Airplane Customer Segmentation

by Oki Samila Rici

OUTLINE

1. DATA BACKGROUND
2. DATA UNDERSTANDING
3. STATISTICAL SUMMARY
4. DATA PREPOCESSING
5. MODELING
6. INSIGHT AND RECOMMENDATION

1. DATA BACKGROUND

This dataset is about aircraft user data on airlines. Group customers by cluster using the K_Means Clustering method, to analyze how each cluster is different. So that it can be known which customers are loyal and which are not. And whether discounts facilitated by airlines can affect the frequency with which users choose this airline.

A. Question

- How many clusters will be formed?
- How to analyze the characteristics of each cluster?
- Does the discount given to have an impact on the customer?

B. Objective

- Create a model to form clusters to determine the performance of each cluster.
- Analyze each cluster to find the best customers.
- Analyze whether the discount facility provided by the airline has an impact on the number of customers.

D. Expected Outcomes

- Knowing the number of clusters formed.
- Can analyze the characteristics of each cluster.
- Knowing how effective the discount effect is on customer orders.

E. Data Dictionary

The dataset has 23 features and 62988 rows.

- MEMBER_NO-b : Member ID
- FFP_DATE : frequent flyer program joining date
- FIRST_FLIGHT_DATE : date of the first flight
- Gender: gender
- FFP_TIER : frequent flyer program level
- WORK_CITY : hometown
- WORK_PROVINSE : province of origin
- WORK_COUNTRY : country of origin
- AGE: customer age
- LOAD_TIME : data retrieval date
- FLIGHT_COUNT : number of customer flights
- BP_SUM : itinerary
- SUM_YR_1 : tariff revenue
- SUM_YR_2 : votes price
- SEG_KM_SUM : total distance (km) flights that have been made.
- LAST_FLIGHT_DATE : date of last flight
- LAST_TO_END : distance of the last flight time to the most recent last flight order.
- AVG_INTERVAL : average time interval
- MAX_INTERVAL : maximum time interval.
- EXCHANGE_COUNT : number of exchanges.
- avg_discount : the average discount that customers get.
- Points_Sum : the number of points earned by the customer.
- Point_NotFlight : points not used by members.

2. DATA UNDERSTANDING

This dataset has 23 features, 69288 rows, and no targets. So this modeling belongs to Machine Learning Unsupervised - Clustering.

A. Numeric Data

There are 15 numerical features among the 23 features.

B. Category Data

There are 8 feature categories, including 4 date type features, namely:
FFP_DATE, FIRST_FLIGHT_DATE, LOAD_TIME, LAST_FLIGHT_DATE.

3. STATISTICAL SUMMARY

A. Numeric Data

Based on median value:

- customer age : 41th
- number of flights: 7 times
- cumulative total distance covered: 9994 km

B. Category Data

Top categorical data is a categorical feature that has the highest frequency of occurrence.

- Gender : Male
- Work City : guangzhou
- Work Province : guangdong
- Work Country : CN

Guangzhou and Guangdong are cities and provinces in China, it can be assumed that this flight data comes from one of the airlines in China.

4. DATA PREPOCESSING

A. Missing Value

Six features have missing values and will drop rows because the percentage is less than 30%.

If more than 30% can handle:
* for numeric data, it can be filled with the mean value for data that is normally distributed and if the data is skewed, it can be filled with the median value.
* for categorical data, it can be filled with the value of the mode, which is the value that occurs frequently. or can be filled with a constant value by forming a new column with the name column is others. However, if the missing value is more than 60% it can drop the column.

The purpose of handling missing values is to make the data easier to analyze and the data to be more accurate as well as the machine learning model is to be made more powerful and without errors.

B. Feature Selection

In general, clustering uses RFM. But in this dataset, RFM will be modified to LRFMC.This is done to find out old customers or new customers for L and the use of discounts by customers for C.

L = LOAD_TIME - FFP_DATE (the end time of observation window - the time of membership)

- The number of months between the time of membership and the end of the observation window

R = LAST_TO_END (the time from the last flight to the end of the observation window)

- The number of months from the last time the customer took the company's aircraft to the end of the observation windows.

