Contents

Overview of Approach

This analysis aimed to segment customers into actionable groups, enabling targeted marketing strategies and improved customer retention. The methodology followed a systematic process:

1. **Data Preprocessing**: Ensuring the dataset's reliability and quality, including cleaning, feature engineering, and handling missing values.
2. **Feature Engineering**: Deriving key metrics such as **Frequency**, **Recency**, **Customer Lifetime Value (CLV)**, **Average Unit Cost**, and **Age** to form the basis of segmentation.
3. **Outlier Detection**: Using the **Isolation Forest** algorithm to identify and analyze anomalies in the data.
4. **Optimal Cluster Selection**: Employing the **Elbow Method** and **Silhouette Score** to determine the most effective number of clusters (kk).
5. **Clustering**: Applying **K-Means Clustering** to segment the dataset into meaningful groups.
6. **Visualization**: Using **boxplots**, **PCA**, and **t-SNE** to understand and validate the clusters visually.

This structured approach ensured that the resulting segments were meaningful, actionable, and aligned with the business objectives.

Data Preprocessing and Feature Engineering

Initial Observations and Cleaning

The dataset consisted of 951,669 entries with 20 features. Key preprocessing steps included:

- **Removal of Duplicates**: Identified and eliminated 21 duplicate rows.
- **Missing Values**: Although three features—**City**, **Postal Code**, and **State Province**—contained missing values, they were not critical to the analysis and were left unaddressed.
- **Aggregation**: Customer data was aggregated into a single row per individual, ensuring a cohesive dataset for segmentation.

Feature Engineering

Key metrics were derived to align with segmentation goals:

- **Frequency**: Number of transactions per customer.
- **Recency**: Days since the last purchase.
- **Customer Lifetime Value (CLV)**: Revenue generated by each customer over their lifetime.
- **Average Unit Cost**: Average cost per item purchased.
- **Age**: Customer age based on their birthdate.

These metrics provided the foundation for effective clustering and analysis.

Exploratory Data Analysis

Distribution Analysis

- **Right-Skewed Distributions**: Most features, including **Frequency**, **Recency**, and **CLV**, exhibited right-skewed distributions, as confirmed by histograms.
- **Age Distribution**: The age feature showed a multimodal distribution, highlighting distinct customer age groups.

Outlier Analysis

Significant outliers were identified, especially in **CLV** and **Average Unit Cost**, representing extreme customer behaviors. These outliers were managed using the **Isolation Forest** algorithm.

Feature Relationships

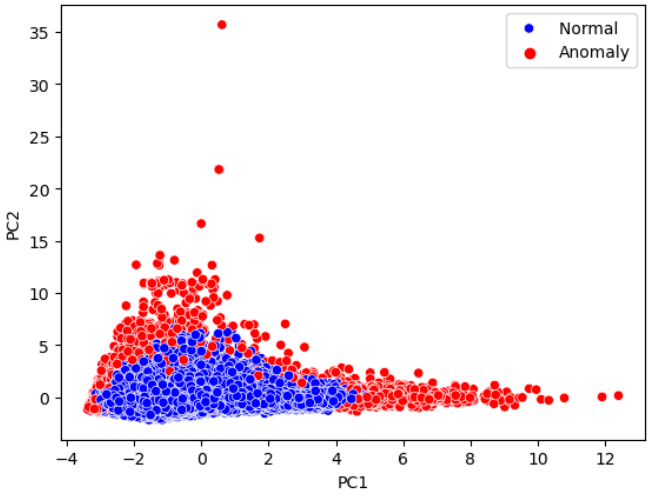
- **Frequency and CLV**: A positive linear relationship indicated that frequent buyers contributed higher lifetime value.
- **Frequency and Average Unit Cost**: No significant relationship was observed, suggesting diverse purchasing patterns among frequent buyers.

Outlier Detection

The **Isolation Forest** algorithm identified approximately 2.5% of the data as anomalies, focusing on extreme values in key metrics.

