

Data Mining Program Assignment 2 Report

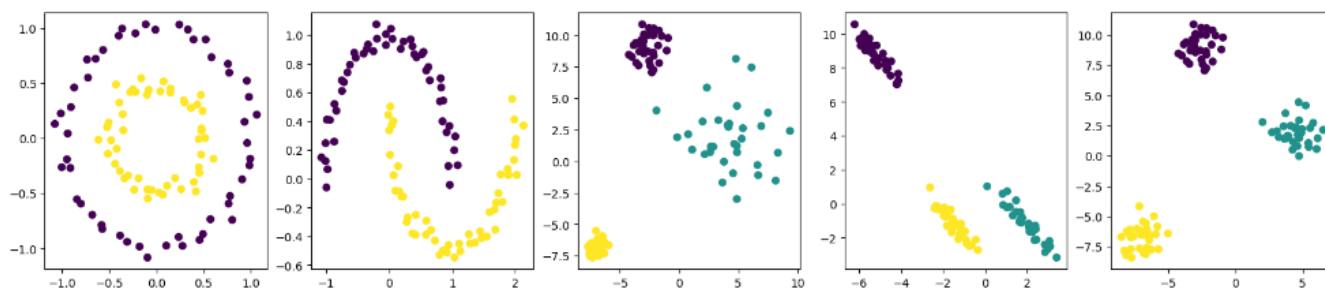
Hunter Garrison

