CS57300: Data Mining

Assignment 5

Name: Mohammad Haseeb
Purdue ID: mhaseeb@purdue.edu

Due: April 19, 2019

Note: Figures appear at different locations in the document due to position issues of LaTeX. I have added reference links that you can click on to take you to that figure. Apologies for the inconvenience.

1 Exploration

(i) Figure 1 shows the visualization for the digit **0**.

Figure 2 shows the visualization for the digit **1**.

Figure 3 shows the visualization for the digit **2**.

Figure 4 shows the visualization for the digit **3**.

Figure 5 shows the visualization for the digit **4**.

Figure 6 shows the visualization for the digit **5**.

Figure 7 shows the visualization for the digit **6**.

Figure 8 shows the visualization for the digit **7**.

Figure 9 shows the visualization for the digit **8**.

Figure 10 shows the visualization for the digit **9**.

(ii) Figure 11 shows the clustering of 1000 randomly selected examples.

2 K-means Clustering

2.1 Code

The output of **kmeans.py** is shown in Figure 12. The code takes approx. 2 minutes to run.

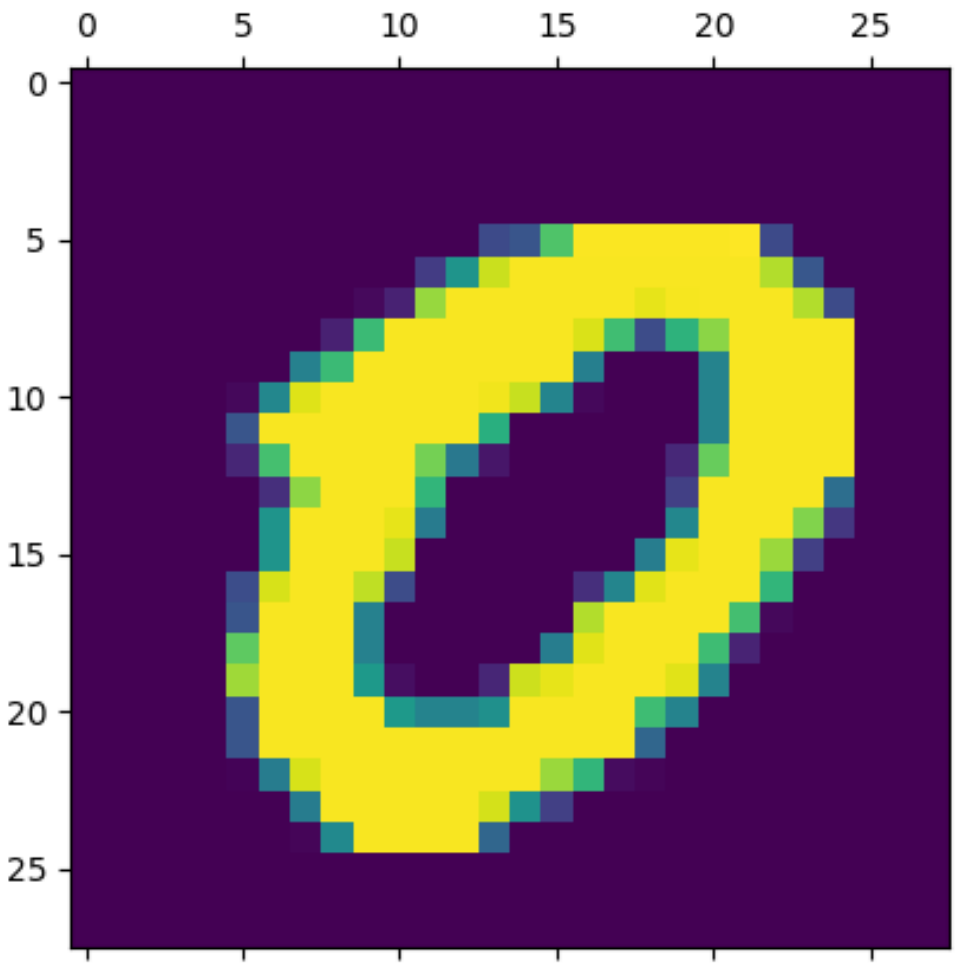

Figure 1: Digit 0

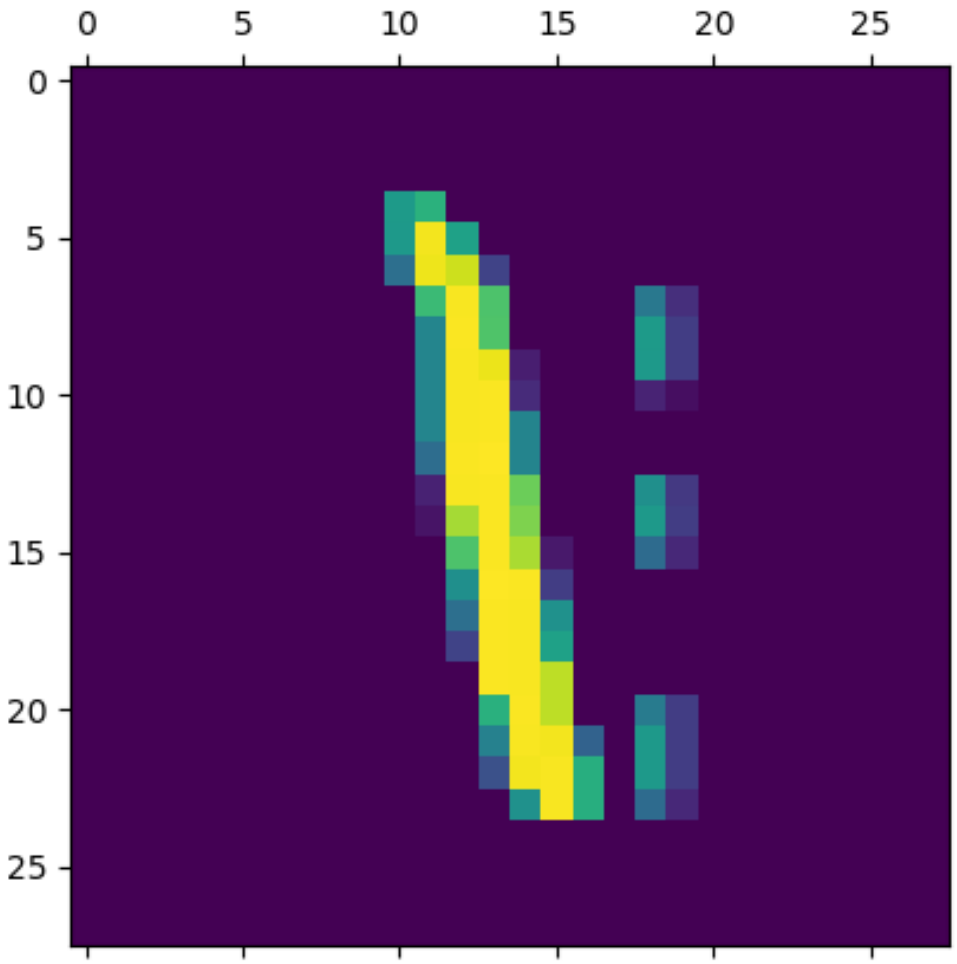

Figure 2: Digit 1

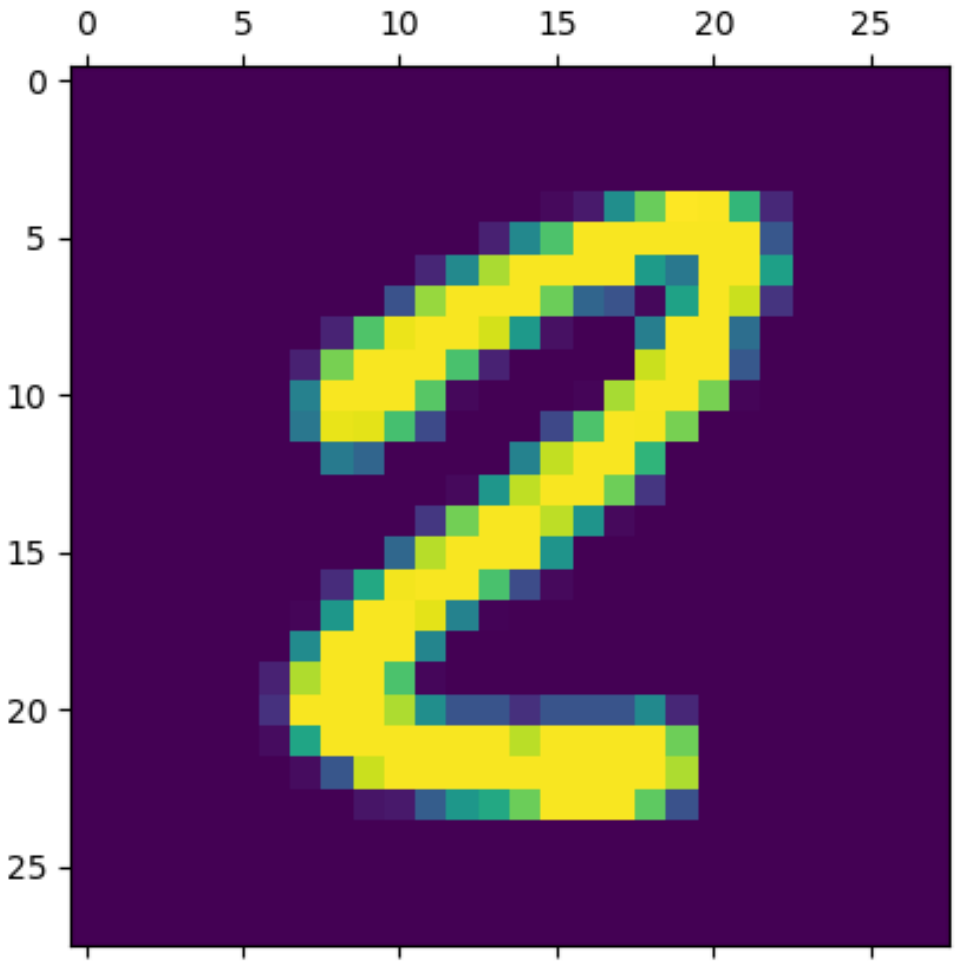

Figure 3: Digit 2

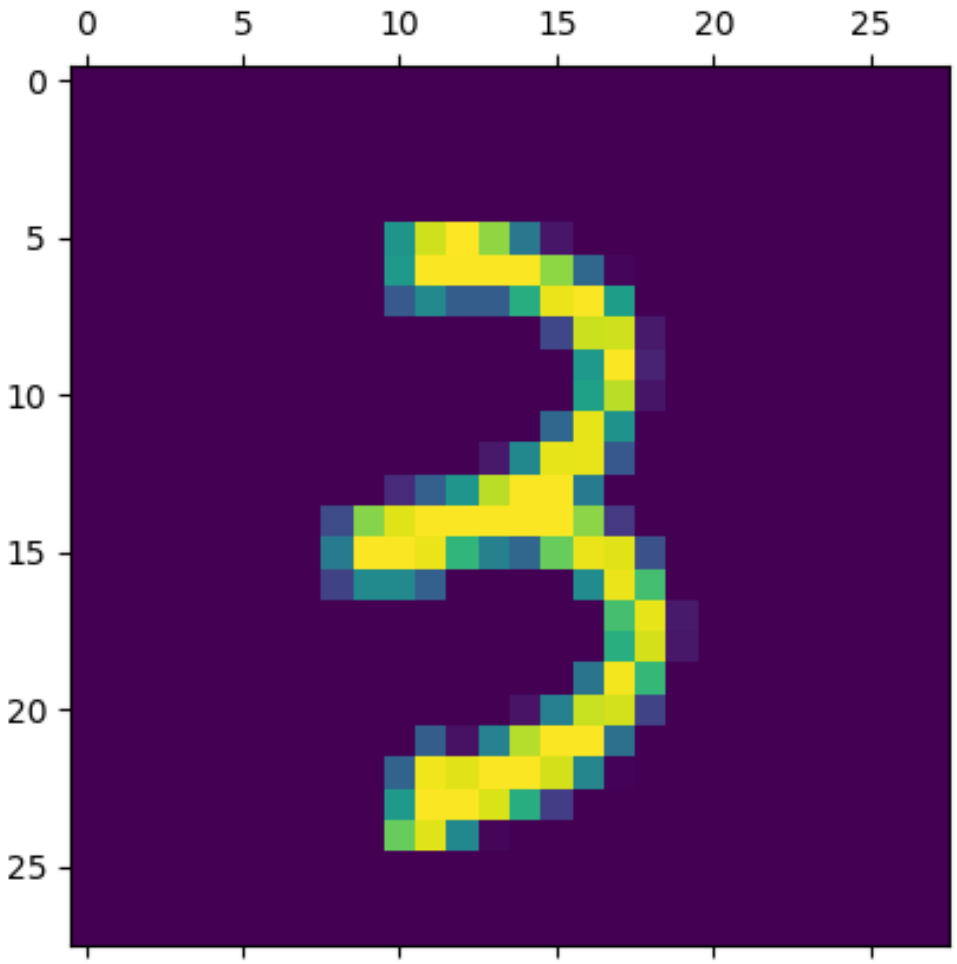

Figure 4: Digit 3

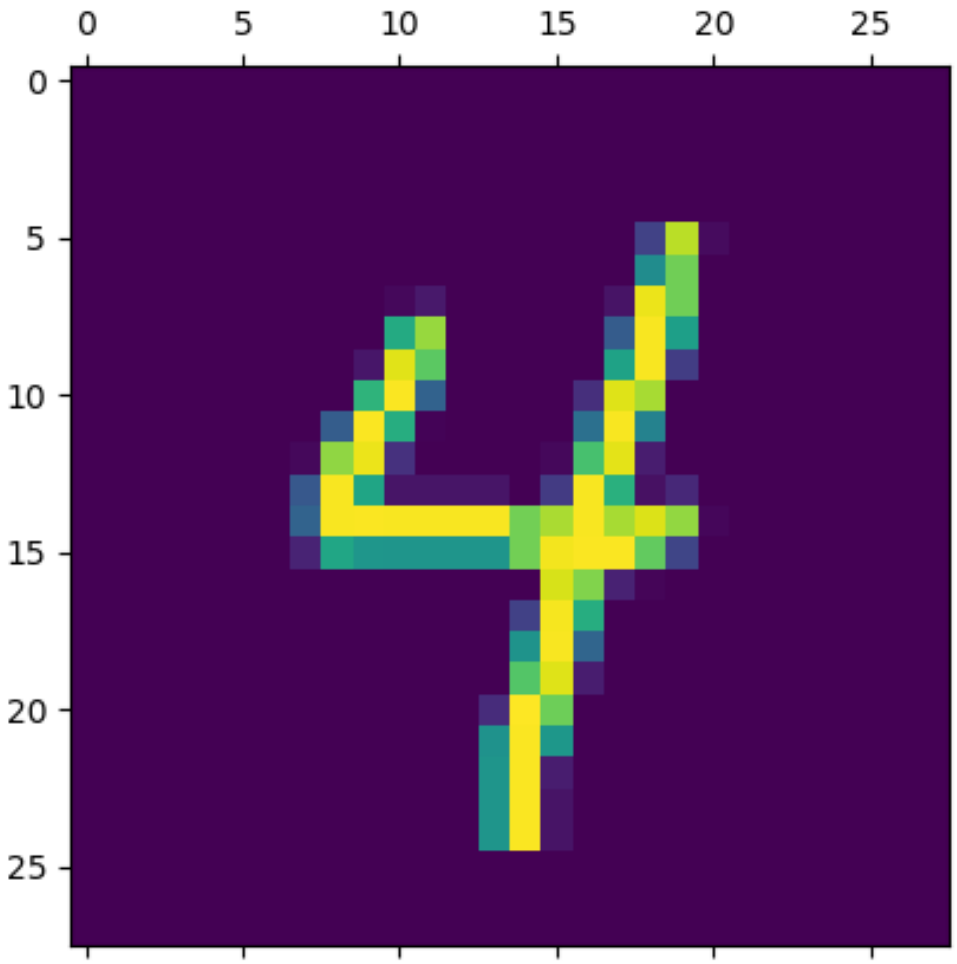

Figure 5: Digit 4

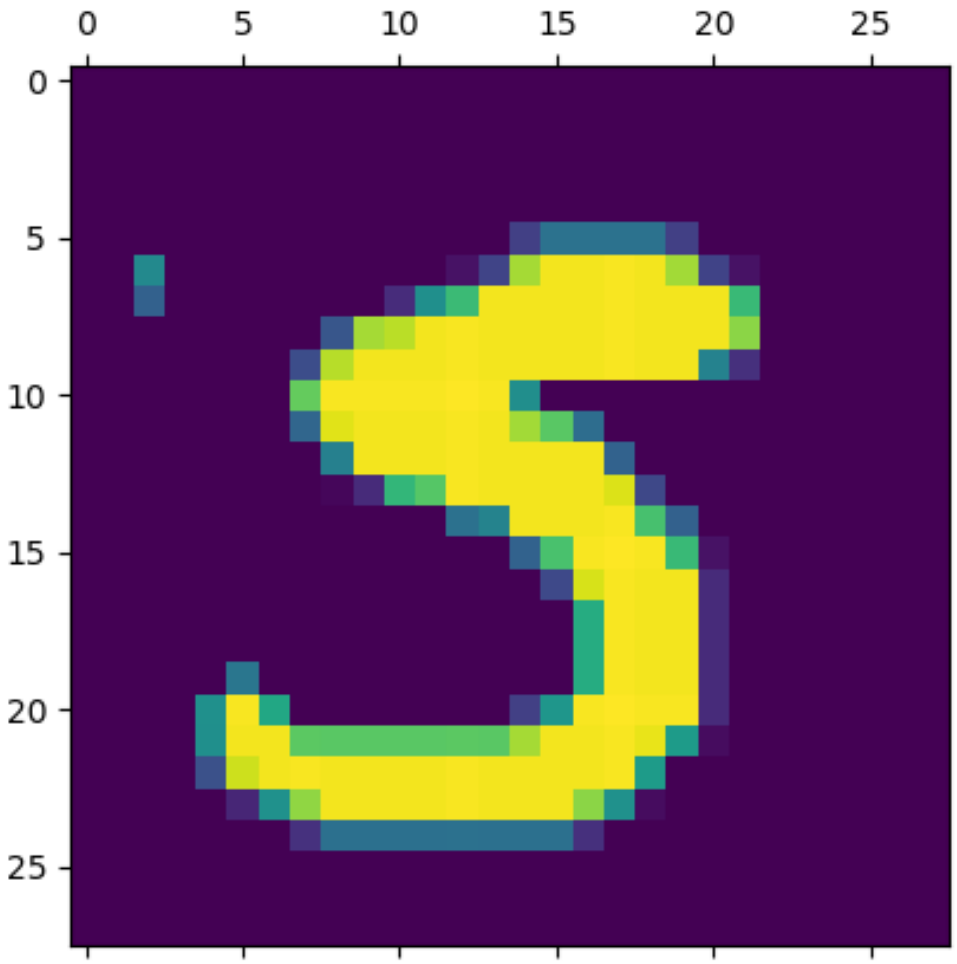

Figure 6: Digit 5

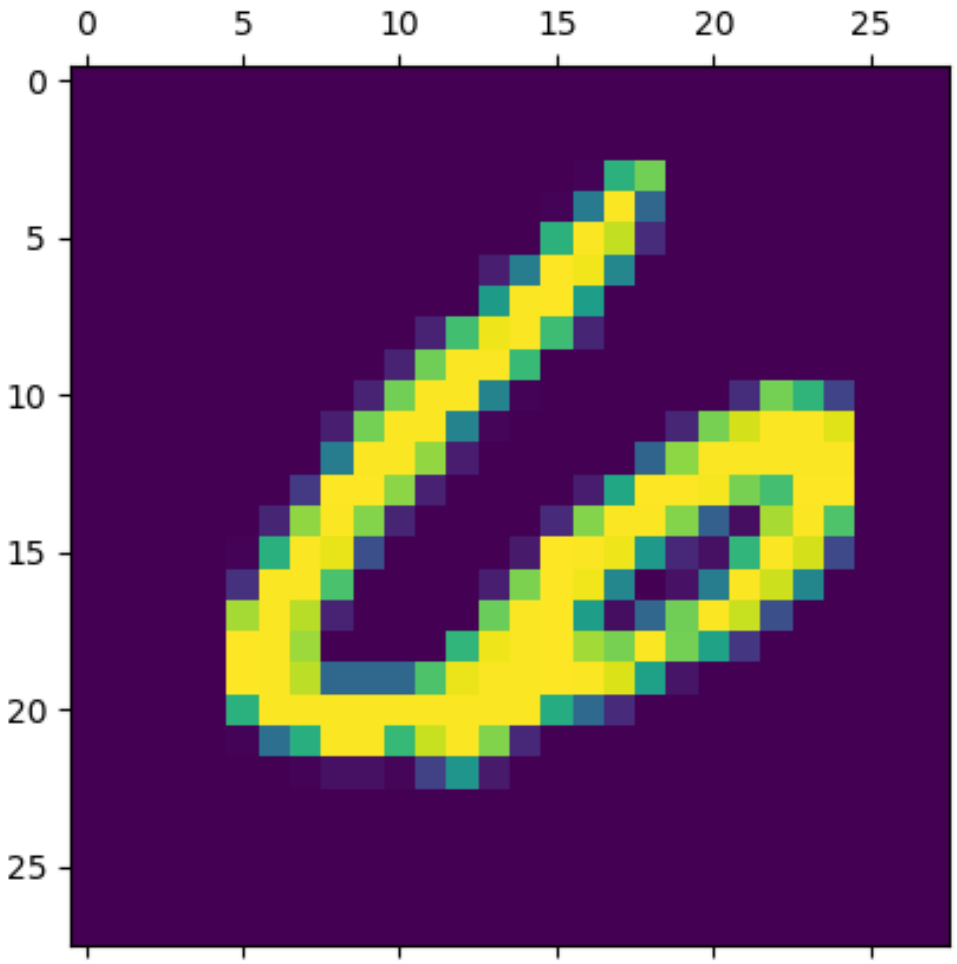

Figure 7: Digit 6

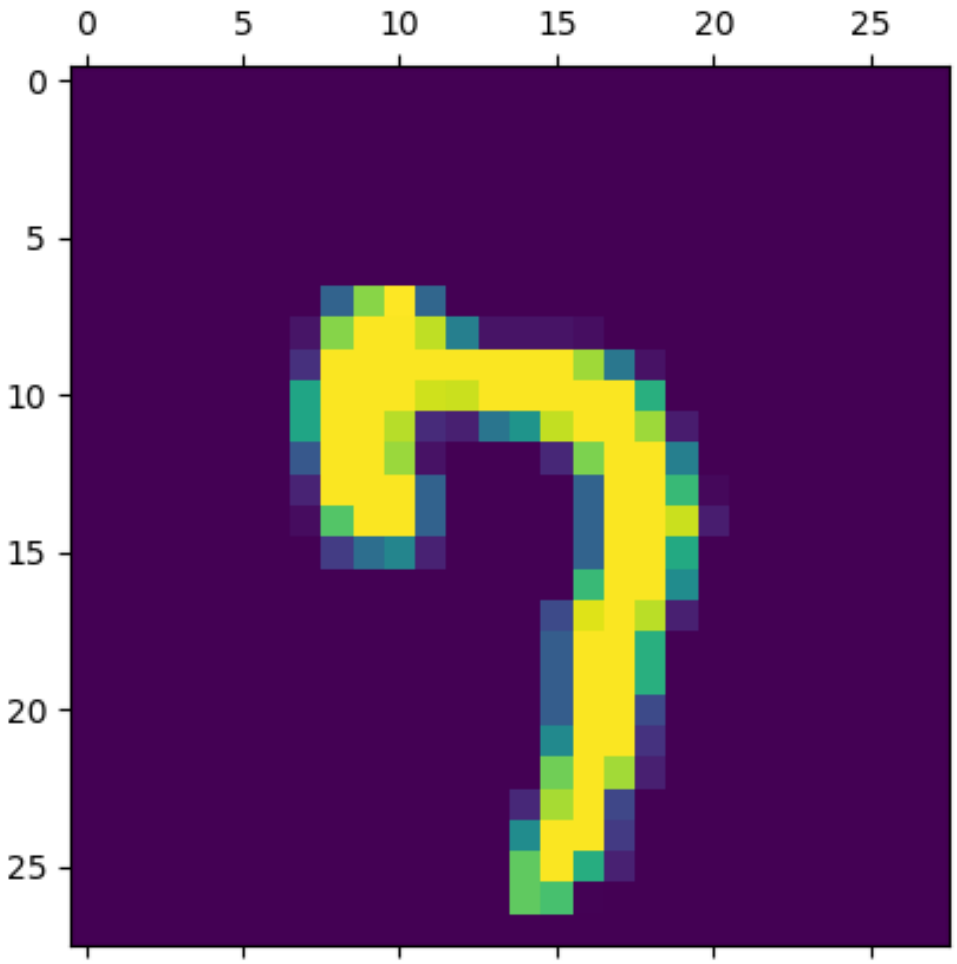

Figure 8: Digit 7

