ZeoTap Data Science Assignment: Customer Segmentation by Clustering

Number of Clusters Formed:

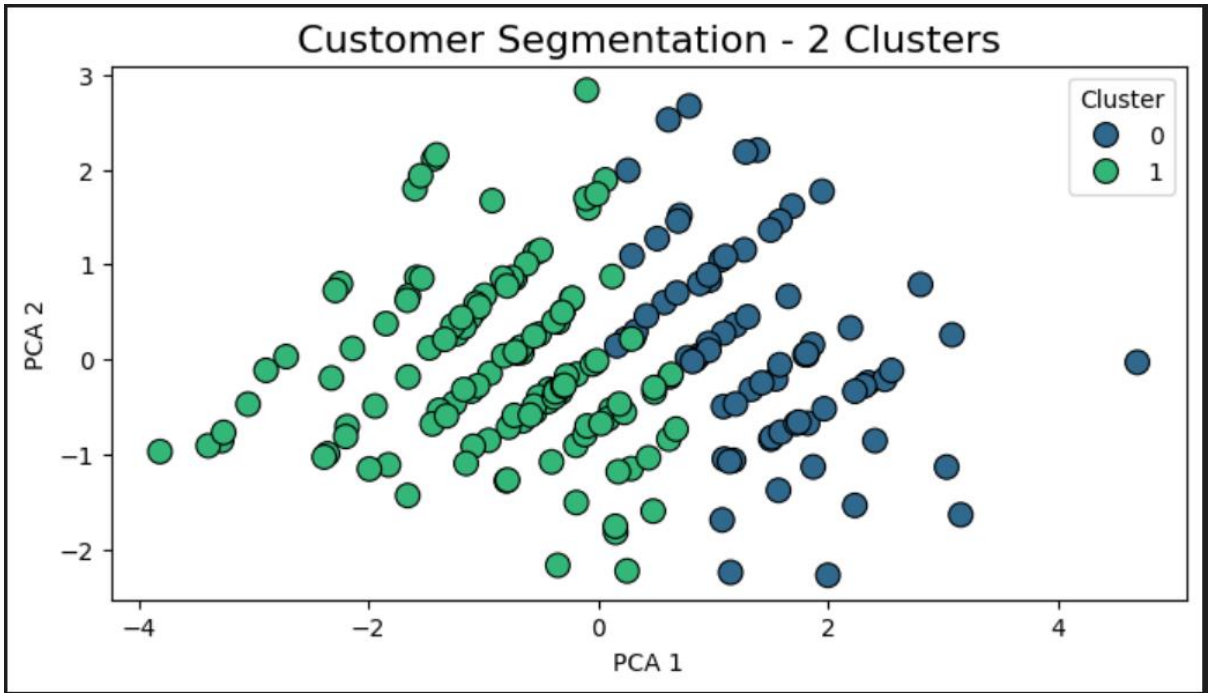
The K-Means clustering method was used to perform customer segmentation in this task. I chose **2 clusters** after trying numerous numbers ranging from 2 to 10 as the optimal segmented number. I chose the 2-cluster number because of the combined assessment of clustering metrics and the way clusters appeared visually. **Silhouette Score** along with **Davies-Bouldin Index** examined the quality of cluster formation.

Clustering Metrics:

- **Silhouette Score**: 0.3055 - The silhouette score tells us how well points match members of their original cluster group relative to alternative cluster groups. Clusters demonstrate better partitioning when the silhouette score reaches higher values. With a score of **0.3055** the clusters maintain some structure although additional improvements may refine their definition. Beside Fraction Location this clustering measure determines cluster cohesion through average similarity toward most analogous groupings.
- **Davies-Bouldin Index**: 1.2253 - The Davies-Bouldin Index measures the average similarity ratio of each cluster with the most similar cluster. Clusters having low DB Index scores show improved clustering results. Evaluation of **1.2253** indicates reliable cluster separation yet intermittent cluster intersection points.

Cluster Visualization:

To visualize the clusters Principal Component Analysis (PCA) generated a two-dimensional space from high dimensional data. Each cluster's central point received notation on the plot representing its position. Visual inspection showed clear divisions among customer groups yet overlapping areas needed refinement through cluster number adjustment or alternate clustering methodology selection.



Customer Segmentation Insights:

The clustering outcomes produced two separate customer segments. Marketing strategies must apply personalized approaches to each behavioral pattern group established by this analysis. The cluster showing maximal total transaction value stands out as the most valuable group among clusters that display lower retention rates and transaction values. Analysis revealed that both Silhouette Score (0.3055) and DB index (1.22) provided evidence of optimal grouping with minimum overlapping clusters when using two segments.

Other Results:

- For 2 clusters, the Silhouette Score was 0.3055 and DB index was 1.22, indicating the best clustering and least overlap.
- For 3 clusters, the Silhouette Score reduced to 0.25 and DB index 1.28.
- For 4 and 5 clusters, the Silhouette Score was around 0.22 and DB index was 1.32, providing the best balance between separation and cohesiveness.
- As the number of clusters increased, the Silhouette Score kept reducing and reached a minimum of 0.13.