Predict Customer Personality Using K-Means

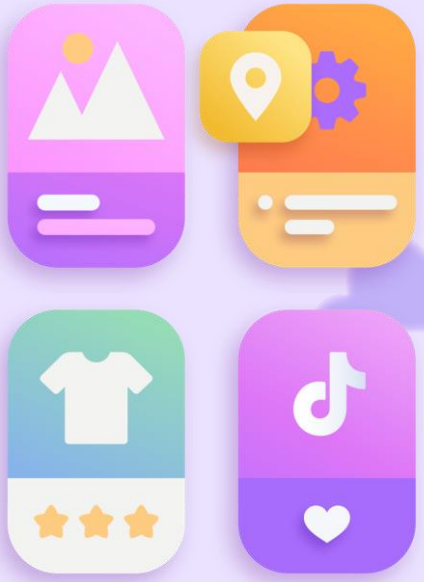
Presented By Radisha Fanni Sianti

About Me

“I am a fresh graduate of Master of Statistics and Bachelor of Mathematics from Institut Teknologi Sepuluh Nopember. I have a passion for continuing to learn and develop myself by constantly trying to broaden my knowledge in statistics and other related disciplines. I’m enthusiastic about deriving valuable insights from data and leveraging them to facilitate well-informed decision-making. ”

Radisha Fanni Sianti

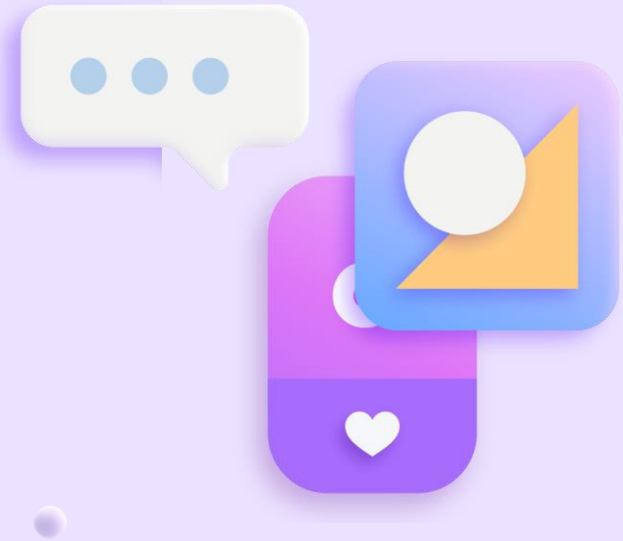
Table of Contents



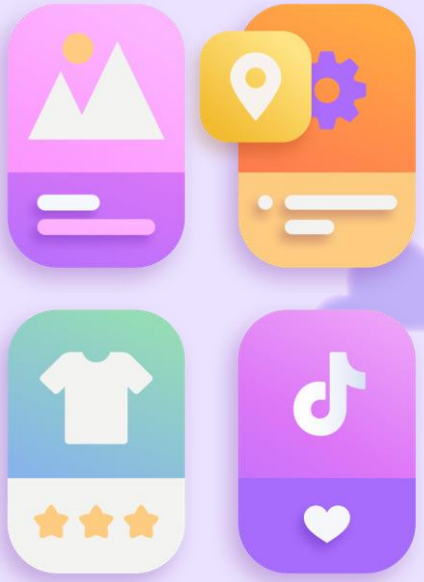
01

Overview

Overview



- Understanding customer behavior and personality is essential for a company's fast and lasting growth.
- Knowing each customer's unique traits enables better services and benefits, especially for potential loyal customers who significantly impact a company's success.
- In this project, we will create predictive customer clusters using their traits to enhance our services and cater to their needs effectively.



