Unsupervised Learning (K-Means Clustering)

ALDY BUDHI ISKANDAR #DataScienceEnthusiast

aldybudhi003@gmail.com

https://www.linkedin.com/in/aldybudhi/

https://github.com/qodym

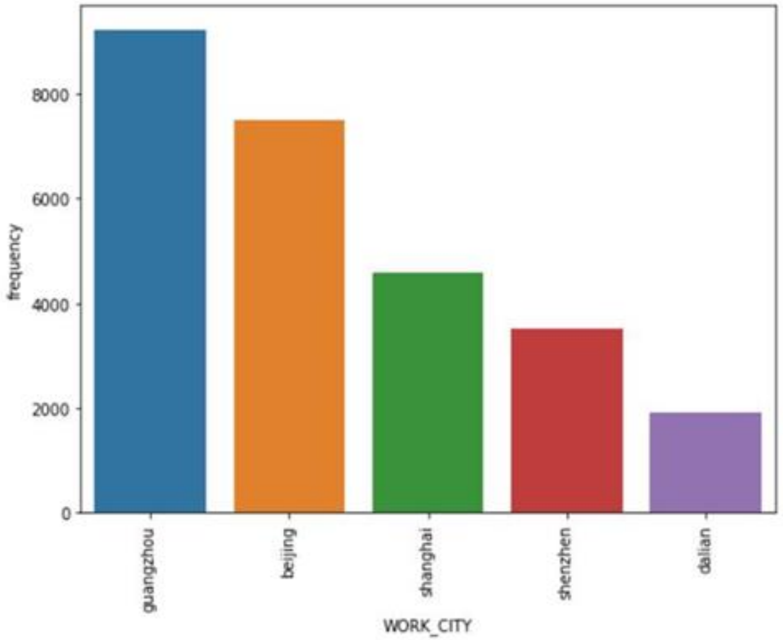
Data Understanding

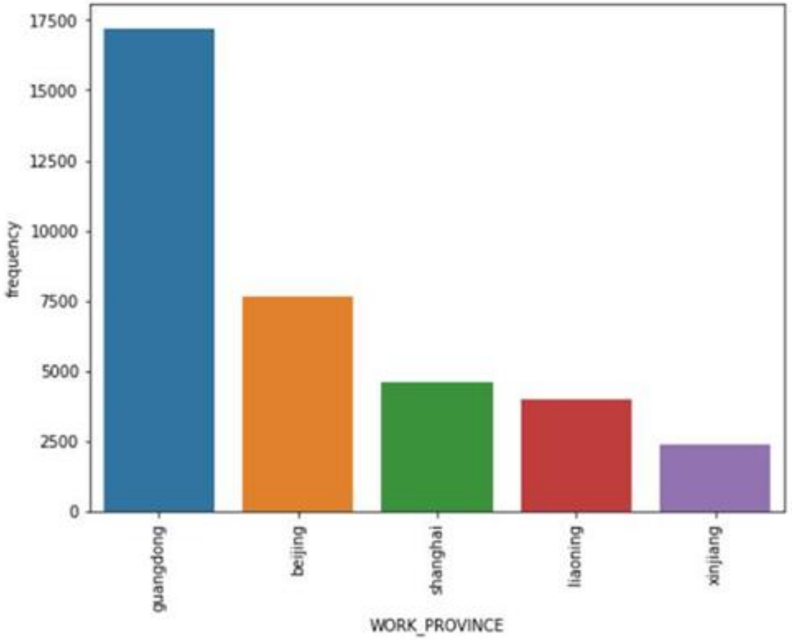
- DESCRIPTIVE STATISTICS
- MATRIX CORRELATION

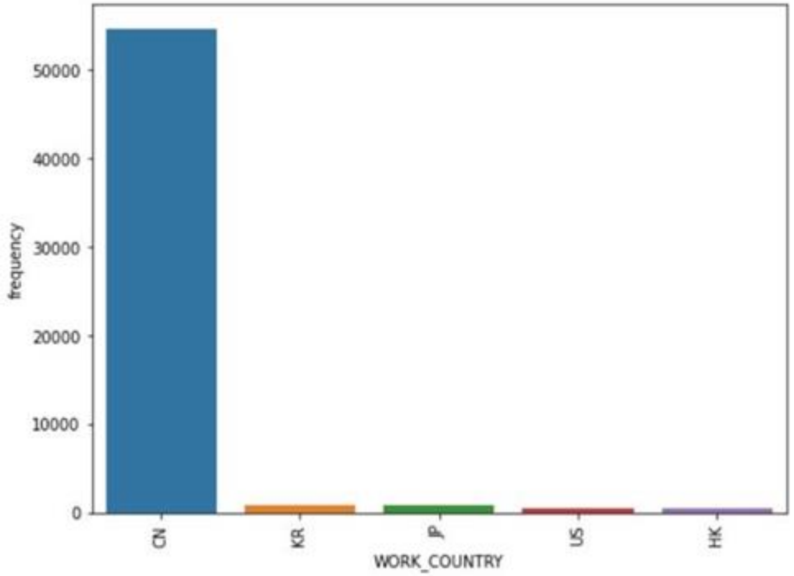
Descriptive Statistics

Top categorical data:

- Based on gender is **Male customer.**
- Based on customers domicile is comes from **Guangzhou, Province of Guangdong, China**.







aldybudhi003gmail.com https://github.com/qodym https://www.linkedin.com/in/aldybudhi/

Descriptive Statistics

Top numerical data:

- The median age of customers is **41 years**.
- The median total flight per customer is **7 flights**.
- The median total distance covered is **9994 Km**.

Median is the middle value of the data.

	AGE	FLIGHT_COUNT	SEG_KM_SUM
count	62568.000000	62988.000000	62988.000000
mean	42.476346	11.839414	17123.878691
std	9.885915	14.049471	20960.844623
min	6.000000	2.000000	368.000000
25%	35.000000	3.000000	4747.000000
50%	41.000000	7.000000	9994.000000
75%	48.000000	15.000000	21271.250000
max	110.000000	213.000000	580717.000000

Descriptive Statistics

Top datetime data:

- Most customers join date, which is on **January 13, 2011**.
- The first flight with the most customers on **February 16, 2013**.

