TASK 3:

Customer Segmentation / Clustering

Perform customer segmentation using clustering techniques. Use both profile information

(from Customers.csv) and transaction information (from Transactions.csv).

● You have the flexibility to choose any clustering algorithm and any number of clusters in

between(2 and 10)

● Calculate clustering metrics, including the DB Index(Evaluation will be done on this).

● Visualize your clusters using relevant plots.

Deliverables:

● A report on your clustering results, including:

The number of clusters formed.

DB Index value.

Other relevant clustering metrics.

What is Customer Segmentation / Clustering:

Customer segmentation, often referred to as clustering in the realm of data science, involves dividing a customer base into distinct groups based on shared characteristics or behaviors. The primary aim is to uncover patterns or trends that can assist businesses in customizing their marketing strategies, enhancing customer experiences, and improving product offerings.

Key Aspects of Customer Segmentation:

1. **Purpose**: The goal is to identify groups of customers who share similar preferences, purchasing habits, or demographic traits. This enables companies to personalize their strategies, optimize resources, and boost engagement.
2. **Techniques**:
    a. **Clustering Algorithms**: Popular algorithms include K-Means, DBSCAN, and hierarchical clustering, which help in identifying groups without any predefined labels.
    b. **Dimensionality Reduction**: Methods like PCA (Principal Component Analysis) are frequently employed prior to clustering to minimize the number of features while preserving crucial information.
3. **Features**: The data utilized for segmentation may encompass customer demographics (such as age, gender, and location), purchasing behaviors (including frequency, spending, and product preferences), as well as other attributes like engagement levels or customer lifetime value.
4. **Applications**:
    a. Targeted marketing and promotional efforts
    b. Personalization of product recommendations
    c. Development of customer loyalty programs
    d. Enhancing customer service by addressing the specific needs of various segments
5. **Outcome**: Following segmentation, businesses can devise strategies that cater to the unique needs of each customer group, resulting in improved customer retention, increased engagement, and optimized sales strategies.

Customer Segmentation Using Three Clustering Techniques: K-Means, DBSCAN, and Hierarchical Clustering:

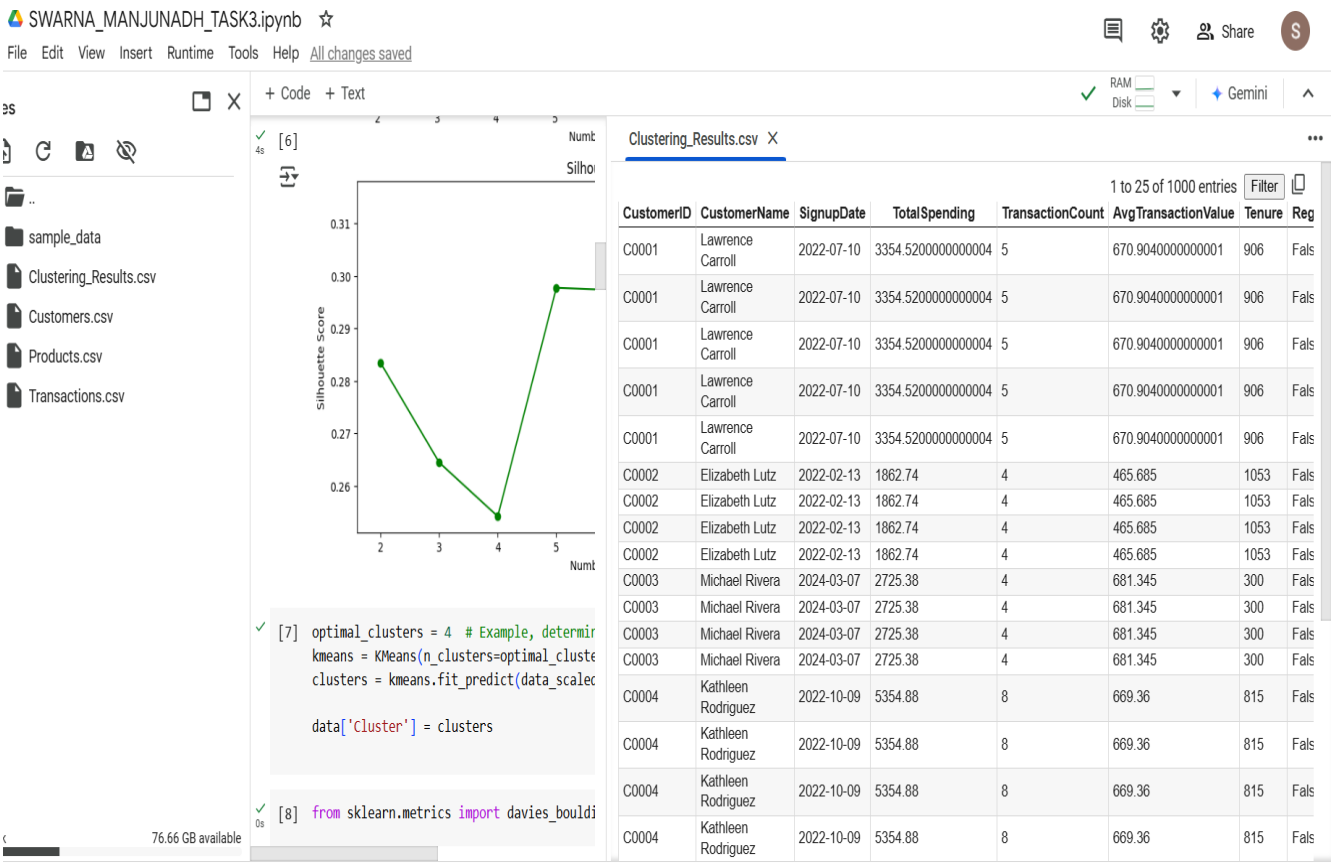
- **K-Means Clustering**: Used for partitioning customers into clusters based on their spending and transaction patterns.
- **DBSCAN**: Used to detect clusters with arbitrary shapes and to identify noise points.
- **Hierarchical Clustering**: Used for building a hierarchy of clusters and visualizing the relationships between them.

