UNIVERSITY OF CAMBRIDGE

INSTITUTE OF CONTINUING EDUCATION

Mini-Project 2:

Customer segmentation with clustering

Arunima Bhattacharya

June 2025

Contents

List of Figures

Page

1 Introduction

Customer segmentation is the process of grouping customers based on shared traits. Figure 1.1 lists some of these traits.

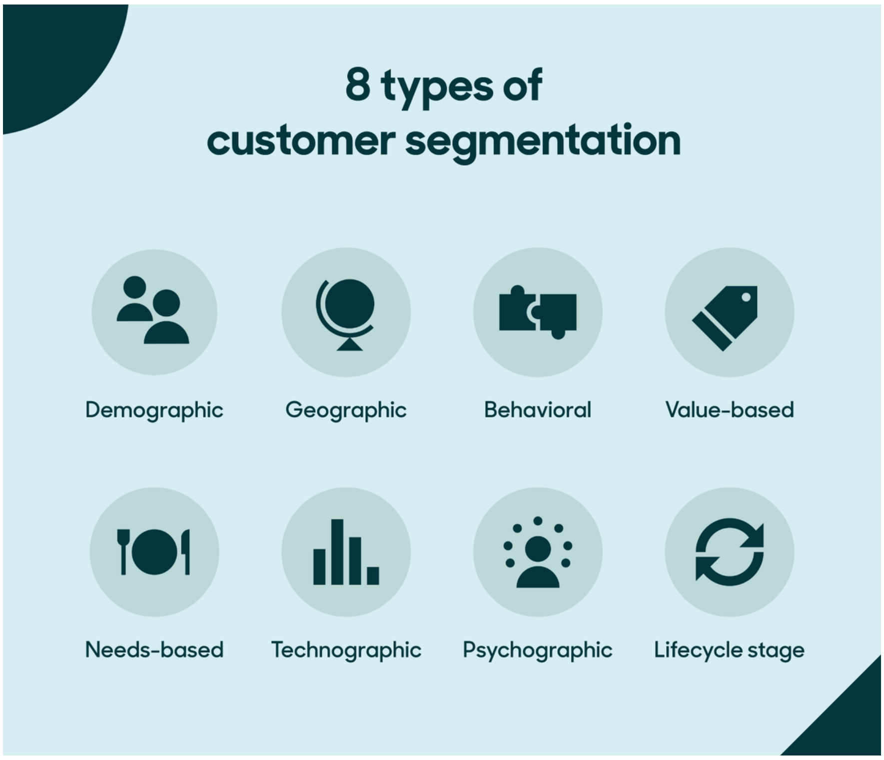


Figure 1.1: Customers can be grouped by different traits like age, location, lifestyle, purchase history, device type, or service preferences.

1.1 Business Context and Problem Statement

This project segments customers from an e-commerce dataset (SAS, 2024) spanning 47 countries across five continents. The goal is to identify meaningful customer groups for strategic actions.

2
Methods

Five features – FREQUENCY, RECENCY, CUSTOMER LIFETIME VALUE (CLV), AVERAGE UNIT COST, and CUSTOMER AGE – were engineered for customer segmentation, from the following original features:

- ✓ Customer ID
- ✓ Order ID
- ✓ Delivery Date
- ✓ Total Revenue
- ✓ Customer Birth Date
- ✓ Unit Cost

Step 1: Checked for missing values in required features.

Step 2: *Feature Generation* — type conversions (e.g., dates to intervals) and aggregations (e.g., revenue per customer for CLV).

Step 3: Clustering is sensitive to scaling and outliers:

- **Step 3a:** *Feature Scaling* — Normalisation/**Standardisation**.
- **Step 3b:** *Outlier Detection* — Interquartile Range (IQR) method.
- **Step 3c:** *Visualisation* — Box plots and KDE histograms.
- **Step 3d:** Entries with ≥ 2 outliers flagged for further analysis.

Step 4: *Clustering* implemented using **k-means** and **hierarchical clustering**.
Figure 2.1 shows the clustering workflow used in this analysis.

