Customer Segmentation Analysis Report

Table of Contents

1. Introduction

Customer segmentation is a fundamental strategy for businesses aiming to enhance customer retention and optimise marketing efforts. By identifying distinct customer groups based on purchasing behaviour, businesses can tailor their strategies to increase engagement and revenue.

This analysis applies **unsupervised machine learning techniques**, specifically **K-Means clustering**, to segment customers into groups based on **frequency of purchases, recency of transactions, customer lifetime value (CLV), average unit cost, and customer age**. The findings provide actionable insights for improving customer engagement.

2. Methodology

Data Preparation and Cleaning

The dataset was first inspected for missing values and duplicates. Missing values in categorical variables (City, Postal Code, and State Province) were replaced with “Unknown” to maintain data completeness. Duplicate entries were removed to prevent data redundancy.

Financial data fields such as total revenue, unit cost, and profit were converted into numerical formats, and date fields were transformed into datetime format for analysis.

Feature Engineering

Several features were created to capture key aspects of customer behaviour:

- **Frequency:** Number of purchases made by each customer.
- **Recency:** Days since the last purchase.
- **Customer Lifetime Value (CLV):** Total revenue per customer.
- **Average Unit Cost:** Mean cost of items purchased.
- **Customer Age:** Derived from the customer’s birth year.

These features were chosen for their relevance in **understanding spending habits, engagement levels, and long-term value**.

Outlier Detection and Handling

Boxplots revealed extreme outliers in **Frequency, CLV, and Average Unit Cost**, which could distort clustering results. To mitigate this, **Winsorization** was applied, capping extreme values at the 95th percentile while preserving overall distribution.

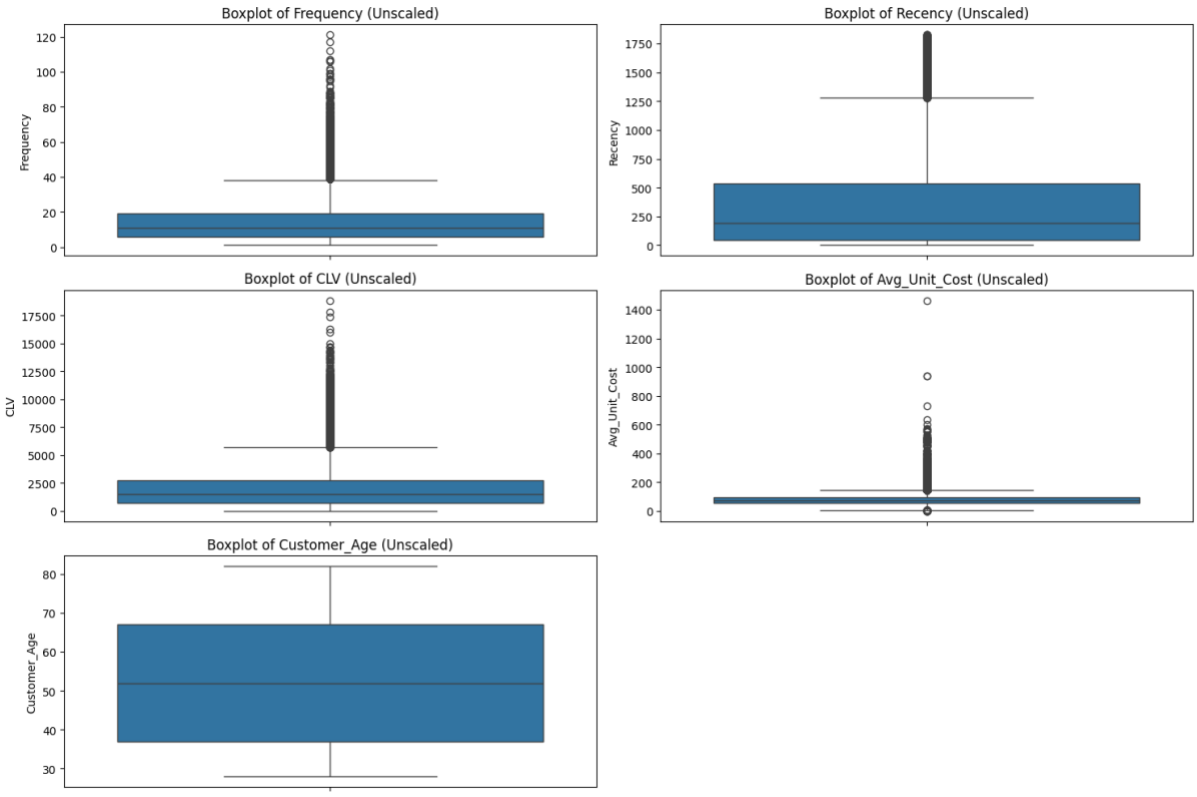

Figure 1: Boxplots Before Winsorization

Feature Scaling

Since clustering algorithms are sensitive to differences in scale, **standardisation** was applied using **StandardScaler**. This transformation ensured all numerical features contributed equally to the clustering process.

