****

**IST 687 Applied Data Science**

**Analysis of Airline Survey dataset**

[Data Cleaning 3](#_Toc41162872)

[Business Questions 5](#_Toc41162873)

[Narrowing down the data set 15](#_Toc41162874)

[Linear Modelling 19](#_Toc41162875)

[Association Rules 22](#_Toc41162876)

[Support Vector Machine 37](#_Toc41162877)

[Decision Tree 40](#_Toc41162878)

[Validation 43](#_Toc41162879)

[Actionable Insights 44](#_Toc41162880)

# Data Cleaning

The dataset for this project consisted of 129880 observations and 26 attributes.

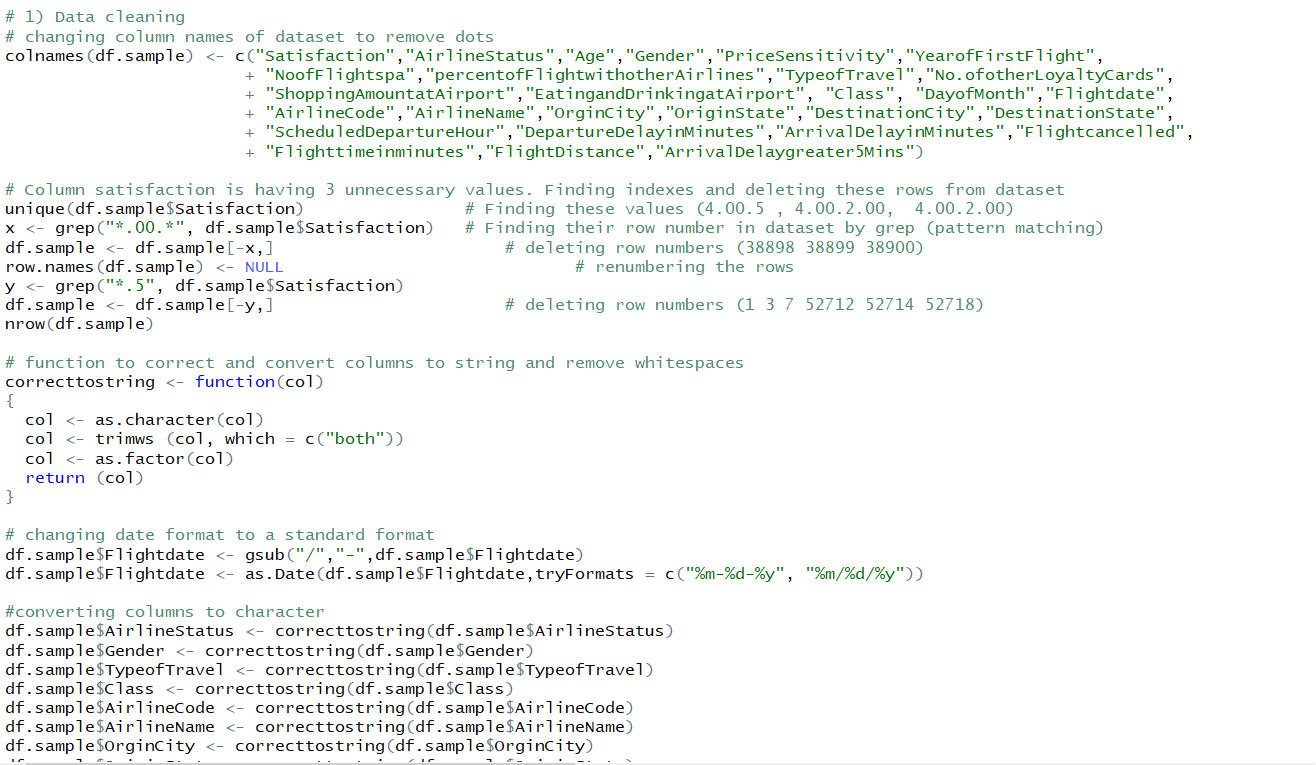
Our approach included detecting and correcting inaccurate records to ease the process of analysis. Data cleaning is necessary before we can explore and create models. We focused on narrowing the data to the airlines with lower median value of customer satisfaction.

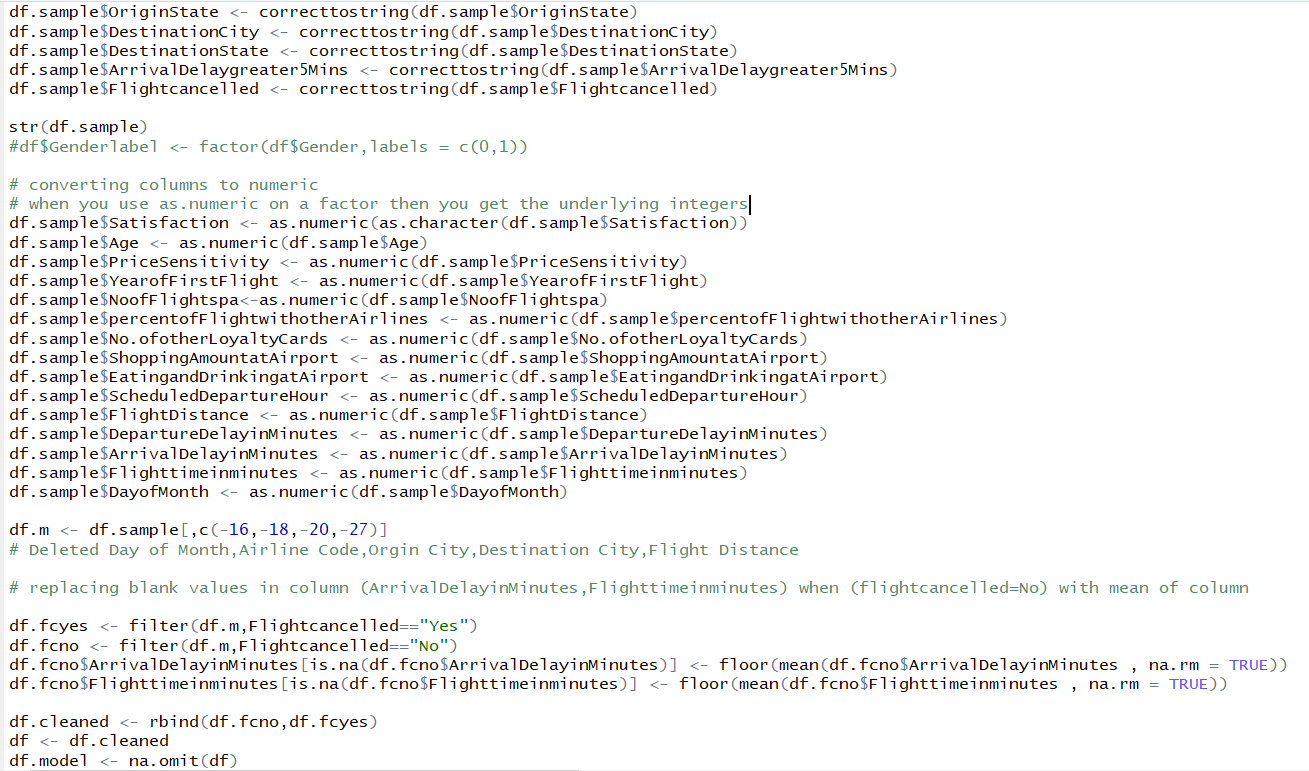
We started the data munging process by altering the names of the dataset columns so that the dots can be removed from column names, deleted values with decimals and then re-numbered the rows. Then we converted the columns to characters and numeric to ease the application of functions on each of them. The date format of flight date column was inconsistent so we changed the format of this column to a standard date format.

When flight cancelled (= yes), there were NA values in 2 columns (ArrivalDelayinMinutes, Flighttimeinminutes), we dropped or omitted these rows (2400) from the dataset.

When flight cancelled (= no), there were NA values in 2 columns (ArrivalDelayinMinutes, Flighttimeinminutes), we changed the NA values to the mean of the column in order to keep the distribution of the column as same.

**Code**:



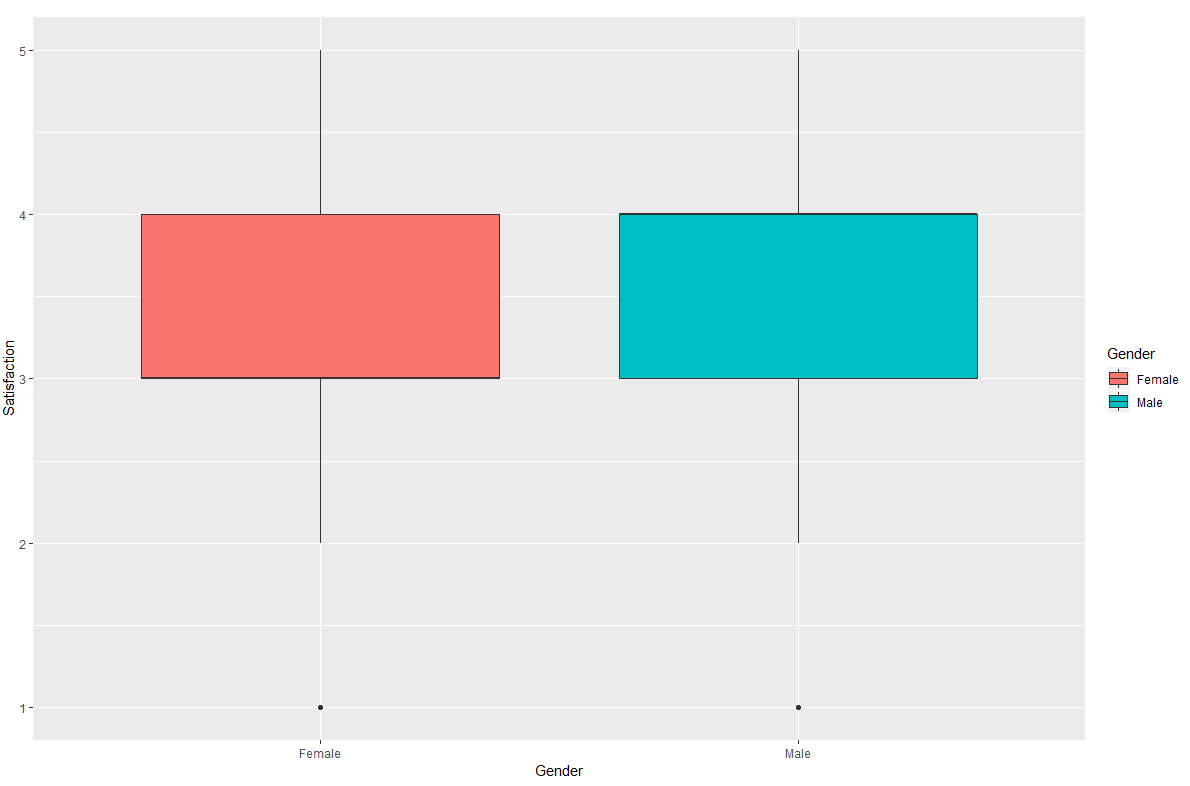


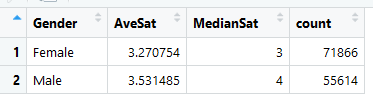


# Business Questions

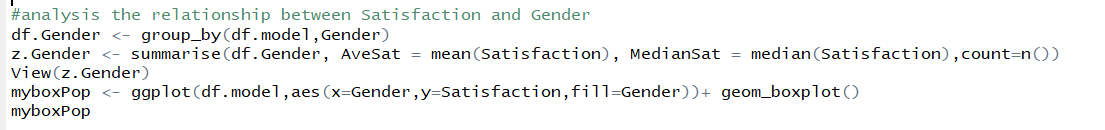
1. **Which gender either male or female has the higher customer satisfaction rating value?**

On plotting a box plot between gender and satisfaction, we observed that the median satisfaction value of males is higher than that of female.

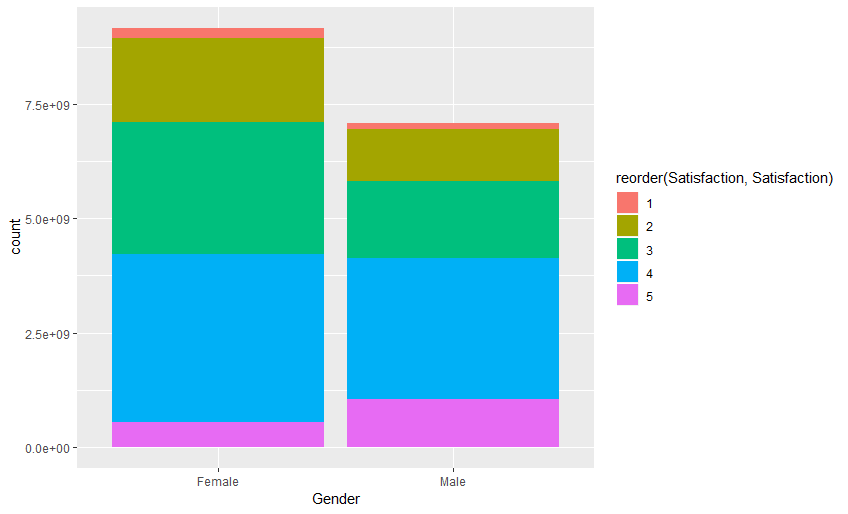




**Code**:

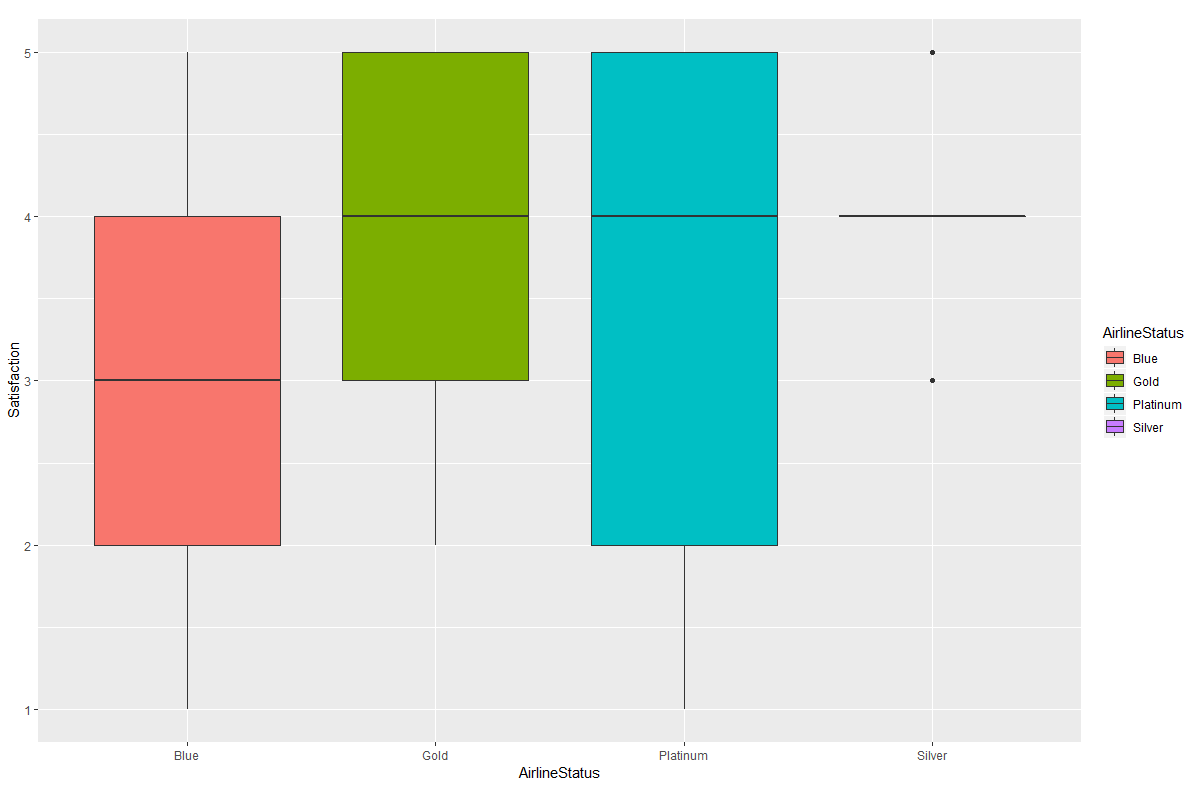


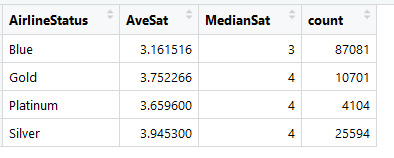
Other than this, we also plotted bar chart for Gender with total count and filled it with satisfaction. It shows females are giving more low ratings of (1, 2, 3) compared to male.



