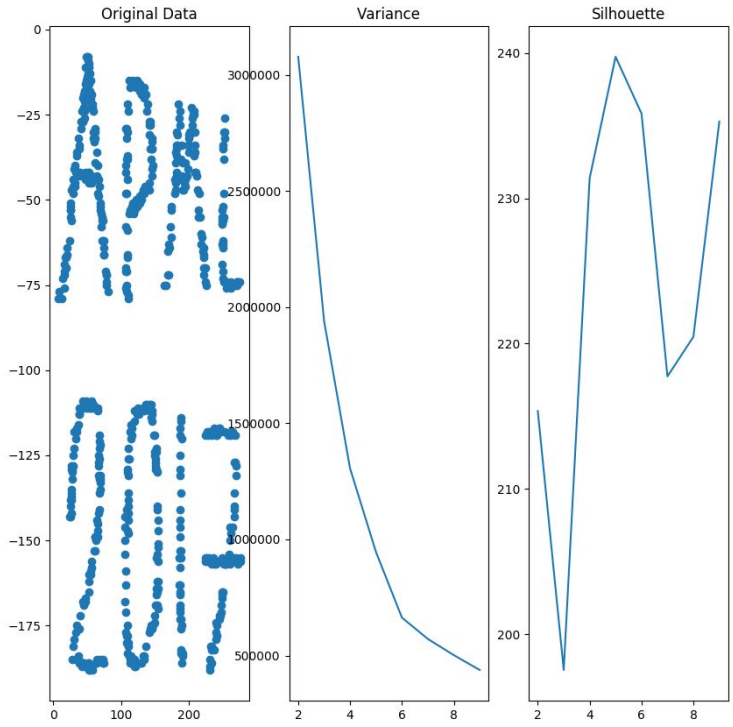
Ex3 - Clustering

lior ziv - 305742611

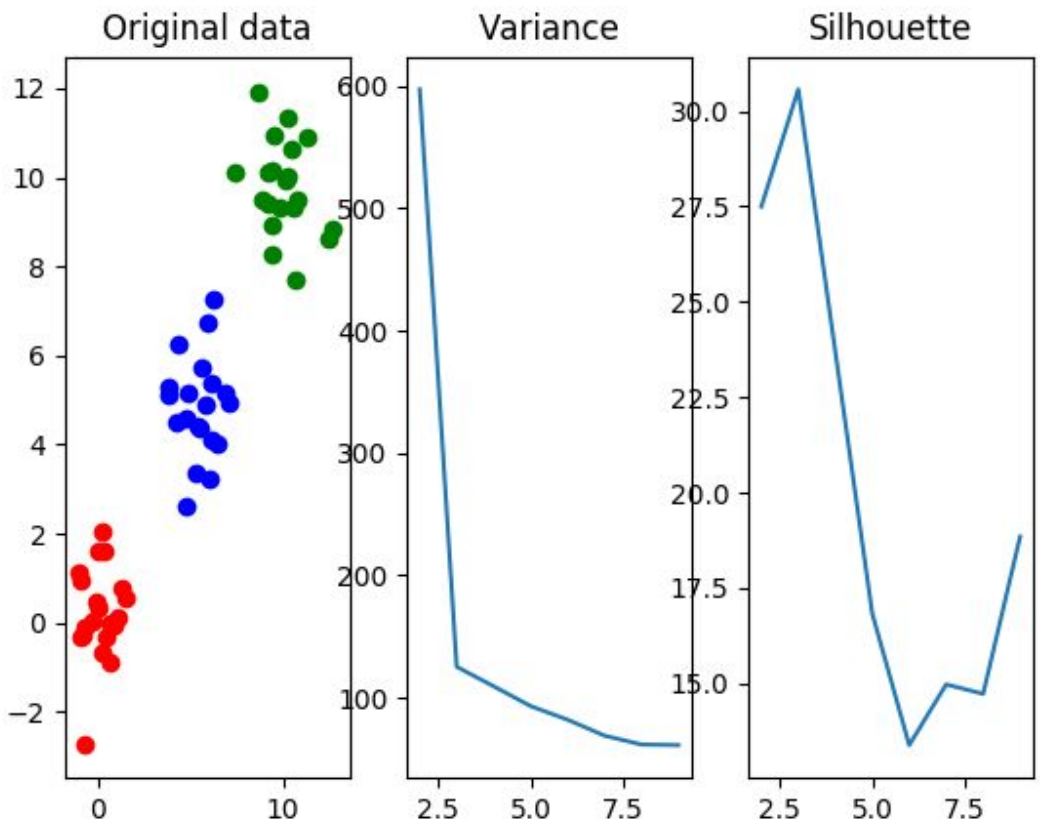
K - selection

There are methods in order to choose the optimal k for our data, here we will demonstrate two data methods.
Elbow method - which is the inner variance of each cluster
Silhouette - which is a method to check how much each point is assigned to its optimal cluster.

I tried those methods on two different data sets



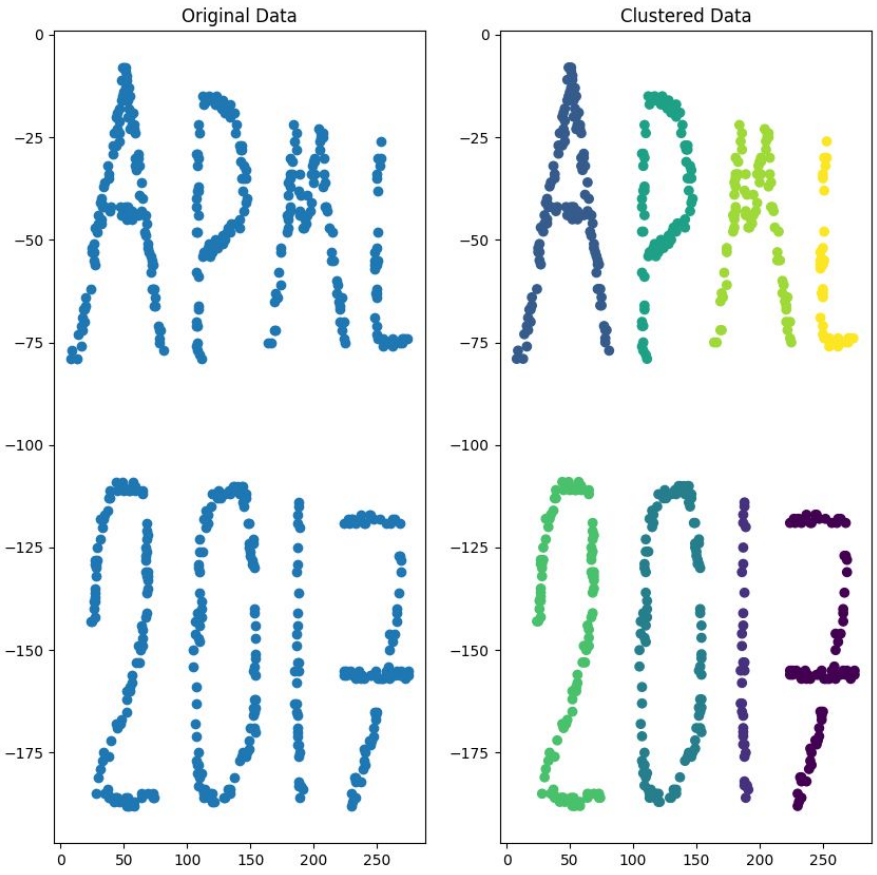
On the AMPL 2017 picture we can see that at around k=6 in the elbow method the inner variance don't change much therefore according to it I would choose k=6.
The silhouette chose 5 as it's maximal value meaning that when we have k=5 we get that the distance of each point from its cluster is minimal compared to the next cluster it would have been assigned to.



