Airline Customer Segmentation

A project by Rafif

https://www.linkedin.com/in/mrafifrbbn/
https://github.com/mrafifrbbn

Project Overview

Goal
To create a customer segmentation using the LRFMC model and to give business recommendations based on the results.

Dataset
This project uses the airline customer dataset from Kaggle.

Focus

- End-to-end project on clustering (unsupervised ML).
- LRFMC model for customer segmentation.
- Finding the optimal number of clusters: elbow method and silhouette score.

Dataset

Contains 23 columns:

Flight information:

- `LOAD_TIME` : The end time of the observation window (observation window: time period of observation)
- `FLIGHT_COUNT` : Number of flights in the observation window
- `SUM_YR_1` : Fare revenue
- `SUM_YR_2` : Votes prices
- `SEG_KM_SUM` : Total flight kilometers in the observation window
- `LAST_FLIGHT_DATE` : Last flight date
- `LAST_TO_END` : The time from the last flight to the end of the observation window
- `AVG_INTERVAL` : Average flight time interval
- `MAX_INTERVAL` : Maximum flight interval
- `avg_discount` : Average discount rate

Basic customer information:

- `MEMBER_NO` : Membership card number (ID)
- `FFP_DATE` : Membership join date
- `FIRST_FLIGHT_DATE` : First flight date
- `GENDER` : Gender
- `FFP_TIER` : Membership card level
- `WORK_CITY` : The city where the customer works
- `WORK_PROVINCE` : The province where the customer works
- `WORK_COUNTRY` : The country where the customer works
- `AGE` : Age

Integral information

- `BP_SUM` : Total basic integral
- `EXCHANGE_COUNT` : Number of points exchanged
- `Points_Sum` : Total cumulative points
- `Point_NotFlight` : points not used by the customer

Data Preprocessing

Data Preprocessing

Missing values:

	columns	missing values	pct
0	WORK_PROVINCE	3248	5.157
1	WORK_CITY	2269	3.602
2	SUM_YR_1	551	0.875
3	AGE	420	0.667
4	SUM_YR_2	138	0.219
5	WORK_COUNTRY	26	0.041
6	GENDER	3	0.005

- Mostly from `WORK_PROVINCE` and `WORK_CITY`.
- In total, only ~8% of the records have missing values.
- Drop them all.

Data Preprocessing

- Standard data cleaning for aviation dataset (Tao, 2020):
  - Discard the records where the fare (`SUM_YR_1` or `SUM_YR_2`) is empty.
  - Discard records where the fare is 0, the average discount rate is non-0 **and** the total flying kilometres is greater than 0.
- There are 58015 records left to be analyzed further.

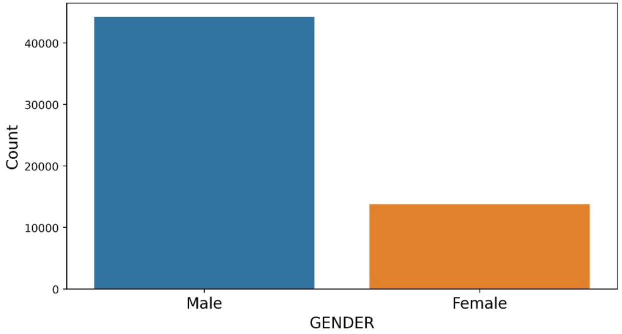
	MEMBER_NO	FFP_DATE	FIRST_FLIGHT_DATE	GENDER	FFP_TIER	WORK_CITY
0	54993	11/2/2006	12/24/2008	Male	6	.
2	55106	2/1/2007	8/30/2007	Male	6	.
3	21189	8/22/2008	8/23/2008	Male	5	Los Angeles
4	39546	4/10/2009	4/15/2009	Male	6	guiyang
5	56972	2/10/2008	9/29/2009	Male	6	guangzhou

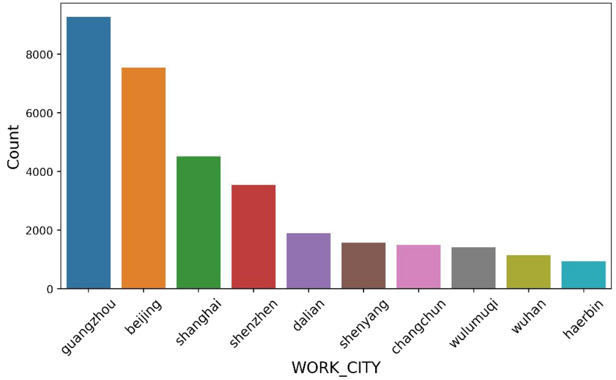
Exploratory Data Analysis

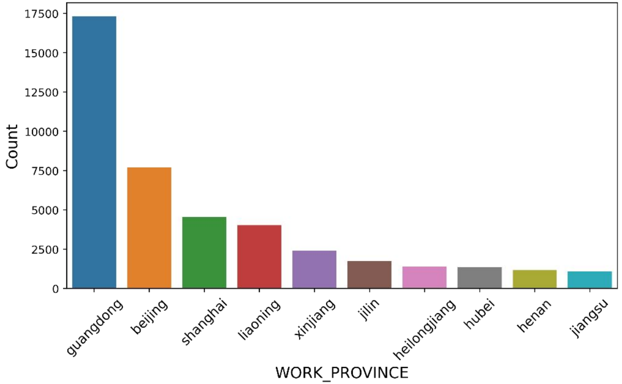
EDA

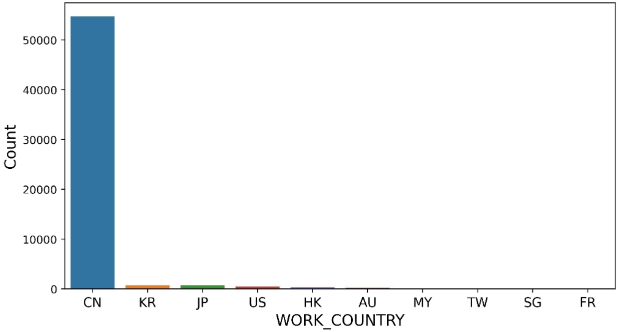
Categorical columns:

- Most customers are male, working in Guangzhou, Guangdong, China.



