CSE 454 Assignment 1

Nurettin Cem Dedetas-171044028

December 6, 2020

1 Problem

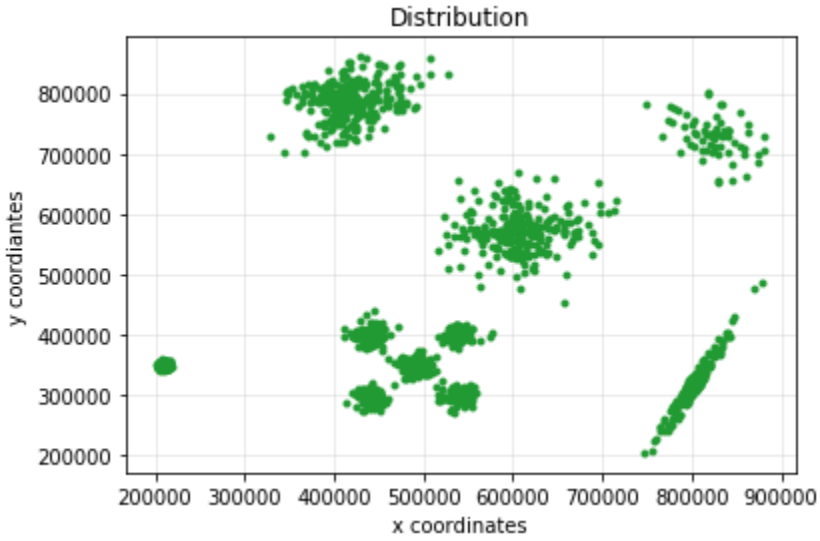
The problem is to "Implement DB-Scan model.". I used the steps in the following section to implement DBSCAN in Python

2 Implementation

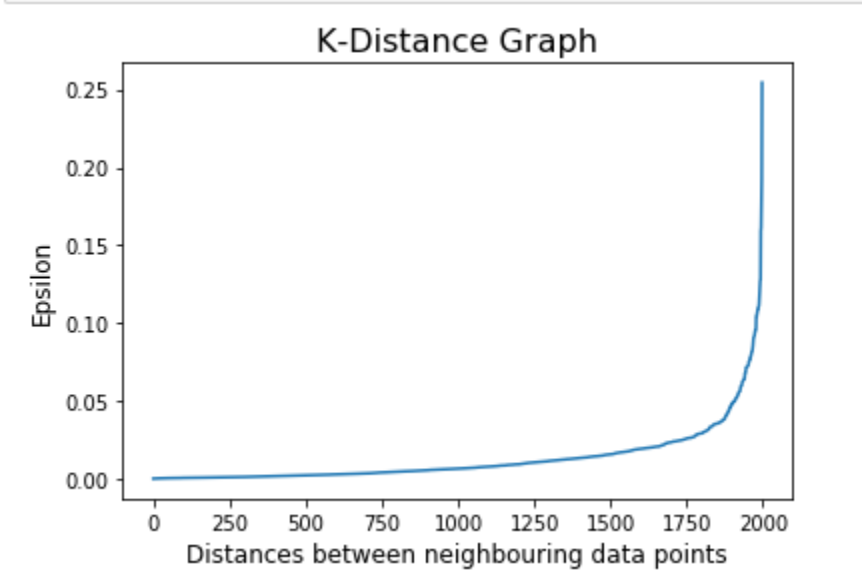
1. Import the example data-set with various clusters.
2. Visualise of raw data.
3. Compute the k-nearest neighbor distances graph.
4. Deduce Epsilon from the graph.
5. Find "min_samples" using trial and error while keeping the outlier count down.
6. Calculate the clusters with previously found parameters

