CSCI 2410 Introduction to Data Analytics Using Python
Homework Assignment #4

HW Programming #4: - Data Analytics with K-Means Clustering

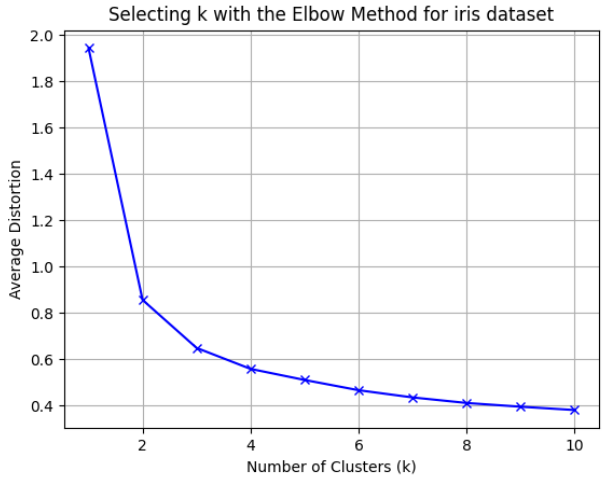
Tasks: Experiment with the **K-means** clustering technique on 'iris.xlsx' dataset.

Assignment Instructions:

1. Run the **K-Means** clustering technique on the sheet 1 of 'iris.xlsx' dataset from your local directory
   (1) **[30%]** Run with k = 2, k= 3, k = 4, k = 5, and k = 6, respectively
   Done. Next page
   (2) **[10%]** Get the **silhouette coefficients** for each run of the k values
   Done. Silhouette coefficients are [0.6808, 0.5526, 0.4978, 0.4885, 0.3682]
   (3) **[30%]** Plot the clustered data in each of the attribute pairs (Total 6 plots) for each run of k
   Done. Next page
   (4) **[30%]** Calculate the optimal **k** number by using the **elbow method**
   Done

   Python libraries needed: numpy, pandas, sklearn.cluster-KMeans, klearn.metrics-silhouette_score, matplotlib.pyplot, scipy.spatial.distance-cdist

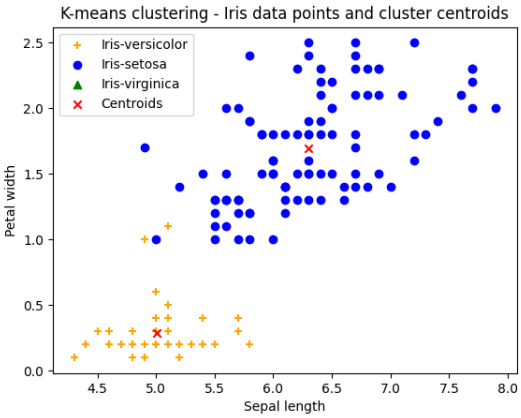
```python
def elbow_method():
    try:
        df = pd.read_excel( io: 'iris.xlsx', sheet_name=f'Sheet1')
        x = df.iloc[:, 0:4].values
        k_range = range(1, 11)
        mean_distortions = []

        for k in k_range:
            kmeans = KMeans(n_clusters=k, random_state=0, n_init=10).fit(x)
            mean_distortions.append(sum(np.min(cdist(x, kmeans.cluster_centers_,
                                                    metric: 'euclidean'), axis=1) / x.shape[0]))

        plt.plot( *args: k_range, mean_distortions, 'bx-')
        plt.xlabel('Number of Clusters (k)')
        plt.ylabel('Average Distortion')
        plt.title('Selecting k with the Elbow Method for iris dataset')
        plt.grid(True)
        plt.show()
    except Exception as e:
        print(e)
```

