

**Southwest Airline Analysis Report**

IST 687 M004 Group 04

**Table of Contents**

[**Summary, Scope and Findings** 3](#_Toc26907824)

[**Project Summary** 3](#_Toc26907825)

[**Business Questions Addressed** 3](#_Toc26907826)

[**Key Findings** 3](#_Toc26907827)

[**Customers** 3](#_Toc26907828)

[**Partners** 4](#_Toc26907829)

[**Recommendations to Improve Satisfaction** 4](#_Toc26907830)

[**Methodology** 5](#_Toc26907831)

[**Data Cleaning** 5](#_Toc26907832)

[**Dataset Description** 5](#_Toc26907833)

[**Treatment for Missing Data** 5](#_Toc26907834)

[**Data Transformation** 6](#_Toc26907835)

[**Getting To Know The Data** 6](#_Toc26907836)

[**Descriptive Statistics & Visualizations** 6](#_Toc26907837)

[**Customers** 6](#_Toc26907838)

[**Partners** 9](#_Toc26907839)

[**Routes** 10](#_Toc26907840)

[**Satisfaction Analysis and Prediction** 10](#_Toc26907841)

[**Modeling Techniques & Visualizations** 10](#_Toc26907842)

[**Data Frame Preparations** 10](#_Toc26907843)

[**Linear Model (Multi-variate Regression)** 11](#_Toc26907844)

[**Apriori** 12](#_Toc26907845)

[**Low Satisfaction Route Mapping** 14](#_Toc26907846)

[**Low Satisfaction Deep-Dive (SVM Modeling)** 17](#_Toc26907847)

[**Annex 1 (Code)** 18](#_Toc26907848)

[**Annex 3 (Text Analysis)** 19](#_Toc26907849)

**Summary, Scope and Findings**

**Project Summary**

The purpose of this project was to conduct analysis on flight data collected through a survey and identify the primary drivers of passenger likelihood to recommend a given airline partner. In short, this analysis attempts to determine how best to improve customer satisfaction. The survey recorded 10282 observations across 32 variables pertaining to the customer demographics, behavior, flight information and most importantly, their likelihood to recommend the airline. We converted the likelihood to recommend values into detractor, passive and promoter categories in order to derive a Net Promoter Score (NPS). A Net Promoter Score is an index ranging from -100 to 100 that measures the willingness of customers to recommend a company’s services to others. Southwest Airline’s current score is a 7, which means there is room for improvement, but the score is on the positive side of the scale. Through descriptive statistics and modelling, we identified key differentiating features between customers, airline partners, and routes and how they relate to NPS. This analysis provides insights that may prove effective in improving customer satisfaction, and in turn, improve collective NPS.

**Business Questions Addressed**

Although many questions were asked throughout this analysis in an attempt to uncover as many insights as possible, the following questions guided the direction and framing of our analysis:

1. What is unique about highly satisfied customers (Age, gender, loyalty status, price sensitivity)? Can we re-create this for more customers?
2. What is unique about our unhappy customers? What is the driving factor of low satisfaction? What’s required to move them to become a promoter?
3. What is the current percentage breakdown of our partners regarding Net Promoter Score? Who is the best, who is the worst (as a percentage of their flights)?
4. Which routes/destination/departure locations are leading to low satisfaction?

**Key Findings**

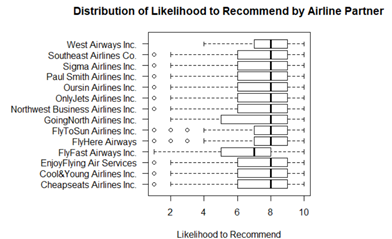
An extensive exploration of the analysis can be found in later sections of this report, but below are the executive level summary insights we believe are most salient to the analysis.

### **Customers**

* The promoters group follows a relatively normal distribution with a high count of passengers aged from 36 to 42 years old and the number gradually decreases toward both ends.
* The count of passives appears to be randomly scattered, we observed high ends and high counts of middle-aged passengers.
* The detractors are overwhelmed with teenagers and the elderly
* Find out young and elderly passengers’ wants and needs, fulfill them to improve their NPS

### **Partners**

The top performing partners regarding the highest average likelihood to recommend are: West Airways at 1st, Fly Here at 2nd, and Fly to Sun at 3rd.  The worst performers are (from worst to better) are Fly Fast, Going North and Cheapseats. As seen in the boxplot below, many partners have the same median score of 8, but there is some difference in the variance of their rating volatility.



## **Recommendations to Improve Satisfaction**

**Customer**

* Continue paying closed attention to middle-aged passengers to maintain and improve their NPS

**Partner**

* Northwest Airways and FlyFast should look into the reasons behind their passengers’ unlikelihood to recommend
* FlyWest should work on improving the needs of its large portion of passive passengers to improve their NPS

**Routes**

* Airline company should pay most attention on the Low Satisfaction Route.
* For Cheapseat Airline Inc, it should try to solve these three routes first: San Jose to Los Angeles, Houston to Chicago, and Orlando to San Juan. Why customer is not satisfied with these airlines.

**Methodology**

Our team’s methodology for approaching this problem was to break the task into a series of steps that allowed us to methodically explore the data and identify the appropriate analysis with the time available for the study. Each part of the analysis was assigned a primary analyst and a secondary analyst. This measure was taken to eliminate the chance of error, and to explore the data in multiple ways. Furthermore, in order to focus our analysis on likelihood to recommend, we organized our analysis along three lines of effort that relate directly to the three main players of air travel: the passengers, the airline, and the route. To that end, we used this three-way breakdown as a lens from which to view the dataset and provide context to the analysis. When appropriate, we grouped all observations by their net promoter stance (detractor, passive, promoter) and studied each subset individually to gain a better appreciation for each subset. This allowed us to compare the things that are going well, satisfied customers, satisfied partners, and satisfied routes, with those that were not.

