Rakamin Academy

UNSUPERVISED LEARNING

Summary Homework

1. EDA & Pre-Processing

a. Tipe data, missing values, duplicated value, range value

```
<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
```

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

```
MEMBER_NO              0
FFP_DATE               0
FIRST_FLIGHT_DATE      0
GENDER                 3
FFP_TIER               0
WORK_CITY           2269
WORK_PROVINCE       3248
WORK_COUNTRY          26
AGE                  420
LOAD_TIME              0
FLIGHT_COUNT           0
BP_SUM                 0
SUM_YR_1             551
SUM_YR_2             138
SEG_KM_SUM             0
LAST_FLIGHT_DATE       0
LAST_TO_END            0
AVG_INTERVAL           0
MAX_INTERVAL           0
EXCHANGE_COUNT         0
avg_discount           0
Points_Sum             0
Point_NotFlight        0
dtype: int64
```

```
df.duplicated().sum()
```

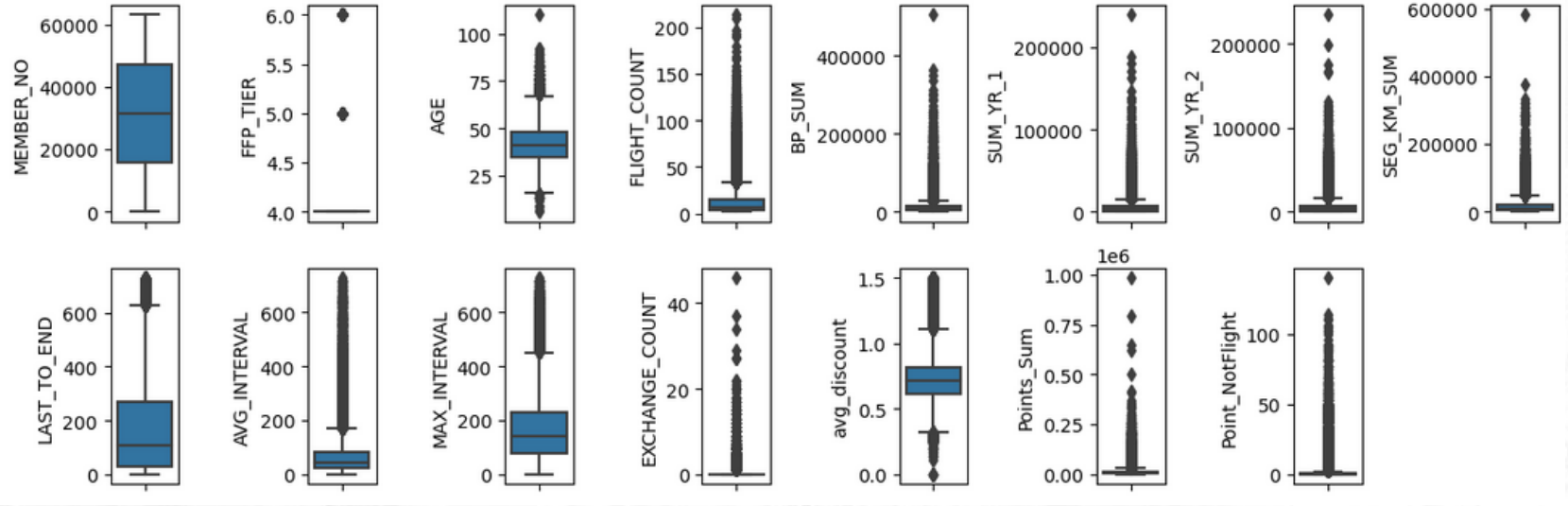
```
0
```

- Data set terdiri dari 62988 baris dan 23 kolom
- Tidak ditemukan ada baris yang duplikat
- Pada kolom GENDER, WORK_CITY, WORK_PROVINCE, WORK_COUNTRY, AGE, SUM_YR_1 dan SUM_YR_2 terdapat missing values
- Data yang mengandung tanggal namun tipedatanya belum sesuai, akan diganti menjadi datetime

1. EDA & Pre-Processing

a. Tipe data, missing values, duplicated value, range value

Outliers



1. EDA & Pre-Processing

Rakamin Academy

b. Statisik numerik & kategorik, distribusi numerik, dan unique value kategorik

Statistik Numerik

	MEMBER_NO	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	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	31494.500000	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	18183.213715	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	1.000000	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%	15747.750000	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%	31494.500000	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%	47241.250000	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	62988.000000	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

