IST 687

Analysis of Airline Data

M004 GROUP 1

Akshay Bhala

Rishabh Agarwal

Sai Praharsha Devalla

Liyang Xue

Xuanran Ji

INDEX

DATA MUNGING

The dataset for South East Airlines consisted 10282 rows and 32 variables.

Our goal is to perform analysis and give actionable insights to South east airlines for improving customer satisfaction.

Our first step was to find missing values and do explanatory analysis in our dataset. After our observation we decided to drop freeText, latitude and longitude column as they are not useful for our analysis.

For Flight Cancelled , we dropped those rows where flight cancelled = yes and where flights were not cancelled we followed the following approach:

First, for NA in Flight.time.in.minutes. If there is data available for the same route from origin to destination having same distance, we will be calculating the average of the Flight.time.in.minutes from the muliple same records or if there is only one record available we will replace Flight.time.in.minutes with the same value of same route.

For Arrival Delay in Minutes we will be using data from Departure.Delay.in.minutes. For example, if the Departure Delay in Minutes has 30 minutes delay. We will be putting the same value in Arrival Delay in Minutes variable.

The next step we followed is to convert each column into numeric or factor as per each characteristics. We converted Likelihood to recommend into 3 groups(Promoters, Passive and Detractors) , age into 3 groups(Age between 15 and 29, Age between 30 and 54 and Age above 54) , arrival delay in minutes into 2 groups (delay >5 mins= Yes else No) and departure delay in minutes into 2 groups (delay >5 mins= Yes else No).

And lastly with the help of Correlation Matrix we removed unsignificant columns.

Code:

```
########################## DATA MUNGING ###################################

# Replacing NA's of Flight time with values which have data for same places
df$Flight.time.in.minutes[14] <- 100
df$Flight.time.in.minutes[124] <- 130
df$Flight.time.in.minutes[359] <- 38
df$Flight.time.in.minutes[1167] <- 112
df$Flight.time.in.minutes[1493] <- 144
df$Flight.time.in.minutes[1541] <- 110
df$Flight.time.in.minutes[1918] <- 68
df$Flight.time.in.minutes[2302] <- 112
df$Flight.time.in.minutes[2672] <- 125
df$Flight.time.in.minutes[3559] <- 115
df$Flight.time.in.minutes[4303] <- 115
df$Flight.time.in.minutes[8431] <- 115
df$Flight.time.in.minutes[5540] <- 106
df$Flight.time.in.minutes[6170] <- 180
df$Flight.time.in.minutes[6718] <- 275
df$Flight.time.in.minutes[7328] <- 115
df$Flight.time.in.minutes[7500] <- 65
df$Flight.time.in.minutes[7864] <- 267
df$Flight.time.in.minutes[7881] <- 118
df$Flight.time.in.minutes[8119] <- 57
df$Flight.time.in.minutes[9208] <- 208
df$Flight.time.in.minutes[9235] <- 63

# Replacing NA's of Arrival delay time with values which have data for same places
df$Arrival.Delay.in.Minutes[(is.na(df$Arrival.Delay.in.Minutes) & df$Flight.cancelled=='No')] <- df$Departure.Delay.in.Minutes[(is.na(df$Arrival.Delay.in.Min

# Removing all the Records where the Flight Cancelled is "Yes"
df <- df[df$Flight.cancelled=='No',]
```

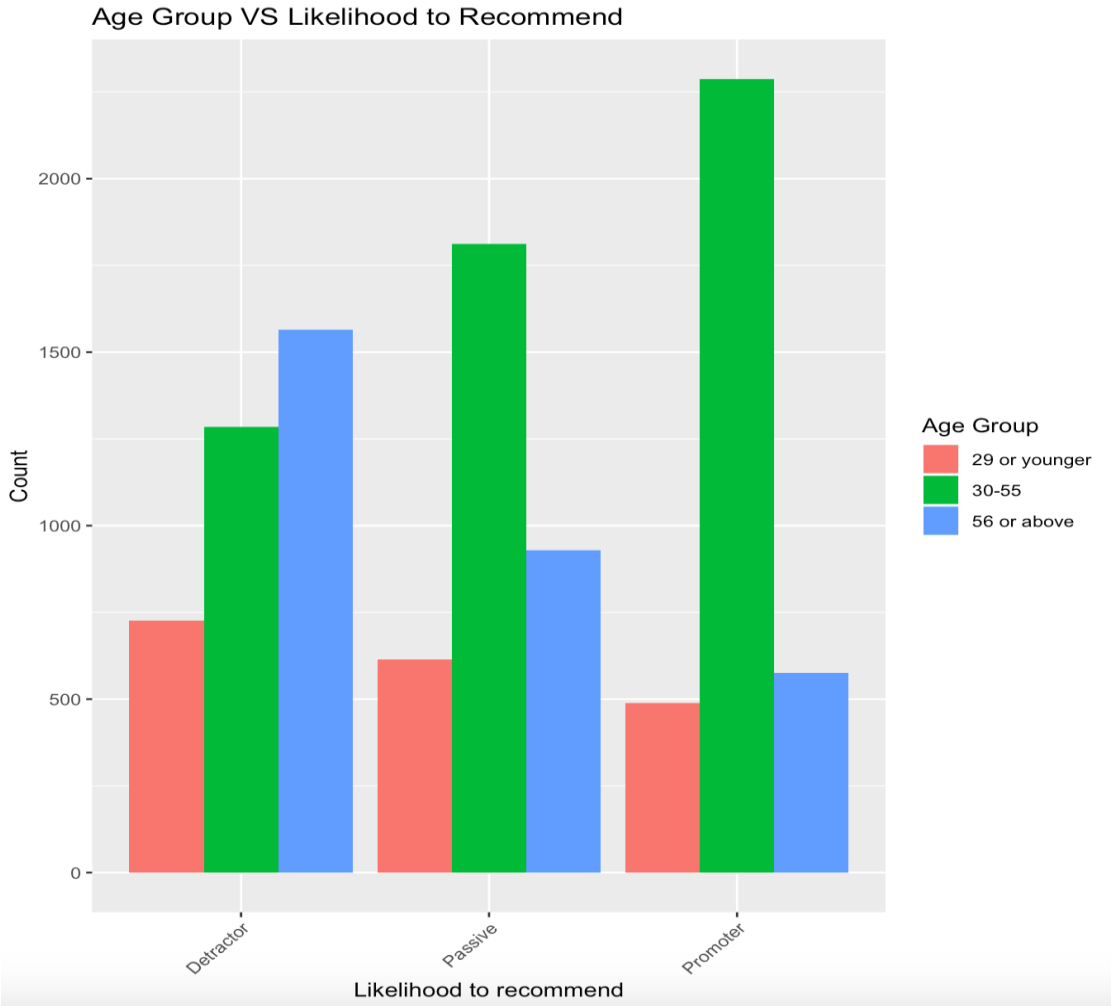
```r
# Converting the Likelihood to recommend to "Detractor,Passive and Promoter"

for (i in 1:length(df$Likelihood.to.recommend))
  {
    if (df$Likelihood.to.recommend[i] < 7)
      {
        df$pr_pa_dt[i] <- "Detractor"
      }
    else if (df$Likelihood.to.recommend[i] >= 7 & df$Likelihood.to.recommend[i] <= 8)
      {
        df$pr_pa_dt[i] <- "Passive"
      }
    else
      {
        df$pr_pa_dt[i] <- "Promoter"
      }
  }

# Converting the Arrival.Delay.in.Minutes if Delay is greater than 5 mins to "Yes" or "No"

for (i in 1:length(df$Arrival.Delay.in.Minutes))
{
  if (df$Arrival.Delay.in.Minutes[i] > 5)
  {
    df$ArrivalDelayof5min[i] <- "Yes"
  }
  else
  {
    df$ArrivalDelayof5min[i] <- "No"
  }
}

# Converting the Departure.Delay.in.Minutes if Delay is greater than 5 mins to "Yes" or "No"

for (i in 1:length(df$Departure.Delay.in.Minutes))
{
  if (df$Departure.Delay.in.Minutes[i] > 5)
  {
    df$DepartDelayof5min[i] <- "Yes"
  }
  else
  {
    df$DepartDelayof5min[i] <- "No"
  }
}


# Coverting the columns to factors

converttostring <- function(vec)
{
  vec <- trimws(as.character(vec))
  vec <- as.factor(vec)
}

# Converting to Factors

df$Airline.Status <- converttostring(df$Airline.Status)
df$Gender <- converttostring(df$Gender)
df$Type.of.Travel <- converttostring(df$Type.of.Travel)
df$Class <- converttostring(df$Class)
df$Partner.Name <- converttostring(df$Partner.Name)
df$Flight.cancelled <- converttostring(df$Flight.cancelled)
df$Price.Sensitivity <- converttostring(df$Price.Sensitivity)
df$ArrivalDelayof5min <- converttostring(df$ArrivalDelayof5min)
df$DepartDelayof5min <- converttostring(df$DepartDelayof5min)
df$pr_pa_dt <- converttostring(df$pr_pa_dt)
```

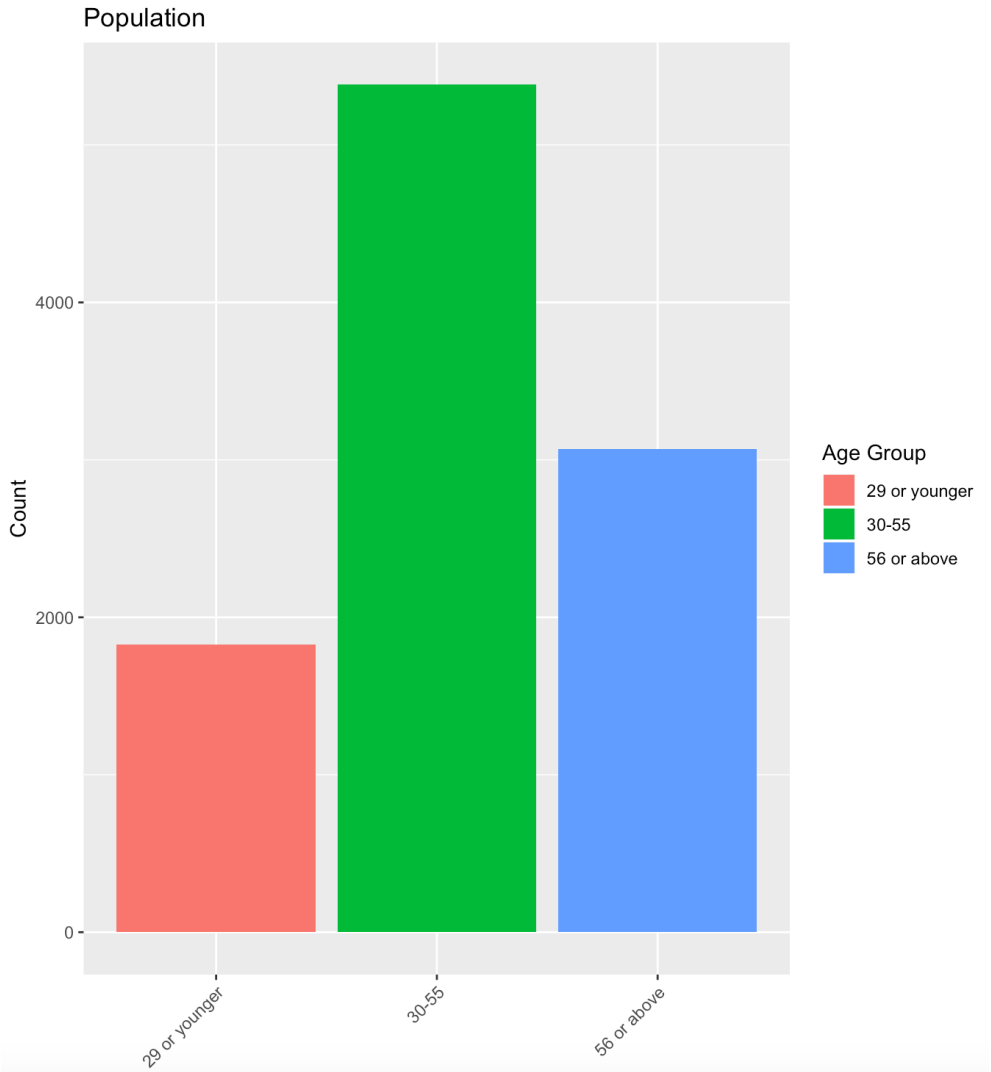
```r
# creating 3 levels for age
for (i in 1:length(data2$Age))
{
  if (data2$Age[i] >=15 & data2$Age[i] <= 29)
  {
    data2$Age_group[i] <- "Age between 15 and 29"
  }
  else if (data2$Age[i] >=30 & data2$Age[i] <= 54)
  {
    data2$Age_group[i] <- "Age between 30 and 54"
  }
  else
  {
    data2$Age_group[i] <- "Age above 54"
  }
}
data2$Age_group <- as.factor(data2$Age_group)
# removing unwanted columns
data2 <- data2[,c(-1,-2,-4,-7,-15,-16,-17,-18,-21,-22,-23,-24,-25,-27,-28,-29,-30,-31,-32)]
data2 <- data2[,-8]
data2 <- data2[,-15]
colnames(data2)
```

BUSINESS QUESTIONS:

1. How different Age groups affect Likelihood to Recommend?

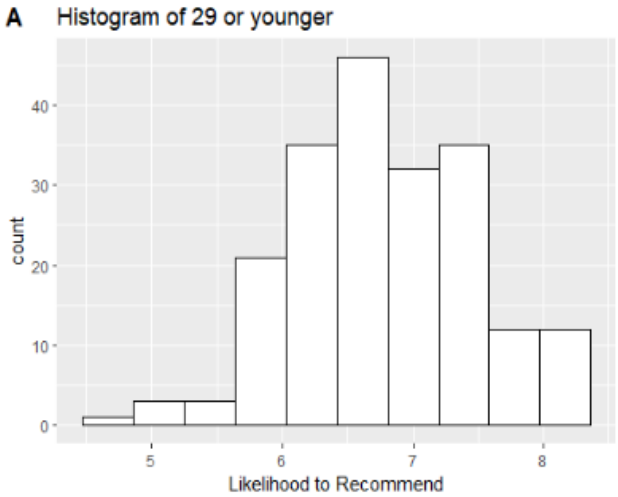
   First, we have divided people in to three age groups: 29 or younger, 30-55, and 56 or above. We have created Age group vs likelihood to recommend graph for you. From this graph, we can tell that in the promoter group, the people from 30 – 55 years old are more likely to be a promoter while people from 56 years old or above group are more likely to become a detractor.
   However, when we look at the population from each group, we have discovered that the population from 30 – 55 years old is larger than any other groups. The younger group has 1829 people, the median age group has 5382 people, the older group has 3071 people. Hence, we decided to go further to investigate if this results still hold.
   First, we generated a boxplot to investigate the raw data. From here, we can tell that the median grade for age group 30 -55 years old is higher than any other groups. The age group from 56 or above is indeed giving lower average grade.

