

UNSUPERVISED LEARNING

H O M E W O R K - W E E K 1 6

Start Slide

Our Great Team

Theofilus Arifin

Christofer Bryan N. K.

Ramlan Apriyansyah

Muhammad Iqbal

Hanifah Arrasyidah

Christopher Stephen

Muhammad Rizq N. A.

Ujang Pian

Exploratory Data Analysis

Data
Info

Berdasarkan pengamatan yang telah dilakukan, ada beberapa kolom yang bertipe data tidak sesuai.

Descriptive
Analysis

```
<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
```

Data 200

Null

Values

Terdapat beberapa kolom kosong yaitu:

- 1.WORK_CITY = 2269
- 2.WORK_PROVINCE = 3248
- 3.WORK_COUNTRY = 26
- 4.AGE = 420
- 5.SUM_YR_1 = 551
- 6.SUM_YR_2 = 138

-

Descriptive Analysis

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	421
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 200

duplikat Values

Tidak ada Values yang duplicated

Descriptive
Analysis

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

```
0
```

Tipe data object yang memiliki unique value banyak

Tipe data Object yang memiliki unique value banyak :

- FFP_DATE : 3068
- FIRST_FLIGHT_DATE : 3406
- WORK_CITY : 3234
- WORK_PROVINCE : 1165
- WORK_COUNTRY : 118
- LAST_FLIGHT_DATE : 731

```
: df.select_dtypes(include='object').nunique()

: FFP_DATE          3068
  FIRST_FLIGHT_DATE 3406
  GENDER             2
  WORK_CITY          3234
  WORK_PROVINCE      1165
  WORK_COUNTRY       118
  LOAD_TIME          1
  LAST_FLIGHT_DATE   731
dtype: int64
```

Merubah tipe data Object ke Date Time

Merubah tipe data Object ke Date Time :

- FFP_DATE
- FIRST_FLIGHT_DATE
- LAST_FLIGHT_DATE
- LOAD_TIME

```
df['FFP_DATE'] = pd.to_datetime(df['FFP_DATE'], errors='coerce')
df['FIRST_FLIGHT_DATE'] = pd.to_datetime(df['FIRST_FLIGHT_DATE'], errors='coerce')
df['LAST_FLIGHT_DATE'] = pd.to_datetime(df['LAST_FLIGHT_DATE'], errors='coerce')
df['LOAD_TIME'] = pd.to_datetime(df['LOAD_TIME'], errors='coerce')
```


Mengelompokkan data Menjadi Numerikal, kategorikal dan Date Time

Kategori =

["GENDER","WORK_CITY","WORK_PROVINCE","WORK_COUNTRY"]

Numbering =

["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"]

Date Time =

["FFP_DATE","FIRST_FLIGHT_DATE","LAST_FLIGHT_DATE","LOAD_TIME"]

```
df.select_dtypes(include='int64'or'float64').columns.tolist()

['MEMBER_NO',
 'FFP_TIER',
 'FLIGHT_COUNT',
 'BP_SUM',
 'SEG_KM_SUM',
 'LAST_TO_END',
 'MAX_INTERVAL',
 'EXCHANGE_COUNT',
 'Points_Sum',
 'Point_NotFlight']

df.select_dtypes(include='object').columns.tolist()

['GENDER', 'WORK_CITY', 'WORK_PROVINCE', 'WORK_COUNTRY']

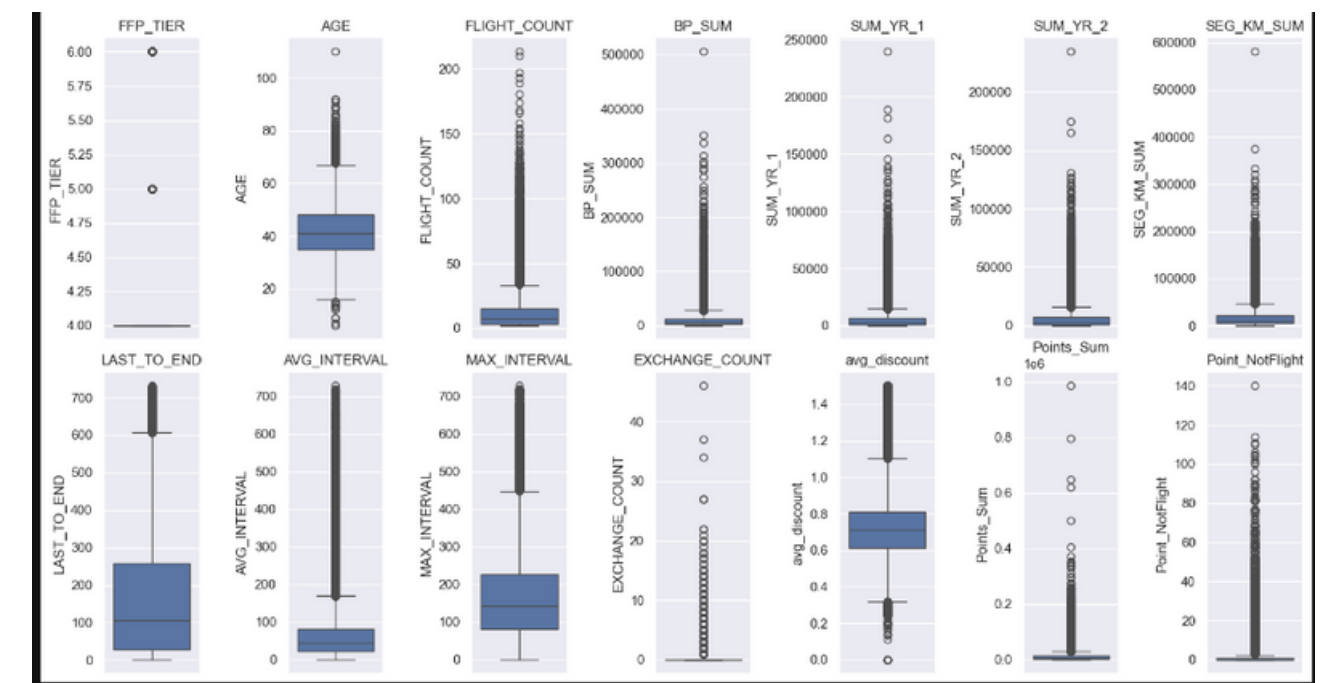
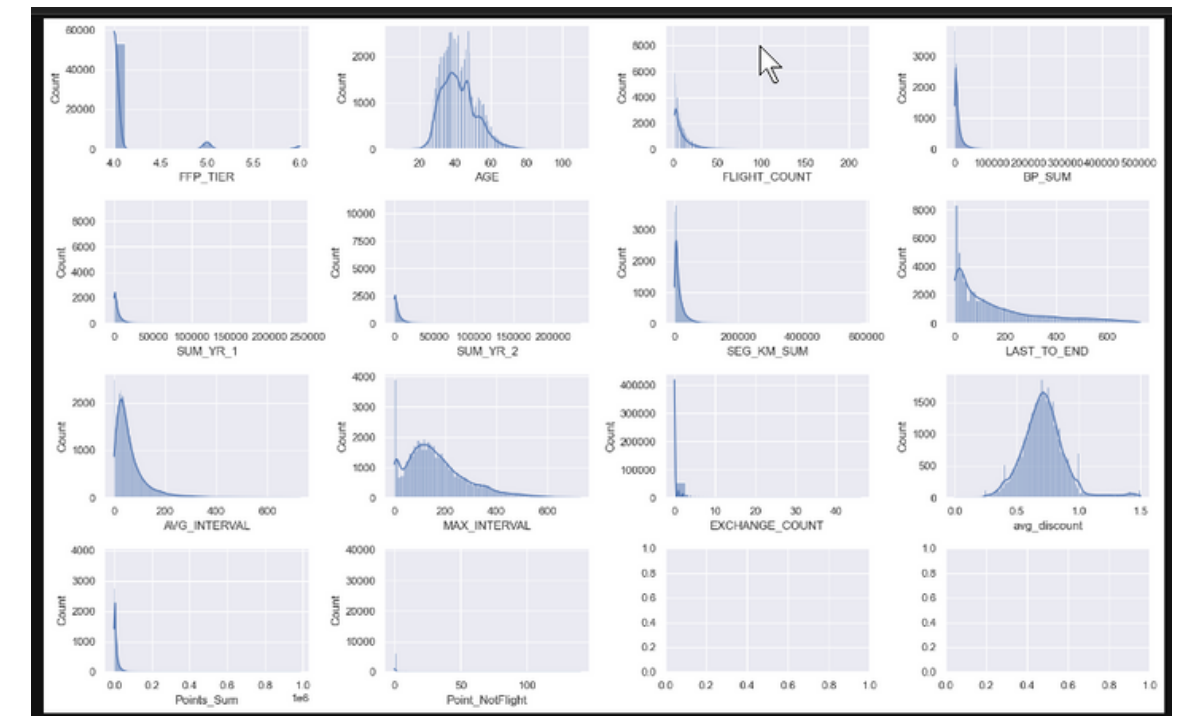
df.select_dtypes(include='datetime64[ns]').columns.tolist()

['FFP_DATE', 'FIRST_FLIGHT_DATE', 'LOAD_TIME', 'LAST_FLIGHT_DATE']
```

