Unsupervised Learning

KELOMPOK 2

Elvis Muh. Rizqy
Fuji Resti M
Ni Kadek Yulia Cyntia Dewi
Haolia
Luthfi Adnan Rahmantyo

CHECKING DATA

```
 #   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)
```

```
MEMBER_NO           0.000000
FFP_DATE            0.000000
FIRST_FLIGHT_DATE   0.000000
GENDER              0.004763
FFP_TIER            0.000000
WORK_CITY           3.602273
WORK_PROVINCE       5.156538
WORK_COUNTRY        0.041278
AGE                 0.666794
LOAD_TIME           0.000000
FLIGHT_COUNT        0.000000
BP_SUM              0.000000
SUM_YR_1            0.874770
SUM_YR_2            0.219089
SEG_KM_SUM          0.000000
LAST_FLIGHT_DATE    0.000000
LAST_TO_END         0.000000
AVG_INTERVAL        0.000000
MAX_INTERVAL        0.000000
EXCHANGE_COUNT      0.000000
avg_discount        0.000000
Points_Sum          0.000000
Point_NotFlight     0.000000
```

Cek Duplikat

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

```
0
```

Melalui fungsi df.info() dan persentase missing values diketahui bahwa:

- Kolom AGE seharusnya int bukan float
- Tipe data Kolom FFP_DATE, FIRST_FLIGHT_DATE, LOAD_TIME, LAST_FLIGHT_DATE seharusnya datetime
- Terdapat missing values pada kolom GENDER, WORK_CITY, WORK_PROVINCE, WORK_COUNTRY, AGE, SUM_YR_1,SUM_YR_2
- Persentase missing values tertinggi pada WORK_PROVINCE sebesar 5.156% dan WORK_CITY sebesar 3.602% sehingga perlu dilakukan imputasi
- Presentase missing values pada kolom GENDER, WORK_COUNTRY, AGE, SUM_YR_1, SUM_YR_2 dibawah 1% sehingga bisa langsung di drop
- Tidak terdapat data duplikat pada dataset flight.csv

DESCRIPTIVE STATISTIC

	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

Melalui fungsi df.describe() pada kolom numerikal diketahui bahwa:

- Terdapat anomali (harga tarif bernilai 0 pada kolom SUM_YR_1 (Fare revenue) dan SUM_YR_2 (votes prices), serta avg-discount yang bernilai 0) pada minimal values beberapa kolom, sehingga perlu dianalisa lebih lanjut.
- Pada kolom AGE terdapat seorang pelanggan yang berusia 110 tahun.Ini bisa di drop karena tidak normal
- Sementara itu, sebagian besar kolom terlihat memiliki distribusi skew positif (Mean > Median), kemungkinan terdapat outlier.

DESCRIPTIVE STATISTIC

	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

Melalui fungsi df.describe() pada kolom kategorikaldiketahui bahwa:

- Mayoritas member adalah Pria dengan frekuensi 48.134
- Mayoritas kolom memiliki banyak unique value, kecuali kolom GENDER dan LOAD_TIME
- Dilihat dari kolom LOAD_TIME (Tanggal data diambil), dataset ini diambil pada tanggal 31-3-2014, bisa digunakan sebagai cut off date

PRE-PROCESSING DATA

Handle Missing Value

- Kolom GENDER, WORK_COUNTRY, AGE, SUM_YR_1, SUM_YR_2 memiliki persentase missing value dibawah 1% sehingga bisa di drop
- Kolom WORK_PROVINCE dan WORK_CITY memiliki persentase missing value diatas 1% sehingga dilakukan imputasi dengan nilai modus karena kolom tersebut bertipe kategorikal.

```
miss_col = ['SUM_YR_1', 'AGE', 'SUM_YR_2', 'WORK_COUNTRY', 'GENDER']
for col in miss_col:
  df_clean.dropna(subset=[col], inplace=True)
```

```
df_clean['WORK_CITY'].fillna(df_clean['WORK_CITY'].mode()[0], inplace=True)
df_clean['WORK_PROVINCE'].fillna(df_clean['WORK_PROVINCE'].mode()[0], inplace=True)
```

	feature	missing_value
0	MEMBER_NO	0
1	FFP_DATE	0
2	FIRST_FLIGHT_DATE	0
3	GENDER	0
4	FFP_TIER	0
5	WORK_CITY	0
6	WORK_PROVINCE	0
7	WORK_COUNTRY	0
8	AGE	0
9	LOAD_TIME	0
10	FLIGHT_COUNT	0
11	BP_SUM	0
12	SUM_YR_1	0
13	SUM_YR_2	0
14	SEG_KM_SUM	0
15	LAST_FLIGHT_DATE	0
16	LAST_TO_END	0
17	AVG_INTERVAL	0
18	MAX_INTERVAL	0
19	EXCHANGE_COUNT	0
20	avg_discount	0
21	Points_Sum	0
22	Point_NotFlight	0