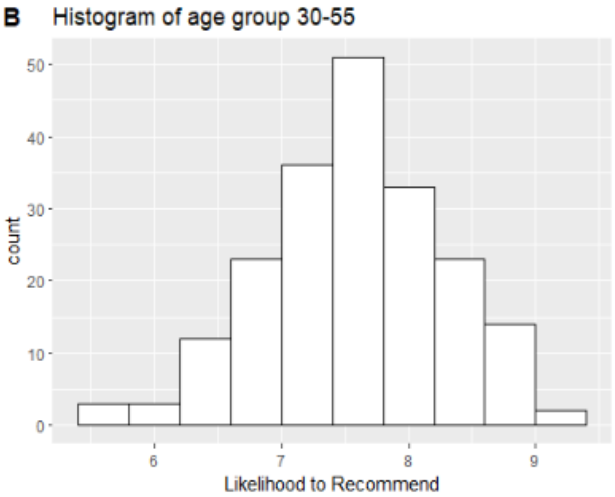


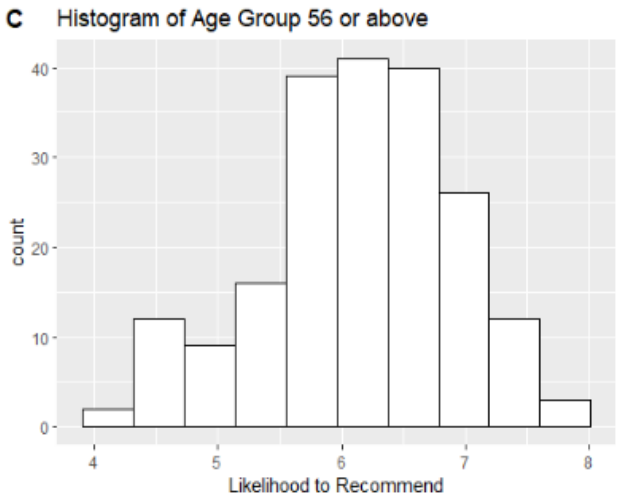



```
> summary(df$Likelihood.to.recommend[df$ageGroup == "29 or younger"])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   5.000   7.000   6.764   9.000  10.000
> summary(df$Likelihood.to.recommend[df$ageGroup == "30-55"])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   7.000   8.000   7.605   9.000  10.000
> summary(df$Likelihood.to.recommend[df$ageGroup == "56 or above"])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   4.000   6.000   6.056   8.000  10.000
```

To further verify this, we were sampling from each group and take average of the grades. We repeat this process 200 times and then we generate the histogram for different age group. And now, the median of each histogram should reveal the true mean of the group. The median score for young age group is 6.8, median age group is 7.7 and old age group is 6.2. Hence, we can say that the people in the median age group tends to give a higher score whereas the people from the older age group tends to give a lower score.

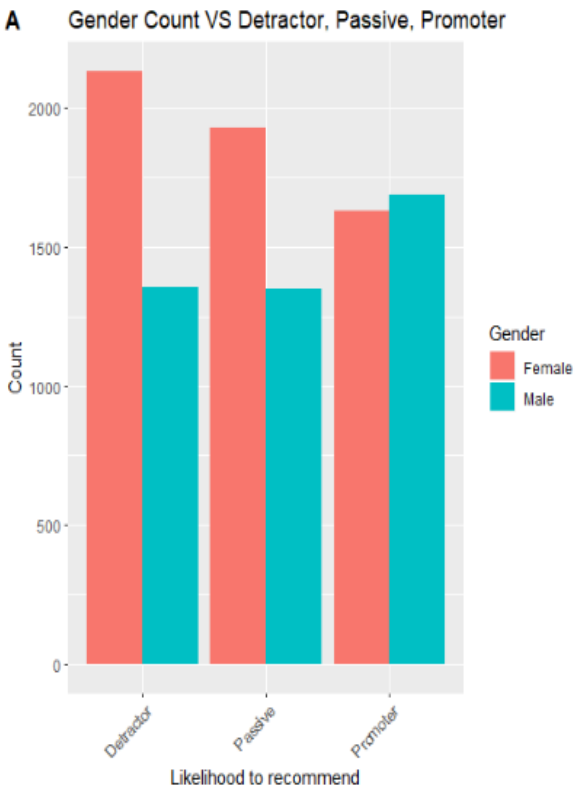


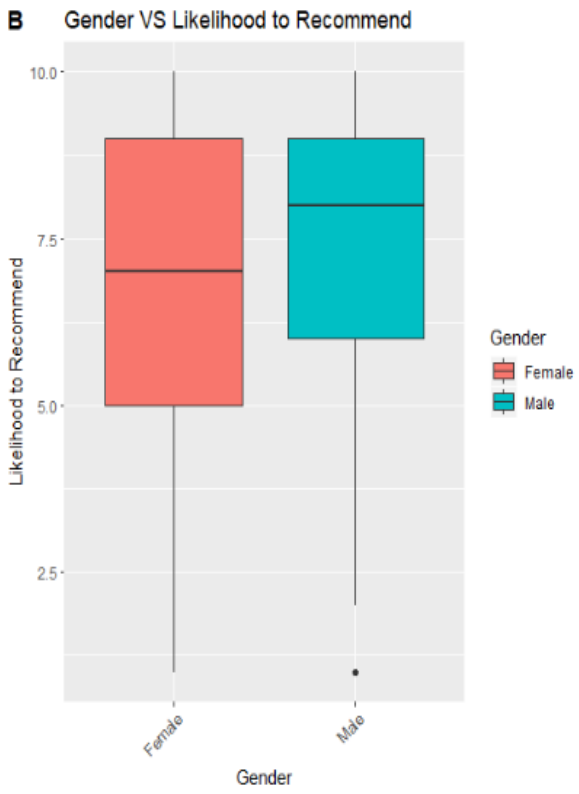


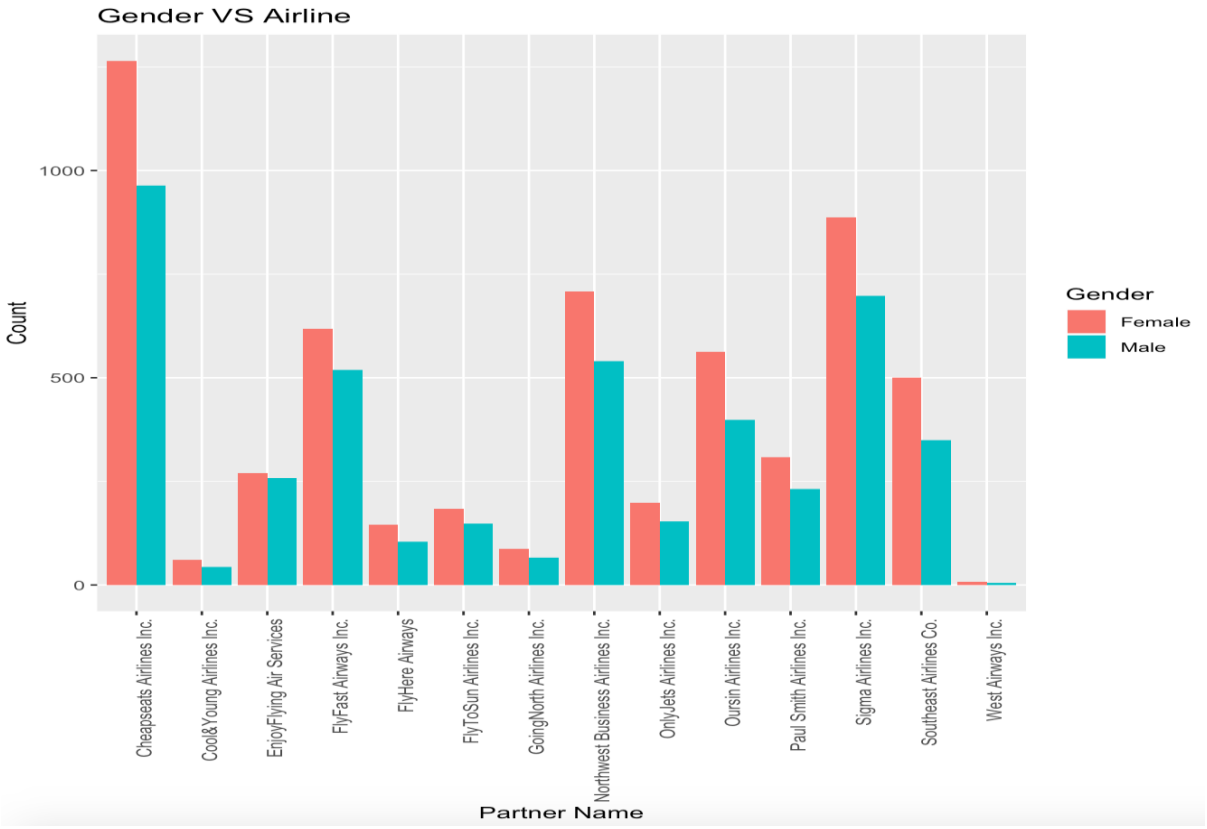


2. Which Gender is more likely to be a Detractor/ Promoter?

   We have investigated the population in both groups. The female has 5804 people, male has 4478 people. From the graph at the left, we see that females are more likely to be a detractor or passive. Males are more likely to be a promoter even though the population for Female is larger than Male. By looking at the picture at the right, which is boxplot. The median score for males is higher than females.

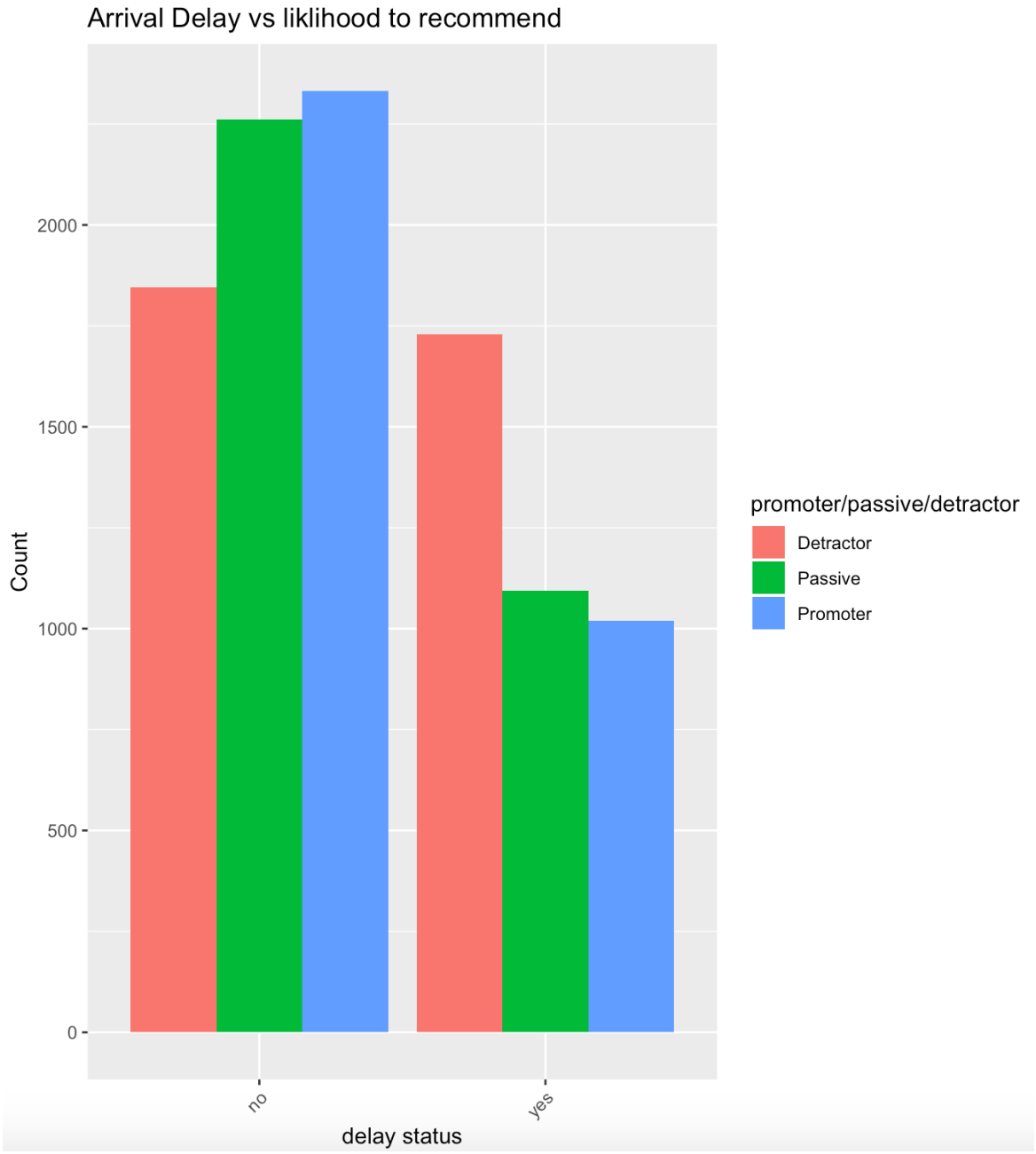

A
Gender Count VS Detractor, Passive, Promoter
Count
2000
1500
1000
500
0
Detractor
Passive
Promoter
Likelihood to recommend
Gender
Female
Male


