# Loyalty, Recency, Frequency, Monetary, Discount (LRFMC) Customer Clustering Analysis

# 1.1 Exploratory Data Analysis (EDA)

## Data Numerik

```
df_num.describe(percentiles=[0.05,0.25,0.50,0.75,0.95])
```

| | 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 | MEMBER_DUR | year | FLIGHT_COUNT/YEAR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 57022.000000 | 57022.000000 | 57022.000000 | 57022.000000 | 57022.000000 | 57022.000000 | 57022.000000 | 57022.000000 | 57022.000000 | 57022.000000 | 57022.000000 | 57022.000000 | 57022.000000 | 57022.000000 | 57022.000000 | 57022.000000 | 57022.000000 | 57022.000000 |
| mean | 31539.439602 | 4.084108 | 41.917242 | 11.758286 | 10096.792764 | 4963.086831 | 5250.246344 | 16847.696223 | 175.849655 | 67.644872 | 166.134211 | 0.304619 | 0.695734 | 11863.402459 | 2.806759 | 47.286574 | 3.934569 | 3.867437 |
| std | 18192.182300 | 0.333306 | 9.613975 | 13.879848 | 14027.097505 | 7053.862895 | 7705.310474 | 20338.286116 | 183.473922 | 77.205280 | 123.275971 | 1.074167 | 0.144009 | 18018.737985 | 7.506145 | 27.476643 | 2.306114 | 4.982851 |
| min | 3.000000 | 4.000000 | 6.000000 | 2.000000 | 0.000000 | 0.000000 | 0.000000 | 368.000000 | 1.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 12.000000 | 1.000000 | 0.000000 |
| 5% | 3207.050000 | 4.000000 | 29.000000 | 2.000000 | 852.000000 | 0.000000 | 0.000000 | 2033.000000 | 4.000000 | 4.000000 | 4.000000 | 0.000000 | 0.436498 | 946.000000 | 0.000000 | 14.000000 | 1.000000 | 0.000000 |
| 25% | 15771.250000 | 4.000000 | 35.000000 | 3.000000 | 2450.000000 | 980.000000 | 770.000000 | 4743.000000 | 30.000000 | 23.438125 | 79.000000 | 0.000000 | 0.605482 | 2718.000000 | 0.000000 | 23.000000 | 2.000000 | 1.000000 |
| 50% | 31620.500000 | 4.000000 | 41.000000 | 7.000000 | 5506.000000 | 2747.500000 | 2711.500000 | 9941.500000 | 107.000000 | 44.666667 | 143.000000 | 0.000000 | 0.703496 | 6152.000000 | 0.000000 | 40.000000 | 3.000000 | 2.000000 |
| 75% | 47315.750000 | 4.000000 | 48.000000 | 15.000000 | 12202.750000 | 6244.000000 | 6572.750000 | 20979.750000 | 267.000000 | 82.000000 | 228.000000 | 0.000000 | 0.794531 | 13700.500000 | 1.000000 | 70.000000 | 6.000000 | 5.000000 |
| 95% | 59900.950000 | 5.000000 | 59.000000 | 39.000000 | 34544.900000 | 17596.900000 | 19447.850000 | 54527.950000 | 579.000000 | 208.500000 | 407.000000 | 2.000000 | 0.924684 | 39927.650000 | 20.000000 | 98.000000 | 8.000000 | 13.000000 |
| max | 62988.000000 | 6.000000 | 110.000000 | 210.000000 | 505308.000000 | 239560.000000 | 234188.000000 | 580717.000000 | 731.000000 | 728.000000 | 728.000000 | 46.000000 | 1.000000 | 985572.000000 | 140.000000 | 112.000000 | 9.000000 | 128.000000 |

