# Evaluation Criteria for Customer Segmentation / Clustering

## 1. Clustering Logic and Metrics

### Clustering Logic:

- **Data Preparation:** The clustering process began with the merging of customer and transaction data to create comprehensive customer profiles. This included aggregating total spending and quantity purchased, as well as encoding categorical variables (e.g., regions) into numerical format.

- **Feature Standardization:** To ensure that all features contributed equally to the clustering process, the data was standardized using StandardScaler. This step is crucial for K-Means clustering, which is sensitive to the scale of the input features.

- **Clustering Algorithm:** K-Means clustering was chosen as the primary algorithm due to its simplicity and effectiveness in partitioning data into distinct groups. The algorithm iteratively assigns data points to clusters based on their proximity to centroids, recalculating centroids until convergence.

- **Optimal Number of Clusters:** The optimal number of clusters was determined through:
    - **Elbow Method:** Evaluated inertia (sum of squared distances from each point to its assigned cluster center) for cluster counts ranging from 2 to 10.

    - **Silhouette Score:** Measured how similar an object is to its own cluster compared to other clusters. Higher scores indicate better-defined clusters.

    - **Davies-Bouldin Index (DB Index):** Calculated to assess the average similarity ratio of each cluster with its most similar cluster. Lower values indicate better clustering quality.

### Clustering Metrics:

- **Inertia Values:** Recorded for different numbers of clusters, indicating how compact the clusters are.

- **Silhouette Scores:** Ranged from approximately 0.2593 (2 clusters) to a maximum of 0.5324 (8 clusters), suggesting varying levels of cluster separation.

- **Davies-Bouldin Index Values:** Ranged from approximately 1.5966 (2 clusters) to a minimum of 0.6317 (8 clusters), indicating that 8 clusters provided the best separation among groups.

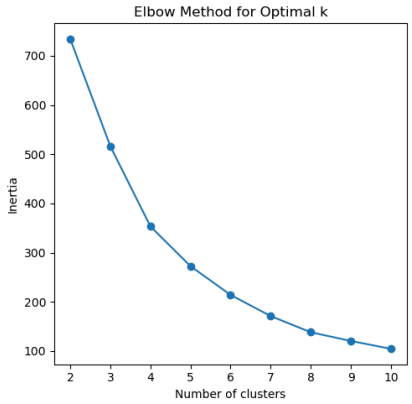
## 2. Visual Representation of Clusters

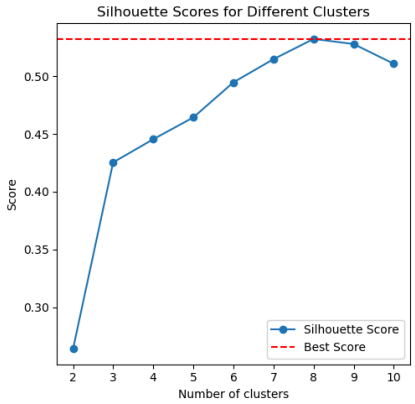
## Visualizations Created:

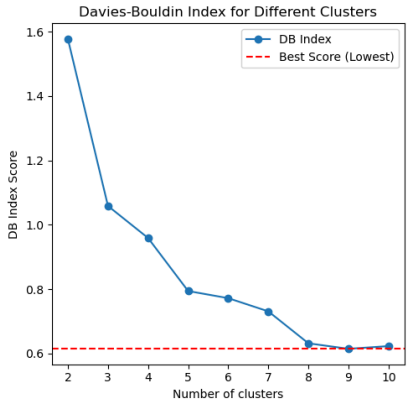
1. **Elbow Method Plot:**
    - A plot showing inertia values against the number of clusters, helping identify the "elbow point" where adding more clusters yields diminishing returns in inertia reduction.

2. **Silhouette Scores Plot:**
    - A plot displaying silhouette scores for each number of clusters, indicating how well-separated the clusters are.