B
Gender VS Likelihood to Recommend
Likelihood to Recommend
10.0
7.5
5.0
2.5
Female
Male
Gender
Gender
Female
Male


Gender VS Airline
Count
1000
500
0
Cheapseats Airlines Inc.
Cool&Young Airlines Inc.
EnjoyFlying Air Services
FlyFast Airways Inc.
FlyHere Airways
FlyToSun Airlines Inc.
GoingNorth Airlines Inc.
Northwest Business Airlines Inc.
OnlyJets Airlines Inc.
Oursin Airlines Inc.
Paul Smith Airlines Inc.
Sigma Airlines Inc.
Southeast Airlines Co.
West Airways Inc.
Partner Name
Gender
Female
Male

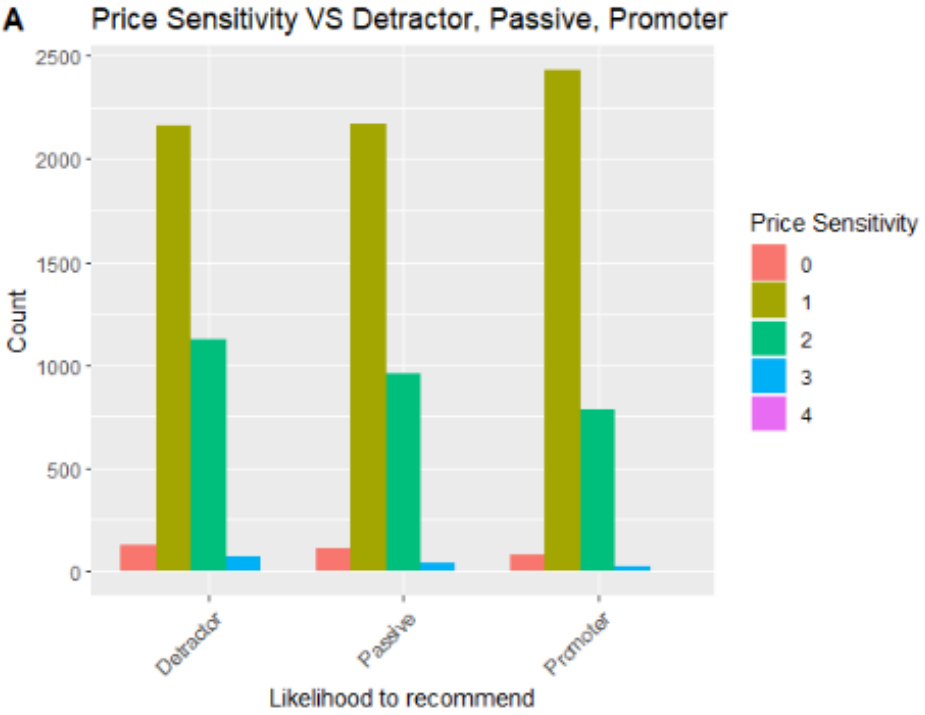
3. How does Arrival Delay affect Likelihood to Recommend?

   Does the arrival delay affect likelihood to recommend? First, for those flights delayed over 5 minutes, we say they are delayed. Otherwise, they are not. From the graph, we can easily see that the delayed flights are having more detractors and those are not delayed would more likely to have more promoters.



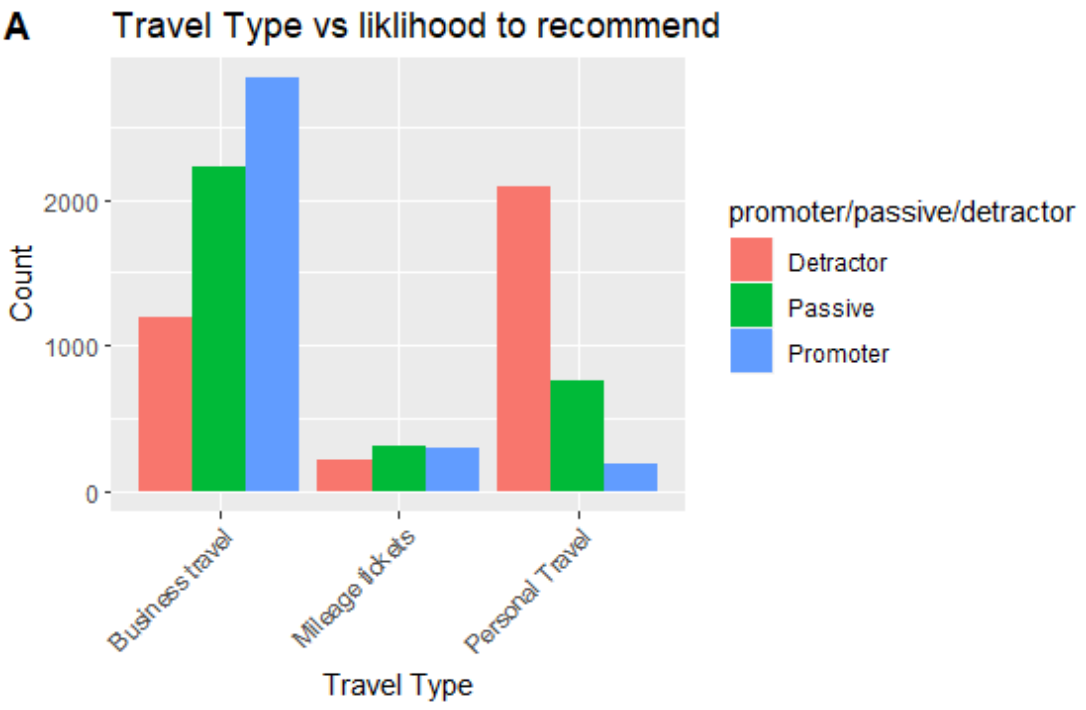
4. How does Price Sensitivity affect Likelihood to Recommend?

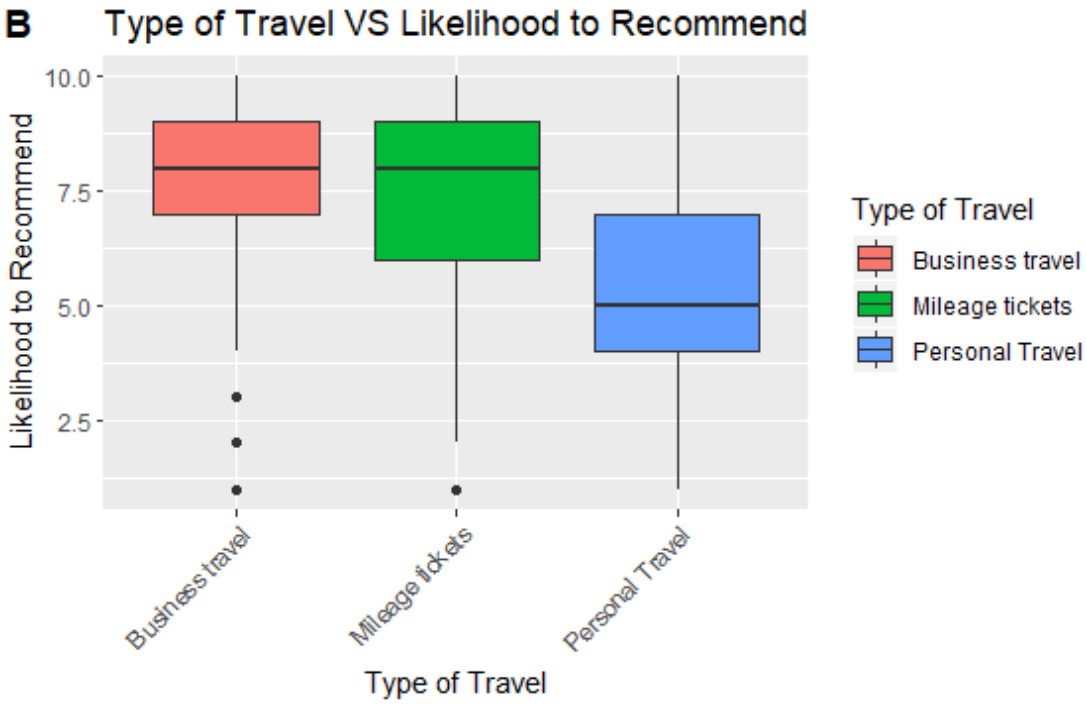
   It is a boxplot for the people have different price sensitivity. It seems like people having sensitivity level of 1 and 4 would more likely to give higher chance to recommend because they are having higher median score, third quantile and lower first quantile.



5. How does Type of travel affect Likelihood to Recommend?

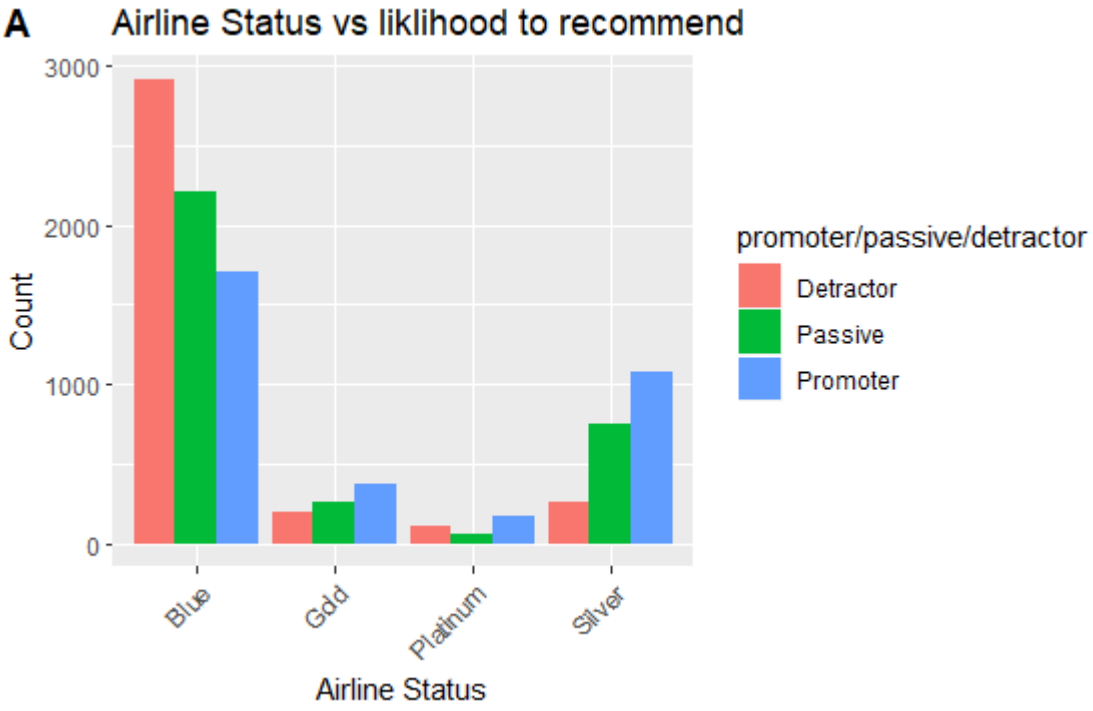
   Business travelers are more likely to become a promoter and less likely to become a detractor. However, the personal travelers are extreme likely to become a detractor and very unlikely to become a promoter. By looking at the boxplot. Business travelers are giving much higher score than the personal travelers in general.

