Customer segmentation with clustering

Contents

Problem Statement

Understanding customer behaviour is critical for businesses aiming to improve marketing efficiency, customer retention, and profitability. This project segments customers for a global e-commerce company using clustering techniques. With nearly one million transactions across five continents, the aim is to group customers into meaningful segments based on behavioural and value-driven metrics. These segments will support more targeted marketing and better resource allocation. The key challenge is identifying relevant features, managing data complexity, and selecting the most effective clustering model to reveal actionable patterns.

Summary of approach

The analysis began with data cleaning and preparation. Two columns that were not relevant to the clustering task ("State_Province" and "Postal_Code") were removed. Missing values and duplicates were addressed, and several columns containing currency symbols or percentage signs were cleaned and converted to numeric format. Outliers were identified using the IQR method but retained to preserve genuine patterns in customer behaviour.

Feature engineering created five variables central to segmentation: Frequency, Recency, Customer Lifetime Value (CLV), Average Unit Cost, and Age. These features reflected both behavioural and value-based aspects of the customer base.

Feature	Description
Frequency	Number of orders placed by each customer (indicates engagement)
Recency	Days since last order (signals churn risk or loyalty)
Customer Lifetime Value (CLV)	Total revenue from each customer over the recorded period
Average Unit Cost	Mean price of items purchased (helps distinguish value vs. budget buyers)
Age	Age of the customer, based on birthdate

Table 1 Description of engineered features used for clustering

Three methods were used to explore the appropriate number of clusters: the Elbow method, Silhouette score, and Hierarchical clustering. These were applied to guide the selection of *k* before running the final clustering model. K-means clustering was then applied using the chosen value of *k*, and the results were explored using boxplots and 2D visualisations.

Key findings from clustering

The clustering process identified four distinct customer segments, each with clear behavioural and demographic differences. Figure 1 shows boxplots for the five engineered features across the clusters, clearly showing the distribution, central tendency, and presence of outliers for each group.

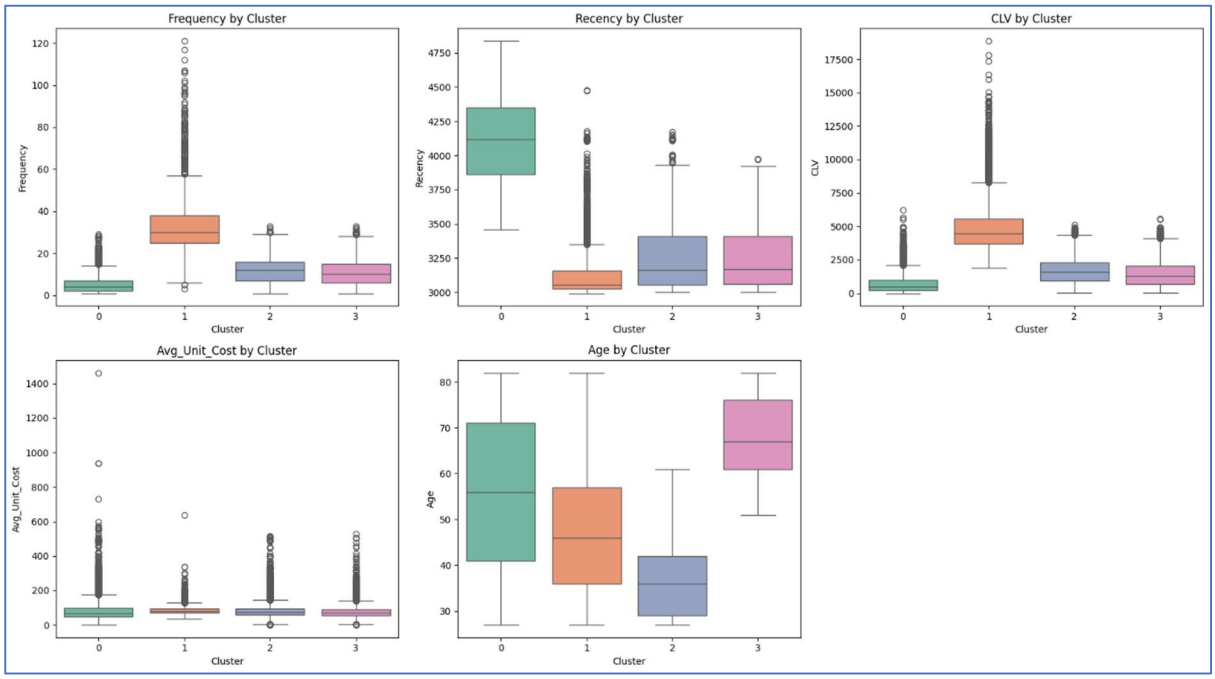

Figure 1 Boxplots of key features by cluster

The boxplots reveal that the clusters differ meaningfully across the engineered features. Cluster 1 stands out for its high frequency and CLV, with very recent purchase activity, suggesting a highly engaged and profitable segment. At the other end of the spectrum, Cluster 3 appears to be the most disengaged group. These customers show very low frequency and CLV, the longest time since last purchase (high recency), and are older on average — indicating a potentially lapsed or churned segment.

