Kelompok 2A

Wahyu Afriza
Hanif Zaki Nur Fauzi
Alyani Noor Septalia
Bilqis Nafida Azza
Muhamad Fahrurrozi
Windy Agelina Manalu
Muhammad Rofi'i
Lingga Atha Khairunnisa

UNSUPERVISED LEARNING

AIRLINE CUSTOMER VALUE ANALYSIS CASE

Dataset Summary

- Data shape: 62988 x 23
- Columns drop: 'MEMBER_NO', 'BP_SUM', 'SUM_YR_1', 'SUM_YR_2', 'SEG_KM_SUM'
- Kolom dengan data kosong:

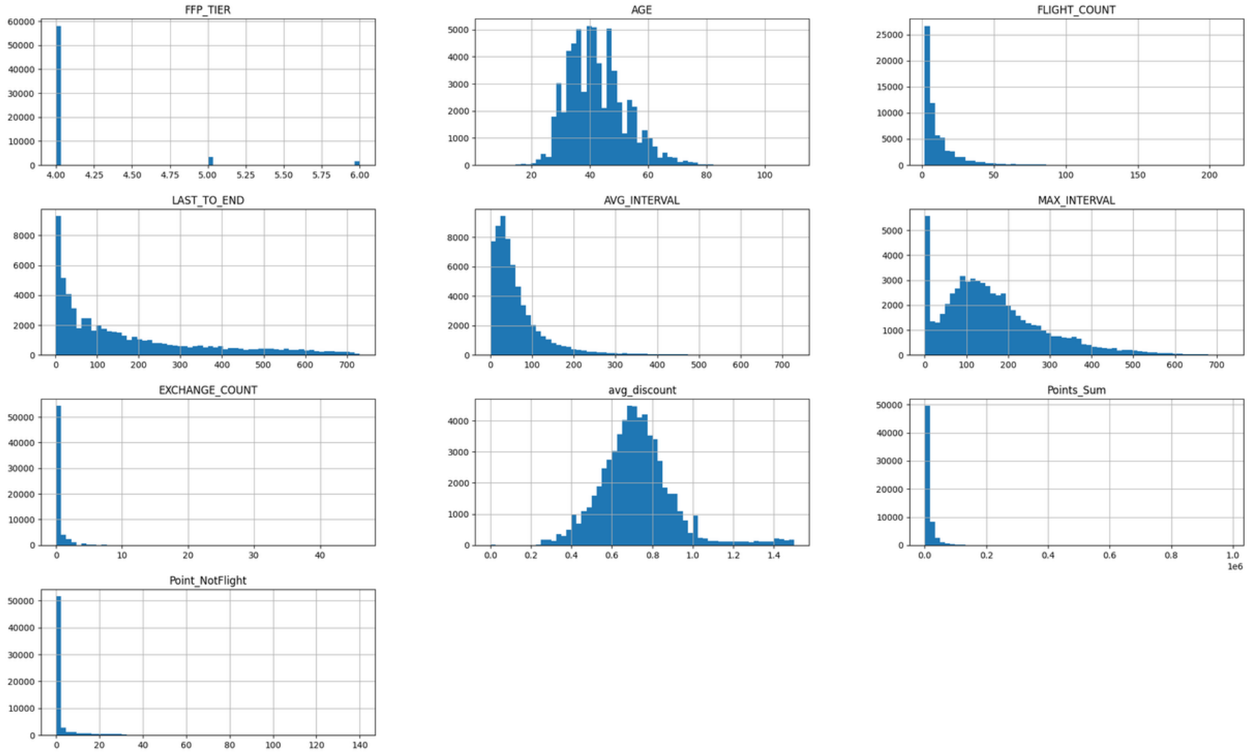
```
data.isna().sum()
```

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
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

