Rakamin Academy

W3.Solutions()

Dokumen
Laporan
Homework
Unsupervised
Learning

1. EDA & Pre-Processing

a. tipe data, missing values, duplicated values, dan 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
```

Tipe data sudah tepat

```
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
```

terdapat missing values

```
[ ]  df.duplicated().sum()

0
```

tidak ada duplicated values

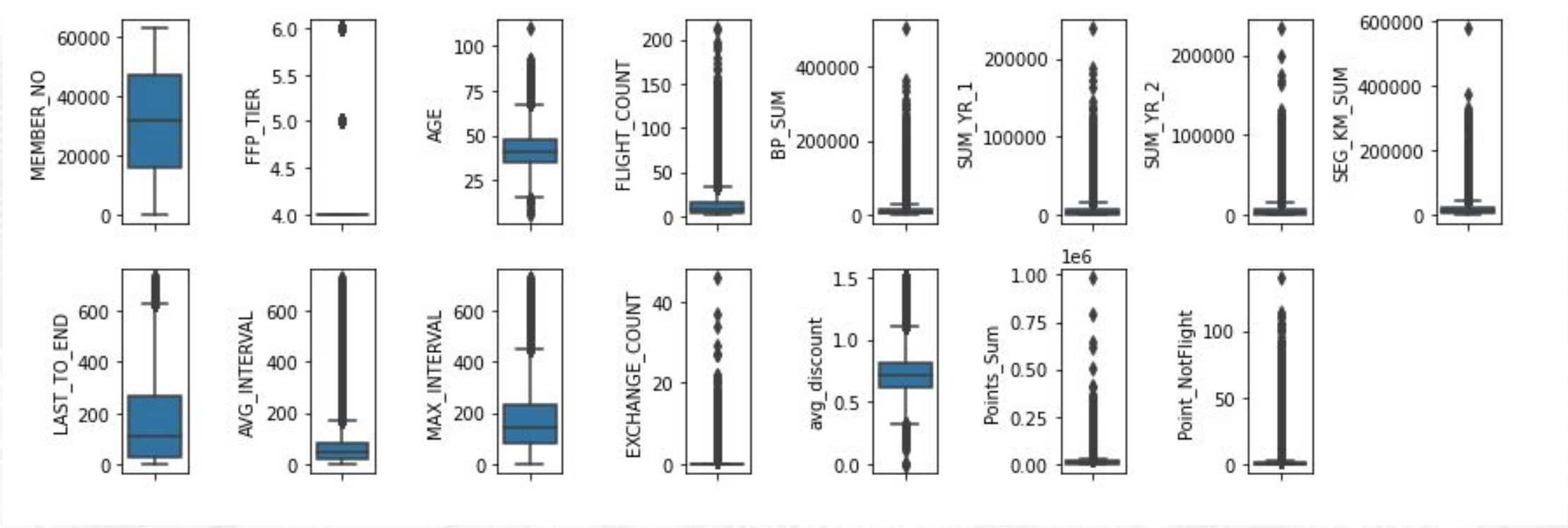
- Pengisian data pada feature **AGE, SUM_YR_1,** dan **SUM_YR_2** dengan **median**
- Drop missing values pada feature kategorikal **WORK_CITY, WORK_PROVINCE, WORK_COUNTRY** dan **GENDER**

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

MEMBER_NO          0
FFP_DATE           0
FIRST_FLIGHT_DATE  0
FFP_TIER           0
AGE                0
LOAD_TIME          0
FLIGHT_COUNT       0
BP_SUM             0
SUM_YR_1           0
SUM_YR_2           0
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
```

data setelah handle duplicated dan missing values

a. tipe data, missing values, duplicated values, dan range value



penampakan outliers

b. **statistik kolom numerik dan kategorikal**, bentuk distribusi kolom (numerik), dan jumlah unique value (kategorikal)

Rakamin Academy

statistik numerikal

```
nums= ['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']
df[nums].describe()
```

