

# Predict Customer Personality to Boost Marketing Campaign by Using Machine Learning



**Created by:**

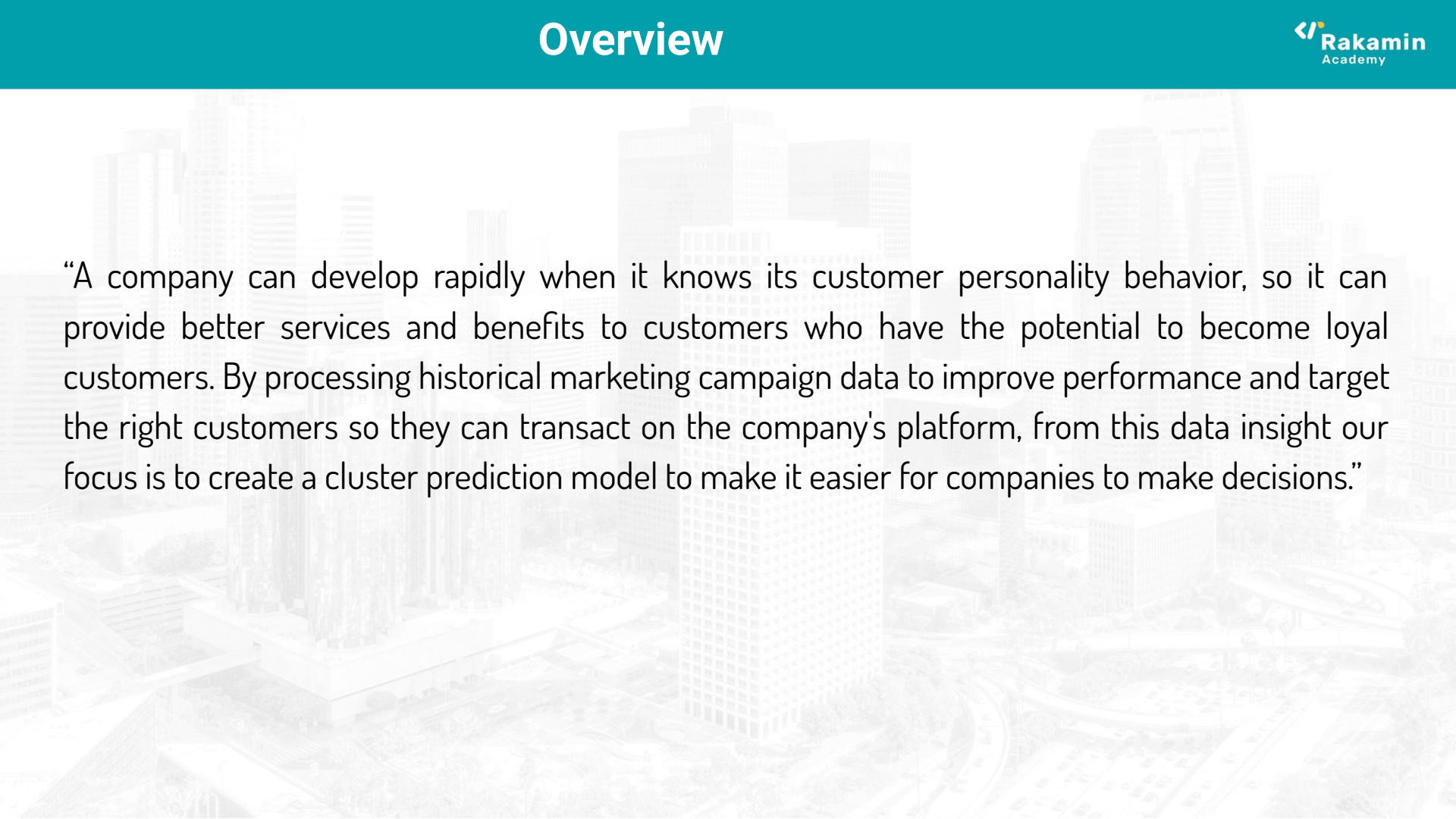
**Siti Hajjah Mardiah**

sitihamardiah1997@gmail.com

<https://www.linkedin.com/in/sitihajjahmardiah/>

Enthusiastic and open to any opportunity in the field of data, especially as a Data Scientist, Data Analyst, Business Analyst, Business Intelligence and so on.

Skilled in using PostgreSQL, Python, Git, Github, Looker Studio, Tableau, Power BI, Microsoft Office (Word, Excel, PowerPoint), etc.

