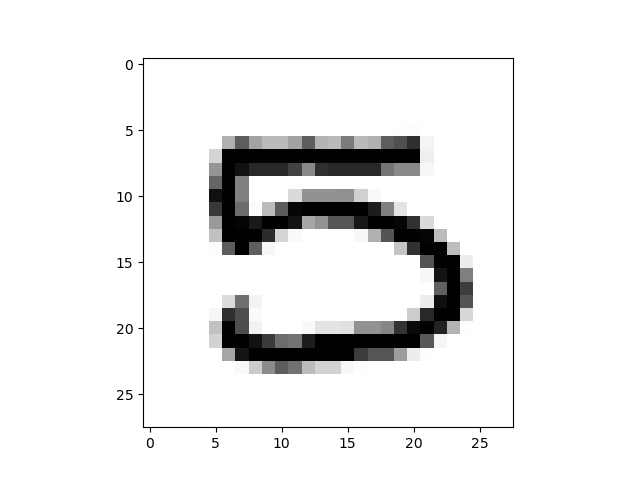
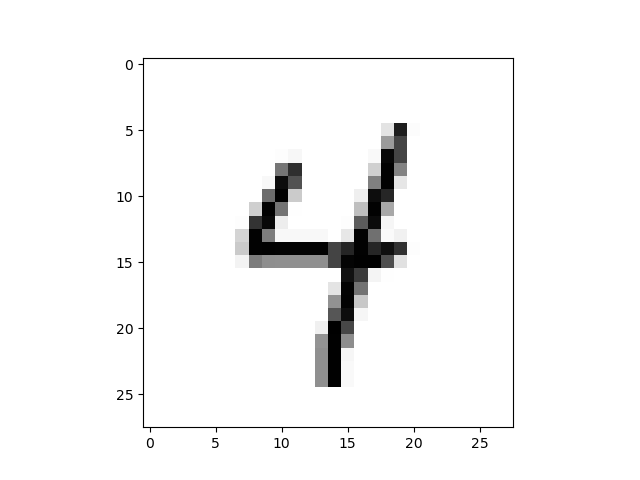
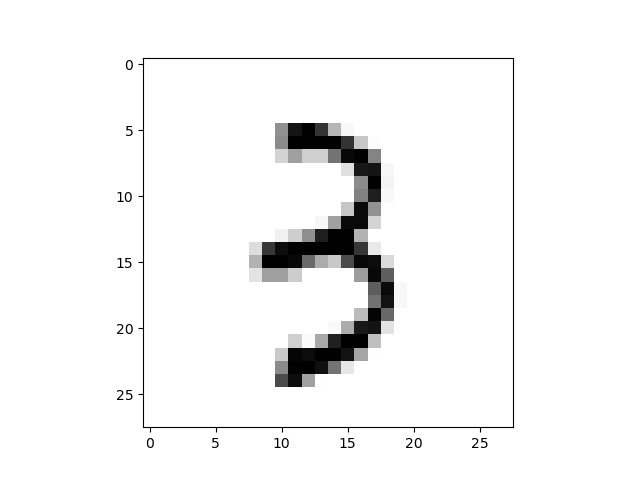
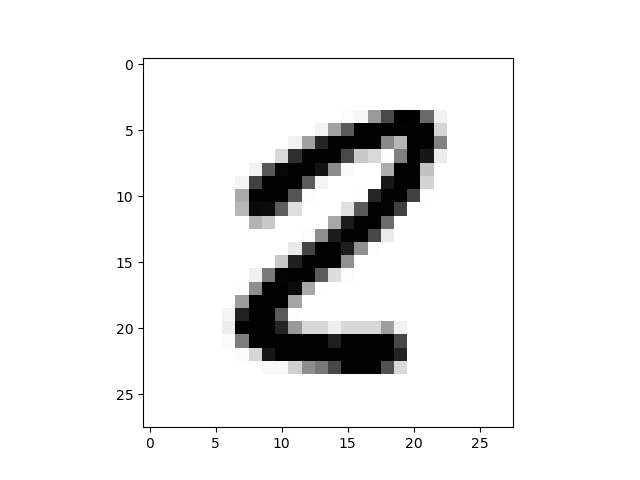
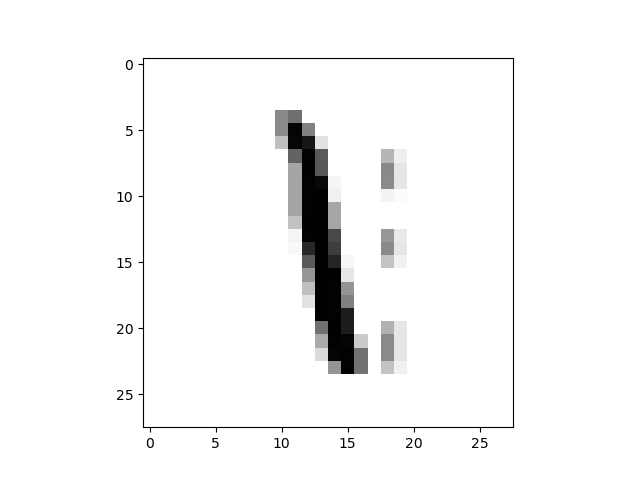
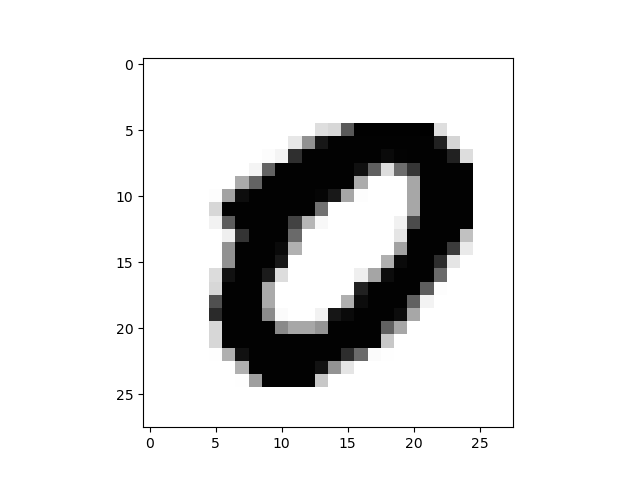
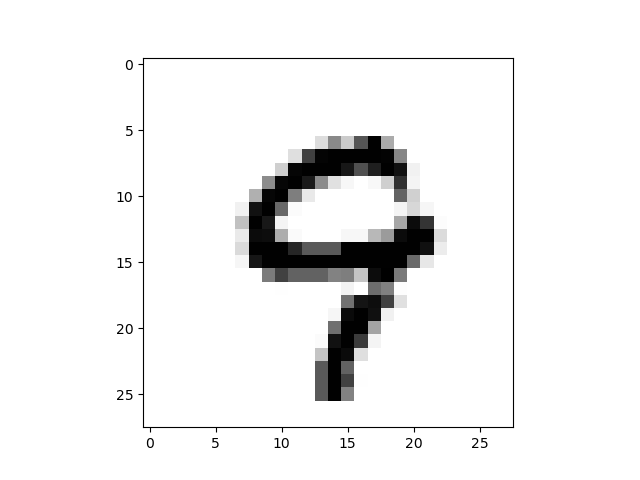
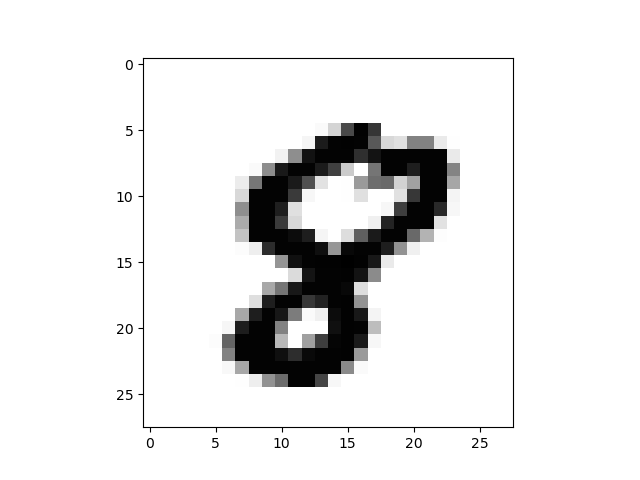
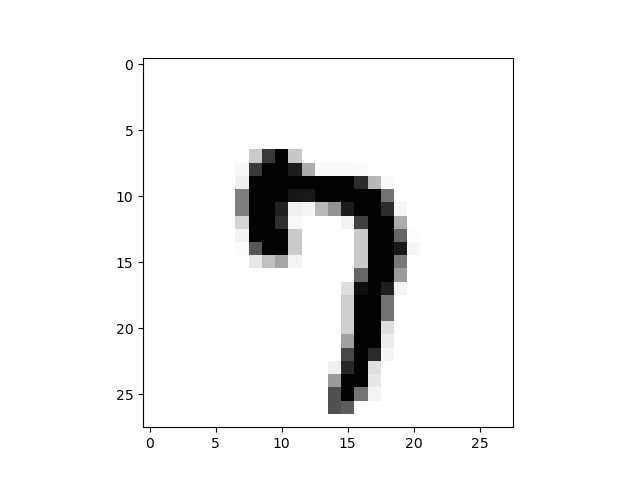
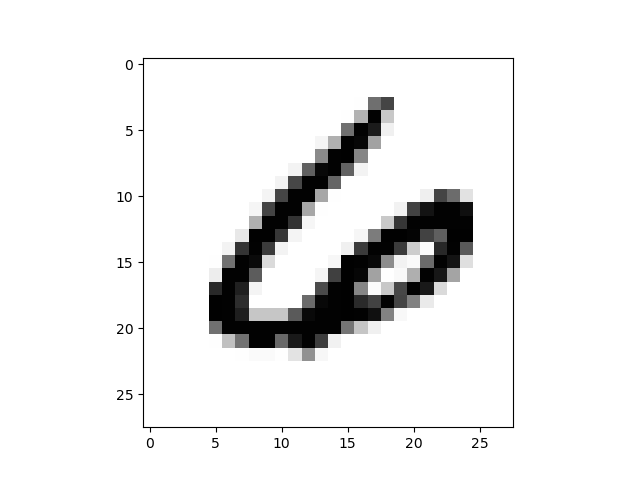
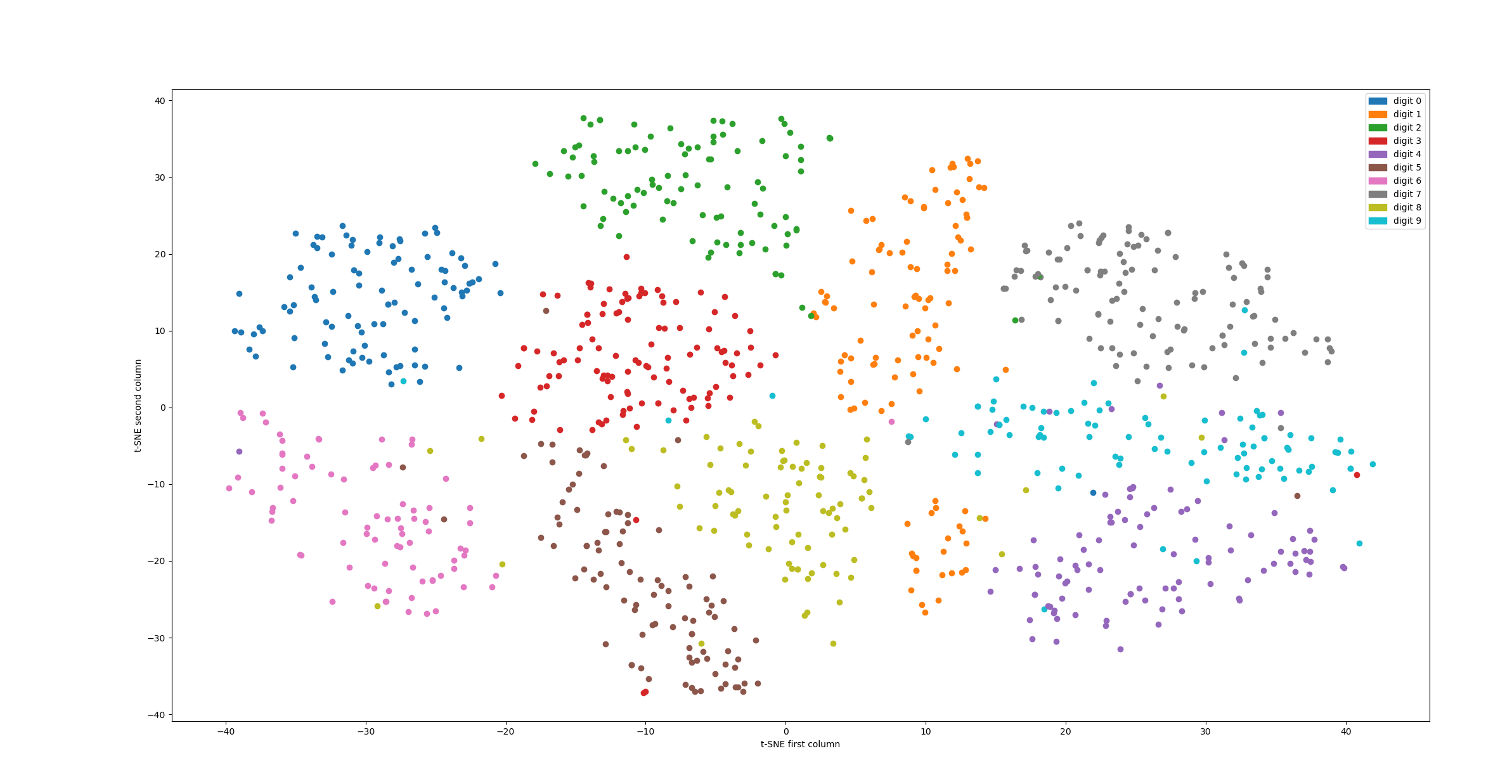
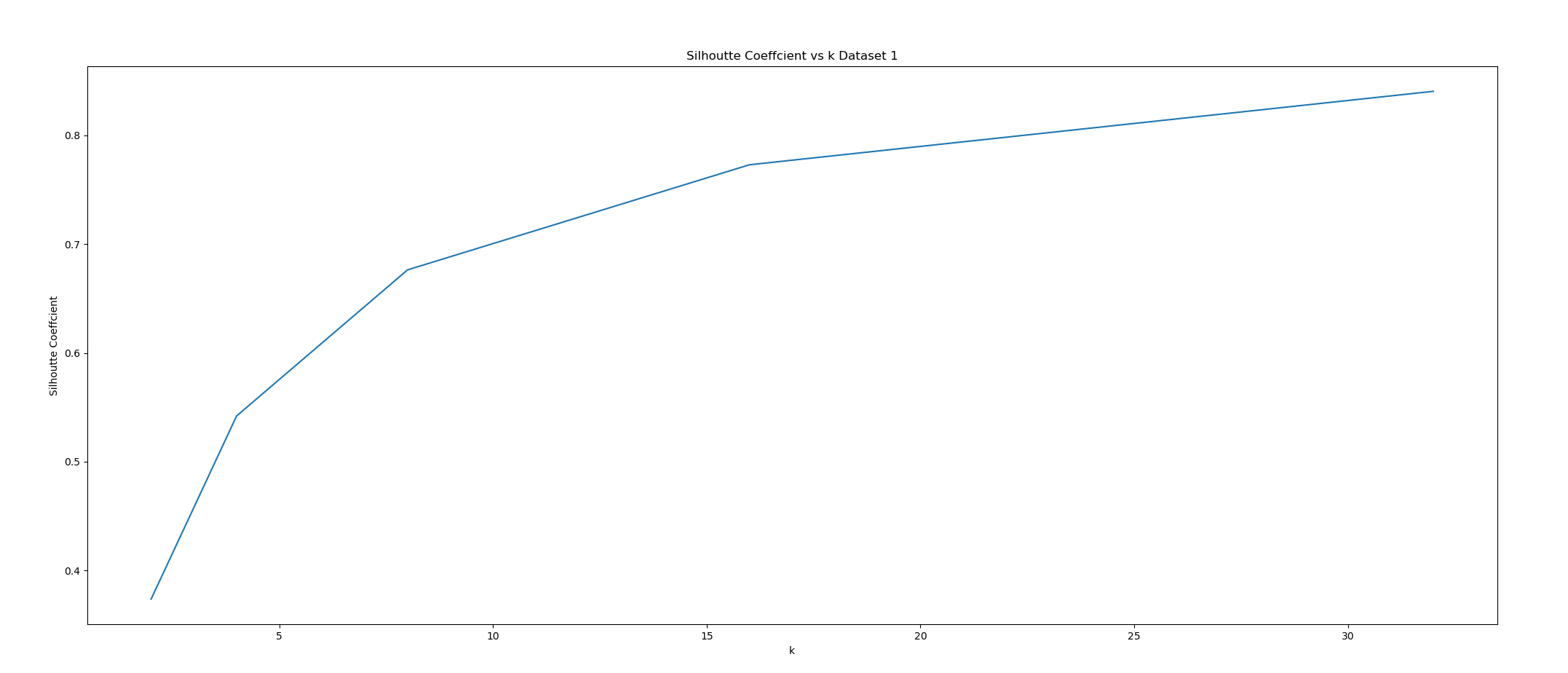
Q1.1