	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

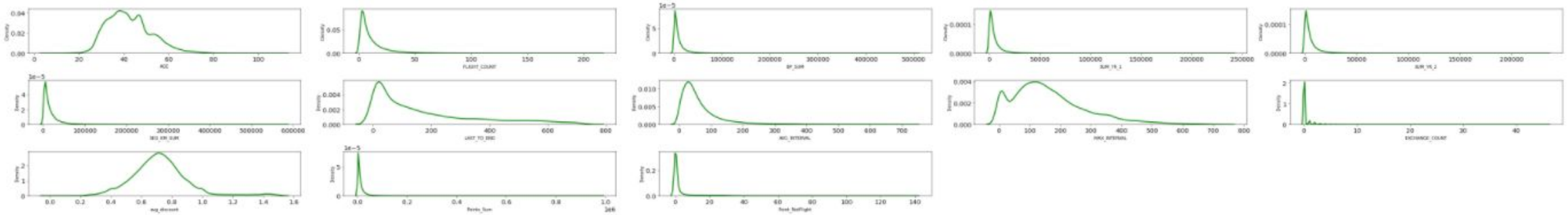
statistik kategorikal

```
cats= ['FFP_DATE','FIRST_FLIGHT_DATE','GENDER','WORK_CITY','WORK_PROVINCE','WORK_COUNTRY','LOAD_TIME','LAST_FLIGHT_DATE']
df[cats].describe()
```

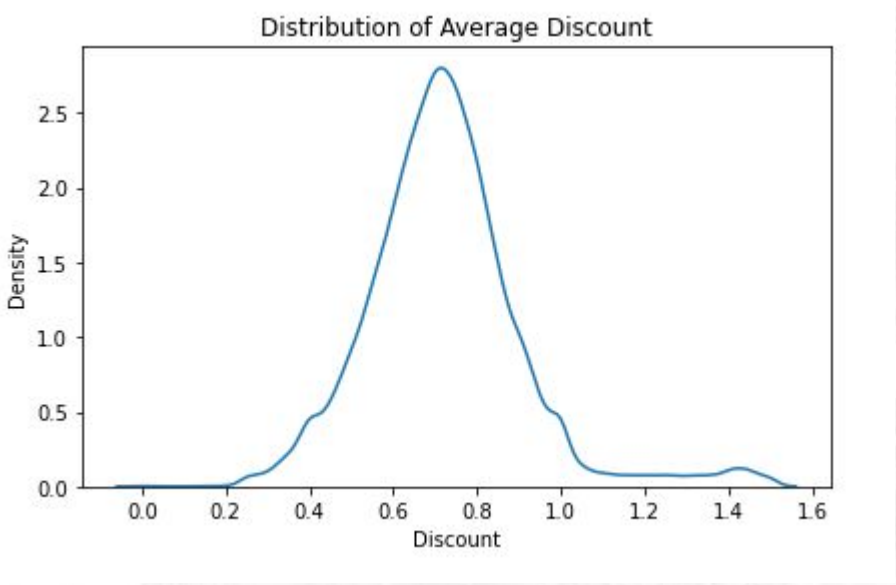
	FFP_DATE	FIRST_FLIGHT_DATE	GENDER	WORK_CITY	WORK_PROVINCE	WORK_COUNTRY	LOAD_TIME	LAST_FLIGHT_DATE
count	62988	62988	62985	60719	59740	62962	62988	62988
unique	3068	3406	2	3234	1165	118	1	731
top	1/13/2011	2/16/2013	Male	guangzhou	guangdong	CN	3/31/2014	3/31/2014
freq	184	96	48134	9386	17509	57748	62988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