F = FLIGHT_COUNT (number of flights in the observation window)

- Number of times the customer takes the company's aircraft in the observation window.

M = SEG_KM_SUM (total flight kilometers of the observation window)

- Accumulated flight history of the customer in observation time.

C = AVG_DISCOUNT (average discount rate)

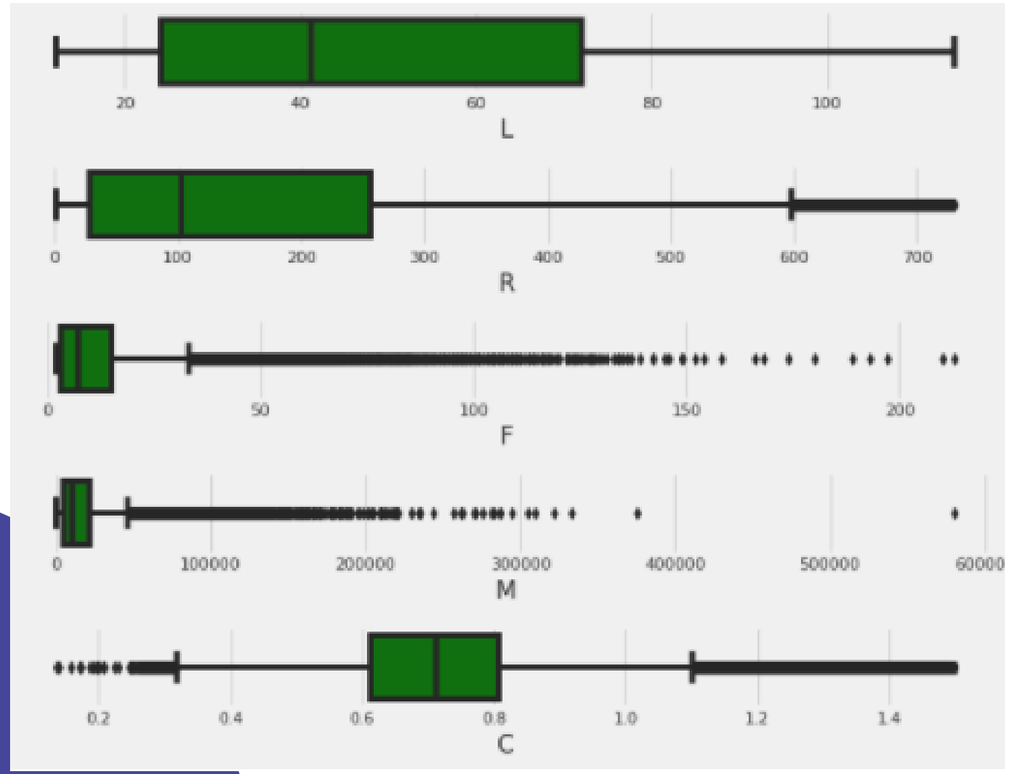
- Average value of the discount coefficient corresponding to the passenger space during the observation time.

