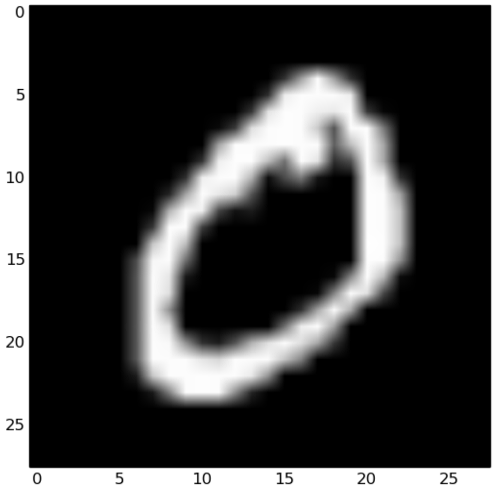
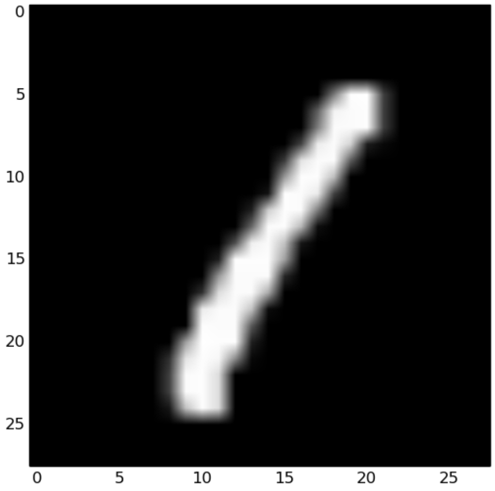
CS57300 - Data Mining

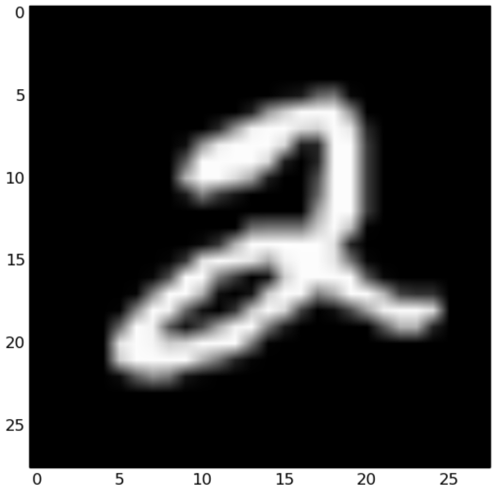
Homework 5

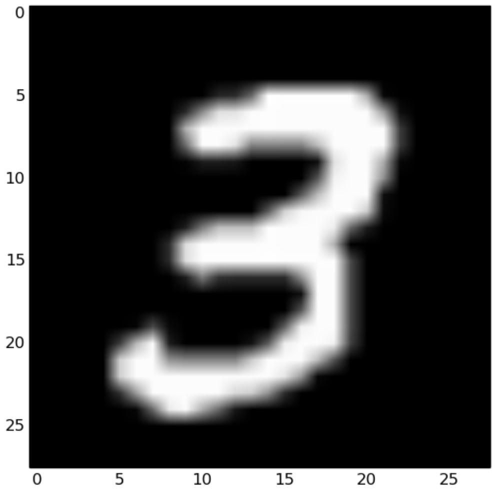
1. **A**

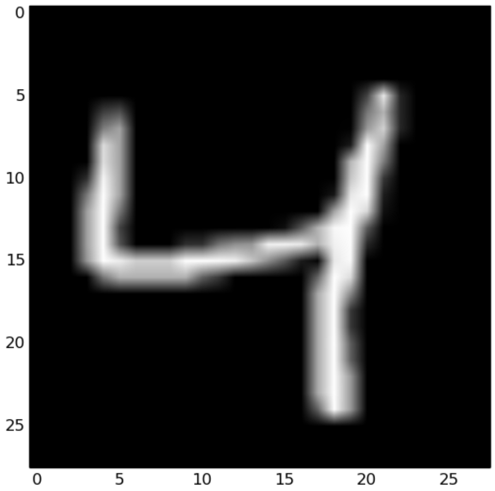
(a) The digits are as shown below:

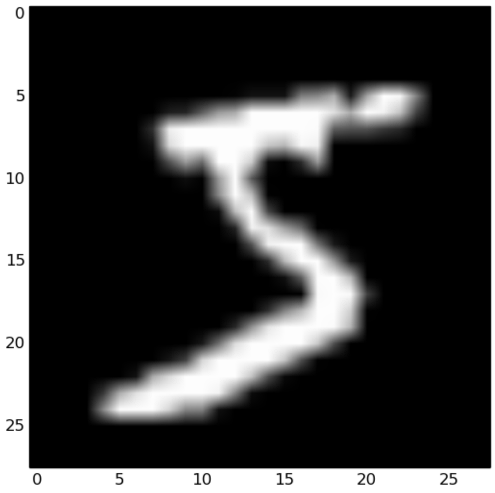


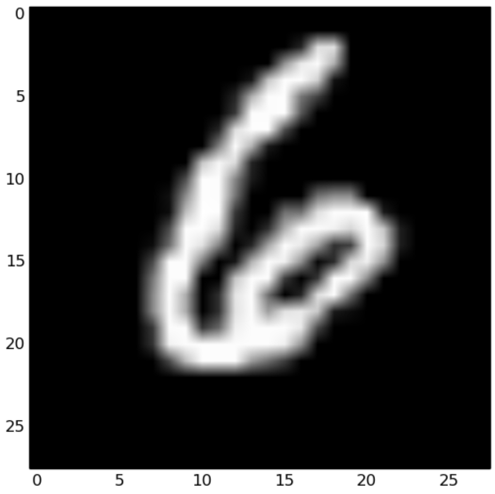


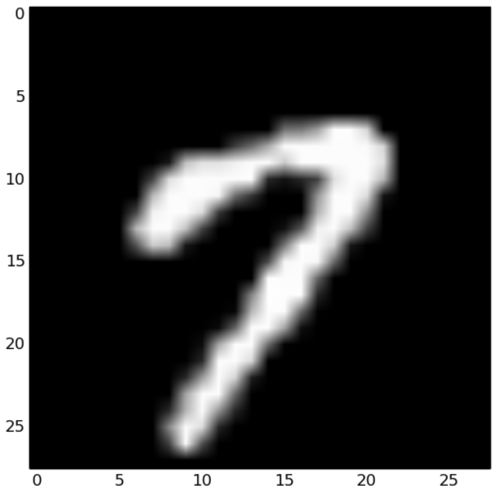

0
5
10
15
20
25
0
5
10
15
20
25

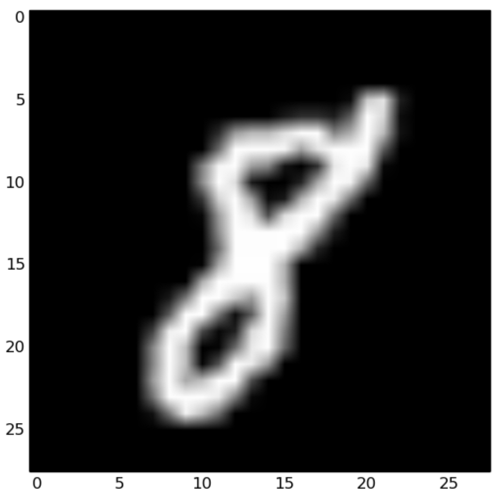

0
5
10
15
20
25
0
5
10
15
20
25

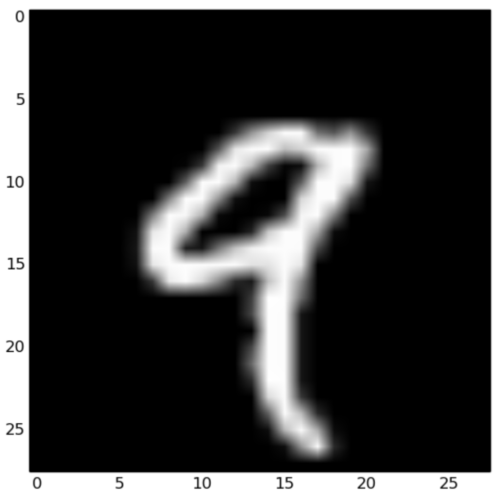

0
5
10
15
20
25
0
5
10
15
20
25


0
5
10
15
20
25
0
5
10
15
20
25

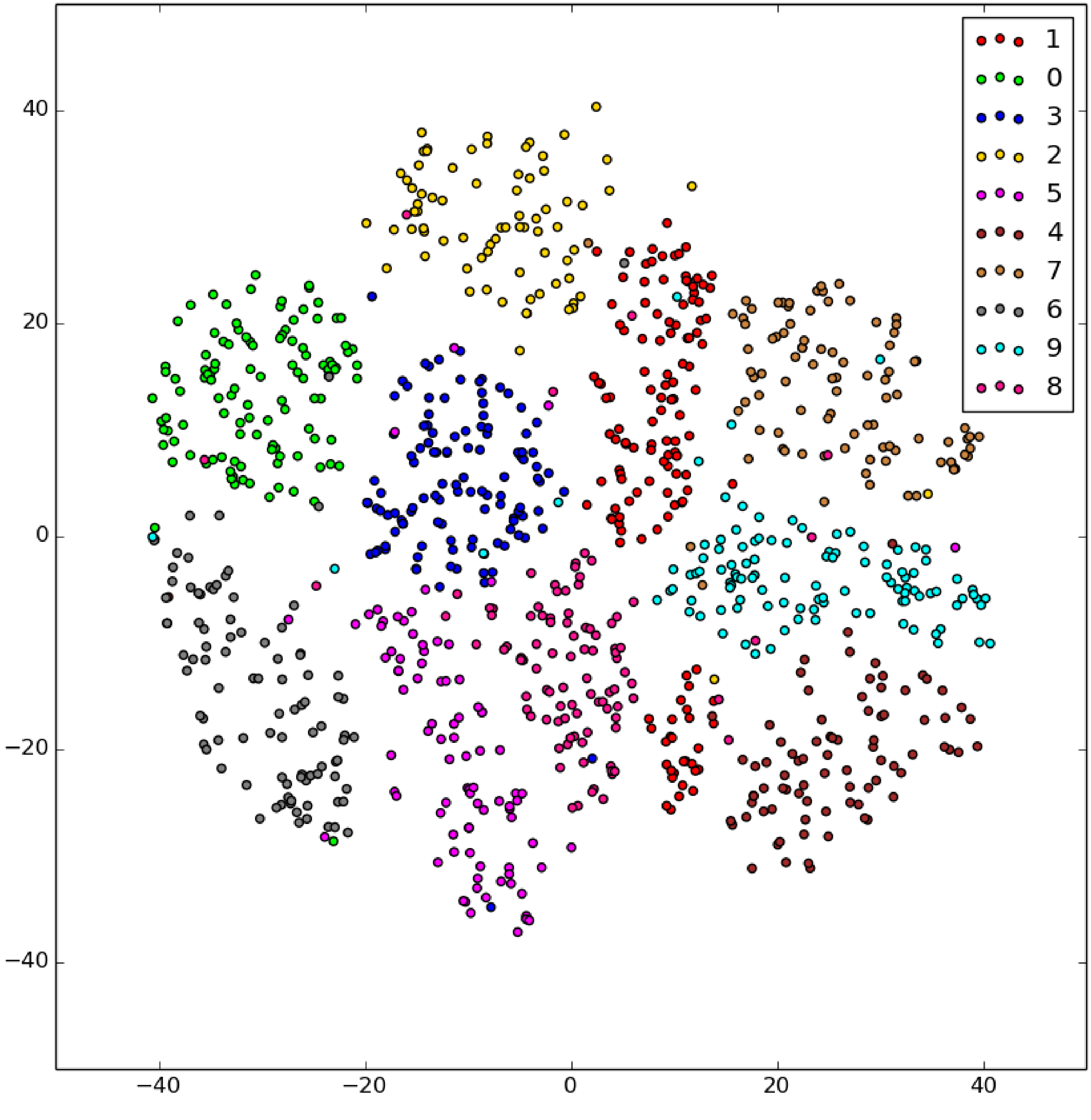

0
5
10
15
20
25
0
5
10
15
20
25


0
5
10
15
20
25
0
5
10
15
20
25



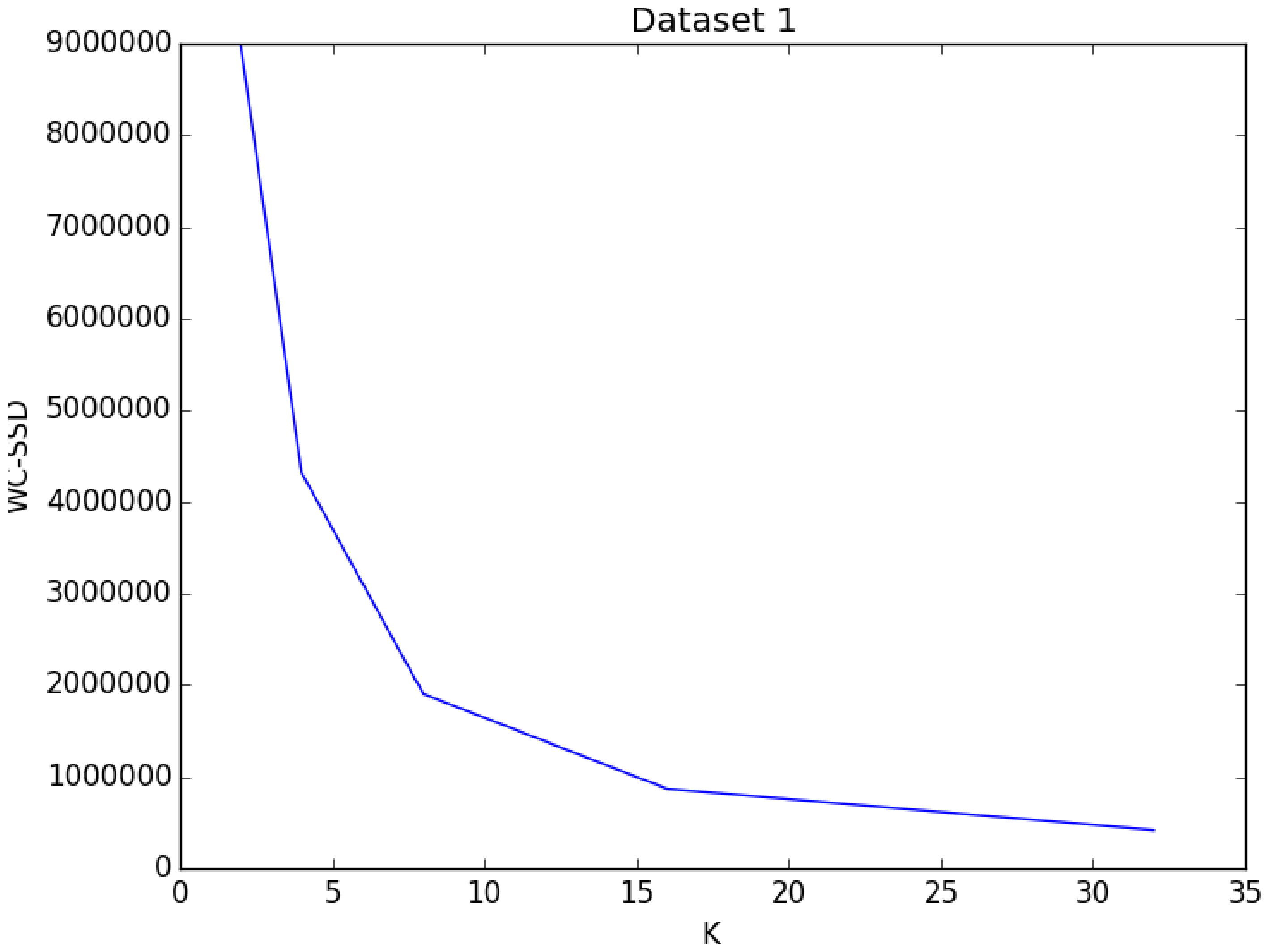


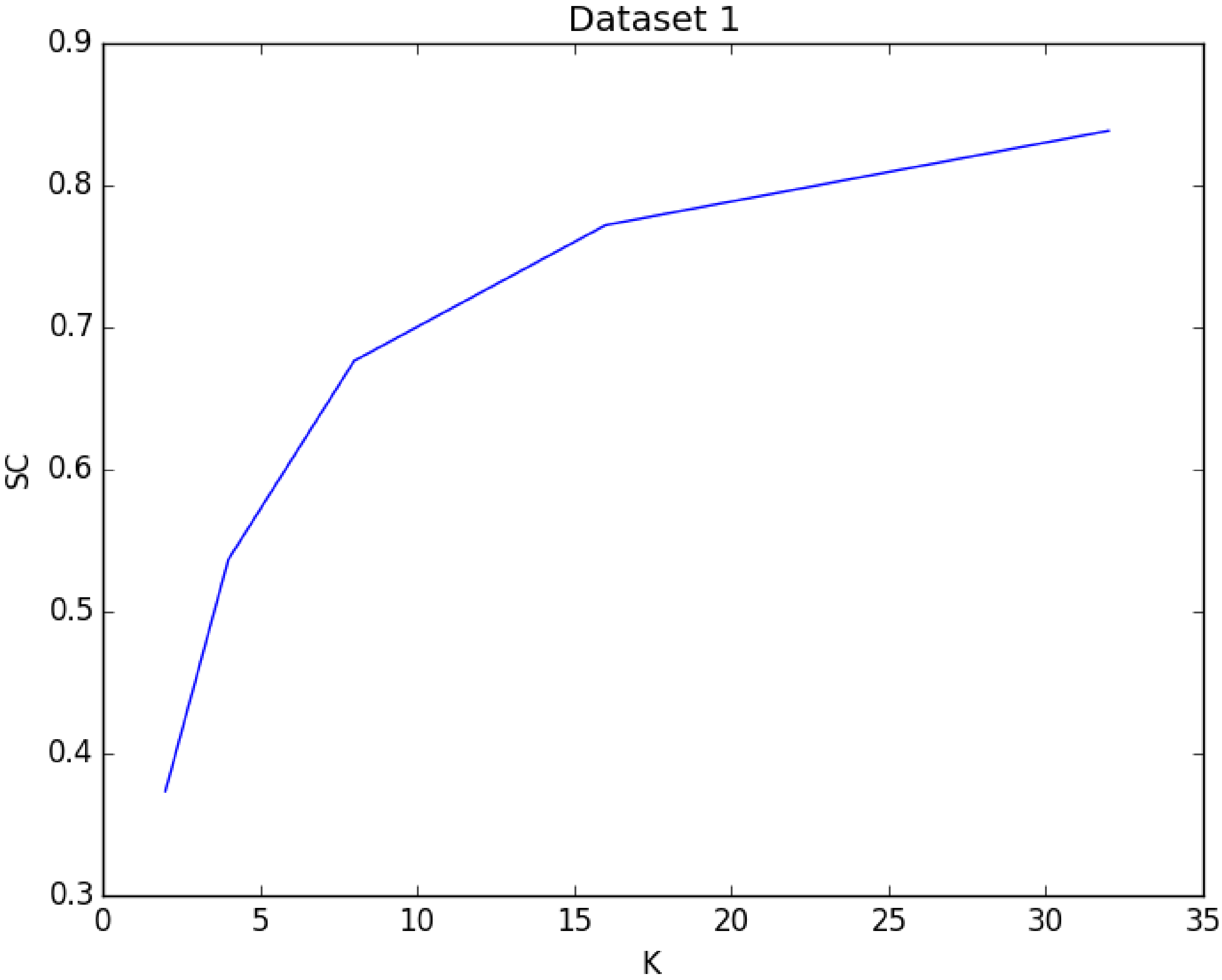
(b) The examples in 2D are:



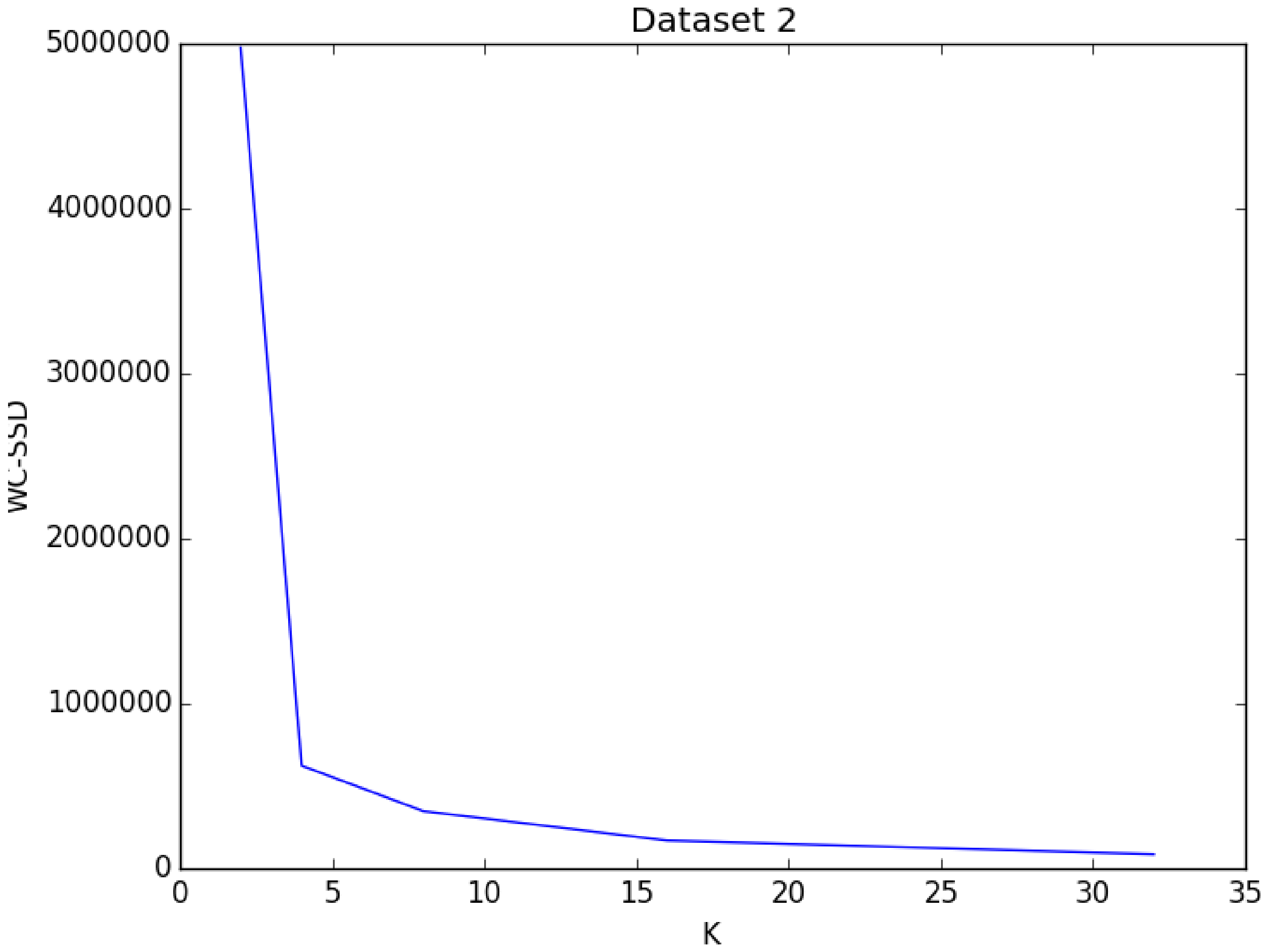
2. **B**

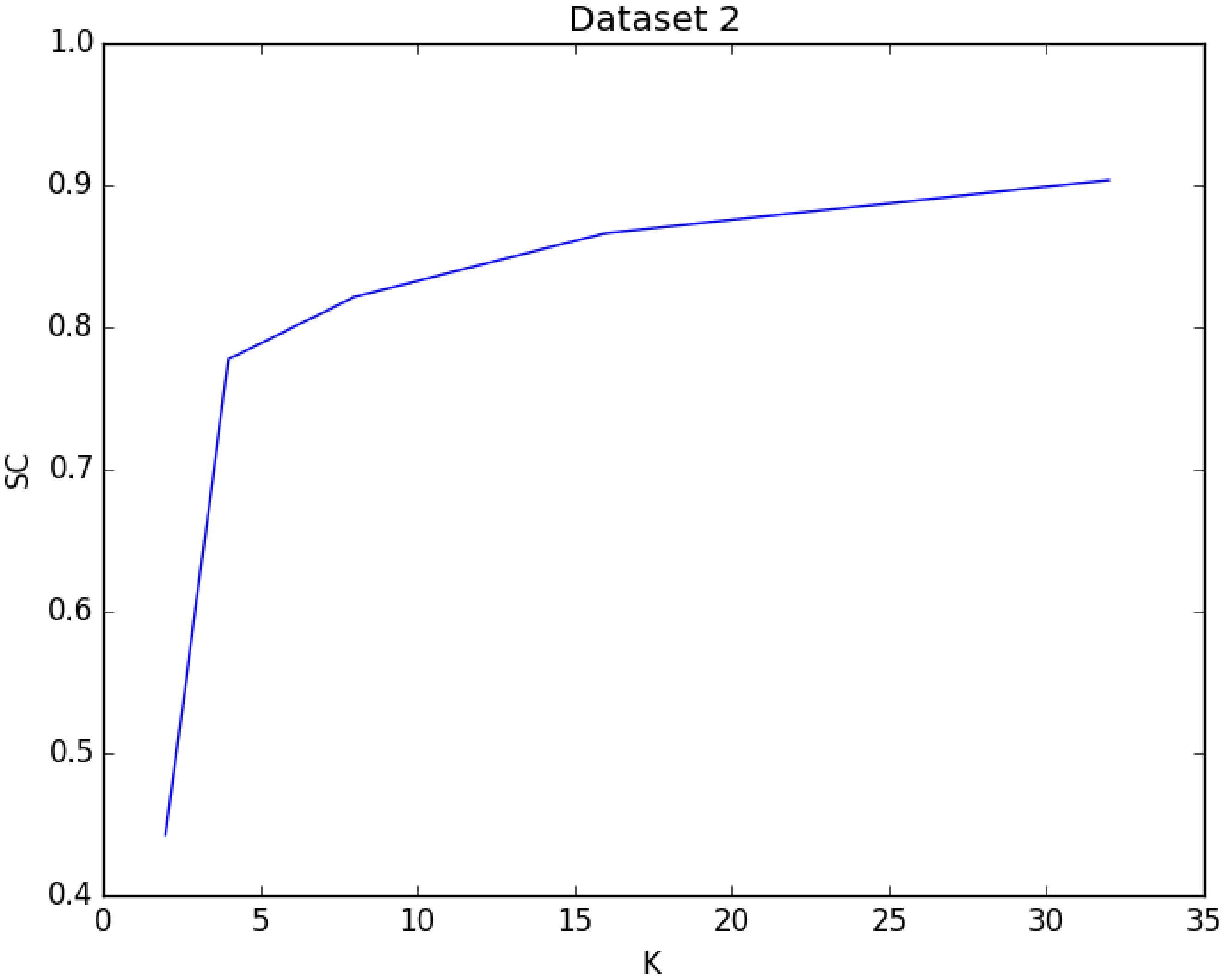
   (a) i. Dataset 1


Dataset 1
WC-SSD
K
9000000
8000000
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
0
5
10
15
20
25
30
35

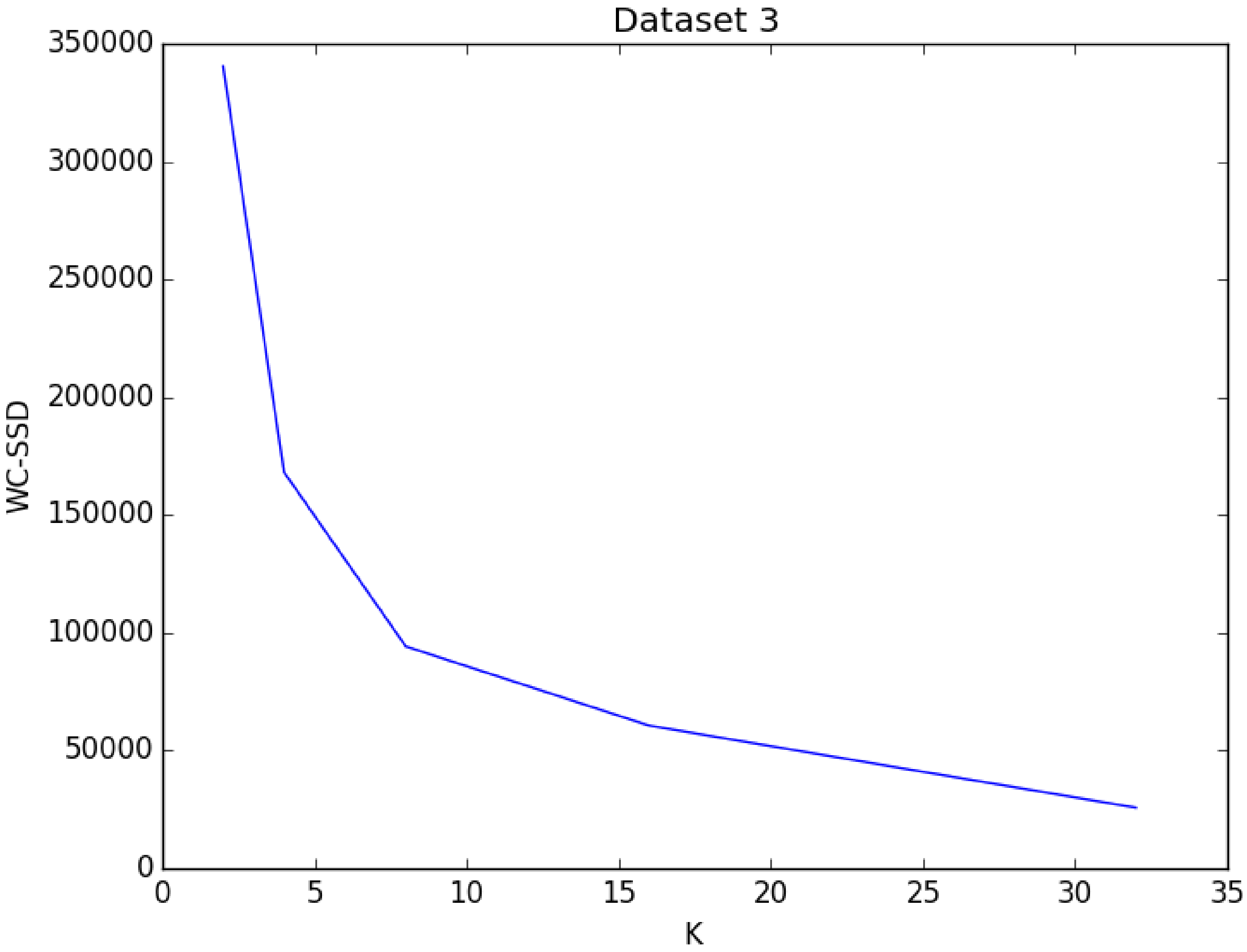


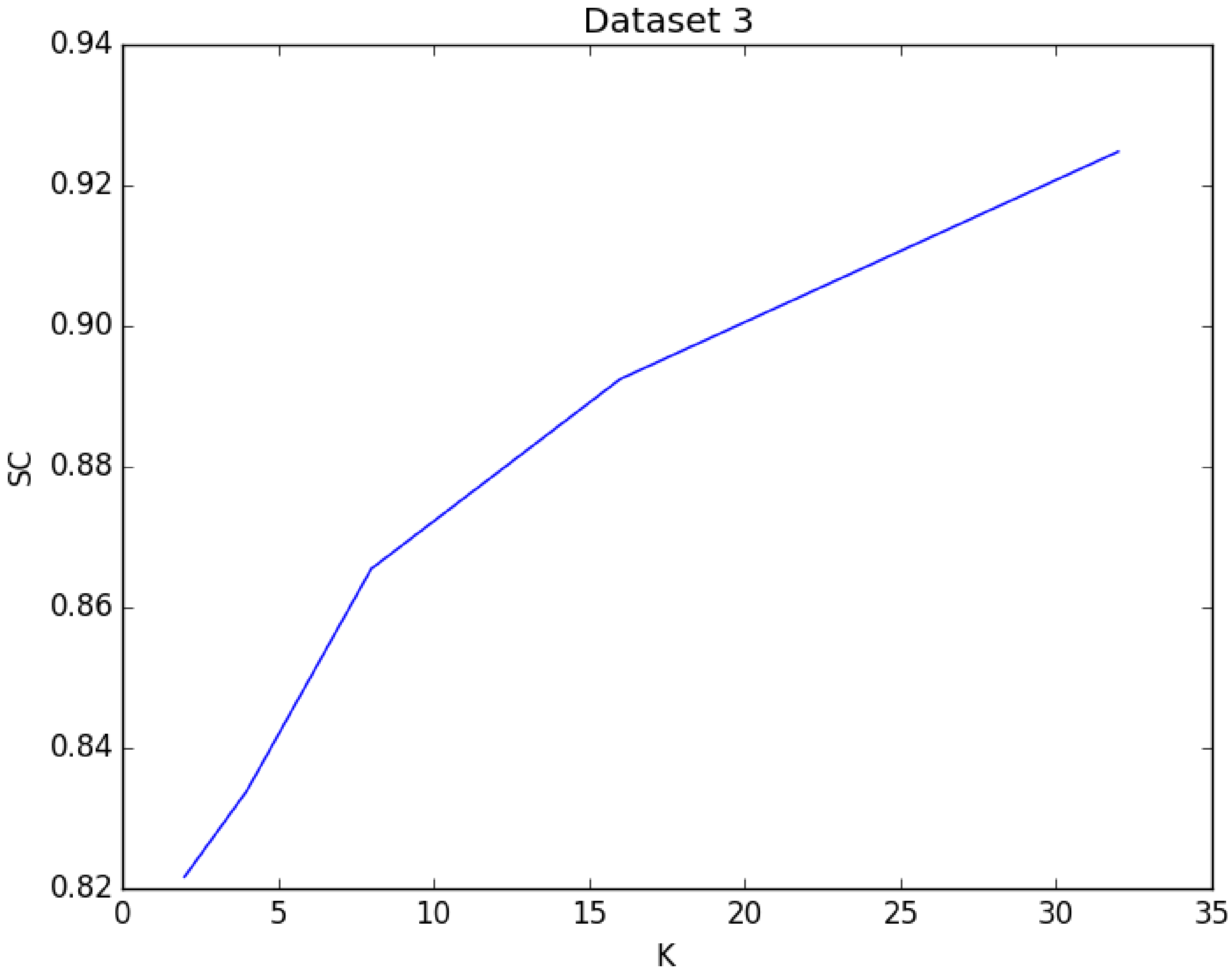
ii. Dataset 2


Dataset 2
WC-SSD
5000000
4000000
3000000
2000000
1000000
0
0
5
10
15
20
25
30
35
K



iii. Dataset 3


Dataset 3
WC-SSD
K
350000
300000
250000
200000
150000
100000
50000
0
0
5
10
15
20
25
30
35

