Name: Rugved Sawant
Roll No: 281065
Batch: A3

Assignment 5

Statement

Q. Perform **Clustering Analysis** on **Mall Customer Data**
Dataset: Mall Customers Dataset

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**.

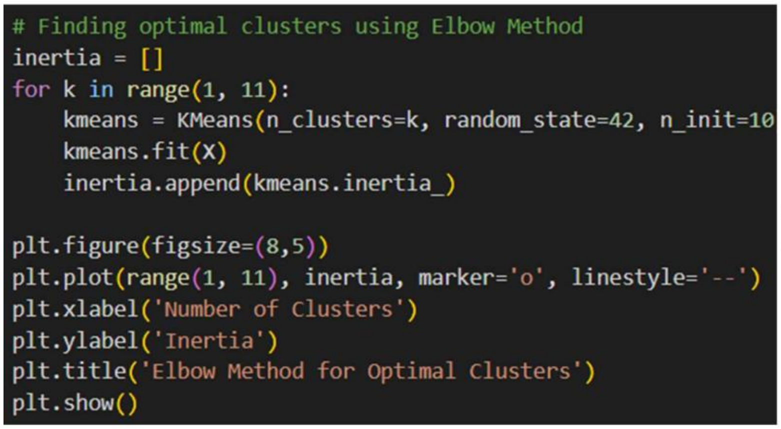
Disadvantages

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

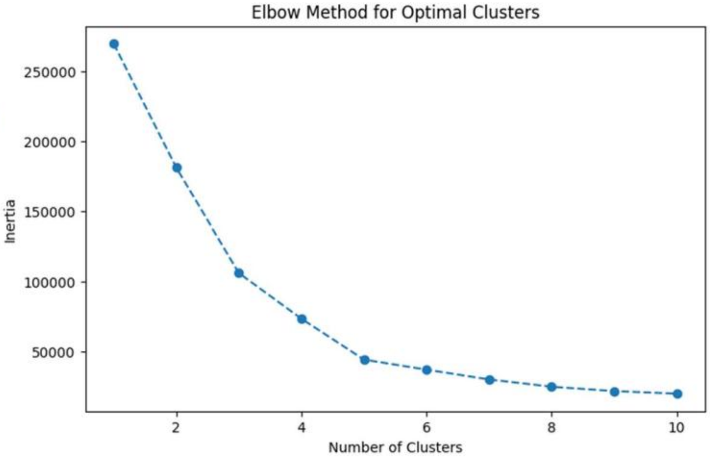
Conclusion

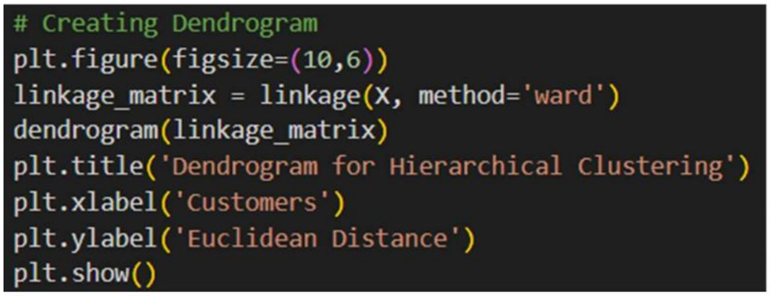
This assignment explored **K-Means** and **Hierarchical Clustering** techniques to segment mall customers based on **Spending Score**. By visualizing and evaluating the clusters using **Silhouette Score**, we gained actionable insights for personalized marketing strategies. The use of **cross-validation**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

