6100 Big Data Analytics for Competitive Advantage

Satisfaction & Member Analysis for Airline Industry

Submitted to: Dr. Dongsong Zhang

Ashwami Dalvi - adalvi4@uncc.edu

Simran Satyavolu - asatyavo@uncc.edu

Syed Muhammad Suffwan - ssuffwan@uncc.edu

Utwej Sai Nalluri - unalluri@uncc.edu

1. **Introduction**

    a. **Introduction & Context**

For most companies, one of the biggest incentives to invest in advertising and different marketing strategies is for the purposes of customer loyalty, especially when that loyalty directly affects the company's reputation and dependability. For airlines in particular, customer loyalty has become one of the most telling signs of an airline company's success. For this reason, it is important to acknowledge the impact of customer loyalty, along with several other factors, in order to be able to predict certain aspects of an airline's attainment of profit. Of course, there are other variables, such as in-flight service, flight distance, the class within which a customer is flying, and the reasons for flying.

    b. **Problem Statement**

Research has been done in recent years concerning the importance of customers' overall satisfaction and loyalty on an airline's ability to have loyal customers. It is shown that good in-flight service does have some sort of impact on customer loyalty and satisfaction. However, is this the only variable to be taken into account? And how can we pinpoint exactly what parts of a flight experience or marketing strategy genuinely make customers more loyal to an airline company?

    c. **Project Aims & Goals**

The main goal of this project is to have a bird's eye view of all of these different features and then analyze them to predict which aspects can be improved in order to increase customer memberships and satisfaction overall.

The dataset that will be used for this project is provided by teejmahal20 on kaggle.com. For the given variables, we would explore the relationship between multiple independent and one dependent variable (Membership & Satisfaction) to study correlation between variables in order to quickly spot patterns and reduce a large amount of random data into a set of meaningful relationships.

We want to be able to determine the significance of different features for an airline to actively modify in order to improve the number of customer membership and satisfaction rate.

2. Literature Review

a. Airline service quality:

A comparison study among China domestic airlines conducted by Jiang and Zhang (2016) explain service quality is a significant factor to influence airlines passenger satisfaction levels and customer loyalty. In this study, a regression model was used to generate factor scores, which were then used as independent variables in probit models. Probit models were used to examine factors that determine customer satisfaction and loyalty.

In many studies, survey methods have been used to measure airline service quality. Most of these studies have been presented to be examined in the context of service quality index (Waguespack & Rhoades, 2014), passenger satisfaction survey (Bellizzi, Eboli, & Mazzulla, 2020) as well as airlines' service quality perception (Suki, 2014). Some studies assess the quality of airline service

included on-time performance, handling baggage, food quality, seat comfort, check-in, and in-flight service (Bellizzi et al., 2020)

b. Airline passenger satisfaction in Comparison:

Understanding passengers' expectations in an airline service industry is essential since passengers compare their performance with their expectations. Research from Hu and Hsiao (2016) using the Kano model in quality risk assessment for Taiwanese airlines case study mentions poor airline service quality causes passenger dissatisfaction. Chen (2008) explains airlines passenger satisfaction is indirectly affected by airlines perceived service quality performance moderated by perceived value. Although this project will not use the same model, similar traits and variables will be studied.

Many studies have been done to assess the service quality of airlines and airline passenger satisfaction levels using traditional statistical testing while others have used multiple-criteria methods to accomplish the same goals. In the field of machine learning method, the study of airline passenger satisfaction is usually measured using sentiment analysis (Khan & Urolagin, 2018; Kumar & Zymbler, 2019; Lucini, Tonetto, Fogliatto, & Anzanello, 2020), which is analyzing texts, tweet or comments to detect positive or negative satisfaction. In other words, this study aims to investigate airline passenger satisfaction using more advanced methods of big data machine learning approach from complex surveys to detect the priority aspect of airlines services.

Our project will be analyzing more quantitative data in order to more accurately understand customer satisfaction in a form that can be measured. In this way, the algorithms that will be implemented in our research project will have a more mathematical approach while still coming to similar conclusions.

3. Data Description

The source that will be used for this project is customer information from American Airlines, provided by teejmahal20 on kaggle.com. This data contains 129,880 rows in total. The dataset has 25 variables. 18 of them are integer variables, 5 are String variables, 1 and an ID variable, and 1 variable was of another type (F1). This F1 variable was rejected, as it was very similar to an ID variable and deemed unimportant. The variables are as follows: Inflight Wifi Service, Inflight Service, Member, Leg Room, Gender, Gate Location, Inflight Entertainment, ID, Type of Travel, Satisfaction, Online Boarding, On Board Service, Seat Comfort, Class, Check in Service, Cleanliness, Arrival Delay in Minutes, Age, Baggage Handling, Flight Distance, Food and Drink, Departure Delay in Minutes, Departure Arrival Time Convenience, and Ease of Online Booking. Below is a snippet of the dataset.

Name	Role	Level	Type	Length
Inflight_wifi_service	**Input**	**Interval**	Numeric	8
Inflight_service	**Input**	**Interval**	Numeric	8
Member	**Input**	**Nominal**	Character	10
Leg_room_service	**Input**	**Interval**	Numeric	8
Gender	**Input**	**Nominal**	Character	6
Gate_location	**Input**	**Interval**	Numeric	8
Inflight_entertainment	**Input**	**Interval**	Numeric	8
id	**ID**	**Nominal**	Numeric	8
Type_of_Travel	**Input**	**Nominal**	Character	15
satisfaction	**Target**	**Nominal**	Character	23
Online_boarding	**Input**	**Interval**	Numeric	8
On_board_service	**Input**	**Interval**	Numeric	8
Seat_comfort	**Input**	**Interval**	Numeric	8
Class	**Input**	**Nominal**	Character	8
Checkin_service	**Input**	**Interval**	Numeric	8
Cleanliness	**Input**	**Interval**	Numeric	8
Arrival_Delay_in_Minutes	**Input**	**Interval**	Numeric	8
Age	**Input**	**Interval**	Numeric	8
Baggage_handling	**Input**	**Interval**	Numeric	8
Flight_Distance	**Input**	**Interval**	Numeric	8
F1	**Rejected**	**Interval**	Numeric	8
Food_and_drink	**Input**	**Interval**	Numeric	8
Departure_Delay_in_Minutes	**Input**	**Interval**	Numeric	8
Departure_Arrival_time_convenien	**Input**	**Interval**	Numeric	8
Ease_of_Online_booking	**Input**	**Interval**	Numeric	8

4. **Data Analytics tools and platforms**
    1. **R/R studio**: Data visualization, data exploration, data wrangling and Classification models.
    2. **SAS Enterprise Guide**: For building, data exploration, testing and evaluating Classification models.
    3. **SAS Enterprise Miner Client:** Data preprocessing and preparing it for respective models or analysis.
    4. **Excel:** Pivot tables and slicers used to visualize data.

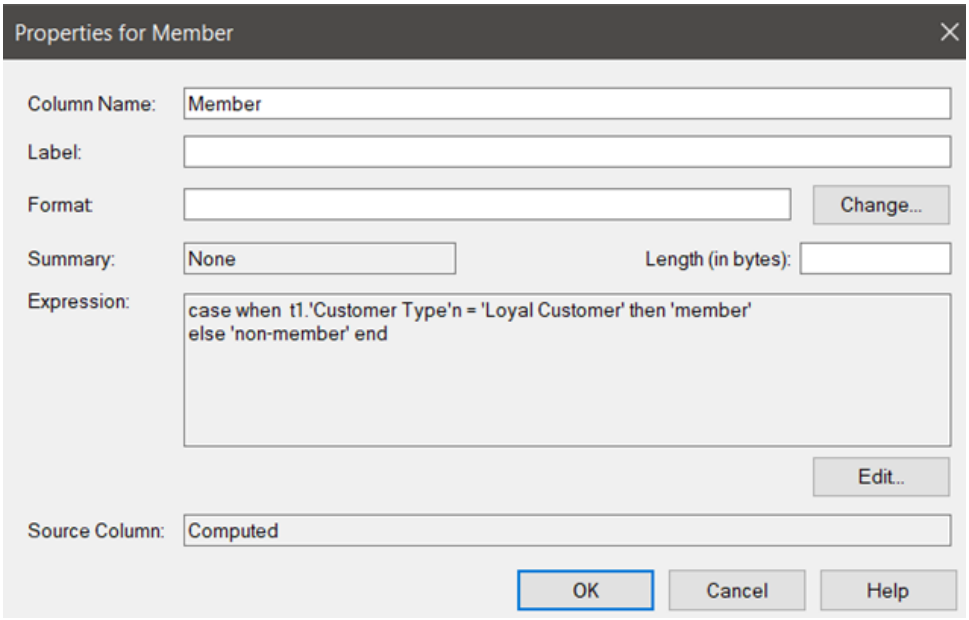
5. Data Preprocessing

a. **Data Merge**:

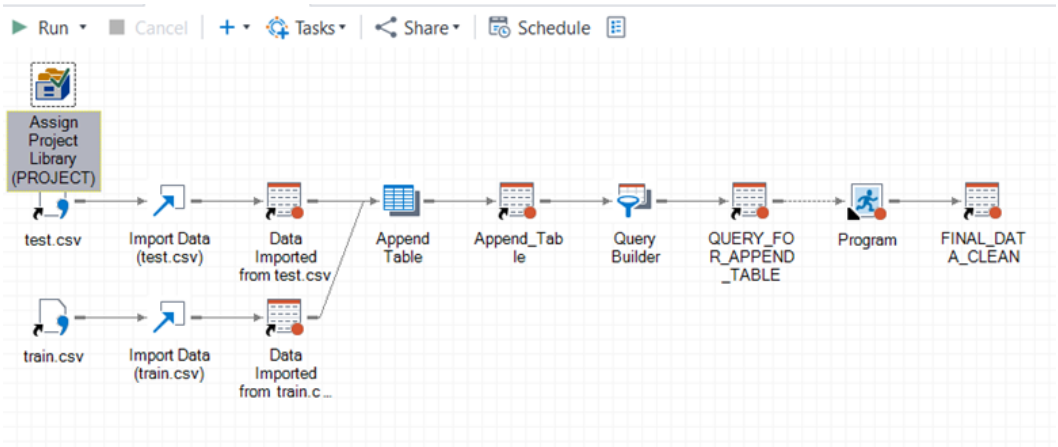
We got our dataset from kaggle split into two .csv files (train and test) (train.csv has got 103904 and consists of 25 variables. test.csv has got 25976 observations and consists of 25 variables. The attributes and the data are explained in the data description part) but we wanted to use the whole dataset for our data exploration and we wanted to create our own splits for modeling purposes to ensure no biases get attached.Hence, we merged the dataset. We performed this by using the ‘Append table’ function in the SAS enterprise guide.

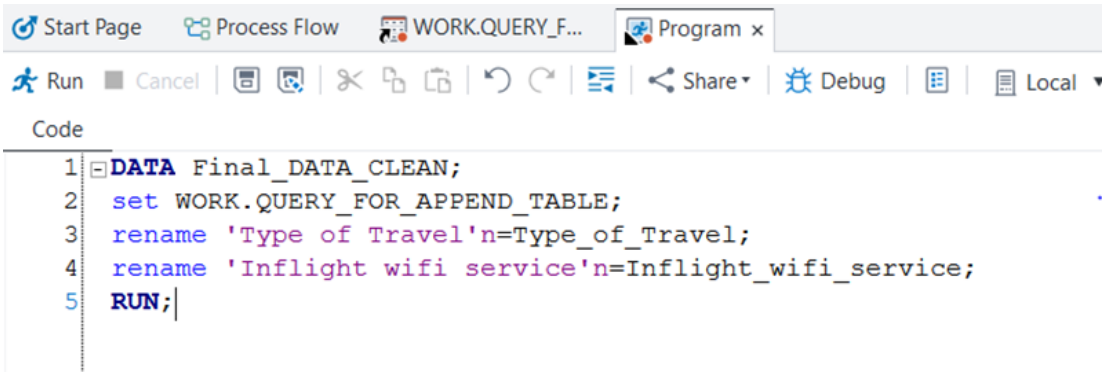
b. **Transforming dataset**:

There is one column, ‘Customer Type’ in our dataset which has two values; ‘Loyal’ and ‘Disloyal’. These two terms are very strong and also ambiguous in the usage so we computed a new column, ‘Member’ based on it.



In order to perform a few data modeling tasks and exploration, we restored to SAS Miner Workstation. But we faced an error in uploading our dataset as the column names had spaces into it. We used a **Program** in SAS Enterprise Guide to write a rename code in order to correct our column names.





c. **Dropping an attribute:**

The ‘F1’ feature is just another column form of ID and we would be using the ‘id’ column as a unique identifier, hence we rejected this attribute.

d. **Data normalization:**

Min-max normalization was performed on all the variables of the dataset, using the ‘Transform Variable’ function in SAS Miner.

Name	Method	Number of Bins	Role	Level
Age	**Default**	**4**	Input	Interval
Arrival_Delay_in_Minutes	**Range Standardization**	**4**	Input	Interval
Baggage_handling	**Range Standardization**	**4**	Input	Interval
Checkin_service	**Range Standardization**	**4**	Input	Interval
Class	**Default**	**4**	Input	Nominal
Cleanliness	**Range Standardization**	**4**	Input	Interval
Departure_Arrival_time_convenien	**Range Standardization**	**4**	Input	Interval
Departure_Delay_in_Minutes	**Range Standardization**	**4**	Input	Interval
Ease_of_Online_booking	**Range Standardization**	**4**	Input	Interval
F1	**Default**	**4**	Rejected	Interval
Flight_Distance	**Range Standardization**	**4**	Input	Interval
Food_and_drink	**Range Standardization**	**4**	Input	Interval
Gate_location	**Range Standardization**	**4**	Input	Interval
Gender	**Default**	**4**	Input	Nominal
Inflight_entertainment	**Range Standardization**	**4**	Input	Interval
Inflight_service	**Range Standardization**	**4**	Input	Interval
Inflight_wifi_service	**Range Standardization**	**4**	Input	Interval
Leg_room_service	**Range Standardization**	**4**	Input	Interval
Member	**Default**	**4**	Input	Nominal
On_board_service	**Range Standardization**	**4**	Input	Interval
Online_boarding	**Range Standardization**	**4**	Input	Interval
Seat_comfort	**Range Standardization**	**4**	Input	Interval
Type_of_Travel	**Default**	**4**	Input	Nominal
satisfaction	**Default**	**4**	Target	Nominal

e. **Data Split:**

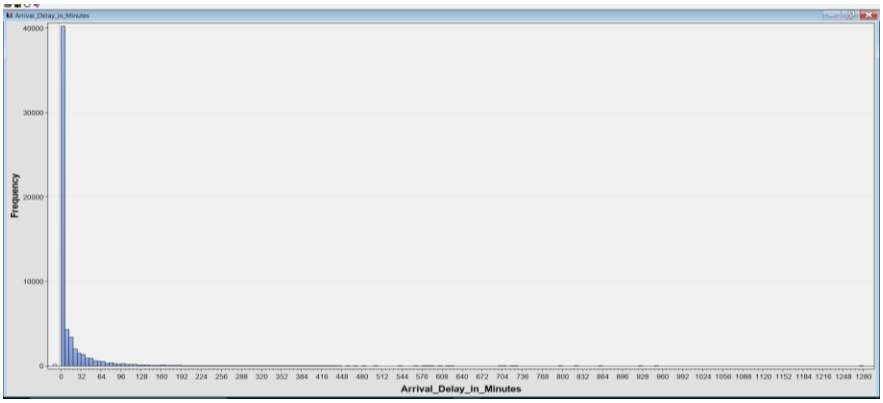
We planned to apply a few modeling techniques using Decision tree and Logistic Regression and hence we used the ‘Data Partition’ to perform a random 80-20 split of our dataset.

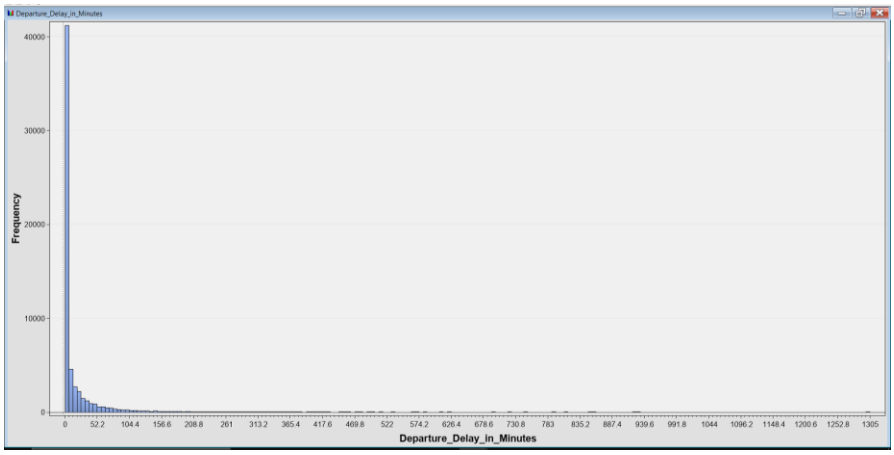
f. **Data Filtering:**

There were 393 missing values of Arrival_Delay_in_Minutes and we dropped the rows. We dropped because it is just 0.3% of the total data.