02

Data Preprocessing

Data

Marketing_campaign_data.csv

#	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

Data Preprocessing

01

Duplicated Data

There is no duplicated data

02

Missing Value

Deleting data that has a missing value of 1.071%

03

Data Type

Dt_Customer should be of the datetime data type

04

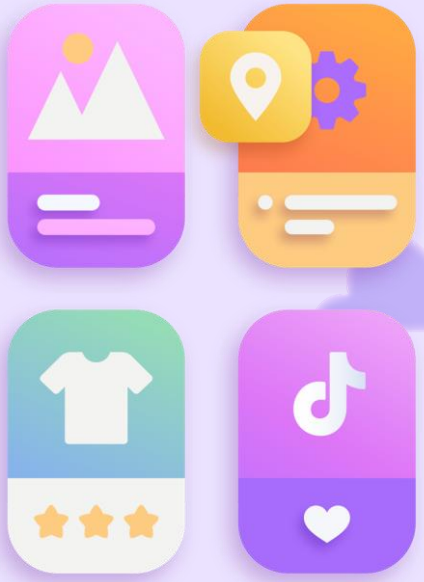
Feature Engineering

- TotalDaysJoined
- Age
- AgeGroup
- TotalChildren
- TotalAmountSpent
- TotalPurchases
- TotalAccepted
- ConversionRate

