

**Introduction**

This report presents the findings of a customer survey conducted by Southeast Airlines. After a comprehensive understanding of the data, we aim to analyze the data to provide some actionable recommendations to Southeast Airlines to lower their customer churn. After data cleansing, our team plan to apply three data modeling techniques covered in this course to identify the attributes which influence the customer’s likelihood to recommend. So, we can focus on these attributes to improve overall customer satisfaction. The recommendations will base on our modeling result.

**Business Question:**

1. What is the meaning of these variables in the survey?
2. Do the existing variables sufficient? Do any new variables need to be added in the future survey?
3. What are the key variables that influence customer satisfaction most?
4. How these variables influence customer satisfaction, negatively or positively?
5. What are some possible actions Southeast airlines can take to improve?

**Data Cleansing**

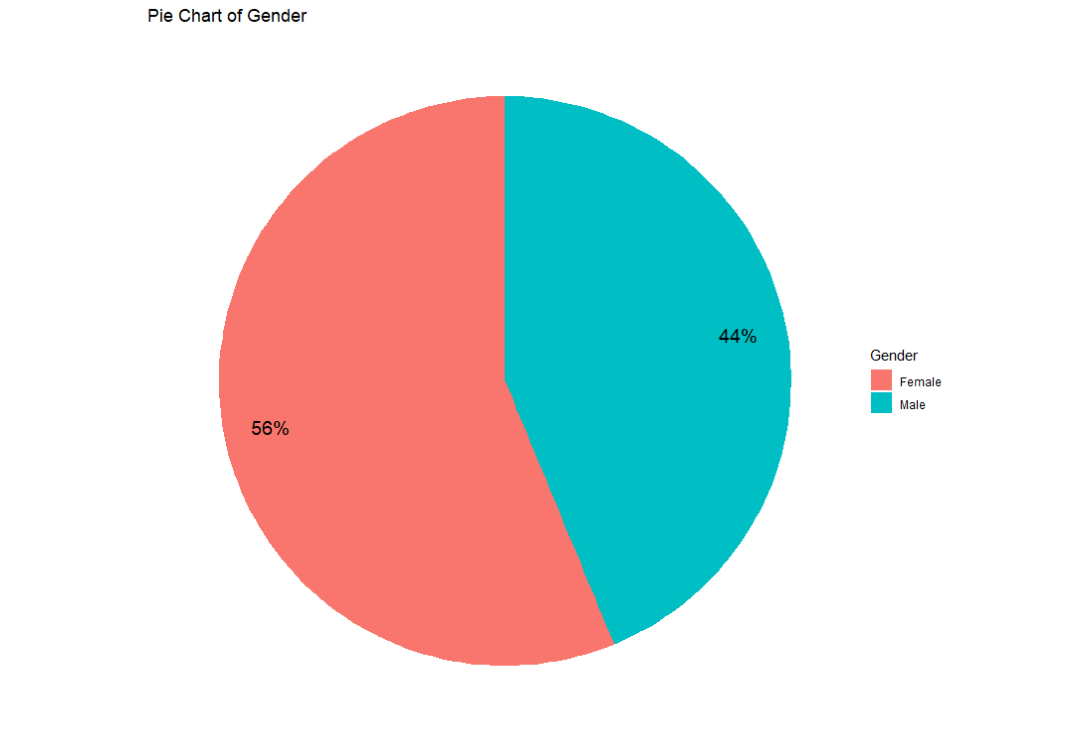
The source data has 5000 observations, 96 of them contains NA values. 92 of 96 is because of flight cancelled, the rest 4 is due to data missing. We decided to abandon these observations since it only occupies 1.92% of the entire data set. It also means our analysis will not include the case that the flight is cancelled.

We keep the remaining 4904 observations as our original data to analyze. And we will not do any further modification for now to remain intact. Any further modification of data frame will be stated in each individual data modeling.