## Data Kategorikal

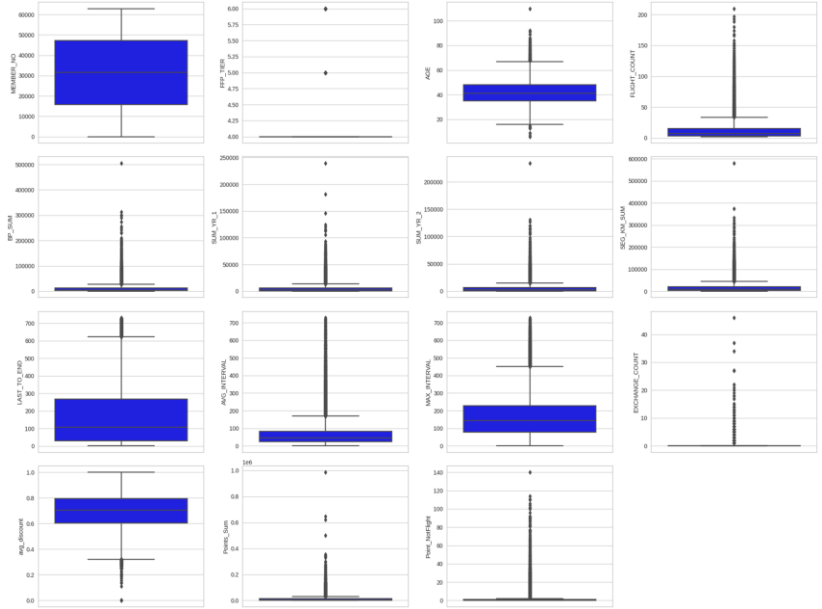
```
[357] df_cat.describe(include =object)
```

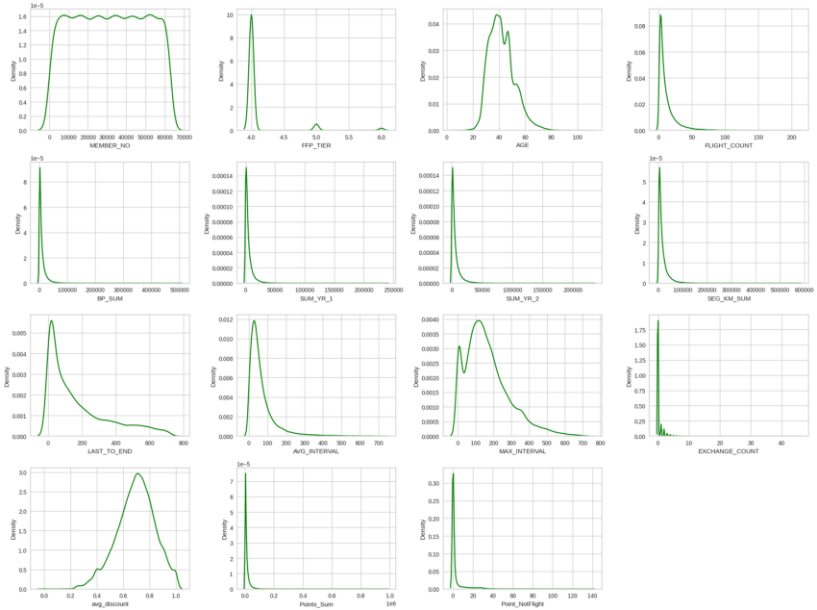
| | FFP_DATE | FIRST_FLIGHT_DATE | GENDER | WORK_CITY | WORK_PROVINCE | WORK_COUNTRY | LOAD_TIME | LAST_FLIGHT_DATE |
|---|---|---|---|---|---|---|---|---|
| count | 57022 | 57022 | 57022 | 57022 | 57022 | 57022 | 57022 | 57022 |
| unique | 3060 | 3392 | 2 | 2825 | 1093 | 103 | 1 | 731 |
| top | 1/13/2011 | 2/16/2013 | Male | guangzhou | guangdong | CN | 3/31/2014 | 3/31/2014 |
| freq | 180 | 89 | 43266 | 9523 | 16919 | 53879 | 57022 | 877 |

