K-Means Clustering: Airline Customer Value Analysis

Awalsyah Rinanto Putra

Table of Contents

Airline Customer Value Dataset

Customer dataset from airline company contains some features to describe each customer's value.

Click to download the dataset

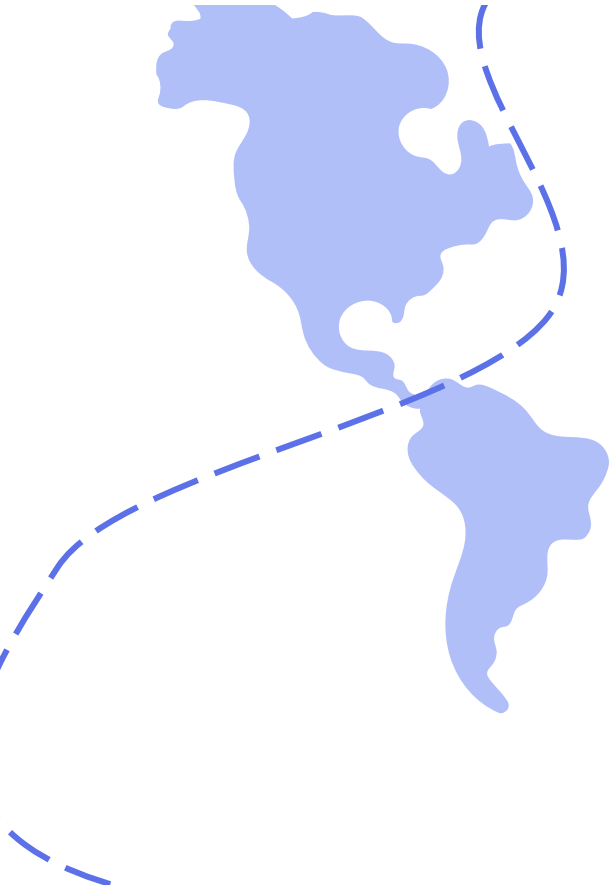
Objective

- To analyze and divide airline customer into segments.
- To make business recommendation based on the cluster model

Goal

Create a clustering model to make customer segmentation

Feature Description



Code	Description
MEMBER_NO-b	: ID Member
FFP_DATE	: Frequent Flyer Program Join Date
FIRST_FLIGHT_DATE	: Tanggal Penerbangan pertama
GENDER	: Jenis Kelamin
FFP_TIER	: Tier dari Frequent Flyer Program
WORK_CITY	: Kota Asal
WORK_PROVINCE	: Provinsi Asal
WORK_COUNTRY	: Negara Asal
AGE	: Umur Customer
LOAD_TIME	: Tanggal data diambil
FLIGHT_COUNT	: Jumlah penerbangan Customer
BP_SUM	: Rencana Perjalanan
SUM_YR_1	: Fare Revenue
SUM_YR_2	: Votes Prices
SEG_KM_SUM	: Total jarak(km) penerbangan yg sudah dilakukan
LAST_FLIGHT_DATE	: Tanggal penerbangan terakhir
LAST_TO_END	: Jarak waktu penerbangan terakhir ke pesanan penerbangan paling akhir
AVG_INTERVAL	: Rata-rata jarak waktu
MAX_INTERVAL	: Maksimal jarak waktu
EXCHANGE_COUNT	: Jumlah penukaran
avg_discount	: Rata rata discount yang didapat customer
Points_Sum	: Jumlah poin yang didapat customer
Point_NotFlight	: point yang tidak digunakan oleh members

01

Exploratory Data Analysis (EDA)

Descriptive, univariate, and multivariate analysis

1. Descriptive Analysis

Data Types

8 Categorical
15 Numerical

Missing Values

Age, fare revenue, vote prices, work city, province and country.

Duplicated Rows

0 duplicated rows

Numerical columns statistical summary

	count	mean	std	min	25%	50%	75%	max
MEMBER_NO	62988.0	31494.500000	18183.213715	1.0	15747.750000	31494.500000	47241.250000	62988.0
FFP_TIER	62988.0	4.102162	0.373856	4.0	4.000000	4.000000	4.000000	6.0
AGE	62568.0	42.476346	9.885915	6.0	35.000000	41.000000	48.000000	110.0
FLIGHT_COUNT	62988.0	11.839414	14.049471	2.0	3.000000	7.000000	15.000000	213.0
BP_SUM	62988.0	10925.081254	16339.486151	0.0	2518.000000	5700.000000	12831.000000	505308.0
SUM_YR_1	62437.0	5355.376064	8109.450147	0.0	1003.000000	2800.000000	6574.000000	239560.0
SUM_YR_2	62850.0	5604.026014	8703.364247	0.0	780.000000	2773.000000	6845.750000	234188.0
SEG_KM_SUM	62988.0	17123.878691	20960.844623	368.0	4747.000000	9994.000000	21271.250000	580717.0
LAST_TO_END	62988.0	176.120102	183.822223	1.0	29.000000	108.000000	268.000000	731.0
AVG_INTERVAL	62988.0	67.749788	77.517866	0.0	23.370370	44.666667	82.000000	728.0
MAX_INTERVAL	62988.0	166.033895	123.397180	0.0	79.000000	143.000000	228.000000	728.0
EXCHANGE_COUNT	62988.0	0.319775	1.136004	0.0	0.000000	0.000000	0.000000	46.0
avg_discount	62988.0	0.721558	0.185427	0.0	0.611997	0.711856	0.809476	1.5
Points_Sum	62988.0	12545.777100	20507.816700	0.0	2775.000000	6328.500000	14302.500000	985572.0
Point_NotFlight	62988.0	2.728155	7.364164	0.0	0.000000	0.000000	1.000000	140.0

Key takes:

