CUSTOMER SEGMENTATION WITH CLUSTERING

Applying Statistical & Machine Learning techniques to do customer segmentation with clustering

Presented by – Poonam Vyas

Date – 20th September 2025

Table of Contents

1. Problem Statement

In today’s competitive e-commerce landscape, understanding customers is one of the most crucial business strategies. Customer segmentation enables companies to categorize their customer base into meaningful groups and tailor marketing strategies accordingly. The dataset provided contains over 950,000 records from transactions across multiple continents.

Each record details customer demographics, order characteristics, and purchase behavior. The challenge is to process this dataset, create meaningful features, identify patterns, and apply clustering techniques to segment customers effectively. The goal is to uncover actionable insights that can guide marketing and resource allocation.

2. Methodology

Data Processing

The initial dataset contained **951,669 rows** and **20 columns**, with each row representing a product ordered by a customer. During preprocessing:

- **Duplicate checks** revealed 21 duplicates.
- Data was aggregated at the **Customer ID** level to ensure one row per customer, reducing the dataset to **68,300 rows**.

Feature Engineering

To support effective segmentation, five new features were created:

- **Frequency** – how often a customer purchases.
- **Recency** – how recently a customer made a purchase.
- **Customer Lifetime Value (CLV)** – the total contribution of the customer over time.
- **Average Unit Cost** – preference for high-cost vs. low-cost products.
- **Customer Age** – derived from birthdate.
- After adding these features, apart from these 5 columns and Customer ID rest all columns were removed. This left a customer-centric dataset with a compact yet informative set of features for clustering.

