Rakamin Academy

Anaconda

Dokumen Laporan Homework Clustering

1. EDA

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 sesuai kecuali yang berkaitan dengan tanggal

```
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
```

Ada missing values

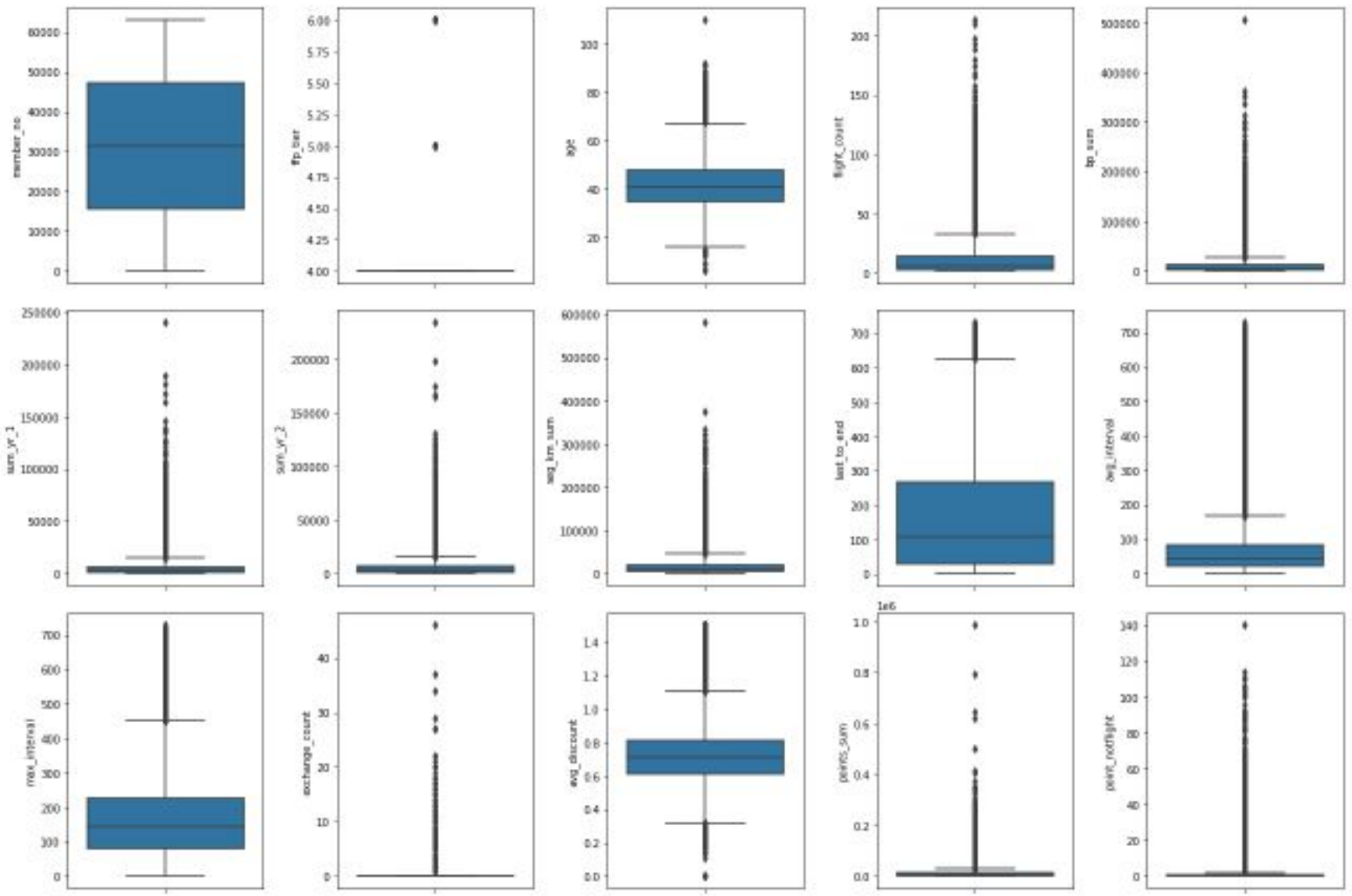
```
df.duplicated().sum()
```

```
0
```

Tidak duplikasi data

Kolom-kolom yang memiliki missing values:

- Numerik: age, sum_yr_1, sum_yr_2
- Kategorikal: gender, work_city, work_province, work_country
- Kolom numerik tersebut akan diimpute dengan median masing-masing sedangkan kolom kategorikal tersebut akan didrop.