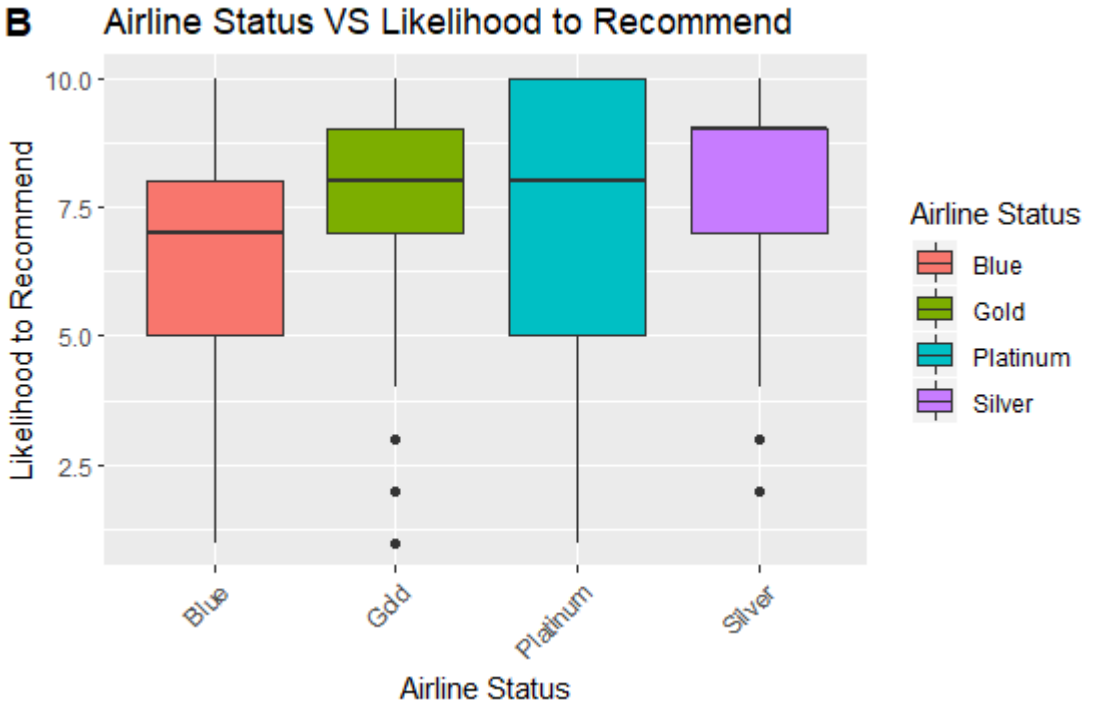




6. Which Airline Status has maximum Detractors?

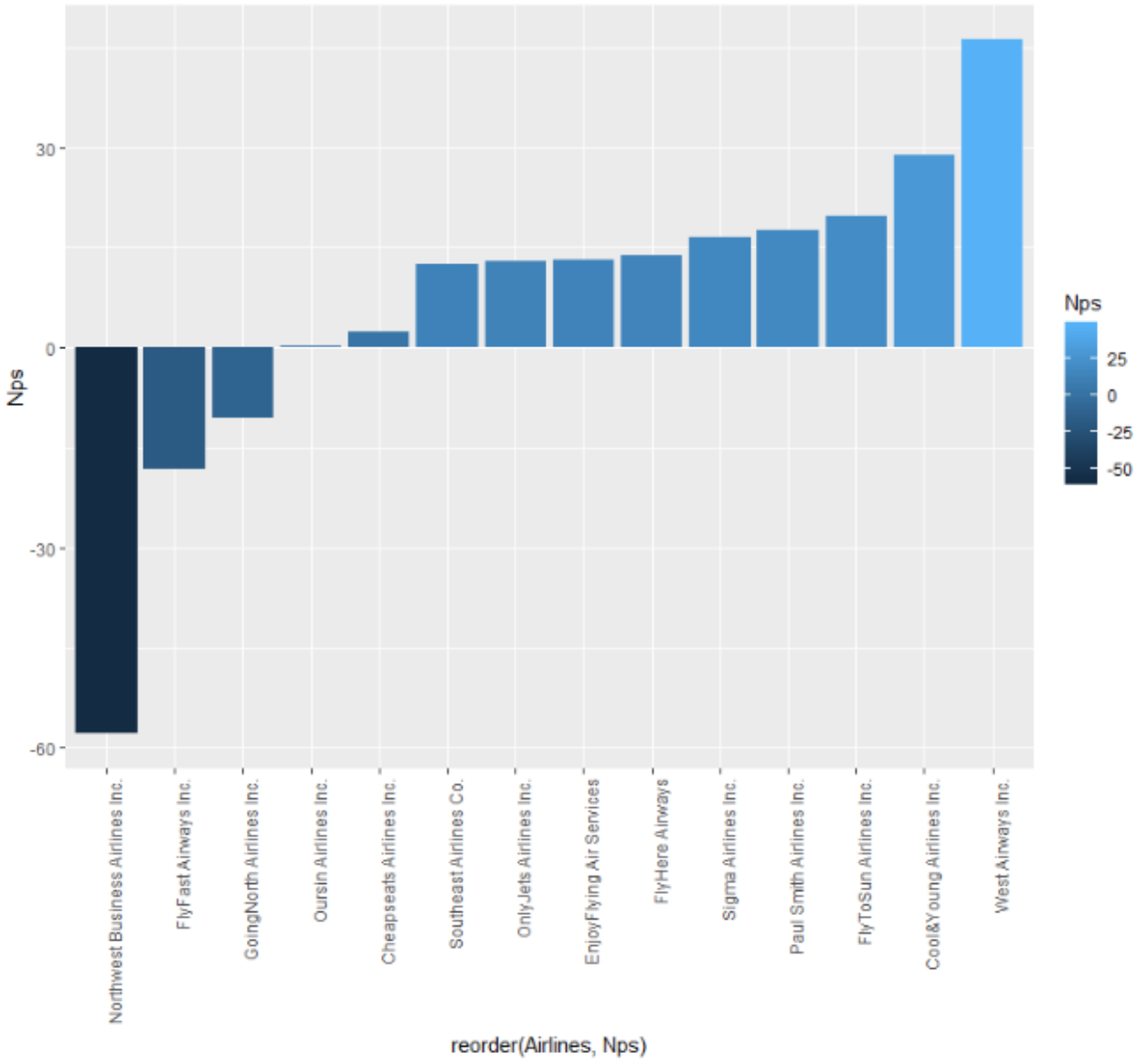
We plotted a boxplot and bar plot between Airline status and Likelihood to recommend. With analysis of this plot we found that the silver airline status, have highest median
value of Likelihood to recommend while blue airline has the lowest.





7. Why we have considered Fly-Fast Airways Inc. and Northwest Business Airlines?

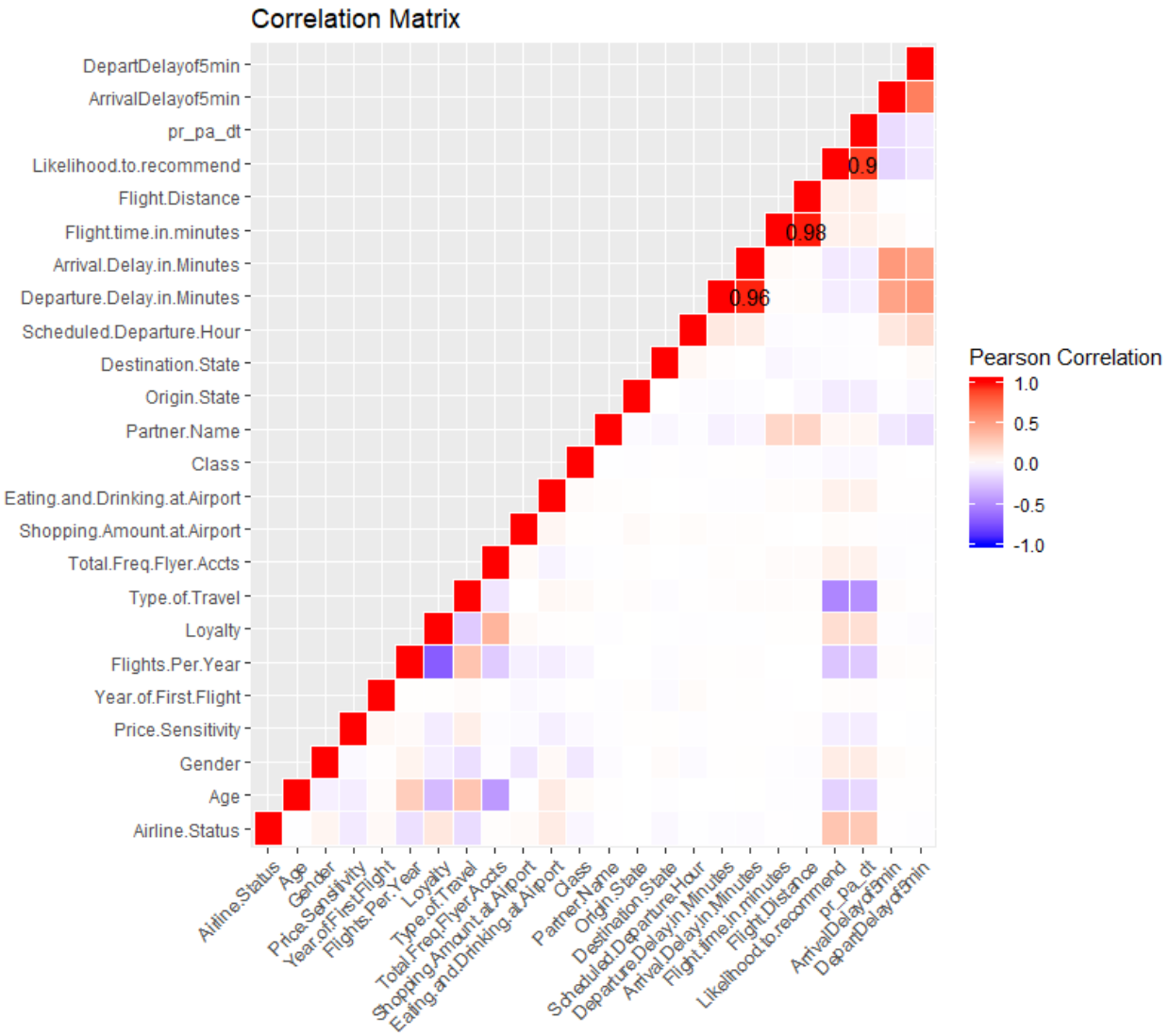
   We have considered the Fly-Fast Airways Inc. and Northwest Business Airlines which have the least NPS score out of all the 13 partner airlines of the South East Airlines.



SIGNIFICANT VARIABLES

Correlation Matrix

Correlation matrix is used to find any dependencies between multiple variables at the same time. It gives a table filled with correlation coefficients with every variable. Correlation coefficient measures the how much one variable variance that can be explained by another variable. The correlation coefficients vary between -1 to 1.
If two variables are positively correlated (same direction) then coefficient is 1 and if two variables are negatively correlated (opposite direction) then coefficient is -1. Correlation of 0 means there is no relationship at all between the two variables.



Based on the correlation coefficients of these variables with likelihood to recommend, we have picked up these variables.

Significant Attributes	Significant Attributes
Airline Status	Eating and Drinking at Airport
Gender	Class
Price Sensitivity	Origin State
Flights Per Year	Destination State
Loyalty	Flight Distance
Type of Travel	Arrival Delay
Total Freq Flyer Accts	Age Group

```
#-----------------------------------Correlation Matrix---------------------------------

library(data.table)
corr <- df1 %>% sapply(., as.numeric) %>% as.data.table()
corr <- cor(corr, use = 'pairwise.complete.obs')
corr[upper.tri(corr)] <- NA
corr <- melt(corr, na.rm = T) %>% as.data.table() %>% setorder(-value)
corr$text <- ifelse(abs(corr$value) >= .8 & corr$value != 1, round(corr$value, 2), '')

ggplot(data = corr, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile(color = 'white') +
  geom_text(aes(label = text)) +
  scale_fill_gradient2(low = 'blue', high = 'red', mid = 'white',
                       midpoint = 0, limit = c(-1, 1),
                       name = 'Pearson Correlation') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = 'Correlation Matrix')
```

LINEAR MODELLING

Linear modelling is used to predict the value of an outcome variable Y based on one or more input predictor variables X. We can obtain a relationship between dependent variable and independent variable in form of a mathematical equation by one or more dependent variable. We have performed the linear modelling on the entire dataset by forward stepwise regression (one-by-one) modelling. In this model we add one-by-one variable and checked the Adjusted R-Squared value. If the Adjusted R-Squared value is increased, then that is a significant variable else that not included in the model. After performing this on entire dataset we are left with 14 variables which is the highest Adjusted R-Squared model after analyzing 30 models.

- Airline Status
- Gender
- Price Sensitivity
- Flights Per Year
- Loyalty
- Type of Travel
- Total Freq Flyer Accts
- Eating and Drinking at Airport
- Class
- Origin State
- Destination State
- Flight Distance
- Arrival Delay
- Age Group

The above factors are independent variables and likelihood to recommend is the dependent variable.

