Customer Segmentation with Clustering and Machine Learning:

A Data-Driven Approach to Retail Optimization

By Shishir. M. Paltanwale

Date: 4th April 2025

TABLE OF CONTENTS

PROBLEM STATEMENT

In the highly competitive retail industry, understanding and effectively serving customers is essential for achieving key strategic marketing objectives such as increased customer satisfaction, improved retention, optimized pricing, and efficient resource allocation. One of the most effective ways to achieve this is through **customer segmentation** — grouping customers based on attributes like demographics, geographics, psychographics, behaviour, technographics, and specific needs.
The goal of this project is to apply **critical thinking** and **machine learning techniques**, particularly **clustering algorithms**, to segment customers meaningfully. By leveraging data-driven customer segmentation, businesses can tailor marketing strategies, personalize offerings, and ultimately drive growth through more customer-centric decision-making.

OVERVIEW APPROACH

The analysis begins by importing the necessary libraries and dataset from the provided URL. The **Data Frame** is then examined for missing values, duplicates. Feature engineering is performed by creating key attributes such as **Frequency, Recency, CLV, Average Unit Cost,** and **Customer Age**, with appropriate and outliers, followed by aggregation to ensure each row represents a unique customer, next scaling and encoding applied as needed. **Exploratory Data Analysis (EDA)** was conducted through visualizations to uncover patterns.

To streamline preprocessing, a **Column Transformer** and **Pipeline** are incorporated. The optimal number of clusters (*k*) is determined using the **Elbow Method** and **Silhouette Score**, and **Hierarchical Clustering** is visualized with a **Dendrogram**. Based on the chosen *k*, k-means clustering is applied, assigning cluster labels to each customer. Boxplots help analyse clusters across key attributes. To enhance visualization, PCA and t-SNE are used for dimensionality reduction, enabling a 2D representation of the clusters with distinct colours.

Finally, key findings and insights are that comparing clustering methods and the optimum number of clusters for this dataset are estimated to be around 6.

DESCRIPTION OF THE ANALYSIS

Data Preprocessing

The dataset contains **951,669 records** and includes a total of **19 original features**. The first step involved loading the dataset and performing exploratory data analysis (EDA) to understand the structure and quality of the data. Several key preprocessing steps were performed:

- **Handling missing values**: Columns with significant missing data were either dropped or imputed based on context.
- **Date parsing**: Features such as "**Delivery Date**" were converted into Recency by calculating the number of days since the last purchase.
- **Duplicate detection**: Duplicate records based on "**Customer ID**" were identified and removed to ensure one row per customer.
- **Outlier detection**: The **Isolation Forest** algorithm was used to detect and remove anomalous data points that could skew clustering results.

Feature Engineering

To capture meaningful customer behaviour, new features were created:

- **Frequency**: Number of transactions per customer.
- **Recency**: Days since last interaction.
- **CLV**: Estimated lifetime value using total spend and frequency.
- **Average Unit Cost**: Mean cost per item per customer.
- **Customer Age**: Derived from available timestamps or inferred data.

All features were scaled using **StandardScaler** to normalize the data for clustering algorithms.

EXPLORATORY DATA ANALYSIS (EDA)

EDA helped uncover patterns and distributions in the data. Histograms and boxplots were created to assess skewness and distribution in numerical features. This step confirmed that engineered features captured meaningful customer behaviours.

METHODS

S. No	Method	Purpose	Description
1	Elbow Method	Optimal K	Plotted inertia (within-cluster sum-of-squares) against different K values. The curve did not have a very sharp bend, but based on visual inspection, 6–8 clusters appeared reasonable.
2	Silhouette Score	Optimal K	Measured how similar a point is to its own cluster versus other clusters. Scores ranged from 0.21 to 0.28, with **K=6** showing a balanced and consistent structure.
3	Hierarchical Clustering	Clustering	An agglomerative approach was used to generate a dendrogram. Due to the dataset size, the number of clusters was capped at 60 for visualization purposes. Outliers were excluded for efficiency.
4	K-Means Clustering	Clustering	Based on the silhouette and elbow analysis, **K=6** was chosen as the optimal number of clusters. The clustering was performed on the scaled dataset, and cluster labels were assigned to each customer
5	PCA (Principal Component Analysis)	Dimensionality Reduction	Reduced high-dimensional data to two principal components. PCA showed a reasonable separation between clusters.
6	t-SNE (t-Distributed Stochastic Neighbor Embedding):	Dimensionality Reduction	Captured nonlinear structures in the data and provided more nuanced visualization of cluster overlap and compactness.