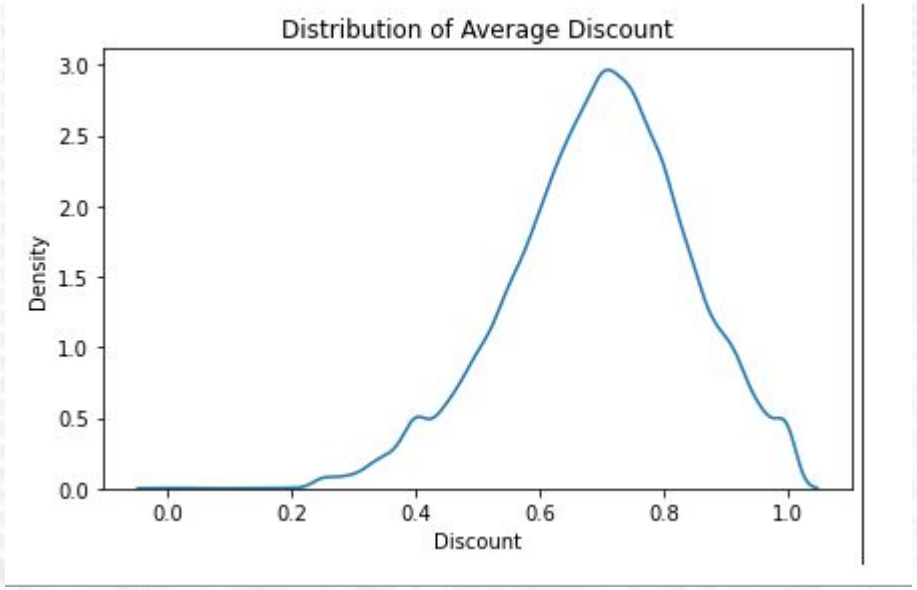
b. statistik kolom numerik dan kategorikal, **bentuk distribusi kolom (numerik)**, dan jumlah unique value (kategorikal)

distribusi kolom numerik



insight:





average discount yang > 1 akan di drop, karena dianggap