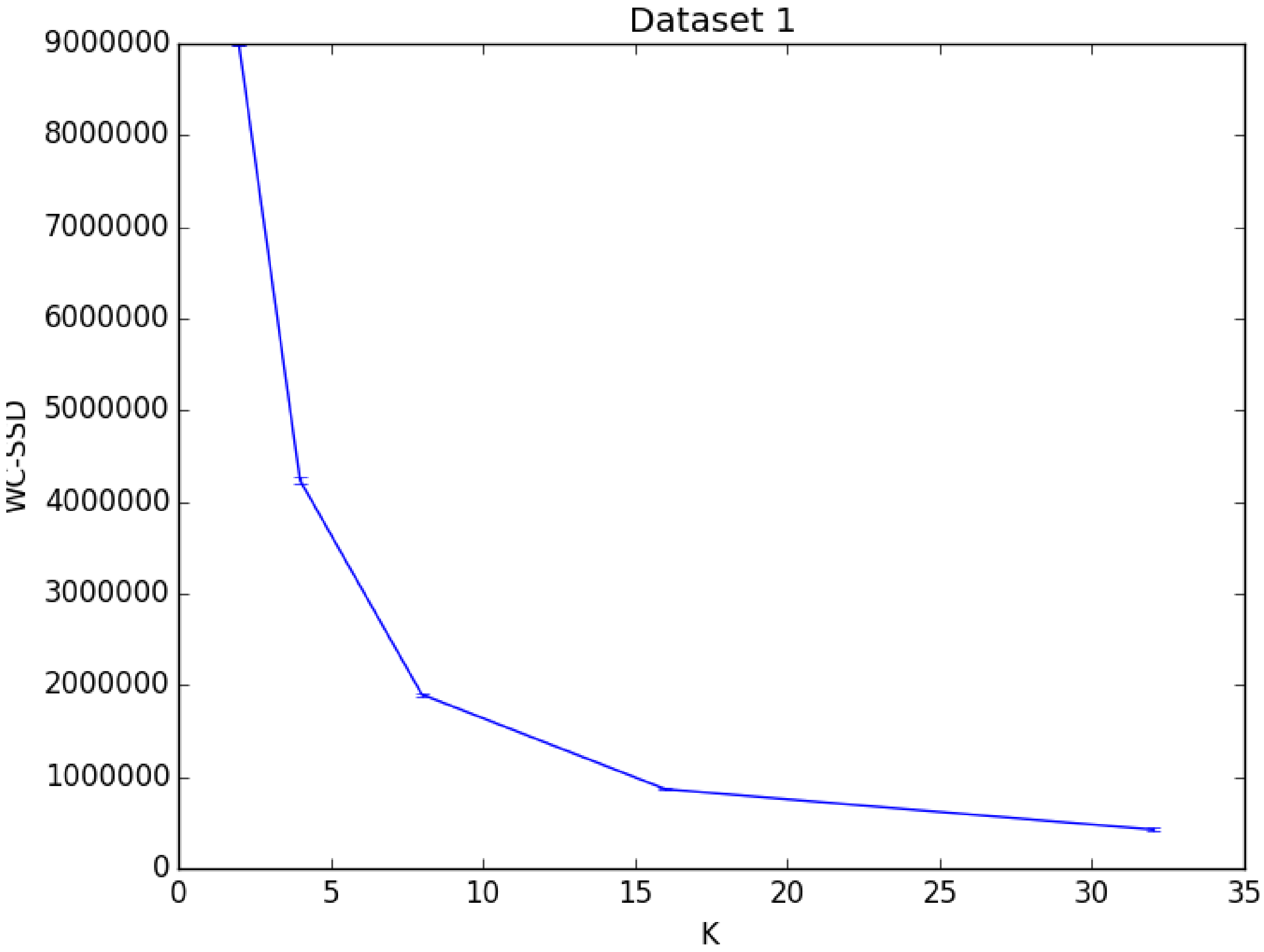


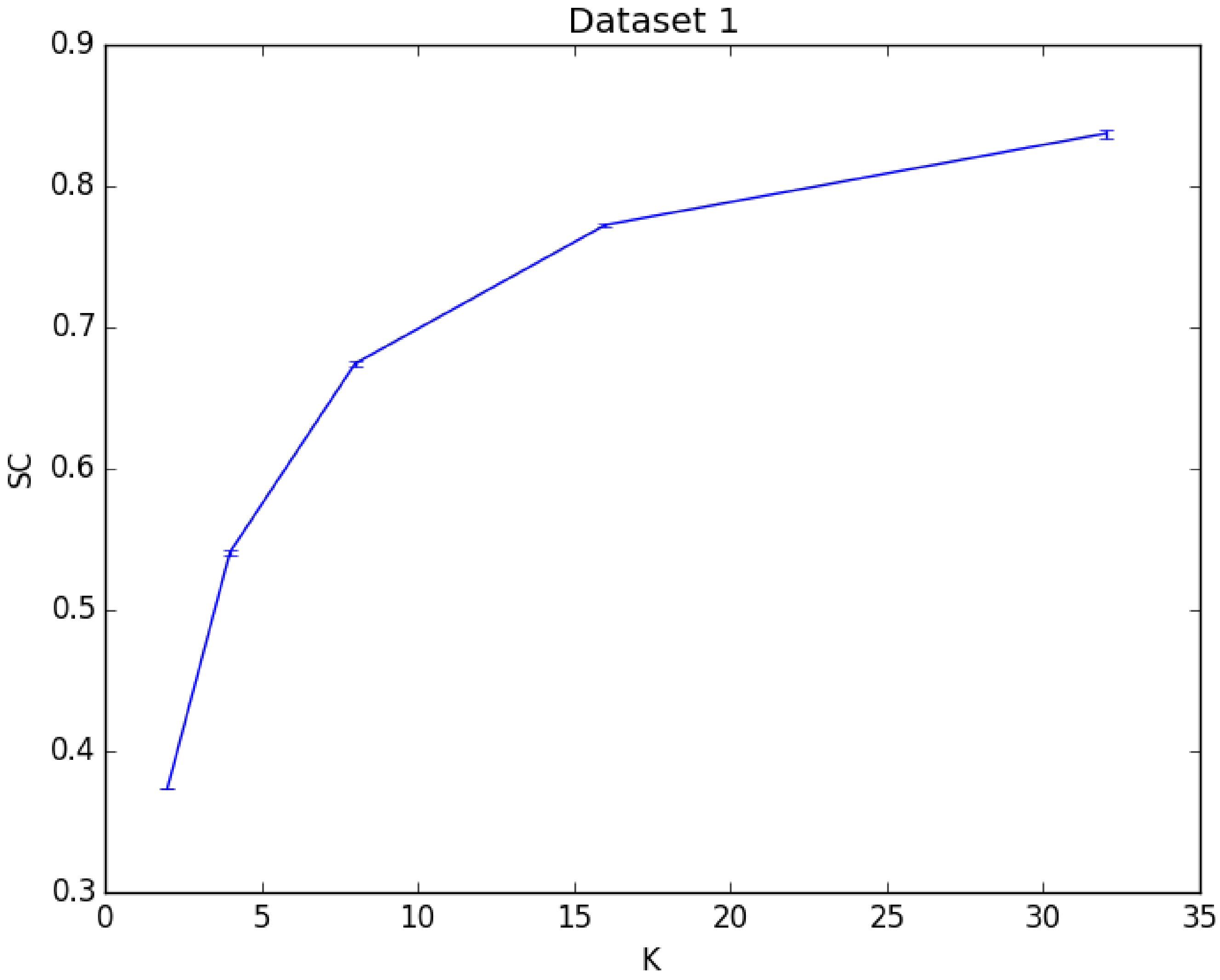
(b) To decide the value of k for each dataset, we can look at the elbow point in the plots of WC-SSD versus k, because beyond this value of k, the reduction in WC-SSD is minimal for an increased computational complexity of running k-means. We also look at the knee-point in the plots of SC versus k (based on the definition of SC used).

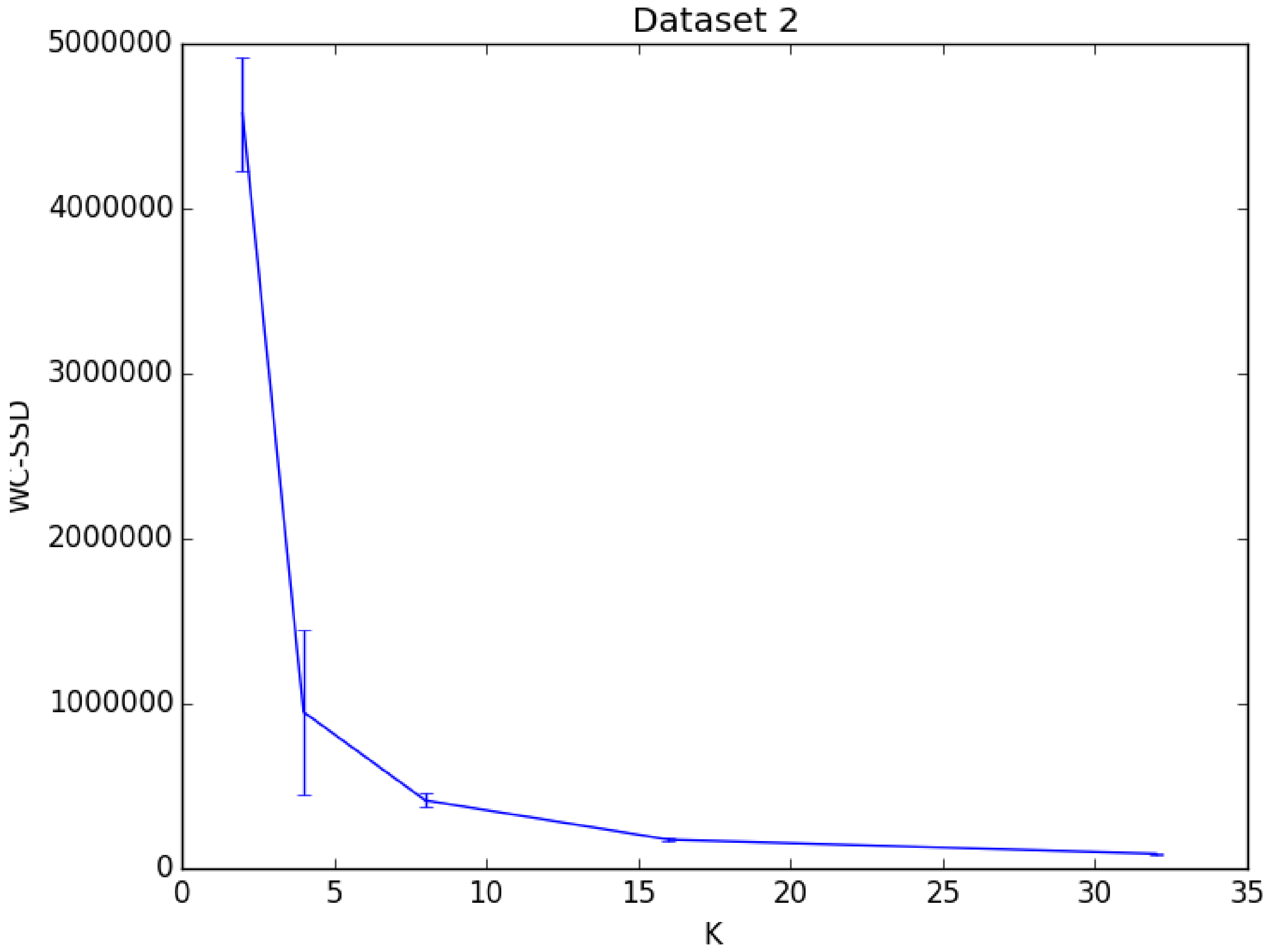
Hence, from the above plots, we can see the appropriate value of K will be 16 for Dataset 1, 8 for Dataset 2, and 2 for Dataset 3.

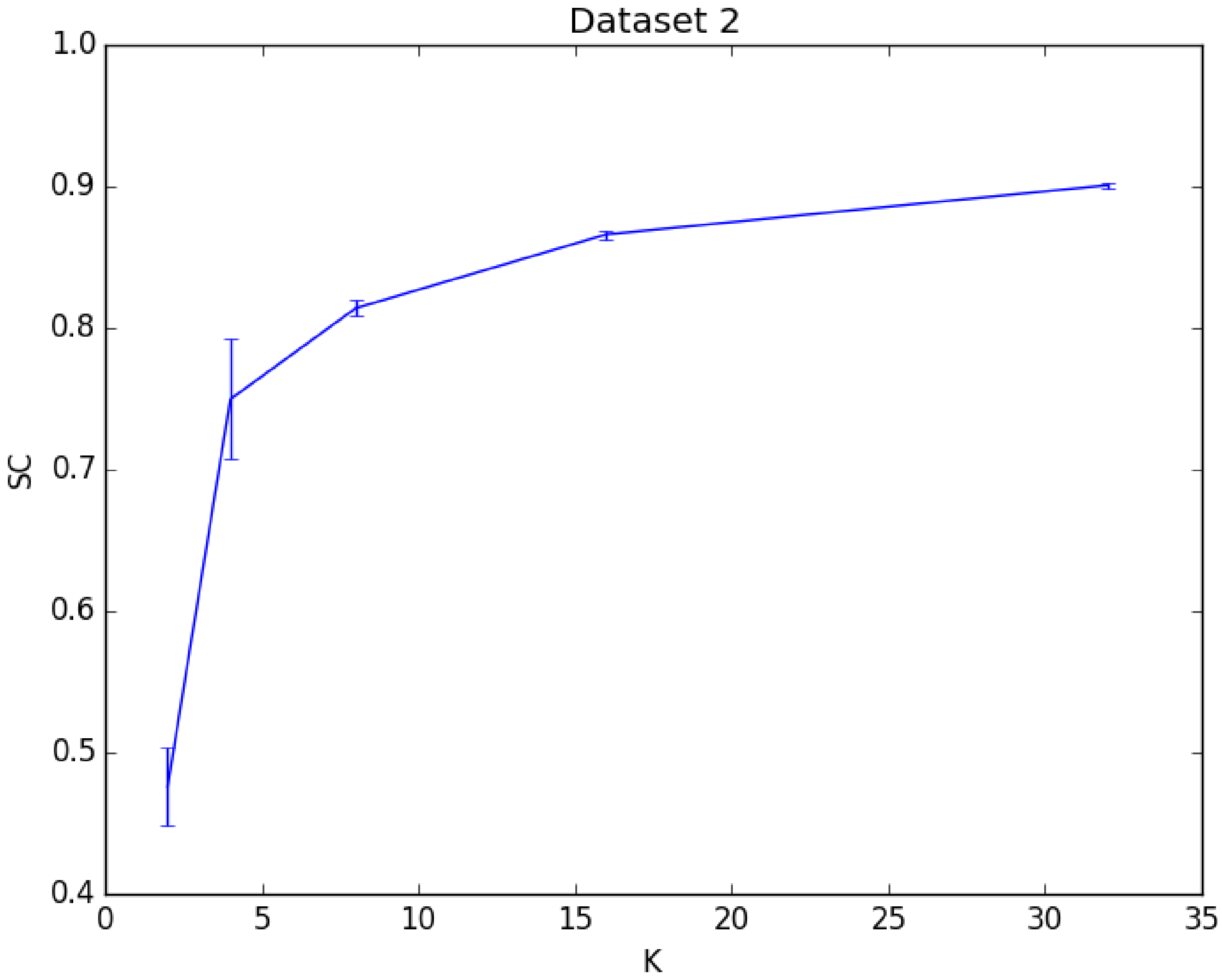
We can see that the selected value of K is close to the number of digits (classes) in the each dataset. Also, we can observe that as the number of digits (classes) decrease, the WC-SSD values also decrease.

