COM S 474/574 Spring 2022
Introduction to Machine Learning

Homework 9

1 Directions:

- **Due: Thursday April 21, 2022 at 9pm.** Late submissions will be accepted for 24 hours with a 15% penalty. (the enforcement is strict, beginning at 9:01pm, except for extreme situations; having a poor wifi connection or minor computer problems is not sufficient for the penalty to be waived.)
- Upload the homework to Canvas as a single pdf file.
- If the graders cannot easily read your submission (writing is illegible, image is too dark, or if the contrast is too low) then you might receive a zero or only partial credit.
- Any non-administrative questions must be asked in office hours or (if a brief response is sufficient) Piazza.

Hamamoto Yuichi

2 Problems

Problem 1. [26 points]

Suppose that we have a data set with four samples. You calculate the pair-wise distances as

$$\begin{array}{c} \\ \{1\} \\ \{2\} \\ \{3\} \\ \{4\} \end{array} \begin{array}{c} \begin{array}{cccc} \{1\} & \{2\} & \{3\} & \{4\} \end{array} \\ \begin{bmatrix} & 0.3 & 0.4 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.4 & 0.5 & & 0.45 \\ 0.7 & 0.8 & 0.45 & \end{bmatrix} \end{array}$$

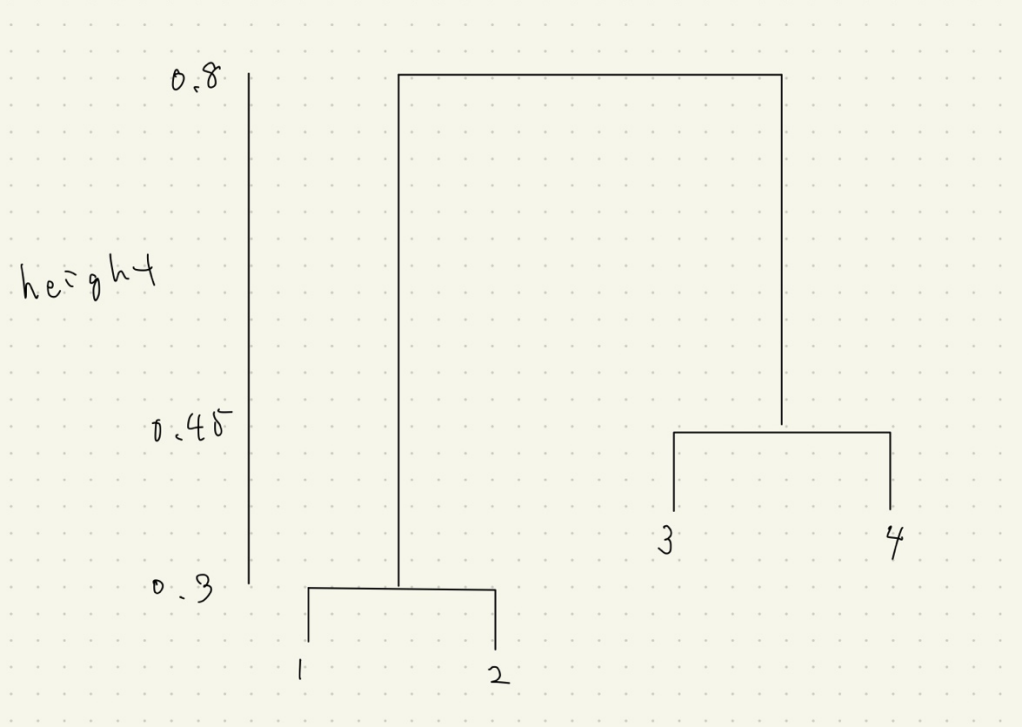
For the following questions, show your work. Recall that "single linkage" means we define the distance of two clusters C_1 and C_2 as the minimum distance of any pair of elements

$$dist(C_1, C_2) := \min_{a \in C_1, b \in C_2} dist(a, b)$$

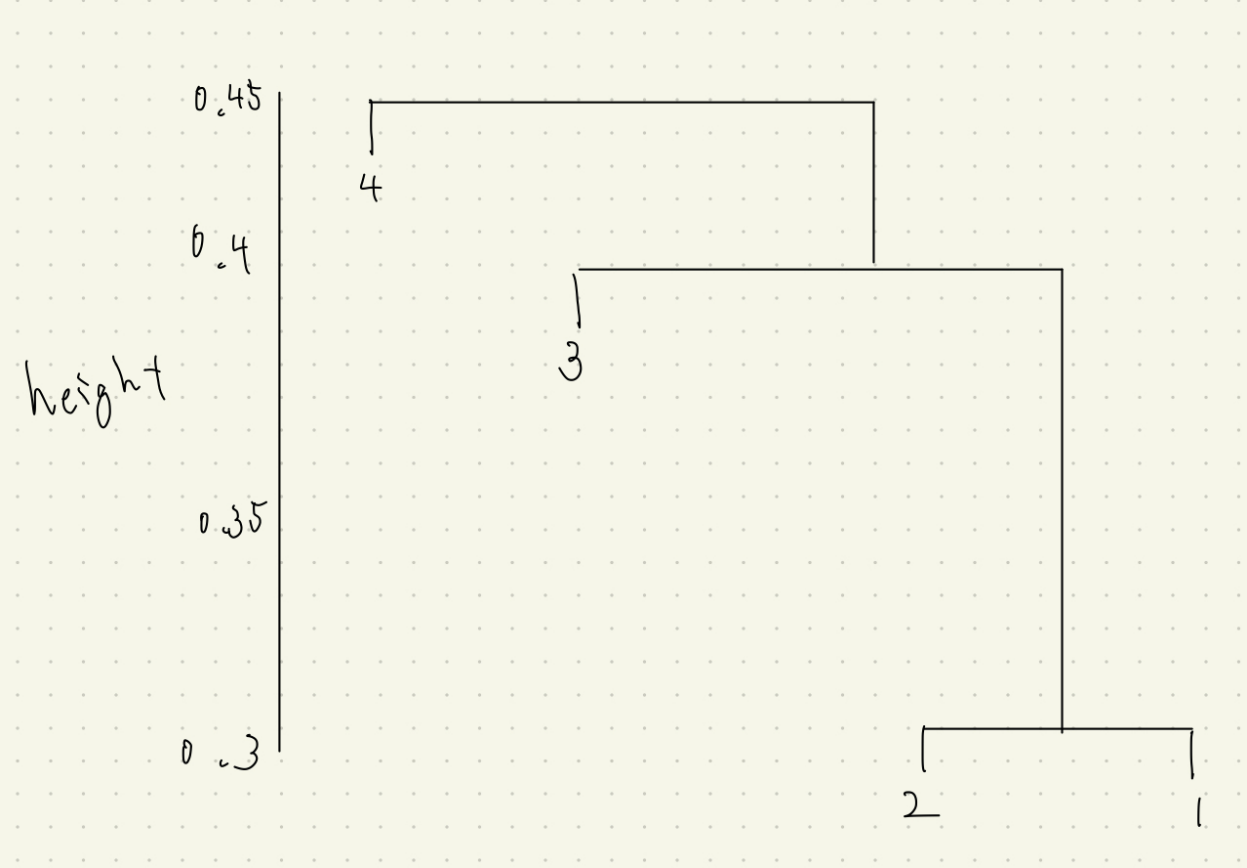
and that "complete linkage" means we define the distance of two clusters C_1 and C_2 as the maximum distance of any pair of elements