Penampakan Outlier.

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

- Numerik

```
df.describe().T
```

	count	mean	std	min	25%	50%	75%	max
member_no	60041.0	31505.292833	18191.245419	1.0	15736.000000	31528.00000	47269.000000	62988.0
ffp_tier	60041.0	4.081727	0.328371	4.0	4.000000	4.00000	4.000000	6.0
age	59652.0	42.174797	9.771595	6.0	35.000000	41.00000	48.000000	110.0
flight_count	60041.0	11.654220	13.756290	2.0	3.000000	7.00000	14.000000	210.0
bp_sum	60041.0	10034.104778	14010.728738	0.0	2432.000000	5457.00000	12133.000000	505308.0
sum_yr_1	59499.0	4974.437205	7059.121002	0.0	968.000000	2700.00000	6285.500000	239560.0
sum_yr_2	59907.0	5210.832374	7667.312231	0.0	755.000000	2676.00000	6531.000000	234188.0
seg_km_sum	60041.0	16772.027931	20335.635374	368.0	4713.000000	9878.00000	20893.000000	580717.0
last_to_end	60041.0	177.260855	184.117824	1.0	30.000000	109.00000	270.000000	731.0
avg_interval	60041.0	67.875748	77.374097	0.0	23.500000	44.87500	82.400000	728.0
max_interval	60041.0	166.267517	123.569936	0.0	79.000000	143.00000	228.000000	728.0
exchange_count	60041.0	0.299629	1.062595	0.0	0.000000	0.00000	0.000000	46.0
avg_discount	60041.0	0.695875	0.144030	0.0	0.605626	0.70339	0.794527	1.0
points_sum	60041.0	11561.295448	17908.780708	0.0	2684.000000	6070.00000	13570.000000	985572.0
point_notflight	60041.0	2.729718	7.402475	0.0	0.000000	0.00000	1.000000	140.0

