TITLE: **Customer Segmentation for E-Commerce Marketing and Customer Retention Optimisation**

NAME: **Joe Achira**

DATE: **08/12/2024**

Customer Segmentation Analysis Report

Overview of Approach

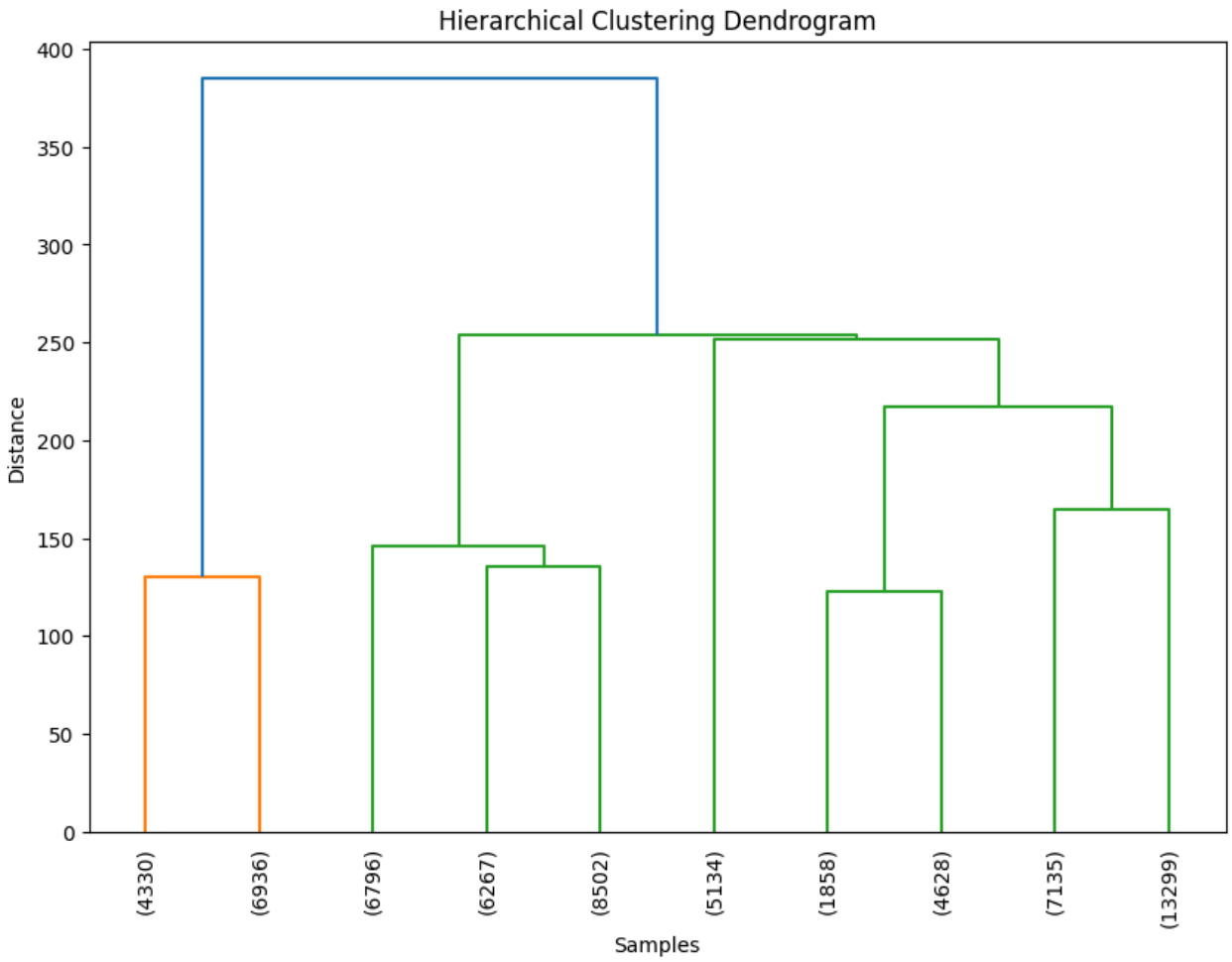
This project focused on developing a robust customer segmentation framework to assist an e-commerce company in optimising its marketing and customer retention strategies. By analysing customer purchasing behaviours, we aimed to provide actionable insights that align with the company's goal of improving its customer-centric approach and operational efficiency.

