EDA Final Project

Customer Segmentation Analysis

Team:

Ayesha Tajammul Ahmed Mulla - amulla@iu.edu
Harshwardhan Patil -hrpatil@iu.edu
Radhika Agarwal - radagar@iu.edu

Abstract

Customer Segmentation is one of the most useful applications of unsupervised learning. It enables the companies to identify and group their customer base into different segments using clustering techniques. These segments are categorized based on various relevant factors such as demographics, purchasing behaviour and lifestyle characteristics that are useful for marketing purposes. By segmenting customers, companies can create targeted marketing strategies for each group, enhancing customer satisfaction and boosting sales.

Clustering, also referred to as a pattern recognition technique, plays a crucial role in discovering knowledge from multidimensional data. The primary objective of clustering is to find patterns of objects that exhibit similar behaviour within a dataset and group these objects under specific categories. With help of this technique we can uncover meaningful insights and hidden patterns from large and complex datasets to improve business decision-making processes.

This report provides an overview of customer segmentation analysis and its benefits including improved customer satisfaction, increases sales and reduced marketing costs. It discusses the different types of methods used in conjunction for analysis such as PCA and KMeans Clustering, etc. The report also states recommendations for businesses looking to implement customer segmentation analysis, emphasizing the need for effective segmentation criteria and the importance of adaptation to changing customer needs.

Objective

This project aims to analyze a customer personality dataset to segment customers based on their spending habits, with the goal of gaining insights into the personality traits and buying behaviors of customers. By doing so, companies can make data-driven business decisions and tailor their product supply and marketing strategies to the specific needs of different customer segments.

To achieve this aim, the project has two main objectives. The first objective involves performing exploratory data analysis (EDA) on the customer personality dataset to identify any patterns, trends, or outliers in the data. The second objective is to investigate the relationships between education and relationship status, age, income, spending habits, and purchasing behavior. Statistical analysis will be performed to identify any correlations and interesting trends.

The project will address several research questions to achieve these objectives. These questions include how a customer's education level and marital status impact their income, spending habits, and purchasing behavior. The project will also seek to identify any other interesting patterns or trends in the data and the different customer segments that can be identified based on their spending habits and other relevant variables.

The results of this project will provide valuable insights into customer segmentation, enabling businesses to make informed decisions based on customer behavior and preferences.

Research Questions

The research questions we aim to answer through the analysis are as follows:

1. In what ways do a customer's education level and marital status influence their income, spending habits, and purchasing behavior?
2. Besides the impact of education and marital status, are there any other interesting patterns or trends in the dataset that can shed light on customer behavior and preferences?
3. What are the various customer segments that can be identified in the dataset based on their spending habits and other relevant variables?

Data Description

The project utilizes a customer personality dataset sourced from Kaggle. The dataset for this project is provided by Dr. Omer Romero-Hernandez. The dataset comprises of a survey of customers of a business, where the purchase and spending information is represented in US dollars. In total, the dataset contains 2240 entries and 29 variables, which will be used to perform exploratory data analysis and statistical analysis in order to gain insights into customer behavior and preferences.

The variables used for analysis are as follows:

Personal Information of the Customer:

- Age: Customer's Age
- Income: Customer's yearly household income
- Education: Customer's education level (Undergraduate, Graduate, Post Graduate),
- Marital Status: Customer's Relationship Status. (Single, Widow, Divorced, Couple)

Channels of purchase made by the Customer:

- Web: Number of purchases made through the company’s website,
- Store: Number of purchases made directly in stores,
- Deals: Number of purchases made using deals,
- Catalog: Number of purchases made using a catalog.

Product Categories purchased by the Customer:

- Fruits: Amount spent on fruits in last 2 years,
- Wines: Amount spent on wine in last 2 years,
- Meat: Amount spent on meat in last 2 years,
- Fish: Amount spent on fish in last 2 years,
- Sweets: Amount spent on sweets in last 2 years,
- Gold: Amount spent on gold in last 2 years.

Methodology

Data Exploration and Analysis:

In order to answer the first research question we analyzed the relationship of customer's relationship status and Education level on Average amount spend on products individually.