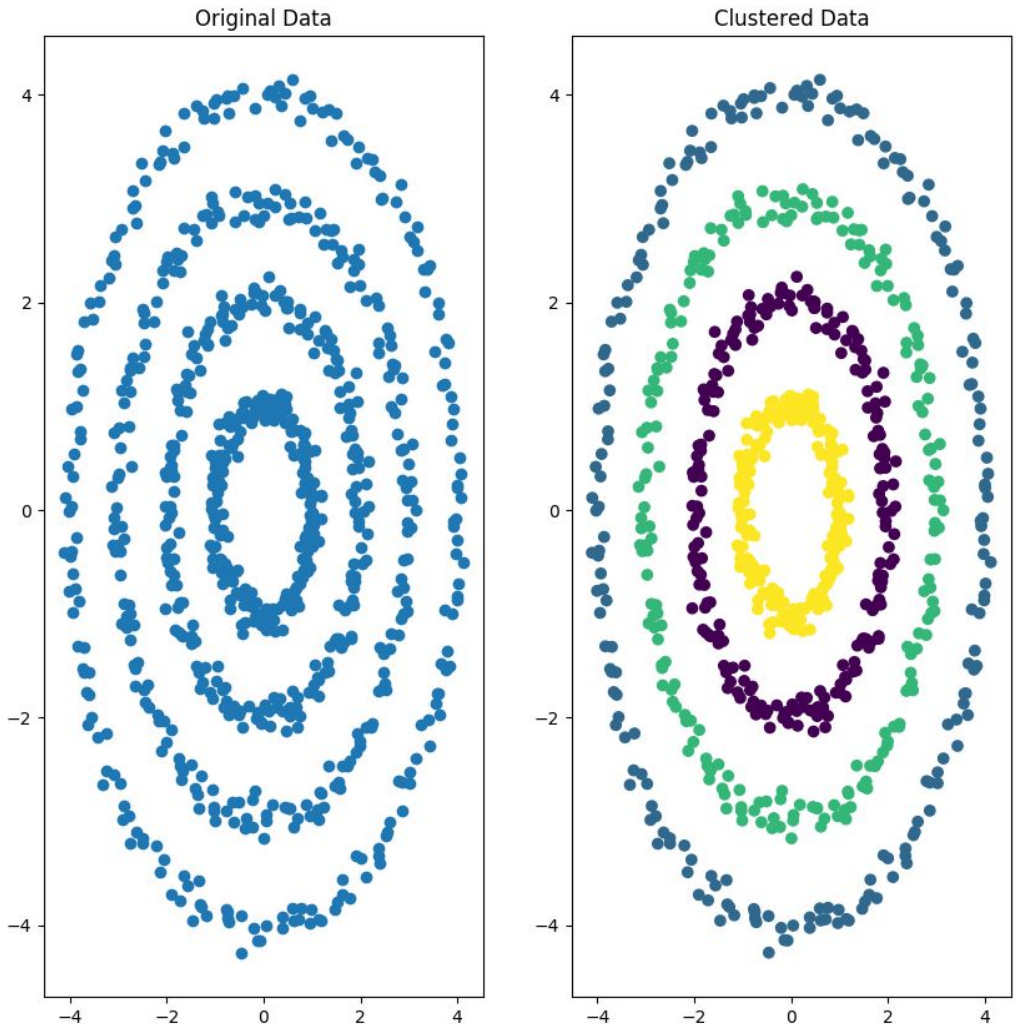
Here the data clusters are a bit more clear therefore we can see that both the elbow method and the silhouette chose k = 3 as their optimal value, which is indeed what we would expect the optimal k will be.

Spectral Clustering

This algorithm first takes the data, separates it into different clusters and send them to distal places in the space, after this stage k-means if performed on the data.
I demonstrated the algorithm outcome with two different data sets :
Using MNN (with k = 8 and number of neighboors = 8)



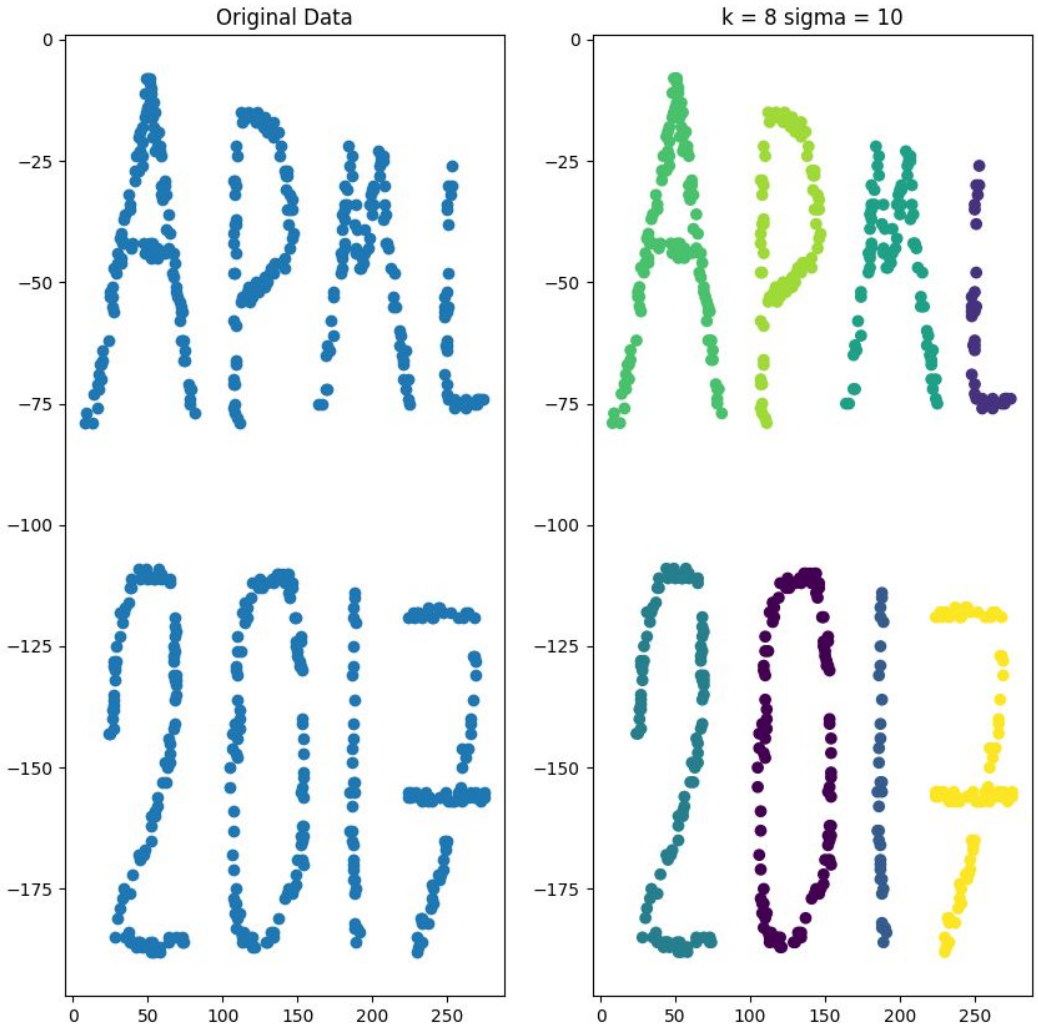
Here you can clearly see that each letter got its own colors since it is in a different cluster according to the spectral cluster algorithm.