**Data Cleaning**

**Dataset Description**

The data for our data set was collected via survey where 10282 observations across 32 variables pertaining to the customer demographics, behavior, flight information and most importantly, their likelihood to recommend the airline we recorded during a time frame covering roughly a two month period from 01 January 2014 until 09 March 2014.

**Treatment for Missing Data**

The following tables contains the variables in the dataset that had no recorded values. The table also shows what measures we used to replace these missing values. In all cases except free text, the mean was selected to replace the missing values. The median would have been a more conservative estimate of central tendency to use, specifically because the distribution of each of these variables was right skewed, however, we selected the mean because the values were still relatively low, and were under the benchmark we have established as a flight qualifying as delayed, which is 15 minutes or less is not considered late, and a flight being +/- 15 minutes is both not unreasonable, and not uncommon. Furthermore, any risk in using the mean over the median is mitigated by the fact that this replacement method impacted so little data (1.9% to 2.2%). No approximation measure was taken for the free text data, the black entries were simply ignored. This approach is reasonable as this is a character field containing stings of text from customer reviews, and no arithmetic was performed on these values.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Number of Occurrences / % of Observations** | **Replaced with what measure** | **Value of Replacement** |
| Departure.Delay.in.Minutes | 200 / 1.9% | Mean | 15.64233 |
| Arrival.Delay.in.Minutes | 226 / 2.2% | Mean | 15.29881 |
| Flight.time.in.minutes | 226 / 2.2% | Mean | 15.8218 |
| freeText | 10000 / 97.3% | None | None |

**Data Transformation**

During our analysis we constructed several subsets of data aligned with specific analysis tasks. For example, we made subsets to organize data by numeric types of values, category (factor) type values, and a free text dataset to capture the free text reviews (only 2.75% of the data had these reviews). Additionally, we made subsets of the data that included all the variables—both original and engineered—but partitioned the data for our three main groups of analysis: Detractors, Passives, and Promoters. We did this so we could better examine the uniqueness of each of these groups to better determine what makes them different. When appropriate (primarily during modeling) we used the complete set of data in order to provide a more robust models that are more prepared to make predictions.

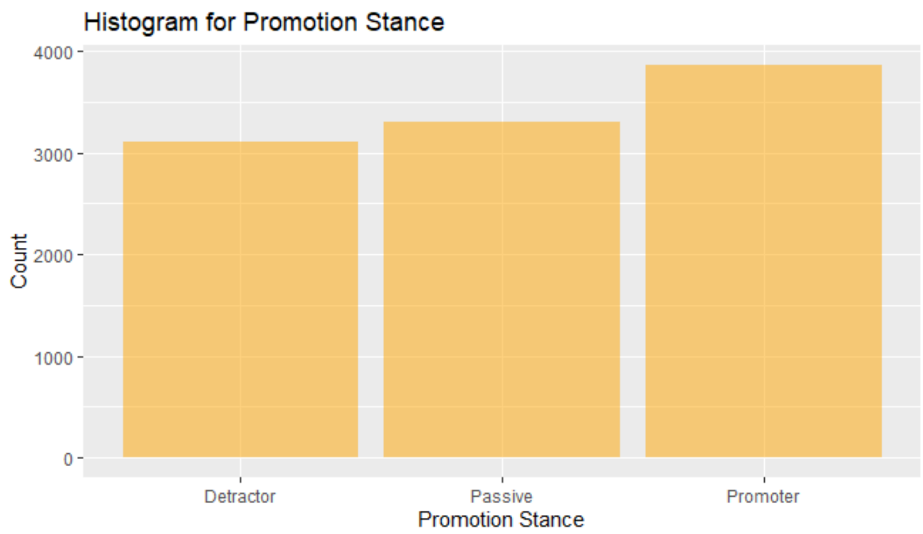
**Feature Engineering**

The original dataset had 32 feature variables but to improve our analysis and make it easier to understand, we built an additional 4 variables. The most important transformation was transforming each customer’s likelihood to recommend into a Promotion Stance variable by assigning all likelihood to recommend values of 6 and below as “Detractors”, 7 and 8 as “Passives”, and 9 and 10 as “Promoters.” The next two variables, Departure Delay Severity and Arrival Delay Severity, represent the transformation of time values slit into delay time increments: “0 to 15 minutes”, “16 to 30 minutes”, “31 to 45 minutes”, “46 to 60 minutes”, “Greater than one hour”. Although this comes at a cost to precisely measuring time delays, it groups the occurrences into logical categories and makes it more usable for caparisons. We also made a variable titled “Routes” to aggregated origin locations with destination locations. We built this feature so we could assess if particular routes are more likely to yield low or high satisfaction. Finally, we created a feature titled Year of Users that captured the time between the survey (2014) and the year of the customer’s first flight. We did this to help determine how long each customer has been a customer.