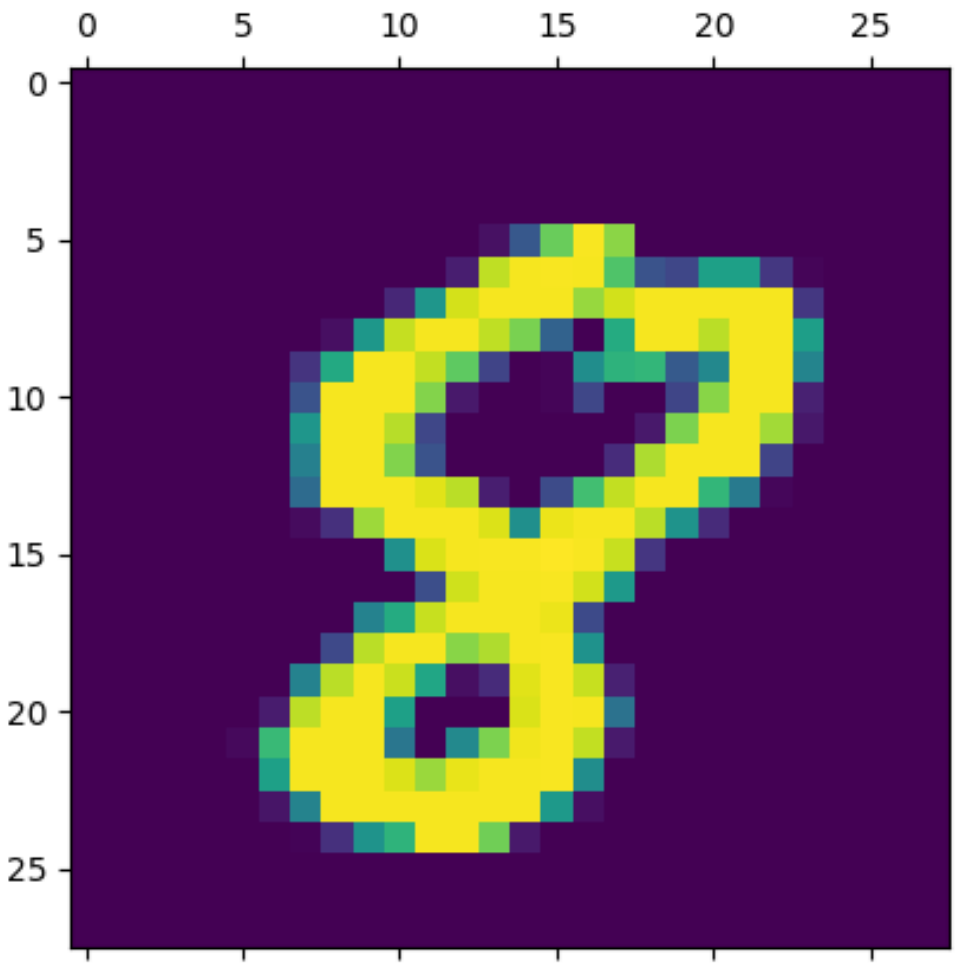

Figure 9: Digit 8

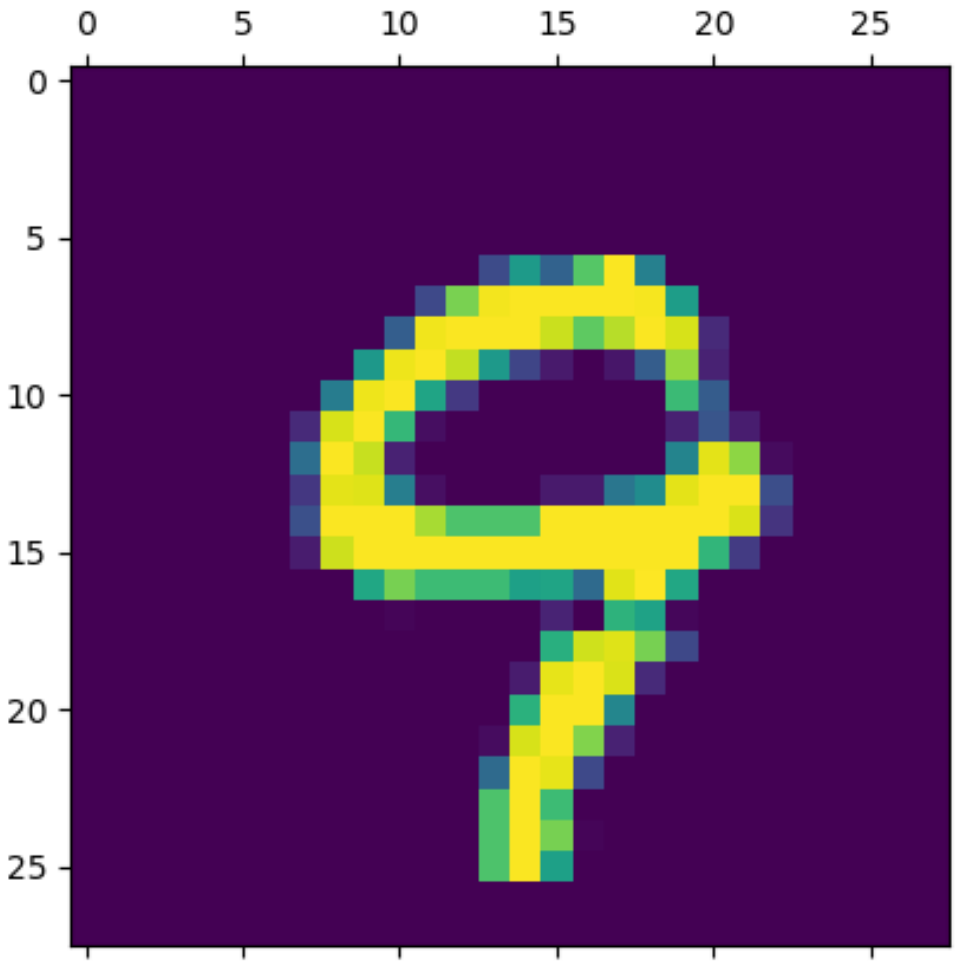

Figure 10: Digit 9

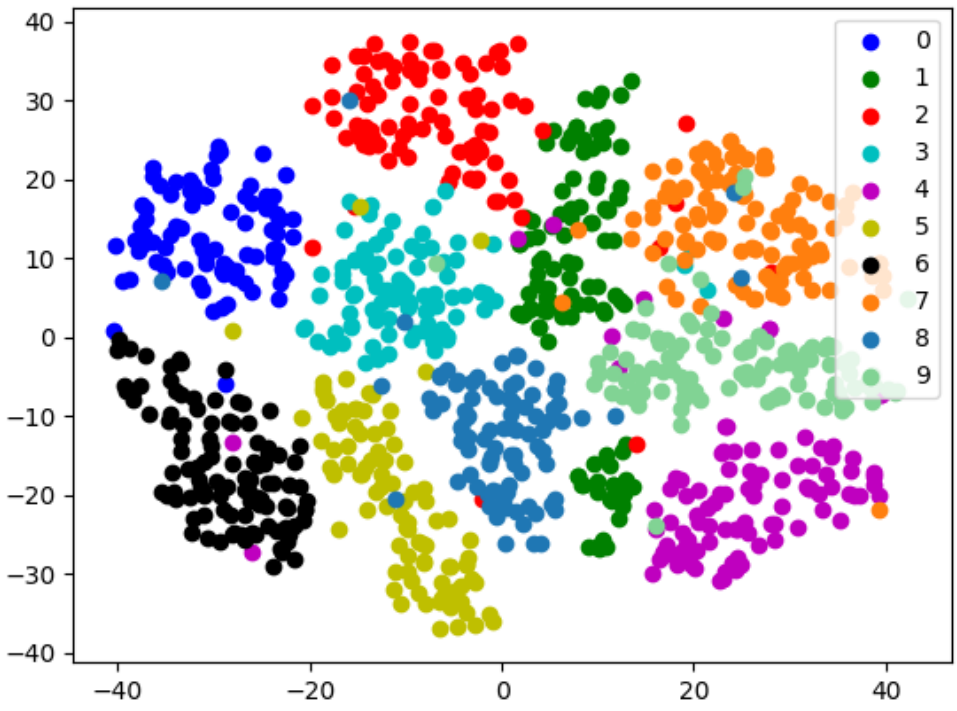

Figure 11: Clustering

```
data 114 $ python3 kmeans.py ../data/digits-embedding.csv 10
WC-SSD: 1433531.469
SC: 0.712
NMI: 0.356
```

Figure 12: Output of running kmeans.py

2.2 Analysis

The code for the analysis and for generating graphs is in **kmeans_analysis.py**. It takes approx. 3 hours to run the analysis on data.cs.purdue.edu.

(i) Figure 13 shows WC-SSD as a function of K for Dataset 1.

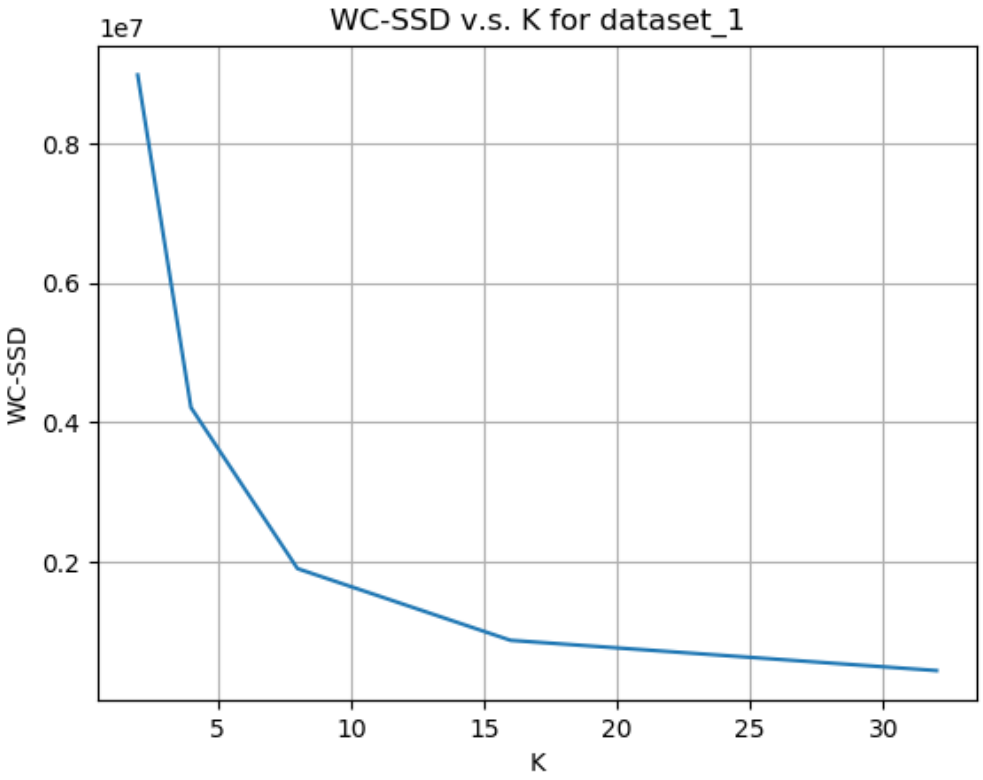

Figure 13: WC-SSD as a function of K for Dataset 1

Figure 14 shows SC as a function of K for Dataset 1.

Figure 15 shows WC-SSD as a function of K for Dataset 2.