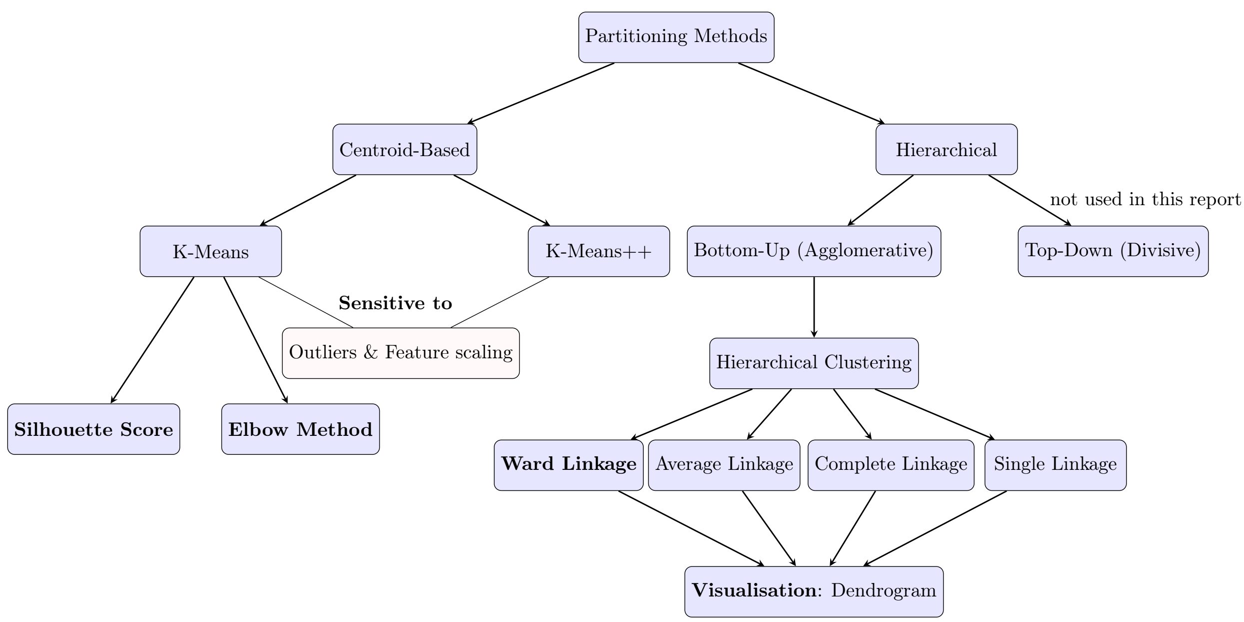


Figure 2.1: Classification of Clustering methods used

3
Results

3.1 Exploratory Data Analysis

1. The dataset contains 121043 missing values (0.64%)
   (a) 135 in 'city'
   (b) 3716 in 'Postal Code'
   (c) 117192 in 'State Province'
2. It includes 21 duplicate samples.

Conclusion:

☞ Columns with missing values were retained, as they weren't used in feature engineering.
☞ Duplicates were removed.

3.2 Outlier Detection using IQR method

The aggregated dataset has 68300 entries and yields the following outlier statistics:

1. Outlier counts for `frequency` vary by scaling method: 2825 (4.14%) with normalisation, 2481 (3.63%) with standardisation.
2. For other features, counts remain consistent across methods
   (a) `recency`: 3353 (4.91%)
   (b) `CLV`: 2590 (3.79%)
   (c) `average unit cost`: 2889 (4.23%)
   (d) `customer age`: 0

Figure 3.1 shows outlier counts using the normalised dataset[1].

3.3 Principal Component Analysis (PCA)

Figure 3.2 shows the cumulative variance explained by PCA components:

1. The first components capture $\sim 62\%$ of variance.
2. First 3 components explain $\approx 84\%$ variance without outliers, and $\approx 82\%$ with outliers.

Conclusion: Outlier removal slightly improves PCA, but the effect is minimal.

3.4 K-Means Clustering Algorithm

3.4.1 Elbow Method and Silhouette score

The silhouette score peaks at 5 clusters, with a 15.88% WCSS drop from 4 to 5, indicating better cohesion and separation. See Figure 3.3 and Table 3.1.

[1] Used for better visualisation; standardised data was used for analysis.

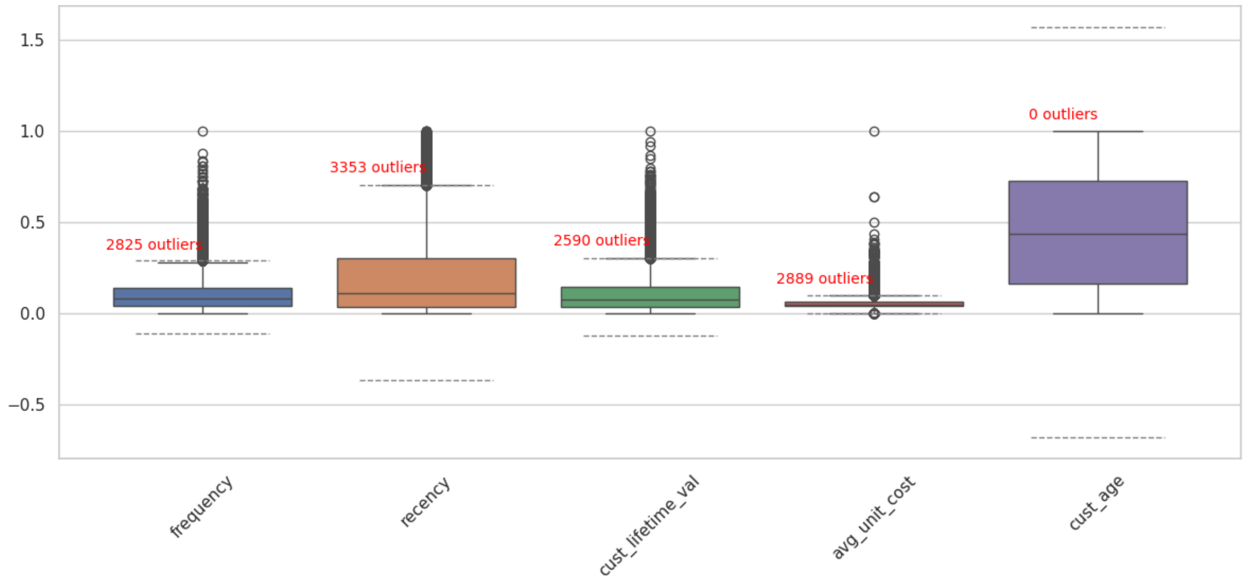


Figure 3.1: Box plot visualisation of IQR results and outliers on normalised data