05

Feature Encoding

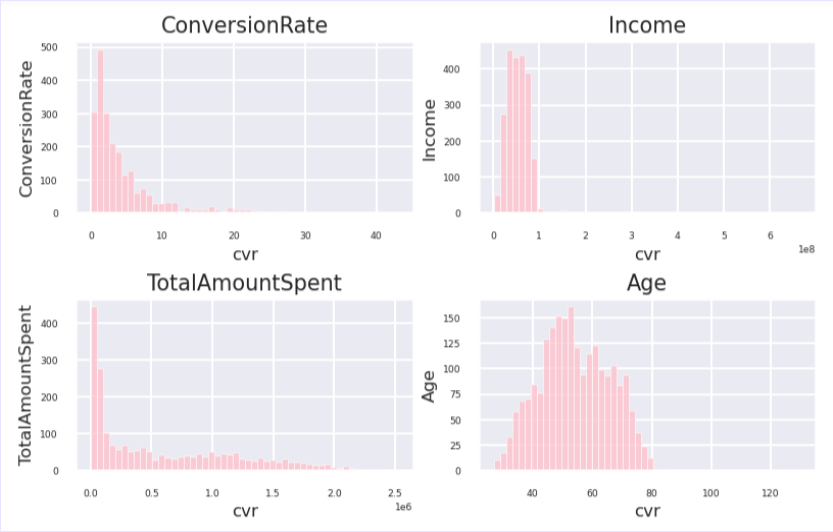
- Education
- Marital_Status
- AgGroup

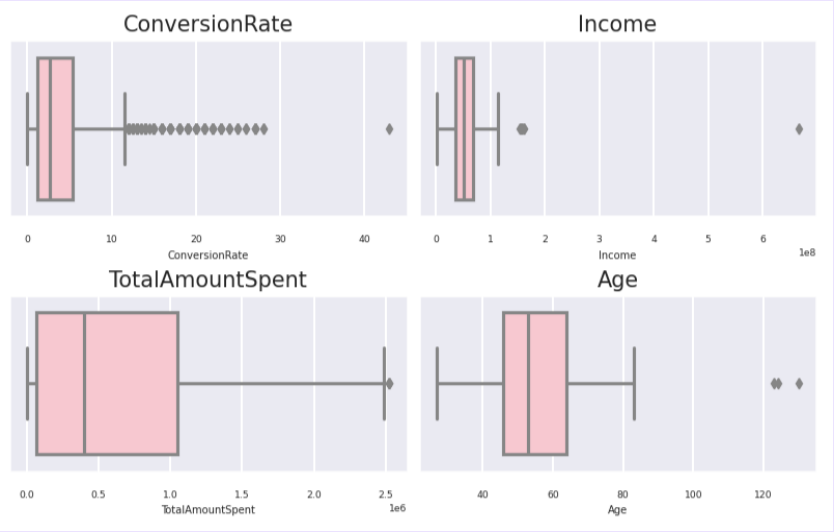


03

EDA

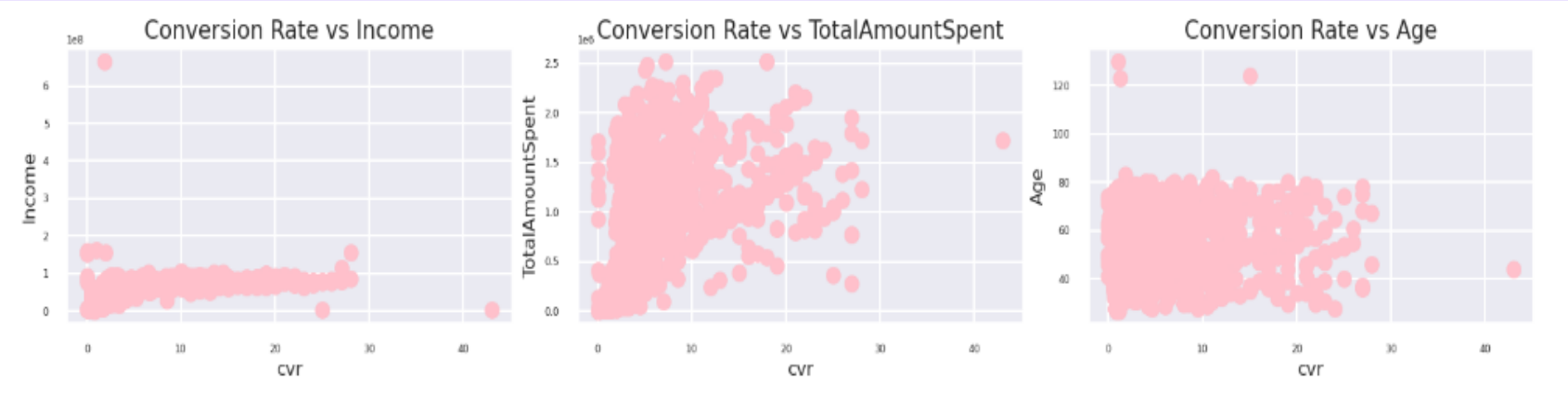
Univariate Analysis





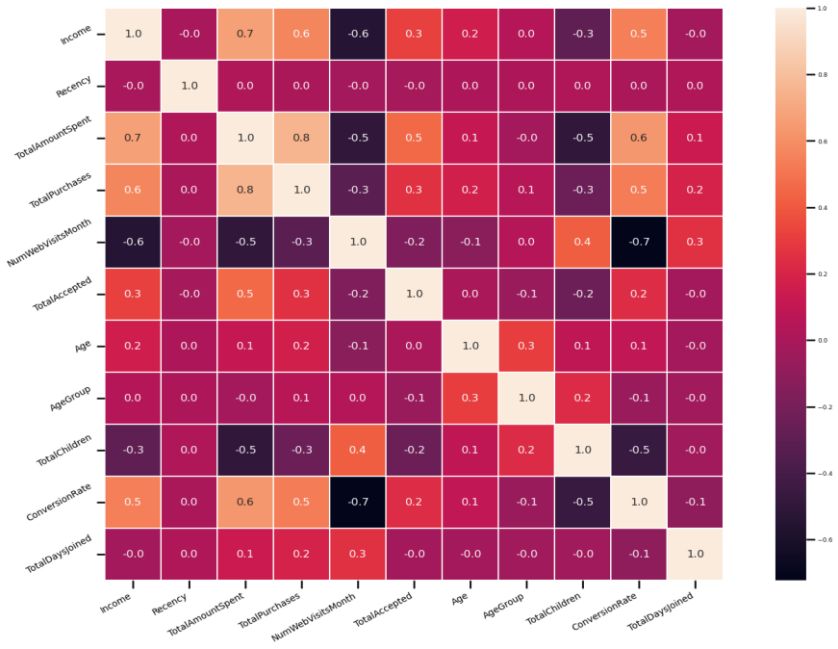
Features ConversionRate, Income, and TotalAmountSpent are right skewed, meaning that most values tend to be low. Still, some very high values make the average more significant than the median. It could indicate that most customers have a low value in this feature, but a few have a very high value. Meanwhile, the Age feature has a more symmetrical distribution, meaning that the average value of the Age feature is almost close to the median value. Based on the boxplot for each feature, several outliers are visible outside the interquartile range (IQR) for each feature. Outliers in a boxplot indicate some extreme or unusual data values.