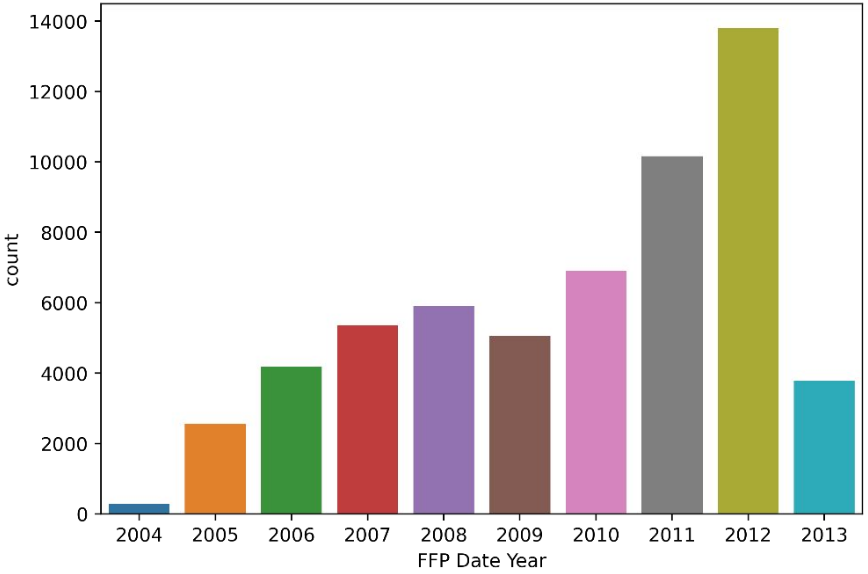


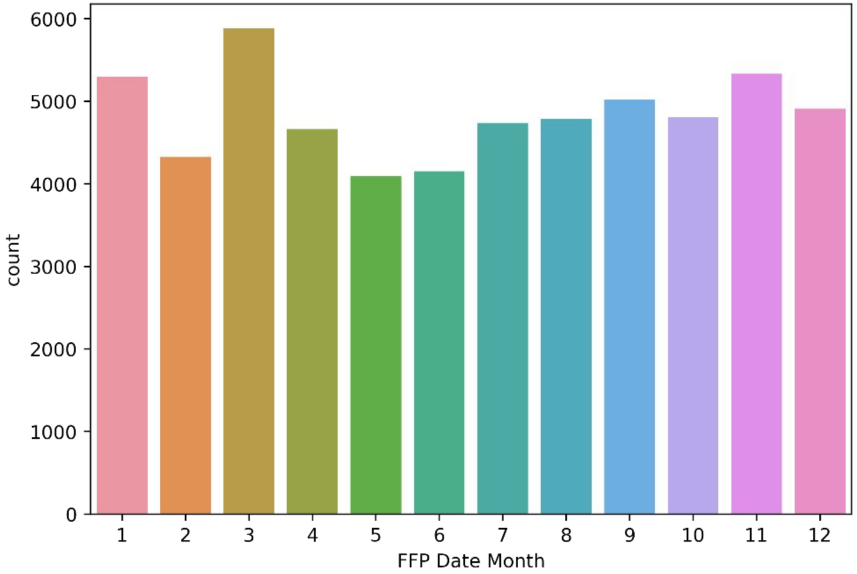




EDA

Date columns (only for `FFP_DATE`) → most members joined in 2012 and in March (not necessarily the same year)



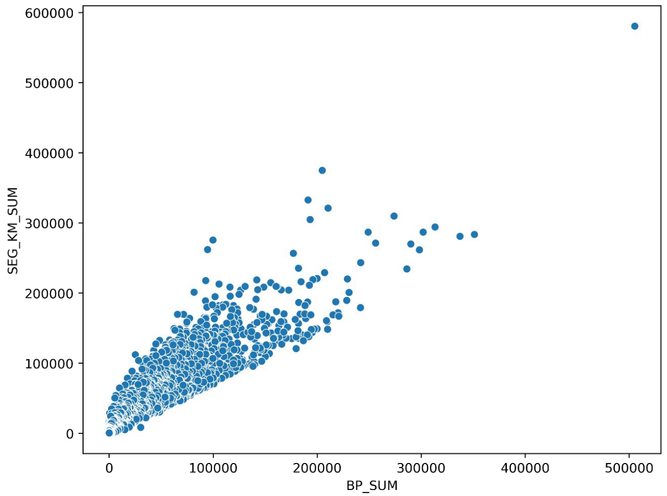


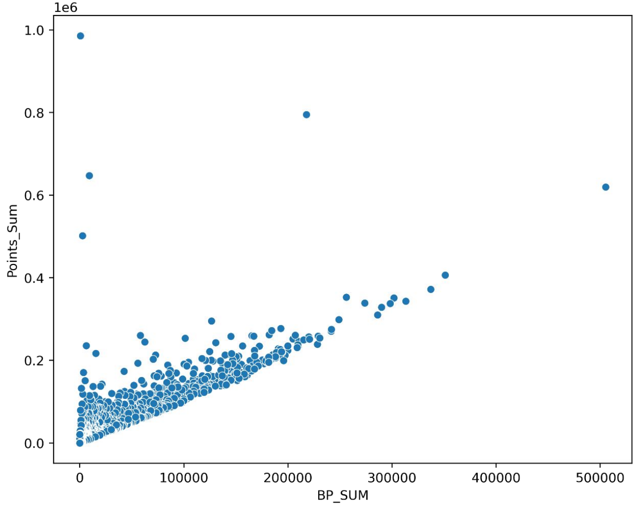
EDA

Numerical columns:

- Strongest correlation between `BP_SUM` and `SEG_KM_SUM`, and `BP_SUM` and `Points_Sum`.

- More total distance → more points (point system based on distance).



