CS573 HOMEWORK 5, Using 1 late day

Mohit Gupta

April 19, 2019

1 Exploration

1.1 Pick Digits

The code for this section is in exploration.py. The digits are visualized as given below:

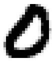

Figure 1: Digit 0

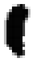

Figure 2: Digit 1

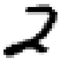

Figure 3: Digit 2

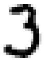

Figure 4: Digit 3

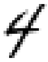

Figure 5: Digit 4

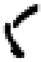

Figure 6: Digit 5

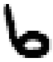

Figure 7: Digit 6

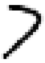

Figure 8: Digit 7

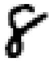

Figure 9: Digit 8

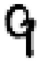

Figure 10: Digit 9

1.2 Visualize

The visualization is given below in figure 11

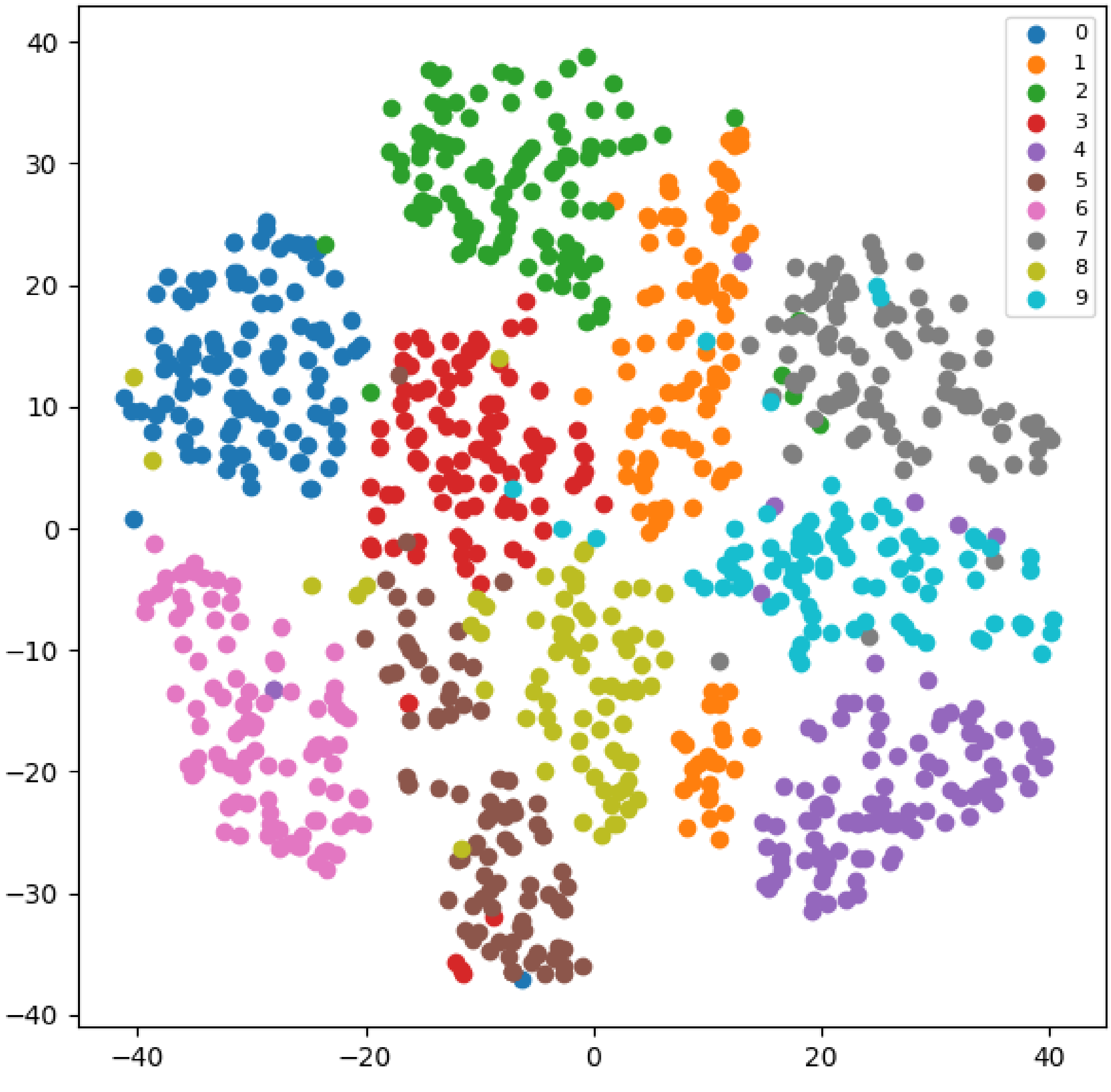

Figure 11: Visualization 1000 examples

2 K-means Clustering

2.1 Code

The code is in kmeans.py. Example usage is given below:
python kmeans.py digits-embedding.csv 10

The scores which I get are given below:
WC-SSD: 1433531.4694124344
SC: 0.7115334025138997
NMI: 0.35591657713508684

2.2 Analysis

2.2.1 Cluster the data with different values of K $\in$ [2,4,8,16,32] and construct a plot showing the within-cluster sum of squared distances (WC SSD) and silhouette coefficient (SC) as a function of K.

