THE iSCHOOL
Syracuse University

IST 687 INTRODUCTION TO DATA SCIENCE

Lab Section M007 | Group 3

DATA ANALYSIS FOR SOUTHEAST AIRLINES

SOUTHEAST AIRLINES
The Spirit of Liberty™

SUBMITTED BY:

HARSH DARJI | HARSH MEHTA | RITIKA SHETTY | URVI MISTRY

Table of contents

1. Description

The project revolves around analyzing the survey data collected from a huge number of customers traveling within the United States and use those data to provide recommendations to Southeast Airlines Co. The recommendations simply answer business questions to increase revenue and provide suggestions for improvement to the airlines for enhanced customer satisfaction.

2. Project Scope and Objective

The scope of this project is to analyse and draw insights from the dataset provided to us which contains data regarding customer satisfaction based on various factors. The data contains survey of people taking different airlines, a total of 14 different airlines were found to be present in the dataset.

We will be concentrating on finding the factors that cause a customer to be unsatisfied and get actionable insights by applying statistical techniques.

The objective of this project is to suggest our client, Southeast Airlines Co. , the areas where they can improve to increase their customer's satisfaction. Also, provide them with insights on what is leading to less satisfaction among their customers compared to other airlines.

3. Project Deliverables

- Perform data cleaning to prepare the data for further analysis so that there are no missing or invalid fields in our dataset.
- Identify the attributes that most affects the satisfaction of the customers by applying linear regression and perform further analysis on those attributes.
- Identification of association rules using Apriori algorithm for suggesting rules that can be used to improve their overall customer satisfaction.

- Predicting the customer satisfaction by applying support vector machine and formulate actionable insights.
- Finally, provide suggestions to the client based on data analysis and interpretation to enhance and improve their customer's satisfaction. Especially targeting then less satisfied group of customers.

4. Data Acquisition

The data set was made available to us by the course instructors. Before any data munging, this data set consisted of approximately 130,000 survey responses of the customers travelling with varied airlines and approximately 25 fields such as Age, Gender, Flight time, Flight distance, etc.

This data was extensively studied to determine the usable variables. After this initial analysis, the data set was forwarded to the preprocessing phase where all the errors in the data were removed in order to make it usable for further analysis.

5. Data preprocessing

Before preprocessing, the data set consisted of 129,889 rows and 28 column variables. All the data in the data set was the survey taken of customers travelling by various airlines.

Firstly, the column names in the data set were not consistent and thus were changed into the same format.

Secondly, all the columns of the data set were summarized to determine errored values and it was found that only Satisfaction column had incorrect values. There were 3 values which were mis typed and thus were replaced with value 4.00.

Thirdly, the NA values in the Arrival delay greater than 5 minutes were replaced with 0. When we analyzed the data with NA values, we found that all those rows of data had value 'no' in the arrival delay greater than 5 minutes column. Thus, replacing these NA values with 0 would not be a feasible option as the mean of those data was greater than 5 minutes. Similarly, the NA values in the Flight time column were replaced with mean value as the time of the flight cannot be 0.

Lastly, the data set consisted of columns which represented data of cancelled flights. These flights had NA values in the various columns such as arrival delay, departure delay, etc. These NA values couldn’t be replaced with 0 or mean values as they would change the entire meaning of cancelled flights. Thus, we subsetted those rows into a different data set.

After data munging the number of rows in the data set were 127488 and had 28 columns.

Code:

```
rawData <- df
colnames(rawData)<-c("Satisfaction", "Airline_status", "Age", "Gender",
"Price_sensitivity", "Year_of_flights", "No_of_flights_pa",
"Percent_of_flights_with_other_airlines", "Type_of_travel", "No_of_other_loyalty_cards",
"Shopping_amount_at_airport", "Eating_and_drinking_at_airport", "Class",
"Day_of_month", "Flight_date", "Airline_code", "Airline_name", "Origin_city",
"Origin_state", "Destination_city", "Destination_state", "Scheduled_departure_hour",
"Departure_delay_in_minutes", "Arrival_delay_in_minutes", "Flight_cancelled",
"Flight_time_in_minutes", "Flight_distance", "Arrival_delay_greater_than_5minutes")
rawData

#Cleaning Satisfaction Column
change <-
c(which(grepl("4.00.2.00",rawData$Satisfaction)),which(grepl("4.00.5",rawData$Satisfaction)))
change
rawData$Satisfaction <- replace(rawData$Satisfaction,list = change,4)
table(rawData$Satisfaction)

#Replacing NA's in Arrival_delay_in_minutes by 0's
index <- which(is.na(cleanData$Arrival_delay_in_minutes))
cleanData$Arrival_delay_in_minutes <- replace(cleanData$Arrival_delay_in_minutes,
is.na(cleanData$Arrival_delay_in_minutes),0)

#Replacing NA's in Flight_time_in_minutes by mean flight time
mean_time <- mean(na.omit(cleanData$Flight_time_in_minutes))
cleanData$Flight_time_in_minutes <- replace(cleanData$Flight_time_in_minutes,
is.na(cleanData$Flight_time_in_minutes),mean_time)
```

```
#Subsetting cancelled flights
wasteData <- rawData[which(rawData$Flight_cancelled=='Yes'),]
wasteData

cleanData <- rawData[which(rawData$Flight_cancelled=='No'),]
cleanData
```

6. Modelling techniques

Various models have been used for accurate results of the information obtained from the data set. These models give an understandable representation of real-life information from the data sets. The following models have been implemented.

6.1. Linear Regression

To begin with, we applied linear modelling to our dataset. By applying Simple linear regression we could summarize and study relationships between two continuous variables in our dataset. The core idea was to obtain a line that best fits the data. The best fit line is the one for which total prediction error are as small as possible.

We found out several significant variables out of which we decided the top three significant variables based on further analysis of the model. This analysis included the following, checking for a high r squared value that is coefficient of determination, a small p value and by analysing the residual plots.

As per our analysis, the following variables had a statistically significant relationship with the dependent variable which for this project was the “ Customer Satisfaction” :

1. Airline status
2. Type of travel
3. Arrival delay greater than 5 minutes
4. Age
5. Number of flights per annum

Out of the above mentioned attributes our analysis suggested that the Age and Number of flights per annum attributes had some hidden pattern which would skew our analyses.

Further analysis of our significant variables:
The following visualization help us to understand our data better with respect to the top three factors that affect our customer satisfaction.

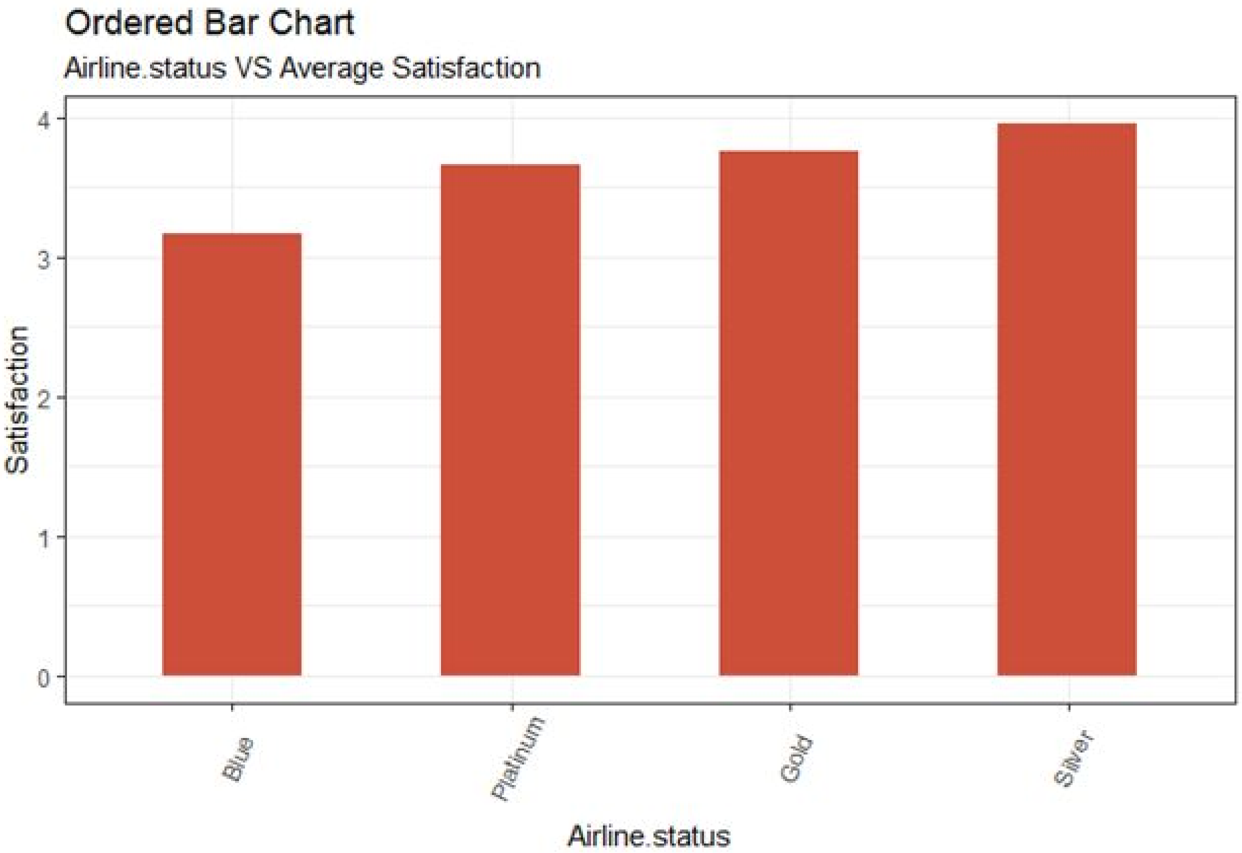


Fig : Bar Chart showing Airline status VS average satisfaction

Insight from graph : It is clearly evident from the graph that Airline status plays an important role in affecting the customer satisfaction. As we can notice the flyers with Airline status : Blue have the lowest mean satisfaction and thus this is one important area to concentrate for improvement.

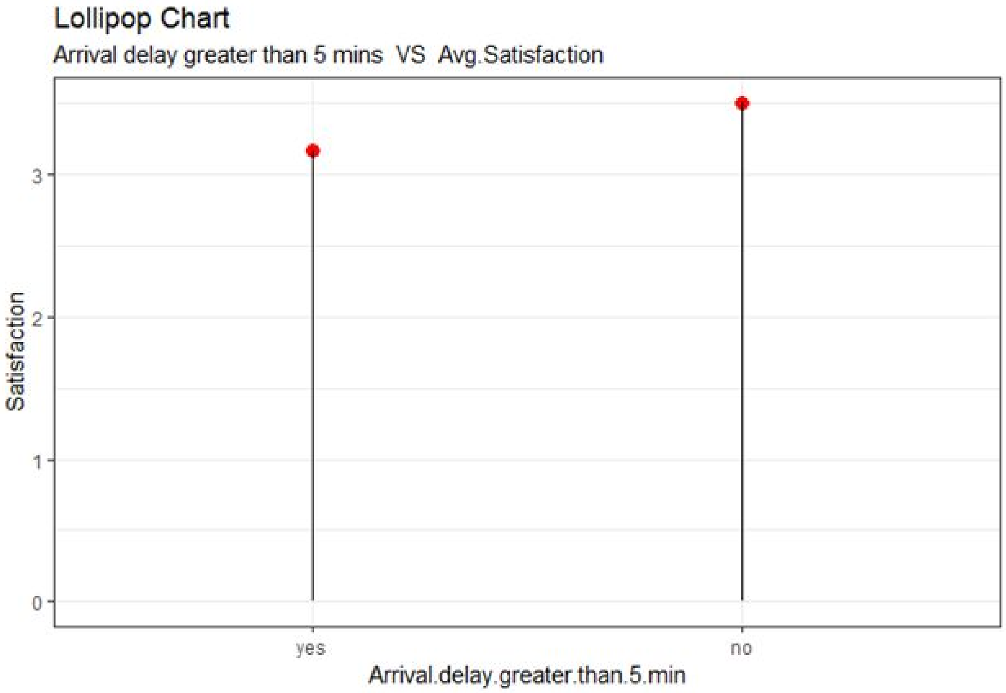


Fig : Chart to plot Arrival delay greater than 5 mins Vs Average Satisfaction

Insights from graph : The above chart depicts how the customers tend to be less satisfied when the arrival delay is greater than 5 minutes. This is another important factor that can be worked upon for better customer experience.