Figure 16 shows SC as a function of K for Dataset 2.

Figure 17 shows WC-SSD as a function of K for Dataset 3.

Figure 18 shows SC as a function of K for Dataset 3.

(ii) As we can see from the graphs, for each dataset, the value of WC-SSD increases as K increases. On the other hand, the value of SC decreases as K increases.

Because the SC v.s. K graphs are not very informative, I will use the WC-SSD v.s. K graphs to find the appropriate value of K for each dataset.

- **Dataset 1**: The elbow point in the WC-SSD v.s. K graph (Figure 13) appears to be 8. Therefore, the most appropriate value of K for dataset 1 is $K = 8$.

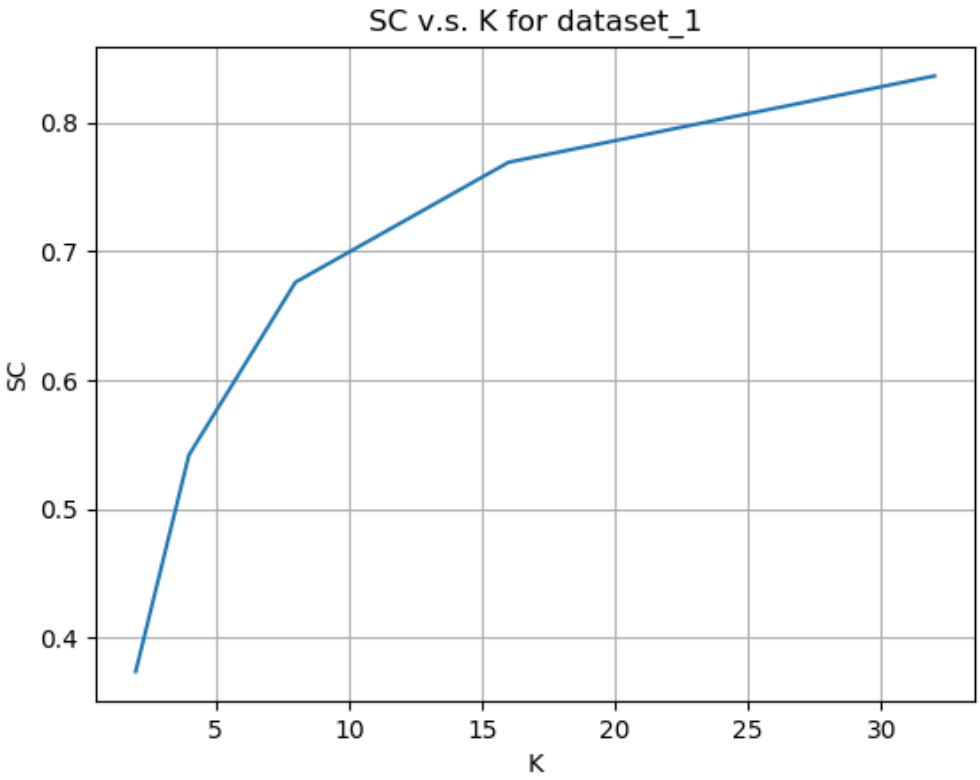

Figure 14: SC as a function of K for Dataset 1

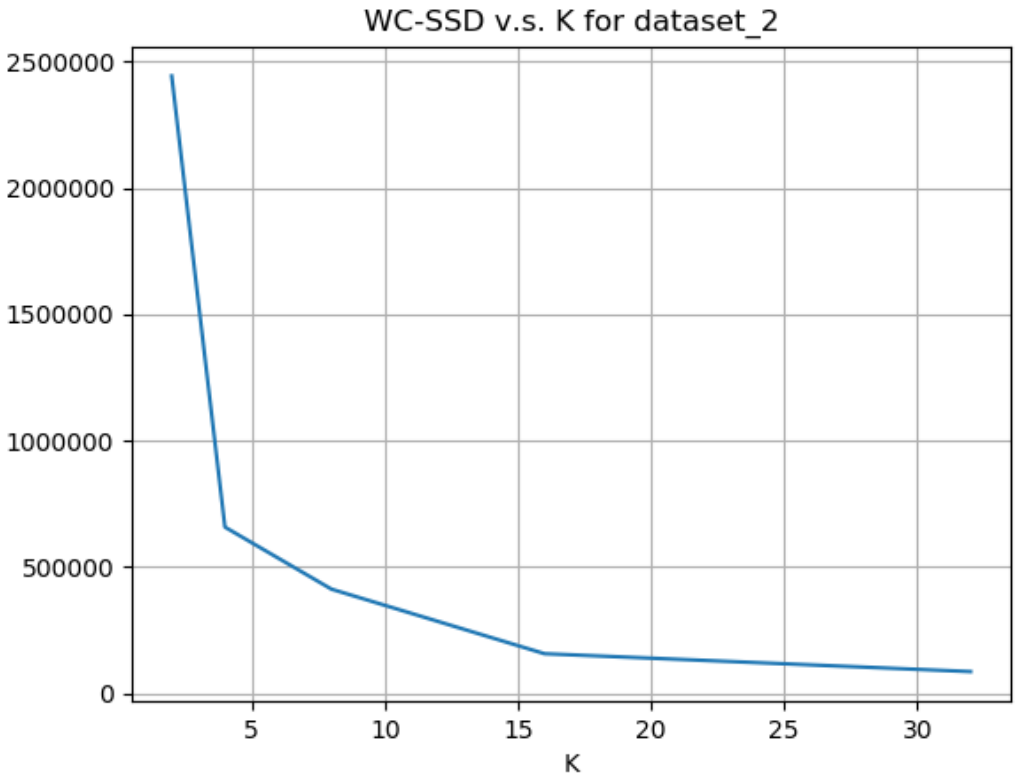

Figure 15: WC-SSD as a function of K for Dataset 2

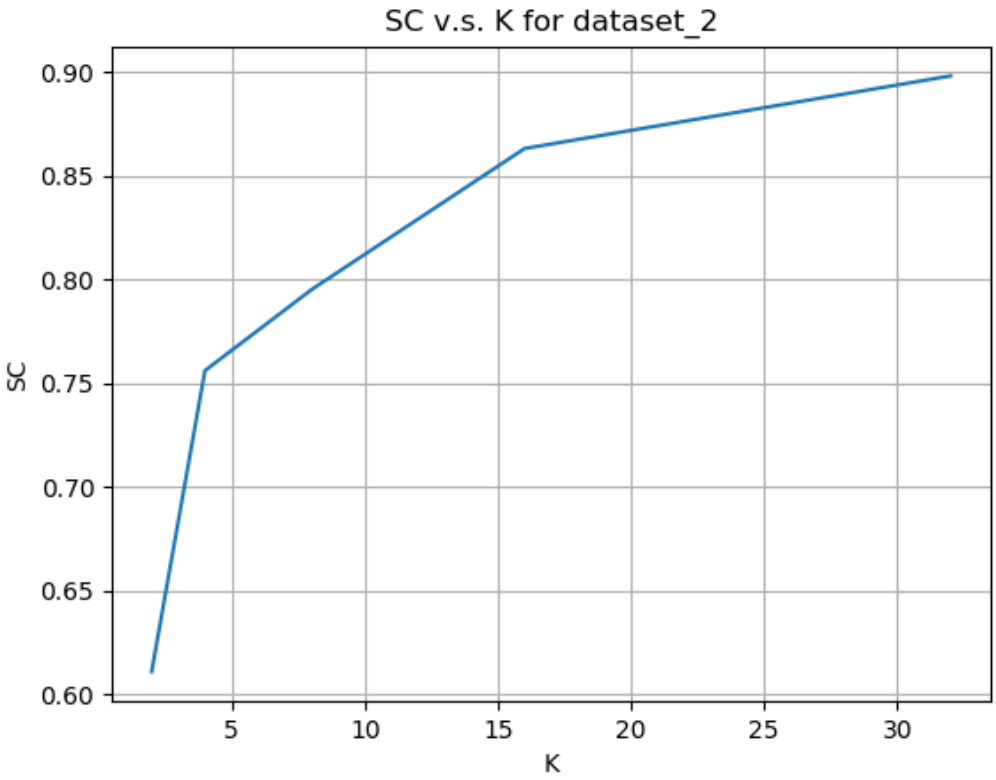

Figure 16: SC as a function of K for Dataset 2

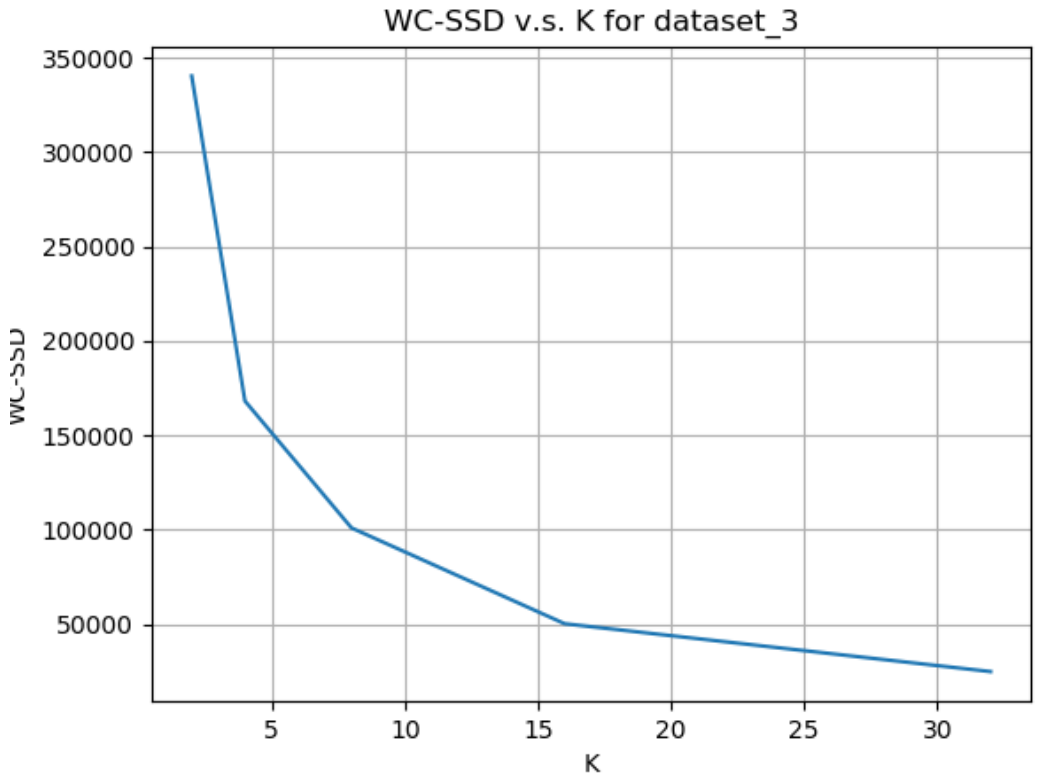

Figure 17: WC-SSD as a function of K for Dataset 3

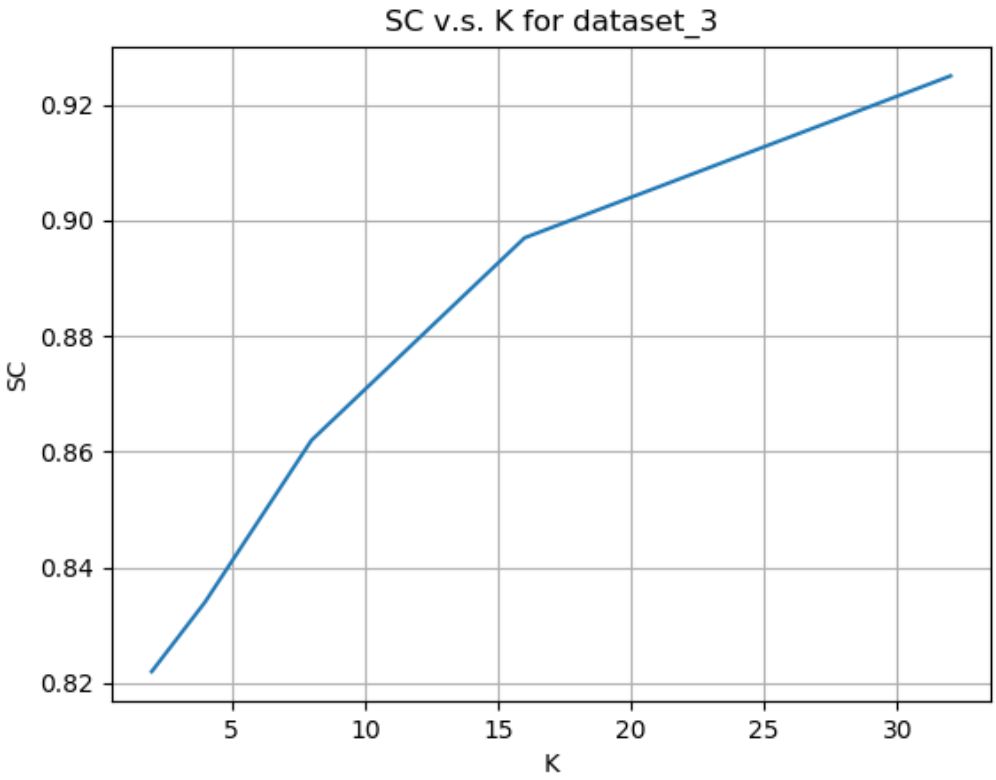

Figure 18: SC as a function of K for Dataset 3

- **Dataset 2**: The elbow point in the WC-SSD v.s. K graph (Figure 15) appears to be 4. Therefore, the most appropriate value of K for dataset 2 is $K = 4$.
- **Dataset 3**: This graph is pretty smooth and it is hard to easily pinpoint an elbow point. However, the elbow point in the WC-SSD v.s. K graph (Figure 17) appears to be 8. Therefore, the most appropriate value of K for dataset 3 is $K = 8$. However, it should actually have been 2 as there as only two classes that we are considering.