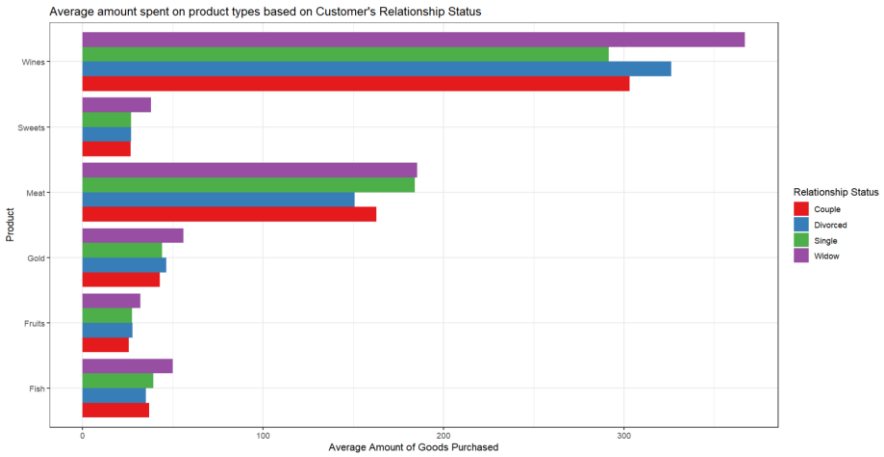

Fig. 1: Average amount spent on products based on Relationship status

Fig. 1 is the graph representing the Average Amount of Different Product types purchased based on Customer's Relationship Status. We can observe that Wine is the most purchased product and the least amount of purchases were made on sweets and fruits irrespective of the relationship status. Widow is the highest purchaser of all the product types with wine being the highly purchased followed by meat purchase. Divorced Customers are the second highest purchasers of products

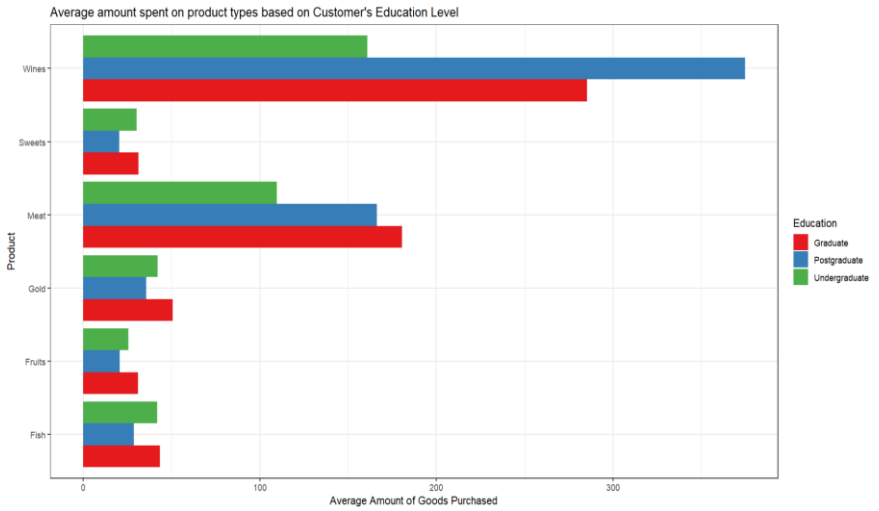

Fig. 2 Average amount spent on products based on Education Level

Fig. 2 is the graph representing the Average Amount of Different Product types Purchased based on Customers Education level. We can observe that similar to Fig. 1 observation, overall, wine is the most purchased product and the least amount of purchases were made on sweets and fruits irrespective of the customer's education followed by meat purchase. In this

graph we can identify that customers with post-graduate level of education purchase the most on products followed by graduates.

We further continued to study and confirm our analysis of the impact relationship status and education level have on customer's purchasing behaviour. We plotted customer's preference of the channel of purchase with respect to the two primary variables.

