Customer Segmentation:
Clustering & Feature Engineering

1 Introduction

Customer segmentation is a powerful strategic tool for understanding customer profiles. By grouping customers into segments, businesses can gain valuable insight into aspects such as geographical distribution, purchasing behaviour, and demographic trends. Such insights can drive optimisation in areas like marketing campaigns, product development, and pricing strategies.

This report utilises the k-means clustering algorithm to segment a large customer dataset into clusters based on patterns and similarities in their features. The segmentation process relies on five engineered features derived from combinations of variables in the original dataset, enabling more meaningful grouping and actionable insights.

2 Overview of Approach

2.1 Dataset overview

The dataset contains 951,669 entries with 20 features, each representing a purchase. Features include geographical data (continent, country, city), order details (quantity, order date, delivery time), customer data (DOB, type, loyalty membership), and monetary data (profit, revenue, unit cost). Missing data was limited to geographical features, further detailed in section 2.2.

2.2 Preprocessing

The state/province column had 12% missing values, primarily from Europe, likely due to differences in address requirements compared to North America which had no missing values in this column. These entries were retained, as the missing data was not critical. Additionally, 21 duplicate entries were removed.

The dataset was aggregated by customer to focus on segmentation by this metric, reducing its size and altering some feature structures (e.g., order dates became lists). To improve clustering performance, five new features were engineered (see section 2.3), and all features were standardised between -1 and 1.

2.3 Feature engineering

Five features were created to better segment the customer profiles:

- **Customer Age**: Derived from DOB for age-group segmentation.

- **Average Unit Cost**: The average company expense per customer.
- **Recency**: Time since last order, indicating retention.
- **Frequency**: Order frequency, reflecting loyalty.
- **Customer Lifetime Value**: Combines age, frequency, and revenue to estimate a customer's future value.

3 Description of Analysis

This section will give an overview and explanation of the steps invloved in the segmentation process.

3.1 Determining optimal k

The key challenge in k-means clustering is selecting an appropriate number of clusters (k). An unsuitable choice can lead to poorly segmented data, potentially missing critical insights. Two methods were used in this report to determine the optimal k:

Elbow method

This approach evaluates various k-values by calculating the **within-cluster sum of squares (WCSS)**, which decreases as k increases. The "elbow point" is identified where the rate of improvement in WCSS sharply diminishes, indicating the optimal k.

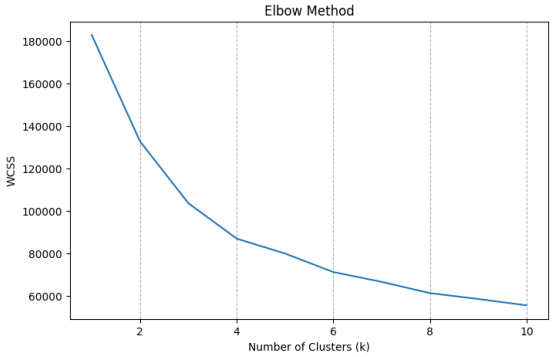

Figure1: Elbow method WCSS

Silhouette score

This method assesses cluster quality for different k-values by measuring how well each data point fits within its assigned cluster compared to others. A higher silhouette score suggests better-defined clusters.

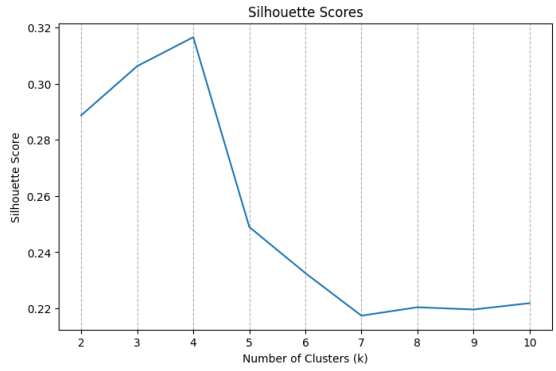

Figure2: Silhouette scores

Both methods identified $k = 4$ as the optimal number of clusters. Supporting analyses are shown in Figures 1 and 2.

3.2 Hierarchical clustering

To validate $k = 4$, **hierarchical clustering** was also applied. This method iteratively groups data points by merging smaller clusters or splitting larger ones based on a linkage criterion, producing a **dendrogram** (Figure 3) to visualize the cluster structure. Hierarchical clustering identified three clusters, differing from the k suggested previously. However, given the majority agreement between the elbow and silhouette methods, $k = 4$ was deemed appropriate, reflecting sufficient granularity for insightful segmentation.

