Retail Customer Segmentation

Filippo Nardi
Laboratory of Customer and Business Analysis - Data Science, University of Trento
https://github.com/filipponardi17/Retail-Customer-Segmentation-2

I. Introduction

In today's competitive business environment, understanding and effectively targeting customer segments is crucial for maximizing the impact of marketing strategies and improving customer satisfaction. Customer segmentation, the process of dividing a customer base into distinct groups based on specific characteristics, allows businesses to tailor their products, services, and marketing efforts to meet the unique needs of each segment.

This analysis will involve several steps: data preprocessing, selection of relevant features, application of various techniques to gain insights, utilization of clustering algorithms, and visualization of the resulting clusters. The results will be interpreted to understand the characteristics of each segment and their implications for business decision-making. Through this empirical analysis, this report will demonstrate how data-driven customer segmentation can be a powerful tool for businesses aiming to optimize their market positioning and achieve better alignment with customer needs.

The data we are using is the Customer Personality Analysis dataset [1], it comprises a wide range of features, including demographic information (e.g., birth year, education, marital status, income), purchasing behaviors (e.g., spending on wines, fruits, meat products), and product preferences. It also includes behavioral indicators such as the number of purchases made through various channels (e.g., web, catalog, store) and interactions with marketing campaigns. By leveraging these features, the dataset enables the identification of distinct customer clusters with shared characteristics, providing businesses with actionable insights to inform targeted marketing strategies, enhance customer relationship management, and drive revenue growth. This comprehensive dataset facilitates a deeper understanding of customer profiles, allowing for the optimization of product offerings and marketing efforts to meet the specific needs of different customer segments.

II. Data Cleaning & Preprocessing

In this section, we detail the processes undertaken to prepare the dataset for analysis. Ensuring the data is clean and properly preprocessed is crucial for reliable results, especially in tasks like clustering and dimensionality reduction.

The dataset initially contained 2,240 entries across 26 features, including demographic information, customer purchase behavior, and responses to various marketing campaigns. An initial inspection revealed that some features had missing values, specifically within the Income column, which had 24 missing entries. Additionally, certain categorical features contained outliers or erroneous data entries that required attention.

To address the issue of missing values, rows containing these omissions were removed. This step reduced the dataset size, ensuring that the dataset was complete and would not introduce biases or distortions in the subsequent analyses.

The categorical features, such as *Education* and *Marital_Status*, were analyzed to understand their distributions. The *Education* feature had several levels, including categories like Graduation, PhD, Master, and 2nd Cycle. The level Basic was also initially present but was removed as it was considered ambiguous. The *Marital_Status* feature contained common levels such as Married, Together, Single, and Divorced, but also included irregular levels such as Absurd and YOLO. These atypical entries were identified as outliers and were removed to maintain the consistency and reliability of the data.

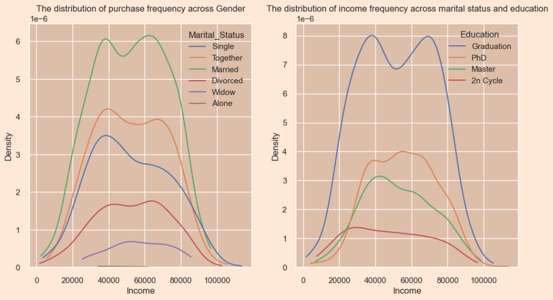

Fig. 1

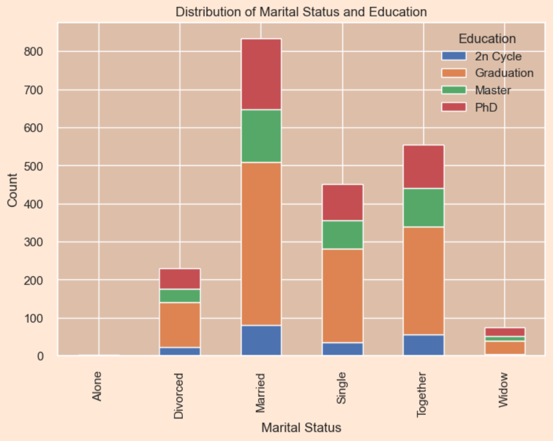

Fig. 2

Descriptive statistics were computed for the numerical features to gain insights into their distributions. For instance, the *Income* feature had a mean value of approximately $52,247, with a significant standard deviation, indicating a wide range of incomes among the customers. The *Year_Birth* feature showed an average year of birth around 1969, with the youngest customers born in 1996 and the oldest in 1893. This analysis helped identify potential outliers and provided a clearer understanding of the data's overall structure.