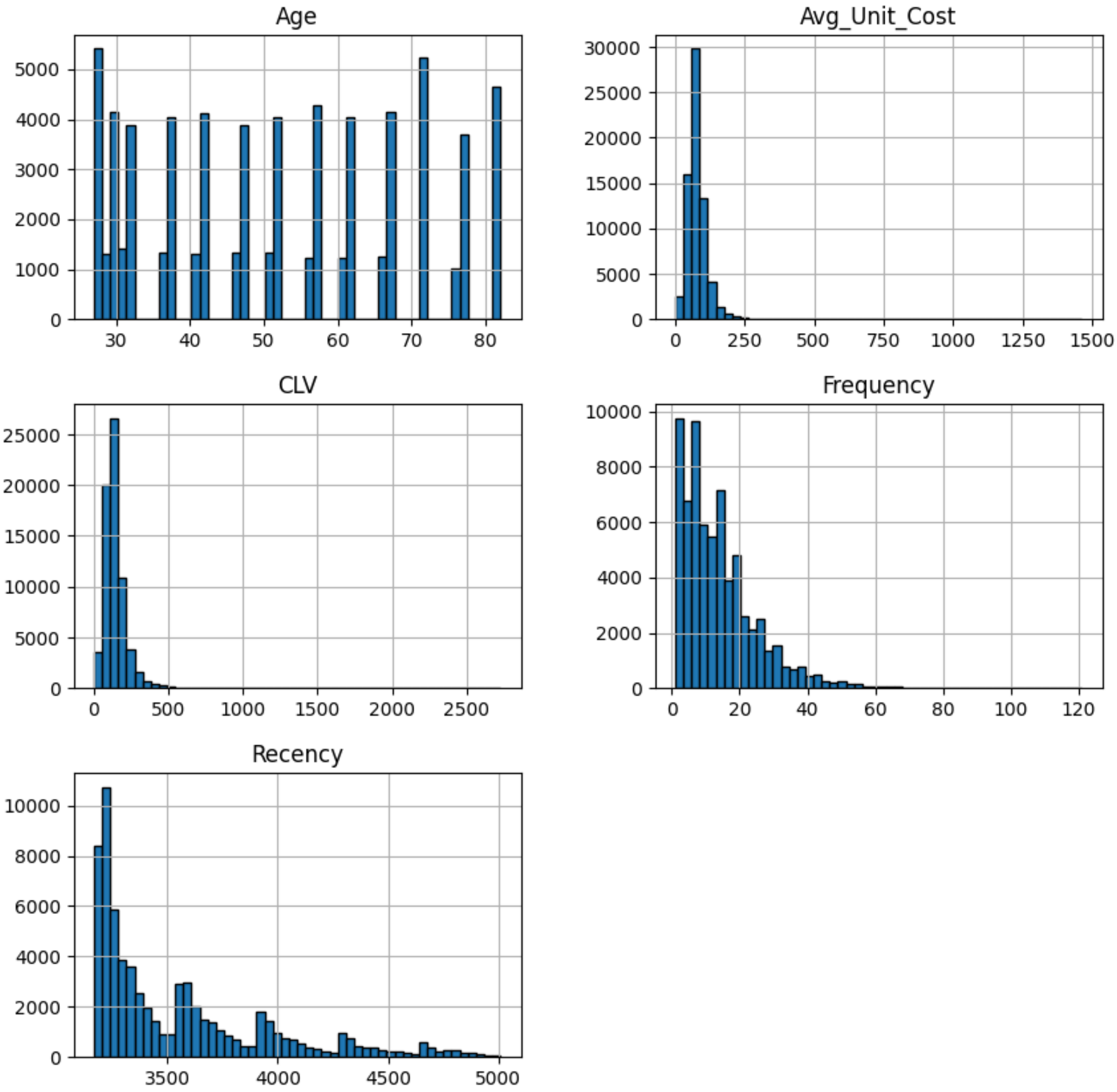
Perform EDA and Visualizations

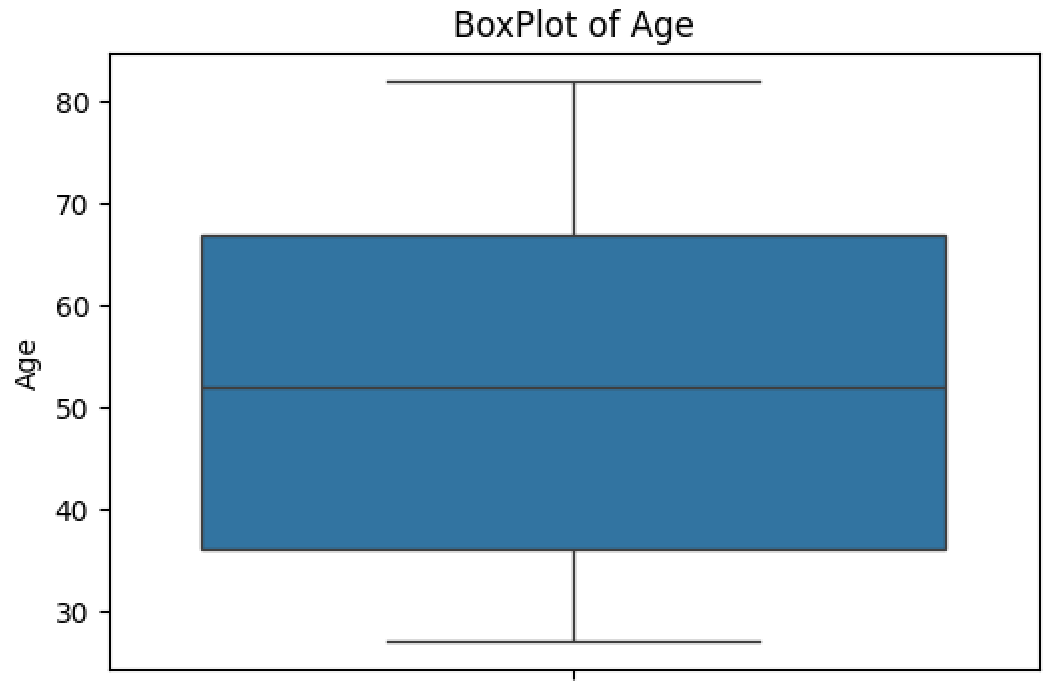
EDA was performed to understand data distribution and highlight anomalies:

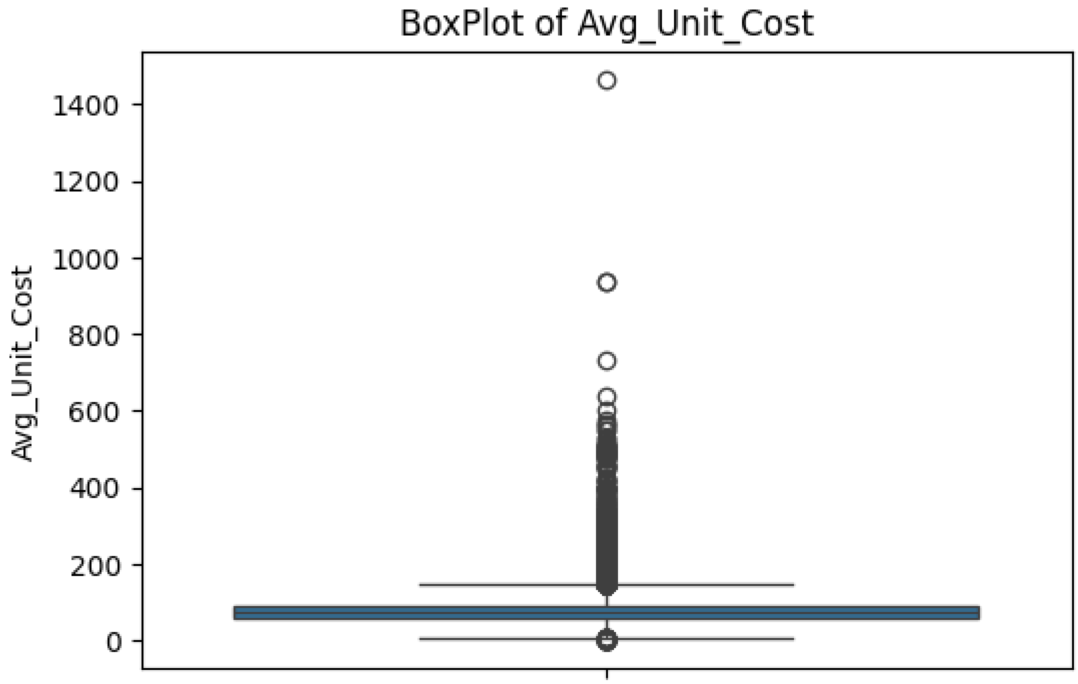
- **Descriptive Statistics**: Calculated mean, median, percentiles, and ranges for the five engineered features. These provided insights into the central tendencies and spread of the data.
- **Histograms**: Used to assess individual feature distributions. Frequency and recency, for instance, showed skewness typical of retail data, with most customers making fewer purchases recently and a small proportion purchasing frequently.
- **Boxplots**: Created for each feature to identify outliers and extreme values. Clusters with high dispersion (such as customer lifetime value) revealed significant differences in purchasing behavior.

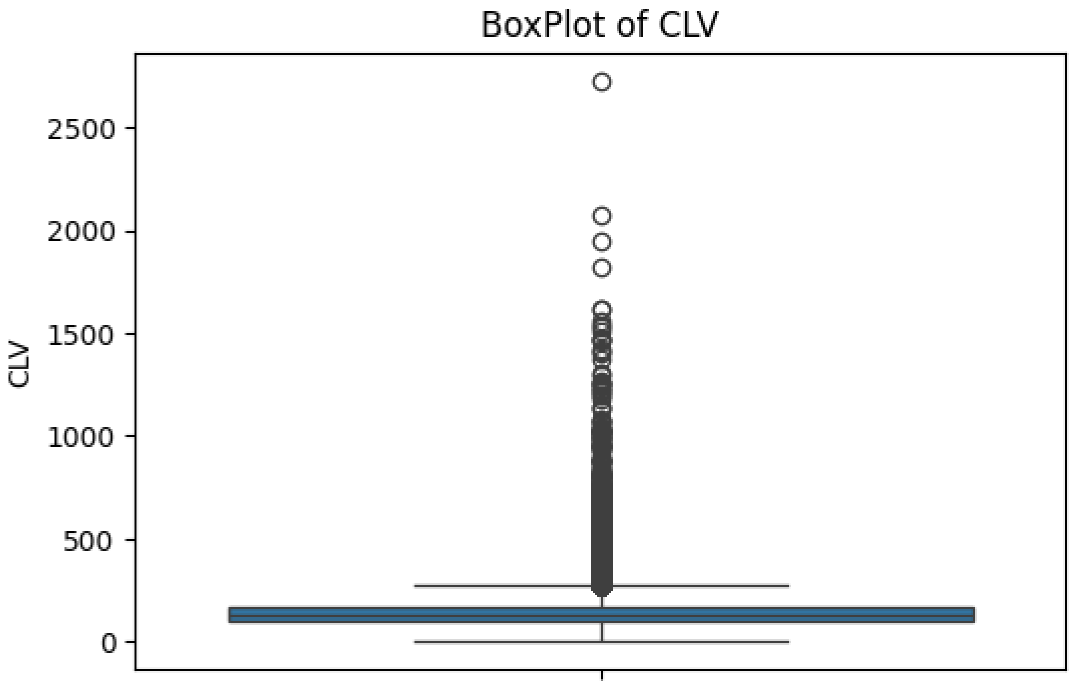
- **Identify Outliers:** Outlier detection using One-Class SVM identified **692** anomalies (nu=0.01 and gamma=0.5), highlighting irregular customer behaviors warranting separate consideration in segmentation.