Visualization

A **PCA-based scatterplot** revealed that outliers were concentrated on the periphery of the data distribution, validating the algorithm's results.



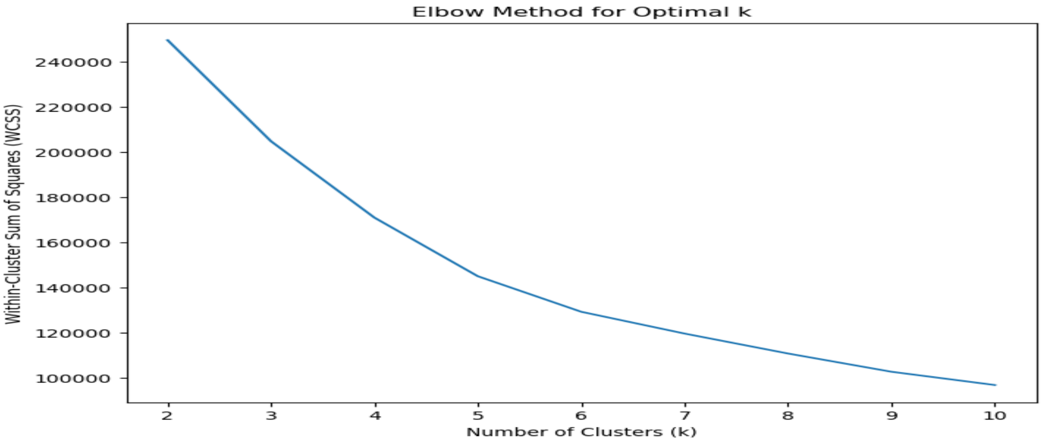
These outliers represent potential high-value customers or exceptions requiring tailored strategies.

Determining the Optimal Number of Clusters

Elbow Method

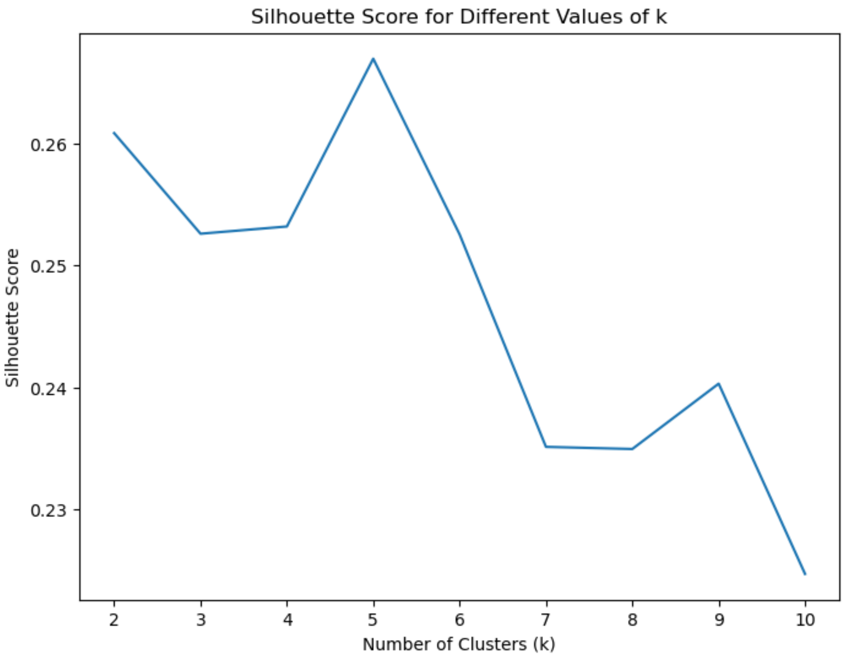
The **Elbow Plot** analyzed the within-cluster sum of squares (WSS) for kk values ranging from 2 to 10.

- The **optimal kk** was determined to be **5**, where the WSS curve began to flatten, indicating diminishing returns.



Silhouette Score

The **Silhouette Score**, which measures cluster cohesion and separation, confirmed k = 5 as the optimal value, achieving the highest score among tested options.



