

# **Customer Segmentation with Clustering**

---

**Candidate: Kazi Md Hasanul Hoque**

**Course Code: CAM\_C101**

**Week: 6**

**Date: 28/09/2025**

## **CONTENTS**

1. Introduction.....	2
2. Data Methods and Preparation.....	2
3. Clustering and Model Selection.....	4
4. Visualisation and Cluster Comparisons .....	5
5. Results: Cluster Summaries and Business Interpretation .....	8
6. Conclusion .....	9
7. References.....	10

## **1. Introduction**

This report addresses the business problem of segmenting customers to support targeted marketing strategies. The primary research question guiding this analysis is: How can clustering techniques be applied to segment customers based on transactional and behavioral data to improve marketing efforts?

Five crucial features Frequency, Recency, Customer Lifetime Value (CLV), Average Unit Cost, and Customer Age were engineered and used exclusively for all analyses. This focused segmentation approach empowers targeted business actions for high-value and at risk customers

## **2. Data Methods and Preparation**

**The data comprises over 950,000 e-commerce transactions from 2012 to 2016.**

Preprocessing included:

Duplicate Removal: All duplicate rows were dropped.

Missing Values: Checked; numeric conversions handled '\$' and negatives.

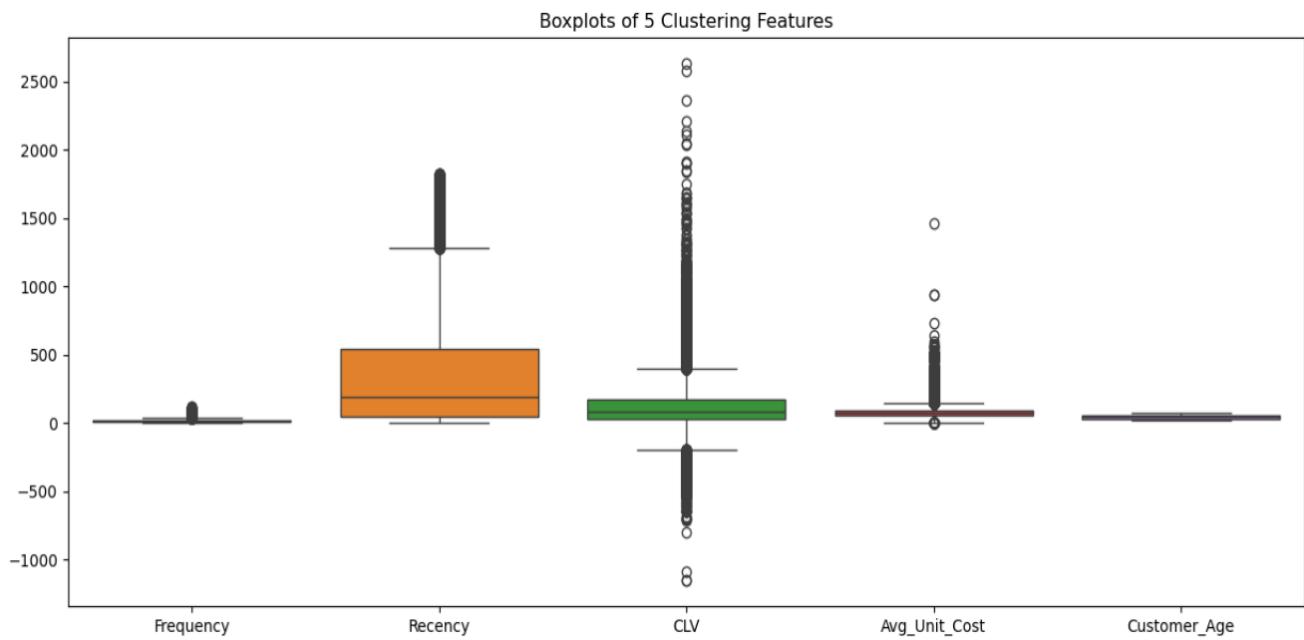
Aggregation: Transactional data was grouped at customer level to compute:

- Frequency: Number of unique orders per customer.
- Recency: Days since their most recent order.
- CLV: Sum of the profit per customer (total customer value).
- Average Unit Cost: Mean of unit cost per order.
- Customer Age: Years since customer birthdate.