$$dist(C_1, C_2) := \max_{a \in C_1, b \in C_2} dist(a, b).$$

a. Sketch the dendrogram that results from hierarchically clustering these four samples using complete linkage. Indicate on the plot the height (distance) at which each fusion occurs, as well as the samples corresponding to each leaf in the dendrogram.



b. Repeat part (a) for single linkage clustering.



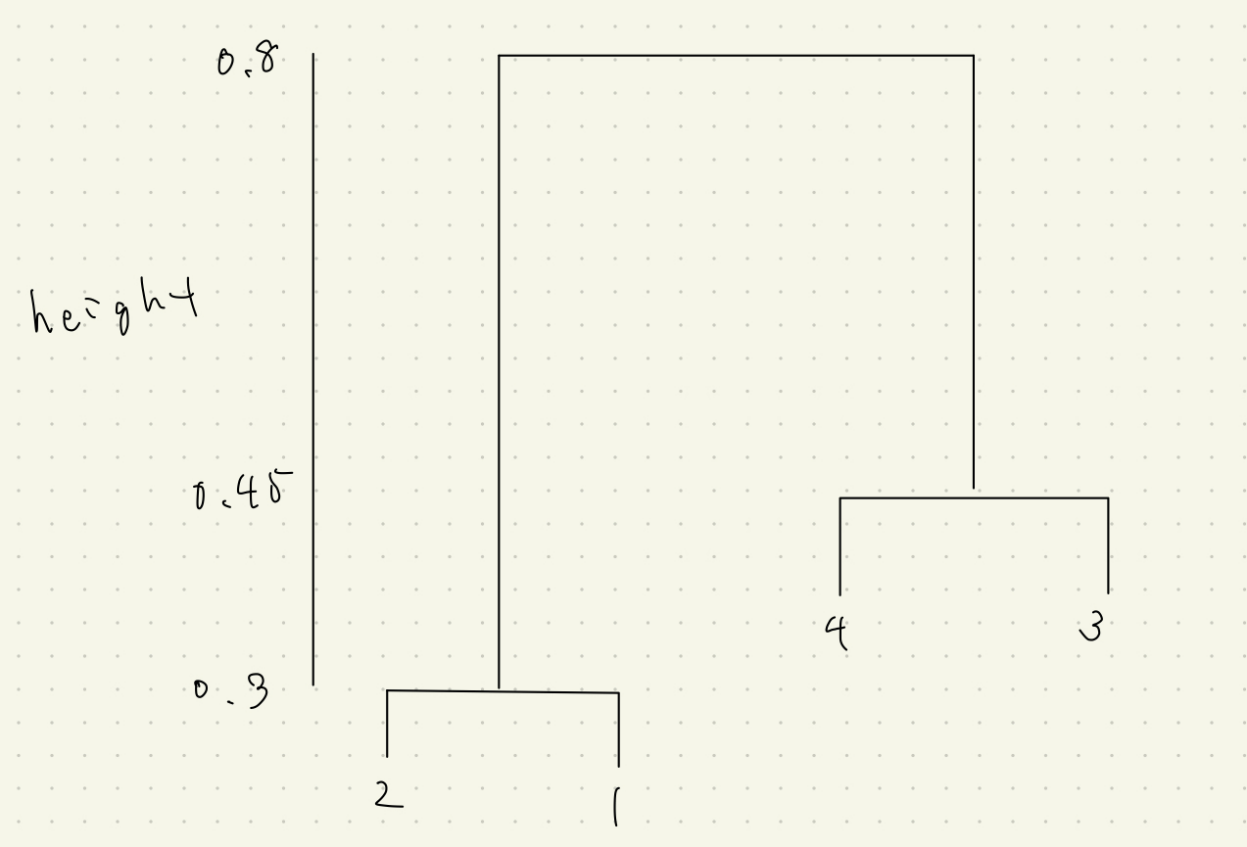
c. Suppose that we cut the dendrogram obtained in (a) such that two clusters result. Which samples are in each cluster?

1,2 and 3,4

d. Suppose that we cut the dendrogram obtained in (b) such that two clusters result. Which samples are in each cluster?