Univariate Analysis menggunakan Histogram Plot & Box Plot untuk tipe data Number

Histogram Plot & Box Plot :

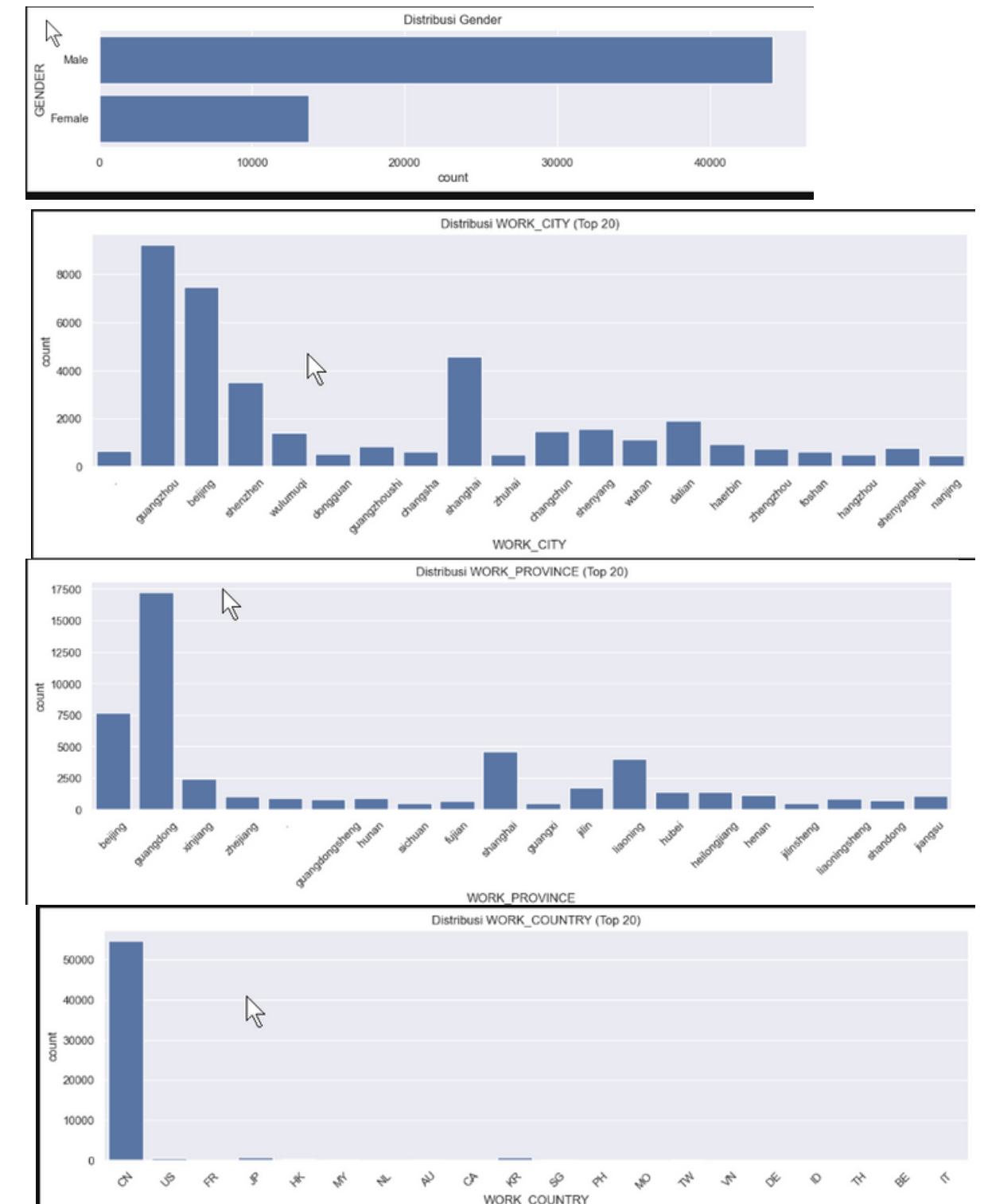
1. Terdapat beberapa kolom dengan persebaran data Positive Skew seperti FLIGHT_COUNT, BP_SUM, SUM_YR_1, SUM_YR_2, SEGMENT_KM_SUM, AVG_INTERVAL dan lain-lain
2. Terdapat beberapa kolom dengan persebaran data Normal seperti AGE dan avg_discount



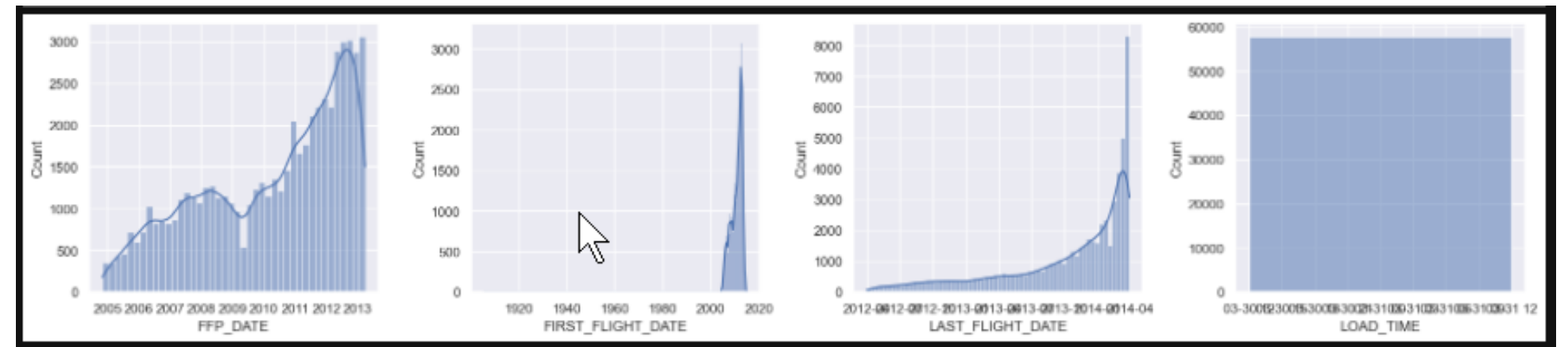
Univariate Analysis menggunakan Bar Plot dan Count Plot untuk tipe data Kategorikal

