Customer segmentation with clustering

Fasone Alessandro

22/06/2025

Table of content

Problem statement

The company requires a robust customer segmentation model to refine its marketing strategies. Absent or inaccurate segmentation poses significant risks, including operational inefficiencies, wasted resources, and a poor understanding of customer needs, thereby hindering customer retention.

This project aims to define customer clusters, providing deep insights for personalizing offers and optimizing promotional campaigns. The objective is to enhance customer satisfaction and drive sustainable business growth through a truly customer-centric approach.

Methods

The adopted approach for customer segmentation is structured into several sequential phases, aiming to delineate each customer's assignment to specific behavioral segments.

Initially, an extensive feature engineering phase was conducted. The original dataset underwent preprocessing and was augmented with five new features, specifically engineered to capture various dimensions of customer behavior. These new features were aggregated at the individual customer level, yielding a final dataset comprising 64,885 unique customers. The new features and their aggregation methodologies are detailed in Table 1.

Table 1: Market segmentation features

Feature	Description	Aggregation
Frequency	Total count of different order made by each client	Maximum
Recency	Days since last order for each client	Minimum
Customer Lifetime Value (CLV)	Total value for the business of each client	Sum
Average Unit Cost	Average profitability from each order made by each client	Mean
Age	Age of the customer	First

Following a preliminary exploratory data analysis (EDA) and dataset cleaning, the optimal number of clusters (k) was determined. To achieve this, several complementary approaches were employed: the Elbow Method, the Silhouette Score, and dendrogram analysis.

Once the optimal number of clusters was identified, k-means clustering was performed and customers were assigned to their respective clusters, and the distinctive characteristics of each group were analyzed in depth.

Finally, to visualize and interpret the cluster structure in lower-dimensional spaces, dimensionality reduction techniques such as PCA and t-SNE were applied.

Results

Upon finalizing the dataset, comprehensive analyses were conducted to determine the optimal number of segments that best represent the company's customer base. The convergence of results from the Elbow Method (Figure 1), Silhouette Score (Table 2), and dendrogram analysis (performed on a representative subset of 25,000 customers) strongly indicated that a segmentation into five clusters (k=4) constitutes the optimal level. Specifically, for k=4, a pronounced reduction in the slope of the inertia curve was observed (Elbow Method), a Silhouette Score of 0.2415 (the third highest value, signifying good within-cluster cohesion and between-cluster separation), and a dendrogram cut (with ward linkage). These convergent indicators robustly support the selection of k=4 as the appropriate number of market segments for the company.

