Predict Customer Personality to Boost Marketing Campaign by Using Machine Learning

Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

Created by:

Afif Surya Pradipta

Let's connect!

- afifsuryapradipta@gmail.com
- https://www.linkedin.com/in/afifsp
- https://github.com/afifsurya

A bachelor with abilities in analyzing and solving problems through fact-based and data-driven decision making which make him proficiency in python, SQL, statistics, machine learning and also had experiences in data analytics and data engineering.

Overview

A company can develop rapidly when it knows the behavior of its customer personality, so that it can provide better services and benefits to customers who have the potential to become loyal customers. By processing historical marketing campaign data to improve performance and target the right customers to be able to do transaction(s) on the company's platform, from these data insights my focus is to create a cluster prediction model so that it makes easier for companies to make decisions.

Tools

PROGRAMMING LANGUAGE

DATA VISUALIZATION

matplotlib

seaborn

PYTHON LIBRARY

pandas

NumPy

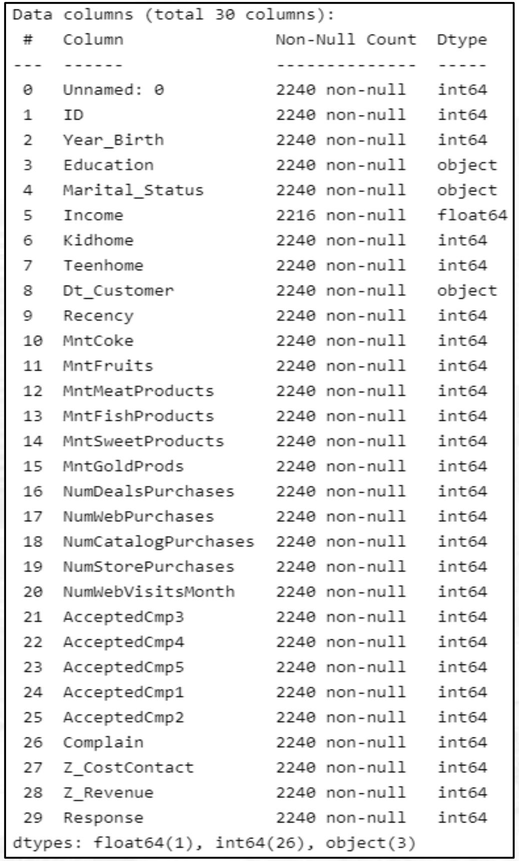
scikit learn

d dython

NOTEBOOK

jupyter

Dataset



```
Data columns (total 30 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Unnamed: 0           2240 non-null   int64
 1   ID                   2240 non-null   int64
 2   Year_Birth           2240 non-null   int64
 3   Education            2240 non-null   object
 4   Marital_Status       2240 non-null   object
 5   Income               2216 non-null   float64
 6   Kidhome              2240 non-null   int64
 7   Teenhome             2240 non-null   int64
 8   Dt_Customer          2240 non-null   object
 9   Recency              2240 non-null   int64
 10  MntCoke              2240 non-null   int64
 11  MntFruits            2240 non-null   int64
 12  MntMeatProducts      2240 non-null   int64
 13  MntFishProducts      2240 non-null   int64
 14  MntSweetProducts     2240 non-null   int64
 15  MntGoldProds         2240 non-null   int64
 16  NumDealsPurchases    2240 non-null   int64
 17  NumWebPurchases      2240 non-null   int64
 18  NumCatalogPurchases  2240 non-null   int64
 19  NumStorePurchases    2240 non-null   int64
 20  NumWebVisitsMonth    2240 non-null   int64
 21  AcceptedCmp3         2240 non-null   int64
 22  AcceptedCmp4         2240 non-null   int64
 23  AcceptedCmp5         2240 non-null   int64
 24  AcceptedCmp1         2240 non-null   int64
 25  AcceptedCmp2         2240 non-null   int64
 26  Complain             2240 non-null   int64
 27  Z_CostContact        2240 non-null   int64
 28  Z_Revenue            2240 non-null   int64
 29  Response             2240 non-null   int64
dtypes: float64(1), int64(26), object(3)
```

DESCRIPTION

Dataset contains customer behavior features who made transactions and interactions on our platform

SHAPE

2.240 data rows, 30 features

DTYPE

Float64 (1 features), int64 (26 features), object (3 features)

MISSING VALUE

1 features that has missing value; Income

DUPLICATED DATA

0 data rows

For the detail of my codes, you can see here

Exploratory Data Analysis (EDA)

FEATURE EXTRACTION

- Total_Acc_Cmp

Total of accepted campaign

- Total_Purchases

Total of item purchases

- cvr

Conversion rate

- Age
- Age_Group

Age classification

- Total_Spent
- NumChildren

Total of children

- Dt_Collected

The day when data collected

- Dt_Days_Customer

How long customer has been a member

DTYPE

Object (3 features)