Top frequent value

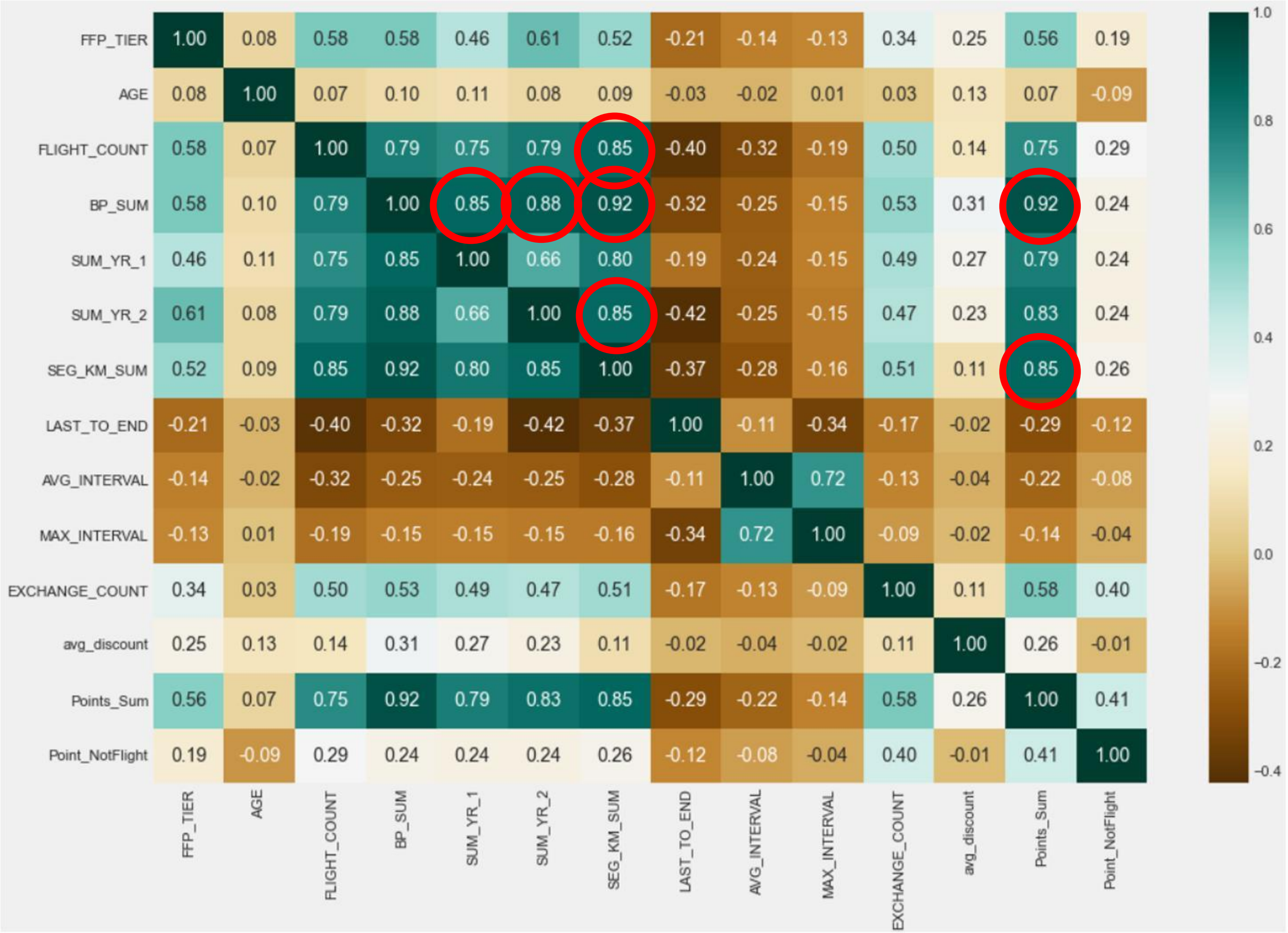
	FFP_DATE	FIRST_FLIGHT_DATE	LAST_FLIGHT_DATE	LOAD_TIME
count	62988	62988	62988	62988
unique	3068	3406	731	1
top	1/13/2011	2/16/2013	3/31/2014	3/31/2014
freq	184	96	959	62988

aldybudhi003gmail.com https://github.com/qodym https://www.linkedin.com/in/aldybudhi/

Matrix Correlation

There are features that have a high correlation with other features, including:

- FLIGHT_COUNT
- BP_SUM
- SUM_YR_1
- SUM_YR_2
- SEG_KM_SUM
- Points_Sum



aldybudhi003gmail.com https://github.com/qodym https://www.linkedin.com/in/aldybudhi/

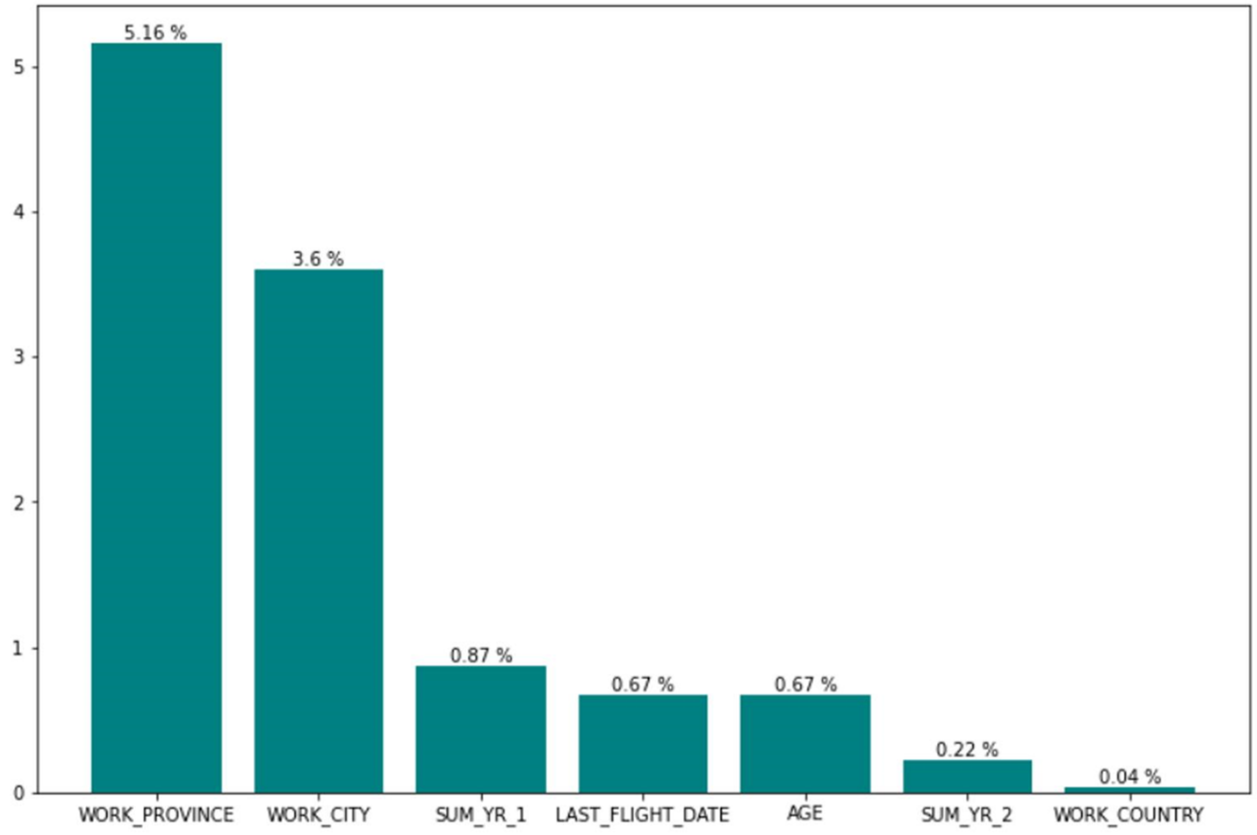
Data Preparation

- MISSING VALUE
- LRMFC FEATURE

Missing Value

Delete rows that have missing values

Because of the small number of missing values.



	feature	missing_value	percentage
0	WORK_PROVINCE	3248	5.16
1	WORK_CITY	2269	3.60
2	SUM_YR_1	551	0.87
3	LAST_FLIGHT_DATE	421	0.67
4	AGE	420	0.67
5	SUM_YR_2	138	0.22
6	WORK_COUNTRY	26	0.04

LRMFC Feature

* **L (Length Relation)** = `LOAD_TIME` - `FFP_DATE`
The number of months since the member ’s membership time from the end of the observation window = end time of the observation window-time to join (unit: month).

* **R (Recency)** = `LAST_TO_END`
The number of months since the customer ’s most recent flight to the end of the observation window = the time from the last flight to the end of the observation window (Unit: month).

* **F (Frequency)** = `FLIGHT_COUNT`
The number of times the customer took the company aircraft in the observation window = the number of flights in the observation window (unit: times].

* **M (Monetary Value)** = `SEG_KM_SUM`
The accumulated mileage of the customer in the company during the observation period = the total number of flight kilometers in the observation window (unit: km).

