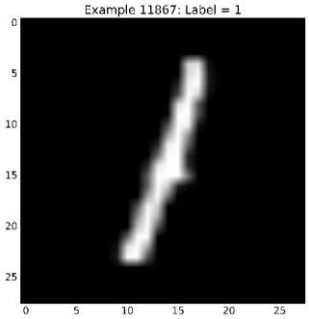
CS 573 Data Mining Homework 5

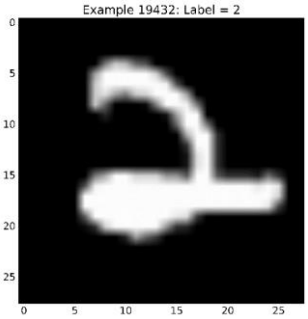
By: Parag Guruji, pguruji@purdue.edu

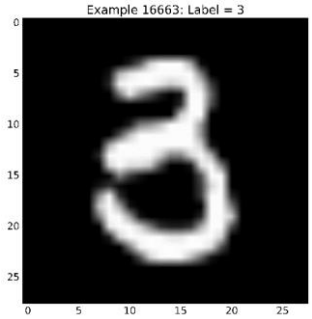
Date: 28 Apr 2017

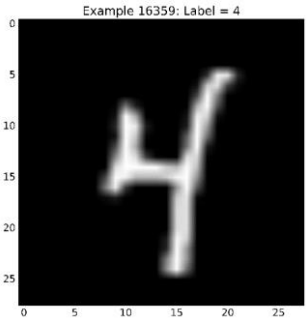
A. Exploration (5 pts)

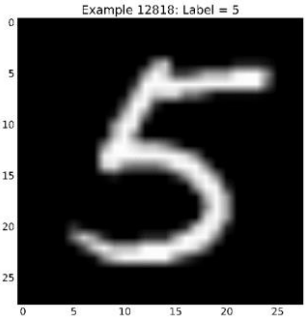
1. Randomly pick one digit from each class in digits-raw.csv and visualize its image as a 28X28 grayscale matrix.