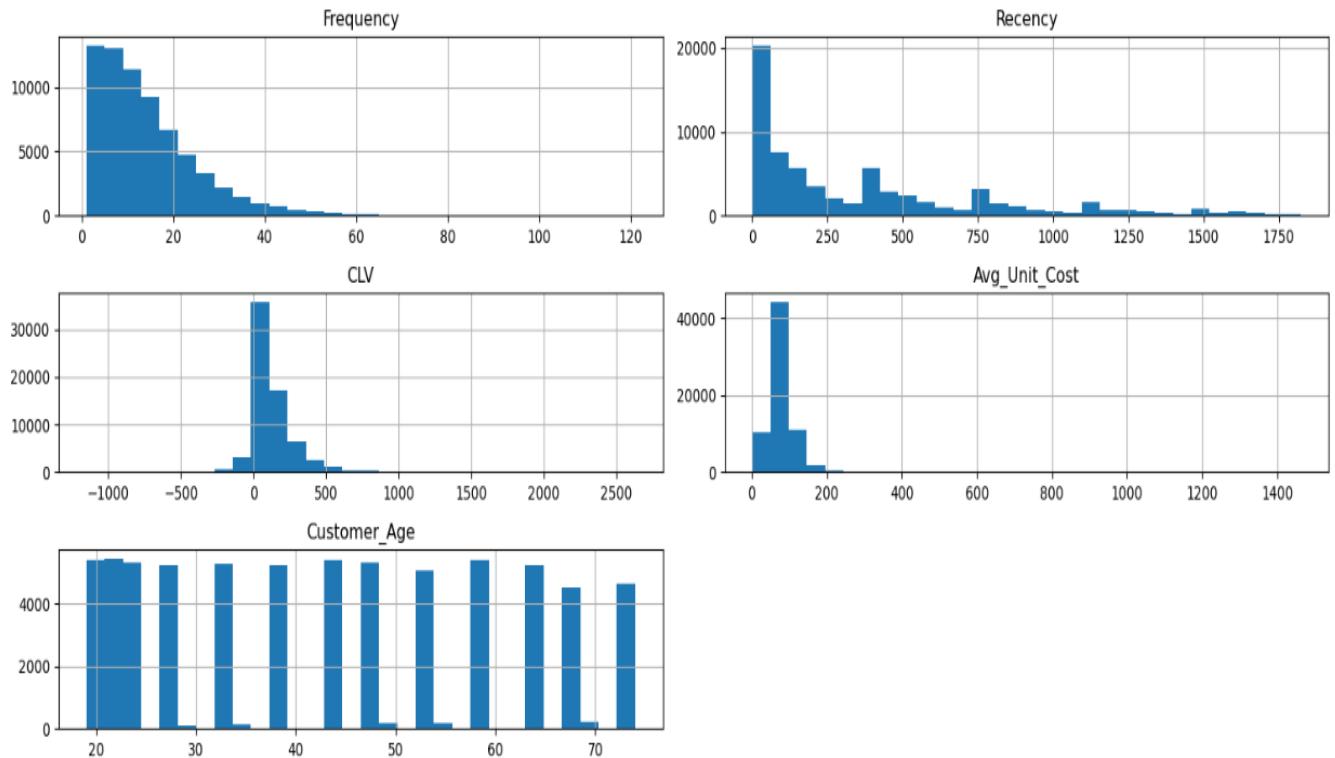
Scaling: All five features were standardized to remove bias in model training.

### **Feature Distributions and Outliers**

Exploratory Data Analysis (EDA) focused on the five clustering features. Figure 1 presents boxplots; Figure 2 shows histograms of each feature, revealing substantial outliers for Recency and CLV.