↓

Object (16 features)

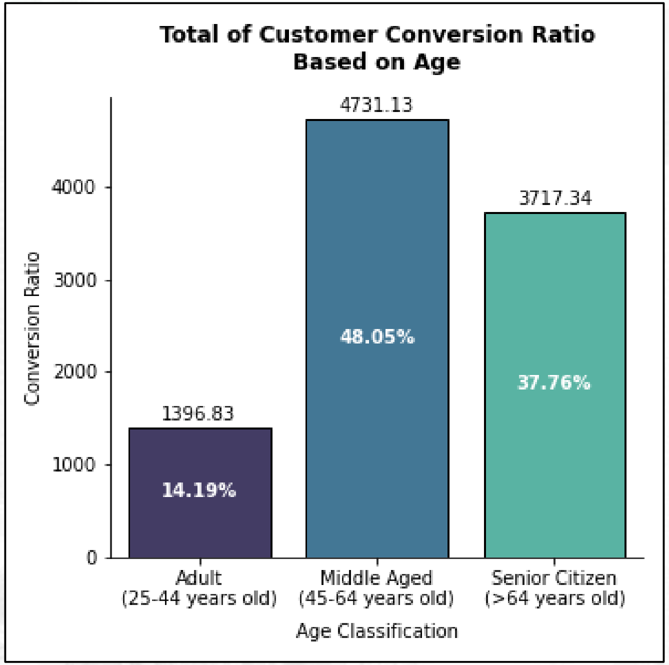
CORRELATION

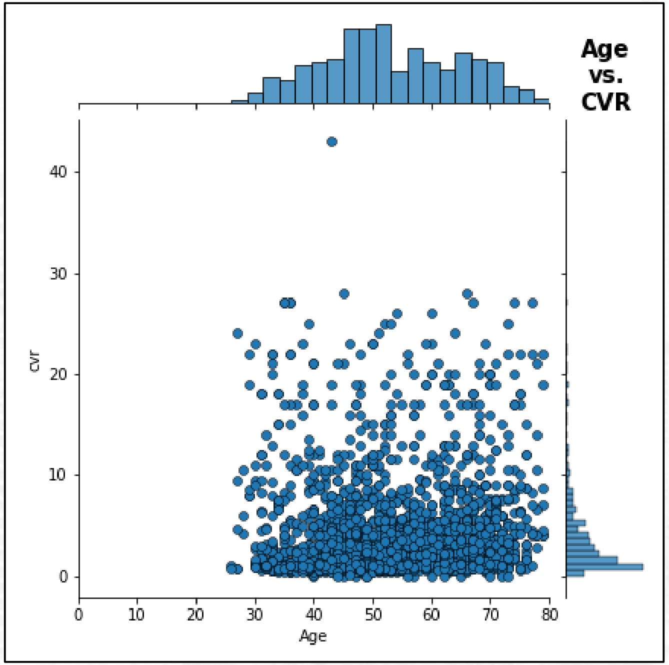
4 features highly corelated

- Age
- Income
- Total_Spent
- cvr

For the detail of my codes, you can see here

Conversion Rate Analysis Based on Income, Spending, and Age

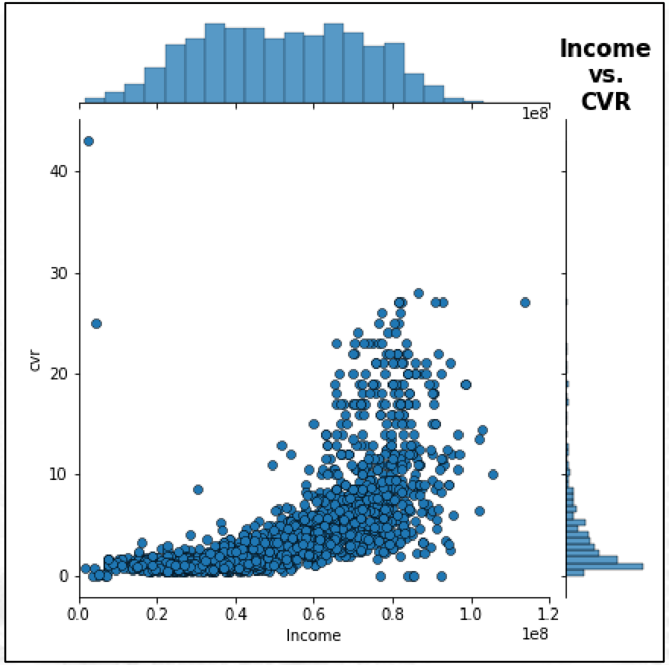


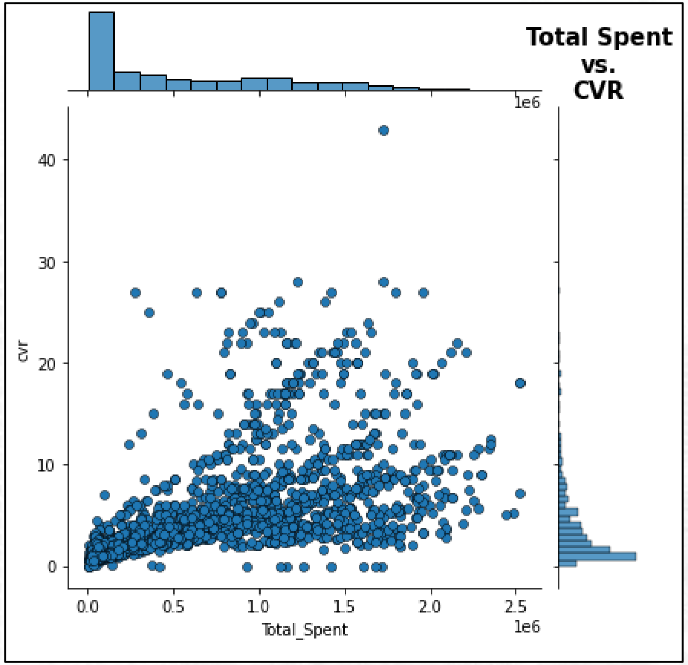


- Based on analysis visualization above, Middle Aged dominated the distribution according to it's cvr (48.05%). Followed by Senior Citizen (37.76%) and Adult (14.19%).
- On this case, we must pay attention for Middle Aged Group to maintain their retention for shopping on our platform. We should give them more personalized ads or specific products to offer for their age.

For the detail of my codes, you can see here

Conversion Rate Analysis Based on Income, Spending, and Age

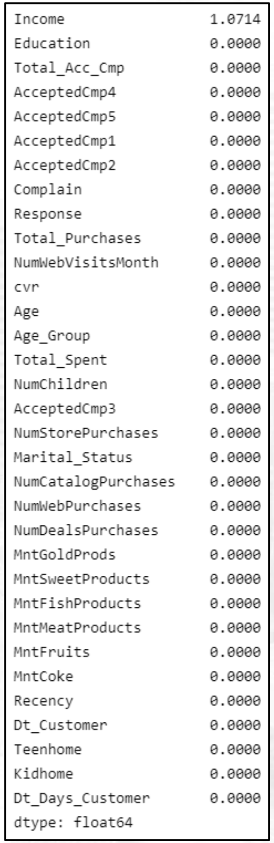




- The higher customer income, the higher cvr they have. The higher cvr dominated by customer who has income >IDR 60M/year.
- Also it's directly proportional to their total spent on our platform. Customers who have total spent >1M/year, their cvr around 5-40.

For the detail of my codes, you can see here

Data Cleaning & Preprocessing



```
Income                 1.0714
Education              0.0000
Total_Acc_Cmp          0.0000
AcceptedCmp4           0.0000