Exploratory Data Analysis

- Distribusi data



Dataset Summary

Exploratory Data Analysis

- Grouping data berdasarkan jenis tipe data:

```
# pengelompokan kolom berdasarkan jenisnya
cats = ['GENDER', 'WORK_CITY', 'WORK_PROVINCE', 'WORK_COUNTRY','AVG_INTERVAL', 'avg_discount']
nums = ['FFP_TIER', 'AGE', 'FLIGHT_COUNT', 'LAST_TO_END','MAX_INTERVAL', 'EXCHANGE_COUNT','Points_Sum', 'Point_NotFlight']
timestamp = ['FFP_DATE', 'FIRST_FLIGHT_DATE', 'LOAD_TIME', 'LAST_FLIGHT_DATE']
```

- Handle missing data:
  a. Kolom cats, timestamp: Diisi dengan modus
  b. Kolom nums: Diisi dengan mean

- Duplicate rows:

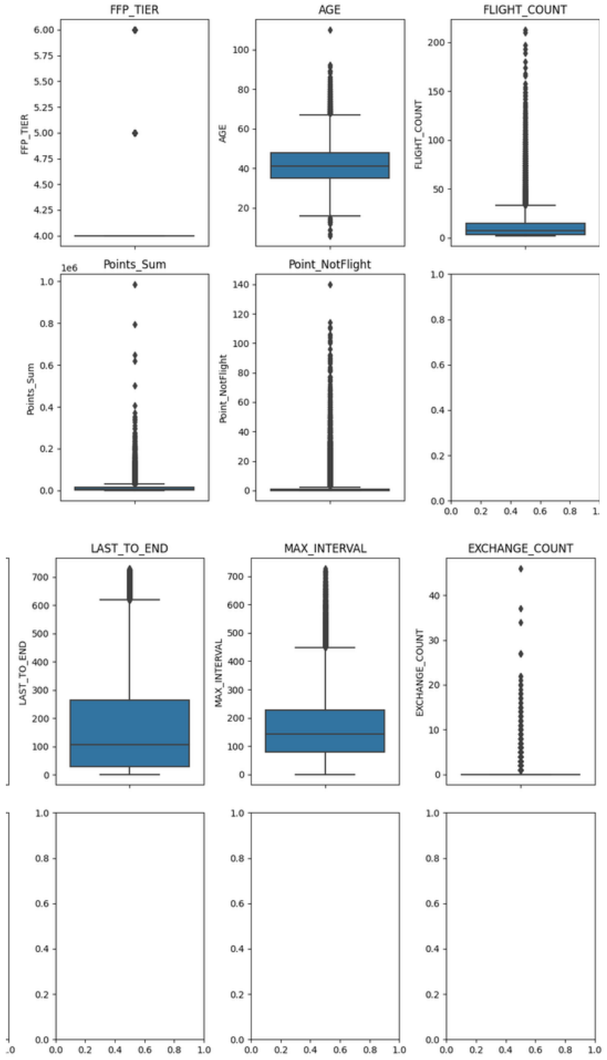
```
duplicate_rows = data[data.duplicated()]
print("Duplicate Rows except first occurrence:")
print(duplicate_rows)
```

```
Duplicate Rows except first occurrence:
       FFP_DATE FIRST_FLIGHT_DATE GENDER  FFP_TIER WORK_CITY WORK_PROVINCE  \
49085  8/11/2012         8/11/2012   Male         4    panjin      liaoning

      WORK_COUNTRY   AGE  LOAD_TIME  FLIGHT_COUNT LAST_FLIGHT_DATE  \
49085           CN  40.0  3/31/2014             2        8/18/2012

       LAST_TO_END  AVG_INTERVAL  MAX_INTERVAL  EXCHANGE_COUNT  avg_discount  \
49085          592           7.0             7               0      0.600021

       Points_Sum  Point_NotFlight
49085        1841                0
```

