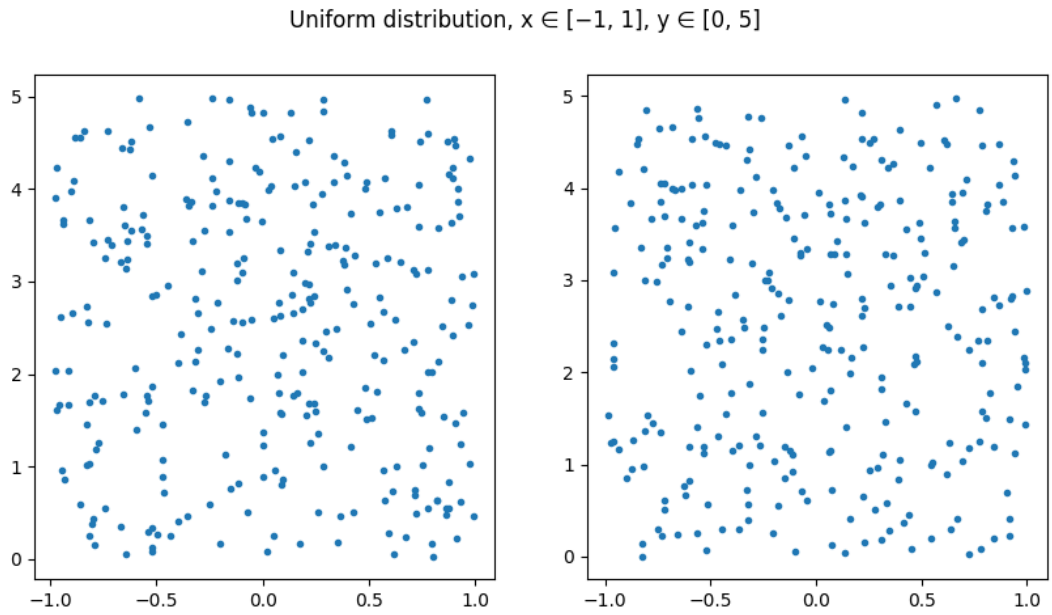
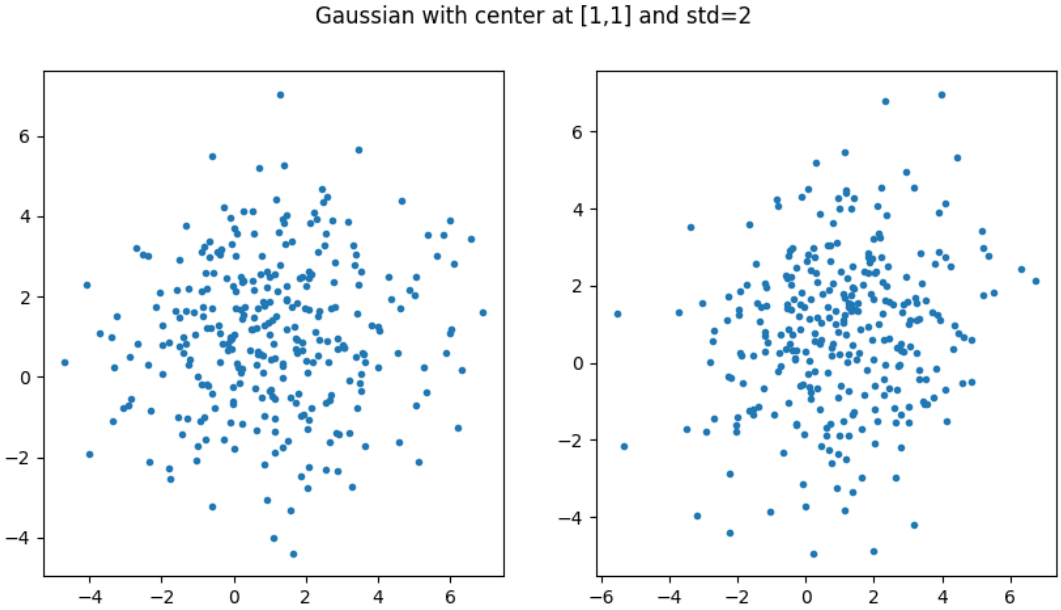
Homework #2: Similar items, Clustering, Community Detection

