Introduction:

Customer experience plays a significant role in the aviation industry. Airlines completely rely on customers for increasing their revenue and market share. This project is about analyzing a large customer survey dataset for airlines which contains 130,000 survey responses of customers flying within the United States of 14 different airlines. This huge dataset has 28 attributes such as customer satisfaction, airline status, age, gender, price sensitivity, type of travel, flight cancelled, arrival and departure delay in minutes, origin city, origin state, destination city, destination state and so on. The main aim of this project is to predict the customers with low satisfaction. Improving the customer satisfaction can increase the revenue for the company and would help in retaining the customers for the airlines.

Description of the attributes:

Given below is the list of 28 attributes which were provided for analysis:

1. **Satisfaction** – it is rated from 1 to 5, that how satisfied is the customer? 5 means higher satisfied, and 1 is lowest level of satisfaction.
2. **Airline Status** – each customer has a different type of airline status or package, which are platinum, gold, silver, and blue.
3. **Age** – the specific customer's age. That is starting from 15 to 85 years old.
4. **Gender** – male or female.
5. **Price Sensitivity** – the grade to which the price affects to customers purchasing. The price sensitivity has a range from 0 to 5.
6. **Year of First Flight** – this attributes shows the first flight of each single customer. The range of year of the first flight for each customer has been started in 2003 until 2012.
7. **No of Flights p. a.** – this could be the number of flights that each customer has taken. The range starting from 0 to 100.
8. **Percent of Flight with other Airlines** – if we were Southeast Airline, we would like to know how many time that customer fly with other Airlines.
9. **Type of Travel** – is provide three traveling purpose for each consumer, which are business travel, mileage tickets that based on loyalty card, and personal travel like to see the family or in vacation
10. **No. Of other Loyalty Cards** – it is kind of membership card of each customer, that for retail establishment to gain a benefits such as, discounts.
11. **Shopping Amount at Airport** – showing the costumer's result of how many products have been purchased. The range of shopping amount is from 0 to 875.

12. **Eating and Drinking at Airport** – it is the quantity eating and drinking per each consumer at the airport. The masseur of how often for eating and drinking, which is 0 to 895.
13. **Class** – it consisted of three different kinds of service level such as, business, and economy plus, economy. Moreover, customers have optional to choose their seat.
14. **Day of Month** – it means the traveling day of each costumer. In this attribute, shows total of 31 days of the month.
15. **Flight date** – all of these data are abbreviate the passenger's flight date travel, which were since 2014 and only in January, February, and March.
16. **Airline Code** – basically, it is unique two or three digits that mean what is the specific type of airline. There are several codes that consumers have been going with. For example, AA, AS, B6, and DL.
17. **Airline Name** – There are several airlines company names such as, West Airways, Southeast Airlines Co, and FlyToSun Airlines Inc. This attribute provide what airline name that passenger have been used.
18. **Origin City** – refers to actual city that customers have departed from. For example, Yuma AZ, Waco TX, and Toledo HO.
19. **Origin State** – same thing as origin city such as, what state that customers have departed from? A good example, Texas, Ohio, Alaska, and Utah.
20. **Destination City** – the place to which passenger travels to. For example, Akron HO, Alpena MI, Austin TX, and Boston MA.
21. **Destination State** – also, it is the same thing as origin city, such as, to what state passenger travel to? Some example of destination states, Alaska, Kentucky, Iowa, and Florida.
22. **Scheduled Departure Hour** – the specific time at which passengers are scheduled to depart. In this data in scheduled departure hour is starting at 1 am until 23 pm.
23. **Departure Delay in Minutes** – which are minutes of departure delayed for each passenger, when compared to schedule. In this data the rage are starting from 0 until 1128 minutes.
24. **Arrival Delay in Minutes** – how many minutes of arrival delayed of each passenger. Rang of delayed minutes in this data are starting from 0 until 1115 minutes.
25. **Flight Cancelled** – occurs when the airline dose not operates the flight at all, and that is for a certain reason.
26. **Flight time in minutes** – indicate to period time to the destination.
27. **Flight Distance** – the extent of space between two places. Also, that means how many minutes are passenger traveling between two different places. Rang in this data starting from 31 until 4983 minutes.
28. **Arrival Delay greater 5 Minutes** – It means the delay of arrival airline time, which is more than 5 minutes per each passenger in the data.

Scope of the project:

For predicting the customers with low satisfaction, we considered performing the following steps:

- Formulating business questions
- Cleaning the data
- Visualizing the data
- Developing models for predicting low satisfaction
- Providing actionable recommendations for improving the customer satisfaction for the airlines

Business Questions:

For the airline which has the maximum number of unhappy customers, given below are some of the business questions that were formulated:

1. What are the factors that will affect customer satisfaction?
2. What are the characteristics of the customers with low satisfaction?
3. What are the recommendations for improving the satisfaction of the customers?

Cleaning the data:

Step 1: Read the dataset which is in .csv format

Step 2: Trimmed the leading and trailing blank spaces and changed the column names to readable format

Step 3: It was observed that NAs accounted to 6% of the entire dataset. Instead of omitting the NAs, replaced the NAs in *Flight_time_in_minutes* attribute with the mean values

```
> sum(is.na(survey)) # 6% of entire dataset has NAs
[1] 7821
```

Step 4: Replaced the NAs in *Arrival_delay_in_minutes* attribute with its mean values.

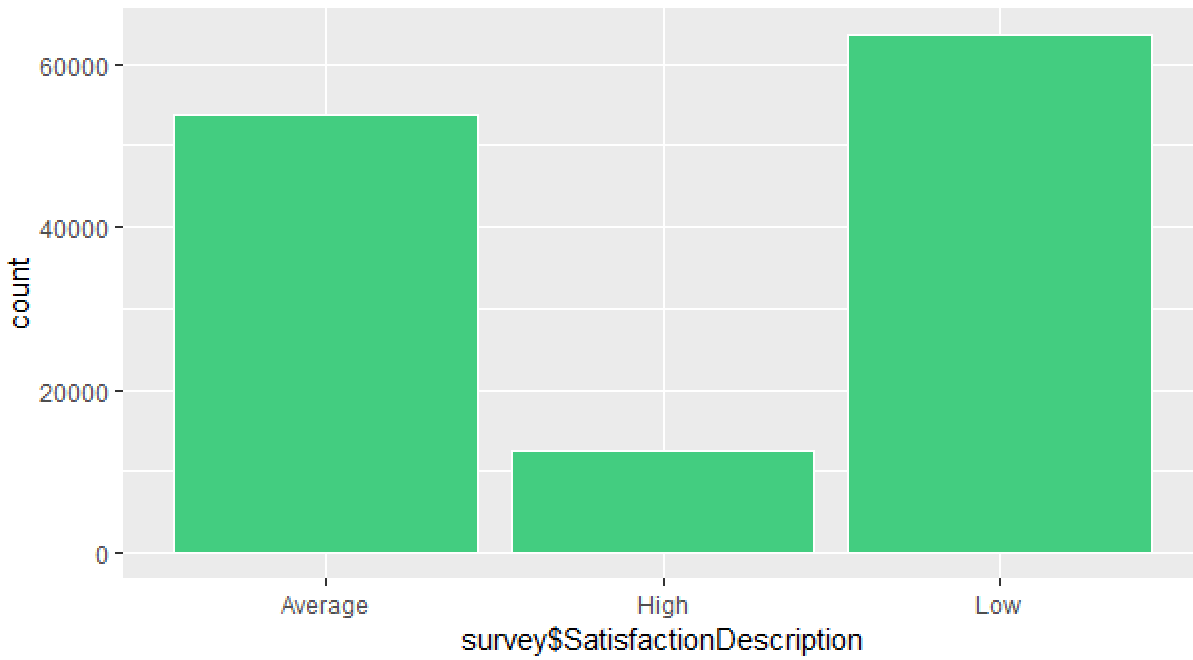
Step 5: Replaced the NAs in *Departure_delay_in_minutes* attribute with its mean values.

Descriptive Statistics:

The following analyses were made on the entire dataset to analyze and observe the attributes:

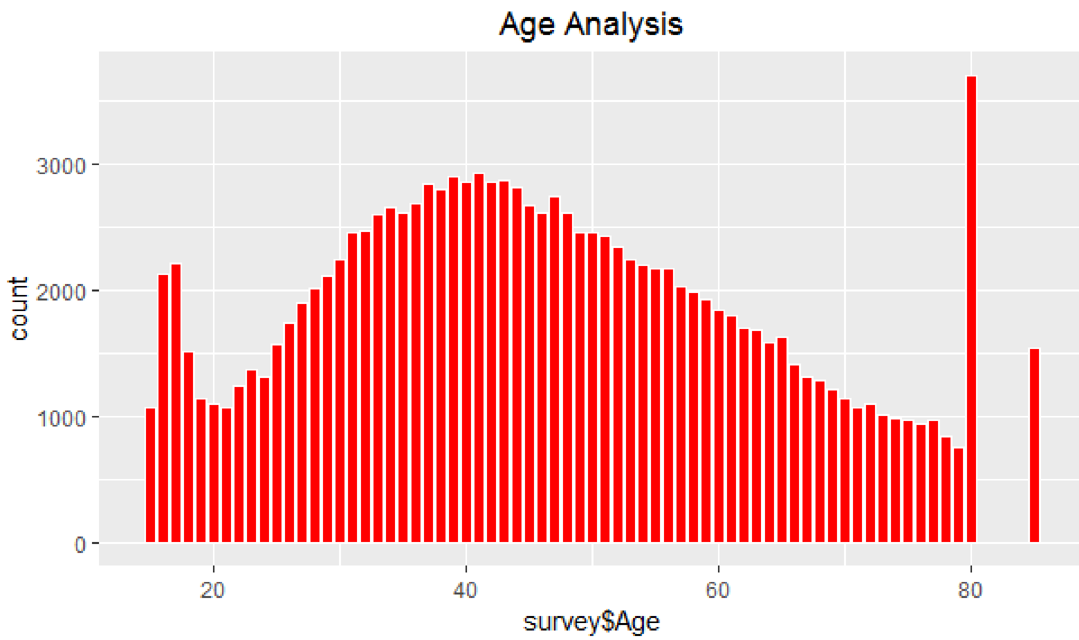
Satisfaction description:

The following plot displays the segregation of customers based on their satisfaction levels i.e. Low, Average and High



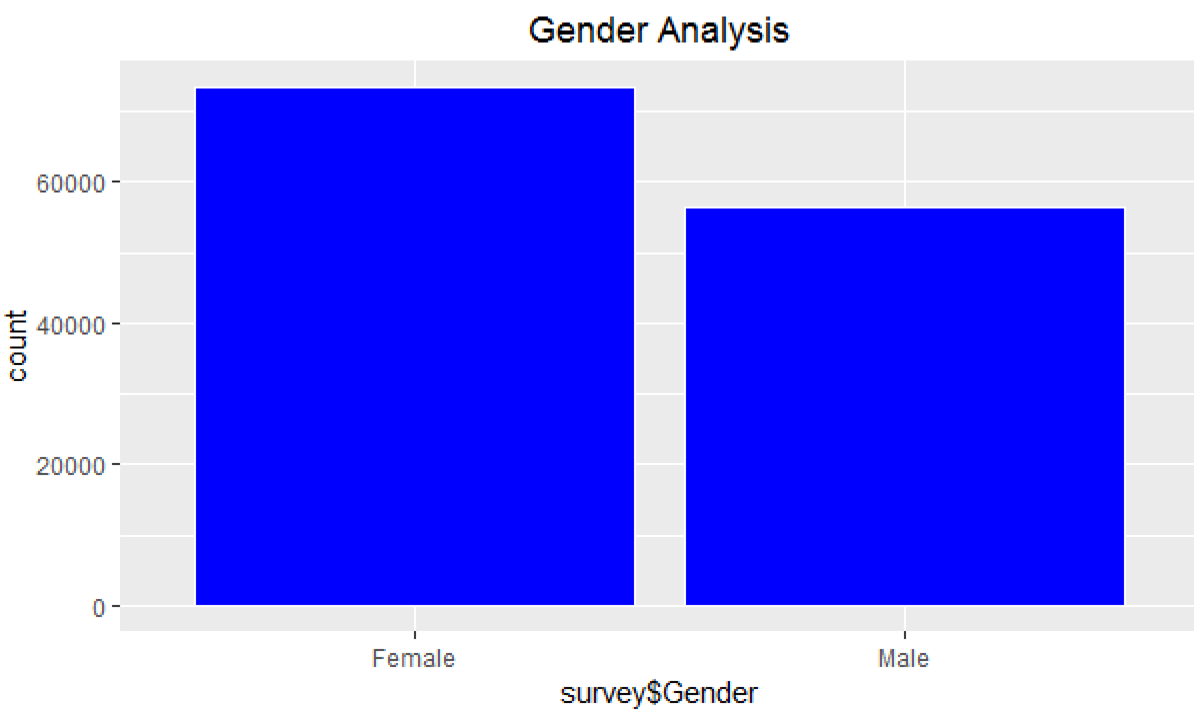
Age analysis:

Based on the age of the customer, the following plot depicts the varied age groups that participated in this survey. It can be seen that other than the outlier age group of 80, customers from the age of 30-55 have more tendency of filling out the survey while travelling depicting that they travel more as compared to other age groups.

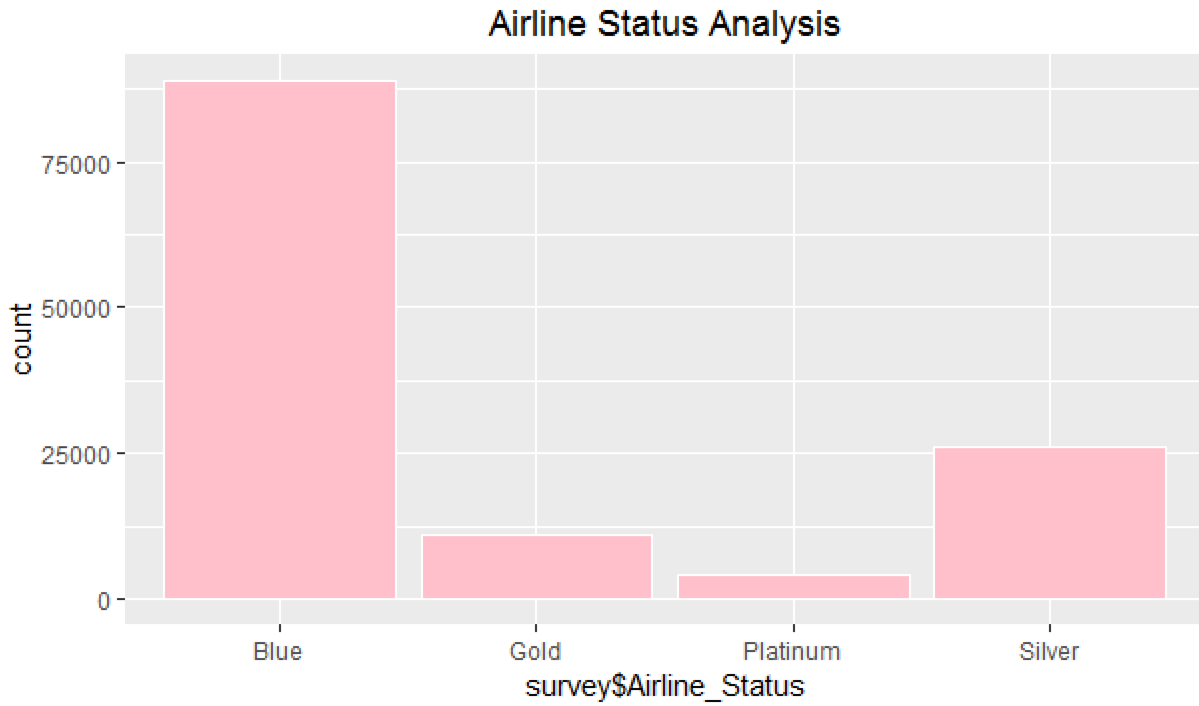


Gender analysis:

Based on the gender of the customers, it can be said that females have more significant contribution towards filling out customer satisfaction survey during their travel.

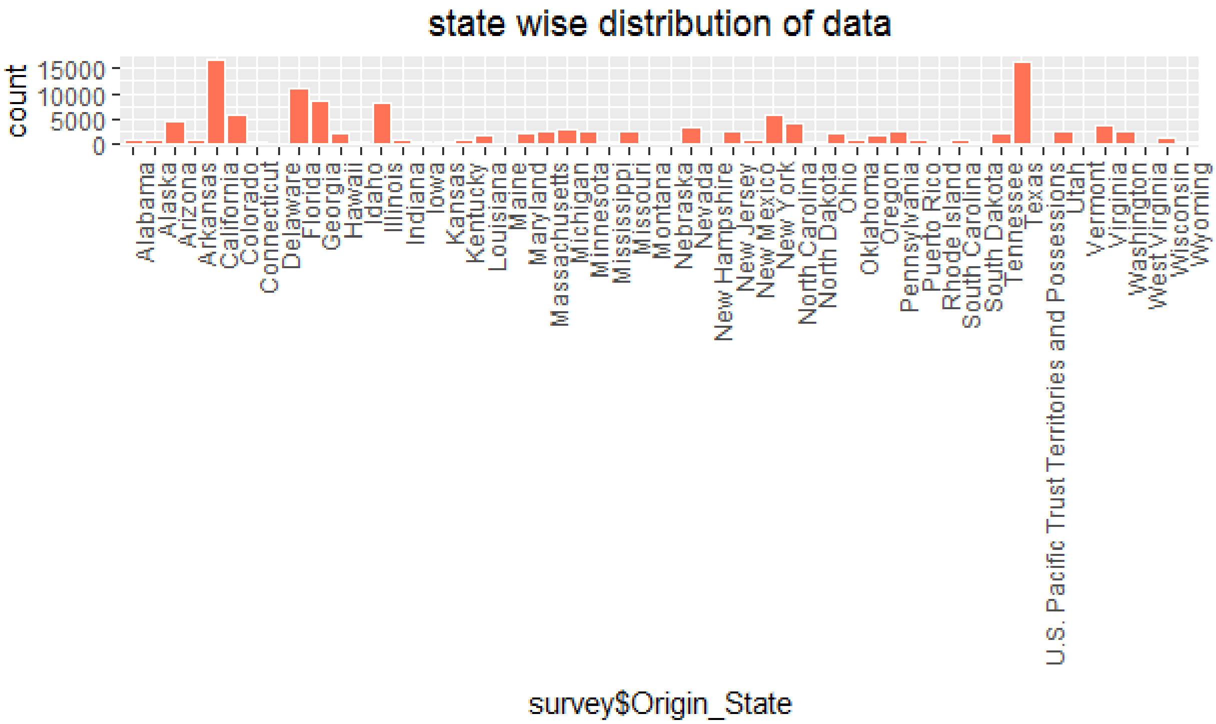


Airline Status analysis:



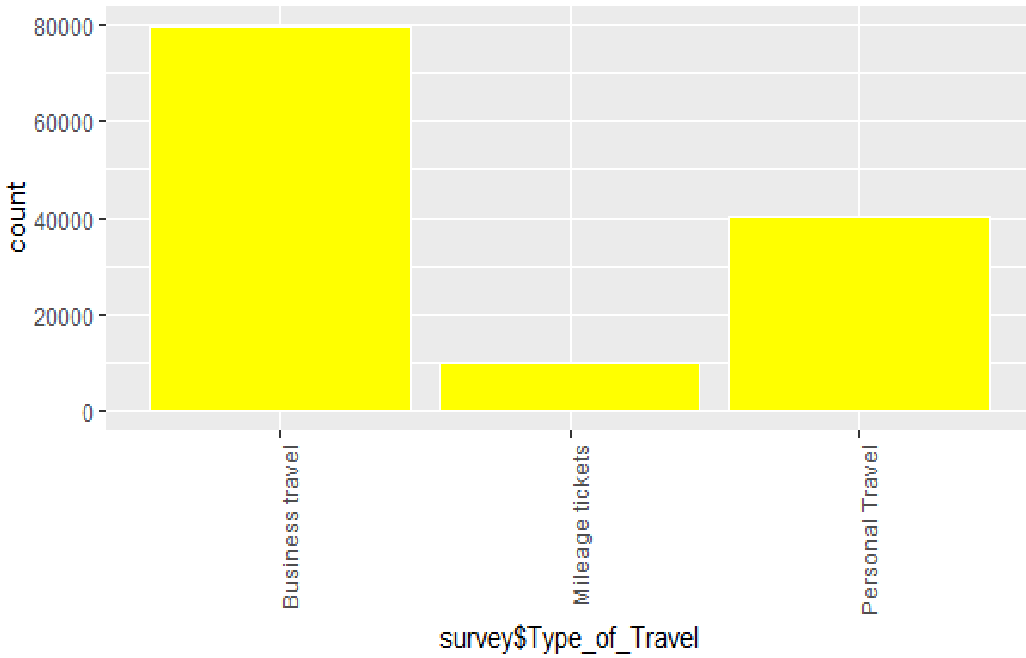
Origin state-wise analysis:

Based on the origin state of the customers analysis was done in the entire dataset and it can be seen that maximum number of customers have their origin state as Arkansas and Tennessee.



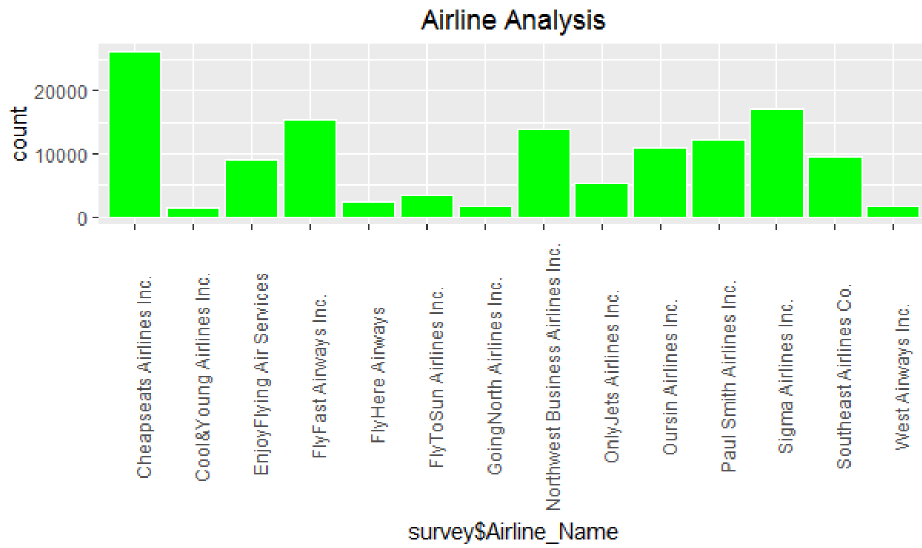
Type of Travel analysis:

Analysis of the type of travel in the entire dataset such as business travel, mileage tickets and personal travel that customers make were observed. It was seen that maximum number of customers fly for their business travel.