3 Testing

3.1 Raw Data



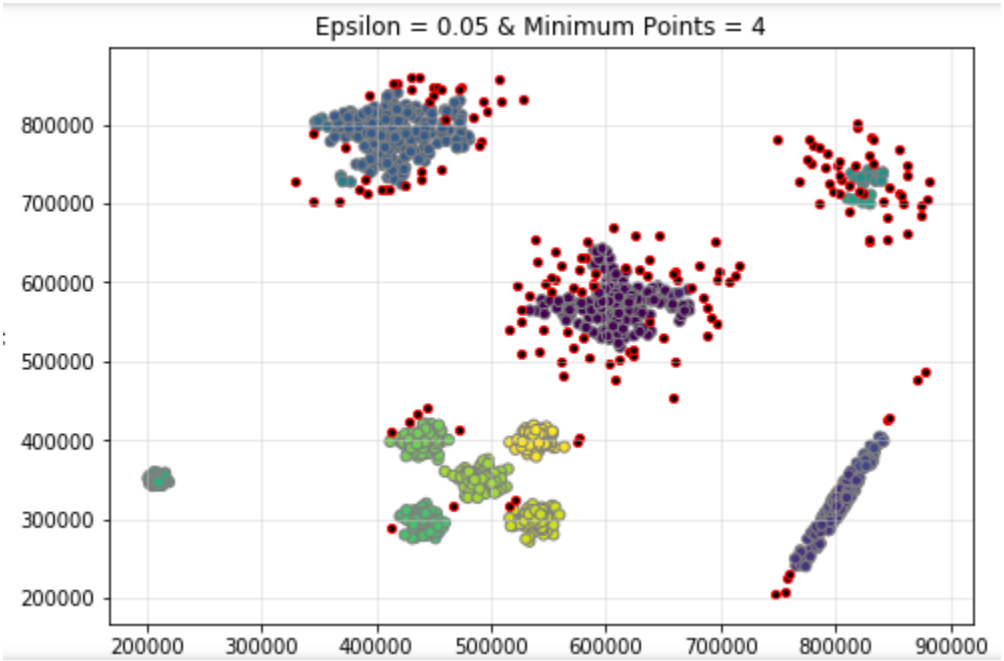
3.2 K-Distance

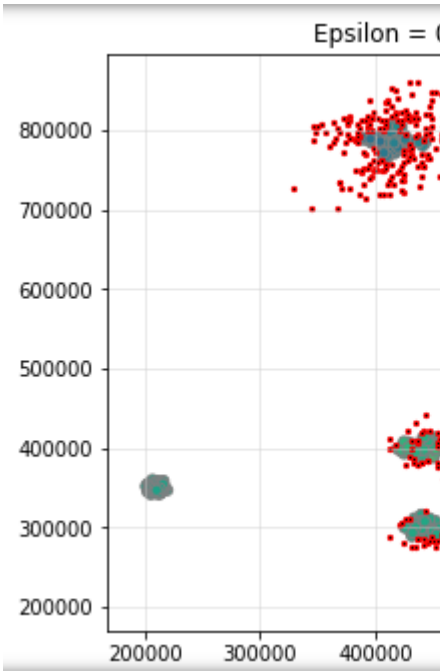


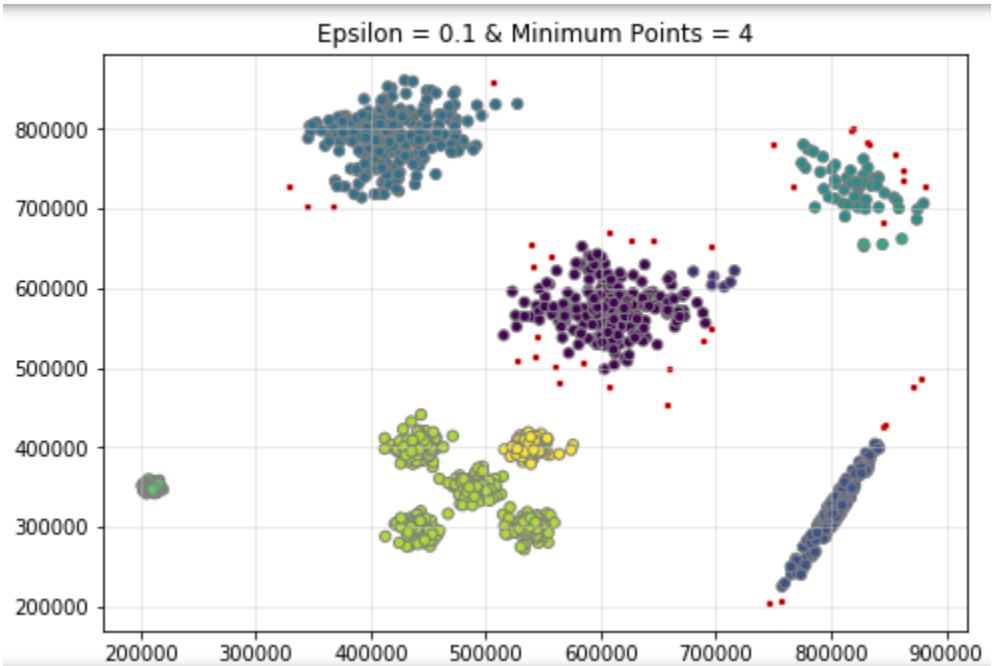
3.3 Difference in Epsilon

The changes in epsilon are ;
If the epsilon is too low the chance of big clusters are very low
The algorithm cannot reach beyond the core cluster because of the low radius

If the epsilon is too big, the defininition between clusters becomes blurry
and two or more clusters can become one.