## Data Descriptive

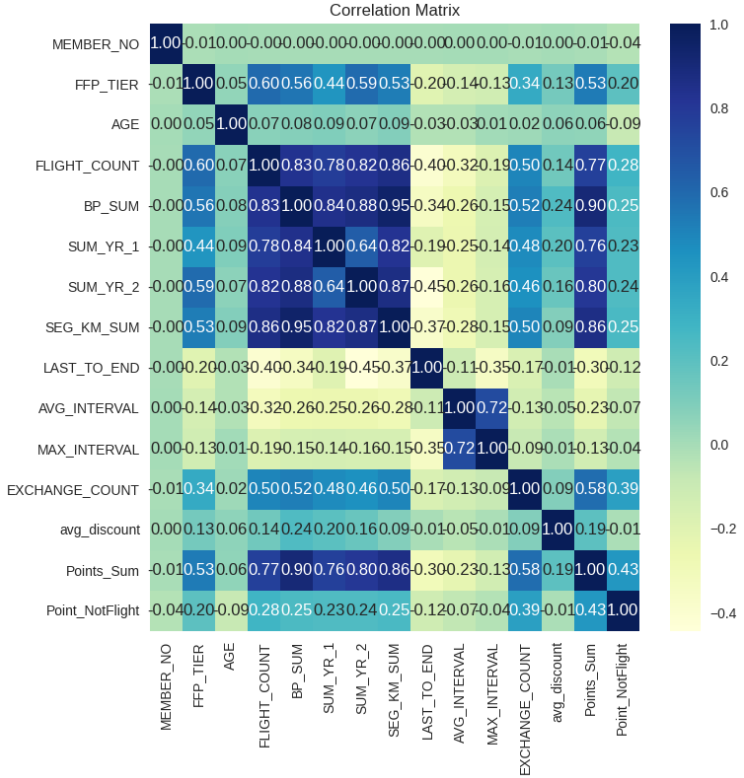
- Customer yang menggunakan layanan memiliki umur berkisar antara 31 sampai 56 tahun, dimana pada rentang umur berikut merupakan orang yang dalam golongan produktif
- FFP_Date / member terbanyak mendaftar pada tanggal 1/13/2011
- Penerbangan didominasi oleh laki-laki
- Customer kebanyakan bekerja di Negara China

# 1.1 Exploratory Data Analysis (EDA)-Lanjutan



# 1.1 Exploratory Data Analysis (EDA)-Lanjutan




**Data Distribution & Correlation**

- Banyak data yang memiliki outlier
- Mayoritas data berdistribusi Positively Skewed kecuali kolom `avg_discount` yang berdistribusi normal
- Golongan kolom yang memiliki korelasi rendah adalah `MEMBER_NO`, `AGE`, `MAX_INTERVAL`, `Point_NotFlight`
- Terdapat banyak data yang memiliki korelasi yang sangat kuat yang kemungkinan menjadi data redundant
- `FLIGHT_COUNT`, `BP_SUM`, `SUM_YR_1`, `SUM_YR_2`,`SEG_KM_SUM`, dan `Points_Sum` berkorelasi kuat satu sama lain.
- `AVG_INTERVAL` dan `MAX_INTERVAL` berkorelasi kuat satu sama lain
- Pada data redundan akan digunakan salah satu dan drop yang lain
- `AGE`, `MEMBER_NO`,`AVG_INTERVAL`, `MAX_INTERVAL` akan di drop karena memiliki korelasi yang rendah dengan fitur lain

# 1.2 Data Pre-Processing

```
df.info()

<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
```

**Before**

```
df_edit.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 57022 entries, 0 to 62987
Data columns (total 23 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   MEMBER_NO          57022 non-null  int64
 1   FFP_DATE           57022 non-null  object
 2   FIRST_FLIGHT_DATE  57022 non-null  object
 3   GENDER             57022 non-null  object
 4   FFP_TIER           57022 non-null  int64
 5   WORK_CITY          57022 non-null  object
 6   WORK_PROVINCE      57022 non-null  object
 7   WORK_COUNTRY       57022 non-null  object
 8   AGE                57022 non-null  float64
 9   LOAD_TIME          57022 non-null  object
 10  FLIGHT_COUNT       57022 non-null  int64
 11  BP_SUM             57022 non-null  int64
 12  SUM_YR_1           57022 non-null  float64
 13  SUM_YR_2           57022 non-null  float64
 14  SEG_KM_SUM         57022 non-null  int64
 15  LAST_FLIGHT_DATE   57022 non-null  object
 16  LAST_TO_END        57022 non-null  int64
 17  AVG_INTERVAL       57022 non-null  float64
 18  MAX_INTERVAL       57022 non-null  int64
 19  EXCHANGE_COUNT     57022 non-null  int64
 20  avg_discount       57022 non-null  float64
 21  Points_Sum         57022 non-null  int64
 22  Point_NotFlight    57022 non-null  int64
dtypes: float64(5), int64(10), object(8)
memory usage: 10.4+ MB
```

**After**

**Summary**

- Dataset yang tersedia terdapat sebanyak 62988 baris yang terdiri dari 23 kolom
- Semua data sesuai dengan jenis datanya masing-masing kecuali `FFP_TIER` kemungkinan kategorikal
- Ada beberapa tipe data yang belum sesuai terutama kolom kelompok time (`FFP_DATE`,`FIRST_FLIGHT_DATE`, `LOAD_TIME`,`LAST_FLIGHT_DATE`), dan akan diubah kolom tersebut ke format datetime.