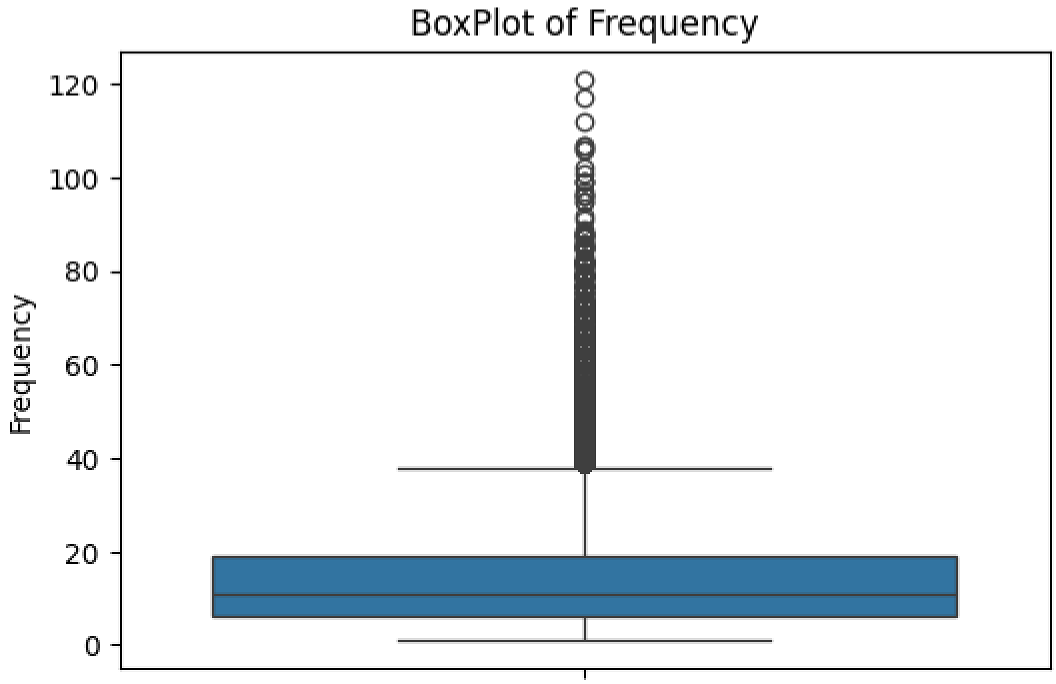
This stage provided a strong understanding of feature distributions and highlighted the presence of outliers, later confirmed in clustering analysis.