* **C (Coefficient Value)** = `AVG_DISCOUNT`
The average value of the discount coefficients corresponding to the passengers who traveled during the observation period = average discount rate (unit: none).

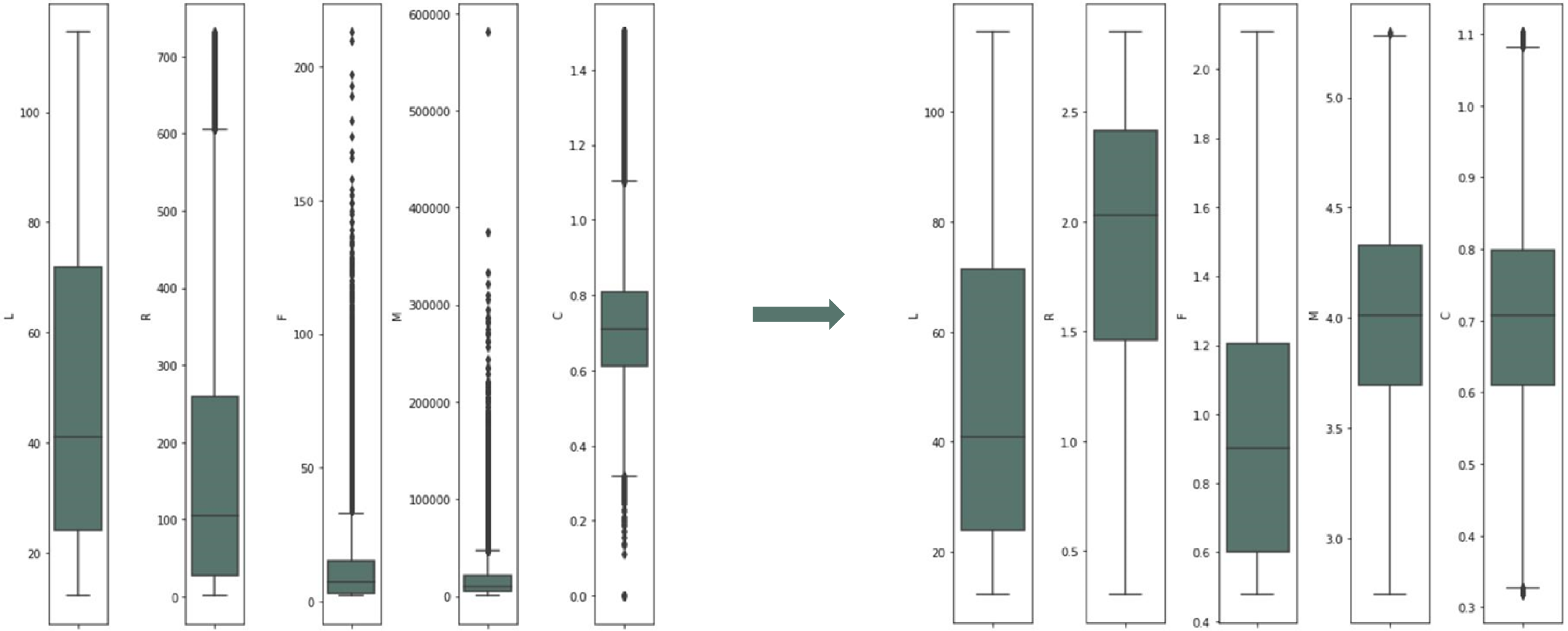
Outlier Handling

- DATA DISTRIBUTION
- DUPLICATE DATA
- SCALING

Data Distribution

There are outliers in numerical data, and it has a skewed distribution.

Implementation of log transformation reduces outliers and removes based on IQR on the data, and it has a distribution that is close to normal.



aldybudhi003gmail.com https://github.com/qodym https://www.linkedin.com/in/aldybudhi/

Duplicate Value

There are 43 duplicate values with the LMRFC feature, in this case all duplicate values are deleted.

Check duplicate data after transformation

```
[195] df_IQR_LRFMC.duplicated().sum()
```

```
43
```

There is **43** duplicate data, just drop it.

```
[196] df_IQR_LRFMC = df_IQR_LRFMC.drop_duplicates().reset_index(drop=True)
      df_IQR_LRFMC.duplicated().sum()
```

```
0
```

There isn't duplicated data.

Scaling

In this case the scaling uses the **StandardScaler** method.

Before scaling

	L	R	F	M	C
0	90.200000	1	210	580717	0.961639
1	87.166667	11	135	283712	1.254676
2	68.233333	97	23	281336	1.090870
3	60.533333	5	152	309928	0.970658
4	74.700000	79	92	294585	0.967692

