

Homework Clustering

Airline Customer Value Analysis

Kelompok 8 - Decentraland

Anggota Kelompok:

- Dharma Setiawan
- Ilham Ibnu A.
- M. Farhan Atmawinanda
- Fikri Diva S.
- Ahmad Ilham H.



Dataset Description

Code	Description
MEMBER_NO-b	: ID Member
FFP_DATE	: Frequent Flyer Program Join Date
FIRST_FLIGHT_DATE	: Tanggal Penerbangan pertama
GENDER	: Jenis Kelamin
FFP_TIER	: Tier dari Frequent Flyer Program
WORK_CITY	: Kota Asal
WORK_PROVINCE	: Provinsi Asal
WORK_COUNTRY	: Negara Asal
AGE	: Umur Customer
LOAD_TIME	: Tanggal data diambil
FLIGHT_COUNT	: Jumlah penerbangan Customer
BP_SUM	: Rencana Perjalanan
SUM_YR_1	: Fare Revenue
SUM_YR_2	: Votes Prices
SEG_KM_SUM	: Total jarak(km) penerbangan yg sudah dilakukan
LAST_FLIGHT_DATE	: Tanggal penerbangan terakhir
LAST_TO_END	: Jarak waktu penerbangan terakhir ke pesanan penerbangan paling akhir
AVG_INTERVAL	: Rata-rata jarak waktu
MAX_INTERVAL	: Maksimal jarak waktu
EXCHANGE_COUNT	: Jumlah penukaran
avg_discount	: Rata rata discount yang didapat customer
Points_Sum	: Jumlah poin yang didapat customer
Point_NotFlight	: point yang tidak digunakan oleh members



WORKFLOW

- EDA
- Data Preprocessing
- Modelling
- Interpretasi Model

EDA

- Descriptive statistics
- Fill missing value and fix wrong value
- Univariate Analysis
- Multivariate Analysis

Descriptive Statistics

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)

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

Descriptive Statistics

- Data terdiri dari **62.988 sampel** (baris).
- **Terdapat null value** di beberapa column.
- **Tidak ada baris yang terduplikasi.**
- Terdapat **23 fitur** dengan 15 fitur numerik dan 8 fitur kategorik.
Untuk data kategorikal terdiri dari 8 feature yang terdiri dari cats = ['FFP_DATE', 'FIRST_FLIGHT_DATE', 'GENDER', 'LOAD_TIME', 'LAST_FLIGHT_DATE', 'WORK_CITY', 'WORK_PROVINCE', 'WORK_COUNTRY']

Untuk data numerikal terdiri dari 14 feature yang terdiri dari nums = ['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']

Fill Missing Value & Fix Wrong Value

	avg_discount
count	62988.000000
mean	0.721558
std	0.185427
min	0.000000
25%	0.611997
50%	0.711856
75%	0.809476
max	1.500000

AGE	420
LOAD_TIME	0
FLIGHT_COUNT	0
BP_SUM	0
SUM_YR_1	551
SUM_YR_2	138

- Pengisian data **AGE, SUM_YR_1, SUM_YR_2** dilakukan dengan menggunakan nilai median
- Untuk data **rata-rata diskon** dilakukan filtering dengan nilai maksimalnya adalah 1 atau 100%

