CUSTOMER SEGMENTATION

Milestone: Project Report Draft

Group 18

Sharayu Thosar

Yishtavi Gedipudi

thosar.sh@northeastern.edu

gedipudi.y@northeastern.edu

Percentage of Effort Contributed by Student 1: 50% Percentage of Effort Contributed by Student 2: 50%

Signature of Student 1: Sharayu Thosar

Signature of Student 2: Yishtavi Gedipudi

Submission Date: 04/07/2023

Table of Contents

I. 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.

The goal of the market segmentation project is to use machine learning techniques to segment customers based on their Personality Analysis, hoping they can tailor their marketing efforts to an ideal customer group and improve sales.

One of the challenges in this project will be dealing with high-dimensional data, as well as handling any missing or incomplete data in the dataset. Additionally, selecting an appropriate number of clusters and determining the optimal parameters for the clustering algorithm may also be challenging.

II. Problem Definition:

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. Overall, the main issues that businesses are attempting to address through customer personality analysis are targeting the right audience and increasing customer loyalty and satisfaction. Questions addressed in the project:

1. What are the ideal customer segments in the company's customer base?
2. Demographics: What is the age, gender, income, education level, etc. of customers?
3. Can the identified customer segments improve marketing targeting and increase sales?
4. What are the customer's purchasing habits and response to marketing campaigns?
5. What are the probable purchases that the customer will make based on purchase history?
6. What are the target products for each customer segment?

III. Data Source:

https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis

IV. Data Description

The dataset includes 29 columns and 2240 columns. 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.

We chose Kaggle's 'marketing campaign.csv' dataset for our customer segmentation project because of its size, variety of variables, and real-world applicability. The dataset contains information on a financial services company's marketing campaign, including valuable demographic, behavioral, and campaign response data for 2,240 customers. Our strategy is to analyze this data to identify patterns and characteristics that distinguish customer segments, allowing businesses to tailor their marketing efforts to each group's specific needs and preferences. We hope to help businesses better understand their customers and develop more effective marketing strategies through customer segmentation analysis. Overall, we believe that the 'marketing-campaign.csv' dataset is a valuable resource for conducting customer segmentation analysis, and that our project will provide useful insights for businesses looking to improve their marketing.

Table 1. Description of Variables

S.no	**Variable**	**Description**
People		
1.	ID	Customer's unique identifier
2.	Year_Birth	Customer's birth year
3.	Education	Education Qualification of customer
4.	Marital_Status	Marital Status of customer

5.	Income	Customer's yearly household income
6.	Kidhome	Number of children in customer's household
7.	Teenhome	Number of teenagers in customer's household
8.	Dt_Customer	Date of customer's enrollment with the company
9.	Recency	Number of days since customer's last purchase
10.	Complain	1 if the customer complained in the last 2 years, 0 otherwise
Products		
11.	MntWines	Amount spent on wine
12.	MntFruits	Amount spent on fruits
13.	MntMeatProducts	Amount spent on meat products
14.	MntFishProducts	Amount spent on fish products
15.	MntSweetProducts	Amount spent on sweet products
16.	MntGoldProds	Amount spent on gold products
Place		
17.	NumWebPurchases	Number of purchases made through the company's website
18.	NumCatalogPurchases	Number of purchases made using a catalog
19.	NumStorePurchases	Number of purchases made directly in stores
20.	NumWebVisitsMonth	Number of visits to company's website in the last month
Promotion		
21.	NumDealsPurchases	Number of purchases made with a discount
22.	AcceptedCmp1	1 if customer accepted the offer in the 1st campaign, 0 otherwise
23.	AcceptedCmp2	1 if customer accepted the offer in the 2nd campaign, 0 otherwise
24.	AcceptedCmp3	1 if customer accepted the offer in the 3rd campaign, 0 otherwise

25.	AcceptedCmp4	1 if customer accepted the offer in the 4th campaign, 0 otherwise
26.	AcceptedCmp5	1 if customer accepted the offer in the 5th campaign, 0 otherwise
27.	Response	1 if customer accepted the offer in the last campaign, 0 otherwise

V. Data Processing

A. Data Cleaning

- Changing the datatype-

  Converting the 'Dt Customer' column's data type from 'object' to 'datetime' is an important data cleaning step for the 'marketing campaign.csv' dataset. We can use the conversion to extract time-based features like day, month, year, and time duration, which are useful in marketing campaign analysis. We can sort and group data more precisely with datetime data types, identify seasonal patterns, and investigate the impact of enrollment dates on customer behavior. This increases the precision of our analysis and allows us to extract more valuable insights from the data.

- Generating columns-

  We can segment our customers based on age and loyalty by extracting these features, allowing us to identify trends and patterns that can be used to optimize marketing strategies.

  1. Age of customer from Date of Birth

     A customer's age can help us understand their purchasing habits and how they change as they get older. When the datetime conversion is finished, we can take the year from the 'Date of Birth' column and subtract it from the current year to get the customer's age.
  2. Years of enrollment from Date of Enrollment

Years of enrollment can assist us in determining the customer's loyalty to the brand.To calculate the number of years of enrollment in the 'Years of Enrollment' column, subtract the enrollment year from the current year.

- Handling Missing Values

  While checking for missing values we found about 24 missing records in the Income column. The steps used for dealing with missing values in the 'Income' column are as follows:

  1. Before we replaced the missing values, we looked for outliers in the 'Income' column. This is because, if there are significant outliers, replacing missing values with the mean or median may not be the best approach.
  2. Created a box plot to visualize the distribution of the 'Income' column and identify outliers.

