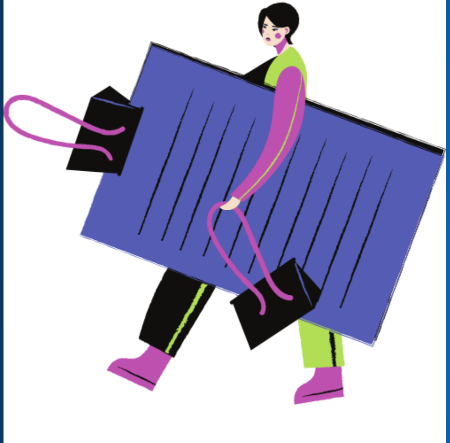
Individual Assignment

Data Cleaning, EDA and Clustering Python

Week 6&7, March 2023
Ichsan Maulana - Section Paris - Team 1

RevoU FSDA 2023

Brief



You are a data analyst in an indonesian startup which makes an application that you can buy and sell mutual funds (indonesian : reksadana). The mutual funds consist of : stocks mutual funds (pasar uang), and mixed-investment mutual funds (campuran). As the name suggests, the types reflect the underlying investment made by the fund manager for the mutual fund

Business Problem

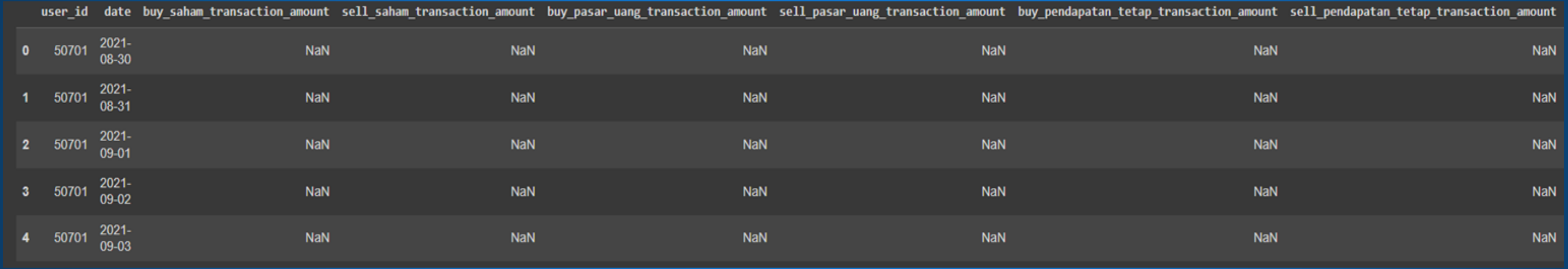
What kind of thematic campaigns that we can recommend to the marketing team for the next month while marketing team wants to create a campaign based on user preference, so you as an analyst will be tasked to create a segmentation for thematic campaign and give recommendations on the themes on each campaign

Important thing :
1. Problem Statement
2. Data Cleaning
3. EDA

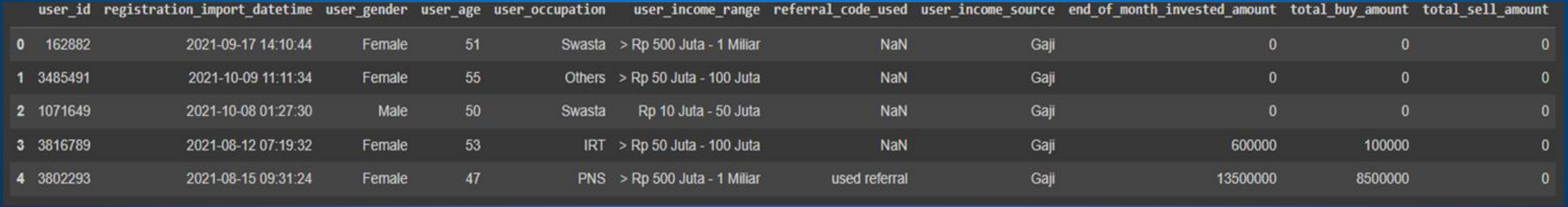
Note : We're working python code in google collab (Milestone 1)

Data Preparation

Upload dataset to collab

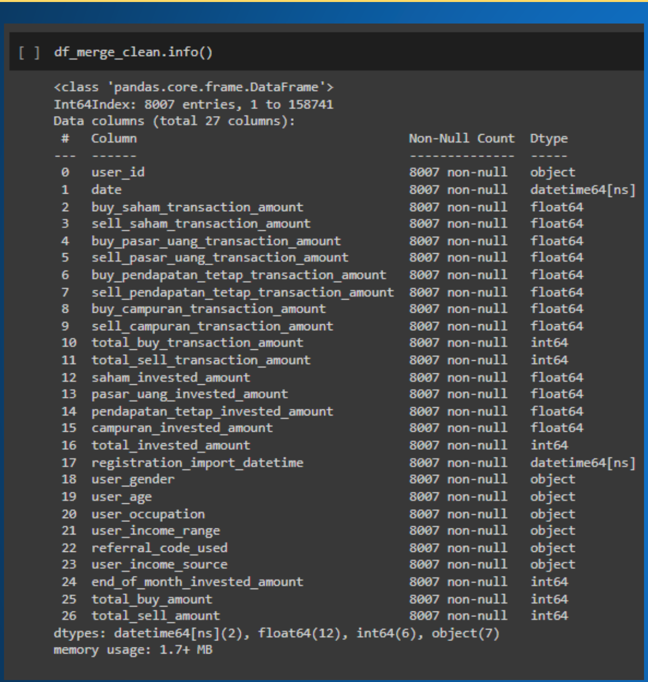


	user_id	date	buy_saham_transaction_amount	sell_saham_transaction_amount	buy_pasar_uang_transaction_amount	sell_pasar_uang_transaction_amount	buy_pendapatan_tetap_transaction_amount	sell_pendapatan_tetap_transaction_amount
0	50701	2021-08-30	NaN	NaN	NaN	NaN	NaN	NaN
1	50701	2021-08-31	NaN	NaN	NaN	NaN	NaN	NaN
2	50701	2021-09-01	NaN	NaN	NaN	NaN	NaN	NaN
3	50701	2021-09-02	NaN	NaN	NaN	NaN	NaN	NaN
4	50701	2021-09-03	NaN	NaN	NaN	NaN	NaN	NaN



