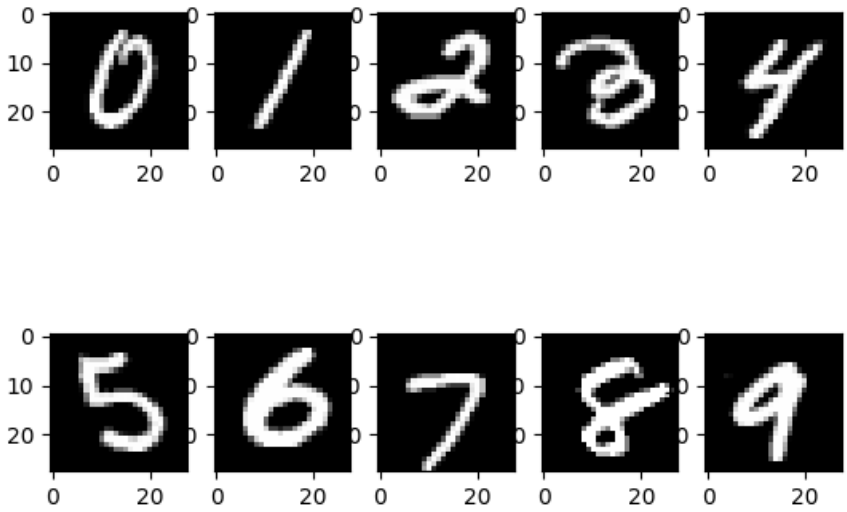
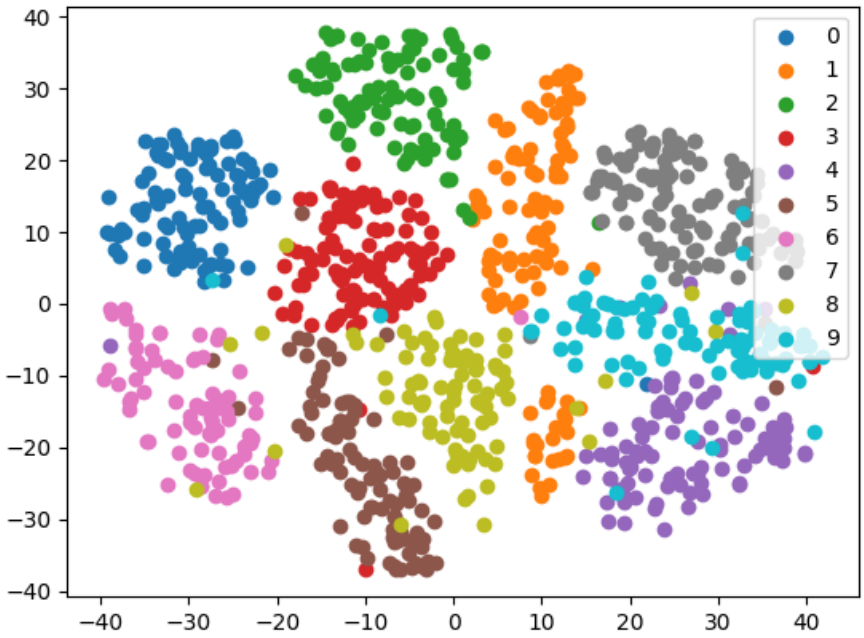
CS573 Assignment5
Sandy Hsiao

