AIRLINE CUSTOMER CLUSTERING

Arya Wiratmaji

OUTLINE

- Data Understanding
- Data Cleaning & Preprocessing
- Exploratory Data Analysis
- Modeling
- Insight + Business Recommendations

DATA UNDERSTANDING

The dataset used contains information about customer flights in the Frequent Flyer program of an airline company. This dataset includes various columns related to customer behavior, flight activities, and other relevant metrics.

BACKGROUND

An airline company wants to understand their customer segments based on the flight behavior data they possess.

PROBLEMS

The airline company wants to improve customer loyalty and retention, but they do not know the existing customer segments and the specific behaviors of each segment. Without this understanding, they find it challenging to make targeted business recommendations.

OBJECTIVE

Using clustering methods to identify customer segments based on LRFMC metrics (Length, Recency, Frequency, Monetary, Customer Cost). This way, the airline can make accurate business recommendations based on customer behavior within each segment.

MISSING & DUPLICATE

MISSING VALUES

Column	Missing Values	Missing Value Percentage
GENDER	3	0.004763
WORK_CITY	2269	3.602273
WORK_PROVINCE	3248	5.156538
WORK_COUNTRY	26	0.041278
AGE	420	0.666794
SUM_YR_1	551	0.874770
SUM_YR_2	138	0.219089

Decisions:

- We will drop WORK_CITY, WORK_PROVINCE, WORK_COUNTRY as they have many unique categorical features and we won't need them for our clustering project.
- We will replace the missing values of GENDER with the most frequent value (mode).
- We will replace the missing values of AGE, SUM_YR_1, SUM_YR_2 with median.

DUPLICATE VALUES

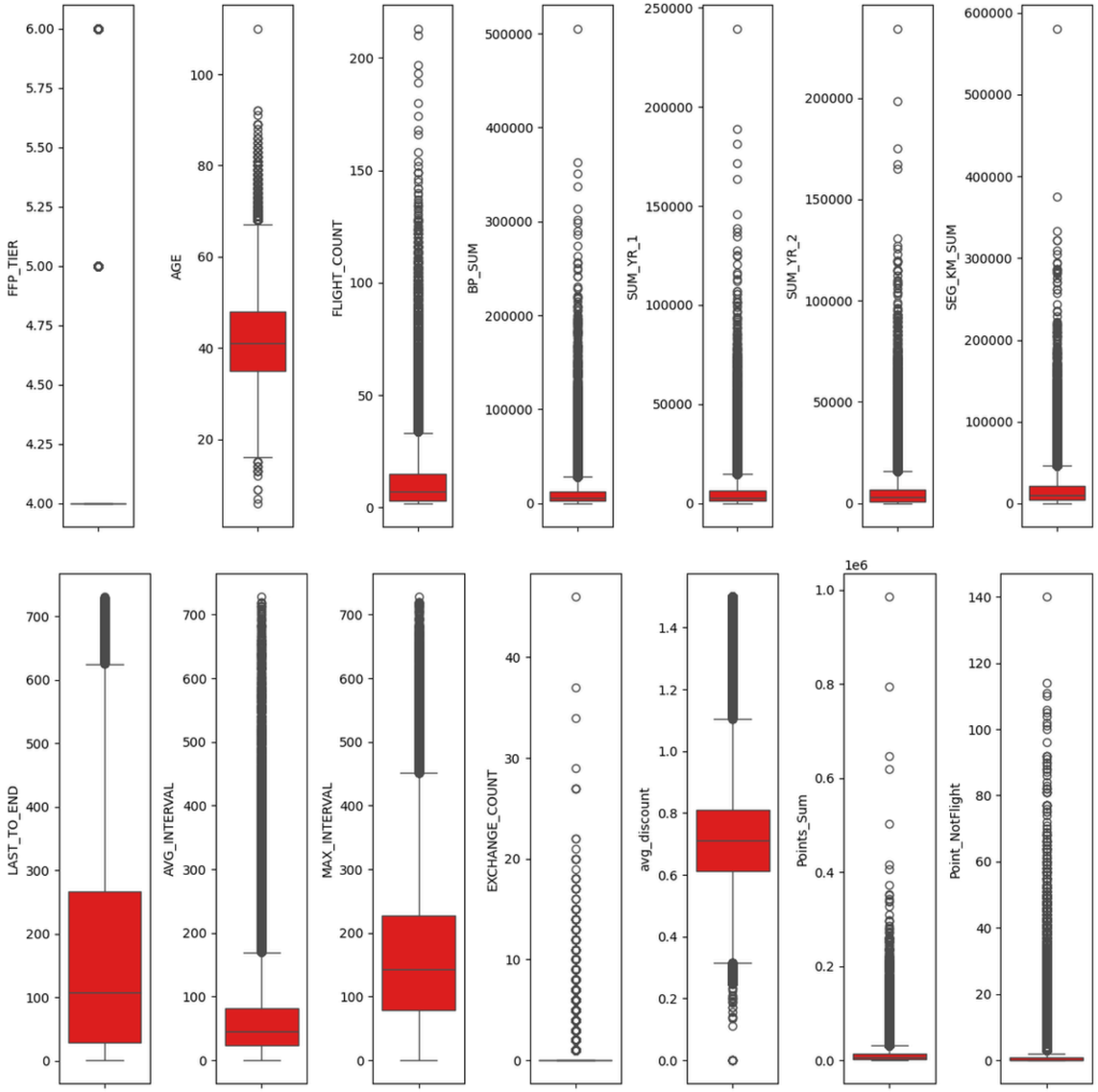
1 Duplicate

	FFP_DATE	FIRST_FLIGHT_DATE	GENDER	FFP_TIER	AGE	LOAD_TIME	FLIGHT_COUNT	BP_SUM	SUM_YR_1	SUM_YR_2	SEG_KM_SUM	LAST_FLI
49070	8/11/2012	8/11/2012	Male	4	40.0	3/31/2014	2	1841	2401.0	0.0	4844	
49085	8/11/2012	8/11/2012	Male	4	40.0	3/31/2014	2	1841	2401.0	0.0	4844	

Decision:

- Drop duplicate data

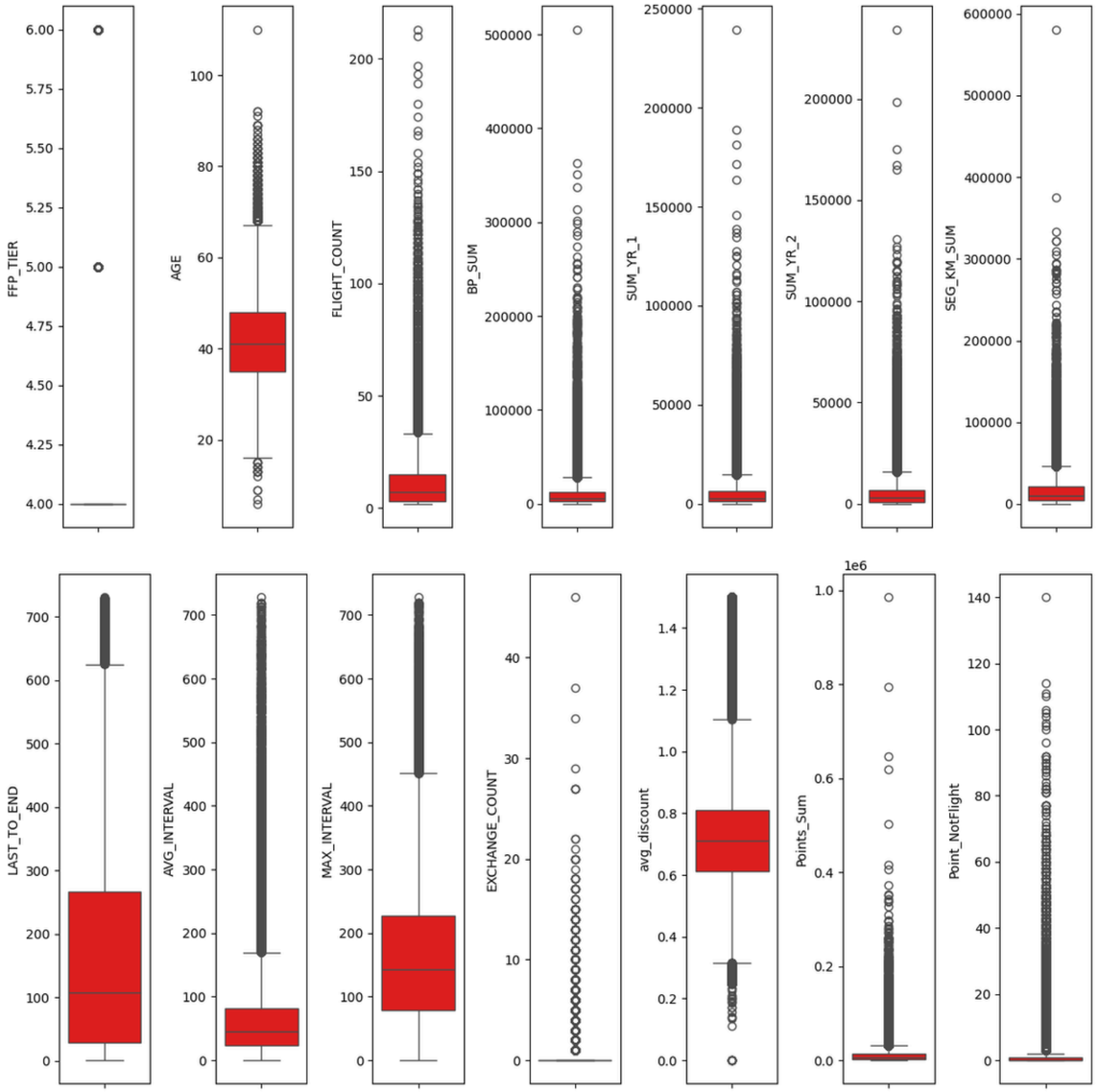
OUTLIERS CHECK



it can be observed that there are numerous outliers detected. However, the majority of these outliers represent the characteristics of the population in the dataset (also known as "true outliers").

Therefore, in this case, I will only drop outliers that are extremely deviant.

OUTLIERS REMOVAL



Outliers removal:

- 1 row of data with AGE > 100, which is 110
- 1 row of data with BP_SUM > 400000, which is 505308
- 1 row of data with SUM_YR_1 > 400000, which is 239560
- 2 rows of data with SUM_YR_2 > 180000, which are 234188 & 198224
- 1 row of data with SEG_KM_SUM > 400000, which is 580717
- 1 row of data with EXCHANGE_COUNT > 40, which is 46
- 2 rows of data with Points_Sum > 700000, which is 795398 & 985572
- 1 row of data with Points_NotFlight > 120, which is 140

STATISTICAL SUMMARY

Feature	Count	Mean	Std. Dev	Min	25%	50%	75%	Max
FFP_TIER	62980	4.10	0.37	4.0	4.00	4.00	4.00	6.0
AGE	62980	42.47	9.85	6.0	35.00	41.00	48.00	92.0
FLIGHT_COUNT	62980	11.83	14.00	2.0	3.00	7.00	15.00	197.0
BP_SUM	62980	10908.94	16168.88	0.0	2518.00	5700.00	12827.00	362480.0
SUM_YR_1	62980	5325.80	8003.43	0.0	1020.00	2800.00	6522.25	188926.0
SUM_YR_2	62980	5589.11	8604.25	0.0	785.00	2773.00	6825.25	174895.0
SEG_KM_SUM	62980	17107.41	20811.85	368.0	4747.00	9993.50	21267.50	375074.0
LAST_TO_END	62980	176.13	183.82	1.0	29.00	108.00	268.00	731.0
AVG_INTERVAL	62980	67.75	77.52	0.0	23.37	44.67	82.00	728.0
MAX_INTERVAL	62980	166.04	123.40	0.0	79.00	143.00	228.00	728.0
EXCHANGE_COUNT	62980	0.32	1.11	0.0	0.00	0.00	0.00	37.0
AVG_DISCOUNT	62980	0.72	0.19	0.0	0.61	0.71	0.81	1.5
POINTS_SUM	62980	12499.35	19694.76	0.0	2775.00	6327.50	14299.00	647113.0
POINT_NOTFLIGHT	62980	2.72	7.33	0.0	0.00	0.00	1.00	114.0
LOAD_TIME	62980	2014-03-31 00:00:00.000000256	N/A	2014-03-31	2014-03-31	2014-03-31	2014-03-31	2014-03-31
FFP_DATE	62980	2010-03-07 01:53:24.445855744	N/A	2004-11-01	2008-04-12	2010-10-08	2012-03-29	2013-03-31
FIRST_FLIGHT_DATE	62980	2010-09-01 19:34:39.517307392	N/A	1905-12-31	2008-12-15	2011-04-23	2012-07-26	2015-05-30
LAST_FLIGHT_DATE	62980	2013-10-07 14:32:09.755477760	N/A	2012-04-01	2013-07-08	2013-12-15	2014-03-03	2014-03-31

STATISTICAL SUMMARY

Anomaly 1

Feature	Count	Mean	Std. Dev	Min	25%	50%	75%	Max
AGE	62980	42.47	9.85	6.0	35.0	41.0	48.0	92.0

On the `AGE` column, the minimum value is 6. According to the research we conducted, the majority of airlines have a policy stating that individuals under the age of 15 are not permitted to fly alone, and must be accompanied by a parent/guardian
(Source: https://www.transportation.gov/sites/dot.gov/files/docs/Kids_Fly_Alone.pdf).

