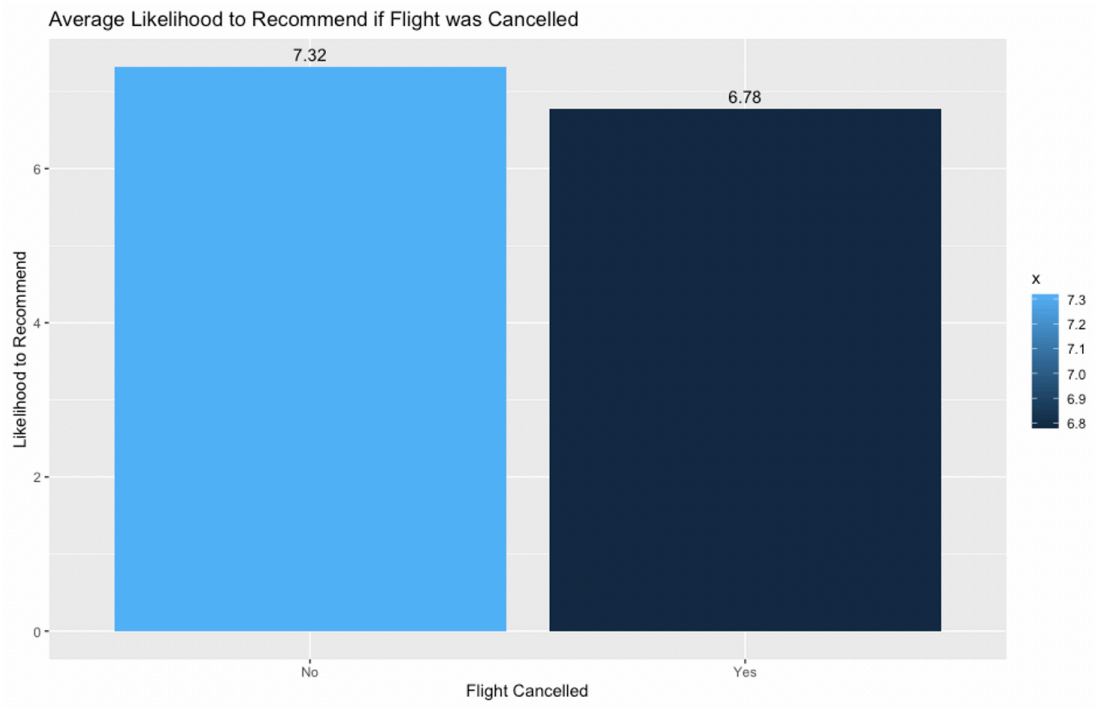


The average likelihood to recommend score varies across the airline partners. The partner with the highest score was West Airways Inc (8.05). The partner with the lowest was FlyFast Airways (6.66). Given Actionable insight: West Airways represents the lowest fraction of the sample, yet its clients provide the highest likelihood to recommend score. To this end, it may be worth exploring West Airway's practices as those could be resulting in superior service delivery for their clients.

Flight Cancelled

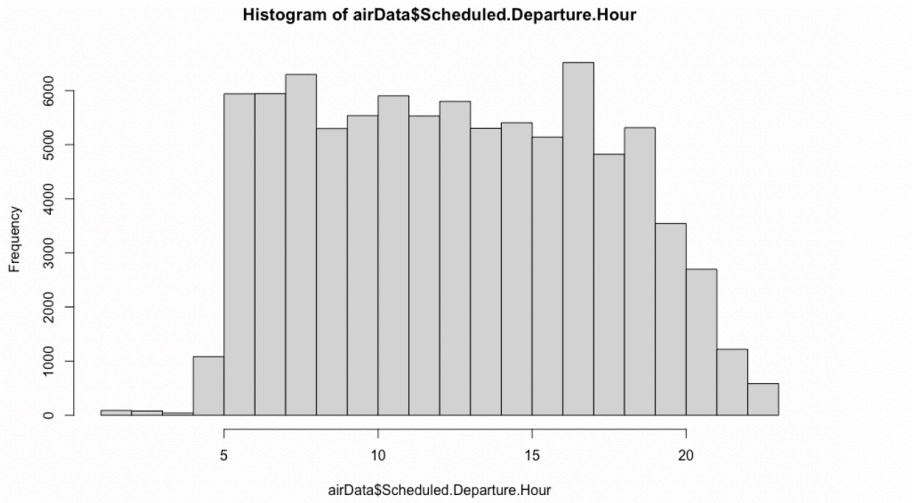


Flight cancellation data is important to examine because it could yield a negative experience for the client. In this survey data, a majority of the flights did not incur a cancellation.

