Unveiling Customer Diversity

By Anna Fenner

Abstract

This project explores the segmentation of customers based on demographic and purchasing data using clustering algorithms. By applying data cleaning, exploratory data analysis (EDA), and feature engineering, we identified key customer segments. Using K-Means and DBSCAN clustering methods, we evaluated the results to enable targeted marketing strategies. The study revealed distinct clusters with unique characteristics, which can inform personalized marketing efforts.

Why

In today's competitive market, understanding customer behavior is crucial for businesses to tailor their marketing strategies. Customer segmentation allows companies to group customers based on similar characteristics, leading to more personalized and effective marketing campaigns. This project aims to leverage clustering algorithms to segment customers based on their demographic and spending data.

Data

I leveraged a Kaggle dataset encompassing diverse customer attributes, including marital status, number of children at home, income, education, and expenditure across various product categories over a span of two years. These categories include wine, fruits, meat products, fish products, sweet products, and gold.

Data Cleaning & EDA

1. Examined feature distributions.
2. Removed outliers where income was over 600,000 per year.
3. Utilized heatmap for feature correlation analysis
4. Explored different relationships in data a. Plotted boxplots and barplots using the two features: education and income.

Pre-processing

1. Developed a cumulative spending column for each customer over the past two years. a. Utilized scatterplots to investigate the relationship between income and total spending, indicating a potential predictive trend. b. Employed boxplots to examine the distribution of total spending across different education levels. c. Utilized KDE plots to visualize the distribution of total spending for each customer.
2. Established an age column for each customer. a. Filtered out entries with ages exceeding 110 years.

Models & Evaluation

Employed 2 different models: k-means and DBSCAN clustering algorithms. Evaluated models through elbow method for kmeans to determine the number of clusters and Silhouette method to see how tight the clusters are.

Results

After meticulous data cleaning and exploratory analysis, I applied clustering algorithms to identify three distinct customer groups within the company's base. Subsequently, I delved deeper into each group's characteristics to gain valuable insights into the company's customer landscape.

Notably, wine emerges as the primary category in terms of total sales across all three customer groups. Specifically, wine sales alone account for a significant portion of the total sales, representing 73.4% ($2,006,147) of the overall sales ($2,732,010). Moreover, Cluster 1 exhibits the highest proportion of wine expenditure, comprising 61.4% of the total wine expenditure among the three groups.

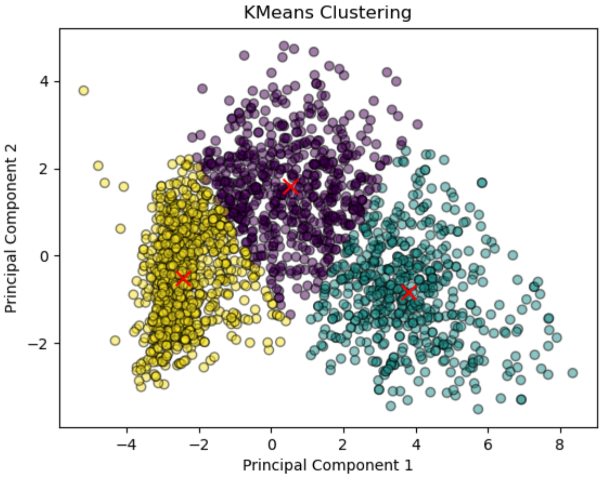

Figure 1: This plot shows the three distinct groups of this company's customer base.

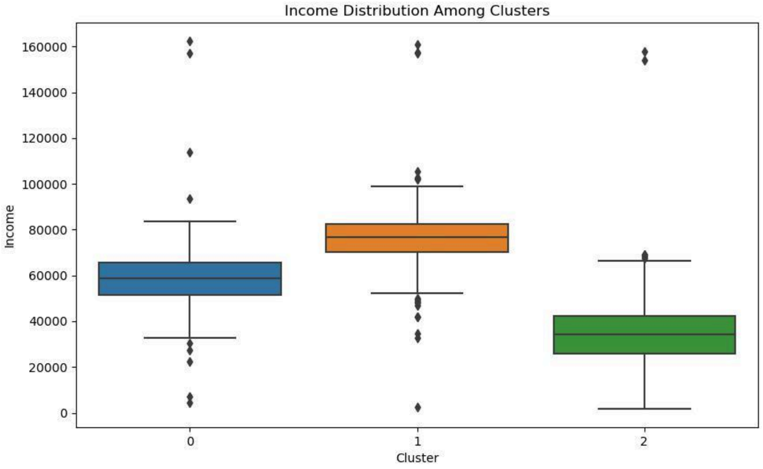

Figure 2: This boxplot shows the distribution of income for each cluster group. Cluster 1 has more customers with higher income. Cluster 2 has customers with the lowest income.

Cluster	Wine	Fruits	Meat	Fish	Sweet Products	Gold Products
0	604,888	13,096	81,893	17,374	13,548	34,837
1	1,231,118	39,244	260,993	57,012	40,567	44,294
2	170,141	5,726	25,859	8,474	5,645	17,541

Table 1: This is the breakdown of the sales of each category for each cluster group e.g. Wine means amount of wine spent in dollars.