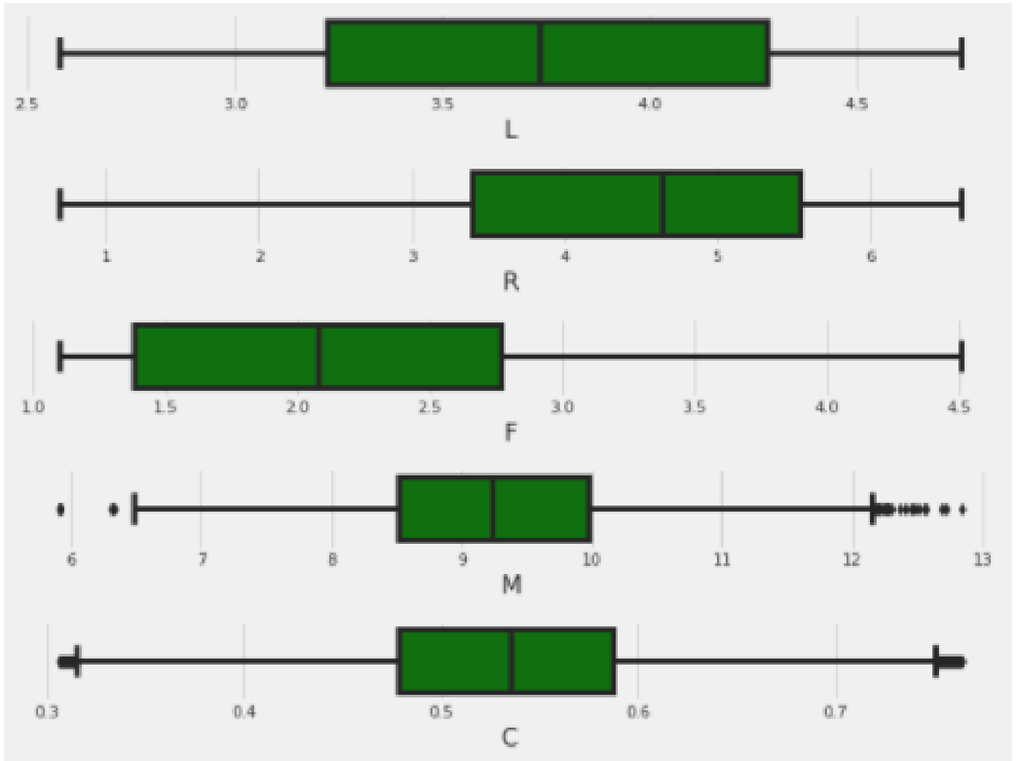
C. Duplicate Value

This dataset has 82 duplicate data removed

D. Outliers and Distribution Data

- There are many outliers in the LRFMC data, so it has a skewed distribution.
- Implementation of log transformation and removal based on IQR to reduce outliers, so that the data has a distribution that is close to normal.





E. Scaling

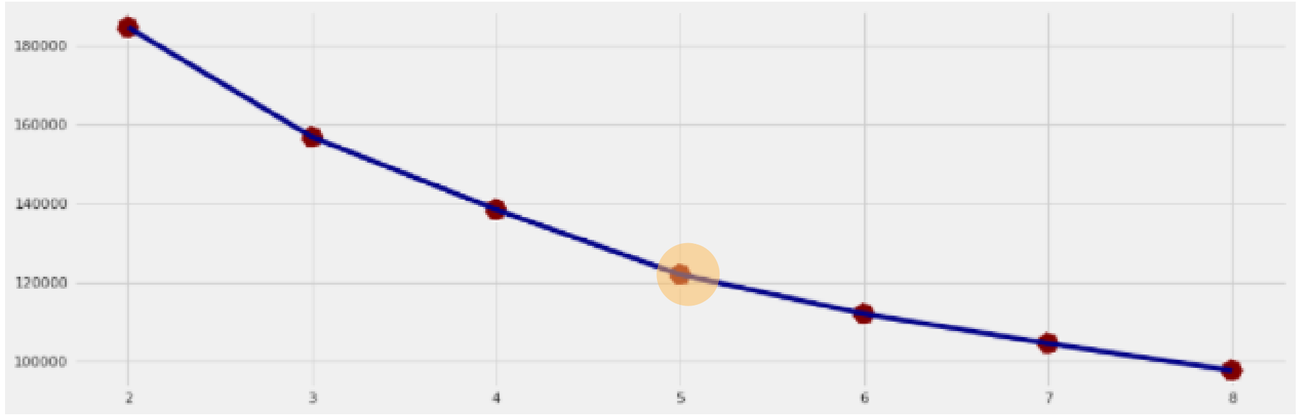
The goal of scaling is to equalize the scale for all values on features. As a result, using the same scale ensures that the learning algorithm treats all features fairly, speeds up the learning algorithm, and makes machine learning models easier to interpret.

The standard scaler is used in this dataset because it is normally distributed by changing the feature values so that the mean = 0 and standard deviation = 1. Meanwhile, normalization is for features that are not normally distributed.

