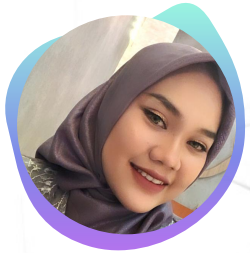
Airline Customer Segmentation

E-NUMPY (BATCH 19A)

THE TEAMS

Ilham Taufiqurrohim

M Chosasih Mahendra



Putri Maylita

Rama Kuswidyawan

Samuel Akwila

EXPLORATORY **DATA ANALYSIS**

Data Understanding

- Feature ffp_date, first_flight_date, load_time, dan last_flight_date seharusnya merupakan tipe data datetime karena keempat feature tersebut berisi tanggal.
- Feature age sebaiknya dalam bentuk integer.
- Feature gender, work_city, work_province, work_country, age, sum_yr_1, sum_yr_2 memiliki data yang hilang.
- Tidak ada data yang duplikat

```
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   member_no         14021 non-null  int64
 1   ffp_date          14021 non-null  object
 2   first_flight_date 14021 non-null  object
 3   gender            14020 non-null  object
 4   ffp_tier          14021 non-null  int64
 5   work_city         13562 non-null  object
 6   work_province     13382 non-null  object
 7   work_country      14017 non-null  object
 8   age               13952 non-null  float64
 9   load_time         14021 non-null  object
 10  flight_count      14021 non-null  int64
 11  bp_sum            14021 non-null  int64
 12  sum_yr_1          14018 non-null  float64
 13  sum_yr_2          14021 non-null  float64
 14  seg_km_sum        14021 non-null  int64
 15  last_flight_date  14021 non-null  object
 16  last_to_end       14020 non-null  float64
 17  avg_interval      14020 non-null  float64
 18  max_interval      14020 non-null  float64
 19  exchange_count    14020 non-null  float64
 20  avg_discount      14020 non-null  float64
 21  points_sum        14020 non-null  float64
 22  point_notflight   14020 non-null  float64
dtypes: float64(10), int64(5), object(8)
```

	feature	missing_value	percentage
0	work_province	639	4.557
1	work_city	459	3.274
2	age	69	0.492
3	work_country	4	0.029
4	sum_yr_1	3	0.021
5	point_notflight	1	0.007
6	points_sum	1	0.007
7	gender	1	0.007
8	avg_discount	1	0.007
9	exchange_count	1	0.007
10	max_interval	1	0.007
11	avg_interval	1	0.007
12	last_to_end	1	0.007

EXPLORATORY **DATA ANALYSIS**

Handle Missing Value

	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

- Pada feature work_city, work_province, work_country menggunakan nilai modus untuk setiap baris yang hilang, karena data yang hilang cukup banyak.
- Pada feature gender juga menggunakan nilai modus.
- Pada feature lainnya, data yang kosong diisi dengan nilai mean.

EXPLORATORY **DATA ANALYSIS**

Categorical Data

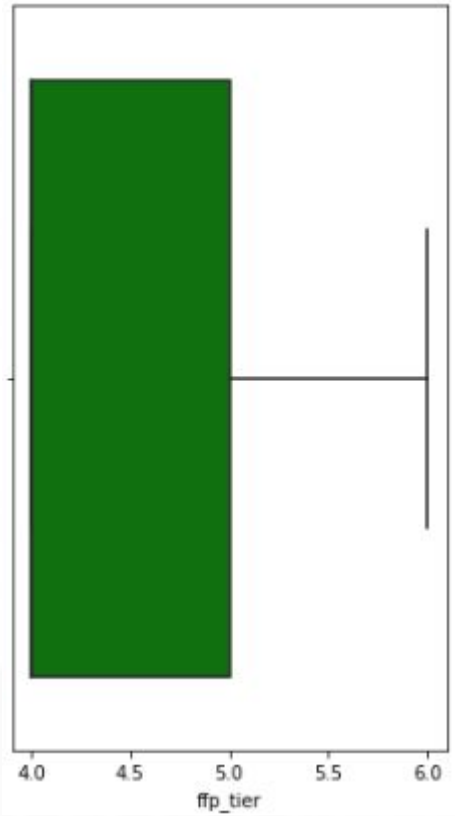
- Kolom categorical memiliki banyak sekali unique value, sehingga sulit untuk diberikan visualisasi.
- Member paling banyak adalah laki-laki.
- Mayoritas feature memiliki unique value yang besar.
- Feature load_time hanya memiliki satu nilai yaitu 2014-03-31/