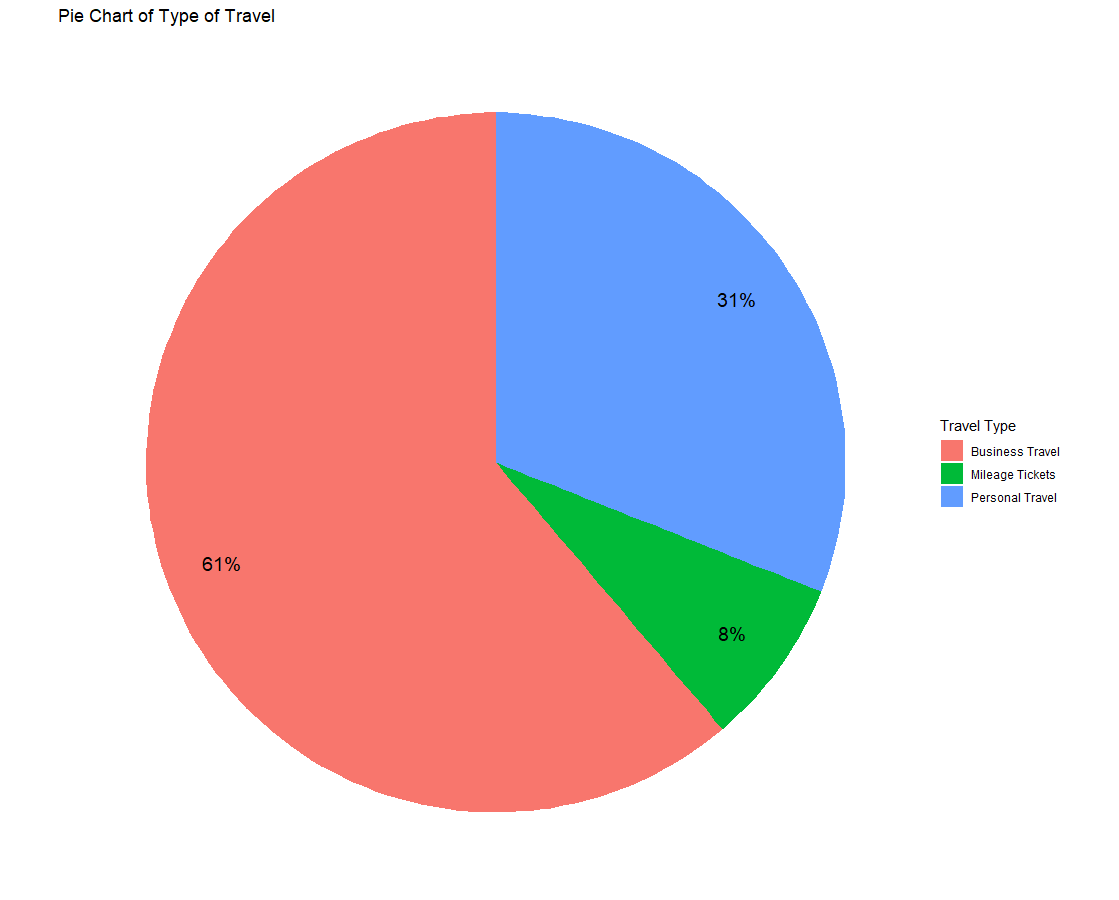
**Data Virtualization**

**Pie Chart:**

From the Pie chart of gender, we know that 56% of the Southeast Airline customers are female and 44% of the Southeast Airline customers are male

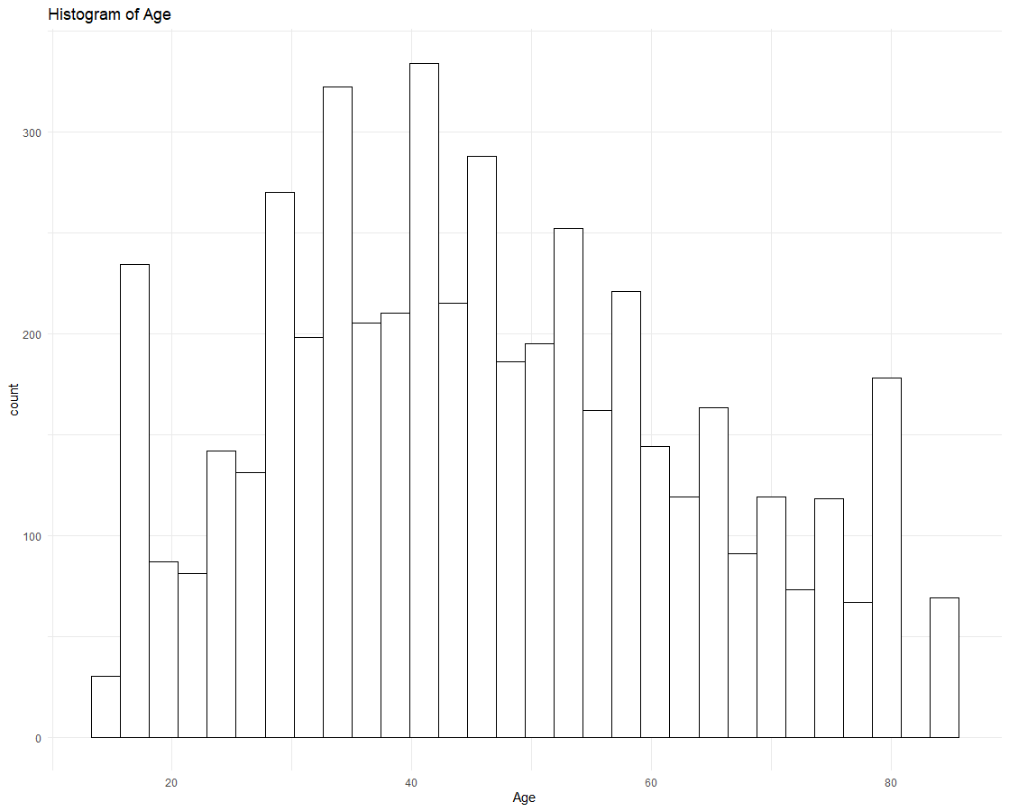


From Pie Chart of Type of Travel, we can tell that 61% of Southeast Airline customers are Business Travelers, 31% of Southeast Airline customers are Personal Travelers, and 8% of Southeast Airline customers are Mileage Tickets travelers.