Therefore, we will drop rows where AGE < 15, as it is not common practice for kids to have membership, and we will not use them for our project.

STATISTICAL SUMMARY

Anomaly 2

Feature	Count	Mean	Min	25%	50%	75%	Max
FIRST_FLIGHT_DATE	62980	2010-09-01 19:34:39.517307392	1905-12-31	2008-12-15	2011-04-23	2012-07-26	2015-05-30

There is an anomaly in the `FIRST_FLIGHT_DATE` column where the minimum value is `1905-12-31`. Upon further investigation, there are 2 rows of data that have this value.

FFP_DATE	FIRST_FLIGHT_DATE	GENDER	FFP_TIER	AGE	LOAD_TIME	FLIGHT_COUNT
2011-02-27	1905-12-31	Male	4	35	2014-03-31	40
2004-11-10	1905-12-31	Female	4	37	2014-03-31	8

The `FFP_DATE` for both of these rows is in the years 2004 and 2011, with ages of 35 and 37, respectively. These two data points are likely system errors because the difference between `FIRST_FLIGHT_DATE` and `FFP_DATE` is nearly 100 years, while their ages are only 35 and 37. Therefore, we will remove both rows of data.

STATISTICAL SUMMARY

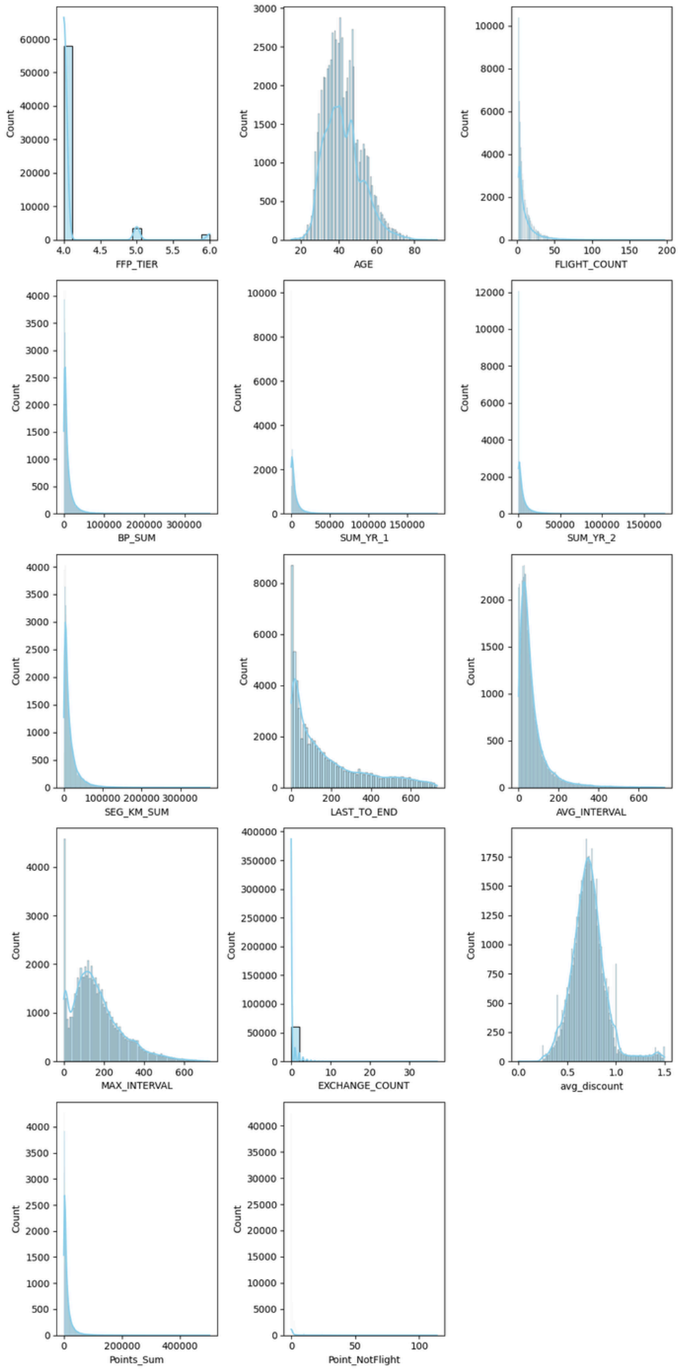
Anomaly 3

FFP_DATE	FIRST_FLIGHT_DATE	GENDER	FFP_TIER	AGE	LOAD_TIME	FLIGHT_COUNT	BP_SUM	SUM_YR_1	SUM_YR_2	SEG_KM_SUM	LAST_FLIGHT_DATE
2008-09-24	2015-03-09	Male	5	47	2014-03-31	25	41312	18347	19595	52083	2014-01-08
2007-06-04	2015-02-15	Male	4	41	2014-03-31	43	33326	14258	22887	63459	2014-03-27
2007-04-01	2015-05-30	Male	4	43	2014-03-31	17	11376	4956	7810	17470	2014-02-17
2006-03-08	2014-07-14	Male	4	38	2014-03-31	11	9113	576	5820	18642	2013-09-15
2011-01-15	2015-04-03	Female	4	48	2014-03-31	8	5502	0	7030	14245	2014-03-24
2006-11-10	2014-09-11	Male	4	48	2014-03-31	2	6540	0	5767	4564	2014-02-09

An anomaly has been identified where there are rows of data with the feature FIRST_FLIGHT_DATE exceeding its LOAD_TIME and LAST_FLIGHT_DATE. Logically, this scenario is not possible because FIRST_FLIGHT_DATE must be less than or equal to LAST_FLIGHT_DATE, as LAST_FLIGHT_DATE represents the last registered flight in the database. There are 6 rows of data exhibiting these anomalous characteristics. These 6 rows of data will be removed.

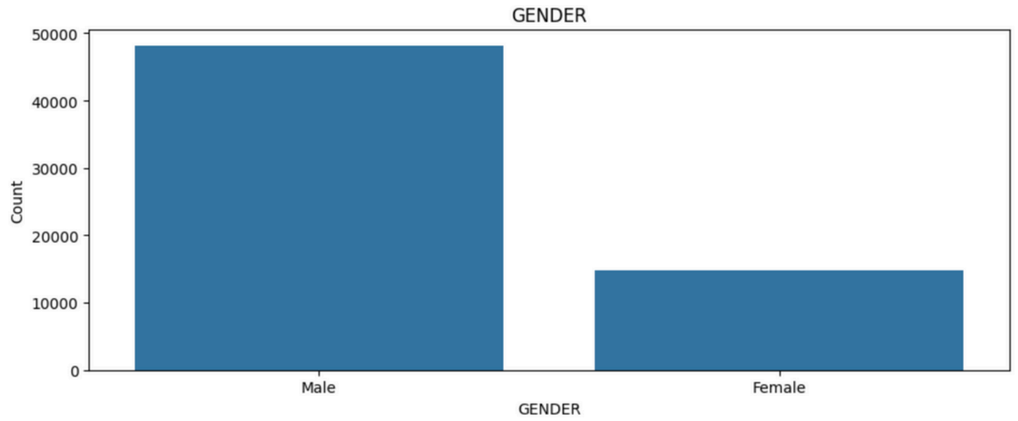
HIST + KDE PLOT

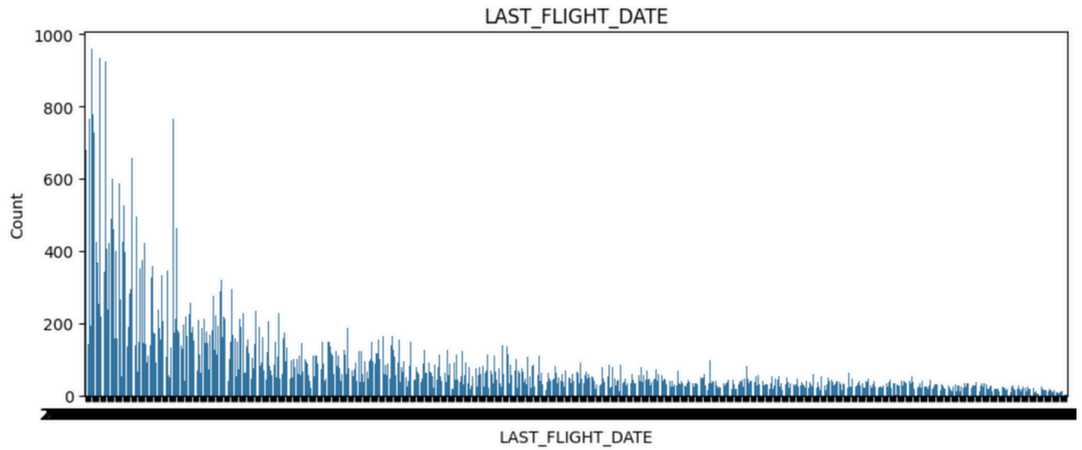
- The majority of customers are at type 4 in the Frequent Flyer program.
- Most customers are between 30 and 50 years old, with a peak around 40 years old.
- Most customers have fewer than 50 flights, with some outliers having a very high number of flights.
- Most customers travel relatively short distances, with some customers traveling very long distances.
- The majority of numerical features showing positive skewness.



