Predict Customer Personality to boost marketing campaign by using Machine Learning

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

Rakamin Academy

Created by:
Dinda Galuh
dindagaluhg@gmail.com
linkedin.com/in/dinda-galuh-guminta

Statistics graduate with passion in data visualization, data analysis, and reporting. Dedicated and hard working person.

I joined the data science bootcamp because I am more interested and have a passion for learning and working in the data field. I like doing data analysis and data visualization. I want to apply my knowledge in the real world.

Business Understanding

Problem Statement:

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..

SuperStore will create a marketing campaign to increase the number of transactions.

1. The marketing campaign costs of $3 per customer proved to be a significant burden.
2. The response rate is only 14.9% from a total of 2240 subscribers. As a result, the revenue generated is still a loss of $ 3,046 from the cost of $ 6,720 camping.

This loss burdens the company and requires handling so that the marketing campaign is more efficient & profitable.

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 the focus of the analysis is to create a cluster prediction model to make it easier for companies to make decisions.

Goals:

Increase the **effectiveness of marketing campaigns** by targeting the right customers so that **response rates and profits increase**.

Objective:

Customer cluster model predictions

Business Metrics:

- Response rate
- Profit

Dataset Marketing Campaign

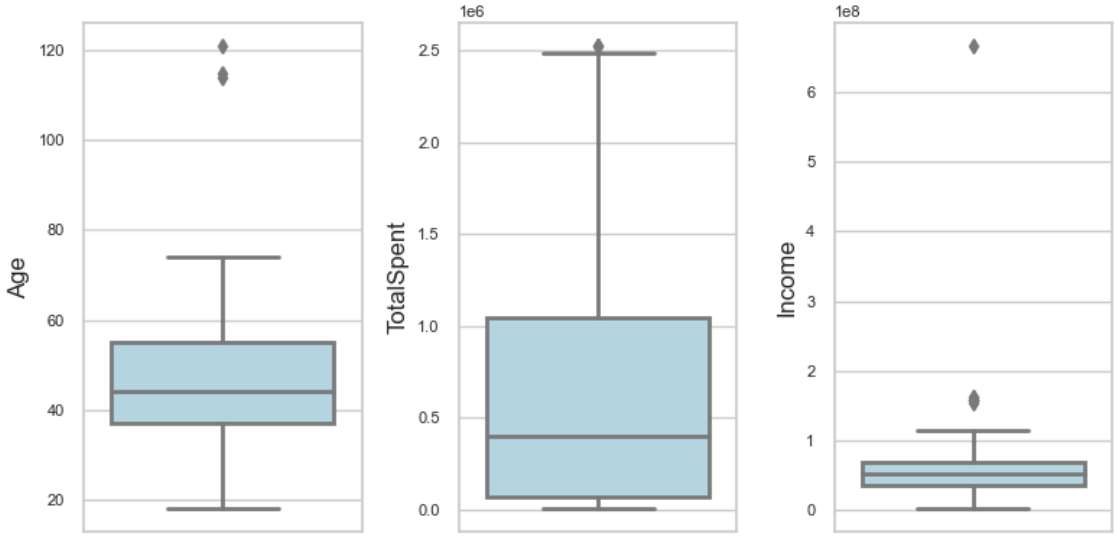
Rakamin Academy

Variable	Description	Type
ID	customer's ID	Numerical
Year_Birth	customer's year of birth	Numerical
DtCustomer	date of customer's enrolment with the company	Date
Education	customer's level of education	Categorical
Marital_Status	customer's marital status	Categorical
Kidhome	number of small children in customer's household	Numerical
Teenhome	number of teenagers in customer's household	Numerical
Income	customer's yearly household income	Numerical
MntFishProducts	amount spent on fish products in the last 2 years	Numerical
MntMeatProducts	amount spent on meat products in the last 2 years	Numerical
MntFruits	amount spent on fruits products in the last 2 years	Numerical
MntSweetProducts	amount spent on sweet products in the last 2 years	Numerical
MntCoke	amount spent on coke products in the last 2 years	Numerical
MntGoldProds	amount spent on gold products in the last 2 years	Numerical
NumDealsPurchases	number of purchases made with discount	Numerical
NumCatalogPurchases	number of purchases made using catalogue	Numerical
NumStorePurchases	number of purchases made directly in stores	Numerical
NumWebPurchases	number of purchases made through company's web site	Numerical
NumWebVisitsMonth	number of visits to company's web site in the last month	Numerical
Recency	number of days since the last purchase	Numerical
Z_CostContact	cost to contact the costumer	Numerical
Z_Revenue	revenue generated by the costumer	Numerical
AcceptedCmp1	1 if customer accepted the offer in the 1st campaign, 0 otherwise	Categorical
AcceptedCmp2	1 if customer accepted the offer in the 2nd campaign, 0 otherwise	Categorical
AcceptedCmp3	1 if customer accepted the offer in the 3rd campaign, 0 otherwise	Categorical
AcceptedCmp4	1 if customer accepted the offer in the 4th campaign, 0 otherwise	Categorical
AcceptedCmp5	1 if customer accepted the offer in the 5th campaign, 0 otherwise	Categorical
Complain	1 if customer complained in the last 2 years	Categorical
Response (target)	1 if customer accepted the offer in the last campaign, 0 otherwise	Categorical

