IE 7275

Data Mining in Engineering

CUSTOMER SEGMENTATION

By- Sharayu Thosar

Yishtavi Gedipudi

Submission Date : 14th April 2023

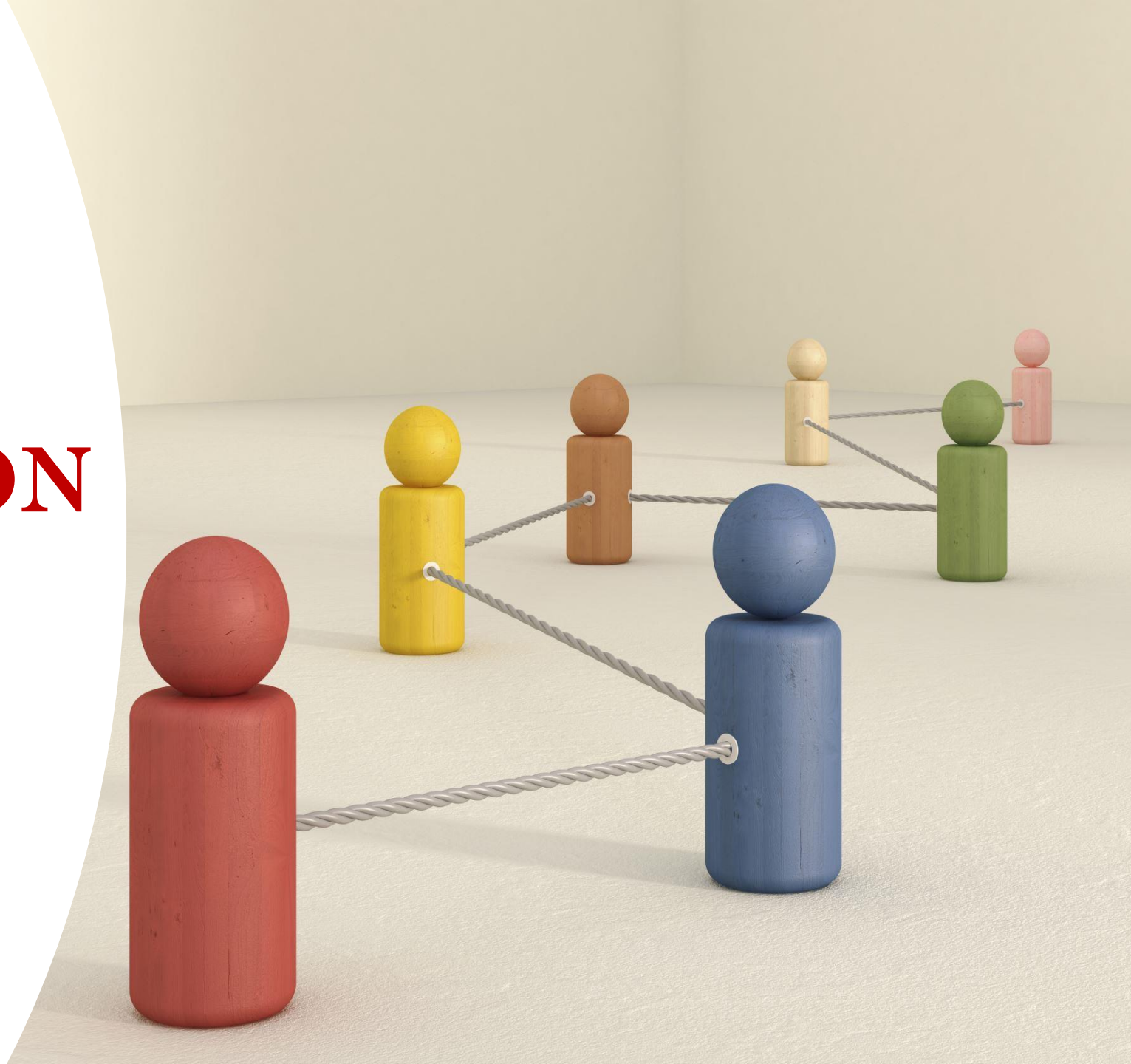


TABLE OF CONTENTS

PROBLEM SETTING

A retail company wants to segment its customer base in order to better target marketing efforts and improve sales.

The company has collected data on customer demographics, purchasing history, and responses to previous marketing campaigns.

By understanding the customers' personalities, businesses can improve their customer service and create a more personalized shopping experience and tailor their marketing efforts and product offerings to specific segments.

DATA DESCRIPTION

The dataset includes 29 columns and 2240 records. The columns include following information about customers-

1. Demographics Customer ID, Education, Age, Income, Marital Status, Number of Kids, etc.
2. Purchases Amounts spent on Wines, Fruits, Fish, Meat, Gold, etc.
3. Responses for 5 Marketing Campaigns
4. Purchase places for example websites, stores, etc.

