Course 101
6.3 Mini-project
Date: May 2024

Customer Segmentation with Clustering

By Joshua Dixon

Executive Summary

This report analyses customer segmentation using clustering techniques on an e-commerce dataset from SAS. Key insights include the identification of three distinct customer clusters using Ward Linkage, which provided a balanced distribution.

The analysis focused on five key features: frequency, recency, customer lifetime value (CLV). Recency appears to be the most distinct feature differentiating the clusters, whilst the remaining had a less notable impact. Dimensionality reduction via PCA and t-SNE showed clear cluster separation, with Ward Linkage with 3 clusters being optimal. Recommendations focus on targeted engagement strategies to maximise customer value and retention.

Recommendations

- Monitor Cluster 0: Address potential churn with loyalty points, tailored promotions, and engagement campaigns. These customers are in the middle range for recency and spending but have outliers with high CLV, indicating a mix of potential high-value customers and those at risk of churn.
- Re-engage Cluster 1: Increase purchase frequency and CLV with incentives, promotions, and targeted marketing strategies. Explore cross-selling and upselling opportunities. Since these customers have the least recent purchases (higher recency values) and the low spending, they may benefit from targeted re-engagement strategies.
- Target Cluster 2: Enhance engagement with loyalty programs, exclusive offers, and tailored communications. These customers have the most recent purchases (lower recency values), indicating recent engagement. Monitoring external influences on their purchasing behavior will help in maintaining their activity and increasing their value.

By focusing on these insights and recommendations, the business can better tailor its marketing and engagement strategies to maximise customer value and retention. Additionally, it is recommended to conduct further detailed analysis incorporating a larger data sample and more of the remaining features.

Table of Contents

Executive Summary 2

Recommendations 2

1 Introduction 4

2 Methodology 5

2.1 Data Source 5

2.2 Key Steps in Analysis 5

3 Results 6

3.1 Clustering with Machine Learning Models 6

3.1.1 K-means clustering 6

3.1.2 Hierarchical Clustering 7

3.1.3 Justification of optimal number of Clusters 8

3.2 Customer Segments 9

3.2.1 Cluster Association with Customer ID 9

3.2.2 Relationship between Clustering and Key Features 10

3.3 Dimensionality Reduction 12

3.3.1 PCA, t-SNE and Explained Variance 12

3.3.2 Key observations in relation to customer segmentation 13

4 Conclusion 13

5 Reference 13

1 Introduction

The primary aim of this project is to identify distinct customer groups within an e-commerce dataset to better understand the customer base and enhance the company's marketing efforts. By segmenting customers based on their purchasing behaviour and demographics, the company can adopt a more customer-centric approach, improve marketing efficiency, and ultimately boost customer satisfaction and retention.

This analysis includes K-means and hierarchical clustering, optimising clusters with the elbow method and silhouette scores, and visualizing clusters with PCA and t-SNE. Outliers will only be commented on.

2 Methodology

2.1 Data Source

Customer segmentation using clustering techniques was conducted on an e-commerce dataset provided by SAS, which includes transactions from customers across 47 countries between January 2012 and December 2016. A sample of 9,000 observations was used for this analysis due to the computational intensity of the full dataset containing 951,668 observations.

2.2 Key Steps in Analysis

The following key steps outline the methodology used for this analysis:

1) Data Preprocessing:
    a. Handle missing values, duplicates, and outliers.
2) Feature Engineering:
    a. Create and select relevant features (Frequency, Recency, CLV, Average Unit Costs, Age), and scale them.
3) Clustering:
    a. Apply K-means and hierarchical clustering, and justify the number of clusters.
4) Customer Segments:
    a. Display clusters in tables and visualisations.
5) Dimensionality Reduction:
    a. Use PCA and t-SNE for visualisations and comment on their effectiveness.

Libraries Used: Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, SciPy.