- Total rows: 129880 - Missing values: 393 - Further used: 129487

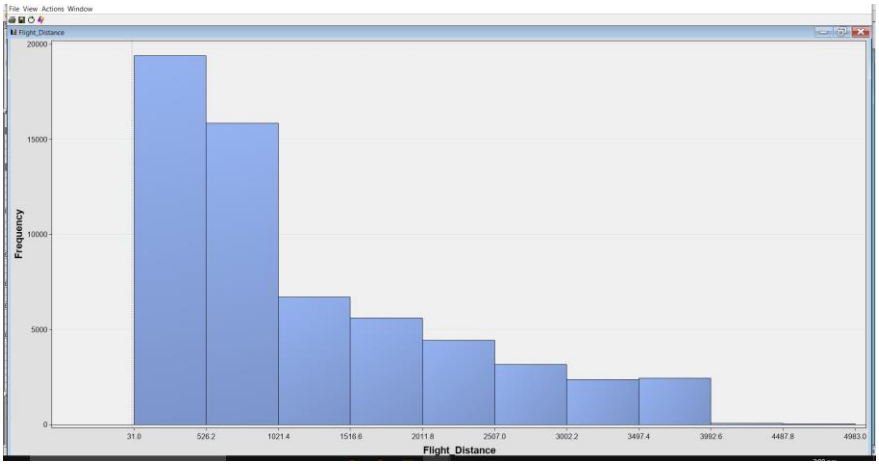
Arrival and Departure delay in minutes are both skewed to the extreme right thus dominating our clustering output.



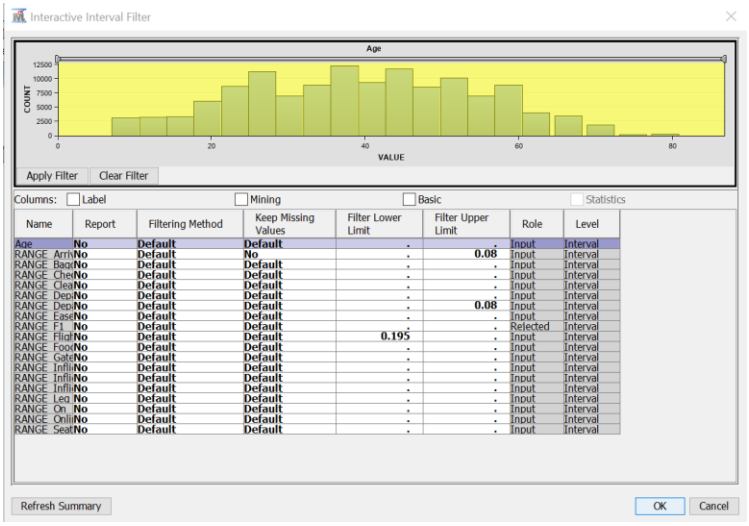


Now if we see the relation between both these variables, a linear relationship exists between them. So for an increase in Arrival delay, the departure delay would also increase.

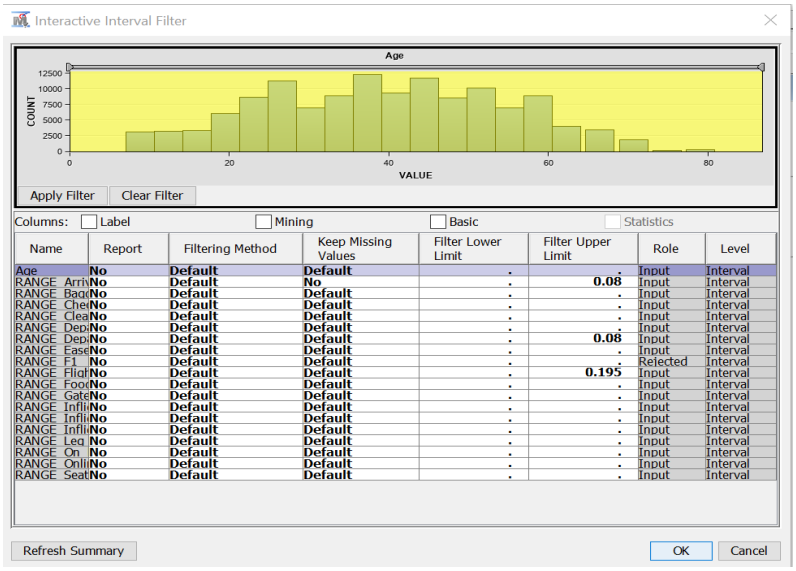
Furthermore, based on common sense and literature review; we know that passenger preferences would change the most with respect to flight distance. We decide to filter our dataset on the basis of flight distances & both of the new cohorts are further filtered to reduce skewness in delay minutes elsewise it would dominate clustering.



Cohort 1: Flighter equal or lesser than 1000 miles. - 73168 instances



Cohort 2: Flights greater than 1000 miles - 56319 instances

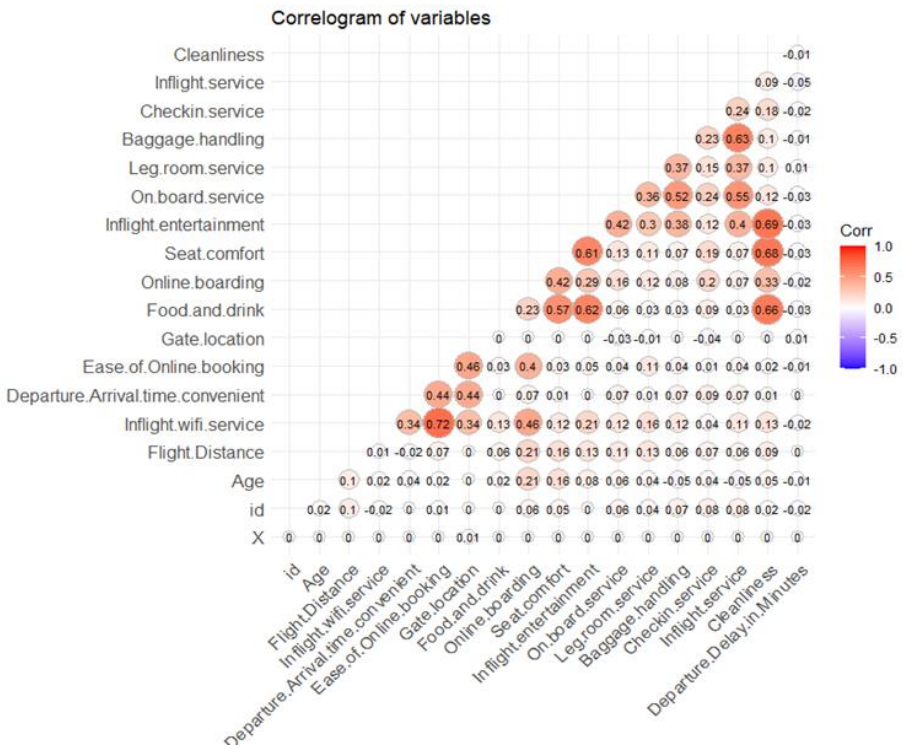


Name	Report	Filtering Method	Keep Missing Values	Filter Lower Limit	Filter Upper Limit	Role	Level
Age	No	Default	Default	.	.	Input	Interval
RANGE_Arri	No	Default	No	.	0.08	Input	Interval
RANGE_Bag	No	Default	Default	.	.	Input	Interval
RANGE_Che	No	Default	Default	.	.	Input	Interval
RANGE_Clea	No	Default	Default	.	.	Input	Interval
RANGE_Dep	No	Default	Default	.	.	Input	Interval
RANGE_Dep	No	Default	Default	.	0.08	Input	Interval
RANGE_Ease	No	Default	Default	.	.	Input	Interval
RANGE_F1_	No	Default	Default	.	.	Rejected	Interval
RANGE_Flig	No	Default	Default	.	0.195	Input	Interval
RANGE_Food	No	Default	Default	.	.	Input	Interval
RANGE_Gate	No	Default	Default	.	.	Input	Interval
RANGE_Infli	No	Default	Default	.	.	Input	Interval
RANGE_Infli	No	Default	Default	.	.	Input	Interval
RANGE_Infli	No	Default	Default	.	.	Input	Interval
RANGE_Leg	No	Default	Default	.	.	Input	Interval
RANGE_On_	No	Default	Default	.	.	Input	Interval
RANGE_Onli	No	Default	Default	.	.	Input	Interval
RANGE_Seat	No	Default	Default	.	.	Input	Interval

6. **Data Exploration**

   a. A correlogram is built to give relation between all the quantitative variables with one another. Categorical variables are further compared using two-way tables and chi square tests.

   b. **Correlation between quantitative variables**: Correlogram is used to study the correlation between all the quantitative variables of the dataset. Here red represents positive correlations (one variable increases with increase in another) and blue represents negative ones (one variable increases with decrease in another).The white ones represent that there is no linear relationship between the two variables considered.



