**Name:** Arya Yemul

**Roll No:** 281015

**Batch:** A1

**Assignment 5**

**Problem Statement:**

Write a program to do following: Data Set:

https://www.kaggle.com/shwetabh123/mall-customers This dataset gives the data of Income and money spent by the customers visiting a shopping mall. The data set contains Customer ID, Gender, Age, Annual Income, Spending Score. Therefore, asa mall owner you need to find the group of people who are the profitable customers for the mall owner. Apply at least two clustering algorithms (based on Spending Score) to find the group ofcustomers.

a) Apply Data pre-processingCO6 L2

b) Perform data-preparation (Train-Test Split)

c) Apply Machine Learning Algorithm

**Objectives:**

1. To preprocess the dataset by handling missing values, encoding categorical variables, and standardizing numerical features.
2. To explore and visualize customer segments using spending score and annual income through clustering techniques.
3. To apply at least two clustering algorithms (e.g., Kmeans,Agglomerative clustering ) for identifying distinct customer groups.
4. To evaluate clustering performance using metrics such as silhouette score and visualizations.
5. To apply cross-validation techniques to assess the consistency and robustness of the clustering models.

**Resources used:**

1) Software used: Jupiter Notebook

2) Libraries used: Pandas, Matplotlib, SKLearn

**Theory:**

1. **Clustering:**

Clustering is an unsupervised learning technique that groups similar data points into clusters. It helps in customer segmentation, anomaly detection, and more.

**2.Clustering Methods**

1. **K-Means Clustering**

* **Type:** Centroid-based
* **Input:** Number of clusters (K)
* **Works by:** Minimizing distance between points and centroids
* **Pros:** Fast, simple
* **Cons:** Sensitive to outliers, needs K

1. **Agglomerative clustering**

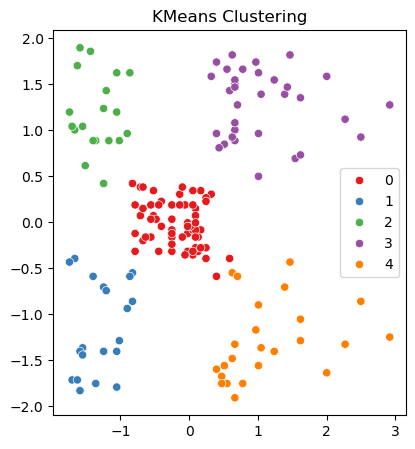
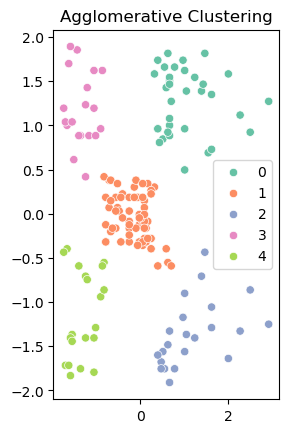
* **Type:** Hierarchical
* **Input:** Number of clusters (can be visualized using a dendrogram)
* **Works by:** Recursively merging the closest pairs of clusters based on a linkage criterion
* **Pros:** No need to specify cluster shape, dendrogram helps visualize structure
* **Cons:** Computationally intensive for large datasets, no automatic way to choose optimal clusters

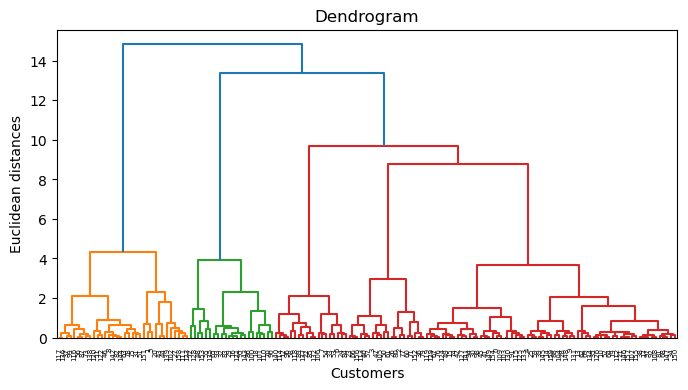
**Methodology:**

**1. Concise Methodology:**

1. Load and preprocess the dataset by handling missing values, encoding categorical data, and standardizing numerical features.
2. Perform exploratory data analysis to understand distributions and relationships between features.
3. Select relevant features (Annual Income, Spending Score) for clustering.
4. Apply KMeans and Agglomerative clustering algorithms to segment customers.
5. Evaluate cluster quality using Silhouette Score and visualizations.
6. Apply cross-validation to assess clustering consistency and robustness.
7. Interpret the results to identify profitable customer segments for strategic decision-making.

**Results:**



Cross-Validation:

K=2, Silhouette Score = 0.321

K=3, Silhouette Score = 0.467

K=4, Silhouette Score = 0.494

K=5, Silhouette Score = 0.555

K=6, Silhouette Score = 0.540

K=7, Silhouette Score = 0.528

K=8, Silhouette Score = 0.455

K=9, Silhouette Score = 0.457

K=10, Silhouette Score = 0.443

**Conclusion:**

Both **KMeans** and **Agglomerative Clustering** methods were effective in identifying distinct customer segments. Evaluation metrics like **Silhouette Score** helped assess cluster quality. The project successfully demonstrates clustering analysis for customer segmentation based on income and spending patterns.