Handle Incorrect Value

- Pada kolom LAST_FLIGHT_DATE terdapat tanggal 2014/2/29 yang tidak sesuai dengan tanggal pada umumnya.
- Presentase data yang memiliki tanggal 2014/2/29 pada kolom LAST_FLIGHT_DATE adalah 0.007% sehingga bisa langsung di drop

```
print('Incorrect LAST_FLIGHT_DATE data percentage: ', end='')
print(str(round(df_clean[df_clean.LAST_FLIGHT_DATE.str.contains('2014/2/29')]['LAST_FLIGHT_DATE'].count()/len(df_clean), 3)), '%')
```

```
df_clean.drop(df_clean[df_clean.LAST_FLIGHT_DATE.str.contains('2014/2/29')].index, inplace = True)
```

- Pada kolom AGE terdapat usia 110 tahun dimana data tersebut harus dihapus karena tidak nomral

```
df_clean.drop(df_clean[df_clean.AGE > 100].index, inplace = True)
```

- Terdapat keanehan pada nilai minimum kolom SUM_YR_1, SUM_YR_2, avg_discount yang bernilai 0 sementara kolom SEG_KM_SUM memiliki nilai yang sangat tinggi, sehingga data dengan keanehan terssebut perlu di drop

```
df_clean.drop(df_clean[(df_clean.SUM_YR_1 == 0) & (df_clean.SUM_YR_2 == 0) & (df_clean.avg_discount == 0) & (df_clean.SEG_KM_SUM > 0)].index, inplace = True)
```

Handle Data Type

- Merubah tipe data kolom AGE dari float menjadi int
- Merubah tipe data kolom FFP_DATE, FIRST_FLIGHT_DATE, LOAD_TIME, LAST_FLIGHT_DATE dari string object menjadi datetime

```
df_clean['AGE'] = df_clean['AGE'].astype(int)
```

```
datee = ['FFP_DATE', 'FIRST_FLIGHT_DATE', 'LOAD_TIME','LAST_FLIGHT_DATE']
for col in datee:
  df_clean[col] = pd.to_datetime(df_clean[col])
```

Feature Engineering

```
df_clean['DURATION'] = ((df_clean['LOAD_TIME'] - df_clean['FFP_DATE'])/np.timedelta64(1, 'M'))
df_clean['DURATION'] = df_clean['DURATION'].astype(int)
df_clean['DURATION'] = pd.to_numeric(df_clean['DURATION'], errors="coerce").fillna(0).astype('int64')
```

- Dari feature LOAD_TIME (tanggal data diambil) dan FFP_DATE (Frequent Flyer Program Join Date) dapat diperoleh feature baru yang menunjukkan durasi atau sudah berapa lama seseorang menjadi member