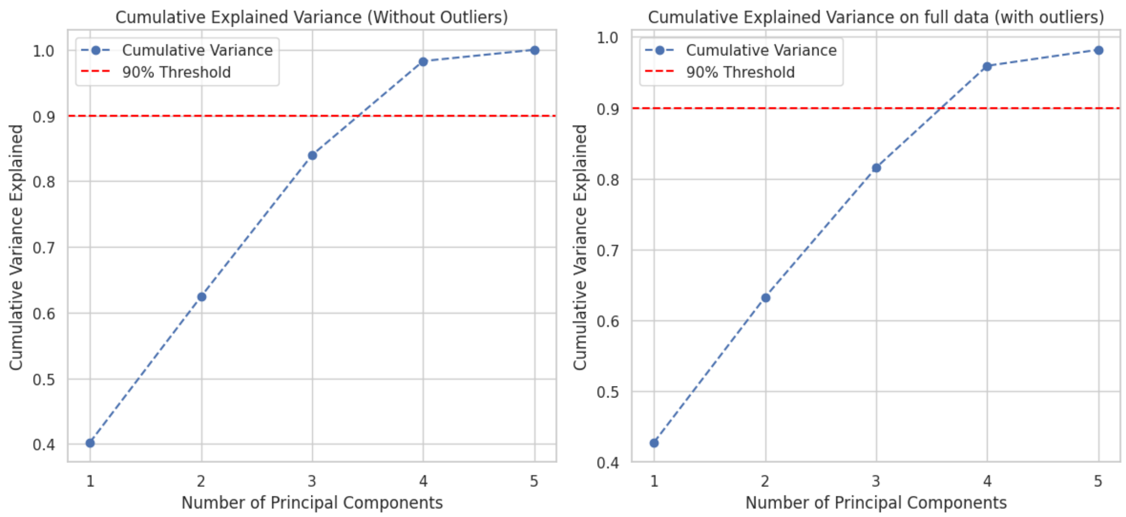


Figure 3.2: Cumulative Variance with the number of principal components

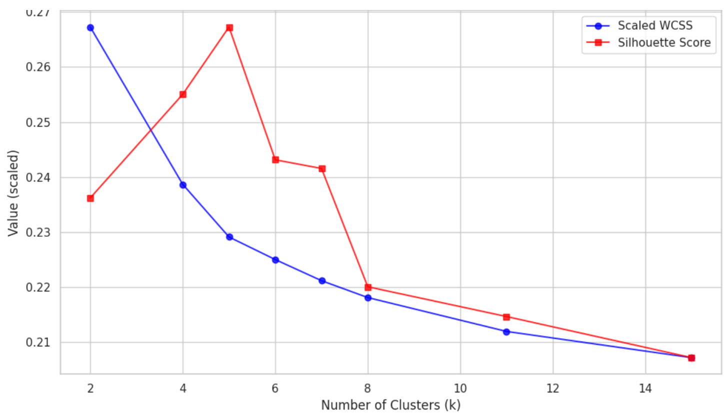


Figure 3.3: Visual comparison of scaled WCSS and silhouette scores across different cluster counts.

Clusters	WCSS	Silhouette score
2	211872.0867	0.2362
4	143664.9072	0.2550
5	120843.3715	0.2672
6	111055.4590	0.2431
7	101902.1495	0.2416
8	94580.9537	0.2201
11	79896.9552	0.2147
15	68535.5317	0.2072

Table 3.1: Numerical comparison of WCSS and silhouette scores

3.4.2 Silhouette Plot, t-SNE and PCA Projection

Figure 3.4 shows silhouette scores with t-SNE and PCA projections for 4, 5 and 6 clusters. The trade-off between overlap, silhouette, and WCSS clearly supports 5 as optimal.

