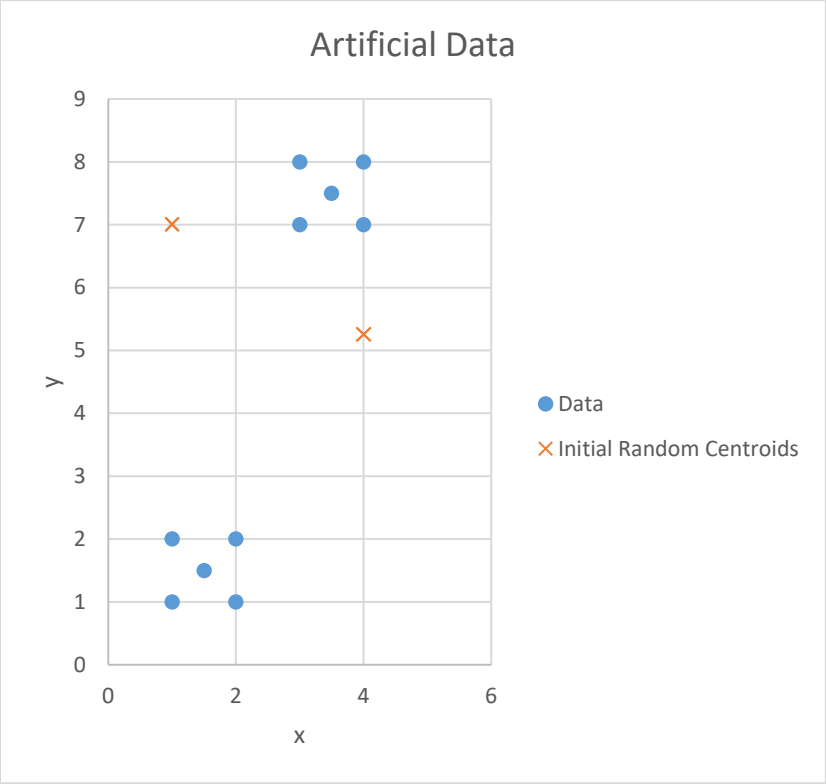
Homework Assignment 1 – [30 points]

STAT430 Unsupervised Learning - Fall 2021

Due: Friday, September 3 on Compass

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 (1,7) and (4,5.25). What will be the NEXT position of the two centroids in the first step of the k-means algorithm? Show your work.

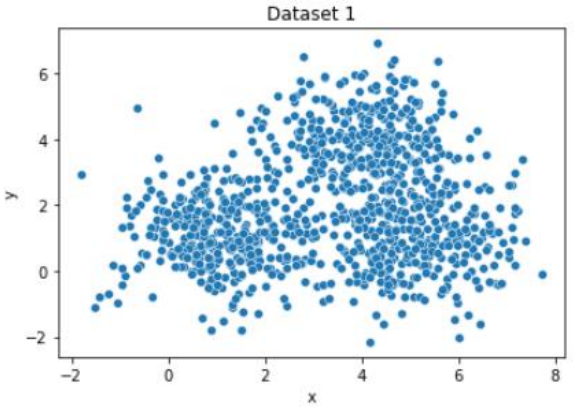
	Data		Additional Information		
	x	y	Squared Distance to Initial Random Centroid 1 (1,7)	Squared Distance to Initial Random Centroid 2 (4,5.25)	With the random initialization point belongs to
Object 1	1	1	36.00	27.06	centroid 2
Object 2	2	2	26.00	14.56	centroid 2
Object 3	1	2	25.00	19.56	centroid 2
Object 4	2	1	37.00	22.06	centroid 2
Object 5	1.5	1.5	30.50	20.31	centroid 2
Object 6	3	7	4.00	4.06	centroid 1
Object 7	3	8	5.00	8.56	centroid 1
Object 8	4	7	9.00	3.06	centroid 2
Object 9	4	8	10.00	7.56	centroid 2
Object 10	3.5	7.5	6.50	5.31	centroid 2