EXPLANATION OF VISUALISATIONS

Histogram

To explore the distribution of key features in your customer segmentation dataset.

Boxplots

Boxplots were generated for each feature across the six clusters to visually assess the distribution. These plots helped identify differences in spending habits, recency, and frequency per cluster, supporting targeted strategy development.

Dendrogram

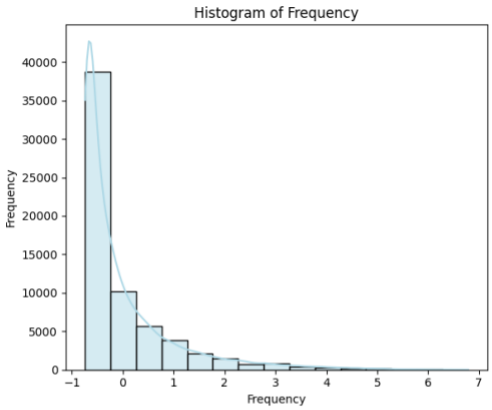
Used to visualize hierarchical clustering and determine where natural separations occurred. While useful for structure, it was limited by dataset size.

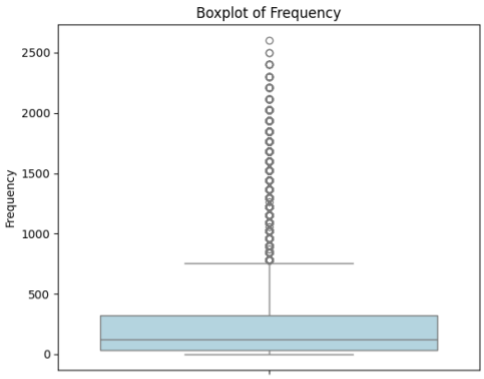
PCA & t-SNE Scatter Plots

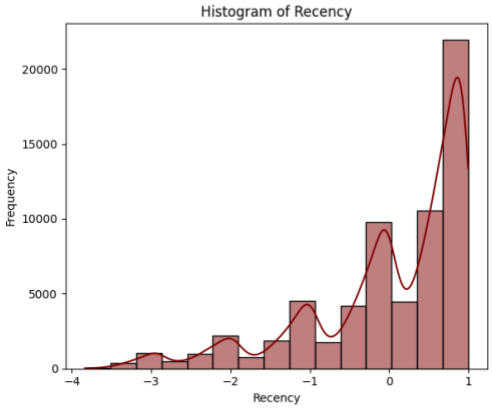
These 2D plots showed cluster separation visually. PCA provided linear insight, while t-SNE captured nonlinear relationships and highlighted overlaps and outliers more effectively.

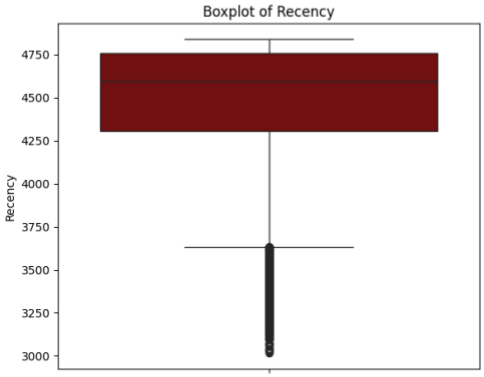
DATA VISUALISATION

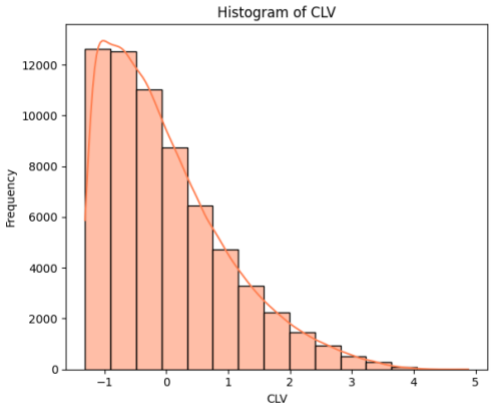
1. Histograms and Boxplots (all features)