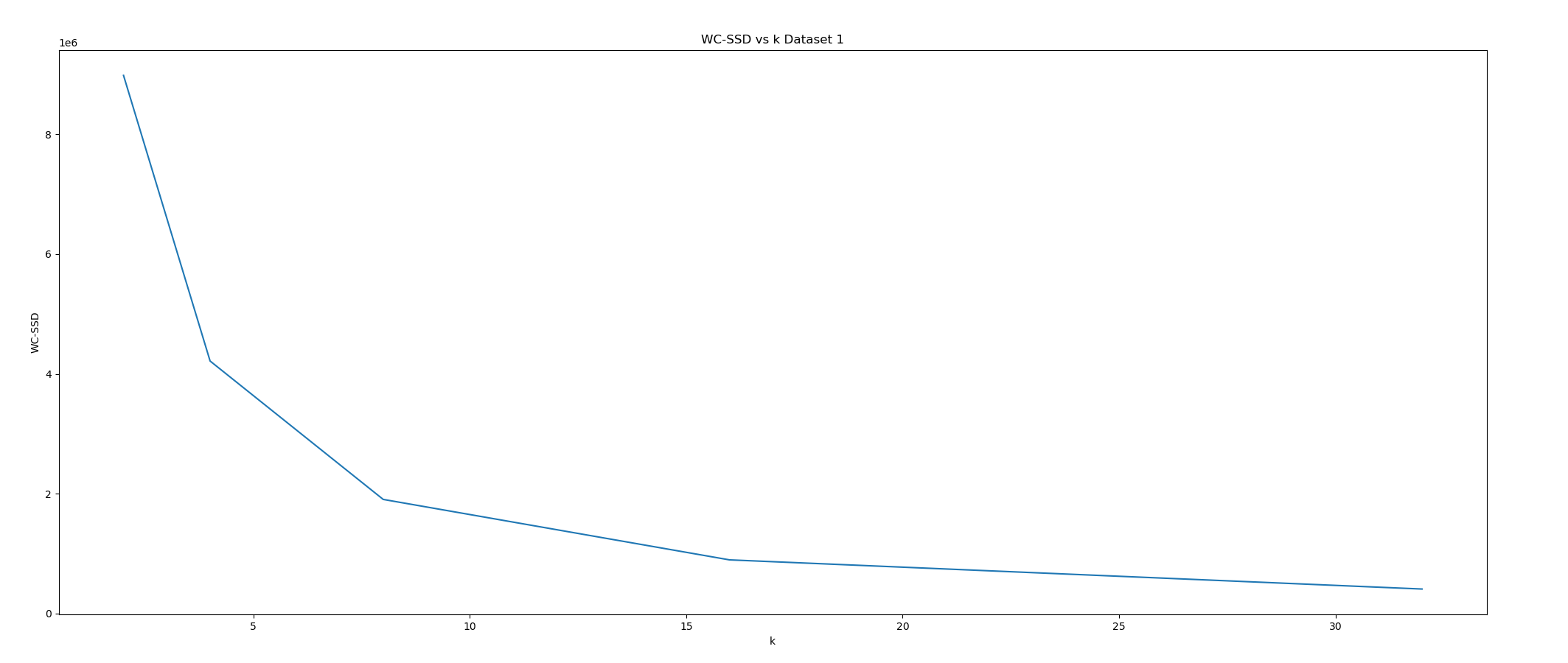


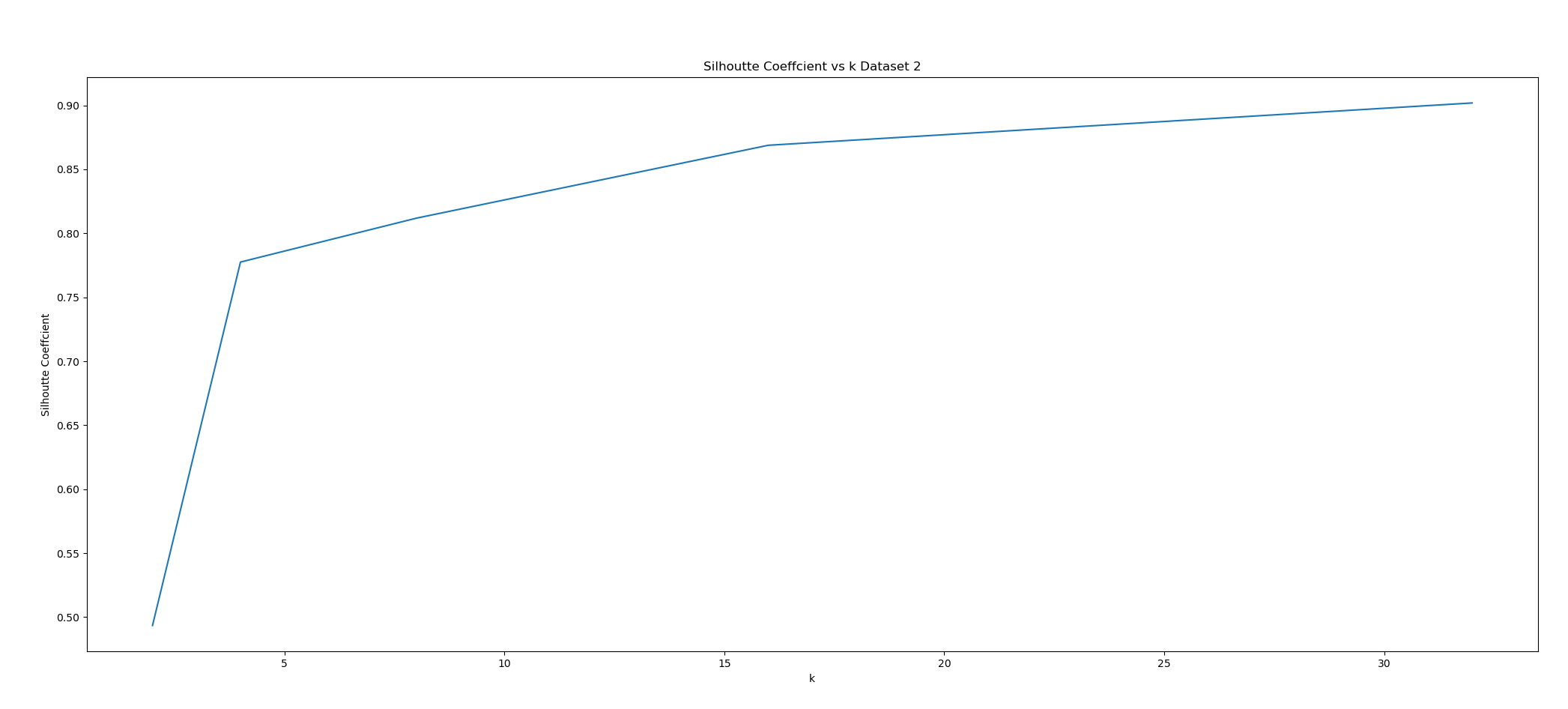
Q1.2

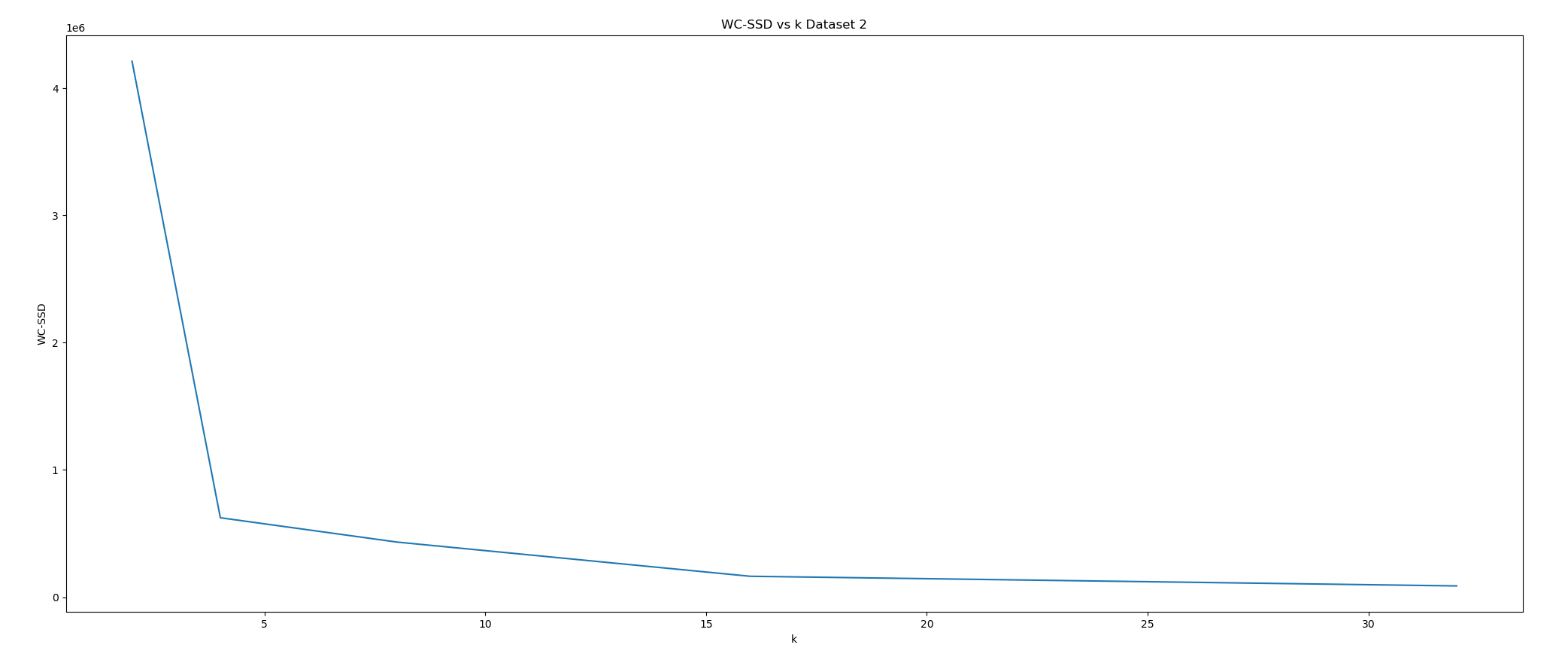


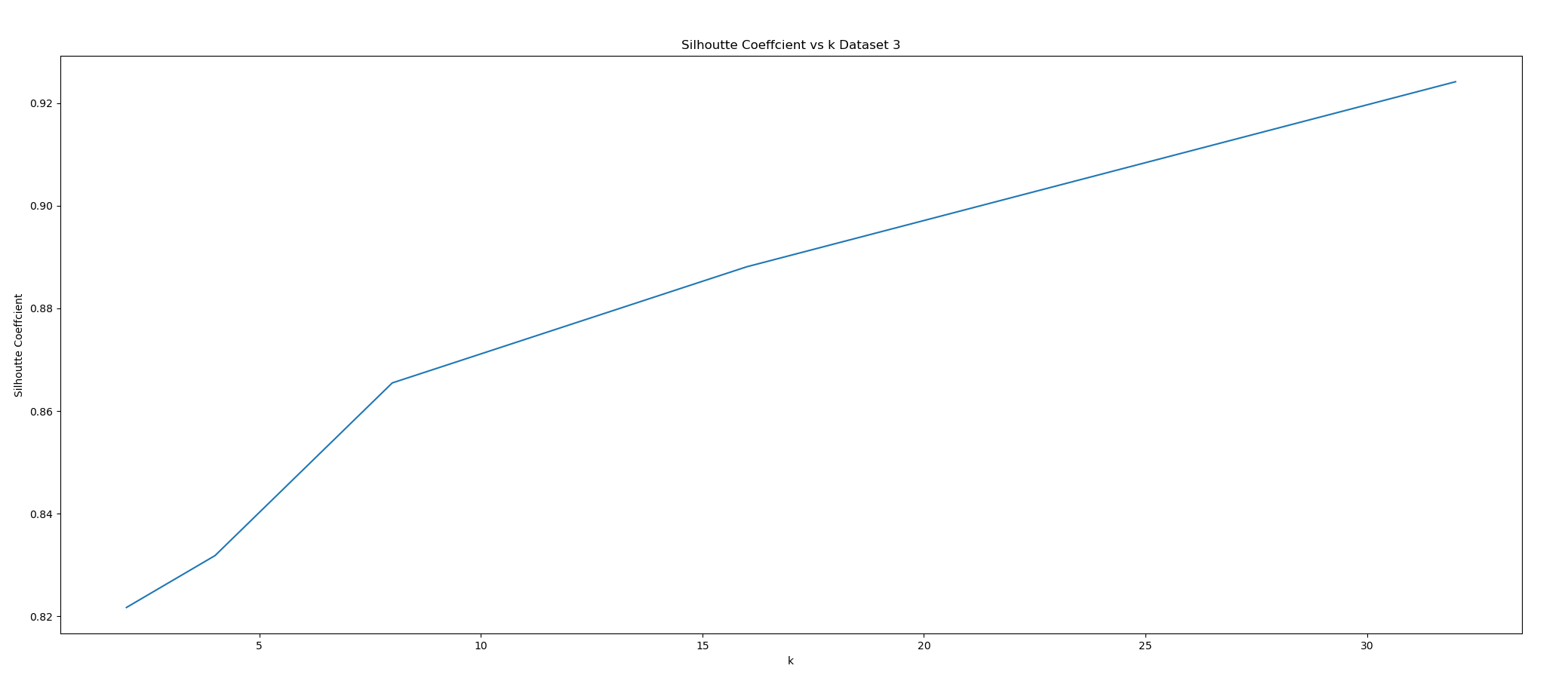
Q2.2.1

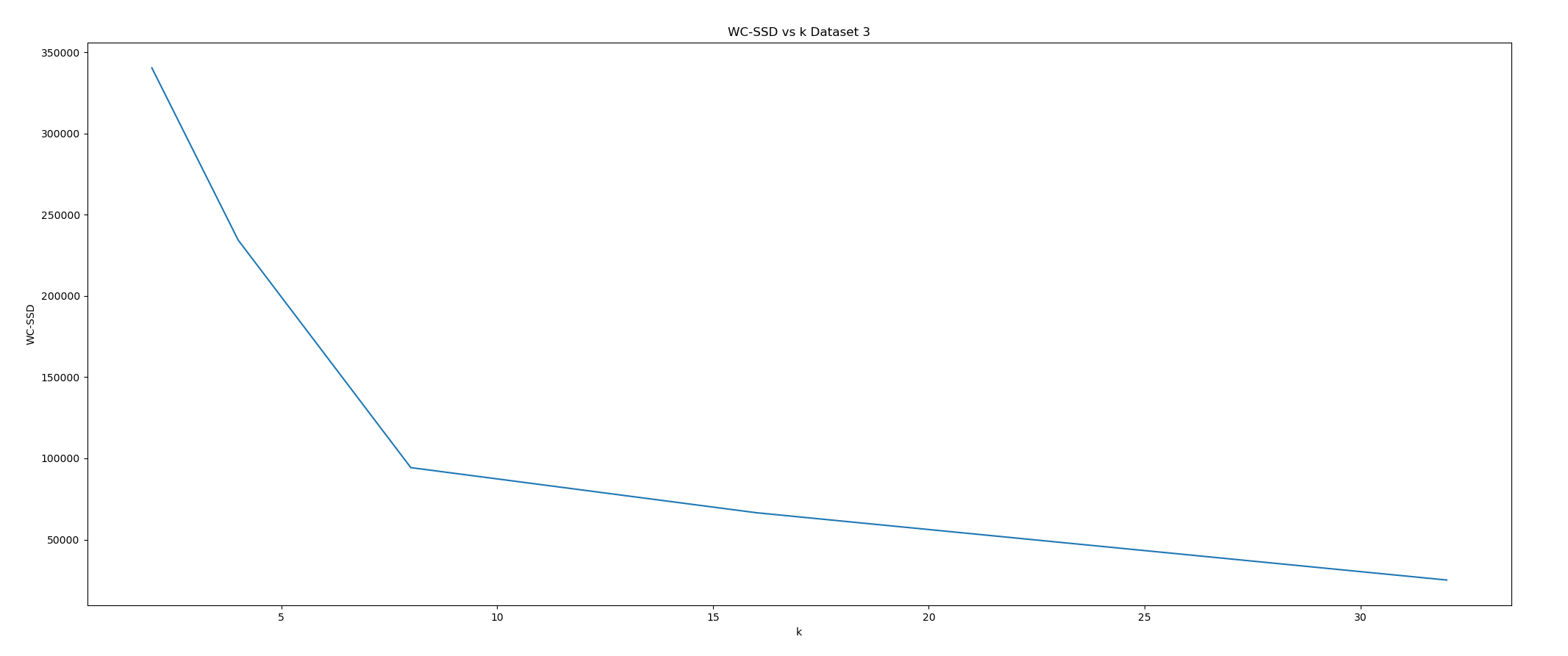












Q2.2.2

We evaluate the elbow and knee point of WC-SSD