PROMOTION

Variable	Description
NumDealsPurchases	Number of purchases made with a discount
AcceptedCmp1	1 if customer accepted the offer in the 1st otherwise
AcceptedCmp2	1 if customer accepted the offer in the 2nd otherwise
AcceptedCmp3	1 if customer accepted the offer in the 3rd otherwise
AcceptedCmp4	1 if customer accepted the offer in the 4th otherwise
AcceptedCmp5	1 if customer accepted the offer in the 5th otherwise
Response	1 if customer accepted the offer in the last otherwise

PEOPLE

Variable	Description
ID	Customer’s unique identifier
Year_Birth	Customer’s unique identifier
Education	Education Qualification of customer
Marital_Status	Marital Status of customer
Income	Customer's yearly household income
Kidhome	Number of children in customer's household
Teenhome	Number of teenagers in customer's household
Dt_Customer	Date of customer's enrollment with the company
Recency	Number of days since customer's last purchase
Complain	1 if the customer complained in the last 2 years, 0 otherwise

PRODUCTS

Variable	Description
MntWines	Amount spent on wine
MntFruits	Amount spent on fruits
MntMeatProducts	Amount spent on meat products
MntFishProducts	Amount spent on fish products
MntSweetProducts	Amount spent on sweet products
MntGoldProds	Amount spent on gold products

PLACES

Variable	Description
NumWebPurchases	Number of purchases made through the company’s website
NumCatalogPurchases	Number of purchases made using a catalog
NumStorePurchases	Number of purchases made directly in stores
NumWebVisitsMonth	Number of visits to company’s website in the last month

DATA PROCESSING

DATA CLEANING

Changing the datatype

Dt Customer

Object -> Date

Generating Columns

1. Age of Customer from Date of Birth.
2. Years of Enrollment from Date of Enrollment

Handling Missing Value

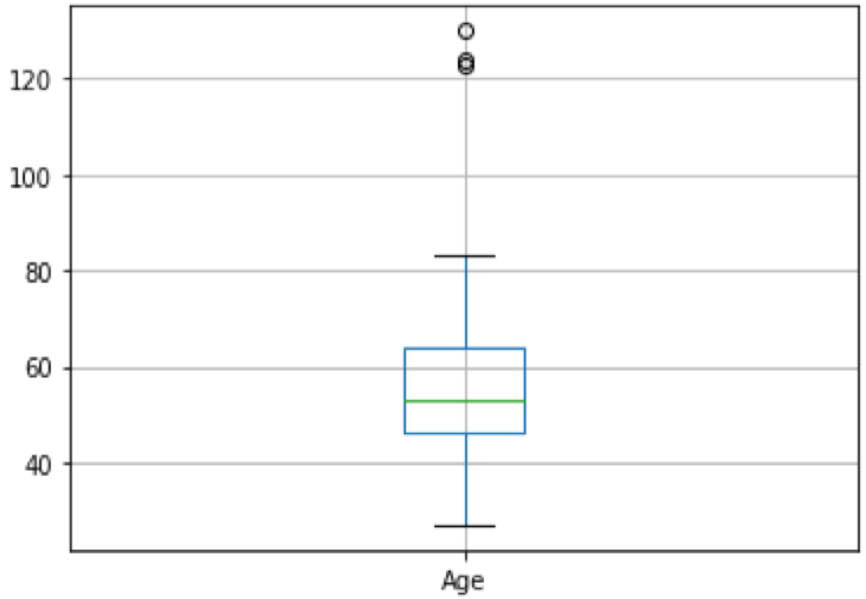
Column with Missing values- Income

Approach- Removing Outliers and substituting with the mean.

DATA PROCESSING

ANALYSING NUMERICAL VARIABLES

- Some values in the ‘Age’ column were above 120, which is not a realistic age for a customer.
- Action – Drop those records



DATA PROCESSING

ANALYSING CATEGORICAL VARIABLES

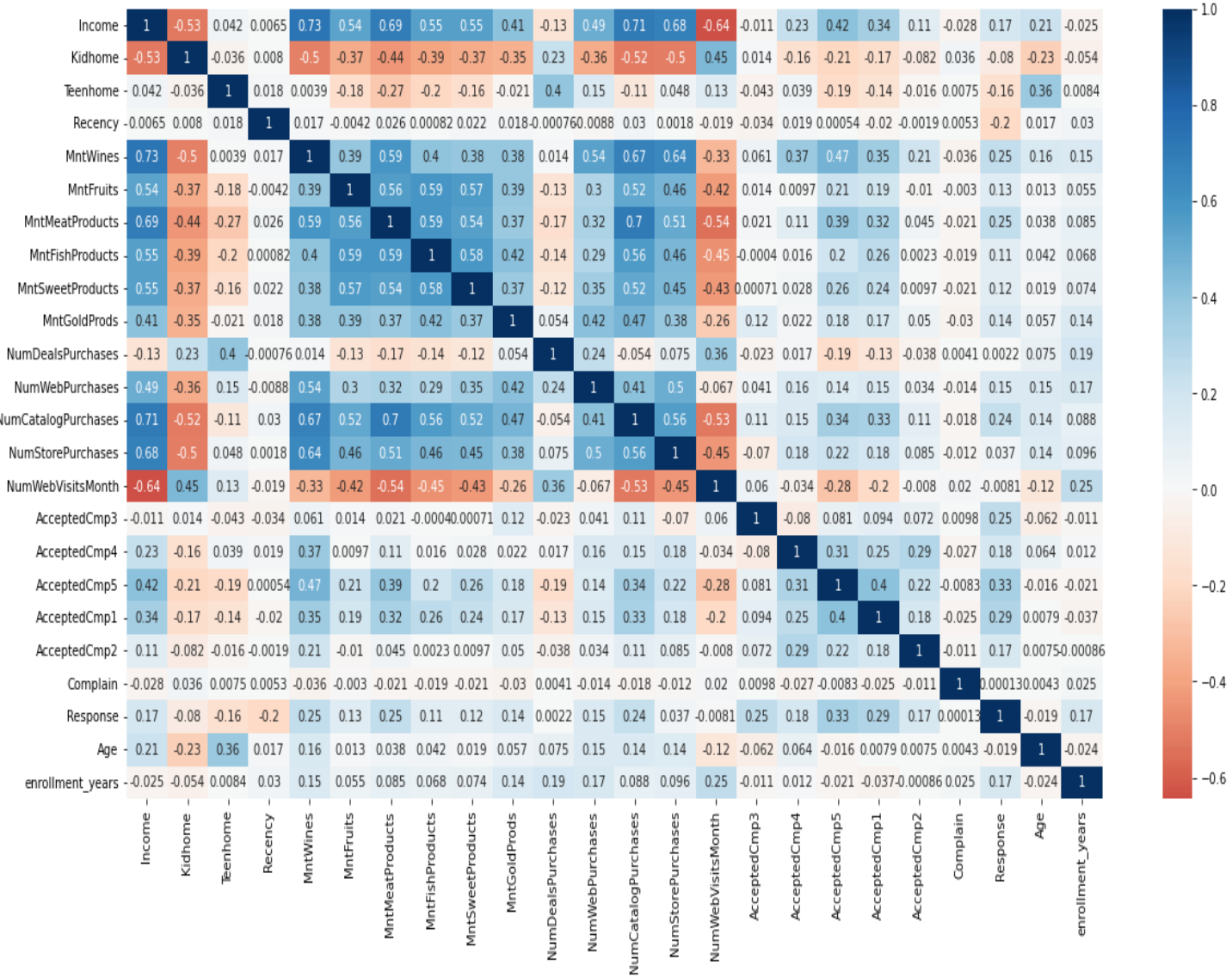
- Checked for irrelevant categories within the variables, such as 'Alone', 'Absurd', and 'YOLO' in Marital Status, which do not accurately describe the customers' marital status.
- Combined or dropped the irrelevant categories to reduce dimensionality and ensure accurate analysis.
- Made necessary substitutions or transformations to the categorical variables to prepare them for further analysis, such as creating dummy variables.