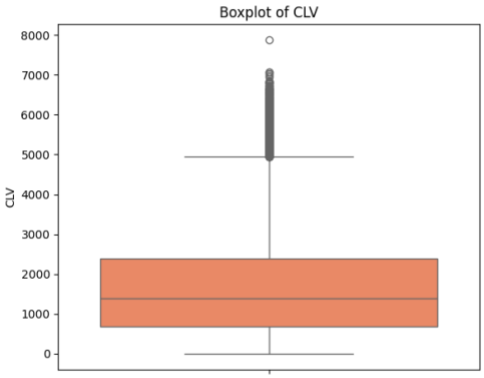

Histogram of Frequency
Frequency
Frequency

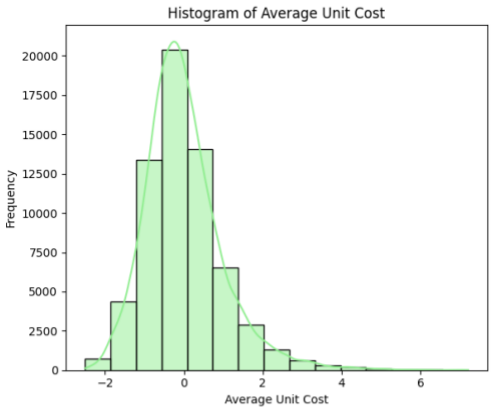

Boxplot of Frequency
Frequency

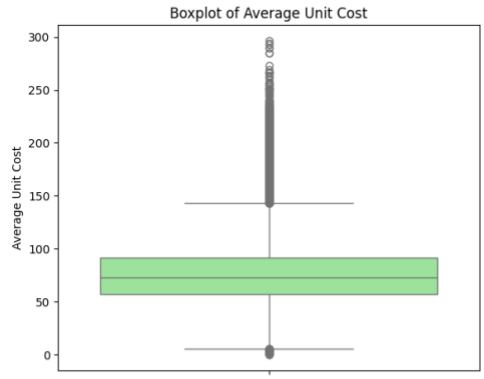

Histogram of Recency
Frequency
Recency

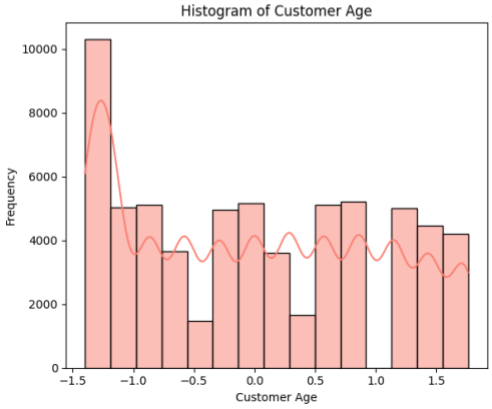

Boxplot of Recency
Recency

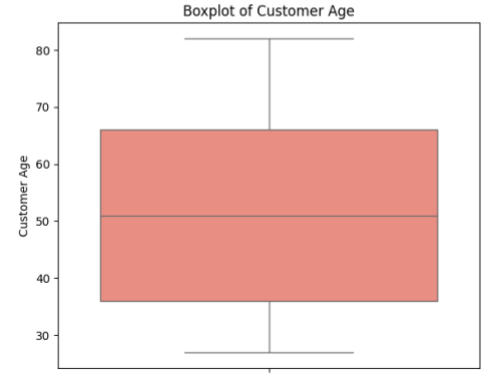

Histogram of CLV
Frequency
CLV


Boxplot of CLV
CLV

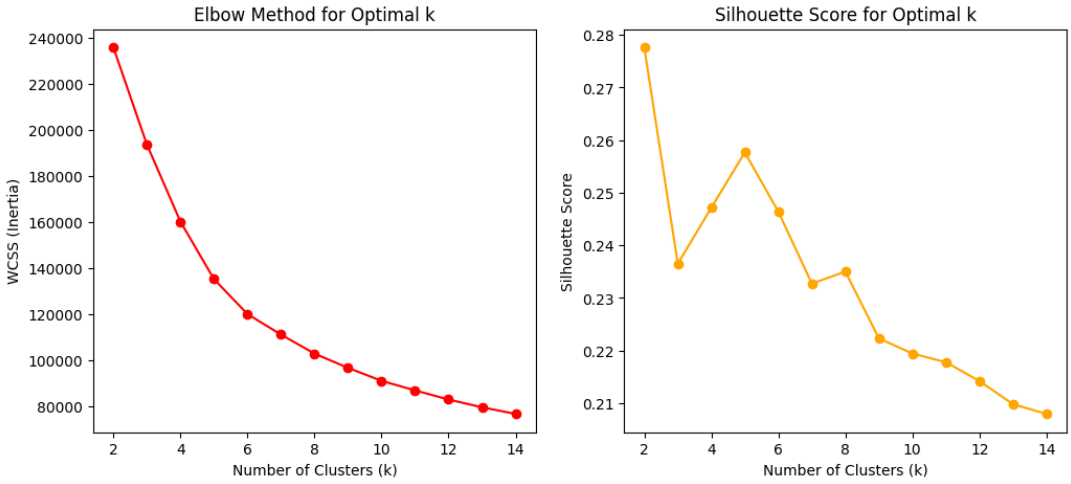

Histogram of Average Unit Cost
Frequency
Average Unit Cost