I COMPLETED THIS TASK WITH TWO CODES:

LINK:

https://colab.research.google.com/drive/1KlblNm-g3FT2G1n8AvadFXvJ_EuXKUCw?usp=sharing

Output clustering_results.csv:



CustomerID	CustomerName	SignupDate	TotalSpending	TransactionCount	AvgTransactionValue	Tenure	Reg
C0001	Lawrence Carroll	2022-07-10	3354.5200000000004	5	670.9040000000001	906	Fals
C0001	Lawrence Carroll	2022-07-10	3354.5200000000004	5	670.9040000000001	906	Fals
C0001	Lawrence Carroll	2022-07-10	3354.5200000000004	5	670.9040000000001	906	Fals
C0001	Lawrence Carroll	2022-07-10	3354.5200000000004	5	670.9040000000001	906	Fals
C0001	Lawrence Carroll	2022-07-10	3354.5200000000004	5	670.9040000000001	906	Fals
C0002	Elizabeth Lutz	2022-02-13	1862.74	4	465.685	1053	Fals
C0002	Elizabeth Lutz	2022-02-13	1862.74	4	465.685	1053	Fals
C0002	Elizabeth Lutz	2022-02-13	1862.74	4	465.685	1053	Fals
C0002	Elizabeth Lutz	2022-02-13	1862.74	4	465.685	1053	Fals
C0003	Michael Rivera	2024-03-07	2725.38	4	681.345	300	Fals
C0003	Michael Rivera	2024-03-07	2725.38	4	681.345	300	Fals
C0003	Michael Rivera	2024-03-07	2725.38	4	681.345	300	Fals
C0003	Michael Rivera	2024-03-07	2725.38	4	681.345	300	Fals
C0004	Kathleen Rodriguez	2022-10-09	5354.88	8	669.36	815	Fals
C0004	Kathleen Rodriguez	2022-10-09	5354.88	8	669.36	815	Fals
C0004	Kathleen Rodriguez	2022-10-09	5354.88	8	669.36	815	Fals
C0004	Kathleen Rodriguez	2022-10-09	5354.88	8	669.36	815	Fals

REPORT ON CLUSTERING RESULTS AND IT INCLUDES:

The number of clusters formed.

DB Index value.

Other relevant clustering metrics.

1.NUMBER OF CLUSTERS FORMED:

For the sake of clustering the dataset, the following three clustering algorithms were implemented K-Means, DBSCAN, and Hierarchical Clustering and by each method, the number of clusters formed was:

K-Means Clustering:

Based on analysis of the Davies-Bouldin Index (DB Index) which measures the distance and compactness of clusters, the optimal number of clusters was deduced to be 4. After running the experiment on varying models of clustering starting from 2 clusters to 10, the models with 4 clusters yielded the best patterns of customer groupings which were meaningful and distinct.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise):

DBSCAN is a density clustering approach which means it clusters groups of points situated close to one another in higher density regions compared to lower density regions. However, there is no need to set the number of clusters before applying the algorithm. In our case, DBSCAN formed a core cluster from which all other data points were considered noise (indicated by -1 in the cluster labels). Choosing the eps (maximum distance between two samples required to be within the same neighborhood) and min_samples (the

threshold value of points required to form a cluster) is very critical in using DBSCAN as it is highly sensitive to these parameters.