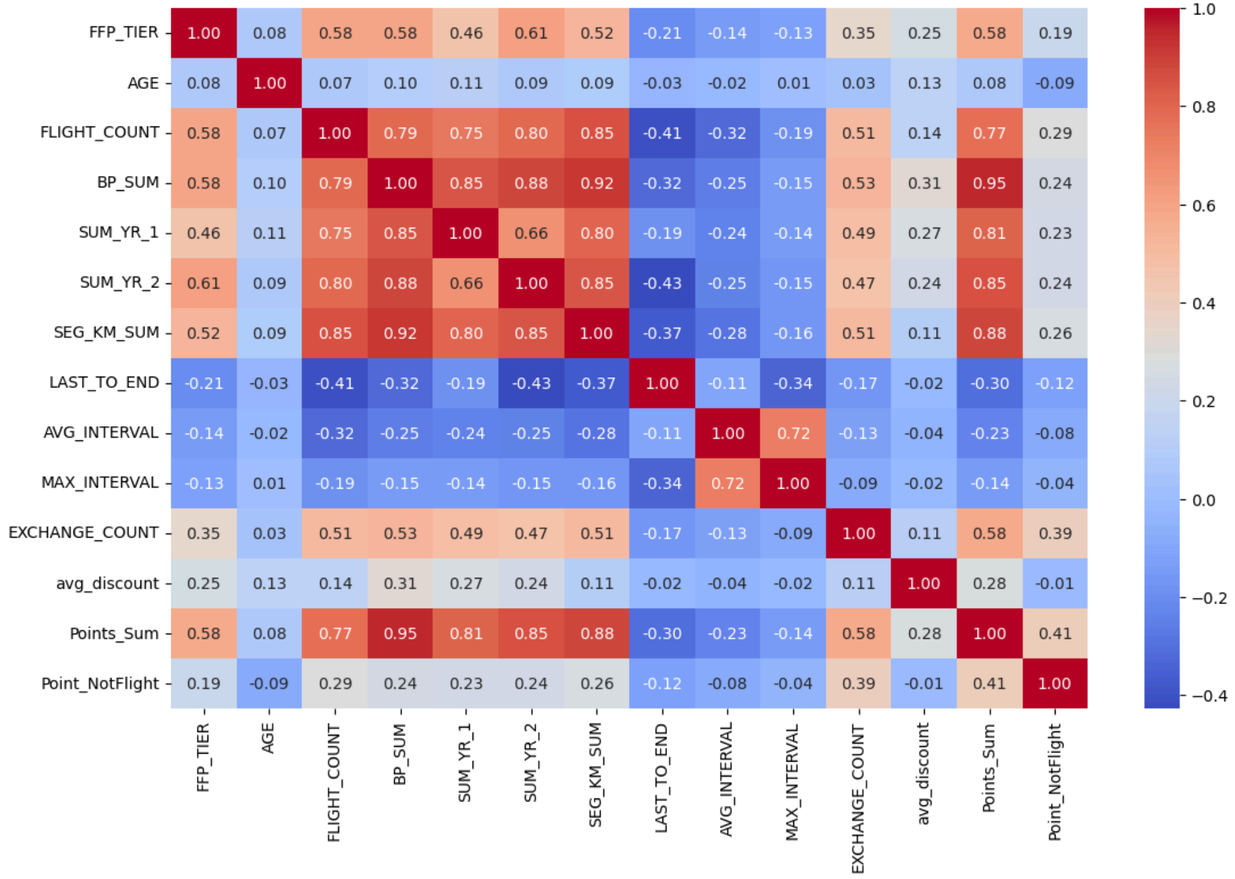
COUNT PLOT

- The majority of airline users on this dataset are male.
- `LAST_FLIGHT_DATE` shows a declining trend over time. My insight: This decline could be due to an increase in the number of customers who no longer use this airline.





MULTICOLLINEARITY CHECK



There are features with high correlation number with each other.

- `FLIGHT_COUNT`, `BP_SUM`, `SUM_YR_1`, `SUM_YR_2`, and `SEG_KM_SUM` are cyclic correlated with each other.
- `Points_Sum` are correlated with `FLIGHT_COUNT`, `BP_SUM`, `SUM_YR_1`, `SUM_YR_2`, and `SEG_KM_SUM`.
- `AVG_INTERVAL` is correlated with `MAX_INTERVAL`.

We will leave them because we will refer LRFMC model and won't use those highly correlated features.

LRFMC

L = LENGTH

R = Recency

F = Frequency

M = Monetary

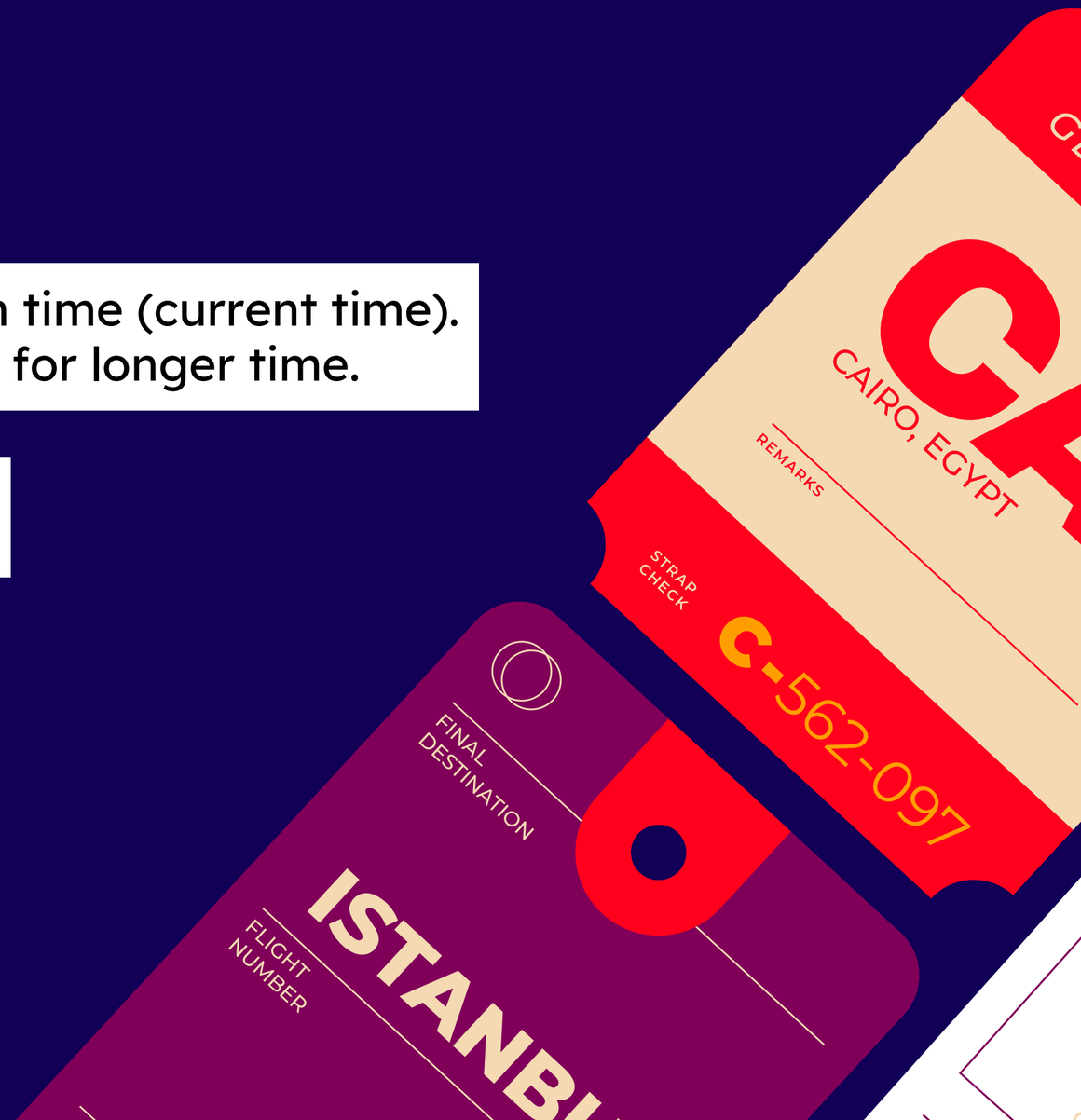
C = Customer Cost

LENGTH

L = LENGTH

Interval of days between register date to observation time (current time).
Larger number means they have been a member for longer time.

`FFP_DATE` - `LOAD_TIME`

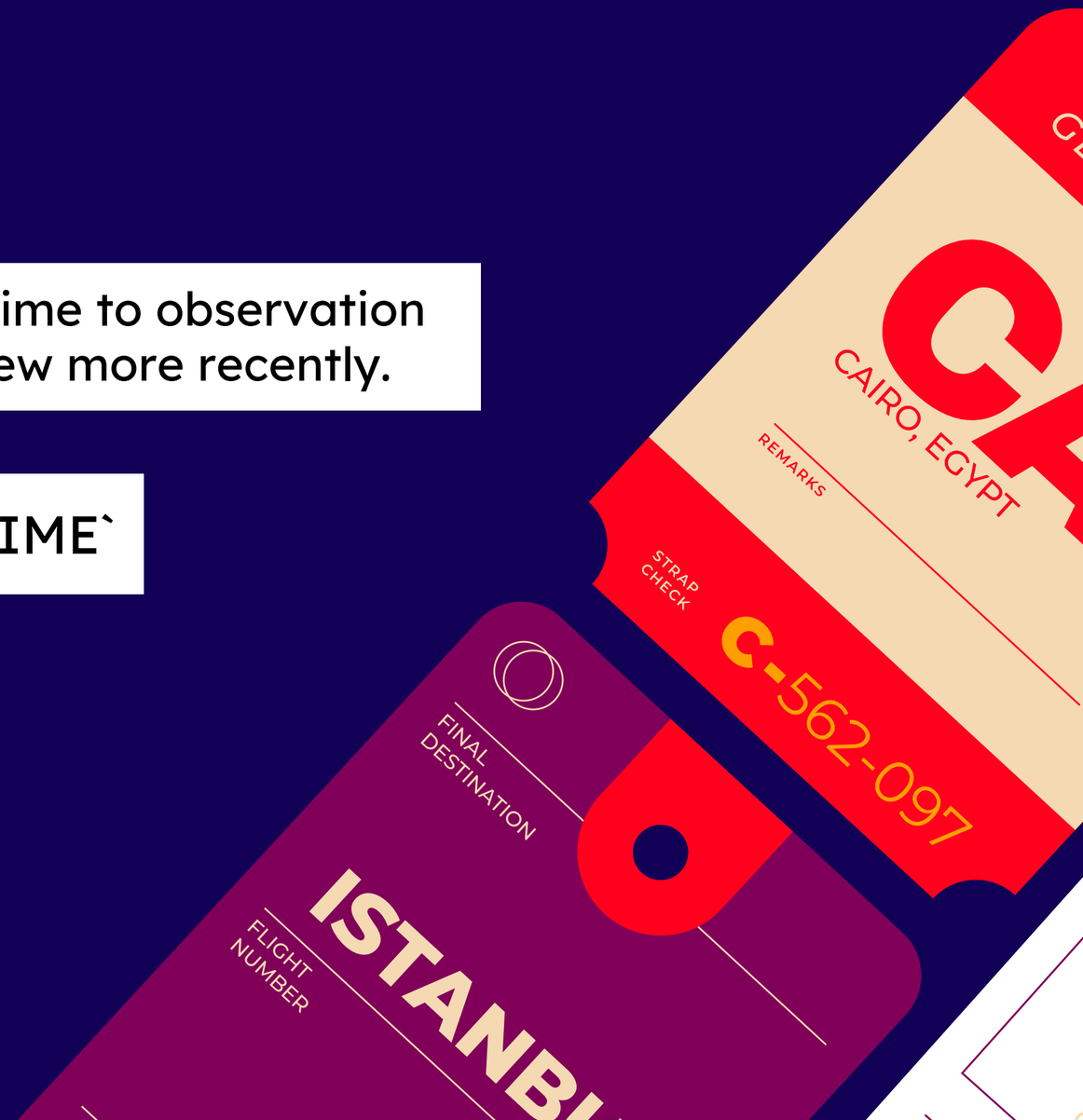


RECENCY

R = RECENCY

The interval of days between the last consumption time to observation time (current time). Smaller number means they flew more recently.

`LAST_FLIGHT_DATE` - `LOAD_TIME`



FREQUENCY

F = Frequency

Number of rides in the observation time window. Bigger number means they flew more frequently.

`FLIGHT_COUNT`

MONETARY

M = Monetary

Number of miles flown within the observation time window. Bigger number means that more flights occurred and/or that those flights covered longer distances

`SEG_KM_SUM`

CUSTOMER COST

C = Customer Cost

Average value of discount factor in the observation time window. Bigger number means they use more discounts.

