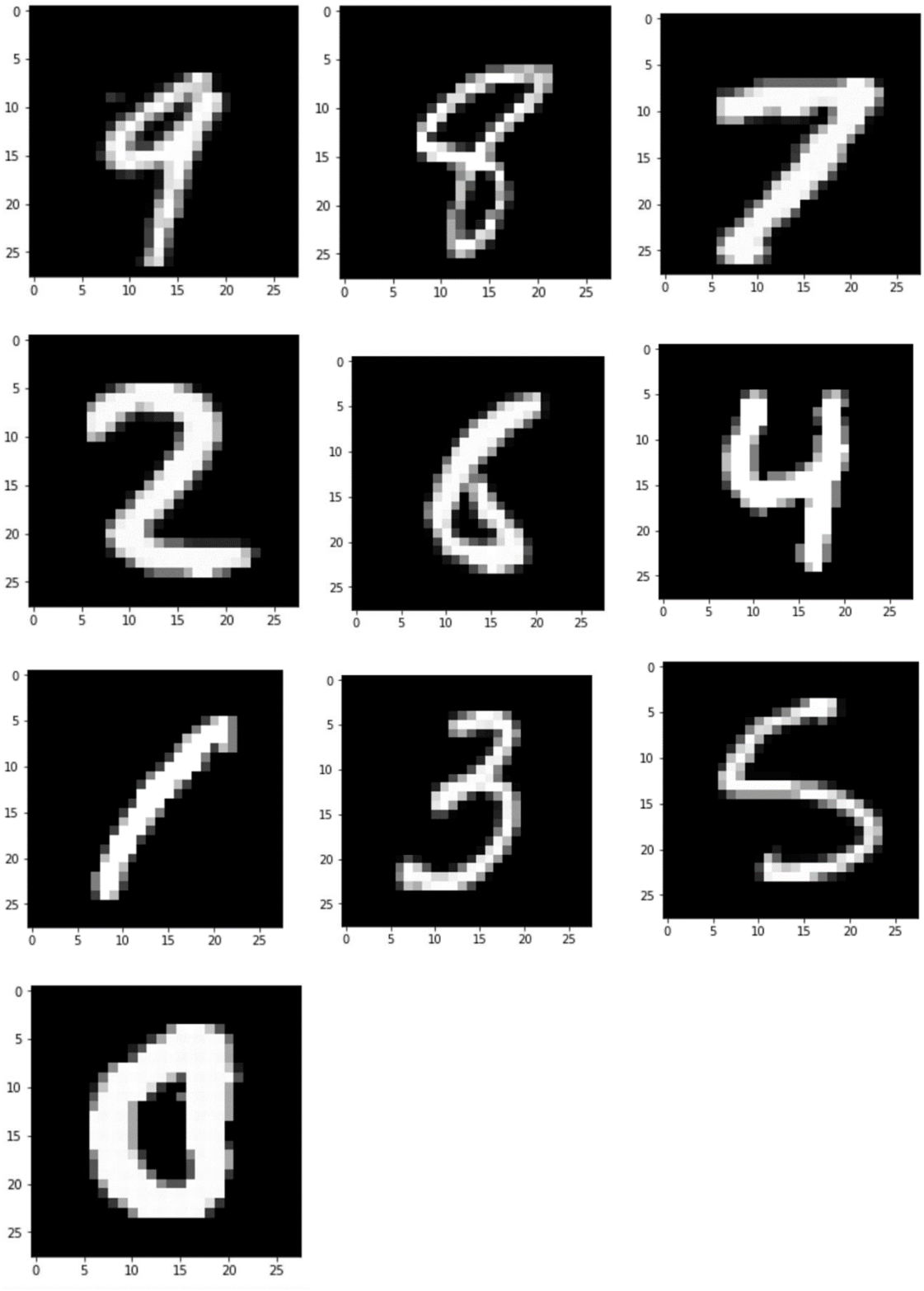
CS 57300 HW5

Hengrui Zhang

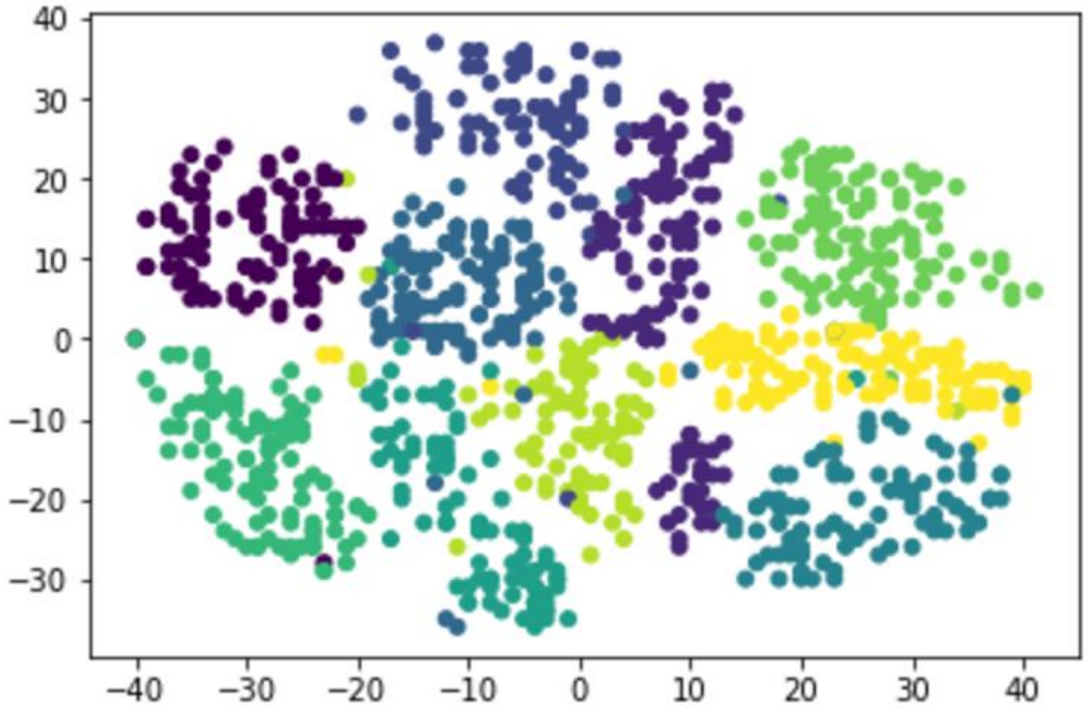
4/29/17

A.Exploration.

