**Name: Abhishek Rajaram Walkoli**

**PRN: 123B2F155**

**Assignment No. 3**

**Problem Statement:**

Elegant description of the K-means problem.

**Objective:**

To implement the K-means clustering algorithm, split data into separate clusters according to similarities, and interpret results using visualization to derive meaningful insights.

**Prerequisite:**

1.Python environment with basic libraries like pandas, numpy, matplotlib, seaborn, and scikit-learn.

2.Internet access (optional for using external datasets).

3.Python basics, machine learning principles, and data visualization methodologies.

**Theory:**

Knowledge of the K-means Algorithm:

●K-means is an unsupervised machine learning algorithm for clustering data points into K clusters.

●It allocates each point to its closest cluster center (centroid) and updates the centroids in an iterative manner to reduce intra-cluster variance.

**Important Components:**

●K (Number of Clusters): The given number of groups into which the dataset is divided.

● Centroid: The typical point of a cluster, typically determined by the mean of all data points in the cluster.

● Inertia: The sum of squared distances from each data point to its closest centroid, quantifying cluster compactness.

Selecting the Best K:

It is important to find the optimal number of clusters (K) to enable effective segmentation.

● Elbow Method:

○Plot inertia (sum of squared distances) against varying values of k.

○The 'elbow point' is where the inertia drops rapidly before it becomes flat, implying an optimal K.

● Silhouette Score: Quantifies the quality of how well each point fits in its respective cluster.

● Gap Statistics: Compares the performance of clustering with the performance of random data distribution.

**Steps of the K-means Algorithm:**

1.Initialization: Choose K centroids at random from the dataset.

2.Assignment Step: Assign every data point to the closest centroid according to Euclidean distance.

3.Update Step: Calculate new centroids as the average positions of assigned data points in each cluster.

4.Iteration: Recursively execute steps 2 and 3 until centroids become stable or a convergence criterion is satisfied.

**Advantages of K-means:**

●Very efficient and scalable for large data sets.

● Simple and easy to understand.

● Well-suited for well-separated clusters.

**Limitations of K-means:**

● Pre-specifies number of clusters (K).

● Sensitive to noise and outliers.

● Can converge to local minima based on centroid initialization.

Visualization of Clusters:

● Scatter Plots: Illustrate how data points cluster around centroids.

● Heatmaps: Visualize cluster densities as well as feature correlations.

● 3D Plots: Apt for high-dimensional data representation.

Practical Applications of K-means Clustering:

● Customer Segmentation: Segregation of customers into groups according to purchasing habits.

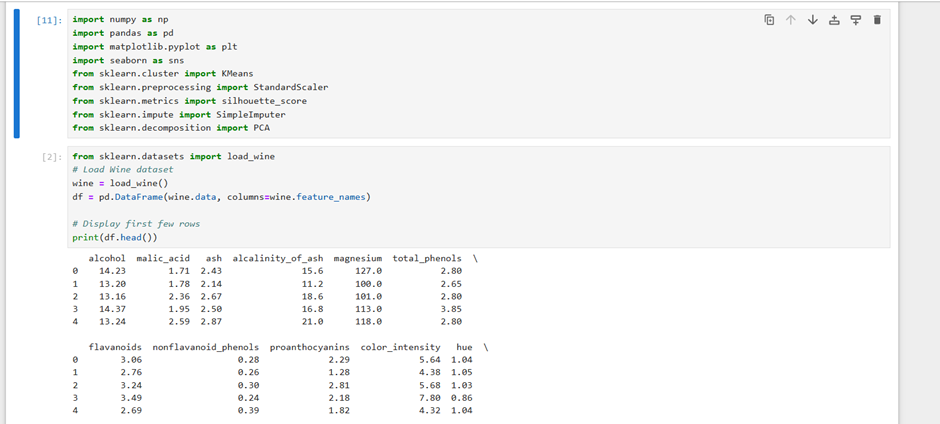
● Anomaly Detection: Detection of unusual patterns in datasets.

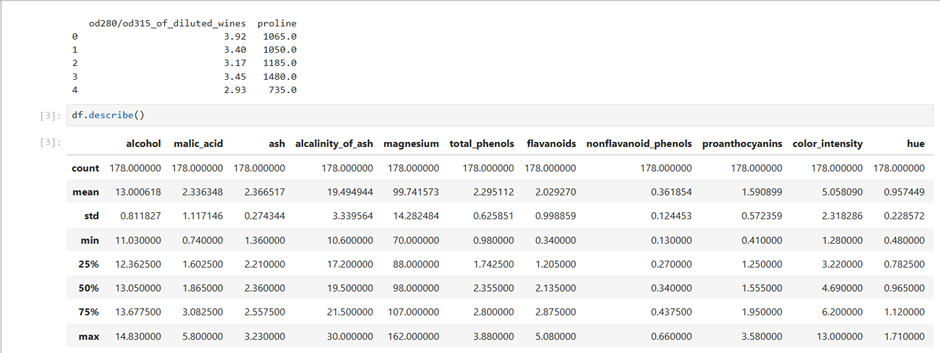
● Image Compression: Decreasing colors in an image by grouping similar pixels.