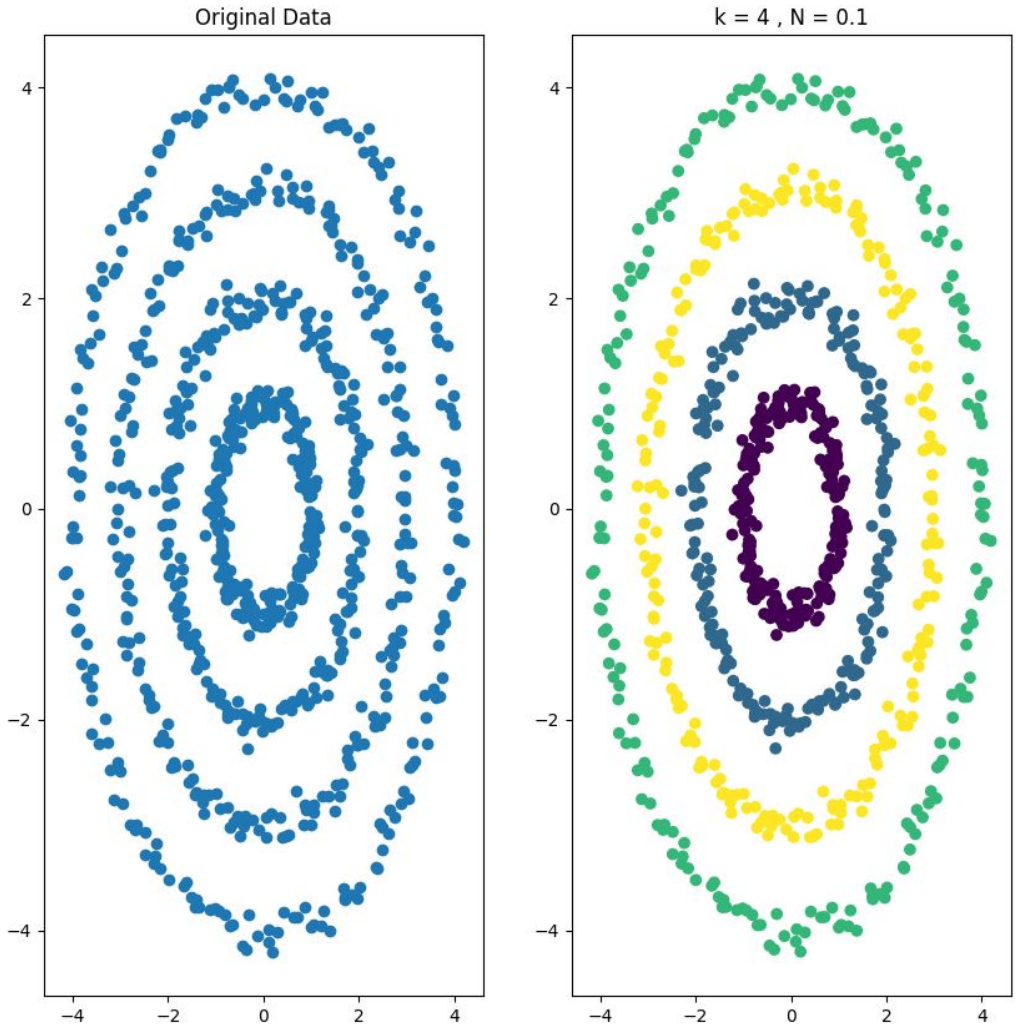
Using MNN (with k = 4 and number of neighboors = 7))

In the circles data I chose k = 4 and the spectral clustering indeed succeeded to cluster the data into the 4 different clusters you would expect when looking at this data.

Using heat kernel (k = 8 and sigma = 10)



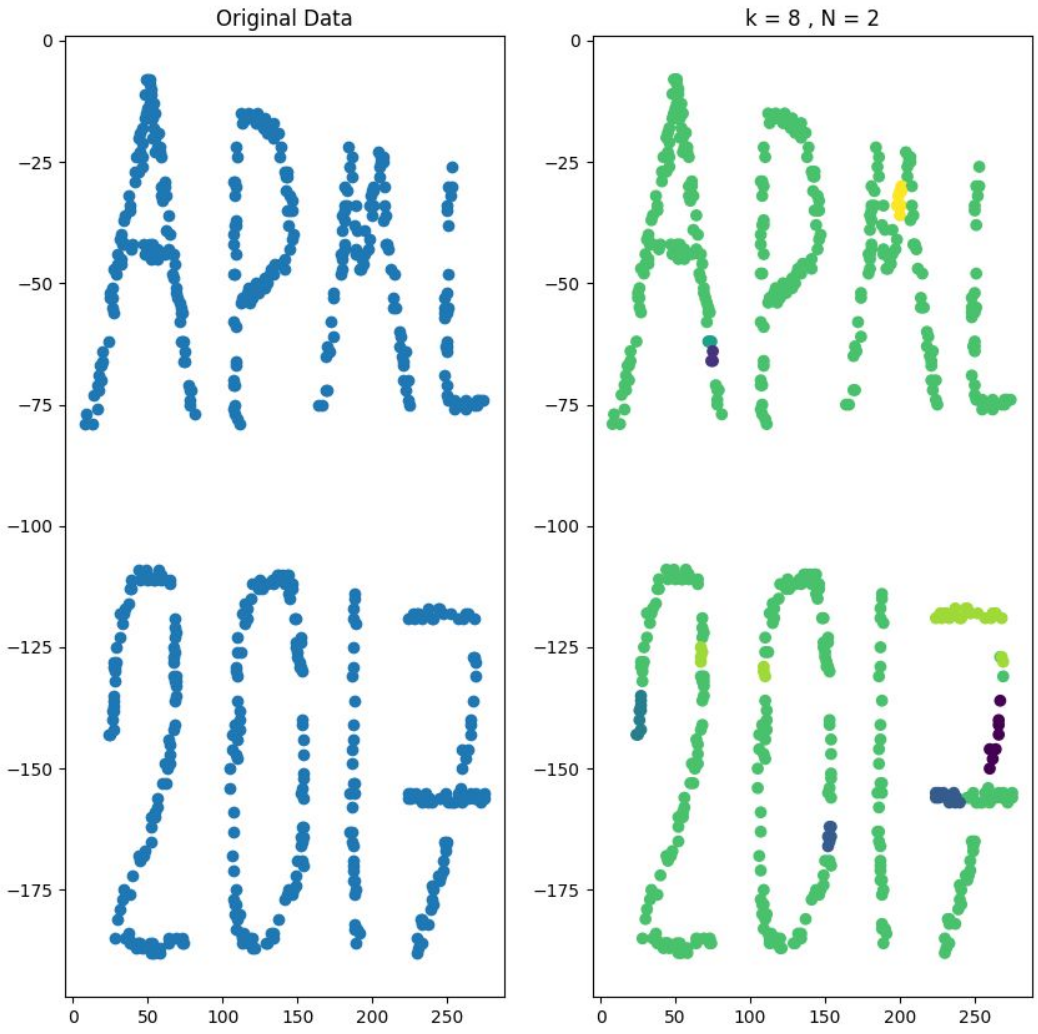
Also here you can see that the heat kernel managed to separate the data into clusters,