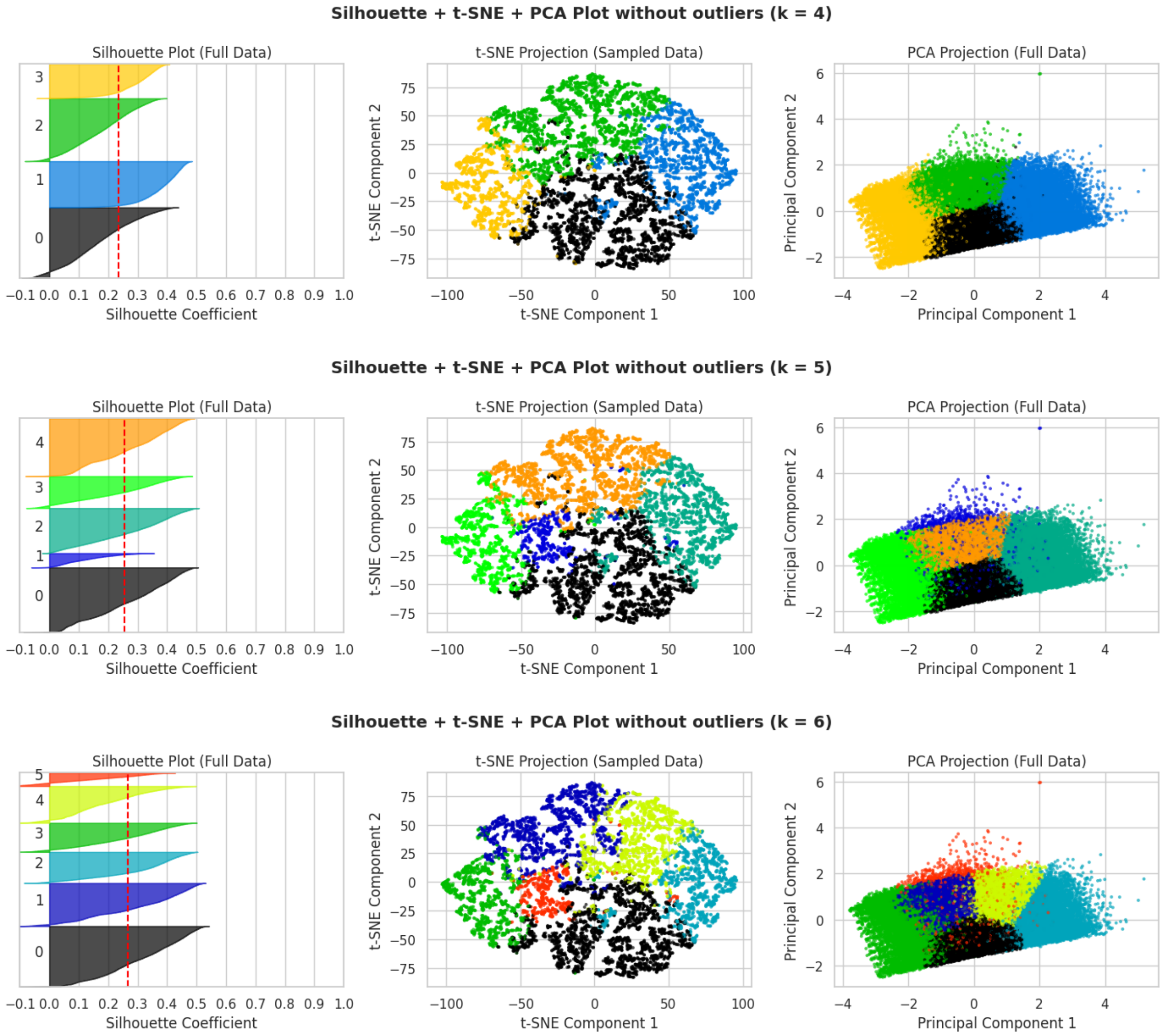


Figure 3.4: Visual comparison of clustering quality for $k = 4$, 5, and 6 using Silhouette plots (left), t-SNE projections (middle), and PCA projections (right). For $k = 4$, silhouette score is 0.26, showing four broad clusters with good separation. For $k = 5$, the score improves to 0.27 with some overlap. For $k = 6$, the score drops to 0.24 with increased overlap in both t-SNE and PCA views.

3.4.3 Hierarchical clustering and Dendrogram

Figure 3.5[2] shows a colour-coded dendrogram with threshold distance 90, yielding a practical $k = 5$ segmentation that preserves meaningful groupings and aligns with the hierarchy, despite not capturing the absolute maximum inter-cluster distance.

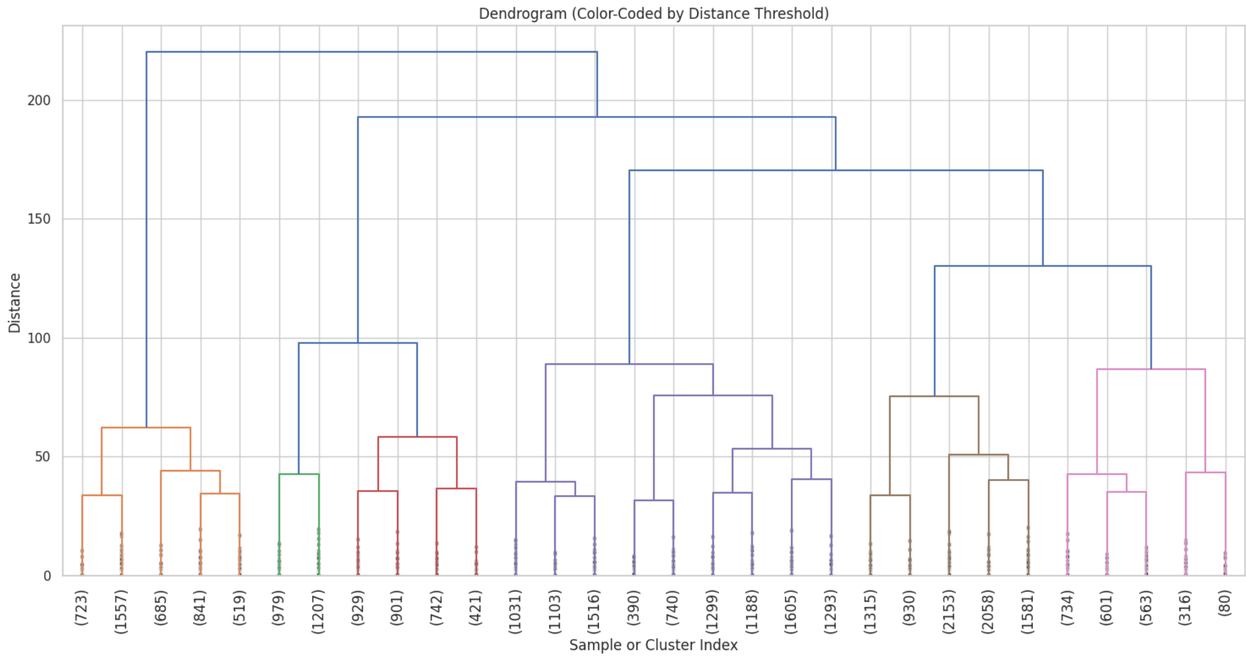


Figure 3.5: Dendrogram generated using the Ward linkage method and Euclidean distance metric

3.5 Feature-wise Statistical Profile of the Five Clusters

To understand the behavioural patterns of each segment, figure 3.6 presents box plots of the key standardised features across the five clusters (excluding outliers).[3] Table 3.2 complements this by summarising each cluster's profile across the key features and highlights which clusters perform best or worst across these dimensions.