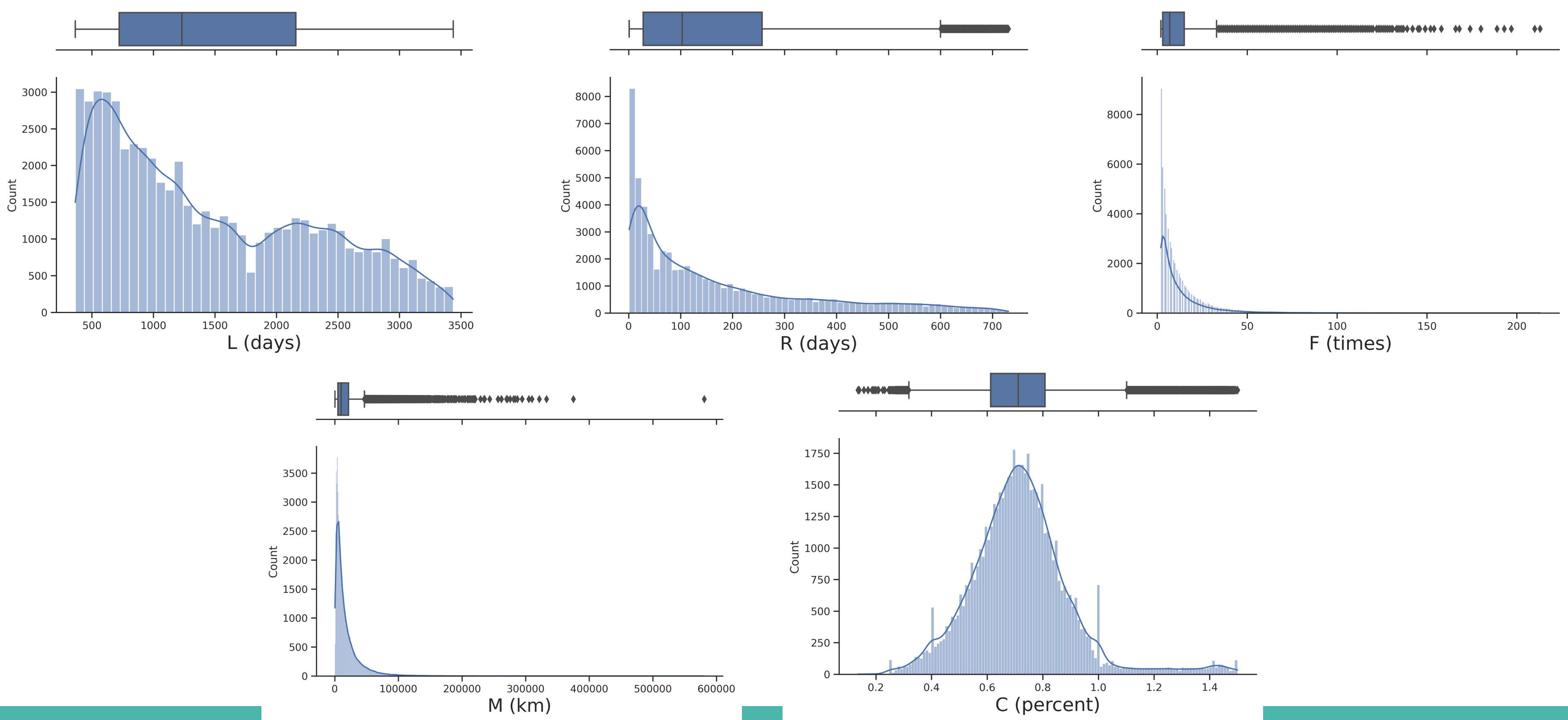


Features: LRFMC model

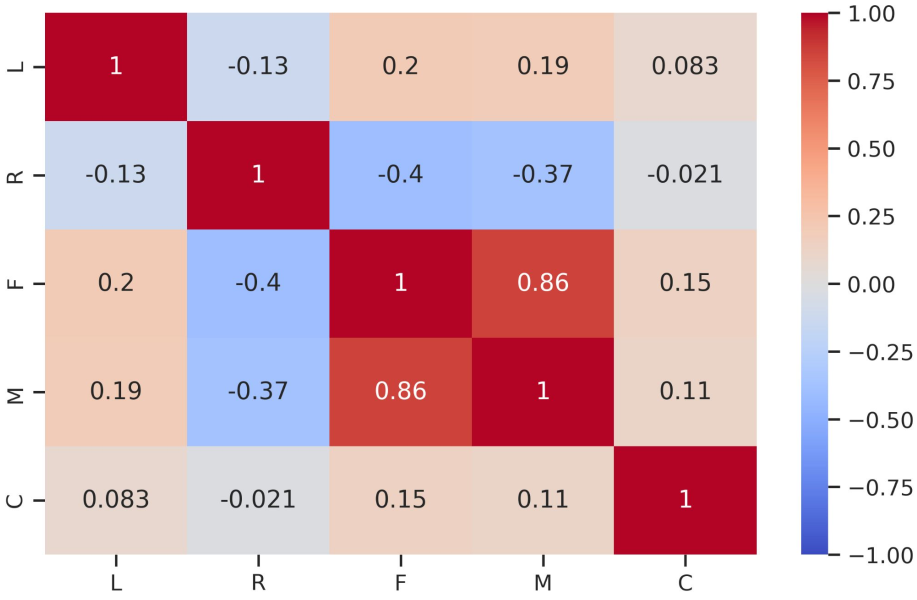
Features: LRFMC model

- The RFM model is often used in customer segmentation problems:
  - Recency (R) → time interval since the last visit/flight
  - Frequency (F) → total number of visits/flights
  - Monetary (M) → total money spent, or total mileage accumulated (for aviation dataset)
- For aviation dataset, two additional features are added (Chen and Wang, 2022):
  - Loyalty (L) →relationship length (how long a customer has been a member)
  - Cabin (C) → average discount price. Larger = higher class in flights
- Ideal customers: high LFMC, low R.
- Using 6 features from the original dataset to extract the LRFMC values: `FFP_DATE`, `LOAD_TIME`, `LAST_TO_END`, `FLIGHT_COUNT`, `SEG_KM_SUM`, and `avg_discount`.

Features: LRFMC model



Features: LRFMC model

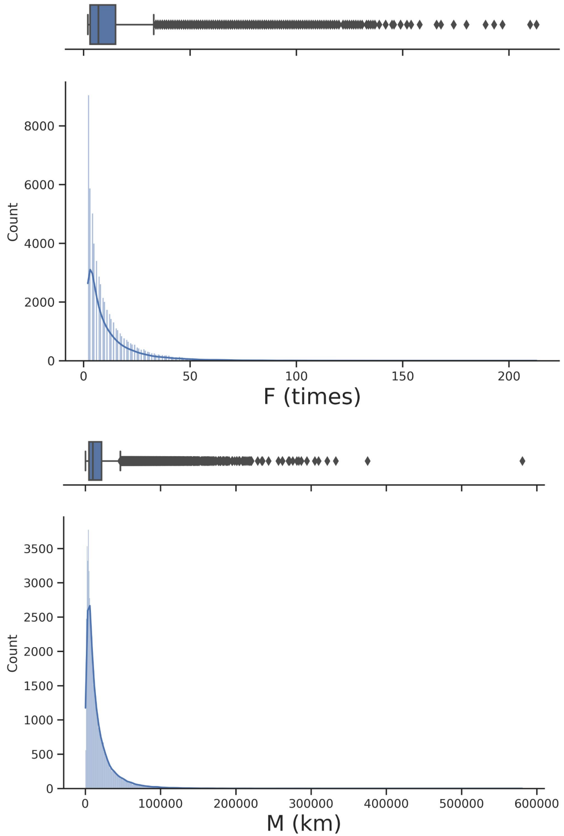


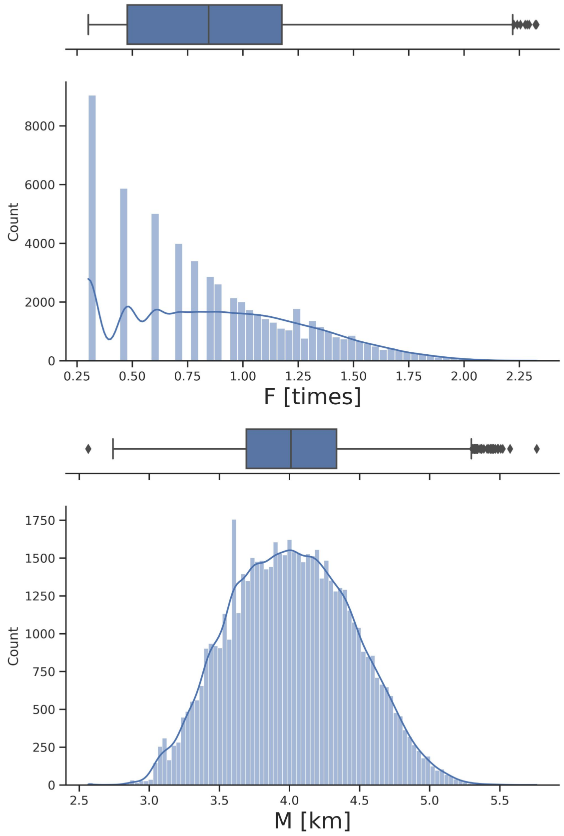
- Strong correlation between `F` and `M` → flying more frequently = more distance covered.
- `R` correlates negatively with the others (especially `F`)→ those who haven't flown in a while rarely flies.

K-Means Clustering

K-Means Clustering

- The `F` and `M` contain a lot of outliers and are heavily skewed, not good for K-means → transform to log units.