and SC plots respectively, for each dataset. Beyond those points, the magnitude of gradient of WC-SSD or SC curve drastically decrease.

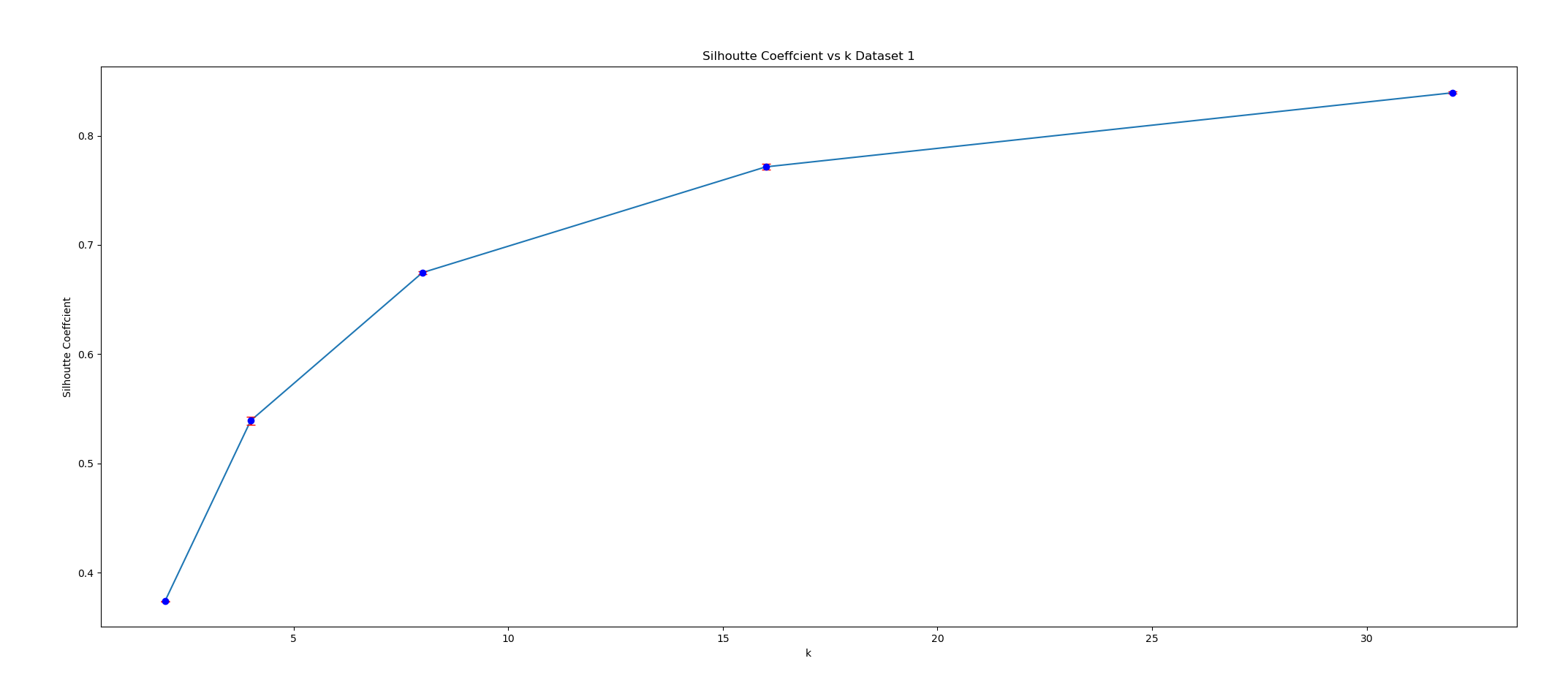
For dataset 1 we chose optimal K=16, for dataset 1 we chose optimal K=4, For dataset 3 we chose optimal K=8

