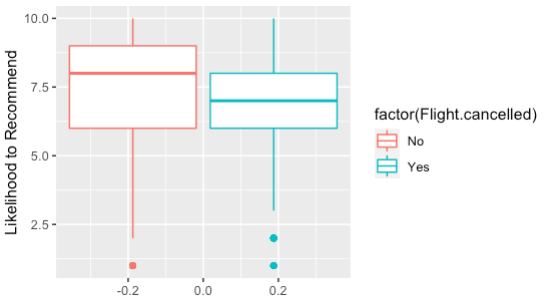
**Data Cleaning:**

Once we installed all the packages we needed for analysis and loaded in the data, the first step was to clean it. Using the jsonlite package, we read in the survey and converted it into a data frame. We thought that the month that a flight takes place might be important as seasonal effects might be present. We created a “month” column by first converting the column “Flight.date” to character, then we used cSplit from the splitstackshape package to separate the values in Flight.date by “/”. After that, we created a “status” column by using an if else statement where if “Likelihood.to.recommend” was greater than 8, the value of status would be “Promoter”, if it was less than 7 it would be “Detractor”, and anything else (7 or 8) would be “Neutral”. Next, we converted a number of columns to factor so that they would be easier to use in a linear model. We got rid of the “freeText” column and the Flight.date column because freeText contained mostly NA values and the specific date of the flight was not important now that we had a month column.

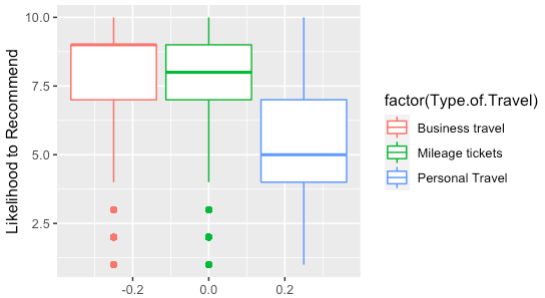
**Boxplots:**

Our first step in our analysis was creating boxplots of Likelihood.to.recommend and grouping it by different variables. Each of our boxplots are created the same way using ggplot2. For each graph, the aes has Likelihood.to.recommend as the y variable, and different variables for group. We set the colour to whatever variable was being grouped by so that we could have a visual distinction for each different category. Lastly, we changed the y label on each of the graphs to “Likelihood to Recommend”, as the column name isn’t visually appealing. Our first boxplot was for whether or not a flight was cancelled:



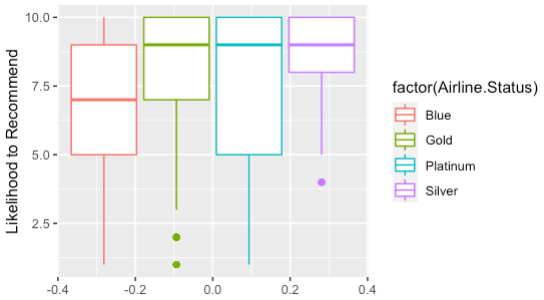
As we can see from this graph, cancelled flights have a lower distribution of Likelihood.to.recommend than flights that aren’t cancelled, as one would expect.

The next boxplot is grouped by “Type.of.Travel”:



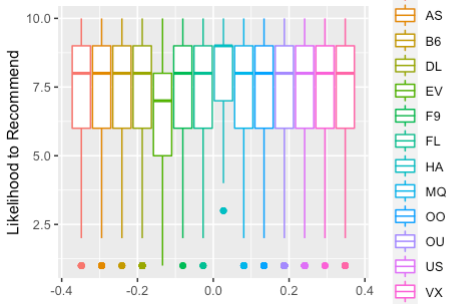
Overall, Business travel has the highest distribution, with a median value of 9, which would be promoters. Mileage tickets have a lower distribution, but still have a median value of 8. The distribution for Personal Travel is concerning, as the median and 3rd quartile are 5 and 7, respectively. This means that majority of passengers on Personal Travel are detractors.

Our third boxplot is grouped by “Airline.Status”:



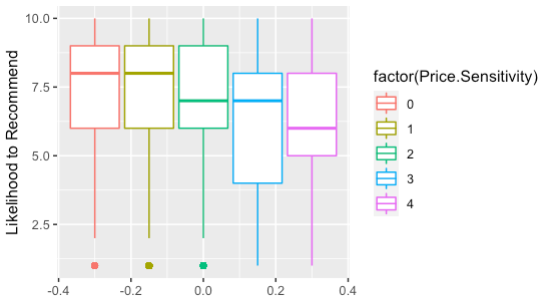
Out of the different Airline statuses, Silver has the most encouraging distribution with a median value of 9 and a 1st quartile value of 8. The Platinum status has a 1st quartile value of only 5, however, the median value is 9 and the mean is 7.462, which means that on average, Platinum members are at worst neutral. The Gold member distribution is closer to Silver with a median value of 9 as well. Lastly, although Blue members have a median value of 7, the mean is 6.846, which suggests that the average Blue member is a detractor. Over half of all passengers are Blue members, which is concerning for Southeast Airlines, however, we suggest improving the perks or benefits that come with the Blue status in hopes that these passengers would be more likely to recommend Southeast.

Next, we have a boxplot for “Partner.Code”:



All but two of the distributions look identical. The first distribution that deviates from the norm is that of “EV”, which is FlyFast Airways Inc. The median value for EV is 7, but the mean is only 6.665. Noticing this, we created table using aggregate where the parameter is Likelihood.to.recommend, grouped by Partner.Code, and using the mean function. We found that only EV, or FlyFast Airways Inc. was below 7, and we would suggest that Southeast Airlines doesn’t extend their contract with this partner airline.