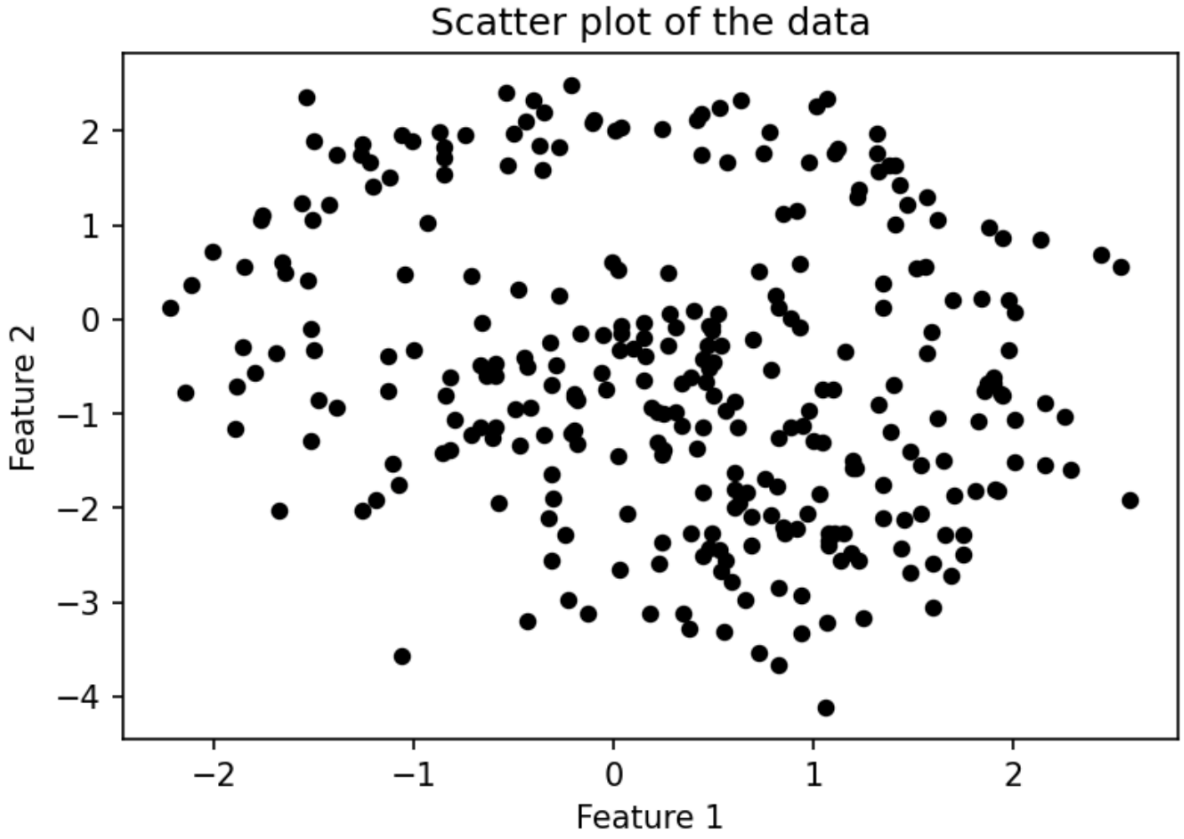
4 and (3, (1,2)

e. Draw a dendrogram that is equivalent to the dendrogram in (a) with a different horizontal arrangement for the samples.