Clusters 0 and 2 fall between these extremes. Cluster 0 includes moderately active customers with average CLV and recent orders, making them a candidate for continued nurturing. Cluster 2 shows slightly lower engagement than Cluster 0, but still displays potential for reactivation through targeted marketing.

These findings are summarised in Table 1 below, which consolidates key patterns across the five segmentation features.

Cluster	Frequency	Recency	CLV	Avg Unit Cost	Age	Summary
0	Medium	Low	Medium	Medium	Lower	Active mid-value group
1	High	Very Low	High	Medium	Mixed	High-value, recent buyers
2	Low-Med	Medium	Low-Med	Medium	Lower	Low activity, re-engageable
3	Very Low	High	Low	Medium	High	Disengaged, likely churned

Table 2 Summary of customer clusters based on behavioural features

Model comparison and selection

Three methods were used to assess the optimal number of clusters. The Elbow method showed a noticeable inflection at $k = 4$, and the dendrogram from hierarchical clustering also supported $k = 4$ as a natural split. The silhouette score was highest at $k = 5$, but only marginally higher than $k = 4$, as shown in Table 3.

Given the minimal difference in silhouette scores and the support from the other two methods, $k = 4$ was selected as the final number of clusters. It balances simplicity with effective group separation and is supported by multiple independent techniques.

Number of cluster (k)	Silhouette score
2	0.2603
3	0.2527
4	0.2532
5	**0.2670**
6	0.2525