	L	R	F	M	C
0	0.839516	0.138300	1.219848	3.307384	2.460742
1	0.278786	-2.057033	2.601830	3.328201	1.701094
2	0.178282	-1.105583	2.442673	3.596113	0.044262

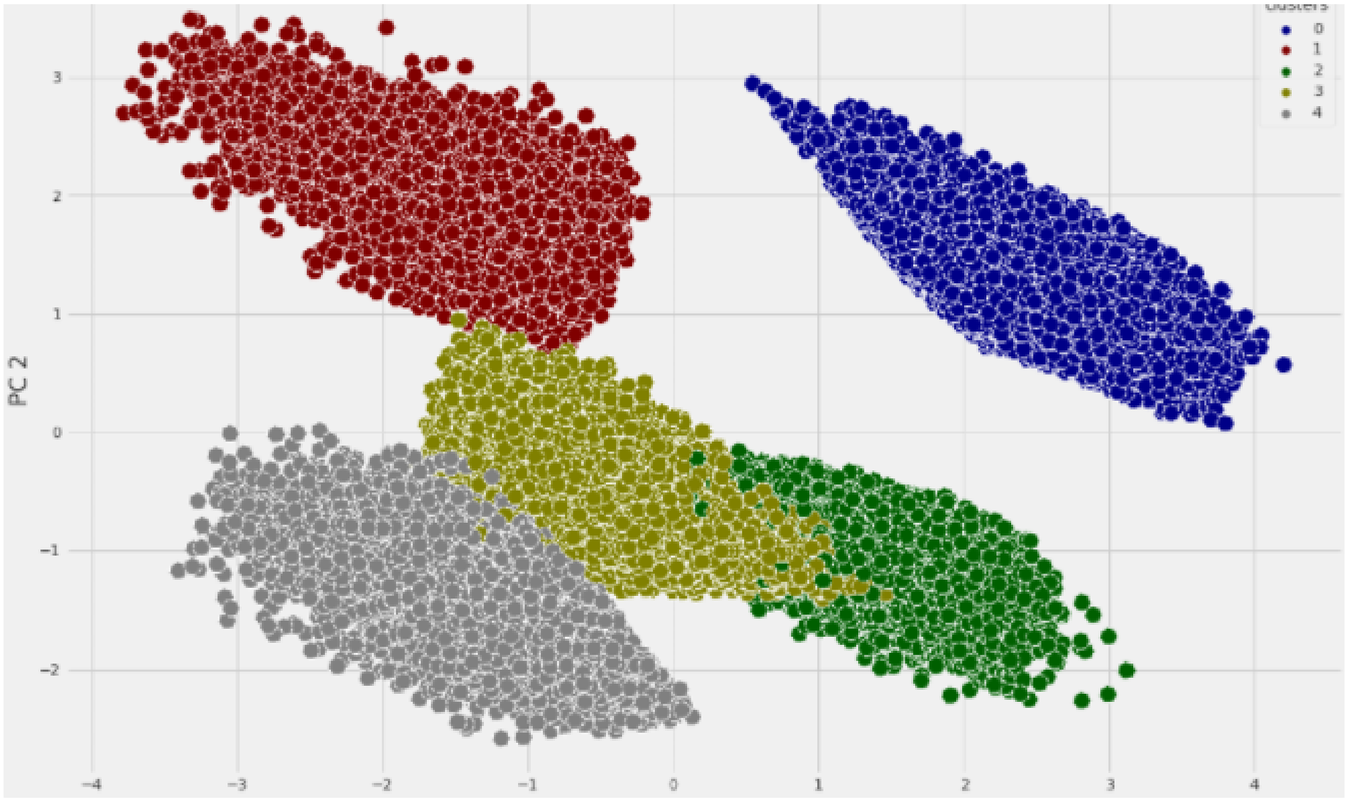
5. MODELING

A. Find The Best K



It can be seen that the value of each point in the elbow method have a similar elbow value, but the value of k=5 has an optimal elbow.

B. Clustering Visualization



This is the result of the visualization of 5 clusters, where in general cluster users are formed because there are users who have high or low values on the LRFMC feature.