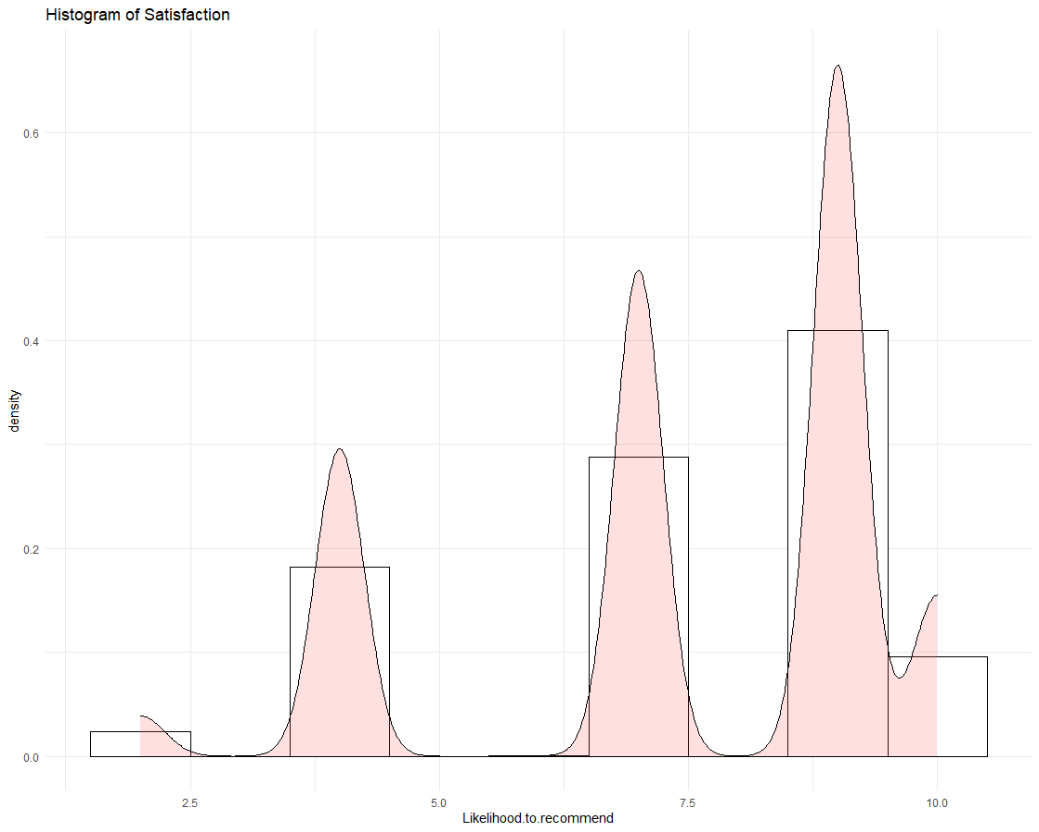


**Histogram:**

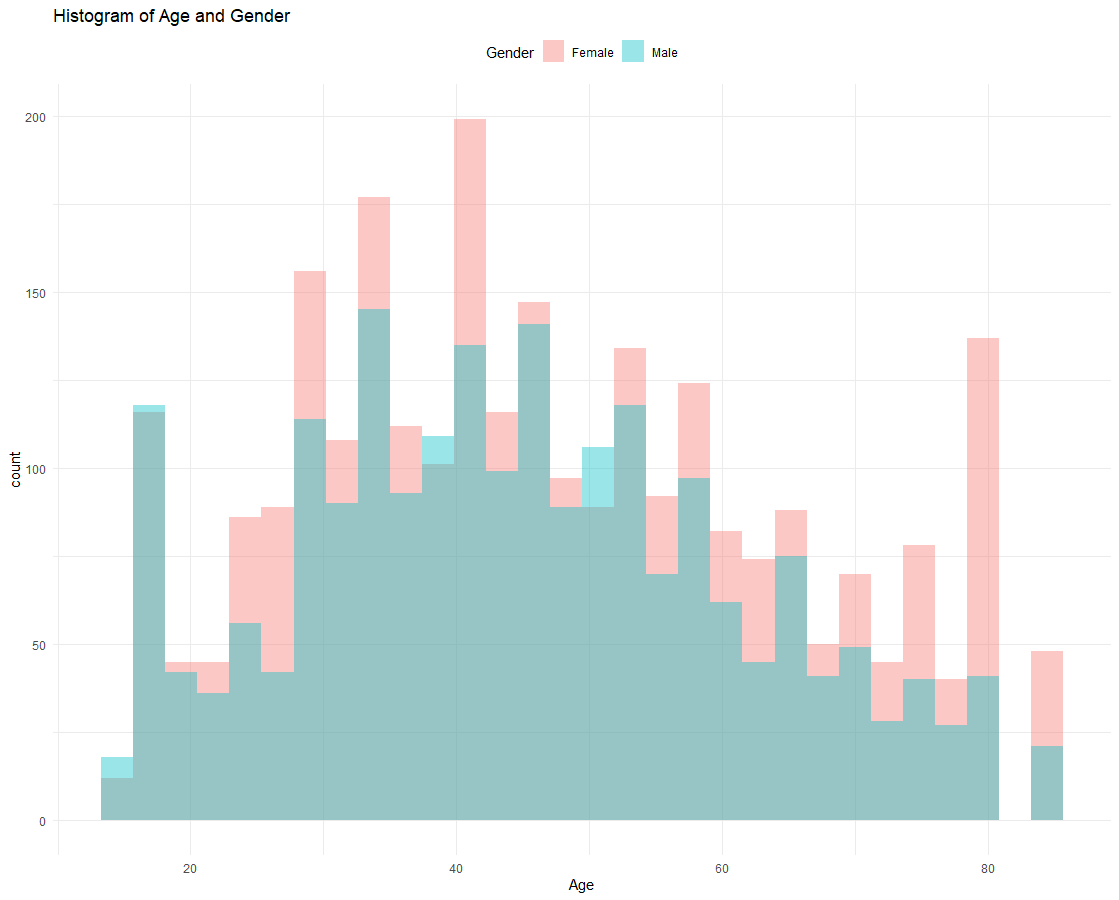
The histogram of Age looks like a Normal distribution graph. Kids and Old people are not likely to travel a lot.