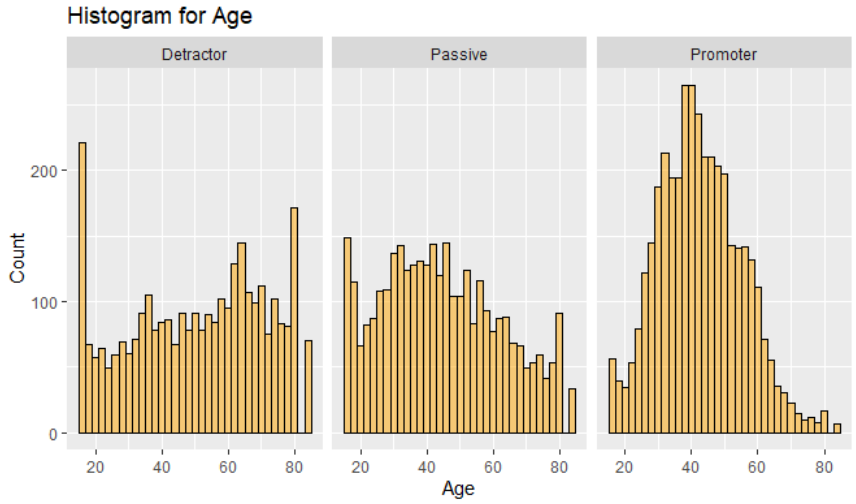
**Getting To Know The Data**

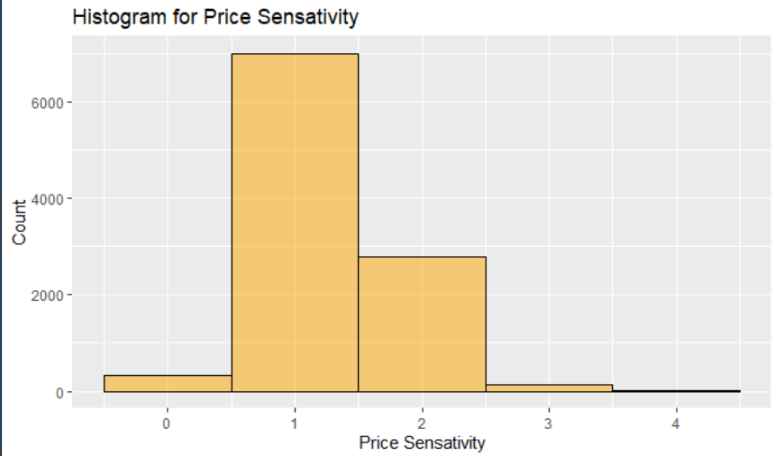
**Descriptive Statistics & Visualizations**

### **Customers**

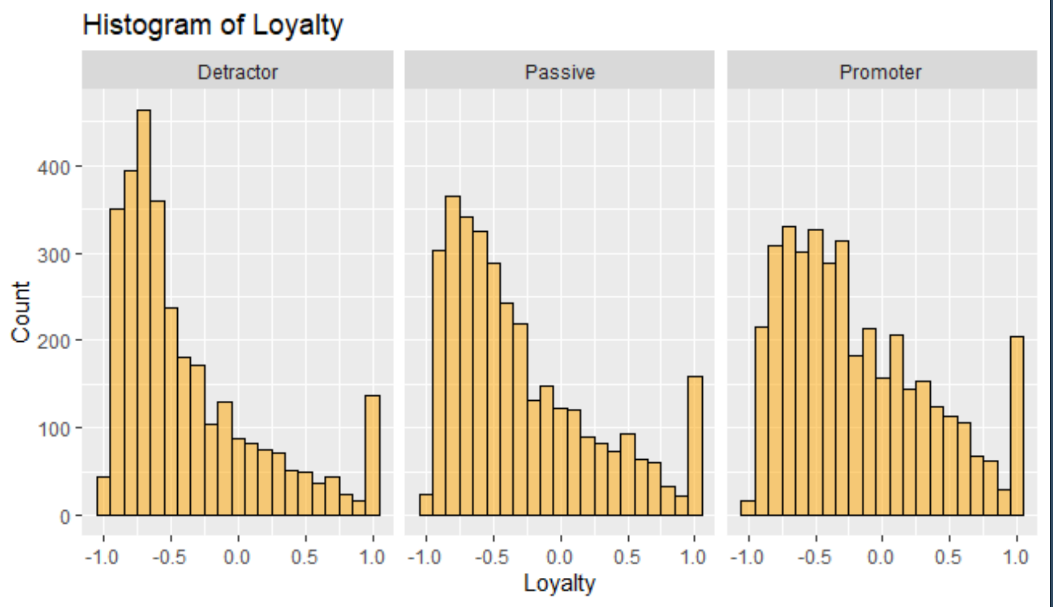
To examine passenger satisfaction, we partitioned the data by likelihood to recommend. To do so, we assigned a score of 1 to 10 from 1 being the least likely to recommend to 10 being the most likely to recommend. We then converted these values into a new variable titled “promotion stance” which binned each customer into three categories: promoters with a score of 9 or 10, passives with a score of 7 or 8, and detractors with a score of 1 to 6. The following variables were used to evaluate customer demographics: Age, Gender, No. Of Flights, Shopping Amount at Airport, Price Sensitivity, Eating and Drinking at Airport, Class, Loyalty, Airline Status, Type of Travel, Eating and Drinking a the Airport, Total Frequent Flier Accounts. Listed below is the summary of some of these variables.

The most import piece of contextual data going forward is promotion stance split across all of the data. The chart to right highlights the relatively even breakdown between Detractors (30.3%), Passives (32.1%), and Promoters (37.6%) across all customers in the data.

Histograms of the three promotion stance groups vs. age finds that the promoter group follows a relatively normal distribution with a high count of passengers aged from 36 to 42 years old and the number gradually decreases toward both ends. The count of passives appears to be randomly scattered, we observed high ends and high counts of middle-aged passengers. The detractors are overwhelmed with teenagers and the elderly, possibly due to their intolerance with airline issues such as a delay.



Price sensitivity reflects the grade to which the price affects a customer purchasing behavior. The price sensitivity has a range from 0 to 5, with 5 being the most price sensitive. We observed that the count of price sensitivity is the highest at 1, with a moderate count at 2 and very little at 0 or 3.

An index of loyalty ranging from -1 to 1 that reflects the proportion of flights taken on other airlines versus flights taken on this airline. A higher index means more loyalty. A substantial number of passengers have a loyalty ranging from -1 to 0 regardless of promotion stance. For those who have a loyalty greater than 0, most of them have a loyalty of 1. This suggests that our customers are not that loyal, but when they are, they are very loyal.