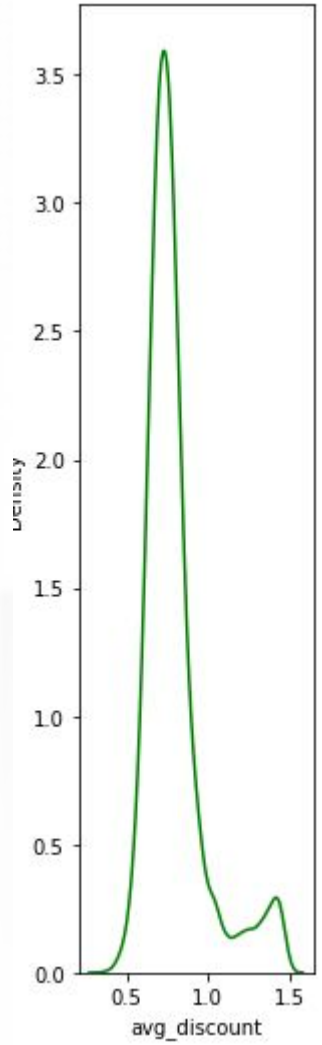
	ffp_date	first_flight_date	gender	work_city	work_province	work_country	load_time	last_flight_date
count	14021	14021	14020	13562	13382	14017	14021	14021
unique	2961	3037	2	1227	460	55	1	517
top	9/9/2005	9/9/2005	Male	guangzhou	guangdong	CN	3/31/2014	3/31/2014
freq	35	28	11538	2971	4821	12764	14021	530

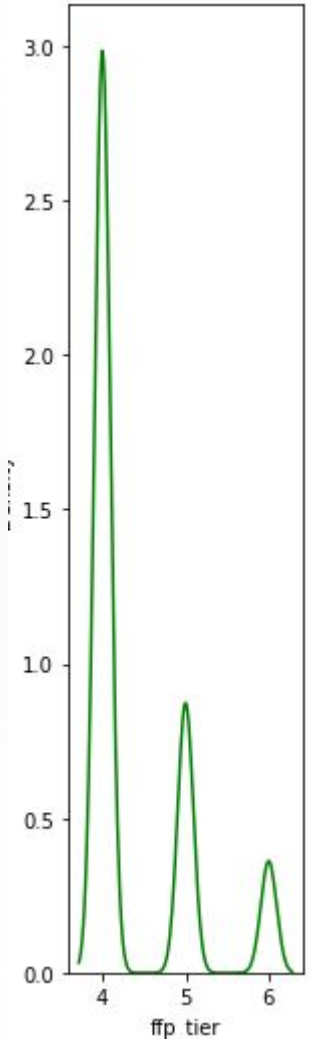
	ffp_date	first_flight_date	gender	work_city	work_province	work_country	load_time	last_flight_date
0	2006-11-02	2008-12-24	Male	.	beijing	CN	2014-03-31	2014-03-31
1	2007-02-19	2007-08-03	Male	guangzhou	beijing	CN	2014-03-31	2014-03-25
2	2007-02-01	2007-08-30	Male	.	beijing	CN	2014-03-31	2014-03-21
3	2008-08-22	2008-08-23	Male	Los Angeles	CA	US	2014-03-31	2013-12-26
4	2009-04-10	2009-04-15	Male	guiyang	guizhou	CN	2014-03-31	2014-03-27

EXPLORATORY **DATA ANALYSIS**

Numerical Data







- Mayoritas feature memiliki distribusi skew positif kecuali ffp_tier dan avg_discount
- Mayoritas feature memiliki outlier kecuali feature ffp_tier.