Statistik & Unique Value Kategorik

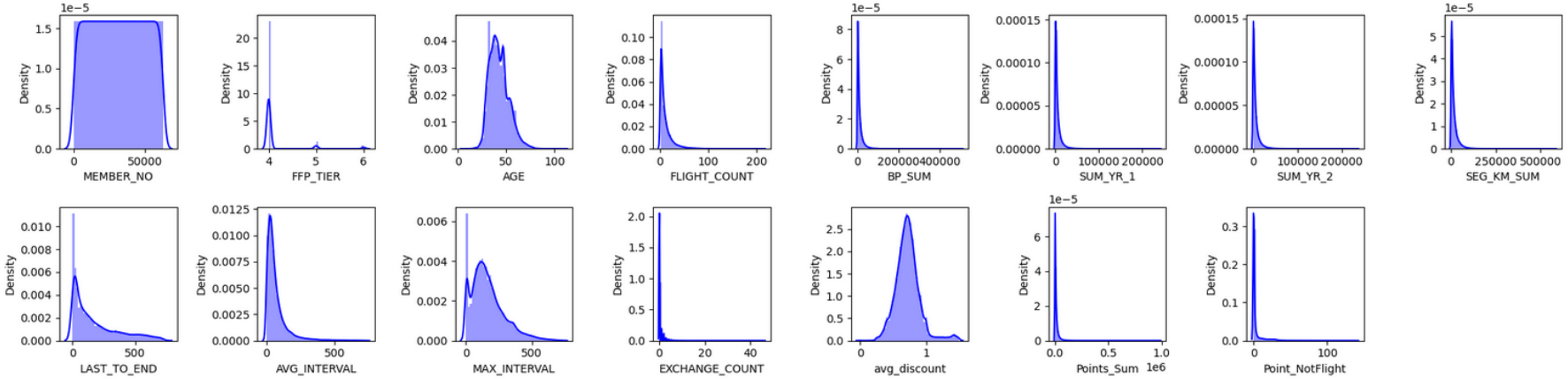
	GENDER	WORK_CITY	WORK_PROVINCE	WORK_COUNTRY	FFP_DATE	FIRST_FLIGHT_DATE	LOAD_TIME	LAST_FLIGHT_DATE
count	62985	60719	59740	62962	62988	62988	62988	62988
unique	2	3234	1165	118	3068	3406	1	731
top	Male	guangzhou	guangdong	CN	1/13/2011	2/16/2013	3/31/2014	3/31/2014
freq	48134	9386	17509	57748	184	96	62988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

- Dari beberapa **kolom numerik** mengindikasikan adanya **outlier** apabila diamati dari perbandingan antara **mean** dan **mediannya**.
- Kolom **AGE** yang menunjukkan umur customer menunjukkan angka yang **lumayan jauh** untuk **min** dan **max** yaitu pada umur 6 dan 110 tahun.
- Kolom **EXCHANGE_COUNT** dan **Point_NotFlight** memiliki nilai yang lumayan sangat aneh dari segi distribusi nilai, **75% customer** memiliki nilai 0, sedangkan **50% customer** diharapkan memiliki poin habis digunakan.

1. EDA & Pre-Processing

b. Statisik numerik & kategorik, distribusi numerik, dan unique value kategorik

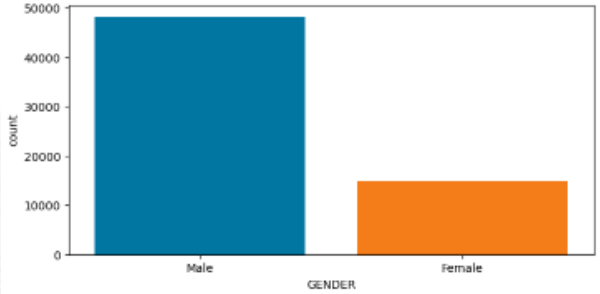
Distribusi Numerik

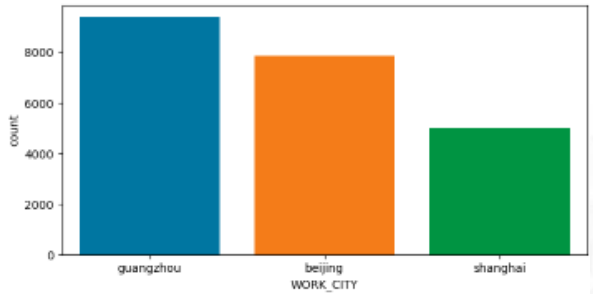


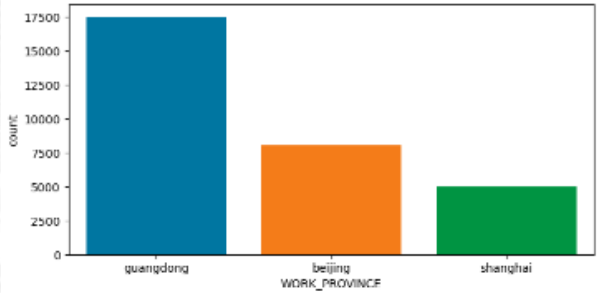
Dari distribusi diatas, menunjukan bahwa sebagian besar data memiliki **skew positif** yang cukup **ekstrim.**