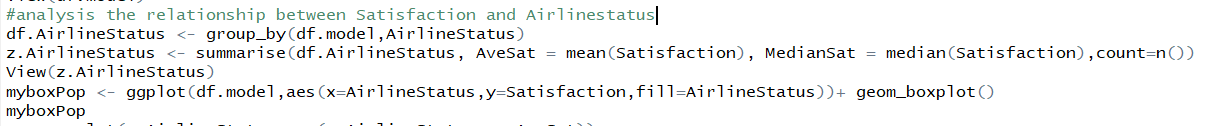
1. **Which type of airline status has the highest customer satisfaction rating?**

We plotted a boxplot between Airlinestatus and customer satisfaction and on analysis of this plot we found that the gold, platinum and silver airline statuses have highest median value of satisfaction while blue airline has the lowest.

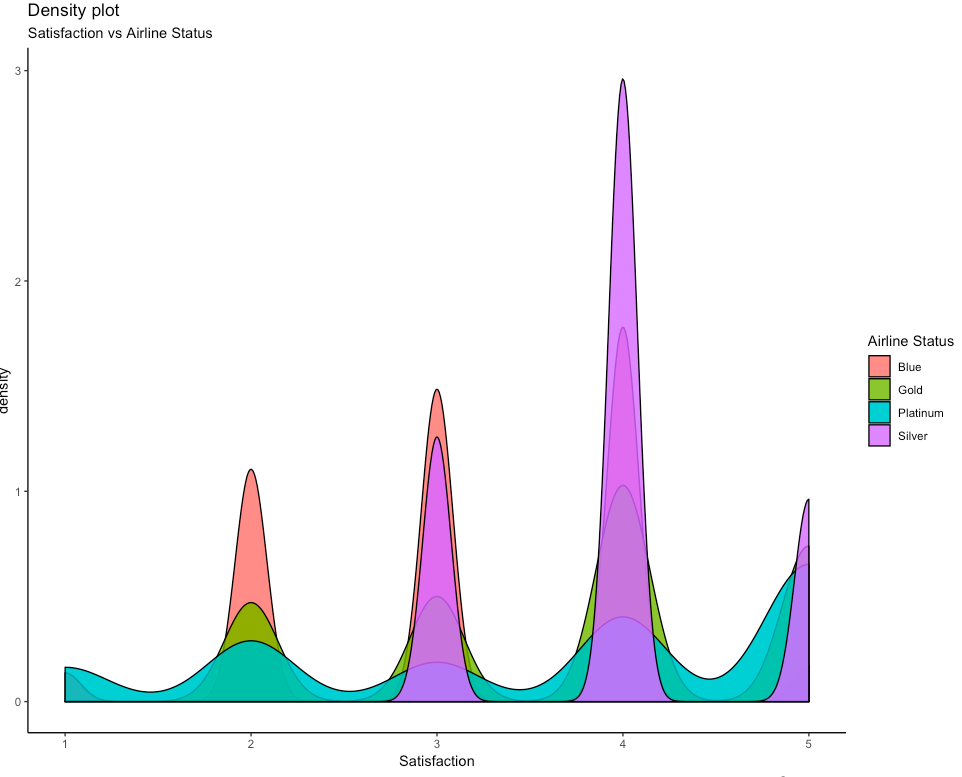




**Code**:

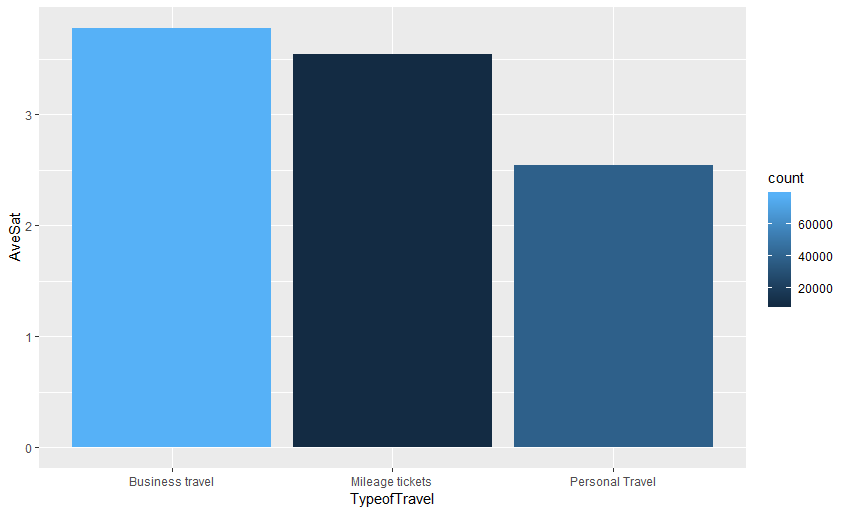


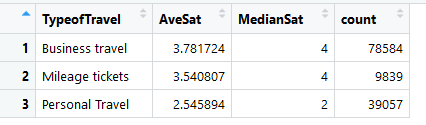
Other than this, we also plotted a density plot for different airline status and observed that for airline status as blue, customer satisfaction is spread in ratings of (1, 2, and 3)



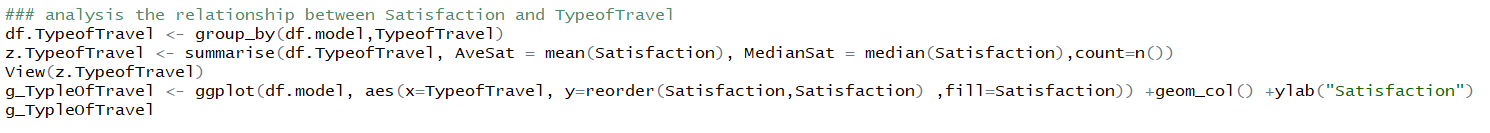
1. **Which travel types have the highest and lowest customer satisfaction rating**?

On plotting a bar chart between type of travel and satisfaction, we can conclude that business and mileage travel type has highest average customer satisfaction rating. While, personal travel type has the lowest average customer satisfaction rating.

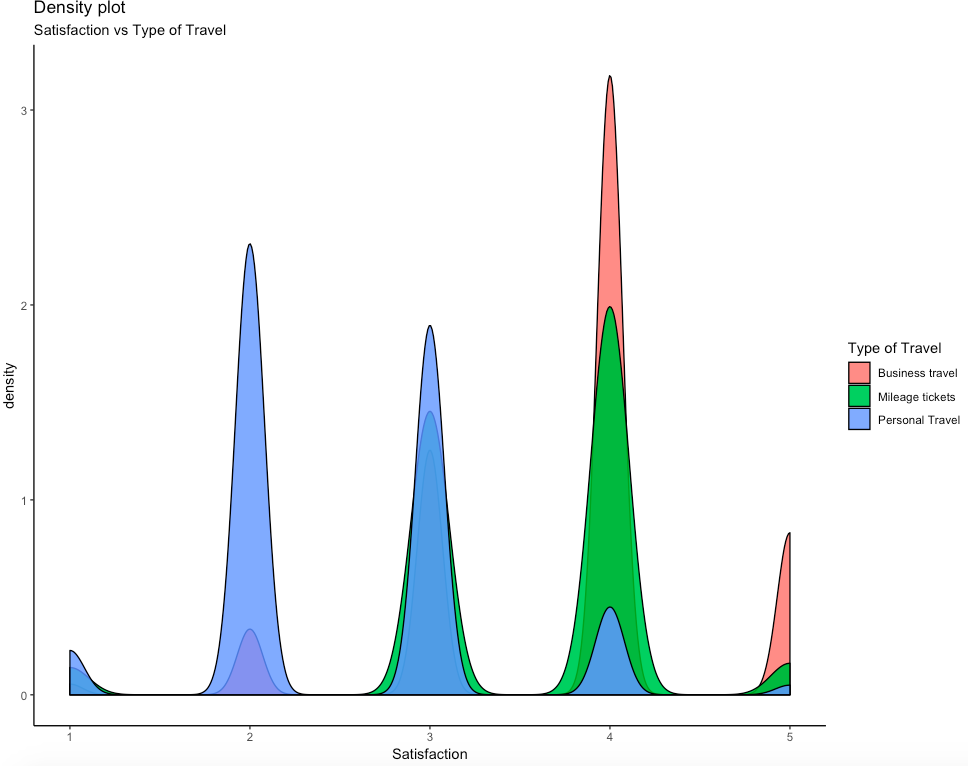




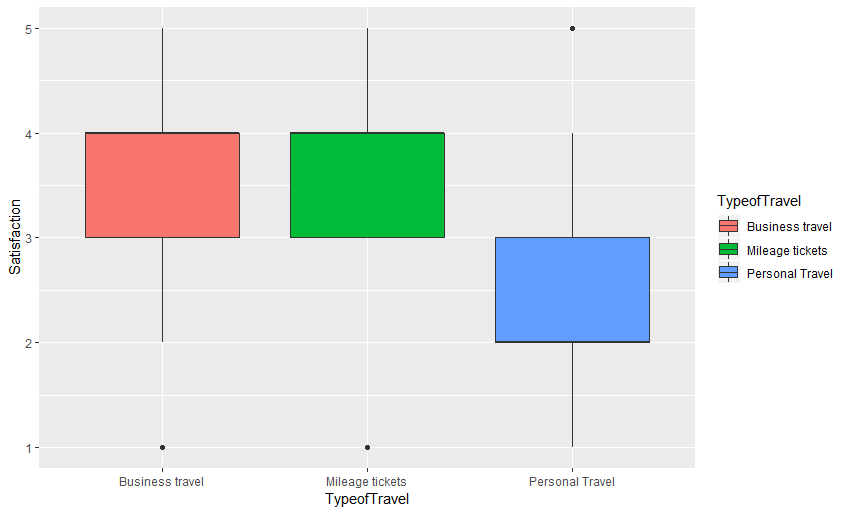
**Code**:



Other than this, we also plotted a density plot for different type of travel and observed that for type of travel as personal, customer satisfaction is mostly spread in the ratings of (1, 2, and 3)

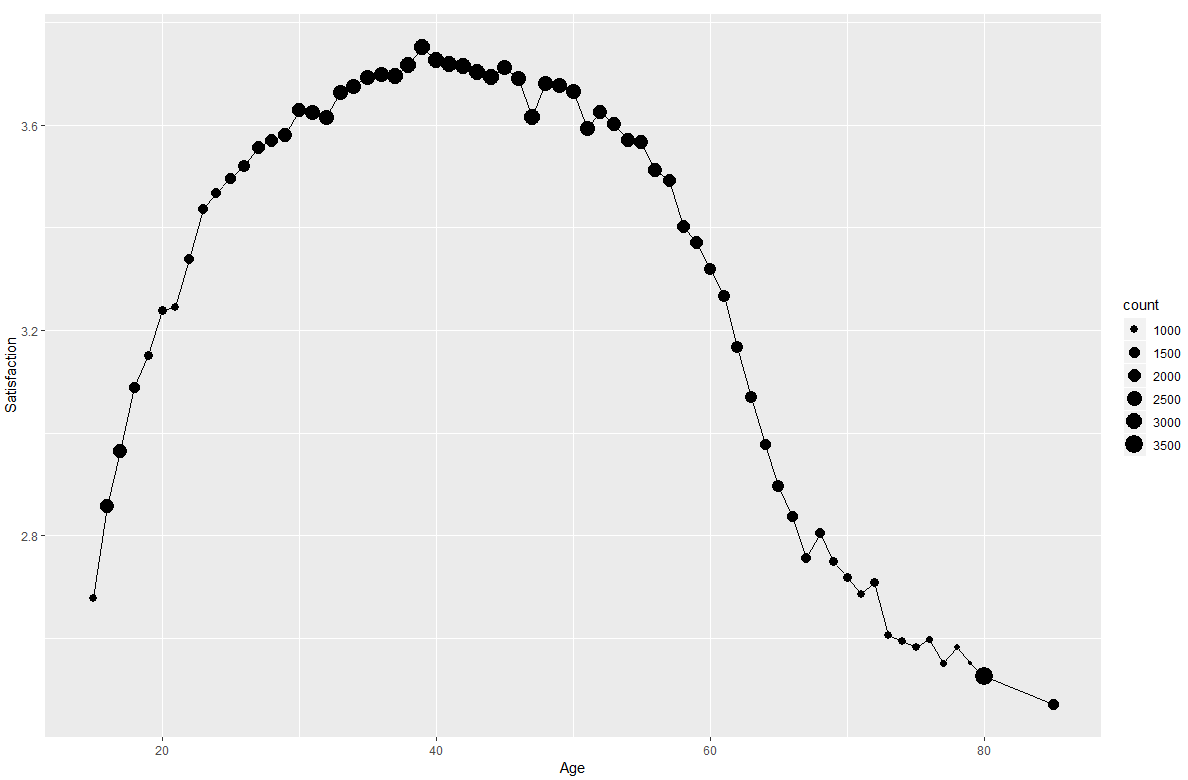


On plotting boxplot, we observed that for personal type of travel median satisfaction rating is low compared to business and mileage tickets.

