Unveiling Customer Diversity

By Anna Fenner

Why:

Utilizing clustering algorithms to group customers is vital for companies aiming to comprehend their customer base thoroughly. This strategic approach facilitates precision-targeted marketing efforts and unveils growth opportunities within each customer segment.

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.

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.

Tools: k-means, DBSCAN, PCA, matplotlib, seaborn

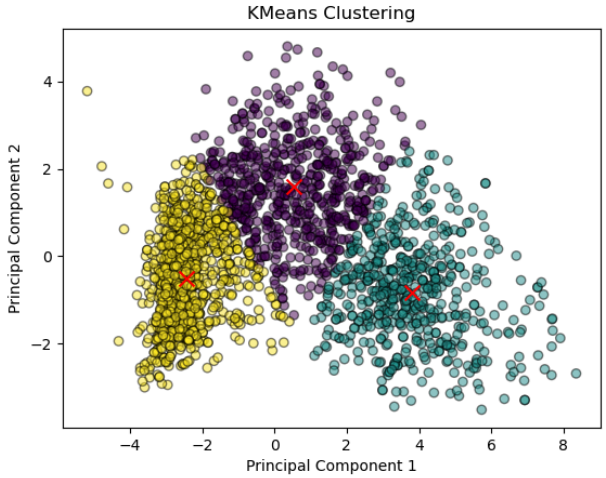

Figure 1: This plot shows the three distinct groups of this company's customer base.

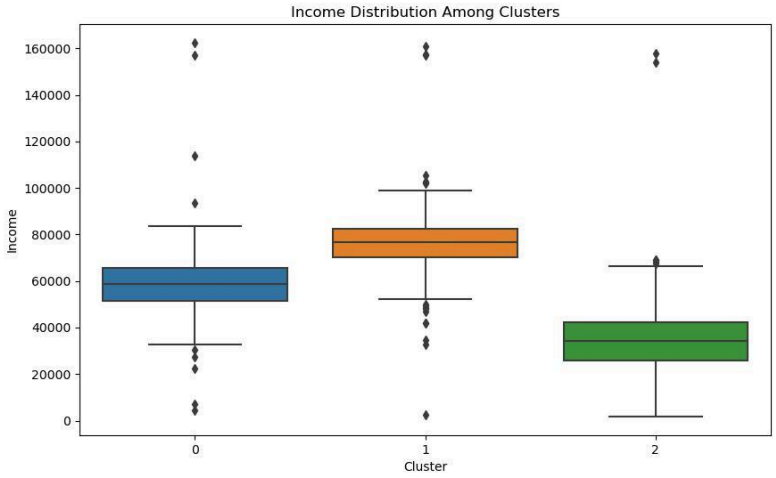

Figure 2: This boxplot shows the distribution of income for each cluster group. Cluster 1 has more customers with higher income. Cluster 2 has the more customers with the least income.

Cluster	MntWine	MntFruits	MntMeat	MntFish	MntSweet Products	MntGoldP rods
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.