Drop column & fix feature Gender

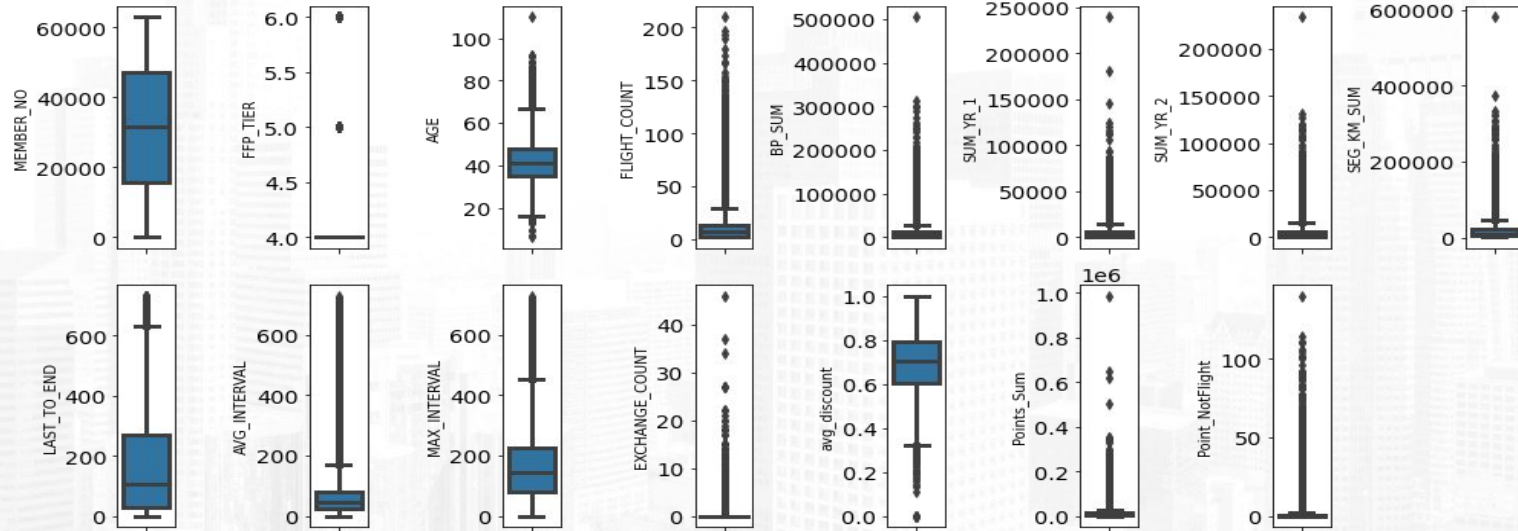
```
df_flight = df_flight.drop(['WORK_CITY', 'WORK_PROVINCE', 'WORK_COUNTRY'], axis = 1)
```

```
df_flight['GENDER'] = df_flight['GENDER'].fillna("Male")
```

MEMBER_NO	0
FFP_DATE	0
FIRST_FLIGHT_DATE	0
GENDER	0
FFP_TIER	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

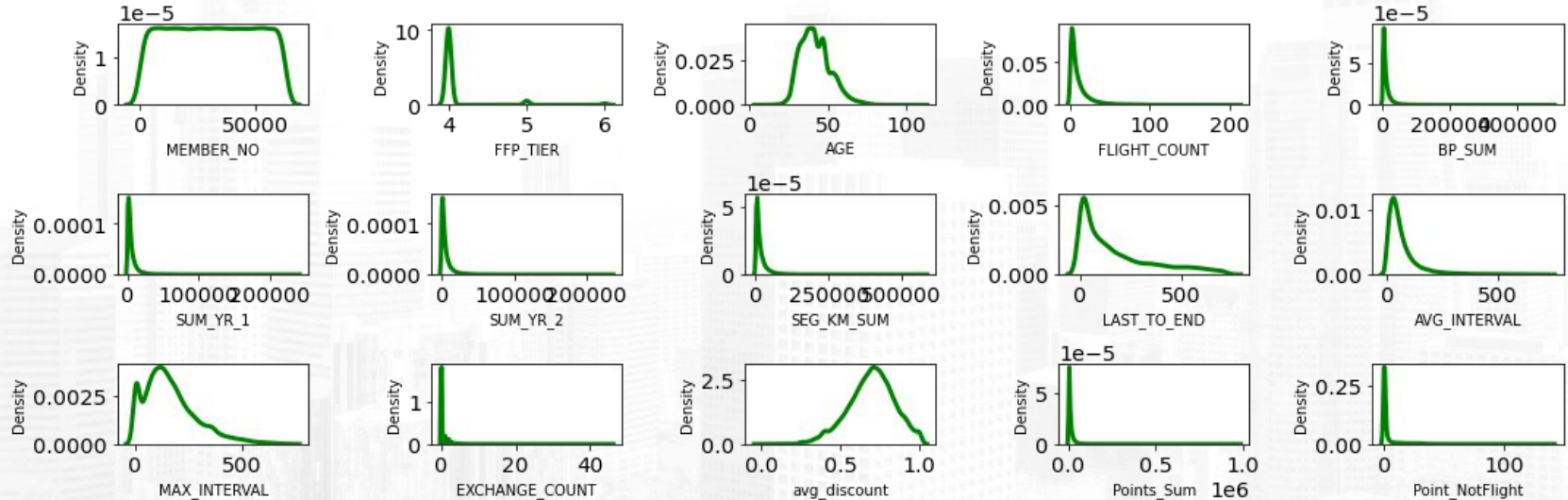
- Data **WORK_CITY, WORK_PROVINCE, WORK_COUNTRY** memiliki nilai unique yang tinggi dan missing value yang banyak sehingga di drop
- Data **Gender** diisi menggunakan modus. Modus dari gender laki-laki