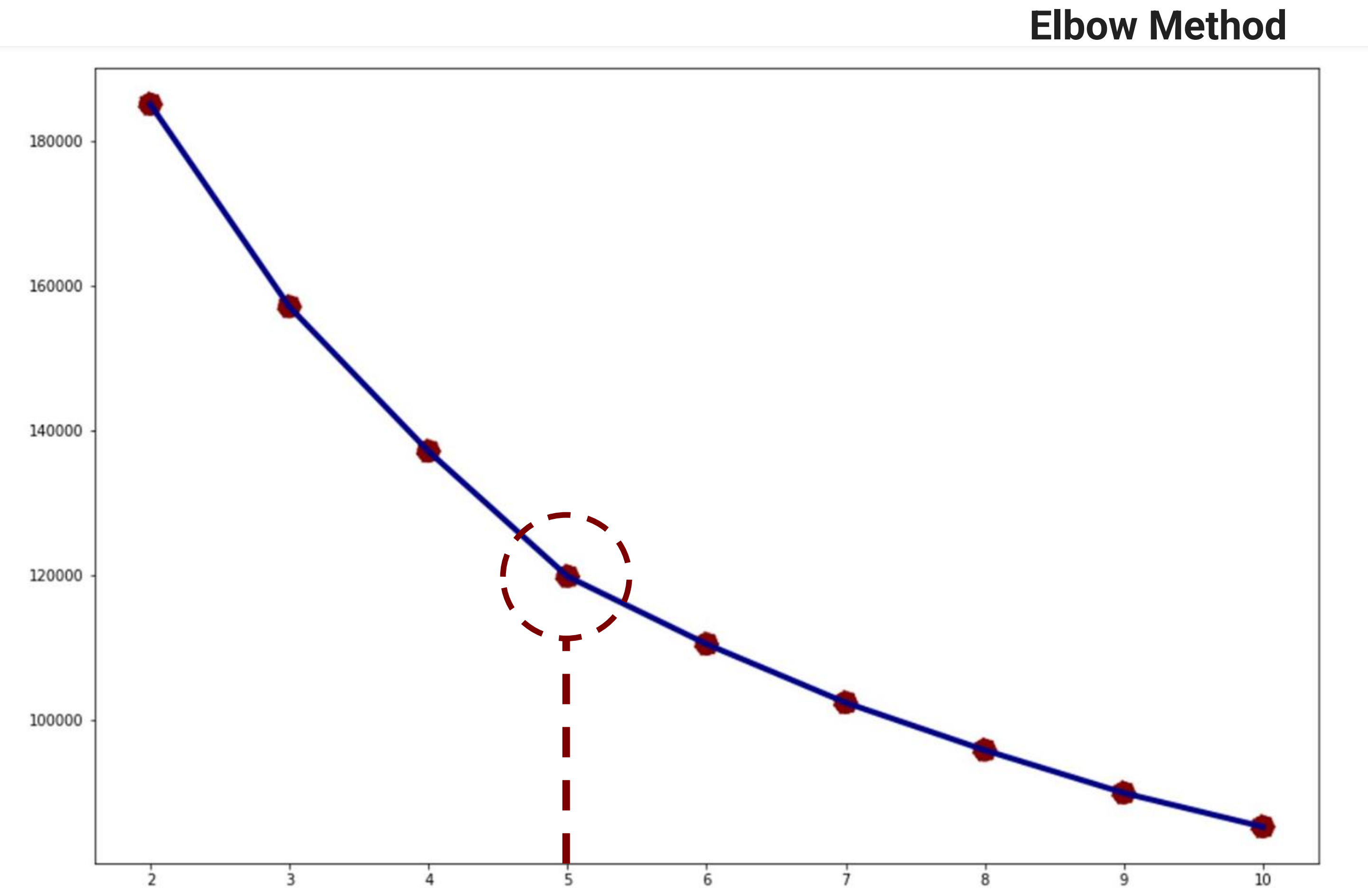
After scaling

	L	R	F	M	C
0	2.244319	-0.572552	1.262283	2.905819	2.679589
1	-0.989378	-2.529205	3.071763	2.894036	2.182716
2	-0.075533	-1.672793	2.747803	2.889941	1.639074
3	-0.061198	-1.500990	1.523403	2.759751	2.470614
4	1.043778	0.065461	1.599295	2.866901	1.685239

Modeling
K-Means Clustering

- CLUSTERING

Clustering (Elbow Method)

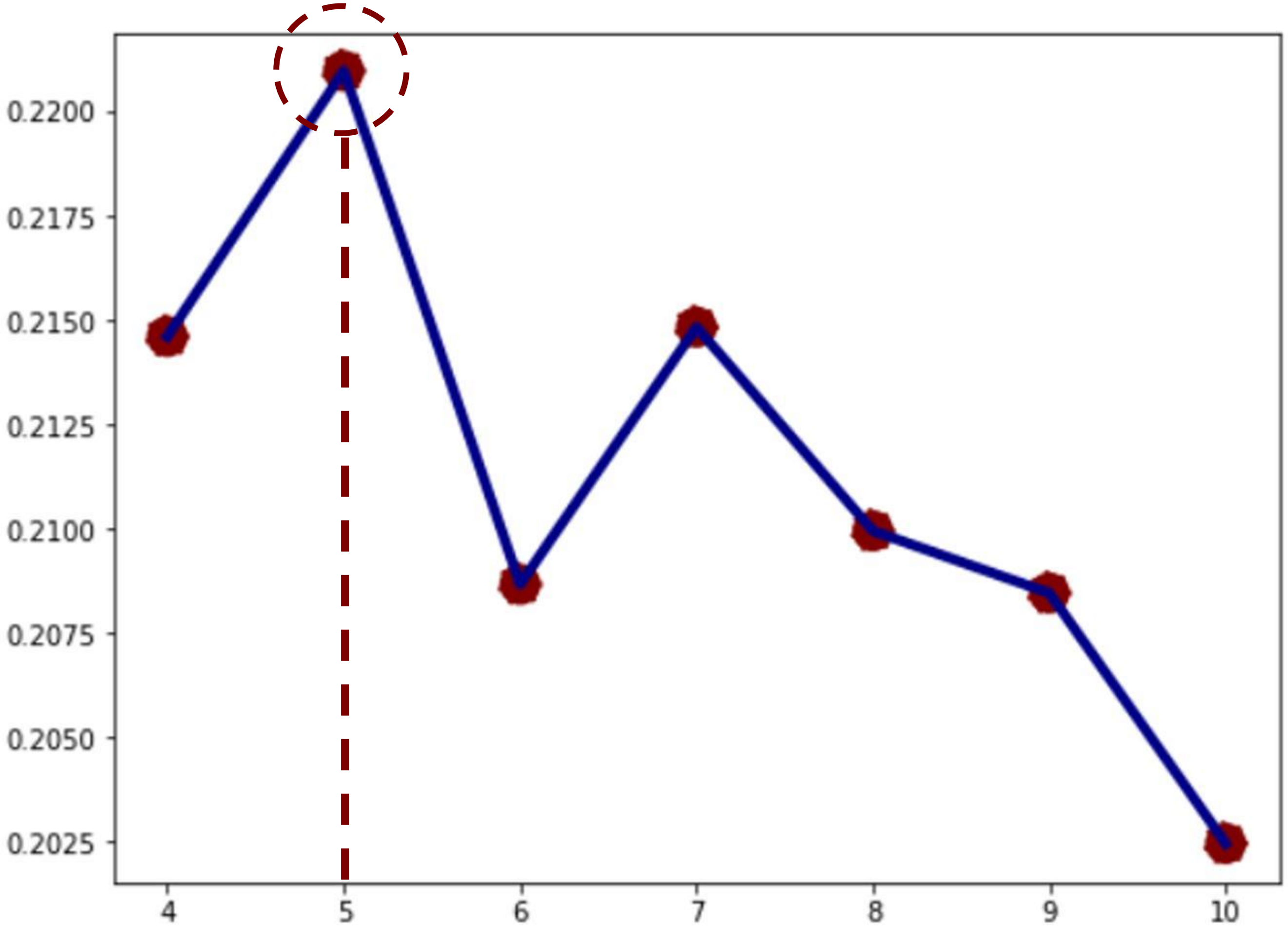


The value of each point in the elbow method has **almost the same** terms, from this graph the best candidates is point **K = 5.**

To make sure the K value, lets try again using **Silhouette Score** method.

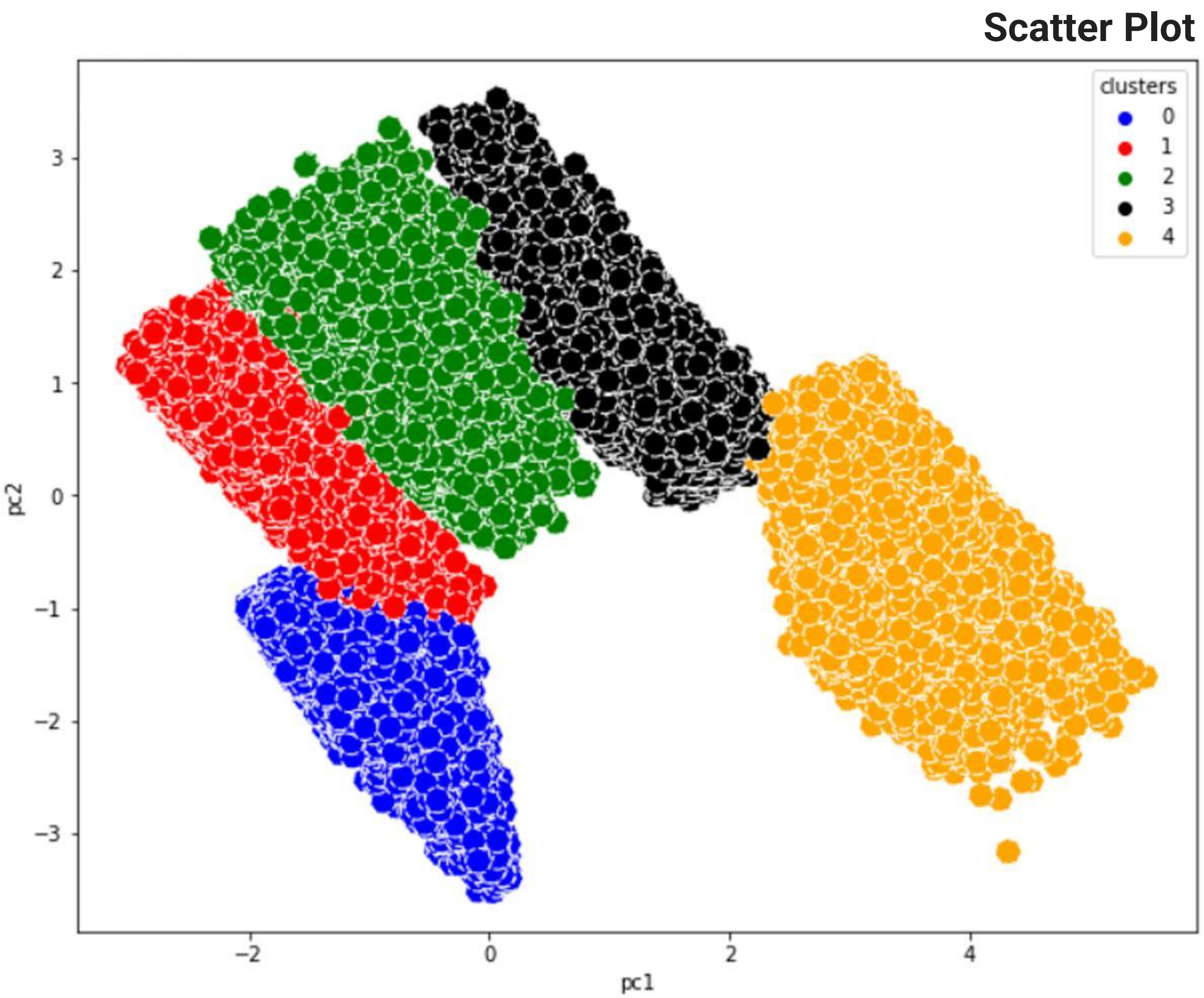
Clustering (Silhouette Score)