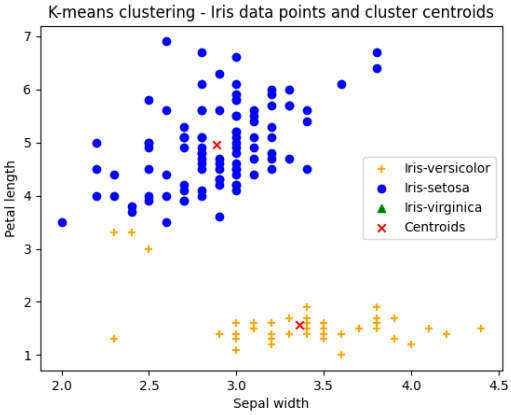


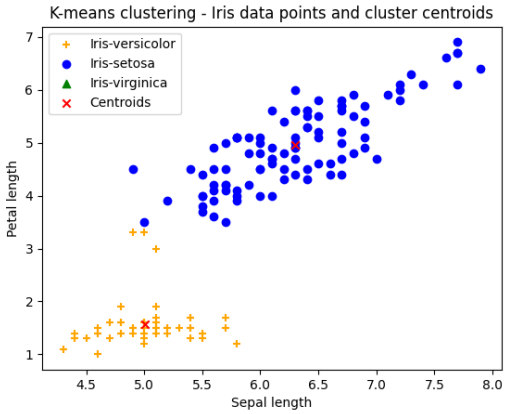
```python
import numpy as np
import pandas as pd
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
import matplotlib.pyplot as plt
from scipy.spatial.distance import cdist
```

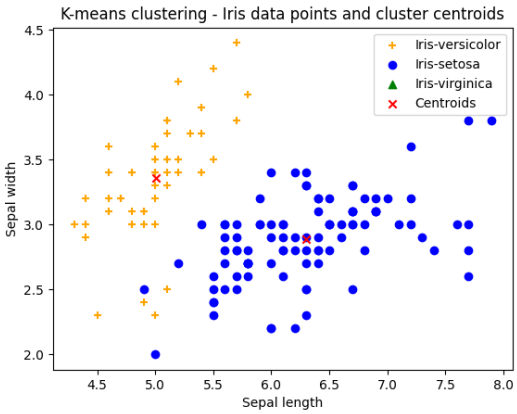
```python
def k_mean_clustering():
    try:
        k = 10
        df = pd.read_excel( io: 'iris.xlsx', sheet_name=f'Sheet1')
        print(f'The length(rows) of dataset: {len(df)}')

        x = df.iloc[:, 0:4].values
        kmeans = KMeans(n_clusters=k, random_state=0, n_init=10).fit(x)
        print(f'Running with the k value of {k}')

        print('The cluster centroids:')
        for centroid in kmeans.cluster_centers_:
            formatted_centroid = [f'{coord:.2f}' for coord in centroid]
            print(formatted_centroid)

        print('The Sample Clusters:')
        print(kmeans.labels_)

        # Define colors and symbols for plotting
        colors = {0: 'orange', 1: 'blue', 2: 'green'}
        symbols = {0: '+', 1: 'o', 2: '^'}
        cluster_species = {0: 'Iris-versicolor', 1: 'Iris-setosa', 2: 'Iris-virginica'}

        # Define pairs of columns to create cluster plots
        column_pairs = [(0, 1), (0, 2), (0, 3), (1, 2), (1, 3), (2, 3)]
        column_names = {0: 'Sepal length', 1: 'Sepal width', 2: 'Petal length', 3: 'Petal width'}

        for pair in column_pairs:
            x_label, y_label = pair
            plt.figure()
            for i in range(3):
                cluster_points = x[kmeans.labels_ == i]
                plt.scatter(cluster_points[:, x_label], cluster_points[:, y_label], c=colors[i], marker=symbols[i],
                            label=f'{cluster_species[i]}')

            plt.scatter(kmeans.cluster_centers_[:, x_label], kmeans.cluster_centers_[:, y_label], c='red', marker='x',
                        label='Centroids')

            plt.xlabel(column_names[x_label])
            plt.ylabel(column_names[y_label])
            plt.title(f'K-means clustering - Iris data points and cluster centroids')
            plt.legend()
            plt.show()

        print(f'The silhouette_score: {silhouette_score(x, kmeans.labels_):.4f}')
    except Exception as e:
        print(e)
```