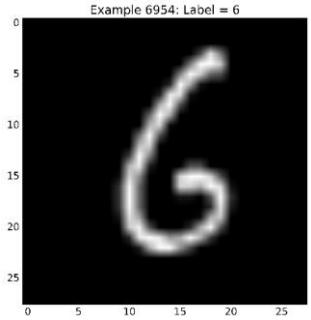


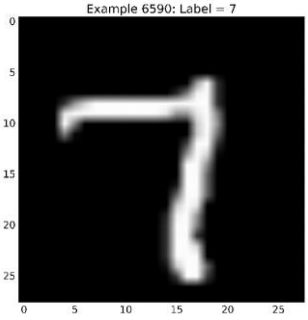


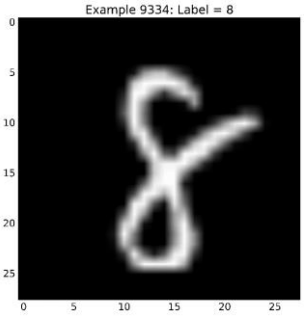


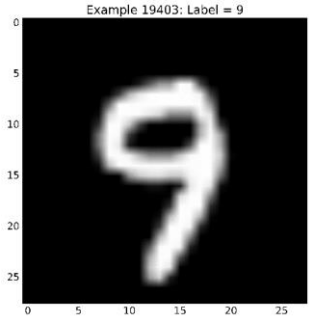


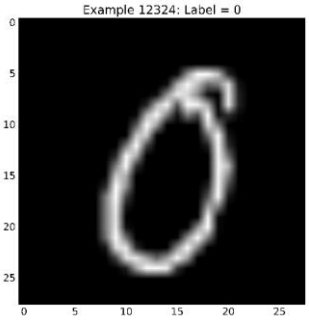




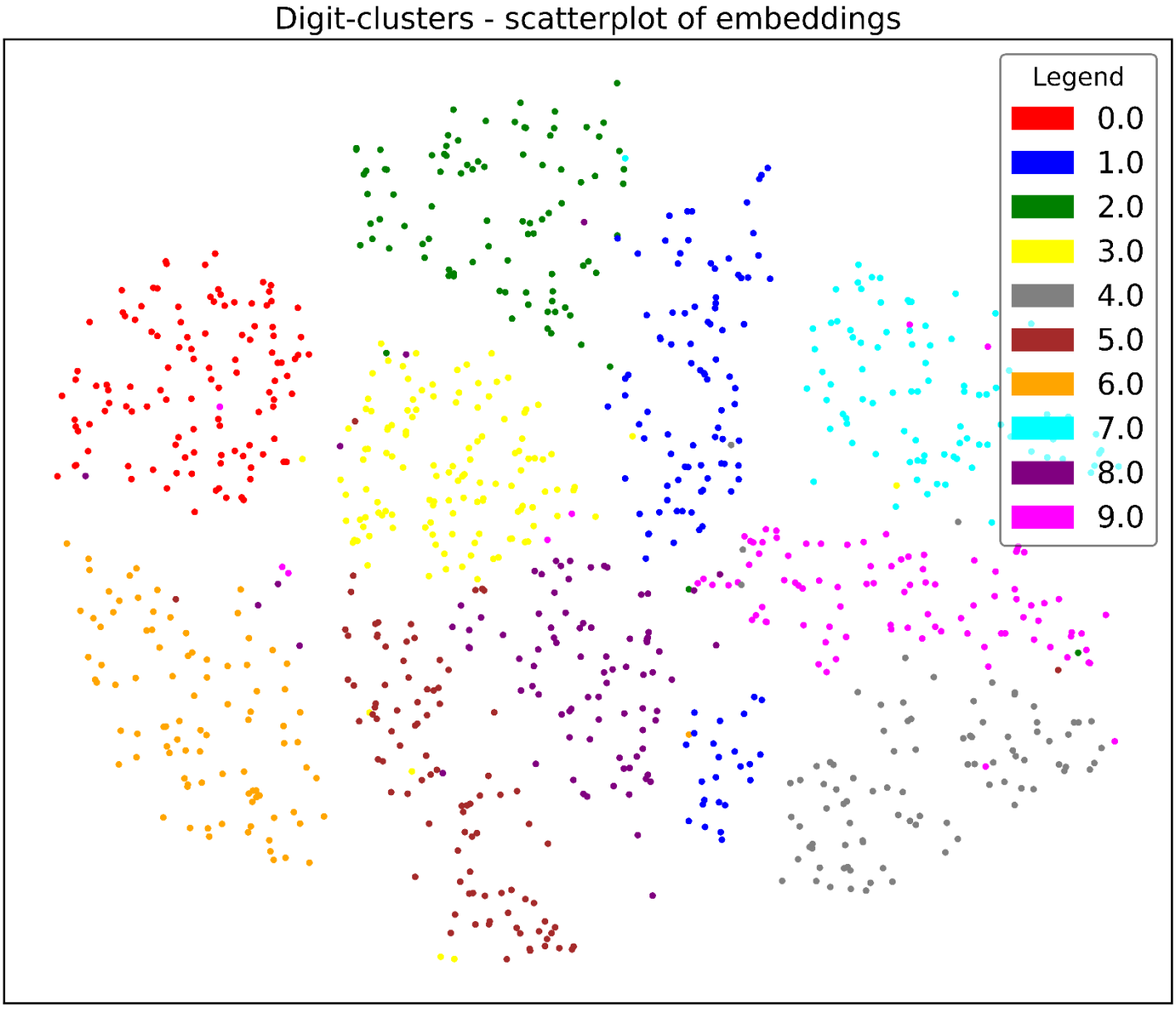








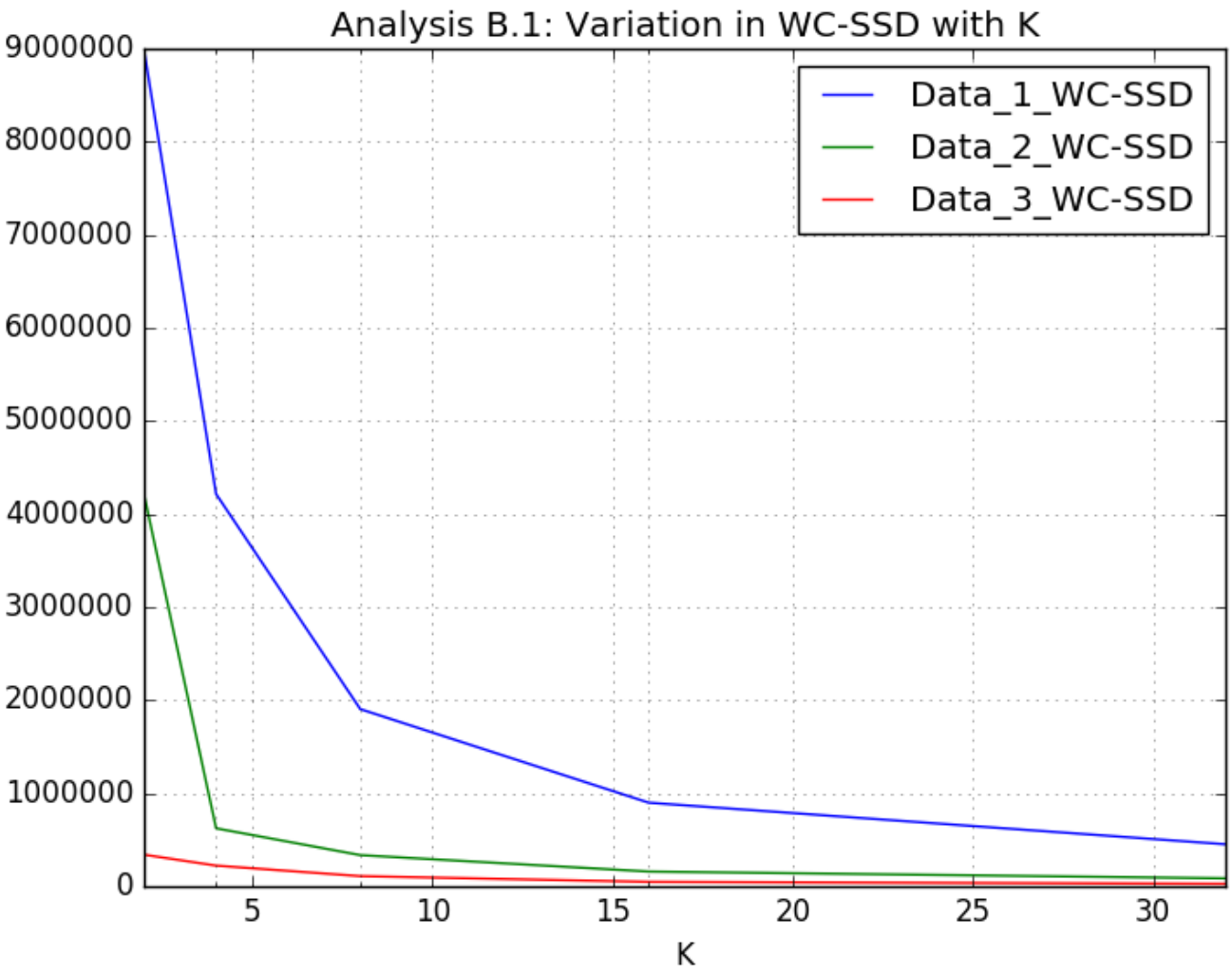
2. Visualize 1000 randomly selected examples in 2d, coloring the points to show their corresponding



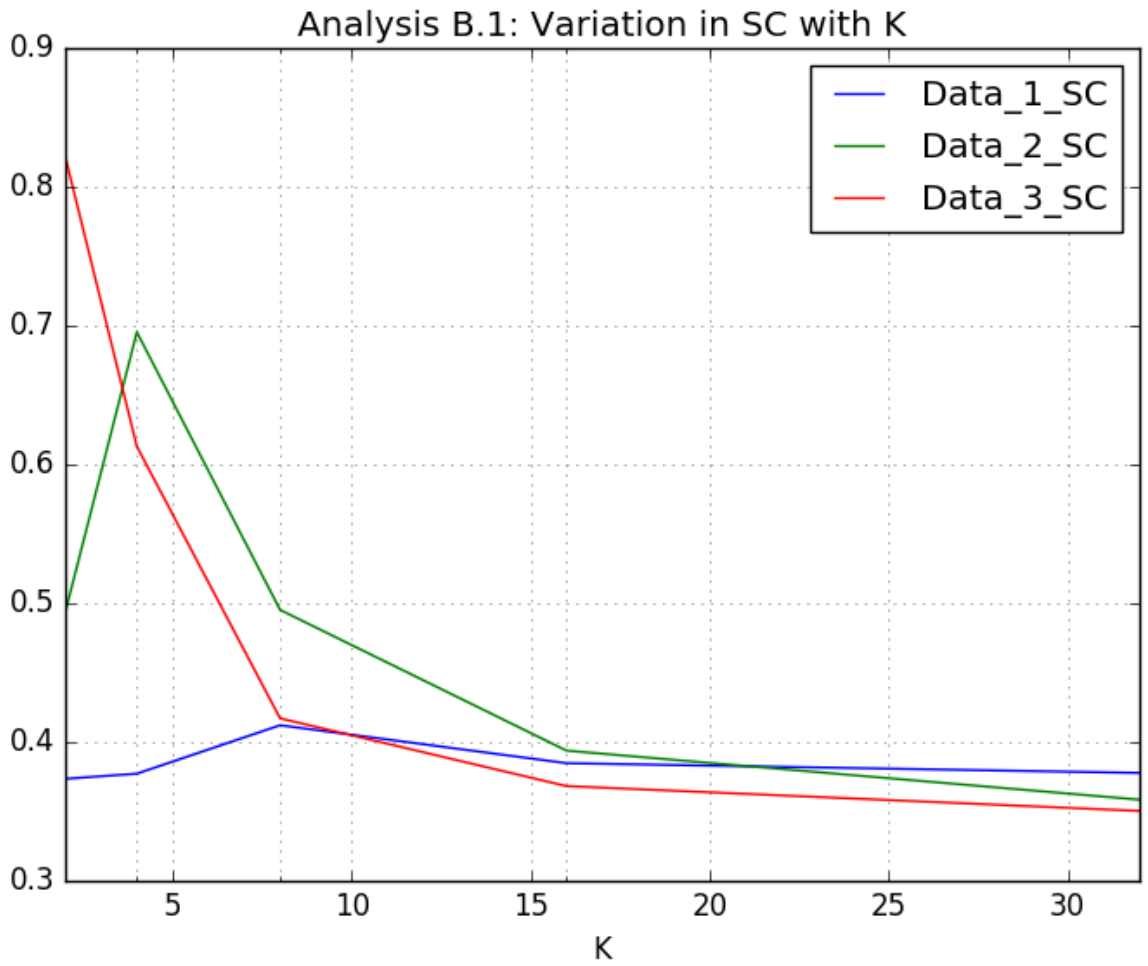
Datasets
Data_1: Full dataset
Data_2: Subset of full dataset where Image labels are 2, 4, 6, 7
Data_3: Subset of full dataset where Image labels are

Note: I am using NMI with average entropy in the denominator (2H)

B. Analysis of k-means
1. Clustering with K = [2; 4; 8; 16; 32] – WC SSD and SC as a function of K.



Plot 1: WC-SSD vs K



Plot 2: SC vs K

2. Choice of K for each dataset:

Argument for Best K:

To choose the appropriate K from **WC-SSD plot, we look at the "Knee"** point in the WC-SSD curve, as it is the point after which, the curve starts flattening – showing only the marginal decrease in WC-SSD for increase in K and thus causing too many clusters.

For finding best K using **SC plot, we look for the "peak"** point since as the SC approaches 1, the quality of cluster improves. On the left side of peak are the K values which represent too few clusters whereas on the right side of the peak are the K values that show too many clusters.