Figure 1: Elbow method

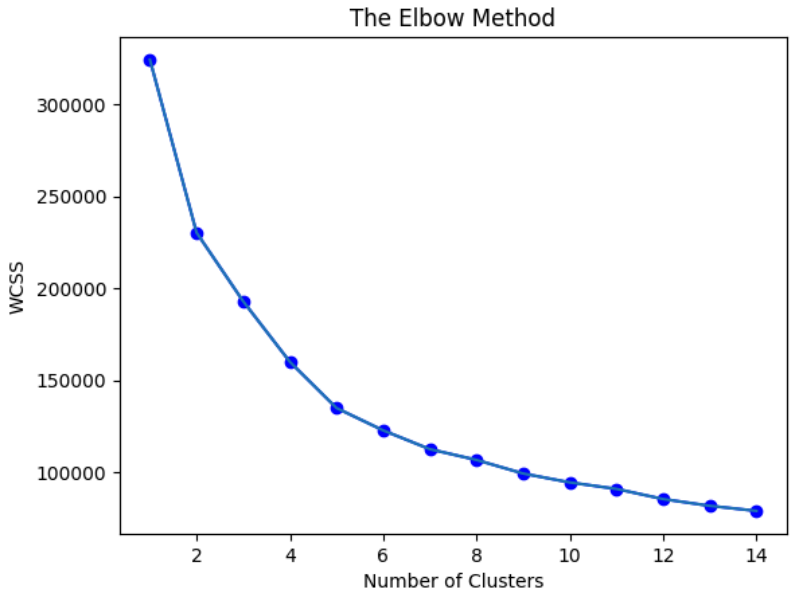


Table 2: Silhouette score

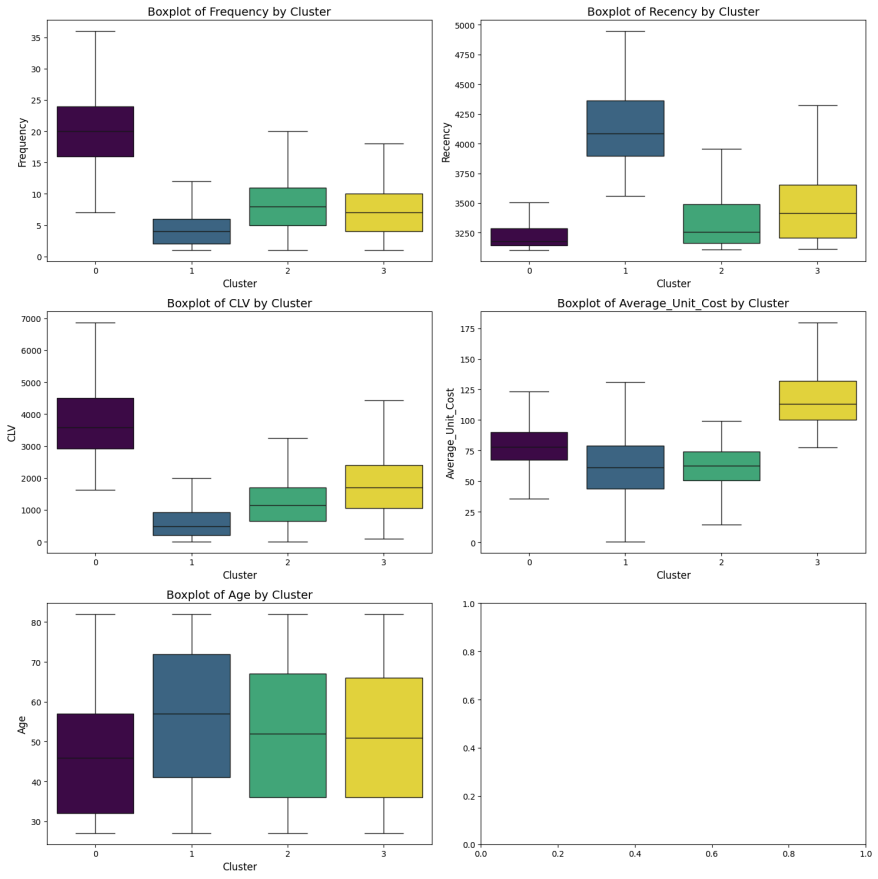
Number of clusters	Silhouette score
2	0.2580
3	0.2309
4	0.2415
5	0.2536

Number of clusters	Silhouette score
6	0.2344
7	0.2220
8	0.2168
9	0.2186

Once the segmentation was established using the K-Means clustering technique, the five distinct clusters were thoroughly analyzed, revealing the following characteristic profiles, with mean values of key metrics provided in parentheses. Illustrative boxplots showcasing the distributions of these features for each cluster are presented in Figure 2.