AcceptedCmp5           0.0000
AcceptedCmp1           0.0000
AcceptedCmp2           0.0000
Complain               0.0000
Response               0.0000
Total_Purchases        0.0000
NumWebVisitsMonth      0.0000
cvr                    0.0000
Age                    0.0000
Age_Group              0.0000
Total_Spent            0.0000
NumChildren            0.0000
AcceptedCmp3           0.0000
NumStorePurchases      0.0000
Marital_Status         0.0000
NumCatalogPurchases    0.0000
NumWebPurchases        0.0000
NumDealsPurchases      0.0000
MntGoldProds           0.0000
MntSweetProducts       0.0000
MntFishProducts        0.0000
MntMeatProducts        0.0000
MntFruits              0.0000
MntCoke                0.0000
Recency                0.0000
Dt_Customer            0.0000
Teenhome               0.0000
Kidhome                0.0000
Dt_Days_Customer       0.0000
dtype: float64
```

HANDLE MISSING VALUE

- 1% missing value on Income

Fill it with median

DUPLICATED DATA

- 0 duplicated data

FEATURE ENCODING

Since the feature I've used for modeling only numeric, so I didn't do any feature encoding

HANDLING OUTLIERS

Using IQR method (Q1=1%; Q3=99%)

FEATURE SELECTION

Using RFMLC method to reduce dimensionality

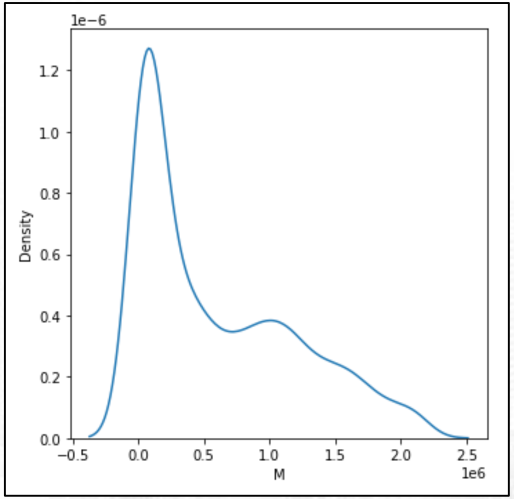
- R (Recency): Recency
- F (Frequency): Total_Purchases
- M (Monetary): Total_Spent
- L (Loyalty): Dt_Days_Customer
- C: Age