Based on this reasoning, following are the "best" K values across 3 data sets and 2 scores:

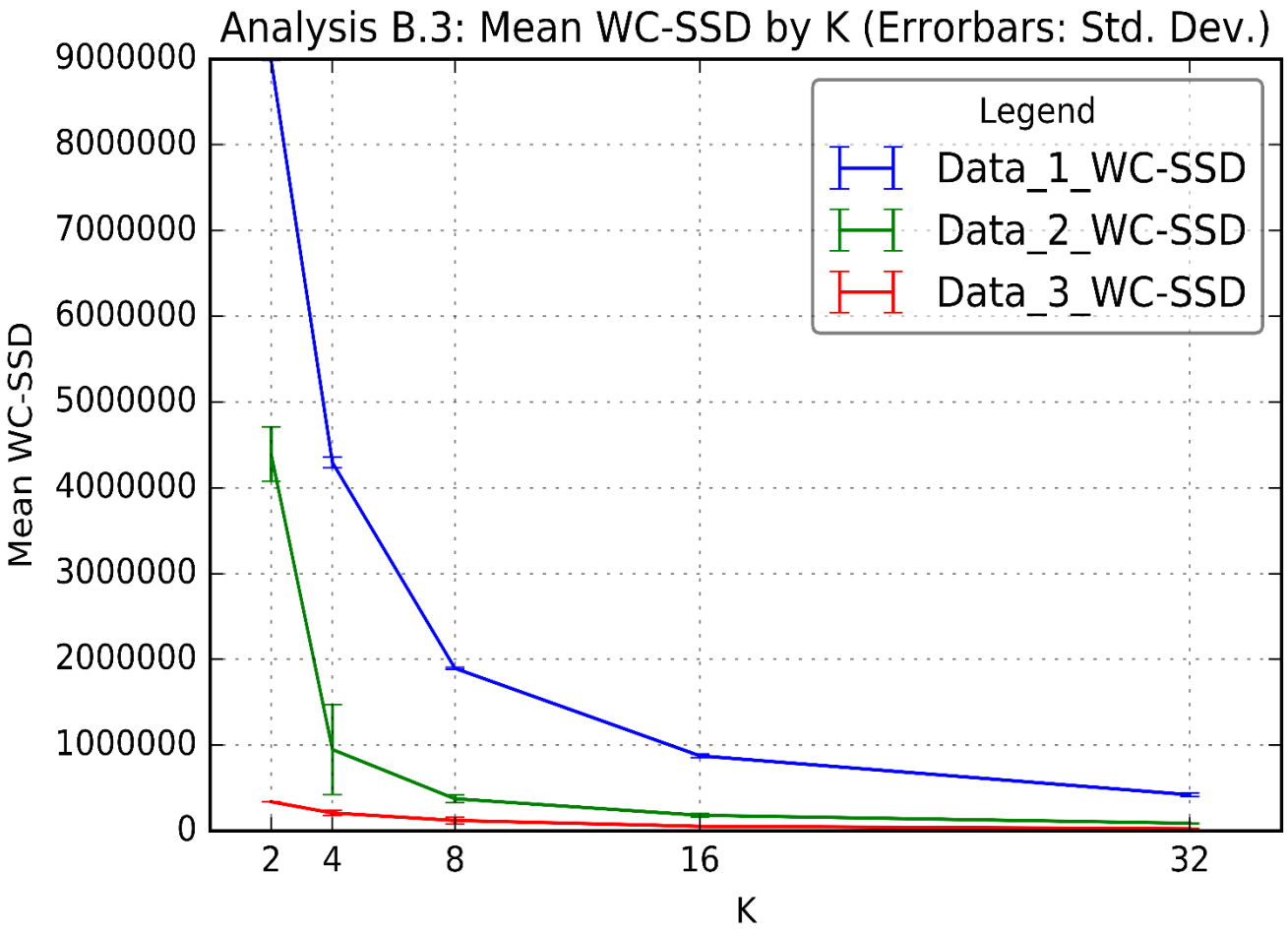
Data	Best K (WC-SSD)	Best K (SC)
Data_1	8	8
Data_2	4	4
Data_3	2	2

Table 1

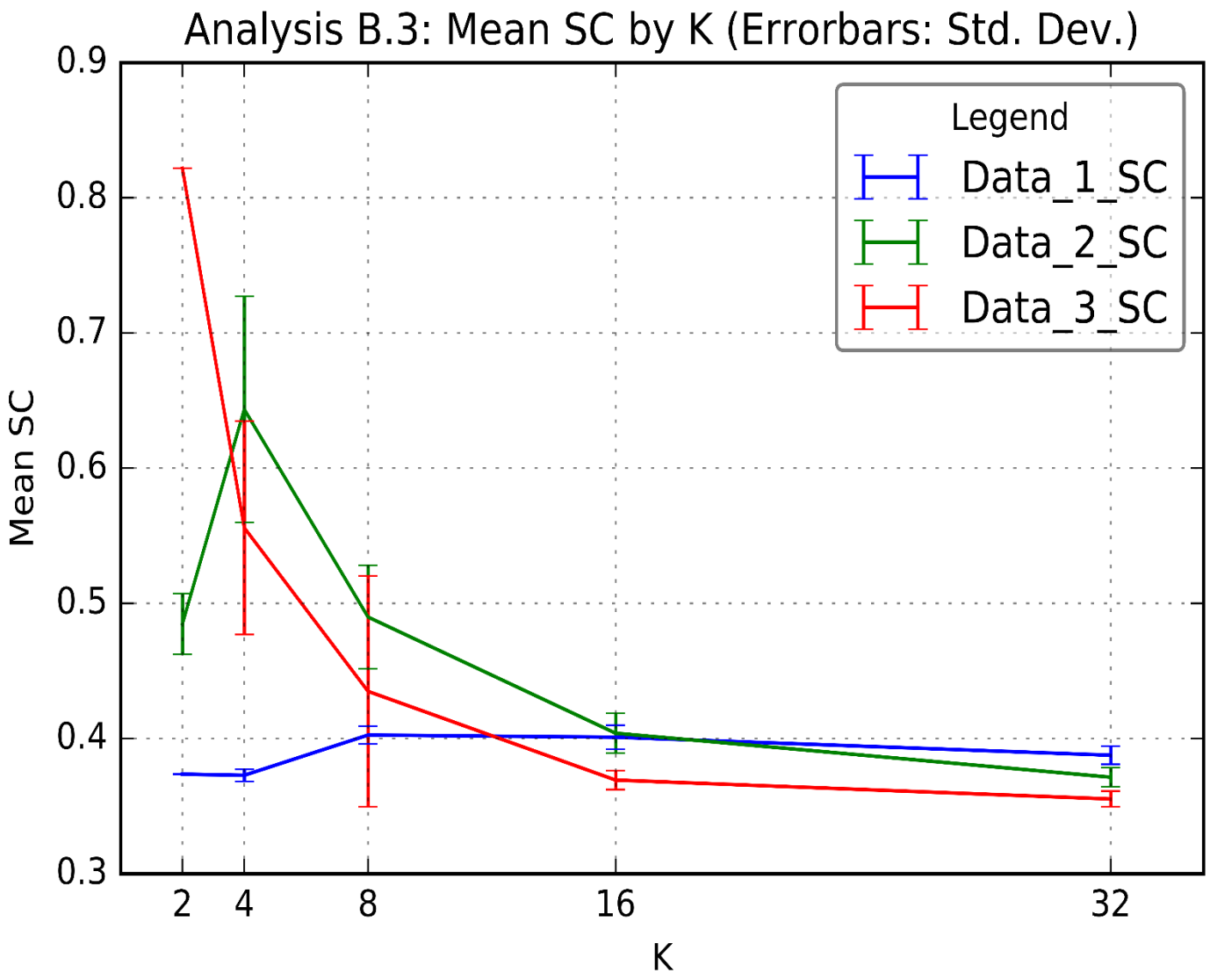
We observe that both the scores agree on suggested best K value for all three datasets. The suggested number of clusters i.e. K varies across the three datasets in alignment with the original number of image labels in the given datasets, which is an indication of good clustering.

