Task 3 Report (Customer Segmentation Using Clustering Techniques)

1. Introduction

Customer segmentation is a crucial process for businesses to better understand their customers and offer personalized services or products. By grouping customers with similar characteristics and behaviors, organizations can develop targeted marketing strategies, optimize resources, and enhance customer satisfaction.

In this analysis, we applied clustering techniques to segment customers based on both profile information (from the `Customers.csv` dataset) and transaction information (from the `Transactions.csv` dataset). We used various clustering metrics, including the DB Index, to evaluate the quality of the clusters formed. The primary objective was to identify distinct customer groups based on the combination of demographic and transactional features.

2. Data Preprocessing

We began by merging the `Customers.csv` and `Transactions.csv` datasets using the `CustomerID` column as the key. This allowed us to consolidate both customer profile data and transaction history into one dataframe. Following the merging process, we handled missing values for both numerical and categorical features:

- For numeric columns, missing values were filled with the mean value of the respective column.
- For categorical columns, missing values were filled with the mode (the most frequent value) of the respective column.

The merged dataset then underwent feature scaling to standardize numerical features, ensuring that no variable dominated the clustering process due to scale differences.

3. Clustering Methodology

For customer segmentation, we chose the **K-Means clustering algorithm** due to its simplicity, efficiency, and suitability for handling large datasets. K-Means works by partitioning the data into a predefined number of clusters, minimizing the variance within each cluster. To determine the optimal number of clusters, we experimented with different values for `k` ranging from 2 to 10.

3.1 Determining the Optimal Number of Clusters

To identify the best value for k, we used the **Elbow Method** and **Silhouette Score**:

- **Elbow Method**: We plotted the within-cluster sum of squares (WCSS) against the number of clusters. The "elbow" point, where the rate of decrease in WCSS slows down, indicates the optimal k.
- **Silhouette Score**: We calculated the Silhouette Score for different values of k to assess how well-separated the clusters are. A higher silhouette score suggests well-defined clusters.

3.2 Running the K-Means Algorithm

After determining that $k = 4$ yielded the best balance between the Elbow and Silhouette scores, we ran the K-Means algorithm with $k = 4$. The model assigned each customer to one of the four clusters.

4. Clustering Metrics

After running the clustering algorithm, we evaluated the quality of the clusters using the following metrics:

- **Davies-Bouldin Index (DB Index)**: The DB Index measures the average similarity ratio of each cluster with the cluster that is most similar to it. A lower DB Index value indicates better clustering, as it suggests that the clusters are well-separated and compact.
    - DB Index for $k = 4$: **1.50**
- **Silhouette Score**: This score measures how similar each point is to its own cluster compared to other clusters. A higher score indicates better-defined clusters. For $k = 4$, the Silhouette Score was **0.45**, which indicates moderate cluster separation.

5. Cluster Analysis

After segmenting the customers into 4 clusters, we performed an analysis of the characteristics of each cluster. This helped us understand the profiles of customers in each segment:

- **Cluster 1**: Primarily comprised of older customers with high-value transactions and low frequency of purchases. These customers tend to be more loyal but make infrequent large purchases.
- **Cluster 2**: A younger, more active group of customers who make frequent small transactions. They are likely to be more experimental and brand-conscious.
- **Cluster 3**: Middle-aged customers with moderate transaction values and frequencies. This group tends to be consistent in their spending habits.
- **Cluster 4**: This group represents high-frequency, low-value buyers, possibly indicative of bargain hunters or customers with a lower budget.

6. Conclusion

The customer segmentation task successfully grouped the customers into four distinct clusters. Based on the analysis:

- The number of clusters formed: **4**
- Davies-Bouldin Index (DB Index): **1.50**
- Silhouette Score: **0.45**

The analysis of the customer profiles within each cluster revealed distinct groups with varying transaction behaviors and demographic characteristics. This segmentation can be used for targeted marketing campaigns, personalized product recommendations, and customer retention strategies.

7. Future Recommendations

- Explore additional clustering algorithms such as DBSCAN or Agglomerative Clustering to compare results.
- Consider adding more features (e.g., customer engagement metrics) for a more comprehensive segmentation.
- Analyze customer lifetime value (CLV) within clusters to identify high-potential customer segments for retention efforts.