Airline Customer Value Analysis through LRFMC Indicators by Performing K-Means Clustering Algorithm

By Selpha Yulida

BACK GROUND

Fierce business competition -
shift from a product-centric approach to a customer-centric approach

Create better relationships with customers -
will lead to increased company profits

Classify customers accurately - formulate personalized service plans

Establish a reasonable customer value evaluation model

Objectives

Customer Value Analysis based on LRFMC Indicators

1. Perform customer segmentation (clustering) through the airline customer dataset. Use LRFMC indicators and perform K-Means clustering algorithm.
2. Analysis of the characteristics of each cluster resulting from segmentation.
3. Provide business insight related to the analysis results.

The Steps

1. Data collection.
2. Data understanding by performing the statistic descriptive of data and check correlation among features by correlation matrix.
3. Data preprocessing that includes handling missing value, feature selection based on LRFMC, handling outlier and also data scaling.
4. Clustering process by K-Means algorithm
5. Analyze the result - customer value analysis

About the Data

The dataset consists of
62988 rows and 23 columns

```
1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 62988 entries, 0 to 62987
Data columns (total 23 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   MEMBER_NO          62988 non-null  int64
 1   FFP_DATE           62988 non-null  object
 2   FIRST_FLIGHT_DATE  62988 non-null  object
 3   GENDER             62985 non-null  object
 4   FFP_TIER           62988 non-null  int64
 5   WORK_CITY          60719 non-null  object
 6   WORK_PROVINCE      59740 non-null  object
 7   WORK_COUNTRY       62962 non-null  object
 8   AGE                62568 non-null  float64
 9   LOAD_TIME          62988 non-null  object
 10  FLIGHT_COUNT       62988 non-null  int64
 11  BP_SUM             62988 non-null  int64
 12  SUM_YR_1           62437 non-null  float64
 13  SUM_YR_2           62850 non-null  float64
 14  SEG_KM_SUM         62988 non-null  int64
 15  LAST_FLIGHT_DATE   62988 non-null  object
 16  LAST_TO_END        62988 non-null  int64
 17  AVG_INTERVAL       62988 non-null  float64
 18  MAX_INTERVAL       62988 non-null  int64
 19  EXCHANGE_COUNT     62988 non-null  int64
 20  avg_discount       62988 non-null  float64
 21  Points_Sum         62988 non-null  int64
 22  Point_NotFlight    62988 non-null  int64
dtypes: float64(5), int64(10), object(8)
memory usage: 11.1+ MB
```

Code	Description
MEMBER_NO	: Member ID
FFP_DATE	: Frequent Flyer Program Join Date
FIRST_FLIGHT_DATE	: Date of First Flight
GENDER	: Gender
FFP_TIER	: Tier of Frequent Flyer Program
WORK_CITY	: City of Origin
WORK_PROVINCE	: Province of Origin
WORK_COUNTRY	: Country of Origin
AGE	: Customer Age
LOAD_TIME	: Date Data was Taken
FLIGHT_COUNT	: Number of Customer Flights
BP_SUM	: Travel Plans
SUM_YR_1	: Fares Revenue
SUM_YR_2	: Votes Prices
SEG_KM_SUM	: Total Distance (Km) Flights that have been done
LAST_FLIGHT_DATE	: Date of Last Flight
LAST_TO_END	: Time Range Between the Last Flight to the Most Recent Flight Booking
AVG_INTERVAL	: Average Time Interval
MAX_INTERVAL	: Maximum Time Interval
EXCHANGE_COUNT	: Exchange Count
avg_discount	: The Average Discount that Customers Get
Points_Sum	: The Total Points that Earned by Customer
Point_NotFlight	: Points not Used by Members

DATA UNDERSTANDING

Statistic Descriptive

Numerical Data Type

- There are 14 numeric columns after dropping `MEMBER_NO` column
- There are some columns that seem to have a normal distribution which mean equal/close to median value. Those columns are `FFP_TIER`, `AGE`, `avg_discount`.
- Other columns seem to have a positive skew distribution which mean > median value.
- Some columns have value of 0. For the example it can be seen in the `SUM_YR_1` and `SUM_YR_2` columns which is a bit strange if fare with 0 value. So those kind of columns need to be checked further.
- `AGE` (Customer Age) column ranged between 6-110, with mean 42yo and median 41yo. A bit strange that there are over 100yo. Looks like it needs to be checked further if required to analyze using that data.

	FFP_TIER	AGE	FLIGHT_COUNT	BP_SUM	SUM_YR_1	SUM_YR_2	SEG_KM_SUM	LAST_TO_END	AVG_INTERVAL	MAX_INTERVAL	EXCHANGE_COUNT	avg_discount	Points_Sum	Point_NotFlight
count	62988.000000	62568.000000	62988.000000	62988.000000	62437.000000	62850.000000	62988.000000	62988.000000	62988.000000	62988.000000	62988.000000	62988.000000	62988.0000	62988.000000
mean	4.102162	42.476346	11.839414	10925.081254	5355.376064	5604.026014	17123.878691	176.120102	67.749788	166.033895	0.319775	0.721558	12545.7771	2.728155
std	0.373856	9.885915	14.049471	16339.486151	8109.450147	8703.364247	20960.844623	183.822223	77.517866	123.397180	1.136004	0.185427	20507.8167	7.364164
min	4.000000	6.000000	2.000000	0.000000	0.000000	0.000000	368.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.0000	0.000000
25%	4.000000	35.000000	3.000000	2518.000000	1003.000000	780.000000	4747.000000	29.000000	23.370370	79.000000	0.000000	0.611997	2775.0000	0.000000
50%	4.000000	41.000000	7.000000	5700.000000	2800.000000	2773.000000	9994.000000	108.000000	44.666667	143.000000	0.000000	0.711856	6328.5000	0.000000
75%	4.000000	48.000000	15.000000	12831.000000	6574.000000	6845.750000	21271.250000	268.000000	82.000000	228.000000	0.000000	0.809476	14302.5000	1.000000
max	6.000000	110.000000	213.000000	505308.000000	239560.000000	234188.000000	580717.000000	731.000000	728.000000	728.000000	46.000000	1.500000	985572.0000	140.000000

Statistic Descriptive

Object & DateTime Data Type

	GENDER	WORK_CITY	WORK_PROVINCE	WORK_COUNTRY
count	62985	60719	59740	62962
unique	2	3234	1165	118
top	Male	guangzhou	guangdong	CN
freq	48134	9386	17509	57748

- Most of customer of this airline is Male which is around 76.42%.
- The top city of origin is Guangzhou which is 15.46%.
- The top province of origin is Guangdong which is 29.31%.
- The top country of origin is CN (China) which is 91.72%.
- Guangzhou is a city in Guangdong Province - China.
- Seems that this is China's airline data.