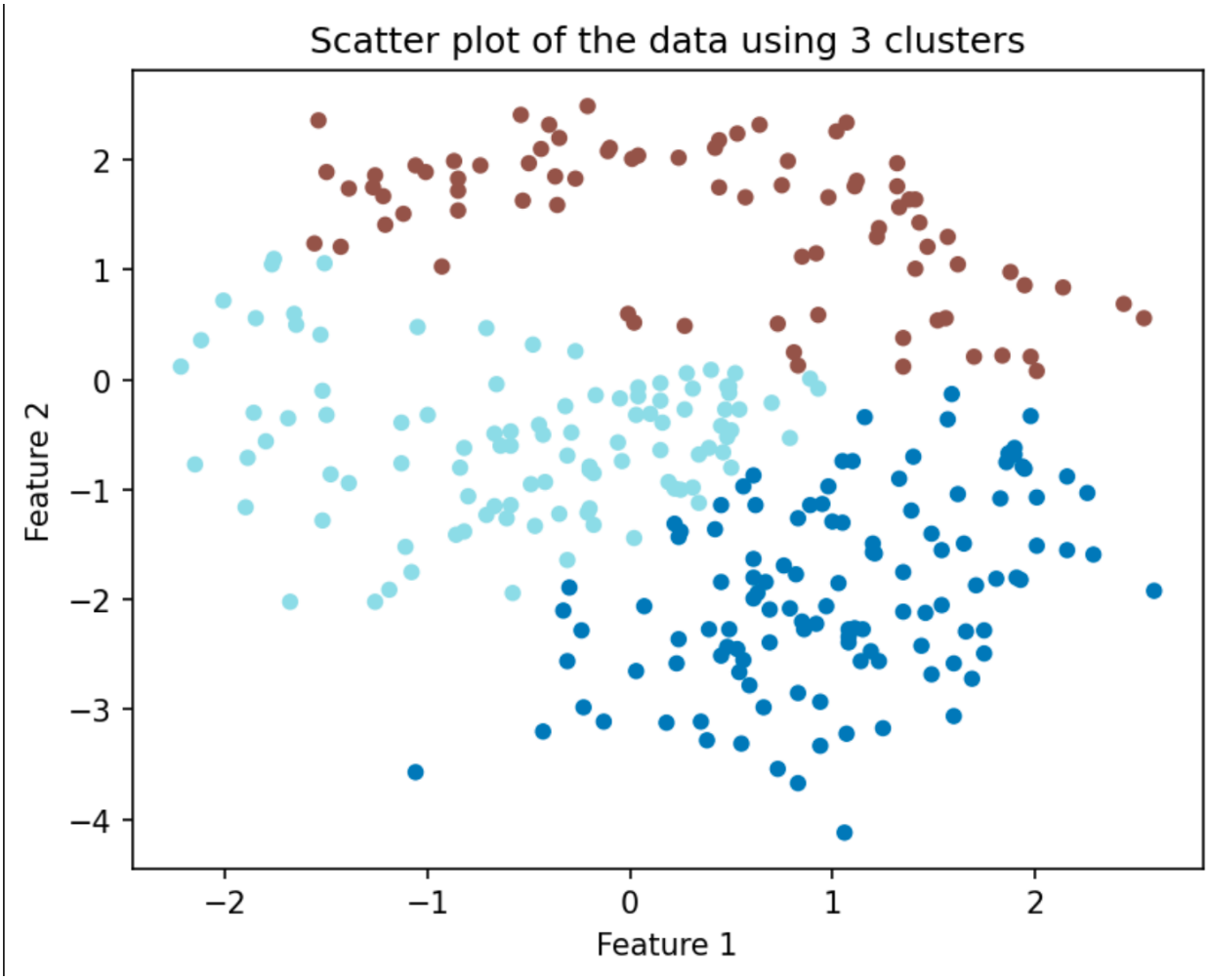


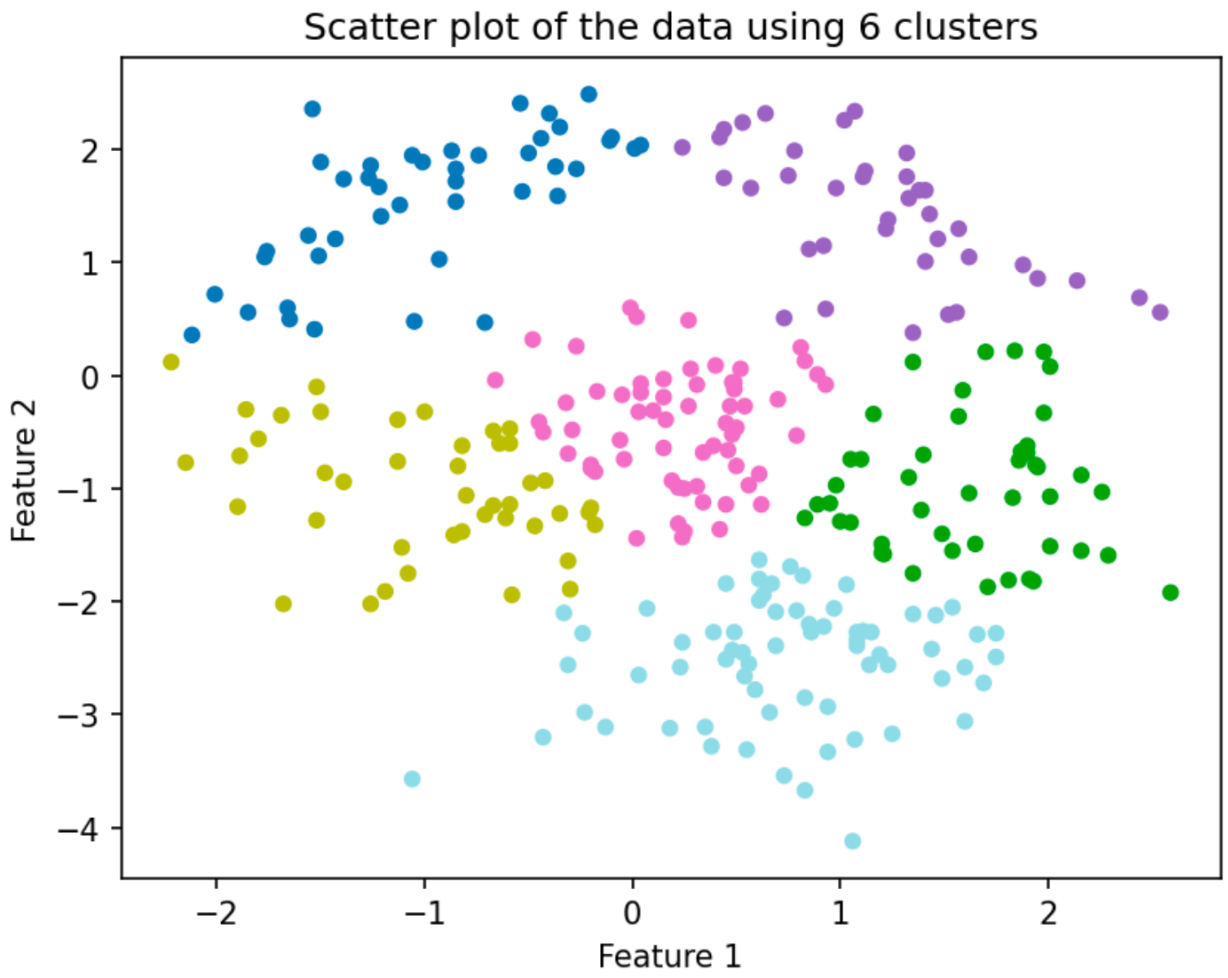
Problem 2. [20 points] In this problem you will run several clustering procedures and compare them on an example data set, HW9data.csv. Example Python code is posted as HW9-cluster-template.ipynb; you may need to make a few, minor changes.

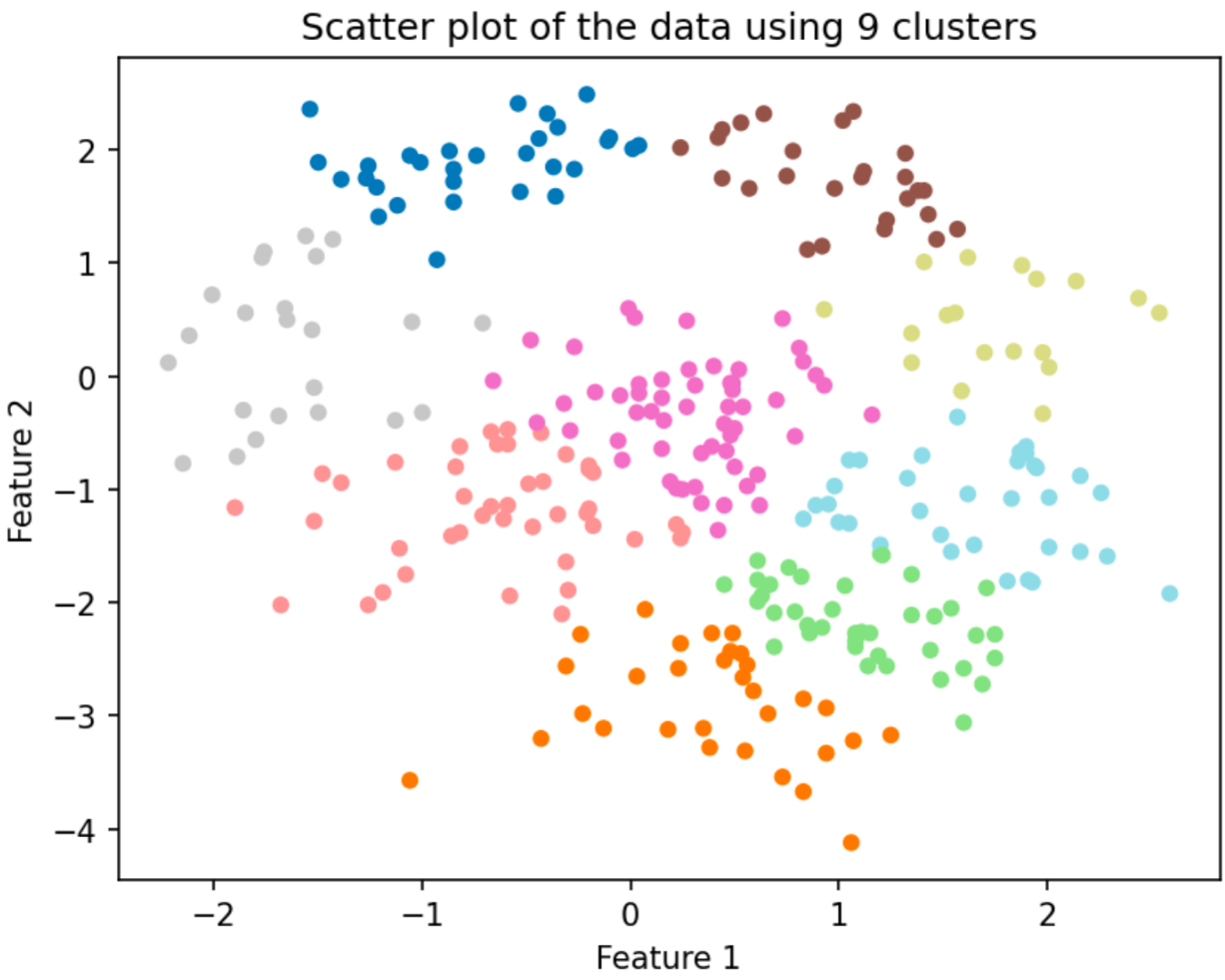
A. Make a scatter-plot of the data, coloring each data point black.