Bar Plot dan Count Plot :

1. Persebaran data pada Kolom GENDER di dominasi oleh Laki-laki , sekitar 45.000 dari keseluruhan data
- 2.pada kolom WORK_CITY, WORK_COUNTRY, & WORK_PROVINCE diambil top 20 dari keseluruhan data



Univariate Analysis menggunakan Histogram Plot untuk tipe data Date Time



Histogram plot :

1. Dapat terlihat dari grafik diatas bahwa terjadi kenaikan trend setiap tahunnya.
2. terdapat keanehan pada Distribusi kolom LOAD_TIME, karena persebaran data yang sama di setiap row dan menjadi pertimbangan untuk tidak menggunakan feature tersebut

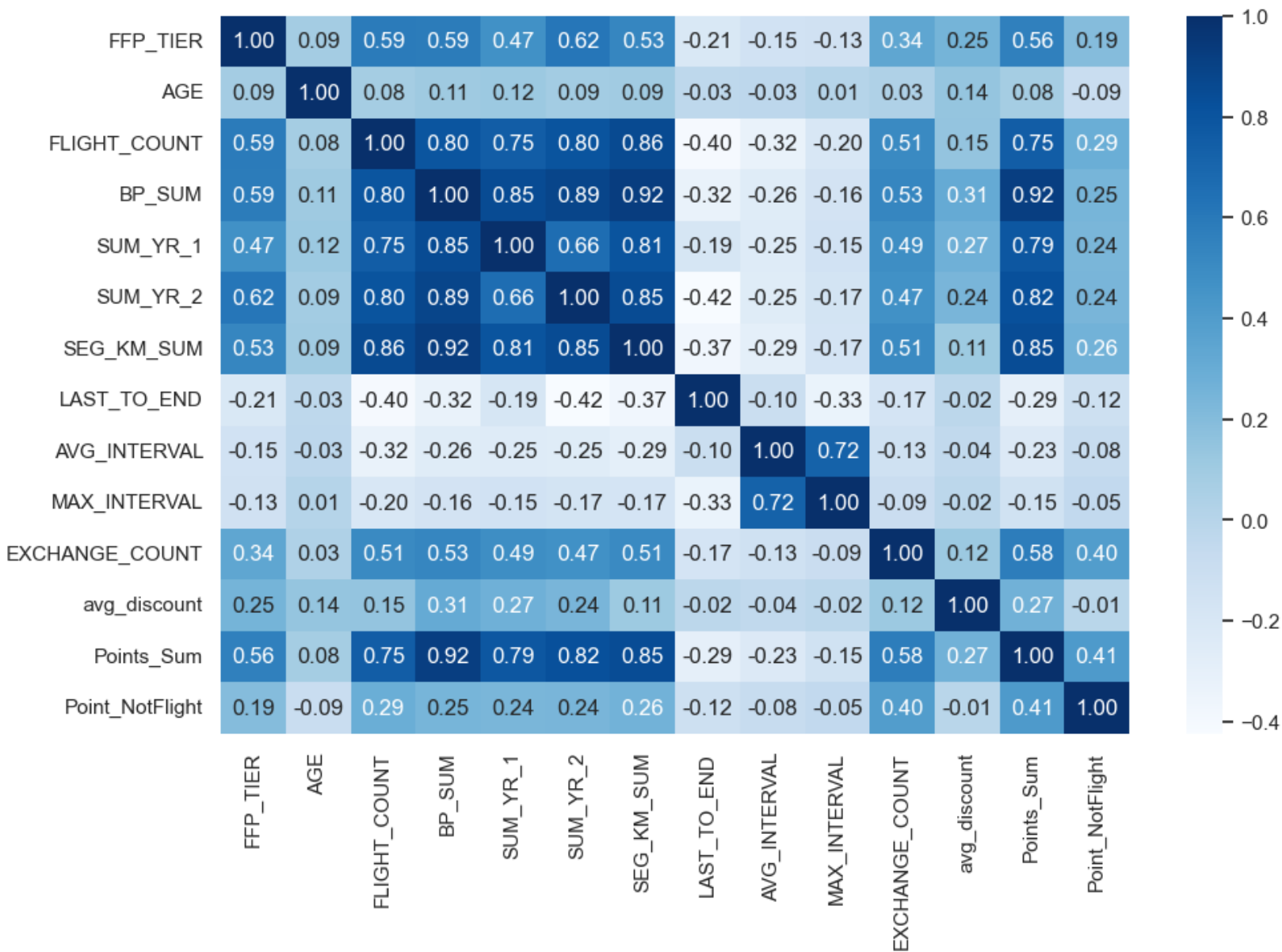
Multivariate Analysis menggunakan Heatmap

Multivariate Analysis

Heatmaps :

Terdapat beberapa fitur yang memiliki korelasi yang besar, maka untuk fitur-fitur yang memiliki nilai korelasi lebih dari 0.85 untuk bisa langsung dieliminasi salah satunya sehingga didapatkan 1 fitur yang tidak redundant. Diantaranya adalah:

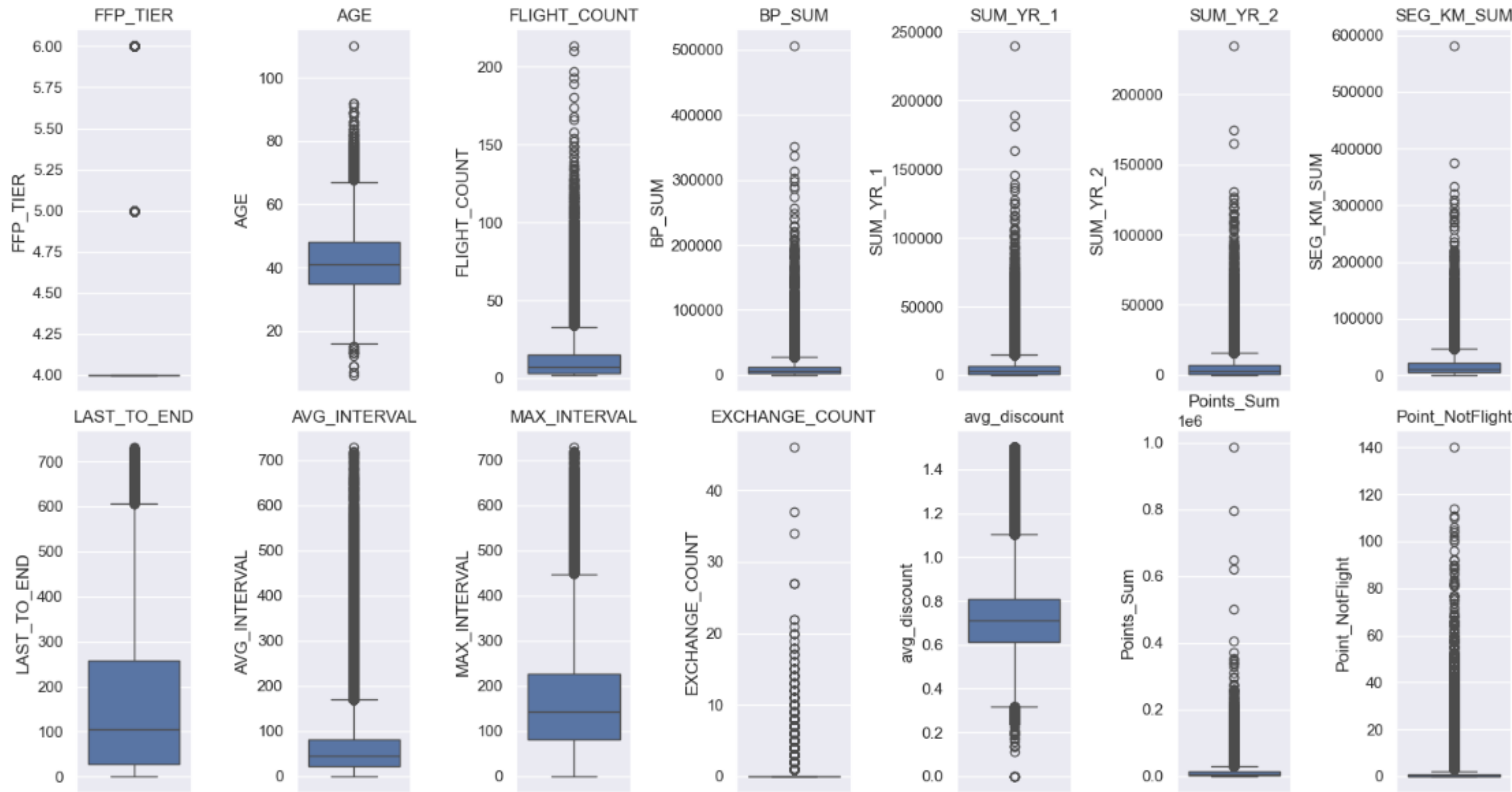
- 1. BP_SUM dan Points_SUM = 0.92
- 2. BP_SUM dan SUM_YR_2 = 0.88
- 3. BP_SUM dan SEG_KM_SUM = 0.92



Feature Extraction

Handling Outliers

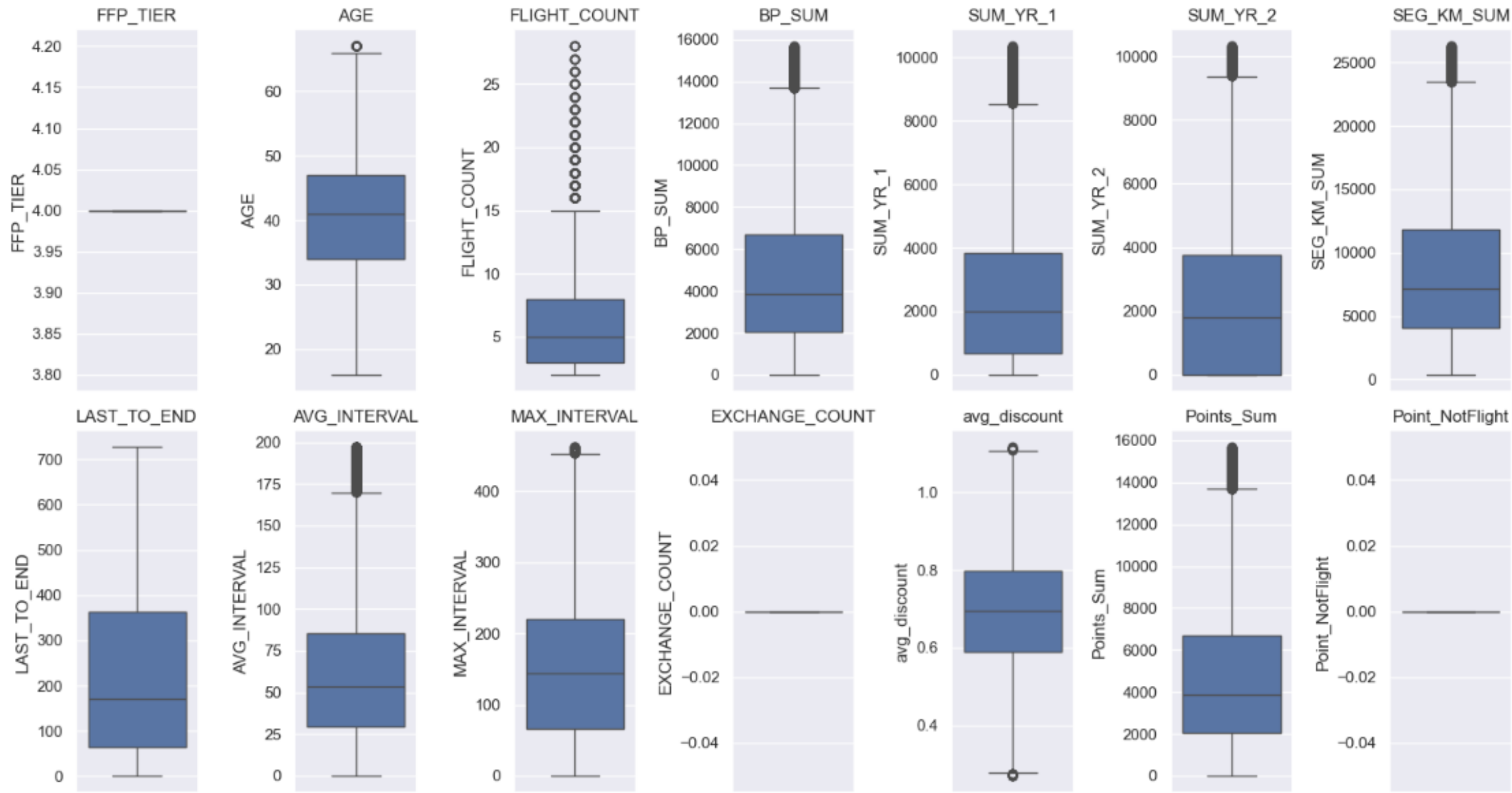
Setelah dilakukan pengecekan outlier pada kolom numerikal dengan menggunakan boxplot didapat bahwa seluruh kolom memiliki outlier,



Handling Outliers

Karena visualisasi outlier dilakukan menggunakan boxplot, maka penghapusan outlier akan menggunakan metode IQR. Berikut adalah hasil dari penghapusan outlier.