Outliers were particularly evident in the *Year_Birth* and *Income* features, as shown in Fig. 3. For example, some customers were born before 1940, and a few had extraordinarily high incomes that could skew the analysis. These outliers were visualized using box plots, which highlighted the extreme values. As a result, data points with birth years before 1940 and incomes exceeding $150,000 were removed, refining the dataset for more accurate and reliable analysis.

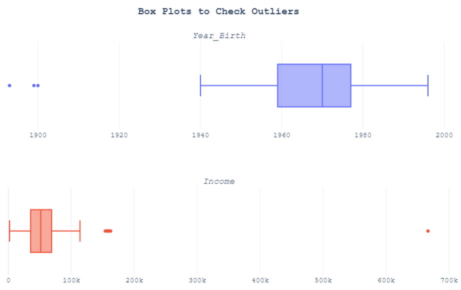

Fig. 3

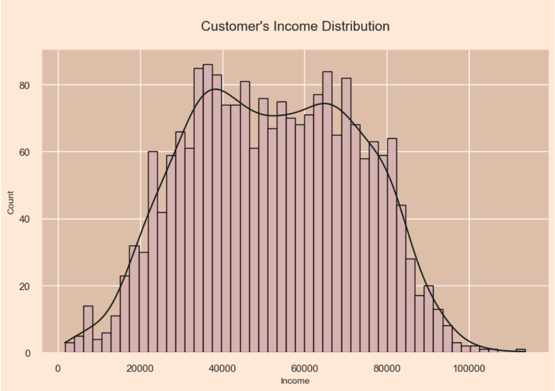

Fig. 4

After cleaning the data, categorical features were converted into numerical formats using label encoding, making them suitable for machine learning algorithms. The dataset was then standardized, ensuring that all features had a mean of zero and a standard deviation of one. This scaling was crucial for clustering algorithms and dimensionality reduction techniques that are sensitive to the magnitude of the features.

III. Feature Selection

Feature selection is a critical step in the data preprocessing pipeline, particularly when preparing data for machine learning models. The goal of feature selection is to identify the most relevant features that contribute to the predictive power of the model, while reducing dimensionality, with the goal of enhancing model performance and interpretability. The dataset initially contained 26 features, including demographic information, customer behavior metrics, and response variables related to marketing campaigns. These features varied in their relevance and contribution to the task of clustering customers for segmentation. To reduce the features set, we implied an algorithm of feature selection based on K-Means clustering: the variance of each feature within each cluster was calculated. Features that showed significant variance across clusters were considered important for distinguishing between customer groups. Features like *Income*, *MntMeatProducts*, and *NumCatalogPurchases* showed high importance, indicating that they are key factors in differentiating customer segments. Based on the analysis, the following features were selected for the clustering model:

- ***Income***: A critical determinant of customer purchasing power, Income was one of the most informative features.
- ***MntWines***: This feature, representing the amount spent on wine, was highly indicative of a customer's spending habits, particularly in the premium segment.
- ***NumCatalogPurchases***: The frequency of catalog purchases provided insights into customer engagement with direct marketing efforts.
- ***NumStorePurchases***: This feature was important for understanding the shopping behavior of customers, particularly their preference for in-store shopping versus online or catalog purchases.
- ***Kidhome***: This feature, indicating the number of children at home, was retained as it provides context on customer lifestyle, which could influence purchasing decisions.
- ***Year_Birth***: Age is a fundamental demographic variable that influences many aspects of consumer behavior.

We reduced our initial dataset to these 6 features which showed the most significant importance.

