DATA CAREER
ACCELERATOR

# Customer segmentation with clustering

Juan Pablo Salazar

2024/12/12


## I. Introduction and problem statement:

Accurate forecasting of book sales is essential for publishers seeking to optimize inventory, minimize costs, and maximize revenue. This project uses real-world data from Nielsen BookScan to analyze and predict weekly sales for two books of interest, “The Alchemist” and “The Very Hungry Caterpillar”. We apply classical time series analytical methods, such as decomposition and ARIMA, alongside machine learning approaches, including XGBoost and Long Short-Term Memory (LSTM) neural networks. To enhance model performance, we employ hyperparameter tuning and explore hybrid forecasting strategies using both sequential and parallel combinations of models. The aim is to provide data-driven reuncover underlying sales patterns, seasonality, and long-term trends for these titles, ultimately providing data-driven recommendations to inform stock management and marketing strategies for publishers and retailers alike.

After erasing the duplicate rows, dropping the unnecessary columns, and creating the five key features with one row per customer, we observed from the Python describe method that there are **68,300** customers. Moreover, after performing the IQR method on each of the features analyzed, we identified **3.89%** outliers for frequency, **4.91%** outliers for recency, **3,79%** outliers for customer lifetime value, **4.23%** outliers for average unit cost, and **0%** outliers for age.

When analyzing the histogram plots of the 5 features (Fig. I and Notebook) we can see that recency, frequency, CLV, and AUC distributions are rightly skewed. However, age shows an almost uniform distribution with the highest bin of the plot being about 30 years old. Since CLV and AUC are rightly skewed the company could create budget-friendly marketing campaigns since most customers have low expenditures.

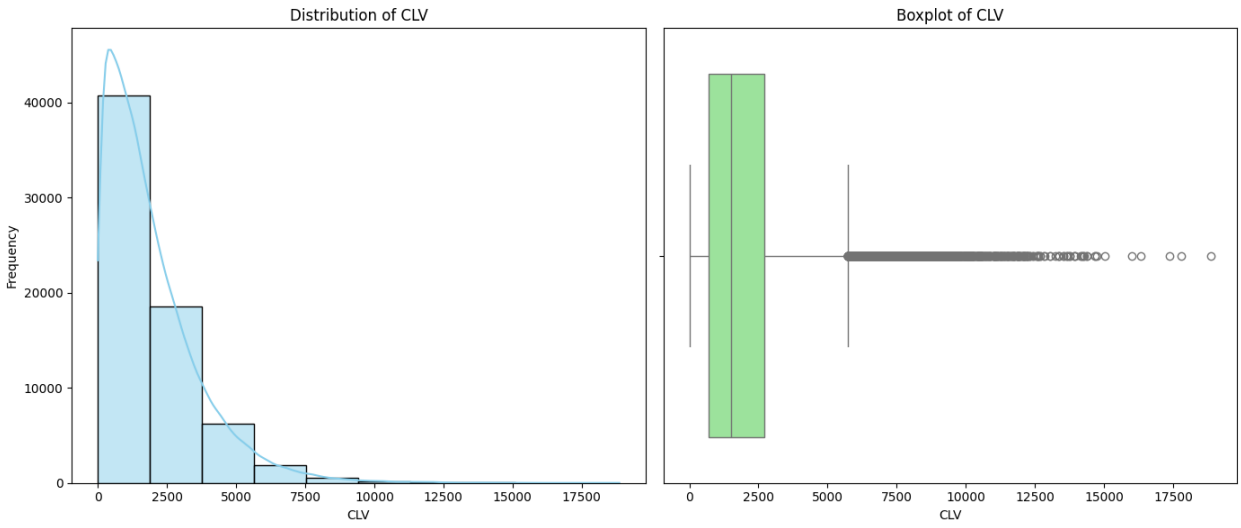


Figure I: Histogram of the distribution of the Customer lifetime value feature and its respective boxplot

