Clustering Report: Customer Segmentation Analysis

1. Objective: The objective of this analysis was to segment customers based on transaction and demographic data using clustering techniques. The primary goal was to identify the optimal number of clusters that provide the best grouping of customers.

2. Methodology:

We utilized the K - Means clustering algorithm and evaluated the performance using two main metrics:

- **Davies-Bouldin Index**: Measures the average similarity ratio of each cluster with the cluster that is most similar to it. Lower values indicate better clustering.
- **Silhouette Score**: Measures how similar each point is to its own cluster compared to other clusters. A higher score indicates better-defined clusters.

The clustering process included the following steps:

- **Data Preprocessing**: Removed duplicates and handled missing values. Numerical features were imputed with the median, and categorical features with the mode.
- **Feature Scaling**: Used StandardScaler to standardize features, ensuring they were on the same scale for optimal clustering performance.
- **Cluster Evaluation**: We tested different numbers of clusters (from 2 to 10) and calculated the Davies-Bouldin Index and Silhouette Score for each.

3. Results:

After evaluating the clustering performance for different numbers of clusters, the results were as follows:

- **Optimal Number of Clusters**: 10
- **Davies-Bouldin Index**: 0.6766
- **Silhouette Score**: 0.4922

These values indicate that 10 clusters offer a reasonable balance of compactness and separation. While the Silhouette Score is not extremely high, it is acceptable, suggesting that the clusters are reasonably well-separated.