Code:

```
#----------------------------------Linear Modelling-------------------------------------

df1_Linear <- lm(formula = Likelihood.to.recommend~Airline.Status+Gender+Loyalty+Eating.and.Drinking.at.Airport+
                 Flight.Distance+Age+Price.Sensitivity+Flights.Per.Year+Type.of.Travel+Origin.State+Destination.State+
                   Class+ArrivalDelayof5min,data = df1)

summary(df1_Linear)
```

Output:

```
Residual standard error: 1.747 on 9970 degrees of freedom
Multiple R-squared:  0.4551,    Adjusted R-squared:  0.4489
F-statistic: 73.69 on 113 and 9970 DF,  p-value: < 2.2e-16
```

We can see that the Adjusted R-Squared value is 0.4489, the variance in the dependent variable is explained by all independent variables present in the model. 45% of the variability in likelihood to recommend are explained by the independent variables.

Linear Modelling on Northwest Business Airlines:

We conducted linear modelling on two airlines, the first one being Northwest Business Airlines with the below factors:

- Airline Status
- Gender
- Price Sensitivity
- Flights Per Year
- Loyalty
- Type of Travel
- Total Freq Flyer Accts
- Eating and Drinking at Airport
- Class
- Origin State
- Destination State
- Flight Distance
- Arrival Delay
- Age Group

For this dataset, the below factors provide the best Adjusted R-Squared model:

- Airline Status
- Gender
- Price Sensitivity
- Type of Travel
- Total Freq Flyer Accts
- Eating and Drinking at Airport
- Class
- Origin State
- Destination State
- Arrival Delay

Code:

```
# Northwest business airlines Linear Model

northwest_bus_air1 <- df1[str_trim(df$Partner.Name)=="Northwest Business Airlines Inc.",]

df1_Linear_north <- lm(formula = Likelihood.to.recommend~Airline.Status+Gender+Loyalty+Eating.and.Drinking.at.Airport+
                         Flight.Distance+Type.of.Travel+Age+Price.Sensitivity+Origin.State+Destination.State+
                         Flights.Per.Year+Class+ArrivalDelayof5min,data = northwest_bus_air1)
summary(df1_Linear_north)
```

Output:

```
Residual standard error: 1.638 on 1139 degrees of freedom
Multiple R-squared:  0.4864,    Adjusted R-squared:  0.449
F-statistic:    13 on 83 and 1139 DF,  p-value: < 2.2e-16
```

We can see that the Adjusted R-Squared value is 0.4490, the variance in the dependent variable is explained by all independent variables present in the model. 45% of the variability in likelihood to recommend are explained by the independent variables.

Linear Modelling on FlyFast Airways:

We conducted linear modelling on two airlines, the second one being Flyfast Airlines with the below factors:

- Airline Status
- Gender
- Price Sensitivity
- Flights Per Year
- Loyalty
- Type of Travel
- Total Freq Flyer Accts
- Eating and Drinking at Airport
- Class
- Origin State
- Destination State
- Flight Distance
- Arrival Delay
- Age Group

For this dataset, the below factors provide the best Adjusted R-Squared model:

- Airline Status
- Price Sensitivity
- Type of Travel
- Eating and Drinking at Airport
- Class
- Origin State
- Destination State
- Arrival Delay

Code:

```
# FlyFast Airways Linear Model

flyfast_air1 <- df1[str_trim(df$Partner.Name)=="FlyFast Airways Inc." & df$Flight.cancelled=='No',]

df1_Linear_flyfast <- lm(formula = Likelihood.to.recommend~Airline.Status+Gender+Loyalty+Eating.and.Drinking.at.Airport+
                         Flight.Distance+Type.of.Travel+Age+Price.Sensitivity+Flights.Per.Year+Class+Origin.State+
                          Destination.State+ArrivalDelayof5min,data = flyfast_air1)

summary(df1_Linear_flyfast)
```

Output:

```
Residual standard error: 1.752 on 1000 degrees of freedom
Multiple R-squared:  0.515,     Adjusted R-squared:  0.4718
F-statistic: 11.93 on 89 and 1000 DF,  p-value: < 2.2e-16
```

We can see that the Adjusted R-Squared value is 0.4718, the variance in the dependent variable is explained by all independent variables present in the model. 47% of the variability in likelihood to recommend are explained by the independent variables.

ASSOCIATION RULES

Association rules are if/then statements for discovering interesting relationships between seemingly unrelated data in a large databases or other information repository. Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently an itemset occurs in a transaction. Given a set of transactions, we can find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction.

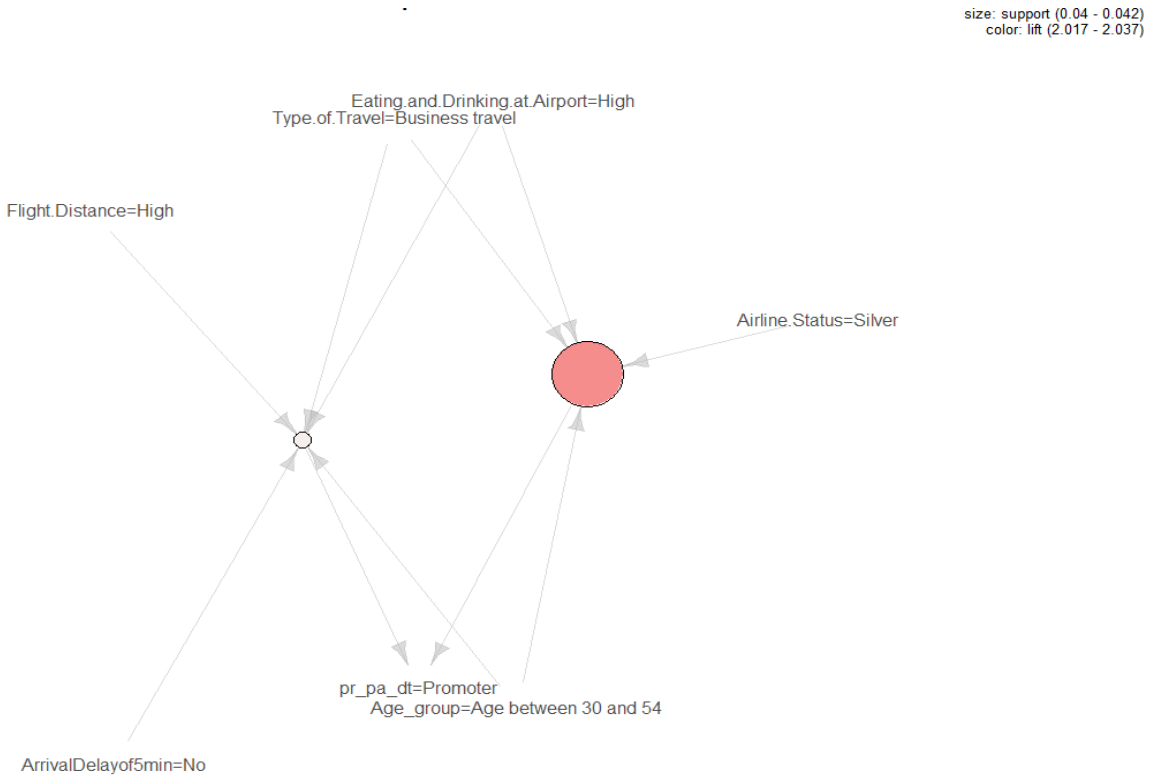
We have used association rules mining to get important insights from the data by analyzing the rules with the highest lift value. Based on the analysis, we have found the following results:

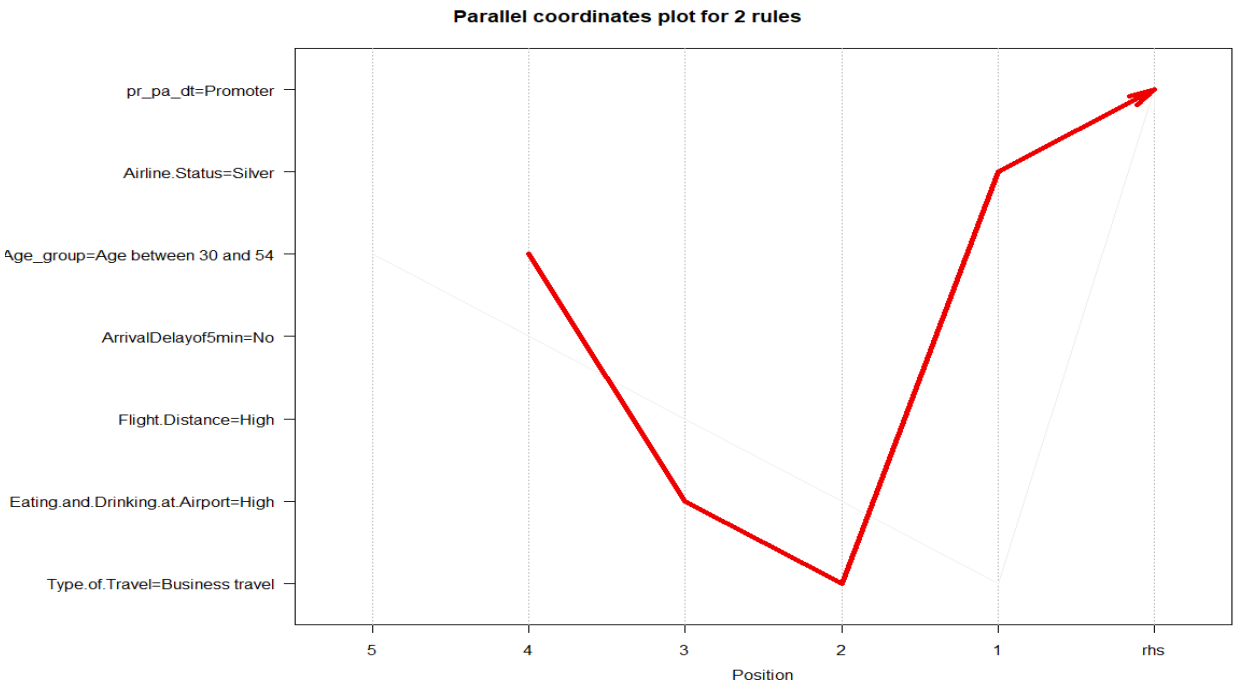
1. For the entire dataset when RHS contains Promoters and LHS is default (the other 14 variables)

LHS	RHS	support	confidence	lift	cour
All	All	All	All	.	All
{Airline.Status=Silver,Type.of.Travel=Business travel,Eating.and.Drinking.at.Airport=High,Age_group=Age between 30 and 54}	{pr_pa_dt=Promoter}	0.042	0.670	2.037	425.0
{Type.of.Travel=Business travel,Eating.and.Drinking.at.Airport=High,Flight.Distance=High,ArrivalDelayof5min=No,Age_group=Age between 30 and 54}	{pr_pa_dt=Promoter}	0.040	0.664	2.017	405.0

So from the first rule we can infer that for passengers between 30-54 years of age spending a high amount at the airport on food and drinks travelling for business purposes on an airline belonging to the silver category are more likely to be a promoter.

From the second rule, passengers between 30-54 years spending a high amount at the airport on food and drinks travelling for business purposes on a flight where the travel distance is high and no arrival delay are more likely to become promoters.


size: support (0.04 - 0.042)
color: lift (2.017 - 2.037)
Eating.and.Drinking.at.Airport=High
Type.of.Travel=Business travel
Flight.Distance=High
Airline.Status=Silver
pr_pa_dt=Promoter
Age_group=Age between 30 and 54
ArrivalDelayof5min=No

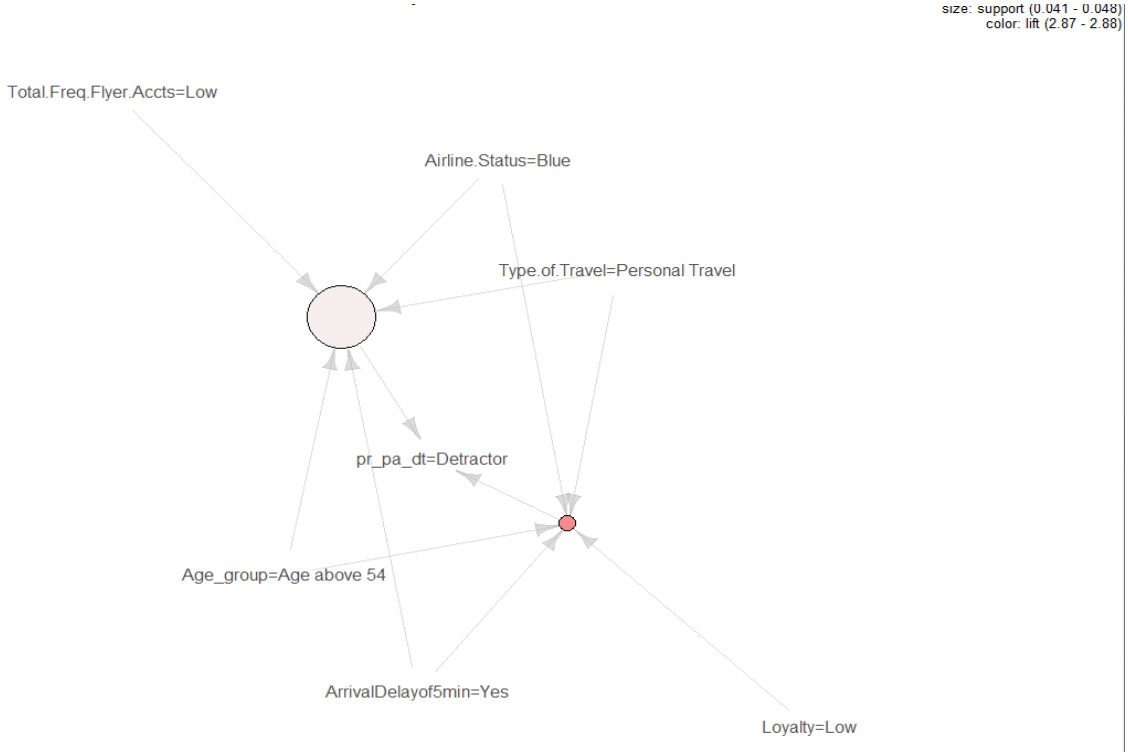

Parallel coordinates plot for 2 rules
pr_pa_dt=Promoter
Airline.Status=Silver
Age_group=Age between 30 and 54
ArrivalDelayof5min=No
Flight.Distance=High
Eating.and.Drinking.at.Airport=High
Type.of.Travel=Business travel
5
4
3
2
1
rhs
Position