Here we can see that Inflight.wifi.service and Ease.of.Online.booking are highly positively correlated. Same is the case with Food.and.Drink and Inflight.Entertainment. While Online.boarding and seat.comfort are slightly positively correlated and hence are represented by a lighter shade of red. Now the box with Age and Food.and.Drinks is in white. It means that although the correlation coefficient is very small, we cannot reject the hypothesis of no correlation and hence white. Similarly the relationship between all the various quantitative variables can be analyzed by this correlogram.

c. **Correlation between qualitative variables:** All permutations and combinations are tried between the qualitative variables. A total of 15 combinations were studied. A few examples are noted down below:

1) Travel Class vs Satisfaction: Business class is more satisfied while eco and eco plus are more dissatisfied.

```
> two_way = table(dat$Class,dat$satisfaction)
> two_way

          neutral or dissatisfied satisfied
  Business                   15185     34480
  Eco                        38044      8701
  Eco Plus                    5650      1844
```

2) Satisfaction Vs Customer.Type: Non-members seem to be more dissatisfied whereas members have a close call between being satisfied and dissatisfied.

```
> two_way = table(dat$satisfaction,dat$Customer.Type)
> two_way

                        disloyal Customer Loyal Customer
  neutral or dissatisfied             14489          44390
  satisfied                            4492          40533
```

3) Satisfaction vs (Type.of.Travel + Customer.Type): A member doing personal travel seems to be more dissatisfied. A member traveling for the purpose of business seems to be more satisfied.

```
                        Business traveldisloyal Customer
  neutral or dissatisfied                            14351
  satisfied                                           4466

                        Business travelLoyal Customer
  neutral or dissatisfied                         15558
  satisfied                                       37280

                        Personal Traveldisloyal Customer
  neutral or dissatisfied                              138
  satisfied                                             26

                        Personal TravelLoyal Customer
  neutral or dissatisfied                         28832
  satisfied                                        3253
```

4) Satisfaction vs (Class + Customer.Type): A member travelling by business class seems to be more satisfied.

	Businessdisloyal Customer	BusinessLoyal Customer
neutral or dissatisfied	4447	10738
satisfied	2909	31571

	Eco Plusdisloyal Customer	Eco PlusLoyal Customer
neutral or dissatisfied	659	4991
satisfied	56	1788

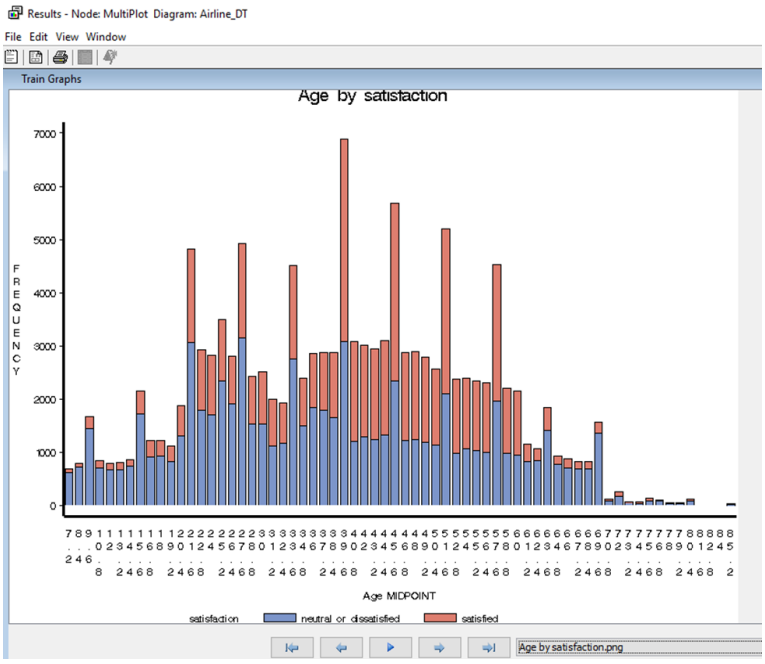
	Ecodisloyal Customer	EcoLoyal Customer
neutral or dissatisfied	9383	28661
satisfied	1527	7174

d. **Correlation between qualitative and quantitative variables:**

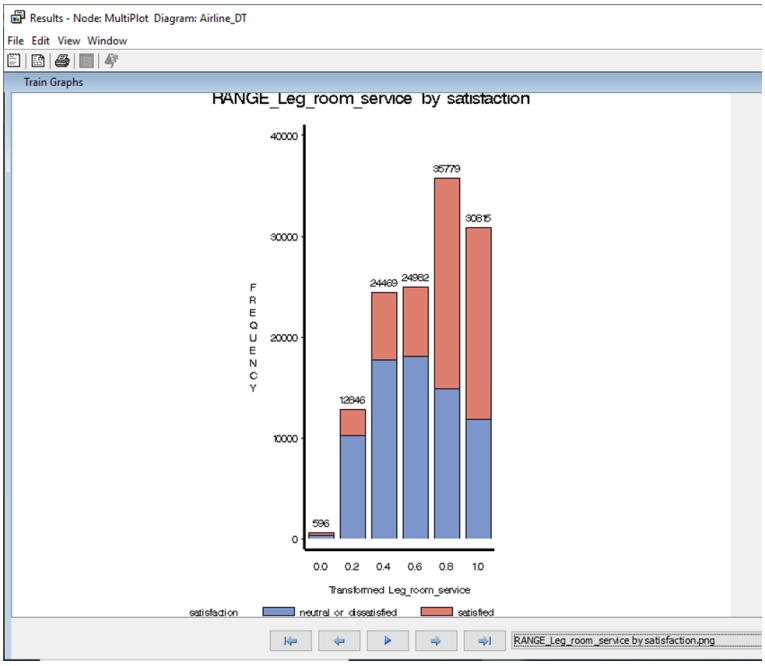
Multiplot node is used to graphically explore wide volumes of data, observe data distribution and examine the relationship amongst each of the independent variables and our target variables each (Satisfaction and Member). So a total of 18 (each quantitative against Satisfaction) and more 18 (each quantitative against Member) scatter plots and bar graphs were generated. Out of the 36 graphs, a few are mentioned below to give a clearer understanding. The range sheet on the basis of which the calculations were done is as follows:

RANGE Arrival Delay in Minutes	(Arrival Delay in Minutes - 0) / (128-0)
RANGE Baggage handling	(Baggage handling - 1) / (5-1)
RANGE Checkin service	(Checkin service - 1) / (5-1)
RANGE Cleanliness	(Cleanliness - 0) / (5-0)
RANGE Departure Arrival time con	(Departure Arrival time convenien - 0) / (5...
RANGE Departure Delay in Minutes	(Departure Delay in Minutes - 0) / (130-0)
RANGE Ease of Online booking	(Ease of Online booking - 0) / (5-0)
RANGE Flight Distance	(Flight Distance - 31) / (1000-31)
RANGE Food and drink	(Food and drink - 0) / (5-0)
RANGE Gate location	(Gate location - 1) / (5-1)
RANGE Inflight entertainment	(Inflight entertainment - 0) / (5-0)
RANGE Inflight service	(Inflight service - 0) / (5-0)
RANGE Inflight wifi service	(Inflight wifi service - 0) / (5-0)
RANGE Leg room service	(Leg room service - 0) / (5-0)
RANGE On board service	(On board service - 0) / (5-0)
RANGE Online boarding	(Online boarding - 0) / (5-0)
RANGE Seat comfort	(Seat comfort - 1) / (5-1)

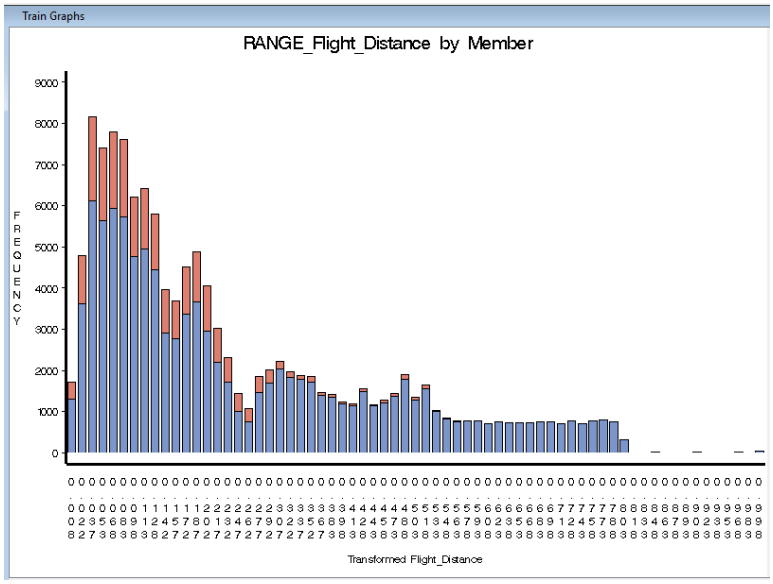
1. Age vs Satisfaction: The customers aged between 38-60 seem to be more satisfied than dissatisfied. Ages above 60 and below 38 seem to be more dissatisfied.



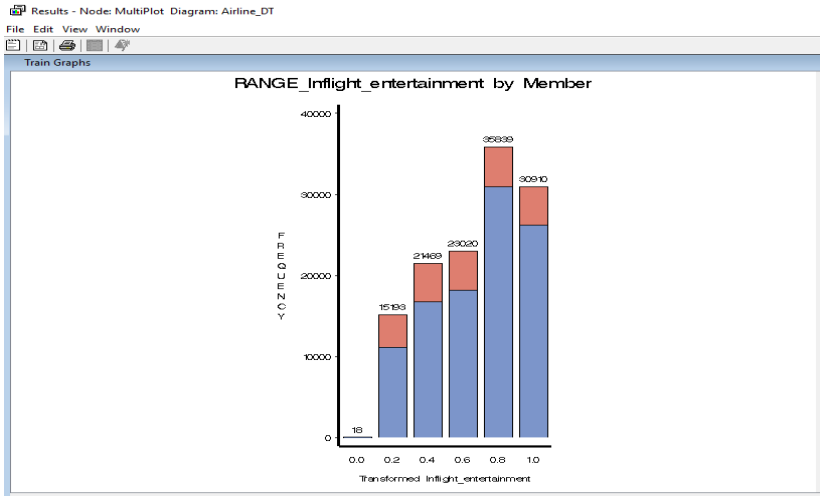
2. Leg Room service vs Satisfaction: At rating 0, the customers seem to be equally satisfied or dissatisfied. Then till 3, they are dissatisfied and above that they are satisfied.



