Ayush Garg
PUID: 0033794803

Homework 5

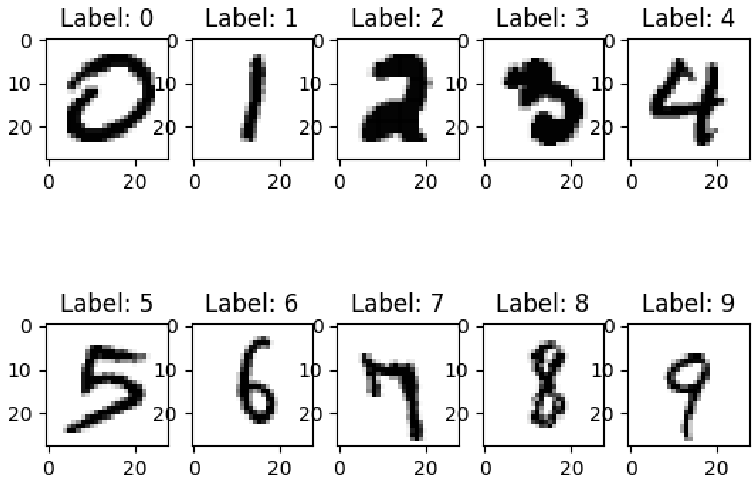
CS 57300

Number of late days used : 1

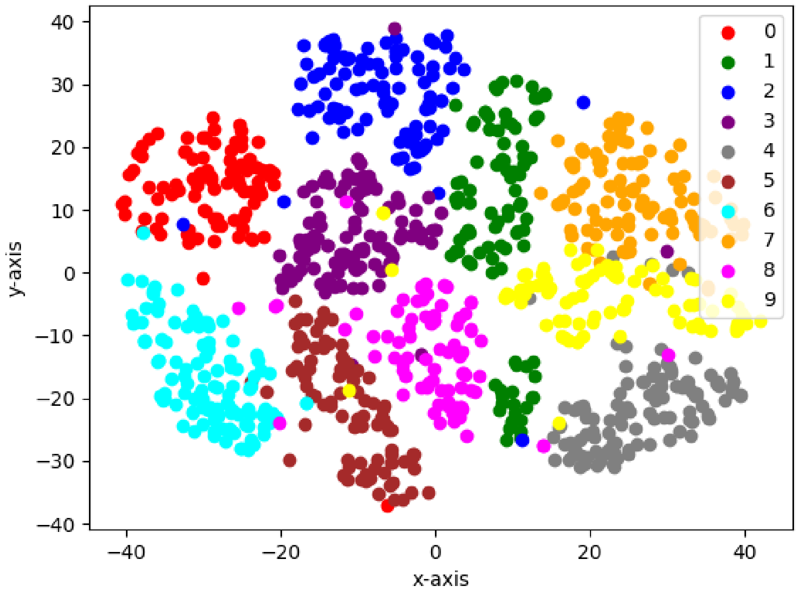
1. Q1

```
Command to be entered on terminal: python exploration.py
Output: Displays two graphs, one for visualising the data point from
each class and other to visualize 1000 randomly selected points
```

