Customer Churn Analysis

IST 687 Final Project

Instructor: Erik Anderson

Group Member:

Yi Shao

Huaiyu Shi

Tzuyang Huang

Brandon Reyes

Wesley Knights

Content

I. Introduction

Southeast Airlines needed to lower customer churn and their loyalty program might not be sufficient in keeping low customer churn. Additionally, customer churn is a lagging indicator actually. Therefore, to reduce the churn and keep a customer, we do the analysis to find out the indicators and factors that would impact customer's choice. In our project, we do some analysis to explore the data, find out association rules and factors which relate to detractors and train different models to predict if a passenger is a detractor.

II. Dataset Overview

We draw the histogram or boxplot graph for each numerical variable. For each graph, we checked if the column has "N/A" data and eliminated them.

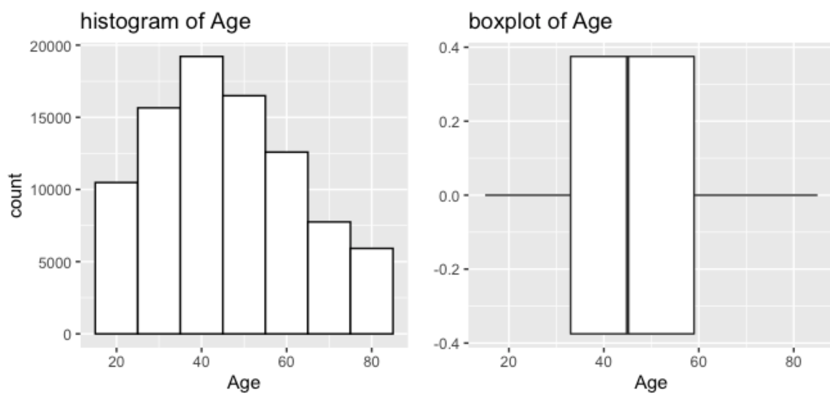


Figure 1

First we create the graph (figure 1) of Age and set the bin width as 10. We can see the people between 35 to 45 years old are the most. The median age is around 42 or 43 years old.

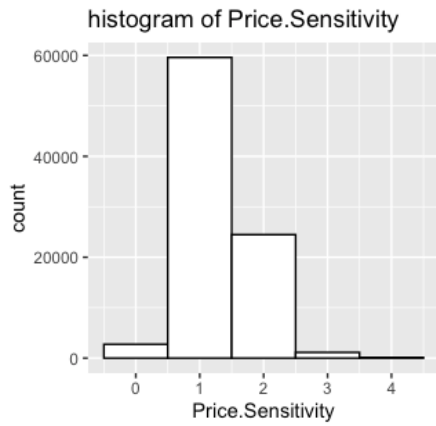


Figure 2

The range of price sensitivity between 0 to 5 which shows how people are affected by the price. In figure 2, there are almost 60,000 customers who have grade 1 price sensitivity, and up to 20,000 customers have grade 2. We can conclude that change of price doesn't affect customers purchasing much.

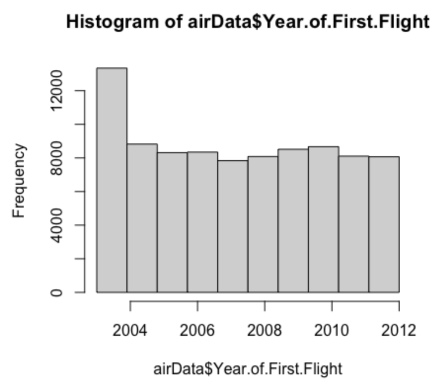


Figure 3

Figure 3 is a histogram of the year each customer first took the flight. The range is between 2003 to 2012. Except for more than 12,000 customers who started their first flight with Southeast Airlines in 2003, there are around 8,000 new customers every year which is quite stable.

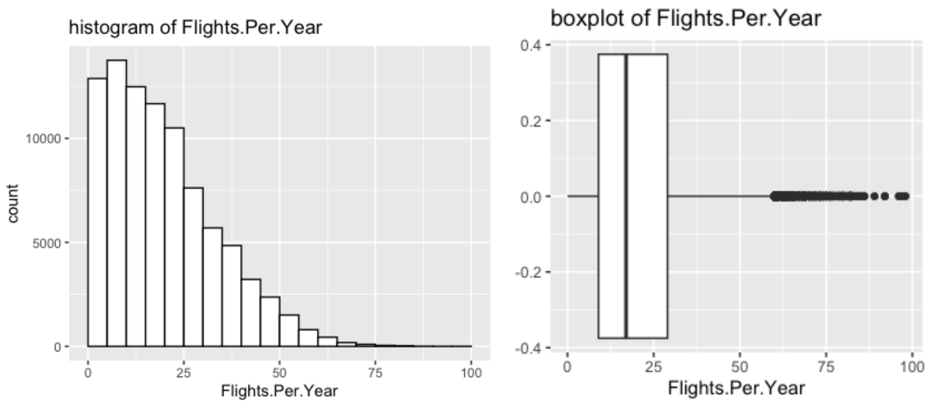


Figure 4

This graph shows how many flights each customer took in the most recent 12 months. The histogram is right-skewed and the median value is less than 25.

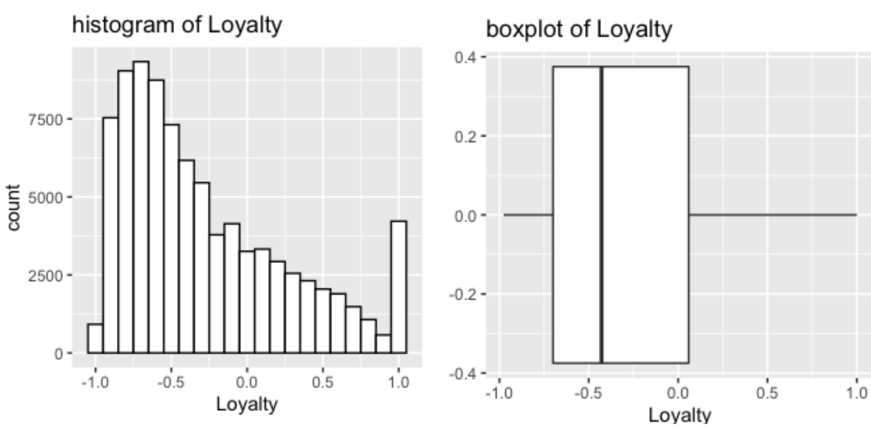


Figure 5

Through the histogram and boxplot (figure 5) of Loyalty, we can see the median value is negative, most people don’t care which airline they would take. Southeast Airlines should think about how to distinguish itself from other airlines.

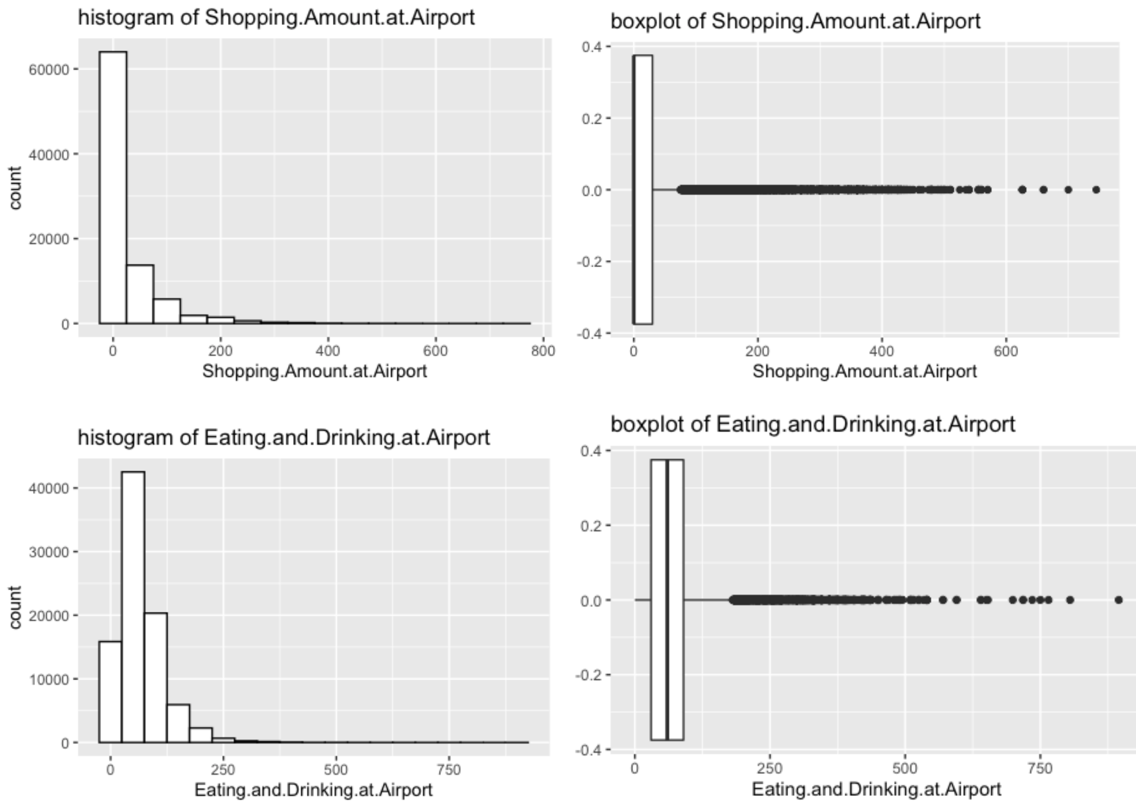


Figure 6

Compare the graph （figure 6） of the amount of shopping and amount of eat and drink, half customers don't spend money on shopping but over 75% customers would spend money on food and drink.

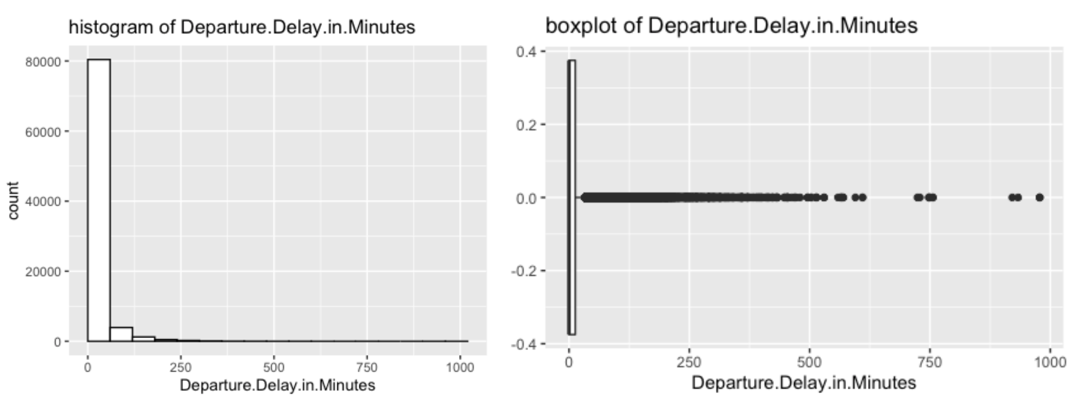
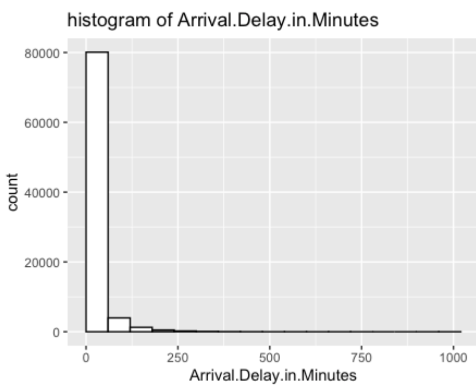


Figure 7

```
ifDepartureDelay
    0     1
57000 29493
```

Figure 8

Southeast Airlines performs well on the departure delay. Over half flights don't have delays (figure 7). When I set the delay minute to 5 mins (figure 8), there are 57000 flights' departure delay less than 5 mins.




```
ifArrivalDelay
    0     1
55912 30350
```

Figure 9

The histogram of arrival delay is similar to histogram of departure delay. When I set the delay minute to 5 mins (figure 9), there are 55912 flights' departure delay less than 5 mins.

