Southeast Airlines Customer Satisfaction Analysis

IST687 M003 – Final Project



Group 4: Hitesh Thadani, Lakshya Gupta, Litin Behata, Yingxue Gao, Sankalp Singh

Instructor: Prof. Erik Anderson

Table of Contents

[1. Introduction 1](#_Toc26904739)

[1.1 Background 1](#_Toc26904740)

[1.2 Problem Statement 1](#_Toc26904741)

[1.3 Data Available 1](#_Toc26904742)

[2. Business Questions addressed 2](#_Toc26904743)

[3. Data Preparation 3](#_Toc26904744)

[3.1 Identified and removed NAs 3](#_Toc26904745)

[3.2 Created a new column 3](#_Toc26904746)

[3.3 Performed transformations on variables 3](#_Toc26904747)

[4. Exploratory Data Analysis 4](#_Toc26904748)

[4.1 Correlation Matrix 4](#_Toc26904749)

[4.2 Variable Analysis 4](#_Toc26904750)

[5. Predictive Modeling 15](#_Toc26904751)

[5.1 Linear Regression 15](#_Toc26904752)

[5.2 Random Forest 17](#_Toc26904753)

[5.3 Support Vector Machines 18](#_Toc26904754)

[5.4 Decision Trees 19](#_Toc26904755)

[5.5 Logistic Regression 21](#_Toc26904756)

[5.6 Association Rules Mining 21](#_Toc26904757)

[5.7 Text Mining 23](#_Toc26904758)

[6. Insights and Recommendations 25](#_Toc26904759)

[6.1 Insights based on variables 25](#_Toc26904760)

[6.2 Recommendations based on insights 25](#_Toc26904761)

[7. Appendix 27](#_Toc26904762)

# **1. Introduction**

## **1.1 Background**

The primary focus of this analysis is to predict the customers with low satisfaction for the Southeast Airlines and to know why the customers are not satisfied. The airlines want to lower their customer churn rate. We have a dataset consisting of feedbacks from around 10000 flyers and we have analyzed the dataset to find out actionable insights which will help Southeast Airlines to lower their customer attrition. Our analysis will help Southeast Airlines to know which customers are about to leave and get ahead of the loss.

## **1.2 Problem Statement**

The main aim of this project is to perform customer churn prediction for the Southeast Airlines. We have used the customer feedbacks from around 10000 customers to find out reasons that makes a customer happy or dissatisfied. We have calculated the Net Promoter Score that will help Southeast Airlines to know ahead of time if a passenger is dissatisfied and they can make amendments accordingly.

## **1.3 Data Available**

We have a customer level survey that basically entails the feedback given by the customer alongside variables that may help us determine the customer’s sentiment. We have customer feedbacks from 10282 customers. This feedback data contains the information regarding the passenger along with the flight details. The passengers have provided their satisfaction scores on a scale from 1 to 10.

# **2. Business Questions addressed**

* We have analyzed whether a particular partner airline is performing well among frequent flyers or not.
* We have tried to find out whether the type of travel affects the net promoter Score
* Find out the most significant variables that lead to customer attrition
* Whether a particular Origin/Destination State affects the customer churn rate.
* To analyze how Southeast Airlines can hold the satisfied customers and attract new flyers to increase the overall revenue.
* Find out combination of factors that lead to lower customer satisfaction.

# **3. Data Preparation**

## **3.1 Identified and removed NAs**

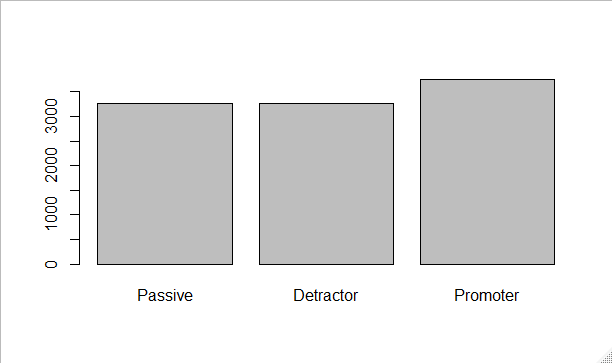
* **Columns with NA’s** – Likelihood.to.recommend, Departure.Delay.in.Minutes, Arrival.Delay.in.Minutes, freeText, Flight.time.in.minutes



* **Likelihood.to.recommend:** Only one NA was present in this column. So, we dropped that one row
* **Departure.Delay.in.Minutes, Arrival.Delay.in.Minutes, Flight.time.in.minutes**: Replaced by median of that column
* **freeText:** Around 98% of values were missing for this variable. So, we did not include this in predictive modeling. We performed sentiment analysis on the freeText column.

## **3.2 Created a new column**

* **“typeOfCustomer”** – with the target categories – Promoters, Passive, Detractors – based on Likelihood.to.recommend column



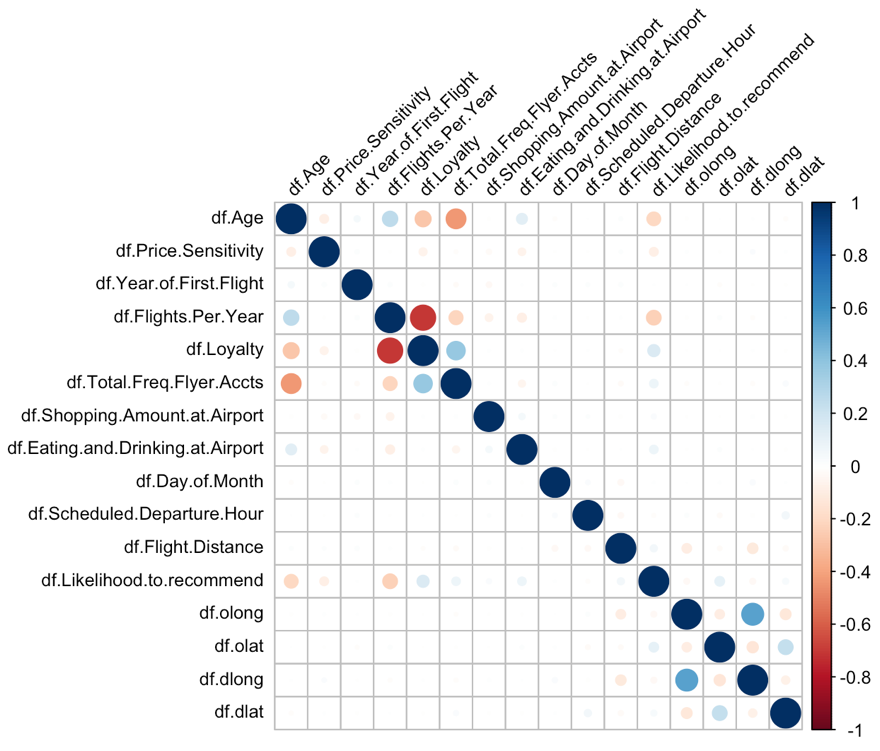
## **3.3 Performed transformations on variables**

Age into buckets, dummy variables- 1-hot encoding for required variables, changing data types accordingly wherever required.

