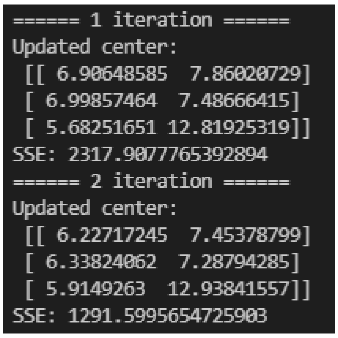
ITSC: ctjlam
SID: 12222973

MSBD5002 Assignment 4 Report

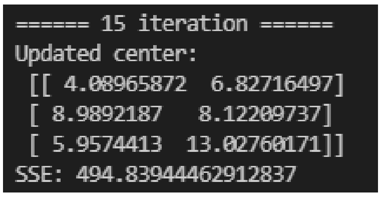
Fuzzy Clustering

SSE and the center for the first two iteration. No hyperparameters has been set for SSE calculation. Initial center was set by the first N (N be number of cluster) input data.

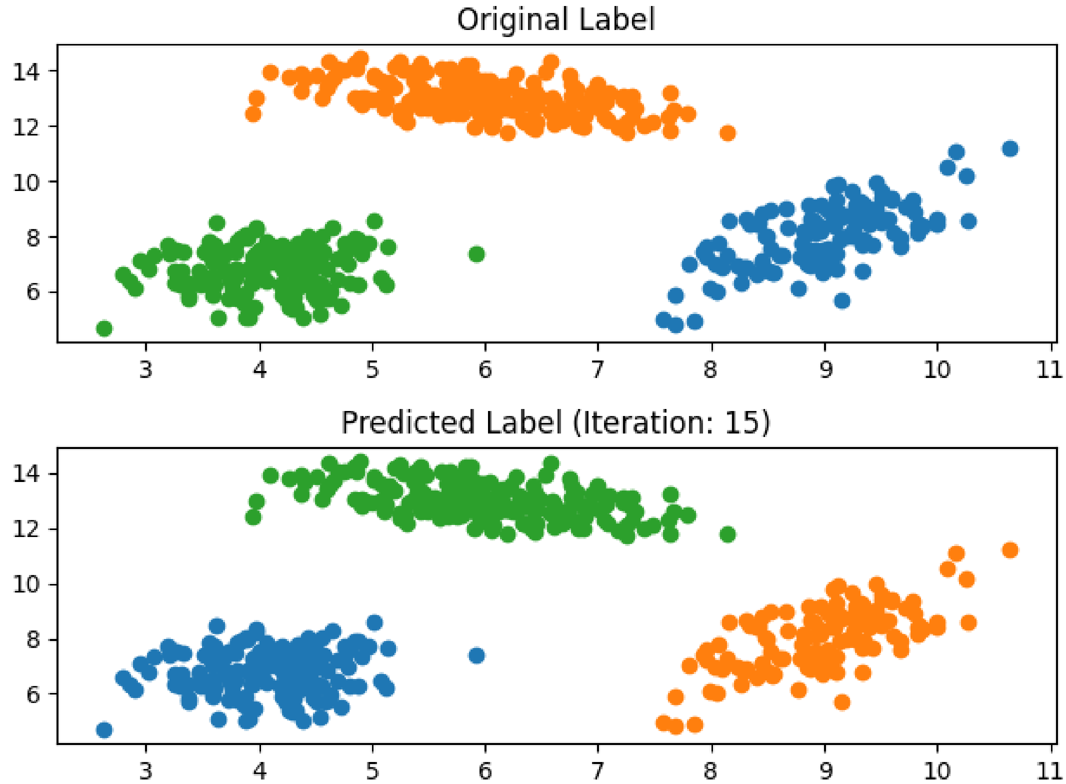


```
====== 1 iteration ======
Updated center:
 [[ 6.90648585  7.86020729]
 [ 6.99857464  7.48666415]
 [ 5.68251651 12.81925319]]
SSE: 2317.9077765392894
====== 2 iteration ======
Updated center:
 [[ 6.22717245  7.45378799]
 [ 6.33824062  7.28794285]
 [ 5.9149263  12.93841557]]
SSE: 1291.5995654725903
```

Final center coverage



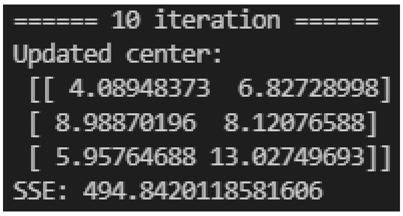
```
====== 15 iteration ======
Updated center:
 [[ 4.08965872  6.82716497]
 [ 8.9892187   8.12209737]
 [ 5.9574413  13.02760171]]
SSE: 494.83944462912837
```