K-Means Clustering

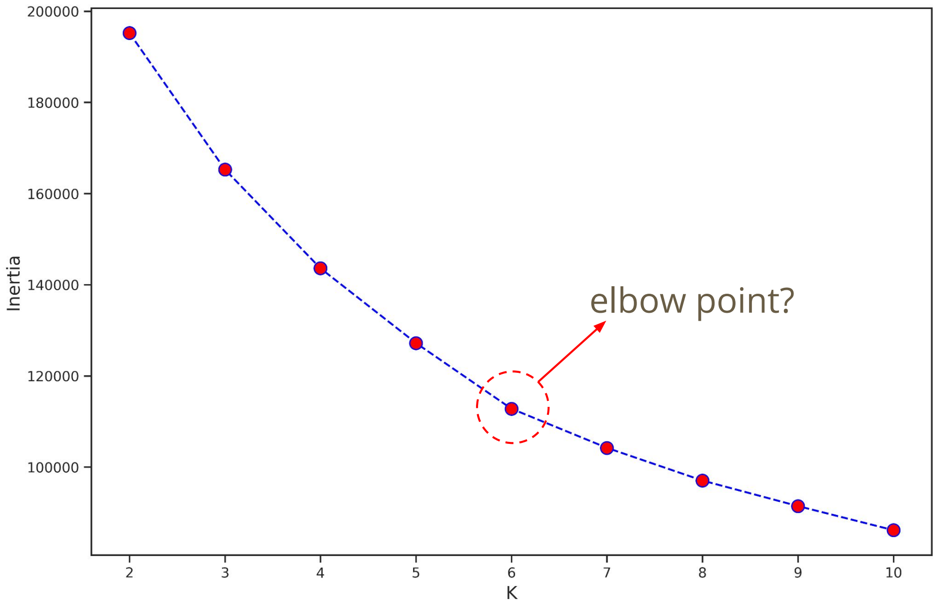
- Scaling with sklearn's `StandardScaler` → mean of 0, variance of 1.

	L	R	F	M	C
0	2706	1	2.322219	5.763965	0.961639
2	2615	11	2.130334	5.452878	1.254676
3	2047	97	1.361728	5.449225	1.090870
4	1816	5	2.181844	5.491261	0.970658
5	2241	79	1.963788	5.469211	0.967692

	L	R	F	M	C
0	1.479608	-0.940166	3.500014	3.965204	1.310440
2	0.695388	-0.409204	1.176379	3.249924	2.014008
3	0.420495	-0.918043	3.160415	3.345454	1.359541
4	0.926251	-0.508760	2.632891	3.295343	1.343397
5	1.747361	-0.940166	2.730950	3.269741	1.330625

K-Means Clustering

- Finding the optimal number of clusters (k value): elbow method



- Plot of inertia or within-cluster sum-of-squares (WCSS) vs. k-value

- Optimal k → the "elbow point". After this point, the inertia decreases linearly (not much improvement, overfitting).

- k=6 seems to be the optimal k, but not very convincing since there is no clear sudden change.

K-Means Clustering

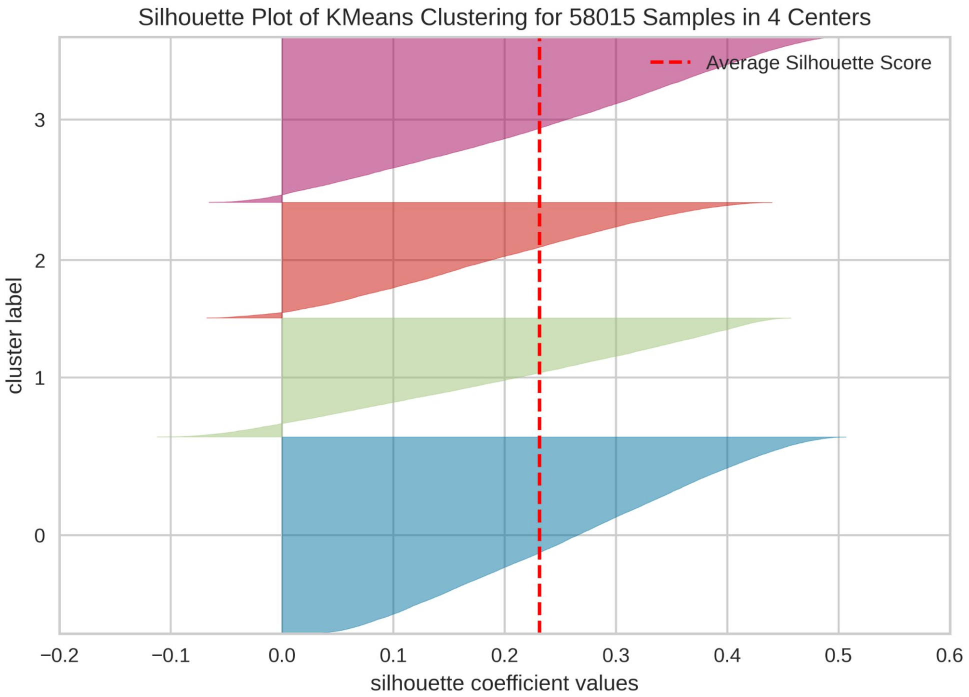
- Another method: silhouette score → uses mean intra-cluster distance a and nearest-cluster distance b for each point.
- For a sample/point, the silhouette coefficient is

$$\frac{b - a}{\max(a, b)}$$

- If $b \gg a$, the nearest-cluster distance is much larger than the cluster size → the clusters are well-separated, the score is ~1.
- If $b \ll a$, the cluster size is much larger than the distance to the nearest cluster → the clusters are mixed together, the score is ~ -1.
- Therefore the range is [-1,1]. Score of 1 is good, -1 is bad.