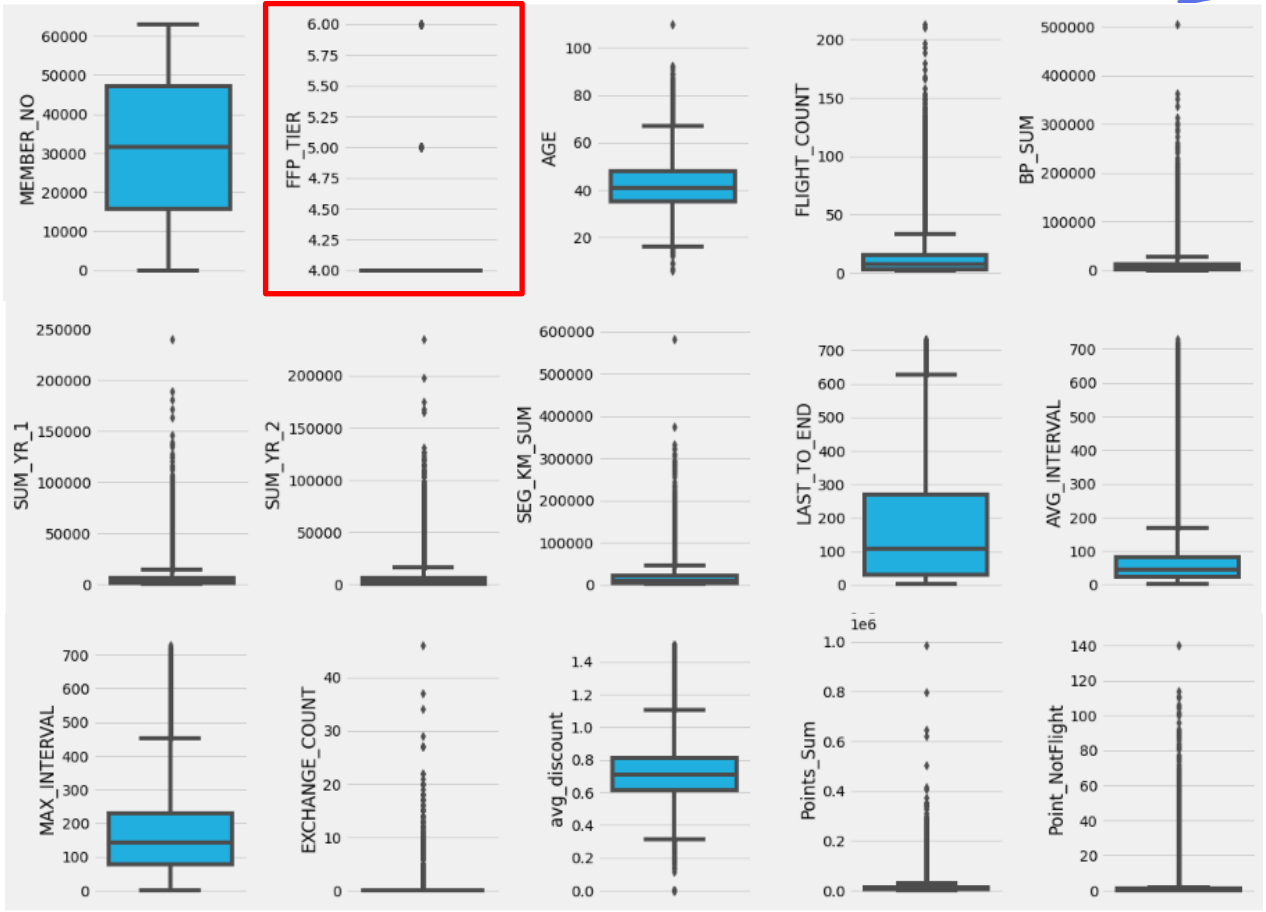
- **Age** has strange maximum values (110)

Categorical columns statistical summary

	FFP_DATE	FIRST_FLIGHT_DATE	GENDER	WORK_CITY	WORK_PROVINCE	WORK_COUNTRY	LOAD_TIME	LAST_FLIGHT_DATE
count	62988	62988	62985	60719	59740	62962	62988	62988
unique	3068	3406	2	3234	1165	118	1	731
top	1/13/2011	2/16/2013	Male	guangzhou	guangdong	CN	3/31/2014	3/31/2014
freq	184	96	48134	9386	17509	57748	62988	959

Key takes:

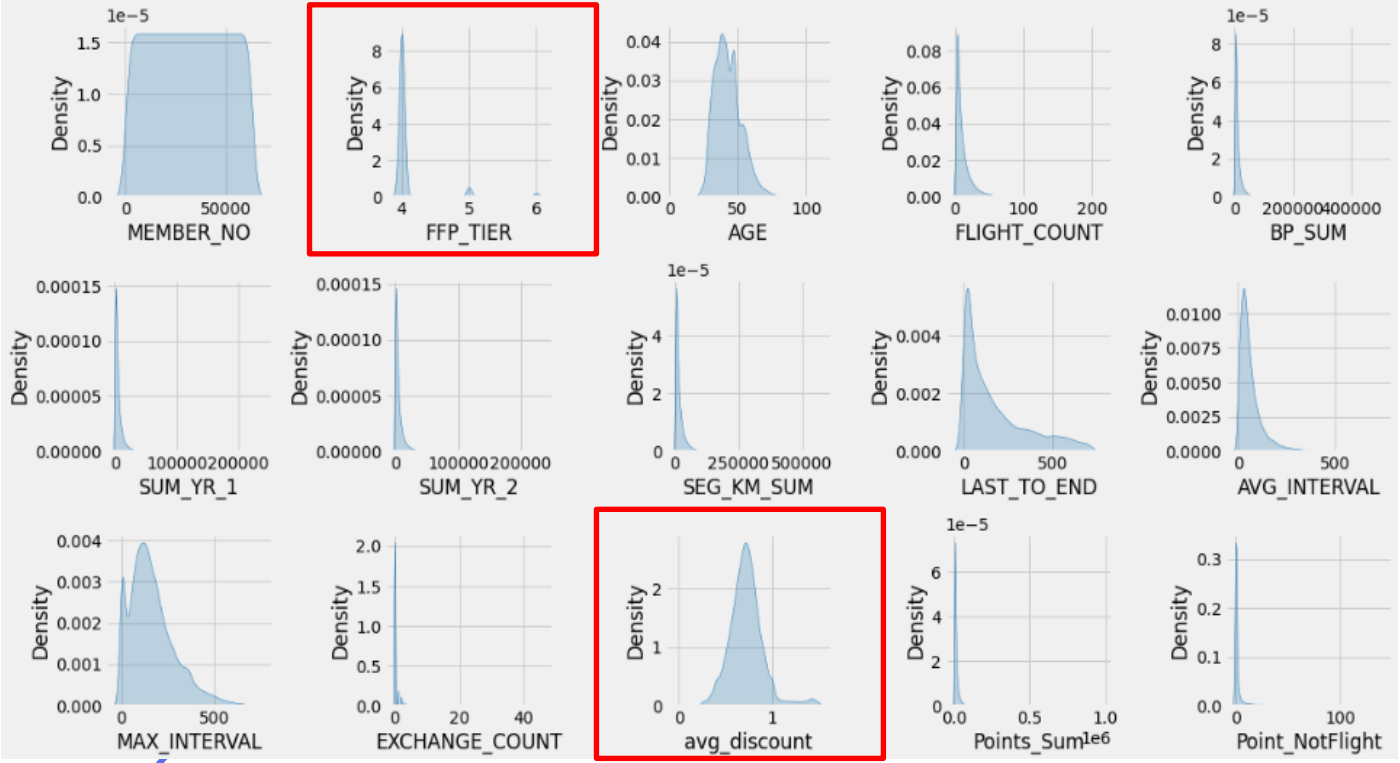
- **Work city, province,** and **country** can be **dropped** because they have **too many unique values** and cant compute with model.
- **FFP_DATE, FIRST_FLIGHT_DATE, LOAD_TIME** and **LAST_FLIGHT_DATE** are **date** types.



2. Univariate Analysis (Boxplot)

Key takes:

- **Most of features** have outliers**.**
- There's an indication **FFP_TIER** have **discreet values.**

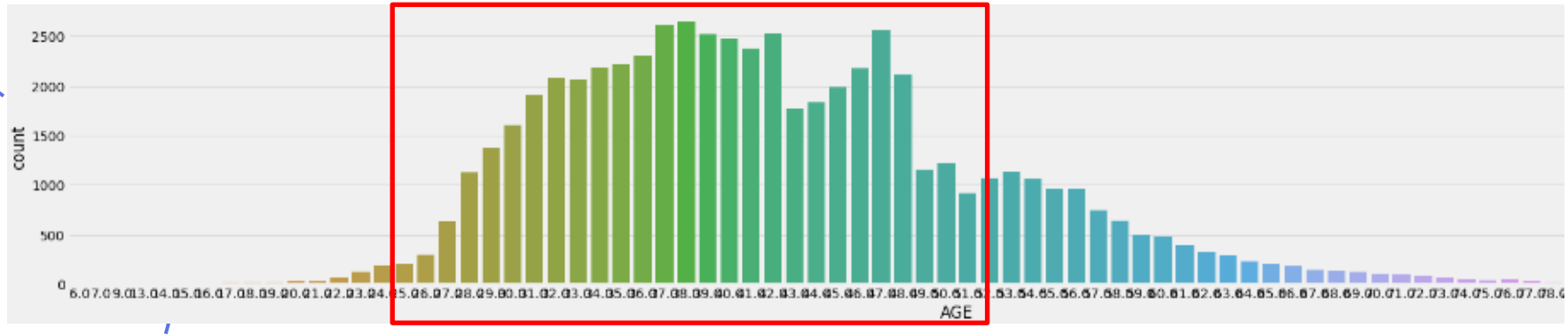


Univariate Analysis (Distplot)

Key takes:

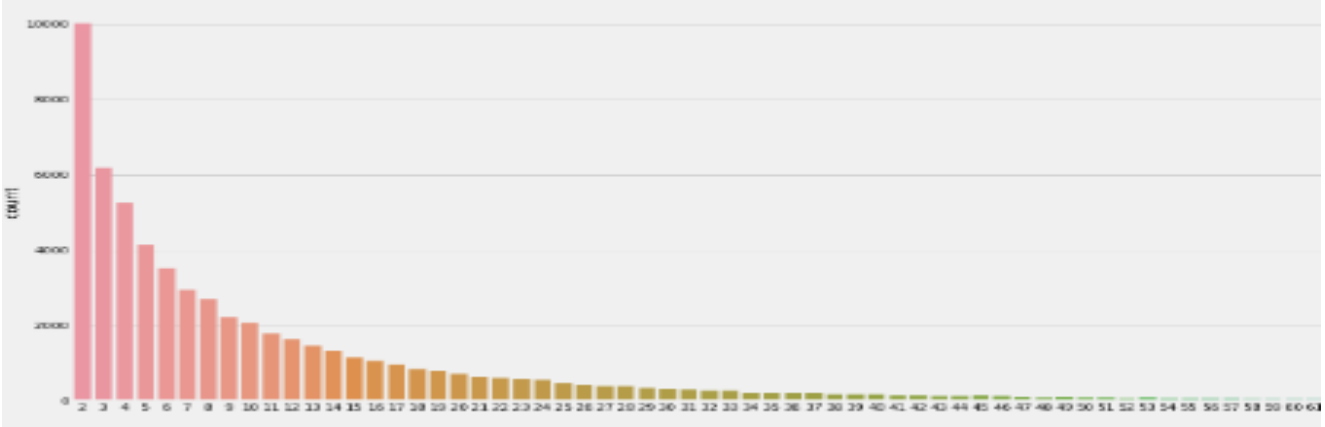
- **Discount** has strange values above 1 (>100%)
- **Most of the features** are **right-skewed** and **have outliers.**