A faded, light-colored background image of a city skyline with various skyscrapers and buildings, overlaid with a grid pattern.

“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.”

## Problem Statements

The company can develop rapidly when it knows the behavior of its customer personality, so that it can provide better services and benefits to potential customers 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.

## Goals

Creating customer segmentation to find out potential customers according to their behavior with the aim of providing the best treatment for customers so that companies can easily make decisions and improve marketing campaign performance.

## Objectives

Creating customer segmentation by cluster prediction model using unsupervised learning to find out potential customers according to their behavior.

## Dataset Description

This dataset contains 2240 samples/rows and 29 features/columns, which is each feature can be grouped into several groups including Accepted/Responses Campaign, Customer Information, Sales Product Type, Number of Purchases Type, Cost and Revenue. The following are for more details:

### Accepted/Responses Campaign

- `AcceptedCmp1` - 1 if customer accepted the offer in the 1st campaign, 0 otherwise
- `AcceptedCmp2` - 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
- `AcceptedCmp3` - 1 if customer accepted the offer in the 3rd campaign, 0 otherwise
- `AcceptedCmp4` - 1 if customer accepted the offer in the 4th campaign, 0 otherwise
- `AcceptedCmp5` - 1 if customer accepted the offer in the 5th campaign, 0 otherwise
- `Response (target)` - 1 if customer accepted the offer in the last campaign, 0 otherwise
- `Complain` - 1 if customer complained in the previous 2 years

### Customer Information

- `ID` - Customer's id
- `Year_Birth` - Customer's year of birth
- `Education` - customer's level of education
- `Marital_Status` - customer's marital status
- `Income` - customer's yearly household income
- `Kidhome` - number of small children in customer's household
- `Teenhome` - number of teenagers in customer's household
- `DtCustomer` - date of customer's enrolment with the company
- `Recency` - number of days since the last purchase

### Sales Product Type

- `MntCoke` - amount spent on coke products
- `MntFishProducts` - amount spent on fish products
- `MntMeatProducts` - amount spent on meat products
- `MntFruits` - amount spent on fruits products
- `MntSweetProducts` - amount spent on sweet products
- `MntGoldProds` - amount spent on gold products

### Number of Purchases Type

- `NumDealsPurchases` - number of purchases made with discount
- `NumWebPurchases` - number of purchases made through company's web site
- `NumCatalogPurchases` - number of purchases made using catalogue
- `NumStorePurchases` - number of purchases made directly in stores
- `NumWebVisitsMonth` - number of visits to company's web site in the last month

### Cost and Revenue

- `Z_CostContact` = 3 (Cost to contact a customer)
- `Z_Revenue` = 11 (Revenue after client accepting campaign)

Link dataset: [marketing\\_campaign\\_data.csv](#)

For more details, can click this link [jupyter notebook](#)



# Data Understanding

## Dataset Info

	Total Null Values	Percentage	Data Type
Income	24	1.071429	int64
ID	0	0.000000	int64
Z_CostContact	0	0.000000	int64
Complain	0	0.000000	object
AcceptedCmp2	0	0.000000	object
AcceptedCmp1	0	0.000000	float64
AcceptedCmp5	0	0.000000	int64
AcceptedCmp4	0	0.000000	int64
AcceptedCmp3	0	0.000000	object
NumWebVisitsMonth	0	0.000000	int64
NumStorePurchases	0	0.000000	int64
NumCatalogPurchases	0	0.000000	int64
NumWebPurchases	0	0.000000	int64
NumDealsPurchases	0	0.000000	int64
Z_Revenue	0	0.000000	int64
MntGoldProds	0	0.000000	int64
MntFishProducts	0	0.000000	int64
MntMeatProducts	0	0.000000	int64
MntFruits	0	0.000000	int64
MntCoke	0	0.000000	int64
Recency	0	0.000000	int64
Dt_Customer	0	0.000000	int64
Teenhome	0	0.000000	int64
Kidhome	0	0.000000	int64
Marital_Status	0	0.000000	int64
Education	0	0.000000	int64
Year_Birth	0	0.000000	int64
MntSweetProducts	0	0.000000	int64
Response	0	0.000000	int64

## Check Duplicate

Data Frame Dimension Before Duplicate Removal: (2240, 29)  
Data Frame Dimension After Duplicate Removal: (2240, 29)

