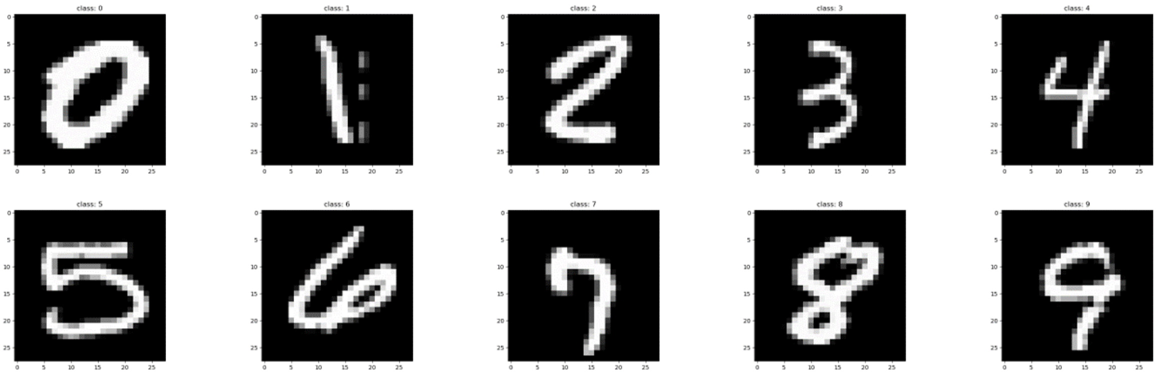
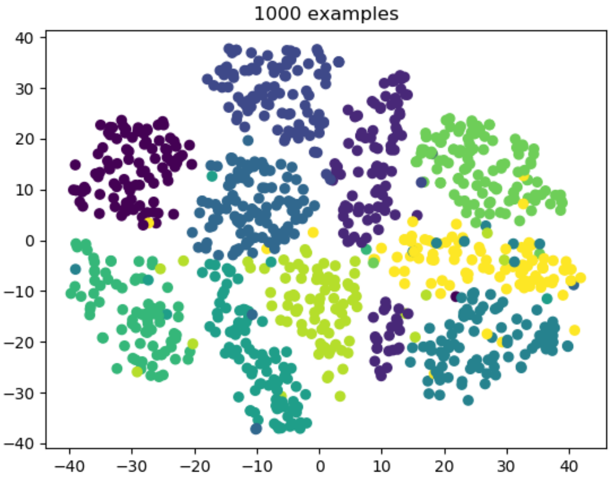
CS57300: Assignment 5

1. exploration.py output:





2. kmeans.py output:

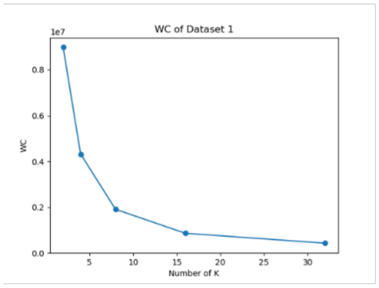
2.1:

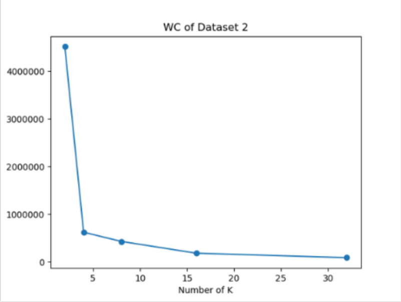
WC-SSD: 1433452.184

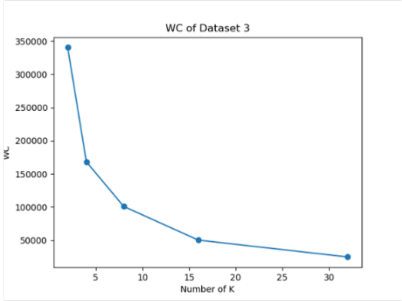
SC: 0.711

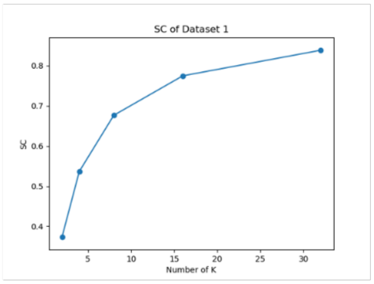
NMI: 0.355

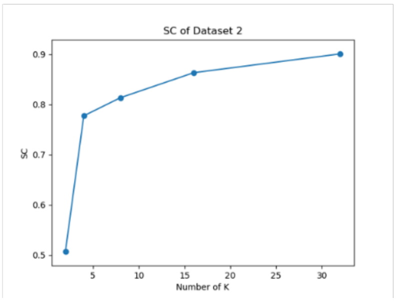
2.2.1:

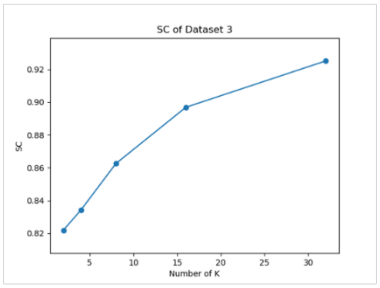












2.2.2

I also wrote a function to calculate the elbow for each dataset and use that results instead of considering SC’s maximum position. The results I got are below:

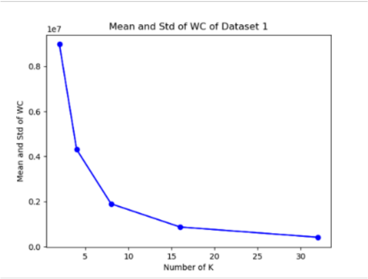
Chosen K for Dataset1 using K-means: 8

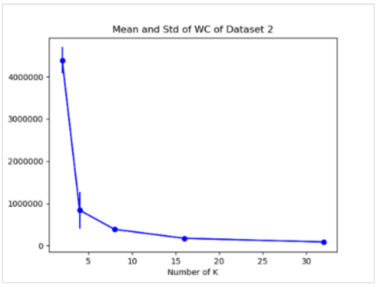
Chosen K for Dataset2 using K-means: 4

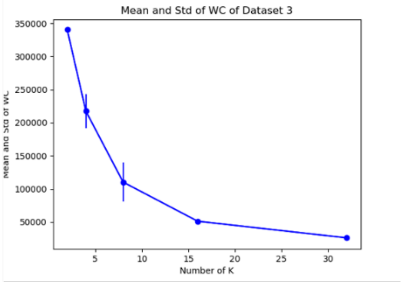
Chosen K for Dataset3 using K-means: 4

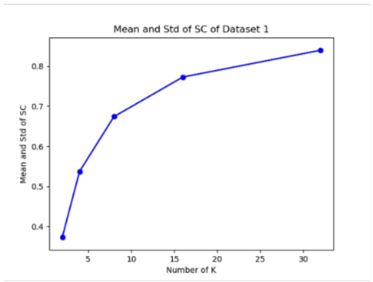
Since the definition of SC given in assignment, the curves of SC are monotonic while K is increasing. So I chose the elbows of WC for K, and you can find that the positions I chose in WC are always the closet ones to origin. Additionally, the positions of K in SC are at the sharp increasing level, which roughly can represent the maximum to some extent.