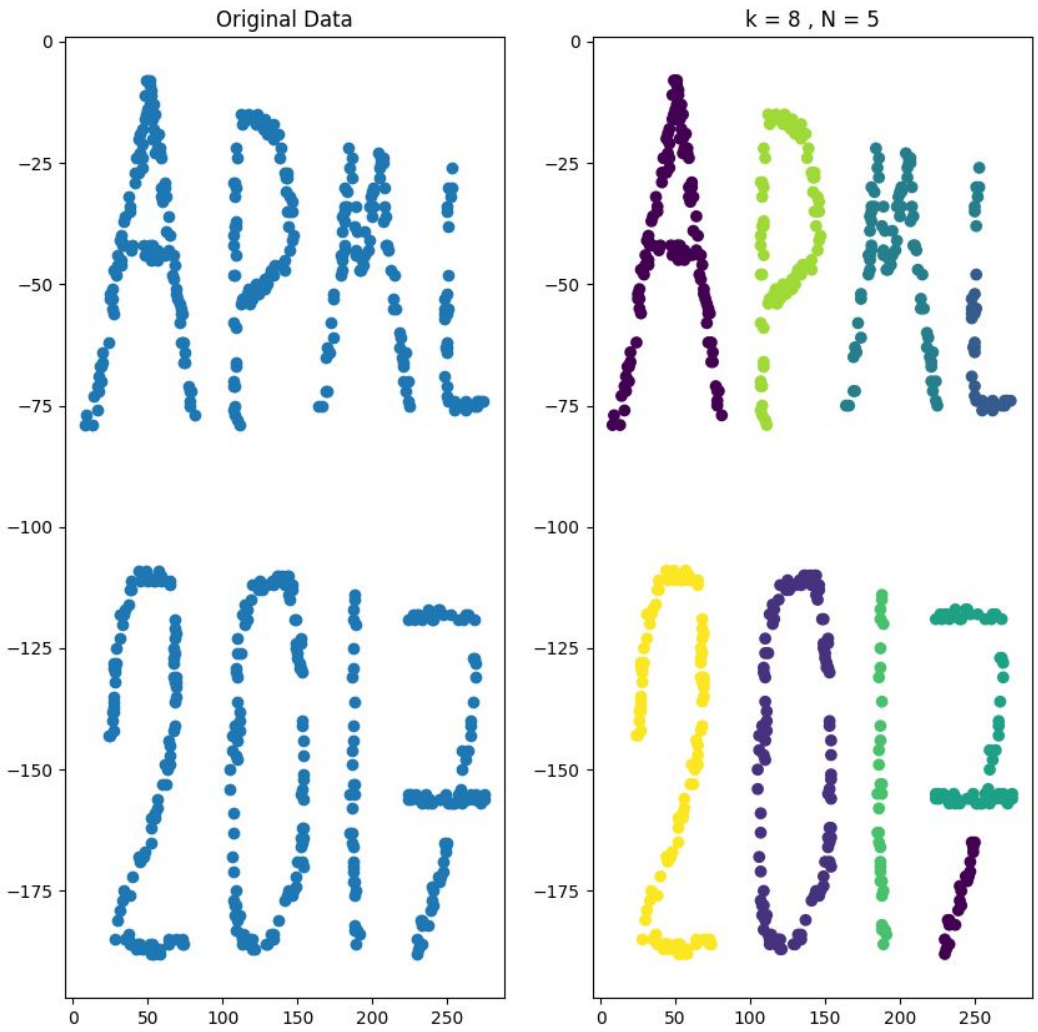
Using heat kernel (k =4 and sigma = 0.1)

We can see that using heat kernel also brings good results using the right parameters.

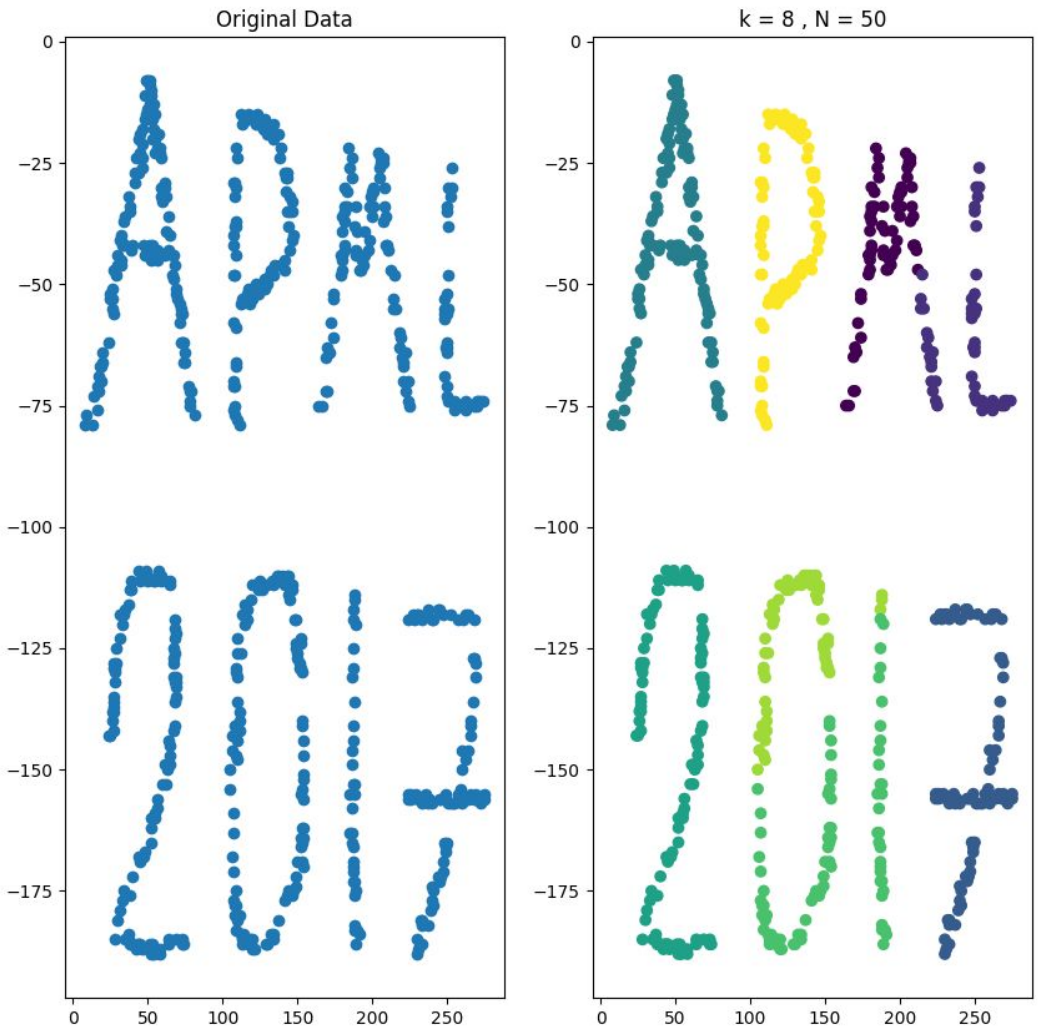
More rounds with diffrent parameters -kmenas