Univariate Analysis



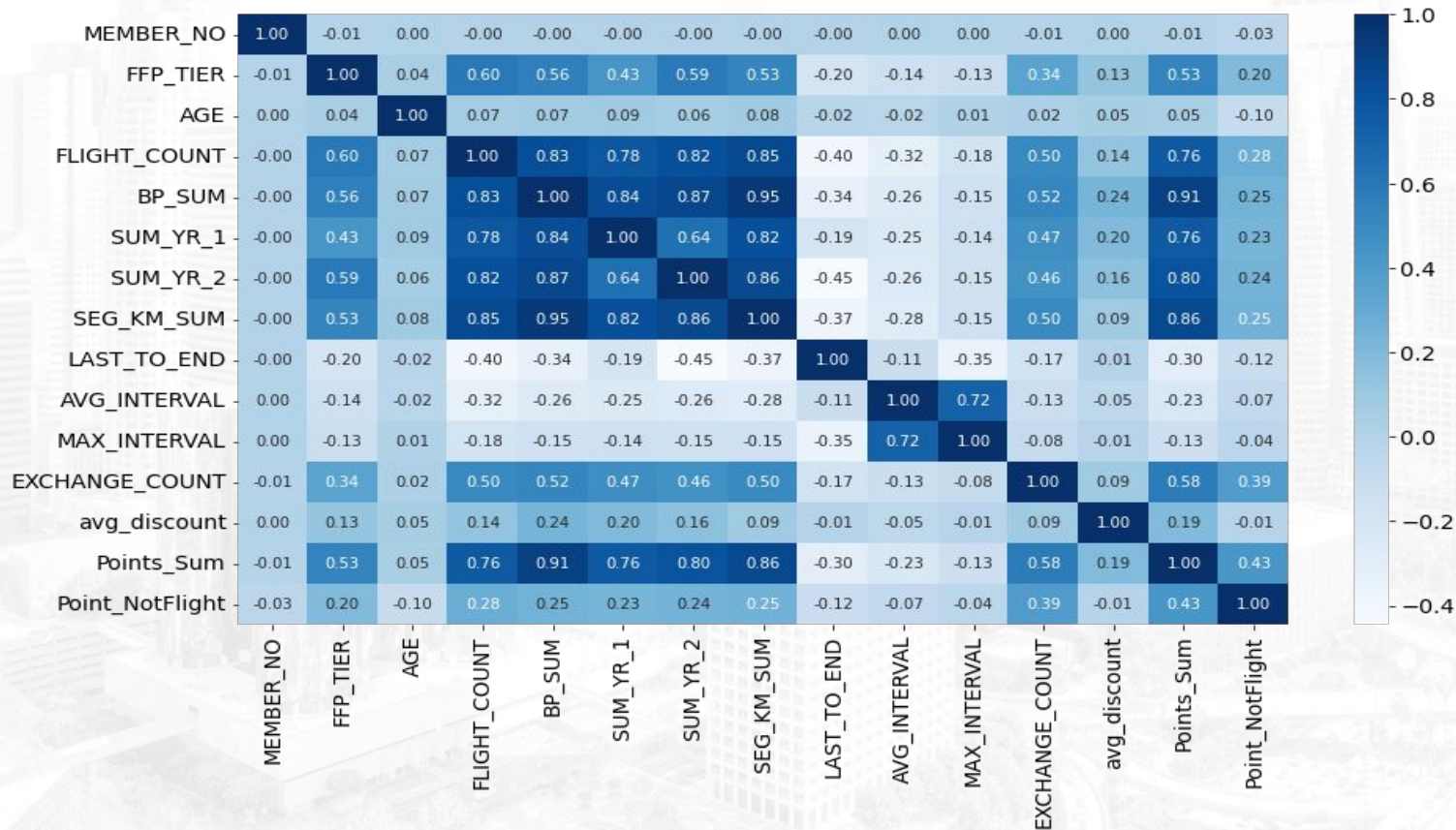
- Terdapat nilai outlier untuk setiap bagian data kecuali data **MEMBER_NO** dikarenakan data tersebut merepresentasikan nomor pelanggan