Each cluster presents distinct and interpretable traits, namely:

- Cluster 4 ⇒ `Senior Low-Value Purchaser`
- Cluster 3 ⇒ `Churned Value-Less Purchaser`
- Cluster 2 ⇒ `Loyal High-Value Purchaser`
- Cluster 1 ⇒ `Infrequent High-Spender`
- Cluster 0 ⇒ `Young Low-Value Purchaser`[4]

[2] Constructed using 30,000 samples due to memory constraints.
[3] See Appendix figure 3 for distributions including outliers.
[4] Clusters 0 and 4 can be merged as `Low-Value Purchasers` in the segmentation without age-based split.

Clust.	**Freq.**	**Recency**	**CLV**	**Avg. Cost**	**Age**	**Summary**
4	Moderate	Low	Low	Moderate	**Oldest**	Older low-value group
3	**Worst**	**Worst**	**Worst**	Lowest	Mixed, older-leaning	Dormant/least engaged
2	High	**Best**	**Best**	Moderate	Mixed, skewed-young	Best: loyal, frequent
1	Worst after 3	Moderate	Moderate	**Highest**	Balanced	Infrequent, premium spenders
0	Moderate	Low	Low	Low	**Youngest**	Young low-value group
-1	**Extremely High**	High	**Extremely High**	Moderate	Mixed, skewed-young	Elite high-frequency outliers

Table 3.2: Feature-wise summary of cluster characteristics (including outliers -1)

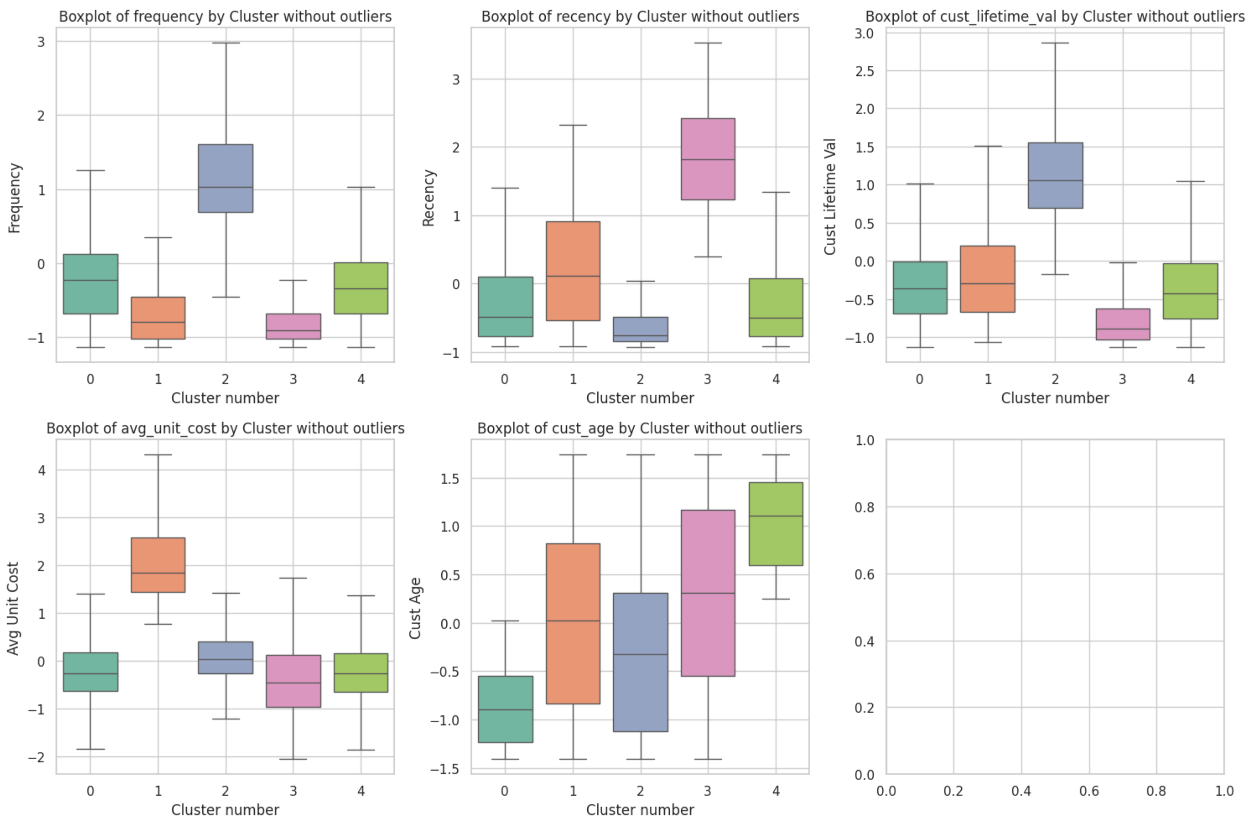


Figure 3.6: Box-plots of feature-wise statistical behaviour after clustering

4
Conclusion

Optimal customer segmentation is achieved with five clusters. Outliers form a distinct segment.[5] The resulting segments are distinct and business-actionable with proportions:

1. **Loyal High-Value Purchasers — 20.44%**
2. **Churned Value-Less Purchasers — 14.46%**
3. **Low-Value Purchasers**: Young is **29.56%**; Senior is **26.18%**
4. **Infrequent High-Spenders — 6.45%**
5. **Elite Group (Outliers) — 2.90%**

Figure 4.1 clearly illustrates this.