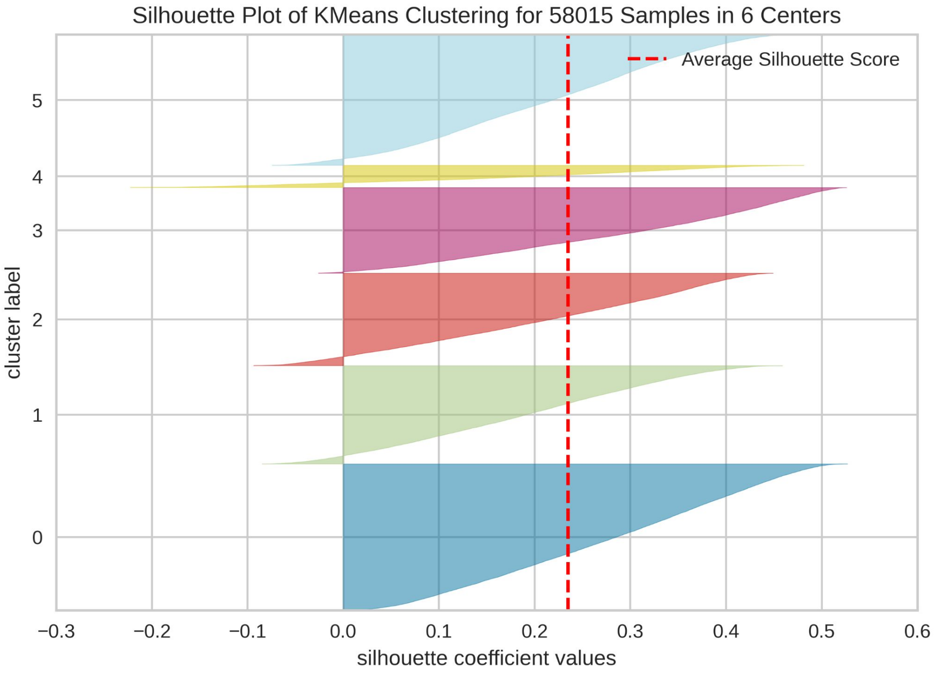
K-Means Clustering

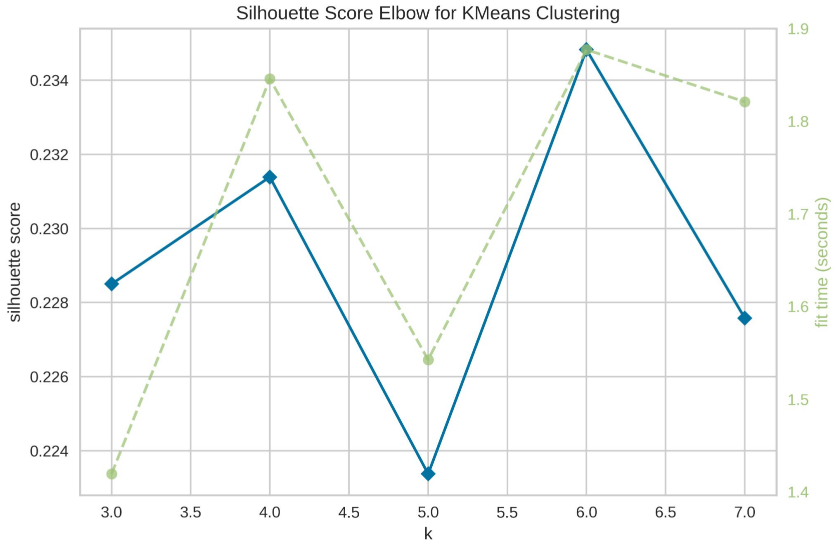
- Silhouette plot: plotting the silhouette score for each point in each cluster, in increasing order.
  - X-axis: silhouette score.
  - Y-axis: cluster member. Thicker = more members in the cluster.
- What we want:
  - Red line (average score) is inside the triangles and as high as possible.
  - The thickness are similar (equal composition in all clusters).



K-Means Clustering

- k=6 yields the largest silhouette score.

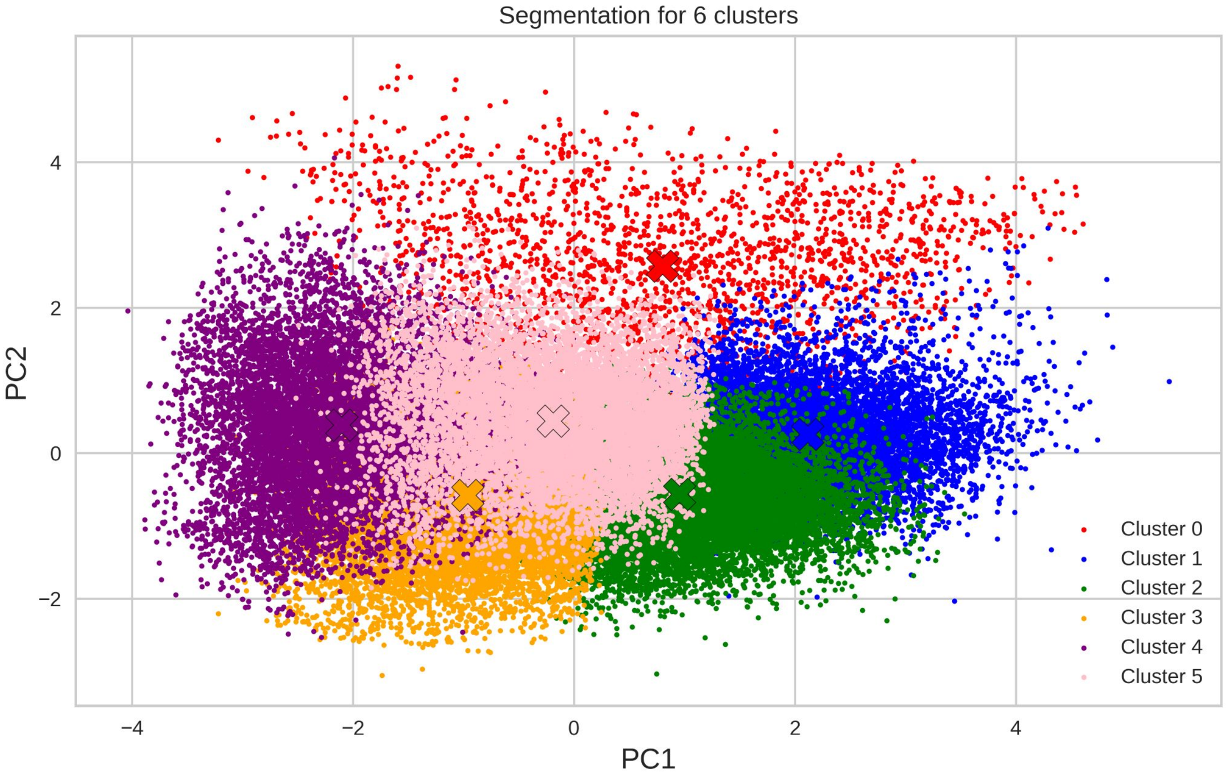




- Cluster 4 has very few members → outlying customers?

- Using k=6 as our benchmark model.

K-Means Clustering

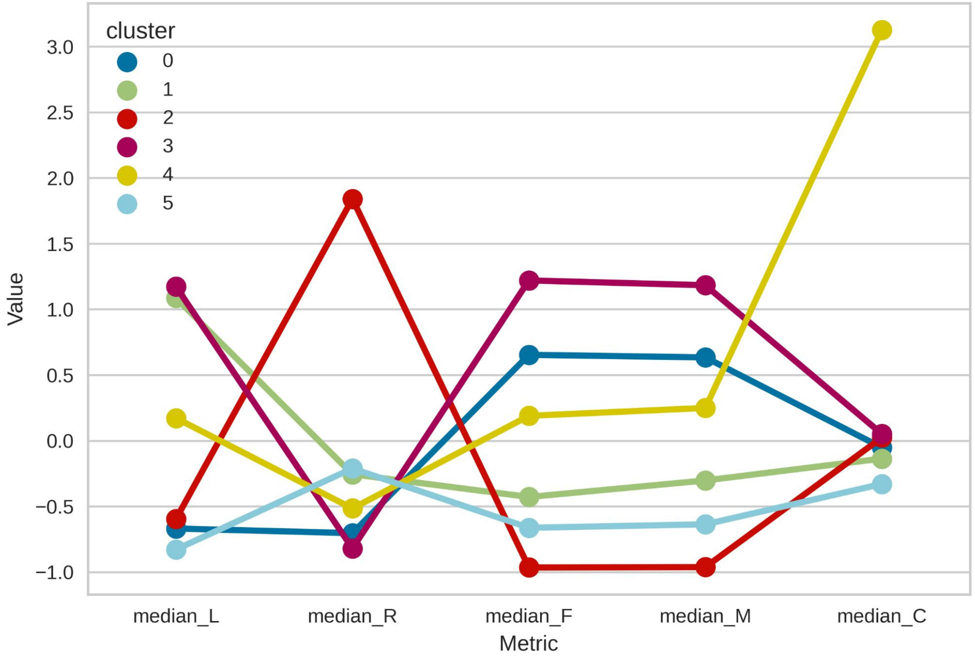


- Visualization with PCA along 2 main PC's → the clusters are very mixed.
- Expected since the average silhouette score is only 0.23 and no signs of multimodality in any of the LRFMC features.