There is no duplicated row

## Handling Missing Value

using imputation with Median, due to Highly Positively Skewed

## Feature Engineering

- Age
- Age Group
- Total Dependents
- Customer Has Children
- Total Spending
- Total Transactions
- Total Accepted Campaign (Accepted Campaign 1-5)
- Ever Accepted Campaign (Accepted Campaign at least 1x)
- Conversion Rate
- Income Segmentation
- Recency Segmentation

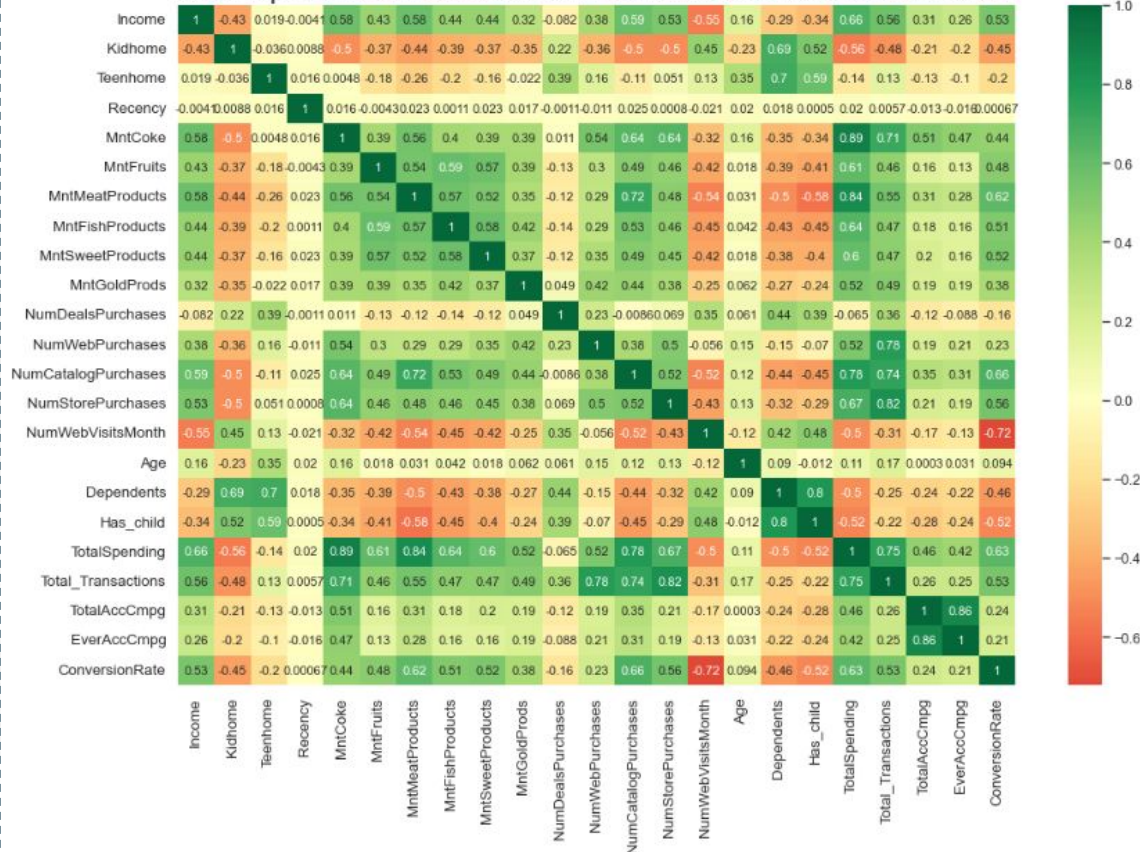
### Observations:

- Dataset has 29 columns and 2240 rows data
- There are 3 types of data types namely: int64, object, float64
- **Income** column has 2216 non-null values, and **24 null / missing values**

For more details, can click this link [jupyter notebook](#)

# Exploration Data Analysis (EDA)

Heatmap Correlation between Conversion Rate and related Attributes

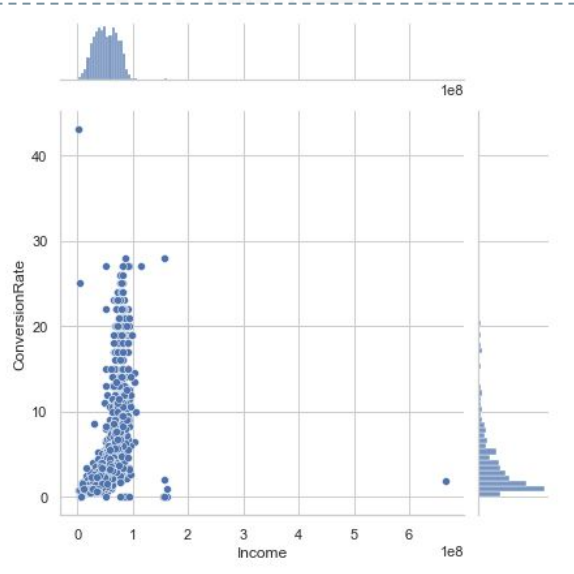


Based on Heatmap Visualization, the following information is obtained:

- Top 10 attributes that have strong correlation with Conversion Rate (the coefficient value  $\geq 0.5$ ):
  - NumWebVisitsMonth - 0.72 - Negatif
  - NumCatalogPurchases - 0.66 - Positif
  - TotalSpending - 0.62 - Positif
  - MntMeatProducts - 0.62 - Positif
  - NumStorePurchases - 0.55 - Positif
  - Total\_Transactions - 0.53 - Positif
  - Income - 0.53 - Positif
  - MntSweetProducts - 0.52 - Positif
  - Has\_child - 0.51 - Negatif
  - MntFishProducts - 0.51 - Positif
- In addition, there are several attributes that have a fairly strong correlation with the Conversion Rate (the coefficientlies between 0.30 and 0.49):
  - MntFruits - 0.48 - Positif
  - Dependents - 0.46 - Negatif
  - Kidhome - 0.44 - Negatif
  - MntCoke - 0.44 - Positif
  - MntGoldProds - 0.37 - Positif

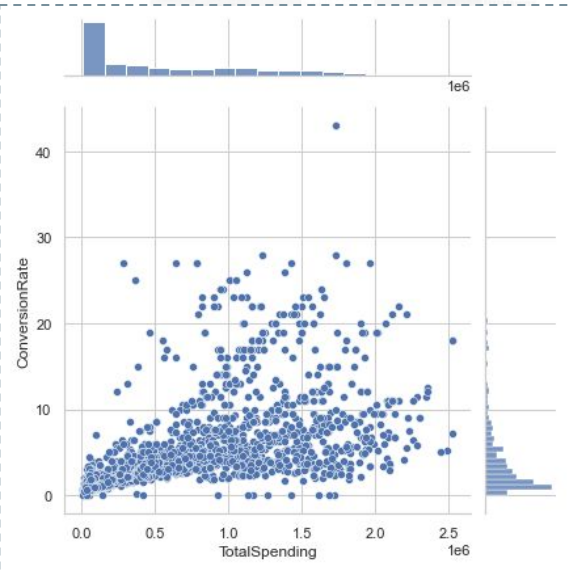
# Conversion Rate Analysis Based on Income, Spending and Age

## Conversion Rate VS Income



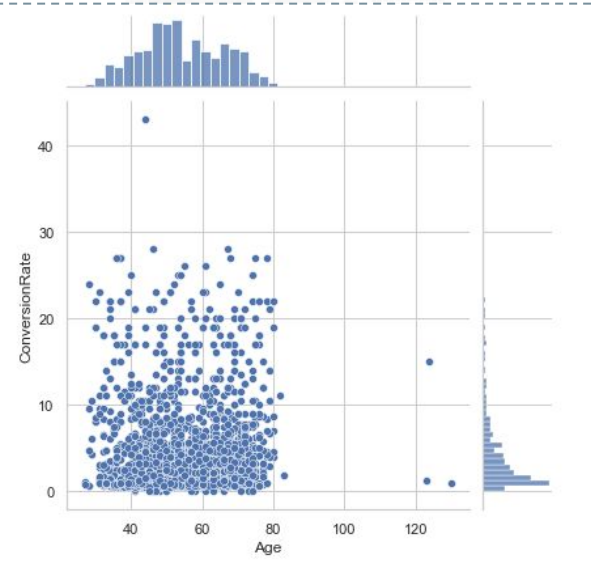
Graph above shows correlation between Conversion Rate and Income has strong **positive correlation**. It can be seen by the data points lie together closer to make a line.

## Conversion Rate VS Spending



Graph above shows correlation between Conversion Rate and Spending has strong **positive correlation**. It can be seen by the data points piles up at first on the left side, but increasingly to the right side data points becomes more spread out.

## Conversion Rate VS Age



Graph above shows correlation between Conversion Rate and Age **does not have correlation**. It can be seen by the data points are spread out and there are no trend to the data.



