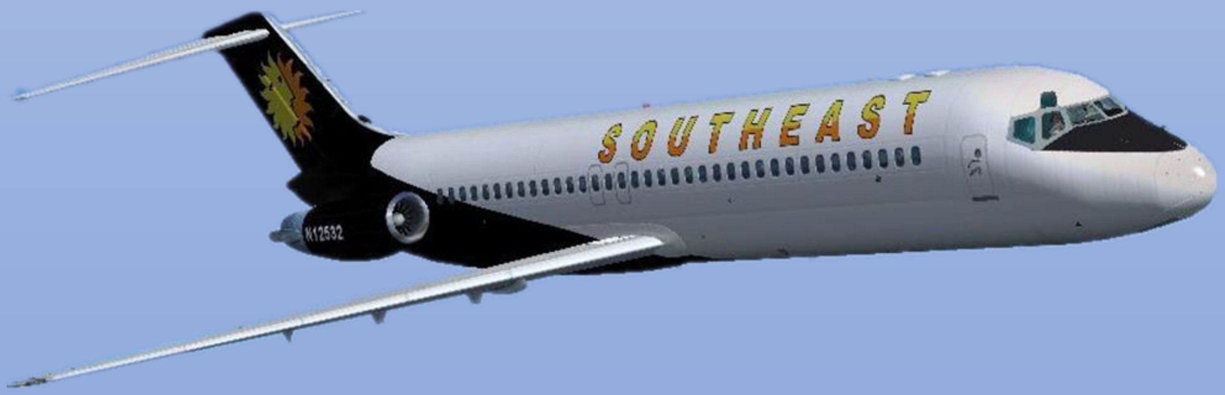
# Southeast Airlines Customer Satisfaction Analysis

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

## 1.1 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.

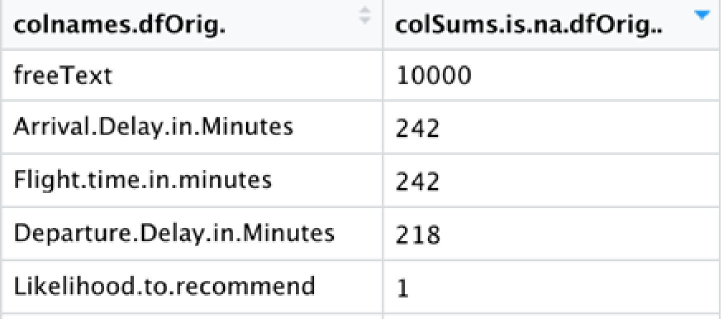
## 2. Business Questions addressed

- 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

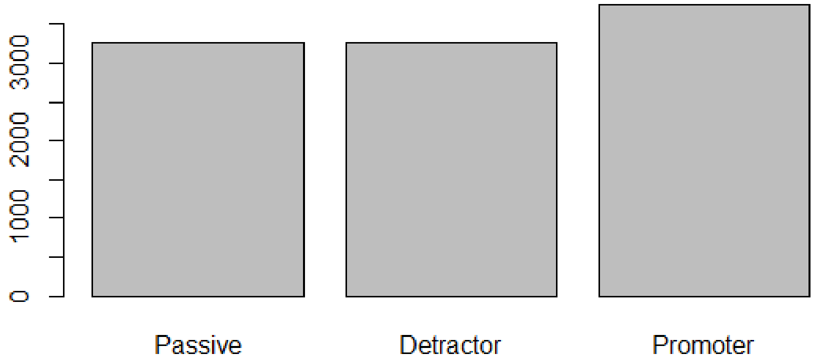


| colnames.dfOrig. | colSums.is.na.dfOrig.. |
|---|---|
| freeText | 10000 |
| Arrival.Delay.in.Minutes | 242 |
| Flight.time.in.minutes | 242 |
| Departure.Delay.in.Minutes | 218 |
| Likelihood.to.recommend | 1 |

- **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



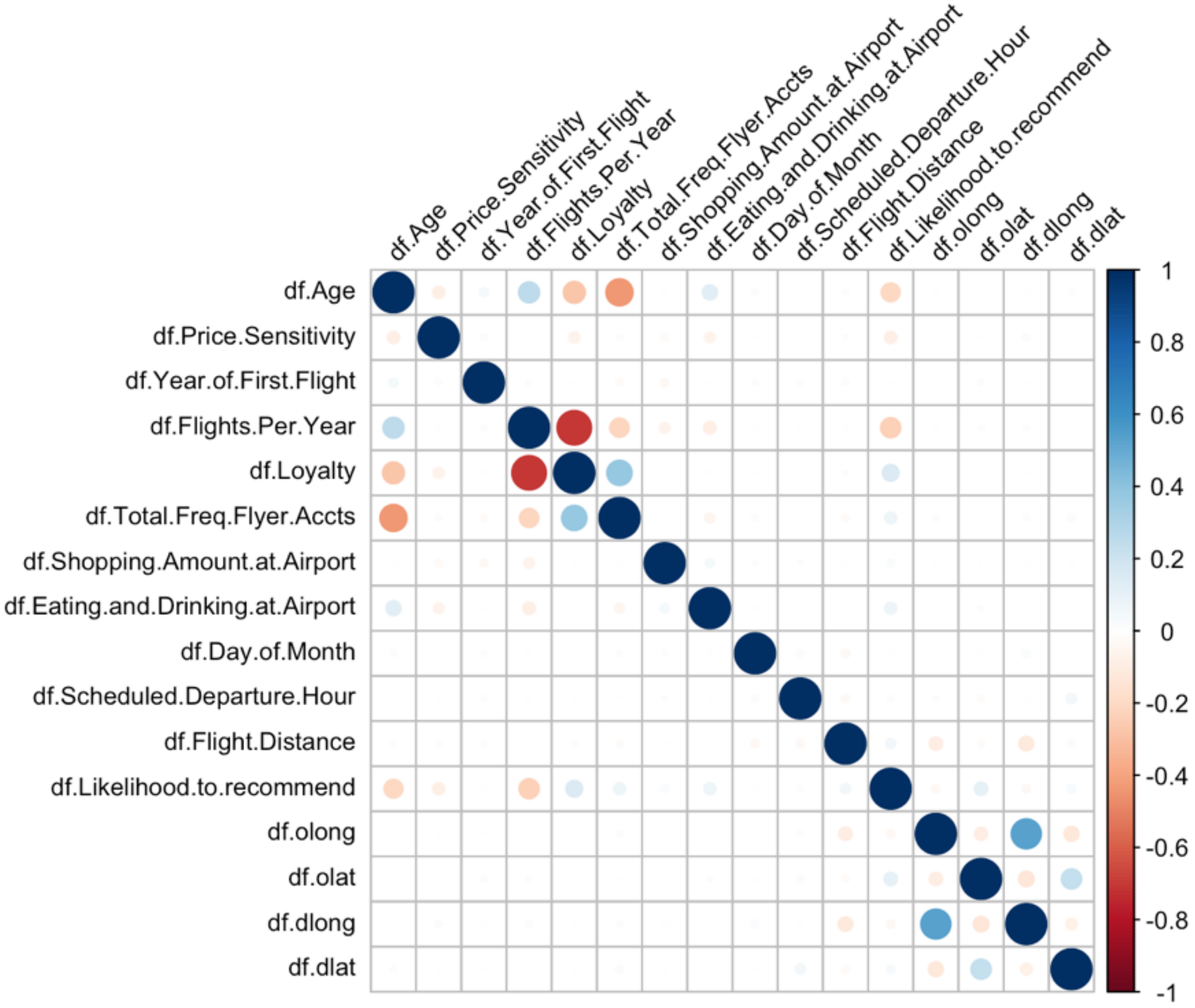
## 3.3 Performed transformations on variables

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:



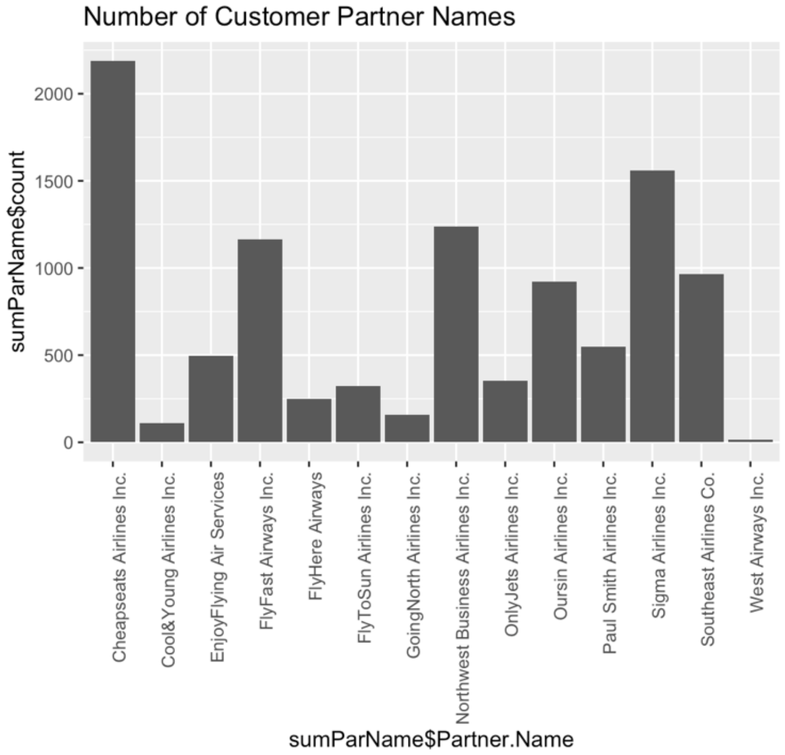
## 4.2 Variable Analysis

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.

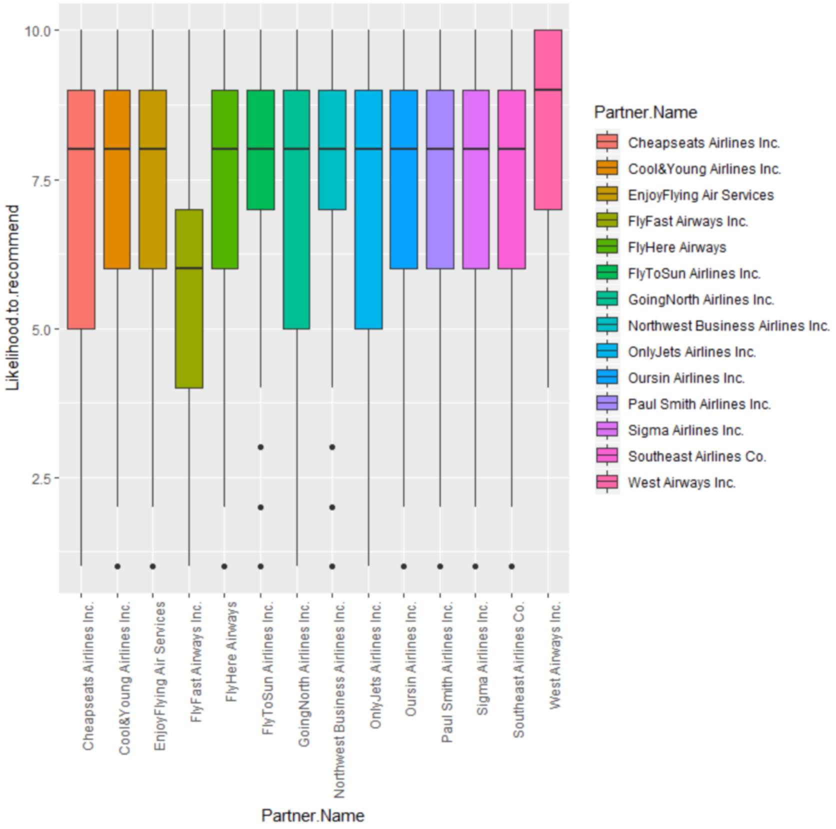
We calculated NPS of variables and plotted it for initial analysis:

### 4.2.1 Partner Airline

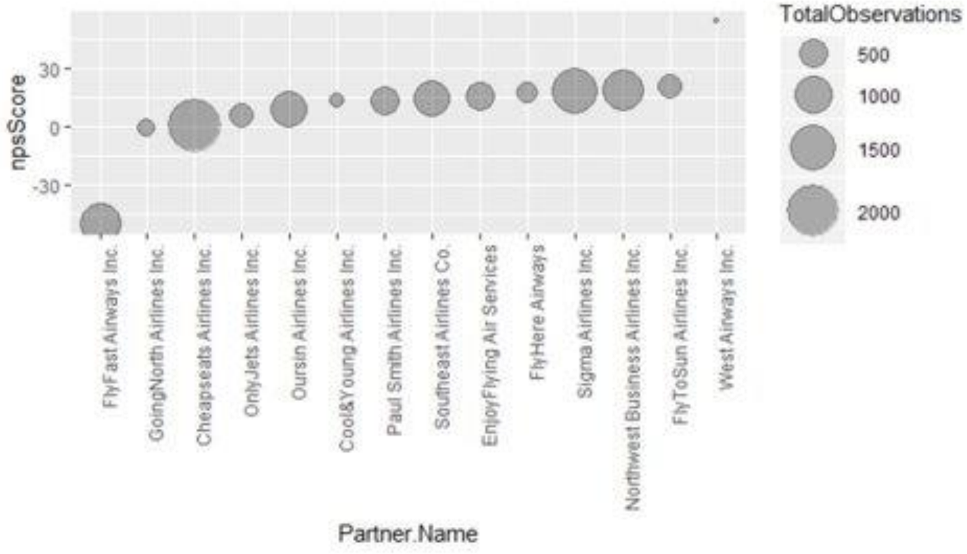
Bar plot of partner airlines displaying their number of observations:



Box plot for partner name vs. likelihood to recommend:



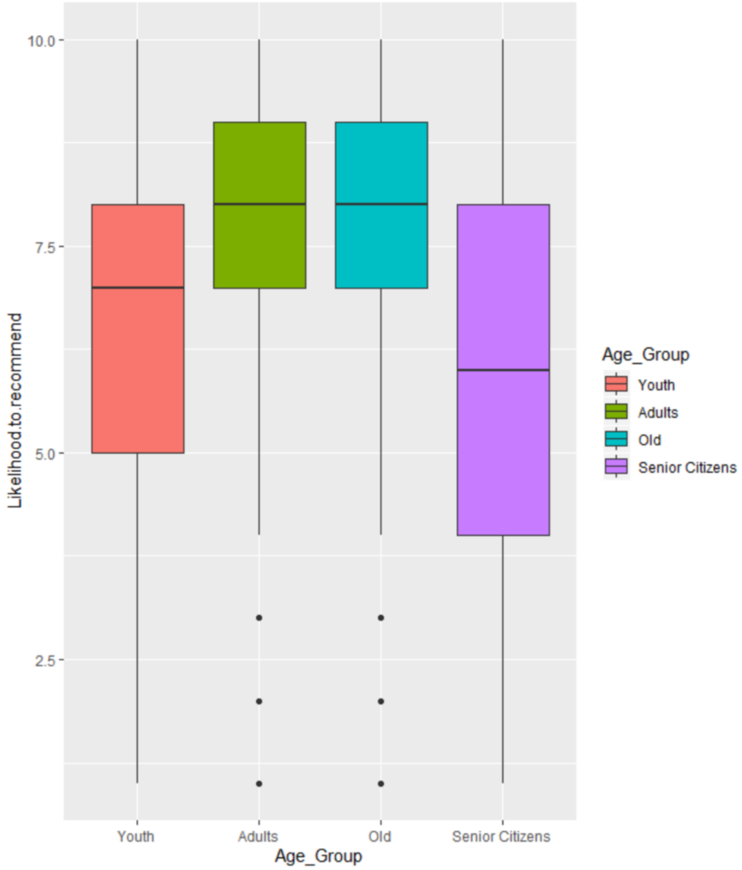
Bubble plot of partner airlines displaying their NPS and number of observations:



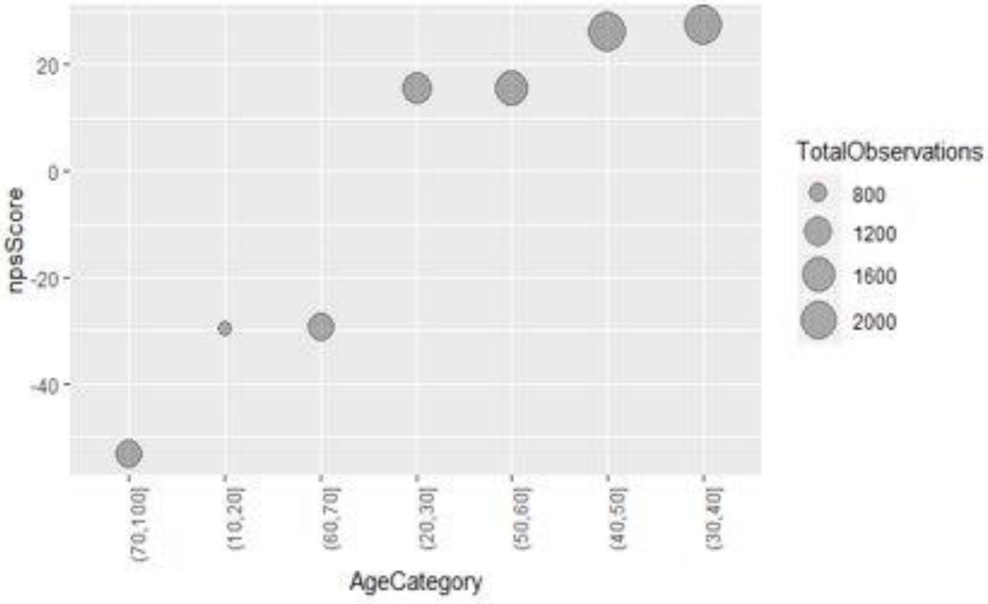
### 4.2.2 Age

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):

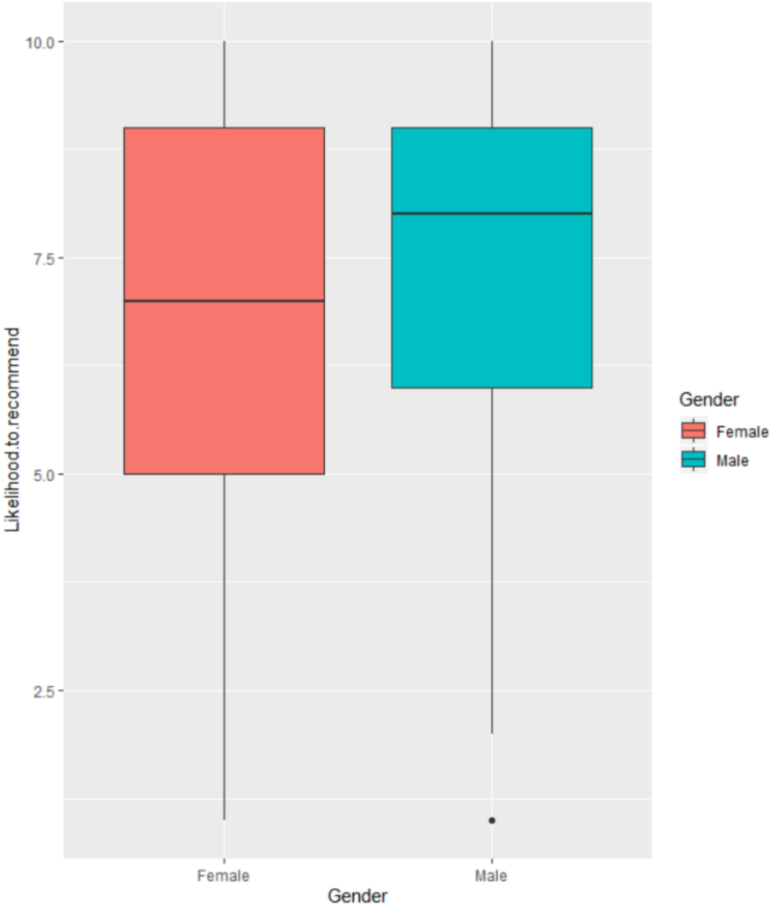


Bubble plot of its NPS and number of observations:

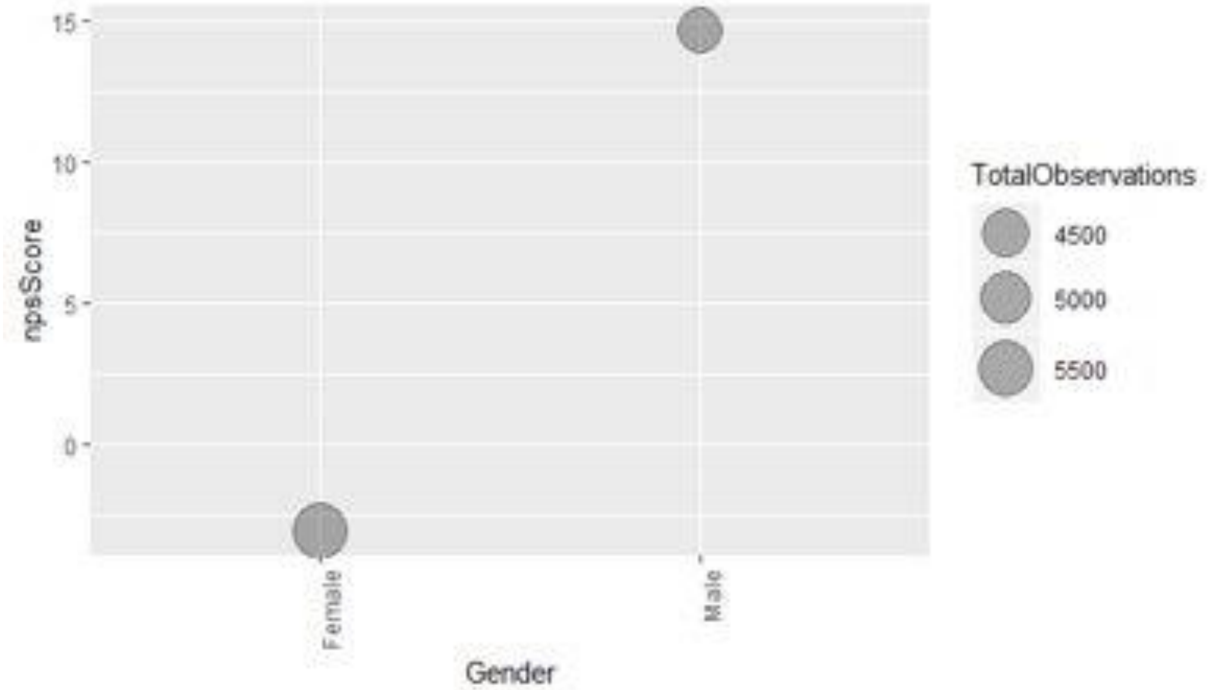


### 4.2.3 Gender

Boxplot for Gender Vs Likelihood to Recommend:

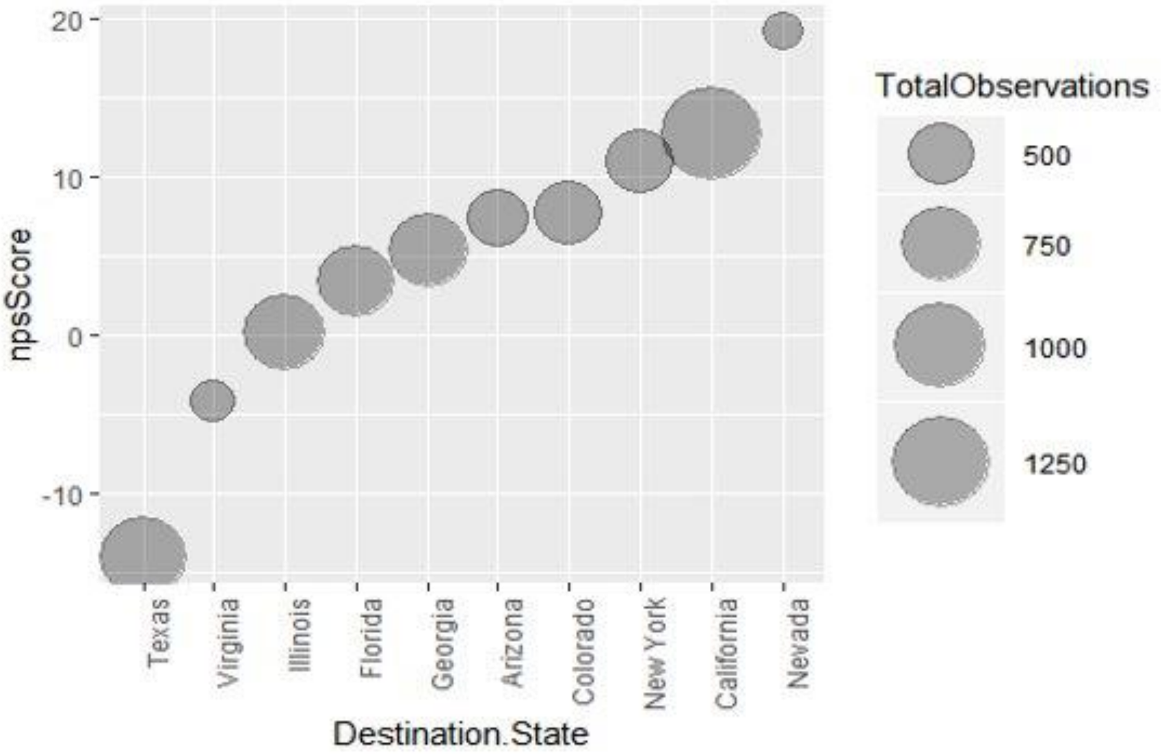


Bubble plot of gender displaying their NPS and number of observations:

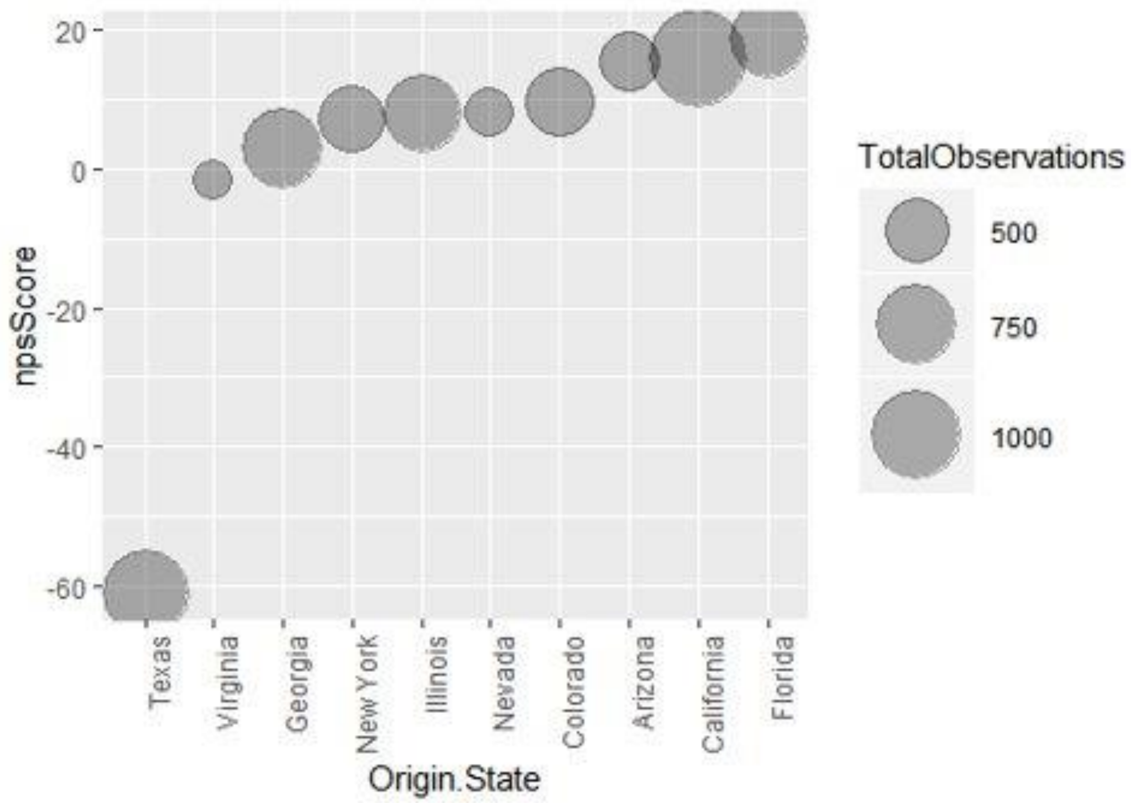


### 4.2.4 Destination and Origin State

Bubble plot of 10 Destination State with lowest NPS:



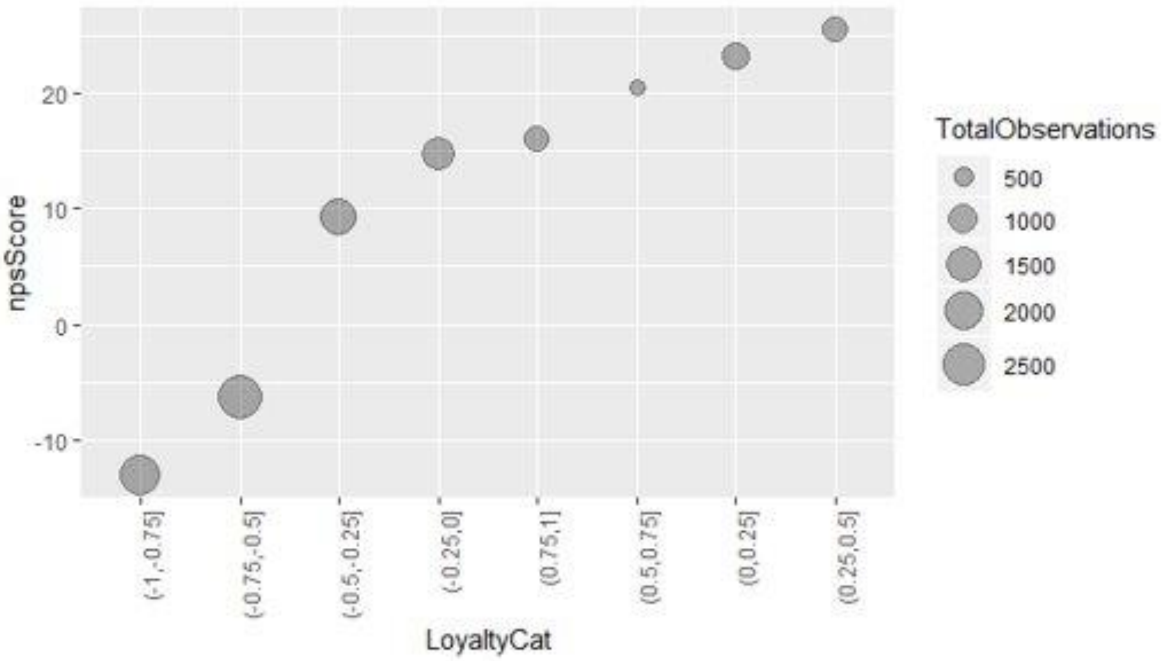
Bubble plot of 10 Origin State with lowest NPS:



### 4.2.5 Loyalty

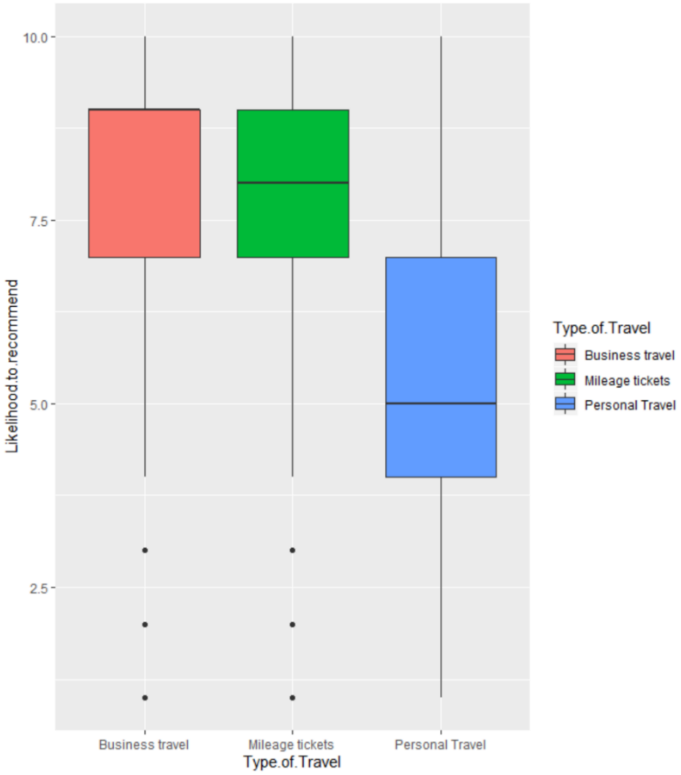
We divided loyalty score into categories i.e. binning.

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



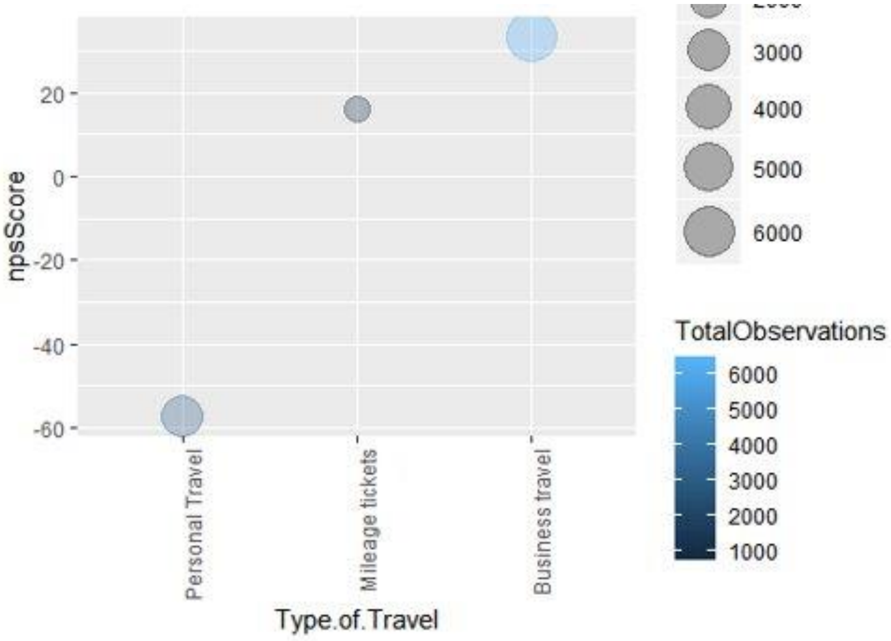
### 4.2.6 Type of travel

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.

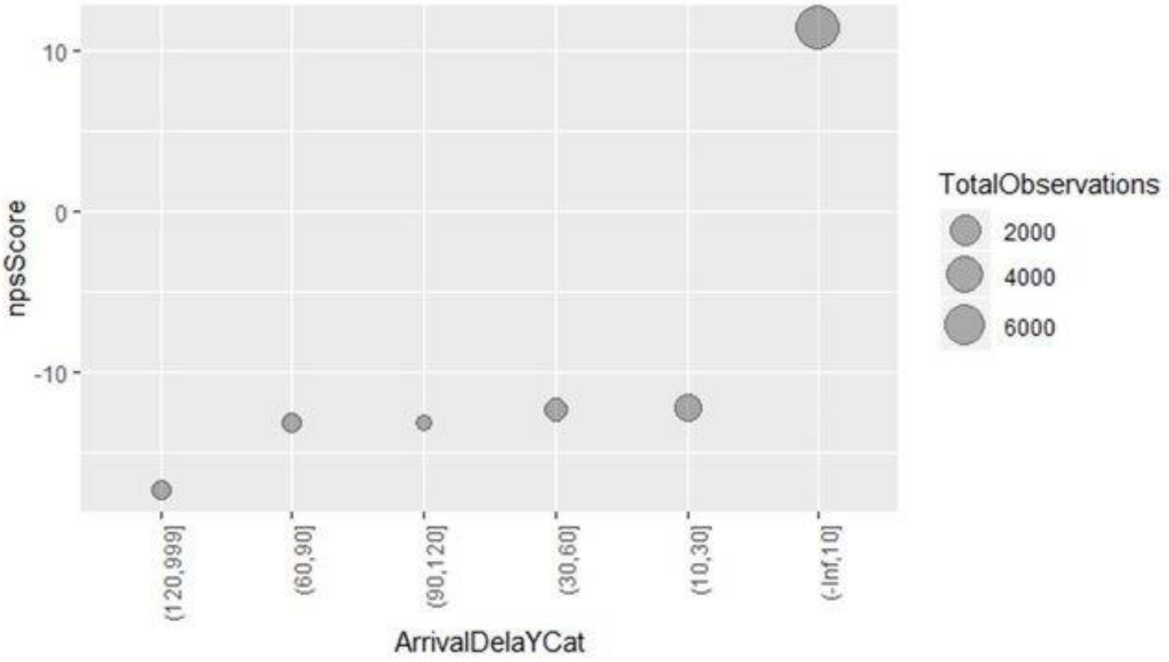
Bubble plot of type of travel categories displaying their NPS and number of observations:



### 4.2.7 Arrival Delay in minutes

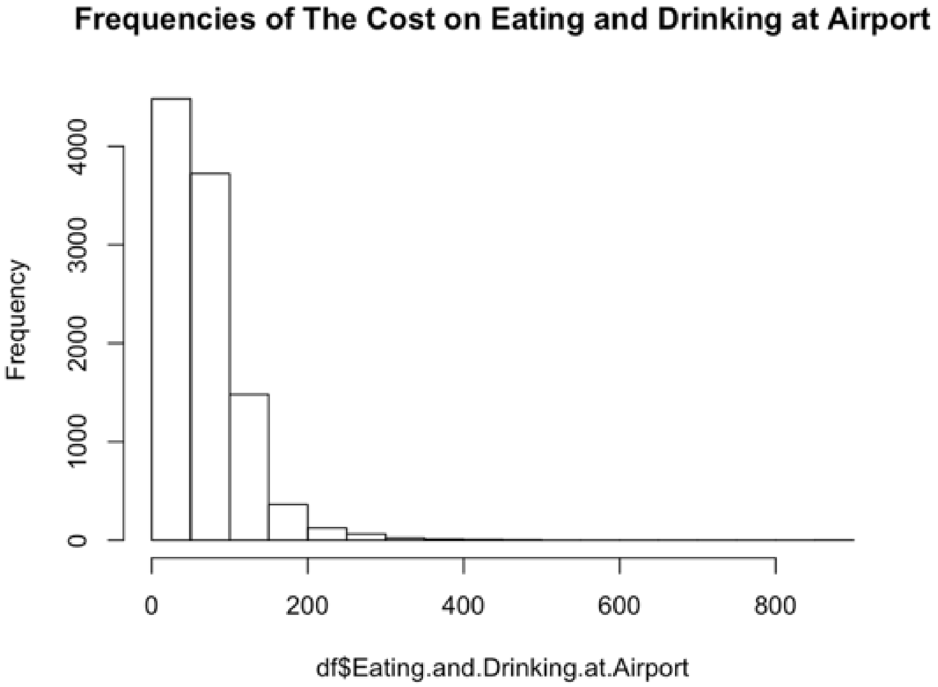
We divided Arrival Delay into categories i.e. binning.

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



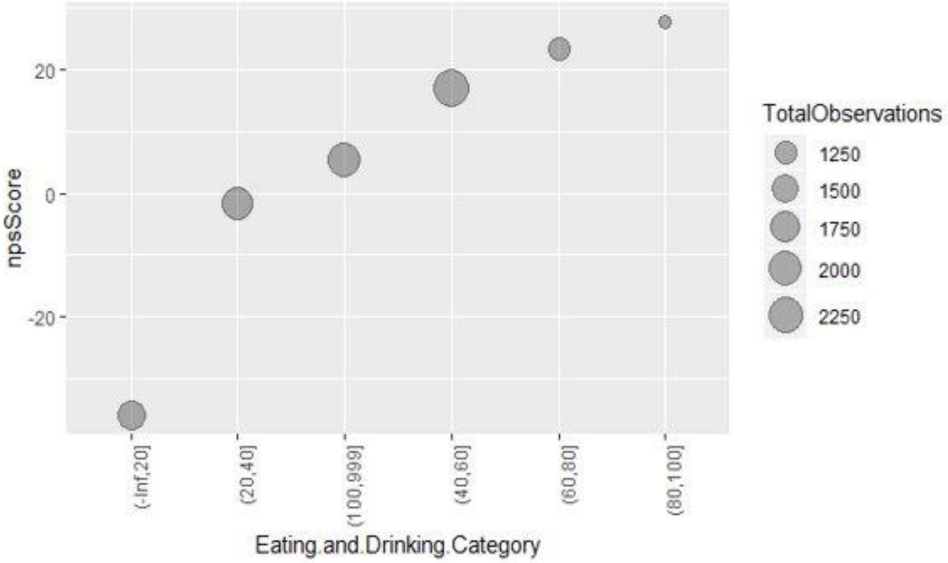
### 4.2.8 Eating and drinking at airport

Histogram of eating and drinking displaying the frequencies:



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

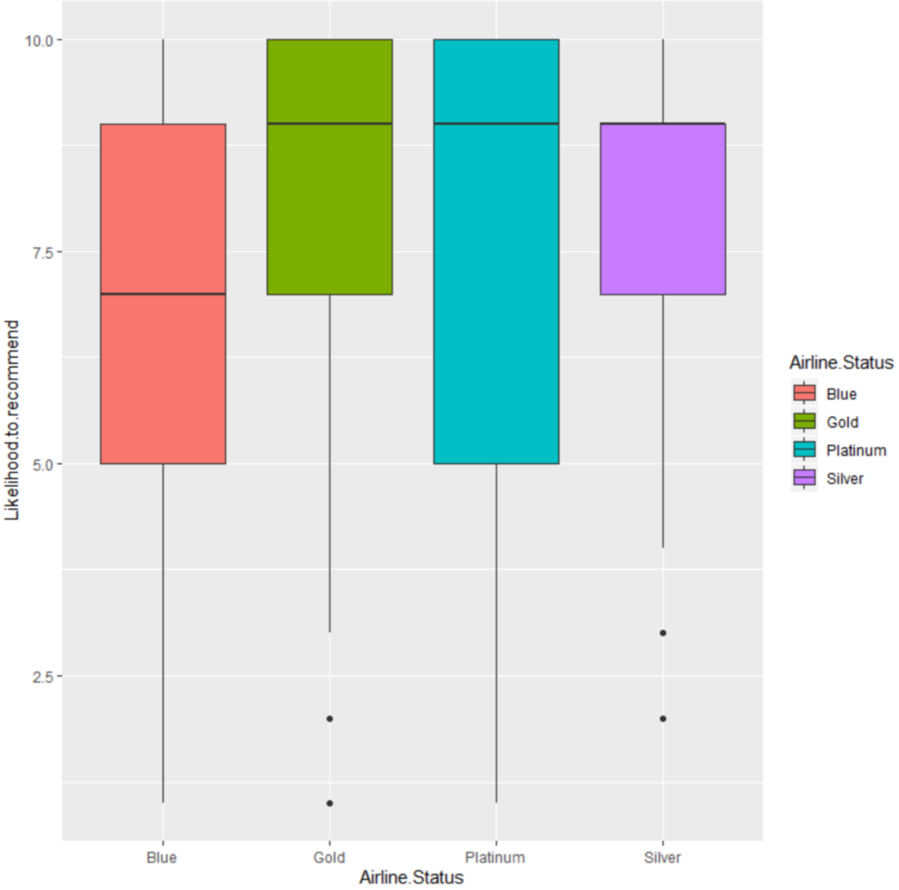
Bubble plot of eating and drinking categories displaying their NPS and number of observations:



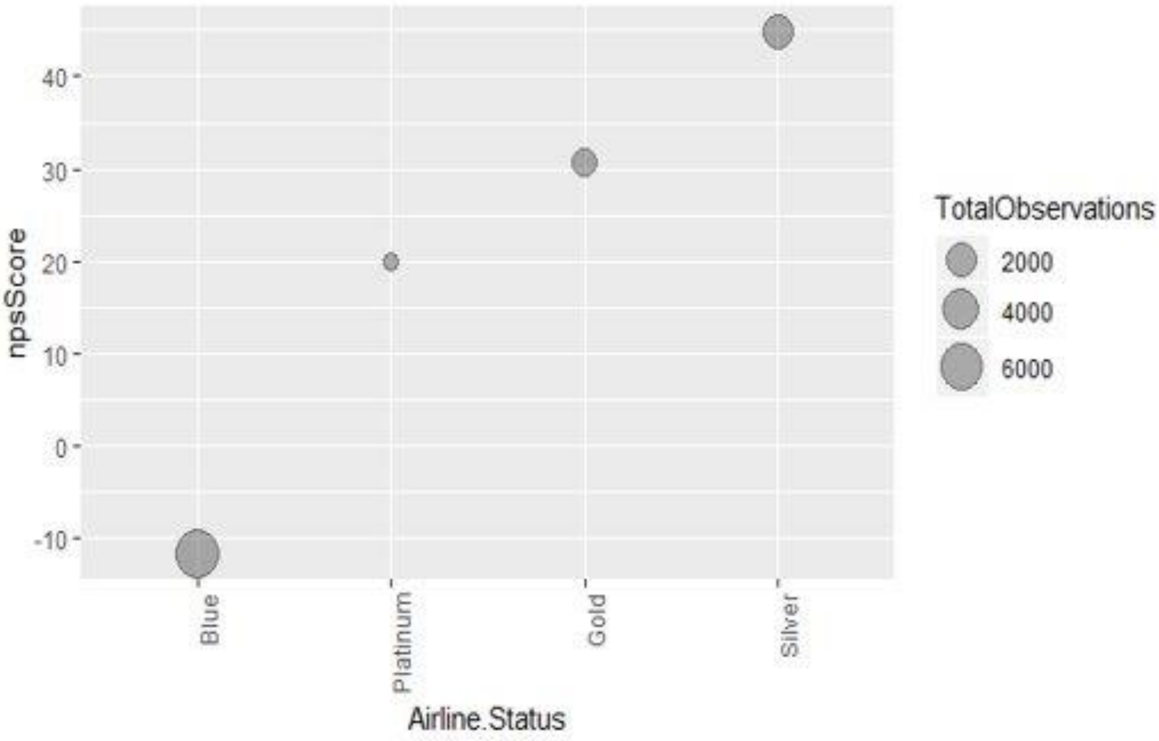
### 4.2.9 Airline Status

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

Box plot for airline status:

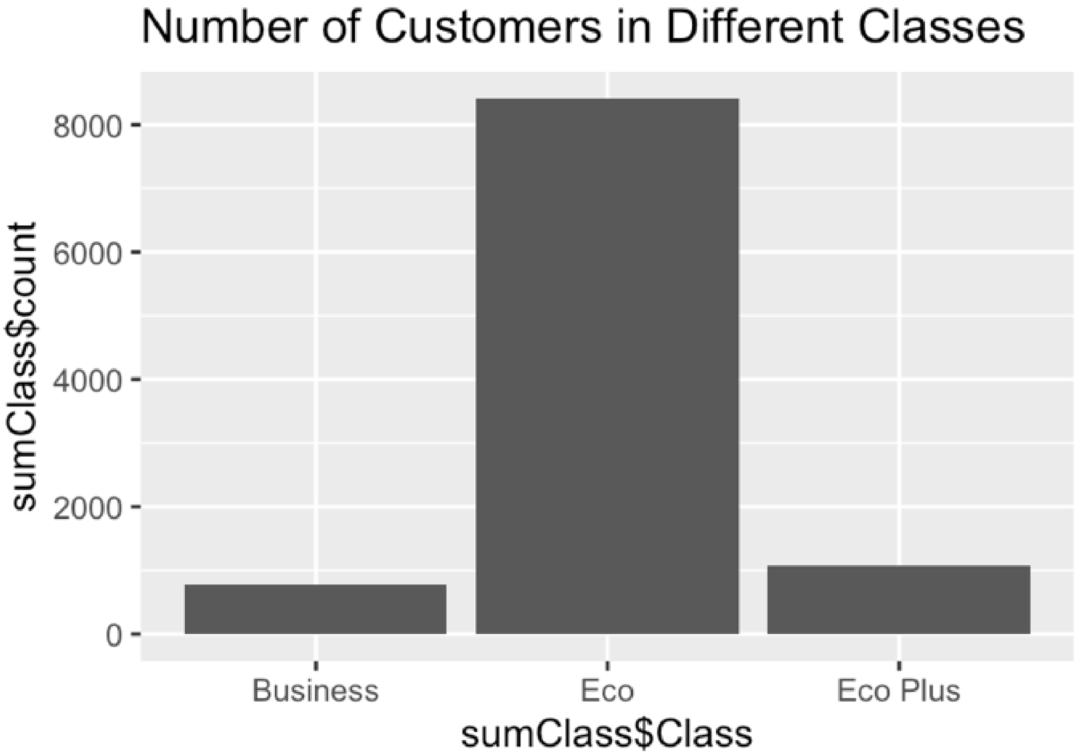


Bubble plot of airline status displaying their NPS and number of observations:

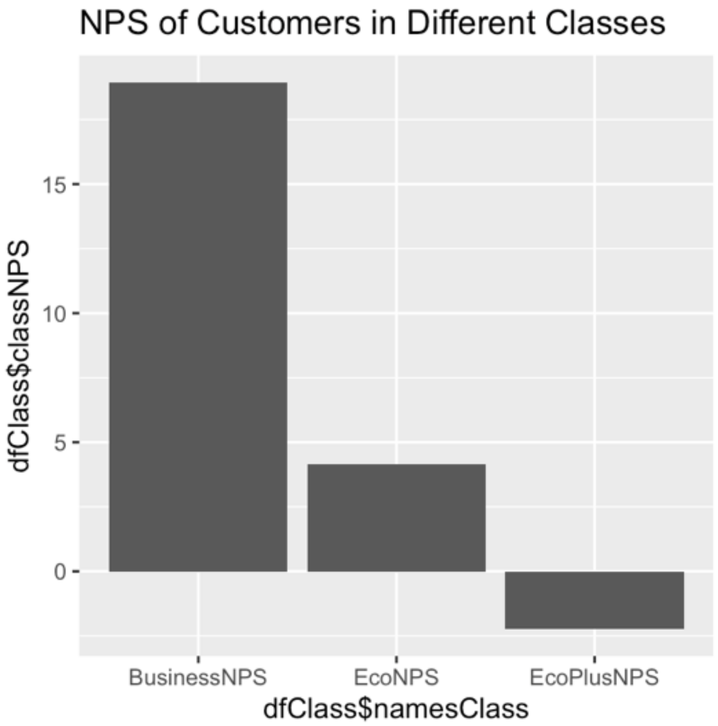


**4.2.10 Class**

Bar chart for different classes:



Bar chart of NPS:



# 5. Predictive Modeling

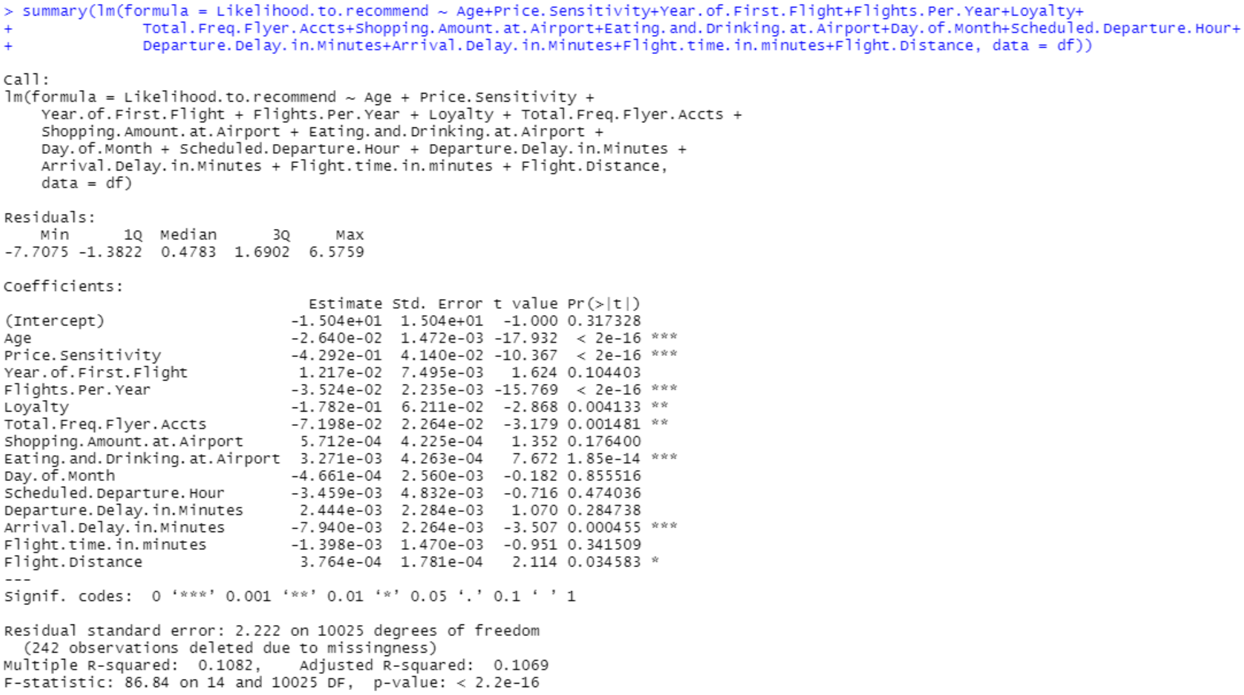
We performed two types of predictive modeling for the models:

1) **Three level classification** – Target categories – Detractors, Passive, Promoters

