SOUTHEAST AIRLINE CORPORATION

IST687-M006-Group2

Xin (Harper) He
Jingxian (Sandra) Sun
Dharmik Gautam Kothari
Sakshi Raghuvanshi

Contents

Introduction 2
Business Questions addressed 3
Data Acquisition, Cleansing, Transformation, Munging 3
    Data Acquisition 3
    Data Cleansing 3
    Data Transformation 4
    Data Munging 5
    Attributes grouping 5
Descriptive statistics & Visualizations 6
Use of modeling techniques & Visualizations 17
Association Rules Mining 17
Linear Modeling 24
        Satisfaction VS Customers characteristic 24
        Satisfaction VS Flight experience characteristic 27
        Satisfaction VS Flight characteristic 29
        Satisfaction VS all the other attributes 31
Support Vector Machine 33
Actionable Insights / Overall interpretation of results 36
Limitation 37
Trello board: 37
Division of work： 38
Appendix – Code 39

Introduction

The aviation sector is a highly competitive environment where high quality services to customers has become the core competitive advantage for an airline's growth. Service quality not only influences an airline’s customer loyalty, but also the profitability and its market share. Southeast Airline is one of the biggest airline company in United State and occupied decent market share. To improve the business performance for Southeast Airline, the research is conducted by accumulating thousands of customers survey and analyzing in statistics techniques.

As we known, customers’ satisfaction reflects consumers’ overall impression of the airline’s services. Therefore, our client need to gain a better understanding of the factors that affect customers’ satisfaction and allocate efforts to improve. Moreover, the results from this research would assist not only Southeast, but industry to develop service quality to achieve a higher level of customer’ satisfaction.

The goal of this research is to examine the satisfaction of customers of different airlines and find the factors that affect it.

This study is also conducted with two objectives:

- To explore the level of customers’ satisfaction with different airlines in terms of three dimensions: Customers characteristic, Flight experience characteristic and Flight characteristic
- To examine the demographic profile of the respondents

We used R to perform data analysis and visualization for the airline industry customer satisfaction survey dataset. The analytic techniques used in this report are Support Vector Machine, Association Rules Mining and Linear Modeling.

Business Questions addressed

1. Rank the Airlines and find the client's rank based on the current survey.
2. Find out satisfaction levels of different age groups across airlines.
3. Do people who fly a lot tend to have a higher satisfaction?
4. Compare the various classes based on the customer satisfaction.
5. Do people who depart from/to certain cities have a lower satisfaction rate?
6. Which attribute influence the satisfaction of costumers the most?

Data Acquisition, Cleansing, Transformation, Munging

Data Acquisition

The source data can be found at this link in the form of a cvs file.

The original data set was originally collected in 2014. It has 129,889 observations and 28 variables. 14 of the variables are of factor type while others of integer type.

There are quite a few numbers of missing values in the dataset. But they have been left blank, or in other words have not been imputed in any form (manipulated in some ways for later use). If the data does not exist it is a blank space filled string, which has been taken care of in data cleansing.

Data Cleansing

To clean the data, we first inspected the source data and noted that there are missing values for some observations for the "Departure.Delay.in.Minutes", "Arrival.Delay.in.Minutes" and "Flight.time.in.minutes" fields, which indicates that the unique customers did not take the plan or take the plane but couldn't remember the flight length. However, these null values are not just missing data, but actually helps us to group customers according to their flight status – "cancelled" and "uncancelled", which correspond to the "Flight.cancelled" field. So, we used the observations with missing values to build a subset and then focus on the observations that don't have missing values.

In addition, we noted that the "Satisfaction" field has 9 data entry errors. So, we remove these 9 observations. Two attributes "Flight data" and "Airline Code" are not used in our study.

After these cleansing, our final dataset contains 129,880 observations and 26 variables. These observations can be divided into different subsets by flight status, airlines and so on.

Below mentioned is the summary of the concerned variables, including data type, number of missing values if any, and value/range:

Data Transformation

For the analysis, some variables had to be changed into some other datatype.

To do the linear modeling and association rules mining, we mapped some attributes of integer to factor type. Attributes that have been converted are listed below:

To do the support vector machines, we converted the attributes that have a numeric range into buckets (ex. low or high). Attributes that have been converted are listed below:

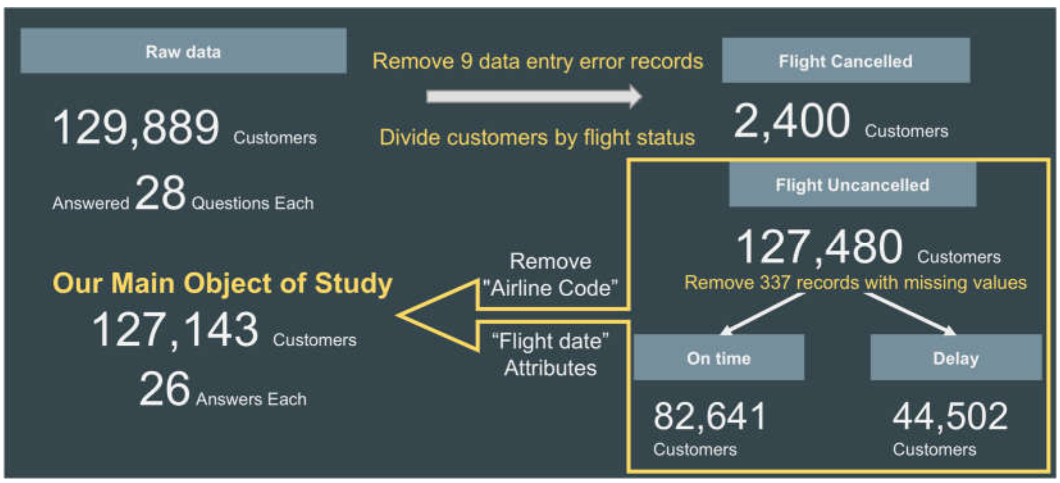
Attribute	Original type	Attribute (after transformation)	Transformed type	Transformation criteria and categories
Age	int	Age.Group	factor	15-18,19-24,25-34,35-44,45-54,55-64,65+
Price.Sensitivity	int	Price.Sensitivity.Group	factor	sensitive (4-5) not sensitive (1-3)
Year.of.First.Flight	int	Year.of.First.Flight.Group	factor	2003-2007 2008-2012
Satisfaction	fac	Satisfaction	num	
Satisfaction	fac	Satisfaction.Group	factor	unsatisfied (1-3) satisfied(4-5)
No.of.Flights.p.a.	int	No.of.Flights.p.a.Group	factor	Low (below 40%) Average (40%-60%) High (above 60%)
No..of.other.Loyalty.Cards	int	No..of.other.Loyalty.Cards.Group	factor	
Departure.Delay.in.Minutes	int	Departure.Delay.in.Minutes.Group	factor	
Flight.time.in.minutes	int	Flight.time.in.minutes.Group	factor	
Flight.Distance	int	Flight.Distance.Group	factor	

Scheduled.Departure.Hour	int	Scheduled.Departure.Hour.Group	factor	early morning(1am-5am)
				morning(6am-11am)
				afternoon(12pm-5pm)
				evening(6pm-11pm)

Data Munging

We create some subsets according to different attributes because for our analysis, sometimes only a small portion of it is useful.

Below is the summary of the subsets and their grouping dimensions:



Among the customers whose flight hasn't been cancelled, we exact the customers of Southeast Airline as "southeast" subset. Our analysis focuses on the "uncancelled" subset and "southeast" subset. This study mainly focuses on the customers whose flight had not been cancelled.

Attributes grouping

For deeper analysis, we group attributes according to their characteristics and use them in modeling techniques.

1) Customers characteristic (5 attributes)
    a) Demographic: Age, Gender
    b) Consuming behavior: Shopping Amount at Airport; Eating and Drinking at Airport; Price Sensitivity
2) Flight experience characteristic (7 attributes)

a) Previous flight experience: Year of First Flight; No of Flights, Percent of Flight with other Airlines, No. Of other Loyalty Cards

b) Current flight experience: Airline Status, Type of Travel, Class,

3) Flight characteristic (12 attributes)

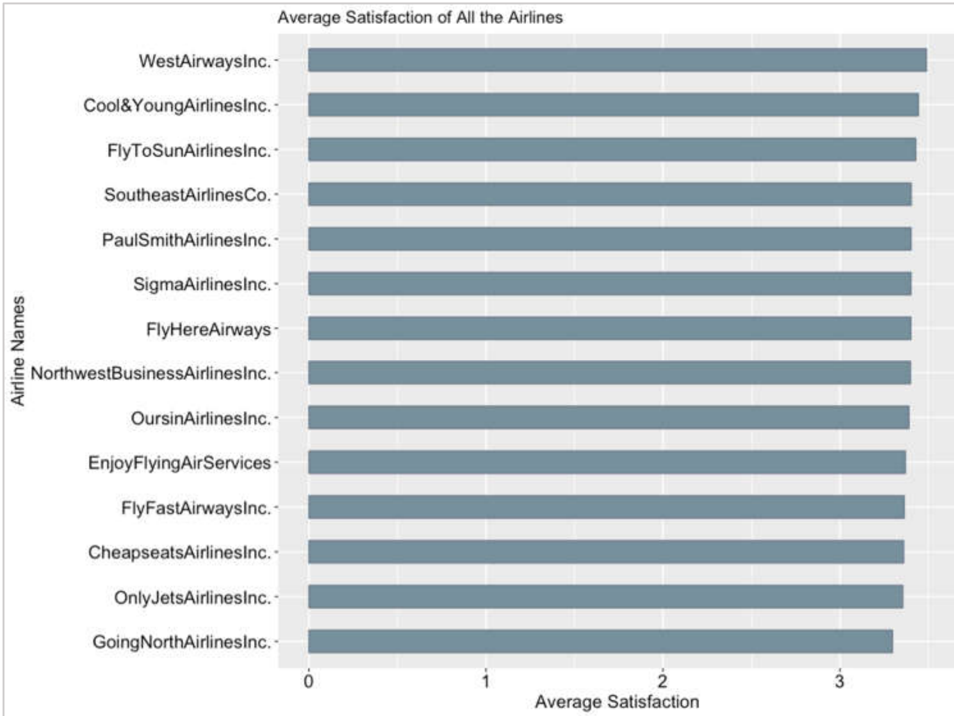
a) Geography: Origin City, Origin State, Destination City, Destination State, Flight Distance

b) Delay and cancellation: Flight date, Scheduled Departure Hour, Departure Delay in Minutes, Arrival Delay in Minutes, Flight time in minutes, Arrival Delay greater 5 Mins, Flight cancelled

Descriptive statistics & Visualizations

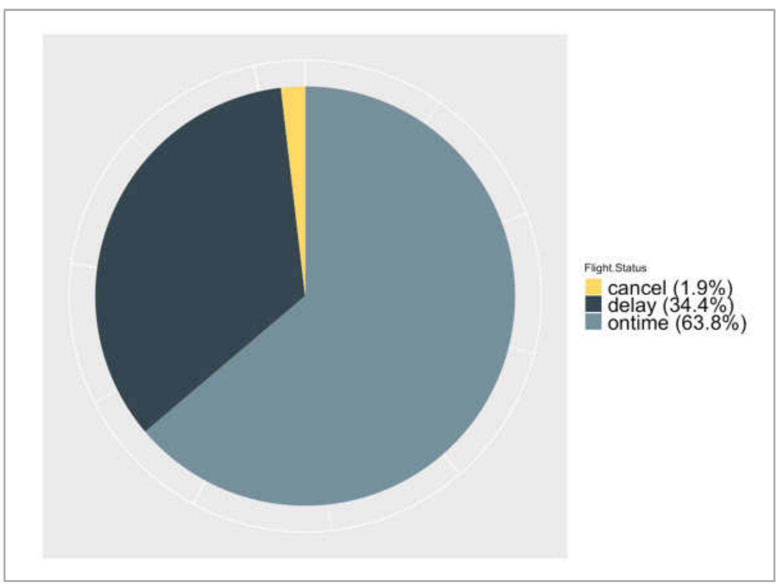
The satisfaction of customers differs on airline companies

Airline.Name	CustomerNumber	AverageSatisfaction
WestAirwaysInc.	1685	3.488427
Cool&YoungAirlinesInc.	1280	3.442969
FlyToSunAirlinesInc.	3372	3.428233
SoutheastAirlinesCo.	9423	3.402738
PaulSmithAirlinesInc.	12051	3.401709
SigmaAirlinesInc.	16801	3.401167
FlyHereAirways	2423	3.400743
NorthwestBusinessAirlinesInc.	13539	3.398405
OursinAirlinesInc.	10800	3.389537
EnjoyFlyingAirServices	8584	3.369874
FlyFastAirwaysInc.	14695	3.363185
CheapseatsAirlinesInc.	25669	3.359850
OnlyJetsAirlinesInc.	5259	3.354820
GoingNorthAirlinesInc.	1562	3.296415



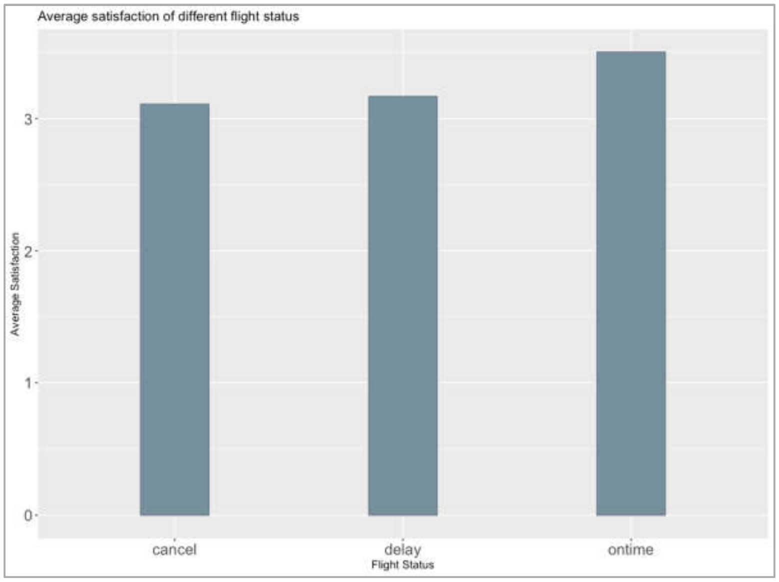
This plot shows how the average satisfaction differs on airline companies. All the airline companies have a satisfaction score between 3.2 and 3.5. There is no obvious difference between different companies. And our client, Southeast Airlines Co. ranked 4 in all the airlines.

The satisfaction of customers differs on flight status



The pie chart shows the proportion of customers based on their flight status: cancelled, delay or on time. Nearly 2% customers' flights are cancelled and more than one third customers have their flights delayed.

Flight.Status	Number.of.customers	Average.Satisfaction
cancel	2400	3.108750
delay	44502	3.166936
ontime	82641	3.501337



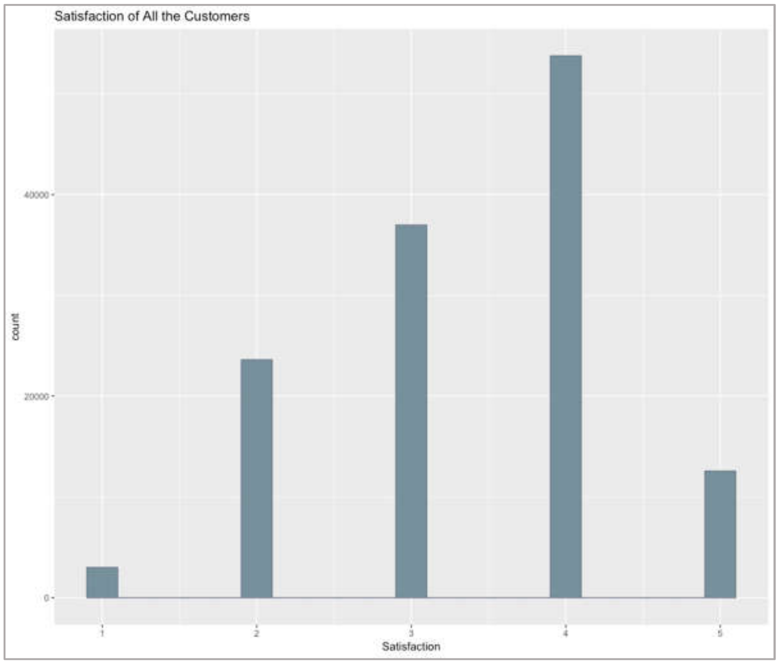
The table and bar chart above show the average satisfaction of customers on different flight status. The interesting phenomenon here is that the satisfaction of customers whose flights have been delayed is closer to those cancelled, which tells us the delay may affect customers' satisfaction strongly.

Summary statistics of satisfaction

Below mentioned is the summary of the most important attribute, the satisfaction:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.000	3.000	4.000	3.379	4.000	5.000

Satisfaction	CustomerNumber
1	2999
2	23587
3	36984
4	53758
5	12552



This plot is a histogram and shows the distribution of Satisfaction of all the airlines. The data spread is from 1 to 5, means the scale of satisfaction is 1 to 5.

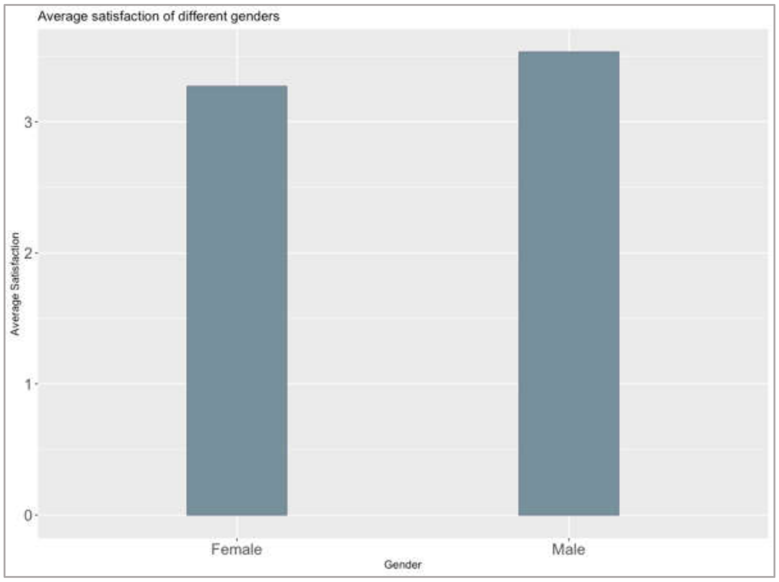
The peak of the data occurs at 4, means most customer rated their satisfaction as 4.

The data are left-skewed, means most of the customer's satisfaction rates are clustered on the left side of the histogram, which is 1 to 4.

There are more than 50000 customers rated 4, more than 35000 rated 3 and more than 20000 rated 2. Over 10000 customers were very satisfied with the airlines and rated 5 while nearly 50000 customers were not satisfied at all.

