Overview of the Process

The goal of this task is to segment customers based on their profile and transaction behavior, using clustering techniques. Unlike KMeans, Agglomerative Clustering provides a hierarchical approach that doesn't require pre-defining the number of clusters initially. This method groups customers step-by-step based on their similarity until all are part of a single cluster.

Key Steps

1. Data Preparation

- **Datasets Used**:
    - Customers.csv for customer demographic information.
    - Transactions.csv for transaction behavior.
- **Merging Data**: The Transactions.csv data was merged with Customers.csv to include demographic details (e.g., region) alongside transaction-related features.

2. Feature Engineering

To represent customers meaningfully, the following features were engineered:

- **Total Spent**: Sum of all transaction values by a customer.
- **Total Transactions**: Number of transactions made by the customer.
- **Average Quantity**: Mean quantity purchased per transaction.
- **Number of Unique Products**: Count of distinct products purchased.
- **Region Encoding**: Converted categorical regions into dummy variables for inclusion in the model.

These features capture both customer demographics and transactional behavior.

---

3. Data Scaling

Agglomerative Clustering relies on distance calculations, so features were standardized using StandardScaler to ensure equal weighting across different scales (e.g., total spent in USD vs. region encoding).

---

4. Hierarchical Clustering

- **Dendrogram**: A dendrogram was created using the Ward linkage method to visualize the hierarchical clustering process. The dendrogram helped determine the optimal number of clusters by observing where large vertical gaps (distance thresholds) occurred.
- **Cluster Assignment**: Using the identified number of clusters (e.g., 5), Agglomerative Clustering was applied to segment the customers.

5. Clustering Metrics

- **Davies-Bouldin Index (DB Index)**: This metric was calculated to evaluate the clustering quality. It measures the compactness of clusters and their separation. A lower DB Index indicates better clustering.

Other clustering metrics can be added if required.

6. Cluster Visualization

- **Pairplots**: Used to visualize how clusters differ across multiple features (e.g., total spent, average quantity, etc.).

7. Results Export

- The clustering results, including cluster assignments, were saved as a CSV file (Customer_Clusters_Agglomerative.csv).
- The CSV contains the original customer IDs, engineered features, and the cluster label for each customer.

Advantages of Agglomerative Clustering

1. **Interpretability**: The dendrogram shows the entire clustering hierarchy, giving insight into how customers are grouped.
2. **Flexibility**: It doesn't require specifying the number of clusters beforehand.
3. **Well-Suited for Small to Medium Datasets**: Hierarchical clustering is computationally expensive but performs well with datasets of moderate size.

Summary

The hierarchical approach segments customers into groups based on their similarity in transactional and demographic features. By analyzing clusters, the business can target groups more effectively (e.g., high spenders, frequent buyers, etc.). The dendrogram and DB Index ensure the model’s logic and quality are data-driven and explainable.