# **4. Exploratory Data Analysis**

## **4.1 Correlation Matrix**

Variables containing numerical data are used to create the correlation matrix:



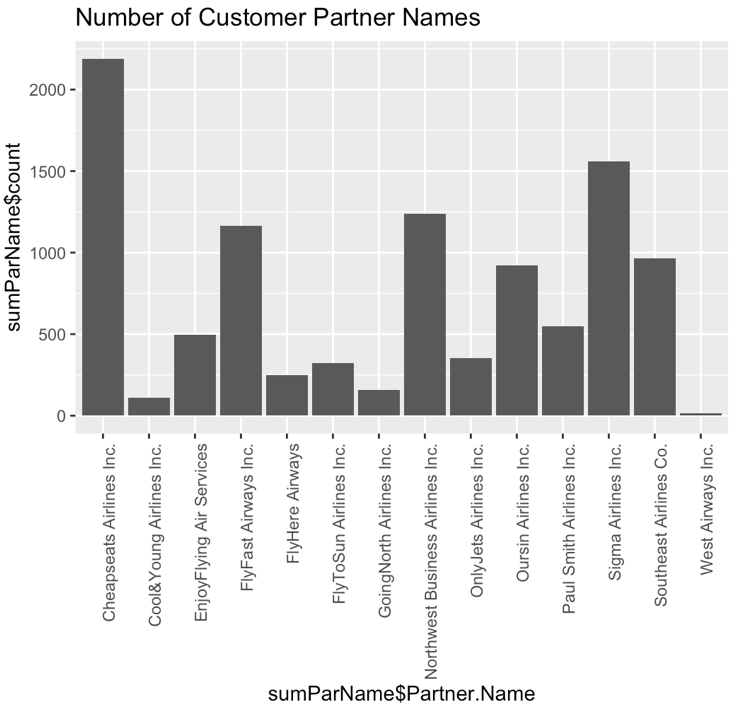
## **4.2 Variable Analysis**

The **Net Promoter Score** (NPS) is an index ranging from -100 to 100 that measures the willingness of customers to recommend a company's products or services to others.

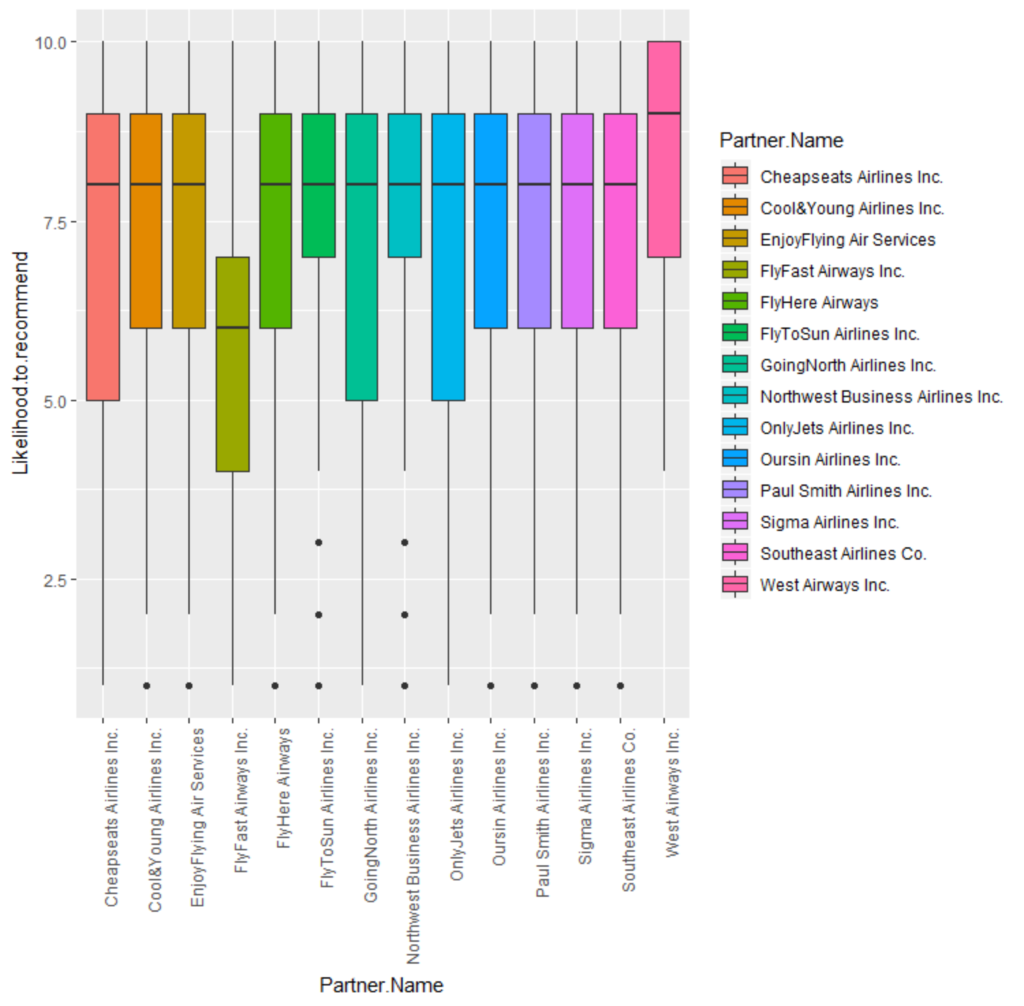
We calculated NPS of variables and plotted it for initial analysis:

**4.2.1 Partner Airline**

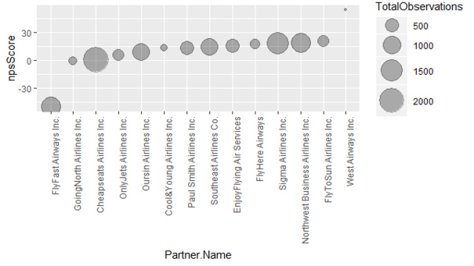
Bar plot of partner airlines displaying their number of observations:



Box plot for partner name vs. likelihood to recommend:



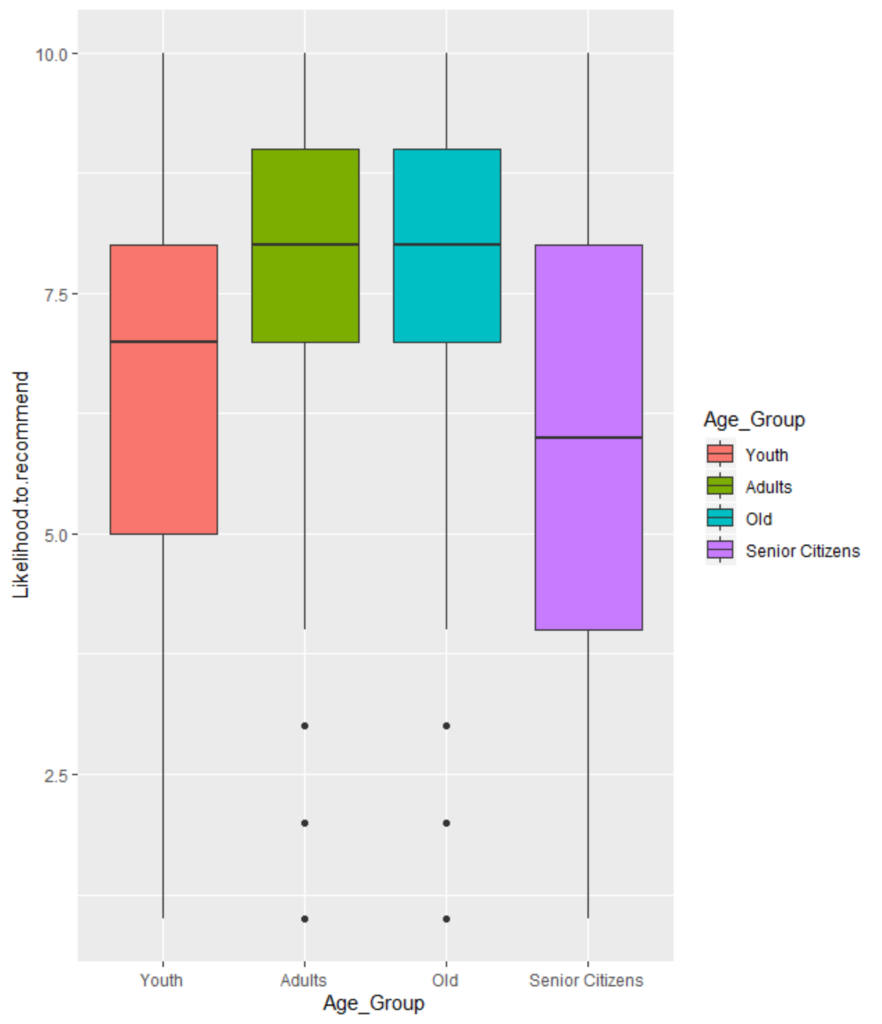
Bubble plot of partner airlines displaying their NPS and number of observations:



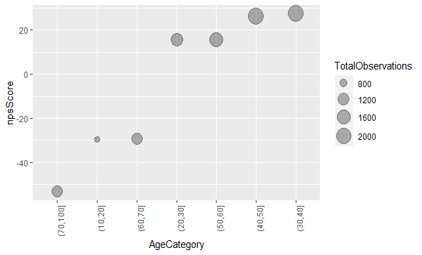
**4.2.2 Age**

We divided age into categories i.e. Binning.

Box plot for the Age group buckets from Age (10 to 20 – Youth, 20 to 40 – Adults, 40 to 60 – Old and 60 above – Senior Citizens):

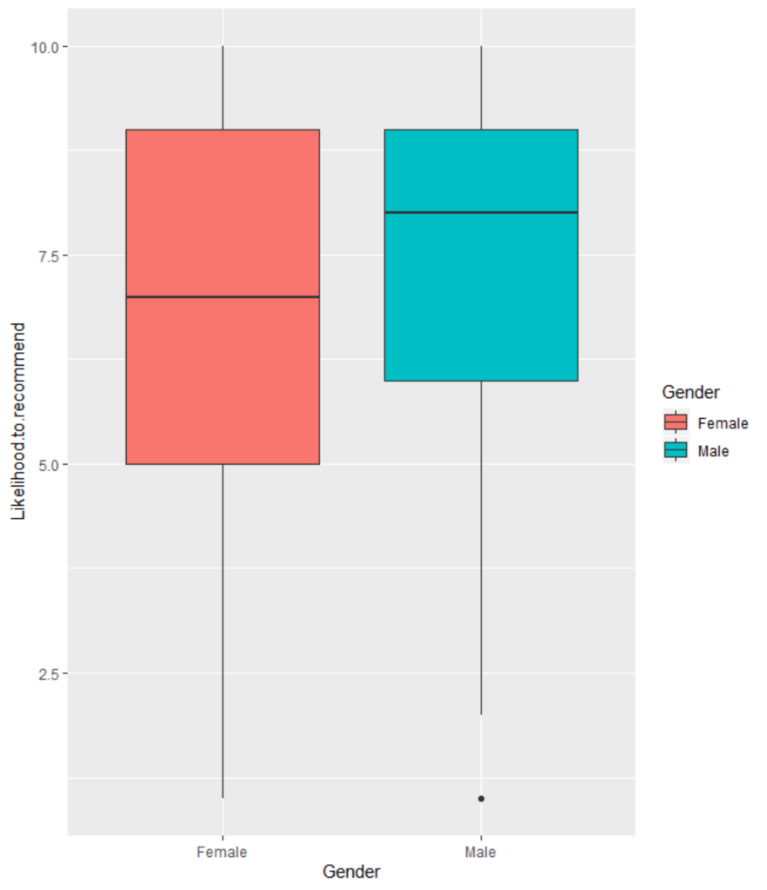


Bubble plot of its NPS and number of observations:

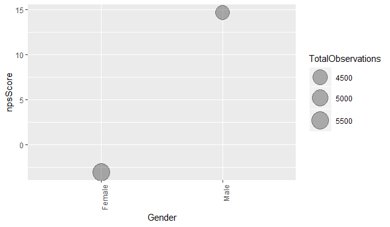


**4.2.3 Gender**

Boxplot for Gender Vs Likelihood to Recommend:

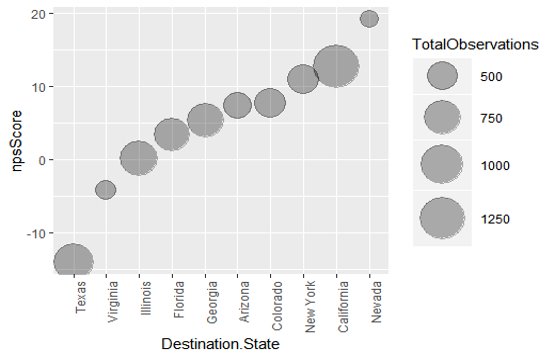


Bubble plot of gender displaying their NPS and number of observations:

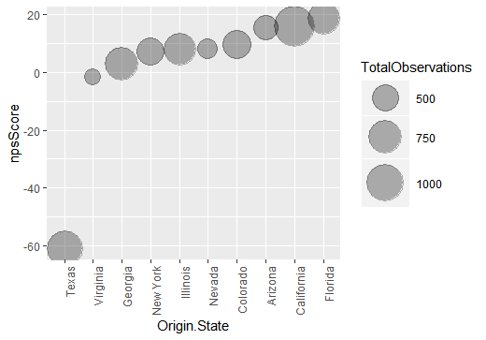


**4.2.4 Destination and Origin State**

Bubble plot of 10 Destination State with lowest NPS:



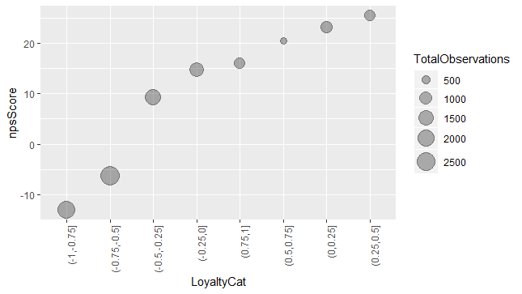
Bubble plot of 10 Origin State with lowest NPS:



**4.2.5 Loyalty**

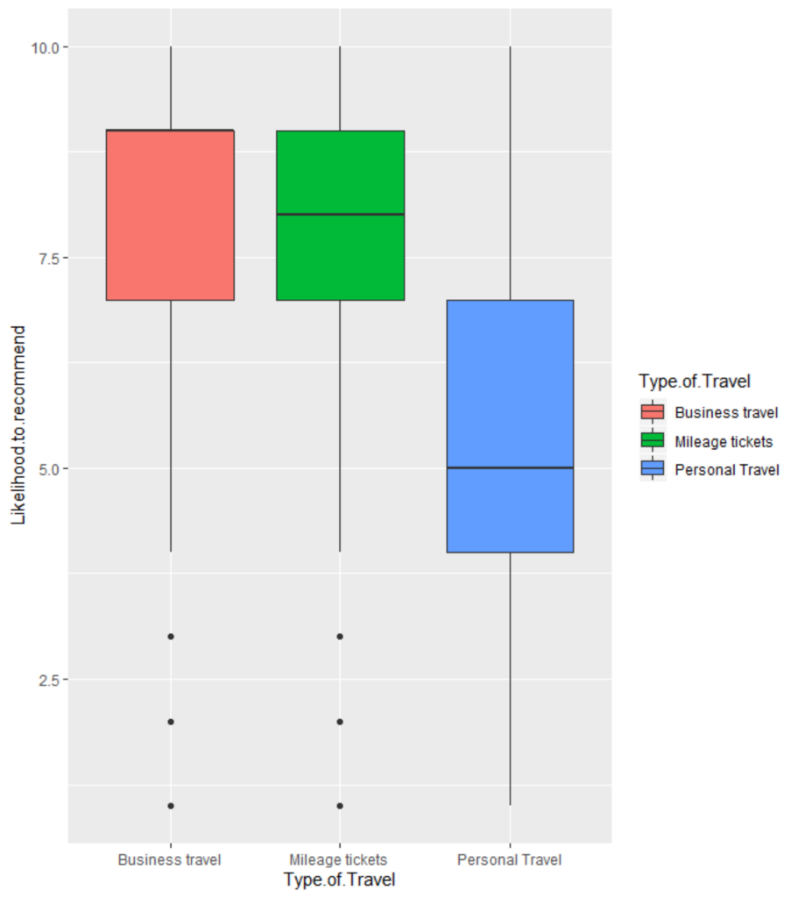
We divided loyalty score into categories i.e. binning.

Bubble plot of loyalty score categories displaying their NPS and number of observations:



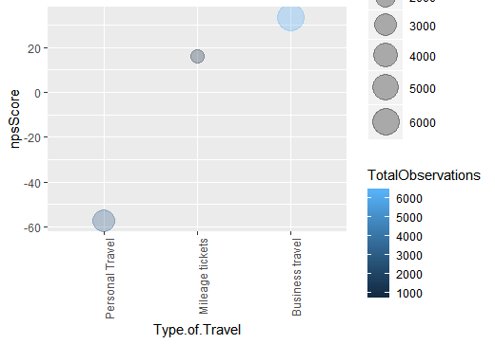
**4.2.6 Type of travel**

Box plot for type of travel:



Personal Travelers give the lowest ratings their 75th Percentile is below the 25th Percentile for the other type of travelers.

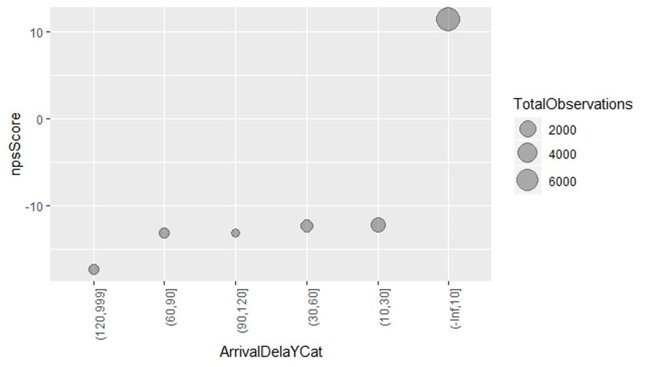
Bubble plot of type of travel categories displaying their NPS and number of observations:



**4.2.7 Arrival Delay in minutes**

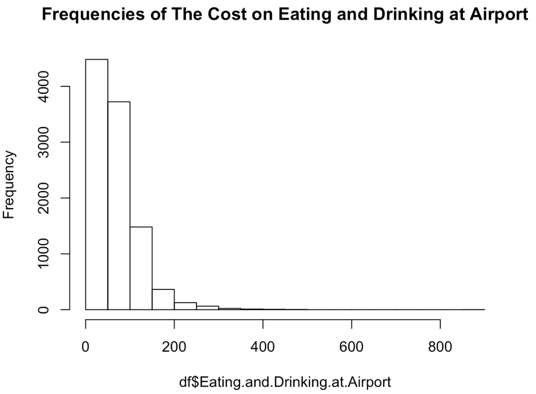
We divided Arrival Delay into categories i.e. binning.

Bubble plot of Arrival Delay categories displaying their NPS and number of observations:



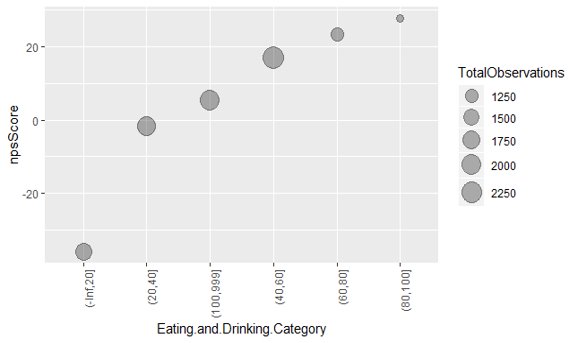
**4.2.8 Eating and drinking at airport**

Histogram of eating and drinking displaying the frequencies:



We divided eating and drinking into categories i.e. binning.

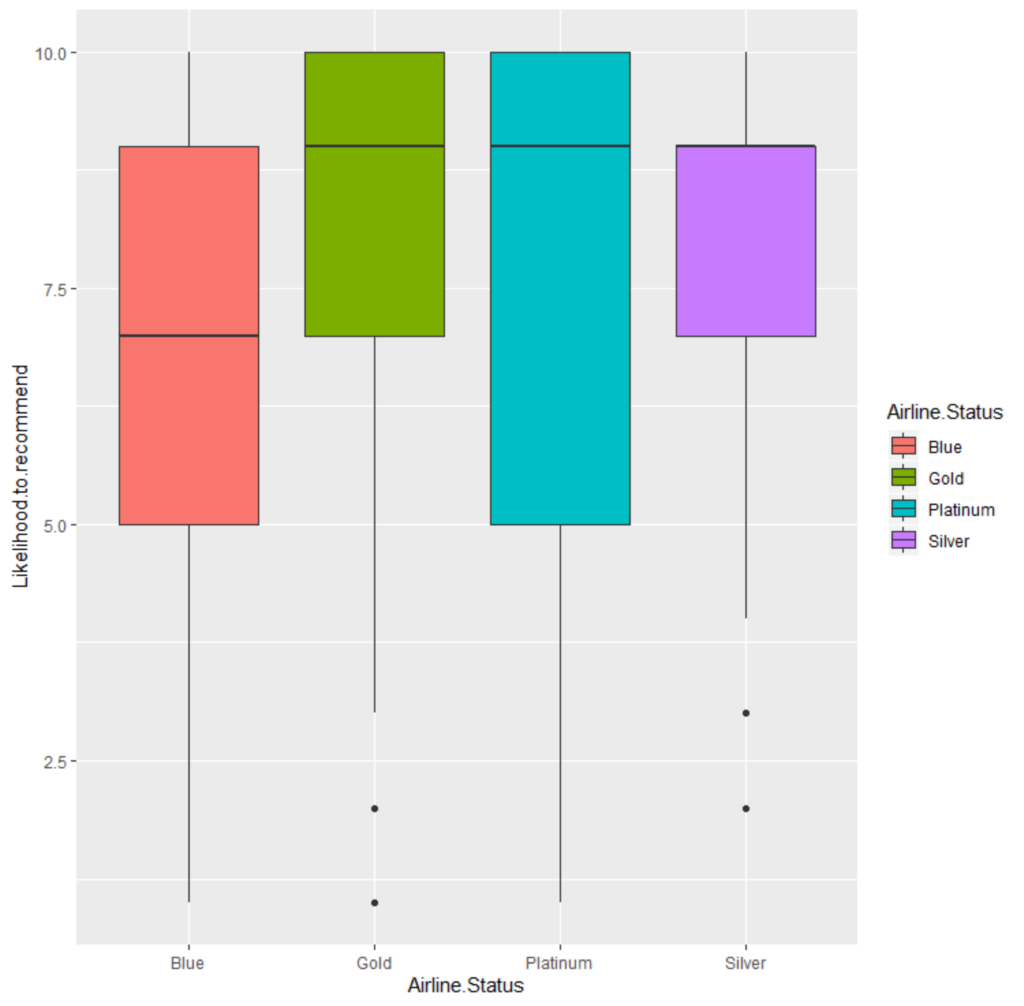
Bubble plot of eating and drinking categories displaying their NPS and number of observations:



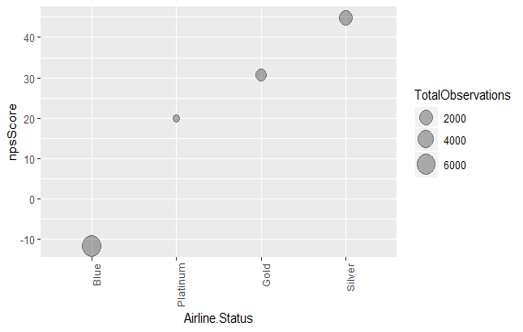
**4.2.9 Airline Status**

We divided airline status into categories i.e. binning.

Box plot for airline status:

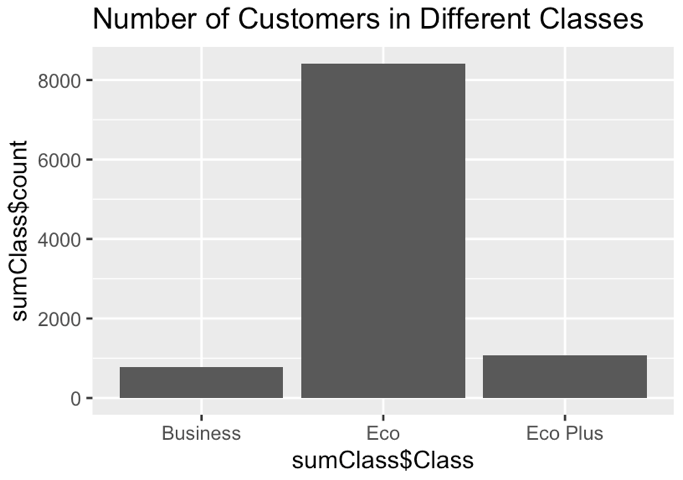


Bubble plot of airline status displaying their NPS and number of observations:

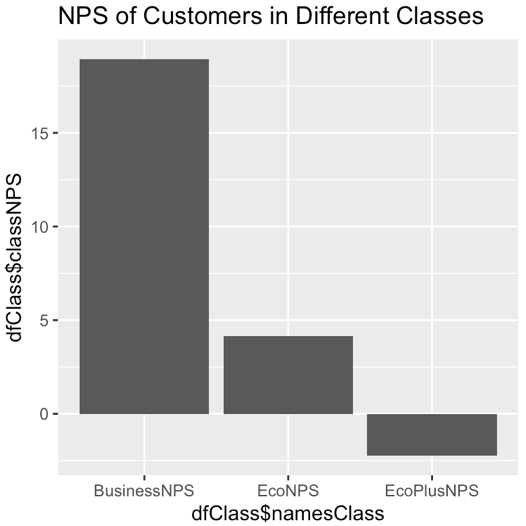


### **4.2.10 Class**

Bar chart for different classes:



Bar chart of NPS:



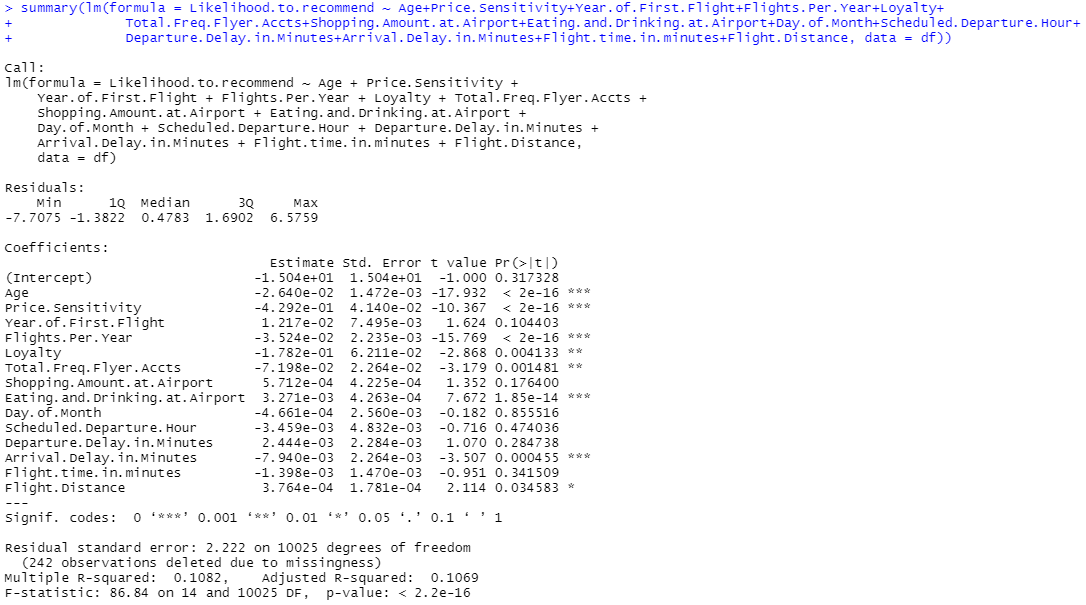
# **5. Predictive Modeling**

We performed two types of predictive modeling for the models:

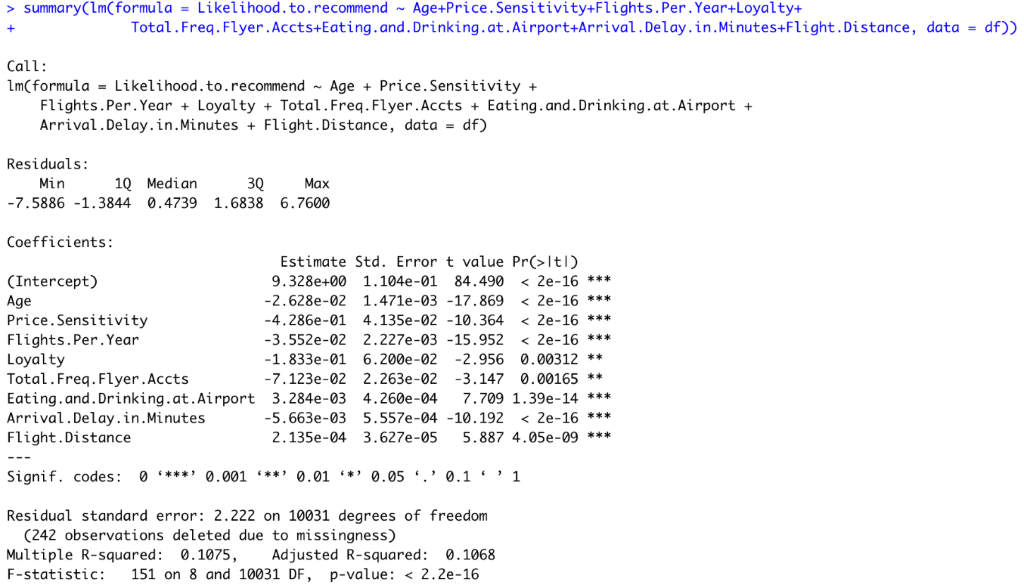
1. **Three level classification** – Target categories – Detractors, Passive, Promoters
2. **Two level classification** – Target categories – Detractors and (Passive+ Promoters) as Non-Detractors