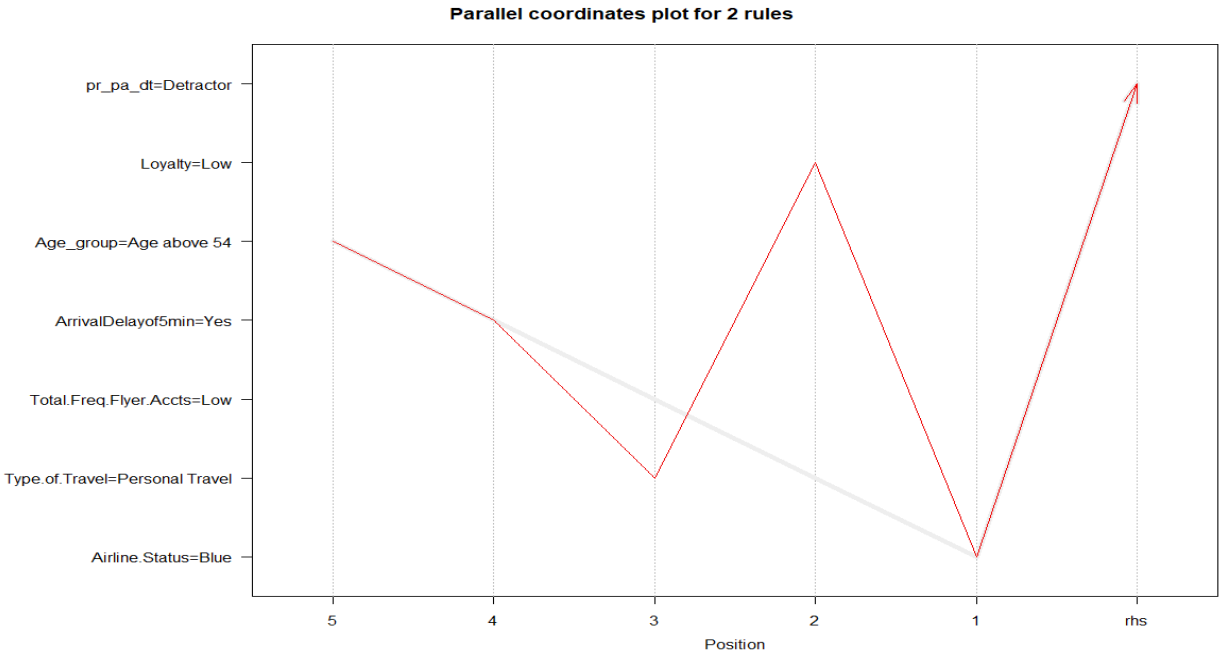
2. For the entire dataset when RHS contains Detractors and LHS is default (the other 14 variables)

LHS	RHS	support	confidence	lift	count
All	All	All	All	All	All
{Airline.Status=Blue,Loyalty=Low,Type.of.Travel=Personal Travel,ArrivalDelayof5min=Yes,Age_group=Age above 54}	{pr_pa_dt=Detractor}	0.041	0.995	2.880	413.000
{Airline.Status=Blue,Type.of.Travel=Personal Travel,Total.Freq.Flyer.Accts=Low,ArrivalDelayof5min=Yes,Age_group=Age above 54}	{pr_pa_dt=Detractor}	0.048	0.992	2.870	479.000

From the first rule, passengers above 54 years of age with a low loyalty index, travelling for personal purposes on a flight belonging to the blue status and delayed by more than 5 minutes are more likely to be detractors.

From the second rule, a passenger above 54 years with less frequent flyer accounts travelling for personal purposes on a flight belonging to the blue status are more likely to be detractors.



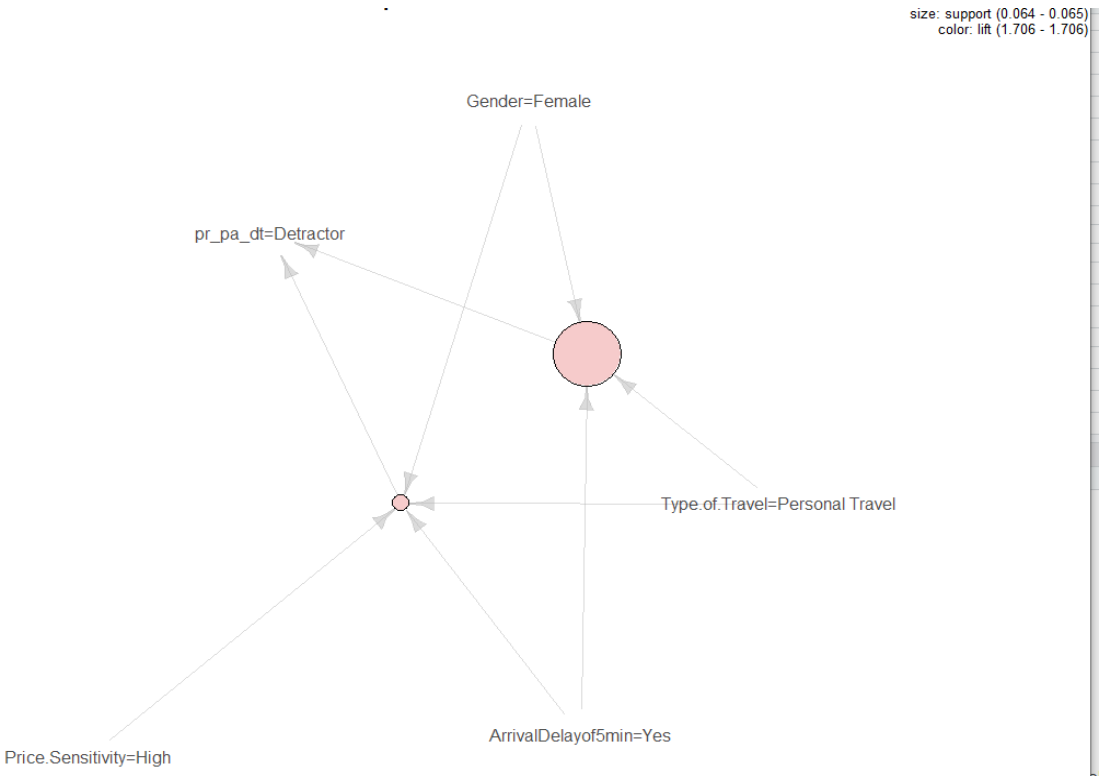


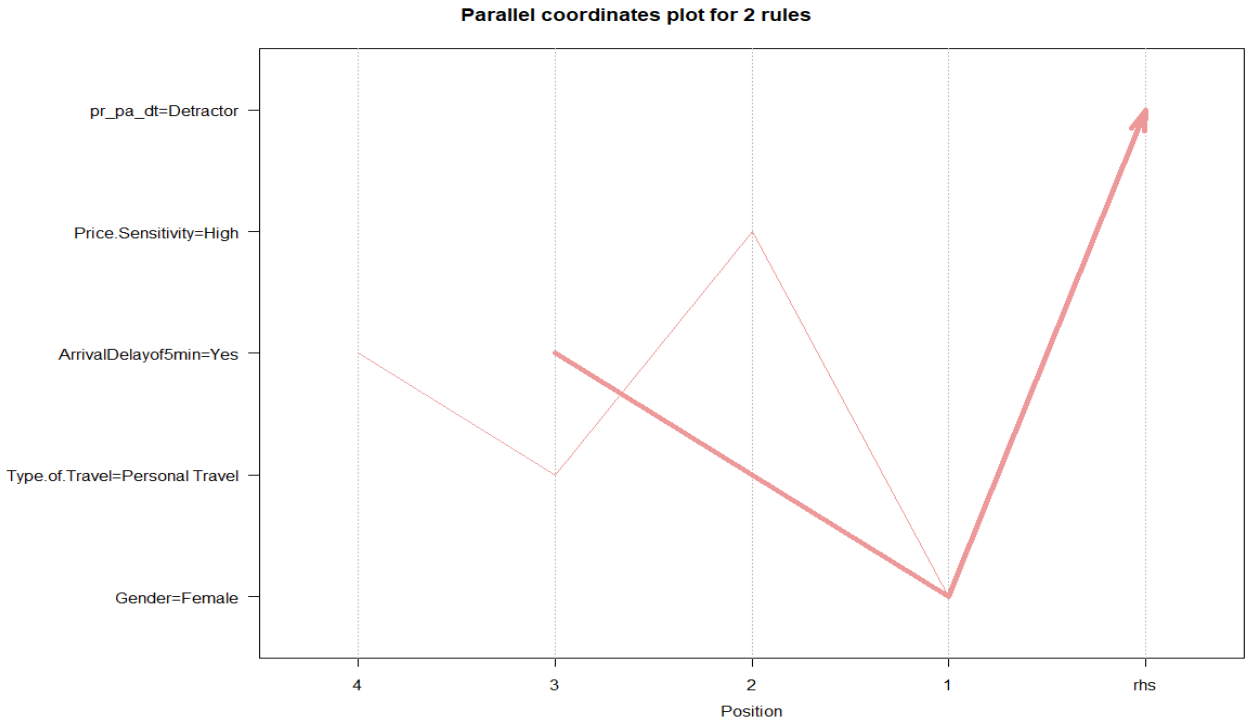
3. For the dataset containing data only for Northwest airlines where RHS has detractors and the LHS is default (the other 14 variables)

LHS	RHS	support	confidence	lift	count
All	All	All	All	All	All
{Gender=Female,Type.of.Travel=Personal Travel,ArrivalDelayof5min=Yes}	{pr_pa_dt=Detractor}	0.065	1.000	1.706	80.000
{Gender=Female,Price.Sensitivity=High,Type.of.Travel=Personal Travel,ArrivalDelayof5min=Yes}	{pr_pa_dt=Detractor}	0.064	1.000	1.706	78.000

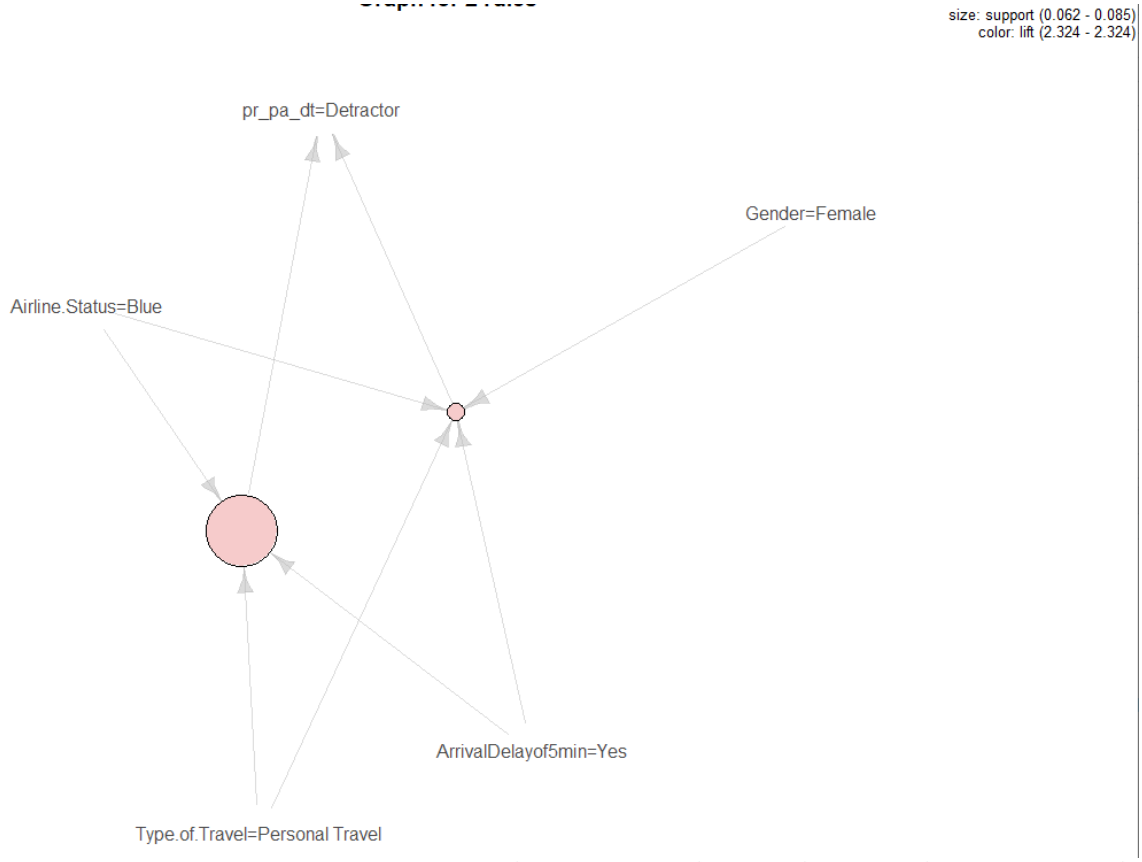
From the first rule, female passengers travelling for personal purposes and flight arriving more than 5 minutes late are more likely to be detractors for Northwest Business Airlines.

