Predict Customer Personality to boost marketing campaign by using Machine Learning

Created by:
Moh. Harwin Prayoga
mh.prayoga11@gmail.com
https://www.linkedin.com/in/mhprayoga/

"An enthusiastic learner, analytical, and flexible graduate of bachelor's degree of Engineering Physics at Institut Teknologi Sepuluh Nopember. I had experience in leadership and teamwork in various organizations and events. Moreover, I have a decent ability in English and operating various data programming software such as MS Excel, Python, SQL, etc. I am excited about seeking a challenge in the field of data where my passion, education, and training background can be fully utilized. "

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

Overview

"A company can develop rapidly when it knows its customer personality behavior, so 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 so they can transact on the company's platform, from this data insight our focus is to create a cluster prediction model to make it easier for companies to make decisions"

Data Exploration

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2240 entries, 0 to 2239
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)
memory usage: 525.1+ KB
```

- The dataset has 2240 rows and 30 columns.
- 'Unnamed: 0' column is only an index that will be dropped.
- Contains 3 data types: float64, int64, object.
- Dt_Customer column has incorrect data type. Can be converted to datetime

For details, see jupyter notebook here

Feature Engineering

Rakamin Academy

Creating new columns :

No	Feature	Explanation
1	Dt_Customer	Convert to datetime data type
2	Age_group	Divided to 5 age category (Senior Adult, Middle-aged Adult, Young Adult, Children, Baby) from Year_Birth column.
3	Total_Children	Sum of Kidhome and Teenhome
4	MaritalStatus	Simplify values to InCouple and Alone
5	Total_Spend	Sum of MntCoke, MntFishProducts, MntFruits, MntMeatProducts, MntSweetProducts, and MntGoldProds
6	TotalAccCmpg	Sum of AcceptedCmp1, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, and AcceptedCmp5
7	TotalPurchases	Sum of NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, and NumWebPurchases
8	Conversion_Rate	From dividing TotalPurchases by NumWebVisitsMonth
9	Lifetime	Count from customers' join date to the end of the year 2014

For details, see jupyter notebook here

EXPLORATORY DATA ANALYSIS (EDA)

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

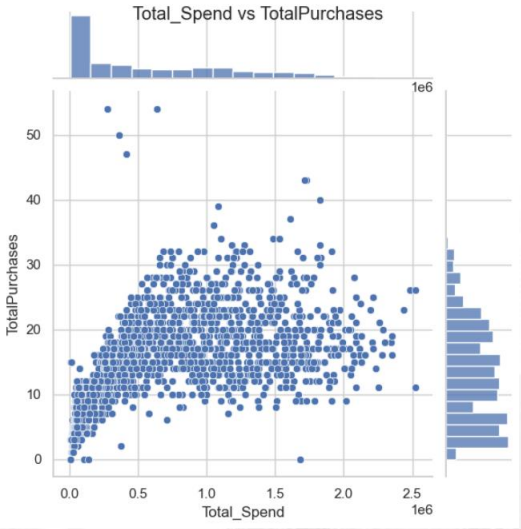
Rakamin Academy

Correlations:

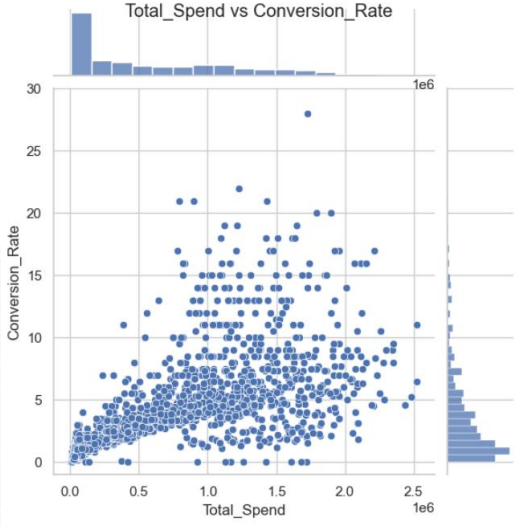
- ✓ Response has medium correlation with TotalAccCmpg
- ✓ Age has no strong correlation with any features.
- ✓ Total_Children has medium correlation with NumDealsPurchases and NumWebVisitsMonth.