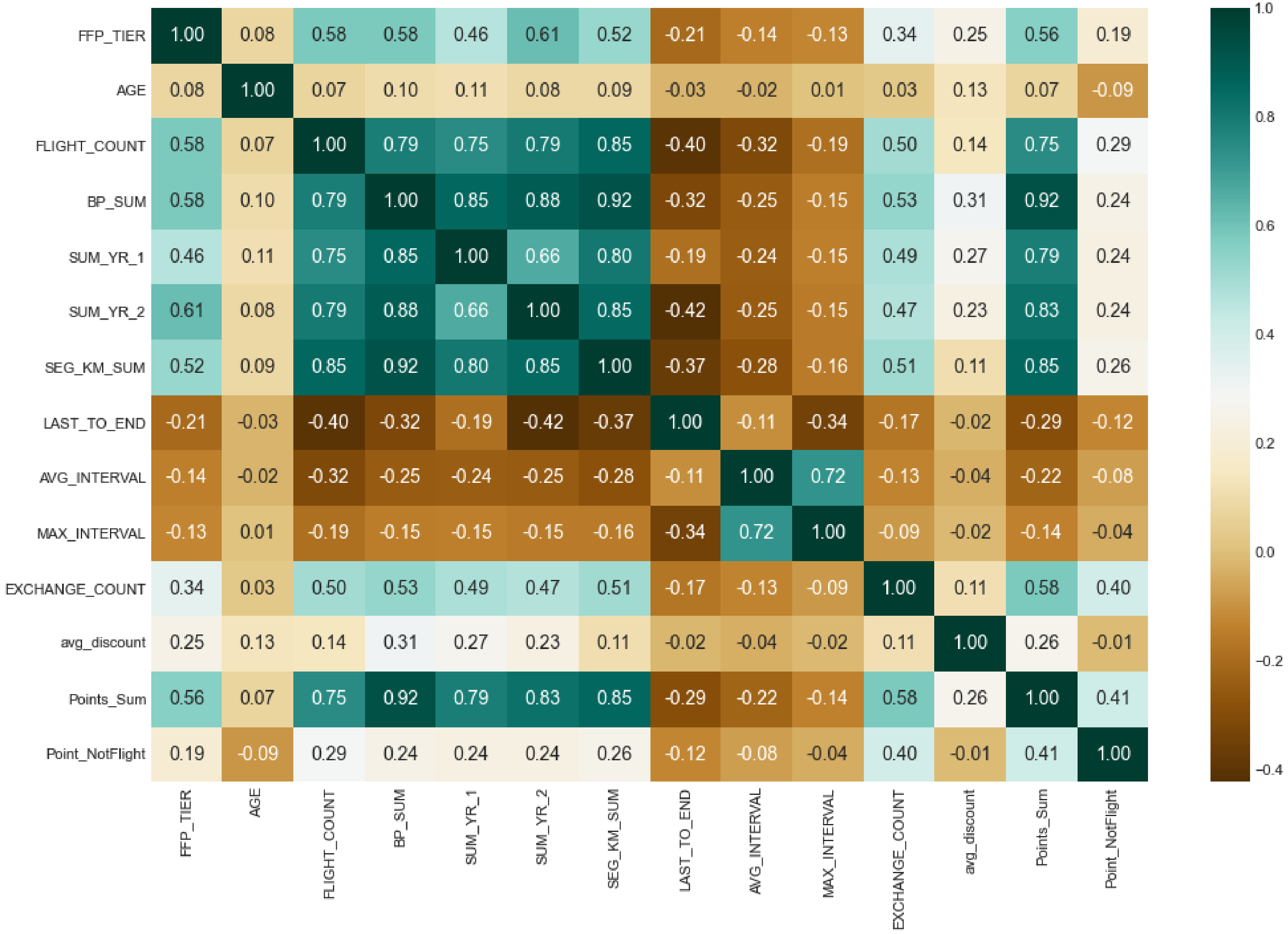
	FFP_DATE	FIRST_FLIGHT_DATE	LOAD_TIME	LAST_FLIGHT_DATE
count	62988	62988	62988	62567
unique	3068	3406	1	730
top	2011-01-13 00:00:00	2013-02-16 00:00:00	2014-03-31 00:00:00	2014-03-31 00:00:00
freq	184	96	62988	959
first	2004-11-01 00:00:00	1905-12-31 00:00:00	2014-03-31 00:00:00	2012-04-01 00:00:00
last	2013-03-31 00:00:00	2015-05-30 00:00:00	2014-03-31 00:00:00	2014-03-31 00:00:00

- From `FFP_DATE` (Frequent Flyer Program Join Date), we know that many customers joined the program on 13 January 2011. with a ddmmyy range from the 1 November 2004 to 31 March 2013.
- Based on column `FIRST_FLIGHT_DATE` (Customer's Date of First Flight), we get the information that many customers have their first flight (on this airline) on 16 February 2013, where the time range ranges from 31 December 1905 to 30 May 2015.
- On column `LAST_FLIGHT_DATE`(Customer Date of Last Flight), it can be seen that many customers have their last flight on 31 March 2014 during time range between 1 April 2012 till 31 March 2014.
- Based on `LOAD_TIME` column, seems that the data was taken on 31 March 2014. This will be the cut off date of this dataset on this project.

Correlation Analysis

From the correlation heatmap, It can be seen that there are some features that have a high correlation with other features, such as:

- FLIGHT_COUNT (Number of Customer Flights)
- BP_SUM (Travel Plans)
- SUM_YR_1 (Fares Revenue)
- SUM_YR_2 (Votes Prices)
- SEG_KM_SUM (Total Distance (Km) Flights that have been done)
- Points_Sum (The Total Points that Earned by Customer)



DATA CLEANING

Handling Missing Values

Because the number of missing values can be categorized as still small, then we will just drop it.