Table 3 Silhouette score for different values of k

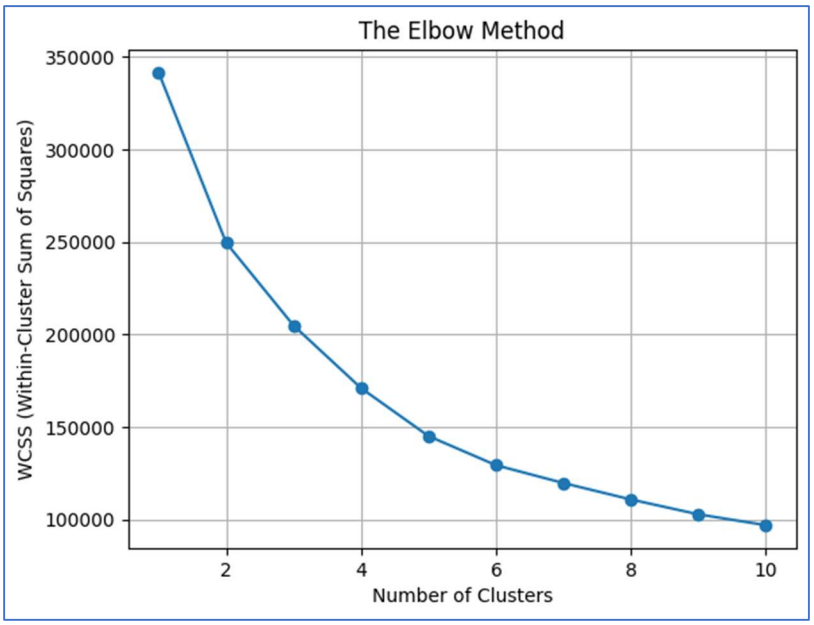

Figure 2 Elbow method WCSS curve

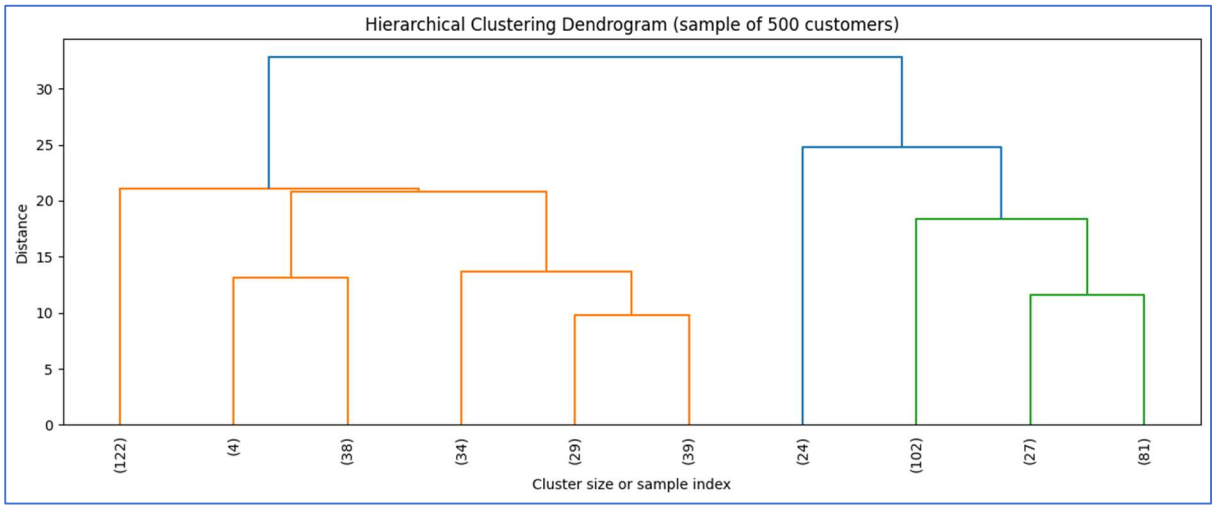

Figure 3 Dendrogram from hierarchical clustering

K-means was selected as the final clustering model for its simplicity, scalability, and consistent performance. The results were interpretable, repeatable, and aligned with insights from the other methods, making it the most effective choice for customer segmentation.

Visualisation effectiveness

To support interpretation of the clustering results, two dimensionality reduction techniques were applied: Principal Component Analysis (PCA) and t-distributed Stochastic Neighbour Embedding (t-

SNE). These projected the multi-feature dataset into two dimensions for visual inspection of cluster separation.

Figure 4 shows the PCA output. While one cluster (Cluster 1) was clearly separated, the remaining three overlapped significantly. This is expected, as PCA is a linear method and may struggle to preserve non-linear relationships in lower dimensions.

