A close up of a logo

Description automatically generated

IST 687: Introduction to Data Science

Data Analysis and Business Intelligence on Airline Data for

South-East Airlines

Akshay Sigar

Dhwani Gandhi

Honglai Chen

Sheetal Chowdhary

**Professors:**

**Jeffrey Saltz**

**Yatish Hege**

**INTRODUCTION**

**Context & Objective**

We are working as a consultant for Southeast Airlines. We have been given a dataset that contains summaries of air travel within the U.S, one row per trip. The focus of our data analysis project is to provide notions and directions to Southeast Airlines so that they can improve their profit and consumer satisfaction. Concrete analysis of the airlines' data helps us generate some useful insights into customers mindsets which can be used by the Airlines to take better business decisions and improve customer service.

**Background**

The airline dataset contains about 10282 records(rows) and has 32 variables(columns).The columns broadly focus on several categories, including customer’s gender, age, number of flights, shopping amount at the airport, type of travel, class, day of travel, etc.

Likelihood To Recommend is the independent variable here which translates to Customer Satisfaction in business world. Highly satisfied customer is very likely to give positive Likelihood to recommend grade and vice versa. Our goal here is to understand what factors influence a customer’s Likelihood To Recommend grade taking a particular flight. There can be different relationships between Likelihood To Recommend and the other 31 variables. Some of them will influence Likelihood To Recommend grade while some will not. Some might influence positively while the others negatively.

In our analysis of the data, we have identified and reviewed the key performance indicators like flight cancellations, flight delays, age, gender etc. that affect Likelihood To Recommend. We created correlational trends between various variables of the data set and Likelihood to Recommend using the analysis models. We did this by creating histograms and bar charts for each variable to achieve this. This gave us an insight on the customer information and helped us analyze it further using different data analysis models.

Moreover, we used system modeling in Correlation Matrix to prepare our data before we do the final modeling part. In this report, we use the Linear Modeling and Association Rule Mining to do analysis of our data, and another model, which is Support Vector Machine to validate our result. This helped us answer the business questions formulated for the data set. On answering the business questions we could infer better insights and solutions for increasing the customer satisfaction and airline services.

We also used values of variables like Origin and Destination Latitudes and Longitudes to create a flight route map for low satisfaction routes.

We also did some sentiment analysis using the Comments column text.

**Scope**

We will only be focusing on following 32 variables provided in the data set. They are defined as follows :

***1****.* ***Likelihood to Recommend*** *– Rated on a scale of 1 to 10, which shows how likely the customer is to recommend the airline to their friends (10 is very likely, and 1 is not very likely).*

***2.******Airline Flyer Status*** *– Each customer has a different type of airline status, which are platinum, gold, silver, and blue (based on level of travel with the airline)*

***3****.* ***Age*** *– The specific customer’s age. Ranging from 15 to 85 years old.*

***4.******Gender*** *– Male or Female.*

***5****.* ***Price Sensitivity*** *– The grade to which the price affects customers purchasing. The price sensitivity has a range from 0 to 5.*

***6****.* ***Year of First Flight*** *– This attribute shows the first flight of each single customer. The range of year of the first flight for each customer has been started in 2003 until 2012.*

***7****.* ***Flights Per Year*** *– The number of flights that each customer has taken in the most recent 12 months. The range starting from 0 to 100.*

***8****.* ***Loyalty*** *– An index of loyalty ranging from -1 to 1 that reflects the proportion of flights taken on other airlines versus flights taken on this airline. A higher index means more loyalty.*

***9****.* ***Type of Travel*** *– One of business travel, mileage tickets, or personal travel (ex. vacation)*

***10****.* ***Total Frequent Flyer Accounts*** *– How many frequent flyer accounts the customer has.*

***11****.* ***Shopping Amount at Airport*** *– The spending on non-food & services at the airport (in $)*

***12****.* ***Eating and Drinking at Airport*** *– The spending on food/drink at the airport (in $).*

***13****.* ***Class*** *– Three different kinds of service level (business, economy plus, and economy).*

***14****.* ***Day of Month*** *–The traveling day of each costumer (ranges from 1 to 31).*

***15****.* ***Flight date*** *– The passenger’s flight date of travel.*

***16****.* ***Partner Code*** *– This airline works with wholly- and partially-owned subsidiary companies to deliver regional flights. For example, AA, AS, B6, and DL.*

