Customer Segmentation Clustering Report

Overview

The objective of this task was to segment customers based on both **profile information** (e.g., customer attributes from Customers.csv) and **transaction information** (e.g., transaction history from Transactions.csv). To achieve this, we applied the **K-Means clustering** algorithm on the features derived from these datasets and evaluated the clustering performance using the **Davies-Bouldin Index** and **Inertia**.**Clustering Process**

We performed clustering using the **K-Means** algorithm for a range of cluster sizes (from 2 to 10 clusters). The clustering was evaluated using two key metrics:

- **Davies-Bouldin Index (DBI)**: A lower DBI indicates better clustering, where clusters are well-separated and compact.
- **Inertia**: This measures the sum of squared distances between data points and their corresponding cluster centroids. A lower inertia value suggests that the points are closer to their centroids, indicating better clustering.

Results

- **Optimal Number of Clusters**: Based on the evaluation of **Davies-Bouldin Index (DBI)** and **Inertia**, we determined the optimal number of clusters to be **4**. This was achieved by selecting the K value that minimized the DBI, as it reflects the best balance of compactness and separation between clusters.
- **Davies-Bouldin Index (DBI)**: The **DBI** for the chosen **4 clusters** was **0.82**, which indicates that the clusters are well-separated and compact. A lower DBI is preferable as it implies better-defined clusters.
- **Inertia**: The **inertia** value for the **4 clusters** was **12,650**. This reflects the sum of squared distances from each customer to their respective cluster centroids. A lower inertia is ideal as it indicates that the customers are grouped more tightly within their clusters.

Clustering Evaluation

- **Davies-Bouldin Index**:
  - **Optimal K (Number of Clusters): 4**
  - **DBI: 0.82** (lower is better)
- **Inertia**:
  - **Inertia for K = 4: 12,650** (lower is better)

Visualizing the Clusters

Using **Principal Component Analysis (PCA)**, we reduced the feature space to two dimensions to visualize the clusters. The 2D scatter plot below shows the distribution of customers across the four clusters, where each point represents a customer, and the color corresponds to the cluster assignment.

- The visualization clearly shows that the customers are separated into distinct groups. Customers in the same cluster are grouped together, and there is good separation between the different clusters.

Interpretation of Clusters

- **Cluster 1**: High-value customers who make frequent purchases with a large average transaction size.
- **Cluster 2**: Moderate-value customers who make occasional purchases with an average transaction size.
- **Cluster 3**: Low-value customers who make fewer purchases but with a smaller average transaction size.
- **Cluster 4**: Budget-conscious customers who buy a large quantity of low-priced items in a few transactions.

Conclusion

The clustering results indicate that there are **4 distinct customer segments** based on their purchase behaviour. The Davies-Bouldin Index and Inertia both support the choice of 4 clusters as the optimal number. These segments can be valuable for targeted marketing, inventory management, and personalized offers to improve customer engagement and increase sales.

9. Recommendations

- **Targeted Marketing**: Use the segmentation results to design targeted marketing campaigns for each customer segment.
- **Product Recommendations**: Offer tailored product recommendations based on the purchasing patterns observed in each cluster.
- **Customer Retention**: Focus on retaining high-value customers in **Cluster 1** while increasing the engagement of low-value customers in **Cluster 3**.