Zeotap Data Science Internship, Clustering Documentation

Mahesh Reddy

January 27, 2025

1 Choice of clustering algorithm

For customer segmentation, K-Means Clustering was chosen due to its efficiency and simplicity in grouping data points based on similarity. It is a centroid-based algorithm that iteratively assigns data points to the nearest cluster center, minimizing intra-cluster distances while maximizing inter-cluster distances. The algorithm is well-suited for this task as it handles numerical and categorical data effectively (with proper preprocessing) and allows clusters to be easily interpreted. The optimal number of clusters was determined using the Davies-Bouldin Index and Silhouette Score to ensure the quality and cohesion of the clusters.

1.1 Considering Principal Component Analysis

The clustering analysis using the K-Means algorithm, with Principal Component Analysis (PCA) for dimensionality reduction, also resulted in the formation of 5 clusters. The clustering achieved a Final Davies-Bouldin Index of 0.917, indicating moderately compact and well-separated clusters, and a Silhouette Score of 0.382, reflecting average cohesion and separation of clusters. PCA was applied to reduce the dataset to two principal components, capturing the majority of the variance in the features. This dimensionality reduction simplified the visualization of clusters while retaining the essential structure of the data. The number of clusters (k) was determined by evaluating values between 2 and 10, with the optimal value of k = 5 chosen based on the Davies-Bouldin Index. While PCA provided a convenient way to visualize the clusters in 2D, the clustering performance metrics remained the same as the non-PCA approach, highlighting that PCA primarily enhanced visualization rather than improving clustering performance in this analysis.

1.2 Without Considering Principal Component Analysis

The clustering analysis using the K-Means algorithm, without applying PCA for dimensionality reduction, resulted in the formation of 5 clusters. The clustering achieved a Final Davies-Bouldin Index of 0.917, indicating moderately compact and well-separated clusters, and a Silhouette Score of 0.382, reflecting average cohesion and separation of clusters. The features used for clustering included Quantity, TotalValue, and Region, which were chosen to capture both transactional and regional variations in customer behavior. The parameters for K-Means were determined by evaluating the number of clusters (k) between 2 and 10, using the Davies-Bouldin Index to identify the optimal value of k = 5. Unlike the PCA-based approach, this method preserved the interpretability of the original features, enabling direct visualization of clusters based on meaningful variables like Quantity and TotalValue. The results demonstrate that PCA was not necessary in this case, as the clustering performance metrics were identical, and skipping PCA provided clearer insights into the dataset's inherent structure.

