Homework Assignment 3 – [30 points]

STAT437 Unsupervised Learning – Spring 2025

Due: Friday, February 14 💔 *on Canvas*

Reference the attached Jupyter notebook for the case study 1 and 2 questions.

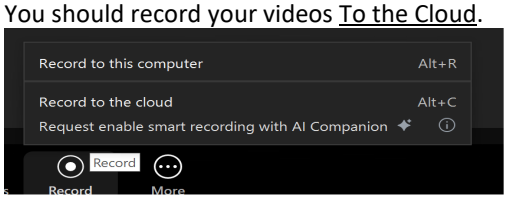
Video Question: Make a short video explaining your answers to question #1 and question #2. What are some potential things that can happen in a cluster analysis that has variables that are higher in scale than others?

IMPORTANT Video Element of ALL Homework Assignments:

- In order to receive points for each video submission, you need to do **ALL** of the following.
  - Have your camera on.
  - Show your FULL screen in Zoom (not just a particular application).
  - We should be able to hear the audio. Make sure to turn your mic on.
  - You should give a good faith attempt to answer the prompt.
  - Your video meet the minimum time requirement.
  - It should not sound like you are just reading off a script.
  - It's ok if your video recording is not the most eloquent. What's important is that you are putting together YOUR authentic thoughts on your particular understanding of the assignment and the lecture content.

How to Submit Videos:

- You should record your videos in your UIUC Zoom client.
- You should record your videos To the Cloud.



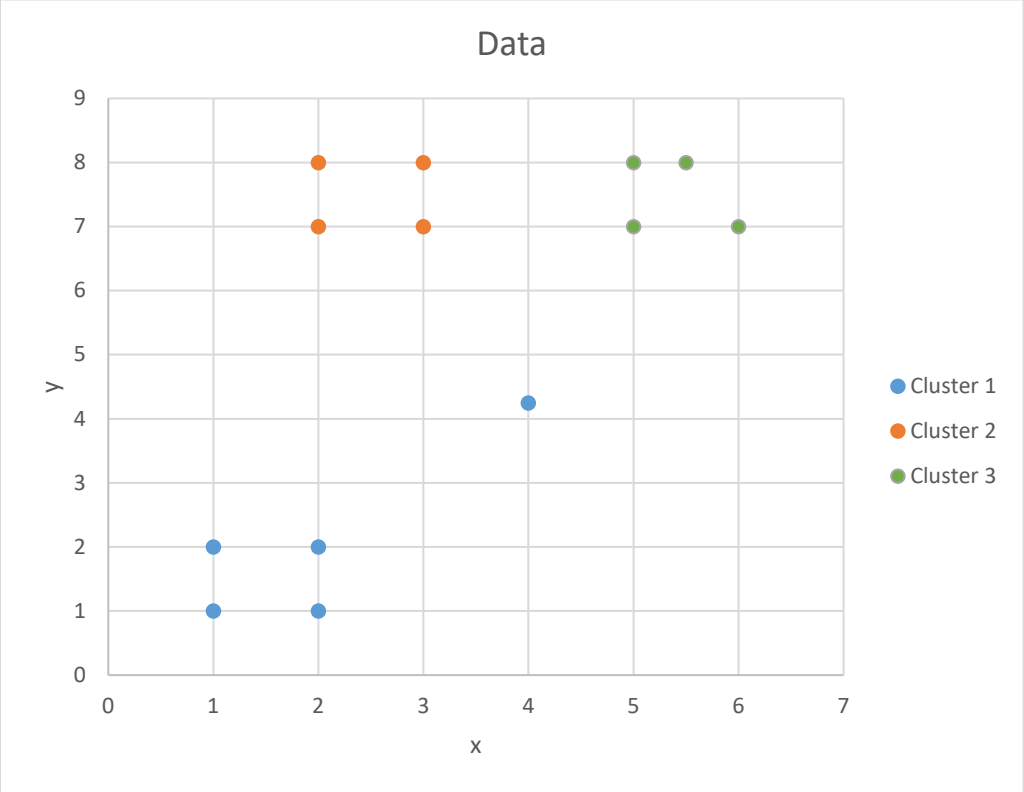
- You can find your recording link at https://illinois.zoom.us/recording/.
- Click on the corresponding video and Copy shareable link to paste the link in Canvas.