## **5.1 Linear Regression**

**Linear regression model with all the numerical variables:**



**Linear regression model with significant variables:**



**Formula:**

Likelihood.to.recommend = (-2.628e-02) \* Age + (-4.286e-01) \* Price.Sensitivity + (-3.552e-02) \* Flights.Per.Year + (-1.833e-01) \* Loyalty + (-7.123e-02) \* Total.Freq.Flyer.Accts + (3.284e-03) \* Eating.and.Drinking.at.Airport + (-5.663e-03) \* Arrival.Delay.in.Minutes + (2.135e-04) \* Flight.Distance + (9.328e+00)

**Three level classification:**

* We also used linear regression to predict the likelihood to recommend and then transformed that predicted score into target categories – Detractors, passive and promoters.

**Accuracy, Precision & Recall for the linear regression model:**

Linear Regression model’s accuracy was around 54%.

A screenshot of a cell phone

Description automatically generated

We also created the linear model by separating the dataset based on whether the flight canceled or not, gender, age, airline status and type of travel, separately.

## **5.2 Random Forest**

**5.2.1 Three level classification**

**Random forest model:**

Parameters used – 500 trees

A screenshot of a cell phone

Description automatically generated

**Accuracy, Precision & Recall**:

Random Forest model had the highest accuracy and it was around 66%.

A screenshot of a cell phone

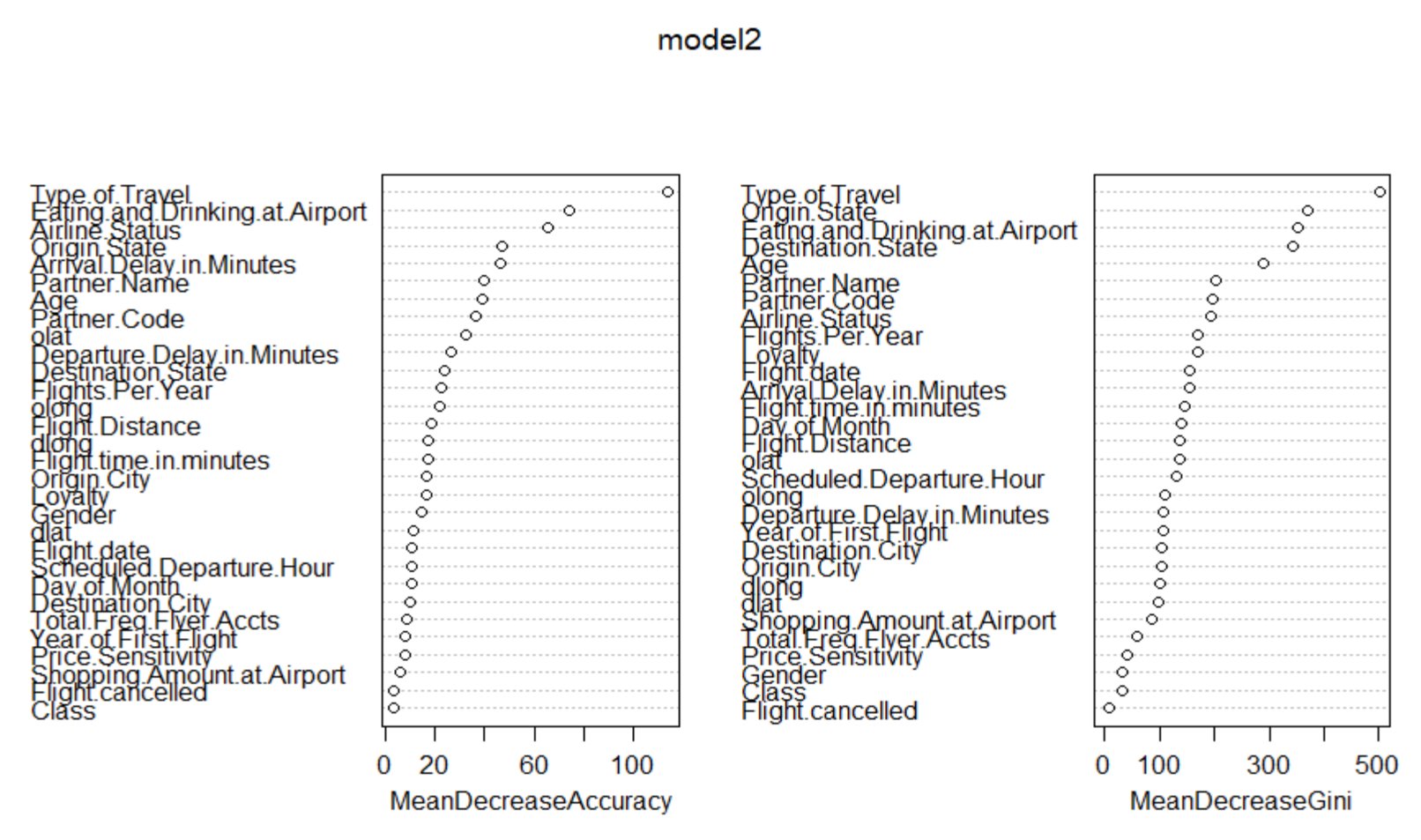
Description automatically generated

We found out the significant variables using the varImp function:

A close up of a newspaper

Description automatically generated

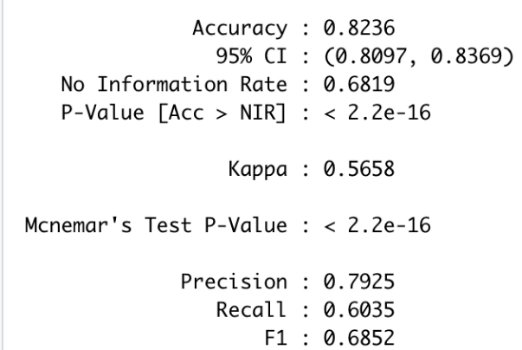
Variable Importance Plot of the top important variables from Random Forest:



### **5.2.2 Two level classification**

**Accuracy, Precision & Recall**:

Random Forest model had an accuracy of 82%.



## **5.3 Support Vector Machines**

**5.3.1 Three level classification**

**SVM model:**

Parameters – kernel used is radial basis function, kpar argument is automatic, cost of constraints is 5, cross validation factor used is 3

A screenshot of a cell phone

Description automatically generated

**Accuracy, Precision & Recall**:

SVM model had an accuracy of around 58%.

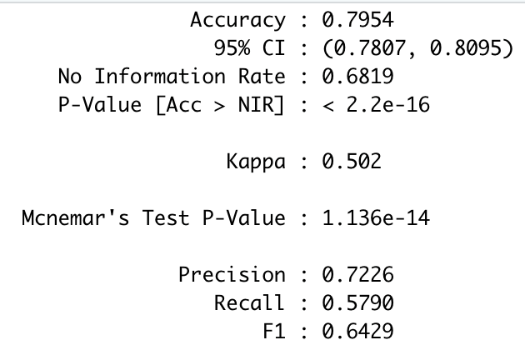
A screenshot of a cell phone

Description automatically generated

**5.3.2 Two level classification**

**Accuracy, Precision & Recall**:

SVM model had an accuracy of around 80%.