Feature **Engineering**

Selecting Column

Dalam project ini, kita menggunakan klustering dengan analisis RFM (Recency, Frequency, and Monetary)

```
# Lamanya member = hari data diambil - hari pertama menjadi member

df_clean['member_time'] = (df_clean['load_time'] - df_clean['ffp_date']).dt.days
```

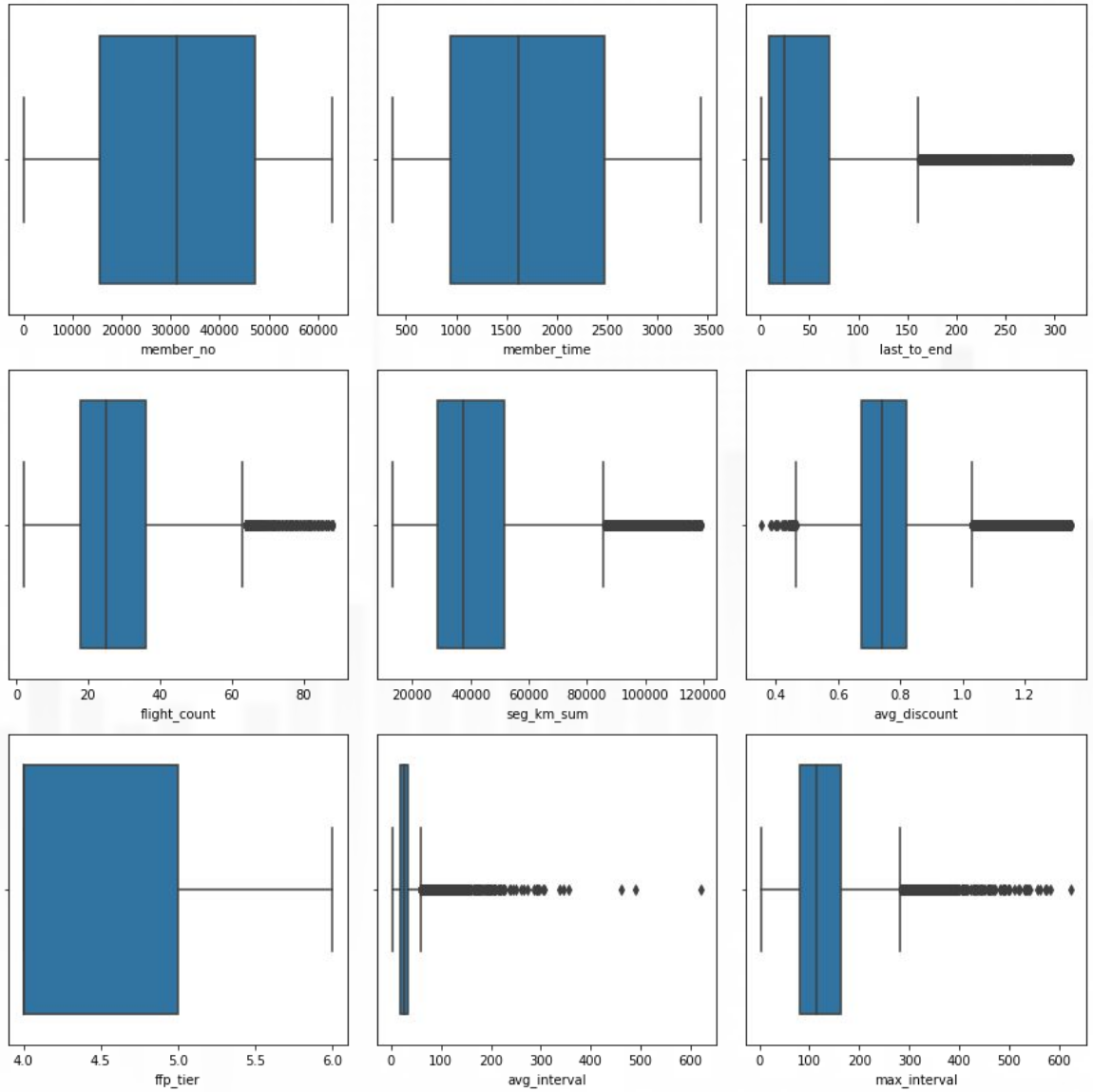
```
#Memilih kolom untuk analisis RFM

df_new = df_clean[['member_no', 'member_time', 'last_to_end', 'flight_count', 'seg_km_sum', 'avg_discount', 'ffp_tier', 'avg_int
erval', 'max_interval']].copy()
```

