25/01/2025

Task 03: Customer Segmentation Report

Introduction:

Customer segmentation is a key technique in data analysis used to divide a customer base into distinct groups or clusters based on shared characteristics, behaviors, or purchasing patterns. It allows businesses to tailor marketing strategies, optimize resource allocation, and provide personalized customer experiences. This report presents the results of customer segmentation performed using **K-Means clustering**, a popular unsupervised learning algorithm.

Theory of Clustering:

Clustering is an unsupervised machine learning technique used to group similar data points together. Unlike supervised learning, where the model is trained on labeled data, clustering algorithms try to find natural groupings within the data without predefined labels.

There are several clustering algorithms, and **K-Means** is one of the most widely used. The algorithm aims to partition a set of data points into **K clusters** by minimizing the variance within each cluster. It does this iteratively by:

1. Assigning each data point to the nearest centroid (representing the cluster).

2. Recalculating the centroids based on the mean of the data points in the cluster.

3. Repeating steps 1 and 2 until convergence (when assignments no longer change).

Clustering Results:

The clustering was performed on customer data using K-Means, and the key outcomes are as follows:

- Number of Clusters Formed: 10

- Davies-Bouldin Index (DB Index): 1.247 (A measure of the quality of clustering, where lower values indicate better separation between clusters.)

K-Means Clustering:

K-Means clustering is widely used because it is relatively simple and efficient, especially for large datasets. In this case, the K-Means algorithm successfully divided the customers into **10 clusters, based on their purchasing behavior and features such as TotalValue (total spend) and Quantity (total items purchased). The algorithm was run with 10 clusters, which was found to be optimal based on the Davies-Bouldin Index.

Davies-Bouldin Index (DB Index):

The Davies-Bouldin Index (DB Index) is a metric used to evaluate the quality of clusters produced by a clustering algorithm. It is defined as the average similarity ratio of each cluster with the cluster that is most similar to it.

A lower Davies-Bouldin Index indicates better separation between the clusters, meaning that the clusters are more distinct. In this case, the Davies-Bouldin Index for the 10 clusters is **1.247**, indicating a moderate level of cluster overlap, which is typical in customer segmentation tasks.

Cluster Distribution:

The distribution of customers across the 10 clusters is as follows:

- Cluster 0: 8 customers (4.02%)
- Cluster 1: 19 customers (9.55%)
- Cluster 2: 23 customers (11.56%)
- Cluster 3: 20 customers (10.05%)
- Cluster 4: 27 customers (13.57%)
- Cluster 5: 24 customers (12.06%)
- Cluster 6: 16 customers (8.04%)
- Cluster 7: 25 customers (12.56%)
- Cluster 8: 22 customers (11.06%)
- Cluster 9: 15 customers (7.54%)

Cluster 4 is the most populous with 27 customers, while Cluster 0 has the fewest with only 8 customers.

- First CustomerID: C0001

- Last CustomerID: C0200

- Total unique CustomerIDs: 199

- Missing CustomerIDs: {'C0180'}

Key Statistical Insights:

- Total Customers: 199

- Number of Clusters: 10

- Most Populous Cluster: Cluster 4 with 27 customers

- Least Populous Cluster: Cluster 0 with 8 customers

These statistics highlight the distribution of customers across the clusters and provide insights into customer groups that may require further analysis, such as identifying high-value customers or potential outliers.

Conclusion:

The K-Means clustering algorithm successfully segmented the customer base into **10 clusters** based on customer behavior. The **Davies-Bouldin Index value of 1.247** suggests that the clusters are reasonably well-separated, though there is some overlap. This clustering can be used to understand customer profiles and create more targeted marketing or customer engagement strategies.

The analysis can be further refined by experimenting with additional features (e.g., recency, frequency, or monetary value of purchases) or other clustering algorithms like **DBSCAN** or **Agglomerative Clustering** for potentially better segmentation.

Prepared by: ***Kaif Tokare***