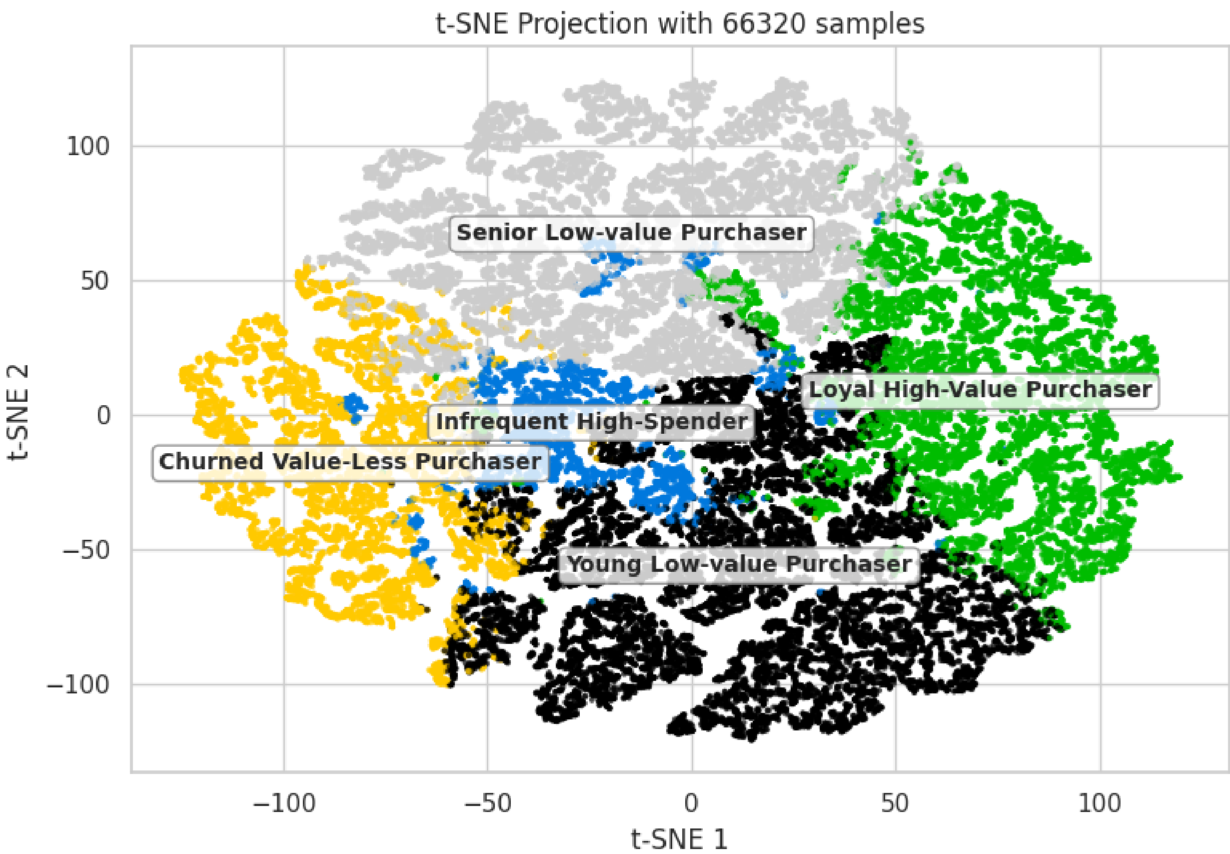
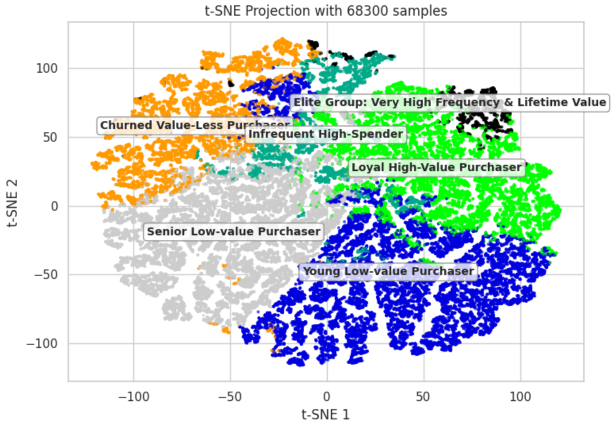


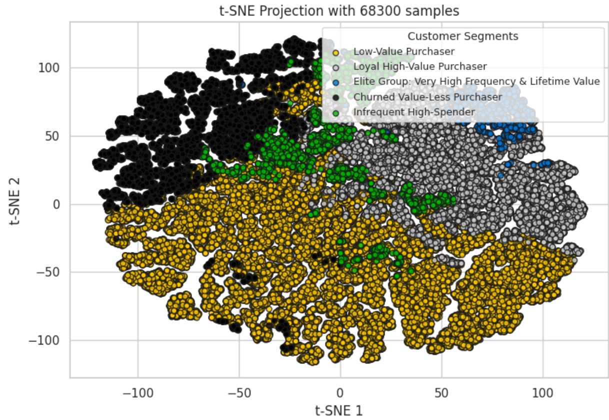
Figure 4.1: t-SNE projections without outliers of the five derived clusters shows clear segmentation and structure

Figure 4.2 presents the t-SNE and PCA projections of the 5 clusters with outliers.