3. K means sensitivity to initial starting conditions:

The initial k cluster-centroids (seeds) are randomly generated for the 10 runs at each K-value 2, 4, 8, 16, 32. The resultant plots are as follows:



Plot 4: Mean WC-SSD vs K



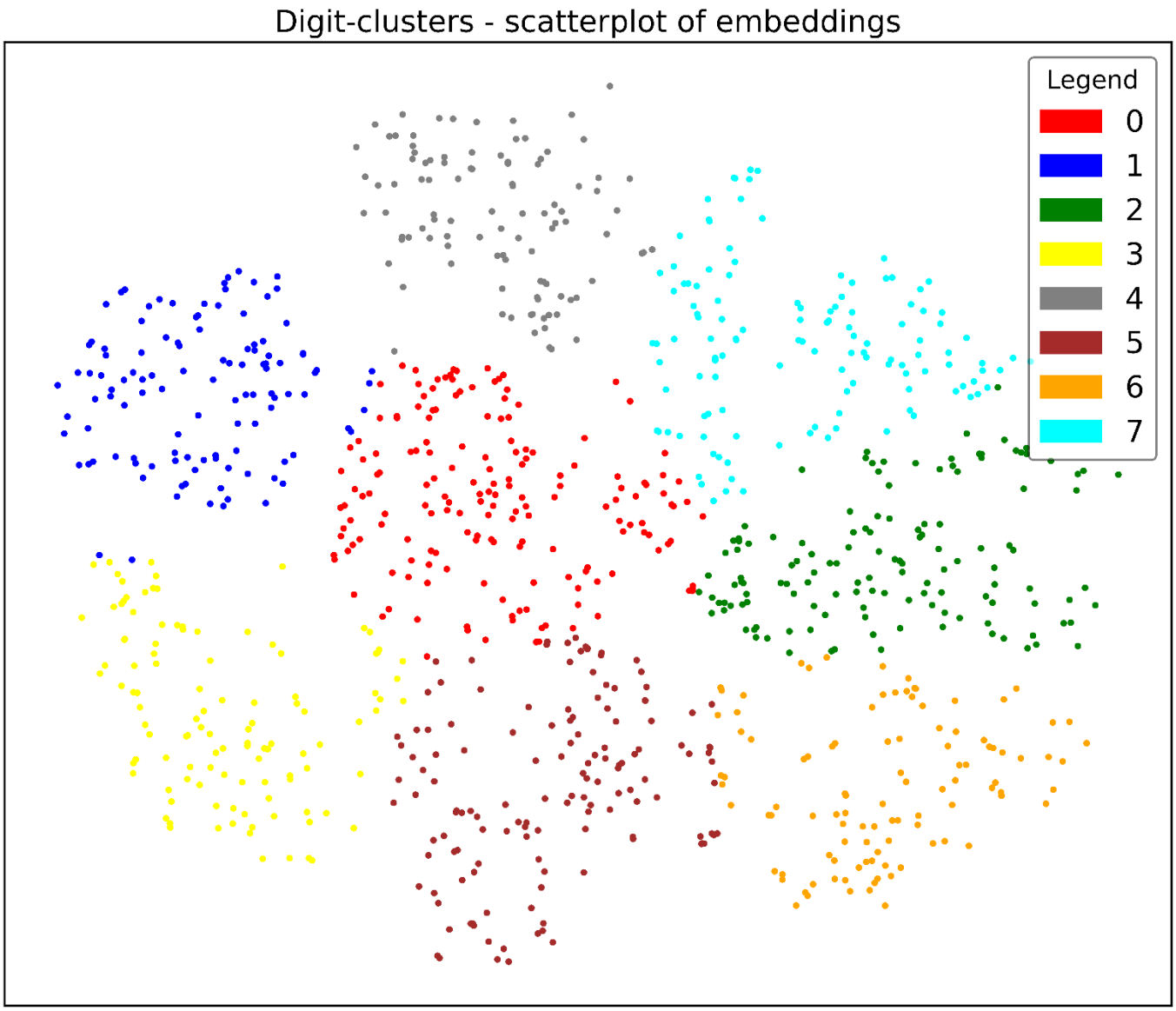
Plot 4: Mean SC vs K

Discussion:

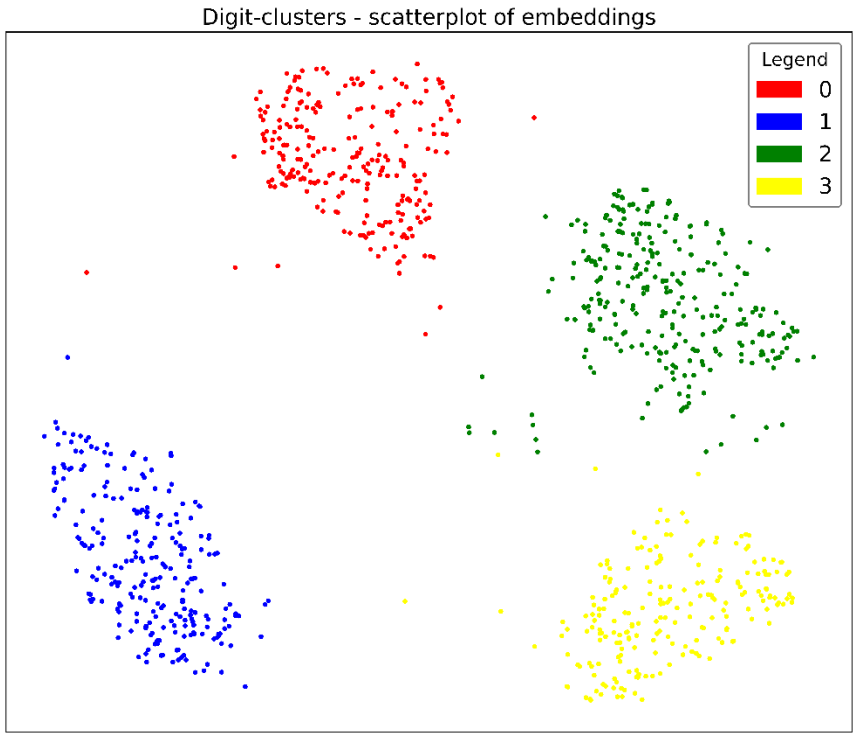
1. WC-SSD is relatively less affected by the initial randomness of seeds than SC since it doesn't have to deal with the randomness of what would come as the 'nearest cluster' as it depends only on points within the cluster, which SC must deal with.
2. The standard deviation appears negligible for the Full dataset – Data_1 for both the scores. An obvious primary reason behind this is its greater sample size.
3. In SC plot, the extreme high K values (16 and 32) show little variance – indicating the more certainty with which the Silhouette Coefficient is identifying them as obvious inappropriate K values with its penalty component.

4. Visualization and NMI of clusters for Chosen Ks:

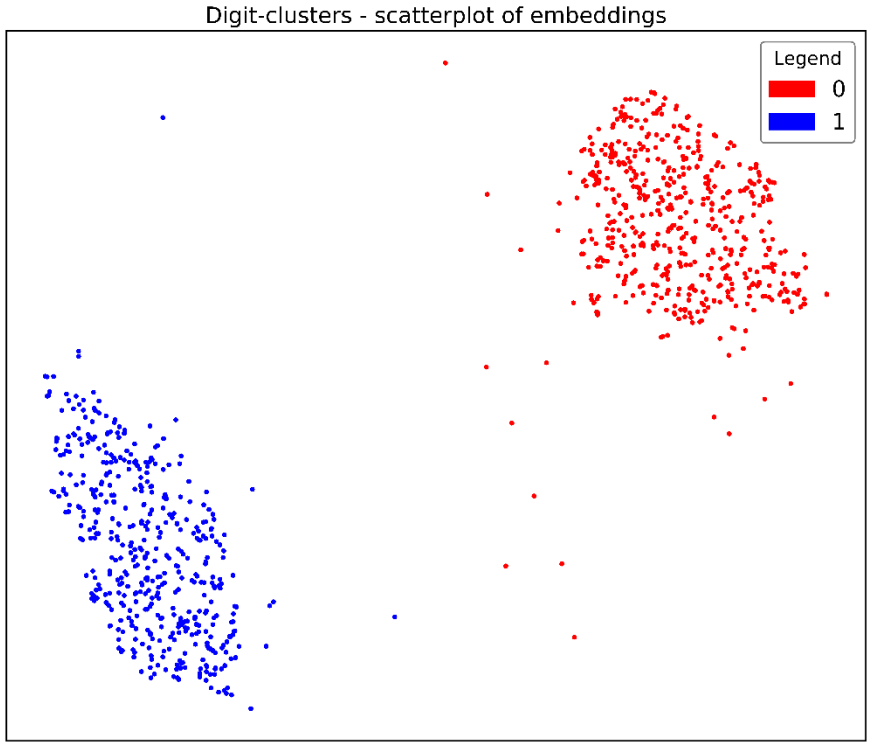
We use the K values as chosen in the Table 1 above. Following are the Scatterplot visualizations for 1000 randomly sampled examples:



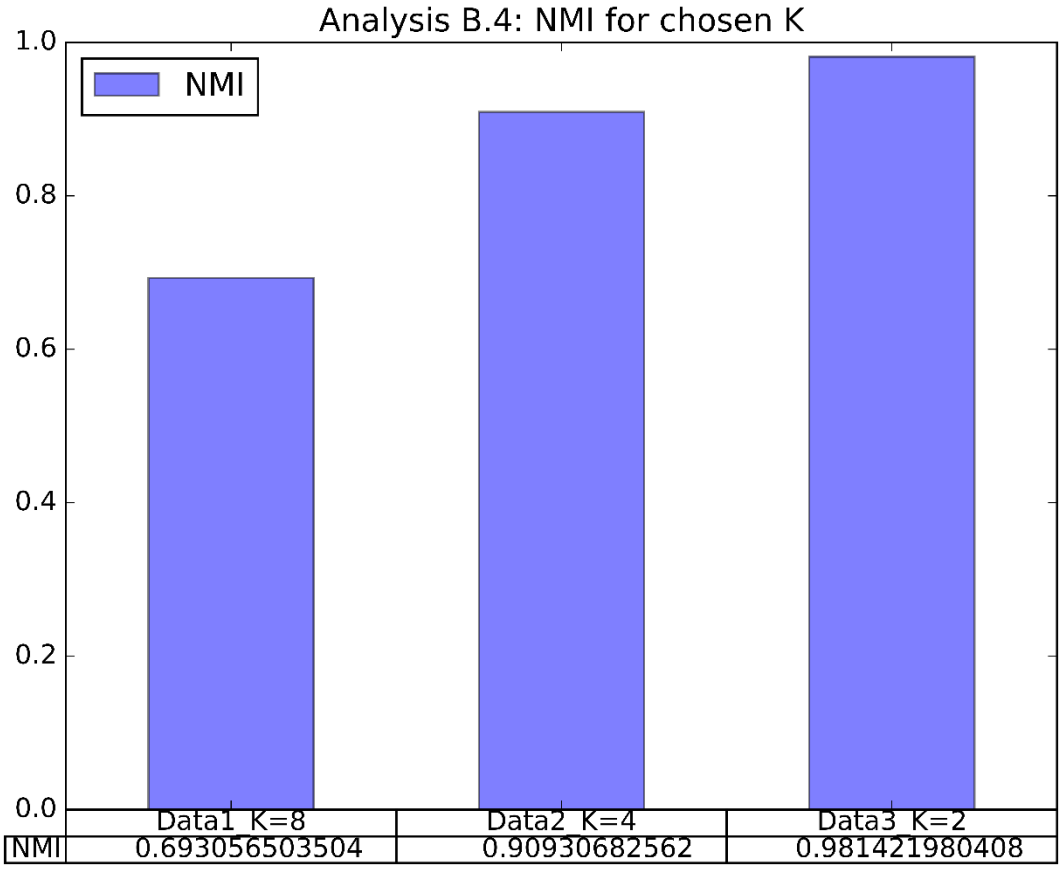
Plot 5: Scatterplot: Data= Data_1, K=8, Sample Size = 1000



Plot 6: Scatterplot: Data= Data_2, K=4, Sample Size = 1000



Plot 7: Scatterplot: Data= Data_3, K=2, Sample Size = 1000



	Data1_K=8	Data2_K=4	Data3_K=2
NMI	0.693056503504	0.90930682562	0.981421980408

Plot 8: NMI across Datasets for chosen K

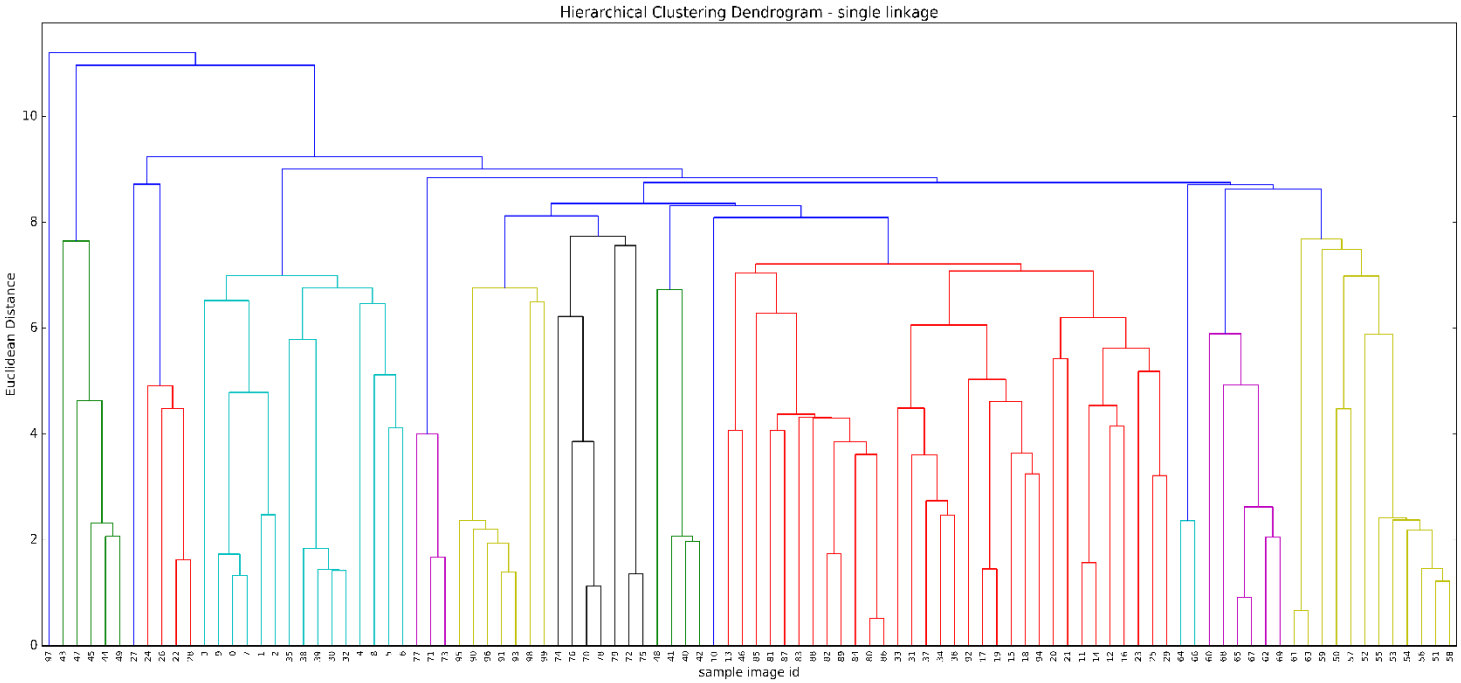
Discussion:

1. All three visualizations show that their respective dataset is visibly clearly well-clustered with the chosen K value.
2. The separation in between different clusters is minimum in Data_1 (full dataset) and is maximum in Data_3 (6, 7) with Data_2 (2, 4, 6, 7) in the middle of these two.
3. The same thing is observed in the NMI bar graph shown above. Higher the NMI, greater the inter-cluster separation, and better is the quality of clustering.
4. Here, we compared the separation of distinct clusters, however we cannot directly compare the intra-cluster closeness from visual observation since the cluster density is implicitly varied across the datasets due to same sample size but different number of clusters. E.g.: Data_3 clusters are bound to be more dense as they get higher number of examples per cluster

C. Comparison with hierarchical clustering

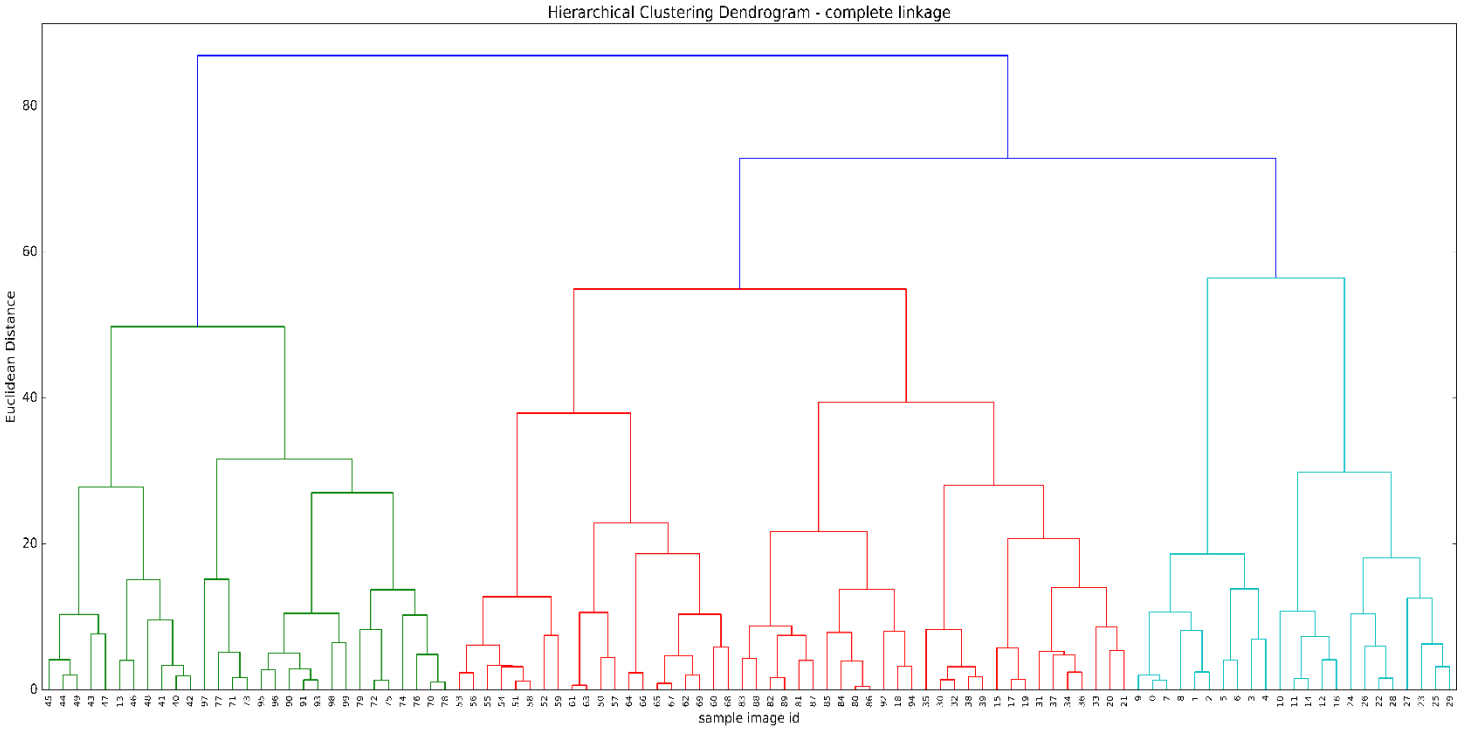
Scipy agglomerative clustering:

1. Single Linkage Dendrogram



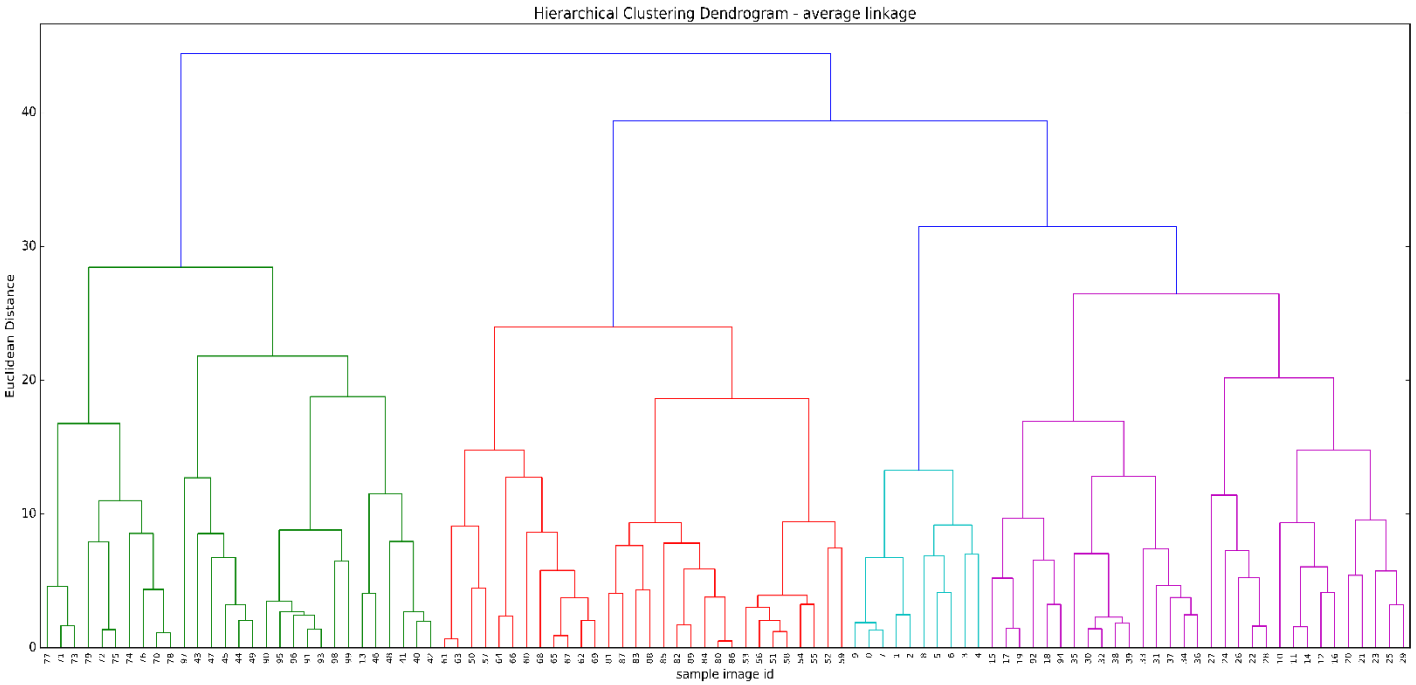
Plot 9: Dendrogram – Single Linkage (Stratified Sample: size = 10 X 10 = 100)