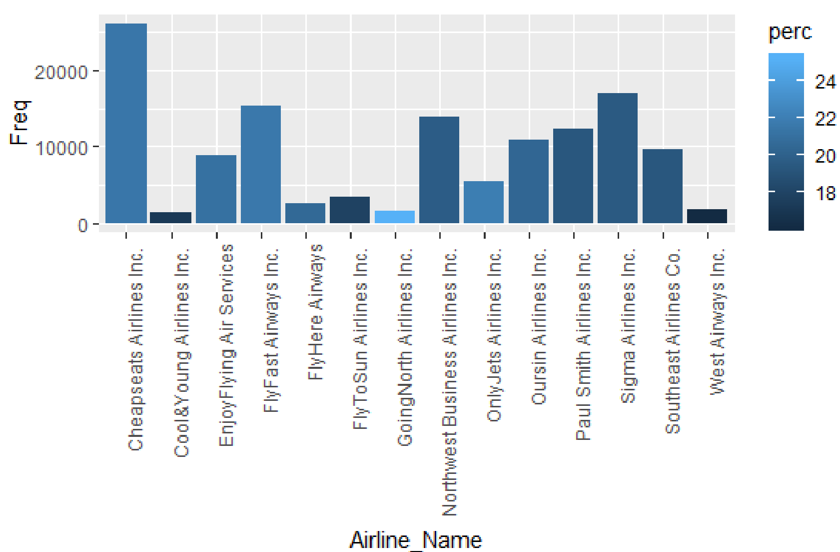


Airline analysis:

Analysis of the 14 different airlines in the entire dataset was made and below is the representation which depicts that Cheapseats airlines has the maximum number of data and Cool&Young airlines, GoingNorth Airlines, WestAirways have the minimum number of data in the entire dataset.

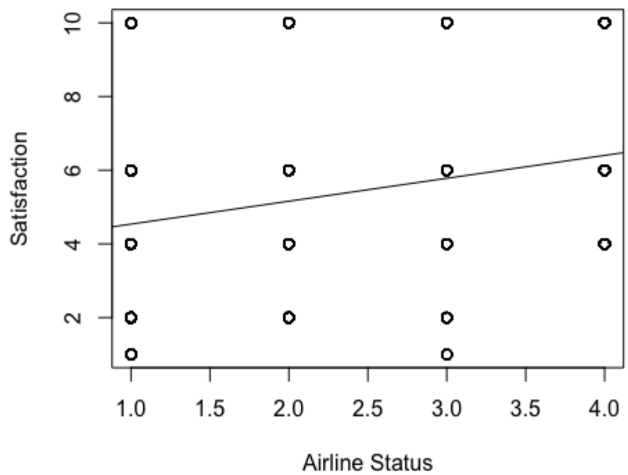


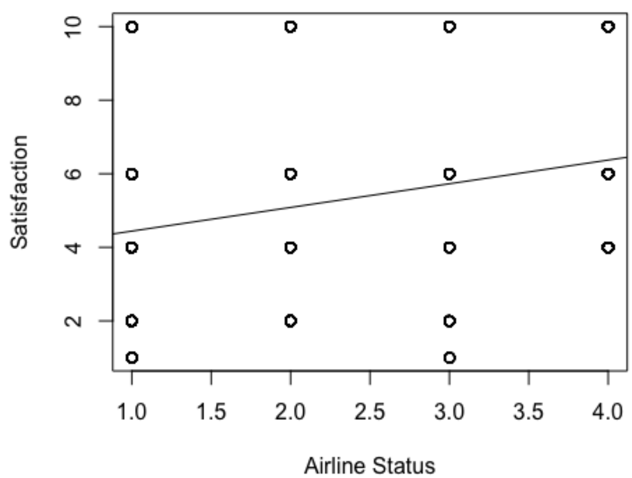
An analysis of the airlines for figuring out the airline with maximum number of unhappy customers was made. It was observed that the Flyfast airlines had a higher percentage of customers with low satisfaction. So Flyfast airways has been chosen for our analysis and has been compared with SouthEast airlines.



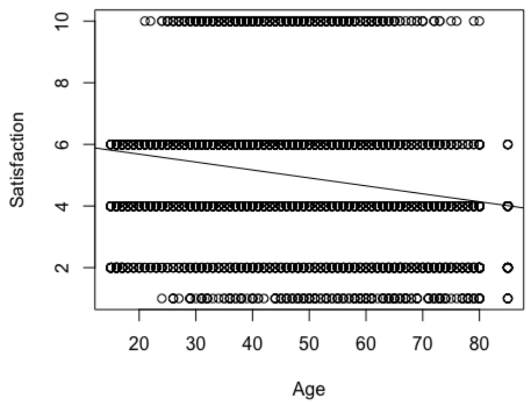
Linear Modelling techniques

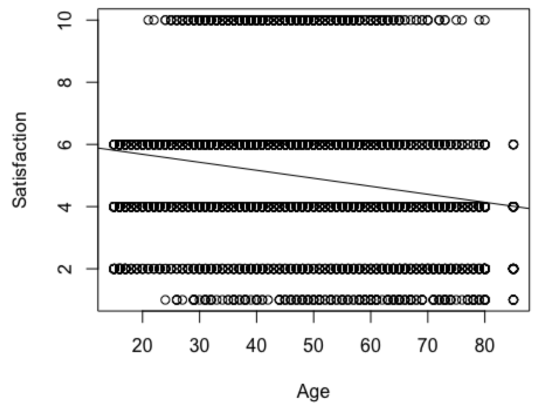
1. SatSatisfaction vs Airline Status



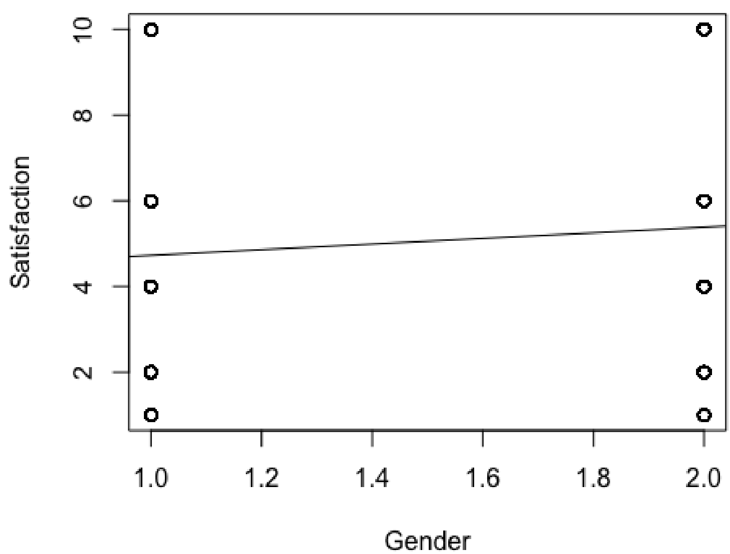


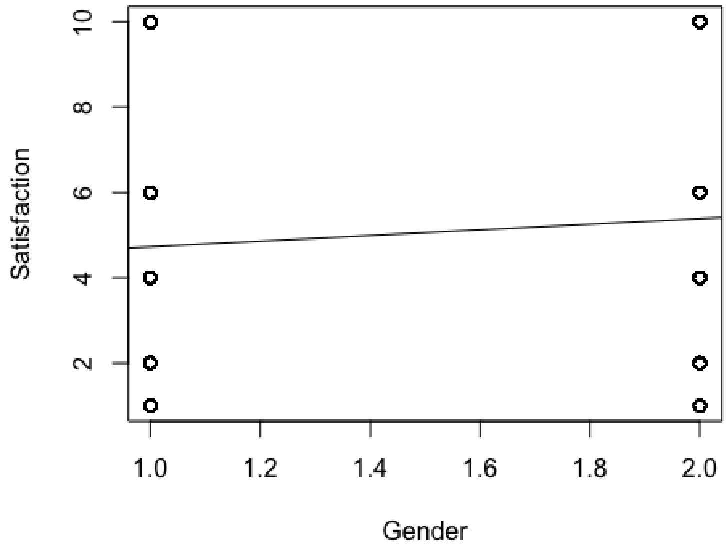
2. Satisfaction vs Age



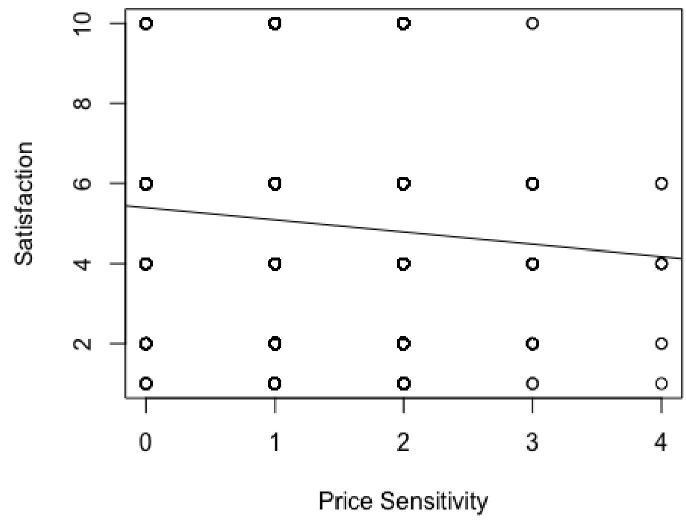


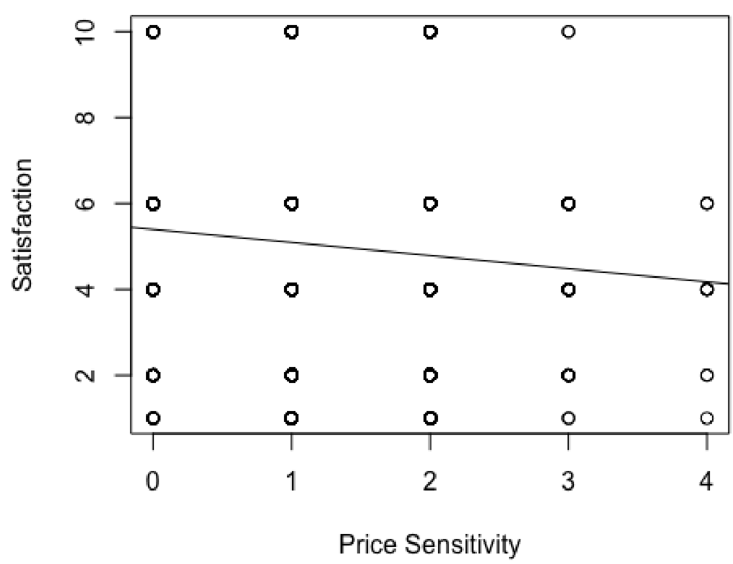
3. Satisfaction vs Gender



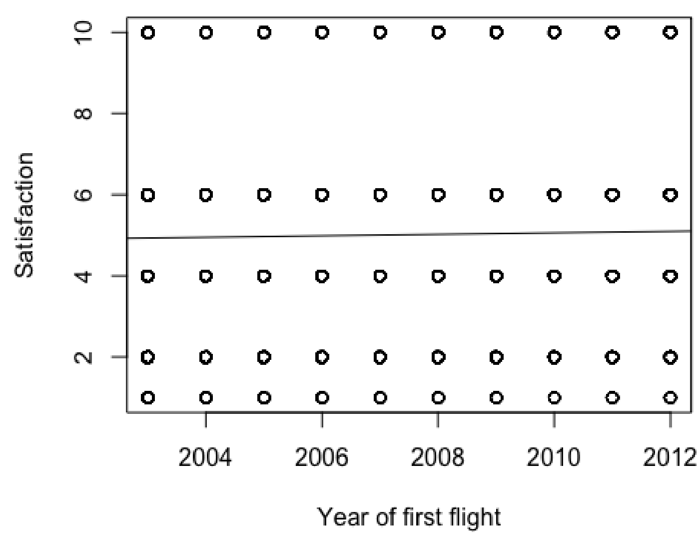


4. Satisfaction vs Price Sensitivity

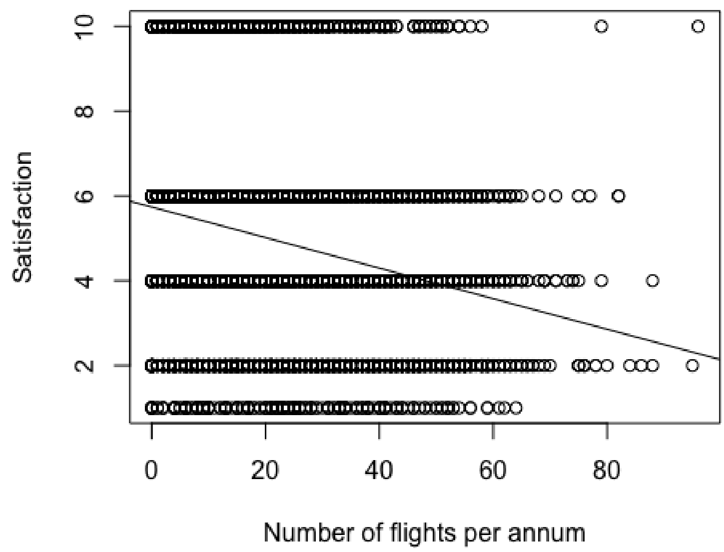


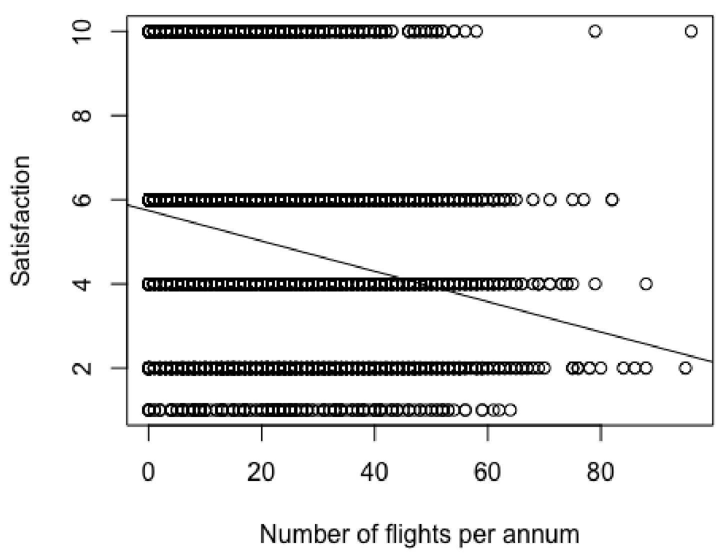


5. Satisfaction vs Year of first flight

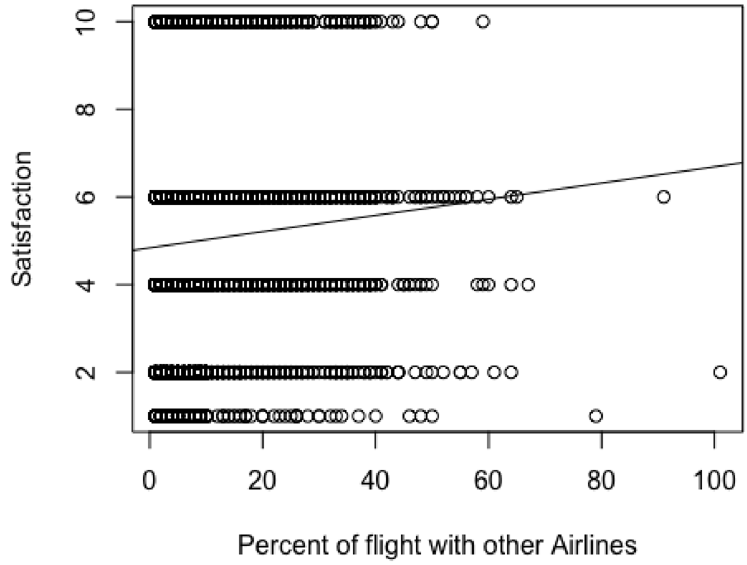


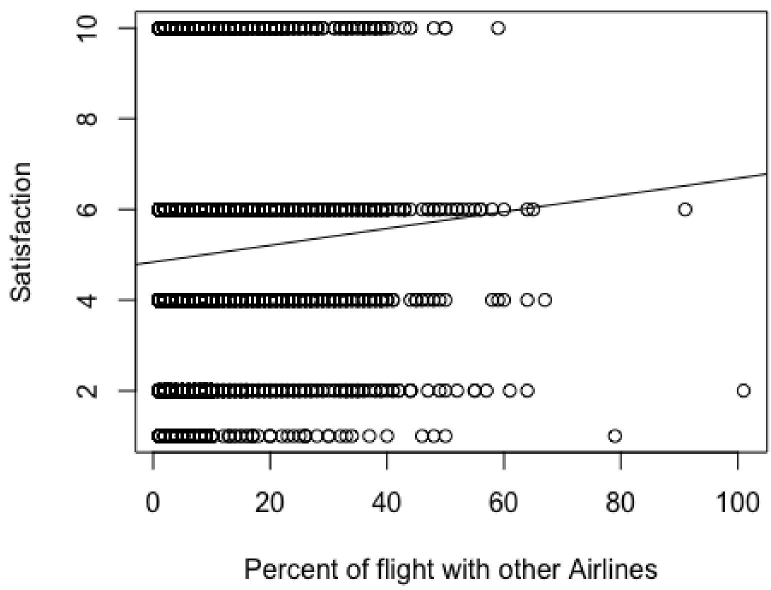
6. Satisfaction vs Number of flights per annum



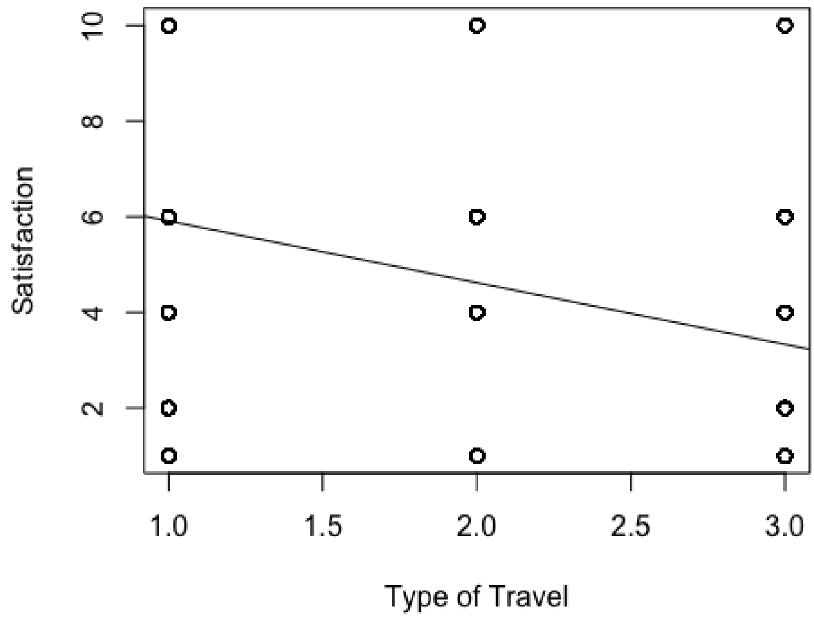


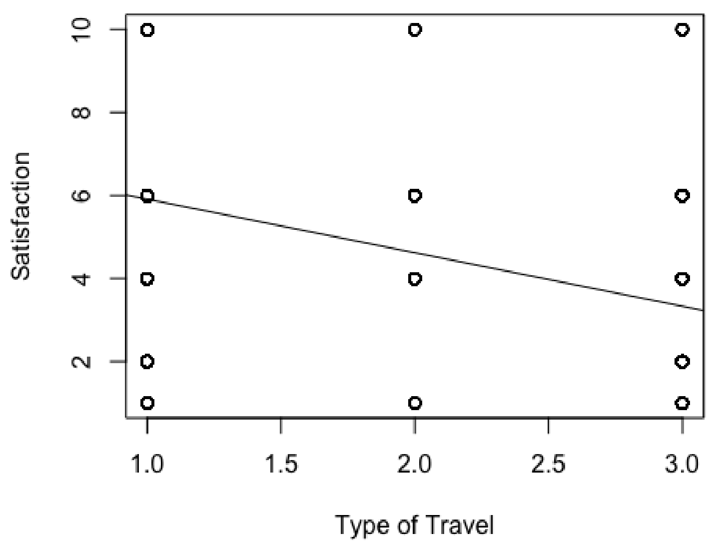
7. Satisfaction vs Percent of Flight with others



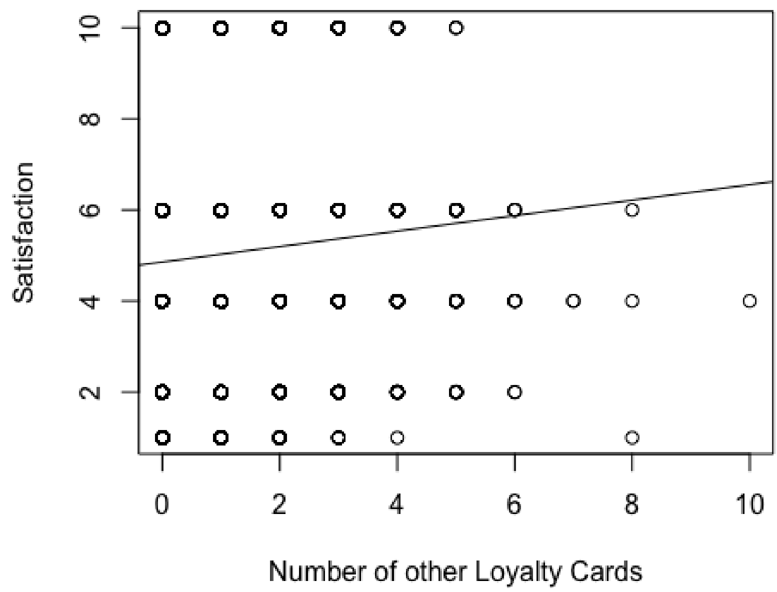


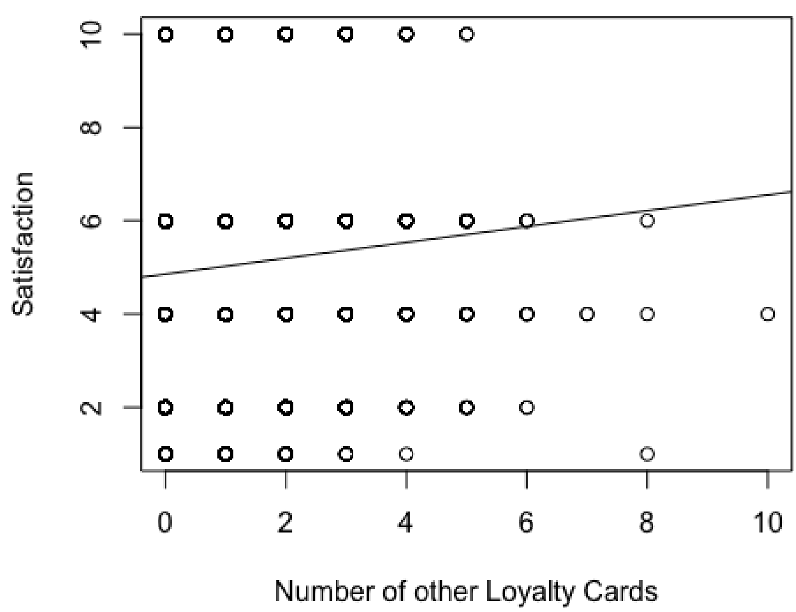
8. Satisfaction vs Type of travel



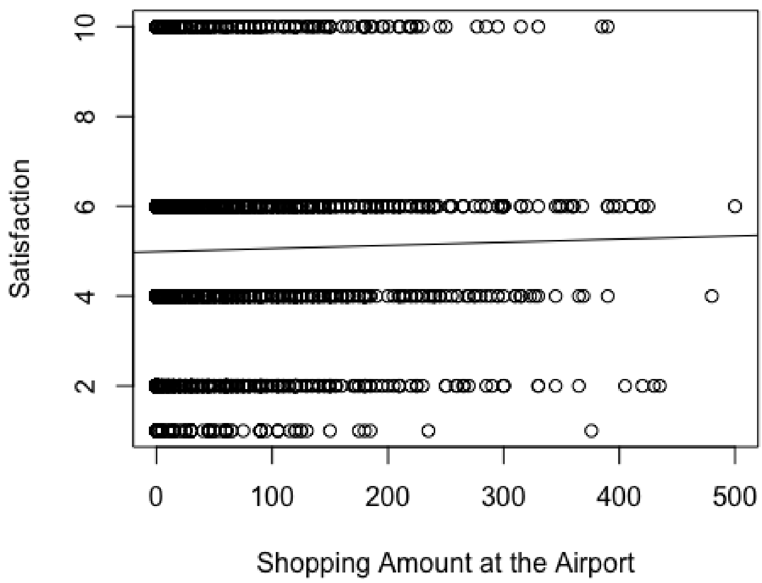


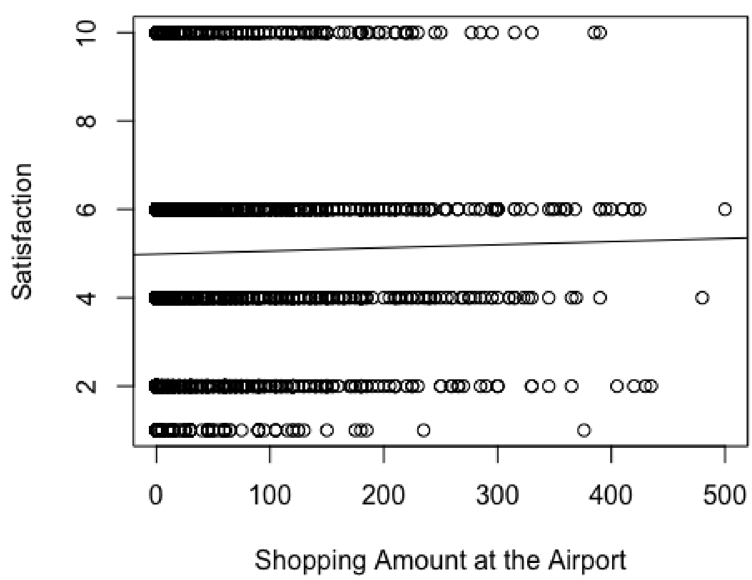
9. Satisfaction vs Number of other loyalty cards