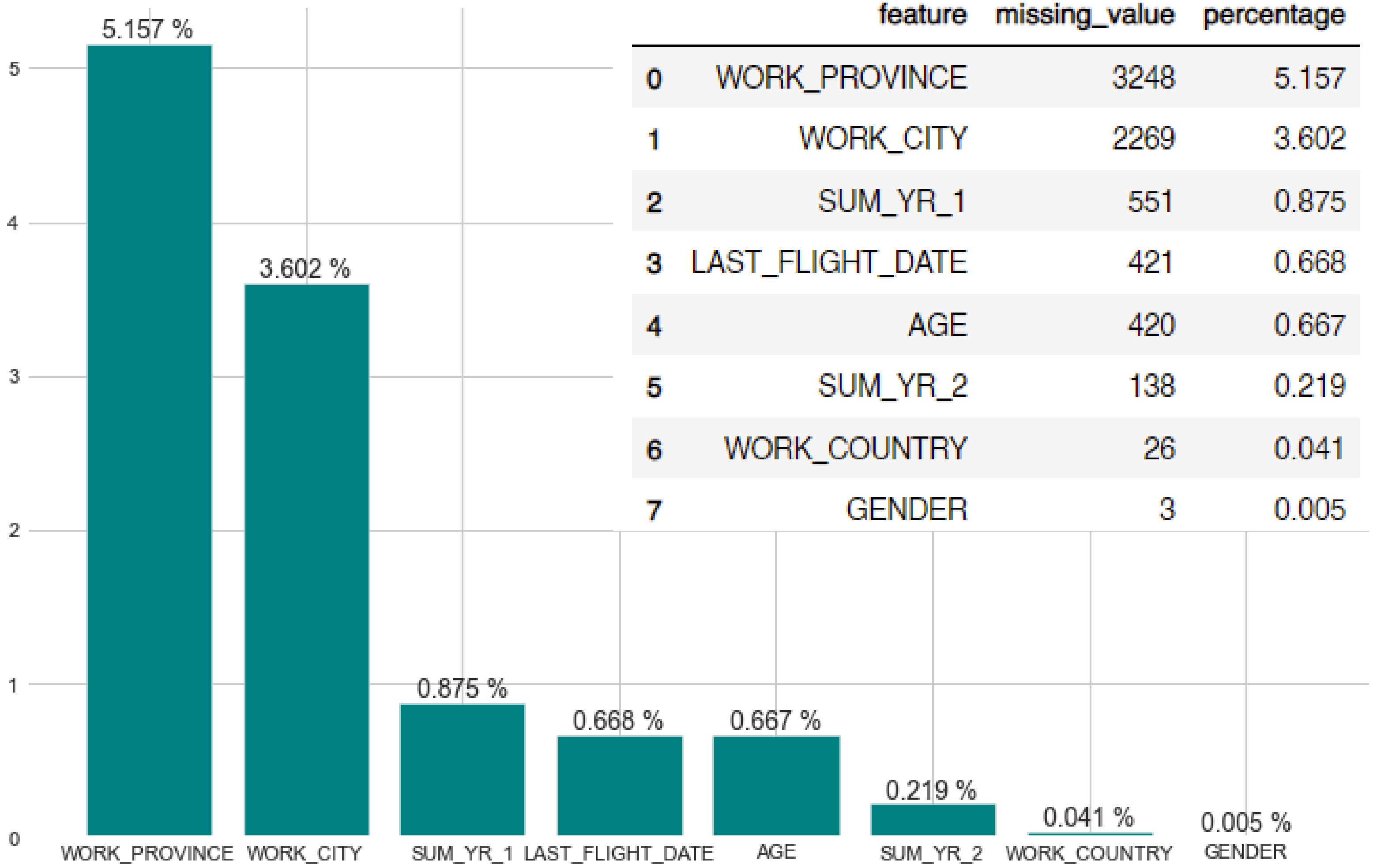
```
# Drop missing values

df = df.dropna().reset_index(drop=True)
df
```

```
# Check rows of data after dropping missing values

df.shape
```

```
(57860, 23)
```



	feature	missing_value	percentage
0	WORK_PROVINCE	3248	5.157
1	WORK_CITY	2269	3.602
2	SUM_YR_1	551	0.875
3	LAST_FLIGHT_DATE	421	0.668
4	AGE	420	0.667
5	SUM_YR_2	138	0.219
6	WORK_COUNTRY	26	0.041
7	GENDER	3	0.005

All missing values have been dropped.
Number of rows become 57860.

FEATURE SELECTION - LRFMC

What is LRFMC ???

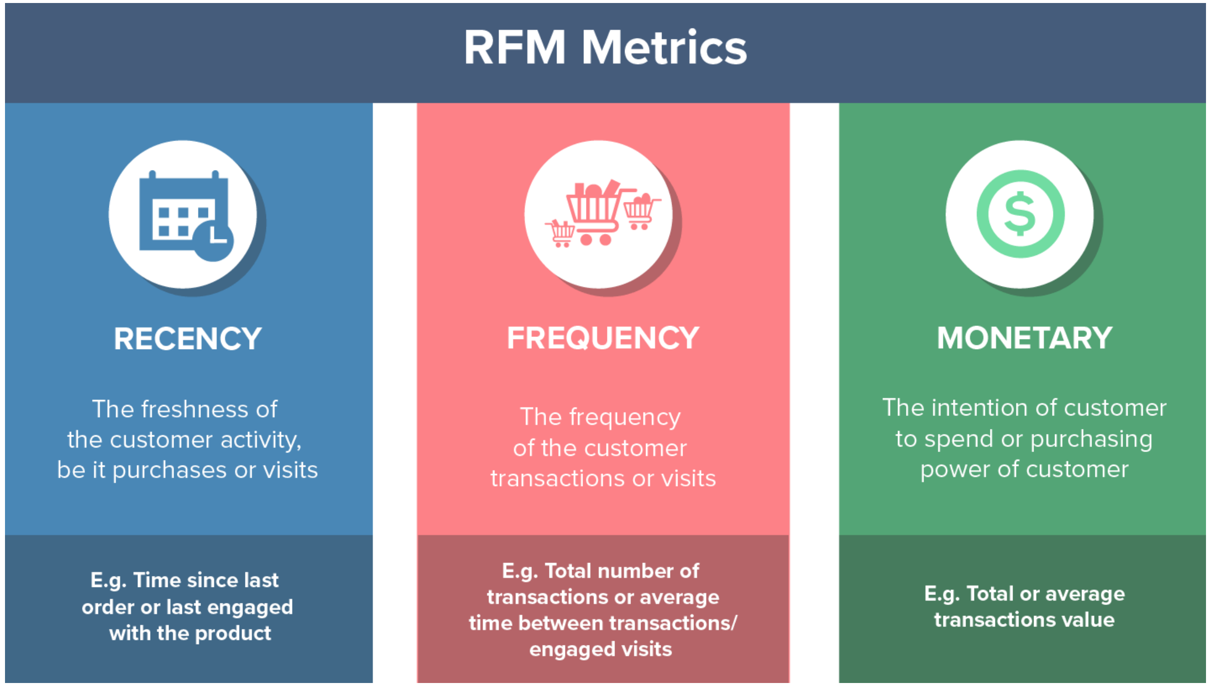


Image source: https://clevertap.com/blog/rfm-analysis/

RFM stands for Recency, Frequency, and Monetary value, each corresponding to some key customer trait. These RFM metrics are important indicators of a customer's behavior because the frequency and monetary value affect a customer's lifetime value, and recency affects retention, a measure of engagement.