From Histogram of Satisfaction, we know that people’s satisfaction is polarized, and most people are satisfied.

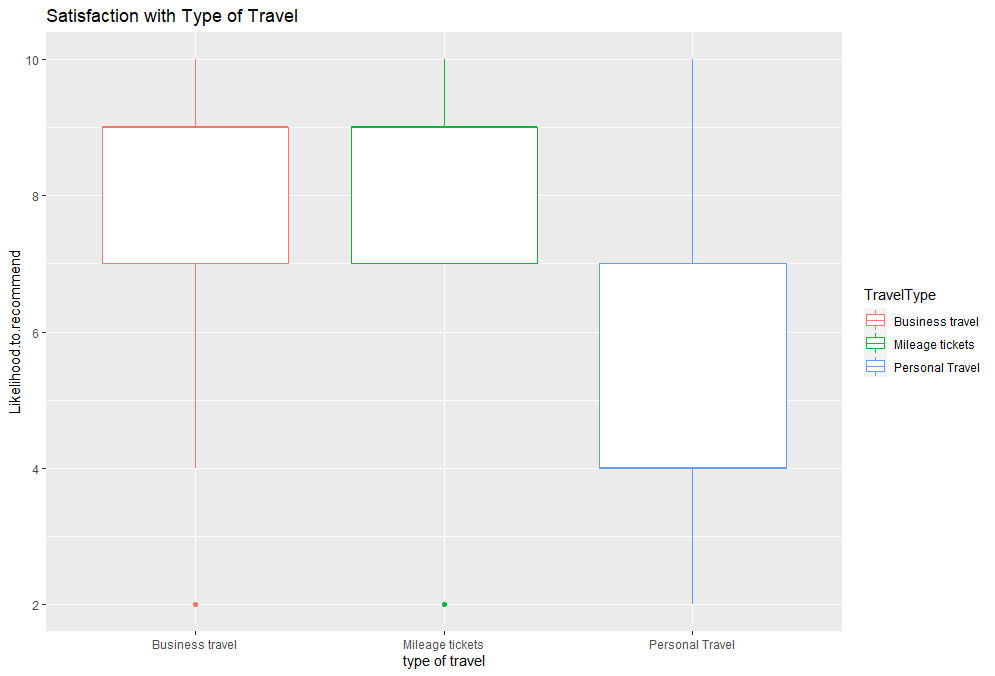


From Histogram of Age and Gender, we know that female customers are slightly larger than male customers from all different ages.

