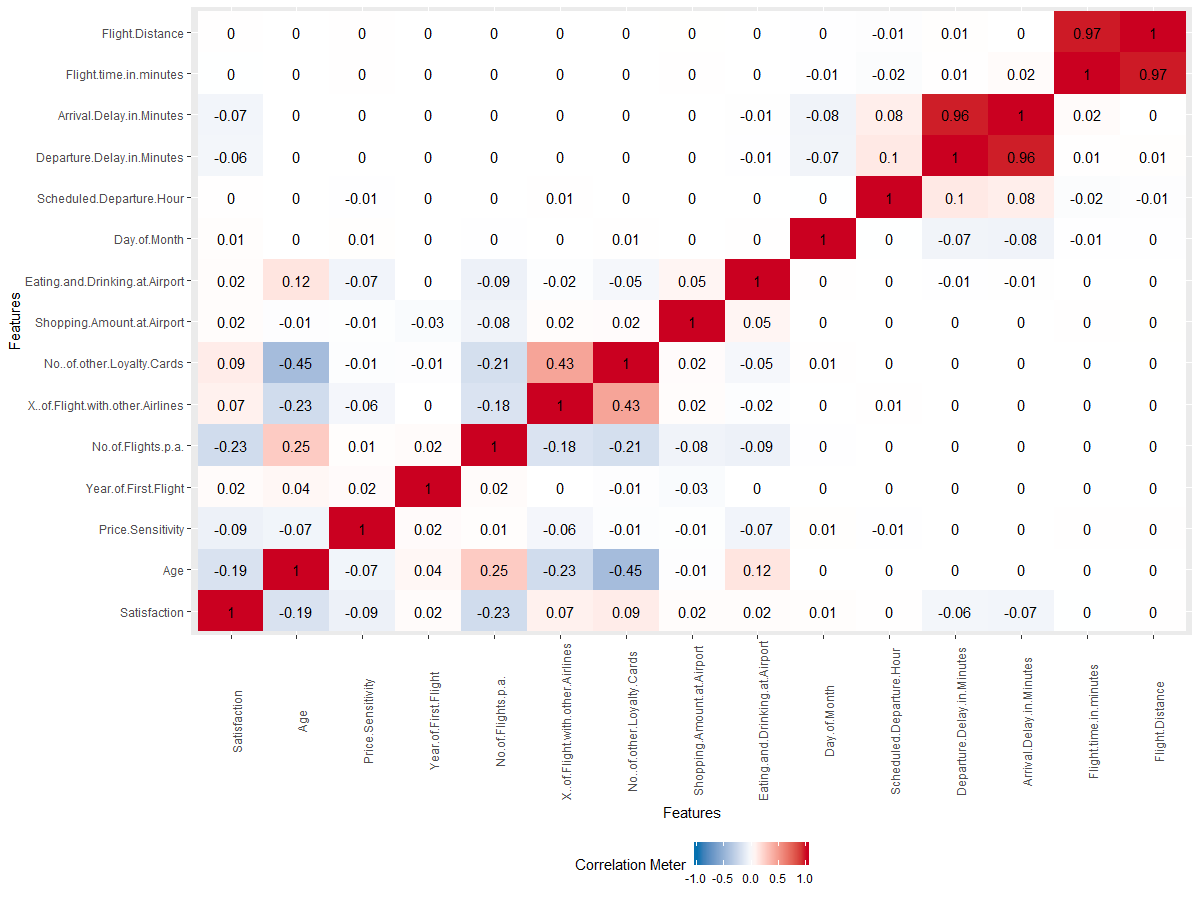
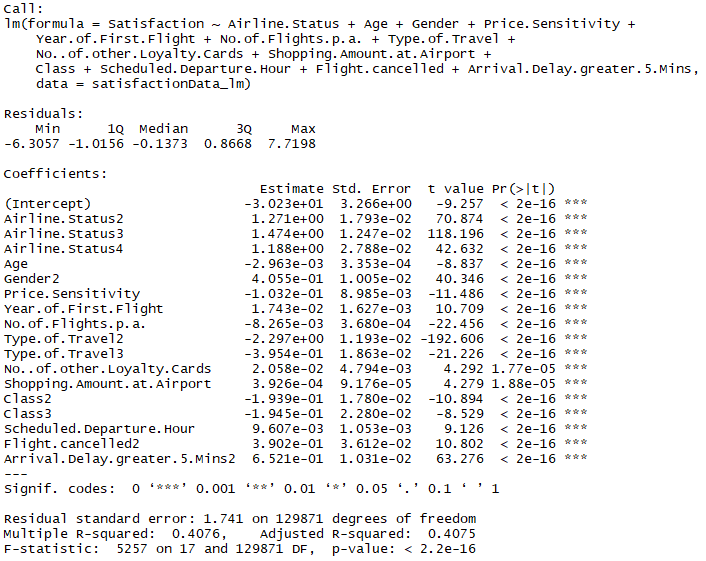
**Multiple Linear Regression Model**