tidak valid (discount max =1 atau 100%)

Unique value categorical :

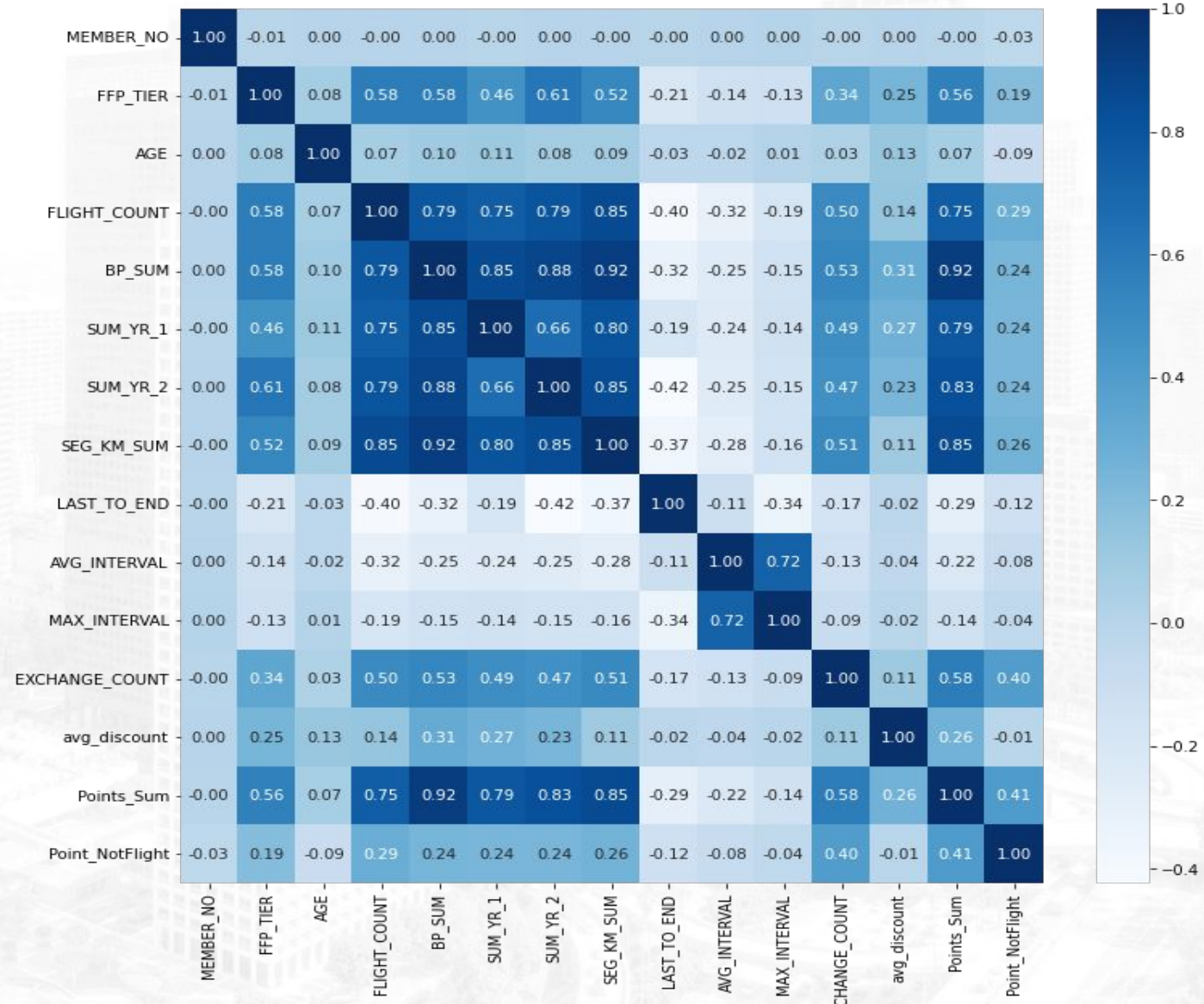
```
cats= ['FFP_DATE','FIRST_FLIGHT_DATE','GENDER','WORK_CITY','WORK_PROVINCE','WORK_COUNTRY','LOAD_TIME','LAST_FLIGHT_DATE']
df[cats].describe()
```