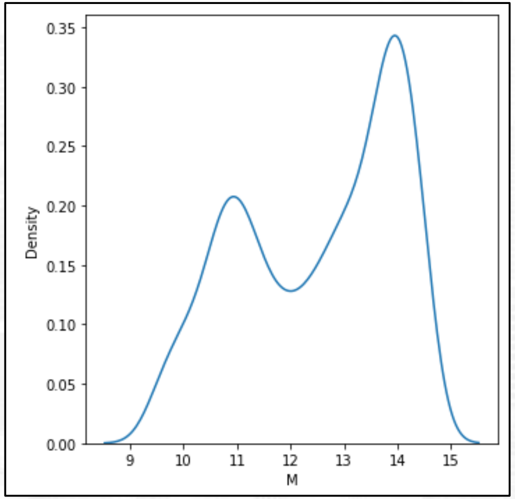
FEATURE TRANSFORMATION

Do standardization to all 5 features for modeling using MinMaxScaler

For the detail of my codes, you can see here

Data Cleaning & Preprocessing

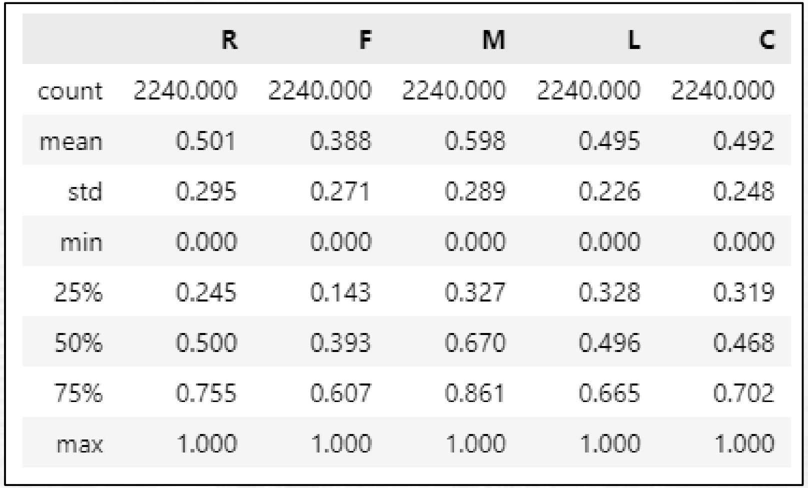




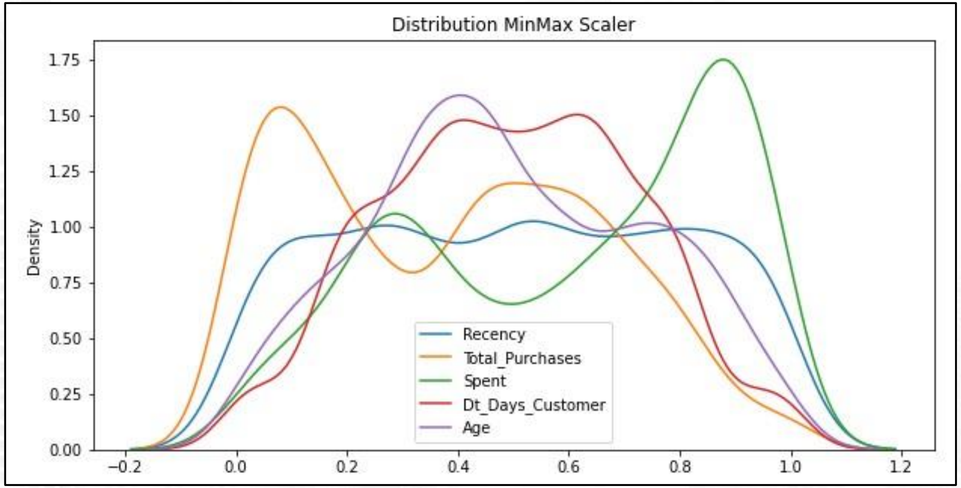
- Column Total_Spent or M is skewed-right, not good for K-Means. So, transform it to using log method.
- As we would like to see on the right one, distribution changed from skewed-right to normal.