***17****.* ***Partner Name*** *– These are the full names of the partner airline companies.*

***18****.* ***Origin City*** *– The place where passenger departed from. For example, Boston MA.*

***19****.* ***Origin State*** *– The place where passenger departed from. For example, Texas.*

***20****.* ***Destination City*** *– The place to which passenger travels to. For example, Boston MA.*

***21****.* ***Destination State*** *– The place to which passenger travels to. For example, Texas.*

***22****.* ***Scheduled Departure Hour*** *– The specific time at which the plane was scheduled to depart.*

***23.******Departure Delay in Minutes*** *– How long the flight’s departure was delayed, when compared to schedule.*

***24****.* ***Minutes*** *– How long the arrival was delayed.*

***25****.* ***Flight Arrival Delay in Cancelled*** *– Occurs when the airline does not operate the flight.*

***26****.* ***Flight time in minutes*** *–The length of time, in minutes, to reach the destination.*

***27****.* ***Flight Distance*** *– The distance between the departure and arrival destination.*

***28****.* ***Comment*** *– A free form text field of the passenger comment, with respect to the flight.*

***29****.* ***olong*** *– Origin City Longitude*

***30****.* ***olat*** *– Origin City Latitude*

***31****.* ***dlong*** *– Destination City Longitude*

***32****.* ***olat*** *– Destination City Latitude*

**BUSINESS QUESTIONS**

We have identified following business questions that we will try to answer by our analysis :

1. What attributes influence the likelihood of customers recommending the airline

services and providing it a good feedback? Is it negatively or positively relevant to the customers’ satisfaction?

1. How is the overall satisfaction of Southeast Airlines for each Partner Airline?
2. Does flight distance, flight time, arrival or departure delay affect customer satisfaction?
3. How does age affect customer satisfaction? Which age group gives lower ratings?
4. Is there any relationship between gender and customer satisfaction?
5. Does Year of first flight and no.of flights taken in a year affect customer satisfaction?
6. Is it possible to associate airline status, type of travel and class with customer rating?
7. How much does Shopping, eating and drinking at the airport affect customer satisfaction?
8. Does date of travel affect customer satisfaction?
9. Do cancelled flights cause customers to give a low customer rating?
10. Does Price Sensitivity affect Customer Satisfaction?
11. How can detractors be converted into promoters based on satisfaction rate?
12. What are the facilities and services provided by airlines that need to be focused

on to increase overall satisfaction rate for airlines?

1. What corrective actions can be done to increase the customer’s satisfaction?

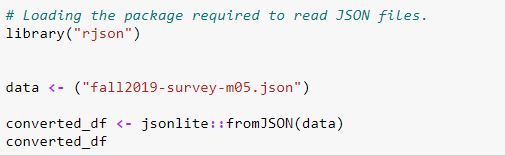
**DATA REQUISITION AND PRE PROCESSING**

Good analysis rests on clean data - The first step of our analysis is to clean up the data that we’ve got. Doing the basic analysis of data and processing it gives us the right direction to our project. Prima Facie, it seems some of the information may not be helpful for our analysis depending on what we want to achieve. So, in this part, we will load the data, clean it and do the munging.

**Data Acquisition**

Our professor provided us with a dataset that contains customer information of Southeast Airlines. This data set is a csv file that contains the required data.

We downloaded the file using the following code:



**Data Cleaning**

Before processing the data further we had to clean the data as many columns like departure delay in minutes, arrival delay in minutes, Flight time in minutes etc. contained a lot of NA values.

Flight time in minutes has 248 missing values, Likelihood.to.recommend had 1, Departure.Delay.in.Minutes had 212 and Arrival.Delay.in.Minutes had 248.

We replaced NAs in a way which could have the least effect on original Mean and Median values of the column and data remains correct in the next phase that is the analysing process such as linear modeling.

We replaced NAs in Flight time in minutes and Likelihood.to.recommend with Mean and NAs in Departure.Delay.in.Minutes and Arrival.Delay.in.Minutes with Median.

Only the columns relevant to the analysis of the data were kept and the remaining were eliminated. Columns such as Origin State and Destination state can be found out by the Origin city and Destination city columns. Thus, eliminating the State columns makes sense.

After eliminating attributes which were not important for the analysis, the dataset is narrowed to limited number of columns which can be used for the analysis and visualization process.