	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

Feature Correlation :

Berdasarkan heatmap pertama:

- Fitur yang nilai korelasinya rendah dan dianggap tidak berhubungan dalam penyelesaian masalah akan didrop dari dataset: ‘member_no’, ‘age’, ‘exchange_count’, ‘sum_yr_1’, ‘sum_yr_2’, ‘point_notflight’, ‘avg_interval’, ‘max_interval’.

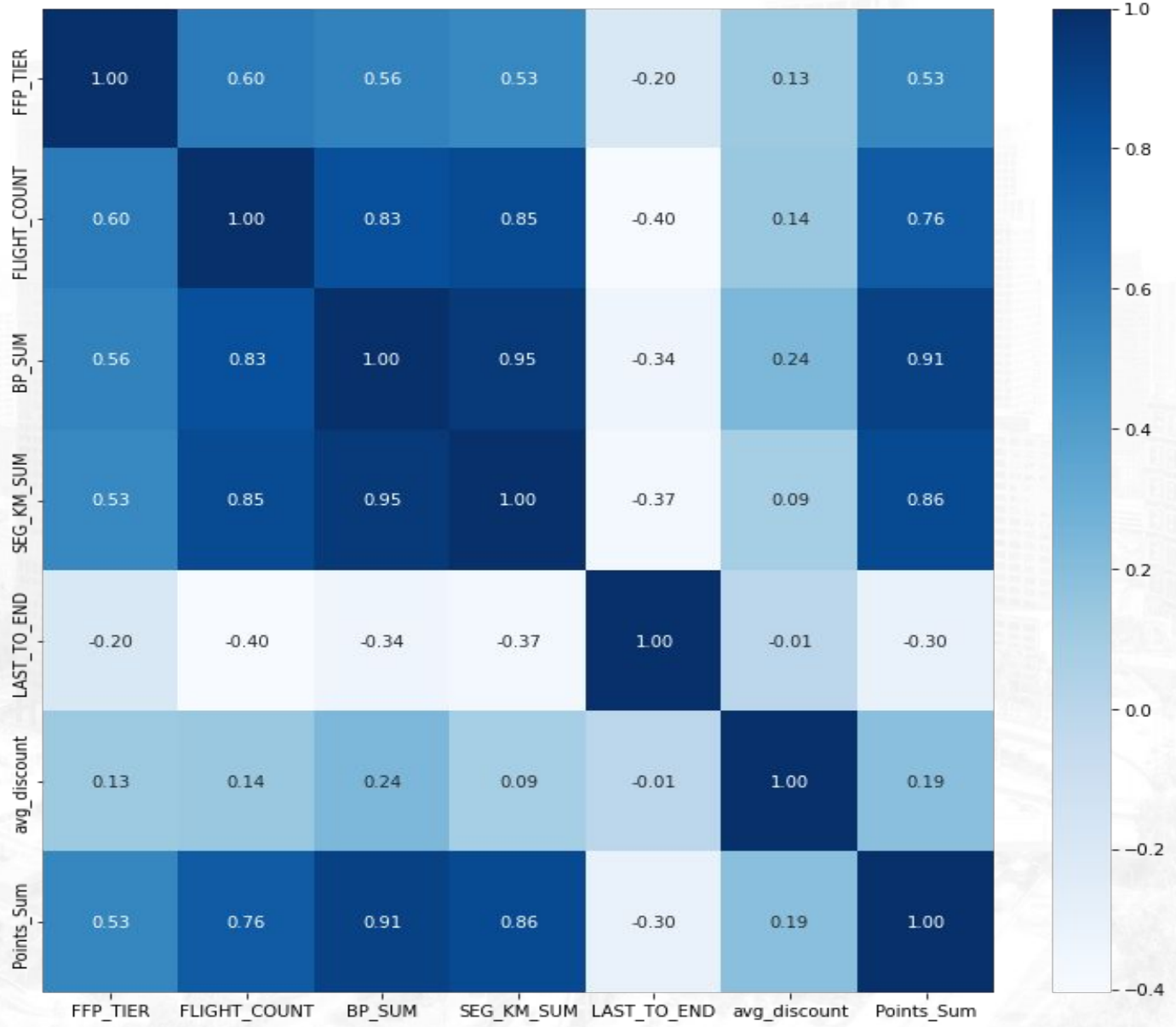


Heatmap 1

Feature Correlation:

Berdasarkan heatmap kedua:

- Fitur penting: 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’.
- Dari EDA (heatmap) dilihat juga fitur yang berkorelasi sangat tinggi seperti ‘bp_sum’, ‘seg_km_sum’, dan ‘point_sum’ sehingga dalam modeling bisa memilih salah satu saja yaitu ‘seg_km_sum’ sehingga ‘bp_sum’ dan ‘point_sum’ akan didrop



Heatmap 2

2. Feature Selection

Jadi 6 fitur yang dipakai adalah :