Bivariate Analysis



ConversionRate vs Income has a positive relationship, and there are several outliers. Meanwhile, ConversionRate vs TotalAmountSpent and ConversionRate vs Age have data distributions that converge in one area. It indicates no significant variation in ConversionRate when looking at TotalAmountSpent or Age. However, some unusual values may be outliers in terms of this relationship.

Multivariate Analysis

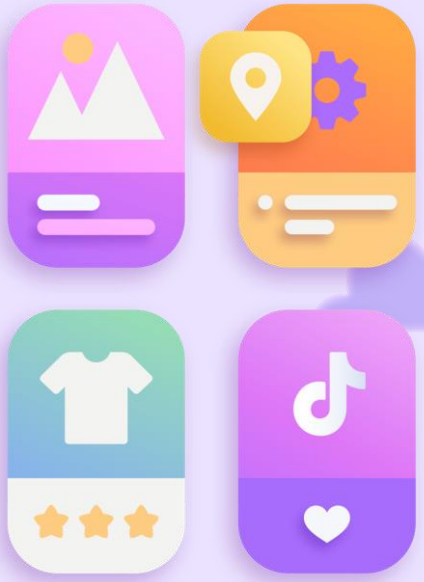


There are several strong positive correlations, including :

- TotalAmountSpent & Income: A positive correlation of 0.7. That is when income increases, total expenditure also tends to increase.
- TotalPurchases & Income: Positive correlation of 0.6. It means that there is a tendency for customers with higher incomes to make more purchases.
- TotalPurchases & TotalAmountSpent: Strong positive correlation of 0.8. It shows that the more spending, the more purchases customers make.
- ConversionRate & TotalAmountSpent: Positive correlation of 0.6. It means that the more money customers spend, the more likely they are to convert.

There are several strong negative correlations, including:

- NumWebVisitMonth & Income : Negative correlation of -0.6. That is, the higher the income, the fewer visits to the website in a month.
- ConversionRate & NumWebVisitMonth: Strong negative correlation of -0.7. It means that the more visits to the website in a month, the lower the probability of customer conversion.



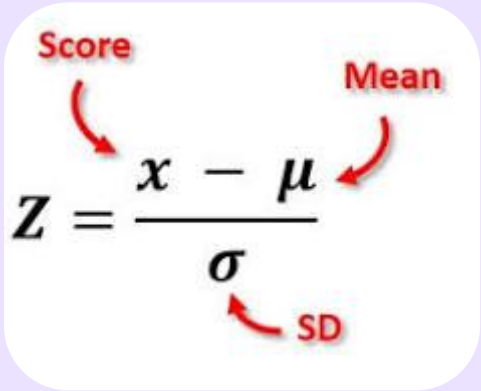
04

Modeling

Standardization

The K-Means algorithm is susceptible to variable scales. Variables with large scales will be dominant in calculating distances. Standardization helps prevent bias that may arise due to differences in scale—standardized features.