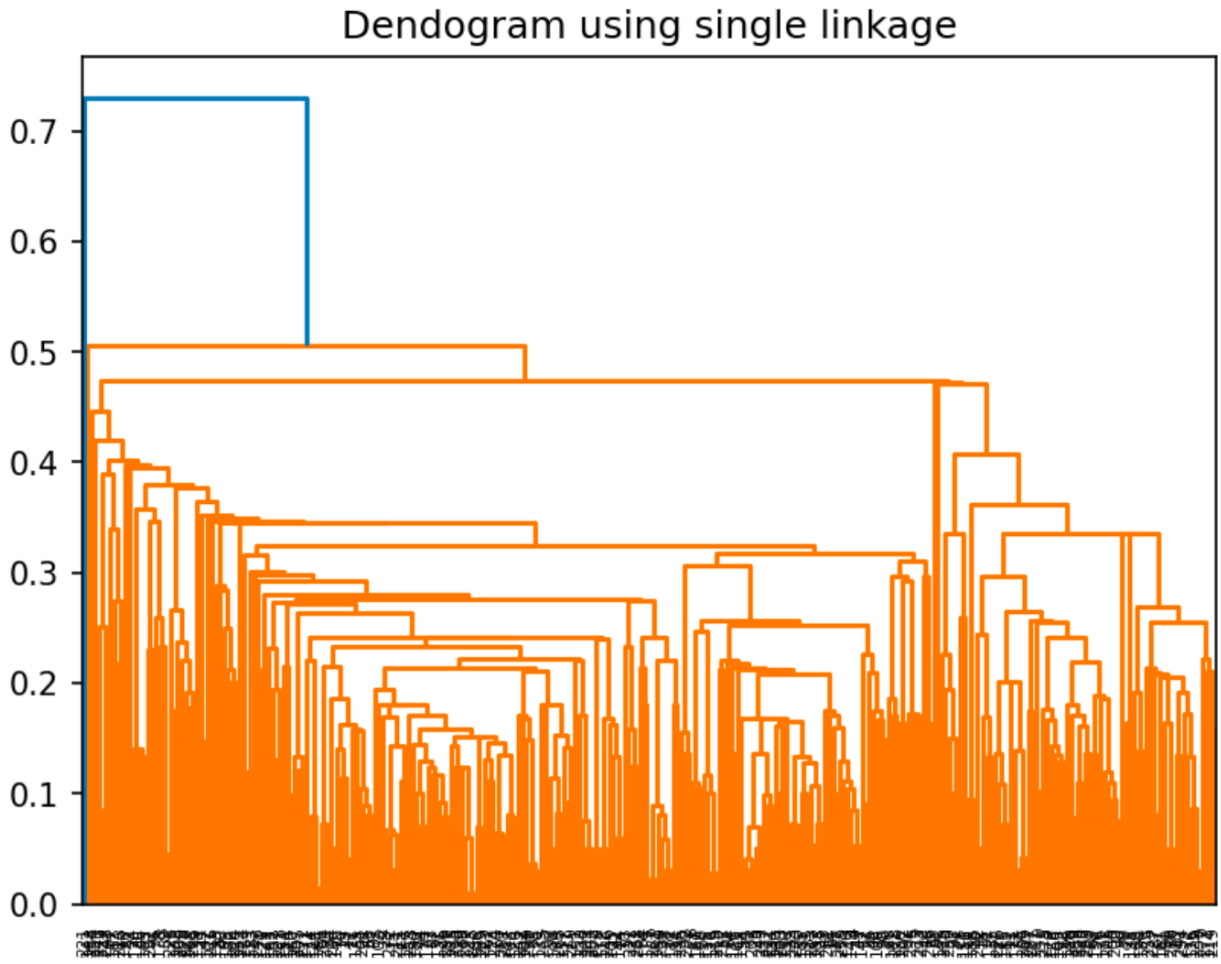
B. For n_clusters=range(1,16), apply K-means clustering. Make a scatter plot for each. You only need to include 3 of them in your homework submission. Select the three pictures whose clusters you think look the best.



