S022 - Spring 2023

Final Project

Wan, Xinyi

Customer Personality Analysis Using K-means Clustering

Customer Personality Analysis Using K-means Clustering

Introduction

Customer Personality Analysis is a comprehensive examination of an organization's preferred customers. It provides businesses with a deeper understanding of their customers, allowing them to customize their products and services to meet the unique needs, behaviors, and concerns of different customer segments. By conducting a customer personality analysis, a company can identify which customer segment is most likely to be interested in a new product, enabling them to focus their marketing efforts on that specific segment rather than targeting all customers in their database. This approach allows for more targeted and efficient marketing strategies, ultimately leading to increased sales and customer satisfaction.

Data

The dataset for this project is from a Kaggle project, provided by Dr. Omar Romero-Hernandez (see https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis). The dataset includes 27 attributes from four dimensions: People, Products, Promotion, and Place (detailed description of the variables can be found in the Appendix). To facilitate analysis, six new variables were created from the original dataset.

- Age: the age of the customer by calculating from “Year_Birth”
- Days_Customer: the number of days since customer’s enrollment with the company
- Accepted: the total number of accepted campaign activities
- Total_Spent: the sum of the amount spent on six product categories
- Child_Num: the number of children and teens at home

The original dataset consists of 2240 observations. A cleaned dataset of 2215 observations was created for the purpose of data analysis. 24 observations with missing values were dropped, together with one observation with unusually high income.

Methods

This project used k-means clustering to segment customers. Three was chosen as the optimal number of clusters using the "within-cluster sum of squares" (WSS) method. The WSS method computes the sum of squared distances between each data point and its corresponding cluster center within each cluster. The optimal number of clusters is visualized as the "elbow point" on the plot where the rate of decrease in WSS starts to level off, as shown in Figure 1. At the point of three, the data is most compact, and the clusters are most distinct from one another.

After the clustering analysis was performed, this project conducted a descriptive data analysis to better understand the purchasing habits and demographic characteristics of customers in each of the identified clusters. The descriptive data analysis aimed to provide a clear picture of the unique traits that distinguish each cluster from the others.

To further explore the relationships between different variables and identify the most significant predictors for each cluster, this project employed the Random Forest algorithm. This algorithm was used to build an ensemble of decision trees based on a subset of the variables, and it helped to identify the variables that were most important in distinguishing between the three clusters. 10-fold cross validation was performed to find the optimal number of variables for each subset.

Results

The K-means clustering algorithm divides the customers in the dataset into three distinct segments. The distribution of customers among these segments is relatively balanced, with 811 customers in cluster 1, 733 customers in cluster 2, and 631 customers in cluster 3.

Figure 2 displays the visualization of k-means clustering on a scatterplot of the first two principal components of the dataset, with the color of the point corresponding to the cluster to which the observation belongs. Figure 3 presents a scatterplot of the relationship between income and total spendings with the same color palette showing the cluster information. The graph shows a distinct separation between the three clusters observed on income levels, with cluster 3 having the highest income and some observations with extremely high income values over $150,000. In contrast, customers in cluster 2 have the lowest income, falling below $40,000. Additionally, the scatterplot revealed a clear positive correlation between income and spending.

Characteristics of each cluster are described in plots from Figure 4 to Figure 9. In Figure 4, the shades of color on each row are used to present the mean value of the corresponding variable of each cluster. For example, the darkest shade on the first row indicates cluster 3 has the highest mean income, while cluster 2 of the shallowest shade has the lowest mean income. Overall, customers in cluster 3 have the highest mean income, the highest mean total spending, the highest mean spending on each product category, and the highest number of accepted campaigns. In contrast, customers belonging to cluster 2 exhibit the lowest average values for the aforementioned variables. However, they have the highest average values for both the complaint rate and the number of web visits made in a single month.

Figure 5 and Figure 6 display the variance of amount spent on grocery and luxury products between the three clusters. The basic necessities for living, such as meat, fish, sugar and fruits, are classified as grocery products. On the other hand, luxury products are generally considered to be more expensive, extravagant, and not essential for survival, examples of which include gold and wines. In accordance to the observations from Figure 4, cluster 3 remains to be the segment of highest spendings, and cluster 2 the lowest. The gaps between cluster are wider on wines and meat products than on other grocery products and gold.