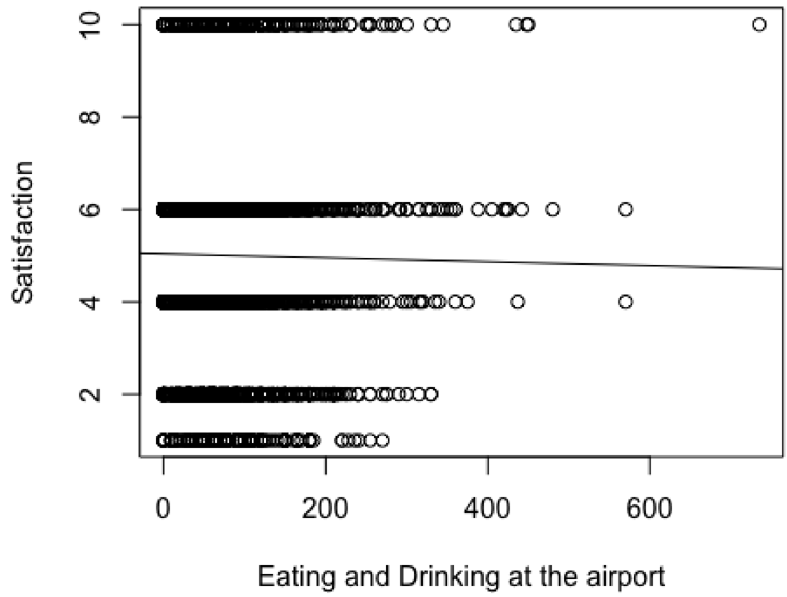


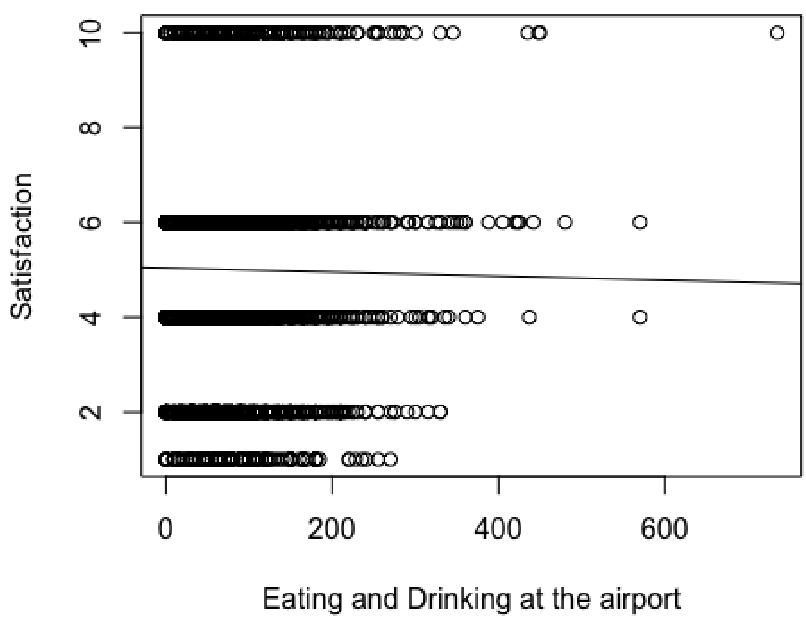
10. Satisfaction vs Shopping amount at airport



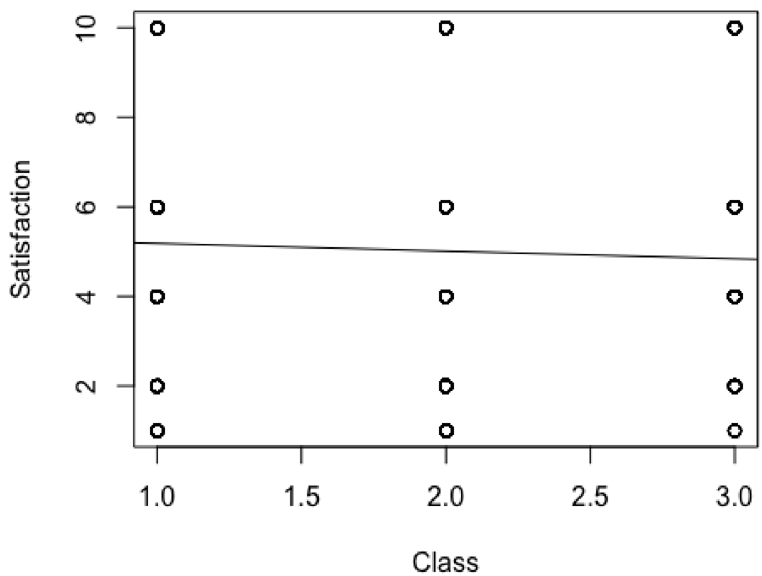


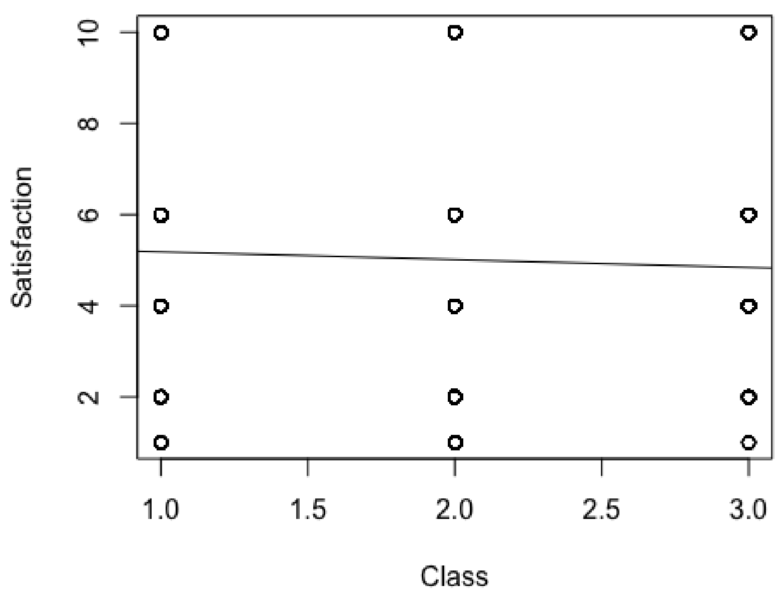
11. Satisfaction vs Eating & Drinking at airport