2. (i) Complete Linkage Dendrogram



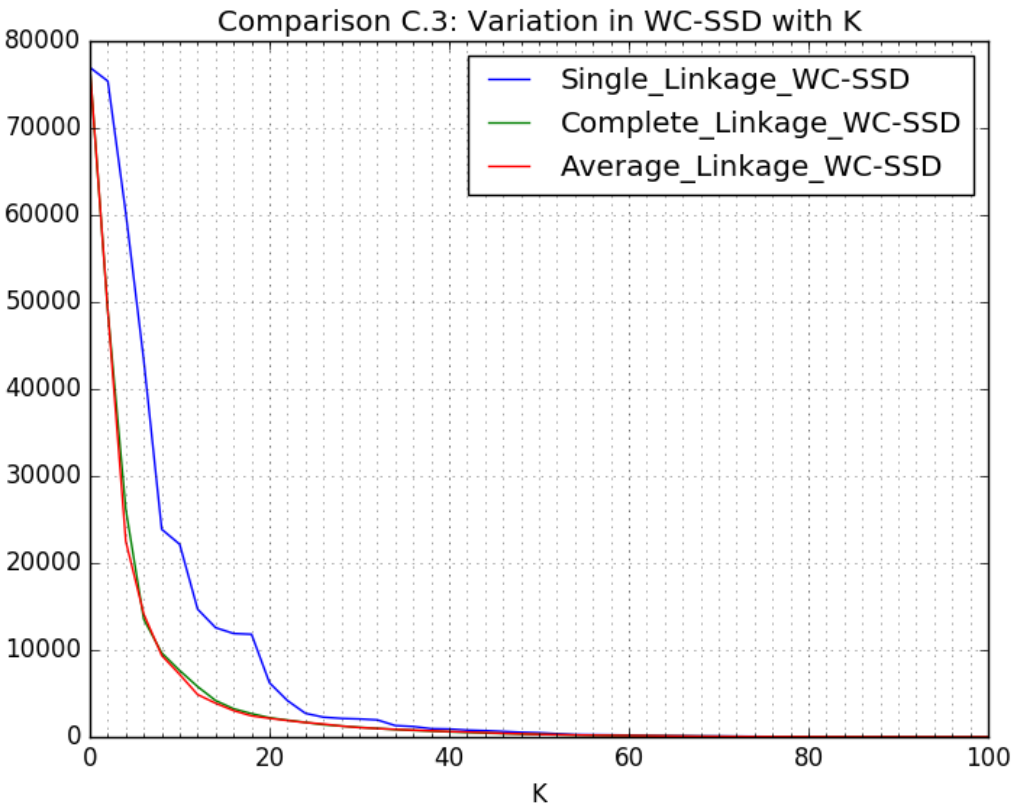
Plot 10: Dendrogram – Complete Linkage (Stratified Sample: size = 10 X 10 = 100)

2. (ii) Average Linkage Dendrogram

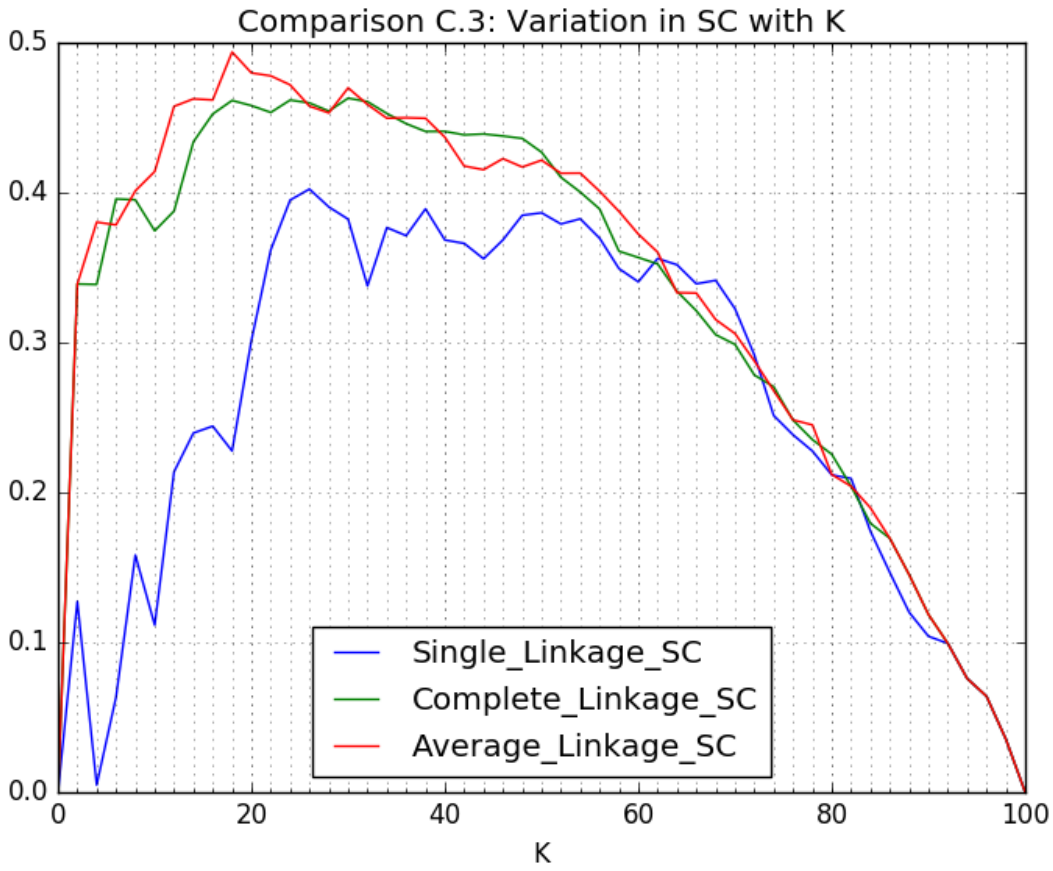


Plot 11: Dendrogram – Average Linkage (Stratified Sample: size = 10 X 10 = 100)

3. Variation of WC-SSD and SC with K – Cutting the Dendrograms:



Plot 12: WC-SSD vs K for 3 Dendrograms (Sample Size = 100)



Plot 13: SC vs K for 3 Dendrograms (Sample Size = 100)

4. Choice of K for each Linkage:

Argument for Best K:

To choose the appropriate K from **WC-SSD plot, we look at the "Knee"** point in the WC-SSD curve, as it is the point after which, the curve starts flattening – showing only the marginal decrease in WC-SSD for increase in K and thus causing too many clusters.

For finding best K using **SC plot, we look for the "peak"** point since as the SC approaches 1, the quality of cluster improves. On the left side of peak are the K values which represent too few clusters whereas on the right side of the peak are the K values that show too many clusters.

If SC and WC-SSD suggest different 'best' values for K, we choose the best K by following rationale:

1. **Give Preference to SC**: Since WC-SSD doesn't penalize over-clustering, if there is one clear best K as suggested by SC, beating its competition by far, we pick that K even if it is a little leftwards of the knee point in WC-SSD

2. **If there are close K values in SC**, look at the corresponding WC-SSD values and position w.r.t. the WC-SSD "Knee" point. If there is steep decrease in WC-SSD for higher K value (among the close competitors from SC-plot), then select the higher K. If the slope of WC-SSD in the interval of competitor K values (from SC-plot) is very less (flat), then choose the lower K.

Based on this reasoning, following are the "best" K values across 3 Linkages are:

Linkage	Best K
Single	26
Complete	18
Average	18

Table 2

Comparison with K chosen in Part B:
Note: Since we are using only the full dataset (all digit labels) for Part C, we can compare it with only Data_1 (full dataset) from Part B

