📅 Day 12 – Hierarchical Clustering (Unsupervised Learning)

📌 Session Overview

Today's session was centered around **Hierarchical Clustering**, a powerful unsupervised learning technique used to find hidden patterns or groupings in unlabeled datasets. Unlike supervised learning, clustering doesn't require pre-defined categories; instead, it identifies natural groupings based on similarity.

We worked with the **Mall Customers dataset**, aiming to uncover customer segments based on **Annual Income** and **Spending Score**. The goal was to group individuals with similar purchasing behavior.

🌱 What is Hierarchical Clustering?

Hierarchical clustering is a method that builds a multi-level hierarchy of clusters, resembling a tree structure called a **dendrogram**. It allows us to see how clusters are formed and related at various levels of similarity.

There are two main strategies:

- 💠 **Agglomerative Clustering (Bottom-Up):** Begins with each data point as a separate cluster and merges them step by step.
- 💠 **Divisive Clustering (Top-Down):** Starts with all points in one cluster and splits them recursively.

In our session, we focused on **Agglomerative Clustering**, which is more commonly used in practice.

▯ Key Concepts Covered

- **Dendrogram**: A tree-like diagram used to decide the number of clusters visually.
- **Linkage Methods**: How the distance between clusters is measured (e.g., single, complete, average, or ward).
- **Affinity Metric**: The distance metric used (e.g., Euclidean distance).

These concepts help in deciding how clusters are formed and what defines "closeness" between data points.

🗁 Dataset in Focus: Mall_Customers.csv

We used a real-world dataset with the following features:

- Customer ID
- Gender
- Age
- **Annual Income (k$)**
- **Spending Score (1–100)**

For this task, we used only **Annual Income** and **Spending Score** to observe behavioral patterns related to earning and spending.

🛠 Step-by-Step Implementation

✅ 1. Importing Libraries

```
import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

import scipy.cluster.hierarchy as sch

from sklearn.cluster import AgglomerativeClustering
```

✅ 2. Data Loading & Selection

```
data = pd.read_csv("Mall_Customers.csv")

X = data.iloc[:, [3, 4]].values
```

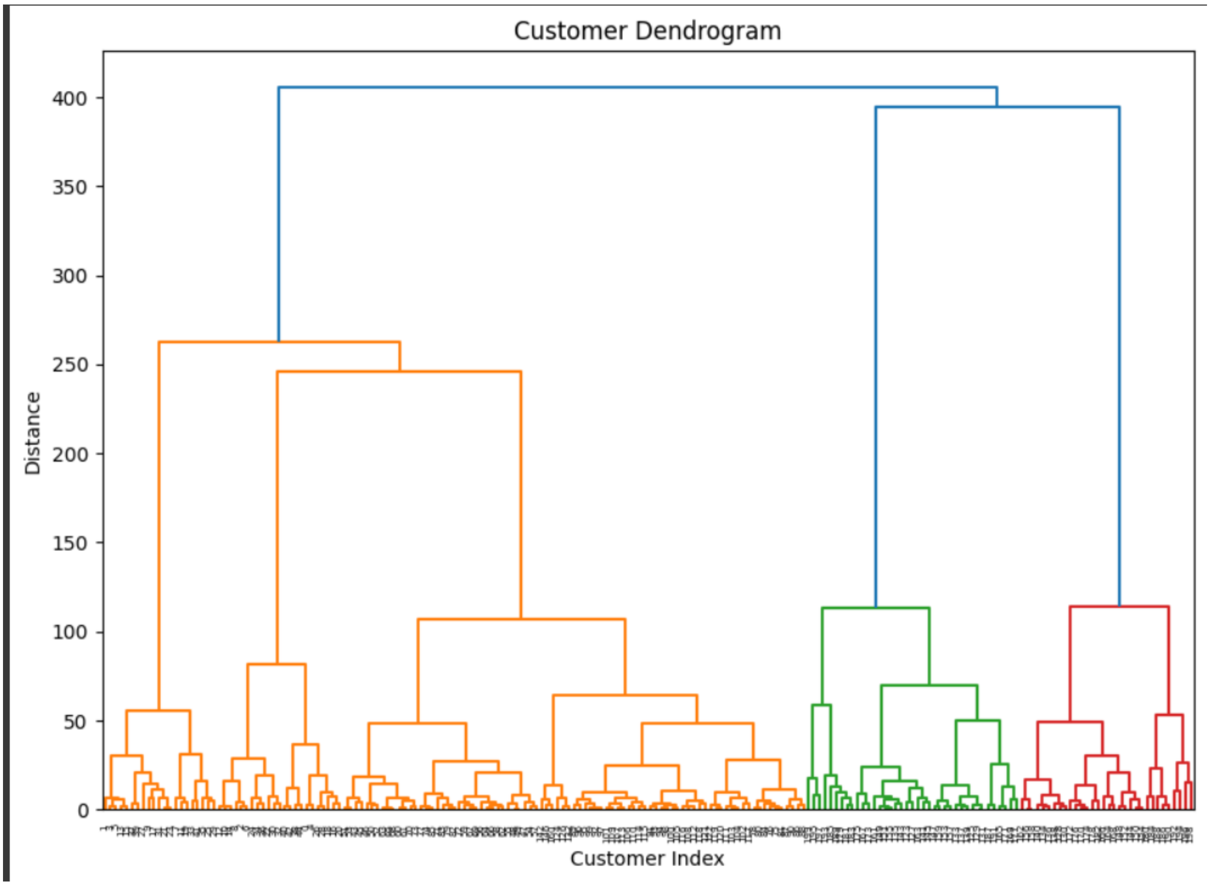
We selected columns 3 and 4 for clustering: **Annual Income** and **Spending Score**.