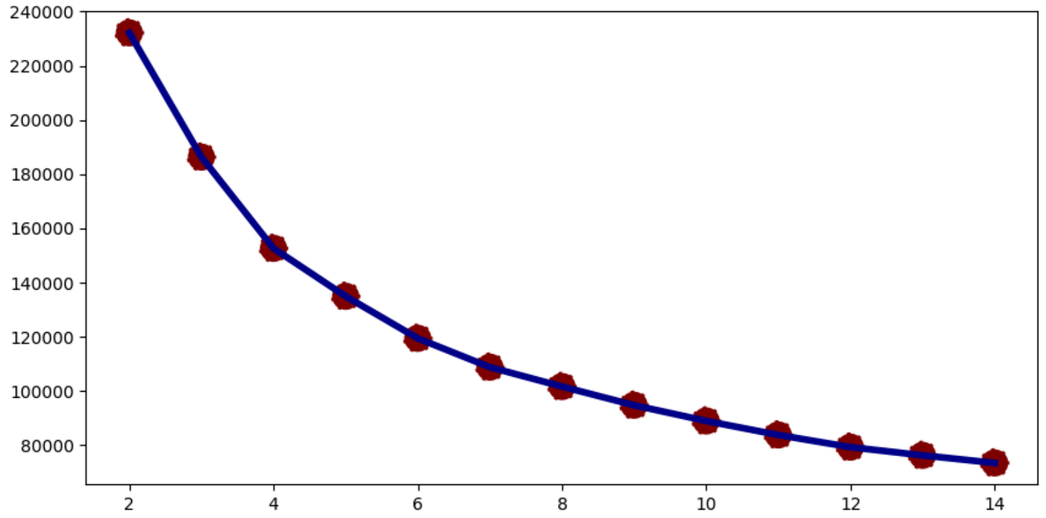
`avg_discount`

K MEANS CLUSTERING

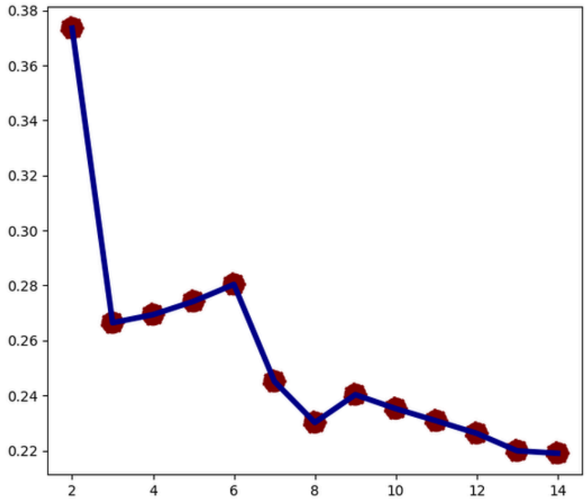
BEST K

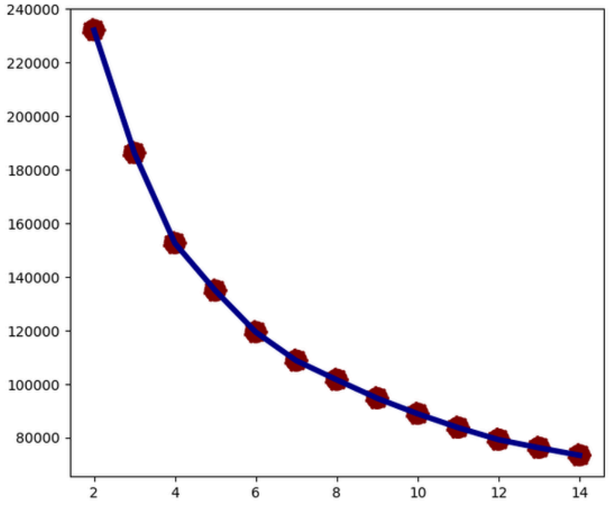
- Based on the elbow method results, clusters ranging from 4 to 6 are potential candidates for the most optimal number of clusters. To confirm this, we will also check using the silhouette method.
- It turns out that the highest silhouette score is achieved with 2 clusters. However, having only 2 clusters may not be very useful for segmenting customers in this case and may not significantly impact business value.
- Other candidates with high silhouette scores and considering the significant drop in inertia from the elbow method are clusters in the following order: 6 -> 5 -> 4.
- Therefore, we will examine clusters with K = 6, K = 5, and K = 4.

Elbow Method



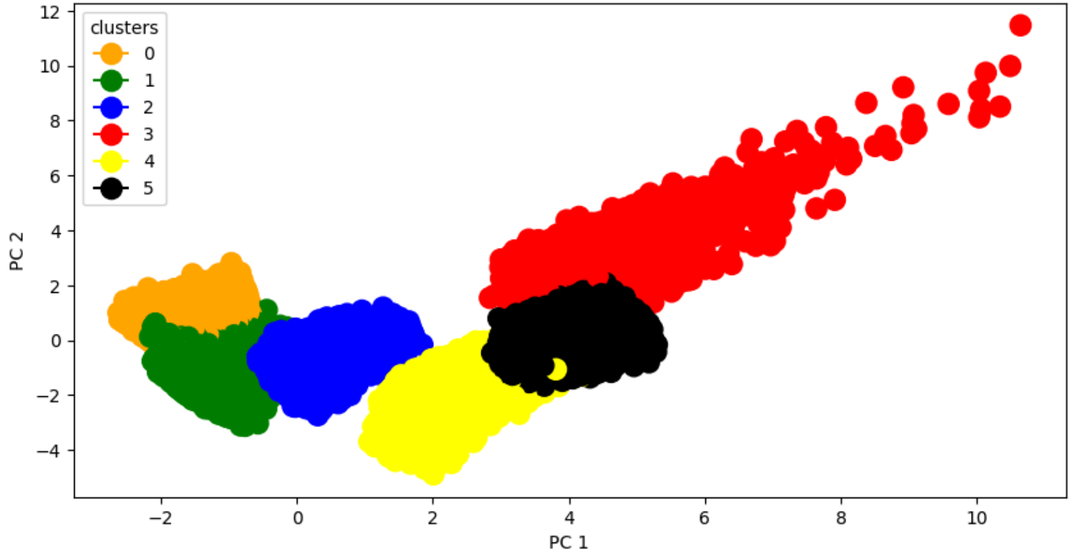
Silhouette Score

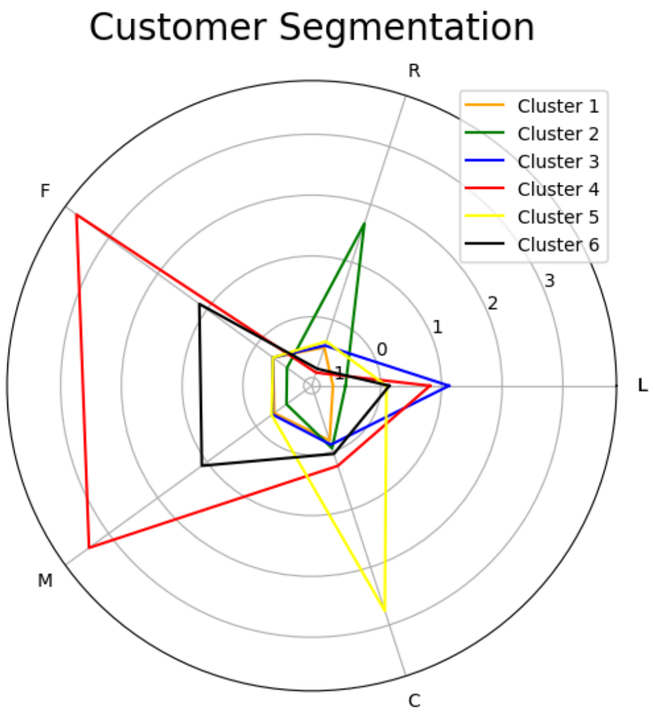




6 CLUSTERS

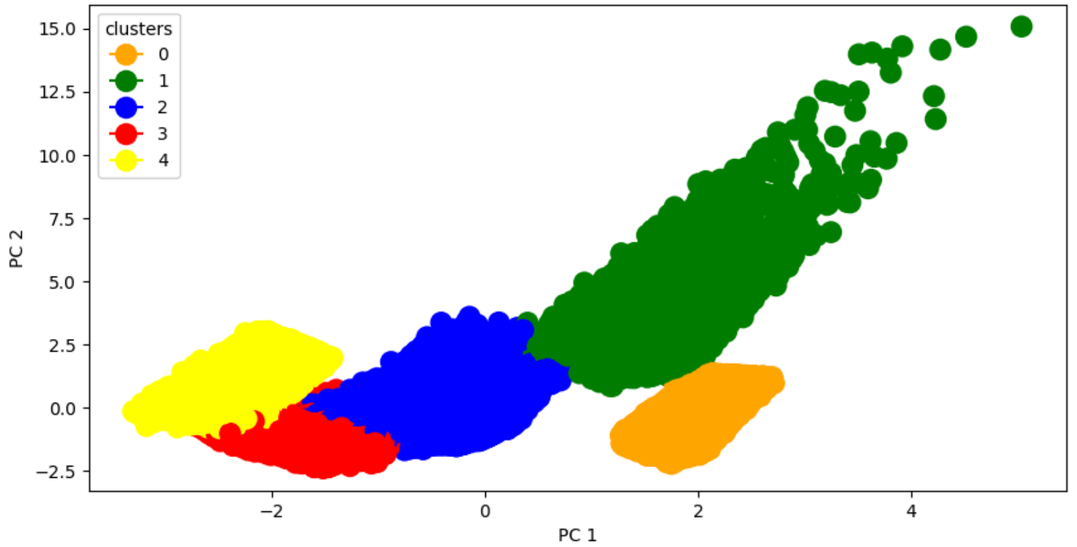
- Cluster with K = 6 has provided a satisfactory segmentation, where each dimension (L, R, F, M, C) has a cluster with high values.
- - However, clusters 6 and 4 overlap, where customers in cluster 6 actually exhibit the same characteristics as cluster 4, but are only differentiated by their ranges. Cluster 6 has lower LRFMC values compared to cluster 4. However, in terms of characteristics, they are the same. Therefore, we will not use cluster K = 6.

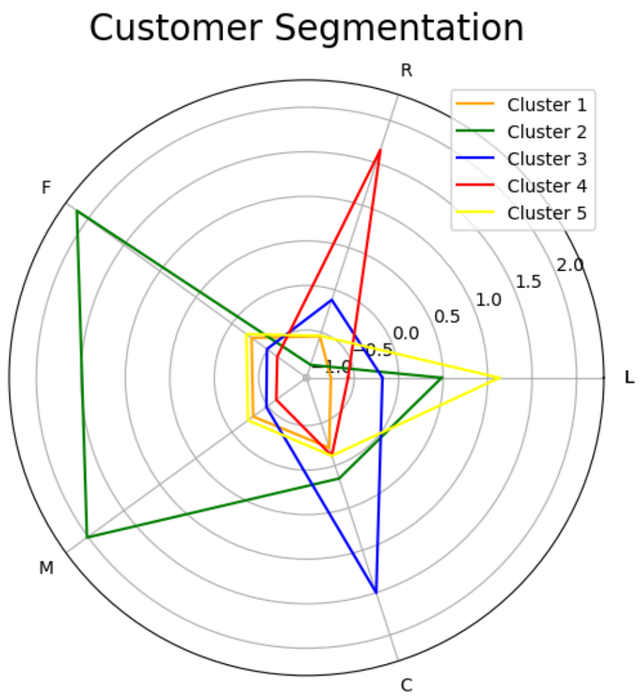




5 CLUSTERS

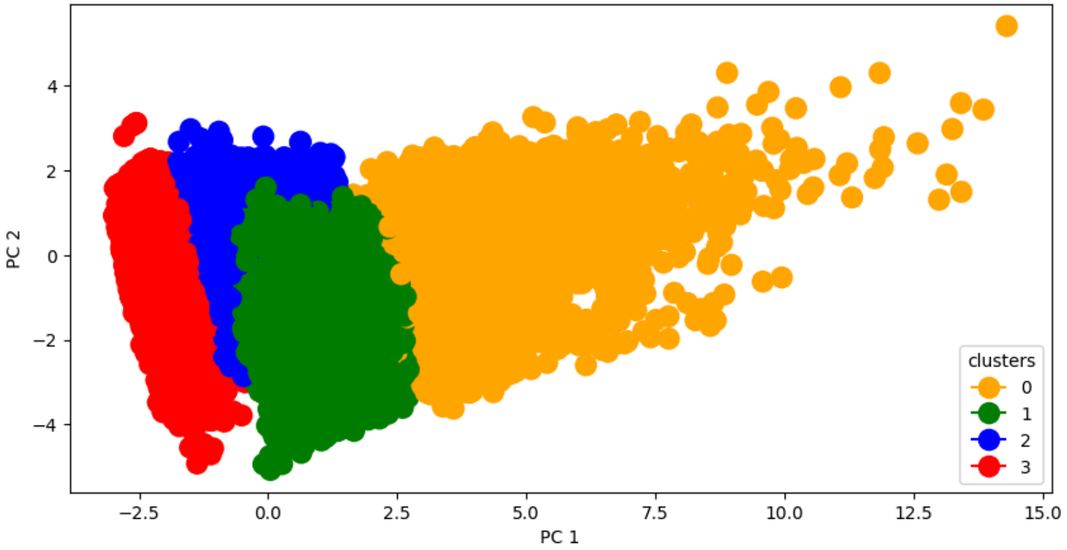
- Cluster with K = 5 has provided a satisfactory segmentation, where each dimension (L, R, F, M, C) has a cluster with high values.
- All clusters also have unique characteristics in customer segmentation, and there is no overlap in characteristics like cluster K = 6. The number of clusters K = 5 is the best candidate for the clustering to be performed in this project.

