Machine Learning
Assignment 1

Name: Ruthvik Reddy Anugu
Student Id: 001096522
NetId: ranugu3
Date: 09/12/23

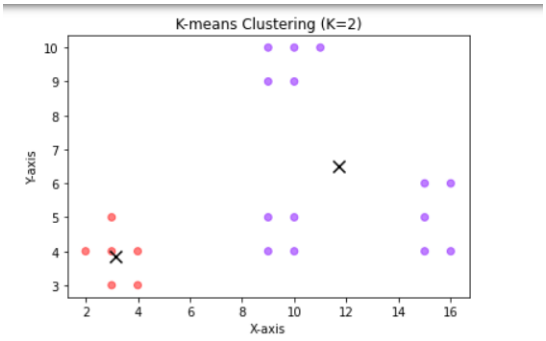
Abstract: This project is divided into two parts:

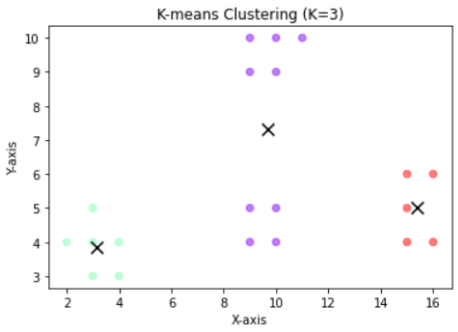
K-means Clustering Analysis: In the first section, the code runs a custom K-means clustering method with a specific number of iterations on the Iris dataset to determine both the best and worst clustering results based on the sum of squared distances (inertia). The original data and the findings are then visualized for comparison.

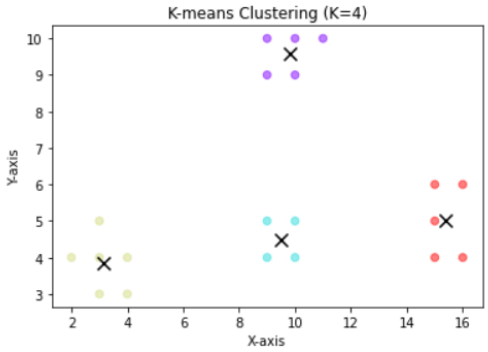
K-means Clustering with sklearn: The second section employs the sklearn package to perform K-means clustering on a new dataset ('kmtest.csv') for a variety of K values. It depicts the data points in various colors to indicate clusters and displays the cluster centers.

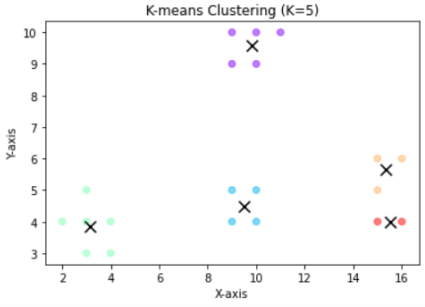
Overall, this research investigates K-means clustering and its effects on various datasets, contrasting bespoke implementation with sklearn.

Source code Output:
1a. Without Normalization:

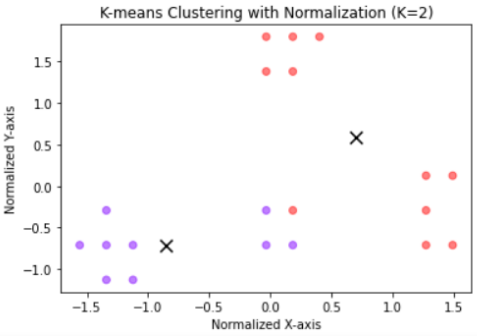


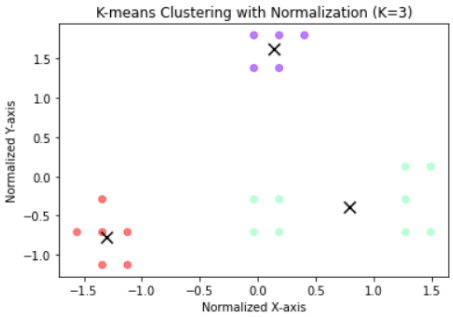


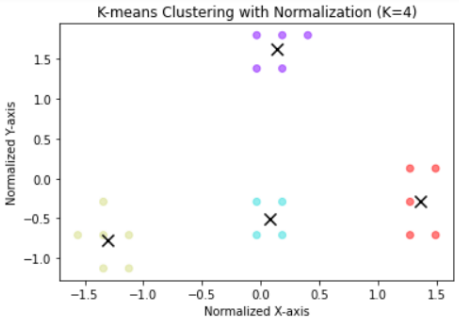


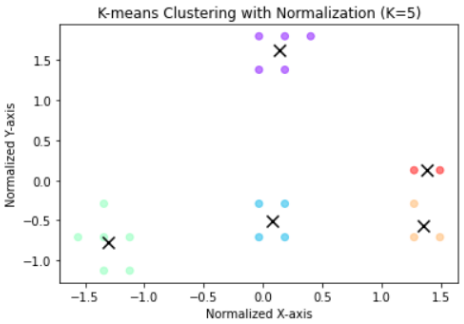


1b. With Normalization:

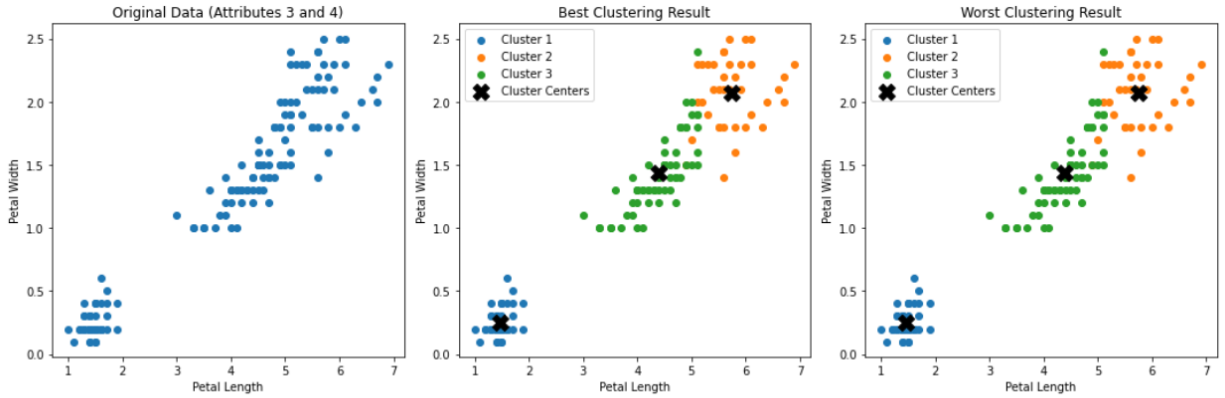








2. a,b,c,d




```
Distance between best result centers and original centers:
[2.99920056 0.1953708  3.16307025]
```

Discussion and conclusion:

Part 1: Personalized K-means Clustering:

The first phase of the project involved developing and applying a proprietary K-means clustering method to the Iris dataset. The programme ran K-means clustering with K=3 5 times to identify the best and worst clustering results based on the sum of squared distances (inertia). The following are the important findings and discussions:

Clustering Results: Best and Worst:

The best clustering result had the lowest inertia, indicating that the data points were firmly packed around the cluster centers.
The worst clustering result had the largest inertia, indicating inadequate cluster separation.
Visualization:

The raw data was represented in a scatter plot, with petal length on the x-axis and petal width on the y-axis.
The best and worst clustering findings were plotted separately, with each cluster colored differently.
Black 'X' marks were used to mark the cluster centers.
Calculate Distance:

The distance between the best clustering result's centers and the original centers (based on attributes 3 and 4) was computed and displayed.
Part 2 - Sklearn K-means Clustering:

The KMeans module of sklearn was used in the second portion of the research to perform K-means clustering on a new dataset ('kmtest.csv') for various values of K. The following are the important findings and discussions:

Clustering using K-means with sklearn:

K-means clustering was carried out for K values of 2, 3, 4, and 5, revealing how varying cluster counts effect data grouping.
Visualization:

Scatter plots were created for each K value, with data points colored according to their cluster assignment.
Black 'X' marks were used to mark the cluster centers.
Discussion:

K-means clustering divides data into clusters based on similarities. The results demonstrated how the choice of K can have a considerable impact on the clustering outcome. Domain knowledge or approaches such as the elbow method should be used to guide the selection of K.

The custom K-means implementation illustrated the significance of clustering initialization and convergence. Because to better initialization and convergence, the best result obtained decreased inertia.

The calculated distance between the best outcome centers and the original centers offered a quantitative indication of how closely the clustering matched the expected centers.
Lessons Discovered:

Using bespoke machine learning algorithms allows you to better grasp their inner workings and limitations.

Visualization is essential for evaluating clustering findings and extracting insights from data.

Experience:

Working on this project allowed me to gain practical expertise with K-means clustering and visualization techniques in both custom and sklearn implementations.
Future Projects:

Future studies could include comparing the performance of other clustering algorithms, such as DBSCAN or hierarchical clustering, against that of K-means.

For better clustering results, more thorough parameter adjustment and feature engineering could be investigated.

Real-world datasets with varying characteristics and complexities could be utilized to investigate the merits and drawbacks of clustering methods in greater depth.