Brian Schramke; Harshita Asnani; Zixuan Liu; Yuying Shi; Nikita Rao Sira

[Email address]

Abstract

[Draw your reader in with an engaging abstract. It is typically a short summary of the document.   
When you’re ready to add your content, just click here and start typing.]

IST 687 Final project Report – group 3

Analysis of Airline Passenger Data

Contents

[Introduction 2](#_Toc26808850)

[Business Questions 2](#_Toc26808851)

[Questions that we tried to answer from the data exploration 2](#_Toc26808852)

[Data Preparation 2](#_Toc26808853)

[NA Identification and elimination 2](#_Toc26808854)

[Addition of columns 3](#_Toc26808855)

[Quantile Grouping 3](#_Toc26808856)

[Grouping of Data in the delay column 3](#_Toc26808857)

[Date Columns 4](#_Toc26808858)

[Promoter.Score Column 4](#_Toc26808859)

[Conversion 4](#_Toc26808860)

[Descriptive Statistics 4](#_Toc26808861)

[Modeling 13](#_Toc26808862)

[Association Rules Mining 13](#_Toc26808863)

[Top Six Airlines 13](#_Toc26808864)

[Northwest Business Airlines 15](#_Toc26808865)

[FlyFast Airways 16](#_Toc26808866)

[Personal Travelers 18](#_Toc26808867)

[Business Travelers 19](#_Toc26808868)

[Text mining and Sentiment Analysis 21](#_Toc26808869)

[Linear Modeling 36](#_Toc26808870)

[Recommendations 45](#_Toc26808871)

# Introduction

# Business Questions

## Questions that we tried to answer from the data exploration

1. Which airline has the most number of delays?
2. What type of customers are majorly promoters/detractors/passive?
3. Why state has the highest number of customers?
4. What locations in US have the highest number of delays?
5. Is there some relationship between delays and the day of week?
6. Which airline status do majority of our customers belong to?
7. Which partner has the highest number of delays?
8. Which partner has the highest number of detractors?
9. Which partner has the greatest number of customers?
10. Which partners are doing well with respect to NPS score and which partners aren’t ?
11. Does gender and likelihood to recommend have any relationship?
12. Does gender influence the amount of shopping/drinking at airport?

# Data Preparation

## NA Identification and elimination

To identify any NAs present in the data, a function was created using a for-loop to iterate through the multiple columns and return the column names with NAs.

check\_na<-function(df){

#Excute for loop to identify where NAs are present in data set

for (Var in names(df)) {

missing <- sum(is.na(df[,Var]))

if (missing > 0) {

print(c(Var,missing))}}}

[1] "Departure.Delay.in.Minutes" "192"

[1] "Arrival.Delay.in.Minutes" "220"

[1] "Flight.time.in.minutes" "220"

[1] "freeText" "10000"

Arrival and departure delays when the flight was cancelled were replaced with NAs to aid in filtering for later calculations. Any NAs for arrival or departure delay where the flight was not cancelled were replaced with zero. Free text NAs where left to aid in filtering. Nested for-loops were used to iterate through passengers with missing flight times and replace it with an average flight time of known flights between the given city pair. The function is given below.

missingflighttime <- df%>%

filter(is.na(df$Flight.time.in.minutes)&df$Flight.cancelled=="No")

#Calculate average flight time from completed segments between city pairs to fill in applicable missing flight times

#Iterate through origin cities with missing flight times

for (var1 in unique(missingflighttime$Origin.City)){

#Iterate through destination cities with missing flight times

for (var2 in unique(missingflighttime$Destination.City)){

#Calculate mean of know flight times between city pairs (origin to destination OR destination to origin) to replace NAs

df$Flight.time.in.minutes[df$Origin.City==var1&df$Destination.City==var2&df$Flight.cancelled=="No"]<-mean(df$Flight.time.in.minutes[(df$Origin.City==var1&df$Destination.City==var2)|(df$Origin.City==var2&df$Destination.City==var1)], na.rm=TRUE)}}

## Addition of columns

### Quantile Grouping

To ease the process of analyzing data we performed quantile grouping on few of the columns in data consisting of numeric values. Quantile grouping divides the data in a column into four classes. The first class ranges from 0th to the 25th percentile, the second class ranges from 26th to 50th percentile, the third class ranges from 50th to the 75th percentile and the last class consists of values above 75th percentile. List of columns which underwent quantile grouping is as under

flightTime

Age

FirstYearOfFlight

EatingAndDrinkingAtAirport

shoppingAmountAtAirport

loyaltyScore

flightsPerYear

departureHour

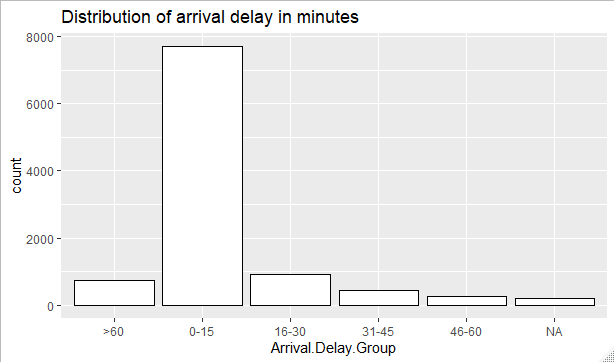
### Grouping of Data in the delay column

The data in the delay column is divided in 5 classes which are:

1. Flights which had 0-15 min delay
2. Flights which had 15-30 min delay
3. Flights which had 30-45 min delay
4. Flights which had 45-60 min delay
5. Flights which delay time greater than 60 mins.

The justification to this division is the distribution of the values in the column. As shown in fig(1) maximum number of flights fall in 0-15 min delay group. Thus, the division of values according to quantile grouping wouldn’t provide insights as appropriate as the above grouping. Also, as one can see the delay column consists of a few NA values which represent the flights which have a delay time but their status is cancelled.

Figure Arrival delay histogram



### Date Columns

The Flight.Date field was converted to a date variable and used to calculate the day of the week a flight occurred. This was done to potentially reveal any patterns associated with certain days of the week.

### Promoter.Score Column

Passengers with likelihood to recommend score between 0-6 are considered detractors, 7-8 are passives and 9-10 are promoters. This column helps in easy division and identification of a passenger based on his/her likelihood.to.recommend score.

## Conversion

**Factorization**

To create linear models, transaction data frames, and ease our understanding of data some columns were converted to factors. The list of columns converted to factors is as follows

DayOfWeek

MonthOfYear

DestinationState

OriginState

AirlineStatus

TypeofTravel

Class

FlightCancelled

PartnerCode

Gender

PriceSensitivity

TotalFreqFlyerAccts

# Descriptive Statistics

We began our analysis with plotting some basic graphs and deriving insights from the same to decide the directions of our analysis path ahead.

1. Age Group

The distribution of data points in all four age groups is almost identical. There are no other useful insights we can derive from this data

Figure Age group histogram

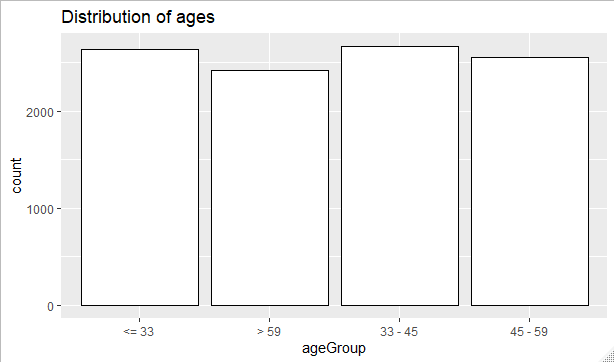
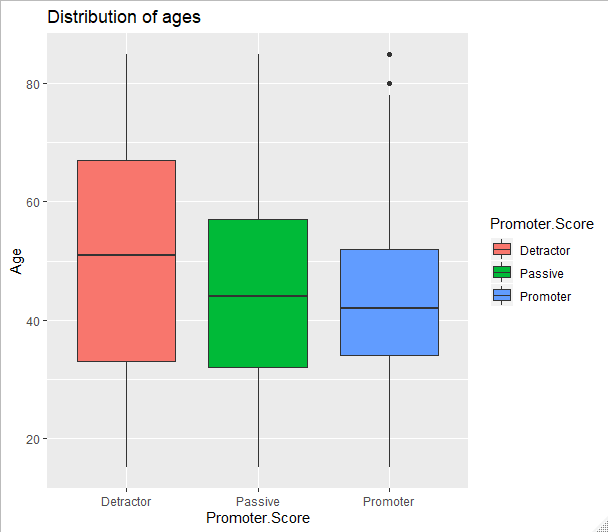


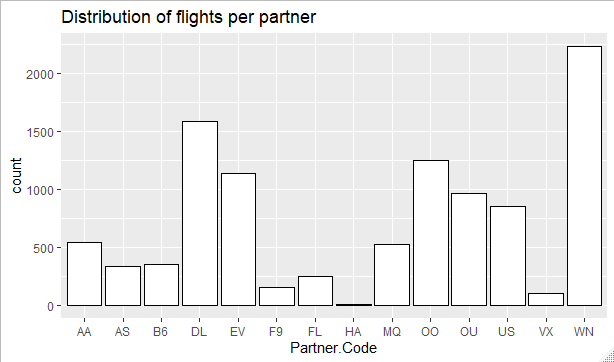
Figure Boxplot for age distribution for each promoter score



1. Number of flights per partner

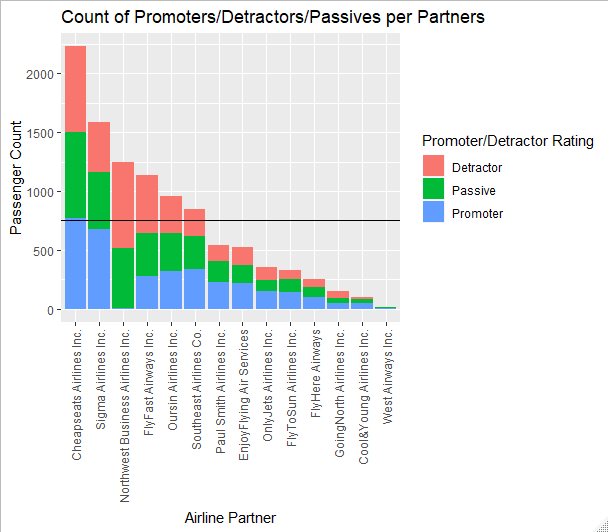
The distribution of data points is very unsymmetrical. As shown in the figure below except the top 6 partner airlines the number of data points available for rest of the partners is very less. Also, the top 6 partners constitute to almost 60% of data present. Thus, we are limiting our data to these 6 partner airlines for analysis.