The table below highlights the split Airline status across each of the promotion stances and provides an idea on opportunities for improvement. For example, most of the Blue status is made up of Detractors. This suggest newer flyers are not overly impressed with their level of service. On the other end of the spectrum, it shows that older customers with a Silver status, are overwhelmingly Promoters. This analysis suggests an audit on benefits associated with level of status should be reviewed to identify potential opportunities to reconciliate differences.

With regard to class of travel, business class travelers disproportionately make-up the Promoter customers. Conversely, customers conducting personal travel are the largest population of Detractors. Further analysis should be done to review the policies and benefits surrounding these traveling categories with a focus on making personal class travel more consistent with business class.

Other Statistics:

* Gender: Female passengers constituted 56.5% with the rest being males.
* Most passengers had their first flight in 2003, with a similar amount of passengers had it from 2004 to 2012.
* Most of them fly within 30 times per year and the number of flights gradually decreases after 30.
* In regard to airport shopping, most people spent less than $100 dollars. For spending on eating and drinking, 75% of the passengers spent less than $90 dollars and 99% spent less than $240.
* The number of frequent flyer accounts averages to one, with 52% of all customers having zero accounts. This pattern is the same across all promotion stances.
* Among all flights, about 2% was canceled.

**Partners**

Splitting passenger’s status by partner, we observed that FlyHere Airlines has the highest percentage of promoter and Northeast Airways has the lowest; FlyWest Airways has the highest percentage of passive and Southeast Airlines has the lowest; Northwest Airways has the highest percentage of detractors and Flyfast Airways has the lowest.



**Routes**

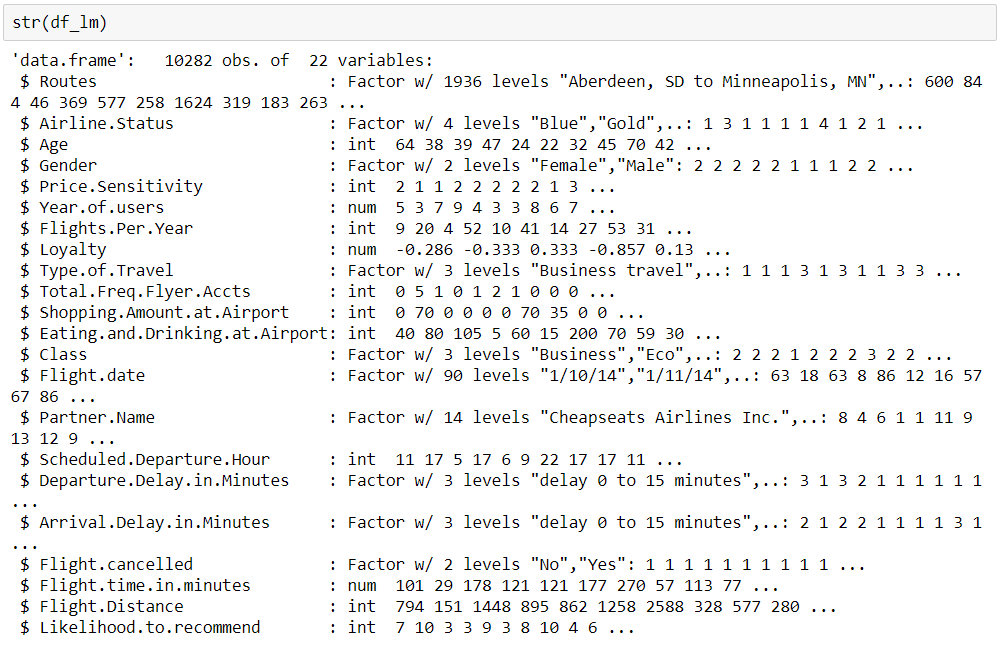
Travel on each day of the month does not vary significantly even though it is the lowest on the two days before the last day of the month. Most scheduled departure hour occurred between 6 am to 8 pm. Roughly 2/3 of the flights departed on time or within 15 minutes and 74% arrived on time or within 15 minutes. A majority of the flights had a flight time within 150 minutes and a flight distance within 1,000 miles.

**Satisfaction Analysis and Prediction**

## **Modeling Techniques & Visualizations**

### **Data Frame Preparations**

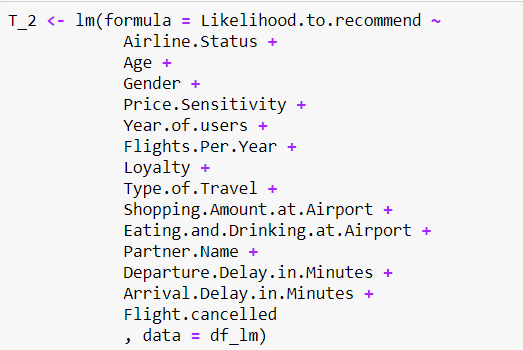
We began our model using the following data frame: (next page)

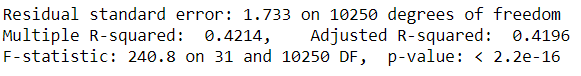
****

Although a Departure Delay Severity feature was already made, for this model we created another feature which is a reduced version of the Departure Delay Severity feature breaking the delay times into only three levels (0-15, 16-30, MoreThan30) on an interval of 15. As we know the actual departure time of a plane is usually a little later than expected, because they need to do some preparation for the flight. And the delay in departure will impact the arrival time, so the arrival delay in minutes can be divided into the same three groups.

### **Linear Model (Multi-variate Regression)**

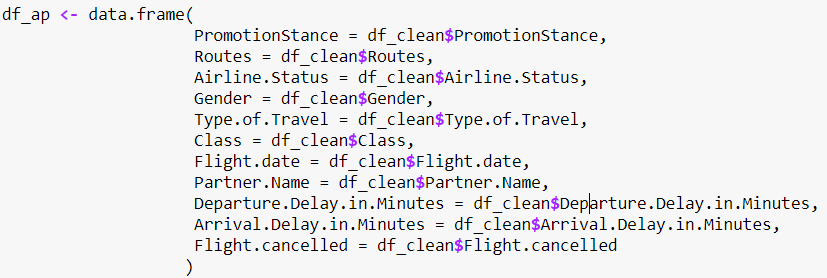
The objective of the linear model was to determine the most relevant variables that explain the variability in a customer’s likelihood to recommend. We used a Stepwise method to optimize the model. This method includes adding variables to the model one at a time to determine how they impact the performance of the model. Below is the final Linear Model and its performance measures.