After

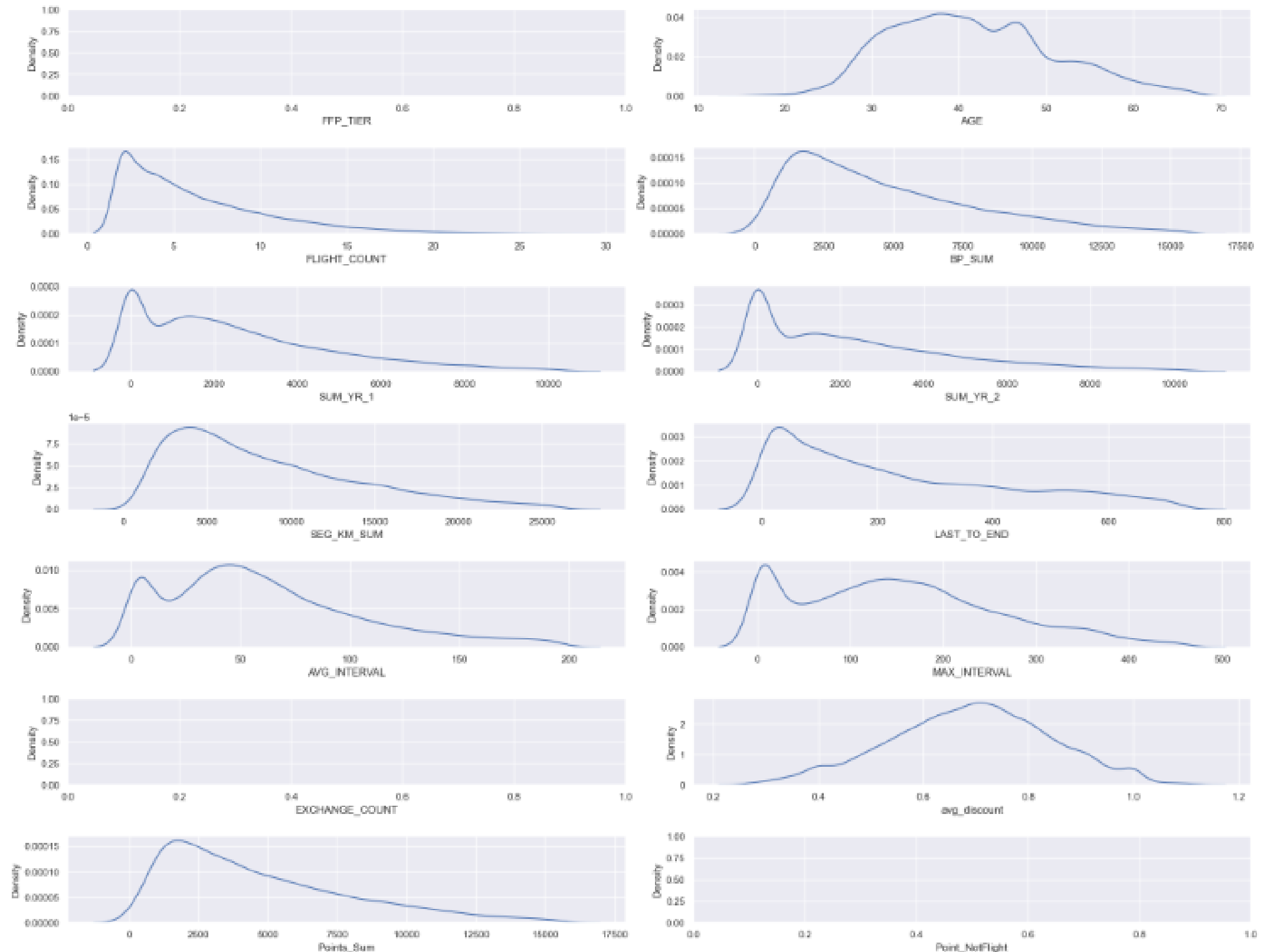


Data 200

Before

Log Transformation

Dapat dilihat bahwa hampir seluruh fitur memiliki skewed distribution dan memiliki skala distribusi yang sangat jauh,

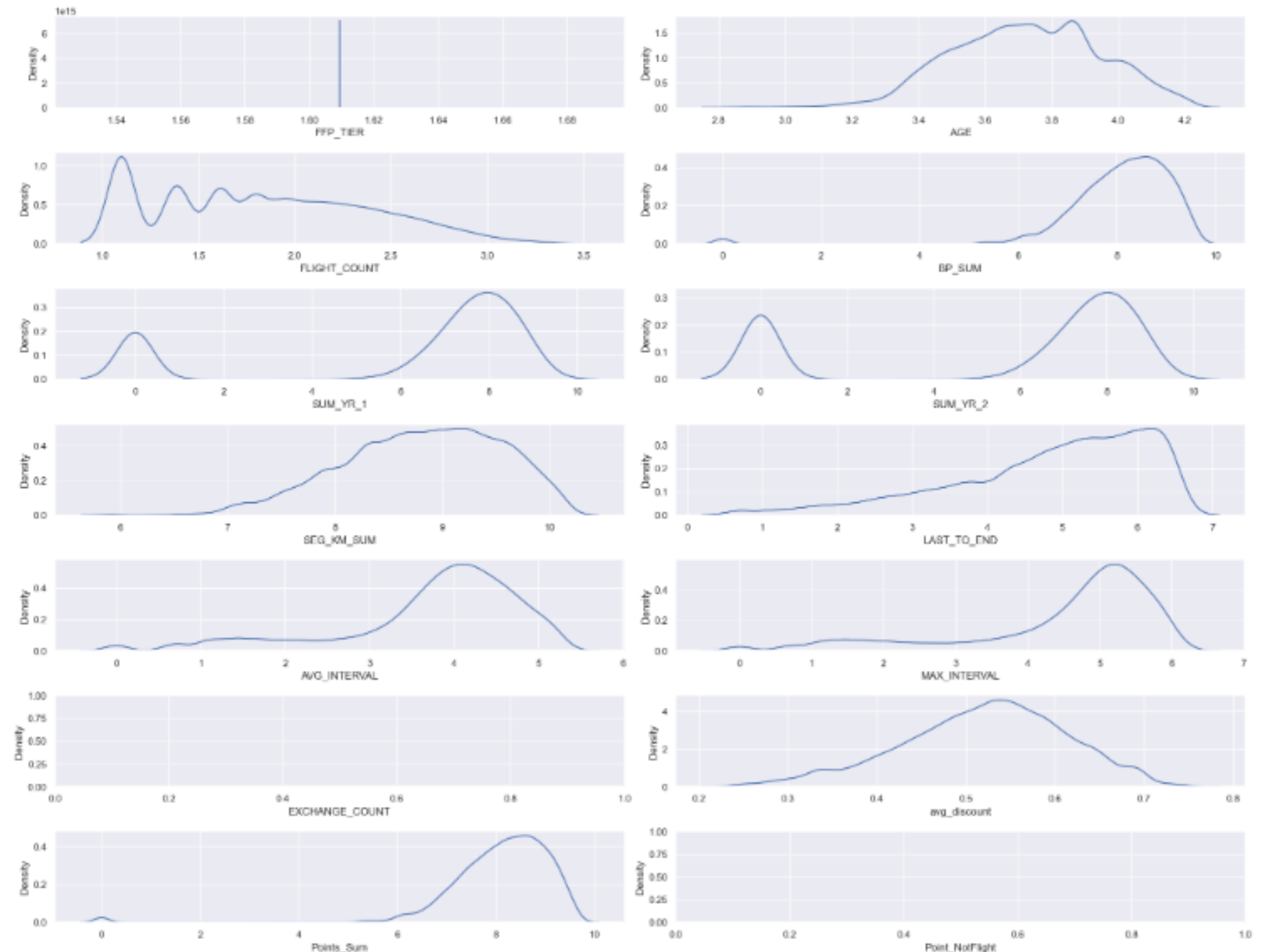


Data 200

After

Log Transformation

Setelah dilakukan log transform, dapat dilihat bahwa hampir seluruh kolom pada kategori numerik sudah memiliki distribusi yang lebih simetrik dan memiliki skala yang sama



Feature Engineering

```
# Membuat mapping dari kategori ke nilai numerik  
gender_mapping = {'Male': 0, 'Female': 1}  
df_filtered['GENDER'] = df_filtered['GENDER'].map(gender_mapping)
```

Feature Encoding

Feature encoding akan dilakukan pada fitur kategorikal yaitu fitur 'Gender' agar fitur dapat direpresentasikan sebagai fitur numerik