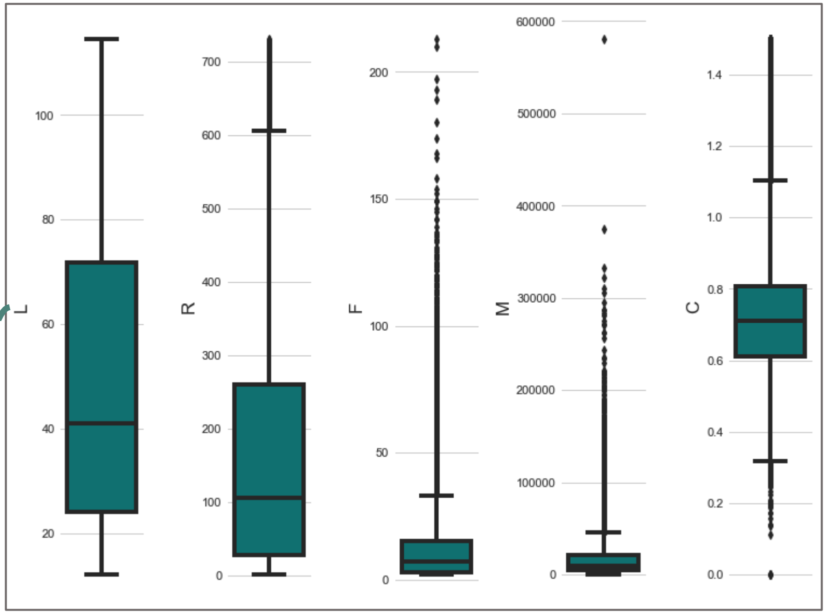
In this project, we extend the RFM model to have 5 indicators namely LRFMC. Since the length of time for airline members to join a meeting can affect customer value to a certain extent, so we add the L indicator. Apart from that, the average value C of the discount coefficient is also used as an airline identification customer value indicator, so that is why we add the C indicator too.

FEATURE SELECTION

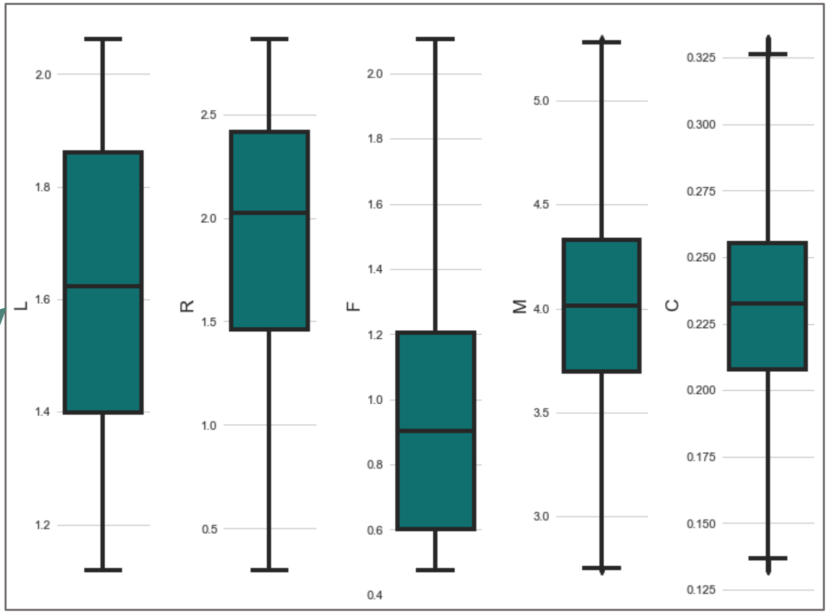
In this project we will use the LRFMC indicator for clustering customers, so we will select only those features.

1. **L = 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]
2. **R = 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]
3. **F = 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]
4. **M = 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]
5. **C = AVG_DISCOUNT**
   The average value of the discount coefficients corresponding to the passengers who traveled during the observation period = average discount rate [Unit: None]

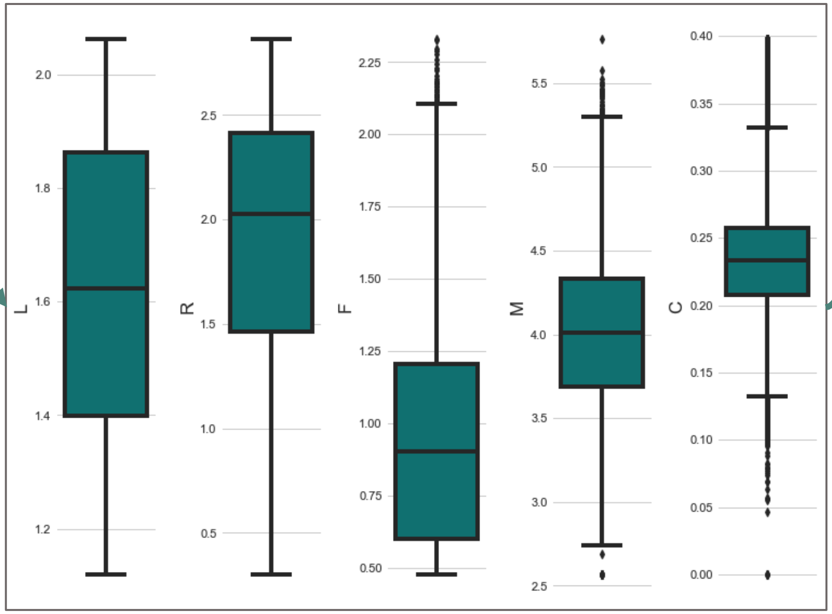
OUTLIER HANDLING & SCALING



Before Handling Outlier



After Remove Outlier based on IQR



After Log Transformation

OUTLIER HANDLING

- rows before IQR outlier filter: 57860
- rows after IQR outlier filter: 55220

```
# Check duplicated value after removing outlier
df_IQR_LRFMC.duplicated().sum()
```