Row: 2.240 customer

Column: 29 feature

Distribution of customer data based on age, income, and spending



It can be seen that the features of age, spending, and income have a positively skewed pattern and have outlier values on the right side.

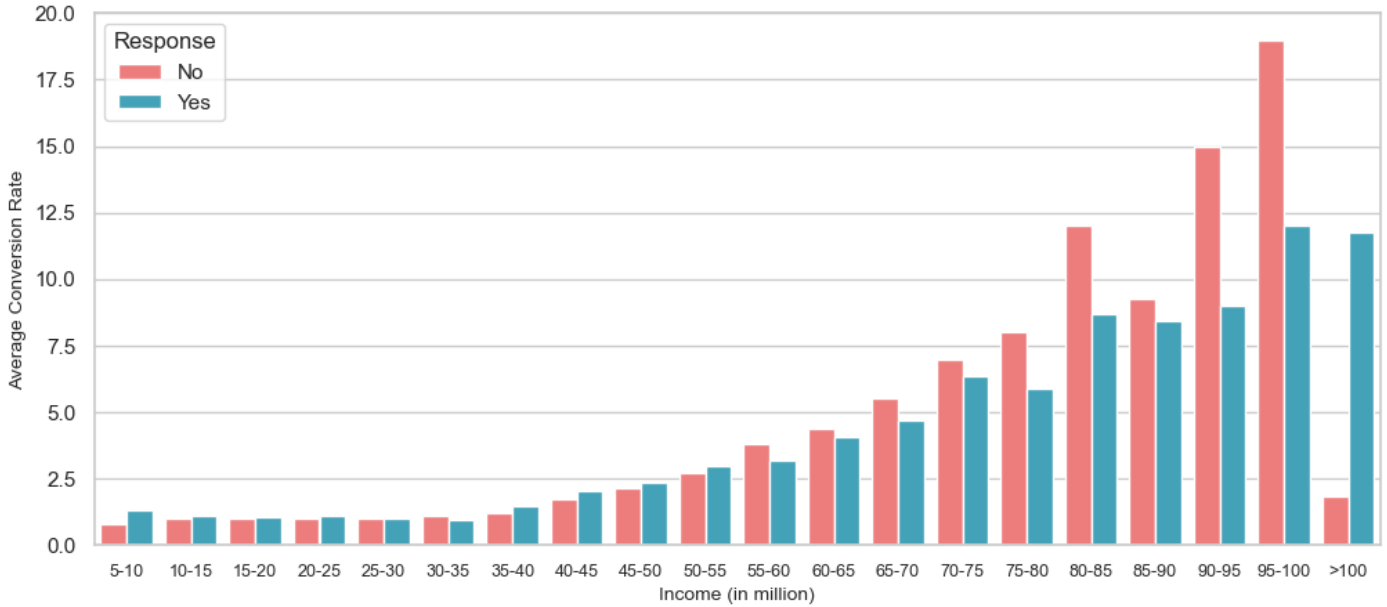
This can happen because most customers have a low age, spending and income, but there are some that are very high.