Countplot – Age feature

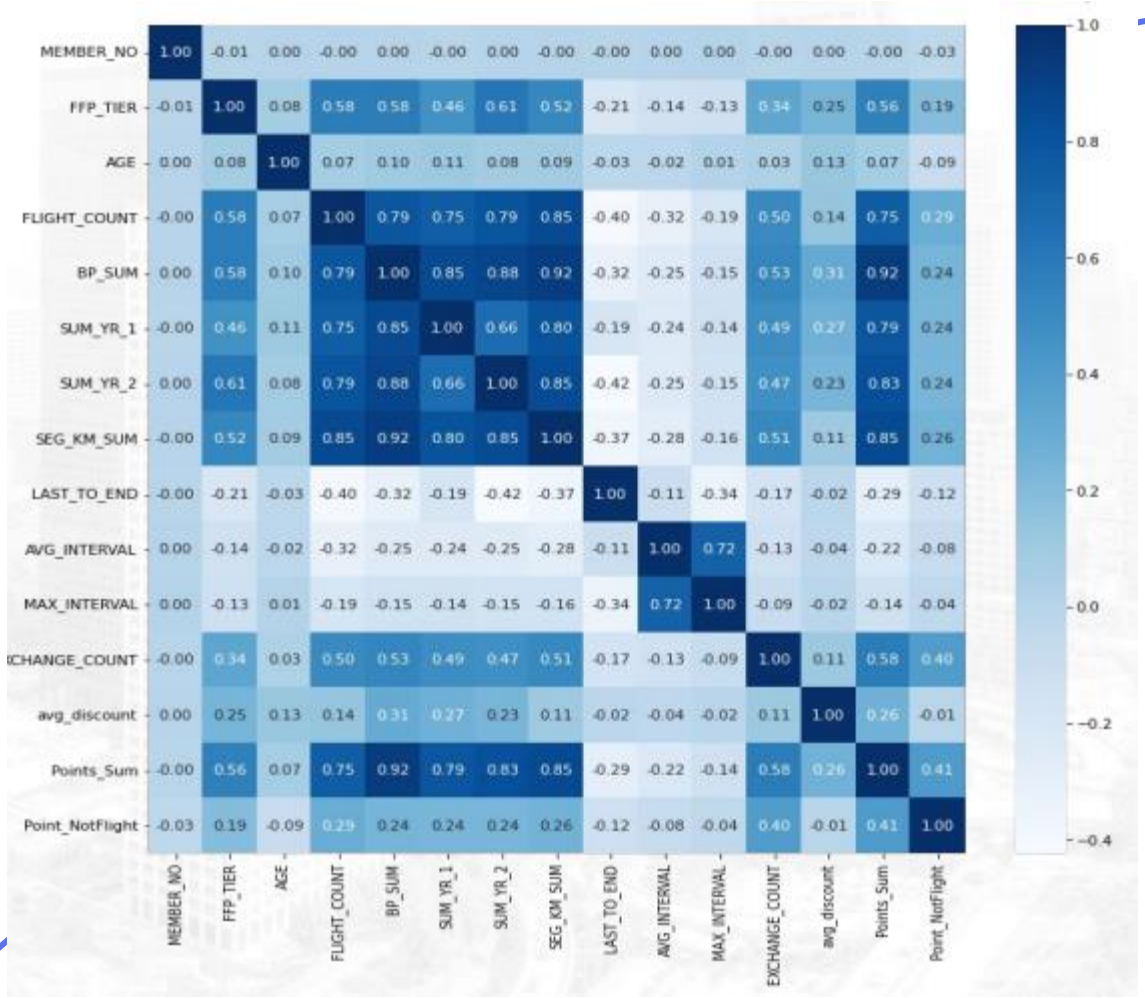


Most of the customer aged between 26-60 (productive age).

Countplot – Flight count feature



Most of the customers have flight only 1-7 times, and then the number is decreasing after that..



3. Multivariate Analysis (Heatmap plot)

- **FLIGHT_COUNT, BP_SUM, SUM_YR_1, SUM_YR_2, SEG_KM_SUM,** and **Points_Sum** are **multicolinearity** columns.
- **AVG_INTERVAL** and **MAX_INTERVAL** are **multicolinearity** columns.
- **AGE** has **very low correlation** with all features