Due: March 21, 2024 @ 11:59PM

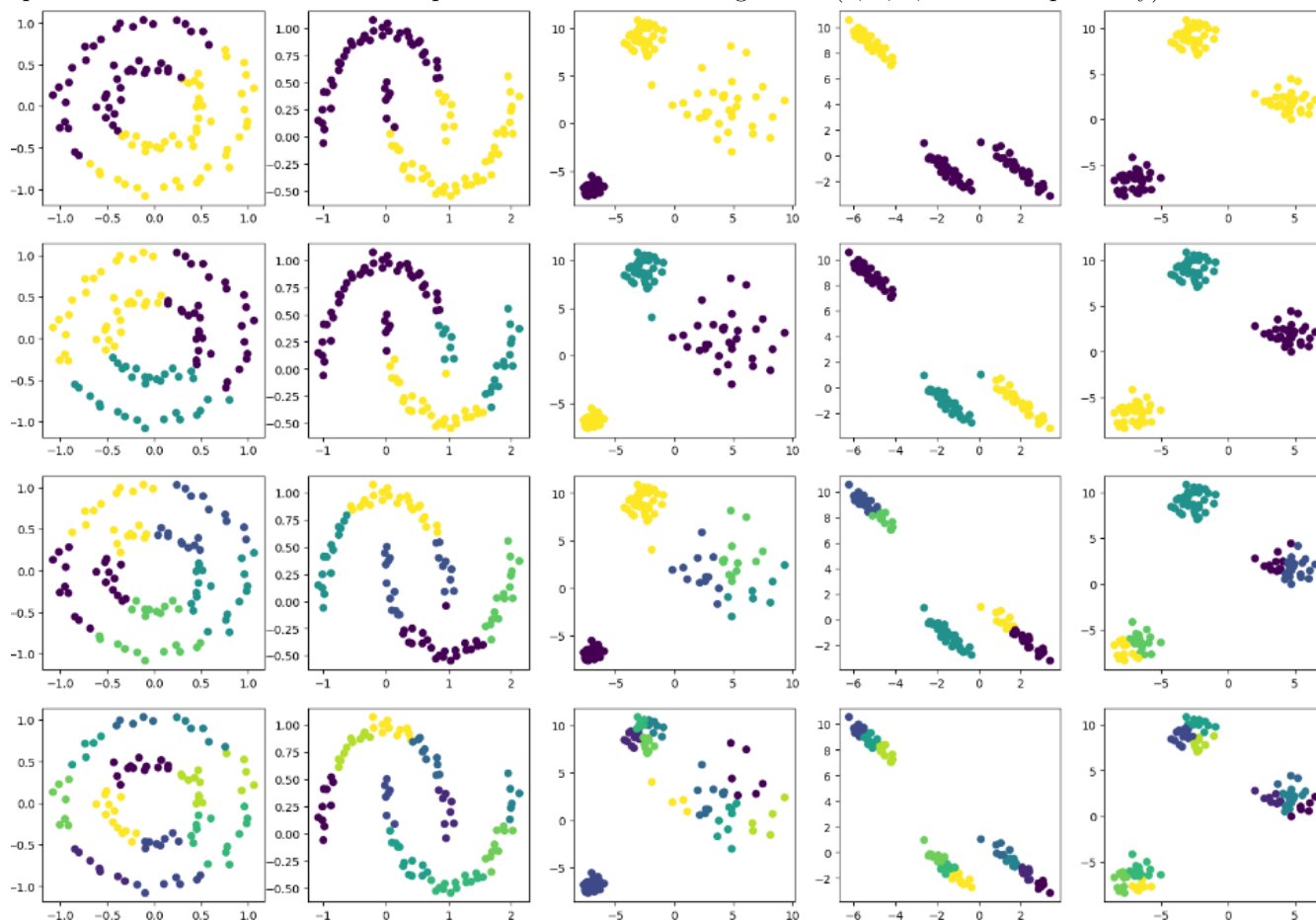
Q1. Evaluation of k-Means over Diverse Datasets