DATA PROCESSING

DIMENSION REDUCTION

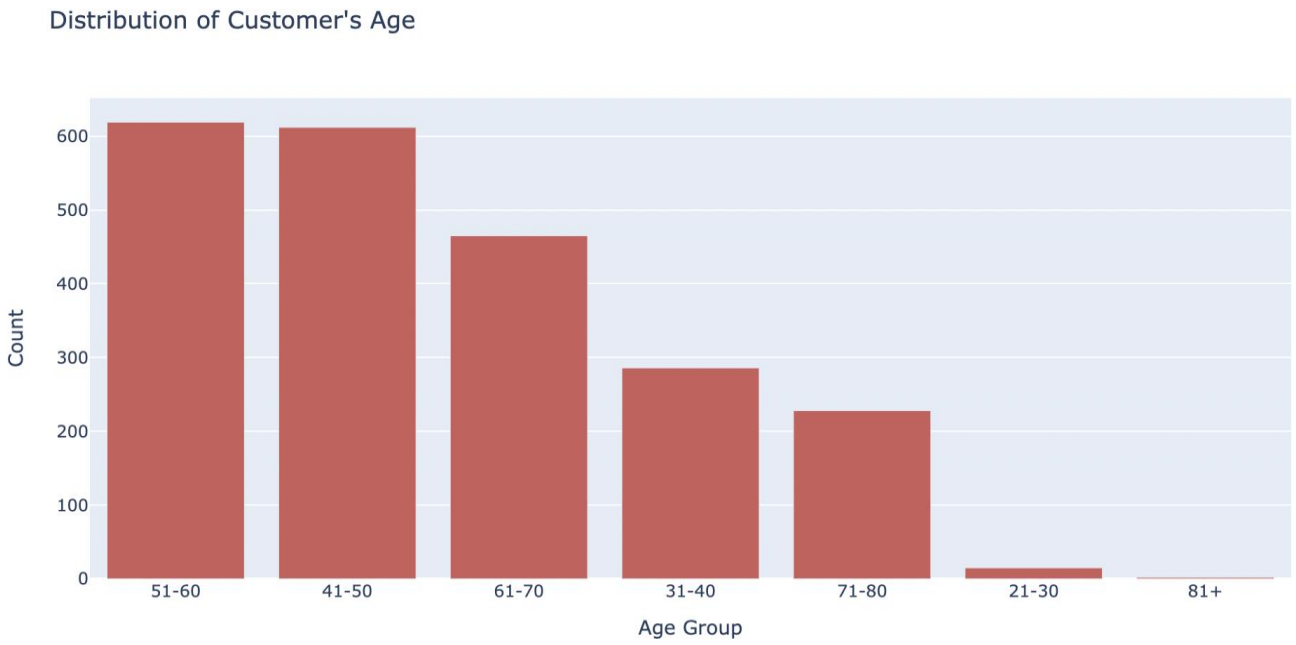
- The variables 'Age' and 'Year Birth' both contain the same data. As a result, we can remove 'Year Birth' to reduce data redundancy.
- The 'Years Enrollment' feature has already been derived from the 'Dt Customer' variable. As a result, we can eliminate the 'Dt Customer' column to reduce dimensions.
- The 'ID' column provides no useful information for clustering customers. As a result, we can remove this column.
- 'Z CostContact' has a fixed value of 3 and will not contribute to customer clustering. As a result, it can be removed to reduce the dataset's dimensionality.
- 'Z Revenue' has a constant value of 11 as well and does not provide any additional information for clustering. As a result, it can be removed to further reduce the dimensionality of the dataset.
- PCA Analysis

DATA EXPLORATION



- The correlation heatmap analysis helped us identify the strong correlation between 'MntWines' and 'Income' columns.
- This information can be used to target customers with higher incomes in wine-related marketing campaigns.