# 1.2 Data Pre-Processing - Lanjutan

```
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
```

**Before**

```
df_edit.isna().sum()
```

```
MEMBER_NO              0
FFP_DATE               0
FIRST_FLIGHT_DATE      0
GENDER                 0
FFP_TIER               0
WORK_CITY              0
WORK_PROVINCE          0
WORK_COUNTRY           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
```

**After**

```
df_edit.duplicated().sum()
```

```
0
```

**Tidak ada data duplikat**

**Summary**

- Ada 7 kolom yang tidak memiliki nilai (Value #N/A) yaitu ‘Gender’ (3), `WORK_CITY` (2269) ,`WORK_PROVINCE` (3248), `WORK_COUNTRY` (26), `AGE` (420), `SUM_YR_1` (551), `SUM_YR_2` (138)
- Data yang memiliki #N/A dihapus pada kolom WORK_PROVINCE sehingga keseluruhan data berkurang sebesar 5.16%
- Masih ada beberapa data yang kosong, data tersebut diisi menggunakan nilai median untuk data numerical dan modus untuk data kategorikal
- Tidak ada data duplikat pada dataset ini
- Kolom `LOAD_TIME` hanya memiliki 1 nilai unique

# 2. Feature Engineering

Pemilihan fitur untuk clustering menggunakan analisis LRFMC. Analisis LRFMC adalah versi lanjutan dari analisis RFM yang telah digunakan dalam industri penerbangan selama bertahun-tahun untuk membagi pelanggan menjadi beberapa segmen.

| Variabel | Deskripsi Variabel | Kolom pada Dataset |
|---|---|---|
| Loyalty (L) | Lama waktu passenger menjadi membership (dalam bulan) | `` `LOAD_DATE` - `FFP_DATE` `` |
| Recency (R) | Jumlah bulan sejak penerbangan terakhir passenger | `` `LAST_TO_END` `` |
| Frequency (F) | Jumlah penerbangan | `` `FLIGHT_COUNT` `` |
| Monetery (M) | Jarak akumulasi penerbangan | `` `SEG_KM_SUM` `` |
| Discount (C) | Rata-rata discount yang digunakan passenger | `` `avg_discount` `` |

# 2.1 Penambahan Feature

```
[ ]  # L (lOYALTY)
     df_edit['FFP_DATE'] = pd.to_datetime(df_edit['FFP_DATE'], format='%m/%d/%Y')
     df_edit['LOAD_TIME'] = pd.to_datetime(df_edit['LOAD_TIME'], format='%m/%d/%Y')
```

```
#`LOAD_DATE` - `FFP_DATE`
selisih_tahun = df_edit['LOAD_TIME'].dt.year-df_edit['FFP_DATE'].dt.year
selisih_bulan = df_edit['LOAD_TIME'].dt.month-df_edit['FFP_DATE'].dt.month
df_num['MEMBER_DUR'] = selisih_tahun*12+selisih_bulan
```

```
[ ]  #F (FREQUENCY)
     df_num['year']=round(df_num['MEMBER_DUR']/12)
     df_num['FLIGHT_COUNT/YEAR'] = round(df_num['FLIGHT_COUNT']/df_num['year'])
     df_num['FLIGHT_COUNT/YEAR'].head()
```

- **“MEMBER_DUR”**

Penambahan feature “MEMBER_DUR” bertujuan untuk mencari value Loyalty (L) dari LRFMC dalam satuan bulan. Oleh karena itu, kolom “FFP_DATE” dan “LOAD_TIME” harus diubah kesatuan bulan terlebih dahulu lalu dikurang untuk mengetahui berapa lama customer telah menjadi member

- **“FLIGHT_COUNT/YEAR”**

Penambahan feature “FLIGHT_COUNT/YEAR” bertujuan untuk mencari value Frequency (F) dari LRFMC dalam rata-rata per tahun.