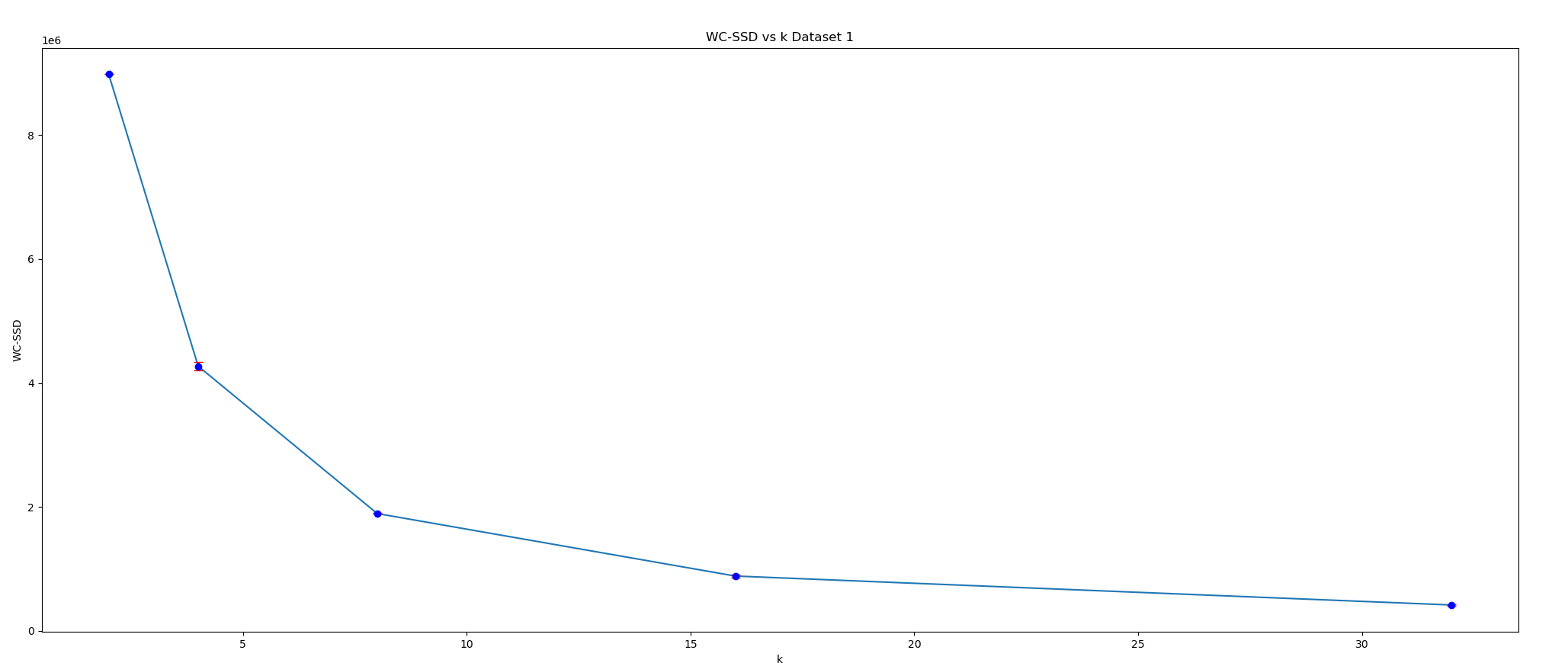
Discuss how the results compare across the two scores and the three versions of the data:

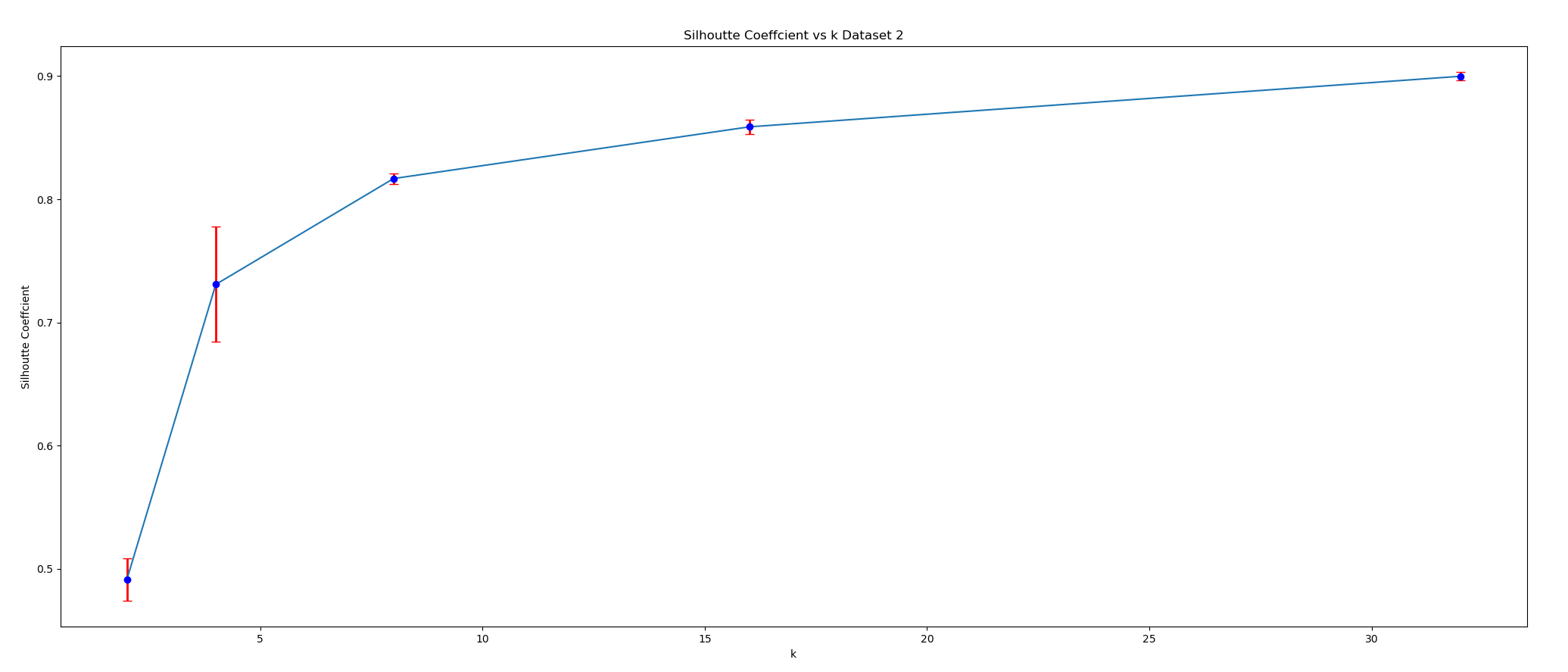
For the WC-SSD plots across the three datasets, its value decrease as the number of classes decreases. Th optimal K in dataset 1 and 2 are somewhat close to the real number of classes, while the optimal K in dataset 3 is not very close to the real number of classes, may be resulted from the uniform distribution of those sample spots in their cluster (loose cluster) which cause some ambiguity in k-means clustering evaluated upon WC-SSD.