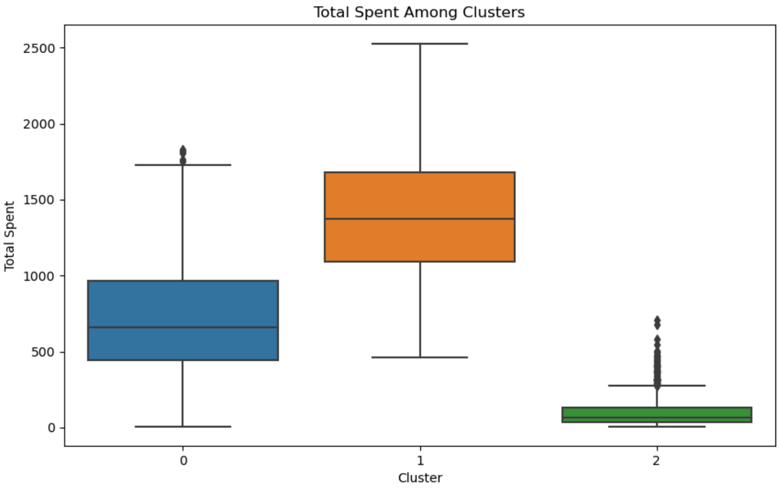

Figure 3: This boxplot reveals the distribution of total spent of each customer per cluster group. Cluster 1 group per customer spends the most.

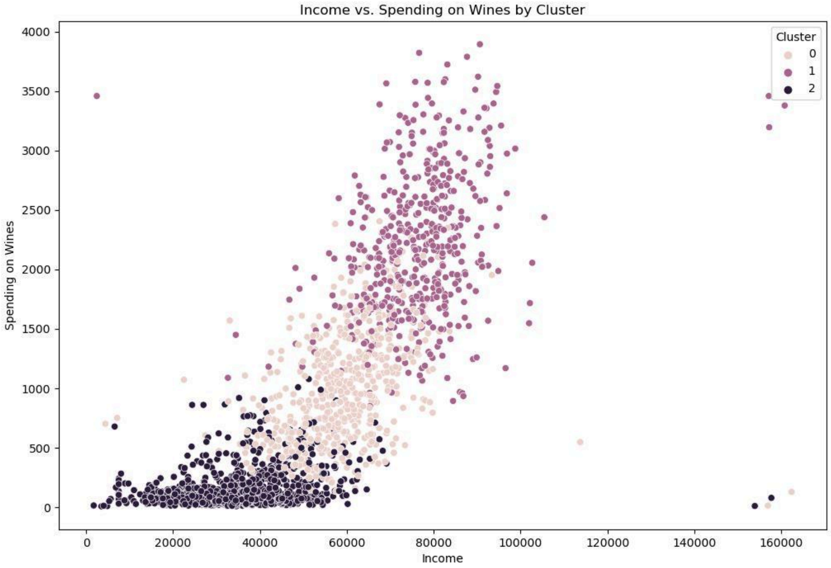

Figure 4: This scatterplot breaks down the data to each customer. Each dot represent a customer and how much they spent on wine alone and their given income. The different colors represent the different clusters. You can see how the cluster 2 group spends the least on wine.

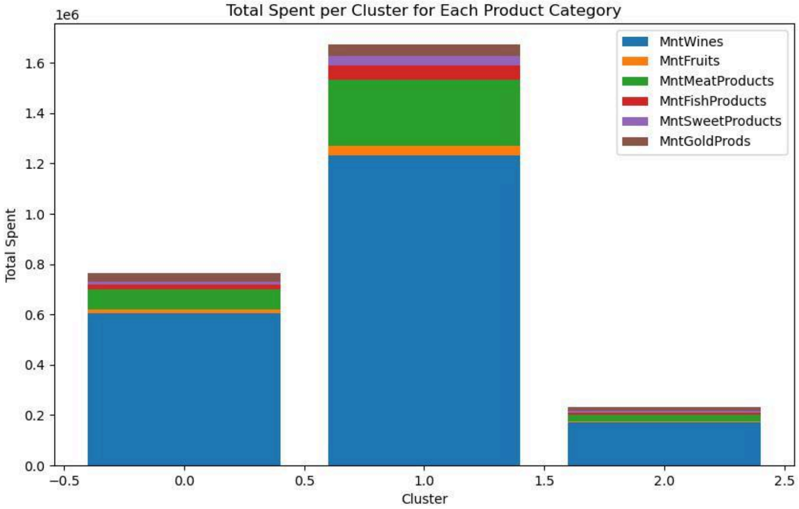

Figure 5: This breaks down the total spending for each cluster on the total for the different sub-categories.

Conclusion

This project employed the k-means clustering algorithm to segment customers based on demographic and spending attributes. The k-means model unveiled three distinct customer groups, demonstrating outstanding performance with a silhouette score of 0.3614 and a Calinski-Harabasz Index of 2845.15. Leveraging these insights, tailored marketing strategies can be devised to meet the diverse needs of our customer base.

Upon analysis, it's evident that Cluster 1 exhibits the highest expenditure, predominantly on wine. Interestingly, while the median income of Cluster 1 is approximately $80K, Cluster 0's median income stands at around $60K. However, Cluster 0's median total expenditure is merely half that of Cluster 1, indicating untapped potential in Cluster 0 sales.

Further exploration could involve dissecting the price per bottle expenditure within each group. This approach could provide valuable insights into devising targeted marketing approaches based on different price points for wine within each customer group. Further exploration and analysis would be to also look at the different items bought together per cluster group to see if there are any trends.