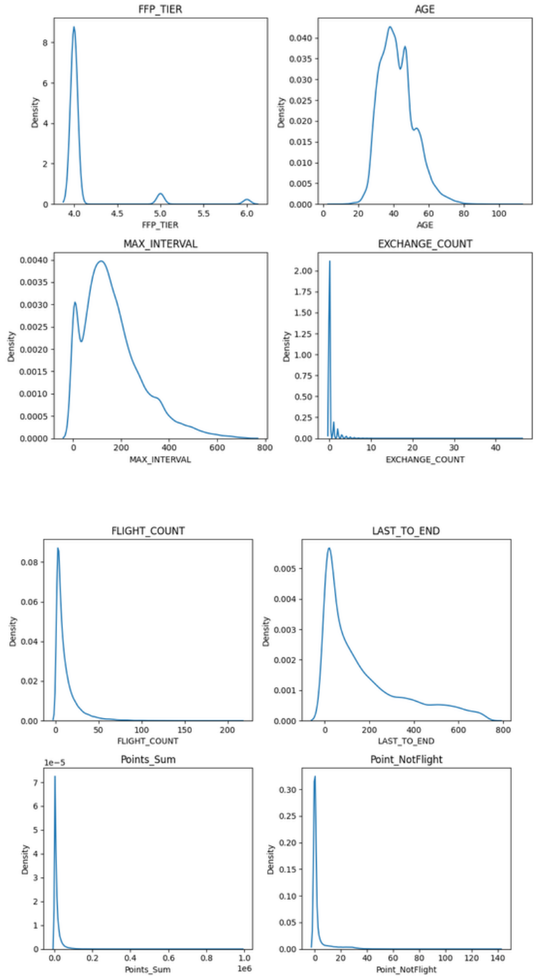
Handling duplicate rows:
drop_duplicates()