Conversion Rate Analysis Based on Income

Rakamin Academy

Higher Incomes, Higher Purchases

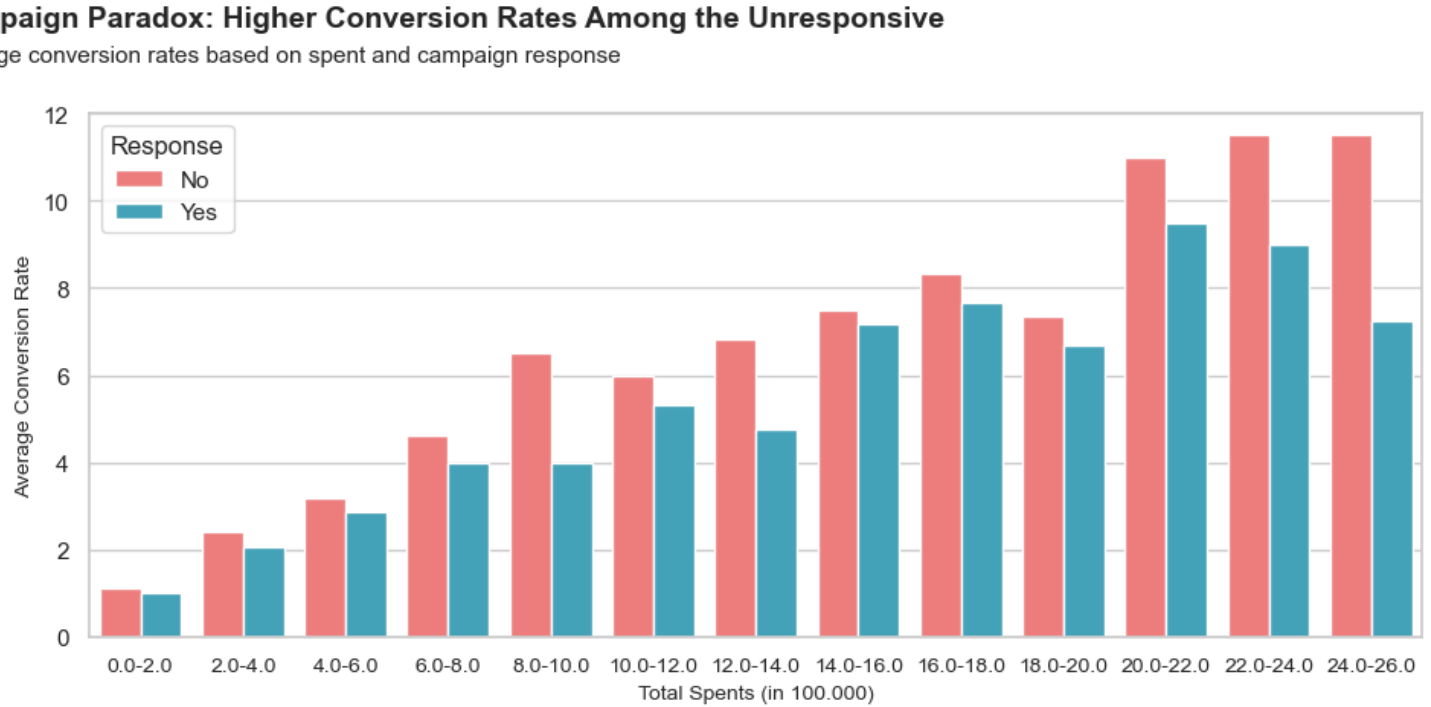
Average conversion rates based on income and campaign response



Conversion rates tend to go up as revenue increases, for both customers who respond to campaigns and those who don't. This suggests that customers with higher incomes are more likely to make purchases.

- For most income groups, the conversion rate of customers who respond to campaigns tends to be higher than those who don't. This is an expected result, as the campaign is designed to encourage purchases..
- However, for the income group of 55 million and above, the conversion rate of customers who do not respond to campaigns is even higher. This may be because these customers do not have much freedom to make impulse purchases and are more selective in their purchases.
- **Recommendation:** target marketing to customers with income > 55 million with a conversion rate > 10%

Conversion Rate Analysis Based on Spending



- The **average conversion rate of customers who don't respond to campaigns is always higher than those who do**. This is quite unusual in that a higher conversion rate is expected from customers responding to the campaign. Maybe there are other factors that affect this or it could be because the campaign is not effective in increasing conversions.
- For both responding and non-responding customers, it appears that conversion rates generally increase as spending increases. That is, **the more money customers spend, the more likely they are to convert**.