1. Visualising the data points:



2. Visualising 1000 random examples:

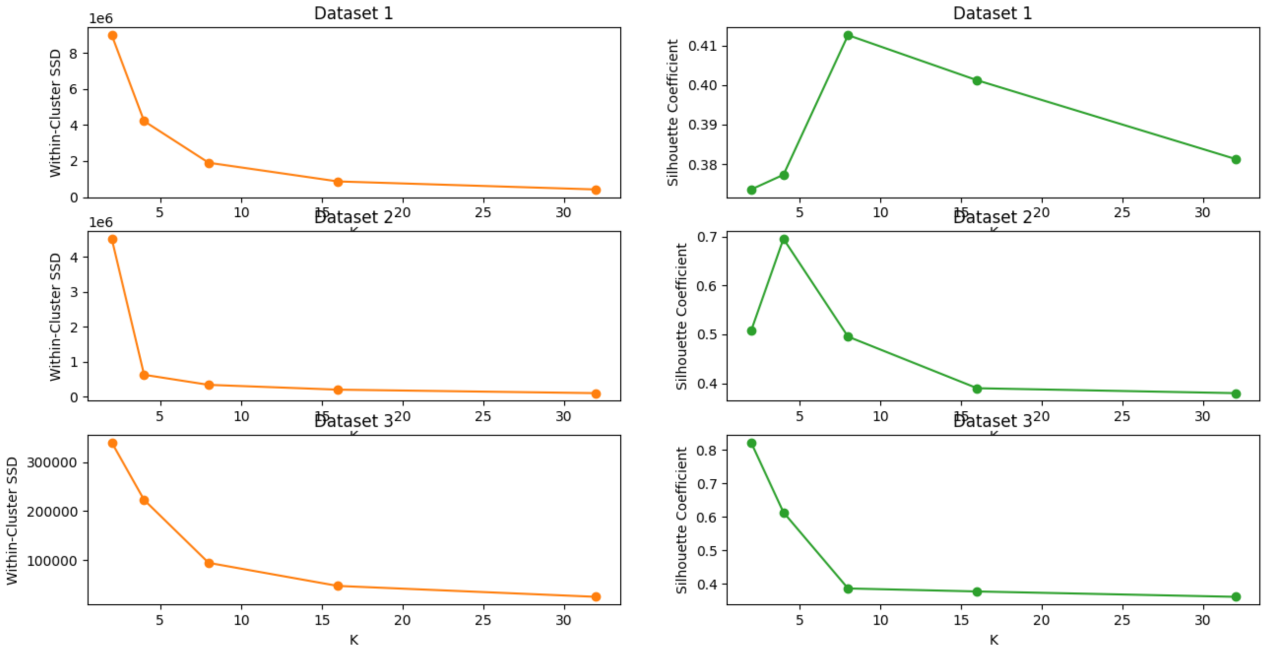


2. Q2

```
1. For running KMeans
   Command to be entered on terminal: python kmeans.py digits-embedding.csv
   Output: WC-SSD: 1489650.532
           SC: 0.40
           NMI: 0.359

2. For analysing KMeans for different values of K
   Command to be entered on terminal: python kmeans_analysis_1.py
   Output: Provides the output plots for WC-SSD and SC for different
   values of K
```

(a) This is the plot obtained for WCSSD and SC for all the 3 datasets as described in the assignment.



(b) After analysing the WC-SSD curves different values of K for dataset 1, we can say that the best value for K is 8 as after that there is not much difference in the value of WC-SSD as the value of K increases more. This value can also be obtained from the SC plot as we can see that we get a peak at the value K=8, so this shows us that K=8 is the best value to cluster the given dataset.

After analysing the WC-SSD curves different values of K for dataset 2, we can say that the best value for K is 4 as after that there is not much difference in the value of WC-SSD as the value of K increases more. This value can also be obtained from the SC plot as we can see that we get a peak at the value K=4, so this shows us that K=4 is the best value to cluster the given dataset.

After analysing the WC-SSD curves different values of K for dataset 3 (also consider the scale of other two graphs), we can say that the best value for K is 2 as after that there is not much difference in the value of WC-SSD as the value of K increases more. This value can also be obtained from the SC plot as we can see that we get a peak at the value K=2, so this shows us that K=2 is the best value to cluster the given dataset.

(c)
```
For analysing KMeans sensitivity to starting points
Command to be entered on terminal: python kmeans_analysis_3.py
```

```
Output: Provides 6 plots corresponding to WC−SSD and SC for
each of the 3 datasets
```

Here are the plots obtained to consider the effect of starting points on the KMeans algorithm. These plots show the standard deviation along with the average values obtained for WC-SSD and SC for eack of the K value for all the 3 datasets. We can see the effect of starting points by analysing the error bar on the plots to see how much deviation is caused by choosing random initial centroid points. Hence, we conclude that KMeans is sensitive to initial points chosen as centroids as the standard deviation is high in the plots.