3 Results

3.1 Clustering with Machine Learning Models

3.1.1 K-means clustering

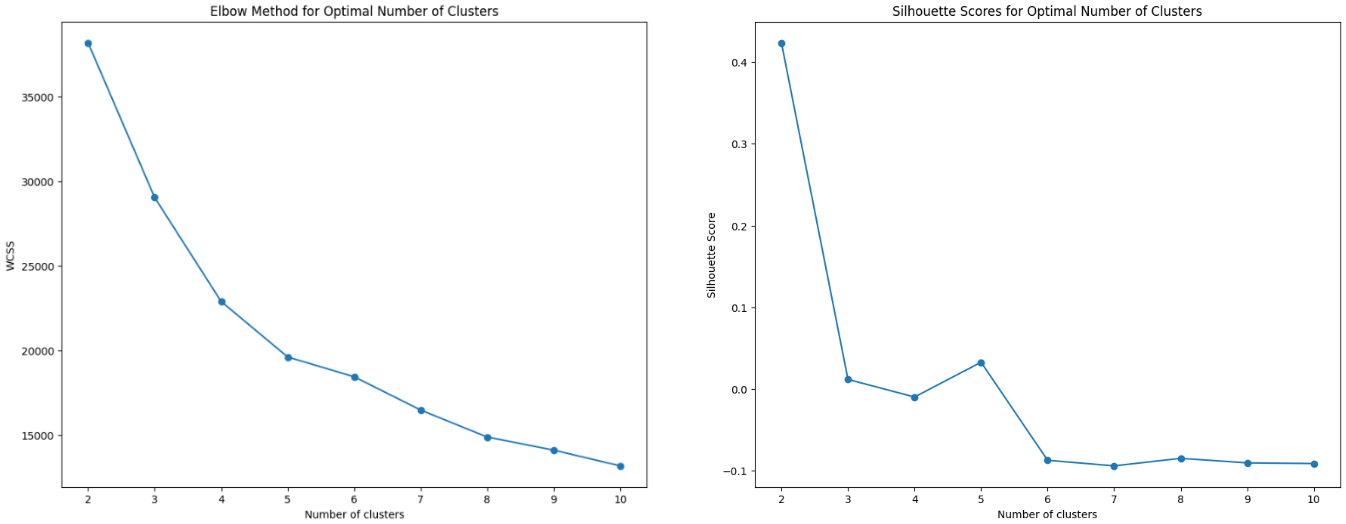

Figure 1: K-Means Clustering, Elbow Method (Left) and Silhouette Method (Right)

- The Elbow Method indicates the optimal number of clusters at around 4 or 5, where adding more clusters does not significantly decrease the WCSS.
- Silhouette scores are highest for 2 clusters, but decline sharply with more clusters, suggesting reduced clustering quality.