You can clearly see that it works bed, the number of neighbors affect the result, it is to little

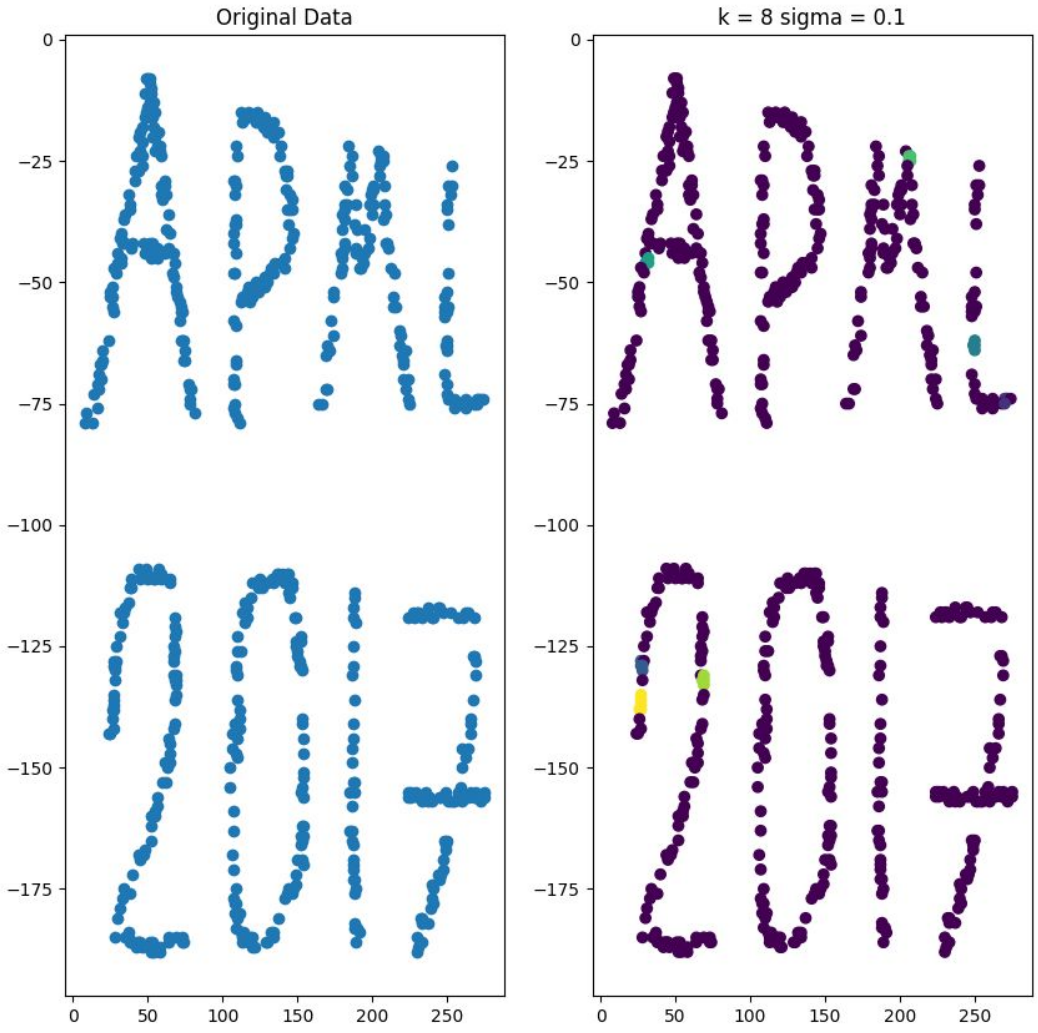


Here it gets better since we have more neighbors, but still not perfect.

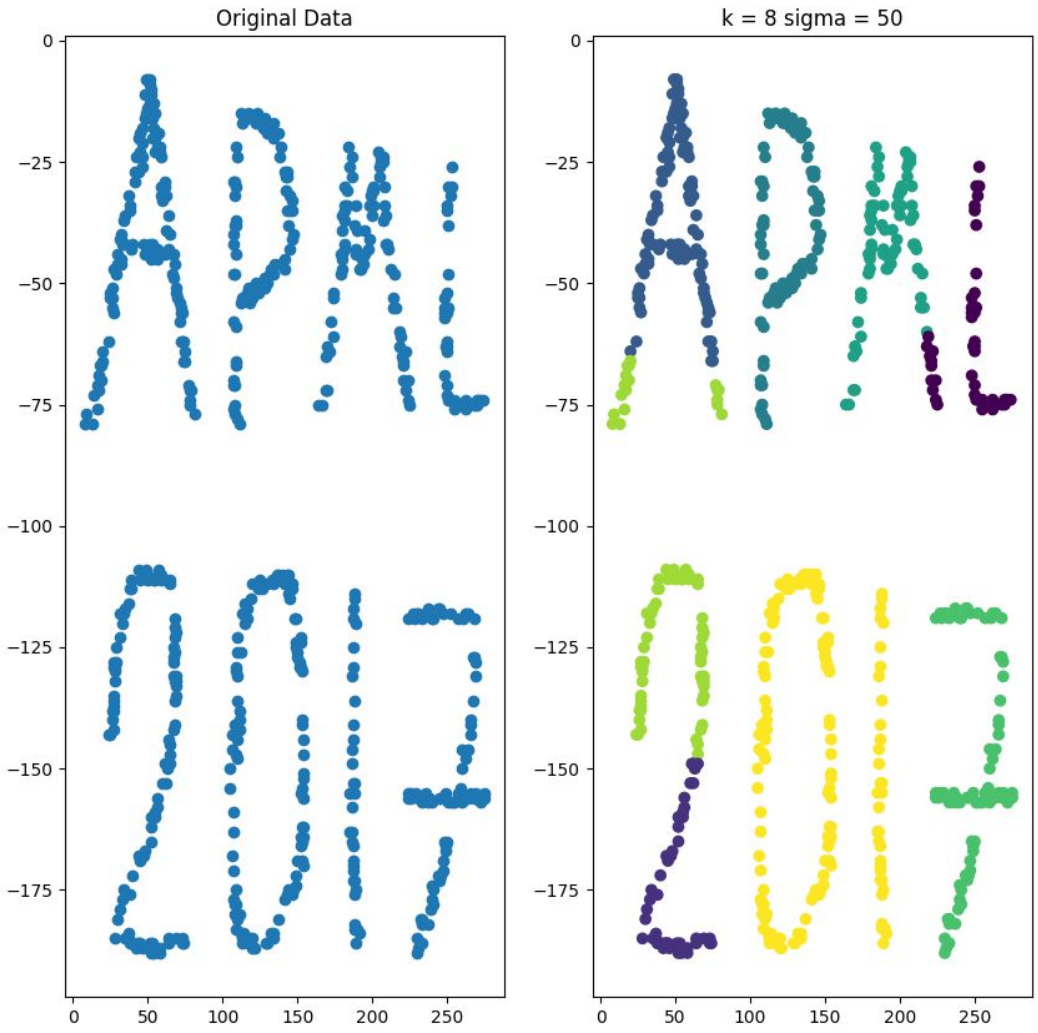


Also with too large number of neighbors we get a cluster which doesn’t look optimal.