****

**C:\Users\katie\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\74770CCE.tmp**

### **Apriori**

After linear regression we used the Apriori function to find factors that impact a customer’s promotion stance. Below is the data frame used for this:

****

Based on the support value of each rules we choose these factors as the predictors:

Airline.Status

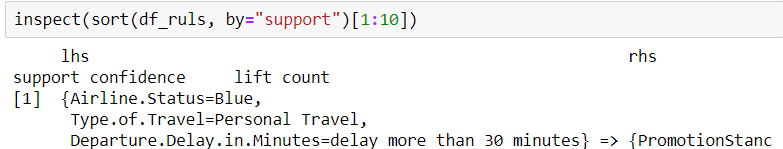
Type.of.Travel

Departure.Delay.in.Minutes(Arrival.Delay.in.Minutes)

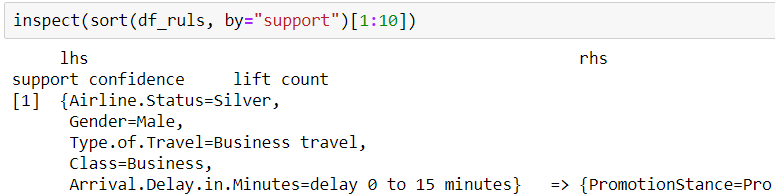
Class

gender

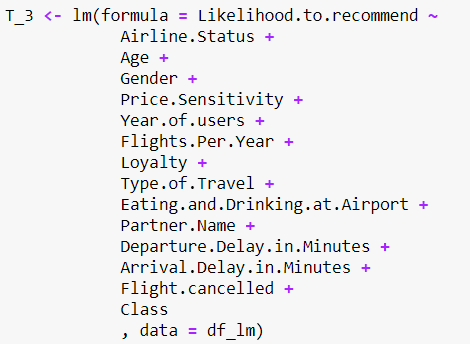
Below is the output for PromotionStance=Detractor:

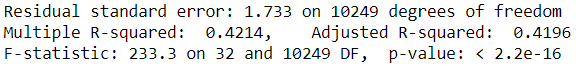
****

Below is the output for PromotionStance= Promoter:

****

Based on the results of our

****

**C:\Users\katie\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\CDD19892.tmp**

Using the adjusted R-squared stay the same, which means that this regression have the similar power as its previous version.

**Low Satisfaction Route Mapping**

In this part we will visualize low satisfaction route map. According to the dataset, almost flights are during in January and April, so in this part, we do not discuss the influence of season.

1. In this chart below, we summarize how many flights of each company in this period. As seen in the table below, the Cheaspseats Airline Inc takes the most Low satisfaction flights.

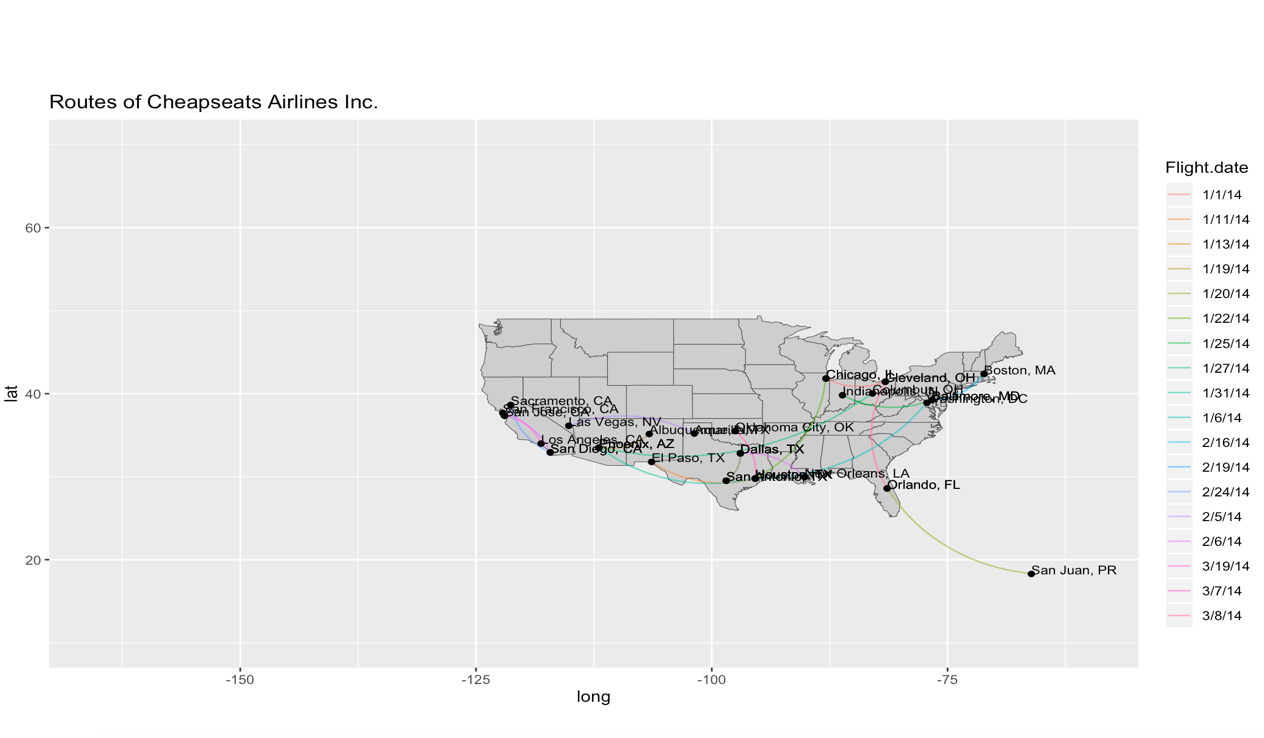
|  |  |
| --- | --- |
| **Partner Name** | **Low satisfaction flight lines** |
| Cheapseats Airlines Inc. | 700 |
| Cool&Young Airlines Inc. | 31 |
| EnjoyFlying Air Services | 141 |
| FlyFast Airways Inc. | 490 |
| FlyHere Airways | 51 |
| FlyToSun Airlines Inc. | 75 |
| GoingNorth Airlines Inc. | 55 |
| Northwest Business Airlines Inc. | 313 |
| OnlyJets Airlines Inc. | 116 |
| Oursin Airlines Inc. | 308 |
| Paul Smith Airlines Inc. | 141 |
| Sigma Airlines Inc. | 447 |
| Southeast Airlines Co. | 243 |
| West Airways Inc. | 1 |