- ✓ Total_Spend has very strong correlation with MntCoke, MntMeatProducts, NumCatalogPurchases and strong correlation with TotalPurchases, Conversion_Rate.
- ✓ TotalAccCmpg has strong correlation with MntCoke and medium correlation with Total_Spend.
- ✓ TotalPurchases has very strong correlation with NumWebPurchases and strong correlation with Conversion_Rate
- ✓ Lifetime has no strong correlation with any features.
- ✓ Conversion_Rate has strong correlation with Income, MntMeatProducts, MntSweetProducts, NumCatalogPurchases.

For details, see jupyter notebook here

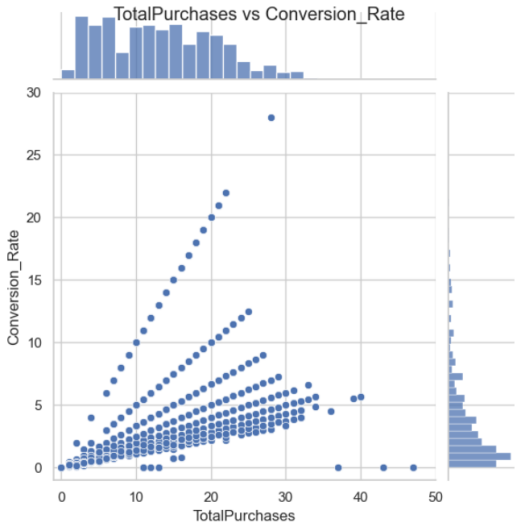
Exploratory Data Analysis (EDA)



The chart above shows that TotalPurchases & Total_Spend has positive correlation



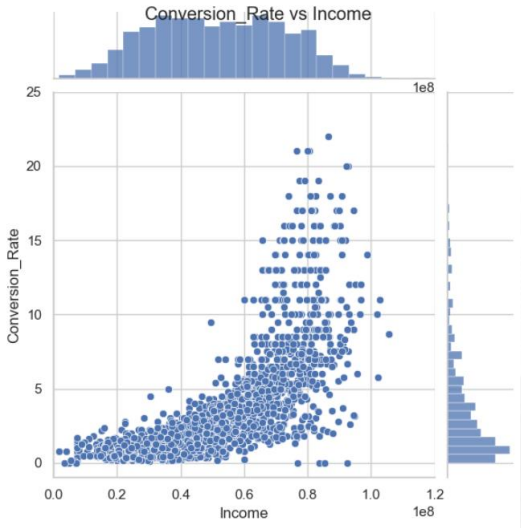
The chart above shows that Total_Spend & Conversion_Rate has positive correlation



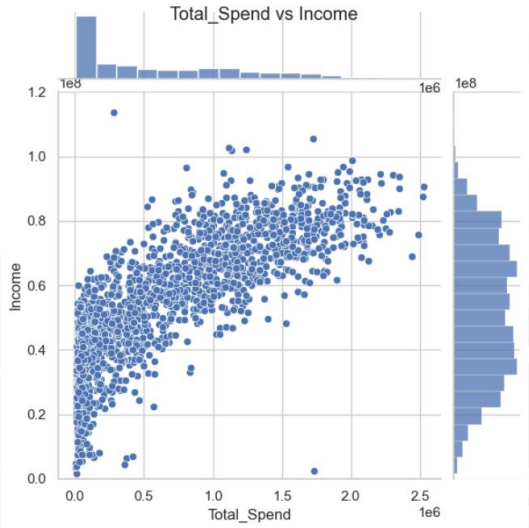
The chart above shows that TotalPurchases & Conversion_Rate has positive correlation.

For details, see jupyter notebook here

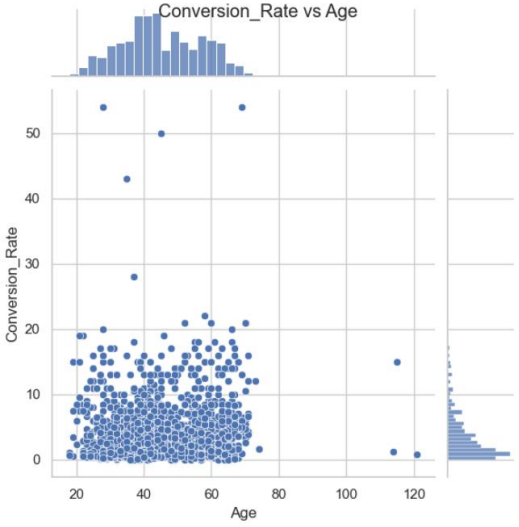
Exploratory Data Analysis (EDA)



The chart above shows that Conversion_Rate & Income has positive correlation



The chart above shows that Total_Spend & Income has positive correlation



Since in correlation heatmap plot and correlation with target, Conversion_Rate and Age has very insignificant correlation score that is only 11%. From Conversion Rate VS Age chart, indicates that the chart above is unable to determine the trend from those features.