1. FFP_TIER,
2. SEG_KM_SUM,
3. LAST_TO_END,
4. avg_discount,
5. Meeting_Time
6. Flight_Count/Year

```
dfa.info()
```

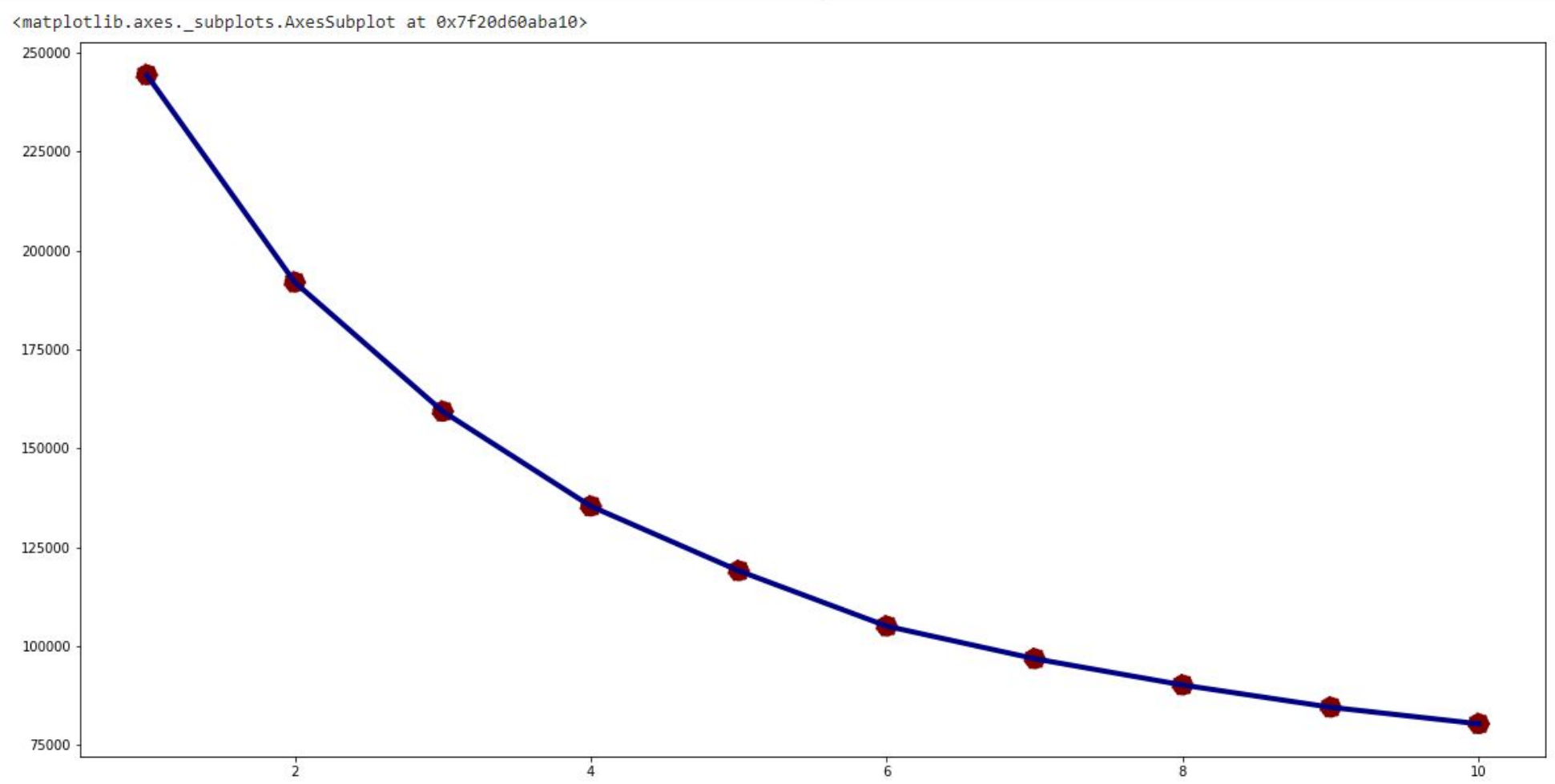
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 60041 entries, 0 to 62987
Data columns (total 6 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   FFP_TIER           60041 non-null  int64  
 1   SEG_KM_SUM         60041 non-null  int64  
 2   LAST_TO_END        60041 non-null  int64  
 3   avg_discount       60041 non-null  float64
 4   Meeting_Time       60041 non-null  float64
 5   Flight_Count/Year  60041 non-null  float64
dtypes: float64(3), int64(3)
memory usage: 3.2 MB
```

Feature Engineering

- Dari feature LAST_FLIGHT_DATE, FIRST_FLIGHT_DATE, dan FLIGHT COUNT, dilakukan perhitungan untuk menentukan rata-rata penerbangan per tahun dengan output pada feature Flight_Count/Year.
- Dibuat feature Meeting_Time untuk menghitung jumlah bulan antara LOAD_TIME dan FPP_DATE.
- Setelah feature engineering, drop feature yang tidak digunakan kembali, dan handling outlier, dilakukan standarisasi pada dataset sehingga data dapat siap untuk modelling.
- Mengubah beberapa feature dengan tipe kategori menjadi tipe berformat tanggal atau *datetime*.

3. Modeling

Dengan menggunakan Elbow method maka jumlah cluster yang digunakan sebanyak **4 cluster**



Clustering K-Means

Melakukan clustering k-Means dengan jumlah **4 cluster**

```
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=4, random_state=0)
kmeans.fit(dfs.values)
```