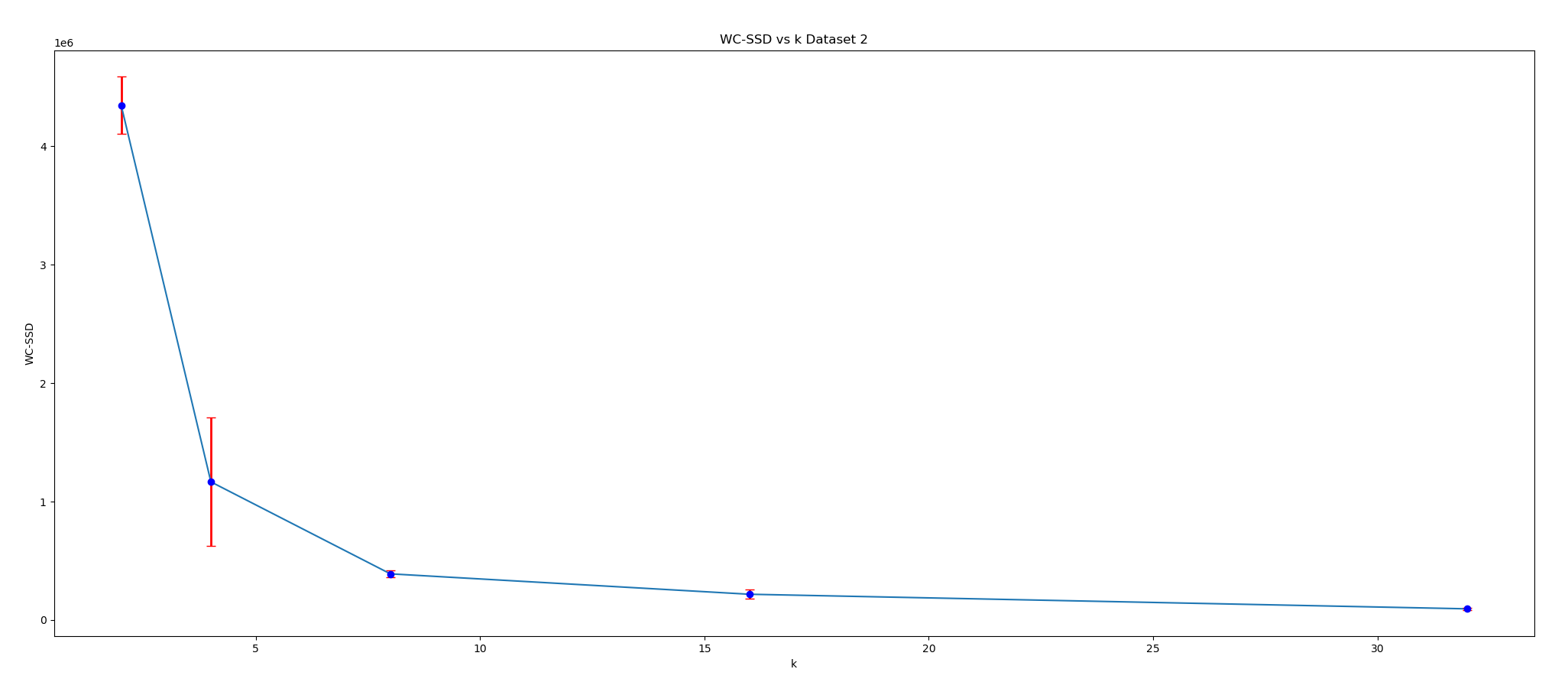
Q2.2.3

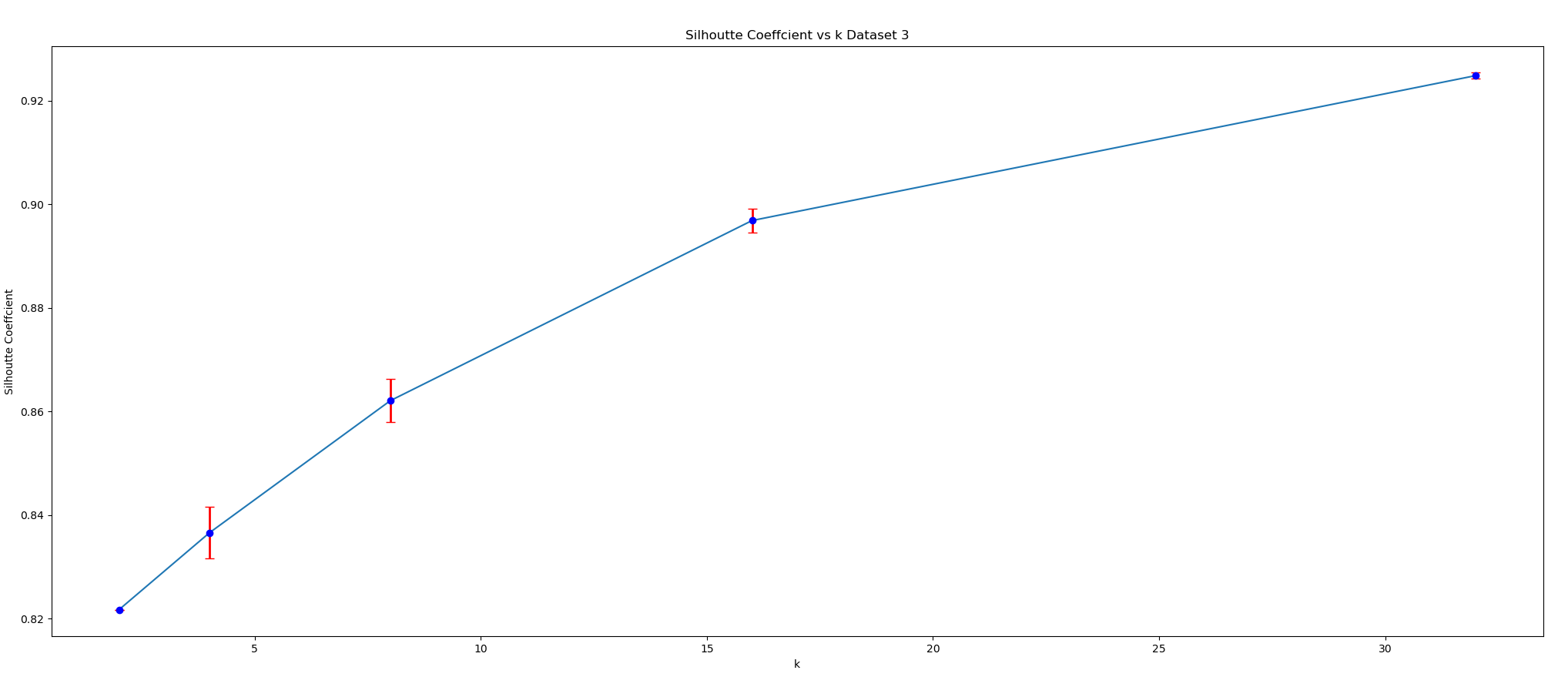
Please see the following plots

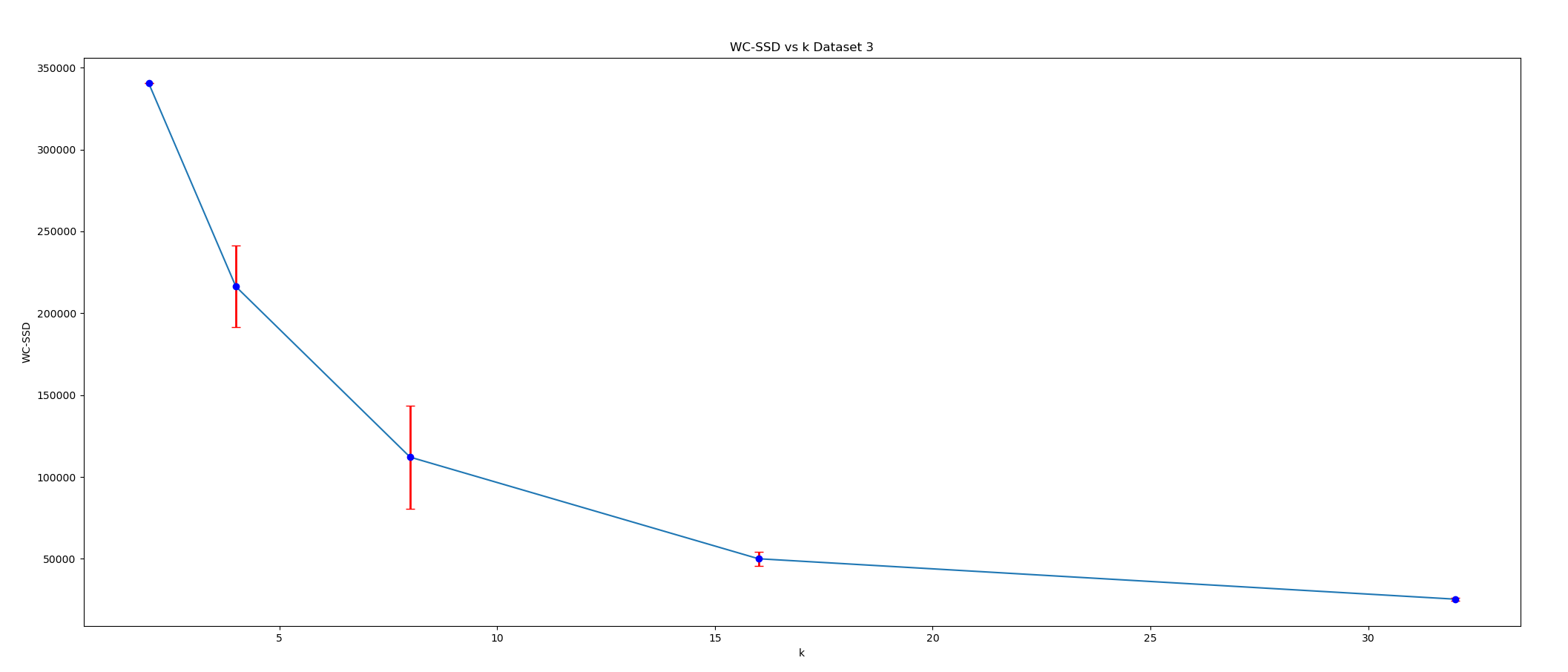












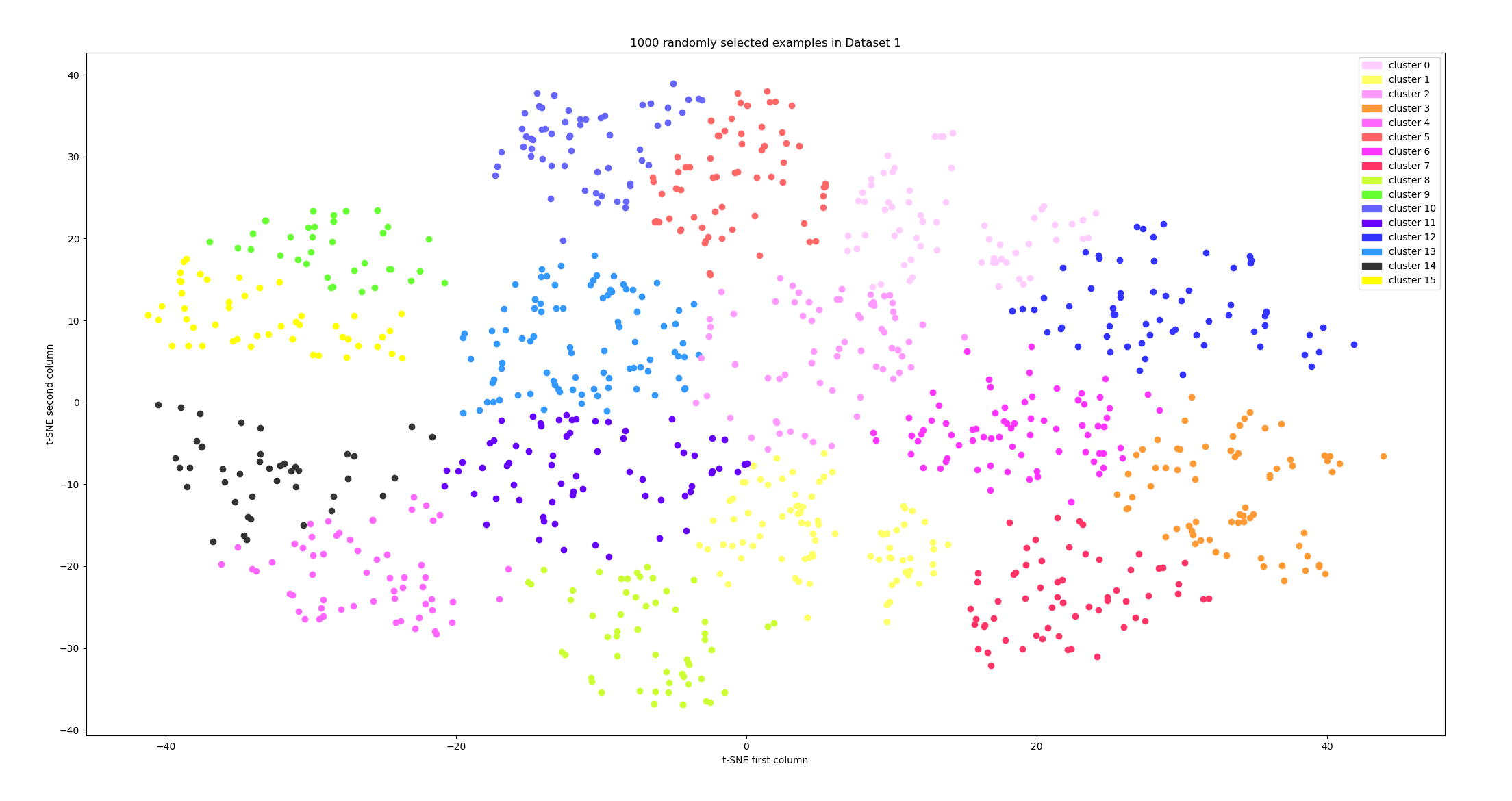
The plots show that kmeans is sensitive to the initial starting conditions. In each dataset the SC and WC-SSD plots show high standard deviation.