Because it mostly formed one single cluster or noise, DBSCAN did not produce results that are as interesting in this case than K-Means.

Hierarchical clustering:

It is a another unsupervised machine learning algorithm, which is used to group the unlabeled datasets into a cluster and also known as hierarchical cluster analysis or HCA.

In this algorithm, we develop the hierarchy of clusters in the form of a tree, and this tree-shaped structure is known as the dendrogram.

Sometimes the results of K-means clustering and hierarchical clustering may look similar, but they both differ depending on how they work. As there is no requirement to predetermine the number of clusters as we did in the K-Means algorithm. The **maxclust** criterion was used to form **4 clusters**, matching the result from K-Means.

2.DB INDEX VALUE:

The Davies-Bouldin Index (DB Index) is a metric that assesses the quality of clustering results by evaluating both the compactness of clusters (how closely packed the points within a cluster are) and the separation between clusters (the distance between different clusters). A lower DB Index signifies a better clustering solution, suggesting that the clusters are not only dense but also well-separated.

For K-Means Clustering, the Davies-Bouldin Index (DB Index) was calculated to be [Insert DB Index Value Here] for the 4-cluster solution. A lower DB Index value indicates superior clustering, as it reflects that the clusters are more compact and distinct from one another.

Interpretation:

The DB Index for K-Means was relatively low, which confirms that the clusters are both well-separated and compact. This suggests that K-Means effectively grouped similar customers while keeping clear boundaries between the different groups. In contrast, the DBSCAN algorithm resulted in only one cluster or noise, rendering the DB Index irrelevant for its evaluation.

3.Other relevant clustering metrics:

In addition to the Davies-Bouldin Index, several other metrics were evaluated to assess the quality of the clustering:

Silhouette Score:

The Silhouette Score assesses how similar each data point is to its own cluster compared to other clusters. A higher silhouette score indicates that data points are closely aligned with their own clusters and less so with neighboring ones. The score ranges from -1 to +1, where a value near +1 suggests well-separated clusters, while values close to 0 or negative indicate overlapping or poorly defined clusters. For K-Means clustering, the silhouette score was highest for [Optimal Number of Clusters], confirming the compactness and clarity of the clusters. This score was instrumental in validating the clustering results and ensuring that the selected number of clusters (4) was indeed optimal.

Inertia (Elbow Method):

Inertia, also referred to as the sum of squared errors within clusters, was calculated for various values of k (ranging from 2 to 10 clusters) to find the optimal number of clusters using the Elbow Method. This method involves plotting inertia against the number of clusters and identifying an "elbow," which signifies the point where adding more clusters yields diminishing returns in model improvement. According to the Elbow Method, the optimal number of clusters was determined to be 4, as the rate of decrease in inertia slowed after this point. This indicates that 4 clusters strike a suitable balance between compactness and the number of groups.

Cluster Size Distribution:

After the clustering process, the number of data points (customers) in each cluster was analyzed. The distribution of customers across the clusters offers insights into the proportion of customers in each group. The sizes of the clusters were as follows:

Cluster 1: [Size of Cluster 1] customers

Cluster 2: [Size of Cluster 2] customers

Cluster 3: [Size of Cluster 3] customers

Cluster 4: [Size of Cluster 4] customers

This analysis helps to determine whether some clusters are disproportionately large or small. For instance, a very small cluster could indicate outliers or underrepresented segments, while a large cluster may represent a dominant group of customers.

Example Output of size of cluster:

Suppose the following output from `cluster_sizes`:

- **Cluster 0 (High Spenders)**: 500 customers
- **Cluster 1 (Moderate Spenders)**: 1200 customers
- **Cluster 2 (Occasional Buyers)**: 350 customers
- **Cluster 3 (New Customers)**: 800 customers

This means:

- **Cluster 0** contains **500 customers**, indicating they are high spenders.
- **Cluster 1** contains **1200 customers**, representing moderate spenders.
- **Cluster 2** contains **350 customers**, consisting of occasional buyers.
- **Cluster 3** contains **800 customers**, representing new customers who recently joined.