DATA EXPLORATION

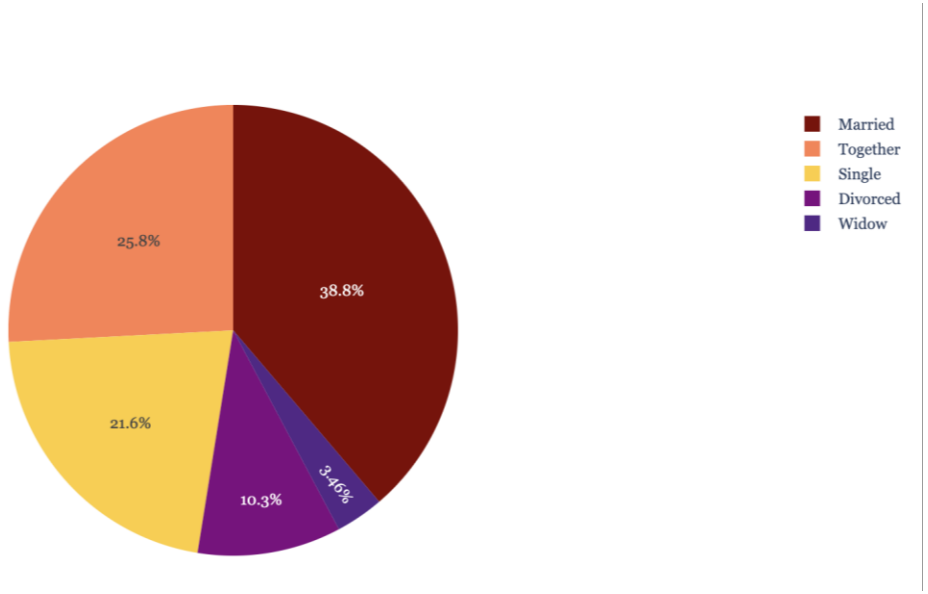


Distribution of Customer's Age

We can see that the largest group of customers falls within the age range of 41-60. This suggests that our loyalty program may be particularly appealing to middle-aged individuals.

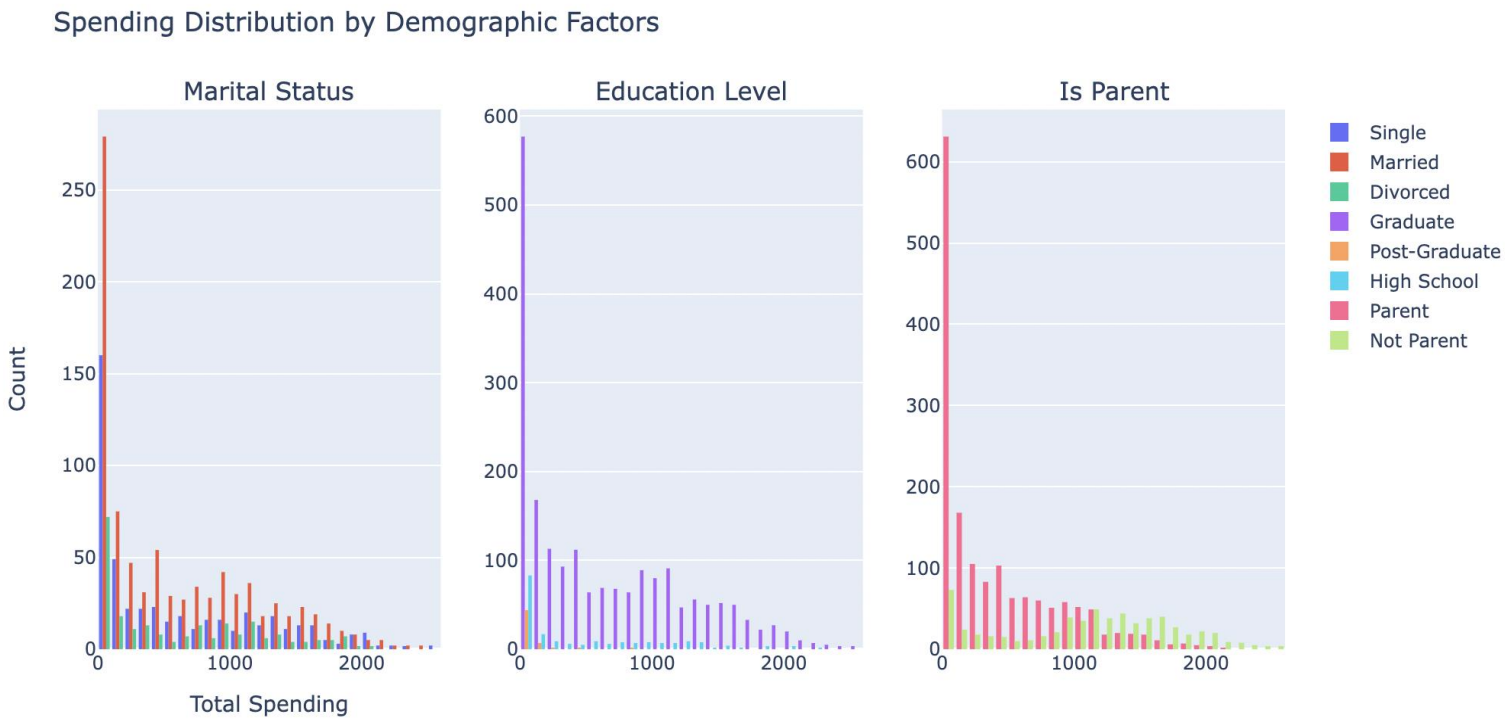
Marital Status Distribution

We can see that majority of our customer base is married, with almost half of all customers falling into this category. Following this, the next most common category is single, which accounts for around a third of customers.