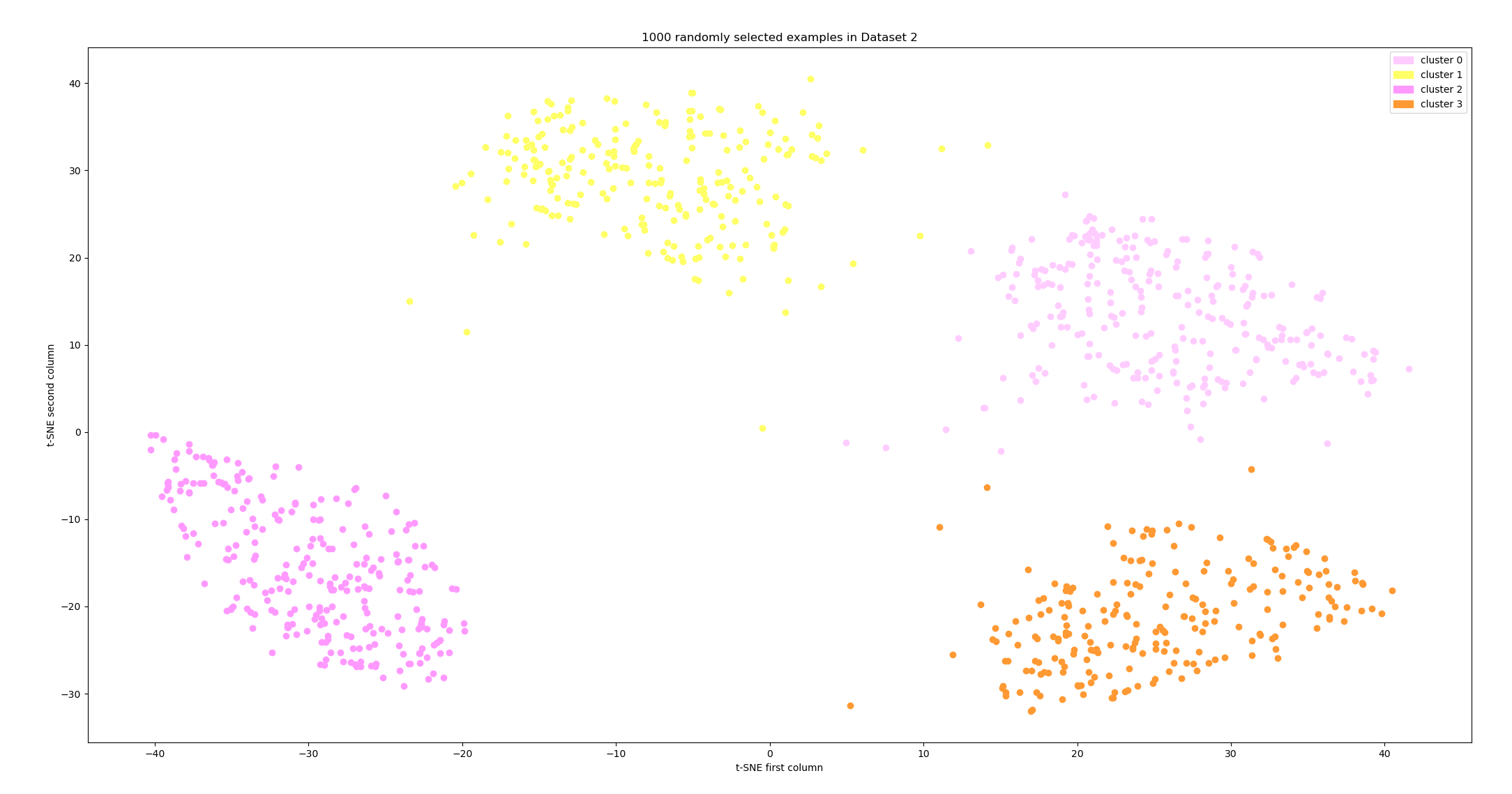
Q2.2.4

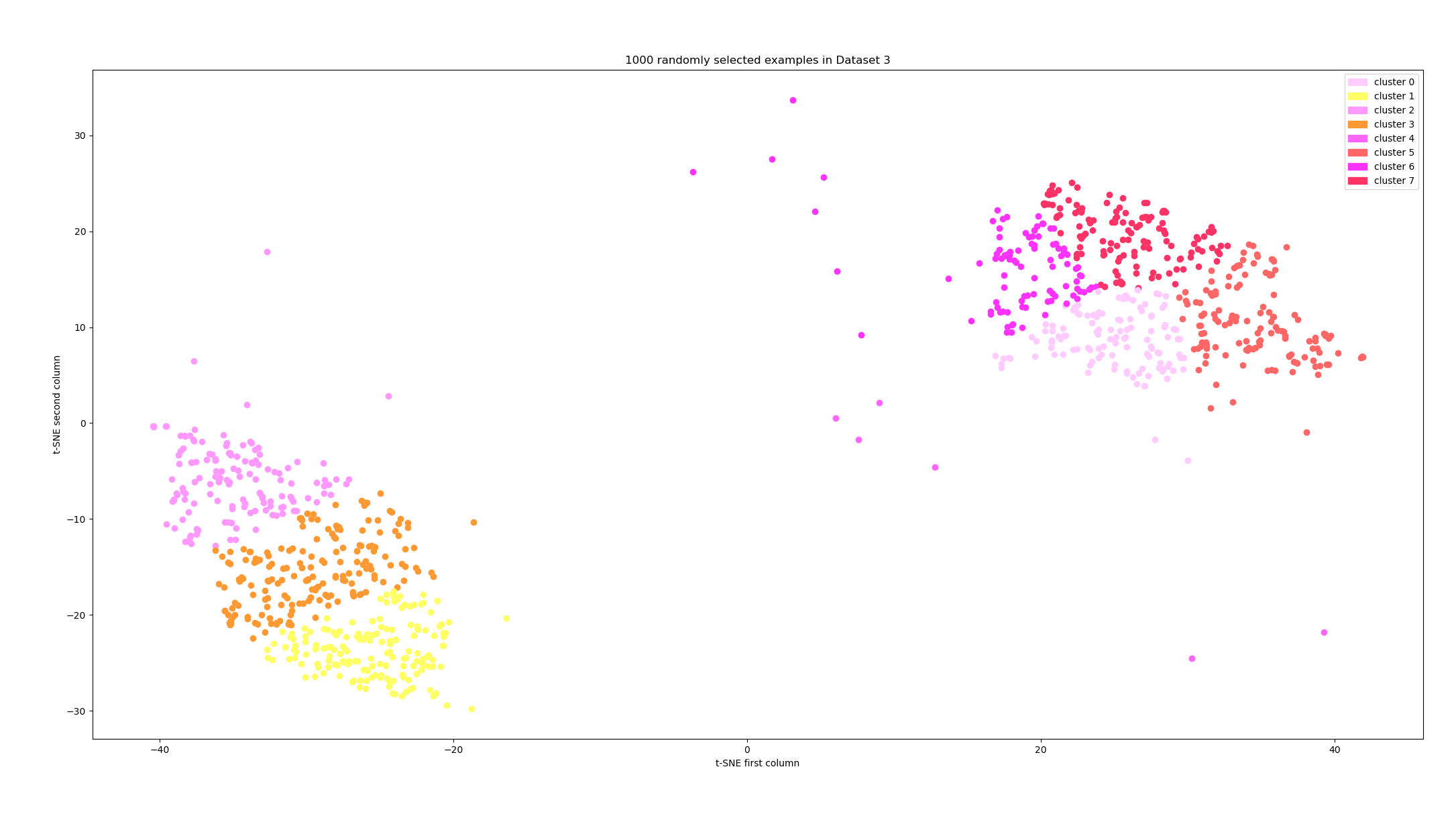


K = 16,4,8 for dataset 1,2 and 3

See the following graph

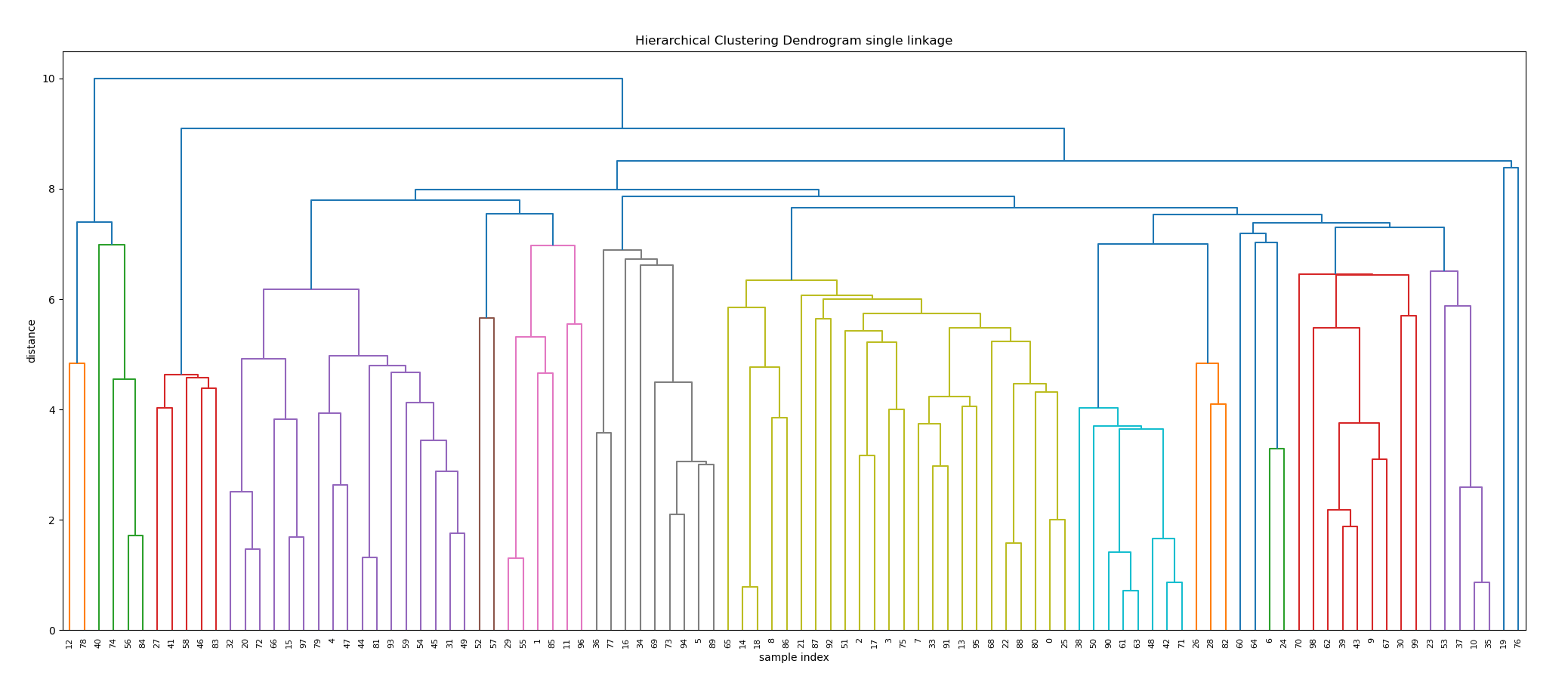


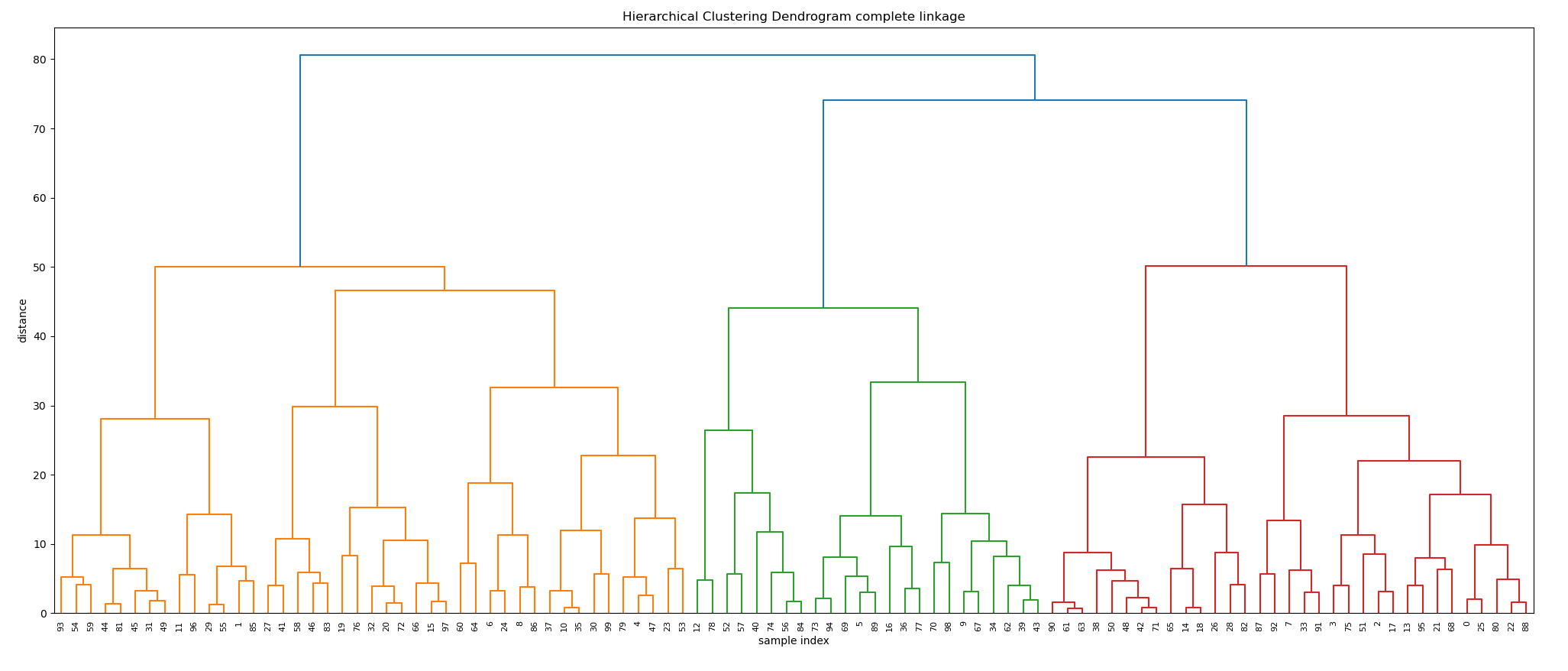


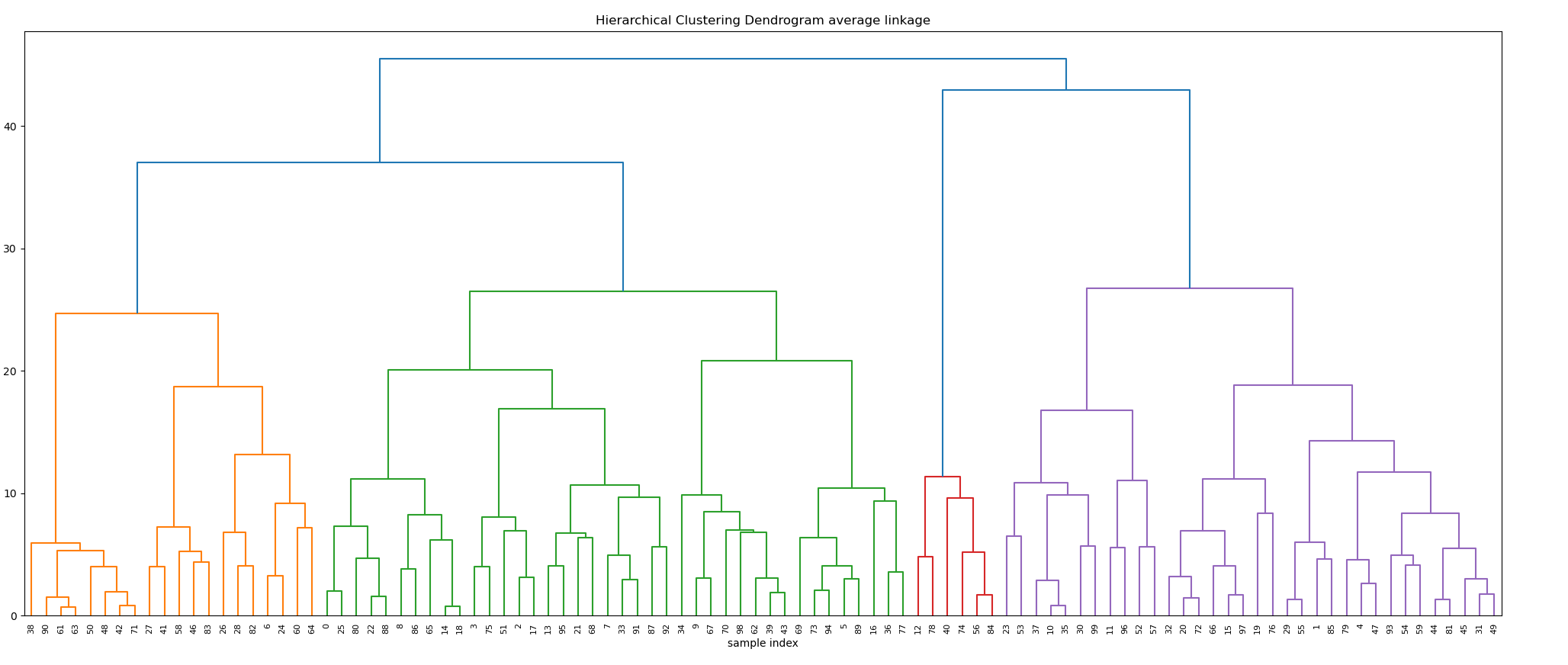


In general the NMI value should be larger if the number of clusters is smaller. In our case, dataset2 has the highest NMI because it has 16 clusters, but for dataset 3 since the clustering is far away from real label numbers, the NMI is a bit abnormal. In the scatter plot, the clusters are more distinguishable with less total cluster numbers. The most separable dataset is 2, then 3, then 1.