Problem 3

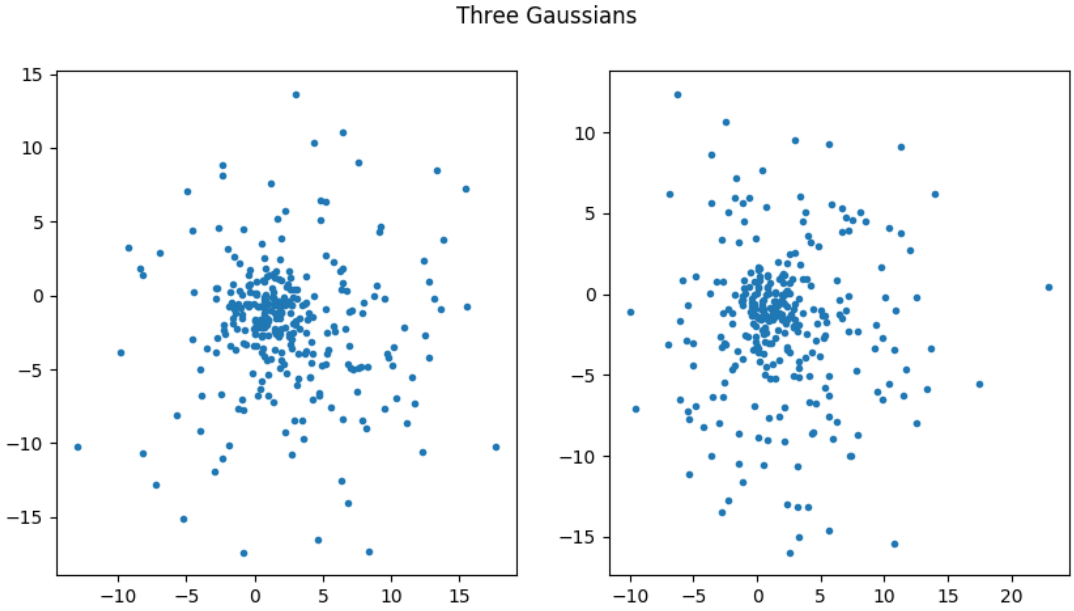
(a)



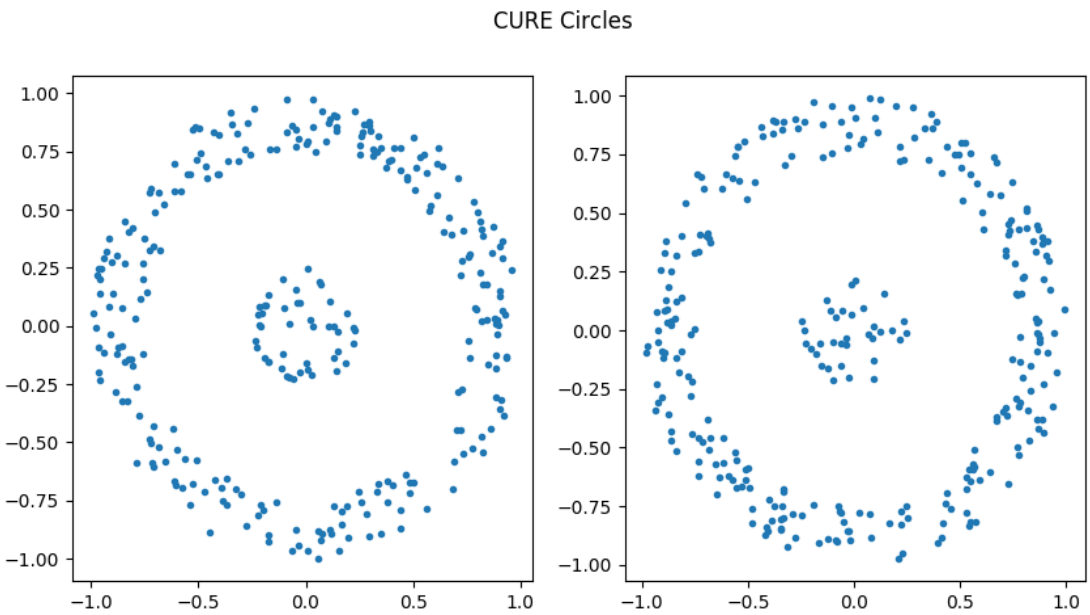
(b)



(c)



(d)