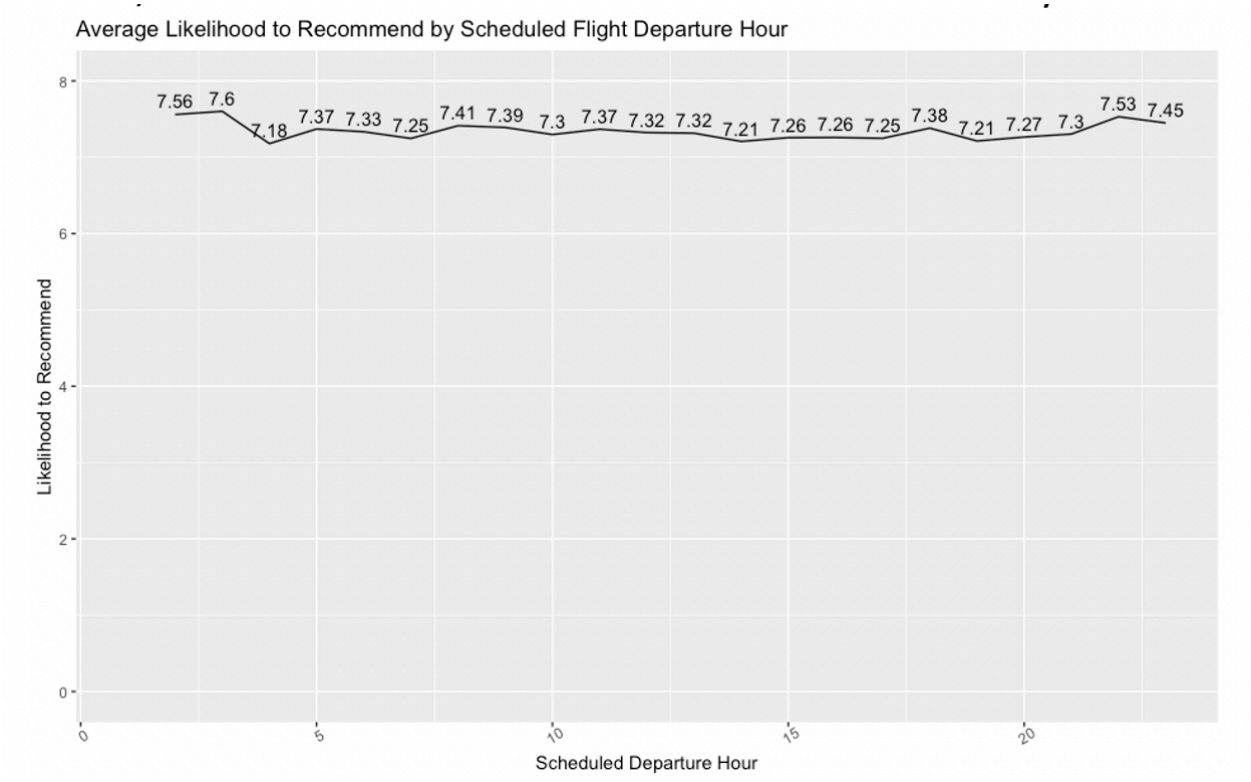


Although very few clients in this survey experienced a flight cancellation, the likelihood to recommend score did not vary drastically between them and those who did not have a cancelled flight. The average score for clients who had a cancelled flight was only 0.54 points less than the average score for clients who didn't have a cancellation. This suggests that other factors are greater influence for likelihood to recommend ratings or that any existing practices to keep clients happy if their flight is cancelled, may be slightly effective.