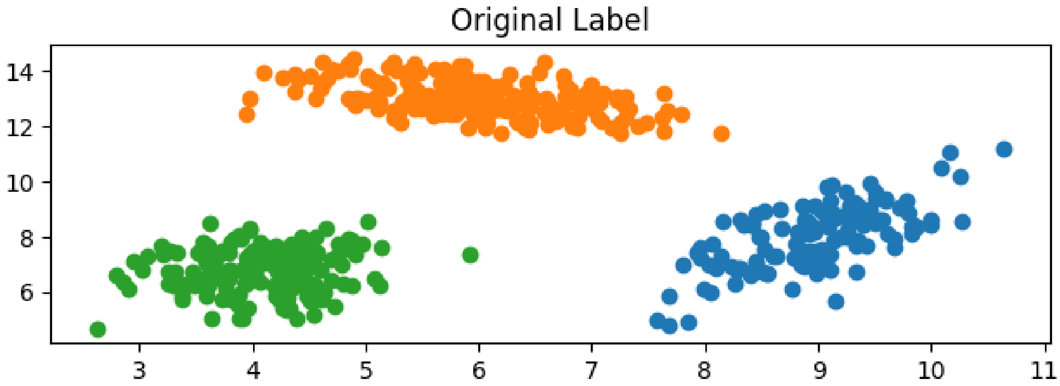
Original Label

Predicted Label (Iteration: 15)

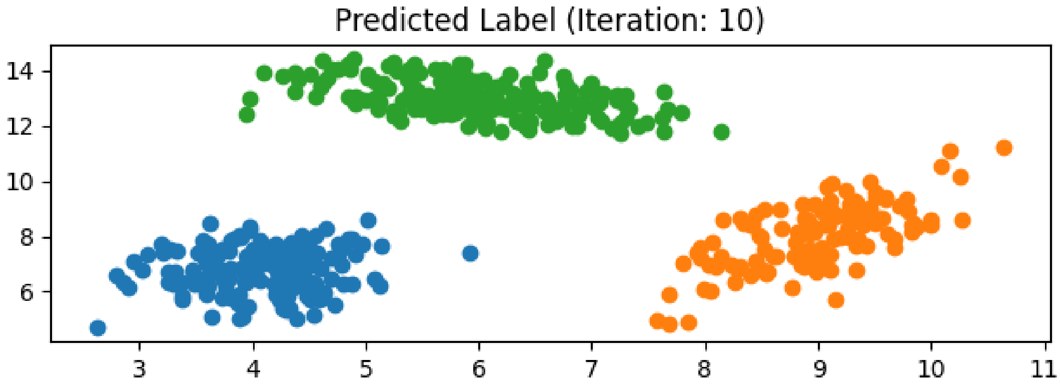
The result is the same with the original label. I have taken 15 iteration for moving, but for the small change in SSE from 10 iteration. As we can see in the graph below, the result is still the same. But we can see for the SSE, and the center, it still have modification for a more precise classification.



```
====== 10 iteration ======
Updated center:
 [[ 4.08948373  6.82728998]
 [ 8.98870196  8.12076588]
 [ 5.95764688 13.02749693]]
SSE: 494.8420118581606
```