02

Data Pre-Processing

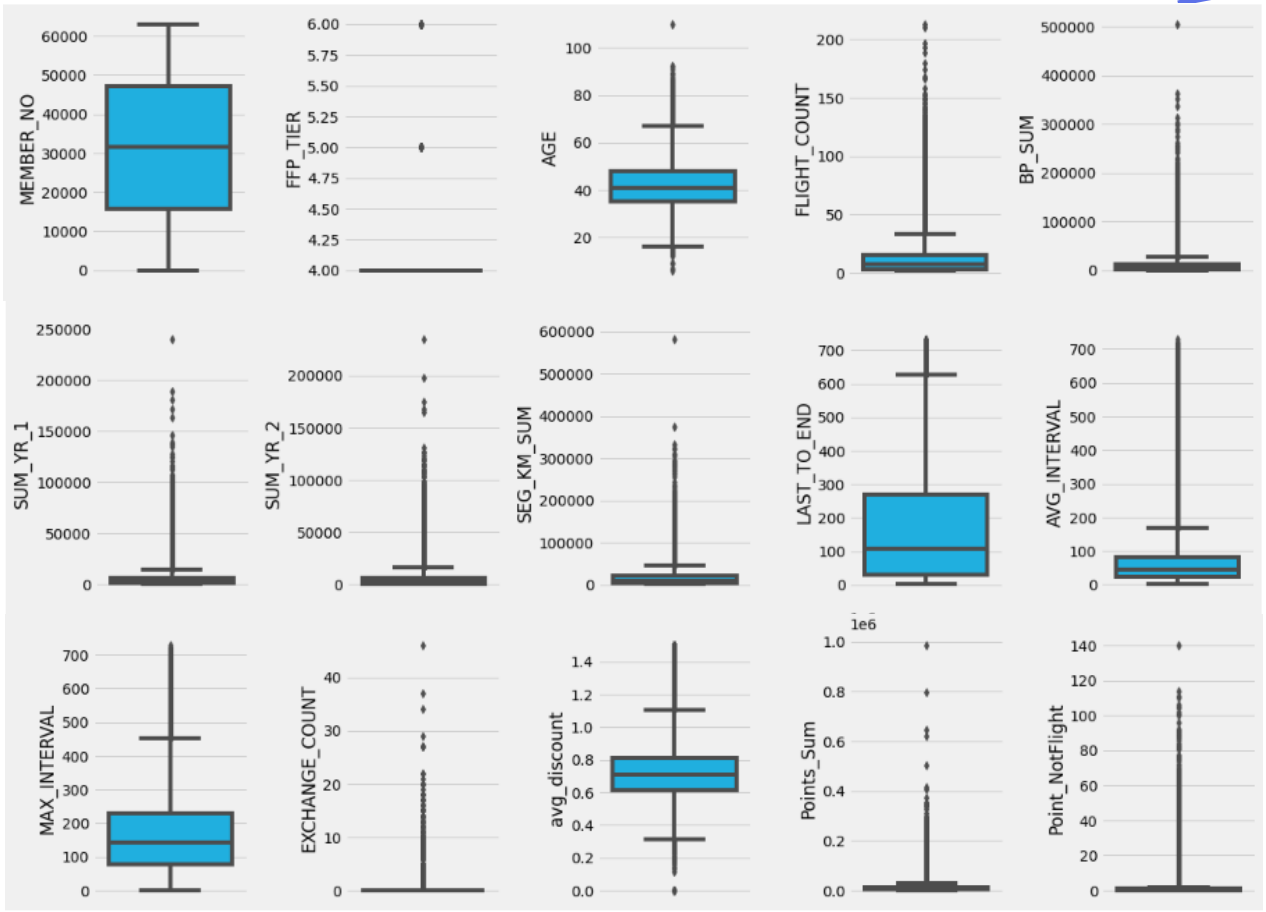
Missing values and outliers handling, standardization, feature selection & engineering

Missing Values

- **Age**: fill with **median**
- **Revenue** (SUM_YR_1 & 2 SUM_YR_2) : fill **0**
- **Drop work city, province and country** because they are **categoricals** and have **too many unique values**

```
df.isna().sum()
```

MEMBER_NO	0
FFP_DATE	0
FIRST_FLIGHT_DATE	0
GENDER	1
FFP_TIER	0
WORK_CITY	2102
WORK_PROVINCE	3019
WORK_COUNTRY	23
AGE	389
LOAD_TIME	0
FLIGHT_COUNT	0
BP_SUM	0
SUM_YR_1	542
SUM_YR_2	134
SEG_KM_SUM	0



Outlier Handling

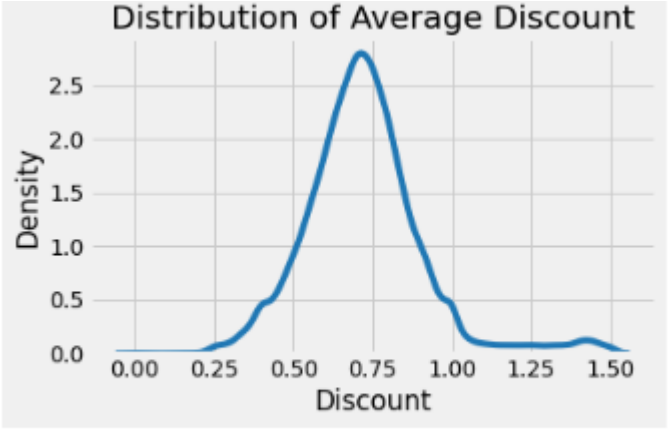
Remove outlier with z-score.

Feature Selection

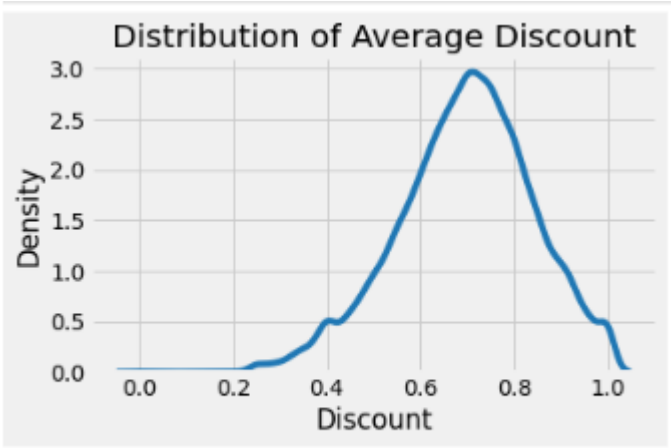
LRFMC analysis: **to divide customers aviation industry into segments**.