1. Exploration

1.1 Plot digits



1.2 Plot 1000 random examples



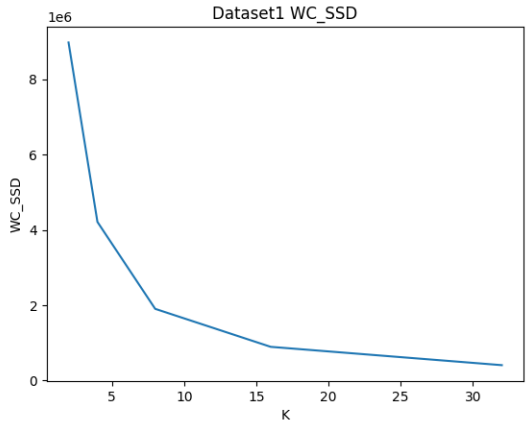
2. K-means

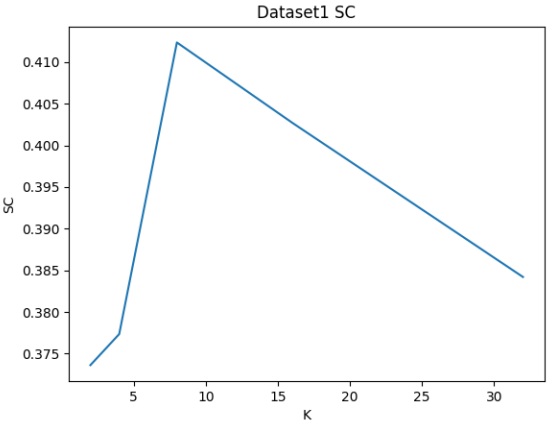
2.1 Code

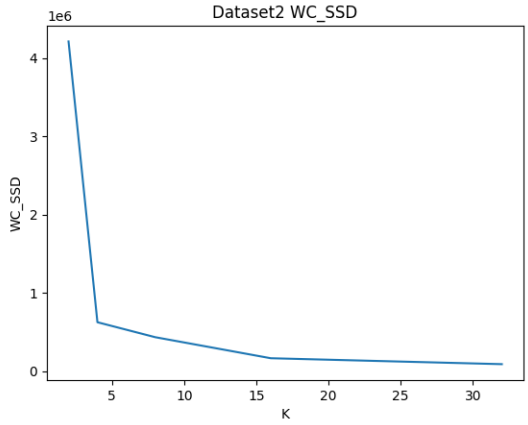
WC_SSD: 1466235.035
SC: 0.411
NMI: 0.469

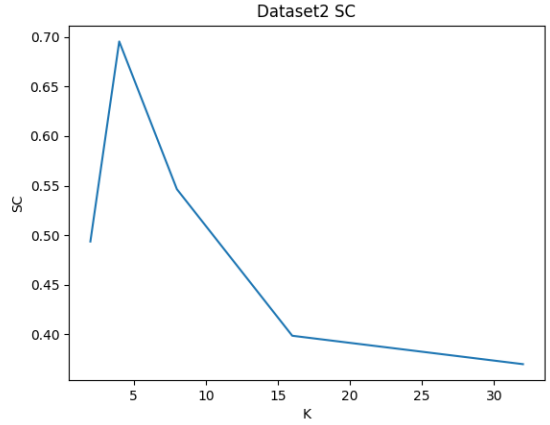
2.2 Analysis

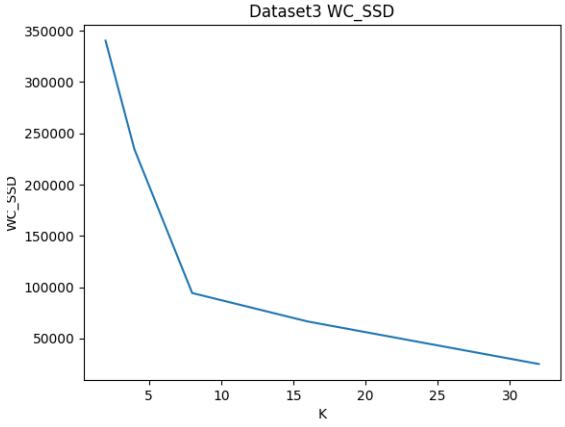
2.2.1 plot

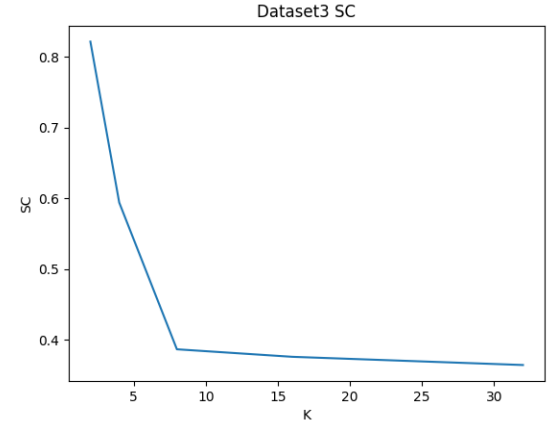












2.2.2 For Dataset1, I chose k=8 because the peak of SC and the elbow point of WC_SSD both occurs at k=8. For Dataset2, I chose k=4 because the peak of SC and the elbow point of WC_SSD both occurs at k=4. For Dataset3, I chose k=2 because the peak of SC occurs at k=2. Though the elbow point occurs at another point, we should choose the point based on SC because SC considers both cohesion and separation of the data while WC_SSD considers only the cohesion of data. Thus, SC is a better mothod to choose k in this case.

