Name: Ayush Gupta

Roll no: 281049

Batch: A2

Assignment 5

Statement:

Q. Clustering Analysis on Mall Customer Data

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. Gain business insights to enhance customer experience and marketing strategies.

Resources Used:

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

Introduction to Clustering: Clustering is an unsupervised learning technique used to group data points based on similarities. In this case, we group customers based on their Spending Score using clustering algorithms such as K-Means and Hierarchical Clustering.

Methodology:

1. **Data Pre-processing:**
   - Load the dataset and inspect structure.
   - Handle missing values if any.
   - Normalize/scale features for better clustering performance.
2. **Data Preparation:**
   - Select relevant features, especially 'Spending Score'.

   - Apply train-test split (if needed for evaluation of clustering performance).
3. **Model Application:**
   - **K-Means Clustering:**
     - Determine optimal number of clusters using Elbow Method.
     - Apply K-Means algorithm and assign cluster labels to data.
   - **Hierarchical Clustering:**
     - Generate dendrogram to visualize cluster formation.
     - Apply Agglomerative Clustering and assign labels.
4. **Model Evaluation:**
   - Use metrics such as Silhouette Score to evaluate clustering quality.
   - Visualize clusters using 2D scatter plots.
5. **Cross-Validation:**
   - Apply techniques like K-Fold cross-validation (if using clustering with supervised metrics post-labeling).
   - Evaluate model consistency across different folds.

Advantages of Clustering:

1. Helps in customer segmentation and targeted marketing.
2. Identifies patterns in customer spending behavior.
3. Assists in personalized service design.

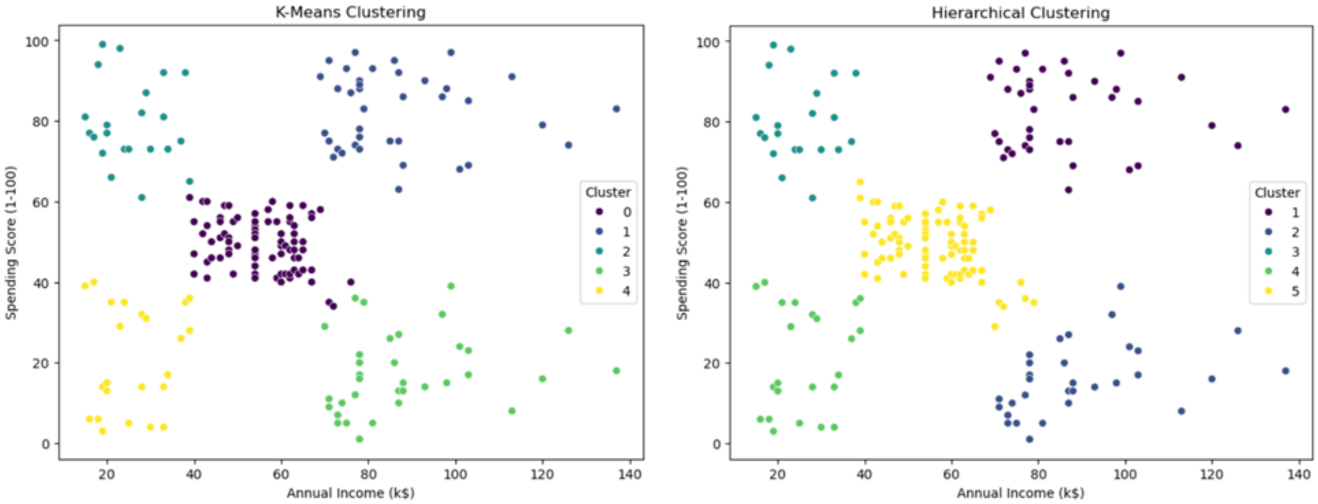
Disadvantages:

1. Sensitive to feature scaling and initial conditions.
2. Difficult to interpret clusters without domain knowledge.

Conclusion:

This assignment demonstrated how clustering techniques like K-Means and Hierarchical Clustering can help segment mall customers based on their Spending Score. These insights can be used to identify profitable customer groups and enhance marketing strategies. The use of model evaluation and cross-validation further validated the reliability of the clustering results.

OUTPUT:


K-Means Clustering
Spending Score (1-100)
Annual Income (k$)
0
20
40
60
80
100
20
40
60
80
100
120
140
Cluster
0
1
2
3
4
Hierarchical Clustering
Spending Score (1-100)
Annual Income (k$)
0
20
40
60
80
100
20
40
60
80
100
120
140
Cluster
1
2
3
4
5