Part C

We start with showing the datasets and their respective true clustering.



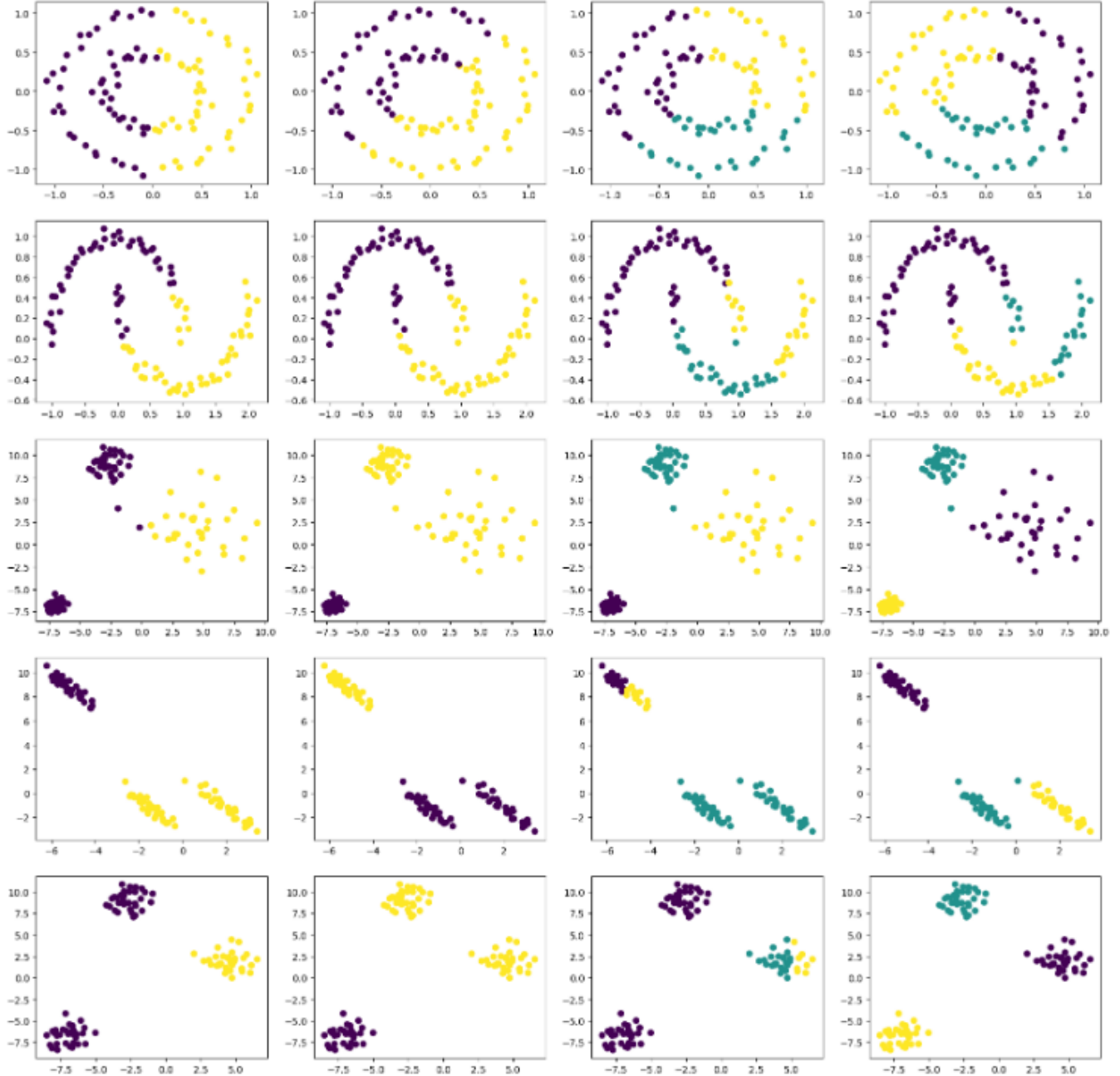
Below is a 5 column by 4 row figure of scatter plots where each column pertains to a certain dataset and each row pertains to the number of clusters specified in the K-Means algorithm (2, 3, 5, and 10 respectively).