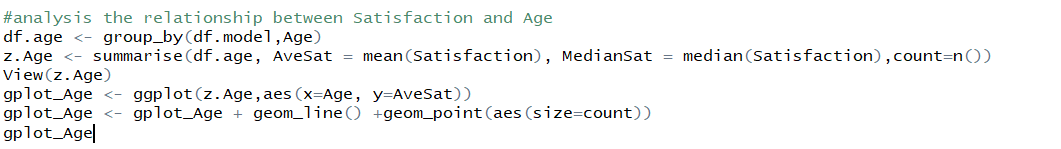


1. **Which age ranges of customers have the highest and lowest customer satisfaction rating?**

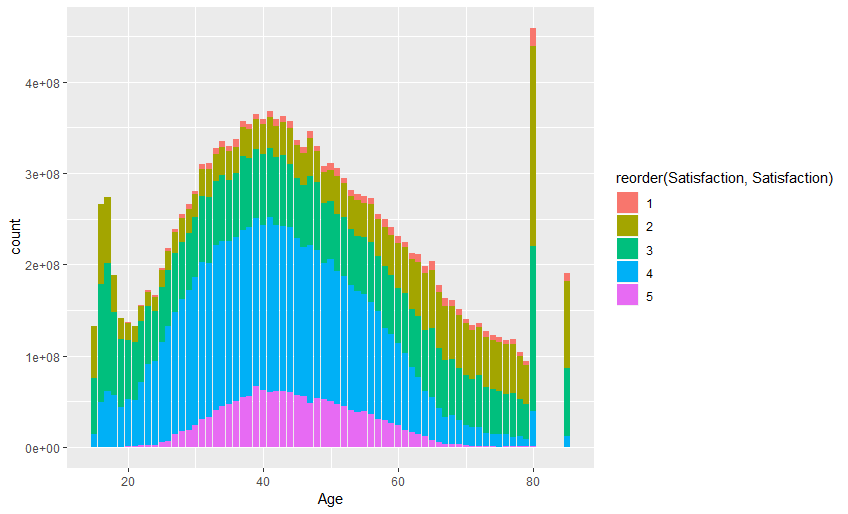
On plotting a scatterplot between ages and average satisfaction, we observed that the average satisfaction increases till the age of 40 and then it starts decreasing with age. Especially, for age group 60-85 average satisfaction is less.



**Code**:

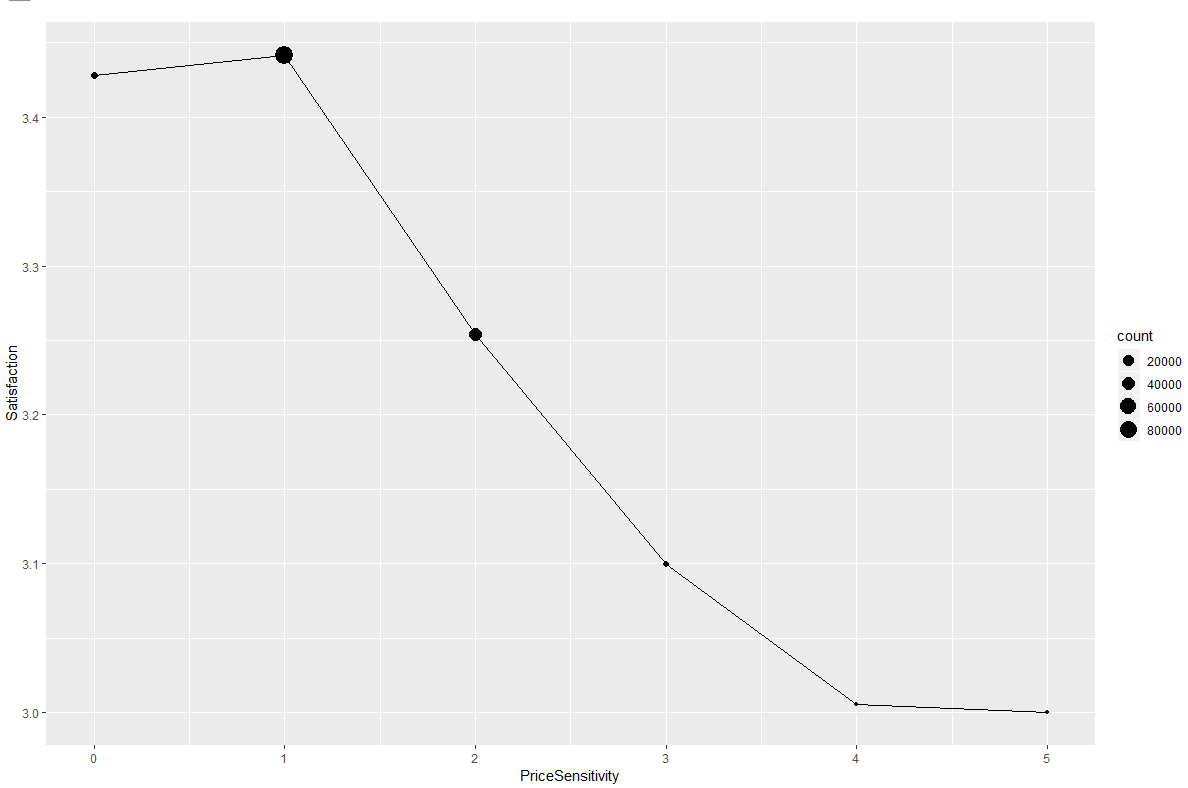


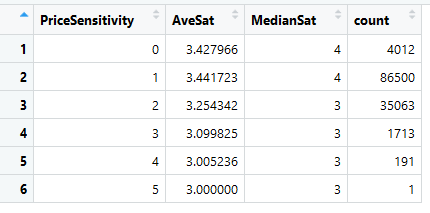
Other than this, we have also plotted a bar chart of Age with total count and filled it with satisfaction. It can be seen that people in age group (60-85) are mostly giving ratings of 1,2 and 3.



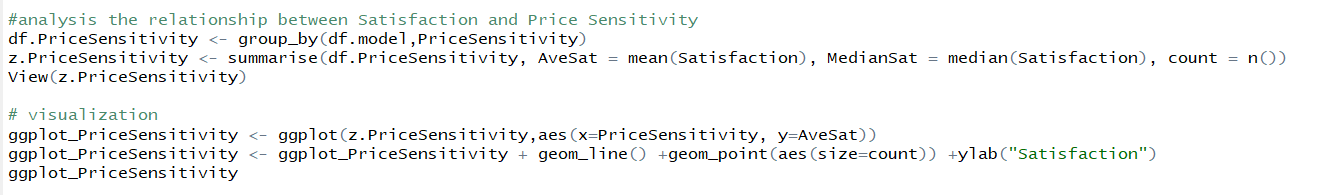
1. **How is price sensitivity related to customer satisfaction?**

On plotting a scatterplot between price sensitivity and customer satisfaction, we observed that the customer satisfaction rating corresponding to price sensitivity of 1 is highest and that of 5 is lowest. We can conclude that if price sensitivity and average customer satisfaction is inversely related.



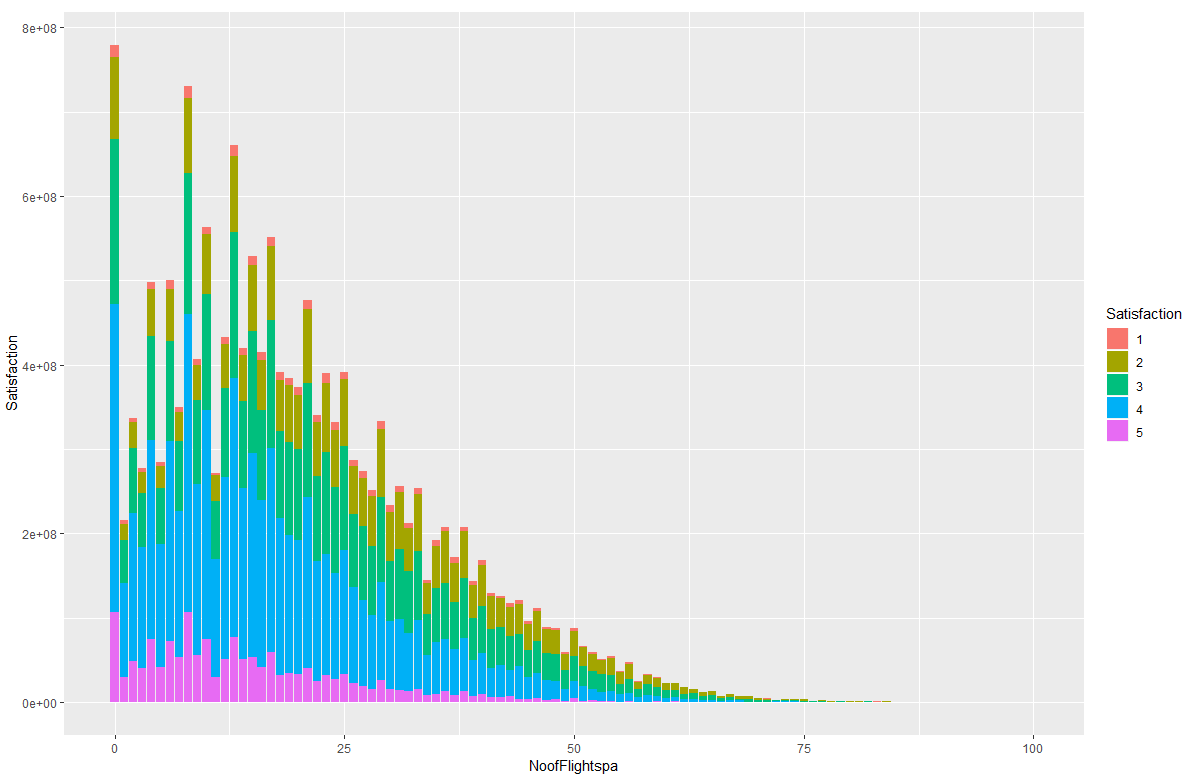


Code:

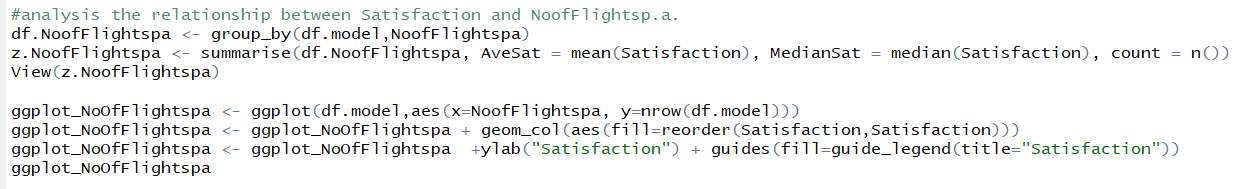


**6. How does number of flights per annum alter the customer satisfaction rating?**

On plotting a barchart between number of flights per annum and customer satisfaction, we observed that when the number of flights per annum decreases customer satisfaction increases.



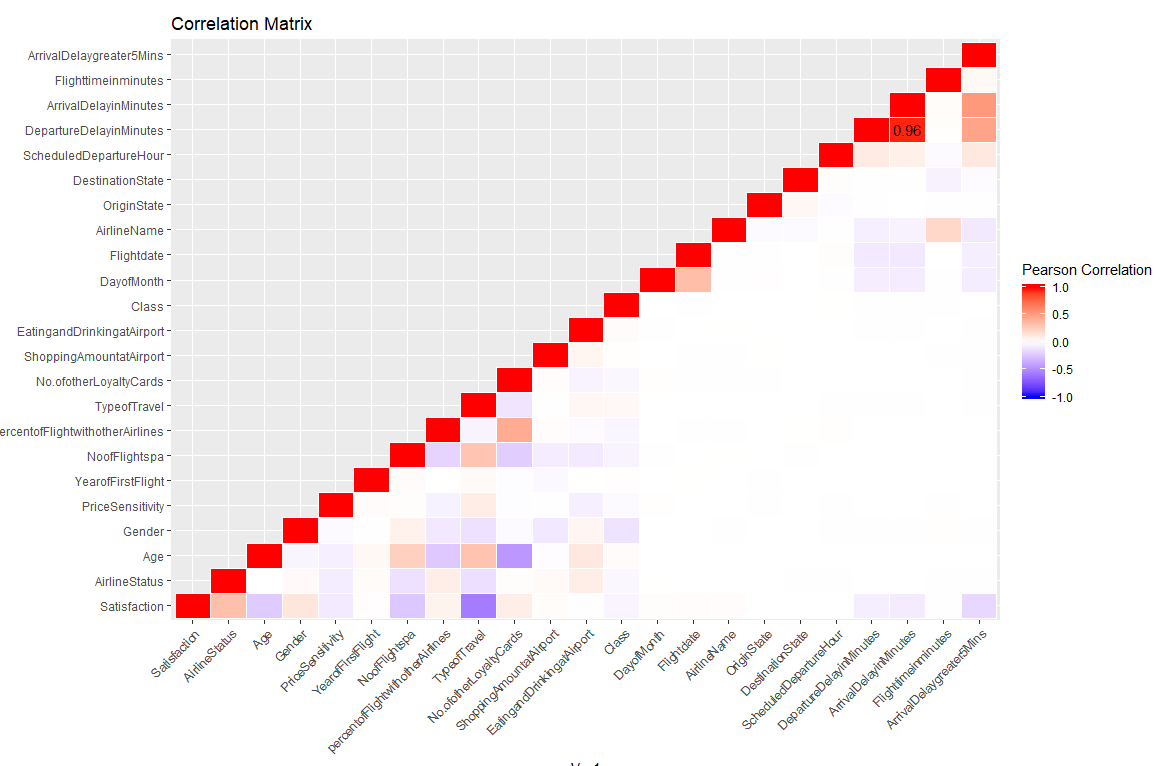
**Code:**



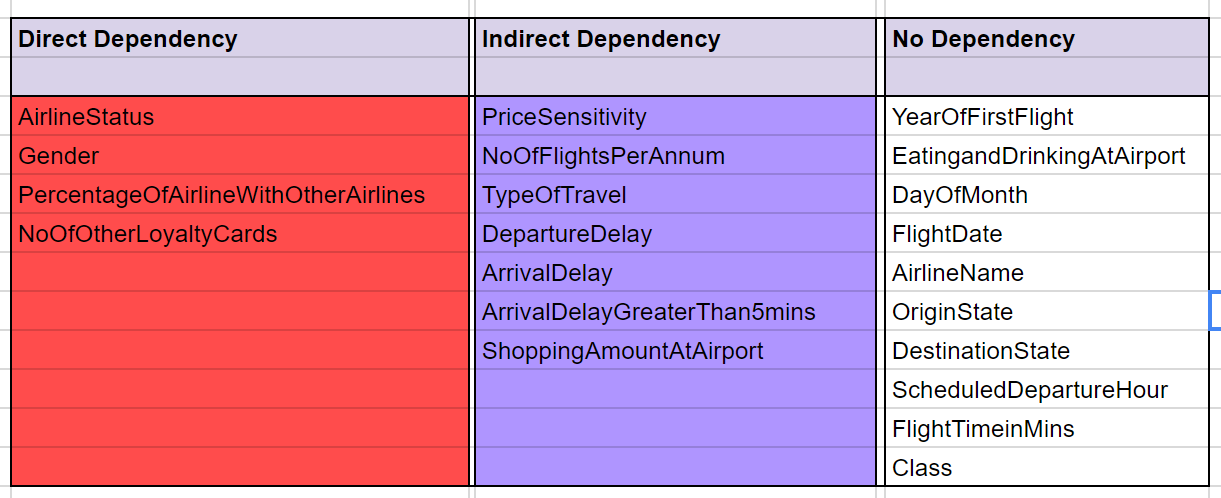
# Narrowing down the data set

**Correlation Matrix**

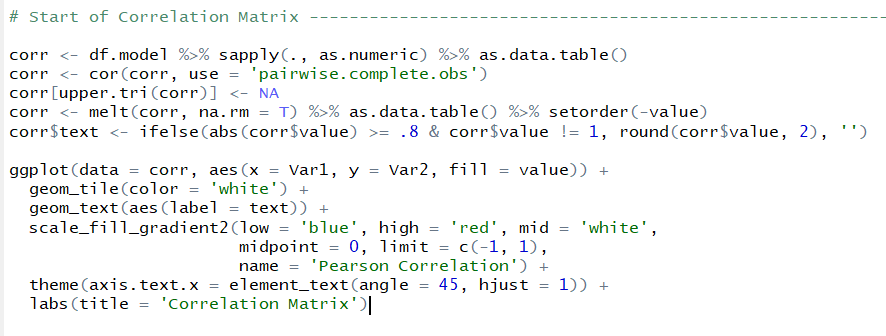
Correlation matrix is used to investigate dependencies between multiple variables at the same time. It results in a table of correlation coefficients between each variables. Correlation coefficient measures the percentage of fluctuation in one variable that can be explained by another variable. A correlation of 1 means the variables move in perfect unison, a correlation of -1 means the variables move in the complete opposite direction, and a correlation of 0 means there is no relationship at all between the two variables.