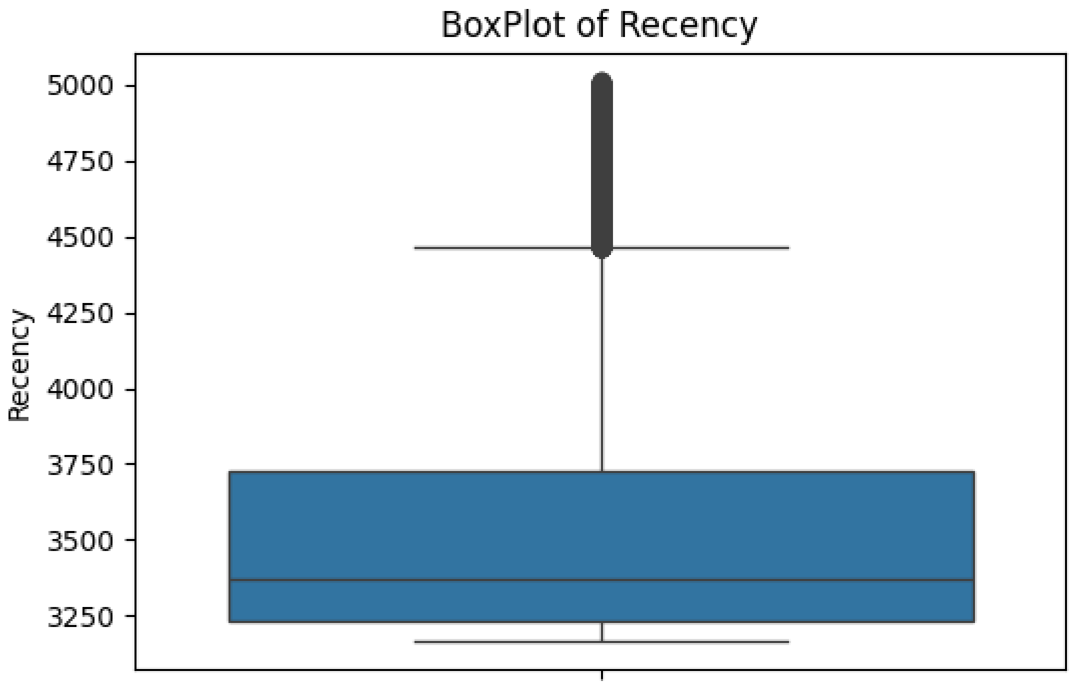



BoxPlot of Age
Age
80
70
60
50
40
30


BoxPlot of Avg_Unit_Cost
Avg_Unit_Cost
1400
1200
1000
800
600
400
200
0


BoxPlot of CLV
CLV
2500
2000
1500
1000
500
0


BoxPlot of Frequency
Frequency
120
100
80
60
40
20
0

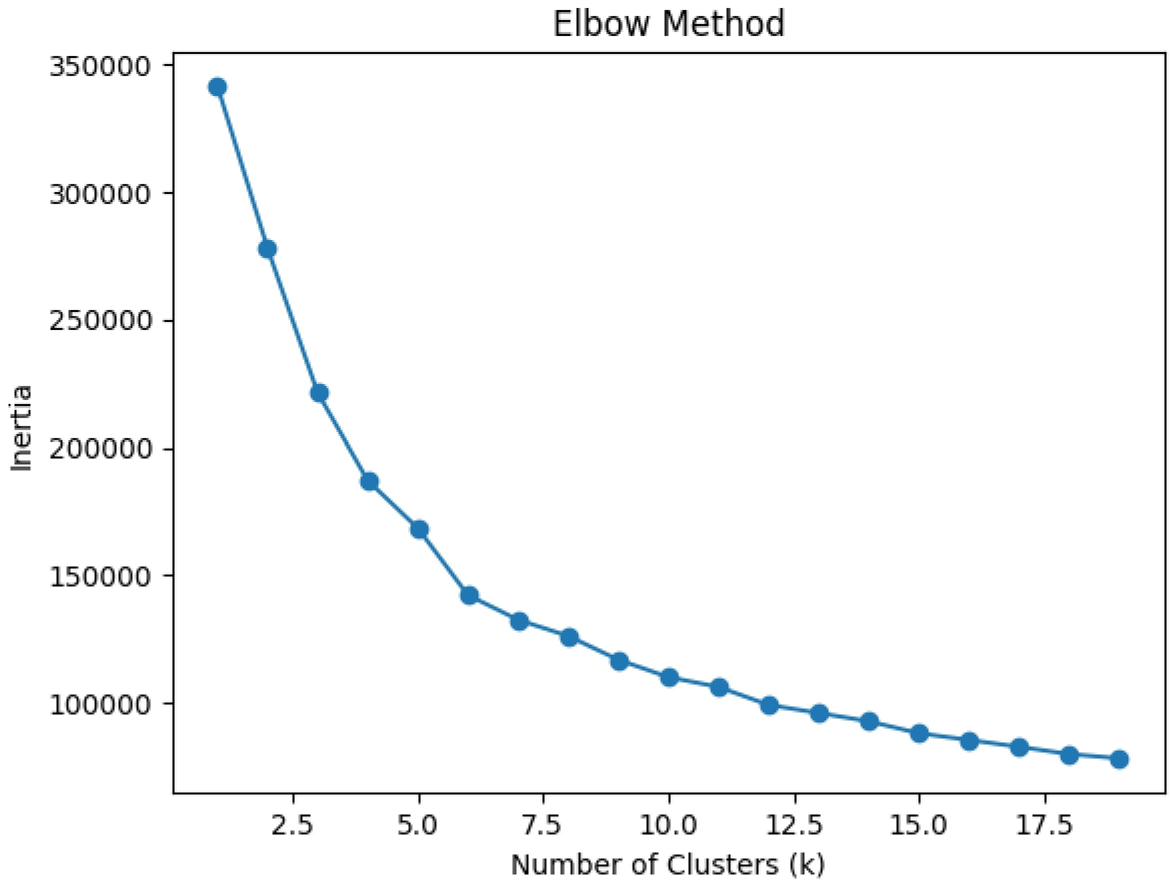

BoxPlot of Recency
Recency
5000
4750
4500
4250
4000
3750
3500
3250

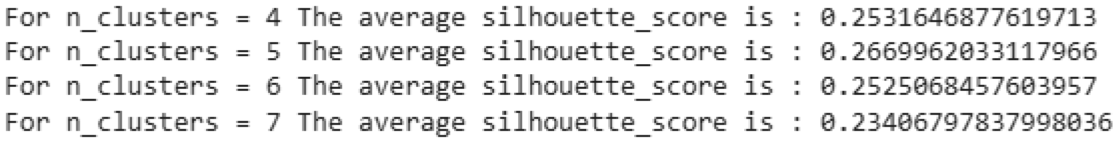
3. Identify Number of Clusters

The clustering approach combined **statistical methods** and **machine learning algorithms**:

Determining Optimal Clusters (k)

- The **Elbow method** was run over ranges 1–10, 1–15, and 1–20. In all cases, the “elbow” consistently appeared around **k = 5 or 6**.
- **Silhouette scores** were computed for k = 4, 5, 6, 7. Scores remained steady around **0.24**, confirming no clear separation but consistent clustering structure.
- Based on these methods, **5 clusters** were chosen as the optimal number.

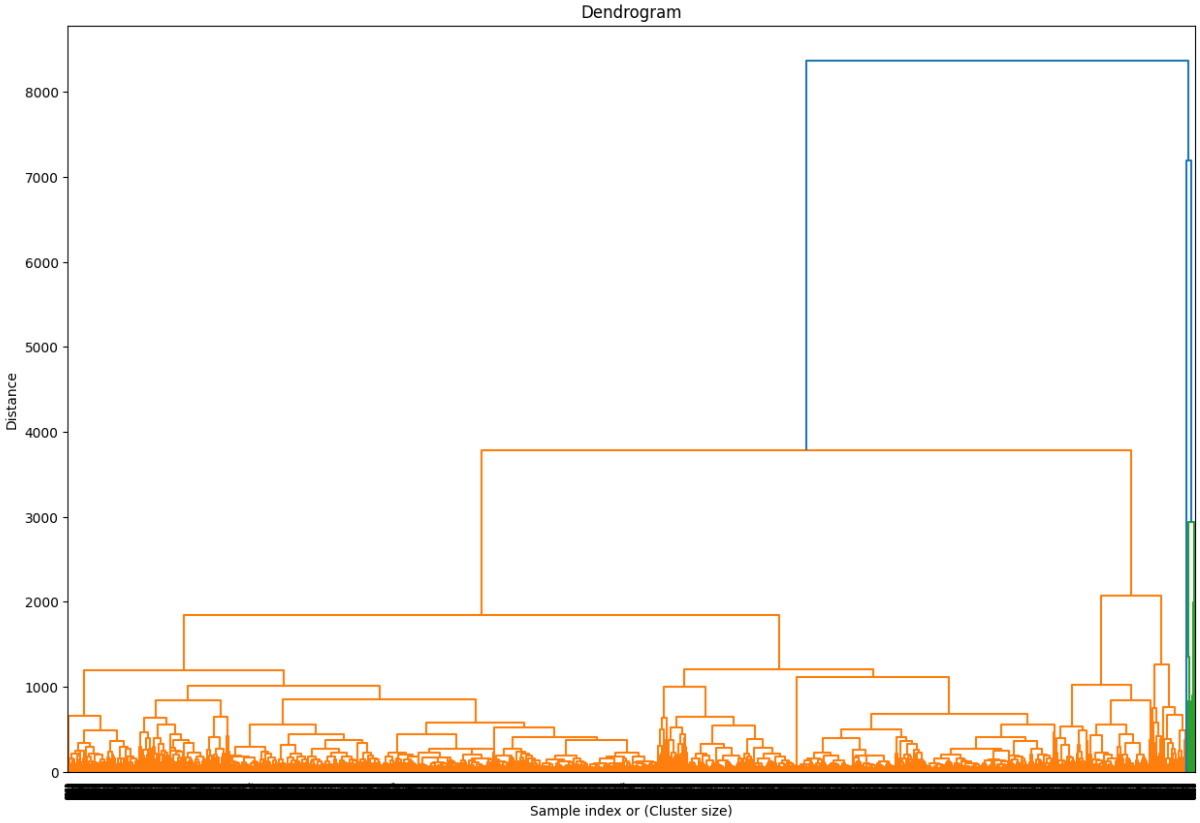




```
For n_clusters = 4 The average silhouette_score is : 0.2531646877619713
For n_clusters = 5 The average silhouette_score is : 0.2669962033117966
For n_clusters = 6 The average silhouette_score is : 0.2525068457603957
For n_clusters = 7 The average silhouette_score is : 0.23406797837998036
```