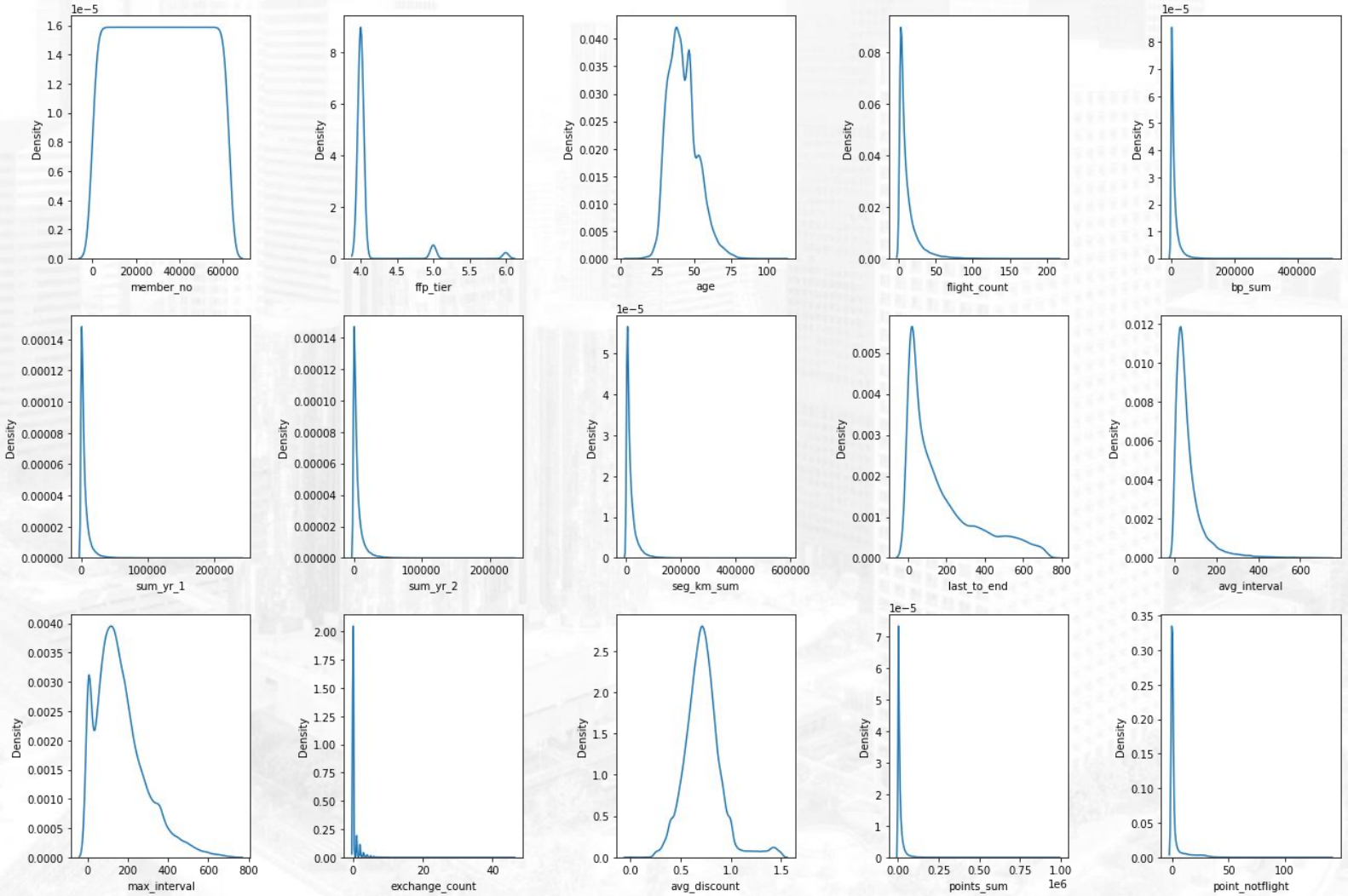
- Kategorikal

```
df.describe(exclude=np.number).T
```

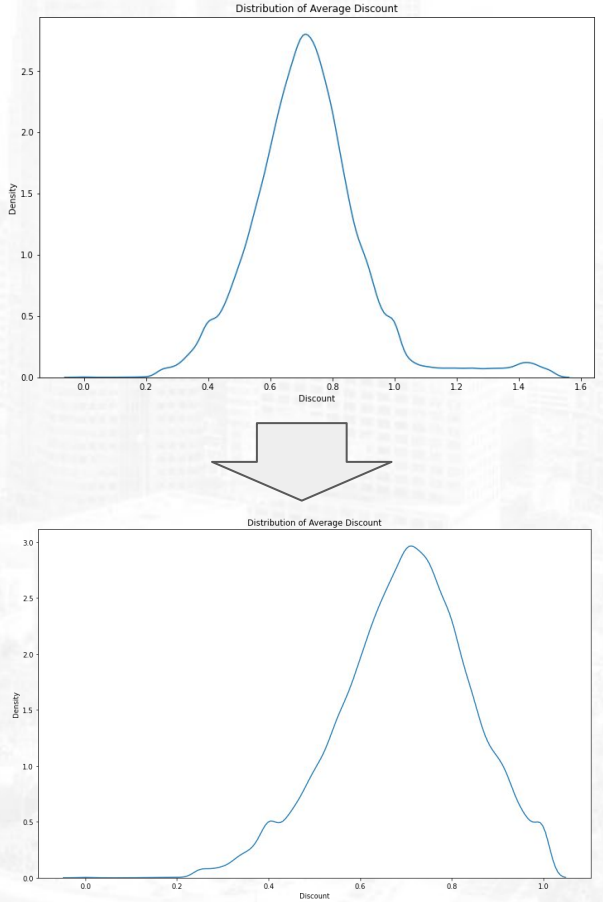
	count	unique	top	freq
ffp_date	62988	3068	1/13/2011	184
first_flight_date	62988	3406	2/16/2013	96
gender	62985	2	Male	48134
work_city	60719	3234	guangzhou	9386
work_province	59740	1165	guangdong	17509
work_country	62962	118	CN	57748
load_time	62988	1	3/31/2014	62988
last_flight_date	62988	731	3/31/2014	959

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

Distribusi kolom numerik.