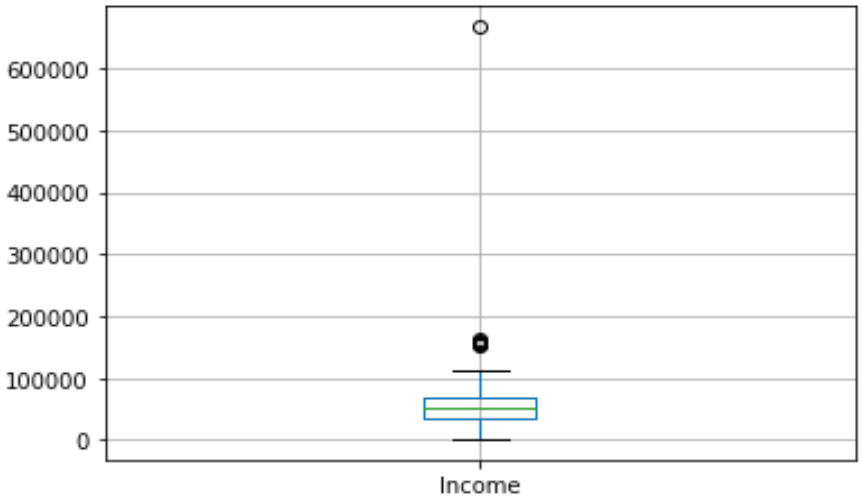

Fig 1: Outliers in the Income variable

3. Outliers in the 'Income' column were to be removed. This could be accomplished through the use of statistical methods such as the interquartile range (IQR) or visually through the use of box plots.

4. Substituted the mean of the remaining values for the missing values in the 'Income' column after removing the outliers. This is a common method for dealing with missing values and ensures that the missing values have no significant impact on the distribution of the 'Income' column.

By following these steps, we can handle missing values in the 'Income' column and ensure that missing data does not affect our analysis.

B. Analyzing Numerical Variables

We noticed that some values in the ‘Age’ column were above 120, which is not a realistic age for a customer.

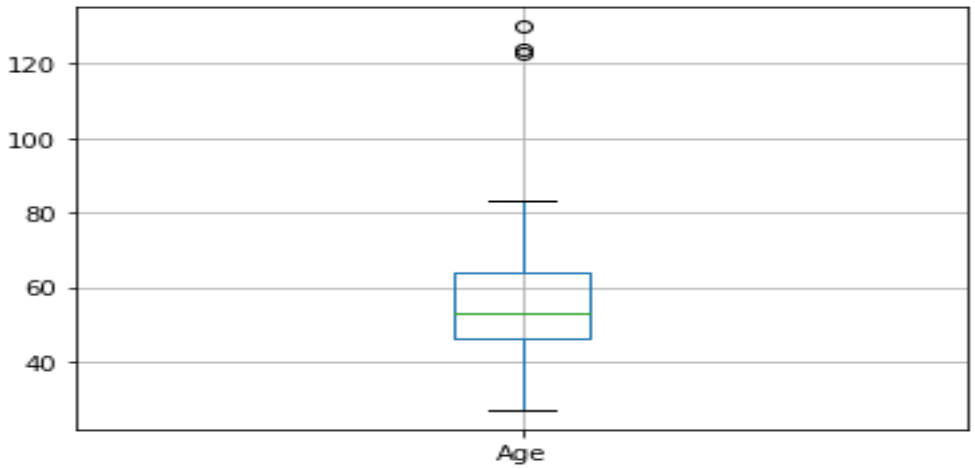

Fig 2 : Outliers in Age variable

As a result, we removed these outliers from the dataset because they would not be useful in our analysis. We can ensure the accuracy and reliability of our analysis by removing outliers and dealing with missing values.

C. Analyzing Categorical Variables

1. Identified the categorical variables that need to be analyzed, such as Education and Marital Status.
2. 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.
3. Combined or dropped the irrelevant categories to reduce dimensionality and ensure accurate analysis.
4. For the 'Alone' category, combined it with the 'Single' category as they had the same meaning.
5. Investigated further for 'YOLO' and 'Absurd' by assessing the corresponding information.
6. For 'YOLO', eliminated the records as they had the same information except for the ID and response.
7. For the 'Absurd' category in Marital Status, combined it with the mode of Marital Status which was 'Married'.
8. Made necessary substitutions or transformations to the categorical variables to prepare them for further analysis, such as creating dummy variables.

D. Dimension Reduction

1. The variables 'Age' and 'Year Birth' both contain the same data. As a result, we can remove 'Year Birth' to reduce data redundancy.
2. 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.
3. The 'ID' column provides no useful information for clustering customers. As a result, we can remove this column.
4. '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.
5. '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.

VI. Data Exploration

1. Classified education into Higher, Basic or Low level of education. And added into a column ’Education_Level’
2. Using Marital _Status to determine the Living status(Alone, Not Alone) of customers and adding it to a column ‘Living_Status’.
3. Analyzing total number of campaigns accepted and added it to a column ‘Total_Campaigns_Accepted’
4. Calculating average spend on products per purchase and adding it to column ‘Average_Spend’
5. Calculating total spent on products and adding it to column ‘Spent’
6. Checking for parent status and labeling 1 for yes and 0 for no in a column ‘Parent_Status’
7. Calculating total spending in the last 2 years and adding it to column ’total_spending(2yrs)’
8. Calculating average monthly visits to the company's website and adding it to ‘avg_web_visits’
9. Calculating ratio of online purchases to total purchases and adding it to ‘online_purchase_ratio’’

A. Data Visualization

Correlation Heatmap of all the numerical variables

After performing a correlation heatmap analysis on the 'marketing_campaign.csv' dataset, we found that most of the columns are not strongly correlated. However, we did find a strong positive correlation between the 'MntWines' and 'Income' columns. This suggests that customers with higher incomes tend to spend more on wines, which can be a useful insight for targeted marketing campaigns.

The correlation heatmap visually represents the correlation between different numerical features of the dataset. In the heatmap, colors range from blue to red, where blue represents negative

correlation, red represents positive correlation, and white represents no correlation. We observed a mostly blue and white heatmap, indicating low to no correlation between most of the columns. However, the 'MntWines' and 'Income' columns were strongly correlated, represented by a bright red square in the heatmap.

In conclusion, 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.