(iii) Figure 19 shows the average and standard deviation of WC-SSD for different values of K for Dataset 1.

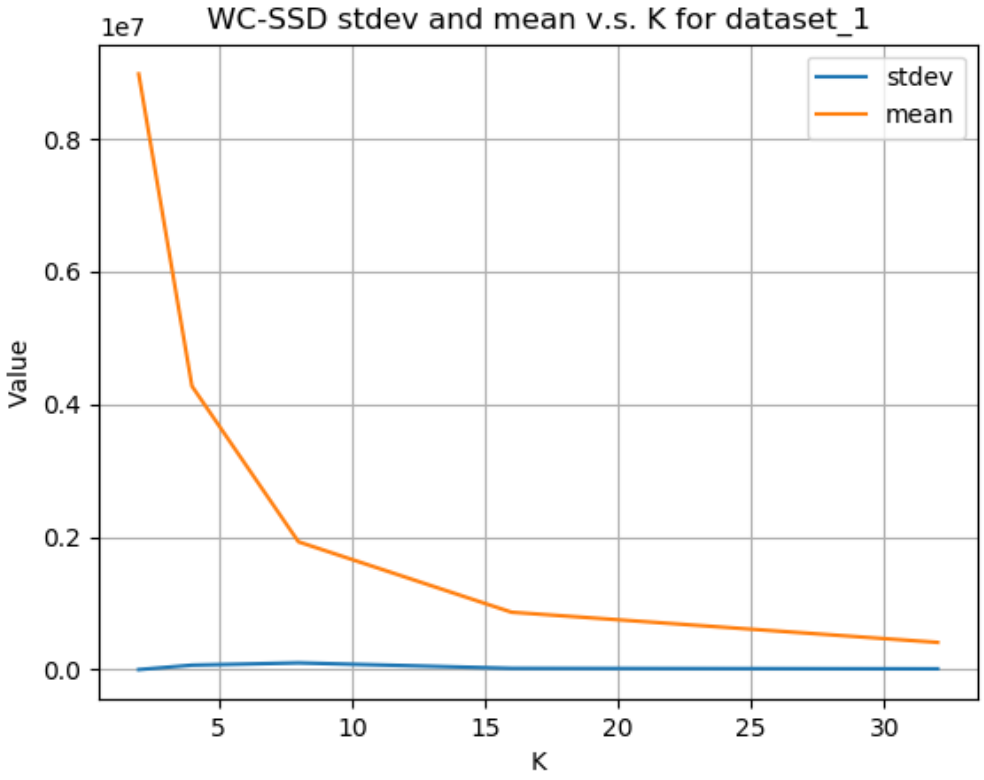

Figure 19: Average and standard deviation of WC-SSD for different values of K for Dataset 1

Figure 20 shows the average and standard deviation of SC for different values of K for Dataset 1.

Figure 21 shows the average and standard deviation of WC-SSD for different values of K for Dataset 2.

Figure 22 shows the average and standard deviation of SC for different values of K for Dataset 2.

Figure 23 shows the average and standard deviation of WC-SSD for different values of K for Dataset 3.

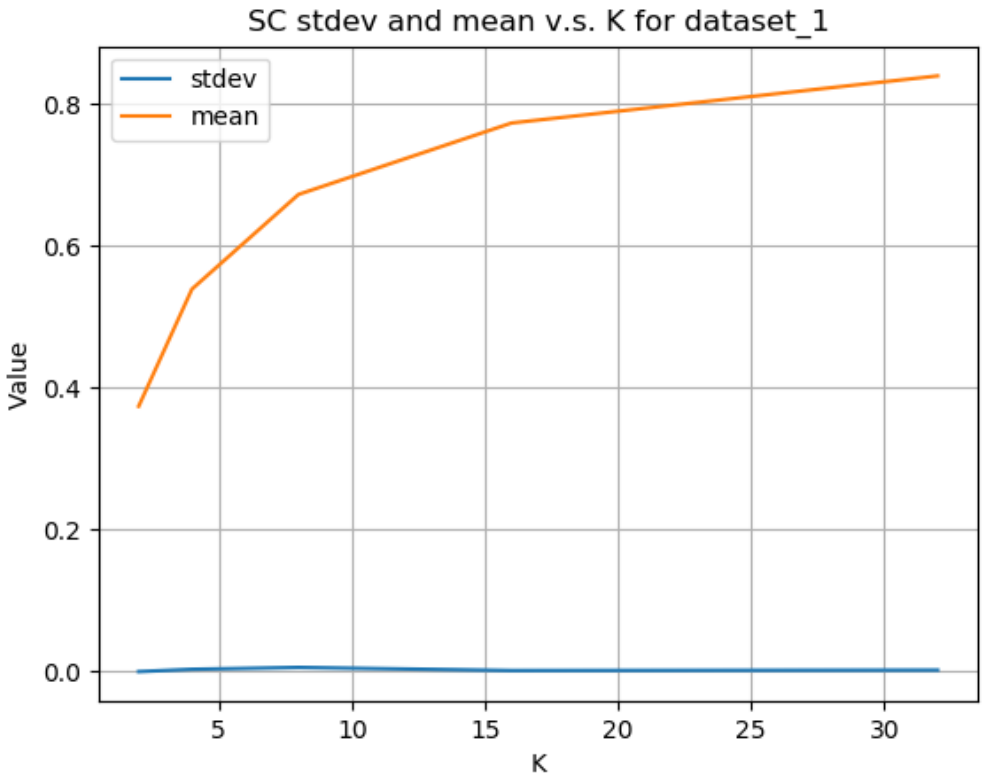

Figure 20: Average and standard deviation of SC for different values of K for Dataset 1

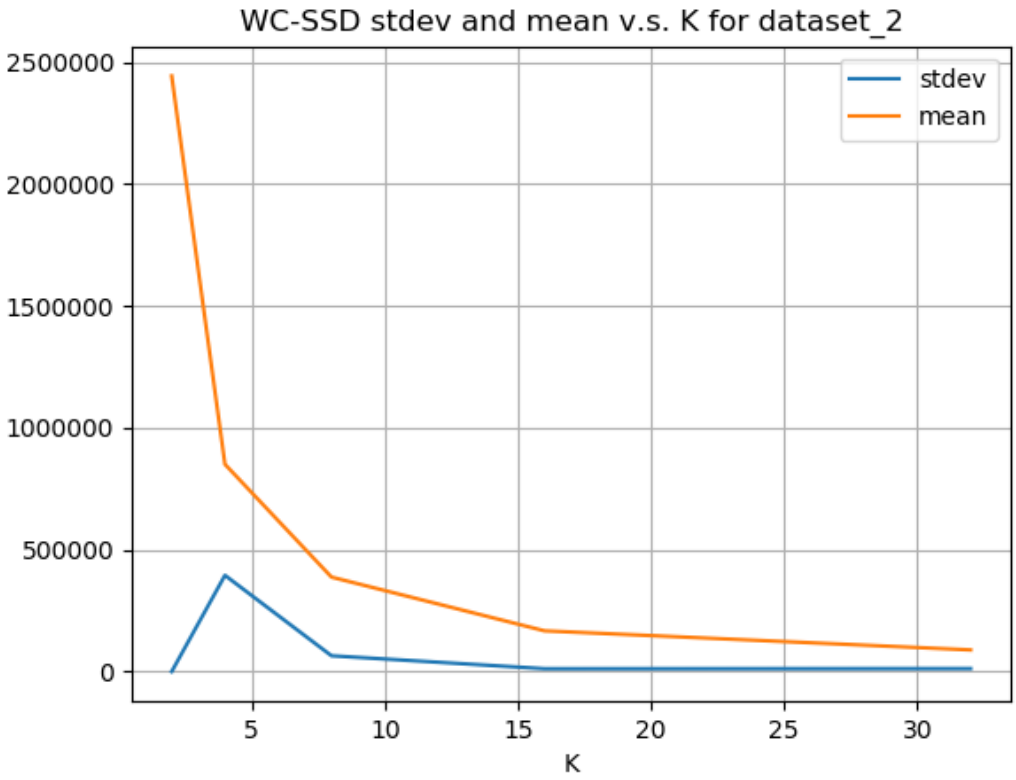

Figure 21: Average and standard deviation of WC-SSD for different values of K for Dataset 2

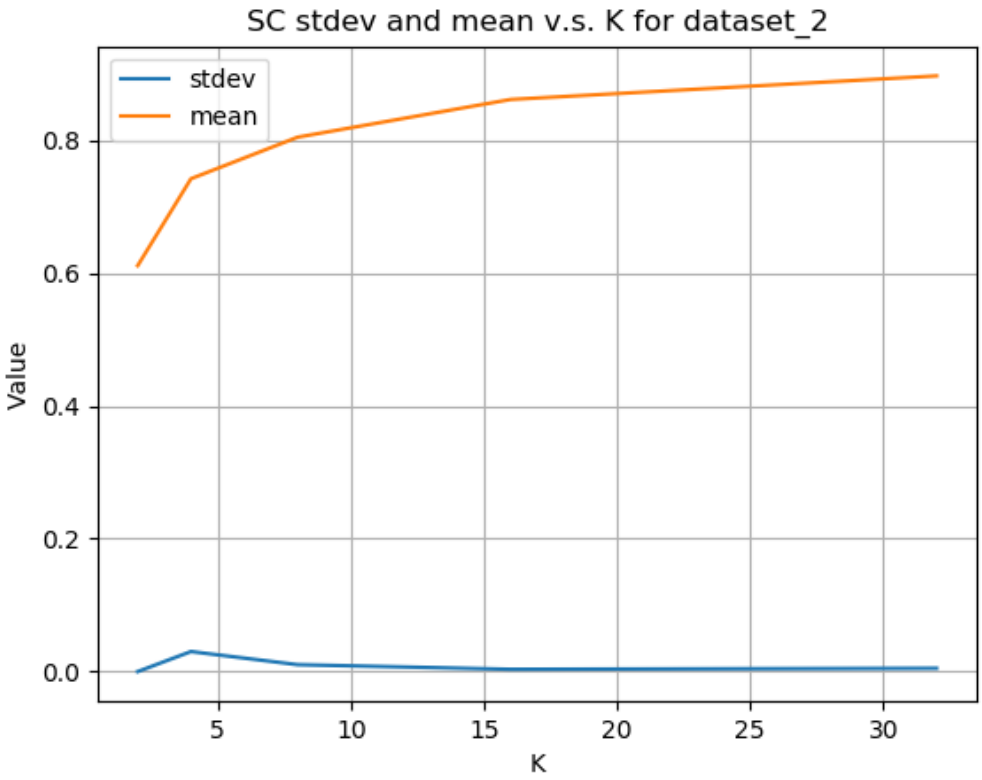

Figure 22: Average and standard deviation of SC for different values of K for Dataset 2

