Homework Assignment 8 [30 points]

STAT437 Unsupervised Learning - Fall 2023

Due: Friday, October 20 on Canvas at 11:59pm CST.

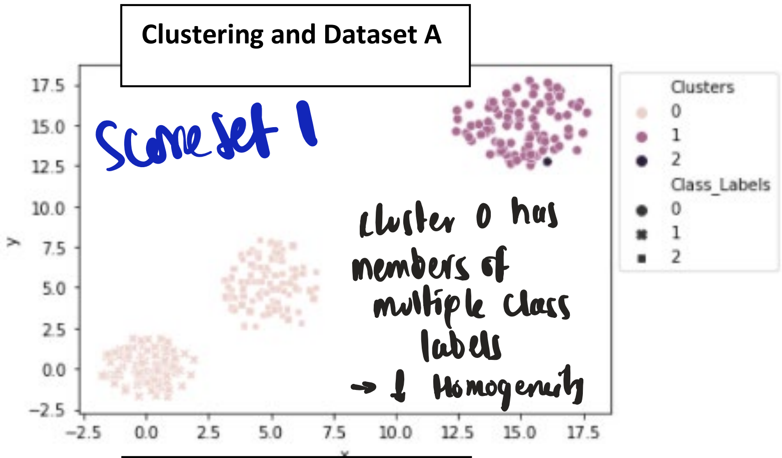
Simón Lizarazo
- Simon13 -

Problem	Points
1	1
2.1	1.25
2.2	3.25
3.1	1.25
3.2	1
4.1.1.	2.25
4.1.2	1.75
4.2	1.25
4.3.1	1.25
4.3.2	0.75
5.1.1	1.5
5.1.2	1.75
5.2.1	1.25
5.2.2	1
5.3.1	1.25
5.3.2	1.25
6	2
7	2.5
8	2.5

Questions #1-#5: Answer the questions in the jupyter notebook.

Question #6:

The three plots below display three sets of clustering labels (shown in the colors) and the class labels (shown by marker type) for the same dataset. For each of these sets of clustering labels and class labels we calculate the completeness score and the homogeneity score. Match the plots to the scores. No explanation required, but they may help if you are wrong.

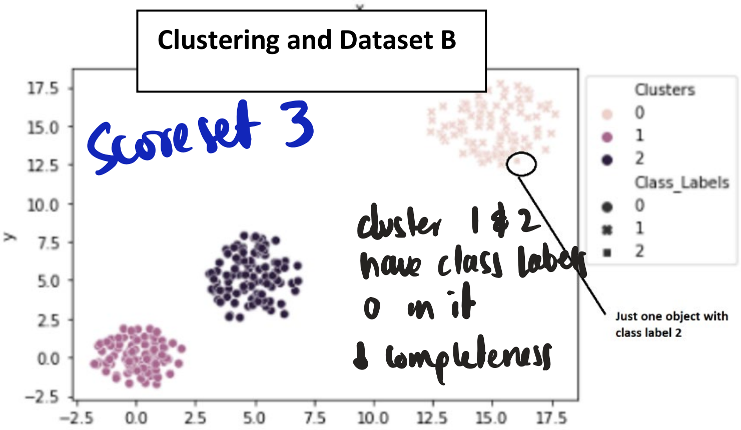


Clustering and Dataset A

Score set 1

cluster 0 has members of multiple class labels → ↓ Homogeneity

Score Set 1	Completeness Score = 0.97
	Homogeneity Score = 0.58
Score Set 2	Completeness Score=1
	Homogeneity Score = 1
Score Set 3	Completeness Score = 0.58
	Homogeneity Score = 0.97

