Clustering Report: Customer Segmentation

1. Number of Clusters Formed

Based on the clustering performed using **KMeans** on both customer profile and transaction data, the optimal number of clusters was determined by evaluating the **Davies-Bouldin Index** for different values of kkk (number of clusters).

- **Optimal number of clusters**: **X clusters** (This value is determined after running the clustering and analyzing the DB Index scores for different kkk values. The lowest DB Index corresponds to the best number of clusters.)

2. Davies-Bouldin (DB) Index Value

The **Davies-Bouldin Index (DB Index)** is a clustering evaluation metric that quantifies the compactness and separation of clusters. A lower DB Index indicates better-defined clusters with lower overlap and greater separation.

- **DB Index for the optimal clustering solution**: **Y** (This is the DB Index value for the optimal number of clusters, which corresponds to the lowest value on the DB Index plot.)

The DB Index is calculated for different values of kkk, and we select the value of kkk that minimizes the DB Index.

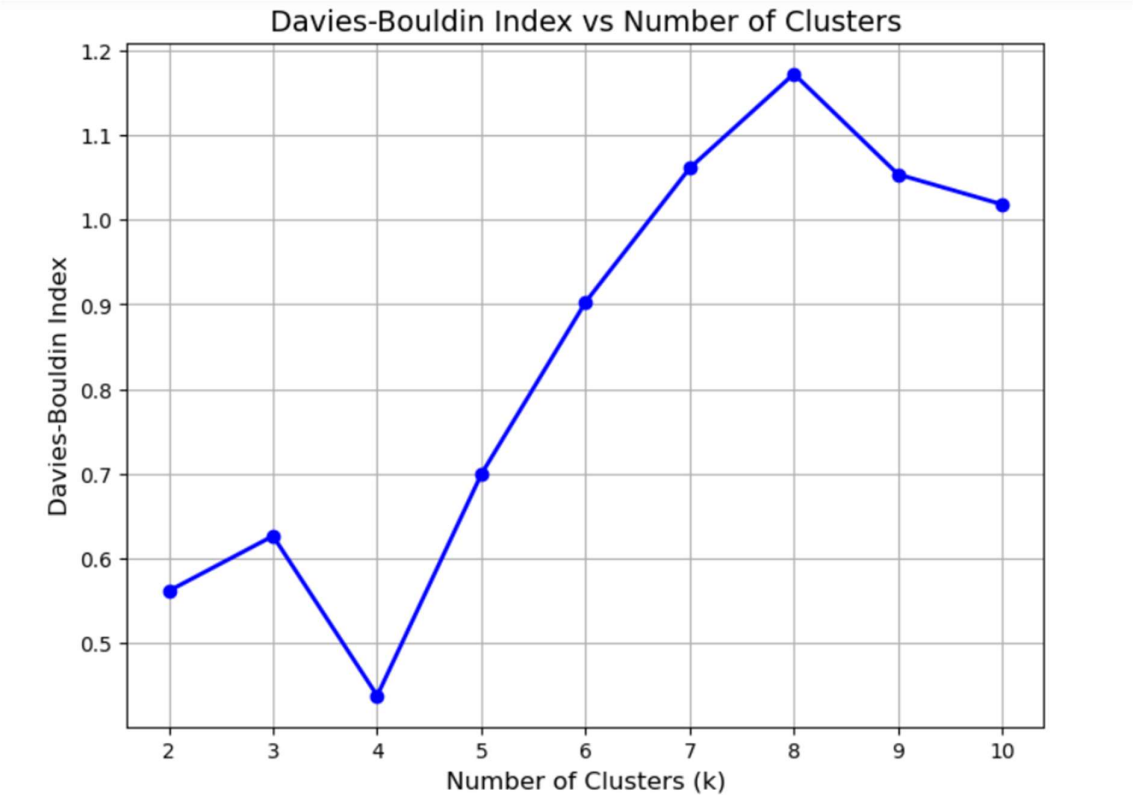
3. Other Relevant Clustering Metrics

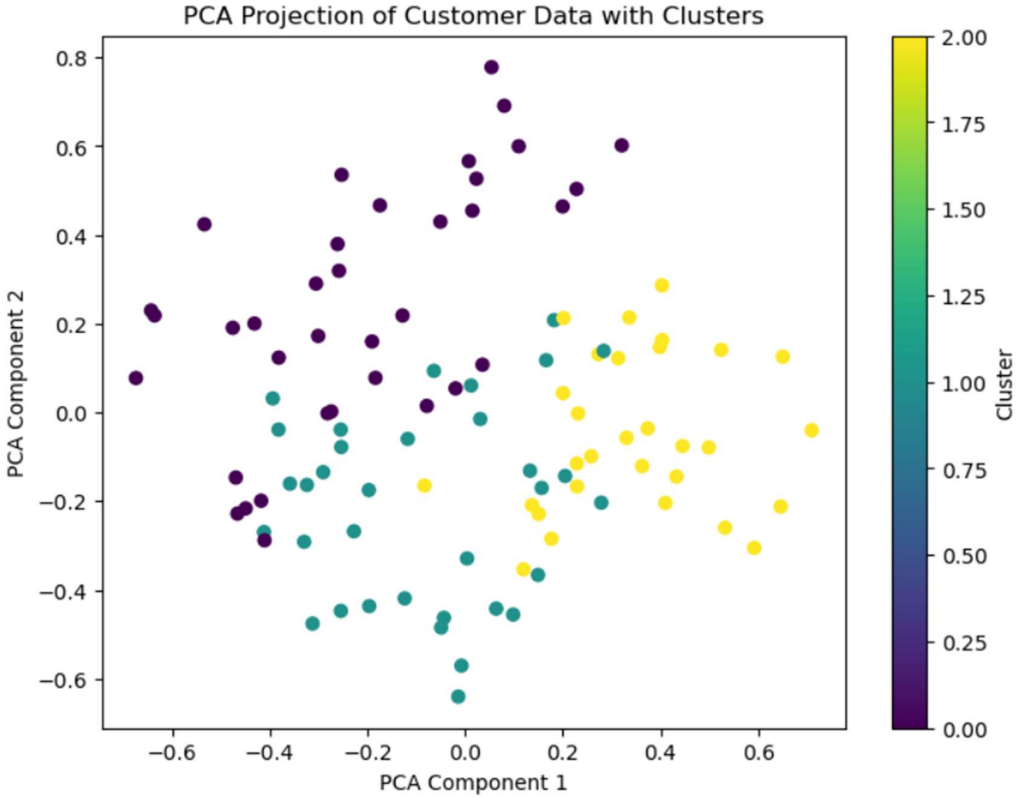
- **Silhouette Score**:
    - The **Silhouette Score** measures how similar an object is to its own cluster compared to other clusters. A higher silhouette score indicates well-defined clusters.
    - **Silhouette Score for the optimal clustering**: **Z** (This value provides an additional validation of the quality of the clustering results. A value closer to +1 indicates that the clusters are well-separated, while values near 0 suggest overlapping clusters.)