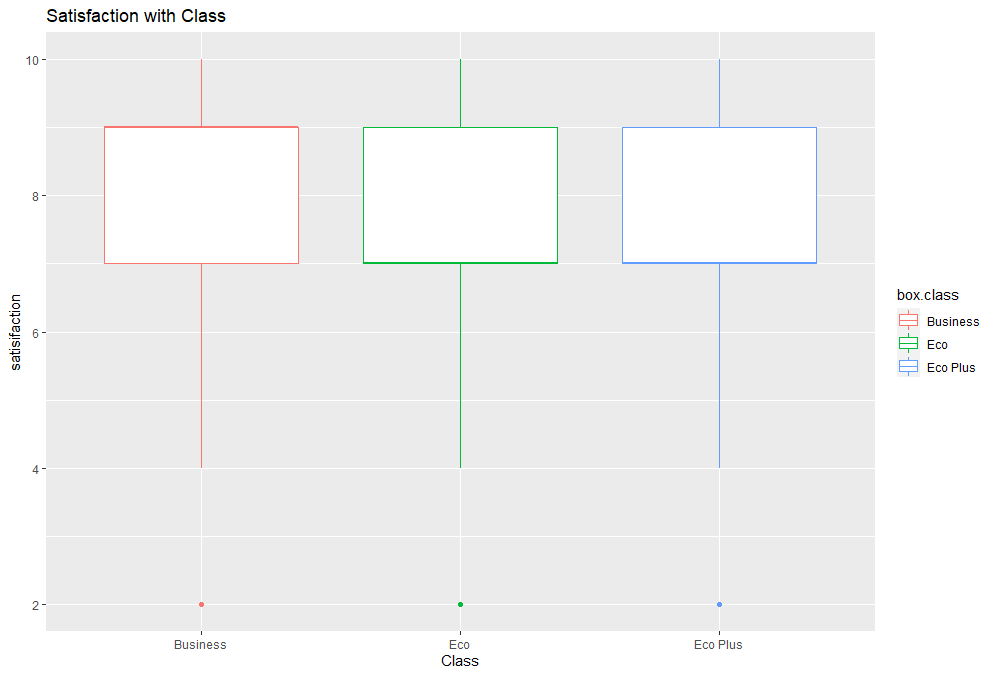


**Boxplot:**

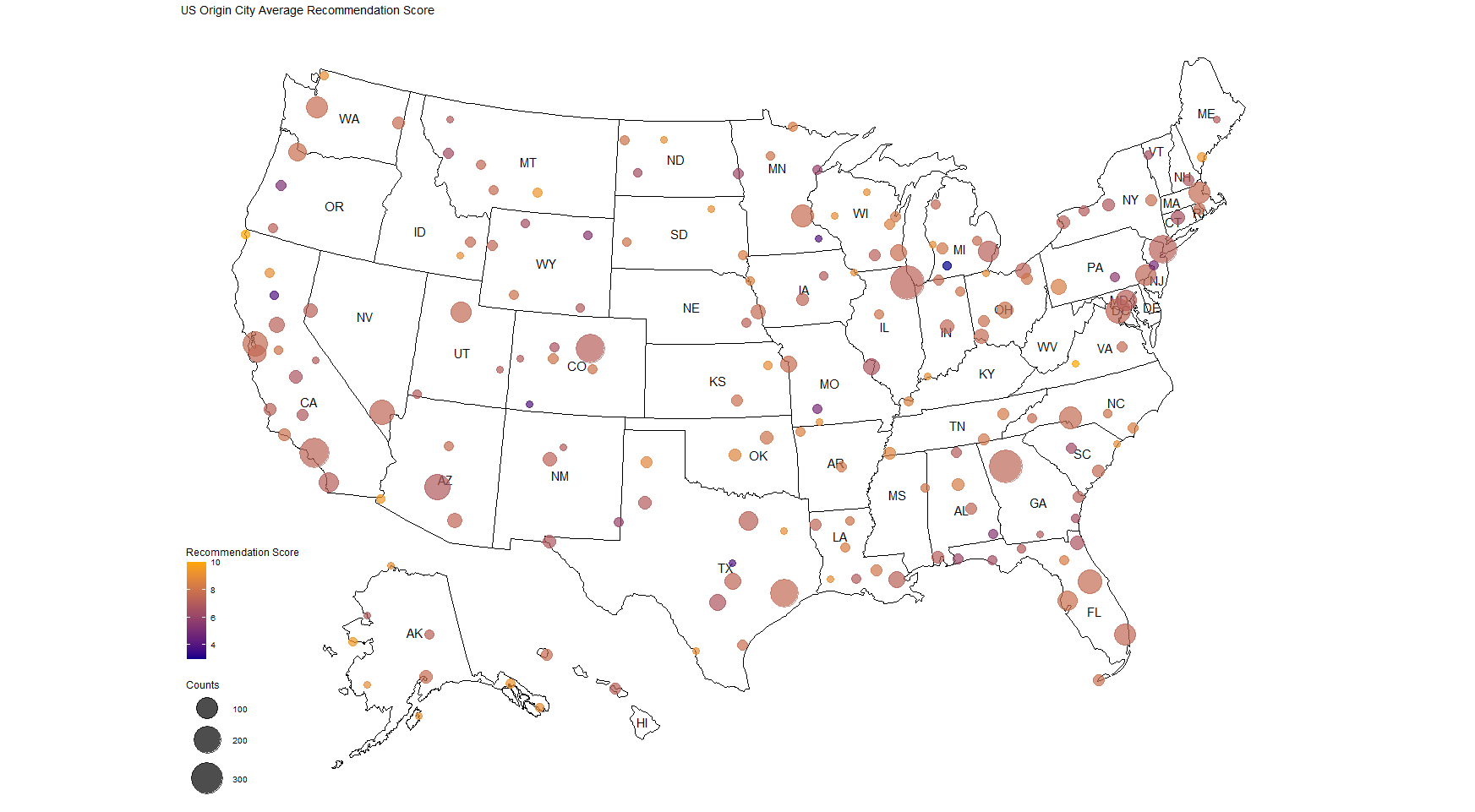
From Satisfaction with Type of Travel, we see that Business Travelers and Mileage Tickets traveler are more satisfied than Personal Travelers.



Satisfaction with Class shows that the medium of Business Class’s satisfaction is higher than Eco and Eco Plus.

