Data Analysis of Southeast Airlines

Project by: Vashisth Bhatt

Table of Contents

Problem Statement

In recent years, the airline industry changed dramatically and became an extremely competitive industry. Because operational excellence became a best practice in the airline industry, airline companies seek to understand customer needs in order to drive and sustain competitive advantage over others. By analyzing customer survey data from a variety of airlines, this project aims to not only provide Southeast Airlines the knowledge and tools needed to understand drivers of customer satisfaction, but also information on how the airline can stand at the forefront of customer service.

Objectives

The goals of this project are to identify the key factors that contribute towards overall customer satisfaction in the airline industry, build a model specifically for Southeast Airlines to predict customer satisfaction, and provide Southeast Airlines with actionable insights for future business operations.

Project Scope

In order to achieve the goals for this project, we have focused on the following major objectives:

1. Generate Business Questions
2. Data Cleaning
3. Data Exploration
4. Data Modeling
5. Data Analysis and Visualizations
6. Offer actionable Recommendations and Insights

Business Questions:

- What impact do the various different airline statuses have on customer satisfaction?
- What is the relationship between the type of travel and overall customer satisfaction?
- Can Southeast Airline improve operational effectiveness in certain states?
- What impact has shopping at the airport on customer satisfaction and what role does gender play in that context?
- How do different age group affect customer satisfaction?

Data Cleaning

The original dataset was received in CSV format and contained a total of 129890 records. After loading the data file into R, we cleaned the data and erased three observations due to format inconsistencies. In detail, three rows showed customer satisfaction ratings in the following format - “4.00.200”. Furthermore, we have removed spaces as well as dots from the column names in order to ease the process of modeling building later on. Moreover, we deleted records of flights that were cancelled for purposes of our delay time analysis.

Data Analysis:

1) Descriptive Statistics:

After cleaning our data, we would like to provide Southeast Airlines with descriptive statistics, for them to better understand the information captured by the dataset. The data contains information on 14 different airlines and 27 attributes that can be classified into three different groups – customer information, travel and flight information, and flight performance. The following visualizations summarize the most relevant attributes for later analysis.

Before diving into the descriptive statistics for each of the attribute groups, figure 1 below depicts the average customer satisfaction by airline and provides Southeast Airlines with a first glance of how they perform compared to other airlines. As seen, there are no major differences in the average customer satisfaction between airlines and the average customer satisfaction across all airlines is 3.37. Simply looking at the average customer satisfaction, we can understand that there is room for improvement across all airlines. Southeast Airline can really differentiate themselves from others by understanding their customer needs and taking initiative.

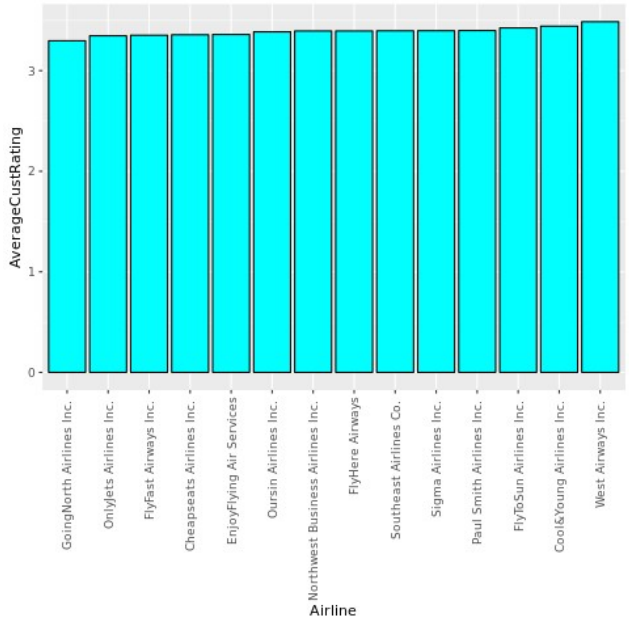


Figure 1: Average customer satisfaction by Airline

A) Customer Information

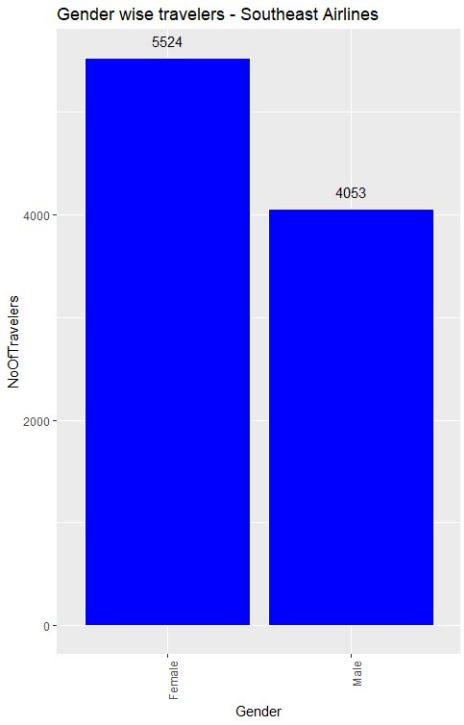


Figure 2: Total Count of Travelers by Gender

Figure 2 shows the total number of Southeast Airline flights taken by gender. The total number of female travelers was 5524, while the total number of male travelers was 4053. The number of female travelers is around 1500 flights higher than the number of male travelers.

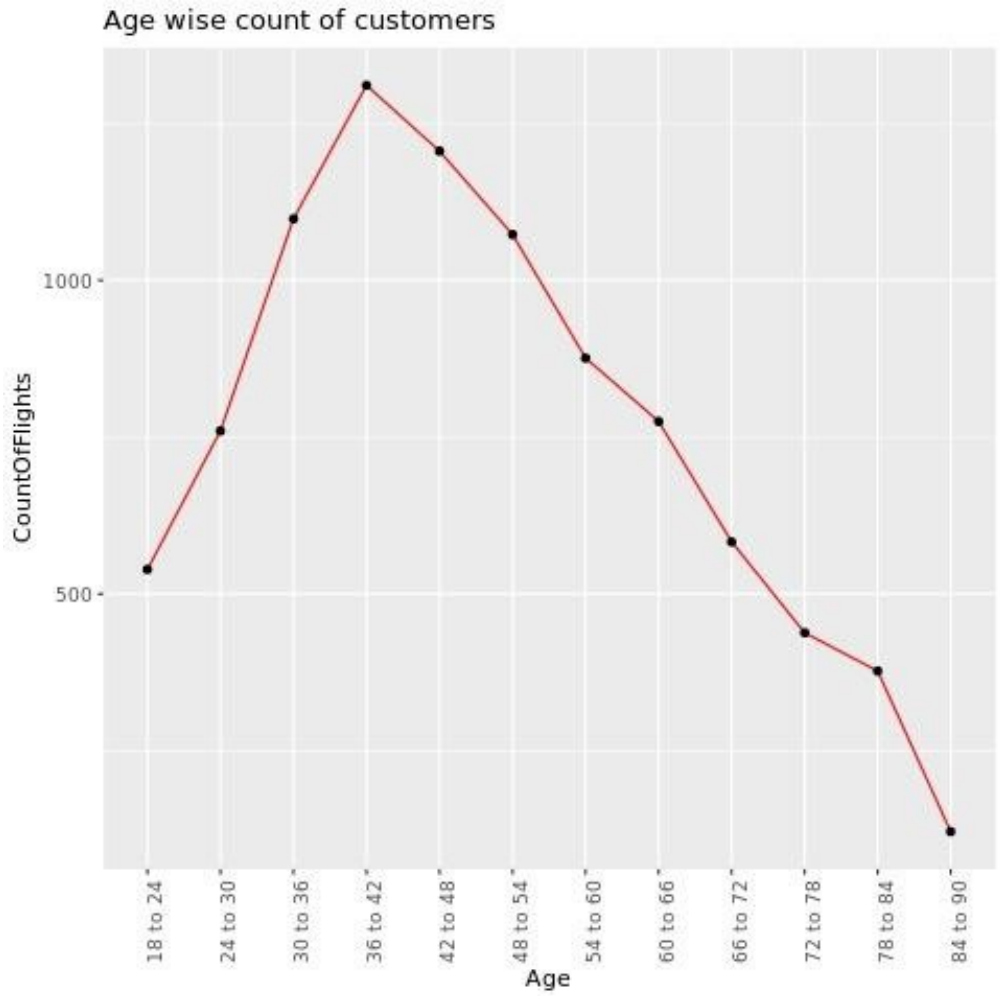


Figure 3: Total Number of Flights by various Age groups

Figure 3 shows the number of flights taken by customer age. We observe that customers age group of 36 to 42 are the ones that fly the most while customers older than 84 are the ones flying the least. What is important to mention is that we see a constant increase in the number of flights taken until the age of 42 and then observe a steady decline in the number of flights taken.

B) Travel and Flight Information

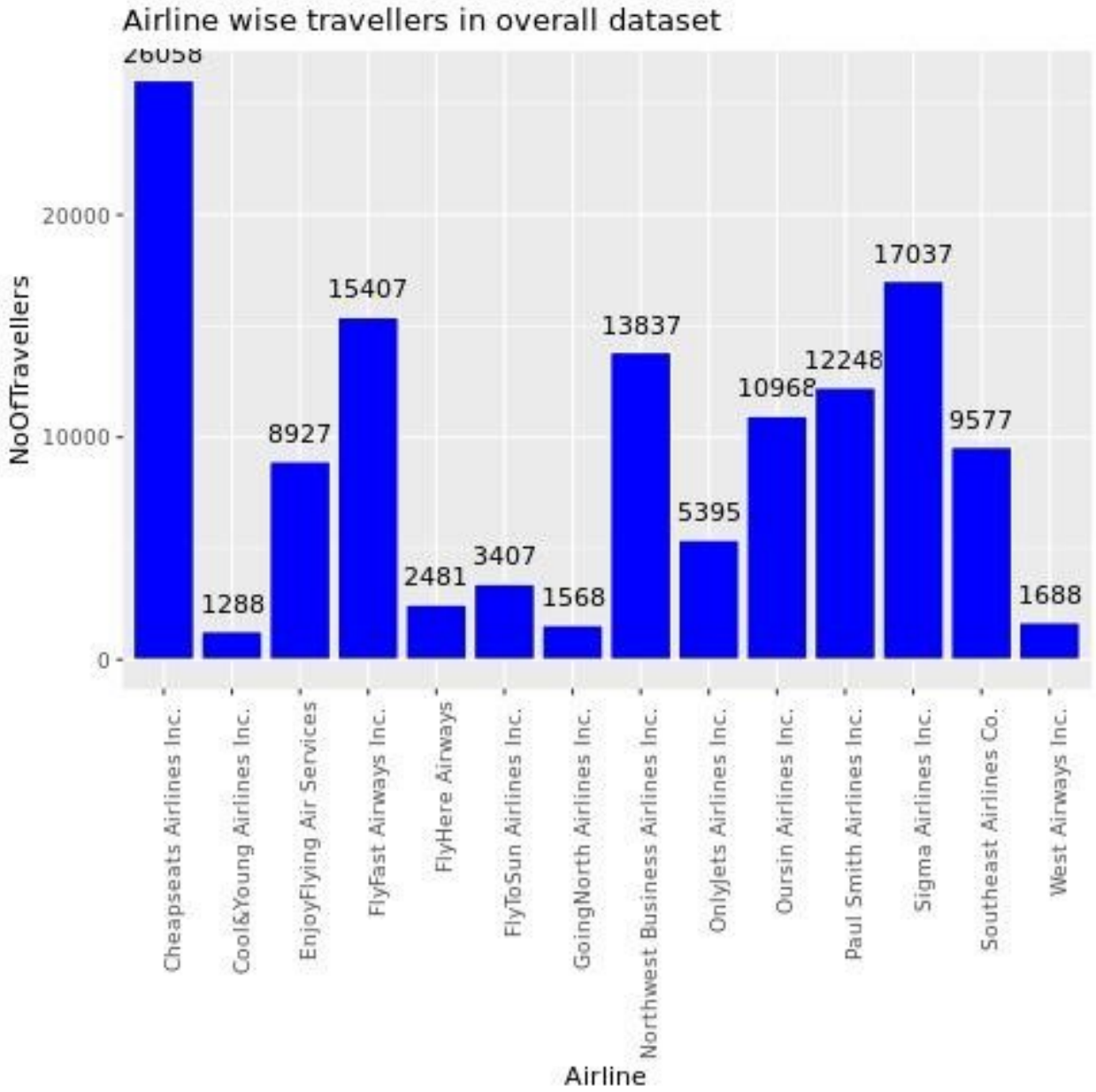


Figure 4: Total Count of Travelers by Airline

Figure 4 depicts the total number of travelers per airline. We can observe that the number of travelers per airline ranges from approximately 1300 to 26000 flights. Southeast Airlines has a total number of travelers of 9577, which is the seventh most travelers out of the 14 airlines.

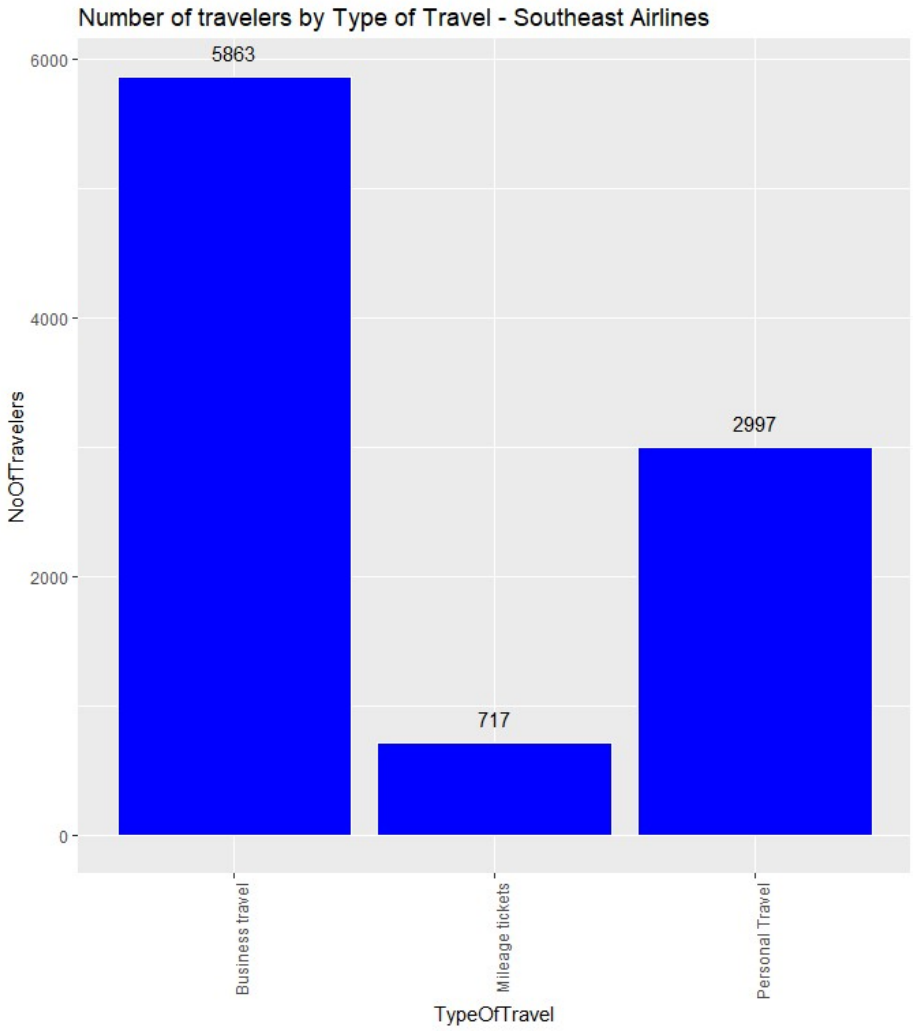


Figure 5: Total Count of Travelers by Type of Travel

Figure 5 below indicates the total number of travelers from Southeast Airlines based on the type of travel. There are three different types of travel; Business, Mileage, and Personal travel. The total number of business traveler amounts 5863 for business travelers, 717 for mileage travelers and 2997 for personal travelers.