	user_id	registration_import_datetime	user_gender	user_age	user_occupation	user_income_range	referral_code_used	user_income_source	end_of_month_invested_amount	total_buy_amount	total_sell_amount
0	162882	2021-09-17 14:10:44	Female	51	Swasta	> Rp 500 Juta - 1 Miliar	NaN	Gaji	0	0	0
1	3485491	2021-10-09 11:11:34	Female	55	Others	> Rp 50 Juta - 100 Juta	NaN	Gaji	0	0	0
2	1071649	2021-10-08 01:27:30	Male	50	Swasta	Rp 10 Juta - 50 Juta	NaN	Gaji	0	0	0
3	3816789	2021-08-12 07:19:32	Female	53	IRT	> Rp 50 Juta - 100 Juta	NaN	Gaji	600000	100000	0
4	3802293	2021-08-15 09:31:24	Female	47	PNS	> Rp 500 Juta - 1 Miliar	used referral	Gaji	13500000	8500000	0

Data Cleaning Overview



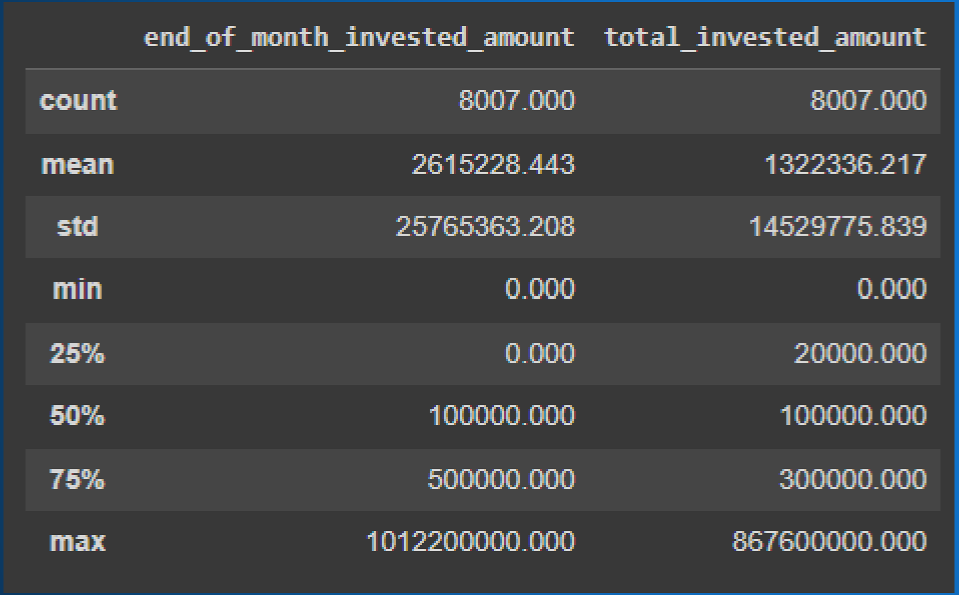
```
[ ]  df_merge_clean.info()

     <class 'pandas.core.frame.DataFrame'>
     Int64Index: 8007 entries, 1 to 158741
     Data columns (total 27 columns):
      #   Column                                  Non-Null Count  Dtype
     ---  ------                                  --------------  -----
      0   user_id                                 8007 non-null   object
      1   date                                    8007 non-null   datetime64[ns]
      2   buy_saham_transaction_amount            8007 non-null   float64
      3   sell_saham_transaction_amount           8007 non-null   float64
      4   buy_pasar_uang_transaction_amount       8007 non-null   float64
      5   sell_pasar_uang_transaction_amount      8007 non-null   float64
      6   buy_pendapatan_tetap_transaction_amount 8007 non-null   float64
      7   sell_pendapatan_tetap_transaction_amount 8007 non-null  float64
      8   buy_campuran_transaction_amount         8007 non-null   float64
      9   sell_campuran_transaction_amount        8007 non-null   float64
      10  total_buy_transaction_amount            8007 non-null   int64
      11  total_sell_transaction_amount           8007 non-null   int64
      12  saham_invested_amount                   8007 non-null   float64
      13  pasar_uang_invested_amount              8007 non-null   float64
      14  pendapatan_tetap_invested_amount        8007 non-null   float64
      15  campuran_invested_amount                8007 non-null   float64
      16  total_invested_amount                   8007 non-null   int64
      17  registration_import_datetime            8007 non-null   datetime64[ns]
      18  user_gender                             8007 non-null   object
      19  user_age                                8007 non-null   object
      20  user_occupation                         8007 non-null   object
      21  user_income_range                       8007 non-null   object
      22  referral_code_used                      8007 non-null   object
      23  user_income_source                      8007 non-null   object
      24  end_of_month_invested_amount            8007 non-null   int64
      25  total_buy_amount                        8007 non-null   int64
      26  total_sell_amount                       8007 non-null   int64
     dtypes: datetime64[ns](2), float64(12), int64(6), object(7)
     memory usage: 1.7+ MB
```

Data Cleaning

- Check data type
- Null values treatment
- Data type treatment
- Duplicated treatment
- Merged dataset