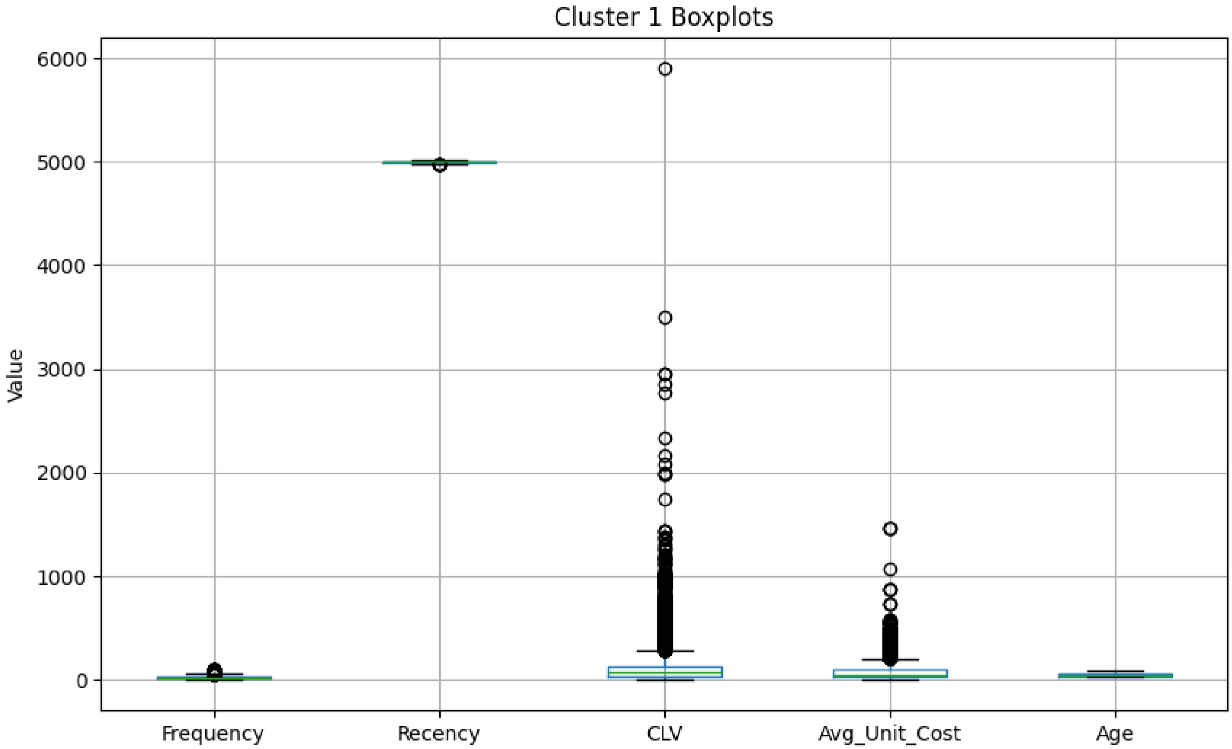
Clustering Methods Applied

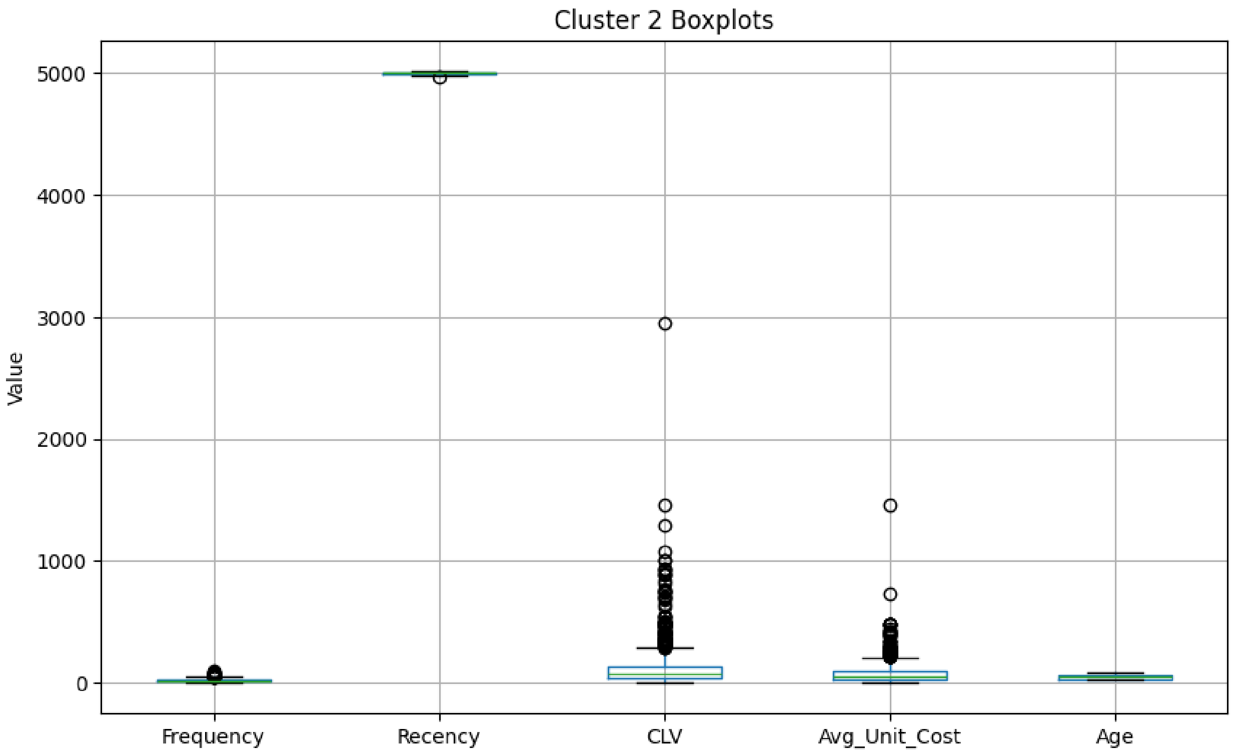
- **K-Means Clustering**: In the K-means clustering analysis, **5 clusters** consistently emerged as the optimal solution across elbow and silhouette methods, indicating stable customer segmentation.
- **Hierarchical Clustering & Dendrogram**: Performed with k = 5. While attempting to generate a dendrogram on the full dataset (68,500 rows), **the notebook repeatedly crashed due to computational limitations**. Therefore, a smaller representative sample of 10,000 rows was used. Even with this reduced dataset, the dendrogram clearly revealed distinct hierarchical relationships between customer groups, supporting the earlier K-means results.