From the second rule, female passengers who are highly price sensitive travelling for personal purposes and flight arriving more than 5 minutes late are likely to be detractors for Northwest Business Airlines.


size: support (0.064 - 0.065)
color: lift (1.706 - 1.706)
Gender=Female
pr_pa_dt=Detractor
Type.of.Travel=Personal Travel
Price.Sensitivity=High
ArrivalDelayof5min=Yes

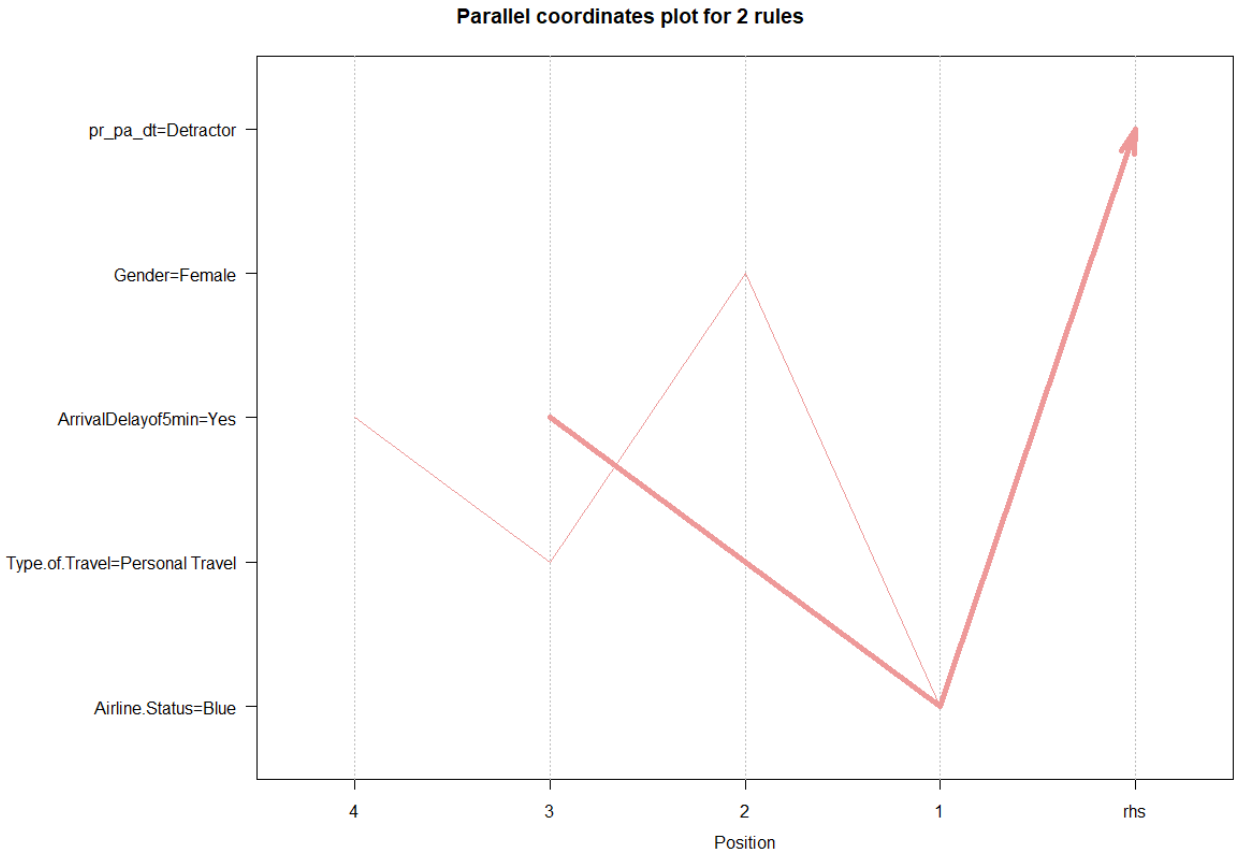

Parallel coordinates plot for 2 rules
pr_pa_dt=Detractor
Price.Sensitivity=High
ArrivalDelayof5min=Yes
Type.of.Travel=Personal Travel
Gender=Female
4
3
2
1
rhs
Position

4. For the dataset containing data only for Flyfast airlines where RHS has detractors and the LHS is default (the other 14 variables)



LHS	RHS	support	confidence	lift	count
All	All	All	All	All	All
{Airline.Status=Blue,Type.of.Travel=Personal Travel,ArrivalDelayof5min=Yes}	{pr_pa_dt=Detractor}	0.085	1.000	2.324	93.000
{Airline.Status=Blue,Gender=Female,Type.of.Travel=Personal Travel,ArrivalDelayof5min=Yes}	{pr_pa_dt=Detractor}	0.062	1.000	2.324	68.000

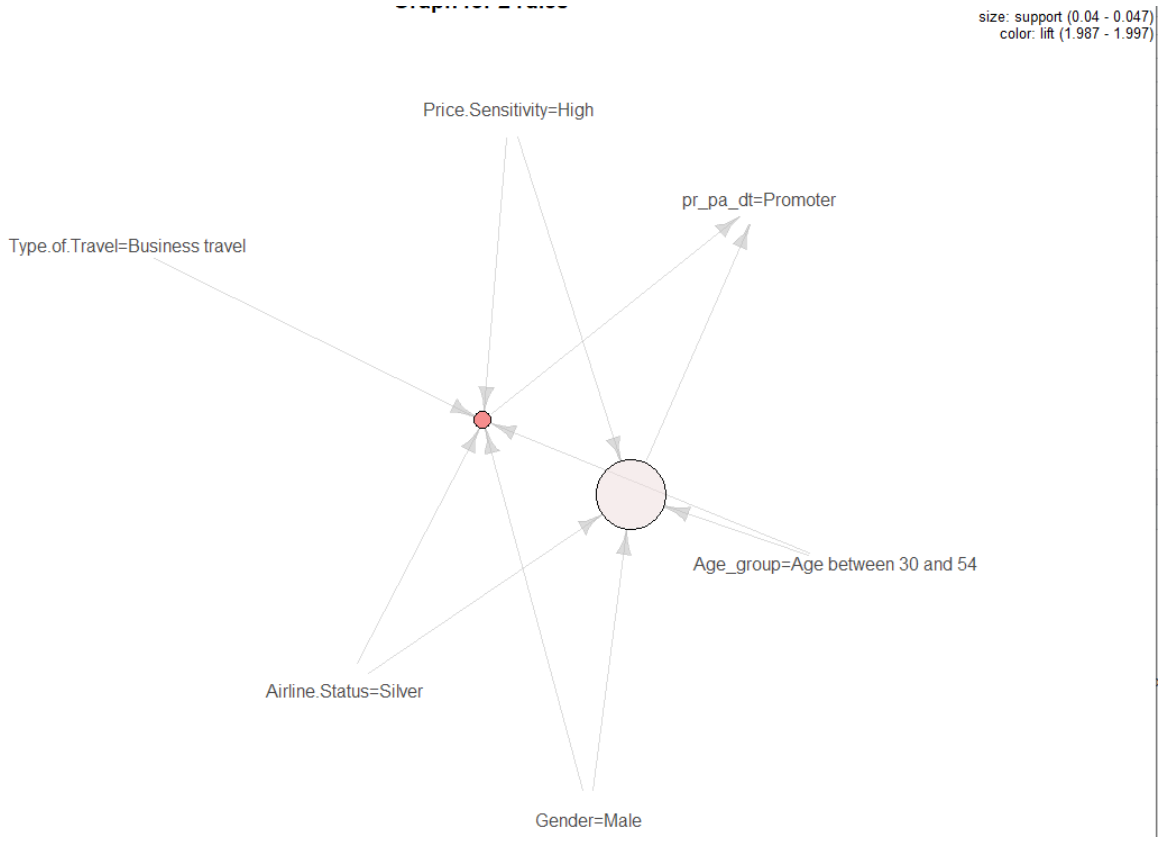
From both the rules, passengers especially female passengers travelling on Flyfast airways (which belongs to the blue category) for personal purposes and an arrival delay of more than minutes are more likely to be detractors.

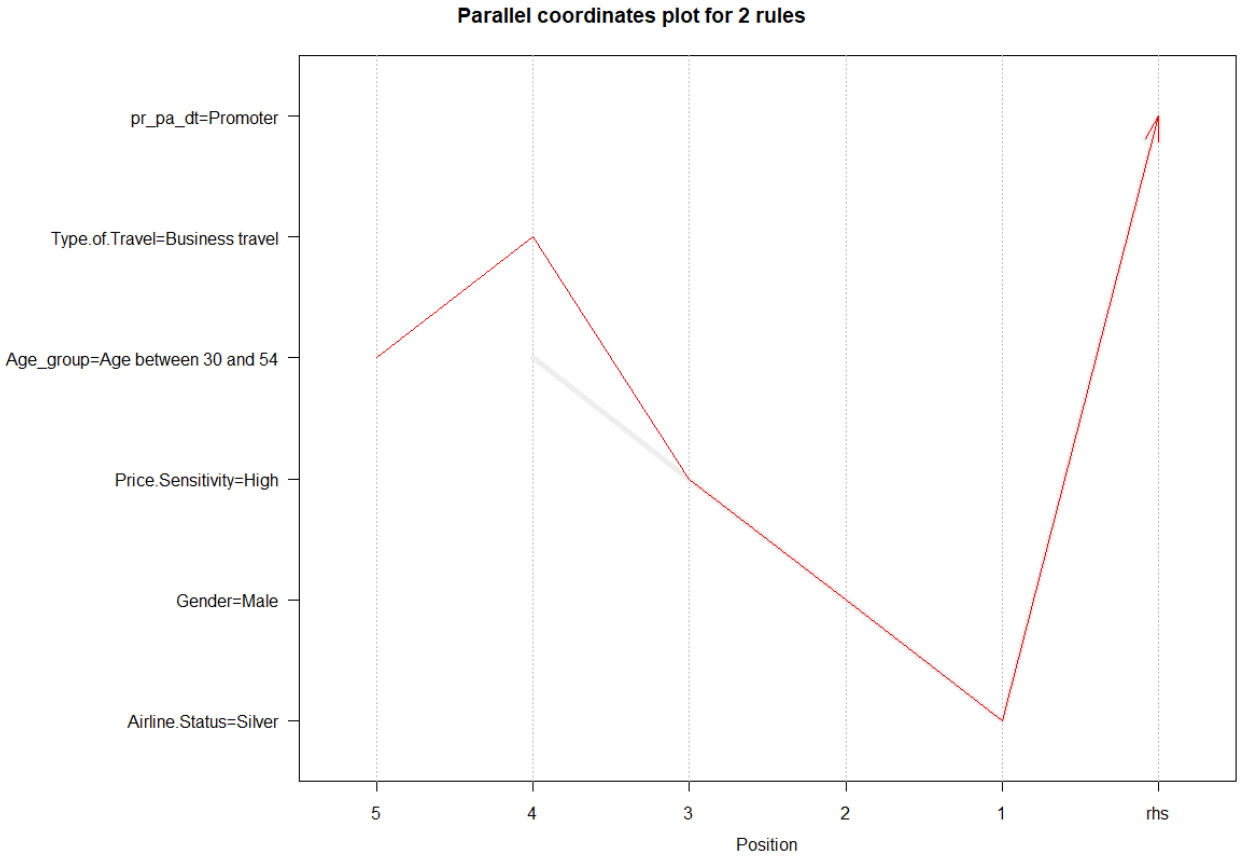


5. For the dataset containing data only for Sigma airlines where RHS has promoters and the LHS is default (the other 14 variables)

LHS	RHS	support	confidence	lift	count
All	All	All	All	.	All
{Airline.Status=Silver,Gender=Male,Price.Sensitivity=High,Type.of.Travel=Business travel,Age_group=Age between 30 and 54}	{pr_pa_dt=Promoter}	0.040	0.863	1.997	63.000
{Airline.Status=Silver,Gender=Male,Price.Sensitivity=High,Age_group=Age between 30 and 54}	{pr_pa_dt=Promoter}	0.047	0.859	1.987	73.000

From both the rules, male passengers between 30-54 years who are highly price sensitive travelling for business purposes are likely to be promoters.


size: support (0.04 - 0.047)
color: lift (1.987 - 1.997)
Price.Sensitivity=High
pr_pa_dt=Promoter
Type.of.Travel=Business travel
Age_group=Age between 30 and 54
Airline.Status=Silver
Gender=Male


Parallel coordinates plot for 2 rules
pr_pa_dt=Promoter
Type.of.Travel=Business travel
Age_group=Age between 30 and 54
Price.Sensitivity=High
Gender=Male
Airline.Status=Silver
5
4
3
2
1
rhs
Position

SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) is a supervised learning technique. The reason SVMs are considered a supervised learning technique is that we train the algorithm on an initial set of data (the supervised phase) and then we test it out on a brand-new set of data.