DATA EXPLORATION

Spending Distribution by Demographic Factors



- The spending distribution by demographic factors reveals interesting insights.
- Among the marital status categories, married customers tend to spend more than other categories.
- Similarly, graduates tend to spend more than other education level categories.
- Furthermore, the analysis shows that people who are parents tend to spend more, compared to those who are not parents. These insights can help in developing targeted marketing strategies to improve customer engagement and retention.

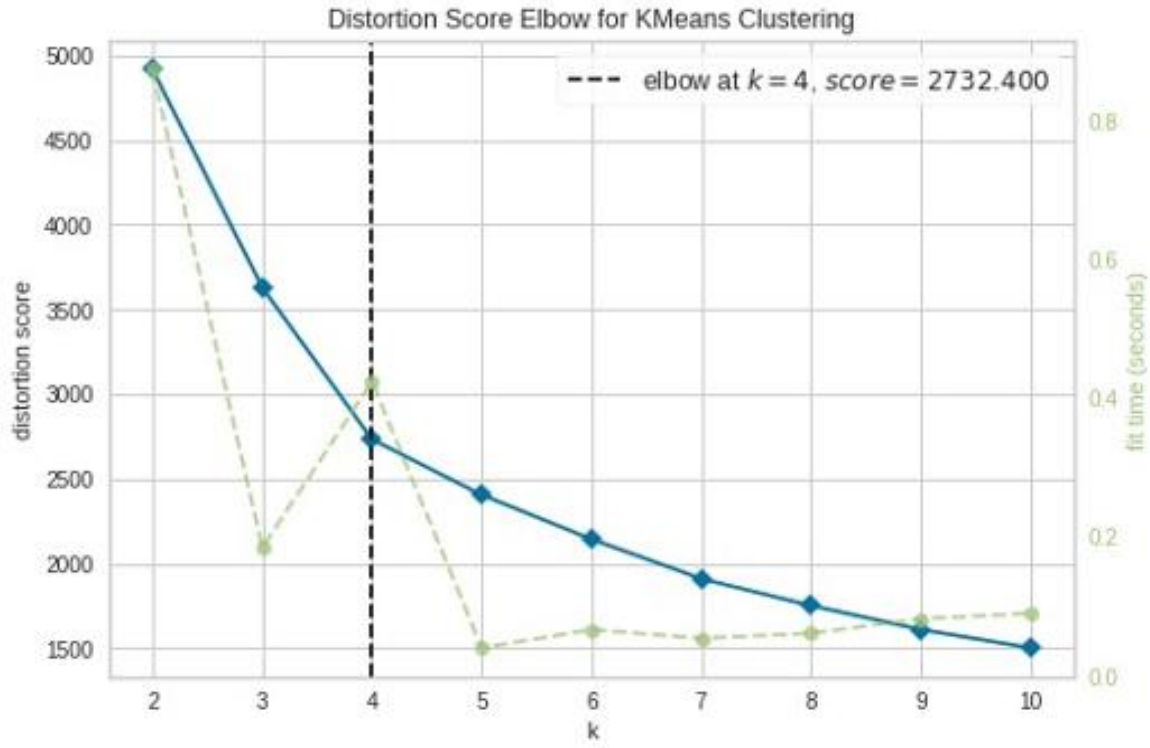
DATA MINING TASKS

- Categorical columns ‘Education’ and ‘Marital Status’ which were replaced by creating dummies.
- Standard Scaler was used to transform data and get all the features between values -1 to 1.
- Principal Component Analysis was performed to reduce the dimensions further.
  - Three components were taken from the PCA which captured approximately 47.63% of variation in the data.