More rounds with diffrent parameters -heat kernel



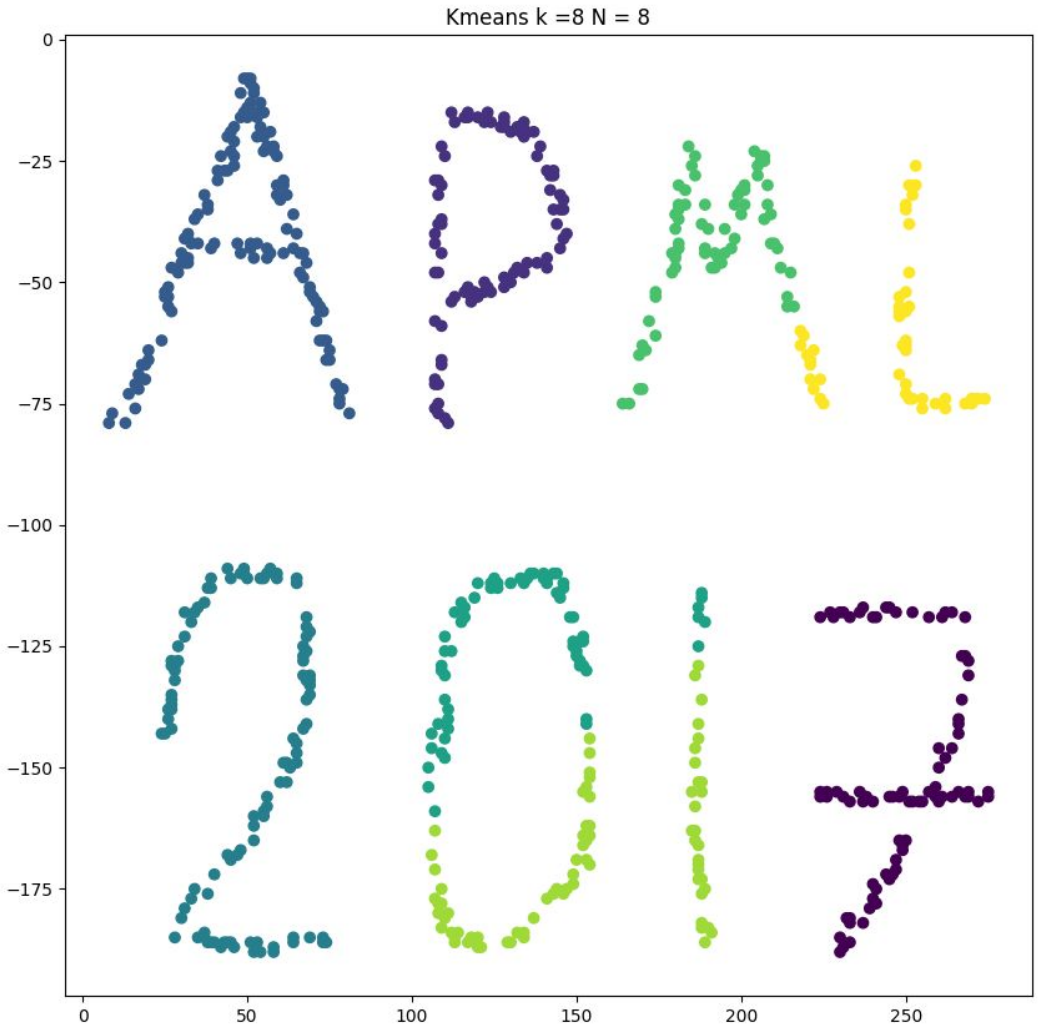
Lower sigma values give a much less good result, you don't really get the wanted result.

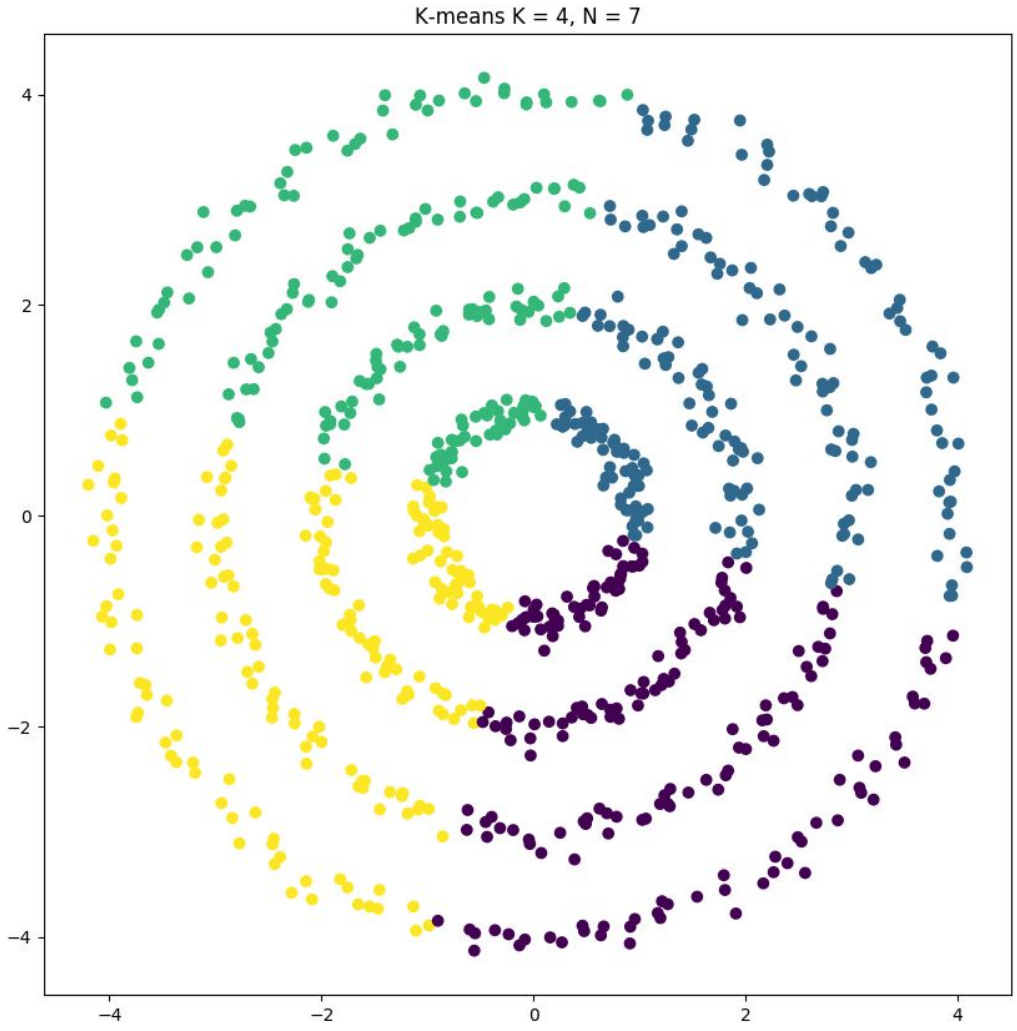


And here we see it that with large sigma we get better results than the 0.1 but still the clusters are not optimal..