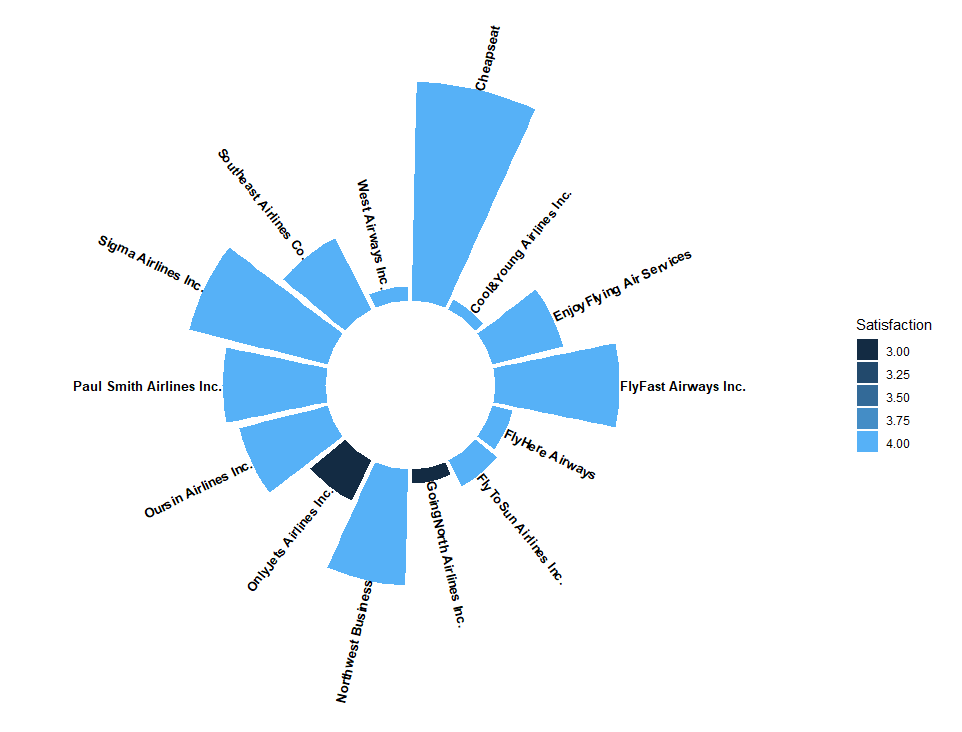
From the below plot, we can observe the below dependencies and relationship between variables with customer satisfaction,



**Code**:

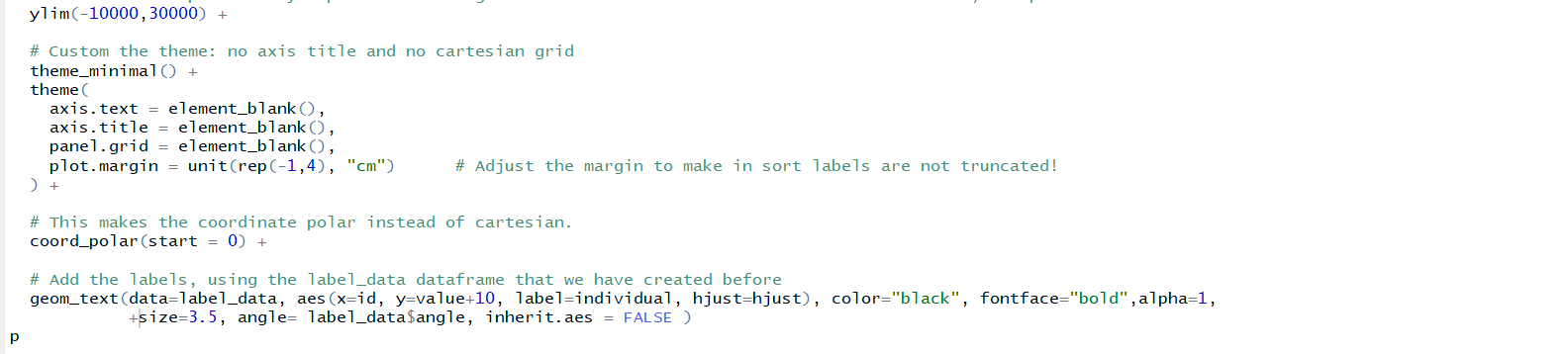


After analyzing and visualizing the dataset, we got a clear idea about the data. Next we moved on focusing on two particular airlines namely OnlyJetInc airlines and GoingNorth airlines because the median customer satisfaction for both these airlines was low as compared to other airlines as can be seen in below circular boxplot. Its scaling is as per the count of each airline and color fill is as per the median satisfaction value. OnlyJet and GoingNorth has lowest median satisfaction value of 3 & hence different in color from the rest.



**Code:**





# Linear Modelling

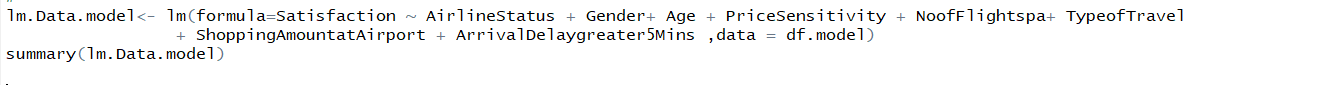
Linear modelling is used to predict the value of an outcome variable Y based on one or more input predictor variables X. Here, we can establish a linear relationship between the two variables where a continuous variable is modelled as a mathematical function of one or more variable which are known.

On the entire dataset, we performed forward linear modelling and added one-by-one variable and kept on checking increasing R square value. Finally, after creating 22 models we got 8 out of 28 columns which were significant and impacting (+ve and –ve manner) customer satisfaction column.

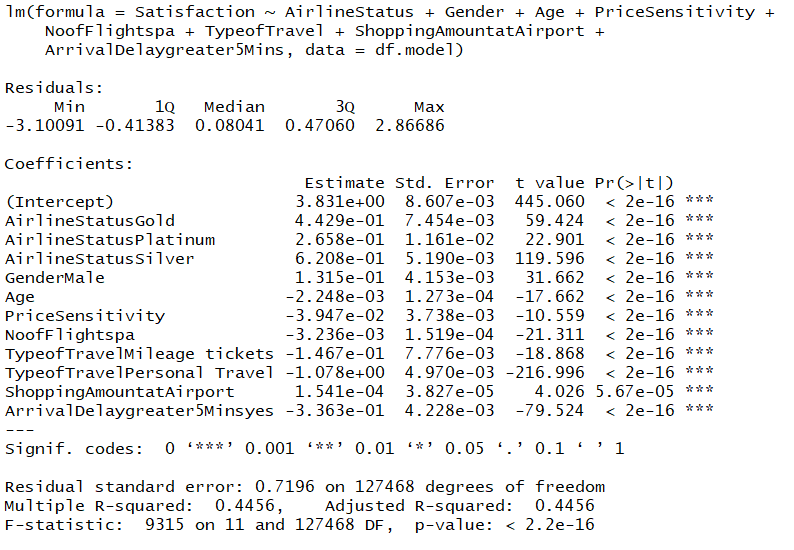
* Airline Status
* Gender
* Age
* Price sensitivity
* Number of flights per annum
* Type of travel
* Shopping Amount at airport
* Arrival delay greater than 5 minutes

The above factors are considered to be independent variables and customer satisfaction will be the dependent variable.

**Code:**



**Output:**



We observe that the R Squared value is 0.4456 whiich means that a combination of all the seven factors account for 44% of the customer satisfaction rating.

**Linear Modeling on OnlyJets Airlines:**

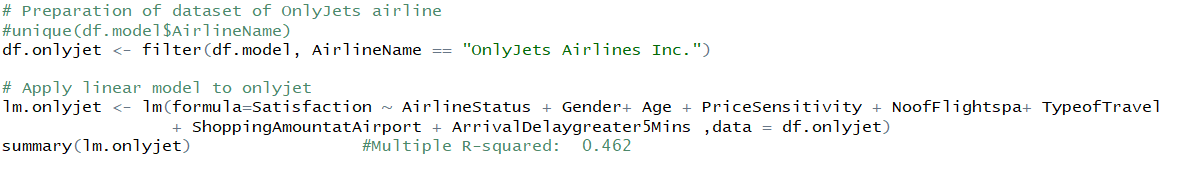
Then, after due deliberation we conducted linear modelling on two airlines , the first one being Only Jet customers with customer satisfaction ratings based on below factors :

* Airline Status
* Gender
* Age
* Price sensitivity
* Number of flights per annum
* Type of travel
* Shopping Amount at airport
* Arrival delay greater than 5 minutes

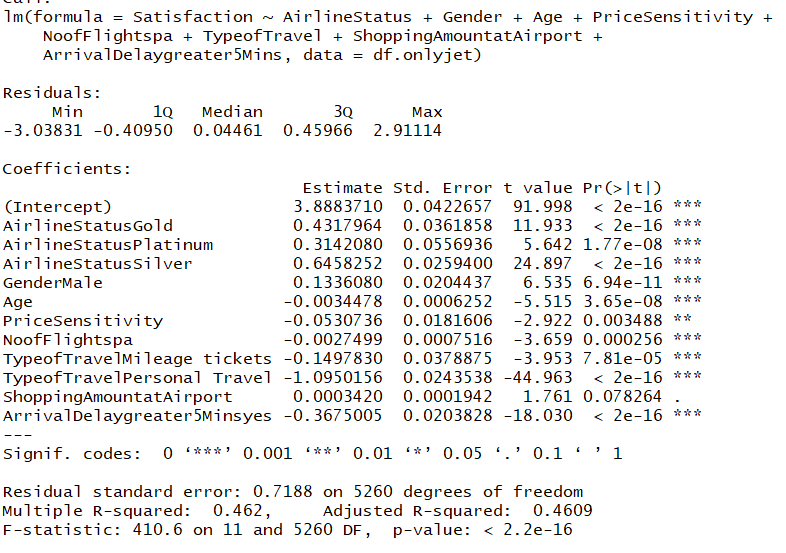
Here, we observed that the R Squared value is highest for the combination of below factors:

* Airline Status
* Gender
* Age
* Price sensitivity
* Number of flights per annum
* Type of travel
* Arrival delay greater than 5 minutes

**Code:**



**Output:**



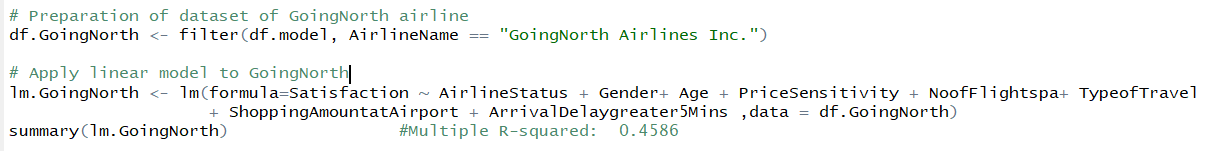
We observe that the R Squared value is 0.462 which means that a combination of all the seven factors account for 46% of the customer satisfaction rating.

**Linear modelling on GoingNorth airlines**

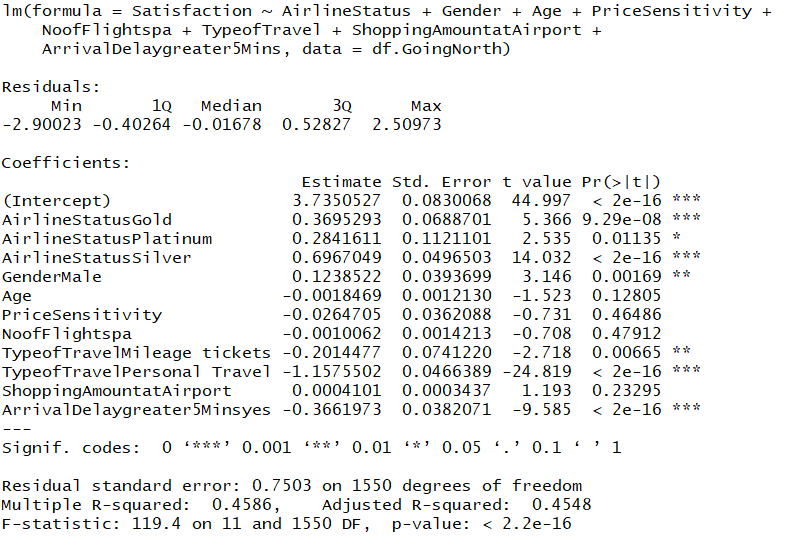
Here, we observed that the R Squared value is highest for the combination of below factors :

* Airline Status
* Gender
* Type of travel
* Arrival delay greater than 5 minutes

**Code:**



**Output:**



We observe that the R Squared value is 0.4586 which means that a combination of all the seven factors account for 45% of the customer satisfaction rating.

# Association Rules

Association Rule Mining is a common technique used to find associations between many variables. It is used when you want to find an association between different objects in a set, find frequent patterns in a transaction database, relational databases or any other information repository.

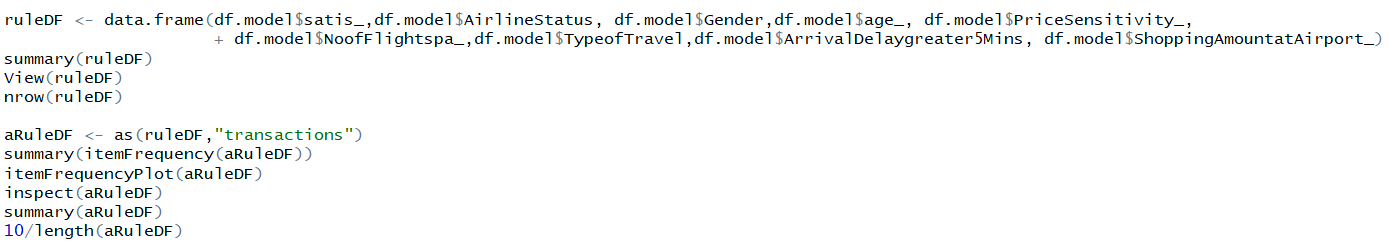
Here, we have used association rule to determine the factors which would have maximum impact on the customer satisfactions. We set RHS in the model as customer satisfaction in order to analyze the factors which made for high or low customer rating. Based on this, we considered some rules and picked up top two rules which had the highest “Lift” value.