RevoU FSDA 2023

Descriptive Statistics

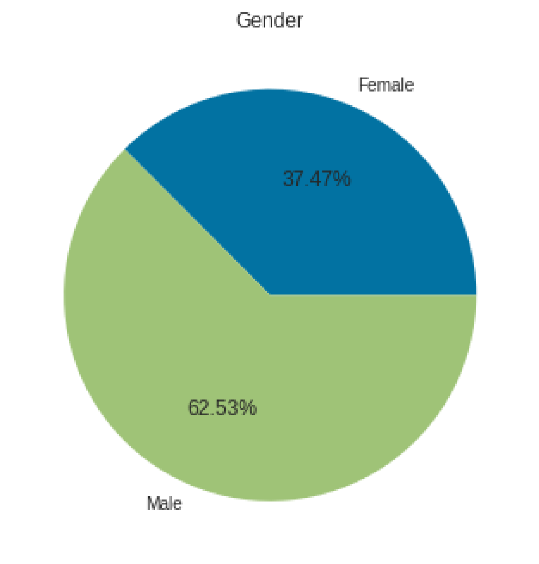


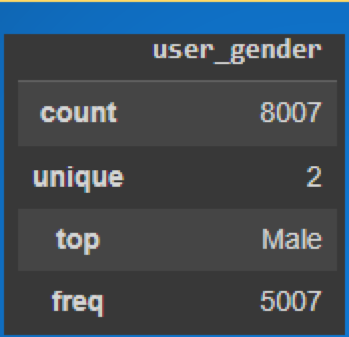
	end_of_month_invested_amount	total_invested_amount
count	8007.000	8007.000
mean	2615228.443	1322336.217
std	25765363.208	14529775.839
min	0.000	0.000
25%	0.000	20000.000
50%	100000.000	100000.000
75%	500000.000	300000.000
max	1012200000.000	867600000.000

- Average end of month invested amount is 2615228.443 million
- Average of total invested amount is 1322336.217 million
- Both minimum value is 0

0 value may be cause of users is sell all of asset or user just joined reksadana

User Gender Population

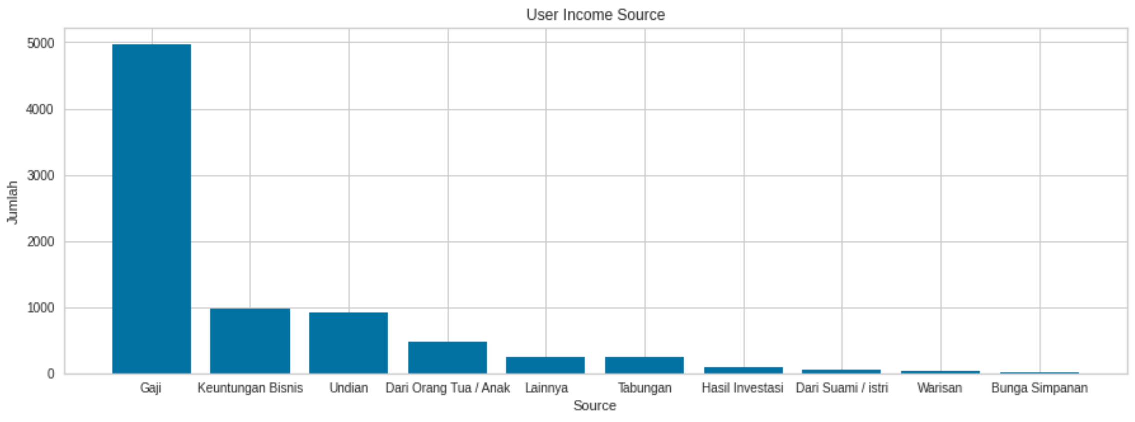




	user_gender
count	8007
unique	2
top	Male
freq	5007

- Male users is dominate the population with 62.53% or 5007 users
- Female users followed by 34.47% or 3000 users

User Income Source



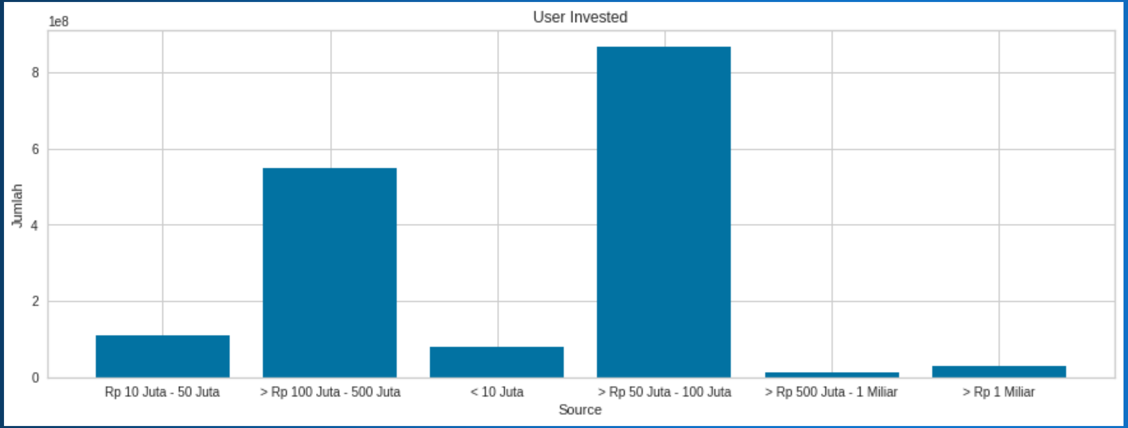
From user income source, we got top 3 income source were invest is

- Gaji
- Keuntungan Bisnis
- Undian

We can assume 4972 or more than 50% of population, working people is the most amount of invest,. Might the workers want to saving their balance by invest

