**Name:** Swaraj Pawar

**Roll no:** 282072

**Batch:** B3

**Assignment – 5**

**Clustering**

**Problem statement:** Write a program to do following: Data Set: https://www.kaggle.com/shwetabh123/mall-customers This dataset givesthe data of Income and money spent by the customers visiting a shopping mall. The data set contains Customer ID, Gender, Age, Annual Income, Spending Score. Therefore, asa mall owner you need to find the group of people who are the profitable customers for the mallowner. Apply at least two clustering algorithms (based on Spending Score) to find the group ofcustomers.

a) Apply Data pre-processing CO6 L2

b) Perform data-preparation (Train-Test Split)

c) Apply Machine Learning Algorithm

**Software used:**

1. Python 3.x
2. Google Colab

**Libraries and packages used:** NumPy, Matplotlib, scikit-learn

**Theory:**

**Methodology:**

Clustering is an unsupervised machine learning technique used to group similar data points together based on certain features. The goal is to ensure that data points in the same cluster are more similar to each other than to those in other clusters.

📌 Why use Clustering?

* To discover patterns or structures in data without pre-existing labels.
* Useful in market segmentation, anomaly detection, image segmentation, recommendation systems, etc.

**1. K-Means Clustering**

K-Means aims to partition data into **K distinct clusters**. It tries to minimize the **intra-cluster variance** (distance from the center).

**📌 How it works:**

1. Choose the number of clusters K.
2. Randomly initialize K centroids.
3. Assign each data point to the nearest centroid.
4. Recalculate centroids based on the mean of points in the cluster.
5. Repeat steps 3-4 until centroids no longer change (or a max iteration is reached).

**🧪 Example Use Case**

* Grouping customers in a shopping mall based on **income** and **spending score**.
* Segmenting users for targeted marketing.

**✅ Advantages**

* Simple and easy to implement.
* Scales well to large datasets.
* Efficient for spherical clusters.

**❌ Disadvantages**

* Must predefine K (number of clusters).
* Sensitive to initial centroid positions.
* Performs poorly with non-spherical (complex-shaped) clusters.
* Affected by outliers.

## 2. Gaussian Mixture Model (GMM)

GMM is a **probabilistic model** that assumes all data points are generated from a mixture of several **Gaussian distributions** with unknown parameters.

Each cluster is modeled as a Gaussian (normal) distribution.

### 📌 ****How it works:****

1. Assume K Gaussian distributions.
2. Use the **Expectation-Maximization (EM)** algorithm:
   * **E-Step:** Estimate the probability of each point belonging to each cluster.
   * **M-Step:** Update parameters (means, variances, and mixing coefficients) based on these probabilities.
3. Repeat until convergence.

### 🧪 ****Example Use Case****

* Customer segmentation when clusters are **overlapping** or **elliptical** in shape.
* Image segmentation, speaker identification, etc.