3. Flight distance vs member: Approximately till the distance of 2596, it looks like the customers are more members than non-members but there are a few non-members, however after that there seems to be only members and no non-members whatsoever.



4. Inflight_entertainment vs Member: For all scores from 0-5, there are more members than non-members with the customers having a score of 4 having the maximum number of members.



7. Method

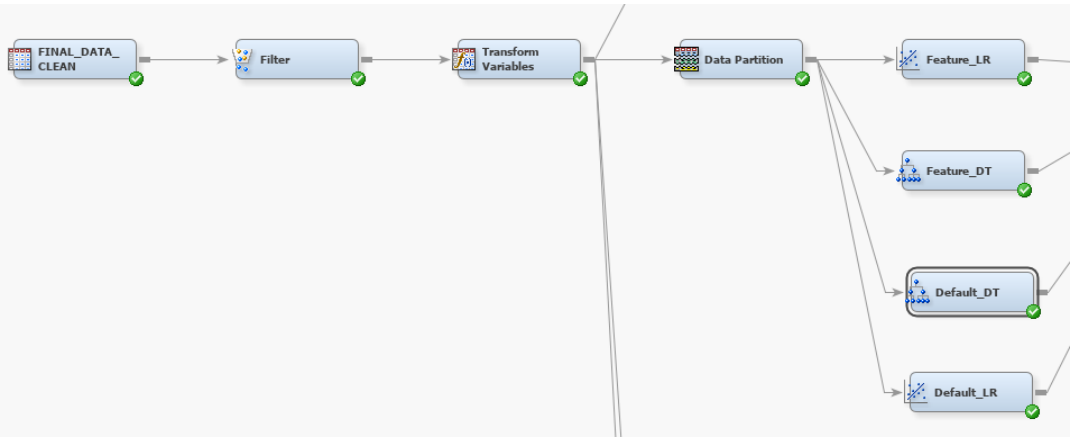
a. Identifying key factors determining Passenger's Satisfaction

We have a total of 23 variables giving us information about the passenger or their flight experience or the flight. We want to see which variables determine a satisfied customer experience. Our desirable model must have the least number of variables (< 23) and be accurate over 90% as our benchmarks from literature review. The target/output for this technique is Satisfaction which is binary (1 = Satisfied & 0 = Dissatisfied or neutral) and our focus is on satisfaction so we would deal neutral as dissatisfied.

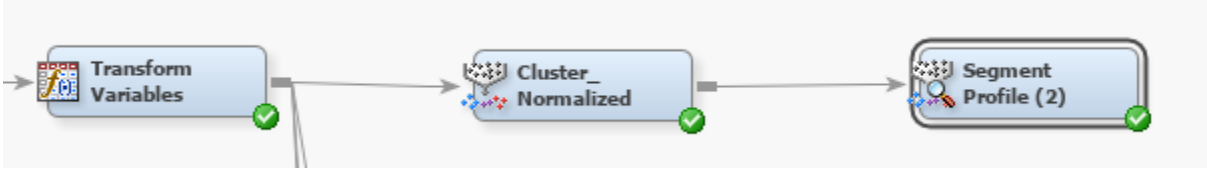
We divided our problem in two cohorts so each would be modelled separately for interpretation and deductions.

Cohort 1:

We went ahead and created Decision Tree (Default_DT) and Logistic Regression(Default_LR) models using all 23 variables to get a baseline model.



And then in order to identify only relevant features we performed a cluster analysis using the automatic K-Means algorithm in SAS Miner.



Mean Statistics

Clustering Criterion	Maximum Relative Change in Cluster Seeds	Improvement in Clustering Criterion	Segment Id	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	Age	Transformed Arrival_Delay_in_Minutes	Transformed Baggage_handling	Transformed Checkin_service	Transformed Cleanliness	Transformed Departure_Arrival_time_convenien	Transformed Departure_Delay_in_Minutes	Transformed Ease_of_Online_booking	Transformed Flight_Distance	Transformed Food_and_drink	Transformed Gate_location	Transformed Inflight_entertainment	Transformed Inflight_service	Transformed Inflight_wifi_service	Transformed Leg_room_service	Transformed On_board_service
0.796329	0.005102	.	1	19997	0.808062	6.828611	4	3.086575	42.37226	0.045789	0.376231	0.43159	0.704286	0.562124	0.040635	0.501205	0.465989	0.711967	0.48211	0.627064	0.501925	0.505356	0.498745	0.456428
0.796329	0.005102	.	2	20352	0.791653	6.604886	1	3.228989	32.51602	0.048799	0.692635	0.555498	0.355611	0.603479	0.043894	0.491352	0.492984	0.355739	0.492151	0.375256	0.758029	0.466853	0.649607	0.657134
0.796329	0.005102	.	3	5785	0.89211	6.711991	1	4.621871	37.97908	0.560446	0.630467	0.531158	0.634641	0.608781	0.544291	0.529231	0.490624	0.6307	0.497969	0.644909	0.704235	0.52847	0.62548	0.633708
0.796329	0.005102	.	4	27034	0.768919	6.420397	1	3.086575	40.27059	0.041033	0.810452	0.661842	0.805282	0.659066	0.038831	0.602833	0.487719	0.779744	0.503098	0.868972	0.855767	0.634061	0.749212	0.807679

We identified 15 Features contributing towards clustering our Cohort and these will be used later in our models. Variables like Gender, Departure Arrival time convenience, Ease of online booking, Flight distance, gate location, type of travel & Class did not significantly contribute to any of our 4 identified clusters. Arrival Time Delay & Departure Time Delay are significant variables to define a cluster with outliers (5785 cases - 4th Cluster).

Now based on our feature engineering, we use the following inputs to create our decision tree (Feature_DT) and Logistic regression (Feature_LR) with the target being ‘Satisfaction’.

Name	Use	Report	Role	Level
Age	Default	No	Input	Interval
Class	No	No	Input	Nominal
F1	No	No	Rejected	Interval
Gender	No	No	Input	Nominal
Member	Default	No	Input	Nominal
RANGE_Arrival_Delay_in_Minutes	Default	No	Input	Interval
RANGE_Baggage_handling	Default	No	Input	Interval
RANGE_Checkin_service	Default	No	Input	Interval
RANGE_Cleanliness	Default	No	Input	Interval
RANGE_Departure_Arrival_time_con	No	No	Input	Interval
RANGE_Departure_Delay_in_Minutes	Default	No	Input	Interval
RANGE_Ease_of_Online_booking	No	No	Input	Interval
RANGE_Flight_Distance	No	No	Input	Interval
RANGE_Food_and_drink	Default	No	Input	Interval
RANGE_Gate_location	No	No	Input	Interval
RANGE_Inflight_entertainment	Default	No	Input	Interval
RANGE_Inflight_service	Default	No	Input	Interval
RANGE_Inflight_wifi_service	Default	No	Input	Interval
RANGE_Leg_room_service	Default	No	Input	Interval
RANGE_On_board_service	Default	No	Input	Interval
RANGE_Online_boarding	Default	No	Input	Interval
RANGE_Seat_comfort	Default	No	Input	Interval
Type_of_Travel	No	No	Input	Nominal
satisfaction	Yes	No	Target	Nominal

Cohort 2:

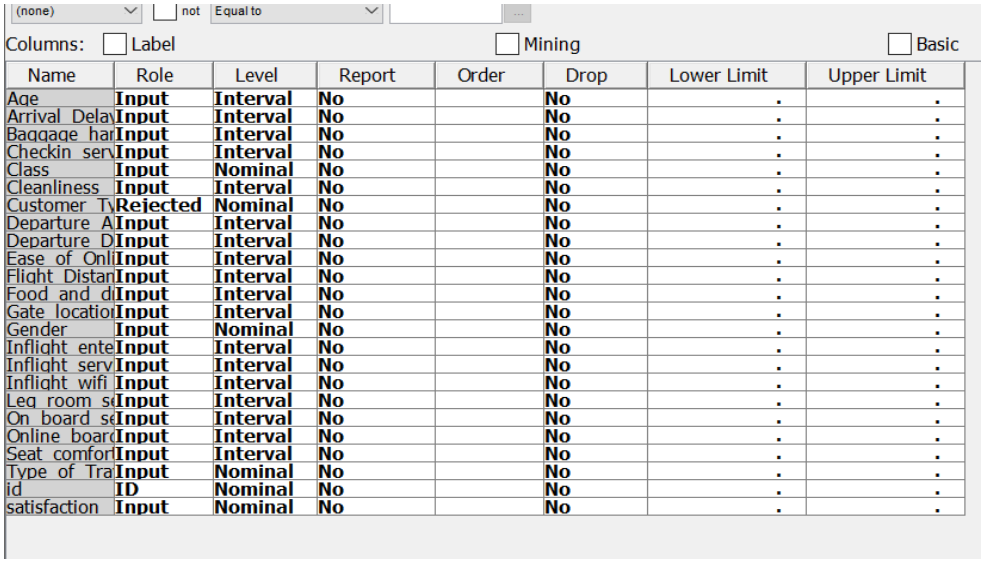
Same procedure from cohort 1 was followed while dealing with the second cohort (flights over > 1000mi). In total 4 models were created; out of which one logistic regression and decision tree were using all the features and the other version of these models were using selective features identified from clustering.