3. Determining the Optimal Number of Clusters

Elbow Method and Silhouette Score

The optimal number of clusters (**K**) was determined using the **Elbow Method**, which plots the sum of squared errors (SSE) against different values of K. The point where SSE starts to flatten indicates the best number of clusters.

Additionally, the **Silhouette Score** was computed to evaluate how well-separated the clusters are, with higher scores indicating better-defined groups.

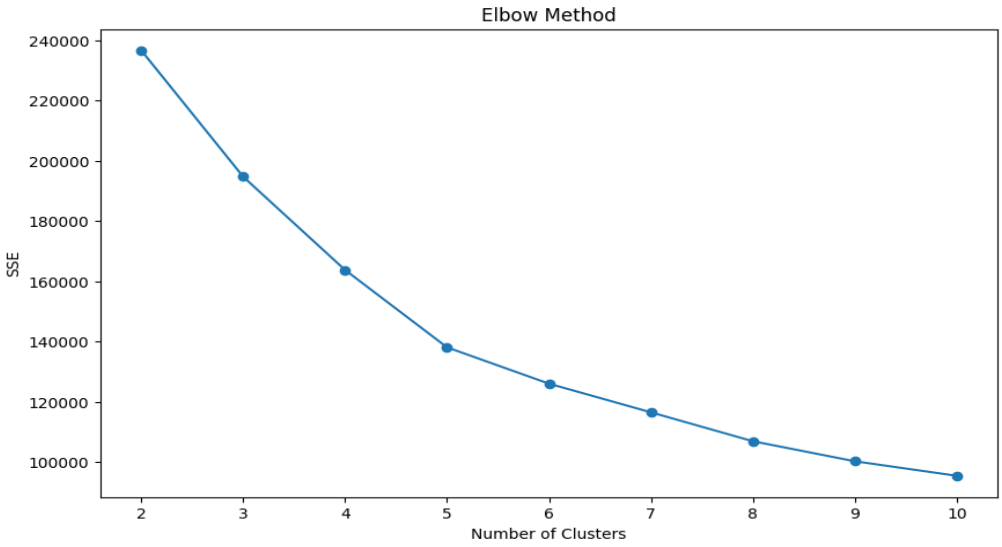

Figure 2: Elbow Method Plot

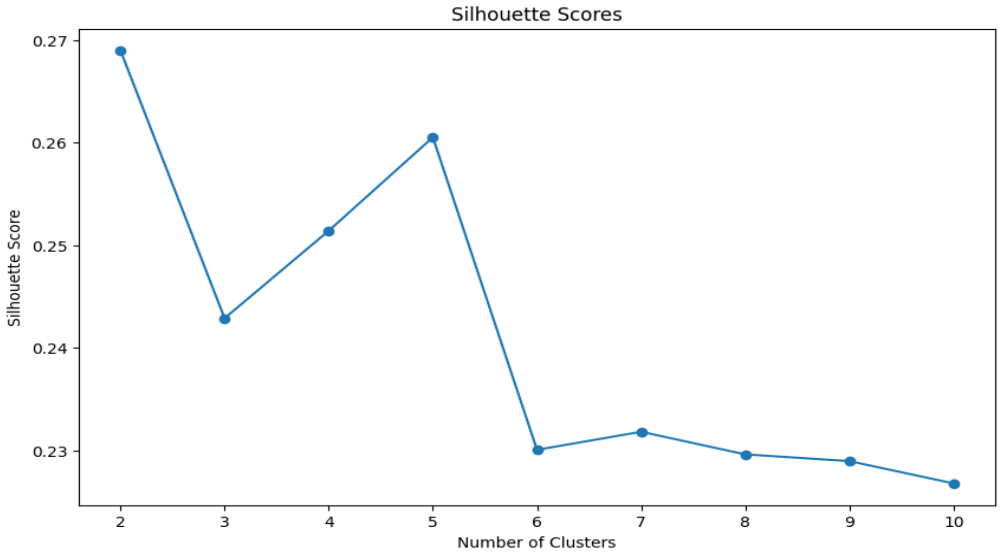

Figure 3: Silhouette Score Plot

Hierarchical Clustering (Dendrogram)

A hierarchical clustering dendrogram was generated to validate segmentation. The clear branching at **K=5** further supported this choice.

