Homework 3
Group-based Assignment

Due May 2nd, 2023

For this homework, we will use the Old Faithful Geyser dataset, which you can download here. This dataset describes the properties of eruptions of the Old Faithful geyser, located in Yellowstone National Park, Wyoming, USA. There are two numeric attributes per instance: the length of time of the eruption, in minutes, and the waiting time until the next eruption, also in minutes. The geyser was named "Old Faithful" because its eruption patterns are very reliable. See here for more information, if you are interested.

Deliverable:

- Python Notebook to be uploaded to GitHub and shared with instructor/TA, or, Google Collab notebook shared with comment option.
- Submit on the blackboard the link either to Github or link to Google Collab notebook
- Please label each of the questions clearly in your notebook

Problem 1 (25 points)

(a) Create and print out a scatter plot of this dataset, eruption time versus waiting time. (10 points)

(b) How many clusters do you see based on your scatter plot? For the purposes of this question, a cluster is a "blob" of many data points that are close together, with regions of fewer data points between it and other "blobs"/clusters. (5 points)

(c) Describe the steps of a hierarchical clustering algorithm. Based on your scatter plot, would this method be appropriate for this dataset? (10 points)

Problem 2 (75 points)

Implement the k-means algorithm in Python, and use it to perform clustering on the Old Faithful dataset. Use the number of clusters that you identified in Problem 1. Be sure to ignore the first column, which contains instance ID numbers. In your notebook, including the following items:

(a) Your source code for the k-means algorithm. **You need to implement the algorithm from scratch.** Don't forget to add comments to your code to make it readable **(**45 points)

(b) A scatter plot of your final clustering, with the data points in each cluster color-coded, or plotted with different symbols. Include the cluster centers in your plot. (10 points)

(c) A plot of the k-means objective function versus iterations of the algorithm. Recall that the objective function is (10 points)

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} \|p - c_i\|^2 \ ,$$

where k is the number of clusters, C_i is the set of instances assigned to the ith cluster, and c_i is the cluster center for the ith cluster. Note that the objective function should always decrease. If this is not the case, look for a bug in your code.

(d) Did the method manage to find the clusters that you identified in Problem 1? If not, did it help to run the method again with another random initialization? (10 points)