For the detail of my codes, you can see here

Data Cleaning & Preprocessing



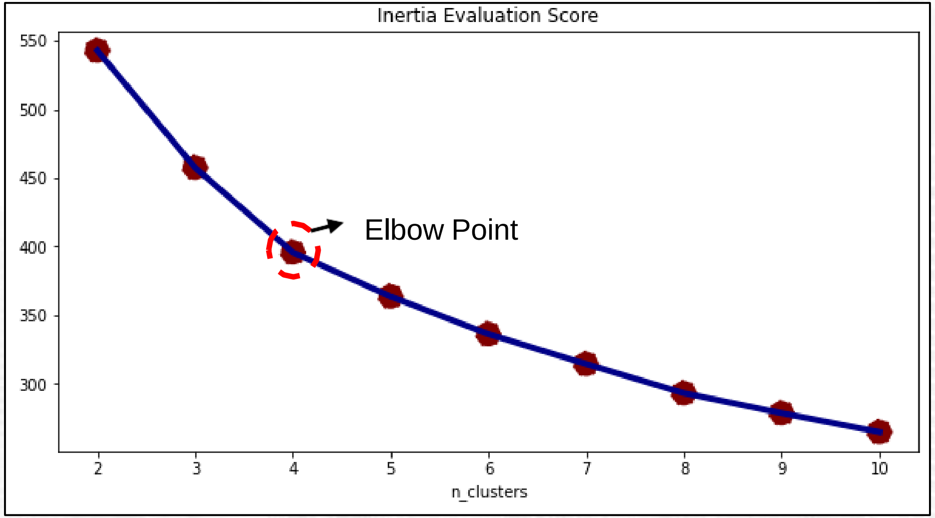
	R	F	M	L	C
count	2240.000	2240.000	2240.000	2240.000	2240.000
mean	0.501	0.388	0.598	0.495	0.492
std	0.295	0.271	0.289	0.226	0.248
min	0.000	0.000	0.000	0.000	0.000
25%	0.245	0.143	0.327	0.328	0.319
50%	0.500	0.393	0.670	0.496	0.468
75%	0.755	0.607	0.861	0.665	0.702
max	1.000	1.000	1.000	1.000	1.000

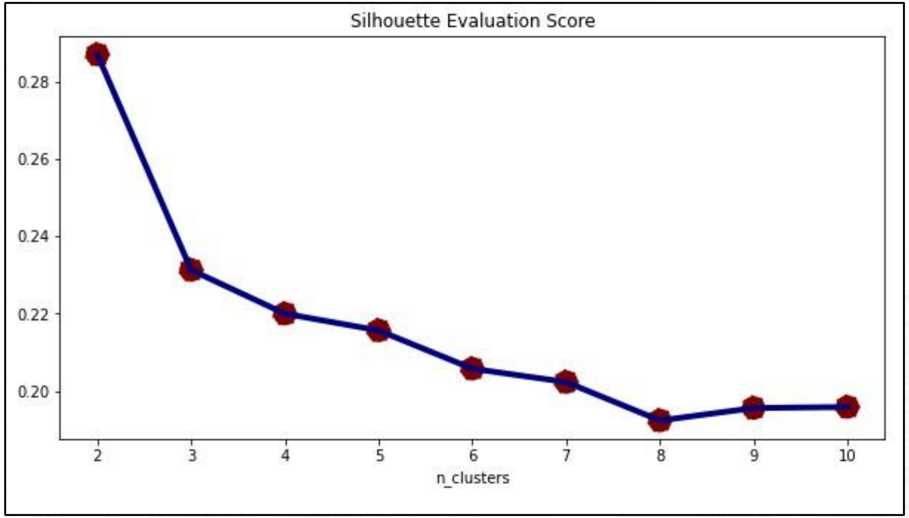


- Data has been standardized using MinMaxScaler.