Univariate Analysis



- Pada distribusi data di atas dijelaskan kalau setiap fitur memiliki ciri skew positif (kanan).
- Akan tetapi, beberapa fitur seperti **avg_discount** dan **AGE** datanya hampir berdistribusi normal

Multivariate Analysis



Multivariate Analysis

- Terdapat beberapa data yang memiliki korelasi rendah dan dianggap tidak berhubungan dalam penyelesaian masalah sehingga data tersebut tidak akan digunakan dalam modelling: **MEMBER_NO, GENDER, EXCHANGE_COUNT, SUM_YR_1, SUM_YR_2, POINT_NOTFLIGHT, AVG_INTERVAL, MAX_INTERVAL, avg_discount**
- Data **BP_SUM, SEG_KM_SUM, POINT_SUM** memiliki nilai korelasi yang tinggi antar fitur sehingga dalam modelling nanti bisa memilih salah satu saja.

Data Preprocessing

- Feature Selection
- Feature Engineering

Feature Selection

- Untuk melakukan segmentasi dari data airline ini dilakukan pemilihan feature menggunakan teknik **LRFM** dari [Jurnal](#).
Apa itu LRFM?
 1. L (Length): lama periode pelanggan berlangganan
 2. R (Retention): jarak waktu penerbangan terakhir ke pesanan penerbangan terakhir
 3. F (Frequency): Berapa kali pelanggan melakukan kegiatan penerbangan
 4. M (Monetary): Total jarak penerbangan yang sudah dilakukan oleh pelanggan

Dari deskripsi tersebut maka kita akan menggunakan feature **Membership (Month)**, **LAST_TO_END**, **FLIGHT_COUNT**, dan **SEG_KM_SUM**

Feature Engineering

- Untuk bagian feature engineering dilakukan pembuatan feature baru yaitu **Membership (Month)**. Fungsi fitur ini adalah untuk mengetahui lama pelanggan berlangganan dari pelanggan bergabung berlangganan hingga tanggal data ini diambil. Maka dari itu diperlukan feature **FFP_DATE** dan **LOAD_TIME** untuk melakukan kalkulasi lama pelanggan berlangganan dalam periode bulan.

4.3 Data Transformation

Data transformation is to transform data into “appropriate” format to meet the needs of mining tasks and algorithms. In this project, the main data transformation method is attribute construction. Because the original data does not directly give the five indicators of LRFMC, the five indicators need to be extracted from the original data. The specific calculation method is as follows:

- $L = \text{LOAD_TIME} - \text{FFP_DATE}$

The number of months between the time of membership and the end of observation window = the end time of observation window - the time of membership [unit: month].

- $R = \text{LAST_TO_END}$

The number of months from the last time the customer took the company's aircraft to the end of the observation window = the time from the last flight to the end of the observation window [unit: month].