Our last boxplot is for “Price.Sensitivity”:



This boxplot shows that as Price.Sensitivity increases, Likelihood.to.recommend decreases. It may be advisable for Southeast Airlines to market more to passengers with a lower Price.Sensitivity.

**Linear Model:**

Before completing our linear model, we decided to break up our original data frame into two new ones. We created one data frame for flights that were cancelled and another for flights that weren’t. Due to the fact that cancelled flights would intuitively have a lower Likelihood.to.recommend than flights that weren’t cancelled, it makes sense to separate them. In the end, only 1637 rows of the original data were cancelled flights compared to 86463 flights that weren’t cancelled. After creating the new data frames, we removed any rows from the not\_cancelled data frame that weren’t complete cases. Out of the 86463 rows, only 205 of them were ultimately removed by being incomplete. After that, we removed columns that we thought were either unnecessary for modeling or we assumed were collinear with other variables.

Our linear model used Likelihood.to.recommend as the dependent variables and the rest of the variables as independent variables. While our model only had an r-squared of about .4, the overall p-value for the model was significant at the .01 level. Our model showed with significance that all other Airline statuses compared to Blue increased Likelihood.to.recommend. Furthermore, Age was significant with a negative coefficient, suggesting the older someone is, the lower their Likelihood.to.recommend. Males also had a Likelihood.to.recommend greater than females at significant level. Like we saw from the Price.Sensitivity boxplot, this variable was significant with a negative coefficient, which suggests the higher the Price.Sensitivity, the lower the Likelihood.to.recommend. Flights.Per.Year and Loyalty were also significant with a negative coefficient. The model also shows that Mileage tickets and Personal Travel were significant and negative compared to Business travel. Furthermore, both shopping amount and eating and drinking at airport were significant and positive. Like type of travel, for Class, Eco and Eco Plus were significant and negative compared to Business. Scheduled Departure Hour was significant and negative, suggesting the later the flight, the lower the Likelihood.to.recommend. While arrival delay was negative, departure delay was surprisingly positive, with both variables being significant. Flight distance was positive and significant, while flight time was negative and significant. Lastly, compared to January, March was significant with a positive coefficient. This makes sense as January has worse weather and may impact the flight experience.

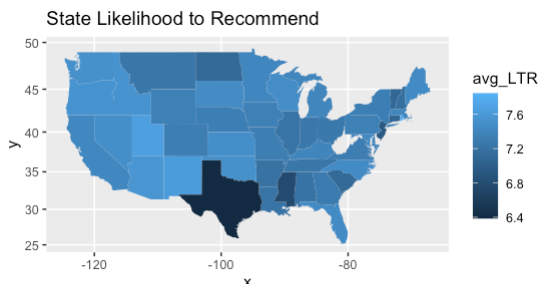
**Association Rules:**

We created two association rules models to see what influences someone being a Detractor or Promoter. The first step was to make all the variables factors that weren’t already. Then, we created a transactions matrix out of our data frame to be used. The first model had the right-hand side as status=Detractor, a support level of 0.1, and a confidence level of .5. Using these levels, the model produced a set of 13 rules. Each of the 13 rules included Personal Travel on the left-hand side, which tells us that Personal Travel is a major indicator for Detractors. The second model had the right-hand side as status=Promoter, a support level of .15, and a confidence level of .5. This produced a set of 12 rules, and each of these rules included Business travel on the left-hand side. Based off of these two models, the indication is that Personal Travel increases customer churn while Business travel decreases it.

To dive further into the differences in between types of travel, we first created three data frames, each one only having rows of one of the types of travel. After creating these, we created proportion tables for the different data frames based on status. This showed us that while 63% of Personal Travel were Detractors, only 7.5% were promoters. Business travel on the other hand had only 14% Detractors and 54% Promoters. Lastly, Mileage tickets only had 41% Promoters, but another 41% were also neutral, leaving only 18% being Detractors. Considering the values of Personal Travel, we would suggest Southeast Airlines take steps to limit Personal Travel and expand Business travel. After some simple calculations, we found that 65% of all Detractors were from Personal Travel, even though Personal Travel is only 30% of the total data. After all, Business travel is majority of data, which makes sense to market more to them and less to Personal Travel.

**US Map:**

Lastly, we created a map of the US, where each state was colored by the average Likelihood.to.recommend of flights where that state is where the flight left from. In order to get these averages, we aggregated the mean Likelihood.to.recommend by Origin.State. Wyoming had an NA value, which was replaced with the average of the rest of the states. We created the map using ggplot2. First, in aes we set the map\_id to state. Next, with geom\_map, we set map equal to us, which is a variable derived from map\_data. Within geom\_map, we set fill equal to the average Likelihood.to.recommend by state. We then used the long and lat variables from us to expand the limits, used the coord\_map to make sure the map looked right, and added a title.



The legend on the right-hand side of the graph shows the color range for average Likelihood.to.recommend. One quick note about the graph is that West Virginia is white. This is because when aggregating by Origin.State, West Virginia did not appear.

The most obvious state on this graph is Texas, which has the lowest average Likelihood.to.recommend of 6.4. The only other two states with an average Likelihood.to.recommend lower than 7 were Mississippi with a value of 6.8, and New Jersey, which was slightly below 7.