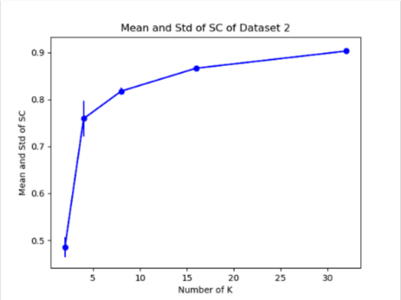
2.2.3

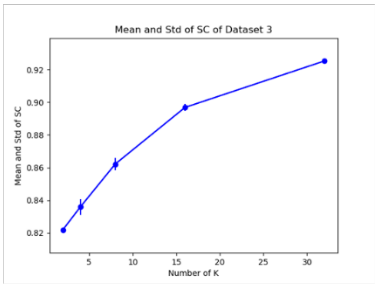






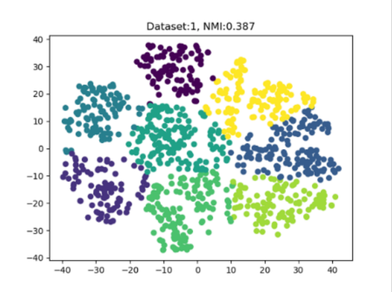


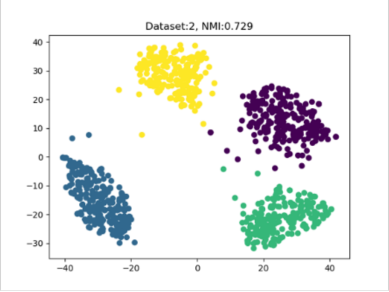


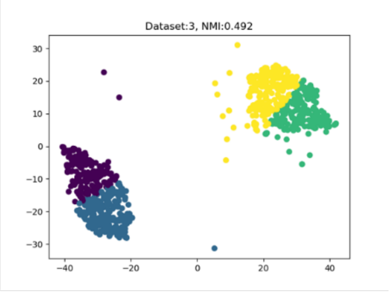


In Dataset 1, we barely can see the standard deviation, which means big dataset can reduce the affect of initial starting conditions. Compared to Dataset 1, in Dataset 2 and 3 we can see some standard deviation when K is small, showing that small dataset and small K will increase the influence of initial starting conditions.