🌳 Building a Dendrogram

Before applying the algorithm, we used a dendrogram to identify the optimal number of clusters.

```
plt.figure(figsize=(10, 7))

dendrogram = sch.dendrogram(sch.linkage(X, method='ward'))

plt.title("Customer Dendrogram")

plt.xlabel("Customer Index")

plt.ylabel("Distance")

plt.show()
```

The dendrogram showed a noticeable **elbow at 5 clusters**, suggesting a natural grouping point.



👹 Applying Agglomerative Clustering

hc = AgglomerativeClustering(n_clusters=5, affinity='euclidean', linkage='ward')

y_hc = hc.fit_predict(X)

We configured the model with:

- n_clusters=5 (from dendrogram)
- affinity='euclidean' (for distance calculation)
- linkage='ward' (to minimize within-cluster variance)

🎨 Visualizing the Clusters

We plotted the resulting clusters to understand how customers are grouped:

plt.figure(figsize=(8,6))

colors = ['red', 'blue', 'green', 'cyan', 'magenta']

for i in range(5):

    plt.scatter(X[y_hc == i, 0], X[y_hc == i, 1], s=100, c=colors[i], label=f'Cluster {i+1}')

plt.title('Customer Segments')

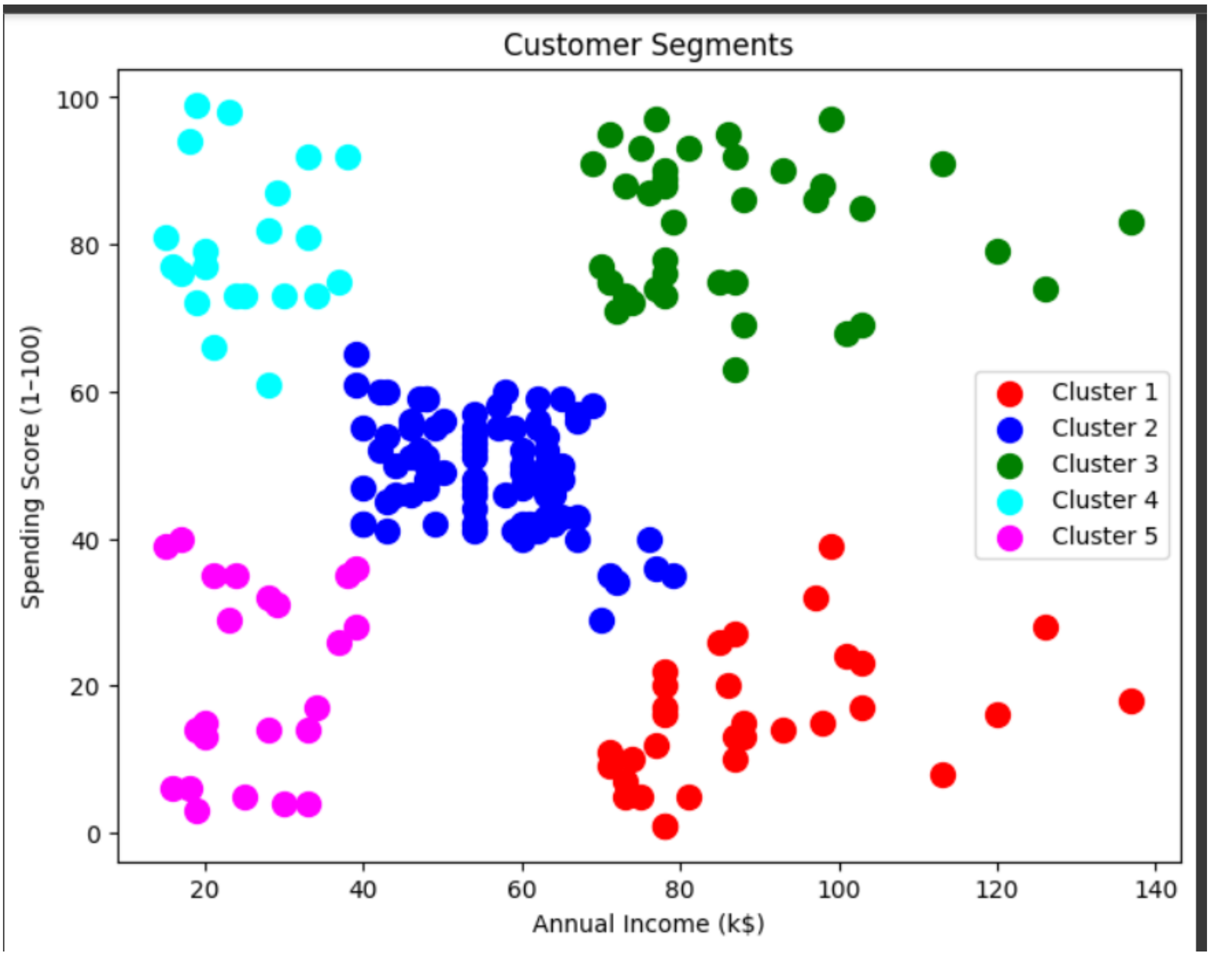
plt.xlabel('Annual Income (k$)')

plt.ylabel('Spending Score (1–100)')

plt.legend()

plt.show()

output :



🔍 Insights from the Clusters

Each cluster displayed distinct customer profiles:

- **Cluster 1**: High income, high spenders – potential premium clients.
- **Cluster 2**: Low income, low spenders – budget-focused.
- **Cluster 3**: Moderate income and spending – average shoppers.
- **Cluster 4**: High income, low spenders – possibly cautious buyers.
- **Cluster 5**: Low income, high spenders – may represent impulsive or young customers.

These clusters help businesses in **customer targeting, product positioning**, and **marketing strategy**.

✅ Why Hierarchical Clustering?

- Doesn't require a pre-defined number of clusters.
- Generates a dendrogram for deep insights into data structure.
- Suitable for small to medium datasets with meaningful structure.

□ Conclusion

Today's session helped us:

- Understand the theory and implementation of **Agglomerative Hierarchical Clustering**.
- Learn how dendrograms assist in identifying cluster count.
- Visualize and interpret customer segments effectively.
- Realize the real-world impact of clustering in business intelligence and customer profiling.

This experience deepened our grasp of **unsupervised learning** techniques and their strategic value in data-driven decision-making.