An Exploratory Factor Analysis (EFA) was conducted using the first three factors, selected based on their eigenvalues exceeding 1. The analysis was performed on the reduced dataset, resulting in the following chart

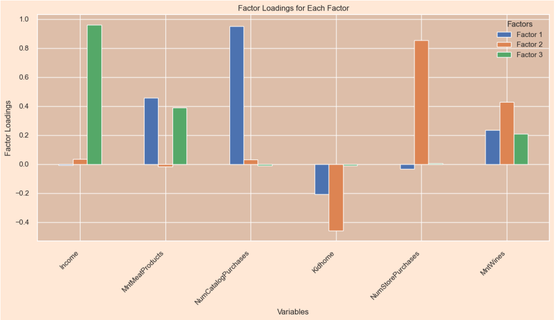

Fig. 5

In Fig. 5 we can see that Factor 1 is strongly associated with *NumCatalogPurchases* and *MntWines*, indicating a focus on catalog and wine purchases. Factor 2 is linked to *NumStorePurchases* and inversely related to *Kidhome*, suggesting a segment based on in-store shopping behavior and family size. Factor 3 correlates with *Income* and *MntMeatProducts*, highlighting a relationship between income and meat product expenditures.

IV. Dimensionality Reduction

Dimensionality reduction is a crucial step in the data preprocessing pipeline, particularly when dealing with high-dimensional datasets. The primary goal of this step is to reduce the number of features while retaining the most important information, simplifying the model, improving computational efficiency, and enhancing visualization without sacrificing significant accuracy.

Three main techniques were employed for dimensionality reduction in this analysis: Principal Component Analysis (**PCA**), t-distributed Stochastic Neighbor Embedding (**t-SNE**) and Uniform Manifold Approximation and Projection (**UMAP**) were used.

PCA is a linear dimensionality reduction technique that projects the data onto a set of orthogonal axes (principal components) that capture the maximum variance in the data. The steps involved in PCA include:

- **Explained Variance**: Initially, PCA was applied to determine how many principal components were necessary to capture a significant portion of the data's variance. The analysis revealed that a small number of principal components could explain most of the variance, indicating that the dataset could be effectively reduced to a lower-dimensional space.
- **Scree Plot**: A scree plot was generated to visualize the explained variance ratio of each principal component. The plot showed that the first 3 components captured a substantial amount of variance, with diminishing returns for each subsequent component. This informed the decision on how many components to retain.

t-SNE is a non-linear dimensionality reduction technique particularly suited for high-dimensional data where complex relationships exist between features.

- **Local Structure Preservation**: t-SNE focuses on preserving the local structure of the data, making it ideal for visualizing clusters that might not be linear in nature. It works by minimizing the divergence between two distributions: one representing pairwise similarities in the original high-dimensional space and the other in the lower-dimensional space.

UMAP is another non-linear dimensionality reduction technique that is similar to t-SNE but often more efficient and scalable.

- **Neighborhood Preservation**:UMAP, which focuses on preserving both the local and global structure of the data, was applied not to validate but to provide an alternative perspective to the dimension reduction techniques such as PCA and t-SNE.

V. Clustering Techniques

K-Means is one of the most widely used clustering algorithms in unsupervised machine learning. Its primary goal is to partition a dataset into K distinct clusters, where each data point belongs to the cluster with the nearest mean. The algorithm works iteratively to assign data points to clusters in a way that minimizes the variance within each cluster. K-Means is particularly effective for partitioning datasets where clusters are spherical and of roughly equal size. In our project, we performed clustering on all three datasets using $k = 3, 4$, and 5. We selected $k = 4$ as it demonstrated the highest performance according to our silhouette score plot (see Fig. 6).