The plot showing WC-SSD with K is given below in figure 12

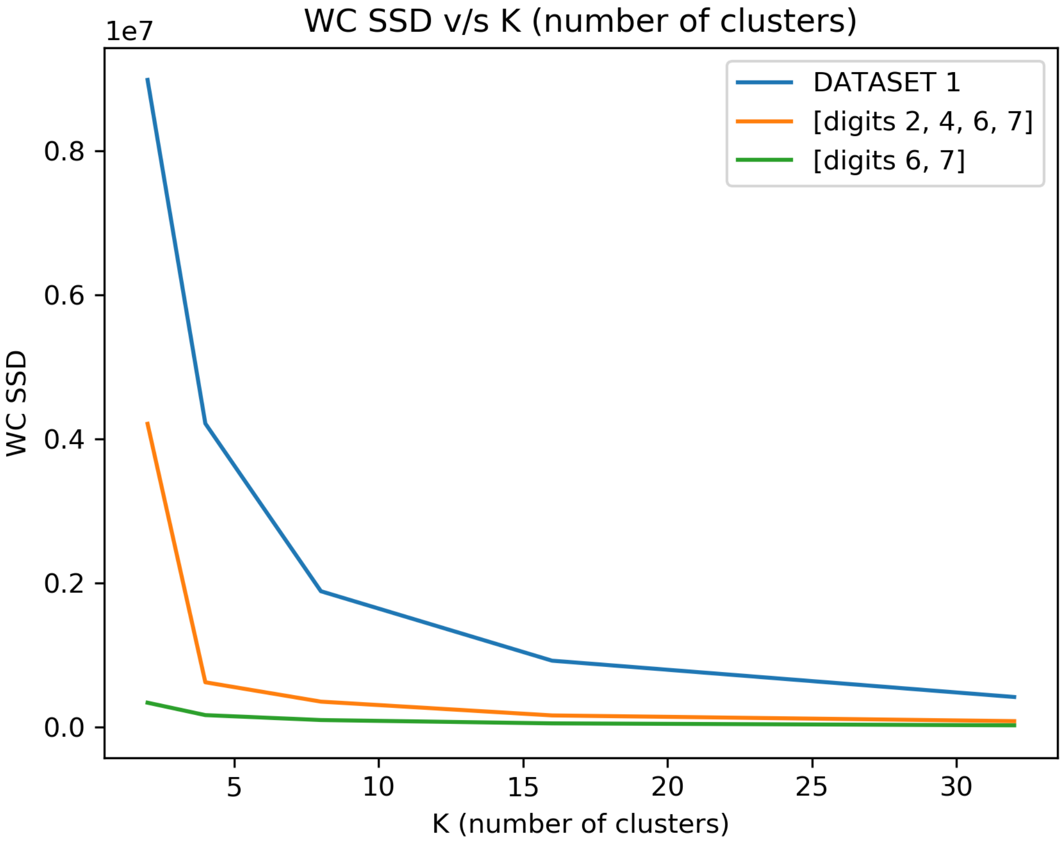

Figure 12: Learning Curve WC-SSD vs K

The plot showing SC with K is given below in figure 13

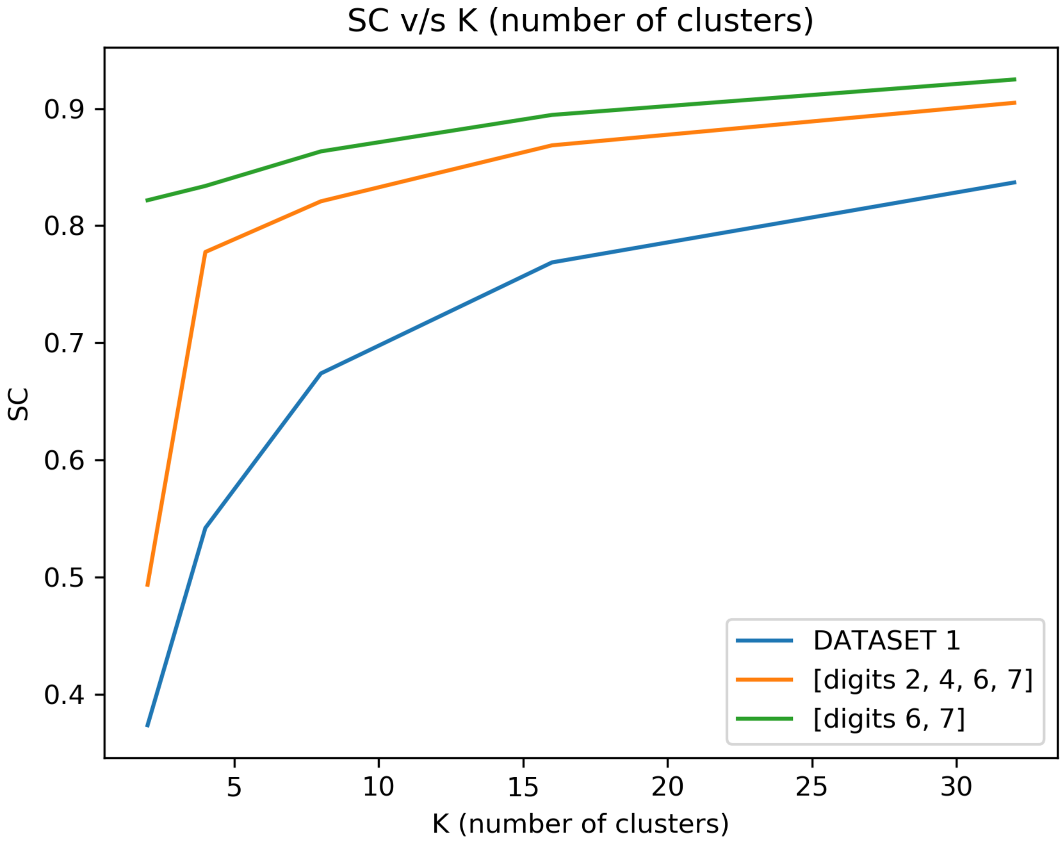

Figure 13: Learning Curve SC vs K

2.2.2 Using the results from Step 1, choose an appropriate K for each dataset and argue why your choice of K is the best. Discuss how the results compare across the two scores and the three versions of the data

As we can see from the elbow curve for WC-SSD vs K, we chose K=8 for dataset 1, K=4 for dataset 2, K=2 for dataset 3.

If we look carefully, we observe that the WC SSD is very high in the starting for Dataset 1 i.e. for k=2, it then decreases sharply till K=8 and then there is relatively small decrease from K=8 to K=16. From this observation from Figure 12, we chose K=8 for dataset 1.

For dataset 2, we see a very sharp decrease from a high value of around 4 million at k=2 to 623865 at k=4. Then there is a relatively small decrease to 355343 at k=8. Observing this, we chose K=4 for dataset 2.

For dataset 3, we observe that the WC SSD for k=2 is just 340372 which sees a relatively small decrease to 168176 at k=4. We notice that its best to chose k=2 for dataset 3.

The scores of WC SSD for dataset 1 are relatively high compared to dataset 2 which in turn is higher than dataset 3.

The SC scores as per the definition given in the slides is observed to have a monotonic increase as shown in figure 13.

So, we use WC SSD to pick k and thus chose k=8 for dataset 1, k=4 for dataset 2 and k=2 for dataset 3.

2.2.3 Repeat Step 1 with 10 times using 10 different random seeds. Measure and report the average and standard deviation (for WC SSD and SC) for the different values of K. Discuss what the results show about k-means sensitivity to initial starting conditions

The plot showing average WC SSD with K along with standard deviation is given below in figure 14.

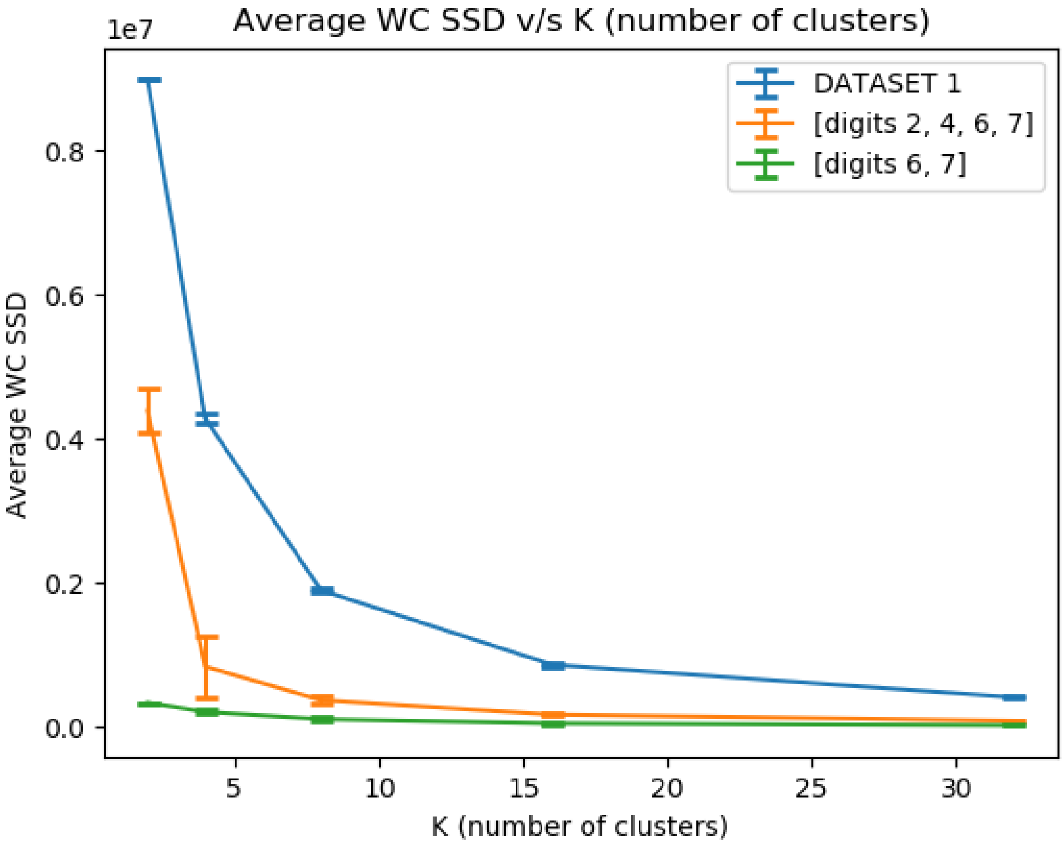

Figure 14: Learning Curve Average WC SSD vs K

The scores for dataset 1 are given below:
Avg WC-SSD list: [8983697.2123934757, 4283094.3622102402, 1900913.1739781355, 866932.64267569443, 413110.93332069973]
Standard Deviation WC-SSD list: [309.76244666027816, 73752.313819248448, 16457.185575483072, 21602.979590098737, 14211.391310006113]