It is reasonable that the appropriate k values are different for the three datasets as the class numbers are different.

2.2.3 I ran the experiments with random.seed ∈ [0, 2, 4, 6, 8, 10, 12, 14, 16, 18]. The results show that these datasets are not sensitivity to the initial starting centroids, which is likely because the clustering result does converge. However, different k and different size of data do make a difference in the clustering result. The peak of average SC for the three data indeed occurs at the k values chosen in 2.2.2. Moreover, the average WC_SSD decreases as the number of cluster increases.

	average	standard deviation
k=2	8982844.086	298.165
k=4	4264636.027	64705.870
k=8	1889227.567	5041.586
k=16	890861.020	24360.820
k=32	420395.656	13318.790

Table 1: WC_SSD for Dataset1

	average	standard deviation
k=2	0.374	6.461e-05
k=4	0.374	0.005
k=8	0.399	0.004
k=16	0.399	0.009
k=32	0.398	0.006

Table 2: SC for Dataset1

	average	standard deviation
k=2	4423793.362	299010.386
k=4	840844.136	433957.649
k=8	382268.703	31252.058
k=16	208882.824	37743.679
k=32	90259.818	6266.430

Table 3: WC_SSD for Dataset2

	average	standard deviation
k=2	0.486	0.023
k=4	0.661	0.069
k=8	0.499	0.035
k=16	0.419	0.028
k=32	0.375	0.008

Table 4: SC for Dataset2

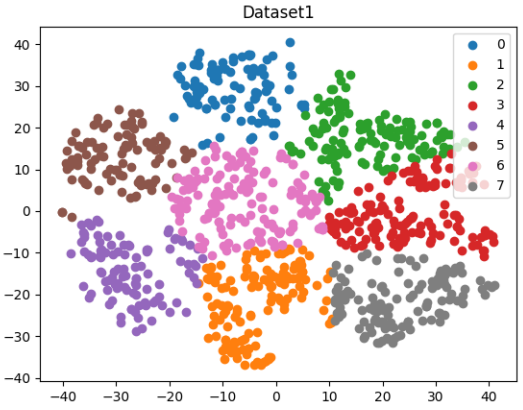
	average	standard deviation
k=2	340372.419	0.000
k=4	180286.767	24353.139
k=8	103022.506	11093.890
k=16	49827.180	2074.694
k=32	26308.991	1029.107

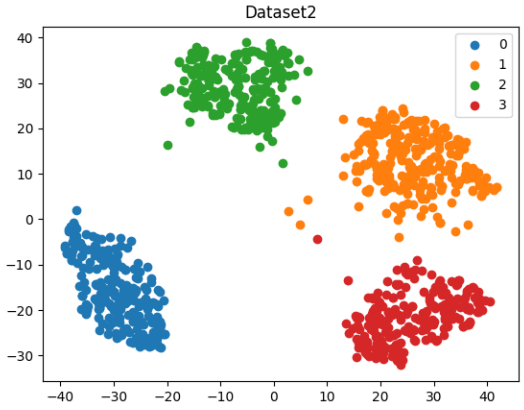
Table 5: WC_SSD for Dataset3

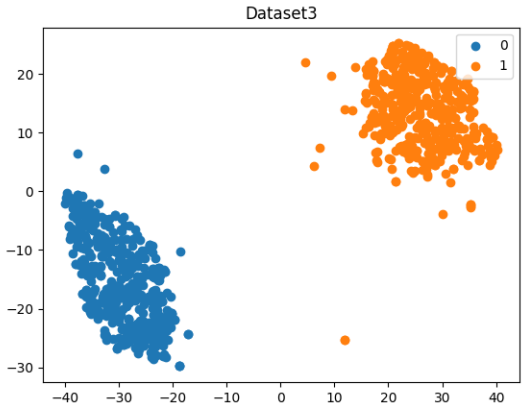
	average	standard deviation
k=2	0.822	0.0
k=4	0.475	0.065
k=8	0.399	0.013
k=16	0.369	0.007
k=32	0.359	0.004