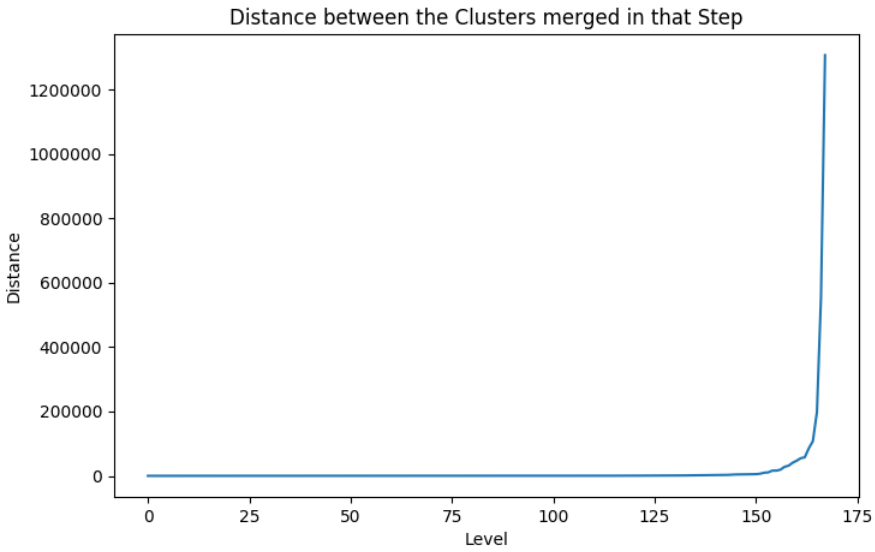
(e)

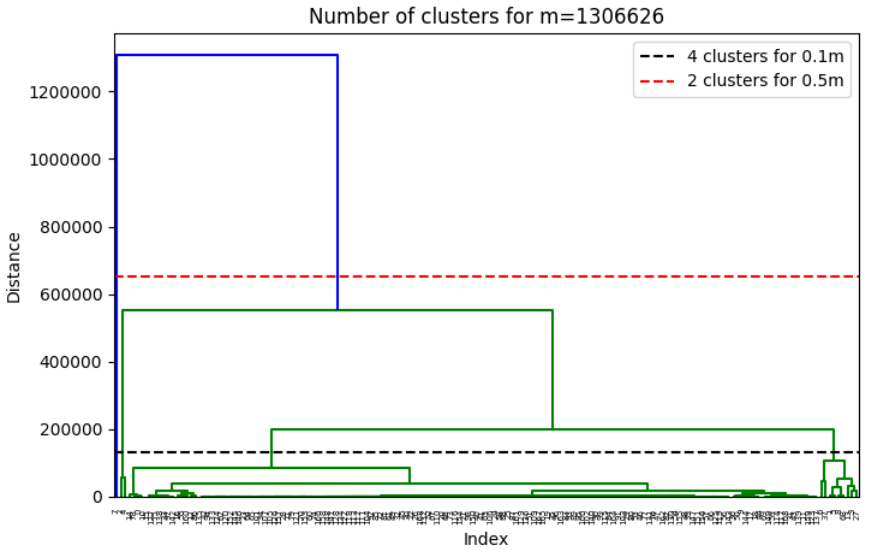

The first letter of both your first names - AF
8
6
4
2
0
−6
−4
−2
0
2
4
8
6
4
2
0
−6
−4
−2
0
2
4

Problem 4

For hierarchical clustering and the plot of the dendrogram we use the package `scipy.cluster.hierarchy`. Function `linkage` provides the hierarchical clustering algorithm for different methods. We use `method='centroid'` in part (a) and `method='complete'` in part (b). Function `dendrogram` is used to plot the dendrogram.

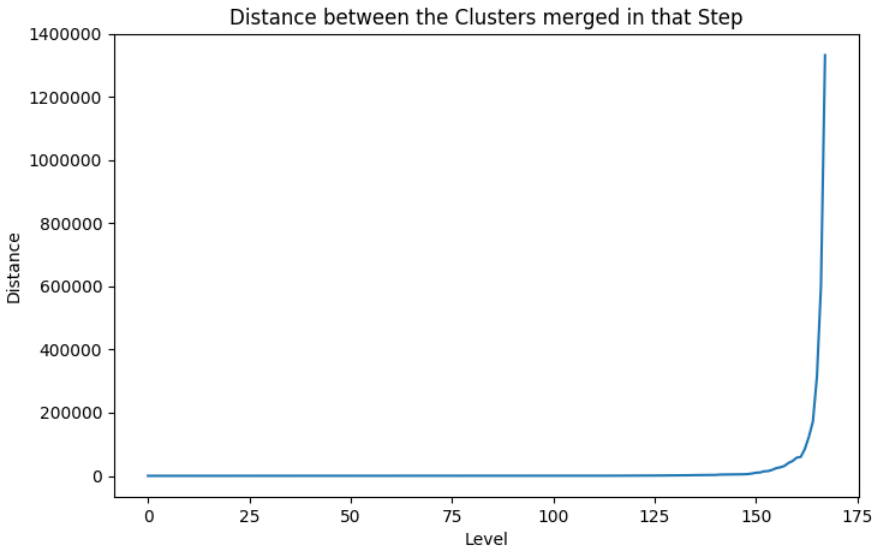
(a) merge clusters with the closest centroids