Multiple linear regression model is the first model our group conducted trying to analyze which factors (attributes) significantly affect the ‘Satisfaction’.

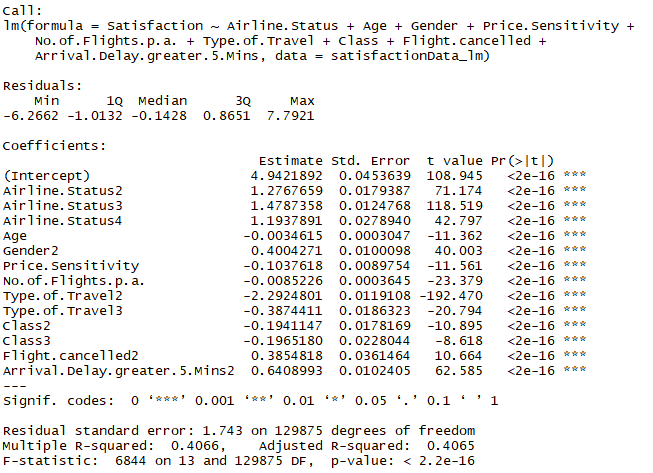
* To build a multiple linear regression model, our first step is to delete 7 meaningless variables from ‘First Flight’ to ‘Destination State’.
* Next, we further detected the correlation between the rest of the variables. The result is shown below.



* From the graph, we could see there are only two correlations which is too high. As we can see, the correlation between ‘Arrival Delay in Minutes’ and ‘Departure Delay in Minutes’ is 0.96 and the correlation between ‘Flight Distance’ and ‘Flight Time in Minutes’ is 0.97. So, we further delete the ‘Arrival Delay in Minutes’ and ‘Flight Time in Minutes’ to avoid autocorrelation issue.
* The next step is to convert categorical variables to dummy variables.
* Then, we used a method called backwardation elimination to optimize our model. By doing this, we firstly included all variables in the model and then, we checked the P-value of each variables. Next, we deleted the variable with the highest P-value each time until we found all the variables are both statistically significant. The result of backwardation elimination is shown below.



* Furthermore, our group moved forward to remove some variables which have no business intelligence such as ‘Scheduled Departure Hour’, ‘Flight Cancelled’, ‘Year of First Flight’, and ‘No. of other Loyalty Cards’. Finally, our final and optimized multiple linear regression model is like this:



As we can see, all the independent variables are significant which means the proposed relationship between the response variable and the set of predictors is statistically reliable.

**Evaluation:**

We evaluated the model by mainly checking its R-square which is 0.4066. It is a goodness-of-it measure for the linear regression model which means 40.66% of the variance in the dependent variable that the independent variables explain collectively. However, R-square has some limitations. The biggest one is that it always increases with the increase of the independent variable. In that case, we should exam the adjusted R-square instead.

**Interpretation:**

Since there are many dummy variables in the model. We firstly tried to interpret the coefficient of the numeric variables.

* For the ‘Age’ variable, its coefficient is about -0.003 which means for every additional age of the customer, the satisfaction of him/her is expected in decrease by an average of -0.003 holding other variables unchanged.
* For the ‘Price Sensitivity’ variable, its coefficient is about -0.104 which means for every unit increase of a customer’s price sensitivity, the satisfaction of him/her is expected in decrease by an average of -0.104 holding other variables unchanged.
* For the ‘No. of Flights p.a.’ variable, its coefficient is about -0.009 which means for every unit increase of number of flight that a certain customer has taken, the satisfaction of him/her is expected in decrease by an average of -0.104 holding other variables unchanged.