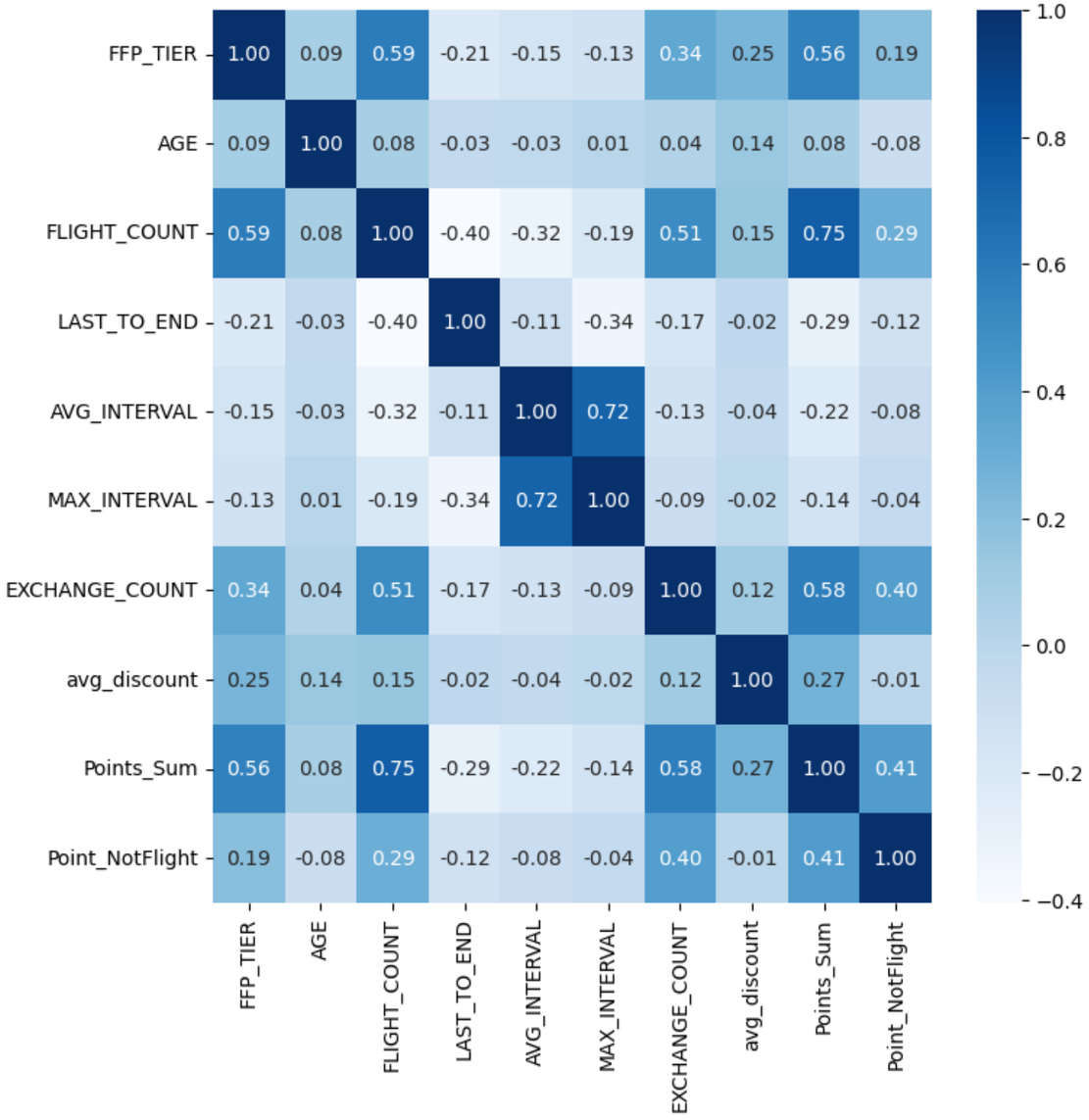
Dataset Summary

- Unique values:
  - FFP_DATE: 3062
  - FIRST_FLIGHT_DATE: 3399
  - GENDER: 2
  - WORK_CITY: 2959
  - WORK_PROVINCE: 1132
  - WORK_COUNTRY: 106
  - LOAD_TIME: 1
  - LAST_FLIGHT_DATE: 731



Exploratory Data Analysis



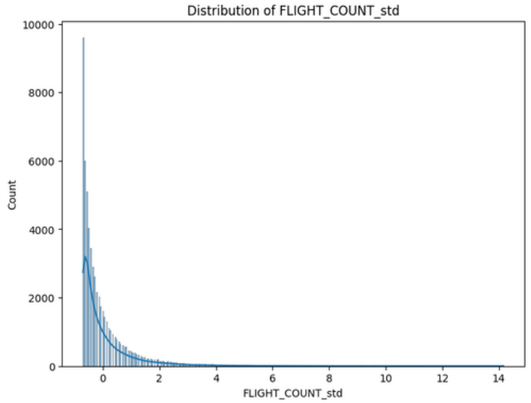


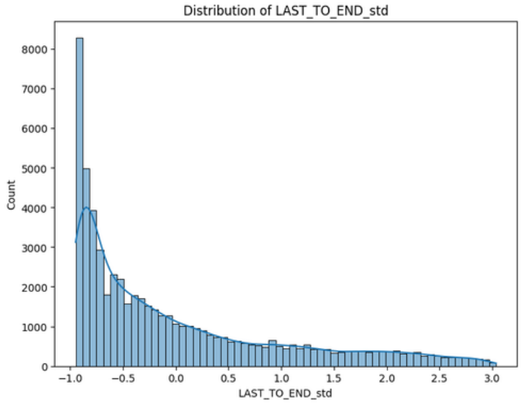
- Kolom-kolom dengan Korelasi Tinggi: ['FLIGHT_COUNT', 'AVG_INTERVAL', 'MAX_INTERVAL', 'Points_Sum']