Scheduled Departure Hour

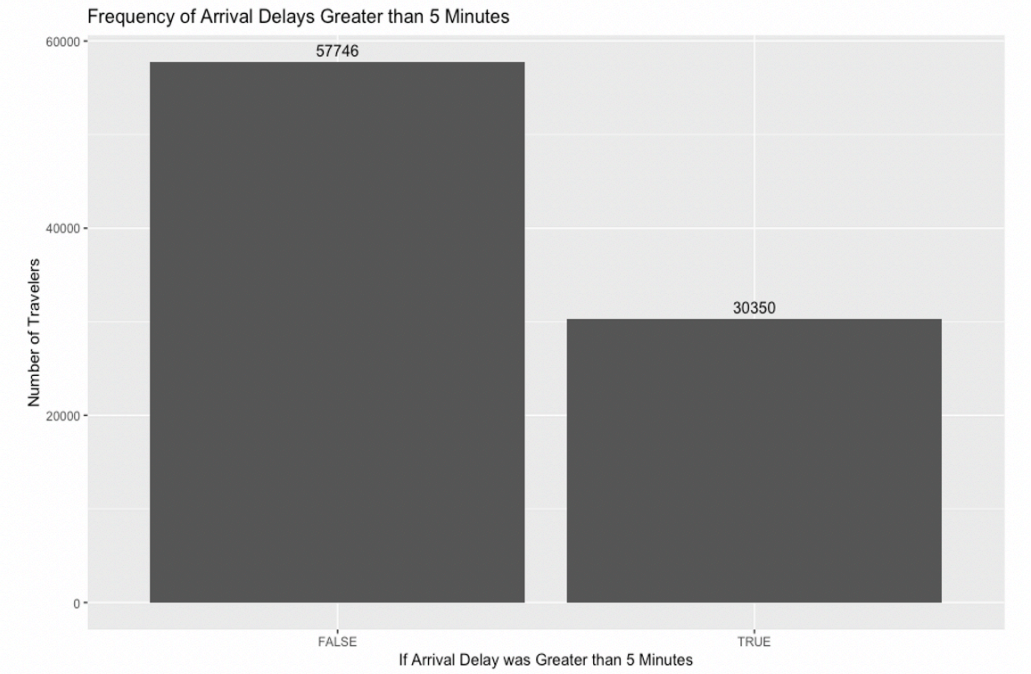


The average scheduled departure hour was 13 (or 1PM). Very few flights take place before 5pm and after 8pm.

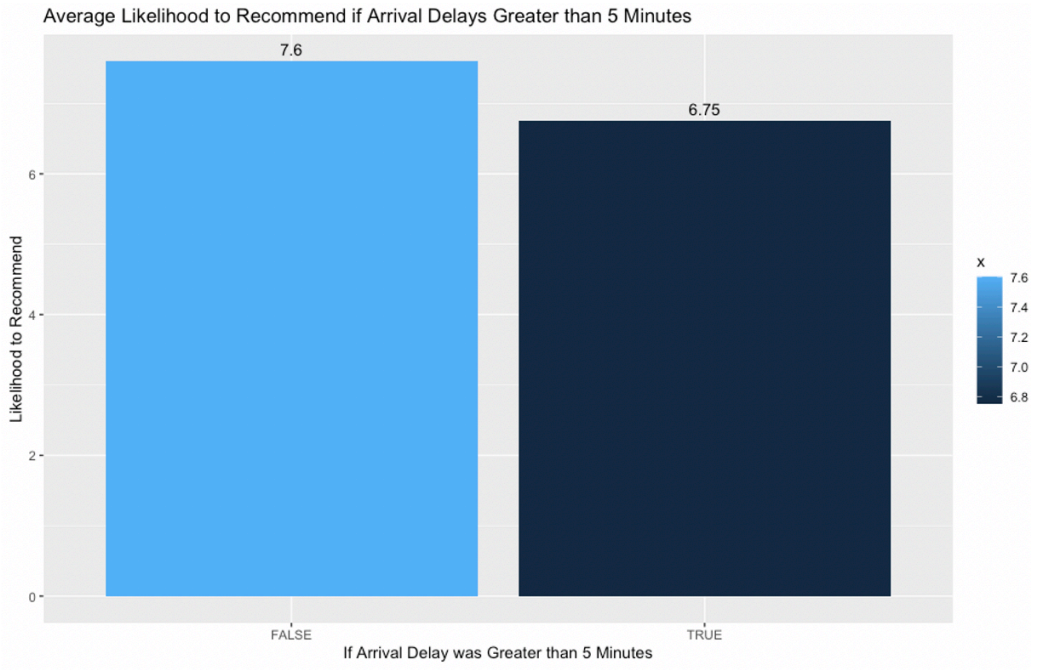


This line chart shows how average likelihood to recommend ratings vary by scheduled departure hour. Overall, there is little variation across the different hours in the day.

Arrival Delay Greater than 5 Minutes: Delay of arrival airline time which is more than 5 minutes per each passenger in the data