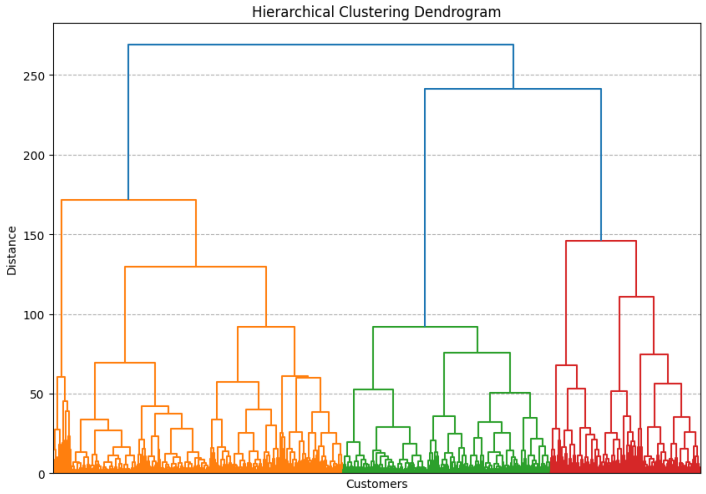

Figure3: Dendrogram of Clusters

3.3 Segmentation process with optimal k

The final segmentation used $k = 4$, supported by the majority consensus and consideration of the added granularity provided by four clusters. The k-means algorithm produced two larger clusters (approximately 25,000 entries each) and two smaller clusters (around 7,000 entries each). Class labels were added to the dataset for further analysis, with cluster distributions visualized in box plots (Figure 4).

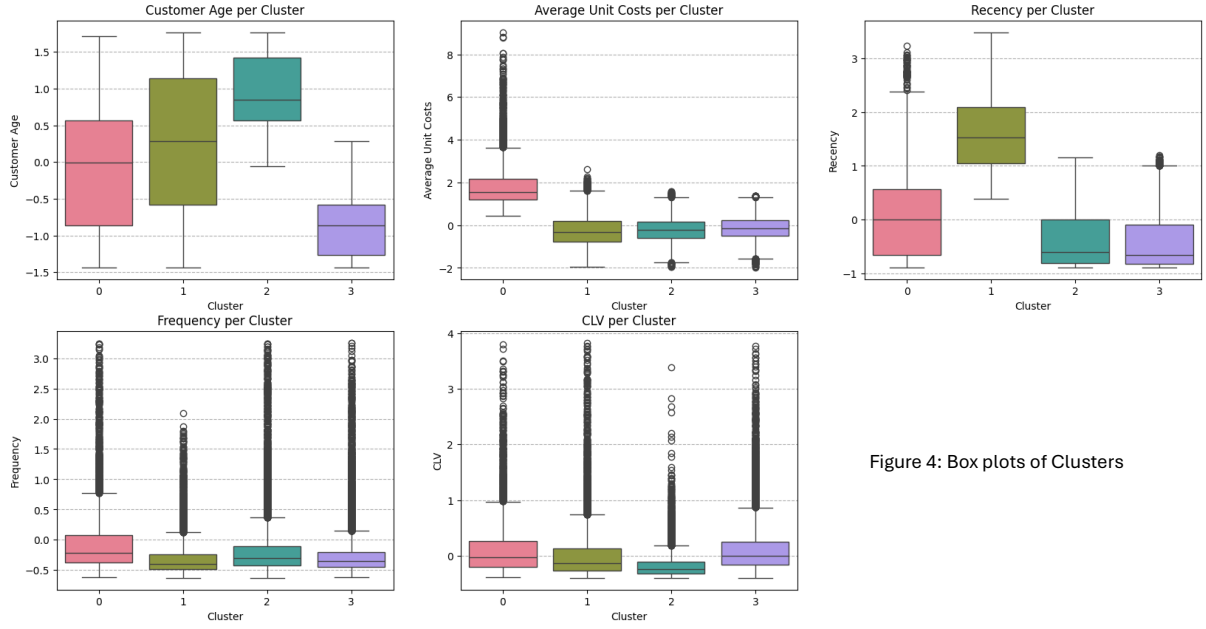

Figure 4: Box plots of Clusters

3.4 Two-dimensional visualisations of the clusters

To visualize customer segmentation, dimensionality reduction was applied using PCA and t-SNE.