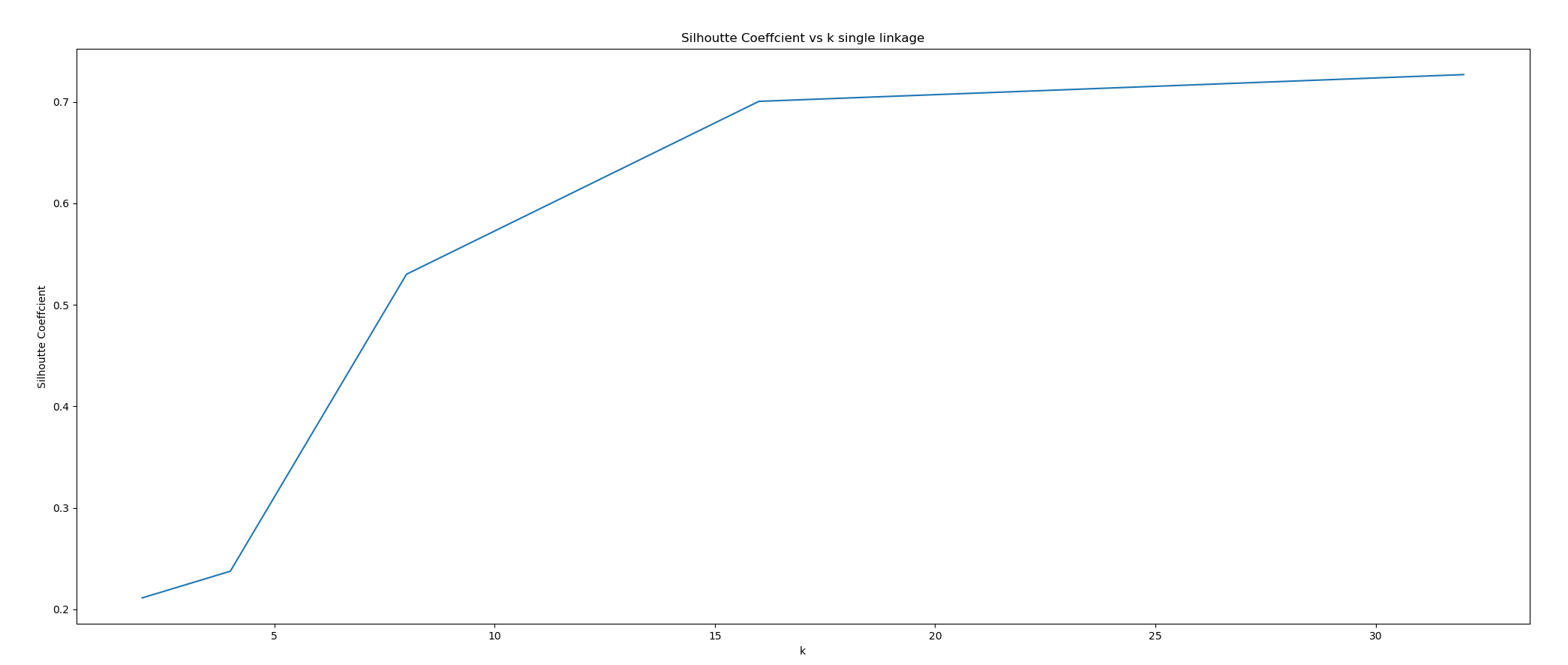
Q3.2, 3.3

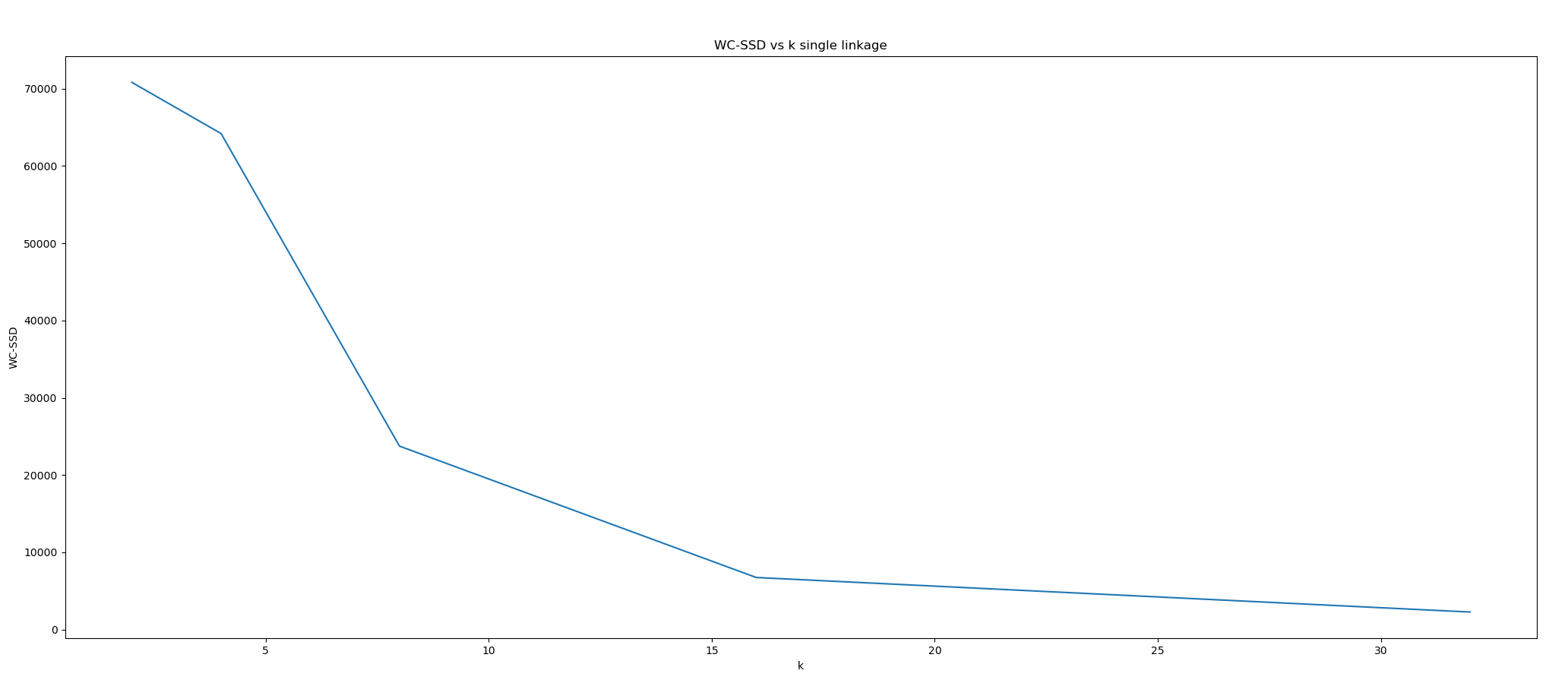


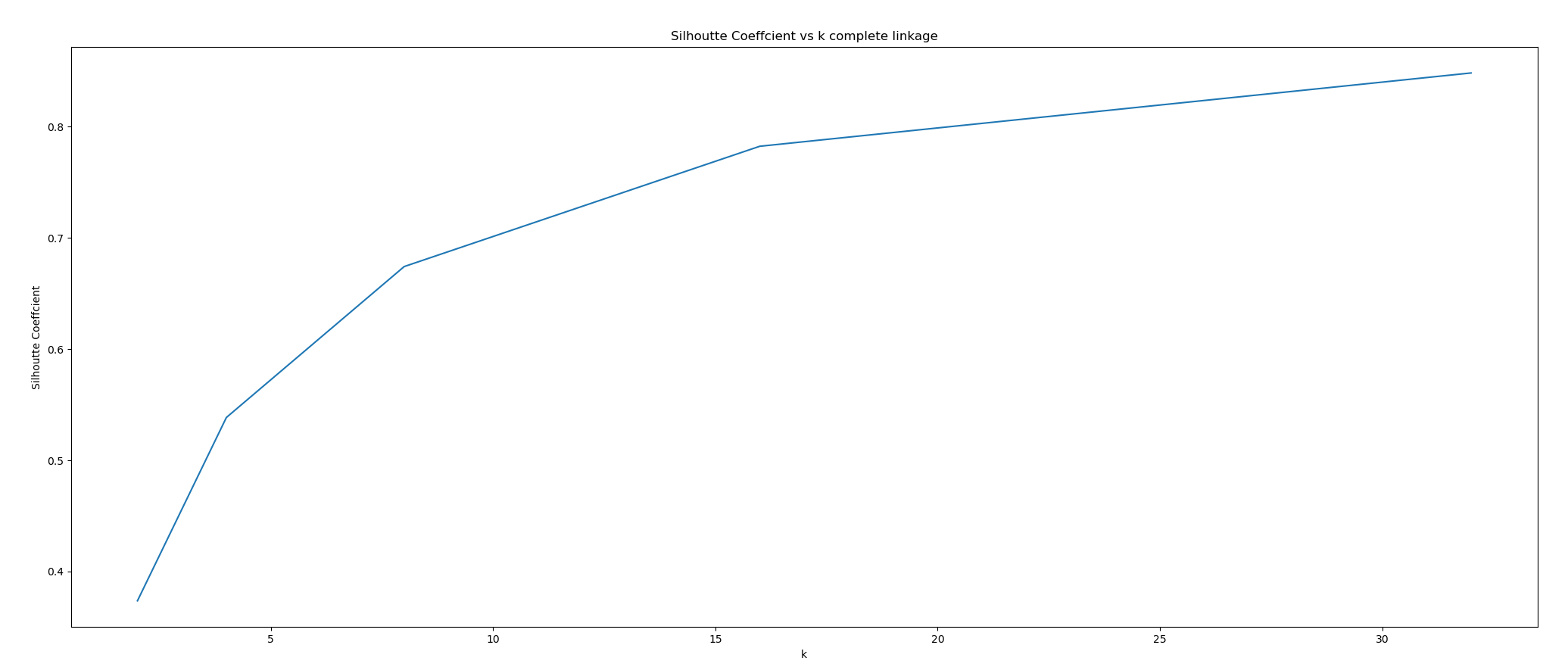


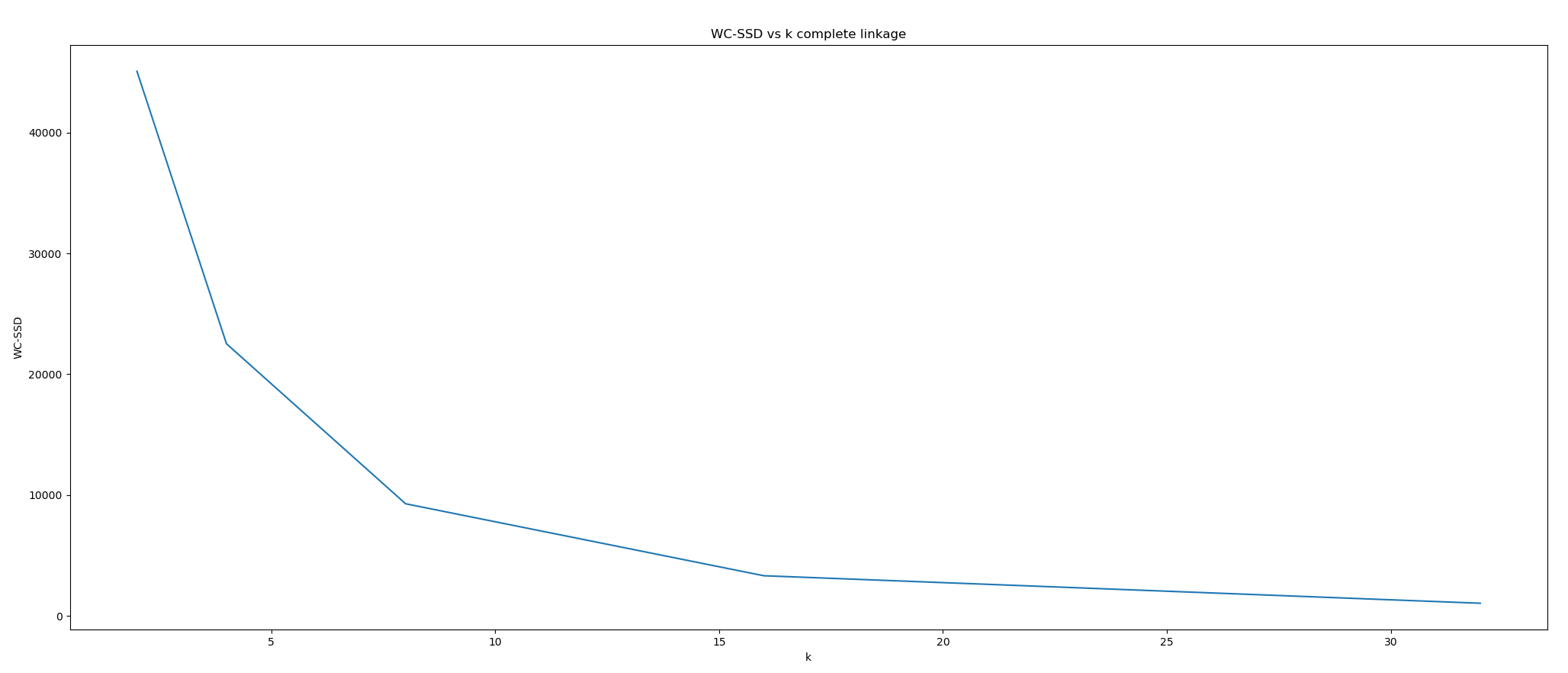


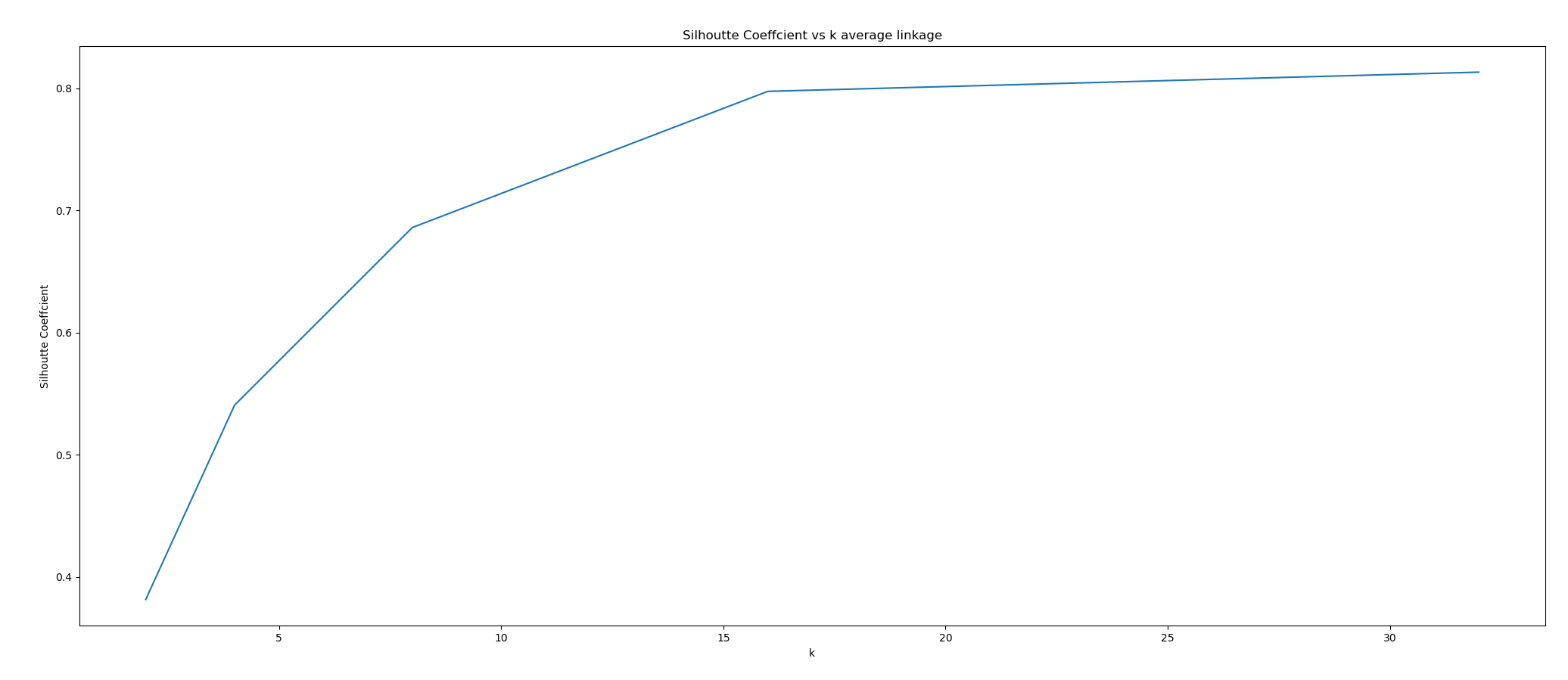
Q3.4

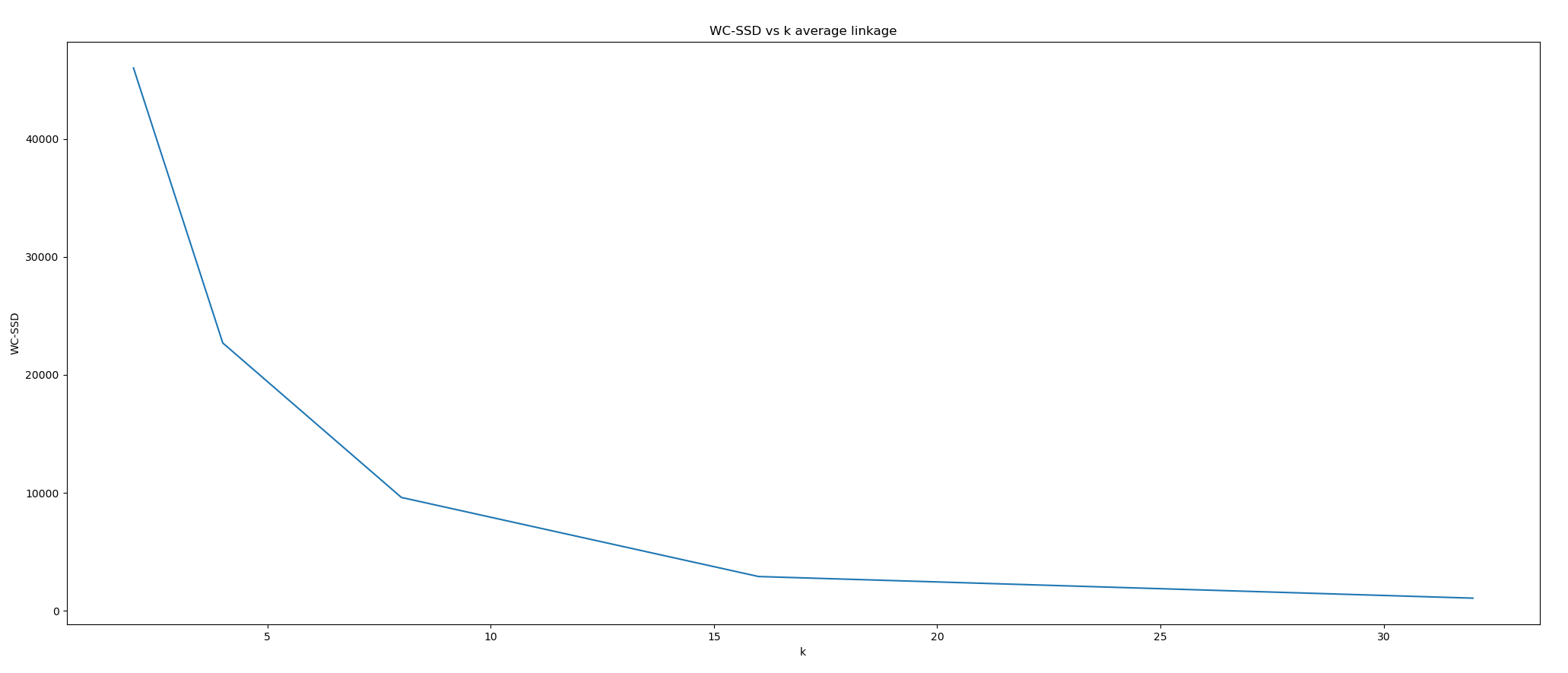












Q3.5

Using the same method used in finding optimal K in kmeans in previous question, we decide the optimal K is 16 for all three likages. The value is the same as what we found using k-means for Dataset 1 in Section 2.

Q3.6

With K=16 for all linkages

Output:

single NMI: 0.3545881948493385

complete NMI: 0.369599539861315

average NMI: 0.390785830258648

The NMI score using average linkage is highest, the second highest is using complete linkage, then single linkage. It seems that the difference between the three scores are not significant and the results will vary a lot with different runs. Their average value is higher than what we got for NMI in using k-means for Dataset 1 in Section2