Satisfaction
10
8
6
4
2
0
200
400
600
Eating and Drinking at the airport

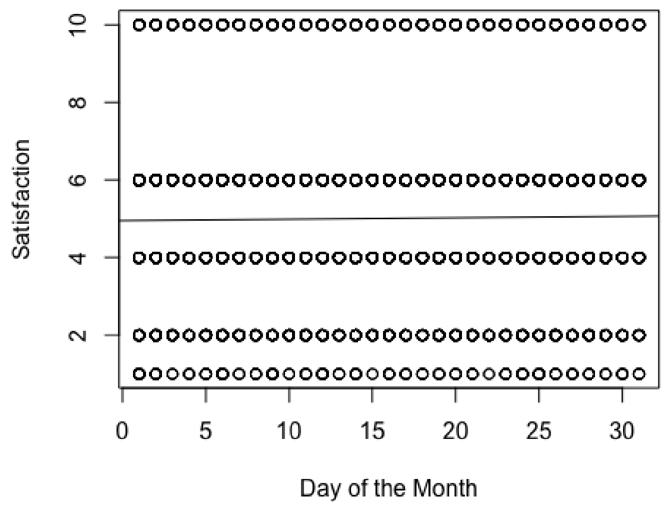

Satisfaction
10
8
6
4
2
0
200
400
600
Eating and Drinking at the airport

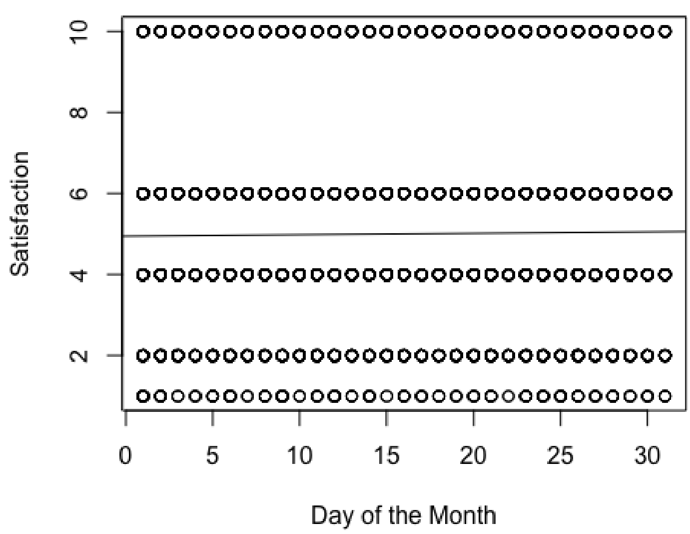
12. Satisfaction vs Class



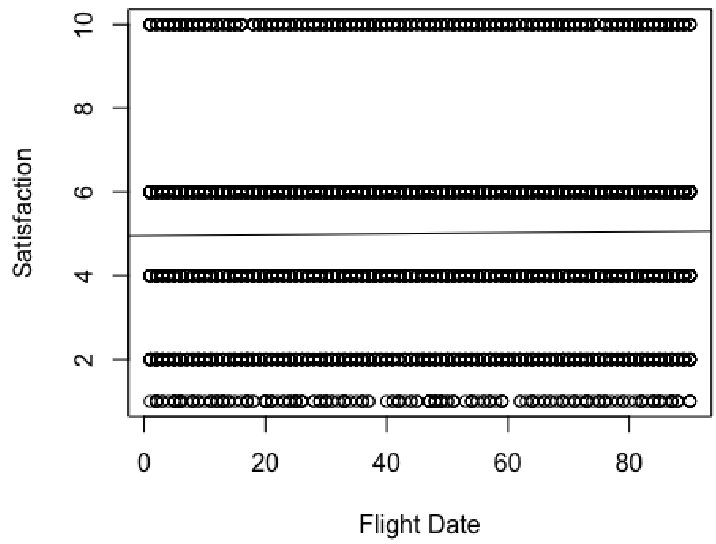


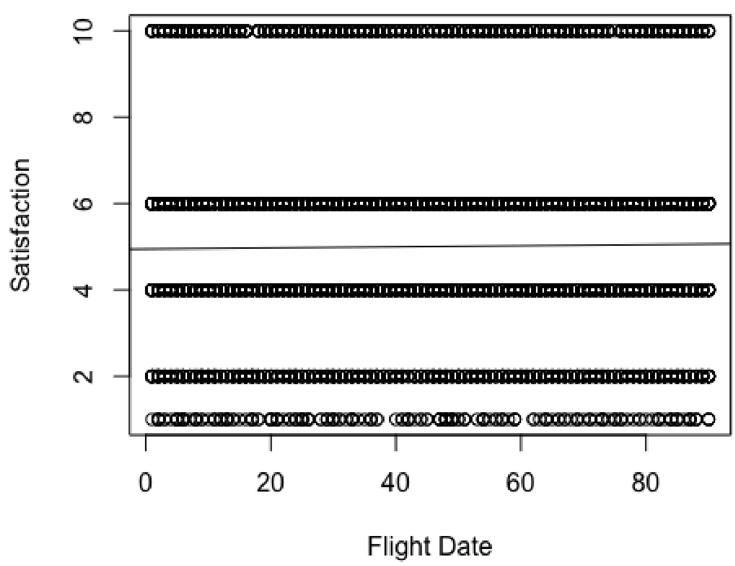
13. Satisfaction vs Day of the month



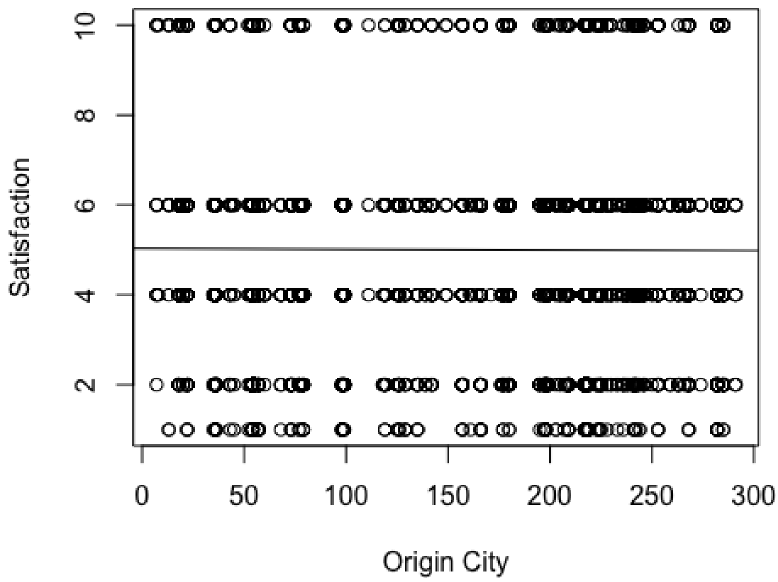


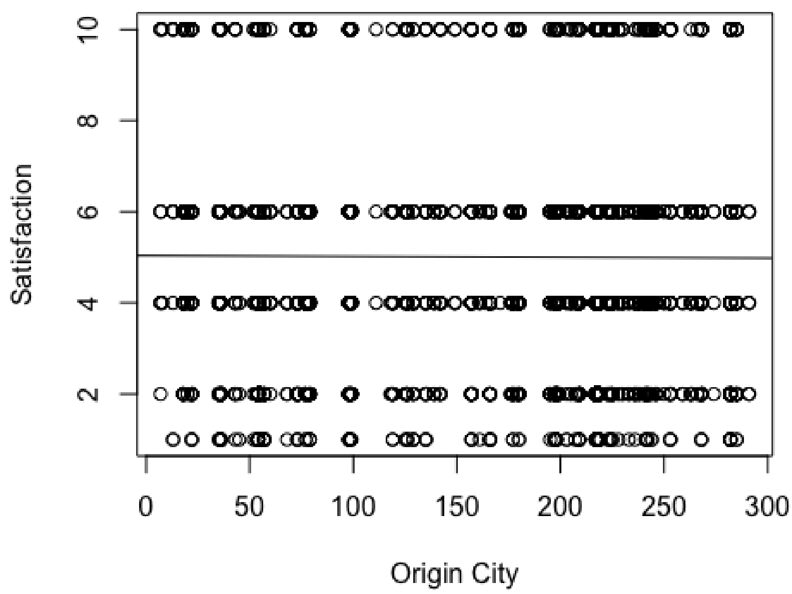
14. Satisfaction vs Flight Date



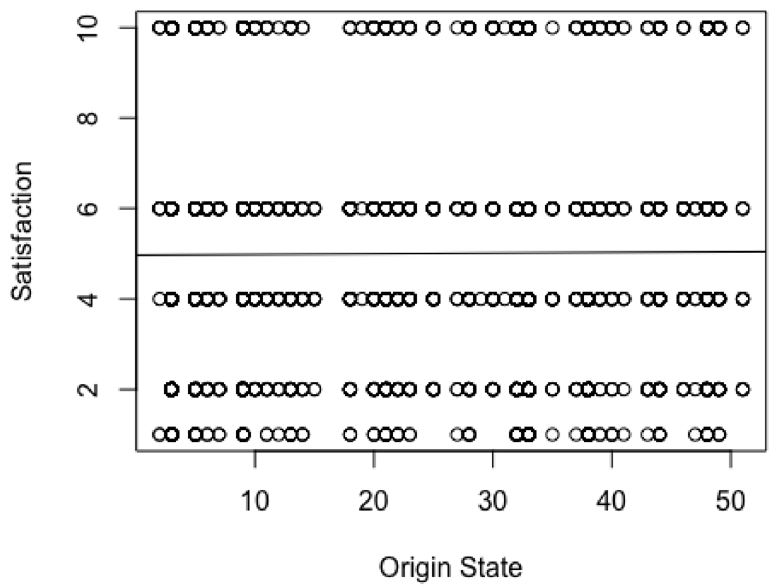


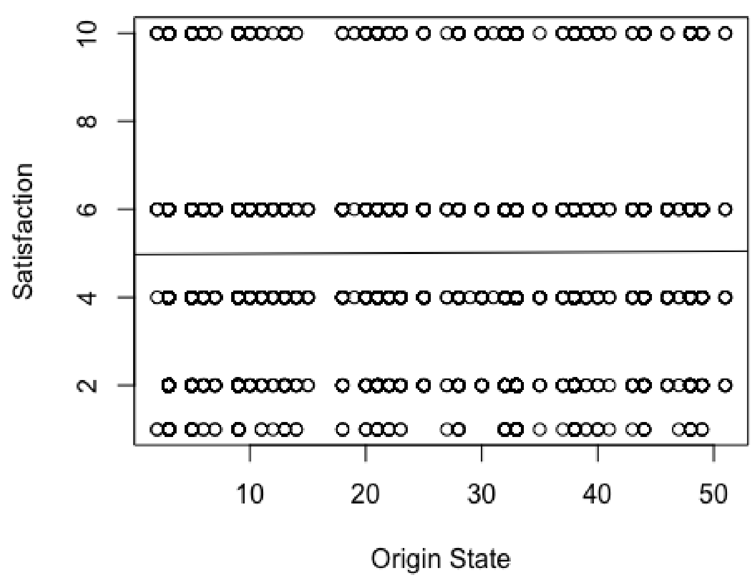
15. Satisfaction vs Origin City



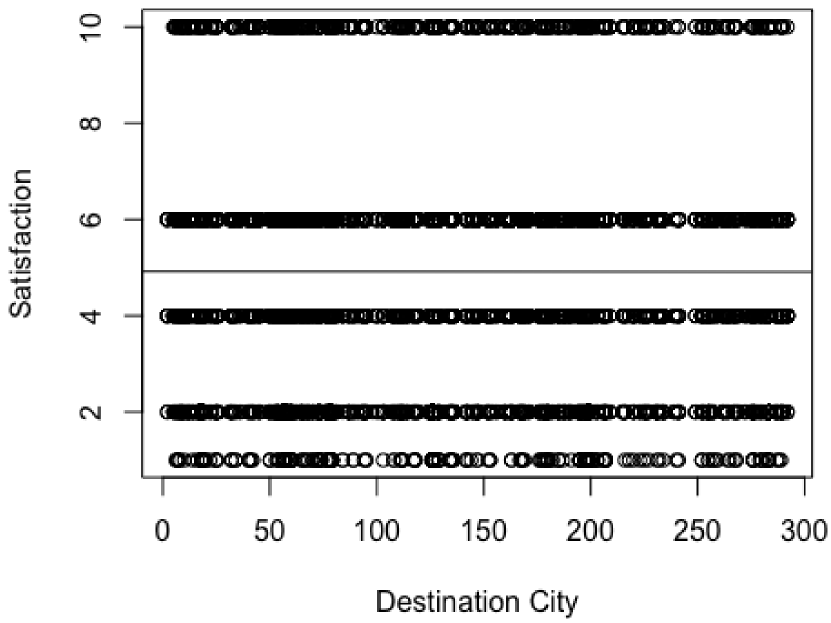


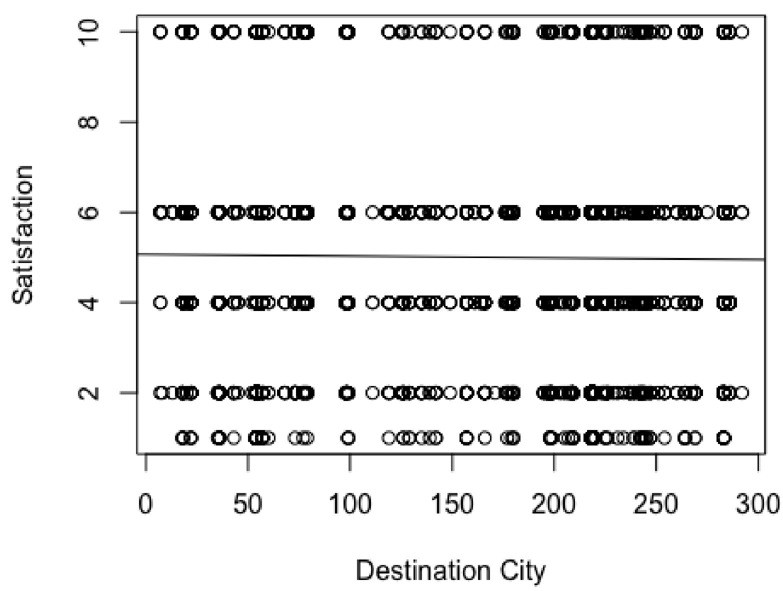
16. Satisfaction vs Origin State



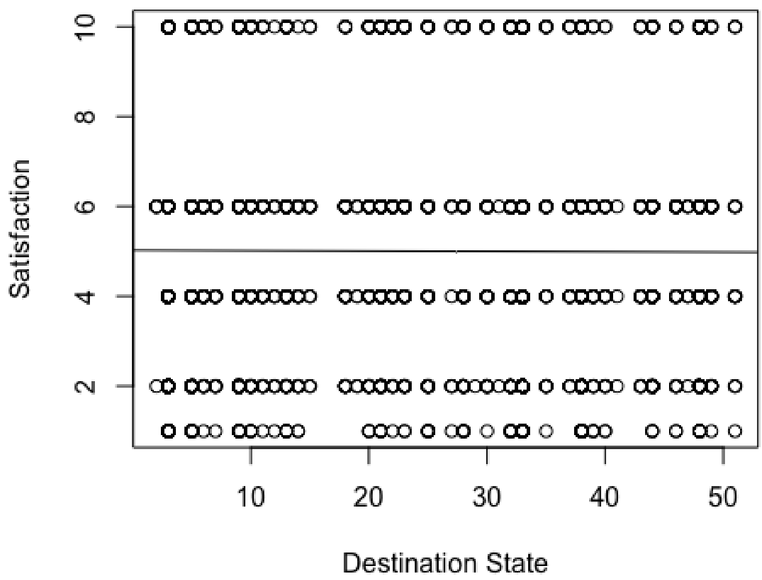


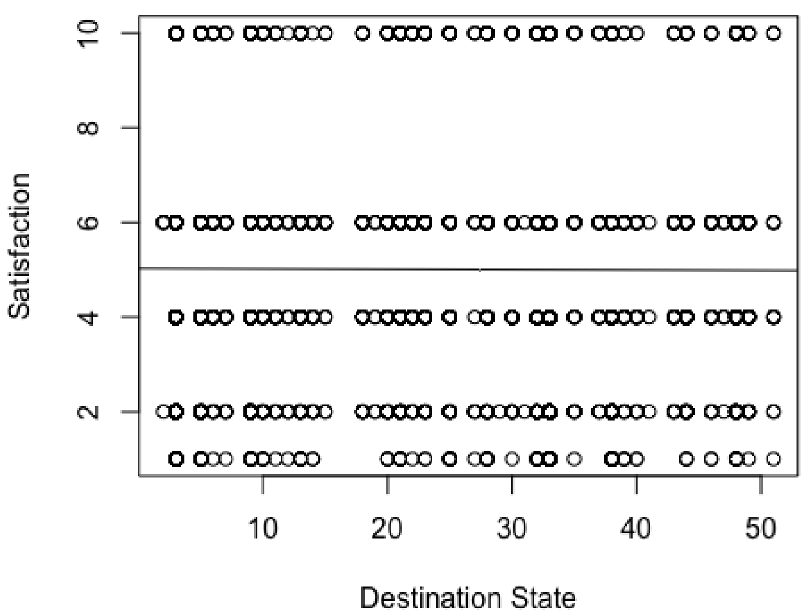
17. Satisfaction vs Destination City


Satisfaction
Destination City
10
8
6
4
2
0
50
100
150
200
250
300

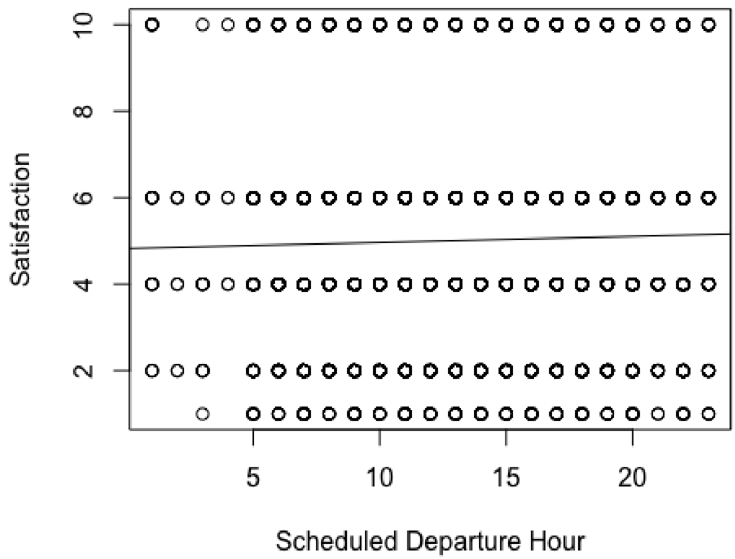

Satisfaction
Destination City
10
8
6
4
2
0
50
100
150
200
250
300

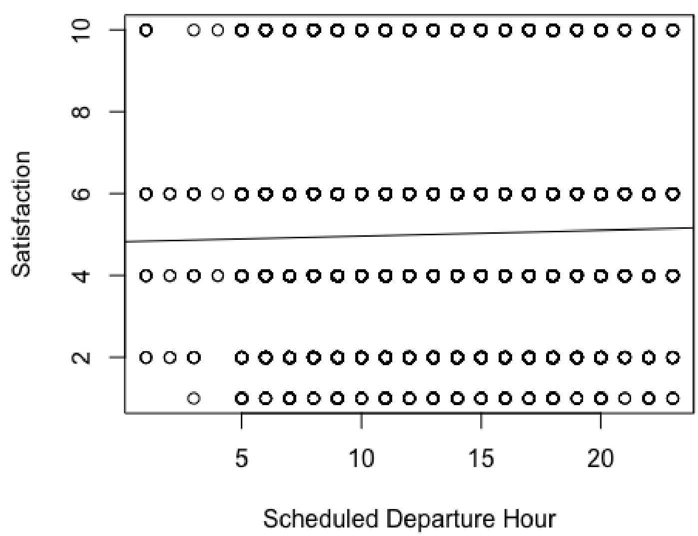
18. Satisfaction vs Destination State



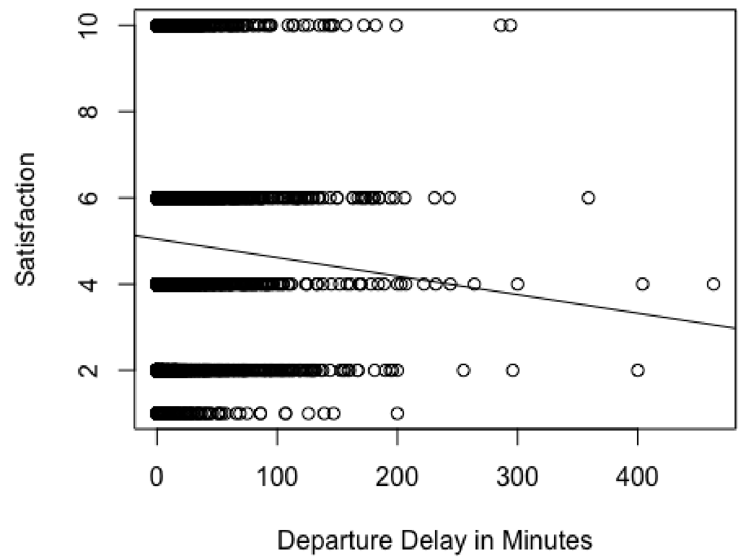


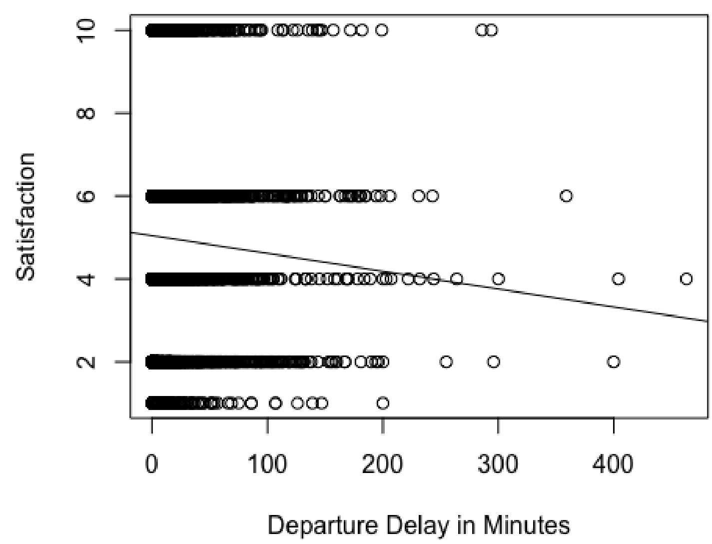
19. Satisfaction vs Scheduled departure Hour



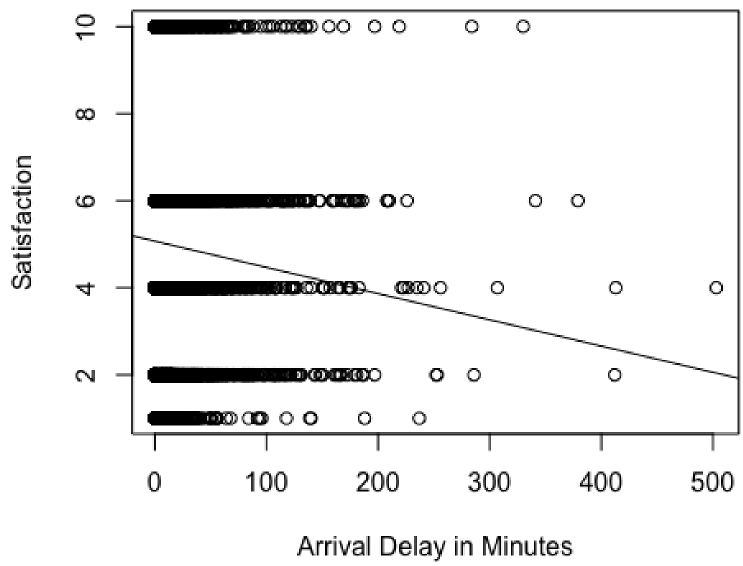


20. Satisfaction vs Departure delay in minutes

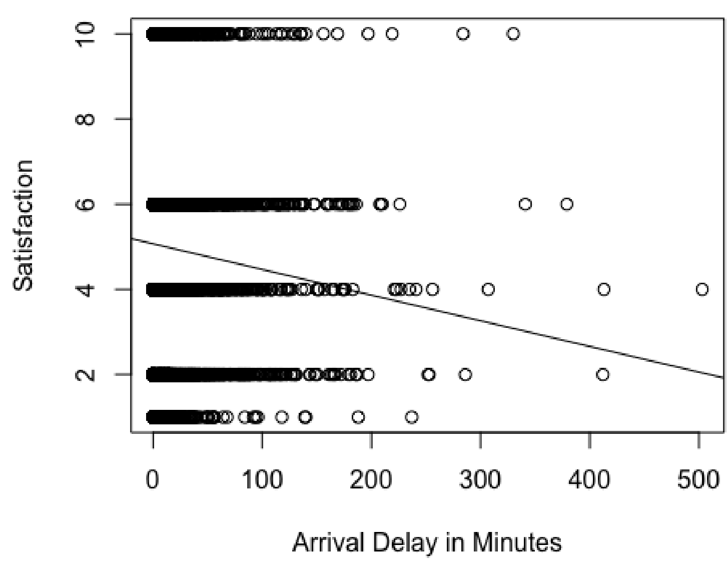




21. Satisfaction vs Arrival delay in minutes

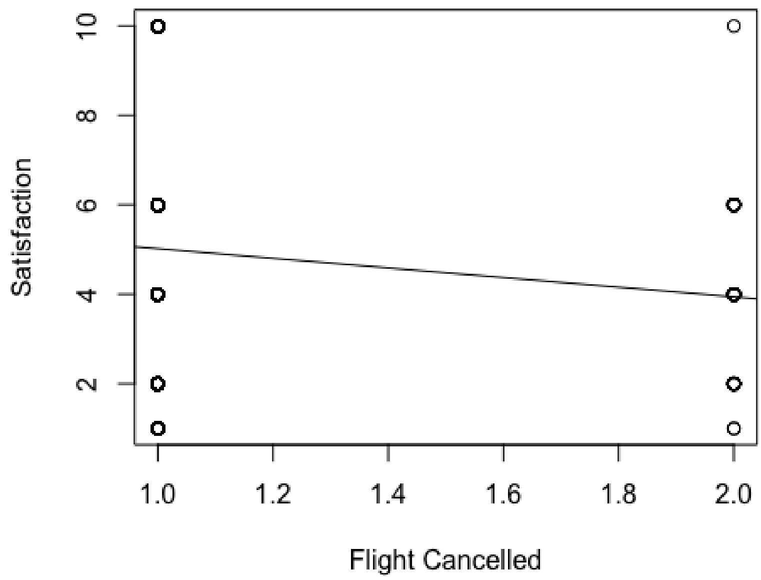


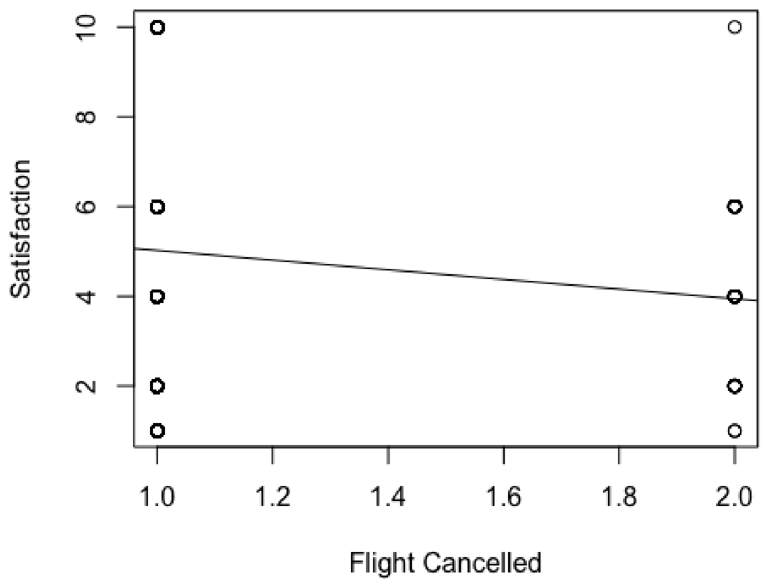
Satisfaction
Arrival Delay in Minutes



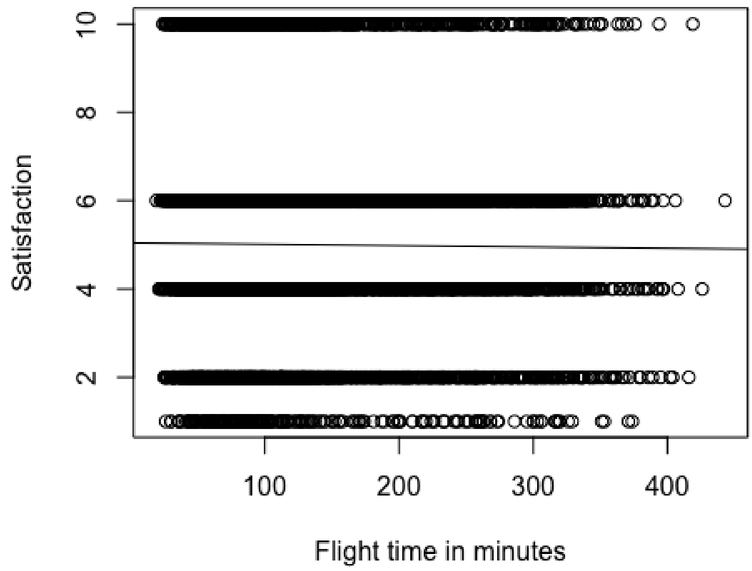
Satisfaction
Arrival Delay in Minutes

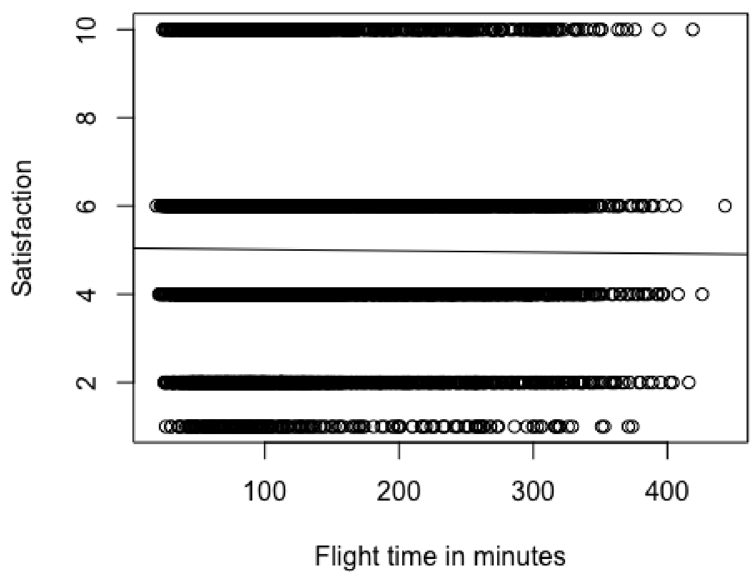
22. Satisfaction vs Flight Cancelled



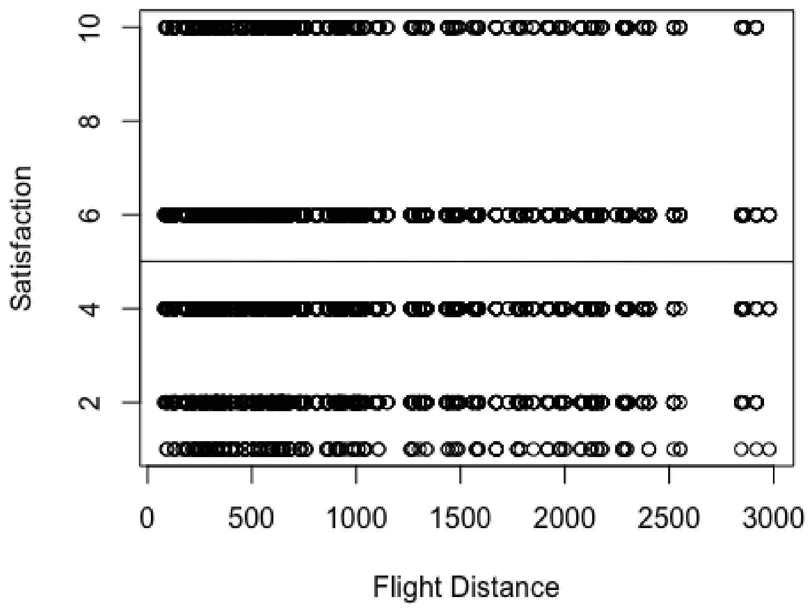


23. Satisfaction vs Flight time in minutes

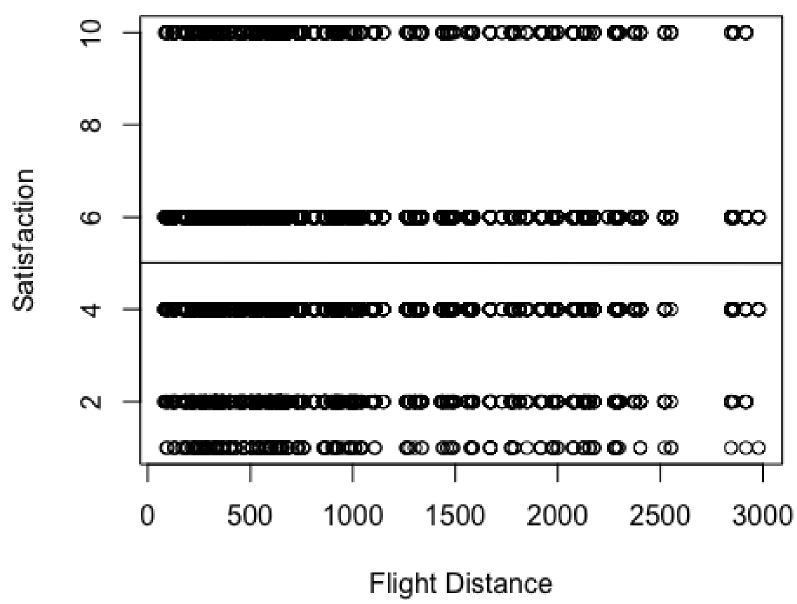




24. Satisfaction vs Flight Distance

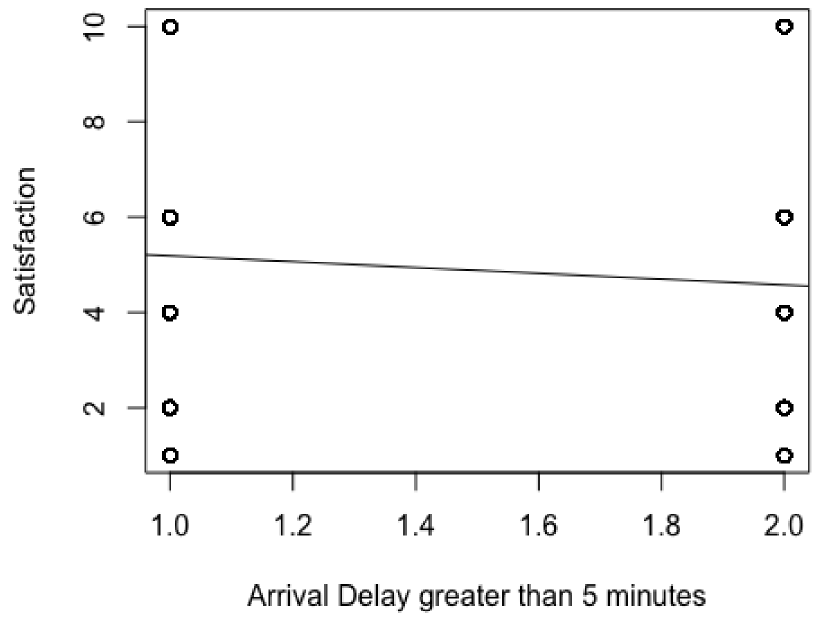


Satisfaction
10
8
6
4
2
0
500
1000
1500
2000
2500
3000
Flight Distance

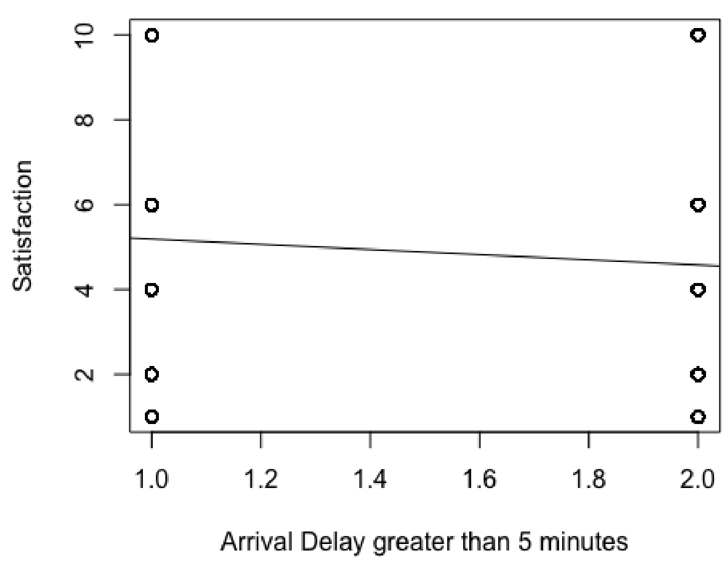


Satisfaction
10
8
6
4
2
0
500
1000
1500
2000
2500
3000
Flight Distance

25. ‘Satisfaction vs Arrival delay greater than 5 minute



s



Comparison of Multiple R squared values between Airline 1 and Airline 2 when any variable is plotted against Satisfaction:

S.r. no	Variable name	Multiple R squared value for A1	Multiple R squared value for A2
1	Airline Status	0.1131	0.04005
2	Age	0.04005	0.0258
3	Gender	0.02129	0.02129
4	Price Sensitivity	0.005561	0.00771
5	Year of first flight	0.0006154	0.0005808
6	No of flights per annum	0.05444	0.05704
7	Percent of flight with other Airlines	0.05444	0.005143
8	Type of Travel	0.2791	0.2908
9	No of other loyalty cards	0.007508	0.007344
10	Shopping amount at airport	0.0002682	0.0005364
11	Eating & Drinking at airport	0.0001009	0.0003544
12	Class	0.001133	0.002636
13	Day of month	0.0001969	8.479e-05
14	Flight Date	0.000228	0.003189
15	Airline Code	0.0002682	0.0002682
16	Airline Name	0.0002682	0.0002682
17	Origin City	4.631e-05	1.78e-05

18	Origin State	9.086e-05	4.872e-05
19	Destination City	0.0002352	1.994e-06
20	Destination State	4.08e-05	0.000124
21	Scheduled Departure Hour	0.0009554	7.699e-06
22	Departure Delay in minutes	0.002393	0.004668
23	Arrival Delay in minutes	0.005191	0.005868
24	Flight Cancelled	0.003189	0.003726
25	Flight time in minutes	0.0001149	0.0004713
26	Flight Distance	1.432e-06	5.785e-06
27	Arrival delay greater than 5 minutes	0.01559	0.0.01525

Comparison of Adjusted R squared values between Airline 1 and Airline 2 when any variable is plotted against Satisfaction:

S.r. no	Variable name	Adjusted R squared value for A1	Adjusted R squared value for A2
1	Airline Status	0.113	0.03999
2	Age	0.039995	0.02573
3	Gender	0.02119	0.02119
4	Price Sensitivity	0.005457	0.007646
5	Year of first flight	0.000511	0.0005159

6	No of flights per annum	0.004883	0.05698
7	Percent of flight with other Airlines	0.05434	0.005078
8	Type of Travel	0.004883	0.2907
9	No of other loyalty cards	0.279	0.00728
10	Shopping amount at airport	0.007405	0.0004715
11	Eating & Drinking at airport	0.0001638	0.0002895
12	Class	-3.532e-06	0.002571
13	Day of month	0.001029	1.988e-05
14	Flight Date	9.244e-05	0.003085
15	Airline Code	0.0001236	0.0001638
16	Airline Name	0.0001638	0.0001638
17	Origin City	0.0001638	-4.711e-05
18	Origin State	-5.812e-05	-1.619e-05
19	Destination City	-1.357e-05	-6.292e-05
20	Destination State	0.0001308	5.908e=05
21	Scheduled Departure Hour	-6.363e-05	-5.721e-05
22	Departure Delay in minutes	0.000851	0.004603
23	Arrival Delay in minutes	0.002289	0.005803
24	Flight Cancelled	0.0030850.005087	0.003663

25	Flight time in minutes	1.052e-05	0.0004064
26	Flight Distance	-0.000103	-5.913e-05
27	Arrival delay greater than 5 minutes	0.01549	0.01519

Multiple Regression :-

For South east Airline: -

We even perform multiple linear regression to check what combination of factors affect the customer satisfaction

Applying Linear modeling on the entire data set for south east airlines

Significant variables as per the analysis is Airline status, Age, Gender, year of first flight, Number of flights per anum, Type of Travel, Scheduled Departure Hour, Flight Cancelled, Arrival delay greater than 5 minutes

We further apply linear regression on the entire data set using the significant variables to check the how significant are these variables to the customer satisfaction.