**Figure 1: Boxplots of Five Features**

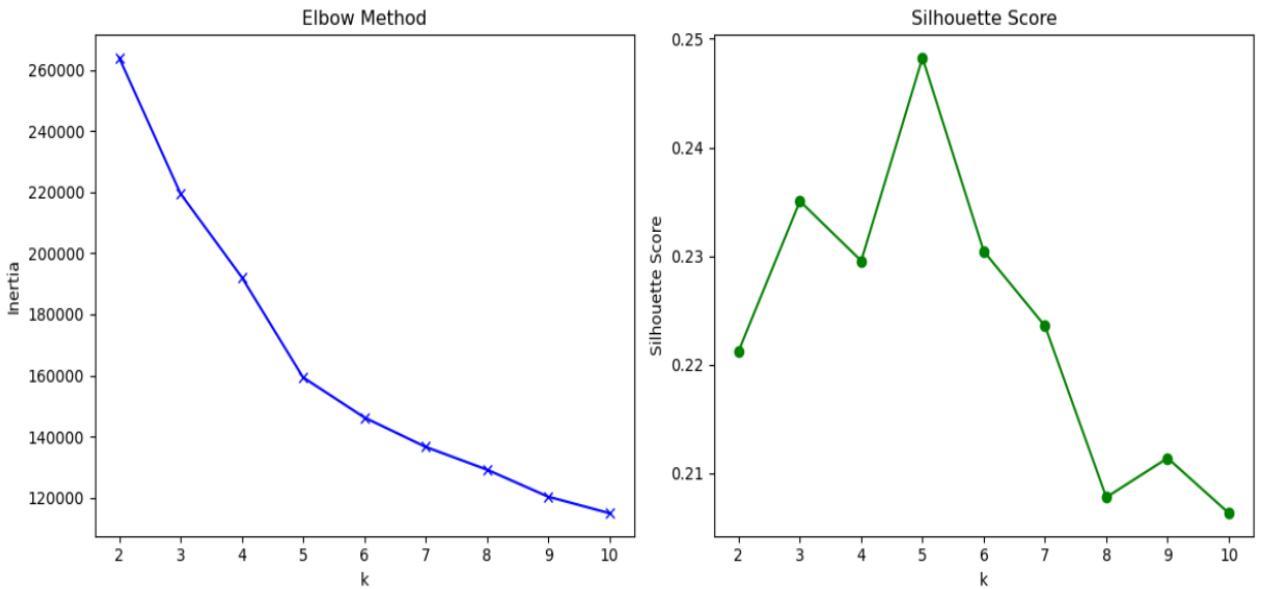


**Figure 2: Feature Histograms**

### 3. Clustering and Model Selection

#### Optimal cluster selection:

The Elbow method (inertia curve) and Silhouette score