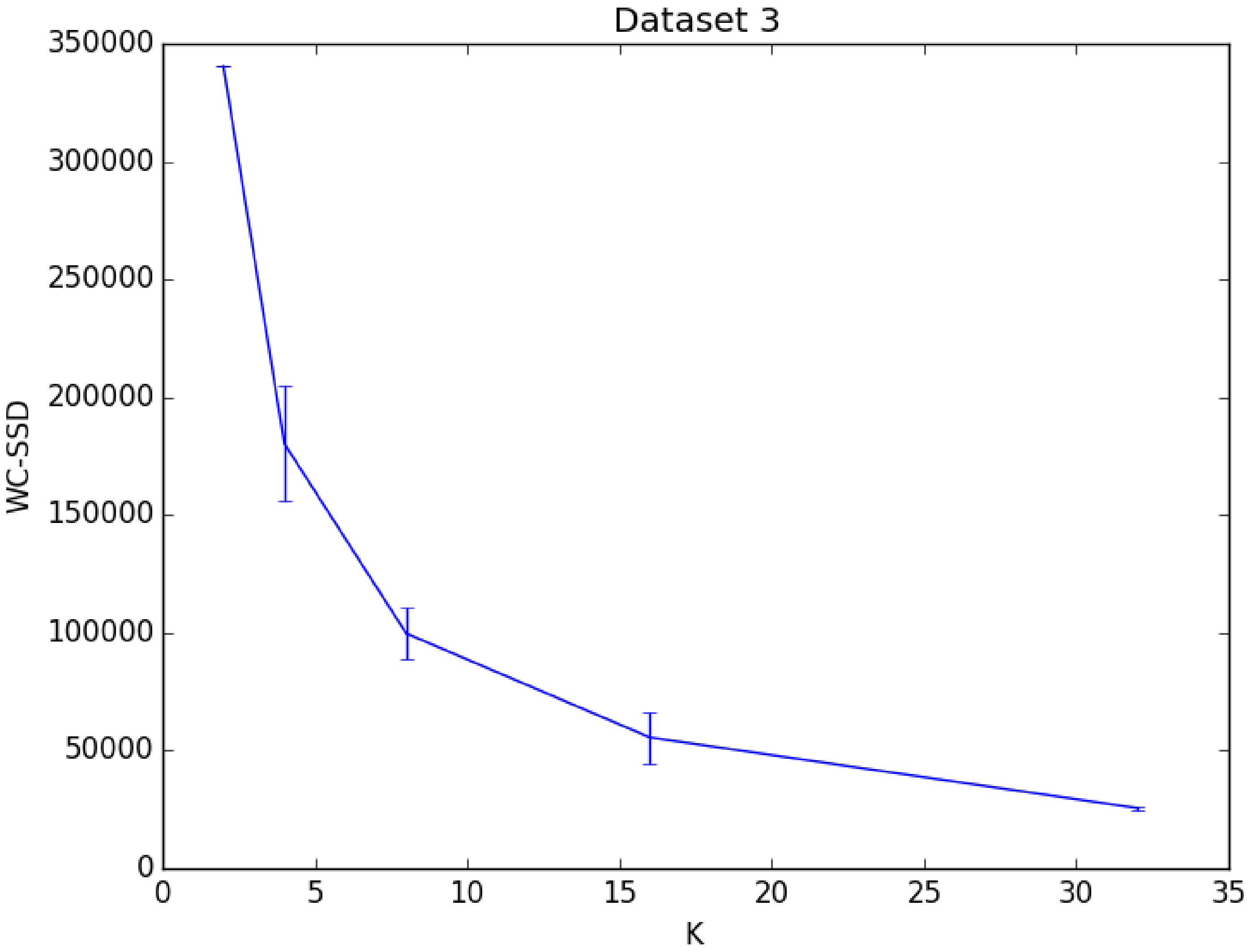
(c) The plots are

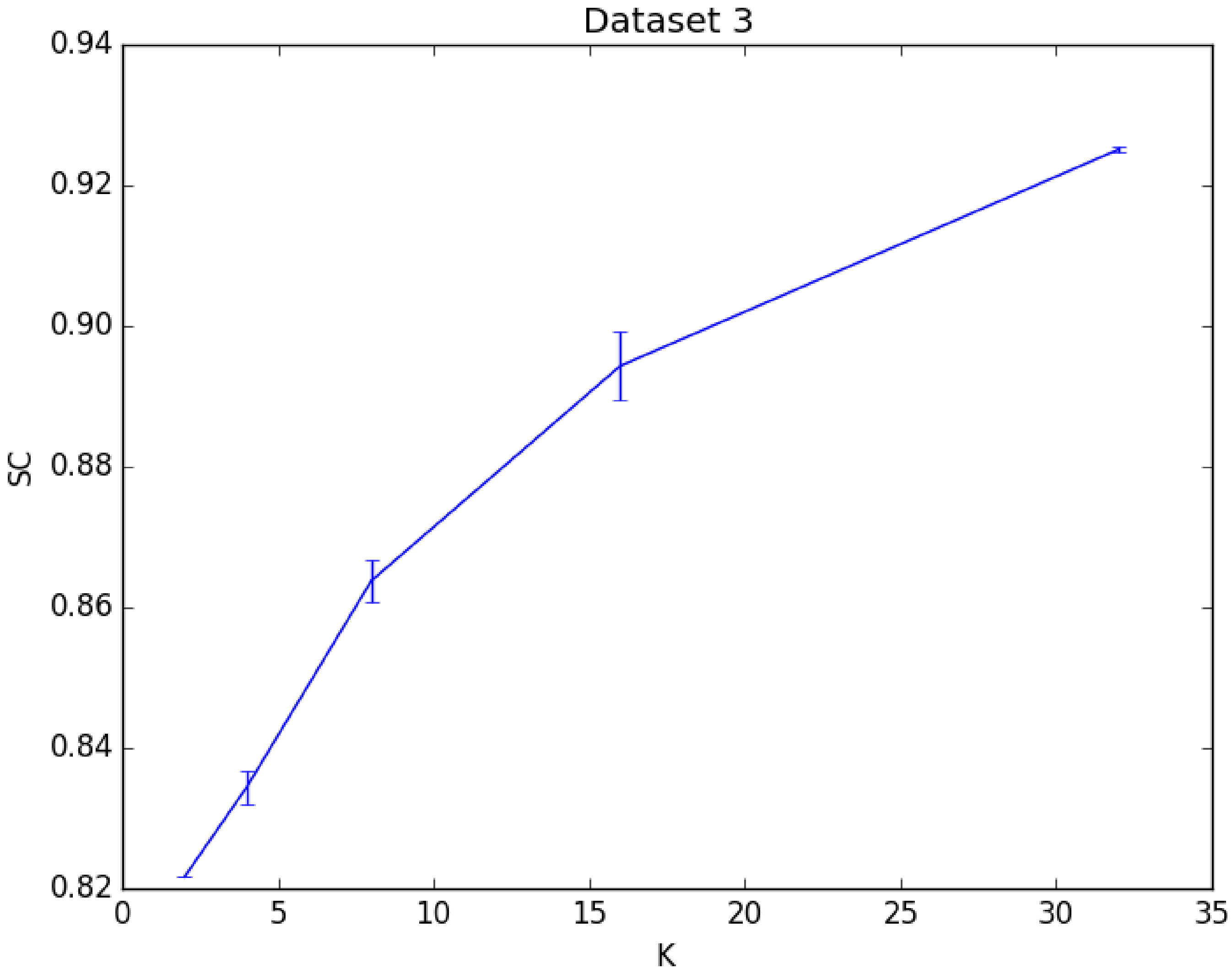

Dataset 1
WC-SSD
9000000
8000000
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
0
5
10
15
20
25
30
35
K


Dataset 1
SC
K
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0
5
10
15
20
25
30
35


Dataset 2
WC-SSD
K
0
1000000
2000000
3000000
4000000
5000000
0
5
10
15
20
25
30
35


Dataset 2
SC
K
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0
5
10
15
20
25
30
35


Dataset 3
WC-SSD
K
350000
300000
250000
200000
150000
100000
50000
0
0
5
10
15
20
25
30
35



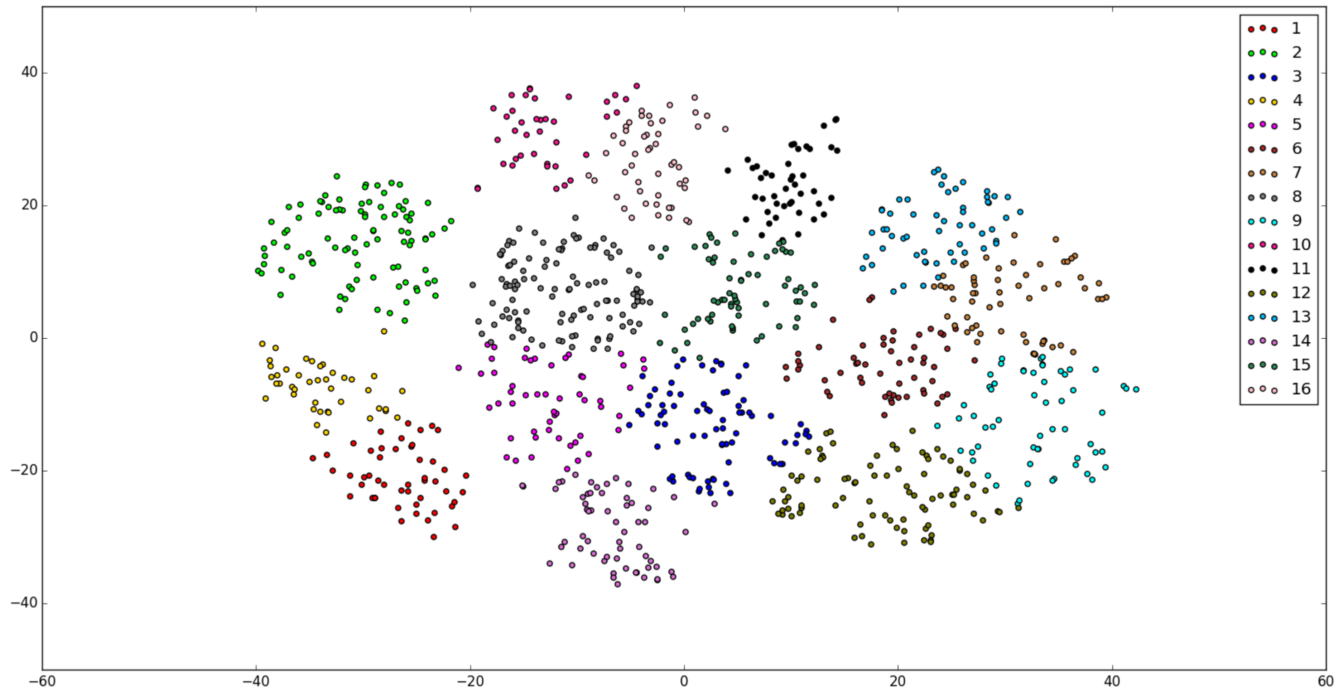
From the above plots, we can conclude that k-means is sensitive to the initial starting conditions, since the standard deviation (or variance) is high in both the plots: WC-SSD versus K, and SC versus K.