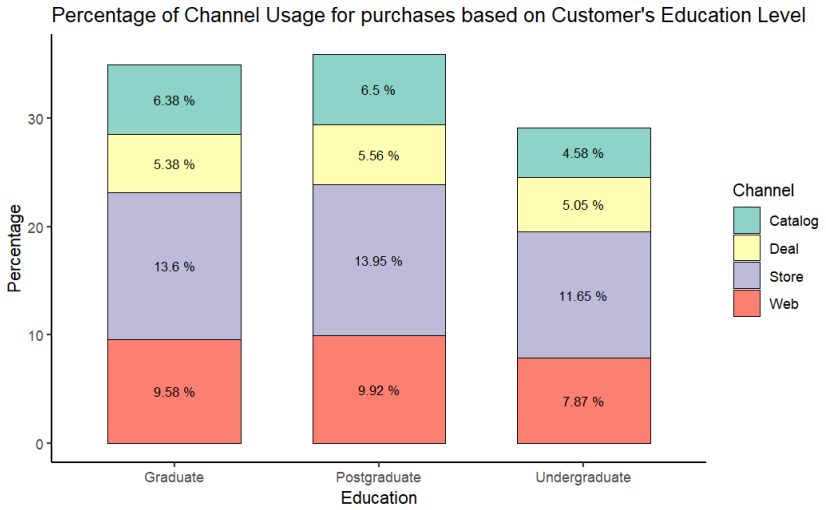

Fig. 3 % of channel usage for purchases by Education

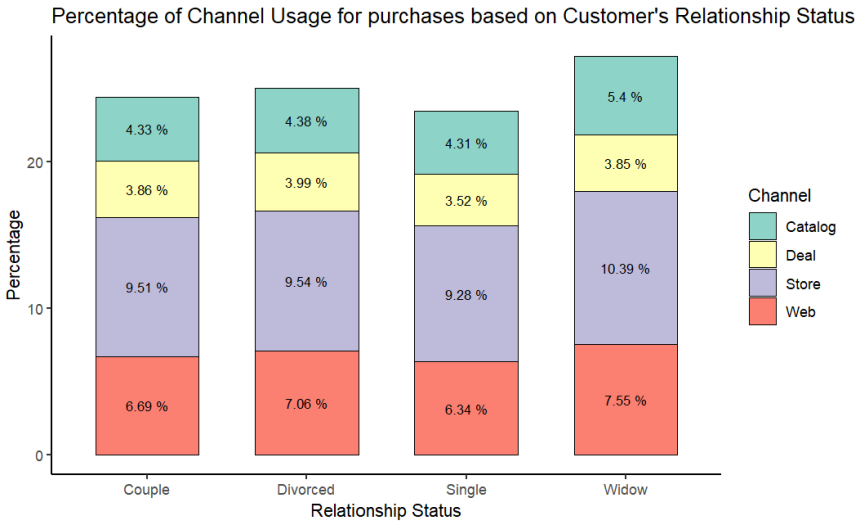

Fig. 4 % of channel usage for purchases by Relationship Status

The above plots (Fig. 3 and Fig. 4) represent the percentage of purchases in the made through different channels by customers in the last 2 years with respect to their education level and relationship status. From both the figures we can observe that store purchases are the highest and deal purchases are the lowest irrespective of the impact of qualifications and relationship status. Specifically, customers with Post Graduate and Graduate level of education have made the highest percentage of purchases by going to the store followed by web purchases. Similarly, Customers who are widows made the highest percentage of store purchases followed by web purchases, whereas all other customers have approximately spent similar amount on shopping

through different channels. Hence, based on the observations and patterns discovered, we can confirm that Education Level and Relationship Status of Customer has an impact on the purchasing behaviour of the customer and are important to identify the clusters in the data.

In order to answer the second research question we analyzed the impact of other variables such as age and combined effect of age with income on the purchasing behaviour of the customer.