Cluster Analysis

Cluster Analysis

- Create ‘snake plot’ → median of the LRFMC for each cluster.

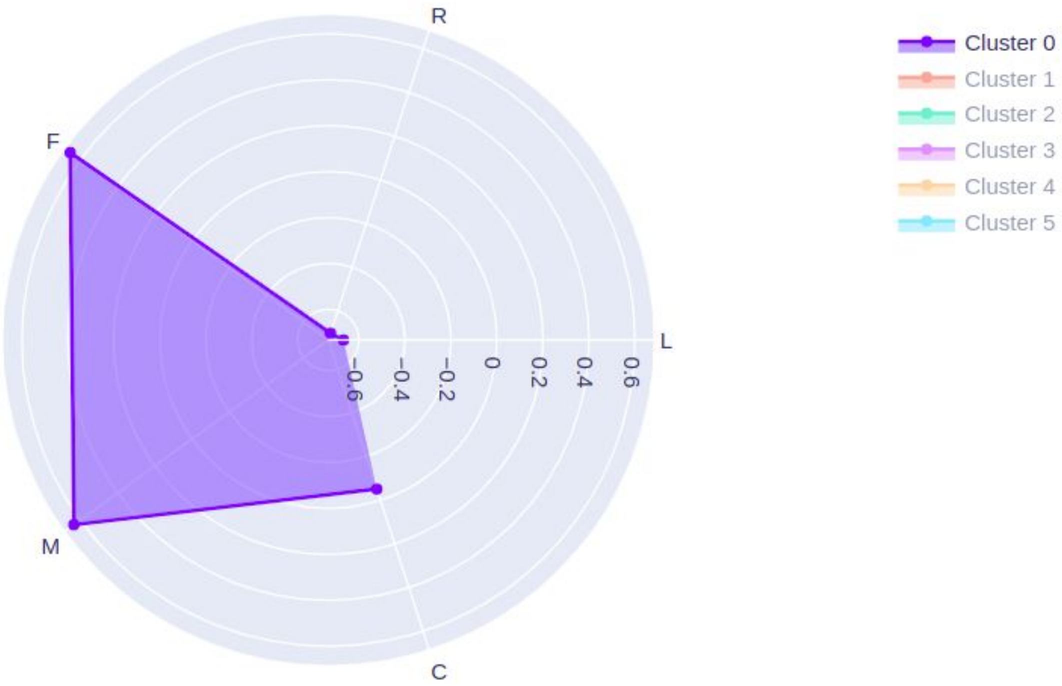


- `L` is grouped into 3: long-time, intermediate-time, and new members.
- `R` is grouped into 2: low and high.
- `F` and `M` are unique for each cluster.
- `C` is grouped into 2: high and low.

	cluster	member
0	2	14134
1	0	9346
2	4	8687
3	3	8271
4	1	8138
5	6	7498
6	5	1941

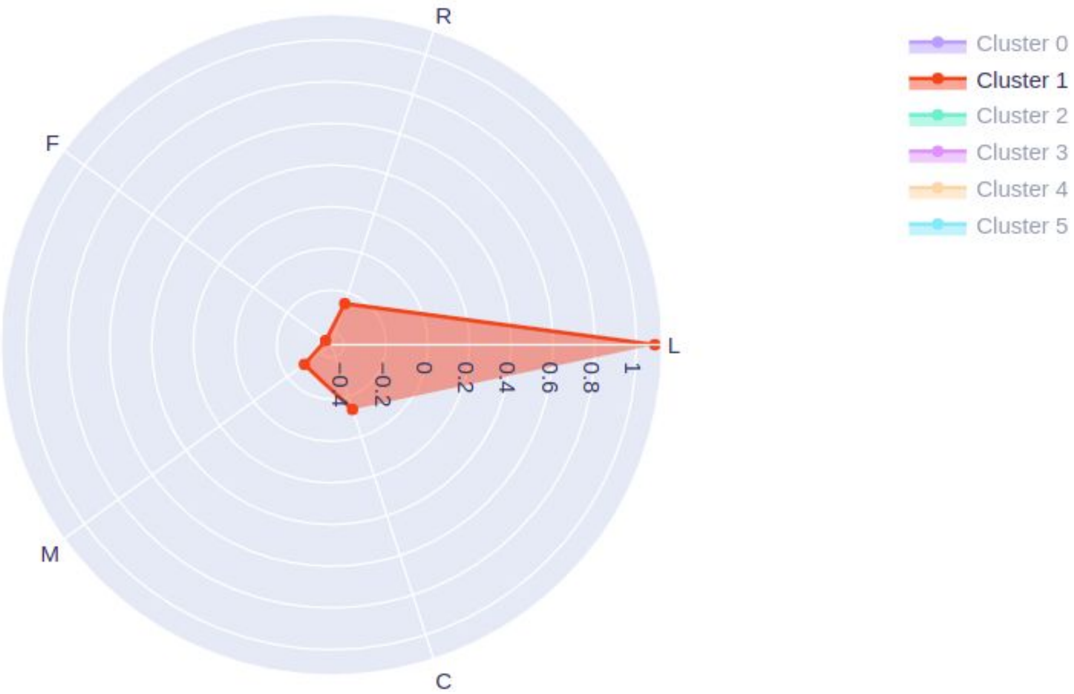
Cluster 2 is the largest, 5 is the smallest.

Cluster Analysis: Cluster 0



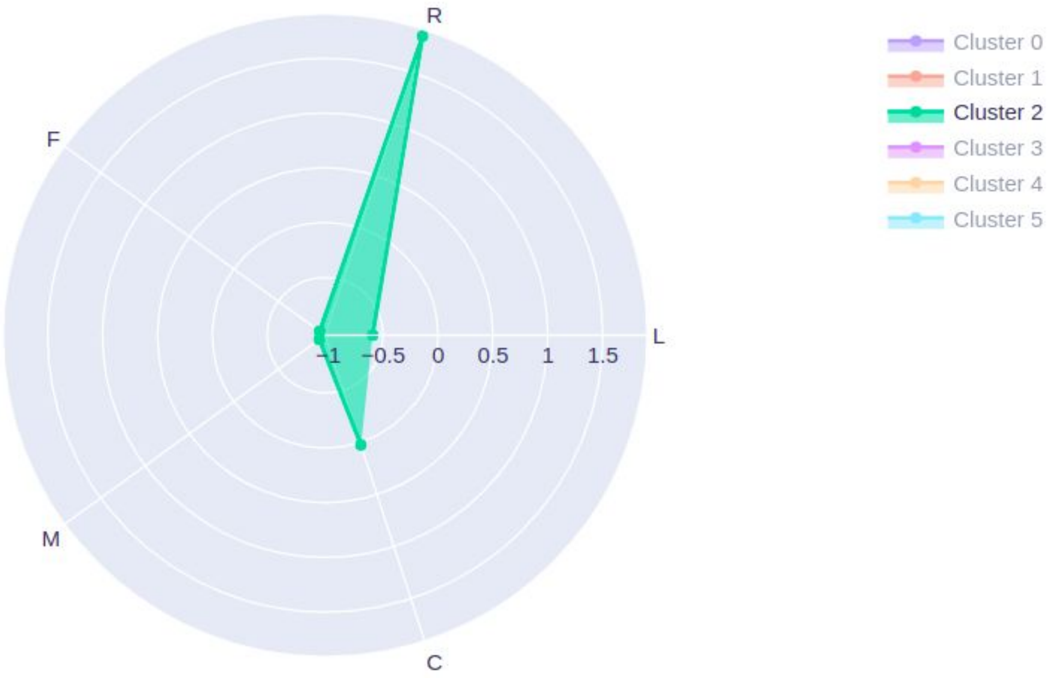
- Low `L`, low `R`, high `FM`, low `C`.
- New members with high consumption → high-value, potential loyal customers.
- Focus on increasing satisfaction and loyalty: extra discounts, free tickets after a certain accumulated mileage, etc.