78

- 78 is considering small qty, so we will just drop them
- rows after drop duplicated values 55142

SCALING - Standardization

```
# Standardize data

std = StandardScaler().fit_transform(df_IQR_LRFMC)
df_std_LRFMC = pd.DataFrame(std, columns = list(df_IQR_LRFMC))
df_std_LRFMC
```

	L	R	F	M	C
0	1.638865	-0.567751	1.254030	2.899821	2.436112
1	-1.074189	-2.522903	3.064036	2.888030	2.021912
2	1.661903	-1.102449	2.233282	2.689878	2.767354
3	0.208871	-1.667148	2.739982	2.883933	1.551613
4	0.222830	-1.495477	1.515226	2.753655	2.263657
...	...	...	...	...	...
55137	0.982026	0.715355	-1.320471	-2.917142	-0.075190
55138	0.409956	0.804703	-1.320471	-2.450463	-2.157817
55139	0.754495	0.656189	-1.320471	-2.446033	-2.199887
55140	0.226301	1.212296	-1.320471	-2.526536	-2.326994
55141	-1.339799	1.247406	-1.320471	-2.619335	-2.326994

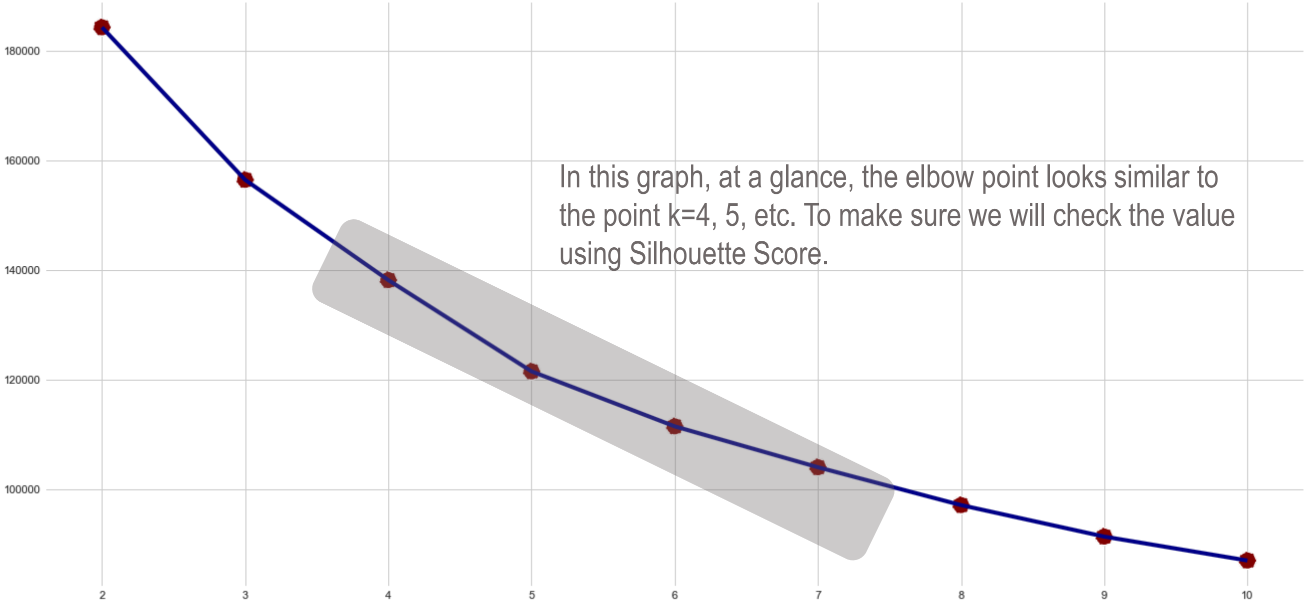
55142 rows × 5 columns

Because K-Means is a distance-based ML algorithm, we need to scale it with StandardScaler.

CLUSTERING (K-Means)

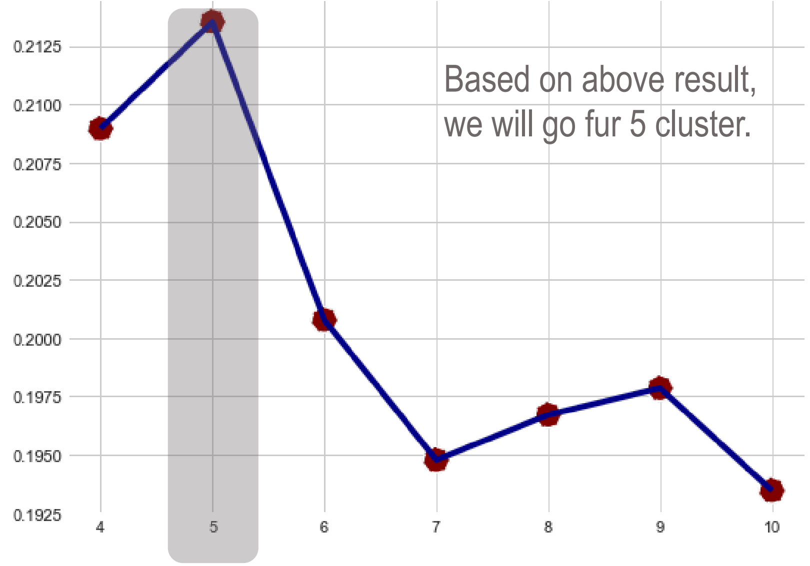
CLUSTERING - K-MEANS

INERTIA



In this graph, at a glance, the elbow point looks similar to the point k=4, 5, etc. To make sure we will check the value using Silhouette Score.

Silhouette Score



Based on above result, we will go fur 5 cluster.