## Insights:

Based on the Heatmap data visualization, it shows that there are several attributes which have strong positive correlation with the Conversion Rate, including TotalSpending, Total\_Transactions, Income, NumCatalogPurchases and NumStorePurchases, MntMeatProducts, MntSweetProducts, MntFishProducts, MntFruits, MntCoke and MntGoldProds.

## Recommendations:

From information above, based on attributes which have strong positive correlation with the conversion rate, 3 conclusions can be drawn:

- Based on customer behavior, there are 3 attributes that most influence the conversion rate, namely TotalSpending, Total\_Transactions and Income. Therefore, further marketing campaigns can be implemented for customers who have high TotalSpending, Total\_Transactions and Income.
- Based on purchase type, it is recommended that more marketing campaigns be implemented through catalogs and shops.
- Based on the product type, it is recommended that the types of products promoted are Meat Products, Sweet Products, Fish Products, Fruit Products, Coke and Gold Products.



## Check Duplicate Rows

```
df.duplicated().sum()
```

```
0
```

```
df.duplicated(subset=["ID"]).sum()
```

```
0
```

```
print(f"Data Frame Dimension Before Duplicate Removal: {df.shape}")  
df = df.drop_duplicates().reset_index(drop=True)  
print(f"Data Frame Dimension After Duplicate Removal: {df.shape}")
```

```
Data Frame Dimension Before Duplicate Removal: (2240, 40)
```

```
Data Frame Dimension After Duplicate Removal: (2240, 40)
```

There is no duplicated row

## Handling Missing Value

## Check Missing Value

Missing values status: True

	Total Null Values	Percentage	Data Type
Income	24	1.071429	object
ID	0	0.000000	int64
AcceptedCmp5	0	0.000000	int64
AcceptedCmp1	0	0.000000	object
AcceptedCmp2	0	0.000000	object
Complain	0	0.000000	float64
Z_CostContact	0	0.000000	int64
Z_Revenue	0	0.000000	int64
Response	0	0.000000	object

### Observations:

- Income column has 24 null / missing values, a proportion of 1.07% of the total data

```
print("Number of missing values in Income column before Imputation =", df["Income"].isna().sum())  
df['Income'].fillna(df['Income'].median(), inplace=True)  
print("Number of missing values in Income column after Imputation =", df["Income"].isna().sum())
```

```
Number of missing values in Income column before Imputation = 24
```

```
Number of missing values in Income column after Imputation = 0
```

For handling missing values for the Income column, using imputation with Median, due to Highly Positively Skewed

Missing values in Income column has been handled successfully, proven by there is no null values after missing value handling.

## Drop Unnecessary Feature

Following are some of the columns that need to be dropped:

- Drop `ID` column as it has multiple categories and is not useful for modeling
- Drop `Year_Birth` column because feature extraction has been done to retrieve `Age` data
- Drop `Dt_Customer` column because it is not needed for modeling
- Drop column `Z_CostContact`, `Z_Revenue` because it only has one value, it doesn't provide significant information to the prediction model
- Drop `Complain` column because it is not needed for modeling
- Drop the `Kidhome` and `Teenhome` columns because the `Has_child` and `Dependents` features already exist to replace them
- Drop the `AcceptedCmp1-5` and `response` fields because there are engineering features `TotalAccCmpg` and `ConversionRate` as replacement features

## Feature Encoding

### Observations:

Based on the result above, there are 5 columns need to be changed using Label Encoding:

- Education => 'S1', 'S3', 'S2', 'SMA', 'D3'
- Age\_group => 'Elder', 'Adult'
- Income\_sgmt => 'Medium', 'High', 'Low', 'None'
- Recency\_Sgmt=> '1 Week - >12 Weeks'
- Marital\_Status, need to be simplified

Reason why all string variable type using Label Encoding is to reduce features addition.

Before performing Label Encoding, it is necessary to simplify values that have the same meaning:

- Single = 'Lajang'
- Couple = 'Bertunangan', 'Menikah'
- Divorce = 'Cerai', 'Janda', 'Duda'

### After Feature Encoding:

	Education	Marital_Status	Age_group	Income_sgmt	Recency_Sgmt
381	2	2	0	1	10
1805	2	2	1	2	3
1544	2	0	1	2	12
110	2	2	0	3	6
1853	2	2	1	2	0
619	3	2	1	1	0
102	2	2	0	3	8
131	3	0	1	2	6
1230	2	0	1	2	7
1791	4	1	1	2	3