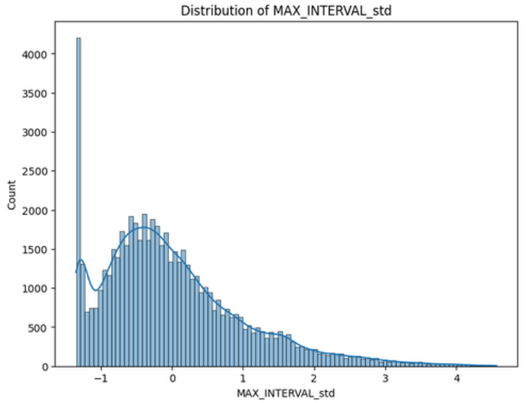
Feature Selection

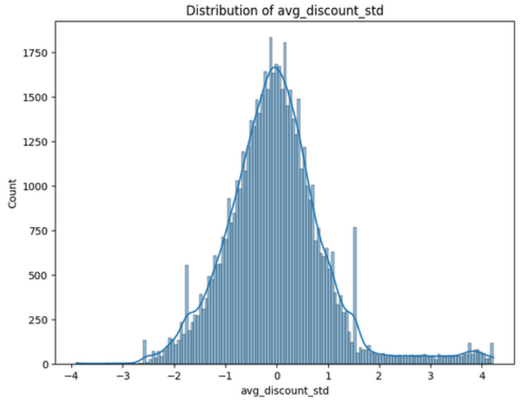
Feature Extraction

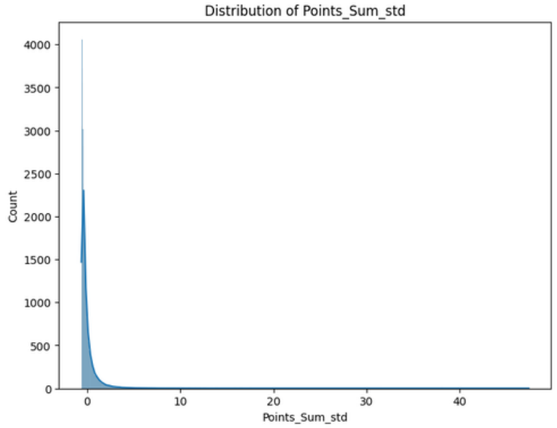
- Membuat kombinasi linier dari beberapa filter yaitu Travel_Score dengan membagi FLIGHT_COUNT dengan LAST_TO_END
- Melakukan standarisasi pada kolom FLIGHT_COUNT, LAST_TO_END, MAX_INTERVAL, EXCHANGE_COUNT, avg_discount, Points_Sum, AVG_INTERVAL, dan Travel_Score
- Men-drop kolom yang tidak dilakukan standarisasi: 'FLIGHT_COUNT_std','LAST_TO_END_std','MAX_INTERVAL_std','EXCHANGE_COUNT_std','avg_discount_std','Points_Sum_std','AVG_INTERVAL_std','Travel_Score_std'