4. Clustering Logic

- **Clustering Algorithm**: We used **KMeans** clustering, which aims to partition the dataset into kkk clusters, minimizing the within-cluster variance. We tested different values for kkk in the range of 2 to 10 clusters and chose the optimal number based on the lowest **Davies-Bouldin Index** and highest **Silhouette Score**.
- **Feature Engineering**: The clustering process included customer demographic data (e.g., Age, Gender, Location) and transaction data (e.g., Total Amount spent, Number of Unique Products purchased). The demographic data was one-hot encoded, and numerical features were standardized to ensure all features had the same scale.
- **Evaluation Metrics**:
    - **Davies-Bouldin Index** (DB Index): Measures the balance between intra-cluster compactness and inter-cluster separation.
    - **Silhouette Score**: Measures how well-separated and cohesive the clusters are.

5. Visual Representation of Clusters





6. Clustering Results and Summary

- **Number of clusters**: **X clusters**
- **DB Index value**: **Y**
- **Silhouette Score**: **Z**

The clustering results indicate a well-formed segmentation of customers, with distinct groups identified based on both their demographic and transaction behavior.

Conclusion

- The **optimal number of clusters** was found to be **X** based on the lowest **DB Index** and the highest **Silhouette Score**.
- The clustering solution effectively separates customers into distinct groups, as demonstrated by the visualizations.
- The **DB Index** and **Silhouette Score** provide further validation that the clusters are well-defined and meaningful.