According to Figure 7, customers belonging to cluster 1 and cluster 2 utilized more discounts for their purchases compared to customers in cluster 3. None of the three clusters showed a distinct mean value of accepted promotional activities. In Figure 8, the purchasing behavior through various mediums (such as catalog, store, and website) was compared across the three clusters, and it was found that in-store purchases were the most common method overall. Customers in cluster 3 had the highest number of purchases through all three methods, which aligns with their highest spending habits. However, the difference between the cluster 1 and cluster 3 was subtle with regards to the number of purchases made on websites.

Customers' demographic information was summarized in Figure 9. In general, there were no substantial differences observed in terms of education level and marital status among the three customer groups. Cluster 2 stands out as the sole cluster comprising customers with a basic

education level. Additionally, it has the highest number of single customers, which corresponds to the observation depicted in Figure 4 where cluster 2 exhibits the lowest mean age among the three clusters.

Table 1. Variable importance from random forest model

Variable	Importance
Income	100.00
MntWines	5.55
Total_Spent	2.65
MntMeatProducts	1.81
NumWebVisitsMonth	0.21
NumDealsPurchases	0.14

*only variables with importance over 0.1 shown (out of 30)

Table 1 shows the results of variable importance from the random forest model with the optimal mtry value selected by 10-folds cross-validation. The numbers represent the variable importance measures for each predictor variable in the random forest model. Higher values indicate greater importance, suggesting that those variables have a stronger impact on the model's accuracy or predictive performance. Income and the spendings on wines were selected as the most important predictors, which align with the observations from Figure 3 and Figure 6.

Conclusion

The plotted features show distinct distinctions between clusters when applying the k-means algorithm. Based on the descriptive analysis of cluster characteristics, the three clusters primarily differ in terms of income and purchasing habits, while the disparities in demographic information are more subtle. In general, the three clusters can be roughly described as average-spend customers, low-spend customers, and high-spend customers.

- Cluster 1 – average-spend customers: Customers in cluster 1 are average-spend customers who have the medium level of income and spendings on all product categories. They are deal-seekers who are more likely to make purchases with discounts. Therefore, tailoring

promotions can be effective in attracting and retaining these customers. Since they are not frequent lookers or surfers, user-friendly and intriguing web interface can be helpful in attracting their attention and converting them into potential customers.

- Cluster 2 – low-spend customers: Customers in cluster 2 are low-spend customers who have the lowest level of income and spendings on all product categories. However, they are potential customers who frequently browse through the company website and look for promotions, which is suggested by the highest web visits history and the large number of purchases made with offers. Campaign activities can be effective in intriguing these customers to make purchases, and the promotions can be applied specifically to low-end products, which aligns with their purchase habits and income level.
- Cluster 3 – high-spend customers: Customers in cluster 3 are high-spend customers with the highest level of income and spendings on all product categories. They are loyal customers and main source of revenue for the company. It is important to examine their purchase habits and build trustworthy relationship with these customers. Getting feedback from them can be valuable to improve brand building, campaign customization and purchase experience. Furthermore, to encourage additional spending, promotions can be targeted towards complementary products based on customers' purchasing habits. For instance, if it is observed that wines are frequently purchased alongside meat products, promotions can be extended to both categories simultaneously.

In conclusion, employing clustering techniques to analyze customer behavior can aid the company in gaining a deeper understanding of customers' purchasing habits and will facilitate the development of customized strategies for each customer cluster.

Greatest Challenge

One of the greatest challenges of this project was to expand my shallow knowledge of clustering through self-learning to make the data-analysis process comprehensive and rigorous. Before, my understanding of clustering was limited to using functions for data segmentation through k-means clustering and hierarchical clustering. I relied on the scree plot to determine the optimal number of clusters and interpreted the results based on mean values. However, I have since expanded my knowledge. I have learned how to visually represent clustering information on scatterplots,

allowing for a more intuitive understanding of the clusters. I have also learned to combine the interpretation process with other types of plots, such as box plots and bar plots, to gain deeper insights. Additionally, I have explored the integration of clustering with random forest algorithms to determine the importance of variables in the clustering process.

