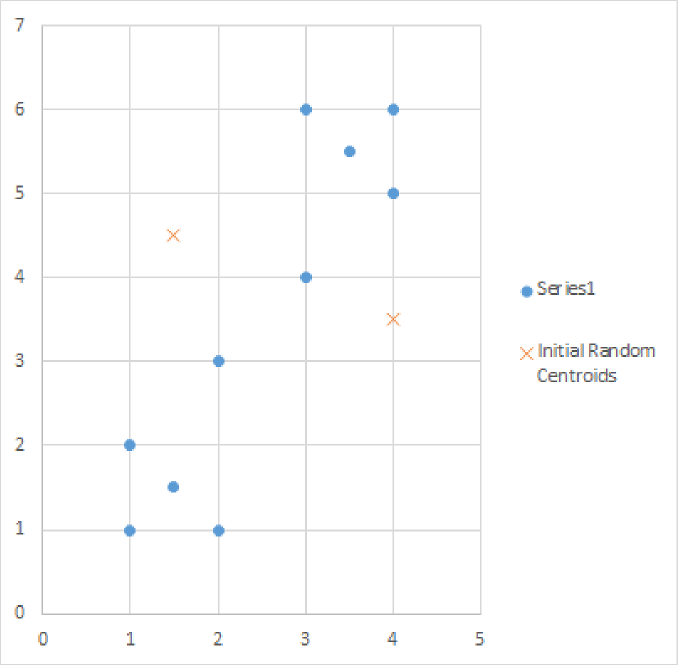
Homework Assignment 1 – [30 points]

STAT430 Unsupervised Learning – Fall 2023

<u>Due</u>: Friday, September 1 11:59pm CST on Canvas

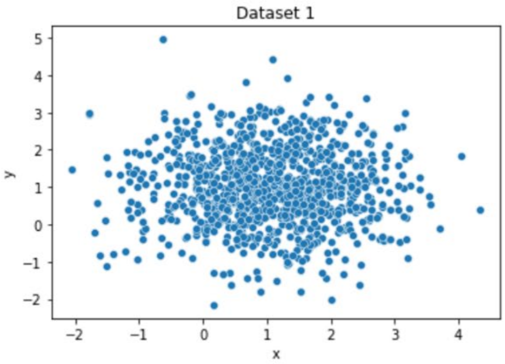
Question #1: [4 pt] Plotted and shown below is a two-dimensional dataset with 10 objects. Also plotted below are two centroids that have been randomly initialized to be (1.5,4.5) and (4,3.5). What will be the NEXT position of the two centroids in the first step of the k-means algorithm? Show your work.

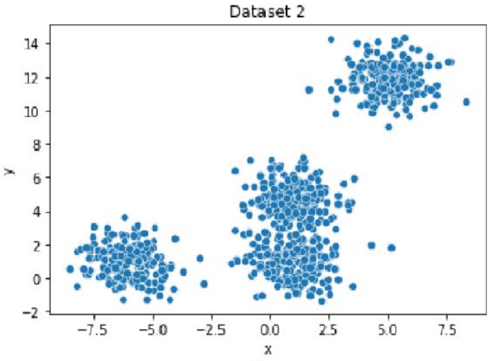
	Data		**Additional Information**	
	x	**y**	**Squared Distance to Initial Random Centroid 1 (1.5,4.5)**	**Squared Distance to Initial Random Centroid 2 (4,3.5)**
Object 1	1	1	12.50	15.25
Object 2	2	3	2.50	4.25
Object 3	1	2	6.50	11.25
Object 4	2	1	12.50	10.25
Object 5	1.5	1.5	9.00	10.25
Object 6	3	4	2.50	1.25
Object 7	3	6	4.50	7.25
Object 8	4	5	6.50	2.25
Object 9	4	6	8.50	6.25
Object 10	3.5	5.5	5.00	4.25

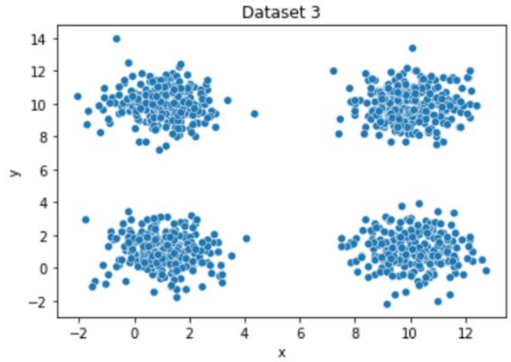


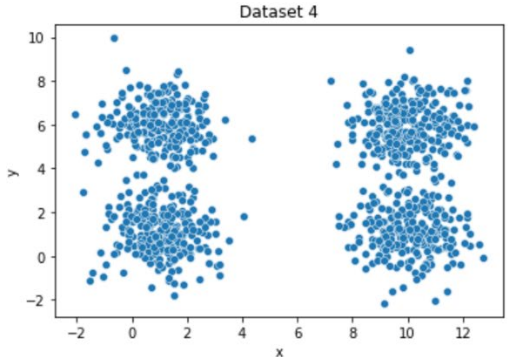
Question #2: [4 pt] Match the dataset (1-4) to the k-means elbow plot (A-D) that was created from this dataset. (Explanations not required, but may help with partial credit if you are wrong.)