	member_no	member_time	last_to_end	flight_count	seg_km_sum	avg_discount	ffp_tier	avg_interval	max_interval
count	14021.000000	14021.000000	14021.000000	14021.000000	14021.000000	14021.000000	14021.000000	14021.000000	14021.000000
mean	31323.849511	1734.432138	58.731669	29.979388	45694.453605	0.791839	4.378932	29.213208	128.159629
std	18194.518213	874.472772	85.980865	19.402562	28105.414013	0.187750	0.638055	23.926082	71.164296
min	2.000000	365.000000	1.000000	2.000000	11424.000000	0.353298	4.000000	2.000000	4.000000
25%	15301.000000	942.000000	9.000000	18.000000	28420.000000	0.681156	4.000000	17.000000	79.000000
50%	31157.000000	1627.000000	25.000000	25.000000	37474.000000	0.750000	4.000000	24.409091	113.000000
75%	47175.000000	2484.000000	74.000000	37.000000	53230.000000	0.833858	5.000000	34.052632	159.000000
max	62988.000000	3437.000000	696.000000	213.000000	580717.000000	1.500000	6.000000	622.000000	622.000000

Feature **Engineering**

Delete Outliers



- Dilakukan penghapusan outliers menggunakan metode Z Score > 3
- Terdapat penghapusan data sebanyak 1328 baris

Feature **Engineering**

Dimensional Reduction

	member_no	recency	frequency	monetary
0	38290	1388.385746	69	75942.270320
1	6088	-854.662289	65	66305.270309
2	26960	1184.404474	45	74133.270309
3	43627	-167.608200	80	62334.270321
4	37021	930.399493	82	63882.270321
...	...	...	...	...
12688	35461	1356.434879	23	-14022.729696
12689	29885	535.199319	24	-25116.729696
12690	10185	759.912318	18	-17687.729696
12691	31670	-327.141530	8	-17212.729696
12692	17887	455.040854	12	-21007.729696

- Mereduksi fitur-fitur untuk analisis RFM (Recency Frequency Monetary)
- 98% informasi fitur masih tersimpan pada pca Recency.
- 100% informasi fitur masih tersimpan pada pca Monetary.

```
#Dimension reduction 1
# Merubah 'member_time', 'last_to_end','avg_interval', 'max_interval' menjadi 'recency'.

from sklearn.decomposition import PCA
reduc1 = df_reduction[['member_time', 'last_to_end','avg_interval', 'max_interval']]
pca = PCA(n_components=1)

pca.fit(reduc1)
pcs1 = pca.transform(reduc1)

data_pca1 = pd.DataFrame(data = pcs1, columns = ['recency'])
df_pca1 = df_reduction.join(data_pca1)

pca.explained_variance_ratio_
```

```
array([0.98816967])
```

```
#Dimension reduction 2
#Merubah 'seg_km_sum', 'avg_discount', 'ffp_tier' menjadi 'monetary'.

reduc2 = df_reduction[['seg_km_sum', 'avg_discount', 'ffp_tier']]
pca = PCA(n_components=1)

pca.fit(reduc2)
pcs2 = pca.transform(reduc2)

data_pca2 = pd.DataFrame(data = pcs2, columns = ['monetary'])
df_pca2 = df_pca1.join(data_pca2)

pca.explained_variance_ratio_
```

```
array([1.])
```