The latitudes and longitudes of origin and destination are not useful for analysis but can be used to create maps for low satisfaction flight routes. So we will keep them.

We have done some Sentiment analysis at later stage so we have kept Comments column and not removed it.

**Data Munging**

Next, Data Munging was used to transform and map data from one raw form into another format with the intent of making it more appropriate and valuable for a variety of downstream modelling.

Columns like Age,Gender,Airline.Status,Type.of.Travel, Class, Day.of.Month, Flight.cancelled, Partner.Code, Partner.Name were converted to Factors.

Columns like Departure.Delay.in.Minutes, Arrival.Delay.in.Minutes, Scheduled.Departure.Hour were converted to numeric.

We converted the Likelihood To Recommend into three categories namely “Promoter”,”Passive” and “Detractor”. So customer with Likelihood To Recommend from 1 to 6 is a Detractor, 7-8 is Passive and 9-10 is a Promoter. We will do further analysis using this new Category column.

**USE OF DESCRIPTIVE STATISTICS AND**

**DATA VISUALIZATION**

In this part, we are trying to find the relationship between the factors by using GGMAP,

and draw the airline information on the map. We also used bar charts, histograms and density graphs in GGPLOT2 to compare the relationship between two factors one of them being Likelihood To Recommend or Category depending on what visualization we are performing.

**Airline Status versus Likelihood To Recommend**

The graphs of Airline status tell us that there is no linear relationship

between these two factors. But we know that we get a higher satisfaction from silver

customer and lowest for Blue

**Age versus Likelihood To Recommend**

From the graphs of Age, we can see that the age between 30-50 will

Are more likely to recommend the flight. And the satisfaction decreases as the age

increases.

**Gender versus Likelihood To Recommend**

Graphs showed us that Males promote the flight more than Females. Females are more of Detractors.

**Price Sensitivity versus Likelihood To Recommend**

From the graphs of Price Sensitivity show that when Price Sensitivity is higher detractors grow.

**Year of First Flight versus Likelihood To Recommend**

Year of First Flight didn’t affect the Likelihhod to recommend grade as concluded from the graphs. For all years we saw almost the same pattern. i.e.Promoters higher than detractors.

**Flight Per Year versus Likelihood To Recommend**

From the graphs of Flights Per Year versus Category, we can see that better feedback appeared when total number of flights of a customer in a year is less than 30.

**Loyalty versus Likelihood To Recommend**

From the graphs we concluded that when flyers who are more loyal are more likely to recommend the airline.

**Type of Travel versus Likelihood To Recommend**

From the graphs we concluded that if travel type was ‘Business Travel’ Likelihood to Recommend was highest. For Personal Travels, there were more detractors than promoters.

**Total Frequency Flyer Accounts versus Likelihood To Recommend**

Total Frequency Flyer Accounts didn’t affect the Likelihood to Recommend grade as concluded from the graphs.

**Shopping Amount At Airport versus Likelihood To Recommend**

Shopping at Airport didn’t affect the Likelihood to Recommend grade as concluded from the graphs. For all amounts we saw almost the same pattern. i.e. Promoters were higher than detractors.

**Eating & Drinking At Airport versus Likelihood To Recommend**

Graphs showed us that if flyer spent money on eating and drinking and if that amount was in a range of 50$ to 150$ they promoted the airline. On the other hand if Flyer spent no money and easting and drinking he was not satisfied.

**Class versus Likelihood To Recommend**

From the graphs for Class, we concluded that ‘Economy’ and Business Class Flyers have higher Promoters compared to Detractors. For some reason, Economy Plus has equal number of Promoters, Passive Flyers and Detractors.

**Day of Month versus Likelihood to Recommend**

Day of the Month didn’t affect the Likelihood to Recommend grade as concluded from the graphs. On all days Promoters were higher than detractors.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |

**Partner Name versus Likelihood To Recommend**

We concluded from the graphs for most of the partners Likelihood to Recommend grade is higher. But for Partner ‘FlyFast Airways’ Detractors are more than Promoters

**Scheduled Departure Hour versus Likelihood To Recommend**

Scheduled Departure Hour didn’t affect the Likelihood to Recommend grade as concluded from the graphs. For all distances we saw almost the same pattern. i.e. Promoters were higher than detractors.