II. Determining the optimal number of clusters with the elbow and silhouette methods.

a) The elbow method

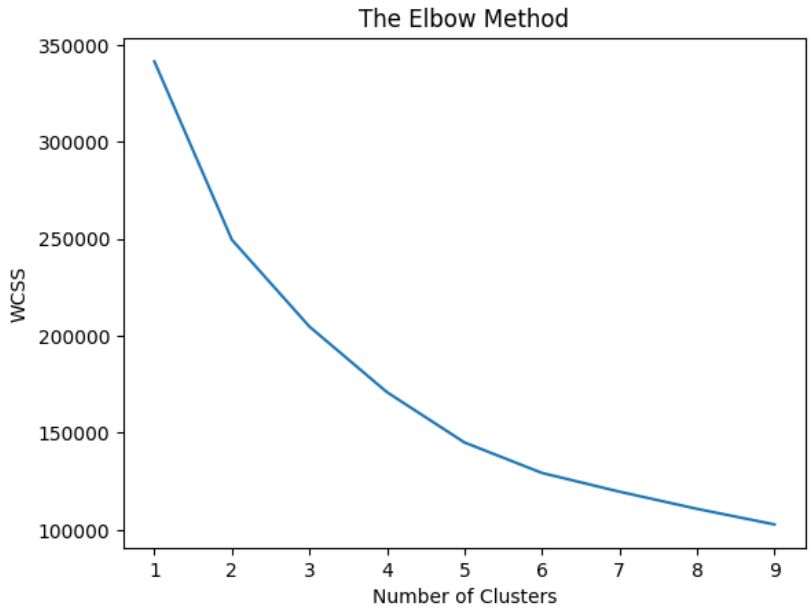


Figure II: Elbow method plot

The elbow method allows us to determine the optimal number of clusters by identifying the point on the plot where the decrease rate of within-cluster variation (WCSS) starts lowering. In Figure

II, it is hard to discern a distinct bent of the curve where the rate diminishes. Nonetheless, the bend in the plot of Figure II suggests the optimal number of clusters is 5 or 6.

b) The silhouette score method

Another method to determine the optimal number of clusters is the silhouette score method, which calculates how well each data point fits the assigned cluster and outputs the mean score, ranging from -1 to 1, with higher values indicating better fitness into the clusters. We have computed the silhouette score for two, three, four, five, and six clusters with **five** clusters showing the highest silhouette score of **0,267**. This suggests that the most cohesive and well-separated configuration of clusters is obtained with five clusters. The visualization of the clustered data in Figure III confirms this, as the five-cluster plot shows less overlap and fewer negative silhouette values. (Compare to figures in Notebook).

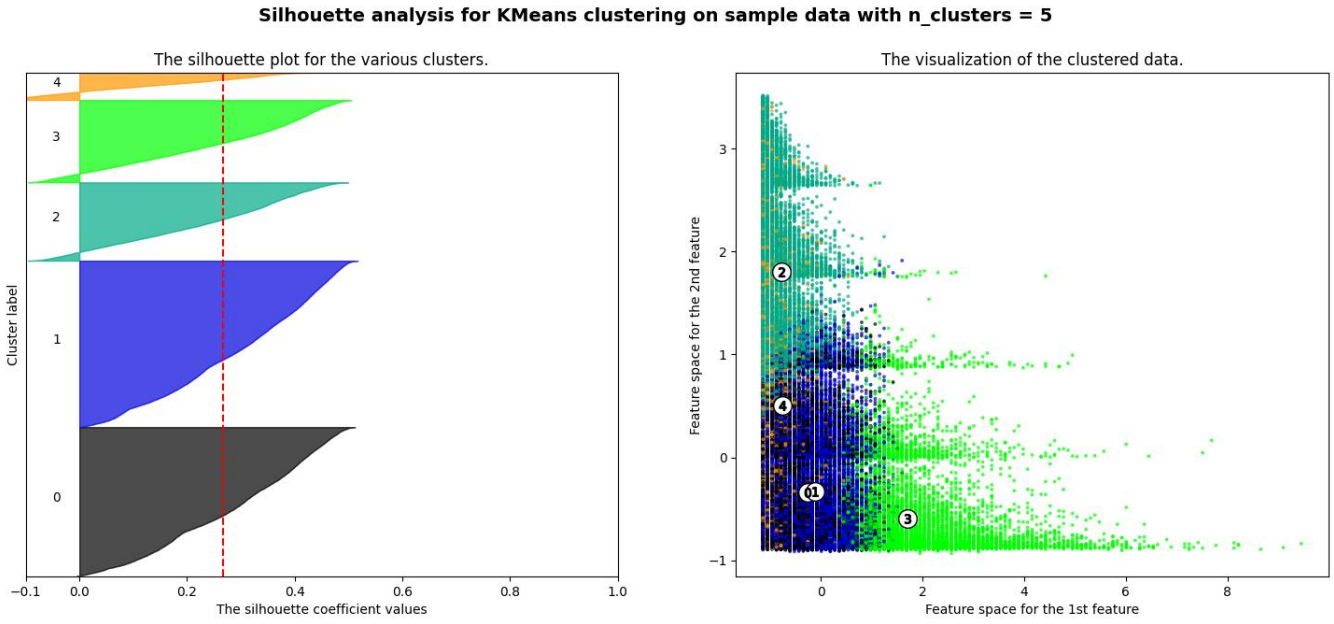


Figure III. Silhouette method analysis for K-means clustering on sample data with 5 clusters

III. Hierarchical clustering method

The hierarchical method was also employed to confirm the optimal number of clusters identified using the silhouette and elbow methods. Hierarchical clustering iteratively combines and divides clusters based on similarities, creating a dendrogram illustrating the relationships between clusters.