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

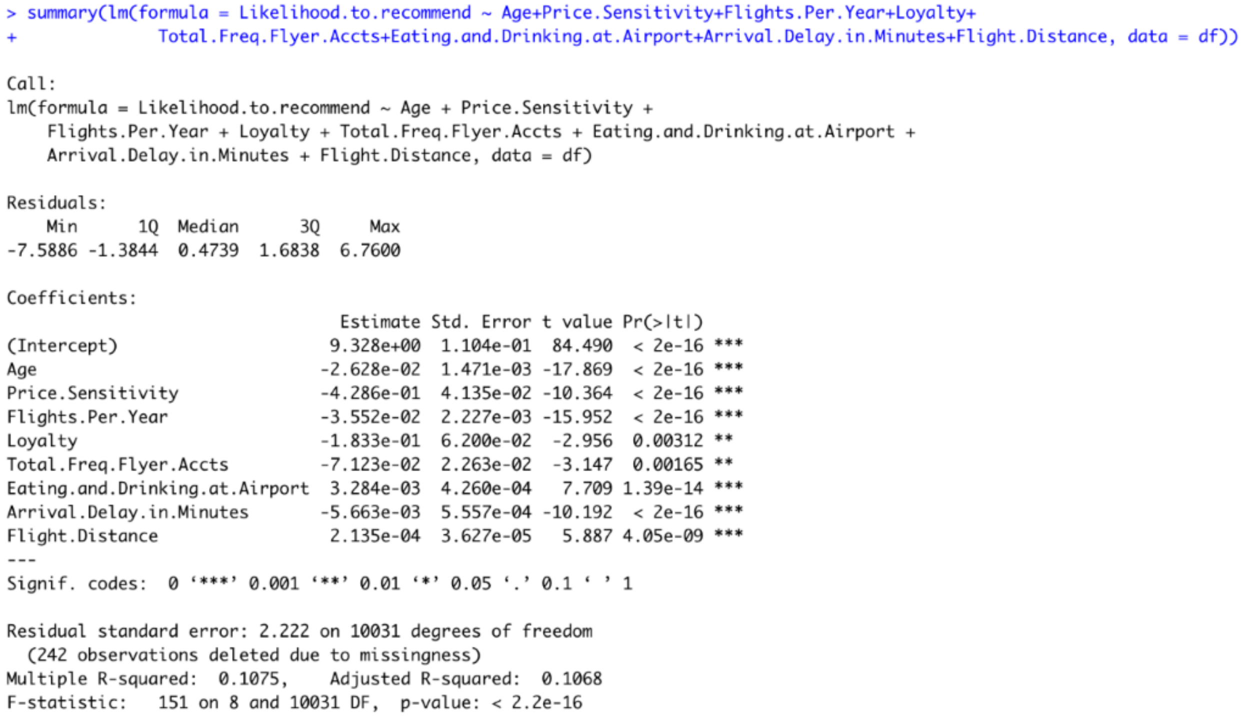
## 5.1 Linear Regression

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



```
> summary(lm(formula = Likelihood.to.recommend ~ Age+Price.Sensitivity+Year.of.First.Flight+Flights.Per.Year+Loyalty+
+              Total.Freq.Flyer.Accts+Shopping.Amount.at.Airport+Eating.and.Drinking.at.Airport+Day.of.Month+Scheduled.Departure.Hour+
+              Departure.Delay.in.Minutes+Arrival.Delay.in.Minutes+Flight.time.in.minutes+Flight.Distance, data = df))

Call:
lm(formula = Likelihood.to.recommend ~ Age + Price.Sensitivity +
    Year.of.First.Flight + Flights.Per.Year + Loyalty + Total.Freq.Flyer.Accts +
    Shopping.Amount.at.Airport + Eating.and.Drinking.at.Airport +
    Day.of.Month + Scheduled.Departure.Hour + Departure.Delay.in.Minutes +
    Arrival.Delay.in.Minutes + Flight.time.in.minutes + Flight.Distance,
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-7.7075 -1.3822  0.4783  1.6902  6.5759

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -1.504e+01  1.504e+01  -1.000 0.317328
Age                            -2.640e-02  1.472e-03 -17.932  < 2e-16 ***
Price.Sensitivity              -4.292e-01  4.140e-02 -10.367  < 2e-16 ***
Year.of.First.Flight            1.217e-02  7.495e-03   1.624 0.104403
Flights.Per.Year               -3.524e-02  2.235e-03 -15.769  < 2e-16 ***
Loyalty                        -1.782e-01  6.211e-02  -2.868 0.004133 **
Total.Freq.Flyer.Accts         -7.198e-02  2.264e-02  -3.179 0.001481 **
Shopping.Amount.at.Airport      5.712e-04  4.225e-04   1.352 0.176400
Eating.and.Drinking.at.Airport  3.271e-03  4.263e-04   7.672 1.85e-14 ***
Day.of.Month                   -4.661e-04  2.560e-03  -0.182 0.855516
Scheduled.Departure.Hour       -3.459e-03  4.832e-03  -0.716 0.474036
Departure.Delay.in.Minutes      2.444e-03  2.284e-03   1.070 0.284738
Arrival.Delay.in.Minutes       -7.940e-03  2.264e-03  -3.507 0.000455 ***
Flight.time.in.minutes         -1.398e-03  1.470e-03  -0.951 0.341509
Flight.Distance                 3.764e-04  1.781e-04   2.114 0.034583 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.222 on 10025 degrees of freedom
  (242 observations deleted due to missingness)
Multiple R-squared:  0.1082,    Adjusted R-squared:  0.1069
F-statistic: 86.84 on 14 and 10025 DF,  p-value: < 2.2e-16
```

### Linear regression model with significant variables:



```
> summary(lm(formula = Likelihood.to.recommend ~ Age+Price.Sensitivity+Flights.Per.Year+Loyalty+
+              Total.Freq.Flyer.Accts+Eating.and.Drinking.at.Airport+Arrival.Delay.in.Minutes+Flight.Distance, data = df))

Call:
lm(formula = Likelihood.to.recommend ~ Age + Price.Sensitivity +
    Flights.Per.Year + Loyalty + Total.Freq.Flyer.Accts + Eating.and.Drinking.at.Airport +
    Arrival.Delay.in.Minutes + Flight.Distance, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-7.5886 -1.3844  0.4739  1.6838  6.7600

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     9.328e+00  1.104e-01  84.490  < 2e-16 ***
Age                            -2.628e-02  1.471e-03 -17.869  < 2e-16 ***
Price.Sensitivity              -4.286e-01  4.135e-02 -10.364  < 2e-16 ***
Flights.Per.Year               -3.552e-02  2.227e-03 -15.952  < 2e-16 ***
Loyalty                        -1.833e-01  6.200e-02  -2.956  0.00312 **
Total.Freq.Flyer.Accts         -7.123e-02  2.263e-02  -3.147  0.00165 **
Eating.and.Drinking.at.Airport  3.284e-03  4.260e-04   7.709 1.39e-14 ***
Arrival.Delay.in.Minutes       -5.663e-03  5.557e-04 -10.192  < 2e-16 ***
Flight.Distance                 2.135e-04  3.627e-05   5.887 4.05e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.222 on 10031 degrees of freedom
  (242 observations deleted due to missingness)
Multiple R-squared:  0.1075,    Adjusted R-squared:  0.1068
F-statistic:   151 on 8 and 10031 DF,  p-value: < 2.2e-16
```

### Formula:

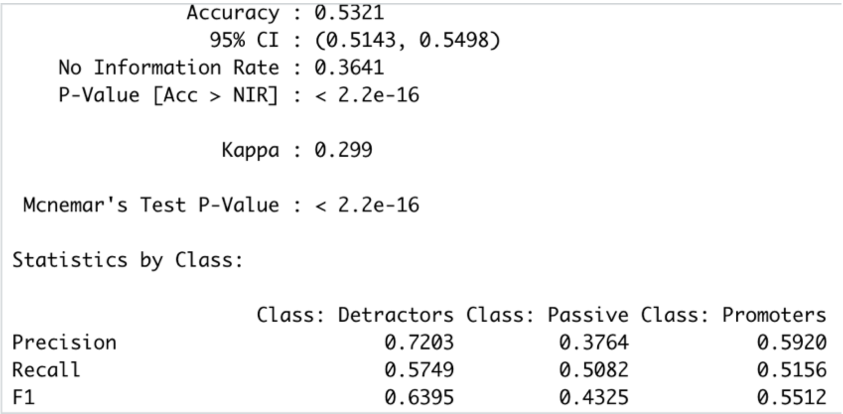
Likelihood.to.recommend = (-2.628e-02) * Age + (-4.286e-01) * Price.Sensitivity + (-3.552e-02) * Flights.Per.Year + (-1.833e-01) * Loyalty + (-7.123e-02) * Total.Freq.Flyer.Accts + (3.284e-03) * Eating.and.Drinking.at.Airport + (-5.663e-03) * Arrival.Delay.in.Minutes + (2.135e-04) * Flight.Distance +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



```
               Accuracy : 0.5321
                 95% CI : (0.5143, 0.5498)
    No Information Rate : 0.3641
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.299

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: Detractors Class: Passive Class: Promoters
Precision                       0.7203         0.3764           0.5920
Recall                          0.5749         0.5082           0.5156
F1                              0.6395         0.4325           0.5512
```

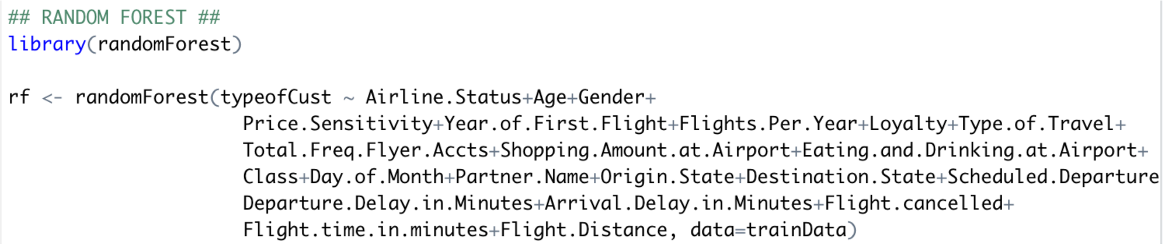
We also created the linear model by separating the dataset based on whether the flight canceled or not, gender, age, airline status and type of travel, separately.

## 5.2 Random Forest

### 5.2.1 Three level classification

**Random forest model:**

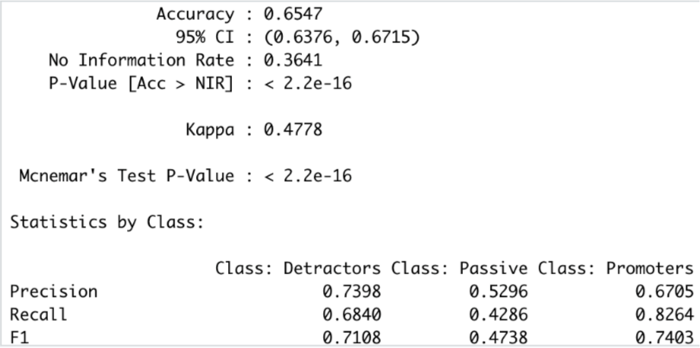
Parameters used – 500 trees



```
## RANDOM FOREST ##
library(randomForest)

rf <- randomForest(typeofCust ~ Airline.Status+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+Day.of.Month+Partner.Name+Origin.State+Destination.State+Scheduled.Departure
                       Departure.Delay.in.Minutes+Arrival.Delay.in.Minutes+Flight.cancelled+
                       Flight.time.in.minutes+Flight.Distance, data=trainData)
```

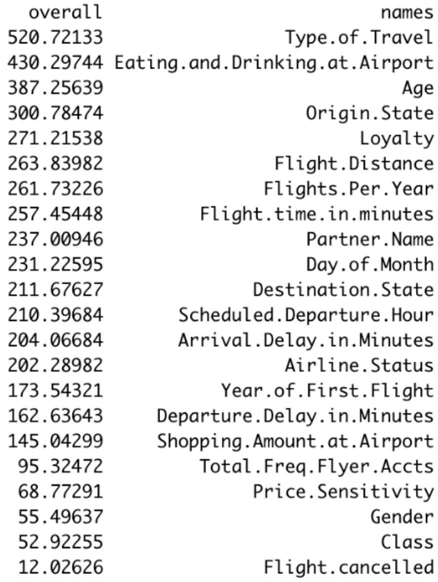
**Accuracy, Precision & Recall**:

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



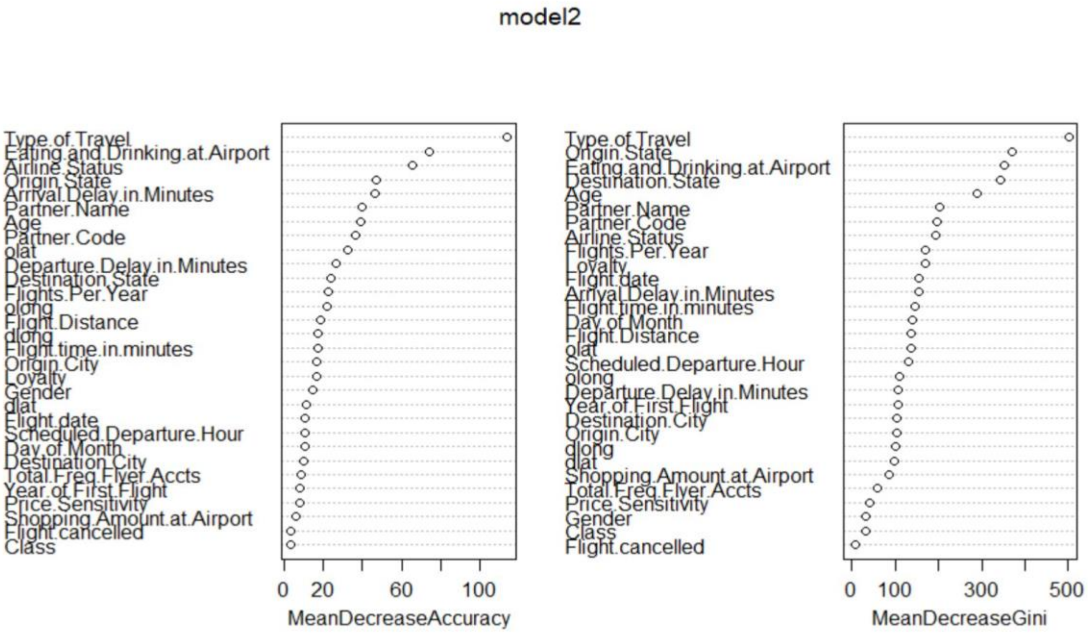
```
               Accuracy : 0.6547
                 95% CI : (0.6376, 0.6715)
    No Information Rate : 0.3641
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4778

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: Detractors Class: Passive Class: Promoters
Precision                       0.7398         0.5296           0.6705
Recall                          0.6840         0.4286           0.8264
F1                              0.7108         0.4738           0.7403
```

We found out the significant variables using the varImp function:



```
   overall                           names
 520.72133                  Type.of.Travel
 430.29744 Eating.and.Drinking.at.Airport
 387.25639                             Age
 300.78474                    Origin.State
 271.21538                         Loyalty
 263.83982                 Flight.Distance
 261.73226                Flights.Per.Year
 257.45448           Flight.time.in.minutes
 237.00946                    Partner.Name
 231.22595                    Day.of.Month
 211.67627               Destination.State
 210.39684        Scheduled.Departure.Hour
 204.06684        Arrival.Delay.in.Minutes
 202.28982                  Airline.Status
 173.54321            Year.of.First.Flight
 162.63643      Departure.Delay.in.Minutes
 145.04299      Shopping.Amount.at.Airport
  95.32472          Total.Freq.Flyer.Accts
  68.77291               Price.Sensitivity
  55.49637                          Gender
  52.92255                           Class
  12.02626                Flight.cancelled
```

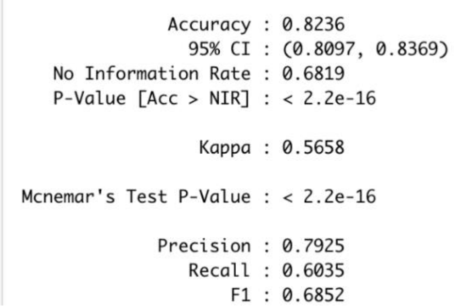
Variable Importance Plot of the top important variables from Random Forest:



### 5.2.2 Two level classification

**Accuracy, Precision & Recall**:

Random Forest model had an accuracy of 82%.