In SVM we have considered variables which we got from the output of Correlation Matrix and Linear Modelling to cross validate the significance of the variables.

The output will give accuracy rate, which will signify how good the selected variables for forecasting are.

SVM on entire dataset:

```
# creating random sample for training and test dataset
randIndex1 <- sample(1:dim(data2)[1])
cutPoint2_3 <- floor(2 * dim(data2)[1]/3)
trainData_1 <- data2[randIndex1[1:cutPoint2_3],]
testData_1 <- data2[randIndex1[(cutPoint2_3+1):dim(data2)[1]],]

# running support vector machine
svmOutput_1 <- ksvm(pr_pa_dt ~.,data=trainData_1,kernel="rbfdot",kpar="automatic",scale=FALSE,C = 5,cross = 3,prob.model = TRUE)
svmOutput_1

#performing prediction
svmPred_1 <- predict(svmOutput_1, testData_1)

# for accuracy
confusionMatrix(svmPred_1,testData_1$pr_pa_dt)
```

Output:

```
> # creating random sample for training and test dataset
> randIndex1 <- sample(1:dim(data2)[1])
> cutPoint2_3 <- floor(2 * dim(data2)[1]/3)
> trainData_1 <- data2[randIndex1[1:cutPoint2_3],]
> testData_1 <- data2[randIndex1[(cutPoint2_3+1):dim(data2)[1]],]
> # running support vector machine
> svmOutput_1 <- ksvm(pr_pa_dt ~.,data=trainData_1,kernel="rbfdot",kpar="automatic",scale=FALSE,C = 5,cross = 3,prob.model = TRUE)
> svmOutput_1
Support Vector Machine object of class "ksvm"

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

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.0671830780319958

Number of Support Vectors : 5290

Objective Function Value : -9833.257 -5196.674 -10164.46
Training error : 0.187891
Cross validation error : 0.399434
Probability model included.
> #performing prediction
> svmPred_1 <- predict(svmOutput_1, testData_1)
> # for accuracy
> confusionMatrix(svmPred_1,testData_1$pr_pa_dt)
Confusion Matrix and Statistics

          Reference
Prediction  Detractor Passive Promoter
  Detractor       764     229       67
  Passive         286     496      244
  Promoter        120     351      805

Overall Statistics

               Accuracy : 0.6142
```

We applied our model on entire dataset taking 2/3rd of data for training and the rest as Test data. Putting Training data in SVM model and test data to predict we get accuracy as 61%.

SVM only on Fly Fast Airways Inc.:

```
flyfast <- data2[str_trim(df$Partner.Name)=="FlyFast Airways Inc." & df$Flight.cancelled=='No',]
randIndex2 <- sample(1:dim(flyfast)[1])
cutPoint2_3_2 <- floor(2 * dim(flyfast)[1]/3)
trainData_2 <- flyfast[randIndex2[1:cutPoint2_3_2],]
testData_2 <- flyfast[randIndex2[(cutPoint2_3_2+1):dim(flyfast)[1]],]

# running support vector machine
svmOutput_2 <- ksvm(pr_pa_dt ~.,data=trainData_2,kernel="rbfdot",kpar="automatic",C = 5,cross = 3,prob.model = TRUE)
svmOutput_2

# performing prediction
svmPred_2 <- predict(svmOutput_2, testData_2)

# to get accuracy
confusionMatrix(svmPred_2,testData_2$pr_pa_dt)
```

Output:

```
> # running support vector machine
> svmOutput_2 <- ksvm(pr_pa_dt ~.,data=trainData_2,kernel="rbfdot",kpar="automatic",C = 5,cross = 3,prob.model = TRUE)
> svmOutput_2
Support Vector Machine object of class "ksvm"

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

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.0673798626831995

Number of Support Vectors : 605

Objective Function Value : -848.8485 -441.8481 -774.8932
Training error : 0.082645
Cross validation error : 0.396694
Probability model included.
>
> # performing prediction
> svmPred_2 <- predict(svmOutput_2, testData_2)
>
> # to get accuracy
> confusionMatrix(svmPred_2,testData_2$pr_pa_dt)
Confusion Matrix and Statistics

          Reference
Prediction  Detractor Passive Promoter
  Detractor       113      31        9
  Passive          43      56       32
  Promoter         11      22       47

Overall Statistics

               Accuracy : 0.5934
                 95% CI : (0.541, 0.6443)
```

We filtered Fly Fast Airways data from the entire dataset and then took 2/3rd of filtered data for training and the rest as Test data. Putting Training data in SVM model and test data to predict we get accuracy as 59%.

SVM only on Northwest Business Airlines Inc.:

```
northwest_bus <- data2[str_trim(df$Partner.Name)=="Northwest Business Airlines Inc."& df$Flight.cancelled=='No',]
randIndex3 <- sample(1:dim(northwest_bus)[1])
cutPoint2_3_3 <- floor(2 * dim(northwest_bus)[1]/3)
trainData_3 <- northwest_bus[randIndex3[1:cutPoint2_3_3],]
testData_3 <- northwest_bus[randIndex3[(cutPoint2_3_3+1):dim(northwest_bus)[1]],]

# running support vector machine
svmOutput_3 <- ksvm(pr_pa_dt ~.,data=trainData_3,kernel="rbfdot",kpar="automatic",C = 5,cross = 3,prob.model = TRUE)
svmOutput_3

svmPred_3 <- predict(svmOutput_3, testData_3)
svmPred_3

# to get accuracy
confusionMatrix(svmPred_3,testData_3$pr_pa_dt)
```

Output:

```
> svmOutput_3 <- ksvm(pr_pa_dt ~.,data=trainData_3,kernel="rbfdot",kpar="automatic",C = 5,cross = 3,prob.model = TRUE)
> svmOutput_3
Support Vector Machine object of class "ksvm"

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

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.0663557306787576

Number of Support Vectors : 536

Objective Function Value : -1297.108 -37.8713 -41.8586
Training error : 0.076074
Cross validation error : 0.280909
Probability model included.
>
> svmPred_3 <- predict(svmOutput_3, testData_3)
>
> # to get accuracy
> confusionMatrix(svmPred_3,testData_3$pr_pa_dt)
Confusion Matrix and Statistics

          Reference
Prediction  Detractor Passive Promoter
  Detractor       192      42        0
  Passive          39     132        3
  Promoter          0       0        0

Overall Statistics

               Accuracy : 0.7941
                 95% CI : (0.7516, 0.8323)
```

We filtered Northwest Business Airlines data from the entire dataset and then took 2/3rd of filtered data for training and the rest as Test data. Putting Training data in SVM model and test data to predict we get accuracy as 79%.

RANDOM FOREST

Random forest, as its name implies, consists of many individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.

In Random Forest, we have considered variables, which we got from the output of Correlation Matrix and Linear Modelling to cross, validate the significance of the variables.

The output will give accuracy rate, which will signify how good the selected variables for forecasting are.

Random Forest Model on entire dataset:

```
#---------------------Randomforest for entire-----------------
# using randomforest and cat tools package to apply random forest model.
rf1<-randomForest(pr_pa_dt ~.,data=trainData_1)
rf1
predrf1 <- predict(rf1, testData_1)
predrf1
cm1 <- table(testData_1$pr_pa_dt,predrf1)
cm1
confusionMatrix(predrf1,testData_1$pr_pa_dt)

#                        randomforest for flyfast
```

Output:

```
> #---------------------Randomforest for entire-----------------
> # using randomforest and cat tools package to apply random forest model.
> rf1<-randomForest(pr_pa_dt ~.,data=trainData_1)
> rf1

Call:
 randomForest(formula = pr_pa_dt ~ ., data = trainData_1)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 37.43%
Confusion matrix:
          Detractor Passive Promoter class.error
Detractor      1582     486      247   0.3166307
Passive         506     970      728   0.5598911
Promoter        104     445     1654   0.2492056
> predrf1 <- predict(rf1, testData_1)
>
> # for determinig accuracy
> confusionMatrix(predrf1,testData_1$pr_pa_dt)
Confusion Matrix and Statistics

           Reference
Prediction  Detractor Passive Promoter
  Detractor       794     234       48
  Passive         248     464      211
  Promoter        128     378      857

Overall Statistics

               Accuracy : 0.6291
                 95% CI : (0.6125, 0.6454)
```

We applied our model on the same training & test dataset used in SVM. Putting Training data in Random forest model and test data to predict we get accuracy as 63%, which is approximately same that we got in SVM.

Random Forest model only on Fly Fast Airways Inc.:

```
#---------------------randomforest for flyfast-----------------
rf2<-randomForest(pr_pa_dt ~.,data=trainData_2)
rf2
predrf2 <- predict(rf2, testData_2)

# for determinig accuracy
confusionMatrix(predrf2,testData_2$pr_pa_dt)
```

Output:

```
> #---------------------randomforest for flyfast-----------------
> rf2<-randomForest(pr_pa_dt ~.,data=trainData_2)
> rf2

Call:
 randomForest(formula = pr_pa_dt ~ ., data = trainData_2)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 38.84%
Confusion matrix:
          Detractor Passive Promoter class.error
Detractor       233      57       12   0.2284768
Passive          80     109       52   0.5477178
Promoter         16      65      102   0.4426230
> predrf2 <- predict(rf2, testData_2)
>
> # for determinig accuracy
> confusionMatrix(predrf2,testData_2$pr_pa_dt)
Confusion Matrix and Statistics

           Reference
Prediction  Detractor Passive Promoter
  Detractor       123      35        9
  Passive          38      54       32
  Promoter          6      20       47

Overall Statistics

               Accuracy : 0.6154
                 95% CI : (0.5633, 0.6656)
```

We filtered Fly Fast Airways data from the entire dataset and then took the same training and test dataset that we took for Fly Fast Airways in SVM. Putting Training data in Random Forest model and test data to predict we get accuracy as 62%, which is approximately same that we got in SVM.

Random Forest only on Northwest Business Airlines Inc.:

```
#---------------------randomforest for northwest-----------------
rf3<-randomForest(pr_pa_dt ~.,data=trainData_3)
rf3
predrf3 <- predict(rf1, testData_3)

# for determinig accuracy
confusionMatrix(predrf3,testData_3$pr_pa_dt)
```

Output:

```
> #---------------------randomforest for northwest-----------------
> rf3<-randomForest(pr_pa_dt ~.,data=trainData_3)
> rf3

Call:
 randomForest(formula = pr_pa_dt ~ ., data = trainData_3)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 25.28%
Confusion matrix:
          Detractor Passive Promoter class.error
Detractor       388      98        0   0.2016461
Passive         102     221        0   0.3157895
Promoter          3       3        0   1.0000000
> predrf3 <- predict(rf1, testData_3)
>
> # for determinig accuracy
> confusionMatrix(predrf3,testData_3$pr_pa_dt)
Confusion Matrix and Statistics

           Reference
Prediction  Detractor Passive Promoter
  Detractor       194       5        0
  Passive          25     136        1
  Promoter         12      33        2

Overall Statistics

               Accuracy : 0.8137
                 95% CI : (0.7725, 0.8503)
```

We filtered Northwest Business Airlines data from the entire dataset and then took the same training and test dataset that we took for Northwest Business Airlines in SVM. Putting Training data in Random Forest model and test data to predict we get accuracy as 81%, which is approximately same that we got in SVM.

INSIGHTS

- The SouthEast Airlines and their partners should focus on needs of the Female customers, by providing them with better services for example mothers travelling with their children should be provided special attention and their needs should be catered to, better seats can be provided (with better leg room) to them, their food preferences can be taken into consideration and more importantly start taking customer feedback etc..

- Airlines belonging to the Silver, Gold and Platinum status achieve relatively higher positive ratings from the customers when compared with airlines in 'Blue' status.
  So, the airlines in "Blue" status should attract the customers with some discount coupons and improving their services such as in-flight entertainment, better food options, so that the passengers feel that they are getting value for their money, as they tend to be price sensitive.

- Airlines should focus on giving better amenities such as wheelchair services, better seats, assistance at the time of check-in and express boarding to customers who are above 54 years as they are more likely to be detractors.

- Airlines who have passengers travelling for business purposes should focus that the flight is running on time as if the flight is delayed there is a high probability that they might give a negative rating so passengers should be provided with lounge access and if the passenger misses his connecting flight provide them with the next immediate flight.

- Airlines should focus on passengers who are frequent flyers and price sensitive by enrolling them in a higher tier of their frequent flyer account so that they can accrue more mileage points and get free tickets, so in this way they can attract customers who are frequent flyers.

Southeast airlines should particularly focus on its regional partners Northwest Business Airlines Inc. and Flyfast Airways as they have the highest relative percentage of its passengers as detractors. We suggest that they should be put on a probationary tenure and be told to focus on the parameters suggested above to improve their customer ratings. If they fail then their contracts shouldn't be renewed.