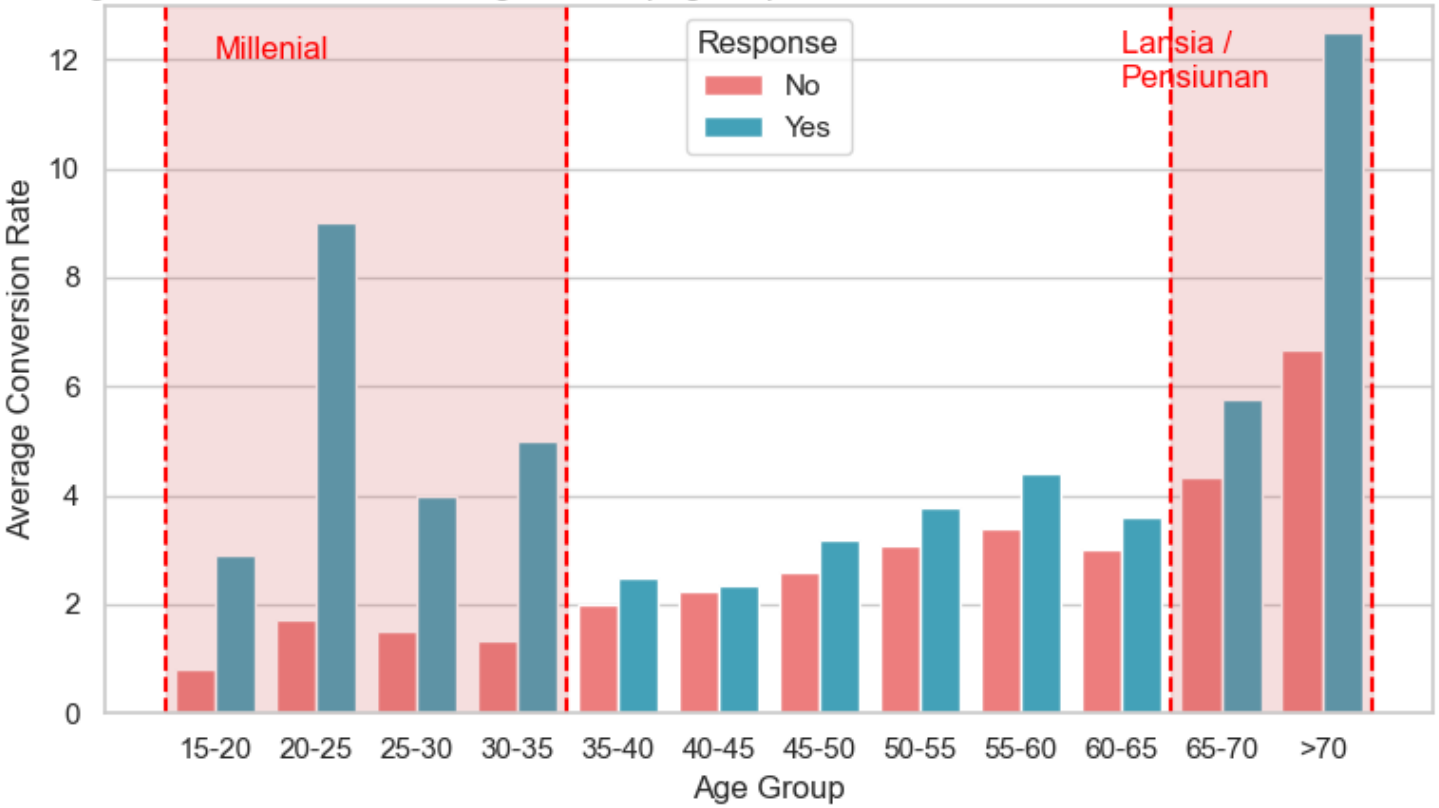
Although conversion rates were higher for customers who weren't responding to campaigns, **the increase in conversion rates appeared to be more steady for customers who did respond to campaigns**. For example, in the spending group 0 to 200,000, their conversion rate is 1.0, and it increases to 2,875 in the 400,000-600,000 group, and peaks at 9.5 in the 2,000,000-2,200,000 group.

Conversion Rate Analysis Based on Age

Rakamin Academy

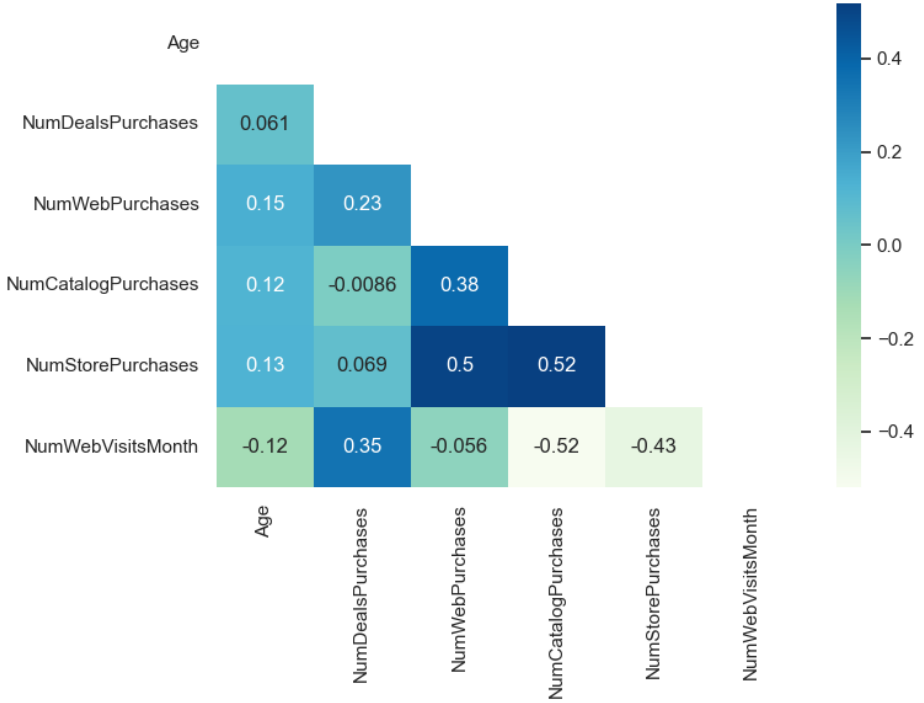
Engaged Audiences: High Conversion Rates among Millennials & Elderly

