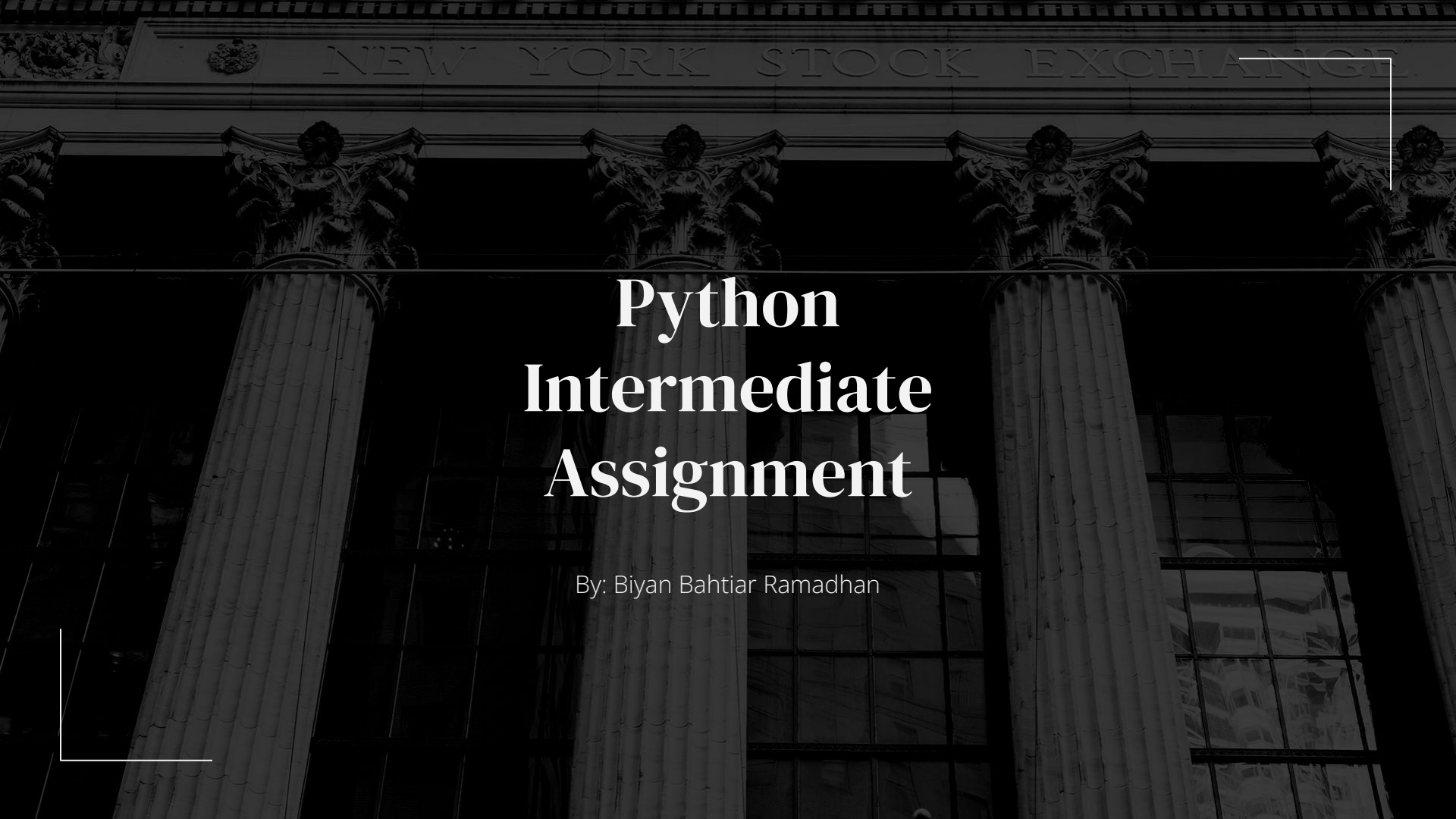


Python Intermediate Assignment

By: Biyan Bahtiar Ramadhan

Milestone 1

01 Business Background

02 Dataset Overview

03 Data Cleaning & Merging

04 Exploratory Data Analysis

05 End of Milestone 1

01 Business Background

Company:
Mutual Fund Investment Application Startup.

Objective:
Recommend **segmented thematic campaign** for next month, based on **customer preference.**

User:
Marketing Team

02 Dataset Overview

users.csv

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14712 entries, 0 to 14711
Data columns (total 11 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   user_id                         14712 non-null  int64
 1   registration_import_datetime    14712 non-null  object
 2   user_gender                     14712 non-null  object
 3   user_age                        14712 non-null  int64
 4   user_occupation                 14712 non-null  object
 5   user_income_range               14712 non-null  object
 6   referral_code_used              5604 non-null   object
 7   user_income_source              14712 non-null  object
 8   end_of_month_invested_amount    14712 non-null  int64
 9   total_buy_amount                14712 non-null  int64
 10  total_sell_amount               14712 non-null  int64
dtypes: int64(5), object(6)
```

daily_user_transaction.csv

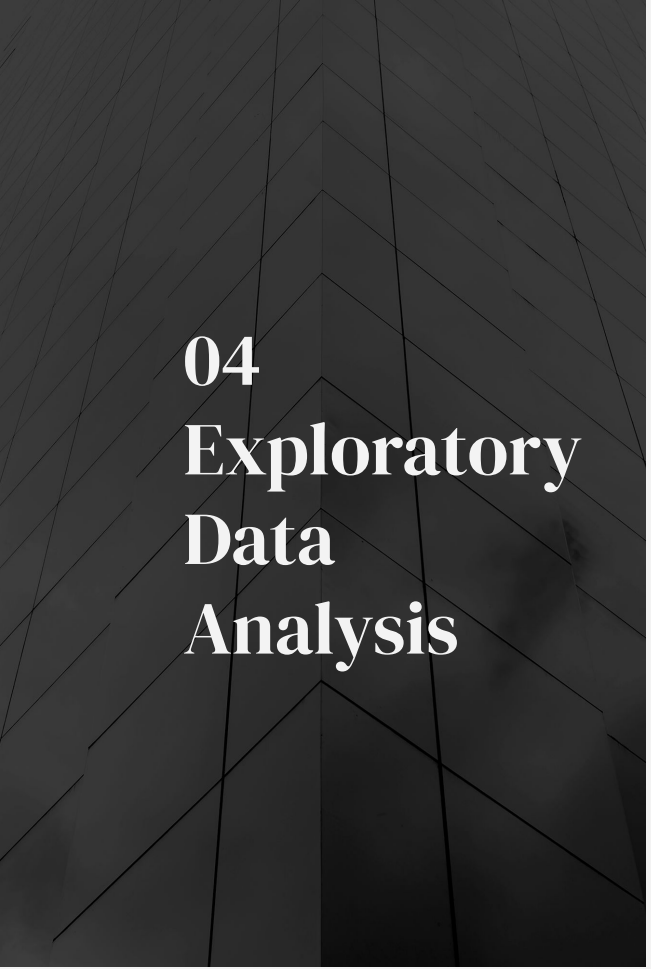
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 158811 entries, 0 to 158810
Data columns (total 17 columns):
 #   Column                                  Non-Null Count   Dtype
---  ------                                  --------------   -----
 0   user_id                                 158811 non-null  int64
 1   date                                    158811 non-null  object
 2   buy_saham_transaction_amount            99031 non-null   float64
 3   sell_saham_transaction_amount           1808 non-null    float64
 4   buy_pasar_uang_transaction_amount       122263 non-null  float64
 5   sell_pasar_uang_transaction_amount      2010 non-null    float64
 6   buy_pendapatan_tetap_transaction_amount 98916 non-null   float64
 7   sell_pendapatan_tetap_transaction_amount 1581 non-null   float64
 8   buy_campuran_transaction_amount         5072 non-null    float64
 9   sell_campuran_transaction_amount        46 non-null      float64
 10  total_buy_transaction_amount            158811 non-null  int64
 11  total_sell_transaction_amount           158811 non-null  int64
 12  saham_invested_amount                   106292 non-null  float64
 13  pasar_uang_invested_amount              131081 non-null  float64
 14  pendapatan_tetap_invested_amount        105946 non-null  float64
 15  campuran_invested_amount                5352 non-null    float64
 16  total_invested_amount                   158811 non-null  int64
dtypes: float64(12), int64(4), object(1)
memory usage: 20.6+ MB
```