Moreover, a cut-off before a significant distance jump in the hierarchical dendrogram (Figure IV) indicates the optimal number of clusters. In this case, a cut-off around a distance of 200 appears optimal, giving us a total cluster count of **five**, which confirms the results from the previous methods.

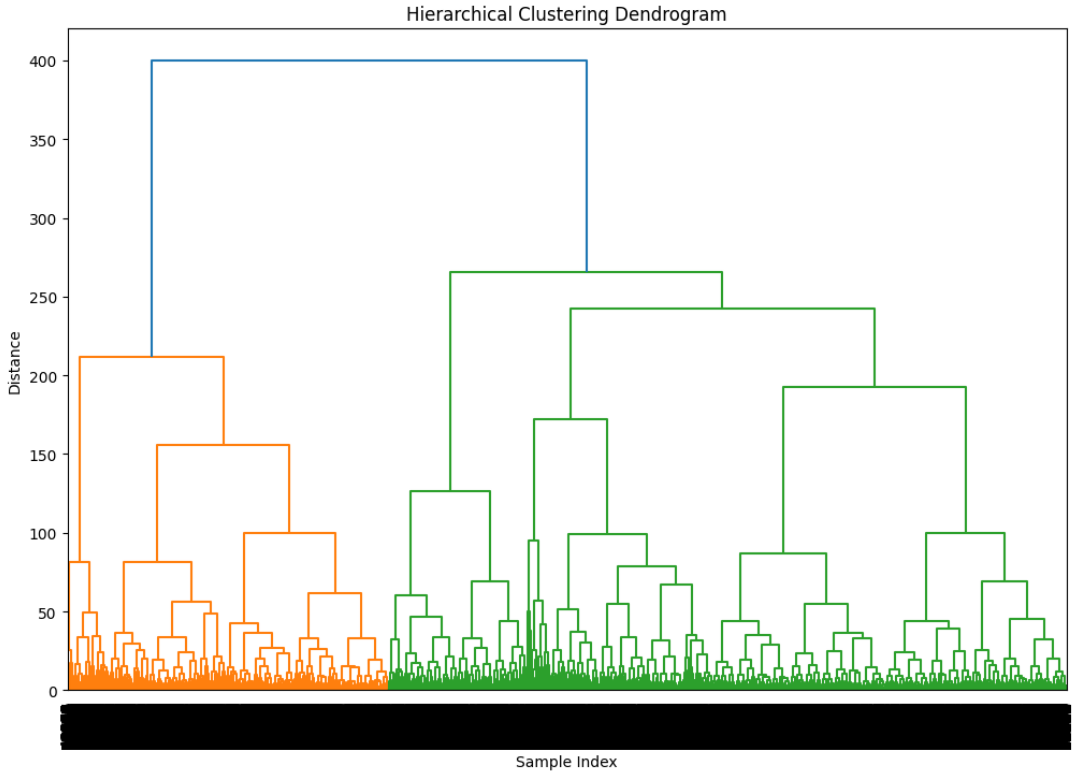


Figure IV. Hierarchical clustering dendrogram

IV. Boxplots of Clusters by Frequency, Recency, CLV, Average Unit Cost, and Customer Age.

The customer age and average unit cost boxplots did not show significantly different results. In contrast, the boxplots of CLV and Frequency in (Fig. V) show higher means for clusters 4 and 2. The company could prioritize these clusters since they represent the most frequent buyers and have the highest lifetime value. For the other clusters, marketing strategies that encourage frequent purchases could be applied. Furthermore, the boxplot of the recency feature shows that cluster 1 has significantly higher values compared to clusters 0, 4, and 2, with the highest mean value. This means the company could create a marketing strategy acknowledging customers from cluster 1 as the most recent buyers.