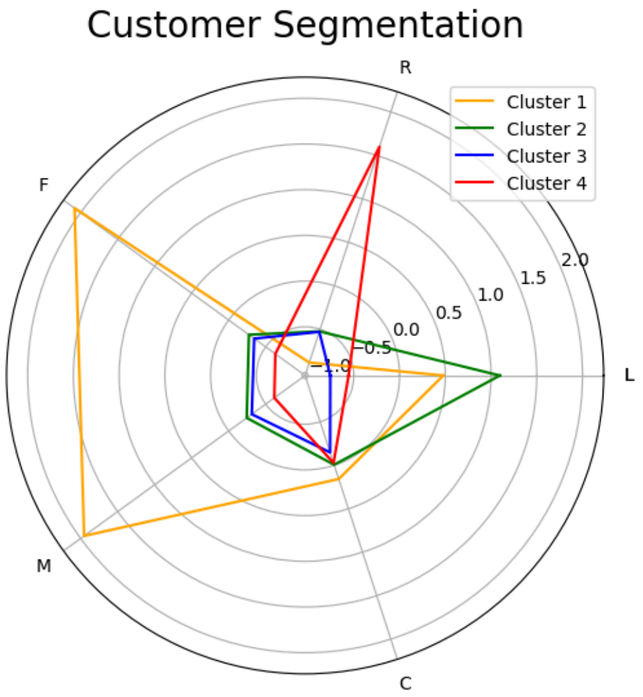




4 CLUSTERS

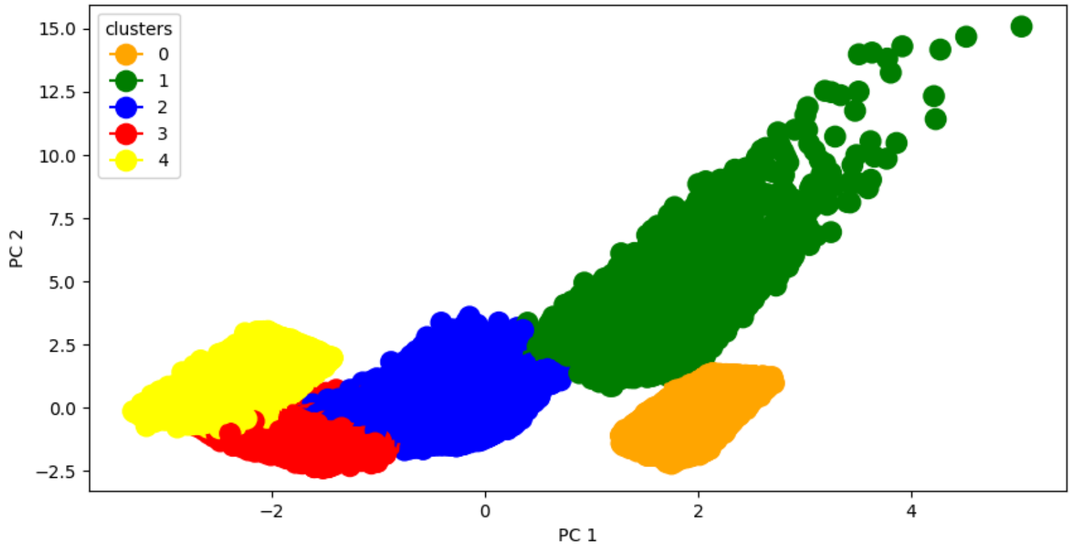
- Cluster with K = 4 does not provide good segmentation results. While L (Length), R (Recency), F (Frequency), and M (Monetary) have clusters with high values, C (Customer Cost) does not have clusters with high values. Therefore, the cluster with K = 4 does not provide good segmentation results and will not be used.

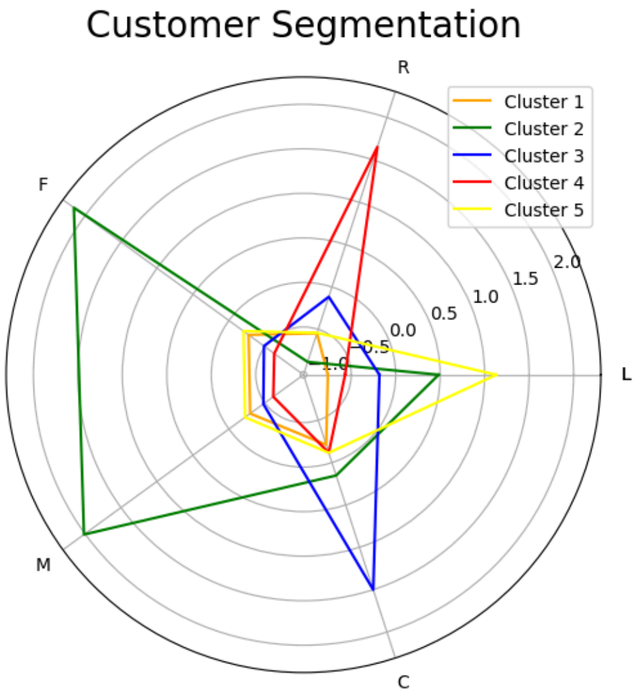




(BEST CLUSTER) 5 CLUSTERS

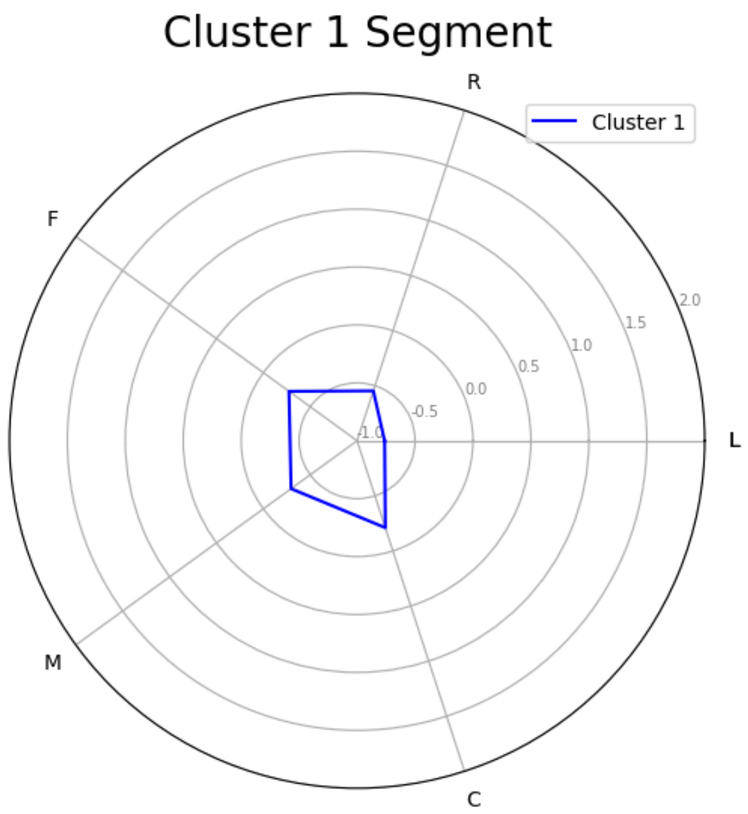
Based on the previous observation results, it was found that K = 5 is the most optimal number of clusters. Therefore, modeling will be conducted using a total of 5 clusters.





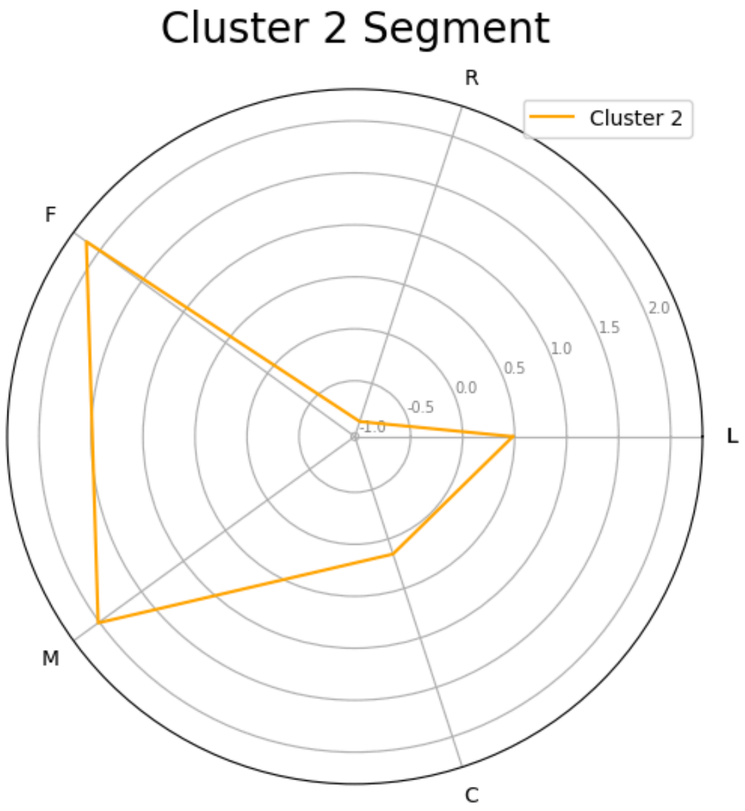
EACH CLUSTER ANALYSIS

CLUSTER 1 (PROSPECTIVE NEW CUSTOMERS)



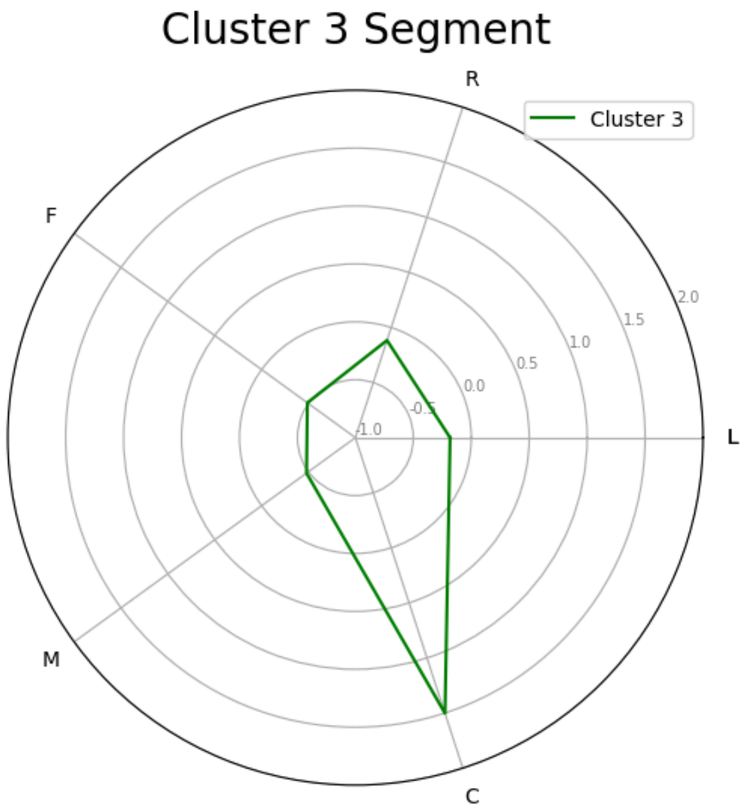
- Cluster 1 has the smallest values of LRFMC compared to other clusters. This cluster can be interpreted as a group of customers who have recently joined the airline.
- The Customer Cost (C) in this cluster is the lowest because they have not yet used the services as much as other clusters. However, it is still considered normal as it is not significantly lower than other clusters and is even competitive.
- The Frequency (F) and Monetary (M) values in this cluster rank third and are almost equivalent to the second-ranked cluster (cluster 5). This indicates that these new users are quite frequent flyers and have a high number of miles (which correlates with high transaction values for the business), making their contribution quite significant to the business revenue
- Cluster 1 can be characterized as **"Prospective New Customers"**.