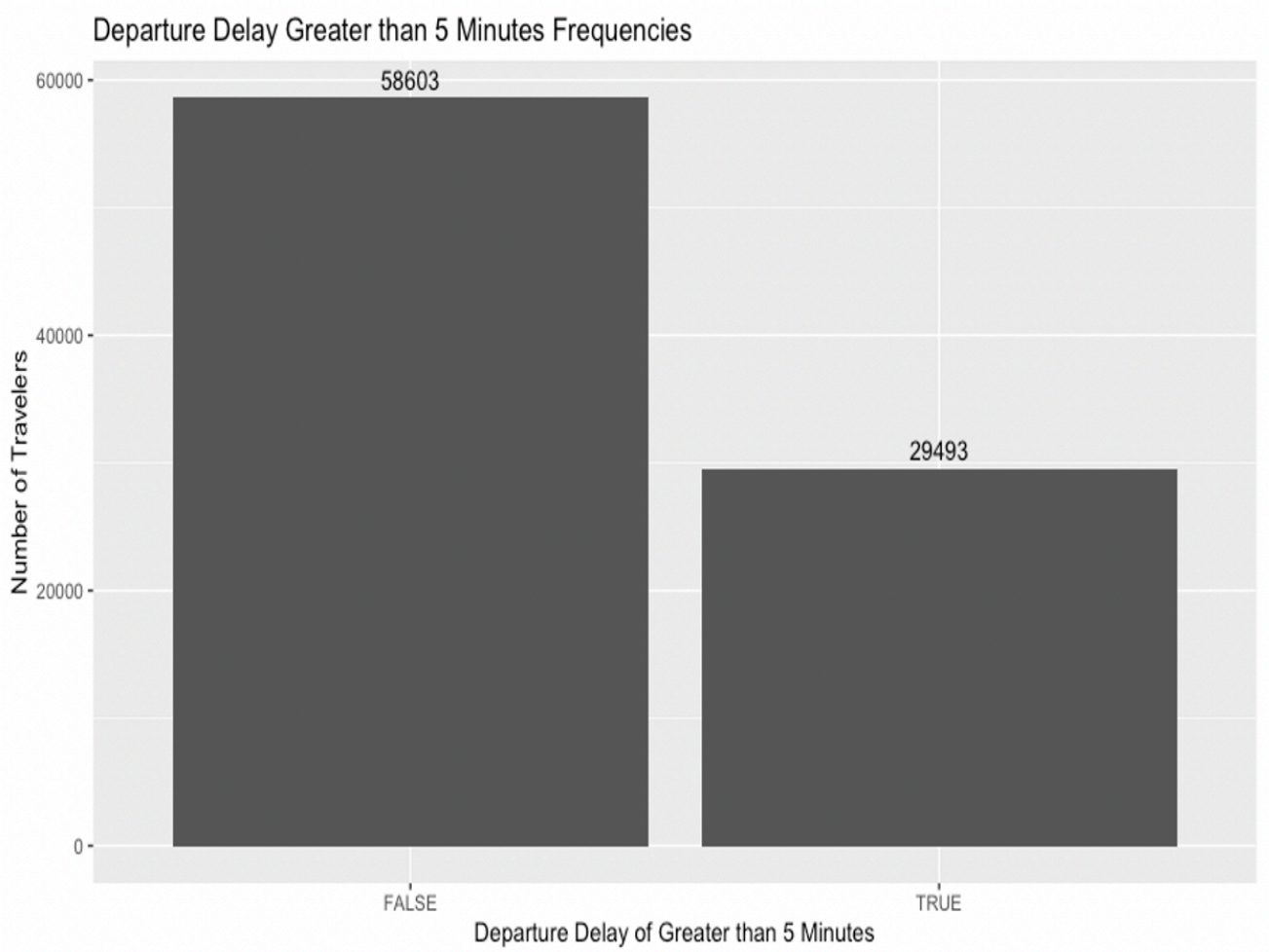


This variable was created from the arrival delay in minutes column of the survey data. The value FALSE represents if the arrival delay in minutes was less than 5. The value TRUE represents if there was an arrival delay in minutes was above 5. A majority of respondents did not have an arrival delay greater than 5 minutes.