Redefine Dataset

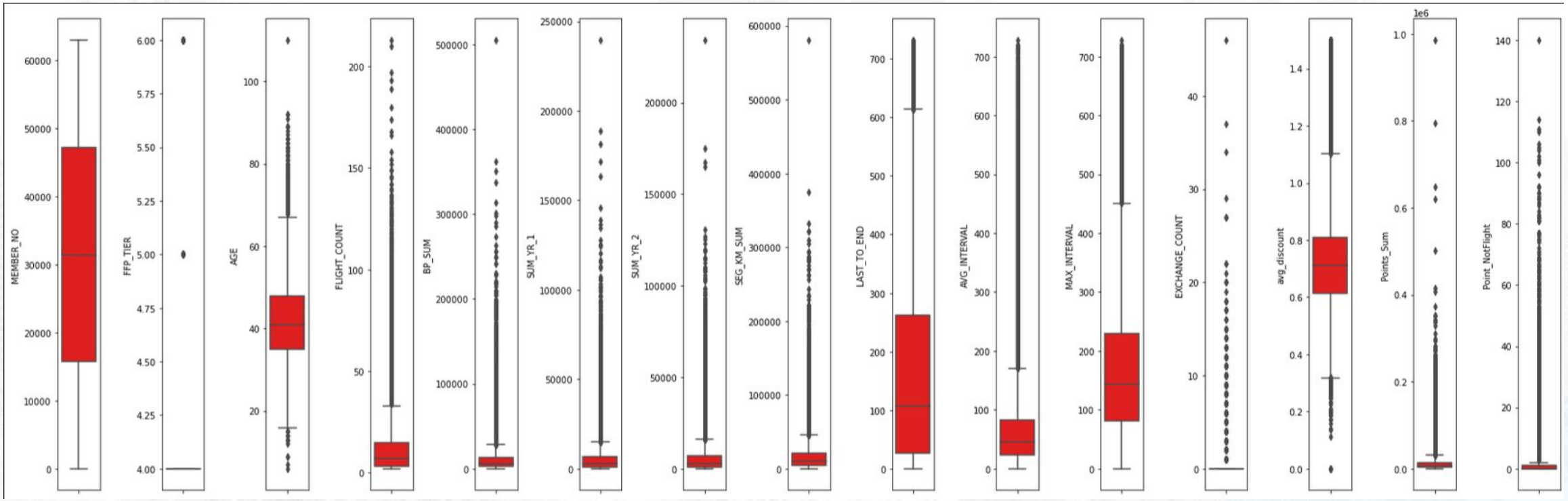
- Setelah Pre-Processing dataset akan dikelompokkan ulang berdasarkan tipe data numerikal, kategorikal, dan datetime

```
numm = [key for key in dict(df_clean.dtypes)
                if dict(df_clean.dtypes)[key]
                    in ['float64','float32','int32','int64']] # Numeric Variable

catt = [key for key in dict(df_clean.dtypes)
            if dict(df_clean.dtypes)[key]
               in ['object'] ] # Categorical Varible

datee = [key for key in dict(df_clean.dtypes)
            if dict(df_clean.dtypes)[key]
               in ['datetime64[ns]'] ] # Categorical Varible
```

```
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   MEMBER_NO          61437 non-null  int64
 1   FFP_DATE           61437 non-null  datetime64[ns]
 2   FIRST_FLIGHT_DATE  61437 non-null  datetime64[ns]
 3   GENDER             61437 non-null  object
 4   FFP_TIER           61437 non-null  int64
 5   WORK_CITY          61437 non-null  object
 6   WORK_PROVINCE      61437 non-null  object
 7   WORK_COUNTRY       61437 non-null  object
 8   AGE                61437 non-null  int64
 9   LOAD_TIME          61437 non-null  datetime64[ns]
 10  FLIGHT_COUNT       61437 non-null  int64
 11  BP_SUM             61437 non-null  int64
 12  SUM_YR_1           61437 non-null  float64
 13  SUM_YR_2           61437 non-null  float64
 14  SEG_KM_SUM         61437 non-null  int64
 15  LAST_FLIGHT_DATE   61437 non-null  datetime64[ns]
 16  LAST_TO_END        61437 non-null  int64
 17  AVG_INTERVAL       61437 non-null  float64
 18  MAX_INTERVAL       61437 non-null  int64
 19  EXCHANGE_COUNT     61437 non-null  int64
 20  avg_discount       61437 non-null  float64
 21  Points_Sum         61437 non-null  int64
 22  Point_NotFlight    61437 non-null  int64
 23  DURATION           61437 non-null  int64
dtypes: datetime64[ns](4), float64(4), int64(12), object(4)
```