The scores for dataset 2 are given below:
Avg WC-SSD list: [4393863.7952239905, 840844.13564789353, 375385.66214356158, 174267.60669973886, 88004.814879763726]
Standard Deviation WC-SSD list: [303779.44147816917, 433957.64895931911, 49944.490165191746, 9660.7556571699224, 5400.0233623331369]

The scores for dataset 3 are given below:
Avg WC-SSD list: [340372.41942807363, 211907.00196273447, 110145.41606064285, 51738.626408906799, 25927.320640480193]
Standard Deviation WC-SSD list: [5.8207660913467407e-11, 29228.849932261273, 29730.482352431151, 3606.851384166459, 895.08752953603801]

We observe that the standard deviation is relatively high for dataset 2 for k=2 and k=4 which indicates k-means is sensitive in these settings. It is relatively robust for dataset 1 and dataset 3 with standard deviation not being as high compared to the values observed in dataset 2 for small k values.

We observe that k-means algorithm is sensitive to initial starting conditions especially for lower k values in dataset 2.

The same pattern can be observed for SC values as shown in the plot in Figure 15.
The Avg SC scores along with standard deviation are given below:

Dataset 1
Avg SC list: [0.37359159965488248, 0.53854633984476863, 0.67421411024398947, 0.77264142603499641, 0.83915126117476324]
Standard Deviation SC list: [6.5434921492337958e-05, 0.0036045210996282807, 0.0020613823143138662, 0.0022356300686984638, 0.0018823532000048231]

Dataset 2
Avg SC list: [0.4847451514428549, 0.7588158565186276, 0.81858959554851718, 0.86624759713879951, 0.90232368277782948]
Standard Deviation SC list: [0.021350942215739865, 0.037478560826040841, 0.0066183592823147452, 0.0019783290813360872, 0.0028146563956198433]

Dataset 3
Avg SC list: [0.82174535051256936, 0.83491216840299742, 0.86182658125033473, 0.89594256426305441, 0.92471913167821462]
Standard Deviation SC list: [0.0, 0.0045019686445715813, 0.0038058659816849888,

0.0019474591171661565, 0.0008510734116287951l]

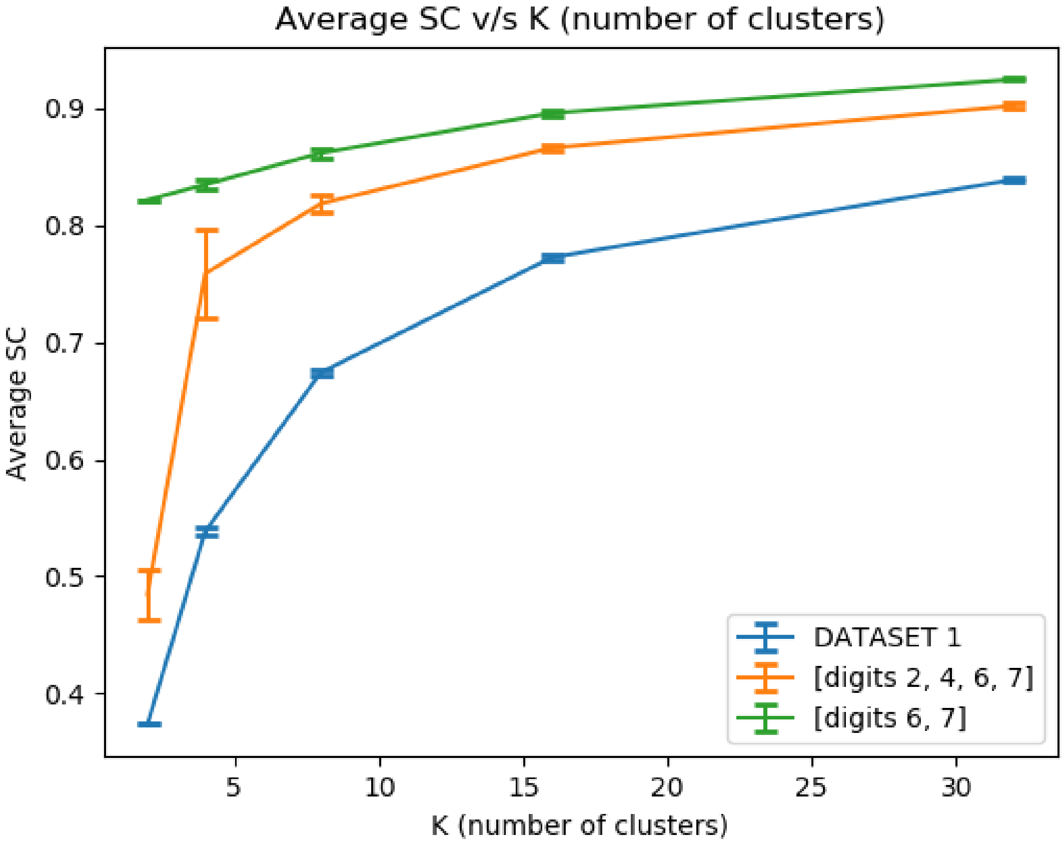

Figure 15: Learning Curve Average SC vs K

2.2.4 For the value of K chosen in Step 2, cluster the data again (a single time) and evaluate the resulting clusters using normalized mutual information gain (NMI). Calculate NMI with respect to the image class labels. Visualize 1000 randomly selected examples in 2d, coloring the points to show their corresponding cluster labels. Discuss how both the NMI and visualization results compare across the three versions of the data

We choose k=8 for dataset 1, k=4 for dataset 2, k=2 for dataset 3
NMI for dataset 1 is 0.346739899006
NMI for dataset 2 is 0.45465341281
NMI for dataset 3 is 0.490710990204
The visualization for dataset 1, dataset 2 and dataset 3 are given below in figure 16, 17 and 18:

We observe from NMI values that clustering has been the best for dataset 3 with a very high NMI value of 0.49. The NMI score is relatively less for dataset 2 with 0.45 compared to dataset 3. Although a score of 0.45 means that it is still very good clustering for dataset 2. The lowest NMI score is observed for dataset 1 i.e. 0.35 which means that the clustering is not as good as dataset 2 and dataset 3.

The same trend can be seen when we observe figures 16, 17 and 18. We see that there are very clear clusters in figure 18 for dataset 3, figure 17 for dataset 2 indicating our chosen k values of 2 and 4 for them are good. For dataset 1, we see that there is not much separation between the leftmost points of cluster 3 and rightmost points of cluster 1 for example. Similarly, we observe a gap between left and right points of clutser 5. We conclude that there could be more clusters formed for dataset 1.