We can see that the only dataset in which the K-Means clustering produced correct clusters was for the blobs dataset (abbreviated "b" in the coding assignment) with denoted clusters of 3. All other datasets were classified incorrectly after using K-Means.

Part D

Below is a 5 row by 4 column figure where each row pertains to a certain dataset and each column denotes specified initial centroids for 2 clusters, random initial centroids for 2 clusters, specified initial centroids for 3 clusters, and initial random centroids for 3 clusters, respectively.



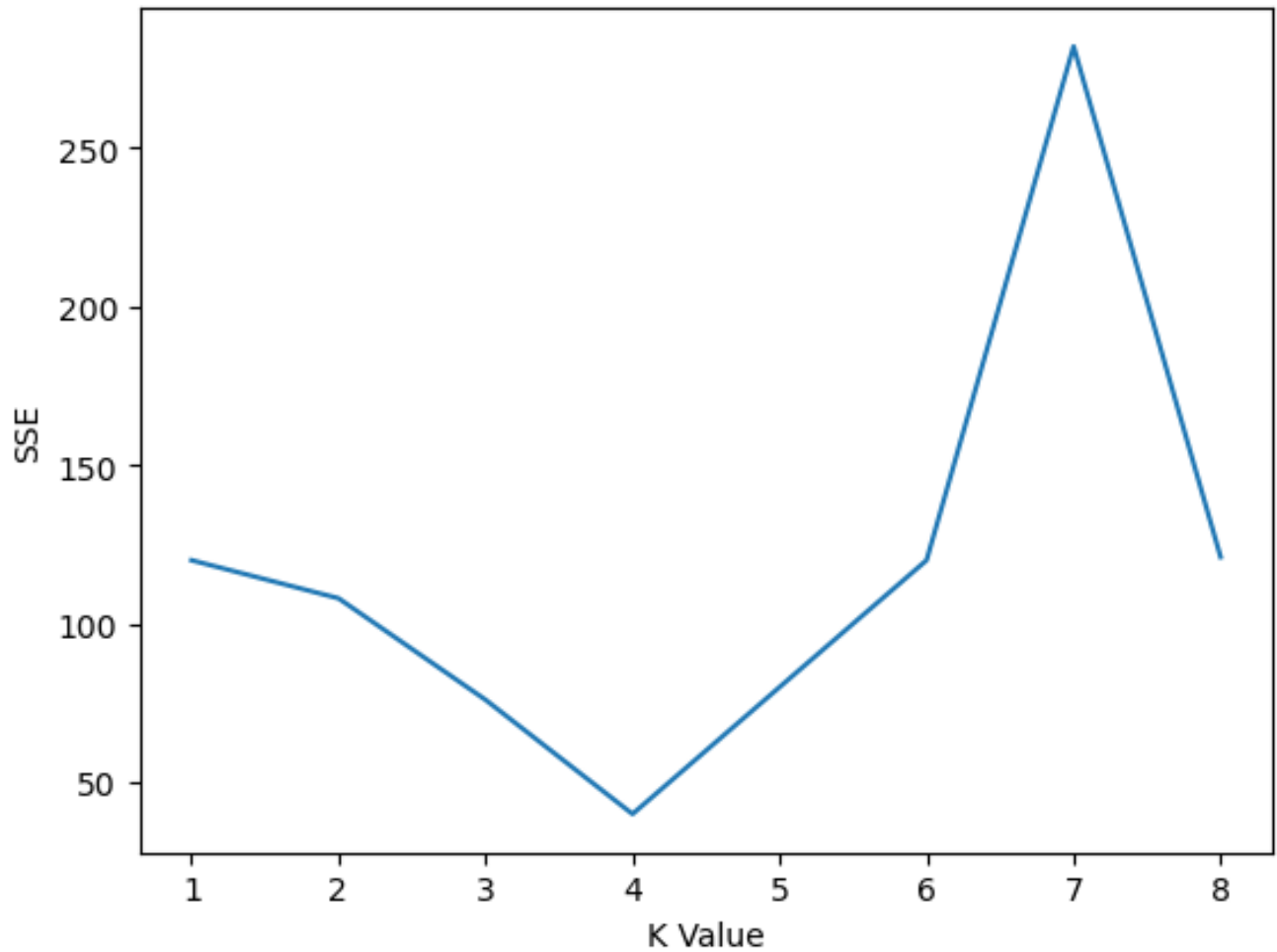
We can see that every single dataset is sensitive to the choice of initialization. For the noisy circles, our clusters for $k=2$ start with the division between the clusters to be a diagonal line, but can be changed to have the division between the clusters to be a line down the middle (centroids were initialized at $(0,0)$ and $(1,0)$). For the noisy moons, our clusters for $k=2$ include an extra point in the left cluster when we initialize centroids as $(0,0)$ and $(1,0)$. For the blobs with varied variances, when we initialize centroids at $(0,0)$ and $(1,0)$ we get the left two clusters merged together whereas random initialization merged the left two clusters together. For the anisotropically distributed data, upon initializing centroids at $(10,0)$, $(0,0)$, and $(0,0)$, we end up splitting the top left cluster into two sections. Finally, for

the blobs we can split the middle right cluster into two sections when initializing centroids at $(0,0)$, $(1,0)$, and $(2,0)$.