This helped us to understand the combination of best factors which would be useful to convert a low customer satisfaction rating to high customer satisfaction rating.

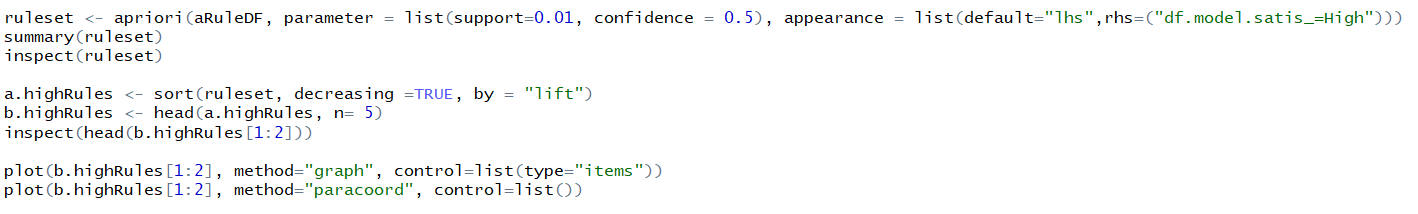
Association Rules on entire data set:

**Code:**

****

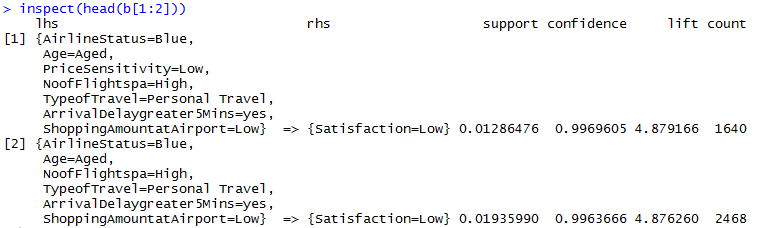
****

**High Customer Satisfaction Rules:**

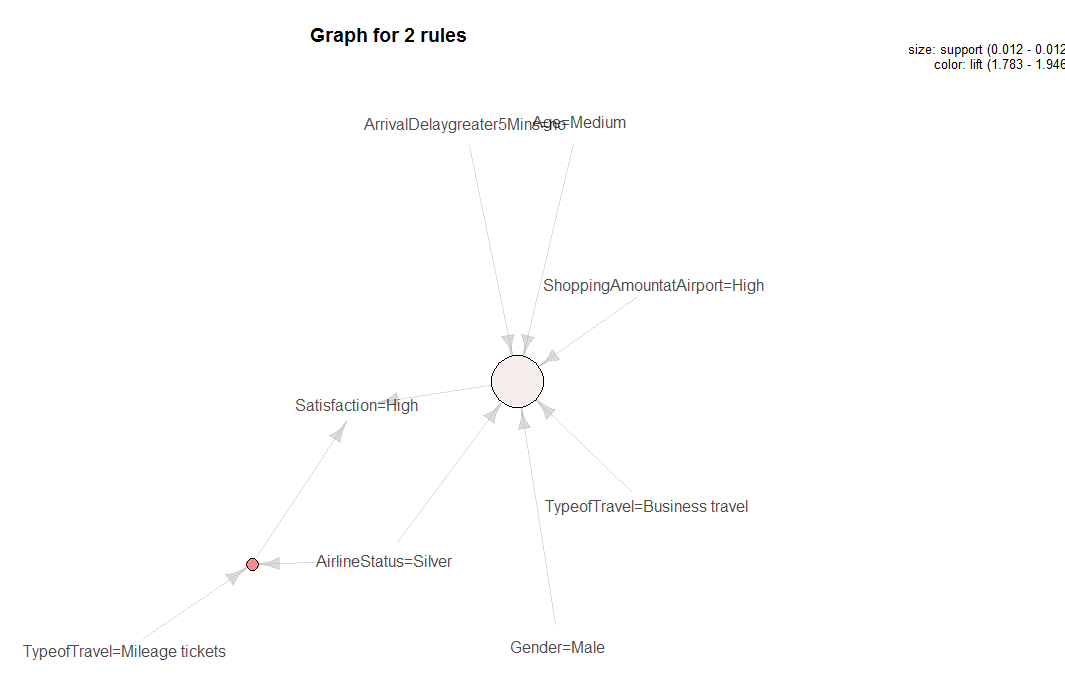
****

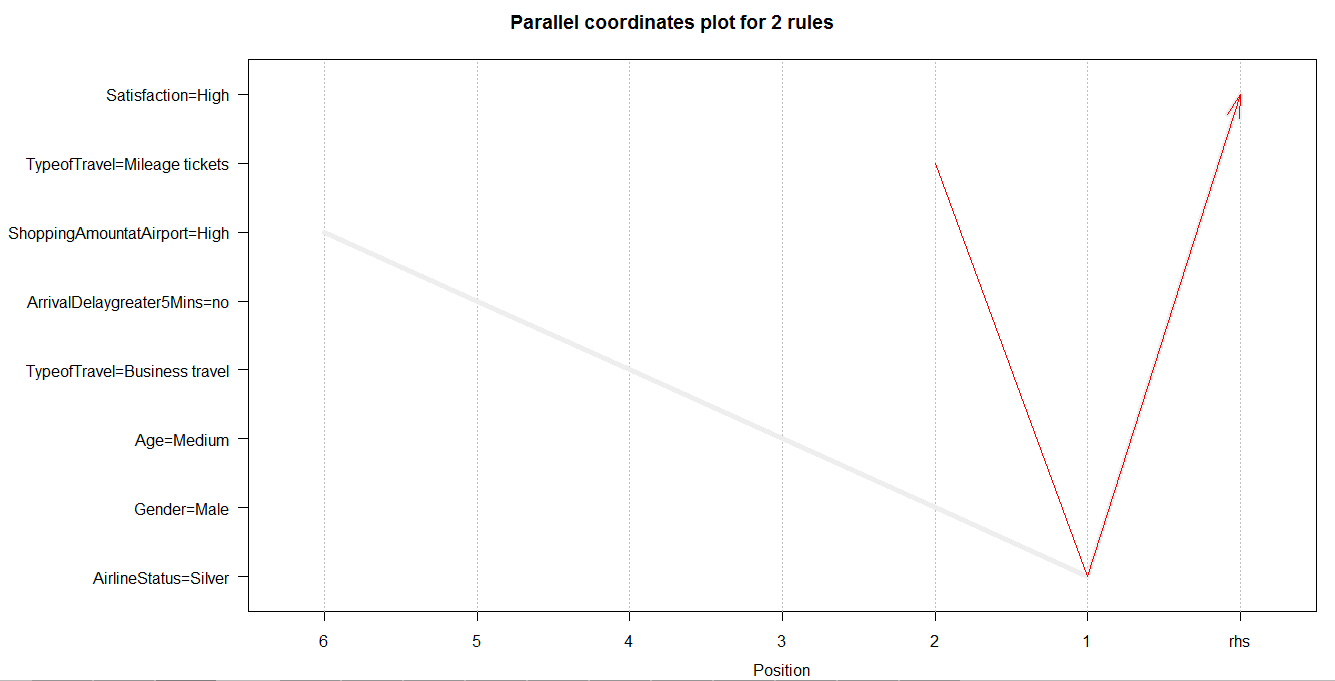
As per our model and combination of our support and confidence values for high customer satisfaction around 2290 rules were generated. Based on the highest values of lift we have taken 2 rules.

Rules:

****

**Plots:**

****

****

On running the association rule mining for highly satisfied customers, we have taken top 2 rules and plotted the below graph to determine the factors which influence customer satisfaction in airlines.

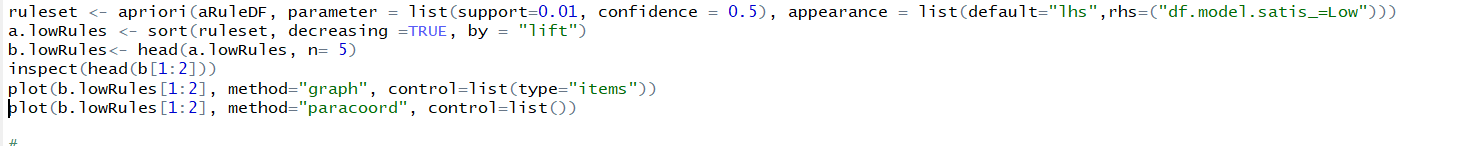
From the above plot we can determine the most significant attributes which influence a customer to become a highly satisfied customer are -

* Age - Medium (30-60)
* Gender - Male
* Type of travel - Business, mileage
* Airline Status - Silver
* Price sensitivity - Low
* Arrival delay greater than 5 minutes - No
* Shopping amount at Airport - High

This shows that the above factors help the customer to be highly satisfied.

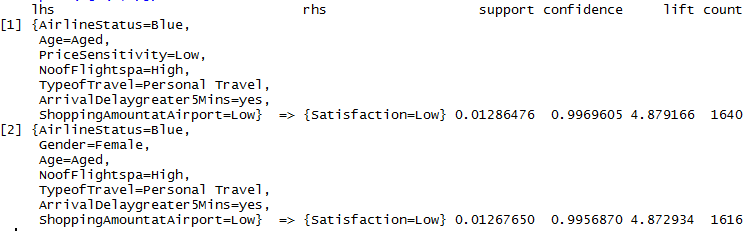
**Low satisfied Customer:**

**Code:**

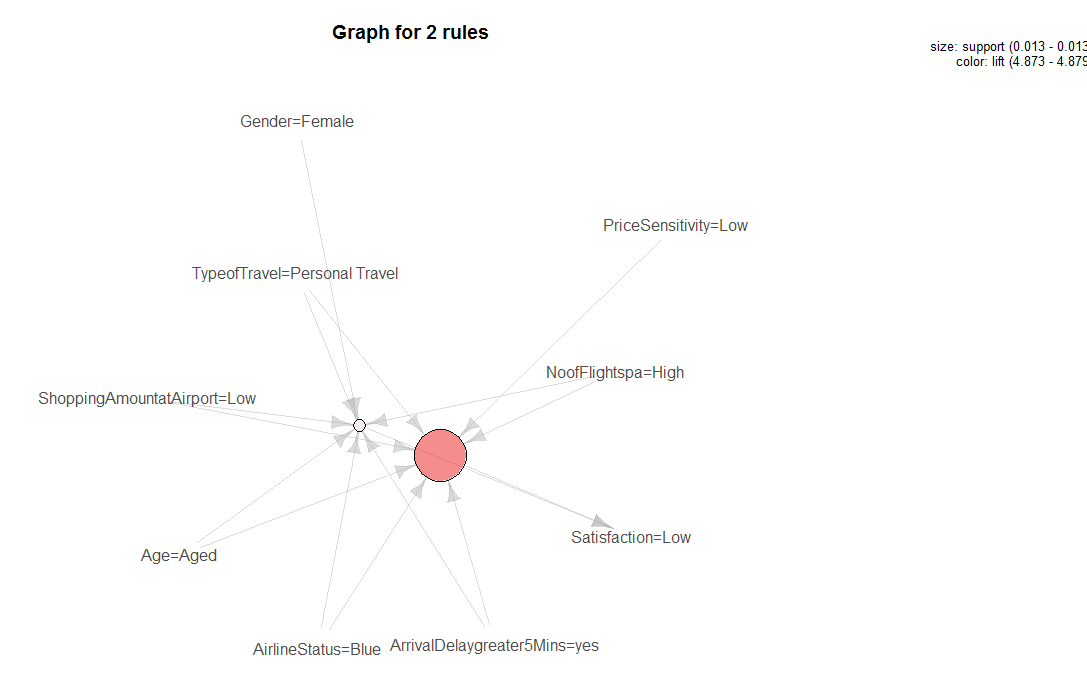
****

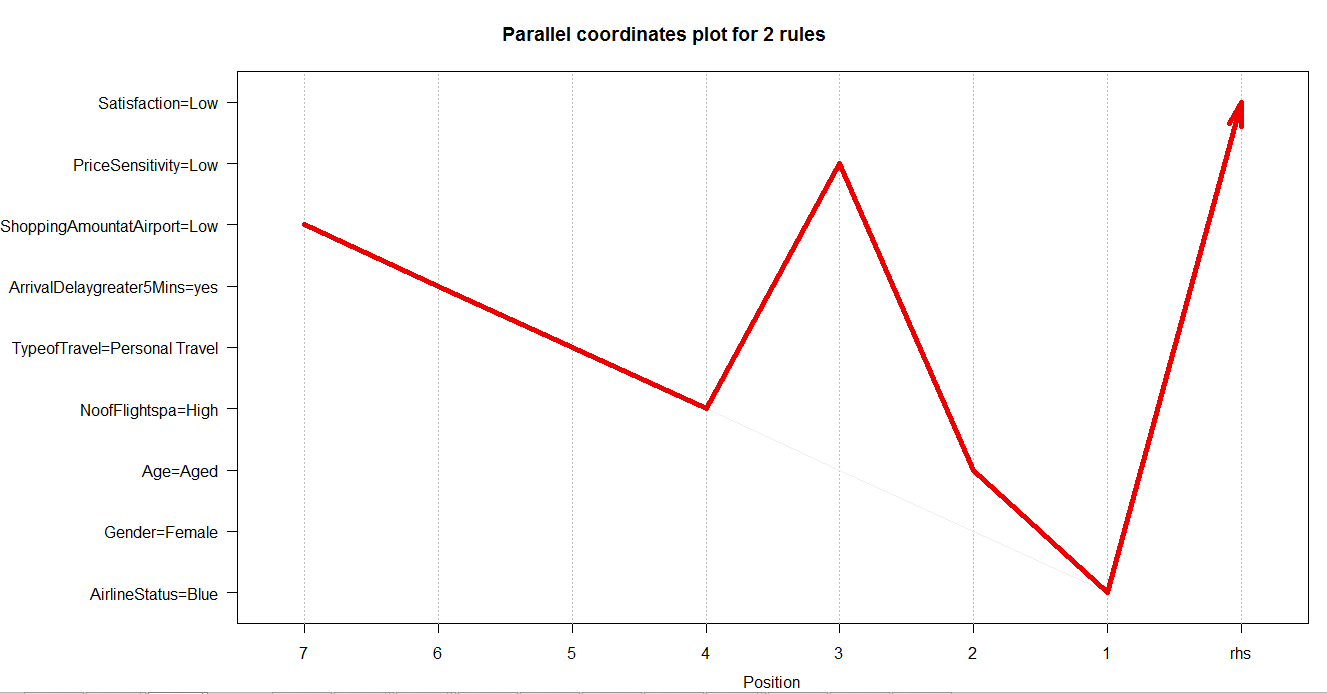
**Rules:**

As per our model and combination of our support and confidence values for high customer satisfaction around 465 rules were generated. Based on the highest values of lift we have taken 1st and 4th rule.