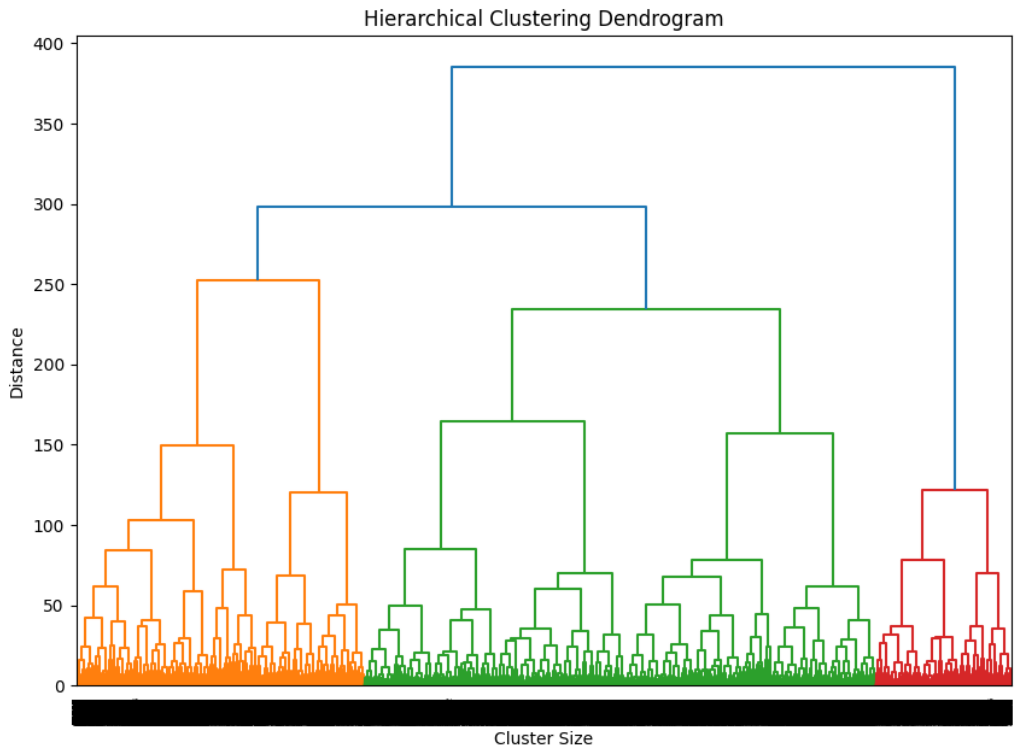

Figure 4: Hierarchical Clustering Dendrogram

4. Customer Segmentation Results

Applying K-Means Clustering

The **K-Means algorithm** was implemented with **K=5**, segmenting customers into five groups based on purchasing behaviour.

Dimensionality Reduction for Visualisation

To visualise clusters effectively, **Principal Component Analysis (PCA)** and **t-SNE** were applied, reducing the dataset's dimensions while maintaining cluster separability.