Figure Histogram for flights per partner



We kept 750 data points as the cut-off value to consider a partner airline for further analysis. Thus, we selected our top six partner airlines as Cheapseats Airlines, Sigma Airlines, Northwest Business Airlines, FlyFast Airlines, Oursin Airlines and Southeast Airlines.

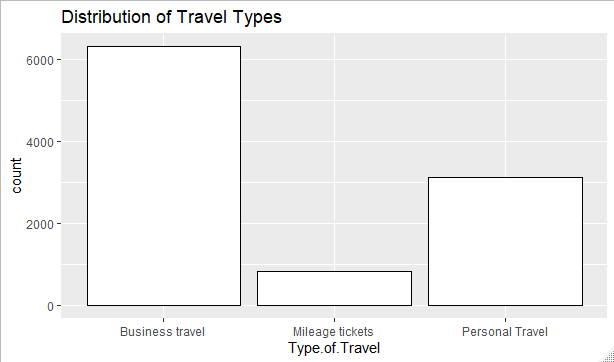
Figure Barplot of promoter score for each partner airline



1. Histogram for count of travel type

As it can be viewed from the plot below most of our data points fall into either business or personal travel types. Also, these two travel types constitute to more than 90% of our data points. Thus, we are limiting our analysis to business and personal travel types for top 6 partners.

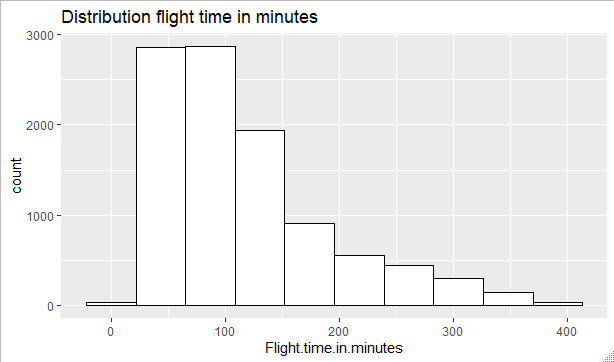
Figure Histogram for type of travel



1. Distribution of Flight time in minutes

The data points for flight time turned out to be a right skewed plot. There was ere not much insights that we could derive from this plot. Also we didn’t use much of this variable to perform our analysis.

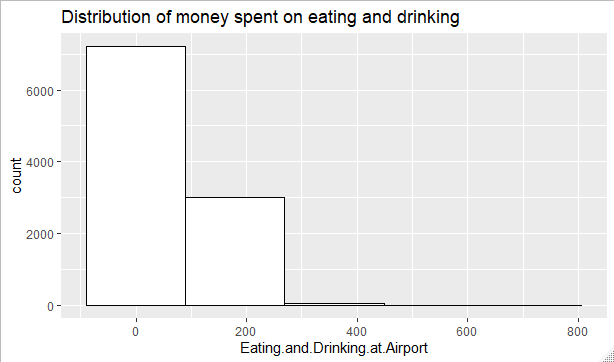
Figure Histogram for duration of flight time



1. Distribution of money spent on eating and drinking

More than 60% of our passengers spent between 0-100 dollars for the purpose of eating and drinking at the airport. No other insights derived.

Figure Histogram for money spent on eating and drinking



1. Promoter Score Distribution

The data points for each category are fairly identical. There is not much we can derive from this data. Thus, we subsetted our data to top 6 partners and focused on improving our likelihood of recommendation for them. We cannot make any conclusions from the below plot as the number of detractors, passive & promoters are approximately similar. We can dive deeper into the distribution of each of the category across airline status, travel type, delay category and other variables to understand more.

Figure Histogram for distribution of promoter score

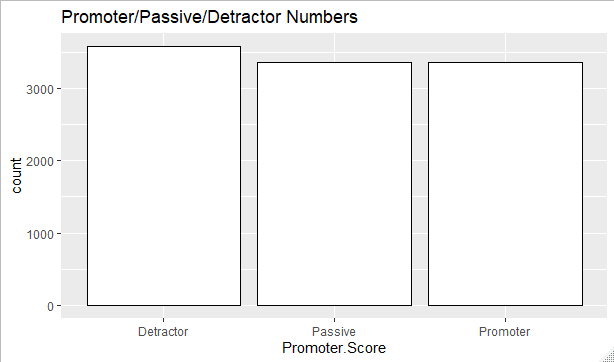
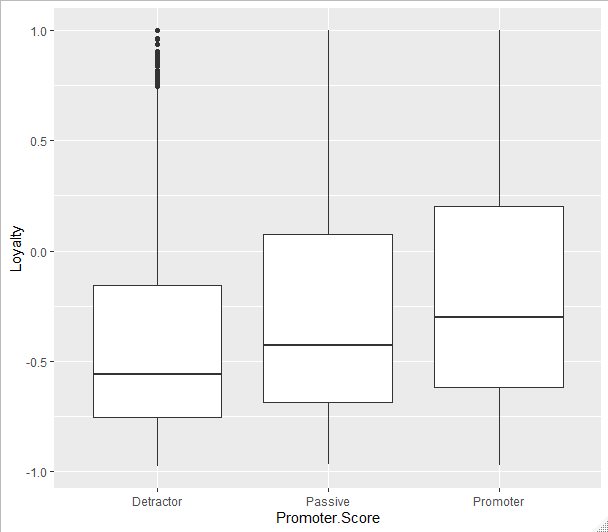


Figure Box plot for promoter score



1. NPS score by type of travel :

From this stacked bar chart, we can clearly see that a huge percentage of our business travel type customers are promoters and a huge proportion of our personsl type travel customers are detractors. Next step would be to dive deeper to understand why is the case.

Figure Bargraph of travel type by NPS Score

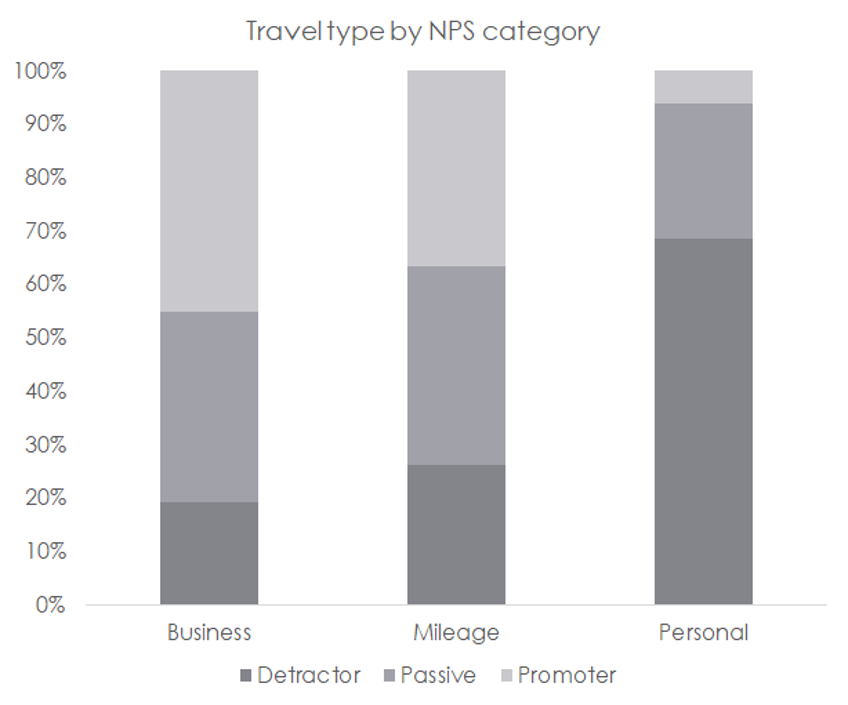
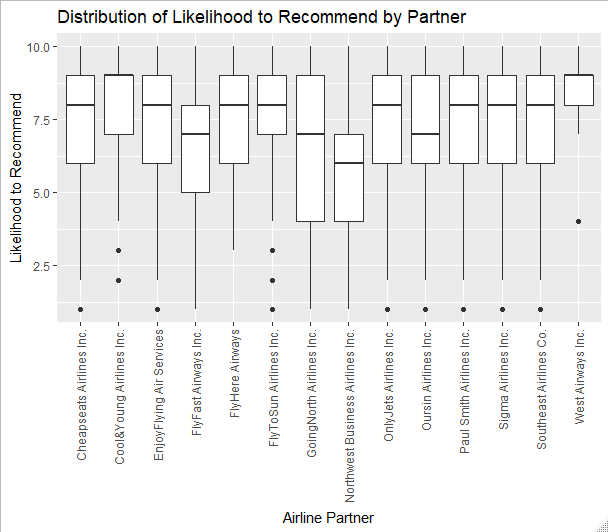
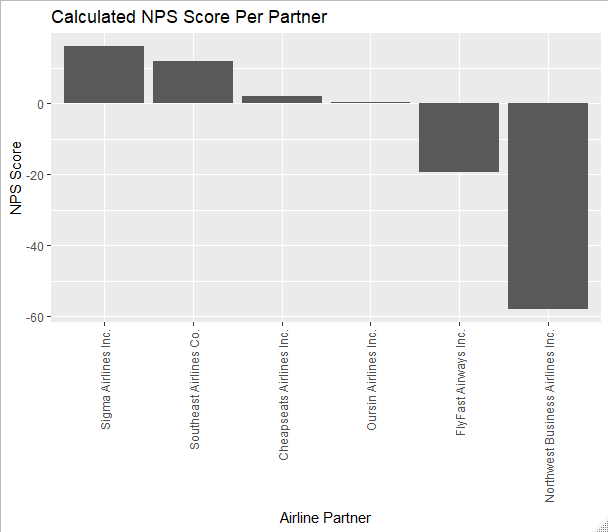


Figure Box plot for likelyhood to recommend by partner airline



We could conclude from the histogram of top six partner airlines that Northeast Business Airlines had the largest number of detractors and thus the least NPS score.

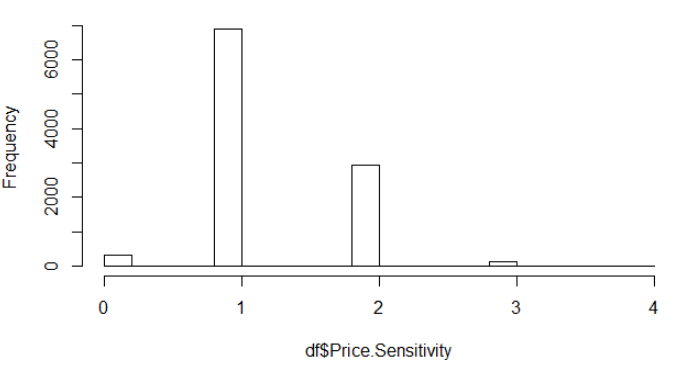
Figure Histogram of NPS Score for top six partners



1. Price senstitivity :

More than 90 % of data points fell into the one or two bucket for price sensitivity. There were no more insights derived.