Q2. Comparison of Clustering Evaluation Metrics

Part C

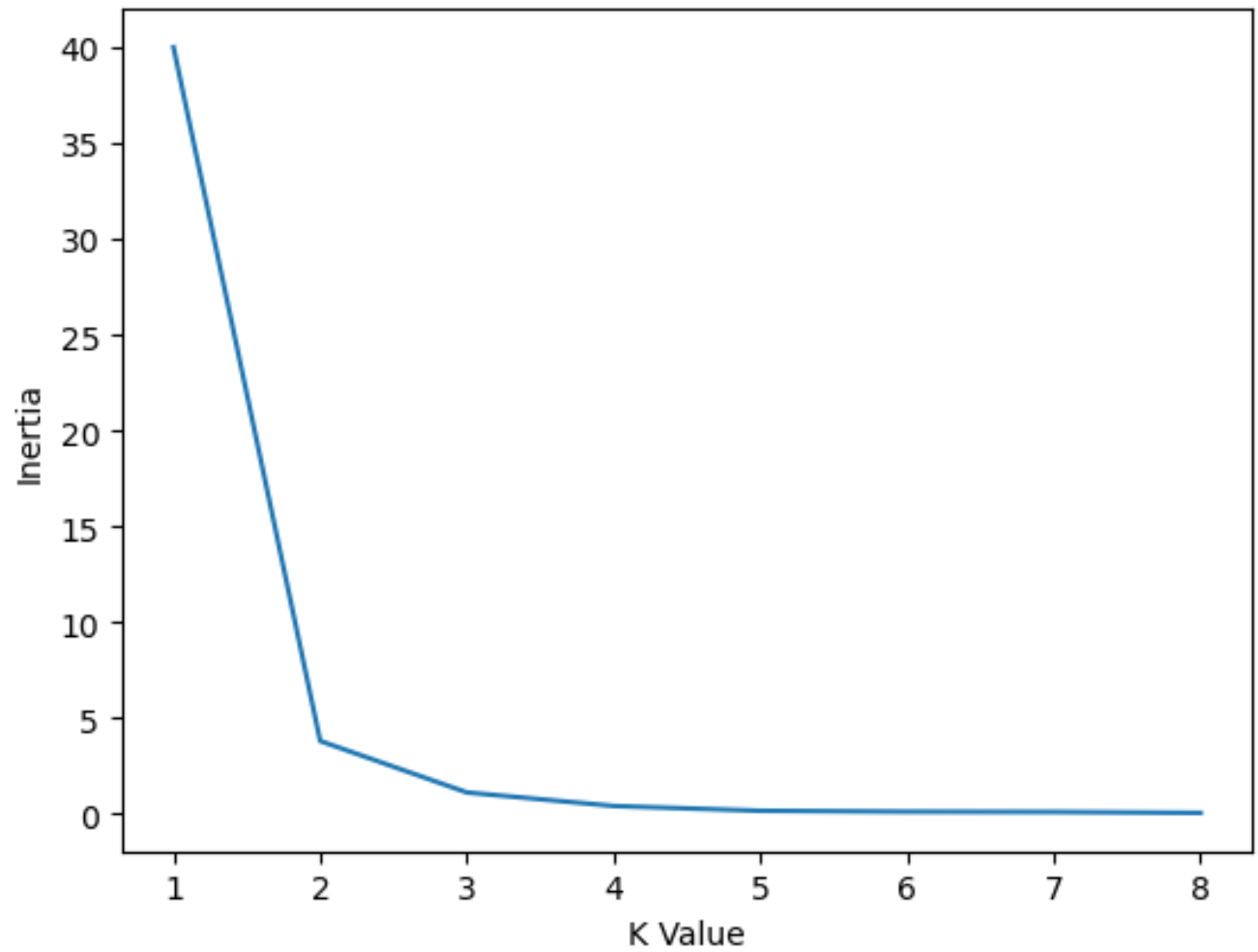
Below is a plot of SSE as a function of k for $k=1, 2, \dots, 8$:



According to the elbow method, we want to choose $k=4$ as our optimal k value.

Part D

Below is a plot of inertia as a function of k for $k=1, 2, \dots, 8$:

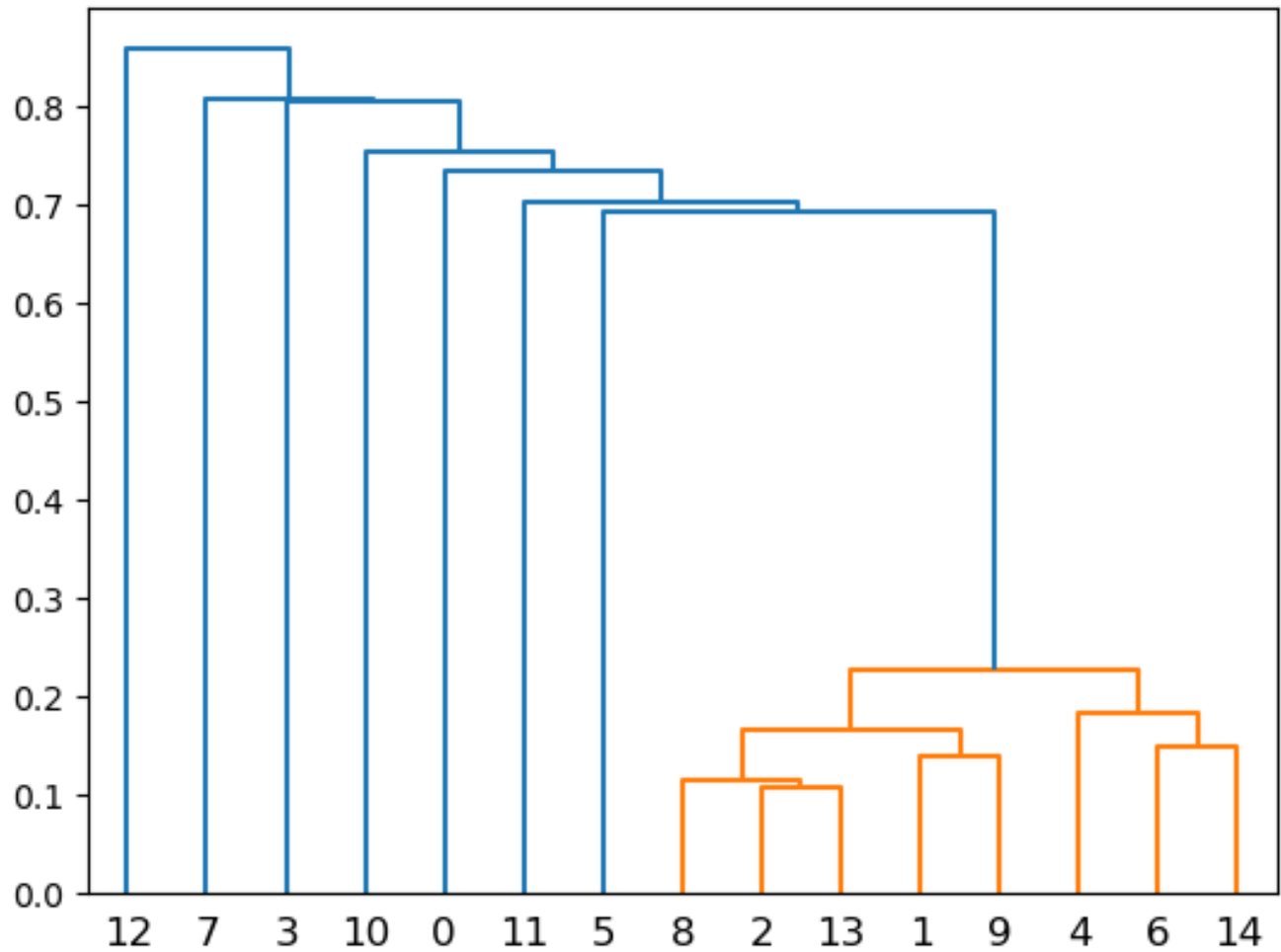


Using the elbow method, it looks like $k=2$ should be our optimal k value, which does not agree with our results from 2.C.