03 Data Cleaning and Merging

Steps	Consideration
Cleaning key column	Both datasets has user_id, I decide to use the column as key column to be cleaned. There is no missing data in both datasets. Duplicates in daily_user_transaction is due to the multiple dates recorded.
Merging	Merging is done first to reduce the amount of cleaning needed (1 table vs 2 table). Since there is difference in unique user_id amount (14712 in users vs 8277 in daily_user_transaction), I decide to use left join
Handling Missing Data	• In referral_code_used I assign the blank cell with 'no' and change 'used referral' value with 'yes' • Missing data in 'date' is leaved as it is since it identifies user that is registered but no transaction yet. • Missing data in int/float type columns are also leaved as it is, because filling will affect the aggregation. I will fill with 0 later when I need to do segmentation.

03 Data Cleaning and Merging

Steps	Consideration
Handling duplicate, typos and change data type	• No duplicates are identified with subset user_id, registration_import_datetime and date • No typos are identified in categorical columns • Change column type: date-like column is changed to datetime64, change int64 column except user_id and user_age to float64
Outlier	Outliers are not removed and will be handled by data transformation if needed.



04 Exploratory Data Analysis

1. Total users in the dataset?

2. Total users in each category? (gender, occupation, referral code used, income, income source and per investment product)

3. General transaction trend and by investment product.

Most user are male, don't use referral code, age 20-25 and pelajar.