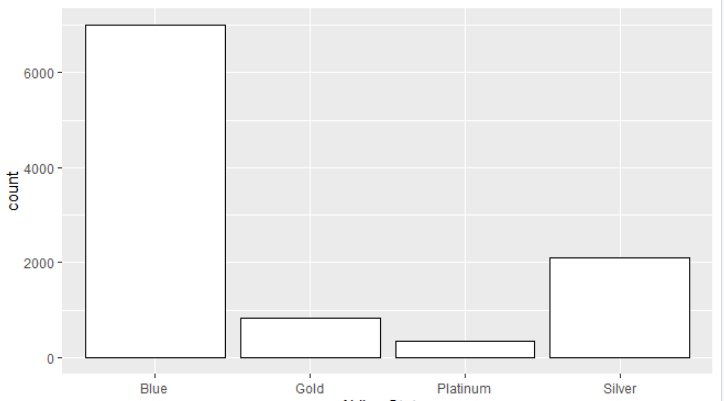
Figure Histogram for price sensitivity



1. Airline Status :

Majority of our customers travelling are of Blue status.

Figure Histogram for airline status



# Modeling

## Association Rules Mining

We began with association rules mining to help determine passenger attributes that predict promoter status. Several iterations were conducted including analyzing the top six partners by passenger count, Northwest Business Airlines, FlyFast Airways, personal travelers, and business travelers.

### Top Six Airlines

We first focused on passengers across the top six partners by passenger count. The first step was calculating the percentage of each type of promoter rating to determine which rules predicted a case at a greater rate than the average.

Detractor Passive Promoter

36.65 33.40 29.95

Rulesets were created by defining the right-hand side as the appropriate promoter rating to identify attributes that predict promoter rating. The rulesets are tabulated below in Table 1: Top Six Partners Promoter Rules, Table 2: Top Six Partners Passive Rules, and Table 3: Top Six Partners Detractor Rules) are below.

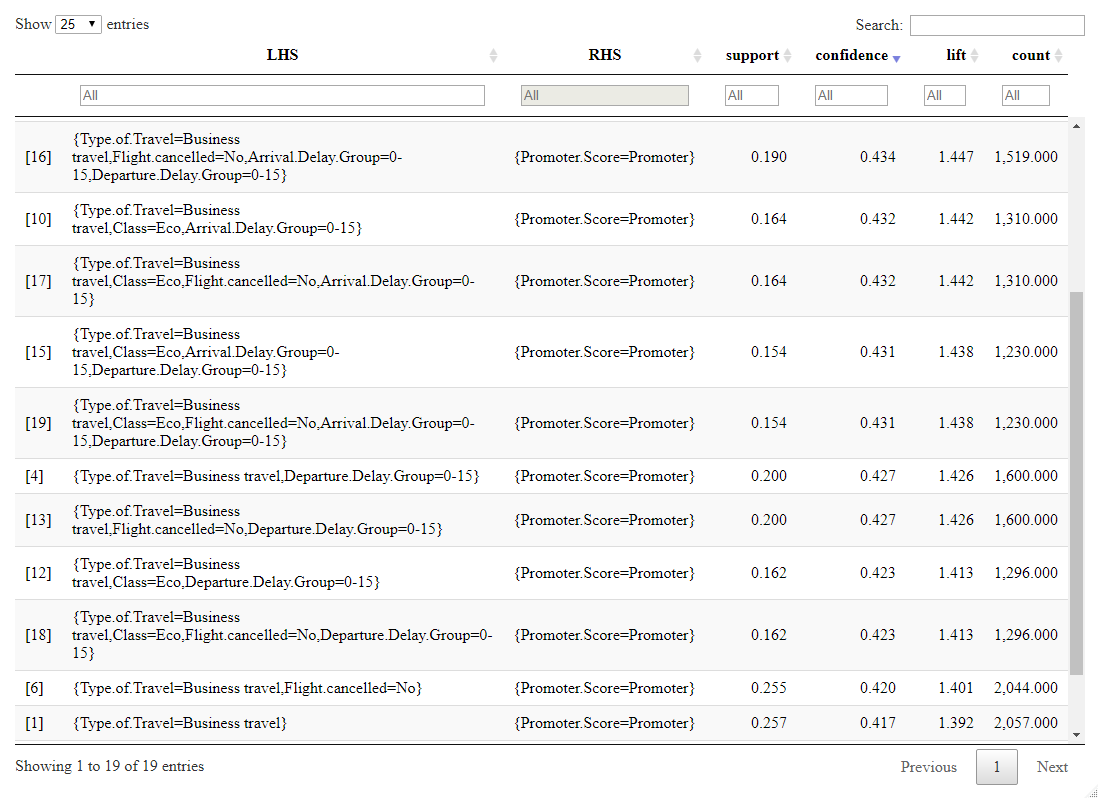
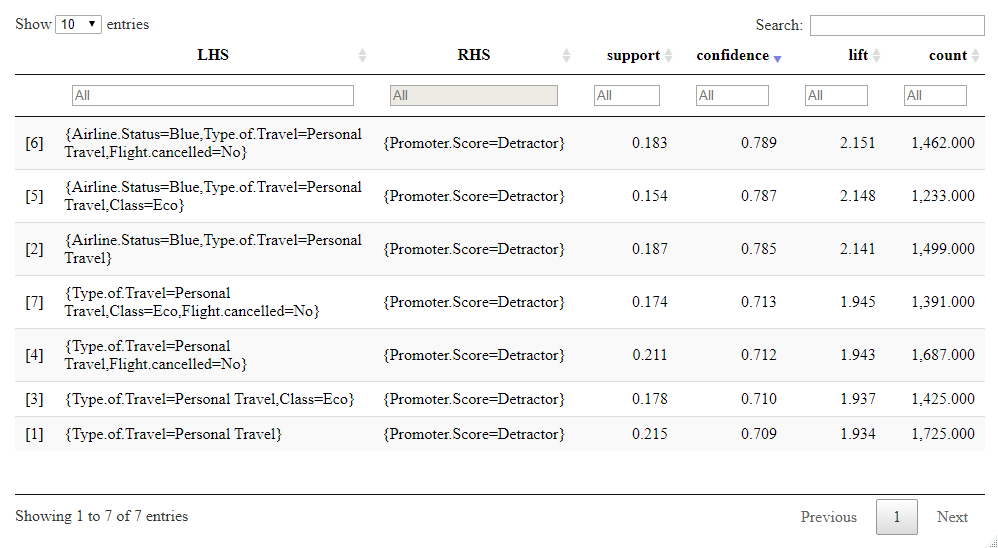
Table : Top Six Partners Promoter Rules

Table : Top Six Partners Passive Rules

Table : Top Six Partners Detractor Rules

The results across the top six airlines by passenger count indicate that business type travel is a strong indicator of having a promoter or passive rating, predicting the outcome 14% and 5% above the respective average. Business type travel is also a common parameter among the strongest rules. The analysis also indicated that personal type travel is a strong predictor of being a detractor, predicting 34% above the average.

### Northwest Business Airlines

Northwest Business Airlines and FlyFast Airways were analyzed as they were the partners with lowest NPS score amongst the top six partners. The percentage of promoter score by partner was calculated the to establish minimum confidence values in Table 4: Promoter Score Percentages by Partner.

Table : Promoter Score Percentages by Partner

Detractor Passive Promoter

Cheapseats Airlines Inc. 32.71 32.57 34.72

FlyFast Airways Inc. 43.54 32.19 24.27

Northwest Business Airlines Inc. 58.73 40.46 0.80

Oursin Airlines Inc. 33.40 32.88 33.71

Sigma Airlines Inc. 26.83 30.24 42.93

Southeast Airlines Co. 27.38 33.25 39.37

Rulesets were created by defining the right-hand side as the appropriate promoter rating to identify attributes that predict promoter rating. There was insufficient data to create rules for promoters on Northwest Business Airlines. The rulesets are tabulated below in Table 5: Northwest Passive Rules and Table 6: Northwest Detractor Rules

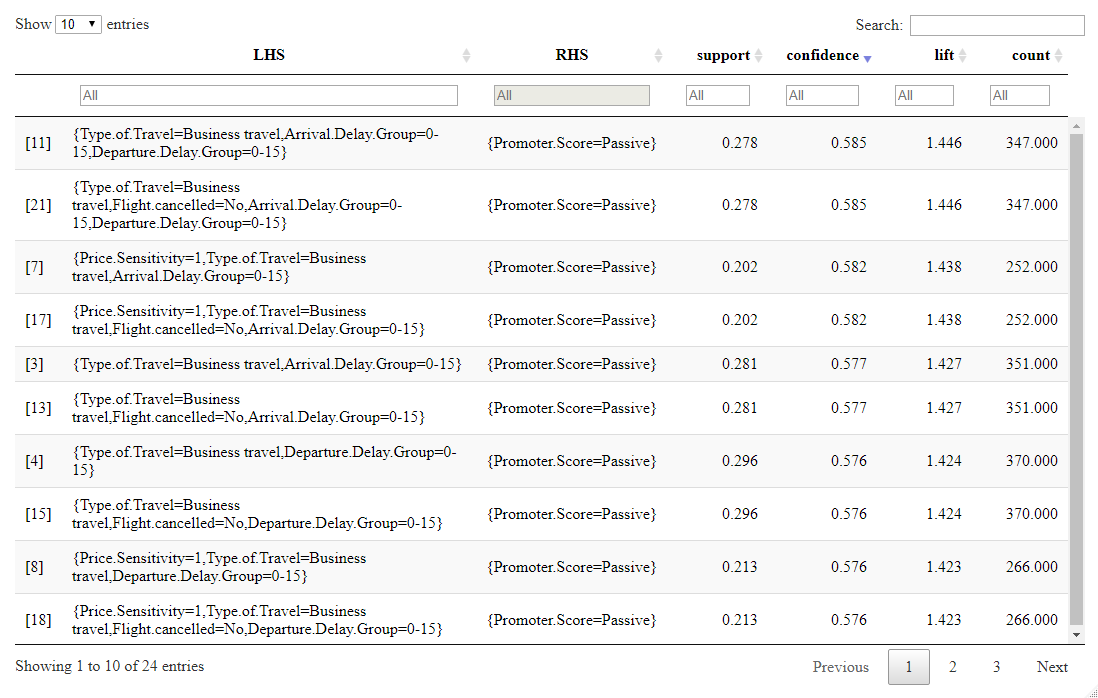
Table : Northwest Passive Rules

Table : Northwest Detractor Rules

The results indicate the business travel is a strong predictor of being a passive customer on Northwest Business Airlines predicting 16% above the average. Personal travel and personal travel coupled with Blue status indicate being a detractor predicting 32% and 37% above the average, respectively.