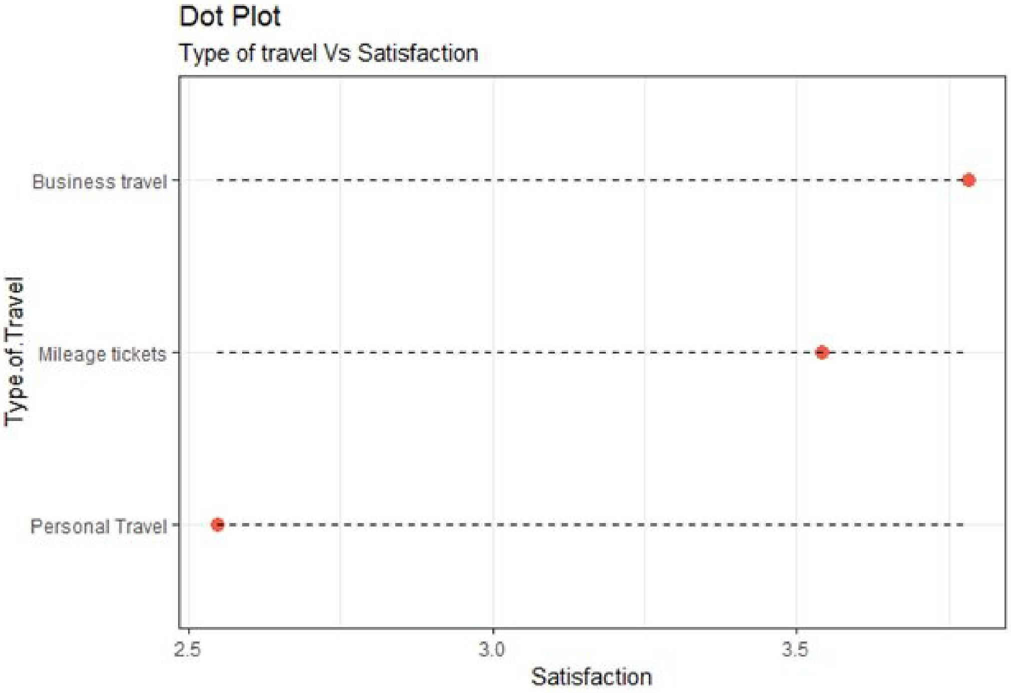


Fig : Plot of Type of travel Vs satisfaction

Insights from graph: The above graph conveys the information that the “personal Travel“ customers are particularly unsatisfied and the airline should be focusing on providing better service to this group of customers.

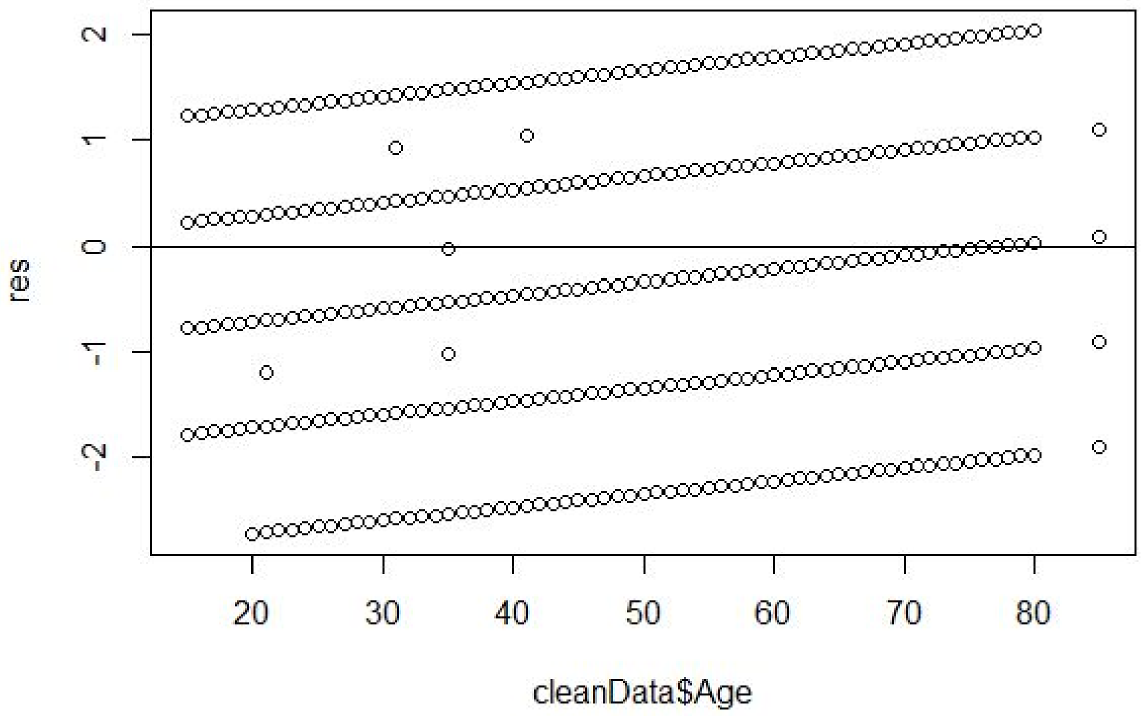


Fig: Residual plot for Age

Insight from graph : We did not consider “Age” as one of our significant variables even when it had a high r squared value and a low p value because the residual plot was not randomly scattered but showed a pattern.

Further analysis of age showed that there were more number of people between age 25 to 60 that travel than the ones above and below this age group. Also, it suggested that we did not have enough data to analyze what caused dissatisfaction among the customers of ages below 25 and above 60 years old. The visualization below explains this clearly.

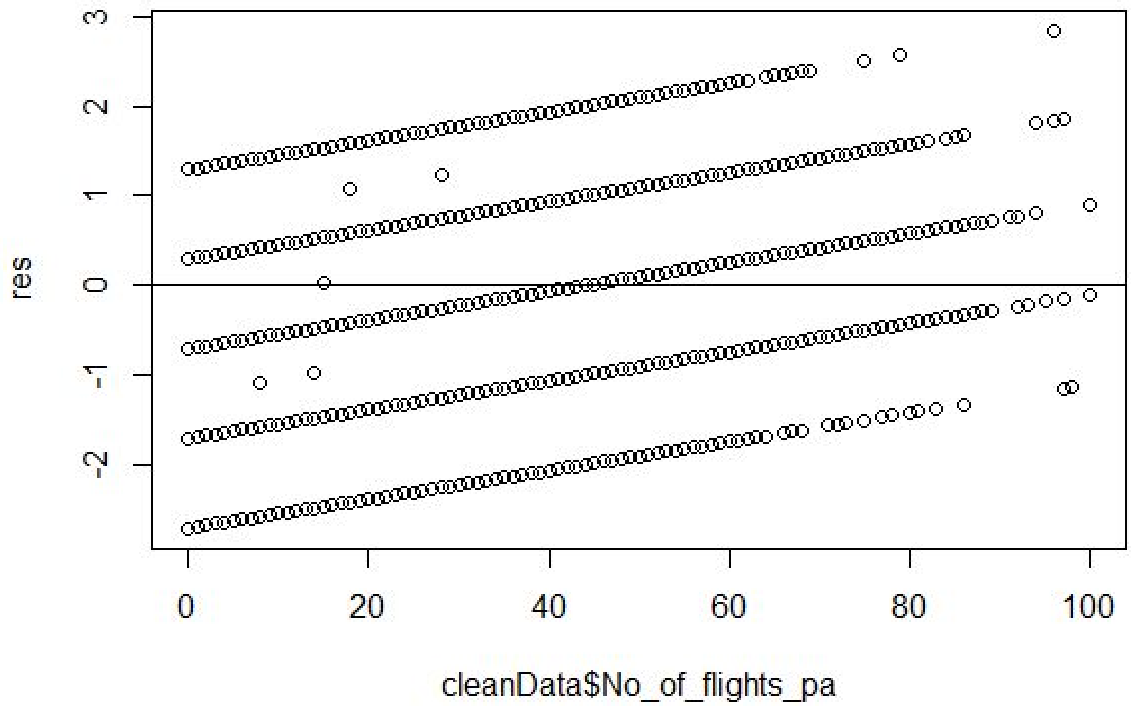


Fig : Residual plot for Number of flights per annum

Insight From Graph : The above graph shows a pattern suggesting hidden pattern in the attribute which might skew our analysis thus the following attribute was not selected as a significant one.

6.2. Association Rules

From the linear modelling, we found 5 major attributes affecting satisfaction of the customers. To be sure about it, we performed Apriori Algorithm on the data set.

The Apriori Algorithm is an influential algorithm for mining frequent item sets for Boolean association rules. Apriori is designed to operate on data sets containing transactions.

We first made plots and tables of all the variables of the data set to categorize them in such a way that there are approximately equal number of values in all the categories. Then, we performed the Apriori Algorithm twice to understand and compare the associations of all the variables of the data set as well as of the 5 major variables derived from the linear regression model whose satisfaction rating was 3.5 and lower.

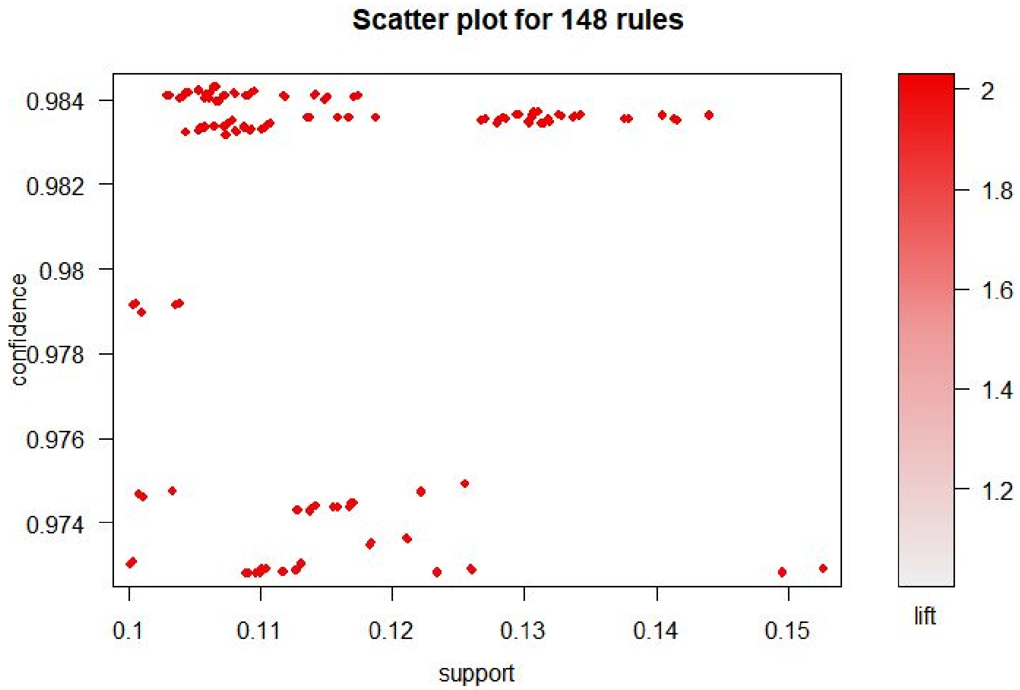


Fig: Scatter plot for 148 rules of all the variables of the data set.