RevoU FSDA 2023

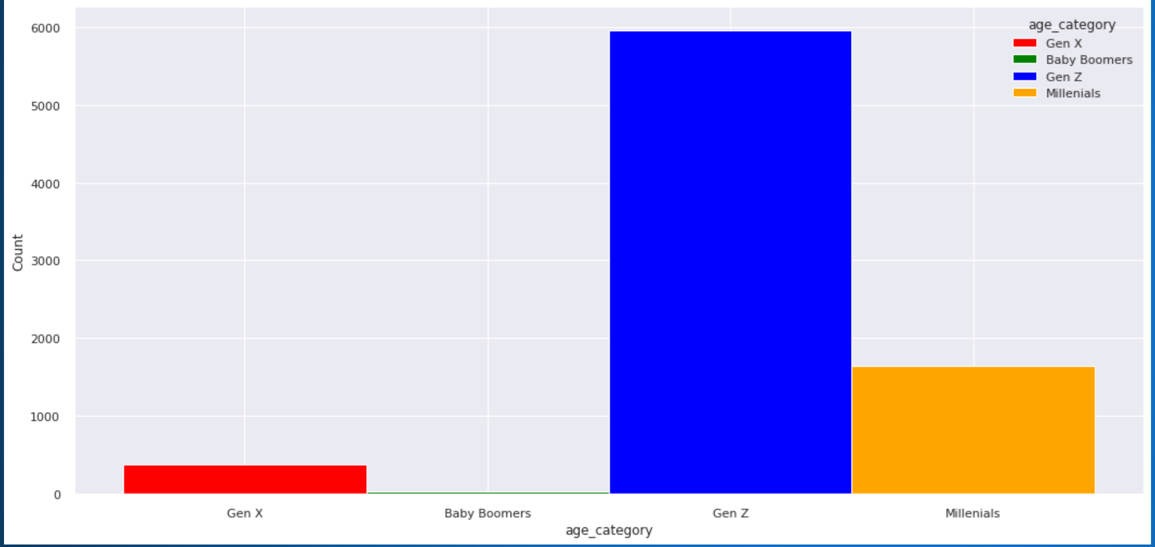
User Income Range



From user income range, we got top 3 income range were invest is

- 50 Juta - 100 Juta
- >100 juta - 500 juta
- 10 juta - 50 juta

User Age

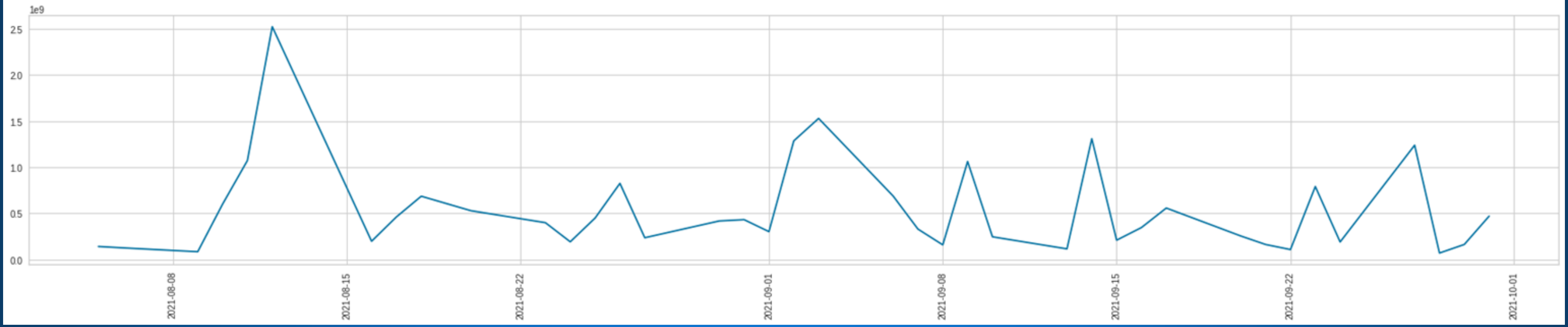


From user age, we got top 4 category by

- Gen Z
- Baby Boomers
- Millenials
- Gen X

From this top 4, we can assume Gen Z and Millenials customers are aware of investation and customers with Baby Boomers we can classified as wealthy retirees who make money by invest.

Trend on end of month invested amount



At 3rd week of august 2021, we can see from the trend line the transactions is rising and tend spiking. We should check what happen at that date if any campaign happen we can deep dive the data and find out the customers activities.

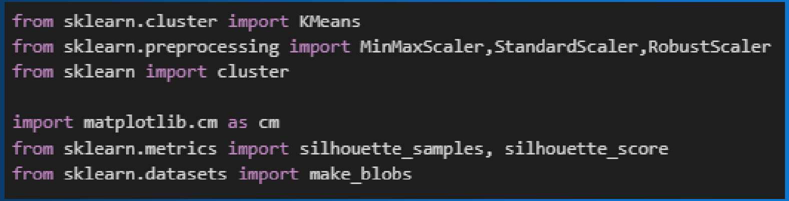
Milestone 2
Customer Segmentation

Ichsan Maulana - Section Paris - Team 1

RevoU FSDA 2023

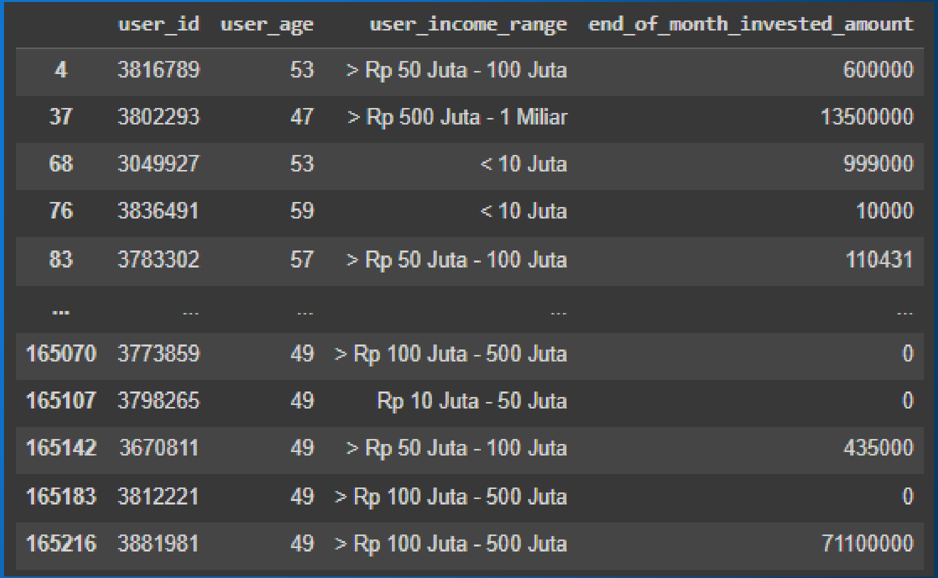
Preparation

Before step up to Clustering, we do prepare the data



```
from sklearn.cluster import KMeans
from sklearn.preprocessing import MinMaxScaler,StandardScaler,RobustScaler
from sklearn import cluster

import matplotlib.cm as cm
from sklearn.metrics import silhouette_samples, silhouette_score
from sklearn.datasets import make_blobs
```