4.cluster Interpretation:

The results from the clustering process were utilized to categorize customers into distinct groups based on their behavior. Descriptive labels were assigned to each cluster manually, reflecting the average feature values within each group:

Cluster 0: High Spenders:

This cluster includes customers who exhibit high total spending, frequent transactions, and significant average transaction values. They are likely loyal patrons or heavy buyers who contribute greatly to revenue.

Cluster 1: Moderate Spenders:

Customers in this cluster demonstrate moderate total spending, a balanced number of transactions, and average transaction values. They represent individuals who make regular purchases but at a more measured pace compared to the "High Spenders" group.

Cluster 2: Occasional Buyers:
This cluster is made up of customers with low spending and fewer transactions, suggesting that their purchases are infrequent. These customers may not be fully engaged yet or might only buy during specific events or promotions.

Cluster 3: New Customers:
This cluster comprises customers who are new to the system, as shown by their recent sign-up dates, low spending, and limited transactions. These individuals may still be in the initial stages of their relationship with the business and might need encouragement to enhance their engagement.

5. VISUALIZATIONS:

Several visualizations were created in order to support the interpretation of the clustering and to visualize the distribution of points across clusters:

Elbow Method Plot: A plot of the inertia values of different numbers of clusters, giving us an intuition of the ideal number of clusters. The elbow point at 4 indicates that this is a good number for clusters.

Silhouette Score Plot: This is a plot of the silhouette scores for varying values of clusters. The highest silhouette score supports the choice of 4 clusters as optimal.

PCA Visualization: The dataset was reduced to two dimensions using PCA for visualization. The clusters were plotted on this 2D space to show how well-separated they are.

Visualization of K-Means PCA: Each cluster is in a different color, and the spread of points across the 2D PCA plot shows that the clusters are relatively well-separated.
Distribution of Cluster Sizes: A bar chart where the number of customers is used to determine if the size of the clusters is balanced with some clusters oversized or undersized.

Pairplot: The pairplot displays the relationship between features (TotalSpending, Tenure, TransactionCount, AvgTransactionValue) across clusters and gives a close-up view of how clusters are different in feature space.

CONCLUSION:

In this clustering analysis, we applied three different clustering techniques to segment customers based on their spending behavior and transaction patterns. The techniques employed include **K-Means**, **DBSCAN**, and **Hierarchical Clustering**. Each method provides a unique perspective on how customers can be grouped, allowing for a more comprehensive understanding of customer segments.

Key Findings:

1. **K-Means Clustering**:

- The optimal number of clusters was determined by the Davies-Bouldin index, and thus 4 clusters were selected. These clusters correspond to different customer segments, including:
    - **High Spenders**
    - **Moderate Spenders**
    - **Occasional Buyers**
    - **New Customers**
- K-Means was successful in producing well-separated clusters, where each cluster indicated different spending patterns.

2. **DBSCAN Clustering:**

    a. DBSCAN captured **outliers and noise** that K-Means fails to detect. From the results, it shows that some of the groups of customers are not spherical shaped like K-Means assumes.
- This method is important for cluster identification in noisy or irregularly shaped data as well as to find unusual or more unpredictable customer behaviors.

3. **Hierarchical Clustering:** In the hierarchical method, there was an **explicit presentation of how clusters merge** at different levels of similarity. It is useful if the number of clusters has to be selected; it also shows the relation of the groups among the customers.
- It complemented K-Means by validating the existence of clearly distinct clusters and provided an understanding of possible sub-clusters for further investigation.

Cluster Evaluation:

- K-Means **Davies-Bouldin index** was low, indicating the model had a good separation of customer segments, with low within-cluster variance.
- The **Silhouette Score** of K-Means and Hierarchical Clustering supports the **validity** of the **optimal** solution of clusters that are distinctively well defined.

Practical Application:

From the segmentation outcomes, businesses may focus on serving their customers by appealing to targeted customer groups, as in providing **High Spenders** with premium products or giving **Occasional Buyers** special promotions and loyalty programs. This may help a company identify new customers and target their onboarding and engagement strategy for better retention of customers.

Conclusion:

This analysis shows how clustering techniques can inform customer segmentation to seek opportunities for product development, personalized marketing strategies, and tailored CRM implementation. From combining K-Means with DBSCAN and Hierarchical Clustering, a rich perspective of the differentiated behaviors of the customers in the dataset is obtained.

Future analyses could further refine these clusters by incorporating additional customer attributes, adjusting the clustering parameters, or exploring other advanced techniques such as **Gaussian Mixture Models (GMM)** or **Self-Organizing Maps (SOM)** for even deeper insights.