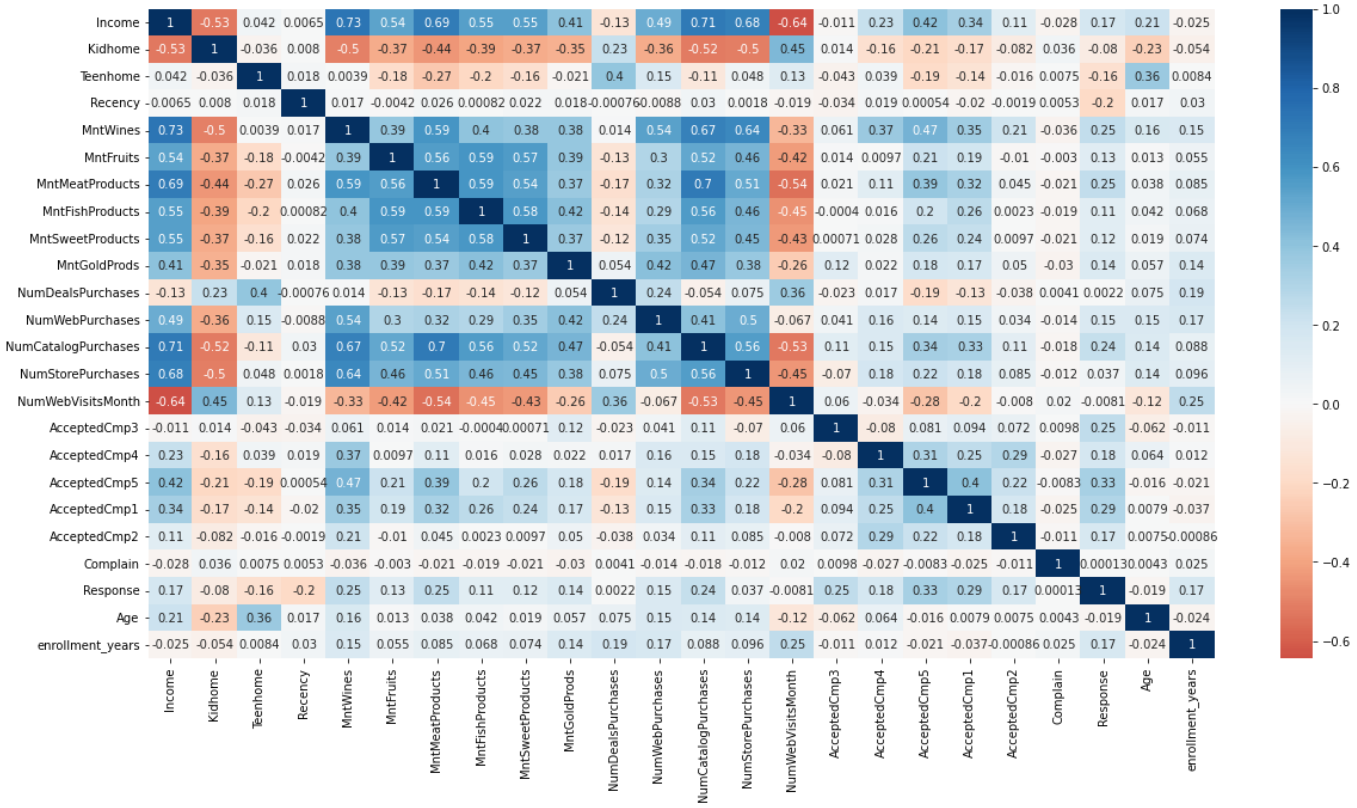

Fig 3: Correlation Heatmap

Distribution of Customer Age

Looking at the distribution of customer age in our dataset, 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. However, we still have a significant number of customers who fall outside this age range, and it is important to consider the different needs and behaviors of customers across different age groups when developing our marketing strategies.

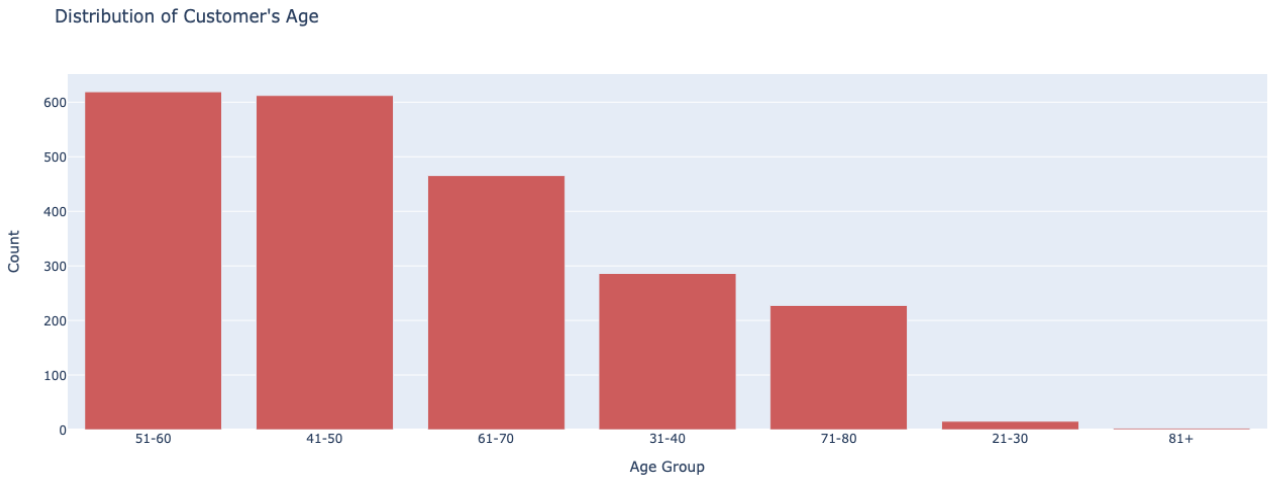

Fig 4 : Distribution of Customer's Age

Marital Status Distribution

When we look at the distribution of marital status among our customers, we can see that the 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. Divorced customers make up just over 10% of our customer base. Interestingly, we can see that we have very few widowed customers, making up less than 5% of the total.