1. Below is the random chosen 10 digits from ‘digits-raw.csv’.



2. Below is the visualization of 1000 randomly selected examples in ‘digits-embedding.csv’

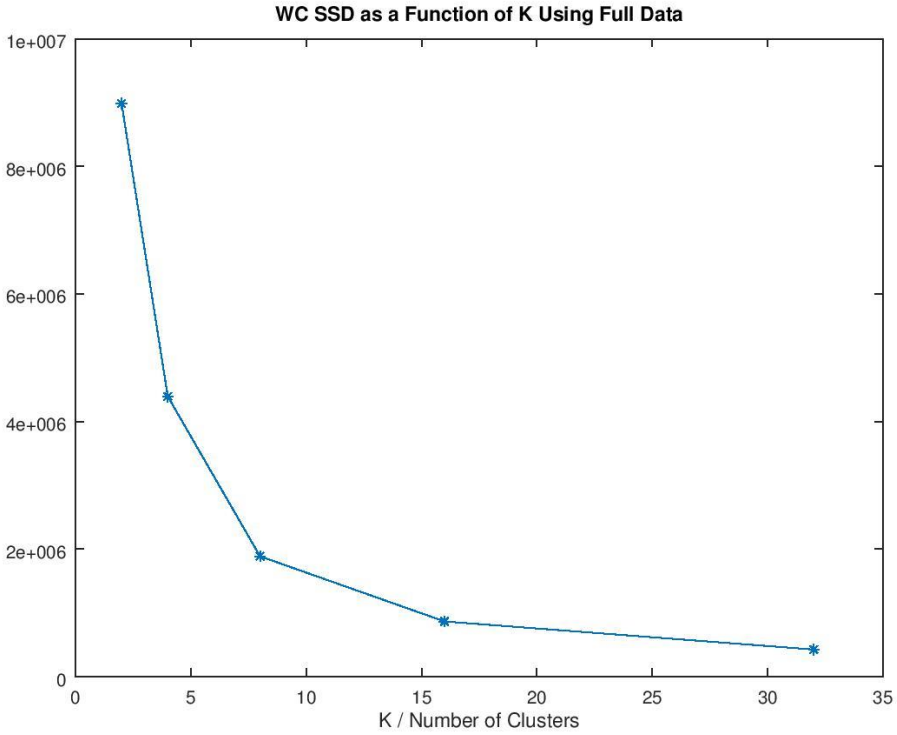


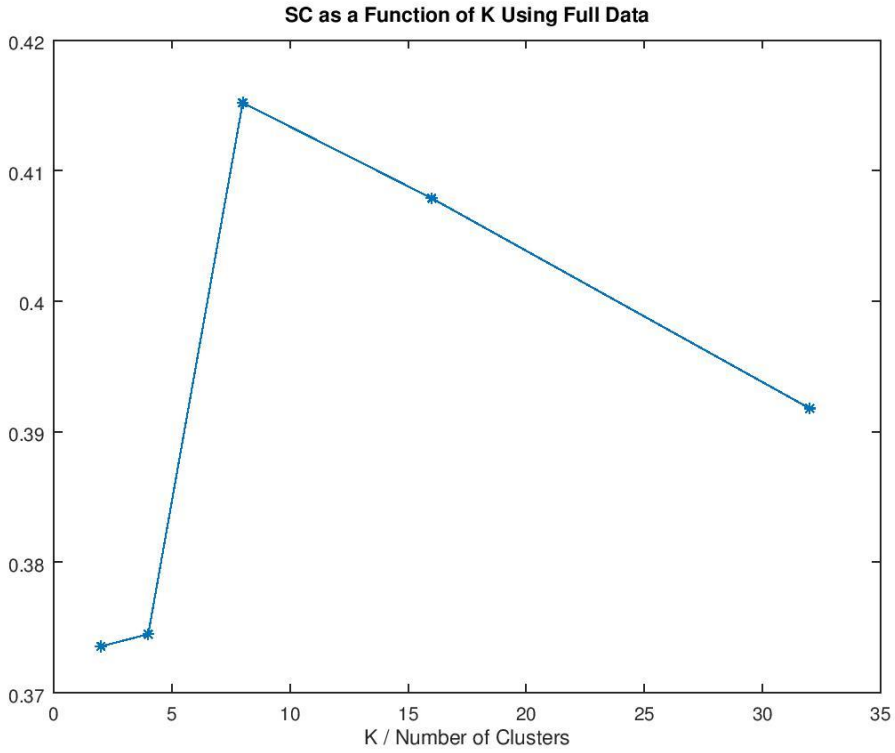
B. Analysis of k-means

1.

(i) Using full data as dataset:

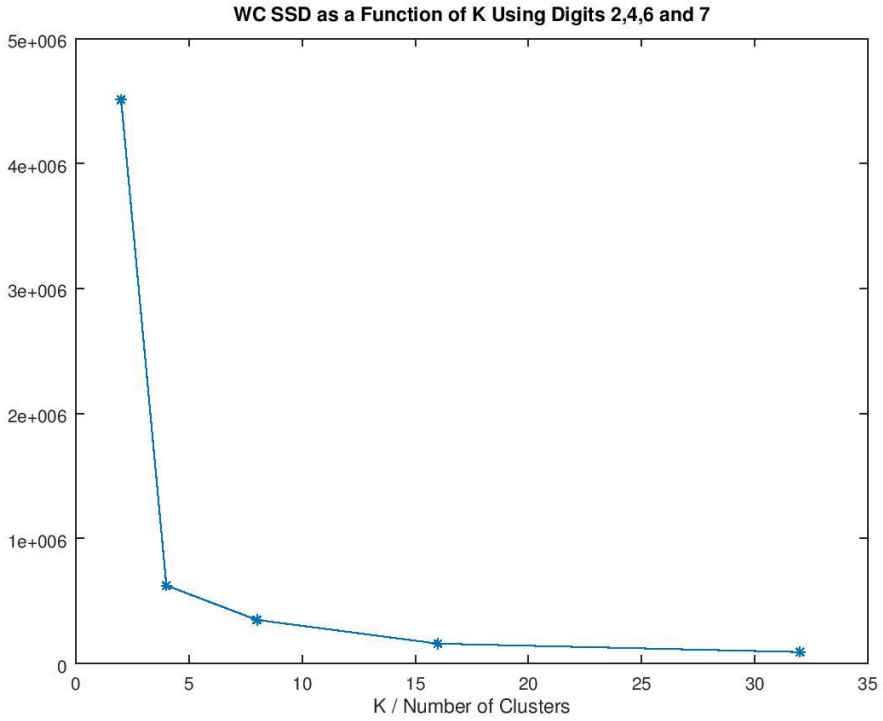
K	2	4	8	16	32
WC SSD	8983899	4393799	1887621	866066	425883
SC	0.37356	0.37448	0.41520	0.4078	0.3917

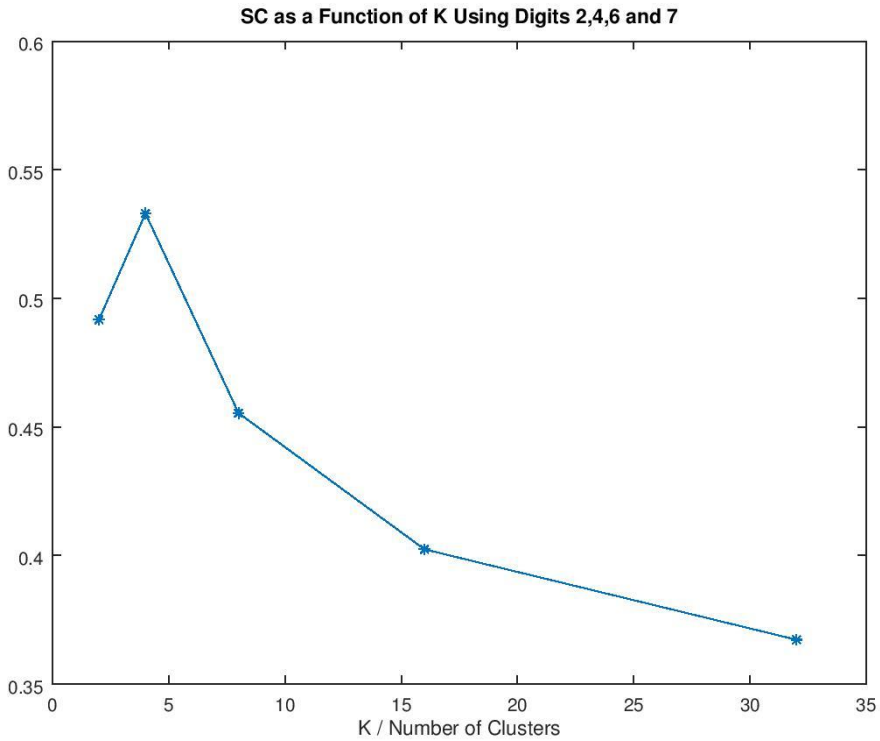

WC SSD as a Function of K Using Full Data
1e+007
8e+006
6e+006
4e+006
2e+006
0
0
5
10
15
20
25
30
35
K / Number of Clusters


SC as a Function of K Using Full Data
0.42
0.41
0.4
0.39
0.38
0.37
0
5
10
15
20
25
30
35
K / Number of Clusters

(ii) Using digits 2,4,6 and 7 as dataset:

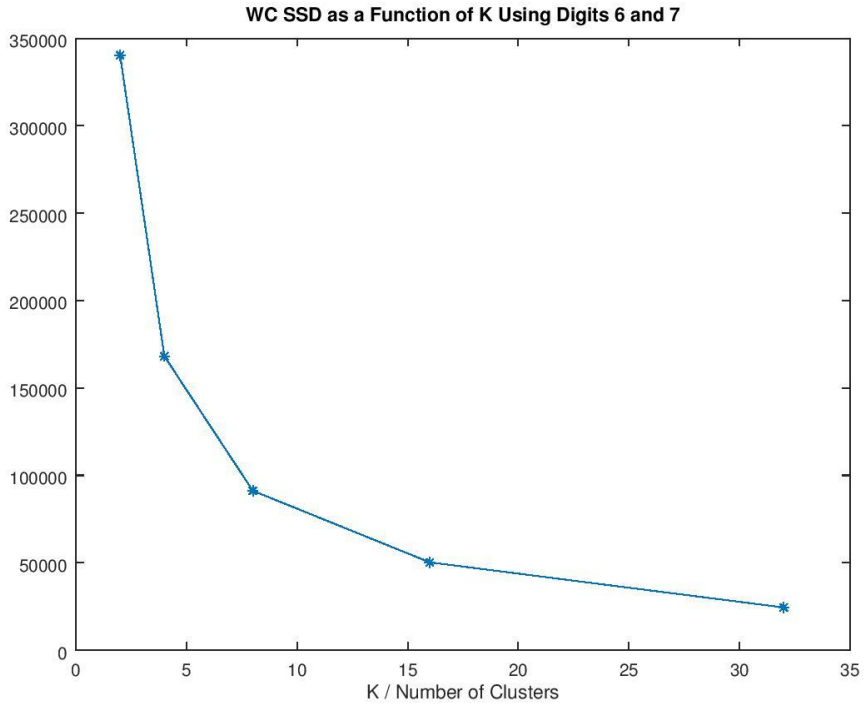
K	2	4	8	16	32
WC SSD	4510451	623865	348349	158332	91034
SC	0.4917	0.53301	0.4554	0.4024	0.36724

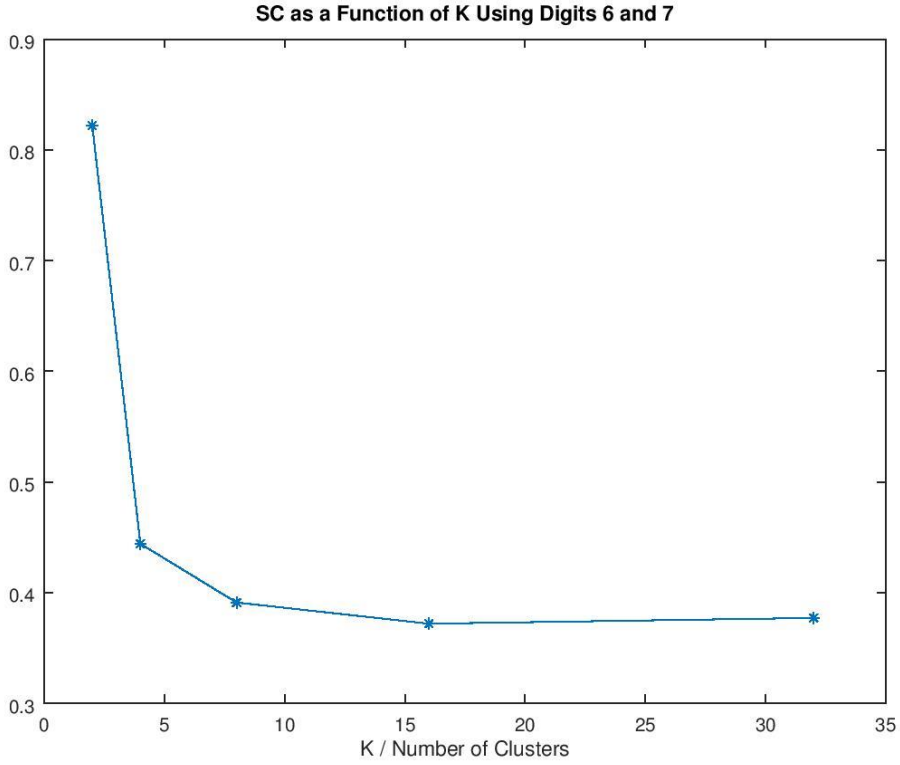




(ii) Using digits 6 and 7 as dataset:

K	2	4	8	16	32
WC SSD	340372	168174	91182	50242	24422
SC	0.821889	0.444182	0.3910377	0.37189	0.377089


WC SSD as a Function of K Using Digits 6 and 7
350000
300000
250000
200000
150000
100000
50000
0
0
5
10
15
20
25
30
35
K / Number of Clusters


SC as a Function of K Using Digits 6 and 7
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0
5
10
15
20
25
30
35
K / Number of Clusters

2.

(i) SC is chosen as the criterion to decide whether the cluster is good or not. Based on the algorithm of WC_SSD, the value of WC_SSD evaluate the distance from each data to their centroids. If each data is more clustered around centroid, the value is small. Hence it can shows whether the cluster is good or not in some aspects. However, if the number of clusters getting larger, the value will become smaller as well. SC is better because SC considers both the separation and cohesion of the distribution of data after clustered. The k-means clustering is the best when K==**8** in full data. The SC value is 0.4152 and WC_SSD value is approximately 1887621.

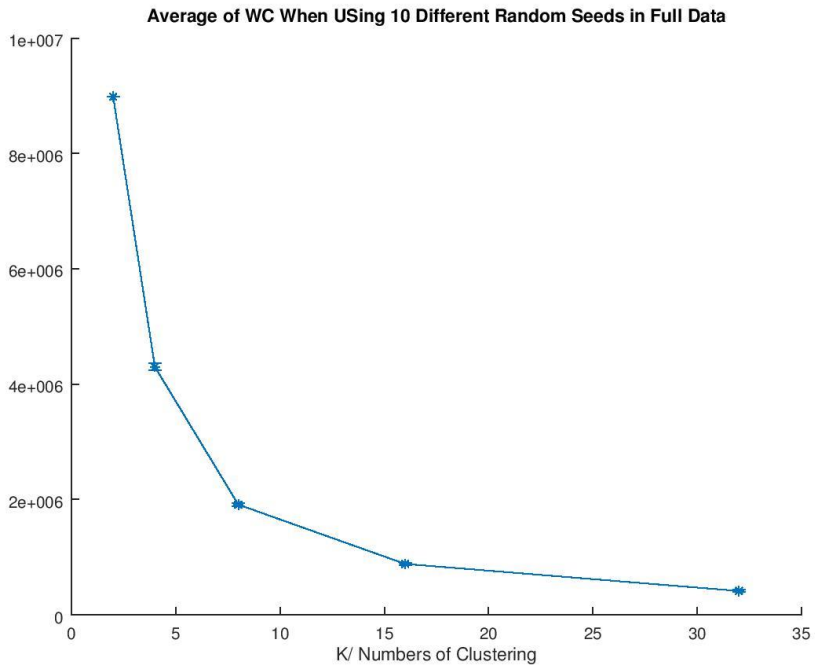
(ii) SC is chosen as the criterion as well when there are only four labels of data. SC shows both the separation and cohesion of the clustering. The k-means clustering of digit 2,4,6 and 7 is the best when K==**4**. The SC value is 0.53301 and WC_SSD value is approximately 623865.

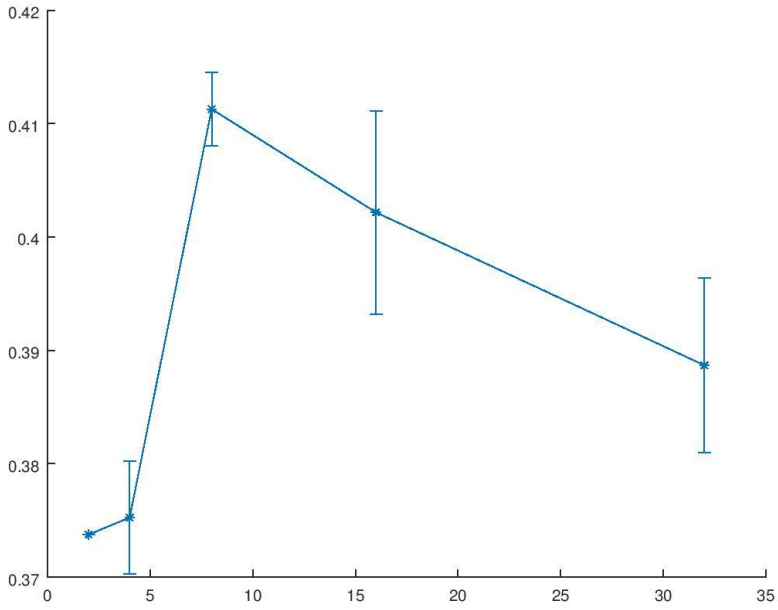
(ii) Since there are only two labels of data in the dataset, we cannot decide which value of k is better by only comparing the value of WC_SSD. So in this case we check WC_SSD and SC together, and notice that when K==**2**, The k-means clustering performs the best. The SC value is 0.821889 and WC_SSD value is approximately 340372.