Another challenge of this project is inferring meanings from the results. Combining clustering methods with background knowledge in business is crucial for extracting meaningful insights and making informed decisions. While clustering algorithms can provide valuable patterns and groupings in the data, the interpretation of these results requires a contextual understanding of the specific business domain. As a non-business major student, this task can be challenging due to the lack of prior exposure to business concepts and industry-specific knowledge. However, I tackled this difficulty through stepwise analysis and visualization on each variable until the results were self-explanatory.

Declaration of Skills Used

- I used fundamental R skills taught in Unit 1. I used ggplot to generate boxplots and barplots, and displayed them with facet_grid() function.
- I used the tidyverse taught in Unit 2. I used pivot_longer() function to prepare my data for visualization.
- I used classic machine learning tools taught in Unit 4. I used random forest to find the most impactful variables in the dataset and used k-fold cross-validation to find the optimal value for mtry.
- I used k-means clustering, which is an extension in line with the course, and found online resources on how to apply the function and how to visualize clustering results.

Appendix – Original Variables

People

- ID: Customer's unique identifier
- Year_Birth: Customer's birth year
- Education: Customer's education level
- Marital_Status: Customer's marital status
- 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

- MntWines: Amount spent on wine in last 2 years
- MntFruits: Amount spent on fruits in last 2 years
- MntMeatProducts: Amount spent on meat in last 2 years
- MntFishProducts: Amount spent on fish in last 2 years
- MntSweetProducts: Amount spent on sweets in last 2 years
- MntGoldProds: Amount spent on gold in last 2 years

Promotion

- NumDealsPurchases: Number of purchases made with a discount
- AcceptedCmp1: 1 if customer accepted the offer in the 1st campaign, 0 otherwise
- AcceptedCmp2: 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
- AcceptedCmp3: 1 if customer accepted the offer in the 3rd campaign, 0 otherwise
- AcceptedCmp4: 1 if customer accepted the offer in the 4th campaign, 0 otherwise
- AcceptedCmp5: 1 if customer accepted the offer in the 5th campaign, 0 otherwise
- Response: 1 if customer accepted the offer in the last campaign, 0 otherwise

Place

- NumWebPurchases: Number of purchases made through the company’s website
- NumCatalogPurchases: Number of purchases made using a catalogue
- NumStorePurchases: Number of purchases made directly in stores
- NumWebVisitsMonth: Number of visits to company’s website in the last month