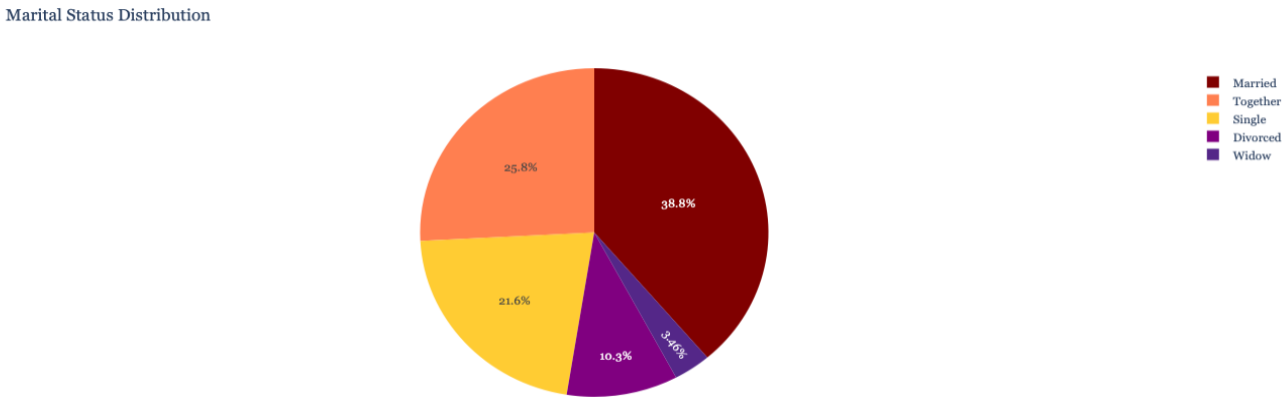

Fig 5 : Marital Status Distribution

Distribution of Average Spending by Marital Status

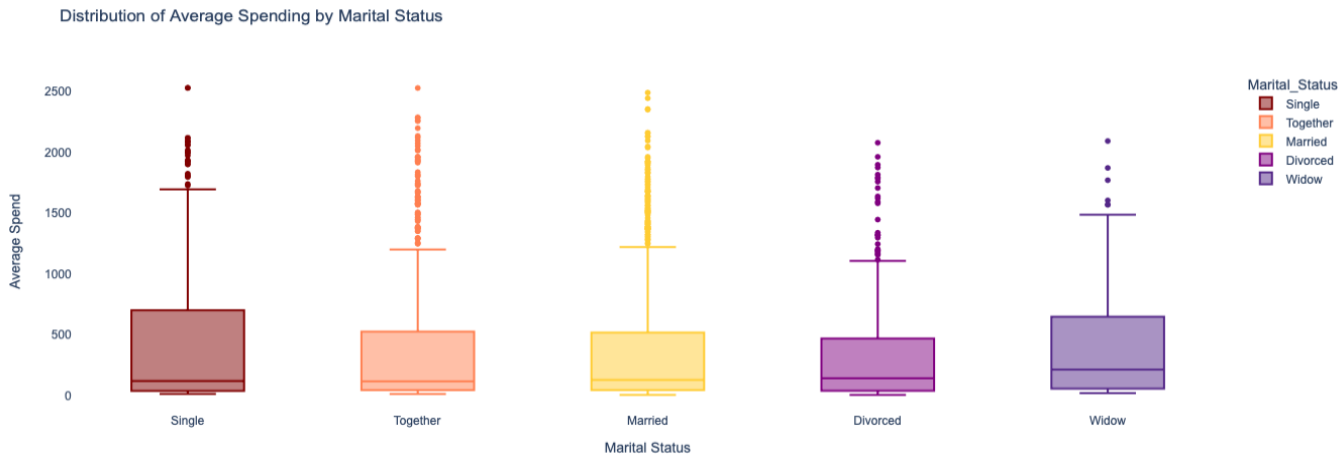

Fig 6 : Distribution of average spending by Marital Status

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.

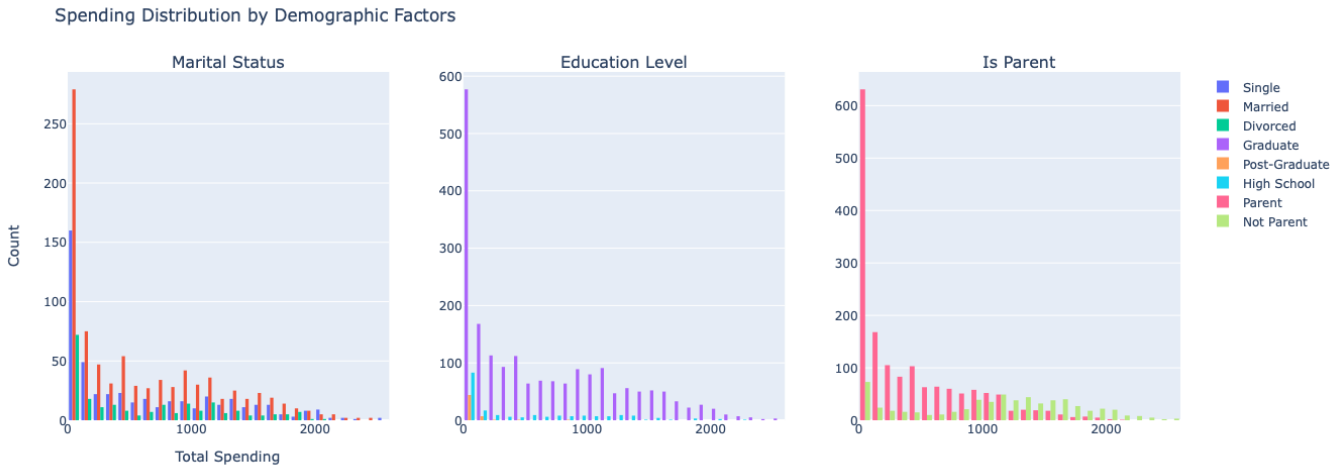

Fig 7 : Spending Distribution by Demographic Factor

VII. Data Mining Tasks

The data as we discussed before had many columns and so we performed dimension reduction methods of combining categories, combining feature, eliminating irrevent features, etc. while data cleaning.