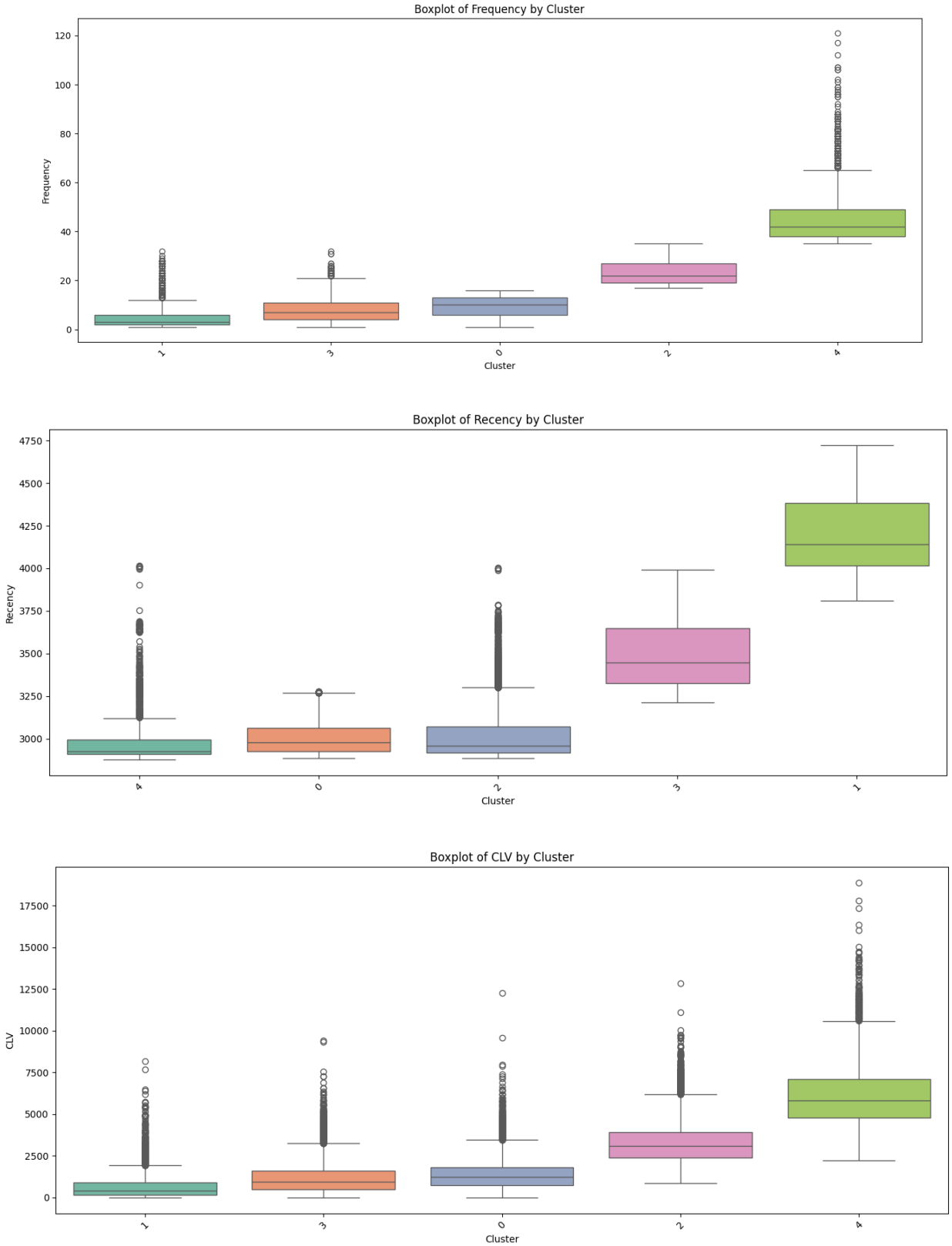


Figure V. Boxplots of the recency, frequency, and customer lifetime value by cluster

(Note that the boxplots of clusters in Figure V were sorted in order of their mean values for each feature, and colors were not assigned to specific clusters)

## V. Dimensionality reduction with PCA and t-SNE

Finally, we applied dimensionality reduction with the PCA and t-SNE methods to visualize the clusters and confirm they are well separated (Figures VI and VII). The PCA visualization shows that the clusters are well separated although clusters 1 and 3 slightly overlap with cluster 2. Additionally, cluster 4 is more spread, which may indicate a higher variance among the features. The t-SNE visualization also shows well-defined and separated clusters. However, a point from cluster 3 is very distant from its respective cluster and there is some overlap of clusters 1 and 3. These overlaps may correspond to noise or outliers from the dataset.