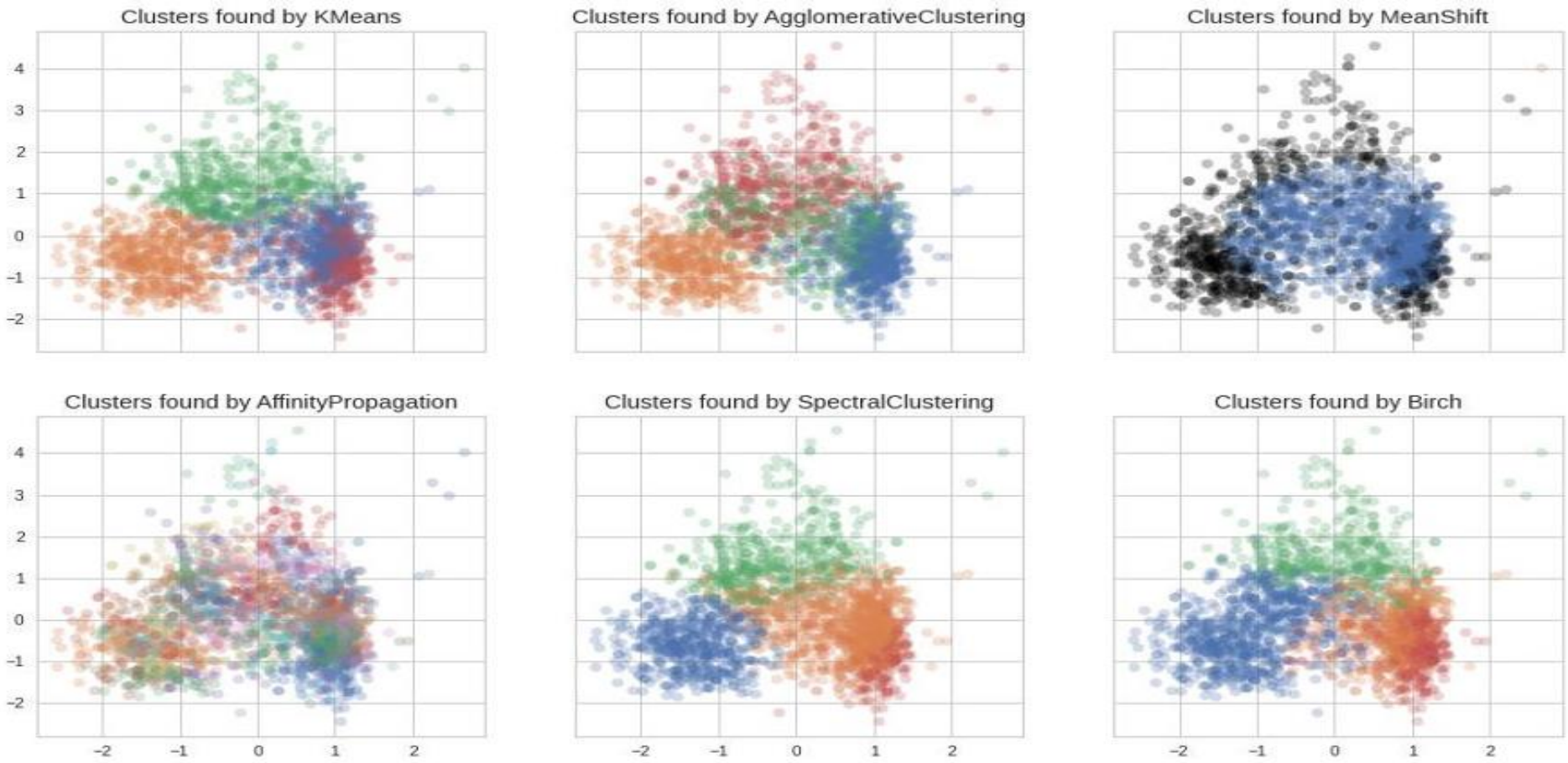
DATA MINING MODELS

- While exploring, we worked with 6 Models that included Kmeans, Hierarchical Clustering, Mean Shift, Affinity Propagation, Spectral Clustering and Birch.

- With number of clusters = 4, different models were implemented and visualized in 2D for PC1 and PC2.



DATA MINING MODELS



PERFORMANCE EVALUATION

Sr. No.	Model	Silhouette Score	Dunn Index
1	Kmeans	0.325	0.028
2	Hierarchical Clustering	0.264	0.019
3	Mean Shift	0.186	0.013
4	Affinity Propagation	0.253	0.016
5	Spectral Clustering	0.228	0.014
6	Birch	0.274	0.018

Higher Silhouette Score and Dunn Index indicates better performance for the clustering.

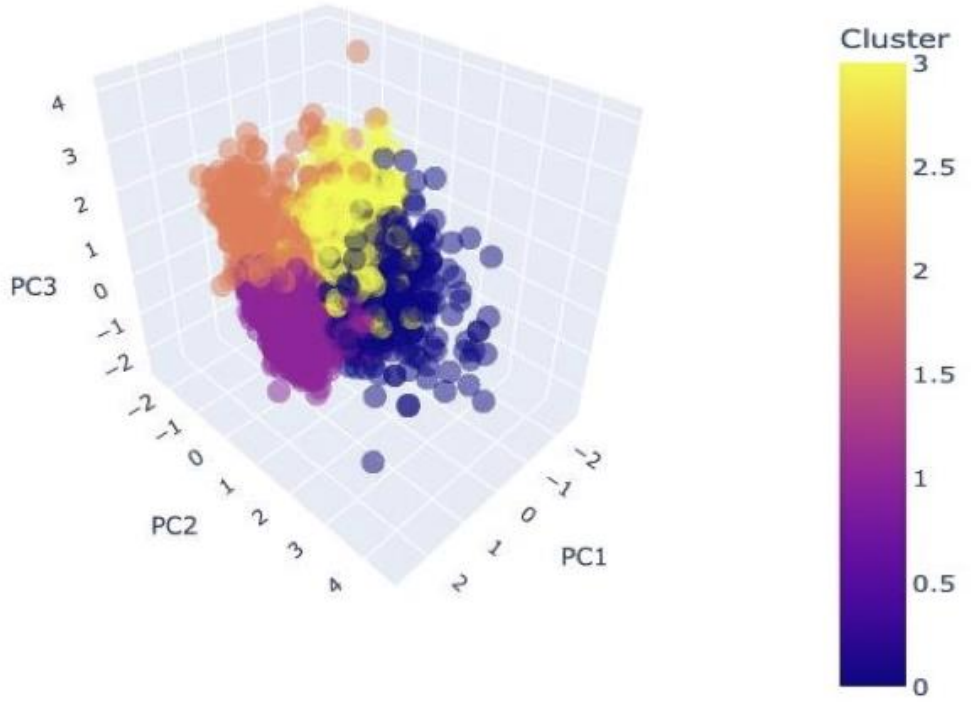
Kmeans being the best model, we have done the evaluation and interpretation of the clusters based on Kmeans Algorithm.