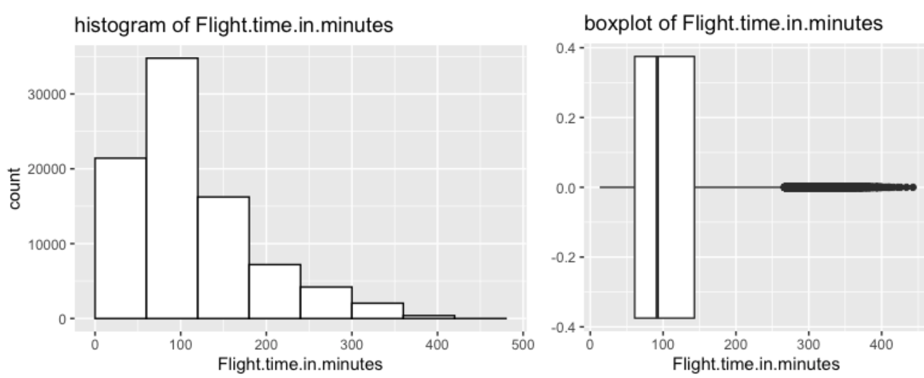


Figure 10

We set the bin width as 60, and center as 30 in histogram. We can see in figure 10, over 20,000 customers' flight time less than 1 hour, up to 35,000 customers' flight time between 1 to 2 hours, over 15,000 customers' flight time between 2 to 3 hours. Boxplot shows the median is less than 100 mins.

```
   Female      Male
0.564563 0.435437
```

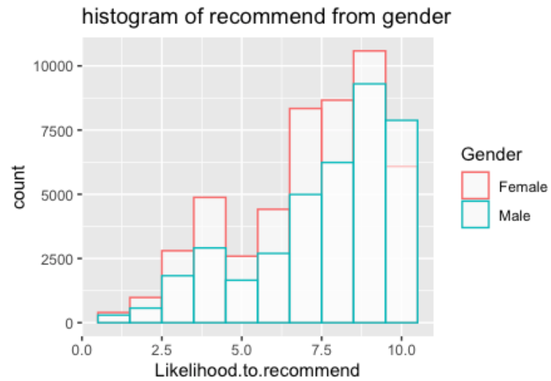


Figure 11

When we separate the customers from gender, there are 56.46% female and 43.54% males. Compare the difference between the recommendations from gender, the percentage of male who give grade 10 is higher than females (figure 11). Then we calculate the percentage of male and female who are promoters, there are 44.79% males, 33.52% female are promoters. This shows that female’s evaluation of Southeast Airlines services is more stringent.

```
Business travel Mileage tickets Personal Travel
     0.61251986      0.07962543      0.30785471
```

Figure 12

Figure 12 shows the percentage of different travel types and the distribution of recommendation from each type. 61.25% customers are business travelers, 30.79% are personal travelers and the rest are defined as mileage tickets. 53.64% customers from business travel are promoters, 41.08% customers who buy the mileage tickets are promoters, but only 7.46% customers from personal travel are promoters. Many personal travel customers didn't give good scores, most of them are detractors.

```
  Business        Eco   Eco Plus 
0.08169126 0.81441544 0.10389330 
```

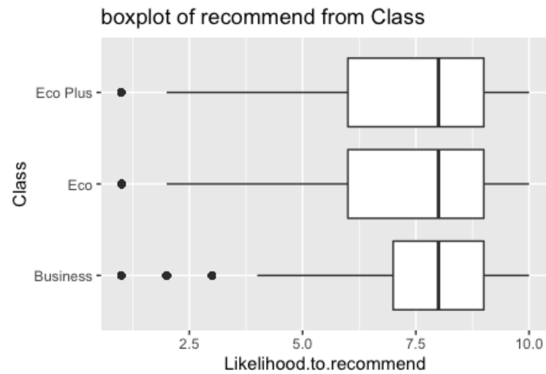


Figure 13

We separate the customers from different classes in figure 13. Most customers are in Eco class. The median grade of recommendation from these three boxplots are quite the same, but the percentage of promoters from Business class is higher than the Eco and Eco Plus.

```
      Blue       Gold   Platinum     Silver 
0.68276958 0.08402951 0.03332577 0.19987514 
```

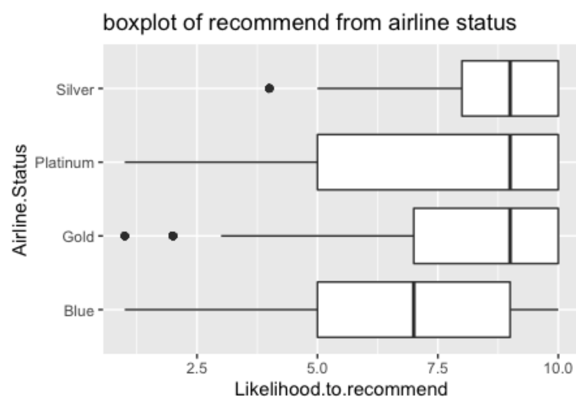


Figure 14

Figure 14 shows the structure of Airline Status. 68.27% customers have Blue status, 8.4% have Gold status, 3.33% have Platinum status and 19.99% have Silver status. The median score from Silver, Platinum and Gold around 8.5, but the median score of Blue is lower than 7.5, which means the half clients from Blue Status are detractors. Southeast Airlines should improve the service for customers who have Blue status since they are the majority group with lower satisfaction.

service	time	good	food	seats	class	delayed	business	staff
113	75	70	66	56	50	49	47	45
one	seat	us	experience	great	crew	luggage	comfortable	first
43	43	40	40	40	38	38	37	32
economy	flew	just	ever	hours	worst	customer	room	trip
32	32	32	30	30	30	29	29	29

Figure 15

We collected all the text from customers and made a word cloud. We add some stopwords like "airline", "southeast", which don't make any sense upon the stopword "english". Then we sort the word from the most frequency to lower. "Service" and "time" are the most frequent words. The word "food", "seat", "class" and "delayed" have also been mentioned many times.

	Detractor	Passive	Promoter
AA	0.0229522339	0.0164366146	0.0125204323
AS	0.0123955685	0.0091491101	0.0072875045
B6	0.0163344533	0.0104431529	0.0082750636
DL	0.0693561569	0.0484017436	0.0374704867
EV	0.0548492554	0.0402401925	0.0190474028
F9	0.0072193970	0.0040410461	0.0033032147
FL	0.0095464039	0.0079345260	0.0056075191
HA	0.0004540501	0.0004086451	0.0004540501
MQ	0.0218057574	0.0151766255	0.0115328732
OO	0.0549060116	0.0382764257	0.0282759717
OU	0.0411709953	0.0288889393	0.0204663095
US	0.0383331820	0.0289797494	0.0208522521
VX	0.0053123865	0.0035529422	0.0027810570
WN	0.0991418453	0.0685956230	0.0478228296

Figure 16

We added a new row "ifPromoter" to check if the customer is a Promoter or a Detractor. We created a table of promoters percentage based on the partner. WN has the highest detractor percentage, then is DL, OO and EV.

III. Linear Model

Converting Data and Cleaning For Predicting Model:

1. Gender: Male = 1, Female = 0
2. Flight.cancelled: Yes = 1, No = 0
3. Class: Business =3, Eco Plus = 2, Eco = 1
4. Deleting all the rows with one or more NULL.

➢ Find out the correlation coefficient between variables and likelihood.to.recommend

	col_name..16.	corr
1	Age	-0.212187433
2	Gender	0.103396448
3	Price.Sensitivity	-0.091483097
4	Flights.Per.Year	-0.237506159
5	Loyalty	0.165212570
6	Total.Freq.Flyer.Accts	0.084016500
7	Shopping.Amount.at.Airport	0.029890072
8	Eating.and.Drinking.at.Airport	0.075899189
9	Class	0.034252052
10	Scheduled.Departure.Hour	-0.009567695
11	Departure.Delay.in.Minutes	-0.089405751
12	Arrival.Delay.in.Minutes	-0.097314943
13	Flight.cancelled	NA
14	Flight.time.in.minutes	0.011207150
15	Flight.Distance	0.016132431

Figure 17

➔ Age, Gender, Loyalty, and Flights.Per.Year are more relative to likelihood.to.recommend. However, the relations are not strong.

➢ Linear model 1:

- Variables:

Age	V	Class	V
Gender	V	Scheduled.Departure.Hour	V
Price.Sensitivity	V	Departure.Delay.in.Minutes	V
Flights.Per.Year	V	Arrival.Delay.in.Minutes	V
Loyalty	V	Flight.cancelled	V
Total.Freq.Flyer.Accts	V	Flight.time.in.minutes	V
Shopping.Amount.at.Airport	V	Flight.Distance	V
Eating.and.Drinking.at.Airport	V		

- Prediction: Likelihood.to.recommend

```
Call:
lm(formula = Likelihood.to.recommend ~ ., data = df1[train_index,
    ])

Residuals:
    Min      1Q  Median      3Q     Max
-7.9699 -1.2853  0.4707  1.5942  6.5961

Coefficients: (1 not defined because of singularities)
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  9.227e+00  6.027e-02 153.096  < 2e-16 ***
Age                         -2.464e-02  6.306e-04 -39.075  < 2e-16 ***
Gender                       4.886e-01  1.932e-02  25.291  < 2e-16 ***
Price.Sensitivity           -4.107e-01  1.744e-02 -23.552  < 2e-16 ***
Flights.Per.Year            -3.327e-02  9.557e-04 -34.814  < 2e-16 ***
Loyalty                     -8.341e-02  2.643e-02  -3.155  0.00160 **
Total.Freq.Flyer.Accts      -7.143e-02  9.645e-03  -7.406 1.32e-13 ***
Shopping.Amount.at.Airport   1.127e-03  1.785e-04   6.316 2.71e-10 ***
Eating.and.Drinking.at.Airport  2.773e-03  1.846e-04  15.025  < 2e-16 ***
Class                        8.814e-02  1.569e-02   5.616 1.96e-08 ***
Scheduled.Departure.Hour    -2.206e-03  2.047e-03  -1.078  0.28107
Departure.Delay.in.Minutes   3.885e-03  9.882e-04   3.931 8.47e-05 ***
Arrival.Delay.in.Minutes    -9.096e-03  9.741e-04  -9.337  < 2e-16 ***
Flight.cancelled                    NA         NA      NA       NA
Flight.time.in.minutes      -1.418e-03  6.368e-04  -2.227  0.02597 *
Flight.Distance              2.292e-04  7.705e-05   2.974  0.00294 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.111 on 49985 degrees of freedom
Multiple R-squared:  0.1212,    Adjusted R-squared:  0.1209
F-statistic: 492.2 on 14 and 49985 DF,  p-value: < 2.2e-16
```

Figure 18

➔ Adjusted R-squared is only 0.1209, which means that the combination of those variables are not relative to likelihood.to.recommend.

➔ The accuracy is 16.35%. This model is not good at predicting the actual rate

➢ Linear model 2:

- Variables:

Age	V	Class	
Gender	V	Scheduled.Departure.Hour	
Price.Sensitivity	V	Departure.Delay.in.Minutes	
Flights.Per.Year	V	Arrival.Delay.in.Minutes	
Loyalty	V	Flight.cancelled	
Total.Freq.Flyer.Accts		Flight.time.in.minutes	V
Shopping.Amount.at.Airport		Flight.Distance	V
Eating.and.Drinking.at.Airport			

- Prediction: Likelihood.to.recommend

```
Call:
lm(formula = Likelihood.to.recommend ~ Age + Gender + Flights.Per.Year +
    Loyalty + Flight.time.in.minutes + Flight.Distance, data = df1[train_index,
    ])

Residuals:
    Min      1Q  Median      3Q     Max
-7.4419 -1.3110  0.4949  1.6333  5.3612

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             8.718e+00  3.570e-02 244.242  < 2e-16 ***
Age                    -2.029e-02  5.822e-04 -34.858  < 2e-16 ***
Gender                  5.075e-01  1.942e-02  26.135  < 2e-16 ***
Flights.Per.Year       -3.587e-02  9.511e-04 -37.713  < 2e-16 ***
Loyalty                -1.147e-01  2.554e-02  -4.490 7.13e-06 ***
Flight.time.in.minutes -3.033e-03  6.174e-04  -4.912 9.03e-07 ***
Flight.Distance         4.218e-04  7.473e-05   5.644 1.67e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.142 on 49993 degrees of freedom
Multiple R-squared:  0.09504,	Adjusted R-squared:  0.09493
F-statistic: 875.1 on 6 and 49993 DF,  p-value: < 2.2e-16
```

Figure 19

➔ Adjusted R-squared is only 0.09493, which means that the combination of those variables are not relative to likelihood.to.recommend.

➔ The accuracy is 16.13%. This model is not good at predicting the actual rate.