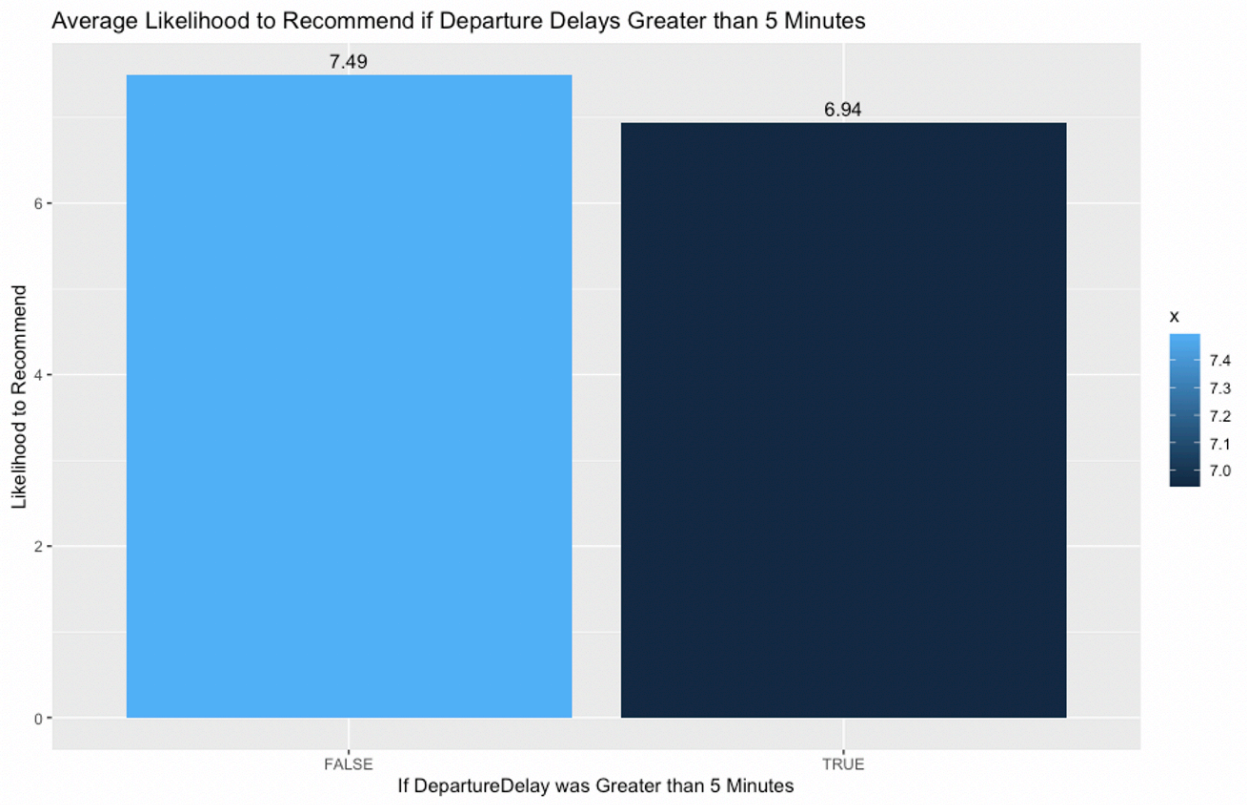


Respondents gave a lower likelihood to recommend rating if their arrival delay was greater than 5 minutes (6.75 vs 7.6). The difference of nearly 1 whole point indicates that delays may be a significant factor for influencing likelihood to recommend scores.

Departure Delay Greater than 5 Minutes: Delay of departure airline time which is more than 5 minutes per each passenger in the data

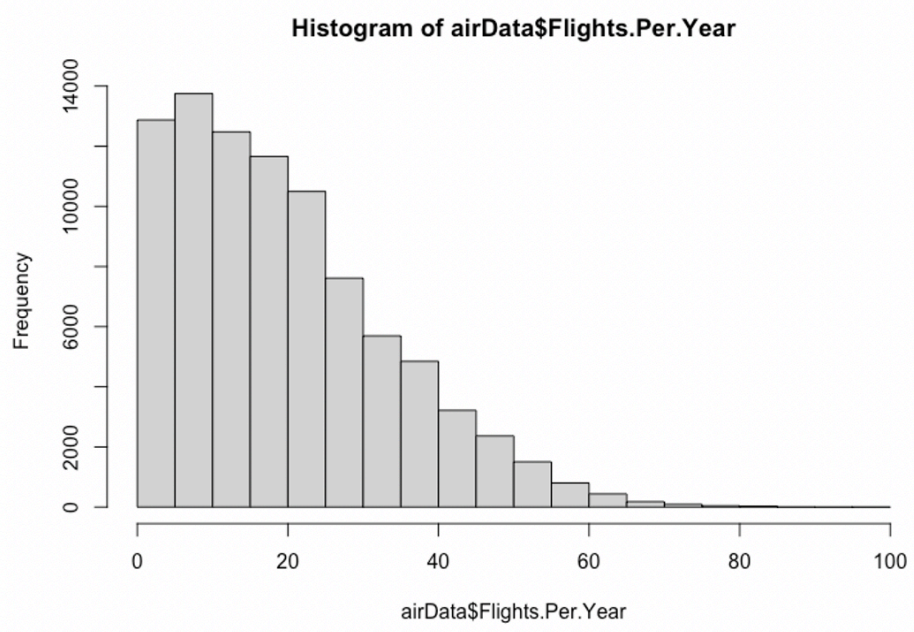


This variable was created from the departure delay in minutes column of the survey data. The value FALSE represents if the departure delay in minutes was less than 5. The value TRUE represents if there was a delay greater than 5 minutes. A majority of respondents did not have a departure delay greater than 5 minutes.