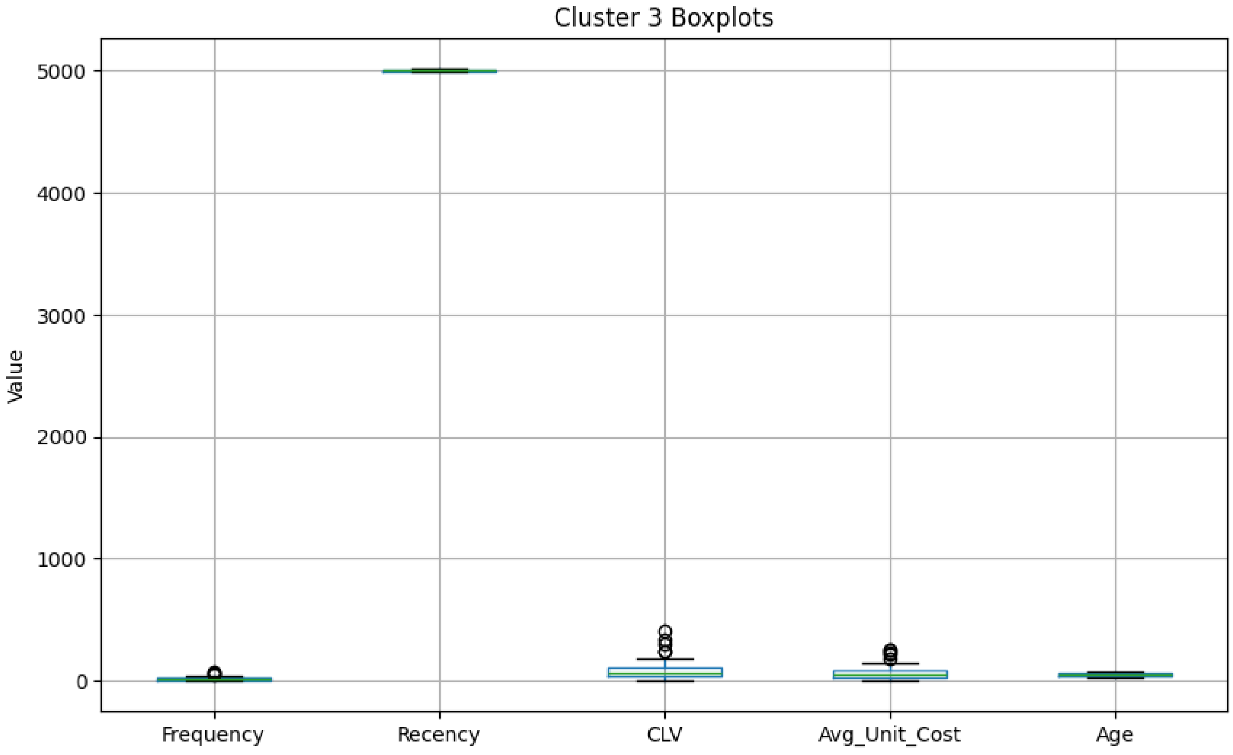


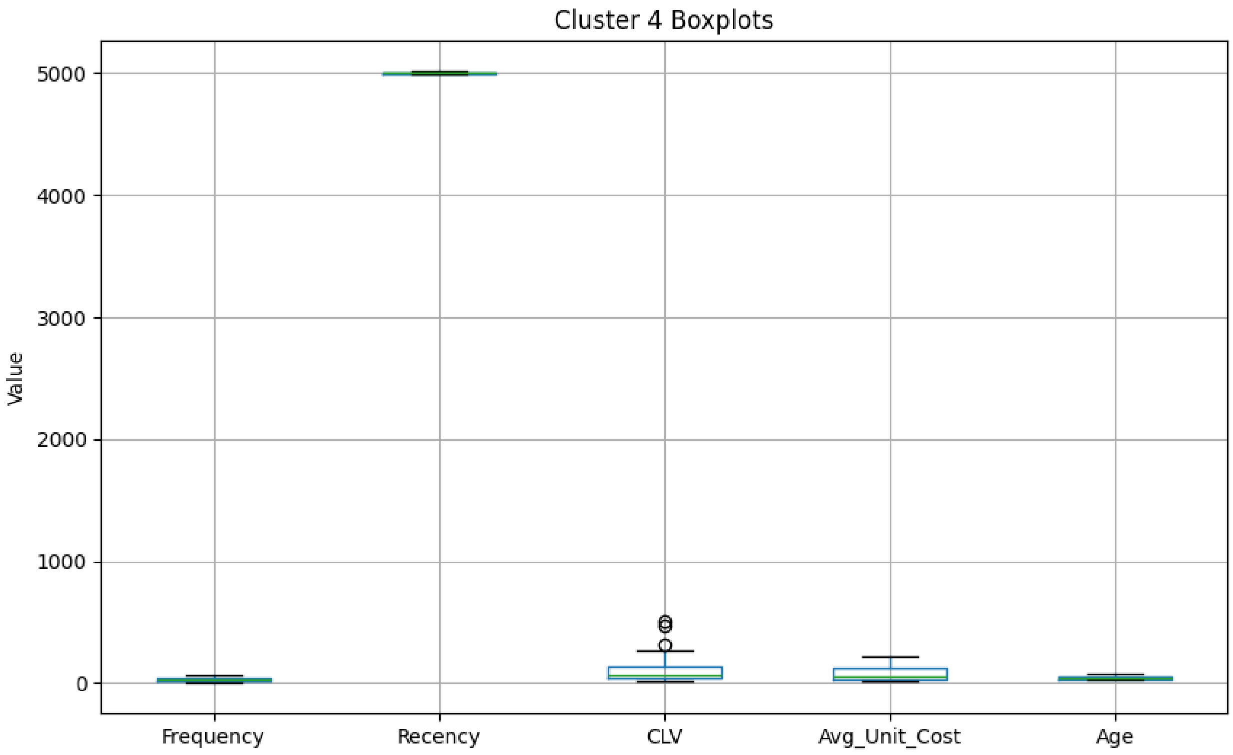
4. Customer Segmentation Results

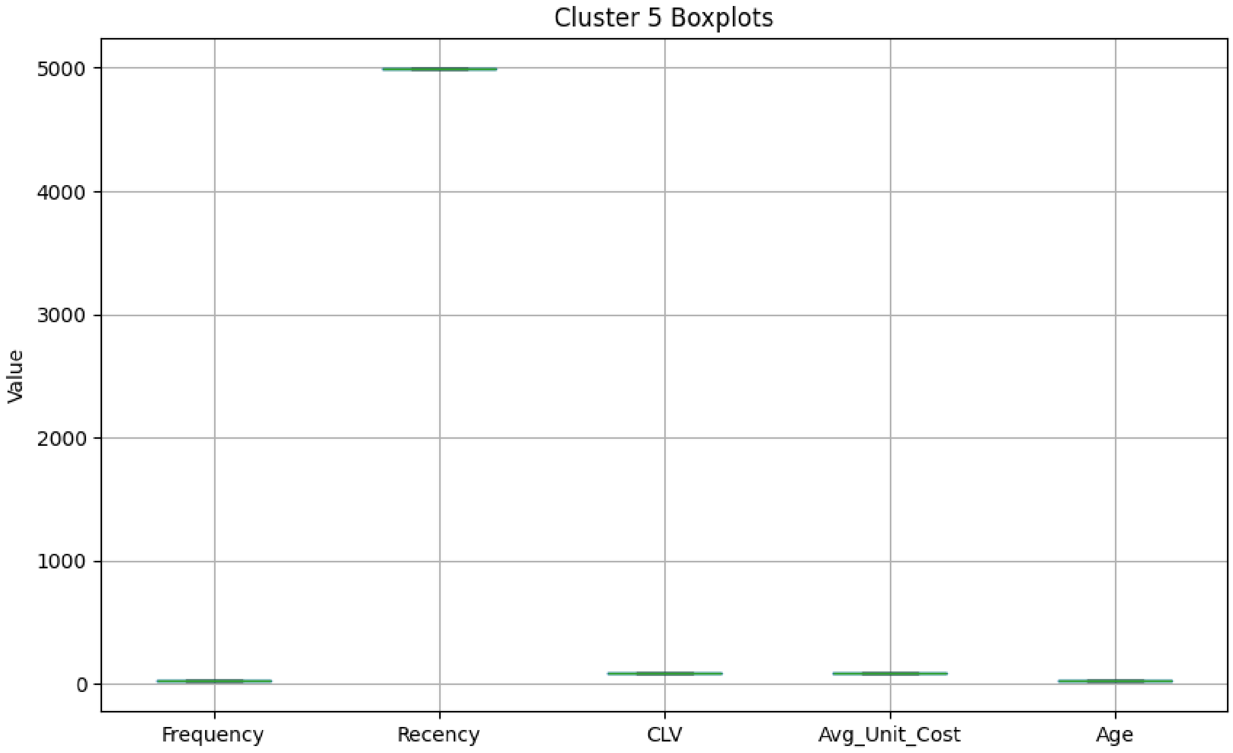
- **Cluster Analysis (Boxplots)**: When visualized, **Clusters 1 and 2** contained higher variability and more outliers, suggesting diverse purchasing behaviors and atypical spending patterns.
  Segment Profiles:
- Some clusters captured **high-CLV customers** representing loyalty and profitability.
- Others represented **low-frequency, low-spend customers**, potentially requiring targeted campaigns.