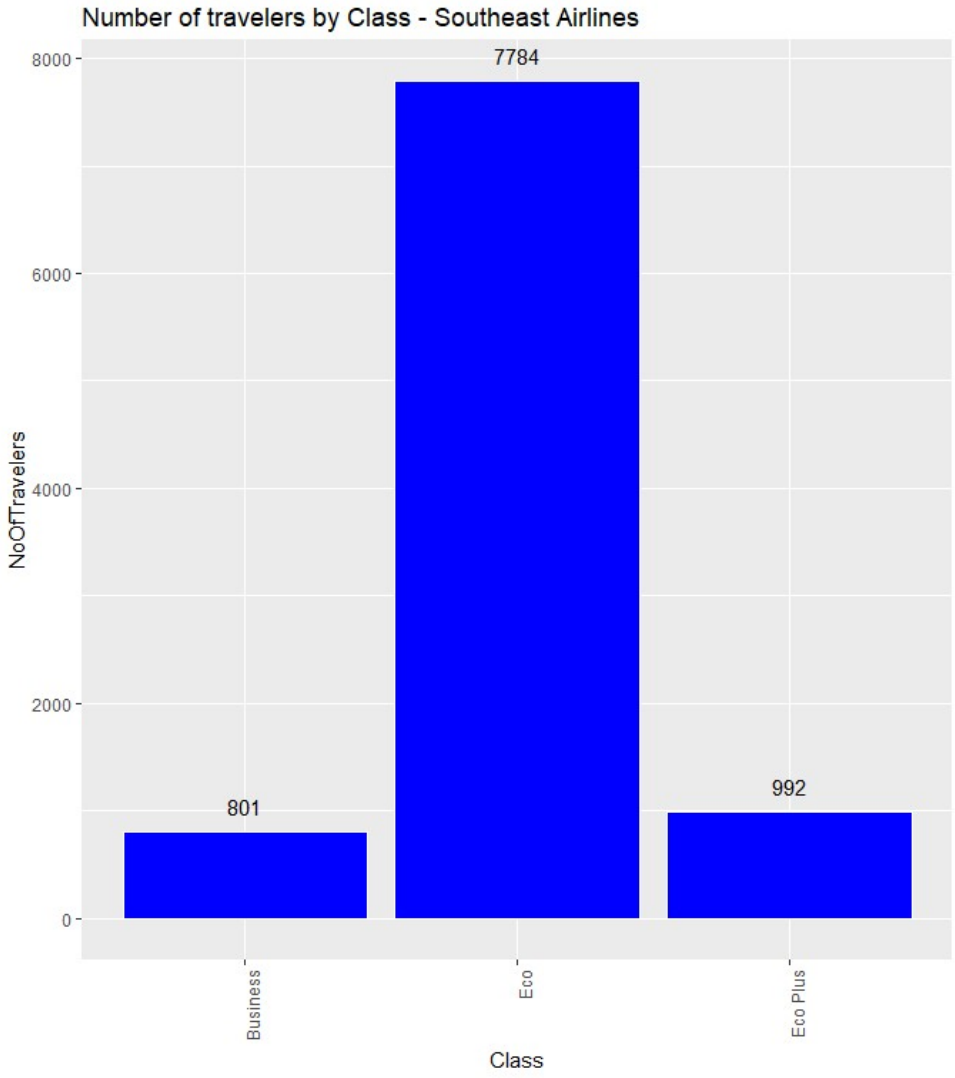


Figure 6: Total Count of Travelers by Class

Figure 6 shows the total number of travelers by class. Economy is the most used class by travelers. In detail, 7784 customers traveled in economy, 801 in business and 992 in economy plus.

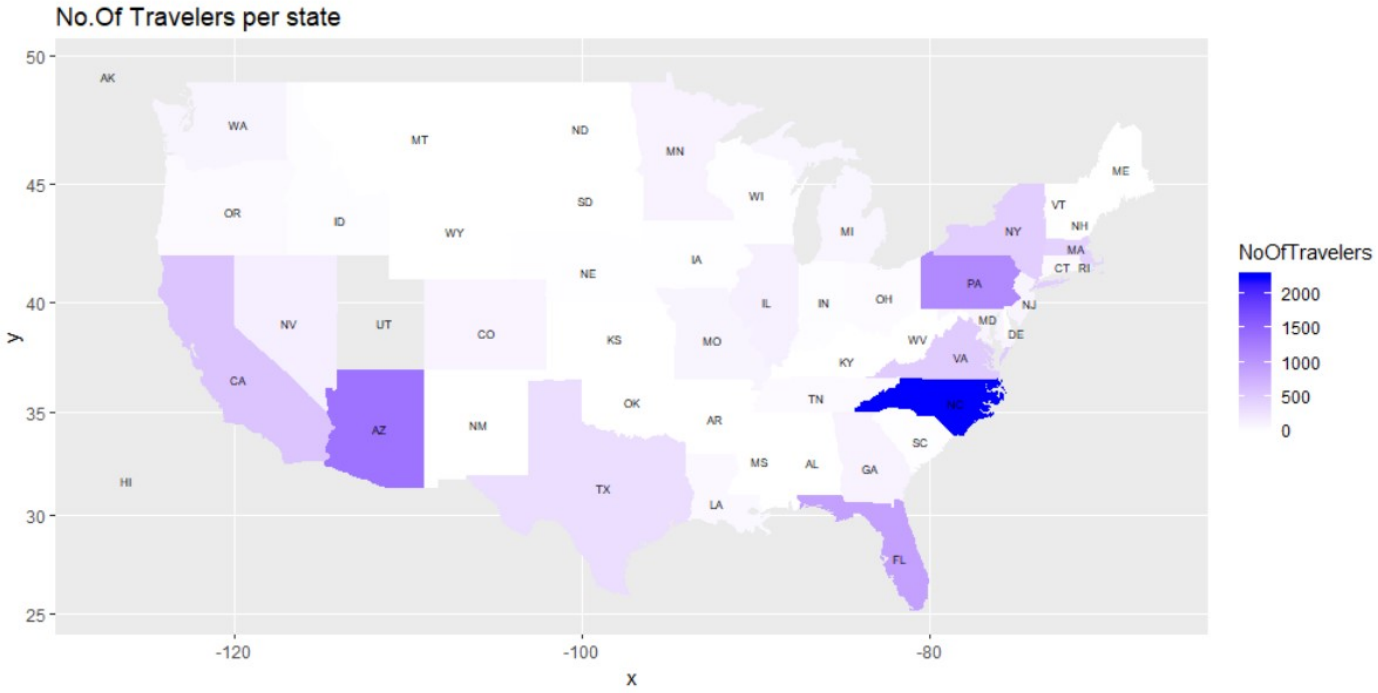


Figure 7: Map of the United States with colored States based on total count of travelers

Figure 7 shows the travelers distribution by state. As reflected on the map we can see that North Carolina and Arizona are the states with the most travelers to depart. Furthermore, we can see that there is a variety of states, such as Wyoming, that have no travelers departing at all.

C) Flight Performance:

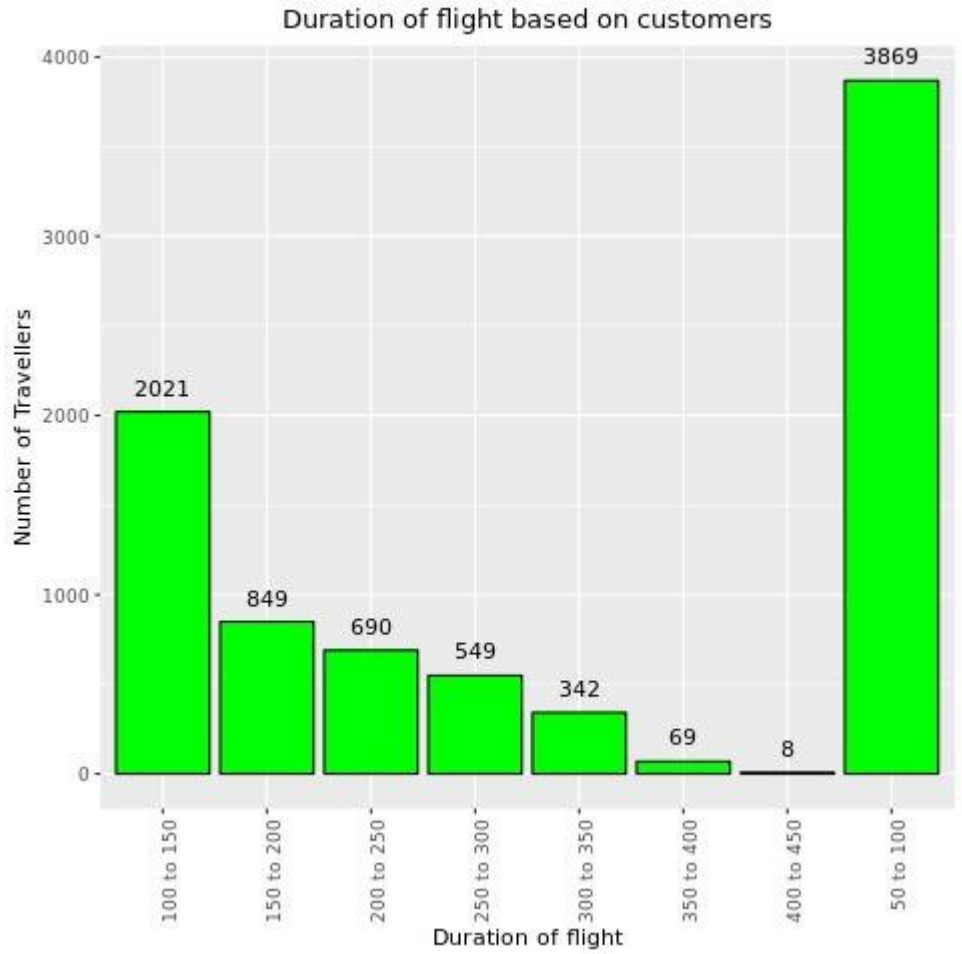


Figure 8: Total count of Travelers by Flight Duration

As figure 8 sows, most of the flights that Southeast Airline operates are of short duration. 3869 travelers used a flight with a total time between 50 and 100 minutes. Around 2021 travelers have traveled on a flight with a total of 100 to 150 minutes

Summary of Descriptive Statistics

- The total number of flights operated by Southeast Airline is 9577.
- Overall customer satisfaction is very similar across airlines. The average customer satisfaction for all customers is 3.37.
- 58 percent of travelers are female, while 42 percent of the customers are male.
- Customers that are in the age group 36 to 42 provide the highest customer satisfaction, while customers that are older than 78 years old give the worst satisfaction rating.
- Business is the most popular type of travel, followed by personal and then mileage.
- Most of the customers fly in economy class. Economy class is followed by economy plus and then business.
- North Carolina and Arizona are the states with the most flights departing.
- The majority of flights that Southeast Airline is operating are of short duration ranging from 50 to 150 minutes.

Data Modeling

In order to derive evidence-based insights and understand which of the variables in the dataset the key drivers for overall customer satisfaction are, we have utilized three different modeling techniques for our data analysis.

A) Linear Regression Model

B) Association Rules Mining

C) Support Vector Machines

A) **Regression Analysis** is the starting point of our analysis as it allows us to monitor the statistical relationship between a variety of variables that are of interest to us. In detail, by creating linear regression models we can understand how selected variables, called independent or explanatory variables, impact one target variable, called the dependent or response variable. Customer satisfaction is going to be the dependent variable for each of our linear regression models as we seek to understand which factors impact customer satisfaction positively or negatively.

For the first linear regression model, we intentionally included all variables captured by the data to see the statistical significance and impact on the response variable – customer satisfaction. In other words, we utilize our first regression analysis to filter and exclude variables that are of no statistical significance. In order to understand whether a variable has statistical significance, we need to consider its p-value. A low p-value indicates that the null hypothesis of a given statistical model can be rejected. In plain language, an explanatory variable that has a low p-value, usually below .05, is considered to have a meaningful impact on the response variable, because changes in the response variable can be explained by changes in the explanatory variable.

Based on the p-values of the regression result below, we notice that the following 11 variables have a statistical impact on customer satisfaction:

- Airline Status, Age, Gender, Price Sensitivity, Years of First Flight, Number of Flights per Anno, Type of Travel, Shopping Amount, Eating and Drinking at Airport, Class, Schedule Departure Hour, and Arrival Delay > 5 Minutes.

Furthermore, we can use the R-squared measure to understand the quality of our model. For this model the adjusted R-Squared is 0.4467, which means that 44.67 percent of the variability in

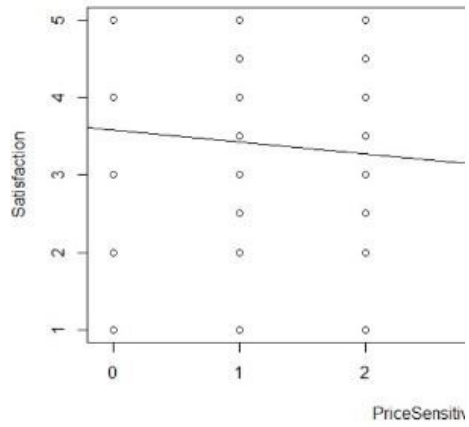
customer satisfaction can be explained by the variability of our explanatory variables. To put the obtained R-Squared in perspective, it is fair to say that it is rather low.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-5.8874834964165	1.3625369975479	-4.32097	0.00001554614832811	***
AirlineStatusGold	0.4417437245769	0.0074890958191	58.98492	< 0.000000000000000222	***
AirlineStatusPlatinum	0.2660931203717	0.0116421865679	22.85594	< 0.000000000000000222	***
AirlineStatusSilver	0.6198936443850	0.0052261472577	118.61389	< 0.000000000000000222	***
Age	-0.0023325424868	0.0001412557023	-16.51291	< 0.000000000000000222	***
GenderMale	0.1319215761721	0.0042229155536	31.23945	< 0.000000000000000222	***
PriceSensitivity	-0.0407848320542	0.0037625913759	-10.83956	< 0.000000000000000222	***
YearofFirstFlight	0.0048592835695	0.0006790146895	7.15637	0.00000000000083283	***
NoofFlightspa	-0.0033084332023	0.0001552960078	-21.30405	< 0.000000000000000222	***
XofFlightwithotherAirlines	-0.0000732184316	0.0002603244689	-0.28126	0.778513	
TypeofTravelMileage tickets	-0.1469139727383	0.0077842173746	-18.87331	< 0.000000000000000222	***
TypeofTravelPersonal Travel	-1.0763959767804	0.0049999433533	-215.28163	< 0.000000000000000222	***
NoofotherLoyaltyCards	-0.0025517015450	0.0021440482847	-1.19013	0.233997	
ShoppingAmountatAirport	0.0001637322676	0.0000383293188	4.27172	0.00001941083424694	***
EatingandDrinkingatAirport	-0.0000864385700	0.0000396110912	-2.18218	0.029098	*
ClassEco	-0.0772388589806	0.0073842666130	-10.45992	< 0.000000000000000222	***
ClassEco Plus	-0.0706253217141	0.0094899413782	-7.44212	0.00000000000009970	***
DayofMonth	-0.0000953520779	0.0002345198163	-0.40658	0.684314	
ScheduledDepartureHour	0.0038134282843	0.0004433079308	8.60221	< 0.000000000000000222	***
Flightcancelled	NA	NA	NA	NA	
DepartureDelayinMinutes	0.0000604668421	0.0002135666252	0.28313	0.777079	
ArrivalDelayinMinutes	0.0000269149478	0.0002167322782	0.12419	0.901169	
Flighttimeinminutes	-0.0000006056712	0.0001389041315	-0.00436	0.996521	
FlightDistance	0.0000045139449	0.0000167835137	0.26895	0.787968	
ArrivalDelaygreater5Minsyes	-0.3446001814449	0.0050959883840	-67.62185	< 0.000000000000000222	***

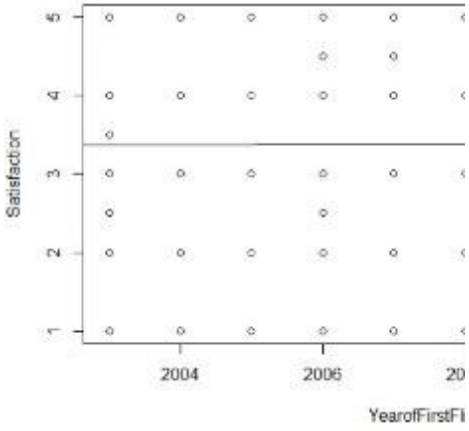
For the second linear regression model, we have included only the explanatory variables that deemed to be significant in the first regression analysis. By including only significant variables, we hoped to see an increase in the adjusted R-squared. In fact, the R-squared increased to .448, which indicates that we can better explain customer satisfaction with the below model and confirms the selected variables are actually the most relevant.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-5.69073409989	1.34879770975	-4.21912	0.000024542811810621	***
AirlineStatusGold	0.43764381068	0.00740644295	59.08961	< 0.000000000000000222	***
AirlineStatusPlatinum	0.25368141654	0.01151245745	22.03538	< 0.000000000000000222	***
AirlineStatusSilver	0.62207152298	0.00515190941	120.74582	< 0.000000000000000222	***
Age	-0.00230894655	0.00012579043	-18.35550	< 0.000000000000000222	***
GenderMale	0.12956484274	0.00414772509	31.23757	< 0.000000000000000222	***
PriceSensitivity	-0.03915252559	0.00370548499	-10.56610	< 0.000000000000000222	***
YearofFirstFlight	0.00475483563	0.00067215039	7.07407	0.000000000001512153	***
NoofFlightspa	-0.00320517482	0.00015074947	-21.26160	< 0.000000000000000222	***
TypeofTravelMileage tickets	-0.14264937838	0.00769176532	-18.54573	< 0.000000000000000222	***
TypeofTravelPersonal Travel	-1.07081635423	0.00491199816	-218.00015	< 0.000000000000000222	***
ShoppingAmountatAirport	0.00016573029	0.00003790053	4.37277	0.000012277506544607	***
ClassEco	-0.07987772219	0.00735015590	-10.86749	< 0.000000000000000222	***
ClassEco Plus	-0.07240307993	0.00940997952	-7.69429	0.000000000000014328	***
ScheduledDepartureHour	0.00366867274	0.00043476665	8.43826	< 0.000000000000000222	***
ArrivalDelaygreater5Minsyes	-0.33649064263	0.00423496839	-79.45529	< 0.000000000000000222	***

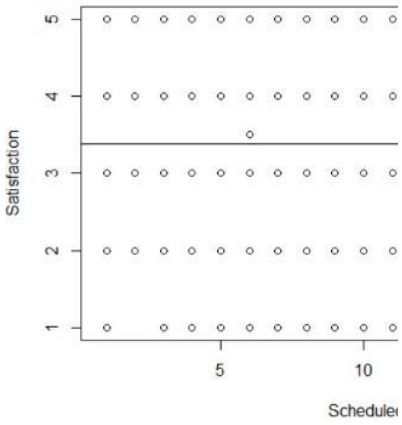
Since analyzing each of the 11 variables would exceed the scope of this project, we attempted to rank variables based on their importance with respect to customer satisfaction. Because both the regression coefficient as well as the p-value cannot be used as indicators for variable importance, we use single regression models and stepwise regression to evaluate variable importance. We manually performed a stepwise regression model to see how the adjusted R-squared changes by adding each of the significant variables individually. The variables that cause the highest increase in the adjusted R-squared are the ones with the greatest importance. Furthermore, we have developed single regression models to better understand the correlation between customer satisfaction and each of the variables. Based on the stepwise regression as well as the single regression models, which are outputted below, we determined that Airline Status and Type of Travel are the two variables with the greatest importance and, therefore, required further indepth analysis.