3.

(i)

K	2	4	8	16	32
WCSSD_avg	8983697	4299584	1911046	882221	413146.876
WCSSD_var	95951.25	3649005649	520296100	279190681	429442729
SC_avg	0.3737460218	0.375249	0.411275	0.402171	0.3886810
SC_var	0	0.0000246	0.0000105287	0.0000803353	0.000059583

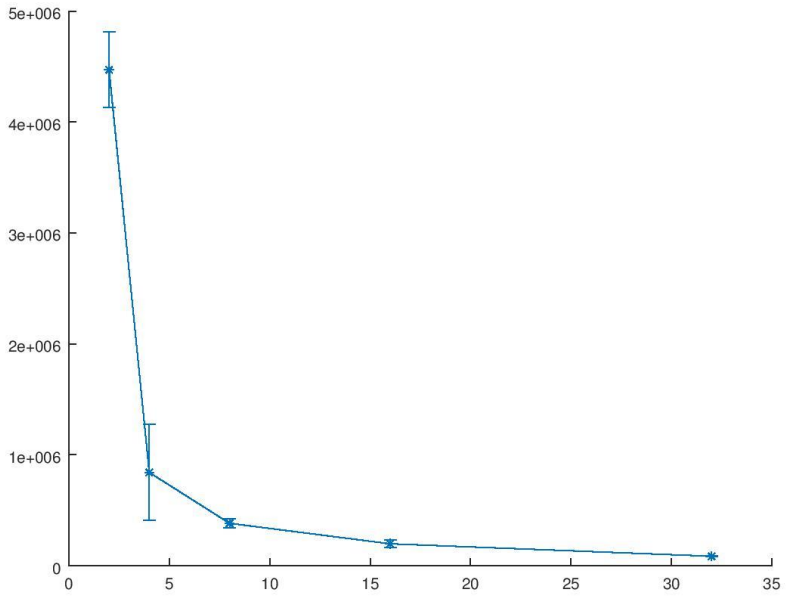


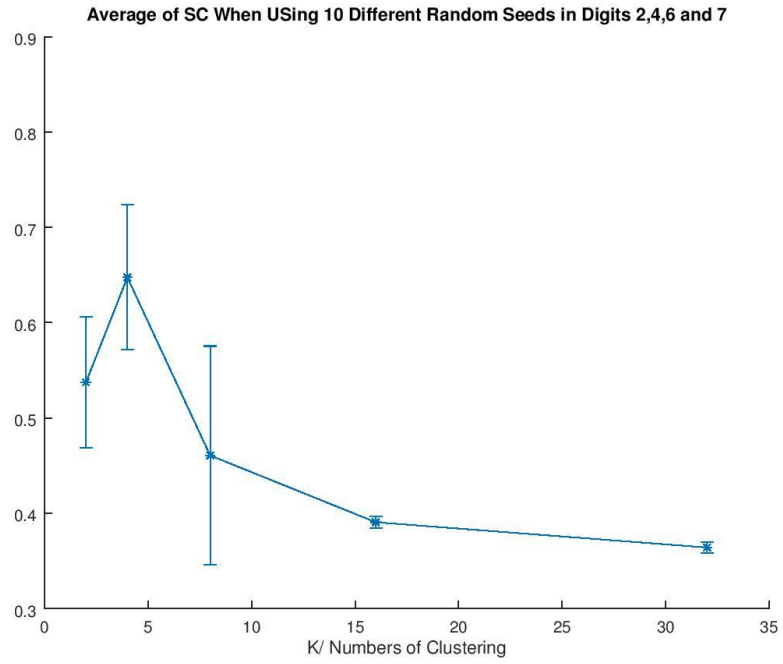


As we can see in the plots above, when the initial starting situation are randomly selecting seeds and full data, the values of the variance of WC SSD are very large. The variance of SC is also not very small, which means k-means are very sensitive.

(ii)

K	2	4	8	16	32
WCSSD_avg	4470253	840843	382229	197370	85263
WCSSD_var	116885353277	188317809988	1588580449	13112365521	7974976
SC_avg	0.5372	0.647466	0.4605424	0.390454	0.363915
SC_var	0.00468636	0.005779	0.01317904	0.0000333044	0.0000351

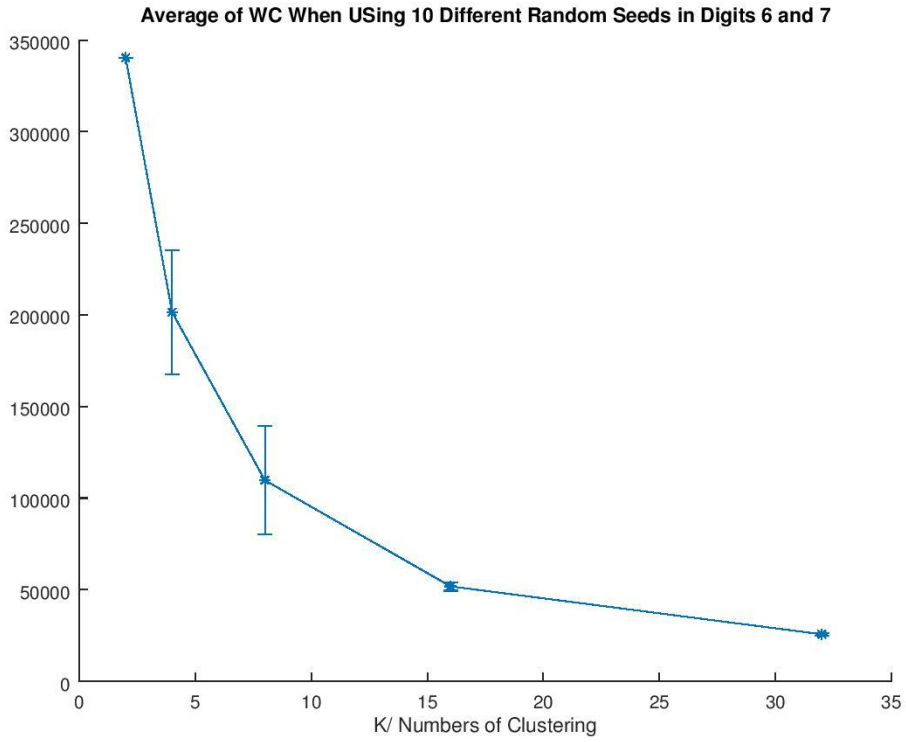


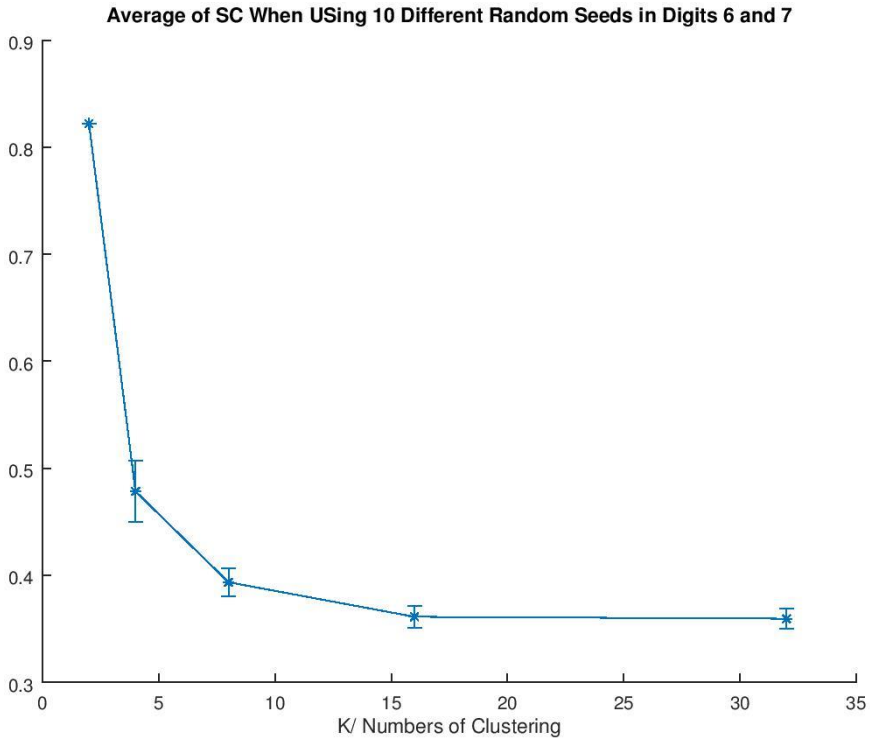


When the dataset is reduced to only have digits 2,4,6 and 7, the values of WC SSD are getting unstable compared to full dataset, because the values of the variance of WC SSD become larger, which means k-means becomes more sensitive. The SCs are still unstable compared to the full dataset situation. K-means clustering is also sensitive in this situation.