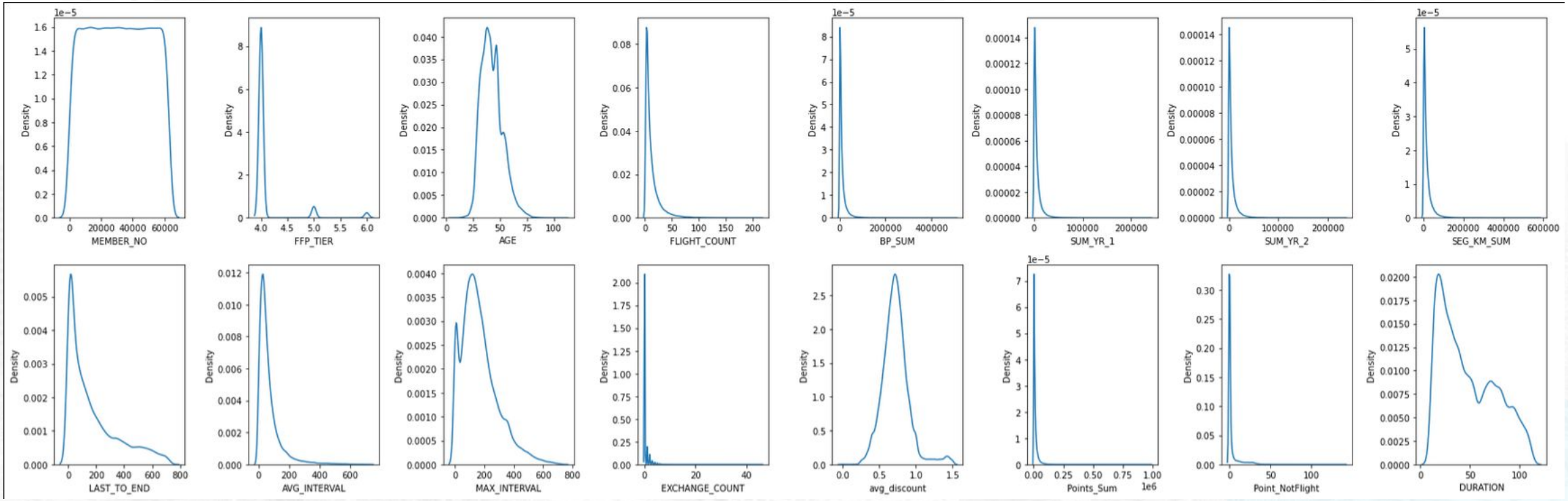
EDA

Univariate Analysis



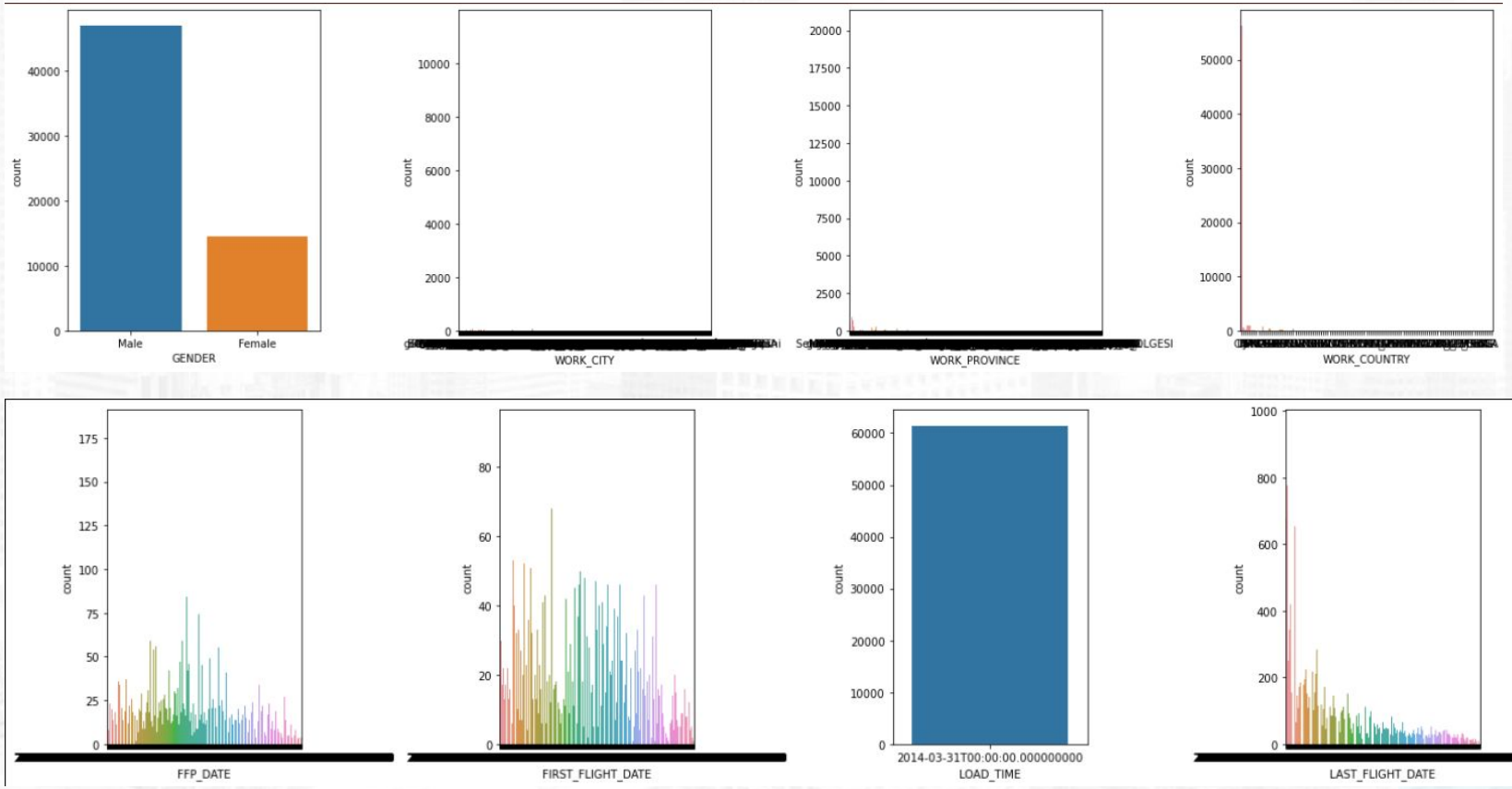
- Hampir semua kolom numerikal punya outliers, kecuali kolom MEMBER_NO, FFP_TIER dan MEMBER_DURATION