- $F = \text{FLIGHT_COUNT}$

Number of times the customer takes the company's aircraft in the observation window = number of flights in the observation window [unit: Times].

- $M = \text{SEG_KM_SUM}$

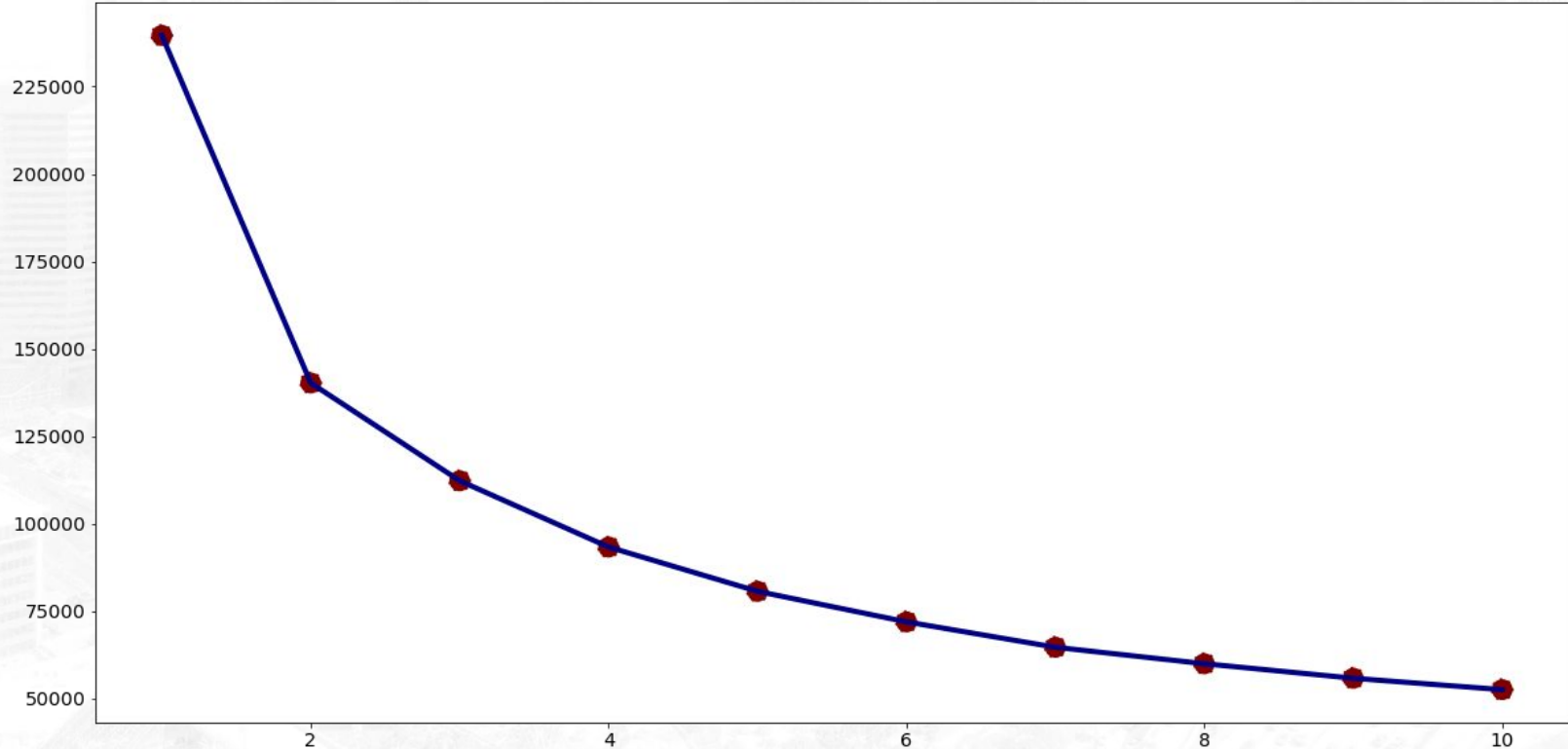
Accumulated flight history of the customer in observation time = total flight kilometers of observation window [unit: km].