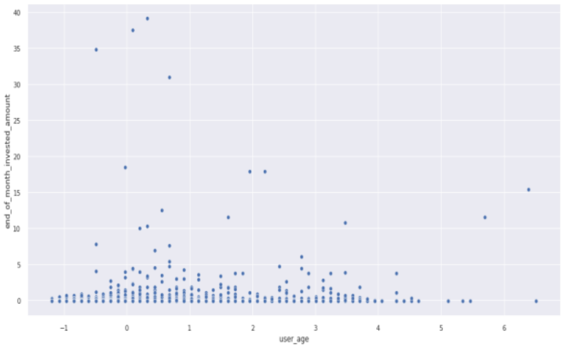
We do change user income range as income range Level 1 - Level 5 and convert to income category, so we can aggregate or calculate the variables.



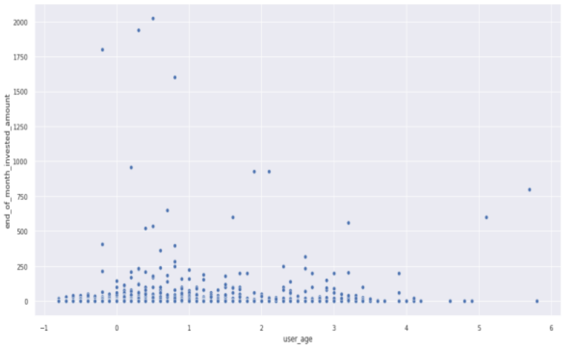
	user_id	user_age	user_income_range	end_of_month_invested_amount
4	3816789	53	> Rp 50 Juta - 100 Juta	600000
37	3802293	47	> Rp 500 Juta - 1 Miliar	13500000
68	3049927	53	< 10 Juta	999000
76	3836491	59	< 10 Juta	10000
83	3783302	57	> Rp 50 Juta - 100 Juta	110431
...	...	...	...	...
165070	3773859	49	> Rp 100 Juta - 500 Juta	0
165107	3798265	49	Rp 10 Juta - 50 Juta	0
165142	3670811	49	> Rp 50 Juta - 100 Juta	435000
165183	3812221	49	> Rp 100 Juta - 500 Juta	0
165216	3881981	49	> Rp 100 Juta - 500 Juta	71100000