K-MEANS Model

```
# Create clusters using K-Means
kmeans = KMeans(n_clusters=5, random_state=0).fit(df_std_LRFMC)

# Assign Cluster
cluster = kmeans.labels_
df_std_LRFMC['clusters'] = cluster
df_IQR_LRFMC['clusters'] = cluster

# see cluster on data after IQR step
df_IQR_LRFMC.head()
```

	L	R	F	M	C	clusters
0	2.049606	1.544068	1.397940	5.278328	0.319568	0
1	1.339783	0.301030	2.045323	5.273207	0.304486	0
2	2.055633	1.204120	1.748188	5.187151	0.331630	0
3	1.675473	0.845098	1.929419	5.271428	0.287361	0
4	1.679125	0.954243	1.491362	5.214849	0.313289	0

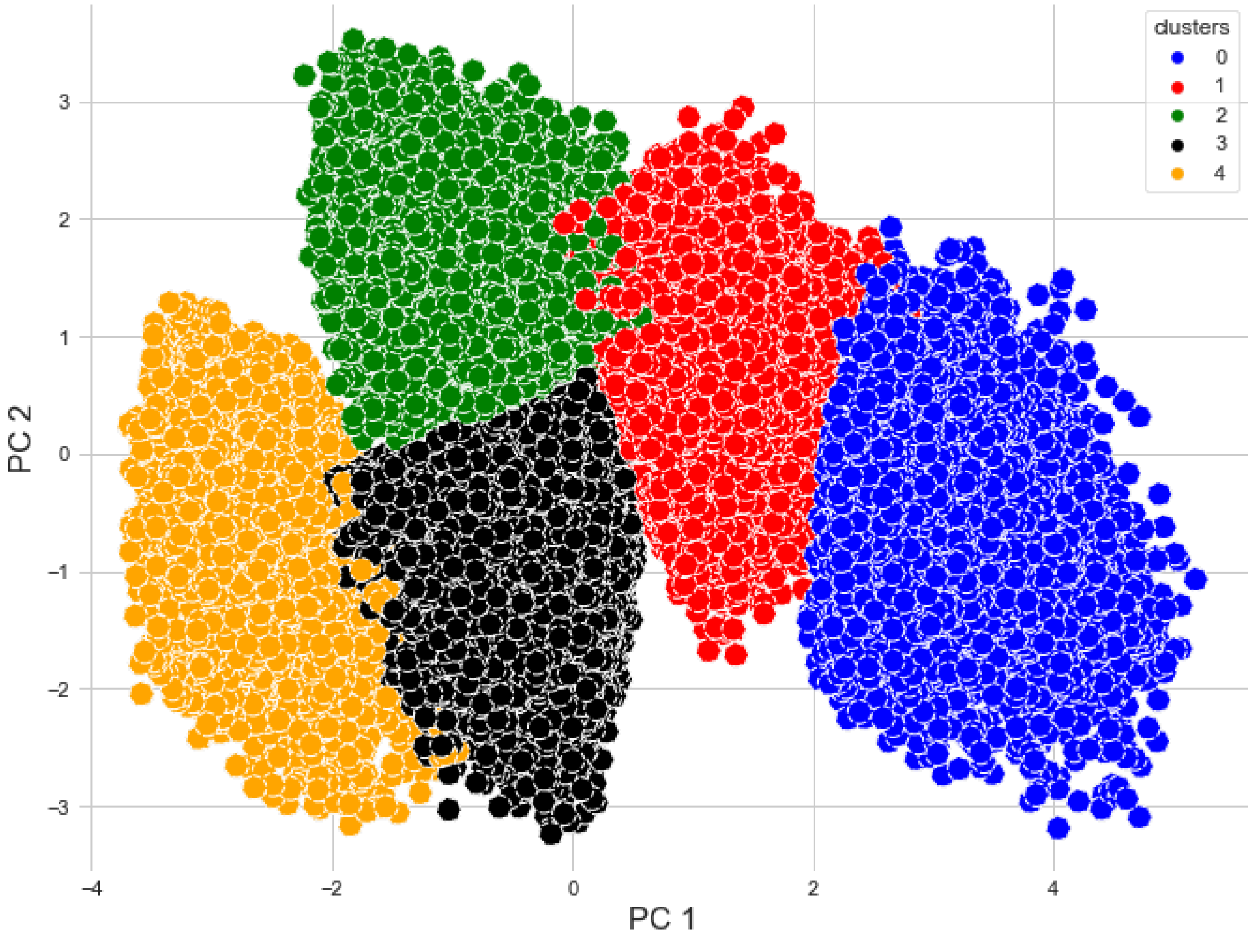
```
# see cluster on data after scaling

df_std_LRFMC.head()
```

	L	R	F	M	C	clusters
0	1.638865	-0.567751	1.254030	2.899821	2.436112	0
1	-1.074189	-2.522903	3.064036	2.888030	2.021912	0
2	1.661903	-1.102449	2.233282	2.689878	2.767354	0
3	0.208871	-1.667148	2.739982	2.883933	1.551613	0
4	0.222830	-1.495477	1.515226	2.753655	2.263657	0

CLUSTERING - K-MEANS

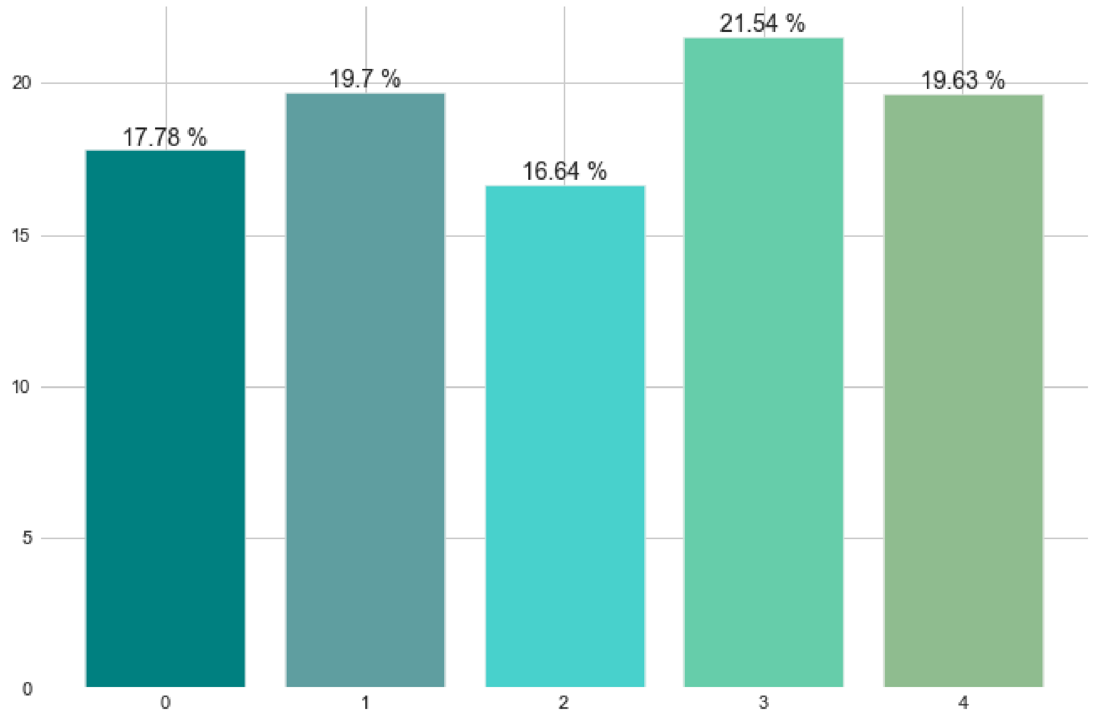
Check Visualization using PCA & Scatter Plot



After evaluating through visualization and compare it with what we've done through inertia and silhouette score, the number of clusters (k=5) is appropriate. The scatter plot shows the data has been clustered quite well.