Silhouette Score Method



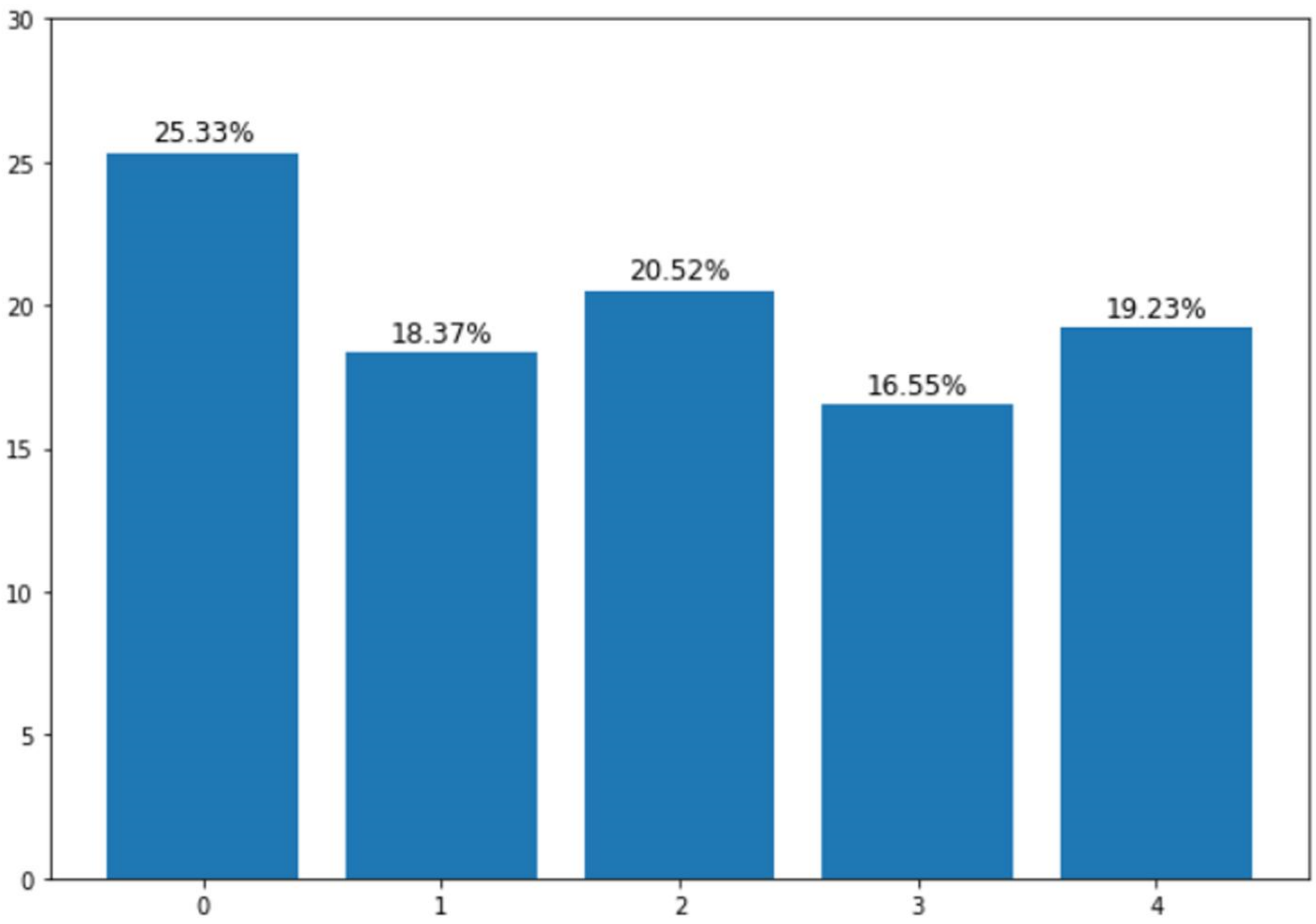
The value **K = 5** is **the right choice** for KMC (K-Means Clustering)**.**

Clustering (Scatter Plot)



When compared with clustering visualization with PCA, then the value of **K = 5 is the right number of clusters**.

Clustering (Bar Plot)



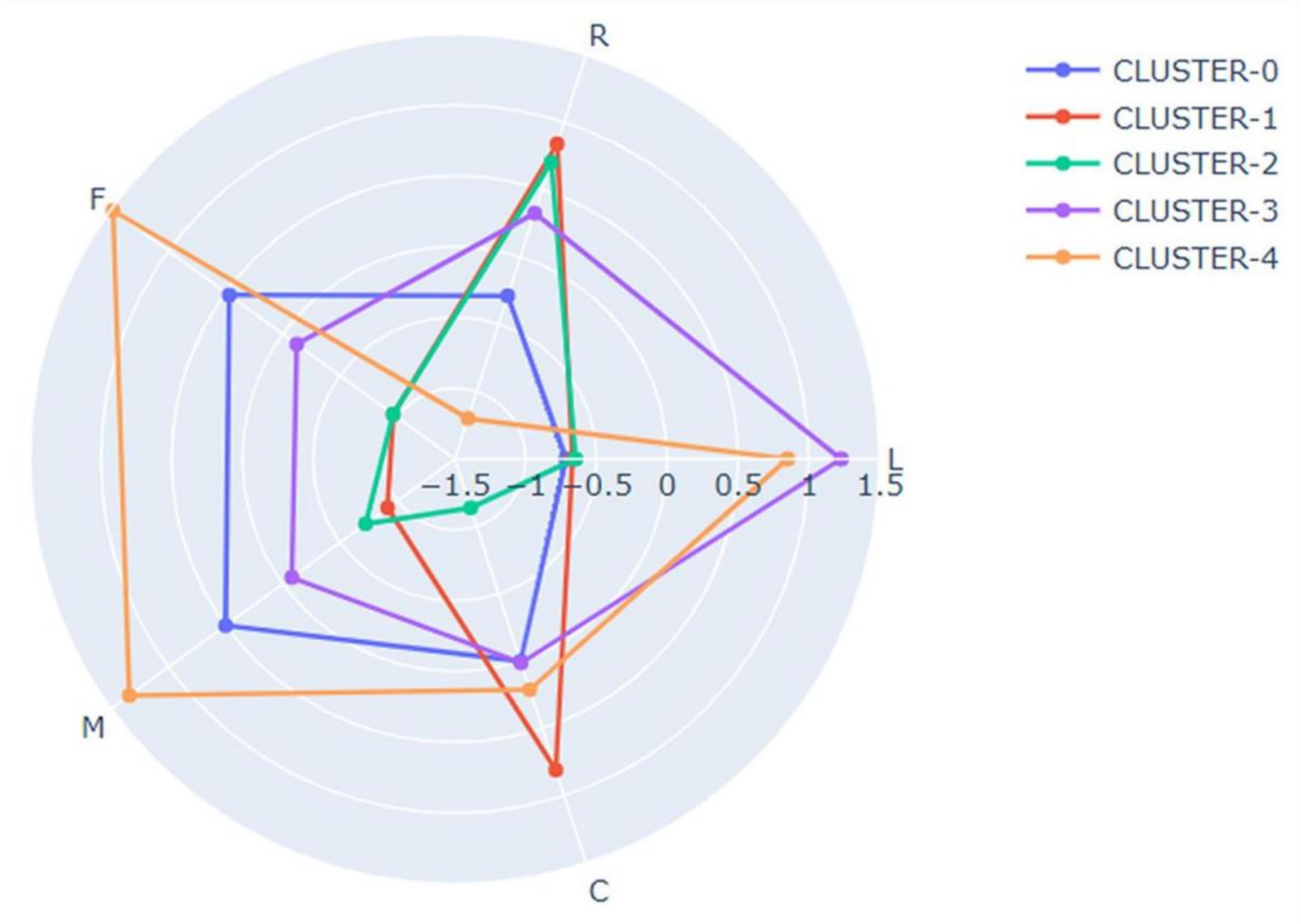
There are 5 clusters:

- Cluster 0 -> **25.33%**
- Cluster 1 -> **18.37%**
- Cluster 2 -> **20.52%**
- Cluster 3 -> **16.55%**
- Cluster 4 -> **19.23%**

Clustering Analysis

- CHARACTERISTIC CLUSTER
- CONCLUSION

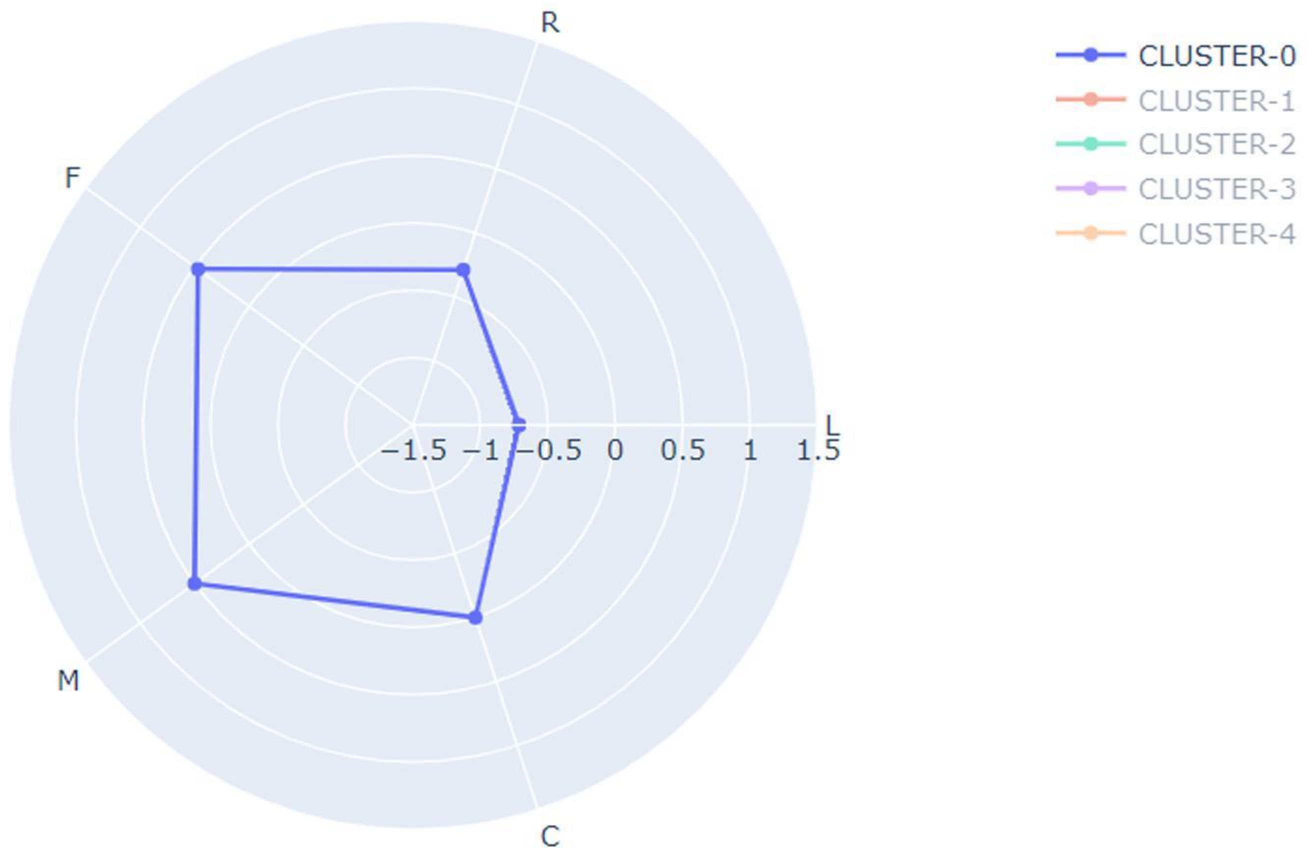
Clustering Characteristic



Cluster 0 = Customer Group 1
Cluster 1 = Customer Group 2
Cluster 2 = Customer Group 3
Cluster 3 = Customer Group 4
Cluster 4 = Customer Group 5

In general, cluster users are formed because there are users who have **high / low** scores on the **LRMFC feature**.