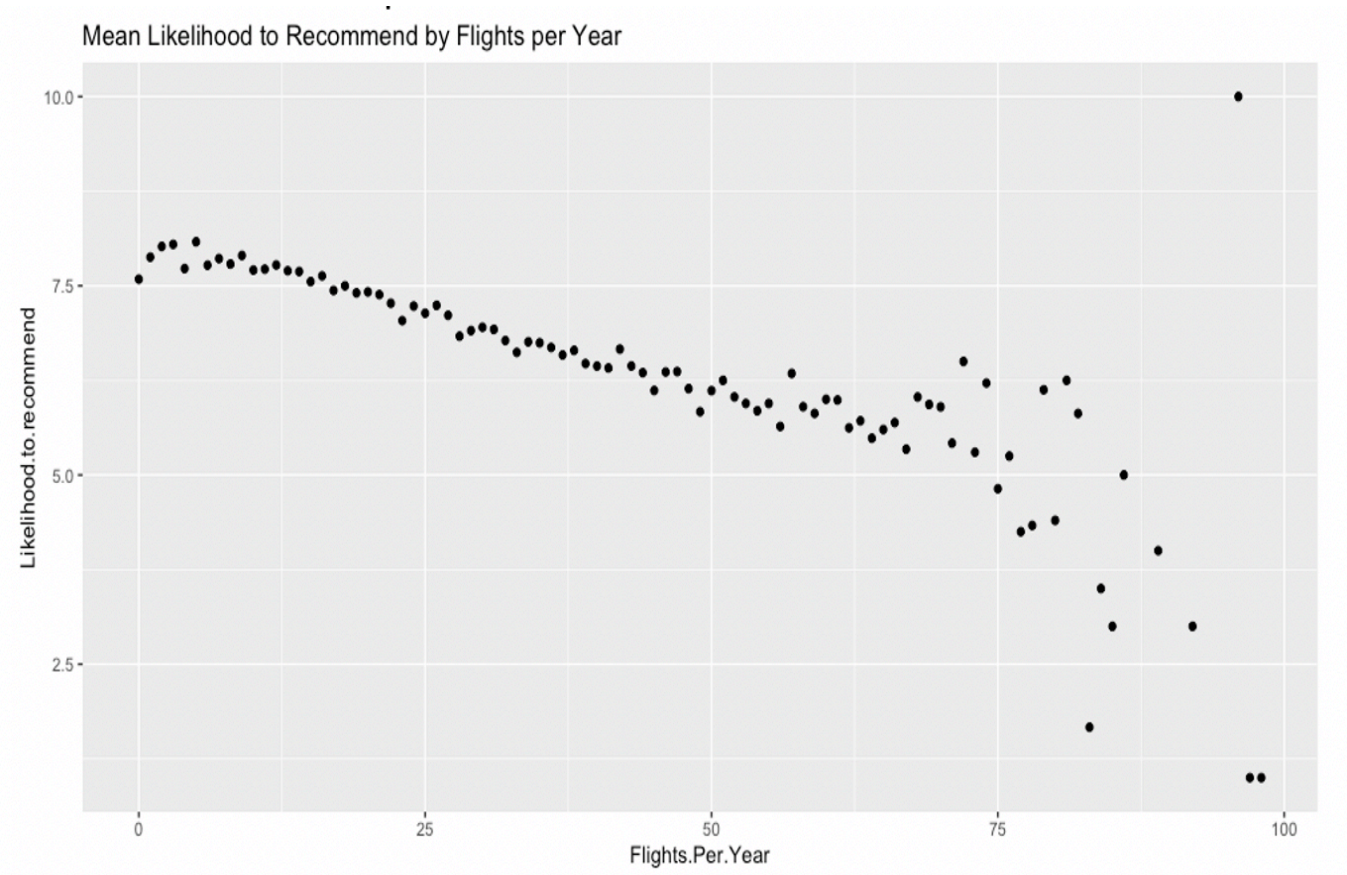


Respondents who had a departure delay greater than 5 minutes gave a lower likelihood to recommend score compared to respondents who did not have a delay (6.94 vs 7.49).

Flights per Year: The number of flights that each customer has taken in the most recent 12 months. The range starting from 0 to 100.