Average conversion rates based on age and campaign response



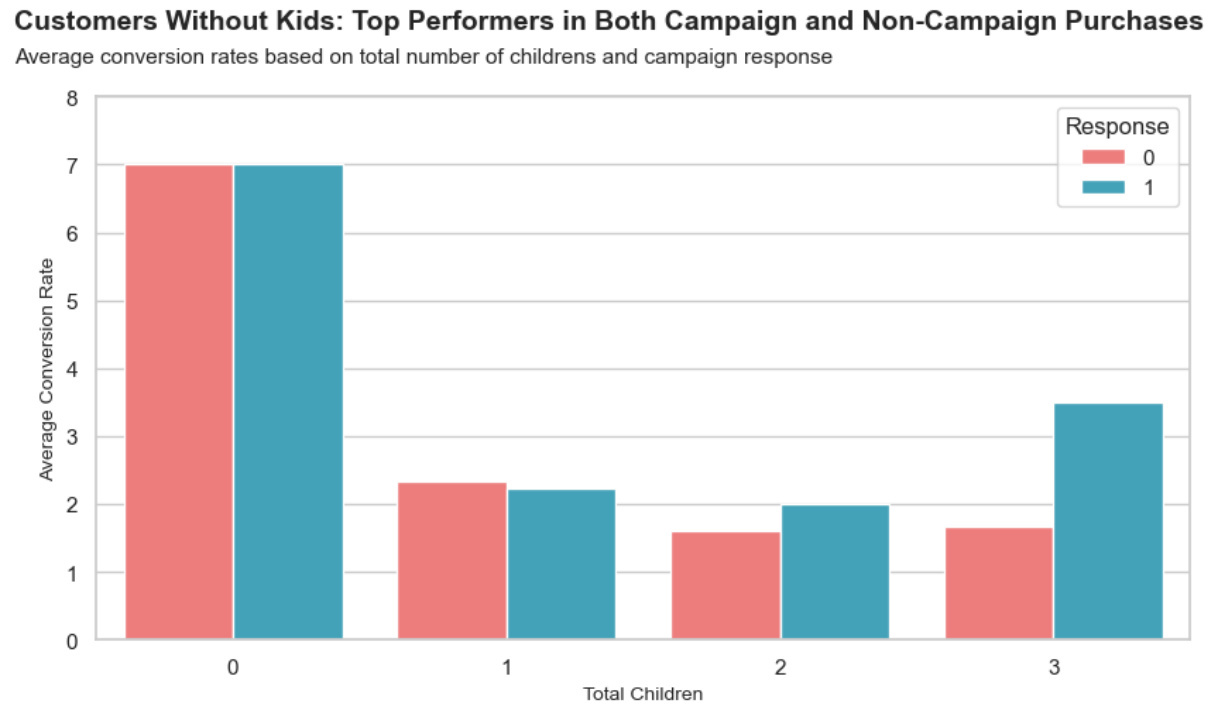
- Conversion rates tend to increase as customers age, both for those who respond and those who don't.
- Customers aged 15-35 and 65 years and over who respond to campaigns have a high conversion rate and are quite significantly different when compared to customers who do not respond to campaigns. This shows that this demographic group is a potential target for future campaigns.
- A high conversion rate in the 20-35 age group can occur because this group is closer to technology and has a lot of free time to respond to campaigns. In general, the >65 year old group is dominated by retirees who have a lot of free time to respond to campaigns and finally make purchases.

Correlation Between Age and Purchase



Age and purchases through websites and catalogs have a positive relationship so that they can support the high conversion rate of millennials, namely 15-35. This customer group responds to campaigns due to the trend towards using technology which makes this group more accessible to campaigns and more likely to respond through purchases.

Conversion Rate Analysis Based on Total Children



- Customers without children have the same conversion rate (7.0) for both responding and non-responding campaigns. This could indicate that their buying decision was less influenced by the campaign, or it could indicate that the campaign was equally effective in both groups, regardless of whether they actively responded or not.
- For customers with 2 or more children, their conversion rate is higher if they respond to the campaign, compared to those who don't. This shows that the campaign has a positive effect on the buying decision of this group.

In general, it appears that marketing campaigns can influence conversion rates, but the degree of influence varies depending on the number of children a customer has.