1. Number of each low satisfaction routes:

|  |  |
| --- | --- |
| **Origin State-Destination State** | **count** |
| California-California | 131 |
| Texas-Texas | 124 |
| Texas-Louisiana | 39 |
| Georgia-Florida | 32 |
| California-Arizona | 31 |
| California-New York | 28 |
| Florida-Georgia | 27 |
| Illinois-California | 27 |
| Texas-California | 27 |
| Texas-Georgia | 27 |
| California-Colorado | 26 |
| Colorado-California | 24 |
| Florida-Florida | 24 |
| New York-Florida | 23 |
| California-Illinois | 21 |
| New York-Illinois | 21 |
| Texas-Colorado | 21 |
| Texas-Oklahoma | 21 |
| California-Nevada | 20 |
| Nevada-California | 20 |
| Texas-Florida | 20 |

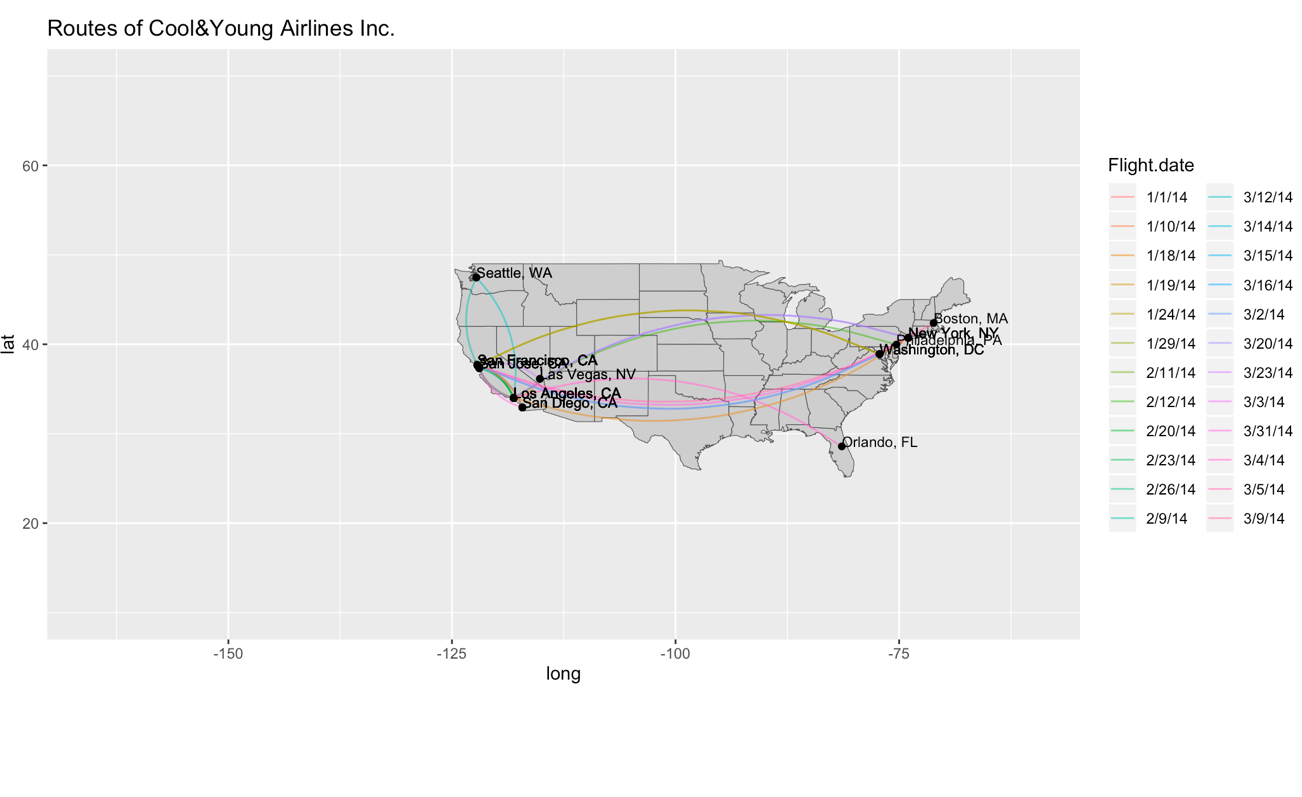
1. Low satisfaction map colored by date and all flights routes distribution in United States

Low satisfaction Routes of Cheapseats Airlines Inc.



There are no specific rules to see from Cheapseats Aillines. Low satisfaction flights of this company distribute all over the country.