Pdf Questions	Points
1	2.5
2	2.5
3	1
Video Question	1
Case Study 1 Questions	**Points**
1.1	0.25
1.2	0.25
1.3	0.5
1.4	1
1.5	0.75
1.6	1
2.1	0.75
2.2	0.25
2.3	0.25
2.4	0.5
2.5	1
2.6	0.5
Case Study 2 Questions	**Points**
1.1	0.25
1.2	0.25
1.3	0.5
1.4	0.75
2.1	0.5
2.2	0.25
2.3	0.5
3	0.5
4.1	0.75
4.2	0.5
5	0.75
6.1	0.75
6.2	0.75
6.3	0.75
6.4	0.75
6.5	0.25
7.1	1.5
7.2	0.75
8	1
9.1	0.5
9.2	0.5
9.3	0.5
9.4	0.75
9.5	0.25
10.1	0.75
10.2	0.75

Question #1:

Calculate the silhouette score of object 5 using the information below. Then interpret what this silhouette score says about object 5 with respect to this clustering.

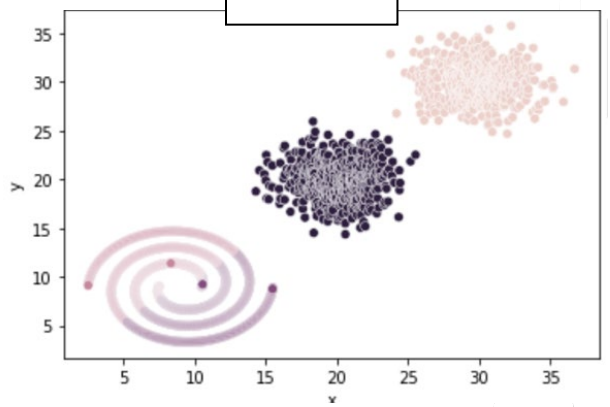
	Data			
		x	y	**Distance Object 5 (4, 4.25) is away from this object.**
Cluster 1	**Object 1**	1	1	4.42
	Object 2	2	2	3.01
	Object 3	1	2	3.75
	Object 4	2	1	3.82
	Object 5	4	4.25	--
Cluster 2	**Object 6**	2	7	3.40
	Object 7	2	8	4.25
	Object 8	3	7	2.93
	Object 9	3	8	3.88
Cluster 3	**Object 10**	5	7	2.93
	Object 11	5	8	3.88
	Object 12	6	7	3.40
	Object 13	5.5	8	4.04



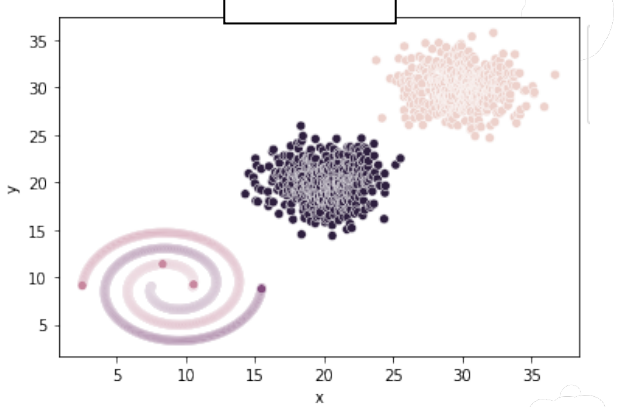
Question #2:

2.1. Matching: Match one of the 6 datasets and clusterings (1-6) to the one of the five silhouette plots (A-F) that were created from one of these datasets and clusterings. *Hint: You'll want to approximate what the silhouette plot for each of these datasets would look like. Pick a few representative points from each dataset to help you.*