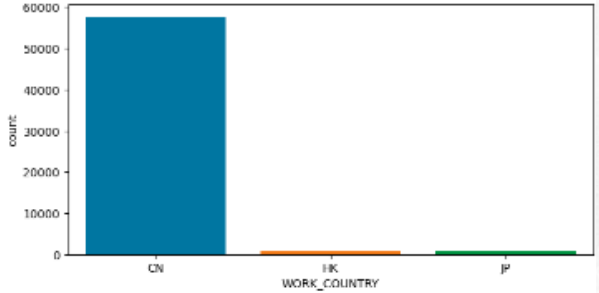
1. EDA & Pre-Processing

b. Statisik numerik & kategorik, distribusi numerik, dan unique value kategorik





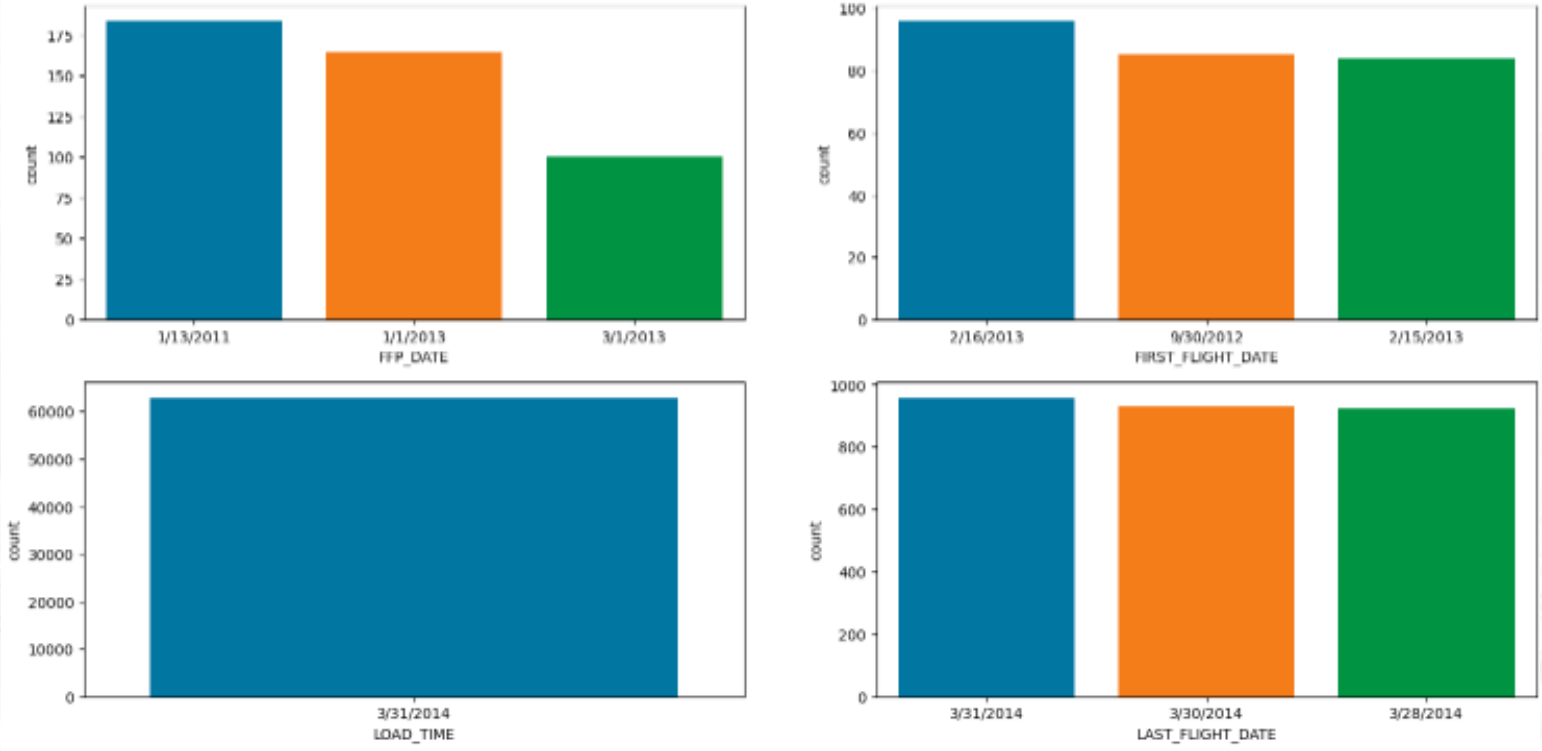




- Kolom **Gender** pada male lebih besar signifikan daripada female.
- Kolom **WORK_CITY**, untuk top 3 ditemukan :
  - guangzhou 9386
  - beijing 7845
  - shanghai 5001
- Kolom **WORK_PROVINCE**, untuk top 3 ditemukan :
  - guangdong 17509
  - beijing 8014
  - shanghai 4998
- Kolom **WORK_COUNTRY**, untuk top 3 ditemukan :
  - CN 57748
  - HK 991
  - JP 875