Modelling

- Clustering K-Means
- Visualisasi clustering dengan PCA
- Evaluasi K-means

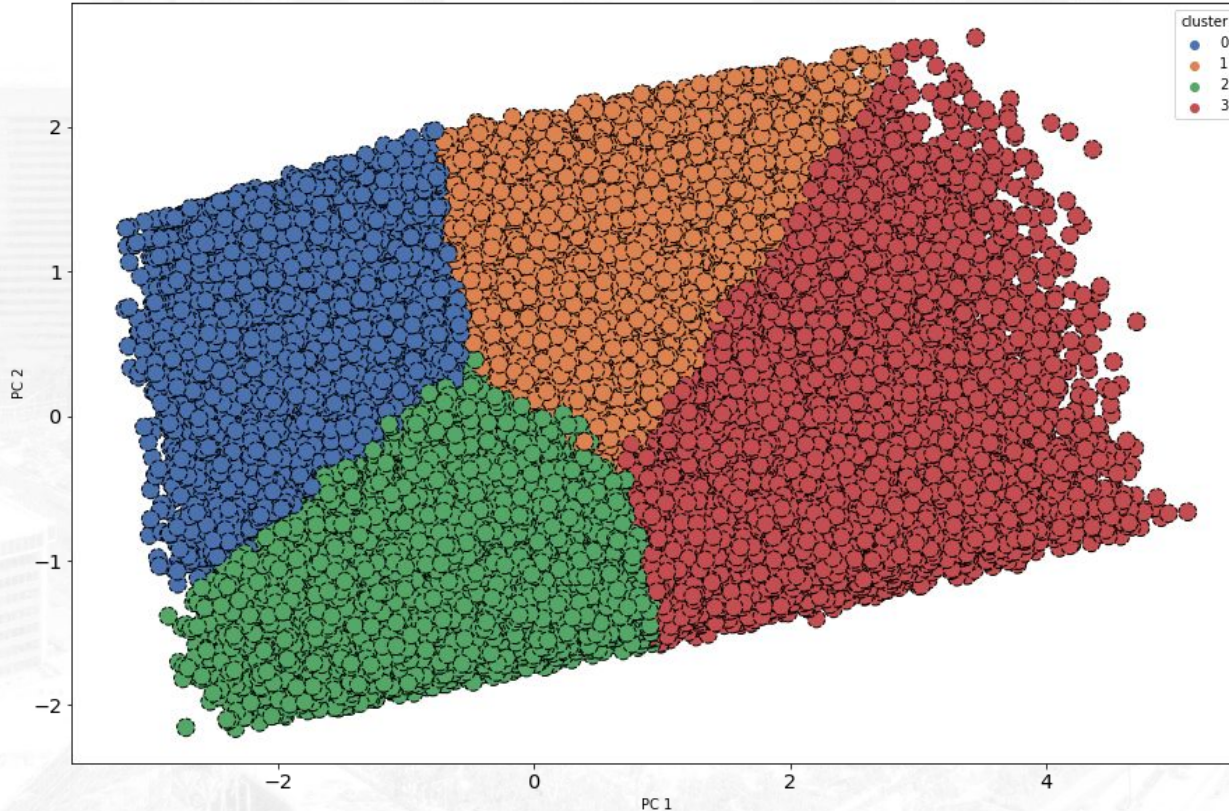
Clustering K-Means

Pada tahap clustering ini kita melakukan dua kali percobaan menggunakan 4 dan 2 cluster



Clustering K-Means

Pada tahap clustering ini kita melakukan dua kali percobaan menggunakan 4 dan 2 cluster