# 2.2 Pemilihan Feature

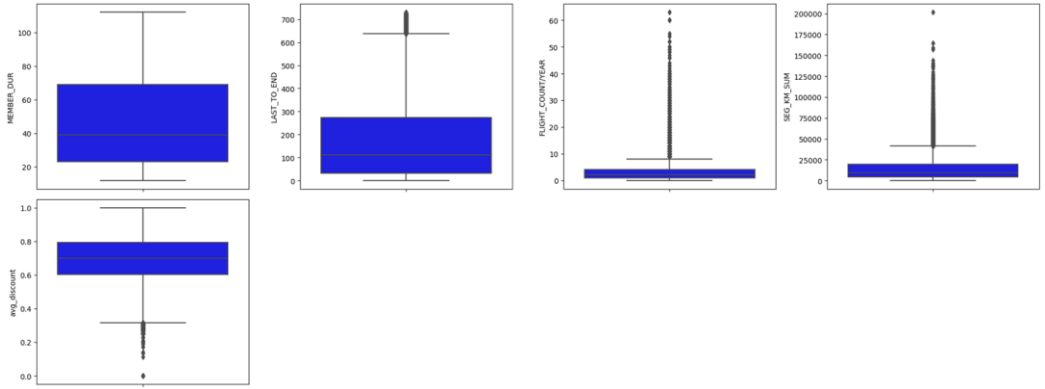
```
x=['MEMBER_DUR','LAST_TO_END','FLIGHT_COUNT/YEAR','SEG_KM_SUM','avg_discount']
df_feature = df_num[x].copy()
```

| Variabel | Kolom pada Dataset Setelah Penambahan Feature |
|---|---|
| Loyalty (L) | `MEMBER_DUR` |
| Recency (R) | `LAST_TO_END` |
| Frequency (F) | `FLIGHT_COUNT/YEAR` |
| Monetery (M) | `SEG_KM_SUM` |
| Discount (C) | `avg_discount` |

# 2.3 Handling Outlier (Z-Score)

```
#Using Z-score for outlier removal

for col in numerical:
    z_scores = np.abs(stats.zscore(df_num[col]))
    filter_mask_z = (z_scores < 3)  # Adjust the threshold as needed
    df_feature = df_feature[filter_mask_z]
```




Setelah feature selection dilakukan, dilakukan handling outlier dengan menggunakan Z-Score dan berhasil me-remove sekitar 13.0%  dari 62988 (data original) jadi 54799 (setelah remove #N/A value serta handling outlier)

# 2.3 Standarisasi

```
[ ]  from sklearn.preprocessing import StandardScaler

     scaler = StandardScaler()
     df_feature_norm = scaler.fit_transform(df_feature)
     df_feature_norm = pd.DataFrame(df_feature_norm, columns=df_feature.columns)

     #output
     df_feature_1 = df_feature_norm
     df_feature_1.describe()
```

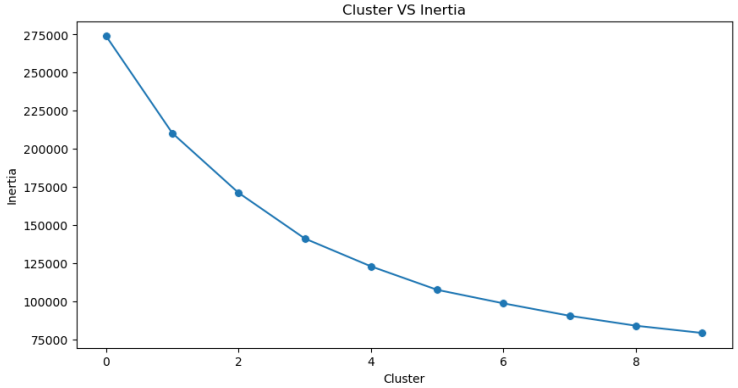
| | MEMBER_DUR | LAST_TO_END | FLIGHT_COUNT/YEAR | SEG_KM_SUM | avg_discount |
|---|---|---|---|---|---|
| count | 5.479900e+04 | 5.479900e+04 | 5.479900e+04 | 54799.000000 | 5.479900e+04 |
| mean | 7.053692e-17 | 1.659692e-17 | 4.149230e-17 | 0.000000 | 1.120292e-15 |
| std | 1.000009e+00 | 1.000009e+00 | 1.000009e+00 | 1.000009 | 1.000009e+00 |
| min | -1.275585e+00 | -9.730505e-01 | -8.410423e-01 | -0.984645 | -4.797907e+00 |
| 25% | -8.717107e-01 | -8.045089e-01 | -6.027097e-01 | -0.688627 | -6.310252e-01 |
| 50% | -2.842565e-01 | -3.695628e-01 | -3.643771e-01 | -0.350064 | 4.901885e-02 |
| 75% | 8.172201e-01 | 5.112030e-01 | 1.122880e-01 | 0.336903 | 6.872340e-01 |
| max | 2.396003e+00 | 2.995833e+00 | 1.417391e+01 | 12.960141 | 2.124071e+00 |

Setelah melakukan handling outlier, dilakukan proses standardisasi dengan menggunakan standardscaler untuk menyamakan skala setiap fitur yang ada.