Insight:



data yang avg_discount > 1 akan di-drop.

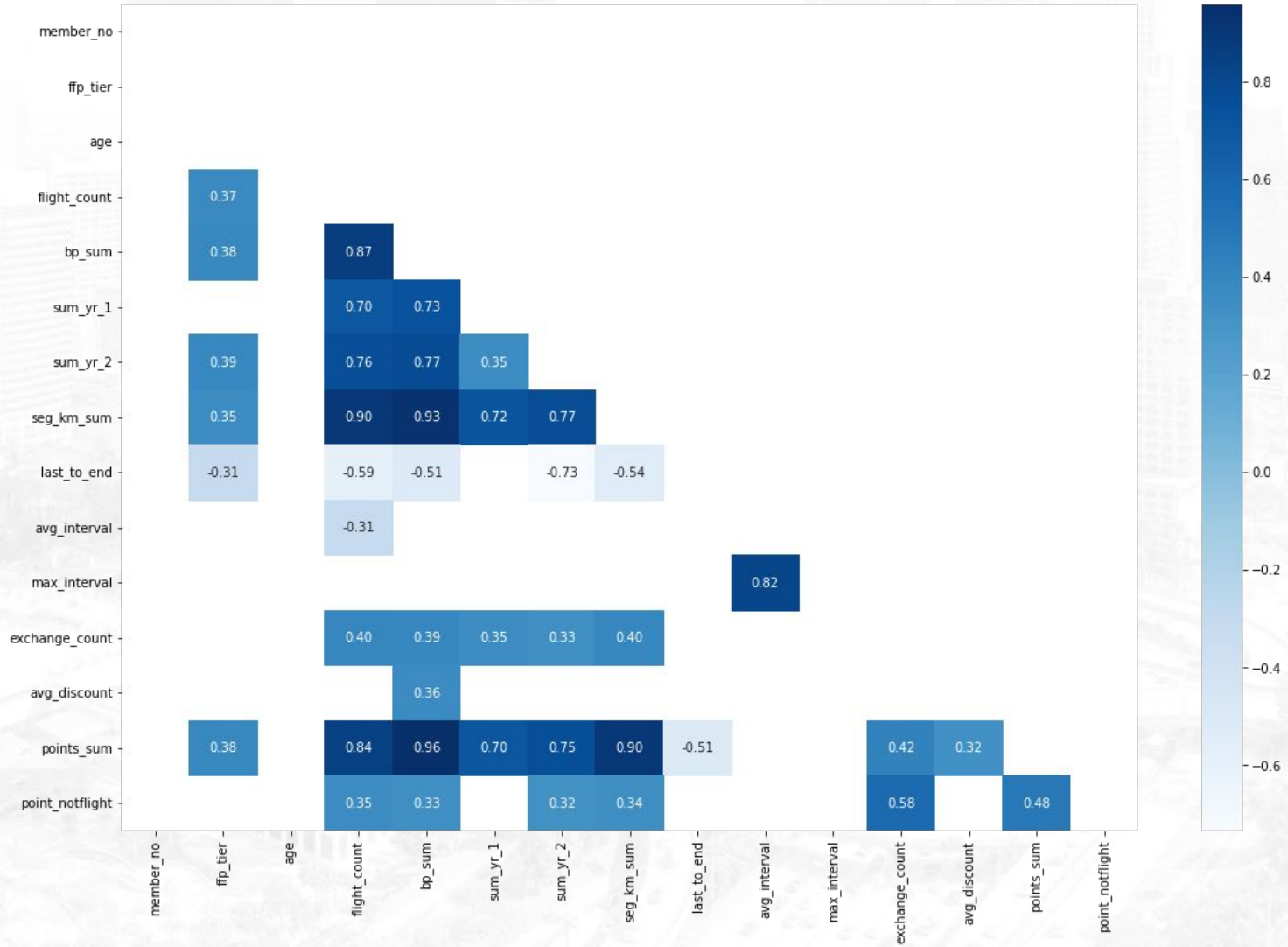
Unique value kategorikal:

```
df.describe(exclude=np.number).T
```

	count	unique	top	freq
ffp_date	62988	3068	1/13/2011	184
first_flight_date	62988	3406	2/16/2013	96
gender	62985	2	Male	48134
work_city	60719	3234	guangzhou	9386
work_province	59740	1165	guangdong	17509
work_country	62962	118	CN	57748
load_time	62988	1	3/31/2014	62988
last_flight_date	62988	731	3/31/2014	959