CUSTOMER VALUE ANALYSIS

	cluster	count	percentage
0	0	10289	17.78
1	1	11400	19.70
2	2	9629	16.64
3	3	12465	21.54
4	4	11359	19.63

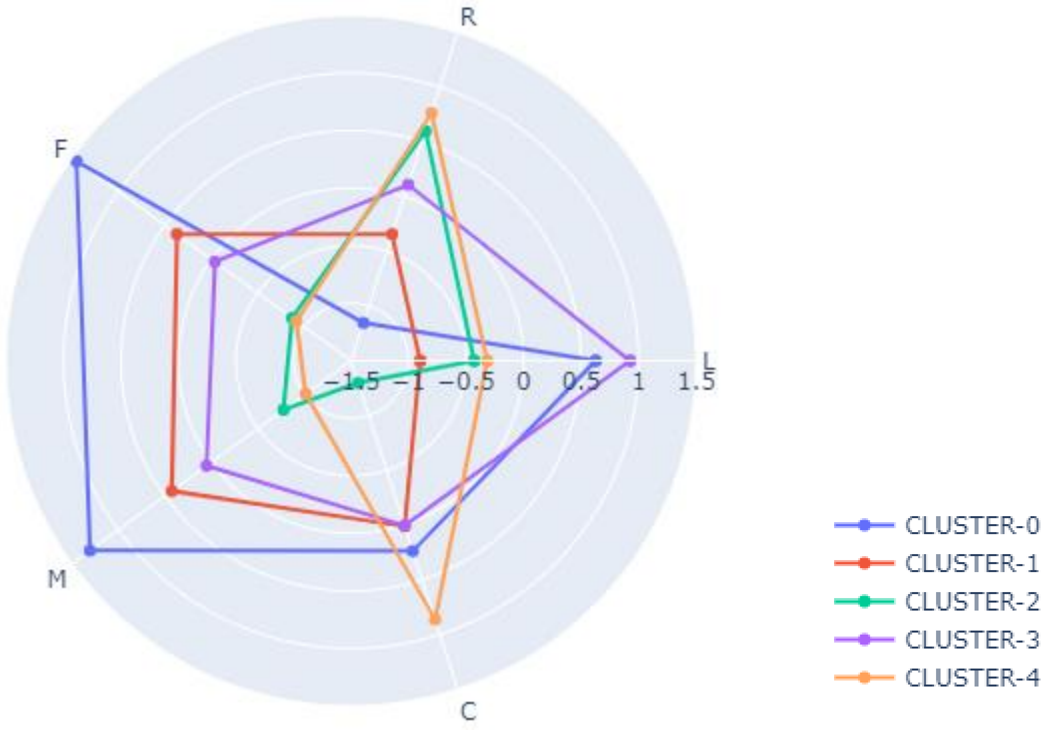


Cluster Count

There are 5 cluster (customer segment/group) with the number of customers from each cluster as follows:

1. Cluster - 0 / Customer Group 1 : 10.289 (17.78 %)
2. Cluster - 1 / Customer Group 2 : 11.400 (19.70 %)
3. Cluster - 2 / Customer Group 3 : 9.629 (16.64 %)
4. Cluster - 3 / Customer Group 4 : 12.465 (21.54 %)
5. Cluster - 4 / Customer Group 5 : 11.359 (19.63 %)

It can be said that this clustering has a fairly even distribution of the number of customers.



ANALYZE USING RADAR CHART

There are five clusters:

- Cluster-0 is Customer Group 1
- Cluster-1 is Customer Group 2
- Cluster-2 is Customer Group 3
- Cluster-3 is Customer Group 4
- Cluster-4 is Customer Group 5

In general, these clusters are formed because there are differences in the value of the LRFMC indicator.

CLUSTER LABEL

1. **Customer Group 1 :**
   - This is a group of customers who actually fly frequently, have a high monetary value because of their high mileage and are customers who have been in the frequent flyer program for a long time. The average discount rate is moderate. And this customer has a good recency because they recently flew with this airline.
   - Let's label this customer as **The Champions**

2. **Customer Group 2 :**
   - This customer group is a new customer (recently joined the frequent flyer program) so it has an RFMC value that is not yet high but looks potential.
   - Let's label this customer as **Potential Loyalists - New Customers**

3. **Customer Group 3 :**
   - This customer group is those who have been in the frequent flyer program for quite a while. However, this group actually does not use this airline very often, has a low monetary or mileage value and the average discount rate is also quite low. They also have not used this airline for a very long time.
   - Let's label this customer as **Hibernating - Low Value Customers**

4. **Customer Group 4 :**
   - This customer group is a customer who has been in the frequent flyer program for a long time and has a moderate RFMC score. They don't fly very often just moderate, so the other values are also moderate but actually have potential.
   - Let's label this customer as **Potential Loyalists - General Customers**

5. **Customer Group 5 :**
   - This customer group has been joining the frequent flyer program for a long time. They have not used this airline for a very long time and have high average discount rate. And this customer group rarely flies and has low monetary value or mileage.
   - Let's label this customer as **Hibernating - Price Sensitive Customers**

No	Customer Group	Dominant Value	Moderate Value	Minimum Value
1	Customer Group 1	FML	C	R*
2	Customer Group 2		R*FMC	L
3	Customer Group 3	R*	L	FMC
4	Customer Group 4	L	R*FMC	
5	Customer Group 5	R*C	L	FM

(*) a low R value means that the customer has recently flown with the airline. And a high R value means that the customer has not used this airline for a long time.
For the value of R we have to be careful because the meaning is the opposite. Low R is actually good from the airline (business) point of view.