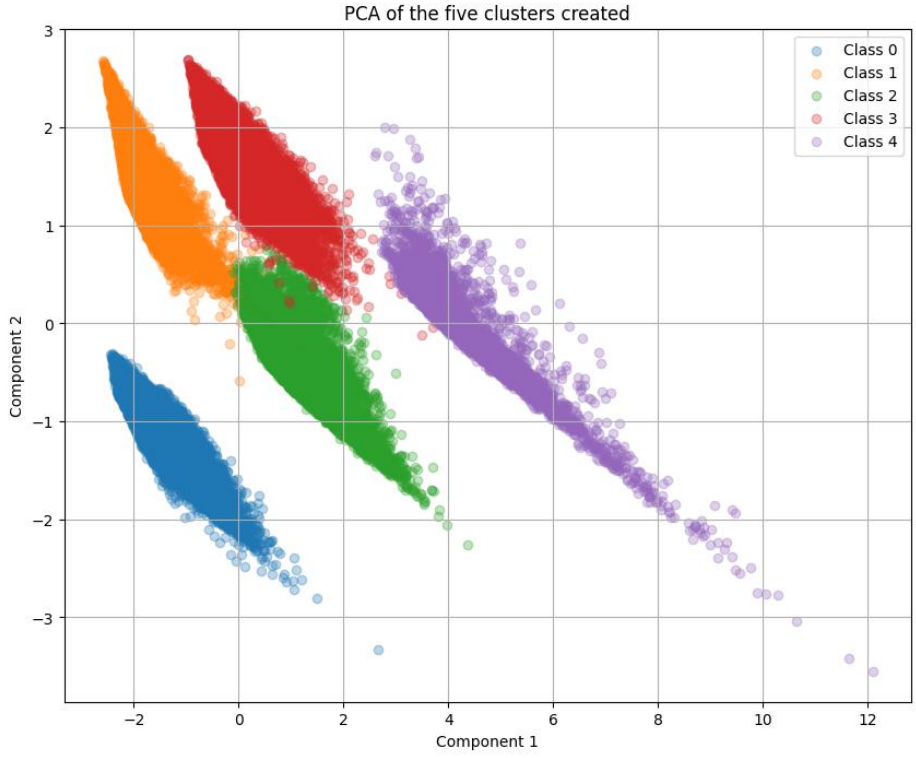


Figure VI. 2D visualization of PCA of the five customer clusters

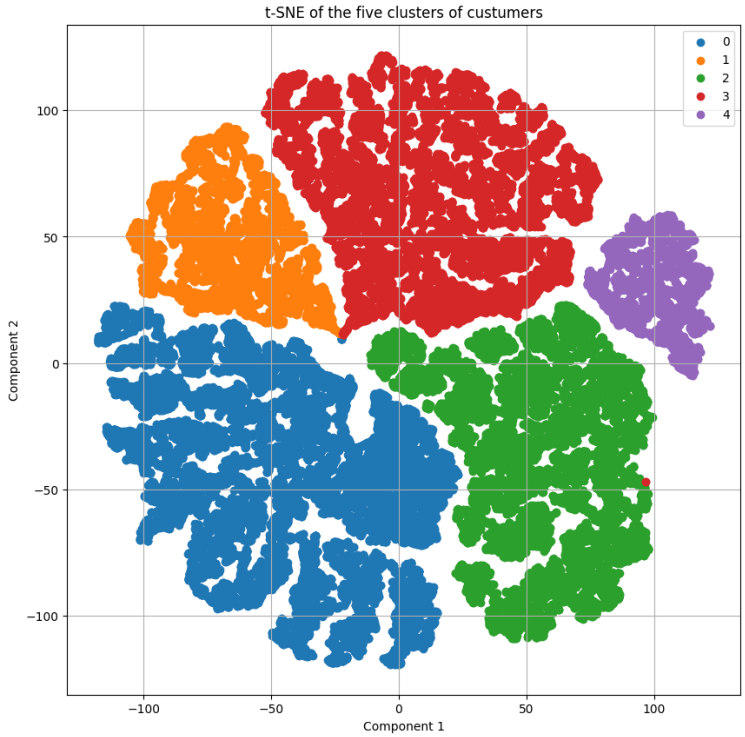


Figure VI. 2D visualization of t-SNE of the five customer clusters

## Conclusion:

The customer segmentation of the company's dataset was conducted to achieve optimal clustering. The best number of clusters was determined by applying the elbow and silhouette methods, which identified five as the optimal cluster number. This was further confirmed by hierarchical clustering. Additionally, the dimensionality reduction techniques, PCA, and t-SNE, visually validated that the clusters were well-defined and separated. Furthermore, boxplots were generated to analyze the buying behavior of the customer segments, providing valuable insights for developing targeted marketing strategies.