- Outlier-heavy clusters indicated customers with unusual order patterns, valuable for further business investigation.




Cluster 2 Boxplots
Value
5000
4000
3000
2000
1000
0
Frequency
Recency
CLV
Avg_Unit_Cost
Age


Cluster 3 Boxplots
Value
5000
4000
3000
2000
1000
0
Frequency
Recency
CLV
Avg_Unit_Cost
Age


Cluster 4 Boxplots
Value
5000
4000
3000
2000
1000
0
Frequency
Recency
CLV
Avg_Unit_Cost
Age

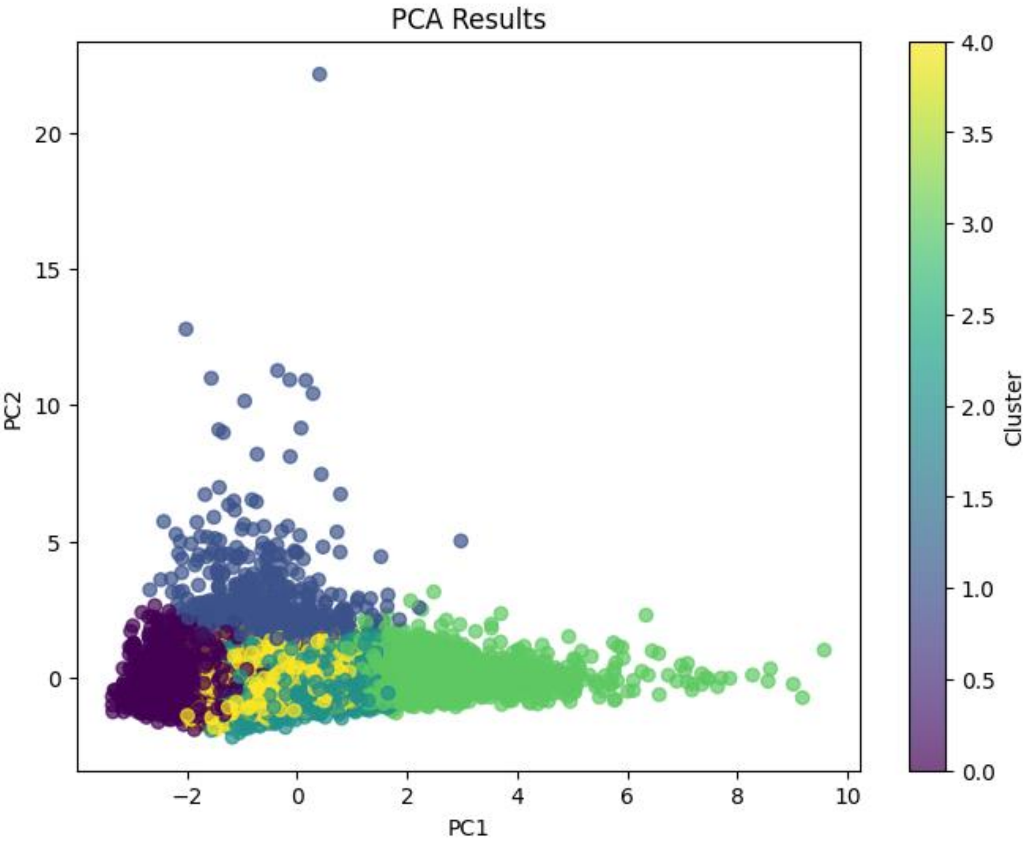

Cluster 5 Boxplots
Value
5000
4000
3000
2000
1000
0
Frequency
Recency
CLV
Avg_Unit_Cost
Age