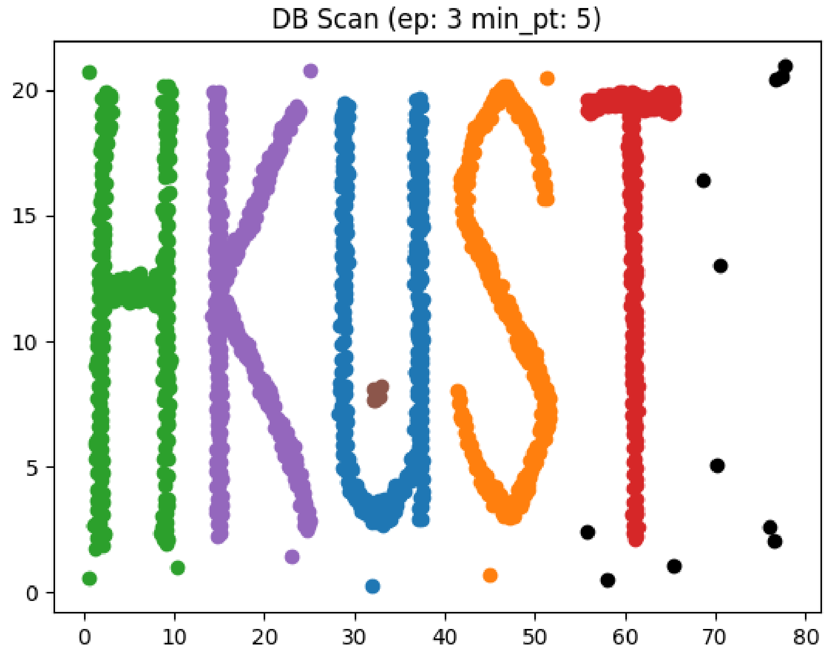
Original Label



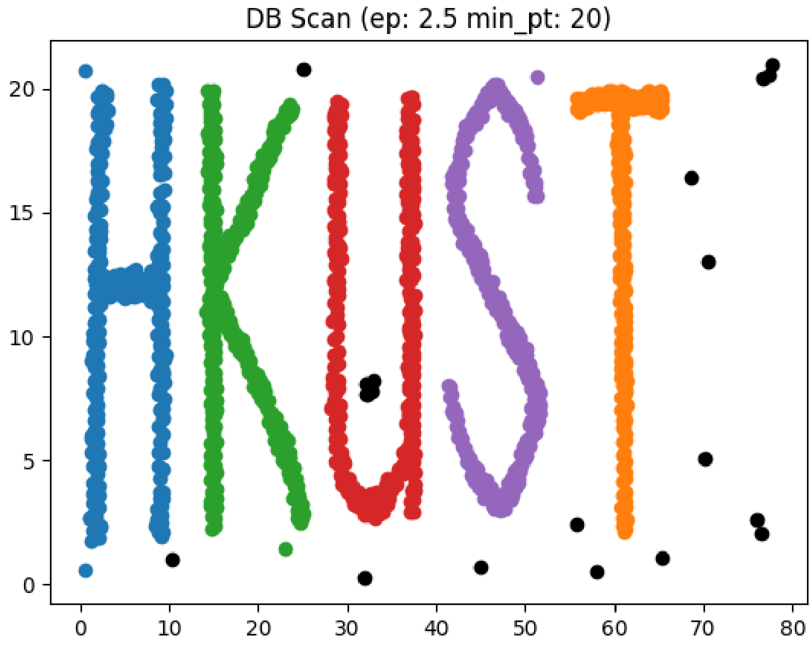
Predicted Label (Iteration: 10)

DBScan

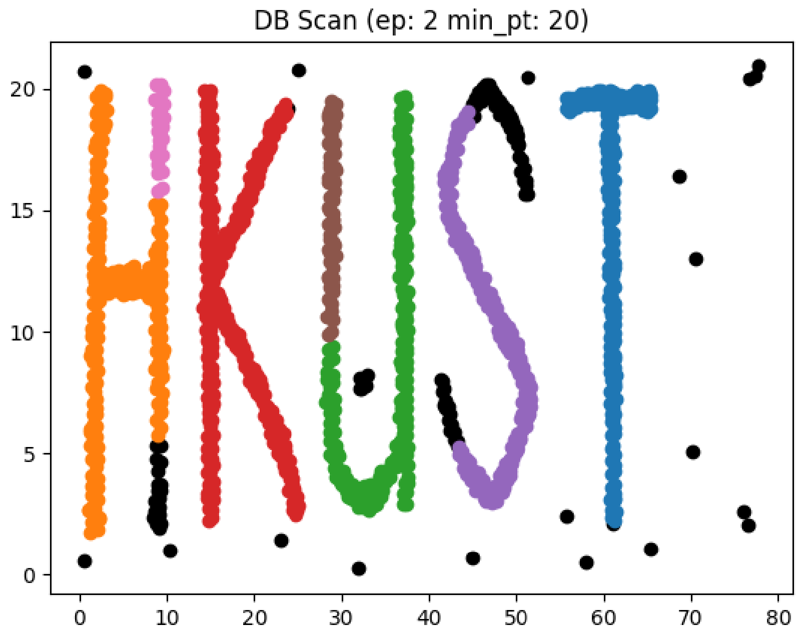
Black will be the outlier points. The parameters are written in the title. All the result list in below.