- Different Combinations of model1:

```
> ols_step_all_possible(model1)
   Index N                                    Predictors    R-Square Adj. R-Square
4      1 1                                 Flights.Per.Year 0.0555014913  0.0554826006
1      2 1                                              Age 0.0455938205  0.0455747316
5      3 1                                          Loyalty 0.0263129187  0.0262934442
2      4 1                                           Gender 0.0103007773  0.0102809825
12     5 1                         Arrival.Delay.in.Minutes 0.0098925983  0.0098727953
11     6 1                       Departure.Delay.in.Minutes 0.0087480181  0.0087281923
3      7 1                                Price.Sensitivity 0.0080190595  0.0079992191
6      8 1                            Total.Freq.Flyer.Accts 0.0067502049  0.0067303391
8      9 1                    Eating.and.Drinking.at.Airport 0.0061888648  0.0061689878
9     10 1                                            Class 0.0012946667  0.0012746918
7     11 1                        Shopping.Amount.at.Airport 0.0008917900  0.0008718070
15    12 1                                  Flight.Distance 0.0002848633  0.0002648682
14    13 1                            Flight.time.in.minutes 0.0001455982  0.0001256003
10    14 1                          Scheduled.Departure.Hour 0.0001146761  0.0000946776
13    15 1                                 Flight.cancelled 0.0000000000  0.0000000000
18    16 2                             Age Flights.Per.Year 0.0806795797  0.0806428047
31    17 2                          Gender Flights.Per.Year 0.0694129833  0.0693757575
62    18 2        Flights.Per.Year Arrival.Delay.in.Minutes 0.0653687559  0.0653313684
61    19 2      Flights.Per.Year Departure.Delay.in.Minutes 0.0641412597  0.0641038231
43    20 2               Price.Sensitivity Flights.Per.Year 0.0629574642  0.0629199802
```

Figure 20

➔ For most combinations, the adjusted R-squares are low

➔ The accuracy is around 16% at most.

★ It’s hard to build up a linear model to predict the likelihood to recommend.

IV. SVM Model

➢ SVM 1:

- Variables:

Age	V	Class	V
Gender	V	Scheduled.Departure.Hour	V
Price.Sensitivity	V	Departure.Delay.in.Minutes	V
Flights.Per.Year	V	Arrival.Delay.in.Minutes	V
Loyalty	V	Flight.cancelled	
Total.Freq.Flyer.Accts	V	Flight.time.in.minutes	V
Shopping.Amount.at.Airport	V	Flight.Distance	V
Eating.and.Drinking.at.Airport	V		

- Prediction: NPS (Promoter, Passive, Detractor)

```
> svm1
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 5

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.0566365191431419

Number of Support Vectors : 40304

Objective Function Value : -108679.1 -79507.03 -112729.5
Training error : 0.38856
Cross validation error : 0.4369
Probability model included.
```

Figure 21

➔ The training error is 0.38856. The accuracy of testing data is 56.31%

- SVM 2:
  - Variables:

Age		Class	V
Gender		Scheduled.Departure.Hour	
Price.Sensitivity		Departure.Delay.in.Minutes	V
Flights.Per.Year		Arrival.Delay.in.Minutes	V
Loyalty	V	Flight.cancelled	
Total.Freq.Flyer.Accts		Flight.time.in.minutes	
Shopping.Amount.at.Airport		Flight.Distance	
Eating.and.Drinking.at.Airport			

  - Prediction: NPS (Promoter, Passive, Detractor)

```
> svm2
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 5

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  2.44793714874362

Number of Support Vectors : 46554

Objective Function Value : -120346.8 -120458 -149126.5
Training error : 0.53882
Cross validation error : 0.56142
Probability model included.
```

Figure 22

→ The training error is 0.53882. The accuracy of testing data is 44.17%

➢ SVM 3:

- Variables:

Age		Class	
Gender		Scheduled.Departure.Hour	
Price.Sensitivity		Departure.Delay.in.Minutes	
Flights.Per.Year	V	Arrival.Delay.in.Minutes	
Loyalty		Flight.cancelled	
Total.Freq.Flyer.Accts	V	Flight.time.in.minutes	
Shopping.Amount.at.Airport		Flight.Distance	V
Eating.and.Drinking.at.Airport	V		

- Prediction: NPS (Promoter, Passive, Detractor)

```
> svm3
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 5

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.22176478315747

Number of Support Vectors : 43092

Objective Function Value : -126164.1 -105057 -129843.6
Training error : 0.49228
Cross validation error : 0.5084
Probability model included.
```

Figure 23

➔ The training error is 0.49228. The accuracy of testing data is 48.78%

Male & Female

➢ SVM_Male: Predicting NPS of male passengers

- Variables:

Age	V	Class	V
Gender		Scheduled.Departure.Hour	V
Price.Sensitivity	V	Departure.Delay.in.Minutes	V
Flights.Per.Year	V	Arrival.Delay.in.Minutes	V
Loyalty	V	Flight.cancelled	
Total.Freq.Flyer.Accts	V	Flight.time.in.minutes	V
Shopping.Amount.at.Airport	V	Flight.Distance	V
Eating.and.Drinking.at.Airport	V		

- Prediction: NPS (Promoter, Passive, Detractor)

```
> svm_male
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 3

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.0678586847913852

Number of Support Vectors : 16689

Objective Function Value : -25249.35 -20193.94 -28197.07
Training error : 0.356159
Cross validation error : 0.407071
Probability model included.
```

Figure 24

➔ The training error is 0.356159. The accuracy of testing data is 59.26%

➢ SVM_Female: Predicting NPS of female passengers

- Variables:

Age	V	Class	V
Gender		Scheduled.Departure.Hour	V
Price.Sensitivity	V	Departure.Delay.in.Minutes	V
Flights.Per.Year	V	Arrival.Delay.in.Minutes	V
Loyalty	V	Flight.cancelled	
Total.Freq.Flyer.Accts	V	Flight.time.in.minutes	V
Shopping.Amount.at.Airport	V	Flight.Distance	V
Eating.and.Drinking.at.Airport	V		

- Prediction: NPS (Promoter, Passive, Detractor)

```
> svm_female
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 3

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.0658240860754297

Number of Support Vectors : 23975

Objective Function Value : -39850.61 -27686.68 -39444.86
Training error : 0.409224
Cross validation error : 0.468376
Probability model included.
```

Figure 25

➔ The training error is 0.409224. The accuracy of testing data is 53.92%

Business Travel vs. Personal Travel vs. Mileage Ticket

➢ SVM_Business: Predicting NPS of business travel passengers

- Variables:

Age	V	Class	V
Gender	V	Scheduled.Departure.Hour	V
Price.Sensitivity	V	Departure.Delay.in.Minutes	V
Flights.Per.Year	V	Arrival.Delay.in.Minutes	V
Loyalty	V	Flight.cancelled	
Total.Freq.Flyer.Accts	V	Flight.time.in.minutes	V
Shopping.Amount.at.Airport	V	Flight.Distance	V
Eating.and.Drinking.at.Airport	V		

- Prediction: NPS (Promoter, Passive, Detractor)

```
> svm_business
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 3

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.0565886801216093

Number of Support Vectors : 27626

Objective Function Value : -29969.16 -27295.88 -57193.46
Training error : 0.35188
Cross validation error : 0.383571
Probability model included.
```

Figure 26

➔ The training error is 0.35188. The accuracy of testing data is 61.35%

➢ SVM_Personal: Predicting NPS of personal travel passengers

- Variables:

Age	V	Class	V
Gender	V	Scheduled.Departure.Hour	V
Price.Sensitivity	V	Departure.Delay.in.Minutes	V
Flights.Per.Year	V	Arrival.Delay.in.Minutes	V
Loyalty	V	Flight.cancelled	
Total.Freq.Flyer.Accts	V	Flight.time.in.minutes	V
Shopping.Amount.at.Airport	V	Flight.Distance	V
Eating.and.Drinking.at.Airport	V		

- Prediction: NPS (Promoter, Passive, Detractor)

```
> svm_personal
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 3

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.0593890978320079

Number of Support Vectors : 12893

Objective Function Value : -30318.93 -7753.941 -7654.735
Training error : 0.320959
Cross validation error : 0.350962
Probability model included.
```

Figure 27

➔ The training error is 0.320959. The accuracy of testing data is 64.26%

➢ SVM_Mileage: Predicting NPS of mileage ticket passengers

- Variables:

Age	V	Class	V
Gender	V	Scheduled.Departure.Hour	V
Price.Sensitivity	V	Departure.Delay.in.Minutes	V
Flights.Per.Year	V	Arrival.Delay.in.Minutes	V
Loyalty	V	Flight.cancelled	
Total.Freq.Flyer.Accts	V	Flight.time.in.minutes	V
Shopping.Amount.at.Airport	V	Flight.Distance	V
Eating.and.Drinking.at.Airport	V		

- Prediction: NPS (Promoter, Passive, Detractor)

```
> svm_mileage
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 3

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.0563804976155871

Number of Support Vectors : 4164

Objective Function Value : -4376.255 -3197.735 -7400.924
Training error : 0.341984
Cross validation error : 0.439095
Probability model included.
```

Figure 28

➔ The training error is 0.341984. The accuracy of testing data is 55.78%

Predicting Detractor

➢ SVM_Detractor: Predicting if the passenger is detractor

- Variables:

Age	V	Class	V
Gender	V	Scheduled.Departure.Hour	V
Price.Sensitivity	V	Departure.Delay.in.Minutes	V
Flights.Per.Year	V	Arrival.Delay.in.Minutes	V
Loyalty	V	Flight.cancelled	
Total.Freq.Flyer.Accts	V	Flight.time.in.minutes	V
Shopping.Amount.at.Airport	V	Flight.Distance	V
Eating.and.Drinking.at.Airport	V		

- Prediction: Detractor (detractor, nondetractor)

```
> svm_detractor
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 5

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.0566025024336909

Number of Support Vectors : 32638

Objective Function Value : -147983.9
Training error : 0.221232
Cross validation error : 0.239616
Probability model included.
```

Figure 29

➔ The training error is 0.221232. The accuracy of testing data is 76.13%

V. Association Rules

We set support to 0.1 and confidence to 0.7, digging 12 rules

```
     lhs                                  rhs                                    support confidence  coverage     lift count
[1]  {Airline.Status=Blue,
      Type.of.Travel=Personal Travel} => {Likelihood.to.recommend=Detractor} 0.1701816  0.7149738 0.2380250 2.421730 14993
[2]  {Airline.Status=Blue,
      Gender=Female,
      Type.of.Travel=Personal Travel} => {Likelihood.to.recommend=Detractor} 0.1141544  0.7237334 0.1577299 2.451400 10057
[3]  {Airline.Status=Blue,
      Type.of.Travel=Personal Travel,
      Class=Eco}                      => {Likelihood.to.recommend=Detractor} 0.1395687  0.7148422 0.1952440 2.421284 12296
[4]  {Airline.Status=Blue,
      Type.of.Travel=Personal Travel,
      Flight.cancelled=No}            => {Likelihood.to.recommend=Detractor} 0.1661635  0.7183375 0.2313167 2.433123 14639
[5]  {Airline.Status=Blue,
      Age=old,
      Type.of.Travel=Personal Travel} => {Likelihood.to.recommend=Detractor} 0.1701816  0.7149738 0.2380250 2.421730 14993
[6]  {Airline.Status=Blue,
      Gender=Female,
      Type.of.Travel=Personal Travel,
      Flight.cancelled=No}            => {Likelihood.to.recommend=Detractor} 0.1113394  0.7271312 0.1531215 2.462909  9809

[7]  {Airline.Status=Blue,
      Age=old,
      Gender=Female,
      Type.of.Travel=Personal Travel} => {Likelihood.to.recommend=Detractor} 0.1141544  0.7237334 0.1577299 2.451400 10057
[8]  {Airline.Status=Blue,
      Type.of.Travel=Personal Travel,
      Class=Eco,
      Flight.cancelled=No}            => {Likelihood.to.recommend=Detractor} 0.1361407  0.7183326 0.1895233 2.433107 11994
[9]  {Airline.Status=Blue,
      Age=old,
      Type.of.Travel=Personal Travel,
      Class=Eco}                      => {Likelihood.to.recommend=Detractor} 0.1395687  0.7148422 0.1952440 2.421284 12296
[10] {Airline.Status=Blue,
      Age=old,
      Type.of.Travel=Personal Travel,
      Flight.cancelled=No}            => {Likelihood.to.recommend=Detractor} 0.1661635  0.7183375 0.2313167 2.433123 14639
[11] {Airline.Status=Blue,
      Age=old,
      Gender=Female,
      Type.of.Travel=Personal Travel,
      Flight.cancelled=No}            => {Likelihood.to.recommend=Detractor} 0.1113394  0.7271312 0.1531215 2.462909  9809
[12] {Airline.Status=Blue,
      Age=old,
      Type.of.Travel=Personal Travel,
      Class=Eco,
      Flight.cancelled=No}            => {Likelihood.to.recommend=Detractor} 0.1361407  0.7183326 0.1895233 2.433107 11994
```

Figure 30

➔ We found out some features of detractors:

- ◆ Airline statue: Blue
- ◆ Type of travel: Personal travel
- ◆ Gender: Female
- ◆ Age: Old (over 45 years old)
- ◆ Class: Eco

VI. MAP & Statistical Analysis

We first check the origin state column, each row indicates a customer, We added a column named frequency for counting the number of cases in each state and each city. Then, based on the data file fifty states, We drew a USA map with all states, and filled the states with blue color based on the case number of each state. From bright to steel blue, the color suggested the case from low to high (Figure 31).

