**Muhammad Abdullah**

**STU-DS-251-930**

**GitHub:** **https://github.com/itxabdullah-code/Wholesale-Cutomer-Data-Clustering.git**

**Wholesale Customer Data Clustering**

**Project Overview**

I conducted a Customer Segmentation project using the Wholesale Customers dataset. The goal was to segment customers into distinct groups based on their annual spending across various product categories, enabling more effective targeting and strategic planning for businesses. I applied unsupervised learning techniques such as K-Means, DBSCAN, and Hierarchical Clustering.

**A. Data Preparation & Preprocessing**

**1. Loading and Exploring the Dataset**

I loaded the dataset using pandas and examined its structure, data types, and summary statistics. This helped me understand the distribution and scale of features such as Fresh, Milk, Grocery, Frozen, Detergents\_Paper, and Delicassen.

**2. Handling Missing or Duplicate Values**

I checked for missing values using isnull().sum() and for duplicates using duplicated().sum(). The dataset contained no missing values but had some duplicate rows, which I removed to maintain data quality.

**3. Encoding Categorical Features**

Although the dataset included Channel and Region as categorical columns, I chose to drop them and focus on numerical features directly related to spending behavior.

**4. Feature Scaling**

I used StandardScaler to normalize the numerical features so they all had a mean of 0 and a standard deviation of 1. This ensured that the clustering algorithms, which rely on distance calculations, performed optimally.

**B. Dimensionality Reduction**

**1. Principal Component Analysis (PCA)**

I applied PCA to reduce the six-dimensional dataset to 2 dimensions for visualization purposes. I reviewed the explained variance to confirm that most of the data’s variance was captured in the top two components.

**2. Visualization of PCA**

I plotted the PCA-transformed dataset using a 2D scatter plot. This visualization provided insights into the natural groupings and separability of customer records.

**C. Clustering Implementation**

**1. K-Means Clustering**

I applied the K-Means algorithm using different values of K. I used the Elbow Method to observe the point where inertia dropped sharply, suggesting the optimal number of clusters. I then applied K-Means with this optimal K value and visualized the results.

**2. DBSCAN**

I implemented DBSCAN to perform density-based clustering. By tuning the eps and min\_samples parameters using k-distance plots, I was able to identify distinct clusters and outlier points, especially in sparse areas.

**3. Hierarchical Clustering**

I used Agglomerative Clustering and plotted a dendrogram to identify the appropriate number of clusters. Based on the dendrogram, I applied the algorithm and visualized the final cluster distribution.

**D. Evaluation**

**1. Silhouette Score**

I calculated the Silhouette Score for each clustering algorithm to evaluate how well the clusters were formed. Higher silhouette scores indicated more coherent clusters.

**2. Davies-Bouldin Index**

I computed the Davies-Bouldin Index to assess cluster separation and compactness. A lower index value indicated better clustering performance.

**3. Comparison and Interpretation**

I compared the clustering results across the three algorithms. K-Means produced the most balanced and distinguishable clusters. DBSCAN helped in identifying outliers and non-spherical clusters. Hierarchical Clustering was useful for understanding the data’s nested structure.

**Conclusion**

Through this project, I successfully segmented wholesale customers based on their annual purchasing patterns. I used multiple clustering algorithms and evaluated them with standard metrics. These clusters can help businesses customize services, optimize resource allocation, and develop targeted marketing campaigns.