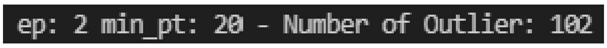


```
ep: 3 min_pt: 5 - Number of Outlier: 11
```

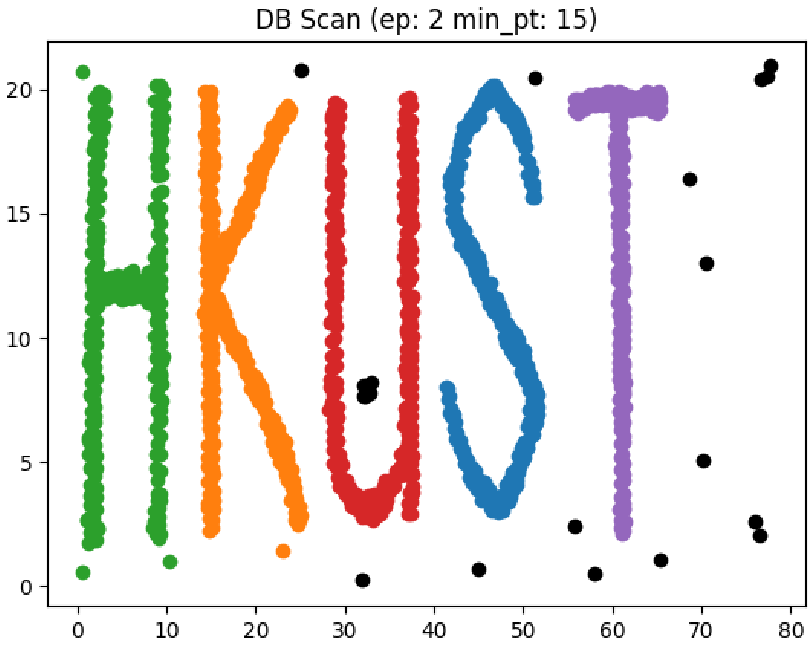


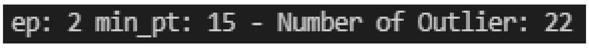
```
ep: 2.5 min_pt: 20 - Number of Outlier: 22
```





```
ep: 2 min_pt: 20 - Number of Outlier: 102
```

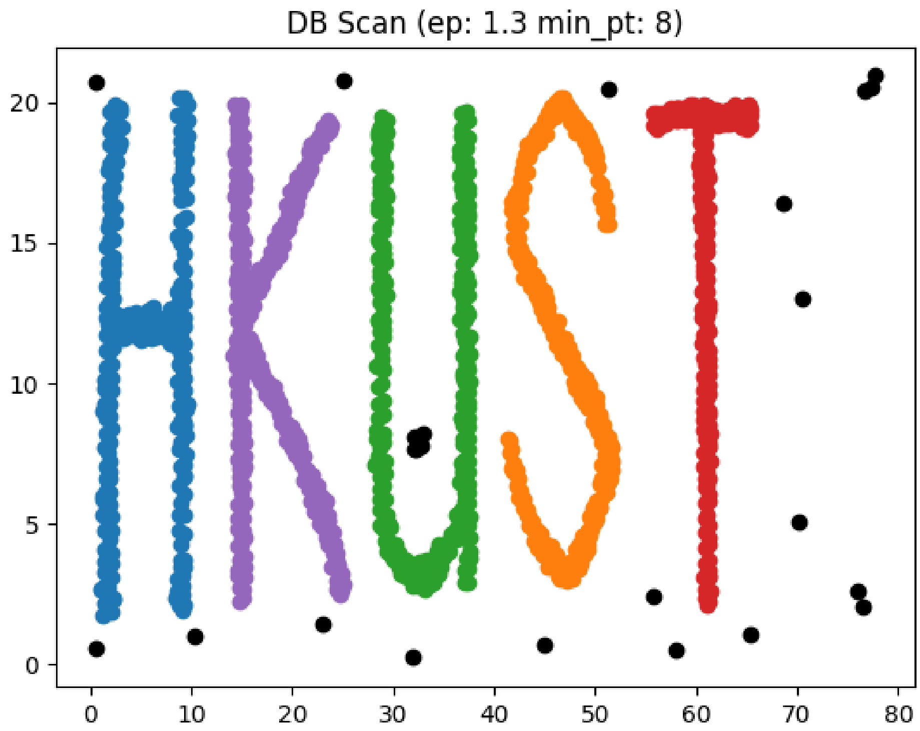


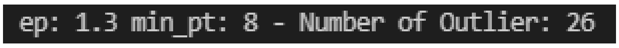


```
ep: 2 min_pt: 15 - Number of Outlier: 22
```

The best result is the below, as all the points do not contain in UST be the outlier.