CLUSTER 2 (LOYAL CUSTOMERS)



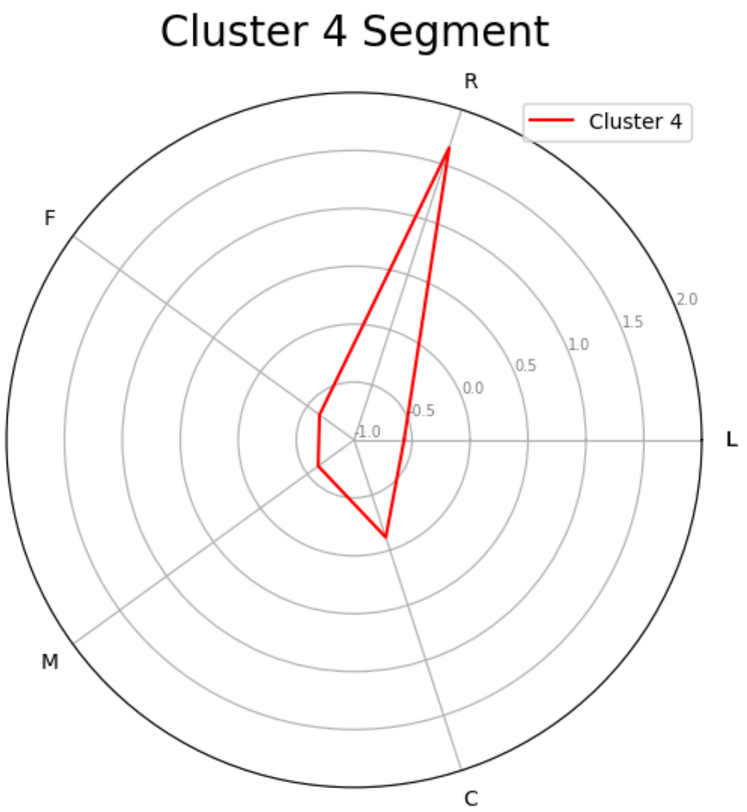
- Cluster 2 has the second highest value of L (Length), with F (Frequency) and M (Monetary) values significantly higher than other clusters. Customers in this cluster have been using the airline's services for a long time and are very active, with a high number of miles. The contribution of customers in cluster 2 to this business is very significant and important.
- The R (Recency) value of this cluster is the smallest, indicating that customers in this cluster recently used the services (more smaller the recency, more recent they used the services).
- Cluster 2 can be characterized as **"Loyal Customers"**.

CLUSTER 3 (DISCOUNT SEEKERS)



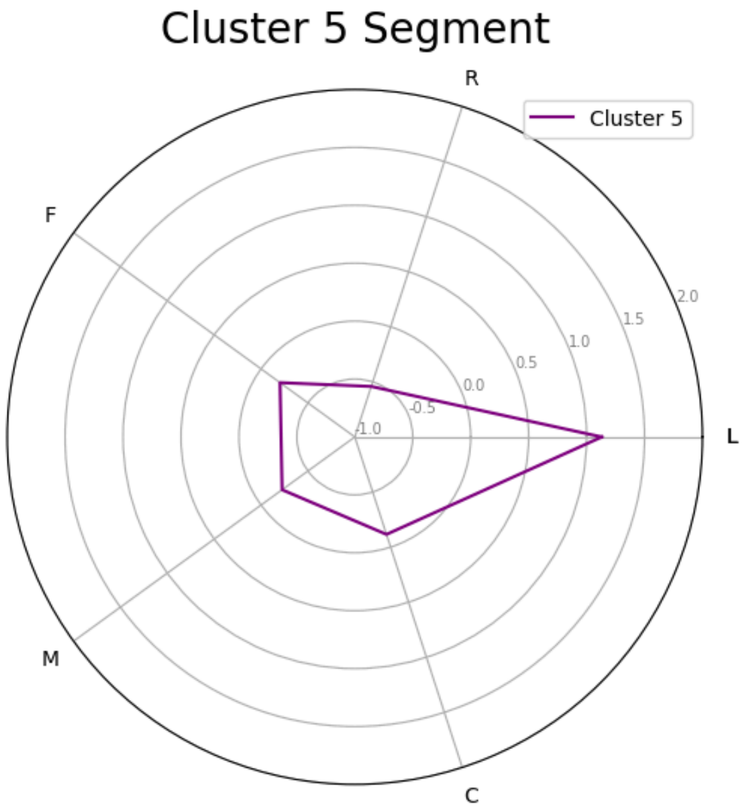
- Cluster 3 has a high value of C (Customer Cost). This indicates that cluster 3 is a segment of customers who frequently use discounts.
- The values of F (Frequency) and M (Monetary) for customers in cluster 3 are very low. This indicates that customers in cluster 3 always take advantage of discounts. It can be said that cluster 3 has a very small contribution to the company.
- Cluster 3 can be characterized as **"Discount Seekers"**.

CLUSTER 4 (DORMANT CUSTOMERS)



- Cluster 4 ranks fourth out of 5 clusters in terms of L (Length), indicating that customers in cluster 4 are relatively new members.
- One important observation is the Recency (R) value of cluster 4, which ranks first and is relatively high. This suggests that customers in cluster 4 have not used the airline's services for quite some time compared to other clusters.
- The values of F (Frequency) and M (Monetary) for cluster 4 are the lowest (ranked 5 out of 5 clusters). Customers in cluster 4 have minimally used the airline's services since joining the membership.
- The Customer Cost (C) value for cluster 4 is average (ranked 3 out of 5 clusters).
- This should be a point of concern, as customers in cluster 4 may be relatively new but are dissatisfied with certain factors, such as service or quality, leading them to rarely or stop using the airline for specific reasons.
- Cluster 4 can be characterized as **"Dormant Customers"**.

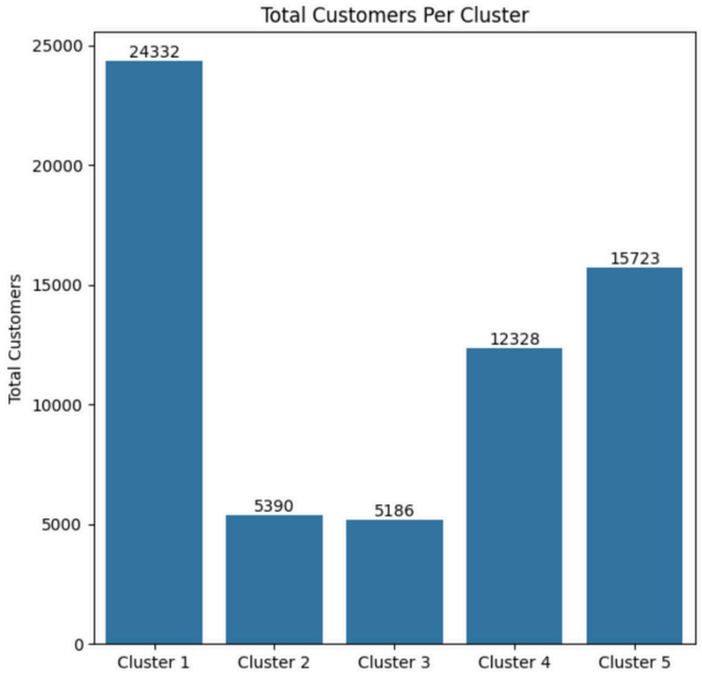
CLUSTER 5
(VETERAN CUSTOMERS WITH OCCASIONAL TRANSACTIONS)

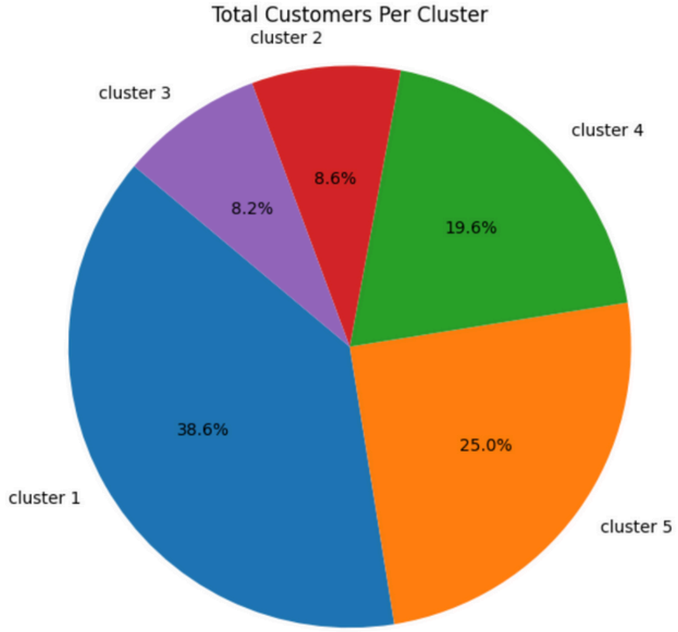


- Cluster 5 has the highest value of L (Length) among the 5 clusters, indicating that customers in cluster 5 are the longest-standing users.
- The values of R (Recency), F (Frequency), M (Monetary), and C (Customer Cost) for cluster 5 are average.
- Cluster 5 can be characterized as "Veteran Customers with Occasional Transactions".