Hasil Clustering dengan 4 Cluster

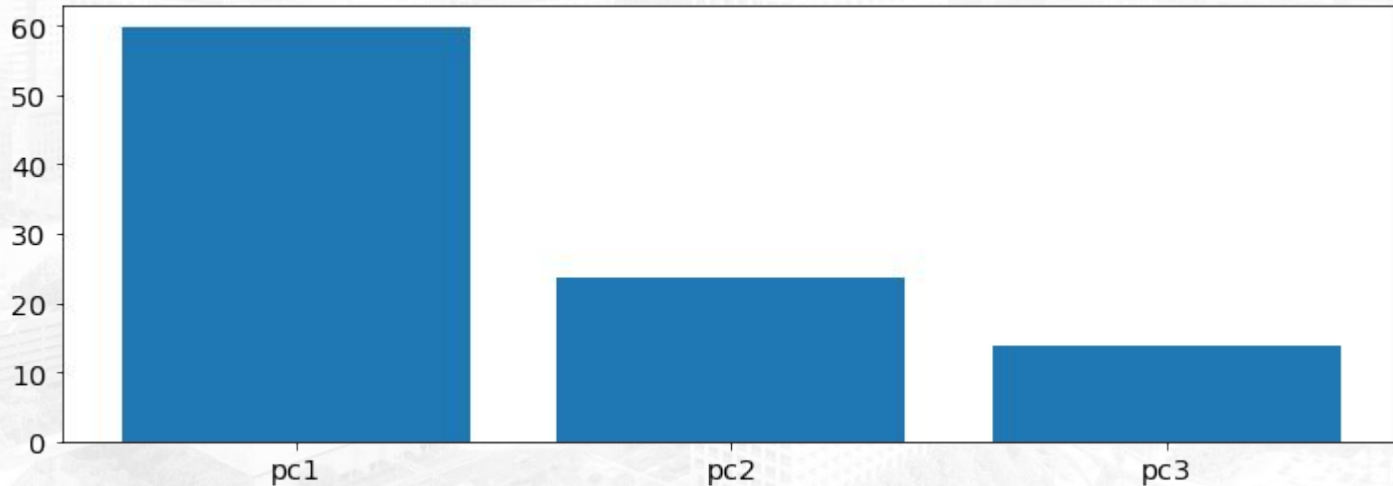
Evaluasi K-Means

Setelah melakukan evaluasi K-Means dengan menggunakan silhouette score, kami mendapati bahwa 2 cluster memiliki nilai silhouette score yang lebih baik dibandingkan dengan 4 cluster. Cluster yang memiliki nilai mendekati 1 adalah cluster yang lebih baik.

Cluster	Silhouette Score
2	0.3507
3	0.2891
4	0.2764
5	0.2618
6	0.2710
7	0.2664
8	0.2470