Normalized Data by Scaling

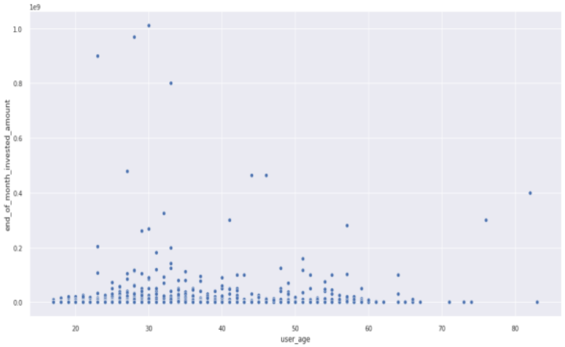
Standard Scaling



Robust Scaling

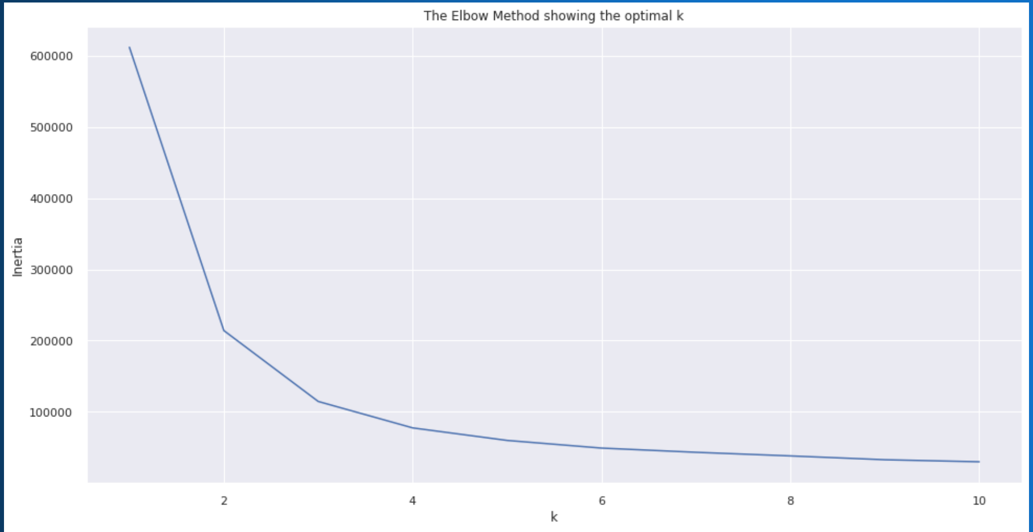


MinMax Scaling



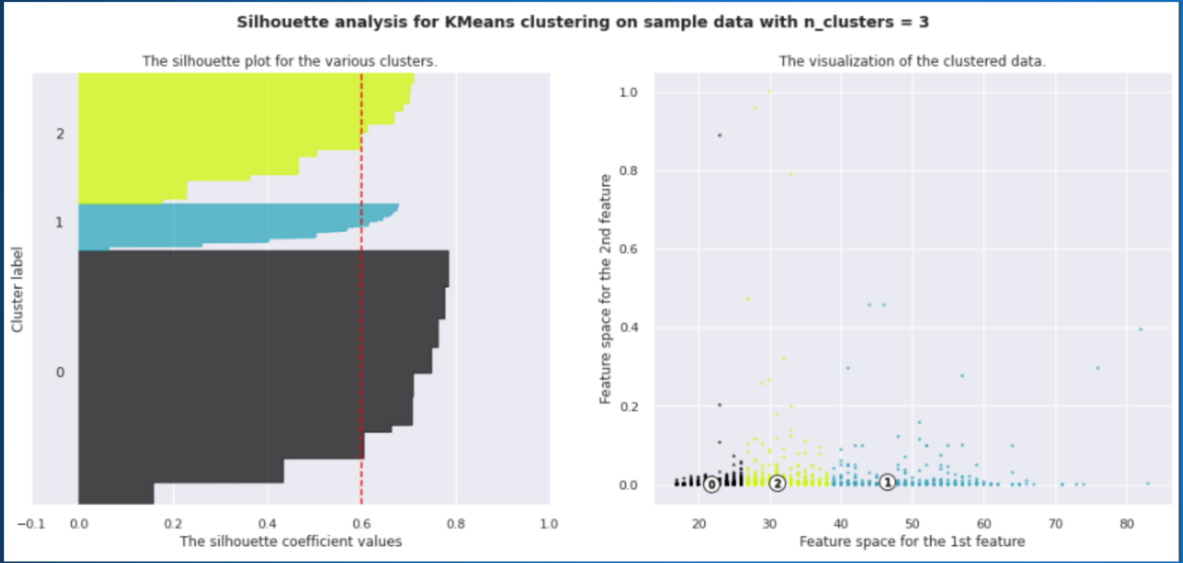
By do normalized with standard, robust and min max scaling we can see the data distribution before we step up to elbow method.

Elbow Method



Looks like elbow forms at 3, 4 and 5. Above 7 is start to show similarity of the clusters

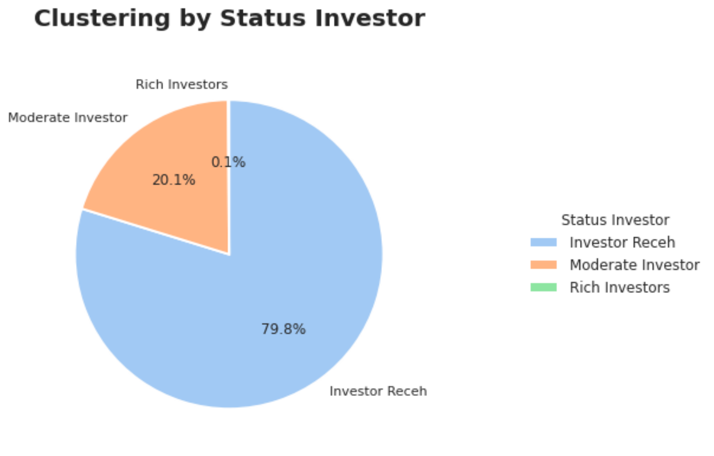
Silhouette Method



Silhouette score shows 0,6. This score is 2nd highest with n_clusters, we choose 3 clusters because population at 3 clusters is enough to describe population we have.

RevoU FSDA 2023

Cluster Population (users)



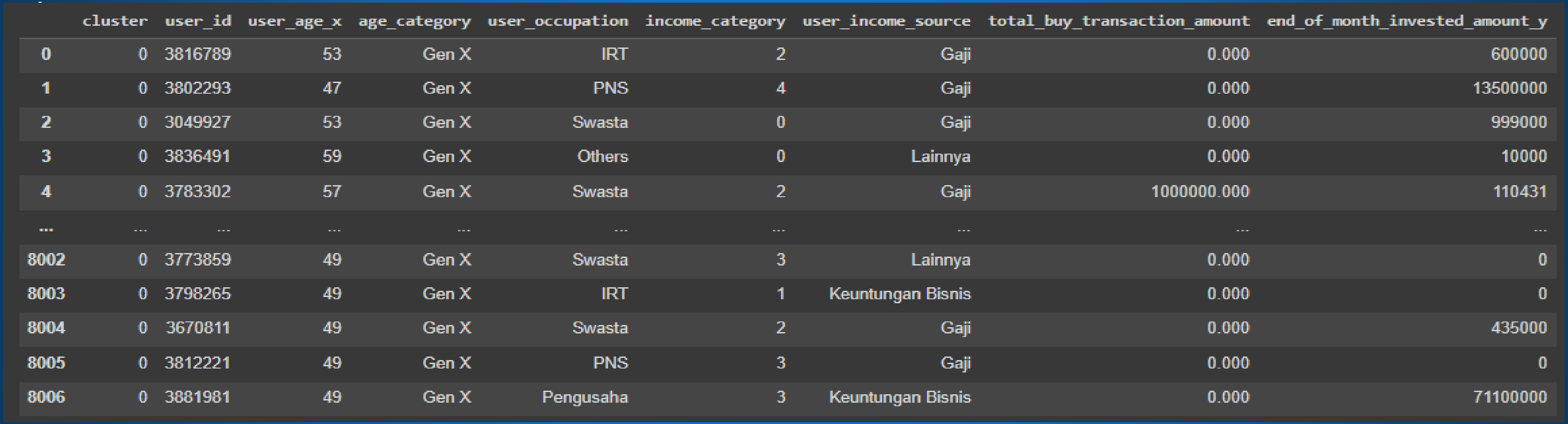
We got 3 clusters

- Cluster 0 (Investor Receh) is dominating by 79.42% or 6359 users population.
- Cluster 1 (Moderate Investor) by 20.48% or 1640 users.
- Cluster 2 (Rich Investor) by 0.10% or 8 users.

Cluster 2 seems like outlier but we can see an interesting facts later.