(iii)

K	2	4	8	16	32
WCSSD_avg	340372	201488	109742	51808	25651
WCSSD_var	0	1140547984	876811321	5650129	408321
SC_avg	0.821889	0.4783	0.3934	0.36110	0.359306
SC_var	0	0.00080656	0.00016926	0.00010201	0.0000849235

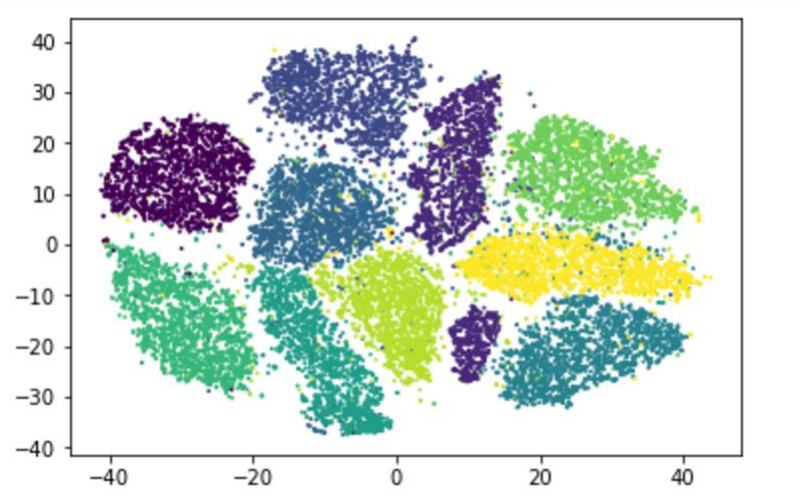




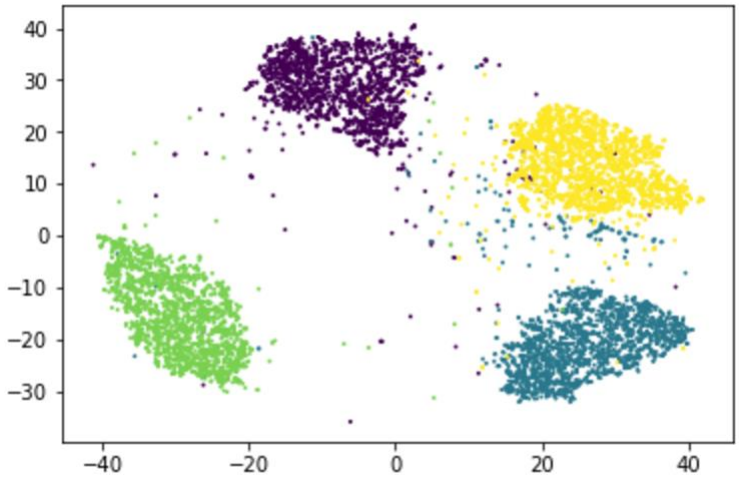
Some of the values of WC SSD becomes unstable reveals that k-means becomes sensitive in these cases (k=2, k=4). The SC remains unstable which also shows that k-means is very sensitive when randomly selecting seeds.

4.

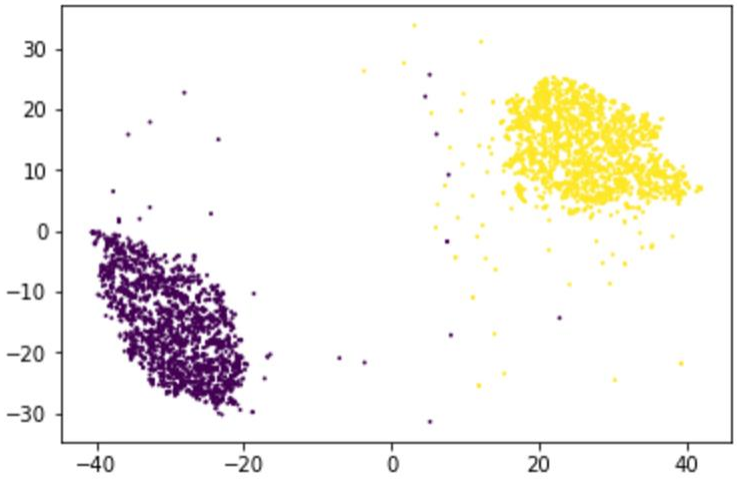
(i) When using the full data, the chosen K is 8, NMI = **0.34649144**



(ii) When using the 2467 digits, the chosen K is 4, NMI = **0.45465341**



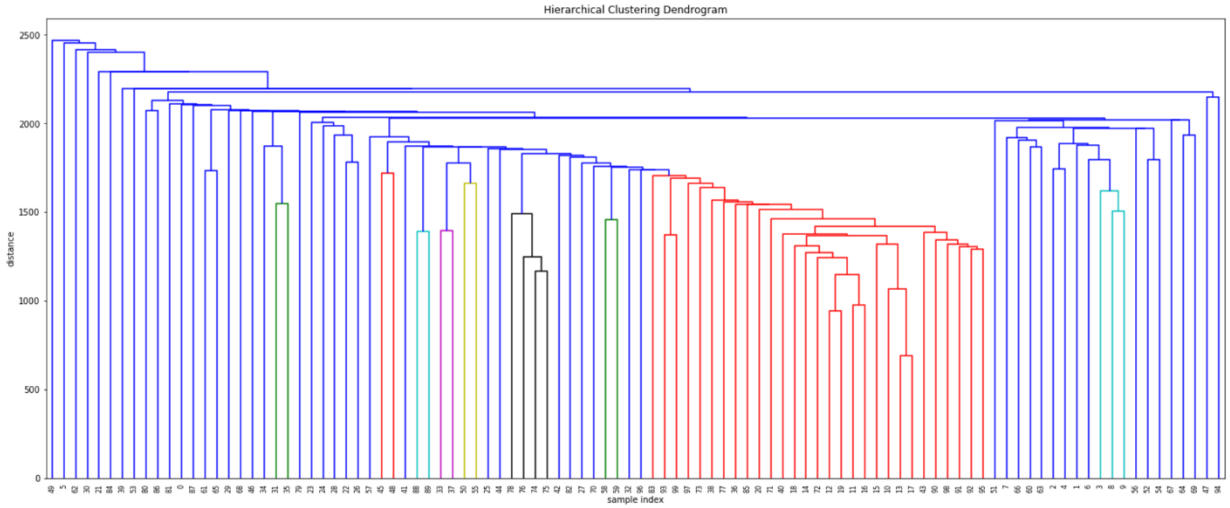
(iii) When using 67 digits, the chosen K is 2, NMI = **0.49071099**



As we can see in the graphs above, when K becomes small, the NMI is getting larger, the cluster of the data has fewer overlap.