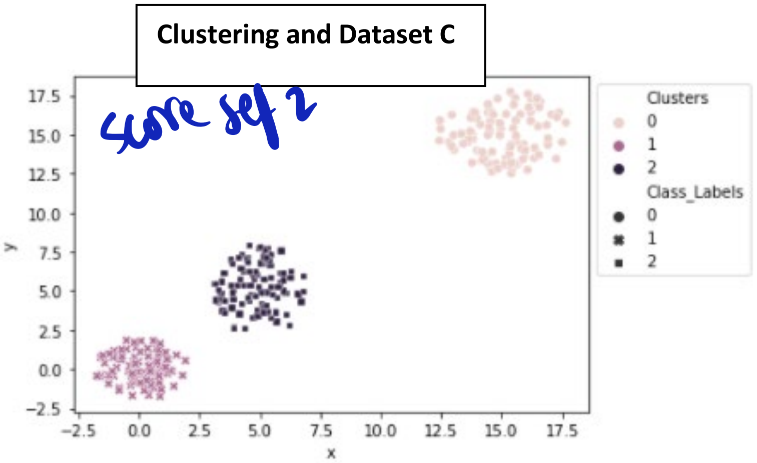


Clustering and Dataset B

Score set 3

cluster 1 & 2 have class label 0 in it ↓ completeness

Just one object with class label 2



Clustering and Dataset C

score set 2

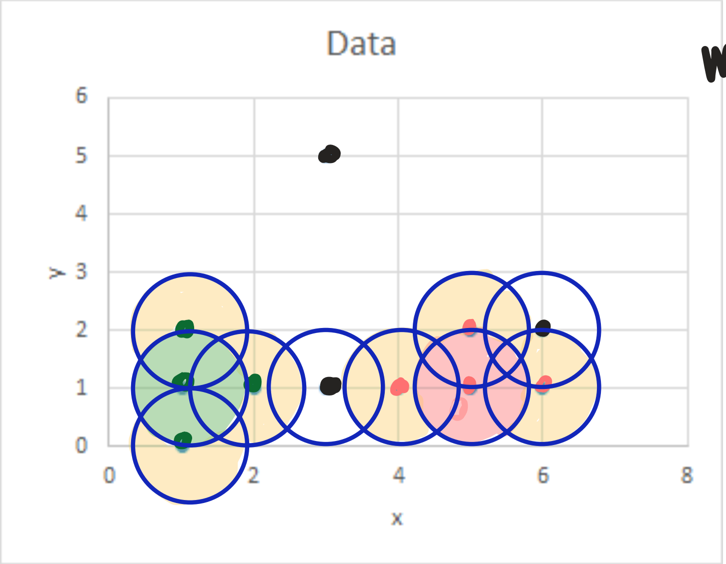
Question #7:

A dataset with 11 objects is shown in the table and the scatterplot below. Select the values of ϵ and $minpts$ in the DBSCAN algorithm that will yield the following desired cluster and noise point assignments shown in the table below. No explanation required, but they may help if you are wrong. *Hint: In the presence of border point ties, the border point can be assigned arbitrarily to either core point.*

	Data		
	x	y	**Desired Assignment**
Object 0	1	0	Cluster 1
Object 1	1	1	Cluster 1
Object 2	1	2	Cluster 1
Object 3	2	1	Cluster 1
Object 4	4	1	Cluster 2
Object 5	5	1	Cluster 2
Object 6	5	2	Cluster 2
Object 7	6	1	Cluster 2
Object 8	3	1	Noise
Object 9	3	5	Noise
Object 10	6	2	Noise