(d) Solution

i. Dataset 1

K = 16

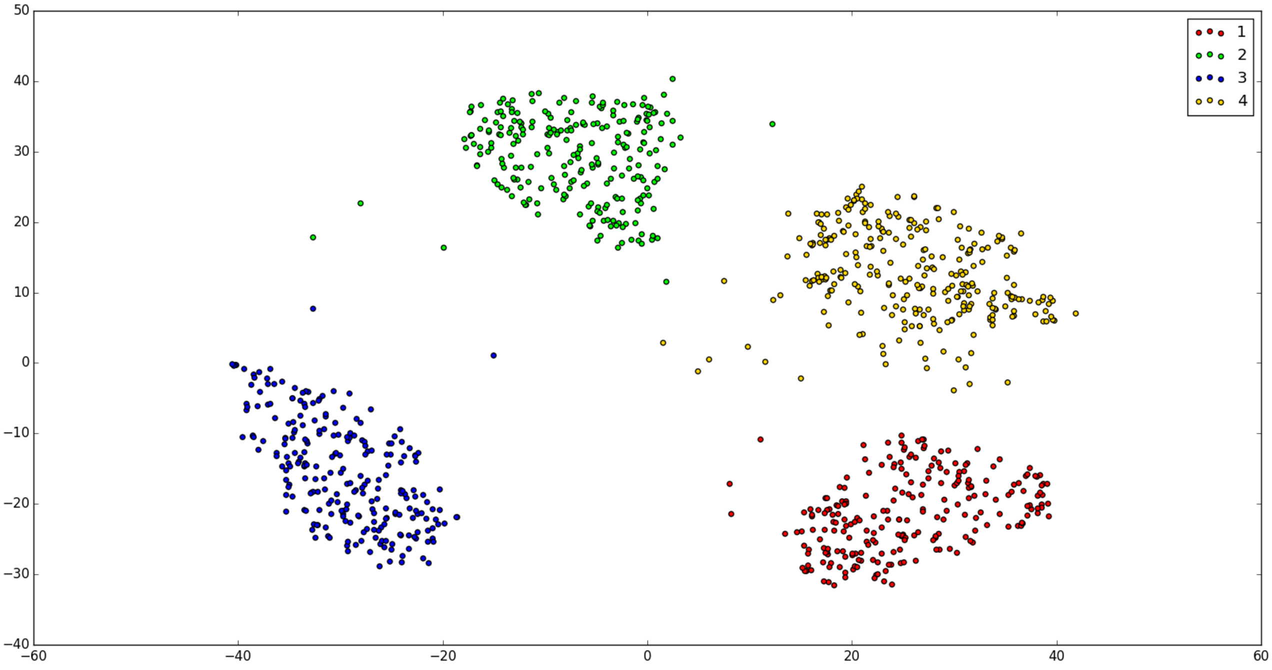
NMI = 0.368292602769



ii. Dataset 2

K = 4

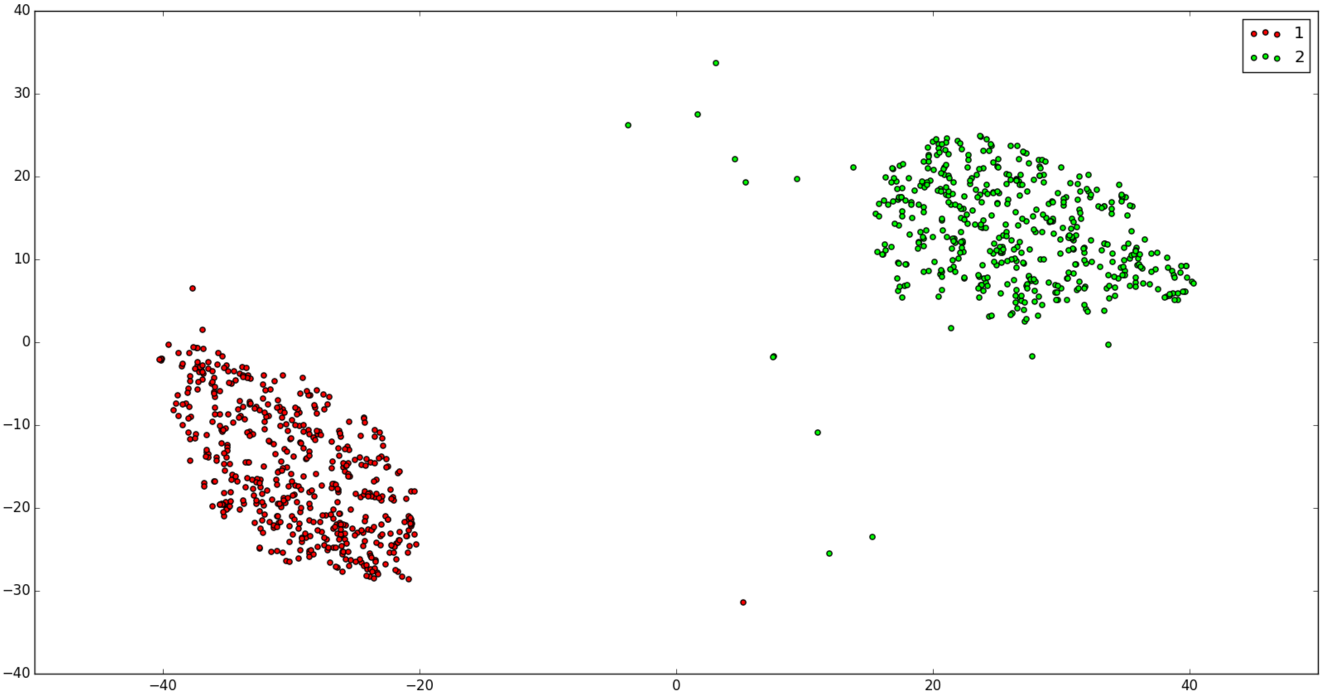
NMI = 0.45465341281



iii. Dataset 3

K = 2

NMI = 0.490710990204

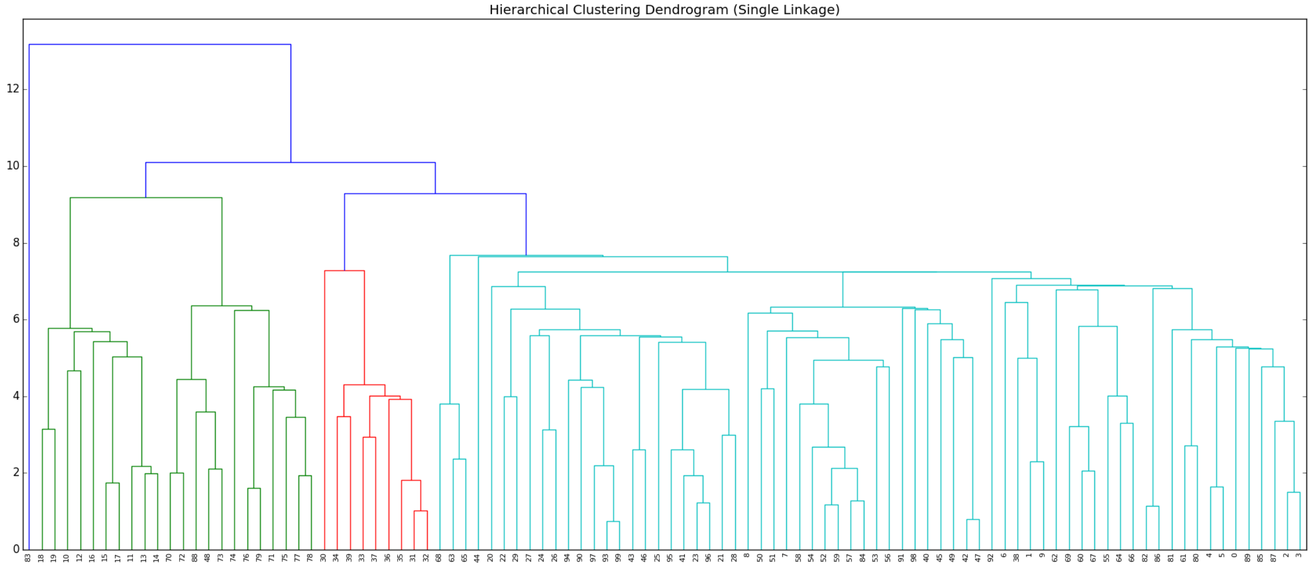


We can see that the NMI value increases as the number of digits (classes) in the dataset decreases. The NMI value is highest for dataset 3 (2 classes), followed by dataset 2 (4 classes) followed by dataset 1 (10 classes).

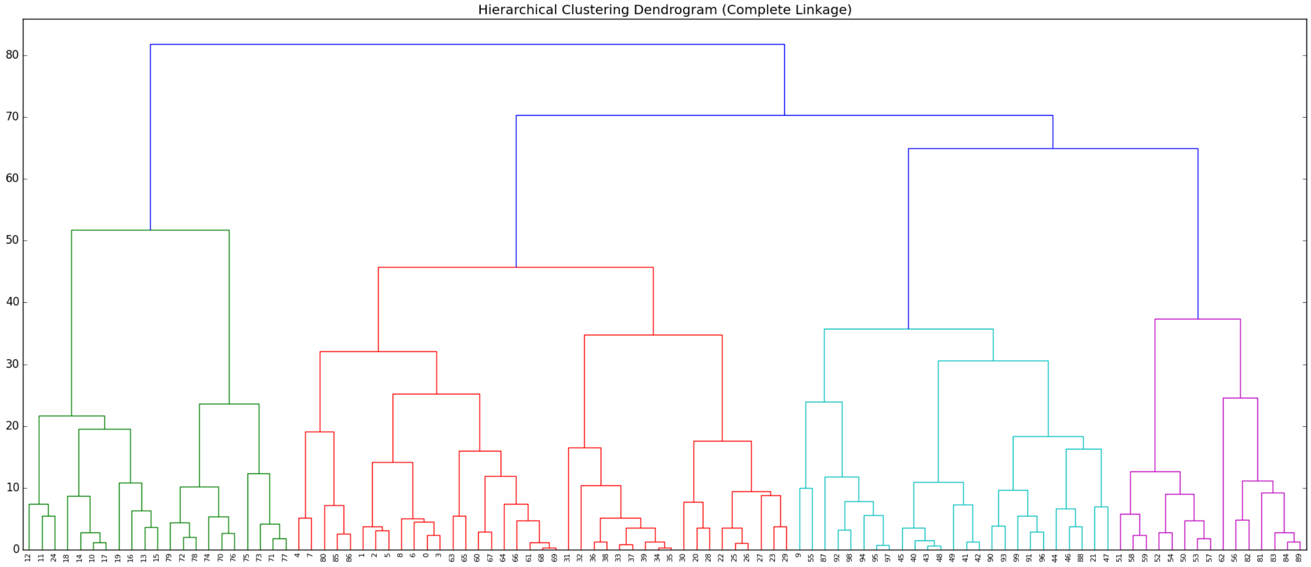
Also, in terms of visualization, the clusters become more separable when the number of classes in the dataset decreases. The clusters in dataset 3 (2 classes) are more clearly separated than those in dataset 2 (4 classes) which are more clearly separated than those in dataset 1 (10 classes).

3. **C**

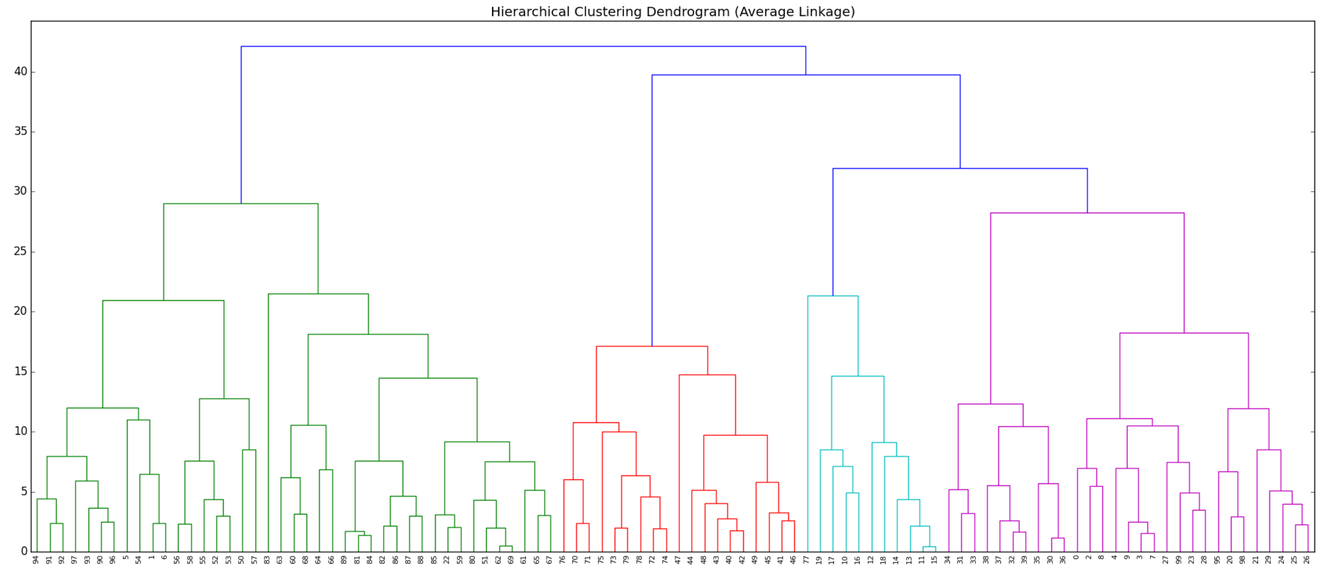
(a) Dendrogram using Single Linkage as the distance measure.