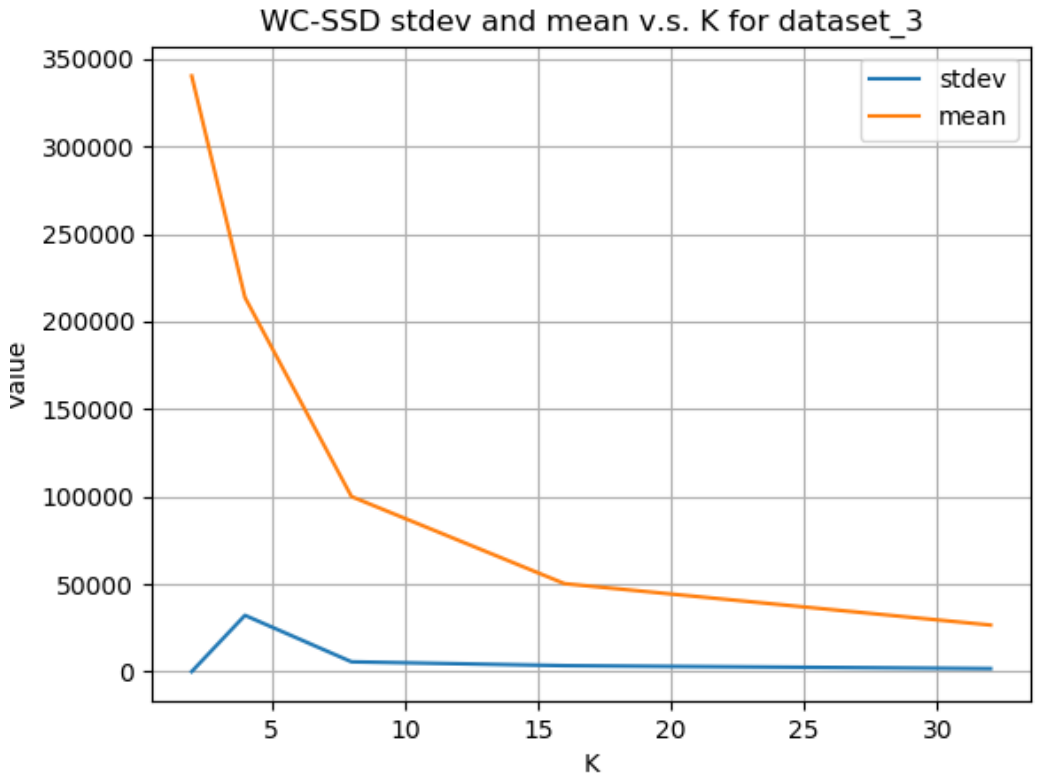

Figure 23: Average and standard deviation of WC-SSD for different values of K for Dataset 3

Figure 24 shows the average and standard deviation of SC for different values of K for Dataset 3.

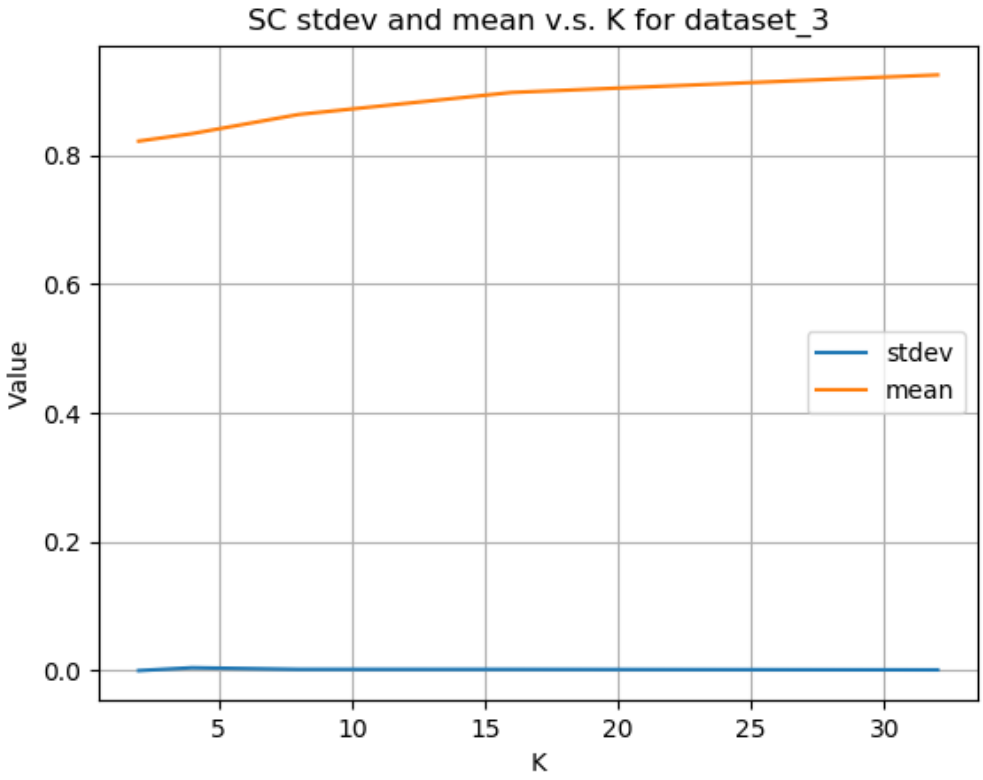

Figure 24: Average and standard deviation of SC for different values of K for Dataset 3

I will again use the WC-SSD graphs as the SC graphs are not very informative. For each of the three datasets, we notice that there is a peak/change in the curve for standard deviation at the value of K we chose for part 2. The standard deviation then levels off for other values of K. This implies that K-Means is sensitive to the initial k-means chosen and therefore, we must choose those as well as we can because if not chosen correctly, it might lead to a local optimum solution.

(iv) NMI values for each of the 3 datasets is shown in Figure 25.

Figure 26 shows 1000 randomly chosen examples visualized with their corresponding clusters colored for dataset 1.

Figure 27 shows 1000 randomly chosen examples visualized with their corresponding clusters colored for dataset 2.

Figure 28 shows 1000 randomly chosen examples visualized with their corresponding clusters colored for dataset 3.

We can see from the visualizations that the clustering for the respective datasets is good as there are distinct clusters showing in the visualizations.

```
=== PART 2.4 ===
NMI for dataset 1 with K = 8: 0.355
NMI for dataset 2 with K = 4: 0.456
NMI for dataset 3 with K = 8: 0.246
```

Figure 25: NMI values for each of the 3 datasets for values of K chosen in part (ii)

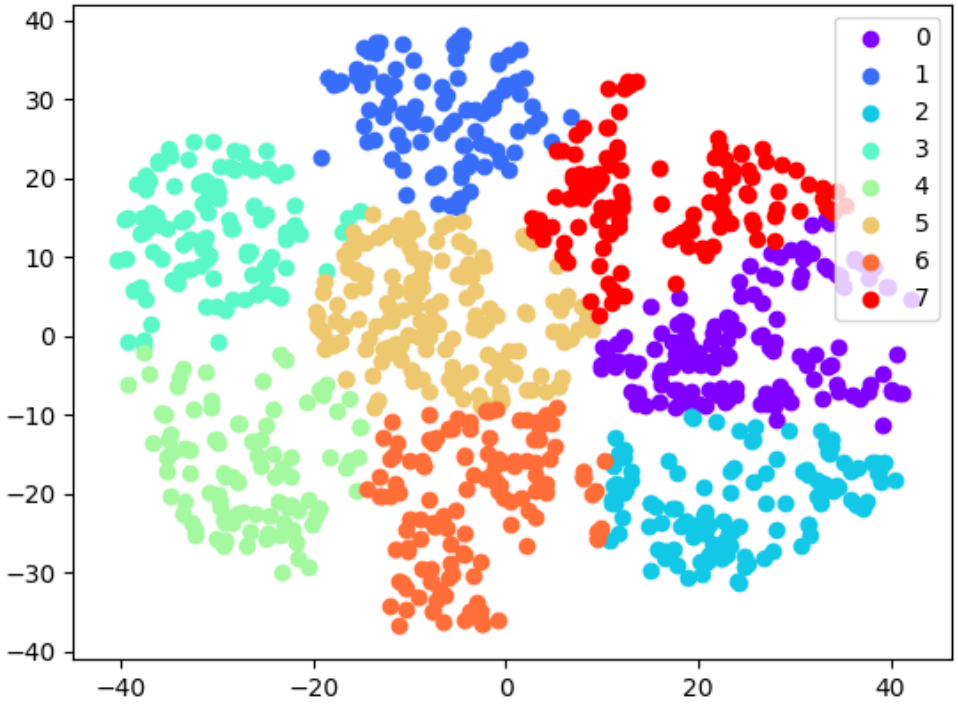

Figure 26: Clustering of Dataset 1 with K = 8

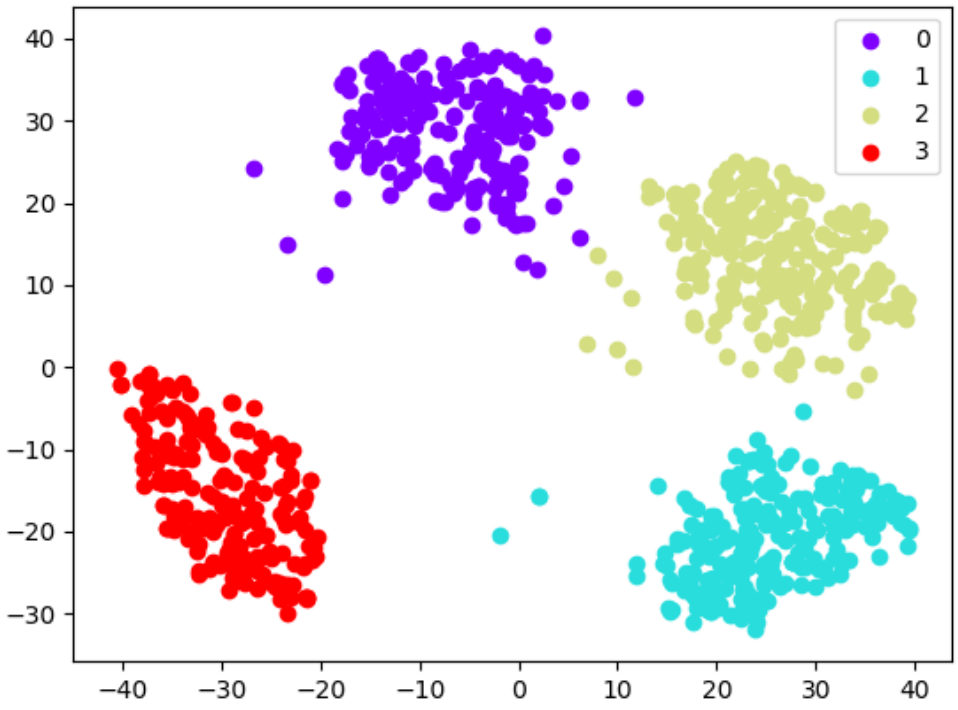

Figure 27: Clustering of Dataset 2 with K = 4

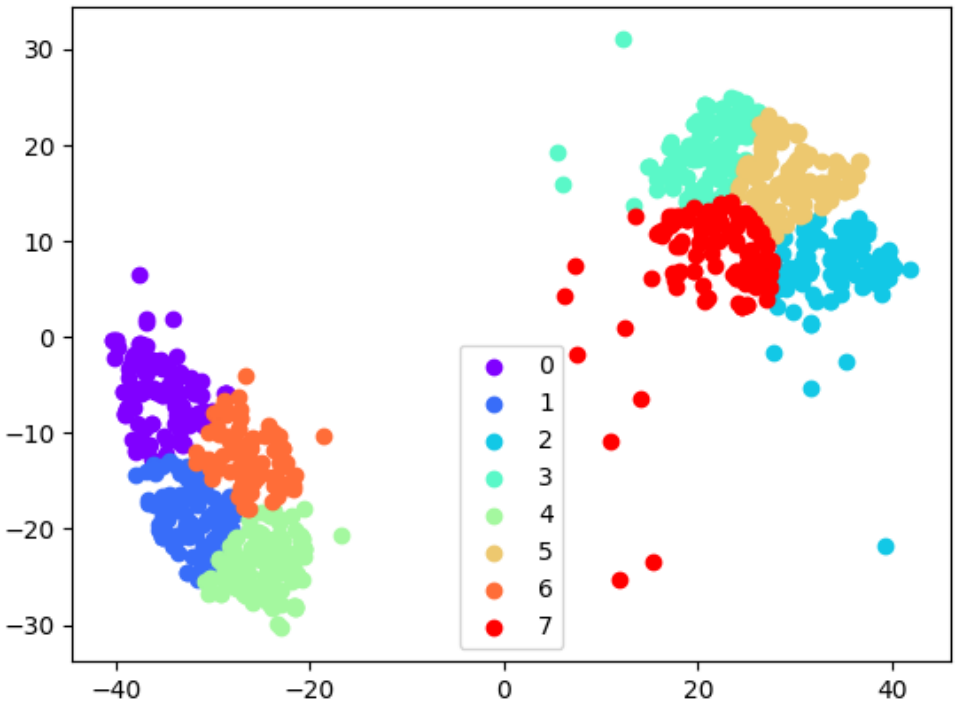

Figure 28: Clustering of Dataset 3 with K = 8

This observation is backed up by the NMI values obtained. A higher NMI value corresponds to a better clustering. For each datasets 1 and 2, the NMI value is > 3.5 and close to 0.5, which indicates that the clustering is good. However, the NMI value for dataset $3 < 3.5$, which implies that this is not a very good clustering and therefore, our choice of K is not very good for this dataset and should have been $K = 2$.