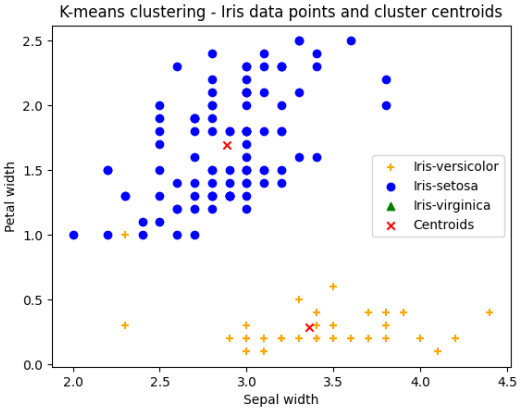
```
The length(rows) of dataset: 150
Running with the k value of 2
The cluster centroids:
['5.01', '3.36', '1.56', '0.29']
['6.30', '2.89', '4.96', '1.70']
The Sample Clusters:
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1]
The silhouette_score: 0.6808
```

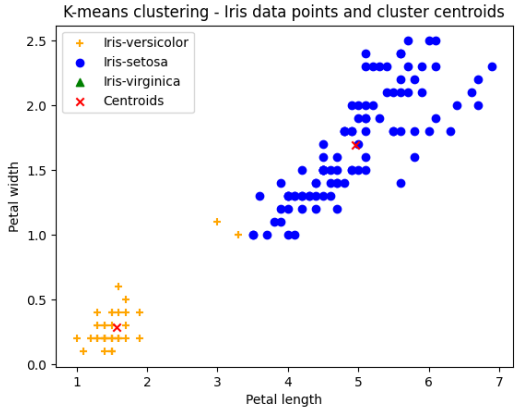




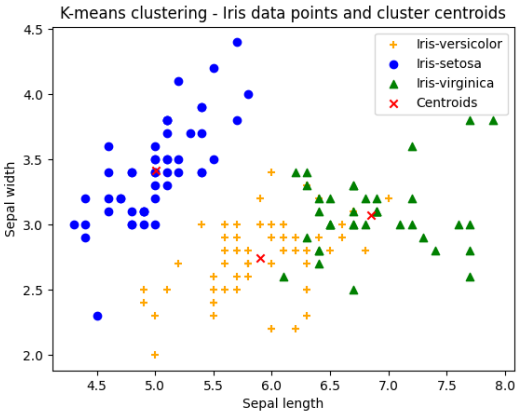


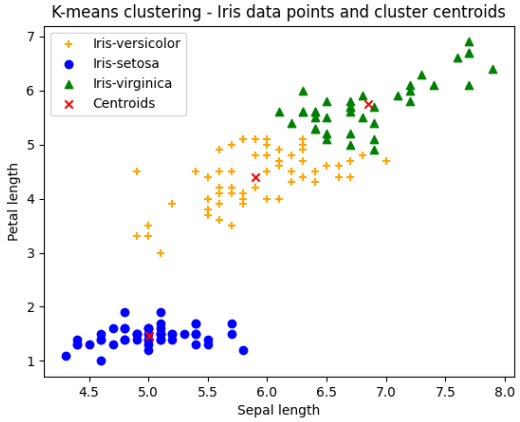


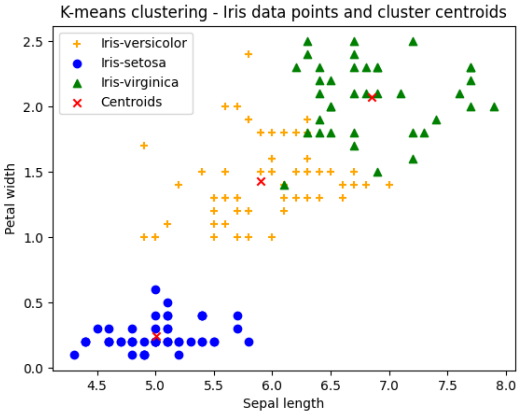