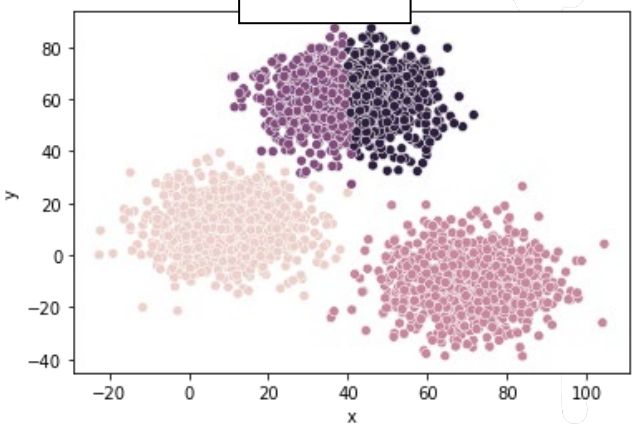
Dataset 1



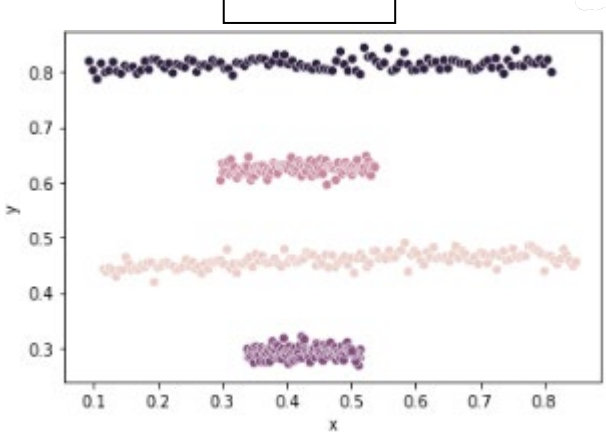
Dataset 2



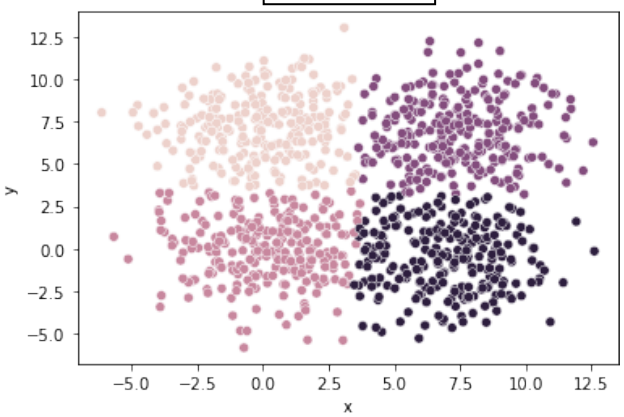
Dataset 3



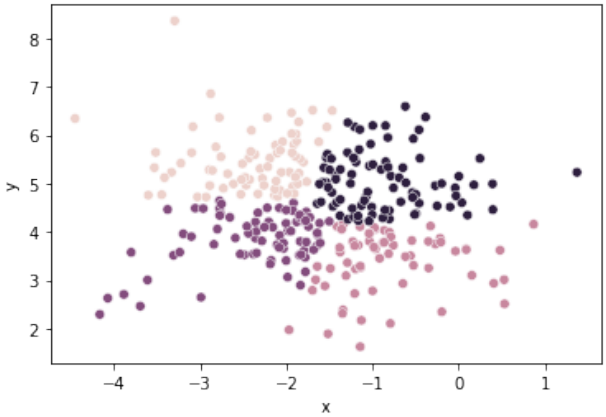
Dataset 4

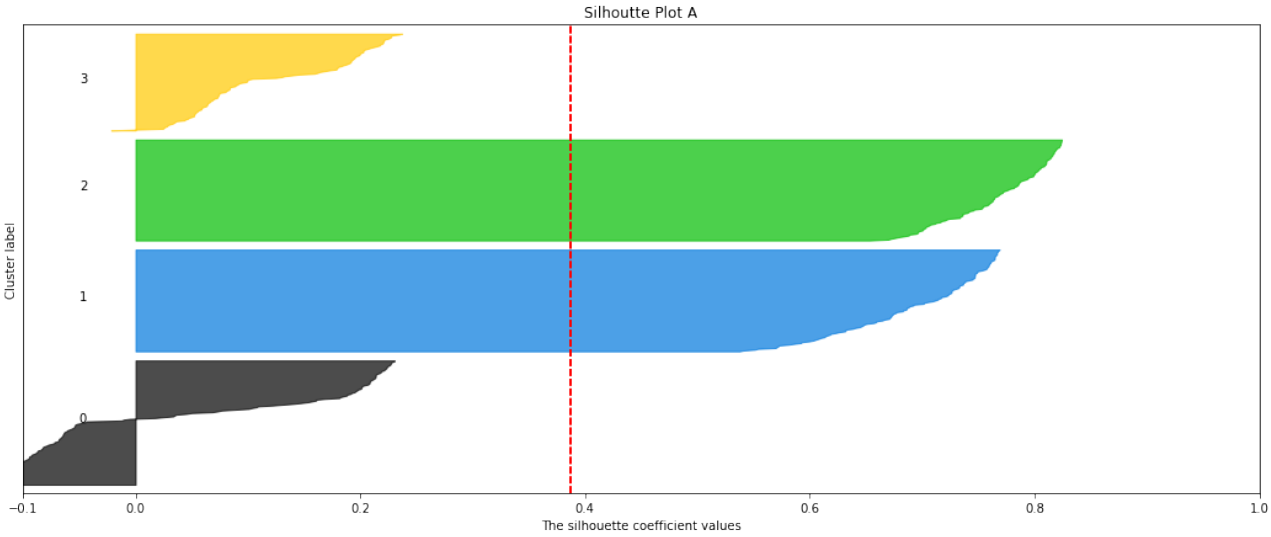


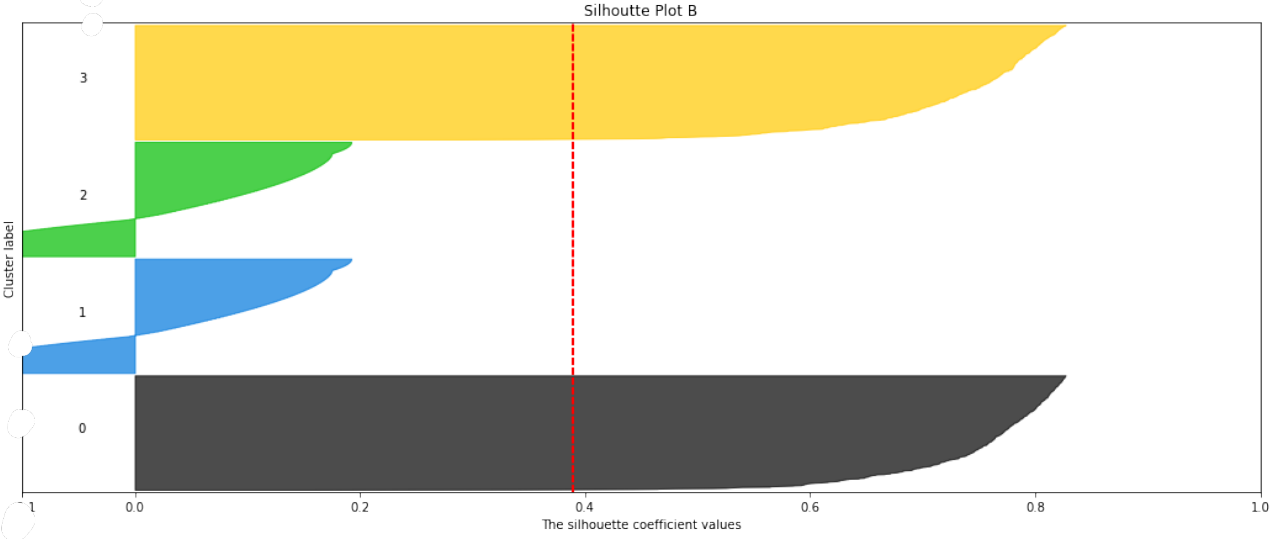
Dataset 5

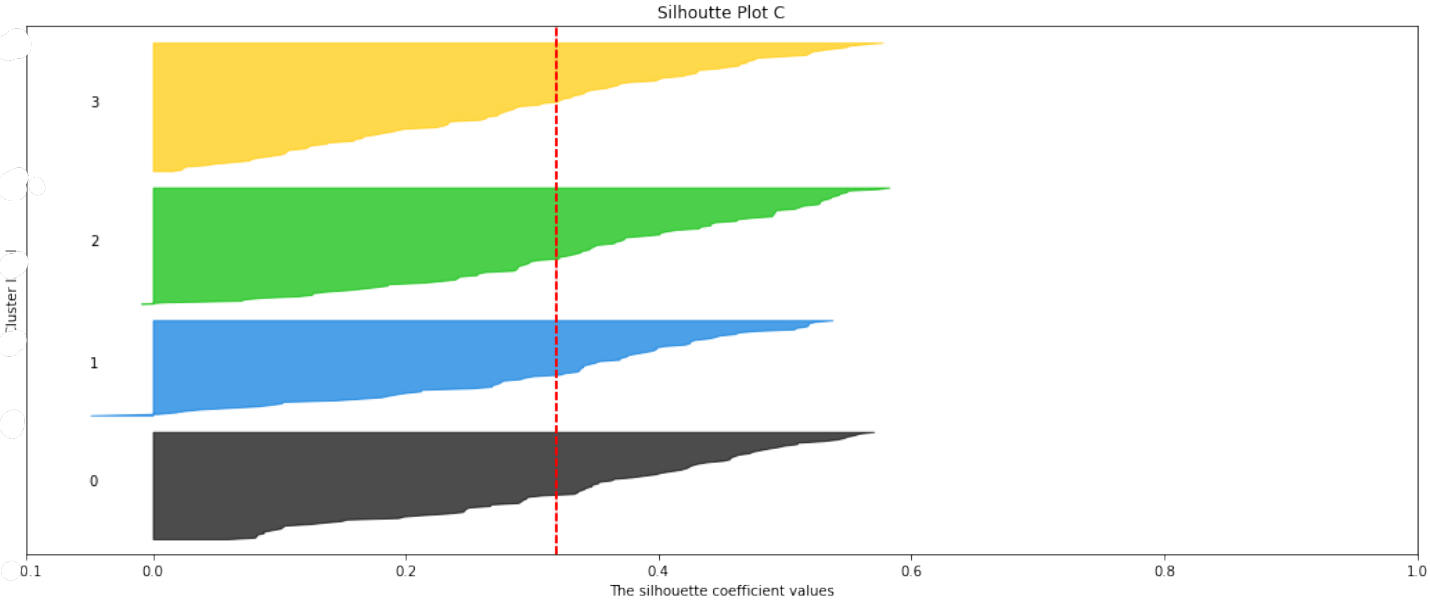


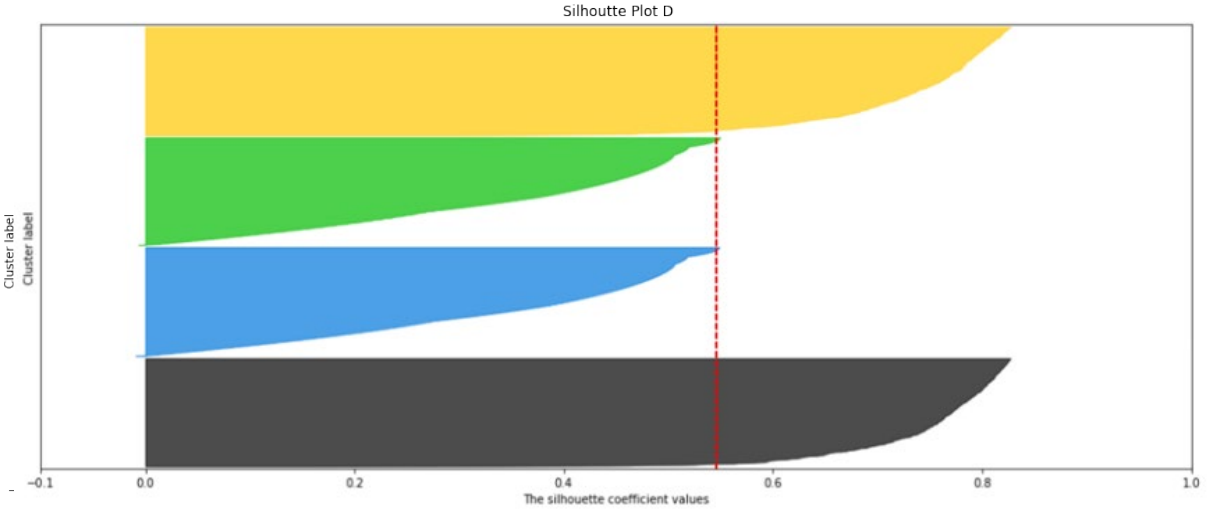
Dataset 6

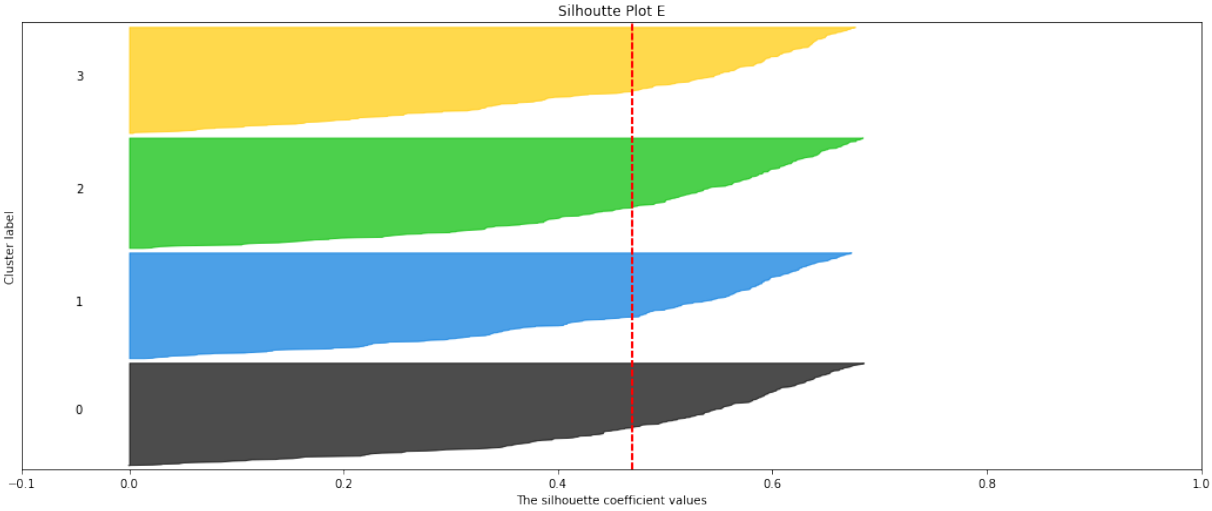