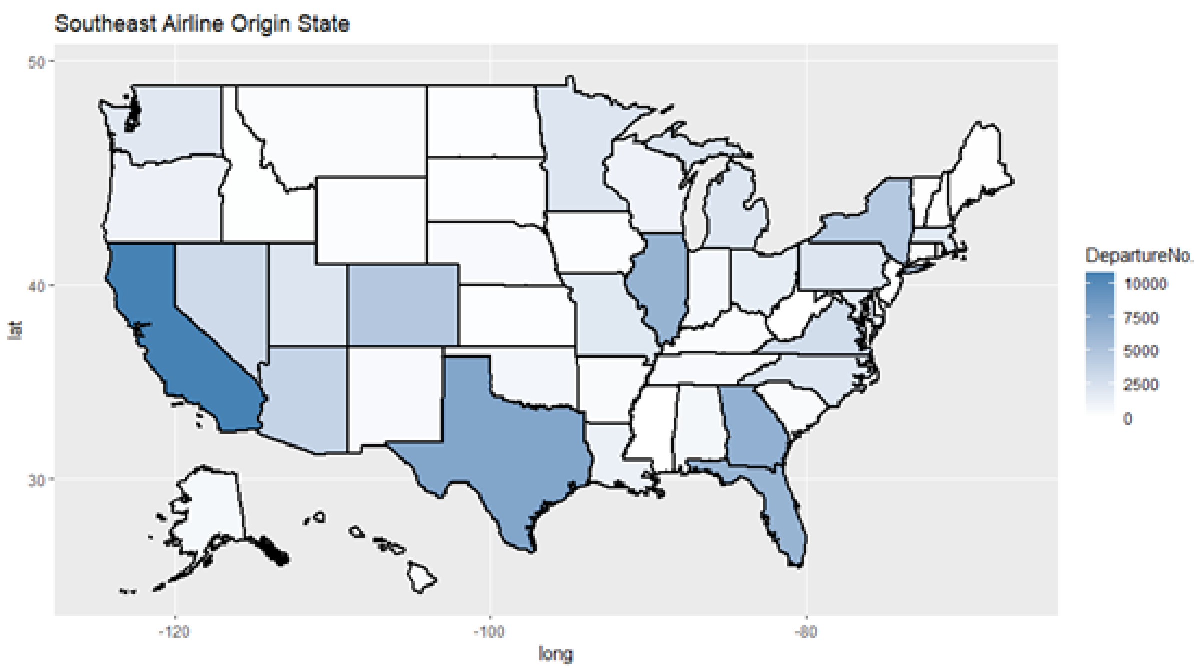


Figure 31

Figure 31 Departure distribution of Southeast Airline within each State. The case numbers from 0 to 100000 were labeled by color from bright to steel blue.

Then we filtered the customers whose “likelihood.to.recommend” were larger than 8 in each state, which indicated the promoters. Dividing the promoter numbers by the case number within each state to find the promoter percentage distribution. The percentage distribution from low to high was labeled by the color of the boundary line of each state from red to dark (Figure 32). The results indicated for those states have large numbers of customers e.g. California, Illinois, Florida have a fair promoter percentage except for Texas.

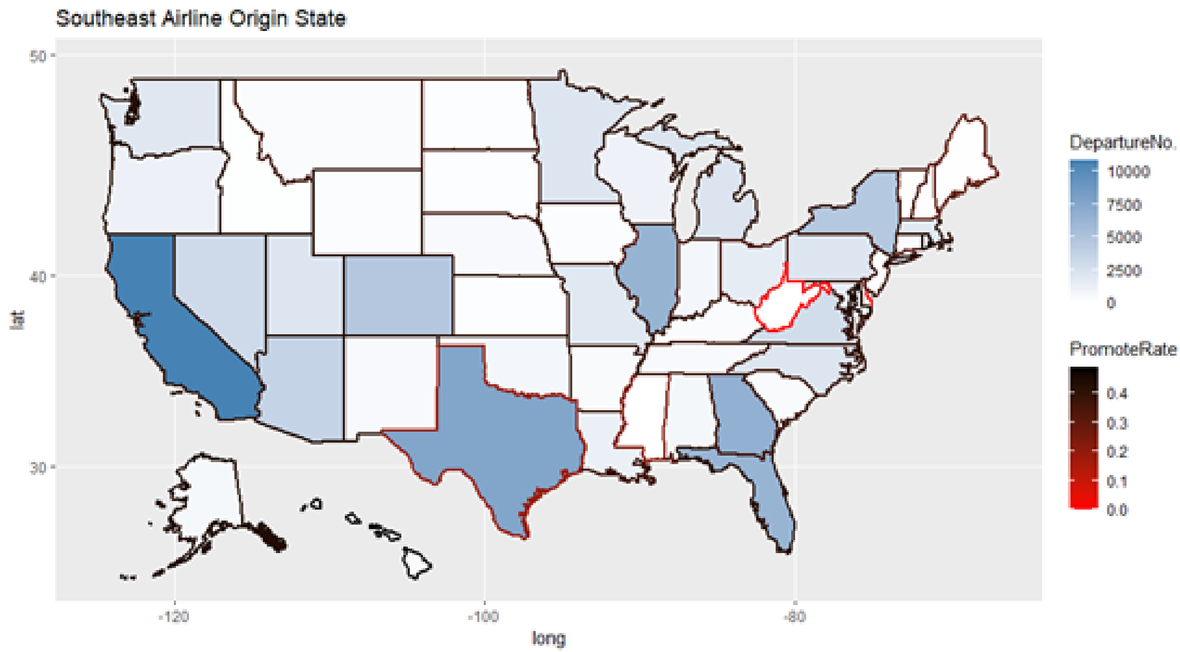


Figure 32

Figure 32 Promoter percentage distribution of Southeast Airline within each state. The percentage of promoters among customers in each state was labeled by the color of the boundary line.

To investigate which airports actually triggered the low promoter rate in Texas, we picked top 30 cities having the largest case number and filtered the customers whose "likelihood.to.recommend" were lower than 7, which is the detractors unsatisfying to the services provided by the airline. The percentage of detractors were calculated within each city, which indicated the sites where the potential customer churn may happen in the future. The detractor percentages were labeled on the map by red dots. The size of dots from small to great suggested the detractor percentages from low to high (Figure 33). The graph showed the 3 largest dots all located in Texas, which was consistent with promoter percentages in the Figure 32.

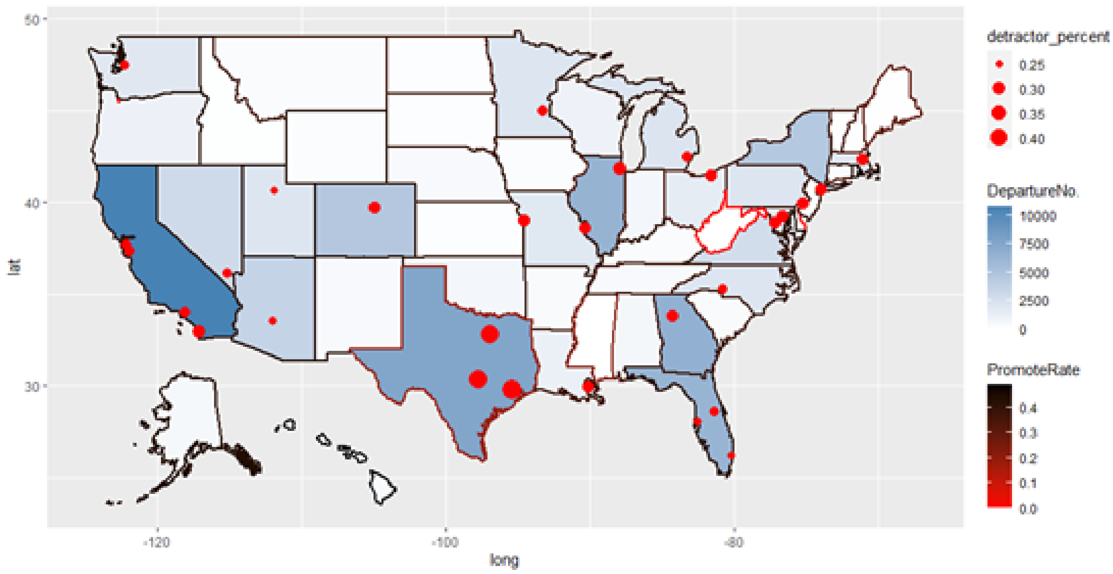


Figure 33

Figure 33 The detractor distribution within top 30 cities with highest customer numbers. The red dots indicated the location of the cities. The size of dots illustrates the detractor percentage. Three cities in Texas have the highest detractor percentage over 0.4. From the above figures, passenger groups departing in Texas are most likely to experience customer churn.

VII. Texas

```
Detractor   Passive  Promoter
0.4951482 0.3632311 0.1416208
```

Figure 34

Then we focused the data on Texas, I created a new data frame called "texas" which only remains customers from Texas. As we can see, 49.51% customers are detractors, 36.32% are passive and only 14.16% are promoters (figure 34).

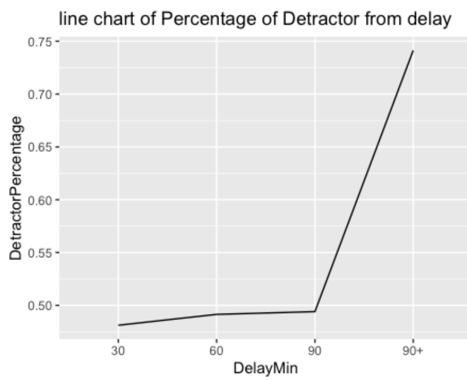


Figure 35

In order to check if the departure delay affects the recommendation grade, we separated customers with 4 ranges, which are delayed less than 30 mins, less than 60 mins, less than 90 mins and greater than 90 mins. We calculated the percentage of detractors from each range. The percentage of detractors didn't increase a lot within 90 mins, but when the flight delay was greater than 90 mins, the percentage rose steeply. We can conclude that departure delay is a factor, especially when the delay time is longer than 90 mins.

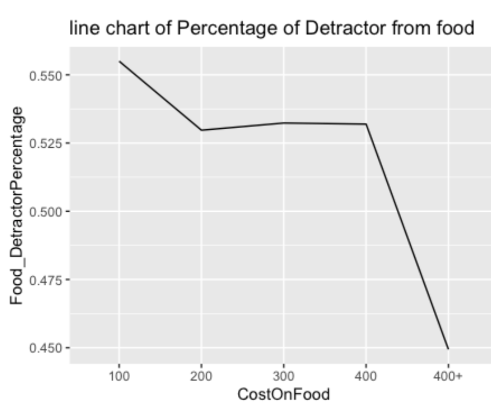


Figure 36

We also check if eating and drinking at the airport affect the recommendation. There are 5 ranges that people spend less than 100, 200, 300, 400 or more. In the line chart, we can see the percentage of detractor is highest when customers spend less than 100 dollars on food, when people spend more than 400 dollars, the percentage of detractor decreases.