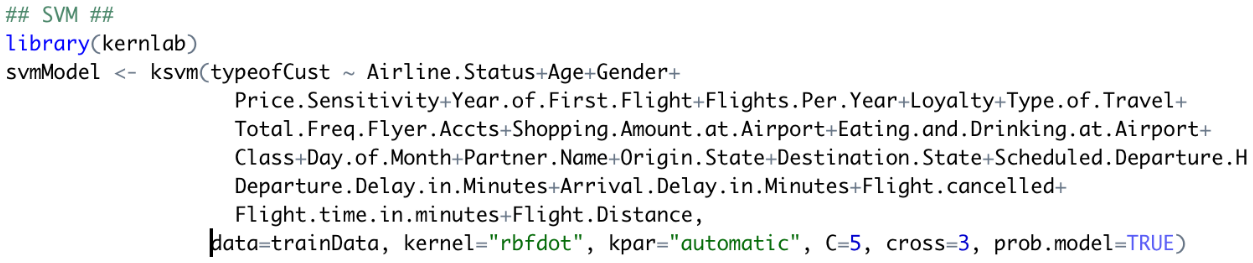
```
               Accuracy : 0.8236
                 95% CI : (0.8097, 0.8369)
    No Information Rate : 0.6819
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5658

 Mcnemar's Test P-Value : < 2.2e-16

              Precision : 0.7925
                 Recall : 0.6035
                     F1 : 0.6852
```

## 5.3 Support Vector Machines

### 5.3.1 Three level classification

**SVM model:**

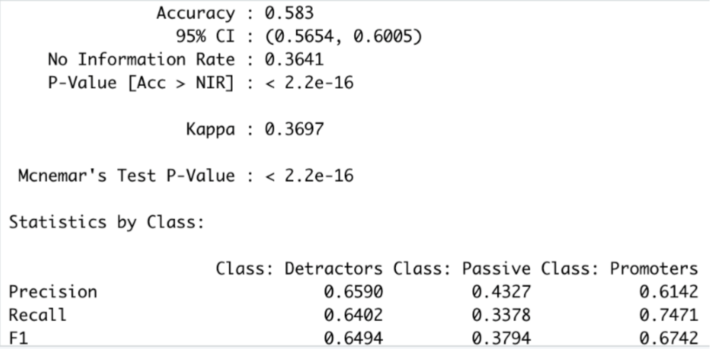
Parameters – kernel used is radial basis function, kpar argument is automatic, cost of constraints is 5, cross validation factor used is 3



```
## SVM ##
library(kernlab)
svmModel <- ksvm(typeofCust ~ Airline.Status+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+Day.of.Month+Partner.Name+Origin.State+Destination.State+Scheduled.Departure.H
                      Departure.Delay.in.Minutes+Arrival.Delay.in.Minutes+Flight.cancelled+
                      Flight.time.in.minutes+Flight.Distance,
                    data=trainData, kernel="rbfdot", kpar="automatic", C=5, cross=3, prob.model=TRUE)
```

**Accuracy, Precision & Recall**:

SVM model had an accuracy of around 58%.

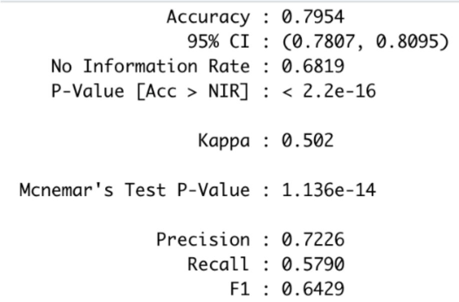


```
                Accuracy : 0.583
                  95% CI : (0.5654, 0.6005)
     No Information Rate : 0.3641
     P-Value [Acc > NIR] : < 2.2e-16

                   Kappa : 0.3697

  Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: Detractors Class: Passive Class: Promoters
Precision                       0.6590         0.4327           0.6142
Recall                          0.6402         0.3378           0.7471
F1                              0.6494         0.3794           0.6742
```

### 5.3.2 Two level classification

**Accuracy, Precision & Recall**:

SVM model had an accuracy of around 80%.

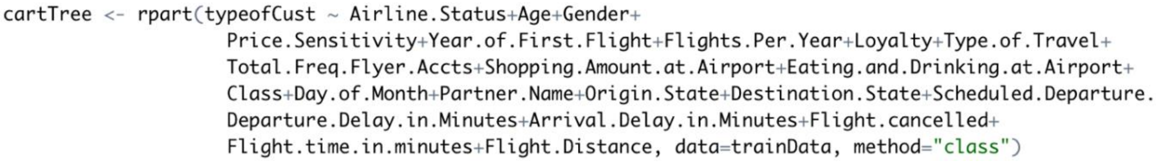


```
               Accuracy : 0.7954
                 95% CI : (0.7807, 0.8095)
    No Information Rate : 0.6819
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.502

 Mcnemar's Test P-Value : 1.136e-14

              Precision : 0.7226
                 Recall : 0.5790
                     F1 : 0.6429
```

## 5.4 Decision Trees

### 5.4.1 Three level classification

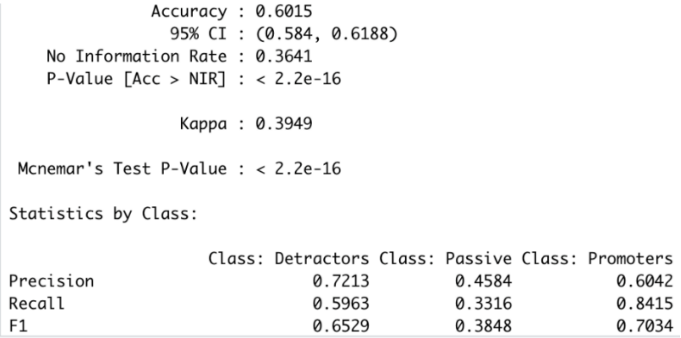
**Classification and regression trees:**



```
cartTree <- rpart(typeofCust ~ Airline.Status+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+Day.of.Month+Partner.Name+Origin.State+Destination.State+Scheduled.Departure.
                      Departure.Delay.in.Minutes+Arrival.Delay.in.Minutes+Flight.cancelled+
                      Flight.time.in.minutes+Flight.Distance, data=trainData, method="class")
```

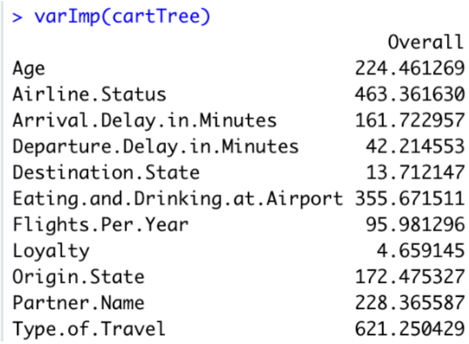
**Accuracy, Precision & Recall**:

Decision trees model had an accuracy of around 60%.



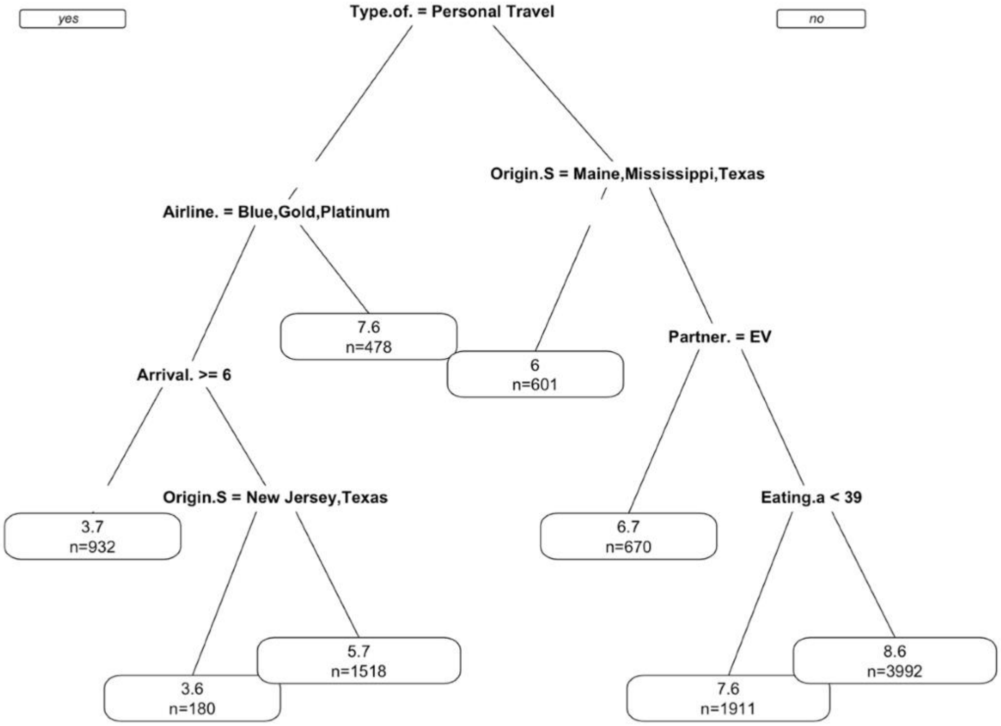
```
                Accuracy : 0.6015
                  95% CI : (0.584, 0.6188)
     No Information Rate : 0.3641
     P-Value [Acc > NIR] : < 2.2e-16

                   Kappa : 0.3949

  Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: Detractors Class: Passive Class: Promoters
Precision                       0.7213         0.4584           0.6042
Recall                          0.5963         0.3316           0.8415
F1                              0.6529         0.3848           0.7034
```

### Significant variables using varImp function:

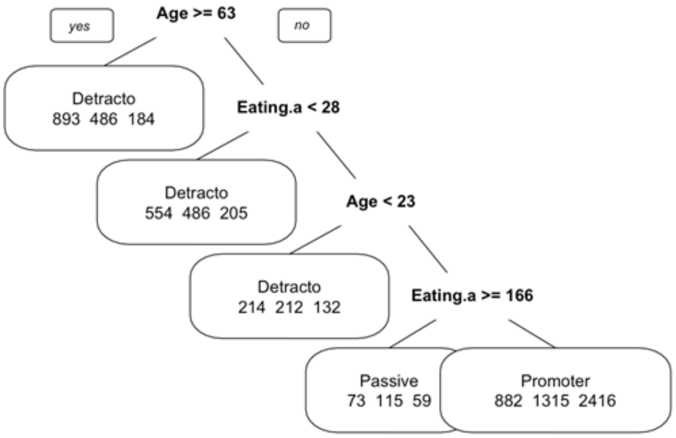


```
> varImp(cartTree)
                                  Overall
Age                            224.461269
Airline.Status                 463.361630
Arrival.Delay.in.Minutes       161.722957
Departure.Delay.in.Minutes      42.214553
Destination.State               13.712147
Eating.and.Drinking.at.Airport 355.671511
Flights.Per.Year                95.981296
Loyalty                          4.659145
Origin.State                   172.475327
Partner.Name                   228.365587
Type.of.Travel                 621.250429
```

### Decision tree model using all the variables:


yes
Type.of. = Personal Travel
no
Origin.S = Maine,Mississippi,Texas
Airline. = Blue,Gold,Platinum
7.6
n=478
6
n=601
Partner. = EV
Arrival. >= 6
Origin.S = New Jersey,Texas
Eating.a < 39
3.7
n=932
6.7
n=670
5.7
n=1518
8.6
n=3992
3.6
n=180
7.6
n=1911

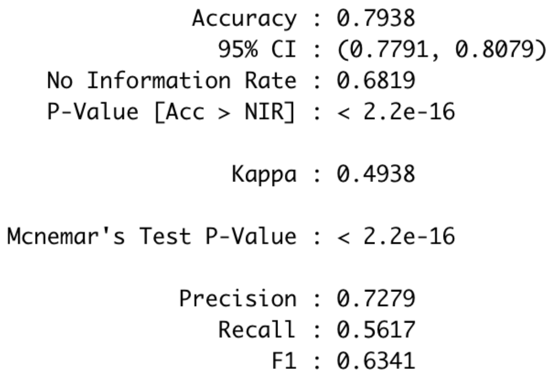
### Decision tree model predicting the target categories:



### 5.4.2 Two level classification

**Accuracy, Precision & Recall**:

Decision trees model had an accuracy of 82%.