Feature **Engineering**

Normalisasi Data

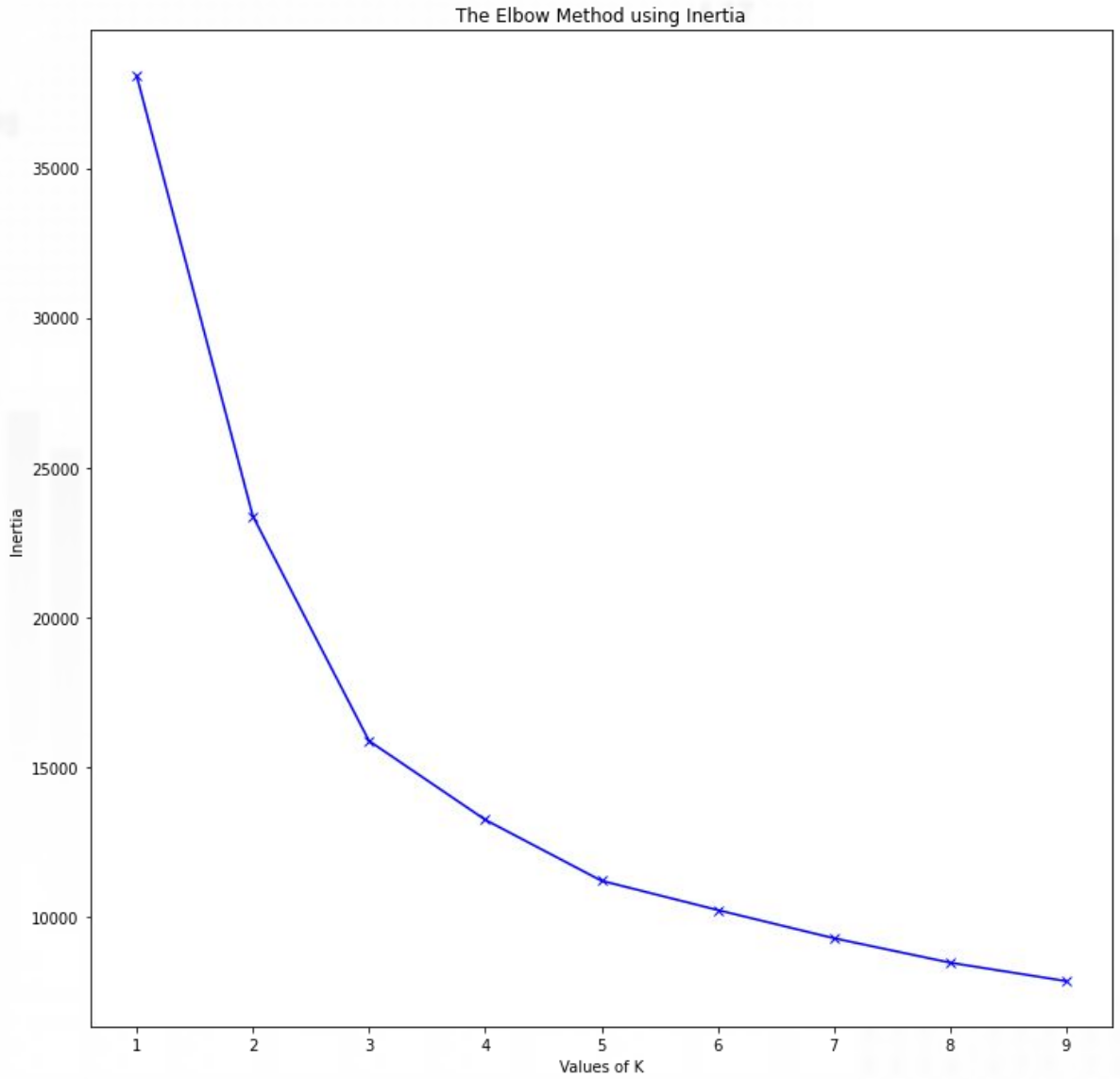
	member_no	recency	frequency	monetary
0	38290	1.594702	2.602777	3.914554
1	6088	-0.981667	2.344573	3.417801
2	26960	1.360409	1.053554	3.821307
3	43627	-0.192515	3.312838	3.213110
4	37021	1.068659	3.441939	3.292904
...	...	...	...	...
12688	35461	1.558004	-0.366567	-0.722822
12689	29885	0.614731	-0.302016	-1.294678
12690	10185	0.872837	-0.689322	-0.911740
12691	31670	-0.375755	-1.334831	-0.887255
12692	17887	0.522661	-1.076627	-1.082874

Normalisasi data dilakukan dengan menggunakan StandardScaler.