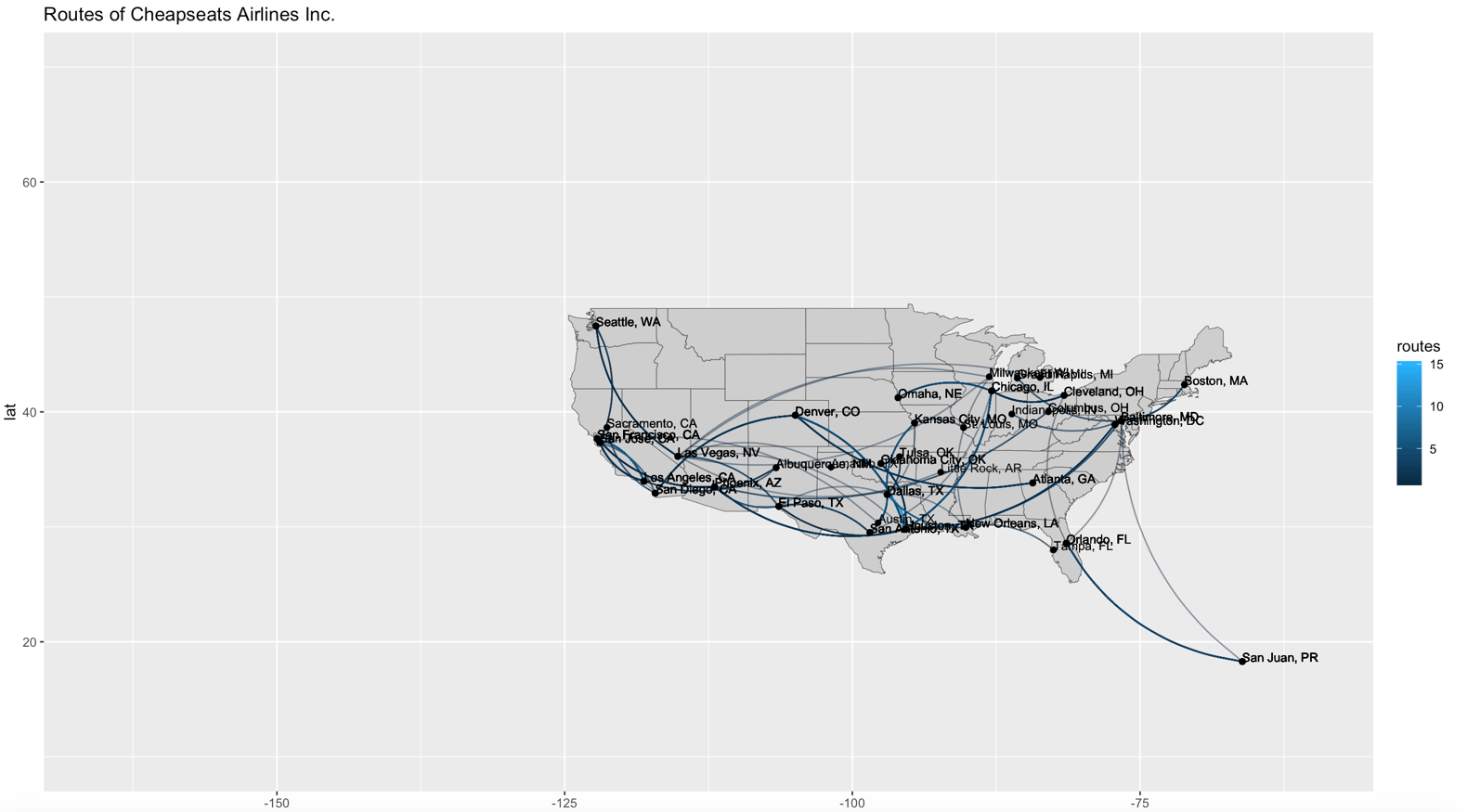
Low satisfaction routes of Cool & Young Airline



Low satisfaction of Cool and Young Airline concentrate on those flights across the East of United States and the West of United States. Flights set off from California states require attention.

1. Low satisfaction map colored by number of routes

Take Cheapseats Airline Inc as example:



From the map above, we can see that routes of San Jose to Los Angeles, Houston to Chicago, and Orlando to San Juan are the worst in this company.

**Low Satisfaction Deep-Dive (SVM Modeling)**

SVM:

Step 1: Customer classification, there are two kinds of customer classification in this model: detractor and promoter. We need to use all of the other variables to predict this customer is detractor or promoter.

Step 2: Sampling data, train\_data takes 70% and test\_data takes 30%.

Step 3: Data mining process:

svmOutput<-ksvm(PromotionStance~.,data=train\_data, kernel="rbfdot", kpar="automatic",C=50,cross=3,prob.model=TRUE )

Step 4: Calculate Accuracy:

test\_data.PromotionStance

svmPred Detractor Promoter

Detractor 701 163

Promoter 212 993

> accuracy

[1] 0.818753

81% accuracy seems pretty good, but we need to see is there any room to increase the accuracy.

Step 5: Dropping variables.

After dropping 3 variables , our model’s accuracy is **82.1653%,** we don’t want this model overfitting, so we keep this model:

PromotionStance~Age+Gender+Price.Sensitivity+Year.of.First.Flight+Flights.Per.Year+Loyalty+Type.of.Travel+Total.Freq.Flyer.Accts+Shopping.Amount.at.Airport+Eating.and.Drinking.at.Airport+Class+Partner.Name.