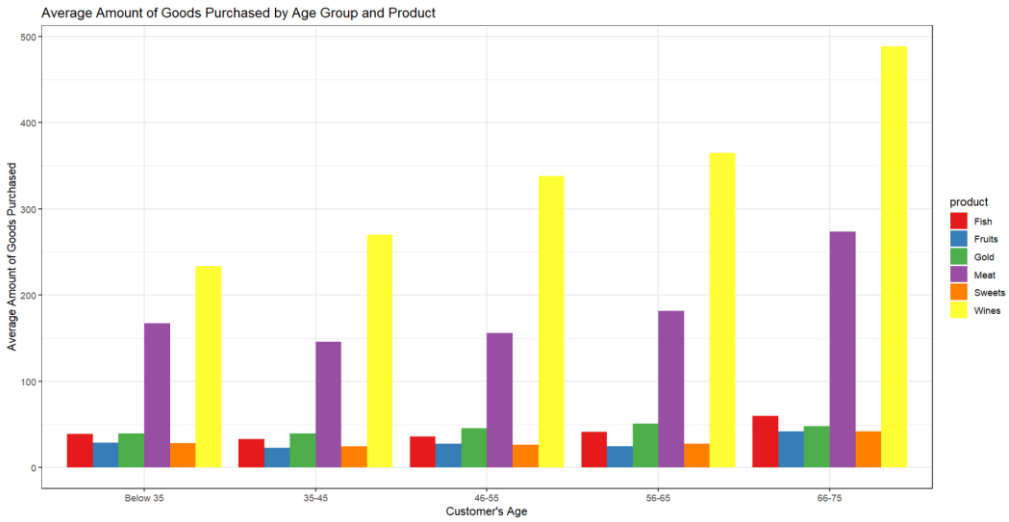

Fig. 5 Average amount of goods purchased by age group and product

The above plot in Fig. 5 is to study the impact of Customer's Age on their purchasing behaviour. The observation from the above plot is that the relationship between age and average amount spent on products have a linear and positive correlation. As the age increases, the expenses also increase with the highest purchase made by senior customers falling in the age group of 66-75 and the lowest amount spent by customers having age below 35. Both categories of customers spent the highest on wine and meat.

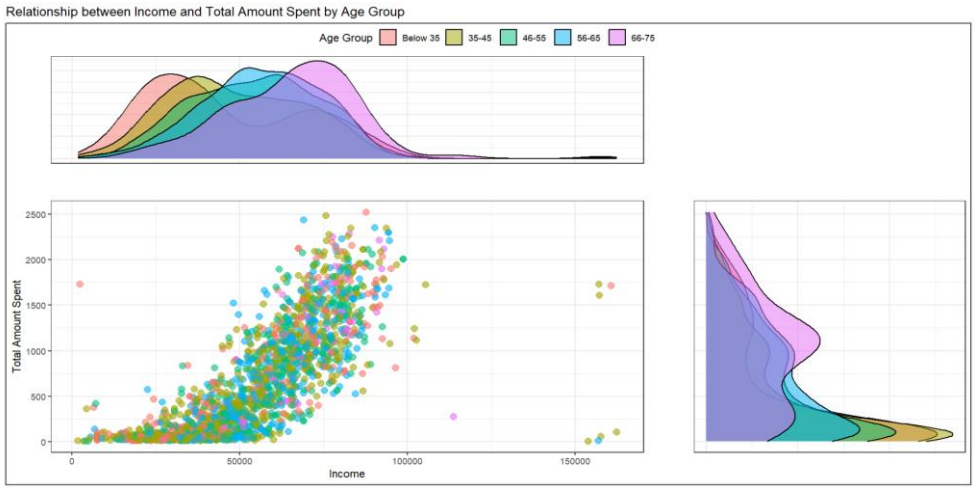

Fig. 6 Relationship between Income and Total Amount spent by Age Group

The graph in Fig. 6 is used to analyze customer behaviour based on both income and age. It can be seen that both the variables have a positive correlation with total amount spent by customers. With increase in age, the income increases and so does the amount spent by the customers.

Customer Segmentation:

The next question is what are the various customer segments represented in the dataset?