Results obtained: -

Multiple R squared: -0.3804 Adjusted R squared: - 0.3798

We perform single variable reduction from the group of the significant variables to check how much the variable holds significance when the analysis is done using groups of variables

1)Without arrival delay greater than 5 minutes

Multiple R squared: -0.3622 Adjusted R squared: - 0.3616

2)Without Flight Cancelled

Multiple R squared: -0.3795 Adjusted R squared: - 0.3789

3)Without Type of Travel

Multiple R squared: -0.2158 Adjusted R squared: - 0.2152

4)Without Number of Flights Per anum

Multiple R squared: -0.3777 Adjusted R squared: - 0.3772

5)Without the Year of first flight

Multiple R squared: -0.3795 Adjusted R squared: - 0.379

6)Without Gender

Multiple R squared: -0.3725 Adjusted R squared: - 0.372

7)Without age

Multiple R squared: -0.3789 Adjusted R squared: - 0.3783

8)Without airline status

Multiple R squared: -0.3149 Adjusted R squared: - 0.3143

9)Without Scheduled Departure hour

Multiple R squared: -0.3795 Adjusted R squared: - 0.3789

By comparing all the adjusted R squared values the best results are obtained when the model is without the Year of first flight. So applying second level of multiple regression

1)Without Arrival greater than 5 minutes

Multiple R squared: -0.3612 Adjusted R squared: - 0.3608

2)Without Flight Cancelled

Multiple R squared: -0.3784 Adjusted R squared: - 0.378

3)Without Scheduled departure hour

Multiple R squared: -0.3786 Adjusted R squared: - 0.3781

4)Without type of travel

Multiple R squared: -0.2153 Adjusted R squared: - 0.2148

5)Without Number of flights per anum

Multiple R squared: -0.3769 Adjusted R squared: - 0.3764

6)Without Gender

Multiple R squared: -0.3716 Adjusted R squared: - 0.3711

7)Without age

Multiple R squared: -0.378 Adjusted R squared: - 0.3776

8)Without airline status

Multiple R squared: -0.3132 Adjusted R squared: - 0.3127

Comparatively the model without the Scheduled Departure hour as per the linear analysis is the best model to determine the customer satisfaction

Similarly, by applying the next level of regression, the best model is with variables Airline Status, Gender, Number of flights per anum, Type of Travel, Arrival Greater than 5 mins.

So the multiple regression analysis for South East airlines gives us the results stating Airline Status, Gender, Number of flights per anum, Type of Travel, Arrival Greater than 5 mins are the variables that affect the customer satisfaction and also those factors that are important in improving

For Far East Airlines :-

Association Rules Mining:

In order to analyze the relationships between the attributes in the customer survey dataset of the airlines and filter out the attributes which play an important role in predicting low customer satisfaction, association rules mining has been used for both the SouthEast and FlyFast airlines. Attributes were categorised into buckets of High, Average and low values and converted to factors.

Adjusting the values of support and confidence, a set of 28 rules were obtained for SouthEast airlines and Flyfast airlines. Given below is the set of rules arranged in the decreasing order based on the life values:

SouthEast airlines:

```
     lhs                                         rhs                  support confidence     lift count
[1]  {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      Class=Eco,
      Flight_cancelled=No}                    => {Satisfaction=Low} 0.1099509  0.5673491 2.927534  1053
[2]  {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      Class=Eco,
      Airline_Name=Southeast Airlines Co. ,
      Flight_cancelled=No}                    => {Satisfaction=Low} 0.1099509  0.5673491 2.927534  1053
[3]  {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      Class=Eco,
      Airline_Code=US,
      Flight_cancelled=No}                    => {Satisfaction=Low} 0.1099509  0.5673491 2.927534  1053
[4]  {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      Class=Eco,
      Airline_Code=US,
      Airline_Name=Southeast Airlines Co. ,
      Flight_cancelled=No}                    => {Satisfaction=Low} 0.1099509  0.5673491 2.927534  1053
[5]  {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      Class=Eco}                              => {Satisfaction=Low} 0.1124569  0.5650577 2.915710  1077
[6]  {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      Class=Eco,
      Airline_Name=Southeast Airlines Co. } => {Satisfaction=Low} 0.1124569  0.5650577 2.915710  1077
[7]  {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      Class=Eco,
      Airline_Code=US}                        => {Satisfaction=Low} 0.1124569  0.5650577 2.915710  1077
[8]  {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      Class=Eco,
      Airline_Code=US,
      Airline_Name=Southeast Airlines Co. } => {Satisfaction=Low} 0.1124569  0.5650577 2.915710  1077
[9]  {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      Flight_cancelled=No}                    => {Satisfaction=Low} 0.1339668  0.5637083 2.908747  1283
```

```
[10] {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      Airline_Name=Southeast Airlines Co. ,
      Flight_cancelled=No}                    => {Satisfaction=Low} 0.1339668  0.5637083 2.908747  1283
[11] {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      Airline_Code=US,
      Flight_cancelled=No}                    => {Satisfaction=Low} 0.1339668  0.5637083 2.908747  1283
[12] {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      Airline_Code=US,
      Airline_Name=Southeast Airlines Co. ,
      Flight_cancelled=No}                    => {Satisfaction=Low} 0.1339668  0.5637083 2.908747  1283
[13] {Airline_Status=Blue,
      Type_of_Travel=Personal Travel}         => {Satisfaction=Low} 0.1367860  0.5619906 2.899883  1310
[14] {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      Airline_Name=Southeast Airlines Co. } => {Satisfaction=Low} 0.1367860  0.5619906 2.899883  1310
[15] {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      Airline_Code=US}                        => {Satisfaction=Low} 0.1367860  0.5619906 2.899883  1310
[16] {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      Airline_Code=US,
      Airline_Name=Southeast Airlines Co. } => {Satisfaction=Low} 0.1367860  0.5619906 2.899883  1310
[17] {Age=High,
      Type_of_Travel=Personal Travel}         => {Satisfaction=Low} 0.1008667  0.5176849 2.671265   966
[18] {Age=High,
      Type_of_Travel=Personal Travel,
      Airline_Name=Southeast Airlines Co. } => {Satisfaction=Low} 0.1008667  0.5176849 2.671265   966
[19] {Age=High,
      Type_of_Travel=Personal Travel,
      Airline_Code=US}                        => {Satisfaction=Low} 0.1008667  0.5176849 2.671265   966
[20] {Age=High,
      Type_of_Travel=Personal Travel,
      Airline_Code=US,
      Airline_Name=Southeast Airlines Co. } => {Satisfaction=Low} 0.1008667  0.5176849 2.671265   966
[21] {Type_of_Travel=Personal Travel,
      No_of_other_Loyalty_Cards=Low,
      Flight_cancelled=No}                    => {Satisfaction=Low} 0.1017020  0.5078206 2.620365   974
[22] {Type_of_Travel=Personal Travel,
      No_of_other_Loyalty_Cards=Low,
      Airline_Name=Southeast Airlines Co. ,
      Flight_cancelled=No}                    => {Satisfaction=Low} 0.1017020  0.5078206 2.620365   974
[23] {Type_of_Travel=Personal Travel,
      No_of_other_Loyalty_Cards=Low,
      Airline_Code=US,
      Flight_cancelled=No}                    => {Satisfaction=Low} 0.1017020  0.5078206 2.620365   974
[24] {Type_of_Travel=Personal Travel,
      No_of_other_Loyalty_Cards=Low,
      Airline_Code=US,
      Airline_Name=Southeast Airlines Co. ,
      Flight_cancelled=No}                    => {Satisfaction=Low} 0.1017020  0.5078206 2.620365   974
[25] {Type_of_Travel=Personal Travel,
      No_of_other_Loyalty_Cards=Low}          => {Satisfaction=Low} 0.1039992  0.5071283 2.616793   996
[26] {Type_of_Travel=Personal Travel,
      No_of_other_Loyalty_Cards=Low,
      Airline_Name=Southeast Airlines Co. } => {Satisfaction=Low} 0.1039992  0.5071283 2.616793   996
[27] {Type_of_Travel=Personal Travel,
      No_of_other_Loyalty_Cards=Low,
      Airline_Code=US}                        => {Satisfaction=Low} 0.1039992  0.5071283 2.616793   996
[28] {Type_of_Travel=Personal Travel,
      No_of_other_Loyalty_Cards=Low,
      Airline_Code=US,
      Airline_Name=Southeast Airlines Co. } => {Satisfaction=Low} 0.1039992  0.5071283 2.616793   996
```

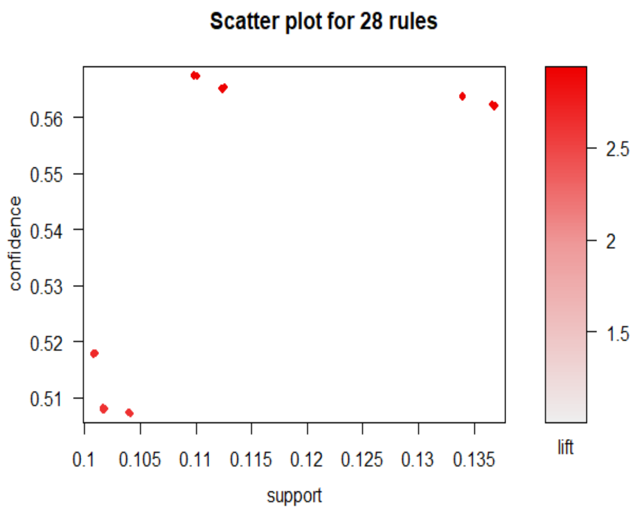
Flyfast airways:

```
    lhs                                     rhs                  support confidence     lift count
[1] {Airline_Status=Blue,
     Gender=Female,
     Type_of_Travel=Personal Travel,
     Flight_cancelled=No}                => {Satisfaction=Low} 0.1015772  0.6597808 3.048049  1565
[2] {Airline_Status=Blue,
     Gender=Female,
     Type_of_Travel=Personal Travel,
     Airline_Code=EV,
     Flight_cancelled=No}                => {Satisfaction=Low} 0.1015772  0.6597808 3.048049  1565
[3] {Airline_Status=Blue,
     Gender=Female,
     Type_of_Travel=Personal Travel,
     Airline_Name=FlyFast Airways Inc. ,
     Flight_cancelled=No}                => {Satisfaction=Low} 0.1015772  0.6597808 3.048049  1565
[4] {Airline_Status=Blue,
     Gender=Female,
     Type_of_Travel=Personal Travel,
     Airline_Code=EV,
     Airline_Name=FlyFast Airways Inc. ,
     Flight_cancelled=No}                => {Satisfaction=Low} 0.1015772  0.6597808 3.048049  1565
[5] {Airline_Status=Blue,
     Gender=Female,
     Type_of_Travel=Personal Travel}     => {Satisfaction=Low} 0.1067697  0.6484036 2.995489  1645
[6] {Airline_Status=Blue,
     Gender=Female,
     Type_of_Travel=Personal Travel,
     Airline_Code=EV}                    => {Satisfaction=Low} 0.1067697  0.6484036 2.995489  1645
[7] {Airline_Status=Blue,
     Gender=Female,
     Type_of_Travel=Personal Travel,
     Airline_Name=FlyFast Airways Inc. } => {Satisfaction=Low} 0.1067697  0.6484036 2.995489  1645
[8] {Airline_Status=Blue,
     Gender=Female,
     Type_of_Travel=Personal Travel,
     Airline_Code=EV,
     Airline_Name=FlyFast Airways Inc. } => {Satisfaction=Low} 0.1067697  0.6484036 2.995489  1645
[9] {Airline_Status=Blue,
     Type_of_Travel=Personal Travel,
     Class=Eco,
     Flight_cancelled=No}                => {Satisfaction=Low} 0.1211787  0.6473648 2.990689  1867
```

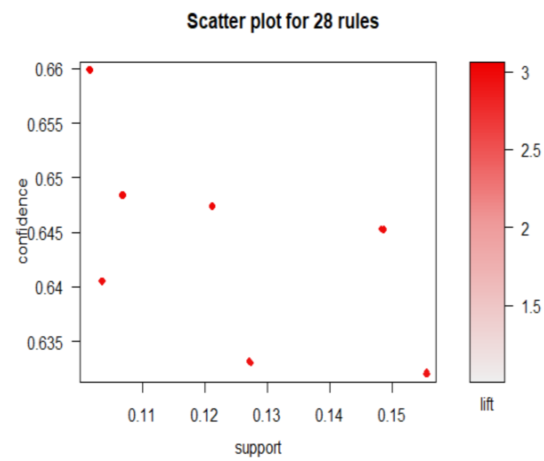
```
[10] {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      Class=Eco,
      Airline_Code=EV,
      Flight_cancelled=No}                 => {Satisfaction=Low} 0.1211787  0.6473648 2.990689  1867
[11] {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      Class=Eco,
      Airline_Name=FlyFast Airways Inc. ,
      Flight_cancelled=No}                 => {Satisfaction=Low} 0.1211787  0.6473648 2.990689  1867
[12] {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      Class=Eco,
      Airline_Code=EV,
      Airline_Name=FlyFast Airways Inc. ,
      Flight_cancelled=No}                 => {Satisfaction=Low} 0.1211787  0.6473648 2.990689  1867
[13] {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      Flight_cancelled=No}                 => {Satisfaction=Low} 0.1485039  0.6452341 2.980846  2288
[14] {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      Airline_Code=EV,
      Flight_cancelled=No}                 => {Satisfaction=Low} 0.1485039  0.6452341 2.980846  2288
[15] {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      Airline_Name=FlyFast Airways Inc. ,
      Flight_cancelled=No}                 => {Satisfaction=Low} 0.1485039  0.6452341 2.980846  2288
[16] {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      Airline_Code=EV,
      Airline_Name=FlyFast Airways Inc. ,
      Flight_cancelled=No}                 => {Satisfaction=Low} 0.1485039  0.6452341 2.980846  2288
[17] {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      No_of_other_Loyalty_Cards=Low}       => {Satisfaction=Low} 0.1035893  0.6404494 2.958742  1596
[18] {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      No_of_other_Loyalty_Cards=Low,
      Airline_Code=EV}                     => {Satisfaction=Low} 0.1035893  0.6404494 2.958742  1596
[19] {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      No_of_other_Loyalty_Cards=Low,
      Airline_Name=FlyFast Airways Inc. } => {Satisfaction=Low} 0.1035893  0.6404494 2.958742  1596
[20] {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      No_of_other_Loyalty_Cards=Low,
      Airline_Code=EV,
      Airline_Name=FlyFast Airways Inc. } => {Satisfaction=Low} 0.1035893  0.6404494 2.958742  1596
[21] {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      Class=Eco}                           => {Satisfaction=Low} 0.1272149  0.6330749 2.924673  1960
[22] {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      Class=Eco,
      Airline_Code=EV}                     => {Satisfaction=Low} 0.1272149  0.6330749 2.924673  1960
[23] {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      Class=Eco,
      Airline_Name=FlyFast Airways Inc. } => {Satisfaction=Low} 0.1272149  0.6330749 2.924673  1960
[24] {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      Class=Eco,
      Airline_Code=EV,
      Airline_Name=FlyFast Airways Inc. } => {Satisfaction=Low} 0.1272149  0.6330749 2.924673  1960
[25] {Airline_Status=Blue,
      Type_of_Travel=Personal Travel}      => {Satisfaction=Low} 0.1554488  0.6320929 2.920137  2395
[26] {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      Airline_Code=EV}                     => {Satisfaction=Low} 0.1554488  0.6320929 2.920137  2395
[27] {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      Airline_Name=FlyFast Airways Inc. } => {Satisfaction=Low} 0.1554488  0.6320929 2.920137  2395
[28] {Airline_Status=Blue,
      Type_of_Travel=Personal Travel,
      Airline_Code=EV,
      Airline_Name=FlyFast Airways Inc. } => {Satisfaction=Low} 0.1554488  0.6320929 2.920137  2395
```

Scatter plot shows the rules which has the high confidence and lift values. Support factor depicts that these rules are applicable to all types of case and confidence factor shows how often these rules are correct.

SouthEast Airlines:

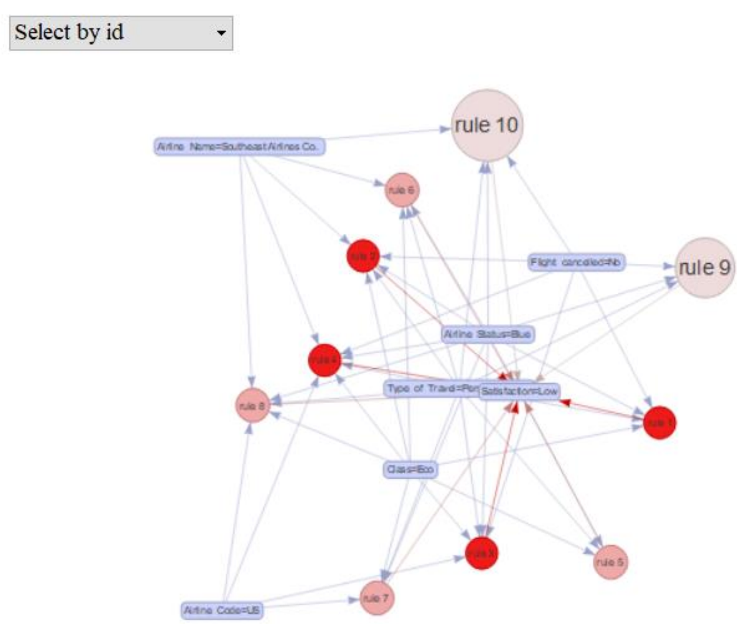


Flyfast Airways:

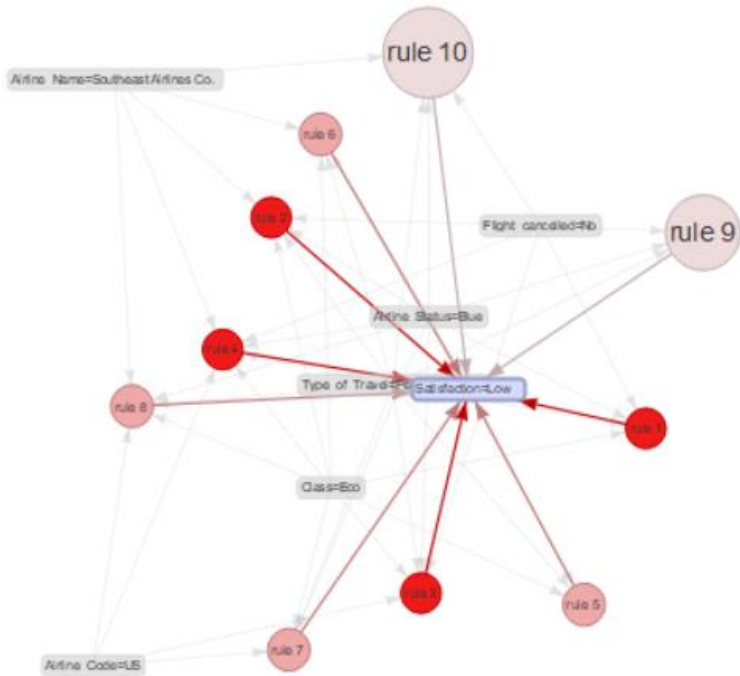


Subsetting the dataset which contains various attributes and trying to see the association between these attributes and low customer satisfaction.