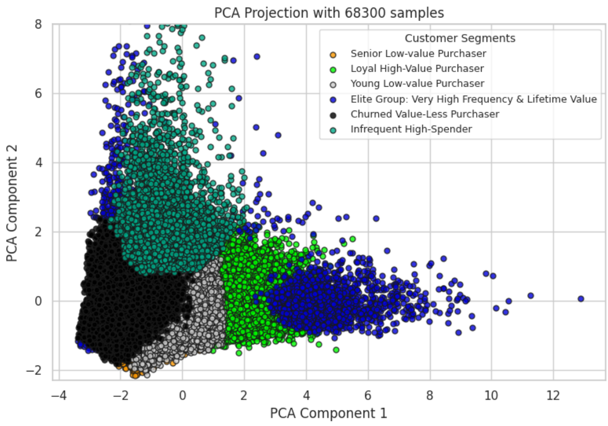
[5]If outliers are detected solely using the IQR method without row-level filtering, the optimal cluster count reduces to four with 13.66% outliers. See figure 4.



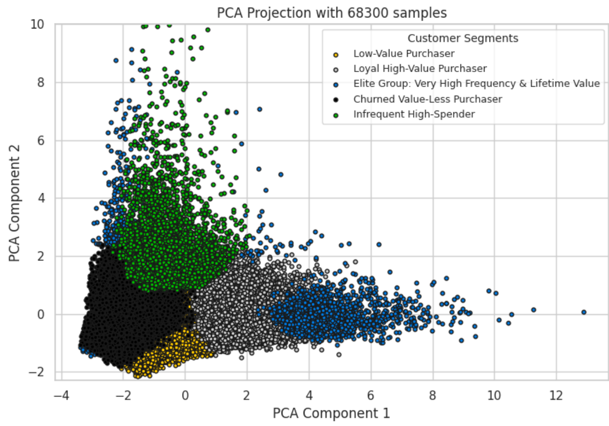
(a) With age-based split



(b) Without age-based split



(c) With age-based split



(d) Without age-based split (not clearly visible)

Figure 4.2: t-SNE and PCA projections of customer clusters across segmentation strategies, highlighting the Outlier "Elite" group and comparing versions with and without age-based splits among low-value customers. Age segmentation is obscured in the PCA projections but enhances granularity in t-SNE with stable core cluster structures.

FEATURE-WISE BOX-PLOTS

with outliers as "cluster -1"