C. Clustering Distribution

There are 5 clusters:

- Cluster 0 : 9.476 users
- Cluster 1 : 10.568 users
- Cluster 2 : 11.203 users
- Cluster 3 : 12.464 users
- Cluster 4 : 11.447 users

D. Cluster Characteristics and Analysis

Cluster 0 (Customer Group 1) :

Judging from feature L which is relatively low compared to the others, it can be assumed that this is a new user with a fairly frequent number of flights for a new user with a total flight distance that is not far, this can be seen in features F and M. This user has a feature value C which is comparable to the value of the F feature, meaning that this user uses promos quite often.

Cluster 1 (Customer Group 2) :

This cluster has a high value of feature L, so it is assumed that they are long-time users with the highest number of flights among others with the highest total flight distance, which can be seen in features F and M. Users also often use promos for their flights which can be seen on feature C.

	cluster	L	R	F	M	C
		median	median	median	median	median
0	0	29.400000	238.0	3.0	4807.0	0.535633
1	1	65.866667	15.0	27.0	36940.5	0.733125
2	2	32.600000	299.0	3.0	3924.0	0.824313
3	3	74.433333	112.0	8.0	10799.0	0.697163
4	4	23.400000	56.0	11.0	15935.0	0.705771

Cluster 2 (Customer Group 3) :

This is a new user that can be seen at low scores on feature L with few flights depicted on feature F. The total flight distance is also not far which can be seen on feature M. However, this user has the highest score on feature C among the others. . So, it can be assumed that even though they are new users with few flights, they are very active using the promo.

	cluster	L	R	F	M	C
		median	median	median	median	median
0	0	29.400000	238.0	3.0	4807.0	0.535633
1	1	65.866667	15.0	27.0	36940.5	0.733125
2	2	32.600000	299.0	3.0	3924.0	0.824313
3	3	74.433333	112.0	8.0	10799.0	0.697163
4	4	23.400000	56.0	11.0	15935.0	0.705771

Cluster 3 (Customer Group 4) :
The cluster that has the highest L feature value, means that it is the oldest user. Even though they are old users, they rarely make flights that can be seen in feature F. However, the value on feature M is quite high, so it is assumed that the total flight distance is long and they are also actively using promos that can be seen in feature C.

Cluster 4 (Customer Group 5) :
This group has the lowest value of feature L which can be assumed as the latest user with a high number of flights for the latest user size that can be seen in feature F. The total flight distance is also high which can be seen in feature M and the value on feature C is also high, it is assumed that This user is active in using the promo.

	cluster	L	R	F	M	C
		median	median	median	median	median
0	0	29.400000	238.0	3.0	4807.0	0.535633
1	1	65.866667	15.0	27.0	36940.5	0.733125
2	2	32.600000	299.0	3.0	3924.0	0.824313
3	3	74.433333	112.0	8.0	10799.0	0.697163
4	4	23.400000	56.0	11.0	15935.0	0.705771

6. Insight and Recommendation

A. Insight

There are 5 clusters:

- Cluster 0 : 9.476 users
- Cluster 1 : 10.568 users
- Cluster 2 : 11.203 users
- Cluster 3 : 12.464 users
- Cluster 4 : 11.447 users

Of the five clusters, the highest number of clusters is in cluster 3 with 12.464 users. Where they are the oldest users who rarely fly, but are active in using discounts.

The best clusters are clusters 1 and 4. Where cluster 1 is the old user who often makes flights and also often uses discounts. Cluster 4 is the newest user who makes frequent flights and also really takes advantage of the discount facility for flights.

B. Recommendation

After analyzing the 5 clusters, it can be seen that each cluster is active in using discounts. So it can be concluded that the discount facilities provided by airlines are effective for customers. The airline can maintain or increase the discount facility that can be an option for customers to choose this airline. In addition to discount facilities, airlines can provide other facilities that can attract customers to choose this airline.

EMAIL ADDRESS
oki.samila3@ymail.com

PHONE NUMBER
(+62) 823 8840 3307

LINKEDIN
www.linkedin.com/in/oki-samila

THANK YOU!