Using K means

Next I tried to run the data with K-means using the optimal parameters from spectral clustering



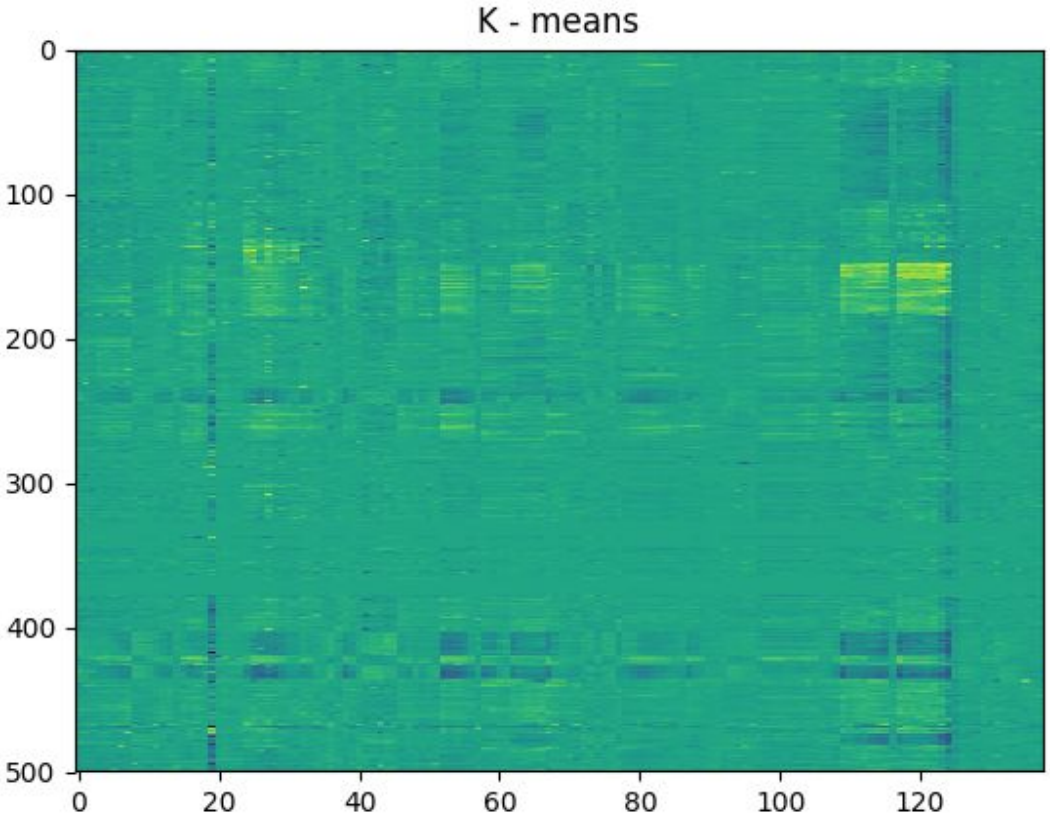


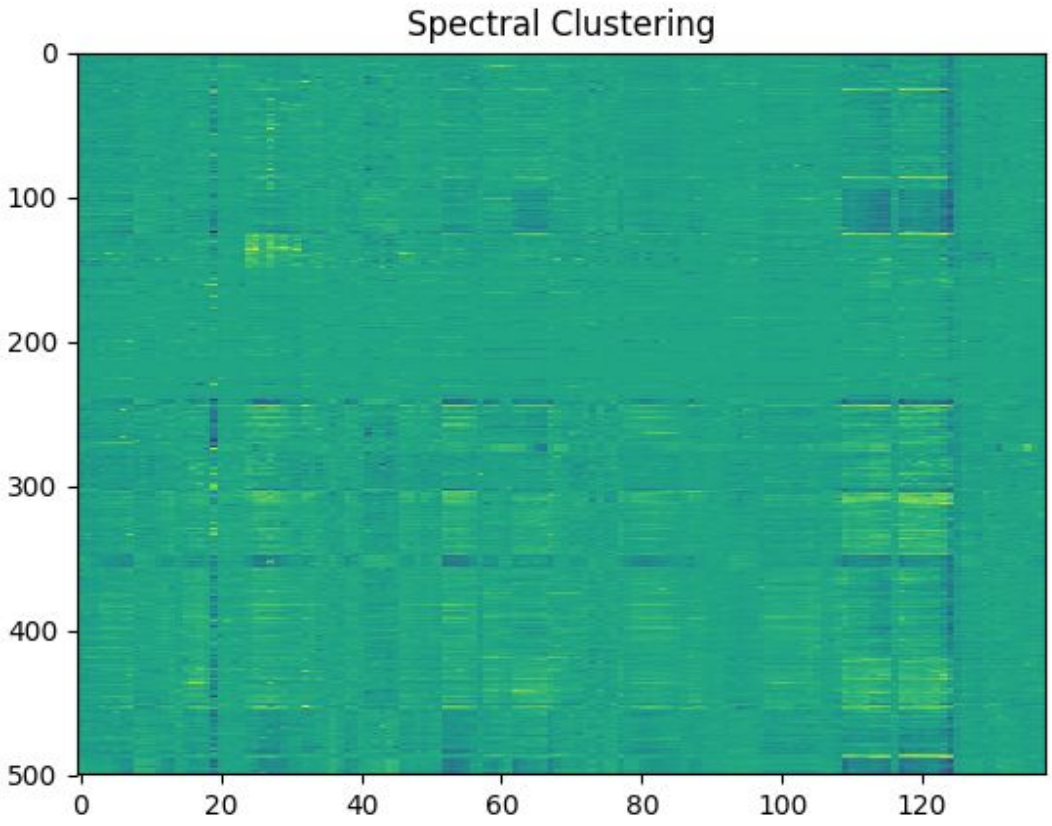
It is clear that k-means gives us results which are much worse than the ones spectral clustering gave. The reason might be that the raw data as we see here it a hard task to cluster, but using the first stage of separation in the spectral clustering where every cluster is being sent as far as it can get (orthogonal directions) from the other cluster. Therefore after this separation perfuming k-means will do a good job, and not as in this example where we only apply k-means.