SouthEast airlines:

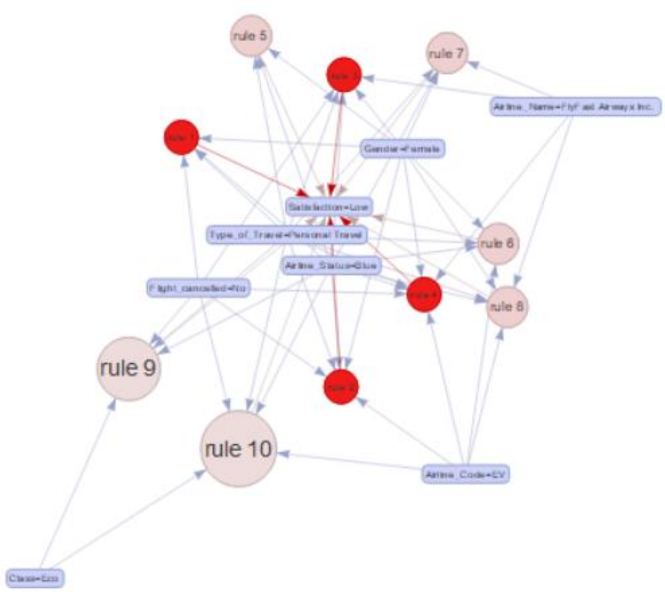


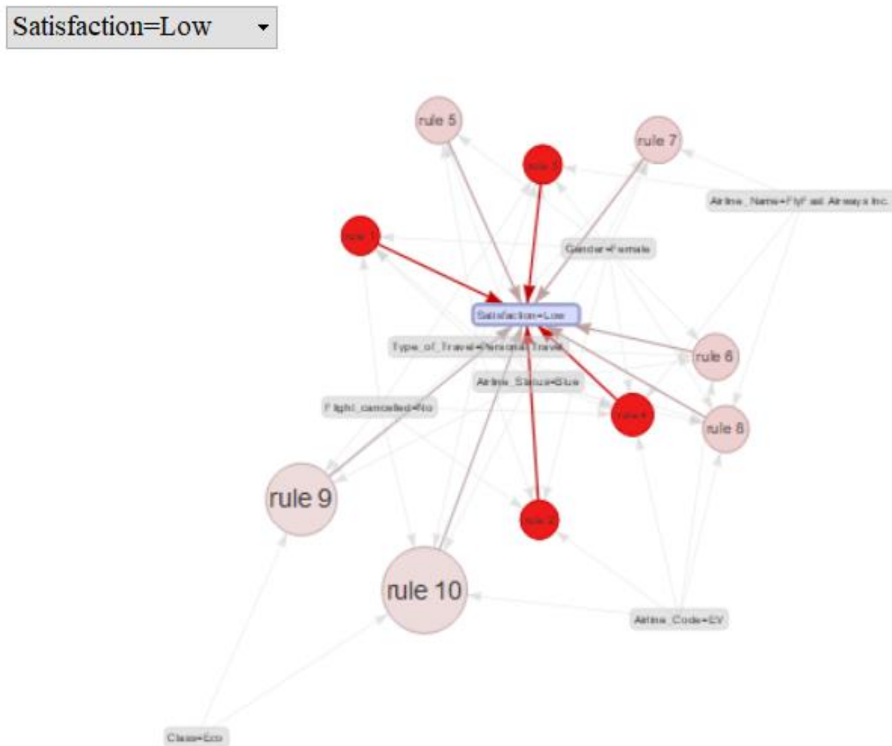
Satisfaction=Low



Flyfast Airways:

Select by id





Support Vector Machine (SVM) models for FlyFast Airways Inc.

For the prediction of low customer satisfaction we are going to add a new variable to our data set called lowSat. This variable/attribute consists of 0s and 1s, 1 being the customers with Satisfaction less than 3.

Also, we excluded the following attributes while modelling to reduce redundancy – Satisfaction, Flight_date, Airline_Code, Airline_Name, Origin_State and Destination_State

All the models are developed using the cost function as 50 because it gave the minimum amount of training and cross validation errors.

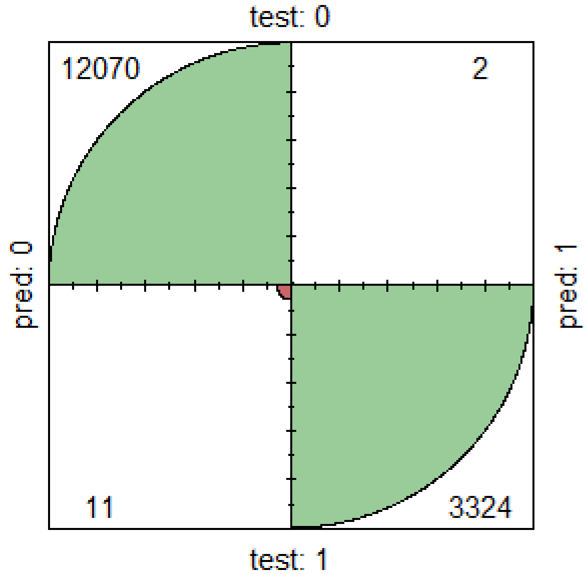
SVM Model 1:

In this model we try to predict the unhappy customers i.e. lowSat, using all the other attributes in our data set giving out the following results –

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

SV type: C-svc  (classification)
 parameter : cost C = 50

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.0510548189886173

Number of Support Vectors : 6337

Objective Function Value : -44558.83
Training error : 0.000844
Cross validation error : 0.158305
Probability model included.
```

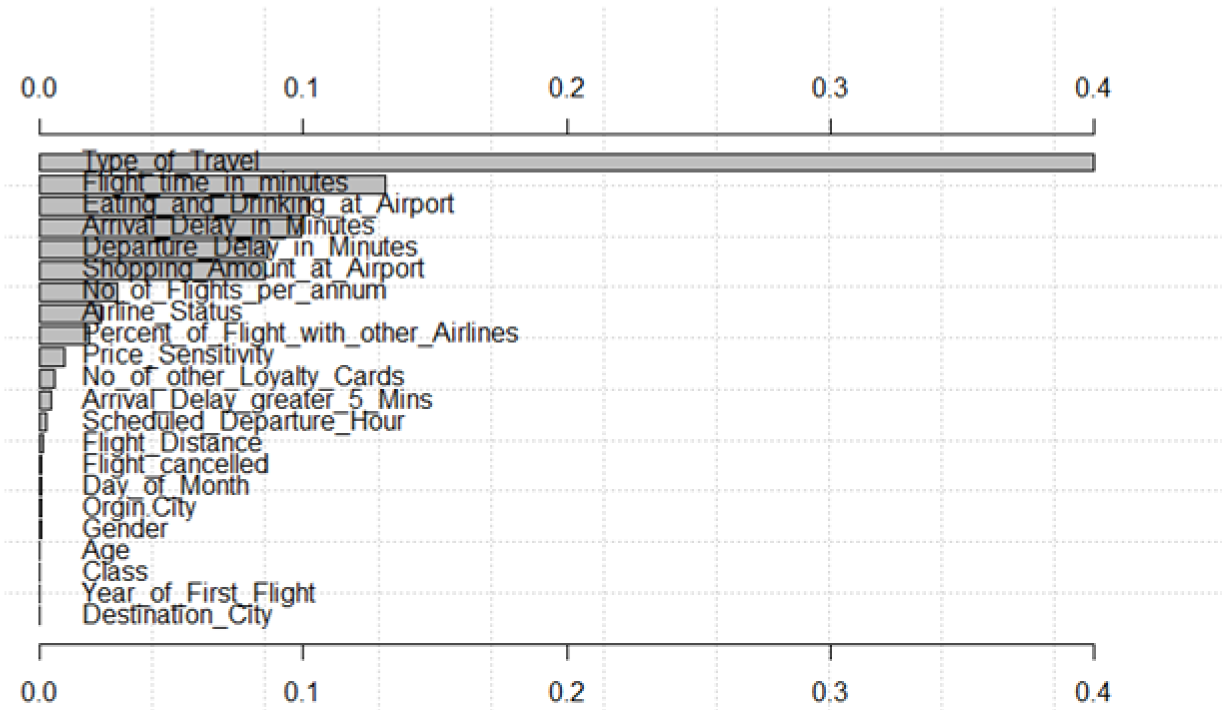
```
        pred
test        0     1
    0 12070     2
    1    11  3324
```

Confusion Matrix for Flyfast Airlines - All Attributes



The above confusion matrix and four-fold plot portray both statistically and visually that the SVM model we designed can successfully predict the categories of customers with low satisfaction with an accuracy of **99.915%**.

Variable importance analysis -
To rank or order the attributes on the basis of their significance in the svm model we did a variable importance analysis using the rminer package –



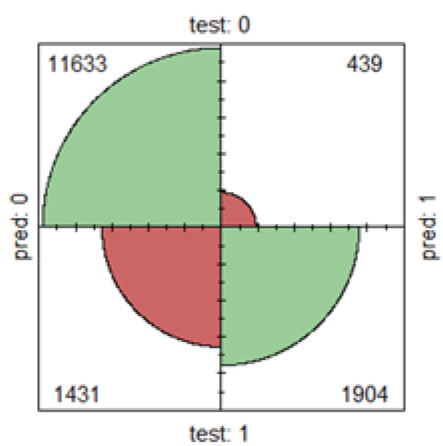
SVM Model 2:

In this model we try to predict the unhappy customers i.e. lowSat, using all the attributes that we got from the variable importance analysis in our data set giving out the following results –

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

SV type: C-svc  (classification)
 parameter : cost C = 50

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.222222222222222

Number of Support Vectors : 4568

Objective Function Value : -191394
Training error : 0.121373
Cross validation error : 0.135003
Probability model included.
```

```
      pred
test     0     1
   0 11633   439
   1  1431  1904
```

Confusion Matrix for Flyfast Airlines - Significant Attributes



The above confusion matrix and four-fold plot portray both statistically and visually that the SVM model we designed can successfully predict the categories of customers with low satisfaction with an accuracy of **87.862%** validating our variable importance study.

Support Vector Machine (SVM) models for Southeast Airlines Co.

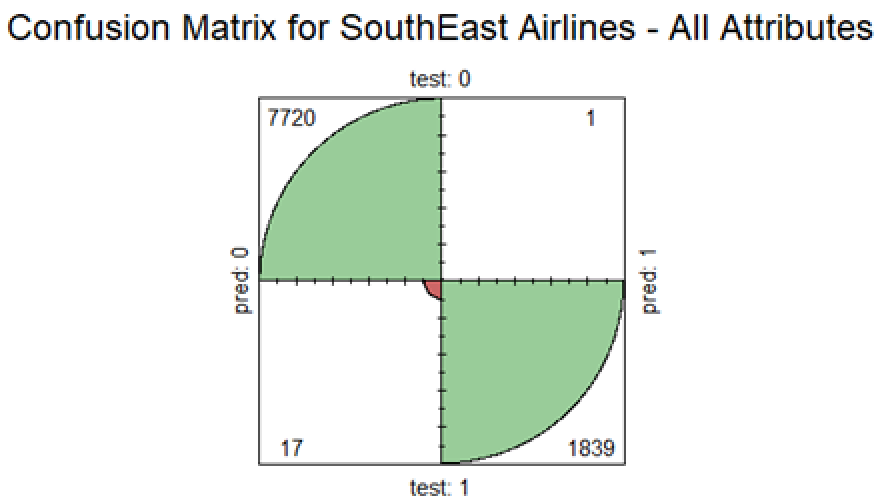
To compare our results with Southeast airlines we did predictive modelling for the dataset of Southeast Airlines too. For the prediction of low customer satisfaction we are going to add a new variable to our data set called lowSat. This variable/attribute consists of 0s and 1s, 1 being the customers with Satisfaction less than 3.
Also, we excluded the following attributes while modelling to reduce redundancy – Satisfaction, Flight_date, Airline_Code, Airline_Name, Origin_State and Destination_State
All the models are developed using the cost function as 50 because it gave the minimum amount of training and cross validation errors.

SVM Model 1:

In this model we try to predict the unhappy customers i.e. lowSat, using all the other attributes in our data set giving out the following results –

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

SV type: C-svc  (classification)
 parameter : cost C = 50

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.0548572282046079

Number of Support Vectors : 3852

Objective Function Value : -30442.76
Training error : 0.00188
Cross validation error : 0.168112
Probability model included.
```

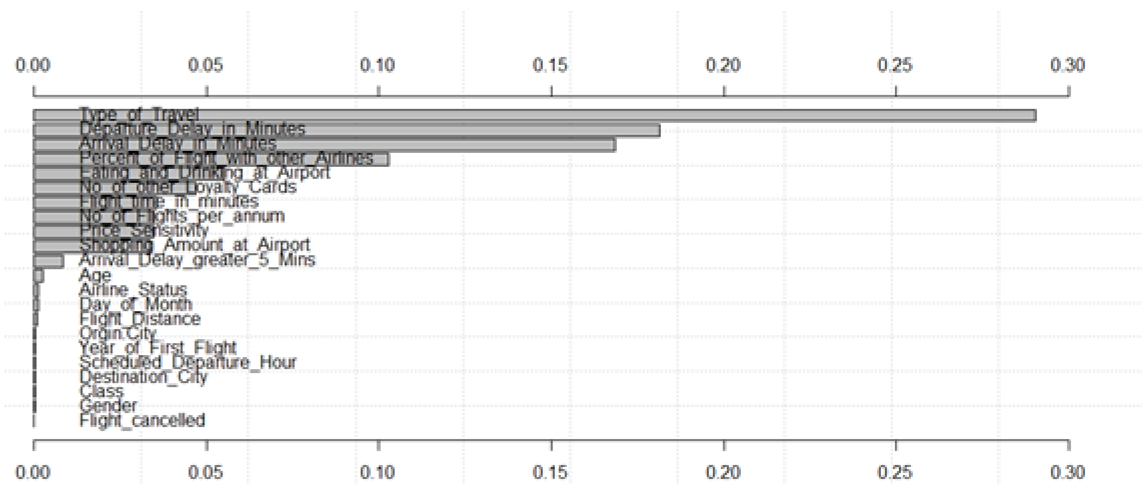
```
      pred
test     0    1
   0 7720    1
   1   17 1839
```



The above confusion matrix and four-fold plot portray both statistically and visually that the SVM model we designed can successfully predict the categories of customers with low satisfaction with an accuracy of **99.812%**.

Variable importance analysis -

To rank or order the attributes of Southeast Airlines on the basis of their significance in the svm model we did a variable importance analysis using the rminer package –



This plot shows the statistically important or significant attributes in the svm model.

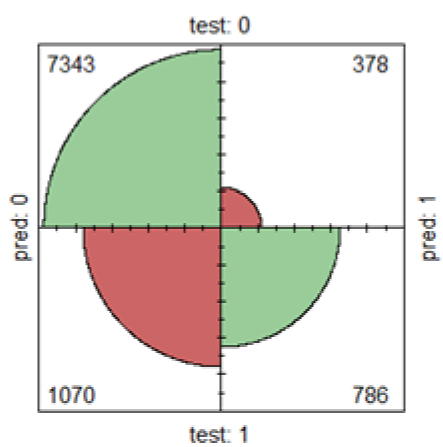
SVM Model 2:

In this model we try to predict the unhappy customers i.e. lowSat, using all the other attributes in our data set giving out the following results –

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

SV type: C-svc  (classification)
 parameter : cost C = 50

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.321428571428571

Number of Support Vectors : 3080

Objective Function Value : -145197.4
Training error : 0.151196
Cross validation error : 0.160801
Probability model included.
```

```
        pred
test       0     1
     0  7343   378
     1  1070   786
```

Confusion Matrix for SouthEast Airlines - Significant Attributes



The above confusion matrix and four-fold plot portray both statistically and visually that the SVM model we designed can successfully predict the categories of customers with low satisfaction with an accuracy of **84.88%** validating our variable importance study.

ACTIONABLE INSIGHTS/ RECOMMENDATIONS:

After analyzing the dataset and focusing on the business questions, a set of recommended solutions are listed below which would improve the customer satisfaction for the airlines. They are given as follows:

- **Attributes influencing customer satisfaction:**
  Flight cancelled, airline status, type of travel, gender, class

  Airlines should focus on avoiding flight cancellations and delays because of which the customers who make personal travel for spending time with their families and friends, and for other leisure activities get affected and this leads to a low customer satisfaction.

- **Characteristics or attributes of unhappy customers:**
  Age, gender, class and type of travel

- Avoid fligh

Appendix
Project Code-

```
setwd("C:/Users/hp/Desktop/Coursework/Fall-2018/IST-687/Final Project")

#Saving the csv file to a dataframe

survey <- read.csv("Satisfaction Survey.csv")

#Analyse the dataset
summary(survey)
str(survey)

#table(survey$Airline.Name)

#trim the dataset to remove unnecessary spaces
trimws(survey$Satisfaction, which = c("both"))
trimws(survey$Airline.Status, which = c("both"))
trimws(survey$Airline.Code, which = c("both"))
trimws(survey$Airline.Name, which = c("both"))
trimws(survey$Age, which = c("both"))
trimws(survey$Gender, which = c("both"))
trimws(survey$Price.Sensitivity, which = c("both"))
trimws(survey$Year.of.First.Flight, which = c("both"))
trimws(survey$No.of.Flights.p.a., which = c("both"))
trimws(survey$No..of.other.Loyalty.Cards, which = c("both"))
trimws(survey$X..of.Flight.with.other.Airlines, which = c("both"))
```

```
trimws(survey$Type.of.Travel, which = c("both"))
trimws(survey$Shopping.Amount.at.Airport, which = c("both"))
trimws(survey$Eating.and.Drinking.at.Airport, which = c("both"))
trimws(survey$Class, which = c("both"))
trimws(survey$Day.of.Month, which = c("both"))
trimws(survey$Flight.Distance, which = c("both"))
trimws(survey$Flight.date, which = c("both"))
trimws(survey$Flight.cancelled, which = c("both"))
trimws( survey$Flight.time.in.minutes,which = c("both"))
trimws(survey$Orgin.City, which = c("both"))
trimws(survey$Origin.State,which = c("both"))
trimws(survey$Destination.City, which = c("both"))
trimws(survey$Destination.State, which = c("both"))
trimws(survey$Scheduled.Departure.Hour, survey$Departure.Delay.in.Minutes)
trimws(survey$Arrival.Delay.in.Minutes, which = c("both"))
trimws(survey$Arrival.Delay.greater.5.Mins, which = c("both"))


#Change the column names
colnames(survey)[colnames(survey)=="Airline.Status"] <- "Airline_Status"
colnames(survey)[colnames(survey)=="Price.Sensitivity"] <- "Price_Sensitivity"
colnames(survey)[colnames(survey)=="Year.of.First.Flight"] <- "Year_of_First_Flight"
colnames(survey)[colnames(survey)=="No.of.Flights.p.a."] <- "No_of_Flights_per_annum"
colnames(survey)[colnames(survey)=="X..of.Flight.with.other.Airlines"] <- "Percent_of_Flight_with_other_Airlines"
colnames(survey)[colnames(survey)=="Type.of.Travel"] <- "Type_of_Travel"
colnames(survey)[colnames(survey)=="No..of.other.Loyalty.Cards"] <- "No_of_other_Loyalty_Cards"
colnames(survey)[colnames(survey)=="Shopping.Amount.at.Airport"] <- "Shopping_Amount_at_Airport"
colnames(survey)[colnames(survey)=="Eating.and.Drinking.at.Airport"] <- "Eating_and_Drinking_at_Airport"
colnames(survey)[colnames(survey)=="Day.of.Month"] <- "Day_of_Month"
colnames(survey)[colnames(survey)=="Flight.date"] <- "Flight_date"
colnames(survey)[colnames(survey)=="Airline.Code"] <- "Airline_Code"
colnames(survey)[colnames(survey)=="Airline.Name"] <- "Airline_Name"
colnames(survey)[colnames(survey)=="Origin.City"] <- "Origin_City"
colnames(survey)[colnames(survey)=="Origin.State"] <- "Origin_State"
colnames(survey)[colnames(survey)=="Destination.City"] <- "Destination_City"
colnames(survey)[colnames(survey)=="Destination.State"] <- "Destination_State"
colnames(survey)[colnames(survey)=="Scheduled.Departure.Hour"] <- "Scheduled_Departure_Hour"
colnames(survey)[colnames(survey)=="Departure.Delay.in.Minutes"] <- "Departure_Delay_in_Minutes"
colnames(survey)[colnames(survey)=="Arrival.Delay.in.Minutes"] <- "Arrival_Delay_in_Minutes"
colnames(survey)[colnames(survey)=="Flight.cancelled"] <- "Flight_cancelled"
colnames(survey)[colnames(survey)=="Flight.time.in.minutes"] <- "Flight_time_in_minutes"
colnames(survey)[colnames(survey)=="Flight.Distance"] <- "Flight_Distance"
colnames(survey)[colnames(survey)=="Arrival.Delay.greater.5.Mins"] <- "Arrival_Delay_greater_5_Mins"

#Converting the NAs in the three columns to their mean values
survey$Flight_time_in_minutes[is.na(survey$Flight_time_in_minutes)] <-
round(mean(survey$Flight_time_in_minutes, na.rm = TRUE))
survey$Arrival_Delay_in_Minutes[is.na(survey$Arrival_Delay_in_Minutes)] <-
round(mean(survey$Arrival_Delay_in_Minutes, na.rm = TRUE))
survey$Departure_Delay_in_Minutes[is.na(survey$Departure_Delay_in_Minutes)] <-
round(mean(survey$Departure_Delay_in_Minutes, na.rm = TRUE))
summary(survey)

# Converting the data types to numeric.
```

```
numberize <- function(survey){
  for (i in 1:28){
    survey[,i] <- as.numeric(survey[,i])


  }
  return(survey)
}

#Verify
str(survey)

#install.packages("dplyr")
#library(dplyr)

res <- survey %>% group_by(Age) %>% summarise(Freq=n(), percLess3= sum(Satisfaction<3, na.rm = TRUE))
res$perc <- (res$percLess3/res$Freq)*100

res_Airline <- survey %>% group_by(Airline_Name) %>% summarise(Freq=n(), percLess3= sum(Satisfaction<3,
na.rm = TRUE))
res_Airline$perc <- (res_Airline$percLess3/res_Airline$Freq)*100

ggplot(res_Airline, aes(x=Airline_Name ,y = Freq, fill=perc))+geom_col()+theme(axis.text.x = element_text(angle =
90, hjust = 1))

ggplot(airlinedf, aes(x=Airline_Name, y=Airline_CustSat)) + geom_col() + theme(axis.text.x = element_text(angle =
90, hjust = 1))

#Data set for the Southeast Airlines Co.
Airline1_df <- filter(survey, Airline_Name == "Southeast Airlines Co. ")

#Data Set for the FlyFast Airways Inc
Airline2_df <- filter(survey, Airline_Name == "FlyFast Airways Inc. ")

#linear regression



 install.packages("dplyr")
 library(dplyr)



 Airline1_df$Satisfaction <- as.numeric(Airline1_df$Satisfaction)

 Airline1_df <- numberize(Airline1_df)

 model1 <- lm(formula = Satisfaction ~., data = Airline1_df)
 summary(model1)

 #Multiple R-squared:  0.3984
 # Adjusted R-squared:  0.3979
```

```
Airline2_df <- filter(survey, Airline_Name == "FlyFast Airways Inc. ")
Airline2_df$Satisfaction <- as.numeric(Airline2_df$Satisfaction)
Airline2_df <- numberize(Airline2_df)
str(Airline2_df)

model2 <- lm(formula = Satisfaction ~., data = Airline1_df)
summary(model2)


#Multiple R-squared:  0.3962
#Adjusted R-squared:  0.3952



#For South East Airline

# Airline_Status, Age ,Gender, Year_of_First_Flight, No_of_Flights_per_annum, Type_of_Travel,
Scheduled_Departure_Hour, Flight_cancelled,Arrival_Delay_greater_5_Mins


#Grouping all the significant variables

model11 <- lm(formula = Satisfaction ~Airline_Status+Age+Gender+Year_of_First_Flight+
No_of_Flights_per_annum+ Type_of_Travel+Scheduled_Departure_Hour+ Flight_cancelled+
Arrival_Delay_greater_5_Mins, data = Airline1_df)
summary(model11)


#Multiple R-squared:  0.3804
#Adjusted R-squared:  0.3798


#without + Arrival_Delay_greater_5_Mins


model12 <- lm(formula = Satisfaction ~Airline_Status+Age+Gender+Year_of_First_Flight+
No_of_Flights_per_annum+ Type_of_Travel+Scheduled_Departure_Hour+ Flight_cancelled, data = Airline1_df)
summary(model12)

#Multiple R-squared:  0.3622
#Adjusted R-squared:  0.3616

# + Flight_cancelled

model13 <- lm(formula = Satisfaction ~Airline_Status+Age+Gender+Year_of_First_Flight+
No_of_Flights_per_annum+ Type_of_Travel+Scheduled_Departure_Hour + Arrival_Delay_greater_5_Mins, data =
Airline1_df)
summary(model13)


#Multiple R-squared:  0.3793
#Adjusted R-squared:  0.3788
```

```
#Scheduled_Departure_Hour+

model14 <- lm(formula = Satisfaction ~Airline_Status+Age+Gender+Year_of_First_Flight+
No_of_Flights_per_annum+ Type_of_Travel+ Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline1_df)
summary(model14)

#Multiple R-squared:  0.3795
#Adjusted R-squared:  0.3789

#Type_of_Travel
model15 <- lm(formula = Satisfaction ~Airline_Status+Age+Gender+Year_of_First_Flight+
No_of_Flights_per_annum +Scheduled_Departure_Hour+ Flight_cancelled+ Arrival_Delay_greater_5_Mins, data =
Airline1_df)
summary(model15)

#Multiple R-squared:  0.2158
#Adjusted R-squared:  0.2152

# + No_of_Flights_per_annum

model16 <- lm(formula = Satisfaction ~Airline_Status+Age+Gender+Year_of_First_Flight +
Type_of_Travel+Scheduled_Departure_Hour+ Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline1_df)
summary(model16)

#Multiple R-squared:  0.3777
#Adjusted R-squared:  0.3772

#Year_of_First_Flight

model17 <- lm(formula = Satisfaction ~Airline_Status+Age+Gender + No_of_Flights_per_annum+
Type_of_Travel+Scheduled_Departure_Hour+ Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline1_df)
summary(model17)

#Multiple R-squared:  0.3795
#Adjusted R-squared:  0.379

# +Gender

model18 <- lm(formula = Satisfaction ~Airline_Status+Age +Year_of_First_Flight+ No_of_Flights_per_annum+
Type_of_Travel+Scheduled_Departure_Hour+ Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline1_df)
summary(model18)

#Multiple R-squared:  0.3725
#Adjusted R-squared:  0.372

#Age

model19 <- lm(formula = Satisfaction ~Airline_Status+Gender +Year_of_First_Flight+ No_of_Flights_per_annum+
Type_of_Travel+Scheduled_Departure_Hour+ Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline1_df)
summary(model19)

#Multiple R-squared:  0.3789
#Adjusted R-squared:  0.3783
```