1. EDA & Pre-Processing

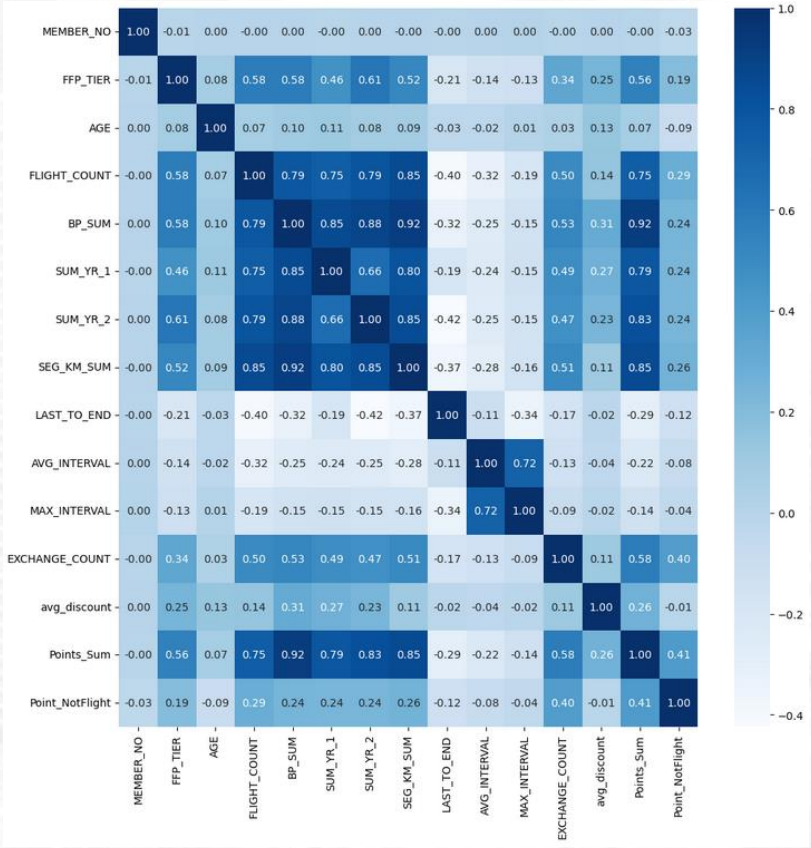
b. Statisik numerik & kategorik, distribusi numerik, dan unique value kategorik



1. EDA & Pre-Processing

Fitur yang berkorelasi apda dataset :

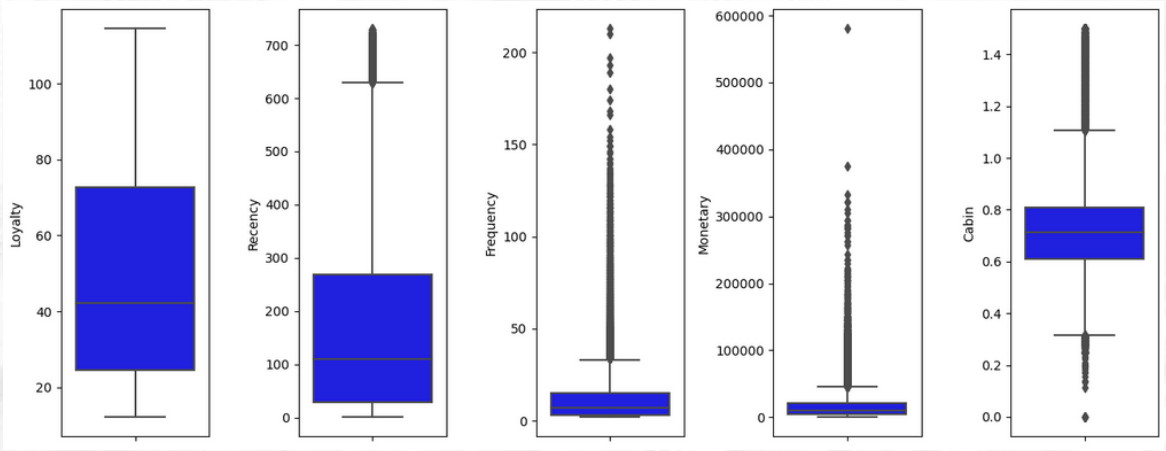
1. FFP_TIER
2. FLIGHT_COUNT
3. BP_SUM
4. SUM_YR_1
5. SUM_YR_2
6. SEG_KM_SUM
7. EXCHANGE_COUNT
8. Points_Sum



2. FEATURE ENGINEERING

Fitur dipilih menggunakan model LRFMC dimana fitur yang digunakan untuk model ini adalah: load_time, ffp_date, last_to_end, flight_count, seg_km_sum, avg_discount

- **Loyalty L** = LOAD_TIME - FFP_DATE : Jumlah bulan sejak customer bergabung hingga waktu pengambilan dataset (melihat apakah customer lama/baru)
- **Recency R** = LAST_TO_END : Jarak waktu penerbangan terakhir ke pesanan penerbangan paling akhir
- **Frequency F** = FLIGHT_COUNT : Jumlah penerbangan customer
- **Monetary M** = SEG_KM_SUM : Total jarak penerbangan yg sudah dilakukan
- **Cabin C** = avg_discount : Rata rata discount yang didapat customer