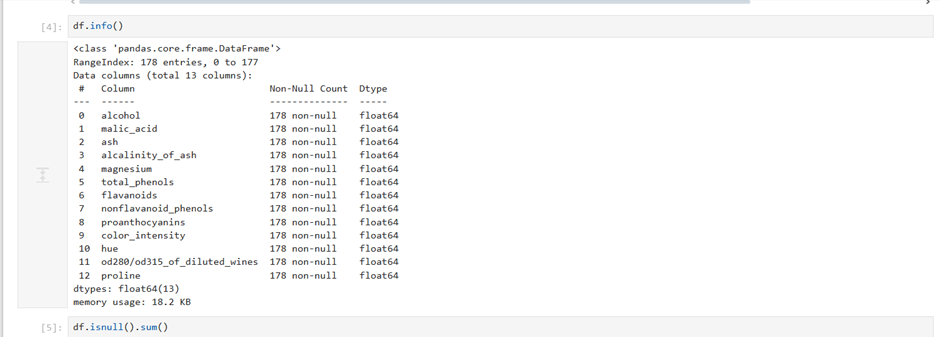
● Genetic Data Analysis: Discovering patterns in biological data sets.

Effective data segmentation is facilitated through these steps, and as a result, K-means clustering is a robust tool for exploratory data analysis as well as applications in real life.

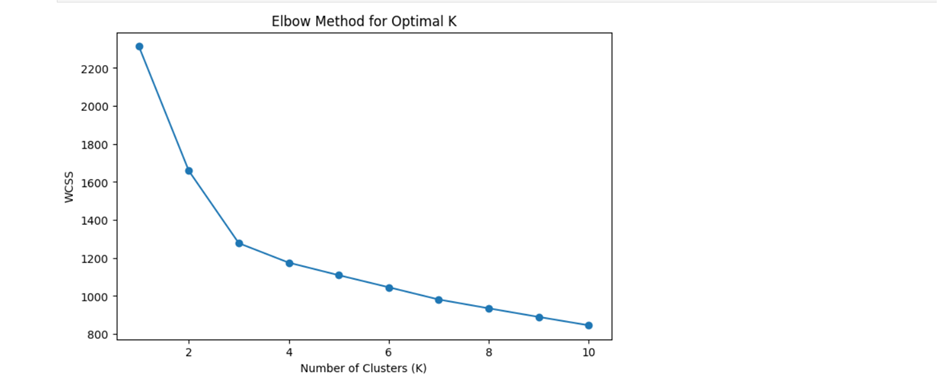
**Code & Output:**

****

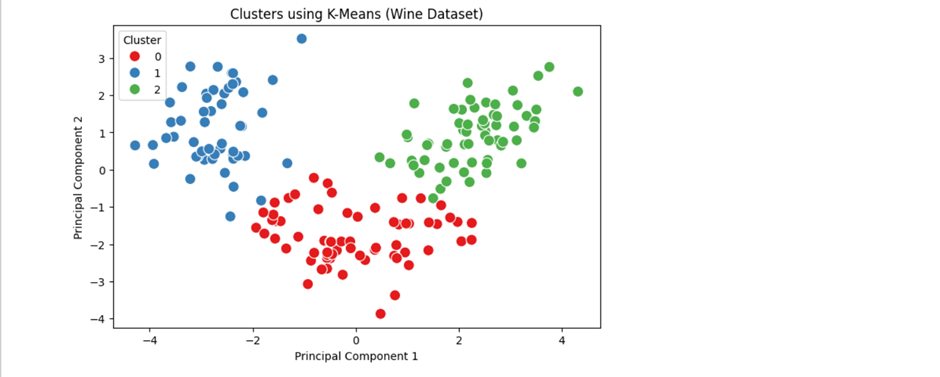
****

****

****

****

****

****

****

**Conclusion:**

K-means clustering effectively grouped the dataset into meaningful clusters. The Elbow Method determined the optimal number of clusters. Visualization showed clear separation between clusters, and the silhouette score validated cluster quality. Further improvements can include tuning initialization or using alternative clustering algorithms for better results.