So, the three numeric variables listed above both have a negative effect for the satisfaction, however, the effects are relatively small. Next, we would explain the meaning of the dummy variables.

* For the ‘Airline Status’ variable, we could see for all categories, the airline status 3 the highest coefficient which means silver package would contribute the most for increasing the satisfaction holding other variables unchanged.
* For the ‘Gender’ variable, we could see for all categories, the gender 2 the highest coefficient which means male would contribute the most for increasing the satisfaction holding other variables unchanged.
* For the ‘Type of Travel’ variable, we could see for all categories, the type of travel 2 the lowest coefficient which means personal travel would contribute the most for decreasing the satisfaction holding other variables unchanged.
* For the ‘Class’ variable, we could see for all categories, the class 3 the lowest coefficient which means economy service level would contribute the most for decreasing the satisfaction holding other variables unchanged.
* For the ‘Flight Cancelled’ variable, we could see for all categories, the Flight Cancelled 2 the highest coefficient which means if a flight is not cancelled, it would contribute the most for increasing the satisfaction holding other variables unchanged.
* For the ‘Arrival Delay Greater 5 mins’ variable, we could see for all categories, the Arrival Delay Greater 5 mins 2 the highest coefficient which means if an arrival delay is not greater than 5 minutes, it would contribute the most for increasing the satisfaction holding other variables unchanged.

In all, by reviewing all factors related to the satisfaction, we would not focus on factors of ‘Age’ and ‘Gender’ because they cannot be controlled by the airline company. Also, we are not focused on factors of ‘Price Sensitivity’ and ‘No. of Flights p.a.’ because their coefficient is relatively too small to affect the satisfaction. So, the rest of factors which are ‘Airline Status’, ‘Type of Travel’, ‘Class’, ‘Flight Cancelled’, and ‘Arrival Delay Greater 5 mins’ are essential for the satisfaction. Our final business rules extracted from multiple regression model is: the airline company should try their best to avoid flight cancellation and arrival delay greater than 5 minutes.; the company should encourage their customers to buy/use silver packages; the company should detect and solve some problems with the economy service and when customers traveled by their own, because those situations would decrease the satisfaction in a great extent.

**SVMs**

The second model we choose is support vector machines which is an algorithms to do the classification.

* We firstly discretized the satisfaction variable and built a new column where if the satisfaction is higher or equal to 6, the customer is classified by ‘happy’. Otherwise, the customer would be classified by ‘notHappy’.
* Next, we split the dataset into training set for building the classifier and testing set for predicting.
* Next, we build a model by picking up 2 variables which are ‘Price Sensitivity’ and ‘No. of Flights p.a.’.
* Next, we made the prediction on the testing dataset. The confusion matrix is shown below.



We calculated the accuracy for the model is around 60.30%.

* Next, we visualized the prediction result shown below.

A picture containing screenshot

Description automatically generated

* Next, we tuned the parameters in the model to see if the accuracy would increase or not. Here, we considered to add variables and to change the kernel in the model. Our final model includes 5 variables which are ‘Airline Status’ , ‘Age’, ‘Gender’, ‘Price Sensitivity’, ‘Class’, and ‘X. of Flight with other Airlines’. We also changed the kernel in the model from ‘linear’ to ‘radial’ which means the separate is no longer linear, instead, it might be curved or circled.
* After we made the prediction on the testing dataset, the confusion matrix is shown below



We calculated the accuracy for the model is around 62.46%.

**Evaluation:**

Our evaluation of SVMs is focused on accuracy. After we tuned the parameter in the model, the best accuracy we got is about 62.46% which means when we use the model to make the prediction, the correct rate would be around 62.46%. We also tried to use cross validation method to exam the model if it has the problem of overfitting.

**Interpretation:**

Unfortunately, we could not extract any business rules from SVMs model, however, we could use it as a model to predict the satisfaction of new customers. For example, if there is a new customer who select our company, with all information needed in the model, we could roughly predict whether he/she would be satisfied or not.