Epsilon 1.3, with minimum point 8.



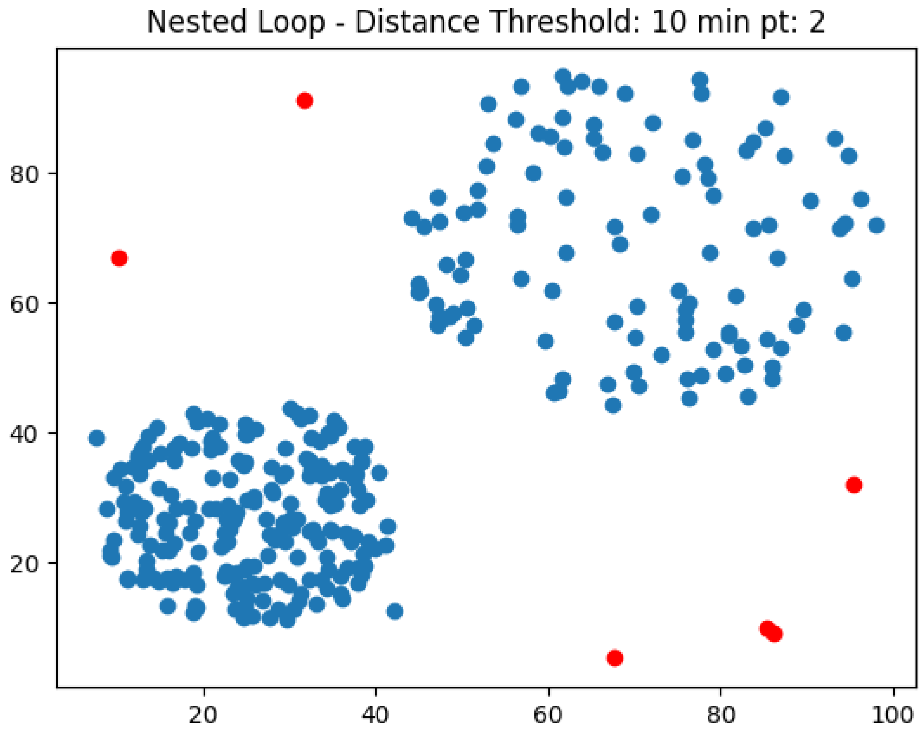


```
ep: 1.3 min_pt: 8 - Number of Outlier: 26
```

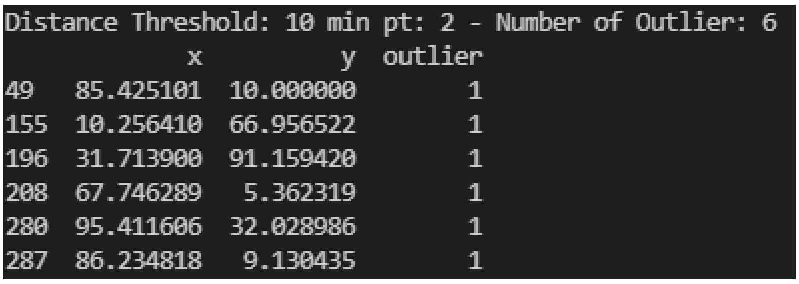
To conclude that, as most of the above setting (list in the specification), outlier point in H and K has been clustered into the cluster, therefore epsilon has been decreased such that it although it is near, it will not belongs to cluster. And we can see large min_pt will affect original cluster (see ep:2, min_pt: 20), and with min_pt 5 has a better result without affecting original cluster. Therefore, min_pt has been kept as small as possible. Therefore, setting ep:1.3, and min_pt: 8 comes after a numerous round of testing.

Nested Loop Outlier Detection

There is two parameter, distance threshold, and minimum numbers of neighbors of the point should have. As we have to count the numbers of points within the specified distance, such that the data does not fulfill the constraint will treat as outlier. It set to 10, and 2 respectively.



Numbers of outliers and its coordinates as follow.



```
Distance Threshold: 10 min pt: 2 - Number of Outlier: 6
             x          y  outlier
49   85.425101  10.000000        1
155  10.256410  66.956522        1
196  31.713900  91.159420        1
208  67.746289   5.362319        1
280  95.411606  32.028986        1
287  86.234818   9.130435        1
```