We identified 17 Features contributing towards clustering of our Cohort 2 and which were used later in our models. Majority of the features remained the same we got for Cohort 1 except Flight distance and type of travel became significant contributors to our 4 identified clusters. Below are the inputs which we used to create our feature models.

Name	Use	Report	Role	Level
Age	Default	No	Input	Interval
Class	Default	No	Input	Nominal
F1	No	No	Rejected	Interval
Gender	No	No	Input	Nominal
Member	Default	No	Input	Nominal
RANGE_Arrival_Delay_in_Minutes	Default	No	Input	Interval
RANGE_Baggage_handling	Default	No	Input	Interval
RANGE_Checkin_service	Default	No	Input	Interval
RANGE_Cleanliness	Default	No	Input	Interval
RANGE_Departure_Arrival_time_con	No	No	Input	Interval
RANGE_Departure_Delay_in_Minutes	Default	No	Input	Interval
RANGE_Ease_of_Online_booking	No	No	Input	Interval
RANGE_Flight_Distance	Default	No	Input	Interval
RANGE_Food_and_drink	Default	No	Input	Interval
RANGE_Gate_location	No	No	Input	Interval
RANGE_Inflight_entertainment	Default	No	Input	Interval
RANGE_Inflight_service	Default	No	Input	Interval
RANGE_Inflight_wifi_service	No	No	Input	Interval
RANGE_Leg_room_service	Default	No	Input	Interval
RANGE_On_board_service	Default	No	Input	Interval
RANGE_Online_boarding	Default	No	Input	Interval
RANGE_Seat_comfort	Default	No	Input	Interval
Type_of_Travel	Default	No	Input	Nominal
dataobs		No	ID	Interval
id		No	ID	Nominal
satisfaction	Yes	No	Target	Nominal

b. **Identifying Key Attributes of Members**

After data preprocessing and exploration we decided to find some trends based on the membership. We only want to identify underlying patterns in the member customers. We have 106100 passengers who are members of the airline and we use these to perform cluster analysis.



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Age	Input	Interval	No		No	.	.
Arrival_Delay	Input	Interval	No		No	.	.
Baggage_han	Input	Interval	No		No	.	.
Checkin_serv	Input	Interval	No		No	.	.
Class	Input	Nominal	No		No	.	.
Cleanliness	Input	Interval	No		No	.	.
Customer_Ty	Rejected	Nominal	No		No	.	.
Departure_A	Input	Interval	No		No	.	.
Departure_D	Input	Interval	No		No	.	.
Ease_of_Onli	Input	Interval	No		No	.	.
Flight_Distan	Input	Interval	No		No	.	.
Food_and_di	Input	Interval	No		No	.	.
Gate_locatio	Input	Interval	No		No	.	.
Gender	Input	Nominal	No		No	.	.
Inflight_ente	Input	Interval	No		No	.	.
Inflight_serv	Input	Interval	No		No	.	.
Inflight_wifi	Input	Interval	No		No	.	.
Leg_room_se	Input	Interval	No		No	.	.
On_board_se	Input	Interval	No		No	.	.
Online_board	Input	Interval	No		No	.	.
Seat_comfort	Input	Interval	No		No	.	.
Type_of_Tra	Input	Nominal	No		No	.	.
id	ID	Nominal	No		No	.	.
satisfaction	Input	Nominal	No		No	.	.

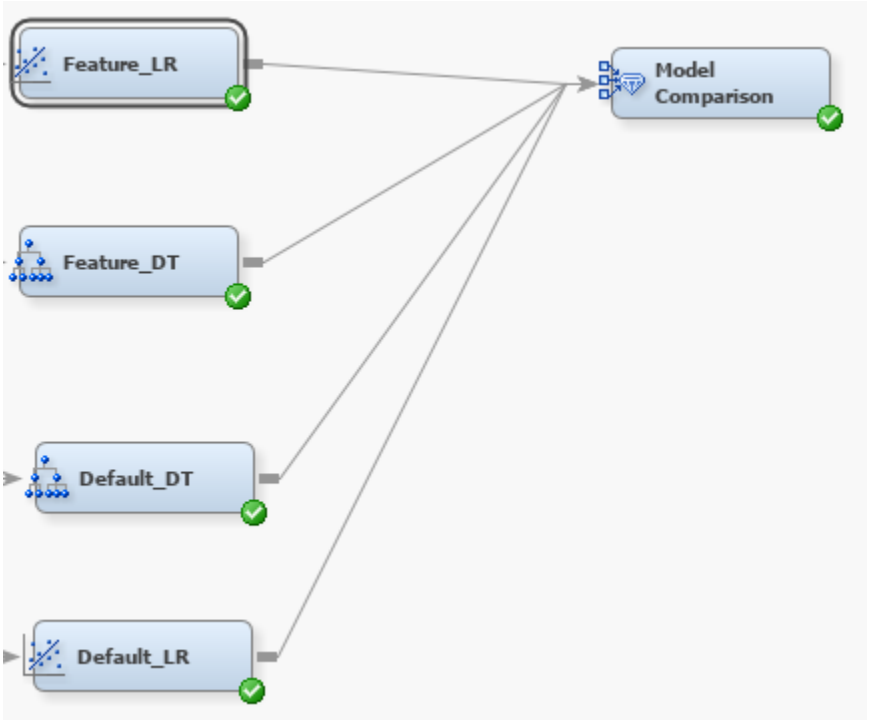
Method used to specify the maximum clusters was automatic. A clustering method used to determine the maximum number of clusters is Ward. And segment profiling is done for evaluating the results.

7. Evaluation and Results:

a. Identifying key factors determining Passenger's Satisfaction

Cohort 1

The models created in the above step were assessed using a model comparison node in SAS Miner and few classification evaluation metrics. 14636 random instances were used in validating these models.


Feature_LR
Model Comparison
Feature_DT
Default_DT
Default_LR

Data Role=Valid

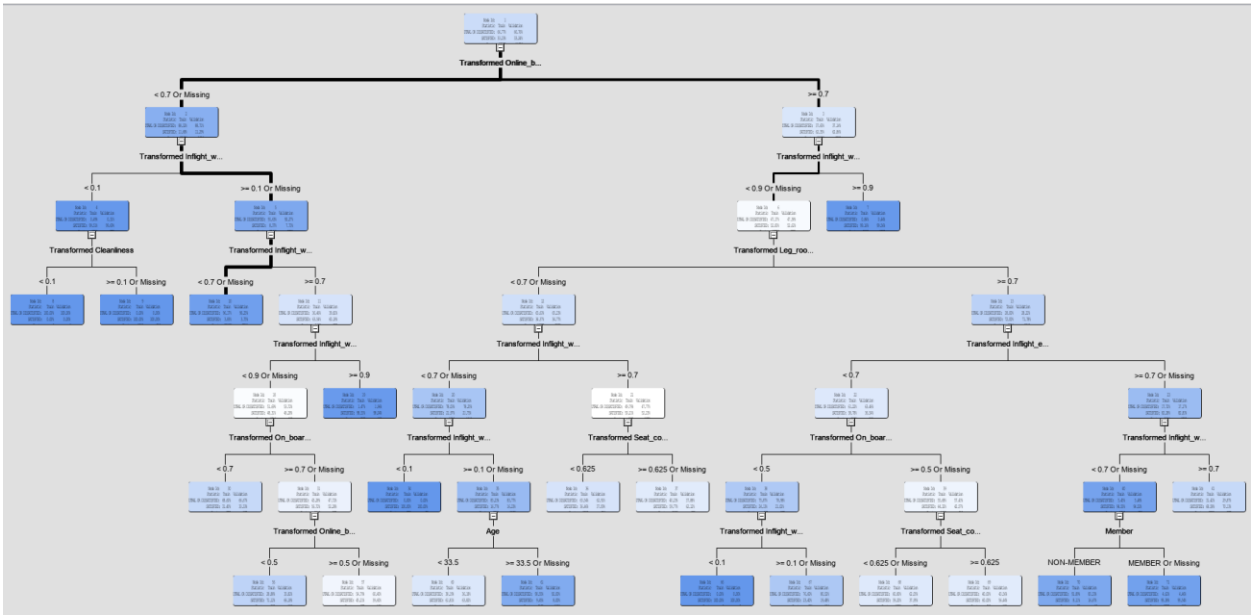
Statistics	Tree2	Tree	Reg	Reg2
Valid: Kolmogorov-Smirnov Statistic	0.83	0.78	0.69	0.58
Valid: Average Squared Error	0.06	0.08	0.10	0.14
Valid: Roc Index	0.97	0.94	0.90	0.84
Valid: Average Error Function	.	.	0.35	0.47
Valid: Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff	0.48	0.31	0.42	0.44
Valid: Cumulative Percent Captured Response	30.09	30.00	29.60	27.52
Valid: Percent Captured Response	15.05	14.98	14.64	13.28
Valid: Divisor for VASE	29272.00	29272.00	29272.00	29272.00
Valid: Error Function	.	.	10384.09	13666.88
Valid: Gain	200.84	199.90	195.91	175.16
Valid: Gini Coefficient	0.94	0.88	0.80	0.67
Valid: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	0.83	0.78	0.69	0.58
Valid: Kolmogorov-Smirnov Probability Cutoff	0.34	0.23	0.37	0.41
Valid: Cumulative Lift	3.01	3.00	2.96	2.75
Valid: Lift	3.01	3.00	2.93	2.65
Valid: Maximum Absolute Error	0.99	0.99	1.00	1.00
Valid: Misclassification Rate	0.07	0.10	0.13	0.18
Valid: Mean Square Error	.	.	0.10	0.14
Valid: Sum of Frequencies	14636.00	14636.00	14636.00	14636.00
Valid: Root Average Squared Error	0.24	0.27	0.32	0.38
Valid: Cumulative Percent Response	100.00	99.69	98.36	91.46
Valid: Percent Response	100.00	99.56	97.27	88.25
Valid: Root Mean Square Error	.	.	0.32	0.38
Valid: Sum of Square Errors	1644.60	2204.30	2957.29	4125.41
Valid: Sum of Case Weights Times Freq	.	.	29272.00	29272.00