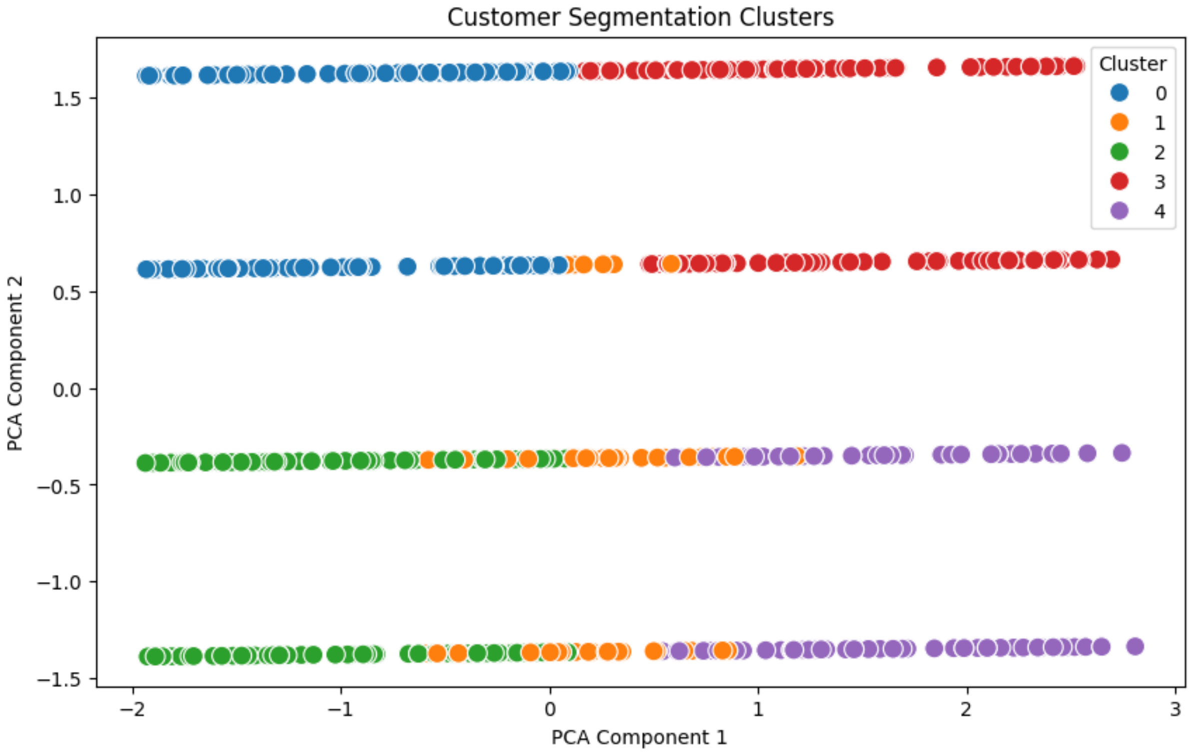

Figure 1: Customer Segmentation Plot (With PCA)

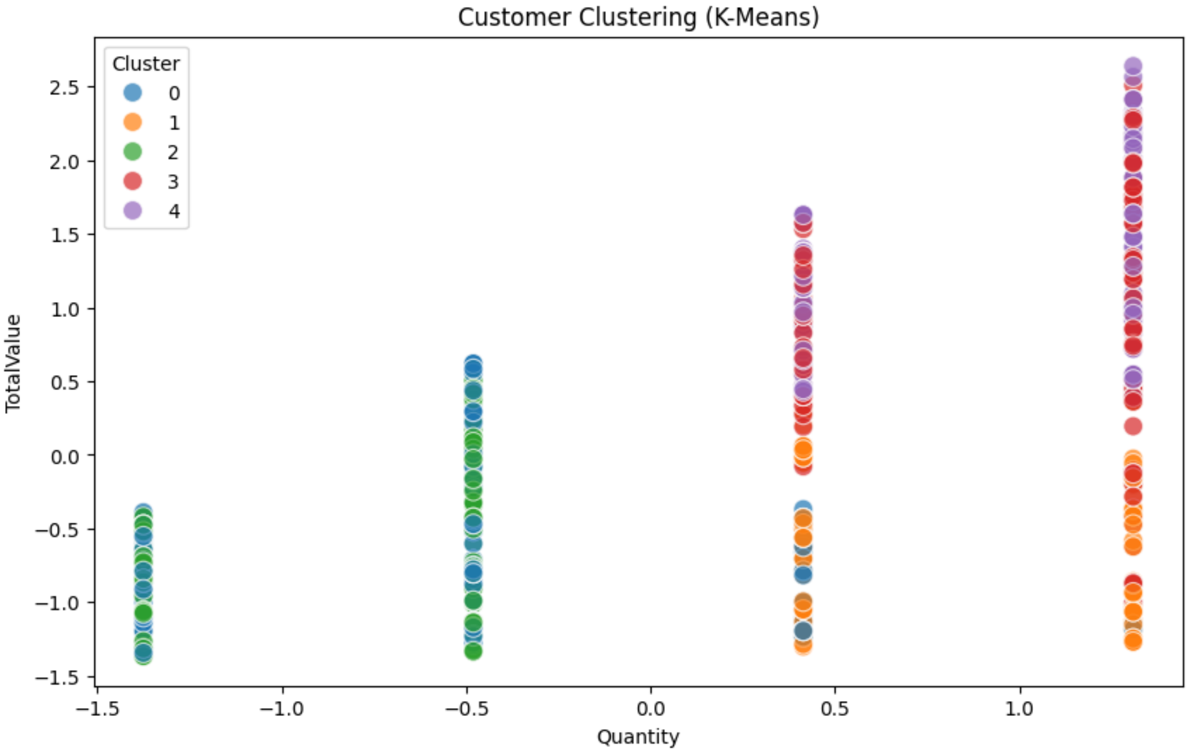

Figure 2: Customer Segmentation Plot (Without PCA)