For the detail of my codes, you can see here

Data Modeling



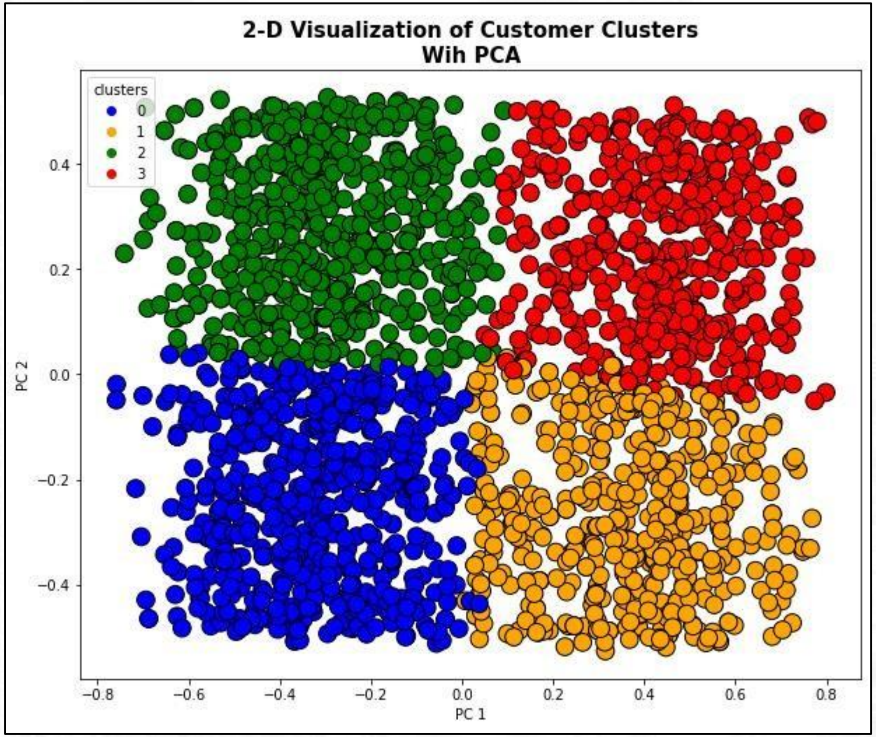


- To find optimal number of clusters, I've used elbow method model evaluation on inertia score then look at silhouette score to validate it.
- From evaluation above, n_clusters =4 is an elbow point, because after this point there isn't much significant decreases on inertia score. Also, on silhouette score n_clusters =4 is better than n_clusters =5 as score isn't closer to 0.
- n_clusters =4 is an optimal number for K-means Clustering Modeling on this dataset.

For the detail of my codes, you can see here

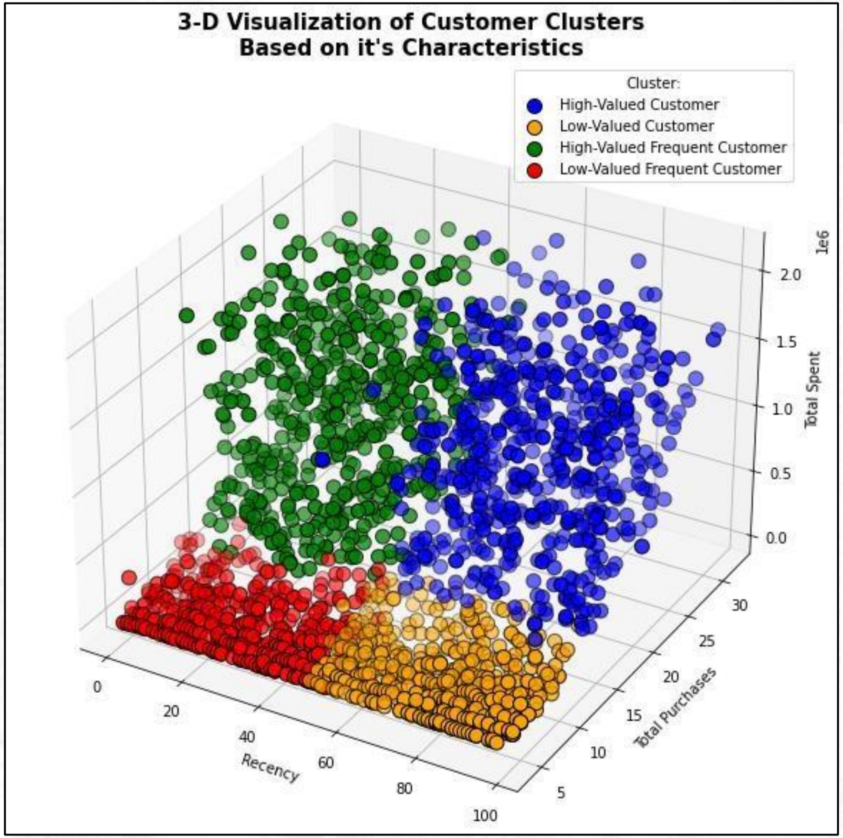
Data Modeling

- According to visualization using PCA with 2 main PC's, the clusters are perfectly separated.
- There's clearly 4 customer clusters that generated by K-Means Clustering algorithm using RFMLC Method for this dataset.