2.2.4

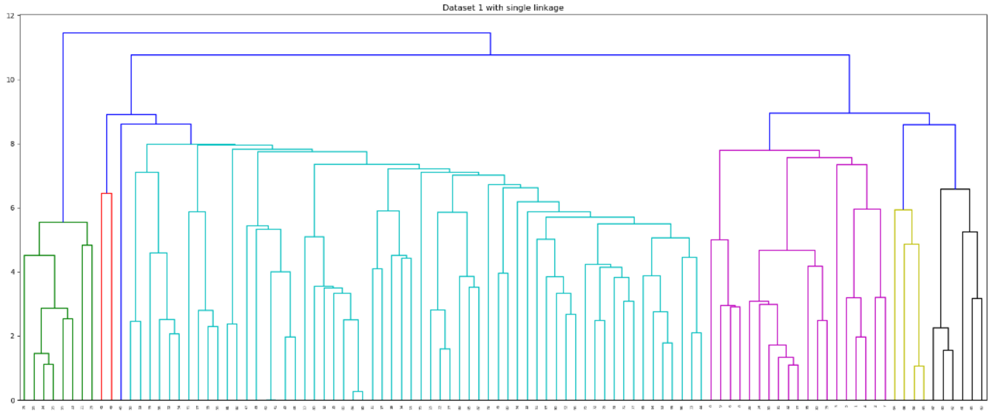




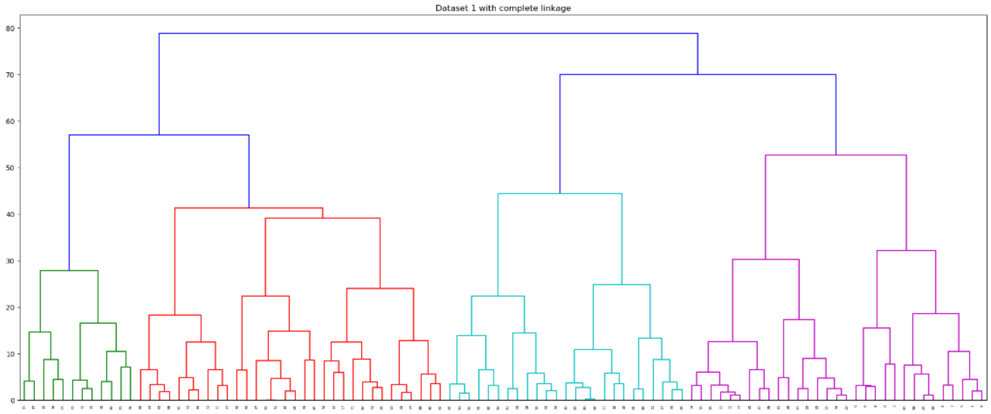


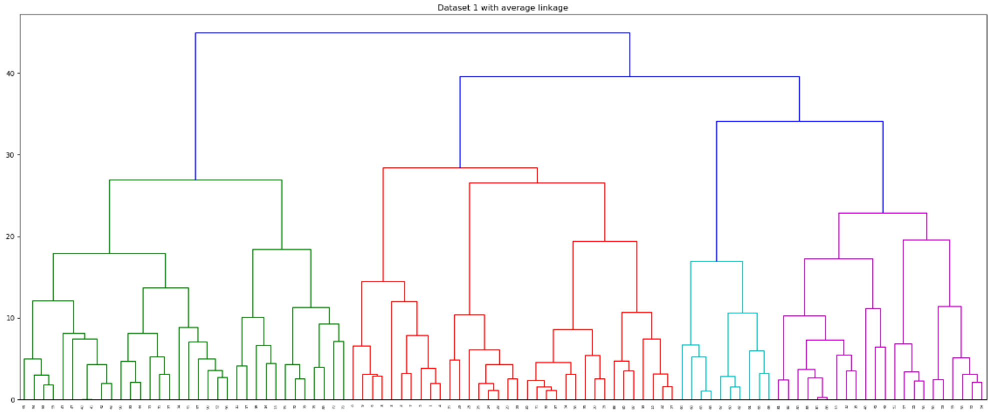
With chosen K, the cluster results are above. In Dataset 1, the result is somehow better than the result of original labels. And in Dataset 2, we can see that the result of K=4 looks well and the NMI is also very high. But in Dataset 3, the K I chose is not really suitable, yet the NMI approaches 0.5.