Feature Correlation

- Karena mayoritas kolom numerik memiliki distribusi right-skewed dan terdapat outlier, maka spearman's correlation lebih tepat digunakan dibandingkan pearson's correlation.
- Banyak multicollinear features sehingga dimensionality reduction bisa diterapkan
- Feature-feature dibawah ini dapat didrop karena memiliki tidak memiliki korelasi yang tinggi antar feature.
  - member_no
  - age
  - last_to_end
  - max_interval
  - avg_interval
  - avg_discount



2. Feature Selection

Berdasarkan paper mengenai Airline Customer Value Analysis (source), indikator yang tepat digunakan untuk membuat airline customers cluster adalah

- Length of customer's membership (L)
- Consumption interval (R)
- Consumption frequency (F)
- Total flight miles (M)
- Mean value of cabin discount coefficient (C)

Kelima indikator tersebut merupakan modifikasi model RFM yang sudah common digunakan untuk memahami customers melalui data.

Berdasarkan indikator-indikator tersebut maka feature yang dipilih sebagai berikut

- ffp_date
- load_time
- flight_count, first_flight_date, last_flight_date
- last_to_end
- avg_discount
- seg_km_sum

```
data_for_clustering.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 60041 entries, 0 to 62987
Data columns (total 8 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   load_time          60041 non-null  datetime64[ns]
 1   ffp_date           60041 non-null  datetime64[ns]
 2   last_to_end        60041 non-null  int64
 3   flight_count       60041 non-null  int64
 4   first_flight_date  60041 non-null  datetime64[ns]
 5   last_flight_date   60041 non-null  datetime64[ns]
 6   seg_km_sum         60041 non-null  int64
 7   avg_discount       60041 non-null  float64
dtypes: datetime64[ns](4), float64(1), int64(3)
memory usage: 4.1 MB
```

Feature Engineering

Feature yang sudah dipilih belum bisa digunakan untuk membangun model clustering. Berikut ini feature engineering yang dilakukan:

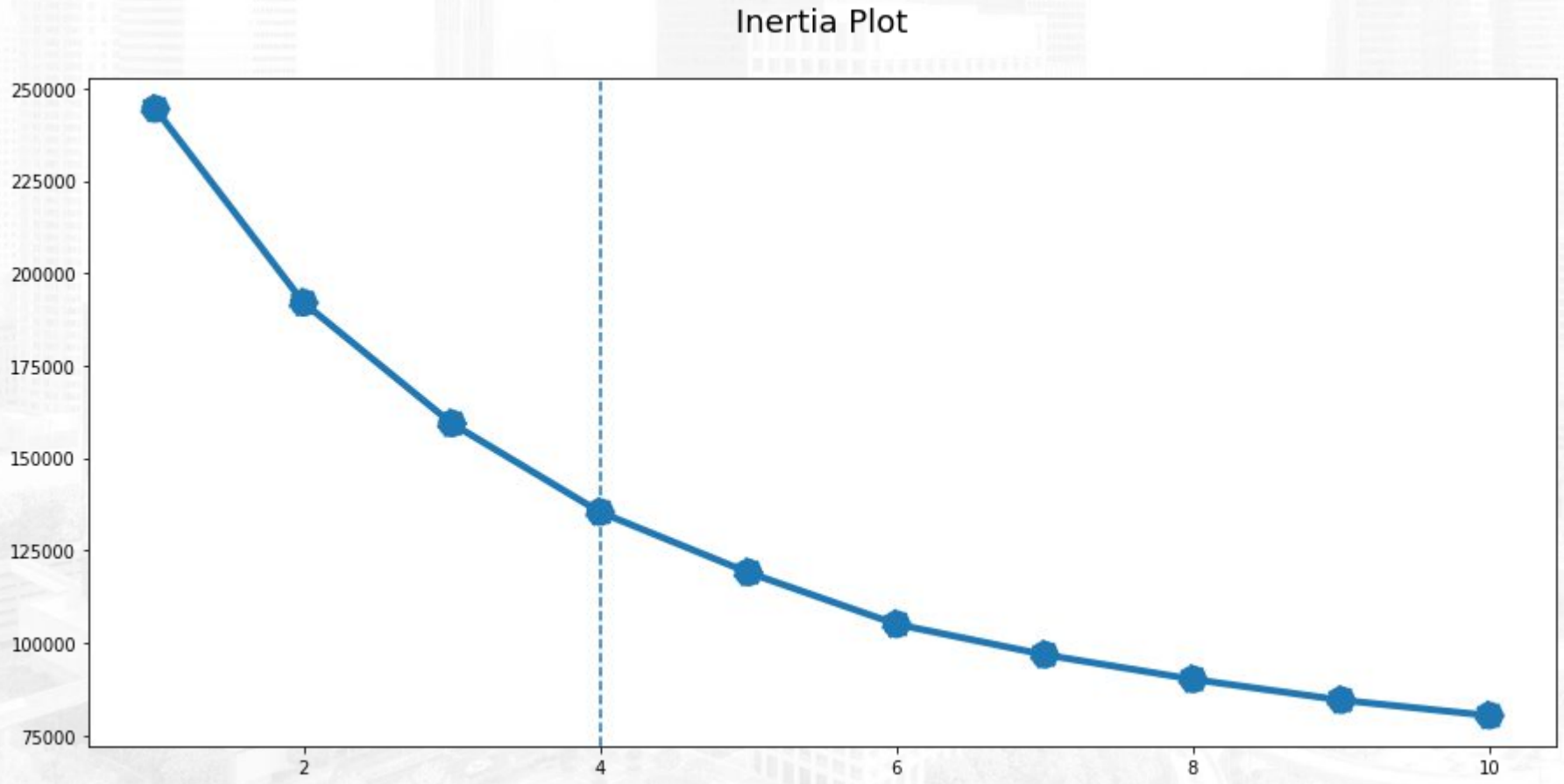
- Feature **Consumption Frequency (F)** diperoleh berdasarkan jumlah penerbangan per tahun mengunakan feature flight_count, last_flight_date dan first_flight_date.
- **Length of customer's membership (L)** diperoleh dengan mengunakan feature load_time dan ffp_date.
- Untuk feature R = last_to_end, M = seg_km_sum, C = avg_discount akan disesuaikan nama kolomnya berdasarkan indikator yang sudah disebutkan di feature selection.
- Setelah feature engineering, drop feature yang tidak akang digunakan kembali, dan handling outlier, dilakukan standarisasi pada dataset sehingga siap untuk modeling.
- Beberapa feature dengan tipe kategori diubah menjadi tipe format tanggal atau datetime.

$$\text{Comsumption Frequency (F)} = \frac{\text{flight\_count}}{\text{last\_flight\_dat} - \text{first\_flight\_date}}$$

Length of customer's membership (L) = load_time – ffp_date

3. Modeling

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



Clustering K-Means

Clustering K-Means dengan 4 cluster.

```
k_means = KMeans(n_clusters=4, random_state=0)
k_means.fit(X_scaled)
```

```
KMeans(n_clusters=4, random_state=0)
```

```
# add the cluster number to the data
data_without_outliers['cluster'] = k_means.labels_
data_without_outliers.sample(5)
```

	L	R	F	M	C	cluster
28185	80.428756	35	0.907435	8627	0.638320	0
40789	82.794308	221	0.338815	3111	0.819855	0
9303	35.516130	23	5.350232	17106	0.881437	3
38157	47.475308	302	1.820750	4152	0.742577	1
8565	24.772583	96	5.533977	21587	0.731138	3

```
data_without_outliers['cluster'].unique()
```

```
array([3, 0, 1, 2], dtype=int32)
```