Total User

14712

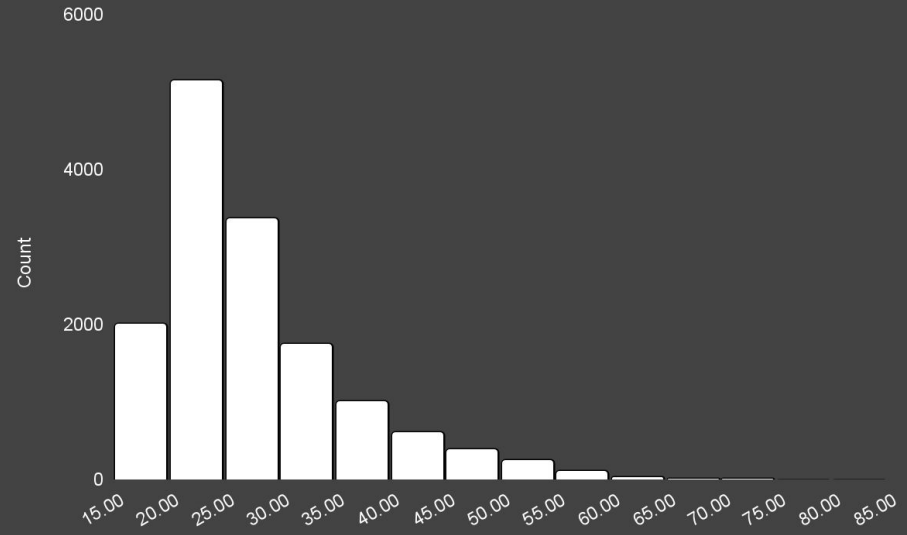
Gender

39% Female | 61% Male

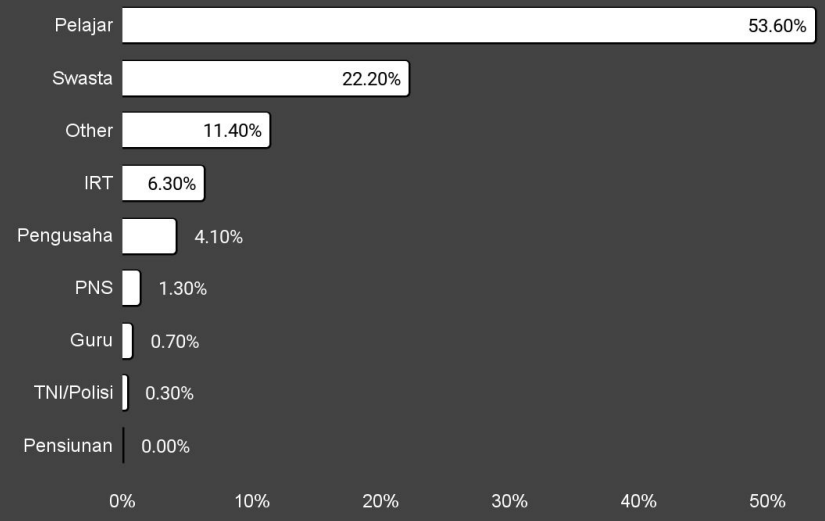
Referral Code Used

38.1% No | 61.9 % Yes

User Age

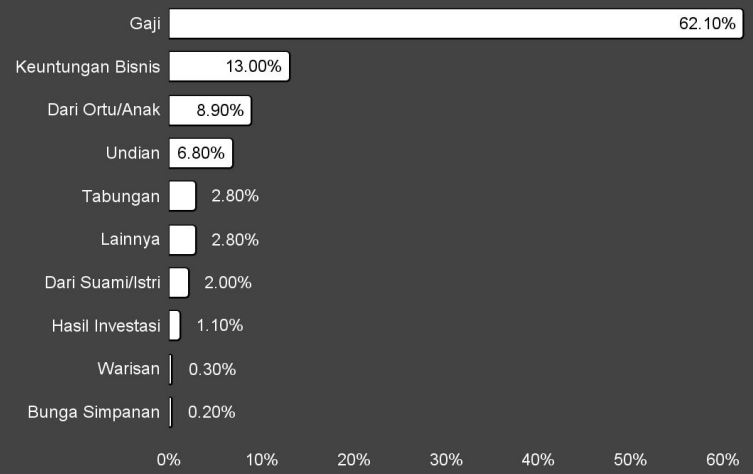


User Occupation

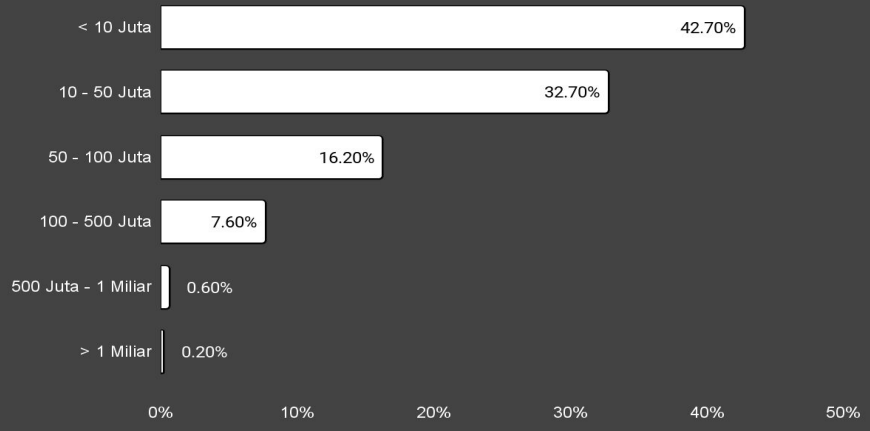


Most user have income source from **gaji**, **<10 juta** and have reksadana **pasar uang** in their portfolio.

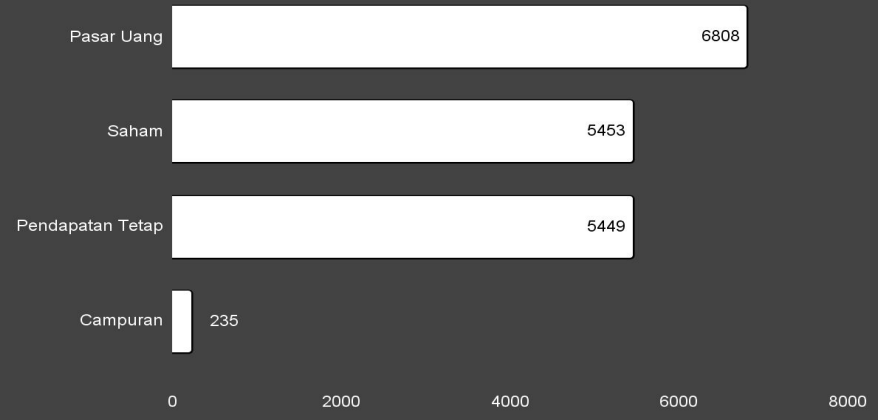
User Income Source



User Income Range

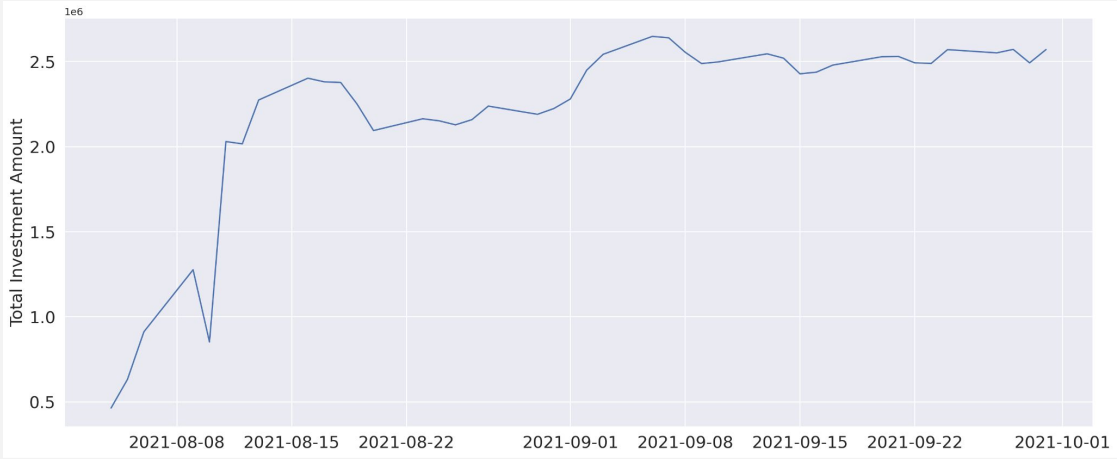


Chosen Investment Product

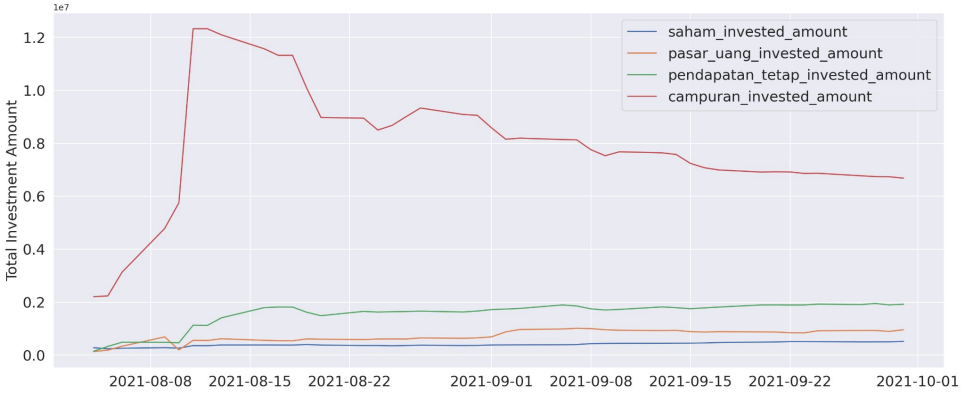


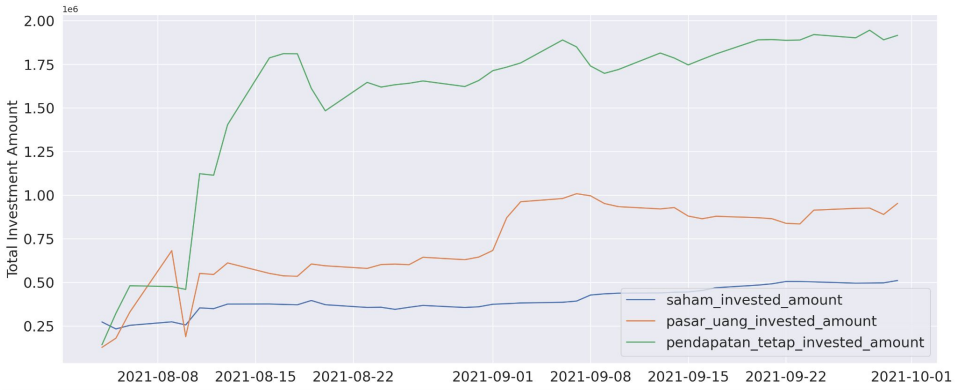
Amount invested on all product is rising over time, except reksadana campuran

Total Investment Amount Over Time



Total Investment Amount Over Time By Product






End of
Milestone 1

Milestone 2

06

Cluster Analysis Steps

07

Cluster Interpretation

08

Campaign Recommendation

06
Cluster Analysis Steps

Feature Selection & Engineering

Assign user to category based on history of transaction per product and owned investment product.