We employed feature engineering, clustering, and dimensionality reduction techniques to segment the customer base effectively. The process included addressing outliers, selecting key features, and validating clustering methods for optimal performance. Visual and quantitative analyses were used to derive meaningful insights from the segmentation.

Description of Analysis

1. **Data Preprocessing and Feature Engineering**
   The application of data pre-processing techniques allowed the whittling down of the dataset from 961558 rows by 20 columns to 68,300 unique customers with no missing values. Five key features were selected for analysis and clustering:
   - **Frequency**: Number of orders per customer.
   - **Recency**: Days since the most recent order.
   - **CLV (Customer Lifetime Value)**: Total revenue per customer.
   - **Avg Unit Cost**: Average cost per unit purchased.
   - **Age**: Customer age derived from birthdate.
2. These features were aggregated to reflect customer-level data:
   - **Frequency, Avg Unit Cost, and Age** were averaged.
   - **Recency** was minimised to capture the most recent orders.
   - **CLV** was summed to identify high-value customers.
3. Outliers were identified using the Isolation Forest algorithm, removing 3,415 rows (5%) to improve model performance. The remaining data was scaled using standardisation to ensure equal weight across features.
4. **Determining the Optimal Number of Clusters (k)**
   To identify the optimal number of clusters, we applied the following methods:
   - **Elbow Method**: The plot indicated a point of inflection at k=5.
   - **Silhouette Method**: Evaluated the quality of clusters for various values of k.
     - k=5 achieved one of the higher silhouette scores (0.255), suggesting well-defined cluster separation compared to other values.
   - **Hierarchical Clustering**: A dendrogram confirmed that cutting at k=5 preserved meaningful cluster structure.
5. These methods collectively validated k=5 as the optimal number of clusters, balancing separation and cohesion.

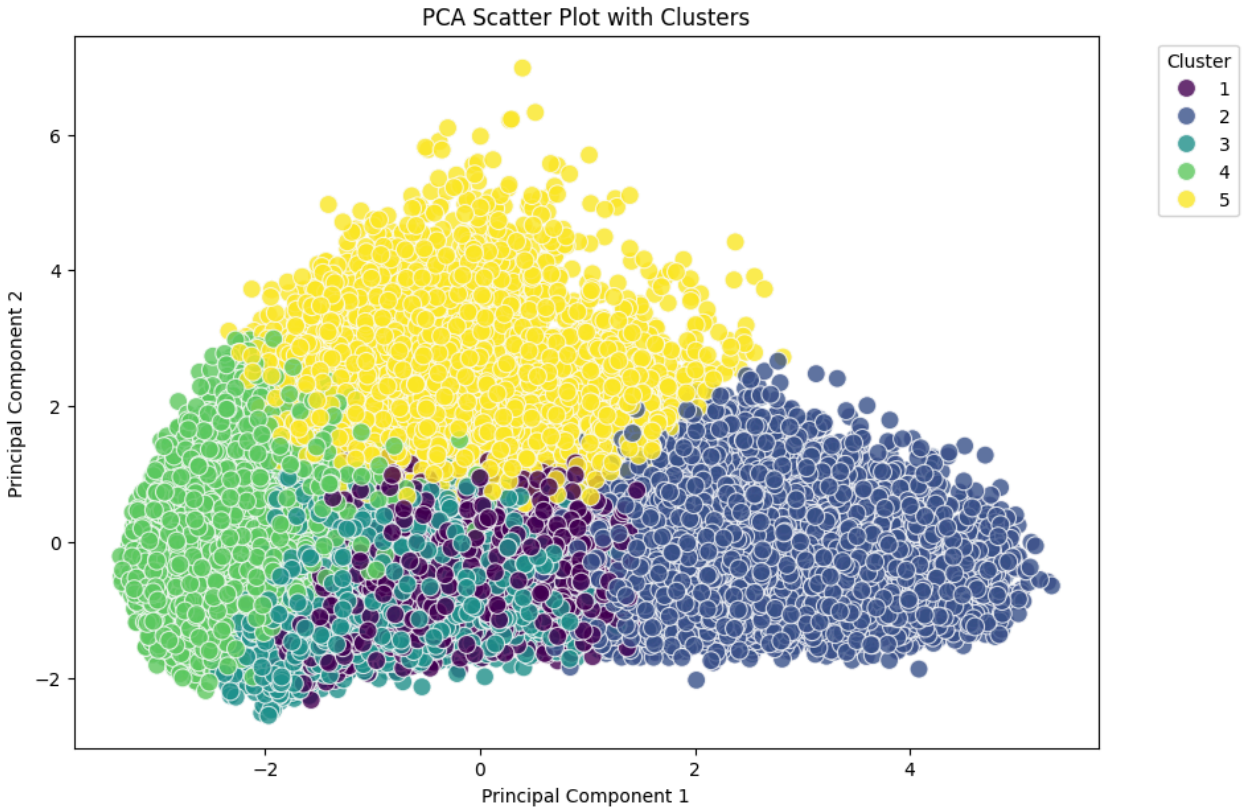
6. **Clustering and Segmentation**
   Using **K-means clustering** with k=5, we segmented the customers into five groups. **Hierarchical clustering** using centroid/ward linkage further confirmed the robustness of this choice.