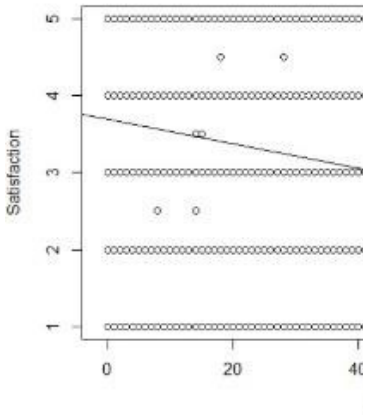
R^2 = 0.007641272



R^2 = 5.270168e-05

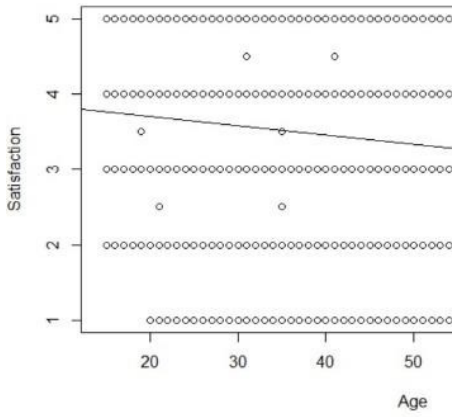


R^2 = -6.981177e-06

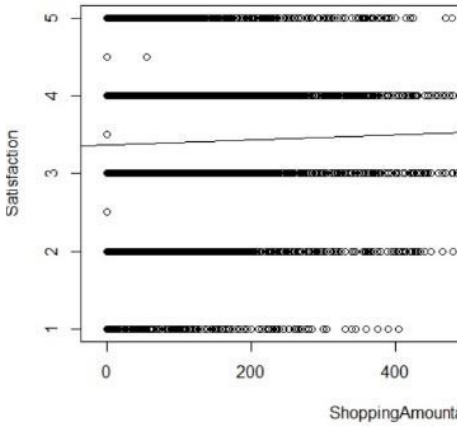


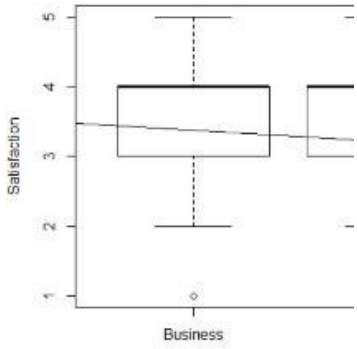
R^2 = 0.05671105

R^2 = 0.0492023



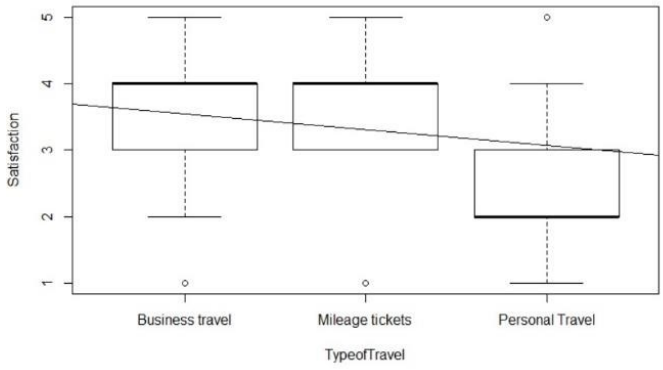
R^2 = 0.0002999279



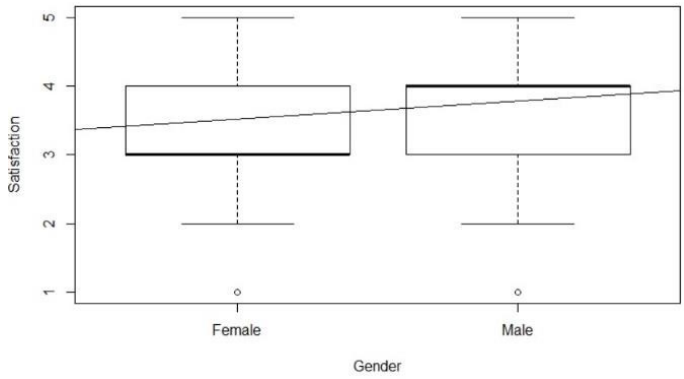


R^2 = 0.002526544

R^2 = 0.3350338

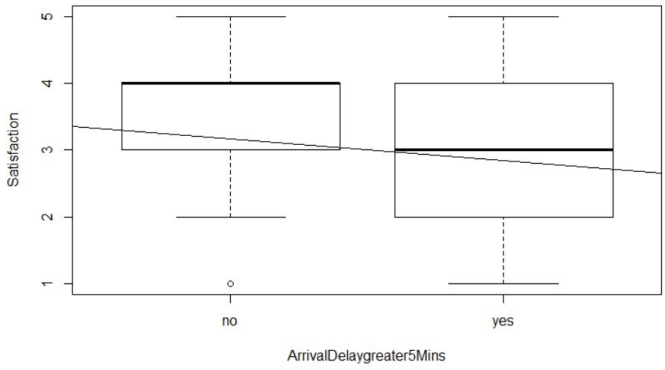


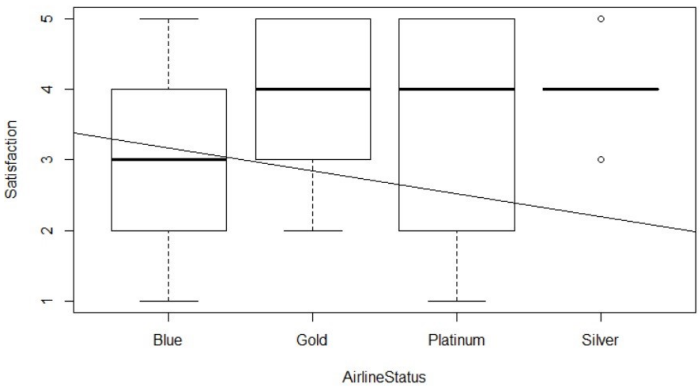
R^2 = 0.01760919



R^2 = 0.0252886

R^2 = 0.1184333





B) **Association Rules Mining:** After determining which the key drivers of customer satisfaction are, we used R's arules package to perform association rules mining for only customers of Southeast Airline. Association rules mining is a data mining technique that is primarily used to discover and understand the co-occurrence of variables. The goal of this analysis was to find a combination of variables that inevitably lead to high or low customer satisfaction. The target variable in association rule mining model cannot be numerical and, therefore, customer satisfaction had to be converted from numerical to nominal. Purely for the purpose of this analysis, we classified customer satisfaction ratings lower than two to low, equal to three as average, and greater than three as high. Our rule selection was primarily

based on the lift parameter, which is a measure of rule importance, to avoid trivial rule generation. The below scatterplot and table depict the 18 most interesting rules we have selected for high customer satisfaction.

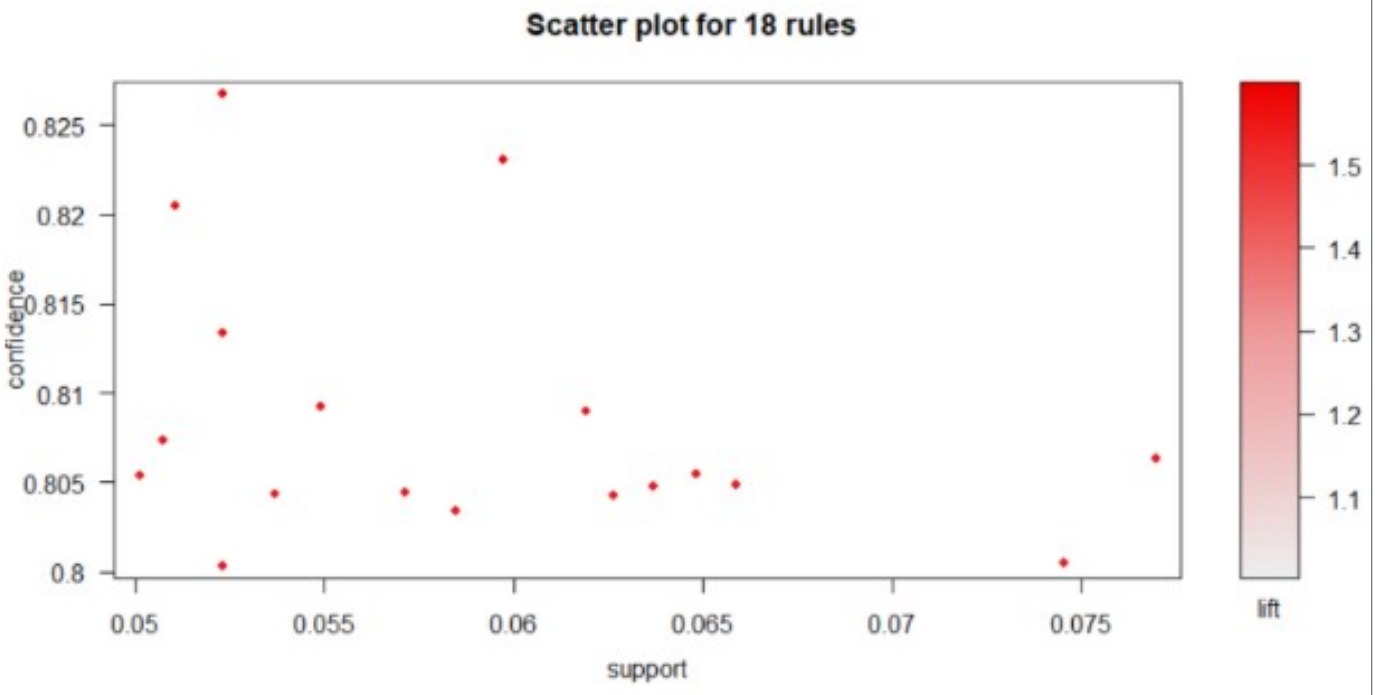


Figure 9: Scatter plot for association rules for high customer satisfaction

```
     lhs                                    rhs                          support   confidence        lift count
[1]  {df.Gender=Male,
      yearoffirstflight=High,
      df.TypeofTravel=Business travel,
      df.ArrivalDelaygreater5Mins=no}   => {Satisfaction=High} 0.05847342592 0.8034433286 1.551325959   560
[2]  {age=Low,
      df.TypeofTravel=Business travel,
      scheduleddepartureHour=High,
      df.ArrivalDelaygreater5Mins=no}   => {Satisfaction=High} 0.06191918137 0.8090040928 1.562062943   593
[3]  {noofflightspa=Low,
      df.TypeofTravel=Business travel,
      scheduleddepartureHour=High,
      df.ArrivalDelaygreater5Mins=no}   => {Satisfaction=High} 0.06369426752 0.8047493404 1.553847668   610
[4]  {df.Gender=Male,
      df.TypeofTravel=Business travel,
      scheduleddepartureHour=High,
      df.ArrivalDelaygreater5Mins=no}   => {Satisfaction=High} 0.05972642790 0.8230215827 1.589128568   572
[5]  {age=Low,
      df.Gender=Male,
      df.TypeofTravel=Business travel,
      df.ArrivalDelaygreater5Mins=no}   => {Satisfaction=High} 0.07455361804 0.8004484305 1.545543270   714
[6]  {df.Gender=Male,
      noofflightspa=Low,
      df.TypeofTravel=Business travel,
      df.ArrivalDelaygreater5Mins=no}   => {Satisfaction=High} 0.07695520518 0.8063457330 1.556930058   737
[7]  {pricesensitive=Low,
      yearoffirstflight=High,
      df.TypeofTravel=Business travel,
      shoppingamount=Low,
      df.ArrivalDelaygreater5Mins=no}   => {Satisfaction=High} 0.05231283283 0.8003194888 1.545294303   501
[8]  {age=Low,
      df.TypeofTravel=Business travel,
      df.Class=Eco,
      scheduleddepartureHour=High,
      df.ArrivalDelaygreater5Mins=no}   => {Satisfaction=High} 0.05231283283 0.8133116883 1.570380250   501
[9]  {pricesensitive=Low,
      noofflightspa=Low,
      df.TypeofTravel=Business travel,
      scheduleddepartureHour=High,
      df.ArrivalDelaygreater5Mins=no}   => {Satisfaction=High} 0.05105983084 0.8204697987 1.584201464   489
[10] {noofflightspa=Low,
      df.TypeofTravel=Business travel,
      df.Class=Eco,
      scheduleddepartureHour=High,
      df.ArrivalDelaygreater5Mins=no}   => {Satisfaction=High} 0.05074658035 0.8073089701 1.558789921   486
[11] {df.Gender=Male,
      df.TypeofTravel=Business travel,
      df.Class=Eco,
      scheduleddepartureHour=High,
      df.ArrivalDelaygreater5Mins=no}   => {Satisfaction=High} 0.05231283283 0.8267326733 1.596294115   501
[12] {df.Gender=Male,
      pricesensitive=Low,
      df.TypeofTravel=Business travel,
      df.Class=Eco,
      scheduleddepartureHour=High}      => {Satisfaction=High} 0.05711600710 0.8044117647 1.553195861   547
[13] {age=Low,
      df.Gender=Male,
      pricesensitive=Low,
      df.TypeofTravel=Business travel,
      df.ArrivalDelaygreater5Mins=no}   => {Satisfaction=High} 0.05492325363 0.8092307692 1.562500620   526
[14] {age=Low,
      df.Gender=Male,
      df.TypeofTravel=Business travel,
      df.Class=Eco,
      df.ArrivalDelaygreater5Mins=no}   => {Satisfaction=High} 0.06484285267 0.8054474708 1.555195651   621
[15] {df.Gender=Male,
      noofflightspa=Low,
      df.TypeofTravel=Business travel,
      shoppingamount=Low,
      df.ArrivalDelaygreater5Mins=no}   => {Satisfaction=High} 0.05012007936 0.8053691275 1.555044382   480
[16] {df.Gender=Male,
      pricesensitive=Low,
      noofflightspa=Low,
      df.TypeofTravel=Business travel,
      df.ArrivalDelaygreater5Mins=no}   => {Satisfaction=High} 0.06265009920 0.8042895442 1.552959872   600
[17] {df.Gender=Male,
      noofflightspa=Low,
      df.TypeofTravel=Business travel,
      df.Class=Eco,
      df.ArrivalDelaygreater5Mins=no}   => {Satisfaction=High} 0.06588702099 0.8048469388 1.554036115   631
```

```
[18] {df.Gender=Male,
      pricesensitive=Low,
      noofflightspa=Low,
      df.TypeofTravel=Business travel,
      df.Class=Eco,
      df.ArrivalDelaygreater5Mins=no}  => {Satisfaction=High} 0.05367025164 0.8043818466 1.553138094   514
```

Since we believe that 18 rules are still too many to understand patterns and key drivers of customer satisfaction, we have included only the top five rules in our final association rules mining model. The visualization of the top five rules below indicate that a combination of the following attributes tends to result in high customer satisfaction:

- Business Flight, Male Traveler, Evening Flight, No Delay, Low Price Sensitivity, and Young Customer.