Characteristic - Cluster 0



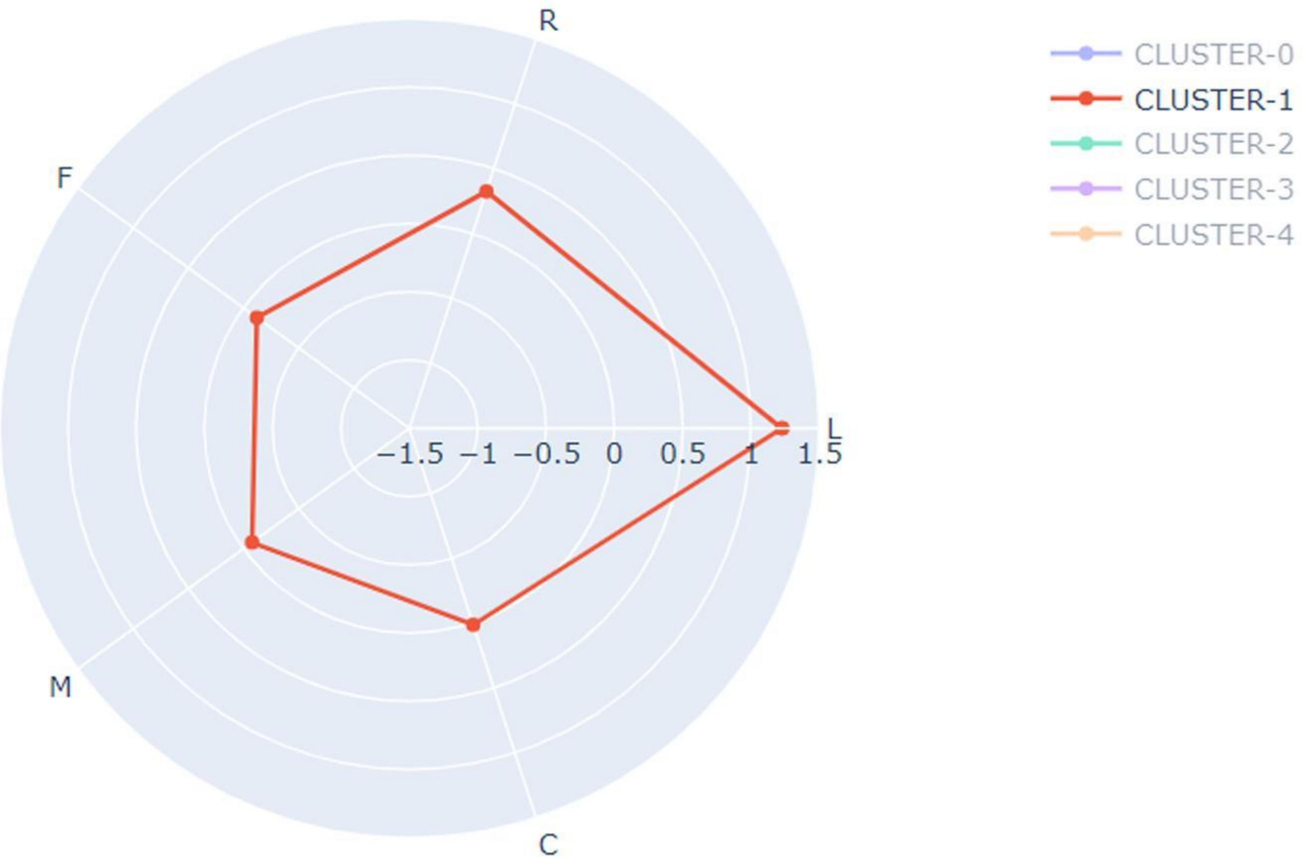
Cluster 0 (Customer Group 1)

Is a new user, because it has a low value on feature L, and can be assumed to be a new user.

New user but has a relatively large number of flights so that it is high on F and M features.

Feature C is quite high, it can be assumed that it is a user who occasionally makes flights because of the promo.

Characteristic - Cluster 1

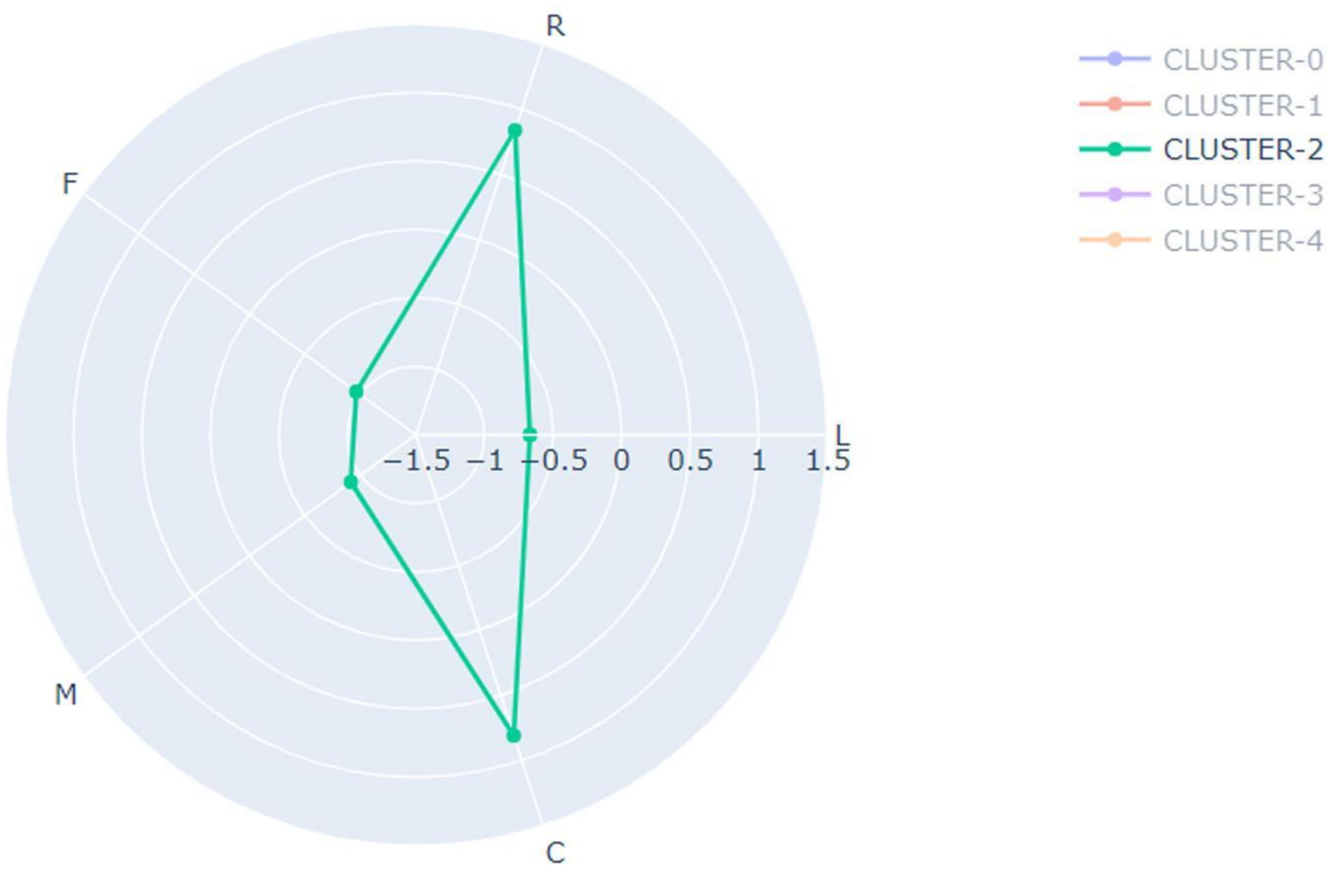


Cluster 1 (Customer Group 2)

This users have LRMFC values that are almost in every feature.

It can be assumed that **old users often fly** and are **promo hunters**, because they have a high C value.