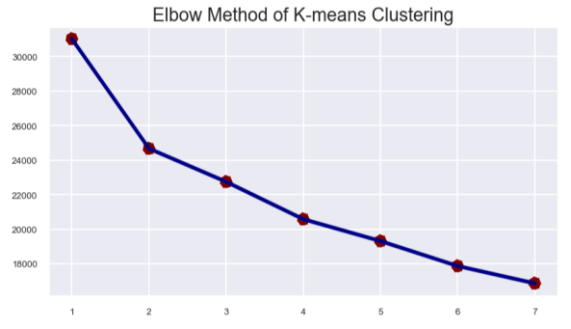
Standardization Method



Features

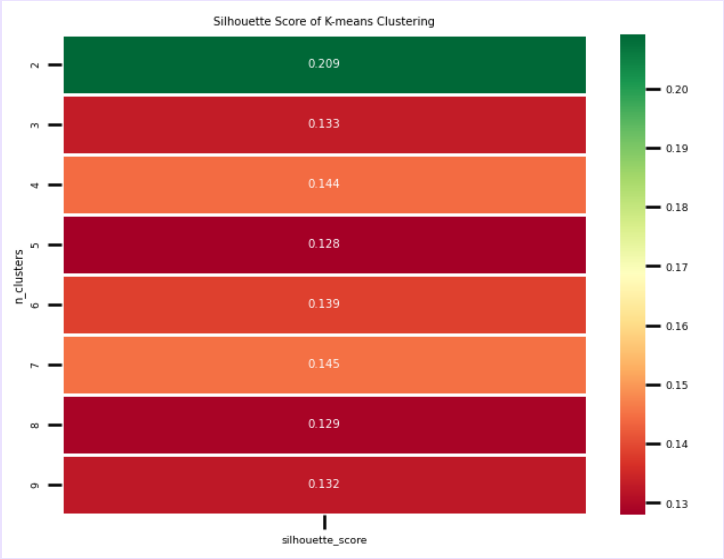
- Education
- Income
- Recency
- NumWebVisitsMonth
- Complain
- Response
- TotalDaysJoined
- AgeGroup
- TotalChildren
- TotalAmountSpent
- TotalPurchases
- TotalAccepted
- ConversionRate
- Have_a_Partner

Evaluate Cluster

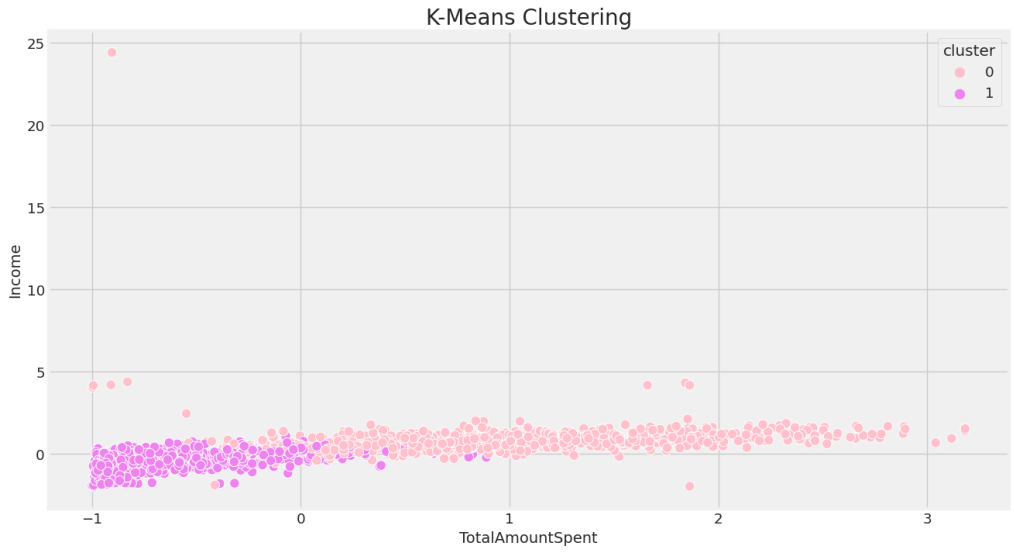


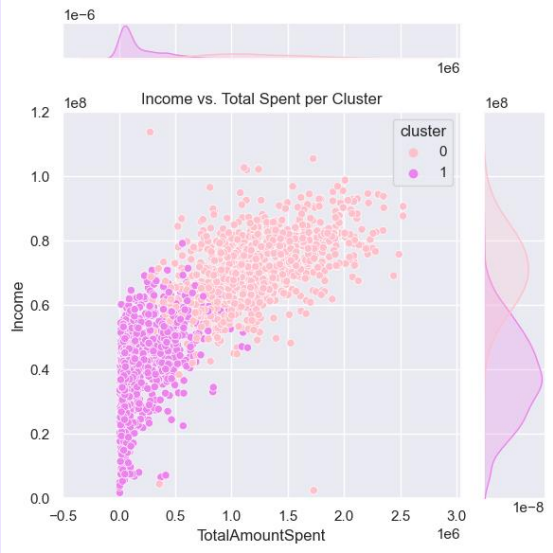
The elbow method determines the optimal number of clusters by identifying "elbow" points in the graph. It can be seen that point 2 is like an elbow on the graph.