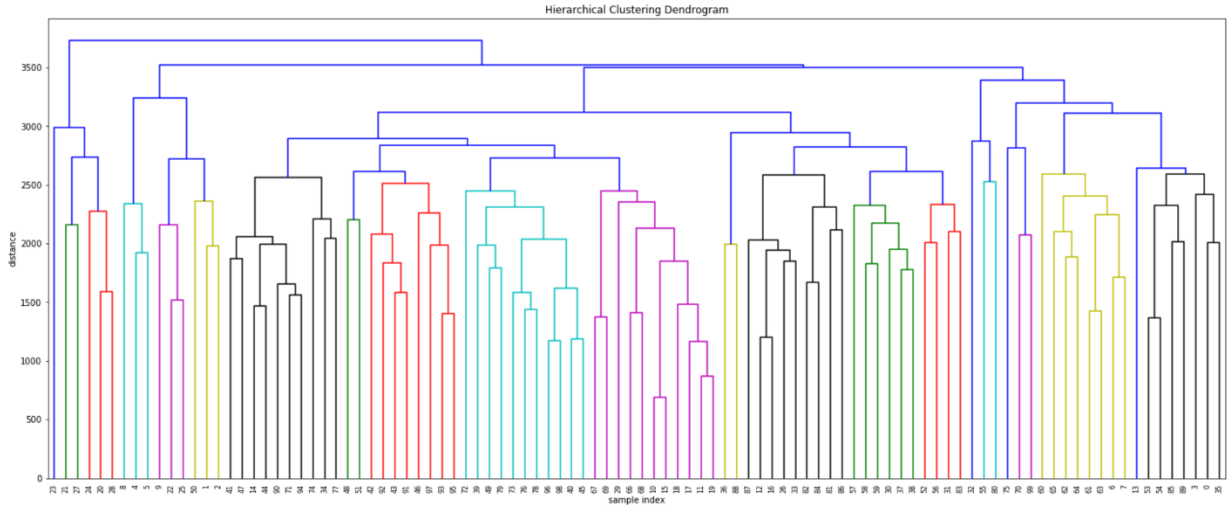
C. Comparison to hierarchical clustering

1.

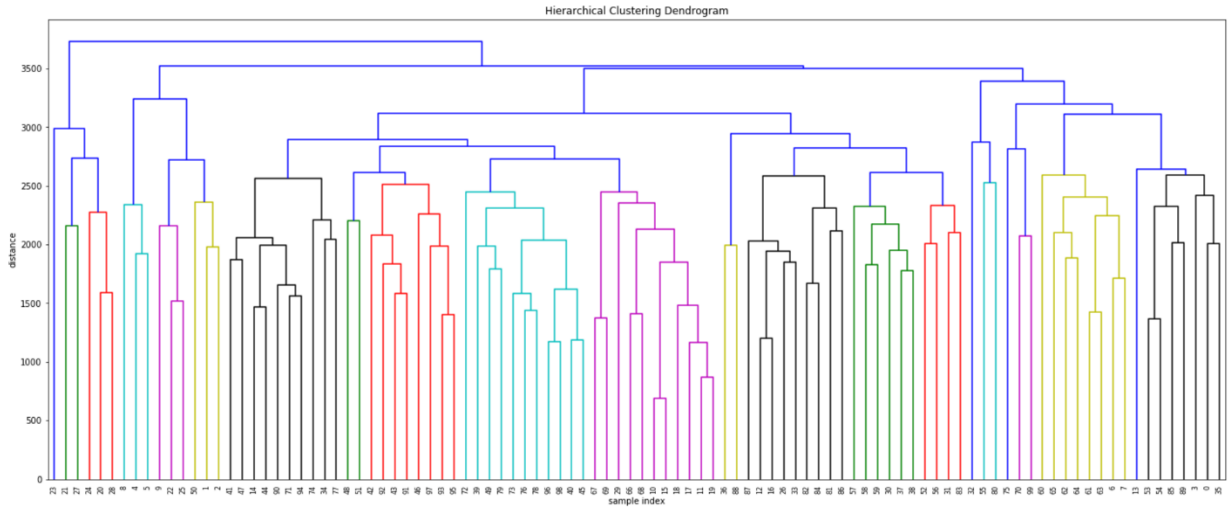


2.

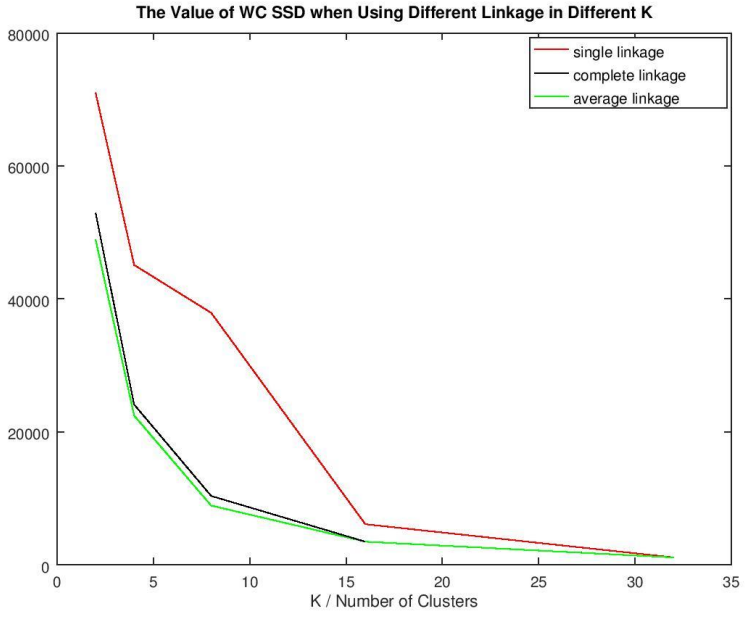
(i)

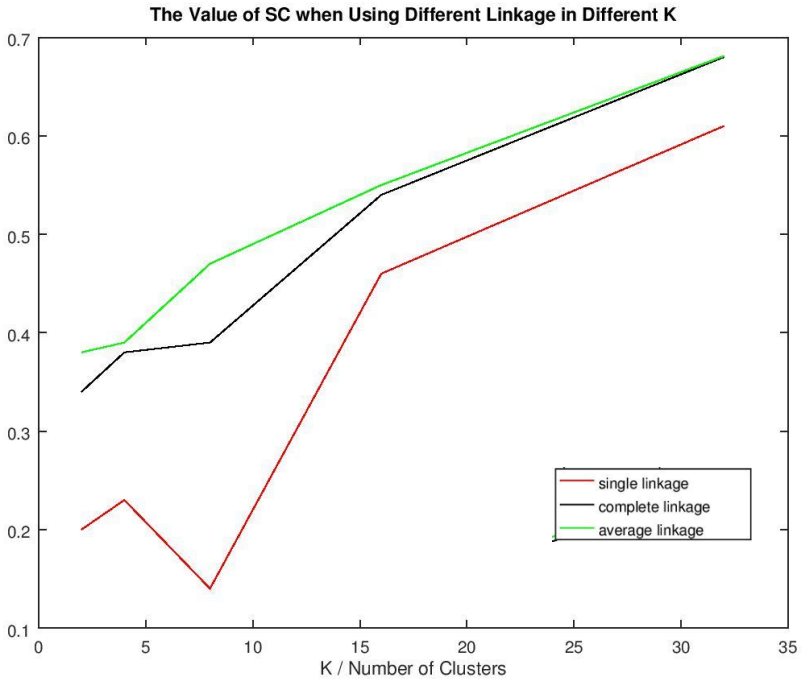


(ii)


Hierarchical Clustering Dendrogram
distance
sample index

3.


The Value of WC SSD when Using Different Linkage in Different K
single linkage
complete linkage
average linkage
K / Number of Clusters



In these plots above, still using k=[2,4,8,16,32].

4.

K=32.

Yes it is different from Part B. The value of SC is larger, but the value of WC SSD is smaller.

5.

The chosen K =32.