USE REAL DATA

To do a better analysis of the five customer groups, we will use real data. In this case we use the data before standardization. However, because the data is log-transformed data, we will return the actual value (antilog) with the exponential formula.

```
# display mean and median for each cluster to get the real value in average and median

display(df_cluster.groupby('clusters').agg(['mean','median']))
```

	L		R		F		M		C	
	mean	median	mean	median	mean	median	mean	median	mean	median
clusters										
0	65.503181	65.666667	26.598698	14.0	32.384100	27.0	44828.602682	37953.0	0.739606	0.733479
1	24.661076	23.400000	88.512982	58.0	12.467281	11.0	18672.256754	15867.0	0.706767	0.704405
2	36.423723	30.266667	274.542632	238.0	3.675460	3.0	5738.858033	4830.0	0.528545	0.534635
3	74.492911	74.433333	148.552908	112.0	8.613317	8.0	12876.017168	11127.0	0.706558	0.702362
4	39.087232	32.266667	323.745576	298.0	3.477771	3.0	4573.088740	3924.0	0.832252	0.820805

```
# we do antilog by creating new dataframe df_cluster

df_cluster = df_IQR_LRFMC.copy()
df_cluster['L'] = 10 ** df_IQR_LRFMC['L'] - 1
df_cluster['R'] = 10 ** df_IQR_LRFMC['R'] - 1
df_cluster['F'] = 10 ** df_IQR_LRFMC['F'] - 1
df_cluster['M'] = 10 ** df_IQR_LRFMC['M'] - 1
df_cluster['C'] = 10 ** df_IQR_LRFMC['C'] - 1
df_cluster
```

	L	R	F	M	C	clusters
0	111.100000	34.0	24.0	189813.0	1.087220	0
1	20.866667	1.0	110.0	187588.0	1.015978	0
2	112.666667	15.0	55.0	153868.0	1.146001	0
3	46.366667	6.0	84.0	186821.0	0.938031	0
4	46.766667	8.0	30.0	164001.0	1.057257	0
...	...	...	...	...	...	...
55137	74.466667	228.0	2.0	564.0	0.690922	4
55138	52.466667	260.0	2.0	900.0	0.420000	2
55139	64.800000	209.0	2.0	904.0	0.415000	2
55140	46.866667	473.0	2.0	834.0	0.400000	2
55141	17.633333	498.0	2.0	760.0	0.400000	2

1.Customer Group 1 (**The Champions**) :

1. It turns out that this customer group has joined ffp for about 65 months, has traveled about 40 thousand km mileage, last flew with an airline about 20 months ago, has flown with an airline about 30 times with an average discount rate of 0.7.

2.Customer Group 2 (**Potential Loyalists - New Customers**) :

2. This customer group is those who have just joined the airline for about 20 months, have only flown with the airline about 80 months before, with a frequency of 12 flights, have traveled 18 thousand km mileage with an average discount rate of 0.7.

3.Customer Group 3 (**Hibernating - Low Value Customers**) :

3. This group is those who have joined ffp for 30 months or more, their last flight was very long about 250 months ago, with a flight frequency of about 3 times, with a total mileage of about 5 thousand km and an average discount rate of 0.5.

4.Customer Group 4 (**Potential Loyalists - General Customers**) :

4. This group is similar to group 2, but this group has joined ffp longer, which is about 70 months ago.

5.Customer Group 5 (**Hibernating - Price Sensitive Customers**) :

5. This group is similar to group 3, but this group has a high average discount rate, which is around 0.8.

BUSINESS RECOMMENDATION

1.The Champions - Customer Group 1

- Airline must really take care of this customer group, because this group contributes well to the business. They can become early adopters for new airline service or program, and will help promote it. One way to keep these customers is through a reward program for this type of customer group. The reward program can also be accompanied by a kind of referral program, with the aim of providing rewards but encouraging them to promote airline brands (or certain programs).
- Example:
  - Special reward discount + post your flight experience for additional discount for the next flight
  - More FFP points (x2/x3) for your flight + booked your flight with friends and get special discount/price
  - Book 1 for 2 - just for you - and triples your point
  - Travelling - Chilling - Healing - Saving. Flying to our new route with xx% discount + triples your ffp point. Catch additional points!!! Get more when your friends booked this route with ur refferal code.
  - My Poin Rewards - Provides convenience and various point reedem options.
  - etc.

2.Potential Loyalists - New Customers - Customer Group 2

- This customer has recently joined the ffp program at this airline. However, they have potential as loyal customers in the future, as can be seen from the good RFMC value as a new customer. By building a good relationship with onboarding support and special offer programs, it may be possible to help increase their frequency and mileage.
- Example:
  - Flight Booking assistant
  - Boost your tier by special flight discount rate with us - double your mileage now
  - Friday escape with firends to upgrade your tier
  - etc.

3.Hibernating - Low Value Customers - Customer Group 3

- This customer has a fairly low value for the airline. There is a possibility that they are the type of customers who fly only because there are certain interests or events. Although the company is not obliged to focus on this type of customer, the company should still try to induce them otherwise they will be completely lost. Airline should make a program to wake them up from hibernation.
- Example:
  - I miss u or I don't wanna lose you program. Give special discount or flight rate.
  - Can't forget u / Let's do it again/ Can't move on Program. Awake them with good memory of their last flight by giving them voucher/ code discount rate with flight routes according to their last flight.
  - etc.

BUSINESS RECOMMENDATION

4.Potential Loyalists - General Customers - Customer Group 4

- This customer group is similar to group 2, but they have been in the frequent flyer program for longer. This customer has a pretty good track record but needs to be improved because they have the potential to contribute more to the company's business. The programs or campaigns that can be offered are more towards increasing engagement by telling them that they are loyalists (even though they are not champions) and influencing them to continue with us and create more shared moments.
- Example:
- Let's be our part forever - Give special rewards point for next booking
  - Fly more get more - doubles the points for this year (particular time period) booking
  - Share your moment with us - for additional points to boost your tier and discover the next treasure
  - etc.

5.Hibernating - Price Sensitive Customers - Customer Group 5

- This customer is similar to group 3, only they seem to have a higher price sensitive. So the approach taken by the company must be more aggressive than group 3. Programs that can be provided to wake them up must be more thought-provoking with special prices or discounts and various benefits that are more attractive to them. The company can provide a typical special program with what they might have gotten before.
- Example:
  - I want to get back together with you - This Deal is just for you
  - Don't you miss this Biggest Deal? Let's do it again!
  - etc.

REFERENCES

REFERENCES

The following are references in working on this project.

https://www.programmersought.com/article/63823799496/

https://www.kaggle.com/code/felixign/airline-clustering/notebook

https://www.kaggle.com/code/amarmaruf/homework-unsupervised-rakamin-ds8/notebook

https://clevertap.com/blog/rfm-analysis/

https://www.moengage.com/blog/rfm-analysis-using-rfm-segments/

THANK YOU