**Assignment No. 7**

**Problem Statement:** Apply the K-Means clustering algorithm to segment customers based on their annual income and spending score using the Mall Customers dataset.

**Objective:**

1. To group mall customers into distinct clusters based on their behavior using unsupervised learning.
2. To identify customer segments that can be used for targeted marketing strategies.
3. To understand the working of the K-Means algorithm and apply it to real-world data.

**Prerequisite :**

1. Basic knowledge of Python programming
2. Familiarity with pandas and NumPy libraries
3. Understanding of data preprocessing techniques
4. Knowledge of K-Means Clustering algorithm
5. Ability to visualize data using matplotlib or seaborn
6. Awareness of clustering evaluation metrics (silhouette score)

**Theory :**

K-Means Clustering is an **unsupervised machine learning algorithm** used to divide a dataset into groups, or clusters, of similar data points. The goal is to group data in such a way that points within the same cluster are very similar to each other, and points in different clusters are quite different. This is especially helpful when we don’t have labeled data and want to explore natural patterns or groupings.

### **Working:**

1. **Choosing the Number of Clusters (K)**  
   The first step is deciding how many clusters (K) you want. This is usually a number that the analyst sets based on their understanding of the problem or by using methods like the Elbow Method or Silhouette Score to find the best value.
2. **Placing the Centroids**  
   K random points are chosen from the data to act as the initial **centroids** (think of these as the centers of your clusters).
3. **Assigning Points to Clusters**  
   Each data point is then assigned to the closest centroid. The “closeness” is measured by calculating the **Euclidean distance** between the data point and each centroid.
4. **Recalculating Centroids**  
   After all data points have been assigned to a cluster, the centroids are updated. Each new centroid is calculated as the **average (mean)** of all the points in its cluster.
5. **Repeating the Process**  
   Steps 3 and 4 are repeated until the centroids don’t change much anymore or the assignments stop changing. This means the algorithm has converged, and the clustering is done.

The strength of K-Means lies in its simplicity. By grouping similar items and reducing the distance within clusters, it effectively uncovers the hidden patterns in the data. It’s especially useful for segmenting customers, grouping search results, organizing inventory, etc.

### **Use Cases in Real Life**

1. **Customer Segmentation**: Businesses often use K-Means to segment their customers based on purchase behavior, age, income, or spending patterns

.

1. **Document Classification**: Articles or emails can be grouped into categories based on topics or word usage.
2. **Image Compression**: It can group similar colors to reduce the size of image files without losing much quality.
3. **City Planning**: Government departments use it to group areas by population, income, traffic levels, etc., to plan better infrastructure.

### **Advantages**

1. Easy to understand and implement
2. Works well on large datasets
3. Efficient in terms of computation
4. Gives clear clusters if the data is well-separated