The satisfaction of customers differs on gender

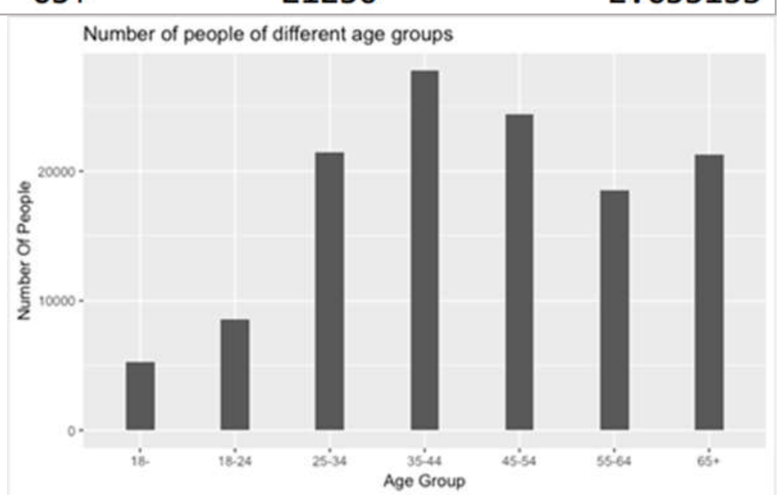
Gender	CustomerNumber	AverageSatisfaction
Female	71683	3.270371
Male	55460	3.531536



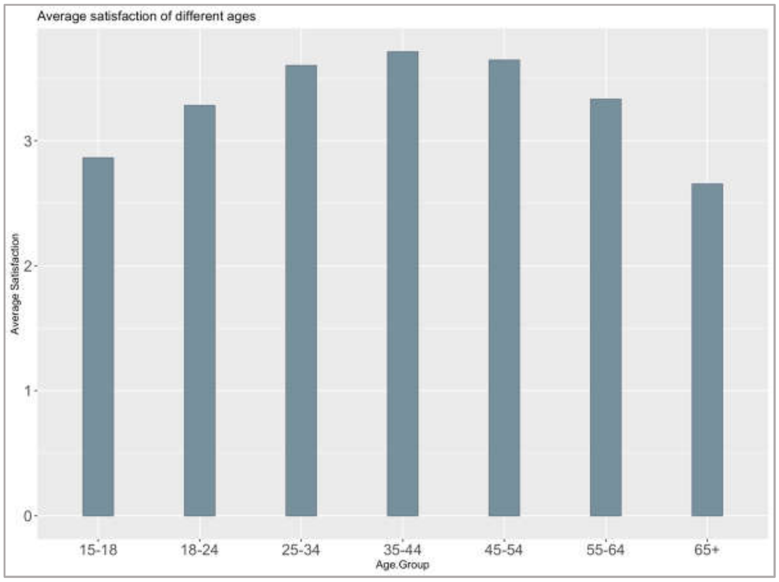
This plot shows the average satisfaction of different genders of all the airlines.
The average satisfaction of male is about 3.5 while for female, the average satisfaction is around 3.25. So, the males tend to have a higher satisfaction compared to female.

The satisfaction of customers differs on age

Age.Group	CustomerNumber	AverageSatisfaction
15-18	5252	2.864242
18-24	8546	3.282003
25-34	21420	3.601541
35-44	27752	3.710940
45-54	24374	3.644867
55-64	18549	3.331986
65+	21250	2.655153



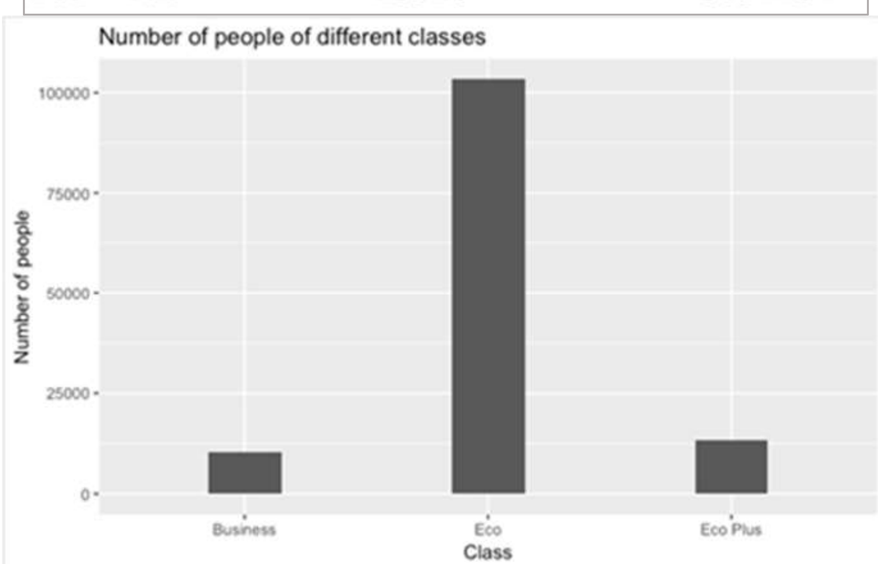
This plot shows the age distribution of all the customers surveyed of all the airlines. Most customers surveyed are among 25 to 54, nearly 90000. There are about 40000 customers surveyed of 55 or above. Other customers surveyed, about 14000 people, are under 24 years old.



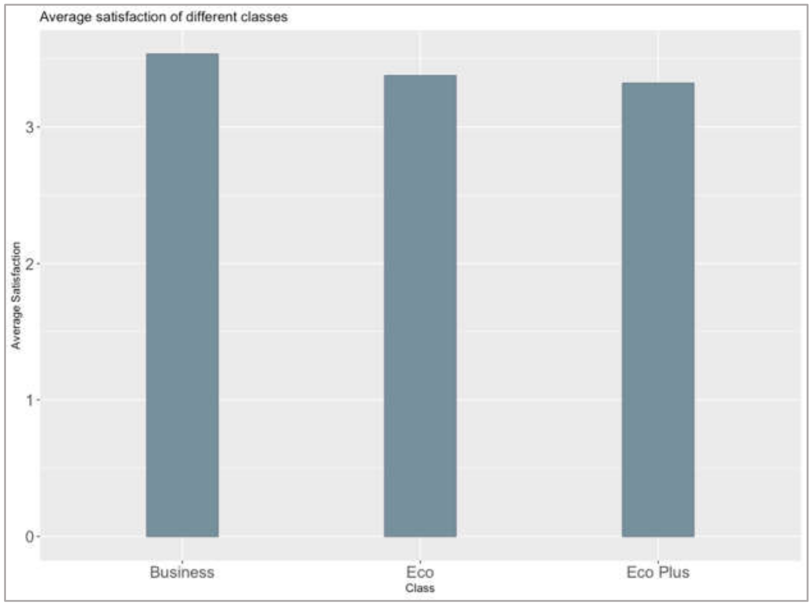
This plot shows the average satisfaction of different age groups of all the airlines. Customers of 18-24 and 55-64 have satisfaction about 3.25 while customers of 25-54 have a higher rate, above 3.5. However, the old tend to have a lowest satisfaction, the teenagers of 14-18 also have a lower satisfaction.

The satisfaction of customers differs on class

Class	CustomerNumber	AverageSatisfaction
Business	10452	3.535113
Eco	103375	3.377074
Eco Plus	13316	3.321944



This plot shows the class distribution of all the customers surveyed of all the airlines. Most customers surveyed took the Eco class, more than 100000. There are about 13000 customers took Eco Plus. The number of customers who took business class is less than 12000.

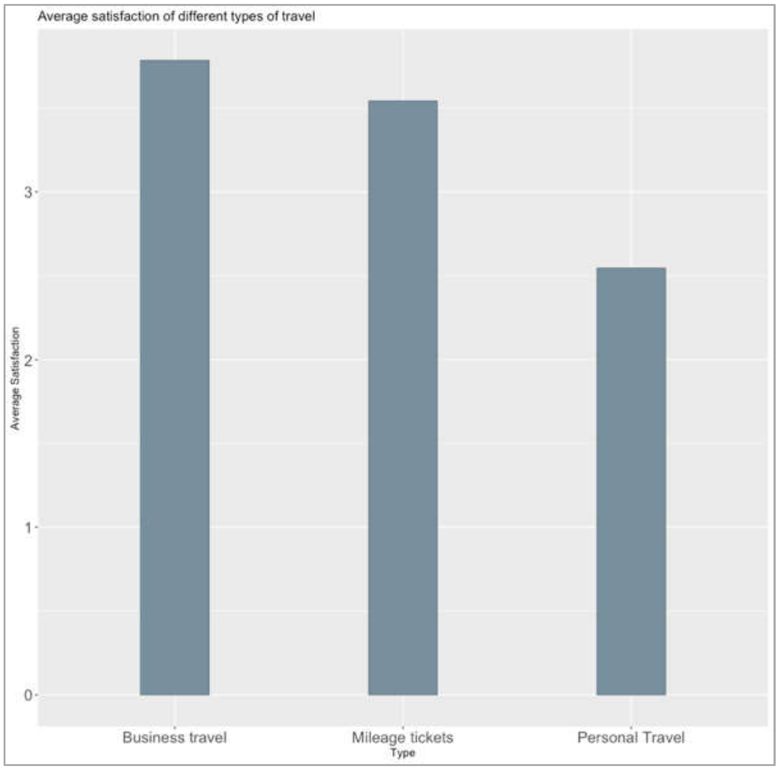


This plot shows the average satisfaction of different classes that customers took of all the airlines.

Customers who took business class tend to have the highest satisfaction, over 3.5. The satisfaction of customers who took Eco class tend to have a lower satisfaction, about 3.4. However, people who took Eco Plus have the lowest satisfaction, around 3.3.

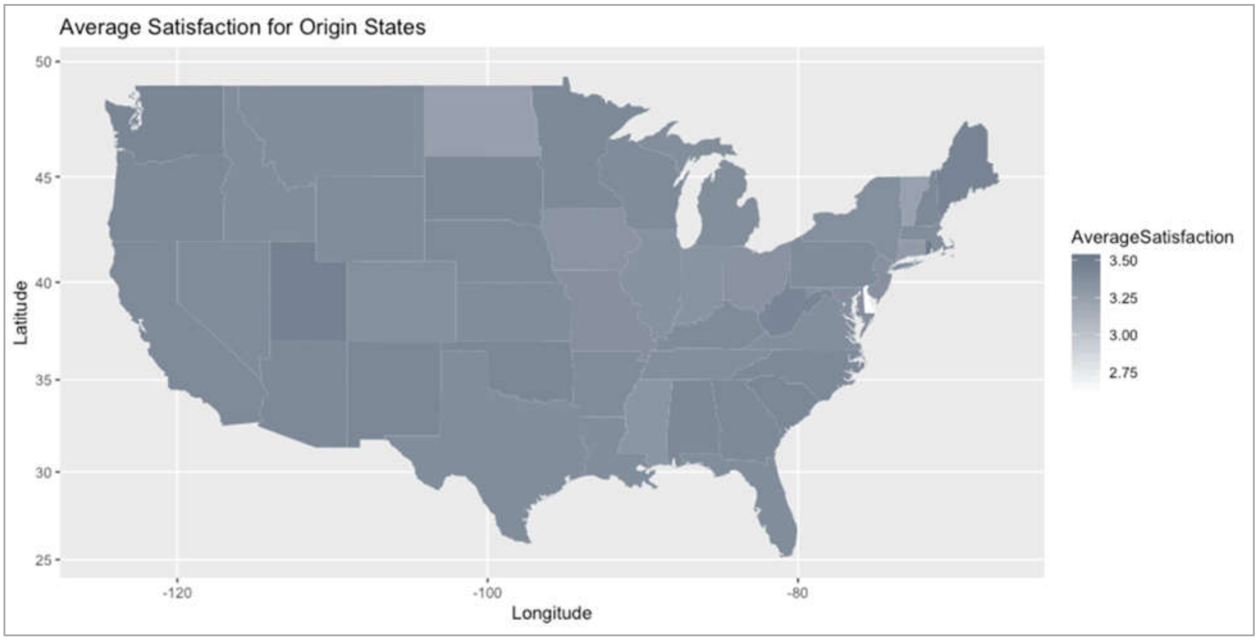
The satisfaction of customers differs on type of travel

Type.of.Travel	CustomerNumber	AverageSatisfaction
Business travel	78379	3.781600
Mileage tickets	9817	3.541000
Personal Travel	38947	2.545228



The satisfaction of customers differs on origin states

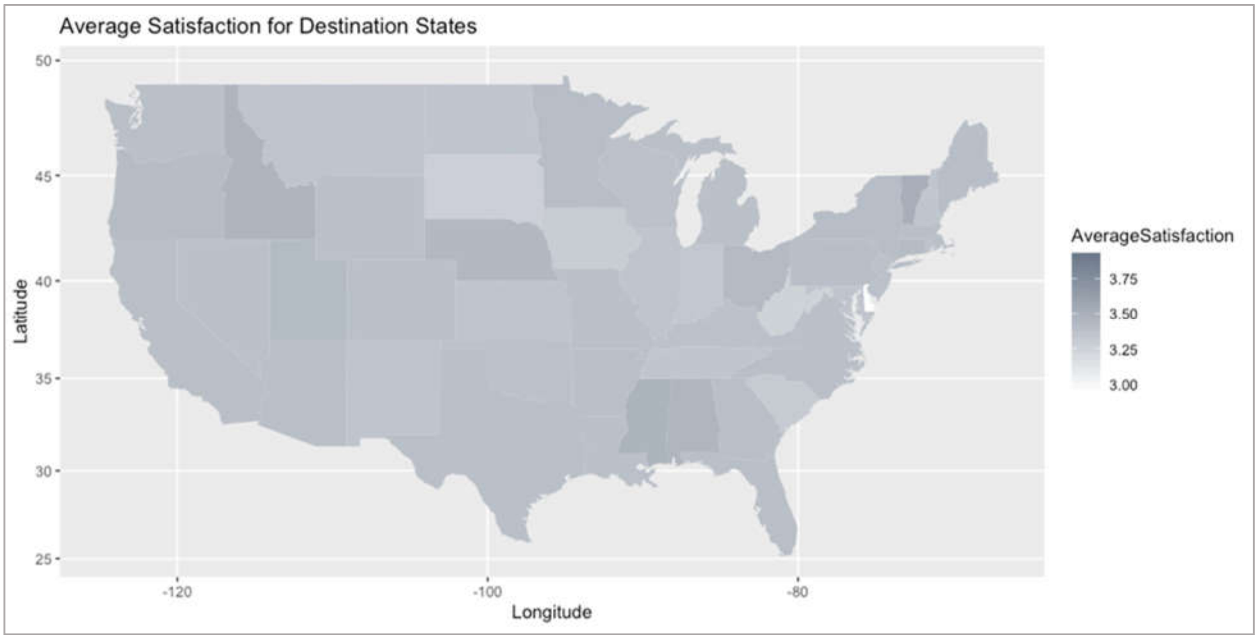
Origin.State	CustomerNumber	AverageSatisfaction
delaware	10	2.600000
vermont	54	3.240741
north dakota	239	3.246862
connecticut	478	3.276151
mississippi	298	3.315436
iowa	374	3.326203
indiana	860	3.333721
ohio	1817	3.336819
illinois	7640	3.341230
new jersey	2387	3.352744



By mapping the origin state, we found ten origin states with the lowest satisfaction rate. We used ggplot to plot and fill the states by the rule "the lighter the color is, the lower the average satisfaction is" as above.

The satisfaction of customers differs on destination states

Destination.State	CustomerNumber	AverageSatisfaction
Delaware	16	2.937500
West Virginia	60	3.250000
South Dakota	242	3.272727
Iowa	388	3.301546
South Carolina	585	3.307692
Indiana	846	3.330969
Tennessee	1901	3.340347
New Hampshire	140	3.342857
North Dakota	242	3.342975
Montana	358	3.349162

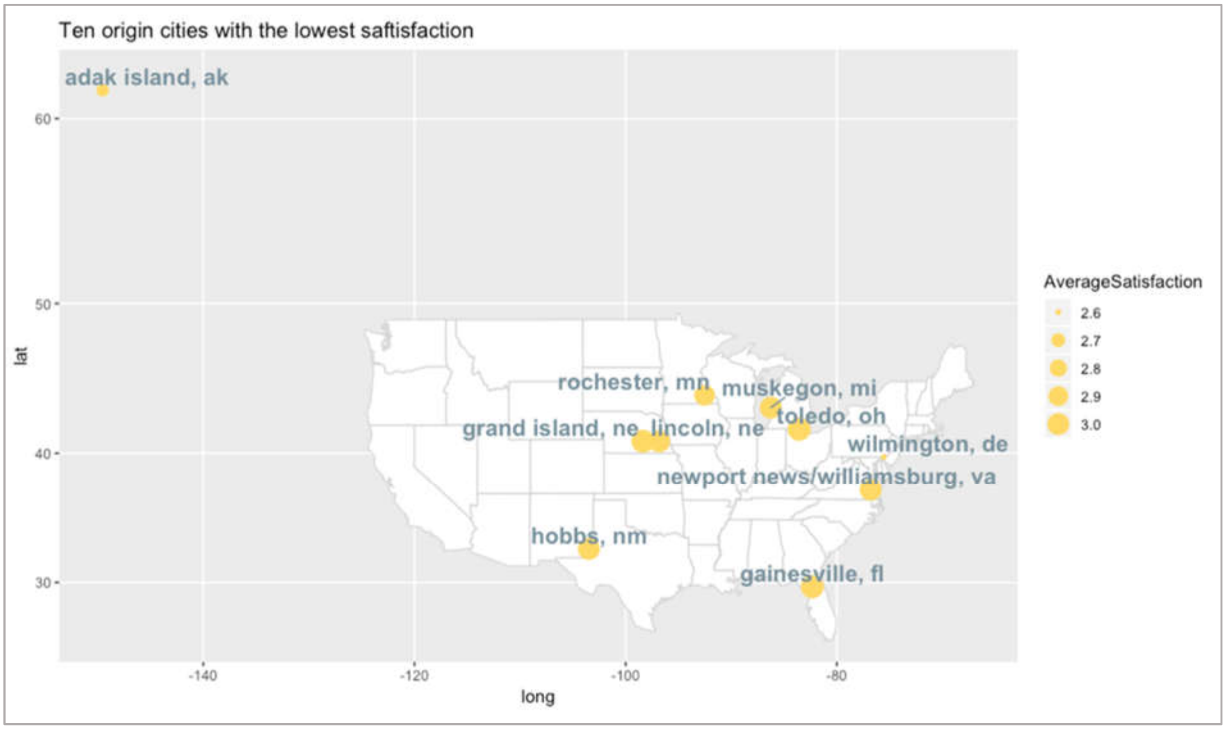


By mapping the destination states, we found ten destination states with the lowest satisfaction rate.

Information in this part helps our client pays attention to the service of these specific states.

The satisfaction of customers differs on origin cities

Orgin.City	CustomerNumber	AverageSatisfaction
Wilmington, DE	10	2.600000
Adak Island, AK	3	2.666667
Rochester, MN	16	2.937500
Hobbs, NM	12	3.000000
Lincoln, NE	35	3.000000
Muskegon, MI	12	3.000000
Newport News/Williamsburg, VA	45	3.000000
Grand Island, NE	21	3.047619
Toledo, OH	21	3.047619
Gainesville, FL	57	3.052632



Mapping out ten origin cities with the lowest average satisfaction level, we found these cities mainly located in the northeast. The reason could be the extreme weather in northeast. Our client can improve the services towards the flights from these cities.

The satisfaction of customers differs on destination cities

Destination.City	CustomerNumber	AverageSatisfaction
La Crosse, WI	11	2.727273
Joplin, MO	16	2.937500
Wilmington, DE	16	2.937500
Trenton, NJ	57	2.947368
Modesto, CA	23	3.000000
Watertown, NY	12	3.000000
Beaumont/Port Arthur, TX	22	3.045455
Williston, ND	41	3.097561
Traverse City, MI	26	3.115385
Toledo, OH	23	3.130435

Mapping out ten destination cities with the lowest satisfaction level, we found Wilmington and Toledo appeared in both "top ten map". Our client should have further improvement on services in these cities.