****

**Plot:**

****

****

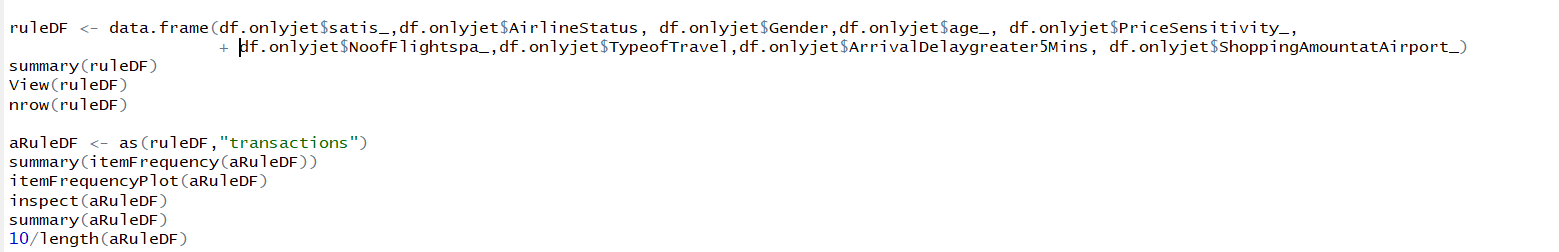
From the above plot we can determine the most significant attributes which influence a customer to become low satisfied customer are -

* Age- Aged(60-85)
* Airline Status - Blue
* Price sensitivity - low
* Number of flights per annum - high
* Arrival delay greater than 5 minutes - Yes
* Shopping amount at Airport - Low
* Type of travel - personal travel
* Gender - Female

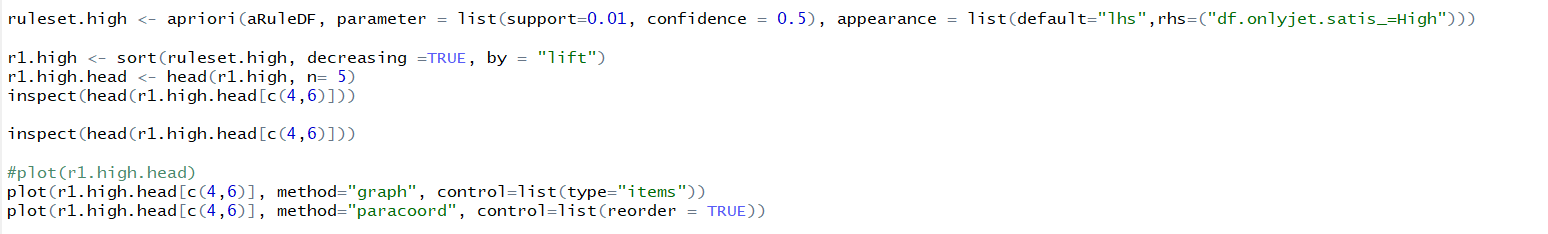
**Association Rules on OnlyJet Airways:**

**Code:**

****

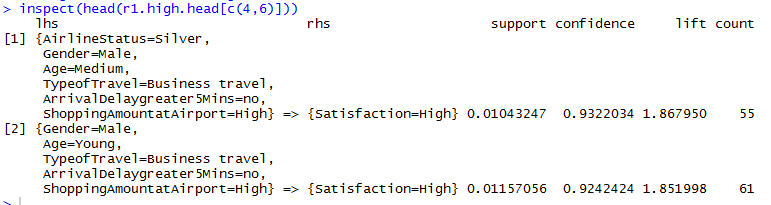
****

**High Customer Satisfaction Rules:**

****

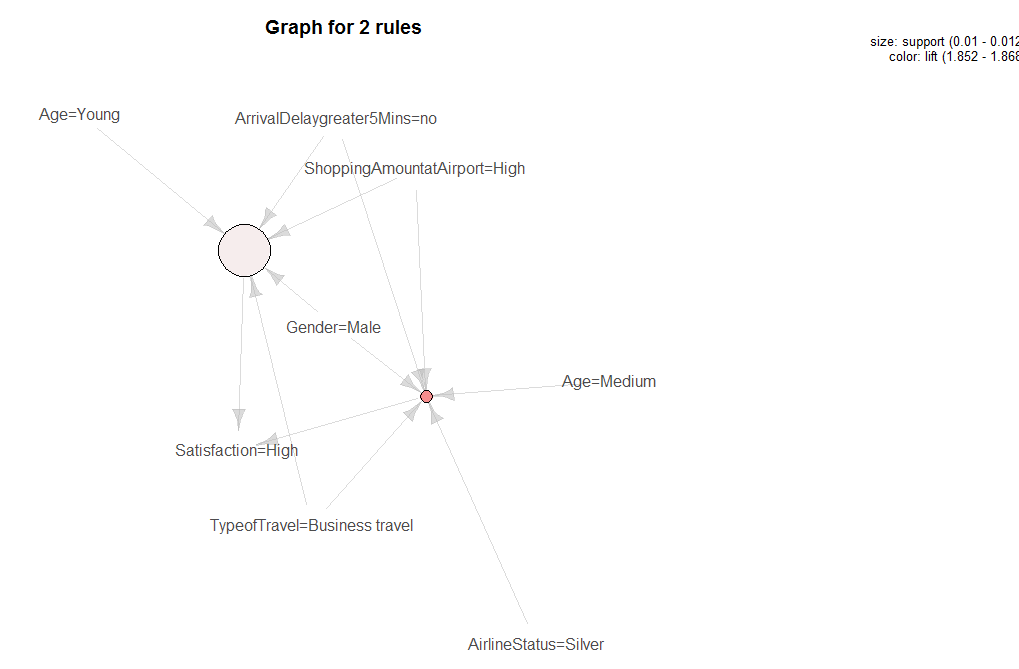
As per our model and combination of our support and confidence values for high customer satisfaction around 2204 rules were generated. Based on the highest values of lift we have taken 2 rules.

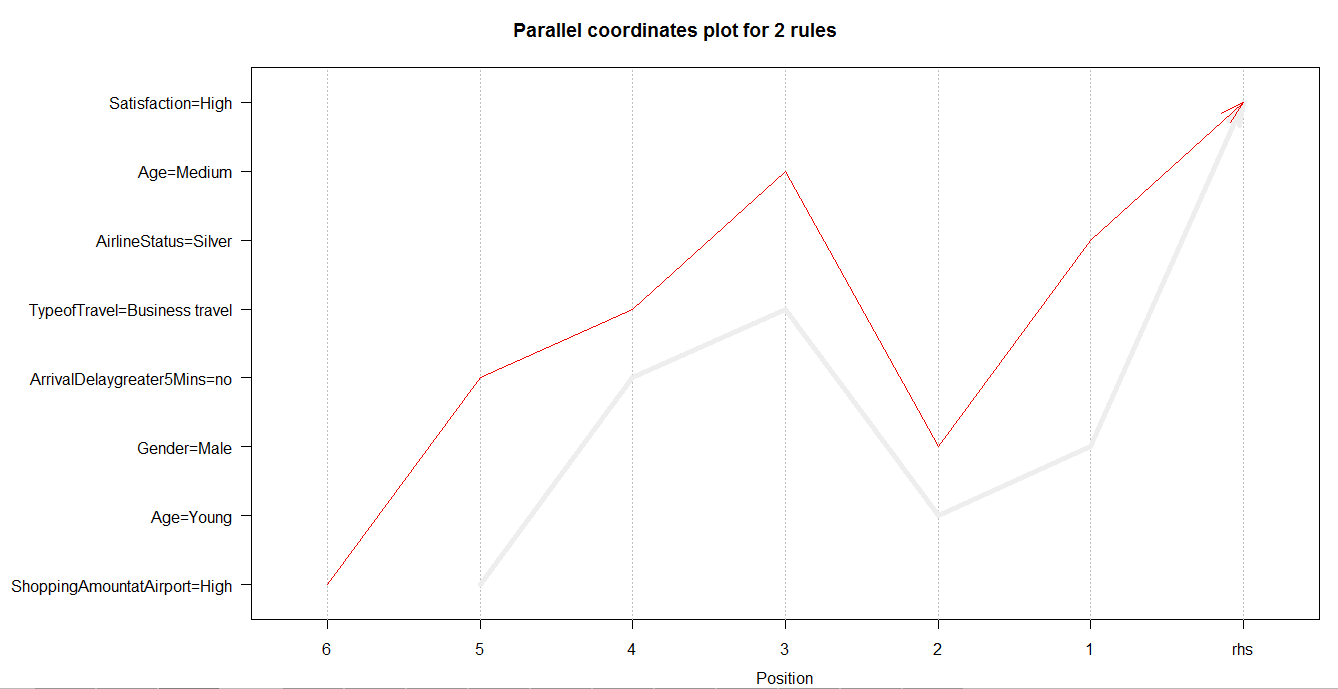
**Rules:**

****

**Plot:**

On running the association rule mining for highly satisfied customers, we have taken top 2 rules and plotted the below graph to determine the factors which influence customer satisfaction in airlines.

****

****

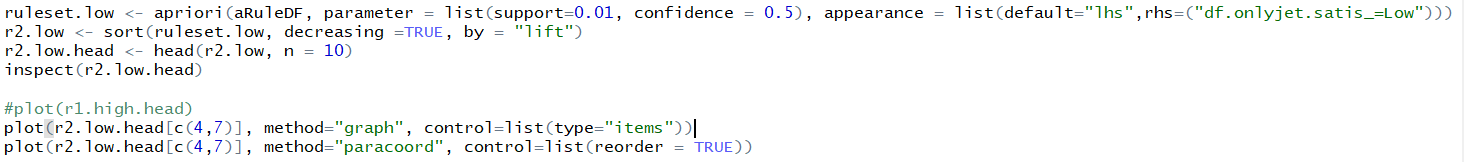
From the above plot we can determine the most significant attributes which influence a customer to become a highly satisfied customer are -

* Age - Young (15-30)
* Gender - Male
* Airline Status - Silver
* Type of Travel - Business travel
* Arrival delay greater than 5 minutes - No
* Shopping amount at Airport - High

This shows that the above factors help the customer to be highly satisfied.

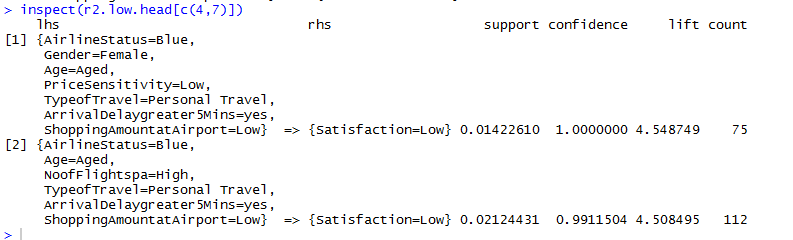
Low Customer Satisfaction Rules:

**Code:**

****

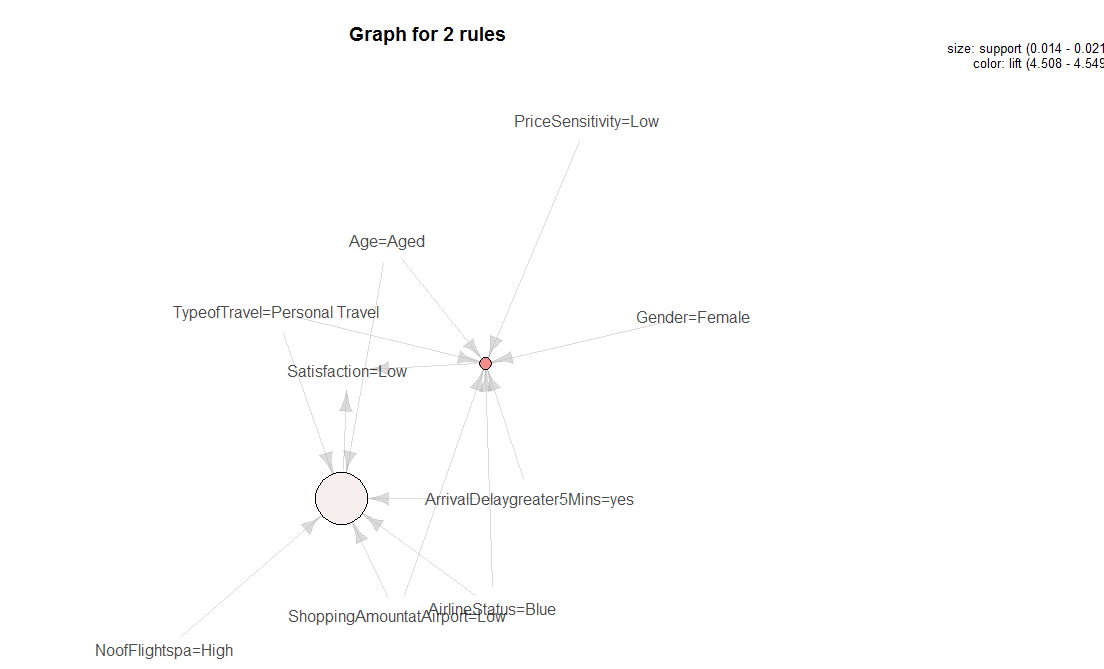
As per our model and combination of our support and confidence values for low customer satisfaction around 563 rules were generated. Based on the highest values of lift we have taken 2 rules.

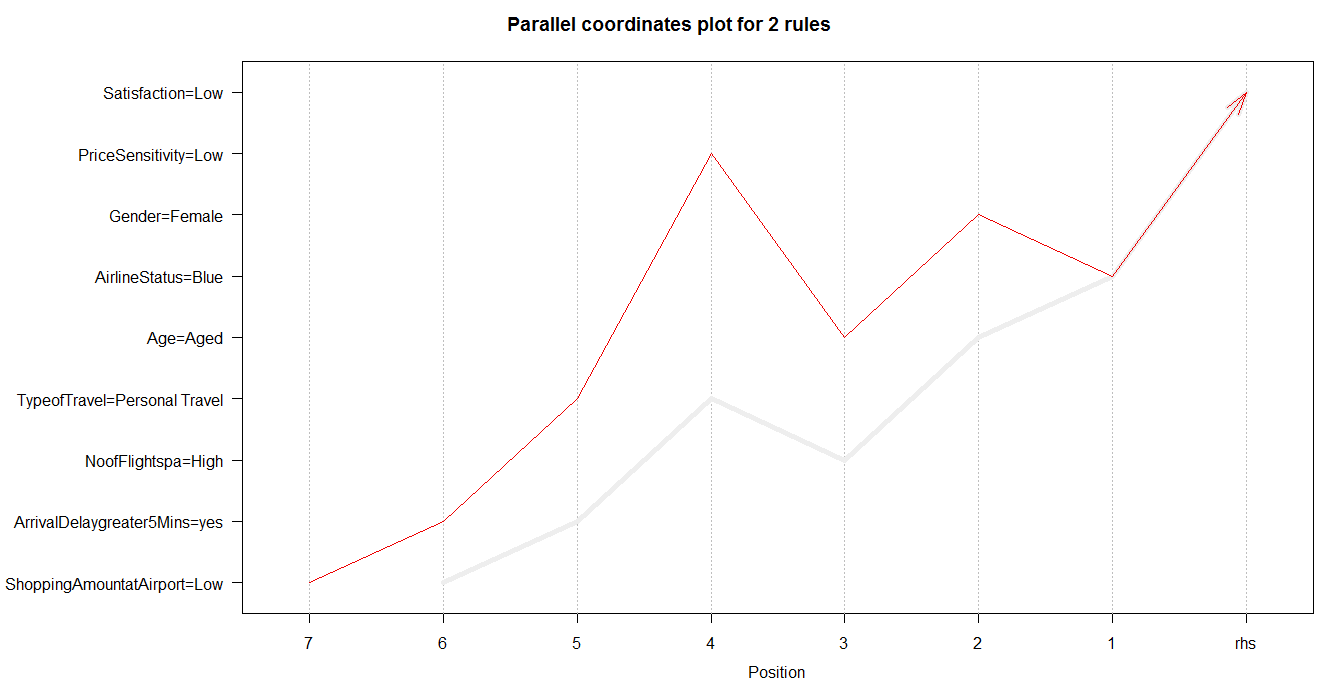
**Rules:**

****

**Plot:**

On running the association rule mining for the Only Jet lowest satisfied customers , we have taken top 2 rules and plotted the below graph to determine the factors which influence customer satisfaction in airlines.