```
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
       n_clusters=4, n_init=10, n_jobs=None, precompute_distances='auto',
       random_state=0, tol=0.0001, verbose=0)
```

```
dfs['cluster'] = kmeans.labels_
dfs.head()
```

	SEG_KM_SUM	LAST_TO_END	avg_discount	Meeting_Time	Flight_Count/Year	cluster
0	2.946823	-0.433412	2.186105	-0.640830	2.364728	1
1	3.151836	-0.776593	1.816499	0.463661	0.441358	1
2	3.006257	-0.242755	2.044803	2.258308	-0.290352	0
3	3.039276	1.225298	1.937261	0.669624	-0.609352	0
4	3.069732	-0.871921	1.869654	0.002352	1.326259	1

Evaluasi Cluster

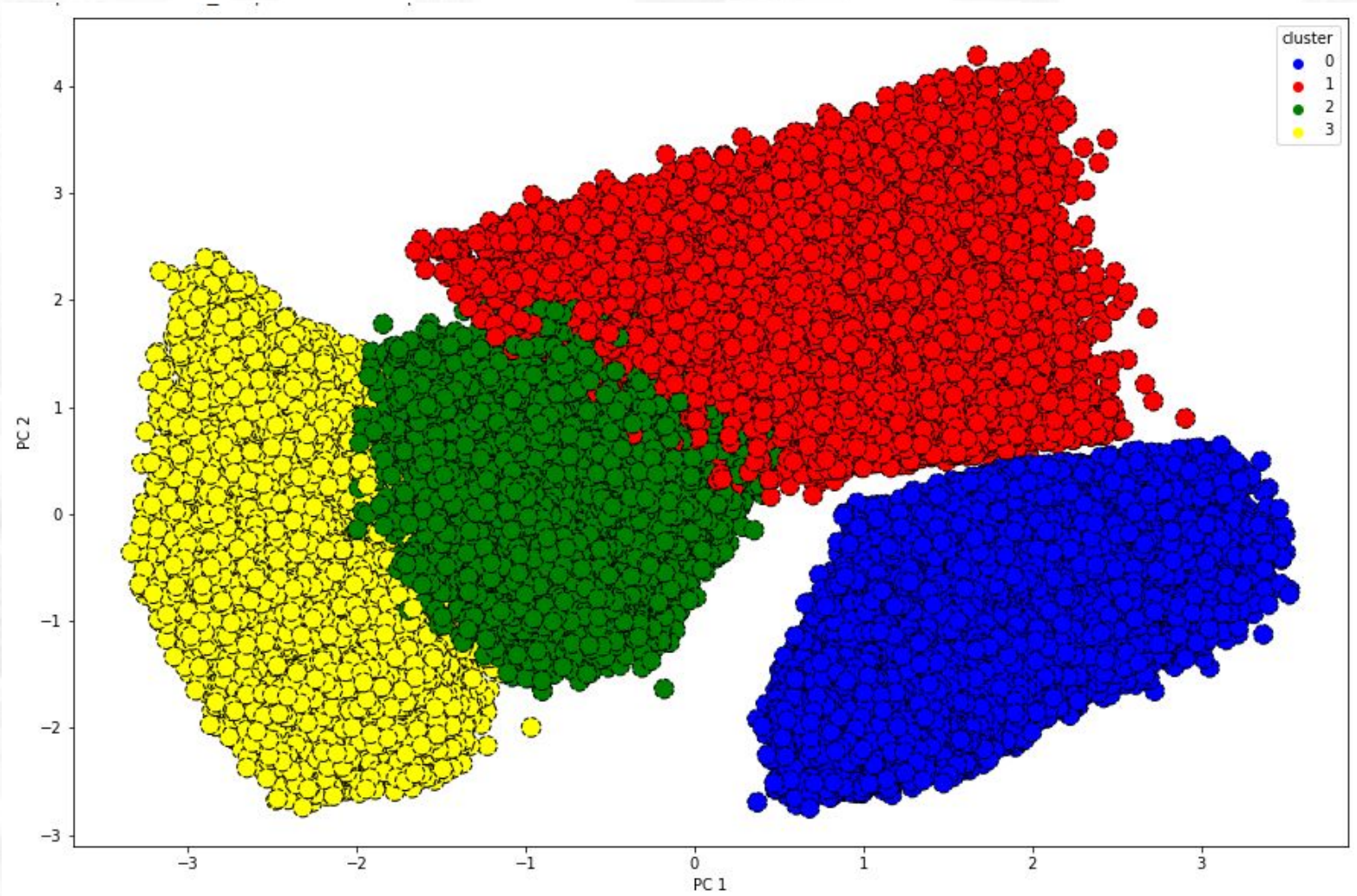
Menggunakan PCA untuk melihat visualisasi hasil clustering