Characteristic - Cluster 2



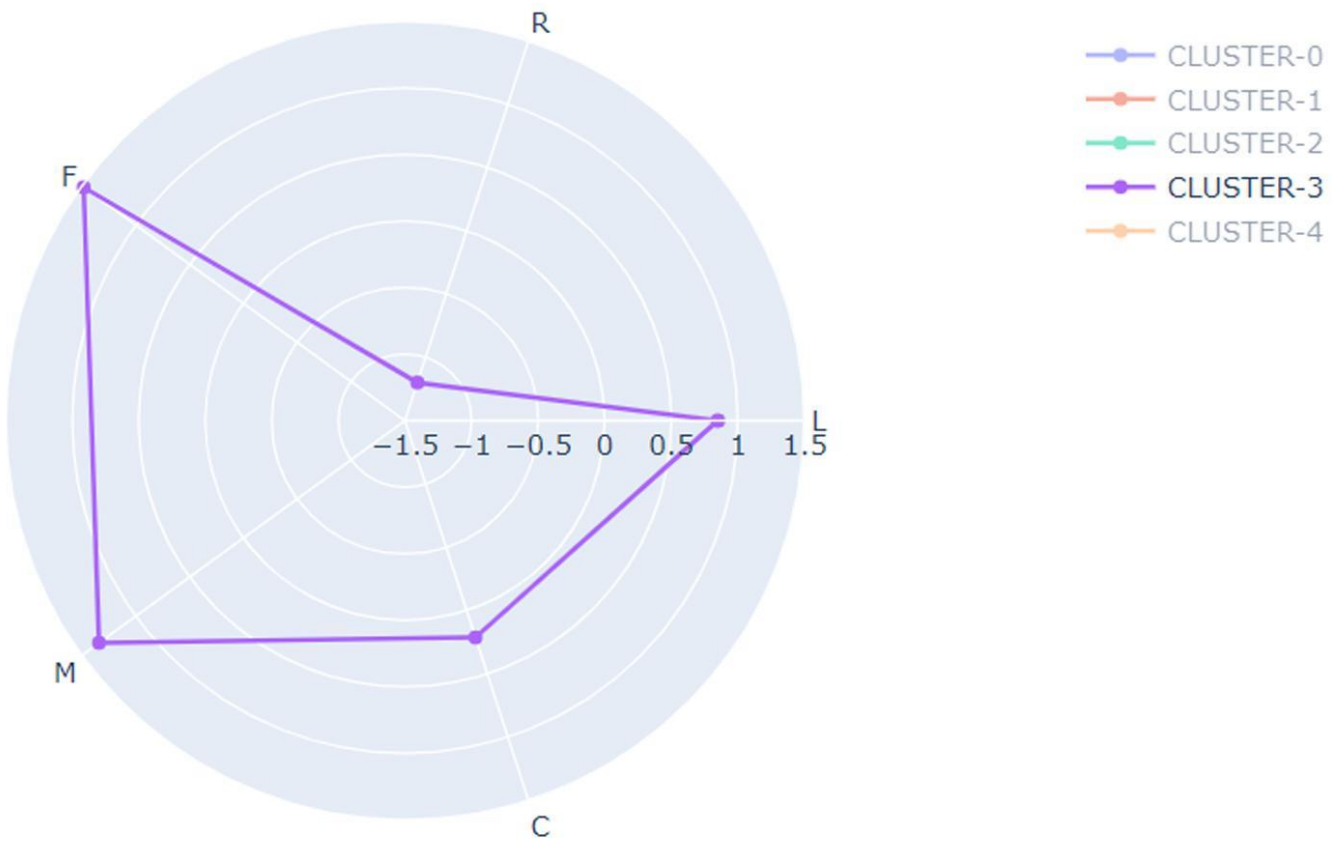
Cluster 2 (Customer Group 3)

Is a new user, because it has a low value on feature L, and can be assumed to be a new user.

New user and has very few flights so very low against F and M features.

but the value of C is quite high, meaning users who **rarely fly but are very interested in the promo**, this is what makes the C value very high.

Characteristic - Cluster 3



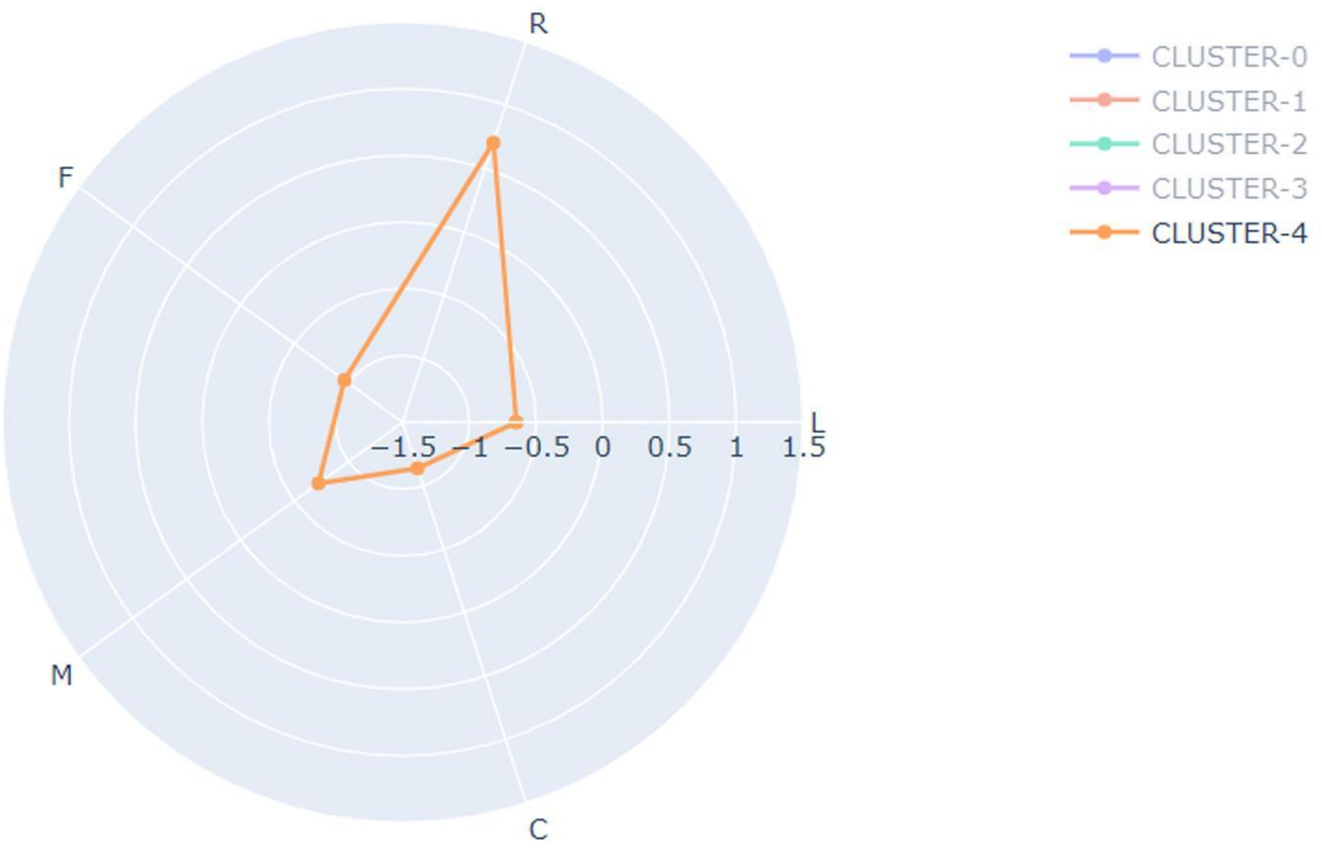
Cluster 3 (Customer Group 4)

Is an user who has been registered for a long time, because it has a high value on the L feature and can be assumed to be an old user.

Old users and have a large number of flights so high on F and M features.

It can be assumed that this user often makes flights and is a loyal customer who sometimes uses promos.

Characteristic - Cluster 4



Cluster 4 (Customer Group 5)

Is a new user, because it has a low value on feature L, and can be assumed to be a new user.

New user and rarely do flights so low on F and M features.

With a low C feature value, it can be assumed that this user is **not interested in the promo**.

Conclusion:

Customer Group 1 -> New user but has a relatively large number of flights.

Customer Group 2 -> Old users often fly and are promo hunters.

Customer Group 3 -> New user and has very few flights, rarely fly but are very interested in the promo.

Customer Group 4 -> Old users and have a large number of flights.

Customer Group 5 -> New user and rarely do flights & not interested in the promo.

Let's Check out my python code Jupyter notebook !!! Don't hesitate to contact me if you want to do some correction or discussion !
#DataScience #ClusteringModelling #DataCleansing #OutlierHandling

aldybudhi003gmail.com https://github.com/qodym https://www.linkedin.com/in/aldybudhi/