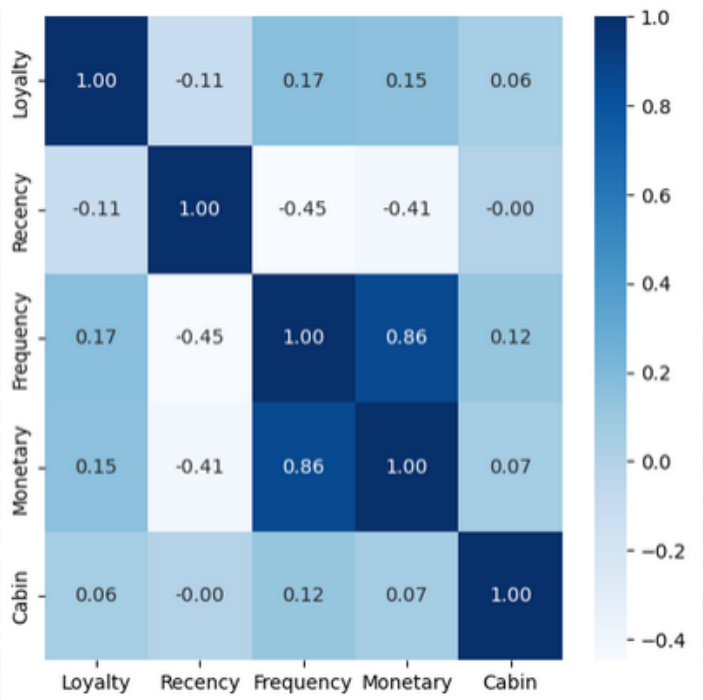


2. FEATURE ENGINEERING

Karena akan digunakan **K-Means** yang merupakan Machine Learning berdasarkan jarak sehingga scalling data yang digunakan adalan **STANDARISASI**.

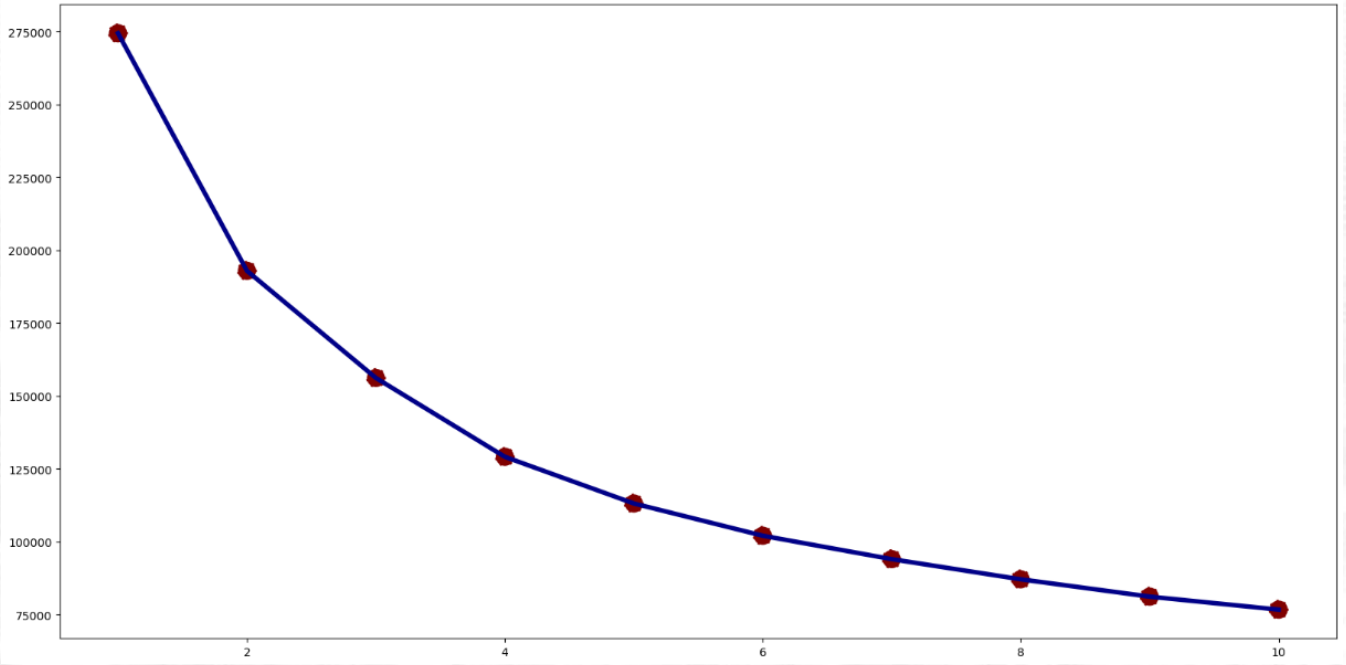
```
# Standardize data
std = StandardScaler().fit_transform(df_LRFMC)
df_LRFMC_std = pd.DataFrame(std, columns = list(df_LRFMC))
df_LRFMC_std
```