For the detail of my codes, you can see here

Customer Personality Analysis for Marketing Retargeting

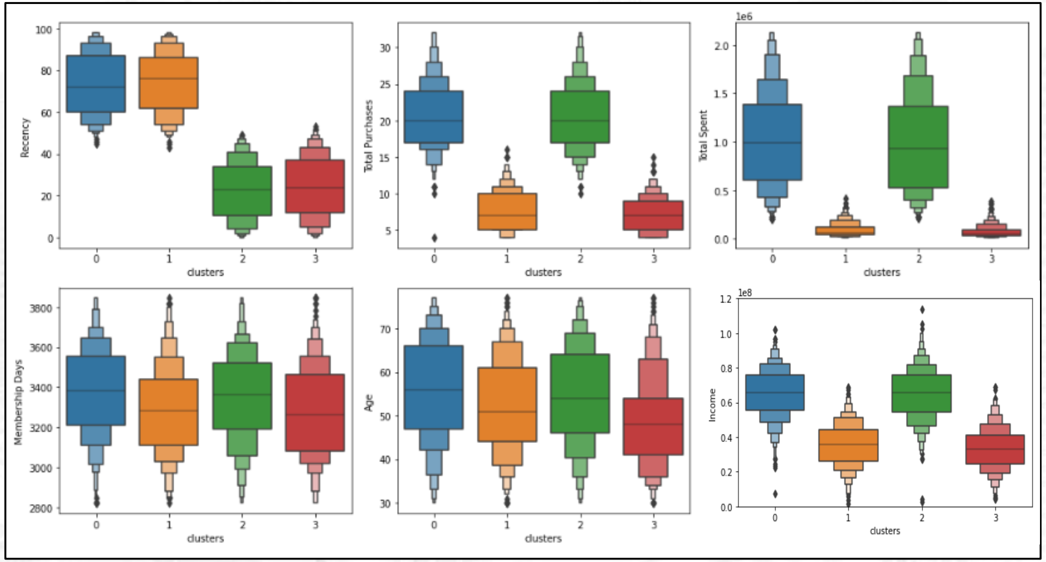


1. High-Valued Customer (Cluster 0):

- There are 648 customers (28.93% of total customers) on this group.
- Customers on this group have `high average recency (73 days)` and `high average of total purchases (21 items)` it means they are not frequent shoppers but `they spend a lot on our platform (around IDR 1M/year)`.
- This group dominated by 48.46% customers at Middle-Aged (45-64 years old), mostly they have 1 children and they have highest average income (around IDR 65M/year) with low average web visits each month (4 times).

For the detail of my codes, you can see here

Customer Personality Analysis for Marketing Retargeting

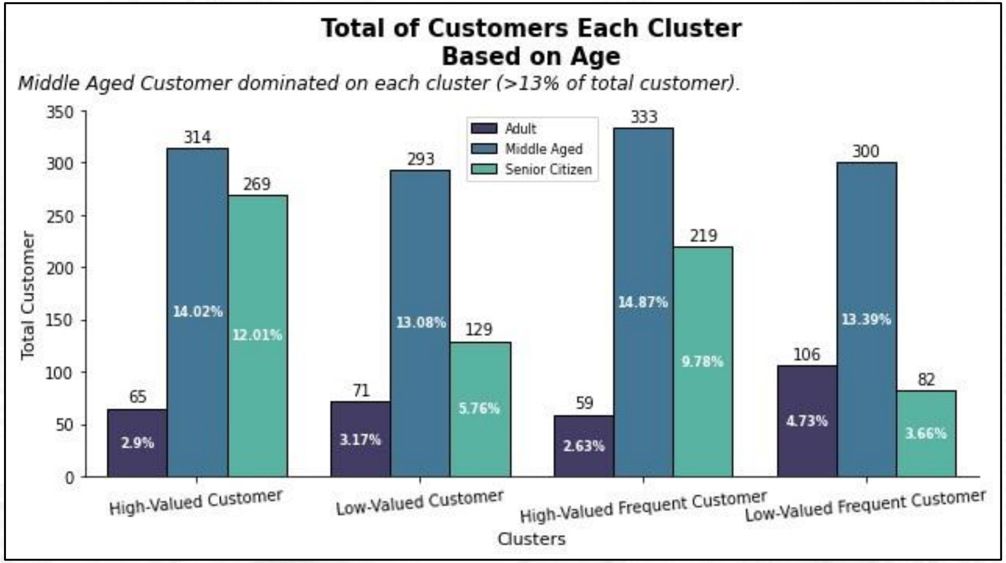


2. Low-Valued Customer (Cluster 1):

- There are 493 customers (22.01% of total customers) on this group.
- Customers on this group have `highest average recency (74 days)` and `low average of total purchases (8 items)` it means they are not frequent shoppers and `they spend a little on our platform (around IDR 92K/year)`.
- This group dominated by 59.43% customers at Middle-Aged (45-64 years old), mostly they have 1 children and they have average income (around IDR 36M/year) with hig average web visits each month (6 times).