However, before performing the actual clustering, following steps were performed to get the data ready.

1. The data obtained after data cleaning had 2 categorical columns 'Education' and
   'Marital Status' which were replaced by creating dummies.
2. At this step we had all features in numerical form but with different scales. Therefore, Standard Scaler was used to transform data and get all the features between values -1 to 1.
3. Before performing clustering, it is important to have minimum features, hence Principal Component Analysis was performed to reduce the dimensions further.
4. Three components were taken from the PCA which captured approximately 47.63% of variation in the data, which is not the best. But visualizing clusters would be harder if higher number of components were chosen.
5. Now, we have a scaled dataset on which PCA has performed and 3 components are chosen. This dataset is used as an input for further computation.

VIII. Data Mining Models

The aim of our project is to get clusters with the use of Unsupervised Machine Learning algorithms as we don't have a target variable to work with. Therefore, we focused on different Unsupervised Machine Learning Algorithms used for Clustering. While exploring, we worked with 6 Models that included Kmeans, Hierarchical Clustering, Mean Shift, Affinity Propagation, Spectral Clustering and Birch.

As a first step in modeling, the appropriate number of Clusters were found by using the Elbow Chart.

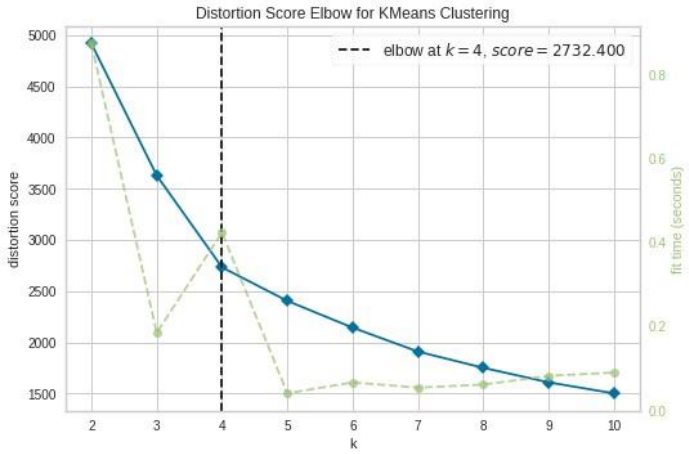

Fig 8 : Elbow Chart representing 4 clusters as optimum

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

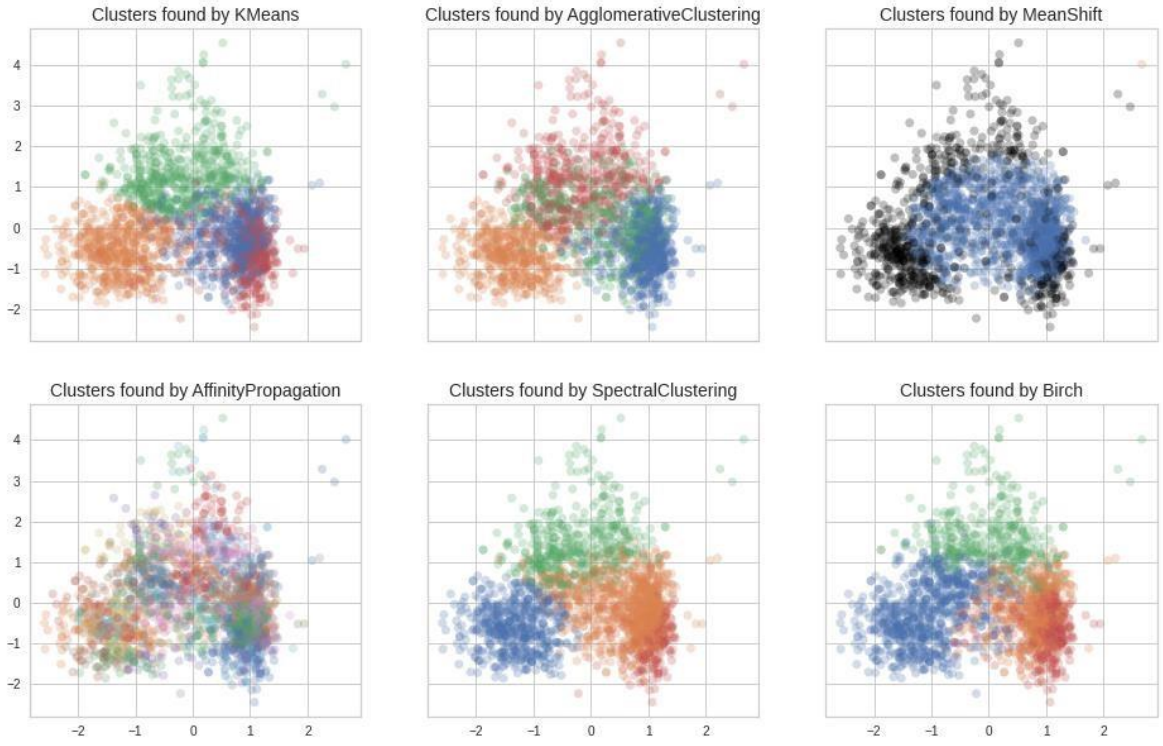

Fig 9 : Clustering formed by different models

IX. Performance Evaluation

To evaluate the performance of these different clustering algorithms, Silhouette Score and Dunn Index were calculated to select the best performing Model.