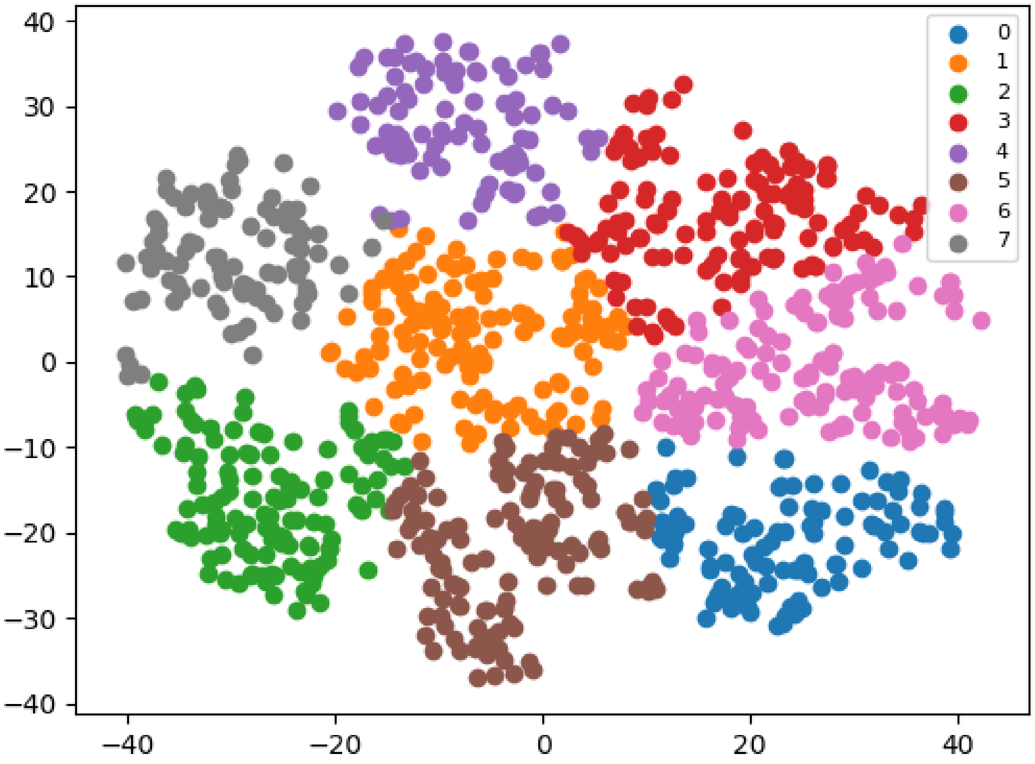

Figure 16: Visualization Dataset 1

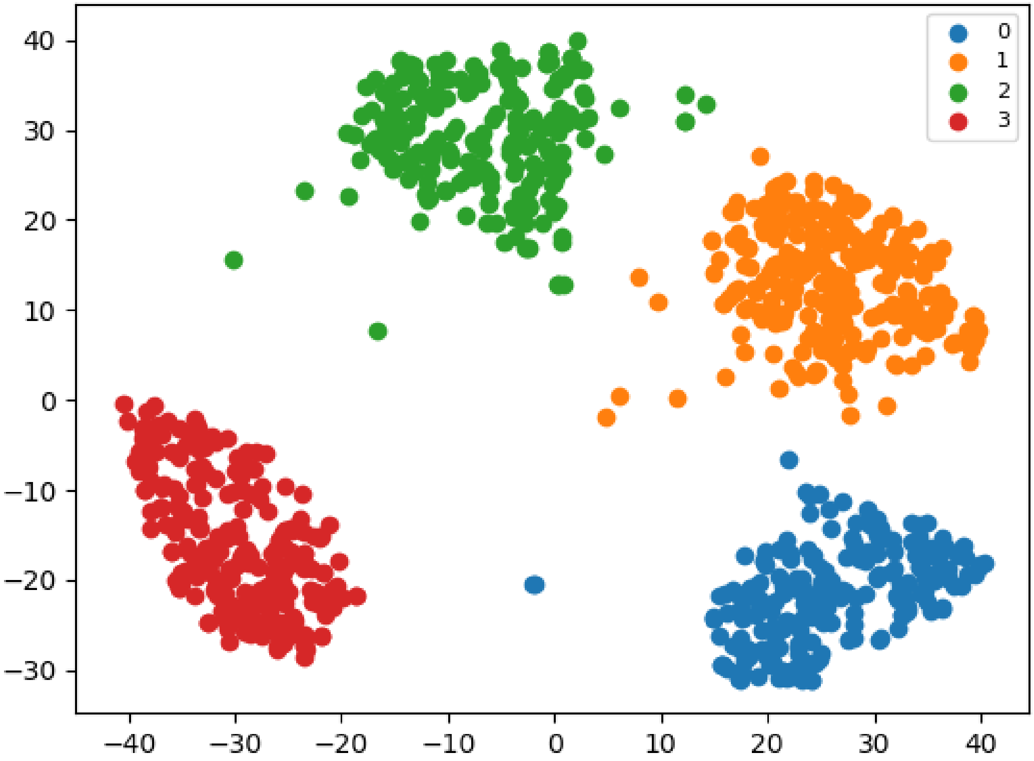

Figure 17: Visualization Dataset 2

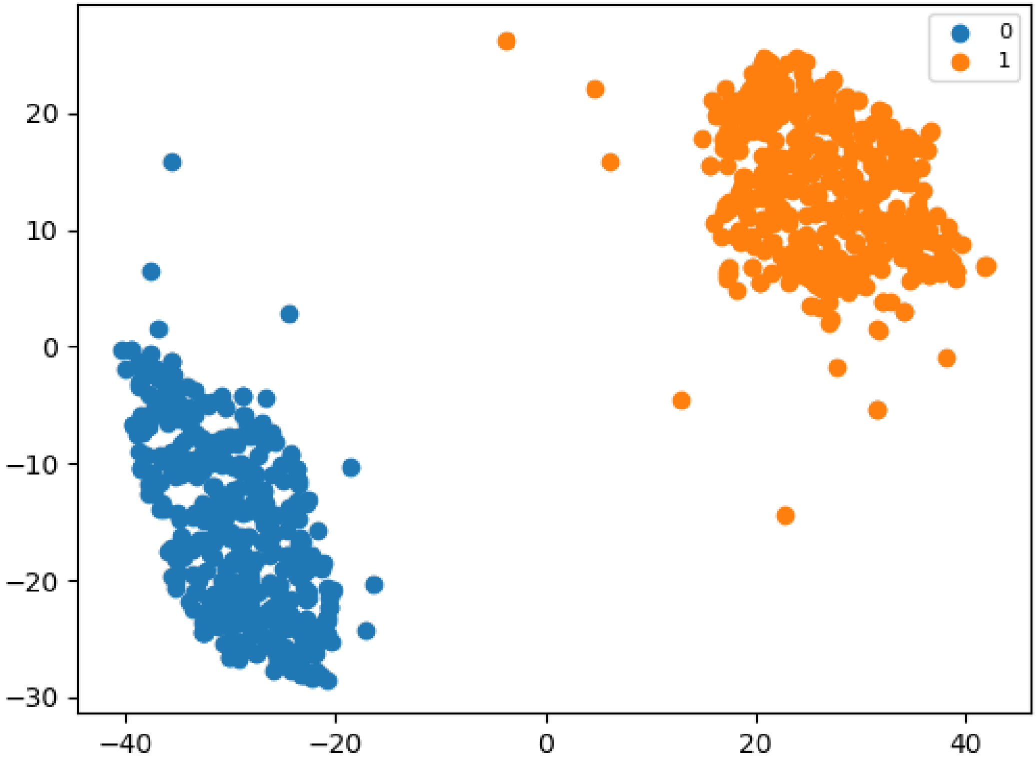

Figure 18: Visualization Dataset 3

3 Hierarchical Clustering

The code is given in hierarchical.py

3.1 Create sub-samples for Dataset 1 in Section 2 by sampling 10 images at random from each digit group (i.e., 100 images in total). Use the scipy agglomerative clustering method to cluster the data using single linkage. Plot the dendrogram