Data 200

Monetary Feature

Pada fitur ini kita dapat melihat berapa total pengeluaran customer ketika melakukan penerbangan

Recency Feature

Pada fitur ini kita dapat melihat kapan terakhir kali customer melakukan penerbangan (Dalam jumlah hari)

AVG_DISTANCE Feature

Pada fitur ini kita dapat melihat jarak rata-rata yang ditempuh customer ketika melakukan penerbangan

Feature Engineering

	monetary	count
0	0.000000	136
1	7.244942	31
2	7.139660	29
3	7.215975	29
4	7.560601	28
...
20316	15.674517	1
20317	16.081355	1
20318	15.286521	1
20319	8.953252	1
20320	5.916202	1

Monetary Feature

	recency	count
0	3	206
1	0	198
2	4	191
3	1	189
4	11	185
...
721	403	4
722	716	4
723	687	4
724	724	3
725	454	1

Recency Feature

AVG_DISTANCE
3.667844
3.838365
3.662903
3.797679
3.588507

AVG_DISTANCE Feature

Feature Engineering

```
# Menggunakan pd.qcut untuk membuat bins pada kolom 'AGE'
df_filtered['age_bin'] = pd.qcut(df_filtered['AGE'], q=5, labels=False)

# Menggunakan pd.qcut untuk membuat bins pada kolom 'avg_discount'
df_filtered['avg_discount_level'] = pd.qcut(df_filtered['avg_discount'], q=5, labels=False)

# Membuat bins untuk mengelompokkan jumlah poin
df_filtered['Points_Sum_level'] = pd.qcut(df_filtered['Points_Sum'], q=5, labels=False)

# Membuat bins untuk mengelompokkan jumlah penukaran poin
df_filtered['exchange_count_bin'] = pd.cut(df_filtered['EXCHANGE_COUNT'], bins=4, labels=False)
```

Binning

Pada fitur 'AGE', 'avg_discount', 'Points_Sum', dan 'EXCHANGE_COUNT' akan dilakukan labeling sesuai dengan quantile masing-masing fitur

Data 200

Insight

Result

Brand

- Terdapat 310 merek unik.
- Mayoritas produk berasal dari merek "SEPHORA COLLECTION".
- Terdapat beberapa merek lainnya tetapi jumlah produk mereka jauh lebih sedikit dibandingkan dengan "SEPHORA COLLECTION".

Numeric Data

Insight

Category

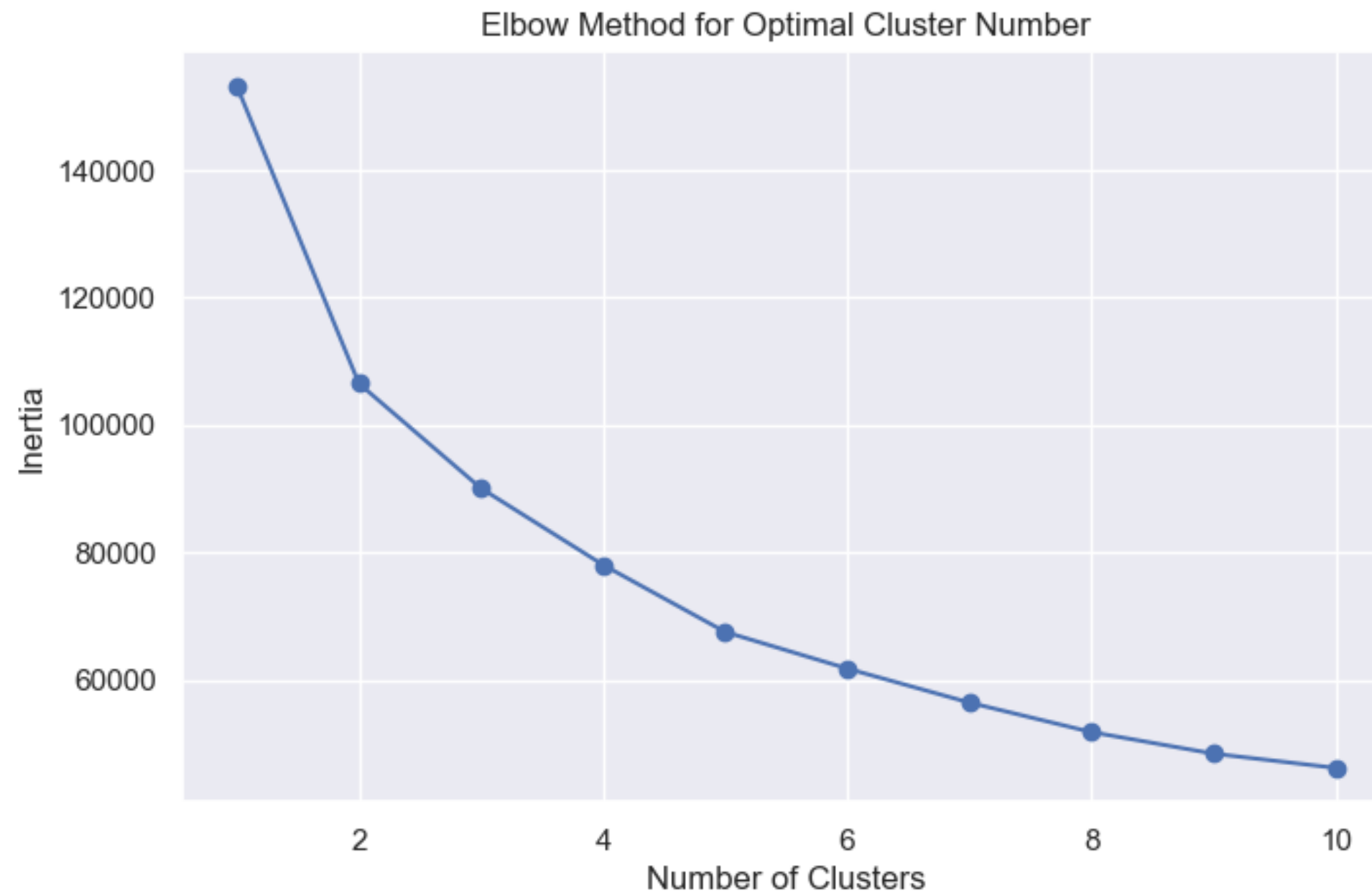
- Terdapat 142 kategori unik
- Mayoritas produk berasal dari merek "Perfumme".
- Terdapat beberapa kategori lainnya tetapi jumlah produk mereka jauh lebih sedikit dibandingkan dengan "Perfumme".

Follow Up

Mengingat sebagian besar produk berasal dari brand dan category tertentu saja mungkin perlu dipertimbangkan untuk mengelompokkan brand dan category jumlah produk yang sangat sedikit ke dalam satu kategori "Lainnya".