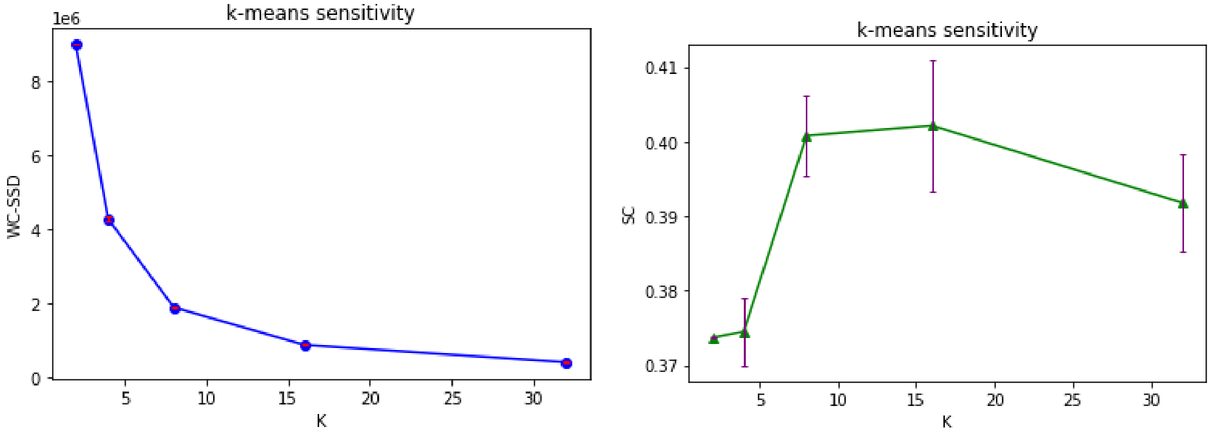


Figure 1: Visualisation for Dataset 1

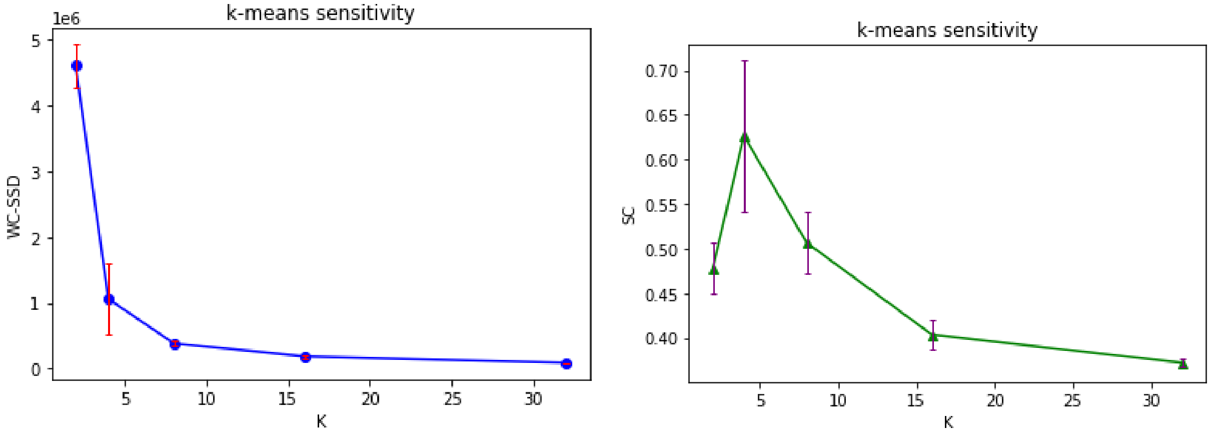


Figure 2: Visualisation for Dataset 2

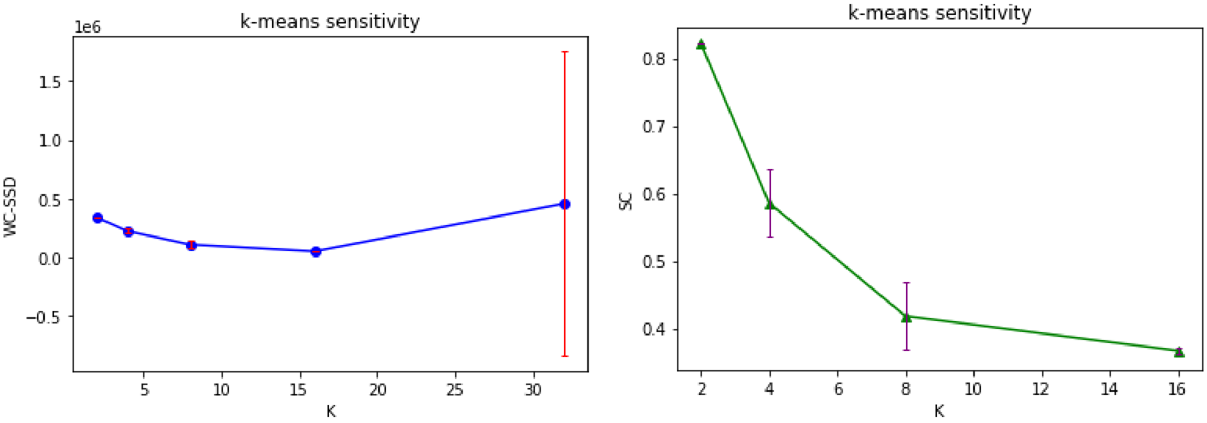


Figure 3: Visualisation for Dataset 3

(d)
```
For analysing NMI values and visualisation results
Command to be entered on terminal: python kmeans_analysis_4.py
Output: Dataset1 NMI: 0.346
        Dataset2 NMI: 0.455
        Dataset3 NMI: 0.491
```

Greater the NMI value, better is the clustering of the dataset. As we can see that we get the highest value of NMI for dataset 3, we should also observe better clusters for that dataset. The graphs shown below help us to conclude the same result.