Dimensionality Reduction

PCA is used to help clustering and visualization

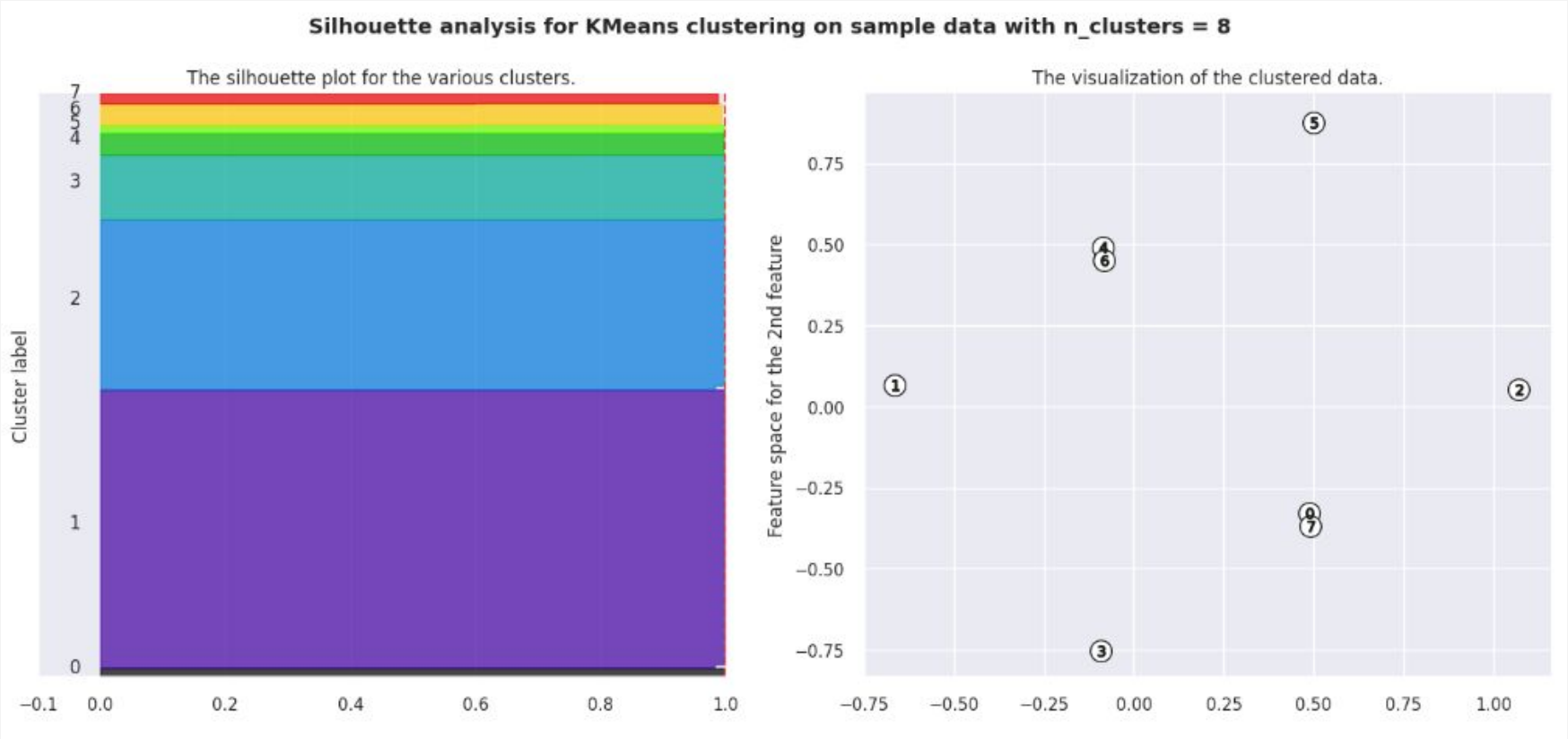
K-means Clustering

Silhouette analysis, PCA visualization and feedback from the results guide the number of cluster needed

Interpretation

Inputting PCA features in K-means result in 8 cluster

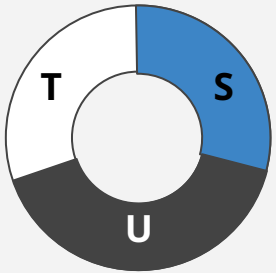
However, one cluster mix user with 0 transaction with user that invest only in Campuran. The cluster is then manually separated



07
Cluster Interpretation

0 - No Investment

6954 User
Users have no history of transaction and don't own any investment product

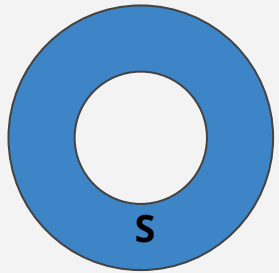


1 - The Most Invested

4294 User
All user in this cluster invest in **Pasar Uang**, **Pendapatan Tetap** and **Saham**, few of them (90) also invest in Campuran

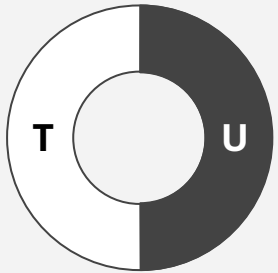
2 - Money Market Investor

1625 user
All user in this cluster invest only in **Pasar Uang**, few of them (15) also invest in Campuran



3 - Equity Investor

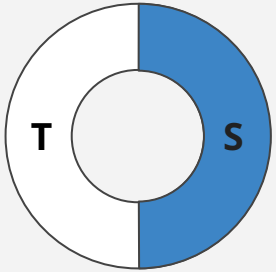
556 User
All user in this cluster invest only in **Saham**, few of them (7) also invest in Campuran



4 - Fixed Income and Money Market Investor

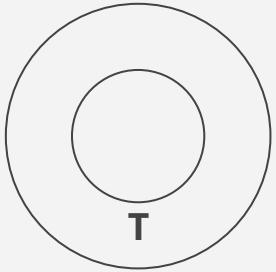
261 User
All user in this cluster invest in **Pasar Uang** and **Pendapatan Tetap,** few of them (11) also invest in Campuran

07
Cluster Interpretation



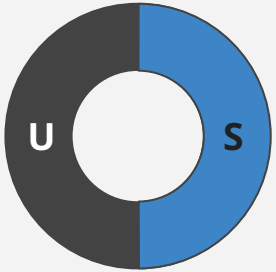
5 - Fixed Income and Equity Investor

182 User
All user in this cluster invest in **Pendapatan Tetap** and **Saham,** few of them (10) also invest in Campuran



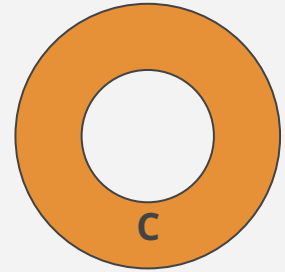
6 - Fixed Income Investor

531 user
All user in this cluster invest only in **Pendapatan Tetap**, few of them (13) also invest in Campuran



7 - Money Market and Equity Investor

222 User
All user in this cluster invest in **Pasar Uang** and **Saham,** few of them (1) also invest in Campuran



8 - Discretionary Fund Investor

87 user
All user in this cluster invest only in **Campuran**

For the one who haven’t invest yet..