Clustering Modeling

Elbow Method



Dari hasil Elbow Method dapat dilihat bahwa:

- **Cluster yang optimalnya dibuat adalah sebanyak 2 cluster**

Clustering menggunakan KMeans

```
df_result = pd.read_csv('../dataset/selected.csv')
df_result['Cluster_Labels'] = labels
df_result
```

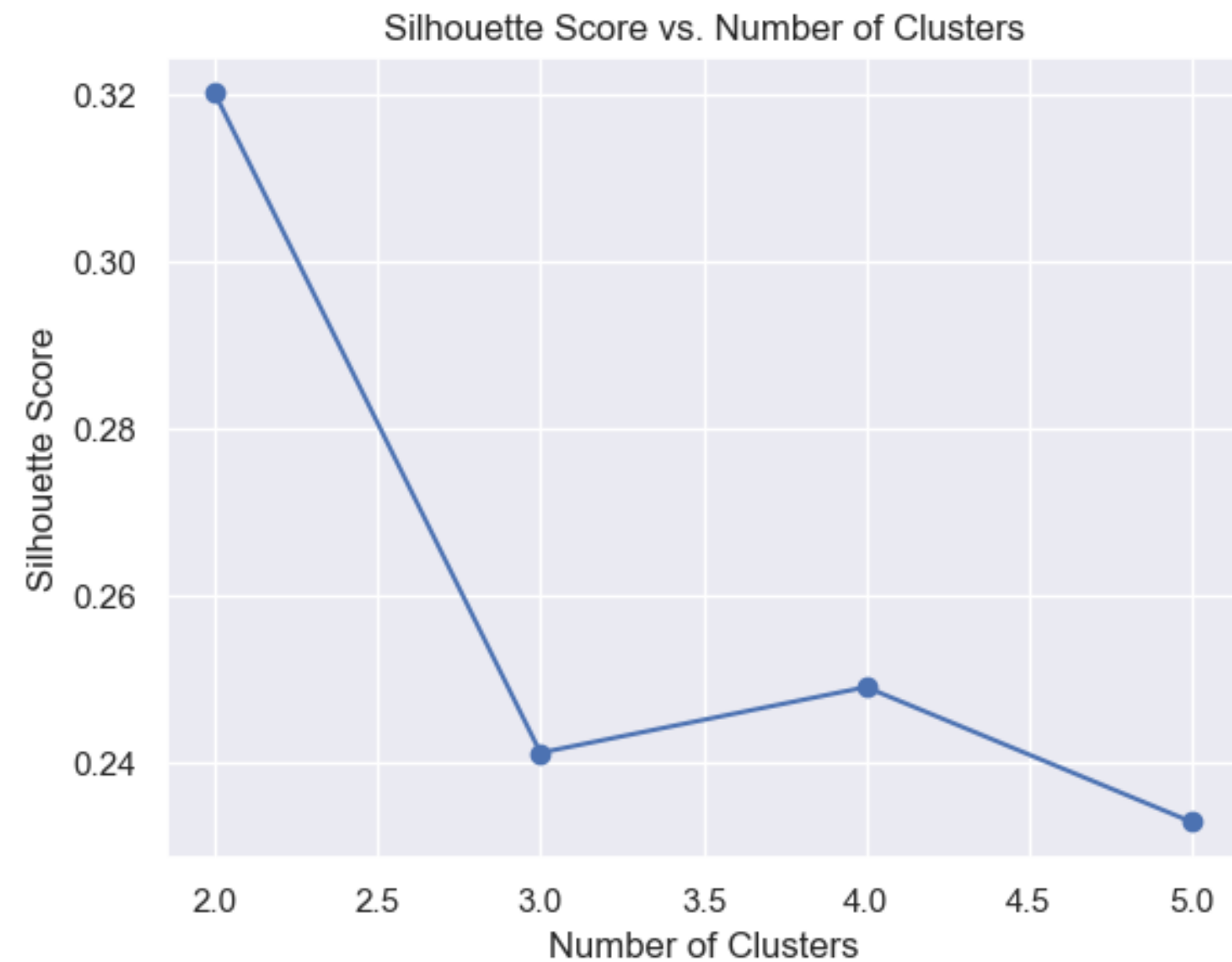
✓ 0.1s

	AGE	FLIGHT_COUNT	recency	monetary	AVG_DISTANCE	Cluster_Labels
0	38.0	13	26	17722.0	1928.846154	0
1	58.0	4	487	7900.0	5930.750000	1
2	43.0	13	85	15875.0	1732.461538	0
3	34.0	14	104	15930.0	1770.357143	0
4	38.0	17	67	16742.0	1359.764706	0
...
30635	40.0	2	414	830.0	184.000000	1
30636	34.0	2	416	830.0	184.000000	1
30637	37.0	2	410	830.0	184.000000	1
30638	38.0	2	119	910.0	184.000000	1
30639	48.0	2	280	370.0	380.000000	1

30640 rows × 6 columns

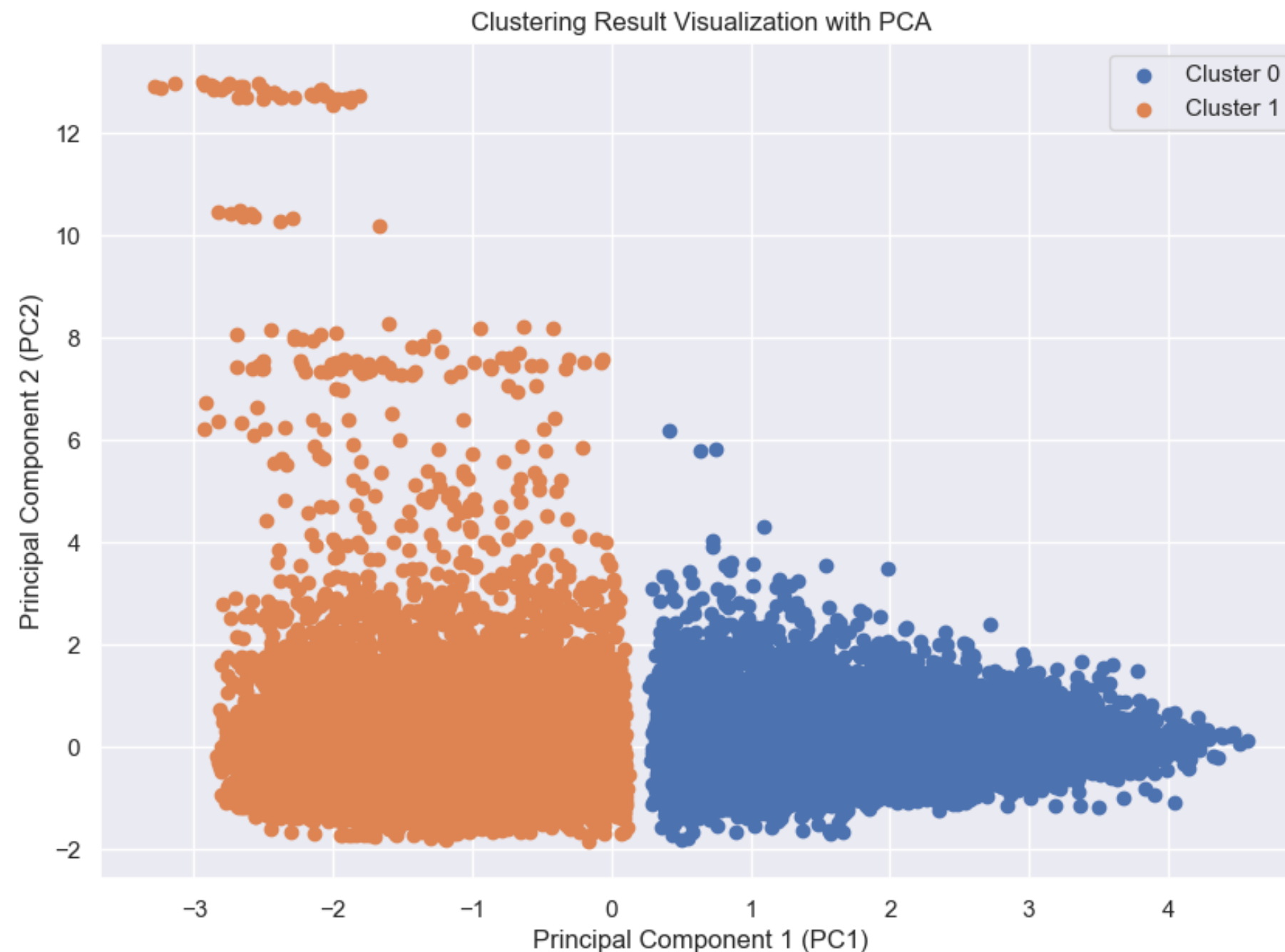
Dari hasil Clustering dapat dilihat bahwa setelah dilakukan clustering menggunakan KMeans data akan terbagi menjadi dua kelompok yang terlihat pada feature Cluster_Labels ada yang 0 dan ada yang 1. Dengan deskripsi yang termasuk di Cluster 0 ada 12965 data, sedangkan Cluster 1 ada 17675 data.