## **5.4 Decision Trees**

### **5.4.1 Three level classification**

**Classification and regression trees:**

A screenshot of a cell phone

Description automatically generated

**Accuracy, Precision & Recall**:

Decision trees model had an accuracy of around 60%.

A screenshot of a cell phone

Description automatically generated

**Significant variables using varImp function:**

A screenshot of text

Description automatically generated

**Decision tree model using all the variables:**

A close up of a map

Description automatically generated

**Decision tree model predicting the target categories:**

A close up of text on a white background

Description automatically generated

**5.4.2 Two level classification**

**Accuracy, Precision & Recall**:

Decision trees model had an accuracy of 82%.

A screenshot of a cell phone

Description automatically generated

## **5.5 Logistic Regression**

**Two level classification:**

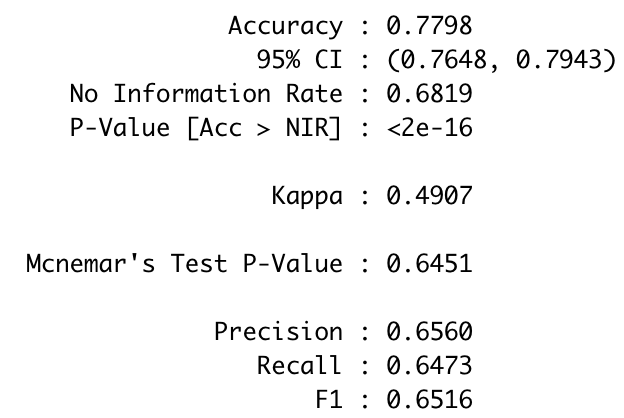
**Logistic regression model:**

A screenshot of a cell phone

Description automatically generated

**Accuracy, Precision & Recall**:

Logistic regression model had an accuracy of 78%.

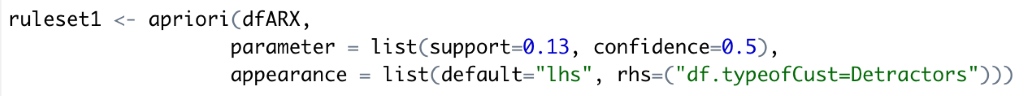


**5.6 Association Rules Mining**

We used apriori rules to predict rules for the detractors.

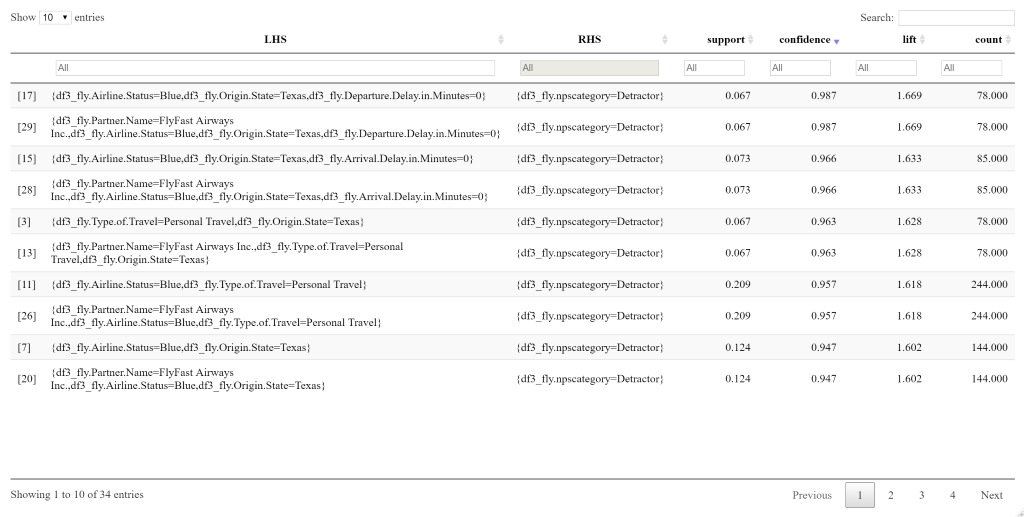
We only included the significant variables while implementing association rules.

