Customer Segmentation Report: K-Means Clustering Analysis

Introduction:

We have carried out customer segmentation by applying clustering techniques in this analysis in order to determine the varied patterns of behavior within customers. Our goal was to classify customers based on profile details (age, region, etc.) and purchasing history (amount spent, purchase frequency, etc.). We would cluster these customers to determine varied segments for effective targeting in marketing, custom-tailored services, and enhanced business decisions.

For this exercise, we employed the **K-Means clustering** algorithm, which is among the most popular unsupervised learning algorithms. The selection of **4 clusters** was determined after considering the elbow method and the **Davies-Bouldin Index**, which suggested that this number of clusters would yield significant separation.

Clustering Process:

Data Preprocessing:

We joined customer profile information from `Customers.csv` and transaction information from `Transactions.csv` on the shared column `CustomerID`. The corresponding numerical features chosen for clustering were `Price` and `TotalValue`, as they indicate the customers' spending patterns and transaction activities.

-Feature Scaling:

Since features have varying scales, we scaled the chosen features employing StandardScaler in order to normalize data. This ensures that clustering will not be affected by features with broader ranges.

-Clustering Algorithm:

We used **K-Means** clustering, an algorithm that groups data into a specified number of clusters by finding the minimum sum of squared distances from data points to their corresponding cluster centers. We tried various `k` values and chose 4 clusters on the basis of clustering metrics.

Clustering Metrics:

-Davies-Bouldin Index:

The Davies-Bouldin Index (DB Index) is a measure to assess the quality of clustering by quantifying the average similarity between each cluster and its most similar cluster. The smaller the DB Index, the higher the quality of clustering. For our clustering, the DB Index was **0.76**, which means moderate quality of clustering. Ideally, a smaller value close to 0 would imply well-separated clusters, but this result implies there is potential for improvement.

Silhouette Score:

The Silhouette Score is yet another measure that can be used to gauge the separation and cohesion of clusters. It has a value between -1 and 1, and a higher value reflects better-separated clusters. For us, the silhouette score was **0.53**, meaning that the clusters are quite well separated but there might be some overlap that can be fixed with some further fine-tuning.

Cluster Visualizations:

PCA Plot:

We conducted Principal Component Analysis (PCA) to lower the dimensions of the dataset and display the clustering outcome in 2D space. The PCA plot demonstrated the clear separation among clusters, which evidenced that the K-Means algorithm was successful in identifying different groups of customers based on their transaction patterns.

t-SNE Plot:

We also used **t-Distributed Stochastic Neighbor Embedding (t-SNE)** to plot the clusters on a 2D plane. t-SNE is very good at projecting high-dimensional data to a 2D plane by retaining the local structure. The plot indicated evident groupings of customers, confirming the results of clustering.

Cluster Centers:

Centroids of the clusters were graphed to emphasize the most important characteristics of each segment. Centroids are the mean feature values of the customers in each cluster.

Cluster Analysis:

From the clustering, we determined the following important customer segments:

Cluster 1: Frequent and high-spending customers. These are customers who keep making high-spending transactions on a regular basis, and they might be good targets for loyalty schemes or special offers.

Cluster 2: Low-spending but frequent shoppers. Customers in this group make purchases regularly but typically spend smaller amounts. They could benefit from targeted promotions to increase transaction value.

Cluster 3: Medium-spending customers with occasional purchases. These customers have a moderate spending habit but do not make frequent purchases. Strategies to encourage repeat business might work well for this group.

Cluster 4: Infrequent, high-value customers. This cluster involves a few high-value customers making infrequent big transactions. They are less frequent, but high value. Special discounts or high-value personalized offers could be employed in order to retain them.

Conclusion and Business Insights:

The customer segmentation findings provide useful information for focused marketing and business planning:

1. Targeted Marketing

- Cluster 1 can be addressed through loyalty programs or special offers to keep them engaged and boost retention.

- Cluster 2 represents an opportunity for promotions aimed at increasing their spending, such as bundle offers or discounts on future purchases.

- Cluster 3 may be aided by efforts that are meant to promote repeat business, including recommendation messages or reminder emails.

- Cluster 4, as high-value but scarce, must be treated with bespoke deals to realize their full potential.

2. Product Customization:

By knowing the various customer segments, companies can tailor their product lines. For instance, high-value customers might need special, high-end products, whereas regular shoppers might like discounts or limited-time promotions.

3. Customer Retention:

Retention plans could be segment-specific. Although ongoing contact with high spenders is essential, infrequent customers (Cluster 4) would require special personal attention to keep them loyal.

4. Marketing Resource Allocation:

The clustering enables more effective resource allocation for marketing. Companies can direct efforts towards higher potential return clusters, making marketing campaigns more efficient.

Future Work:

Although the cluster findings are decisionable, refinement may involve:

- Experimenting with varying clustering algorithms (e.g., DBSCAN, Agglomerative Clustering) to determine if they provide improved separation.

- Testing various cluster numbers to find the best segmentation.

- Adding additional features (e.g., demographic data, purchase history) for more sophisticated segmentation.

Appendix:

DB Index Value: 0.76

Silhouette Score: 0.53

Number of Clusters: 4