Modeling

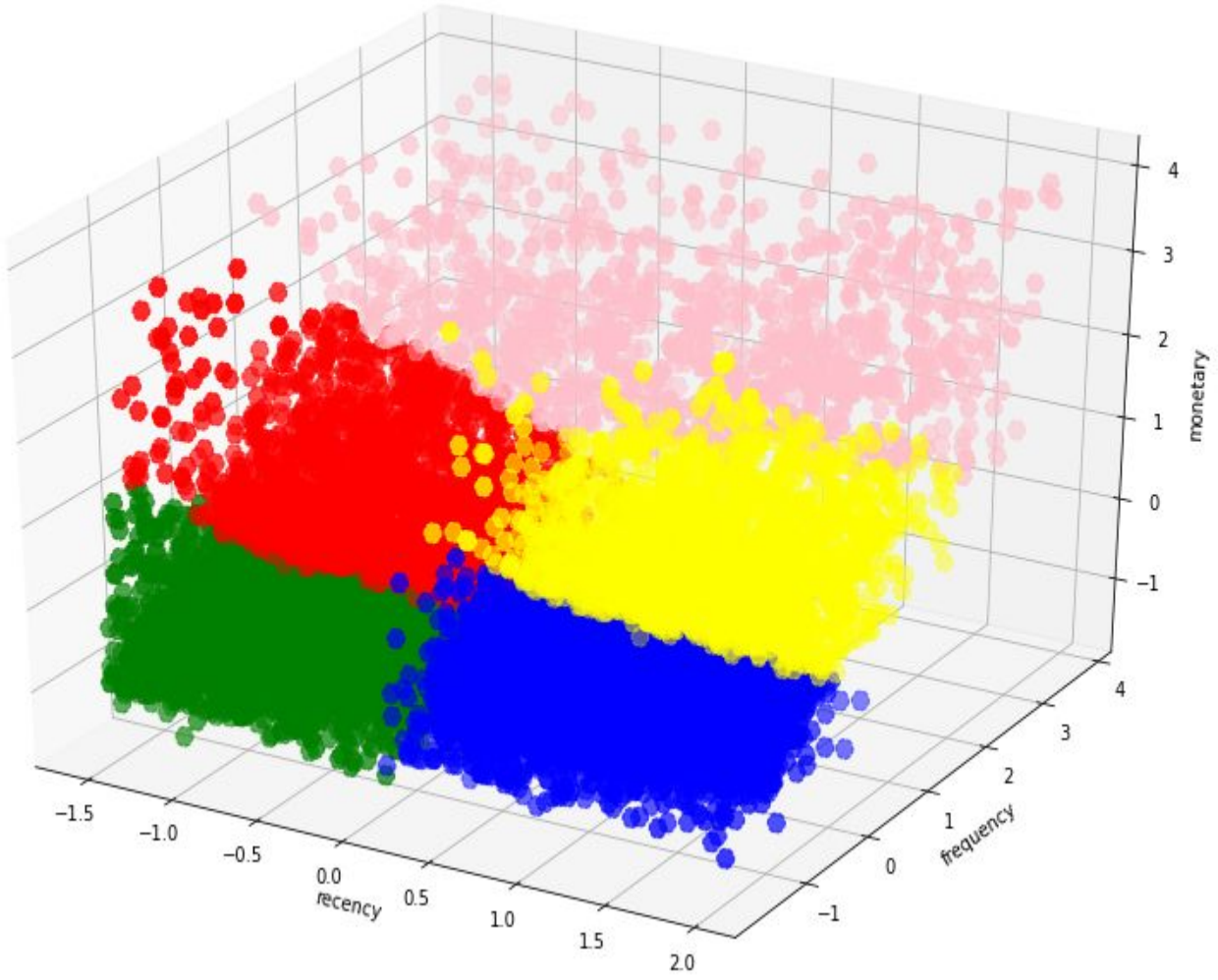
Best n Cluster

Jumlah cluster yang optimum ada di antara 4 karena setelahnya penurunan nilainya sudah tidak signifikan.



Modeling

Visualize Cluster



Modeling

Penggabungan Dataset

```
segment_member = df_norm.merge(rfm,
                               on = ['recency', 'frequency', 'monetary'],
                               how ='outer')
segment_member.drop(columns=(['recency', 'frequency', 'monetary']), inplace=True)
segment_member.drop_duplicates(inplace=True)
segment_member
```