For single linkage, NMI = **0.37326362**

For complete linkage, NMI = **0.379802872**

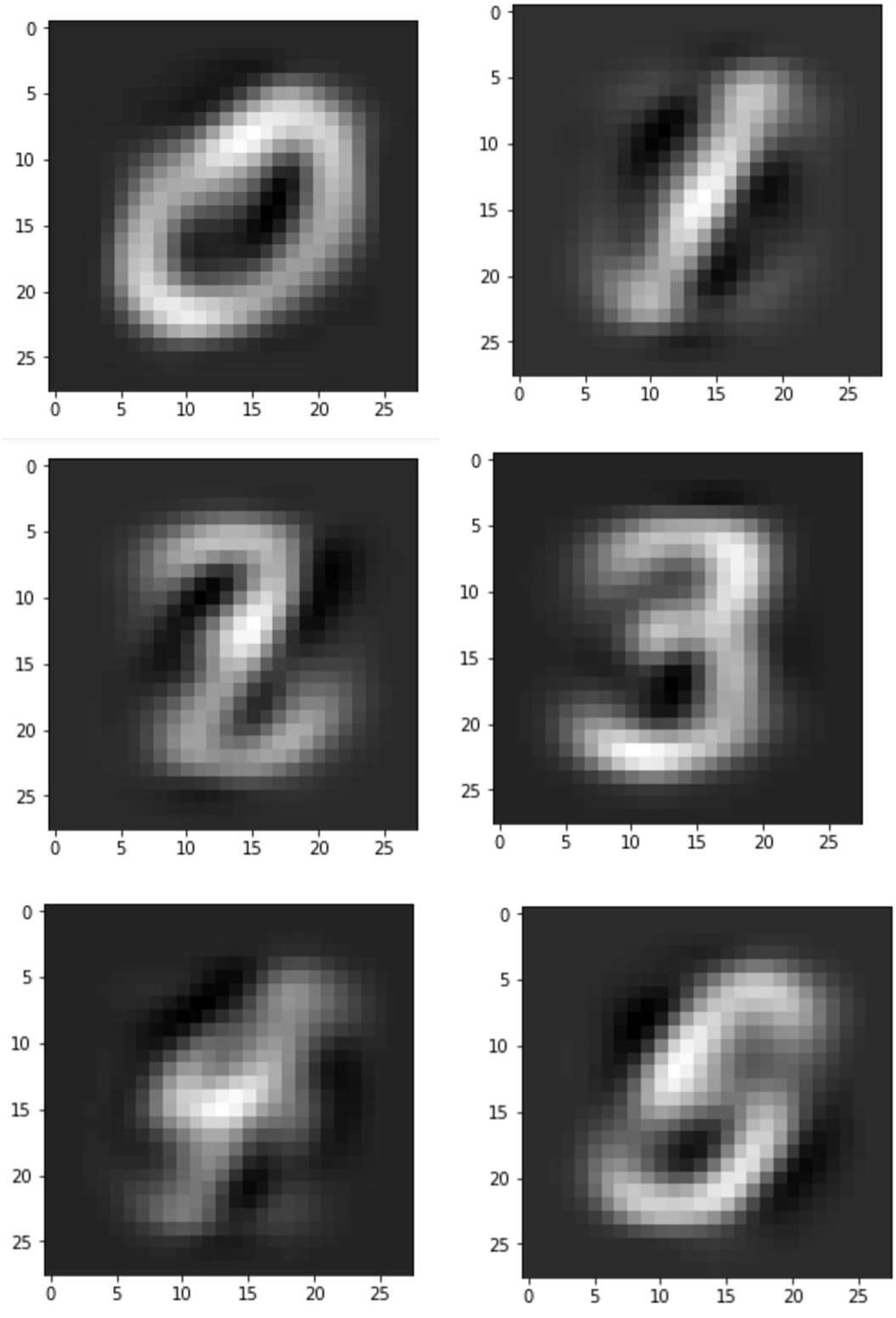
For average linkage, NMI = **0.380982633**

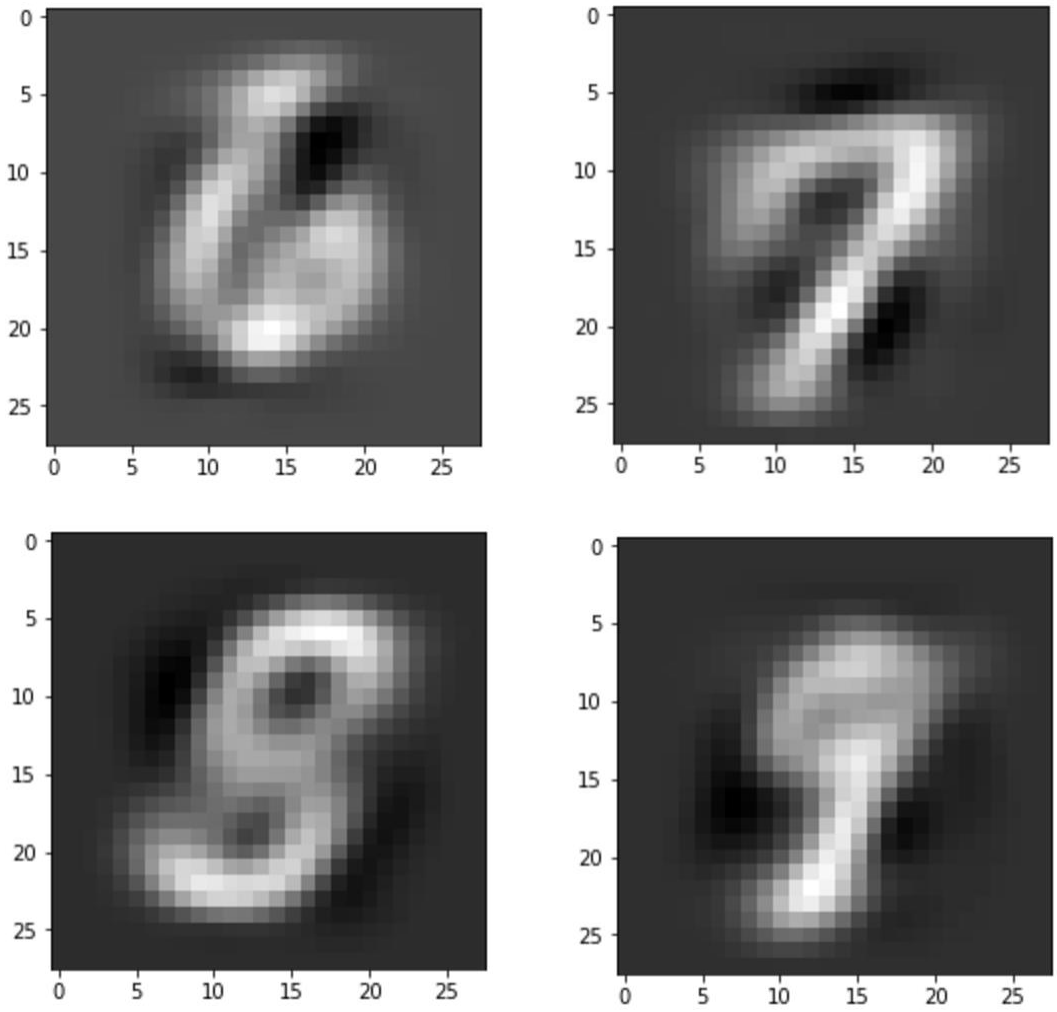
In part B, the chosen K is 9 and NMI = 0.34649144. Hence the NMI is better here than part B.

Bonus

1. Please check the code of hw5.py to see the answer of this part.

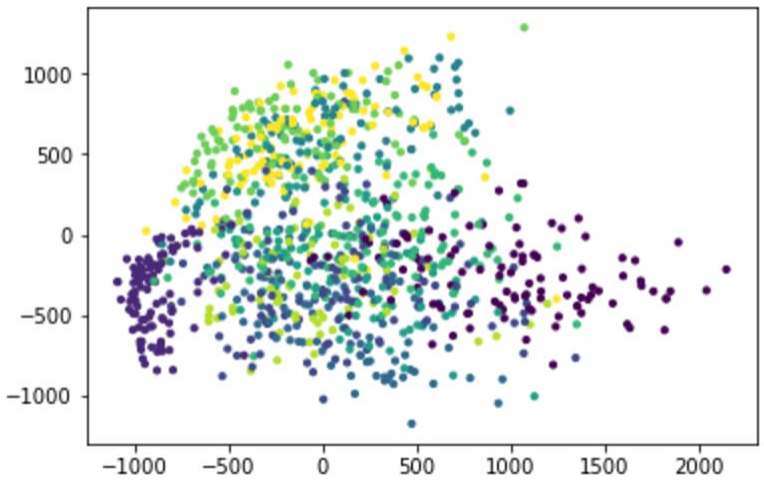
2. Below are the plots of each digit using PCA when principle component is 10.


0 5 10 15 20 25



3.

Below is the graph of 1000 randomly selected examples using the first two principle components. Different colors show the classes of each example respectively.