Data Preprocessing

Rakamin Academy

Check missing values & duplicated row

Feature Selection

Feature Encoding

Feature Standardization

```
5   Income          2216 non-null   float64
```

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

```
0
```

- Missing value in the income column is filled with the median value
- There are no duplicate rows

```
df_prep = df[['ID', 'Education', 'Marital_Status','Income','Age','TotalChildren','IsParent','Recency','CustomerDays',
              'MntCoke','MntFruits', 'MntMeatProducts','MntFishProducts','MntSweetProducts', 'MntGoldProds',
              'TotalSpent','TotalCampaign','NumPurchase','ConversionRate','Complain','Response',
              'NumDealsPurchases','NumWebPurchases', 'NumCatalogPurchases','NumStorePurchases','NumWebVisitsMonth']]
```

```
# feature encoding pada education dan marital status
# label encoder pada Education
df_prep['Education'] = df_prep['Education'].astype('category').cat.codes

# one hot encoding pada marital status karena tidak memiliki urutan
status_df_prep = pd.get_dummies(df_prep['Marital_Status'], prefix='Status')
df_prep = df_prep.join(status_df_prep)
```

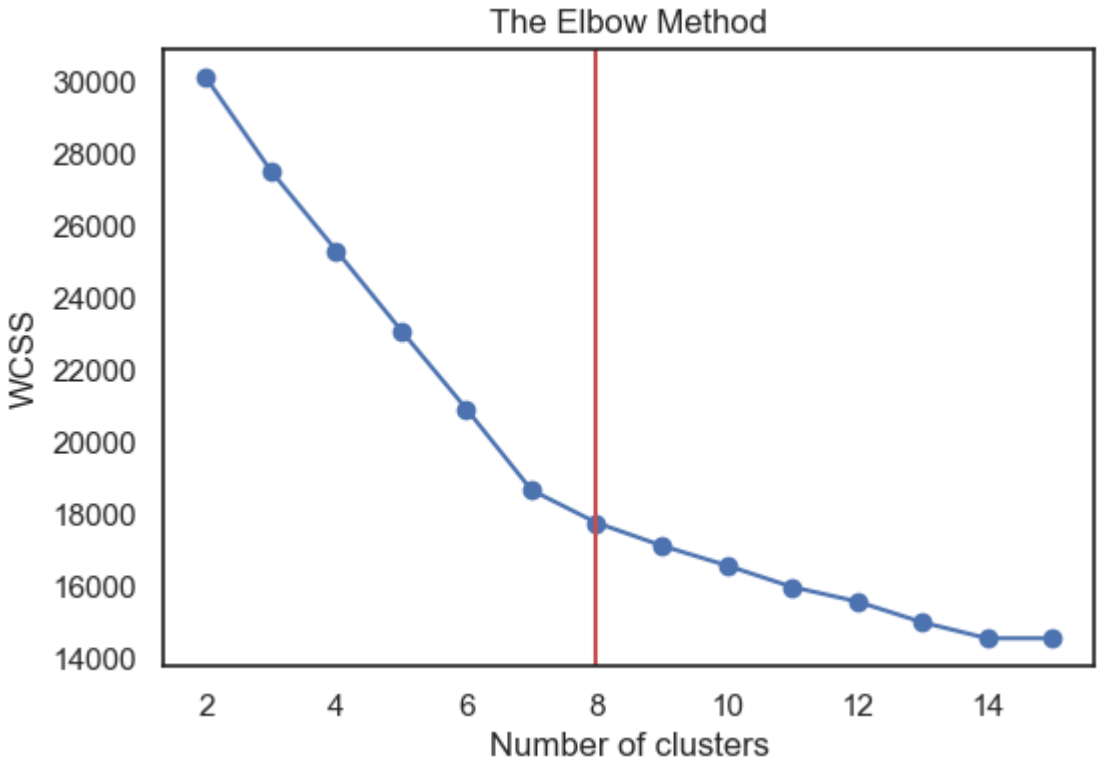
```
nums = ['Income', 'Age', 'TotalChildren', 'Recency', 'CustomerDays', 'MntCoke', 'MntFruits',
        'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts','MntGoldProds', 'TotalSpent', 'TotalCampaign', 'NumPurchase',
        'ConversionRate', 'NumDealsPurchases','NumWebPurchases', 'NumCatalogPurchases', 'NumStorePurchases','NumWebVisitsMonth']
```

```
from sklearn.preprocessing import StandardScaler
ss = StandardScaler()

for n in nums:
  scaler = ss.fit(df_prep[[n]])
  df_prep[n] = scaler.transform(df_prep[[n]])
```

for details, see jupyter notebook here

Data Modeling



The elbow method is a technique used in determining the optimal number of clusters in the k-means clustering algorithm. Here are the steps for the elbow method :

1. Running K-Means Clustering. For example, the number of clusters is between 2 and 15.
2. Calculates Within-Cluster-Sum-of-Squares (WCSS): the sum of the squares of the distance between each cluster member and its cluster center.