Univariate Analysis



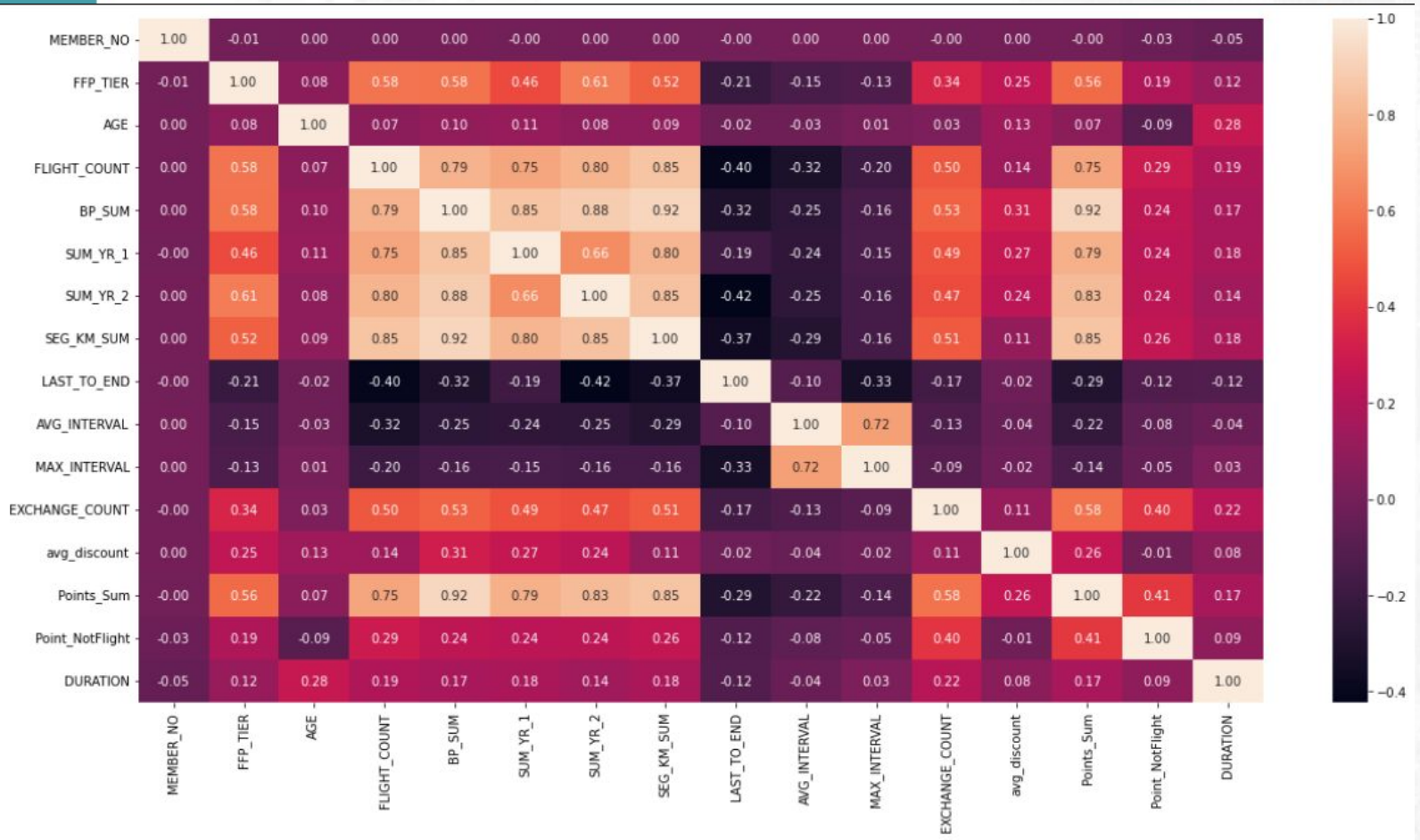
- Hampir semua kolom numerikal menghasilkan distribusi positif kecuali kolom MEMBER_NO, FFP_TIER dan avg_discount