Reg2: Default_LR Reg: Feature_LR
Tree2: Defualt_Tree Tree: Feature_Tree

Overall, the decision tree performs better than Regression for classifying the satisfaction class. The main reason behind this could be the dominant presence of variables which are either categorical or they are in range forms (0 till 5), hence the decision tree tends to perform better on such a dataset. Tree2 outperforms our Tree model with a slightly higher ROC index and the reason behind it is the number of predictors used by Tree2 are much more than Tree (Feature_DT). We finalised the Feature_DT model to carry on our study on the basis of minute difference in ROC

index from Tree2 and being more parsimonious as it has a lesser number of leaves as well.

Our final model looks like this;



We got a total of 20 leaves

	True Positive	True Negative
Predicted Positive	4083	715
Predicted Negative	782	9056

Measure	Value
Sensitivity/Recall	0.84
Specificity	0.93
Accuracy	0.90
F1 Score	0.84
Precision	0.85

Following are the major findings from the model regarding what factors drive customer satisfaction for flights within 1000 mi.

1. Excellent Online Boarding, Leg room and Inflight Wifi would most likely determine customer satisfaction.

2. For a member to be satisfied on shorter flights; inflight entertainment, online boarding, and legroom are most important factors for satisfied experience

3. One of the major causes of dissatisfaction is poor online boarding service.

Cohort 2:

We repeat the similar assessing techniques to finalise a model based on ROC index and simplicity as Cohort 1.

Data Role=Valid

Statistics	Tree2	Tree	Reg	Reg2
Valid: Kolmogorov-Smirnov Statistic	0.86	0.83	0.79	0.79
Valid: Average Squared Error	0.05	0.07	0.08	0.08
Valid: Roc Index	0.97	0.95	0.95	0.95
Valid: Average Error Function	.	.	0.29	0.29
Valid: Bin-Based Two-Way Kolmogorov-Smirnov Probability Cutoff	0.56	0.50	0.79	0.78
Valid: Cumulative Percent Captured Response	16.86	16.73	17.27	17.27
Valid: Percent Captured Response	8.35	8.35	8.63	8.63
Valid: Divisor for VASE	21320.00	21320.00	21320.00	21320.00
Valid: Error Function	.	.	6134.78	6133.46
Valid: Gain	68.62	67.26	72.69	72.69
Valid: Gini Coefficient	0.93	0.90	0.90	0.90
Valid: Bin-Based Two-Way Kolmogorov-Smirnov Statistic	0.86	0.82	0.78	0.79
Valid: Kolmogorov-Smirnov Probability Cutoff	0.37	0.34	0.73	0.71
Valid: Cumulative Lift	1.69	1.67	1.73	1.73
Valid: Lift	1.67	1.67	1.73	1.73
Valid: Maximum Absolute Error	1.00	1.00	1.00	1.00
Valid: Misclassification Rate	0.06	0.08	0.12	0.12
Valid: Mean Square Error	.	.	0.08	0.08
Valid: Sum of Frequencies	10660.00	10660.00	10660.00	10660.00
Valid: Root Average Squared Error	0.23	0.26	0.29	0.29
Valid: Cumulative Percent Response	97.65	96.86	100.00	100.00
Valid: Percent Response	96.74	96.74	100.00	100.00
Valid: Root Mean Square Error	.	.	0.29	0.29
Valid: Sum of Square Errors	1163.67	1462.57	1802.20	1808.49
Valid: Sum of Case Weights Times Freq	.	.	21320.00	21320.00

Reg2: Default_LR Reg: Feature_LR
Tree2: Defualt_Tree Tree: Feature_Tree

For Cohort 2, Tree2 outperforms our rest of our models based on the above model evaluation statistics. But, Tree 2 being a more complex model with only slightly better performance (ROC index), we can exclude this model from consideration. Reg2 can also be ignored as it is not only complex but it also has a lower ROC index. We can see that Tree(Feature_Tree) and Reg(Feature_LR) have equal ROC index hence we consider other classification metrics to opt our final model between the two for this cohort.

Feature_Tree			
Predicted	Actual		
		0	1
	0	4030	397
	1	457	5776

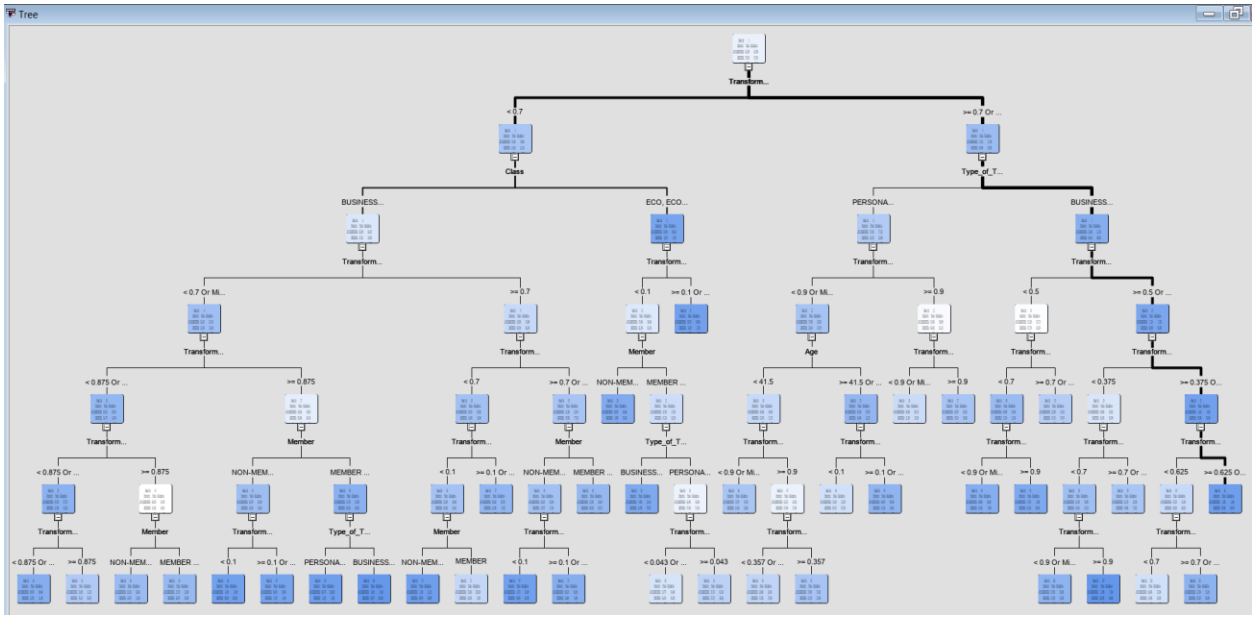
Feature_LR			
Predicted	Actual		
		0	1
	0	3838	597
	1	649	5776

10660 values used in validating models.

Measure	Feature_DT Value	Feature_LR Value
Sensitivity/Recall	0.94	0.90
Specificity	0.90	0.86
Accuracy	0.92	0.88
F1 Score	0.93	0.90
Precision	0.93	0.89

With respect to above model evaluation metrics, we can safely decide that the Feature_DT is most parsimonious and accurate in our all 4 models.