**Departure Delay In Minutes versus Likelihood To Recommend**

Graphs for Departure Delay In Minutes showed us that that most airplanes usually depart on time and the Likelihood To Recommend grade is almost stable after the delay time is more than around 20 minutes. This means that the Likelihood To Recommend level is relatively high only when departure delay time is less than 20 minutes.

**Arrival Delay In Minutes versus Likelihood To Recommend**

Graphs for Arrival Delay In Minutes showed us that that most airplanes arrived on time and the Likelihood To Recommend grade is almost stable after the delay time is more than around 10 minutes. This means that the Likelihood To Recommend level is relatively high only when arrival delay time is less than 20 minutes.

|  |  |
| --- | --- |

**Flight Cancelled versus Likelihood To Recommend**

From the graphs of Flight Cancelled we could conclude that most of flights did not get cancelled and customers whose flights were not cancelled had a higher Likelihood to Recommend grade and were Promoters

**Flight Time In Minutes versus Likelihood To Recommend**

Flight Time In Minutes didn’t affect the Likelihood to recommend grade as concluded from the graphs. For all times we saw almost the same pattern. i.e. Promoters were higher than detractors.

**Flight Distance versus Likelihood To Recommend**

Flight Distance didn’t affect the Likelihood to recommend grade as concluded from the graphs. For all distances we saw almost the same pattern. i.e. Promoters were higher than detractors.

**USE OF MODELLING TECHNIQUES AND**

**DATA VISUALIZATION**

**Correlation Matrix**

A correlation matrix is a table where the correlation coefficients between two variables

is presented at the crossing point of the certain row and column. Each cell in the table

suggests how relevant it is between two variables. A correlation matrix is often applied

as a way to summarize data, as an input into a more advanced analysis, and as a

diagnostic for advanced analysis.

Therefore, the correlation matrix can be the first step for us to briefly examine the correlation between Likelihood to recommend and other attributes, helping us to select the most relevant attributes for further analysis and study. We converted all attributes to numeric and then created a Correlation Matrix.

What we concluded is Age,Flights per year,Departure delay in minutes, Arrival delay in minutesPrice Sensitivity are negative factors which means as they increase Likelihood to recommend decreases.

AirlineStatus, TypeofTravel,Loyalty are positive factors which means as they increase Likelihood to recommend increases. It also shows some positive relationship with Eating and Drinking at Airport and Total Freq Flyer Accounts.

For Gender we set Female=0 and Male=1 and it showed positive correlation which means Males are more likely to recommend than Females.

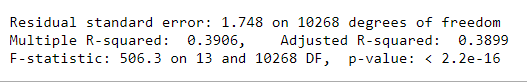
We have high level idea about how factors affect now which we will confirm in following models.

**Linear Modelling**

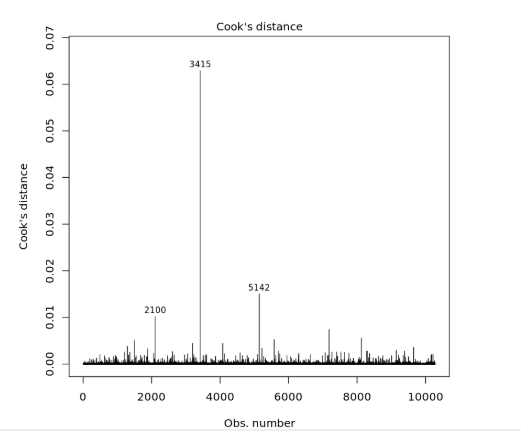
Next, we applied linear modeling on the airline dataset. By using the linear regression we could study and understand the relation between the variables in the dataset. The purpose of using linear regression was to find the line that best fits the data. The best fit line is the one where the prediction error is low. Out of all the variable, we could find some of the significant variables which are responsible for determining the satisfaction of a customer. We analyzed the adjusted R-squared value which needs to be high, the p-value which needs to be less than 0.05 and the residual plots that need to be symmetric. The significant variables which we got from the linear model are:

1. Flights.Per.Year
2. Class
3. Age
4. Gender
5. Price Sensitivity
6. Arrival Delay in Minutes
7. Departure Delay in Minutes
8. Type of Travel
9. Airline Status

Here is out output :

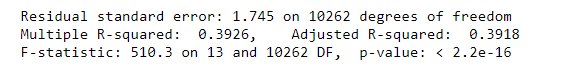


We plotted some graphs as shown below:



We removed some rows identified in the linear modelling plots that might be affecting the R square value. After removing them our R square improved somewhat from 38.99 to 39.18.