Cluster Analysis: Cluster 1



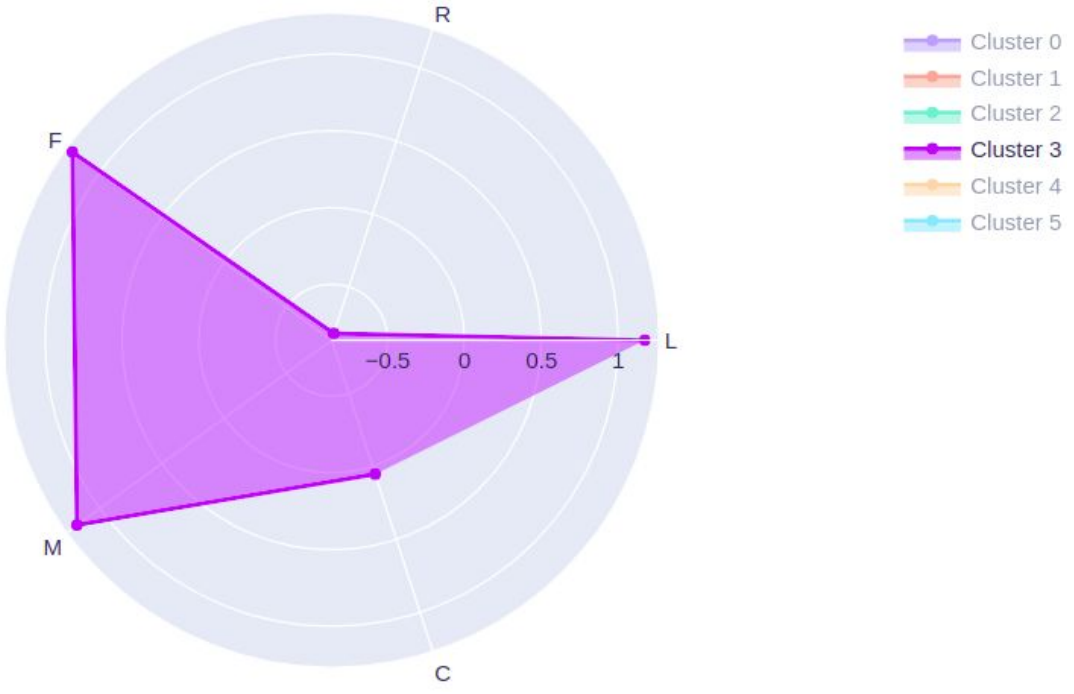
- High `L`, low `R`, low `FM`, low `C`→ long-time members that rarely use our service.
- Low-value customers.
- Encourage consumption? Also may not worth the effort.

Cluster Analysis: Cluster 2



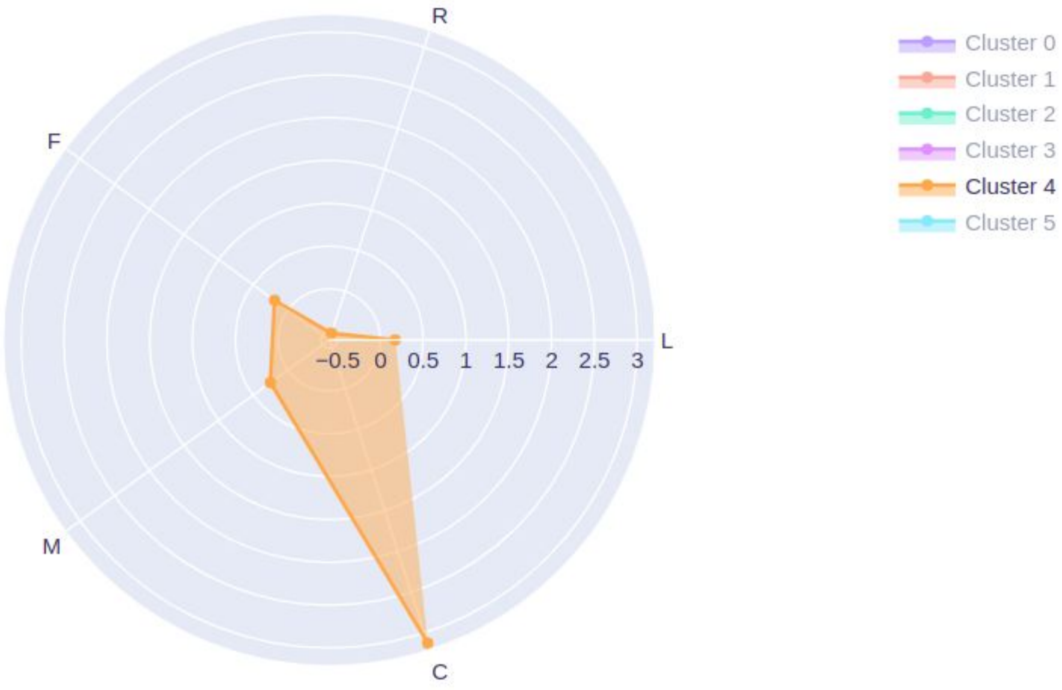
- Very high `R` → churned customers.
- Low-value customers.
- Attract them back? May not worth the effort.
- Interesting to hear their feedback (if they respond!).