Epsilon = 1

minpoints = 3



🟢 This point is a core point

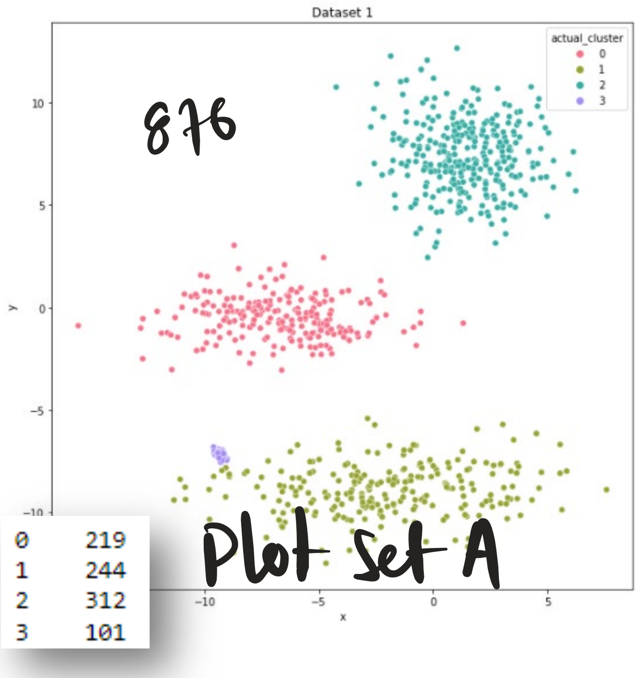
🔴 This point is a core point

🟡 This is a border point

Question #8:

For each of the four datasets shown below, we created some sorted k-nearest neighbors distance plots (k=2,3,4,5). Match each of the datasets (1-4) to it's corresponding plot sets (A-D). Explanations are not required, but may help for partial credit.

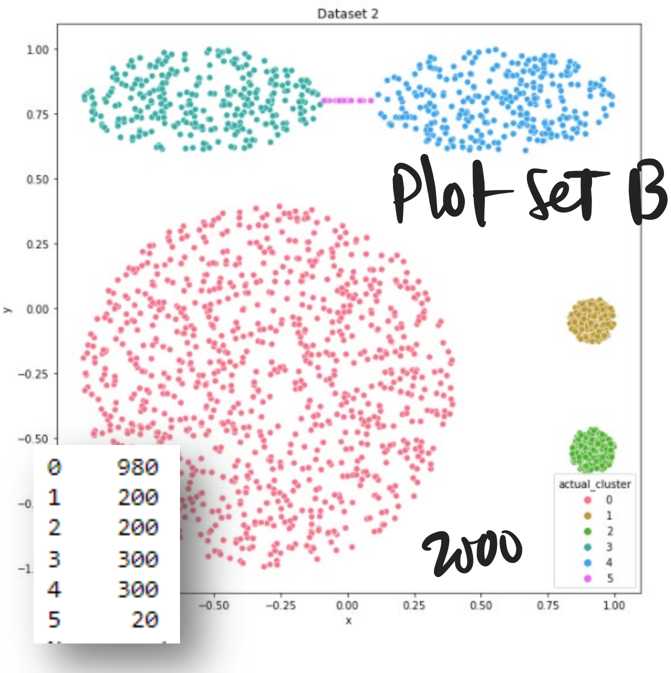
<u>Hint</u>: *For each of the datasets, we have colored coded each of the objects by a certain class (ie. 'actual_cluster'). To help you with this problem, we have also provided the number of objects that correspond to each class.*



```
0    219
1    244
2    312
3    101
```

876

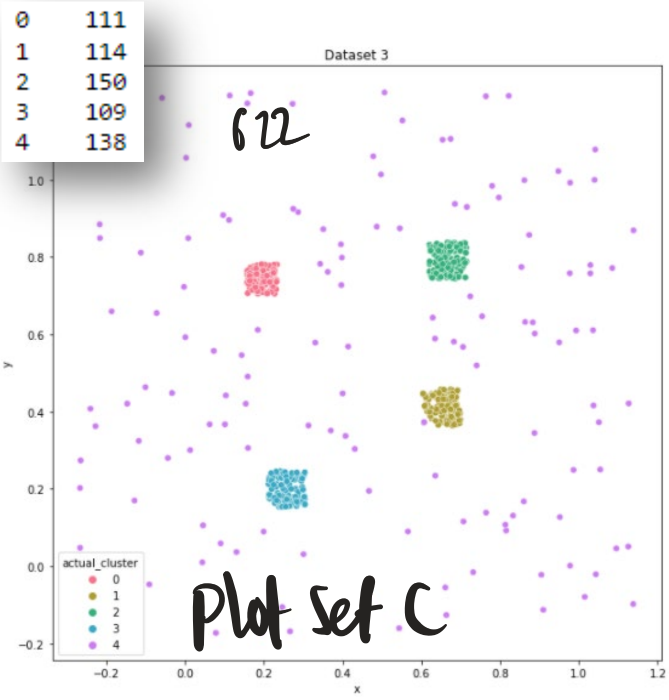
Plot set A



```
0    980
1    200
2    200
3    300
4    300
5     20
```