Output

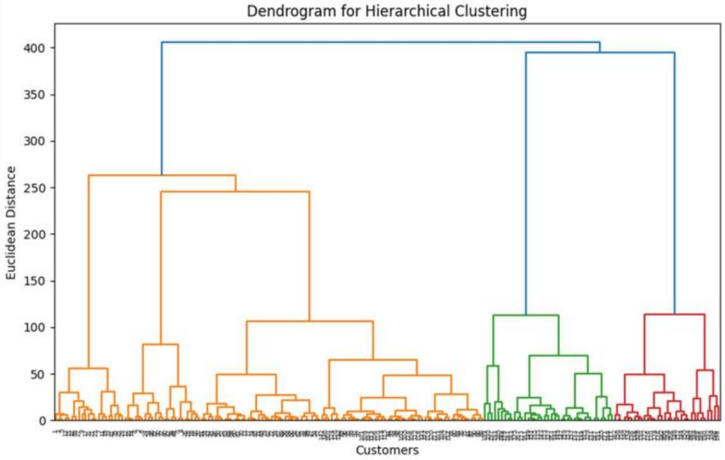


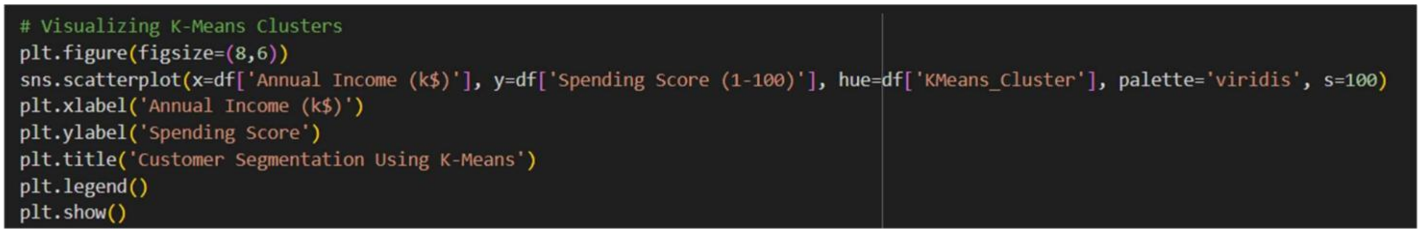
```
# Finding optimal clusters using Elbow Method
inertia = []
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k, random_state=42, n_init=10
    kmeans.fit(X)
    inertia.append(kmeans.inertia_)

plt.figure(figsize=(8,5))
plt.plot(range(1, 11), inertia, marker='o', linestyle='--')
plt.xlabel('Number of Clusters')
plt.ylabel('Inertia')
plt.title('Elbow Method for Optimal Clusters')
plt.show()
```