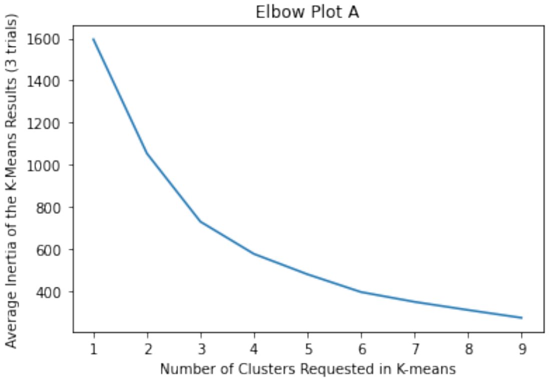
<u>Hint</u>: Pay close attention to the axes in these graphs.

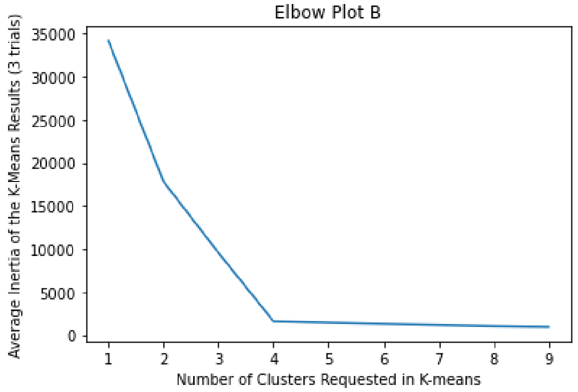


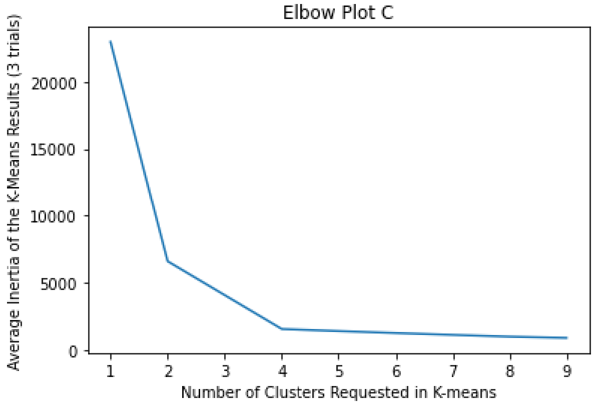


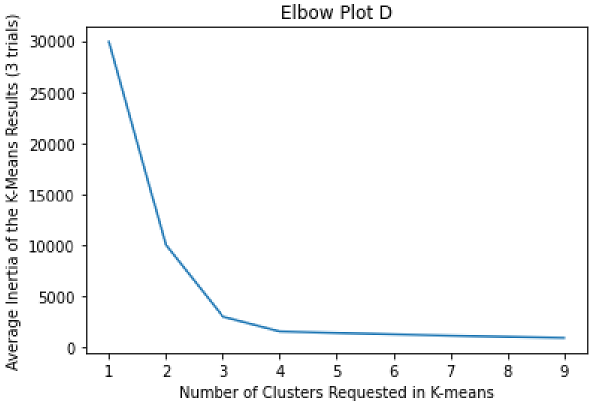




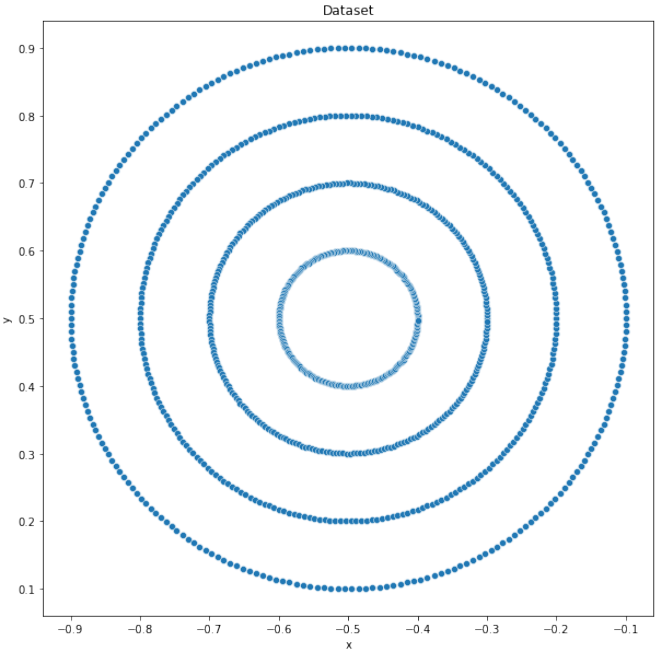






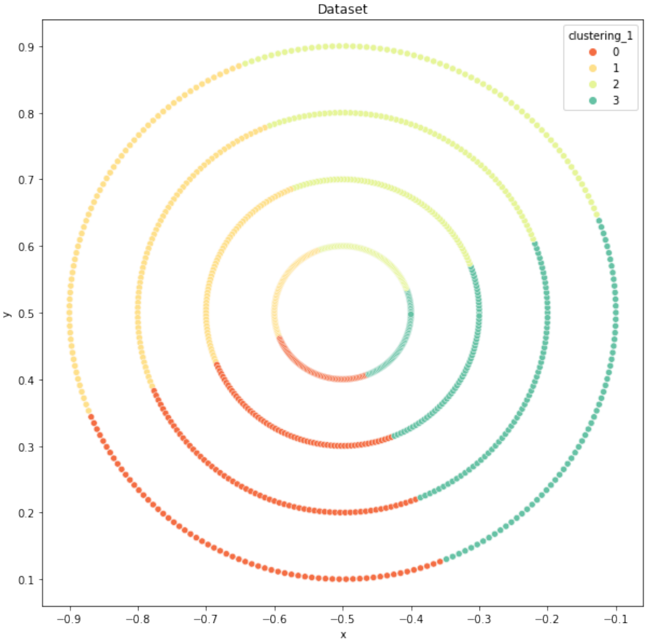


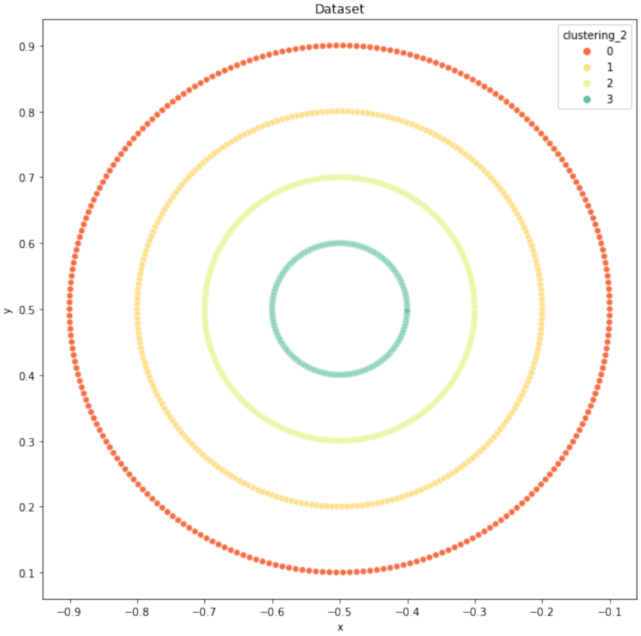
Question #3: [5 pt] (1 point for each problem) The data displayed below shows four “clusters”. We can see that in a dataset such as this, a meaningful cluster is one of the four concentric circles of points.



Displayed below are two potential clusterings of the same dataset (ie. Clustering 1 and Clustering 2). Each clustering has 4 clusters, which are color-coded.

We can see that clustering 2 successfully identifies the 4 clusters that we would ideally be looking for a clustering algorithm to return in a dataset such as this.





a. For Clustering 1, approximate where the centroids of the four clusters would be (drawing on the graph or giving an approximate numerical point is fine).

b. For Clustering 2, approximate where the centroids of the four clusters would be (drawing on the graph or an approximate numerical point is fine).

c. One of these clusterings has an inertia of 75 and the other clustering has an inertia of 24. Which inertia do you think corresponds to which clustering? Explain why.

d. Which final clustering do you think the k-means clustering algorithm is more likely to return: clustering 1 or clustering 2? Why?

e. Rather than using the Euclidean distance to measure the “similarity” between two objects, come up with another way to measure “similarity” between two objects in which:
   a. objects in the <u>same cluster</u> in clustering 2 are considered “similar” to one another.
   b. objects in <u>different clusters</u> in clustering 2 are considered “dissimilar” to one another.

Explain. There are many possible answers for this!

Question #4 [2.5 pts]:

Complete the “Getting to Know you Survey” on Canvas quizzes.

Question #5 [14.5 pts]:

1. Download the Assignment_01.zip file from Canvas.
2. Edit the Jupyter notebook (.ipynb) file to complete/answer questions 4.1-4.10.
3. Submit your completed Jupyter notebook (.ipynb) file as well as any other files you used to answer Questions 1-3 to Canvas.