3.1.2 Hierarchical Clustering

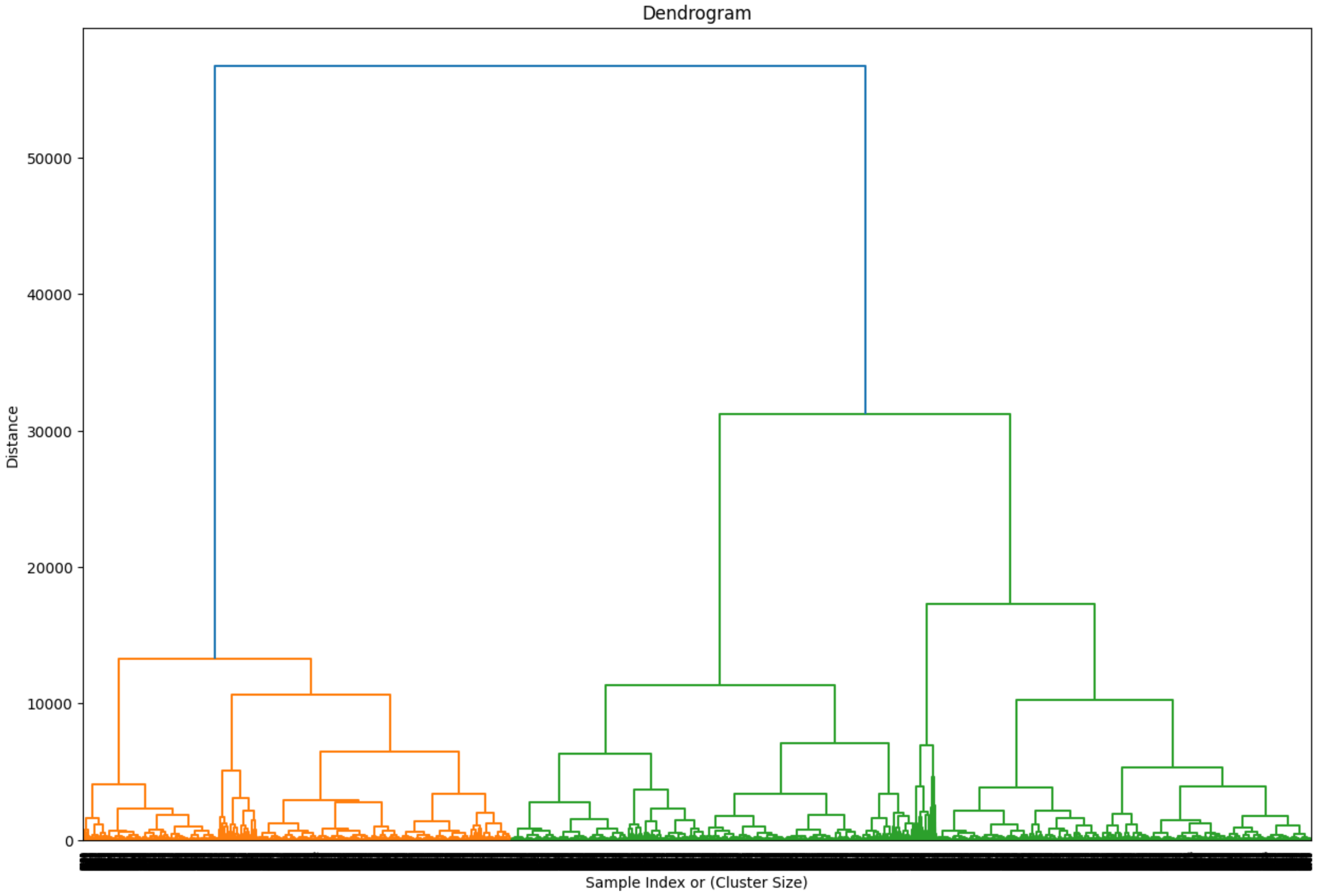

Figure 2: Dendrogram using Ward Linkage

- Figure 2 shows the dendrogram using the Ward Linkage method, visualizing the optimal number of clusters.
- Hierarchical clustering with Average and Ward Linkage found Ward Linkage provided a more balanced distribution, ideal for practical applications like marketing.
- Visualising the dendrogram suggests three distinct clusters of approximately equal size.

3.1.3 Justification of optimal number of Clusters

Table 1: Linkage and Cluster Number Impact on Cluster Size/Distribution

Cluster	Average Linkage with 5 Clusters	Average Linkage with 2 Clusters	Ward Linkage with 5 Clusters	Ward Linkage with 2 Clusters	Ward Linkage with 3 Clusters
0	36	8998	2956	5857	2901
1	11	2	2748	3143	3143
2	8950	-	2143	-	2956
3	2	-	153	-	-
4	1	-	1000	-	-

Table 2: Linkage and Number of Cluster Impact on Silhouette Score

Linkage Method	Number of Clusters	Silhouette Score
Average Linkage	5	0.5401
Average Linkage	2	0.7943
Ward Linkage	5	0.4052
Ward Linkage	2	0.4895
Ward Linkage	3	0.4487

- Table 1 shows the impact of varying linkage methods and cluster numbers on cluster size and distribution. It suggests that Ward Linkage with 3 Clusters is optimal due to the even spread of observations across clusters.
- Table 2 indicates that Average Linkage yielded higher silhouette scores, signifying well-separated clusters. However, the uneven distribution of cluster sizes undercuts their practical utility.
- Combined analysis of K-Means, Hierarchical Clustering, and Elbow and Silhouette Methods indicate Ward Linkage with 3 clusters as the best model.
- Although Ward Linkage with 2 clusters had similar silhouette scores, the 3-cluster model is preferred for business relevance and strategy granularity.

3.2 Customer Segments

3.2.1 Cluster Association with Customer ID

Customer IDs and their associated clusters have been tabulated in a csv file for information. An extract of the first 10 rows has been provided in this report for reference.