Plot set B

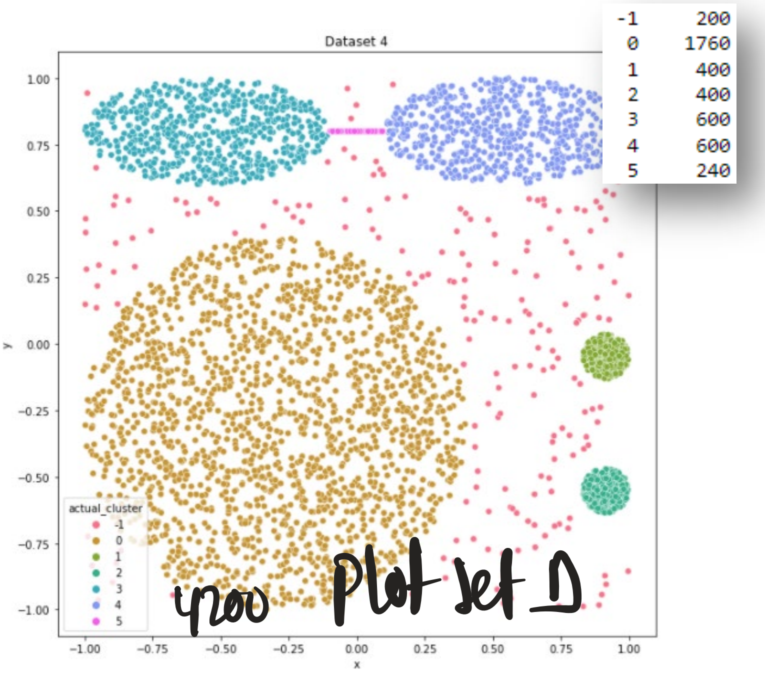
2000



```
0    111
1    114
2    150
3    109
4    138
```