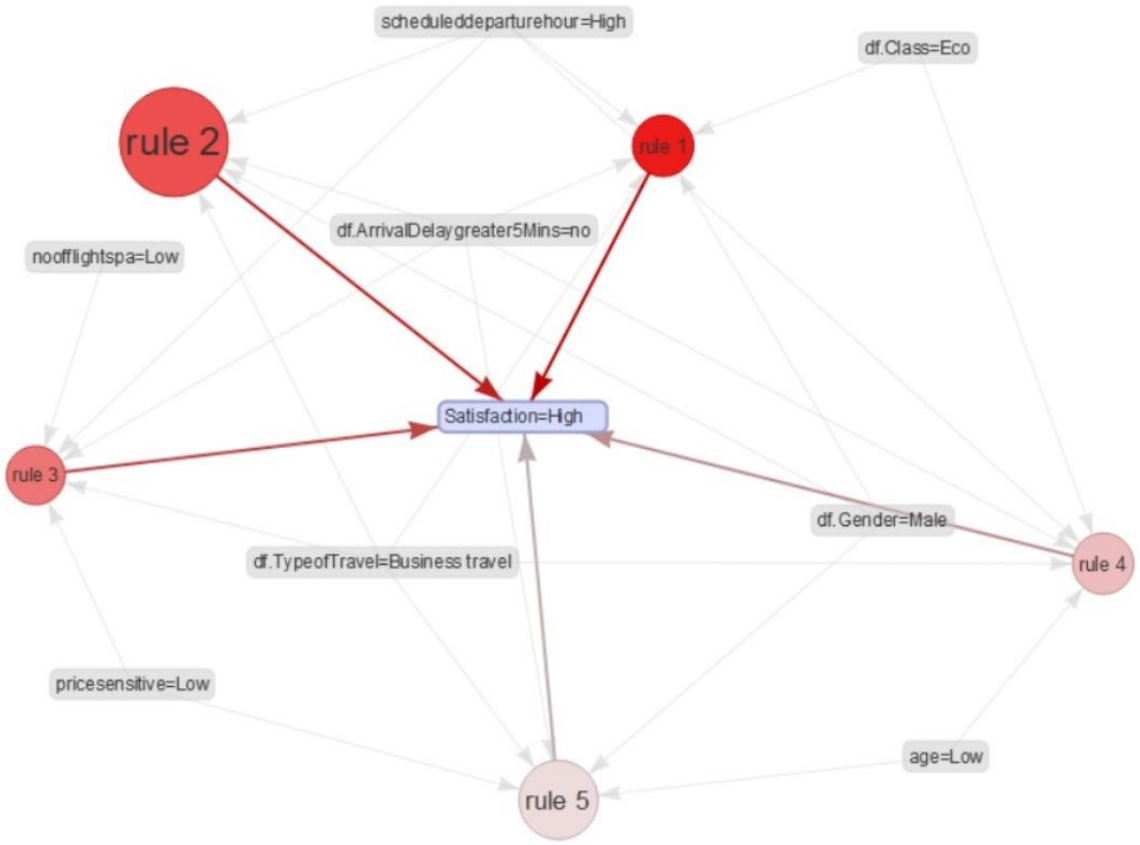


Figure 10: Visualization of top 5 rules with high customer satisfaction

Besides generating rules for high customer satisfaction, we also focused on generating rules to predict what leads to low customer satisfaction. A similar modeling approach was used for the association rules mining model when predicting low customer satisfaction. However, as reflected in the confidence level in the scatterplot below, it was much harder to generate rules for low customer satisfaction. Therefore, we have chosen a lower confidence interval.

The following scatter plot and table show the 20 rules that contribute to low customer satisfaction.

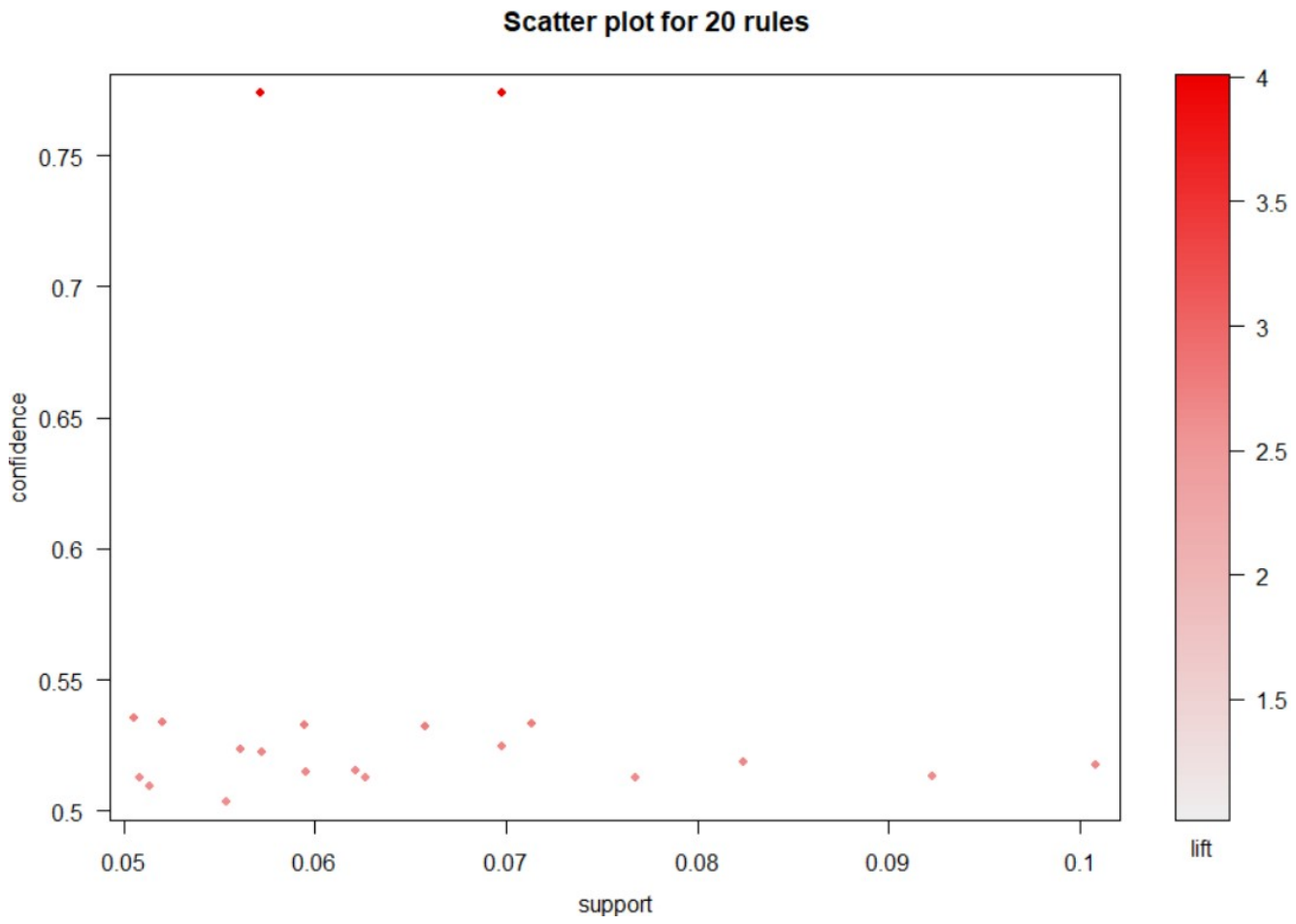


Figure 11: Scatter plot of association rules with low customer satisfaction

```
     lhs                                  rhs                     support    confidence        lift count
[1]  {df.TypeofTravel=Personal Travel,
      df.ArrivalDelaygreater5Mins=yes} => {Satisfaction=Low} 0.06975044377 0.7740440324 3.994083889   668
[2]  {age=High,
      df.ArrivalDelaygreater5Mins=yes} => {Satisfaction=Low} 0.05534092096 0.5033238367 2.597161845   530
[3]  {noofflightspa=High,
      df.TypeofTravel=Personal Travel} => {Satisfaction=Low} 0.09230447948 0.5130586187 2.647393530   884
[4]  {age=High,
      df.TypeofTravel=Personal Travel} => {Satisfaction=Low} 0.10086665971 0.5176848875 2.671265176   966
[5]  {df.TypeofTravel=Personal Travel,
      df.Class=Eco,
      df.ArrivalDelaygreater5Mins=yes} => {Satisfaction=Low} 0.05711600710 0.7736916549 3.992265614   547
[6]  {age=High,
      noofflightspa=High,
      df.TypeofTravel=Personal Travel} => {Satisfaction=Low} 0.07131669625 0.5331772053 2.751205870   683
[7]  {noofflightspa=High,
      df.TypeofTravel=Personal Travel,
      shoppingamount=Low}              => {Satisfaction=Low} 0.05951759424 0.5149051491 2.656921666   570
[8]  {df.Gender=Female,
      noofflightspa=High,
      df.TypeofTravel=Personal Travel} => {Satisfaction=Low} 0.05722042393 0.5224022879 2.695607064   548
[9]  {pricesensitive=Low,
      noofflightspa=High,
      df.TypeofTravel=Personal Travel} => {Satisfaction=Low} 0.06212801504 0.5151515152 2.658192921   595
[10] {noofflightspa=High,
      df.TypeofTravel=Personal Travel,
      df.Class=Eco}                    => {Satisfaction=Low} 0.07674637152 0.5125523013 2.644780921   735
[11] {age=High,
      df.TypeofTravel=Personal Travel,
      shoppingamount=Low}              => {Satisfaction=Low} 0.06265009920 0.5128205128 2.646164898   600
[12] {age=High,
      df.Gender=Female,
      df.TypeofTravel=Personal Travel} => {Satisfaction=Low} 0.06578260416 0.5320945946 2.745619576   630
[13] {age=High,
      pricesensitive=Low,
      df.TypeofTravel=Personal Travel} => {Satisfaction=Low} 0.06975044377 0.5247446976 2.707693949   668
[14] {age=High,
      df.TypeofTravel=Personal Travel,
      df.Class=Eco}                    => {Satisfaction=Low} 0.08238488044 0.5183968463 2.674938899   789

[15] {age=High,
      pricesensitive=Low,
      noofflightspa=High,
      df.TypeofTravel=Personal Travel} => {Satisfaction=Low} 0.05053774668 0.5353982301 2.762666406   484
[16] {age=High,
      noofflightspa=High,
      df.TypeofTravel=Personal Travel,
      df.Class=Eco}                    => {Satisfaction=Low} 0.05941317740 0.5327715356 2.749112606   569
[17] {pricesensitive=Low,
      noofflightspa=High,
      df.TypeofTravel=Personal Travel,
      df.Class=Eco}                    => {Satisfaction=Low} 0.05085099718 0.5126315789 2.645189995   487
[18] {age=High,
      df.TypeofTravel=Personal Travel,
      shoppingamount=Low,
      df.Class=Eco}                    => {Satisfaction=Low} 0.05137308134 0.5093167702 2.628085511   492
[19] {age=High,
      df.Gender=Female,
      df.TypeofTravel=Personal Travel,
      df.Class=Eco}                    => {Satisfaction=Low} 0.05199958233 0.5337620579 2.754223722   498
[20] {age=High,
      pricesensitive=Low,
      df.TypeofTravel=Personal Travel,
      df.Class=Eco}                    => {Satisfaction=Low} 0.05607183878 0.5233918129 2.700713034   537
```

Similarly, we chose the top five “most interesting” rules based on the lift factor for low customer satisfaction. The major contributor to low customer satisfaction is a combination of the following attributes:

 Delay greater > 5 Minutes, Personal Travel, Old Customer, Female, Number of Flights per Anno

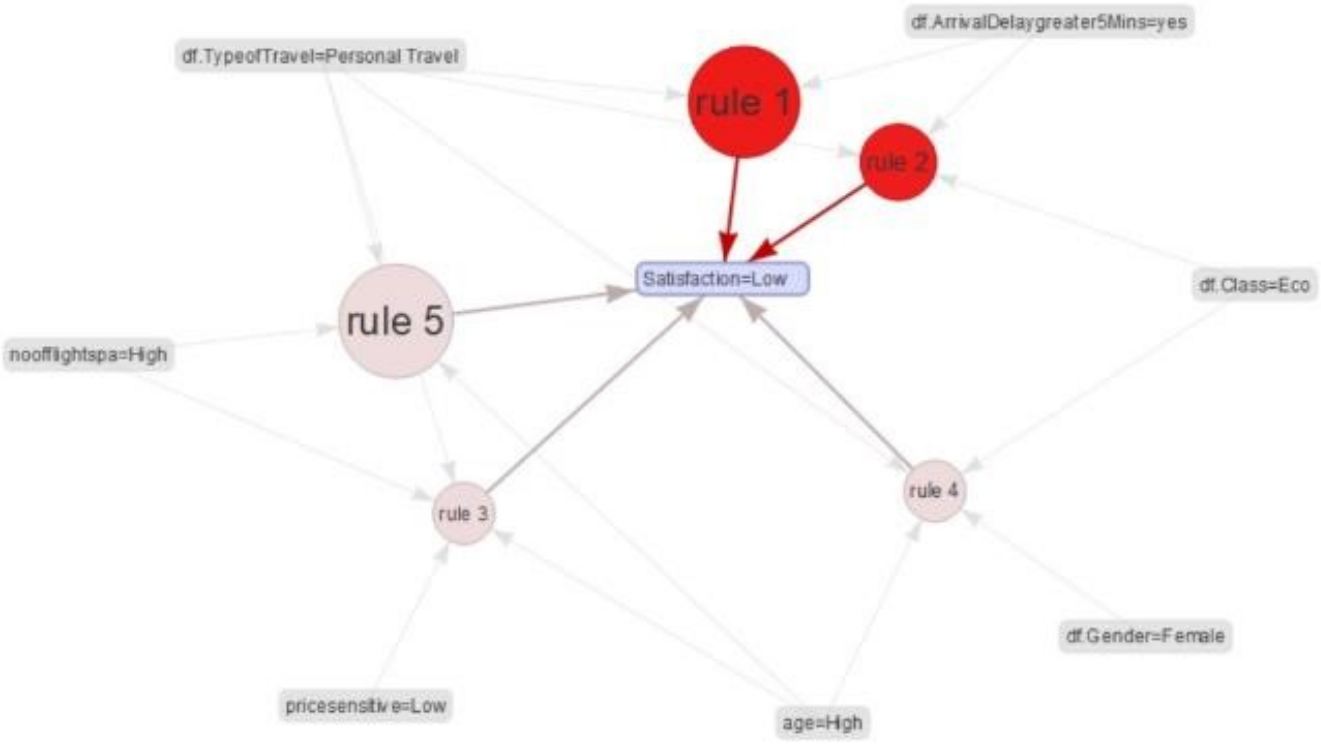


Figure 12: Visualization of top 5 rules with low customer satisfaction

B) **Support Vector Machine (SVM):** SVM is a supervised machine-learning algorithm that is used in data mining for classification purposes. SVM is considered supervised because the output dataset is already known. In this context, we are intending to build a classification model with customer satisfaction as the target variable. In other words, we are seeking to predict whether a customer has low, average, or high customer satisfaction based on the personal customer information, the flight details, and the flight performance. Assuming that the accuracy of the model is satisfying, Southeast Airline can use this SVM classification model as

a tool to continuously track customer satisfaction without generating any further surveys in the future. That means Southeast Airlines is longer dependent on customer feedback as they are able to generate their own predicted customer satisfaction rating.

The below confusion matrix shows us both statistically as well as visually that the SVM classification model was implemented successfully. Testing our SVM model on a test dataset resulted in an accuracy of approximately 79.6%. Out of the 3193 customers contained in the test data, the SVM model was able to predict 1527 satisfied customers correctly, while incorrectly predicting 113 customers. Moreover, the model correctly classified 1014 customers as dissatisfied while incorrectly classifying 539.