Insight from the graph: We found that there were 148 rules which had lift values greater than 2.0. The above graph shows the resulting confidence and support of all the rules.

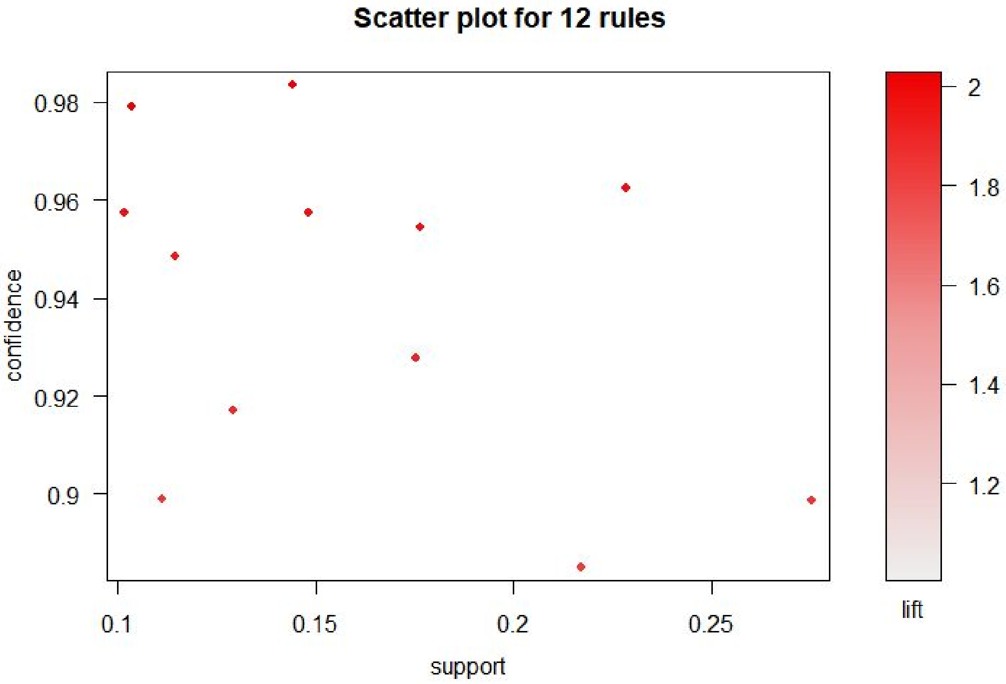


Fig: Scatter plot for the variables computed from the linear regression model

Insight from the graph: The rules i.e. the points which have lift values greater than 1.8. Also, the rules that have higher confidence and support are the best rules.

Rules:

```
    lhs                                            rhs            support confidence     lift count
[1]   {cleanData.Type_of_travel=Personal  Travel}              => {satisfied=no} 0.2753592
0.8987685 1.848398 35105
[2]  {cleanData.Type_of_travel=Personal Travel,
      cleanData.Arrival_delay_greater_than_5minutes=yes} => {satisfied=no} 0.1017586
0.9574876 1.969159 12973
[3]  {cleanData.Type_of_travel=Personal Travel,
      Age=High}                                    => {satisfied=no} 0.1754440  0.9278218
1.908149 22367
[4]  {cleanData.Airline_status=Blue,
        cleanData.Type_of_travel=Personal  Travel}             => {satisfied=no} 0.2282960
0.9624351 1.979334 29105
[5]  {cleanData.Type_of_travel=Personal Travel,
      Flightspa=low}                               => {satisfied=no} 0.2169459  0.8848861
1.819848 27658
[6]  {cleanData.Type_of_travel=Personal Travel,
     Age=High,
      cleanData.Arrival_delay_greater_than_5minutes=no}  => {satisfied=no} 0.1110928
0.8989527 1.848777 14163
[7]  {cleanData.Airline_status=Blue,
     cleanData.Type_of_travel=Personal Travel,
      Age=High}                                    => {satisfied=no} 0.1441234  0.9836188
2.022900 18374
[8]  {cleanData.Type_of_travel=Personal Travel,
     Flightspa=low,
      Age=High}                                    => {satisfied=no} 0.1291886  0.9170379
1.885971 16470
[9]  {cleanData.Airline_status=Blue,
     cleanData.Type_of_travel=Personal Travel,
      cleanData.Arrival_delay_greater_than_5minutes=no}  => {satisfied=no} 0.1483826
0.9575319 1.969250 18917
[10] {cleanData.Airline_status=Blue,
     cleanData.Type_of_travel=Personal Travel,
      Flightspa=low}                              => {satisfied=no} 0.1763146  0.9545609
1.963140 22478
[11] {cleanData.Airline_status=Blue,
     cleanData.Type_of_travel=Personal Travel,
     Flightspa=low,
      Age=High}                                   => {satisfied=no} 0.1036725  0.9791821
2.013776 13217
[12] {cleanData.Airline_status=Blue,
     cleanData.Type_of_travel=Personal Travel,
     Flightspa=low,
      cleanData.Arrival_delay_greater_than_5minutes=no}  => {satisfied=no} 0.1143637
0.9485395 1.950757 14580
```

All these above rules had one thing in common i.e.

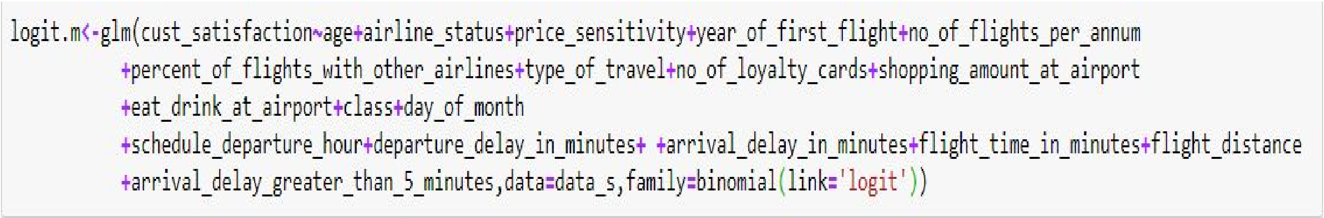
- Type of travel as Personal travel
- Airline Status as Blue.

Also, they had other variables such as price sensitivity, number of flights per annum, arrival delay greater than 5 minutes, etc. as low which means that these were not the factors affecting satisfaction but just the two of them i.e customer going for personal travel with blue status.

6.3. Logistic Regression

We use one more modelling technique to validate our data for predicting the customer satisfaction.We run a logistic regression on the full dataset of Southeast Airlines.And then we take the significant variables and find out the predicting power and the accuracy.

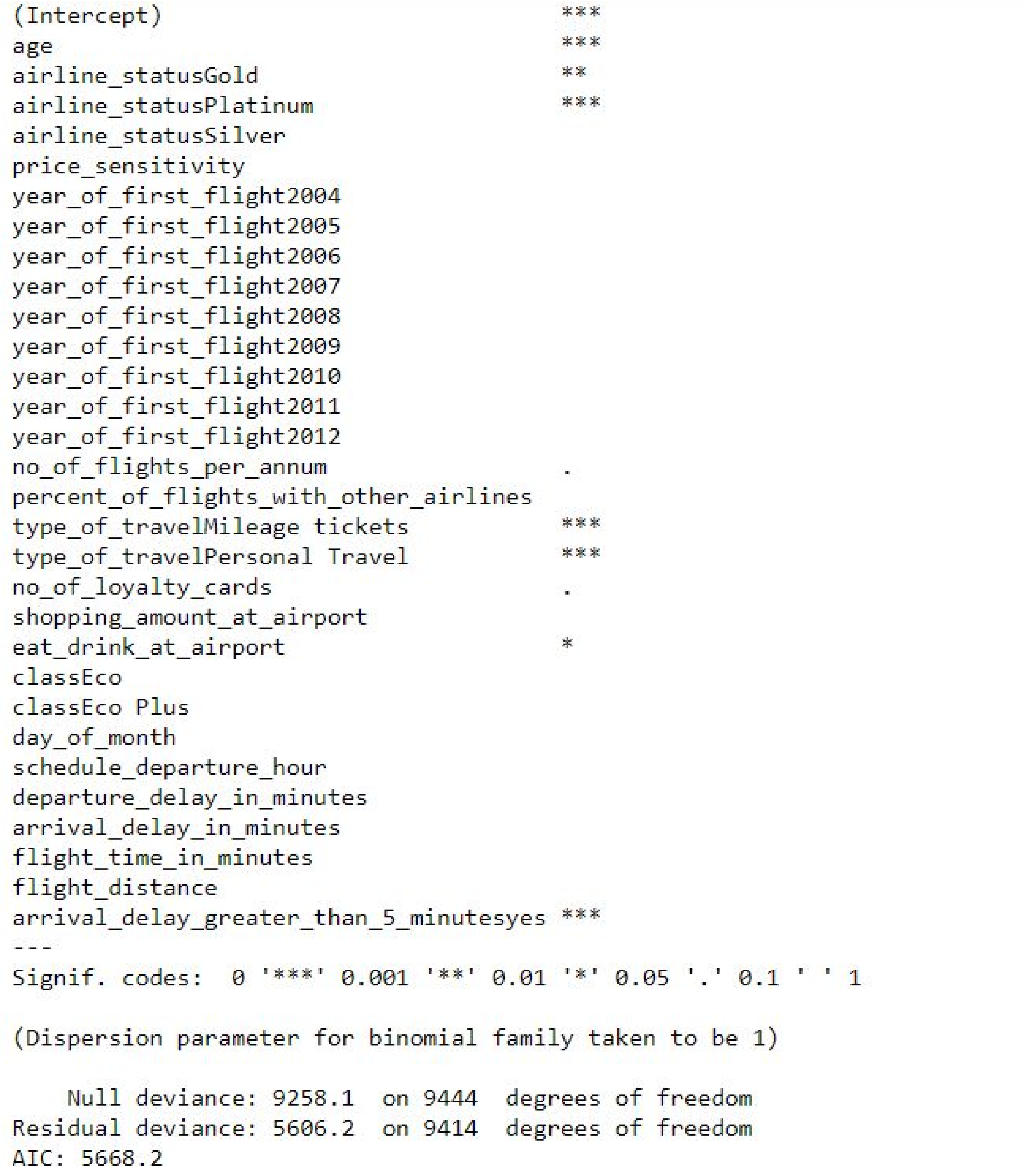
Code:



```
logit.m<-glm(cust_satisfaction~age+airline_status+price_sensitivity+year_of_first_flight+no_of_flights_per_annum
          +percent_of_flights_with_other_airlines+type_of_travel+no_of_loyalty_cards+shopping_amount_at_airport
          +eat_drink_at_airport+class+day_of_month
          +schedule_departure_hour+departure_delay_in_minutes+ +arrival_delay_in_minutes+flight_time_in_minutes+flight_distance
          +arrival_delay_greater_than_5_minutes,data=data_s,family=binomial(link='logit'))
```