### ✅ ****Advantages****

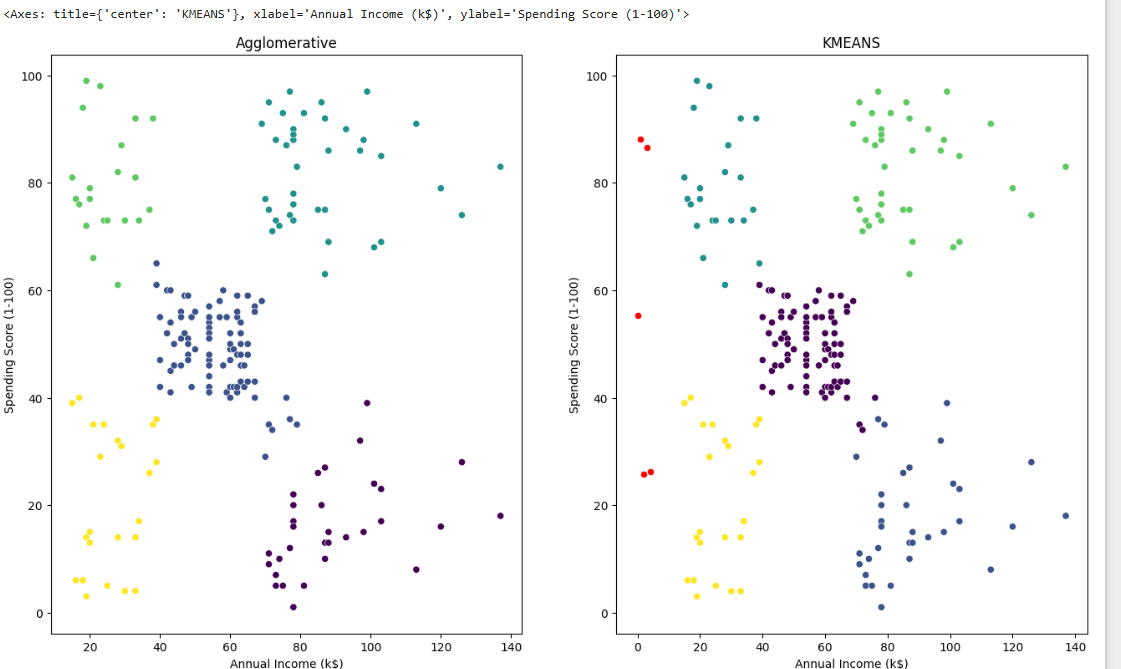
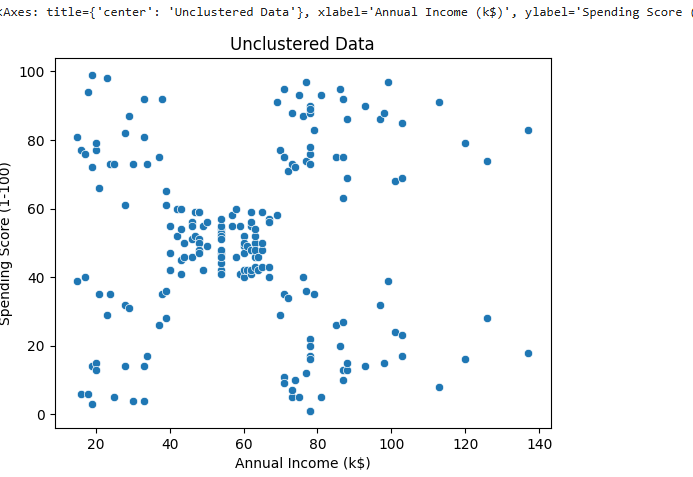
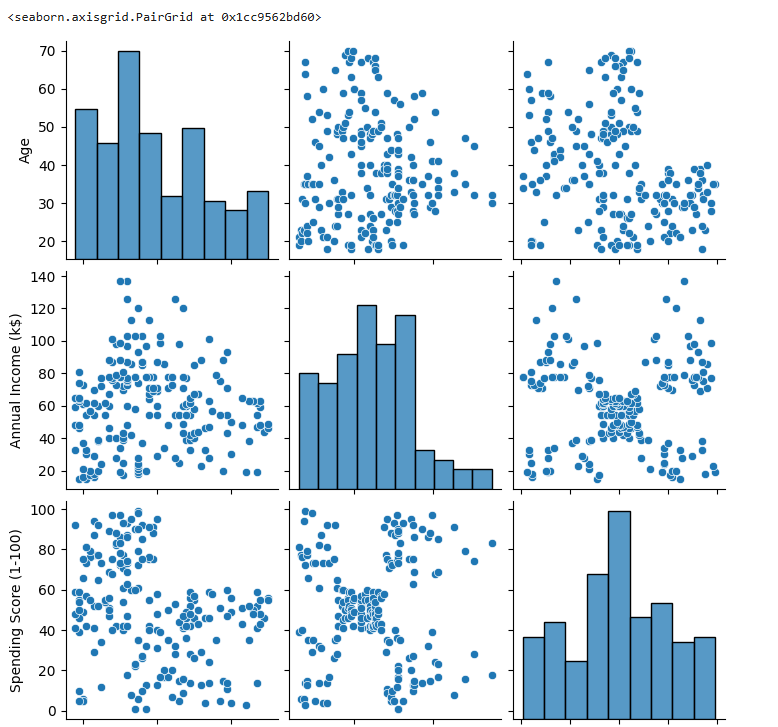
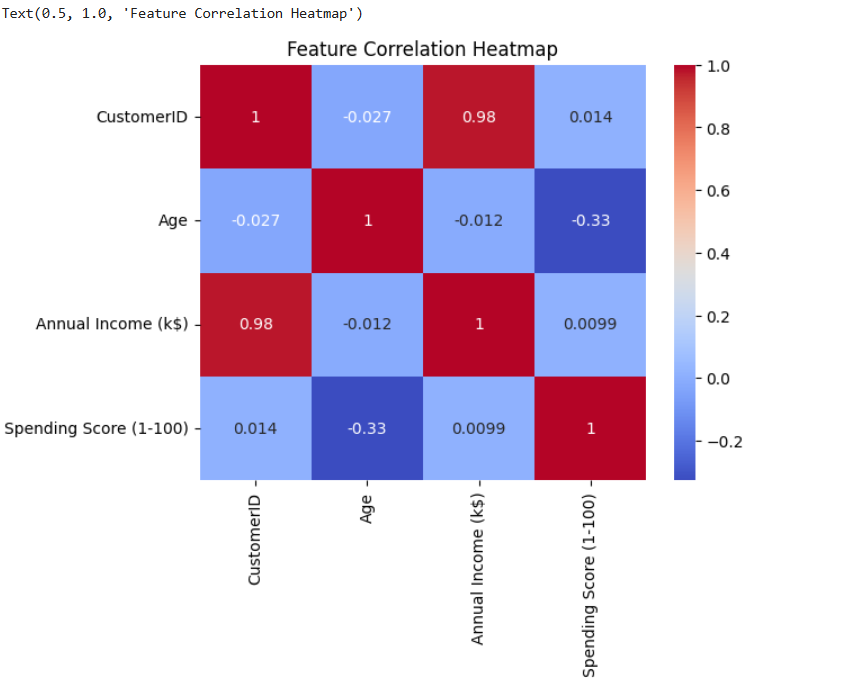
* More flexible than K-Means (supports elliptical clusters).
* Provides **probability** for cluster membership (soft clustering).
* Handles overlapping clusters well.