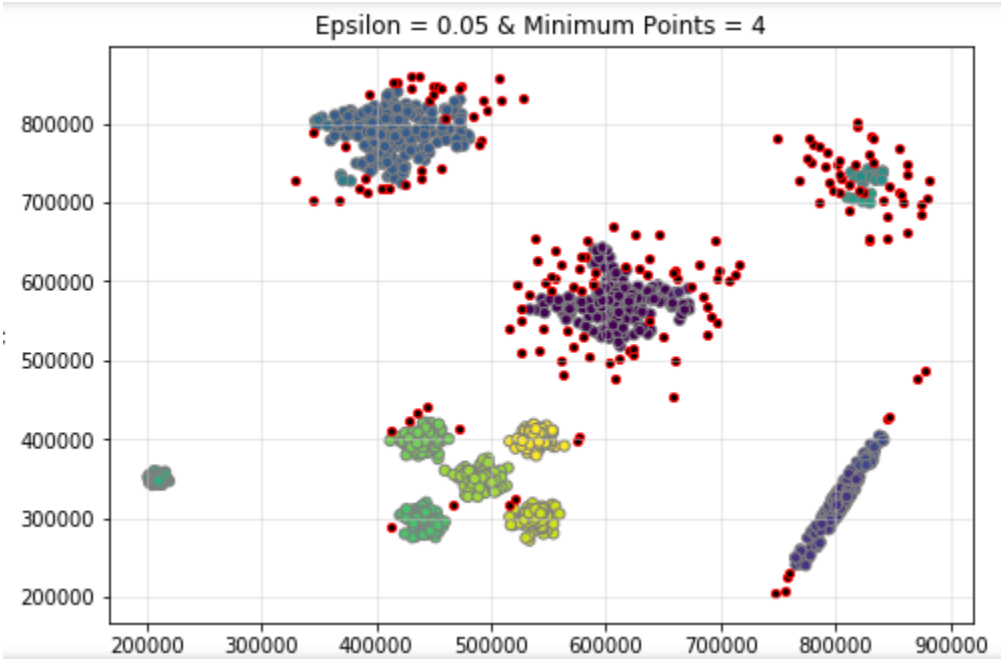


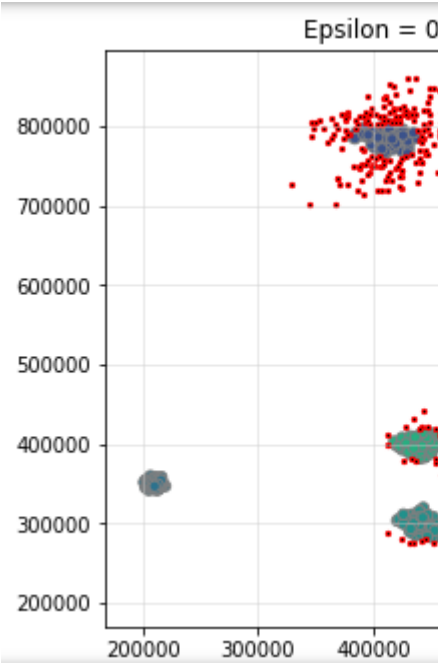


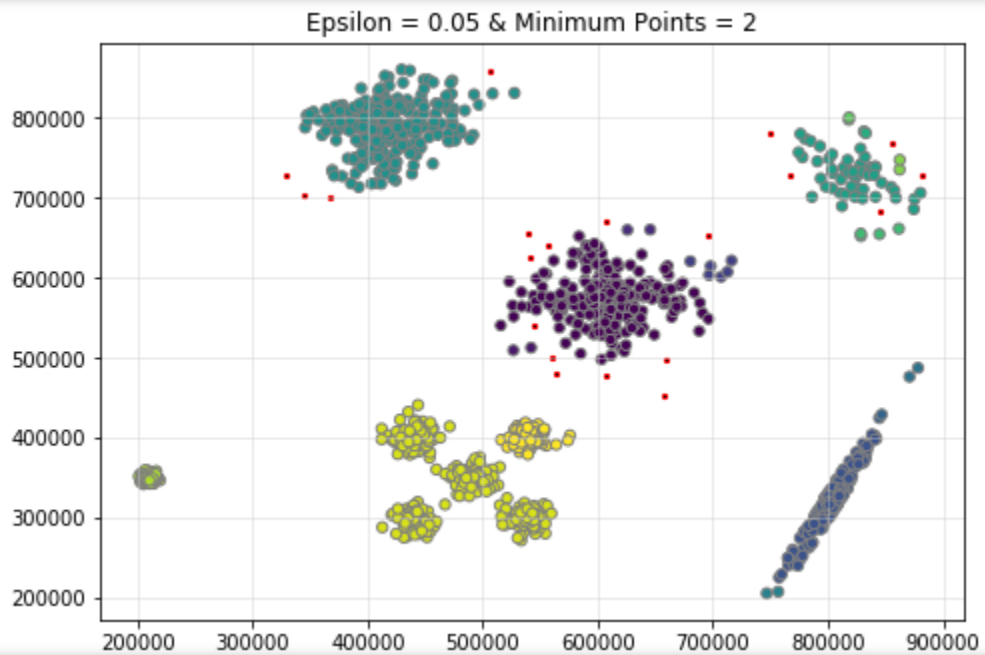
3.4 Difference in Minimum amount of samples

The changes in min sample are ;
If the min samples is too low, number of clusters radically increase
and the outlier number decreases as the outlier become clusters of their own

If the min samples is too big, it becomes harder to be defined as a cluster as there are not enough points to define a core group.






Epsilon = 0.05 & Minimum Points = 2
800000
700000
600000
500000
400000
300000
200000
200000
300000
400000
500000
600000
700000
800000
900000