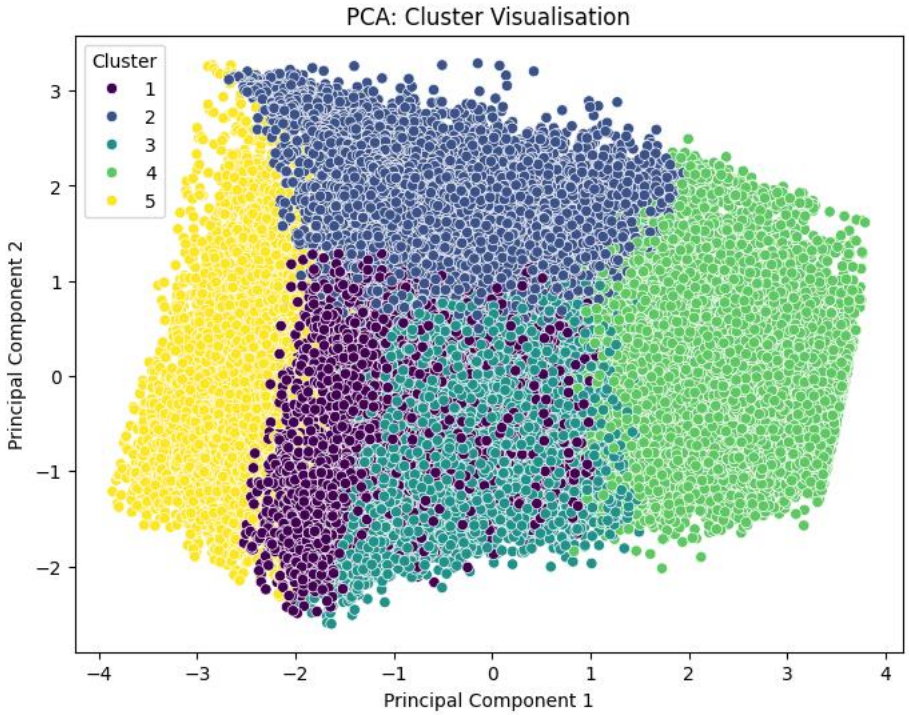

Figure 5: PCA Visualisation

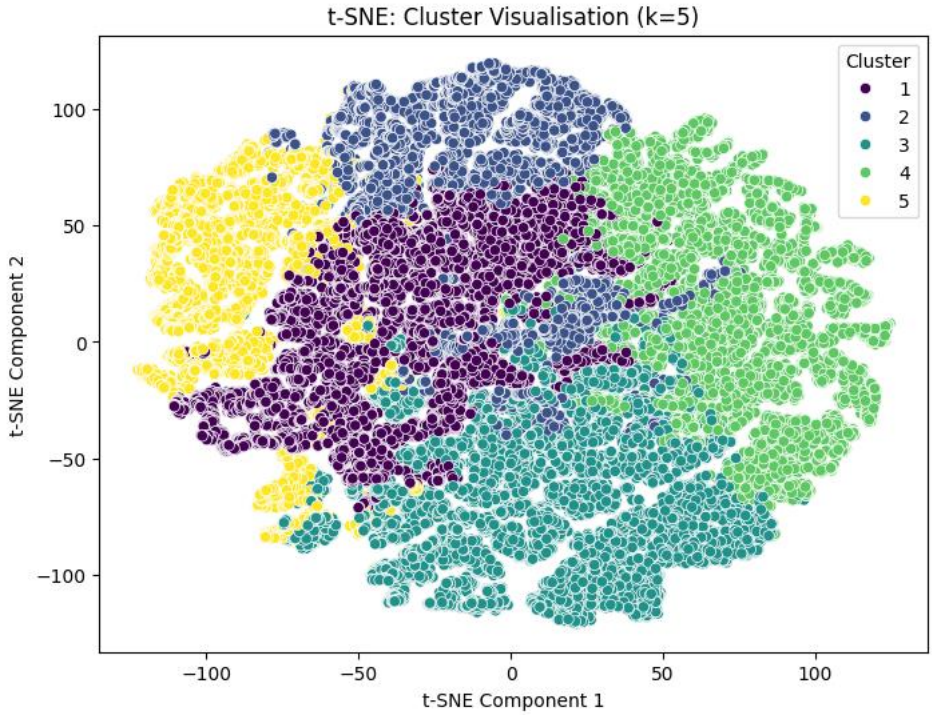

Figure 6: t-SNE Visualisation

Cluster Profiles and Insights

Each cluster represents a distinct customer segment with unique behavioural characteristics:

Cluster	Frequency	Recency	CLV	Avg Unit Cost	Customer Age
1	10.3	229.6	1219.5	64.5	69.4
2	7.3	491.5	1625.1	122.03	52.8
3	11.7	235.9	1397.7	66.49	36.6
4	27.6	130.7	4063.1	82.23	47.9
5	4.6	1143.7	509.7	56.93	57.1

Table 1: Mean Customer Metrics for Each Identified Cluster

Each group reflects varying degrees of purchasing engagement and spending power.

5. Business Implications and Recommendations

The segmentation results can be leveraged to **develop targeted marketing strategies**:

- **Cluster 1 (Loyal Frequent Buyers):** Engaged customers who purchase regularly. Loyalty programs and exclusive discounts could increase retention.
- **Cluster 2 (Premium Buyers):** Customers who spend more per item but purchase less frequently. Luxury promotions and high-end product recommendations would be effective.
- **Cluster 3 (Young, High-Engagement Shoppers):** These customers are active and could benefit from subscription models or influencer-based marketing.
- **Cluster 4 (High-Value Customers):** The most valuable segment, purchasing frequently and spending significantly. VIP memberships and priority customer service could enhance engagement.
- **Cluster 5 (Inactive Customers):** A segment that purchases infrequently and has low spending. Re-engagement campaigns through special promotions may help reactivate them.

By implementing **segment-specific marketing strategies**, businesses can increase **customer retention, revenue, and long-term loyalty**.

6. Conclusion

This analysis successfully segmented customers using **K-Means clustering**, identifying five distinct groups. The choice of **K=5** was validated using the **Elbow Method, Silhouette Score, and Dendrogram analysis**. PCA and t-SNE provided clear visualisations of cluster separability.

The insights derived from this segmentation highlight opportunities for **customer engagement and revenue growth**. Businesses can now implement **data-driven marketing strategies** tailored to each segment, ensuring better **customer experience and profitability**.

Future work could involve **refining feature selection** or integrating additional data sources to further enhance segmentation accuracy.