Univariate Analysis



- Hampir semua kolom kategorikal dan datetime memiliki unique value yang tinggi kecuali GENDER dan LOAD_TIME
- Kolom LOAD_TIME hanya memiliki 1 value yaitu 2014-03-31
- Member terbanyak adalah MALE

Multivariate Analysis



* Korelasi tertinggi 0.92 pada
  - Points_Sum dengan BP_SUM,
  - SEG_KM_SUM dengan BP_SUM
* Korelasi 0.88 pada SUM_YR_2 dengan BP_SUM
* Korelasi 0.86 pada FIRST_FLIGHT_DATE dgn FFP_DATE
* Korelasi 0.85 pada
  - SUM_YR_1 dengan BP_SUM,
  - SUM_YR_2 dengan SEG_KM_SUM,
  - SEG_KM_SUM dengan FLIGHT_COUNT,
  - Points_Sum dengan SEG_KM_SUM ,
* Korelasi 0.83 pada SUM_YR_2 dgn Point_sum
* Korelasi 0.80 pada
  - SUM_YR_2 dengan FLIGHT_COUNT,
  - SUM_YR_1 dengan SEG_KM_SUM
* Korelasi 0.79 pada
  - BP_SUM dgn FLIGHT_COUNT,
  - SUM_YR_1 dgn Point_Sum
* Korelasi 0.75 pada
  - SUM_YR_1 dgn FLIGHT_COUNT,
  - Points_Sum dgn FLIGHT_COUNT
* Korelasi 0.72 pada MAX_INTERVAL dgn AVG_INTERVAL

FEATURE SELECTION

Menggunakan metode segmentasi RFM

- Recency (Kebaruan) -> Kapan terakhir kali member melakukan penerbangan? Dapat diketahui dari kolom LAST_TO_END (jarak waktu pemesanan penerbangan terakhir ke penerbangan terbaru)
- Frequency(Frekuensi) -> Berapa kali member sudah melakukan penerbangan FLIGHT_COUNT
- Moneytary -> Berapa banyak pengeluaran yg sudah dilakukan member tsb disesuaikan menjadi berapa SEG_KM_SUM (TOTAL JARAK(km)) penerbangan yang sudah dilakukan

Feature tambahan

- DURATION -> Menunjukkan LOYALTY dari membership
- AVG_DISCOUNT -> Menunjukkan VALUE CUSTOMER (high atau low customer) ketika memilih kabin

Normalization/Standardization

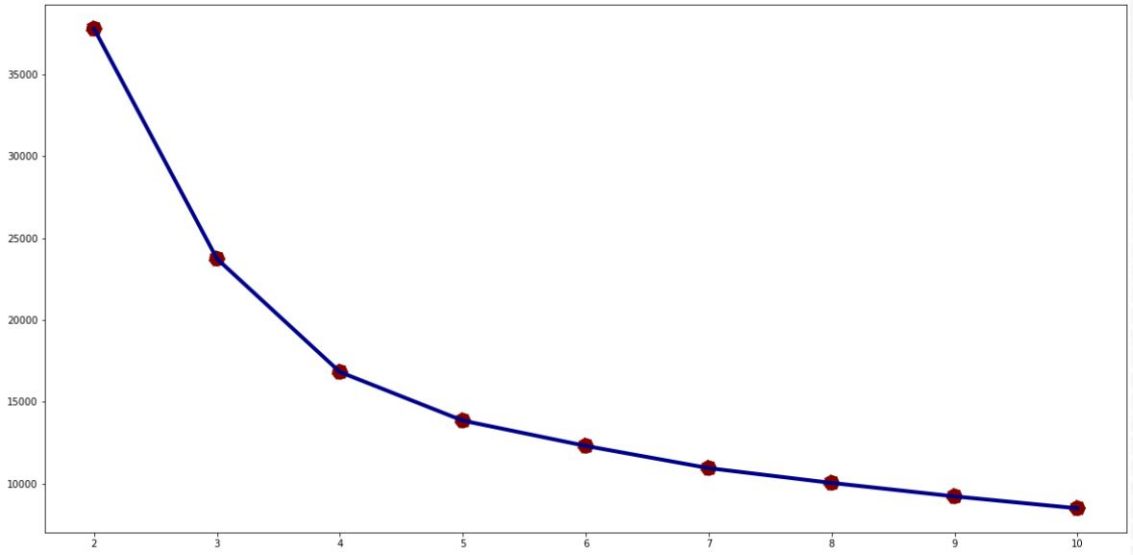
	MEMBER_DURATION	LAST_TO_END	FLIGHT_COUNT	SEG_KM_SUM	avg_discount	MEMBER_DURATION_NORM	LAST_TO_END_NORM	FLIGHT_COUNT_NORM	SEG_KM_SUM_NORM	avg_discount_STD
count	61429.000000	61429.000000	61429.000000	61429.000000	61429.000000	61429.000000	61429.000000	61429.000000	61429.000000	6.142900e+04
mean	48.215761	173.550245	11.940956	17277.610917	0.721726	0.368473	0.236370	0.047114	0.029137	1.890226e-15
std	27.818950	181.473954	14.119509	21055.460423	0.184697	0.275435	0.248594	0.066917	0.036281	1.000008e+00
min	11.000000	1.000000	2.000000	368.000000	0.112043	0.000000	0.000000	0.000000	0.000000	-3.301023e+00
25%	24.000000	28.000000	3.000000	4850.000000	0.612569	0.128713	0.036986	0.004739	0.007723	-5.910148e-01
50%	41.000000	107.000000	7.000000	10150.000000	0.711807	0.297030	0.145205	0.023697	0.016855	-5.370799e-02
75%	71.000000	262.000000	15.000000	21433.000000	0.808997	0.594059	0.357534	0.061611	0.036297	4.725116e-01
max	112.000000	731.000000	213.000000	580717.000000	1.500000	1.000000	1.000000	1.000000	1.000000	4.213829e+00