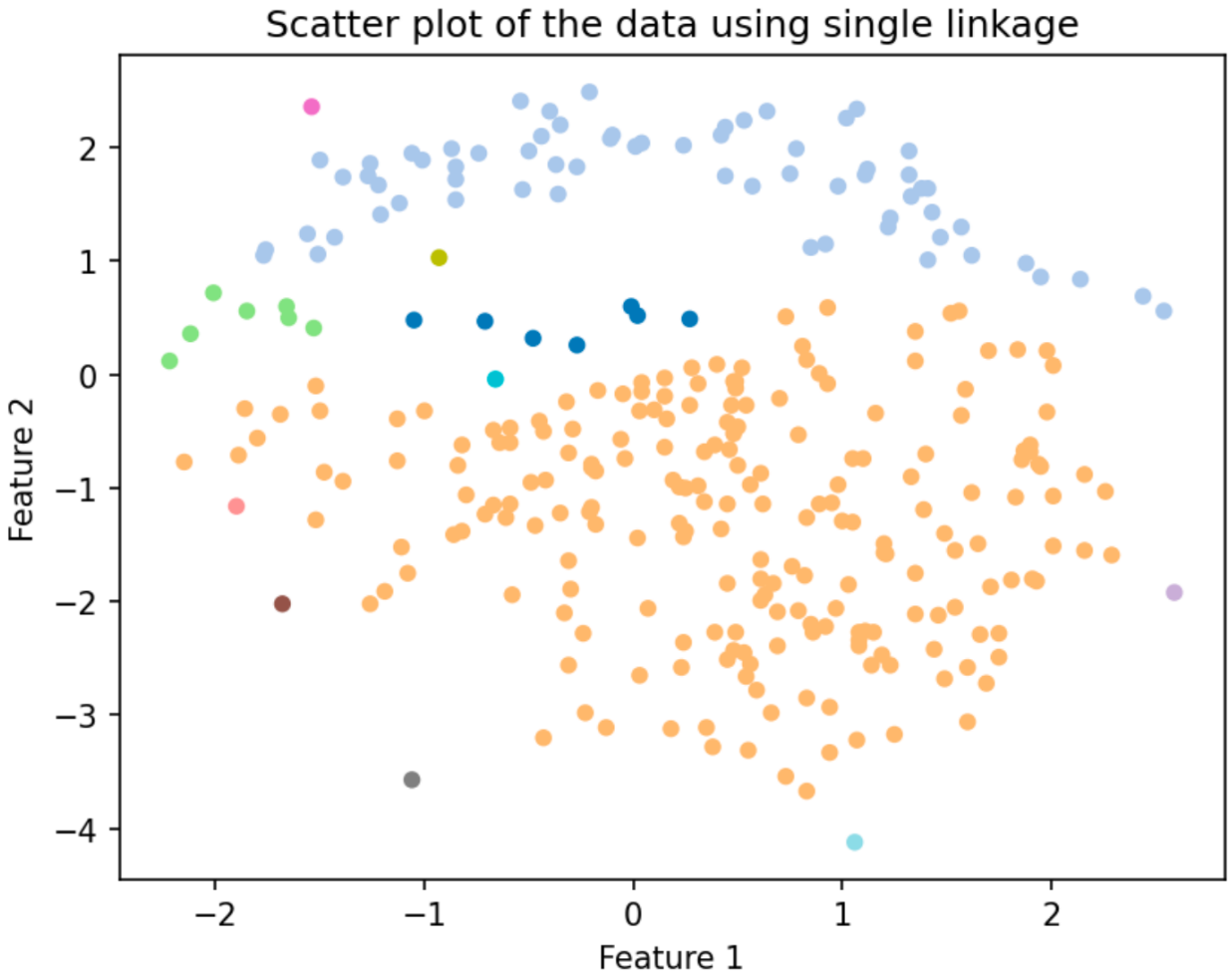

Scatter plot of the data using 6 clusters
Feature 2
Feature 1
2
1
0
−1
−2
−3
−4
−2
−1
0
1
2


Scatter plot of the data using 9 clusters
Feature 2
Feature 1
2
1
0
−1
−2
−3
−4
−2
−1
0
1
2

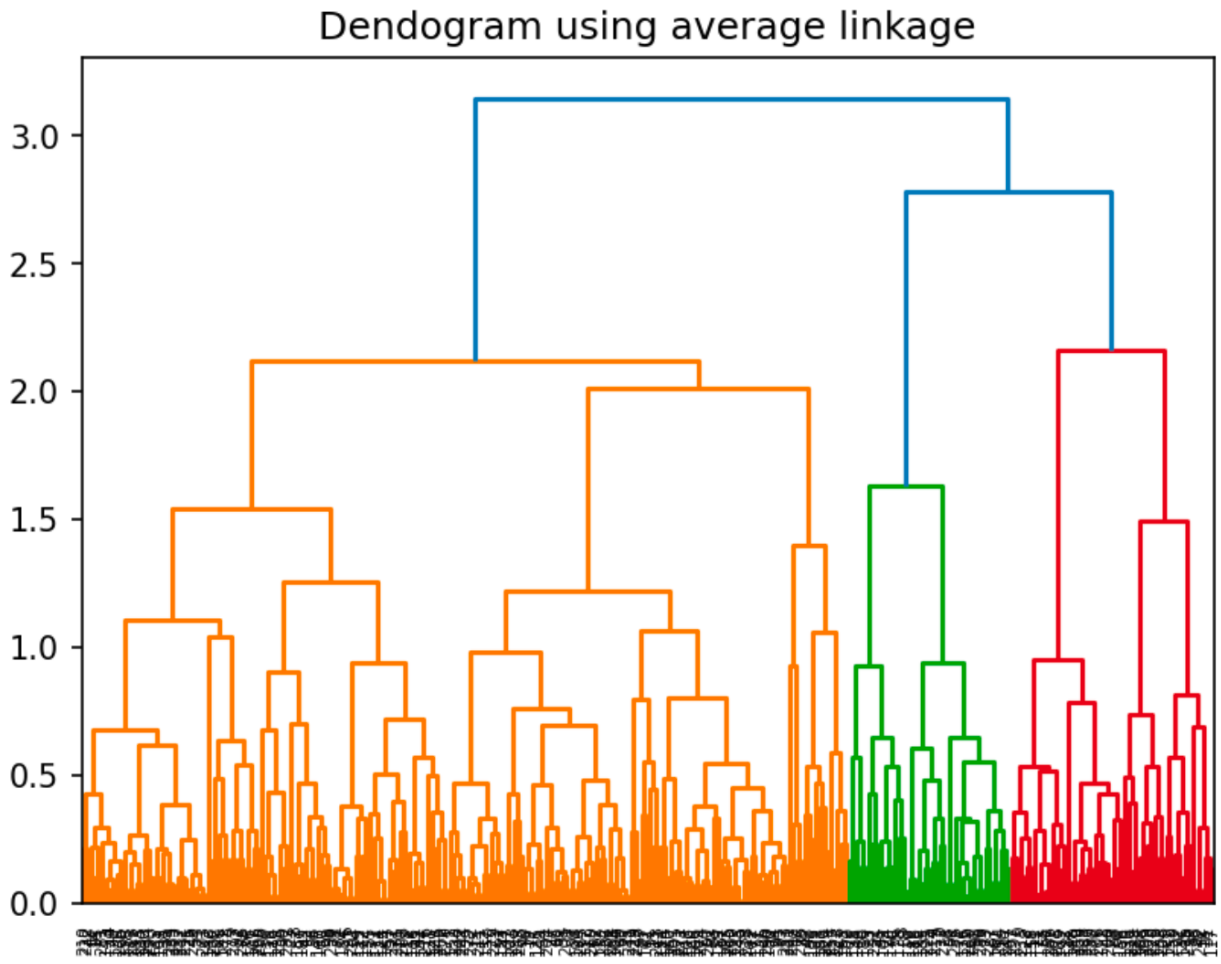
C. For bottom-up hierarchical clustering (aka ‘Agglomerative Clustering’), make a dendo- gram using ‘single’ linkage.