Sr. No.	Model	Hyper Parameters	Silhouette Score	Dunn Index
1	Kmeans	`'n_clusters':4`	0.325	0.028
2	Hierarchical Clustering	`'n_clusters':4, 'linkage':'ward'`	0.264	0.019
3	Mean Shift	`'cluster_all':False`	0.186	0.013
4	Affinity Propagation	`'damping' : 0.9`	0.253	0.016
5	Spectral Clustering	`'n_clusters':5`	0.228	0.014
6	Birch	`'threshold' : 0.01, 'n_clusters':4`	0.274	0.018

All algorithms except for Mean Shift have similar Silhouette Score. But, it is highest for
the Kmeans Model. Higher Silhouette Score and Dunn Index indicates better performance for the clustering.

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

X. Implementing the selected Kmeans Model

When we implement the Kmeans model, the performance metrics are as follows-

Silhouette Score = 0.325 Dunn Index = 0.028

Visualizing the Clustering Groups by applying the Kmeans model on the scaled data and principle components.

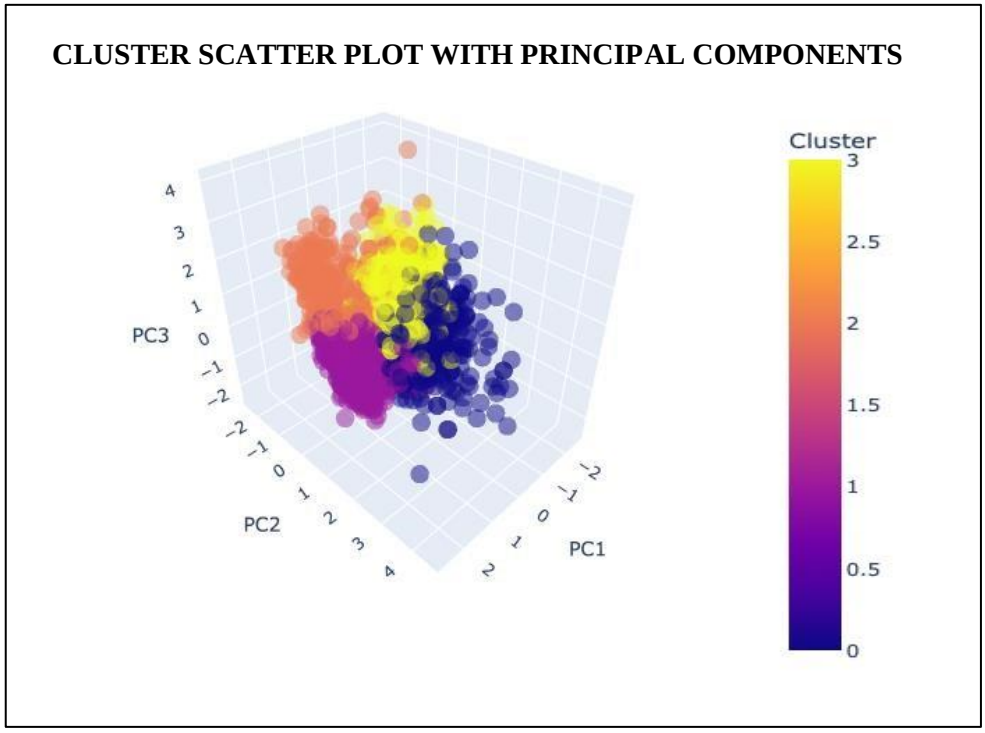

Fig 10 : Clustering formed by different models

XI. Interpretation of Clusters for Segmentation

Following graphs were used to understand the performance of the clusters on different parameters. Accordingly, they were categorized based on the customer behavior.

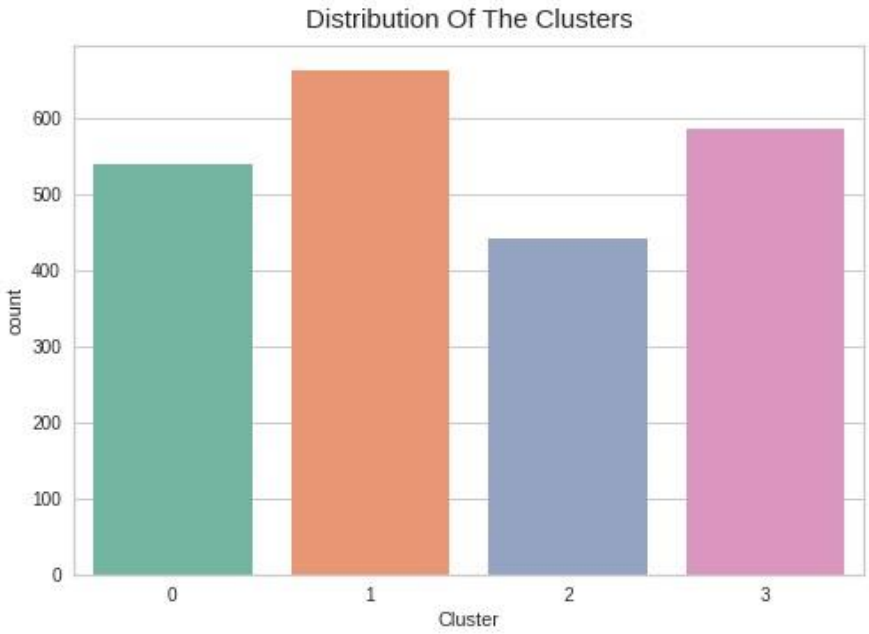

Fig 11 : Distribution of Clusters

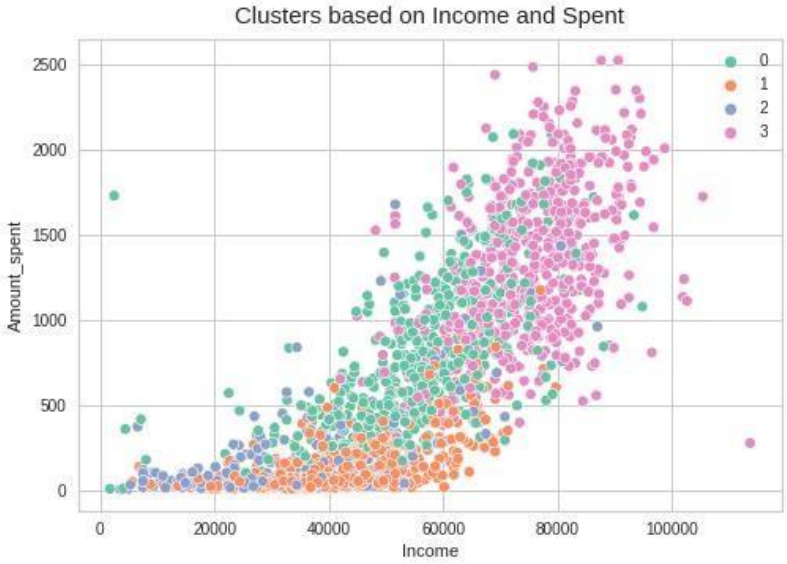

Fig 12: Clusters based on Income and Spent

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.