- **L (Length of joining member**) : The number of months since the member's joining time => **LOAD_TIME - FFP_DATE**
- **R (Recent flight)** : Number of months since the member's last flight => **LAST_TO_END**
- **F (Flight Count)** : The total number of times the member has flown => **FLIGHT_COUNT**
- **M (Miles Accumulated)** : Miles accumulated => **SEG_KM_SUM**
- **C (Discount Used)** : The average value of the discount used by the member => **avg_discount**

Feature engineering



It makes no sense that `avg_discount` has values more than 1 (100%)



No discount > 100%

Drop rows with discount more than 100%

Standardization

```
# remove rows nan, inf, dan -inf

df_cust =df_cust[~df_cust.isin([np.nan, np.inf, -np.inf]).any(1)]
```

```
from sklearn.preprocessing import StandardScaler

std = StandardScaler()
custvalue_std = std.fit_transform(df_cust)
custvalue_std
```

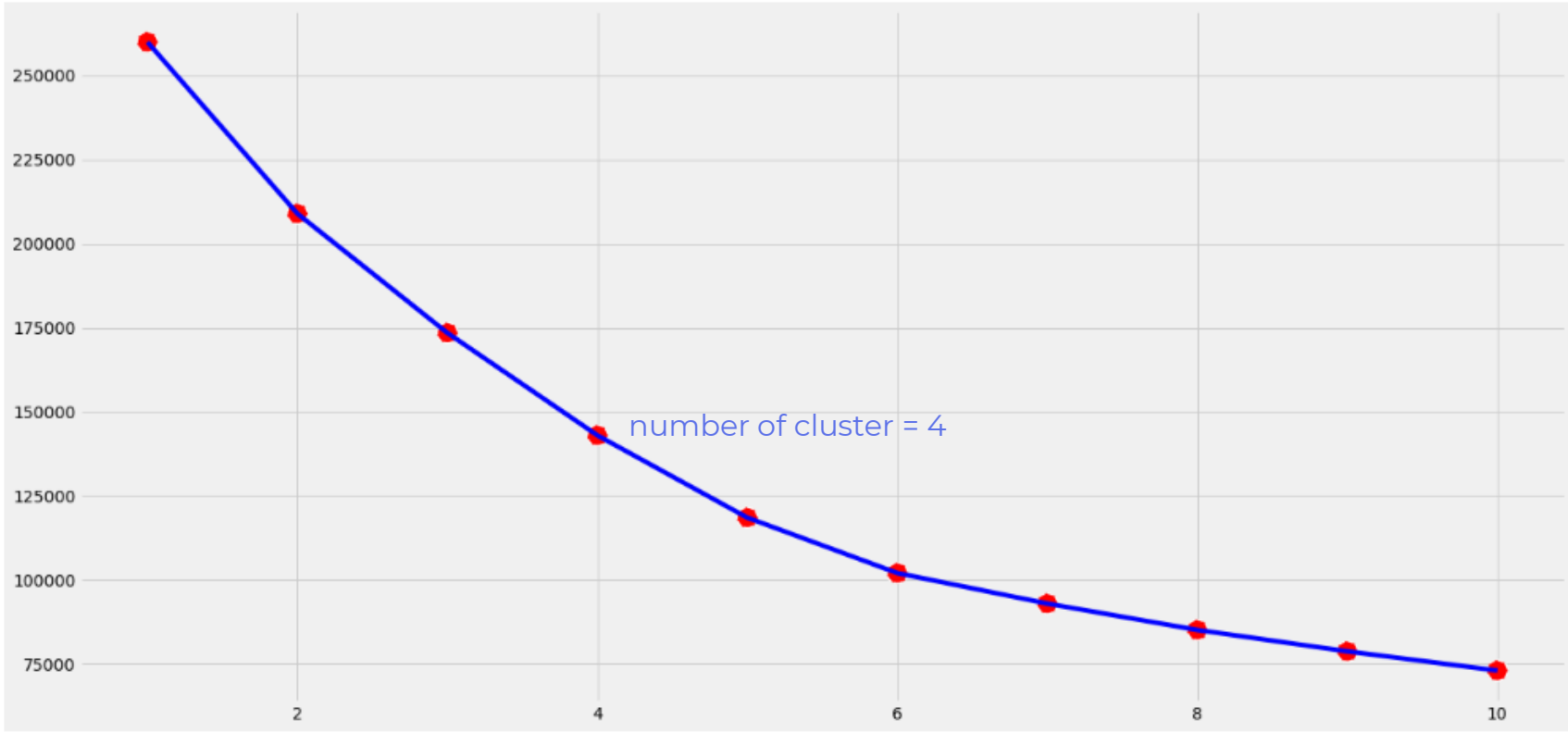
	L	R	F	M	C
0	0.143953	-0.783974	0.933512	4.901182	0.651321
1	0.293691	-0.634441	-0.079031	4.799460	0.742712
2	-0.386717	-0.698527	0.033474	4.897951	0.616131
3	0.111610	-0.997593	3.071103	5.035603	0.366655
4	0.897433	-0.356737	0.258483	4.835153	0.522968

03

Modelling

K-Means Clustering

Elbow Method



Clustering K-Means

```
# cluster n=4
kmeans = KMeans(n_clusters = 4, random_state = 0)

# fit model
kc = kmeans.fit(dfcust_std)
cluster_labels = kc.labels_

# add cluster
datacust_cluster = dfcust_std.assign(K_Cluster = cluster_labels)
datacust_cluster.head()
```