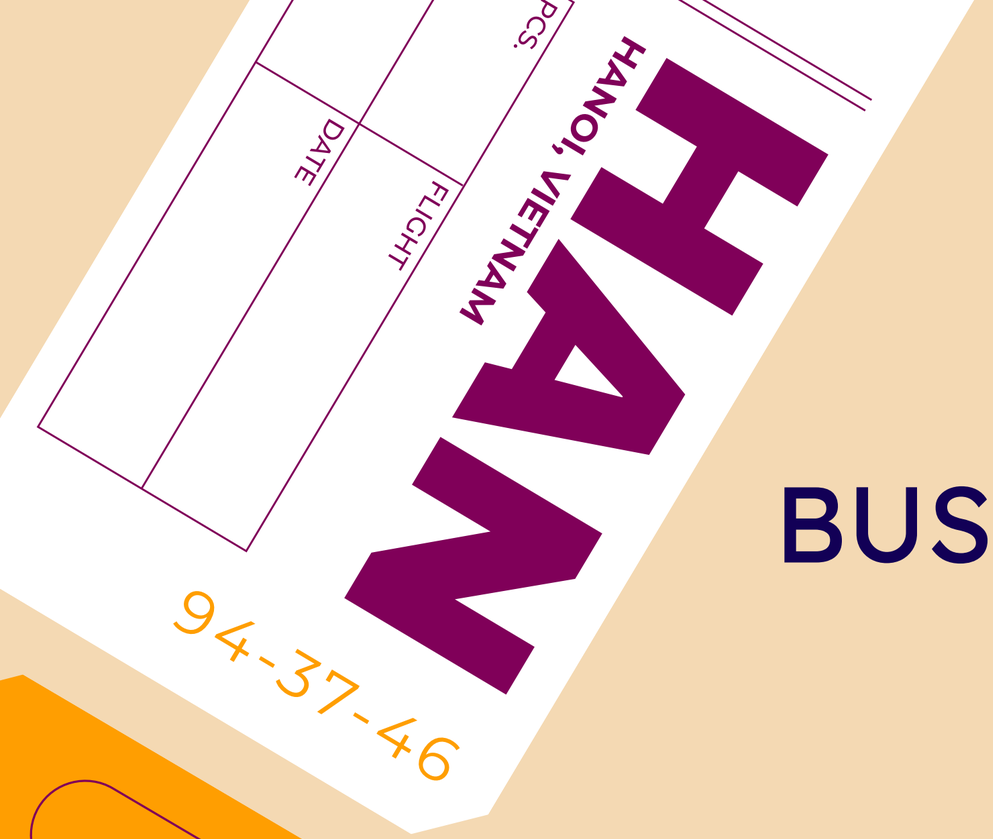
CLUSTERS OBSERVATIONS

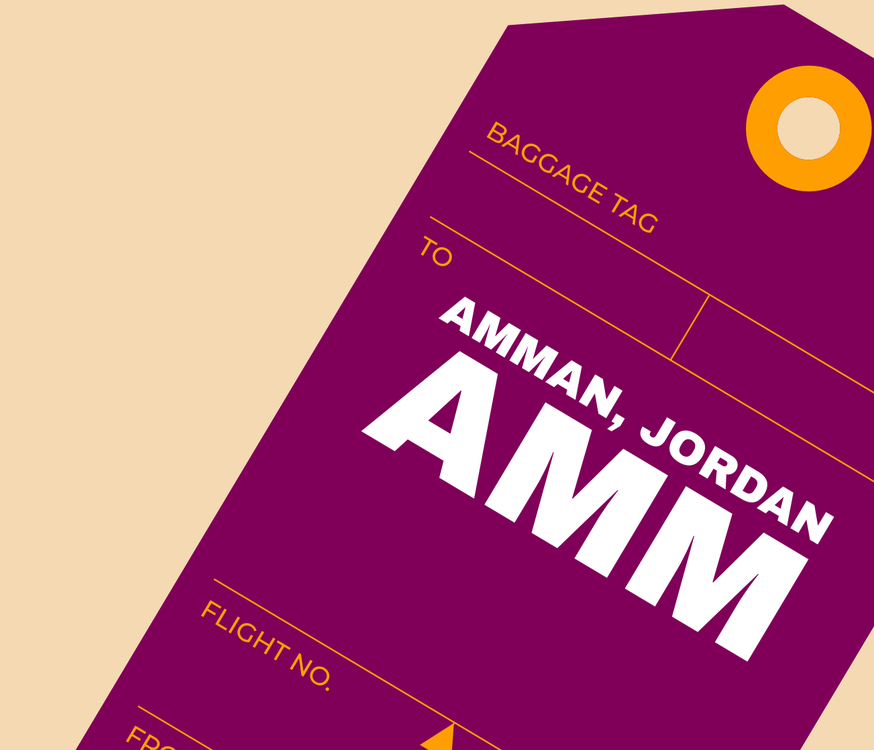
- **Cluster 1, or "Prospective New Customers"**, has the highest total of **24,332 customers, ranking 1st** out of the total number of customers.
- **Cluster 2, or "Loyal Customers"**, has a total of **5,390 customers, ranking 4th** out of the total number of customers.
- **Cluster 3, or "Discount Seekers"**, has a total of **5,186 customers, ranking 5th** out of the total number of customers.
- **Cluster 4, or "Dormant Customers"**, has a total of **12,328 customers, ranking 3rd** out of the total number of customers.
- **Cluster 5, or "Veteran Customers with Occasional Transactions"**, has a total of **15,723 customers, ranking 2nd** out of the total number of customers.





BUSINESS RECOMMENDATIONS

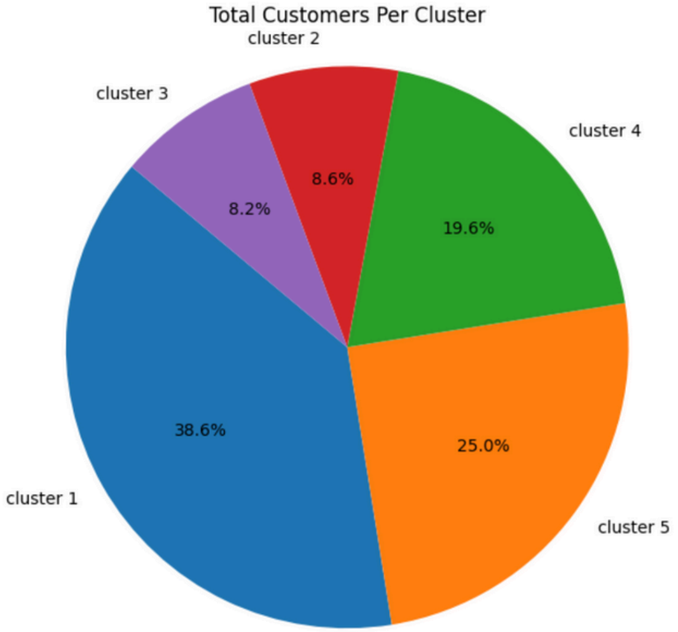


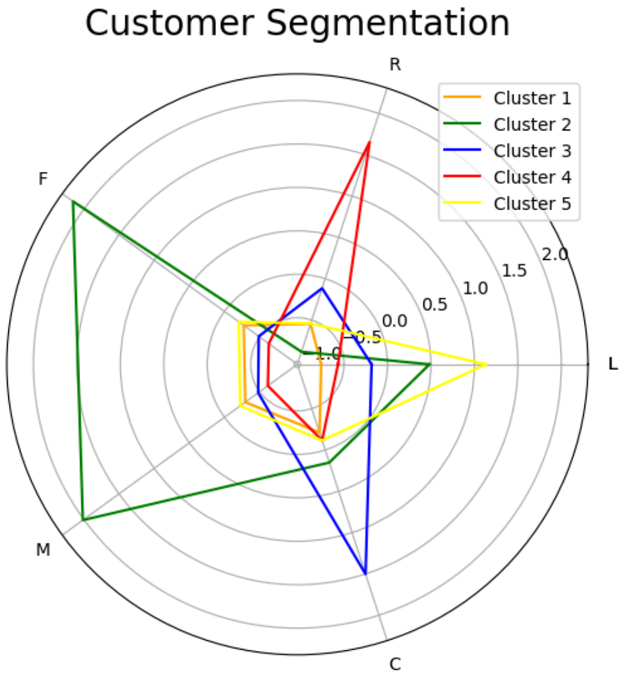


RECOMMENDATION 1

Cluster 1, or "Prospective New Customers", accounts for 38.6% of the population and represents the majority. Looking at their F (Frequency) and M (Monetary) values, they closely resemble cluster 5 ("Veteran Customers with Occasional Transactions"), indicating that cluster 1 is highly prospective for the company.

The company should devise strategies to increase the consumption level of these "Prospective New Customers" (encouraging them to fly more with the airline). This can be achieved by offering more promotions and privileges to new customers, making them feel comfortable and fostering loyalty to the airline.

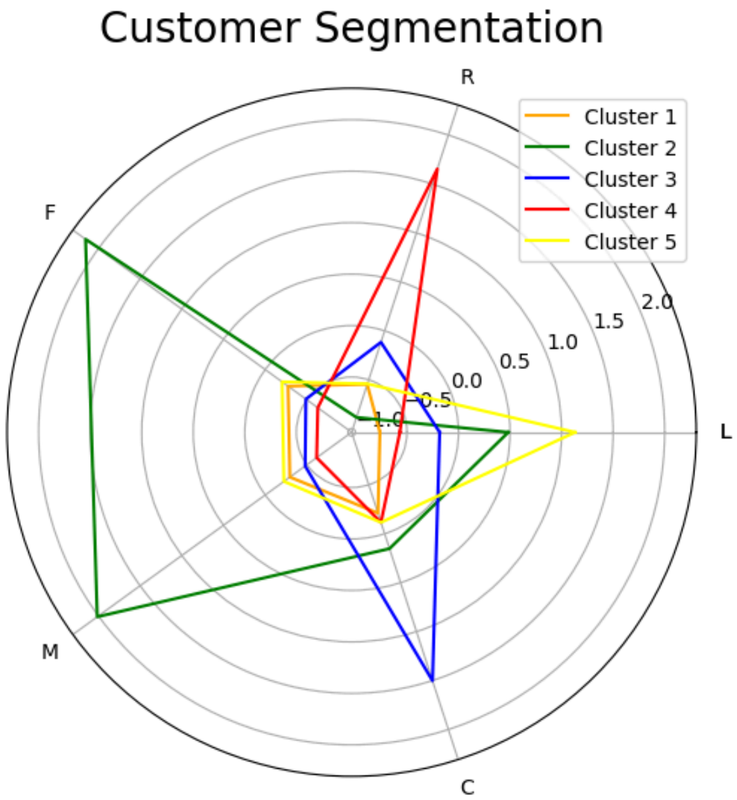




RECOMMENDATION 2

Despite being recent joiners, the consumption level of **cluster 1, or "Prospective New Customers"**, is quite good, as seen from their F (Frequency) and M (Monetary) values. They should be retained to become loyal customers of the company. It's essential to ensure they don't prioritize another airline or stop using this one (Target: R or Recency should be low).

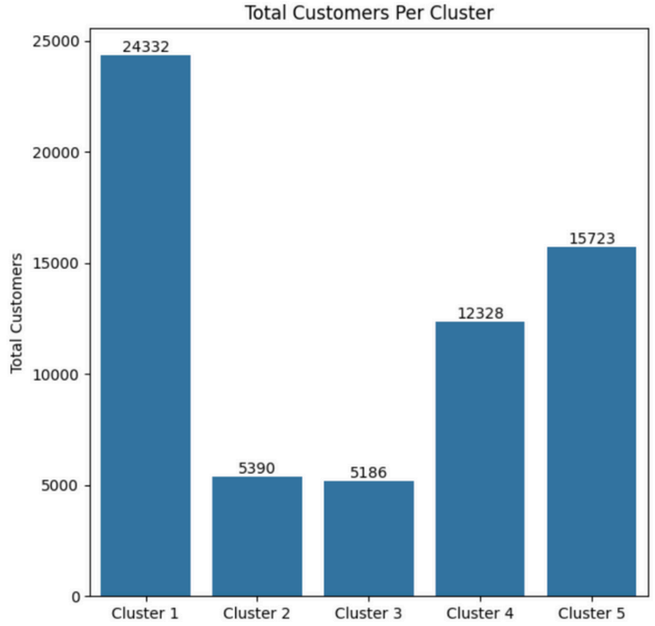
The airline may frequently offer membership benefit programs to make customers in cluster 1 feel special or valued, encouraging loyalty to the company.

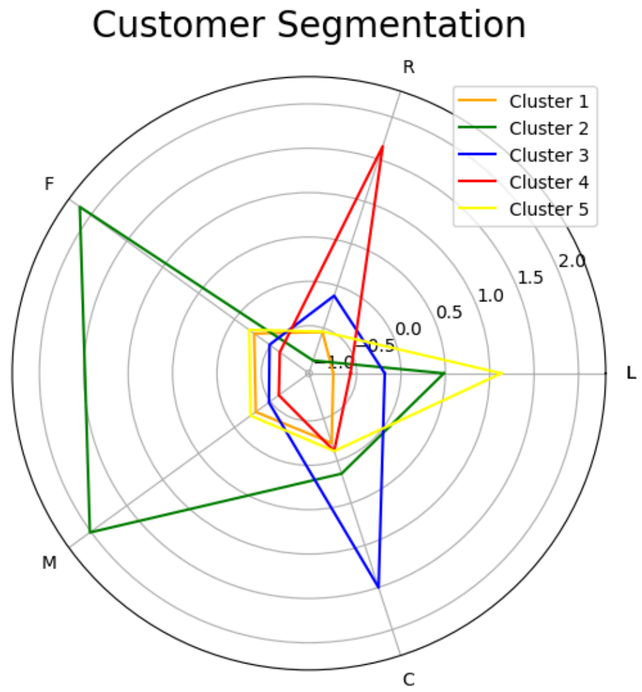


RECOMMENDATION 3

Cluster 2, or "Loyal Customers", occupies the 4th position with 8.6% of the total population. Customers in this segment are significant contributors to this airline company. The company must not lose them.