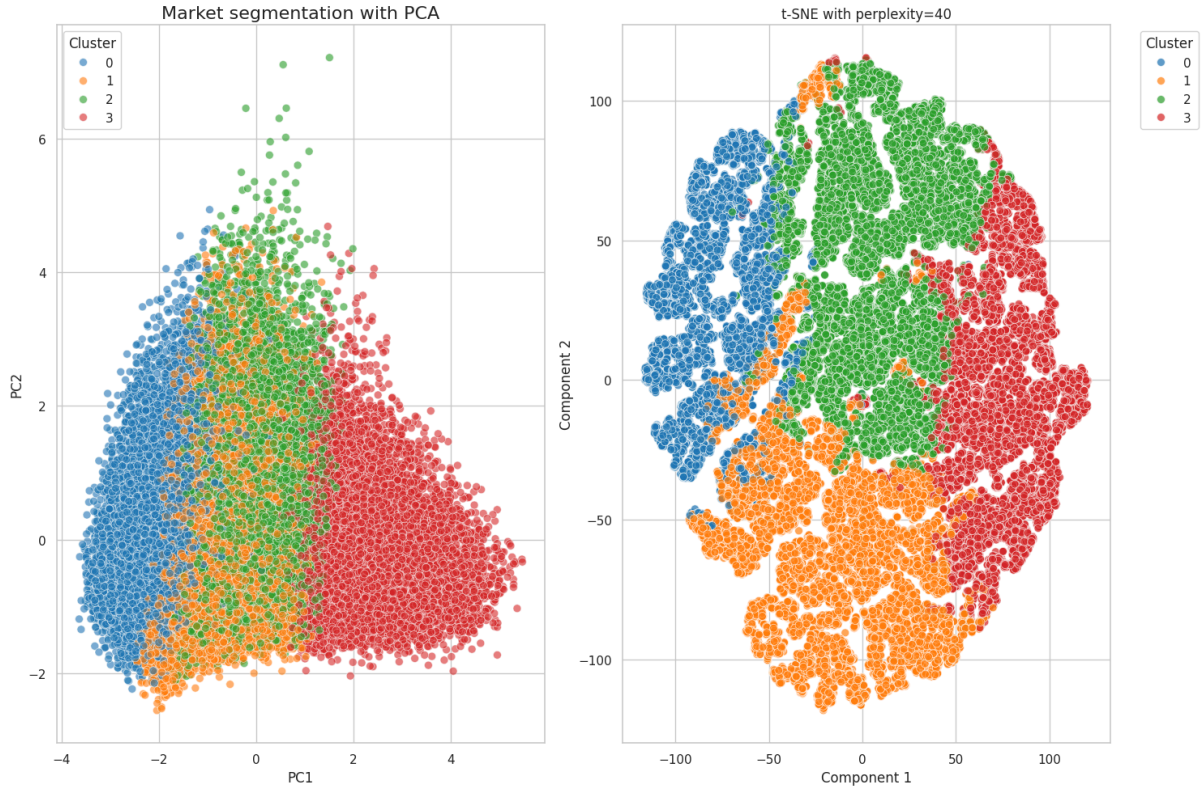
- Cluster 0: This segment represents the youngest demographic (mean age = 46.78 years) and the most active clients, exhibiting the highest purchase frequency (mean = 20.85) and the lowest recency (mean = 3255 days). Maintaining a high average unit cost (79.66$), they also yield the highest Customer Lifetime Value (CLV) (mean = 3806.64$).
- Cluster 1: Comprising the oldest customer population (mean age = 56.41 years), these are the least active customers, characterized by the highest recency (mean = 4184.87 days) and lowest frequency (mean = 4.39). Despite an average unit cost (mean = 62.18$) similar to other clusters, they exhibit the lowest Customer Lifetime Value (mean = 643.07$).
- Cluster 2: Customers in this cluster tend to make more frequent purchases, though with a lower average unit cost (mean = 61.78$). This results in a medium Customer Lifetime Value (mean = 1185.93$). Targeting these customers effectively requires focusing on lower-cost goods.
- Cluster 3: This cluster represents customers with the most expensive purchasing behaviors, evident from their highest Average Unit Cost (mean = 119.67$) and second-highest Customer Lifetime Value (mean = 1773.30$), alongside an average frequency (mean = 6.94). Identifying this segment allows the company to focus marketing campaigns on higher-value goods tailored to their preferences.

Figure 2: Cluster Exploratory Data Analysis



To graphically visualize the separation and distribution of the different groups, two dimensionality reduction techniques were applied, as illustrated in Figure 3. The representation via Principal Component Analysis (PCA) reveals that, although the clusters are conceptually distinct, they tend to visually overlap due to the inherent linear nature of PCA. In contrast, for t-Distributed Stochastic Neighbor Embedding (t-SNE), a perplexity of 40 was chosen, which yielded a more distinct and less ambiguous separation between the clusters. Consequently, in this context, t-SNE proved more effective in graphically representing the cluster division, by preserving local proximities and revealing nonlinear data structures, thereby offering a robust visual validation of the K-Means generated clusters.

Figure 3: Cluster visualisation



Conclusion

In conclusion, clustering analysis identified four distinct customer segments, each characterized by specific purchasing habits. Implementing this strategy is crucial for the company to adopt a truly customer-centric growth plan.

Specifically, Cluster 0 emerged as the most strategically relevant segment, representing the core clientele with the highest frequency, lowest recency, and highest CLV. Loyalty policies are essential to maximize their long-term value. Cluster 3 comprises customers with fewer but higher-value purchases, indicating a potentially more affluent clientele requiring tailored approaches. Conversely, Cluster 2 shows distinct behavior with higher purchase frequency but lower average unit costs, leading to a medium CLV, suggesting a focus on value-driven

offerings. The remaining Cluster 1, characterized by lower activity and CLV, necessitates targeted marketing efforts to boost engagement and value.

Overall, these segmentation-based strategies will optimize campaign effectiveness, allocate resources more efficiently, and foster more targeted and sustainable business growth.