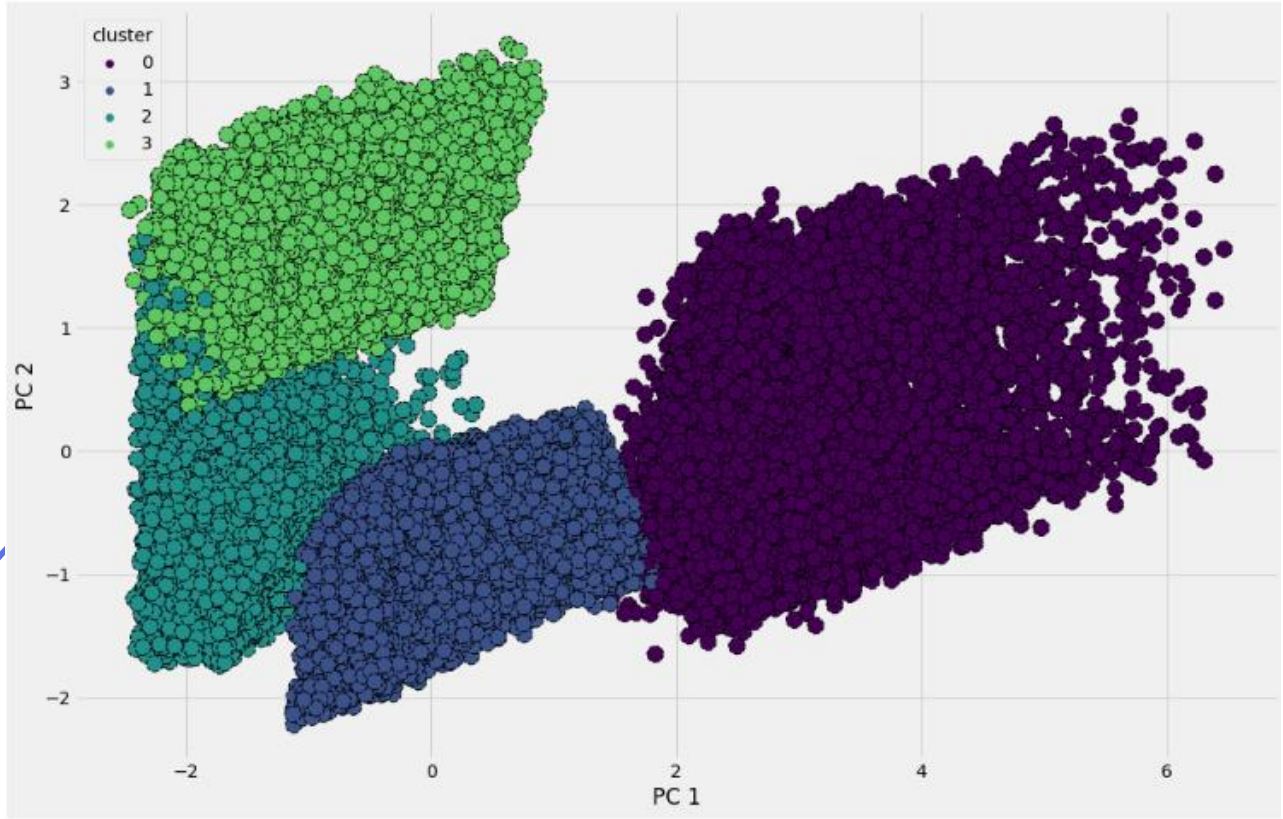
- Syntax

```
# add cluster to df_cust
cust_cluster = df_cust.assign(K_Cluster = cluster_labels)
cust_cluster.head()
```

	L	R	F	M	C	K_Cluster
1486	1561	41	18	76005	0.786950	0
1490	1686	69	9	74714	0.799971	0
1508	1118	57	10	75964	0.781936	0
1594	1534	1	37	77711	0.746389	0
1609	2190	121	12	75167	0.768661	0