Table 3: Extract of "customer_clusters_ward_3.csv"

Customer ID	Cluster
15142	1
32441	0
86361	1
464	0
87565	0
2029	0
27559	1
51612	0
38390	0
84954	2

3.2.2 Relationship between Clustering and Key Features

Figure 3 overleaf displays the visual plots for the clusters in relation to key features. The following inferences can be drawn:

Summary:

- Frequency: Most customers have a low purchase frequency (1 purchase), which progressively reduces as the count increases.
- Recency: Cluster 2 has the most recent purchases, while Cluster 1 has the least. Cluster 0 is in the middle, with outliers on both high and low ends. Recency appears to be the most distinct feature captured between clusters.
- Customer Lifetime Value (CLV): CLV is relatively consistent across clusters, but Cluster 0 has a higher concentration and greater spread of outliers.
- Unit Cost: Unit costs are similar across clusters, with outliers present in all clusters. Cluster 0 has outliers that are notably higher than Cluster 1 and 2.
- Age: The age distribution is uniform across all clusters, indicating age does not significantly impact clustering.

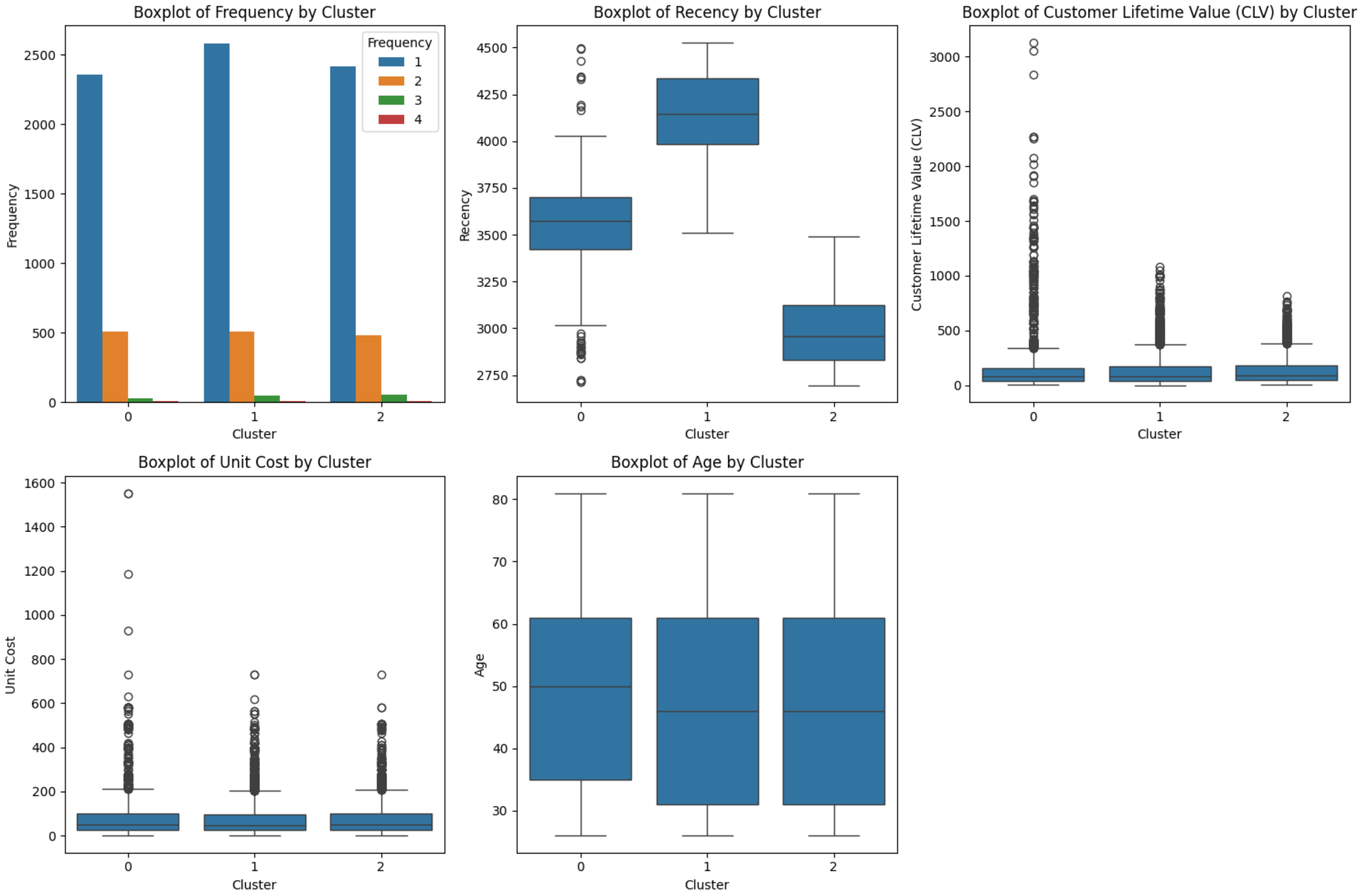

Figure 3: Visual Plots of Clusters Associated with Key Features

3.3 Dimensionality Reduction

3.3.1 PCA, t-SNE and Explained Variance

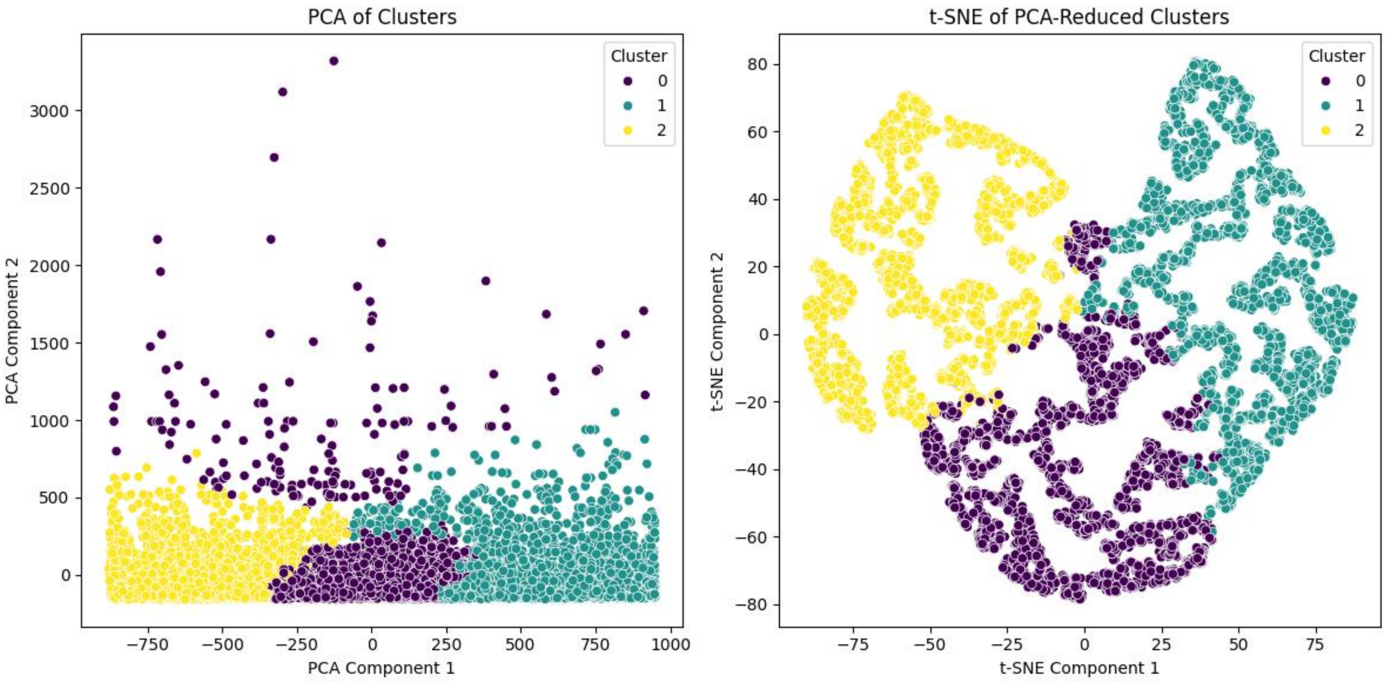

Figure 4: PCA of Clusters(left) and t-SNE of PCA-Reduced Clusters (right)