For the detail of my codes, you can see here

Customer Personality Analysis for Marketing Retargeting

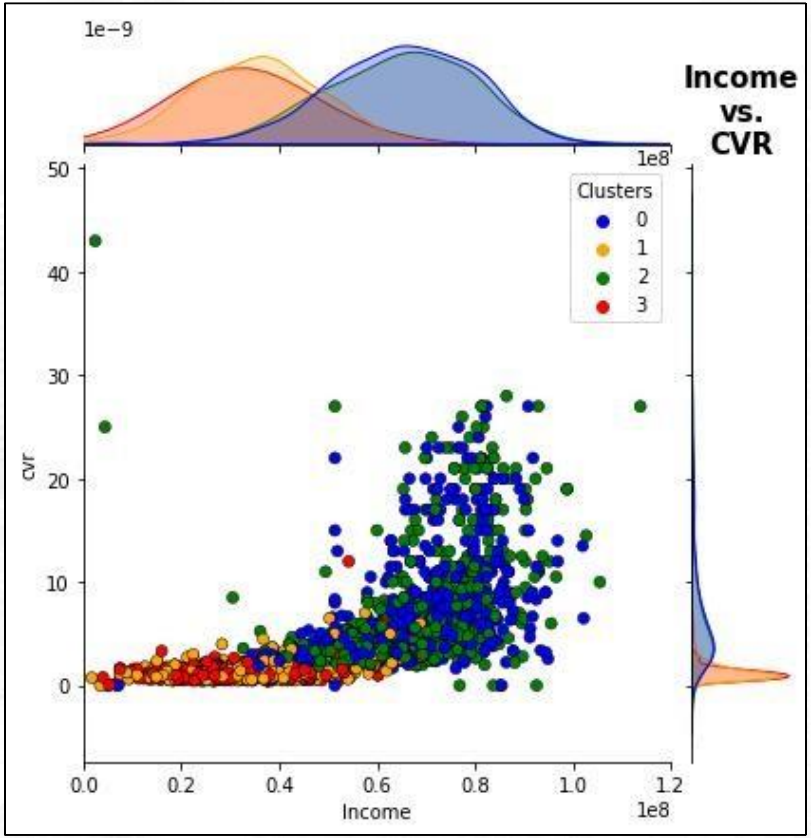


3. High-Valued Frequent Customer (Cluster 2):

- There are 611 customers (27.28% of total customers) on this group.
- Customers on this group have `low average recency (23 days)` and `high average of total purchases (21 items)` it means they are frequent shoppers and `they spend a lot on our platform (around IDR 989K/year)`.
- This group dominated by 54.5% customers at Middle-Aged (45-64 years old), mostly they have 1 children and they have average income (around IDR 65M/year) with low average web visits each month (4 times).

For the detail of my codes, you can see here

Customer Personality Analysis for Marketing Retargeting



4. Low-Valued Frequent Customer (Cluster 3):

- There are 488 customers (21.79% of total customers) on this group
- Customers on this group have `high average recency (24 days)` and `lowest average of total purchases (7 items)` it means they are frequent shoppers but `they spend a little on our platform (around IDR 75K/year)`
- This group dominated by 61.48% customers at Middle-Aged (45-64 years old),mostly they have 1 children and they have average income (around IDR 35M/year) with high average web visits each month (6 times)

For the detail of my codes, you can see here

Recommendation

***Actionable Insights*:**

1. Develop a membership tier program to enhance customer retention and incentivize increased shopping activity on our platform. We can establish four membership tiers - Platinum, Gold, Silver, and Bronze - each offering distinct privileges to customers. The level of membership will correspond to customer clusters, with Platinum reserved for high-value customers, Gold for highly frequent customers, Silver for moderately frequent customers, and Bronze for customers with lower value.
2. To minimize churn risk, our primary focus should be on the High-Valued Customers group. We need to closely monitor their purchasing trends and implement strategies to retain their loyalty. This can be achieved by improving our services, providing excellent after-sales support, maintaining product quality, and optimizing our mobile apps. Additionally, we can offer them the top-tier Platinum membership, entitling them to exclusive benefits such as higher discounts, promotions, and free shipping. These perks will encourage them to shop more frequently on our platform.

For the detail of my codes, you can see here

Recommendation

***Actionable Insights*:**

3. To further boost engagement from the High-Valued Frequent Customer group, we should provide them with additional promotions and coupons for free shipping through our membership tier program. By doing so, we aim to incentivize them to shop more often on our platform, leveraging their loyalty and strengthening their connection with our brand.
4. Since the Low-Valued Frequent Customer and Low-Valued Customer segments have the lowest overall spending, we should concentrate on creating personalized advertisements, promotions, and campaigns that target affordable products. This approach aims to attract these customer groups to our platform. By tailoring our marketing efforts to their specific preferences and needs, we have the potential to increase their frequency of purchases and overall spending on lower-cost items, ultimately enhancing their engagement levels.

For the detail of my codes, you can see here

Recommendation

Potential Impact (Quantitative):

If we keep prioritize on Customer Groups/Clusters and they do not turn to churn, we still have potential GMV around IDR 1.3B/year (High-Valued Customer=IDR 670M/year; Low-Valued Customer=IDR 46M/year; Low-Valued Frequent Customer=IDR 604M/year; Low-Valued Customer=IDR 47M/year).

For the detail of my codes, you can see here