### FlyFast Airways

The analysis was repeated for Fly fast airways using minimum confidence values determined by the averages in Table 4: Promoter Score Percentages by Partner. The rulesets are tabulated in Table 7: FlyFast Promoter Rules, Table 8: FlyFast Passive Rules, and Table 9: FlyFast Detractor Rules.

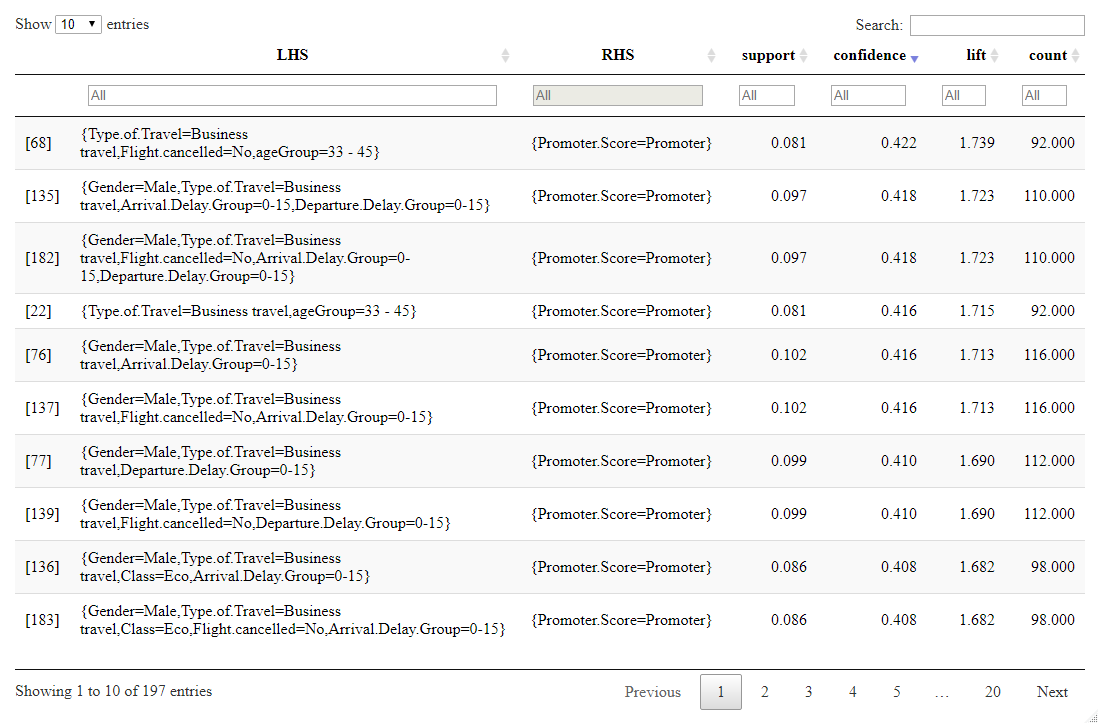
Table : FlyFast Promoter Rules

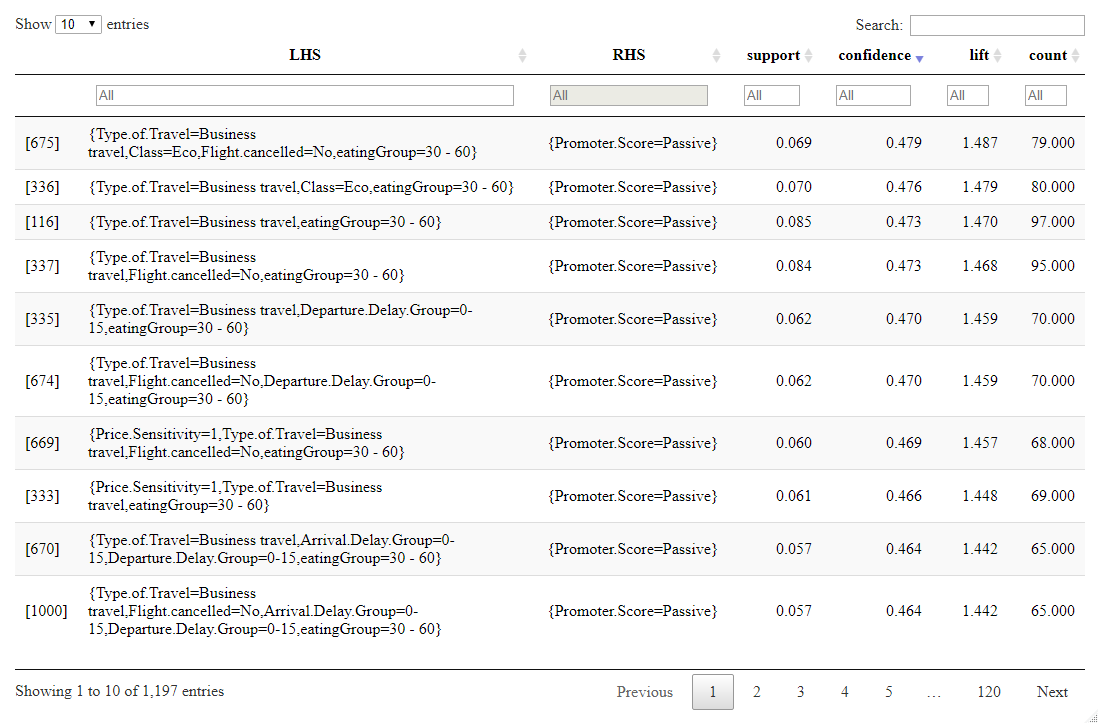
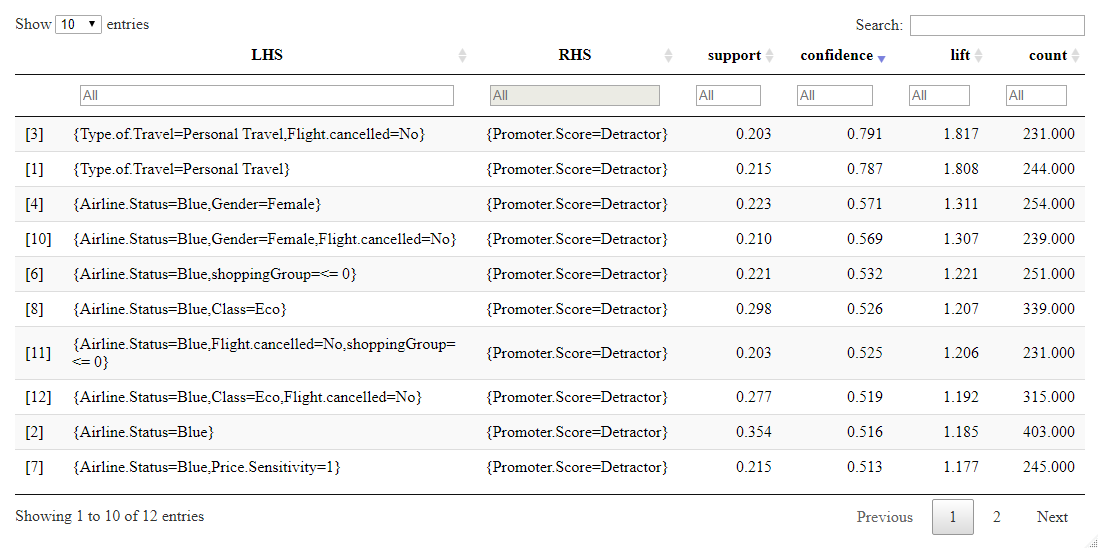
Table : FlyFast Passive Rules

Table : FlyFast Detractor Rules

The results for FlyFast generally mirror the top six airlines with business type travel being a strong predictor of being a promoter or passive and personal travel predicting being a detractor. Additionally, being age 33-45 predicts being a promoter and being female and or blue status indicate being a detractor.

### Personal Travelers

Guided by previous analysis, attributes that predict passive and detractor status among the personal travel population were determined in order to identify potential improvement areas. The minimum confidence values were determined from the average promoter score by travel type in Table 10: Promoter Score Percentage by Travel Type. The rulesets are tableted in Table 11: Personal Travel Passive Rules and Table 12: Personal Travel Detractor Rules.

