**Name: Kedar Ramchandra Kolase**

**Roll no: 282013**

**Batch: B1**

**Assignment 5**

**Problem Statement:**

Analyze the Mall Customers dataset to identify profitable customer segments using clustering techniques. Apply data preprocessing, prepare the data, and use at least two clustering algorithms (e.g., KMeans, Agglomerative) based on Spending Score and Income. Evaluate the clusters using silhouette scores and visualizations, and perform cross-validation to ensure model robustness.

**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 and Agglomerative) 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

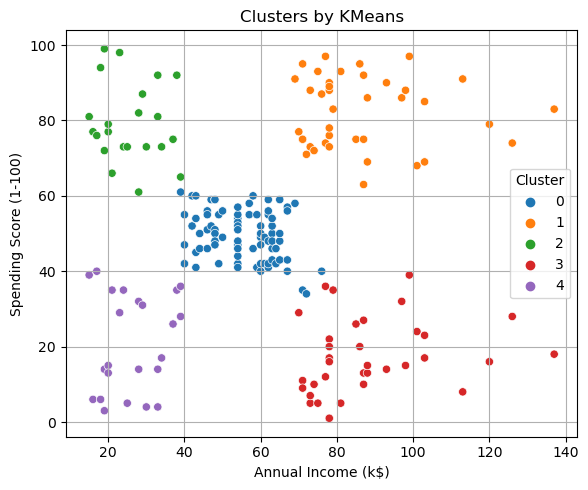
**II. Agglomerative Clustering (Hierarchical Clustering - Bottom-Up)**

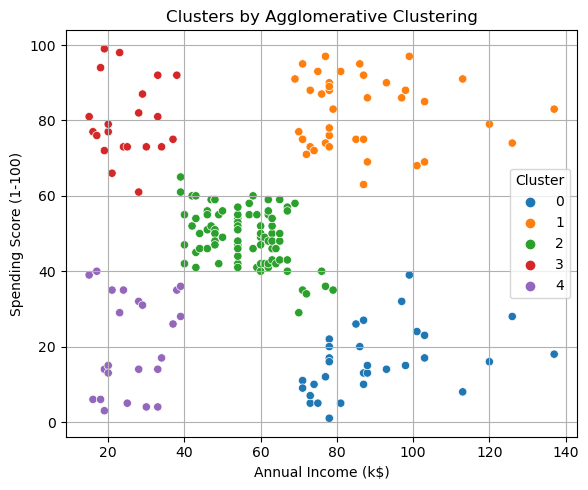
* **Type:** Hierarchical-based
* **Input:** Number of clusters, linkage criteria (e.g., single, complete, average, ward)
* **Works by:** Initially treats each point as its own cluster, then repeatedly merges the closest pair of clusters based on linkage criteria until the desired number of clusters is reached
* **Pros:**
  + Dendrogram gives insights into cluster structure
  + No need to pre-define cluster shapes
* **Cons:**
  + Computationally expensive for large datasets
  + Requires number of clusters as input
  + Sensitive to noise and outliers

**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:**



**Conclusion:**

Clustering techniques applied to the Mall Customers dataset effectively identified distinct customer segments based on income and spending behavior. KMeans and Agglomerative revealed profitable customer groups, supported by silhouette scores and visualizations. The results offer valuable insights for targeted marketing and strategic business decisions.