For details, see jupyter notebook here

Insights & Recommendations

Insight :

- The greater customers spend total, the greater amount of customers purchase.
- The greater customers spend total, the greater number of customers conversion rate.
- The greater customers purchase total, the greater number of customers conversion rate.
- The greater customers income, the greater number of customers conversion rate.
- The greater customers income, the greater amount of spend.
- Customers age doesn't have an affection to customers conversion rate.

Recommendation :

Conversion Rate has a positive correlation against Total_Spend and Income. Hence, we can focus on customers with more than 6,000,000 in income and over 1,000,000 in total spending.

For details, see jupyter notebook here

DATA PREPROCESSING

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

Data Cleaning & Preprocessing

Rakamin Academy

Handling Outliers

Age and **Income** features are manually trimmed

Missing Values Handling

Inputted **Income** column missing value with median

Duplicated Handling

No duplicated data

Drop Features

Dropping '**Unnamed: 0**', '**ID**', '**Year_Birth**', '**Kidhome**', '**Teenhome**', '**Dt_Customer**', '**Marital_Status**'

For details, see jupyter notebook here

Data Cleaning & Preprocessing

Rakamin Academy

Feature Encoding

Do a feature encoding to **Education**, **Age_group**, **MaritalStatus** columns.

And then rename **Marital Status** to **Has_Couple**

```
Education have 5 unique values: int64
Education values: [2 4 3 0 1]
----------------------------------------
Has_Couple have 2 unique values: int64
Has_Couple values: [0 1]
----------------------------------------
Age_group have 3 unique values: int64
Age_group values: [2 0 1]
----------------------------------------
```

Feature Scaling

Do a feature **standardization**

```python
# for standardization
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
scaler.fit(df2_scal)
df2_scal = pd.DataFrame(scaler.transform(df2_scal), columns= df2_scal.columns )
df2_scal.describe()
```

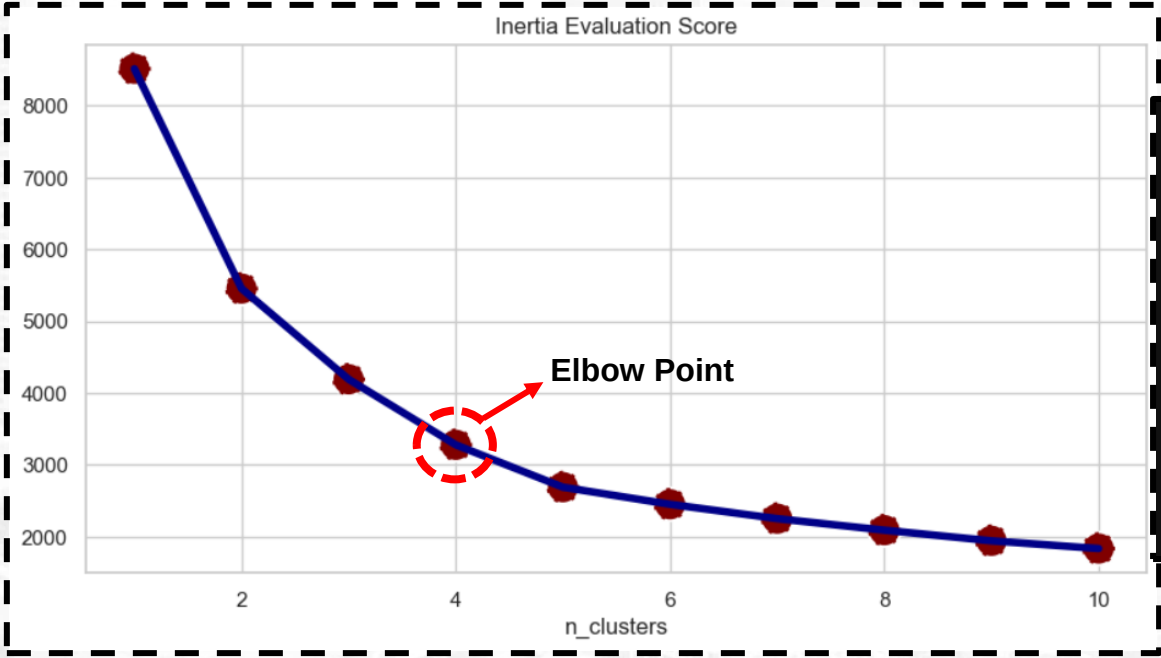
For details, see jupyter notebook here

DATA MODELLING

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

Data Modelling

❖ Elbow Method of K-means Clustering



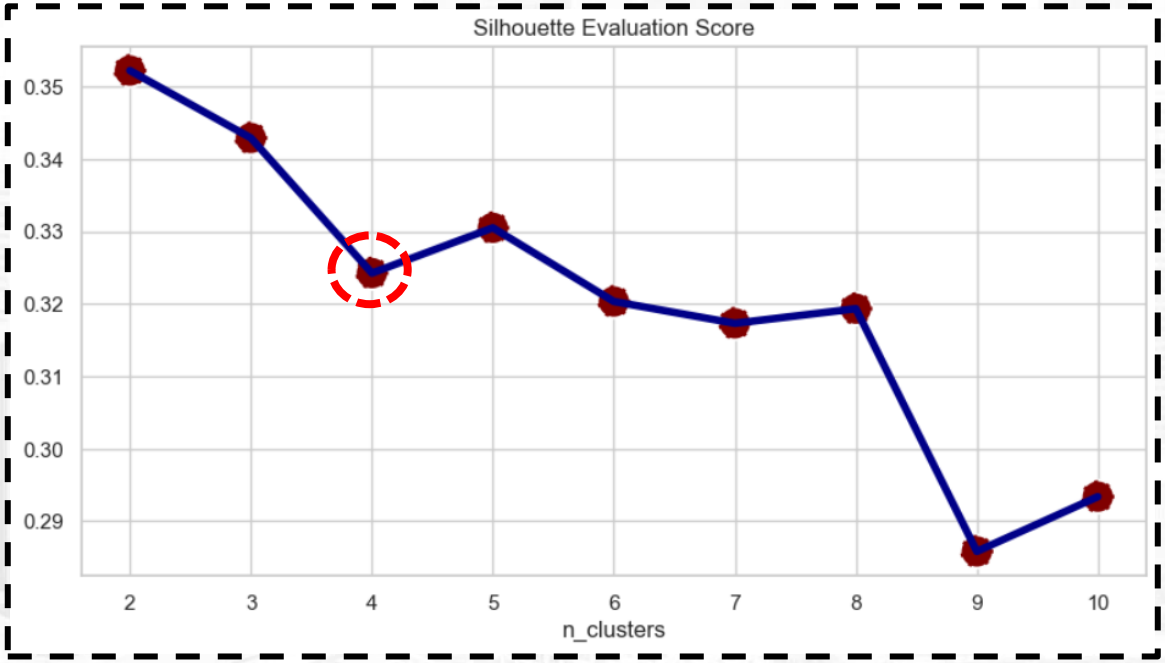

```
0    35.949648
1    23.176797
2    21.782225
3    18.002626
4     8.869346
5     8.163618
6     7.137277
7     7.095766
8     5.553627
9          NaN
dtype: float64
```

The **Elbow Method** is used in purpose to find the proper amount of clusters of K-means Clustering.