	Loyalty	Recency	Frequency	Monetary	Cabin
0	1.208373	-0.591563	4.289657	4.380582	3.307799
1	-0.895824	-0.959739	3.572130	4.260724	3.426038
2	-0.471888	-0.981397	4.187153	4.061673	3.451065
3	-0.995854	-0.938082	3.367123	4.045578	3.281898
4	-0.170608	-0.927253	3.367123	3.966527	3.288521
...	...	...	...	...	...
59470	0.591523	-0.331674	-0.835531	-1.010816	-0.222155
59471	-0.121784	-0.775651	-0.835531	-0.961391	-3.033753
59472	-1.188768	1.525451	-0.835531	-0.972644	-2.875798
59473	-0.455217	0.540038	-0.835531	-0.982899	-2.686252
59474	-1.291180	1.065231	-0.835531	-0.983896	-2.875798



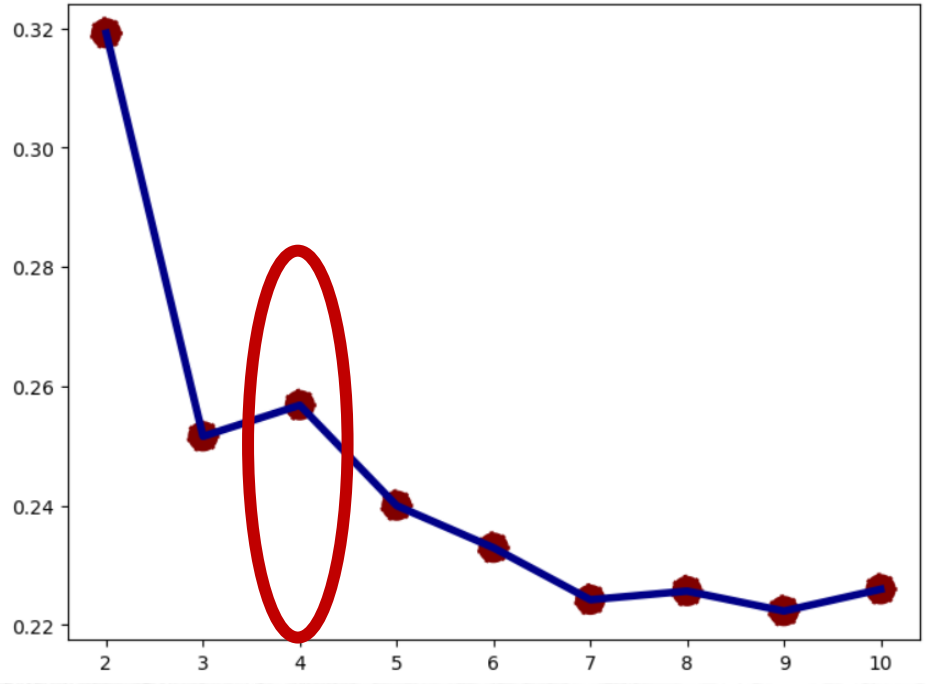
3. Modelling

Berikut adalah plot untuk “Elbow Method” algoritma K-means untuk menemukan jumlah cluster yang tepat.