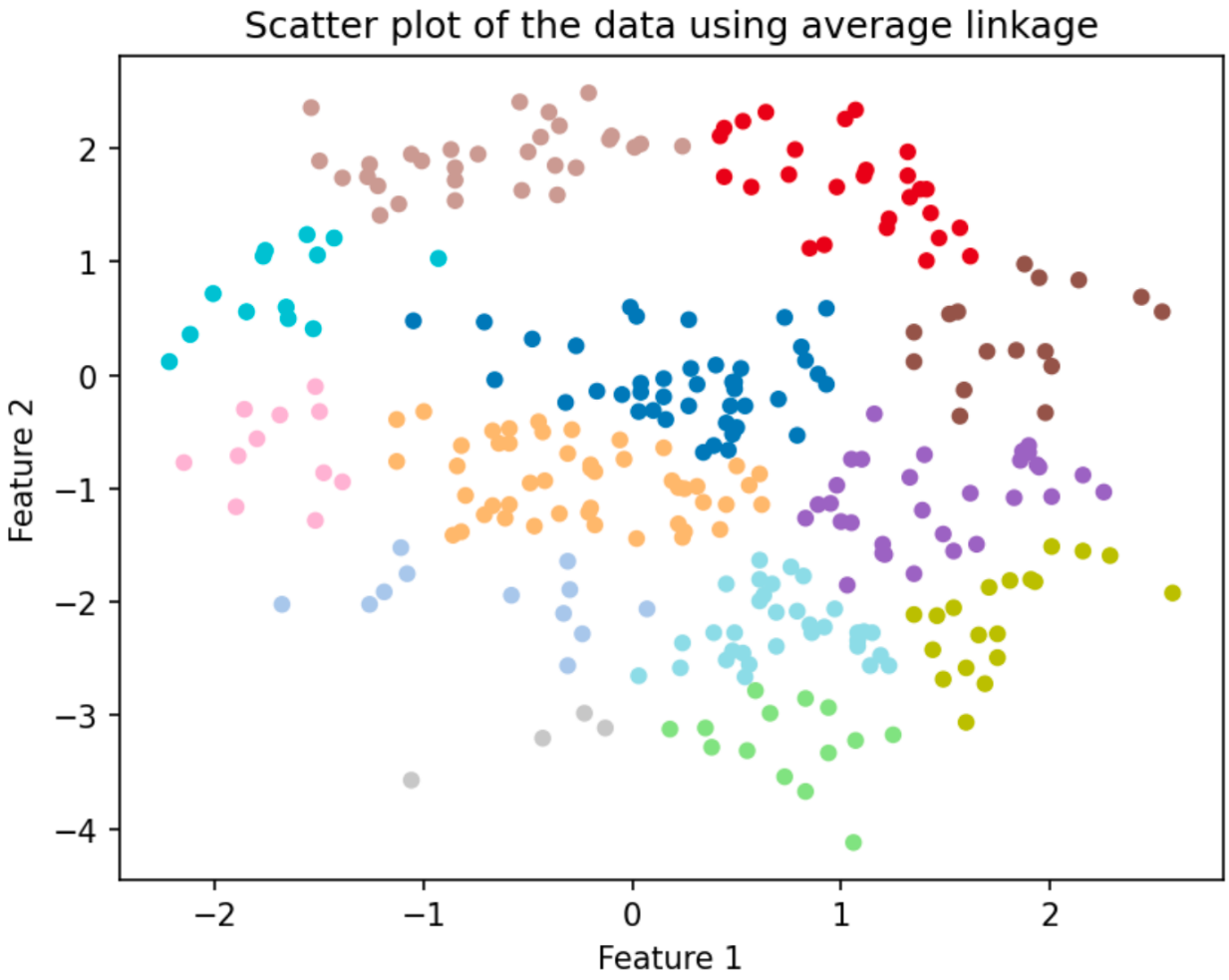


D. Manually select and report a distance threshold for single linkage. Look for regions in the dendogram where there are few mergers (eg a big vertical gap in distance threshold between mergers). Use that to make a scatter plot of the data clustered based on that threshold.

