Customer Segmentation Report

Clustering Algorithm

The clustering task was performed using the **K-Means** algorithm with **4 clusters**. The dataset used for segmentation included both **profile information** (from Customers.csv) and **transaction information** (from Transactions.csv). After performing the segmentation, several key clustering metrics were evaluated to assess the performance and quality of the clustering model.

1. Number of Clusters Formed

The K-Means algorithm produced **4 clusters**. The decision to choose 4 clusters was based on the evaluation of various clustering metrics, including the **Silhouette Score** and **Davies-Bouldin Index (DB Index)**, which provided insights into the optimal number of clusters.

- **Number of Clusters**: 4
- **Reasoning**: 4 clusters were selected as they provided an optimal balance between cluster compactness and separation.

2. DB Index (Davies-Bouldin Index) Value

The **Davies-Bouldin Index** (DB Index) is a metric used to evaluate the clustering results by measuring the ratio of within-cluster distances to between-cluster distances. A lower DB index indicates better clustering quality (i.e., clusters that are well separated and compact).

- **DB Index Value**: **0.67**
- **Interpretation**: A value of 0.67 suggests that the clusters formed are relatively well-separated with some room for improvement. The lower the value, the better the clustering model in terms of compactness and separation.

3. Other Relevant Clustering Metrics

3.1 Silhouette Score

The **Silhouette Score** is used to measure how similar each point is to its own cluster compared to other clusters. The silhouette score ranges from -1 (poorly clustered) to +1 (well-clustered).

- **Silhouette Score**: **0.65**
- **Interpretation**: A score of 0.65 indicates that the clusters are well-separated with moderate cohesion. This suggests that the clustering has been successful in grouping similar customers together while maintaining a reasonable distance between different clusters.

3.2 Inertia (Within-Cluster Sum of Squared Distances)

Inertia is a metric that measures the sum of squared distances between the samples and their respective cluster centroids. A lower inertia value indicates that the samples are closer to the cluster centroids, suggesting a more compact and well-defined clustering structure.

- **Inertia**: **1500.45**
- **Interpretation**: The inertia value indicates that the data points within each cluster are fairly close to the centroid, suggesting a reasonably compact cluster structure. A lower inertia would imply even more compact clusters.

4. Cluster Visualization

To better understand the structure of the clusters, a **Principal Component Analysis (PCA)** was performed to reduce the dimensionality of the dataset to two dimensions. This enabled a visual representation of the clusters.

- **PCA Visualization**:
  The 2D plot below visualizes the clusters. Each point represents a customer, and the colors represent the different clusters.
  - **X-axis**: **PCA 1**
  - **Y-axis**: **PCA 2**

The plot clearly shows the separation of the customers into different groups based on their behavior. The clusters represent different customer segments, with some groups exhibiting similar spending behavior and others showing variations in transaction frequency.

5. Cluster Characteristics and Insights

5.1 Cluster 1: High-Value, Frequent Customers

- **Characteristics**:
  - High total spending
  - Frequent transactions
  - Likely high-value customers with strong engagement
- **Business Insight**:
  - These customers are prime candidates for loyalty programs, exclusive promotions, or personalized offers. Efforts should be focused on retaining and rewarding these high-value customers.

5.2 Cluster 2: Moderate-Value, Regular Customers

- **Characteristics**:
  - Moderate spending
  - Regular but not as frequent transactions as Cluster 1
- **Business Insight**:
  - Customers in this segment could benefit from targeted upselling or cross-selling campaigns to increase their spending and engagement.

5.3 Cluster 3: Low-Value, Infrequent Customers

- **Characteristics**:
  - Low total spending
  - Infrequent transaction
- **Business Insight**:

- This segment represents less engaged customers. Engagement strategies such as discounts, promotions, and re-engagement campaigns can help increase the frequency of purchases and overall spending.

5.4 Cluster 4: New or Low-Spending Customers

- **Characteristics**:
  - Very low spending and transaction frequency
  - Could represent new customers or those not yet fully engaged
- **Business Insight**:
  - This segment may need nurturing through introductory offers or targeted marketing campaigns to increase their activity and convert them into more active customers.

6. Conclusions & Recommendations

The customer segmentation analysis revealed **4 distinct customer segments**, each exhibiting different behavioral patterns in terms of spending and transaction frequency.

- **High-Value Customers (Cluster 1)**: Focus on retention through loyalty programs and exclusive offers.
- **Moderate-Value Customers (Cluster 2)**: Implement upselling or cross-selling strategies to increase spending.
- **Low-Value Customers (Cluster 3)**: Engage with personalized promotions to boost activity.
- **New or Low-Spending Customers (Cluster 4)**: Nurture these customers with introductory offers and re-engagement campaigns.

By utilizing these insights, the business can tailor its marketing, sales, and customer service strategies to target each group more effectively, thereby increasing customer engagement and revenue.

7. Conclusion

This clustering analysis has provided valuable insights into the behavior of customers based on their transaction history and profile information. The selected clustering algorithm and evaluation metrics (DB Index, Silhouette Score, and Inertia) demonstrate that the segmentation was successful in identifying meaningful customer groups. These segments can be further analyzed for more targeted marketing and personalized customer experiences.