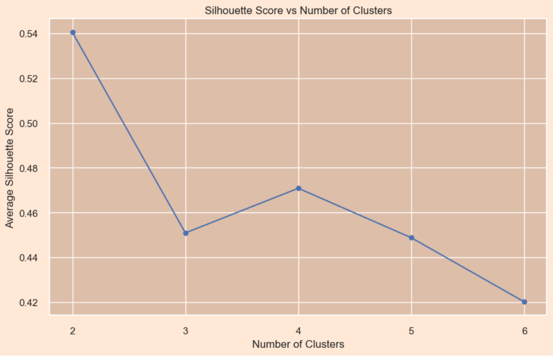

Fig. 6

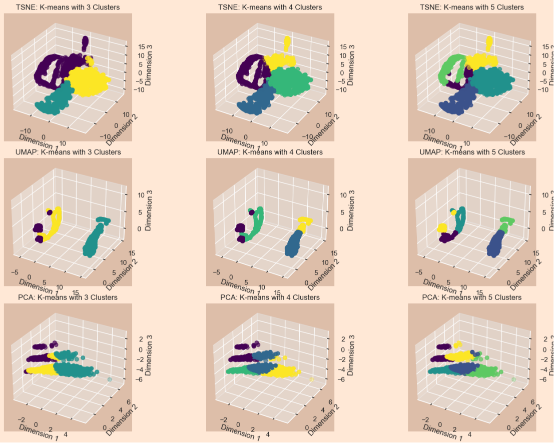

Fig. 7

As can also be seen in Fig. 7, it makes sense to proceed with clustering for all three datasets generated by the dimensionality reduction algorithms. The results show a clear separation between clusters, regardless of the dimensionality reduction method used, suggesting that each of these approaches provides valuable and complementary insights

VI. Empirical Results

To quantitatively evaluate the clustering results, we examined the mean values of key features within each cluster. This was done by aggregating the data within each cluster and calculating the average for the variables. This analysis helped us understand the characteristics of each cluster and compare them across the different dimensionality reduction methods.

The tables generated for each clustering approach (PCA, t-SNE, UMAP) provided insights into how well-separated and distinct the clusters were. For example:

- **PCA-Based Clustering**: The table for PCA clusters showed clear differences in average income and spending patterns between clusters. One cluster might represent high-income, high-spending customers, while another might indicate lower-income customers with specific purchasing preferences.

- **t-SNE-Based Clustering**: The clusters from t-SNE displayed similar patterns but often with tighter and more distinct averages, suggesting that t-SNE was effective at capturing more nuanced differences between customer groups.
- **UMAP-Based Clustering**: The UMAP-based clusters provided a balance between local and global structure, with mean values that were often consistent with those from t-SNE but showed slightly different emphasis on certain features, indicating subtle variations in customer behavior.

These tables allowed us to verify that the clusters made logical sense and corresponded to differences in customer characteristics.

In addition to the quantitative tables, visualizations were employed to assess the distribution of key features within each cluster. In particular, box plots were used to assess the differences between the clusters and variables. These plots helped visually confirm that the clusters were distinct and that the model was effectively segmenting the customers.

Performing a Biplot using PCA also give us this chart showing distinct customer segments based on purchasing behavior, with *Income*, *NumCatalogPurchases* and *NumMeatProducts* contributing strongly to Dimension 1, indicating a segment focused on direct marketing shopping, especially those with higher income levels and a preference for specialized product categories such as meat. The arrows indicate that these variables are positively correlated, while Dimension 2 captures variation related to *Kidhome* and *NumStorePurchases*, suggesting a segment influenced by the presence of children and in-store purchases.

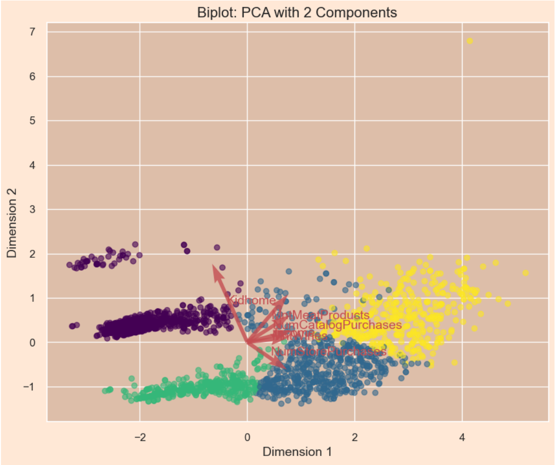

Fig. 8

VII. Clusters Description

In our case study, we believe that the dimensionality reduction method that performed the best in terms of separation and segmentation was PCA as can be seen in Fig. 6, and it produced the following insights:

- **Cluster 0 – Lower Income Segment**: The lower income of Cluster 0 corresponds to their lower spending across all product categories. Their higher number of children in the household could indicate financial constraints, leading to more conservative spending. This cluster might respond well to discounts, offers on essential goods, and value-oriented marketing. Since they have fewer purchases, it may also be beneficial to explore what barriers might exist for this group, such as accessibility or awareness of certain products.
- **Cluster 1 – Family-Oriented Segment**: Cluster 1, while having a slightly lower income than Cluster 3, shows a higher number of children in the household. This suggests that their spending might be more focused on family needs, and they might prioritize products and services that cater to households with children. Marketing strategies aimed at this cluster could emphasize value for money, family bundles, and promotions that appeal to a broader family-oriented lifestyle.
- **Cluster 2 – Middle Ground**: Cluster 2 represents a middle ground in terms of income and consumption patterns. They have fewer children, which could mean more disposable income per capita in the household, yet their spending remains moderate across categories. This cluster might be a good target for upselling strategies, where the goal is to increase their spending on premium products or encourage them to explore more product categories.
- **Cluster 3 – High Spenders**:Cluster 3 has the highest income, which is reflected in their spending patterns. This group has a strong preference for high-value items such as wine and meat products, and they are active in making purchases through various channels, including catalog and store purchases. This cluster may represent a target for premium products and loyalty programs, given their spending capacity and engagement across multiple purchasing platforms.

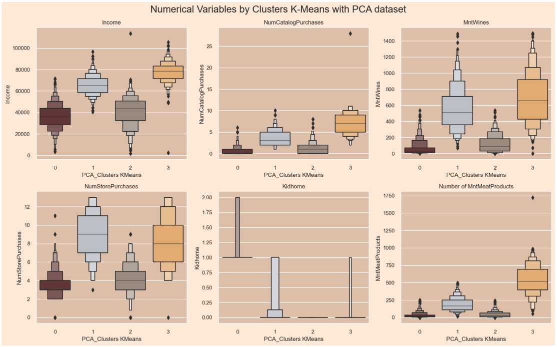

Fig. 9

Additionally we produced the visualitation for the other two dimensionality reduction datasets. The results are shown in Fig. 9 and Fig. 10.

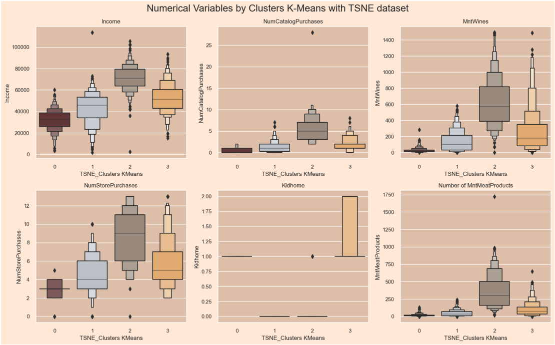

Fig. 10

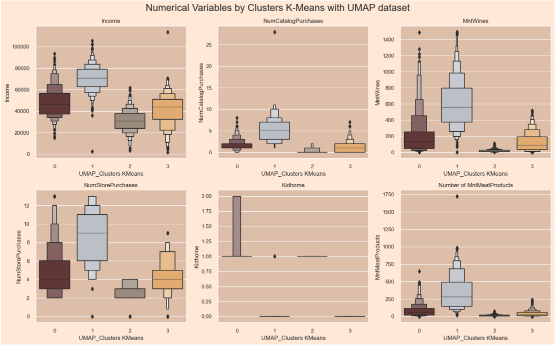

Fig. 11

VIII. Conclusion

Based on the findings presented in the analysis, it is evident that customer segmentation is a powerful tool that can significantly enhance marketing strategies and business decision-making. By employing various clustering techniques and dimensionality reduction methods such as PCA, t-SNE, and UMAP, we were able to identify distinct customer segments that exhibit unique behaviors and preferences. Each segment, from the lower-income, family-oriented clusters to the high-income, high-spending groups, presents opportunities for targeted marketing and personalized engagement.

The project highlights the importance of understanding customer demographics, purchasing behaviors, and responses to marketing campaigns in creating effective segmentation strategies. For instance, the identification of a high-spending cluster that prefers premium products suggests that businesses can focus on luxury goods and loyalty programs to maximize revenue from this segment. Conversely, the lower-income segment may benefit from value-oriented offerings and promotions.

Furthermore, the comparison of clustering results across different dimensionality reduction techniques demonstrates the value of using multiple approaches to gain comprehensive insights. While PCA provided clear segmentation, t-SNE and UMAP offered additional perspectives on the nuances within the data, reinforcing the robustness of the clustering outcomes.

References

[1] Vishakh Dapat. *Customer Segmentation Clustering - Dataset*. 2023. URL: https://www.kaggle.com/datasets/vishakhdapat/customer-segmentation-clustering (visited on 08/31/2024).