Appendix – Plots

Figure 1. Optimal cluster number using WSS method

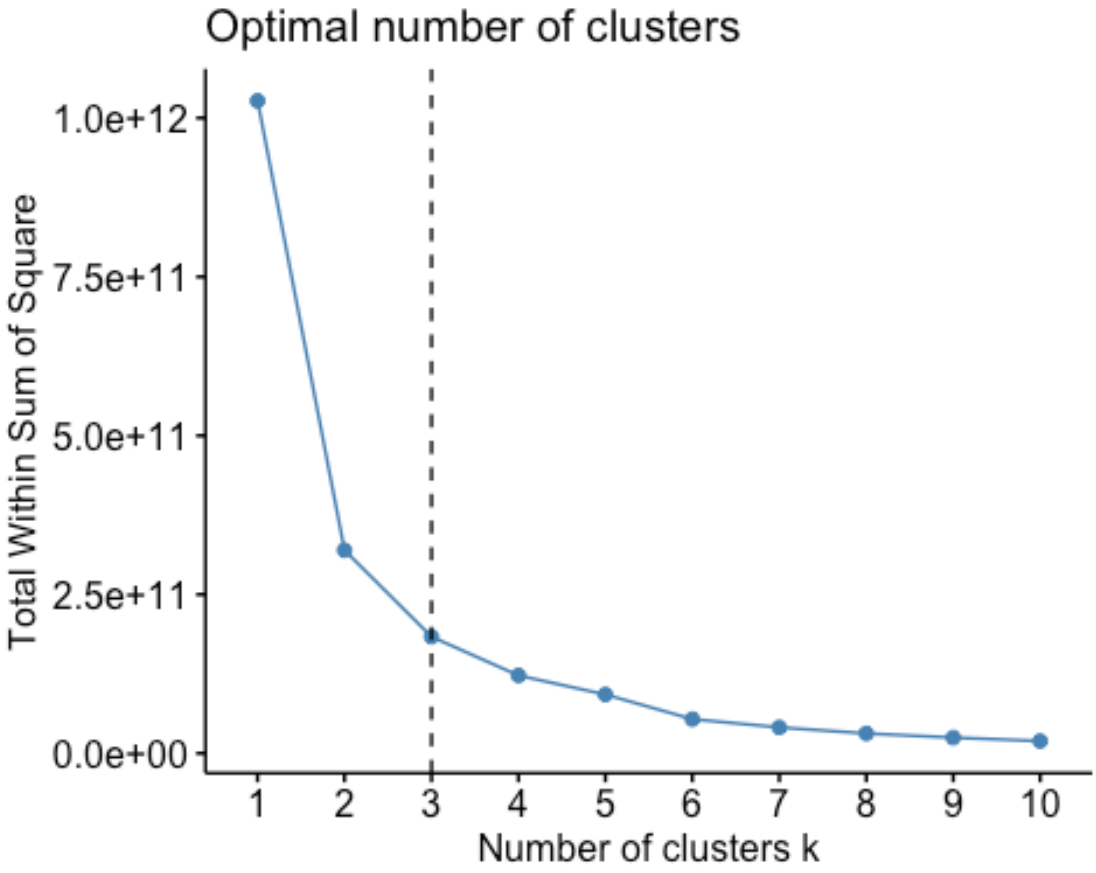

Figure 2. K-means clustering

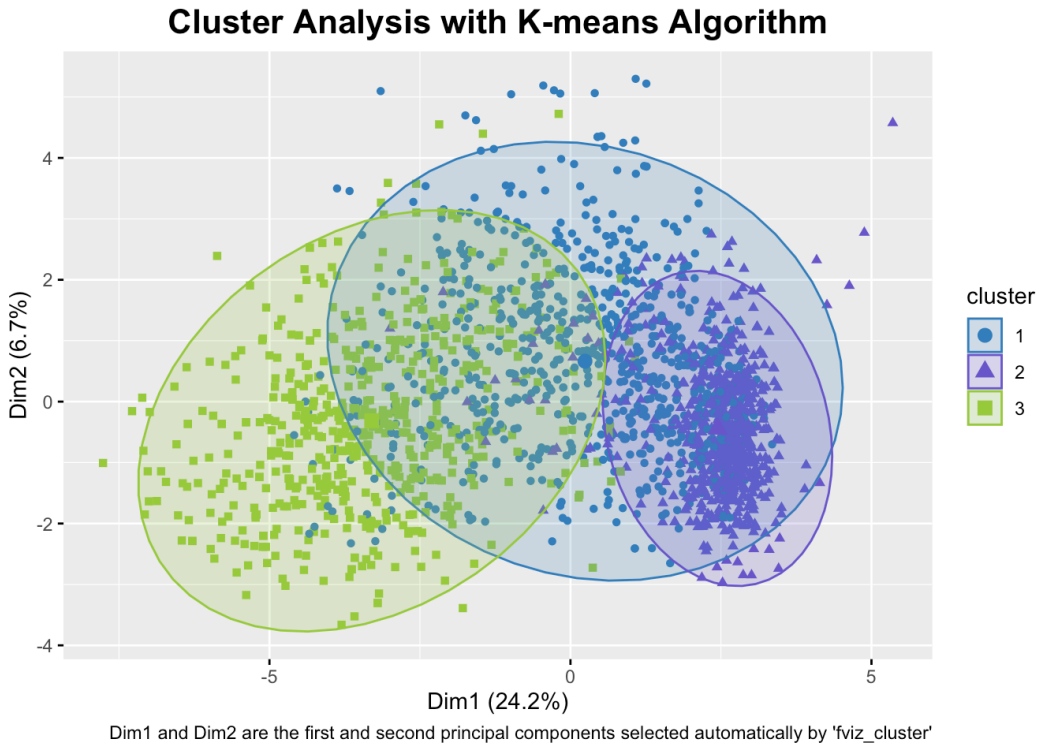

Figure 3. K-means clustering on income and total spendings