Use of modeling techniques & Visualizations

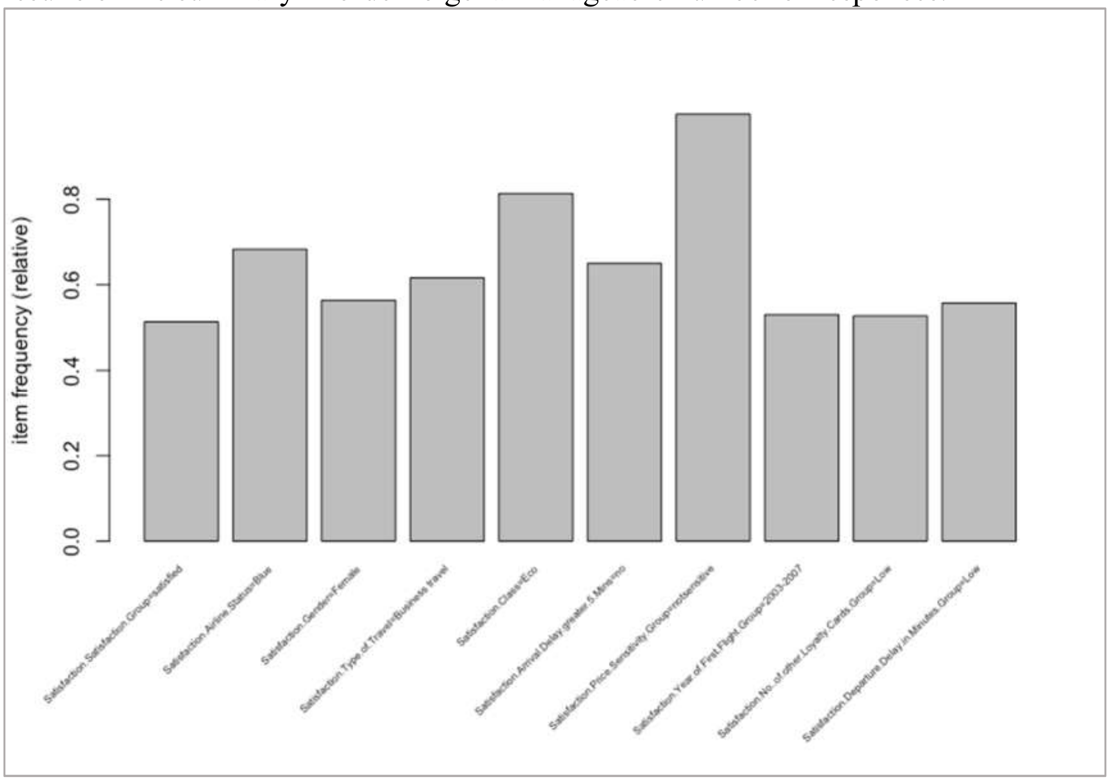
Association Rules Mining

```
transactions as itemMatrix in sparse format with
 127143 rows (elements/itemsets/transactions) and
 46 columns (items) and a density of 0.326087

most frequent items:
Satisfaction.Price.Sensitivity.Group=notsensitive                              Satisfaction.Class=Eco
                                           126951                                              103375
                 Satisfaction.Airline.Status=Blue      Satisfaction.Arrival.Delay.greater.5.Mins=no
                                            86848                                               82641
      Satisfaction.Type.of.Travel=Business travel                                             (Other)
                                            78379                                             1428951

element (itemset/transaction) length distribution:
sizes
    15
127143

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     15      15      15      15      15      15

includes extended item information - examples:
                                       labels                       variables      levels
1   Satisfaction.Satisfaction.Group=satisfied Satisfaction.Satisfaction.Group   satisfied
2 Satisfaction.Satisfaction.Group=unsatisfied Satisfaction.Satisfaction.Group unsatisfied
3            Satisfaction.Airline.Status=Blue      Satisfaction.Airline.Status        Blue

includes extended transaction information - examples:
  transactionID
1             1
2             2
3             3
```

To conduct the association rules mining, we create "SatiTrans", an itemMatrix object in sparse format. It is a rectangular data structure with 127,143 rows and 46 columns.
The output also shows us which responses occur in satisfaction survey most

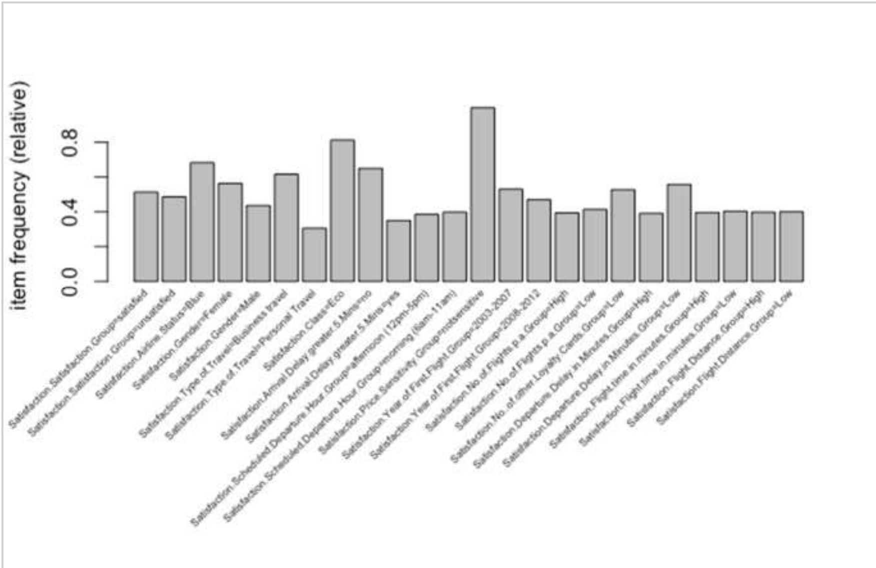
frequently. For example, there are 126,951 customers that are not sensitive to price, that is the grade to which the price affects to customers purchasing is not greater than 3.

Below is a bar graph that shows the relative frequency of occurrence of different responses in the matrix. We choose 0.5 as the minimum level of support based on the results of the summary in order to get a manageable number of responses.



We can see the relative frequency of occurrence of different items in the matrix. For example, the frequency of occurrence of “satisfied” response is more than 15% and the “business travel” response about 60%.

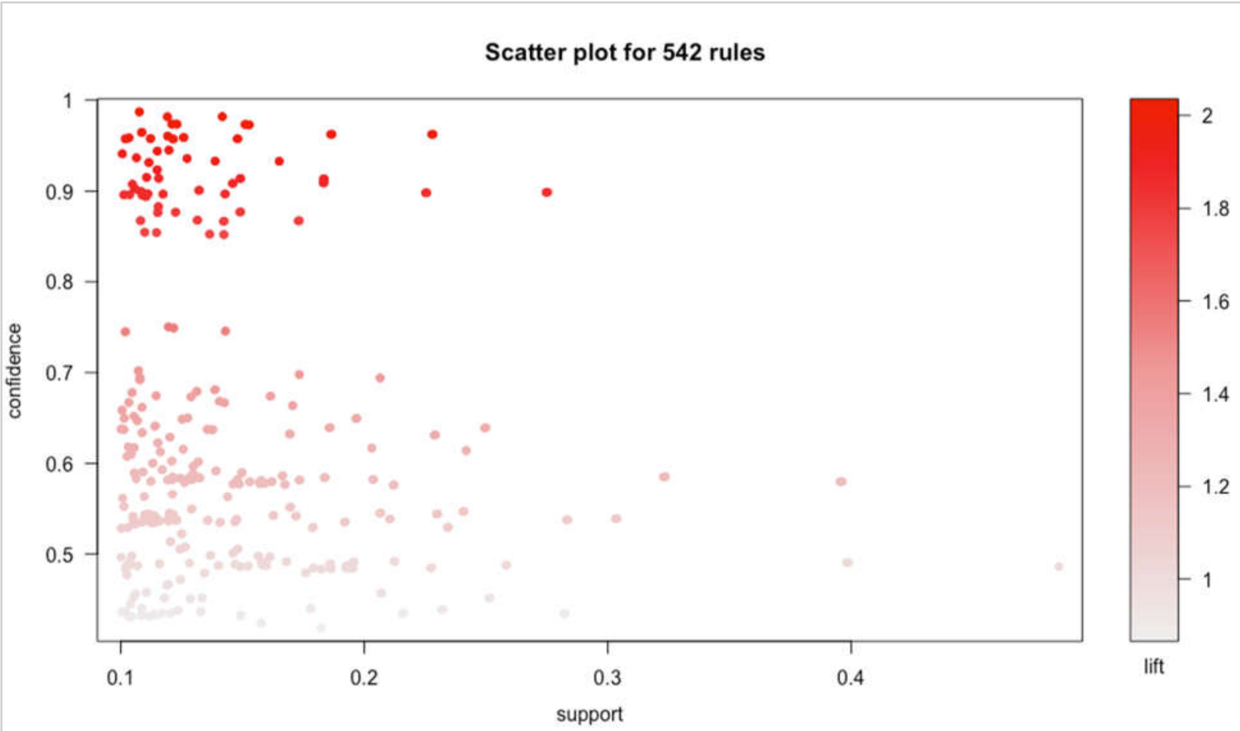
We also use different levels of support to get a sense of the other common responses in the data set. Below is an example with minimum level of support as 0.3.



This time it yields about 24 items on the X-axis. By trying different levels of support, we can guess a minimum support that will give us quite a substantial number of items that can potentially be part of a rule.

Below is a scatter plot of rules with minimum support as 0.1 and confidence as 0.4. It creates more than 500 rules. Rules that have high support, confidence and lift are what we want. On this plot, the lift is shown by the darkness of a dot. The darker the dot, the closer the lift of that rule is to 2, which appears to be the highest lift value among these 542 rules.

The support of 542 rules ranges from below 10% all the way up above 40%, all the rules with high lift seem to have support below 30% and confidence above 80%. On the other hand, there are rules with high lift and high confidence.



Based on this, we focus on a smaller set of rules that have only the very highest levels of lift. Below is a subset of the larger set of rules by choosing only those rules that have lift higher than 2.

From these results, we can conclude that:

1） Customers whose airline status is blue, having personal travel and having taken flights for many times tend to be unsatisfied.

2） Customers whose airline status is blue, having personal travel and having few other companies' loyalty cards tend to be unsatisfied.

3）Female customers whose airline status is blue having personal travel tend to be unsatisfied.

4) Customers whose airline status is blue, having personal travel, having few other companies' loyalty cards and having taken flights more frequently tend to be unsatisfied.

5) Customers whose airline status is blue, having personal travel, being in economy class and having few other companies' loyalty cards tend to be unsatisfied.

6) Customers whose airline status is blue, having personal travel, being not price sensitive and having taken flights more frequently tend to be unsatisfied.

7) Customers whose airline status is blue, having personal travel, being in economy class and having few other companies' loyalty cards tend to be unsatisfied.

8) Customers whose airline status is blue, having personal travel, being not price sensitive and having few other companies' loyalty cards tend to be unsatisfied.

9）Female customers whose airline status is blue having personal travel and being in the economy class tend to be unsatisfied.

10) Female customers whose airline status is blue having personal travel, being not price sensitive tend to be unsatisfied.

11) Customers whose airline status is blue, having personal travel, being in economy class and having few other companies' loyalty cards tend to be unsatisfied.

12）Customers whose airline status is blue, , having personal travel, being in economy class, being not price sensitive and having taken flights more frequently tend to be unsatisfied.

13) Customers whose airline status is blue, having personal travel, being in economy class, being not price sensitive and having few other companies loyalty cards tend to be unsatisfied.

14) Female customers whose airline status is blue having personal travel, being in economy class and being not price sensitive tend to be unsatisfied.

```
     lhs                                                   rhs                                          support confidence     lift count
[1]  {Satisfaction.Airline.Status=Blue,
      Satisfaction.Type.of.Travel=Personal Travel,
      Satisfaction.No.of.Flights.p.a.Group=High}        => {Satisfaction.Satisfaction.Group=unsatisfied} 0.1417380  0.9818033 2.018685 18021
[2]  {Satisfaction.Airline.Status=Blue,
      Satisfaction.Type.of.Travel=Personal Travel,
      Satisfaction.No..of.other.Loyalty.Cards.Group=Low} => {Satisfaction.Satisfaction.Group=unsatisfied} 0.1513729  0.9733475 2.001299 19246
[3]  {Satisfaction.Airline.Status=Blue,
      Satisfaction.Gender=Female,
      Satisfaction.Type.of.Travel=Personal Travel}      => {Satisfaction.Satisfaction.Group=unsatisfied} 0.1528830  0.9728729 2.000323 19438
[4]  {Satisfaction.Airline.Status=Blue,
      Satisfaction.Type.of.Travel=Personal Travel,
      Satisfaction.No.of.Flights.p.a.Group=High,
      Satisfaction.No..of.other.Loyalty.Cards.Group=Low} => {Satisfaction.Satisfaction.Group=unsatisfied} 0.1076190  0.9870158 2.029402 13683
[5]  {Satisfaction.Airline.Status=Blue,
      Satisfaction.Type.of.Travel=Personal Travel,
      Satisfaction.Class=Eco,
      Satisfaction.No.of.Flights.p.a.Group=High}        => {Satisfaction.Satisfaction.Group=unsatisfied} 0.1191886  0.9816674 2.018406 15154
[6]  {Satisfaction.Airline.Status=Blue,
      Satisfaction.Type.of.Travel=Personal Travel,
      Satisfaction.Price.Sensitivity.Group=notsensitive,
      Satisfaction.No.of.Flights.p.a.Group=High}        => {Satisfaction.Satisfaction.Group=unsatisfied} 0.1417380  0.9818033 2.018685 18021
[7]  {Satisfaction.Airline.Status=Blue,
      Satisfaction.Type.of.Travel=Personal Travel,
      Satisfaction.Class=Eco,
      Satisfaction.No..of.other.Loyalty.Cards.Group=Low} => {Satisfaction.Satisfaction.Group=unsatisfied} 0.1231605  0.9735762 2.001769 15659
[8]  {Satisfaction.Airline.Status=Blue,
      Satisfaction.Type.of.Travel=Personal Travel,
      Satisfaction.Price.Sensitivity.Group=notsensitive,
      Satisfaction.No..of.other.Loyalty.Cards.Group=Low} => {Satisfaction.Satisfaction.Group=unsatisfied} 0.1509560  0.9732759 2.001152 19193
[9]  {Satisfaction.Airline.Status=Blue,
      Satisfaction.Gender=Female,
      Satisfaction.Type.of.Travel=Personal Travel,
      Satisfaction.Class=Eco}                           => {Satisfaction.Satisfaction.Group=unsatisfied} 0.1212650  0.9736044 2.001827 15418
[10] {Satisfaction.Airline.Status=Blue,
      Satisfaction.Gender=Female,
      Satisfaction.Type.of.Travel=Personal Travel,
      Satisfaction.Price.Sensitivity.Group=notsensitive} => {Satisfaction.Satisfaction.Group=unsatisfied} 0.1524976  0.9728062 2.000186 19389
[11] {Satisfaction.Airline.Status=Blue,
      Satisfaction.Type.of.Travel=Personal Travel,
      Satisfaction.Price.Sensitivity.Group=notsensitive,
      Satisfaction.No.of.Flights.p.a.Group=High,
      Satisfaction.No..of.other.Loyalty.Cards.Group=Low} => {Satisfaction.Satisfaction.Group=unsatisfied} 0.1076190  0.9870158 2.029402 13683
[12] {Satisfaction.Airline.Status=Blue,
      Satisfaction.Type.of.Travel=Personal Travel,
      Satisfaction.Class=Eco,
      Satisfaction.Price.Sensitivity.Group=notsensitive,
      Satisfaction.No.of.Flights.p.a.Group=High}        => {Satisfaction.Satisfaction.Group=unsatisfied} 0.1191886  0.9816674 2.018406 15154
[13] {Satisfaction.Airline.Status=Blue,
      Satisfaction.Type.of.Travel=Personal Travel,
      Satisfaction.Class=Eco,
      Satisfaction.Price.Sensitivity.Group=notsensitive,
      Satisfaction.No..of.other.Loyalty.Cards.Group=Low} => {Satisfaction.Satisfaction.Group=unsatisfied} 0.1227751  0.9734955 2.001603 15610
[14] {Satisfaction.Airline.Status=Blue,
      Satisfaction.Gender=Female,
      Satisfaction.Type.of.Travel=Personal Travel,
      Satisfaction.Class=Eco,
      Satisfaction.Price.Sensitivity.Group=notsensitive} => {Satisfaction.Satisfaction.Group=unsatisfied} 0.1209347  0.9735343 2.001683 15376
```

Since we want to focus our attention on responses that occur with some meaningful frequency in the survey data set. Considering the size of the data set as well as the potential application of the rules, we set the minimum support as 0.25 and increase the level of confidence to check other rules. Below are the rules with support equal to or greater than 0.25 and confidence equal to or greater than 0.4.

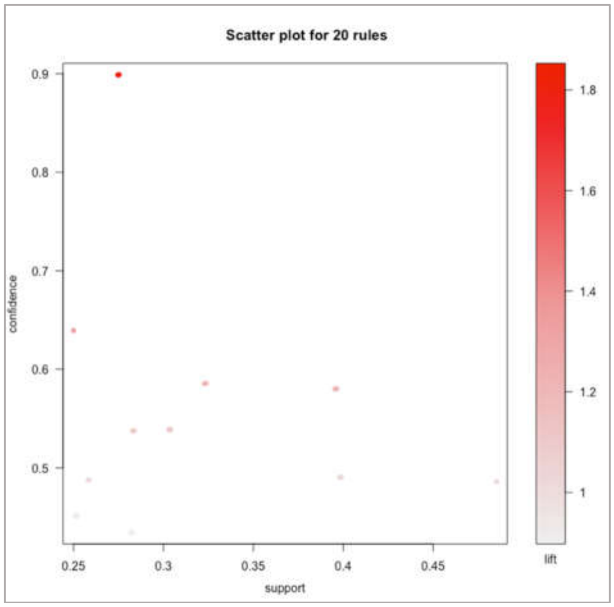
```
     lhs                                                    rhs                                           support confidence      lift count
[1]  {Satisfaction.Type.of.Travel=Personal Travel}       => {Satisfaction.Satisfaction.Group=unsatisfied} 0.2753435  0.8988626 1.8481505 35008
[2]  {Satisfaction.No..of.other.Loyalty.Cards.Group=Low} => {Satisfaction.Satisfaction.Group=unsatisfied} 0.2836648  0.5376565 1.1054751 36066
[3]  {Satisfaction.Year.of.First.Flight.Group=2003-2007} => {Satisfaction.Satisfaction.Group=unsatisfied} 0.2586379  0.4877340 1.0028294 32884
[4]  {Satisfaction.Departure.Delay.in.Minutes.Group=Low} => {Satisfaction.Satisfaction.Group=unsatisfied} 0.2517244  0.4516334 0.9286030 32005
[5]  {Satisfaction.Gender=Female}                        => {Satisfaction.Satisfaction.Group=unsatisfied} 0.3038547  0.5389423 1.1081188 38633
[6]  {Satisfaction.Arrival.Delay.greater.5.Mins=no}      => {Satisfaction.Satisfaction.Group=unsatisfied} 0.2824536  0.4345543 0.8934867 35912
[7]  {Satisfaction.Airline.Status=Blue}                  => {Satisfaction.Satisfaction.Group=unsatisfied} 0.3963097  0.5801861 1.1929201 50388
[8]  {Satisfaction.Class=Eco}                            => {Satisfaction.Satisfaction.Group=unsatisfied} 0.3989130  0.4906312 1.0087864 50719
[9]  {Satisfaction.Price.Sensitivity.Group=notsensitive} => {Satisfaction.Satisfaction.Group=unsatisfied} 0.4853354  0.4860694 0.9994069 61707
[10] {Satisfaction.Type.of.Travel=Personal Travel,
      Satisfaction.Price.Sensitivity.Group=notsensitive} => {Satisfaction.Satisfaction.Group=unsatisfied} 0.2746592  0.8987055 1.8478276 34921
[11] {Satisfaction.Price.Sensitivity.Group=notsensitive,
      Satisfaction.No..of.other.Loyalty.Cards.Group=Low} => {Satisfaction.Satisfaction.Group=unsatisfied} 0.2829806  0.5374010 1.1049498 35979
[12] {Satisfaction.Price.Sensitivity.Group=notsensitive,
      Satisfaction.Year.of.First.Flight.Group=2003-2007} => {Satisfaction.Satisfaction.Group=unsatisfied} 0.2581660  0.4875093 1.0023674 32824
[13] {Satisfaction.Price.Sensitivity.Group=notsensitive,
      Satisfaction.Departure.Delay.in.Minutes.Group=Low} => {Satisfaction.Satisfaction.Group=unsatisfied} 0.2511660  0.4512938 0.9279048 31934
[14] {Satisfaction.Airline.Status=Blue,
      Satisfaction.Gender=Female}                        => {Satisfaction.Satisfaction.Group=unsatisfied} 0.2500177  0.6393532 1.3145735 31788
[15] {Satisfaction.Gender=Female,
      Satisfaction.Price.Sensitivity.Group=notsensitive} => {Satisfaction.Satisfaction.Group=unsatisfied} 0.3032019  0.5386185 1.1074529 38550
[16] {Satisfaction.Arrival.Delay.greater.5.Mins=no,
      Satisfaction.Price.Sensitivity.Group=notsensitive} => {Satisfaction.Satisfaction.Group=unsatisfied} 0.2818559  0.4342810 0.8929248 35836
[17] {Satisfaction.Airline.Status=Blue,
      Satisfaction.Class=Eco}                            => {Satisfaction.Satisfaction.Group=unsatisfied} 0.3236592  0.5856376 1.2041290 41151
[18] {Satisfaction.Airline.Status=Blue,
      Satisfaction.Price.Sensitivity.Group=notsensitive} => {Satisfaction.Satisfaction.Group=unsatisfied} 0.3953737  0.5799043 1.1923406 50269
[19] {Satisfaction.Class=Eco,
      Satisfaction.Price.Sensitivity.Group=notsensitive} => {Satisfaction.Satisfaction.Group=unsatisfied} 0.3980007  0.4903154 1.0081370 50603
[20] {Satisfaction.Airline.Status=Blue,
      Satisfaction.Class=Eco,
      Satisfaction.Price.Sensitivity.Group=notsensitive} => {Satisfaction.Satisfaction.Group=unsatisfied} 0.3228176  0.5853144 1.2034644 41044
```

As we can see in the result, both support and confidence seem low, but even a rule with low support and smallish confidence might help airline companies to find out how customers really feel about their services and identify areas of improvement.