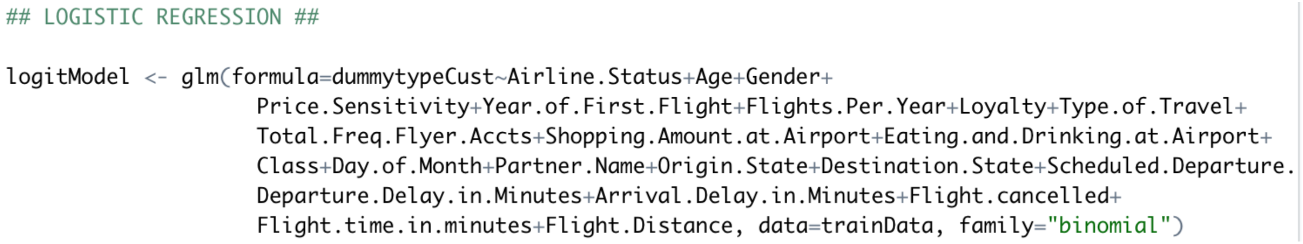


```
               Accuracy : 0.7938
                 95% CI : (0.7791, 0.8079)
    No Information Rate : 0.6819
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4938

 Mcnemar's Test P-Value : < 2.2e-16

              Precision : 0.7279
                 Recall : 0.5617
                     F1 : 0.6341
```

## 5.5 Logistic Regression

**Two level classification:**

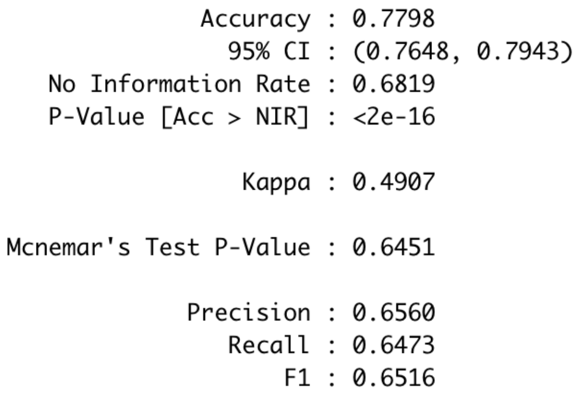
**Logistic regression model:**



```
## LOGISTIC REGRESSION ##

logitModel <- glm(formula=dummytypeCust~Airline.Status+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+Day.of.Month+Partner.Name+Origin.State+Destination.State+Scheduled.Departure.
                    Departure.Delay.in.Minutes+Arrival.Delay.in.Minutes+Flight.cancelled+
                    Flight.time.in.minutes+Flight.Distance, data=trainData, family="binomial")
```

**Accuracy, Precision & Recall**:

Logistic regression model had an accuracy of 78%.



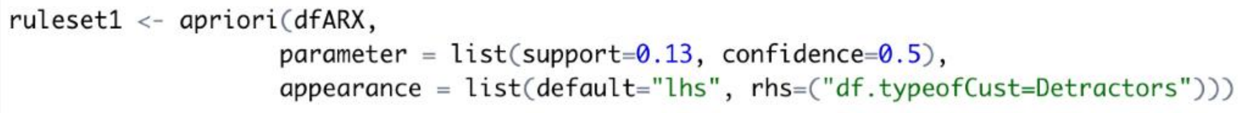
```
               Accuracy : 0.7798
                 95% CI : (0.7648, 0.7943)
    No Information Rate : 0.6819
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.4907

 Mcnemar's Test P-Value : 0.6451

              Precision : 0.6560
                 Recall : 0.6473
                     F1 : 0.6516
```

## 5.6 Association Rules Mining

We used apriori rules to predict rules for the detractors.

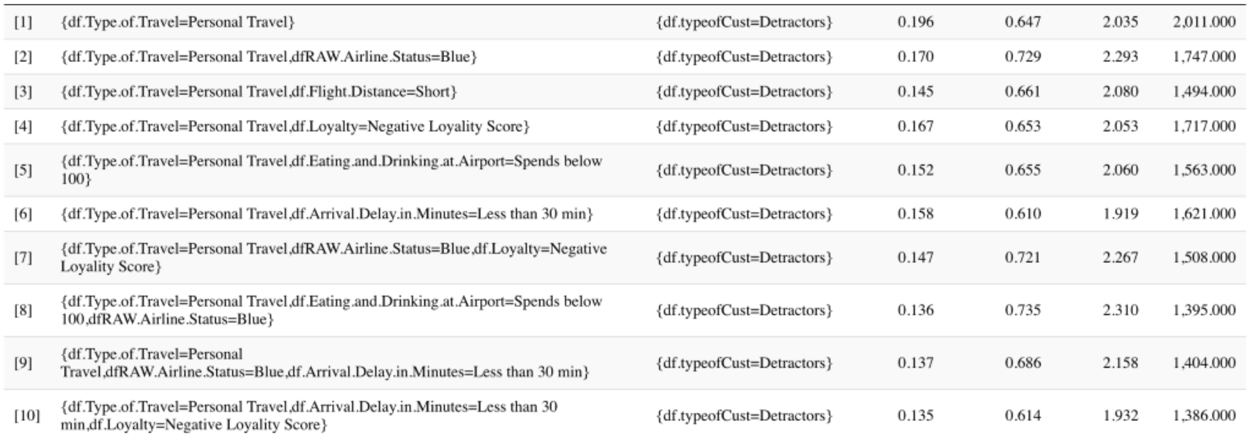
We only included the significant variables while implementing association rules.

## Association rules:

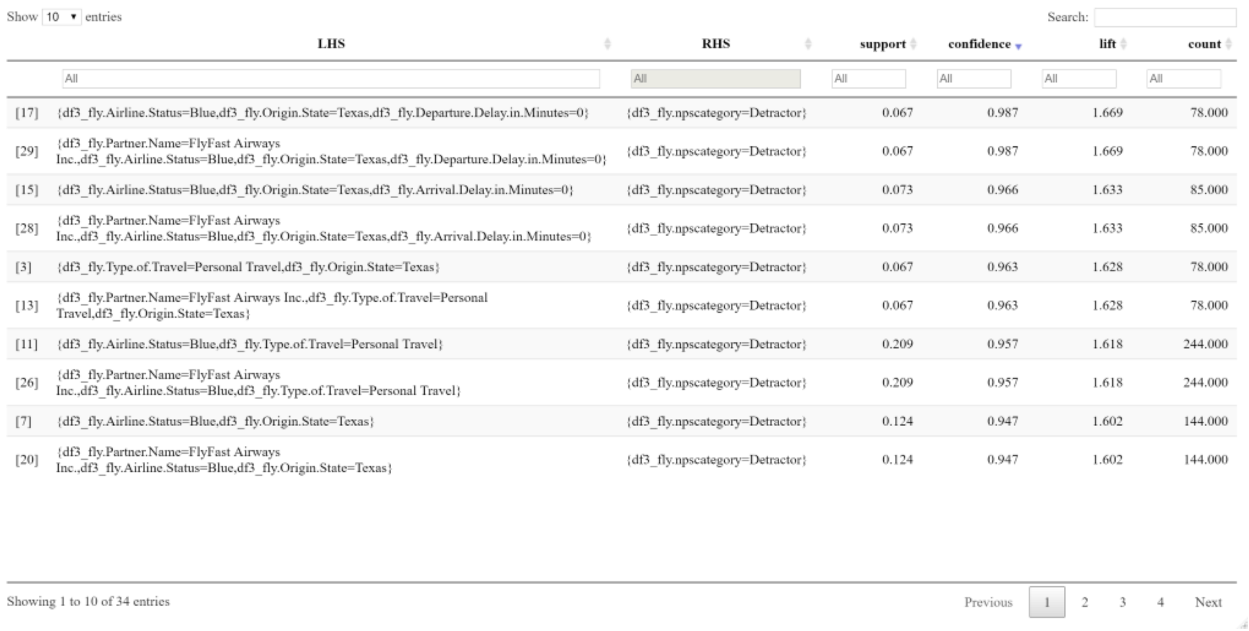


```
ruleset1 <- apriori(dfARX,
                    parameter = list(support=0.13, confidence=0.5),
                    appearance = list(default="lhs", rhs=("df.typeofCust=Detractors")))
```

## Ruleset for detractors:

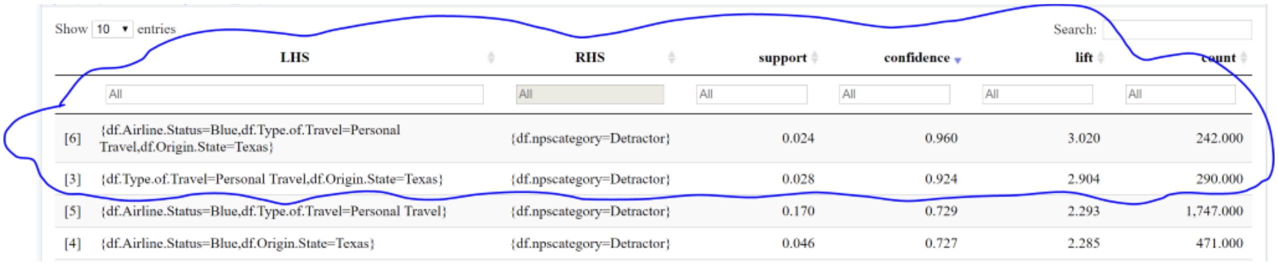


| | | | | | | |
|---|---|---|---|---|---|---|
| [1] | {df.Type.of.Travel=Personal Travel} | {df.typeofCust=Detractors} | 0.196 | 0.647 | 2.035 | 2,011.000 |
| [2] | {df.Type.of.Travel=Personal Travel,dfRAW.Airline.Status=Blue} | {df.typeofCust=Detractors} | 0.170 | 0.729 | 2.293 | 1,747.000 |
| [3] | {df.Type.of.Travel=Personal Travel,df.Flight.Distance=Short} | {df.typeofCust=Detractors} | 0.145 | 0.661 | 2.080 | 1,494.000 |
| [4] | {df.Type.of.Travel=Personal Travel,df.Loyalty=Negative Loyality Score} | {df.typeofCust=Detractors} | 0.167 | 0.653 | 2.053 | 1,717.000 |
| [5] | {df.Type.of.Travel=Personal Travel,df.Eating.and.Drinking.at.Airport=Spends below 100} | {df.typeofCust=Detractors} | 0.152 | 0.655 | 2.060 | 1,563.000 |
| [6] | {df.Type.of.Travel=Personal Travel,df.Arrival.Delay.in.Minutes=Less than 30 min} | {df.typeofCust=Detractors} | 0.158 | 0.610 | 1.919 | 1,621.000 |
| [7] | {df.Type.of.Travel=Personal Travel,dfRAW.Airline.Status=Blue,df.Loyalty=Negative Loyality Score} | {df.typeofCust=Detractors} | 0.147 | 0.721 | 2.267 | 1,508.000 |
| [8] | {df.Type.of.Travel=Personal Travel,df.Eating.and.Drinking.at.Airport=Spends below 100,dfRAW.Airline.Status=Blue} | {df.typeofCust=Detractors} | 0.136 | 0.735 | 2.310 | 1,395.000 |
| [9] | {df.Type.of.Travel=Personal Travel,dfRAW.Airline.Status=Blue,df.Arrival.Delay.in.Minutes=Less than 30 min} | {df.typeofCust=Detractors} | 0.137 | 0.686 | 2.158 | 1,404.000 |
| [10] | {df.Type.of.Travel=Personal Travel,df.Arrival.Delay.in.Minutes=Less than 30 min,df.Loyalty=Negative Loyality Score} | {df.typeofCust=Detractors} | 0.135 | 0.614 | 1.932 | 1,386.000 |

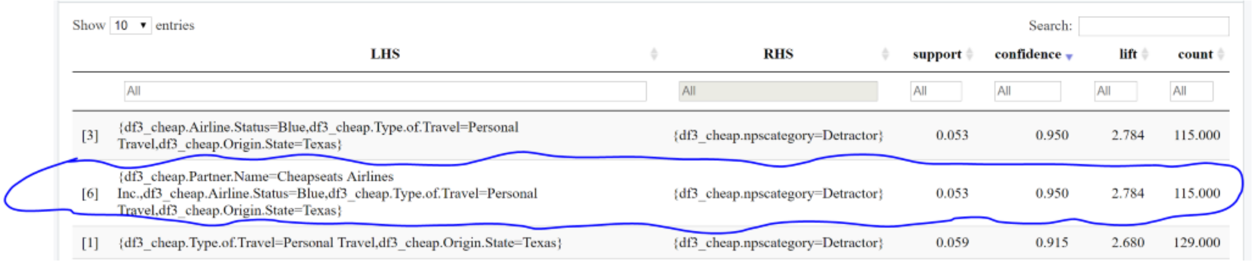


| | LHS | RHS | support | confidence | lift | count |
|---|---|---|---|---|---|---|
| | All | All | All | All | All | All |
| [17] | {df3_fly.Airline.Status=Blue,df3_fly.Origin.State=Texas,df3_fly.Departure.Delay.in.Minutes=0} | {df3_fly.npscategory=Detractor} | 0.067 | 0.987 | 1.669 | 78.000 |
| [29] | {df3_fly.Partner.Name=FlyFast Airways Inc.,df3_fly.Airline.Status=Blue,df3_fly.Origin.State=Texas,df3_fly.Departure.Delay.in.Minutes=0} | {df3_fly.npscategory=Detractor} | 0.067 | 0.987 | 1.669 | 78.000 |
| [15] | {df3_fly.Airline.Status=Blue,df3_fly.Origin.State=Texas,df3_fly.Arrival.Delay.in.Minutes=0} | {df3_fly.npscategory=Detractor} | 0.073 | 0.966 | 1.633 | 85.000 |
| [28] | {df3_fly.Partner.Name=FlyFast Airways Inc.,df3_fly.Airline.Status=Blue,df3_fly.Origin.State=Texas,df3_fly.Arrival.Delay.in.Minutes=0} | {df3_fly.npscategory=Detractor} | 0.073 | 0.966 | 1.633 | 85.000 |
| [3] | {df3_fly.Type.of.Travel=Personal Travel,df3_fly.Origin.State=Texas} | {df3_fly.npscategory=Detractor} | 0.067 | 0.963 | 1.628 | 78.000 |
| [13] | {df3_fly.Partner.Name=FlyFast Airways Inc.,df3_fly.Type.of.Travel=Personal Travel,df3_fly.Origin.State=Texas} | {df3_fly.npscategory=Detractor} | 0.067 | 0.963 | 1.628 | 78.000 |
| [11] | {df3_fly.Airline.Status=Blue,df3_fly.Type.of.Travel=Personal Travel} | {df3_fly.npscategory=Detractor} | 0.209 | 0.957 | 1.618 | 244.000 |
| [26] | {df3_fly.Partner.Name=FlyFast Airways Inc.,df3_fly.Airline.Status=Blue,df3_fly.Type.of.Travel=Personal Travel} | {df3_fly.npscategory=Detractor} | 0.209 | 0.957 | 1.618 | 244.000 |
| [7] | {df3_fly.Airline.Status=Blue,df3_fly.Origin.State=Texas} | {df3_fly.npscategory=Detractor} | 0.124 | 0.947 | 1.602 | 144.000 |
| [20] | {df3_fly.Partner.Name=FlyFast Airways Inc.,df3_fly.Airline.Status=Blue,df3_fly.Origin.State=Texas} | {df3_fly.npscategory=Detractor} | 0.124 | 0.947 | 1.602 | 144.000 |