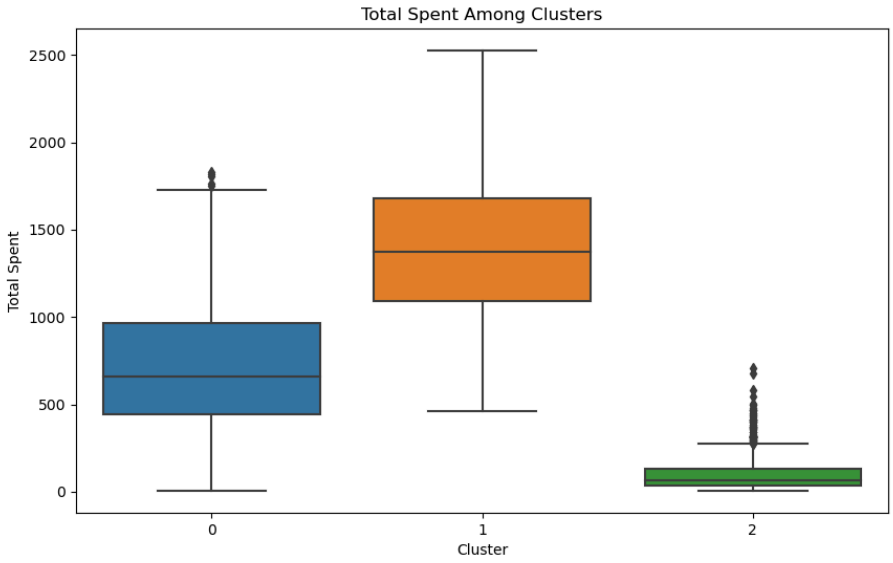

Figure 3: This boxplot reveals the distribution of total spent of each customer per cluster group. Cluster 1 group per customer spends the most.

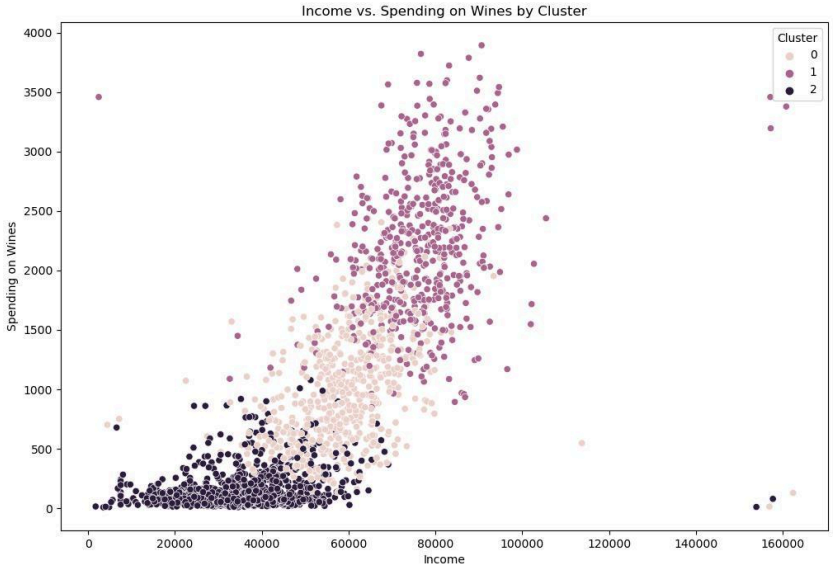

Figure 4: This scatterplot breaks down the data to each customer. Each dot represent a customer and how much they spent on wine alone and their given income. The different colors represent the different clusters. You can see how the cluster 2 group spends the least on wine.

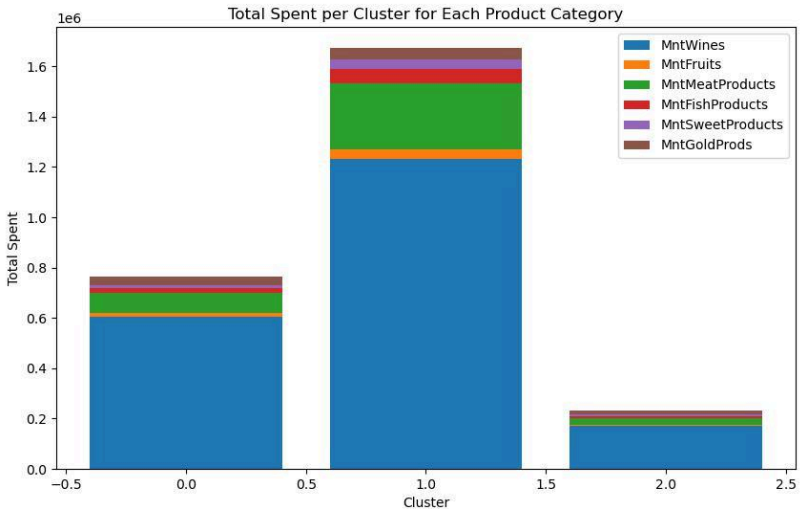

Figure 5: This breaks down the total spending for each cluster on the total for the different sub-categories.

Summary:

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 is any trends.