The support refers to the frequency of cooccurrence of LHS and RHS together. In this case, the frequency of LHS and RHS occur together are greater than 25% in the survey. For example, the first rule has a support of 0.2753, it means that the frequency of “a customer’s type of travel is person” and “a customer is not satisfied” happening together is 27.53%

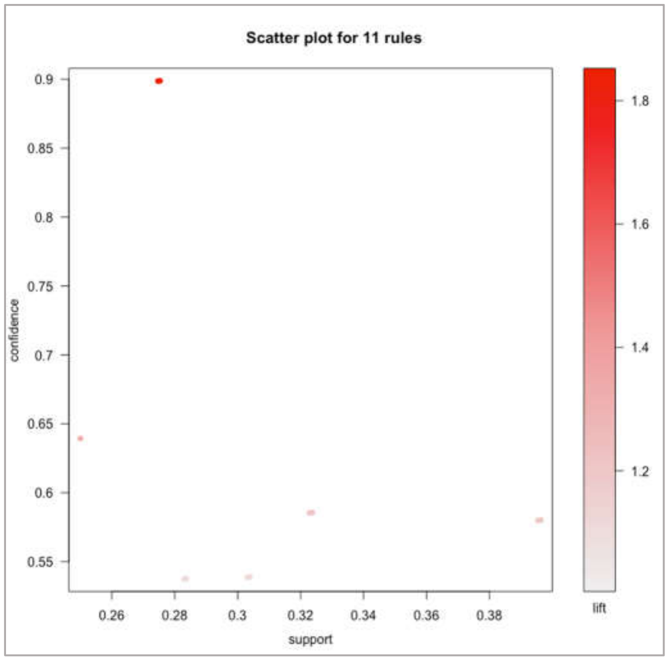
The confidence of a rule refers to the proportion of the time that LHS and RHS occur together versus the total number of appearances of LHS. For example, the fifth rule has a confidence of 0.8988, it means that the proportion of “a customer is female and unsatisfied” versus the total number of female customers is 89.88%. In other words, it indicates that given that the customer is female, the probability of this customer is unsatisfied is 89.88%.

Below is a scatter plot for these 20 rules.



Below are the rules whose lift is higher than 1.1. This high lift means although LHS and RHS are not abundant, they always happen together.

```
     lhs                                                  rhs                                         support confidence     lift count
[1]  {Satisfaction.Type.of.Travel=Personal Travel}        => {Satisfaction.Satisfaction.Group=unsatisfied} 0.2753435  0.8988626 1.848150 35008
[2]  {Satisfaction.No..of.other.Loyalty.Cards.Group=Low} => {Satisfaction.Satisfaction.Group=unsatisfied} 0.2836648  0.5376565 1.105475 36066
[3]  {Satisfaction.Gender=Female}                         => {Satisfaction.Satisfaction.Group=unsatisfied} 0.3038547  0.5389423 1.108119 38633
[4]  {Satisfaction.Airline.Status=Blue}                   => {Satisfaction.Satisfaction.Group=unsatisfied} 0.3963097  0.5801861 1.192920 50388
[5]  {Satisfaction.Type.of.Travel=Personal Travel,
      Satisfaction.Price.Sensitivity.Group=notsensitive} => {Satisfaction.Satisfaction.Group=unsatisfied} 0.2746592  0.8987055 1.847828 34921
[6]  {Satisfaction.Price.Sensitivity.Group=notsensitive,
      Satisfaction.No..of.other.Loyalty.Cards.Group=Low} => {Satisfaction.Satisfaction.Group=unsatisfied} 0.2829806  0.5374010 1.104950 35979
[7]  {Satisfaction.Airline.Status=Blue,
      Satisfaction.Gender=Female}                         => {Satisfaction.Satisfaction.Group=unsatisfied} 0.2500177  0.6393532 1.314573 31788
[8]  {Satisfaction.Gender=Female,
      Satisfaction.Price.Sensitivity.Group=notsensitive} => {Satisfaction.Satisfaction.Group=unsatisfied} 0.3032019  0.5386185 1.107453 38550
[9]  {Satisfaction.Airline.Status=Blue,
      Satisfaction.Class=Eco}                             => {Satisfaction.Satisfaction.Group=unsatisfied} 0.3236592  0.5856376 1.204129 41151
[10] {Satisfaction.Airline.Status=Blue,
      Satisfaction.Price.Sensitivity.Group=notsensitive} => {Satisfaction.Satisfaction.Group=unsatisfied} 0.3953737  0.5799043 1.192341 50269
[11] {Satisfaction.Airline.Status=Blue,
      Satisfaction.Class=Eco,
      Satisfaction.Price.Sensitivity.Group=notsensitive} => {Satisfaction.Satisfaction.Group=unsatisfied} 0.3228176  0.5853144 1.203464 41044
```



From these results, we can conclude that:

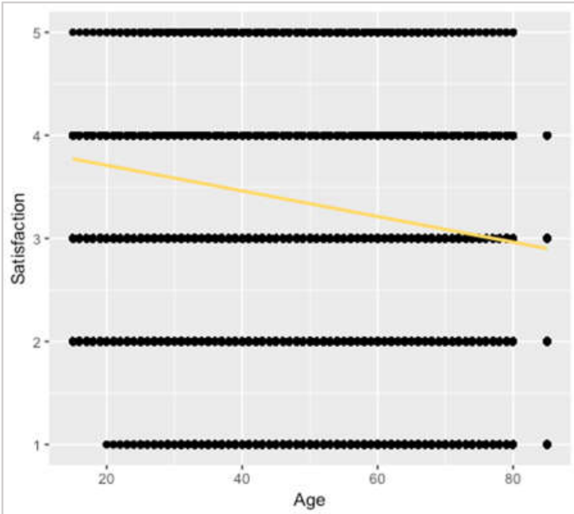
1) Customers whose travel type is personal travel tend to be unsatisfied.
2) Customers who have small amounts of other airlines' loyalty cards tend to be unsatisfied.
3) Female customers tend to be unsatisfied.
4) Customers whose airline status is blue tend to be unsatisfied.
5) Customers whose travel type is personal travel and are not sensitive to price tend to be unsatisfied.
6) Customers who have small amounts of other airlines' loyalty cards and are not sensitive to price tend to be unsatisfied.
7) Female customers whose airline status is blue tend to be unsatisfied.
8) Female customers who are not sensitive to price tend to be unsatisfied.
9) Customers whose airline status is blue, and class is economy tend to be unsatisfied.
10) Customers whose airline status is blue and who are not sensitive to price tend to be unsatisfied.
11) Customers whose airline status is blue, class is economy and who are not sensitive to price tend to be unsatisfied.

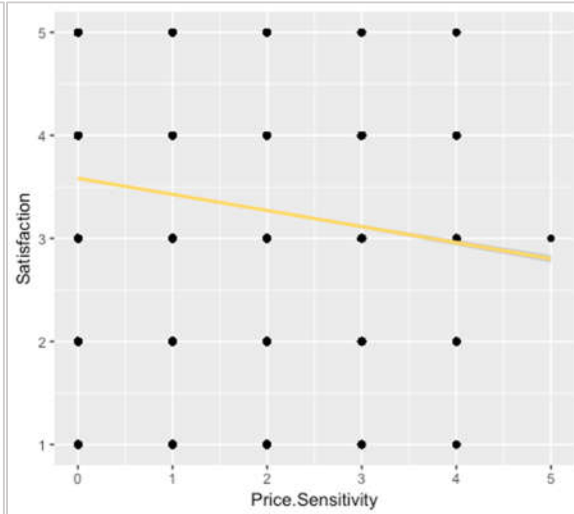
Linear Modeling

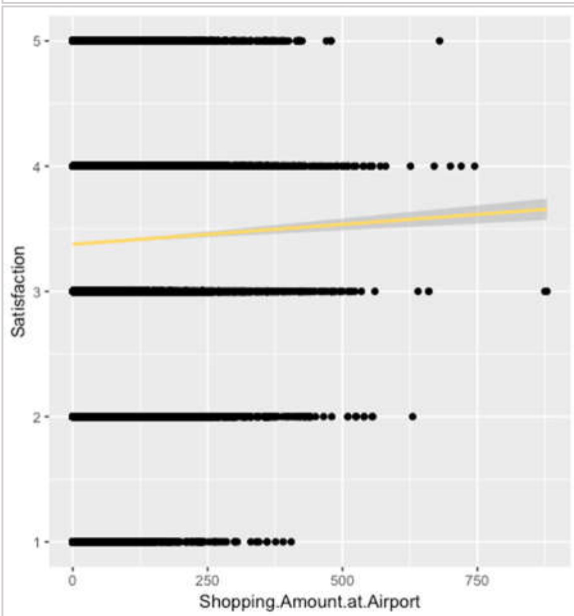
We use both simple and multiple linear modeling to understand the relationships between Satisfaction and all the other attributes in terms of three dimensions: Customers characteristic, Flight experience characteristic and Flight characteristic. We also visualize some results. The full model is developed but the ideal model with the highest adjust R square is developed after we conduct the association rules mining. The interpretation is listed in the end of this part.

Satisfaction VS Customers characteristic

Simple linear modeling







In the first model, the “age” is negatively correlated with the “satisfaction”. The elder the customers are, the lower the satisfaction level they have.

In the second model, the “price sensitivity” is negatively correlated with the “satisfaction”. The higher the price sensitivity the customers have, the lower the satisfaction level will be.

In the third model, the “shopping amount at airport” is positively correlated with the “satisfaction”. The more goods the customers buy in the airport, the higher the satisfaction level will be.

Multiple linear modeling

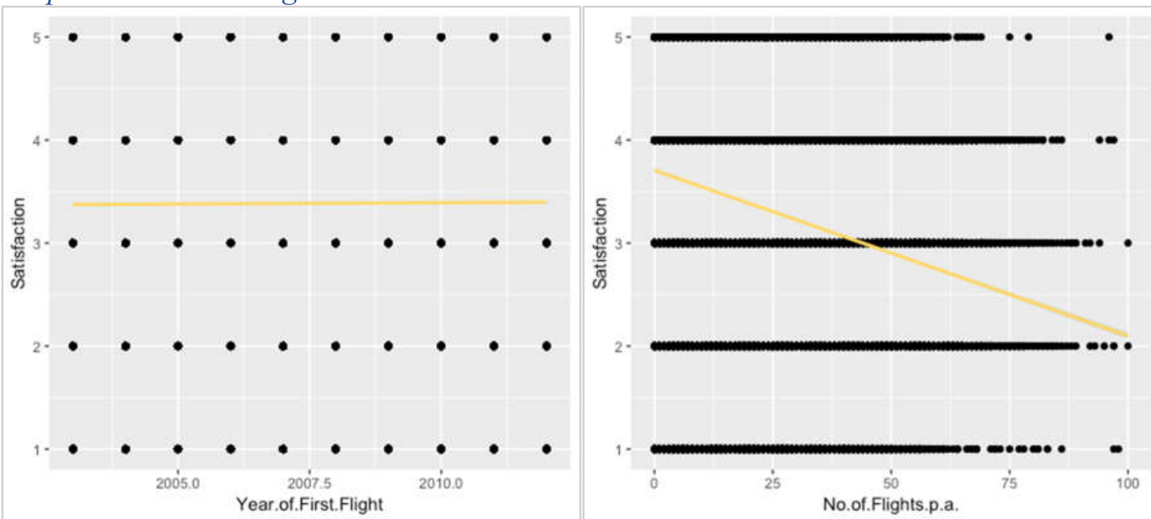
```
Call:
lm(formula = Satisfaction ~ Age + Gender + Price.Sensitivity +
    Shopping.Amount.at.Airport + Eating.and.Drinking.at.Airport,
    data = Satisfaction)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.0160 -0.6495  0.1991  0.6296  2.3849 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     4.052e+00  1.079e-02 375.519  < 2e-16 ***
Age                            -1.262e-02  1.526e-04 -82.675  < 2e-16 ***
GenderMale                      2.422e-01  5.294e-03  45.757  < 2e-16 ***
Price.Sensitivity              -1.762e-01  4.792e-03 -36.776  < 2e-16 ***
Shopping.Amount.at.Airport      4.630e-04  4.941e-05   9.371  < 2e-16 ***
Eating.and.Drinking.at.Airport  3.162e-04  5.051e-05   6.259 3.88e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9291 on 127137 degrees of freedom
Multiple R-squared:  0.07596,	Adjusted R-squared:  0.07592 
F-statistic:  2090 on 5 and 127137 DF,  p-value: < 2.2e-16
```

The model above consists of “age”, “gender”, “price sensitivity”, “shopping amount at airport” and “eating and drinking at airport”. Holding other variables constant, a year elder the age is, the satisfaction level will fall 0.01262; while males tend to have 0.2422 satisfaction level higher than female. Holding other variables constant, a level higher the grade to which price affects customers purchasing will bring the level of satisfaction 0.1762 lower. In term of our result, how many goods or foods consumed by each customer at the airport have slight change on the level of satisfaction.

Model above has a small R-squared, and only around 7% of the data was explained within the model. Even though variables in this model are having low P-values, the explanatory power of this model is weak.

Satisfaction VS Flight experience characteristic

Simple linear modeling



In the first model, the “Year of first flight” is not correlated with the “satisfaction”. It means that the satisfaction level is not changing by the year of the first flight the customers have taken in.

In the second model, the “no. of flights p.a.” is negatively correlated with the “satisfaction”. The more flights customers have taken, the lower satisfaction level it is.

Multiple linear modeling

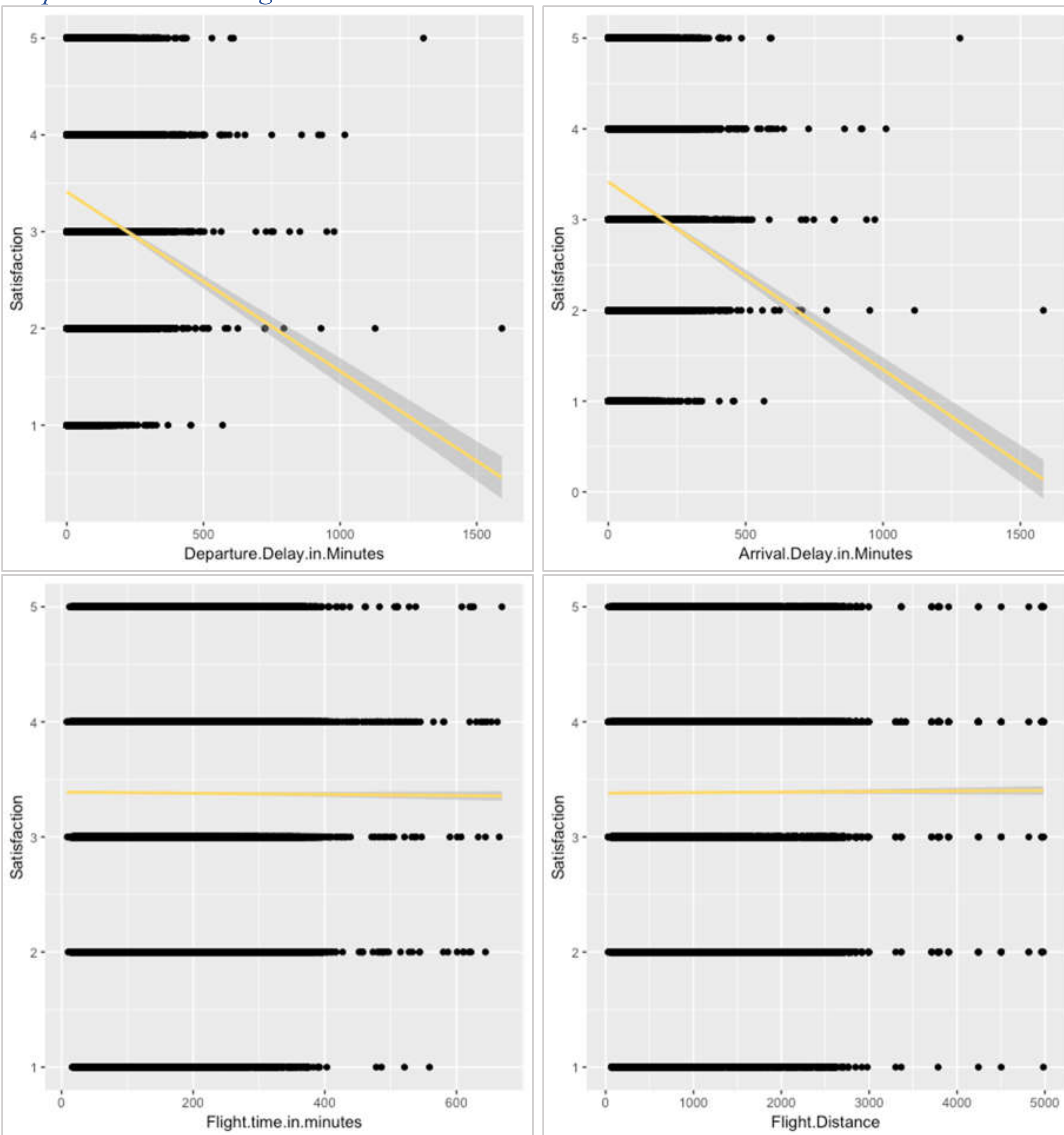
```
Call:
lm(formula = Satisfaction ~ Year.of.First.Flight + No.of.Flights.p.a. +
    X..of.Flight.with.other.Airlines + Type.of.Travel + No..of.other.Loyalty.Cards +
    Class + Airline.Status, data = Satisfaction)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9897 -0.4522 -0.0134  0.4918  2.6973

Coefficients:
                                  Estimate Std. Error  t value Pr(>|t|)
(Intercept)                     -4.9746228  1.4021415   -3.548 0.000388 ***
Year.of.First.Flight             0.0043210  0.0006986    6.185 6.23e-10 ***
No.of.Flights.p.a.              -0.0030824  0.0001569  -19.648  < 2e-16 ***
X..of.Flight.with.other.Airlines -0.0004954  0.0002661   -1.862 0.062617 .
Type.of.TravelMileage tickets   -0.1507234  0.0080002  -18.840  < 2e-16 ***
Type.of.TravelPersonal Travel   -1.1254894  0.0048569 -231.731  < 2e-16 ***
No..of.other.Loyalty.Cards       0.0121052  0.0020546    5.892 3.83e-09 ***
ClassEco                        -0.0771544  0.0076071  -10.142  < 2e-16 ***
ClassEco Plus                   -0.0962946  0.0097202   -9.907  < 2e-16 ***
Airline.StatusGold               0.4332488  0.0076685   56.497  < 2e-16 ***
Airline.StatusPlatinum           0.2791653  0.0119123   23.435  < 2e-16 ***
Airline.StatusSilver             0.6236381  0.0053571  116.413  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7406 on 127131 degrees of freedom
Multiple R-squared:  0.413,     Adjusted R-squared:  0.4129
F-statistic:  8130 on 11 and 127131 DF,  p-value: < 2.2e-16
```