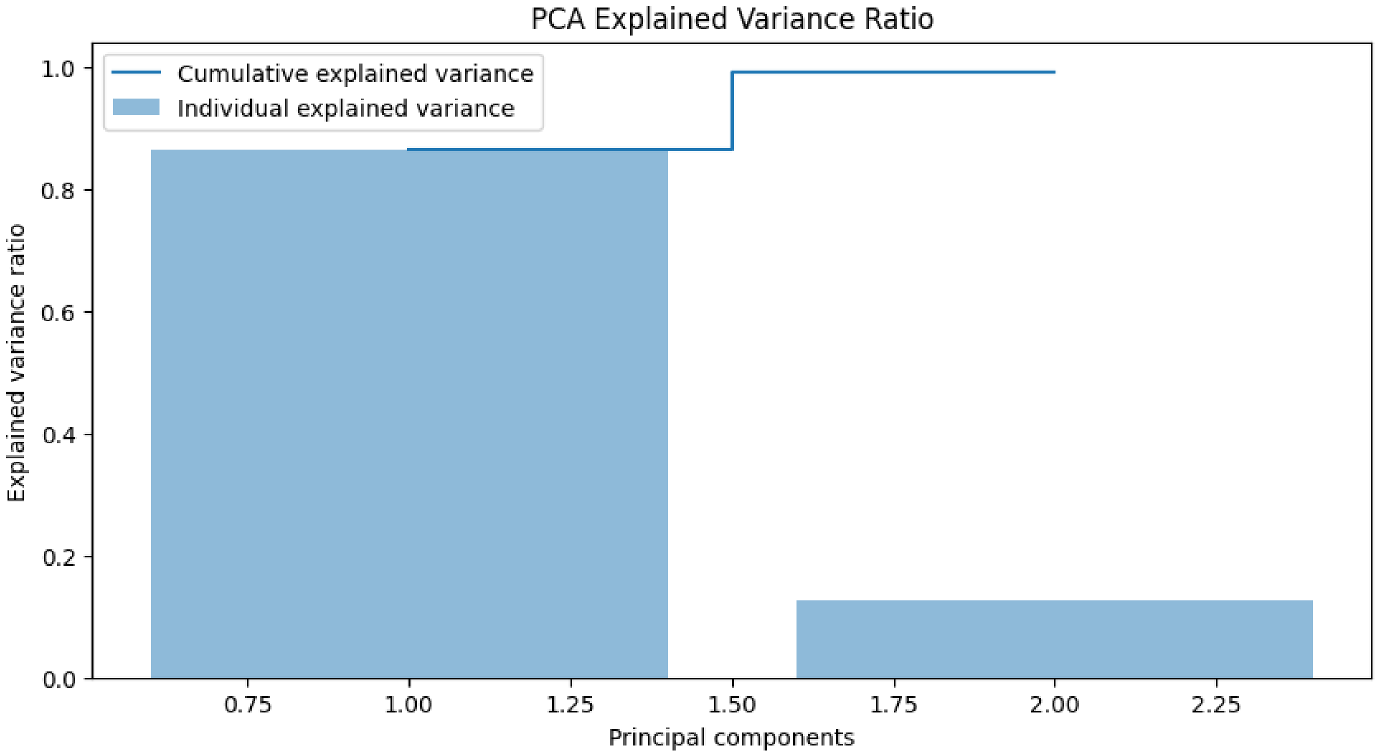

Figure 5: PCA Explained Variance Ratio

3.3.2 Key observations in relation to customer segmentation

PCA Component 1 and Component 2 show a clear separation of clusters, indicating distinct characteristics captured by linear combinations of the original features. The t-SNE plot provides a non-linear projection of the PCA-reduced data, offering a more nuanced view of cluster separations, which are well-defined and distinct. The first principal component explains approximately 85% of the variance, while the second explains about 15%. Combined, they capture nearly 100% of the variance in the data.

4 Conclusion

This report presents customer segmentation from a SAS dataset. K-Means and Hierarchical Clustering identified three optimal clusters with balanced distributions and distinct patterns. Recency is the key distinguishing feature. Dimensionality reduction using PCA and t-SNE confirmed clear cluster separation. Implementing the recommended strategies can enhance customer engagement, increase CLV, and optimise marketing spend, leading to cost savings and improved customer retention.

5 Reference

SAS, 2024. CUSTOMERS_CLEAN [Data set]. SAS. Last revised on 15 December 2021. [Accessed 20 February 2024].