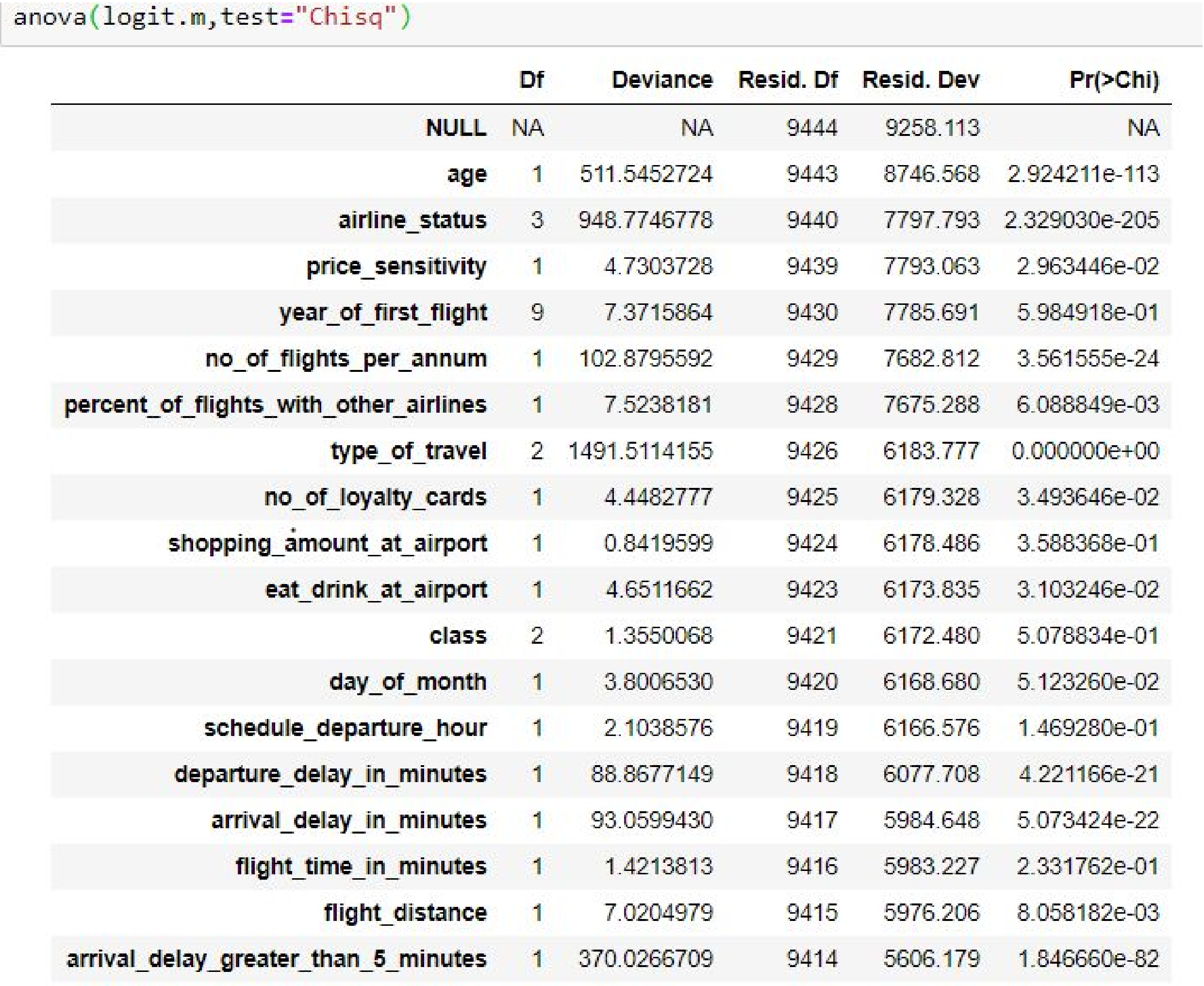
Output:

summary(logit.m)



```
(Intercept)                             ***
age                                     ***
airline_statusGold                      **
airline_statusPlatinum                  ***
airline_statusSilver
price_sensitivity
year_of_first_flight2004
year_of_first_flight2005
year_of_first_flight2006
year_of_first_flight2007
year_of_first_flight2008
year_of_first_flight2009
year_of_first_flight2010
year_of_first_flight2011
year_of_first_flight2012
no_of_flights_per_annum                 .
percent_of_flights_with_other_airlines
type_of_travelMileage tickets           ***
type_of_travelPersonal Travel           ***
no_of_loyalty_cards                     .
shopping_amount_at_airport
eat_drink_at_airport                    *
classEco
classEco Plus
day_of_month
schedule_departure_hour
departure_delay_in_minutes
arrival_delay_in_minutes
flight_time_in_minutes
flight_distance
arrival_delay_greater_than_5_minutesyes ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9258.1  on 9444  degrees of freedom
Residual deviance: 5606.2  on 9414  degrees of freedom
AIC: 5668.2
```

Now, we compare with the null model:



```
anova(logit.m,test="Chisq")
```

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	9444	9258.113	NA
age	1	511.5452724	9443	8746.568	2.924211e-113
airline_status	3	948.7746778	9440	7797.793	2.329030e-205
price_sensitivity	1	4.7303728	9439	7793.063	2.963446e-02
year_of_first_flight	9	7.3715864	9430	7785.691	5.984918e-01
no_of_flights_per_annum	1	102.8795592	9429	7682.812	3.561555e-24
percent_of_flights_with_other_airlines	1	7.5238181	9428	7675.288	6.088849e-03
type_of_travel	2	1491.5114155	9426	6183.777	0.000000e+00
no_of_loyalty_cards	1	4.4482777	9425	6179.328	3.493646e-02
shopping_amount_at_airport	1	0.8419599	9424	6178.486	3.588368e-01
eat_drink_at_airport	1	4.6511662	9423	6173.835	3.103246e-02
class	2	1.3550068	9421	6172.480	5.078834e-01
day_of_month	1	3.8006530	9420	6168.680	5.123260e-02
schedule_departure_hour	1	2.1038576	9419	6166.576	1.469280e-01
departure_delay_in_minutes	1	88.8677149	9418	6077.708	4.221166e-21
arrival_delay_in_minutes	1	93.0599430	9417	5984.648	5.073424e-22
flight_time_in_minutes	1	1.4213813	9416	5983.227	2.331762e-01
flight_distance	1	7.0204979	9415	5976.206	8.058182e-03
arrival_delay_greater_than_5_minutes	1	370.0266709	9414	5606.179	1.846660e-82

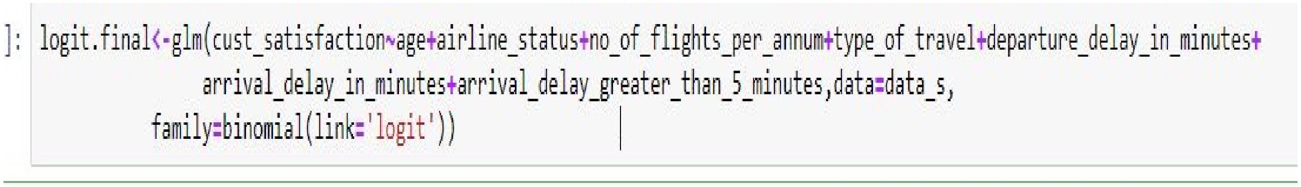
The difference between the null deviance and the residual deviance shows how our model is doing against the null model (a model with only the intercept). The wider this gap, the better.

Adding Age, airline_status and type_of_travel significantly reduces the residual deviance. The other variables seem to improve the model less even though class and arrival_delay_greater_than_5_minutes has a low p-value.

A large p-value here indicates that the model without the variable explains more or less the same amount of variation.

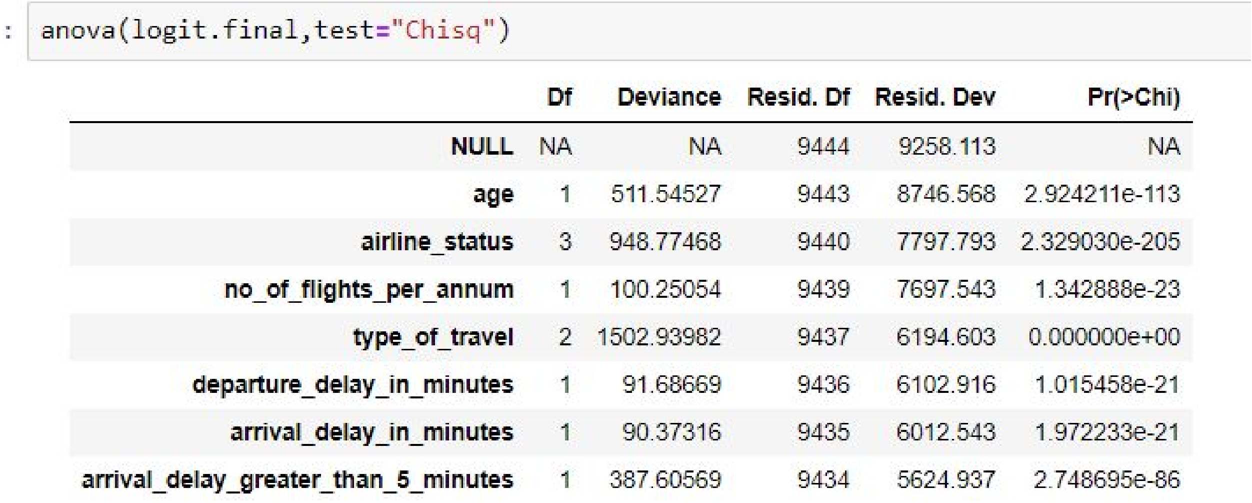
Ultimately what you would like to see is a significant drop in deviance and the AIC.
Here(logistic regression) doesn't have equivalent R^2 value like in Linear regression, but the McFadden R2 index can be used to assess the model fiy.

Now, we select only the significant variables in our model:



```
]: logit.final<-glm(cust_satisfaction~age+airline_status+no_of_flights_per_annum+type_of_travel+departure_delay_in_minutes+
                arrival_delay_in_minutes+arrival_delay_greater_than_5_minutes,data=data_s,
        family=binomial(link='logit'))
```

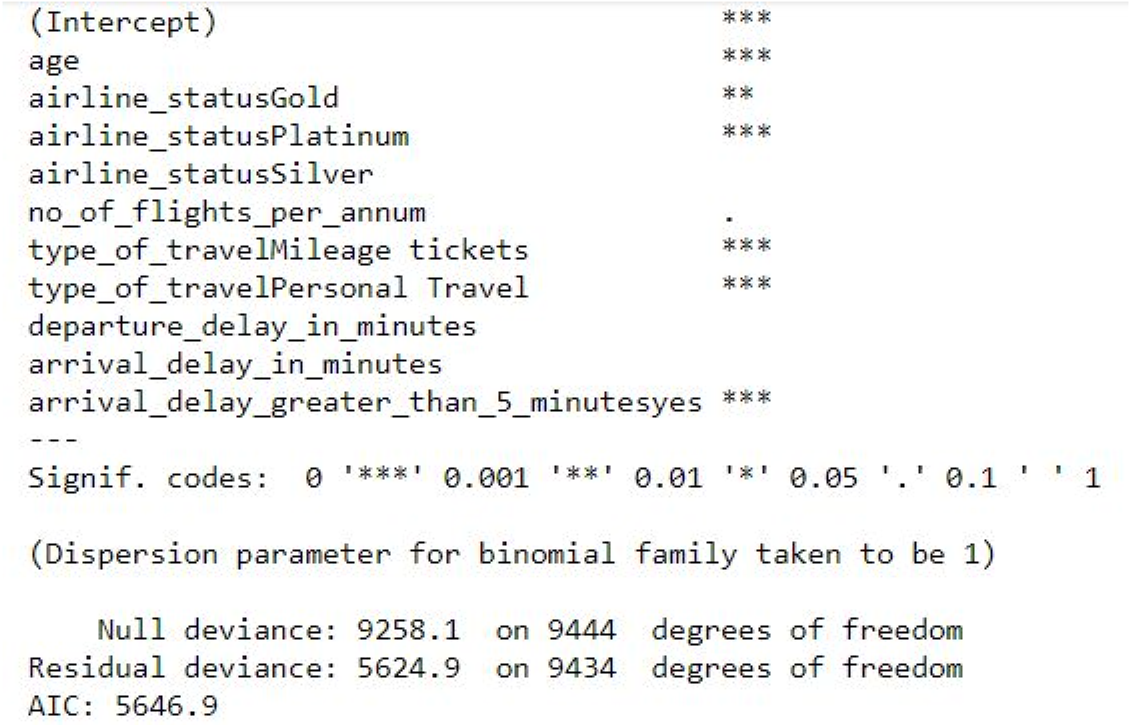
Comparing with null model:



```
: anova(logit.final,test="Chisq")
```