For details, see jupyter notebook here

Data Modelling

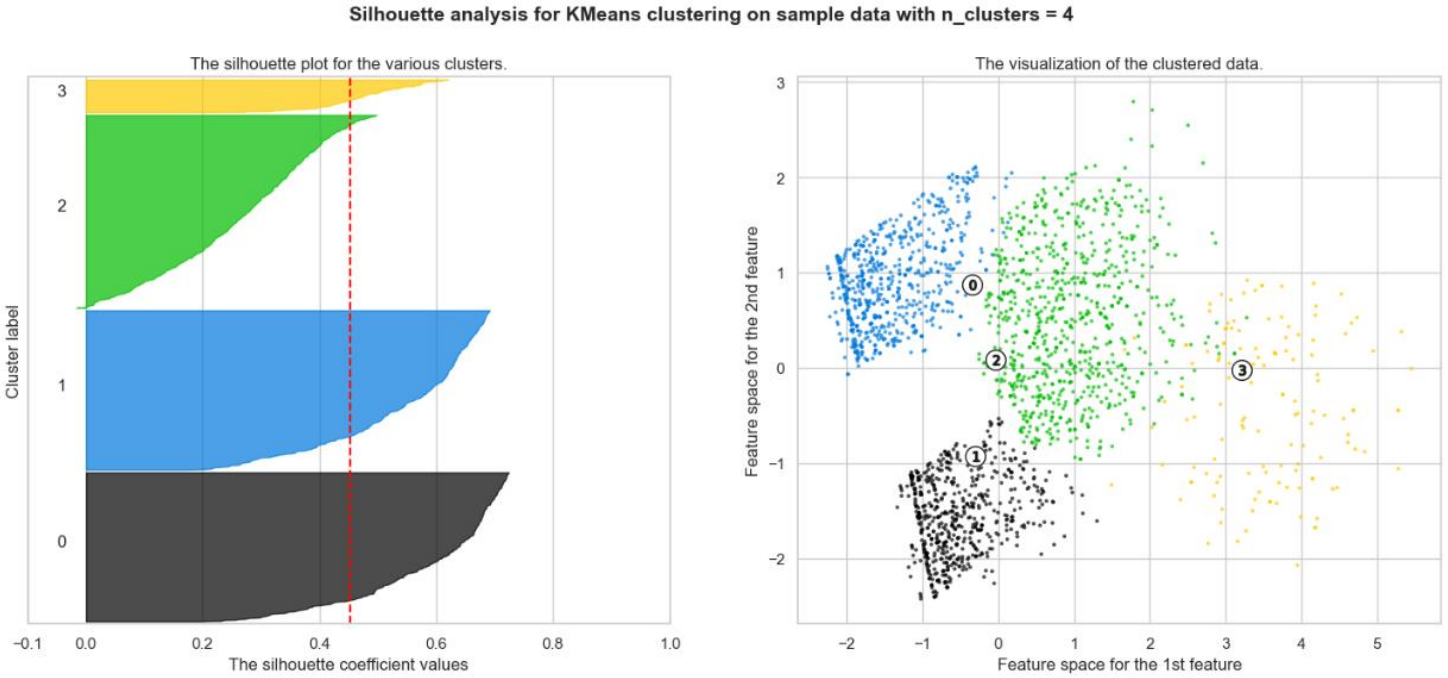
❖ Elbow Method by Silhouette Score



From the analysis by considering both evaluation scores in the charts, we decided to **divide the customers into 4 clusters** (n_clusters = 4)

For details, see jupyter notebook here

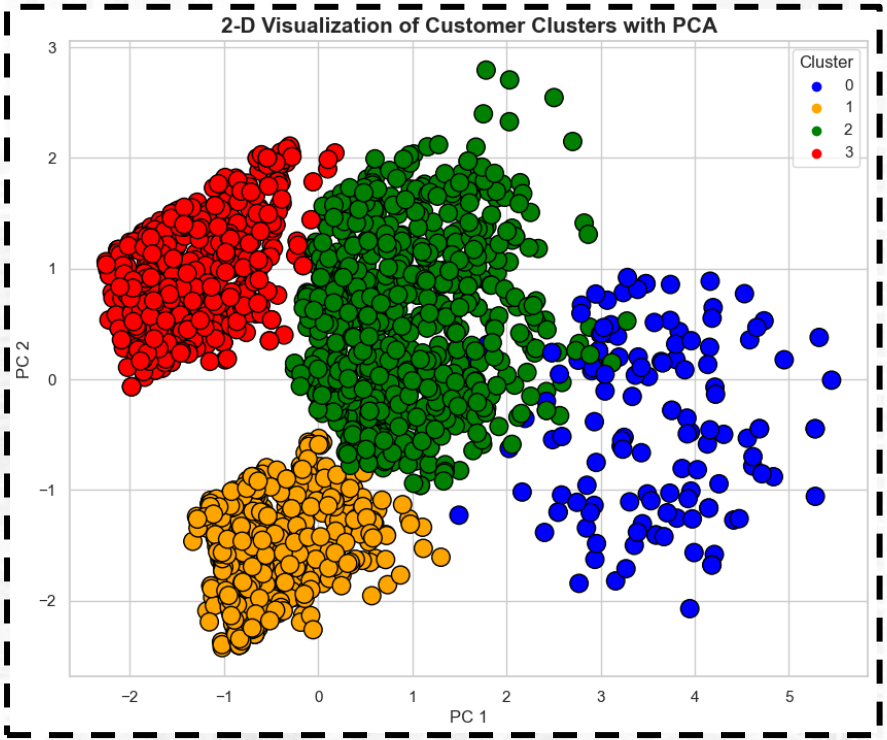
Silhouette Score Plot with PCA