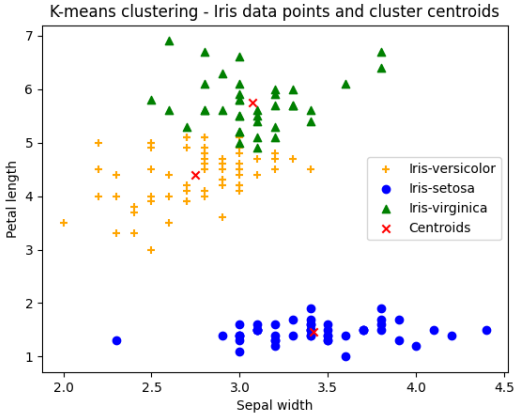


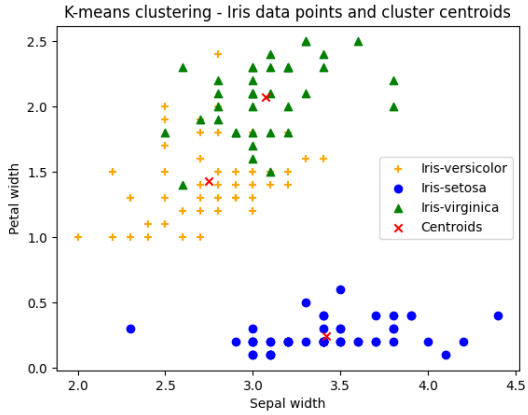
```
The length(rows) of dataset: 150
Running with the k value of 3
The cluster centroids:
['5.90', '2.75', '4.39', '1.43']
['5.01', '3.42', '1.46', '0.24']
['6.85', '3.07', '5.74', '2.07']
The Sample Clusters:
[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 2 2 2 2 0 2 2 2 2
 2 2 0 0 2 2 2 2 0 2 0 2 0 2 2 0 0 2 2 2 2 2 0 2 2 2 2 0 2 2 2 0 2 2 2 0 2
 2 0]
The silhouette_score: 0.5526
```