3 Hierarchical Clustering

(i) Figure 29 shows the dendogram made using single linkage.

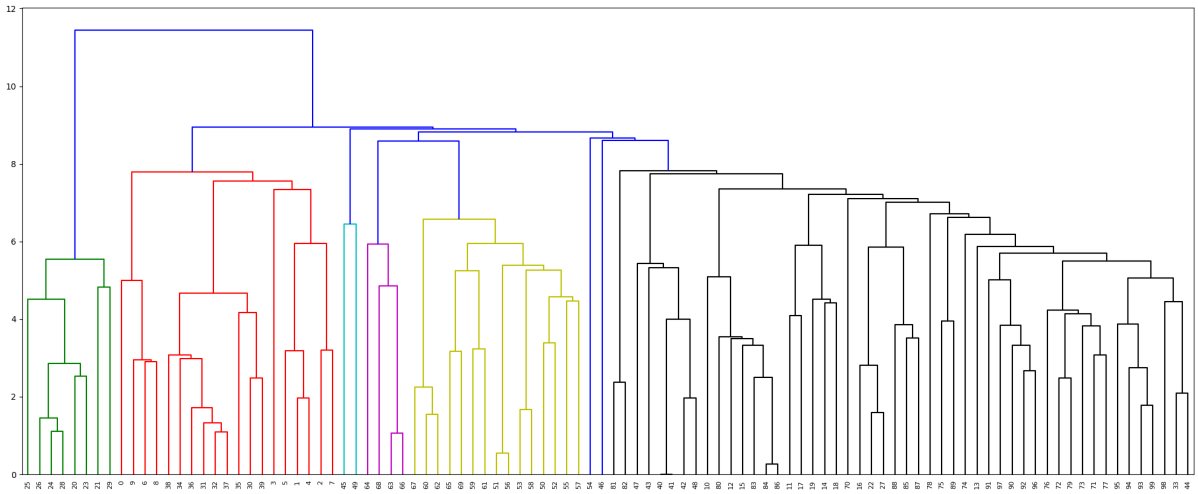

Figure 29: Dendogram using single linkage

(ii) Figure 30 shows the dendogram made using complete linkage.

Figure 31 shows the dendogram made using average linkage.

(iii) Figure 32 shows WC-SSD as a function of K for Single Linkage.

Figure 33 shows SC as a function of K for Single Linkage.

Figure 34 shows WC-SSD as a function of K for Complete Linkage.

Figure 35 shows SC as a function of K for Complete Linkage.

Figure 36 shows WC-SSD as a function of K for Average Linkage.

Figure 37 shows SC as a function of K for Average Linkage.

(iv)
- **Single Linkage**: The elbow point in the WC-SSD v.s. K graph (Figure 32) appears to be 8. Therefore, the most appropriate value of K for with Single Linkage is $K = 8$. This choice of K does not differ from my choice of K for Dataset 1 in Section 2.