Table : Promoter Score Percentage by Travel Type

Detractor Passive Promoter

Business travel 19.25 35.62 45.14

Mileage tickets 26.18 37.15 36.67

Personal Travel 68.43 25.40 6.17

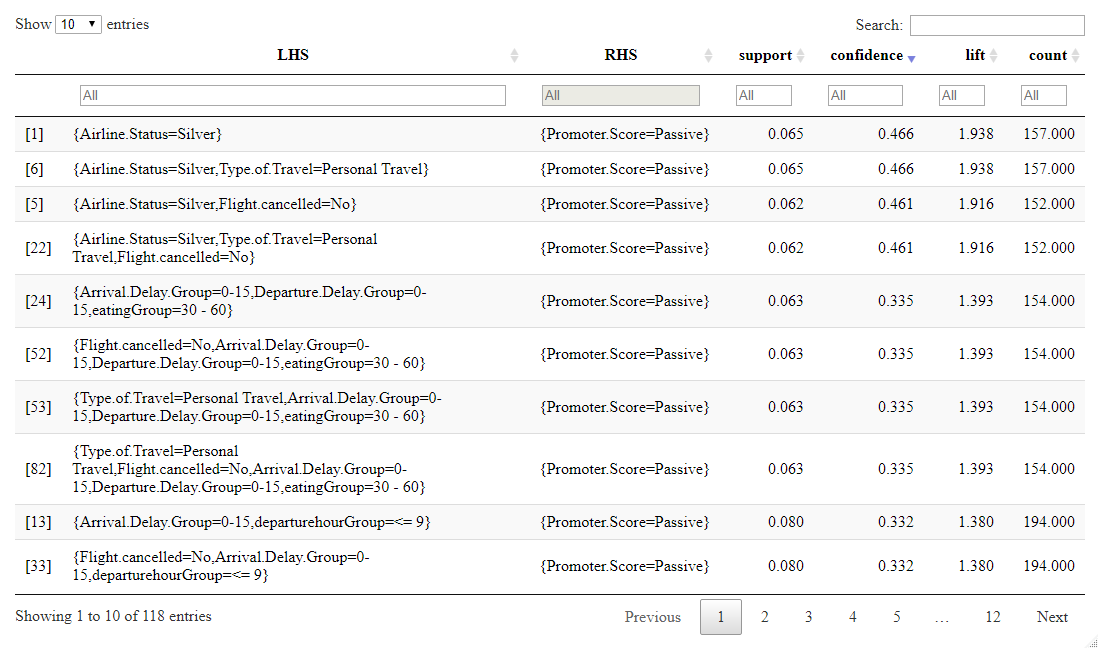
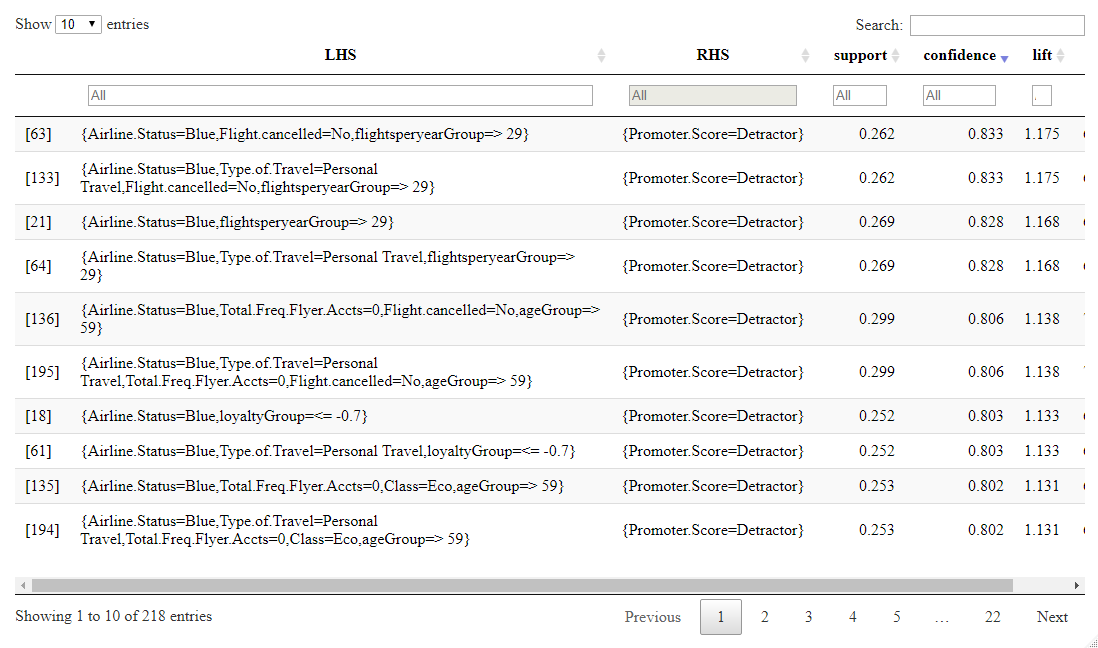
Table : Personal Travel Passive Rules

Table : Personal Travel Detractor Rules

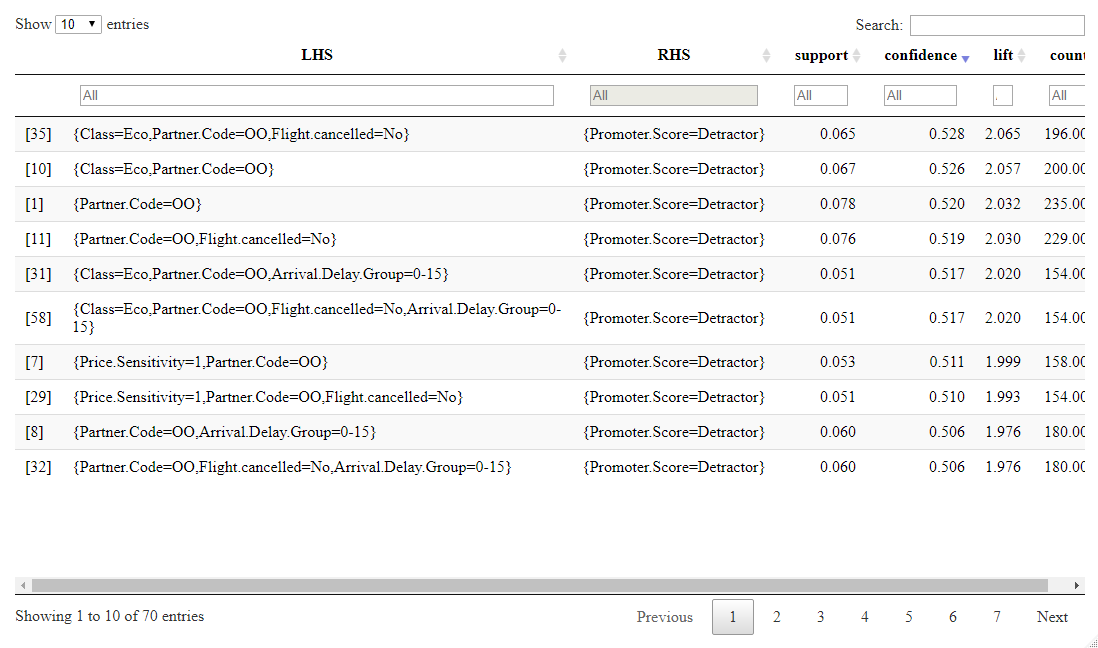


Comparing the results of the rulesets, passenger status seems to have significant impact on determining a passive or detractor rating. Being greater than or equal to 59 years of age also predicts personal travelers being detractors.

### Business Travelers

Attributes that predict detractor status among the business travel population were analyzed in order to identify weaknesses in a strong demographic. The minimum confidence values were determined from the average promoter score by travel type in Table 10: Promoter Score Percentage by Travel Type. The rulesets are tableted in Table 13: Business Travel Detractor Rules.

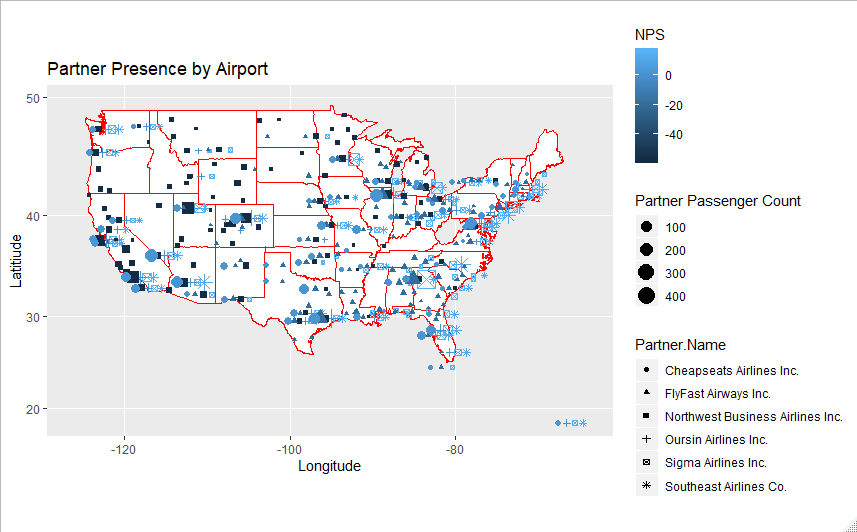
Table : Business Travel Detractor Rules



Northwest Business Airlines (partner code OO) has a strong relationship with business travelers being detractors. This indicates Northwest Business Airlines is an underperforming airline, failing to satisfy customers in a strong demographic.

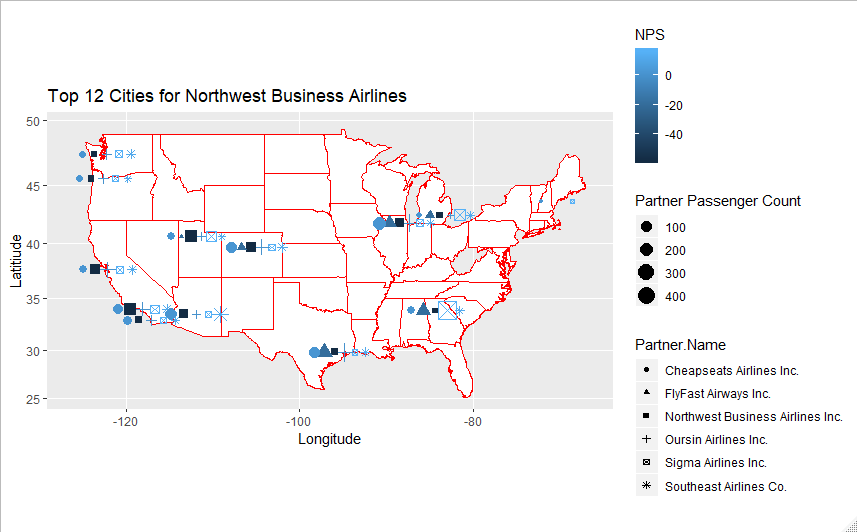
We plotted the top 6 partners across all cities in continent US to compare Northwest with other partners

Figure Map for Partner airline presence by airport



Here are the top 12 cities for Northwest business airlines that have co-located partners.

Figure Map for top 12 cities for Northwest Business Airline



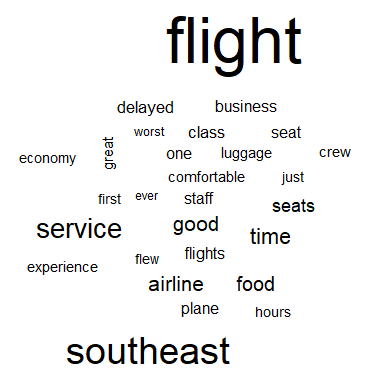
## Text mining and Sentiment Analysis

Text Mining and sentiment analysis is also an important model. From the dataframe, we can find a column named “freeText” store the comment of passengers. Actually, those who have really strong positive or negative emotion of the flight are more likely to comment on their flight and their feeling. So even though we only have around 200 comments among all 10282 records, these comments are representative enough to show the emotion of passengers.

**1.Text mining and sentiment analysis to the whole comment**

First, we want to find the frequency of appearance of all the words in all comments, no matter emotional or not. We filter the records that do have comments and form a new dataframe called “comment”. Then, we do some data cleaning on it: change uppercase letter to lowercase, remove punctuation, remove numbers, and remove words that appear a lot but meaningless. Then we form a Term Document Matrix and use this count the number of appearance of each term and sort from the most popular to least popular. Then we create a dataframe to store these information and create a word cloud to show these term (only focus on those who appear more than 30 times).

Figure Word cloud for all words



From this word cloud, we can find that there are too much description about flight noun which is useless. So we start to use sentiment analysis on all the comments. We scan the positive txt and negative txt, clean the useless headers, then use match() function to find the matched positive/negative word and their frequency of appearance. We create a dataframe and form a wordcloud to visualize it. We also count the positive ratio (0.08772164) and negative ratio (0.06759099). This shows that the overall comments are more positive.

Figure Word cloud for positive words from all comments



Figure Word cloud for negative words from all comments



From the positive wordcloud, we can find out that the comment are focused on several overall parts. They feel good about the friendly and polite service from attendant, comfortable atmosphere and so on. From the negative wordcloud, we can find out that they complain about delays, uncomfortable atmosphere and rude service from attendants.