Each cluster is above-average of silhouette scores which means the **number of clusters is optimal**

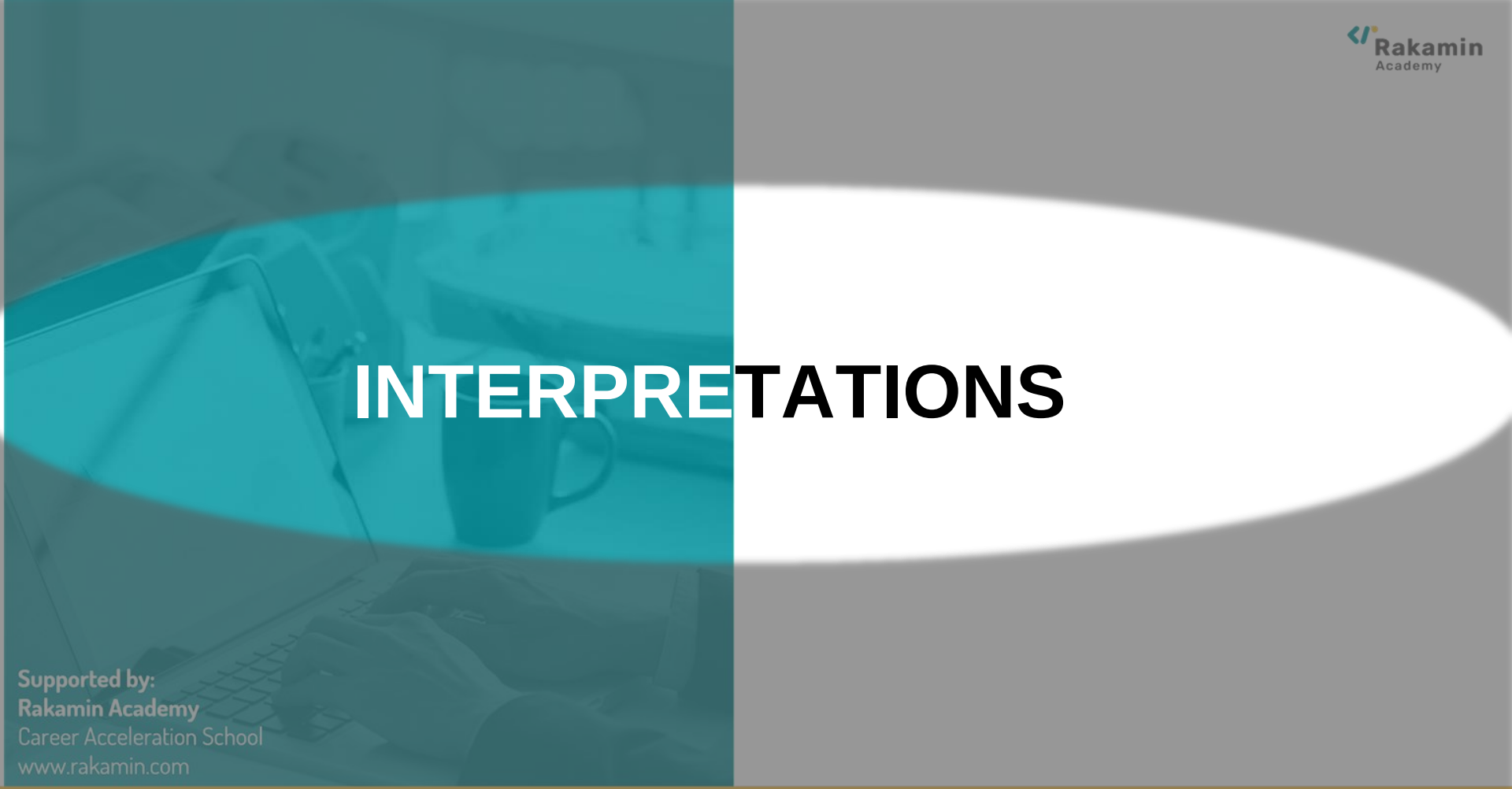
For details, see jupyter notebook here

Visualization of Customer Clusters with PCA

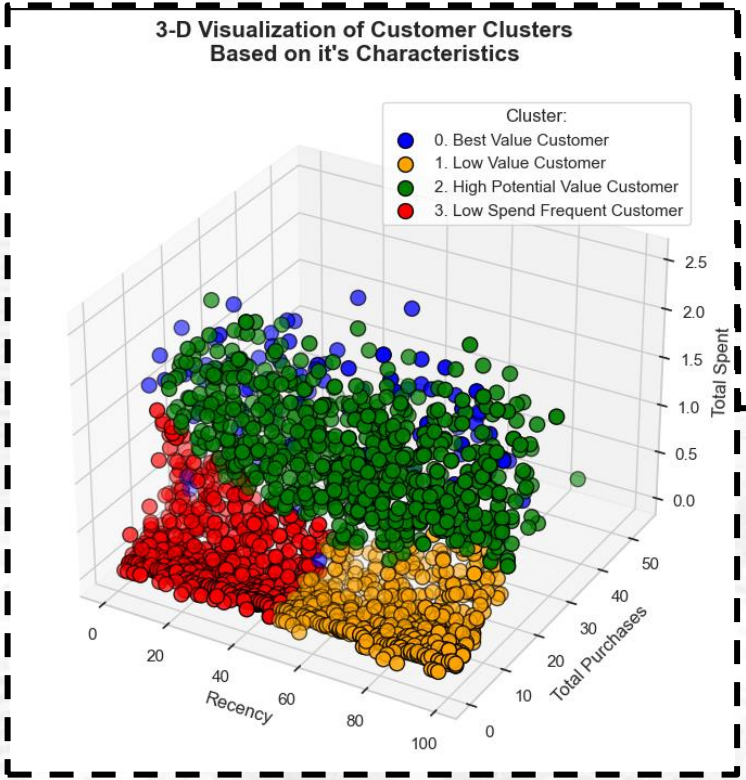


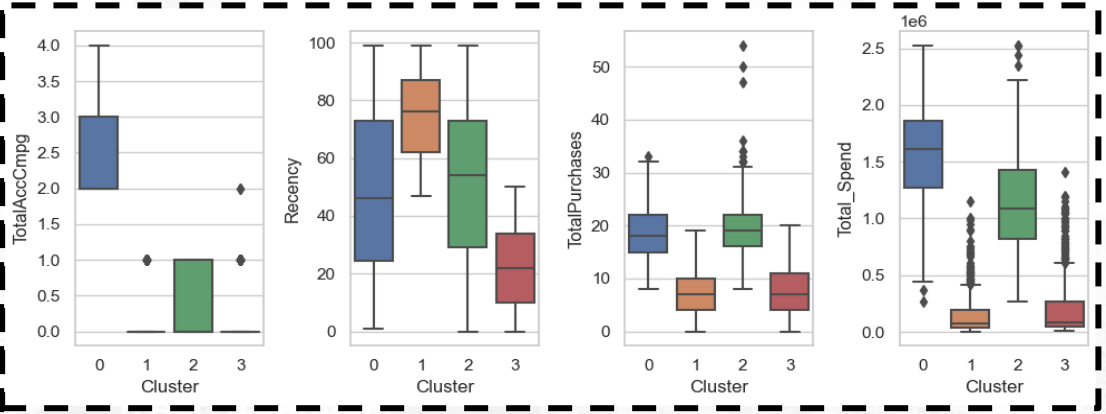
By visualizing customers clusters with Principal Component Analysis (PCA), we can see that each cluster were well segmented, so that each cluster has distinguishable characteristics

For details, see jupyter notebook here


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

Visualization of Customer Clusters Based on its Characteristics





Customer characteristics were decided desiring the amount of customer spends and how often transactions are made.