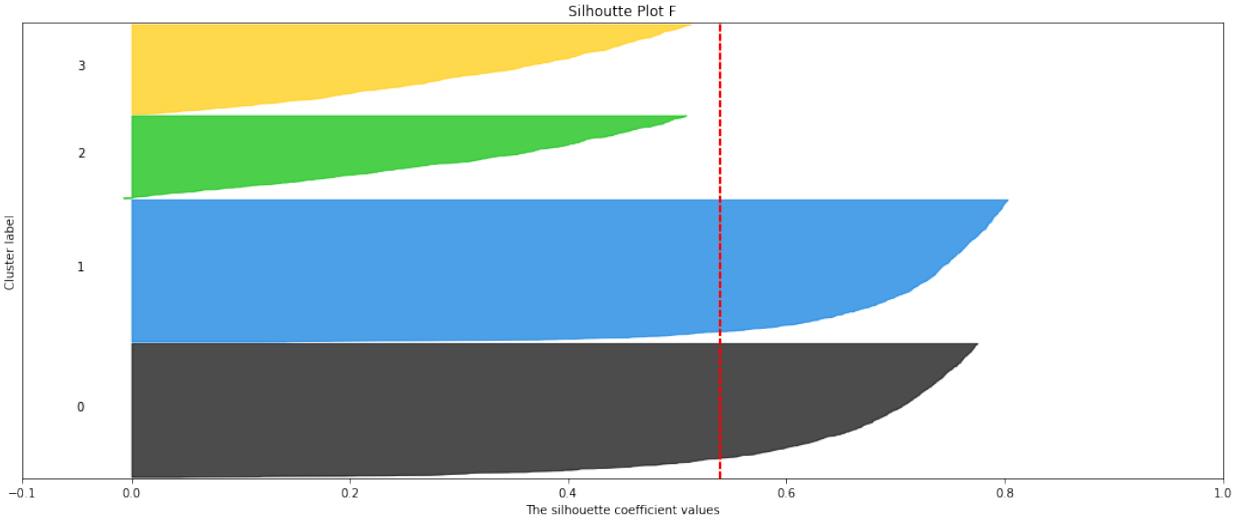

Silhoutte Plot A
Cluster label
3
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values


Silhoutte Plot B
Cluster label
3
2
1
0
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values


Silhoutte Plot C
3
2
1
0
0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values


Silhoutte Plot D
Cluster label
Cluster label
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values


Silhoutte Plot E
3
2
1
0
Cluster label
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values


Silhoutte Plot F
3
2
1
0
Cluster label
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values

2.2. Silhouette Plot Suitability for Assesing a Clustering Result: Select **two datasets/clustering pairs** from (1,2,..,6) above in which:

- **a.** the clustering successfully identifies the inherent clusters in the dataset BUT
- **b.** the corresponding silhouette plot suggest that at least one of the clusters is not cohesive and well-separated.

Question #3:

Suppose that we cluster the following dataset below using k-means into k=2, 3, and 4 clusters respectively. Which one of these clusterings will have the highest average silhouette score? (We're using Euclidean distance). Explain.

Do you think that the average silhouette score alone will be useful in helping us select the clustering that identified the 4 "inherent" clusters in this dataset?