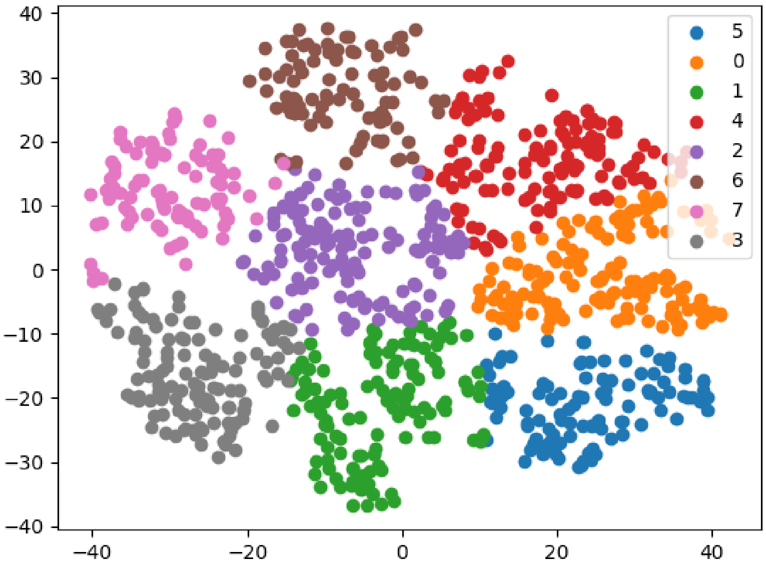


Figure 4: Visualisation for Dataset 1 with K=8

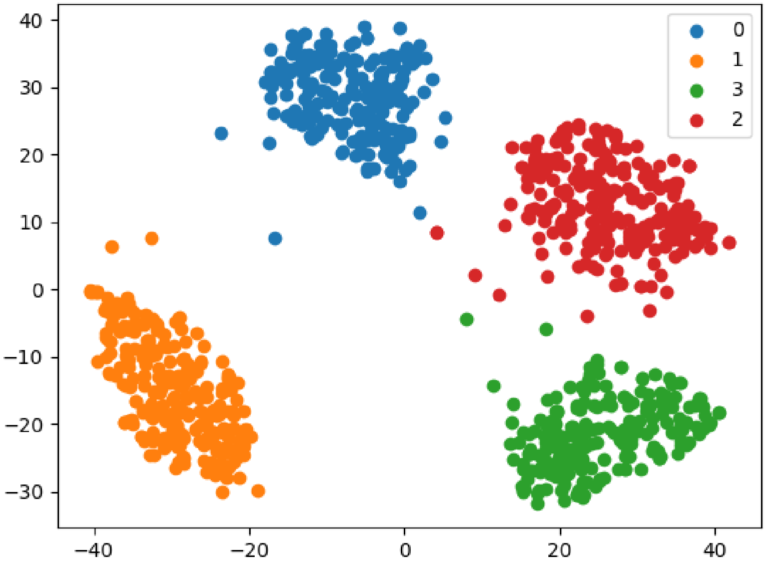


Figure 5: Visualisation for Dataset 2 with K=4

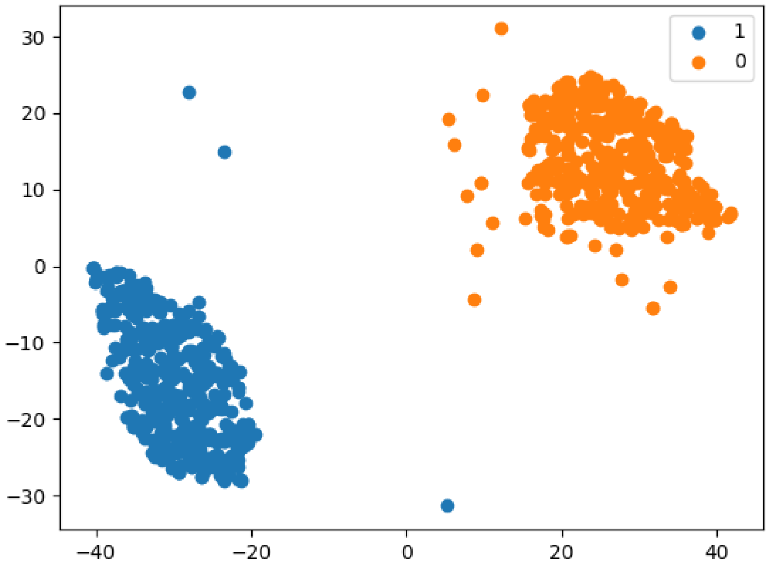


Figure 6: Visualisation for Dataset 3 with K=2

3. Q3