Showing 1 to 10 of 34 entries



| | LHS | RHS | support | confidence | lift | count |
|---|---|---|---|---|---|---|
| | All | All | All | All | All | All |
| [6] | {df.Airline.Status=Blue,df.Type.of.Travel=Personal Travel,df.Origin.State=Texas} | {df.npscategory=Detractor} | 0.024 | 0.960 | 3.020 | 242.000 |
| [3] | {df.Type.of.Travel=Personal Travel,df.Origin.State=Texas} | {df.npscategory=Detractor} | 0.028 | 0.924 | 2.904 | 290.000 |
| [5] | {df.Airline.Status=Blue,df.Type.of.Travel=Personal Travel} | {df.npscategory=Detractor} | 0.170 | 0.729 | 2.293 | 1,747.000 |
| [4] | {df.Airline.Status=Blue,df.Origin.State=Texas} | {df.npscategory=Detractor} | 0.046 | 0.727 | 2.285 | 471.000 |

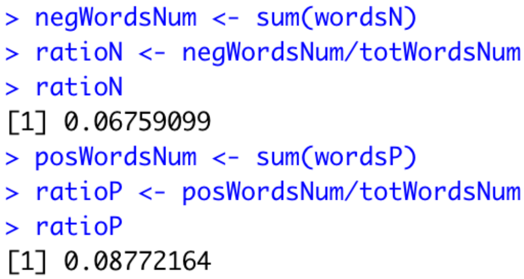


Show 10 entries Search:

| | LHS | RHS | support | confidence | lift | count |
|---|---|---|---|---|---|---|
| | All | All | All | All | All | All |
| [3] | {df3_cheap.Airline.Status=Blue,df3_cheap.Type.of.Travel=Personal Travel,df3_cheap.Origin.State=Texas} | {df3_cheap.npscategory=Detractor} | 0.053 | 0.950 | 2.784 | 115.000 |
| [6] | {df3_cheap.Partner.Name=Cheapseats Airlines Inc.,df3_cheap.Airline.Status=Blue,df3_cheap.Type.of.Travel=Personal Travel,df3_cheap.Origin.State=Texas} | {df3_cheap.npscategory=Detractor} | 0.053 | 0.950 | 2.784 | 115.000 |
| [1] | {df3_cheap.Type.of.Travel=Personal Travel,df3_cheap.Origin.State=Texas} | {df3_cheap.npscategory=Detractor} | 0.059 | 0.915 | 2.680 | 129.000 |

# 5.7 Text Mining

## 5.7.1 Sentiment Analysis

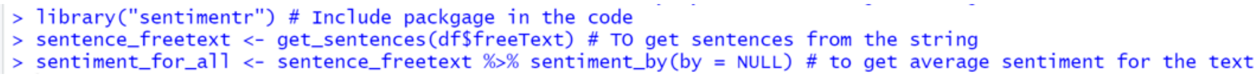


```
> negWordsNum <- sum(wordsN)
> ratioN <- negWordsNum/totWordsNum
> ratioN
[1] 0.06759099
> posWordsNum <- sum(wordsP)
> ratioP <- posWordsNum/totWordsNum
> ratioP
[1] 0.08772164
```

More positive words than negative words in the freeText.

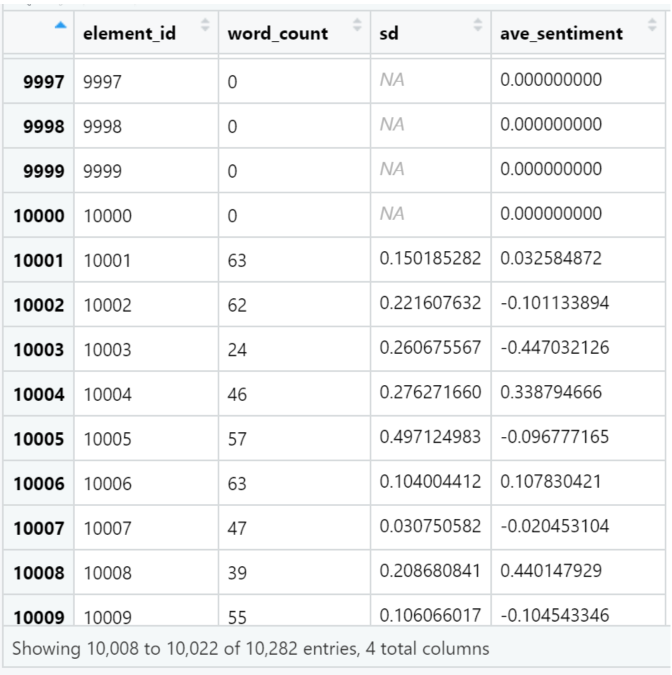
## 5.7.2 Individual Free Text Column Sentimental Analysis

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.



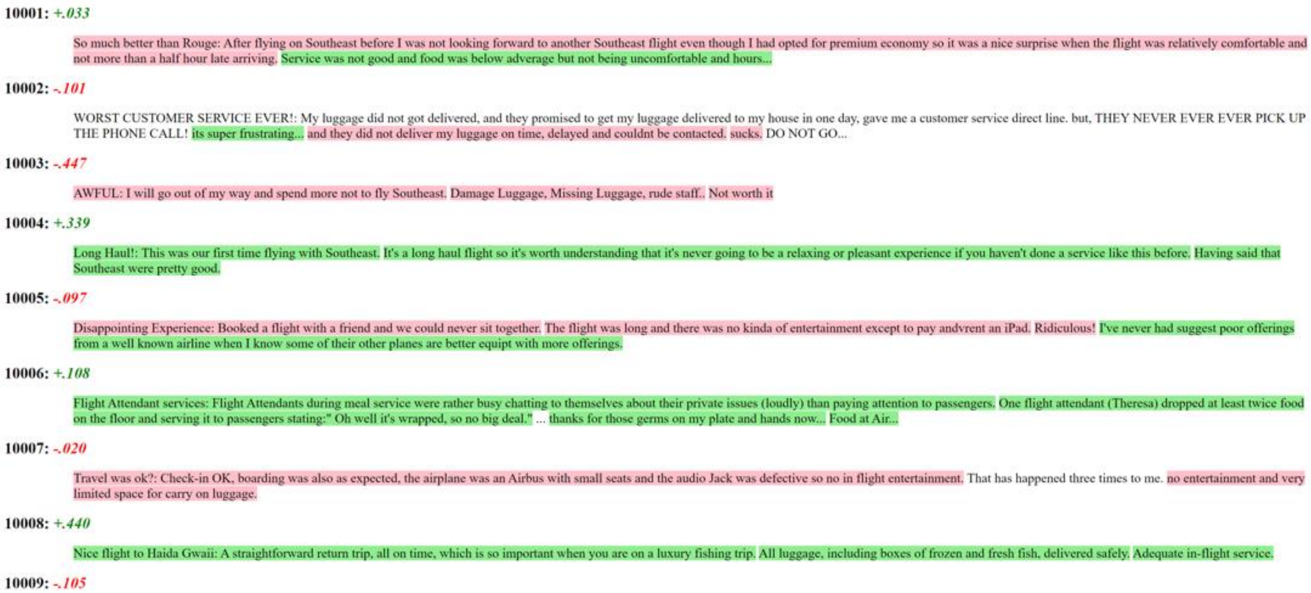
```
> library("sentimentr") # Include packgage in the code
> sentence_freetext <- get_sentences(df$freeText) # TO get sentences from the string
> sentiment_for_all <- sentence_freetext %>% sentiment_by(by = NULL) # to get average sentiment for the text
```

Examples:



| | element_id | word_count | sd | ave_sentiment |
|---|---|---|---|---|
| 9997 | 9997 | 0 | NA | 0.000000000 |
| 9998 | 9998 | 0 | NA | 0.000000000 |
| 9999 | 9999 | 0 | NA | 0.000000000 |
| 10000 | 10000 | 0 | NA | 0.000000000 |
| 10001 | 10001 | 63 | 0.150185282 | 0.032584872 |
| 10002 | 10002 | 62 | 0.221607632 | -0.101133894 |
| 10003 | 10003 | 24 | 0.260675567 | -0.447032126 |
| 10004 | 10004 | 46 | 0.276271660 | 0.338794666 |
| 10005 | 10005 | 57 | 0.497124983 | -0.096777165 |
| 10006 | 10006 | 63 | 0.104004412 | 0.107830421 |
| 10007 | 10007 | 47 | 0.030750582 | -0.020453104 |
| 10008 | 10008 | 39 | 0.208680841 | 0.440147929 |
| 10009 | 10009 | 55 | 0.106066017 | -0.104543346 |

Showing 10,008 to 10,022 of 10,282 entries, 4 total columns



**10001:** *+.033*

So much better than Rouge: After flying on Southeast before I was not looking forward to another Southeast flight even though I had opted for premium economy so it was a nice surprise when the flight was relatively comfortable and not more than a half hour late arriving. Service was not good and food was below adverage but not being uncomfortable and hours...

**10002:** *-.101*

WORST CUSTOMER SERVICE EVER!: My luggage did not got delivered, and they promised to get my luggage delivered to my house in one day, gave me a customer service direct line. but, THEY NEVER EVER EVER PICK UP THE PHONE CALL! its super frustrating... and they did not deliver my luggage on time, delayed and couldnt be contacted. sucks. DO NOT GO...

**10003:** *-.447*

AWFUL: I will go out of my way and spend more not to fly Southeast. Damage Luggage, Missing Luggage, rude staff.. Not worth it

**10004:** *+.339*

Long Haul!: This was our first time flying with Southeast. It's a long haul flight so it's worth understanding that it's never going to be a relaxing or pleasant experience if you haven't done a service like this before. Having said that Southeast were pretty good.

**10005:** *-.097*

Disappointing Experience: Booked a flight with a friend and we could never sit together. The flight was long and there was no kinda of entertainment except to pay andvrent an iPad. Ridiculous! I've never had suggest poor offerings from a well known airline when I know some of their other planes are better equipt with more offerings.

**10006:** *+.108*

Flight Attendant services: Flight Attendants during meal service were rather busy chatting to themselves about their private issues (loudly) than paying attention to passengers. One flight attendant (Theresa) dropped at least twice food on the floor and serving it to passengers stating:" Oh well it's wrapped, so no big deal." ... thanks for those germs on my plate and hands now... Food at Air...

**10007:** *-.020*

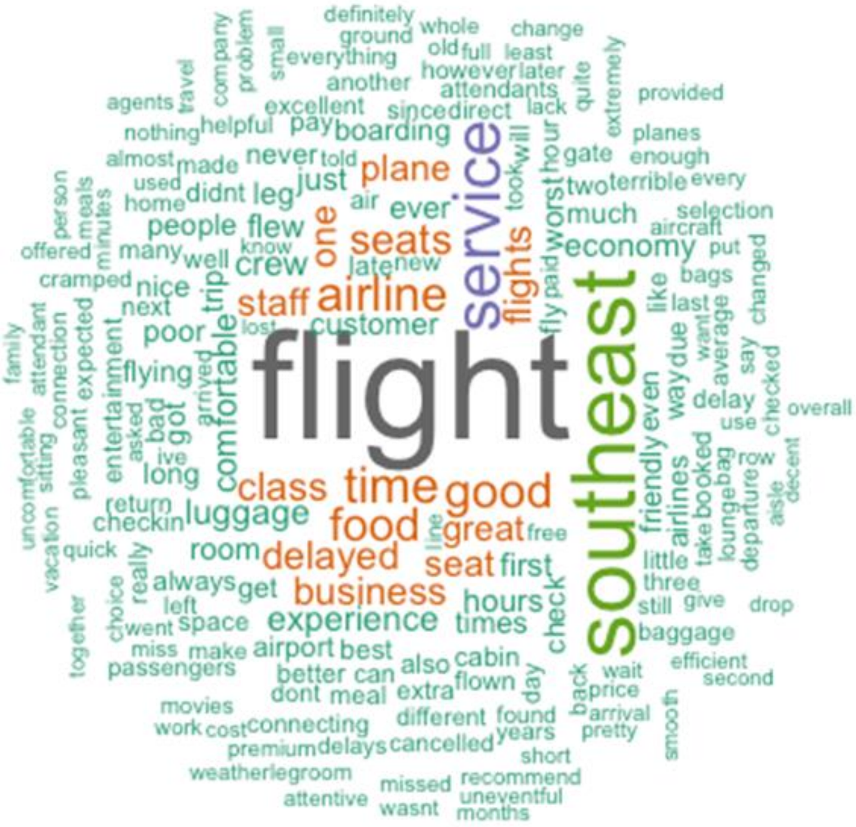
Travel was ok?: Check-in OK, boarding was also as expected, the airplane was an Airbus with small seats and the audio Jack was defective so no in flight entertainment. That has happened three times to me. no entertainment and very limited space for carry on luggage.

**10008:** *+.440*

Nice flight to Haida Gwaii: A straightforward return trip, all on time, which is so important when you are on a luxury fishing trip. All luggage, including boxes of frozen and fresh fish, delivered safely. Adequate in-flight service.

**10009:** *-.105*

### 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