The number of clusters is said to be optimum when the decrease in WCSS starts to slow down significantly and becomes flat (this point is often called elbow).

- The decrease between 7 and 8 clusters, and between 8 and 9 clusters, was quite significant, but after that the decline was not that big.
- So, based on the elbow method, we might choose between 7 and 9 as the optimal number of clusters.
- **In this analysis, the optimum number of clusters was selected, namely 8 clusters**

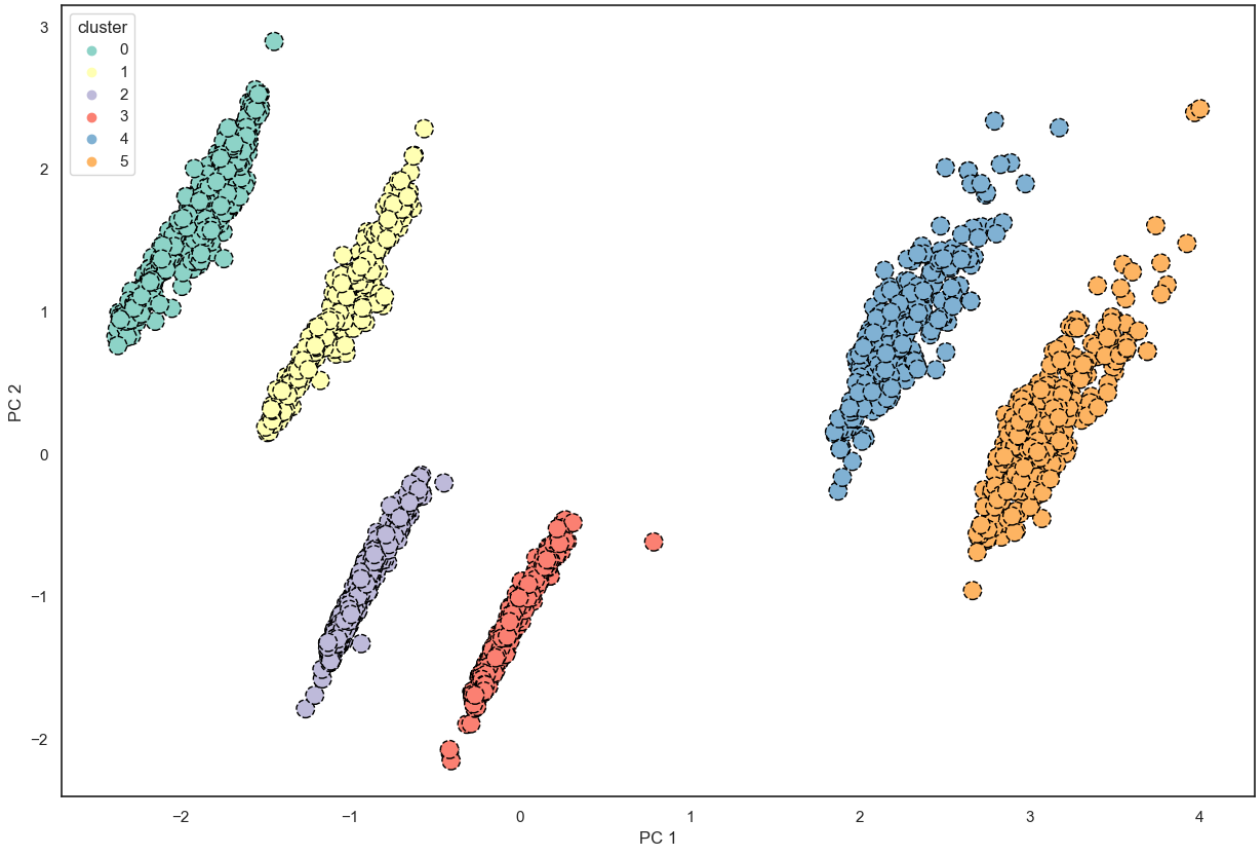
Cluster Evaluation using Silhouette Score

```
For n_clusters = 2. The average silhouette_score is : 0.12948997700460652
For n_clusters = 3. The average silhouette_score is : 0.05489556939071352
For n_clusters = 4. The average silhouette_score is : 0.08750161609581243
For n_clusters = 5. The average silhouette_score is : 0.08407095893330167
For n_clusters = 6. The average silhouette_score is : 0.14193376023695667
For n_clusters = 7. The average silhouette_score is : 0.14640191077503129
For n_clusters = 8. The average silhouette_score is : 0.15252320101356376
For n_clusters = 9. The average silhouette_score is : 0.11569897849516105
For n_clusters = 10. The average silhouette_score is : 0.12398787575897074
For n_clusters = 11. The average silhouette_score is : 0.1311032673847169
For n_clusters = 12. The average silhouette_score is : 0.12237438114485819
For n_clusters = 13. The average silhouette_score is : 0.12411068249205684
For n_clusters = 14. The average silhouette_score is : 0.10532086716535298
For n_clusters = 15. The average silhouette_score is : 0.1280338138034653
```