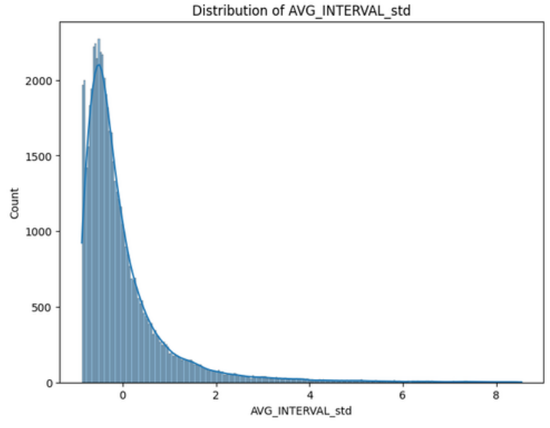


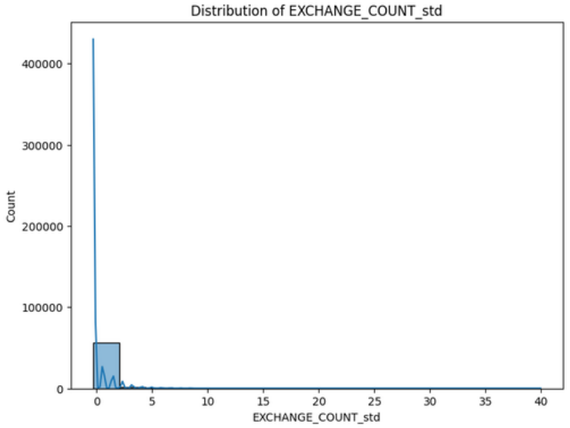


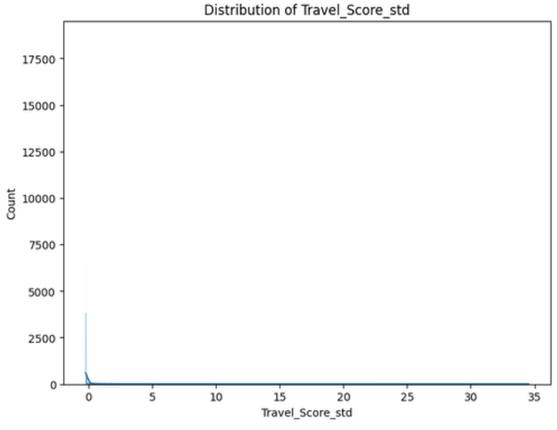




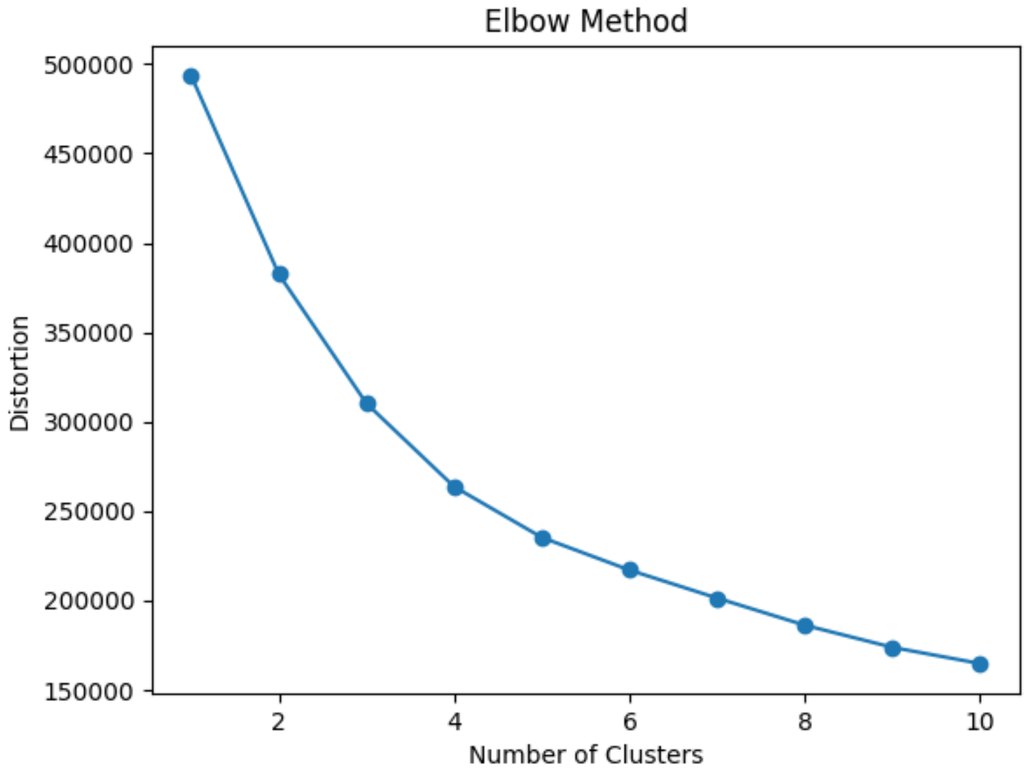




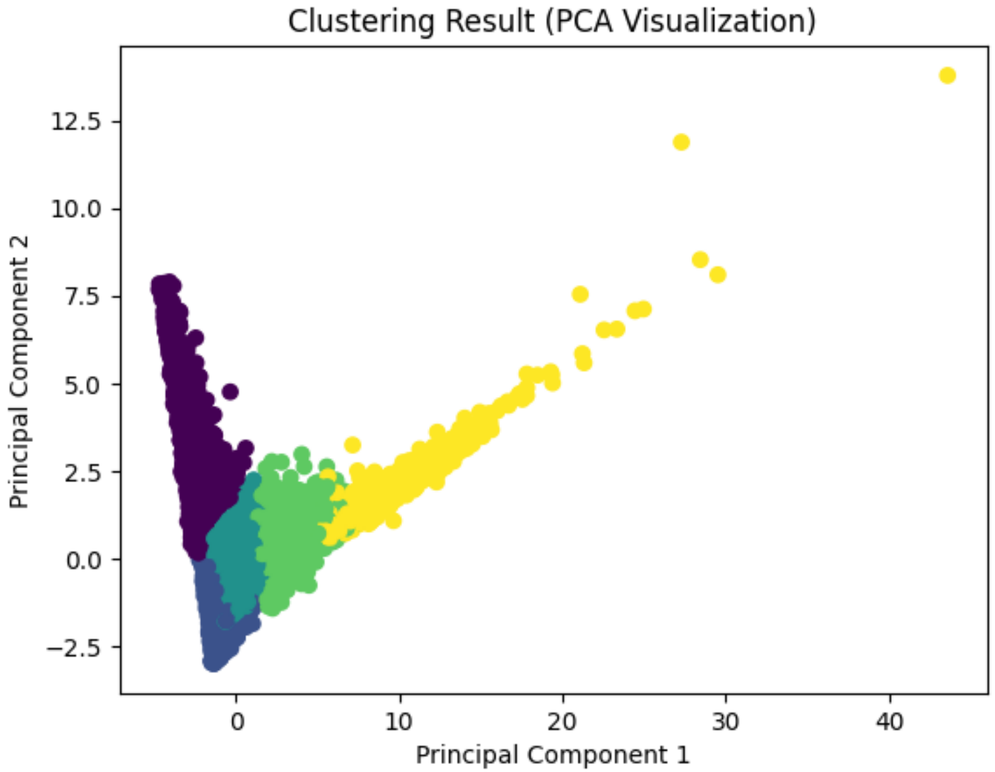




Modelling



Dipilih Cluster sebanyak 5 dengan kondisi tidak menggunakan Cluster yang terlalu banyak namun nilai cukup baik karena nilai setelah tidak terlalu menggambarkan grafik yang terlalu signifikan



Silhouette Score: 0.28825371310470 48

Modelling

Cluster Centers:

	FLIGHT_COUNT_std	LAST_TO_END_std	MAX_INTERVAL_std	EXCHANGE_COUNT_std	\
0	-0.560713	1.573742	-0.826149	-0.242371	
1	-0.073654	-0.432708	-0.000331	-0.166547	
2	1.741547	-0.748856	-0.485031	1.109962	
3	4.486854	-0.907560	-0.818827	4.312208	
4	-0.561065	-0.242786	1.706053	-0.231082	

	avg_discount_std	Points_Sum_std	AVG_INTERVAL_std	Travel_Score_std	\
0	-0.046672	-0.415803	-0.422996	-0.224094	
1	-0.074897	-0.125196	-0.175139	-0.116286	
2	0.502972	1.449561	-0.596862	0.604115	
3	0.784947	4.725832	-0.727132	5.924837	
4	-0.105567	-0.401336	1.784124	-0.193733	

	Cluster
0	1.000000e+00
1	2.000000e+00