- Table

Clustering K-Means with PCA Visualization

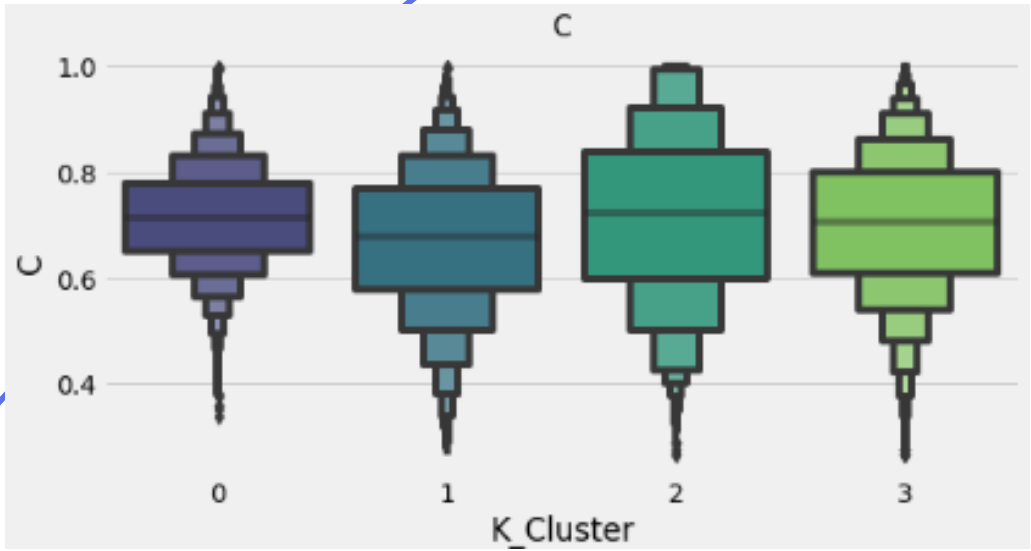


Interpretation

Summary statistics

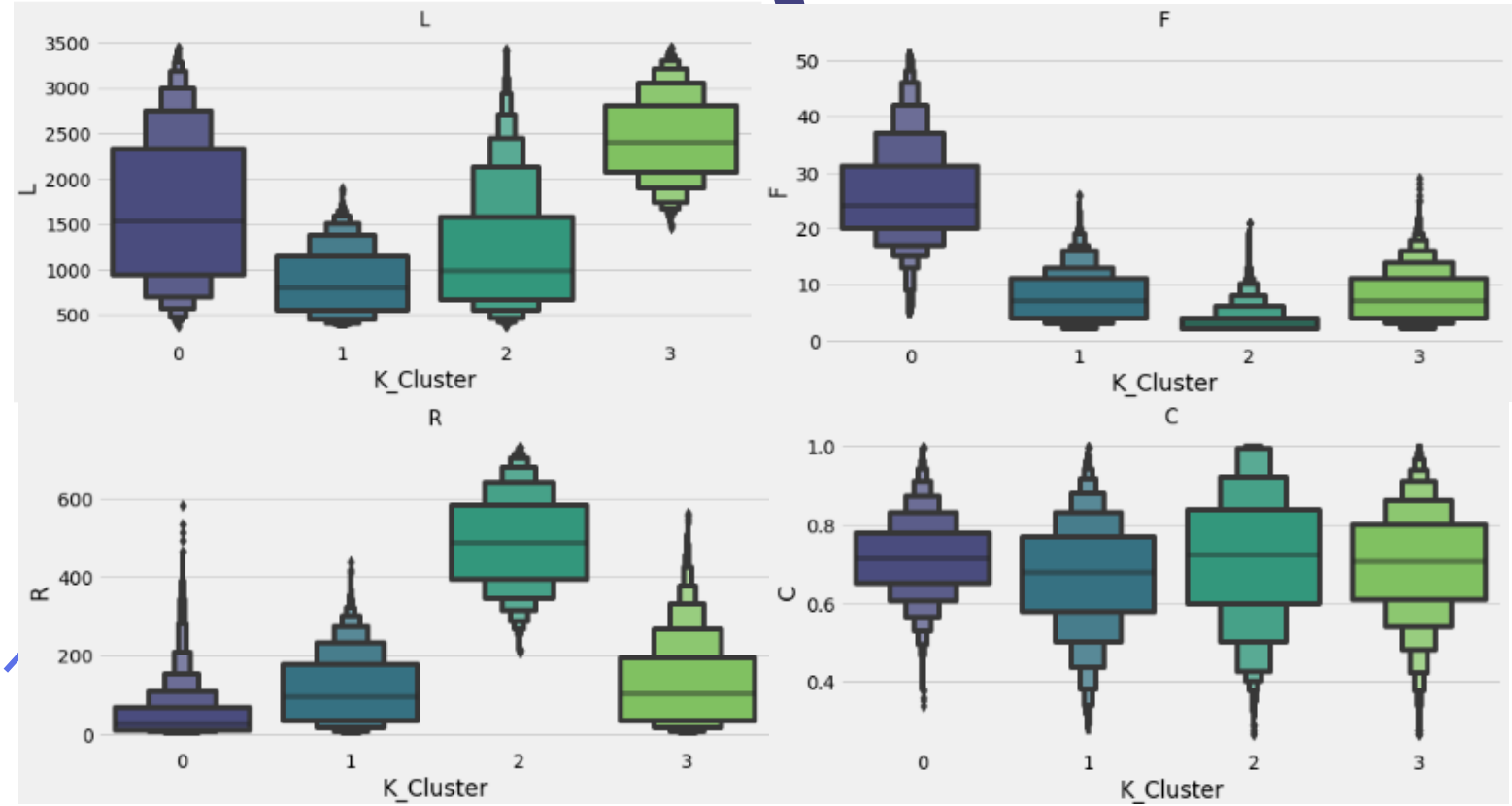
K_Cluster		L		R		F		M		C	
		mean	median	mean	median	mean	median	mean	median	mean	median
0	0	1649.972379	1535.0	49.594812	25.0	26.004563	24.0	36818.721028	34325.0	0.715935	0.714094
1	1	860.074690	797.0	114.092317	96.0	7.605338	7.0	10820.552208	9412.0	0.669370	0.678153
2	2	1186.243756	978.0	488.576439	487.0	3.602480	3.0	5416.018158	4153.0	0.715294	0.725000
3	3	2433.846578	2393.0	128.454835	103.0	7.780847	7.0	10922.126416	9591.5	0.700152	0.706322

LRFMC Boxplot Visualization



The difference of average discount used is not too significant among clusters

LRFMC Boxplot Visualization



04

Analysis and Recommendation

Customer characteristics and business recommendation

Cluster Characteristics

Cluster 0 (The most loyal member)

- The 2nd oldest member
- The shortest recency
- The highest flying frequency
- The highest flying distance

Cluster 1 (New member but fly often)

- The newest member
- The 2nd recency, mostly have flight in recent time
- Medium flying count
- Medium flying distance

Cluster 2 (Potential churned customer)

- The second newest member
- Havent't flight recently (> 1 year)
- Lowest flight frequency
- Lowest miles accumulated

Cluster 3 (Casual Customer)

- The oldest member
- Recency, Frequency, and flying distance are almost identic with cluster 1

Business Recommendation

- Loyalty Membership
- Avoid Churn
- Engage Cluster 1 to Become More Loyal Customer

Loyalty Membership

Create loyalty membership based on this clustering.
Example:

- Gold: Cluster 0
- Silver: Cluster 1 and 3
- Bronze: Cluster 2

Avoid Churn

Focus to **Cluster 2:**

- Push notification
- Special promotion if install airline apps or book flight for the first time since long time ago.
- Give local flight discount (considering their lowest distance accumulated)

Engage More Loyalty

Engage Cluster 1 to become loyal customers by getting more flights by **suggesting discount voucher/bundling package.**

Since **Cluster 1** are new members but have flight in recent time.

Thank you

Awalsyahrp@gmail.com
+62821 1393 2861
linkedin.com/in/awalsyahrp/
github.com/Awalsyah1995

Visit my repository to access the source code and dataset

Thanks

Do you have any questions?
youremail@freepik.com
+91 620 421 838
yourcompany.com