#Airline_Status

model20 <- lm(formula = Satisfaction ~Age +Gender +Year_of_First_Flight+ No_of_Flights_per_annum+ Type_of_Travel+Scheduled_Departure_Hour+ Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline1_df)
summary(model20)

#Multiple R-squared: 0.3149
#Adjusted R-squared: 0.3143

##Year_of_First_Flight

#model17 <- lm(formula = Satisfaction ~Airline_Status+Age+Gender + No_of_Flights_per_annum+ Type_of_Travel+Scheduled_Departure_Hour+ Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline1_df)
#summary(model17)

#Multiple R-squared: 0.3795
#Adjusted R-squared: 0.379

model171 <- lm(formula = Satisfaction ~Airline_Status+Age+Gender + No_of_Flights_per_annum+ Type_of_Travel+Scheduled_Departure_Hour+ Flight_cancelled, data = Airline1_df)
summary(model171)

#Multiple R-squared: 0.3612
#Adjusted R-squared: 0.3608

#Flight_cancelled
model172 <- lm(formula = Satisfaction ~Airline_Status+Age+Gender + No_of_Flights_per_annum+ Type_of_Travel+Scheduled_Departure_Hour+ Arrival_Delay_greater_5_Mins, data = Airline1_df)
summary(model172)

#Multiple R-squared: 0.3784
#Adjusted R-squared: 0.378

#+Scheduled_Departure_Hour
model173 <- lm(formula = Satisfaction ~Airline_Status+Age+Gender + No_of_Flights_per_annum+ Type_of_Travel + Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline1_df)
summary(model173)
#Multiple R-squared: 0.3786
#Adjusted R-squared: 0.3781

#+ Type_of_Travel
model174 <- lm(formula = Satisfaction ~Airline_Status+Age+Gender + No_of_Flights_per_annum +Scheduled_Departure_Hour+ Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline1_df)
summary(model174)
#Multiple R-squared: 0.2153
Adjusted R-squared: 0.2148
#+ No_of_Flights_per_annum

model175 <- lm(formula = Satisfaction ~Airline_Status+Age+Gender + Type_of_Travel+Scheduled_Departure_Hour+ Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline1_df)
summary(model175)

```
 #Multiple R-squared:  0.3769
 #Adjusted R-squared:  0.3764


 #+Gender
 model176 <- lm(formula = Satisfaction ~Airline_Status+Age + No_of_Flights_per_annum+
Type_of_Travel+Scheduled_Departure_Hour+ Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline1_df)
 summary(model176)
 #Multiple R-squared:  0.3716
 #Adjusted R-squared:  0.3711

 #+Age
 model177 <- lm(formula = Satisfaction ~Airline_Status +Gender + No_of_Flights_per_annum+
Type_of_Travel+Scheduled_Departure_Hour+ Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline1_df)
 summary(model177)
 #Multiple R-squared:  0.378
 #Adjusted R-squared:  0.3776

 # Airline_Status+
 model178 <- lm(formula = Satisfaction ~ Age+Gender + No_of_Flights_per_annum+
Type_of_Travel+Scheduled_Departure_Hour+ Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline1_df)
 summary(model178)

 #Multiple R-squared:  0.3132
 #Adjusted R-squared:  0.3127



 #+Scheduled_Departure_Hour
 #model173 <- lm(formula = Satisfaction ~Airline_Status+Age+Gender + No_of_Flights_per_annum+ Type_of_Travel
+ Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline1_df)
 #summary(model173)
 #Multiple R-squared:  0.3786
 #Adjusted R-squared:  0.3781


 #Airline_Status+Age+Gender + No_of_Flights_per_annum+ Type_of_Travel + Flight_cancelled+
Arrival_Delay_greater_5_Mins


 #+ Arrival_Delay_greater_5_Mins
 model1731 <- lm(formula = Satisfaction ~Airline_Status+Age+Gender + No_of_Flights_per_annum+ Type_of_Travel
+ Flight_cancelled, data = Airline1_df)
 summary(model1731)

 #Multiple R-squared:  0.3606, Adjusted R-squared:  0.3602

 # Flight_cancelled
 model1732 <- lm(formula = Satisfaction ~Airline_Status+Age+Gender + No_of_Flights_per_annum+ Type_of_Travel
+ Arrival_Delay_greater_5_Mins, data = Airline1_df)
 summary(model1732)
```

```
 #Multiple R-squared:  0.3775, Adjusted R-squared:  0.3771


 #+ Type_of_Travel
 model1733 <- lm(formula = Satisfaction ~Airline_Status+Age+Gender + No_of_Flights_per_annum +
Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline1_df)
 summary(model1733)

 #Multiple R-squared:  0.2141, Adjusted R-squared:  0.2136
 # + No_of_Flights_per_annum
 model1734 <- lm(formula = Satisfaction ~Airline_Status+Age+Gender + Type_of_Travel + Flight_cancelled+
Arrival_Delay_greater_5_Mins, data = Airline1_df)
 summary(model1734)
 #Multiple R-squared:  0.376, Adjusted R-squared:  0.3756
 #+Gender
 model1735 <- lm(formula = Satisfaction ~Airline_Status+Age + No_of_Flights_per_annum+ Type_of_Travel +
Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline1_df)
 summary(model1735)
 #Multiple R-squared:  0.3707, Adjusted R-squared:  0.3703
 # +Age
 model1736 <- lm(formula = Satisfaction ~Airline_Status +Gender + No_of_Flights_per_annum+ Type_of_Travel +
Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline1_df)
 summary(model1736)

 #Multiple R-squared:  0.3771, Adjusted R-squared:  0.3767

 # Airline_Status

 model1737 <- lm(formula = Satisfaction ~ Age+Gender + No_of_Flights_per_annum+ Type_of_Travel +
Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline1_df)
 summary(model1737)

 #Multiple R-squared:  0.3121, Adjusted R-squared:  0.3117

 #Airline_Status+Age+Gender + No_of_Flights_per_annum+ Type_of_Travel + Arrival_Delay_greater_5_Mins

#For Fareast airline

 #Grouping all the significant variables

 model11 <- lm(formula = Satisfaction ~Airline_Status+Age+Gender+Year_of_First_Flight+
No_of_Flights_per_annum+ Type_of_Travel+Scheduled_Departure_Hour+ Flight_cancelled+
Arrival_Delay_greater_5_Mins, data = Airline2_df)
 summary(model11)

 #Multiple R-squared:  0.3944, Adjusted R-squared:  0.394



 #without + Arrival_Delay_greater_5_Mins


 model12 <- lm(formula = Satisfaction ~Airline_Status+Age+Gender+Year_of_First_Flight+
No_of_Flights_per_annum+ Type_of_Travel+Scheduled_Departure_Hour+ Flight_cancelled, data = Airline2_df)
```

summary(model12)

#Multiple R-squared: 0.3754
#Adjusted R-squared: 0.3751

+ Flight_cancelled

model13 <- lm(formula = Satisfaction ~Airline_Status+Age+Gender+Year_of_First_Flight+
No_of_Flights_per_annum+ Type_of_Travel+Scheduled_Departure_Hour + Arrival_Delay_greater_5_Mins, data =
Airline2_df)
summary(model13)

#Multiple R-squared: 0.3931
#Adjusted R-squared: 0.3928

#Scheduled_Departure_Hour+

model14 <- lm(formula = Satisfaction ~Airline_Status+Age+Gender+Year_of_First_Flight+
No_of_Flights_per_annum+ Type_of_Travel+ Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline2_df)
summary(model14)

#Multiple R-squared: 0.3942
#Adjusted R-squared: 0.3939

#Type_of_Travel
model15 <- lm(formula = Satisfaction ~Airline_Status+Age+Gender+Year_of_First_Flight+
No_of_Flights_per_annum +Scheduled_Departure_Hour+ Flight_cancelled+ Arrival_Delay_greater_5_Mins, data =
Airline2_df)
summary(model15)

#Multiple R-squared: 0.2213
#Adjusted R-squared: 0.2209

+ No_of_Flights_per_annum

model16 <- lm(formula = Satisfaction ~Airline_Status+Age+Gender+Year_of_First_Flight +
Type_of_Travel+Scheduled_Departure_Hour+ Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline2_df)
summary(model16)

#Multiple R-squared: 0.391
#Adjusted R-squared: 0.3906

#Year_of_First_Flight

model17 <- lm(formula = Satisfaction ~Airline_Status+Age+Gender + No_of_Flights_per_annum+
Type_of_Travel+Scheduled_Departure_Hour+ Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline2_df)
summary(model17)

#Multiple R-squared: 0.3936
#Adjusted R-squared: 0.3933

+Gender

```
model18 <- lm(formula = Satisfaction ~Airline_Status+Age +Year_of_First_Flight+ No_of_Flights_per_annum+
Type_of_Travel+Scheduled_Departure_Hour+ Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline2_df)
summary(model18)

#Multiple R-squared:  0.386
#Adjusted R-squared:  0.3857

#Age

model19 <- lm(formula = Satisfaction ~Airline_Status+Gender +Year_of_First_Flight+ No_of_Flights_per_annum+
Type_of_Travel+Scheduled_Departure_Hour+ Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline2_df)
summary(model19)

#Multiple R-squared:  0.394
#Adjusted R-squared:  0.3937

#Airline_Status

model20 <- lm(formula = Satisfaction ~Age +Gender +Year_of_First_Flight+ No_of_Flights_per_annum+
Type_of_Travel+Scheduled_Departure_Hour+ Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline2_df)
summary(model20)

#Multiple R-squared:  0.3277
#Adjusted R-squared:  0.3273

#Scheduled_Departure_Hour+

#model14 <- lm(formula = Satisfaction ~Airline_Status+Age+Gender+Year_of_First_Flight+
No_of_Flights_per_annum+ Type_of_Travel+ Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline2_df)
#summary(model14)

#Multiple R-squared:  0.3942
#Adjusted R-squared:  0.3939

#Airline_Status+Age+Gender+Year_of_First_Flight+ No_of_Flights_per_annum+ Type_of_Travel+
Flight_cancelled+ Arrival_Delay_greater_5_Mins

#+ Arrival_Delay_greater_5_Mins
model141 <- lm(formula = Satisfaction ~Airline_Status+Age+Gender+Year_of_First_Flight+
No_of_Flights_per_annum+ Type_of_Travel+ Flight_cancelled, data = Airline2_df)
summary(model141)

#Multiple R-squared:  0.3754, Adjusted R-squared:  0.3751

#+ Flight_cancelled
model142 <- lm(formula = Satisfaction ~Airline_Status+Age+Gender+Year_of_First_Flight+
No_of_Flights_per_annum+ Type_of_Travel + Arrival_Delay_greater_5_Mins, data = Airline2_df)
summary(model142)

#Multiple R-squared:  0.393, Adjusted R-squared:  0.3927

# Type_of_Travel

model143 <- lm(formula = Satisfaction ~Airline_Status+Age+Gender+Year_of_First_Flight+
```

No_of_Flights_per_annum+ Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline2_df)
summary(model143)

#Multiple R-squared: 0.2212, Adjusted R-squared: 0.2208

# No_of_Flights_per_annum+
model144 <- lm(formula = Satisfaction ~Airline_Status+Age+Gender+Year_of_First_Flight+ Type_of_Travel+ Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline2_df)
summary(model144)

#Multiple R-squared: 0.3908, Adjusted R-squared: 0.3905

#+Year_of_First_Flight
model145 <- lm(formula = Satisfaction ~Airline_Status+Age+Gender + No_of_Flights_per_annum+ Type_of_Travel+ Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline2_df)
summary(model145)

#Multiple R-squared: 0.3934, Adjusted R-squared: 0.3932

+Gender

model146 <- lm(formula = Satisfaction ~Airline_Status+Age +Year_of_First_Flight+ No_of_Flights_per_annum+ Type_of_Travel+ Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline2_df)
summary(model146)
#Multiple R-squared: 0.3858, Adjusted R-squared: 0.3855

#+Age

model147 <- lm(formula = Satisfaction ~Airline_Status +Gender+Year_of_First_Flight+ No_of_Flights_per_annum+ Type_of_Travel+ Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline2_df)
summary(model147)

#Multiple R-squared: 0.3938, Adjusted R-squared: 0.3935

Airline_Status+

model148<- lm(formula = Satisfaction ~ Age+Gender+Year_of_First_Flight+ No_of_Flights_per_annum+ Type_of_Travel+ Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline2_df)
summary(model148)

#Multiple R-squared: 0.3275, Adjusted R-squared: 0.3271

#+Age

#model147 <- lm(formula = Satisfaction ~Airline_Status +Gender+Year_of_First_Flight+ No_of_Flights_per_annum+ Type_of_Travel+ Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline2_df)
#summary(model147)

#Multiple R-squared: 0.3938, Adjusted R-squared: 0.3935

```
 #Airline_Status +Gender+Year_of_First_Flight+ No_of_Flights_per_annum+ Type_of_Travel+ Flight_cancelled+
Arrival_Delay_greater_5_Mins

 #+ Arrival_Delay_greater_5_Mins
 model1471 <- lm(formula = Satisfaction ~Airline_Status +Gender+Year_of_First_Flight+ No_of_Flights_per_annum+
Type_of_Travel+ Flight_cancelled, data = Airline2_df)
 summary(model1471)

 #Multiple R-squared:  0.375, Adjusted R-squared:  0.3747
 #+ Flight_cancelled
 model1472 <- lm(formula = Satisfaction ~Airline_Status +Gender+Year_of_First_Flight+ No_of_Flights_per_annum+
Type_of_Travel + Arrival_Delay_greater_5_Mins, data = Airline2_df)
 summary(model1472)
 #Multiple R-squared:  0.3926, Adjusted R-squared:  0.3923
 # Type_of_Travel

 model1473 <- lm(formula = Satisfaction ~Airline_Status +Gender+Year_of_First_Flight+ No_of_Flights_per_annum+
Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline2_df)
 summary(model1473)

 #Multiple R-squared:  0.2011, Adjusted R-squared:  0.2008

 # No_of_Flights_per_annum+
 model1474 <- lm(formula = Satisfaction ~Airline_Status +Gender+Year_of_First_Flight+ Type_of_Travel+
Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline2_df)
 summary(model1474)

 #Multiple R-squared:  0.3899, Adjusted R-squared:  0.3896


 #+Year_of_First_Flight
 model1475 <- lm(formula = Satisfaction ~Airline_Status +Gender + No_of_Flights_per_annum+ Type_of_Travel+
Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline2_df)
 summary(model1475)
 #Multiple R-squared:  0.393, Adjusted R-squared:  0.3928
 # +Gender

 model1476 <- lm(formula = Satisfaction ~Airline_Status +Year_of_First_Flight+ No_of_Flights_per_annum+
Type_of_Travel+ Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline2_df)
 summary(model1476)

 #Multiple R-squared:  0.3853, Adjusted R-squared:  0.3851
 # Airline_Status+

 model1477<- lm(formula = Satisfaction ~ Gender+Year_of_First_Flight+ No_of_Flights_per_annum+
Type_of_Travel+ Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline2_df)
 summary(model1477)

 #Multiple R-squared:  0.3274, Adjusted R-squared:  0.3272
 #+Year_of_First_Flight
 #model1475 <- lm(formula = Satisfaction ~Airline_Status +Gender + No_of_Flights_per_annum+ Type_of_Travel+
Flight_cancelled+ Arrival_Delay_greater_5_Mins, data = Airline2_df)
 #summary(model1475)
 #Multiple R-squared:  0.393, Adjusted R-squared:  0.3928
```

#Airline_Status +Gender + No_of_Flights_per_annum+ Type_of_Travel+ Flight_cancelled+ Arrival_Delay_greater_5_Mins