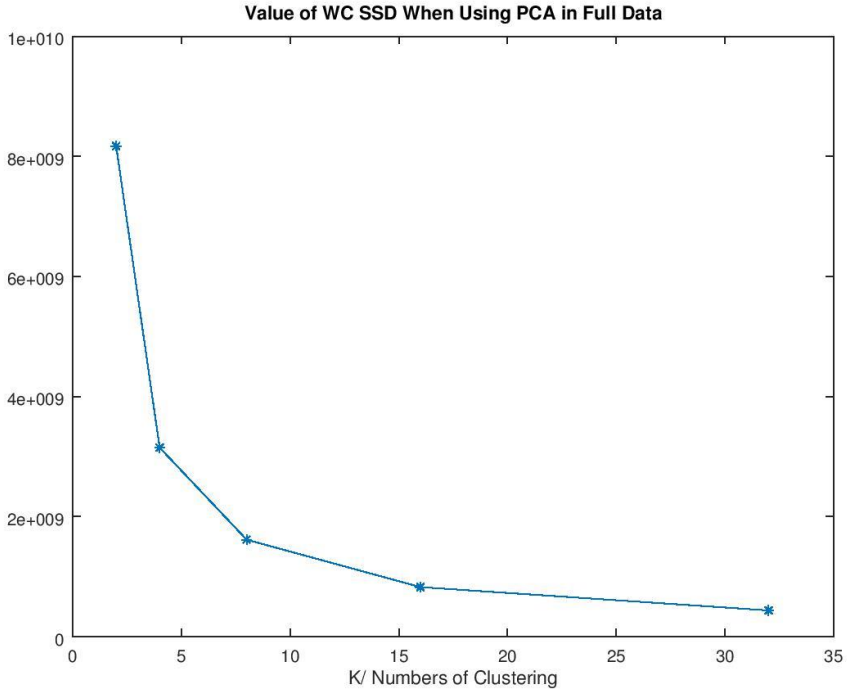


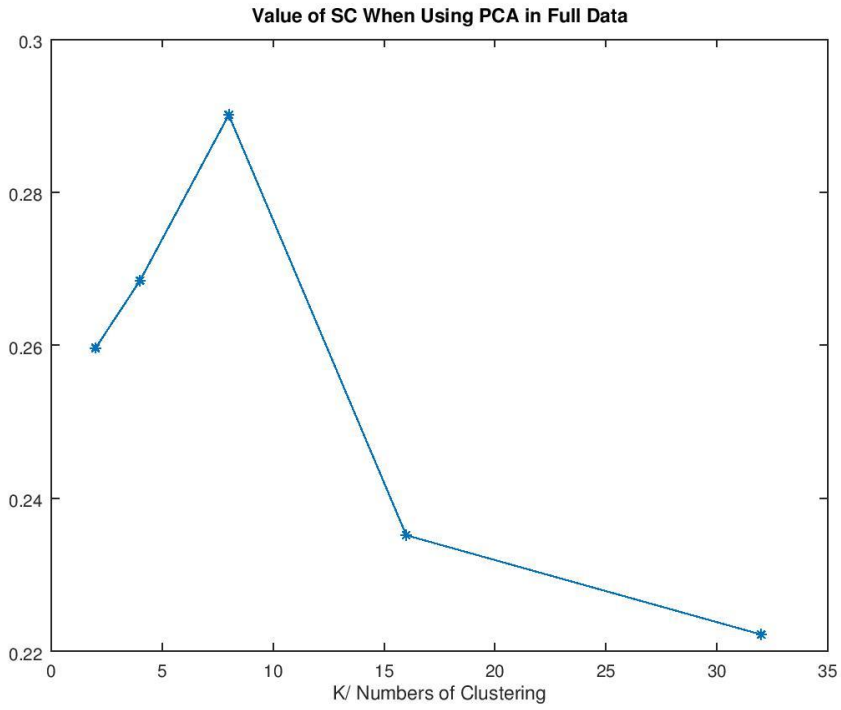
Compared to the tSNE embedding method, the clusters are more chaotic when using 2 principle components in PCA.

4.

Repeat question B.1:

K	2	4	8	16	32
WCSSD	8173013797	3149009918	1616056334	825346933	437140857
SC	0.259647811	0.26845387	0.29011847	0.23515859	0.22217264





Repeat question B.2:

The chosen K here is 8, because when K==8, the SC value is the highest, which shows that the clustering is the best. Compared to the result from the tSNE, the tendency of WC SSD and SC is basically the same : WC SSD are getting lesser and SC is the highest when K equals 8. However, the SC when using PCA is a smaller than using tSNE, which means PCA is not good as tSNE.

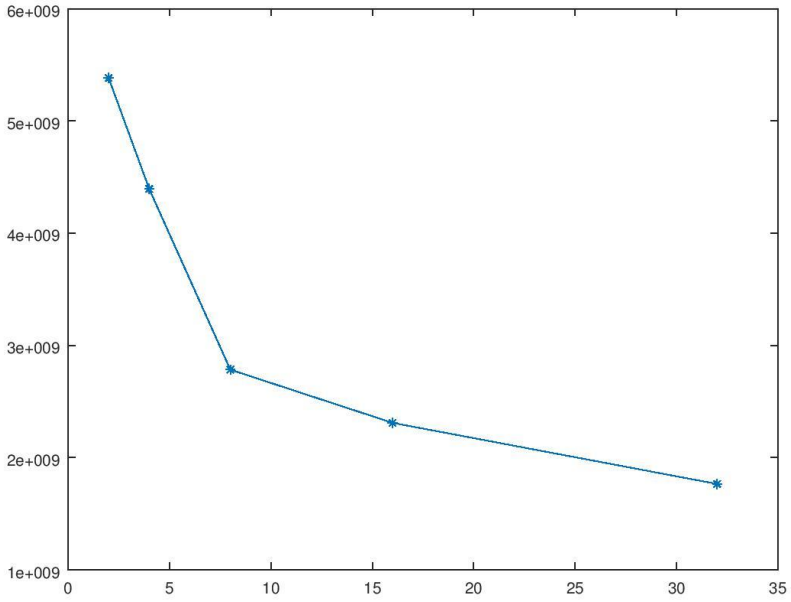
Repeat question B.4:

Using K equals 8, the calculated NMI = **0.27935372232.**

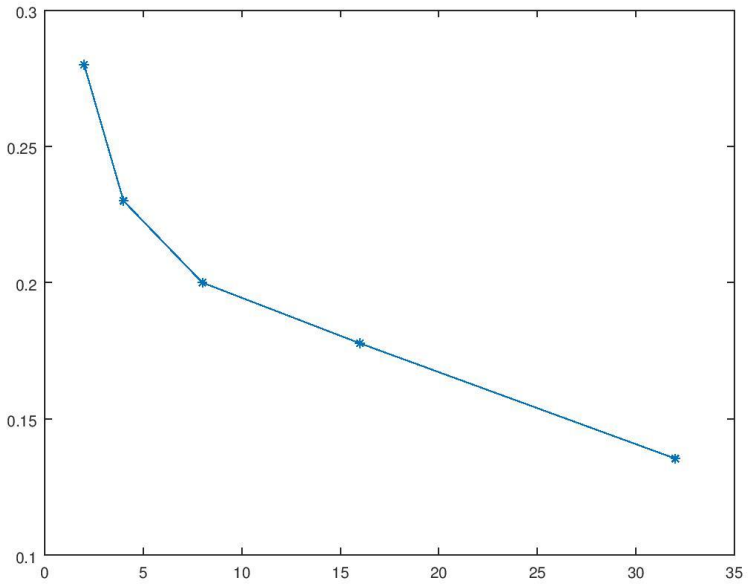
Notice that the NMI value when using PCA is less than using tSNE, which also shows that the PCA is not good as tSNE.

5.

(i) In the situation that dataset only has digits 6 and 7.



WC SSD



SC

From the plot above we can tell that when K=2.these two value reach the peak. The value of WC SSD is bigger than B1, and SC is smaller than B1. Since we mostly based our decision on SC, it performs worse than tSNE.

Compute NMI, we get NMI=**0.42238493202,** it is smaller in B part when K =2.

In summary, when only have digits 6 and 7, the PCA performs weakly compared to tSNE

(ii) In the situation that dataset only has digits 2,4, 6 and 7.

Don’t have enough time to do this part. ☹