- **Cluster 0** = **Best Value Customer**, a customer that spent plenty and did do transactions quite frequently.
- **Cluster 1** = **Low Value Customer**, a customer that spent little and didn't do transactions frequently.
- **Cluster 2** = **High Potential Value Customer**, a customer that spent quite plenty and did do transactions quite frequently.
- **Cluster 3** = **Low Spend Frequent Customer**, a customer that spent little and did do transactions frequently.

For details, see jupyter notebook here

Customer Clusters Summary

Cluster 0 (Best Value Customer)

- Has 131 customers.
- Dominant with Middle-aged Adults and Senior Adults.
- Has the highest (around IDR 81 million) income per year.
- Has the highest (around IDR 1,61 million) total yearly spend.
- Has the lowest (far lower than other clusters) number of deals/promo purchases.
- Has the highest conversion rate.
- Is the **champions** cluster.

Cluster 1 (Low Value Customer)

- Has 594 customers.
- Dominant with Middle-aged Adults and Senior Adults.
- Has the lowest (around IDR 37 million) income per year.
- Has the lowest (around IDR 0,07 million) total spend per year.
- Has the third highest number of deals/promo purchases.
- Has the lowest conversion rate.
- Is in **need of attention**.

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

For details, see jupyter notebook here

Customer Clusters Summary

Cluster 2 (High Potential Value Customer)

- Has 769 customers.
- Dominant with Senior Adults.
- Has second highest (around IDR 68 million) income per year.
- Has second highest (around IDR 1,08 million) total yearly spend.
- Has the highest number of deals/promo purchases.
- Has the second-highest conversion rate.
- Is the **potential loyalist**.

Cluster 3 (Low Spend Frequent Customer)

- Has 635 customers.
- Dominant with Middle-aged Adults.
- Has second lowest (around IDR 38 million) income per year.
- Has second lowest (around IDR 0,09 million) total spend per year.
- Has the second highest number of deals/promo purchases.
- Has second lowest conversion rate.
- Is at **risk of churn**.

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

For details, see jupyter notebook here

Recommendations

1. Create a **membership tier program** to increase client retention. Membership tier items will also entice users to spend more on our platform. Assume we have four membership tiers (Platinum, Gold, Silver, and Bronze), and each membership category provides various customer benefits. The higher their membership tier, the more benefits they will receive. We can assign membership tiers based on customer clusters in this scenario (Platinum: High-Valued Customer, Gold: High-Valued Frequent Customer, Silver: Low-Valued Frequent Customer, Bronze: Low-Valued Customer).
2. Since we have **High Value Customers** (Best Value Customer & High Potential Value Customer), try to **prioritize focusing on their segment to avoid churn**. Continue to track their purchasing trends and retain them by improving our service, after-sales care, and the quality of our products and apps. Furthermore, we can provide them with the highest membership tier (Platinum Tier), in which case we can provide them with more discounts, promotions, and free shipping costs than any other membership tier in order to encourage them to buy on our platform more frequently.

For details, see jupyter notebook here

Recommendations

3. To avoid Low Spend Frequent Customers to do churn, provide more **promotions** or **free shipping cost coupons** to our **High-Valued Frequent Customer** group via our membership tier program to encourage them to shop on our platform more frequently.
4. Because **Low-Valued Frequent Customer** and **Low-Valued Customer** have the lowest overall spend on our platform, we should **produce more personalized ads, specials, and campaigns** for low-cost products to entice these groups to buy on our platform. This method may boost their recency (to low) and total number of purchases (to high) on varying products.

For details, see jupyter notebook here

Potential Impact

By calculating every **Total_Spend** from all clusters (assumpting won't do churn), we can estimate potential upcoming yearly GMV.

Total Spent of Best Value Customer: IDR 203.733.000

Total Spent of Low Value Customer: IDR 87.155.000

Total Spent of High Potential Value Customer: IDR 867.500.000

Total Spent of Low Spend Frequent Customer: IDR 124.327.000

Total Spent from All Clusters: **IDR 1.282.715.000**

For details, see jupyter notebook here