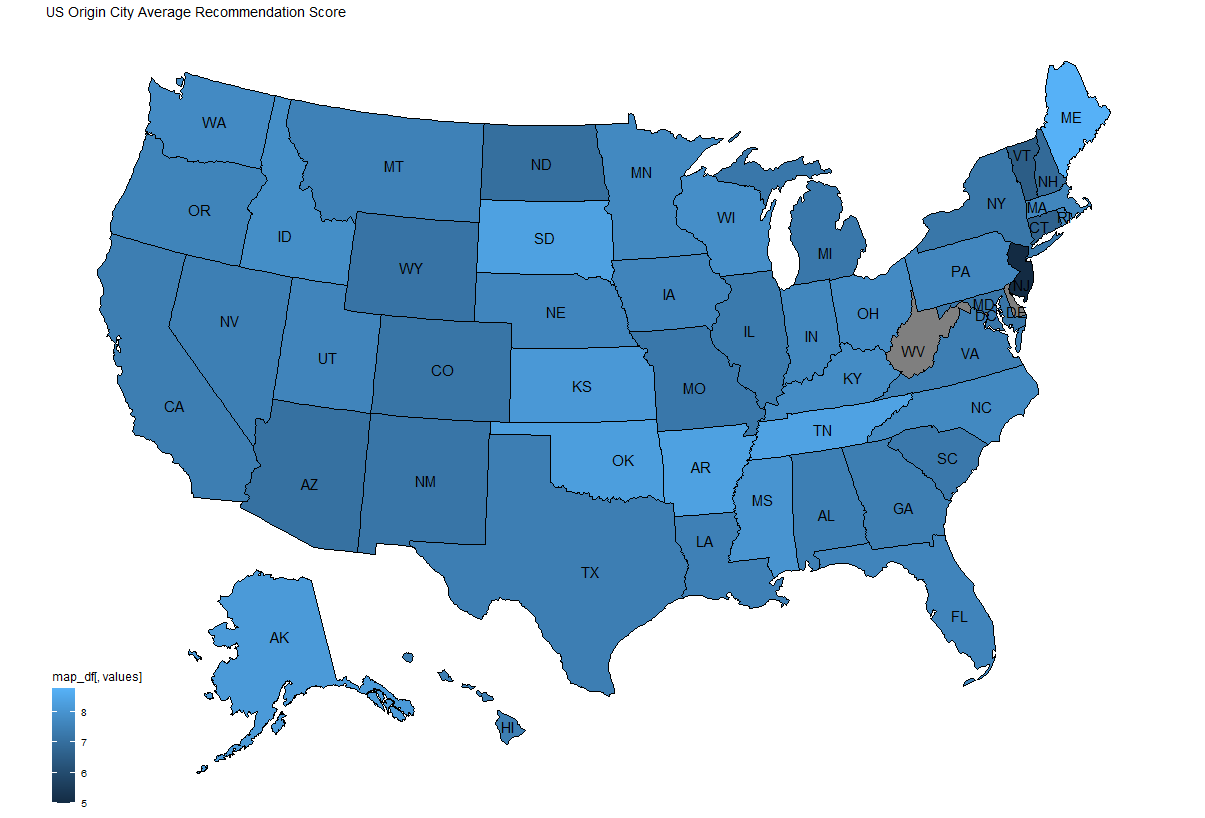


**Maps:**

From the following map, we can see that yellow-colored dots are more than blue-colored dots, which means that normally people are satisfied. Also, dots are bigger on west and east coast, which means that flights depart more on these areas (especially in South California, Florida, and NYC). During data cleaning we realized that there is thirty-seven data point that’s outside of the U.S territory, at olong above -68. Which is Puerto Rico.

This map shows the average recommendation score in color and frequency count in size.

From the following map, we can see that some states’ travelers are usually more satisfied, like SD, OK, AR, and TN



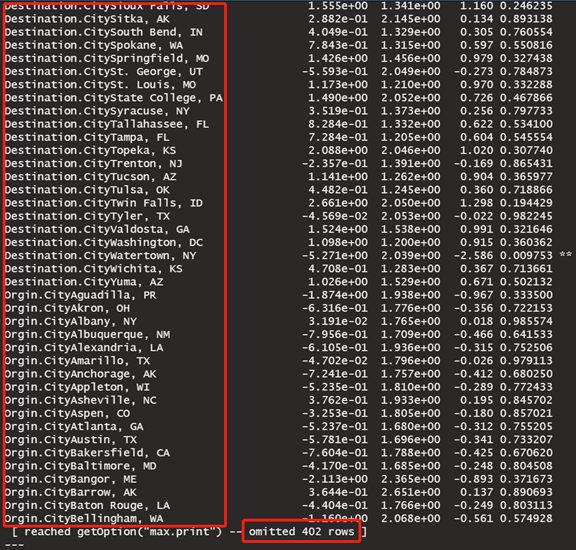
**Data Modeling**

**Linear Regression**

After cleaning the original data, the data frame contains only flights that were not be canceled. So, the column "Flight. Cancelled" will be removed and not considered a useful predictor. By applying the linear regression technique, we aim to observe which factors influence customers' score willingness to recommend Southeast Airlines to others. Hence, the model's dependent value will be "Likelihood to recommend," and independent values or predictors will be the rest of the columns.

**Model 1:**

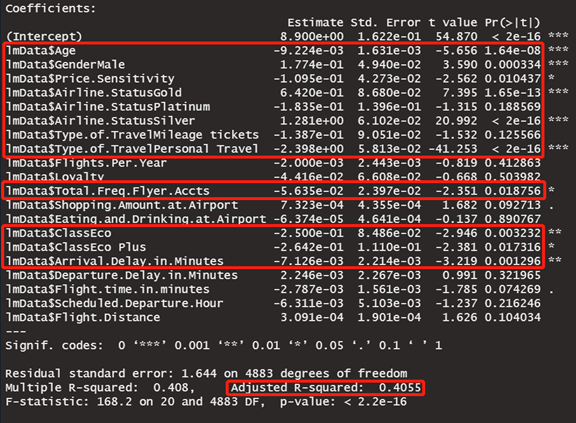
Model 1 is the basic model. We generally put all the rest of the predictors into the model. The model’s result is a mess. The model contains too many dummy variables, especially origin cities and destination cities make the model too complicated. The Adjusted R-squared of model 2 is 0.4069.