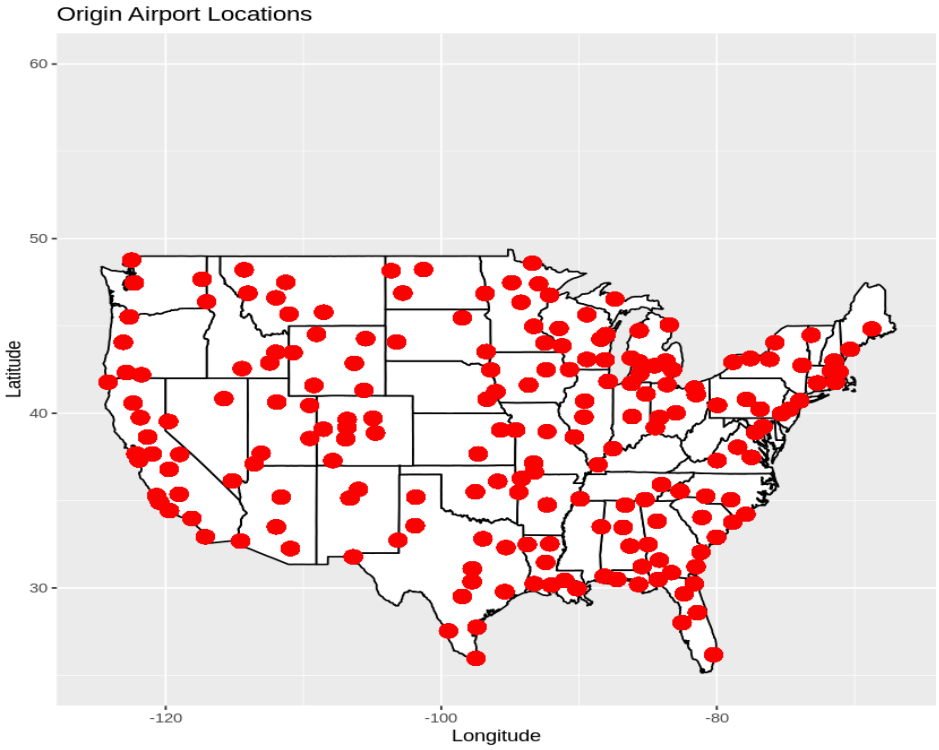
This histogram for flights per year is skewed to the right. Most respondents had taken either no flights or less than 10 in the past year. The average number of flights among respondents was 20 and the median was 17. The maximum was 98 while the minimum was 0. Overall, most respondents had not flown more than once per month prior to the flight recorded in the survey.

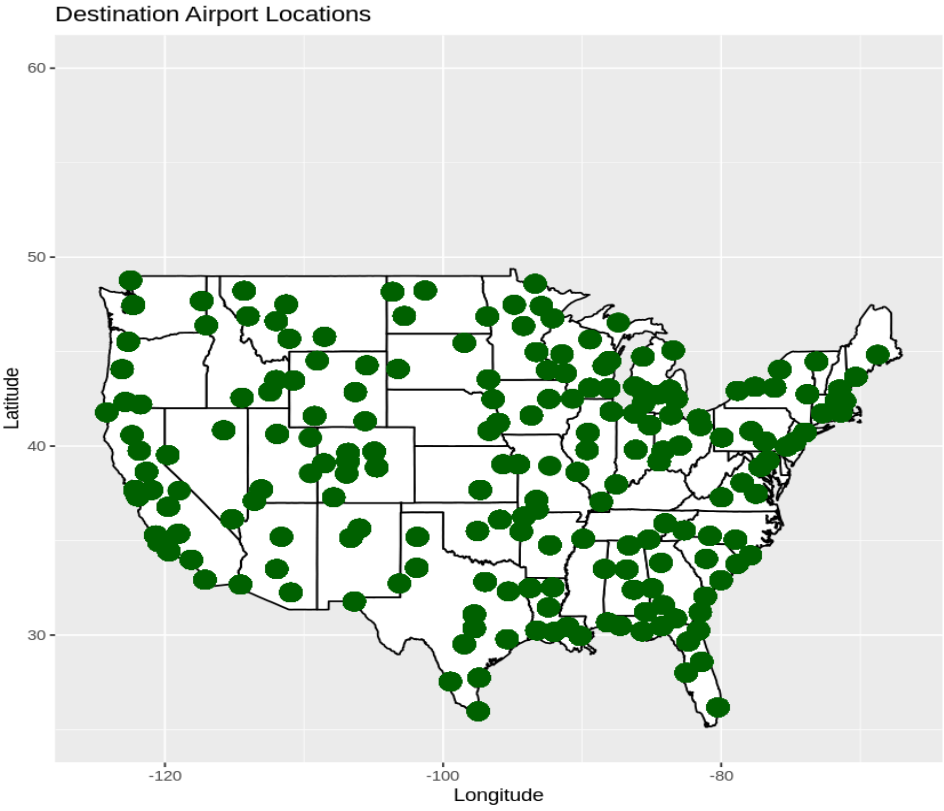


This graph suggests that the mean likelihood to recommend decreases as the number of flights per year increases. Most respondents who flew fewer than 25 times in the previous year had a mean score of approximately 7.5. But scores for respondents who travel more appear to decrease their score as the number of flights increase.