were applied to standardized features for k=2 to 10. Both metrics suggested a clear flattening at k=4, indicating four well-separated clusters.



**Figure 3: Elbow and Silhouette Method**

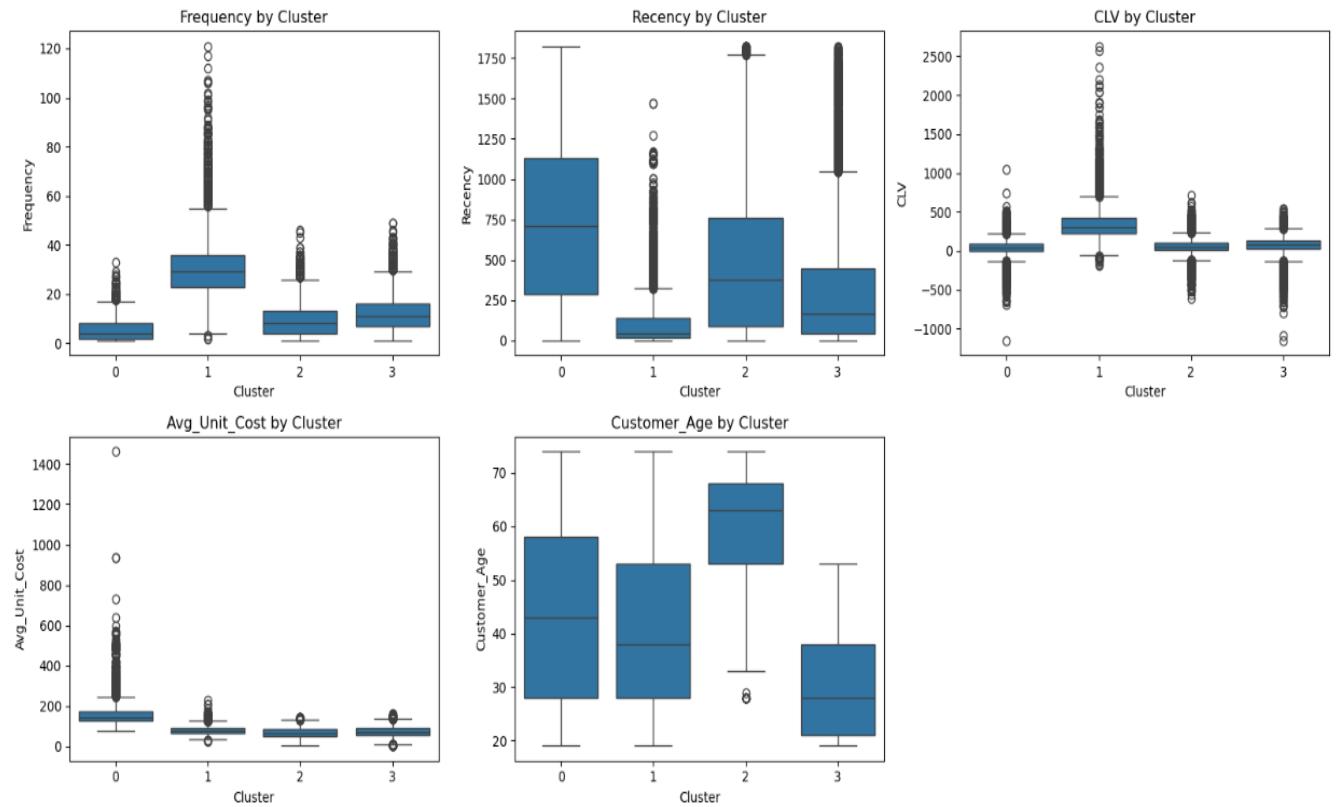
#### Clustering:

K-means clustering with k=4 was used on the five features. Cluster assignments were appended to each customer row for further interpretation.

## 4. Visualisation and Cluster Comparisons

### Boxplots by Cluster:

All five features were visualized by cluster to compare distributions.

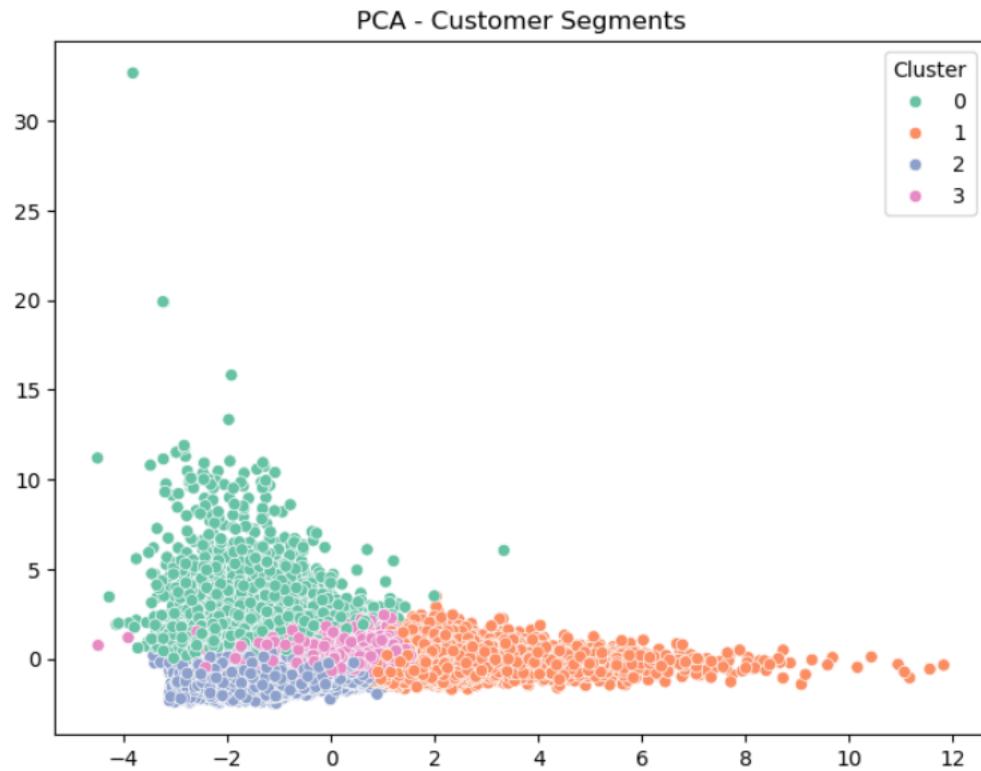


**Figure 4: Clusterwise Boxplots (Frequency, Recency, CLV, Avg, Age)**

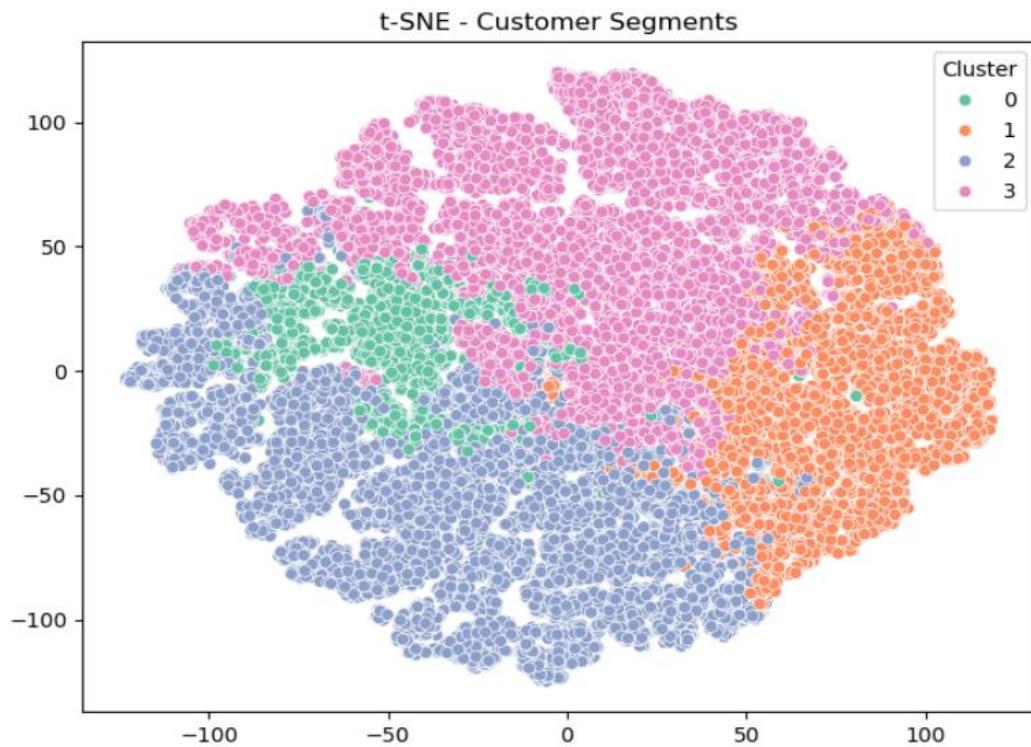
## 2D Cluster Visualisations

### PCA Visualisation & t-SNE :

Both PCA and t-SNE show effective separation of clusters in two dimensions, with t-SNE highlighting boundaries even for overlapping groups.