Remind Them!

Pop-up message, remind them that they have the app and they haven’t invested.

Make Them Motivated!

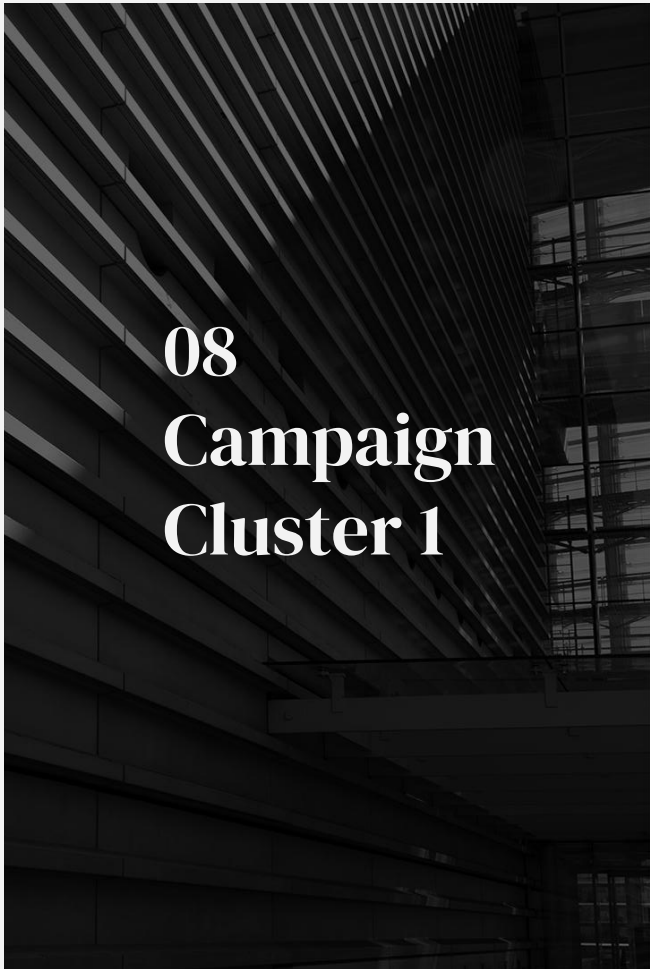
“Do you know that if you invest Rp 1,000,000 now, and you routinely invest monthly, you could get Rp 200,000,000 in ten year!! Tap to see how”

08 Campaign Cluster 0

For the most invested..

Congratulate and Motivate

Congratulation! Your investment have grown by %! Keep investing to earn more!



For money market investor..

Recommend them Pendapatan Tetap, Saham and Campuran

"Reksadana Campuran XXX give annual return 20% last year $$$ Tap to invest now!"

Motivate to Invest Regularly!

"Do you know that if you invest Rp 1,000,000 now, and you routinely invest monthly, you could get Rp 200,000,000 in ten year!! Tap to see how"

08 Campaign Cluster 2

For equity investor..

Recommend them Pendapatan Tetap, Pasar Uang and Campuran

"Reksadana saham give the highest return but you should diversify to protect your portfolio! Tap to learn how"

Motivate to Invest Regularly!

"Do you know that if you invest Rp 1,000,000 now, and you routinely invest monthly, you could get Rp 200,000,000 in ten year!! Tap to see how"

For fixed income and money market investor..

Recommend them Saham and Campuran

“You know you could get higher return if you invest in Saham and Campuran? Let’s invest there now!”

Motivate to Invest Regularly!

“Do you know that if you invest Rp 1,000,000 now, and you routinely invest monthly, you could get Rp 200,000,000 in ten year!! Tap to see how”

08 Campaign Cluster 4

For fixed income and equity investor..

Recommend them Pasar Uang and Campuran

"Reksadana Campuran XXX is less fluctuative but give spectacular return! Tap to invest now"

Motivate to Invest Regularly!

"Do you know that if you invest Rp 1,000,000 now, and you routinely invest monthly, you could get Rp 200,000,000 in ten year!! Tap to see how"

08 Campaign Cluster 5

For fixed income investor..

Recommend them Pasar Uang, Saham and Campuran

"You know you could get higher return if you invest in Saham and Campuran? Let's invest there now!"

Motivate to Invest Regularly!

"Do you know that if you invest Rp 1,000,000 now, and you routinely invest monthly, you could get Rp 200,000,000 in ten year!! Tap to see how"

08 Campaign Cluster 6

For money market and equity investor..

Recommend them Pendapatan Tetap and Campuran

"Reksadana Campuran XXX is less fluctuative but give spectacular return! Tap to invest now"

Motivate to Invest Regularly!

"Do you know that if you invest Rp 1,000,000 now, and you routinely invest monthly, you could get Rp 200,000,000 in ten year!! Tap to see how"

08 Campaign Cluster 7

For discretionary fund investor..

Recommend them Pendapatan Tetap, Saham and Pasar Uang

"To get more stable return, let's invest in Pendapatan Tetap and Pasar Uang! Tap to know more"

Motivate to Invest Regularly!

"Do you know that if you invest Rp 1,000,000 now, and you routinely invest monthly, you could get Rp 200,000,000 in ten year!! Tap to see how"

08 Campaign Cluster 8

Thanks!

Reach me out on linkedin! and email

Google Colab

Python Advanced Assignment

By: Biyan Bahtiar Ramadhan

Table of Contents

01 Business Background

Company:
Mutual Fund Investment Application Startup.

Objective:

- **Select 30% the most potential customer** to benefit from marketing campaign.
- Create projection of **how profitable the campaign is.**

User:
Marketing Team.

02 Dataset Overview

users.csv

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14712 entries, 0 to 14711
Data columns (total 11 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   user_id                         14712 non-null  int64
 1   registration_import_datetime    14712 non-null  object
 2   user_gender                     14712 non-null  object
 3   user_age                        14712 non-null  int64
 4   user_occupation                 14712 non-null  object
 5   user_income_range               14712 non-null  object
 6   referral_code_used              5604 non-null   object
 7   user_income_source              14712 non-null  object
 8   end_of_month_invested_amount    14712 non-null  int64
 9   total_buy_amount                14712 non-null  int64
 10  total_sell_amount               14712 non-null  int64
dtypes: int64(5), object(6)
```

users.csv

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8484 entries, 0 to 8483
Data columns (total 2 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   user_id  8484 non-null   int64
 1   churn    8484 non-null   int64
dtypes: int64(2)
memory usage: 132.7 KB
```