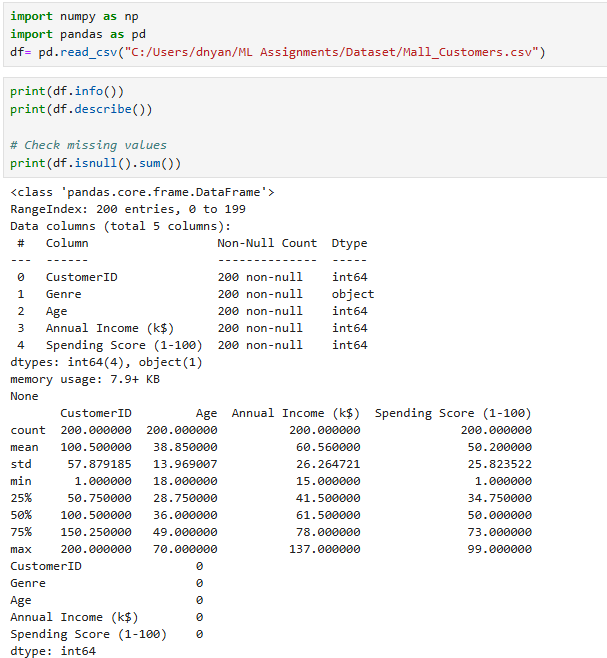
### **Limitations**

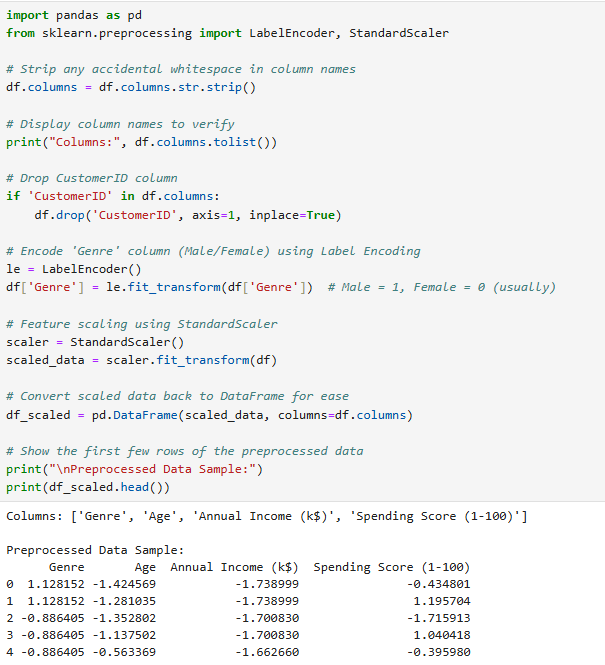
1. You need to know the value of K in advance
2. It doesn’t perform well with non-spherical clusters or data with different densities
3. Sensitive to outliers and noise
4. If the initial centroids are poorly chosen, results can vary (though K-Means++ helps with better initial centroid selection)

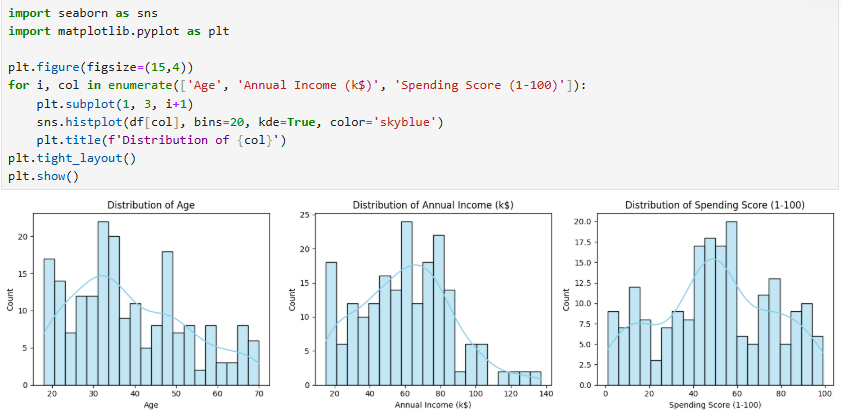
### **Key Concepts**

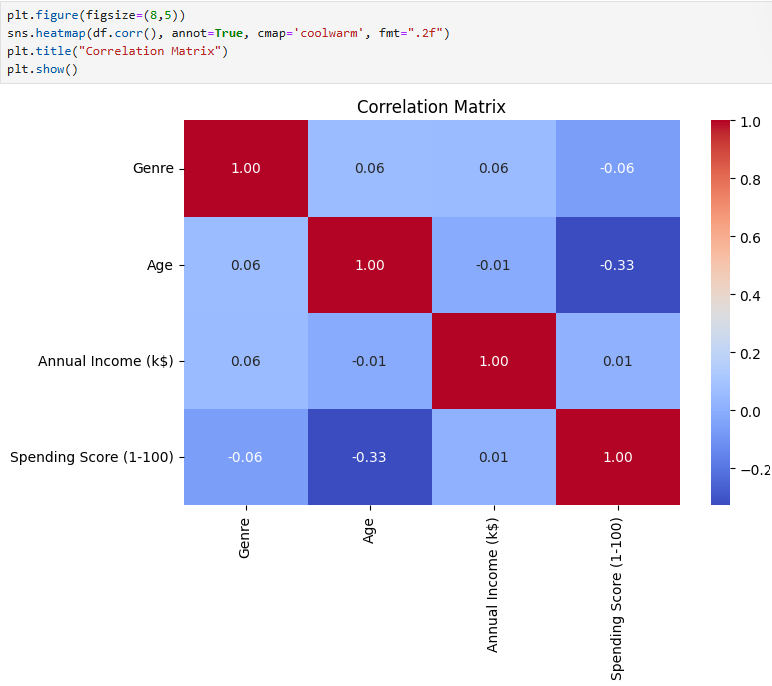
1. **Euclidean Distance**: Measures how far a data point is from the centroid (as the crow flies).
2. **Centroid**: The average position of all the points in a cluster.
3. **Inertia**: A measure of how tightly the data points are grouped in each cluster (lower is better).