Q3. Hierarchical Clustering

Part B

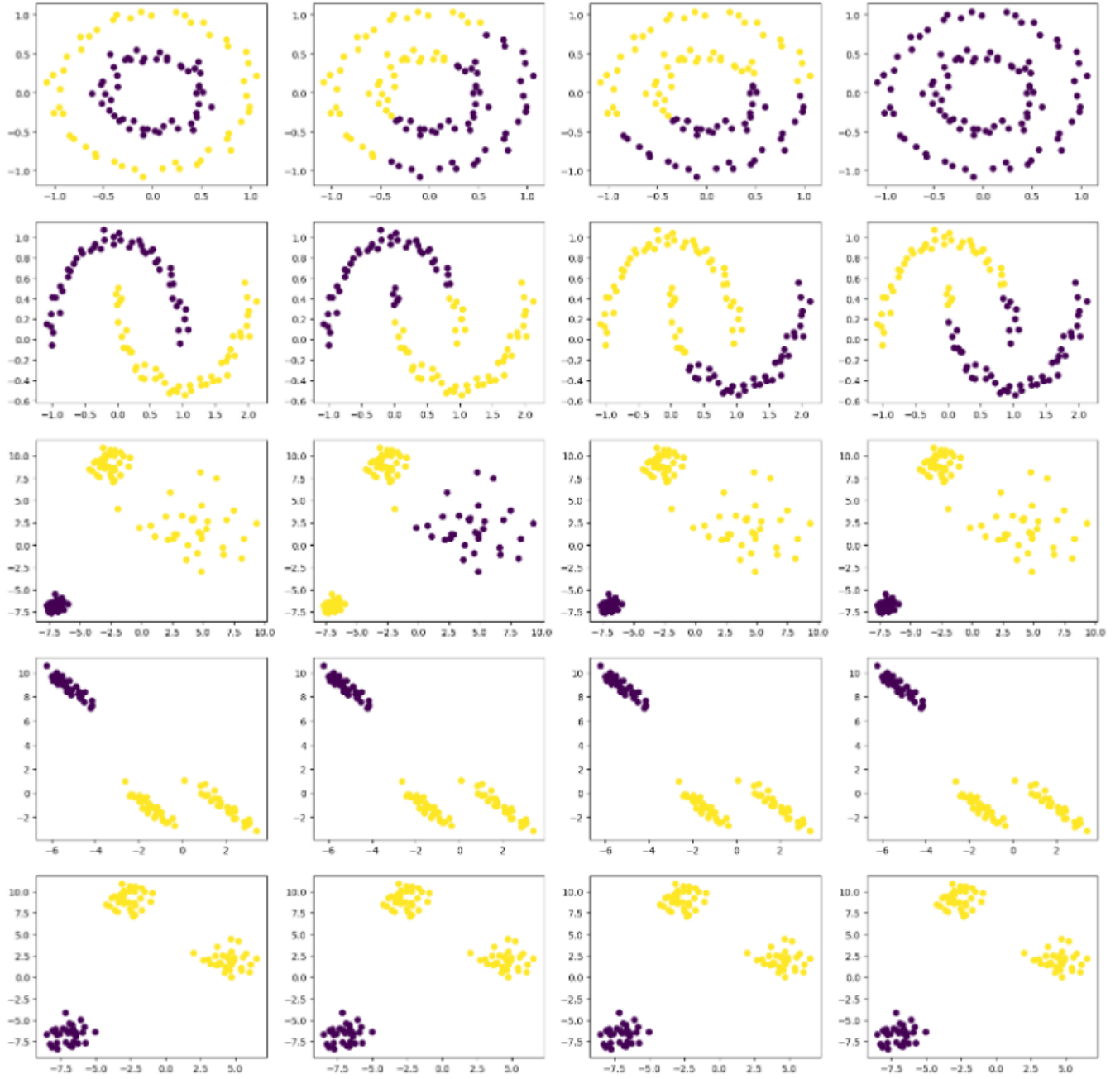
Below is a plot of the dendrogram from the hierarchical toy dataset:



Q4. Evaluation of Hierarchical Clustering over Diverse Datasets

Part B

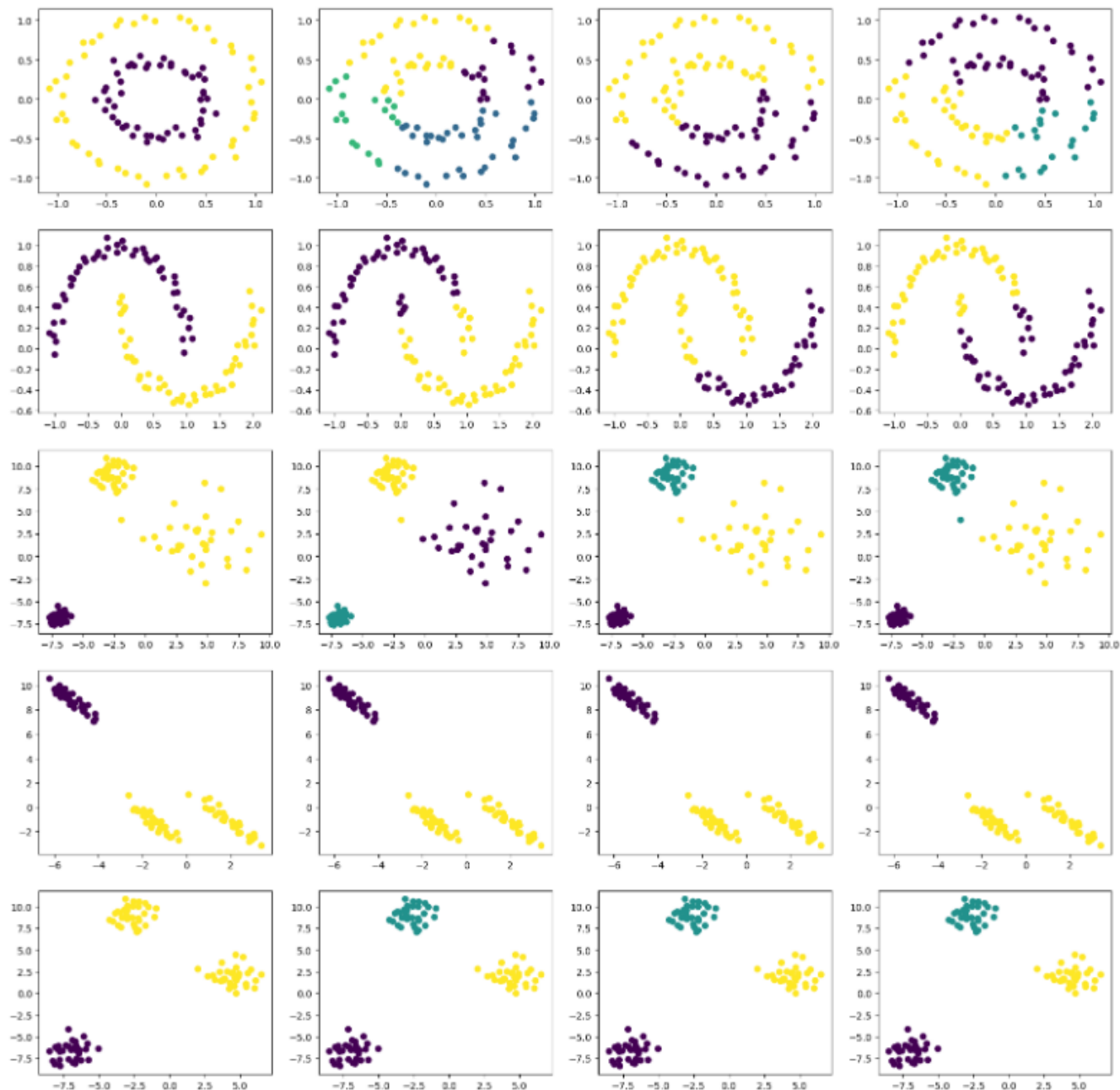
Below is a 5 row by 4 column plot where each row pertains to a specific dataset and each column pertains to single, complete, ward, and centroid linkage type, respectively, using agglomerative clustering:



The noisy circles and noisy moons datasets are now clustered correctly when using the single linkage type.

Part C

Below is a 5 row by 4 column plot where each row pertains to a specific dataset and each column pertains to single, complete, ward, and centroid linkage type, respectively, using agglomerative clustering with cut-off distance:



We see that we can correctly identify the clusters with each dataset (with varying link types) except for the "add" dataset.