Principal Component Analysis (PCA) efficiently projects data onto directions of maximum variance but assumes linear separability, which may not suit this dataset's non-linear relationships. As seen in Figure 5, PCA provided some distinction between clusters but struggled with Cluster 0, blending it with others.

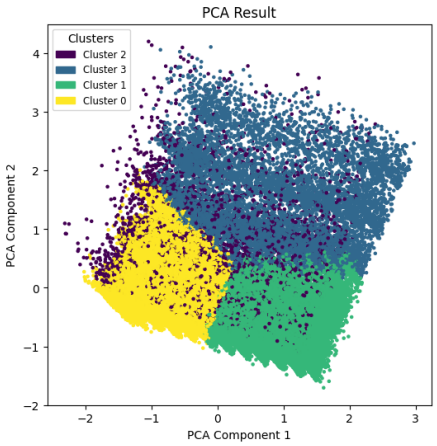

Figure 5: PCA distribution of Clusters

t-Distributed Stochastic Neighbor Embedding (t-SNE), though computationally slower, preserves non-linear relationships better. In Figure 6, t-SNE identified Cluster 2 more distinctly, positioning it centrally among other clusters. This highlights t-SNE's suitability for datasets with complex patterns, offering improved cluster separation over PCA.

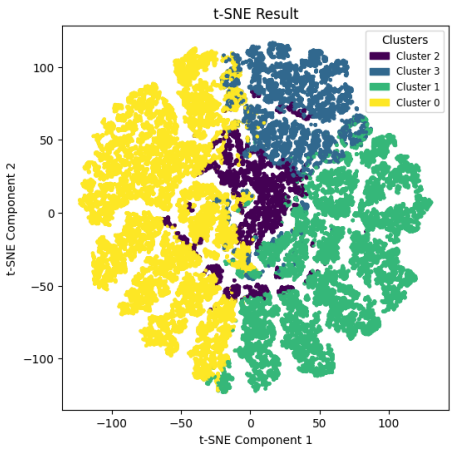

Figure 6: t-SNE distribution of Clusters

4 Explanation of Insights

4.1 Cluster characteristics

Analysis of the box plots (Figure 4) reveals meaningful distinctions between the clusters, shedding light on important customer characteristics regarding the engineered features.

Age

Cluster 0 comprises mostly lower-middle-aged customers, while Cluster 1 represents upper-middle-aged individuals. Cluster 2 is dominated by older customers, and Cluster 3 consists mainly of young customers.

Unit Cost

Cluster 0 has the highest unit costs, making it the most expensive group for the company. In contrast, the other clusters exhibit similar, lower unit costs.

Recency (lower values indicate more recent purchases)

Clusters 2 and 3 have the most recent customers, with similar and significantly lower mean recency compared to Clusters 0 and 1. Cluster 0 includes a mix of

recent and less recent customers, while Cluster 1 skews toward less recent activity.

Frequency (lower values indicate less frequent purchases)

Most clusters show similar frequencies, except Cluster 1, which stands out with higher frequency. Interestingly, these frequent customers are also among the least recent, suggesting possible issues with loyalty or retention over time.

Customer Lifetime Value (CLV)

Cluster 2, comprising older customers, has the lowest average CLV, aligning with their lower potential value to the company. Clusters 0 and 3 show the highest average CLV, with Cluster 3 representing younger customers with significant future potential and Cluster 0 combining frequency and value within a similar age range.

4.2 Geographical trends and patterns

The bar graph in Figure 7 shows that the largest customer base is in Europe, followed by North America, with significantly smaller segments in other continents. The cluster distribution is similar across Europe and North America, with Cluster 0 being the smallest and Cluster 3 the largest.

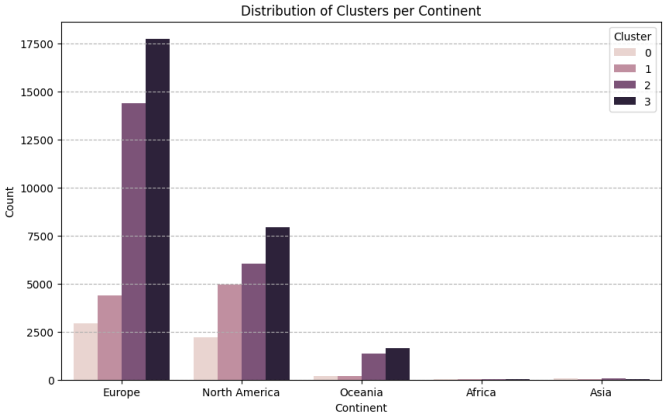

Figure 6: Cluster distribution per continent

Europe

In Europe's top seven countries (Figure 8), cluster distributions are consistent, except in Germany, where Cluster 0 is more prominent. However, Cluster 3 remains the largest group across all countries.