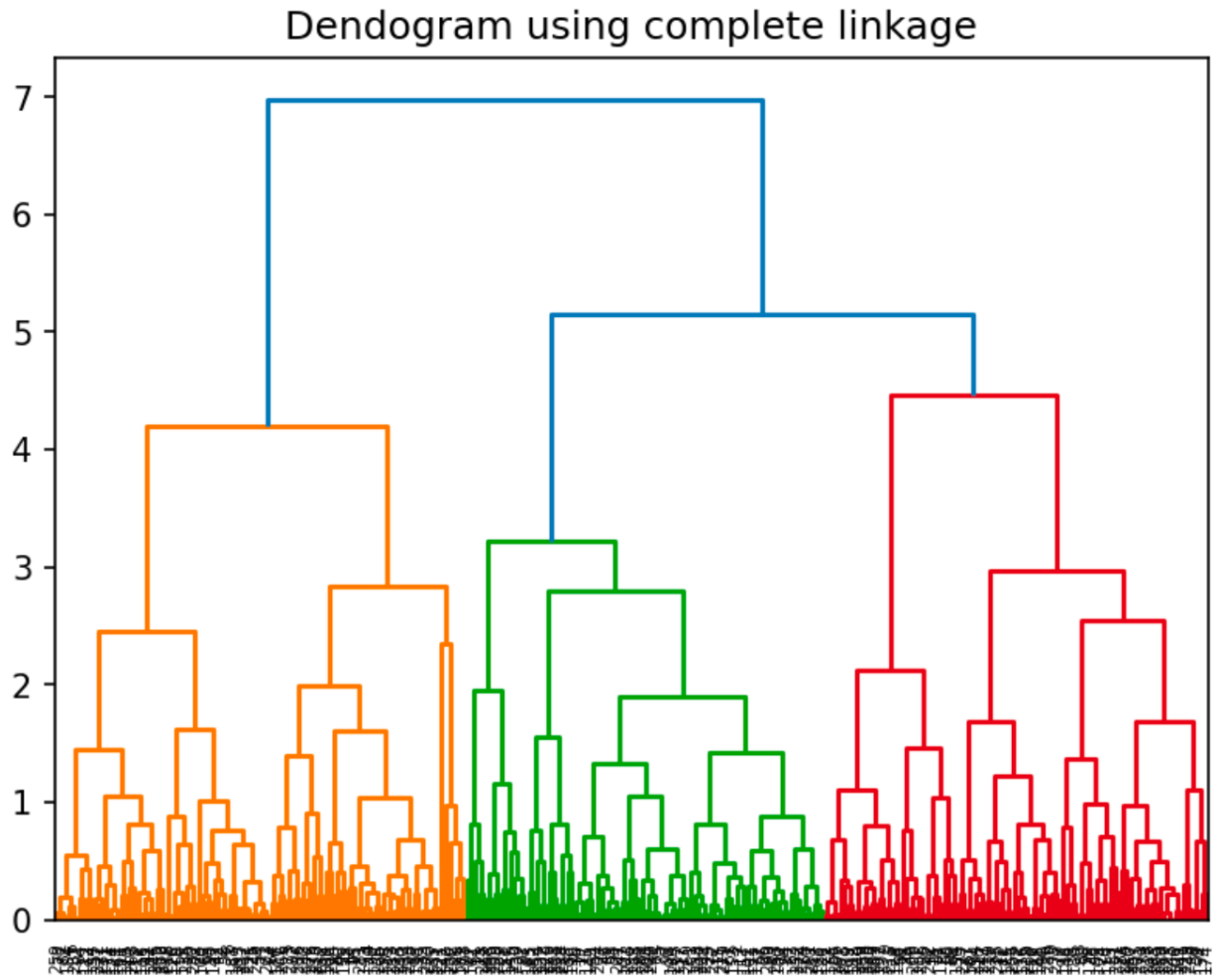


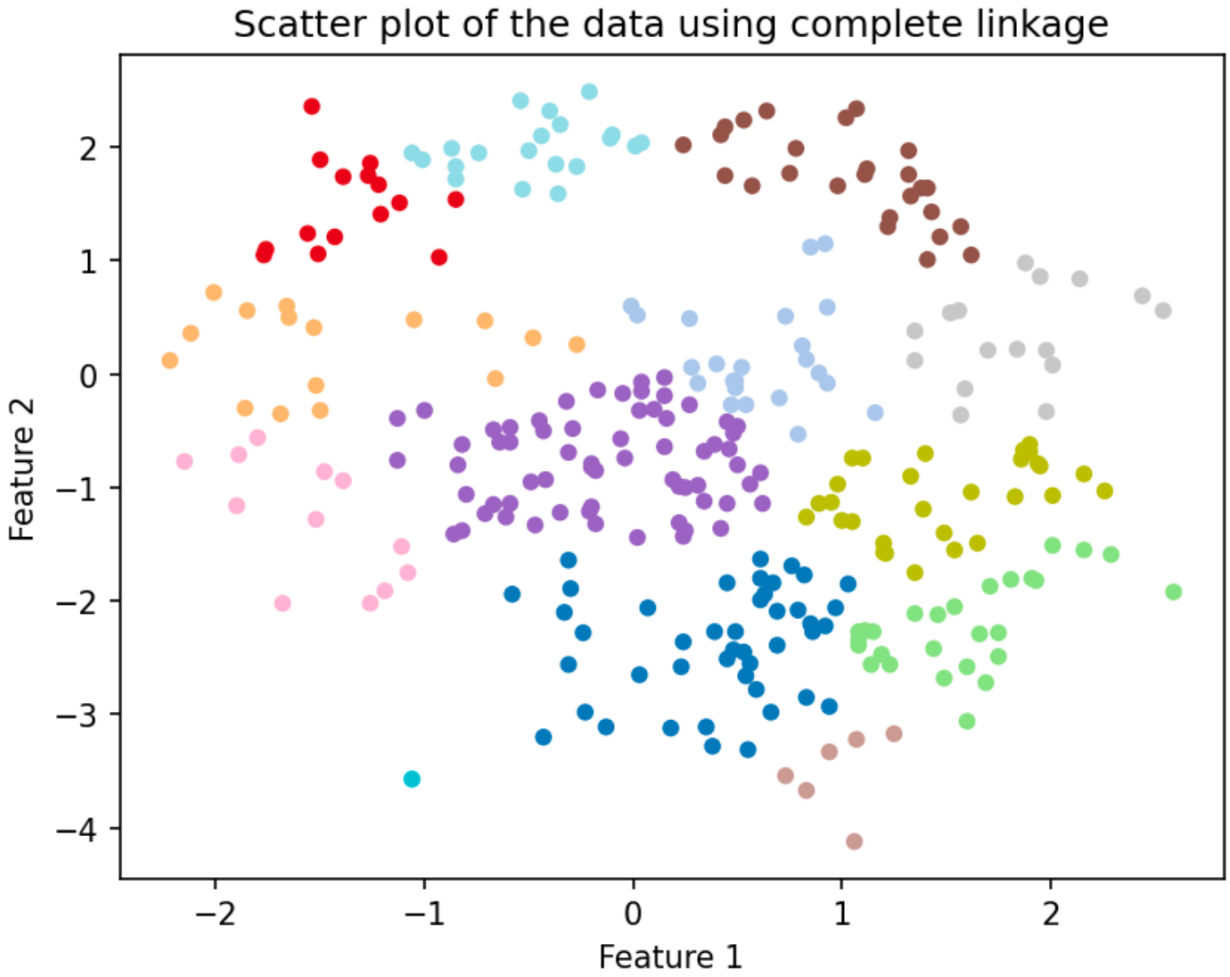
E. Repeat the previous two steps, using ‘average’ linkage.




Scatter plot of the data using average linkage
Feature 2
Feature 1
2
1
0
−1
−2
−3
−4
−2
−1
0
1
2

F. Repeat the previous two steps, using ‘complete’ linkage.




Scatter plot of the data using complete linkage
Feature 2
Feature 1
2
1
0
−1
−2
−3
−4
−2
−1
0
1
2

G. In a few sentences for each clustering method (k-means, single linkage aggl., average linkage aggl., and complete linkage aggl.), comment on the clusters found by the methods and how they compared or differed. Which clustering method do you think resulted in the best clusters for this data set and why?

Average one and complete one look fairly similar and nicely clustered whereas single one is not nicely spread out. I think the one using average linkage is better than others for this dataset. It is because compared to the rest of them, this one is nicely evenly spread out.