The “Customer segmentation” is the process of dividing customers into groups based on their common characteristics such as demographics, buying behaviour, preferences, and needs. It helps businesses to better understand their customers and create targeted and more effective marketing campaigns. One way to perform customer segmentation is through clustering, a technique that groups together data points with similar attributes.

Clustering algorithms seek to identify similarities and differences between data points, and group them into clusters based on those similarities and differences. There are various clustering algorithms available, including k-means, hierarchical clustering, and DBSCAN. In this project, we performed customer segmentation using the k-means algorithm in R.

K-means is a type of unsupervised machine learning algorithm that groups similar observations together into clusters based on their similarities in terms of chosen features. The "k" in k-means represents the number of clusters that the algorithm will create, which must be specified by the user.

Before performing customer segmentation, we pre-processed the data by removing missing values and outlier. We also performed correlation analysis to identify highly correlated features and removed them to avoid multicollinearity in the model. We set the cut-off at 70% so that the highly correlated features can be removed from the data.

After removing the highly dependent features, for further operations the columns in the data are:

Income, Recency, Tenure, Age, Wines, Fruits, Meat, Fish, Sweets, Gold, RelStatus, LevEd, WebVisits, Web, Deal, Store

After performing correlation analysis, we normalised the data. Also, we performed Principal Component Analysis (PCA) to reduce the dimensionality of the data and visualize the clusters in two-dimensional space. Then, we selected the above columns from the processed dataset and applied the K-means clustering algorithm to the data, selecting the optimal number of clusters using the elbow method.

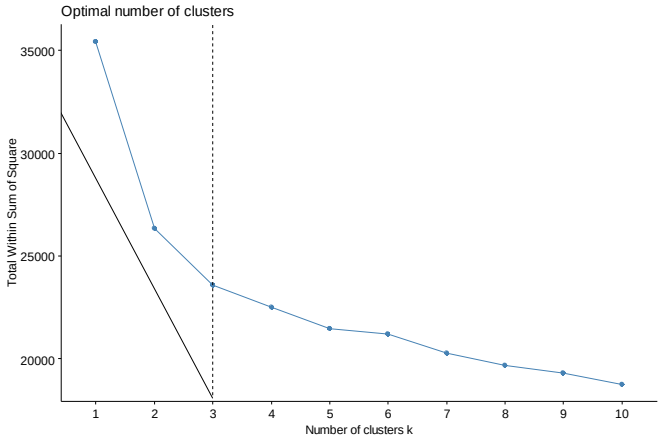

Fig. 7 Optimal Number of Clusters

The elbow method works by plotting the variance explained as a function of the number of clusters and then selecting the number of clusters at the "elbow" of the plot, where the increase in variance explained begins to level off. This method helped us to determine the number of clusters. Here, we got the cluster number as 3, and using K=3, we visualised the clusters in 2D space.

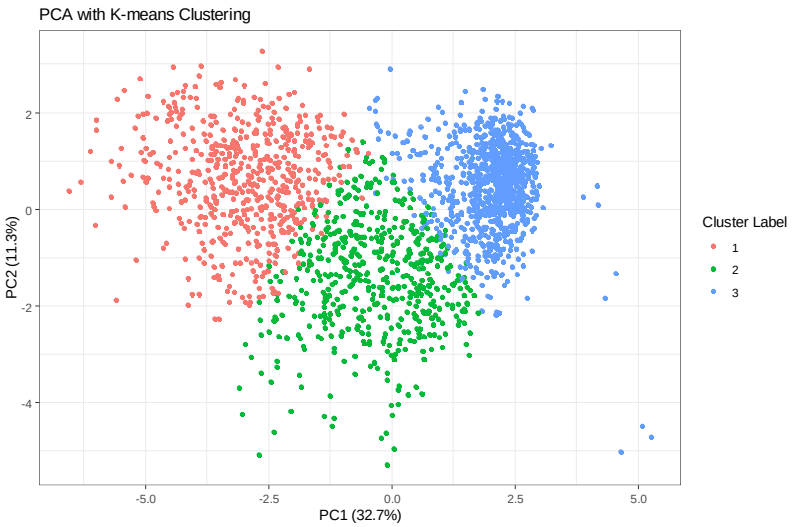

Fig. 8 PCA with K-Means Clustering

In order to do the analysis and find more information on the characteristics of the 3 clusters, we can plot a boxplot with respect to all selected variables.