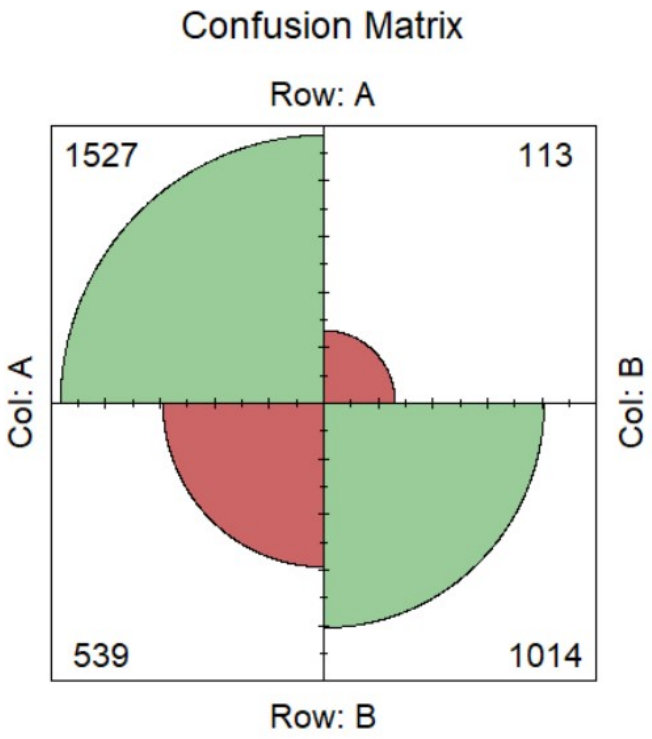


Figure 13: Confusion Matrix of 1st SVM classification model

As seen in the graph as well as mentioned above, we have two different types of errors. The satisfied customers that are classified incorrectly as dissatisfied and the dissatisfied customers that are classified incorrectly as satisfied. These two type of errors introduce the notion of precision and recall. In a perfect scenario, we would like to eliminate both of the errors. However, since that is not possible we need to balance our model and understand which of the two errors is preferred in our context and against which error specifically we want to protect our model. As a service provider, Southeast Airline wants to understand and analyze the customers that give low customer satisfaction in order to improve its business. Therefore, we are more interested in correctly classifying customers that are dissatisfied as opposed to correctly classifying customers that are satisfied. Based on that assumption we

have adjusted our SVM model by changing the C parameter. Below is the graph of our second SVM classification model. We can see that while the overall accuracy of our model decreased to 76.1%, we are not able to better at classifying dissatisfied customers.

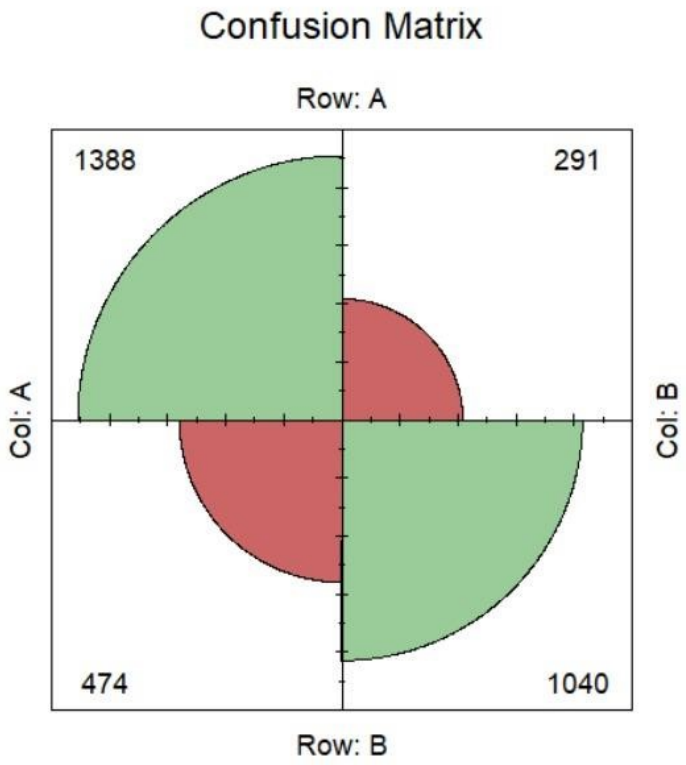


Figure 14: Confusion Matrix of 2nd SVM classification model

Data Analysis and Visualization

Since our regression analysis showed that both airline status, as well as type of travel, are the two major drivers of customer satisfaction we attempted to shed some more light on which affect the specific airline statuses – blue, silver, gold, or platinum – and types of travel – business, mileage, or personal – have on overall customer satisfaction.

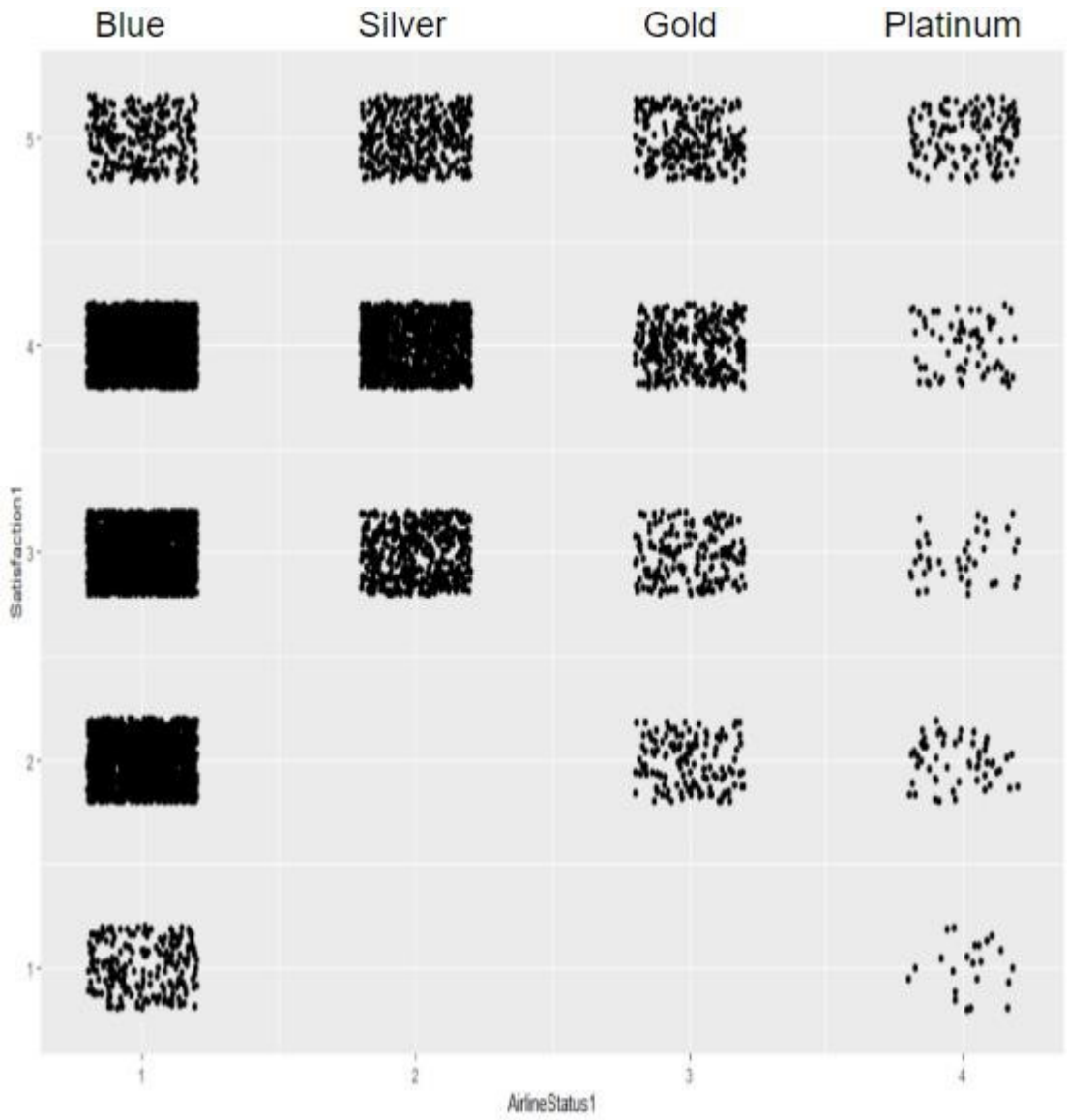


Figure 15: Scatter plot of Satisfaction and Airline status

The above graph depicts a scatterplot of each customers satisfaction rating based on the airline status he or she belongs to. It becomes clear that not only do most customers fly with airline status blue, but also blue customers tend to give the lowest customer satisfaction rating. As the density of the plot shots, travelers belonging to airline status blue are inclined to give a rating between two and four. Airline status silver is doubtlessly the status with the most success. Not a single customer chose to give a bad rating (one or two) and most customers give a rating of four. A similar pattern can be seen for airline status gold. Not a single customer gave a rating of lower than two, while most customers gave a rating of four or above. Counterintuitively, platinum, which is the most prestigious airline status, offered by Southeast Airlines shows no clear relationship with overall customer satisfaction. While most people tend to give a rating of five there is still a considerably high number of customers giving ratings below three. Because the

number of low customer ratings for airline status platinum was unexpected, we decided to dive deeper into the analysis of that particular airline status.

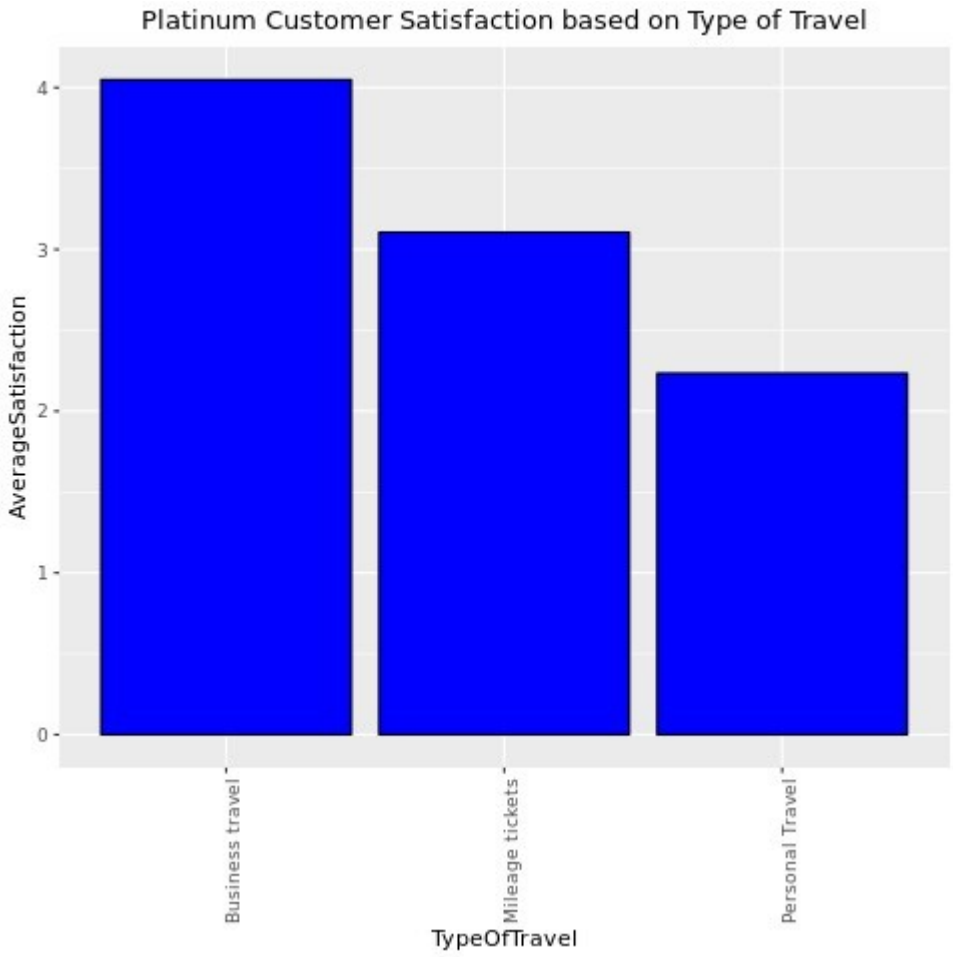


Figure 16: Bar chart for average customer satisfaction and type of travel

The above graph shows the average customer satisfaction based on the type of travel for customers belonging to airline status platinum. We can observe that personal travelers give a substantially lower satisfaction rating than both mileage and business travelers. Further rootcause analysis is required regarding this particular point, but it appears that personal travelers, who pay for their ticket and have achieved platinum status, are not quite satisfied with the benefits of being a platinum member. Achieving the highest airline status comes with a lot of flights and money spent and, therefore, customers seem to expect more for their commitment to Southeast Airlines.

A similar visualization can be made for the type of travel.

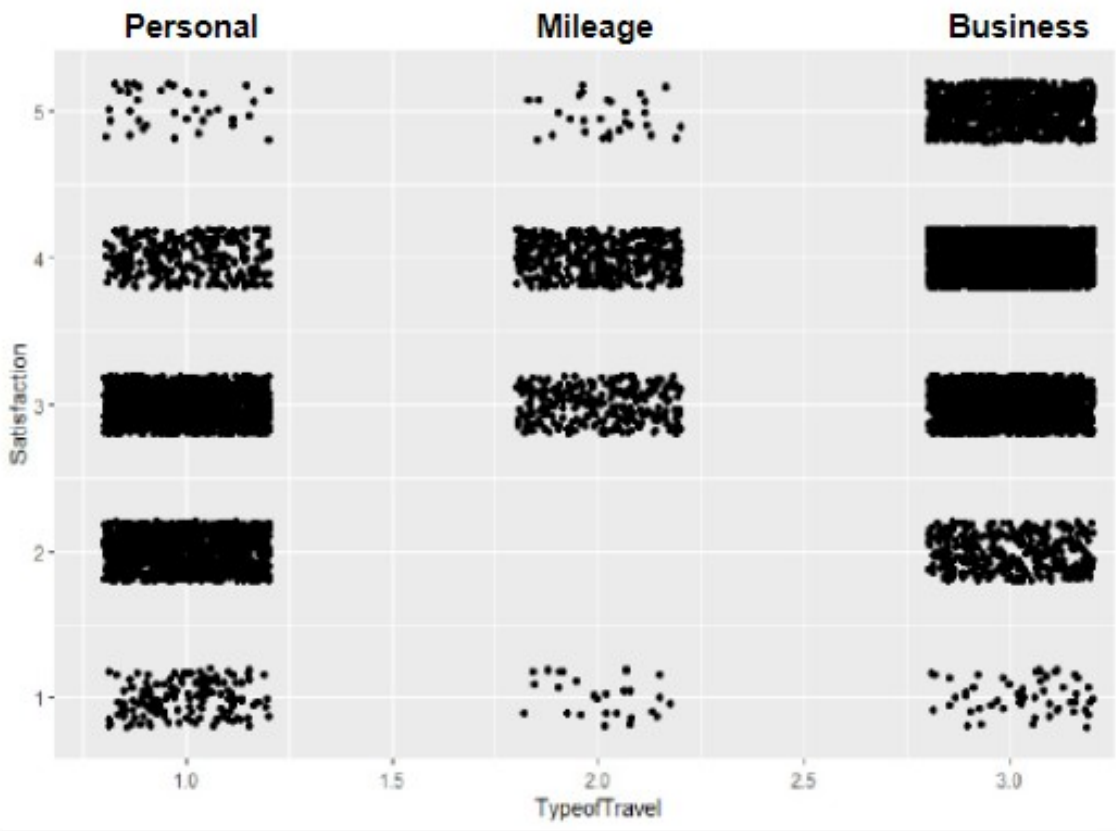


Figure 17: Scatter plot of satisfaction and type of travel

From the graph above, we can see that business travelers are by far the most satisfied across all airline statuses. The density is highest from three to five implying that most of the business traveler give positive satisfaction ratings. Similarly, for the mileage travelers, we can notice that most of them give a rating of three or four and only in rare case provide really good or really bad feedback. Contrary, personal travelers tend to give a rating of three or lower. While most of the customers are traveling business, it is still imperative to understand that personal travelers generally are dissatisfied with their experience with Southeast Airlines.

Based on the results of our linear regression model we were able to see that gender was also a statistically significant variable. Looking at the chart below we can see that female customer tend to give a lower satisfaction rating than men. This not only validates the findings of our association rules mining model but also makes us think about which factors make the experience for women different than for men.