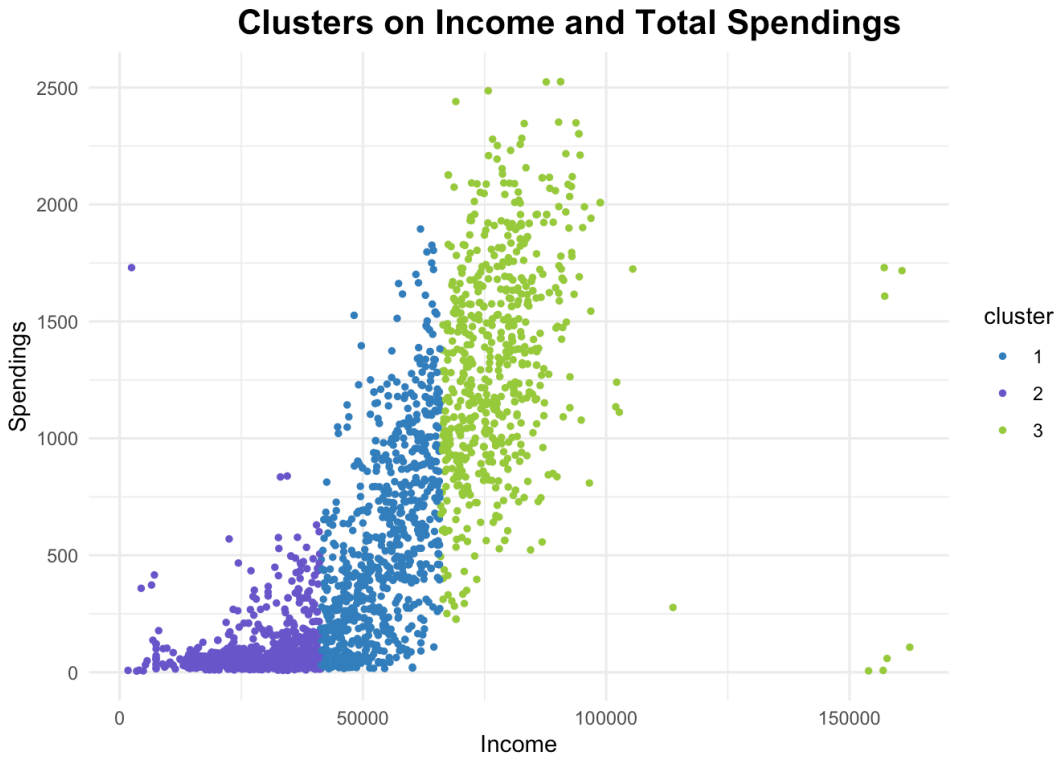

Figure 4. Between-cluster differences on each variable

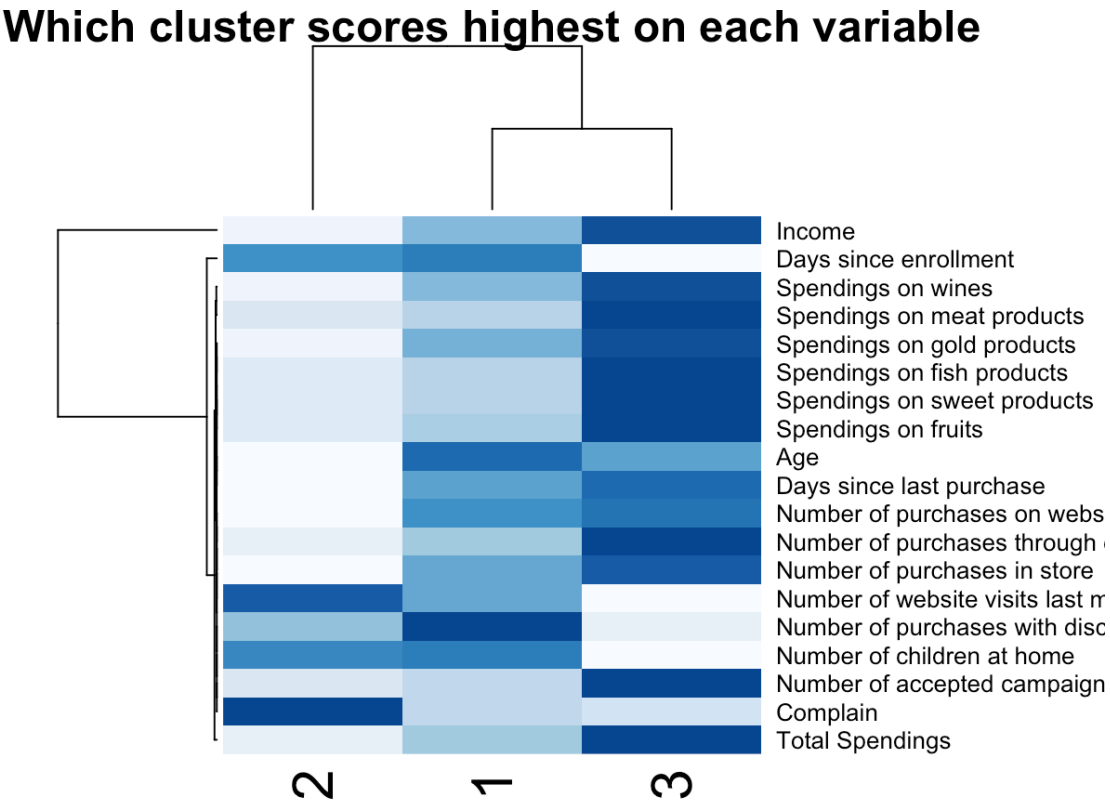

Figure 5. Variance of amount spent on grocery products between clusters