****

****

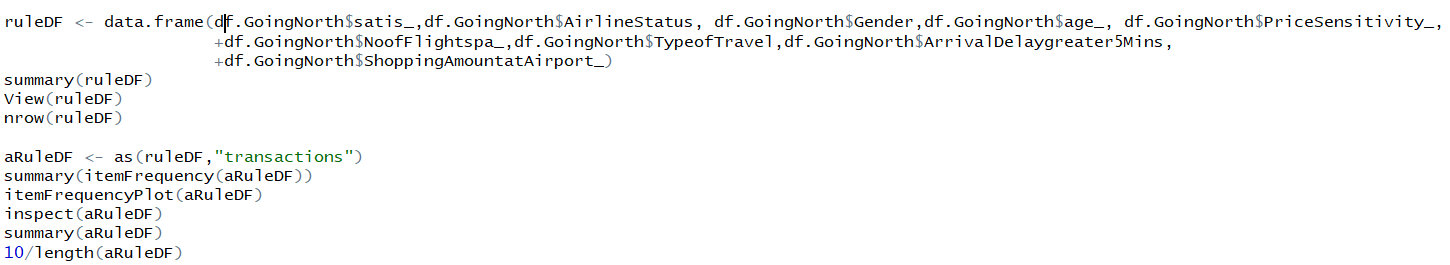
From the above plot we can determine the most significant attributes which influence a customer to become low satisfied customer are -

* Age- Aged(60-85)
* Gender- Female
* Airline Status - Blue
* Price sensitivity - low
* Arrival delay greater than 5 minutes - Yes
* Shopping amount at Airport - Low
* Type of travel - personal travel

**Association Rules on GoingNorth Airways:**

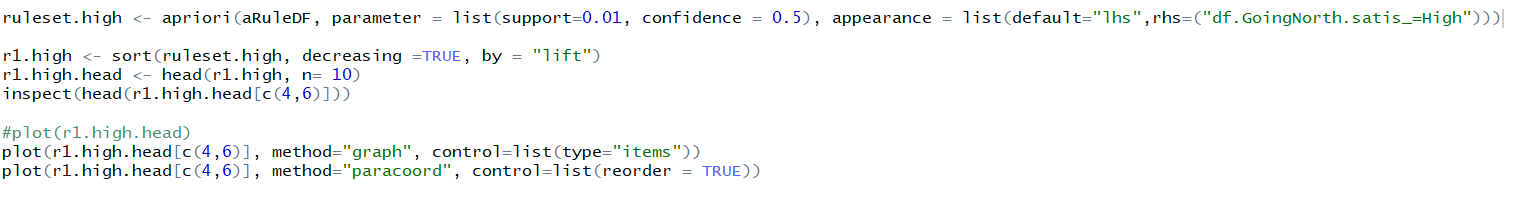
**Code:**

****

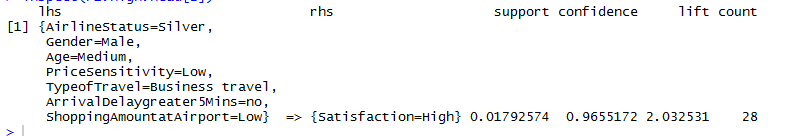
****

**High Customer satisfaction Rules:**

**Code:**

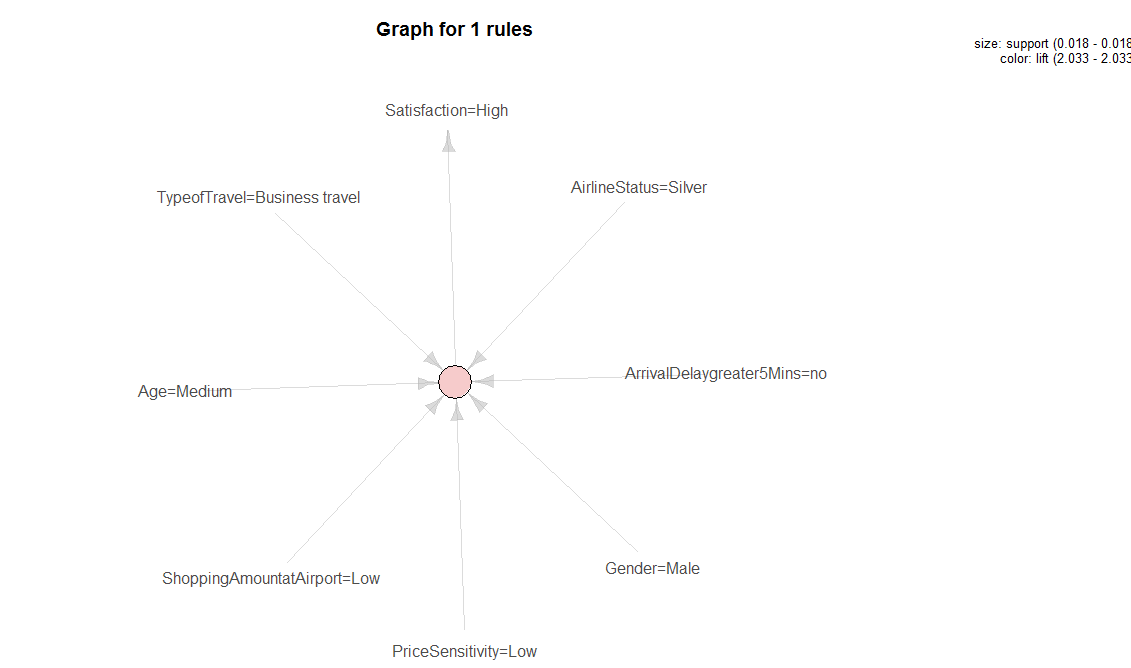
****

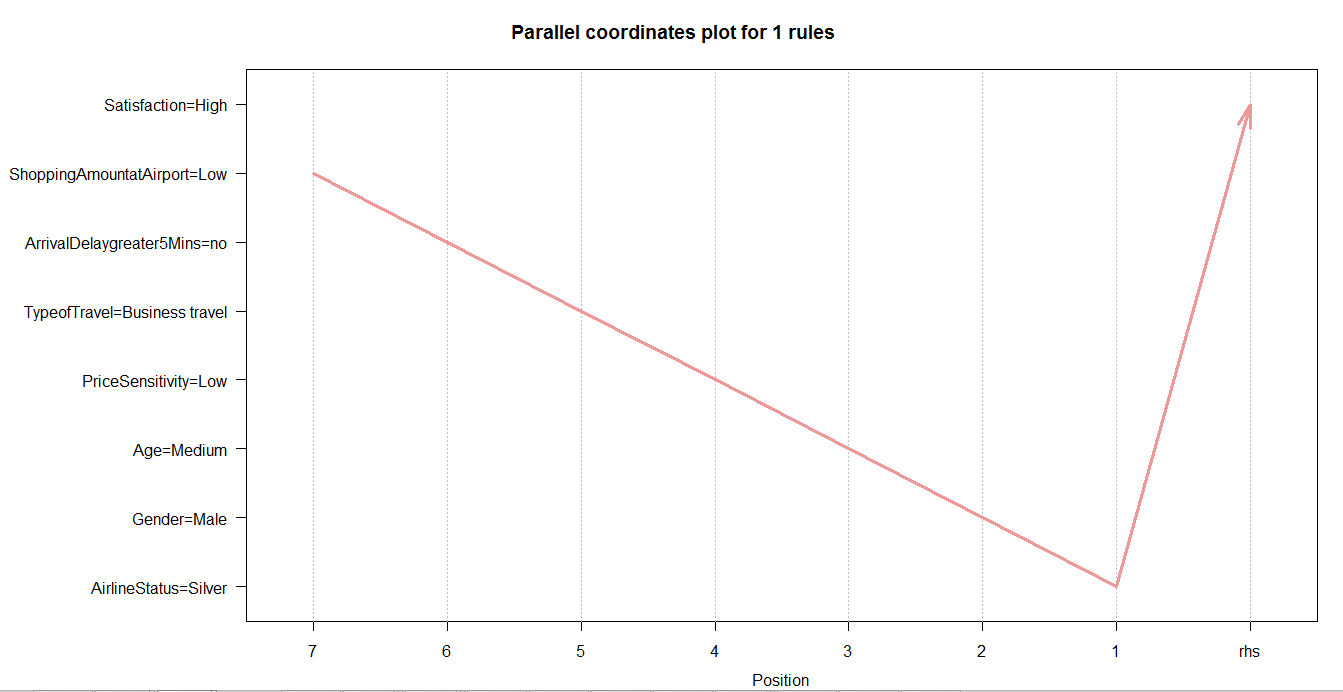
As per our model and combination of our support and confidence values for low customer satisfaction around 2054 rules were generated. Based on the highest values of lift we have taken 1 rule.

****

**Plot:**

On running the association rule mining for the GoingNorth highly satisfied customers , we have taken top 1 rules and plotted the below graph to determine the factors which influence customer satisfaction in airlines.

****

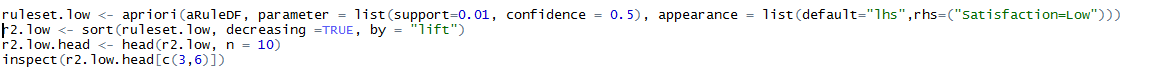
****

From the above plot we can determine the most significant attributes which influence a customer to become highly satisfied customer are -

* Age- Medium(30-60)
* Gender- Male
* Airline Status - Silver
* Price sensitivity - low
* Arrival delay greater than 5 minutes - No
* Shopping amount at Airport - Low
* Type of travel - Business travel

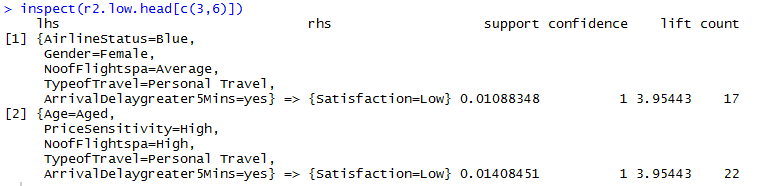
**Low Customer Satisfaction Rules:**

**Code**

****

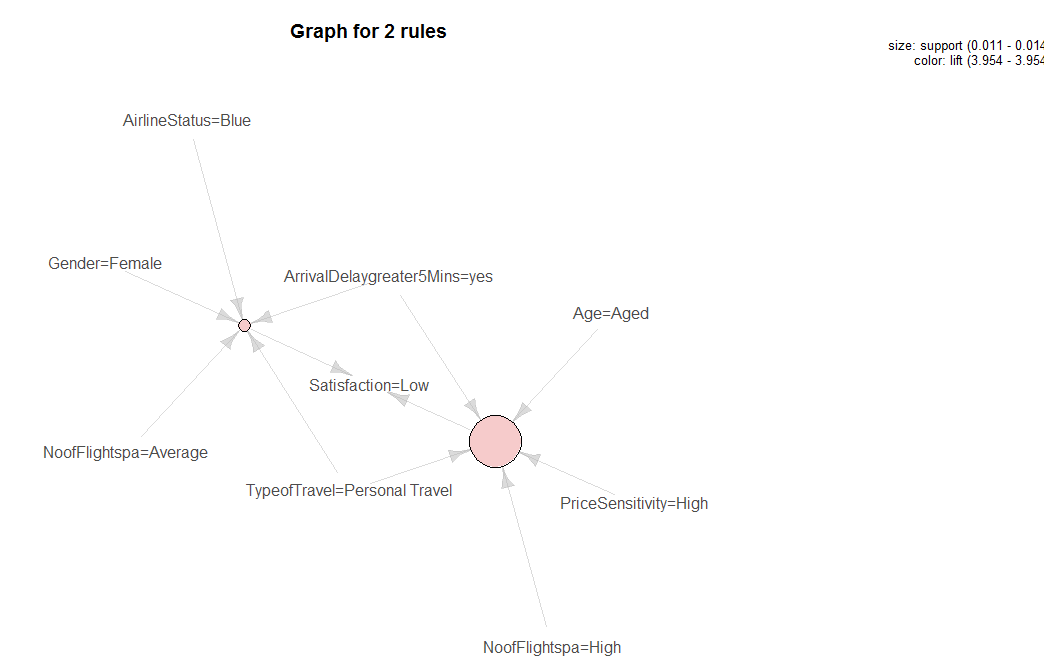
As per our model and combination of our support and confidence values for low customer satisfaction around 721 rules were generated. Based on the highest values of lift we have taken 2 rules.

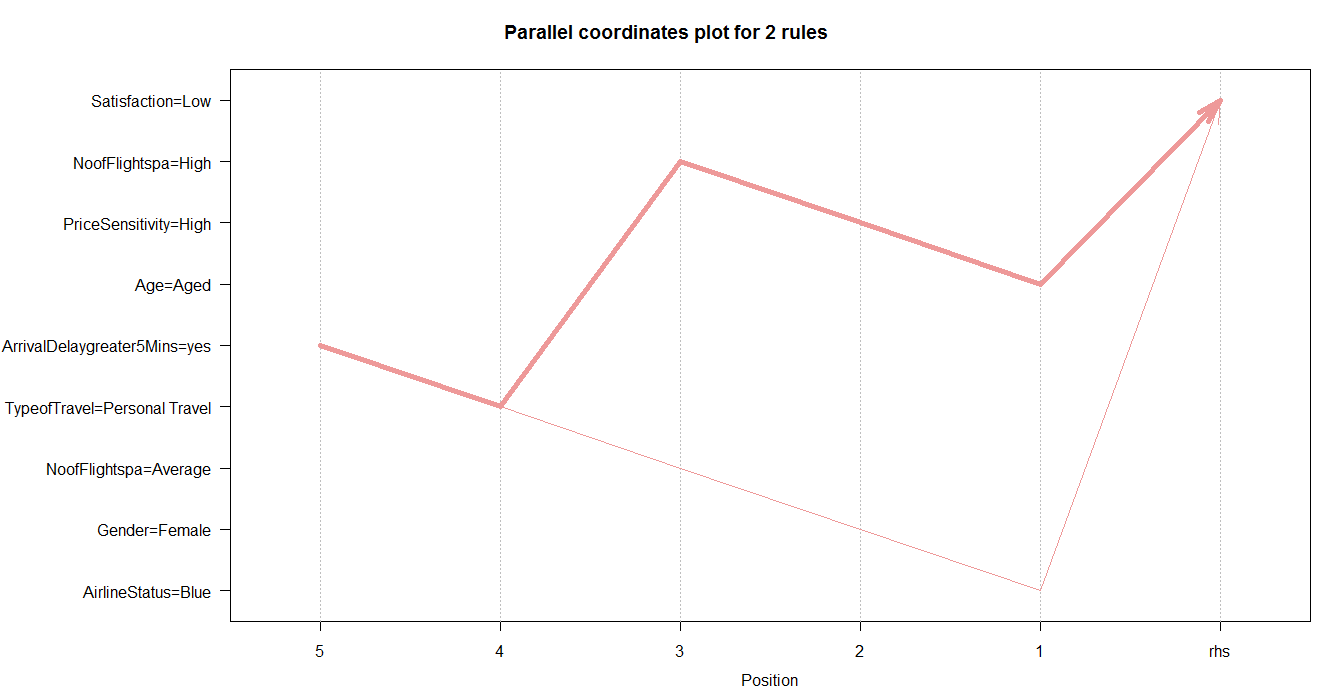
**Rules:**

****

**Plot:**

On running the association rule mining for the Only Jet lowest satisfied customers , we have taken top 2 rules and plotted the below graph to determine the factors which influence customer satisfaction in airlines.