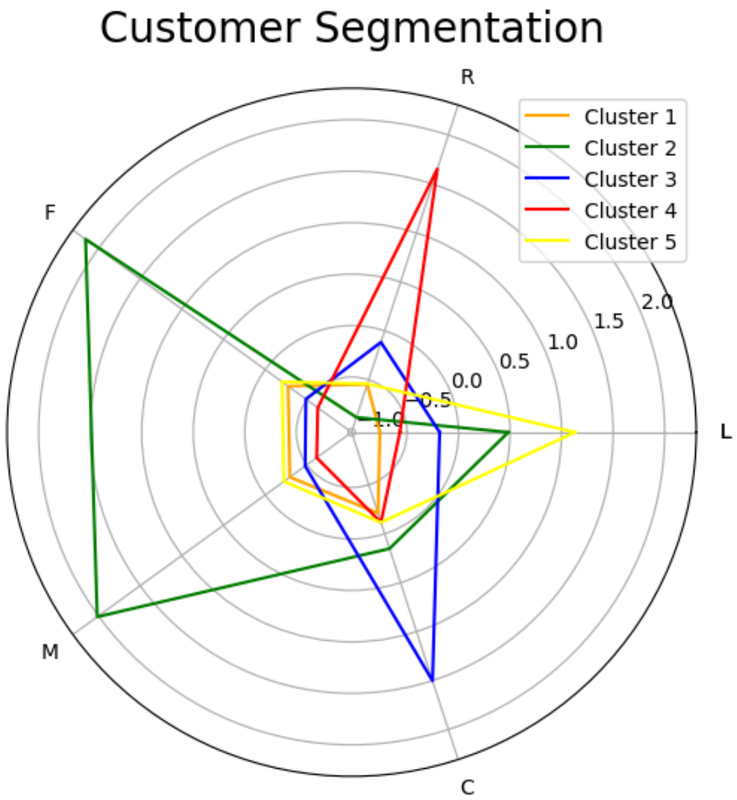
The company can enhance or expand its loyalty program to provide more incentives to loyal customers. This could include special offers or exclusive gifts to strengthen the emotional connection with customers.





RECOMMENDATION 4

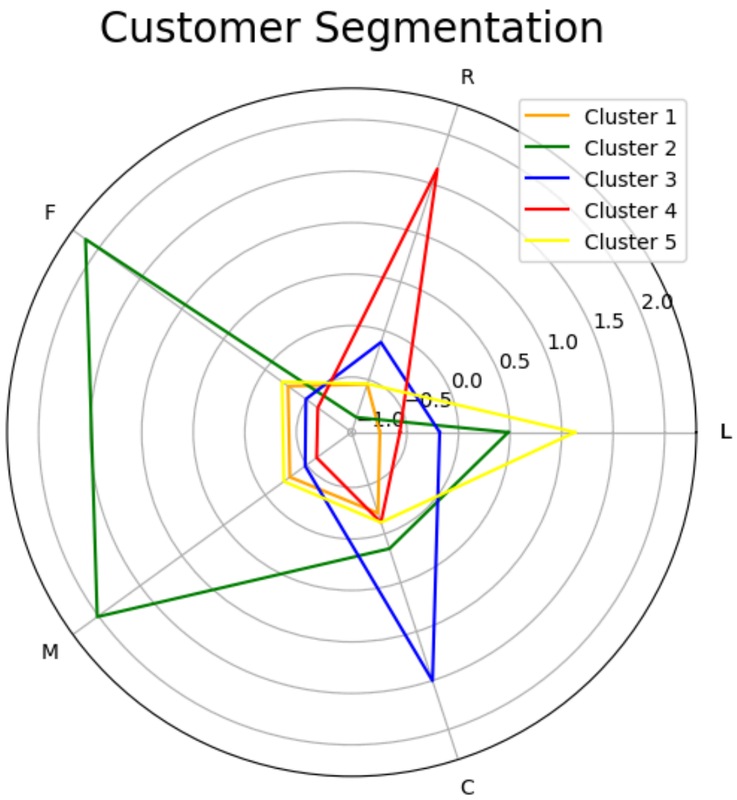
Cluster 2, or "Loyal Customers", undoubtedly represents customers who are fond of this company. This can be leveraged to expand the company's user base by introducing a referral program. This program can encourage loyal customers to refer their friends and relatives to use the company's services, serving as an effective way to expand the customer base. The company can also provide incentives to referring customers, such as additional discounts or special rewards.



RECOMMENDATION 5

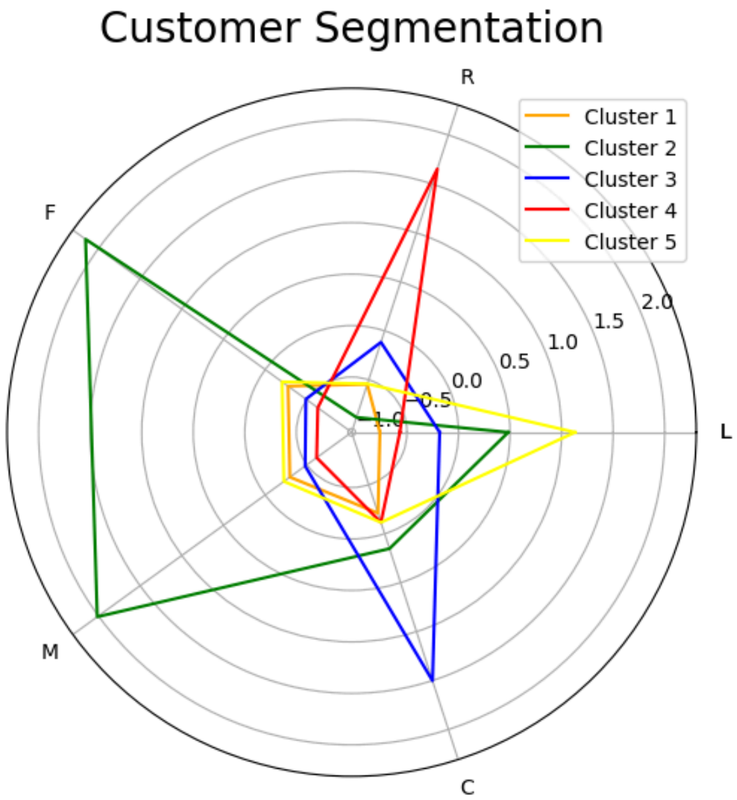
Cluster 3, or "Discount Seekers", consists of customers with high values of C (Customer Cost) but relatively low values of F (Frequency) and M (Monetary).

From these characteristics, the company can consider implementing a progressive discount system for customers in cluster 3. Progressive discounts could be based on purchase frequency or accumulation based on flight miles. This is expected to increase the shopping value of customers in cluster 3 and reduce dependence on shopping solely based on discounts.



RECOMMENDATION 6

The company can also implement a discount system based on bundling for **Cluster 3, or "Discount Seekers"**. Discounts are only provided if customers purchase bundled flight tickets (Round Trip), or discounts can also be given for ticket upgrades from regular to business/first class. This approach can increase the transaction value of customers in this cluster.

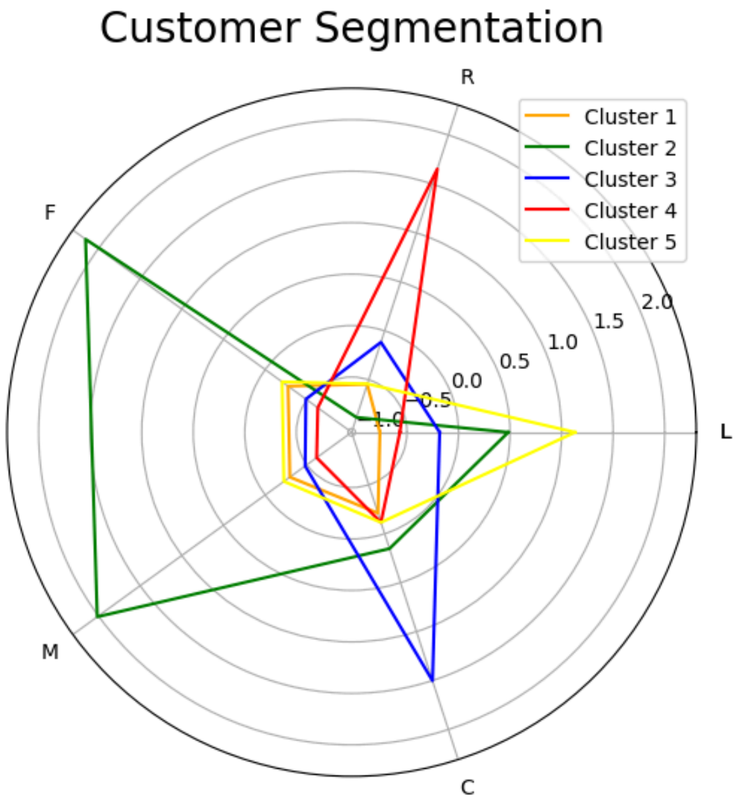


RECOMMENDATION 7

Cluster 4, or "Dormant Customers", consists of customers with high values of R (Recency). Customers in this cluster have not used the airline's services for a long time, or have even uninstalled the airline's app for ticket purchases.

The company may consider offering "welcome back" discounts or other benefits to this cluster.

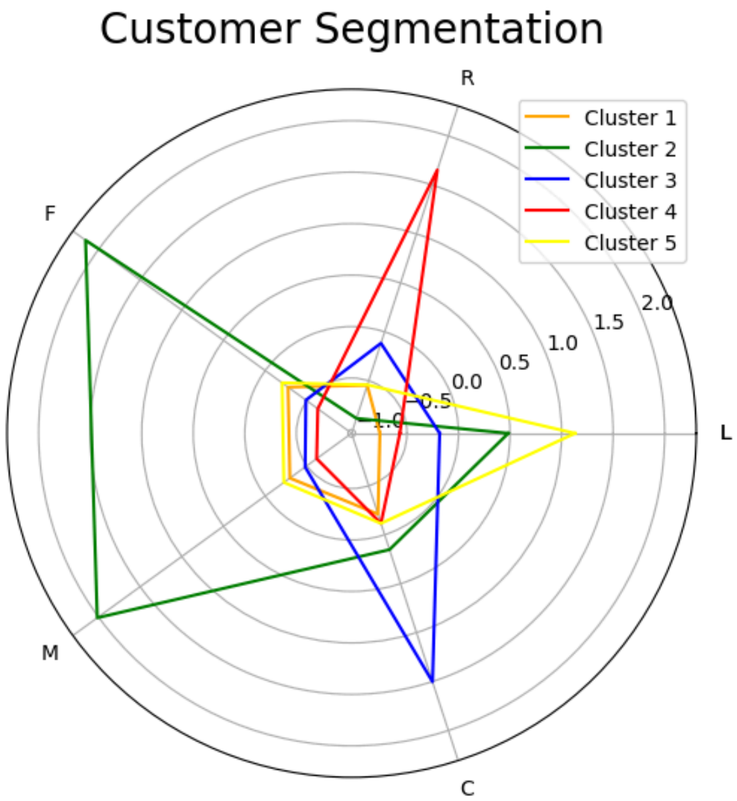
In addition to offering discounts, the information about discounts must reach the customers. Channels that can be used to send discount informations to this cluster include social media, email, SMS, or pop-up messages (useful if the customer has not uninstalled the app).



RECOMMENDATION 8

Cluster 5, or "Veteran Customers with Occasional Transactions", comprises customers with high values of L (Length) but average values of R (Recency), F (Frequency), and M (Monetary).

For cluster 5, a deeper analysis can be conducted regarding the months when customers in this cluster make ticket bookings. Once the booking pattern is understood, the company can consider offering discounts or special membership benefits during months outside their usual booking pattern. This is done to encourage customers in cluster 5 to increase their consumption beyond their usual habits. This can provide significant benefits to the company, as this cluster accounts for 25% of the total population.



THANK YOU


WEIGHT
SEOUL
SOUTH KOREA
FLIGHT NO.
DATE
90-46-28
SEOUL
90-46-28
MANILA, PHILIP
FLIGHT NUMBER
PRINTED IN DUBAI
DCG
SA
BAGGAGE TAG
TO:
DUBAI, UAE
DC
FLIGHT NO.
PCS.
AIR TRAVEL EXPRESS
WT.
PRG
CZECH REPUBLIC
FLIGHT
TO
PRAGUE
AA
ABC AI