Table 6: SC for Dataset3

2.2.4 With the k values chosen in 2.2.2., the kmeans clustering produces similar NMI for the three datasets. However, Dataset1 has the highest NMI score, which is likely because the dataset contains more examples and the k valuse are larger; thus, the clustering is more accurate.
Dateset1 NMI (k=8): 0.522
Dateset2 NMI (k=4): 0.493
Dateset3 NMI (k=2): 0.499

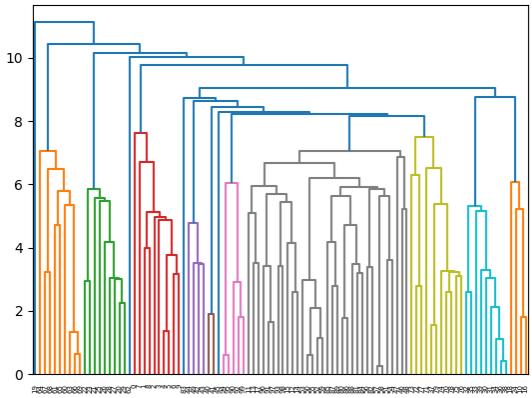




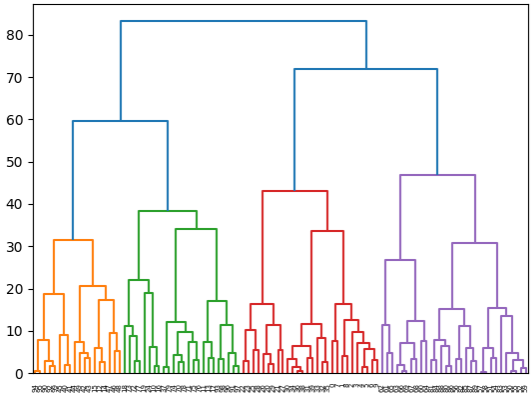


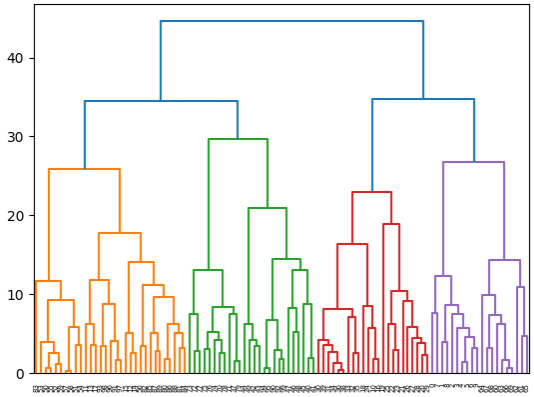
3. Hierarchical Clustering

3.1 single-linkage

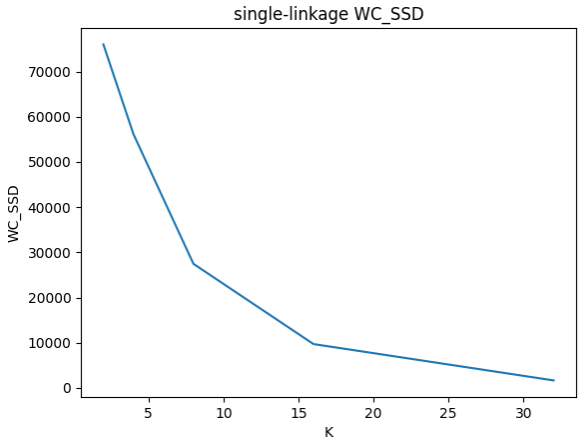


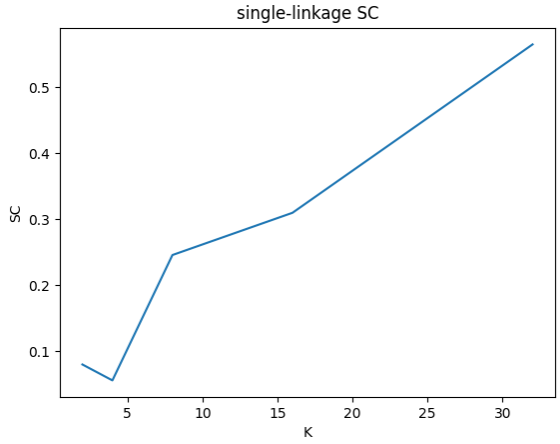
3.2 complete-linkage(left), average-linkage(right)