3.1

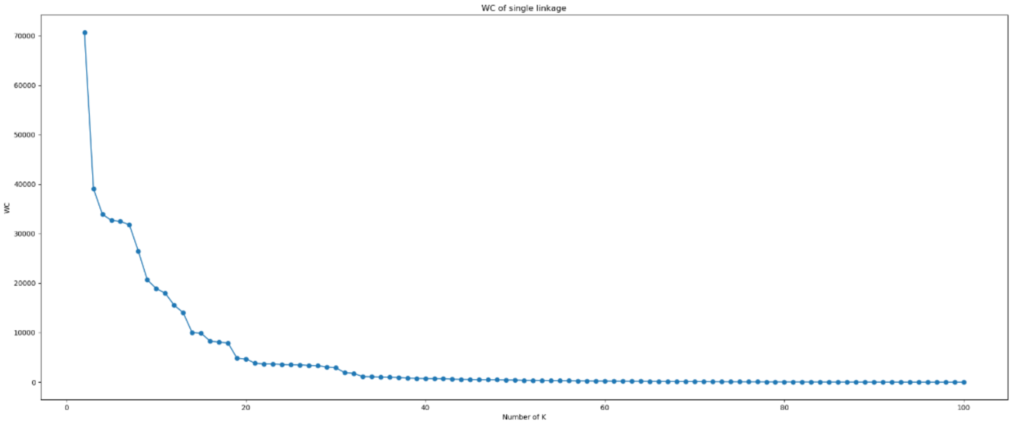

Dataset 1 with single linkage
0
2
4
6
8
10
12

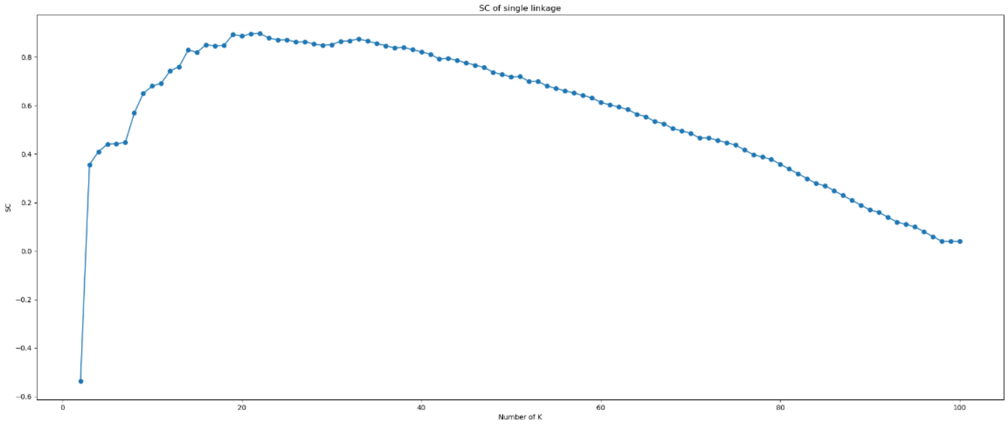
3.2

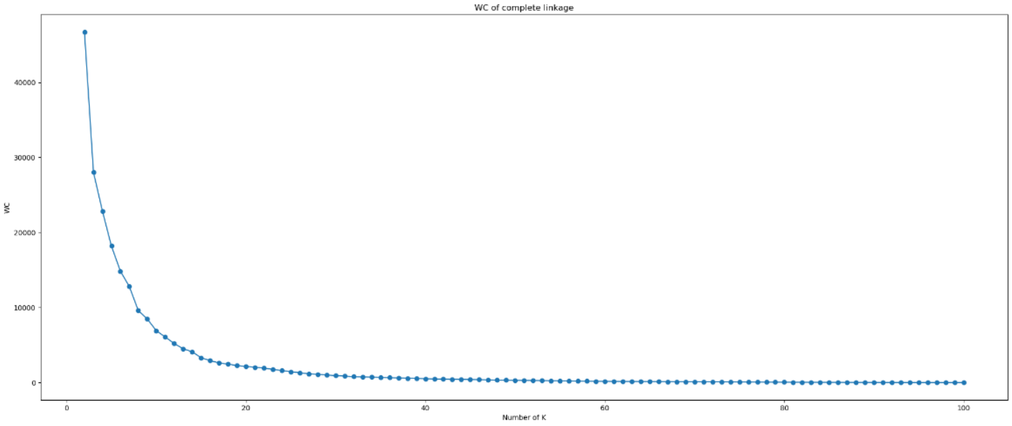

Dataset 1 with complete linkage
0
10
20
30
40
50
60
70
80

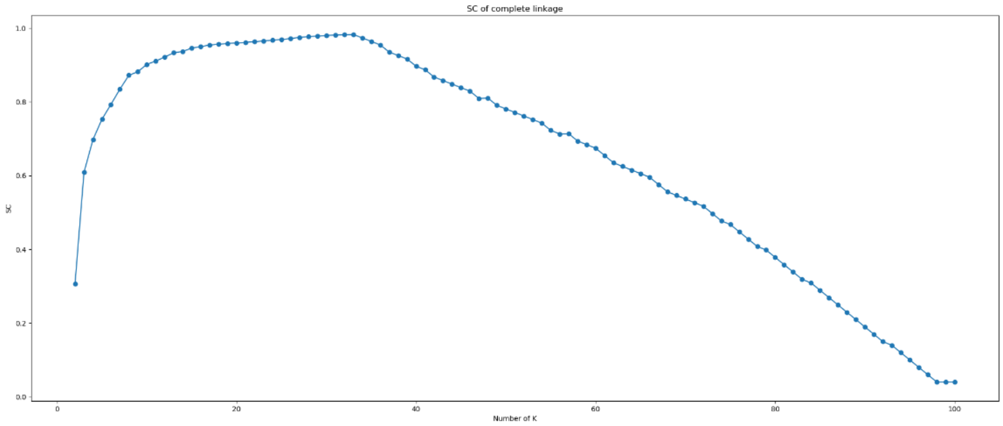

Dataset 1 with average linkage
0
10
20
30
40

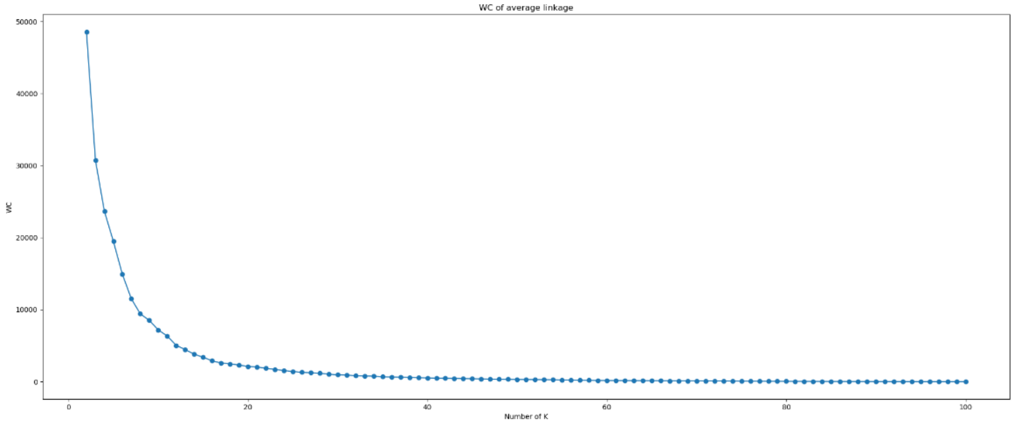
3.3

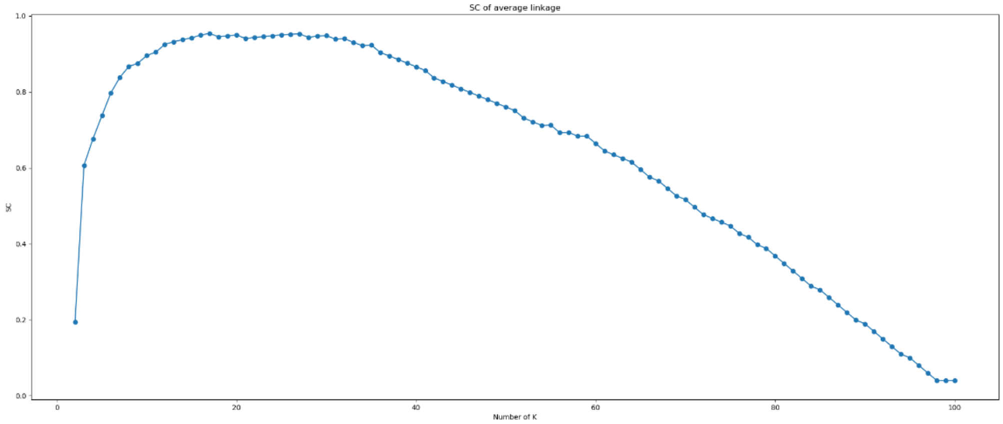

WC of single linkage
WC
Number of K
0
10000
20000
30000
40000
50000
60000
70000
0
20
40
60
80
100


SC of single linkage
SC
Number of K
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
0
20
40
60
80
100


WC of complete linkage
WC
Number of K
0
10000
20000
30000
40000
0
20
40
60
80
100


SC of complete linkage
SC
Number of K
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100


WC of average linkage
WC
Number of K
0
10000
20000
30000
40000
50000
0
20
40
60
80
100


SC of average linkage
SC
Number of K
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100

3.4

In this K, I also used the same function to find the elbows in WC and get the results below:

Chosen K for Dataset1 using single linkage: 19

Chosen K for Dataset1 using complete linkage: 15

Chosen K for Dataset1 using average linkage: 14

In this case, we roughly get the best choice in different linkage methods. Compared to the result of Dataset 1 in section 2, those Ks are closer to 16, which are slightly different from my choice, 8.

3.5

The results of Ks I chose are below:

NMI of 19 for Dataset1 using single linkage: 0.384

NMI of 15 for Dataset1 using complete linkage: 0.381

NMI of 14 for Dataset1 using average linkage: 0.397

All three NMI values are close, but all of them are not as high as ones in section 2.