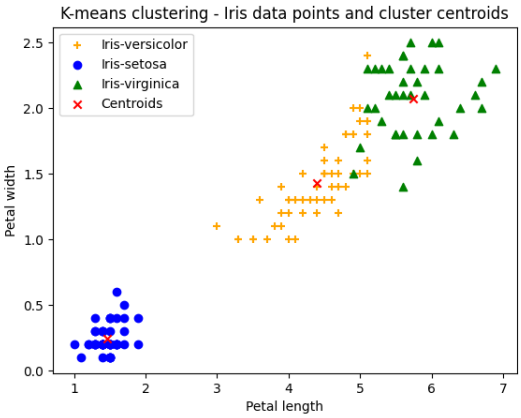




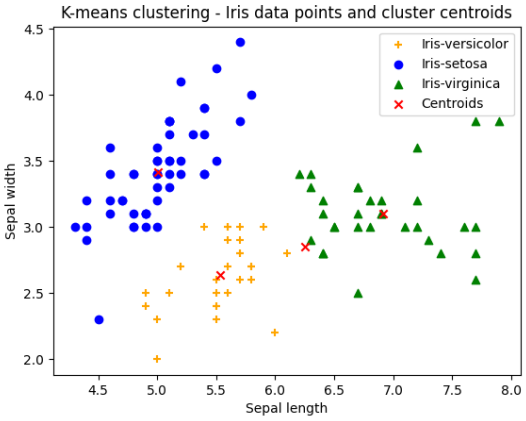


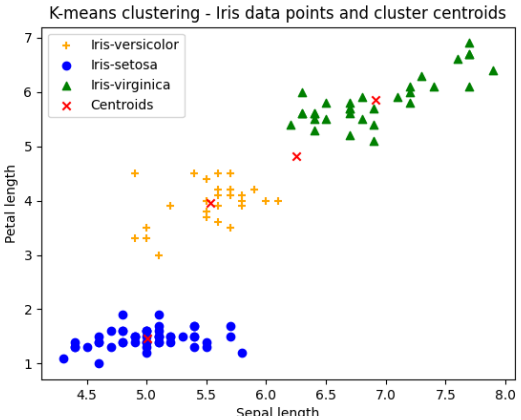


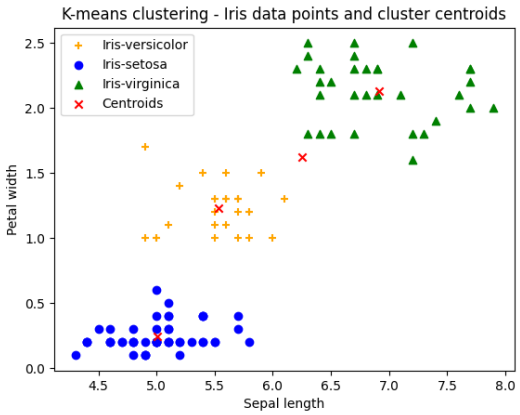


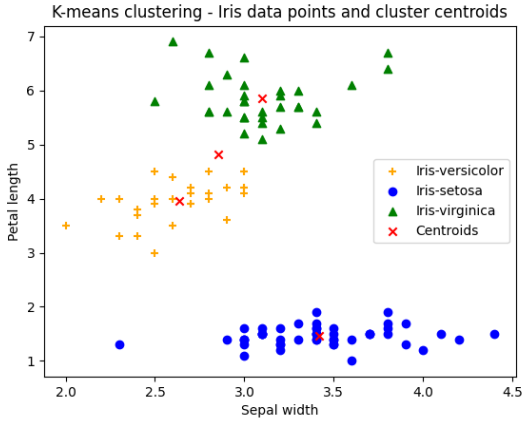


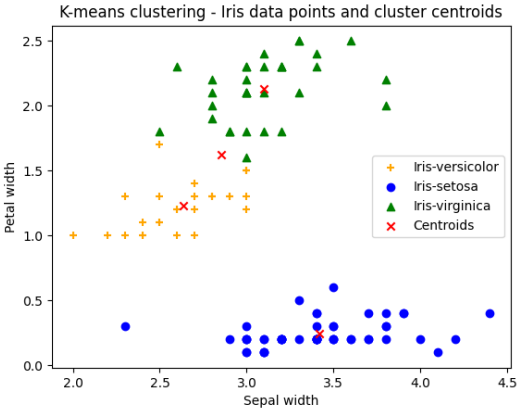
```
The length(rows) of dataset: 150
Running with the k value of 4
The cluster centroids:
['5.53', '2.64', '3.96', '1.23']
['5.01', '3.42', '1.46', '0.24']
['6.91', '3.10', '5.85', '2.13']
['6.25', '2.85', '4.82', '1.62']
The Sample Clusters:
[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 3 0 3 0 3 0 3 0 0 0 0 3 0 3 0 0 3 0 3 0 3 3
 3 3 3 3 3 0 0 0 0 3 0 3 3 3 0 0 0 3 0 0 0 0 3 0 0 2 3 2 2 2 2 0 2 2 2 3
 3 2 3 3 2 2 2 2 3 2 3 2 3 2 2 3 3 2 2 2 2 2 3 3 2 2 2 3 2 2 2 3 2 2 2 3 3
 2 3]
The silhouette_score: 0.4978
```