This indicates the importance of providing a positive experience to individuals on their first flights, as any novel positive experiences may imprint more strongly and yield a higher likelihood to recommend, than clients who are frequent fliers.

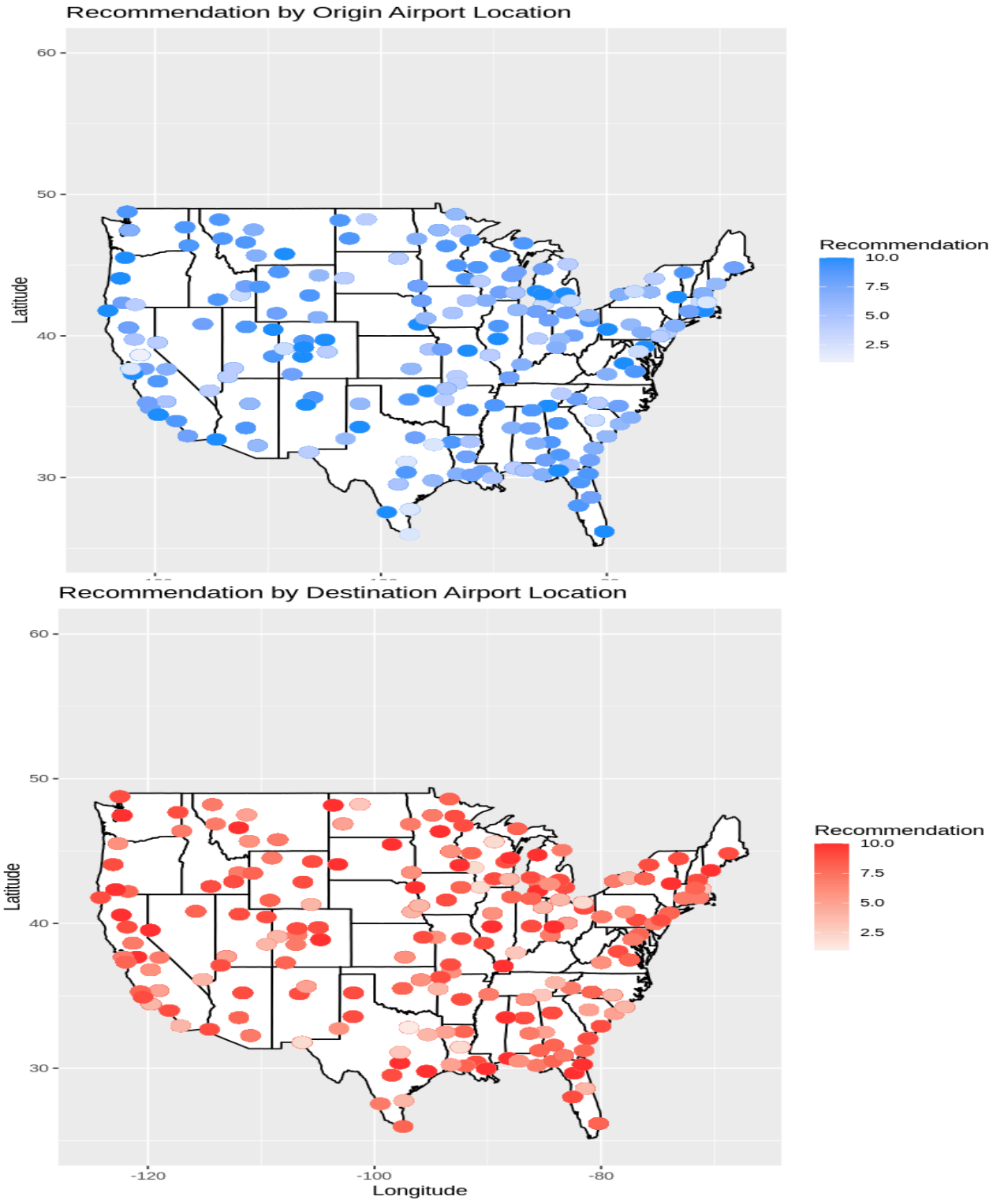
U.S Map of Origin and Destination Flights





We created U.S maps based on the origin and departure locations to get a better geographical visual sense of where customers travel to the most. It was pretty scattered around, but at least we can see the geographical regions of the most condensed flights which are the east and west cost, the southeast, and the north midwest.

U.S Map of Origin and Destination Flights by Score Rating



After sketching out the flights on the U.S maps for both arrival and departure flights, we factored in the recommendation score based on the customer's arrival and departure flight, respectively. The color of the dot indicates the value of the score, with the lightest color being the lowest score, and the darkest color being the highest score. The dots matched up closely with our previous bar charts that displayed the recommendation scores based on origin states and the destination states.