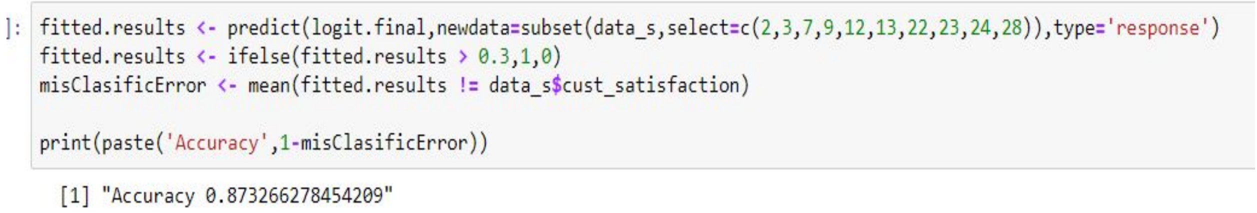
	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	9444	9258.113	NA
age	1	511.54527	9443	8746.568	2.924211e-113
airline_status	3	948.77468	9440	7797.793	2.329030e-205
no_of_flights_per_annum	1	100.25054	9439	7697.543	1.342888e-23
type_of_travel	2	1502.93982	9437	6194.603	0.000000e+00
departure_delay_in_minutes	1	91.68669	9436	6102.916	1.015458e-21
arrival_delay_in_minutes	1	90.37316	9435	6012.543	1.972233e-21
arrival_delay_greater_than_5_minutes	1	387.60569	9434	5624.937	2.748695e-86

Output:



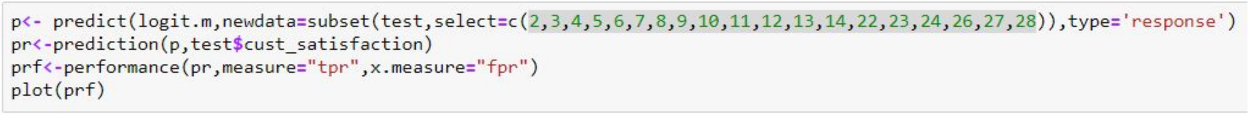
```
(Intercept)                           ***
age                                   ***
airline_statusGold                    **
airline_statusPlatinum                ***
airline_statusSilver
no_of_flights_per_annum               .
type_of_travelMileage tickets         ***
type_of_travelPersonal Travel         ***
departure_delay_in_minutes
arrival_delay_in_minutes
arrival_delay_greater_than_5_minutesyes ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9258.1  on 9444  degrees of freedom
Residual deviance: 5624.9  on 9434  degrees of freedom
AIC: 5646.9
```

Next, step is to find out accuracy:



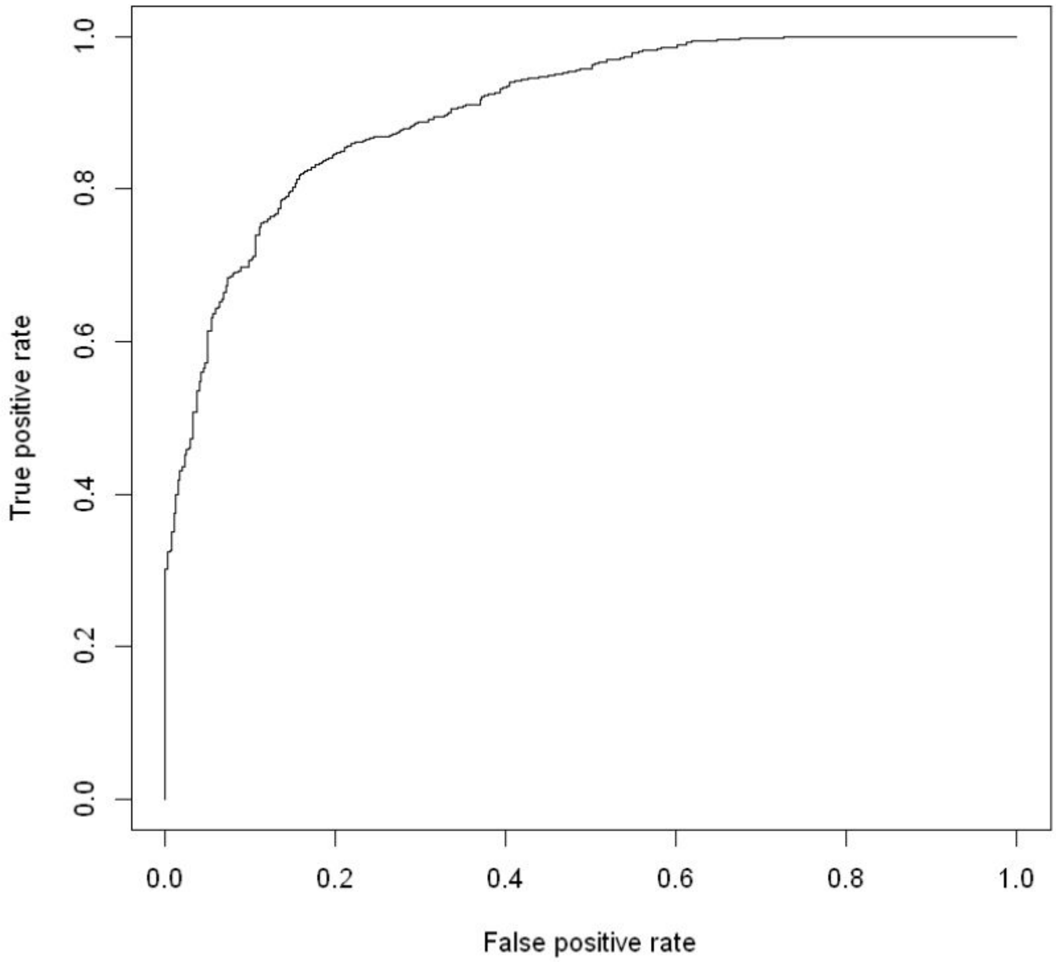
```
]: fitted.results <- predict(logit.final,newdata=subset(data_s,select=c(2,3,7,9,12,13,22,23,24,28)),type='response')
   fitted.results <- ifelse(fitted.results > 0.3,1,0)
   misClasificError <- mean(fitted.results != data_s$cust_satisfaction)

   print(paste('Accuracy',1-misClasificError))

     [1] "Accuracy 0.873266278454209"
```

Area under the curve:
library(ROCR)
Code:

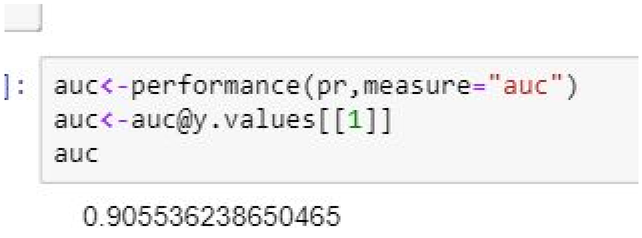


```
p<- predict(logit.m,newdata=subset(test,select=c(2,3,4,5,6,7,8,9,10,11,12,13,14,22,23,24,26,27,28)),type='response')
pr<-prediction(p,test$cust_satisfaction)
prf<-performance(pr,measure="tpr",x.measure="fpr")
plot(prf)
```

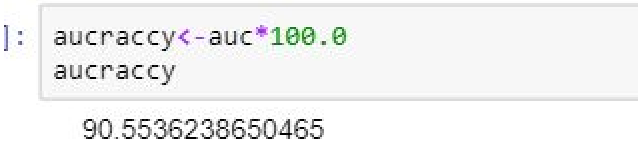
Output:



Accuracy:



```
]: auc<-performance(pr,measure="auc")
   auc<-auc@y.values[[1]]
   auc

     0.905536238650465
```

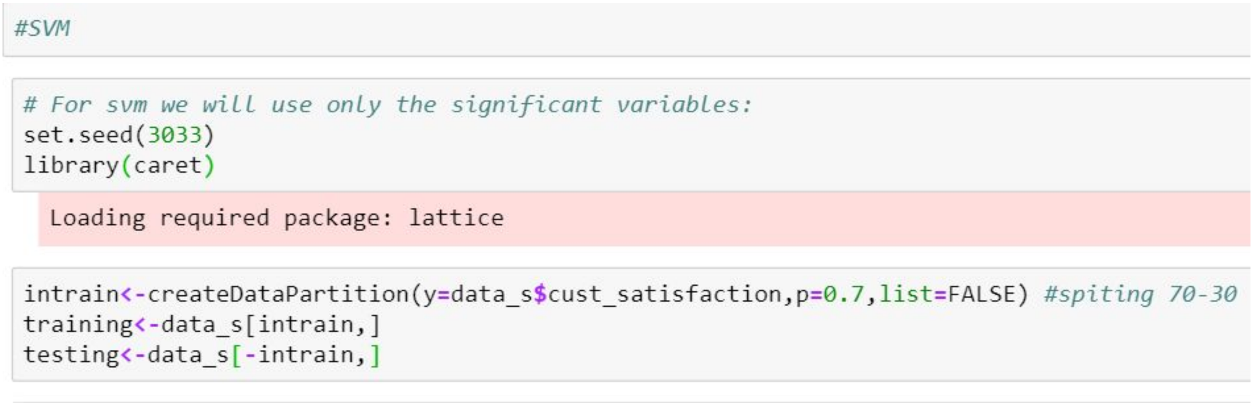


```
]: aucraccy<-auc*100.0
   aucraccy

     90.5536238650465
```

6.4. Support Vector Machine

We use svm modeling techniques to predict the customer satisfaction by using using various significant variables from our model.