New output is as follows :



**Association Rule Learning**

Apriori is an algorithm for frequent item set mining and association rule learning over relational databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis.

In this model, Apriori Algorithm helps to define how different combinations of variables will affect the satisfaction (Likelihood.to.recommend ) with the a filter to select minimum 0.001 support and 0.5 confidence.

The results shows that the customers with “Personal Travel”, “Agedness”, “Female”, “Blue Airline Status”, “Flight.cancelled = NO” will have a higher probability to be “Detractors”.

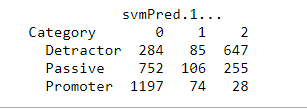
Following is the output from Apriori model :

|  |  |  |
| --- | --- | --- |
| Attributes | Probability | Num |
| Type.of.Travel=Personal Travel | 0.1968 | 2024 |
| Type.of.Travel=Personal Travel  Flight.cancelled=No | 0.1920 | 1974 |
| Airline.Status=Blue  Type.of.Travel=Personal Travel | 0.1731 | 1780 |
| Type.of.Travel=Personal Travel  Airline.Status=Blue  Flight.cancelled=No | 0.1688 | 1736 |
| Type.of.Travel=Personal Travel  Class=Eco | 0.1636 | 1682 |
| Type.of.Travel=Personal Travel,  Airline.Status=Blue,  Flight.cancelled=No  Class=Eco | 0.1408 | 1448 |
| Gender=Female,  Type.of.Travel=Personal Travel, | 0.1320 | 1357 |
| Gender=Female,  Type.of.Travel=Personal Travel,  Flight.cancelled=No | 0.1287 | 1323 |
| Age=Agedness | 0.1181 | 1215 |
| Flight.cancelled=No  Age=Agedness | 0.1149 | 1181 |

**Support Vector Machine (SVM Model)**

Support Vector Machine creates a hyperplane to divide our model. It separates the data into three categories called Detractors (Likelihood to recommend is between 1-6), Passive(Likelihood to recommend is between 7-8) and Promoters(Likelihood to recommend is between 9-10). We divide the data into two i.e. the train data set and the test data set. The variables used to train the data are Flights Per Year, Price Sensitivity, Class, Age, Gender, Arrival Delay in Minutes, Departure Delay in Minutes, Type of Travel, and Airline Status. After training the model we predict the number of detractors, passive and promoters using the test data. We build a confusion matrix and using this we calculate the accuracy rate. We received an accuracy rate of 56.88%

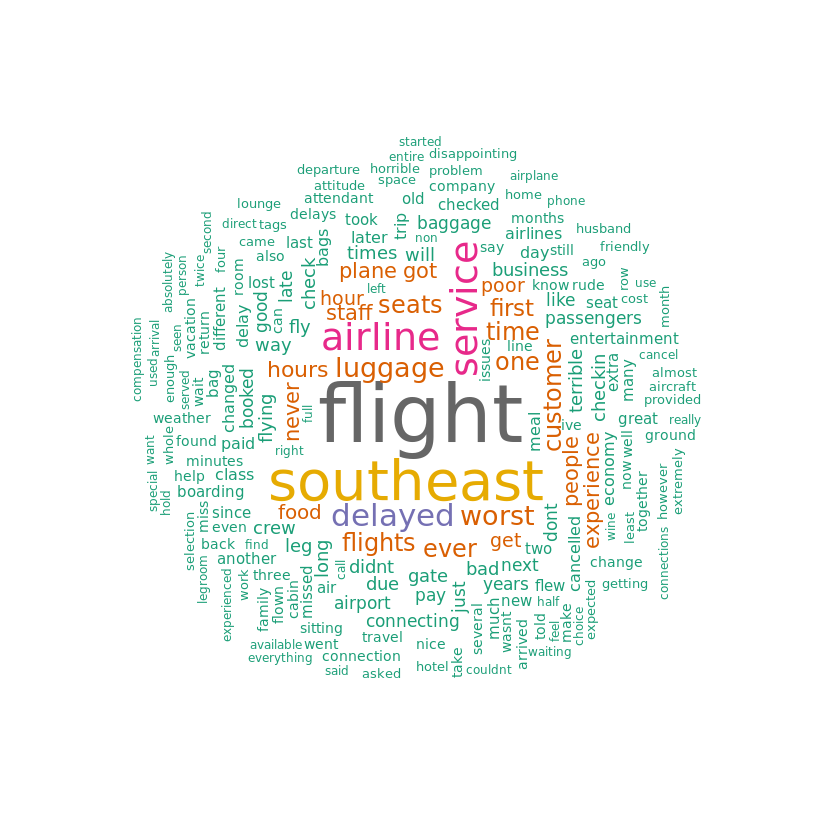
Output of the modelling :



The highlighted numbers tells us the number of times the model predicted the customer satisfaction accurately.