2	3.000000e+00
3	4.000000e+00
4	-9.237056e-14

```
[ ]  # Evaluasi cluster dengan silhouette score
     silhouette_avg = silhouette_score(df_filter_std, df['Cluster'])
     print(f'Silhouette Score: {silhouette_avg}')

     Silhouette Score: 0.2904077415166286

     from sklearn.metrics import davies_bouldin_score

     # Evaluasi cluster dengan Davies-Bouldin Index
     db_index = davies_bouldin_score(df_filter_std, df_filter_std['Cluster'])
     print(f'Davies-Bouldin Index: {db_index}')

     Davies-Bouldin Index: 1.1303101628898262
```

- Nilai 0.29 pada Silhouette, menunjukkan bahwa clustering tersebut memiliki sejumlah besar overlapping antar cluster, namun tetap menunjukkan sejauh mana setiap data point berada dalam clusternya masing-masing.
- Nilai 1.13 pada Davis Bouldin menunjukkan tingkat overlap yang moderat antara cluster dan sejauh mana cluster-cluster tersebut berbeda satu sama lain.
- Meskipun Silhouette Score menunjukkan adanya overlap antar cluster, Davies-Bouldin Index menunjukkan bahwa cluster tersebut masih cukup terpisah dan terdefinisi dengan baik.

Evaluation