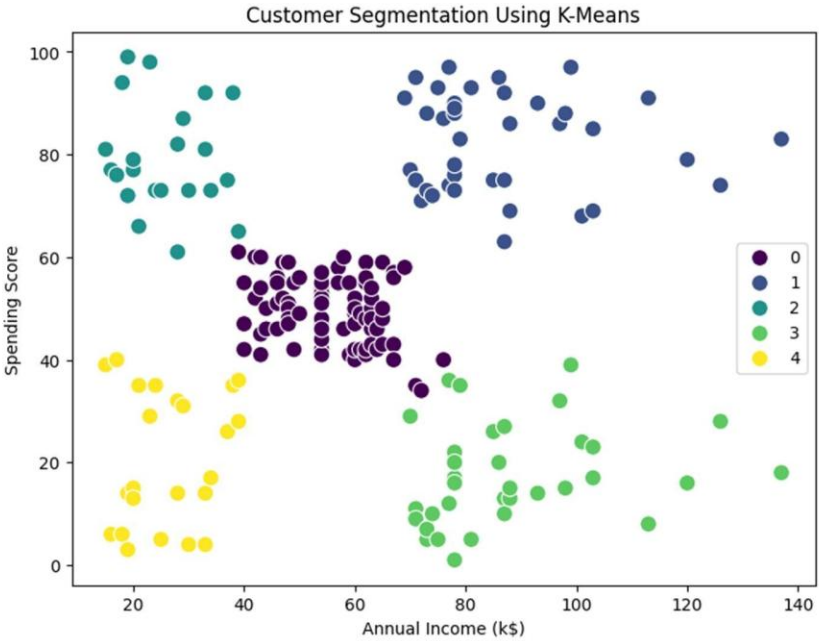


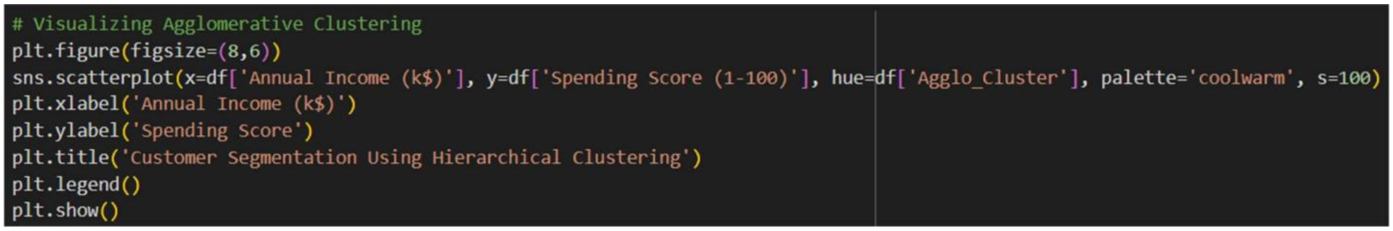
```
# Creating Dendrogram
plt.figure(figsize=(10,6))
linkage_matrix = linkage(X, method='ward')
dendrogram(linkage_matrix)
plt.title('Dendrogram for Hierarchical Clustering')
plt.xlabel('Customers')
plt.ylabel('Euclidean Distance')
plt.show()
```



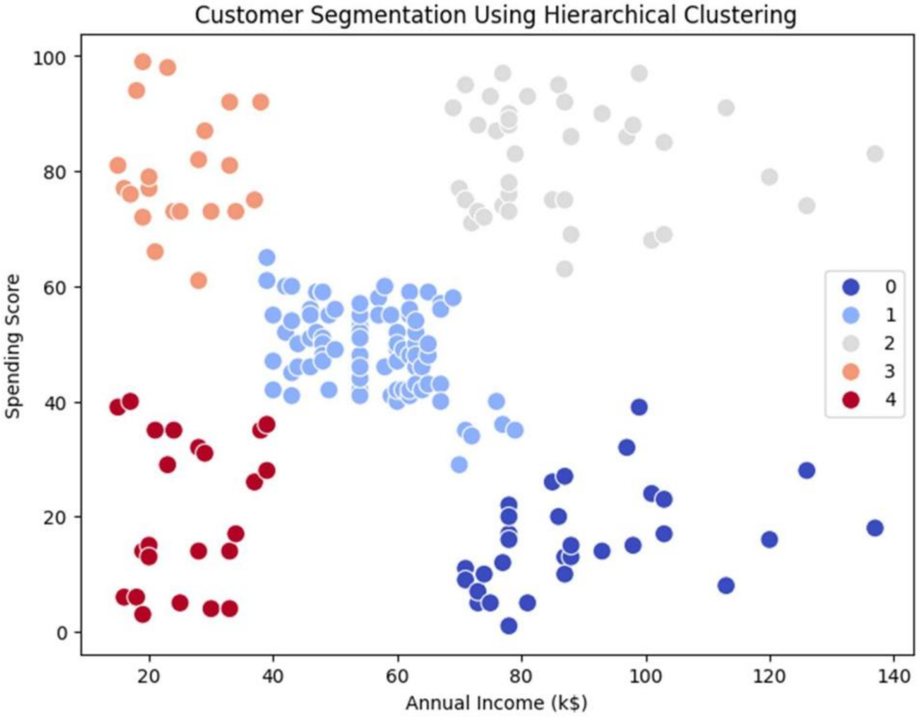


```
# Visualizing K-Means Clusters
plt.figure(figsize=(8,6))
sns.scatterplot(x=df['Annual Income (k$)'], y=df['Spending Score (1-100)'], hue=df['KMeans_Cluster'], palette='viridis', s=100)
plt.xlabel('Annual Income (k$)')
plt.ylabel('Spending Score')
plt.title('Customer Segmentation Using K-Means')
plt.legend()
plt.show()
```





```
# Visualizing Agglomerative Clustering
plt.figure(figsize=(8,6))
sns.scatterplot(x=df['Annual Income (k$)'], y=df['Spending Score (1-100)'], hue=df['Agglo_Cluster'], palette='coolwarm', s=100)
plt.xlabel('Annual Income (k$)')
plt.ylabel('Spending Score')
plt.title('Customer Segmentation Using Hierarchical Clustering')
plt.legend()
plt.show()
```


Customer Segmentation Using Hierarchical Clustering
Spending Score
Annual Income (k$)
● 0
● 1
● 2
● 3
● 4