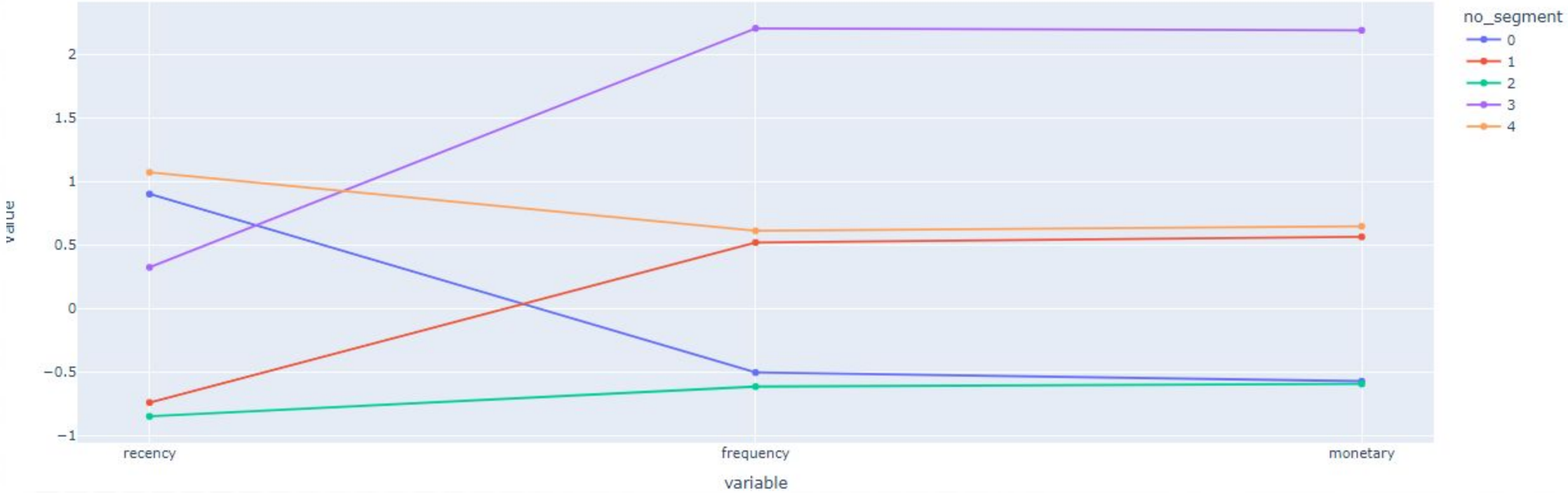
```
#Menggabungkan dataset
airline_segment = df_clean.merge(segment_member,
                               on ='member_no',
                               how ='outer')
airline_segment['segment'] = airline_segment['segment'].fillna(5)
airline_segment
```

	segment	member_no
0	0.0	3415
1	1.0	2127
2	2.0	4386
3	3.0	1009
4	4.0	1756
5	5.0	1328

- **Menggabungkan hasil segmentasi ke dataset awal**
- **Menggolongkan outlier menjadi segment 5**
- **Melihat jumlah konsumen per segment.**

Intrepertasi

Snake Plot



- Segment 0 (Potential Churn/Low Value Customer)
- Segment 1 (Potential Customer)
- Segment 2 (New Customer)
- Segment 3 (High Value Customer)
- Segment 4 (General Customer)
- Segment 5 (Outlier)

Rekomendasi

- **Segment 0 (Potential Churn/Low Value Customer)**

Memberikan personalisasi marketing campaign terkait dengan produk/tiket penerbangan yang sama pada saat terakhir kali konsumen bertransaksi.

- **Segment 1 (Potential Customer)**

Memberikan informasi loyalty program kepada konsumen jenis ini. Dengan total nilai dan frekuensi pembelian yang sedang, konsumen segmen ini cocok untuk diberikan program diskon serta informasi benefit lainnya agar konsumen ini tetap loyal kepada perusahaan.

- **Segment 2 (New Customer)**

First impression merupakan hal yang penting bagi konsumen baru. Memberikan email welcome offer, tips menggunakan produk-produk penerbangan dan informasi yang membantu bagi konsumen baru merupakan beberapa langkah yang bisa dilakukan.

- **Segment 3 (High Value Customer)**

Fokus untuk memberikan informasi produk baru dan loyalty program secara berkala. Karena konsumen ini memiliki kebiasaan membeli dengan frekuensi serta nilai yang tinggi, maka tidak perlu diberikan diskon yang tinggi. Tetapi lebih berfokus kepada pemberian informasi terkait produk premium dan produk baru dari perusahaan.

- **Segment 4 (General Customer)**

Fokus dalam meyakinkan konsumen untuk melakukan pembelian produk lagi. Dengan waktu pembelian terakhir yang cukup lama, marketing campaign berupa follow up email, penawaran baru, dan berbagai diskon merupakan langkah yang bisa dilakukan untuk menargetkan konsumen ini.

- **Segment 5 (Outlier)**

Perlu dilakukan analisis lebih lanjut kepada konsumen dengan segmen ini.

Thank **You**