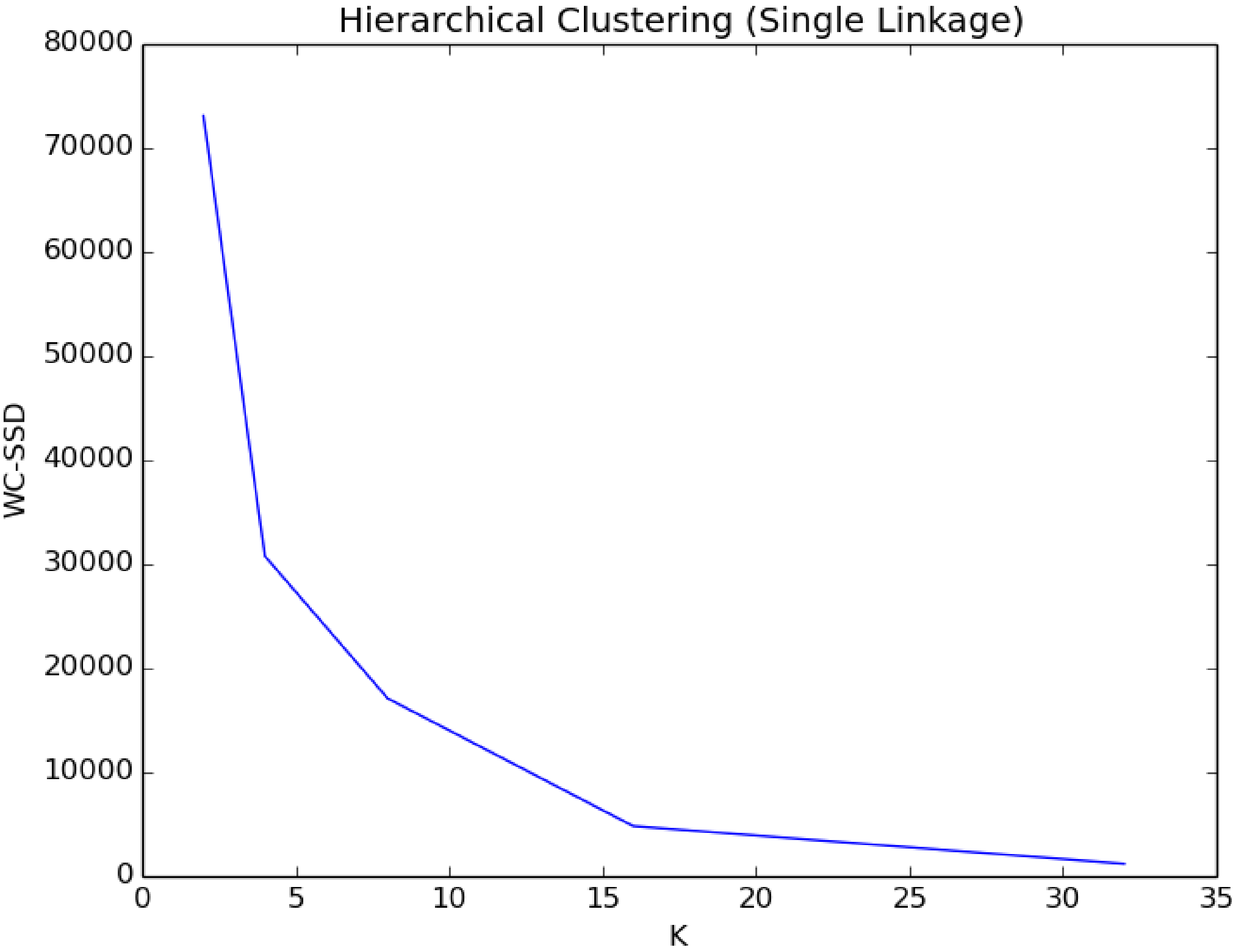
(b) i. Dendrogram using Complete Linkage as the distance measure.

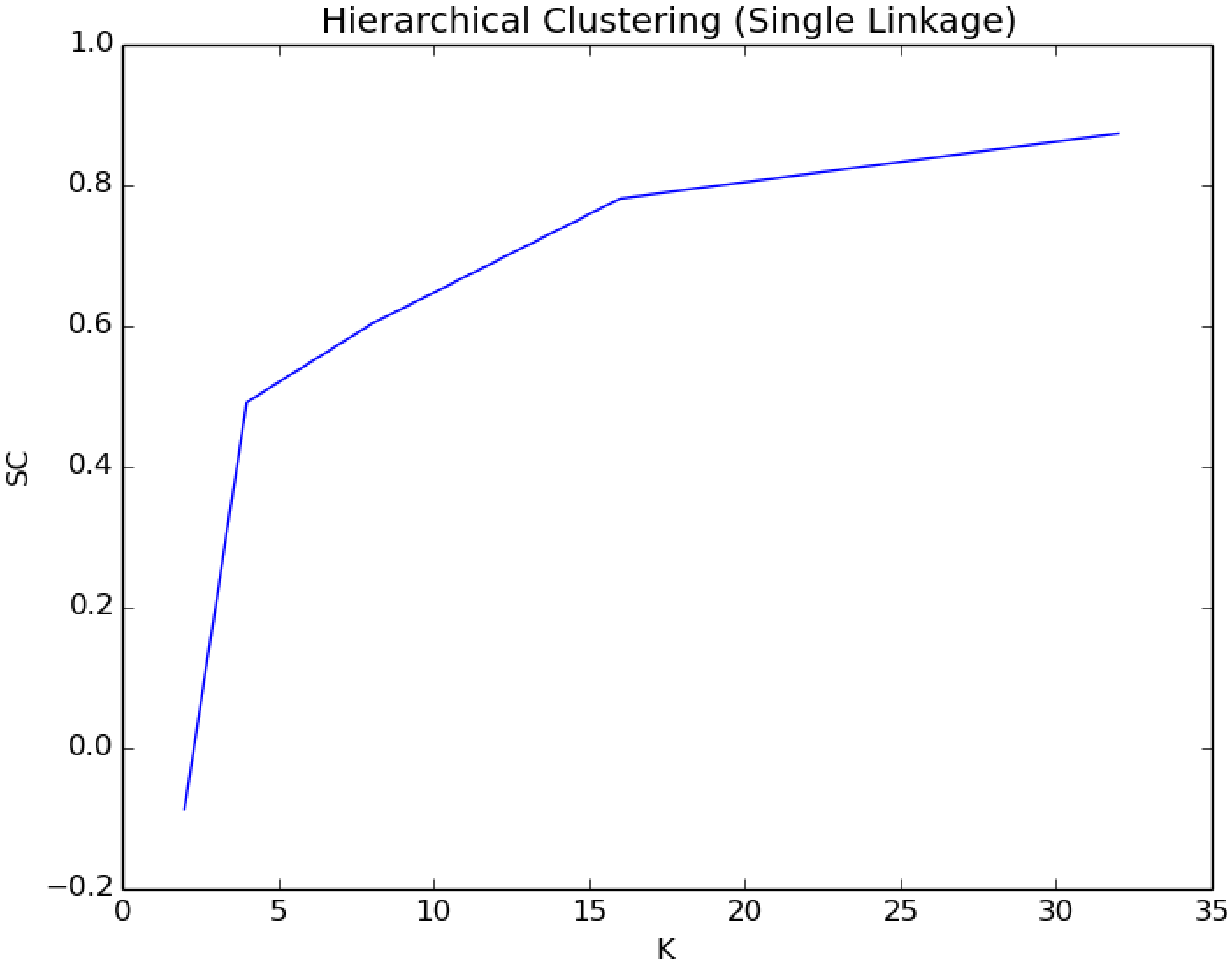


ii. Dendrogram using Average Linkage as the distance measure.