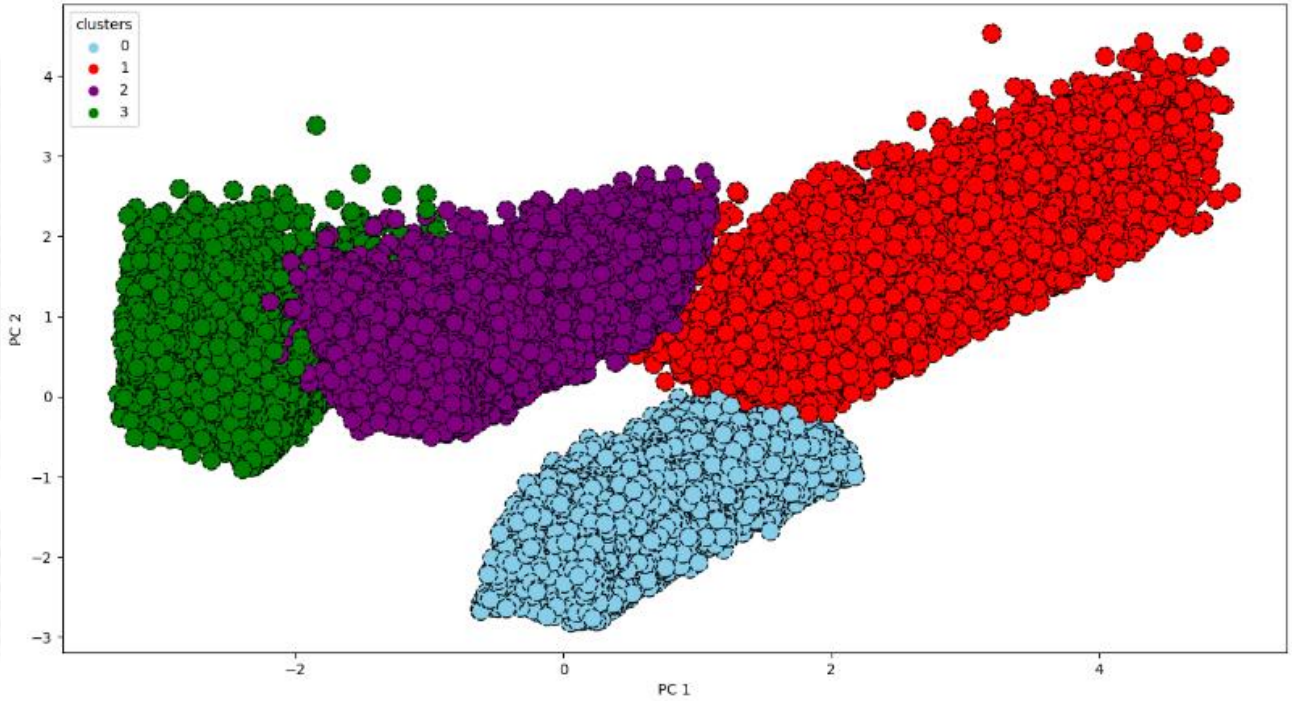
3. Modelling

Berdasarkan “Silhouette Score” menggunakan “Elbow Method” di bawah, ditemukan jumlah cluster yang paling optimal yaitu berjumlah 4 cluster. Apabila dipilih lebih dari 4 cluster, maka penurunan nilai inersia akan tidak terlalu besar.



3. Clustering using PCA

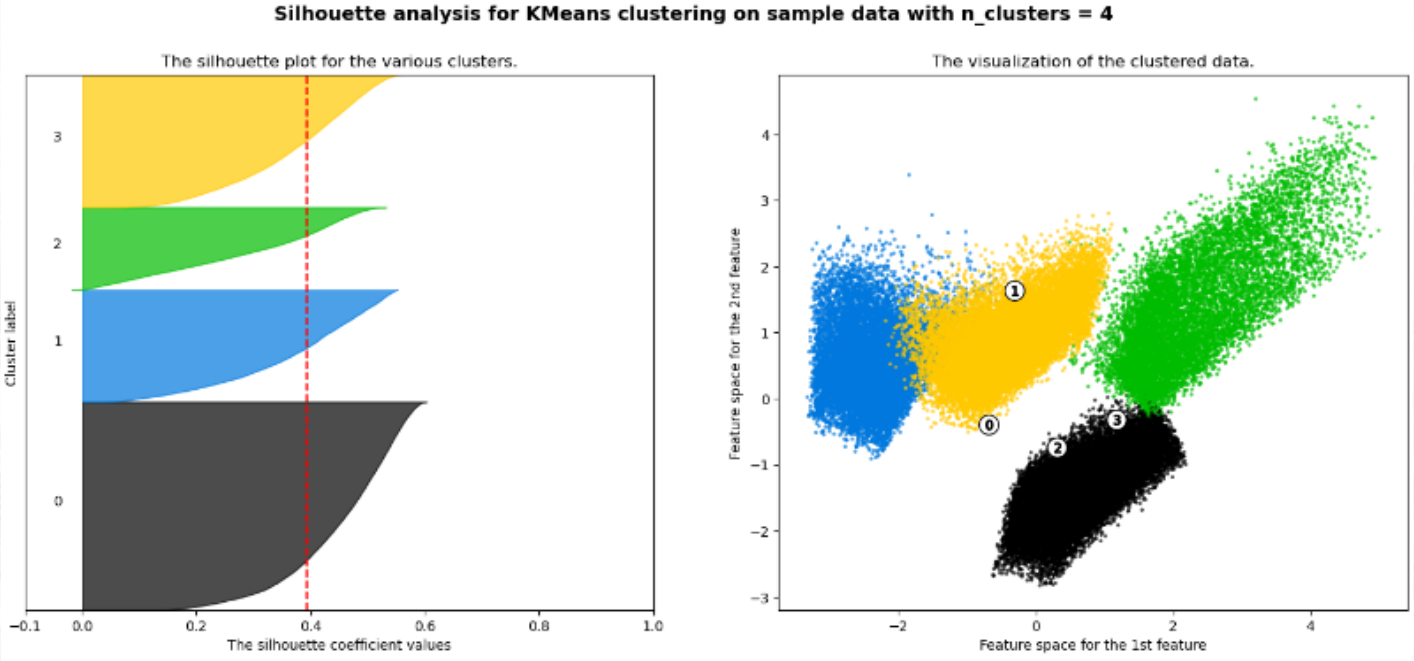
Berikut adalah hasil clustering dengan PCA terbagi menjadi 4 cluster