Cluster Intepretation

After cluster is decided, we can interpret each

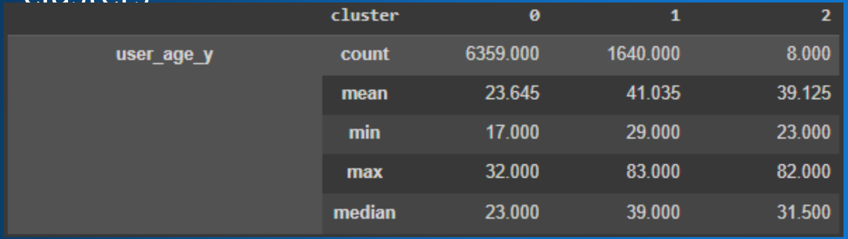


	cluster	user_id	user_age_x	age_category	user_occupation	income_category	user_income_source	total_buy_transaction_amount	end_of_month_invested_amount_y
0	0	3816789	53	Gen X	IRT	2	Gaji	0.000	600000
1	0	3802293	47	Gen X	PNS	4	Gaji	0.000	13500000
2	0	3049927	53	Gen X	Swasta	0	Gaji	0.000	999000
3	0	3836491	59	Gen X	Others	0	Lainnya	0.000	10000
4	0	3783302	57	Gen X	Swasta	2	Gaji	1000000.000	110431
...	...	...	...	...	...	...	...	...	...
8002	0	3773859	49	Gen X	Swasta	3	Lainnya	0.000	0
8003	0	3798265	49	Gen X	IRT	1	Keuntungan Bisnis	0.000	0
8004	0	3670811	49	Gen X	Swasta	2	Gaji	0.000	435000
8005	0	3812221	49	Gen X	PNS	3	Gaji	0.000	0
8006	0	3881981	49	Gen X	Pengusaha	3	Keuntungan Bisnis	0.000	71100000

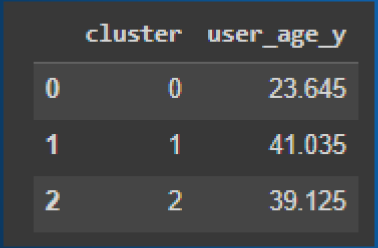
We got the data we want to describe and cluster we already generate

User Age

After cluster is decided, we can interpret each clusters



	cluster	0	1	2
user_age_y	count	6359.000	1640.000	8.000
	mean	23.645	41.035	39.125
	min	17.000	29.000	23.000
	max	32.000	83.000	82.000
	median	23.000	39.000	31.500



	cluster	user_age_y
0	0	23.645
1	1	41.035
2	2	39.125

Cluster 0

- Dominate the population by 6359 data
- Fill by population by age 17 - 32 with average 23 years old (Gen Z)

Cluster 1

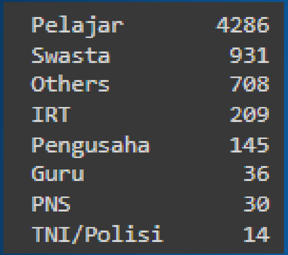
- Total 1640 customer in cluster 1
- Fill by population by age 29 - 83 with average arround 41 years old (Millenials)

Cluster 2

- Only 8 customer in cluster 2
- Fill by population by age 23 - 82 with average 39 years old (Millenials)

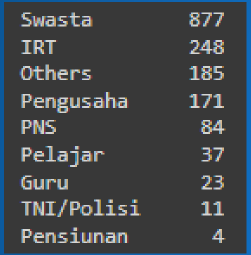
User Occupation

Cluster 0



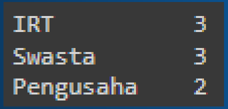
```
Pelajar        4286
Swasta          931
Others          708
IRT             209
Pengusaha       145
Guru             36
PNS              30
TNI/Polisi       14
```

Cluster 1



```
Swasta         877
IRT            248
Others         185
Pengusaha      171
PNS             84
Pelajar         37
Guru            23
TNI/Polisi      11
Pensiunan        4
```

Cluster 2



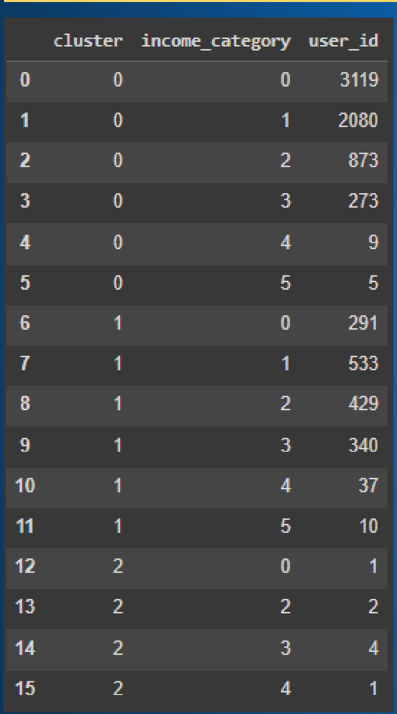
```
IRT          3
Swasta       3
Pengusaha    2
```

Cluster 1
- Dominate the population by pelajar by 4286 customers
- Swasta 931 customers
- Others 708 customers

Cluster 2
- Dominate the population by swasta by 877 customers
- IRT 248 customers
- Others 185 customers

Cluster 3
- IRT by 3 customers
- Swasta by 3 customers
- Pengusaha by 2 customers

User Income



	cluster	income_category	user_id
0	0	0	3119
1	0	1	2080
2	0	2	873
3	0	3	273
4	0	4	9
5	0	5	5
6	1	0	291
7	1	1	533
8	1	2	429
9	1	3	340
10	1	4	37
11	1	5	10
12	2	0	1
13	2	2	2
14	2	3	4
15	2	4	1