1. Hierarchial Clustering (single linkage)

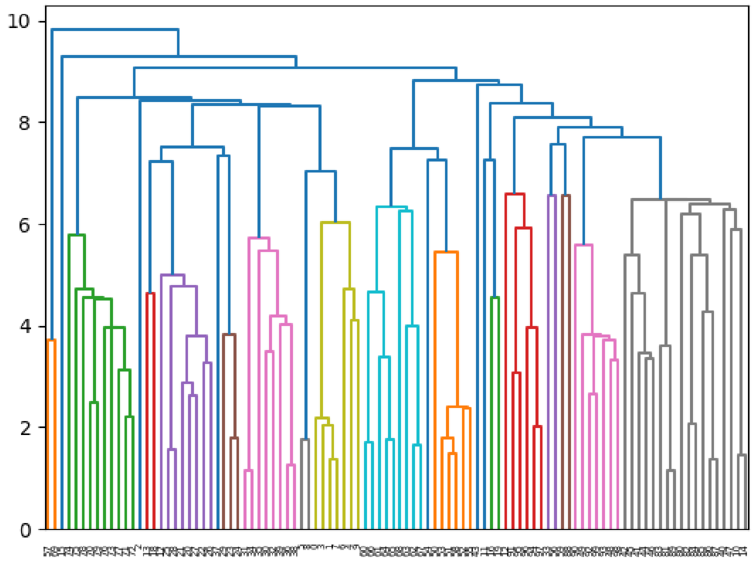


Figure 7: Single linkage hierarchial clustering

2. Hierarchial Clustering (complete linkage)

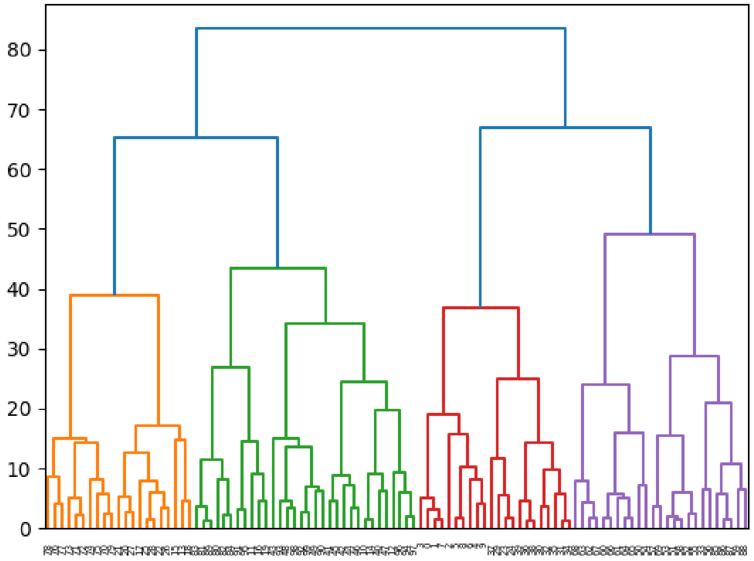


Figure 8: Complete linkage hierarchial clustering

Hierarchial Clustering (average linkage)