Table 2. Cluster Segmentation

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

Based on these Customer Segmentation, further interpretations were made based on distribution for different parameters in the dataset.

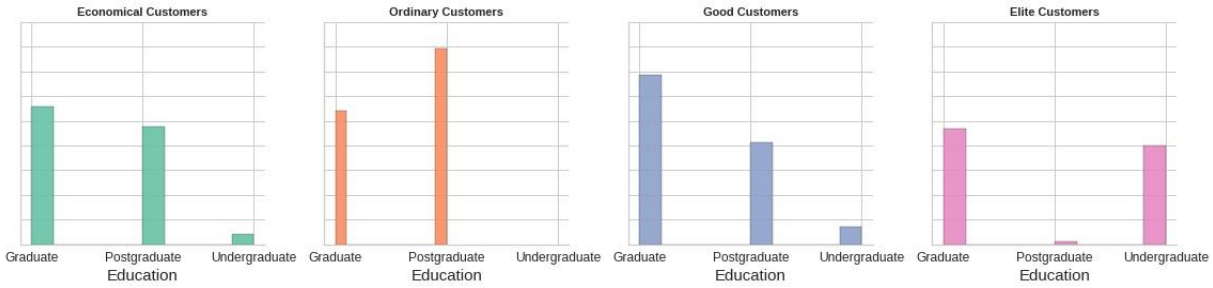

Fig 13: Education for different customer segments

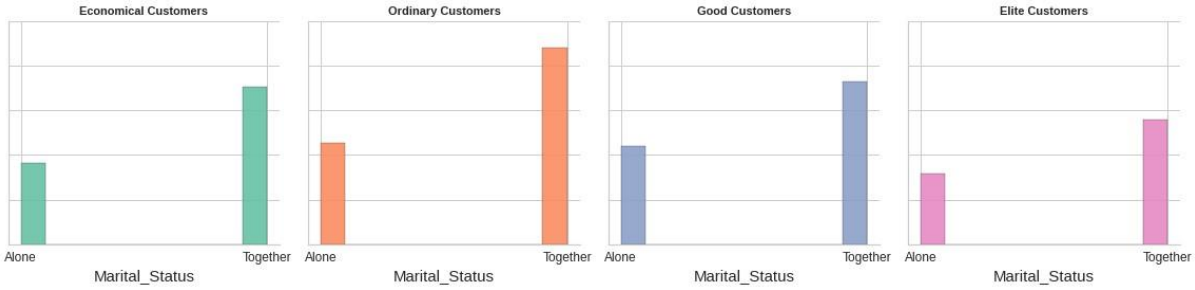

Fig 14: Marital Status for Different Customer Segments

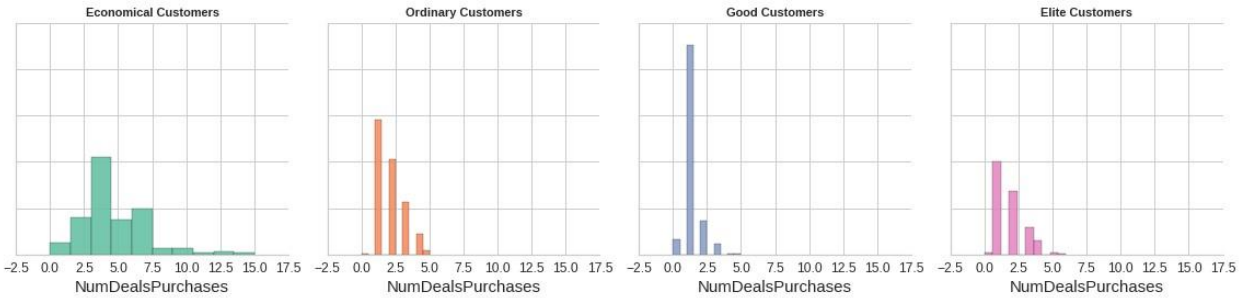

Fig 15: Deal Purchases for different customer segments

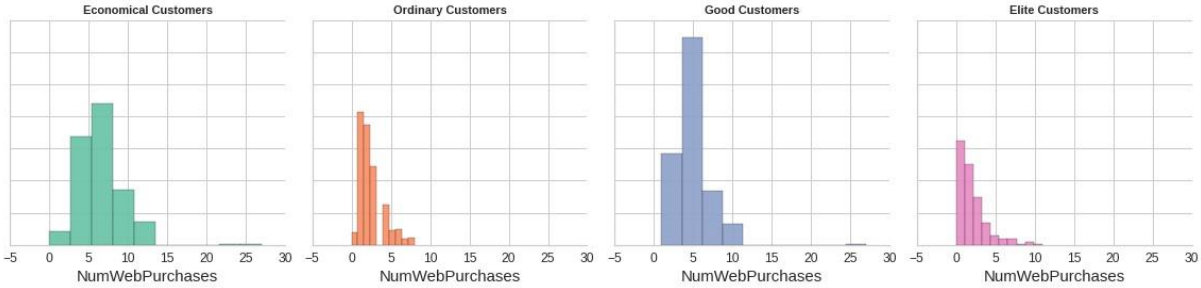

Fig 16: Web purchases for different customer segments

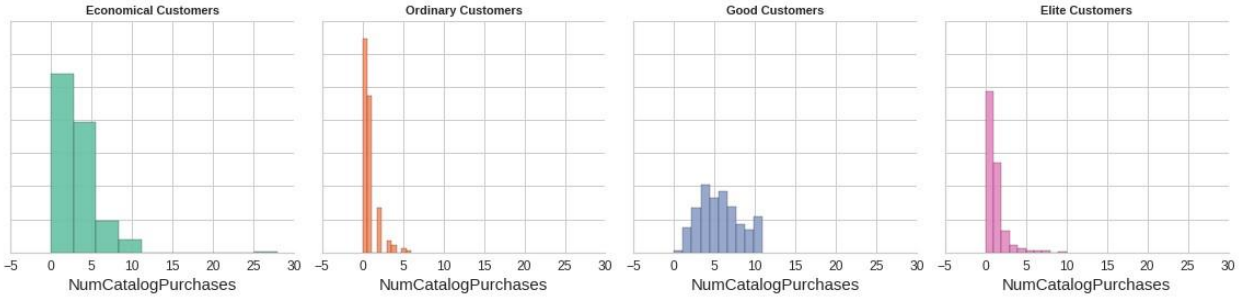

Fig 17: Catalog Purchases for Different Customer Segments

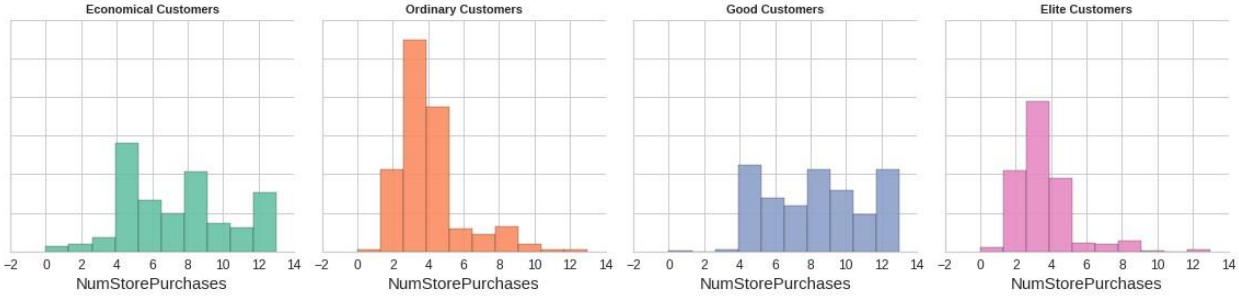

Fig 18: Store purchases for Different Customer Segments

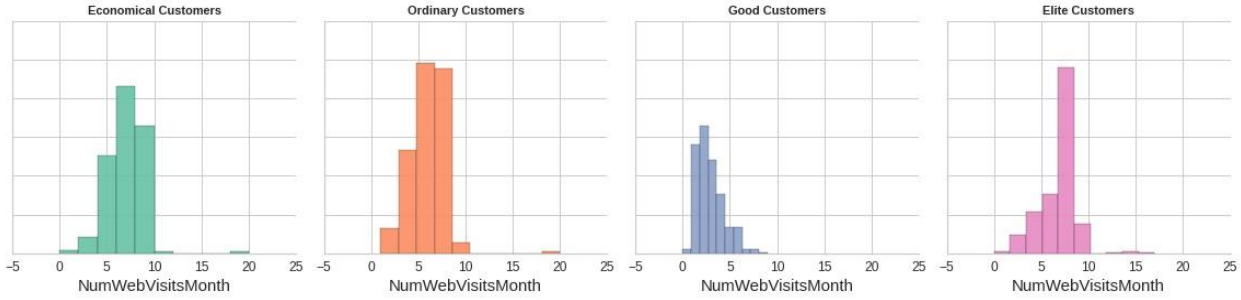

Fig 19: Web visits per month for different Customer Segments

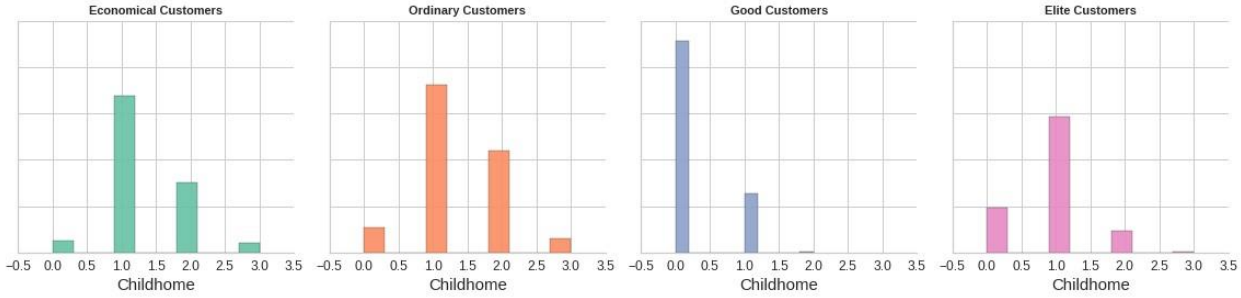

Fig 20: Children at Home for dsifferent Customer Segments

XII. Project Results

Kmeans algorithm along with PCA was used to perform the analysis. The performance of the model was evaluated by Silhouette Score and Dunn Index which came out to be 0.325 and 0.028 respectively.

The Clusters formed from these models were then categorized into different segments based on their nature of spending, Income, Children at home, Martial Status, etc. The graphs in previous interpretation resulted in 4 Customer Segments and their properties are as follows.

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.

XIII. Impact of Project Outcomes

Our aim of the project was to aid the Marketing Sector by generating different customer segments and then apply focused strategy for each of the group.

The data mining process provided us with the 4 clusters and the interpretations as we discussed before. These 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. Marketing efforts should be tailored to their unique needs and preferences. 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. Additionally, you can use email marketing and social media platforms to reach out to them with personalized deals and offers.