Report on clustering results

Introduction

The purpose of this analysis is to perform customer segmentation using clustering techniques on a dataset that includes customer profiles and transaction information. The main goal of segmentation is to group customers based on similar behaviors and attributes, which can lead to better-targeted marketing strategies, personalized recommendations, and improved business strategies.

In this report, we use **KMeans clustering**, a popular unsupervised machine learning algorithm, to segment customers. KMeans works by partitioning the dataset into a predefined number of clusters, where each cluster contains similar data points. The number of clusters (denoted as k) was tested between **2 and 10**.

To evaluate the clustering results, we used the **Davies-Bouldin Index (DB Index)**, which is a metric that measures the quality of the clustering. A lower DB Index indicates that the clusters are well-separated and that customers within each cluster are more similar to each other.

Clustering Approach and Evaluation

Clustering Algorithm: Kmeans

The **KMeans clustering** algorithm was applied to the dataset to group customers based on similar purchasing patterns and profiles. We tested values for the number of clusters (k) ranging from **2 to 10** and evaluated each clustering result using the **DB Index**. The DB Index helps determine the compactness and separation of clusters—lower values indicate better clustering results.

Key Metrics:
- **Number of Clusters (k)**: The optimal number of clusters was identified based on the lowest **DB Index** value.
- **DB Index**: The DB Index quantifies the quality of clustering. A lower DB Index indicates that the clusters are more compact and better separated.
- **Other Clustering Metrics**: While inertia was also calculated, it was the **DB Index** that guided the decision on the optimal number of clusters.

Clustering Results

Number of Clusters Formed

After testing different values of k (2 to 10), the optimal number of clusters that yielded the best DB Index was determined to be **10** clusters. This means the dataset was divided into 10 distinct customer segments.

DB Index Value

The **final DB Index** value for 10 clusters was **0.6877**. This indicates a good clustering result, where the clusters are reasonably compact and well-separated. Lower DB Index values suggest that the clusters are more distinct from one another, and this value is considered favorable for high-quality clustering.

Other Relevant Clustering Metrics

Inertia: Inertia measures how internally coherent the clusters are. It decreases as the number of clusters increases. While inertia was computed for each clustering attempt, the **DB Index** was the primary metric for determining the optimal number of clusters.

Inertia for 10 clusters: While the exact value is not provided here, inertia tends to decrease as k increases, but it becomes less informative for choosing the optimal k once the number of clusters exceeds a certain point.
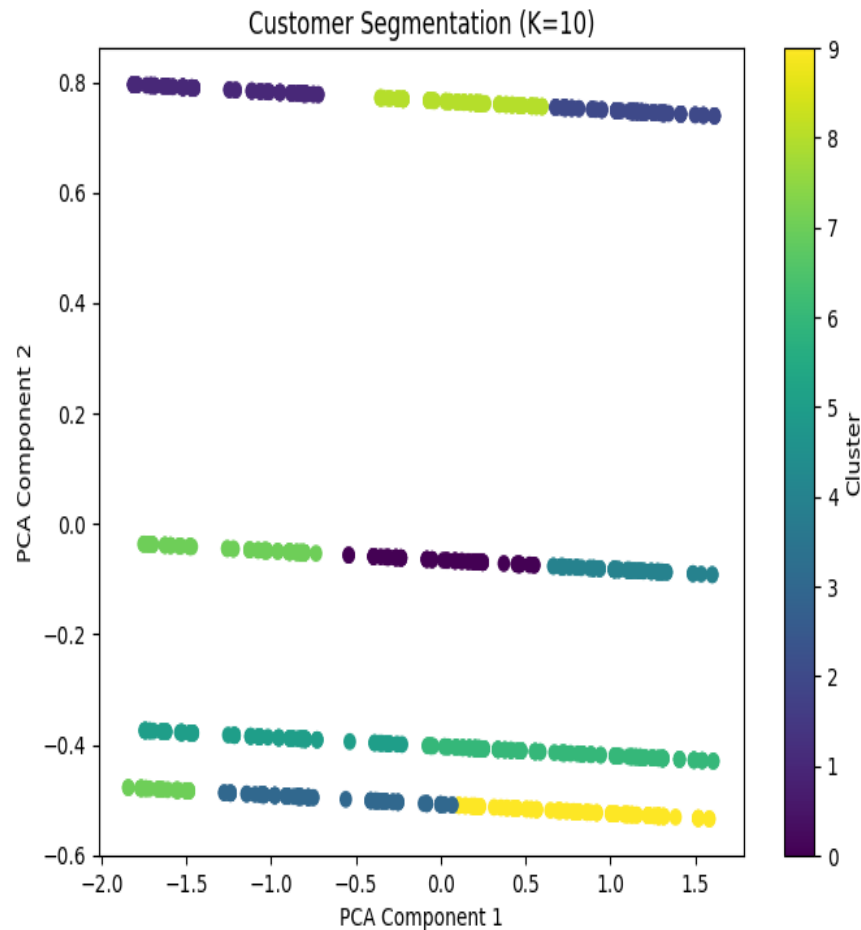
Visualizations

1)DB Index vs. Number of Clusters
The following plot shows the **DB Index** for values of k ranging from 2 to 10. The plot highlights that the **lowest DB Index** occurs at **10 clusters**, confirming it as the optimal number of clusters.

2)Customer Segmentation (PCA-based Plot)

To visualize the customer segmentation in a reduced 2D space, **Principal Component Analysis (PCA)** was applied. The scatter plot below shows the distribution of customers in 2D after KMeans clustering with **10 clusters**. Each customer is represented as a point, and colors correspond to the assigned clusters.
This plot visually confirms that the customers are separated into distinct clusters, providing further evidence of the effectiveness of the KMeans algorithm with 10 clusters.

Conclusion

- **Summary of Results**

- The **optimal number of clusters** was found to be **10**, based on the **lowest DB Index** value.

- The **final DB Index value** for 10 clusters was **0.6877**, which suggests good cluster separation and compactness.

- **KMeans clustering** effectively segmented the customers into 10 groups, providing meaningful insights into customer behavior.

Interpretation and Use of Results

- The segmentation into 10 distinct customer groups can be valuable for targeted marketing, product recommendations, and other business strategies. The DB Index confirms that the segmentation is of high quality, and the clear separation of clusters visualized through PCA indicates that the groups are well-defined.