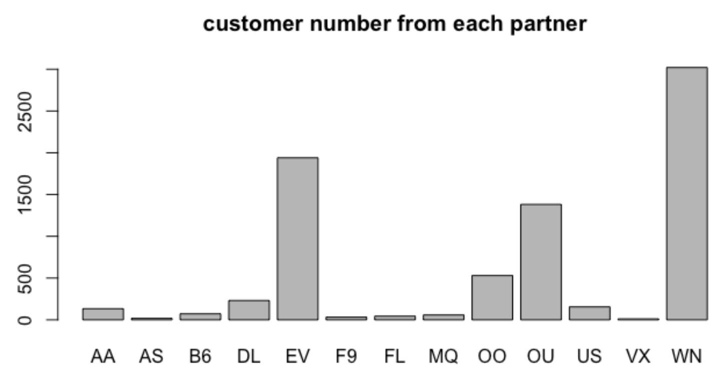


Figure 37

Figure 37 shows the number of customers from each partner airline. WN, EV and OU are top 3 airline Texas customers take. According to figure 15, WN and EV both have high detractor percentages. We can believe that partner is a factor which causes Texas a high detractor percentage.

```
> foodFit <- aov(Likelihood.to.recommend~foodFactor, data = texasFood)
> summary(foodFit)
              Df Sum Sq Mean Sq F value   Pr(>F)    
foodFactor     4    213   53.20    9.69 8.16e-08 ***
Residuals   7621  41843    5.49                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> TukeyHSD(foodFit)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Likelihood.to.recommend ~ foodFactor, data = texasFood)

$foodFactor
                      diff         lwr        upr     p adj
100-200-0-100    0.2425981  0.05381558  0.4313806 0.0041736
200-300-0-100   -0.8686943 -1.39042550 -0.3469631 0.0000553
300-400-0-100   -0.6955341 -1.81149613  0.4204279 0.4336193
400+-0-100      -0.5288675 -3.14018290  2.0824480 0.9816483
200-300-100-200 -1.1112924 -1.65374437 -0.5688404 0.0000002
300-400-100-200 -0.9381322 -2.06393055  0.1876661 0.1535544
400+-100-200    -0.7714655 -3.38699971  1.8440686 0.9292425
300-400-200-300  0.1731602 -1.05321111  1.3995315 0.9953407
400+-200-300     0.3398268 -2.32054600  3.0001997 0.9968400
400+-300-400     0.1666667 -2.67071793  3.0040513 0.9998523
```

Figure38

We added a new column “foodFactor” which separates customers by cost on food in the texasFood data frame. There are five ranges “0-100”, “100-200”, “200-300”, “300-400” and “400+”. Through the ANOVA Tukey's test (Figure 38), we can see the likelihood to recommend of people who cost 100~200 are significantly higher than those whose costs were 0-100, 100-200, and 200-300. No significant difference has been found among groups 0-100, 300-400 and 400+.

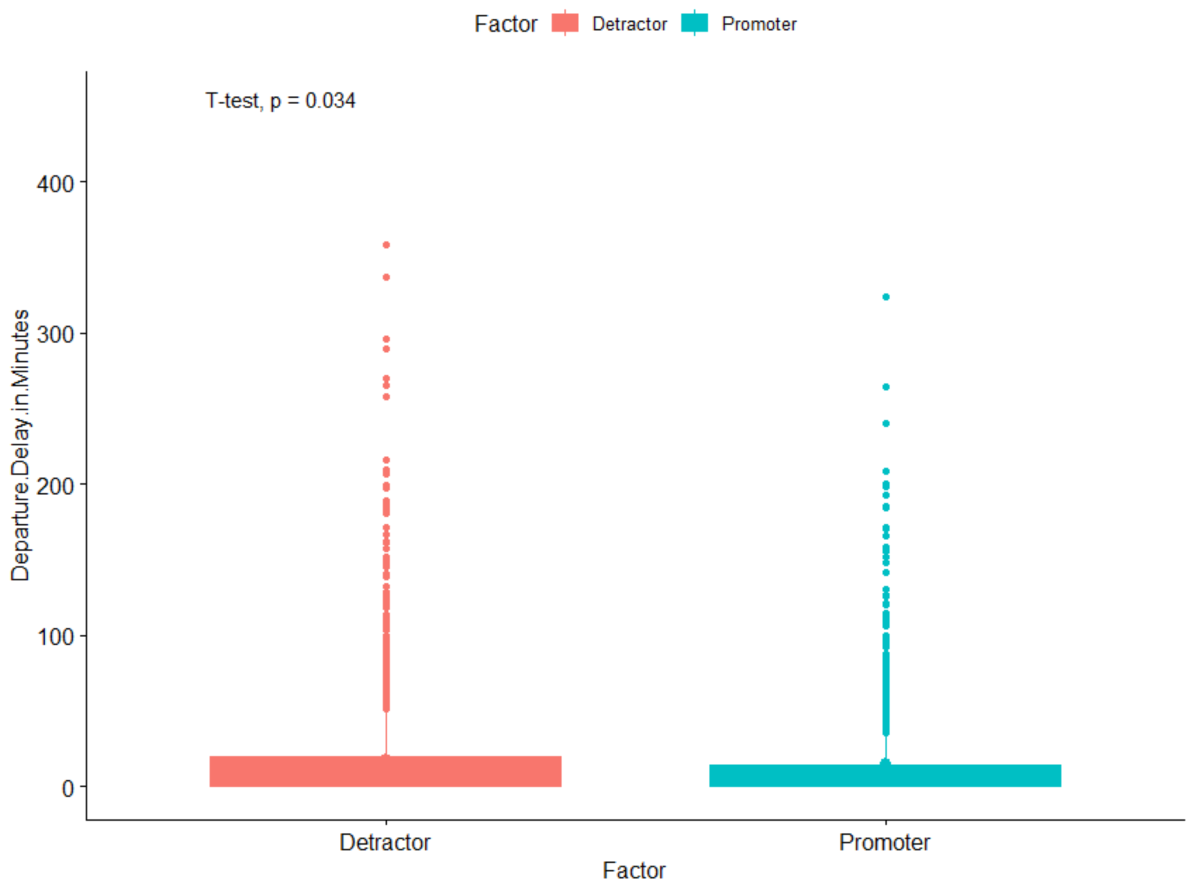


Figure 39

Figure 39 Results of student's t test. The average delay time in group promoter was 15.38 while the average delay time in group detractor was 17.64. The p-value is 0.03397, suggesting a significant difference between two groups.

We also examined whether the dissatisfaction of the passenger group was related to the flight delay time. The promoter group and detractor group were found by filtering the “likelihood.to.recommend”. To clean the data, the rows with missing value in column Departure Delay in Minutes were dropped. Then the departure delay time of two groups (promoter and detractor) was examined by two-tailed student’s t test (Figure 39). As expected, the delay time in detractor was significantly longer than the promoter group, suggesting the dissatisfaction is related to the difference in the flight delay time.

Though the t-test analysis result indicated the significant difference, there are some issues lying there, the priority is that the samples were not typical normal distributions. So we did Kruskal-Wallis test to investigate the relationship between “Likelihood.to.Recommendation” and flight delay time. According to the results, the likelihood.to.recommend of people whose flight delay time shorter than 30 minutes was significantly higher than the groups delay time higher than 30 minutes, which suggested the flight delay time was a critical factor to affect customers’ satisfaction in Texas (Figure 40).

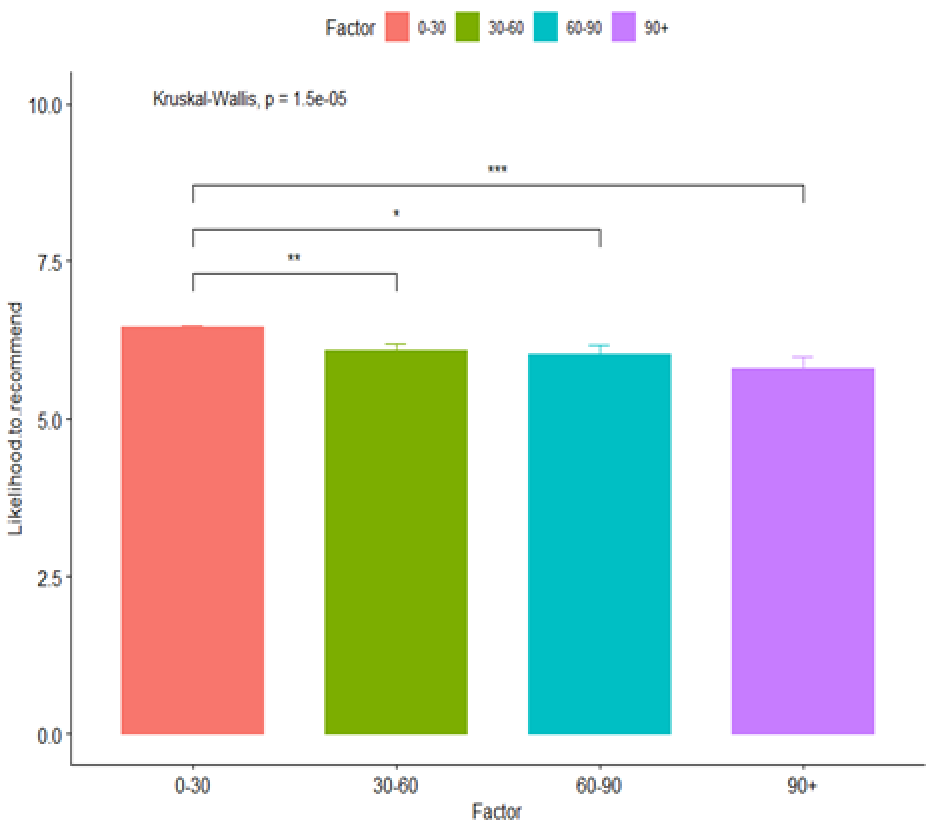


Figure 40

*Figure 40 Results of Kruskal-Wallis test. The sample used in the statistical test is the data where the original state was Texas. Graph illustrated the mean value of likelihood to recommend with standard deviation of each group. * $p < 0.05$, ** $p < 0.005$, ***$p < 0.001$.*

We also did the same tests towards the variable “Eating.and.Drinking.at.Airport”. The result of t-test indicated that promoters had higher food cost in airports, which suggested the food cost and the satisfaction of customers are relational (Figure 41). Nevertheless, it still cannot be confirmed whether the higher food cost contributed to promoters, or the promoters tend to be more likely to consume more food. Therefore, we ran another Kruskal-Wallis test by grouping customers into 5 levels based on the spending. The results shown in Figure 42. Interestingly, we found that the likelihood to recommend people whose spending were between 100-200 were more satisfied. The likelihood to recommend customers in groups with higher spending (e.g. group 200-300, 300-400, 400+) was not actually significantly higher than people had lower spending (e.g. 0-100, 100-200).

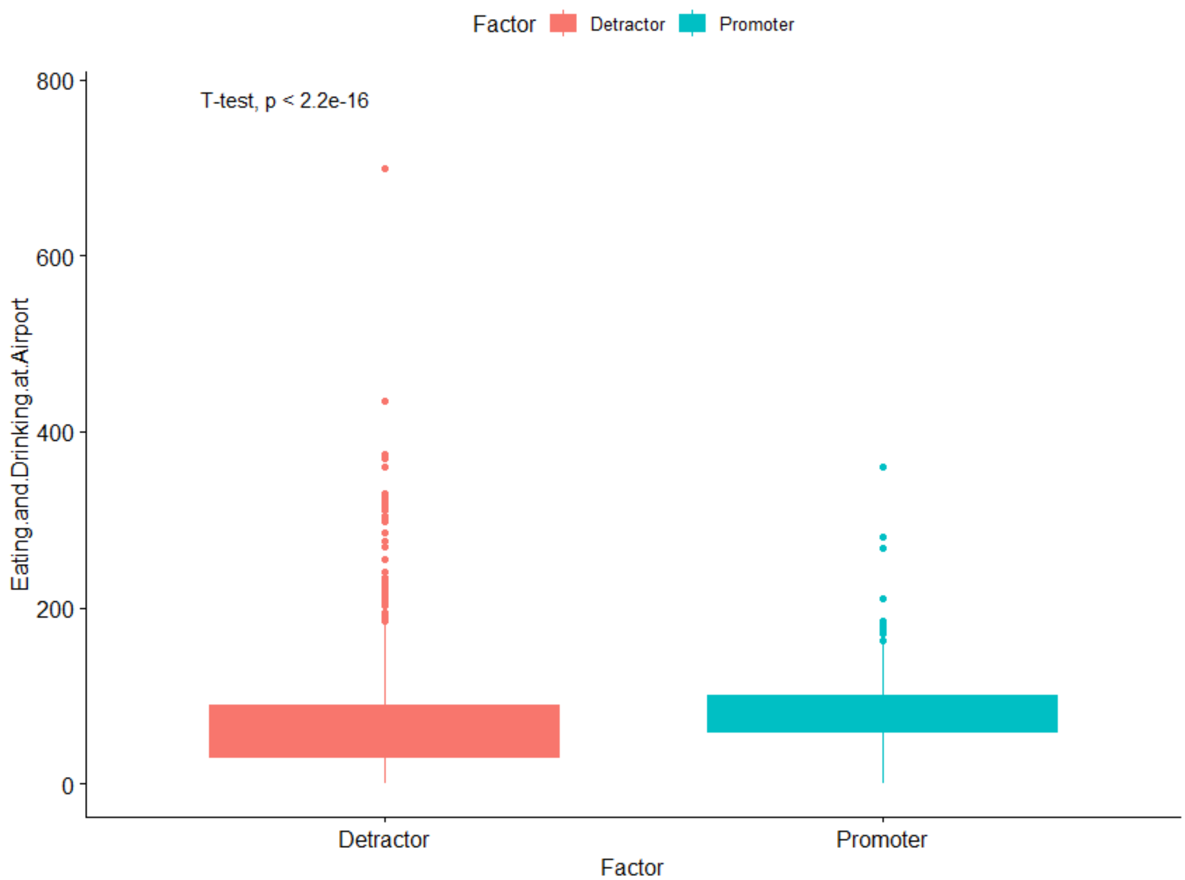


Figure 41

Figure 41 Results of student's t test. The food spending in group promoter was significantly higher than the spending in group detractor.