It can be seen that the formation of 8 clusters produces the highest silhouette score

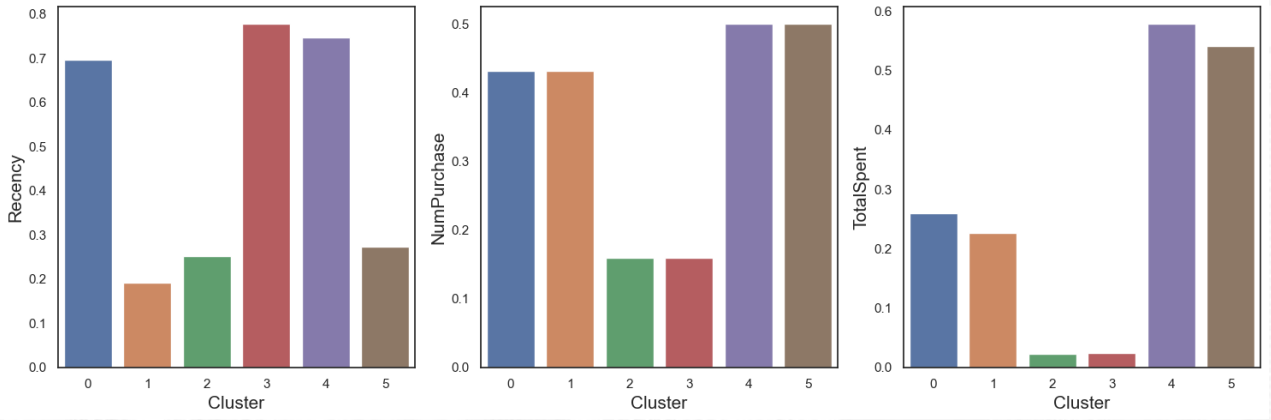
- Silhouette Coefficient values range between -1 and 1.
- Values close to 1 indicate that the sample is far from neighboring clusters. Values close to 0 indicate that the point is at or very close to the decision boundary between two adjacent clusters. Negative values generally indicate that the sample has been placed in the wrong cluster.
- **The optimum number of clusters is determined by the highest Silhouette Coefficient value**.

Clustering Result



There are 6 optimal clusters using the K-Means method

Cluster Characteristics



INSIGHT

- **Cluster 0: Churning Risk** - Customers in this cluster have a high risk of churn because they haven't interacted or made a purchase in a long time. Their purchase frequency is moderate, and their total spending is moderate.
- **Cluster 1: Steady Customers** - Customers in this cluster are consistent in making purchases, but they make purchases with moderate frequency and their total expenses are moderate. They are consistent shoppers but don't spend much.
- **Cluster 2: Casual Buyers** - Customers in this cluster may be casual buyers or new buyers who have not shopped in large quantities or frequently.
- **Cluster 3: Inactive Customers** - Customers in this cluster may be customers who are inactive or have stopped making purchases.
- **Cluster 4: High-Value Customers** - Customers in this cluster often shop and spend a lot of money. They are the most valuable customers.
- **Cluster 5: Active High Spenders** - Customers in this cluster have recently made high-frequency purchases, and their total spend has been high. They are the most active customers.

SUGGESTIONS

- **Cluster 0: Churning Risk** - focuses on customer retention strategies such as special offers, discounts or loyalty programs to revive purchase interest.
- **Cluster 1: Steady Customers** - cross-selling additional products and services.
- **Cluster 2: Casual Buyers** - provide relevant and attractive offers or discounts, or try to better understand their preferences and needs.
- **Cluster 3: Inactive Customers** - take a more personal approach to understanding why they are inactive and trying to persuade them to shop again. Customer satisfaction surveys, direct calls, or personal emails can be done.
- **Cluster 4: High-Value Customers** - try to maintain a good relationship with them, ensure that they feel valued, and try to better understand their needs and wants in order to serve them better.
- **Cluster 5: Active High Spenders** - strive to maintain good relationships and ensure that they remain satisfied. Can through good customer service, exclusive offers, or respond quickly and effectively to their feedback and questions