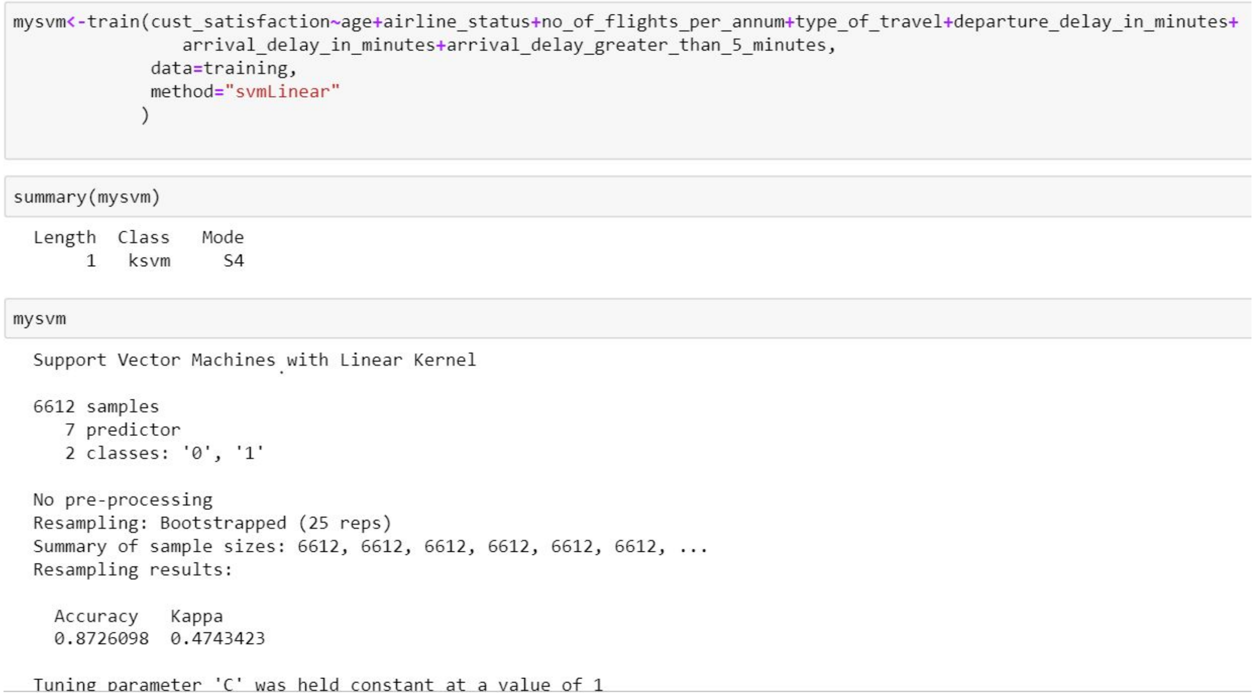
We divide the dataset into training and testing so that we can check and validate our results.



```
#SVM

# For svm we will use only the significant variables:
set.seed(3033)
library(caret)

  Loading required package: lattice

intrain<-createDataPartition(y=data_s$cust_satisfaction,p=0.7,list=FALSE) #spiting 70-30
training<-data_s[intrain,]
testing<-data_s[-intrain,]
```

Set.seed will help to get same results when you run the svm.

Output:



```
mysvm<-train(cust_satisfaction~age+airline_status+no_of_flights_per_annum+type_of_travel+departure_delay_in_minutes+
                arrival_delay_in_minutes+arrival_delay_greater_than_5_minutes,
             data=training,
             method="svmLinear"
            )
```

```
summary(mysvm)
```

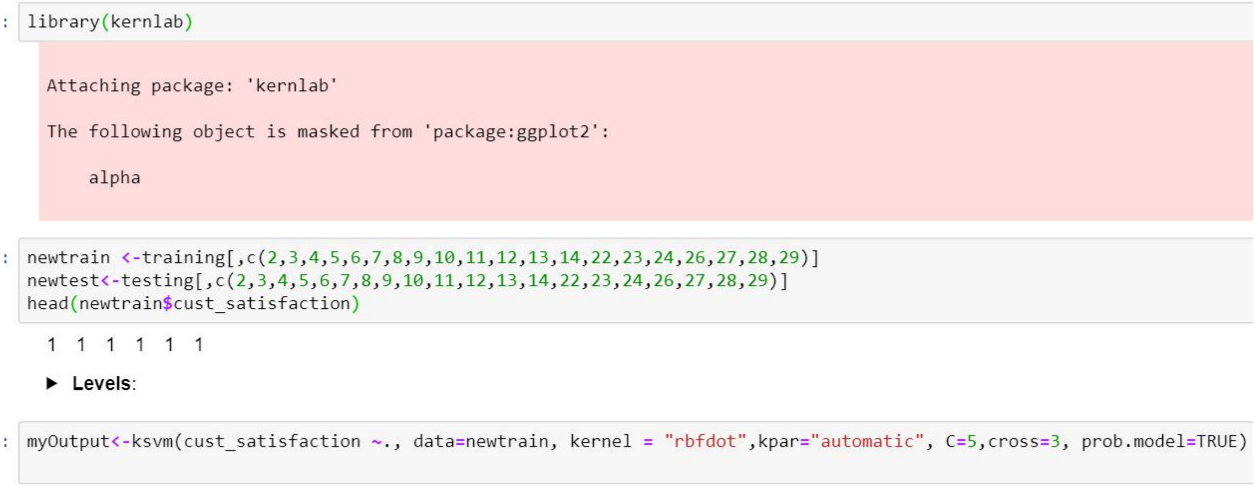
```
  Length  Class   Mode
       1   ksvm     S4
```

```
mysvm
```

```
  Support Vector Machines with Linear Kernel

  6612 samples
     7 predictor
     2 classes: '0', '1'

  No pre-processing
  Resampling: Bootstrapped (25 reps)
  Summary of sample sizes: 6612, 6612, 6612, 6612, 6612, 6612, ...
  Resampling results:

    Accuracy   Kappa
    0.8726098  0.4743423

  Tuning parameter 'C' was held constant at a value of 1
```

Using Kernlab library for SVM:

Code:



```
: library(kernlab)
```

```
  Attaching package: 'kernlab'

  The following object is masked from 'package:ggplot2':

      alpha
```

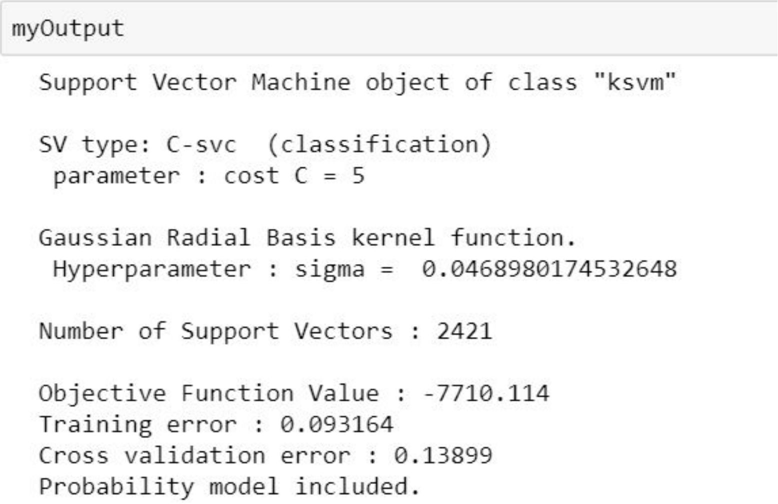
```
: newtrain <-training[,c(2,3,4,5,6,7,8,9,10,11,12,13,14,22,23,24,26,27,28,29)]
  newtest<-testing[,c(2,3,4,5,6,7,8,9,10,11,12,13,14,22,23,24,26,27,28,29)]
  head(newtrain$cust_satisfaction)
```

```
  1  1  1  1  1  1
```

▶ **Levels**:

```
: myOutput<-ksvm(cust_satisfaction ~., data=newtrain, kernel = "rbfdot",kpar="automatic", C=5,cross=3, prob.model=TRUE)
```

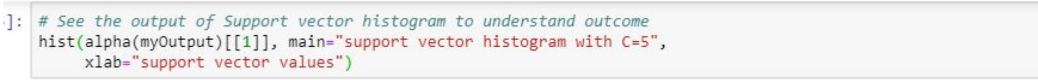
Output:



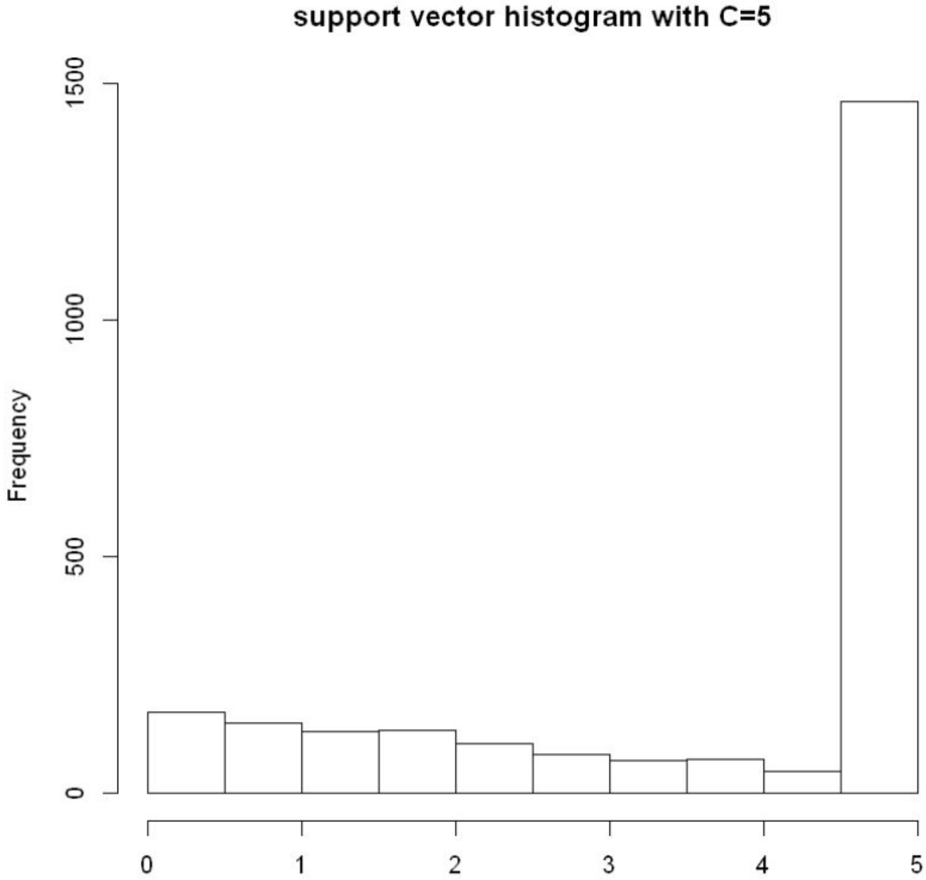
```
myOutput
```

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

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

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.0468980174532648

Number of Support Vectors : 2421

Objective Function Value : -7710.114
Training error : 0.093164
Cross validation error : 0.13899
Probability model included.
```

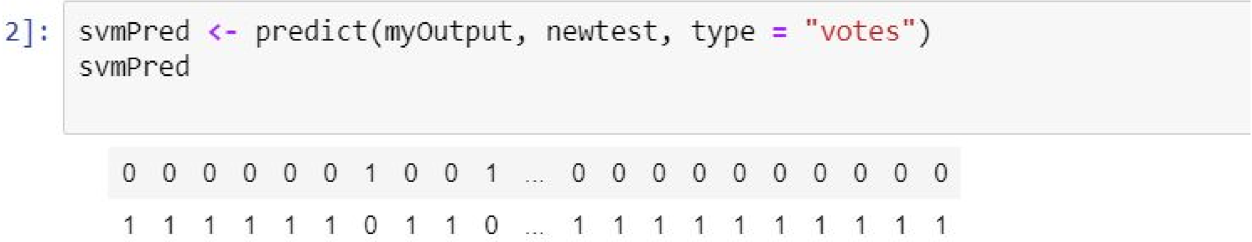
Supporting Histogram:



```
]: # See the output of Support vector histogram to understand outcome
   hist(alpha(myOutput)[[1]], main="support vector histogram with C=5",
        xlab="support vector values")
```

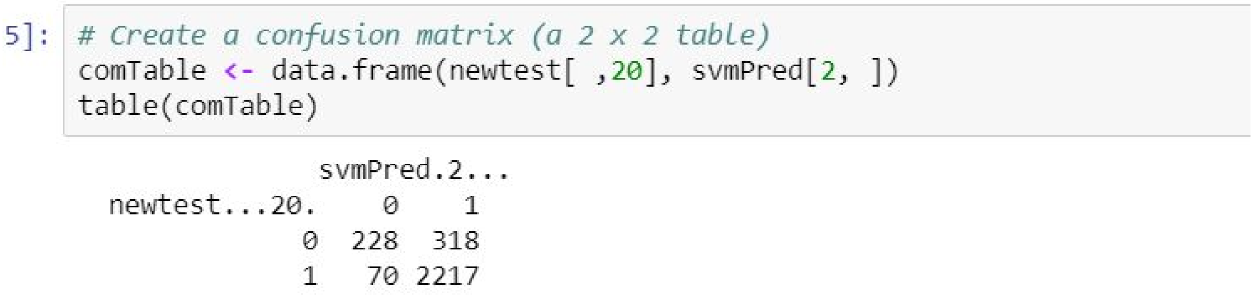


Prediction And Accuracy:



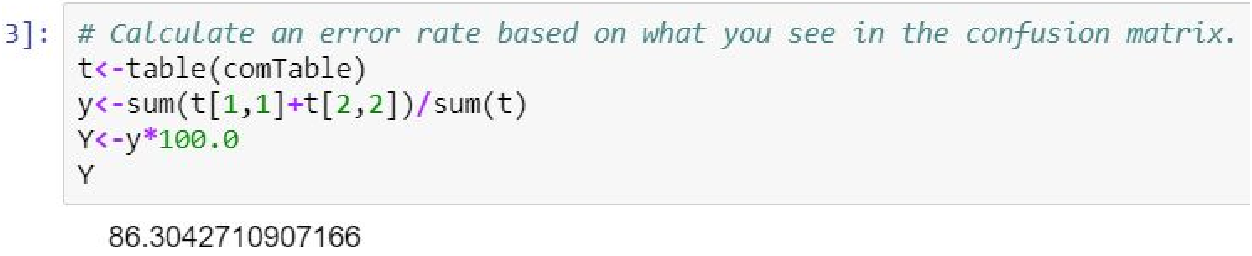
```
2]:  svmPred <- predict(myOutput, newtest, type = "votes")
     svmPred
```

```
0  0  0  0  0  0  1  0  0  1  …  0  0  0  0  0  0  0  0  0  0
1  1  1  1  1  1  0  1  1  0  …  1  1  1  1  1  1  1  1  1  1
```



```
5]:  # Create a confusion matrix (a 2 x 2 table)
     comTable <- data.frame(newtest[ ,20], svmPred[2, ])
     table(comTable)
```

```
              svmPred.2...
newtest...20.    0    1
            0  228  318
            1   70 2217
```



```
3]:  # Calculate an error rate based on what you see in the confusion matrix.
     t<-table(comTable)
     y<-sum(t[1,1]+t[2,2])/sum(t)
     Y<-y*100.0
     Y
```