5. Dimensionality Reduction (PCA & t-SNE)

PCA (Principal Component Analysis)

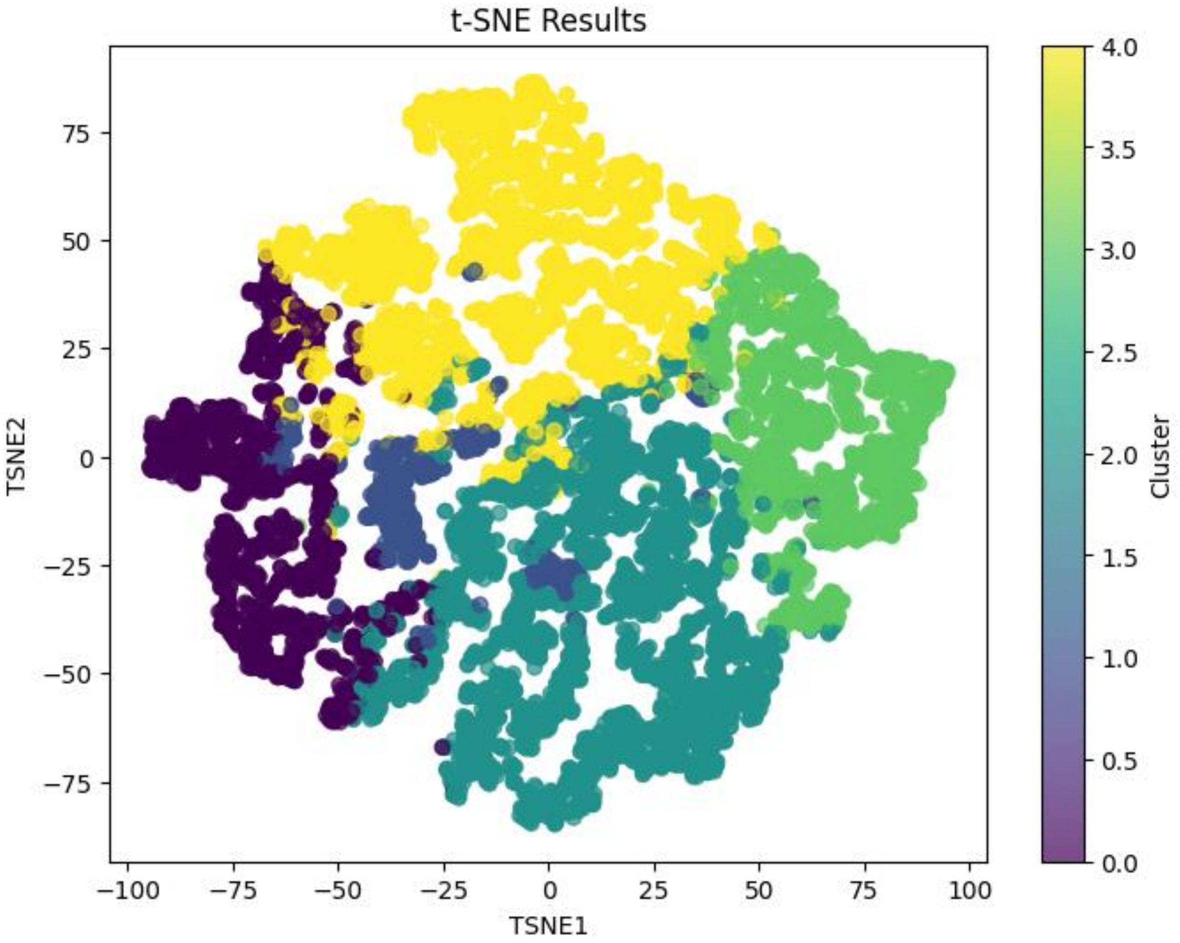
Reduced data into two principal components.

- The PCA scatterplot shows clusters somewhat overlapping, which suggests that while PCA is good for dimensionality reduction, it does not separate the customer groups.
- This overlap indicates that customers may share similarities across multiple features, and their segmentation boundaries are not strictly linear.
- The spread of points along **PC2** shows variation in customer behavior.
- **Cluster 3 (Green)** appears the most widespread along **PC1**, which may represent a diverse group of customers with varying purchase behaviors.
- The dense overlap around the origin suggests many customers share average behavior, with fewer customers showing extreme values (those spread farther on the axes).



t-SNE (t-distributed Stochastic Neighbor Embedding)

- The t-SNE plot provides much **clearer separation of clusters** compared to PCA, highlighting that t-SNE captures the non-linear relationships better.
- Clusters (especially yellow and teal) form more distinct, non-overlapping groups, which may indicate strong differences in their buying behavior or demographics.
- The circular pattern in t-SNE suggests the presence of dense "core" groups.
- **Clusters 2 (Teal) and 4 (Yellow)** appear the largest and most cohesive, meaning these groups of customers may be dominant and potentially most valuable for targeted marketing.



6. Conclusion

This study demonstrated the effectiveness of clustering in customer segmentation:

- **Insights gained**:
  - The dataset was consolidated into **68,300 unique customers**, with five engineered features providing strong explanatory power for segmentation.
  - Outliers concentrated in certain clusters suggested niche or atypical customer groups.
- **Best clustering technique**:
  - Both elbow and silhouette methods supported **five clusters** as optimal.
  - **K-means clustering** provided clearer, scalable results compared to hierarchical clustering, which was computationally intensive on larger subsets.
- **On visualizations**:
  - PCA is helpful for quick dimensionality reduction and variance explanation, but **t-SNE gives more meaningful, interpretable clusters** for customer segmentation.
  - Quantitative methods (silhouette, elbow, clustering performance) are more reliable indicators than 2D visuals alone.