`run_on_microarray`

MicroArray data

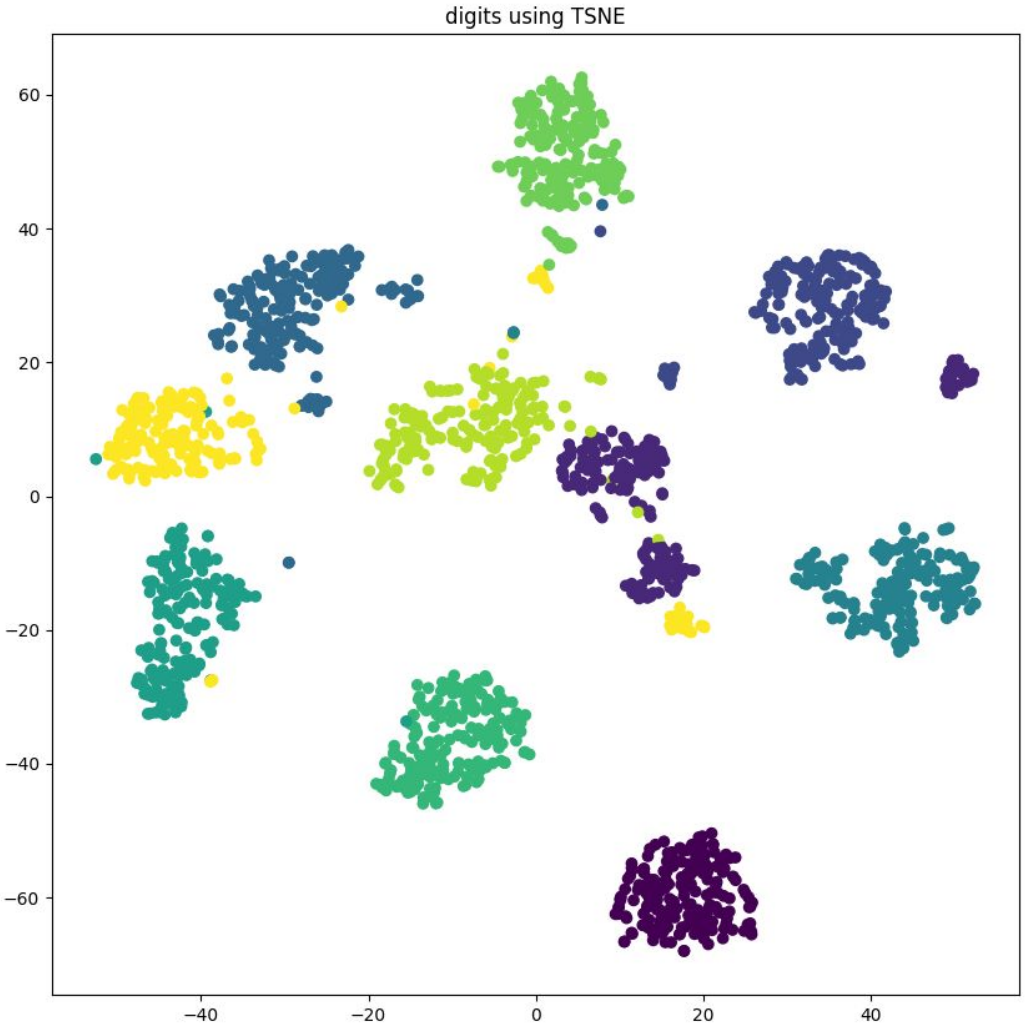
In this part we were asked to use k-means and the spectral clustering with microarray data, I tried to use all the data set by my computer had hard time computing it all(so i sampled 500 values randomly.



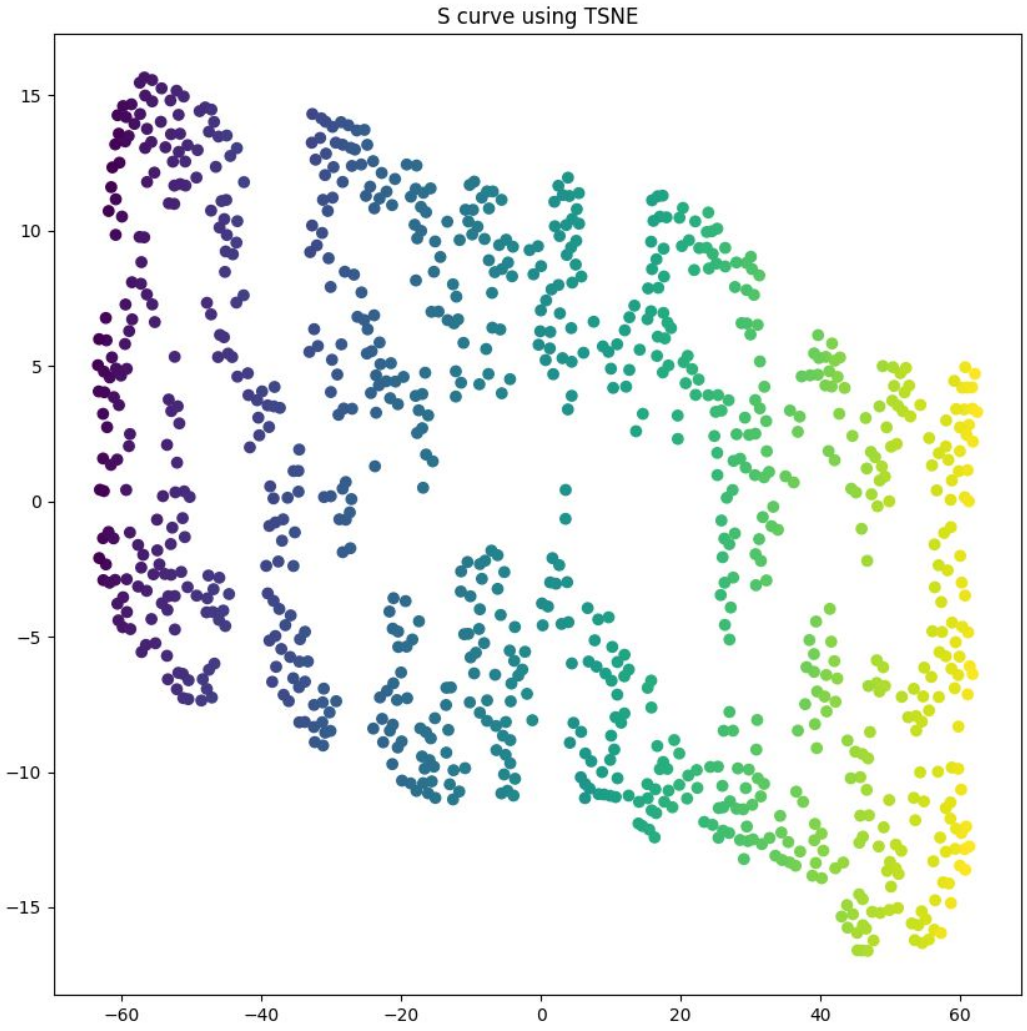


We can see clusters in both of the algorithms, they are also pretty similar. I used here k =30 .
Just to give an example in the spectral clustering we can see group around 200 and 120.

Using TSNE



Using TSNE gives nice results on the digits data, you can see that it separated the data into pretty clear clusters.



The second structure I chose is an s curve, which also here you can see good result on the TSNE, it works really good, even better than some of the methods we used in the first exercise.