Evaluasi Cluster

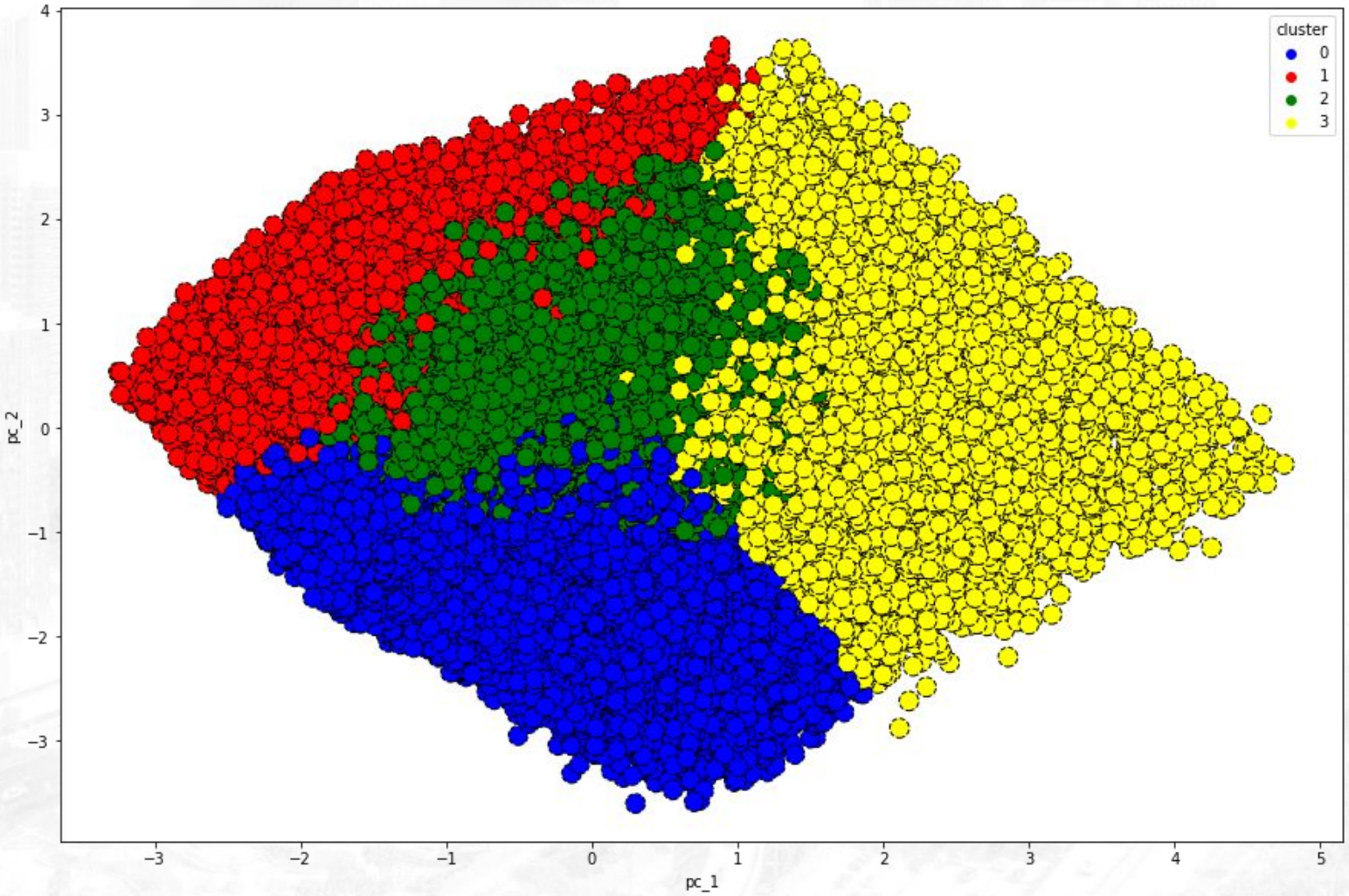
Dengan PCA, kita dapat melihat hasil visualisasi clustering

```
# reduce the dimension to visualize each cluster
from sklearn.decomposition import PCA

pca = PCA(n_components=2, random_state=0)
pca.fit(X_scaled)
pcs = pca.transform(X_scaled)

data_for_visualization = pd.DataFrame(
    pcs,
    columns=['pc_1', 'pc_2']
)

# append to data without outliers
for col in ['pc_1', 'pc_2']:
    data_without_outliers[col] = data_for_visualization[col].copy()

data_without_outliers.sample(3)
```