# 3.1 Clustering - Elbow Method

```
df_inertia = pd.DataFrame(inertia, columns=['inertia'])
df_inertia['delta_inertia'] = round(df_inertia.inertia - df_inertia['inertia'].shift(-1),1)
df_inertia['persen_delta'] = round(df_inertia['delta_inertia']/df_inertia['inertia'] *100,1)
df_inertia
```

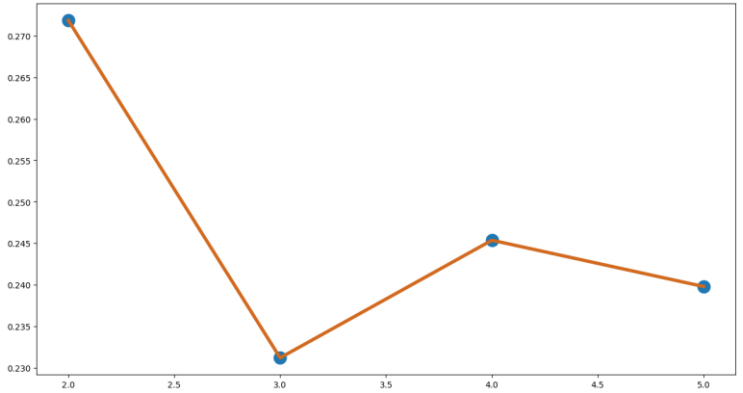
| | inertia | delta_inertia | persen_delta |
|---|---|---|---|
| 0 | 273995.000000 | 63675.5 | 23.2 |
| 1 | 210319.462560 | 39155.4 | 18.6 |
| 2 | 171164.105473 | 29980.9 | 17.5 |
| 3 | 141183.182133 | 18275.2 | 12.9 |
| 4 | 122908.031318 | 15393.7 | 12.5 |
| 5 | 107514.362432 | 8866.0 | 8.2 |
| 6 | 98648.397195 | 8156.3 | 8.3 |
| 7 | 90492.109264 | 6520.2 | 7.2 |
| 8 | 83971.872375 | 4741.3 | 5.6 |
| 9 | 79230.527793 | NaN | NaN |




Jumlah kelompok optimal terletak pada jumlah kelompok sebelum terjadi peningkatan yang signifikan dalam inersia. Dalam hal ini, kelompok optimal terletak pada k = 4, dimana penurunan inersia mulai melambat.

# 3.2 Clustering - Silhouette Score

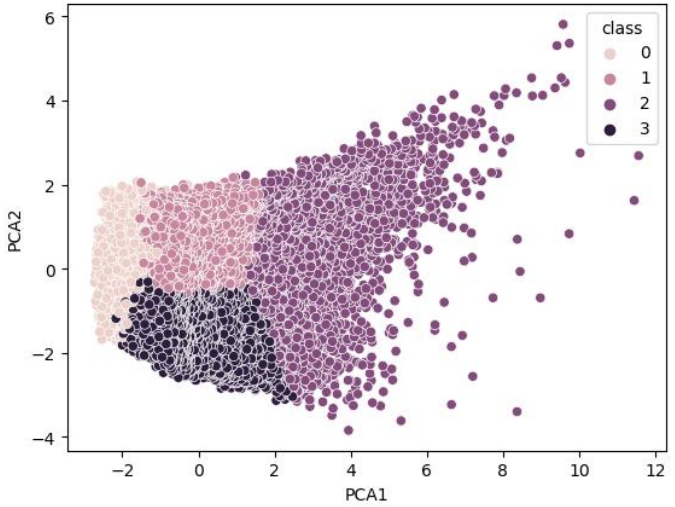
```
For n_clusters = 2, silhouette score is 0.2736635356025175)
For n_clusters = 3, silhouette score is 0.23133431840056243)
For n_clusters = 4, silhouette score is 0.245364569251633)
For n_clusters = 5, silhouette score is 0.24050309229531774)
```




Nilai silhouette score tertinggi ialah n_clusters = 2 lalu diikuti oleh n_cluster = 4, menunjukkan bahwa pembagian data menjadi 2 atau 4 kelompok adalah pilihan yang baik berdasarkan score. Tetapi jika dicocokkan dengan elbow method sebelumnya, jumlah elbow method yang optimal ialah 4. Oleh karena itu, jumlah penentuan cluster dalam case ini tetap menjadi 4 cluster.