K-Means Clustering Results

Cluster Assignments

The K-Means algorithm segmented customers into five clusters with the following distribution:

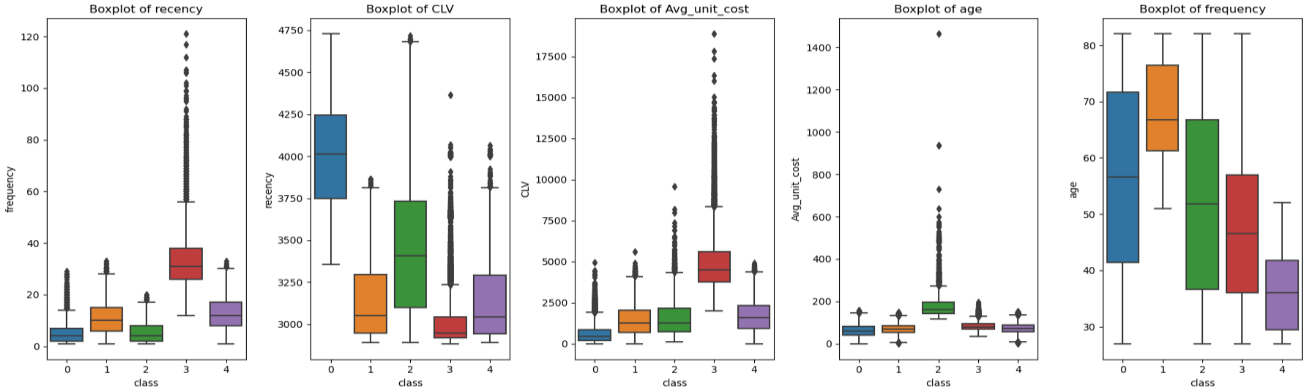
- **Cluster 1**: 22,592 customers
- **Cluster 3**: 20,296 customers
- **Cluster 4**: 11,111 customers
- **Cluster 2**: 10,622 customers
- **Cluster 0**: 3,679 customers

Each cluster represented distinct customer behaviors and characteristics.

Boxplot Analysis

Boxplots compared clusters based on key metrics, providing the following insights:

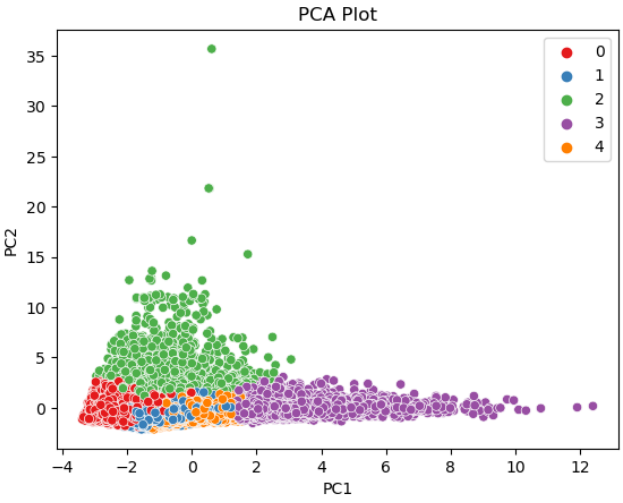
1. **Frequency**: Consistent across clusters, indicating it may not be a primary differentiator.
2. **Recency**: Significant variability suggests differing levels of customer engagement.
3. **CLV**: Wide variability highlights clusters with high-value customers.
4. **Average Unit Cost**: Variability reflects diverse purchasing behaviors.
5. **Age**: Younger customers were prominent in Clusters 1 and 4.