**Sentiment Analysis**

We have done some sentiment analysis on comments of Detractors and identified the frequently used words.

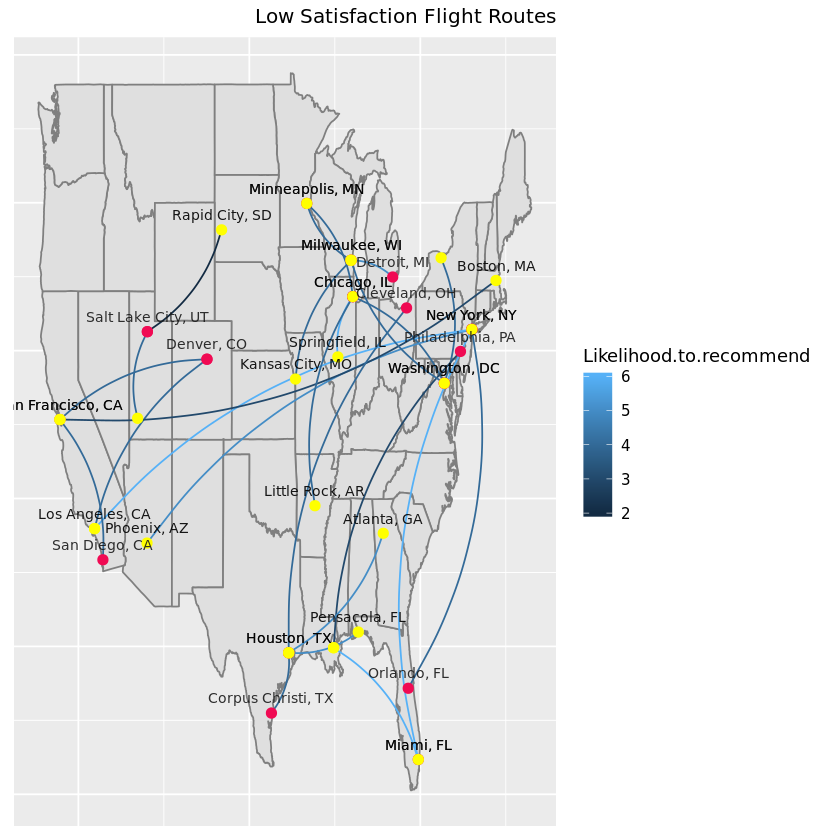


We can see negative words like poor,delayed,worst,never,airline, service that were used by detractors which probably means detractors were not happy with Southeast Airline Services.

This model was not very helpful for gaining any data insights because there were very few records that had text for comments .

**Visualization for Low Satisfaction Routes**

We identified unique routes taken by detractors and did sampling of the routes so that we can have good visualization of the routes on the map.

****

**RECOMMENDATIONS**

* Southeast Airlines should cancel the contract with Flyfast Airways or should reduce the flight frequency associated with them.
* Airline should provide free lounge access to customers when the flight delays.
* Provide wheelchairs/escort vehicle and personnel to help the senior citizens.
* Provide teenagers with some goodies and entertainment services with high speed WiFi.
* Female customers should be given more preferences for window seat.
* Give customers with blue airline status some incentive points to upgrade their airline status.
* Provide customers making personal trips with family plan packages.
* Provide economy class customers with better leg space and with occasional upgrades in the class.
* Airline should provide discounts to customers when the prices are high.

**MIDST USAGE/INFO**

We used the MIDST tool as instructed by our professor to divide our code into chunks and assign tasks to each team member. To approach our problem statement, we divided our code into necessary chunks according to each task as explained earlier.

Our MIDST link is provided below:

**MIDST link:** <https://midst-test.syr.edu/projects>

Project Name - [IST687M005 Team2](https://midst-test.syr.edu/project/5d9bbb91053d35ab1e8d7e9f)