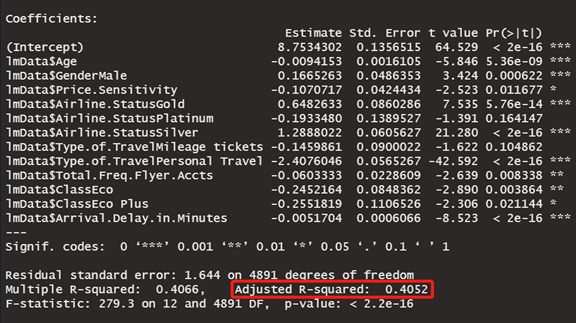
**Model 2:**

We filtered multiple variables to make the model simpler. Because we are not interested in analyzing each flight route, any variables involving origin and destination will be abandoned. Furthermore, any variables that relate to flight date and year will also be abandoned. There are 16 variables left in the data frame. The Adjusted R-squared of model 2 is 0.4055. The variables labeled in the result have p-Values less than 0.05.



**Model 3:**

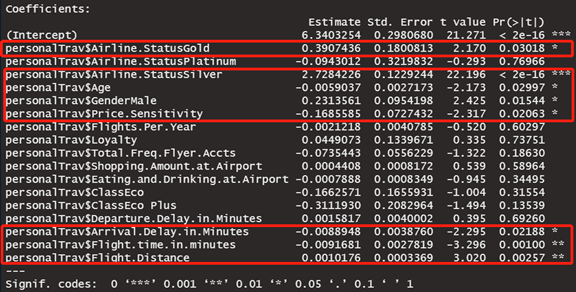
Model 3 is optimized based on p-Values. We wipe off all the variables having p-Values larger than 0.05, which means they are not statistically significant. The remaining variables are: Age, Gender, Price Sensitivity, Airline Status, Type of Travel, Total Frequent Flyer Accounts, Class, Arrival Delay in Minutes. The Adjusted R-squared is 0.4052.

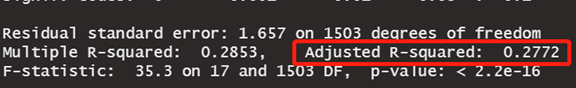


The coefficients in the model indicate the relationship between variables and satisfaction. The intercept is approximately 8.75 means if the customer at the age of 0, gender is female, price sensitivity is 0, airline status is blue, type of travel is business travel, total frequent flyer account is 0, class is business, arrival delay in minutes is 0, her satisfaction will be 8.75. If gender is male, others remain the same, and his satisfaction will increase by approximately 0.167. Every price sensitivity increases by 1; his/her satisfaction will decrease by approximately 0.11. If the airline statues are gold or silver, the satisfaction will increase by approximately 1.289 or 0.648. However, satisfaction will decrease by approximately 0.19 if the airline status is platinum. If the type of travel is mileage tickets, the satisfaction will decrease by approximately 0.15. The satisfaction will drop significantly by approximately 2.41 if the type of travel type is personal travel. Each one more frequent flyer accounts the customer has, his/her satisfaction will decrease by approximately 0.06. If the class is Eco or Eco plus, the customer’s satisfaction will also drop by about 0.25. Finally, everyone minute of arrival delay will lead to a satisfaction decrease by 0.005.

**Model 4:**

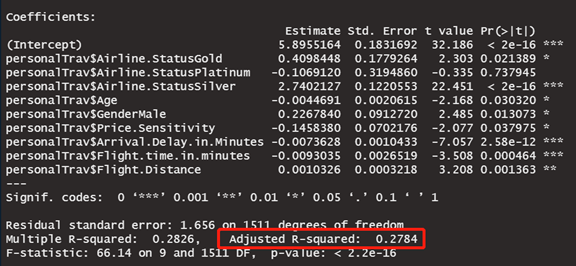
Since personal travel customers drop the likelihood to recommend significantly, we will only focus on personal travel customers to find out if there is some possible way to fix it.





The Adjusted R-squared of model 4 is 0.2772. Like model 2, we will keep the variables labeled in the result with p-Values less than 0.05 in the following model 5.

Model 5:



As we can see from the model 5 result, this model's variables are not exactly the same as that of model 3, meaning the variables influencing most of the overall customer and personal travel customer are not the same. The base value of likelihood to recommend in this customer segment is only 5.896, under the condition that the customer is 0 years old female with blue airline status, 0 price sensitivity, 0 arrival delay, 0 flight time, and 0 flight distance. The critical point here is if the customer is a silver airline status holder, the estimated likelihood to recommend increases by 2.74. The customer turns from a detractor to a promoter immediately.

