**Name: Ninad Chavan**

**Roll no: 281036**

**Batch: A2**

**Assignment 5**

**Statement**

**Q.** Perform **Clustering Analysis** on **Mall Customer Data**  
**Dataset**: [Mall Customers Dataset](https://www.kaggle.com/shwetabh123/mall-customers)

The dataset includes details such as Customer ID, Gender, Age, Annual Income, and Spending Score. As a mall owner, the aim is to **identify customer segments** based on their **Spending Score** using clustering techniques.

**Tasks:** a) Apply Data Pre-processing  
b) Perform Data Preparation (Train-Test Split)  
c) Apply Machine Learning Algorithms  
d) Evaluate the Model  
e) Apply Cross-Validation and Evaluate the Model

**Objective**

1. Identify customer segments based on spending behavior.
2. Use clustering algorithms to group similar customers.
3. Derive business insights to improve customer service and marketing strategies.

**Resources Used**

* **Software**: Google Colab
* **Libraries**: Pandas, Scikit-learn, Matplotlib, Seaborn

**Introduction to Clustering**

Clustering is an **unsupervised machine learning technique** used to group data points with similar characteristics. In this assignment, clustering helps group **mall customers** based on their **Spending Score**, enabling targeted business actions.  
We primarily use:

* **K-Means Clustering**
* **Hierarchical Clustering**

**Methodology**

1. **Data Pre-processing**
   * Load the dataset and inspect the structure.
   * Handle missing values (if any).
   * Normalize/scale features for optimal clustering performance.
2. **Data Preparation**
   * Select relevant features (e.g., Age, Annual Income, Spending Score).
   * Apply **train-test split** if evaluating clustering with supervised metrics post-labeling.
3. **Model Application**
   * **K-Means Clustering**:
     + Use the **Elbow Method** to determine the optimal number of clusters.
     + Apply the **K-Means algorithm** and assign cluster labels to each customer.
   * **Hierarchical Clustering**:
     + Create a **dendrogram** to visualize the cluster formation.
     + Apply **Agglomerative Clustering** and assign cluster labels.
4. **Model Evaluation**
   * Evaluate clustering quality using **Silhouette Score**.
   * Visualize clusters with **2D scatter plots** for insights.
5. **Cross-Validation**
   * Use techniques like **K-Fold Cross-Validation** (especially if evaluating using labeled outcomes).
   * Check model consistency across folds.

**Advantages of Clustering**

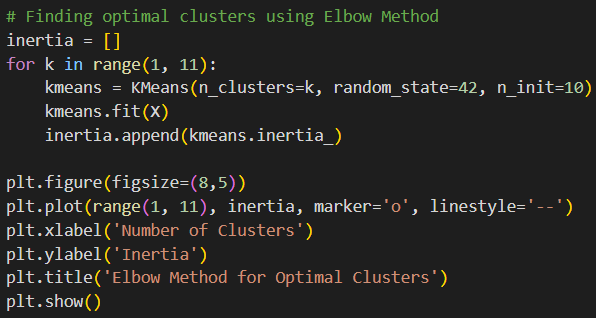
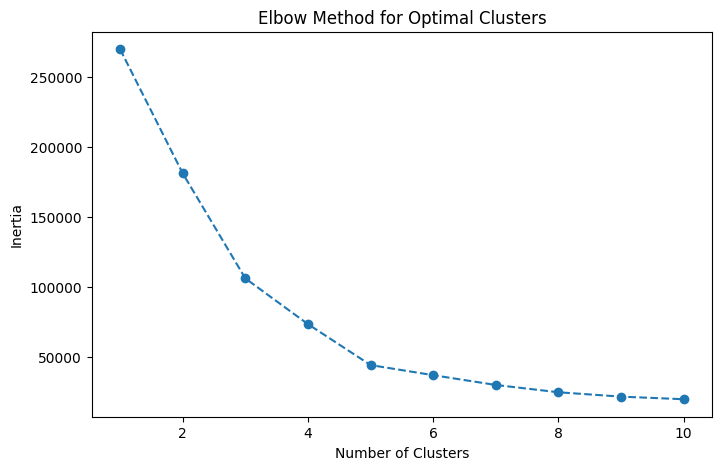
1. Aids in **customer segmentation** and **targeted marketing**.
2. Helps discover hidden **patterns in spending behavior**.
3. Enables the design of **personalized customer services**.

**Disadvantages**

1. Sensitive to **feature scaling** and **initial conditions**.
2. Interpretation of clusters may require **domain expertise**.

**Conclusion**

This assignment explored **K-Means** and **Hierarchical Clustering** techniques to segment mall customers based on **Spending Score**. By visualizing and evaluating the clusters using **Silhouette Score**, we gained actionable insights for personalized marketing strategies. The use of **cross-validation** improved confidence in the model’s reliability and consistency.

******Output**