# **Annex 1 (Code)**

All code published on MIDST

**Annex 2 (Description of attributes in the survey dataset)**

Attributes we engineered

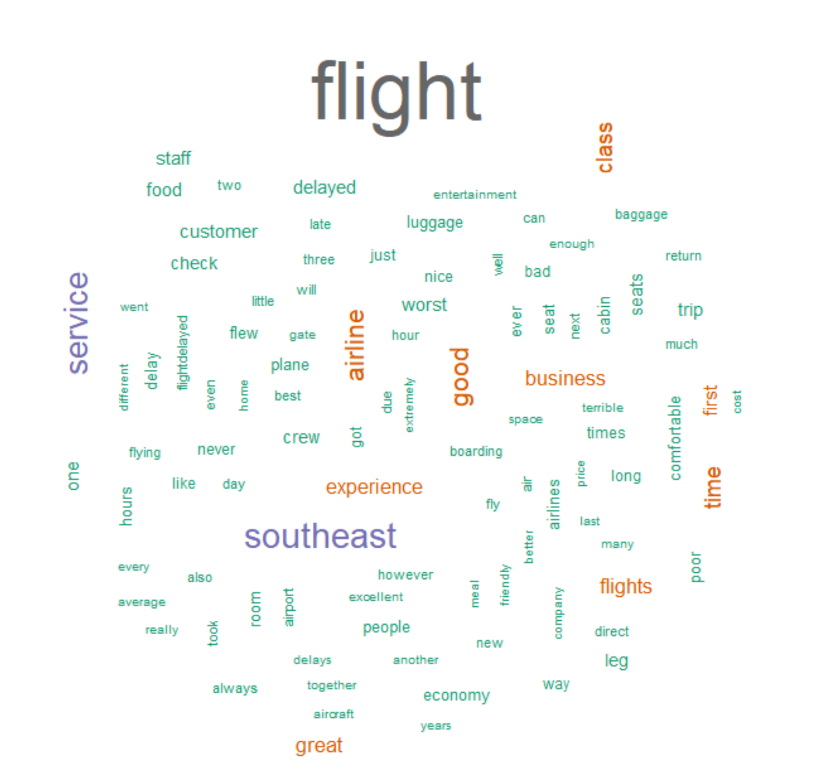
1. **Promotion Stance -** Assigned likelihood to recommend values of 6 and below as “Detractors”, 7 and 8 as “Passives”, and 9 and 10 as Promoters.  ​
2. **Departure Delay Severity**and **Arrival Delay Severity**, represent the transformation of time values slit into delay time increments: 0 to 15 minutes, 16 to 30 minutes, 31 to 45 minutes, 46 to 60 minutes, Greater than one hour.
3. **Routes**- Aggregated origin locations with destination locations.  We built this feature so we could assess if particular routes are more likely to yield low or high satisfaction.​
4. **Year of Users** - Captured the time between the survey (2014) and the year of the customer’s first flight.  We did this to help determine how long each customer has been a

Attributes provided in the dataset

1. **Likelihood to Recommend –** rated on a scale of 1 to 10, which shows how likely the customer is to recommend the airline to their friends (10 is very likely, and 1 is not very likely).
2. **Airline Flyer Status** – each customer has a different type of airline status, which are platinum, gold, silver, and blue (based on level of travel with the airline)
3. **Age** – the specific customer’s age. Ranging from 15 to 85 years old.
4. **Gender** – male or female.
5. **Price Sensitivity** – the grade to which the price affects to customers purchasing. The price sensitivity has a range from 0 to 5.
6. **Year of First Flight** – this attribute shows the first flight of each single customer. The range of year of the first flight for each customer has been started in 2003 until 2012.
7. **Flights Per Year** – The number of flights that each customer has taken in the most recent 12 months. The range starting from 0 to 100.
8. **Loyalty** – An index of loyalty ranging from -1 to 1 that reflects the proportion of flights taken on other airlines versus flights taken on this airline. A higher index means more loyalty.
9. **Type of Travel** – One of business travel, mileage tickets, or personal travel (ex. vacation)
10. **Total Frequent Flyer Accounts** – How many frequent flyer accounts the customer has.
11. **Shopping Amount at Airport** – The spending on non-food & services at the airport (in $)
12. **Eating and Drinking at Airport** – The spending on food/drink at the airport (in $).
13. **Class** – three different kinds of service level (business, economy plus, and economy).
14. **Day of Month** –the traveling day of each costumer (ranges from 1 to 31).
15. **Flight date** – the passenger’s flight date of travel.
16. **Partner Code** – This airline works with wholly- and partially-owned subsidiary companies to deliver regional flights. For example, AA, AS, B6, and DL.
17. **Partner Name** – These are the full names of the partner airline companies.
18. **Origin City** – the place where passenger departed from. For example, Boston MA.
19. **Origin State** – the place where passenger departed from. For example, Texas.
20. **Destination City** – the place to which passenger travels to. For example, Boston MA.
21. **Destination State** – the place to which passenger travels to. For example, Texas.
22. **Scheduled Departure Hour** – the specific time at which the plane was scheduled to depart.
23. **Departure Delay in Minutes** – How long the flight’s departure was delayed, when compared to schedule.
24. **Arrival Delay in Minutes** – How long the arrival was delayed.
25. **Flight Cancelled** – occurs when the airline does not operate the flight.
26. **Flight time in minutes** –the length of time, in minutes, to reach the destination.
27. **Flight Distance** – the distance between the departure and arrival destination.
28. **Comment** – a free form text field of the passenger comment, with respect to the flight.

# **Annex 3 (Text Analysis)**

The most powerful text analysis that could be performed is sentiment analysis of the reviews. However, the likeliness to recommend value is the most informative sentiment indicator. As such, we decided to conduct a broader approach to the text analysis by using a Term Document Matrix (TDM) to look at the frequency of certain words, and a word cloud to visualize the results. This analysis is perhaps more artful than insightful, the TDM does help illustrate some of the most frequently used terms in the free text reviews.

TermDocumentMatrix (terms: 2733, documents: 1)  
Non-/sparse entries: 2733/0  
Sparsity : 0%  
Maximal term length: 54  
Weighting : term frequency (tf)  
Sample :

Docs

Terms 1  
 airline 50  
 business 32  
 class 39  
 flight 209  
 flights 34  
 good 50  
 great 33  
 service 68  
 southeast 75  
 time 36