Evaluasi Cluster



Dari hasil Evaluasi Cluster dapat dilihat bahwa jumlah cluster yang optimal mungkin ada di bawah atau di atas angka yang dievaluasi, karena pada titik tersebut silhouette score tertinggi.

Evaluasi Cluster



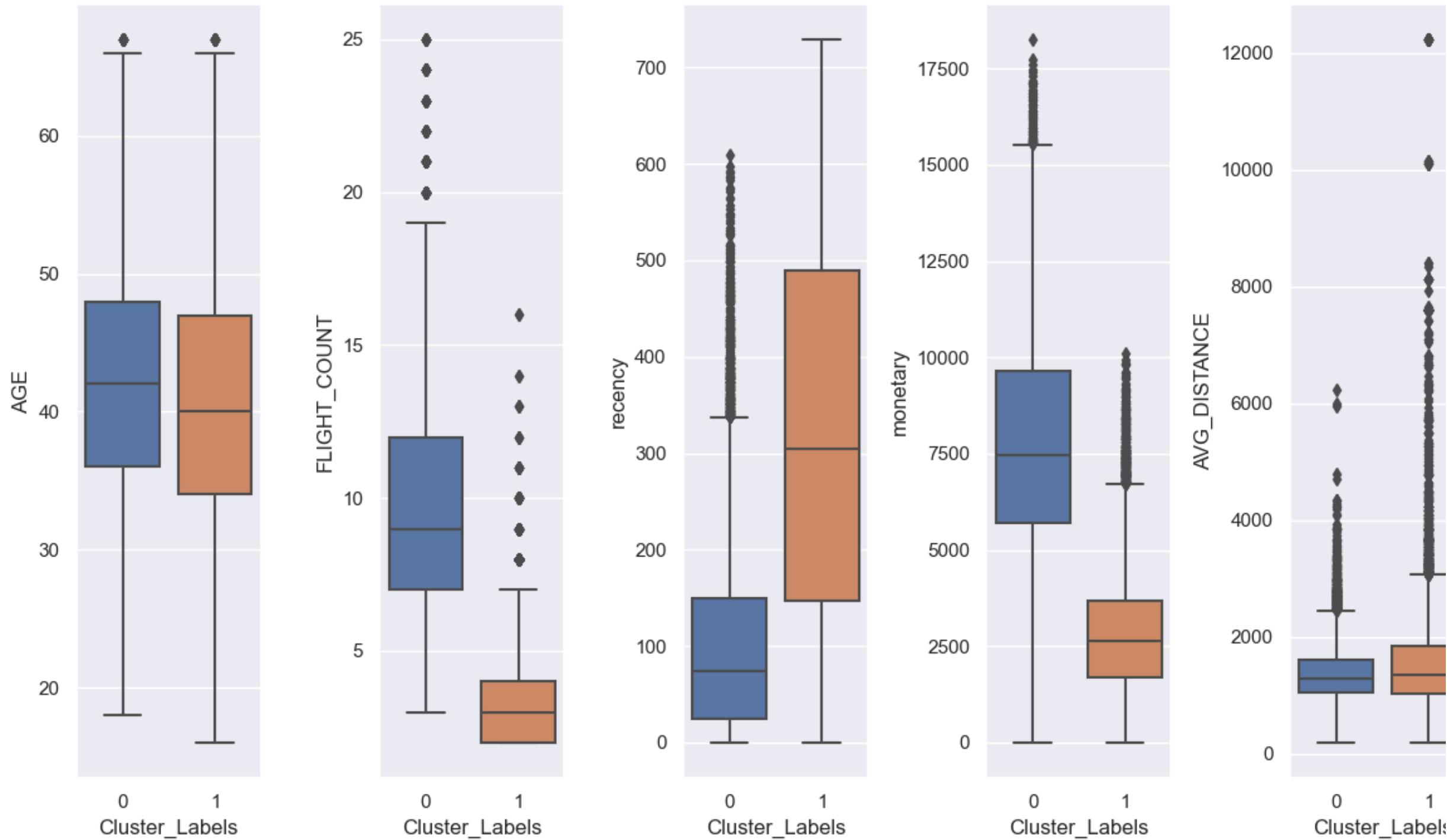
Dari hasil Evaluasi Cluster yang menggunakan PCA dapat dilihat bahwa:

- **Berdasarkan gambar, tiga titik terlihat berada pada satu tingkat kompresi dan saling berdekatan. Maka dapat dikatakan bahwa kedua komponen ini sangat penting dalam memahami hubungan antar data dalam dataset.**
- **Gambar ini juga menunjukkan bahwa cluster tersebut sangat mungkin terjadi pergerakan dari satu cluster ke cluster lainnya. Hal ini dapat terlihat dari sebaran titik yang menggantung di sepanjang akhir PCA dimensi yang lainnya.**

Business Recommendation

Deskripsi Statistik pada Cluster

	AGE	FLIGHT_COUNT	recency	monetary	AVG_DISTANCE
Cluster_Labels					
0	42.815254	10.043979	102.633199	7819.485873	1360.563768
1	41.040271	3.512908	318.551419	2824.404059	1539.282807



Dari visualisasi di atas dapat disimpulkan sebagai berikut:

1. Penggemar traveling (cluster 0)

Orang-orang pada cluster 0 ini memiliki karakteristik sebagai berikut:

- Lebih sering melakukan penerbangan dengan rata-rata jumlah penerbangan sebanyak 10 penerbangan.
- Terakhir kali melakukan penerbangan kurang dari 3 bulan yang lalu saat data ini diambil.
- Tingkat spend untuk traveling cukup tinggi, dengan rata-rata sebesar USD 7,819.

2. Bukan penggemar traveling (cluster 1)

Orang-orang pada cluster 1 ini memiliki karakteristik sebagai berikut:

- Lebih jarang melakukan penerbangan dengan rata-rata jumlah penerbangan sebanyak 3 penerbangan.
- Terakhir kali melakukan penerbangan lebih dari 10 bulan yang lalu saat data ini diambil.
- Tingkat spend untuk traveling rendah, dengan rata-rata sebesar USD 2,824.

Rekomendasi Bisnis:

1. Memberikan promo khusus kepada customers di cluster 1 dengan tujuan meningkatkan intensitas penerbangan yang dilakukan.
2. Memberikan kartu VIP kepada customers di cluster 0 dengan tujuan memberikan eksklusivitas agar mereka tetap mempertahankan loyalitas kepada perusahaan.

THANK YOU

HOMWORK -
UNSUPERVISED LEARNING