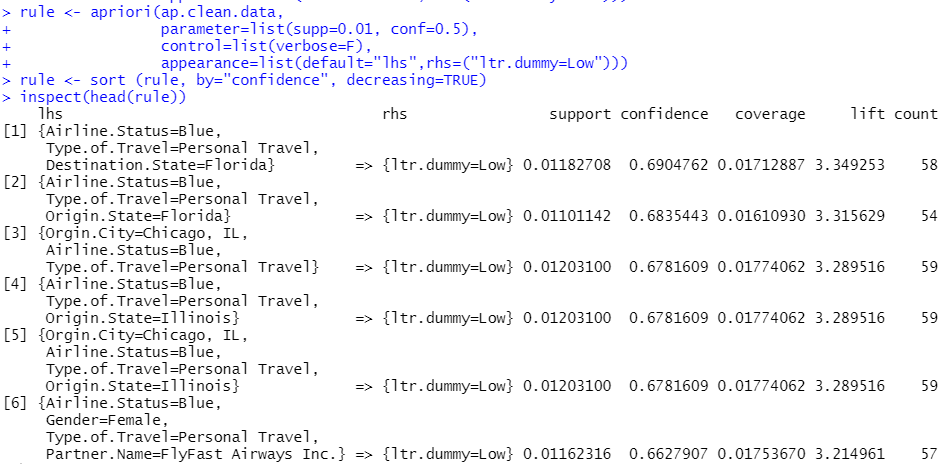
In summary, we will apply other data modeling techniques to analyze the data more. According to what we get from linear regression modeling, we will keep an eye on the key predictors' likelihood to recommend most, especially those with negative correlation.

Predictive Modeling

**Model 1:**

The following model uses different support value to predict the outcomes.

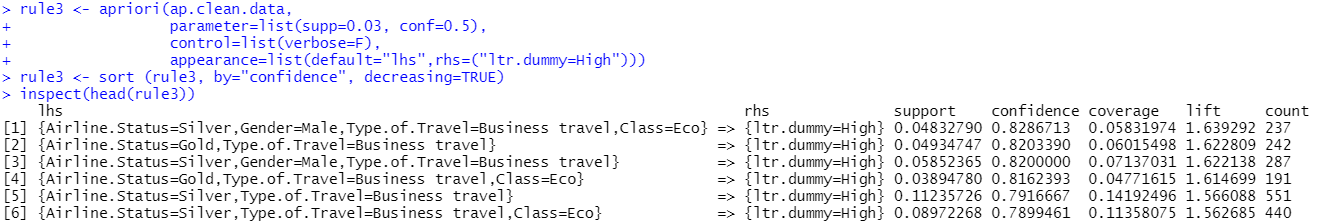
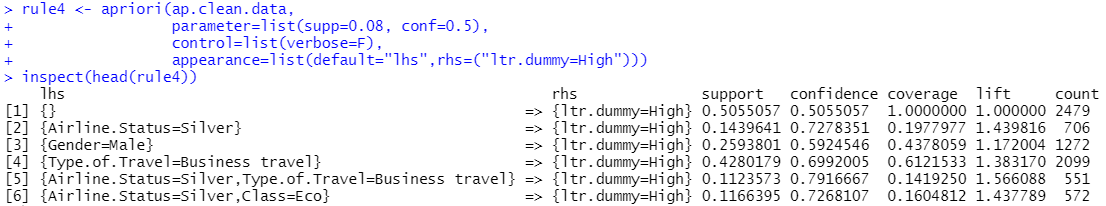
The outcome shows that Personal Travelers are not usually satisfied, as well as Blue status Airlines.



**Model 2:**

The below two apriori test is looking on the item that will predict a high recommendation score. We picked 0.03 support for the first model and 0.08 for the second model, and we picked confidence over 0.5 to be significant as it means over half of the confidence.

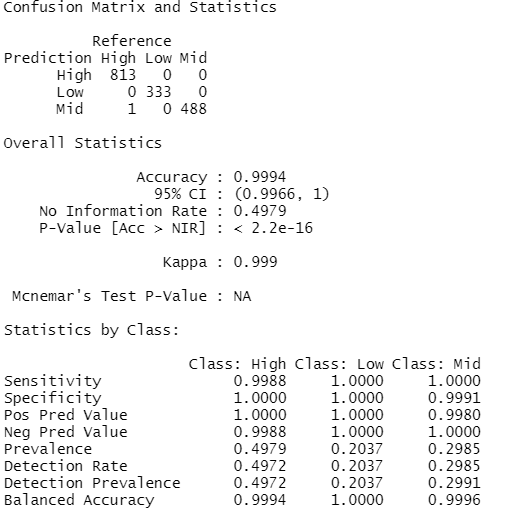
In the first test we sorted the result in confidence order, and the highest outcome for both test comes to have silver status, male gender, and business travel.



**Model 3:**

Our survey data don’t have a continuous range of factor, we either need to drop the factor level that we are missing value in, or we need to use some other way of data cleaning here. In this section we decided to use summarization method from dplyr to mutate a new column that’s three factors with low mid and high.

In this model we can see it only missed one value and the accuracy rate is 99.94 percent.



Suggestions for Southeastern Airline:

According to the result, there are four facts:

* Blue Card Members’ satisfaction is much lower than Silver and Golden Card member.
* Female passengers’ satisfaction is lower than male passengers.
* Personal Traveling passengers usually has low satisfaction.
* Eco Plus class passengers’ evaluation is even lower than Eco class passengers.

To improve the service and the passengers' experience:

1. Improve the experience of Blue card passengers appropriately.
2. Improve the service for female passengers, or design special service for female passengers
   1. The lavatories in cabin would comply with a stricter cleaning regime.
   2. Prepare feminine hygiene products.
3. Provide a higher quality cabin service in Eco Plus class.

Except the facts those listed above, there are still some minor matters like elder passengers, the service for them is important as well which could increase the whole evaluation of passengers.