IMPLEMENTATION

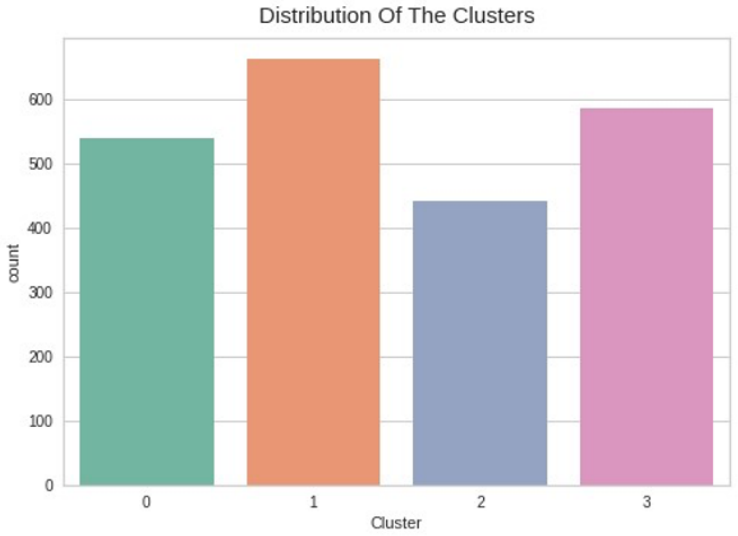
K-MEANS ALGORITHM

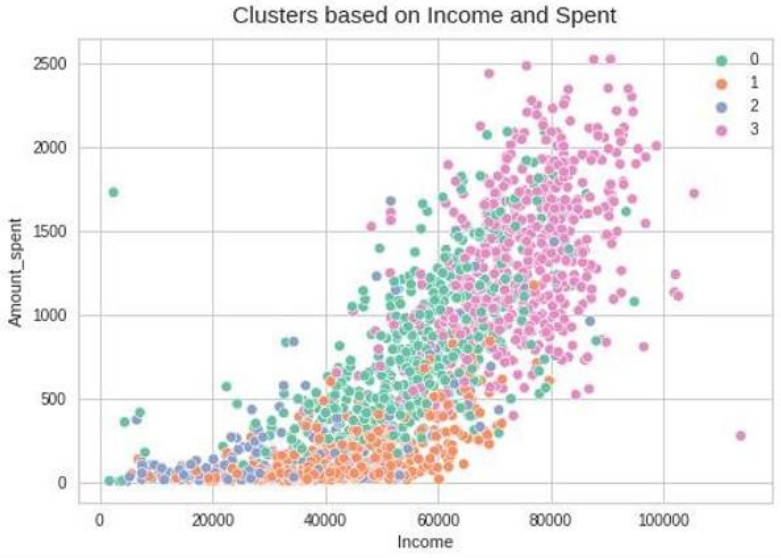
- Performance Metrics
  - Silhouette Score = 0.325
  - Dunn Index = 0.028

SCATTER PLOT WITH 3 PC'S



CLUSTER SEGMENTATION





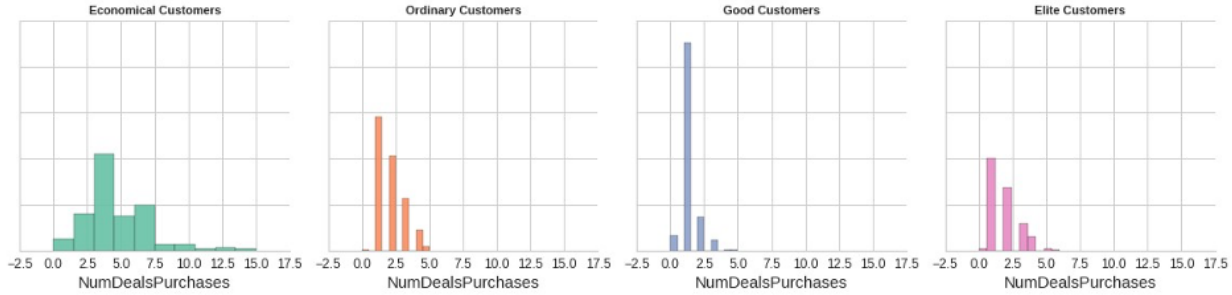
Customers have patterns when clustered based on the income and amount spent.

Hence, we try to categorize these customers into different segments and then make further interpretation for each customer segment.

ELITE CUSTOMERS	Cluster 3
GOOD CUSTOMERS	Cluster 0
ECONOMIC CUSTOMERS	Cluster 1
ORDINARY CUSTOMERS	Cluster 2