```
data_pca = pd.DataFrame(data = pcs, columns = ['PC 1', 'PC 2'])
data_pca['cluster'] = dfs['cluster']
data_pca.head()
```

	PC 1	PC 2	cluster
0	1.850496	2.894874	1
1	2.324234	0.969372	1
2	3.152077	-0.889668	0
3	1.945420	-0.335250	0
4	2.249357	1.870184	1

Hasil Cluster PCA

Dan mendapatkan hasil clustering dengan PCA sebagai berikut :



4. Interpretasi Cluster

Dari hasil clustering dengan k-means diperoleh hasil statistik mean dan median dari setiap fitur di setiap cluster adalah sebagai berikut:

	index		SEG_KM_SUM		LAST_TO_END		avg_discount		Meeting_Time		Flight_Count/Year	
	mean	median	mean	median	mean	median	mean	median	mean	median	mean	median
cluster												
0	26435.201628	23625	16563.558510	14608.0	87.372948	67	0.710600	0.713211	80.687529	80.395901	2.259611	1.875443
1	19195.172600	15982	22257.667047	21543.0	102.254593	59	0.712065	0.715332	29.596675	25.101131	10.550579	9.961159
2	40131.050767	39889	8195.435842	7196.0	113.475673	98	0.657686	0.662702	32.427532	30.850736	3.673931	3.290473
3	44748.749663	46522	5888.867640	4663.5	424.043820	419	0.735520	0.746035	51.683586	47.688864	2.944816	1.727888

Deskripsi Masing-Masing Cluster

Adapun penjelasan terhadap ke-4 cluster yang terbentuk adalah sebagai berikut:

1. Cluster 0 merupakan pelanggan yang telah mendaftar sebagai member cukup lama (rata-rata 81 bulan) dengan frekuensi terbang sedang (rata-rata 16.564 km) atau sering melakukan penerbangan jarak jauh mengingat jumlah penerbangan per tahunnya cenderung rendah (rata-rata 2 kali per tahun).
2. Cluster 1 merupakan pelanggan baru (rata-rata 30 bulan) dengan frekuensi terbang tinggi (rata-rata 11 penerbangan per tahun), dengan penerbangan jarak jauh (rata-rata 22.258 km).
3. Cluster 2 merupakan pelanggan dengan durasi menjadi member, frekuensi terbang, dan jarak penerbangan sedang namun memiliki jarak waktu penerbangan terakhir ke pesanan penerbangan paling akhir rendah.
4. Cluster 3 merupakan pelanggan dengan durasi menjadi member, frekuensi terbang, dan jarak penerbangan sedang namun memiliki jarak waktu penerbangan terakhir ke pesanan penerbangan paling akhir tinggi.

Rekomendasi Strategi Bisnis

1. Memberikan apresiasi kepada member lama maupun member baru seperti diskon ataupun promo khusus untuk member agar mereka tetap berlangganan. promo ataupun diskon disesuaikan dengan segmentasi mereka seperti member yang suka berpergian jauh maka dikasih promo dengan jarak pererbangan jauh, begitu pula jarang penerbangan sedang maupun dekat.
2. Perusahaan dapat menarik pelanggan baru dan mempertahankan pelanggan lama dengan cara meningkatkan kualitas pelayanan dan memberikan inovasi berkelanjutan sesuai dengan segmentasi mereka.