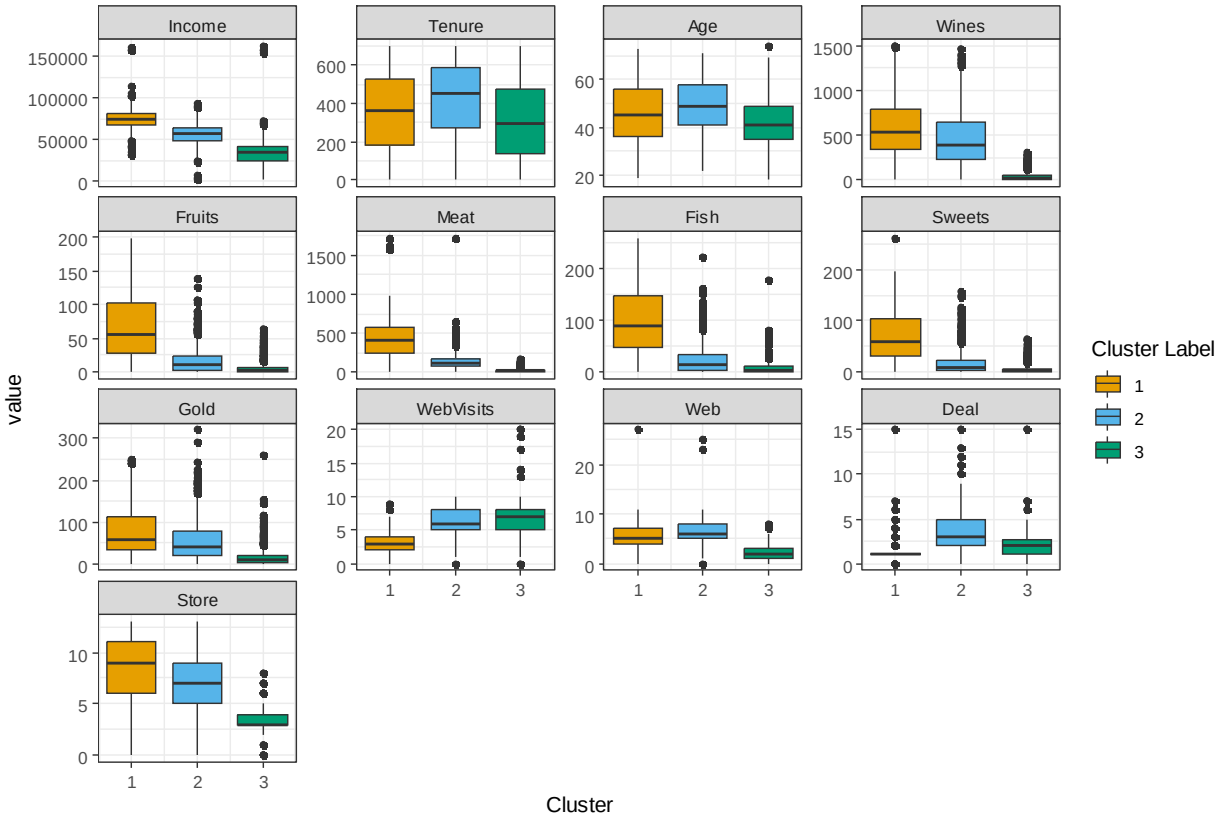

Fig. 9 Boxplot of variables by Cluster Label

Our analysis identified three distinct customer segments:

Cluster 1:

- They have highest income.
- Their median age is about 45.
- They spend highest in all the product categories.
- In general, they tend to spend more money on purchasing wine and meat.
- They make the highest store purchases.
- They do not shop via deals.

Thus, we will call the customers in cluster 1 as impulsive spenders.