In the two steps above, we only focus on the “freeText” column itself. Then we want to focus more on other columns to find out where these positive word and negative word come from. This is helpful for us to know better about our passengers and make some improvement. It is also helpful for us to select a subset based on the criteria we find further.

We created a column that separate the passengers into three categories: Promoter, Passive and Detractor. After we create the histogram for records that contain positive words, we found that not all of them are promoter. Similarly, not all negative words come from detractors.

Figure Histogram for promoter scores with positive words in comments

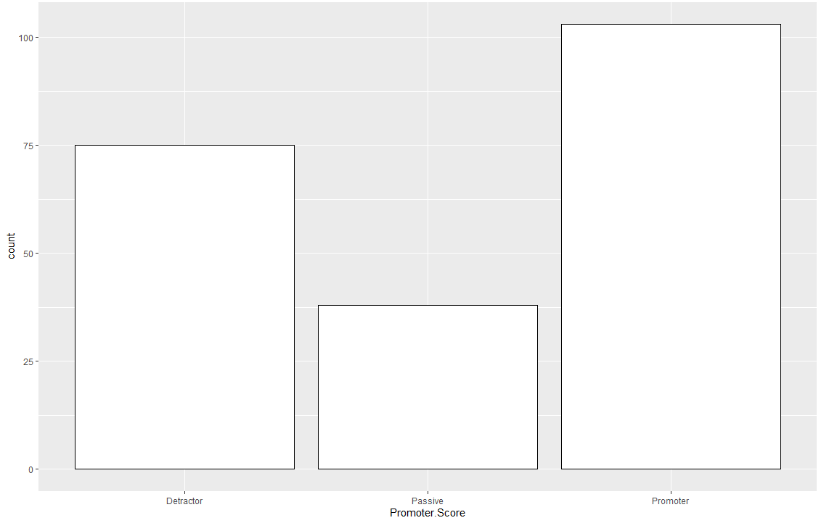
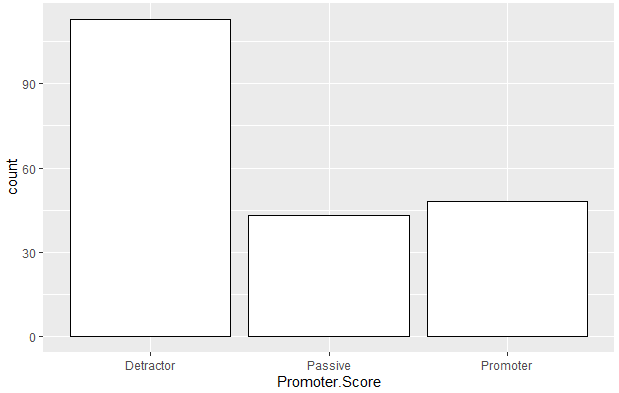


Figure Histogram for promoter score with negative words in comments



Actually, this make sense, because the levels of emotion inside every positive words are different. Some passangers may give positive words, but their emotion may not be strong enough to make them a promoter – they may only be passive. The message given by promoter and detractor are actually more useful. So to positive part, we want to focus only on the promoter; to negative part, we want to focus only on the detractor. We run the sentiment analysis again after filter the data.

Figure Word cloud for promoters with positive comments



Figure Word cloud for promoters with negative comments



Since several codes of sentiment analysis are used repeatedly, we create one formula named “positive\_function” for positive part and one formula named “negative\_function” for negative part. Then we can reuse this formula in the next part without type the same codes repeatedly.

**2. sentiment analysis to top 6 companies (business travel/personal travel)**

Since we have place a 750 passenger threshold and filter top 6 companies that contain 77% of all the passengers, we start to filter the dataframe to only focus on these top 6 companies and do sentiment analysis on them. From the associated rules mining, we find that business travel & personal travel have a strong relationship with likelihood to recommand. So we focus on the top 6 companies, and analyze positive word from promoter and negative from detractor to both business traveler and personal traveler.

**(1)Sentiment analysis for business traveler**

Figure Top 6 business travel – positive word



Figure Top 6 business travel – Promoter



Figure Top 6 business travel – negative word



Figure Top 6 business travel – detractor



From positive aspect, service(friendly, decent, polite) have a high appearance. From negative aspect, apart from delays, business travelers also complain the flight is so cold and slow. Some complain about rude and hostile also appear.

**(2)Sentiment analysis for personal traveler**

Figure Top 6 personal travel – positive word



Figure Top 6 personal travel – Promoter



Figure Top 6 personal travel – negative word



Figure Top 6 personal travel – detractor



From positive aspect, personal traveler think these companies provide a friendly and polite service, and they also think the flight is more quiet. From negative aspect, personal traveler face lots of delays. They also face some emergency issue that make them mention death.

**3. sentiment analysis to top 6 companies (male/female)**

We want to find out if there are some difference between female’s review and male’s review.

**(1)Sentiment analysis for male**

Figure Top 6 male – positive word



Figure Top 6 male – Promoter



Figure Top 6 male – negative word

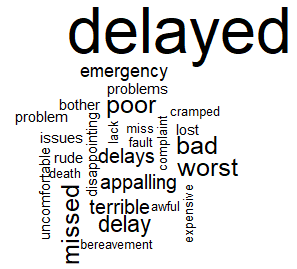


Figure Top 6 male – detractor



**(2)Sentiment analysis for female**

Figure Top 6 female – positive word



Figure Top 6 female – Promoter

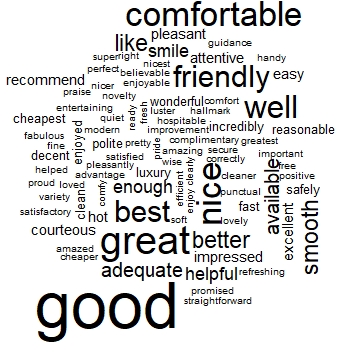


Figure Top 6 female – negative word

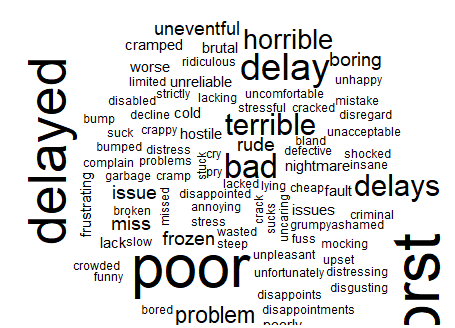


Figure Top 6 female – detractor



By comparing positive feedback between male and female, we find that male focus more on comfortable, and the attitude of attendants(friendly,decent), but actually this also works for female. From negative aspect, males focus more on things like delays that is more actual, while female care more about the feeling(worst). Overall, there are no much difference between male and female.

**4. sentiment analysis to Cheapseats**

Next we focus on two representative companies. First, we focus on Cheapseats Airline. There are 39 comments in total for Cheapseats Airline, 25 comments contain positive word, 14 from promoters with positive word; 30 comments contain negative word, 18 from detractors with negative word. Here we will place the “positive word” and “positive word from promoter” on one row. The graph from the right is more powerful.

Figure Positive word for Cheapseats



Figure Positive word for cheapseat from promoters



Figure Negative word for Cheapseats

****

Figure Negative word for cheapseat from detractor

****

We can look closely to the result. For the positive part, the atmosphere provided by cheapseat is very comfortable and clean, and customers like the efficient and friendly of attendants. For the negative part, we can find that delay is a big issue to cheapseat. The customer also face some emergency and terrible accident. So we need to improve the safety of cheapseat and control delays (Actually, this can be show from the linear regression that delay actually does not have much effect on likelihood to recommend.)

**5. sentiment analysis to Northwest**

We also focus on Northwest Business Airline. There are 22 comments in total for Northwest, 14 comments among them contain positive words, 9 records have positive words from promoters. 17 comments contain negative words, and 9 comments have negative words from detractors.

Figure Positive word for Northwest

****

Figure Positive word for Northwest from promoters

****

From positive aspect, we can find that friendly, polite and courteous appear more. This shows that the service provided by Northwest is very great!

Figure Negative word for Northwest

****

Figure Negative word for Northwest from detractor

****

From negative aspect, we can find that delays actually does not appear that much. But “Boring” do have a big font. We can increase some entertainment service for Northwest.

## Linear Modeling

For this part, we try to use linear models to figure out some hidden relationships between different variables and quantify these relationships, in addition, to prove the results that we got from the association rules mining.

1. Principal component analysis

At first, I try to use principal component analysis to deduce the number of variables. However, most of our variables are categorical variables like destination city, origin city, airline status and so on, and we also have some variables that have no significance to be calculated. So I chose these variables in principal component analysis: Age, Year.of.First.Flight, Flights.Per.Year, Shopping.Amount.at.Airport, Eating.and.Drinking.at.Airport, Day.of.Month, Scheduled.Departure.Hour and Fight.Distance. By calculating the correlation matrix, we can get such the results:



And then I did principal component analysis with these variables, getting the results like this:

Principal Components Analysis

Call: principal(r = cor\_matrix, nfactors = 5)

Standardized loadings (pattern matrix) based upon correlation matrix

RC1 RC2 RC3 RC5 RC4 h2 u2 com

Age 0.79 0.27 -0.01 0.04 -0.02 0.70 0.30 1.2

Year.of.First.Flight 0.00 -0.03 0.01 0.85 0.13 0.75 0.25 1.0

Flights.Per.Year 0.78 -0.26 0.00 -0.03 0.02 0.68 0.32 1.2

Shopping.Amount.at.Airport -0.14 0.49 -0.07 -0.28 0.31 0.44 0.56 2.6

Eating.and.Drinking.at.Airport 0.07 0.83 0.05 0.08 -0.11 0.72 0.28 1.1

Day.of.Month -0.05 0.09 -0.68 0.34 -0.22 0.64 0.36 1.8

Scheduled.Departure.Hour 0.02 -0.01 0.00 0.13 0.90 0.82 0.18 1.0

Flight.Distance -0.05 0.08 0.75 0.26 -0.18 0.67 0.33 1.4

RC1 RC2 RC3 RC5 RC4

SS loadings 1.27 1.09 1.03 1.02 1.01

Proportion Var 0.16 0.14 0.13 0.13 0.13

Cumulative Var 0.16 0.30 0.42 0.55 0.68

Proportion Explained 0.23 0.20 0.19 0.19 0.19

Cumulative Proportion 0.23 0.44 0.63 0.81 1.00

Mean item complexity = 1.4

Test of the hypothesis that 5 components are sufficient.