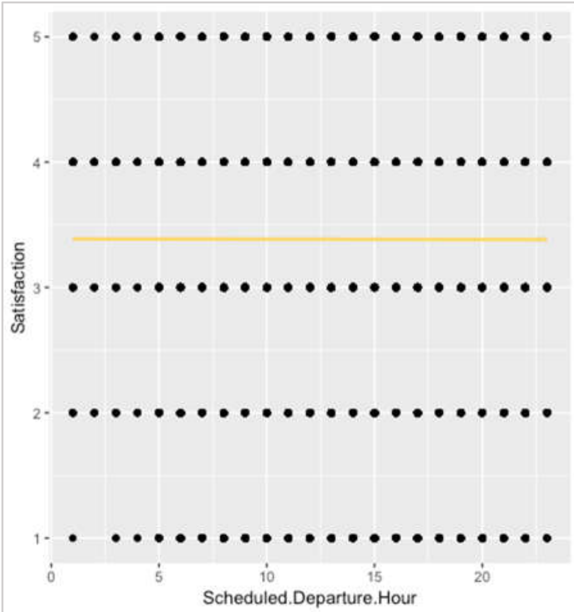
The model consists of "the year of customers' first flight", "the number of flights each customer has taken", "type of travel", "number of loyalty cards", "class" and "the airline status". Having a low P-value, "type of travel" is significantly correlated to customers' satisfaction level by dividing into three groups: "mileage tickets", "business travel" and "personal travel". Comparing to people taking business travel, people with mileage will have 0.1507 lower satisfaction level, while people having personal travel will have 1.1255 lower satisfaction level. In this model, the "airline status", whose type are platinum, gold, silver and blue, has significant influence on satisfaction. Other factors in this model affect the satisfaction level slightly.

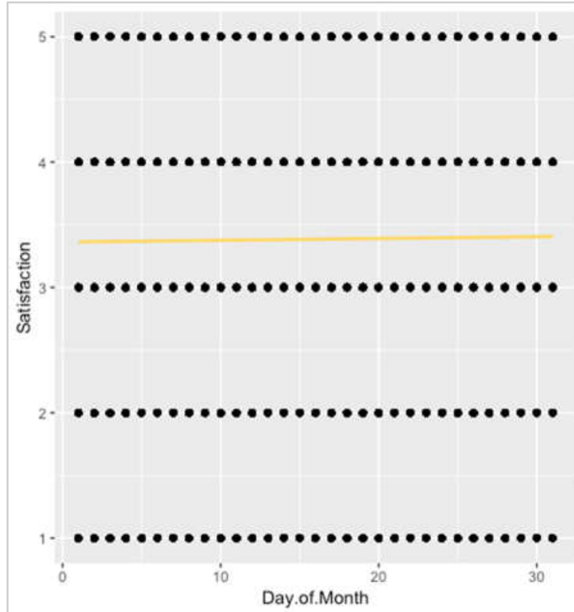
The R-squared shows that this model can explain 41.3% of the data, with most of the variables having low P-values, this model is relatively good in explaining the factors affect customers' satisfaction level.

Satisfaction VS Flight characteristic

Simple linear modeling







In the first model, the “departure delay in minutes” is negatively correlated with the “satisfaction”. In the second model, being the same as departure, the “arrival delay in minutes” is negatively correlated with the “satisfaction”. In other words, the delay will contribute to the low satisfaction level of customers.
In the third model, the “flight time in minutes” is not correlated with the “satisfaction”.
In the fourth model, the “flight distance” is not correlated with the “satisfaction”.
In the fifth model, the “Scheduled departure hour” is not correlated with the “satisfaction”.
In the last model, the “day of month” at airport is not correlated with the “satisfaction”.
Therefore, we found that “flight time in minutes”, “flight distance”, “Scheduled departure hour” and “day of month” are less necessary to affect the satisfaction level.

Multiple linear modeling

```
Call:
lm(formula = Satisfaction ~ Day.of.Month + Scheduled.Departure.Hour +
    Departure.Delay.in.Minutes + Arrival.Delay.in.Minutes + Flight.time.in.minutes +
    Flight.Distance, data = Satisfaction)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5419 -0.4295  0.5403  0.5972  3.9989

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 3.413e+00  1.072e-02 318.269  < 2e-16 ***
Day.of.Month                6.985e-04  3.137e-04   2.227    0.026 *
Scheduled.Departure.Hour    5.652e-04  5.898e-04   0.958    0.338
Departure.Delay.in.Minutes  1.932e-03  2.840e-04   6.803 1.03e-11 ***
Arrival.Delay.in.Minutes   -3.865e-03  2.806e-04 -13.773  < 2e-16 ***
Flight.time.in.minutes     -1.125e-03  1.846e-04  -6.092 1.12e-09 ***
Flight.Distance             1.371e-04  2.231e-05   6.146 7.96e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9628 on 127136 degrees of freedom
Multiple R-squared:  0.007882,  Adjusted R-squared:  0.007835
F-statistic: 168.3 on 6 and 127136 DF,  p-value: < 2.2e-16
```

Satisfaction VS all the other attributes

Model below is developed by all the attributes that we assumed are related to satisfaction and have a P-value relatively significant.

```
Call:
lm(formula = Satisfaction ~ Age.Group + Gender + Price.Sensitivity +
    Year.of.First.Flight + No.of.Flights.p.a. + Type.of.Travel +
    No..of.other.Loyalty.Cards + Shopping.Amount.at.Airport +
    Eating.and.Drinking.at.Airport + Class + Scheduled.Departure.Hour.Group +
    Arrival.Delay.greater.5.Mins, data = Satisfaction)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2516 -0.4674  0.0732  0.4248  2.9923

Coefficients:
                                                     Estimate Std. Error  t value Pr(>|t|)
(Intercept)                                        -1.191e+01  1.422e+00   -8.380  < 2e-16 ***
Age.Group18-24                                      1.115e-01  1.348e-02    8.272  < 2e-16 ***
Age.Group25-34                                      3.561e-01  1.188e-02   29.977  < 2e-16 ***
Age.Group35-44                                      4.404e-01  1.161e-02   37.945  < 2e-16 ***
Age.Group45-54                                      3.844e-01  1.197e-02   32.112  < 2e-16 ***
Age.Group55-64                                      2.368e-01  1.252e-02   18.914  < 2e-16 ***
Age.Group65+                                        3.021e-02  1.258e-02    2.401   0.0163 *
GenderMale                                          1.372e-01  4.385e-03   31.293  < 2e-16 ***
Price.Sensitivity                                  -5.417e-02  3.921e-03  -13.817  < 2e-16 ***
Year.of.First.Flight                                7.791e-03  7.084e-04   10.998  < 2e-16 ***
No.of.Flights.p.a.                                 -4.263e-03  1.612e-04  -26.448  < 2e-16 ***
Type.of.TravelMileage tickets                      -1.456e-01  8.166e-03  -17.825  < 2e-16 ***
Type.of.TravelPersonal Travel                      -1.032e+00  5.491e-03 -187.971  < 2e-16 ***
No..of.other.Loyalty.Cards                         -2.468e-02  2.196e-03  -11.234  < 2e-16 ***
Shopping.Amount.at.Airport                          1.972e-04  4.001e-05    4.928 8.33e-07 ***
Eating.and.Drinking.at.Airport                      3.469e-04  4.121e-05    8.417  < 2e-16 ***
ClassEco                                           -8.110e-02  7.707e-03  -10.523  < 2e-16 ***
ClassEco Plus                                      -8.227e-02  9.904e-03   -8.307  < 2e-16 ***
Scheduled.Departure.Hour.Groupearly morning (1am-5am) 9.829e-03  1.839e-02    0.535   0.5930
Scheduled.Departure.Hour.Groupevening (6pm-11pm)    4.354e-03  5.778e-03    0.753   0.4512
Scheduled.Departure.Hour.Groupmorning (6am-11am)   -2.488e-02  4.778e-03   -5.206 1.93e-07 ***
Arrival.Delay.greater.5.Minsyes                    -3.420e-01  4.444e-03  -76.965  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7503 on 127121 degrees of freedom
Multiple R-squared:  0.3975,    Adjusted R-squared:  0.3974
F-statistic:  3994 on 21 and 127121 DF,  p-value: < 2.2e-16
```

Below is the full linear model based on the result of associating rules mining, the first one is developed by using numeric attributes while the second one is by using factor attributes transformed from numeric.

```
Call:
lm(formula = Satisfaction ~ Airline.Status + Gender + Price.Sensitivity +
    Year.of.First.Flight + Type.of.Travel + No..of.other.Loyalty.Cards +
    Class + Departure.Delay.in.Minutes + Arrival.Delay.greater.5.Mins,
    data = Satisfaction)

Residuals:
     Min       1Q   Median       3Q      Max
-3.15016 -0.41374  0.08469  0.47387  2.90318

Coefficients:
                                 Estimate Std. Error  t value Pr(>|t|)
(Intercept)                    -4.590e+00  1.366e+00   -3.360 0.000779 ***
Airline.StatusGold              4.457e-01  7.447e-03   59.846  < 2e-16 ***
Airline.StatusPlatinum          2.664e-01  1.161e-02   22.942  < 2e-16 ***
Airline.StatusSilver            6.282e-01  5.190e-03  121.021  < 2e-16 ***
GenderMale                      1.223e-01  4.167e-03   29.349  < 2e-16 ***
Price.Sensitivity              -3.103e-02  3.740e-03   -8.296  < 2e-16 ***
Year.of.First.Flight            4.144e-03  6.807e-04    6.088 1.15e-09 ***
Type.of.TravelMileage tickets  -1.613e-01  7.767e-03  -20.763  < 2e-16 ***
Type.of.TravelPersonal Travel  -1.132e+00  4.608e-03 -245.725  < 2e-16 ***
No..of.other.Loyalty.Cards      1.956e-02  1.790e-03   10.926  < 2e-16 ***
ClassEco                       -7.816e-02  7.412e-03  -10.546  < 2e-16 ***
ClassEco Plus                  -5.790e-02  9.501e-03   -6.095 1.10e-09 ***
Departure.Delay.in.Minutes      1.153e-04  6.014e-05    1.916 0.055308 .
Arrival.Delay.greater.5.Minsyes -3.408e-01  4.820e-03  -70.693  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7215 on 127129 degrees of freedom
Multiple R-squared:  0.4428,    Adjusted R-squared:  0.4427
F-statistic:  7771 on 13 and 127129 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Satisfaction ~ Airline.Status + Gender + Price.Sensitivity.Group +
    Year.of.First.Flight.Group + Type.of.Travel + No..of.other.Loyalty.Cards.Group +
    Class + Departure.Delay.in.Minutes.Group + Arrival.Delay.greater.5.Mins,
    data = Satisfaction)

Residuals:
     Min       1Q   Median       3Q      Max
-3.15333 -0.42233  0.09167  0.46141  2.89314

Coefficients:
                                     Estimate Std. Error  t value Pr(>|t|)
(Intercept)                          3.738423   0.012416  301.094  < 2e-16 ***
Airline.StatusGold                   0.450097   0.007434   60.543  < 2e-16 ***
Airline.StatusPlatinum               0.271404   0.011598   23.401  < 2e-16 ***
Airline.StatusSilver                 0.631159   0.005181  121.827  < 2e-16 ***
GenderMale                           0.124962   0.004168   29.983  < 2e-16 ***
Price.Sensitivity.Groupsensitive    -0.086215   0.052109   -1.654    0.098 .
Year.of.First.Flight.Group2008-2012  0.019742   0.004058    4.864 1.15e-06 ***
Type.of.TravelMileage tickets       -0.164166   0.007754  -21.172  < 2e-16 ***
Type.of.TravelPersonal Travel       -1.127172   0.004629 -243.520  < 2e-16 ***
No..of.other.Loyalty.Cards.GroupHigh -0.006201  0.005982   -1.037    0.300
No..of.other.Loyalty.Cards.GroupLow -0.074867   0.005424  -13.804  < 2e-16 ***
ClassEco                            -0.078535   0.007409  -10.600  < 2e-16 ***
ClassEco Plus                       -0.053675   0.009494   -5.654 1.57e-08 ***
Departure.Delay.in.Minutes.GroupHigh  0.004997  0.009900    0.505    0.614
Departure.Delay.in.Minutes.GroupLow -0.007226   0.009350   -0.773    0.440
Arrival.Delay.greater.5.Minsyes     -0.343762   0.005374  -63.973  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7213 on 127127 degrees of freedom
Multiple R-squared:  0.4432,    Adjusted R-squared:  0.4432
F-statistic:  6747 on 15 and 127127 DF,  p-value: < 2.2e-16
```

As we can see, the second model has the highest adjusted R-squared among all the models we have developed. It can explain 44.32% of the data, having most of the variables significant. Since the model is developed by associating rules mining, it is eligible in explaining the factors that affect satisfaction level.

Holding other variables constant, males will have 0.1249 satisfaction level higher than females. As for "the airline status", comparing to the satisfaction level of the Blue status customers, the Silver customers' satisfaction level will be 0.6312 higher while Gold customers' will be 0.4501 higher. When it comes to Platinum status, the satisfaction level only

improves 0.2714. As we can see, the higher the status is, the slower the satisfaction level increases. It can be explained by the marginal utility that the additional satisfaction a consumer gains (from consuming one more unit of better service brings by higher status) is getting less.

Holding all the variables as the same, compared to people traveling for business, customers who travel by the mileage tickets that based on loyalty card will have 0.1642 lower satisfaction than others; customers who travel for personal reasons like seeing the family or being in vacation will have 1.1272 lower. It might be caused by customers' expectation of having good experience in the flight. For people who take business trip, they might have a lower expectation by regarding the flight as a task but not a journey.

Support Vector Machine

Based on the result of association rules, we used "Southeast" subset which contains 9,423 observations to predict the unsatisfied customers.

By using the table function we find that there are 4,916 satisfied and 4,507 unsatisfied customers.

To make the analysis, we use two thirds of the data set to train and the remainder to test. So, we have 6,282 observations in train data set and 3,141 in the test data set.

Below is the command we use to train our support vector model based on the training data set:

```
> SatisvmOutput6 <- ksvm(Satisfaction.Group ~ Airline.Status+Gender+
+                         Price.Sensitivity.Group+Year.of.First.Flight.Group+
+                         Type.of.Travel+No..of.other.Loyalty.Cards.Group+
+                         Class+Departure.Delay.in.Minutes.Group+
+                         Arrival.Delay.greater.5.Mins,
+                       data=SatiTrainData, kernel= "rbfdot", kpar = "automatic",
+                       C = 5, cross = 3, prob.model = TRUE)
```

We have the "Satisfaction.Group" variable as the outcome variable that our model predicts. And we use "Price.Sensitivity.Group","Year.of.First.Flight.Group" ,"Type.of.Travel","No..of.other.Loyalty.Cards.Group","Class", "Departure.Delay.in.Minutes.Group" and "Arrival.Delay.greater.5.Mins" as variables to try to predict customers' satisfaction.

We set the parameter C as 5 means we allow the model to make some classification mistakes to get a generalizable model. We also use threefold cross-validation to avoid overfitting.

Below is the output:

```
> SatisvmOutput6
Support Vector Machine object of class "ksvm"

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

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.229166666666667

Number of Support Vectors : 2960

Objective Function Value : -13000.55
Training error : 0.200255
Cross validation error : 0.210761
Probability model included.
```

As we can see, the training error at about 20% is acceptable. And a 21 % cross-validation error rate is not bad for predicting customers' overall impression of the airline's services.

Then we use the support vector model we just generated to predict the outcomes in the test data set. Then we compare the result of our prediction with the ground truth, "Satisfied.Group" variable.

Below is the result:

```
                                 SatiSvmPrediction6.2...
SatiTestData.Satisfaction.Group     0    1
                      satisfied  1480  136
                    unsatisfied   517 1008
```

As we can see, the left-hand list has one (1) for an unsatisfied vote and zero (0) for a satisfied vote. 517 cases that were unsatisfied but were classified as satisfied, and 136 cases that were satisfied but were classified as unsatisfied by the support vector matrix.

Then we calculate the accuracy of the prediction, summing error cases (517+136=653) and dividing by the total cases (3,141) for a total error rate of about 20.8%. Although the error rate looks high, it is not bad because in the real word, individuals are very heterogonous in their attitudes and behaviors.