# 3.2 Clustering - Evaluation Using PCA

```
# Pemilihan Jumlah Cluster
cluster = 4
kmeans = KMeans(n_clusters=cluster,random_state=0)
kmeans.fit(df_feature_1)
```




Berdasarkan graph PCA di atas, dapat terlihat distribusi dari setiap cluster (atau class) setelah di fit menggunakan k-means dengan jumlah cluster = 4

# 4.1) Statistik Fitur & Deskripsi Setiap Cluster

| | MEMBER_DUR | | LAST_TO_END | | FLIGHT_COUNT/YEAR | | SEG_KM_SUM | | avg_discount | |
|---|---|---|---|---|---|---|---|---|---|---|
| | mean | median | mean | median | mean | median | mean | median | mean | median |
| class | | | | | | | | | | |
| 0 | 39.327788 | 33.0 | 484.719831 | 479.0 | 1.438430 | 1.0 | 5513.649503 | 4154.0 | 0.702406 | 0.714273 |
| 1 | 28.958872 | 28.0 | 114.830240 | 98.0 | 3.491627 | 3.0 | 10259.891455 | 8641.0 | 0.671388 | 0.681353 |
| 2 | 37.884870 | 30.0 | 44.918688 | 22.0 | 11.406117 | 10.0 | 40607.418762 | 37639.0 | 0.714648 | 0.713487 |
| 3 | 79.730259 | 79.0 | 104.814007 | 76.0 | 1.676362 | 1.0 | 15627.579985 | 12592.0 | 0.706179 | 0.710842 |

**Deskripsi Setiap Cluster**

**1. Kelompok 0: "Low-Activity Loyals"**

- Pelanggan dalam kelompok ini memiliki durasi keanggotaan yang cukup lama, tetapi aktivitas terbang dan jarak terbang yang rendah. Mereka memberikan diskon yang moderat, tetapi mungkin memiliki kecenderungan untuk tetap setia meskipun dengan aktivitas terbang yang rendah.

**2. Kelompok 1: "Occasional Flyers"**

- Pelanggan dalam kelompok ini memiliki tingkat aktivitas yang lebih rendah, terlihat dari jumlah penerbangan per tahun dan jarak terbang yang rendah. Mereka juga cenderung memberikan diskon yang lebih tinggi. Kelompok ini mungkin terdiri dari pelanggan yang hanya terbang secara sporadis atau untuk tujuan tertentu.

**3. Kelompok 2: "Variety Explorers"**

- Pelanggan dalam kelompok ini memiliki variasi tinggi dalam semua atribut, termasuk durasi keanggotaan, waktu terakhir terbang, jumlah penerbangan per tahun, jarak terbang, dan diskon yang diberikan. Mereka cenderung memiliki pengalaman penerbangan yang beragam.

**4. Kelompok 3: "Long-Term Explorers"**

- Pelanggan dalam kelompok ini memiliki durasi keanggotaan yang relatif lama, tetapi terlihat bahwa aktivitas terbang (jumlah penerbangan per tahun) dan diskon yang diberikan lebih rendah. Mereka mungkin merupakan pelanggan setia yang tidak terlalu sering terbang.

# 4.2) Business Recommendation

1. Untuk kelompok Low-Activity Loyals diberlakukan pengurangan jumlah discount dikarenakan walaupun jumlah discount yang diberikan relatif tinggi namun kelompok ini masih tidak berminat untuk meningkatkan frekuensi atau aktivitas terbang. Hal ini bertujuan untuk mengurangi marketing cost yang tidak efektif.

2. Untuk kelompok Variety Explorers dan Long-Term Explorers diberlakukan program membership khusus atau program loyalitas khusus yang berbeda setiap kelompoknya. Kelompok Long-Term Explorers hanya melakukan penerbangan pada saat-saat tertentu seperti liburan panjang dimana penumpang membawa banyak barang sehingga diberikan discount biaya penambahan kapasitas bagasi. Kelompok Variety Explorers dengan frekuensi penerbangan yang cukup tinggi, diberikan executive lounge access pada setiap keberangkatan.