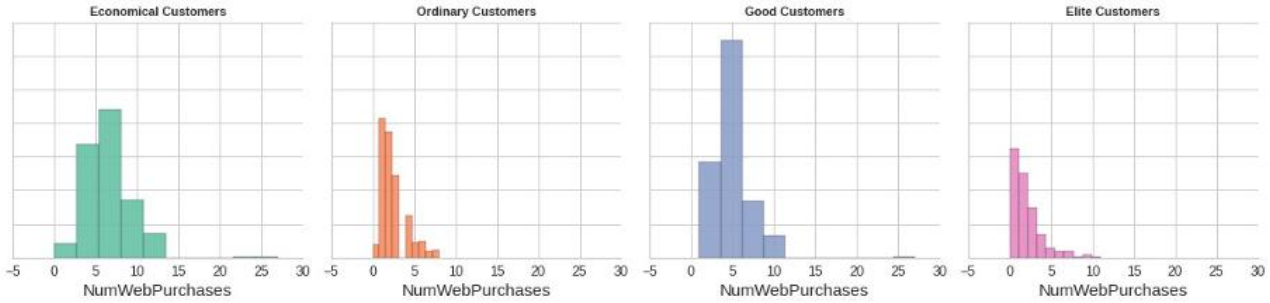
CLUSTER INTERPRETATION

Deal Purchases for different customer segments

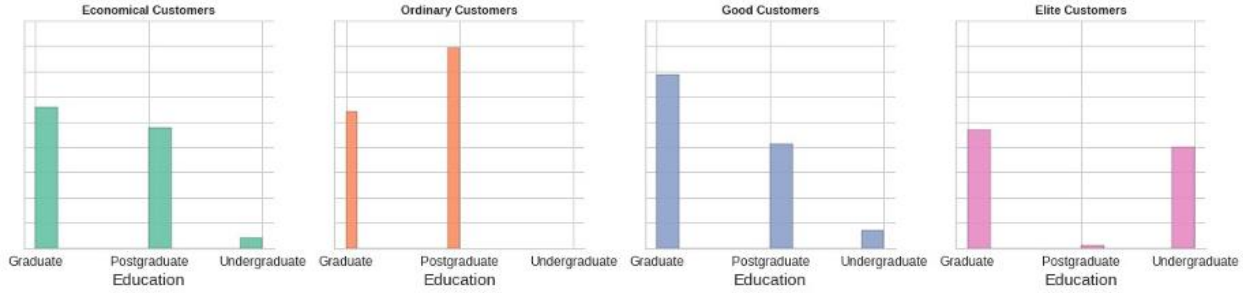


Store purchases for Different Customer Segments

Web purchases for different customer segments

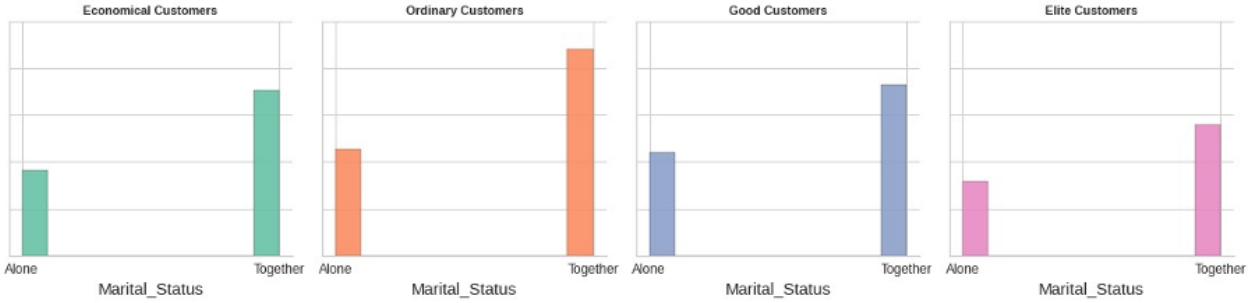


Education for different customer segments



Catalog Purchases for Different Customer Segments

Marital Status for Different Customer Segments



RESULTS

ELITE CUSTOMERS

These are customers who have the highest income, are graduates, have 0 or 1 child, and spend the highest amount. They tend to make more catalog purchases and have campaigns with positive performance. They have the least number of web visits but make more web purchases. This group represents the highest spending and most profitable customer segment.

GOOD CUSTOMERS

This group includes customers with high income and amount spent, mostly parents with 1 or 2 kids. They tend to purchase more in deals compared to other customers and have a dominant presence on the web. This segment represents an important customer base that values deals and convenience.

ECONOMIC CUSTOMERS

These are customers with lower income who have the most number of children. Most of them are together, representing people with more members in the family. They tend to buy more in deals and very few catalog purchases. This group represents a value-oriented segment of customers who prioritize practicality over luxury.

ORDINARY CUSTOMERS

This group includes customers with the lowest income and amount spent, less education, and an overall lower number of purchases. It includes comparatively fewer children and people from marital status as both together and alone. This segment represents a diverse group of customers with varied preferences and purchasing power.

IMPACT AND CONCLUSION

Interpretations can be used to gain insights and improve the marketing as follows

ELITE CUSTOMERS

Since this group is the most profitable, it's essential to retain them by offering exclusive deals, promotions, and personalized experiences.

- The catalog purchases they make show that they appreciate quality and luxury products, so marketing efforts should highlight the premium quality of your products.
- Also, they tend to make more web purchases, so it's essential to have a seamless and user-friendly online shopping experience to cater to their preferences.

GOOD CUSTOMERS

This group values deals and convenience, so marketing efforts should emphasize the affordability and value proposition of your products.

You can create loyalty programs and rewards to incentivize them to make repeat purchases.

- Since they have a dominant presence on the web, it's essential to have a strong online presence and offer a seamless e-commerce experience. You can also use email marketing and social media platforms to reach out to them with personalized offers.

ECONOMIC CUSTOMERS

This segment represents value-oriented customers who prioritize practicality over luxury.

- Marketing efforts should emphasize the affordability and value proposition of your products while highlighting their practicality and usefulness in everyday life.
- You can offer deals and discounts that align with their budget, and highlight products that cater to families and larger households. Since they make fewer catalog purchases, you can focus on email marketing and social media platforms to reach out to them with personalized deals and offers.

ORDINARY CUSTOMERS

This group represents a diverse group of customers with varied preferences and purchasing power.

- You can segment this group based on their demographic and psychographic factors and offer targeted promotions that appeal to their specific needs.
- Since they have a lower income and spending power, you can offer affordable products and focus on value propositions to cater to their budget.

THANK YOU !