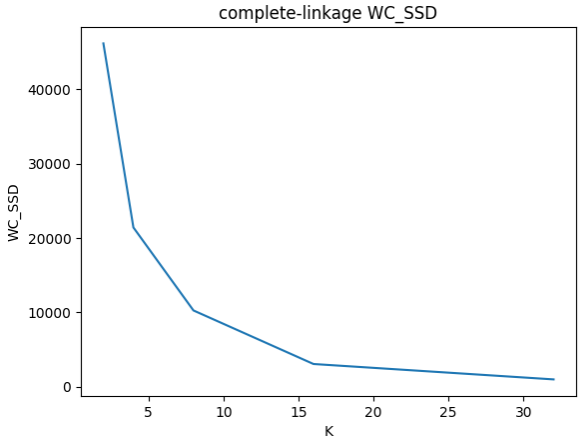


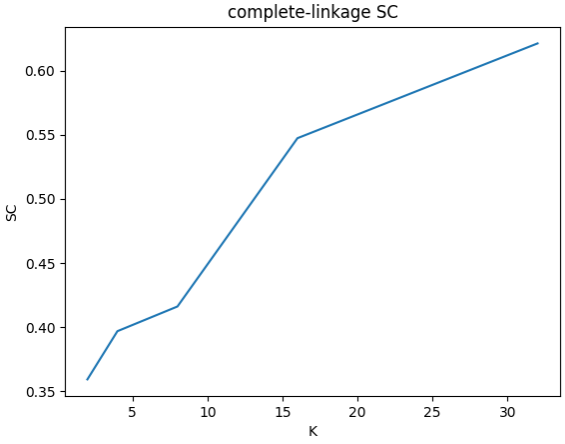


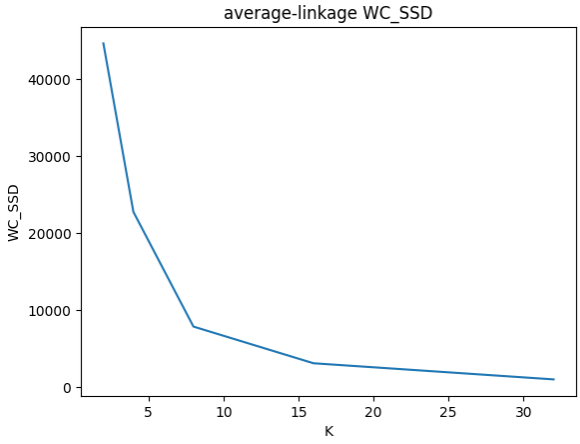
3.3 WC_SSD & SC for various K

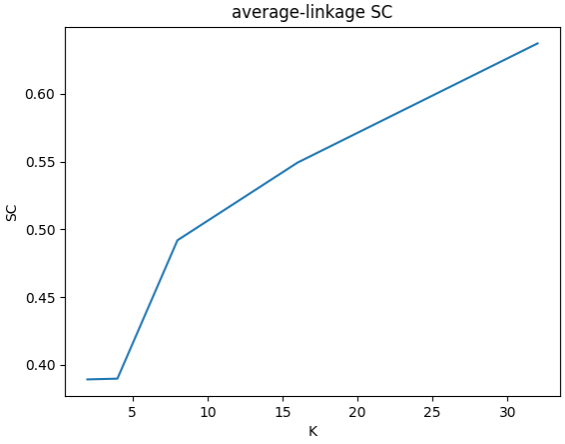












3.4 Based on the WC_SSD and SC results shown in 3.3, I would choose k=16 for all three linkage methods based on WC_SSD because there is no peak value in SC for all the linkage methods. The k value chosen differs from that in section 2, where the k chosen is 8. Though the elbow point in section 2 could also be 16, there is a peak in SC. Thus, 8 is more appropriate in section 2. The difference in k could be attributed to the difference in the examples used for section 2 and section 3. While section 2 used the full dataset, section 3 used only a fraction of the dataset.

3.5 single NMI: 0.364
complete NMI: 0.355
average NMI: 0.366
The NMI for k=8 in section 2 is 0.522, which is clearly higher than that in section 3. Thus, it is better to use k=8 for clustering dataset1 as the nmi is higher and 8 is also closer to the actual number of class labels in the dataset.