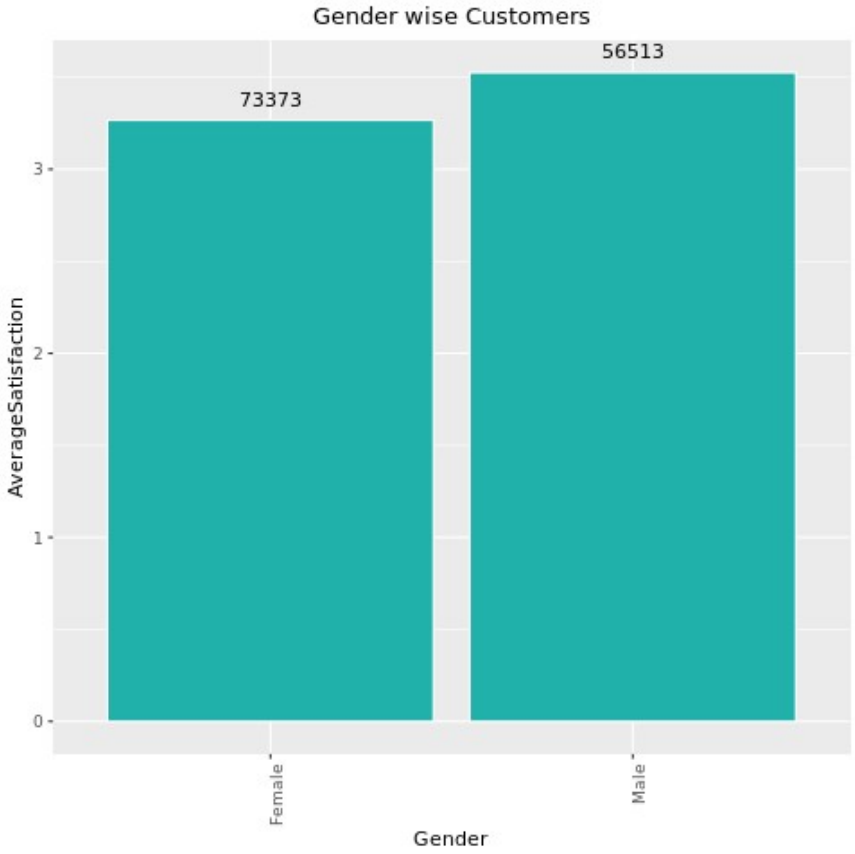


Figure 18: Bar chart of average customer satisfaction by gender

In order to better understand the differences between customers satisfactions based on gender we are going to add another variable to the plot – shopping at airport.

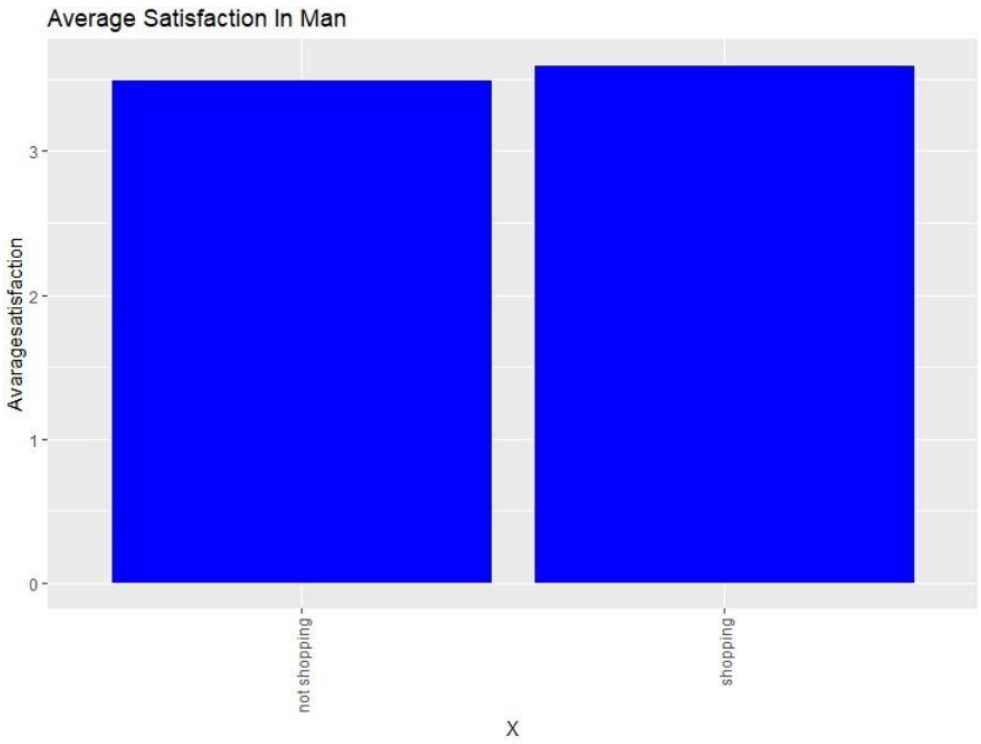


Figure 19: Average customer satisfaction for men that do and don't shop at airport

Figure 19 shows the average customer satisfaction for men based on whether or not they shop at the airport. While we can see that men are more satisfied when they spend money, it is fair to say that shopping at the airport has a limited effect on average customer satisfaction.

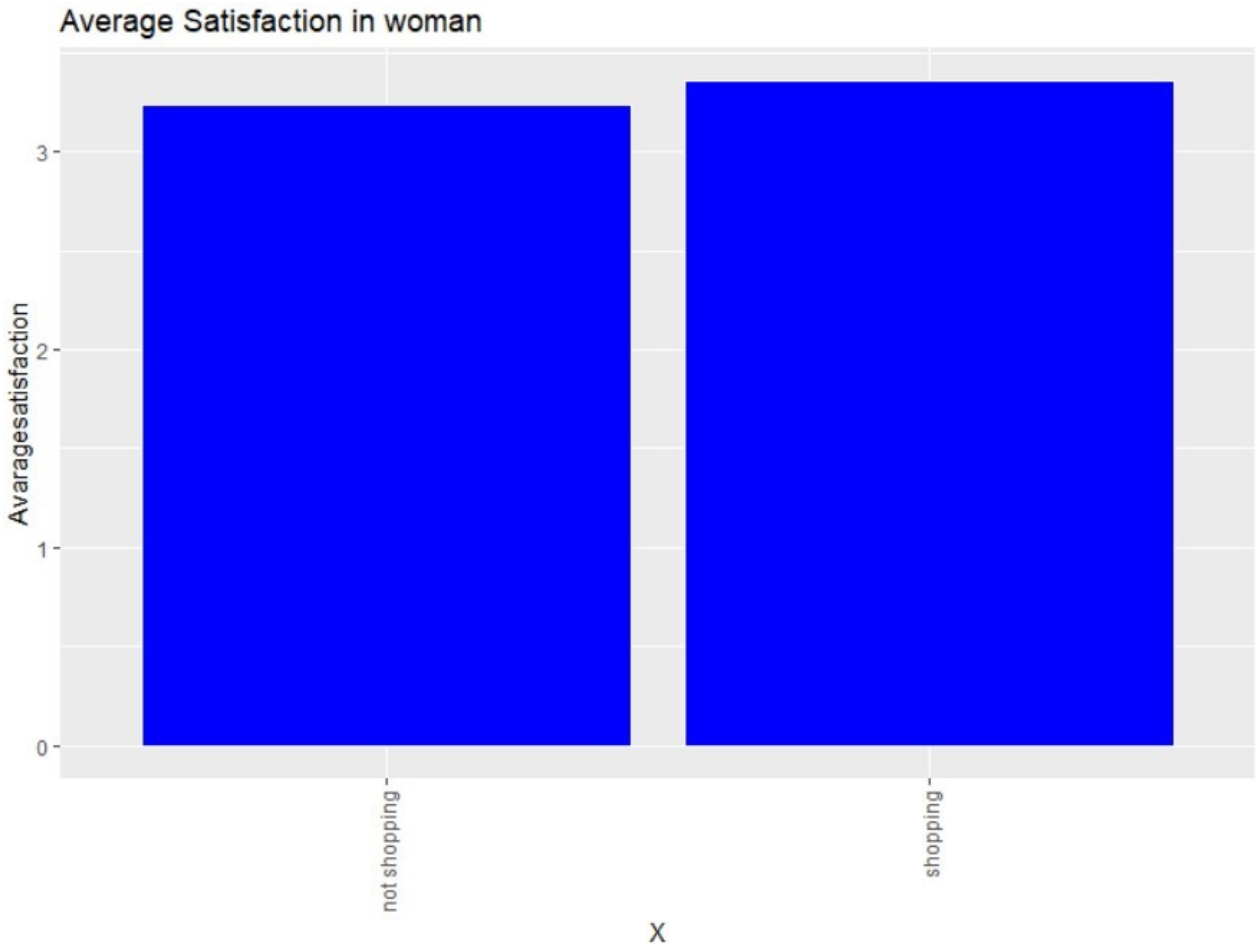


Figure 20: Average customer satisfaction for women that do and don't shop at airport

A similar graph can be created for women. Similarl to men, we can notice that women that go shopping at the airport are slightly more satisfied than women that don't shot. As with men, however, the difference is not very significant. In conclusion, we can observe a slight tendency that women and men that shop more are slightly more satisfied with their experience than women or men that don't go shopping.

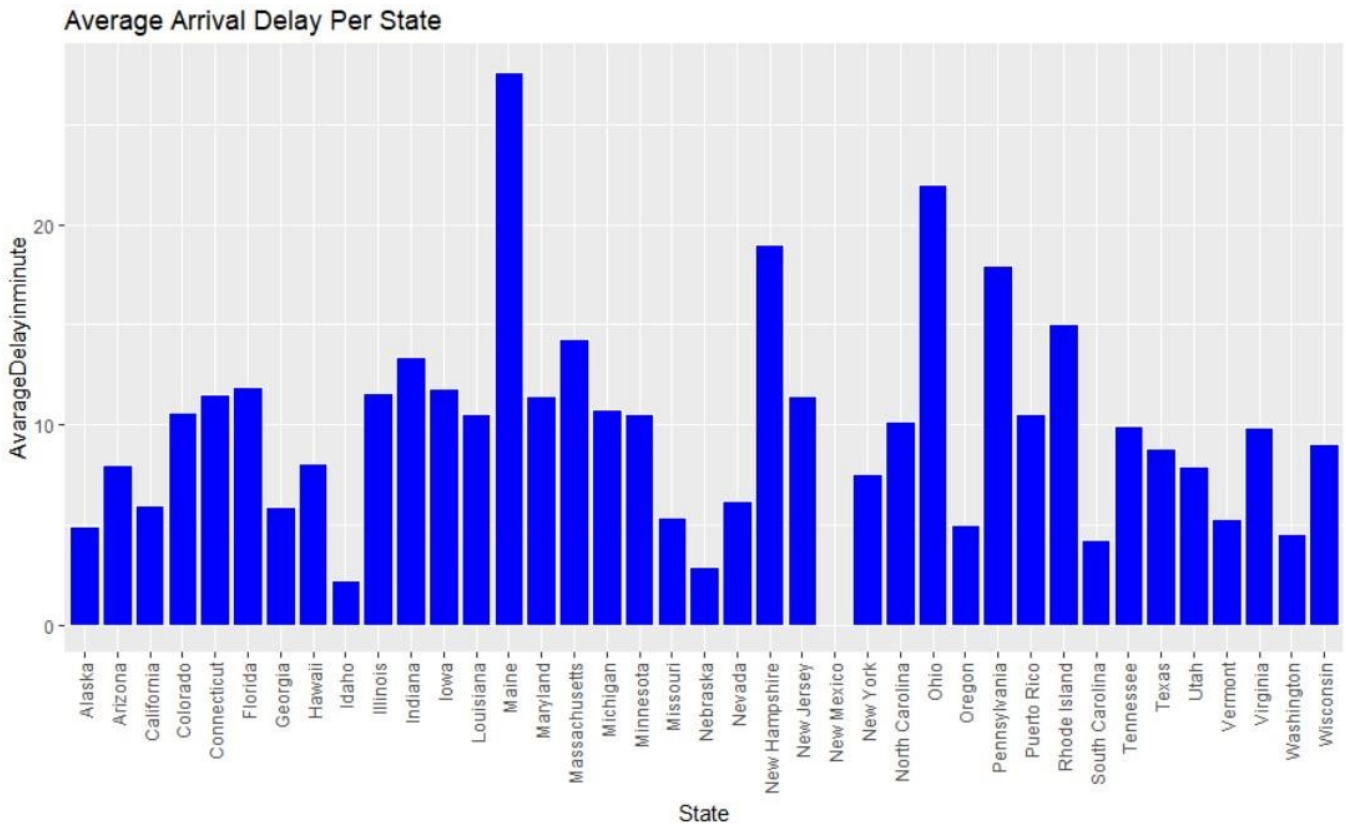


Figure 21: Average arrival delay per state

Figure 21 depicts the average arrival delay per state. What becomes noticeable right away is the fact that the two states, Maine and Ohio, are the states with the highest average arrival delay.

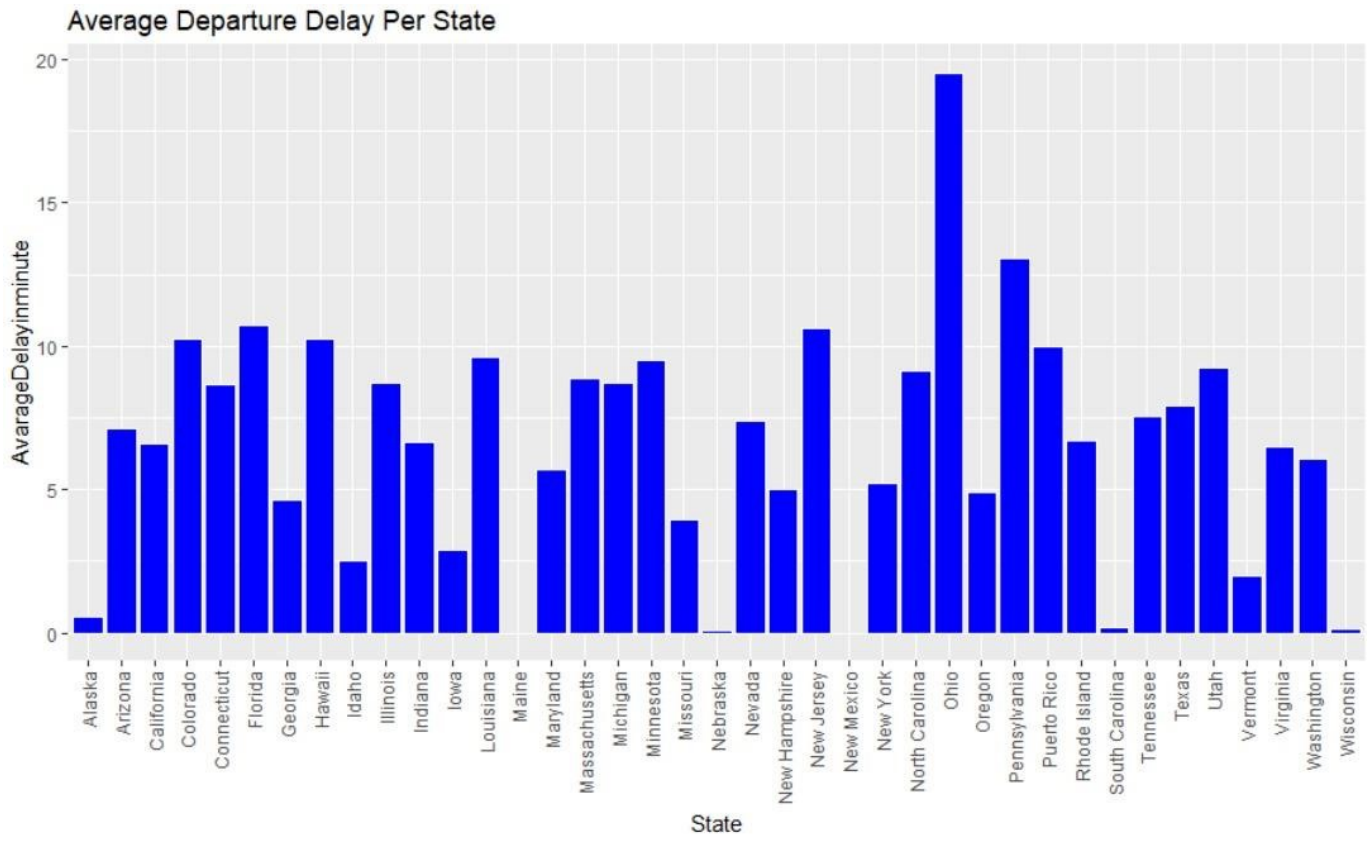


Figure 22: Average departure delay per state

Figure 22 shows the average departure delay per state. We can see that Ohio as well as Pennsylvania are the states with the greatest departure delay.