Cluster Analysis: Cluster 3



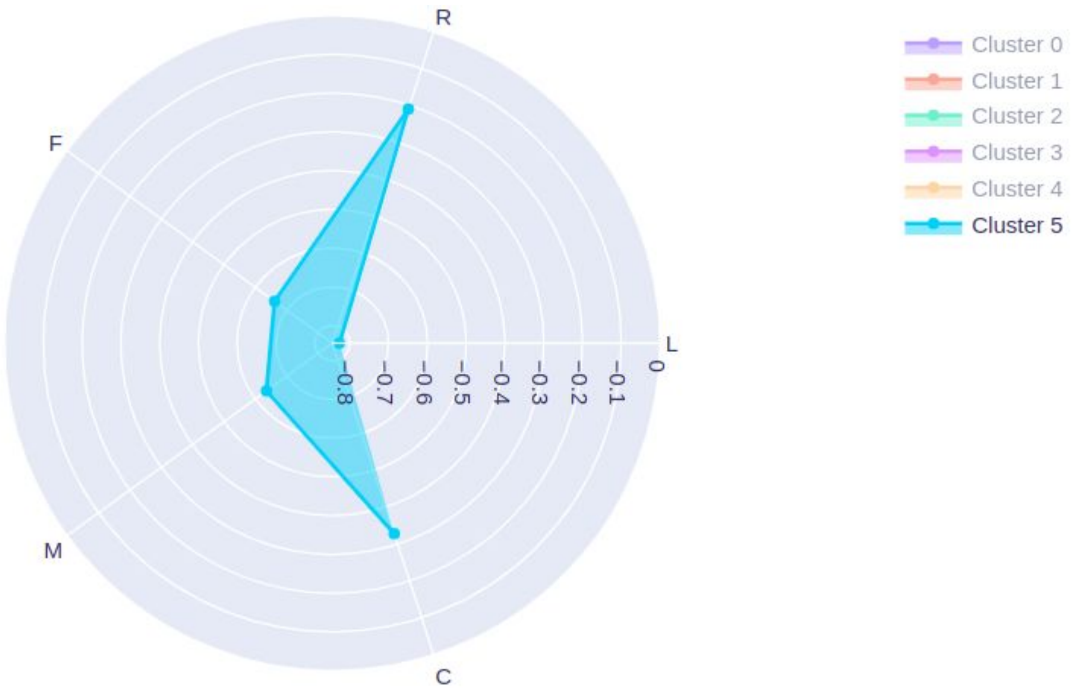
- High `L`, low `R`, very high `FM`, low `C`.
- Our ideal customers → high value, loyal.
- Retain their satisfaction and loyalty with extra service: free food, souvenirs, extra discount for higher seat class.

Cluster Analysis: Cluster 4



- High `C` → uses high class seats, e.g. first/business class (Wang and Chen, 2022).
- Potential VIP customers.
- Differentiated management and one-to-one marketing.

Cluster Analysis: Cluster 5



- Low `L`, moderate `R`, average `FM`, low `C`.
- New members with uncertain status .
- May need to wait to see how they develop.
- Encourage consumption by increasing discount.

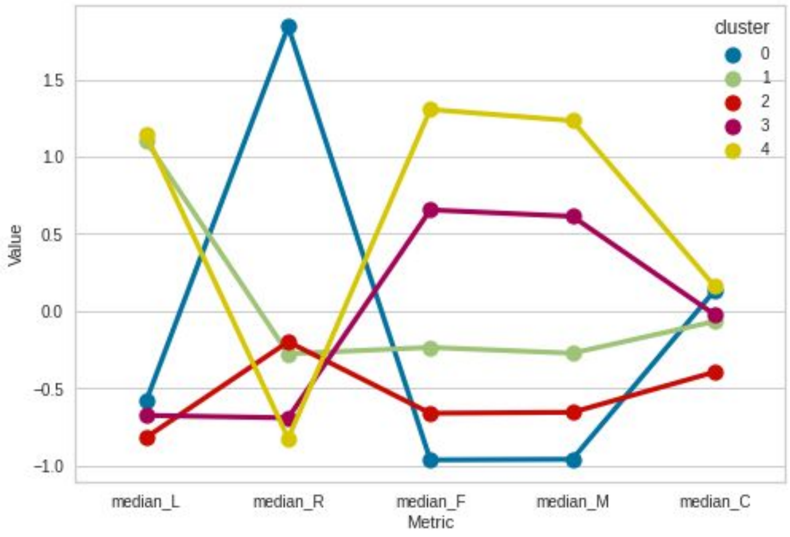
Cluster Analysis: Recommendations

- Implement membership levels: VIP, platinum, diamond, gold, silver, ordinary member, with increasing benefits.
- Point system to obtain higher membership level. Customers can gather points from flight count or accumulated mileage.
- These points expire after a certain period → pushing consumption. Give reminders before the points expire.
- Differentiated management and one-to-one marketing for the VIP, potential, and loyal customers → increase sense of belonging.
- Questionnaire to gain feedback from the low-value customers. If too much effort, just stop promoting to them to cut spending.

Case k=5 and k=7

Case k=5

- Testing out different k-value to see the results. Reducing the cluster to k=5.



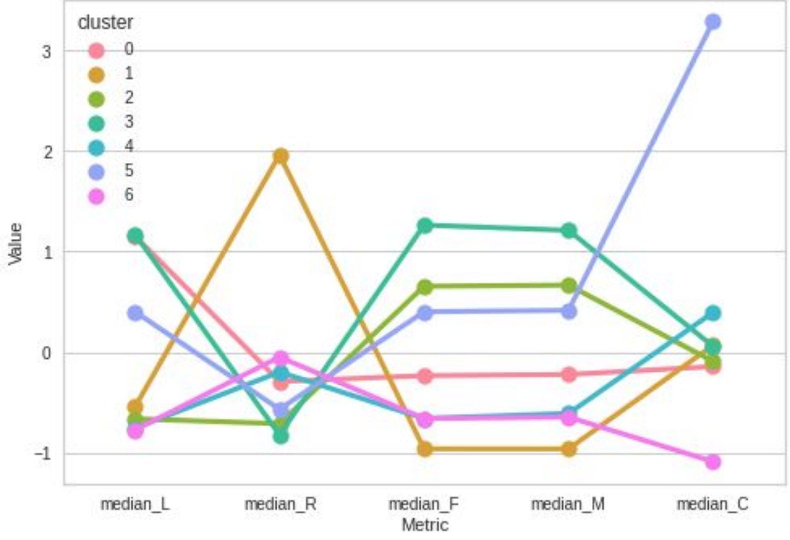
	cluster	member
0	3	15510
1	2	13303
2	1	10560
3	0	9537
4	4	9105

- We lose the potential VIP customers.

- Not optimal to use.

Case k=7

- Increasing the cluster to k=7.



	cluster	member
0	2	14134
1	0	9346
2	4	8687
3	3	8271
4	1	8138
5	6	7498
6	5	1941

- A new segment with very low `C`.
- The other LRFM features are redundant with cluster 4.
- Not worth implementing, overfitting.

Conclusions

- We use the LRFMC model: loyalty, recency, frequency, monetary, and cabin to create customer segmentation for aviation dataset.
- Using the elbow method in combination with the silhouette score to determine optimum k, we get k=6 as our benchmark model.
- We recommend implementing increasing membership level and point system to push consumption and increase loyalty.
- By omitting an additional cluster (k=5), we lose the potential VIP customers. By adding an additional cluster (k=7), we don't gain much information.

References

- RFM Segmentation in E-Commerce: https://towardsdatascience.com/rfm-segmentation-in-e-commerce-e0209ce8fcf6 by Pararawendy Indarjo (Towards Data Science).
- RFM Model for Customer Value of Air Company: https://www.kaggle.com/code/vinzzhang/rfm-model-for-customer-value-of-air-company/data?select=air_data.csv by Vincent Zhang (Kaggle).
- Customer modeling and analysis of civil aviation industry based on Python data analysis: https://pythonmana.com/2021/12/202112130116081138.html by user Mr. thirteen Po (pythonmana).
- Chen, T. & Wang, P. (2022). **IJRES** vol. 10 issue 4 pp. 05-13.
- Tao, Y. (2020). ICPCSEE 2020. Communications in Computer and Information Science, vol 1257. Springer, Singapore. https://doi.org/10.1007/978-981-15-7981-3_7

Also check out the notebook in my GitHub:

https://github.com/mrafifrbbn/airline_customer_segmentation

Contact me on LinkedIn:

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

Fin