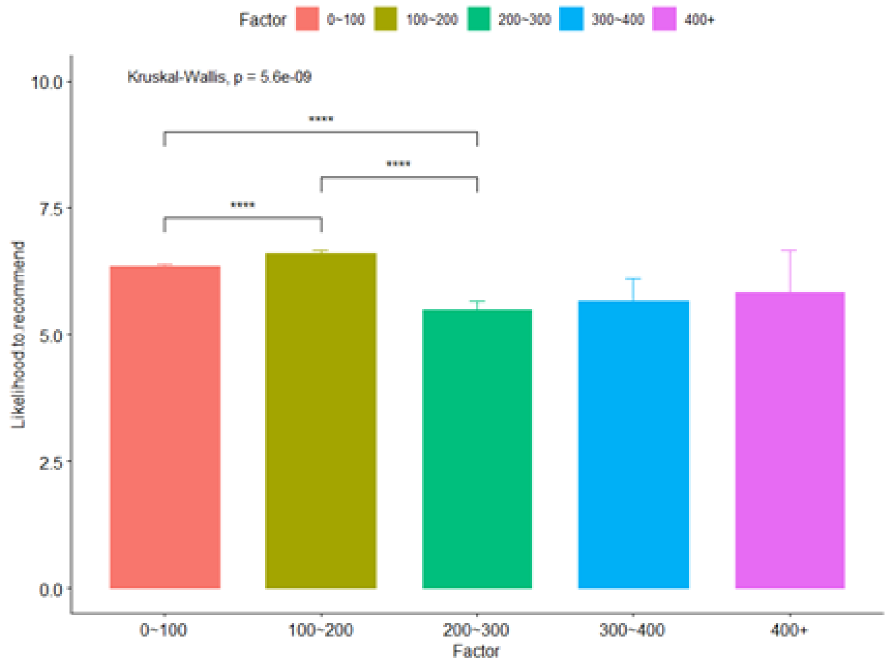


Figure 42

*Figure 42. Results of Kruskal-Wallis test. The sample used in this statistical test is based on data of Texas. Graph illustrated the mean value of likelihood to recommend with standard deviation of each group. * p < 0.05, ** p < 0.005, ***p < 0.001*

To find out the other variables that may affect clients attitude, we then establish a SVM model to predict the detractors and promoters. We set the ‘detractor’ and ‘promoter’ as the factors which need to be predicted. Then we let ‘Type.of.Travel’, ‘Flight.Distance’, ‘Flight.cancelled’, and ‘Partner.Name’ come into play. The result is shown in Figure 43. The accuracy reached 0.75, suggesting that the partner airline, whether flight cancelled, flight distance and type of travel also contributed to differentiate customers into the detractors and promoters.

```
pred         Detractor Promoter
  Detractor        768      120
  Promoter         229      318

               Accuracy : 0.7568
                 95% CI : (0.7337, 0.7788)
    No Information Rate : 0.6948
    P-Value [Acc > NIR] : 1.119e-07

                  Kappa : 0.464

 Mcnemar's Test P-Value : 7.421e-09

            Sensitivity : 0.7703
            Specificity : 0.7260
         Pos Pred Value : 0.8649
         Neg Pred Value : 0.5814
             Prevalence : 0.6948
         Detection Rate : 0.5352
   Detection Prevalence : 0.6188
      Balanced Accuracy : 0.7482

       'Positive' Class : Detractor
```

Figure 43

Figure 43. SVM model to predict the detractor groups and promoter groups. The variables involved in the model were 'Type.of.Travel', 'Flight.Distance', 'Flight.cancelled', and 'Partner.Name', with parameters cross=3, cost(C)=5.

We also looked through the association rules to check what was associated with the detractors. We set the factors 'Destination.City', 'Origin.City', 'Airline.Status', 'Likelihood.to.recommend', 'Gender', 'Departure.Delay.in.Minutes', 'Partner.Code', 'Eating.and.Drinking.at.Airport' to dig up what were related to the detractors. Before running the model, we first defined the likelihood.to.recommend less than 7 as the 'Detractor'. For column 'Eating.and.Drinking.at.Airport', we defined the values less than 100 as '0~100', the values within 100~200 as '100~200', the values within 200~300 as '200~300', the values within 300~400 as '300~400', and the values greater than 400 as '400+'. In the column 'Departure.Delay.in.Minutes', the values less than 30 were defined as 'On time' while the values larger than 30 were defined as 'Delayed'. The result shows 12 rules in Figure 44. The 4th rule has the highest lift 1.44, indicating that customers who dealt with partner airline EV and had Blue status are more likely to be the detractors.

```
     lhs                                     rhs                                  support confidence  coverage     lift count
[1]  {Partner.Code=EV}                    => {Likelihood.to.recommend=Detractor} 0.1401495  0.5573248 0.2514682 1.256161  1050
[2]  {Airline.Status=Blue}                => {Likelihood.to.recommend=Detractor} 0.3711959  0.5317400 0.6980779 1.198495  2781
[3]  {Origin.City=Houston, TX,
      Partner.Code=EV}                    => {Likelihood.to.recommend=Detractor} 0.1042445  0.5515537 0.1890016 1.243153   781
[4]  {Airline.Status=Blue,
      Partner.Code=EV}                    => {Likelihood.to.recommend=Detractor} 0.1114522  0.6393568 0.1743193 1.441053   835
[5]  {Departure.Delay.in.Minutes=On time,
      Partner.Code=EV}                    => {Likelihood.to.recommend=Detractor} 0.1043780  0.5503167 0.1896690 1.240365   782
[6]  {Airline.Status=Blue,
      Partner.Code=WN}                    => {Likelihood.to.recommend=Detractor} 0.1461559  0.5157796 0.2833689 1.162521  1095
[7]  {Airline.Status=Blue,
      Gender=Female}                      => {Likelihood.to.recommend=Detractor} 0.2365190  0.5828947 0.4057662 1.313793  1772
[8]  {Origin.City=Houston, TX,
      Airline.Status=Blue}                => {Likelihood.to.recommend=Detractor} 0.2172985  0.5392514 0.4029632 1.215425  1628
[9]  {Airline.Status=Blue,
      Departure.Delay.in.Minutes=On time} => {Likelihood.to.recommend=Detractor} 0.2633476  0.5083741 0.5180192 1.145830  1973
[10] {Origin.City=Houston, TX,
      Airline.Status=Blue,
      Gender=Female}                      => {Likelihood.to.recommend=Detractor} 0.1384143  0.5932494 0.2333155 1.337131  1037
[11] {Airline.Status=Blue,
      Gender=Female,
      Departure.Delay.in.Minutes=On time} => {Likelihood.to.recommend=Detractor} 0.1689802  0.5545335 0.3047250 1.249869  1266
[12] {Origin.City=Houston, TX,
      Airline.Status=Blue,
      Departure.Delay.in.Minutes=On time} => {Likelihood.to.recommend=Detractor} 0.1477576  0.5103734 0.2895088 1.150336  1107
```

Figure 44

Figure 44 Association rules using data in Texas. 12 rules were digged up by the parameters set as: support =0.1, confidence=0.5, default= 'lhs' and rhs=("Likelihood.to.recommend=Detractor")

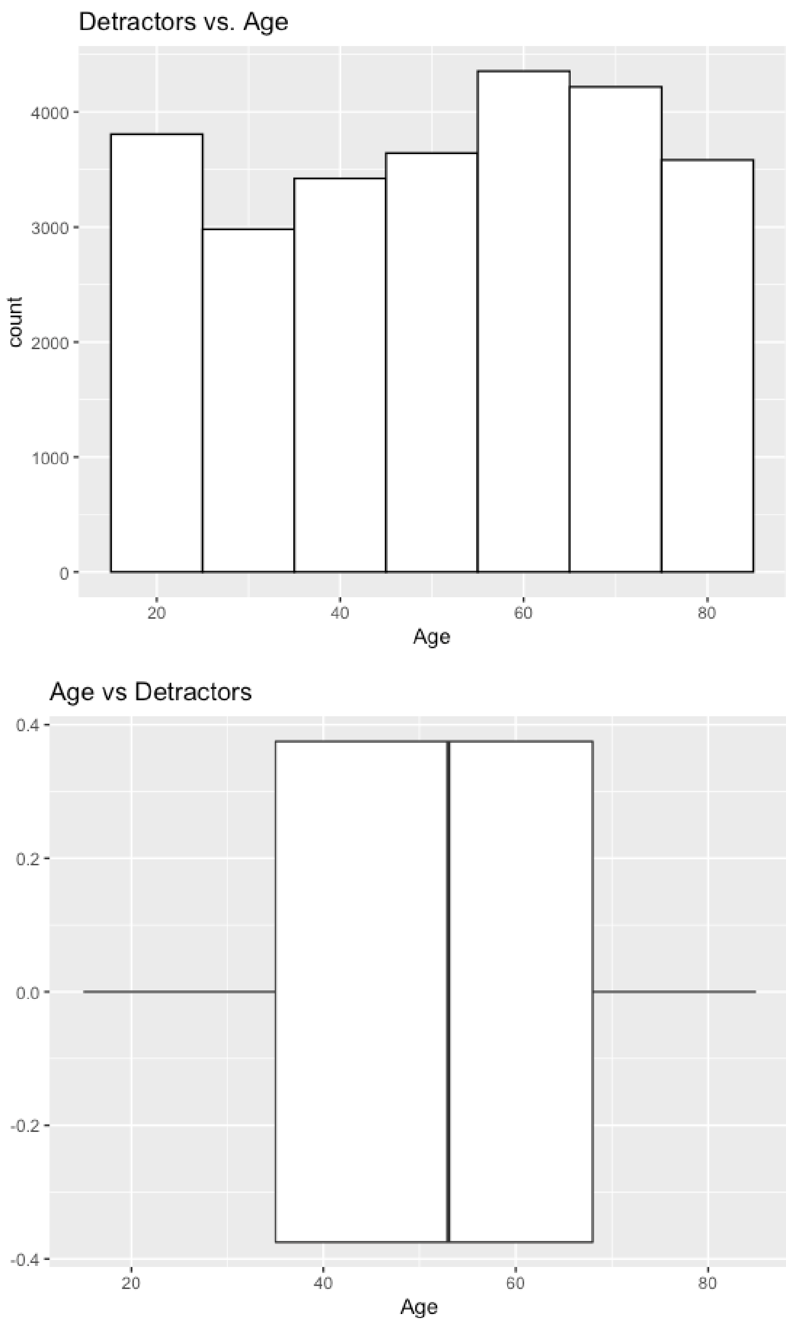
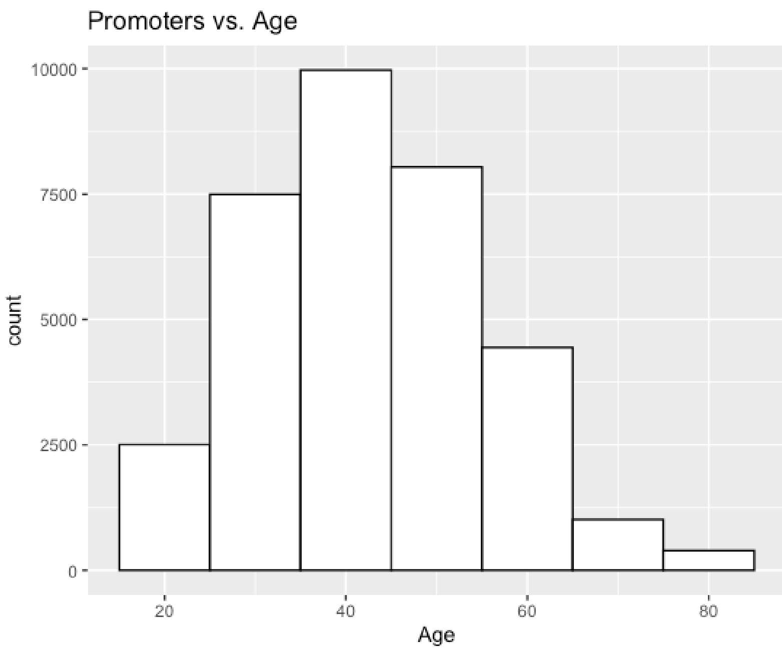


Figure 45

Figure 45 gives us a look at the comparison between how likely it is to be a detractor based on your age. We can see that the least amount of detractors are around 30 years old while the greatest amount are around 60 years old. I was a bit surprised to see that 20 year olds held such a large amount of detraction. I suspect it has something to do with always wanting something new and more advanced. We are in the age of information and it is likely that a millennial would do their research to know what is expected of an airline.