The root mean square of the residuals (RMSR) is 0.16

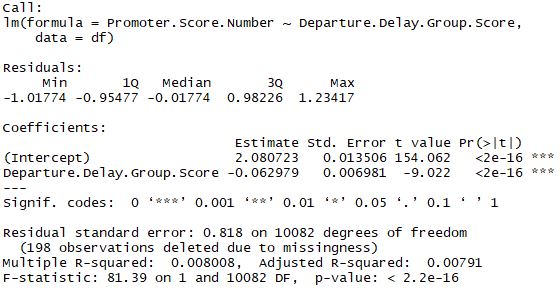
Fit based upon off diagonal values = -6.02

We can see from the result that the first component (RC1) is highly correlated with customers’ ages and their years of first flights. We can understand why the computer put ages and years of first flights together, because the older people are, the earlier their first flighted. However, this component is not useful for us to predict the likelihood to recommend. Therefore, although we decrease the number of variables, it’s not helpful for our linear models and find the relationships between likelihood and other variables. So we drop this method off.

1. Linear models on delays

In this module, we want to figure out the solution to improve the company, by then, I created several linear models to get the relationship between delays and the promoter scores.

In specific, according to the range of the variable “Departure.Delay.in.Minutes”, we can separate the minutes into 5 groups. First is less than 15 minutes. Second group is between 15 minutes to 30 minutes. Third one is between 30 minutes to 45 minutes. Forth one is between 45 minutes to 60 minutes. And the last one is more than 60 minutes. After grouping “Departure.Delay.in.Minutes”, we also group the variable “Likelihood.to.recommend”. At first, we divided it into “detractor”, “passive” and “promoter”. Considering the independent variables in a linear model are numeric variables, we change the categories into different groups. Then we build the linear model between the two new variables. And we get the results:



From the results, we can see the p-values are less than 0.05, which proves that the delay statistical significantly effects the promoter scores, in other words, the delays and promoter scores indeed have relationship with each other.

Therefore, from the company perspective, we suggest that all partners need to avoid delaying for all airlines.

1. Linear models on the top 6 partners

From the rules mining, by sorting the count of promoters of each companies, we can find that the top 6 companies are Cheapseats Airlines Inc.(WN), FlyFast Airways Inc.(EV), Northwest Business Airlines Inc.(OO), Oursin Airlines Inc.(OU), Sigma Airlines Inc.(DL) and Southeast Airlines Co(US).

To figure out other factors which need to improve for the 6 companies, I select several variables which might be relative to improvement of companies: “Departure.Delay.Group.Score”, “Price.Sensitivity.Score” and “Flight.cancelled.Dummy”. I will describe the meanings of these variables. “Departure.Delay.Group.Score” is the variable that we used in the previous step. “Price.Sensitivity.Score” is the transformation of the variable “Price.Sensitivity” from 1 to 5. And the variable “Flight.cancelled.Dummy” is a dummy variable whether the flight is cancelled. The reason why I selected these 3 variables is that among all the variables, there are only three variables related with companies. In other words, we only know companies from these 3 aspects.

By then, I made 6 linear models to see if all considering variables indeed need to improve by checking p-value or giving the order of improvement by significance.

Cheapseats Airlines Inc.(WN)

Call:

lm(formula = Promoter.Score.Number ~ Departure.Delay.Group.Score +

Price.Sensitivity.Score + Flight.cancelled.Dummy, data = df4)

Residuals:

Min 1Q Median 3Q Max

-1.30126 -0.92617 -0.05621 0.87876 1.31887

Coefficients: (1 not defined because of singularities)

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

(Intercept) 2.54630 0.07743 32.883 < 2e-16 \*\*\*

Departure.Delay.Group.Score -0.06502 0.01378 -4.719 2.52e-06 \*\*\*

Price.Sensitivity.Score -0.18002 0.03156 -5.704 1.33e-08 \*\*\*

Flight.cancelled.Dummy NA NA NA NA

---

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

Residual standard error: 0.8114 on 2196 degrees of freedom

(30 observations deleted due to missingness)

Multiple R-squared: 0.02495, Adjusted R-squared: 0.02406

F-statistic: 28.1 on 2 and 2196 DF, p-value: 8.949e-13

FlyFast Airways Inc.(EV)

Call:

lm(formula = Promoter.Score.Number ~ Departure.Delay.Group.Score +

Price.Sensitivity.Score + Flight.cancelled.Dummy, data = df9)

Residuals:

Min 1Q Median 3Q Max

-0.9450 -0.8042 0.1254 0.3990 1.4694

Coefficients: (1 not defined because of singularities)

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

(Intercept) 2.06622 0.10859 19.028 < 2e-16 \*\*\*

Departure.Delay.Group.Score -0.05080 0.01831 -2.774 0.00563 \*\*

Price.Sensitivity.Score -0.07041 0.04368 -1.612 0.10723

Flight.cancelled.Dummy NA NA NA NA

---

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

Residual standard error: 0.8011 on 1087 degrees of freedom

(47 observations deleted due to missingness)

Multiple R-squared: 0.009223, Adjusted R-squared: 0.0074

F-statistic: 5.06 on 2 and 1087 DF, p-value: 0.006499

Northwest Business Airlines Inc.(OO)

Call:

lm(formula = Promoter.Score.Number ~ Departure.Delay.Group.Score +

Price.Sensitivity.Score + Flight.cancelled.Dummy, data = df3)

Residuals:

Min 1Q Median 3Q Max

-0.5685 -0.4621 -0.3556 0.5379 1.6444

Coefficients: (1 not defined because of singularities)

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

(Intercept) 1.70334 0.06668 25.547 < 2e-16 \*\*\*

Departure.Delay.Group.Score -0.02840 0.01354 -2.097 0.036208 \*

Price.Sensitivity.Score -0.10644 0.02746 -3.877 0.000112 \*\*\*

Flight.cancelled.Dummy NA NA NA NA

---

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

Residual standard error: 0.505 on 1220 degrees of freedom

(25 observations deleted due to missingness)

Multiple R-squared: 0.01582, Adjusted R-squared: 0.0142

F-statistic: 9.803 on 2 and 1220 DF, p-value: 5.979e-05

Oursin Airlines Inc.(OU)

Call:

lm(formula = Promoter.Score.Number ~ Departure.Delay.Group.Score +

Price.Sensitivity.Score + Flight.cancelled.Dummy, data = df2)

Residuals:

Min 1Q Median 3Q Max

-1.17565 -0.94600 -0.06958 0.93042 1.31869

Coefficients: (1 not defined because of singularities)

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

(Intercept) 2.35226 0.11584 20.306 < 2e-16 \*\*\*

Departure.Delay.Group.Score -0.07055 0.02255 -3.129 0.00181 \*\*

Price.Sensitivity.Score -0.10606 0.04726 -2.244 0.02504 \*

Flight.cancelled.Dummy NA NA NA NA

---

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

Residual standard error: 0.8146 on 948 degrees of freedom

(10 observations deleted due to missingness)

Multiple R-squared: 0.01551, Adjusted R-squared: 0.01343

F-statistic: 7.465 on 2 and 948 DF, p-value: 0.000607

Sigma Airlines Inc.(DL)

Call:

lm(formula = Promoter.Score.Number ~ Departure.Delay.Group.Score +

Price.Sensitivity.Score + Flight.cancelled.Dummy, data = df6)

Residuals:

Min 1Q Median 3Q Max

-1.2600 -0.9223 -0.1603 0.7898 1.1275

Coefficients: (1 not defined because of singularities)

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

(Intercept) 2.38186 0.09427 25.268 < 2e-16 \*\*\*

Departure.Delay.Group.Score -0.07196 0.01944 -3.701 0.000222 \*\*\*

Price.Sensitivity.Score -0.04985 0.03782 -1.318 0.187666

Flight.cancelled.Dummy NA NA NA NA

---

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

Residual standard error: 0.8165 on 1559 degrees of freedom

(22 observations deleted due to missingness)

Multiple R-squared: 0.009468, Adjusted R-squared: 0.008197

F-statistic: 7.451 on 2 and 1559 DF, p-value: 0.000602

Southeast Airlines Co(US)

Call:

lm(formula = Promoter.Score.Number ~ Departure.Delay.Group.Score +

Price.Sensitivity.Score + Flight.cancelled.Dummy, data = df1)

Residuals:

Min 1Q Median 3Q Max

-1.1662 -0.9807 -0.1381 0.8478 1.0905

Coefficients: (1 not defined because of singularities)

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

(Intercept) 2.23743 0.13099 17.081 <2e-16 \*\*\*

Departure.Delay.Group.Score -0.05715 0.02722 -2.100 0.0361 \*

Price.Sensitivity.Score -0.01406 0.05319 -0.264 0.7916

Flight.cancelled.Dummy NA NA NA NA

---

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

Residual standard error: 0.8083 on 834 degrees of freedom

(14 observations deleted due to missingness)

Multiple R-squared: 0.005341, Adjusted R-squared: 0.002955

F-statistic: 2.239 on 2 and 834 DF, p-value: 0.1072

Because the dependent variable “Promoter.Score.Number” is a discrete variable and one of the independent variable “Flight.cancelled.Dummy” is a binary variable, we cannot make a perfect linear model for this situation and we get NAs for “Flight.cancelled.Dummy” in the results.

Besides that, we can see the results of regression.

For Cheapseats Airlines Inc.(WN), we can see the p-values of two coefficients are all statistically significant, which means both delay and price effect the scores. Two coefficient are respectively -0.06502 and -0.18002, which shows that the price will effect more on the scores.

For FlyFast Airways Inc.(EV), we can see only the p-values of “Departure.Delay.Group.Score” coefficient is statistically significant, the p-value of “Price.Sensitivity.Score” is more than 0.05, which means delay effects the scores, the customers on this airline don’t care about price of tickets and we can appropriately increase the price of tickets.

For Northwest Business Airlines Inc.(OO), we can see the p-values of two coefficients are all statistically significant, which means both delay and price effect the scores. Two coefficients are respectively -0.02840 and -0.10644, which shows the price will effect more on the final score.

For Oursin Airlines Inc.(OU), we can see the p-values of wo coefficients are all statistically significant, which means both delay and price effect the scores.

For Sigma Airlines Inc.(DL), we can see only the p-values of “Departure.Delay.Group.Score” coefficient is statistically significant, the p-value of “Price.Sensitivity.Score” is more than 0.05, which means delay effects the scores, the customers on this airline don’t care about price of tickets and we can appropriately increase the price of tickets.

For Southeast Airlines Co(US), we can see only the p-values of “Departure.Delay.Group.Score” coefficient is statistically significant, the p-value of “Price.Sensitivity.Score” is more than 0.05, which means delay effects the scores, the customers on this airline don’t care about price of tickets and we can appropriately increase the price of tickets.