Figure 6. Variance of amount spent on luxury products between clusters

Figure 7. Promotion variance by cluster

Figure 8. Purchase place variance by cluster

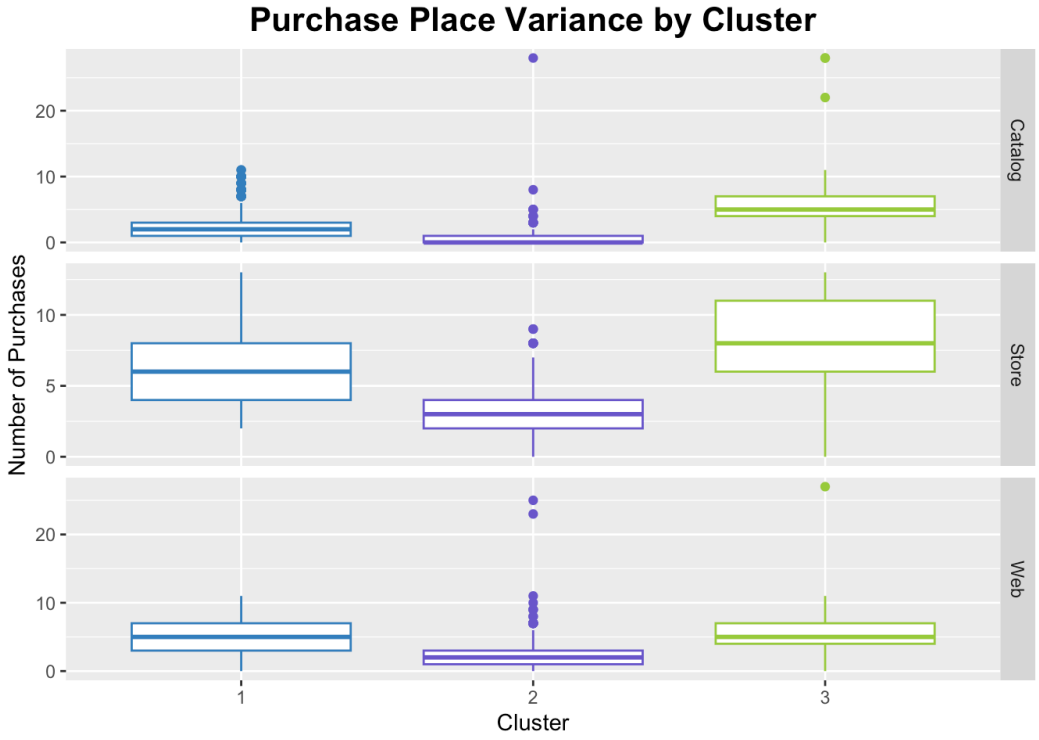

Figure 9. Education and marital status variance by cluster