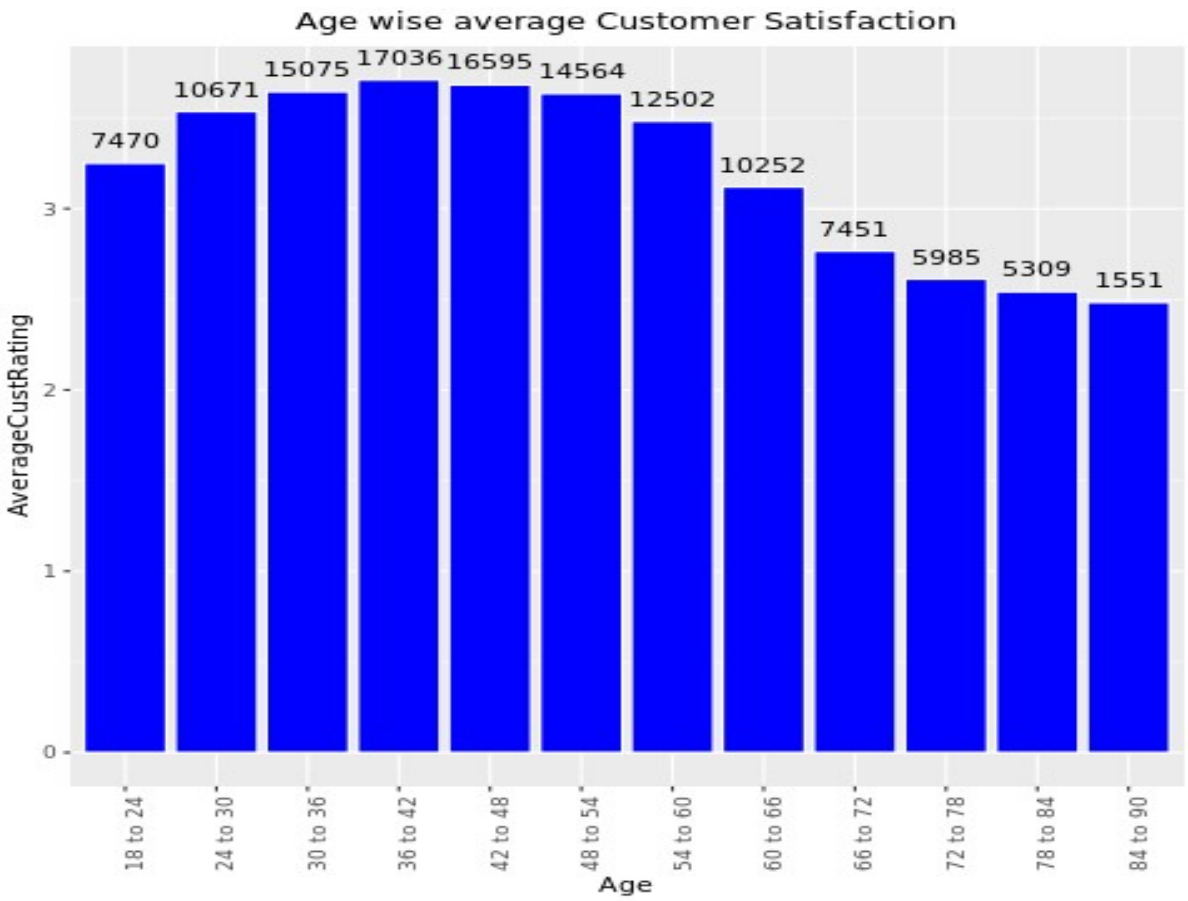


Figure 23: Average customer rating by age group

Visibly people in the higher age group are less satisfied. The highest satisfaction was among the people who belonged to the middle age group. Southeast airlines should look into improving the amenities provided to the higher age group people.

Key Drivers for Low Customer Satisfaction:

- Type of Travel: Personal Travel
- Airline Status: Blue and Platinum
- Age greater than 70 years
- Delay greater than 5 minutes
- Women give worse rating than men

Response: Business Questions

1) Airline status is the second most important driver after type of travel. Certain airline statuses such as silver or platinum tend to lead to higher customer satisfaction while other airline statuses such as blue and platinum lead to lower customer satisfaction.
2) Type of travel is the strongest explanatory variable in the dataset for customer satisfaction. People that travel business are constantly giving better ratings than people that travel with miles or personal.
3) Yes, certain states such as Ohio show both high departure as well as arrival delay. Delays are often times correlated with lack of operational efficiency.
4) The impact of shopping at the airport is limited. While we can see that shopping leads to higher satisfaction for both men and women, the difference is minimal.
5) Age, in fact, plays a quite significant role with respect to customer satisfaction. People within the age group of 36 to 42 are the travelers that give the highest customer satisfaction. Furthermore, we identified that people that are older than 70 years old are giving the worse ratings.

Recommendations and Actionable Insights:

1) Analyze which services are offered to customers with airline status silver and gold and understand which of those services can be offered at low cost to customers of airline status blue.

2) Analyze the difference between the airline statuses gold and platinum and consider providing platinum travelers with additional services, as they are not pleased by the current services provided.

3) Build partnerships with airport stores to provide Southeast Airline customers with store discounts to improve their shopping experience, which inevitably leads to higher customer satisfaction.
4) Delays are a critical driver for low customer satisfaction and, therefore, Southeast airline should attempt to minimize flight delays.
5) Improve the operational efficiency at airports in states that consistently show delays.

   Example list of States:

   - Ohio, New Hampshire, Pennsylvania, Massachusetts

6) Consider the impact customers of 70 years have on satisfaction and potentially provide them with extra amenities such as blankets or pillows to increase their comfortability.
7) Southeast Airline should implement a pricing strategy that differentiates between personal and business travelers. Personal travelers are price sensitive and tend to give a low customer satisfaction when the price is high.

Appendix: R code

Data Cleaning: str(raw_data)

```
CleanSatisfaction<-raw_data[(raw_data$Satisfaction=="1" |
raw_data$Satisfaction=="1.5" |          raw_data$Satisfaction=="2"
|          raw_data$Satisfaction=="2.5" |
raw_data$Satisfaction=="3" |          raw_data$Satisfaction=="3.5"
```

```
|           raw_data$Satisfaction=="4" |
raw_data$Satisfaction=="4.5" |           raw_data$Satisfaction=="5"
),]
#clean_data <- subset(raw_data,trimws(raw_data$Satisfaction)==c(1:5))
#clean_data df<-CleanSatisfaction
newCol<-colnames(CleanSatisfaction)
newCol<-gsub("\\.", "", newCol)
newCol colnames(df)<-newCol

a <- sub("No","0",df$Flightcancelled) b
<- sub("Yes","1",a) df$Flightcancelled
<- b
df$Flightcancelled<-as.numeric(df$Flightcancelled)

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

Dataset Subsetting:

```
fulldf<-df str(fulldf)
#Custdf<-subset(df,AirlineName=="Southeast")
Custdf<-df[df$AirlineName == 'Southeast Airlines Co. ',]
#summary(df)
#positive = posWords[which(posWords >=2)]
#df$AirlineName
df<-Custdf str(fulldf)
```

Association Rule Mining:

```
library(methods) summary(df) createBuckets<-
function(vec){   q <- quantile(vec, c(0.4, 0.6))
```

```
vBuckets <- replicate(length(vec), "Average")
vBuckets[vec <= q[1]] <- "Low"
vBuckets[vec > q[2]] <- "High"
return(vBuckets)
}
vBuckets<-replicate(length(df$Satisfaction),"Median")
vBuckets[df$Satisfaction>3]<-"High"
vBuckets[df$Satisfaction<3]<-"Low" Satisfaction<-
as.factor(vBuckets) age<-createBuckets(df$Age)
pricesensitive<-createBuckets(df$PriceSensitivity)
yearoffirstflight<-createBuckets(df$YearofFirstFlight)
noofflightspa<-createBuckets(df$NoofFlightspa)
shoppingamount<-
createBuckets(df$ShoppingAmountatAirport)
scheduleddepartureh our<-
createBuckets(df$ScheduledDepartureHour) library(arules)
library(arulesViz) ruleDF<-
data.frame(Satisfaction,df$AirlineStatus,age,df$Gender,prices
ensitive,yearoffirstflight,noofflight
spa,df$TypeofTravel,shoppingamount,df$Class,scheduleddepa
rturehour,df$ArrivalDelaygreater 5Mins)
#ruleDF<-
data.frame(Satisfaction,df$AirlineStatus,age,df$Gender,pricesensitive,yearoffirstflight,noofflight
spa,df$TypeofTravel,shoppingamount,df$Class,scheduleddepartureh our,df$ArrivalDelaygreater
5Mins)
hotelSurveyruleDFSE<-as(ruleDF,"transactions")
rulesetsoutheastH<- apriori(hotelSurveyruleDFSE, parameter=list(support=0.05,
confidence=0.8),appearance = list(default="lhs", rhs=("Satisfaction=High")))
goodrulesH<-sort(rulesetsoutheastH,by="lift")[1:5] inspect(goodrulesH)
```

```
plot1<-plot(goodrulesH, method = "graph", engine = "htmlwidget") rulesetsoutheastL<-
apriori(hotelSurveyruleDFSE, parameter=list(support=0.05,
confidence=0.5),appearance = list(default="lhs", rhs=("Satisfaction=Low")))
goodrules2<-sort(rulesetsoutheastL,by="lift")[1:10]
plot2<-plot(goodrules2, method = "graph", engine = "htmlwidget")
```

Linear Modeling:

```
LM1<-
lm(Satisfaction~AirlineStatus+Age+Gender+PriceSensitivity+YearofFirstFlight+NoofFlightspa
+XofFlightwithotherAirlines+TypeofTravel+NoofotherLoyaltyCards+ShoppingAmountatAirpor
t+EatingandDrinkingatAirport+Class+DayofMonth+ScheduledDepartureHour++Flightcancelled
+DepartureDelayinMinutes+ArrivalDelayinMinutes+Flighttimeinminutes+FlightDistance+Arriv
alDelaygreater5Mins,data=df) summary(LM1)
LM2<-
lm(Satisfaction~AirlineStatus+Age+Gender+ShoppingAmountatAirport+PriceSensitivity+Yearo
fFirstFlight+NoofFlightspa+TypeofTravel+Class+ScheduledDepartureHour+ArrivalDelaygreate
r5Mins,data=df) summary(LM2)
dfp<-predict(LM2,interval="prediction")
dfp<-merge(df$Satisfaction,dfp) draw(dfp)
LM3<-
lm(Satisfaction~AirlineStatus+Age+Gender+Age*Gender+ShoppingAmountatAirport+PriceSen
sitivity+YearofFirstFlight+NoofFlightspa+TypeofTravel+Class+ScheduledDepartureHour+Arri
v alDelaygreater5Mins,data=df) summary(LM3)
LM4<-
lm(Satisfaction~AirlineStatus+Age+Gender+Age*Gender+ShoppingAmountatAirport+PriceSen
sitivity+Gender*PriceSensitivity+YearofFirstFlight+NoofFlightspa+TypeofTravel+Class+Sched
uledDepartureHour+ArrivalDelaygreater5Mins,data=df) summary(LM4)
LM6<-
lm(Satisfaction~AirlineStatus+Age+Gender+Age*Gender+ShoppingAmountatAirport+PriceSen
sitivity+Gender*PriceSensitivity+YearofFirstFlight+Gender*NoofFlightspa+NoofFlightspa+Ty
p eofTravel+Class+ScheduledDepartureHour+ArrivalDelaygreater5Mins,data=df)
summary(LM6)
```

```
LM5<-lm(Satisfaction~NoofFlightspa,data=df) summary(LM5)

plot(Satisfaction~NoofFlightspa,xlab="NoofFlightspa",ylab="Satisfaction",data=df)
abline(LM5)

#NoofFlightspa=0.05671105

LM5<-lm(Satisfaction~TypeofTravel,data=df) summary(LM5)

plot(Satisfaction~TypeofTravel,xlab="TypeofTravel",ylab="Satisfaction",data=df)

#TypeofTravel=0.3350338

LM5<-lm(Satisfaction~ShoppingAmountatAirport,data=df) summary(LM5)

plot(Satisfaction~ShoppingAmountatAirport,xlab="ShoppingAmountatAirport",ylab="Satisfacti
on",data=df)


#ShoppingAmountatAirport=0.0002999279  LM5<-

lm(Satisfaction~Class,data=df) summary(LM5)

plot(Satisfaction~Class,xlab="Class",ylab="Satisfaction",data=df)


#Class=0.002526544

LM5<-lm(Satisfaction~ScheduledDepartureHour,data=df) summary(LM5)

plot(Satisfaction~ScheduledDepartureHour,xlab="ScheduledDepartureHour",ylab="Satisfaction
",data=df)


#ScheduledDepartureHour=-6.981177e-06

LM5<-lm(Satisfaction~ArrivalDelaygreater5Mins,data=df)

plot(Satisfaction~ArrivalDelaygreater5Mins,xlab="ArrivalDelaygreater5Mins",ylab="Satisfactio
n",data=df)


summary(LM5)

#ArrivalDelaygreater5Mins=0.02528861

#Linear Model with Airline Status as predictor
```

LMAirlineStatus<-lm(Satisfaction~AirlineStatus,data=df) summary(LMAirlineStatus)

plot(Satisfaction~AirlineStatus,xlab="AirlineStatus",ylab="Satisfaction",data=df)

#AirlineStatus=0.1184333

#Linear Model with Age as predictor LMAge<-

lm(Satisfaction~Age,data=df) summary(LMAge)

plot(Satisfaction~Age,xlab="Age",ylab="Satisfaction",data=df) abline(LMAge)

#Age=0.0492023

#Linear Model with Gender as predictor LMGender<-

lm(Satisfaction~Gender,data=df) summary(LMGender)

plot(Satisfaction~Gender,xlab="Gender",ylab="Satisfaction",data=df) abline(LMGender)

#Gender=0.01760919

#Linear Model with Price Sensitivity as predictor

LMPriceSensitivity<-lm(Satisfaction~PriceSensitivity,data=df) summary(LMPriceSensitivity)

plot(Satisfaction~PriceSensitivity,xlab="PriceSensitivity",ylab="Satisfaction",data=df)

abline(LMPriceSensitivity)

#PriceSensitivity=0.007641272

#Linear Model with Airline Status Year of first flight as predictor LMFirstFlight<-

lm(Satisfaction~YearofFirstFlight,data=df) summary(LMFirstFlight)

plot(Satisfaction~YearofFirstFlight,xlab="YearofFirstFlight",ylab="Satisfaction",data=df) abline(LMFirstFlight)

#YearofFirstFlight=5.270168e-05

SVM Model: #svm df$happy<-df$Satisfaction df$happy[df$happy>=4]<-"happy"

df$happy[df$happy<4]<-"unhappy"

df1<data.frame(df$happy,df$AirlineStatus,dfAge,dfGender,df$PriceSensitivity,df$YearofFirstFlig

ht,df$NoofFlightspa,df$TypeofTravel,df$ShoppingAmountatAirport,df$Class,df$ScheduledDepartureHour,df$ArrivalDelaygreater5Mins) cutPoint2_3 <- floor(2 * dim(df1)[1]/3) randIndex <-

```
sample(1:dim(df1)[1]) trainData <- df1[randIndex[1:cutPoint2_3],] testData <-

df1[randIndex[(cutPoint2_3+1):dim(df1)[1]],] happy<-testData$df.happy table(happy)

# happy unhappy

# 1675    1518

1675/(1518+1675)

dim(testData) dim(trainData)

library(kernlab)


#try to lower error rate.

svmOutput <- ksvm(df.happy ~ df.TypeofTravel, data=trainData, kernel =
"rbfdot",kpar="automatic",C=5,cross=3, prob.model=TRUE)

svmOutput <- ksvm(df.happy ~ df.TypeofTravel+df.AirlineStatus, data=trainData, kernel =
"rbfdot",kpar="automatic",C=5,cross=3, prob.model=TRUE)

svmOutput <- ksvm(df.happy ~ df.TypeofTravel+df.AirlineStatus+df.Age, data=trainData,
kernel = "rbfdot",kpar="automatic",C=5,cross=3, prob.model=TRUE)

svmOutput <- ksvm(df.happy ~ df.TypeofTravel+df.AirlineStatus+df.Age+df.Gender,

data=trainData, kernel = "rbfdot",kpar="automatic",C=5,cross=3, prob.model=TRUE)

svmOutput <- ksvm(df.happy ~

df.TypeofTravel+df.AirlineStatus+df.Age+df.Gender+df.PriceSensitivity, data=trainData, kernel
= "rbfdot",kpar="automatic",C=5,cross=3, prob.model=TRUE) str(trainData)

svmOutput <- ksvm(df.happy ~ ., data=trainData, kernel =
"rbfdot",kpar="automatic",C=250,cross=3, prob.model=TRUE)


svmPred <- predict(svmOutput, testData, type = "votes") compTable

<- data.frame(testData$df.happy,svmPred[2,]) table(compTable)

table(compTable,testData$df.happy)


ctable <- as.table(matrix(c(1388,291,474,1040), nrow = 2, byrow = TRUE))

fourfoldplot(ctable, color = c("#CC6666", "#99CC99"),conf.level = 0, margin = 1, main =
"Confusion Matrix")
```

Visualization:

```
AgeSat<-aggregate(df2[, 1], list(df2$Age), mean)

AgeSat<-data.frame(AgeSat)

#CompOverallSat

colnames(AgeSat) <- c("Age", "AverageCustRating")

AgeSat<-merge(x = AgeSat, y = countvar, by = "Age", all = TRUE)

#AgeSat

plot2<-ggplot(AgeSat, aes(x=Age, y=AverageCustRating, label=CountOfFlights)) +
geom_text(aes(label=CountOfFlights), vjust=-1.0) +
geom_bar(stat="identity",colour="white",fill="blue") +theme(axis.text.x = element_text(angle =
90, hjust = 1))+ ggtitle("Age wise average Customer Satisfaction") + theme(plot.title=
element_text(hjust=0.5))

#plot2<-ggplot(AgeSat, aes(x=Age, y=AverageCustRating)) + geom_bar(stat="identity") +
theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))