We also use the entire "southeast" subset to model the support vector and use subsets of other airlines (West Airways Inc. and Enjoy Flying Air Services) to predict in order to see our model's prediction power. Below are the results and our model's error rate is around 20%. It is a relatively low error rate to predict human attitudes and behaviors.

```
> # Build a model with Southeast airlines
> SatisvmOutput7 <- ksvm(Satisfaction.Group ~ Airline.Status+Gender+
+                            Price.Sensitivity.Group+Year.of.First.Flight.Group+
+                            Type.of.Travel+No..of.other.Loyalty.Cards.Group+
+                            Class+Departure.Delay.in.Minutes.Group+
+                            Arrival.Delay.greater.5.Mins,
+                          data=SESubset, kernel= "rbfdot", kpar = "automatic",
+                          C = 5, cross = 3, prob.model = TRUE)
> SatisvmOutput7
Support Vector Machine object of class "ksvm"

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

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.229166666666667

Number of Support Vectors : 4350

Objective Function Value : -19505.96
Training error : 0.204181
Cross validation error : 0.208638
Probability model included.
> # Create a subset of WestAirwaysInc
> WASubset <- data.frame(filter(Satisfaction, Airline.Name=="WestAirwaysInc."))
> # Making a Prediction variable based on number of votes
> SatiSvmPrediction7 <- predict(SatisvmOutput7, WASubset, type = "votes")
> str(SatiSvmPrediction7)
 num [1:2, 1:1685] 1 0 1 0 1 0 1 0 1 0 ...
> head(SatiSvmPrediction7[2,])
[1] 0 0 0 0 0 0
> # Creating a composite table based on satisfied customers and SVM Prediction
> SatiCompTable7<-data.frame(WASubset$Satisfaction.Group,SatiSvmPrediction7[2,])
> # Creating a confusion matrix
> ConfusionMatrix7<-table(SatiCompTable7)
> ConfusionMatrix7
                           SatiSvmPrediction7.2...
WASubset.Satisfaction.Group   0   1
                satisfied   882  72
                unsatisfied 272 459
> # Creating a dataframe containing sum of errors
> SatiErrorSum7 <- ConfusionMatrix7[1,2]+ConfusionMatrix7[2,1]
> # Creating percentage of error rate
> SatiErrorRate7<-SatiErrorSum7/sum(ConfusionMatrix7)*100
> SatiErrorRate7
[1] 20.41543
```

```
> # use EnjoyFlyingAirServices to predict
> WFSubset <- data.frame(filter(Satisfaction, Airline.Name=="EnjoyFlyingAirServices"))
> # Making a Prediction variable based on number of votes
> SatiSvmPrediction8 <- predict(SatisvmOutput7, WFSubset, type = "votes")
> str(SatiSvmPrediction8)
 num [1:2, 1:8584] 1 0 1 0 1 0 1 0 1 0 ...
> head(SatiSvmPrediction8[2,])
[1] 0 0 0 0 0 0
>
> # Creating a composite table based on satisfied customers and SVM Prediction
> SatiCompTable8<-data.frame(WFSubset$Satisfaction.Group,SatiSvmPrediction8[2,])
> # Creating a confusion matrix
> ConfusionMatrix8<-table(SatiCompTable8)
> ConfusionMatrix8
                           SatiSvmPrediction8.2...
WFSubset.Satisfaction.Group    0    1
               satisfied    3883  445
               unsatisfied 1436 2820
> # Creating a dataframe containing sum of errors
> SatiErrorSum8 <- ConfusionMatrix8[1,2]+ConfusionMatrix8[2,1]
> # Creating percentage of error rate
> SatiErrorRate8<-SatiErrorSum8/sum(ConfusionMatrix8)*100
> SatiErrorRate8
[1] 21.91286
```

Besides, we use the Customers characteristic, Flight experience characteristic and Flight characteristic to build the model, but the error rate are around 31%, 39% and 45%, the error rate is too high to predict if a customer is satisfied or nor. So we keep it in our appendix, not in the report.

Actionable Insights / Overall interpretation of results

The linear modelling and support vector machine modelling indicate that the combination of “Airline.Status”, “Gender”, “Price.Sensitivity”, “Year.of.First.Flight.Group”, “Type of Travel”,”No..of.other.Loyalty.Cards.Group”, ”Class” ,”Departure.Delay.in.Minutes.Group” and “Arrival.Delay.greater.5.Mins” factors provides highest adjusted R square and accuracy. The adjusted R square is 0.4432, which means the 44.32% data can be explained by these factors. And the error rate of this model to predict of new data 21%. These numbers validate our model’s power to predict the satisfaction of customers.

From the descriptive analysis, modeling analysis and other quantitative research we have done, we found that the satisfaction level are mainly correlative with the experience that the customers have in their flights. Some specific personal features also contributed to the satisfaction level, like the “airline status” or “gender”. To help our client, Southeast Airline Company, get bigger share in the competitive airline market, we give out the suggestions as following:

1. For the customers whose status is blue, they are more likely to have lower satisfaction level. Southeast company can offer them multiple way to upgrade their status, for example, offering double credit points in typical holidays, like Christmas or Thanksgiving Day. Having higher airline status can help them earn better service so that the customers loyalty will be improved with higher satisfaction levels.
2. Customers who have their traveling for personal purpose, Southeast company can corporate with hotels and tourist spots and offer customers discounted accommodations

or tickets. In this way, customers traveling for personal affairs might prefer our company and have higher satisfaction.

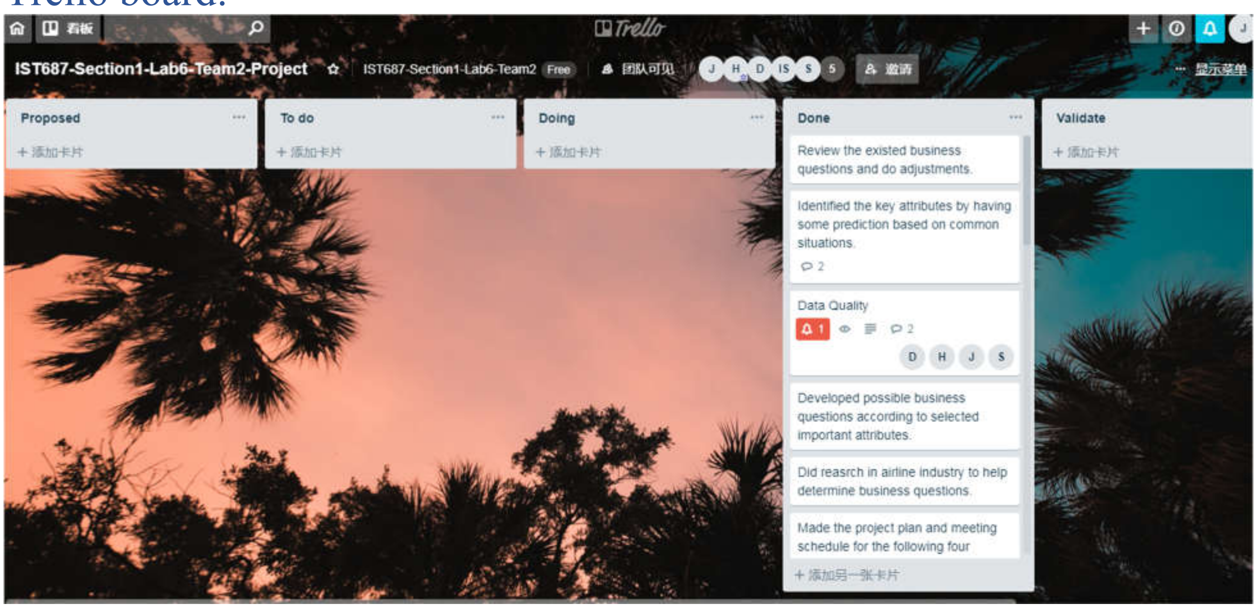
3. For the elder people, Southeast can improve the specific services, called “senior citizens privilege”, for example, offering fast pass card when they need to wait in lines.
4. For the customers who have few loyalty cards of other airline companies, they may prefer switch airline companies when they are traveling by plane. For those have not signed up membership in Southeast company, we can offer them gifts or discount to encourage them sign up, and then stay taking us as the first choice in the future traveling. For others who have signed up for our company and not having other companies’ loyalty cards, marketing department can send them niche targeting survey to get to know the reasons of dissatisfaction.
5. Customers who taking flights more frequent are more likely to have lower satisfaction. Therefore, our company can upgrade the classes of people who accumulate enough mileage.
6. Since people who experience delay tend to have lower satisfaction, our company should offer food and accommodations for the customers if flights delay.

As the research shown, people being in economy class tend to have lower satisfaction, the design of the airplane can be important for these customers to have a better experience, for example, offering extra pillows or upgrading the film and television equipment.

Limitation

Not having enough details about the reasons why customers are not satisfied with the services, it is not easy to have accurate suggestions about how to make the customers feel better compared to their previous experience. If the satisfaction level has more dimensions, for example, the satisfactions of cabin crew service, front desk services or online webpage service.

Trello board:



Division of work :

Tasks		People Conducted
1	Update 1	Harper He, Jingxian Sun
2	Update 2	Harper He,
3	Update 3	Harper He, Jingxian Sun
Report:		
4	Introduction	Harper He, Jingxian Sun
5	Business Questions	Harper He, Jingxian Sun, Dharmik Gautam Kothari, Sakshi Raghuvanshi
6	Data Accqucition Interpretation	Harper He
7	Data Cleaning Interpretation	Harper He
8	Data Transformation Interpretation	Jingxian Sun, Harper He
9	Data Mining Interpretation	Harper He
10	Descriptive Statistics and visualization	Jingxian Sun, Harper He
11	Associate Rules Analysis	Harper He, Jingxian Sun,
12	Modeling	Jingxian Sun, Harper He
13	SVM	Harper He
14	Actionable Insight	Jingxian Sun, Harper He
15	Limitation	Jingxian Sun
16	Cover Page/ Division of Job/Format	Jingxian Sun
Coding		
17	Data Cleaning	Harper He
17.1	Data Cleaning [not used]	Dharmik Gautam Kothari
18	Modeling	Harper He
18.1	Modeling [not used]	Jingxian Sun
19	Associate Rules Analysis	Harper He
19.1	Associate Rules Analysis [not used]	Dharmik Gautam Kothari
19.2	Associate Rules Analysis [not used]	Jingxian Sun
20	SVM	Harper He
20.1	SVM [not used]	Sakshi Raghuvanshi

Please Check Links:
17.1: https://drive.google.com/file/d/15fj6Ak1ECtIBfJTqc7_yql53pcA_kB2n/view?usp=sharing
18.1: https://drive.google.com/file/d/1wuJ3tJRBg-VrBVZn8fM_rA0IcfpfiUfR/view?usp=sharing
19.1:
https://drive.google.com/file/d/1s2EDca_St7gsBboe_GkqW8144kbpz1kd/view?usp=sharing
https://drive.google.com/file/d/1CvmZvyKcyhPo4QahQo5BCbBWBptcr1wM/view?usp=sharing
19.2:
https://drive.google.com/file/d/1ZZsMp7_fT82YYAfVJJ9ixUKcX5Szwd_n/view?usp=sharing
20.1:
https://drive.google.com/file/d/1s2EDca_St7gsBboe_GkqW8144kbpz1kd/view?usp=sharing
https://drive.google.com/file/d/1PZC4NVsW1UxJnOkqIsfaxvEyqC7nAR3W/view?usp=sharing