****

****

From the above plot we can determine the most significant attributes which influence a customer to become low satisfied customer are -

* Age- Aged(60-85)
* Gender- Female
* Airline Status - Blue
* Price sensitivity - High
* Arrival delay greater than 5 minutes - Yes
* Type of travel - personal travel

# Support Vector Machine

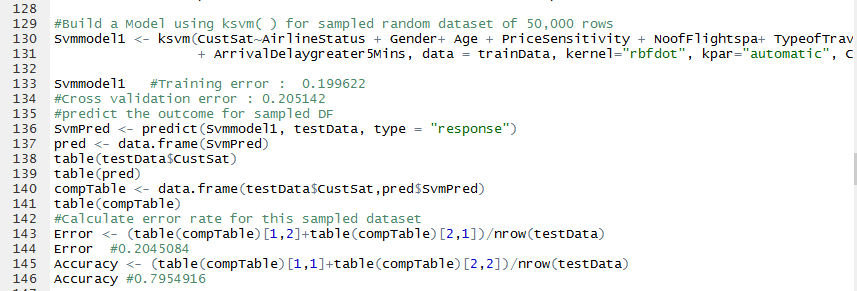
Support Vector Machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.

Here, we have considered variables from the above three methods i.e. linear modelling, association rules mining and correlation matrix and applied them to support vector machine to cross validate the significance of the variables.

The outcomes will highlight if the accuracy of SVM model is high then the selected variables would be highly significant.

**Support Vector Machine on sampled dataset (of 50,000 rows)**

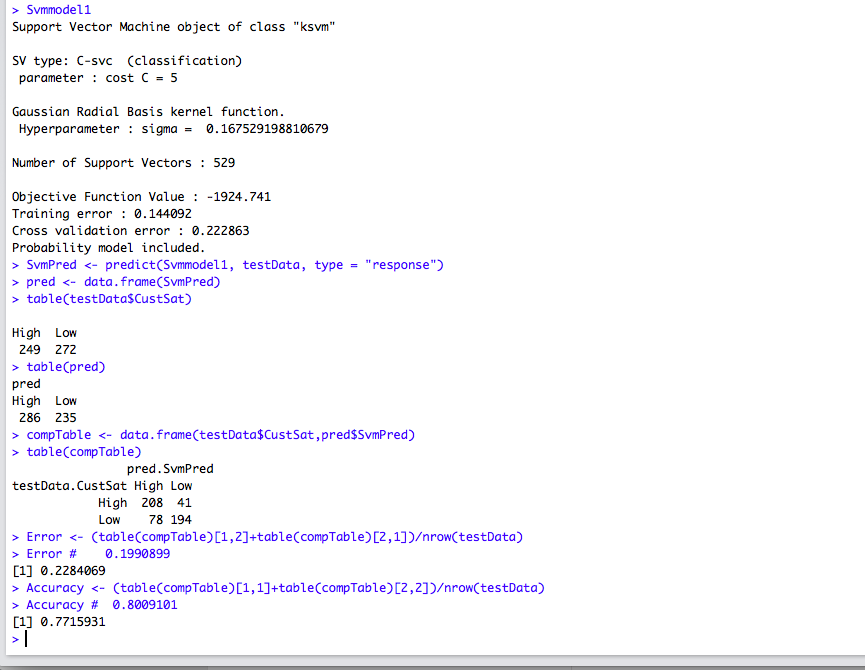
**Code:**



Here, we have sampled 50000 attributes from the whole dataset and divided into test data and training data. Then we used this training data to build SVM model and applied test data to predict our results i.e. the accuracy reflects out to be 79%.

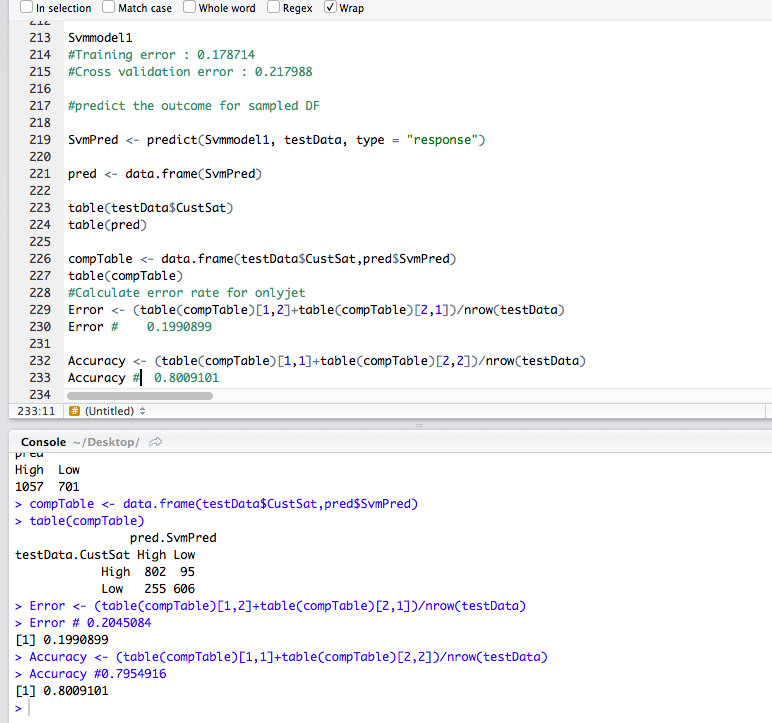
**Support Vector Machine on Going North Airline**

****

****

Here, we filtered the dataset for Going North airline and divided into test data and training data. Then we used this training data to build SVM model and applied test data to predict our results i.e. the accuracy reflects out to be 77%.

**Support Vector Machine on OnlyJets**

****

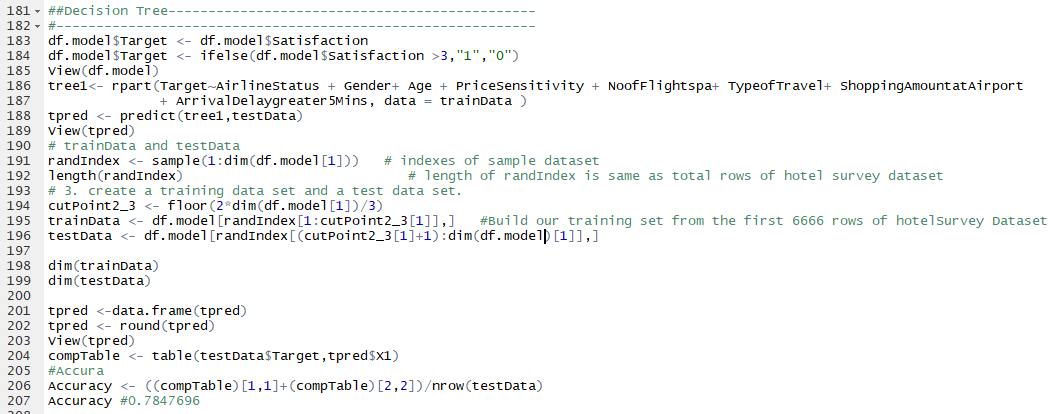
Here, we filtered the dataset for OnlyJets airline and divided into test data and training data. Then we used this training data to build SVM model and applied test data to predict our results i.e. the accuracy reflects out to be 80%.

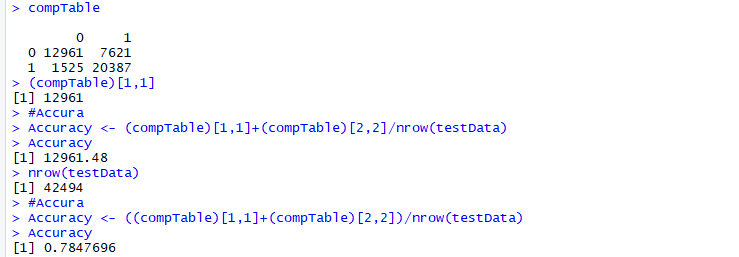
# Decision Tree

**Decision Tree on entire dataset**

Decision tree algorithm helps us explore the structure of a set of data, while developing easy to visualize decision rules for predicting a categorical (classification tree) or continuous (regression tree) outcome. A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.

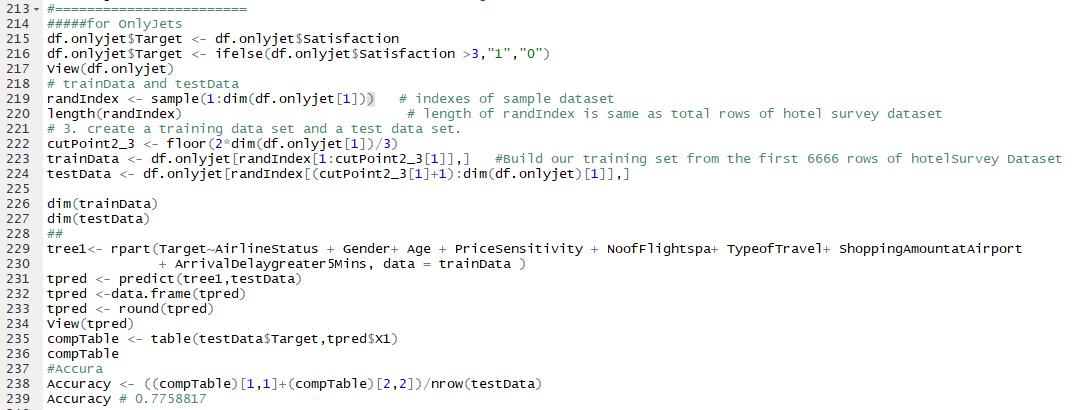
To further validate our results, we used decision tree model to check the accuracy of the significant variables.

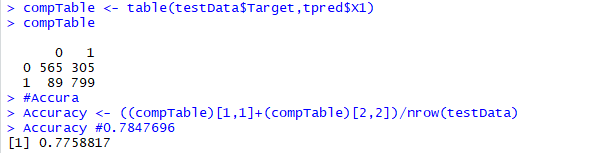
****

****

So here, with decision tree model, a tree of all variables is constructed and then variable condition is checked that starts from the root of the tree and progresses towards the stem of the tree.

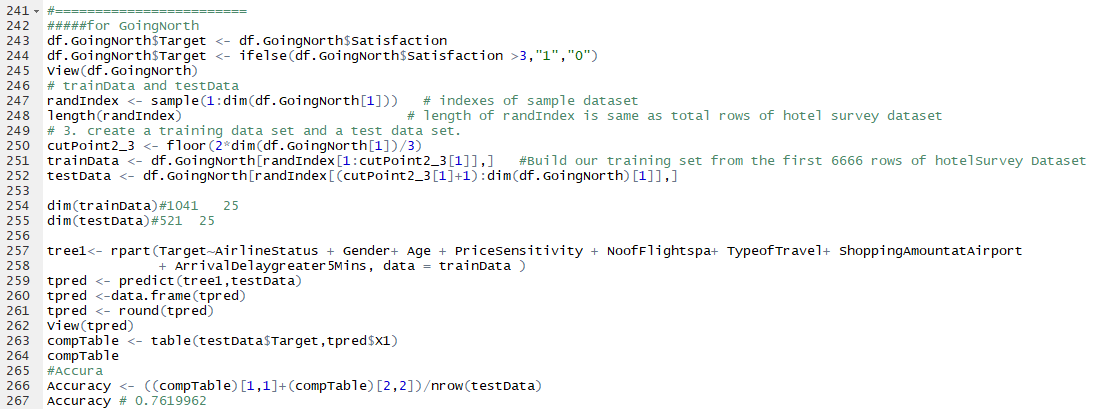
**Decision Tree on Only Jets**

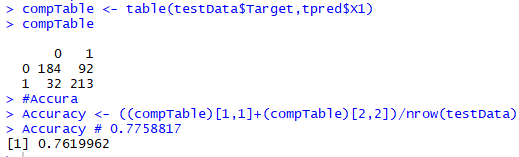
****

****

Here, with decision tree model, a tree of Only Jet dataset is constructed and then variable condition is checked that starts from the root of the tree and progresses towards the stem of the tree.

**Decision Tree on Going North**

****

****

Here, with decision tree model, a tree of Going North dataset is constructed and then variable condition is checked that starts from the root of the tree and progresses towards the stem of the tree.

# Validation

* Significant variables observed from Linear modelling, visualizations and correlation matrix (mentioned below) are validated using 3 modelling techniques of Association rules, Support vector machine and Decision tree.
* Significant Variables on which these models are applied:
  + Airline Status
  + Gender
  + Age
  + Price sensitivity
  + Number of flights per annum
  + Type of travel
  + Shopping Amount at airport
  + Arrival delay greater than 5 minutes
* Models:

1. **In Association rules**, we got combination of rules with these significant variables.

Customer Satisfaction = Low when,

Age = Age range (60-85)

Gender = Female

Airline Status = Blue

Price sensitivity = High

Arrival delay greater than 5 minutes = Yes

Type of travel = personal travel

This proves the insights we are getting from visualizations of these columns.

1. **In SVM & Decision tree**, we built a model using these significant variables as input (training data) and checked accuracy of the models with prediction on testing data

Prediction Efficiency from SVM: 80.09%

Prediction Efficiency from Decision Tree: 77.58

With higher accuracy, these variables are proved to be important and impacting customer satisfaction the most.

# Actionable Insights

Through Visualizations & modelling these are inferences,

* Female customers are giving lower satisfaction ratings and could be offered better deals for improvement
* Aged (60-85) customers are using “Personal type of travel” & their average satisfaction is low. Hence, personal travel amenities for senior customers should be increased and they should be well-maintained.
* Premium status classes likes Gold, Platinum are having good customer satisfaction whereas Blue has lower ratings. Therefore, extra services should be provided for convenience of customers travelling through Blue status.
* Customers taking more flights per year are giving low satisfaction ratings. Thus, loyal customers should be provided with discounts and points which might satisfy them in long run.

**Attached R code:**

****