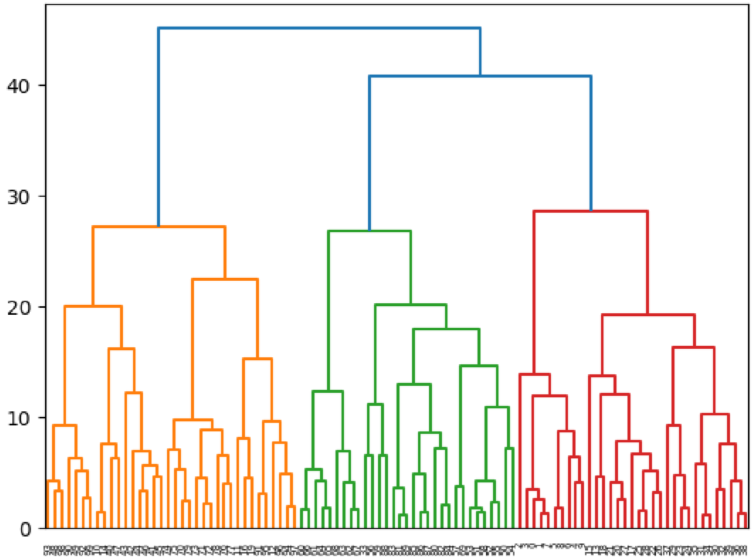


Figure 9: Average linkage hierarchial clustering

3. Here are the respective plots showing the within-cluster sum of squared distances (WC-SSD) and silhouette coefficient (SC) as a function of K:

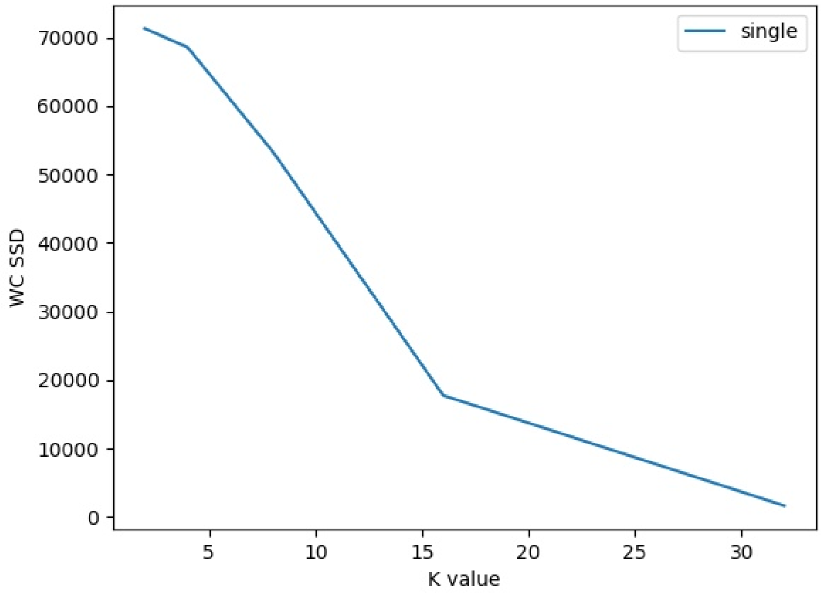


Figure 10: Single linkage WC-SSD

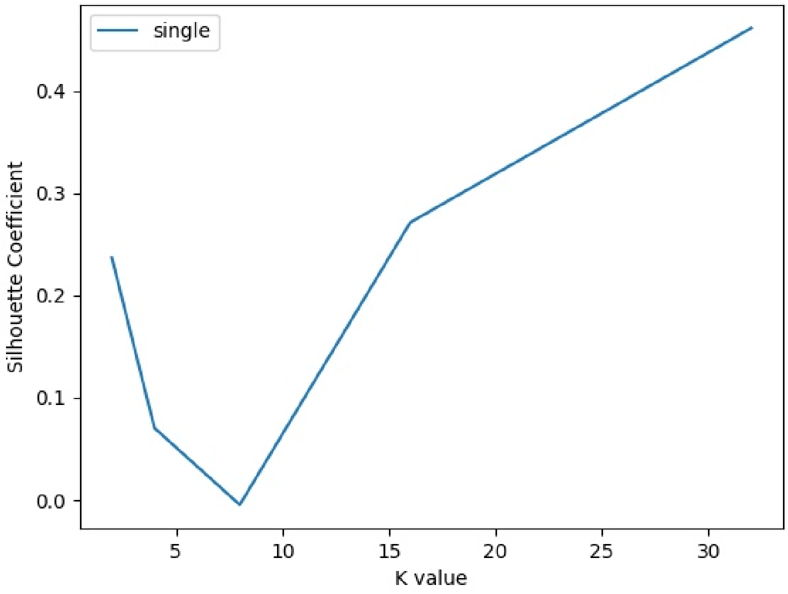


Figure 11: Single linkage SC

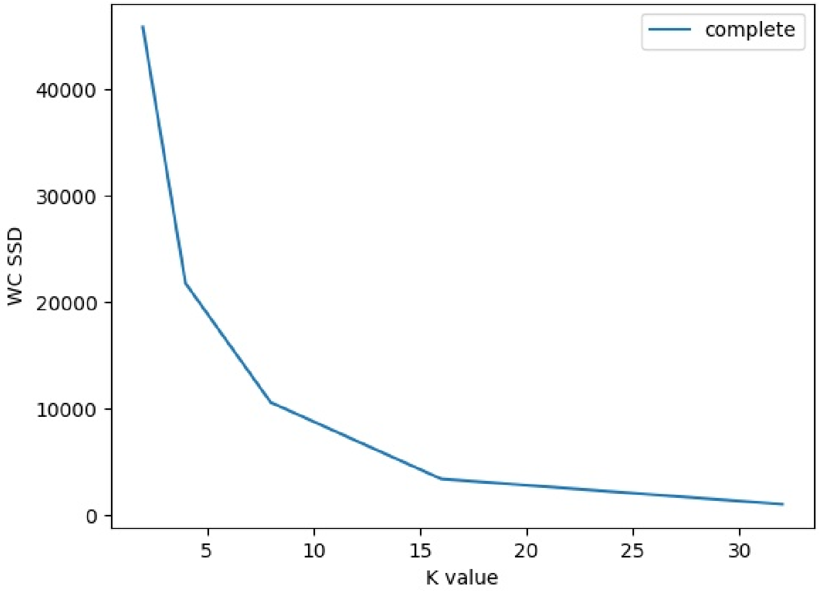


Figure 12: Complete linkage WC-SSD

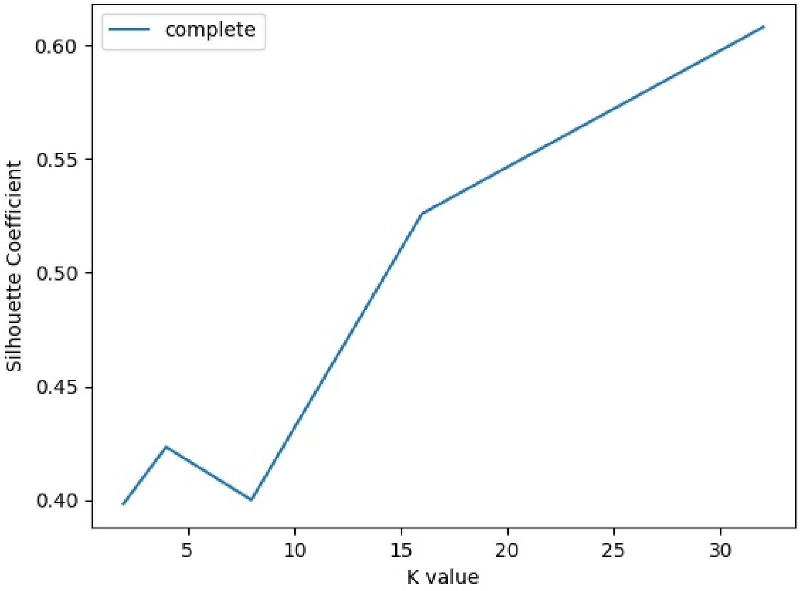


Figure 13: Complete linkage SC

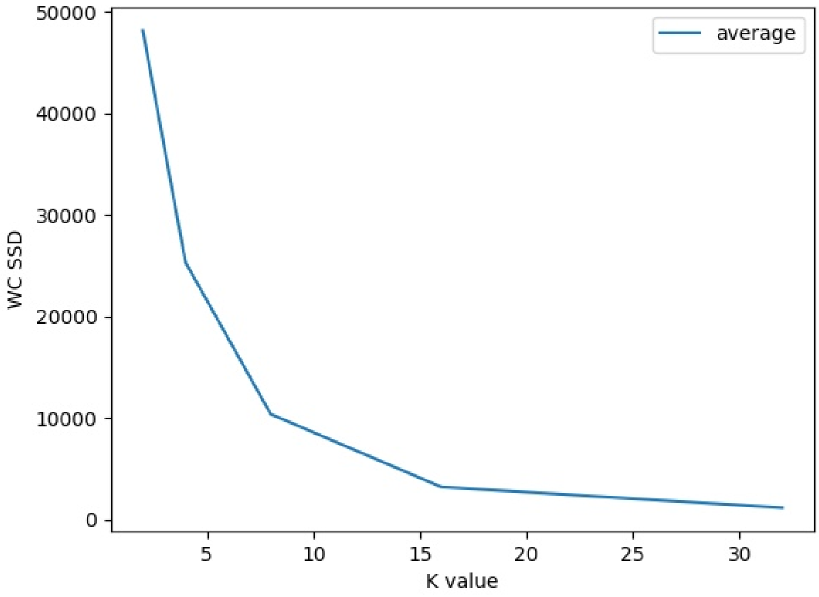


Figure 14: Average linkage WC-SSD

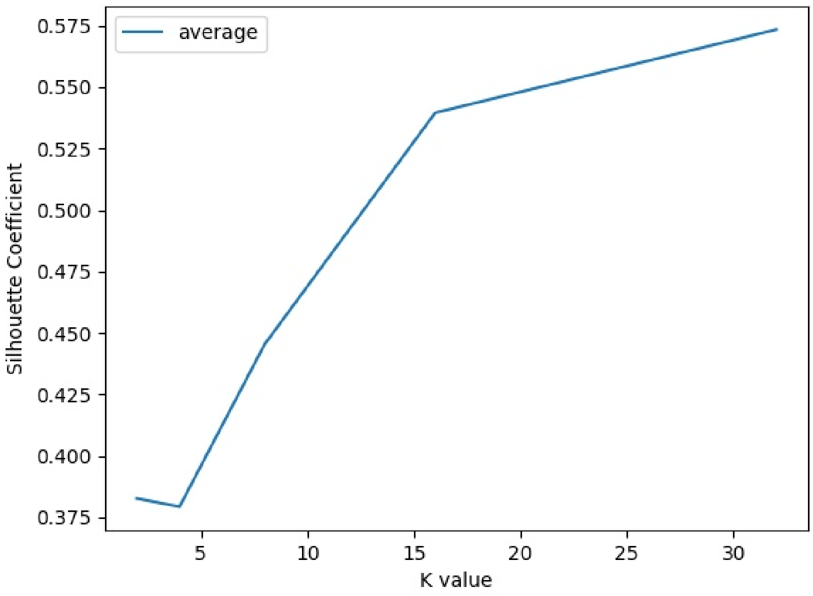


Figure 15: Average linkage SC

4. Upon analysing the above graphs, it seems that choosing a value of K=16 is a better option for all the three linkages shown above as at that point, we can observe that WC-SSD curve has an elbow around that value in each plot.
   This selected value is different from the one which was chosen for the K-Means clustering method. Hence, we can say that we have two different no of optimal clusters for the given dataset based on the two approaches we used.

5.
   ```
   For computing NMI values for each of the dendrogram
   Output: NMI_single: 0.37476587930890864
           NMI_complete: 0.4096988613663228
           NMI_average: 0.404185471150106
   ```

   The NMI score is highest for complete linkage followed by average linkage and then comes the single linkage.
   Compared to KMeans, we have higher value of NMI for the hierarchial clustering method, although the variation is pretty small and might be due to case that we are analysing just 100 examples here instead of the complete dataset.