--- # association rules - flyfast:---

```
# airline 2
Airline2_df <- filter(survey, Airline_Name == "FlyFast Airways Inc. ")

Airline2_df$Satisfaction <- as.character(Airline2_df$Satisfaction)
Airline2_df$Satisfaction <- as.numeric(Airline2_df$Satisfaction)

library(dplyr)
library(arulesViz)

# arules

createBucketSat <- function(vec){
  vBuckets <- replicate(length(vec), "Average")
  vBuckets[vec > 3] <- "High"
  vBuckets[vec < 3] <- "Low"
  return(vBuckets)
}

createBuckets <- function(vec){
  q <- quantile(vec, c(0.3, 0.6), na.rm = "TRUE")
  vBuckets <- replicate(length(vec), "Average")
  vBuckets[vec <= q[1]] <- "Low"
  vBuckets[vec > q[2]] <- "High"
  return(vBuckets)
}

createBucketsPrice <- function(v) {
  vBucket <- replicate(length(v), "Average")
  vBucket[v>1] <- "High"
  vBucket[v<1] <- "Low"
  return(vBucket)
}

createBucketsFPA <- function(v) {
  vBucket <- replicate(length(v), "Average")
  vBucket[v>24] <- "High"
  vBucket[v<24] <- "Low"
  return(vBucket)
}

CreateBucketsFWOA <- function(v) {
  vBucket <- replicate(length(v), "Average")
  vBucket[v>6] <- "High"
  vBucket[v<6] <- "Low"
  return(vBucket)
}
```

```
new_arules_df <- Airline2_df
new_arules_df$Year_of_First_Flight <- NULL
new_arules_df$Day_of_Month <- NULL
new_arules_df$Scheduled_Departure_Hour <- NULL

new_arules_df$Satisfaction <- createBucketSat(new_arules_df$Satisfaction)
new_arules_df$Satisfaction <- as.factor(new_arules_df$Satisfaction)
View(new_arules_df)
new_arules_df$Price_Sensitivity <- createBucketsPrice(new_arules_df$Price_Sensitivity)
new_arules_df$Price_Sensitivity <- as.factor(new_arules_df$Price_Sensitivity)
new_arules_df$Age <- createBuckets(new_arules_df$Age)
new_arules_df$Age <- as.factor(new_arules_df$Age)
new_arules_df$No_of_other_Loyalty_Cards <- createBuckets(new_arules_df$No_of_other_Loyalty_Cards)
new_arules_df$No_of_other_Loyalty_Cards <- as.factor(new_arules_df$No_of_other_Loyalty_Cards)
new_arules_df$Shopping_Amount_at_Airport <- createBuckets(new_arules_df$Shopping_Amount_at_Airport)
new_arules_df$Shopping_Amount_at_Airport <- as.factor(new_arules_df$Shopping_Amount_at_Airport)
new_arules_df$No_of_Flights_per_annum <- createBucketsFPA(new_arules_df$No_of_Flights_per_annum)
new_arules_df$No_of_Flights_per_annum <- as.factor(new_arules_df$No_of_Flights_per_annum)
new_arules_df$Percent_of_Flight_with_other_Airlines <-
CreateBucketsFWOA(new_arules_df$Percent_of_Flight_with_other_Airlines)
new_arules_df$Percent_of_Flight_with_other_Airlines <-
as.factor(new_arules_df$Percent_of_Flight_with_other_Airlines)
new_arules_df$Eating_and_Drinking_at_Airport <- createBuckets(new_arules_df$Eating_and_Drinking_at_Airport)
new_arules_df$Eating_and_Drinking_at_Airport <- as.factor(new_arules_df$Eating_and_Drinking_at_Airport)
new_arules_df$Departure_Delay_in_Minutes <- createBuckets(new_arules_df$Departure_Delay_in_Minutes)
new_arules_df$Departure_Delay_in_Minutes <- as.factor(new_arules_df$Departure_Delay_in_Minutes)
new_arules_df$Arrival_Delay_in_Minutes <- createBuckets(new_arules_df$Arrival_Delay_in_Minutes)
new_arules_df$Arrival_Delay_in_Minutes <- as.factor(new_arules_df$Arrival_Delay_in_Minutes)
new_arules_df$Flight_time_in_minutes <- createBuckets(new_arules_df$Flight_time_in_minutes)
new_arules_df$Flight_time_in_minutes <- as.factor(new_arules_df$Flight_time_in_minutes)
new_arules_df$Flight_Distance <- createBuckets(new_arules_df$Flight_Distance)
new_arules_df$Flight_Distance <- as.factor(new_arules_df$Flight_Distance)
View(new_arules_df)
library(arules)

rulesetLow <- apriori(new_arules_df, parameter = list(support = 0.1, confidence = 0.6), appearance =
list(rhs=c("Satisfaction=Low")))
inspect(rulesetLow)
sorted_rules <- sort(rulesetLow, decreasing=TRUE, by="lift")
inspect(sorted_rules)
plot(sorted_rules)



subrules1 <- head(rulesetLow, n = 10, by = "lift")
plot(subrules1, method = "graph", engine = "htmlwidget")

#airline1
#Data set for the Southeast Airlines Co.
Airline1_df <- filter(survey, Airline_Name == "Southeast Airlines Co. ")

Airline1_df$Satisfaction <- as.character(Airline1_df$Satisfaction)
Airline1_df$Satisfaction <- as.numeric(Airline1_df$Satisfaction)
```

```
#Converting the NAs in the three columns to their mean values
Airline1_df$Flight_time_in_minutes[is.na(Airline1_df$Flight_time_in_minutes)] <-
round(mean(Airline1_df$Flight_time_in_minutes, na.rm = TRUE))
Airline1_df$Arrival_Delay_in_Minutes[is.na(Airline1_df$Arrival_Delay_in_Minutes)] <-
round(mean(Airline1_df$Arrival_Delay_in_Minutes, na.rm = TRUE))
Airline1_df$Departure_Delay_in_Minutes[is.na(Airline1_df$Departure_Delay_in_Minutes)] <-
round(mean(Airline1_df$Departure_Delay_in_Minutes, na.rm = TRUE))
summary(Airline1_df)

new_arules <- Airline1_df
new_arules$Year_of_First_Flight <- NULL
new_arules$Day_of_Month <- NULL
new_arules$Scheduled_Departure_Hour <- NULL

new_arules$Satisfaction <- createBucketSat(new_arules$Satisfaction)
new_arules$Satisfaction <- as.factor(new_arules$Satisfaction)
View(new_arules)
new_arules$Price_Sensitivity <- createBucketsPrice(new_arules$Price_Sensitivity)
new_arules$Price_Sensitivity <- as.factor(new_arules$Price_Sensitivity)
new_arules$Age <- createBuckets(new_arules$Age)
new_arules$Age <- as.factor(new_arules$Age)
new_arules$No_of_other_Loyalty_Cards <- createBuckets(new_arules$No_of_other_Loyalty_Cards)
new_arules$No_of_other_Loyalty_Cards <- as.factor(new_arules$No_of_other_Loyalty_Cards)
new_arules$Shopping_Amount_at_Airport <- createBuckets(new_arules$Shopping_Amount_at_Airport)
new_arules$Shopping_Amount_at_Airport <- as.factor(new_arules$Shopping_Amount_at_Airport)
new_arules$No_of_Flights_per_annum <- createBucketsFPA(new_arules$No_of_Flights_per_annum)
new_arules$No_of_Flights_per_annum <- as.factor(new_arules$No_of_Flights_per_annum)
new_arules$Percent_of_Flight_with_other_Airlines <-
CreateBucketsFWOA(new_arules$Percent_of_Flight_with_other_Airlines)
new_arules$Percent_of_Flight_with_other_Airlines <- as.factor(new_arules$Percent_of_Flight_with_other_Airlines)
new_arules$Eating_and_Drinking_at_Airport <- createBuckets(new_arules$Eating_and_Drinking_at_Airport)
new_arules$Eating_and_Drinking_at_Airport <- as.factor(new_arules$Eating_and_Drinking_at_Airport)
new_arules$Departure_Delay_in_Minutes <- createBuckets(new_arules$Departure_Delay_in_Minutes)
new_arules$Departure_Delay_in_Minutes <- as.factor(new_arules$Departure_Delay_in_Minutes)
new_arules$Arrival_Delay_in_Minutes <- createBuckets(new_arules$Arrival_Delay_in_Minutes)
new_arules$Arrival_Delay_in_Minutes <- as.factor(new_arules$Arrival_Delay_in_Minutes)
new_arules$Flight_time_in_minutes <- createBuckets(new_arules$Flight_time_in_minutes)
new_arules$Flight_time_in_minutes <- as.factor(new_arules$Flight_time_in_minutes)
new_arules$Flight_Distance <- createBuckets(new_arules$Flight_Distance)
new_arules$Flight_Distance <- as.factor(new_arules$Flight_Distance)
View(new_arules)
library(arules)

rulesetLow_SE <- apriori(new_arules, parameter = list(support = 0.1, confidence = 0.5), appearance =
list(rhs=c("Satisfaction=Low")))
inspect(rulesetLow_SE)
sorted_rules <- sort(rulesetLow_SE, decreasing=TRUE, by="lift")
inspect(sorted_rules)
plot(sorted_rules)

subrules2 <- head(rulesetLow_SE, n = 10, by = "lift")
plot(subrules2, method = "graph", engine = "htmlwidget")
```

```
#-------------------------------------------Support Vector Machine--------------------------------------

#Function for test set

createBuckets <- function(vec){
  q <- quantile(vec, c(0.3, 0.6), na.rm = "TRUE")
  vBuckets <- replicate(length(vec), "Average")
  vBuckets[vec <= q[1]] <- "Low"
  vBuckets[vec > q[2]] <- "High"
  return(vBuckets)
}

createBucketsPrice <- function(v) {
  vBucket <- replicate(length(v), "Average")
  vBucket[v>1] <- "High"
  vBucket[v<1] <- "Low"
  return(vBucket)
}

createBucketsFPA <- function(v) {
  vBucket <- replicate(length(v), "Average")
  vBucket[v>24] <- "High"
  vBucket[v<24] <- "Low"
  return(vBucket)
}

CreateBucketsFWOA <- function(v) {
  vBucket <- replicate(length(v), "Average")
  vBucket[v>6] <- "High"
  vBucket[v<6] <- "Low"
  return(vBucket)
}

CreateBucketsSDH <- function(v) {
  vBucket <- replicate(length(v), "Afternoon")
  vBucket[v>16] <- "Night"
  vBucket[v<8] <- "Morning"
  return(vBucket)
}

Training_set <- function(Airline1_df){
  Airline1_df$Satisfaction <- as.character(Airline1_df$Satisfaction)
  Airline1_df[,1] <- as.numeric(Airline1_df[,1])
  Airline1_df$lowSat <- ifelse(Airline1_df[,1]<3, 1, 0)
  randIndex <- sample(1:dim(Airline1_df)[1])   #creating a random sample of Index from the dataset
  cutpoint_2_3 <- floor(2*dim(Airline1_df)[1]/3)  #specifying a cut point in the data set index to segregate the dataset
  trainData <- Airline1_df[randIndex[1:cutpoint_2_3],]
  trainData <- Airline1_df[,c(-1,-15,-16,-17,-19,-21)]

  trainData$lowSat <- as.factor(trainData$lowSat)
```

```
  return(trainData)
}

#Function for test set
Test_set <- function(Airline1_df){
  Airline1_df$Satisfaction <- as.character(Airline1_df$Satisfaction)
  Airline1_df[,1] <- as.numeric(Airline1_df[,1])
  Airline1_df$lowSat <- ifelse(Airline1_df[,1]<3, 1, 0)
  randIndex <- sample(1:dim(Airline1_df)[1])   #creating a random sample of Index from the dataset
  cutpoint_2_3 <- floor(2*dim(Airline1_df)[1]/3)  #specifying a cut point in the data set index to segregate the dataset
  testData <- Airline1_df[randIndex[(cutpoint_2_3+1):dim(Airline1_df)[1]],]
  testData <- Airline1_df[,c(-1,-15,-16,-17,-19,-21)]

  testData$lowSat <- as.factor(testData$lowSat)


  return(testData)
}

#creating a function for confusion matrix
confusion_matrix <- function(test_set, pred_set){
  compLowSat<- data.frame(test_set, pred_set)
  colnames(compLowSat) <- c("test", "Pred")
  cfmat <- table(test=compLowSat$test, pred=compLowSat$Pred)
  return(cfmat)
}

#Function for error rate
error_rate <- function(cfmat){
  FP_plus_FN <- cfmat[2,"0"] + cfmat[1, "1"]                              #saving the sum of False positive and False negative
in one variable.
  FPFNTPTN <- cfmat[2,"0"] + cfmat[1, "1"] + cfmat[1,"0"] + cfmat[2,"1"]
  error_rt <- (FP_plus_FN/FPFNTPTN)*100
  return(error_rt)
}

#Function for Accuracy
Accuracy <- function(cfmat){
  Tp_plus_TN <- cfmat[1,"0"] + cfmat[2,"1"]
  FPFNTPTN <- cfmat[2,"0"] + cfmat[1, "1"] + cfmat[1,"0"] + cfmat[2,"1"]
  accuracy <- (Tp_plus_TN/FPFNTPTN)*100
  return(accuracy)
}

-------------------------------------------------------------South east SVM--------------------------------------------

#creating a variable for low satisfaction
trainData <- Training_set(Airline1_df)
testData <- Test_set(Airline1_df)

trainData$Price_Sensitivity <- createBucketsPrice(trainData$Price_Sensitivity)
trainData$Price_Sensitivity <- as.factor(trainData$Price_Sensitivity)
trainData$Age <- createBuckets(trainData$Age)
```

```
trainData$Age <- as.factor(trainData$Age)
trainData$No_of_other_Loyalty_Cards <- createBuckets(trainData$No_of_other_Loyalty_Cards)
trainData$No_of_other_Loyalty_Cards <- as.factor(trainData$No_of_other_Loyalty_Cards)
trainData$Shopping_Amount_at_Airport <- createBuckets(trainData$Shopping_Amount_at_Airport)
trainData$Shopping_Amount_at_Airport <- as.factor(trainData$Shopping_Amount_at_Airport)
trainData$No_of_Flights_per_annum <- createBucketsFPA(trainData$No_of_Flights_per_annum)
trainData$No_of_Flights_per_annum <- as.factor(trainData$No_of_Flights_per_annum)
trainData$Percent_of_Flight_with_other_Airlines <-
CreateBucketsFWOA(trainData$Percent_of_Flight_with_other_Airlines)
trainData$Percent_of_Flight_with_other_Airlines <- as.factor(trainData$Percent_of_Flight_with_other_Airlines)
trainData$Eating_and_Drinking_at_Airport <- createBuckets(trainData$Eating_and_Drinking_at_Airport)
trainData$Eating_and_Drinking_at_Airport <- as.factor(trainData$Eating_and_Drinking_at_Airport)
trainData$Departure_Delay_in_Minutes <- createBuckets(trainData$Departure_Delay_in_Minutes)
trainData$Departure_Delay_in_Minutes <- as.factor(trainData$Departure_Delay_in_Minutes)
trainData$Arrival_Delay_in_Minutes <- createBuckets(trainData$Arrival_Delay_in_Minutes)
trainData$Arrival_Delay_in_Minutes <- as.factor(trainData$Arrival_Delay_in_Minutes)
trainData$Flight_time_in_minutes <- createBuckets(trainData$Flight_time_in_minutes)
trainData$Flight_time_in_minutes <- as.factor(trainData$Flight_time_in_minutes)
trainData$Flight_Distance <- createBuckets(trainData$Flight_Distance)
trainData$Flight_Distance <- as.factor(trainData$Flight_Distance)
trainData$Scheduled_Departure_Hour <- CreateBucketsSDH(trainData$Scheduled_Departure_Hour)
trainData$Scheduled_Departure_Hour <- as.factor(trainData$Scheduled_Departure_Hour)

testData$Price_Sensitivity <- createBucketsPrice(testData$Price_Sensitivity)
testData$Price_Sensitivity <- as.factor(testData$Price_Sensitivity)
testData$Age <- createBuckets(testData$Age)
testData$Age <- as.factor(testData$Age)
testData$No_of_other_Loyalty_Cards <- createBuckets(testData$No_of_other_Loyalty_Cards)
testData$No_of_other_Loyalty_Cards <- as.factor(testData$No_of_other_Loyalty_Cards)
testData$Shopping_Amount_at_Airport <- createBuckets(testData$Shopping_Amount_at_Airport)
testData$Shopping_Amount_at_Airport <- as.factor(testData$Shopping_Amount_at_Airport)
testData$No_of_Flights_per_annum <- createBucketsFPA(testData$No_of_Flights_per_annum)
testData$No_of_Flights_per_annum <- as.factor(testData$No_of_Flights_per_annum)
testData$Percent_of_Flight_with_other_Airlines <-
CreateBucketsFWOA(testData$Percent_of_Flight_with_other_Airlines)
testData$Percent_of_Flight_with_other_Airlines <- as.factor(testData$Percent_of_Flight_with_other_Airlines)
testData$Eating_and_Drinking_at_Airport <- createBuckets(testData$Eating_and_Drinking_at_Airport)
testData$Eating_and_Drinking_at_Airport <- as.factor(testData$Eating_and_Drinking_at_Airport)
testData$Departure_Delay_in_Minutes <- createBuckets(testData$Departure_Delay_in_Minutes)
testData$Departure_Delay_in_Minutes <- as.factor(testData$Departure_Delay_in_Minutes)
testData$Arrival_Delay_in_Minutes <- createBuckets(testData$Arrival_Delay_in_Minutes)
testData$Arrival_Delay_in_Minutes <- as.factor(testData$Arrival_Delay_in_Minutes)
testData$Flight_time_in_minutes <- createBuckets(testData$Flight_time_in_minutes)
testData$Flight_time_in_minutes <- as.factor(testData$Flight_time_in_minutes)
testData$Flight_Distance <- createBuckets(testData$Flight_Distance)
testData$Flight_Distance <- as.factor(testData$Flight_Distance)
testData$Scheduled_Departure_Hour <- CreateBucketsSDH(testData$Scheduled_Departure_Hour)
testData$Scheduled_Departure_Hour <- as.factor(testData$Scheduled_Departure_Hour)




#Significant variables
#Airline_Status, Age, Gender, Price_Sensitivity, Year_of_First_Flight, Flight_cancelled
```

```
#No_of_Flights_per_annum, Type_of_Travel,Eating_and_Drinking_at_Airport, Arrival_Delay_greater_5_Mins

#For Complete Model
svmModOutput_all <- ksvm(lowSat~., data=trainData, kernel= "rbfdot", kpar = "automatic", C = 50, cross = 3,
prob.model = TRUE)
svmModOutput_all
svmPred <- predict(svmModOutput_all, testData)
str(svmPred)
head(svmPred)
#Airline_Status+Age+Gender+No_of_Flights_per_annum+Type_of_Travel+Arrival_Delay_greater_5_Mins
#Confusion matrix
conf_mat <- confusion_matrix(testData$lowSat, svmPred)
conf_mat
fourfoldplot(conf_mat, color = c("#CC6666", "#99CC99"),
        conf.level = 0, margin = 1, main = "Confusion Matrix for SouthEast Airlines - All Attributes")
error_rate(conf_mat)
Accuracy(conf_mat)
hist(alpha(svmModOutput_all)[[1]])

M_1 <- fit(lowSat~., data=trainData, model="svm", kpar=list(sigma=0.10), C=2)
VariableImportance_1 <- Importance(M_1, data=trainData, method = "sensv")

L_1=list(runs=1,sen=t(VariableImportance_1$imp),sresponses=VariableImportance_1$sresponses)
mgraph(L_1,graph="IMP",leg=names(trainData),col="gray",Grid=10)

#For Significant variables
svmModOutput <-
ksvm(lowSat~Type_of_Travel+Departure_Delay_in_Minutes+Arrival_Delay_in_Minutes+Percent_of_Flight_with_othe
r_Airlines+Eating_and_Drinking_at_Airport+No_of_other_Loyalty_Cards, data=trainData, kernel= "rbfdot", kpar =
"automatic", C = 50, cross = 3, prob.model = TRUE)
svmModOutput
svmPred_reg <- predict(svmModOutput, testData)
conf_mat_reg <- confusion_matrix(testData$lowSat, svmPred_reg)
conf_mat_reg
fourfoldplot(conf_mat_reg, color = c("#CC6666", "#99CC99"),
        conf.level = 0, margin = 1, main = "Confusion Matrix for SouthEast Airlines - Significant Attributes")
Accuracy(conf_mat_reg)

#Significant variables
#Type_of_Travel+Departure_Delay_in_Minutes+Arrival_Delay_in_Minutes+Percent_of_Flight_with_other_Airlines+E
ating_and_Drinking_at_Airport+No_of_other_Loyalty_Cards
#Accuracy - 84.88044




#------------------------------------------For FlyFast Airlines----------------------------------------
trainData_1 <- Training_set(Airline2_df)
testData_1 <- Test_set(Airline2_df)

trainData_1$Price_Sensitivity <- createBucketsPrice(trainData_1$Price_Sensitivity)
trainData_1$Price_Sensitivity <- as.factor(trainData_1$Price_Sensitivity)
trainData_1$Age <- createBuckets(trainData_1$Age)
```

```
trainData_1$Age <- as.factor(trainData_1$Age)
trainData_1$No_of_other_Loyalty_Cards <- createBuckets(trainData_1$No_of_other_Loyalty_Cards)
trainData_1$No_of_other_Loyalty_Cards <- as.factor(trainData_1$No_of_other_Loyalty_Cards)
trainData_1$Shopping_Amount_at_Airport <- createBuckets(trainData_1$Shopping_Amount_at_Airport)
trainData_1$Shopping_Amount_at_Airport <- as.factor(trainData_1$Shopping_Amount_at_Airport)
trainData_1$No_of_Flights_per_annum <- createBucketsFPA(trainData_1$No_of_Flights_per_annum)
trainData_1$No_of_Flights_per_annum <- as.factor(trainData_1$No_of_Flights_per_annum)
trainData_1$Percent_of_Flight_with_other_Airlines <-
CreateBucketsFWOA(trainData_1$Percent_of_Flight_with_other_Airlines)
trainData_1$Percent_of_Flight_with_other_Airlines <- as.factor(trainData_1$Percent_of_Flight_with_other_Airlines)
trainData_1$Eating_and_Drinking_at_Airport <- createBuckets(trainData_1$Eating_and_Drinking_at_Airport)
trainData_1$Eating_and_Drinking_at_Airport <- as.factor(trainData_1$Eating_and_Drinking_at_Airport)
trainData_1$Departure_Delay_in_Minutes <- createBuckets(trainData_1$Departure_Delay_in_Minutes)
trainData_1$Departure_Delay_in_Minutes <- as.factor(trainData_1$Departure_Delay_in_Minutes)
trainData_1$Arrival_Delay_in_Minutes <- createBuckets(trainData_1$Arrival_Delay_in_Minutes)
trainData_1$Arrival_Delay_in_Minutes <- as.factor(trainData_1$Arrival_Delay_in_Minutes)
trainData_1$Flight_time_in_minutes <- createBuckets(trainData_1$Flight_time_in_minutes)
trainData_1$Flight_time_in_minutes <- as.factor(trainData_1$Flight_time_in_minutes)
trainData_1$Flight_Distance <- createBuckets(trainData_1$Flight_Distance)
trainData_1$Flight_Distance <- as.factor(trainData_1$Flight_Distance)
trainData_1$Scheduled_Departure_Hour <- CreateBucketsSDH(trainData_1$Scheduled_Departure_Hour)
trainData_1$Scheduled_Departure_Hour <- as.factor(trainData_1$Scheduled_Departure_Hour)

testData_1$Price_Sensitivity <- createBucketsPrice(testData_1$Price_Sensitivity)
testData_1$Price_Sensitivity <- as.factor(testData_1$Price_Sensitivity)
testData_1$Age <- createBuckets(testData_1$Age)
testData_1$Age <- as.factor(testData_1$Age)
testData_1$No_of_other_Loyalty_Cards <- createBuckets(testData_1$No_of_other_Loyalty_Cards)
testData_1$No_of_other_Loyalty_Cards <- as.factor(testData_1$No_of_other_Loyalty_Cards)
testData_1$Shopping_Amount_at_Airport <- createBuckets(testData_1$Shopping_Amount_at_Airport)
testData_1$Shopping_Amount_at_Airport <- as.factor(testData_1$Shopping_Amount_at_Airport)
testData_1$No_of_Flights_per_annum <- createBucketsFPA(testData_1$No_of_Flights_per_annum)
testData_1$No_of_Flights_per_annum <- as.factor(testData_1$No_of_Flights_per_annum)
testData_1$Percent_of_Flight_with_other_Airlines <-
CreateBucketsFWOA(testData_1$Percent_of_Flight_with_other_Airlines)
testData_1$Percent_of_Flight_with_other_Airlines <- as.factor(testData_1$Percent_of_Flight_with_other_Airlines)
testData_1$Eating_and_Drinking_at_Airport <- createBuckets(testData_1$Eating_and_Drinking_at_Airport)
testData_1$Eating_and_Drinking_at_Airport <- as.factor(testData_1$Eating_and_Drinking_at_Airport)
testData_1$Departure_Delay_in_Minutes <- createBuckets(testData_1$Departure_Delay_in_Minutes)
testData_1$Departure_Delay_in_Minutes <- as.factor(testData_1$Departure_Delay_in_Minutes)
testData_1$Arrival_Delay_in_Minutes <- createBuckets(testData_1$Arrival_Delay_in_Minutes)
testData_1$Arrival_Delay_in_Minutes <- as.factor(testData_1$Arrival_Delay_in_Minutes)
testData_1$Flight_time_in_minutes <- createBuckets(testData_1$Flight_time_in_minutes)
testData_1$Flight_time_in_minutes <- as.factor(testData_1$Flight_time_in_minutes)
testData_1$Flight_Distance <- createBuckets(testData_1$Flight_Distance)
testData_1$Flight_Distance <- as.factor(testData_1$Flight_Distance)
testData_1$Scheduled_Departure_Hour <- CreateBucketsSDH(testData_1$Scheduled_Departure_Hour)
testData_1$Scheduled_Departure_Hour <- as.factor(testData_1$Scheduled_Departure_Hour)



#For the complete data set
svmModOutput_all_1 <- ksvm(lowSat~., data=trainData_1, kernel= "rbfdot", kpar = "automatic", C = 50, cross = 3,
prob.model = TRUE)
```

```
svmModOutput_all_1
hist(alpha(svmModOutput_all_1)[[1]])
svmPred_1 <- predict(svmModOutput_all_1, testData_1)
conf_mat_1 <- confusion_matrix(testData_1$lowSat, svmPred_1)
conf_mat_1
fourfoldplot(conf_mat_1, color = c("#CC6666", "#99CC99"),
        conf.level = 0, margin = 1, main = "Confusion Matrix for Flyfast Airlines - All Attributes")
Accuracy(conf_mat_1)
#Airline_Status +Gender + No_of_Flights_per_annum+Type_of_Travel+ Flight_cancelled+
Arrival_Delay_greater_5_Mins

M <- fit(lowSat~., data=trainData_1, model="svm", kpar=list(sigma=0.10), C=2)
VariableImportance <- Importance(M, data=trainData_1, method = "sensv")

L=list(runs=1,sen=t(VariableImportance$imp),sresponses=VariableImportance$sresponses)
mgraph(L,graph="IMP",leg=names(trainData_1),col="gray",Grid=10)

#For Significant variables
svmModOutput_1 <-
ksvm(lowSat~Type_of_Travel+Flight_time_in_minutes+Shopping_Amount_at_Airport+Airline_Status+Departure_Del
ay_in_Minutes+Arrival_Delay_in_Minutes+Eating_and_Drinking_at_Airport, data=trainData_1, kernel= "rbfdot", kpar
= "automatic", C = 50, cross = 3, prob.model = TRUE)
svmModOutput_1
hist(alpha(svmModOutput_1)[[1]])
svmPred_reg_1 <- predict(svmModOutput_1, testData_1)
conf_mat_reg_1 <- confusion_matrix(testData_1$lowSat, svmPred_reg_1)
conf_mat_reg_1
fourfoldplot(conf_mat_reg_1, color = c("#CC6666", "#99CC99"),
        conf.level = 0, margin = 1, main = "Confusion Matrix for Flyfast Airlines - Significant Attributes")
Accuracy(conf_mat_reg_1)
```

.