In a word, we conclude that for Cheapseats Airlines Inc.(WN), Northwest Business Airlines Inc.(OO) and Oursin Airlines Inc.(OU), both price and delay are important for customers’ scores, what’s more, price of tickets has more effect than delay on scores, so we suggest these three companies appropriately decrease the price of tickets and try to eliminate the delay.

While for FlyFast Airways Inc.(EV), Sigma Airlines Inc.(DL) and Southeast Airlines Co(US), only delay effects the customers’ scores, and for customers who travel by these airlines, they think the price of tickets doesn’t matter. So we suggest that these companies should try to eliminate delays and appropriately increase the price of ticket in order to get more revenue.

1. Linear models on consumption at airports

In this part, we try to find whether there is impact on NPS when customers shopped, ate or drank at airports when there was a delay or no delay for top 6 companies.

First, we detected shopping amount at airport and flight delays for the top 6 companies, then got the result like this:

Call:

lm(formula = Shopping.Amount.at.Airport ~ Departure.Delay.Dummy,

data = df11)

Residuals:

Min 1Q Median 3Q Max

-30.03 -30.03 -27.80 2.59 404.97

Coefficients:

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

(Intercept) 30.032 2.336 12.857 <2e-16 \*\*\*

Departure.Delay.Dummy -2.237 3.206 -0.698 0.485

---

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

Residual standard error: 56.59 on 1249 degrees of freedom

(22 observations deleted due to missingness)

Multiple R-squared: 0.0003897, Adjusted R-squared: -0.0004106

F-statistic: 0.4869 on 1 and 1249 DF, p-value: 0.4854

And then we detected the relationship between eating and drinking at airport and delay for the top 6 companies.

Call:

lm(formula = Eating.and.Drinking.at.Airport ~ Departure.Delay.Dummy,

data = df11)

Residuals:

Min 1Q Median 3Q Max

-67.96 -37.63 -7.96 22.37 650.04

Coefficients:

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

(Intercept) 67.9642 2.1974 30.93 <2e-16 \*\*\*

Departure.Delay.Dummy -0.3332 3.0161 -0.11 0.912

---

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

Residual standard error: 53.24 on 1249 degrees of freedom

(22 observations deleted due to missingness)

Multiple R-squared: 9.771e-06, Adjusted R-squared: -0.0007909

F-statistic: 0.0122 on 1 and 1249 DF, p-value: 0.9121

We can see that delay has no effect on shopping amount or eating and drinking amount at airport. We have insufficient evidence to prove whether the airline is delayed or not effects the consumption of customers in airports.

Then I try to find the relationship between promoter scores and the consumption in airport and got the result like this:

Call:

lm(formula = Promoter.Score.Number ~ Shopping.Amount.at.Airport +

Eating.and.Drinking.at.Airport, data = df11)

Residuals:

Min 1Q Median 3Q Max

-1.46231 -0.89547 0.07564 1.03480 1.13386

Coefficients:

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

(Intercept) 1.871e+00 3.841e-02 48.709 <2e-16 \*\*\*

Shopping.Amount.at.Airport -7.612e-05 4.048e-04 -0.188 0.8509

Eating.and.Drinking.at.Airport 8.255e-04 4.278e-04 1.930 0.0538 .

---

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

Residual standard error: 0.8124 on 1270 degrees of freedom

Multiple R-squared: 0.002936, Adjusted R-squared: 0.001366

F-statistic: 1.87 on 2 and 1270 DF, p-value: 0.1545

The p-values are still greater than 0.05, which means that the consumption at airport doesn’t effect customers’ score.

1. Linear models on business travel and personal travel

From the result of association rules mining, we found that customers for business travel tend to give higher scores while customers for personal travel tend to give lower scores. So in this part, we try to figure out the relationship between business travel and promoter scores and the relationship between personal travel and promoter scores.

At first, I try to find if the business or personal travel significantly effects promoter score.

Call:

lm(formula = Promoter.Score.Number ~ Business.Dummy, data = df22)

Residuals:

Min 1Q Median 3Q Max

-1.2057 -0.4828 -0.2057 0.7943 1.5172

Coefficients:

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

(Intercept) 1.48283 0.03294 45.01 <2e-16 \*\*\*

Business.Dummy 0.72283 0.04214 17.15 <2e-16 \*\*\*

---

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

Residual standard error: 0.7329 on 1271 degrees of freedom

Multiple R-squared: 0.188, Adjusted R-squared: 0.1874

F-statistic: 294.3 on 1 and 1271 DF, p-value: < 2.2e-16

Call:

lm(formula = Promoter.Score.Number ~ Personal.Dummy, data = df22)

Residuals:

Min 1Q Median 3Q Max

-1.1955 -0.3067 -0.1955 0.8045 1.6933

Coefficients:

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

(Intercept) 2.19548 0.02362 92.94 <2e-16 \*\*\*

Personal.Dummy -0.88878 0.04279 -20.77 <2e-16 \*\*\*

---

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

Residual standard error: 0.7027 on 1271 degrees of freedom

Multiple R-squared: 0.2534, Adjusted R-squared: 0.2528

F-statistic: 431.5 on 1 and 1271 DF, p-value: < 2.2e-16

From the results, we can see that the type of travel indeed effects the prompter score. I built another linear model to prove the personal travel have more effects than business travel. The coefficient of Business.Dummy is >0, while the coefficient of Personal.Dummy is <0, so we can conclude that for business travel, people tend to give a high score and for personal travel, people tend to give a lower score.

> model23 <- lm(Promoter.Score.Number~Business.Dummy+Personal.Dummy,data=df22)

> summary(model23)

Call:

lm(formula = Promoter.Score.Number ~ Business.Dummy + Personal.Dummy,

data = df22)

Residuals:

Min 1Q Median 3Q Max

-1.2057 -0.3067 -0.2057 0.7943 1.6933

Coefficients:

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

(Intercept) 2.12150 0.06793 31.232 <2e-16 \*\*\*

Business.Dummy 0.08416 0.07245 1.162 0.246

Personal.Dummy -0.81479 0.07672 -10.620 <2e-16 \*\*\*

---

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

Residual standard error: 0.7026 on 1270 degrees of freedom

Multiple R-squared: 0.2542, Adjusted R-squared: 0.2531

F-statistic: 216.5 on 2 and 1270 DF, p-value: < 2.2e-16

To figure out what leads this result, we put more variables in the linear model and use different subset.

First, subset1=top 6 and business travel:

> model25 <- lm(Promoter.Score.Number~Price.Sensitivity.Score+Flight.Distance+Departure.Delay.Group.Score,data=df23)

> summary(model25)

Call:

lm(formula = Promoter.Score.Number ~ Price.Sensitivity.Score +

Flight.Distance + Departure.Delay.Group.Score, data = df23)

Residuals:

Min 1Q Median 3Q Max

-1.5471 -0.3101 -0.1173 0.7737 1.0675

Coefficients:

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

(Intercept) 2.2136599 0.1275079 17.361 < 2e-16 \*\*\*

Price.Sensitivity.Score -0.0516585 0.0502188 -1.029 0.304

Flight.Distance 0.0002026 0.0000505 4.013 6.59e-05 \*\*\*

Departure.Delay.Group.Score -0.0283271 0.0226304 -1.252 0.211

---

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

Residual standard error: 0.7483 on 765 degrees of freedom

(9 observations deleted due to missingness)

Multiple R-squared: 0.02405, Adjusted R-squared: 0.02022

F-statistic: 6.283 on 3 and 765 DF, p-value: 0.0003256

We can see that for business travel, the distance of flight is statistically significant for the promoter scores. In other words, people don’t care about the price or delay, they care about the distance of flight. However, the coefficient of Flight.Distance is so small that I think it effects the score little.

In a word, I think among all the factors, the distance of flight indeed effects the promoter score but the effect is so little that the promoter scores almost unchanged.

Second, subset2=top 6 and personal travel:

> model26 <- lm(Promoter.Score.Number~Price.Sensitivity.Score+Flight.Distance+Departure.Delay.Group.Score,data=df24)

> summary(model26)

Call:

lm(formula = Promoter.Score.Number ~ Price.Sensitivity.Score +

Flight.Distance + Departure.Delay.Group.Score, data = df24)

Residuals:

Min 1Q Median 3Q Max

-0.4748 -0.3298 -0.2964 0.5817 1.7185

Coefficients:

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

(Intercept) 1.367e+00 1.284e-01 10.651 <2e-16 \*\*\*

Price.Sensitivity.Score -1.114e-02 4.665e-02 -0.239 0.8114

Flight.Distance 7.095e-05 4.666e-05 1.520 0.1292

Departure.Delay.Group.Score -6.088e-02 2.505e-02 -2.430 0.0156 \*

---

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

Residual standard error: 0.5514 on 377 degrees of freedom

(7 observations deleted due to missingness)

Multiple R-squared: 0.02085, Adjusted R-squared: 0.01306

F-statistic: 2.676 on 3 and 377 DF, p-value: 0.04695

We can see that there is no factor that is statistically significant except for flights delay. Similarly to business travel, we can not conclude if there is a factor effect a lot on people’s scores.

Compared with two models, we can find the intercept is greatly different. The intercept for business travel is 2.2136599, while the intercept for personal travel is 1.367. As we set before, the promoter score is between 1 to 3.

So why the intercepts for two types of travel are greatly different? From my point of view, I think the type of travel indeed a factor, and maybe there are other variables effecting the promoter score but we don’t know.

# Recommendations

1. Focus marketing & operations on attracting Business travelers

* From our preliminary exploration and association rules mining, business travelers were stronger indicators of being promoters

1. Improve in-flight entertainment for personal travel and re-structure the loyalty programs so that personal travelers take lesser time to reach higher status

* From our sentiment analysis for personal travelers in Northwest business airlines (with lowest NPS), we found out that the most common customer sentiment was “boring”. We predicted it to be an unsatisfactory flight experience either due to lack of entertainment or necessary amenities
* Also, the main difference in predictor for being a passive & detractor was the status. From our association rules mining, silver status was a predictor of being passive; Blue status was a predictor of being a detractor. Therefore, moving a personal traveler from blue to silver status would increase the likelihood of being a passive

1. Renegotiating partner contracts by airport for Northwest business airlines
   * We identified that Northwest was the partner with the lowest NPS score
   * We then identified the top cities it has the greatest presence in with co-located partners to check if there is an opportunity for shift
   * After analyzing the top 12 airports, we recommend the contract negotiation team to shift contracts from Northwest to other partners in those 12 cities which are San Francisco, Denver, Los Angeles, Houston, Seattle, Chicago, San Diego, Salt Lake city, Phoenix, Detroit, Portland & Atlanta
2. Reducing delays should not be a priority
   * Though it appears to be a major issue from the sentiment analysis, the linear model built doesn’t support it.
   * Even in association rules mining, delay didn’t show up to be a predictor of NPS category