Report on Clustering Results

Introduction:

This report presents the results of customer segmentation performed using clustering techniques. The analysis was conducted to understand customer groups better and tailor marketing strategies accordingly. Both profile information (from Customers.csv) and transaction information (from Transactions.csv) were used in the clustering process.

Number of Clusters Formed:

The K-Means clustering algorithm was applied to segment the customers into 5 distinct clusters. The choice of 5 clusters was determined based on the analysis of clustering metrics and visualizations.

DB Index Value

The Davies-Bouldin Index (DB Index) is a metric used to evaluate the clustering performance. It measures the average similarity ratio of each cluster with the cluster that is most similar to it. A lower DB Index indicates better clustering. For this analysis, the DB Index value is **0.9344500774507883**. This value suggests that the clusters formed are reasonably well-defined.

Silhouette Score

The Silhouette Score is another metric used to assess the quality of the clusters. It measures how similar an object is to its own cluster compared to other clusters. The score ranges from -1 to 1, where a higher score indicates better-defined clusters. The Silhouette Score for this analysis is **0.3221539961837519**. This score indicates moderate clustering quality, with some overlap between clusters.

Other Relevant Clustering Metrics

The Within-Cluster Sum of Squares (WCSS) measures the sum of squared distances from each point to the centroid in a cluster. Lower WCSS values indicate more compact clusters. The WCSS values for different numbers of clusters are as follows:

- For 1 cluster: **597.0000000000001**
- For 2 clusters: **346.6724462401302**

- For 3 clusters: **238.90579183627435**
- For 4 clusters: **186.57672834998087**
- For 5 clusters: **158.6801917739819**
- For 6 clusters: **133.9805098110021**
- For 7 clusters: **108.33016904701638**
- For 8 clusters: **100.73157565124612**
- For 9 clusters: **88.00127003049502**
- For 10 clusters: **83.61839849863799**

Visualizations

Scatter Plot of Clusters

The scatter plot titled **"Customer Segmentation"** illustrates the relationship between **Total Value (USD)** and **Quantity**. The X-axis represents the total value in USD, ranging from 0 to 10,000, and the Y-axis represents the quantity, ranging from 0 to around 35. Data points are colored according to five different clusters (0 to 4), represented in the legend:

- Cluster 0 (Purple)
- Cluster 1 (Blue)
- Cluster 2 (Teal)
- Cluster 3 (Yellow)
- Cluster 4 (Green)

There is an upward trend, indicating that as the total value increases, the quantity also tends to increase. Different clusters show varying densities; for instance, Cluster 0 (purple) appears more concentrated at lower total values, while Clusters 3 (yellow) and 4 (green) are more prominent at higher values. This plot helps in visualizing customer segments based on their purchasing behavior, illustrating how different groups of customers might differ in terms of quantity purchased relative to their total spending.

Pair Plot of Clusters

The pair plot titled **"Pair Plot of Clusters"** visualizes the relationships between **Total Value**, **Quantity**, and **Signup Date**, segmented by different clusters. The plot contains multiple subplots showing pairwise relationships, with diagonal plots showing the distribution of

each variable. Data points are categorized into five clusters (0 to 4), illustrated with different colors:

- Cluster 0 (Light pink)
- Cluster 1 (Pink)
- Cluster 2 (Light purple)
- Cluster 3 (Dark purple)
- Cluster 4 (Dark blue)

A distinct linear relationship suggests that higher total values correlate with greater quantities purchased. Different clusters exhibit varying densities, with Cluster 3 (dark purple) particularly prevalent at higher values and quantities. There are distinct trends over time, with peaks indicating periods when customers tended to sign up and make purchases, pointing to seasonal purchasing behaviors or promotional campaigns. This plot reveals how the quantity purchased changes over time, helping identify periods of peak purchasing activity among various clusters. Cluster densities vary, suggesting differing engagement levels based on signup dates.

Silhouette Plot

The silhouette plot visualizes the silhouette coefficients for various clusters, helping to evaluate the clustering quality. The Y-axis lists the different cluster labels (0 through 4), and the X-axis ranges from 0.0 to 0.6, indicating the value of the silhouette coefficients. Each colored bar corresponds to a cluster, with colors representing specific clusters:

- Cluster 0 (gray)
- Cluster 1 (blue)
- Cluster 2 (cyan)
- Cluster 3 (green)
- Cluster 4 (orange)

The length of each bar represents the range of silhouette coefficients for that cluster, with longer bars indicating better-defined clusters. A vertical red dashed line at approximately 0.3 represents the average silhouette coefficient across all clusters, serving as a reference to evaluate the clustering quality. A value closer to 1 indicates better-defined clusters, while values around 0 suggest overlapping clusters. Overall, the plot effectively visualizes the separation of clusters, providing insight into the clustering structure.

Elbow Method Plot

The elbow method plot helps determine the optimal number of clusters based on the Within-Cluster Sum of Squares (WCSS). The X-axis, labeled "Number of Clusters," ranges from 1 to 10, indicating the different numbers of clusters being assessed, while the Y-axis, labeled "WCSS," shows the Within-Cluster Sum of Squares, with values ranging from just below 100 to 600. The plot shows data points for each number of clusters, with a line connecting them. There is an initial sharp decrease in WCSS as the number of clusters increases from 1 to about 4, after which the decrease becomes more gradual, suggesting diminishing returns in reducing WCSS with additional clusters. The "elbow" point around 4 clusters indicates the optimal number of clusters, where the rate of decrease in WCSS slows down significantly. This point suggests that adding more clusters beyond this may not provide significantly better clustering. Overall, the graph effectively visualizes the relationship between the number of clusters and WCSS, helping to identify the optimal cluster count for the dataset.

Conclusion

The clustering analysis provides valuable insights into customer groups, which can help in tailoring marketing strategies and improving customer engagement. The Davies-Bouldin Index value and visualizations indicate the effectiveness of the clustering algorithm in segmenting the customers. By understanding customer segments better, businesses can develop targeted marketing strategies, improve customer retention, and enhance overall customer satisfaction.