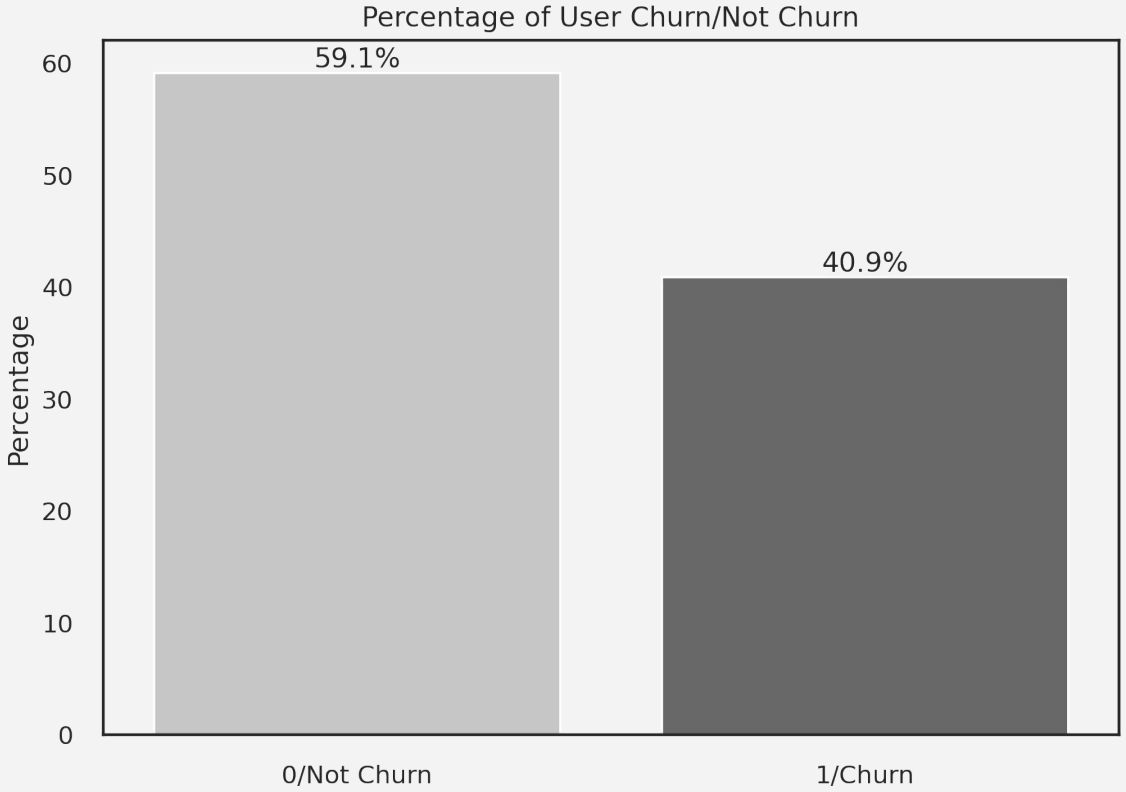
daily_user_transaction.csv

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 158811 entries, 0 to 158810
Data columns (total 17 columns):
 #   Column                                    Non-Null Count   Dtype
---  ------                                    --------------   -----
 0   user_id                                   158811 non-null  int64
 1   date                                      158811 non-null  object
 2   buy_saham_transaction_amount              99031 non-null   float64
 3   sell_saham_transaction_amount             1808 non-null    float64
 4   buy_pasar_uang_transaction_amount         122263 non-null  float64
 5   sell_pasar_uang_transaction_amount        2010 non-null    float64
 6   buy_pendapatan_tetap_transaction_amount   98916 non-null   float64
 7   sell_pendapatan_tetap_transaction_amount  1581 non-null    float64
 8   buy_campuran_transaction_amount           5072 non-null    float64
 9   sell_campuran_transaction_amount          46 non-null      float64
 10  total_buy_transaction_amount              158811 non-null  int64
 11  total_sell_transaction_amount             158811 non-null  int64
 12  saham_invested_amount                     106292 non-null  float64
 13  pasar_uang_invested_amount                131081 non-null  float64
 14  pendapatan_tetap_invested_amount          105946 non-null  float64
 15  campuran_invested_amount                  5352 non-null    float64
 16  total_invested_amount                     158811 non-null  int64
dtypes: float64(12), int64(4), object(1)
memory usage: 20.6+ MB
```

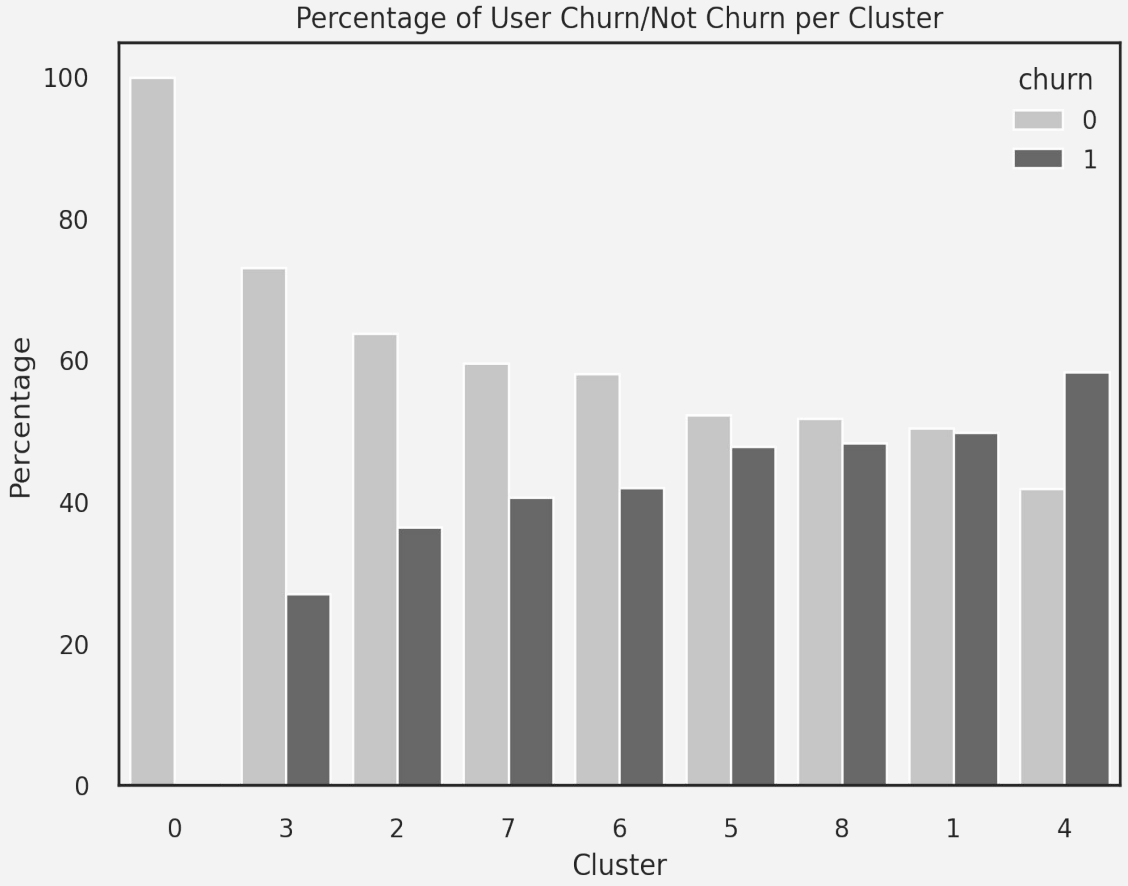
03 Data Cleaning and Merging