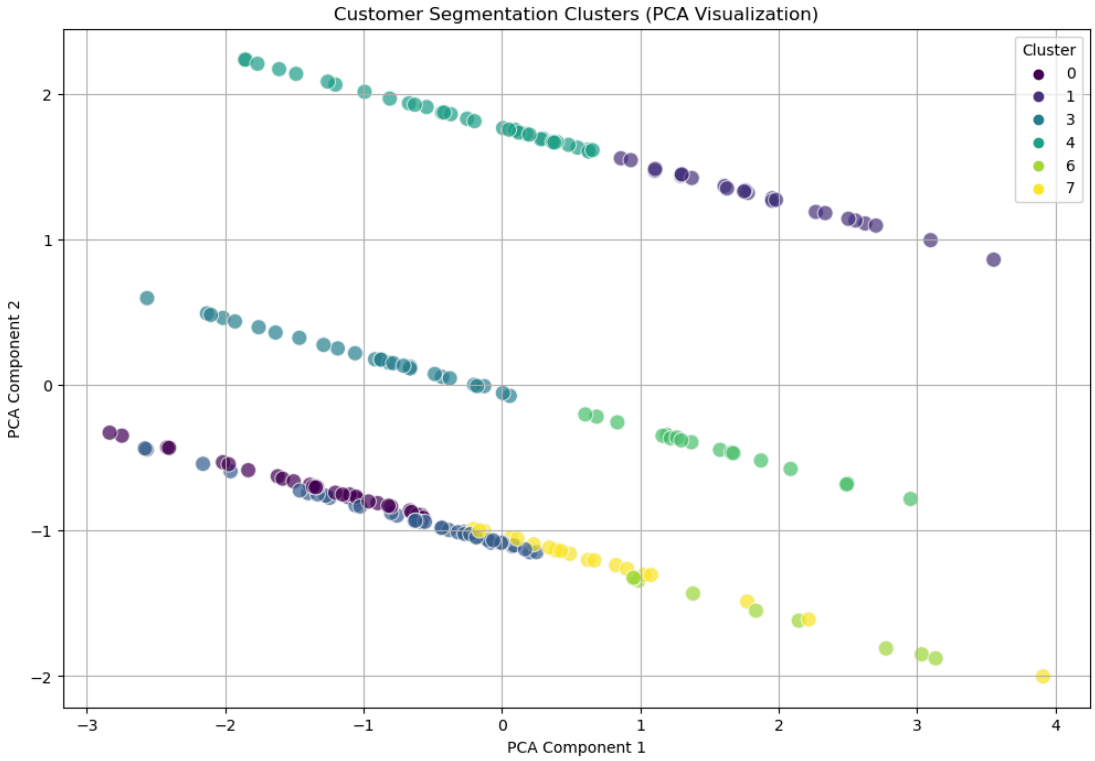
3. **Davies-Bouldin Index Plot:**
    - A plot illustrating DB Index values for different cluster counts, highlighting the optimal number of clusters with the lowest index.



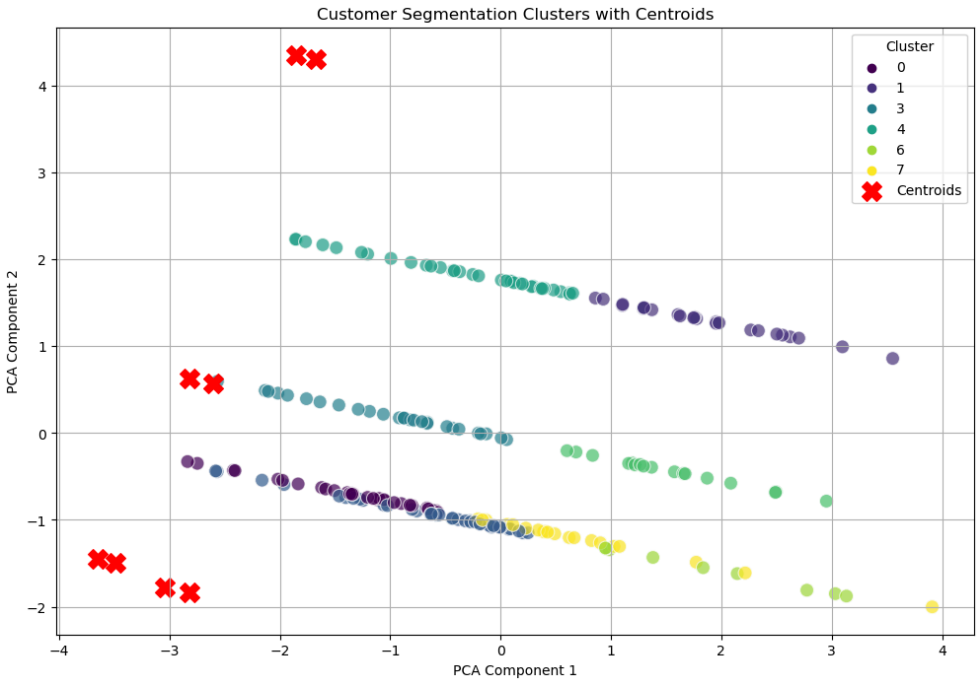




4. **PCA Visualization of Clusters:**
    - A scatter plot visualizing customer segments in two-dimensional PCA space, color-coded by cluster assignments. This visualization helps in understanding how well-defined and separated the clusters are.




## 5. Cluster Centers Visualization:

- A scatter plot similar to the PCA visualization but includes red 'X' markers representing the centroids of each cluster, providing insights into where each cluster center lies in relation to its members.



## Conclusion

The clustering analysis effectively segmented customers into distinct groups based on their purchasing behavior and demographic information. The choice of K-Means clustering, combined with thorough evaluation metrics and visualizations, provided a clear understanding of customer segments, which can inform targeted marketing strategies and business decisions.