The dendogram is given below in figure 19:

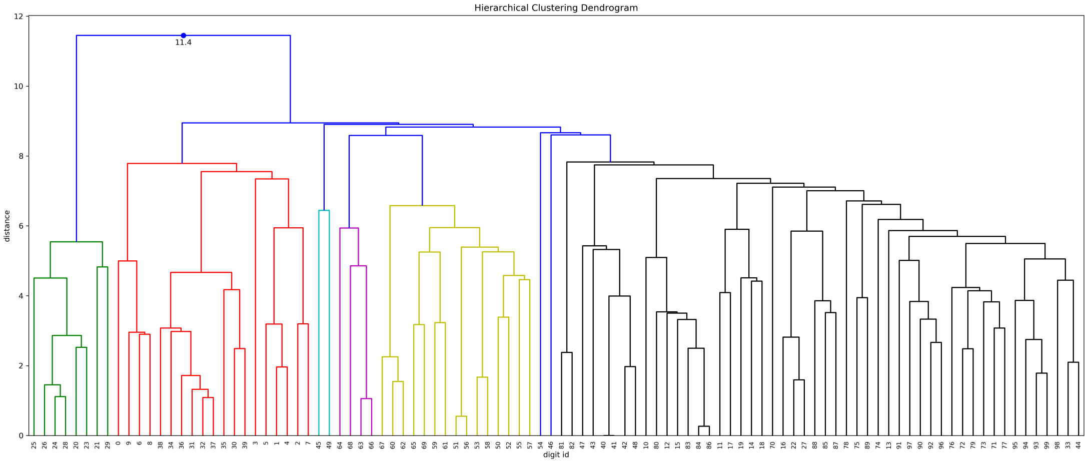

Figure 19: Dendogram Single Linkage

3.2 Cluster the data again, but this time using (i) complete linkage, and (ii) average linkage. Plot the associated dendrograms

The dendogram for complete linkage, average linkage is given below in figure 20,21 respectively.

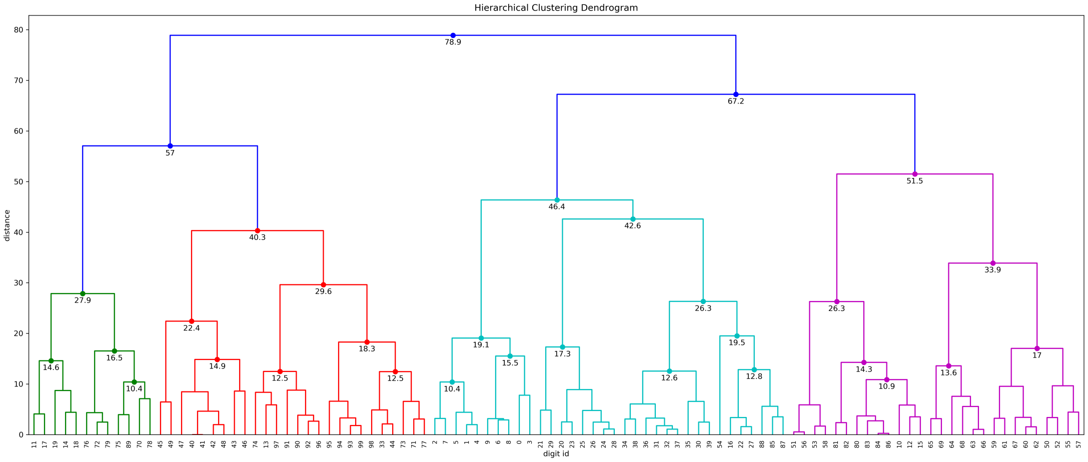

Figure 20: Dendogram Complete Linkage

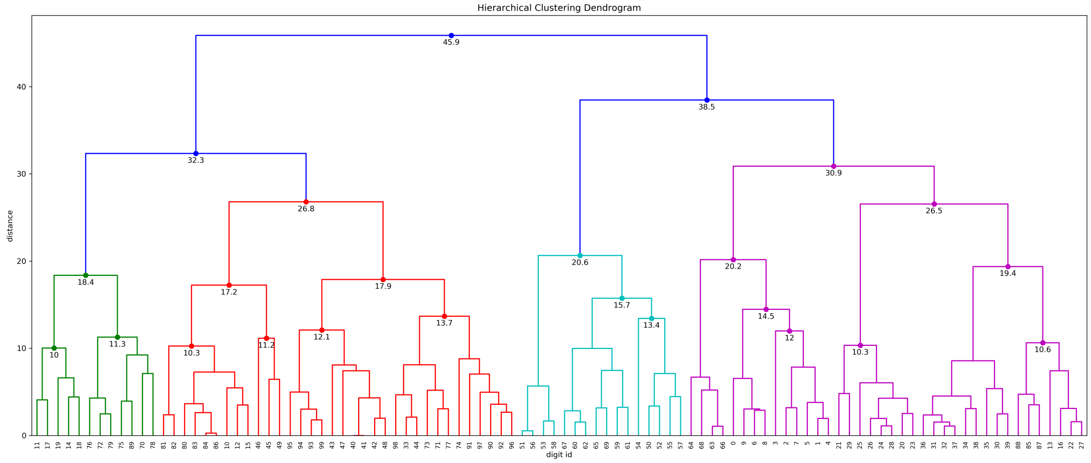

Figure 21: Dendogram Average Linkage

3.3 Consider cutting each of the dendrograms at successive levels of the hierarchy to produce parti- tions of different sizes (i.e., vary choice of K). Construct a plot showing the within-cluster sum of squared distances (WC SSD) and silhouette coefficient (SC) as a function of K

We vary k as [2,4,8,16,32]. The plot showing WC SSD vs K is given below in figure 22,23,24 for single, complete and average linkages.

The plot showing SC vs K is given in figure 25,26,27 for single, complete and average linkages.