Steps	Consideration
Inspecting key column	I use merged dataset from intermediate assignment and new churn dataset. Both datasets has user_id, I decide to use that column as key column to be cleaned before joining. Inner join is used to merge.
Merging	New merged datasets have 8484 rows.
Handling Missing Data, Duplicates, Data Type	No missing data, no duplicates and all data are in acceptable type for EDA

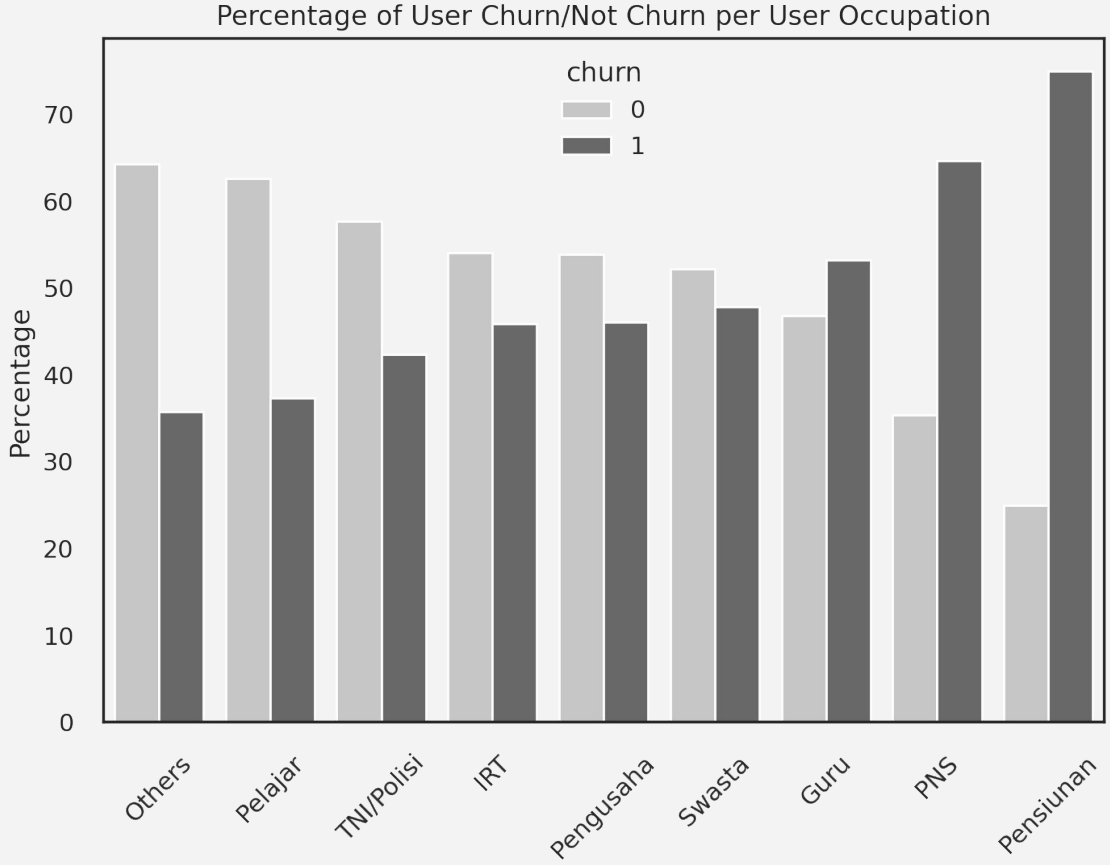
The churn data is balanced.



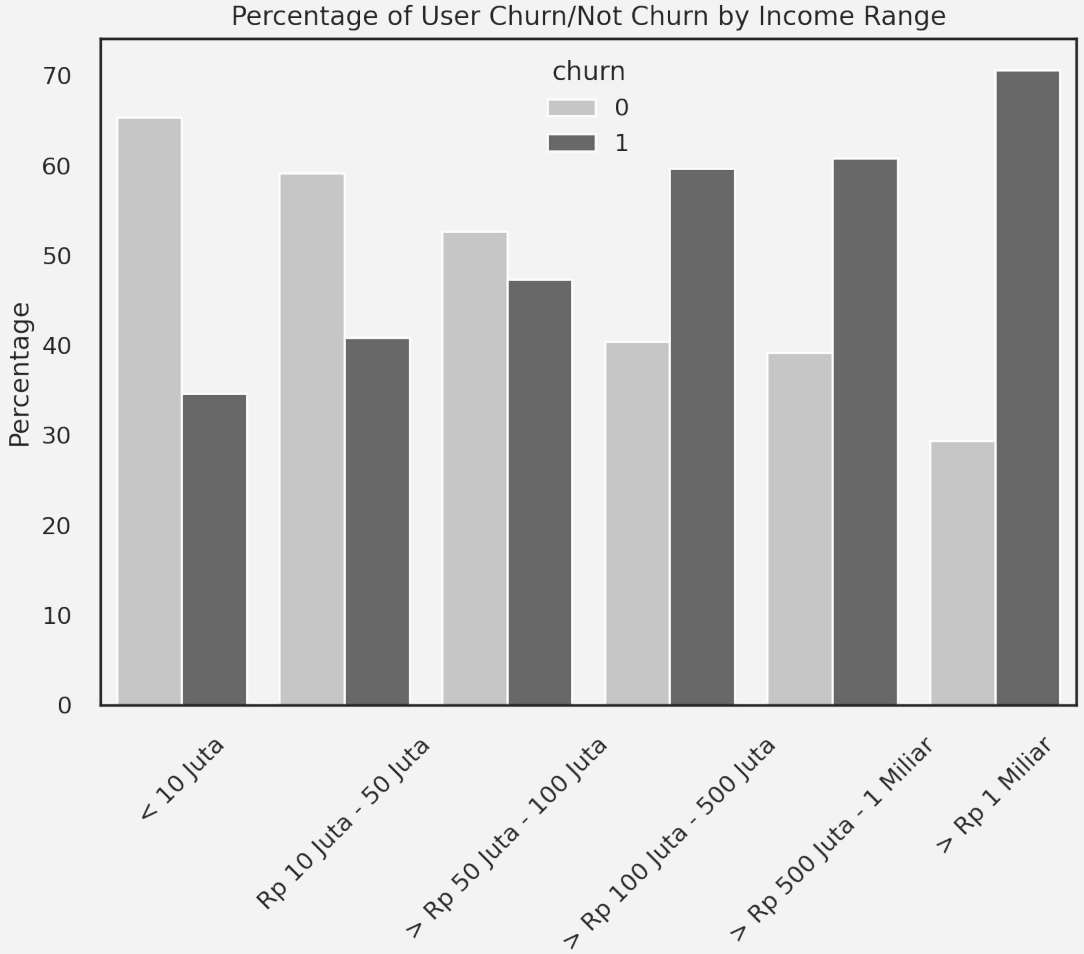
Almost everyone in cluster 0 don't churn, and cluster 4 have the highest percentage of churn.



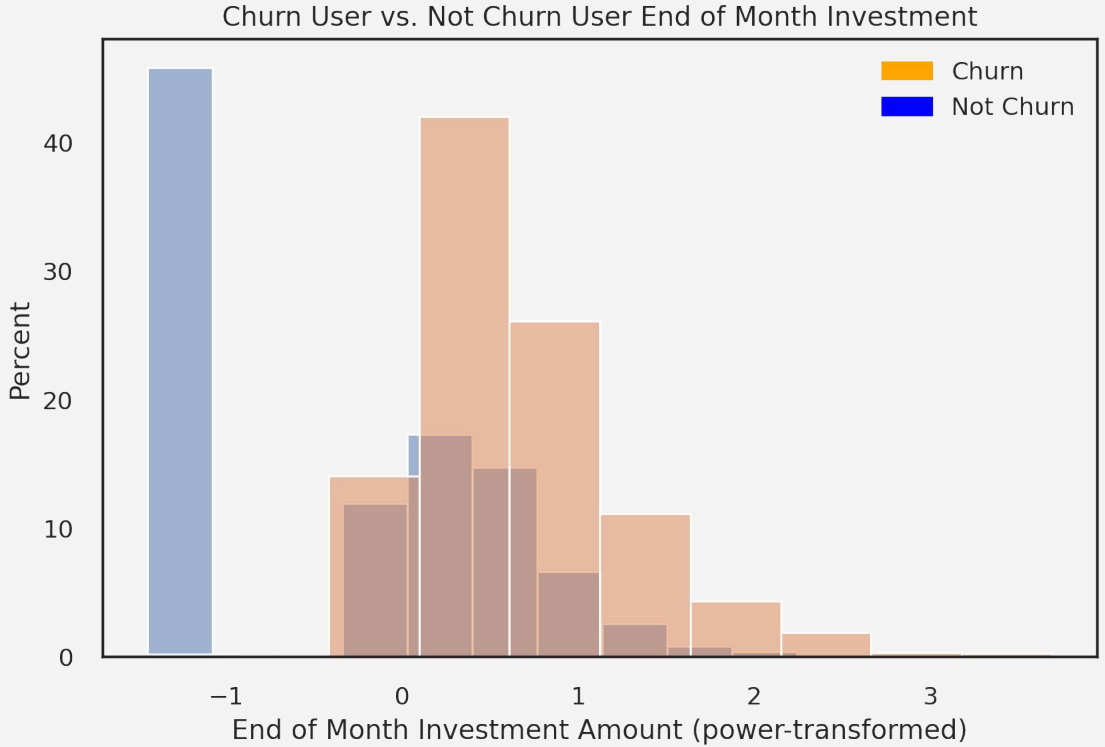
Pensiunan, PNS and Guru churn more than other occupation.



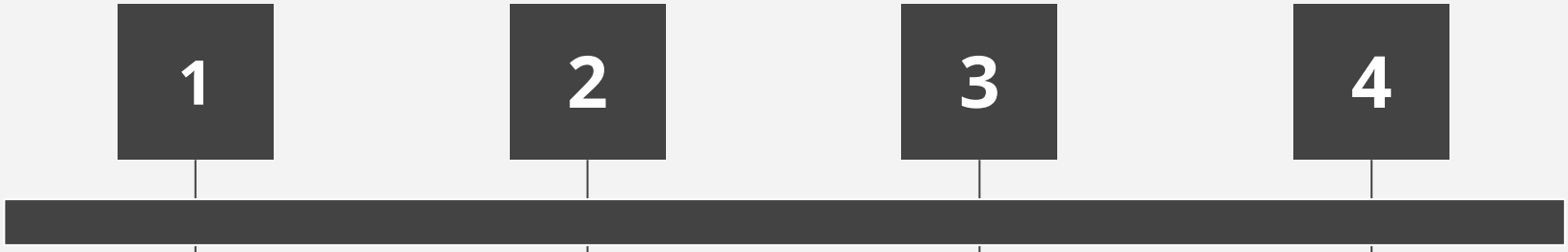
User with income >100 juta tend to churn.



User having investment at the end of month tend to not to do transaction again in the future.



05a- Build Classification Model



Feature Engineering

Convert object and categorical column to numerical by label encoding or one-hot encoding

Feature Selection

Remove column with high multicolinearity. This will also reduce the number of feature and improve model speed.

Model Selection and Evaluation

Test 3 model: Logistic Regression, Gradient Boosting and Random Forest on equal test set.

Obtain Prediction Probability

Select top 30% prediction probability and calculate cost benefit analysis.

05b- Model Selection and Evaluation Pipeline

3a

Train Test Split

3b

Power Transform High Skew Column

Power transform on training data. Obtain lambda for test set and future prediction.

3c

Recursive Feature Elimination