Boxplot of Average Unit Cost
Average Unit Cost

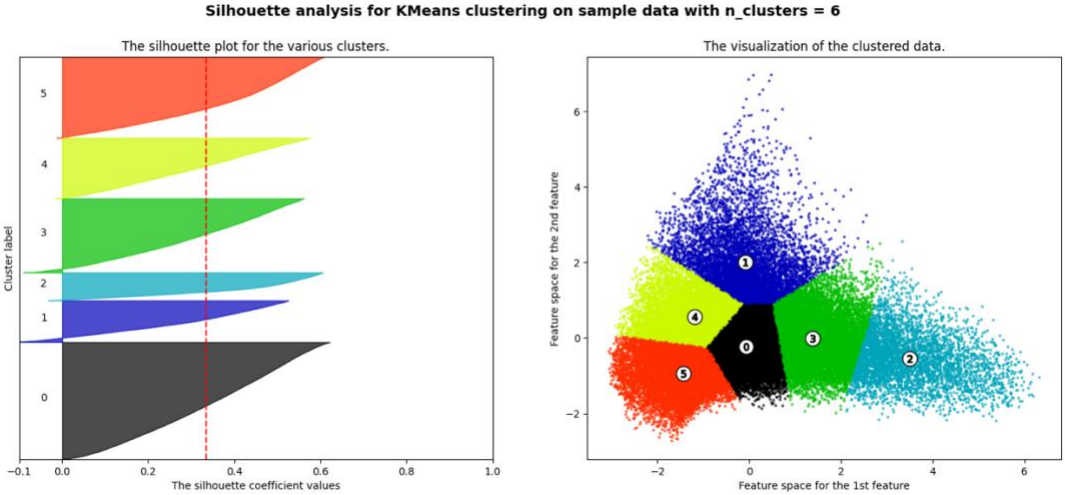

Histogram of Customer Age
Frequency
Customer Age


Boxplot of Customer Age
Customer Age

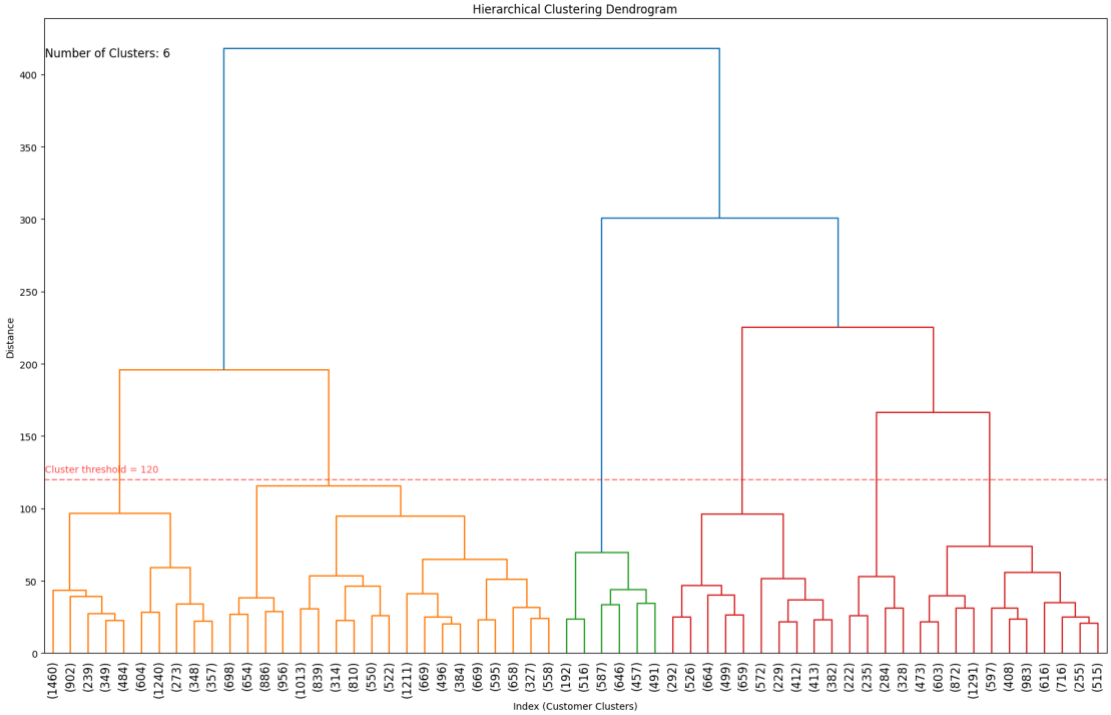
2. Elbow method with Silhouette Score



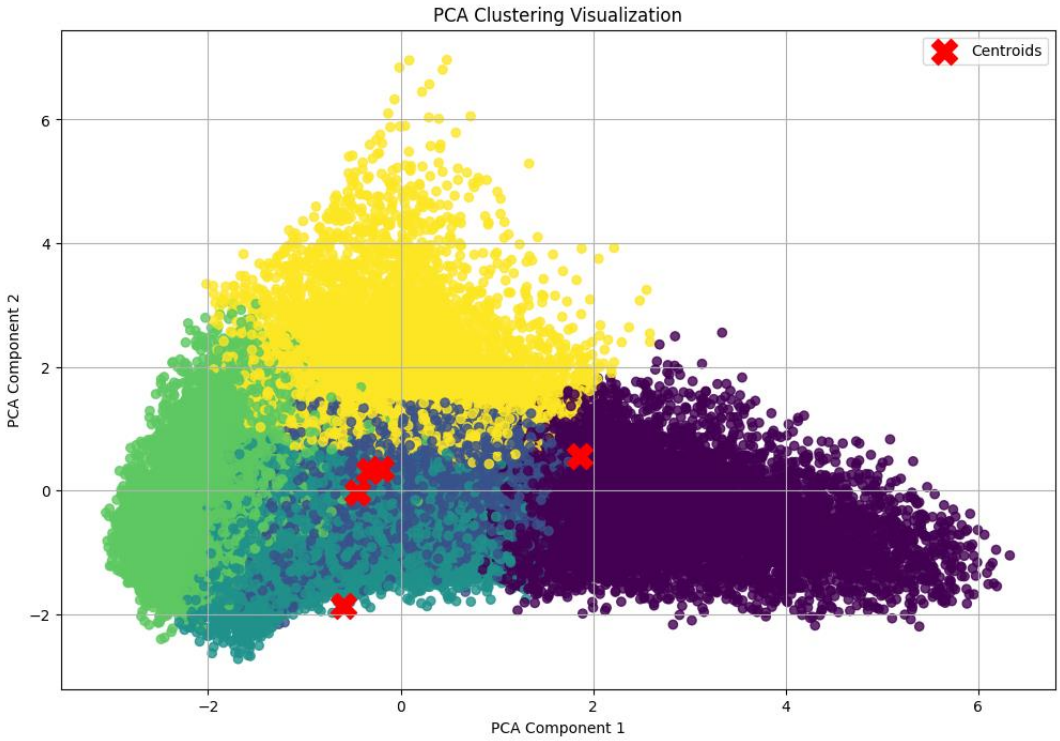
3. Silhouette Plots



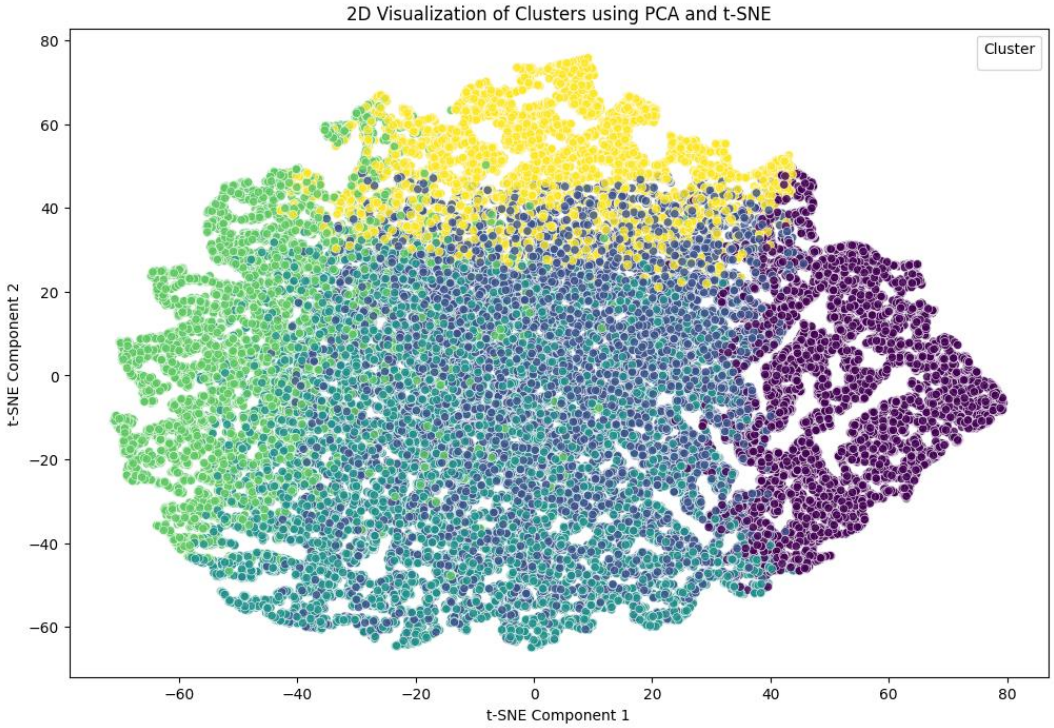
4. Dendrograms



5. PCA visualisation with Centroids:



6. t-SNE Visualisation:



RESULT

Both the **Elbow Method** and **Silhouette Analysis** were helpful, but the **Silhouette Score** provided a more quantitative metric for choosing the number of clusters. Despite close scores among K=5, 6, and 7, **K=6** was selected due to its slightly higher silhouette score and better separation seen in PCA and t-SNE visualizations. It also aligned well with business logic, producing clear and interpretable customer segments.

CONCLUSION

This customer segmentation project successfully applied unsupervised learning and clustering techniques to extract valuable insights from a large retail dataset. By engineering meaningful features and using clustering methods supported by dimensionality reduction and visualization, we identified distinct customer groups that can be targeted with tailored strategies.

The **K-Means algorithm with K=6** provided the best overall performance, supported by both visual and metric-based analysis. The approach demonstrates the power of machine learning in driving customer-centric decision-making in the retail sector.

REFERENCES

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