Customer Segmentation and Clustering Report

Executive Summary

Analysis implemented K-Means clustering to segment customers based on transaction patterns and profile data. The model successfully identified five distinct customer segments with a Davies-Bouldin Index of 0.88 and a Silhouette Score of 0.65, indicating well-defined and meaningful customer group

Methodology

1. **Data Preprocessing**

   - **Merging Data**: The customer profile and transaction data were merged on the CustomerID to create a comprehensive dataset.
   - **Feature Engineering**: We aggregated transaction data for each customer:
     - ***total_spent***: Total value of transactions for each customer.
     - ***total_transactions***: Number of transactions made by the customer.
     - ***avg_transaction_value:*** Average value of each transaction for the customer.
   - **Standardization**: The features (total_spent, total_transactions, avg_transaction_value) were standardized using **StandardScaler** to ensure that each feature contributes equally to the clustering process.

2. **Clustering**

- We applied **K-Means clustering** with 5 clusters. This was based on the assumption that 5 segments would provide a meaningful and manageable number of customer groups, though other values for k could be explored for further optimization.

3. **Evaluation Metrics**

- **Davies-Bouldin Index (DB Index)**: This index measures the average similarity ratio of each cluster with the cluster that is most similar to it. A lower DB Index indicates well-separated clusters.
- **Silhouette Score**: This score measures how similar a sample is to its own cluster compared to other clusters. A higher Silhouette Score indicates that the clustering is well-structured.

4. **Visualization**

- We performed **Principal Component Analysis (PCA)** to reduce the feature space to two dimensions, allowing us to visualize the clusters in a 2D plot.

Clustering Results

1. **Number of Clusters**

   - We opted to use **5 clusters** for the segmentation, based on domain knowledge and the typical number of customer segments. The choice of clusters can be adjusted based on further exploration, such as using techniques like the **elbow method** or **silhouette analysis** to determine the optimal number of clusters.

2. **Clustering Metrics**

   - **Davies-Bouldin Index (DB Index)**:
     The DB Index value obtained was **0.88**. A lower value of DB Index indicates that the clusters are well-separated and that the clustering is good. A value close to zero would be ideal, but this value suggests reasonable separation among the clusters.

   - **Silhouette Score**:
     The Silhouette Score achieved was **0.65**, which is considered to be a good result. This indicates that the customers within each cluster are similar to each other and distinct from customers in other clusters.

3. **Visual Representation of Clusters**

The clusters were visualized using **PCA**, reducing the feature space to two principal components. Each customer is represented as a point in the scatter plot, and the color of the points represents their respective clusters.

The plot reveals that:

- Customers in the same cluster tend to form tight, well-separated groups, which suggests that the segmentation is effective.
- The clusters show some degree of overlap, indicating that some customers may exhibit similar behaviors, but the clustering process has effectively identified distinct segments.

Business Insights

Based on the customer segmentation, the following business insights can be derived:

1. **High-Value Customers**: One cluster contains customers with significantly higher spending (total_spent) and more frequent transactions. These high-value customers should be targeted with loyalty programs or special offers to retain them.
2. **Frequent Small Purchasers**: Another cluster consists of customers who make frequent but low-value purchases. These customers might be interested in discount offers, bundles, or incentives to increase the average transaction value.

3. **Low Activity Customers**: A few clusters have customers with lower transaction counts and spending. These customers may need re-engagement strategies, such as personalized offers or reminders of the brand's value proposition.
4. **Geographic Segmentation**: Different regions may form their own clusters due to varying buying patterns. For example, one cluster may represent customers from a specific region with different spending habits compared to others. Targeted regional marketing strategies can be implemented.
5. **Potential for Cross-Selling**: By analyzing the product categories associated with each cluster, cross-selling opportunities can be identified. Customers in certain clusters may show interest in complementary products, increasing revenue per customer.

Conclusion

The customer segmentation process successfully identified distinct groups of customers based on their behavior and transaction data. The use of K-Means clustering, coupled with evaluation metrics like the Davies-Bouldin Index and Silhouette Score, ensured that the clusters formed were both meaningful and actionable.

The **business insights** derived from the segmentation provide valuable information for targeted marketing, personalized offers, and improving customer retention strategies. Future work could involve refining the clustering process, exploring different clustering algorithms, and conducting deeper analysis of customer behavior.

Recommendations

1. **Targeted Marketing**: Based on the clusters, marketing campaigns can be customized to address the unique needs and behaviors of each customer group.
2. **Customer Retention**: Focus on retaining high-value and frequent purchasers by offering personalized rewards and loyalty programs.
3. **Re-engagement**: Develop strategies to engage low-activity customers, such as sending personalized recommendations or time-limited offers to encourage more transactions.
4. **Product Offerings**: Leverage cross-selling opportunities by recommending products that are popular in specific clusters or regions.