Cluster 2:

- They have average income.
- Their median age is almost 50.
- They have been with company the longest.
- They spend moderate amount in each product category.
- They give visits to websites more often than cluster 1.
- They shop via all modes i.e., web, Deals and store.

Thus, we will call the customers in cluster 2 as Loyal Customers.

Cluster 3:

- These are the customers with lowest income.
- These customers are the most recent members of the company and relatively younger people (with approx. median age 41).
- They tend to visit the website more than other 2 clusters but tend to buy the lowest amount via web.
- Also, they are lowest spenders in all product categories.

Thus, we will call the customers in cluster 3 as Window Shoppers or Surfers.

By categorizing customers into these segments (Impulsive spenders, Loyal customers, Surfers), companies can tailor their marketing and product strategies to better meet the needs of each group. For instance, impulsive spenders may respond well to promotions and discounts, while loyal customers may benefit from personalized services and loyalty programs. Additionally, companies can prioritize improving website user experience for surfers. These insights can help companies make data-driven decisions and better allocate resources to increase customer satisfaction and ultimately, drive revenue growth.

Results

The results of the analysis showed that customer education and relationship status had a substantial impact on income, spending habits, and purchasing behavior. Moreover, it was

observed that the mode of purchase varied depending on the customer's level of education and relationship status.

Based on the spending habits and other relevant variables, the customer personality dataset was segmented into three distinct customer groups: Impulsive spenders, loyal customers, and Surfers. These groups exhibited unique characteristics, preferences, and behavior patterns that could be used by businesses to tailor their marketing strategies and improve customer satisfaction.

Conclusion

Based on the analysis of the customer information, their purchasing behaviour and methods used for clustering the data we could successfully identify that customers can be segmented into 3 clusters. Based on the characteristics of the clusters we have named them as “Impulsive Spenders”, “Loyal Customers'”, and “Surfers”.

“Impulsive spenders” are customers who make impulsive purchases and have the highest income and tend to spend more money on purchasing various products, including wine and meat. They make the highest store purchases and do not shop via deals, indicating that they may not be concerned about finding the best deals or bargains.

On the other hand “Loyal customers” have an average income, are around 50 years old, and have been with the company for a long time, indicating their loyalty to the brand. These customers spend moderately on products but frequently visit the company's website and shop via all modes. Their spending behavior suggests that they prioritize value and quality over luxury and are not impulsive shoppers. Overall, Loyal Customers are valuable to the company and should be rewarded through personalized experiences, promotions, and rewards programs to encourage repeat business.

And the last type of customers is “Surfers”. They are characterized by having the lowest income, being relatively young with a median age of around 41, and being the most recent members of the company. While they tend to visit the company's website more frequently than the other clusters, they tend to purchase the lowest amount via the web. Additionally, they are the lowest spenders across all product categories. The behavior of these customers suggests that they are interested in the brand and its products, but may not have the financial means to make significant purchases. Companies could benefit from engaging with these customers through targeted marketing and promotions that cater to their budget constraints and interests, which may convert them from window shoppers to actual customers.

Based on this interpretation, companies can modify their product supply and marketing strategies to cater to each cluster's specific needs and behaviors. For example, they can offer more discounts and promotional offers to the impulsive spenders, provide personalized services and loyalty programs to loyal customers, and focus on improving the website's user experience for surfers. They are more likely to make purchases using catalogs and visit the company's website, but they are less likely to make purchases directly in the stores.

Future Scope

To further understand customer behavior and preferences, additional surveys could be conducted to gather more data. This could provide valuable insights into customer needs, wants, and motivations.

Moreover, incorporating additional customer attributes such as gender and exploring the effect of combined relationship status and qualifications could lead to a more comprehensive understanding of the factors influencing customer behavior.

It would also be interesting to examine the impact of customer satisfaction levels on spending habits and purchasing behavior. This could help businesses improve their products and services, ultimately leading to greater customer satisfaction and loyalty.

Reference

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

https://www.researchgate.net/publication/356756320_Customer_Segmentation_Using_Machine_Learning

https://www.qualtrics.com/experience-management/brand/customer-segmentation/