**Code & Output**

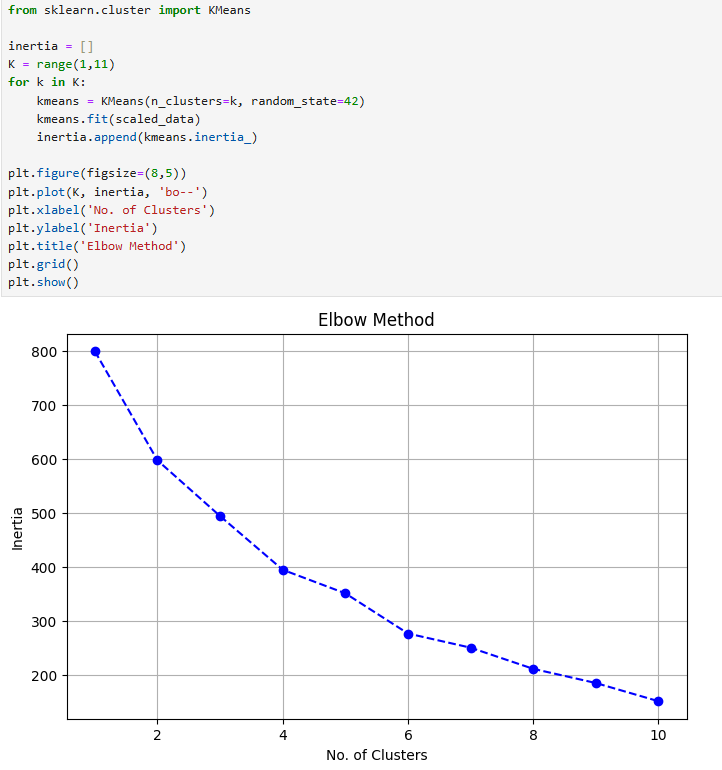
****

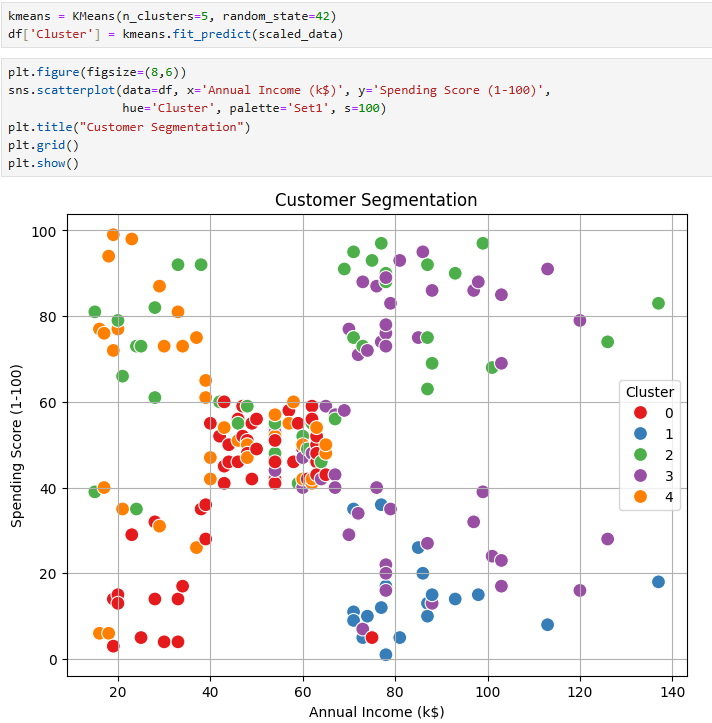
****

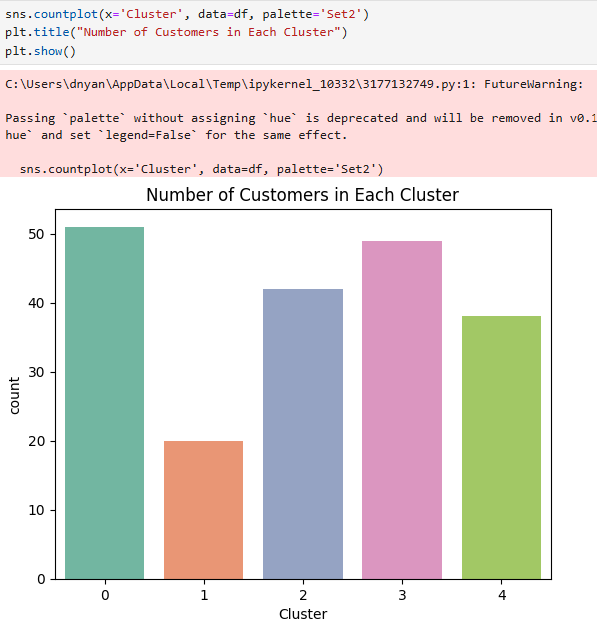


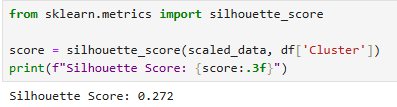












**Github :** [**https://github.com/dnyaneshwardhere/ML**](https://github.com/dnyaneshwardhere/ML)

**Conclusion:**

K-Means clustering effectively grouped mall customers based on their spending habits and income levels. It provided clear insights into different customer segments, which can be valuable for targeted marketing and strategic business decisions. The algorithm proved to be efficient, easy to implement, and interpretable, making it suitable for real-world customer segmentation tasks.