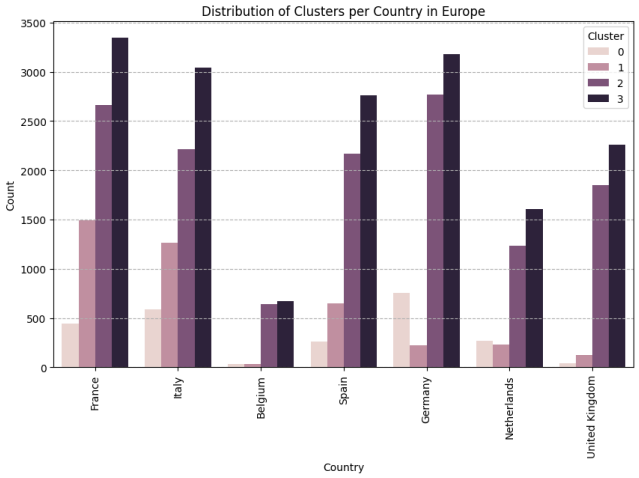

Figure 6: Cluster distribution per Country (Europe

North America

The United States has a much larger customer base than Canada (Figure 9), and both countries share a similar cluster distribution, with Cluster 3 being the largest. Within the U.S., the top 10 states (Figure 10) show an even cluster distribution, with California having the largest customer base and Cluster 3 dominating across all states.

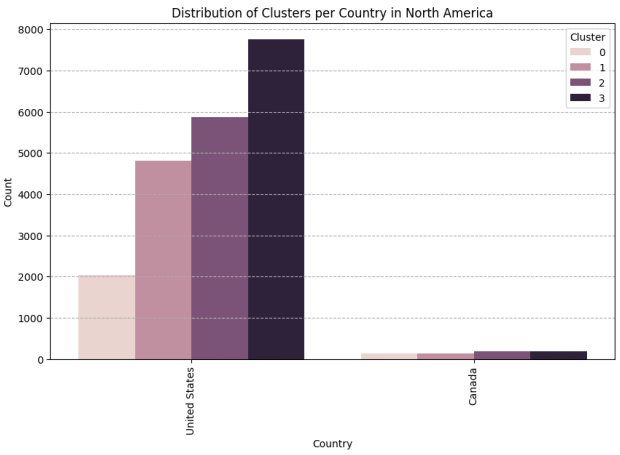

Figure 9: Cluster distribution per Country (North America)

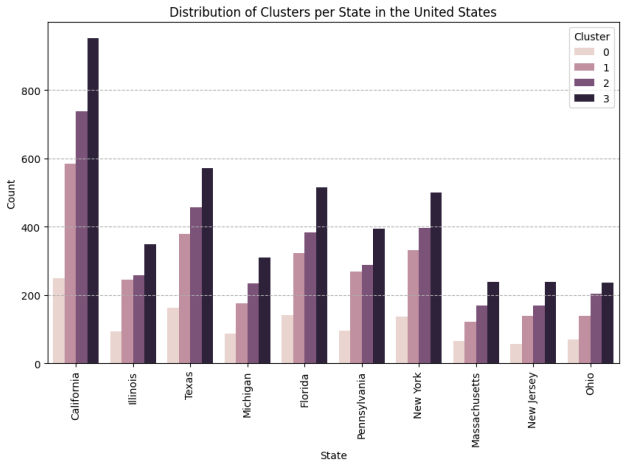

Figure 10: Cluster distribution per US state

5 Conclusion and Recommendations

K-means clustering effectively identified meaningful customer segments, offering insights into who the most valuable customers are, where they are located, and their key behaviors.

Focusing on **Customer Lifetime Value (CLV)**, a key driver of long-term growth, it is recommended to prioritize expanding Cluster 0. While Clusters 0 and 3 share similar average CLV values, Cluster 0 is consistently the smallest. Attracting more Cluster 0 customers, particularly younger individuals, can leverage their high potential value and recover lost profits. Additionally, strengthening the company's existing presence in Cluster 3 can enhance market share and drive sustained growth.