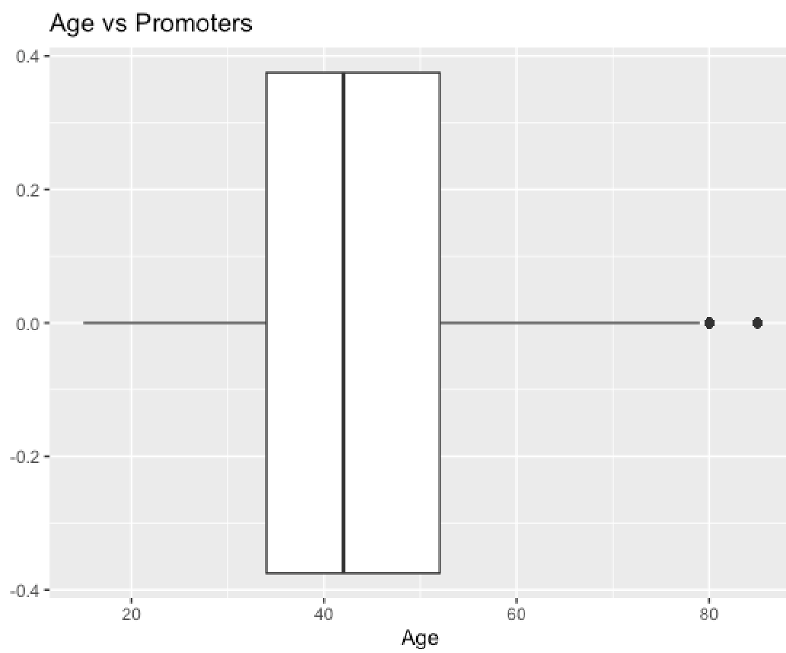


Figure 46

Figure 46 shows us the frequency of promoters with respect to their ages. We can see from the graphs that middle aged people are more likely to be a promoter. They feel comfortable with what they already know. The reason why I believe senior citizens might not be promoters is because of the level of comfort or the closeness of everything. It is perhaps too much work for them to get around or find something. The airline is not majorly focused on the old. They must adopt an out with the old in with the new perspective in order to thrive and stay afloat in today's world.

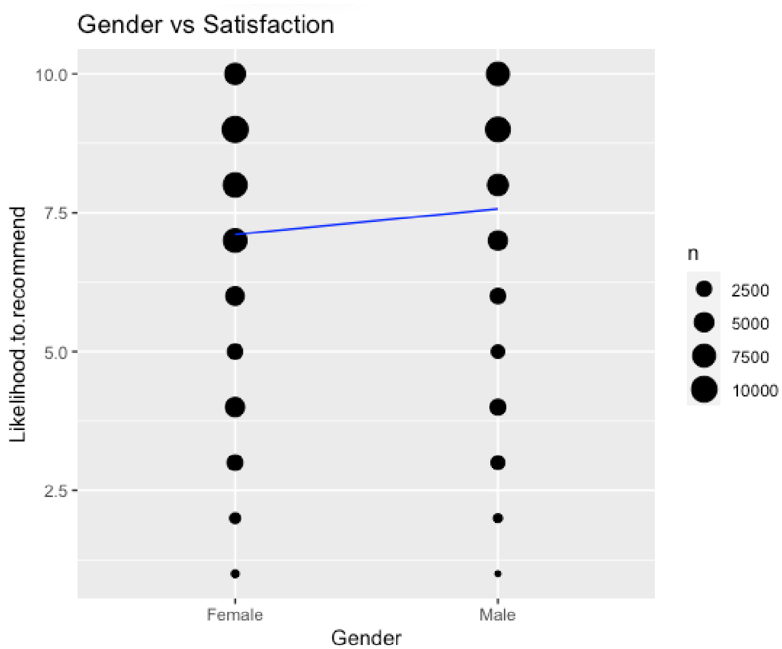


Figure 47

In figure 47 we tried to see the difference between gender and their likelihood to recommend. A likelihood to recommend score of less than 7 meant they were detractors, meaning they would likely not use the service again. A score above 8 meant they were promoters, meaning they would speak fondly of the service they received to their friends and colleagues. Here we can see that there isn't much of a difference. In fact it is a bit skewed because less male gave a score of 8 and below but more women gave a score of 9 while more men gave a perfect score of 10. This did not lead us to anything substantial.

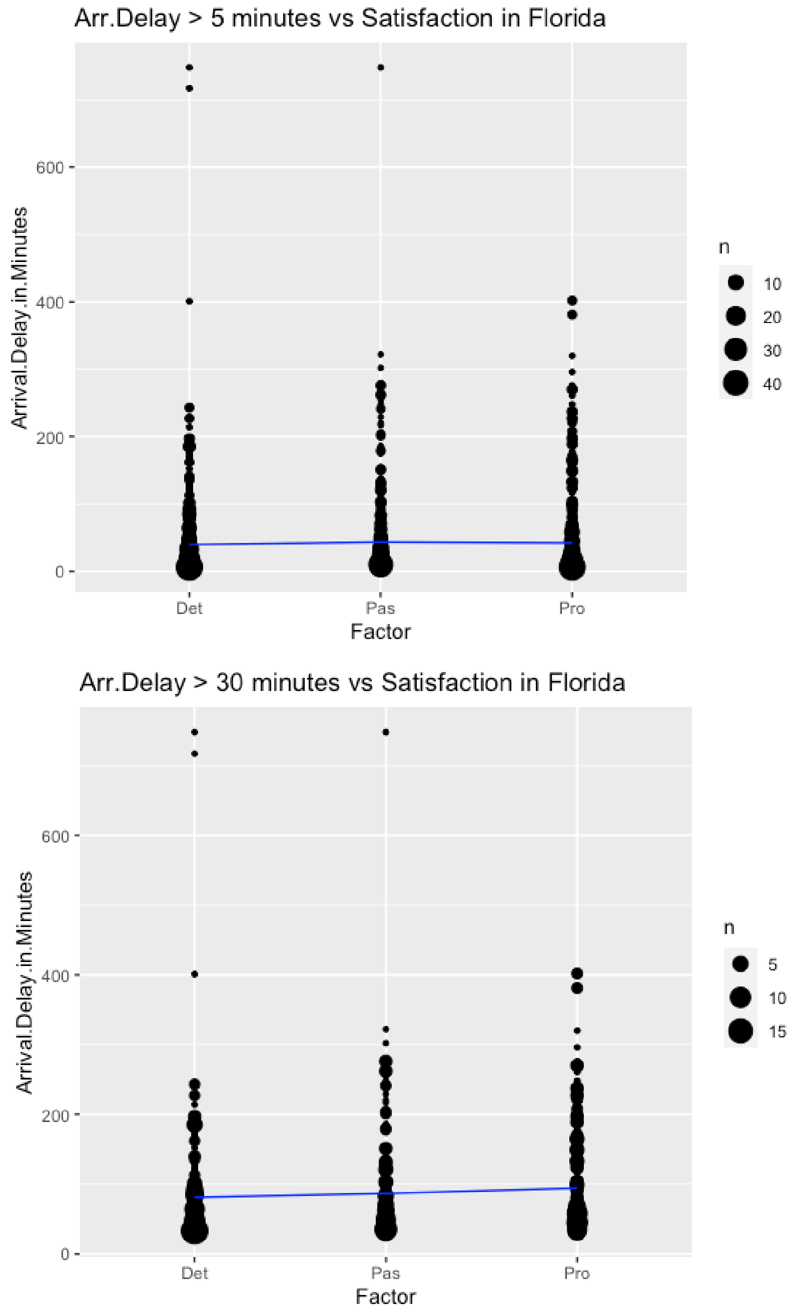


Figure 48

Figure 48 shows us the significance of arrival delays of more than 5 minutes and more than 30 minutes, specifically in Florida. We chose these to see if there was any difference between the time intervals. This was quite interesting to see because there were more promoters with an experienced delay of 400 minutes than there were detractors for the same time delay. There was little correlation here. The average number of people however was greater for detractors at a lesser time than it was for promoters at a greater time. That was interesting.

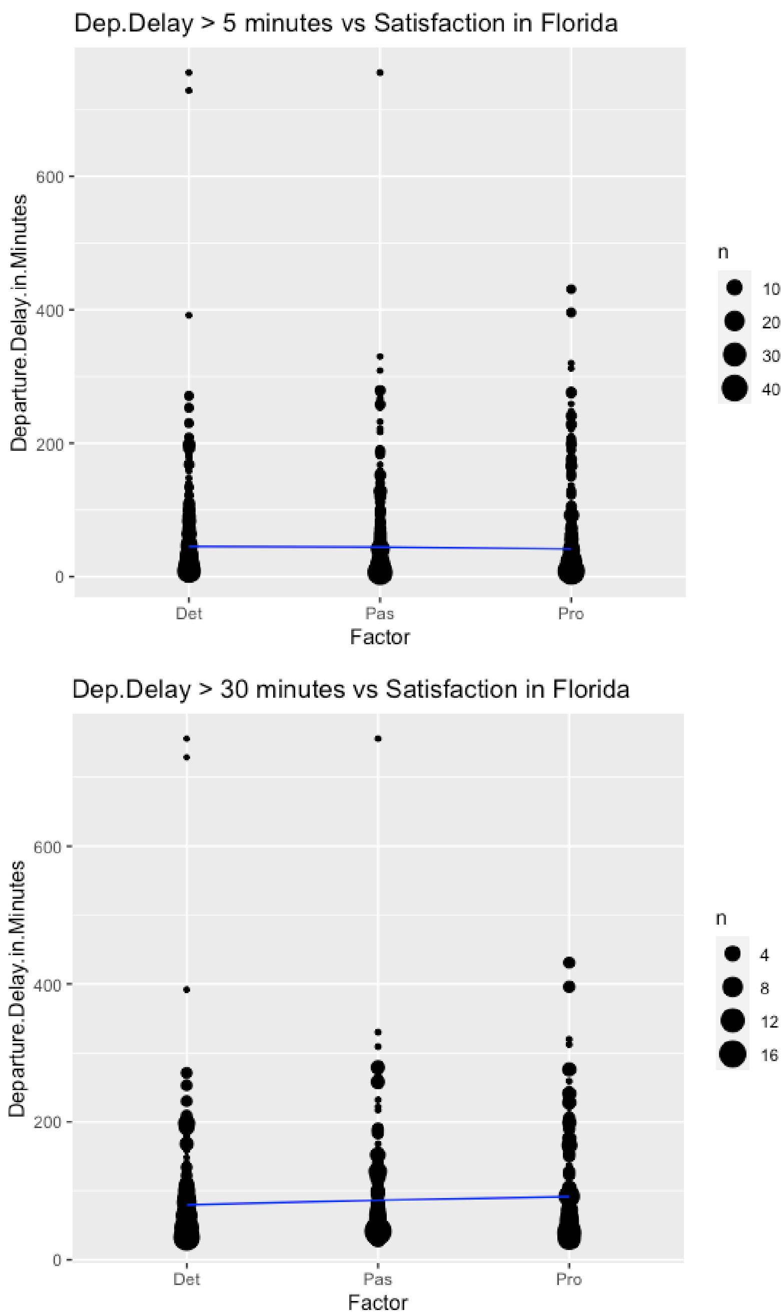


Figure 49

In figure 49 we checked to see if it would be the same for a delay in departure. I am still not sure why there are more people promoting than there are detracting for the same amount of time. However, again, there were more people detracting with a lesser amount of time in delay. This might be because some people run late to the airport because of traffic or because they forgot another thing back home, and so they appreciate that extra time they have. Yet and still, the question that arises to the mind is what exactly is influencing this disparity?