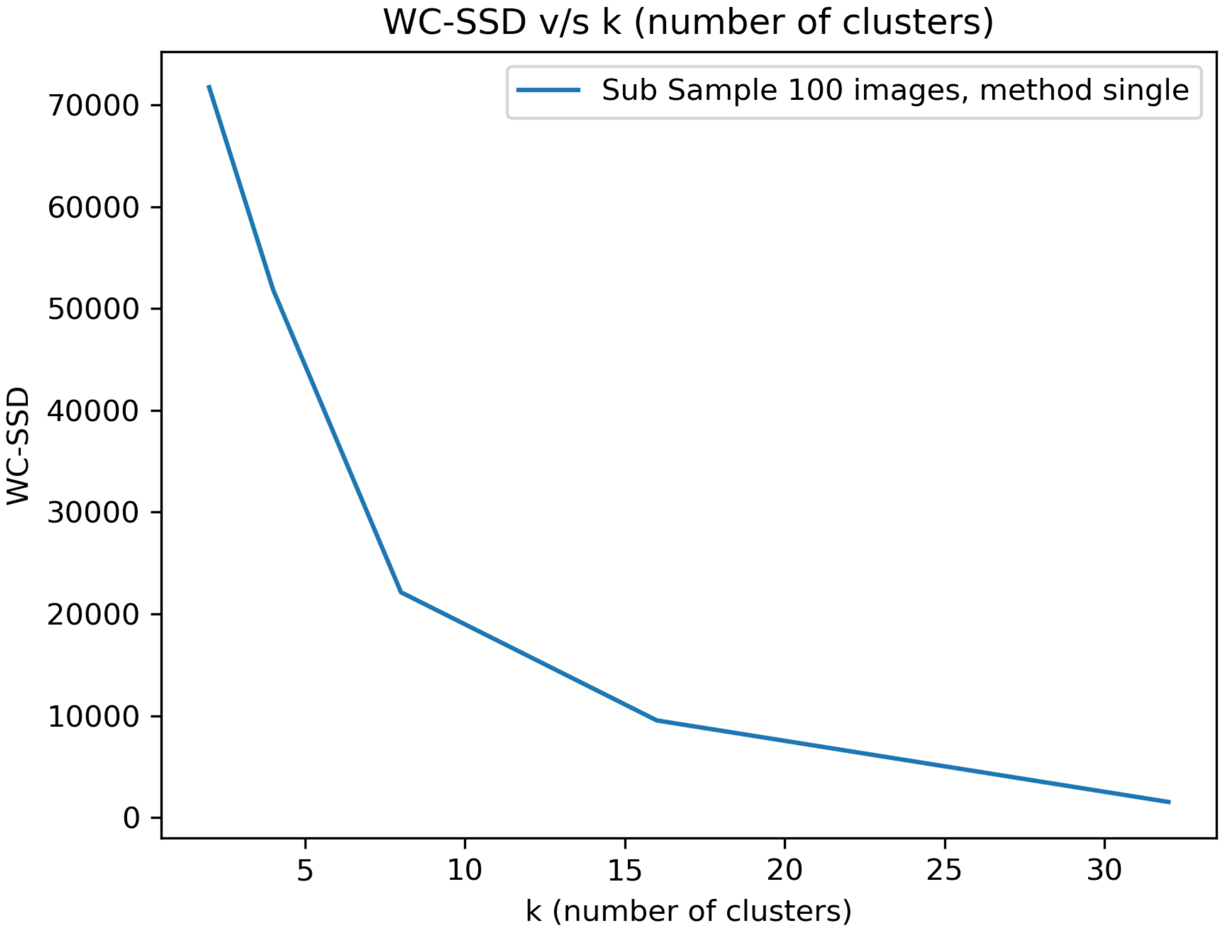

Figure 22: WC SSD vs K

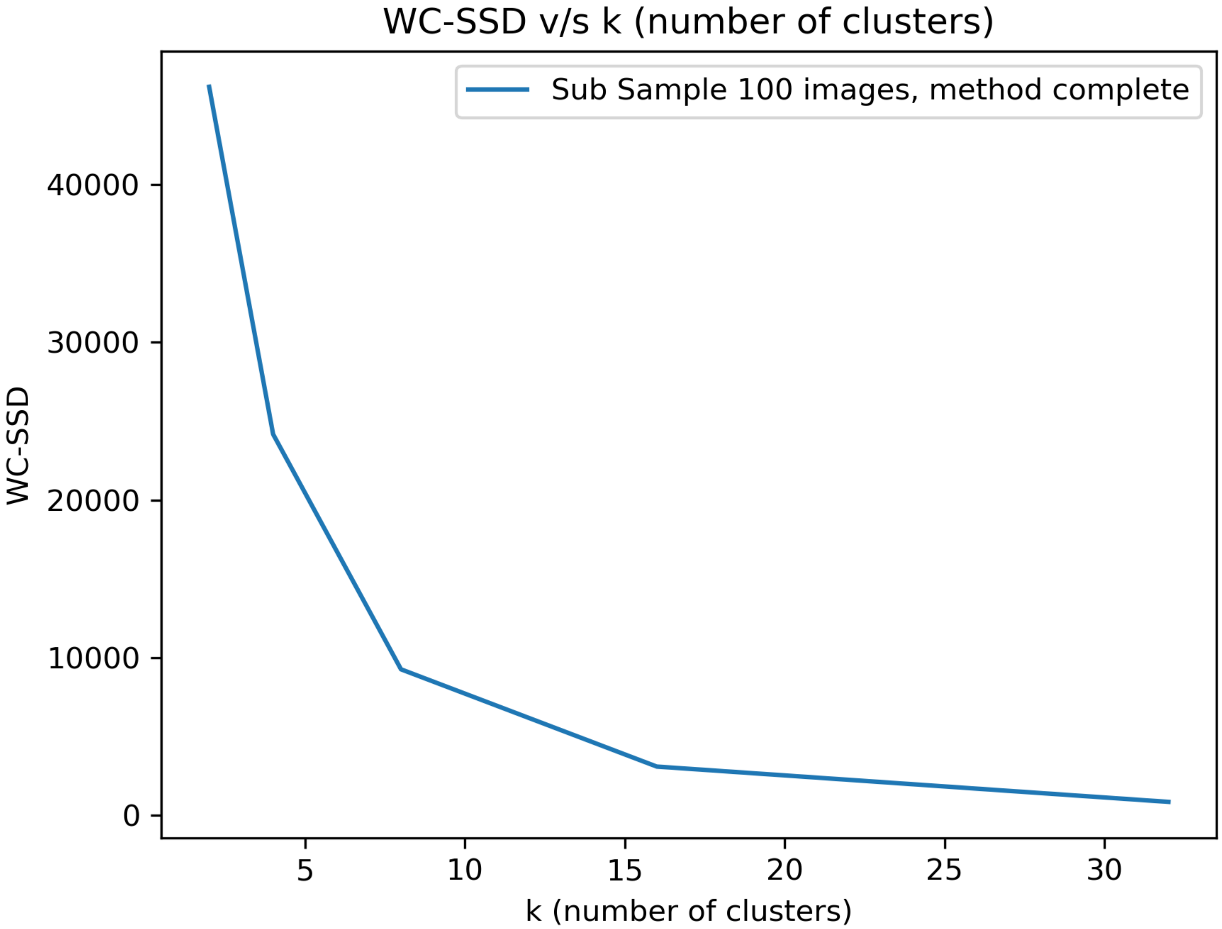

Figure 23: WC SSD vs K

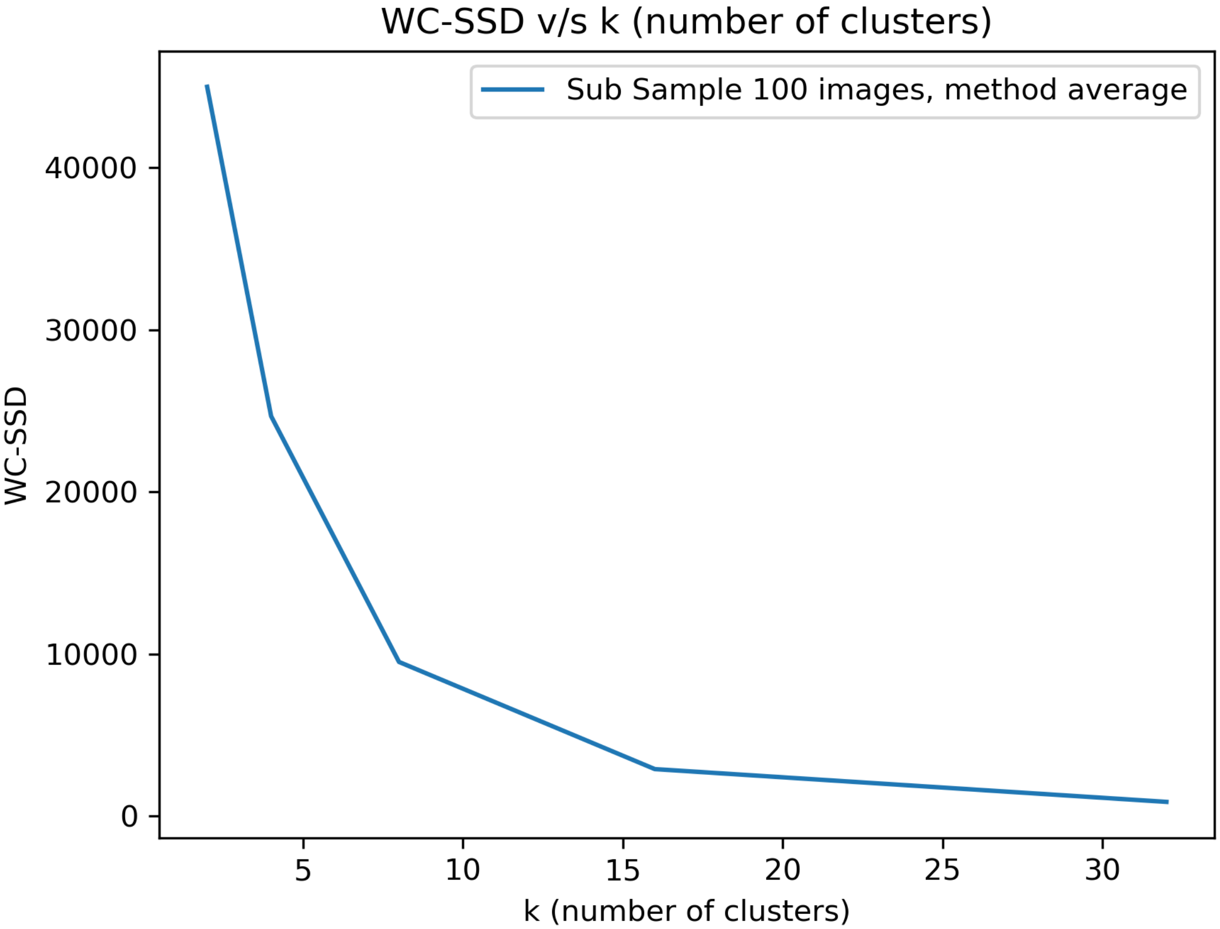

Figure 24: WC SSD vs K

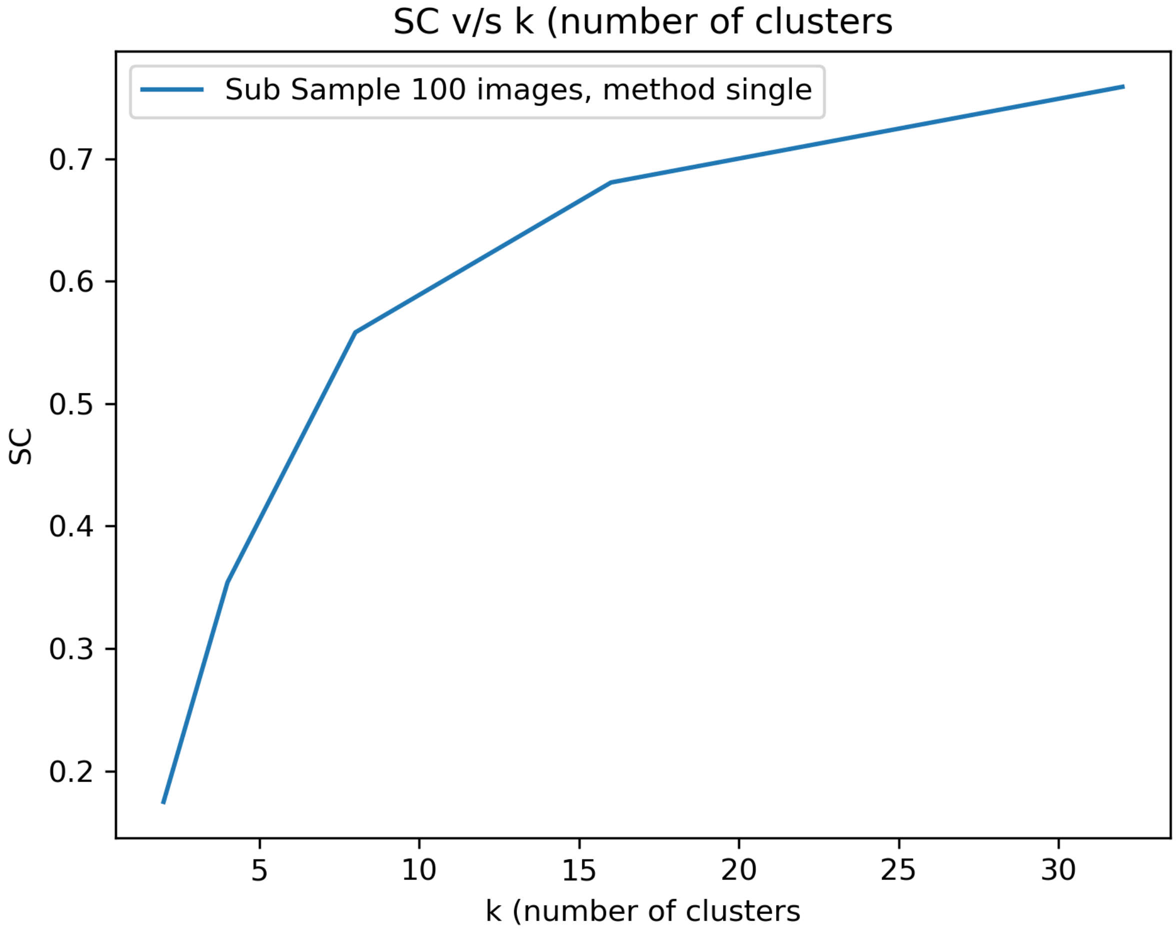

Figure 25: SC vs K

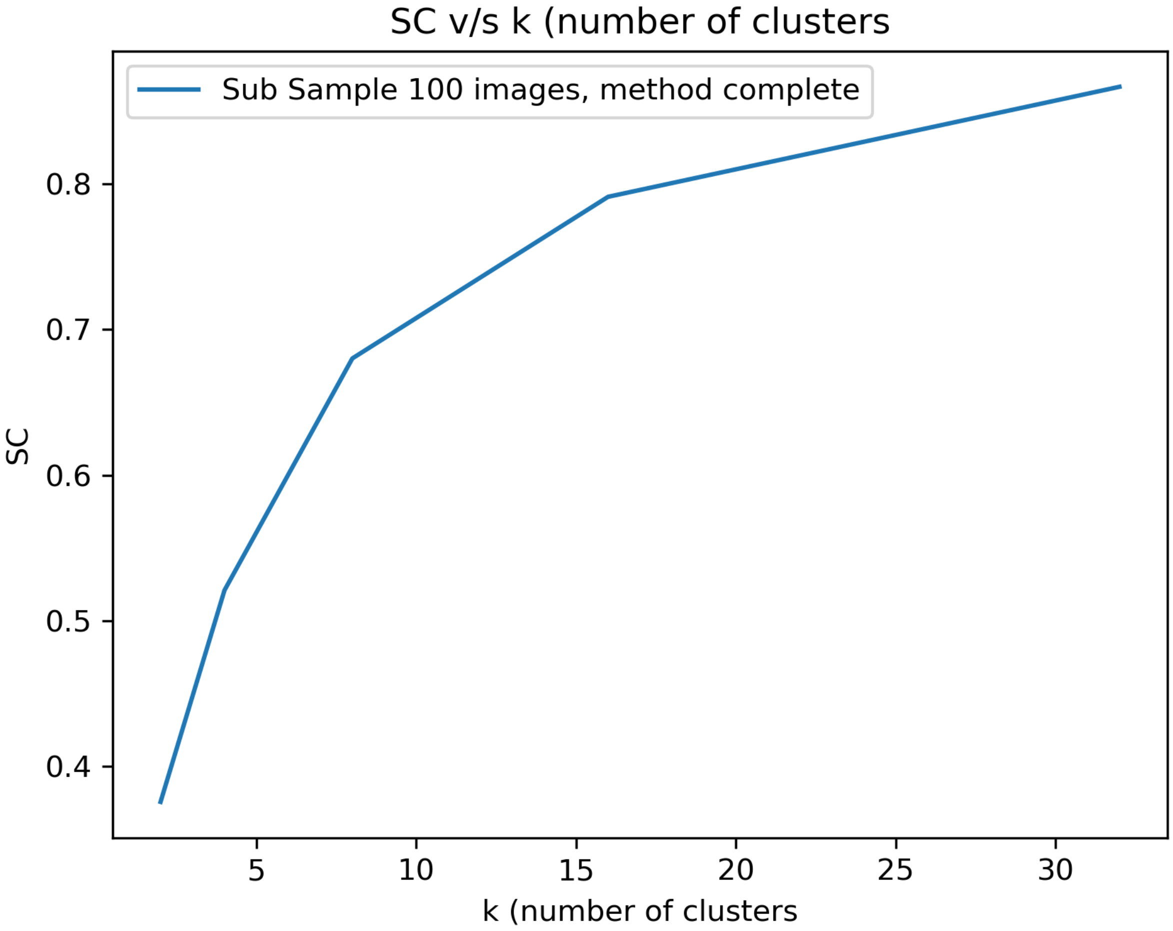

Figure 26: SC vs K

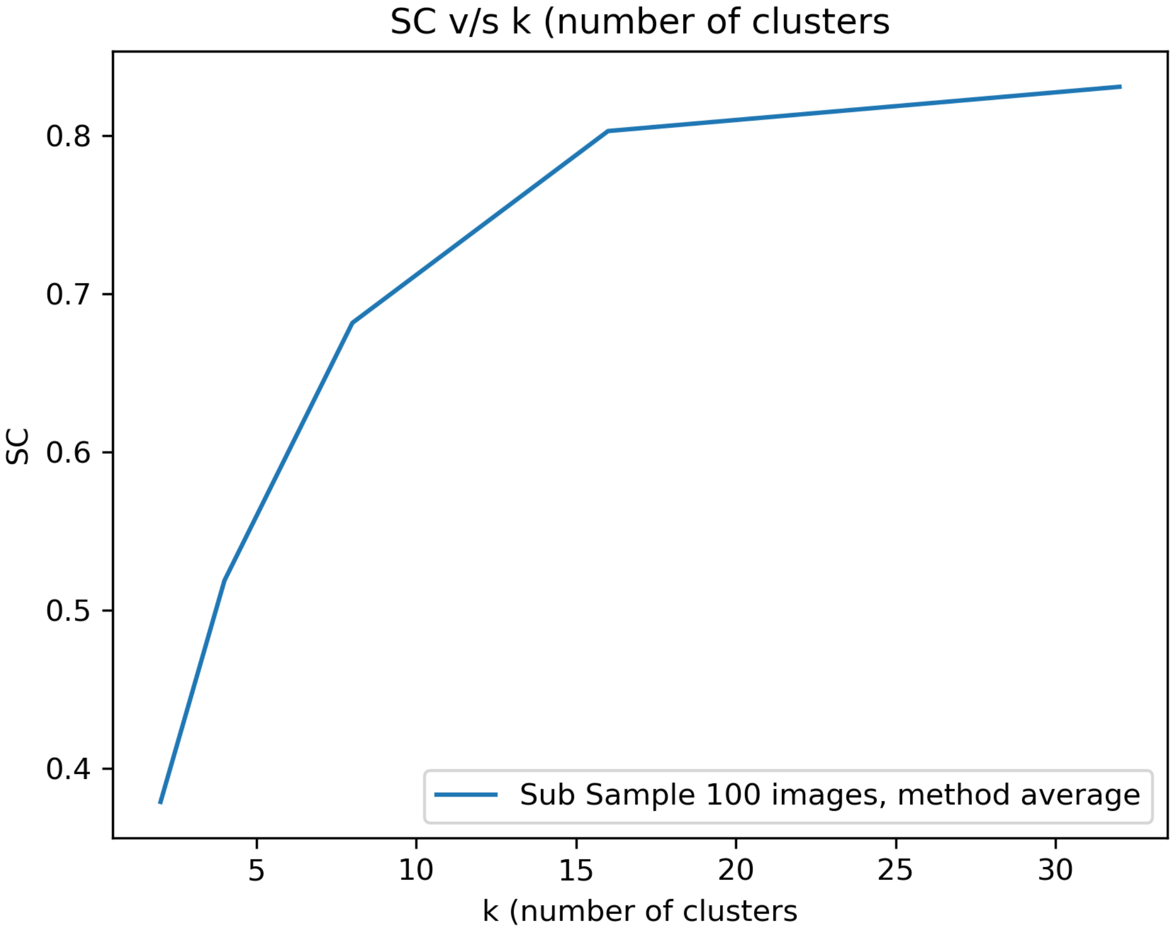

Figure 27: SC vs K

3.4 Discuss what value you would choose for K (for each of single, complete, and average linkage) and whether the results differ from your choice of K using k-means for Dataset 1 in Section 2

From the figures we can observe that in plots for WC-SSD vs k, we see elbowing at k=8 i.e. there is a relatively very sharp decrease till k=8 in WC SSD but then there is a relatively smaller decrease from k=8 to k=16. Thus, we choose k=8 for all single, complete and average linkages in hierarchical clustering for dataset 1.

The results are the same as k-means algorithm for dataset 1 in section 2. There too, we chose k=8 out of [2,4,8,16,32]. This is expected as out of the given k, k=8 seems to be most reasonable choice for clustering dataset 1.

3.5 For your choice of K (for each of single, complete, and average linkage), compute the NMI with respect to the image class labels. Discuss how the results compare across distance measures and how they compare to the results from k-means on Dataset 1 in Section 2

The NMI values are given below:
For method single NMI: 0.316043793012
For method complete NMI: 0.359605889127
For method average NMI: 0.339063556355

We observe that the NMI values are comparable to NMI value for k-means algorithm which was 0.346739899006 for k=8 as seen in section 2.2.4.
This means that the clustering quality is similar for both hierarchical clustering and k-means clustering. Though, we also observe that complete linkage method has the best NMI score of around 0.36 which is better than k-means NMI score of around 0.35.

Across other evaluation measures, we observe that WC SSD value for single linkage is around 20000 as seen in figure 22, 10000 for complete and average linkages as seen in figure 23, 24. This is small in comparison to the value around 1887699 observed for kmeans for dataset 1. We note that this is mostly because in kmeans we are taking the entire set of 20000 points. In Hierarchical clustering, we just take a sub sample of 100 points.