## Feature Scaling (Standardization)

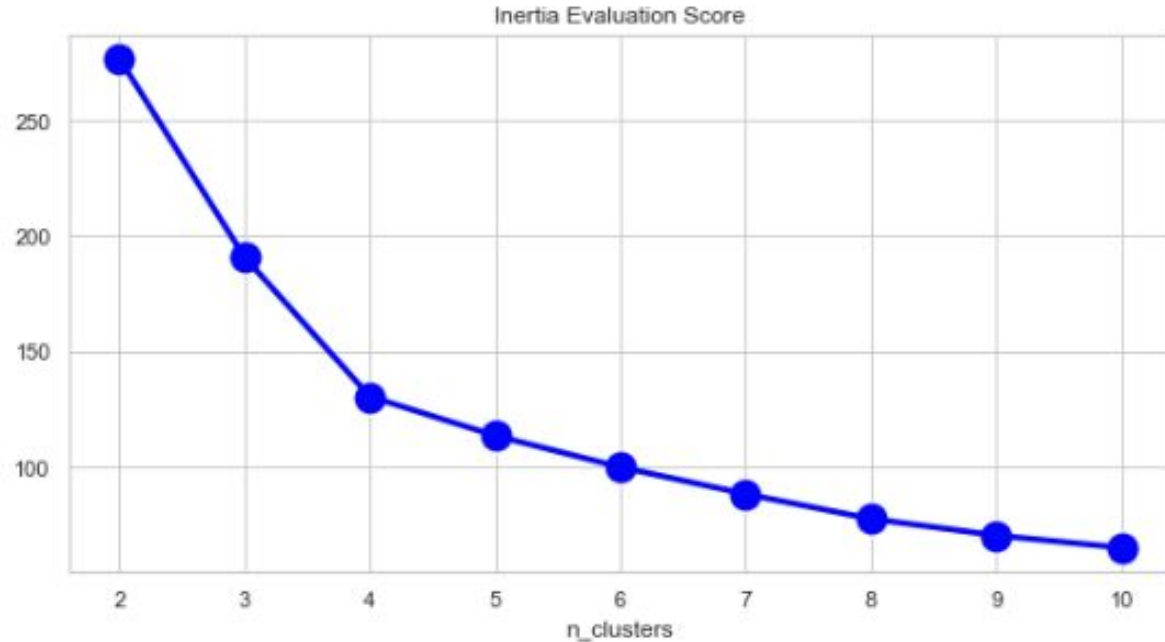
```
# standardization
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
scaler.fit(df)
df_std = pd.DataFrame(scaler.transform(df), columns= df.columns )
```

### After Scaling:

Education	Marital_Status	Income	Recency	MntCoke	MntFruits	MntMeatProducts
-0.458383	-0.695118	0.235696	0.307039	0.983781	1.551577	1.679702
-0.458383	-0.695118	-0.235454	-0.383664	-0.870479	-0.636301	-0.713225
-0.458383	0.680380	0.773999	-0.798086	0.362723	0.570804	-0.177032
-0.458383	0.680380	-1.022355	-0.798086	-0.870479	-0.560857	-0.651187
1.533425	0.680380	0.241888	1.550305	-0.389085	0.419916	-0.216914



## Elbow Method



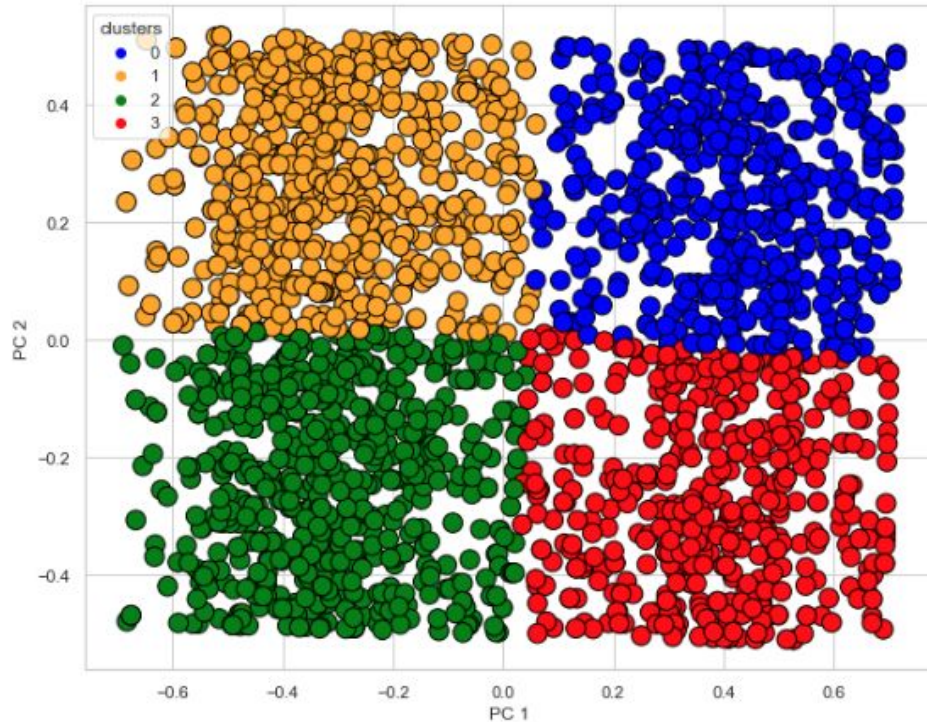
## Observation:

To find optimal number of clusters, can use Elbow Method model evaluation on inertia score.

Based on the Elbow Method results,  $n_{\text{cluster}} = 4$  will be used to divide the cluster because the inertia score is not change significantly.

## K Means Clustering

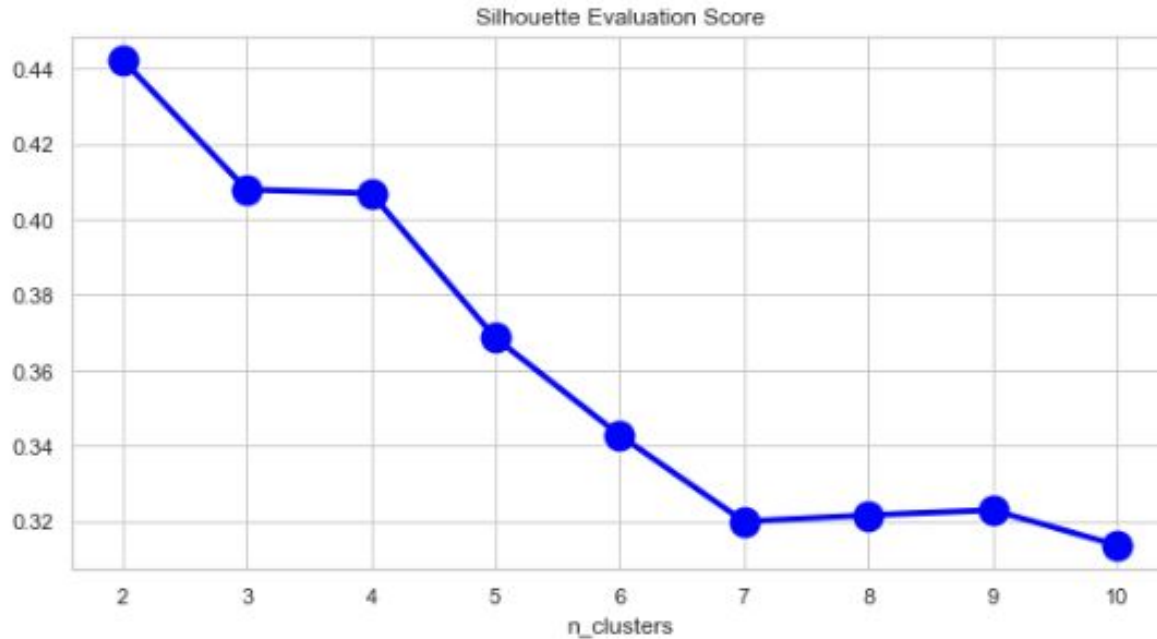
2-D Visualization of Customer Clusters  
With PCA



### Observation:

Based on the visualization, there is clearly 4 clusters that generated by K-Means Clustering algorithm using RFM Method for this dataset