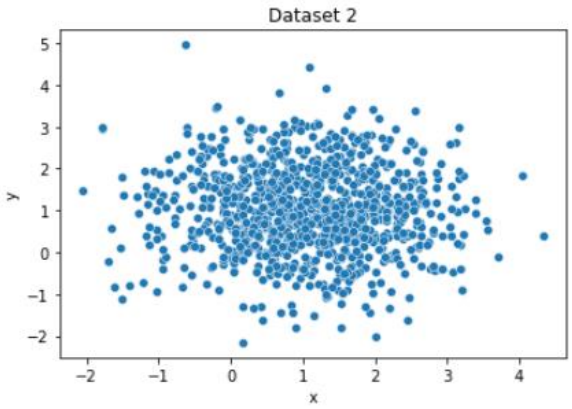


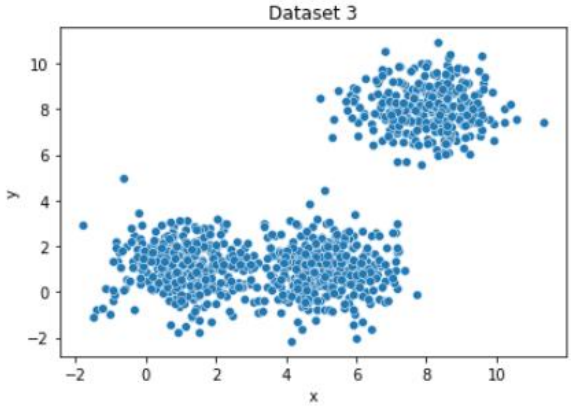
new centroid 1 position:
x1 = (3+3)/2 = 3
y1 = (7+8)/2 = 7.5

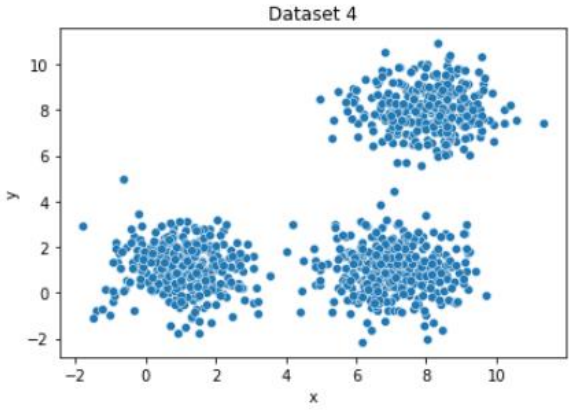
new centroid 2 position:
x2 = (1+2+1+2+1.5+4+4+3.5)/8 = 2.375
y2 = (1+2+2+1+1.5+7+8+7.5)/8 = 3.75

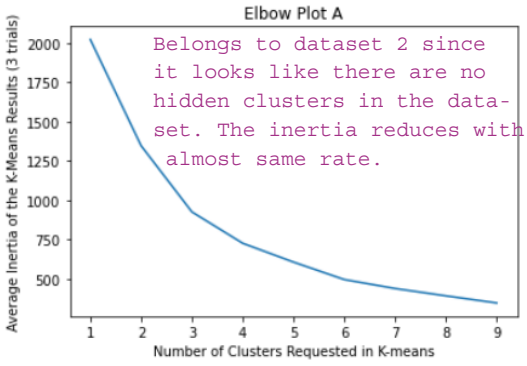
Question #2: [4 pt] Match the dataset to the k-means elbow plot that was created from this dataset. (Explanations not required, but may help with partial credit if you are wrong.)