	L	R	F	M	C	cluster	pc_1	pc_2
48112	61.307214	387	0.701041	977	1.000000	1	-1.942179	0.757024
23984	19.055833	117	5.498274	11731	0.584590	2	0.690070	0.703543
39605	31.507834	93	1.576586	3687	0.753686	2	-0.490918	0.428652

Hasil Cluster PCA

Berikut hasil clustering PCA.



4. Interpretasi Cluster

Interpretasi cluster yang dihasilkan secara bisnis dan berikan rekomendasi yang sesuai dengan cluster yang dihasilkan

Berdasarkan hasil analisis dan modeling, diperoleh beberapa poin penting sebagai berikut.

1. Feature-feature yang berpengaruh dalam segmentasi pelanggan di bidang airline ada lima yaitu
   - jarak penerbangan (km)
   - jarak waktu penerbangan terakhir ke pesanan penerbangan terkini (bulan)
   - rata-rata diskon (persentase)
   - durasi pelanggan menjadi member (month)
   - rata-rata jumlah penerbangan per tahun
2. Berdasarkan hasil modeling, karakteristik setiap cluster dapat dilihat berdasarkan tabel dibawah ini. Pada kolom C (rata-rata diskon) terlihat bahwa diskon yang diperoleh pelanggan tidak terlalu mempengaruhi segmentasi pelanggan karena nilai mean dan median tiap cluster tidak berbeda jauh.

```
display(data_without_outliers.groupby('cluster').agg(['mean','median']).round(2))
```

	L		R		F		M		C		pc_1		pc_2	
	mean	median	mean	median	mean	median	mean	median	mean	median	mean	median	mean	median
cluster														
0	80.70	80.43	86.41	66.0	2.28	1.90	16742.50	14812.0	0.71	0.71	-0.49	-0.57	-1.30	-1.19
1	52.04	48.03	423.74	418.0	2.90	1.70	5888.18	4660.0	0.73	0.74	-1.27	-1.34	1.07	1.04
2	32.64	31.02	113.20	98.0	3.64	3.26	8193.56	7187.5	0.66	0.66	0.01	0.01	0.41	0.39
3	29.31	24.94	104.02	61.0	10.59	9.99	22041.94	21288.0	0.71	0.71	1.88	1.74	0.23	0.19

Deskripsi Masing-Masing Cluster

Berikut ini penjelasan setiap cluster.

- **Cluster 0** merupakan customer yang telah menjadi member cukup lama dengan rata-rata 81 bulan dan jumlah penerbengan per tahun cenderung rendah dengan rata-rata 2 kali per tahun tetapi rata-rata jarak tempuh yang cenderung tinggi sekitar 16 ribu km.
- **Cluster 1** merupakan customer baru dengan durasi rata-rata menjadi member hanya 30 bulan. Namun, customer ini sering melakukan penerbangan dengan rata-rata 11 penerbangan per tahun dan rata-rata jarak tempuh yang jauh sekitar 22 ribu km.
- **Cluster 2** merupakan customer dengan durasi member, frekuensi terbang, dan jarak terbang yang sedang namun memiliki jarak waktu penerbangan akhir ke pesanan penerbangan terkini rendah.
- **Cluster 3** merupakan customer dengan durasi member, frekuensi terbang, dan jarak terbang yang sedang namun memiliki jarak waktu penerbangan akhir ke pesanan penerbangan terkini paling tinggi.

Rekomendasi Strategi Bisnis

Membuat program membership yang berjenjang berdasarkan jumlah penerbangan per tahun dan jarak tempuh penerbangan. Apabila jenjang semakin tinggi maka perusahaan airline tersebut akan memberikan benefit yang semakin tinggi juga misalnya seperti mendapat voucher diskon untuk membeli makanan di pesawat.