```
86.3042710907166
```

7. Business Questions

A vast survey of customers flying within the United States was provided which gives information regarding the factors that drive the overall customer satisfaction. Analyzing these factors would help Southeast Airlines to improve their quality and further boost their business.

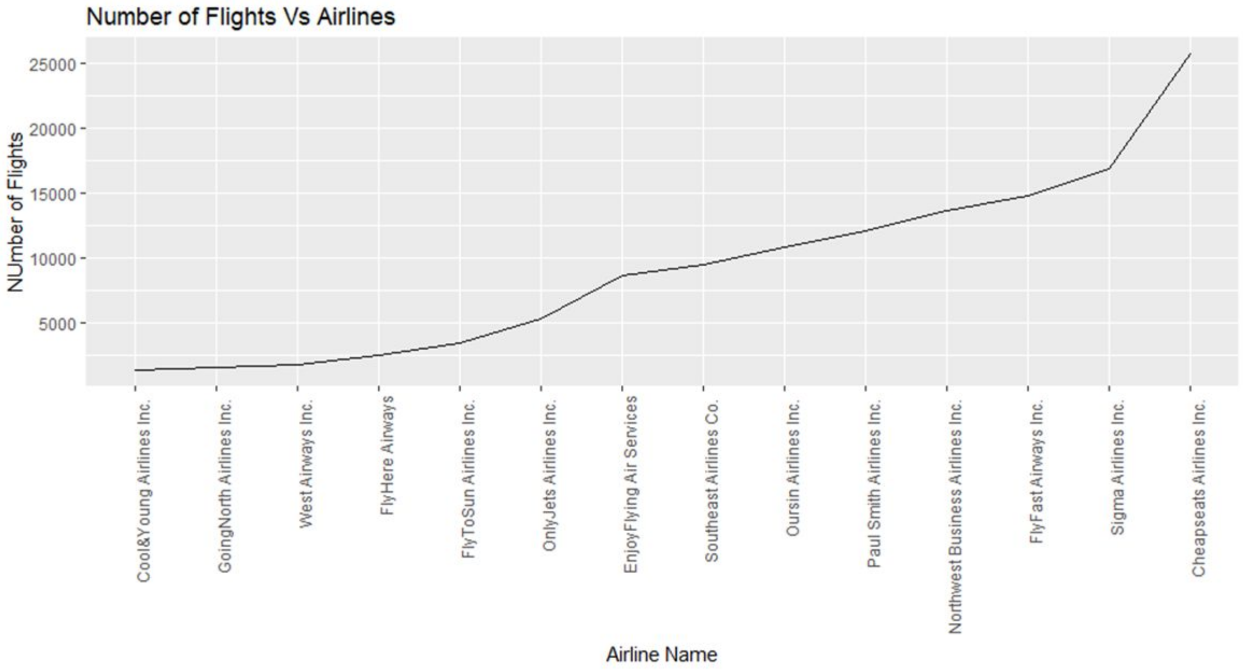
The business questions that have been recognized and answered through the project are as follows:

1. Does type of travel, loyalty cards and price sensitivity have an impact on rating by customer?
2. Does delay in arrival and delay of flights causes dissatisfaction?
3. How does age and gender affects satisfaction and are the significant for Southeast and Other Airlines?
4. Are the customers of Southeast Airlines satisfied as compared to other Airlines?
5. How does the delay in flight by more than 5 minutes affects the ratings?
6. Do particular city or state causes dissatisfaction?
7. Does date of travel and day of month affects customer satisfaction?
8. Is it possible to associate no of flights taken, flight distance and flight time to ratings given by customer?
9. How much does shopping and drinking at airport causes variation in rating given by customer?

8. Descriptive Statistics and Visualizations

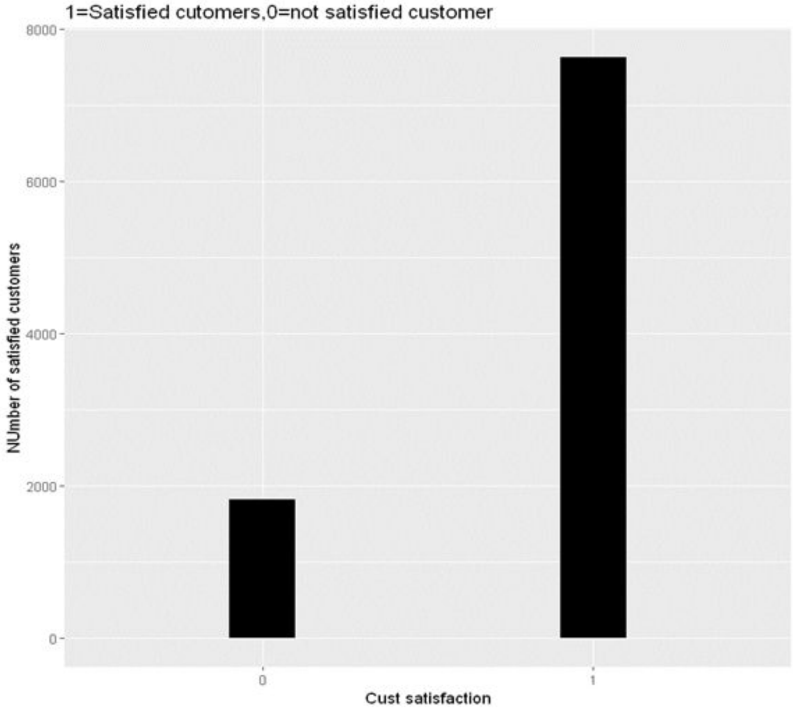
The different types of descriptive analysis performed to understand the data and gain some insights from it are as follows:

1. Number of flights vs Airlines

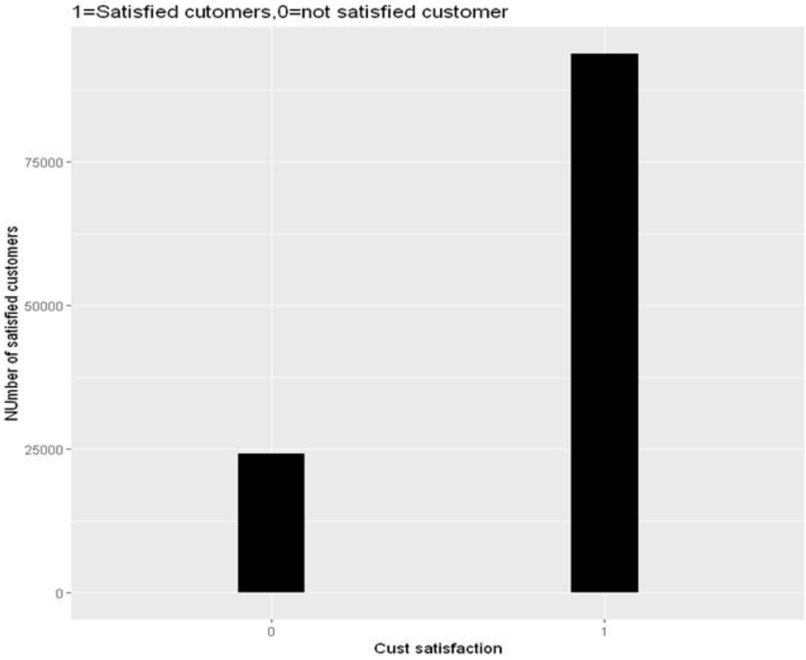


This graph helps us to understand the position of Southeast Airlines in the industry. The number of flights by Southeast Airlines is average and there are many other airlines which has more number of flights. This implies that Southeast Airlines needs to attract more customers to pose a competition. This can be done by analyzing the customer survey and increasing the overall customer satisfaction.

2. Customer Satisfaction



Southeast Airlines

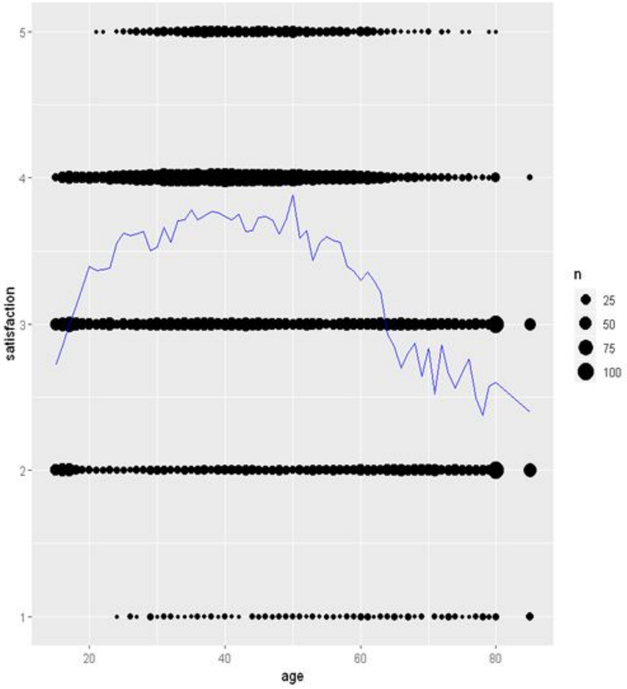


Other Airlines

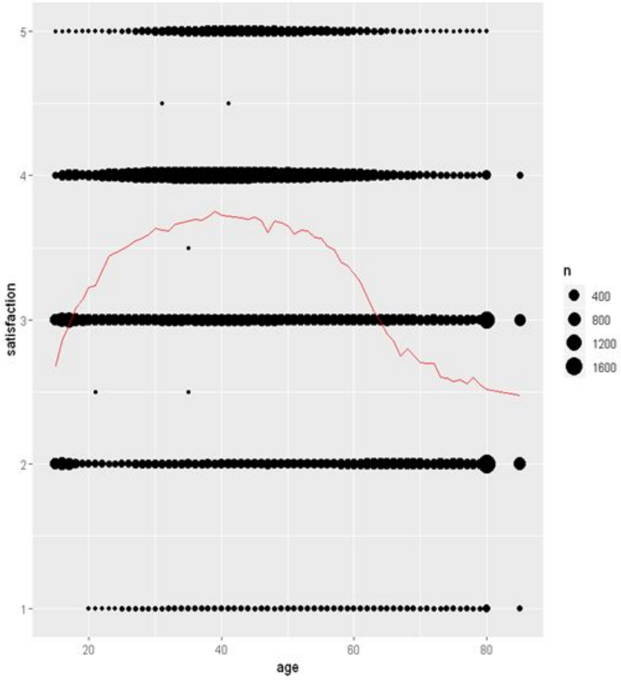
Here we analyzed the overall customer satisfaction of all the other airlines and Southeast Airlines. We found that the average number of unsatisfied customers for

Southeast Airlines is low than that of other airlines,. The average number of satisfied customers is quite similar.

3. Age vs Satisfaction



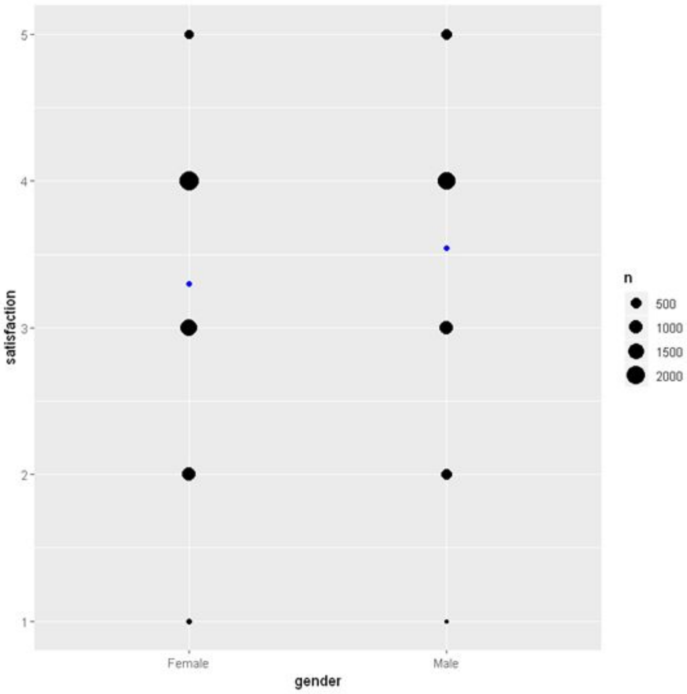
Southeast Airlines



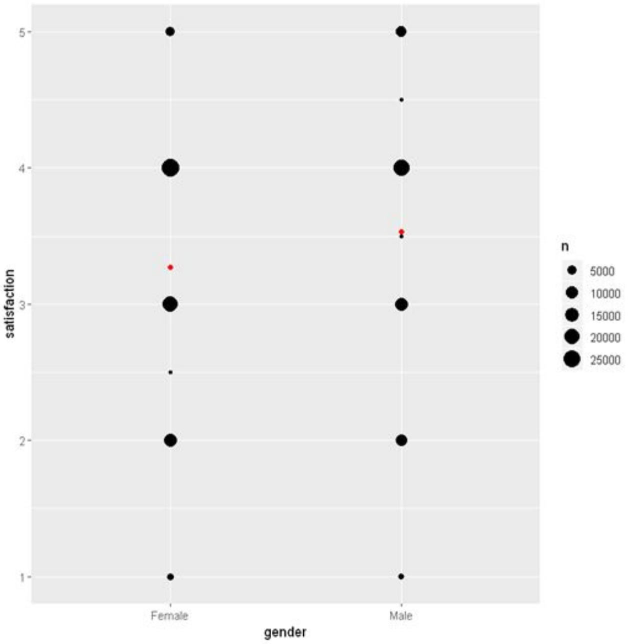
Other airlines

This graph shows that how different age groups affect the overall satisfaction. We found that people who are below 30 tend to be less satisfied as compared to other age groups. The line passing through the graph indicates the average satisfaction The overall satisfaction remains constant for age group 30-60 and decreases for the age group of 60-80. Age vs Satisfaction is very similar for both Southeast airlines and other airlines.

4. Gender vs Satisfaction



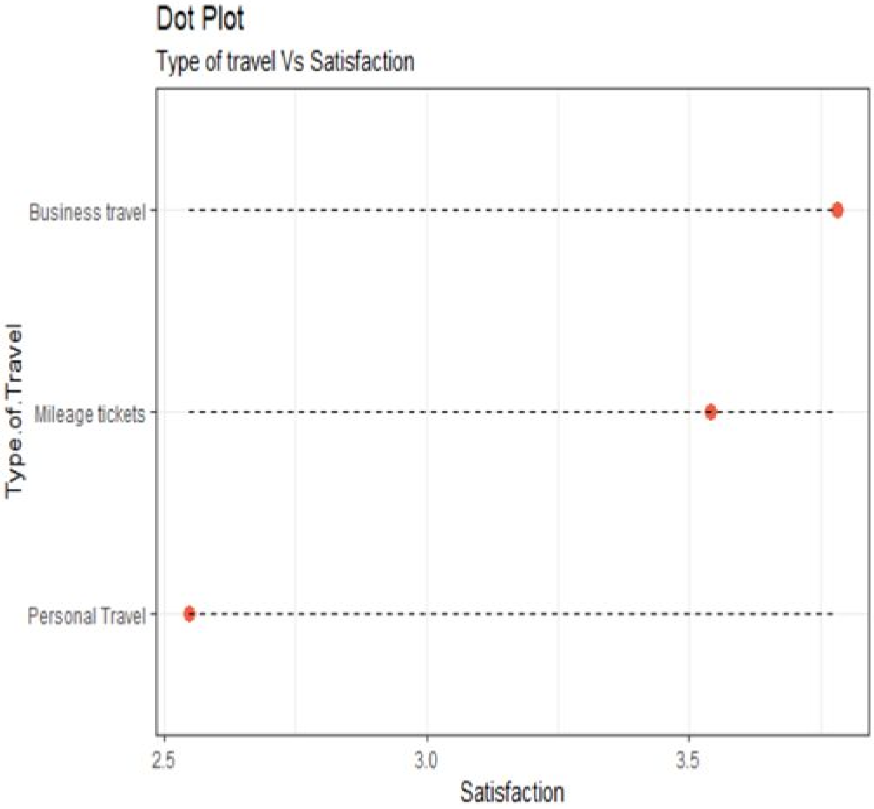
Southeast Airlines



Other airlines

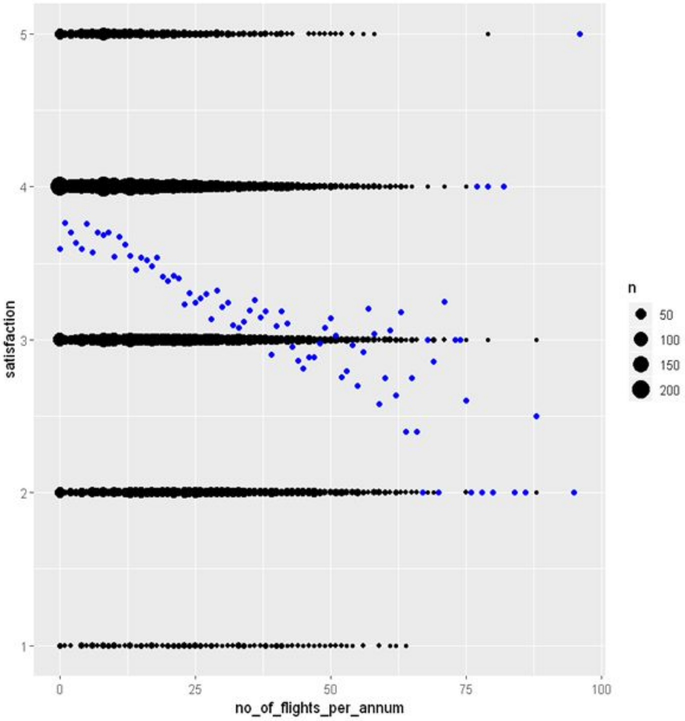
The blue and the red dot shows the average satisfaction for that gender. From the graph, we can infer that women are less satisfied than men for all the airlines. As there are more female customers than male customers we need to focus on factors responsible for their satisfaction.

5. Type of travel vs Satisfaction

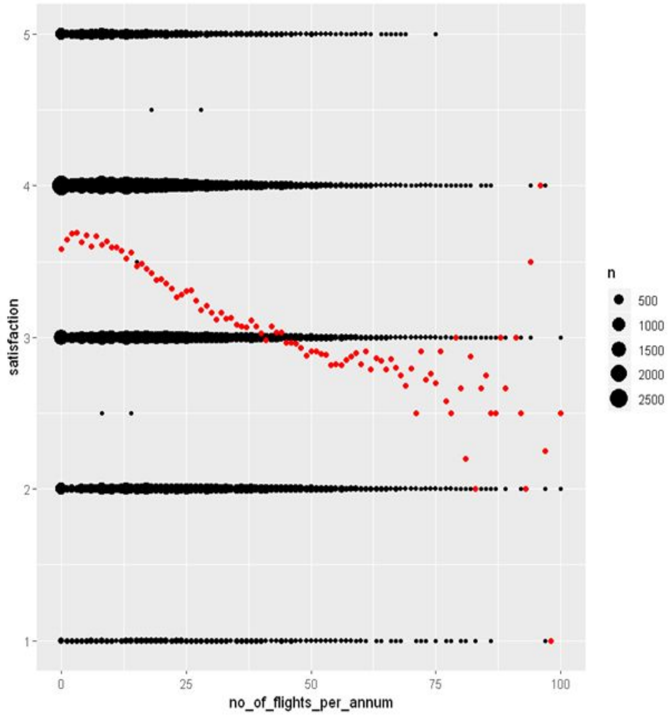


This graph shows the overall satisfaction of customers based on type of travel. This shows that the business travel customers has the highest satisfaction and personal travel customers has the lowest satisfaction. Since majority of the customers belong to personal travel group we need to improve our quality provided to personal travel customers.

6. Number of flights per annum vs Satisfaction



Southeast Airlines



Other Airlines

This graph shows that as the number of flights per annum increases the satisfaction goes on decreasing. This implies that the satisfaction of frequently flying customers should be increased.

9. Interpretations

From the above graphs, we interpreted the following points:

- Satisfaction tend to reduce as arrival delay increases
- Satisfaction reduces by around 0.5 If the arrival delay is greater than 5 mins for South east and other airlines.
- Departure delay affects the Satisfaction rating by a small amount. There is evidence that the satisfaction reduces as departure delay increases, but this is not true for all cases.
- Age affects satisfaction in both Southeast and other airlines. For the age group 20 to 50, the satisfaction increases. The satisfaction decreases as the age increases from 50.
- When the customers travel for personal reasons, their satisfaction is the lowest which is around 2.5
- For Southeast airlines, satisfaction tends to increase as the number of flights p.a. decrease.
- In both Southeast Airlines and other airlines, customers travelling by Blue status have the lowest satisfaction rating, followed by Platinum status.
- Customers who shop between 600-800 give a high customer satisfaction rating for other airlines whereas those who shop for 500-600 give a high customer rating for Southeast airlines.

10. Project link

https://midst.syr.edu/project/data_science_project