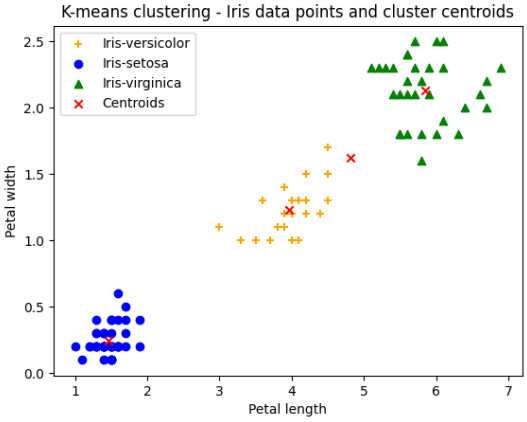




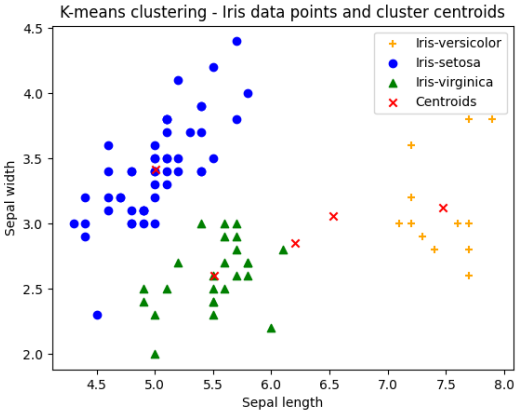


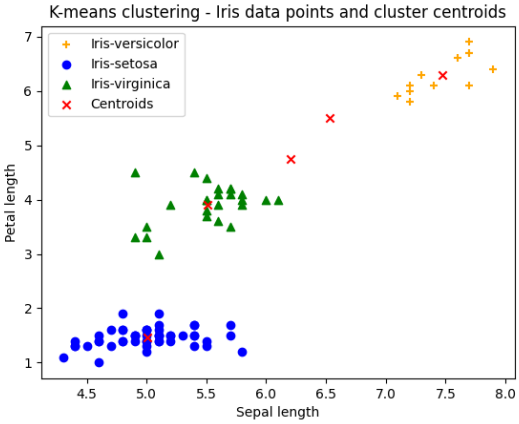


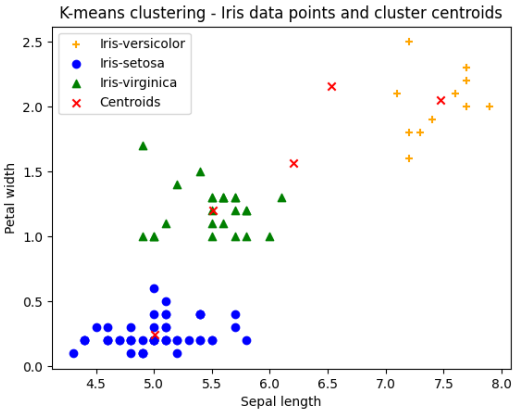


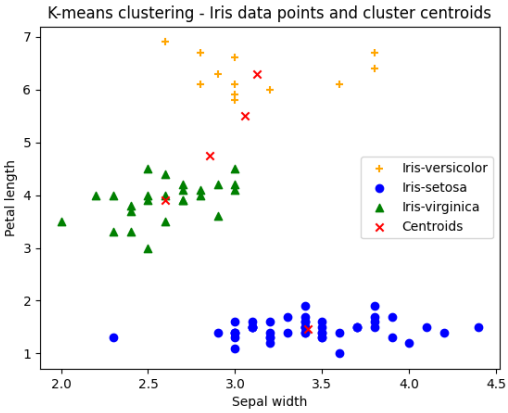


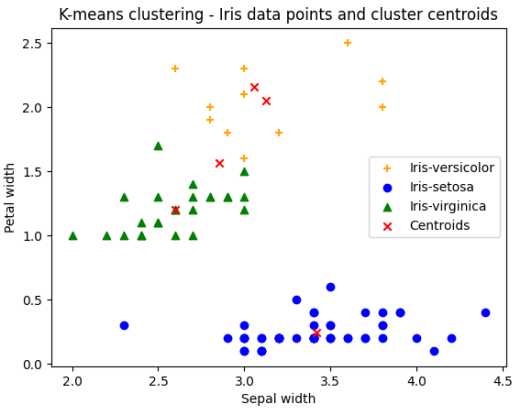
```
The length(rows) of dataset: 150
Running with the k value of 5
The cluster centroids:
['7.47', '3.12', '6.30', '2.05']
['5.01', '3.42', '1.46', '0.24']
['5.51', '2.60', '3.91', '1.20']
['6.53', '3.06', '5.51', '2.16']
['6.21', '2.85', '4.75', '1.56']
The Sample Clusters:
[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 4 4 4 2 4 4 4 2 4 2 2 4 2 4 2 4 4 2 4 2 4 2 4 4
 4 4 4 4 4 2 2 2 2 4 2 4 4 4 2 2 2 4 2 2 2 2 2 4 2 2 3 4 0 3 3 0 2 0 3 0 3
 3 3 4 3 3 3 0 0 4 3 4 0 4 3 0 4 4 3 0 0 0 3 4 4 0 3 3 4 3 3 3 4 3 3 3 4 3
 3 4]
The silhouette_score: 0.4885
```