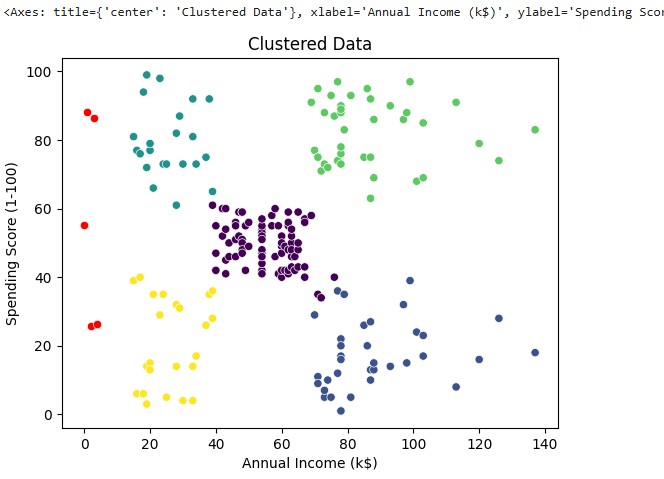
### ❌ ****Disadvantages****

* Computationally more expensive than K-Means.
* Can overfit if too many components are chosen.
* Assumes data follows Gaussian distribution.



Output :





**Conclusion:**

In this assignment, we explored customer segmentation using unsupervised machine learning techniques on the *Mall Customers* dataset. By focusing on features like **Annual Income** and **Spending Score**, we aimed to identify distinct customer groups that would be most profitable for the mall owner.

We applied two clustering algorithms:

1. **K-Means Clustering** – A simple and fast algorithm that grouped customers into distinct, non-overlapping clusters based on their proximity to cluster centroids.
2. **Gaussian Mixture Model (GMM)** – A more flexible, probabilistic model that allowed for overlapping and elliptical clusters, providing soft assignments based on likelihood.