	PC 1	PC 2	clusters
0	4.704661	4.421910	1
1	4.511718	2.920081	1
2	4.723844	3.317943	1
3	4.295256	2.677554	1
4	4.244723	3.111765	1

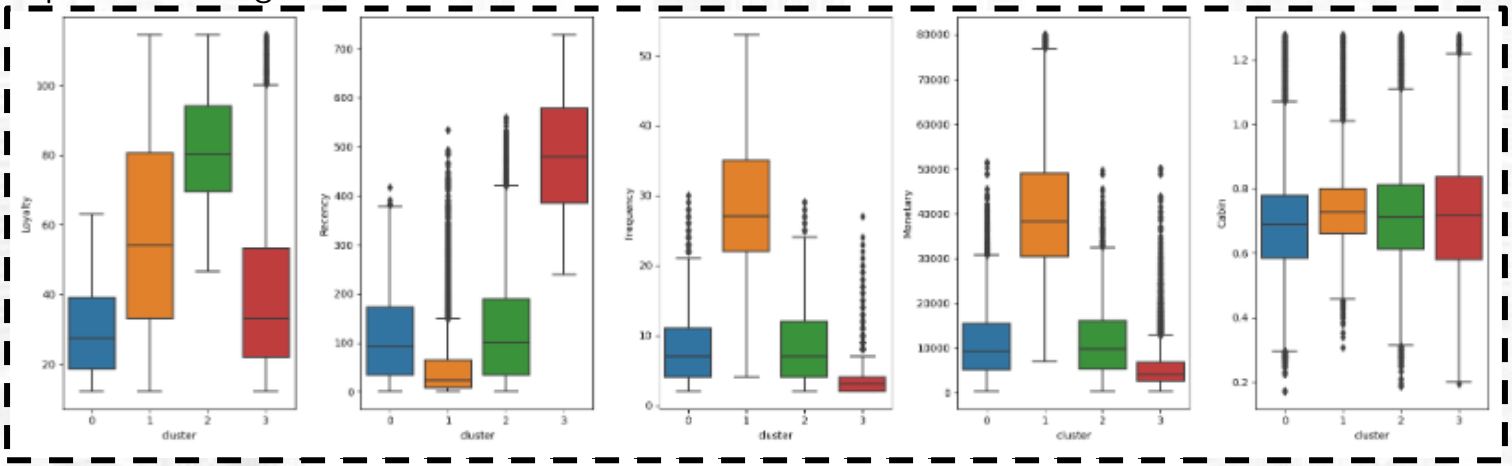
3. Silhouette Score Plot

Di bawah ini adalah Silhouette Plot untuk melihat pembagian menjadi 4 cluster



3. Interpretasi Cluster

Dari hasil clustering menggunakan k-means diperoleh hasil statistic mean dan median dari setiap fitur setiap cluster sebagai berikut:



	Loyalty		Recency		Frequency		Monetary		Cabin	
	mean	median	mean	median	mean	median	mean	median	mean	median
cluster										
0	29.294559	27.400000	109.635853	92.0	7.720209	7.0	11008.085697	9313.5	0.683366	0.687689
1	57.426506	54.066667	46.864909	23.0	28.503214	27.0	40732.175400	38196.0	0.740208	0.727464
2	81.546363	80.166667	124.829328	101.0	8.043041	7.0	11412.216633	9835.0	0.714499	0.712327
3	39.965579	33.166667	485.225599	480.0	3.624859	3.0	5585.111594	4154.0	0.708767	0.716667

3. Deskripsi Cluster

- **Cluster 0** : menunjukkan customer yang sudah bergabung sekitar hampir 7 tahun, jarang melakukan penerbangan dan jarak penerbangan yang dilakukan tidak jauh (retained customer)

- **Cluster 1** : menunjukkan customer yang sudah bergabung sekitar hampir 5 tahun, aktif dalam menggunakan jasa penerbangan dan sering berpergian dengan jarak yang cukup jauh (High-value customer)

- **Cluster 2 :** menunjukkan customer yang sudah bergabung sekitar 2 tahun (tergolong baru), sering melakukan penerbangan dan jaraknya cukup jauh. (potential customer)

- **Cluster 3** : menunjukkan customer yang sudah bergabung sekitar 3 tahun (tergolong baru) namun sangat jarang menggunakan jasa penerbangan dan sekali menggunakan juga tidak terlalu jauh (low-value customer)

4. Rekomendasi Strategi Bisnis

- Fokus mengintensifkan strategi pemasaran yang tergabung dalam cluster 1. Untuk meningkatkan jumlah penerbangan yang masih sedikit dan sudah lama tidak melakukan penerbangan, strategi pemasaran dapat dilakukan dengan memberikan promo diskon atau kerjasama dengan travel agent untuk memberikan liburan yang menarik. paket promo, agar para pelanggan ini tertarik untuk terbang kembali menggunakan maskapai ini.

- Fokus pada mempertahankan pelanggan yang tergabung dalam cluster 2 dengan menawarkan keanggotaan (membership) maskapai premium.