622

Plot set C

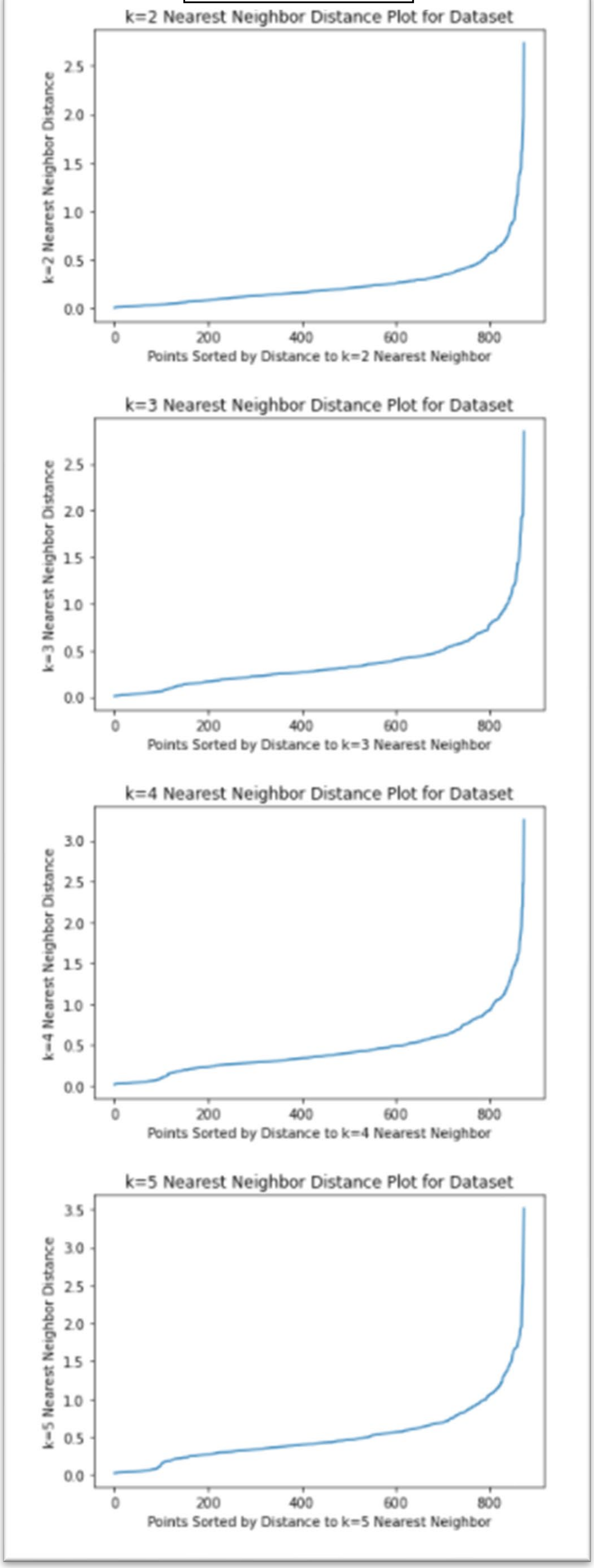


```
-1    200
 0   1760
 1    400
 2    400
 3    600
 4    600
 5    240
```