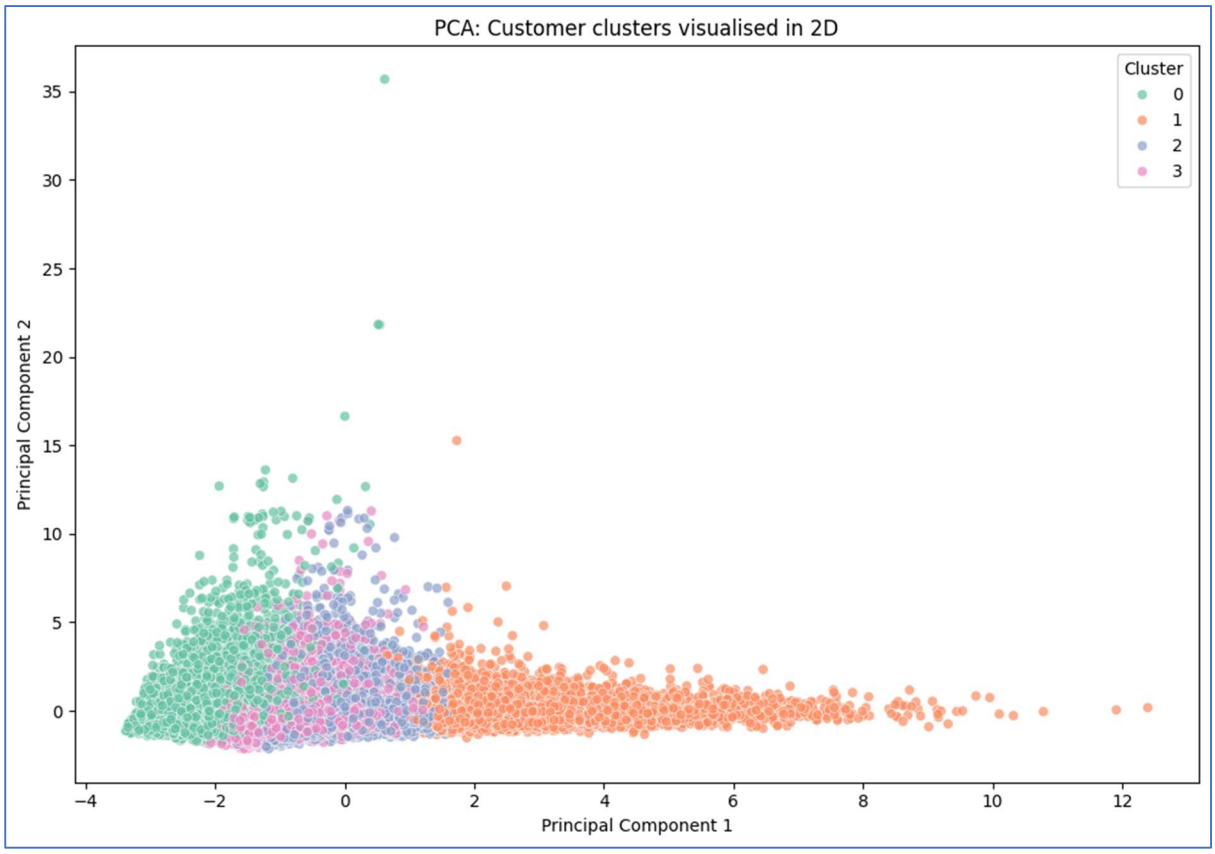

Figure 4 PCA visualisation of clusters in 2D

t-SNE, by contrast, produced a more informative result. As shown in Figure 5, all four clusters appeared visually distinct, with well-formed groupings and minimal overlap. This enabled a clearer interpretation of how customers were segmented by K-means.

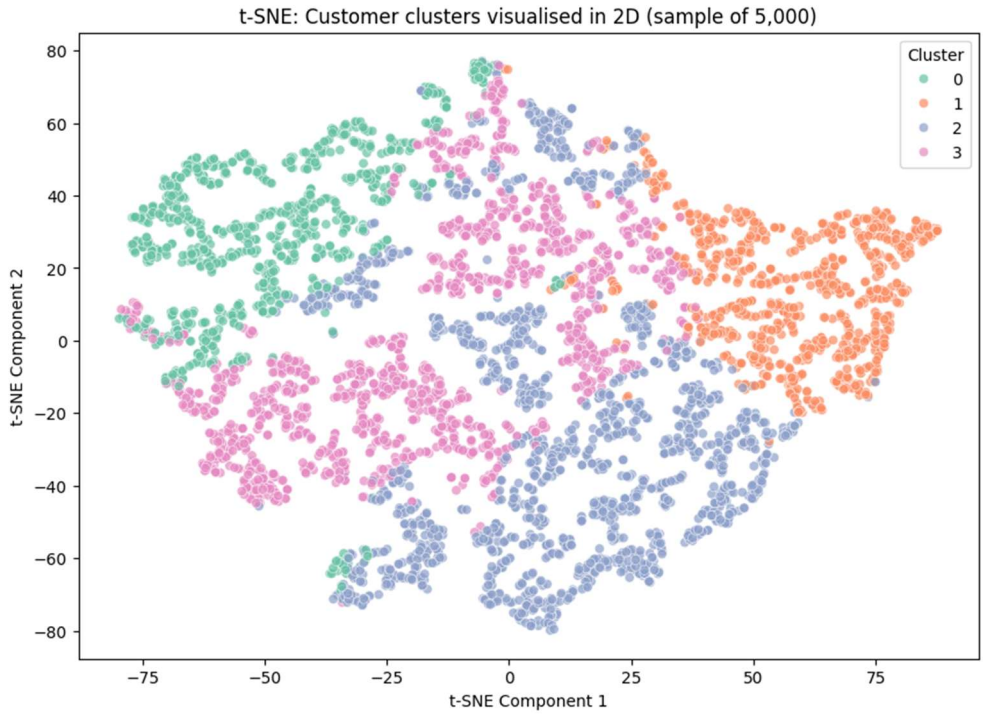

Figure 5 t-SNE visualisation showing separation between clusters

Overall, t-SNE proved more effective for visualising clustering performance. While PCA was helpful for initial exploration, t-SNE offered a clearer view of the underlying cluster structure.

Conclusion

This project successfully segmented a global e-commerce customer base into four meaningful groups using K-means clustering. The workflow combined data cleaning, feature engineering, statistical analysis, and machine learning to derive behavioural and value-based customer profiles.

The segmentation revealed clear differences between groups, from high-value, loyal customers to inactive or potentially churned individuals. These insights offer immediate value for marketing — for example, focusing retention on Cluster 1, nurturing Clusters 0 and 2, and deprioritising Cluster 3.

The modelling approach was guided by analytical techniques (Elbow method, Silhouette score, and hierarchical clustering), with K-means selected for its effectiveness and interpretability. Visualisations, particularly t-SNE, provided strong support for the chosen segmentation by clearly showing cluster separation.

This analysis equips the business with a customer-centric lens for targeting and decision-making, enabling more efficient resource allocation, stronger retention, and greater revenue potential.