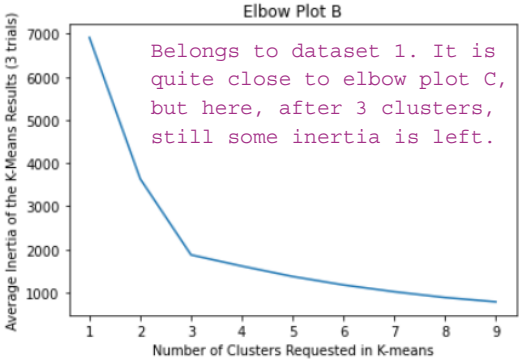


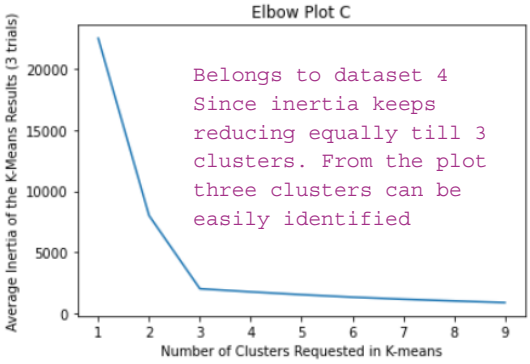


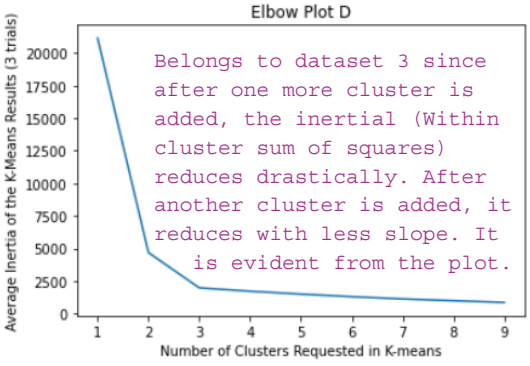




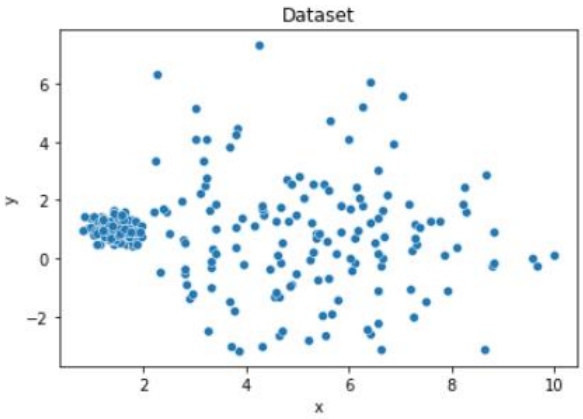




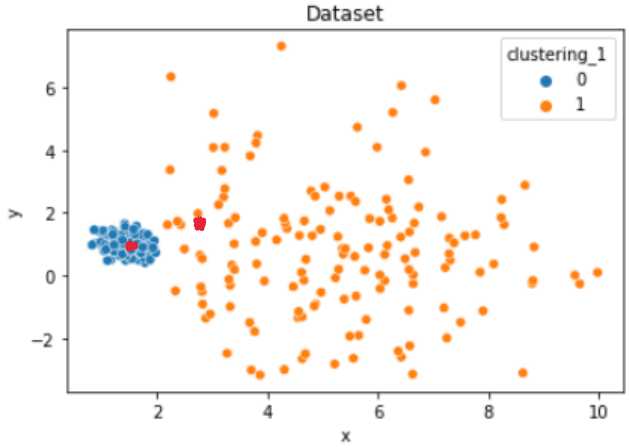


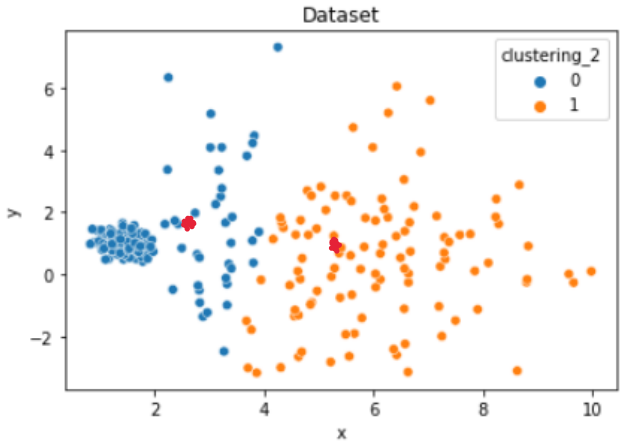


Question #3: [4 pt] The data displayed below shows two clusters. The dense cluster on the left contains about 150 objects and the sparse cluster on the right also contains about 150 objects.



Displayed below are two clusterings of the same dataset (ie. Clustering 1 and Clustering 2).





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

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

c. One of these clusterings has an inertia of 917 and the other clustering has an inertia of 1105. Which inertia do you think corresponds to which clustering? Explain why.

- The inertia 917 belongs to clustering_1 since its blue cluster is tightly packed and its centroid is very close to all the points. Hence, even though the orange cluster has high inertia, blue cluster has very low inertia.
- The inertia 1105 belongs to clutering_2 since blue cluster points are more sparse than clustering_1.

d. Do you think the k-means clustering algorithm will work well for this dataset? Why or why not?

```
k-means clustering algorithm will not work here due to the following reasons:
1. The clusters are not spherical (even though one is spherical, the other one is not).
2. The clusters do not have same sparsity.
```

Question #4:

1. Download the Assignment_01.zip file from Compass.
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 compass.