## Silhouette Score

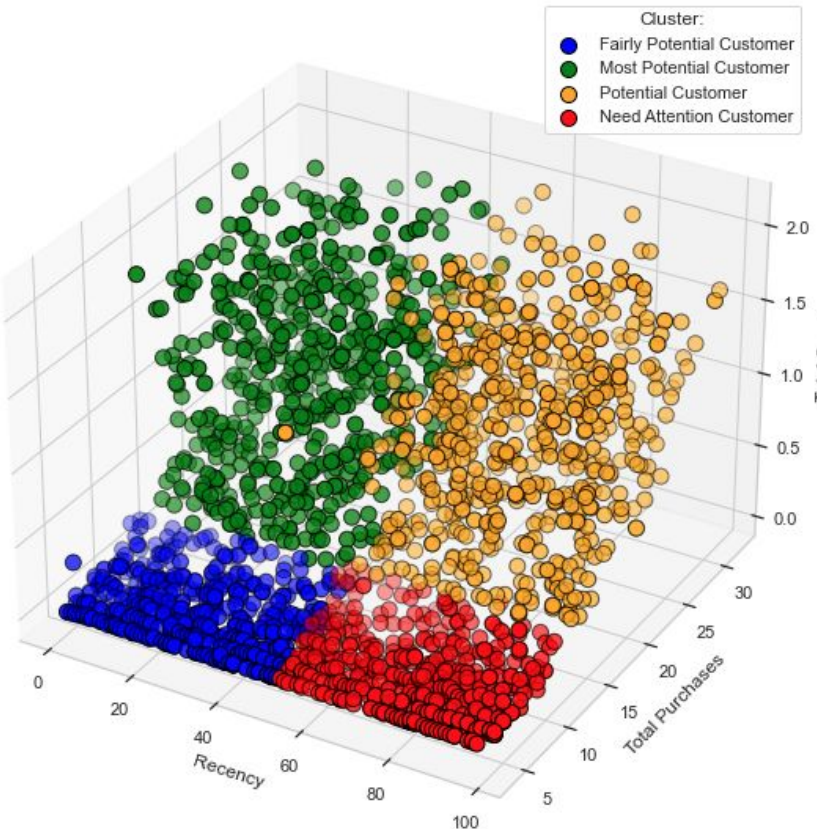


## Observation:

Based on that visualization,  $n\_cluster=4$  is an optimal number for K-Means Clustering in this dataset



3-D Visualization of Customer Clusters  
Based on it's Characteristics



## Observation:

Customer behavior based on Recency, Total Transactions/Purchases and Total Spending.

### 4 Clusters:

- Fairly Potential Customer (Cluster 0)
- Most Potential Customer (Cluster 1)
- Potential Customer (Cluster 2)
- Need Attention Customer (Cluster 3)

## Insights

### 1. Fairly Potential Customer (Cluster 0)

- There are 494 customers (22.05% of total customers) in this group
- Customers in this group have average recency 24 days, average of total transactions 7 items and average total spending money 76,469
- This group dominated by Adult customers 11.47% (20-50 years old) have average income around 35,433,907

### 2. Most Potential Customer (Cluster 1)

- There are 617 customers (27.54% of total customers) in this group
- Customers in this group have average recency 23 days, average of total transactions 20 items and average total spending money 996,356
- This group dominated by Elder customers 17.07% (>50 years old) have average income around 65,403,923

### 3. Potential Customer (Cluster 2)

- There are 650 customers (29.02% of total customers) in this group
- Customers in this group have average recency 73 days, average of total transactions 20 items and average total spending money 1,019,693
- This group dominated by Elder customers 19.15% (>50 years old) have average income around 65,133,940

### 4. Need Attention Customer (Cluster 3)

- There are 479 customers (21.38% of total customers) in this group
- Customers in this group have average recency 74 days, average of total transactions 7 items and average total spending money 86,970
- This group dominated by Elder customers 11.4% (>50 years old) have average income around 35,109,462

## Recommendations

### 1. Fairly Potential Customer (Cluster 0)

- Characteristic: Very active for shopping, quite high shopping intensity, spending quite much of money
- Recommendations: Give Discount/Flash Sale, Promo Bundling/Special Offer, Buy 1 Get 1 strategy
  - Example: If customers buy Coke Product with this promo, so they will get 1 free Coke

### 2. Most Potential Customer (Cluster 1)

- Characteristic: Very active for shopping, spending quite much of money, very high shopping intensity
- Recommendations: Vouchers/Rewards, Promo Bundling/Special Offer, Mix & Match bundling strategy
  - Example: Free to choose Coke from 3 different types/brands if meet the minimum purchase

### 3. Potential Customer (Cluster 2)

- Characteristic: Active for shopping, spending very much of money, very high shopping intensity
- Recommendations: Vouchers/Rewards, Promo Bundling/Special Offer with Cross Selling bundling
  - Example: Every purchase of Coke & Meat Products at the same time can get a 10% cheaper price

### 4. Need Attention Customer (Cluster 3)

- Characteristic: Quite active for shopping, spending quite much of money, quite high shopping intensity
- Recommendations: Discount/Flash Sale, Promo Bundling/Special Offer, Buy 1 Get 1 and Limited Edition Bundling (giving urgency sense to purchase)
  - Example: Get Coke Product promo with 55% discount at 9 AM - 11 AM only