Appendix – Code

```
# Load the necessary pacakges
library(dplyr)
library(ggplot2)
library(ggrepel)
require(ggmap)
require(maps)
library(arules)
library(arulesViz)
library(kernlab)
library("ggthemes")
library("RColorBrewer")

##### Data Acquisition
# read the data
RawData <- read.csv(file="/Users/harperhe/Documents/IST 687/Project/Satisfaction
Survey.csv", header=TRUE, sep=",")
str(RawData)
# Find the columns containing NAs
colSums(is.na(RawData))

##### Data Cleansing and Munging
# remove 9 unusual satisfaction values
CleanData <- RawData[ ! RawData$Satisfaction %in% c('4.00.5', '4.00.2.00',1.5,2.5,3.5,4.5), ]
# Remove the Airline.Code and Flight.date attribute
CleanData <- CleanData[,-(15:16)]
# check the data
str(CleanData)
# Delete the white spece
CleanData$Airline.Name <- gsub('\\s+', '', CleanData$Airline.Name)
# Find the columns containing NAs
colSums(is.na(CleanData))
# transform attribute "satisfaction" to numeric
CleanData$Satisfaction <- as.numeric(as.character(CleanData$Satisfaction))

# Build subsets and clean NAs
# Build a subset for customers whose flights have been cancelled
CancelledSubset <- CleanData[which(CleanData$Flight.cancelled == "Yes"), ]
str(CancelledSubset)
colSums(is.na(CancelledSubset))

# Build a subset for customers whose flights have not been cancelled
UncancelledSubset <- CleanData[which(CleanData$Flight.cancelled == "No"), ]
str(UncancelledSubset)
colSums(is.na(UncancelledSubset))

# Remove the rows containing NAs
```

```
Satisfaction <- na.omit(UncancelledSubset, cols=c("Arrival.Delay.in.Minutes",
"Flight.time.in.minutes"))
str(Satisfaction)

##### Data Transformation
# Group some attributes for descriptive analysis and linear regression
Satisfaction$Age.Group <-ifelse(Satisfaction$Age < 18,'15-18',
                          ifelse(Satisfaction$Age >=18 & Satisfaction$Age <=24,'18-24',
                             ifelse(Satisfaction$Age >=25 & Satisfaction$Age <=34,'25-34',
                                ifelse(Satisfaction$Age >=35 & Satisfaction$Age <=44,'35-44',
                                   ifelse(Satisfaction$Age >=45 & Satisfaction$Age <=54,'45-54',
                                      ifelse(Satisfaction$Age >=55 & Satisfaction$Age
<=64,'55-64','65+'
                             ))))))
Satisfaction$Age.Group <- as.factor(Satisfaction$Age.Group)
str(Satisfaction)
Satisfaction$Scheduled.Departure.Hour.Group <-
ifelse(Satisfaction$Scheduled.Departure.Hour >=1 & Satisfaction$Scheduled.Departure.Hour
<=5,'early morning (1am-5am)',
                                       ifelse(Satisfaction$Scheduled.Departure.Hour >=6 &
Satisfaction$Scheduled.Departure.Hour <=11,'morning (6am-11am)',
                                          ifelse(Satisfaction$Scheduled.Departure.Hour >=12
& Satisfaction$Scheduled.Departure.Hour <=17,'afternoon (12pm-5pm)','evening (6pm-11pm)'
                                          )))
Satisfaction$Scheduled.Departure.Hour.Group  <-
as.factor(Satisfaction$Scheduled.Departure.Hour.Group)
# Map each numeric attribute to a category
#
Price.Sensitivity,Year.of.First.Flight,No.of.Flights.p.a,No..of.other.Loyalty.Cards,Departure.Delay
.in.Minutes,Flight.time.in.minutes, Flight.Distance
Satisfaction$Price.Sensitivity.Group <- as.factor(ifelse(Satisfaction$Price.Sensitivity >=4,
"sensitive", 'notsensitive'))
Satisfaction$Year.of.First.Flight.Group <- as.factor(ifelse(Satisfaction$Year.of.First.Flight <=
2007,"2003-2007","2008-2012"))
Satisfaction$Satisfaction.Group <- as.factor(ifelse(Satisfaction$Satisfaction >=
4,"satisfied","unsatisfied"))
FlightFeature1 <- function(v){
  vBuckets <- v
  q <- quantile(v, c(0.4, 0.6))
  vBuckets <- replicate(length(v), "Average")
  vBuckets[v <= q[1]] <- "Low"
  vBuckets[v > q[2]] <- "High"
  return(vBuckets)
}

Satisfaction$No.of.Flights.p.a.Group <- as.factor(FlightFeature1(Satisfaction$No.of.Flights.p.a.))
Satisfaction$No..of.other.Loyalty.Cards.Group <-
as.factor(FlightFeature1(Satisfaction$No..of.other.Loyalty.Cards))
Satisfaction$Departure.Delay.in.Minutes.Group <-
as.factor(FlightFeature1(Satisfaction$Departure.Delay.in.Minutes))
```

```
Satisfaction$Flight.time.in.minutes.Group <-
as.factor(FlightFeature1(Satisfaction$Flight.time.in.minutes))
Satisfaction$Flight.Distance.Group <- as.factor(FlightFeature1(Satisfaction$Flight.Distance))

str(Satisfaction)

# Build a subset for customers whose flights have been delayed
DelaySubset <- Satisfaction[which(Satisfaction$Arrival.Delay.greater.5.Mins == "yes"), ]
str(DelaySubset)

# Build a subset for customers whose flights have not been delayed
NoDelaySubset <- Satisfaction[which(Satisfaction$Arrival.Delay.greater.5.Mins == "no"), ]
str(NoDelaySubset)

# Create a subset of the customers whose flight haven't been cancelled of SE airlines.
SESubset <-  as.data.frame(filter(Satisfaction, Airline.Name=="SoutheastAirlinesCo."))
str(SESubset)
NoSESubset <- as.data.frame(filter(Satisfaction, Airline.Name!="SoutheastAirlinesCo."))
str(NoSESubset)

##### Descriptive statistics & Visualizations

#### flight status
# Calculate average satisfaction on different flight status
CancelledSati <- mean(CancelledSubset$Satisfaction)
CancelledSati
DelaySati<-mean(DelaySubset$Satisfaction)
DelaySati
NoDelaySati<-mean(NoDelaySubset$Satisfaction)
NoDelaySati

# Create a dataframe showing the flight status, number of customers and their Average
Satisfaction
SatiByFlightStatus <-data.frame("Flight.Status"=c("cancel","delay","ontime"),
"Number.of.customers"=c(nrow(CancelledSubset),nrow(DelaySubset),nrow(NoDelaySubset)))
SatiByFlightStatus$Average.Satisfaction <- c(CancelledSati,DelaySati,NoDelaySati)
str(SatiByFlightStatus)
SatiByFlightStatus

# Sample flight status distribution - using pie chart
label_value <- paste('(',
round(SatiByFlightStatus$Number.of.customers/sum(SatiByFlightStatus$Number.of.customers)
* 100, 1), '%)', sep = '')
label_value
label <- paste(SatiByFlightStatus$Flight.Status, label_value, sep = " ")
label
FlightStatusPieChart <- ggplot(data = SatiByFlightStatus, mapping = aes(x = 'Content', y =
Number.of.customers, fill = Flight.Status))+
  geom_bar(stat = 'identity', position = 'stack', width = 1)+
```

```
  coord_polar(theta = 'y') + labs(x = '', y = '', title = '')+
  theme(axis.text = element_blank()) + theme(axis.ticks = element_blank())+
  scale_fill_manual(breaks = SatiByFlightStatus$Flight.Status, labels = label,values =
c("#FFD966", "#37474F","#77909C"))+
  theme(legend.text = element_text(size=20))
FlightStatusPieChart

# Average satifaction of flight status - using bar chart
FlightStatusSatiCol <- ggplot(SatiByFlightStatus, aes(x=Flight.Status,y=Average.Satisfaction))+
  geom_col(width = 0.3,fill="#77909C", colour="#6E7B8B")+
  labs(title="Average satisfaction of different flight status",x="Flight Status", y="Average
Satisfaction")+
  theme(legend.text = element_text(size=20),axis.text =element_text(size=10))
FlightStatusSatiCol

#### Satisfaction distribution
summary(CleanData$Satisfaction)
SatiDistHist <- ggplot(CleanData, aes(x= Satisfaction))+
  geom_histogram(binwidth = 0.2,fill="#77909C", colour="#6E7B8B")+
  labs(title="Satisfaction of All the Customers")
SatiDistHist
SatiDist <- as.data.frame(CleanData %>%
                             group_by(Satisfaction) %>%
                             summarize(CustomerNumber=n()))
SatiDist

#### Satisfaction of different genders
# Calculate the average satisfaction of different genders
SatiByGender <- as.data.frame(Satisfaction %>%
                             group_by(Gender) %>%
                             summarize(CustomerNumber=n(),AverageSatisfaction =
mean(Satisfaction)))
SatiByGender

# Barchart to decribe the average satisfaction of different genders
GenSatiCol <- ggplot(SatiByGender, aes(x=Gender,y=AverageSatisfaction))+
  geom_col(width = 0.3,fill="#77909C", colour="#6E7B8B")+
  labs(title="Average satisfaction of different genders",x="Gender", y="Average Satisfaction")+
  theme(legend.text = element_text(size=20),axis.text =element_text(size=10))
GenSatiCol

#### Satisfaction of different ages
# Calculate the average satisfaction of different ages
SatiByAge <- as.data.frame(Satisfaction %>%
                             group_by(Age.Group) %>%
                             summarize(CustomerNumber=n(),AverageSatisfaction =
mean(Satisfaction)))
SatiByAge
# Barchart to decribe the average satisfaction of different ages
AgeSatiCol <- ggplot(SatiByAge, aes(x=Age.Group,y=AverageSatisfaction))+
```

```
  geom_col(width = 0.3,fill="#77909C", colour="#6E7B8B")+
  labs(title="Average satisfaction of different ages",x="Age.Group", y="Average Satisfaction")+
  theme(legend.text = element_text(size=20),axis.text =element_text(size=15))
AgeSatiCol

#### Satisfaction of different classes
# Calculate the average satisfaction of different classes
SatiByClass <- as.data.frame(Satisfaction %>%
                           group_by(Class) %>%
                           summarize(CustomerNumber=n(),AverageSatisfaction = mean(Satisfaction)))
SatiByClass
# Barchart to decribe the average satisfaction of different class
ClassSatiCol <- ggplot(SatiByClass, aes(x=Class,y=AverageSatisfaction))+
  geom_col(width = 0.3,fill="#77909C", colour="#6E7B8B")+
  labs(title="Average satisfaction of different classes",x="Class", y="Average Satisfaction")+
  theme(legend.text = element_text(size=20),axis.text =element_text(size=15))
ClassSatiCol

#### Satisfaction of different airlines
SatByAirlines <- data.frame(Satisfaction %>%
                            group_by(Airline.Name) %>%
                            summarize(CustomerNumber=n(),AverageSatisfaction =
mean(Satisfaction)))
str(SatByAirlines)
SatByAirlines <- SatByAirlines[order(-SatByAirlines$AverageSatisfaction),]
SatByAirlinesPlot <- ggplot(SatByAirlines,aes(x=reorder(Airline.Name,AverageSatisfaction),
y=AverageSatisfaction))+
  geom_col(fill="#77909C", colour="#6E7B8B",width=0.5)+
  labs(title="Average Satisfaction of All the Airlines", x="Airline Names", y="Average
Satisfaction")+
  coord_flip()+
  theme(axis.text.x = element_text(size = 14,color="black"),axis.text.y = element_text(size =
14,color="black"))+
  theme(axis.title.x = element_text(size = 14),axis.title.y = element_text(size = 14))
SatByAirlinesPlot

# Satisfaction of different type of travels
SatiByType <- data.frame(Satisfaction %>%
                            group_by(Type.of.Travel) %>%
                            summarize(CustomerNumber=n(),AverageSatisfaction =
mean(Satisfaction)))
str(SatiByType)
SatiByType
# Bar chart to decribe the average satisfaction of different types
TypeSatiCol <- ggplot(SatiByType, aes(x=Type.of.Travel,y=AverageSatisfaction))+
  geom_col(width = 0.3,fill="#77909C", colour="#6E7B8B")+
  labs(title="Average satisfaction of different types of travel",x="Type", y="Average
Satisfaction")+
  theme(legend.text = element_text(size=20),axis.text =element_text(size=15))
TypeSatiCol
```

```
#### Satisfaction of different locations
# Visualization of origin and destinations
states <- map_data("state")
# Satisfaction of different Origin.States
# Calculate mean 'satisfaction' of guests grouped by variable 'Origin.States'.
SatByOriStates <- data.frame(Satisfaction %>%
                             group_by(Origin.State) %>%
                             summarize(CustomerNumber=n(),AverageSatisfaction =
mean(Satisfaction)))
str(SatByOriStates)
SatByOriStates <- SatByOriStates[order(SatByOriStates$AverageSatisfaction),]
SatByOriStates$Origin.State <- tolower(SatByOriStates$Origin.State)

SatByOriStatesMap <- ggplot(SatByOriStates, aes(map_id = Origin.State))+
  geom_map(map = states, aes(fill = AverageSatisfaction))+
  expand_limits(x=states$long, y=states$lat)+
  coord_map() + ggtitle("Average Satisfaction for Origin States") + labs (x="Longitude",
y="Latitude")+
  scale_fill_gradient(high = "#6E7B8B",low = "white")
SatByOriStatesMap

# Satisfaction of different Destination.States
SatByDesStates <- data.frame(Satisfaction %>%
                               group_by(Destination.State) %>%
                             summarize(CustomerNumber=n(),AverageSatisfaction =
mean(Satisfaction)))
str(SatByDesStates)
SatByDesStates <- SatByDesStates [order(SatByDesStates$AverageSatisfaction),]
SatByDesStates
SatByDesStates$Destination.State<- tolower(SatByDesStates$Destination.State)

SatByDesStatesMap <- ggplot(SatByDesStates, aes(map_id = Destination.State))+
  geom_map(map = states, aes(fill = AverageSatisfaction))+
  expand_limits(x=states$long, y=states$lat)+
  coord_map() + ggtitle("Average Satisfaction for Destination States") + labs (x="Longitude",
y="Latitude")+
  scale_fill_gradient(high = "#6E7B8B",low = "white")
SatByDesStatesMap

# Satisfaction of different Origin.Cities
# Calculate mean 'satisfaction' of guests grouped by variable 'Origin.Cities'.
SatByOriCity <- data.frame(Satisfaction %>%
                           group_by(Orgin.City) %>%
                           summarize(CustomerNumber=n(),AverageSatisfaction = mean(Satisfaction)))
str(SatByOriCity)
SatByOriCity <- SatByOriCity[order(SatByOriCity$AverageSatisfaction),]
SatByOriCity
LowSatOriCity <- SatByOriCity[1:10,]
```

```
# Draw the map of the origin cities with the lowest satisfaction
LowSatOriCityGeo <- cbind(geocode(as.character(LowSatOriCity$Orgin.City),source = "dsk"),
LowSatOriCity)
LowSatOriCityGeo$Orgin.City <- tolower(LowSatOriCityGeo$Orgin.City)
LowSatOriCityGeo
str(LowSatOriCityGeo)

LowSatOriCityMap <- ggplot(data = states) +
  geom_polygon(aes(x = long, y = lat, group = group),colour = alpha("grey", 1/2), fill = "white") +
  coord_map() +
  geom_point(data=LowSatOriCityGeo, aes(x=lon, y=lat, size=AverageSatisfaction),
color="#FFD966")+
  geom_text_repel(data=LowSatOriCityGeo, aes(x=lon, y=lat, label=Orgin.City),size=5,
color="#77909C",fontface = "bold",vjust=1)+
  ggtitle("Ten origin cities with the lowest saftisfaction ")
LowSatOriCityMap

# Calculate mean 'satisfaction' of guests grouped by variable 'Des.Cities'.
SatByDesCity <- data.frame(Satisfaction %>%
                           group_by(Destination.City) %>%
                           summarize(CustomerNumber=n(),AverageSatisfaction = mean(Satisfaction)))
str(SatByDesCity)
SatByDesCity <- SatByDesCity[order(SatByDesCity$AverageSatisfaction),]
SatByDesCity
LowSatDesCity <- SatByDesCity[1:10,]

# Draw the map of the destination cities with the lowest satisfaction
LowSatDesCityGeo <- cbind(geocode(as.character(LowSatDesCity$Destination.City),source =
"dsk"), LowSatDesCity)
LowSatDesCityGeo$Destination.City <- tolower(LowSatDesCityGeo$Destination.City)
LowSatDesCityGeo
str(LowSatDesCityGeo)

LowSatDesCityMap <- ggplot(data = states) +
  geom_polygon(aes(x = long, y = lat, group = group),colour = alpha("grey", 1/2), fill = "white") +
  coord_map() +
  geom_point(data=LowSatDesCityGeo, aes(x=lon, y=lat, size=AverageSatisfaction),
color="#FFD966")+
  geom_text_repel(data=LowSatDesCityGeo, aes(x=lon, y=lat, label=Destination.City),size=3,
color="#77909C",fontface = "bold",vjust=1)+
  ggtitle("Ten destination cities with the lowest saftisfaction")
LowSatDesCityMap



##### Association rules mining
# transforma the data frame to transactions
SatiTrans <- data.frame(Satisfaction$Satisfaction.Group,
```

```
                        Satisfaction$Airline.Status,
                        Satisfaction$Gender,
                        Satisfaction$Type.of.Travel,
                        Satisfaction$Class,
                        Satisfaction$Arrival.Delay.greater.5.Mins,
                        Satisfaction$Age.Group,
                        Satisfaction$Scheduled.Departure.Hour.Group,
                        Satisfaction$Price.Sensitivity.Group,
                        Satisfaction$Year.of.First.Flight.Group,
                        Satisfaction$No.of.Flights.p.a.Group,
                        Satisfaction$No..of.other.Loyalty.Cards.Group,
                        Satisfaction$Departure.Delay.in.Minutes.Group,
                        Satisfaction$Flight.time.in.minutes.Group,
                        Satisfaction$Flight.Distance.Group)
str(SatiTrans)
colSums(is.na(SatiTrans))
SatiTrans <- as(SatiTrans,"transactions")
class(SatiTrans)
str(SatiTrans)

# Use the inspect( ), itemFrequency( ), and itemFrequencyPlot( ) commands to explore the
contents of SatiTrans.
inspect(head(SatiTrans))
inspect(tail(SatiTrans))
summary(SatiTrans)
itemFrequency(SatiTrans)
itemFrequencyPlot(SatiTrans, support=0.5, cex.names=0.5)
itemFrequencyPlot(SatiTrans, support=0.3, cex.names=0.5)

# Run the apriori command to try and predict satisfied customers (as defined by their overall
satisfaction being high – above 7).
SatirRuleset2 <- apriori(SatiTrans, parameter = list(support=0.1, confidence=0.4,minlen=2),
appearance = list(rhs ="Satisfaction.Satisfaction.Group=unsatisfied"))
ruleFeature2 <- inspect(SatirRuleset2)
plot(SatirRuleset2,jitter=0)
# Find the rules with high lift
GoodSatiRules2 <- SatirRuleset2[quality(SatirRuleset2)$lift > 2]
GoodSatiRules2
inspect(GoodSatiRules2)
plot(GoodSatiRules2,jitter=0)

# Improve the level of support and confidence to run the apriori command
SatirRuleset1 <- apriori(SatiTrans, parameter = list(support=0.25, confidence=0.4,minlen=2),
appearance = list(rhs ="Satisfaction.Satisfaction.Group=unsatisfied"))
ruleFeature1 <- inspect(SatirRuleset1)
plot(SatirRuleset1)
# Find the rules with high lift
GoodSatiRules1 <- SatirRuleset1[quality(SatirRuleset1)$lift > 1.15]
GoodSatiRules1
inspect(GoodSatiRules1)
```

```
plot(GoodSatiRules1)
# Find the rules with high support
HighSuppRules <- SatirRuleset1[quality(SatirRuleset1)$support > 0.35]
inspect(HighSuppRules)
# Find the rules with high confidence
HighConfiRules <- SatirRuleset1[quality(SatirRuleset1)$confidence > 0.6]
inspect(HighConfiRules)

##### Linear Model
#### Simple linear model
### Customers characteristic
# SatiVsAge
lm.SatiVsAge <- lm(formula= Satisfaction~ Age, data=Satisfaction)
summary(lm.SatiVsAge)
ggplot(Satisfaction,aes(x=Age, y=Satisfaction))+
  geom_point()+
  stat_smooth(method = "lm", col="#FFD966")

# SatiVsAgeGroup
lm.SatiVsAgeGroup <- lm(formula= Satisfaction~ Age.Group, data=Satisfaction)
summary(lm.SatiVsAgeGroup)
ggplot(Satisfaction,aes(x=Age.Group, y=Satisfaction))+
  geom_point()+
  stat_smooth(method = "lm", col="#FFD966")

# SatiVsPriceSensitivity
lm.SatiVsPriceSensitivity <- lm(formula= Satisfaction~ Price.Sensitivity, data=Satisfaction)
summary(lm.SatiVsPriceSensitivity)
ggplot(Satisfaction,aes(x=Price.Sensitivity, y=Satisfaction))+
  geom_point()+
  stat_smooth(method = "lm", col="#FFD966")

# Sati Vs Consuming behavior in airport
lm.SatiVsConsume <- lm(formula= Satisfaction~
Shopping.Amount.at.Airport+Eating.and.Drinking.at.Airport, data=Satisfaction)
summary(lm.SatiVsConsume)

# Sati Vs Shopping.Amount.at.Airport
lm.SatiVsShoppingAmount <- lm(formula= Satisfaction~ Shopping.Amount.at.Airport,
data=Satisfaction)
summary(lm.SatiVsShoppingAmount )
ggplot(Satisfaction,aes(x=Shopping.Amount.at.Airport, y=Satisfaction))+
  geom_point()+
  stat_smooth(method = "lm", col="#FFD966")

### Flight experience characteristic
# SatiVsYearFirst
lm.SatiVsYearFirst <- lm(formula= Satisfaction~ Year.of.First.Flight, data=Satisfaction)
summary(lm.SatiVsYearFirst)
```

```
ggplot(Satisfaction,aes(x=Year.of.First.Flight, y=Satisfaction))+
  geom_point()+
  stat_smooth(method = "lm", col="#FFD966")

# Sati Vs No.of.Flights.p.a.
lm.SatiVsNoFlight <- lm(formula= Satisfaction~ No.of.Flights.p.a., data=Satisfaction)
lm.SatiVsNoFlight

ggplot(Satisfaction,aes(x=No.of.Flights.p.a., y=Satisfaction))+
  geom_point()+
  stat_smooth(method = "lm", col="#FFD966")

### Flight characteristic

# Sati Vs Delay
lm.SatiVsDelay <- lm(formula= Satisfaction~ Departure.Delay.in.Minutes +
                      Arrival.Delay.in.Minutes, data=Satisfaction)
summary(lm.SatiVsDelay)

# Sati Vs Departure Delay
ggplot(Satisfaction,aes(x=Departure.Delay.in.Minutes, y=Satisfaction))+
  geom_point()+
  stat_smooth(method = "lm", col="#FFD966")

# Sati Vs Arrival Delay
ggplot(Satisfaction,aes(x=Arrival.Delay.in.Minutes, y=Satisfaction))+
  geom_point()+
  stat_smooth(method = "lm", col="#FFD966")

# Sati Vs Length of travel
OnwaySatilm <- lm(formula= Satisfaction~ Flight.time.in.minutes + Flight.Distance,
data=Satisfaction)
summary(OnwaySatilm)