Dimensional Reduction for Visualization

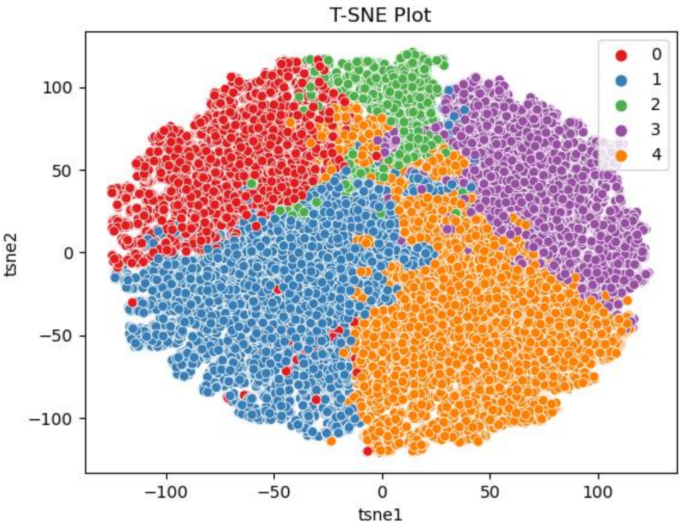
PCA Visualization

A **PCA-based 2D plot** showed distinct cluster separations, with some overlap in **Cluster 4**, suggesting shared traits with other clusters.



t-SNE Visualization

The **t-SNE visualization** confirmed the PCA results, highlighting clear separations with nuanced local structures. The stability of the t-SNE output across multiple perplexity values reinforced its reliability.



Insights and Business Implications

Cluster-Specific Observations

1. **High CLV Clusters**: Prioritize retention efforts and upselling strategies for these valuable customers.
2. **Young Customers**: Develop targeted campaigns for younger customers in Clusters 1 and 3, focusing on trendy or tech-oriented products.
3. **Engaged Customers**: Clusters with higher recency values indicate engaged customers, suitable for loyalty programs and personalized offers.

Outlier Management

Outliers represent unique opportunities or risks. For example:

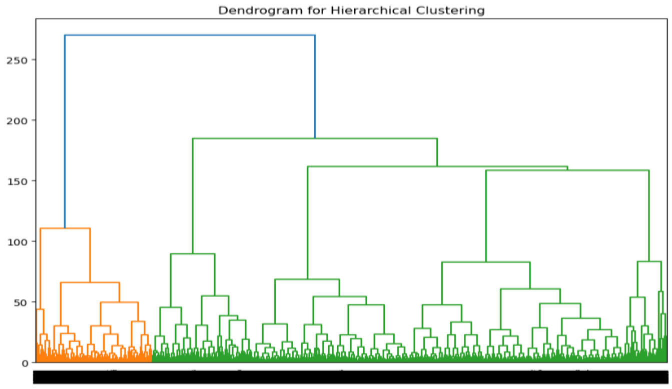
- **High CLV Outliers**: May warrant exclusive offers or VIP programs.
- **Low Recency Outliers**: Require re-engagement strategies.

Hierarchical Clustering and Dendrogram Analysis

To complement the K-Means clustering, hierarchical clustering was performed using the Ward method, which minimizes variance within clusters. A dendrogram was generated to visualize the clustering process and determine the optimal number of clusters.

Dendrogram Insights

The dendrogram revealed a clear cutoff point at 5 clusters, confirmed by the distinct separation of branches at this level. This aligns with the results of the Elbow and Silhouette analyses from K-Means.



Agglomerative Clustering Results

Using the cutoff point from the dendrogram, agglomerative clustering assigned each customer to one of five clusters. This method further validated the segmentation by producing consistent clusters aligned with previous analyses.

The dendrogram and hierarchical clustering provide additional confidence in the robustness of the customer segmentation, supporting targeted strategies based on the identified clusters.

Conclusion

This analysis demonstrated a robust approach to customer segmentation, leveraging clustering techniques and advanced visualization tools. Key takeaways include:

1. The identification of five distinct customer clusters.
2. Validation of clustering through metrics and dimensional reduction techniques.
3. Actionable insights to guide marketing strategies, customer retention, and revenue growth.

By focusing on the unique characteristics of each segment, businesses can develop tailored strategies to maximize customer satisfaction and profitability.