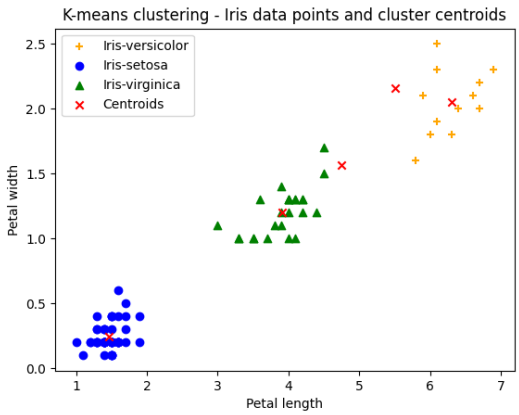




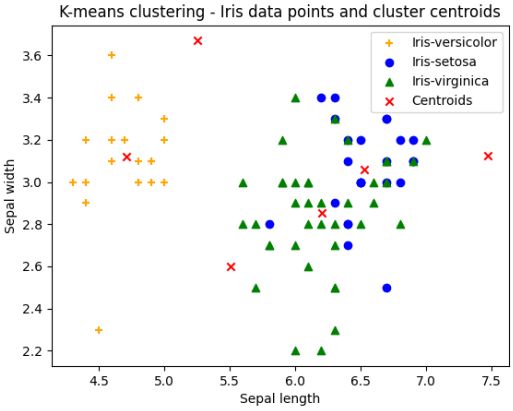


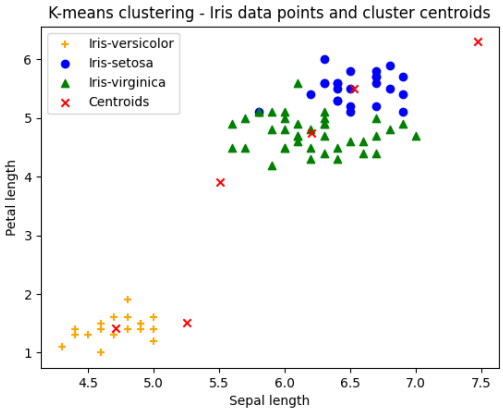


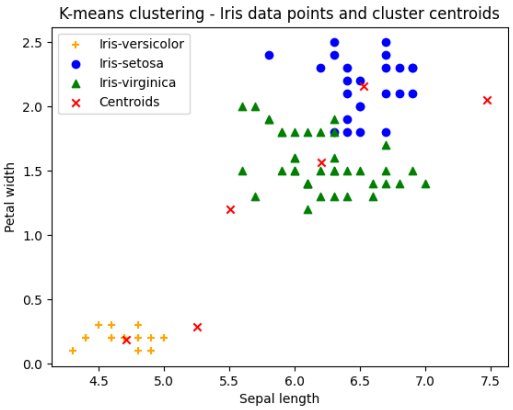


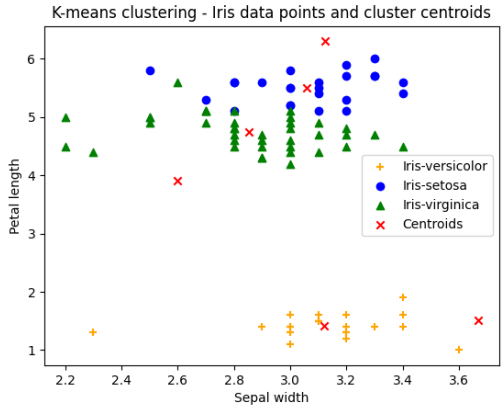


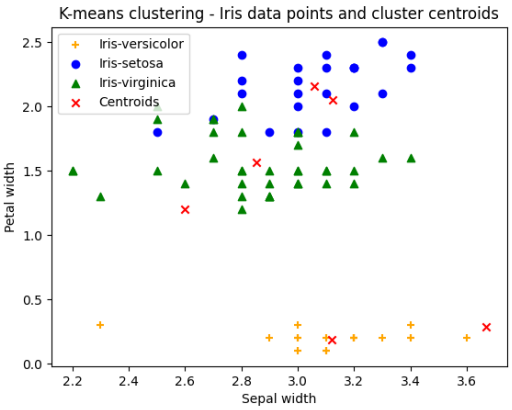
```
The length(rows) of dataset: 150
Running with the k value of 6
The cluster centroids:
['4.71', '3.12', '1.42', '0.19']
['6.53', '3.06', '5.51', '2.16']
['6.21', '2.85', '4.75', '1.56']
['7.47', '3.12', '6.30', '2.05']
['5.51', '2.60', '3.91', '1.20']
['5.26', '3.67', '1.50', '0.29']
The Sample Clusters:
[5 0 0 0 5 5 0 5 0 0 5 0 0 0 5 5 5 5 5 5 5 5 0 5 0 0 5 5 5 0 0 5 5 5 0 0 5
 0 0 5 5 0 0 5 5 0 5 0 5 0 2 2 2 4 2 2 2 4 2 4 4 2 4 2 4 2 2 4 2 4 2 4 2 2
 2 2 2 2 4 4 4 4 2 4 2 2 2 4 4 4 2 4 4 4 4 4 2 4 4 1 2 3 1 1 3 4 3 1 3 1
 1 1 2 1 1 1 3 3 2 1 2 3 2 1 3 2 2 1 3 3 3 1 2 2 3 1 1 2 1 1 1 2 1 1 1 2 1
 1 2]
The silhouette_score: 0.3682
```

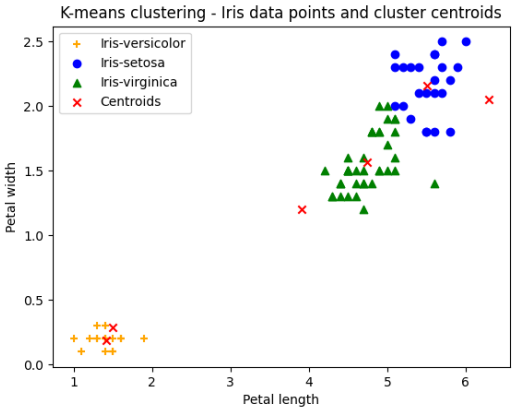












Requirements for the Submission of Programming/Homework Assignments

1. Well-documented program list (the .py files)
   20% of total points if no .py file submitted.
   Done

2. Three annotated program test and run examples (screenshots) that **show different and representative test cases** with **input, output, and the parameter settings of the program runs clearly marked/annotated**. You can do the annotations by
   (1) Pasting the screenshots into a WORD document,
   Done
   (2) Editing on the WORD document pages for the required marks and annotations,
   Done.
   Testing and running examples, as well as annotations, were provided inside the screenshots.
   (3) Converting the document to pdf for submission (it is ok to submit the WORD file directly without converting to pdf).
   Done
   20% of total points will be taken off if run examples are not representative.
   20% of total points will be taken off if run examples are not clearly marked/annotated.

3. A discussion page
   (a) Hardware and software used by your program,
   I completed this assignment using my personal computer with PyCharm Professional Version: 2023.2.1.
   (b) Features of your program, e.g., data structures, algorithms, programming styles, etc.
   The program incorporates various data analysis and visualization features. It employs the K-means clustering algorithm to analyze the Iris dataset and includes an elbow method for determining the optimal number of clusters. The program effectively loads, analyzes, and visualizes data.
   (c) Problems you encountered during your work, and
   None
   (d) Assigned discussion problems, if there is any.
   No assigned discussion problems
   (e) Fill in the following table and submit it along with your above submissions.

Total (approximate) time spent on the assignment	14 hours	Total (approximate) time for the correction part	2 hours

Problems and difficulties encountered	None
Reflections (good and bad) on the assignment	Good: A snippet of lines of code was provided Bad: None
Any comments and suggestions	None

20% of total points will be taken off if no discussion page is submitted.