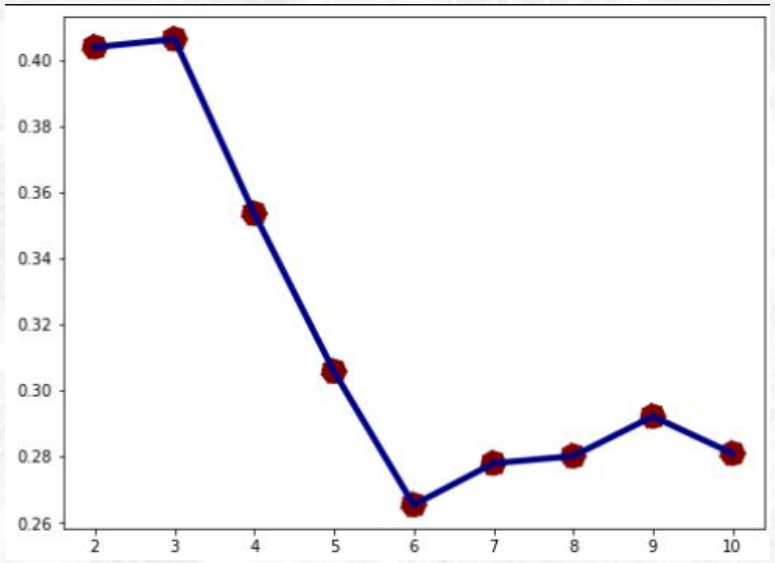
- Normalisasi dilakukan karena kolom MEMBER_DURATION, LAST_TO_END, FLIGHT_COUNT, SEG_KM_SUM memiliki rentang nilai yang berbeda jauh dengan avg_discount
- Standardisasi pada kolom avg_discont digunakan untuk menskalakan ulang data sehingga memberikan grafik normal