Silhouette score calculates how similar or different a data is to the cluster where the data is, compared to its nearest neighboring cluster. The highest silhouette score occurs when there are 2 clusters. Therefore, the optimal number of clusters is 2.



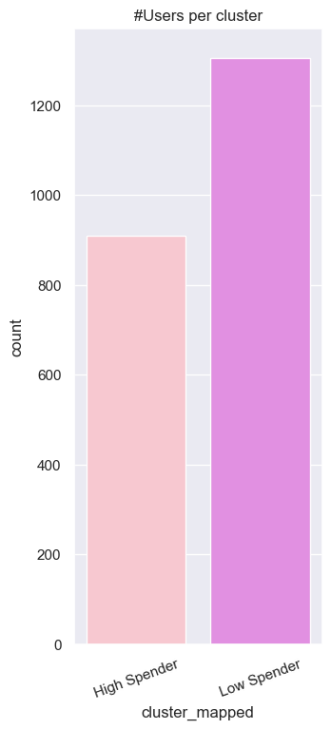
K-Means Clustering





The results of the clustering process show that in cluster '0', there are customers who can be identified as High spenders, while in cluster '1', there are customers who can be identified as low spenders.

Interpretation

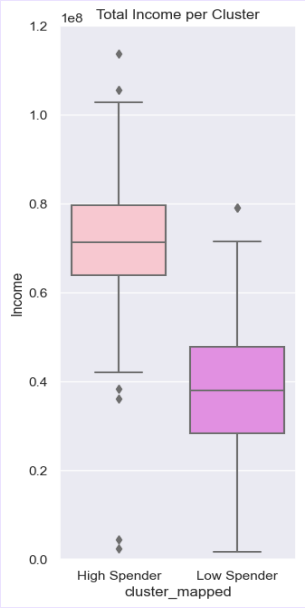


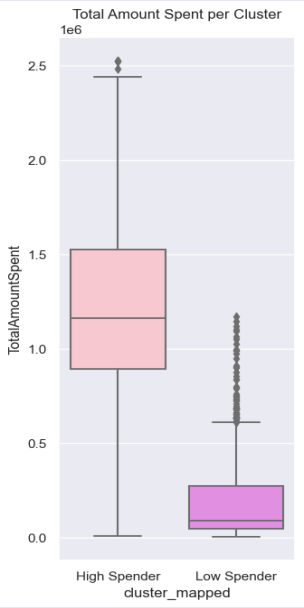
The interpretation of the clustering results provides an essential understanding of customer behavior in two different groups :

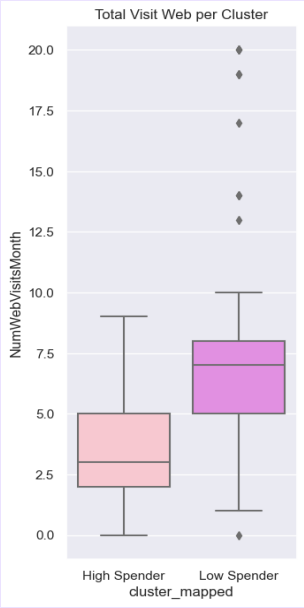
- Cluster '0' (High Spender) : This cluster describes a group of customers who tend to make lower-value purchases or spend less money. It could indicate that customers in this cluster prefer a more affordable product or service or have a more conservative purchase level.

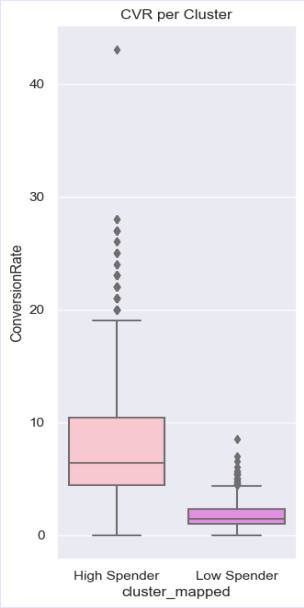
- Cluster '1' (Low Spender) : This cluster characterizes a group of customers who tend to make high-value purchases or spend more money. It could indicate that customers in this cluster tend to buy more expensive products or services or have a tendency to make more extensive and frequent purchases.