The selected model; Feature_DT has 34 leaves.



Following are the major findings from the model regarding what factors drive customer satisfaction for flights over 1000 mi.

1. A Non-Member travelling Business class would most likely be satisfied with smooth Check-in service.

2. A Member passenger travelling Business Class for their Business reasons would just care about Check in service to be satisfied.

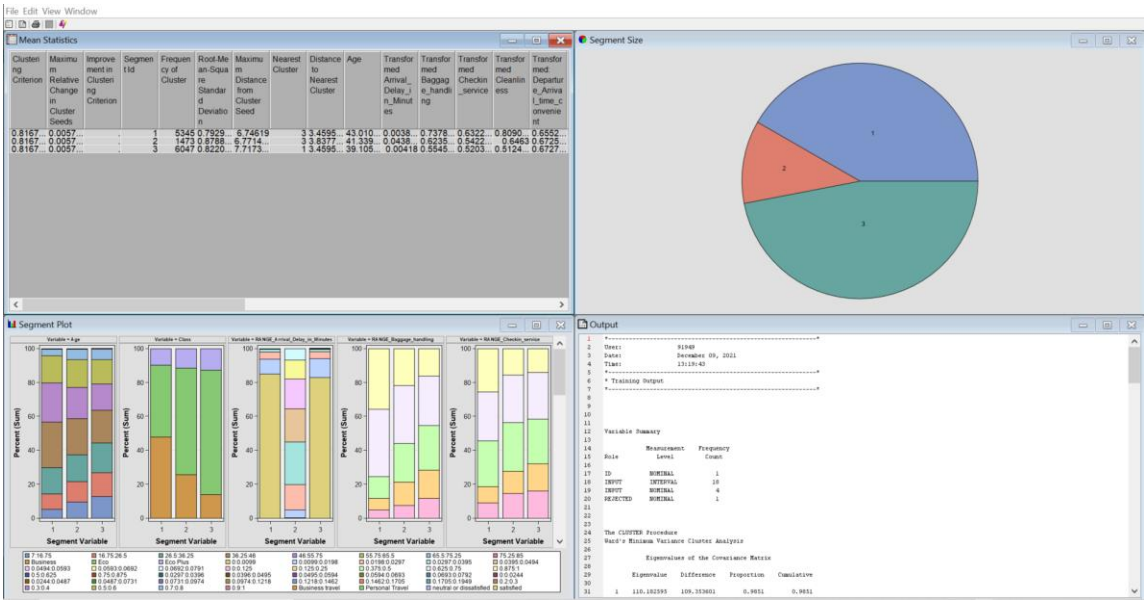
3. A Member passenger travelling Business Class would prefer pleasant inflight entertainment

4. Passengers(non-members as well members) travelling business class require; online boarding, check in service, seat comfort and cleanliness to be more than mediocre for a satisfied experience.

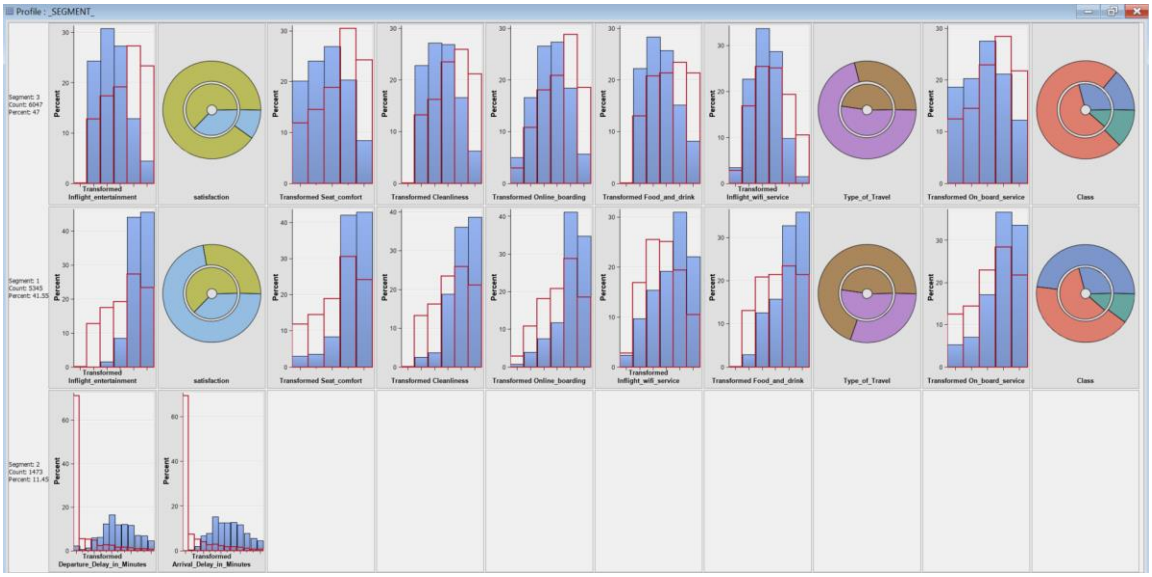
5. A Member passenger traveling for personal reasons in a business class wants more than pleasant checkin service, a decent Online_boarding and entertainment.

6. Passengers travelling in Eco or Eco plus class would be dissatisfied if you don't provide them with decent online services like online boarding and check-in.

b. **Identifying Key Attributes of Members**



After performing the cluster analysis we got 3 clusters. One cluster is formed because of skewness in the data due to arrival delay in minutes and departure delay in minutes (Segment 2). The other two clusters are representing membership passengers and their distinct behaviors.



Segment profiling indicates that Inflight entertainment services, seat comfort, cleanliness, online boarding, food and drinks and Inflight wifi services are the important parameters for members.

Segment 1: Cluster 1 has more satisfied customers than dissatisfied members. Segment profiling shows that these members are satisfied and they score Inflight entertainment services, seat comfort, cleanliness, online boarding, food and drinks and Inflight wifi services more than 3.

Segment 3: Cluster 3 has the most number of dissatisfied members. Segment profiling shows that these members are dissatisfied and they score the following variables less than the mean of the whole dataset; Inflight entertainment services, seat comfort, food and drinks and Inflight wifi services.

8. Discussion

a. Identifying key factors determining Passenger's Satisfaction

The airline must focus upon improving inflight entertainment, inflight wifi, online boarding and online check-in services irrespective of flight distance and travel class. Inflight entertainment can be improved by adding more audio and video streaming platforms and online services can be improved by introducing interactive mobile applications.

The major future advancement in this aspect of the project would be dividing the dataset into cohorts on the bases of Age groups or purpose of travel instead of just distance as we did in this project.

b. Identifying Key Attributes of Members

The findings from this analysis would help the airline in identifying their potential non-member passengers who can be converted into members if targeted effort is made. From our clustering results we found, all Business class travellers can most likely be converted to Members.

Moreover, any non-member passenger who is pleased with Inflight entertainment services, seat comfort, cleanliness, online boarding, food and drinks and Inflight wifi services also has high potential to be converted into a member.

Hence, the airline must target the non-member passengers having the above mentioned trait more specifically rather than targeting all

non-member passengers (targeted marketing to reduce expense and increase effectiveness).

Member Contribution:

1. Ashwami Dalvi: Conducted and wrote **Data preprocessing** and **Data Exploration**.

2. Simran Satyavolu: Wrote and presented; **Project introduction**, **Literature Review** and **Dataset description**.

3. Syed Muhammad Suffwan: Conducted and wrote the section: **Identifying key factors to determine Satisfaction** [Model, Results/Evaluation and Discussion]

4. Utwej Sai Nalluri: Conducted and wrote the section: **Identifying key attributes of a Member** [Model, Results/Evaluation and Discussion]

**Each member contributed the same parts in the presentation as well.*

** Feedback from the presentation is also incorporated in the final report.*

9. **References**

1.https://www.mastercardservices.com/en/expert-insights/future-airline-industry-innovating-customer-loyalty

2.Airline customer satisfaction and loyalty: impact of in-flight service quality https://link.springer.com/content/pdf/10.1007/s11628-009-0068-4.pdf

3.An, M., Noh, Y. Airline customer satisfaction and loyalty: impact of in-flight service quality. *Serv Bus* 3, 293–307 (2009). https://doi.org/10.1007/s11628-009-0068-4

4. Jin-Woo Park, Rodger Robertson, Cheng-Lung Wu, The effect of airline service quality on passengers' behavioural intentions: a Korean case study, Journal of Air Transport Management, Volume 10, Issue 6, 2004, Pages 435-439, ISSN 0969-6997, https://doi.org/10.1016/j.jairtraman.2004.06.001.

5. Clement Kong Wing Chow, Customer satisfaction and service quality in the Chinese airline industry, Journal of Air Transport Management, Volume 35, 2014, Pages 102-107, ISSN 0969-6997, https://doi.org/10.1016/j.jairtraman.2013.11.013.

6. B., K., The Influence Factors of Online Purchase on Customer Satisfaction in Mongolian Airlines, Vol 57, Issue 15, 2012, http://www.ipedr.com/vol57/015-ICBMG2012-B00031.pdf