Rakamin Academy

K-MEANS CLUSTERING

Inertia & Silhouette Score

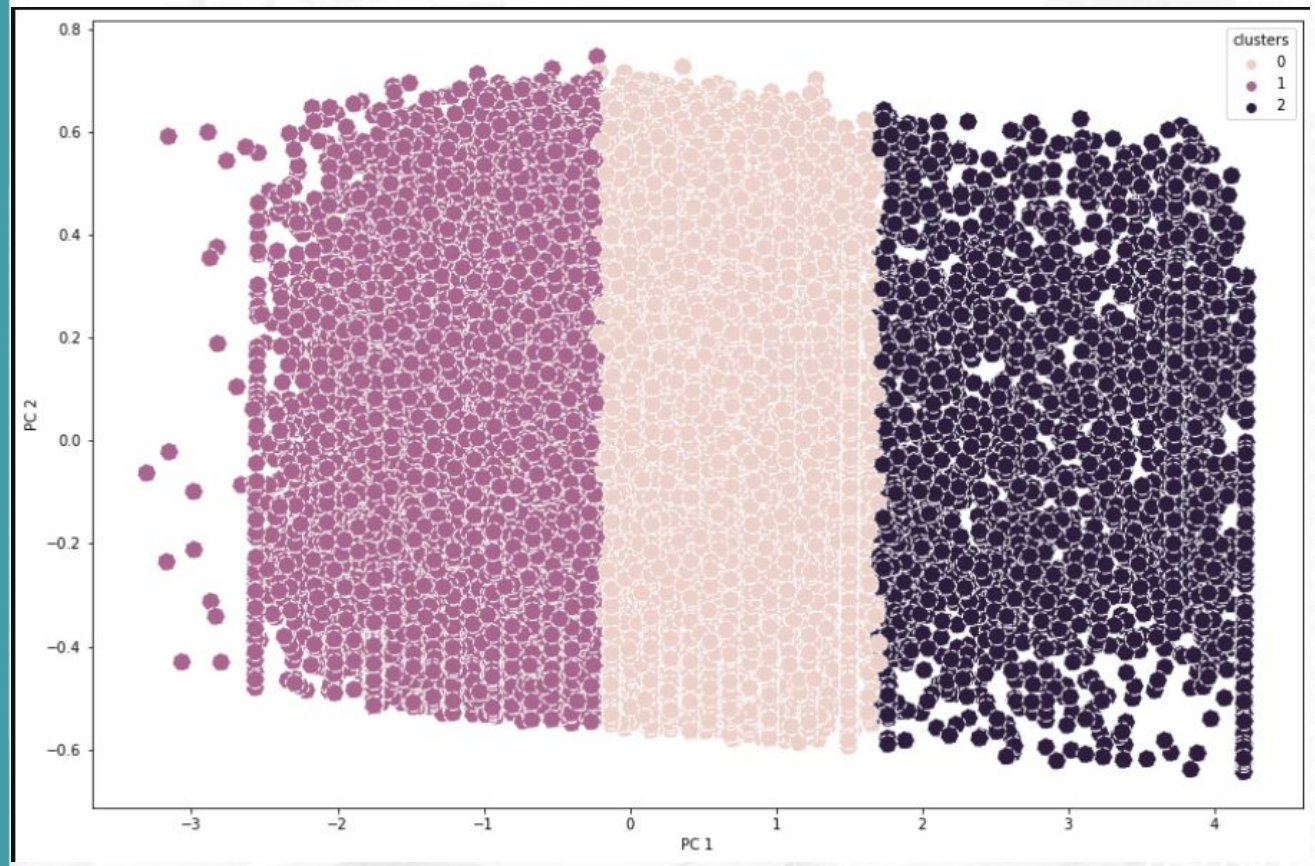




- Elbow Method digunakan untuk mencari jumlah cluster optimal dengan mencari nilai Inertia yaitu nilai yang menunjukkan total jarak setiap titik ke pusatnya.

- Dilihat dari nilai inertia dan silhouette score, informasi yang didapat bahwa nilai cluster yang ideal adalah 3

Clusters Analysis



- Menggunakan metode PCA untuk mereduksi dimensi dataset fitur clustering

	PC 1	PC 2
0	1.322036	0.536434
1	2.889383	0.421894
2	2.901836	0.416626
3	2.005285	0.185232
4	1.359489	0.235990

Analysis

clusters	MEMBER_DURATION	LAST_TO_END	FLIGHT_COUNT	SEG_KM_SUM	avg_discount
0	43.0	102.0	8.0	10846.0	0.782476
1	39.0	114.0	6.0	9176.0	0.591321
2	53.0	78.0	9.0	12898.0	1.275279

	clusters	total_members
0	0	33067
1	1	25792
2	2	2570

Dari 3 cluster yang dihasilkan diketahui hal-hal sebagai berikut:

- Cluster 0 (middle value customer)
  - Member pada cluster 0 memiliki durasi membership selama 43 bulan yang memiliki jarak penerbangan sebesar 10.846 kilometer yang memiliki rata rata penerbangan diatas 8 kali
- Cluster 1 (low value customer)
  - Member pada cluster 1 memiliki durasi membership selama 39 bulan yang memiliki jarak penerbangan sebesar 9.176 kilometer yang memiliki rata rata penerbangan diatas 6 kali
- Cluster 2 (high value customer)
  - Member pada cluster 2 memiliki durasi membership selama 53 bulan yang memiliki jarak penerbangan sebesar 12.898 kilometer yang memiliki rata rata penerbangan diatas 9 kali

BUSINESS RECOMMENDATION

- Mempertahankan high class customer dengan memberikan fasilitas first class agar dapat meningkatkan kepuasan mereka terhadap maskapai
- Memberikan promo yang menarik pada middle class dan low class customer agar dapat meningkatkan transaksi pembelian tiket pesawat dan memberi reedem point yang bertujuan untuk penukaran tiket pesawat gratis