# Sati Vs Flight.time.in.minutes
ggplot(Satisfaction,aes(x=Flight.time.in.minutes, y=Satisfaction))+
  geom_point()+
  stat_smooth(method = "lm", col="#FFD966")

# Sati Vs Flight.Distance
ggplot(Satisfaction,aes(x=Flight.Distance, y=Satisfaction))+
  geom_point()+
  stat_smooth(method = "lm", col="#FFD966")

# Sati Vs Scheduled.Departure.Hour
lm.SatiVsScheduled.Departure.Hour <- lm(formula= Satisfaction~ Scheduled.Departure.Hour,
data=Satisfaction)
```

```
summary(lm.SatiVsScheduled.Departure.Hour)
# Sati Vs Scheduled.Departure.Hour
ggplot(Satisfaction,aes(x=Scheduled.Departure.Hour, y=Satisfaction))+
  geom_point()+
  stat_smooth(method = "lm", col="#FFD966")

# Sati Vs Day.of.Month
lm.SatiVsDay.of.Month <- lm(formula= Satisfaction~ Day.of.Month, data=Satisfaction)
summary(lm.SatiVsDay.of.Month)
# Sati Vs Day.of.Month
ggplot(Satisfaction,aes(x=Day.of.Month, y=Satisfaction))+
  geom_point()+
  stat_smooth(method = "lm", col="#FFD966")

#### Multiple linear model
### Grouping attributes
## 1)    Customers characteristic
lm.CustomersCharacteristic<-
lm(formula=Satisfaction~Age+Gender+Price.Sensitivity+Shopping.Amount.at.Airport+Eating.an
d.Drinking.at.Airport, data = Satisfaction)
summary(lm.CustomersCharacteristic)

## 2) Flight experience characteristic
# a)     Previous flight experience: Year of First Flight; No of Flights, Percent of Flight with other
Airlines, No. Of other Loyalty Cards
# b)     Current flight experience: Airline Status, Type of Travel, Class,
str(Satisfaction)
lm.ExperienceCharacteristic <-
lm(formula=Satisfaction~Year.of.First.Flight+No.of.Flights.p.a.+X..of.Flight.with.other.Airlines+T
ype.of.Travel+No..of.other.Loyalty.Cards+Class+Airline.Status, data = Satisfaction)
summary(lm.ExperienceCharacteristic)

## 3)    Flight characteristic (12 attributes)
# a)     Geography: Flight Distance
# b)     Delay and cancellation: Scheduled Departure Hour, Departure Delay in Minutes, Arrival
Delay in Minutes, Flight time in minutes,
lm.FlightCharacteristic<-
lm(formula=Satisfaction~Day.of.Month+Scheduled.Departure.Hour+Departure.Delay.in.Minutes
+Arrival.Delay.in.Minutes+Flight.time.in.minutes+Flight.Distance, data = Satisfaction)
summary(lm.FlightCharacteristic)

### A full model
## Below are the models developed using stepwise
# Using 'Arrival.Delay.greater.5.Mins' instead of 'Arrival.Delay.in.Minutes' -- Adjusted R-
squared:  0.3791
lmAllSati <-
lm(formula=Satisfaction~Age+Gender+Price.Sensitivity+Year.of.First.Flight+No.of.Flights.p.a.+
X..of.Flight.with.other.Airlines+Type.of.Travel+No..of.other.Loyalty.Cards+Shopping.Amount.at.
Airport+Eating.and.Drinking.at.Airport+Class+Day.of.Month+Scheduled.Departure.Hour.Group+
Flight.time.in.minutes+Flight.Distance+Arrival.Delay.greater.5.Mins, data = Satisfaction)
```

```
summary(lmAllSati)
# Keep the significant attributes -- Adjusted R-squared: 0.3791
lmSigSati <-
lm(formula=Satisfaction~Age+Gender+Price.Sensitivity+Year.of.First.Flight+No.of.Flights.p.a.+
X..of.Flight.with.other.Airlines+Type.of.Travel+No..of.other.Loyalty.Cards+Shopping.Amount.at.
Airport+Eating.and.Drinking.at.Airport+Class+Scheduled.Departure.Hour.Group+Arrival.Delay.g
reater.5.Mins, data = Satisfaction)
summary(lmSigSati)
# Only using the numeric attribute to do the linear regression--Adjusted R-squared:  0.1048
lmAllNumSati <-
lm(formula=Satisfaction~Age+Price.Sensitivity+Year.of.First.Flight+No.of.Flights.p.a.+X..of.Fligh
t.with.other.Airlines+No..of.other.Loyalty.Cards+Shopping.Amount.at.Airport+Eating.and.Drinkin
g.at.Airport+Day.of.Month+Scheduled.Departure.Hour+Departure.Delay.in.Minutes+Arrival.Dela
y.in.Minutes+Flight.time.in.minutes+Flight.Distance, data = Satisfaction)
summary(lmAllNumSati)
# Using 'Arrival.Delay.in.Minutes' instead of 'Arrival.Delay.greater.5.Mins' -- Adjusted R-
squared:  0.3589
lmAllSati2 <-
lm(formula=Satisfaction~Age+Gender+Price.Sensitivity+Year.of.First.Flight+No.of.Flights.p.a.+
X..of.Flight.with.other.Airlines+Type.of.Travel+No..of.other.Loyalty.Cards+Shopping.Amount.at.
Airport+Eating.and.Drinking.at.Airport+Class+Day.of.Month+Scheduled.Departure.Hour.Group+
Flight.time.in.minutes+Flight.Distance+Arrival.Delay.in.Minutes, data = Satisfaction)
summary(lmAllSati2)
# Using 'Age.Group' instead of 'Age' & 'Arrival.Delay.greater.5.Mins' instead of
'Arrival.Delay.in.Minutes' --Adjusted R-squared:  0.3974
lmAllSati3 <-
lm(formula=Satisfaction~Age.Group+Gender+Price.Sensitivity+Year.of.First.Flight+No.of.Flights
.p.a.+X..of.Flight.with.other.Airlines+Type.of.Travel+No..of.other.Loyalty.Cards+Shopping.Amou
nt.at.Airport+Eating.and.Drinking.at.Airport+Class+Day.of.Month+Scheduled.Departure.Hour.Gr
oup+Flight.time.in.minutes+Flight.Distance+Arrival.Delay.greater.5.Mins, data = Satisfaction)
summary(lmAllSati3)
# Keep the significant attributes -- Adjusted R-squared: 0.3974
lmSigSati3 <-
lm(formula=Satisfaction~Age.Group+Gender+Price.Sensitivity+Year.of.First.Flight+No.of.Flights
.p.a.+Type.of.Travel+No..of.other.Loyalty.Cards+Shopping.Amount.at.Airport+Eating.and.Drinki
ng.at.Airport+Class+Scheduled.Departure.Hour.Group+Arrival.Delay.greater.5.Mins, data =
Satisfaction)
summary(lmSigSati3)
# Using 'Age.Group' instead of 'Age' &'Arrival.Delay.in.Minutes' instead of
'Arrival.Delay.greater.5.Mins' --Adjusted R-squared:  0.3772
lmAllSati4 <-
lm(formula=Satisfaction~Age.Group+Gender+Price.Sensitivity+Year.of.First.Flight+No.of.Flights
.p.a.+X..of.Flight.with.other.Airlines+Type.of.Travel+No..of.other.Loyalty.Cards+Shopping.Amou
nt.at.Airport+Eating.and.Drinking.at.Airport+Class+Day.of.Month+Scheduled.Departure.Hour.Gr
oup+Flight.time.in.minutes+Flight.Distance+Arrival.Delay.in.Minutes, data = Satisfaction)
summary(lmAllSati4)

## Below are the models developed based on the association rules mining results
# linear regression based on the results of ARS (using numeric attributes)- Adjusted R-
squared:  0.4427
```

```
lmAllSati5 <-
lm(formula=Satisfaction~Airline.Status+Gender+Price.Sensitivity+Year.of.First.Flight+Type.of.Tr
avel+No..of.other.Loyalty.Cards+Class+Departure.Delay.in.Minutes+Arrival.Delay.greater.5.Min
s, data = Satisfaction)
summary(lmAllSati5)

# linear regression based on the results of ARS (using category attributes) - Adjusted R-
squared: 0.4432
lmAllSati6 <-
lm(formula=Satisfaction~Airline.Status+Gender+Price.Sensitivity.Group+Year.of.First.Flight.Gro
up+Type.of.Travel+No..of.other.Loyalty.Cards.Group+Class+Departure.Delay.in.Minutes.Group
+Arrival.Delay.greater.5.Mins, data = Satisfaction)
summary(lmAllSati6)

# So the lmAllSati6 is the best model

##### Support Vector Machines
# Considering the data volumn, we'd better only use the Southeast Airlines' data to do the
modeling.
table(SESubset$Satisfaction.Group)
# Create training and test data sets
Satinrows <- nrow(SESubset)
Sati.random.indexes <- sample(1:Satinrows, replace=FALSE)
SatiCutPoint <- floor(Satinrows/3*2)
SatiTrainData <- SESubset[Sati.random.indexes[1:SatiCutPoint],]
SatiTestData <- SESubset[Sati.random.indexes[(SatiCutPoint+1):Satinrows],]
# Use the dim( ) function to demonstrate that the resulting training data set and test data set
contain the appropriate number of cases.
dim(SatiTrainData)
str(SatiTrainData)
dim(SatiTestData)
str(SatiTrainData)

#### Build a support vector model using the ksvm( ) function using all the variables to predict an
unsatisfied customer.
SatisvmOutput <- ksvm(Satisfaction.Group ~
Age+Gender+Price.Sensitivity+Year.of.First.Flight+Type.of.Travel+No..of.other.Loyalty.Cards+S
hopping.Amount.at.Airport+Eating.and.Drinking.at.Airport+Class+Scheduled.Departure.Hour+Ar
rival.Delay.greater.5.Mins, data=SatiTrainData, kernel= "rbfdot", kpar = "automatic", C = 5,
cross = 3, prob.model = TRUE)
SatisvmOutput

# Making a Prediction variable based on number of votes
SatiSvmPrediction <- predict(SatisvmOutput, SatiTestData, type = "votes")
str(SatiSvmPrediction)
head(SatiSvmPrediction[2,])

# Creating a composite table based on satisfied customers and SVM Prediction
SatiCompTable<-data.frame(SatiTestData$Satisfaction.Group,SatiSvmPrediction[2,])
```

```
# Creating a confusion matrix
ConfusionMatrix<-table(SatiCompTable)
ConfusionMatrix
# Creating a dataframe containing sum of errors
SatiErrorSum <- ConfusionMatrix[1,2]+ConfusionMatrix[2,1]
# Creating percentage of error rate
SatiErrorRate<-SatiErrorSum/sum(ConfusionMatrix)*100
SatiErrorRate

#### Use only numeric variables
SatisvmOutput2 <- ksvm(Satisfaction.Group ~
Age+Price.Sensitivity+Year.of.First.Flight+No.of.Flights.p.a.+X..of.Flight.with.other.Airlines+No..
of.other.Loyalty.Cards+Shopping.Amount.at.Airport+Eating.and.Drinking.at.Airport+Day.of.Mont
h+Scheduled.Departure.Hour+Departure.Delay.in.Minutes+Arrival.Delay.in.Minutes+Flight.time.i
n.minutes+Flight.Distance, data=SatiTrainData, kernel= "rbfdot", kpar = "automatic", C = 5,
cross = 3, prob.model = TRUE)
SatisvmOutput2

SatiSvmPrediction2 <- predict(SatisvmOutput2, SatiTestData, type = "votes") # Making a
Prediction variable based on number of votes
str(SatiSvmPrediction2)
head(SatiSvmPrediction2[2,])

# Creating a composite table based on satisfied customers and SVM Prediction
SatiCompTable2<-data.frame(SatiTestData$Satisfaction.Group,SatiSvmPrediction2[2,])
# Creating a confusion matrix
ConfusionMatrix2<-table(SatiCompTable2)
ConfusionMatrix2
# Creating a dataframe containing sum of errors
SatiErrorSum2 <- ConfusionMatrix2[1,2]+ConfusionMatrix2[2,1]
# Creating percentage of error rate
SatiErrorRate2<-SatiErrorSum2/sum(ConfusionMatrix2)*100
SatiErrorRate2

#### change different variables to check the error rate
SatisvmOutput3 <- ksvm(Satisfaction.Group ~
No..of.other.Loyalty.Cards+Gender+Type.of.Travel+Age+Price.Sensitivity+Year.of.First.Flight+
No.of.Flights.p.a.+X..of.Flight.with.other.Airlines+Class+Shopping.Amount.at.Airport+Eating.and
.Drinking.at.Airport+Scheduled.Departure.Hour, data=SatiTrainData, kernel= "rbfdot", kpar =
"automatic", C = 5, cross = 3, prob.model = TRUE)
SatisvmOutput3

SatiSvmPrediction3 <- predict(SatisvmOutput3, SatiTestData, type = "votes")
# Making a Prediction variable based on number of votes
str(SatiSvmPrediction3)
head(SatiSvmPrediction3[2,])

# Creating a composite table based on satisfied customers and SVM Prediction
SatiCompTable3<-data.frame(SatiTestData$Satisfaction.Group,SatiSvmPrediction3[2,])
```

```
# Creating a confusion matrix
ConfusionMatrix3<-table(SatiCompTable3)
ConfusionMatrix3
# Creating a dataframe containing sum of errors
SatiErrorSum3 <- ConfusionMatrix3[1,2]+ConfusionMatrix3[2,1]
# Creating percentage of error rate
SatiErrorRate3<-SatiErrorSum3/sum(ConfusionMatrix3)*100
SatiErrorRate3

SatisvmOutput4 <- ksvm(Satisfaction.Group ~
Gender+Type.of.Travel+No..of.other.Loyalty.Cards+Class+Airline.Status+Price.Sensitivity,
data=SatiTrainData, kernel= "rbfdot", kpar = "automatic", C = 5, cross = 3, prob.model = TRUE)
SatisvmOutput4
SatiSvmPrediction4 <- predict(SatisvmOutput4, SatiTestData, type = "votes") # Making a
Prediction variable based on number of votes
str(SatiSvmPrediction4)
head(SatiSvmPrediction4[2,])
# Creating a composite table based on satisfied customers and SVM Prediction
SatiCompTable4<-data.frame(SatiTestData$Satisfaction.Group,SatiSvmPrediction4[2,])
# Creating a confusion matrix
ConfusionMatrix4<-table(SatiCompTable4)
ConfusionMatrix4
# Creating a dataframe containing sum of errors
SatiErrorSum4 <- ConfusionMatrix4[1,2]+ConfusionMatrix4[2,1]
# Creating percentage of error rate
SatiErrorRate4<-SatiErrorSum4/sum(ConfusionMatrix4)*100
SatiErrorRate4

#### Use grouped attributes to build model
### 1) Customers characteristic (5 attributes)
#a)     Demographic: Age, Gender
#b)     Consuming behavior: Shopping Amount at Airport; Eating and Drinking at Airport; Price
Sensitivity
SvmCC <- ksvm(Satisfaction.Group ~
Age+Gender+Price.Sensitivity+Shopping.Amount.at.Airport+Eating.and.Drinking.at.Airport,
data=SatiTrainData, kernel= "rbfdot", kpar = "automatic", C = 5, cross = 3, prob.model = TRUE)
SvmCC
# Making a Prediction variable based on number of votes
SvmCCPrediction <- predict(SvmCC, SatiTestData, type = "votes")
str(SvmCCPrediction)
head(SvmCCPrediction[2,])

# Creating a composite table based on satisfied customers and SVM Prediction
SvmCCCompTable<-data.frame(SatiTestData$Satisfaction.Group,SvmCCPrediction[2,])
# Creating a confusion matrix
SvmCCConfusionMatrix<-table(SvmCCCompTable)
SvmCCConfusionMatrix
# Creating a dataframe containing sum of errors
SvmCCErrorSum <- SvmCCConfusionMatrix[1,2]+SvmCCConfusionMatrix[2,1]
# Creating percentage of error rate
```

```
SvmCCErrorRate<-SvmCCErrorSum/sum(SvmCCConfusionMatrix)*100
SvmCCErrorRate

SvmCC2 <- ksvm(Satisfaction.Group ~
Age+Price.Sensitivity+Shopping.Amount.at.Airport+Eating.and.Drinking.at.Airport,
data=SatiTrainData, kernel= "rbfdot", kpar = "automatic", C = 5, cross = 3, prob.model = TRUE)
SvmCC2
SvmCCPrediction2 <- predict(SvmCC2, SatiTestData, type = "votes") # Making a Prediction
variable based on number of votes
str(SvmCCPrediction2)
head(SvmCCPrediction2[2,])

# Creating a composite table based on satisfied customers and SVM Prediction
SvmCCCompTable2<-data.frame(SatiTestData$Satisfaction.Group,SvmCCPrediction2[2,])
# Creating a confusion matrix
SvmCCConfusionMatrix2<-table(SvmCCCompTable2)
SvmCCConfusionMatrix2
# Creating a dataframe containing sum of errors
SvmCCErrorSum2 <- SvmCCConfusionMatrix2[1,2]+SvmCCConfusionMatrix2[2,1]
# Creating percentage of error rate
SvmCCErrorRate2<-SvmCCErrorSum2/sum(SvmCCConfusionMatrix2)*100
SvmCCErrorRate2

### 2)  Flight experience characteristic (7 attributes)
#a)      Previous flight experience: Year of First Flight; No of Flights, Percent of Flight with other
Airlines, No. Of other Loyalty Cards
#b)      Current flight experience: Airline Status, Type of Travel, Class,
str(SatiTrainData)
SvmFE <- ksvm(Satisfaction.Group
~Year.of.First.Flight+No.of.Flights.p.a.+X..of.Flight.with.other.Airlines+No..of.other.Loyalty.Card
s , data=SatiTrainData, kernel= "rbfdot", kpar = "automatic", C = 5, cross = 3, prob.model =
TRUE)
SvmFE
SvmFEPrediction <- predict(SvmFE, SatiTestData, type = "votes") # Making a Prediction
variable based on number of votes
str(SvmFEPrediction)
head(SvmFEPrediction[2,])

# Creating a composite table based on satisfied customers and SVM Prediction
SvmFECompTable<-data.frame(SatiTestData$Satisfaction.Group,SvmFEPrediction[2,])
# Creating a confusion matrix
SvmFEConfusionMatrix<-table(SvmFECompTable)
SvmFEConfusionMatrix
# Creating a dataframe containing sum of errors
SvmFEErrorSum <- SvmFEConfusionMatrix[1,2]+SvmFEConfusionMatrix[2,1]
# Creating percentage of error rate
SvmFEErrorRate<-SvmFEErrorSum/sum(SvmFEConfusionMatrix)*100
SvmFEErrorRate

### 3)  Flight characteristic (12 attributes)
```

```
#a)	Geography: Origin City, Origin State, Destination City, Destination State, Flight Distance
#b)	Delay and cancellation: Scheduled Departure Hour, Departure Delay in Minutes, Arrival Delay in Minutes, Flight time in minutes, Arrival Delay greater 5 Mins, Flight cancelled
SvmFC <- ksvm(Satisfaction.Group ~Flight.Distance+Scheduled.Departure.Hour+Departure.Delay.in.Minutes+Arrival.Delay.in.Minutes+Flight.time.in.minutes, data=SatiTrainData, kernel= "rbfdot", kpar = "automatic", C = 5, cross = 3, prob.model = TRUE)
SvmFC
SvmFCPrediction <- predict(SvmFC, SatiTestData, type = "votes") # Making a Prediction variable based on number of votes
str(SvmFCPrediction)
head(SvmFCPrediction[2,])

# Creating a composite table based on satisfied customers and SVM Prediction
SvmFCCompTable<-data.frame(SatiTestData$Satisfaction.Group,SvmFCPrediction[2,])
# Creating a confusion matrix
SvmFCConfusionMatrix<-table(SvmFCCompTable)
SvmFCConfusionMatrix
# Creating a dataframe containing sum of errors
SvmFCErrorSum <- SvmFCConfusionMatrix[1,2]+SvmFCConfusionMatrix[2,1]
# Creating percentage of error rate
SvmFCErrorRate<-SvmFCErrorSum/sum(SvmFCConfusionMatrix)*100
SvmFCErrorRate

### Using the result of Association Rules Mining to build the model
# Use the numeric variables
SatisvmOutput5 <- ksvm(Satisfaction.Group ~ Airline.Status+Gender+Price.Sensitivity+Year.of.First.Flight+Type.of.Travel+No..of.other.Loyalty.Cards+Class+Departure.Delay.in.Minutes+Arrival.Delay.greater.5.Mins, data=SatiTrainData, kernel= "rbfdot", kpar = "automatic", C = 5, cross = 3, prob.model = TRUE)
SatisvmOutput5

# Making a Prediction variable based on number of votes
SatiSvmPrediction5 <- predict(SatisvmOutput5, SatiTestData, type = "votes")
str(SatiSvmPrediction5)
head(SatiSvmPrediction5[2,])

# Creating a composite table based on satisfied customers and SVM Prediction
SatiCompTable5<-data.frame(SatiTestData$Satisfaction.Group,SatiSvmPrediction5[2,])
# Creating a confusion matrix
ConfusionMatrix5<-table(SatiCompTable5)
ConfusionMatrix5
# Creating a dataframe containing sum of errors
SatiErrorSum5 <- ConfusionMatrix5[1,2]+ConfusionMatrix5[2,1]
# Creating percentage of error rate
SatiErrorRate5<-SatiErrorSum5/sum(ConfusionMatrix5)*100
SatiErrorRate5

### Using the result of Association Rules Mining to build the model
# Use the category variables
```

```
SatisvmOutput6 <- ksvm(Satisfaction.Group ~ Airline.Status+Gender+
                        Price.Sensitivity.Group+Year.of.First.Flight.Group+
                        Type.of.Travel+No..of.other.Loyalty.Cards.Group+
                        Class+Departure.Delay.in.Minutes.Group+
                        Arrival.Delay.greater.5.Mins,
                      data=SatiTrainData, kernel= "rbfdot", kpar = "automatic",
                      C = 5, cross = 3, prob.model = TRUE)
SatisvmOutput6

# Making a Prediction variable based on number of votes
SatiSvmPrediction6 <- predict(SatisvmOutput6, SatiTestData, type = "votes")
str(SatiSvmPrediction6)
head(SatiSvmPrediction6[2,])

# Creating a composite table based on satisfied customers and SVM Prediction
SatiCompTable6<-data.frame(SatiTestData$Satisfaction.Group,SatiSvmPrediction6[2,])
# Creating a confusion matrix
ConfusionMatrix6<-table(SatiCompTable6)
ConfusionMatrix6
# Creating a dataframe containing sum of errors
SatiErrorSum6 <- ConfusionMatrix6[1,2]+ConfusionMatrix6[2,1]
# Creating percentage of error rate
SatiErrorRate6<-SatiErrorSum6/sum(ConfusionMatrix6)*100
SatiErrorRate6

### Use SESubest to model, use WestAirways to predict
# Build a model with Southeast airlines
SatisvmOutput7 <- ksvm(Satisfaction.Group ~ Airline.Status+Gender+
                        Price.Sensitivity.Group+Year.of.First.Flight.Group+
                        Type.of.Travel+No..of.other.Loyalty.Cards.Group+
                        Class+Departure.Delay.in.Minutes.Group+
                        Arrival.Delay.greater.5.Mins,
                      data=SESubset, kernel= "rbfdot", kpar = "automatic",
                      C = 5, cross = 3, prob.model = TRUE)
SatisvmOutput7
# Create a subset of WestAirwaysInc
WASubset <- data.frame(filter(Satisfaction, Airline.Name=="WestAirwaysInc."))
# Making a Prediction variable based on number of votes
SatiSvmPrediction7 <- predict(SatisvmOutput7, WASubset, type = "votes")
str(SatiSvmPrediction7)
head(SatiSvmPrediction7[2,])
# Creating a composite table based on satisfied customers and SVM Prediction
SatiCompTable7<-data.frame(WASubset$Satisfaction.Group,SatiSvmPrediction7[2,])
# Creating a confusion matrix
ConfusionMatrix7<-table(SatiCompTable7)
ConfusionMatrix7
# Creating a dataframe containing sum of errors
SatiErrorSum7 <- ConfusionMatrix7[1,2]+ConfusionMatrix7[2,1]
# Creating percentage of error rate
SatiErrorRate7<-SatiErrorSum7/sum(ConfusionMatrix7)*100
```

```
SatiErrorRate7

# use EnjoyFlyingAirServices to predict
WFSubset <- data.frame(filter(Satisfaction, Airline.Name=="EnjoyFlyingAirServices"))
# Making a Prediction variable based on number of votes
SatiSvmPrediction8 <- predict(SatisvmOutput7, WFSubset, type = "votes")
str(SatiSvmPrediction8)
head(SatiSvmPrediction8[2,])

# Creating a composite table based on satisfied customers and SVM Prediction
SatiCompTable8<-data.frame(WFSubset$Satisfaction.Group,SatiSvmPrediction8[2,])
# Creating a confusion matrix
ConfusionMatrix8<-table(SatiCompTable8)
ConfusionMatrix8
# Creating a dataframe containing sum of errors
SatiErrorSum8 <- ConfusionMatrix8[1,2]+ConfusionMatrix8[2,1]
# Creating percentage of error rate
SatiErrorRate8<-SatiErrorSum8/sum(ConfusionMatrix8)*100
SatiErrorRate8
```