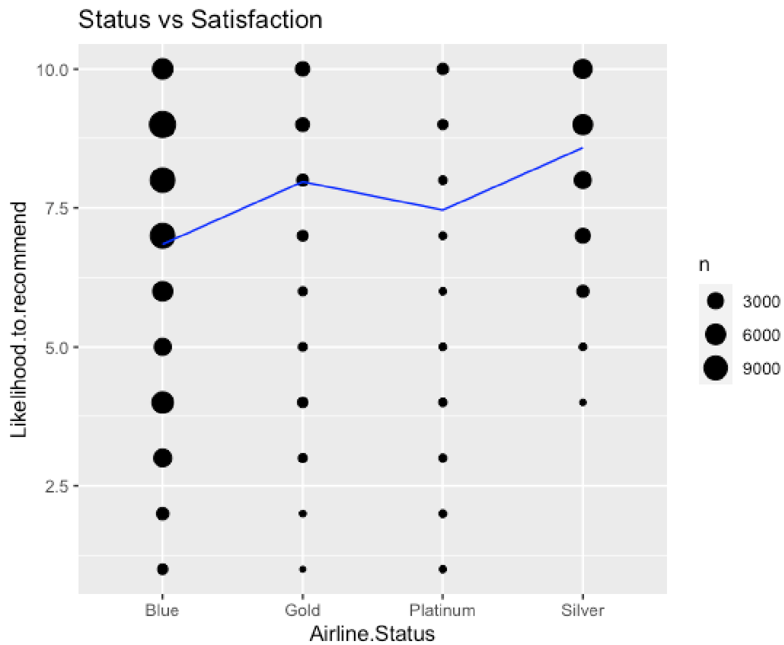


Figure 50

Figure 50 takes a look at the likelihood to recommend when compared to their airline status. There are 4 statuses ranging from low to high levels of Blue, Gold, Platinum, and Silver. We found that there are more members with a Blue status promoting and detracting than all others. This is more than likely due to the fact that there are simply more people with a Blue status than there are with higher statuses. However the interesting part is that there are those with a Silver status who are more likely to not recommend, with a score of ~6, than there are for Platinum and Gold. That is intriguing because we would expect for you to recommend at a higher rate if you have the highest status, therefore reaping more of the benefits.

Figure 51

- Scatter Plot showing detractors with a 0.15 support and a 0.5 confidence level. The shaded lift shows how well these rules ranked in the survey dataset.

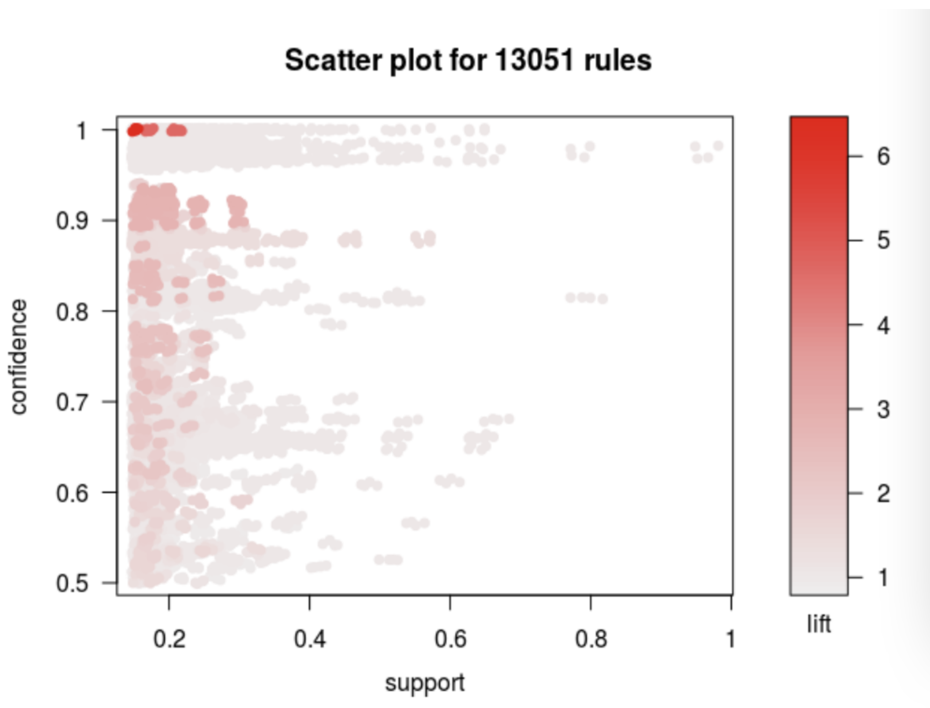


Figure 52

- Scatter Plot showing data from the airData with a 0.15 support and a 0.5 confidence level. The shaded lift shows how well these rules ranked in the survey dataset.

```
> summary(airData)
 Destination.City   Origin.City        Airline.Status          Age          Gender          Price.Sensitivity
 Length:88100       Length:88100       Length:88100       Min.   :15.00   Length:88100       Min.   :0.000
 Class :character   Class :character   Class :character   1st Qu.:33.00   Class :character   1st Qu.:1.000
 Mode  :character   Mode  :character   Mode  :character   Median :45.00   Mode  :character   Median :1.000
                                                          Mean   :46.22                      Mean   :1.277
                                                          3rd Qu.:59.00                      3rd Qu.:2.000
                                                          Max.   :85.00                      Max.   :4.000

 Year.of.First.Flight Flights.Per.Year    Loyalty          Type.of.Travel     Total.Freq.Flyer.Accts
 Min.   :2003         Min.   : 0.00    Min.   :-0.97619   Length:88100       Min.   : 0.0000
 1st Qu.:2004         1st Qu.: 9.00    1st Qu.:-0.70000   Class :character   1st Qu.: 0.0000
 Median :2007         Median :17.00    Median :-0.42857   Mode  :character   Median : 0.0000
 Mean   :2007         Mean   :20.04    Mean   :-0.27419                      Mean   : 0.8899
 3rd Qu.:2010         3rd Qu.:29.00    3rd Qu.: 0.05882                      3rd Qu.: 2.0000
 Max.   :2012         Max.   :98.00    Max.   : 1.00000                      Max.   :12.0000

 Shopping.Amount.at.Airport Eating.and.Drinking.at.Airport    Class             Day.of.Month
 Min.   :  0.00             Min.   :  0.00                 Length:88100       Min.   : 1.00
 1st Qu.:  0.00             1st Qu.: 30.00                 Class :character   1st Qu.: 8.00
 Median :  0.00             Median : 60.00                 Mode  :character   Median :16.00
 Mean   : 26.62             Mean   : 67.99                                    Mean   :15.69
 3rd Qu.: 30.00             3rd Qu.: 90.00                                    3rd Qu.:23.00
 Max.   :745.00             Max.   :895.00                                    Max.   :31.00

 Flight.date        Partner.Code       Partner.Name       Origin.State       Destination.State
 Length:88100       Length:88100       Length:88100       Length:88100       Length:88100
 Class :character   Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character


 Scheduled.Departure.Hour Departure.Delay.in.Minutes Arrival.Delay.in.Minutes Flight.cancelled
 Min.   : 1.00            Min.   :  0.00             Min.   :  0.00           Length:88100
 1st Qu.: 9.00            1st Qu.:  0.00             1st Qu.:  0.00           Class :character
 Median :13.00            Median :  0.00             Median :  0.00           Mode  :character
 Mean   :13.02            Mean   : 15.04             Mean   : 15.38
 3rd Qu.:17.00            3rd Qu.: 13.00             3rd Qu.: 13.00
 Max.   :23.00            Max.   :978.00             Max.   :970.00
                          NA's   :1607               NA's   :1838
 Flight.time.in.minutes Flight.Distance  Likelihood.to.recommend     olong             olat
 Min.   : 13.0          Min.   :  67.0   Min.   : 1.000          Min.   :-165.39   Min.   :18.02
 1st Qu.: 61.0          1st Qu.: 373.0   1st Qu.: 6.000          1st Qu.:-111.93   1st Qu.:33.56
 Median : 92.0          Median : 628.0   Median : 8.000          Median : -90.14   Median :37.67
 Mean   :113.1          Mean   : 807.7   Mean   : 7.309          Mean   : -95.33   Mean   :37.08
 3rd Qu.:143.0          3rd Qu.:1024.0   3rd Qu.: 9.000          3rd Qu.: -81.64   3rd Qu.:40.72
 Max.   :443.0          Max.   :3414.0   Max.   :10.000          Max.   : -66.12   Max.   :71.29
 NA's   :1838                            NA's   :4
     dlong             dlat          freeText
 Min.   :-165.39   Min.   :18.02   Length:88100
 1st Qu.:-111.93   1st Qu.:33.82   Class :character
 Median : -90.14   Median :37.67   Mode  :character
 Mean   : -95.33   Mean   :37.08
 3rd Qu.: -81.64   3rd Qu.:40.72
 Max.   : -66.12   Max.   :71.29
```

Figure 53

- Summary statistics of the different columns of data in airData.

```
> inspect(tail(rules))
    lhs                                  rhs                                  support confidence  coverage      lift count
[1] {Airline.Status=Blue,
     Price.Sensitivity=[1,4],
     Class=Eco,
     Flight.cancelled=No}             => {Arrival.Delay.in.Minutes=[0,7)}  0.3487741  0.6646406 0.5247560 1.0216855 30727
[2] {Airline.Status=Blue,
     Price.Sensitivity=[1,4],
     Shopping.Amount.at.Airport=[0,15),
     Class=Eco}                       => {Flight.cancelled=No}             0.3581271  0.9777798 0.3662656 0.9962921 31551
[3] {Airline.Status=Blue,
     Shopping.Amount.at.Airport=[0,15),
     Class=Eco,
     Flight.cancelled=No}             => {Price.Sensitivity=[1,4]}         0.3581271  0.9661033 0.3706924 0.9972197 31551
[4] {Airline.Status=Blue,
     Price.Sensitivity=[1,4],
     Shopping.Amount.at.Airport=[0,15),
     Flight.cancelled=No}             => {Class=Eco}                       0.3581271  0.8165795 0.4385698 1.0026572 31551
[5] {Price.Sensitivity=[1,4],
     Shopping.Amount.at.Airport=[0,15),
     Class=Eco,
     Flight.cancelled=No}             => {Airline.Status=Blue}             0.3581271  0.7000288 0.5115891 1.0252783 31551
[6] {Airline.Status=Blue,
     Price.Sensitivity=[1,4],
     Class=Eco,
     Flight.cancelled=No}             => {Shopping.Amount.at.Airport=[0,15)} 0.3581271  0.6824641 0.5247560 1.0353186 31551
>
```

Figure 54

- Inspect the rules created from apriori and sorting confidence by decreasing value from airData. Doesn't show very substantial relationships between the lhs and rhs from the data given.

VIII. Conclusion

National Scale

- It’s hard to train a linear model to predict the likelihood to recommend perfectly
- For a support vector machine, the accuracy of predicting detractor is 76.13% .
- Detractor’s features: airline status is blue, type of travel is personal travel, female, over 45 years old, and Eco class. Therefore, Southeast Airline should put more attention on these customers.

Texas

- Flight delay may be a critical factor in annoying clients.
- Find the receipts of people whose food spending was in 100~200

Based on the result from models and tests, we have some suggestions for Southeast Airlines. Since females have lower recommendation grades than males, airlines can consider providing some more detailed services, such as helping women place their luggage when boarding or providing some childcare services when mothers go to the washroom.

Flight delay may be a critical factor in annoying clients. It could be a good choice to give customers some compensation. If the flight delay is less than 1 hour, the airline can provide some snacks; if the flight delay is more than 1 hour, they can issue some coupon that can be used at the airport.

Due to the association rules, old people is another factor to be a detractor. Southeast Airline can provide some specific meals to old people who suffer from diseases (such as diabetes). They can also give priority to the old people when checking in.

Southeast Airline also needs to improve its services to those customers who choose Eco class. They can try to give Eco more meal options, and they can give customers eye masks and toothbrushes on long-haul flights. For customers who have blue status, Southeast Airline could consider reducing restrictions of flight change policy.