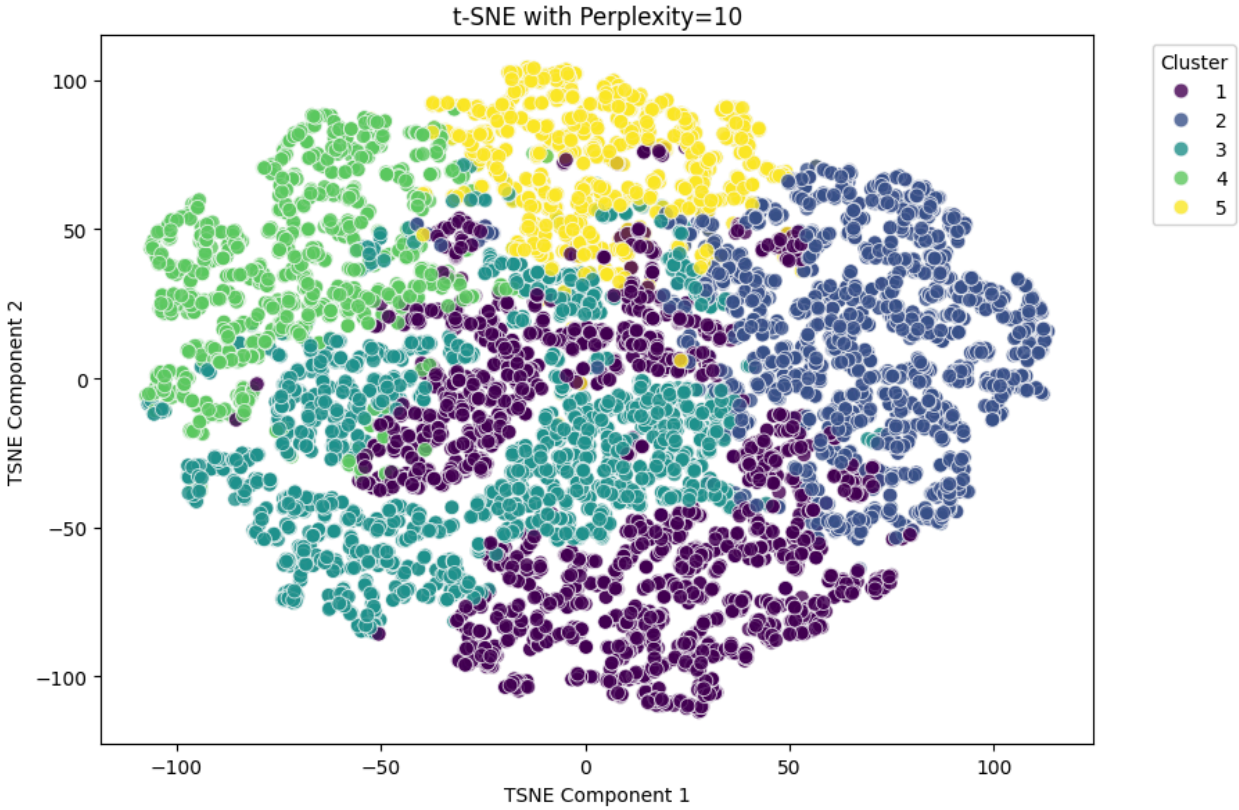
   The clusters were analysed using feature distributions across each group:

   - **Box Plots**: Provided a detailed understanding of the variability in `Frequency`, `Recency`, `CLV`, `Avg Unit Cost`, and `Age`. Clusters displayed distinct patterns, suggesting meaningful segmentation.

7. **Dimensionality Reduction for Visualisation**
   To visualise the clusters effectively:

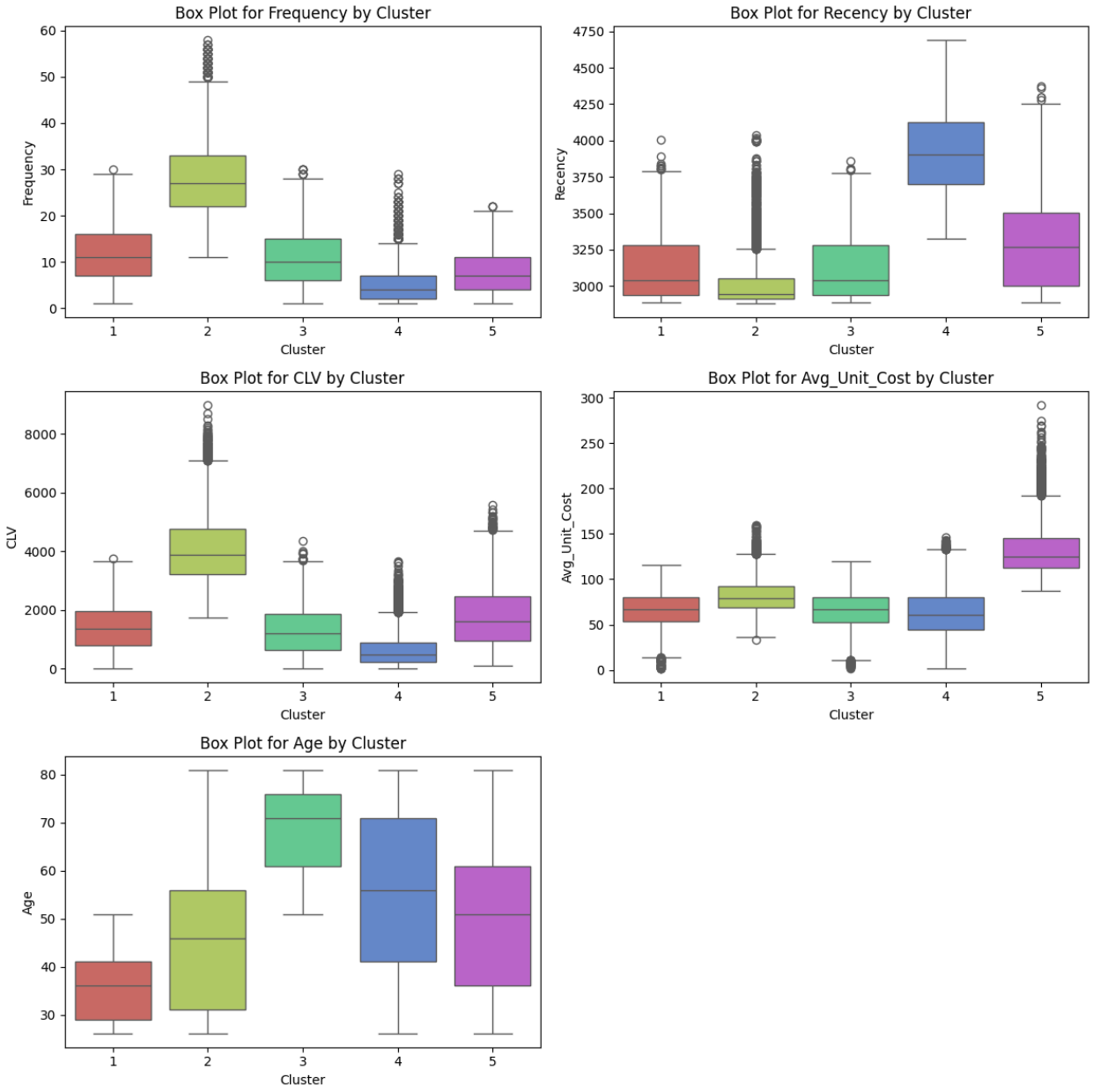


- **PCA (Principal Component Analysis)**: Reduced the high-dimensional data to two components, showing clear separation for Clusters 2, 4, and 5. Clusters 1 and 3 were partially overlapping, indicating shared characteristics.
- **t-SNE (t-Distributed Stochastic Neighbor Embedding)**: Explored cluster distributions at different perplexities (10, 25, 35).
  - Clusters 2, 4, and 5 remained distinct across perplexities.
  - Clusters 1 and 3 showed partial overlap, consistent with PCA findings.



8. These visualisations confirmed the clustering results and highlighted areas where clusters were well-defined or overlapped.

Insights Identified



1. **Cluster Analysis**
    - **Cluster 2 (High-Value Customers)**: This group exhibits the highest order frequency, most recent purchases, and the greatest lifetime value. Customers are younger to middle-aged, representing the most valuable segment for retention and upselling campaigns.

    - **Cluster 4 (Inactive Customers)**: With the lowest frequency, oldest recency, and minimal lifetime value, these middle-aged customers likely represent lapsed buyers. Targeted reactivation efforts may yield results.
    - **Cluster 3 (Traditional Customers)**: Older customers with moderate frequency and below-average revenue. This segment may represent long-standing customers who engage periodically. Loyalty initiatives could strengthen relationships.
    - **Cluster 1 (Emerging Customers)**: Younger customers with average frequency and lower revenue. As a growth segment, this group could benefit from targeted marketing campaigns.
    - **Cluster 5 (Highly Profitable Customers)**: Middle-aged customers with low frequency but the highest average unit cost. Premium product offers and exclusive deals may maximise profitability here.
2. **Feature-Level Insights**