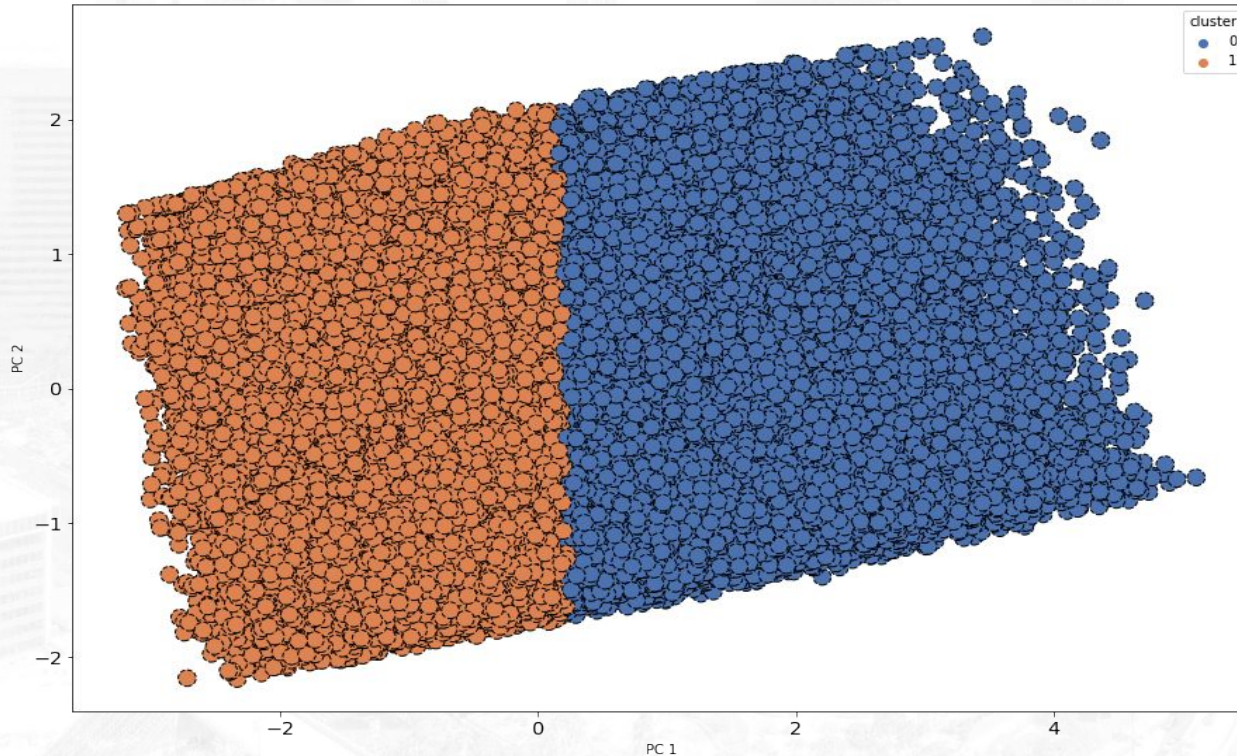
Visualisasi Clustering dengan PCA

- PCA dapat membantu untuk mencari sumbu yang tepat. Pc1 dan pc2 sudah cukup untuk dibuat sumbu karena sudah mewakili 83% data. Hal ini sudah melebihi rule of thumb dari PCA dimana kita harus mengambil 80-90% data.



Clustering K-Means

Pada tahap clustering ini kita melakukan dua kali percobaan menggunakan 4 dan 2 cluster



Hasil Clustering dengan 2
Cluster

Business Insight

- Statistik Tiap Fitur
- Makna Setiap Cluster
- Rekomendasi Bisnis

cluster	MEMBERSHIP_MONTH_COUNT_exp		LAST_TO_END_exp		FLIGHT_COUNT_exp		SEG_KM_SUM_exp	
	mean	median	mean	median	mean	median	mean	median
0	54.836576	51.0	57.750391	30.0	20.977145	16.0	29922.730036	23145.5
1	42.790210	35.0	266.903181	221.0	4.365704	4.0	6395.403108	5326.0

Keterangan column:

MEMBERSHIP_MONTH_COUNT_exp: Lama pelanggan berlangganan dari tanggal pelanggan berlangganan hingga data ini diambil

LAST_TO_END_exp: jarak waktu penerbangan terakhir ke pesanan penerbangan paling akhir

FLIGHT_COUNT_exp: jumlah penerbangan yang dilakukan oleh pelanggan

SEG_KM_SUM_exp: Total jarak penerbangan yang sudah dilakukan oleh pelanggan

Makna dari Cluster

1. Cluster 0 merupakan pelanggan yang sudah lama berlangganan dengan jumlah penerbangan yang banyak dan total jarak penerbangan yang panjang.
2. Cluster 1 merupakan pelanggan member yang memiliki durasi penerbangan yang sedikit dan total jarak penerbangan yang pendek.

1. Untuk customer di cluster 0 harus lebih diperhatikan dengan peningkatan customer service dan juga bisa diberikan banyak penawaran-penawaran yang menarik. Customer di cluster ini lebih baik diberi penawaran berupa poin karena mereka adalah customer yang sering berpergian. **Business Metrics: Retention Rate.**
2. Untuk customer di cluster 1 harus lebih diperhatikan dengan pemberian diskon dan promo karena mereka adalah tipe customer yang cenderung tidak terlalu loyal dan bepergian hanya sesekali.
Business Metrics: Churn Rate