Interpretation

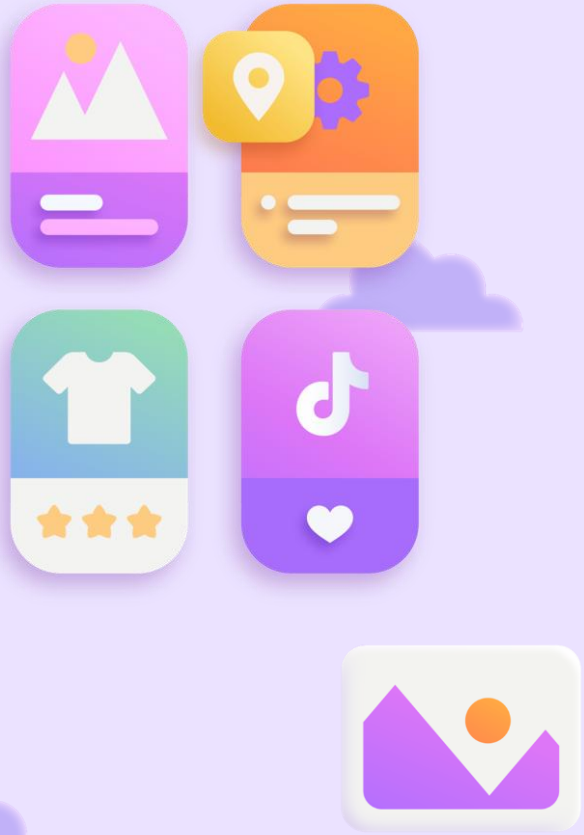








This data distribution occurs in high-spender and low-spender clusters based on the Income, TotalAmountSpent, NumWebVisitMonth, and ConversionRate features. It can be seen that there are still some outliers in each cluster.



05

Insight & Recomendation

Insight

1. High Spender

This group is dominated by customers aged 45-65 who are married and do not have children. In terms of income and total expenditure, this cluster has high income and total expenditure, around 72 million per year for revenue and 1.2 million per year for total expenditure. Customers in this cluster respond most to company campaigns and complain less frequently. Even though they don't visit the website often, their conversion rate is high.

2. Low Spender

This group is dominated by customers aged 45-65 who are married and have one child. In terms of income and total expenditure, this cluster has low income and total expenditure, around 13.5 million per year for income and 185 thousand per year for total expenditure. Customers in this cluster make the most complaints and respond the least to company campaigns. This cluster visits websites more frequently than the High Spenders cluster, but their conversion rates are low.

Recommendations

1. Targeting High Spender Customers

Companies can offer special promotions, exclusive products, or incentives to attract High Spender customers. Strive to satisfy High Spender customers with superior service, responsive customer support, and a customized experience. They are a significant source of income.

2. Increasing Low Spender Customer Engagement

Identify the preferences and needs of Low Spender customers, offer promotions that are better suited to their budgets, and create more attractive campaigns for Low Spender customers to increase their engagement. Campaigns with discounts, reward points, or loyalty programs can help.

3. Fixed Low Spender Conversion Rate

Conduct surveys or deeper analysis to understand why Low Spender customers have low conversion rates. It can help in adjusting marketing and product strategies. Offer more relevant product recommendations based on the purchase history of Low Spender customers. It may encourage them to make additional purchases.

Recommendations

4. Website Analysis and Complaints

Companies can improve the user experience on their websites, including better navigation, more informative content, and more engaging features to increase visits and conversions. For handling complaints, customer complaints can be researched and handled quickly and effectively. Customers who are satisfied with handling their complaints are more likely to remain loyal.

5. Personalization and Segmentation

Use customer data to personalize communications and offers. Customers tend to respond better if they feel personally cared for. When customers feel personally cared for, they are more likely to interact with the company, feel valued, and even increase their satisfaction. It can positively impact customer retention, conversion, and customer loyalty to a company's brand or product.

Thanks!

radishafs@gmail.com

linkedin.com/in/radishafannis