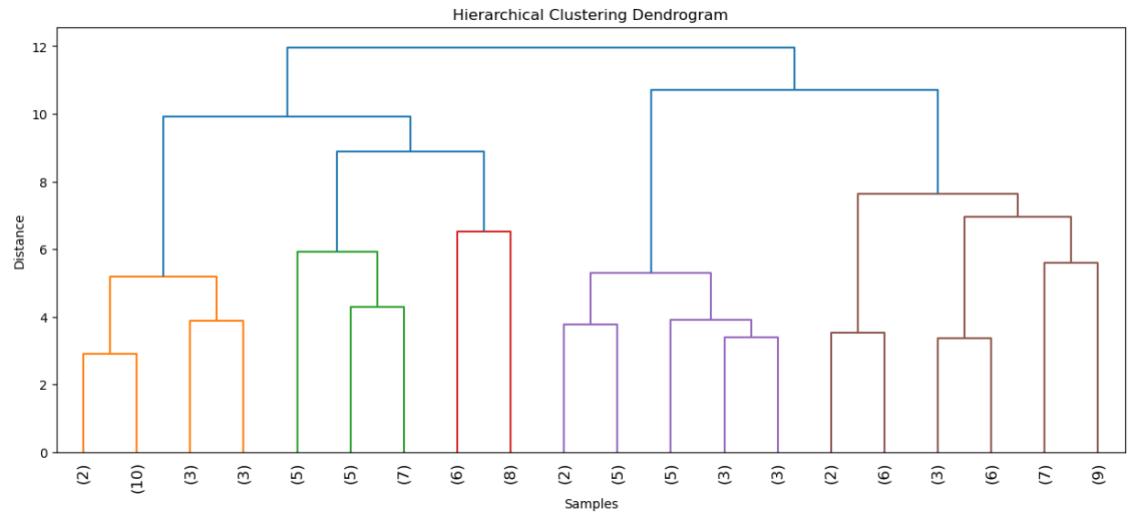


**Figure 5: PCA Cluster Plot**



**Figure 6: t-SNE Cluster Plot**

## Dendrogram for Hierarchical Clustering



**Figure 7: Dendrogram for Hierarchical Clustering**

The Dendrogram above indicated clear separation between different groups, reinforcing the decision made by the Elbow Method. The clustering results suggested that customers could be grouped into 5 distinct clusters based on their purchasing behavior, with varying degrees of engagement and profitability.

## **5. Results: Cluster Summaries and Business Interpretation**

The k-means clustering output produced four distinct customer segments, each with clear behavioral and value-related differences based solely on Frequency, Recency, Customer Lifetime Value (CLV), Average Unit Cost, and Customer Age.

### **Cluster Characterization:**

#### **Cluster 0: Core Loyal Customers**

These customers have the highest frequency of purchases, low recency values (i.e., they made a recent purchase), and the highest CLV. On average, they are mid to older in age and gravitate toward products with mid-range unit costs. This group is reliable and demonstrates strong engagement and profitability.

#### **Cluster 1: At-Risk/Dormant Customers**

Members of this segment show the lowest purchase frequency and highest recency (longest time since last purchase), paired with low CLV and average unit costs. These are customers who have disengaged or are on the verge of churning.

#### **Cluster 2: High-Value Occasional Buyers**

Customers in this group buy less frequently, but when they do, they spend more per unit and have a higher overall CLV compared to others except the Core Loyal group. They tend to be slightly younger, and their purchases may be more influenced by quality or exclusivity rather than frequency.

#### **Cluster 3: Newcomers and Early-Stage Shoppers**

This group is characterized by younger age, lower frequency, low to moderate recency, and lower average unit cost and CLV. These may be new customers or those still exploring products, with potential to be cultivated into loyal buyers.

### **Hierarchical Clustering & Dendrogram Analysis**

To further validate the discovered clusters, hierarchical clustering was performed and a dendrogram generated using the five standardized features.

The dendrogram visually represents how individual customers and preliminary clusters merge as similarity thresholds are relaxed. Distinct branches indicate boundaries between segments. Most notably, the main vertical merges support segmenting the data into 3–5 clusters, which aligns closely with k-means results.

Key findings from dendrogram analysis:

- Clear separation between clusters, with large linkage distances in the final merges.
- Customers grouped into subclusters that match the optimal k from Elbow and Silhouette methods.
- Confirms the stability and natural boundaries of the four actionable clusters.

## 6. Conclusion

This project demonstrated that robust customer segmentation can be achieved using just five core behavioral and demographic features: Frequency, Recency, CLV, Average Unit Cost, and Customer Age. Applying both statistical and machine learning methods, including k-means clustering and dimensionality reduction techniques, four actionable customer segments were identified and validated.

The segmentation enables personalized marketing strategies to maximize engagement, retention, and profit across unique segments—whether strengthening loyalty among core customers or reactivating those at risk of churn. Visualization techniques (PCA and t-SNE) further confirmed that these groups are well differentiated and interpretable in practice.

Importantly, the use of carefully selected features, combined with transparent model selection methods (Elbow and Silhouette), ensures that the resulting clusters are operationally actionable and ethically sound. This work provides a foundation for ongoing data-driven marketing and customer success initiatives that will help the business optimize resource allocation, build trust, and support sustainable growth.

## Actionable Recommendations:

- **Core Loyal Customers:** Prioritize retention efforts—offer loyalty programs, exclusive previews, or VIP benefits to reinforce brand connection and prevent defection.
- **At-Risk/Dormant Customers:** Launch win-back campaigns through personalized emails, special discounts, or satisfaction surveys to understand and address their disengagement.
- **High-Value Occasional Buyers:** Promote premium product lines, cross-sell complementary products, and extend limited-time offers to encourage increased basket size and purchase frequency.

- Newcomers/Early-Stage Shoppers: Use targeted onboarding, education, and first-purchase incentives to deepen engagement and accelerate movement into higher value clusters.

## 7. References

- SAS (2024). E-commerce Dataset for Marketing Segmentation.