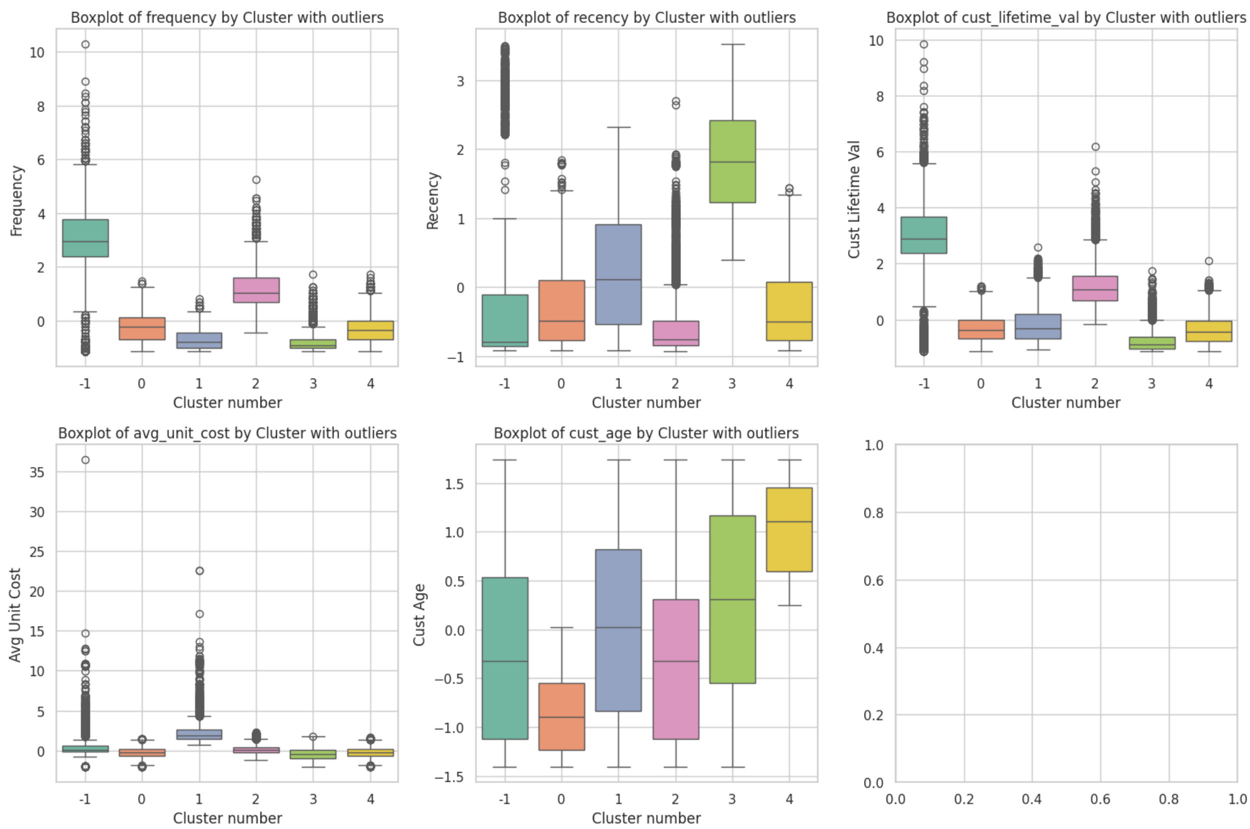


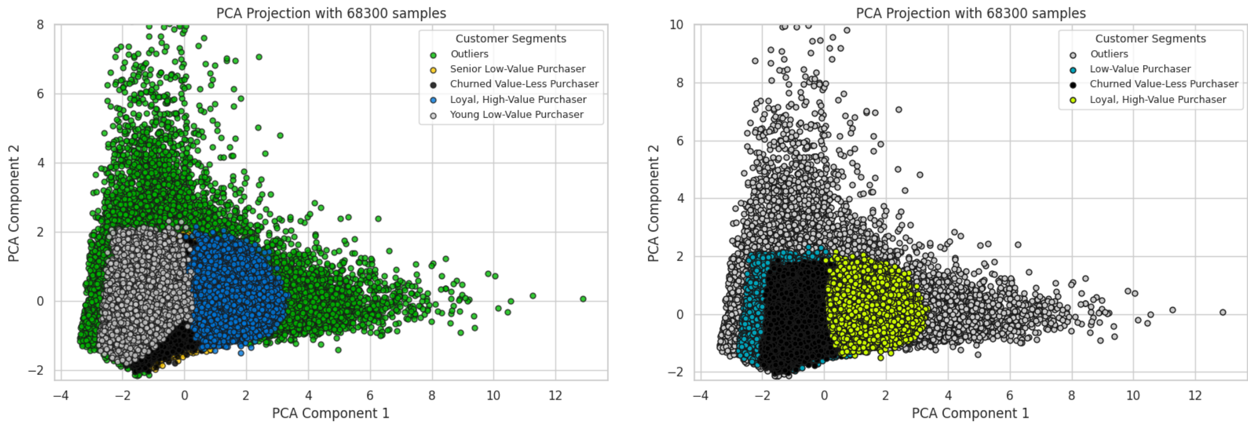
Figure 3: Box-plots showing feature-wise statistical behaviour with outliers labelled as "-1".

The 2.9% outliers comprise a distinct group of `Exceptionally High-Frequency High-Value Customers`. They most closely resemble Cluster 2 (`Loyal High-Value Purchaser`) but surpass it in engagement and value. This might be worth retaining, studying, and modelling.

CLUSTERING

without row-wise Outlier Filtering

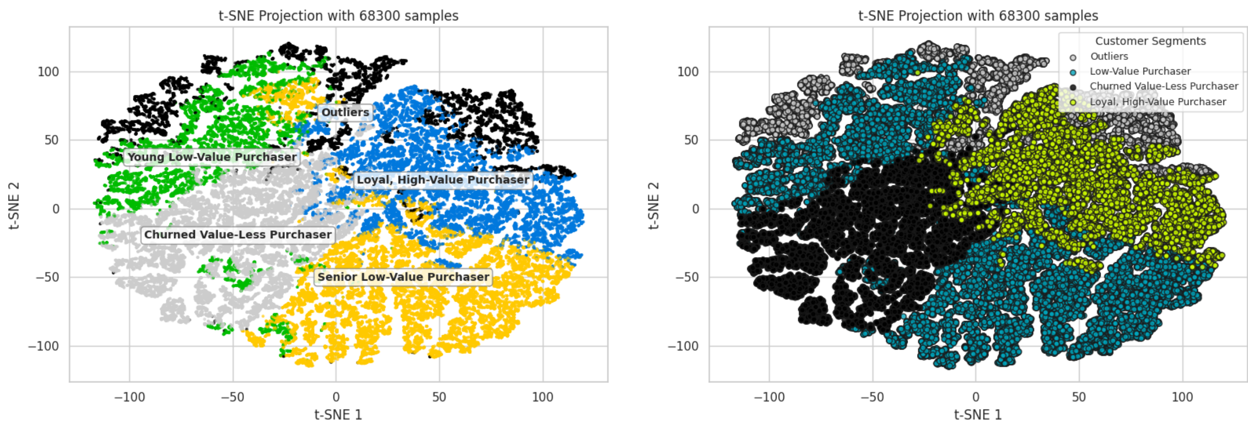
PCA Projections



(a) With age-based split (not clearly visible) **(b)** Without age-based split

Figure 4: PCA projections of customer clusters with outliers, under different segmentation strategies.

t-SNE Projections



(a) With age-based split **(b)** Without age-based split

Figure 5: t-SNE projections of customer clusters with outliers, under different segmentation strategies.

Conclusion: Aggressive outlier flagging eliminated the Infrequent High-Spender segment entirely and over-restricted the outlier group.