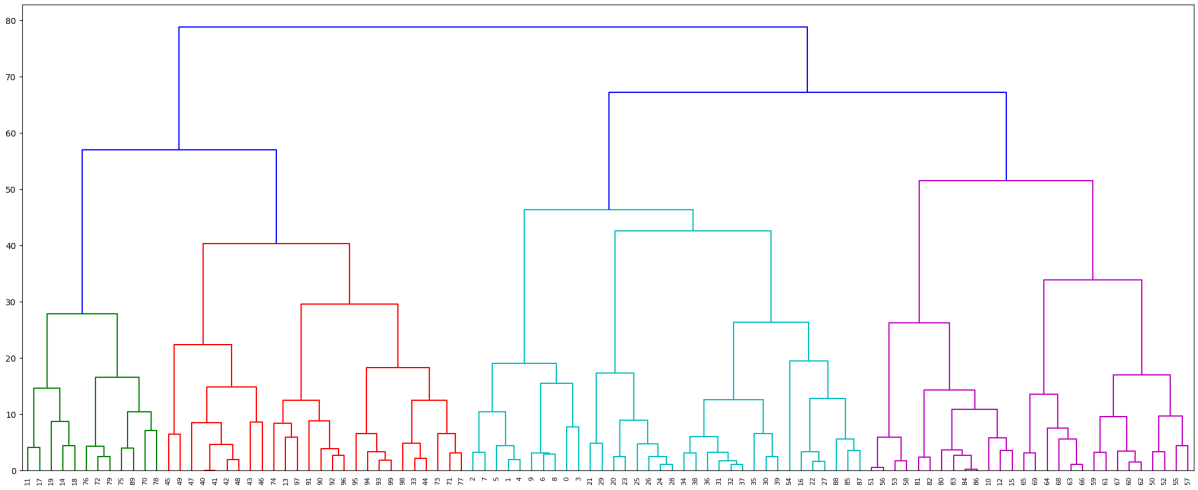

Figure 30: Dendogram using complete linkage

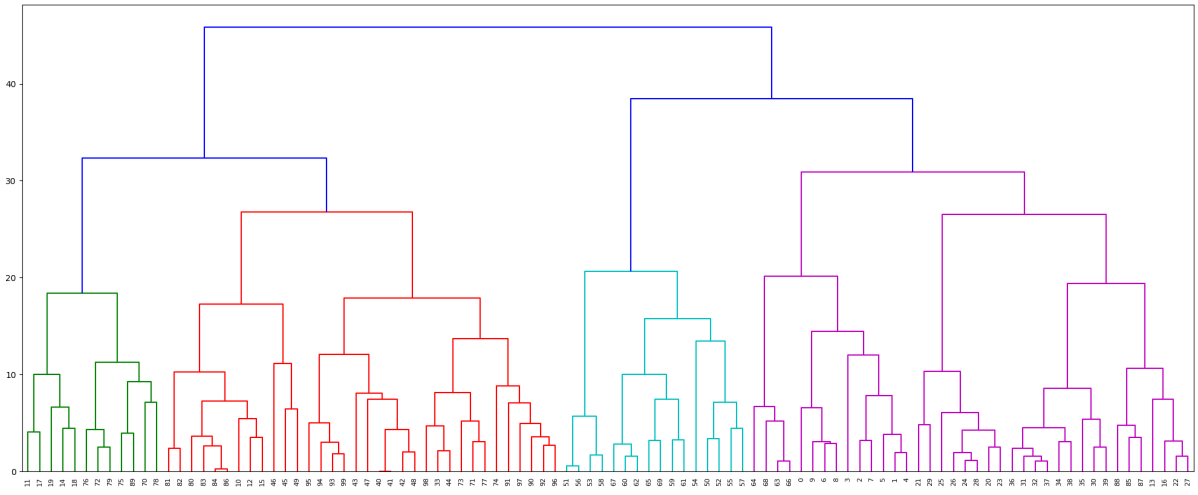

Figure 31: Dendogram using average linkage

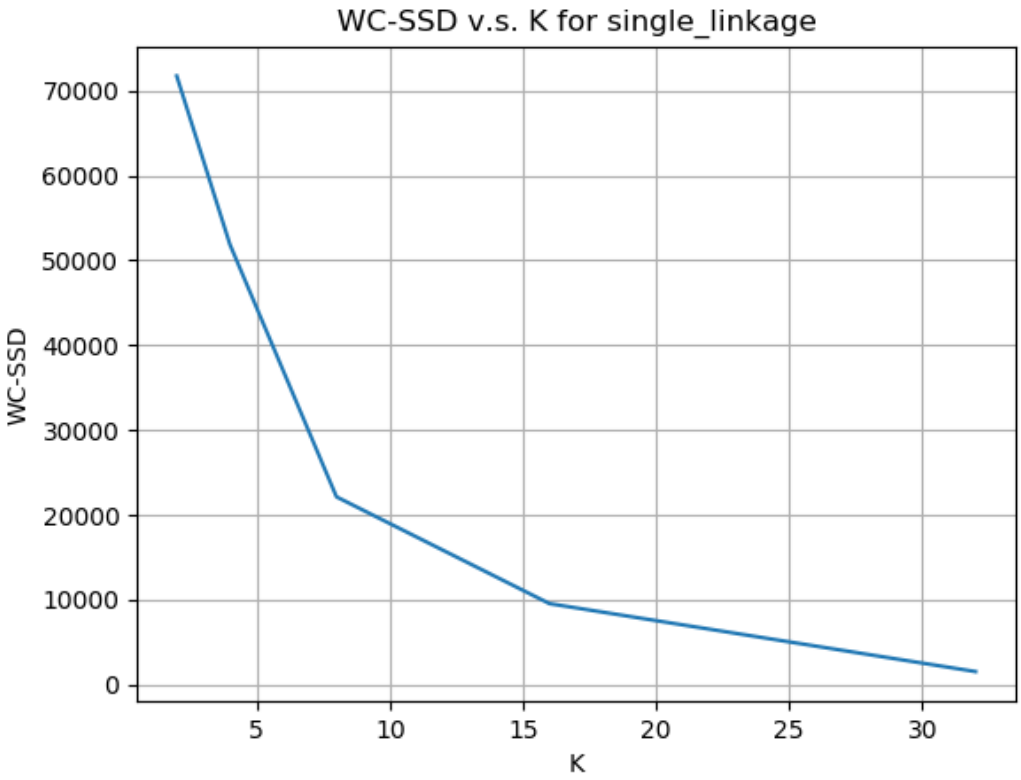

Figure 32: WC-SSD as a function of K for Single Linkage

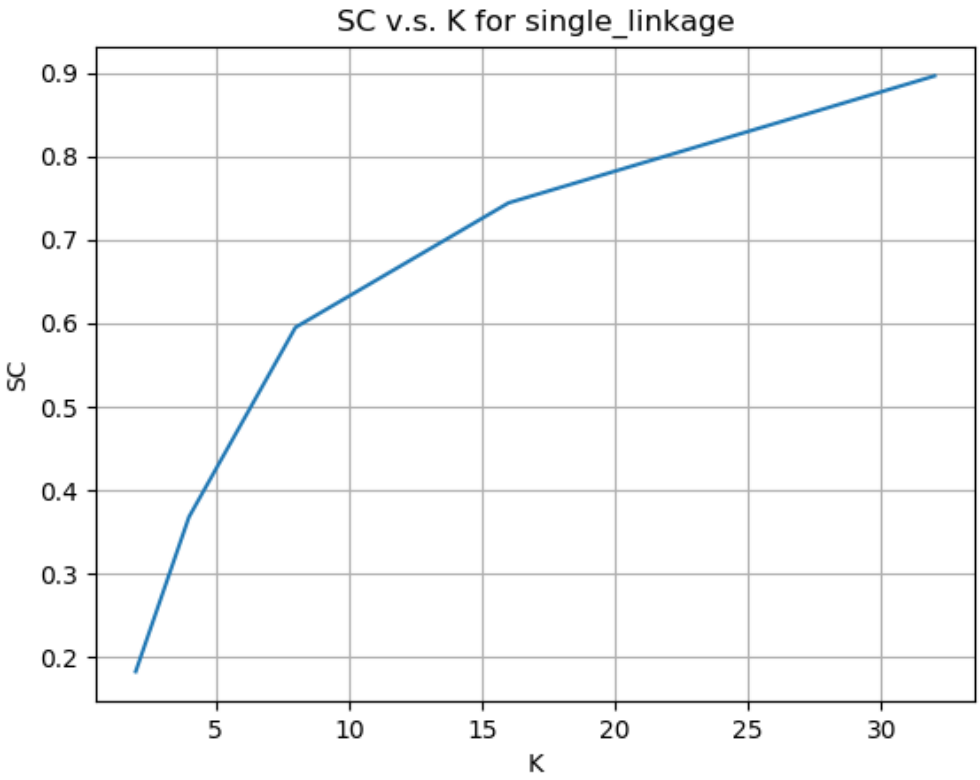

Figure 33: SC as a function of K for Single Linkage

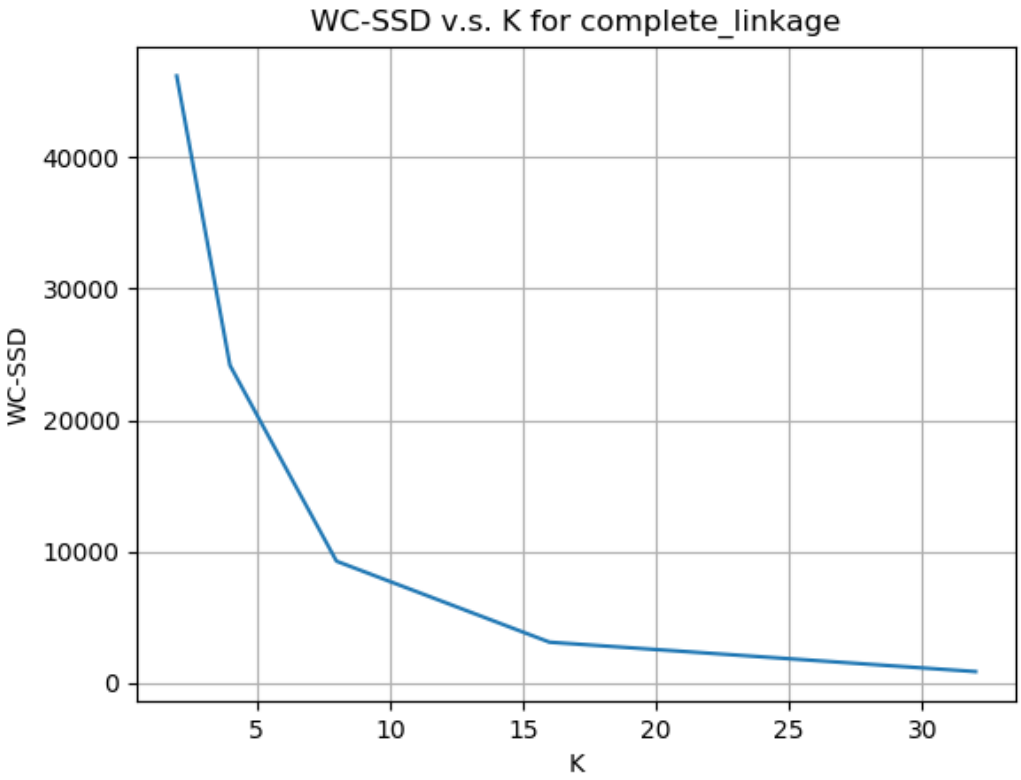

Figure 34: WC-SSD as a function of K for Complete Linkage

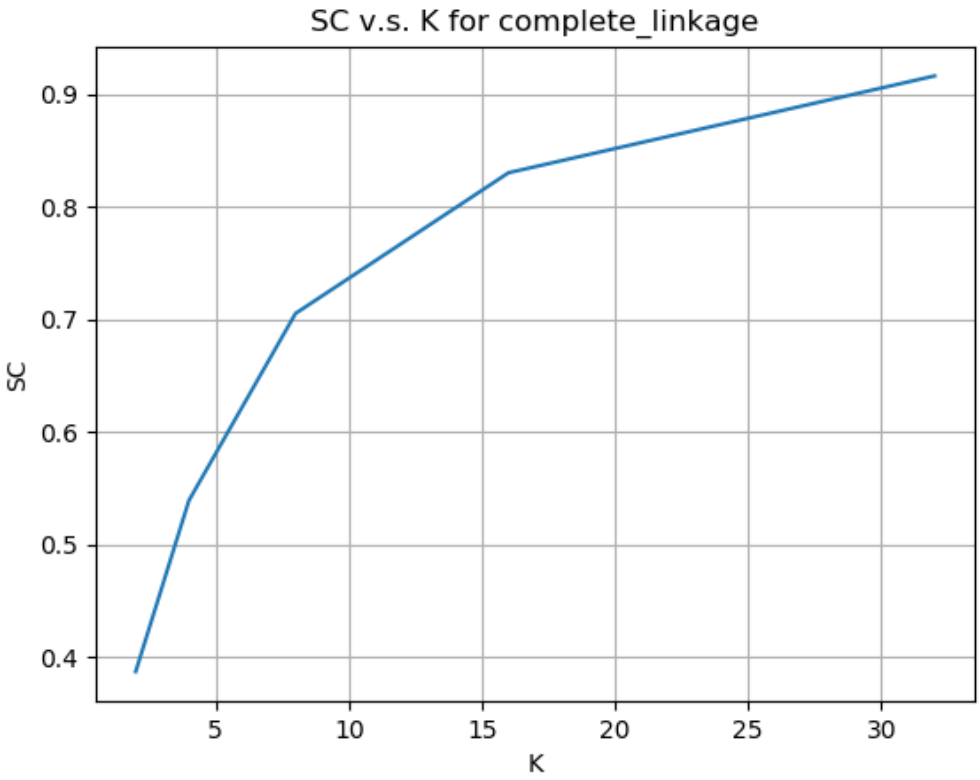

Figure 35: SC as a function of K for Complete Linkage

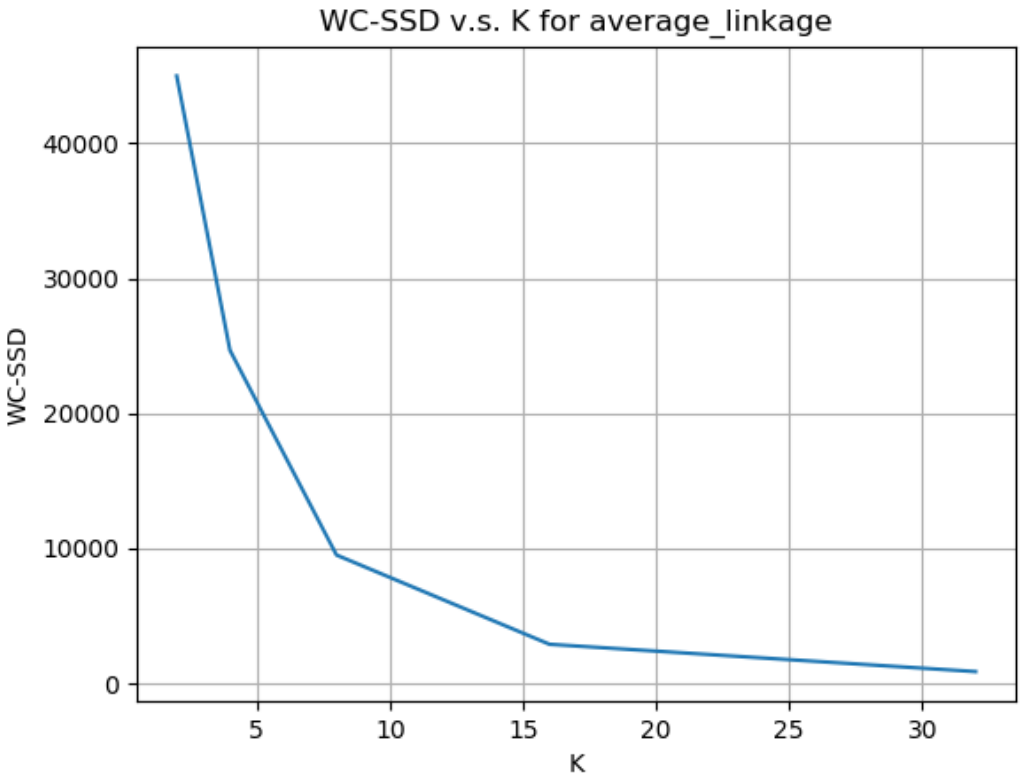

Figure 36: WC-SSD as a function of K for Average Linkage

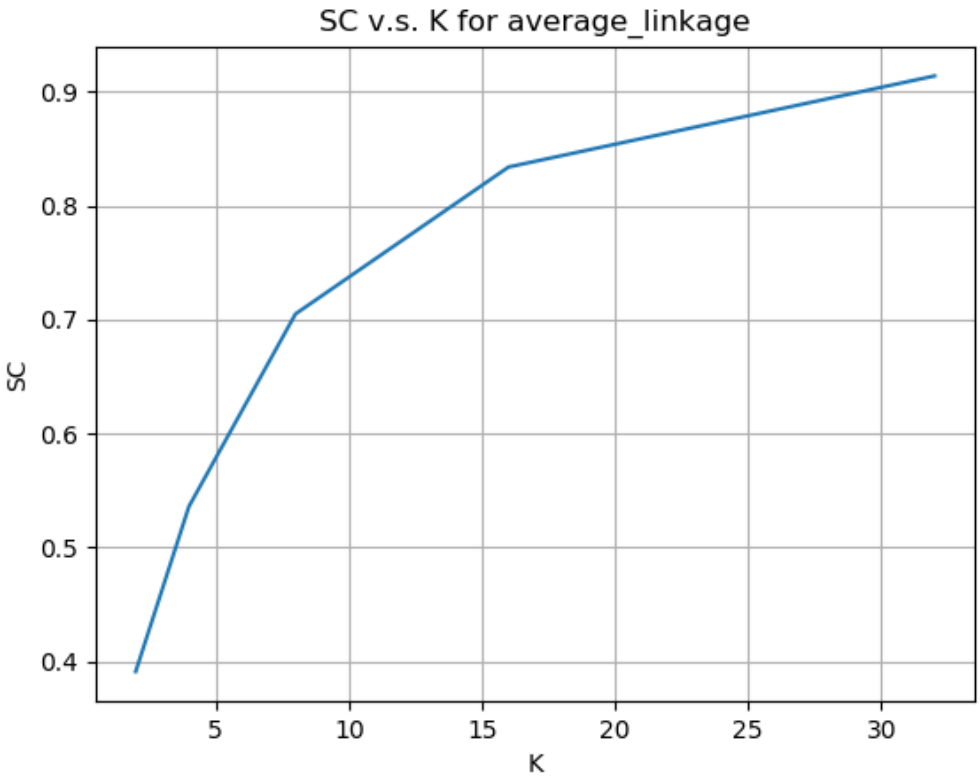

Figure 37: SC as a function of K for Average Linkage

- **Complete Linkage**: The elbow point in the WC-SSD v.s. K graph (Figure 34) appears to be 8. Therefore, the most appropriate value of K for Complete Linkage is $K = 8$. This choice of K does not differ from my choice of K for Dataset 1 in Section 2.
- **Average Linkage**: The elbow point in the WC-SSD v.s. K graph (Figure 36) appears to be 8. Therefore, the most appropriate value of K for Average Linkage is $K = 8$. This choice of K does not differ from my choice of K for Dataset 1 in Section 2.

(v) NMI values for each of the 3 measures is shown in Figure 38. The best NMI is achieved with single linkage and then complete linkage. As compared to the results from k-means on Dataset 1 in Section 2, where NMI was equal to 3.55, we get a better NMI with hierarchical clustering using any of the three methods.

```
=== PART 3.5 ===
NMI for algo/distance measure = single_linkage with K = 8: 0.509
NMI for algo/distance measure = complete_linkage with K = 8: 0.385
NMI for algo/distance measure = average_linkage with K = 8: 0.373
```

Figure 38: NMI values for each of the 3 distance measures for values of K chosen in part (iv).