#plot2 #

Gender plot

countvar<-data.frame(table(df$Gender)) colnames(countvar)

<- c("Gender", "NoOfTravelers") countvar1<-aggregate(df[,

1], list(df2$Gender), mean) colnames(countvar1) <-

c("Gender", "AverageSatisfaction")

countvar<-merge(x = countvar, y = countvar1, by = "Gender", all = TRUE)

#countvar

plot3<-ggplot(countvar, aes(x=Gender, y=AverageSatisfaction)) +
geom_text(aes(label=NoOfTravelers), vjust=-1.0) +
geom_bar(stat="identity",colour="white",fill="lightseagreen") +theme(axis.text.x =
element_text(angle = 90, hjust = 1))+ ggtitle("Gender wise Customers") + theme(plot.title=
element_text(hjust=0.5))  # Airline Status
```

```
grouped_data <- aggregate(df, by=list(df$AirlineStatus, df$Satisfaction), FUN=length);

grouped_data <-grouped_data[,c(1:3)]

#grouped_data

colnames(grouped_data) <- c("AirlineStatus", "Satisfaction","NoOfTravelers") grouped_data

plot4<-ggplot(grouped_data, aes(factor(Satisfaction), NoOfTravelers, fill = AirlineStatus)) +
geom_bar(stat = "identity", width = 0.2, position = "dodge") +   labs(list(x = "Satisfaction", y =
"Number of Travellers",fill = "group"))

#Type of travel

TypeTravel<-aggregate(df[, 1], list(df2$TypeofTravel), mean)

colnames(TypeTravel) <- c("TypeOfTravel", "AverageSatisfaction")

TypeTravel1<-data.frame(table(df$TypeofTravel)) colnames(TypeTravel1)

<- c("TypeOfTravel", "NoOfCustomers")

TypeTrav<-merge(x = TypeTravel, y = TypeTravel1, by = "TypeOfTravel", all = TRUE)

TypeTrav

plot5<-ggplot(TypeTrav, aes(x=TypeOfTravel, y=AverageSatisfaction)) +
geom_text(aes(label=NoOfCustomers), vjust=-1.0) +
geom_bar(stat="identity",colour="white",fill="red") +theme(axis.text.x = element_text(angle =
90, hjust = 1))+ ggtitle("Customer Satisfaction based on Type of Travel") + theme(plot.title=
element_text(hjust=0.5))



str(df)

a <- sub("Blue","1",df$AirlineStatus) b

<- sub("Silver","2",a) c<-

sub("Gold","3",b) d<-

sub("Platinum","4",c)

df$AirlineStatus<-d

df$AirlineStatus<-as.numeric(df$AirlineStatus)

df$AirlineStatus<-jitter(df$AirlineStatus)

df$Satisfaction<-jitter(df$Satisfaction) library(ggplot2)

g<-ggplot(df,aes(x=AirlineStatus,y=Satisfaction))+geom_point() g
```

```
a <- sub("Personal Travel","1",df$TypeofTravel) b
<- sub("Mileage tickets","2",a) c<-sub("Business
travel","3",b) df$TypeofTravel<-c
df$TypeofTravel<-as.numeric(df$TypeofTravel)
df$TypeofTravel<-jitter(df$TypeofTravel)
g1<-ggplot(df,aes(x=TypeofTravel,y=Satisfaction))+geom_point() g1
#1 Blue, 2 Gold,3 Platinum, 4 silver df$TypeofTravel1<-as.integer(df$TypeofTravel)
df$TypeofTravel1<-jitter(df$TypeofTravel1)
g1<-ggplot(df,aes(x=TypeofTravel1,y=Satisfaction1))+geom_point() g1



#graph about satsifaction based on state.
df<-df[df$AirlineCode=="US",] df1<-df
x1<-gsub('.*\\,', '', df1$OrginCity) df1$OrginCity<-x1
#CustPerCity<-data.frame(table(df1$OrginCity)) SatState<-aggregate(df1[,
1], list(df1$OrginCity), mean) colnames(SatState)<-c("state","sat")
g<-
ggplot(SatState,aes(x=row.names(SatState),y=sat))+geom_bar(stat="identity")+theme(plot.title=
element_text(hjust=0.5)) g


df<-df[df$AirlineStatus=="Platinum",]
LM7<-
lm(Satisfaction~Age+Gender+PriceSensitivity+YearofFirstFlight+NoofFlightspa+XofFlightwit
h
otherAirlines+TypeofTravel+NoofotherLoyaltyCards+ShoppingAmountatAirport+EatingandDri
nkingatAirport+Class+DayofMonth+Flightdate+ScheduledDepartureHour++Flightcancelled+De
partureDelayinMinutes+ArrivalDelayinMinutes+Flighttimeinminutes+FlightDistance+ArrivalD
e laygreater5Mins,data=df) summary(LM7)
LM7<-lm(Satisfaction~Age+TypeofTravel,data=df) summary(LM7)


AgeGroups<-cut(df$Age, breaks=c(18, 24,30,36,42,48,54,60,66,72,78,84,90), right = FALSE)
```

```
#AgeGroups
AgeGroups<-gsub(',', ' to ', AgeGroups)
AgeGroups<-gsub('\\[', '', AgeGroups) AgeGroups<-gsub('\\)',
'', AgeGroups)

df$Age<-AgeGroups



table(df$Gender)
#Female   Male
#139    173
173/(173+139) #0.5544871795
table(df$TypeofTravel)
#Business travel Mileage tickets Personal Travel
#246              19              47
47/294
#0.1598639456     table(df$Class)
#Business      Eco Eco Plus
#     26      259       27
259/294
#0.880952381 table(noofflightspa)
#Average    High     Low
#   55     123     134
123/294 #
0.4183673469
library(dplyr)
count<-aggregate(df[, 9], list(df$AirlineName),count)
count<-data.frame(count) g<-ggplot(df) df<-
as(df,"transaction") barplot(df$TypeofTravel)
```

```
#PLATINUM ANALYSIS

plat1<-df

plat1<-plat1[plat1$AirlineStatus == "Platinum",] str(plat1)


vBuckets<-replicate(length(plat1$Satisfaction),"Median")

vBuckets[plat1$Satisfaction>3]<-"High" vBuckets[plat1$Satisfaction<3]<-"Low"


plat1$Satisfaction<-as.factor(vBuckets)


plat1Agg<-aggregate(plat1[, 11], list(plat1$Satisfaction), mean)#ShoppingAmountatAirport
plat1Agg

plat1Agg<-aggregate(plat1[, 12], list(plat1$Satisfaction), mean)#EatingandDrinkingatAirport

plat1Agg

plat1Agg<-aggregate(plat1[, 5], list(plat1$Satisfaction), mean) #PriceSensitivity plat1Agg



#summary(fulldf) fdf<-fulldf

#fulldf$AirlineName

# plotting var charts

CompOverallSat<-aggregate(fdf[, 1], list(fdf$AirlineName), mean)

CompOverallSat<-data.frame(CompOverallSat)

#CompOverallSat

colnames(CompOverallSat) <- c("Airline", "AverageCustRating")

library(ggplot2) CompOverallSat

plot1<-ggplot(CompOverallSat, aes(x=Airline, y=AverageCustRating)) +
geom_bar(stat="identity") + theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))

#CompOverallSat <- CompOverallSat[order(CompOverallSat$AverageCustRating),]
```

```
CompOverallSat

CompOverallSat$Airline <- factor(CompOverallSat$Airline, levels =
CompOverallSat$Airline[order(CompOverallSat$AverageCustRating)]) plot2<-
ggplot(CompOverallSat, aes(x=Airline, y=AverageCustRating)) +
geom_bar(stat="identity") + theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))

GenderData <- data.frame(table(df$Gender)) colnames(GenderData) <- c("Gender",

"NoOfTravelers")

GenderDistribution <- ggplot(GenderData, aes(x=Gender, y=NoOfTravelers)) +
geom_text(aes(label=NoOfTravelers), vjust=-1.0) +
geom_bar(stat="identity",colour="white",fill="blue") +theme(axis.text.x = element_text(angle =
90, hjust = 1))+ ggtitle("Gender wise travelers - Southeast Airlines")

newCol<-colnames(df) newCol<-gsub("\\.", "", newCol)

colnames(df)<-newCol

a <- sub("No","0",df$Flightcancelled) b

<- sub("Yes","1",a) df$Flightcancelled

<- b

df$Flightcancelled<-as.numeric(df$Flightcancelled)

df$Satisfaction<- as.numeric(as.character(df$Satisfaction)) us

<- map_data("state")

dfSoutheast<-df[df$AirlineCode=="US",]

str(dfSoutheast) install.packages("mice")

library(VIM)

dfSoutheast<-dfSoutheast[complete.cases(dfSoutheast),] dffemale<-

dfSoutheast[dfSoutheast$Gender=="Female",]

dffemale$ShoppingAmountatAirport[dffemale$ShoppingAmountatAirport>0]<-"shopping"

dffemale$ShoppingAmountatAirport[dffemale$ShoppingAmountatAirport==0]<-"not shopping"

delay<-aggregate(dffemale[, 1], list(dffemale$ShoppingAmountatAirport), mean) delay<-

aggregate(dfSoutheast[, 23], list(dfSoutheast$OriginState), mean) colnames(delay)<-

c("X","Avaragesatisfaction") plot<-ggplot(delay, aes(x=X, y=Avaragesatisfaction)) +

geom_bar(stat="identity",colour="white",fill="blue") +

theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))+ggtitle("Average Satisfaction In
```

```
Man")

plot


delay<-aggregate(dfSoutheast[, "ArrivalDelayinMinutes"], list(dfSoutheast$OriginState), mean)
mergedf<-merge(dfSoutheast$OriginState)

GenderData <- data.frame(table(dfSoutheast$OriginState))

colnames(GenderData) <- c("stateName", "NoOfTravelers") str(GenderData)

View(GenderData)

GenderData$State<-tolower(GenderData$State)

GenderData <- GenderData[-39,]

GenderData <- GenderData[-45,]

GenderData$state.abb<-statemap$state.abb

GenderData$stateName<-tolower(GenderData$stateName)

statemap <- data.frame(state.abb) statemap$Lon <-

state.center$x statemap$Lat <- state.center$y

GenderData$lon<-statemap$Lon GenderData$lat<-statemap$Lat

map<-ggplot(GenderData,aes(map_id=stateName))+geom_map(map=us,
aes(fill=NoOfTravelers))+expand_limits(x = us$long, y = us$lat)+coord_map() + ggtitle("No.Of
Travelers per state")+geom_text(aes(x=GenderData$lon, y=GenderData$lat,
label=GenderData$state.abb), size=2)+scale_fill_gradient(low = "white", high = "blue", guide =
"colorbar") map

statemap <- data.frame(state.abb)

statemap$Lon <- state.center$x

statemap$Lat <- state.center$y str(statemap)

GenderData$StateAbb <- state.abb[match(GenderData$State, state.name)]

View(GenderData)


GenderData$Lon <- state.center$x

GenderData$Lat <- state.center$y

View(GenderData)
```

```
str(raw_data)


CleanSatisfaction<-raw_data[(raw_data$Satisfaction=="1" |
raw_data$Satisfaction=="1.5" |          raw_data$Satisfaction=="2"
|          raw_data$Satisfaction=="2.5" |
raw_data$Satisfaction=="3" |          raw_data$Satisfaction=="3.5"
|          raw_data$Satisfaction=="4" |
raw_data$Satisfaction=="4.5" |          raw_data$Satisfaction=="5"
),]
#clean_data <- subset(raw_data,trimws(raw_data$Satisfaction)==c(1:5))
#clean_data df<-CleanSatisfaction
newCol<-colnames(CleanSatisfaction)
newCol<-gsub("\\.", "", newCol)
newCol colnames(df)<-newCol



a <- sub("No","0",df$Flightcancelled) b
<- sub("Yes","1",a) df$Flightcancelled
<- b
df$Flightcancelled<-as.numeric(df$Flightcancelled)

m <- mode(df$Age) m

}
```

```
vBuckets<-replicate(length(df$Satisfaction),"Median")
vBuckets[df$Satisfaction>3]<-"High" vBuckets[df$Satisfaction<3]<-"Low"

Satisfaction<-as.factor(vBuckets)

age<-createBuckets(df$Age)
pricesensitive<-createBuckets(df$PriceSensitivity) yearoffirstflight<-
createBuckets(df$YearofFirstFlight) noofflightspa<-createBuckets(df$NoofFlightspa)
shoppingamount<-createBuckets(df$ShoppingAmountatAirport) scheduleddeparturehour<-
createBuckets(df$ScheduledDepartureHour) hotelSurveyruleDFSE<-as(ruleDF,"transactions")

rulesetsoutheastH<- apriori(hotelSurveyruleDFSE, parameter=list(support=0.05,
confidence=0.8),appearance = list(default="lhs", rhs=("Satisfaction=High")))

goodrulesH<-sort(rulesetsoutheastH,by="lift")[1:5] inspect(goodrulesH)
plot1<-plot(goodrulesH, method = "graph", engine = "htmlwidget")  rulesetsoutheastL<-
apriori(hotelSurveyruleDFSE, parameter=list(support=0.05,
confidence=0.5),appearance = list(default="lhs", rhs=("Satisfaction=Low")))
goodrules2<-sort(rulesetsoutheastL,by="lift")[1:10]
plot2<-plot(goodrules2, method = "graph", engine = "htmlwidget")

.
svmOutput <- ksvm(df.happy ~ df.TypeofTravel, data=trainData, kernel =
"rbfdot",kpar="automatic",C=5,cross=3, prob.model=TRUE)
svmOutput <- ksvm(df.happy ~ df.TypeofTravel+df.AirlineStatus, data=trainData, kernel =
"rbfdot",kpar="automatic",C=5,cross=3, prob.model=TRUE)
svmOutput <- ksvm(df.happy ~ df.TypeofTravel+df.AirlineStatus+df.Age, data=trainData,
kernel = "rbfdot",kpar="automatic",C=5,cross=3, prob.model=TRUE)
svmOutput <- ksvm(df.happy ~ df.TypeofTravel+df.AirlineStatus+df.Age+df.Gender,
data=trainData, kernel = "rbfdot",kpar="automatic",C=5,cross=3, prob.model=TRUE)
svmOutput <- ksvm(df.happy ~
```

```
df.TypeofTravel+df.AirlineStatus+df.Age+df.Gender+df.PriceSensitivity, data=trainData, kernel
= "rbfdot",kpar="automatic",C=5,cross=3, prob.model=TRUE)

svmOutput <- ksvm(df.happy ~ ., data=trainData, kernel =

"rbfdot",kpar="automatic",C=5,cross=3, prob.model=TRUE)

svmOutput <- ksvm(df.happy ~ ., data=trainData, kernel =

"rbfdot",kpar="automatic",C=1,cross=3, prob.model=TRUE)

svmPred <- predict(svmOutput, testData, type = "votes") compTable

<- data.frame(testData$df.happy,svmPred[2,]) table(compTable)

plot3<-ggplot(countvar, aes(x=Gender, y=AverageSatisfaction)) +
geom_text(aes(label=NoOfTravelers), vjust=-1.0) +
geom_bar(stat="identity",colour="white",fill="lightseagreen") +theme(axis.text.x =
element_text(angle = 90, hjust = 1))+ ggtitle("Gender wise Customers") + theme(plot.title=
element_text(hjust=0.5))


# Airline Status   grouped_data <- aggregate(df, by=list(df$AirlineStatus,

df$Satisfaction), FUN=length); grouped_data <-grouped_data[,c(1:3)]

#grouped_data

colnames(grouped_data) <- c("AirlineStatus", "Satisfaction","NoOfTravelers") grouped_data

plot4<-ggplot(grouped_data, aes(factor(Satisfaction), NoOfTravelers, fill = AirlineStatus)) +
geom_bar(stat = "identity", width = 0.2, position = "dodge") +   labs(list(x = "Satisfaction", y =
"Number of Travellers",fill = "group"))

#Type of travel

TypeTravel<-aggregate(df[, 1], list(df2$TypeofTravel), mean)

colnames(TypeTravel) <- c("TypeOfTravel", "AverageSatisfaction")

TypeTravel1<-data.frame(table(df$TypeofTravel)) colnames(TypeTravel1)

<- c("TypeOfTravel", "NoOfCustomers")

TypeTrav<-merge(x = TypeTravel, y = TypeTravel1, by = "TypeOfTravel", all = TRUE)
TypeTrav  plot5<-ggplot(TypeTrav, aes(x=TypeOfTravel, y=AverageSatisfaction)) +
geom_text(aes(label=NoOfCustomers), vjust=-1.0) +
geom_bar(stat="identity",colour="white",fill="red") +theme(axis.text.x = element_text(angle =
90, hjust = 1))+ ggtitle("Customer Satisfaction based on Type of Travel") + theme(plot.title=
element_text(hjust=0.5))
```