    - **Recency**: Cluster 4 showed the highest median value, indicating older purchases, while Cluster 2 displayed the lowest, suggesting active buyers.
    - **Frequency & CLV**: Clusters 2 and 5 had the most significant impact, with Cluster 2 excelling in frequency and CLV, whereas Cluster 5 achieved profitability through high unit costs.
    - **Age**: Cluster 3 represented the oldest customers (median 70), while Cluster 1 had the youngest, reflecting varied life stages and likely purchasing behaviours.
3. **Dimensionality Reduction Observations**
    - **Distinct Clusters**: Clusters 2, 4, and 5 formed well-separated groups in both PCA and t-SNE plots, highlighting their unique characteristics.
    - **Overlap**: Clusters 1 and 3 were less distinct, suggesting shared traits or transitional customer behaviours.

Summary of Results for Determining k

The combination of elbow, silhouette, and hierarchical clustering methods identified k=5 as the optimal number of clusters. This choice was validated through visualisation (PCA and t-SNE) and subsequent analysis, which showed meaningful and actionable segmentation.

Recommendations

1. **Focus on High-Value and Highly Profitable Customers**
    - **Cluster 2**: Implement loyalty programs and personalised campaigns to retain and expand this segment.
    - **Cluster 5**: Offer premium and exclusive products to maximise profitability.
2. **Reactivation and Growth Strategies**

- **Cluster 4**: Engage inactive customers through targeted discounts or reactivation campaigns.
- **Cluster 1**: Leverage digital marketing and referral programs to cultivate emerging customers.

Future Work

1. **Expand Features for Deeper Insights**
   - Incorporate regional and product-level data to identify localised trends.
   - Analyse customer behaviour over time to refine segmentation further.
2. **Integrate Predictive Analytics**
   - Develop churn prediction models to prevent customer loss.
   - Forecast customer lifetime value for strategic prioritisation.

This analysis has provided a clear framework for understanding customer behaviours and actionable insights for segmentation. By leveraging these findings, the company can effectively target its marketing efforts, boost customer engagement, and maximise revenue growth.