Using training data, let each model select the best feature to include based on feature importance or coefficient. Select n feature that give the best cross validated F1 score.

3d

Hyperparameter Tuning

Only on Gradient Boosting and Random Forest, if applicable

3e

Compare F1-score on test set

Logistic Regression is the chosen model

Model	F1-score
Logistic Regression	**0.716**
Gradient Boosting Classifier	0.58
Random Forest Classifier	0.58

06-Cost Benefit Analysis

Potential return is calculated using:

(Benefit of making people do transaction) - (Cost of marketing campaign)

Notes:

- Company get profit **0.15% per buy/sell transaction.**
- Cost of marketing campaign per user is **Rp 1000,-**
- Average buy and sell per month is obtained based on each cluster.
- Calculated on **top 0.3 percentile user who predicted to churn.**

	total_buy_amount			total_sell_amount		
	mean	median	max	mean	median	max
cluster_kmeans						
0	0.00	0.00	0.00	0.00	0.00	0.00
1	1441804.34	30000.00	326000000.00	328237.43	0.00	185000000.00
2	706589.16	0.00	433800000.00	136624.31	0.00	32500000.00
3	92256.86	0.00	10010000.00	107992.35	0.00	10000000.00
4	6705851.42	45000.00	452950000.00	1468279.52	0.00	112500000.00
5	1840834.39	0.00	207000000.00	230790.49	0.00	20000000.00
6	4893724.95	0.00	799500000.00	332156.39	0.00	80000000.00
7	319258.27	0.00	10000287.00	118707.76	0.00	4300000.00
8	2911000.00	0.00	50000000.00	1218390.80	0.00	50000000.00

06-Cost Benefit Analysis

Cluster	Potential Return	Cluster	Potential Return
0	No churn user.	5	Rp. 68.297
1	Rp. 1.337.148	6	Rp. 676.237
2	- Rp. 39.597	7	- Rp. 33.978
3	- Rp. 60.088	8	Rp. 159.776
4	Rp. 1.089.120		

Recommendation:
Do marketing campaign only on cluster 1, 4, 5, 6 and 8

Thanks

Google Colab