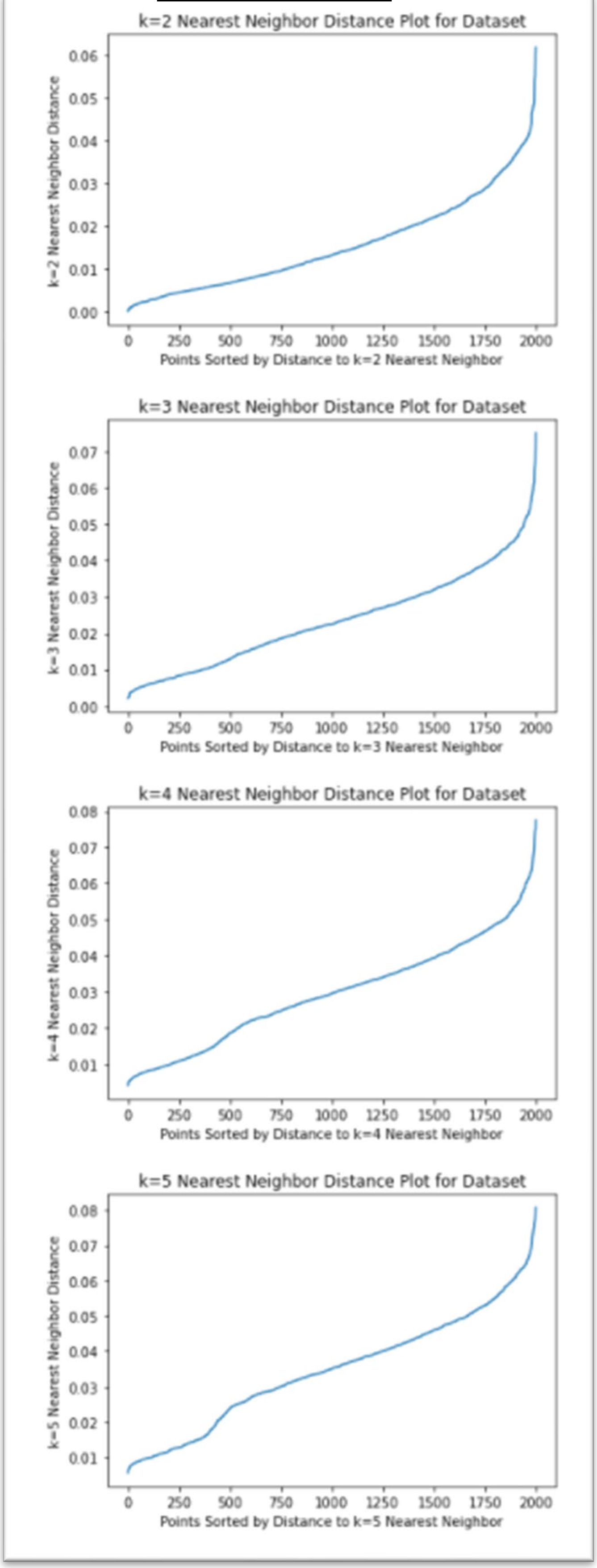
4200

Plot set D

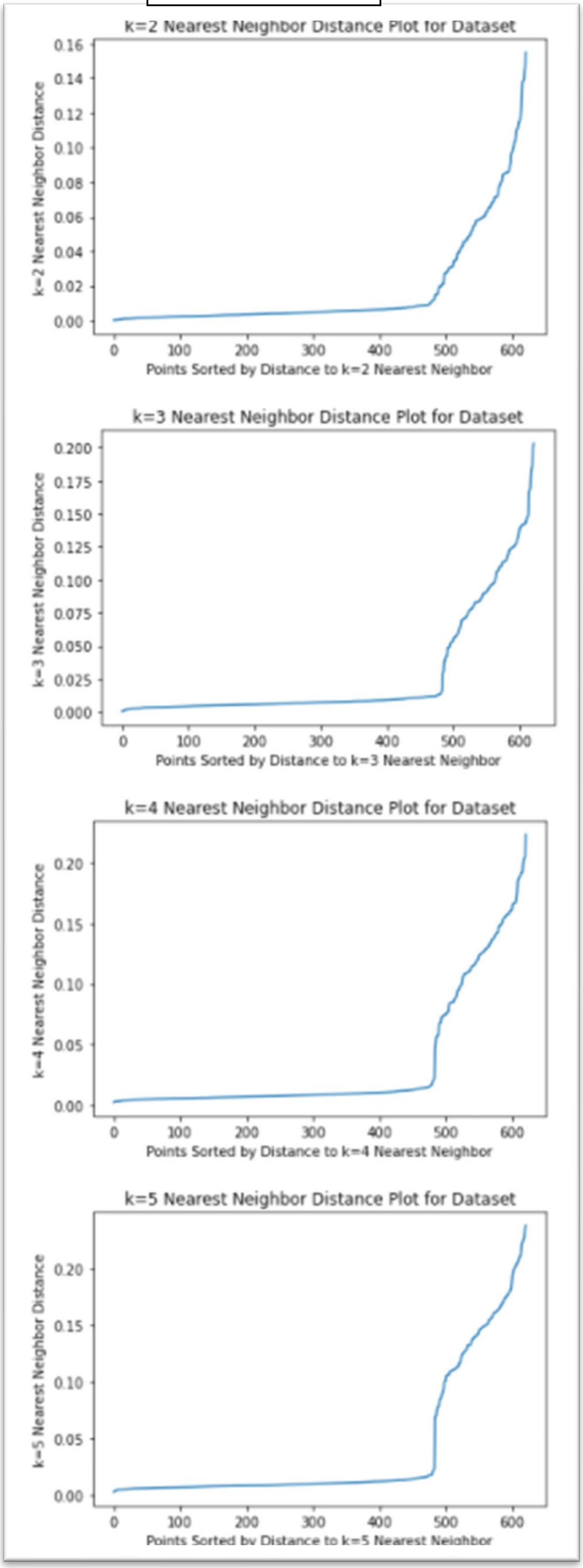
Plot Set A



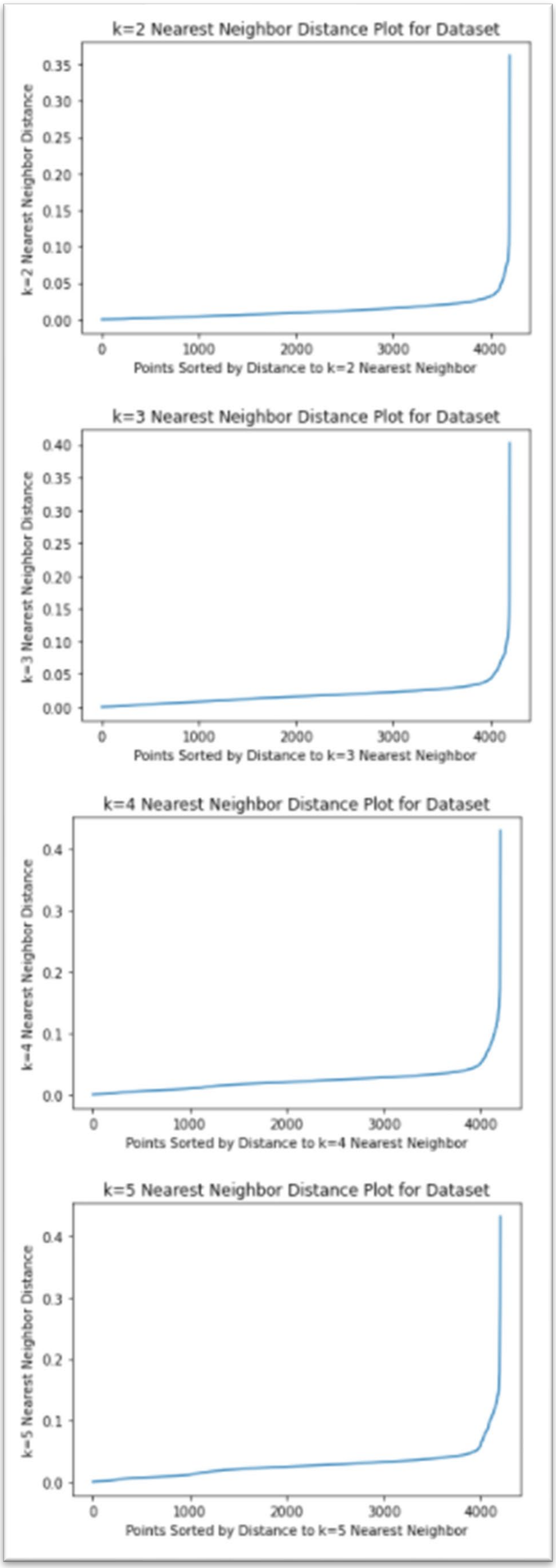
Plot Set B



Plot Set C



Plot Set D


k=2 Nearest Neighbor Distance Plot for Dataset
k=2 Nearest Neighbor Distance
Points Sorted by Distance to k=2 Nearest Neighbor
k=3 Nearest Neighbor Distance Plot for Dataset
k=3 Nearest Neighbor Distance
Points Sorted by Distance to k=3 Nearest Neighbor
k=4 Nearest Neighbor Distance Plot for Dataset
k=4 Nearest Neighbor Distance
Points Sorted by Distance to k=4 Nearest Neighbor
k=5 Nearest Neighbor Distance Plot for Dataset
k=5 Nearest Neighbor Distance
Points Sorted by Distance to k=5 Nearest Neighbor