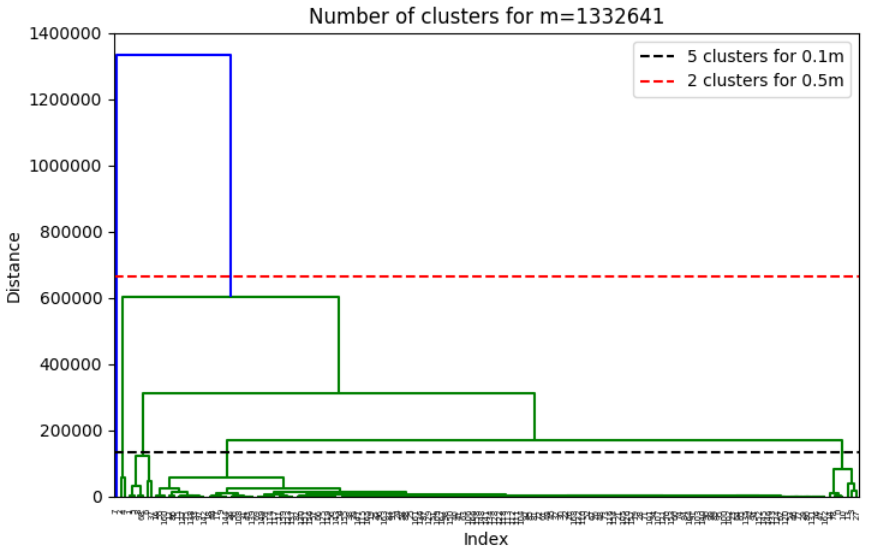




The Dendrogram shows that we would get 4 clusters if we decide to stop clustering at a distange of 0.1m and 2 clusters if we decide to stop at 0.5m.

(b) merge clusters so that the new diameter is the smallest among all options



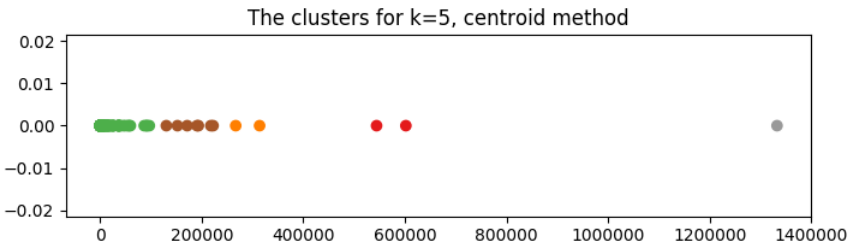


The Dendrogram shows that with this method we would get 5 clusters if we decide to stop clustering at a distange of 0.1m and 2 clusters if we decide to stop at 0.5m.

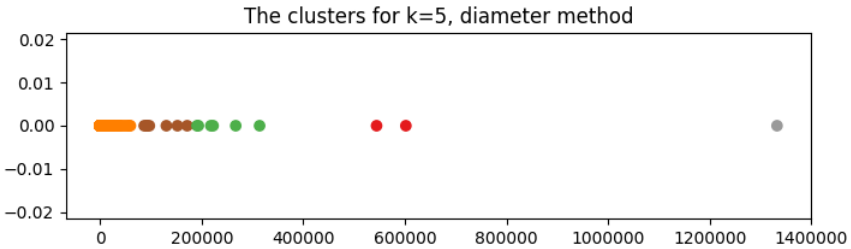
Comparison

When we evaluate the Dendograms of the clustering methods, we conclude in both cases that our data is best represented by k=5 clusters. In this task we have the special case of just one dimensional data that we can visualize easily and compare the outcome with our expectation (since humans are incredibly good in clustering easy patterns).

In the case of centroid clustering the clusters represent the structure of our data pretty well.



Whereas in the diameter case, the algorithm assigns data points to clusters that don't seem to make sense from our (visual) intuition.



This shows that the choice of method and how you measure the distance always depends on the question that is asked about the data.