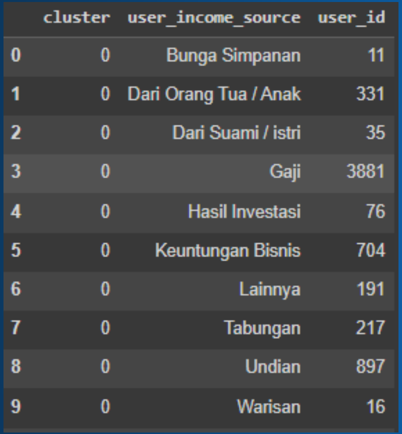
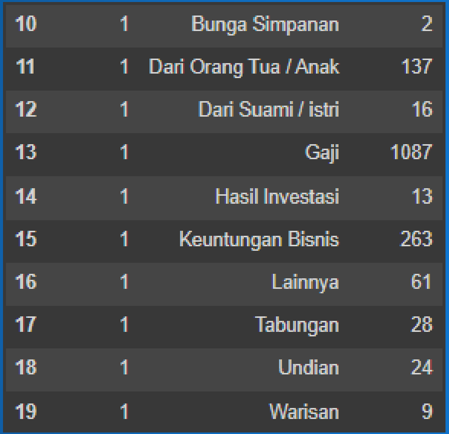
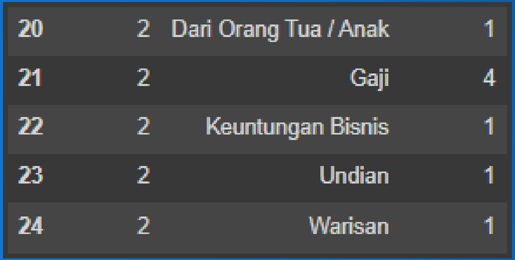
Index	Range
0	<10 juta
1	10 - 50 juta
2	50 - 100 juta
3	100 - 500 juta
4	500 juta - 1 Milliar
5	>1 Milliar

1. At cluster 0 we got top 3 user income range
   - 3119 customers with income <10 juta
   - 2080 customers with income 10 - 50 juta
   - 873 customers with income 50 - 100 juta

1. At cluster 1 we got top 3 income range
   - 291 customers with income <10 juta
   - 533 customers with income 10 - 50 juta
   - 429 customers with income 50 - 100 juta

1. At cluster 2 we got a little population but there is 4 customers with income range 100 - 500 juta

User Income Source

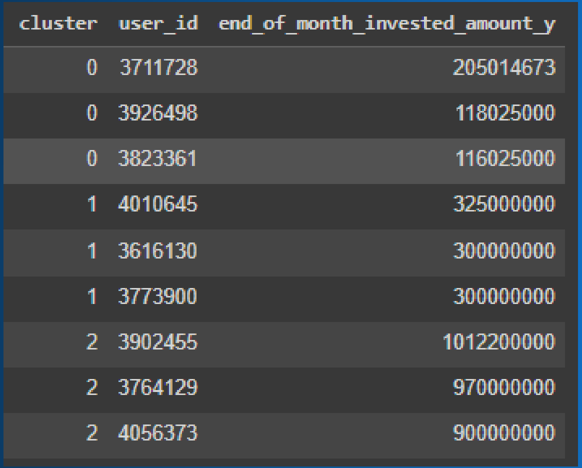




	cluster	user_income_source	user_id
0	0	Bunga Simpanan	11
1	0	Dari Orang Tua / Anak	331
2	0	Dari Suami / istri	35
3	0	Gaji	3881
4	0	Hasil Investasi	76
5	0	Keuntungan Bisnis	704
6	0	Lainnya	191
7	0	Tabungan	217
8	0	Undian	897
9	0	Warisan	16
10	1	Bunga Simpanan	2
11	1	Dari Orang Tua / Anak	137
12	1	Dari Suami / istri	16
13	1	Gaji	1087
14	1	Hasil Investasi	13
15	1	Keuntungan Bisnis	263
16	1	Lainnya	61
17	1	Tabungan	28
18	1	Undian	24
19	1	Warisan	9
20	2	Dari Orang Tua / Anak	1
21	2	Gaji	4
22	2	Keuntungan Bisnis	1
23	2	Undian	1
24	2	Warisan	1

- At cluster 0, user income source is dominating by Gaji, 3881 users
- Cluster 1 also dominate by Gaji with 1087 users
- And cluster 2 is same with cluster 0 and 1, Gaji. But this one, has an enormous value

All clusters is have similar income source domination, we can draw the conclusion users with Gaji / salary income source is stocks savvy or the users rich wanna be.

End of Month Invested Amount



cluster	user_id	end_of_month_invested_amount_y
0	3711728	205014673
0	3926498	118025000
0	3823361	116025000
1	4010645	325000000
1	3616130	300000000
1	3773900	300000000
2	3902455	1012200000
2	3764129	970000000
2	4056373	900000000

We got top 3 users of End of Month Invested Amount and we see uniqueness here. Even the cluster 0 and cluster 1 is dominating the population, in fact end of month invested amount at cluster 3 have the greatest value

Cluster	End of Month Invested Amount (in Rupiah)
0	439.064.673
1	925.000.000
2	2.882.200.000

Business Recomendation

After all the clustering we can suggest thematic campaign for marketing division

Cluster	Campaign
0 (Investor Receh)	We can give loyalty point to this populations with minimum top up, this clusters is dominating the population so we can consider this cluster as ‘Investor Receh’.
1 (Moderate Investor)	Cluster 1 is middle up investors we can consider this cluster as ‘Moderate Investor’, we can bundle every purchasing with ecommerce voucher or household needs.
2 (Rich Investor)	We can consider the cluster 2 as ‘Rich Investors’. We can provide them with strategic bonds and some stuff like dashboard which connect to capital market and give good suggestions what’s recommendation line to invest.

Thankyou for your attention

Merci pour votre attention

Ichsan Maulana - Section Paris - Team 1

Appendix

Ichsan Maulana - Section Paris - Team 1

Dataset and Data Dictionary

Link to dataset (Users) :
https://docs.google.com/spreadsheets/d/1VC90M2jmTCyN9yotawFWn_brmbkMkTgWwH8fOSJG2NM/edit?usp=sharing

Link to dataset (Daily User Transactions) :
https://docs.google.com/spreadsheets/d/1PZDdudn15RXZznkldknu58_jwjZtaMW57SH8mNyjCZ8/edit?usp=sharing

Link to data dictionary : https://docs.google.com/spreadsheets/d/1ADOss-zoude5rSk73gMzh5fBIZbIVjL-iXRWlUGJf2s/edit?usp=sharing

Link to collab :
https://colab.research.google.com/drive/1JvSzVw4fIikzOaBLQo3O9hMNGm4Ho1J9?usp=share_link