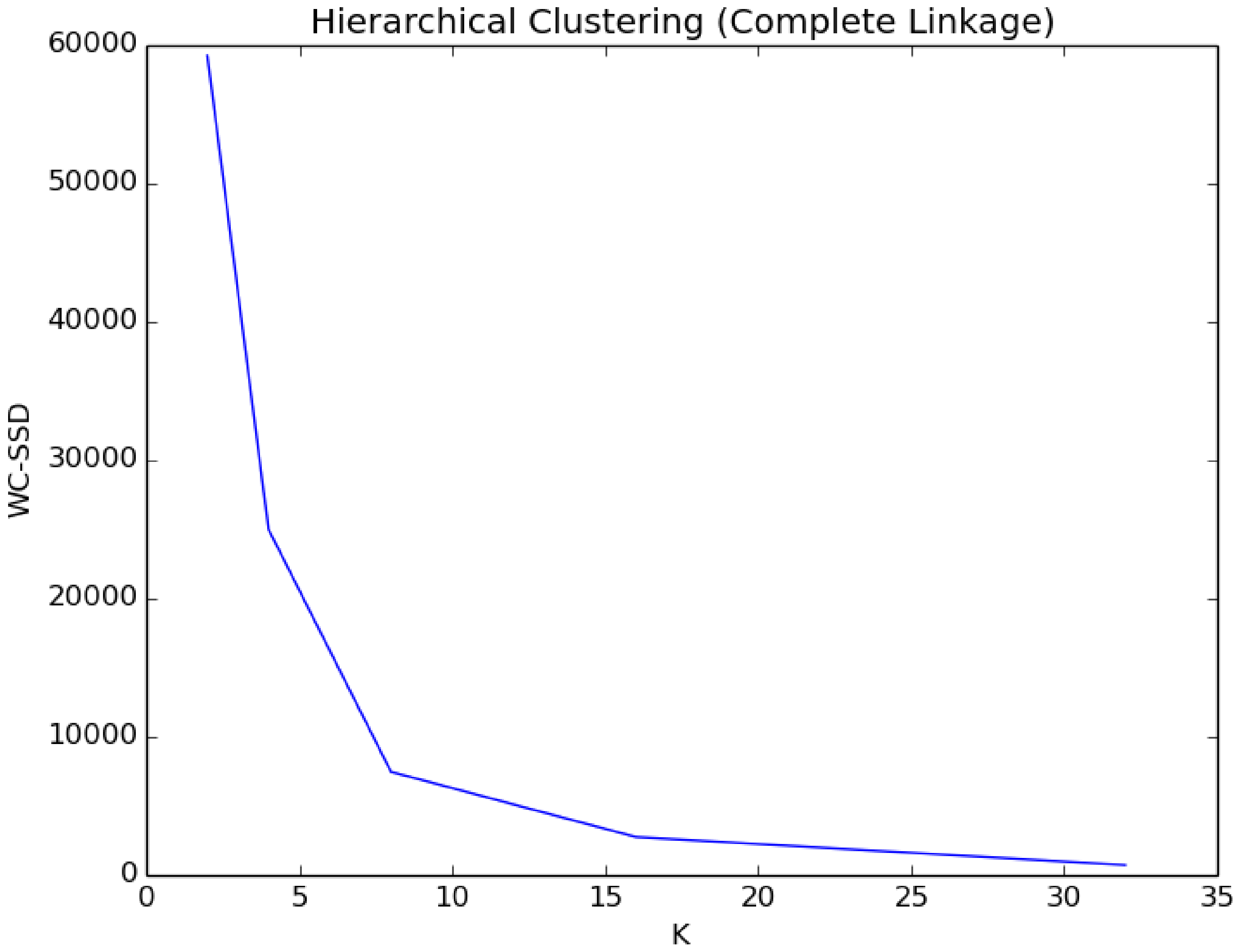


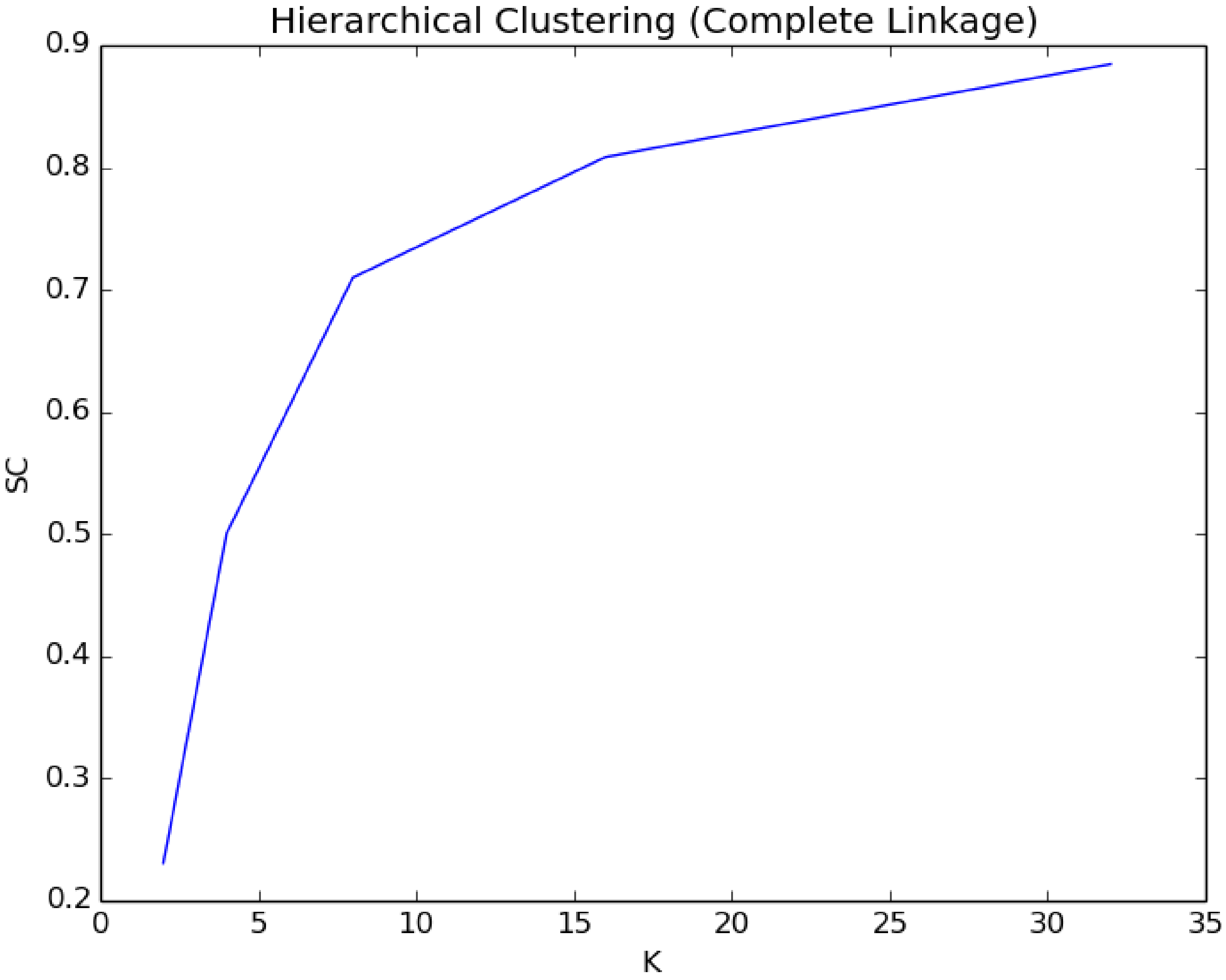
(c) i. Single Linkage


Hierarchical Clustering (Single Linkage)
80000
70000
60000
50000
40000
30000
20000
10000
0
WC-SSD
0
5
10
15
20
25
30
35
K

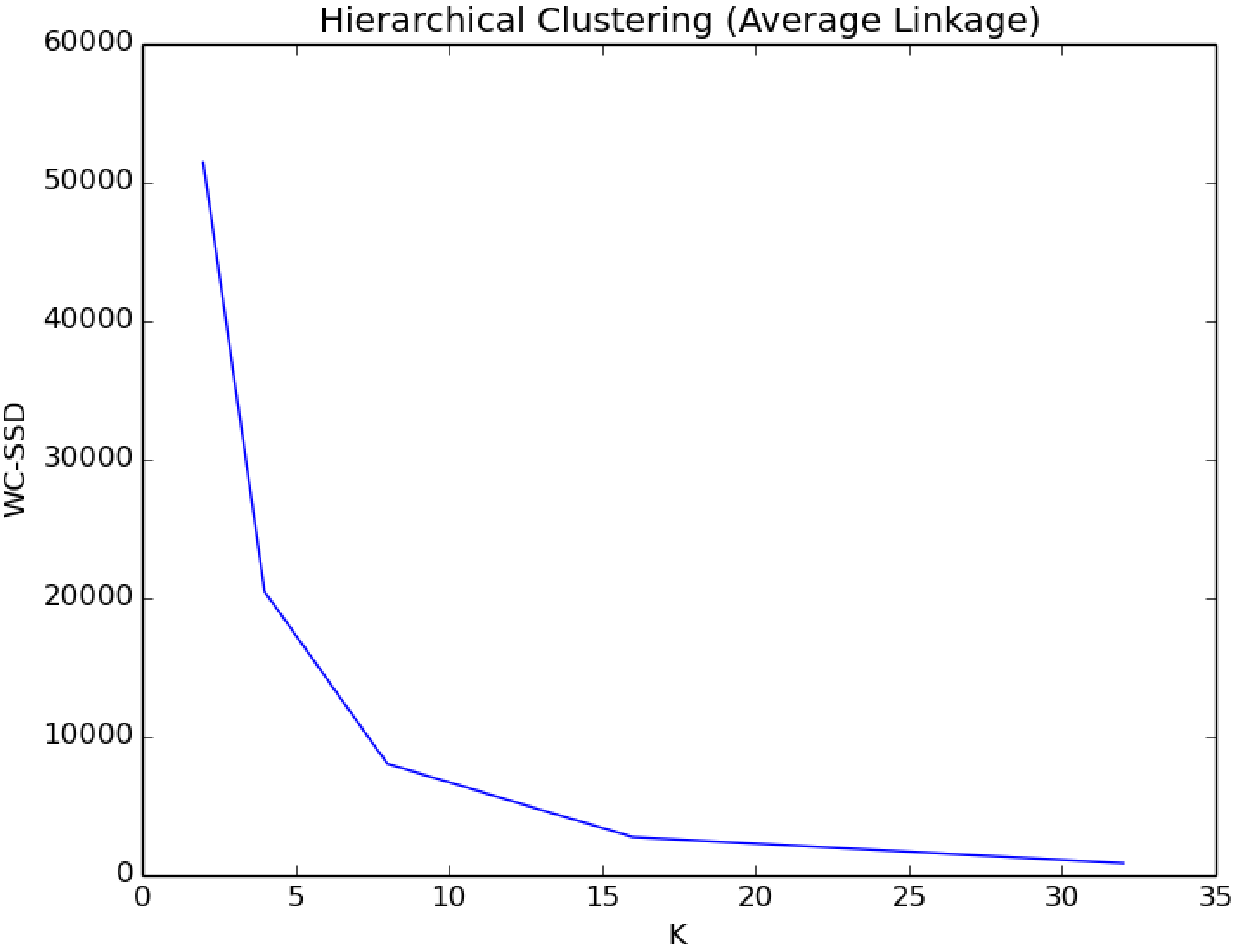


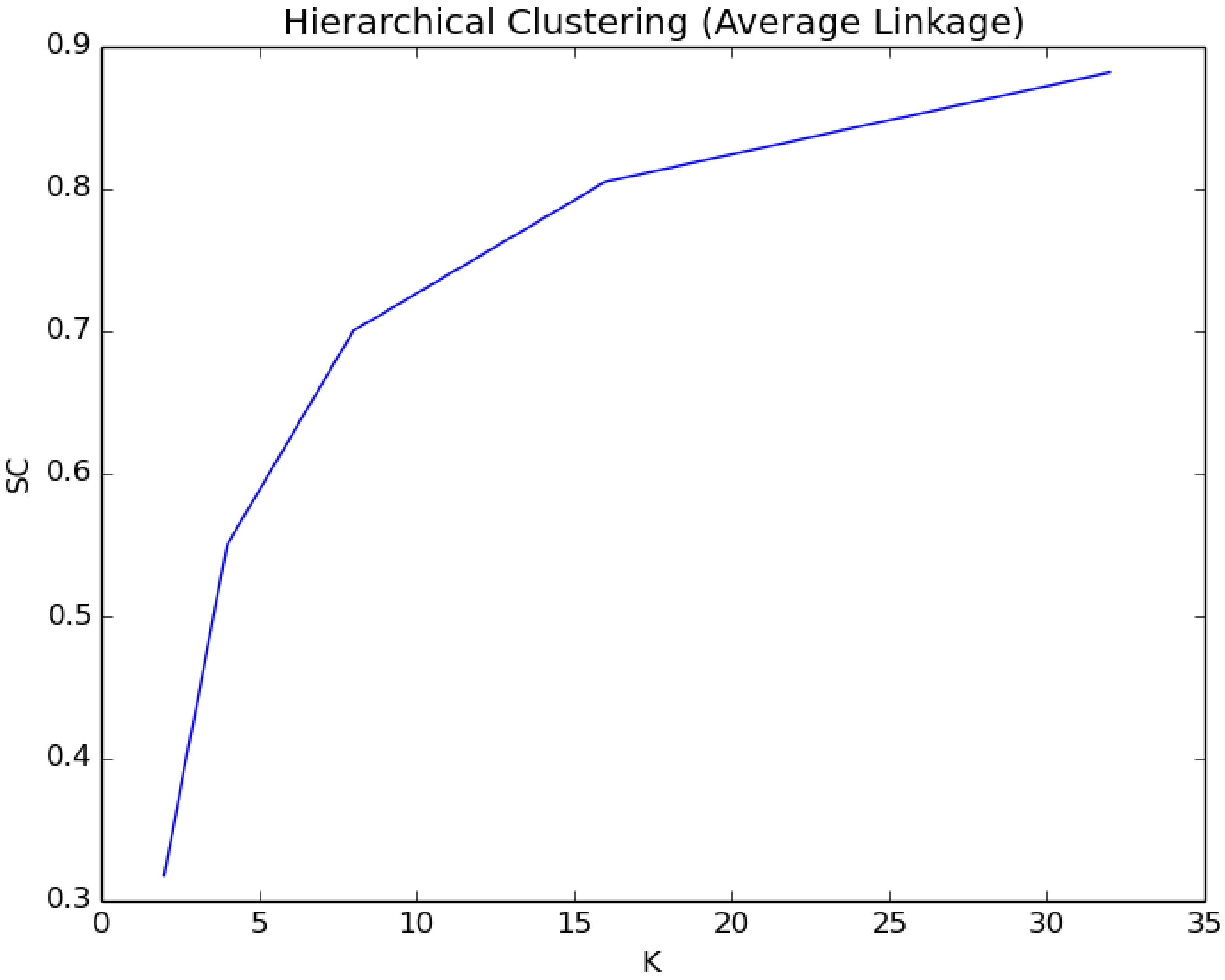
ii. Complete Linkage


Hierarchical Clustering (Complete Linkage)
WC-SSD
60000
50000
40000
30000
20000
10000
0
0
5
10
15
20
25
30
35
K



iii. Average Linkage


Hierarchical Clustering (Average Linkage)
WC-SSD
K
0
10000
20000
30000
40000
50000
60000
0
5
10
15
20
25
30
35



(d) To decide the value of k for each dataset, we can look at the elbow point in the plots of WC-SSD versus k, because beyond this value of k, the reduction in WC-SSD is minimal for an increased computational complexity of running k-means. We also look at the knee-point in the plots of SC versus k (based on the definition of SC used).

Hence, from the above plots, we can see the appropriate value of K will be 16 for all types of linkages : single, complete and average.

The selected value of K, which is 16, is same as that for the K chosen for kmeans which was also 16, for dataset 1 (complete dataset with 10 digits).

(e) The value of K chosen is 16 for all types of linkages: single, average and complete. The NMI values are:

Single Linkage: 0.3782359727

Complete Linkage: 0.4088758492

Average Linkage: 0.4373220577

The NMI score is highest for average linkage, followed by complete linkage, followed by single linkage. But, the differences in scores are not significant, and the dataset is too small in size (100 examples), and there is a lot of variation in the results across different runs (since the points are chosen randomly). Hence, these scores are rough estimates.

The NMI scores for hierarchical clustering using all (single, complete and hierarchical) linkages are higher than the NMI scores for k-means in part B.