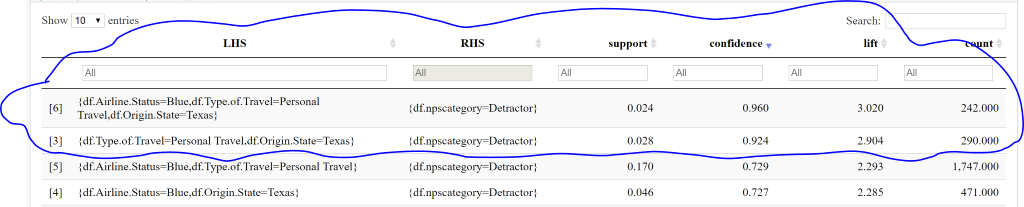
**Association rules:**

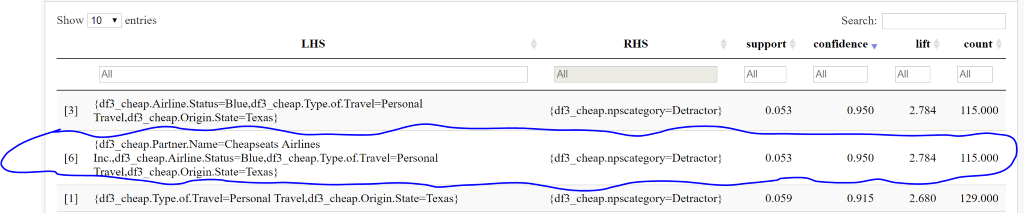


**Ruleset for detractors:**



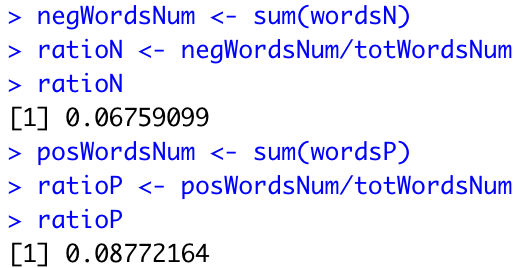






**5.7 Text Mining**

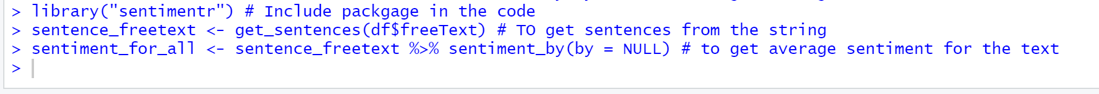
### **5.7.1 Sentiment Analysis**



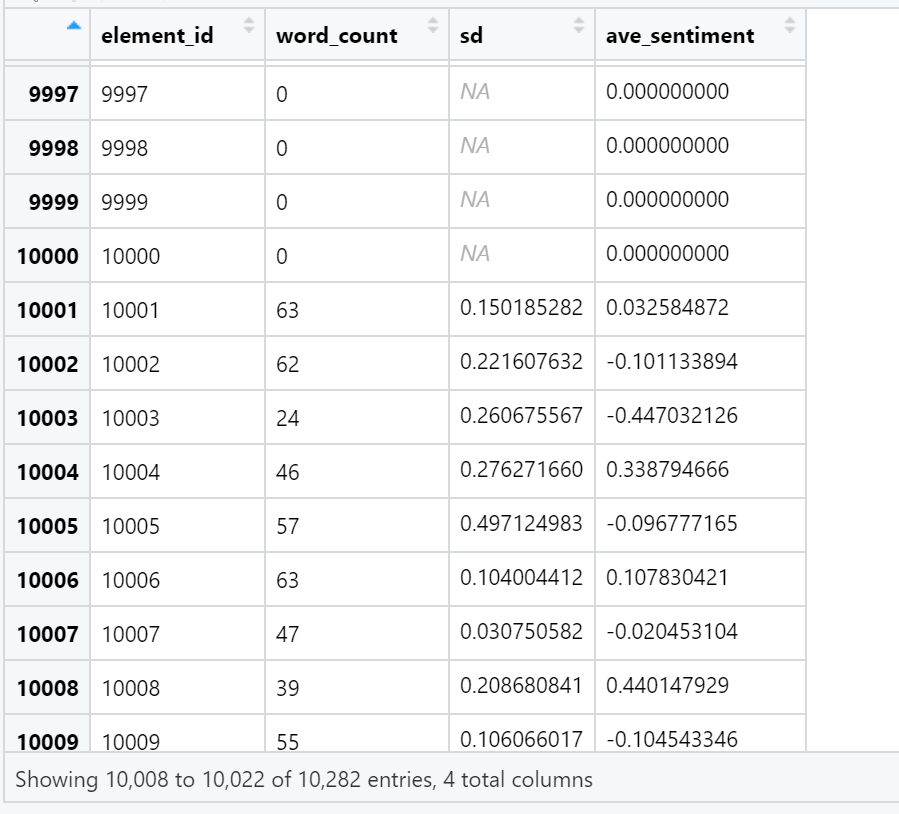
More positive words than negative words in the freeText.

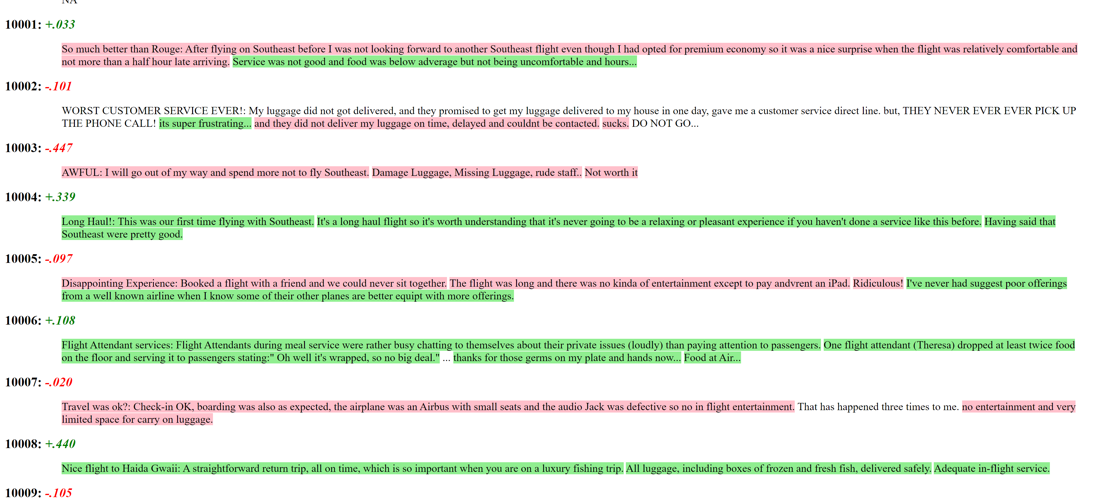
### **5.7.2 Individual Free Text Column Sentimental Analysis**

We have tried to analyze the Free Text Column Separately to find the average sentiment for each of the text which is a feedback from the customer. Each text has row number, word count and average sentiment how much the feedback is positive or negative.

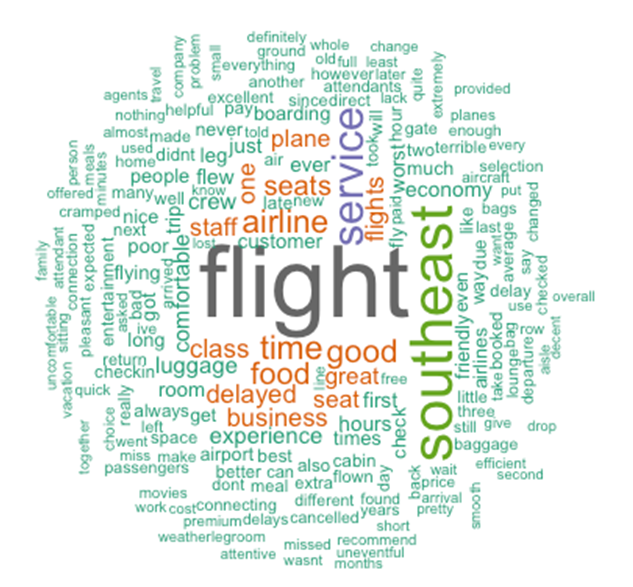


Examples:





### **5.7.3 World Cloud**



# **6. Insights and Recommendations**

## **6.1 Insights based on variables**

* **Age**
  + Most of the passengers flying are in the age group 30-50.
  + Highest NPS is for the age group 30-40.
  + Lowest NPS is for the passengers above 70.
* **Eating and Drinking amount at Airport**
  + Eating and Drinking amount spent by passive customers is more for short duration flights.
* **Origin & Destination State**
  + Most of the flights are flying in and out of Texas State which has lowest NPS and most detractors.
* **Loyalty**
  + 45% of loyal customers are Promoters.
* **Airline Status**
  + Airline Status Blue has highest number of flyers but lowest NPS score.
  + Highest NPS is for Airline Status Silver.
* **Partner Name**
  + FlyFast Airways has the lowest NPS.
  + Maximum passengers flying via Cheapseats Airlines.
* **Gender**
  + Female flyers more than male flyers. NPS is low for female passengers.
* **Type of Travel**
  + Most flyers are Business travelers.
  + Personal travelers are the detractors.
* **Arrival Delay**
  + NPS score decreases with increase in Arrival Delay.

## **6.2 Recommendations based on insights**

* For passengers flying with Fly Fast Airlines whose type of travel is personal and airline status is blue are the most dissatisfied customers. You need to take good care of them.
* For passengers flying with Cheap Seats Airlines whose type of travel is personal and airline status is blue and origin state is Texas, they are most likely to be detractors. Thus, you need to improve the services for those passengers.
* For flyers flying out of Texas state who are going on a personal trip and whose airline status is blue have lowest satisfaction score. You need to provide some offers to passengers flying out of Texas on personal trips.
* Female passengers have lower net promoter score compared to male passengers. So, you should do a survey and find out why that is happening.
* You need to provide more assistance to senior citizens during their check ins in order to attract more senior flyers.
* Passengers who spend less than $100 on eating and drinking at the airport and whose type of travel is personally have low satisfaction rate. You can provide some food and drink coupons to these passengers.
* If passengers going on personal trips experience delays in flights, then that also leads to low satisfaction scores. You need to decrease the delays in all the flights.
* People with negative loyalty score and whose airline status is blue are detractors. You need to improve services for blue passengers whose loyalty score is low.
* Male passengers in the age group 30-50 whose type of travel is personal are a big chunk of your overall flyers. You should maintain the good services that you are providing to them.

# **7. Appendix**

Please find the R code in the files:

* SEAirlines\_exploratoryAnalysis.R
* SEAirlines\_LinearMode.R
* SEAirlines\_PredictiveModels.R
* SEAirlines\_Apriori.R
* SEAirlines\_TextMining.R