For the plot of Data_1 generated in Part B the best K values observed was K = 8

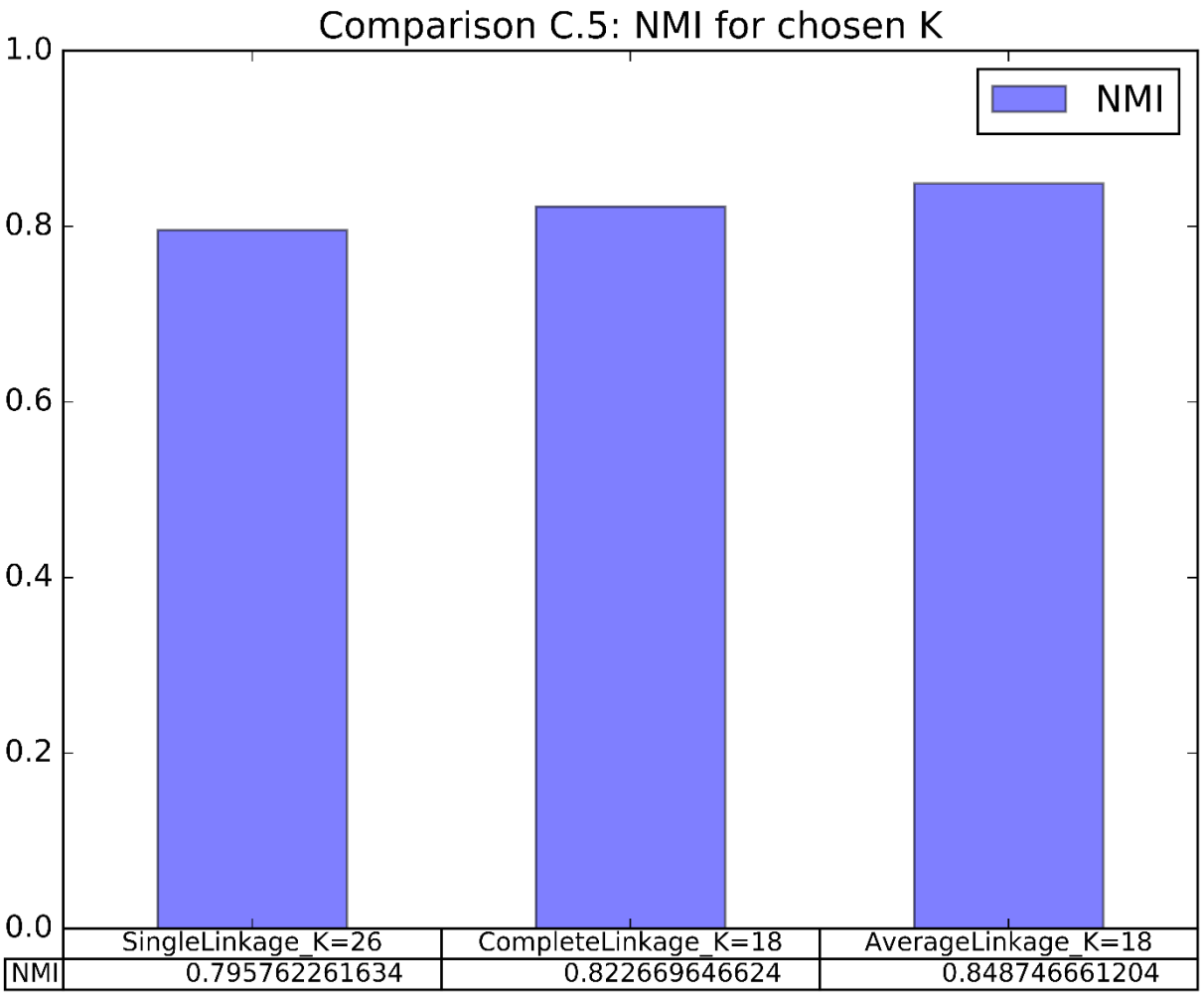
For part C however, the best K values observed were as follows:
K = 26 (Single Linkage)
K = 18 (Complete and Average Linkages)

1. Here, we can clearly see the clear and significant difference in the Best K value across K means and Hierarchical clustering.
2. Another observation is that at the K=8 i.e. the K from part B, in the Single Linkage Plots for SC and WC-SSD the exist local maxima and minima resp.
3. One of the primary potential reason of high variation in SC plots of the Hierarchical clustering is the very low sample size (100) as compared to that in Part B (sample size = 20000)

5. Comparison of NMI:
For Part B as the cluster size increases the NMI value tends to decrease.
Observed NMI values across clusters were as follows:
K = 8, NMI = 0.6930
K = 4, NMI = 0.9093
K = 2, NMI = 0.9814

Since we are using only the full dataset (all digit labels) for Part C, we can compare it with only Data_1 (full dataset) from Part B

NMI for chosen K for Different linkages:

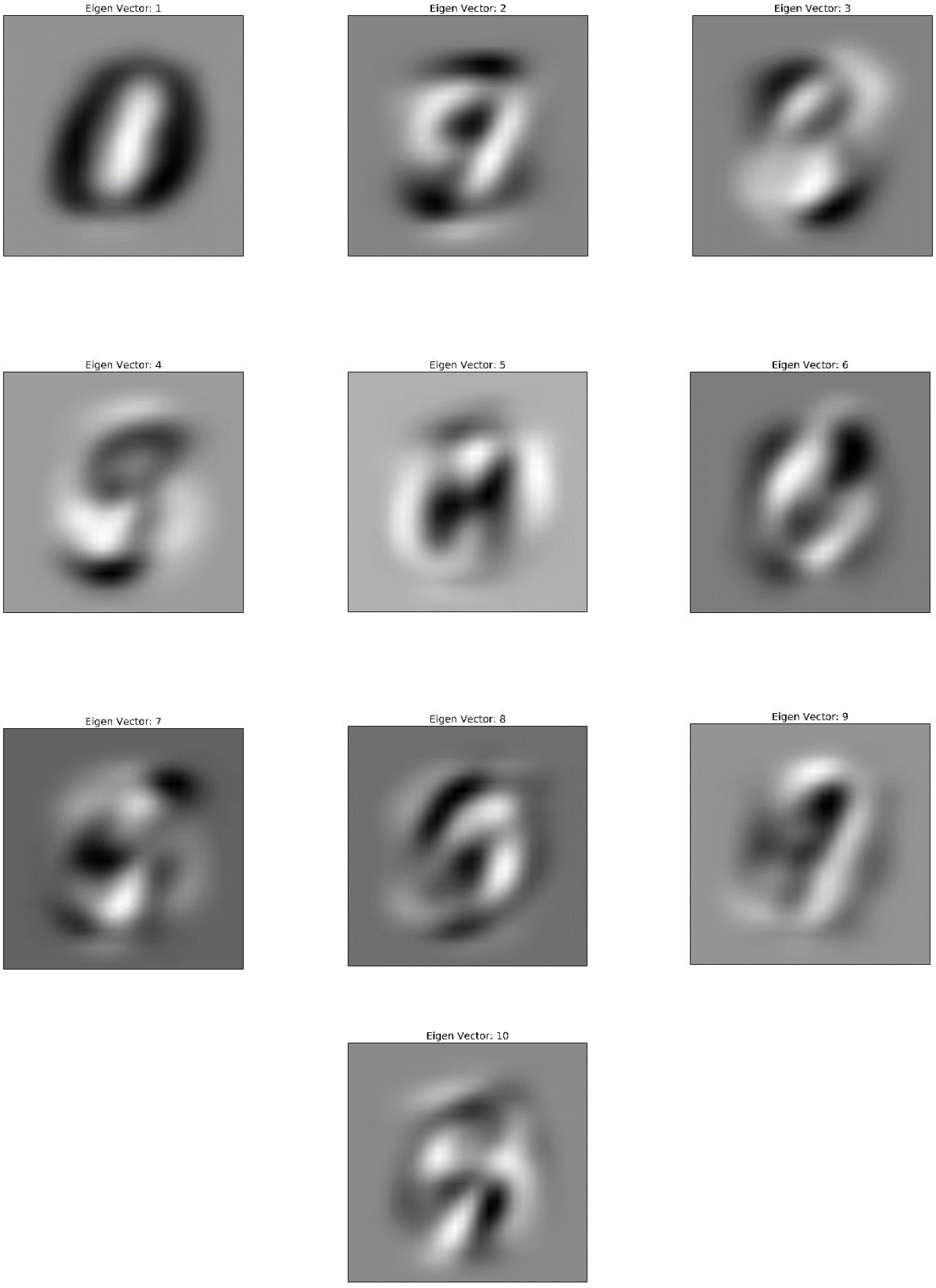


Plot 14: NMI for chosen K for Different Linkages

For Part C we observe that across distance measures i.e. single linkage, complete linkage and average linkage the NMI value is mostly stable and consistent.
Observed NMI values across clusters were as follows:
Single Linkage: K = 26, NMI = 0.7957
Complete Linkage: K = 18, NMI = 0.8226
Average Linkage: K = 18, NMI = 0.8487

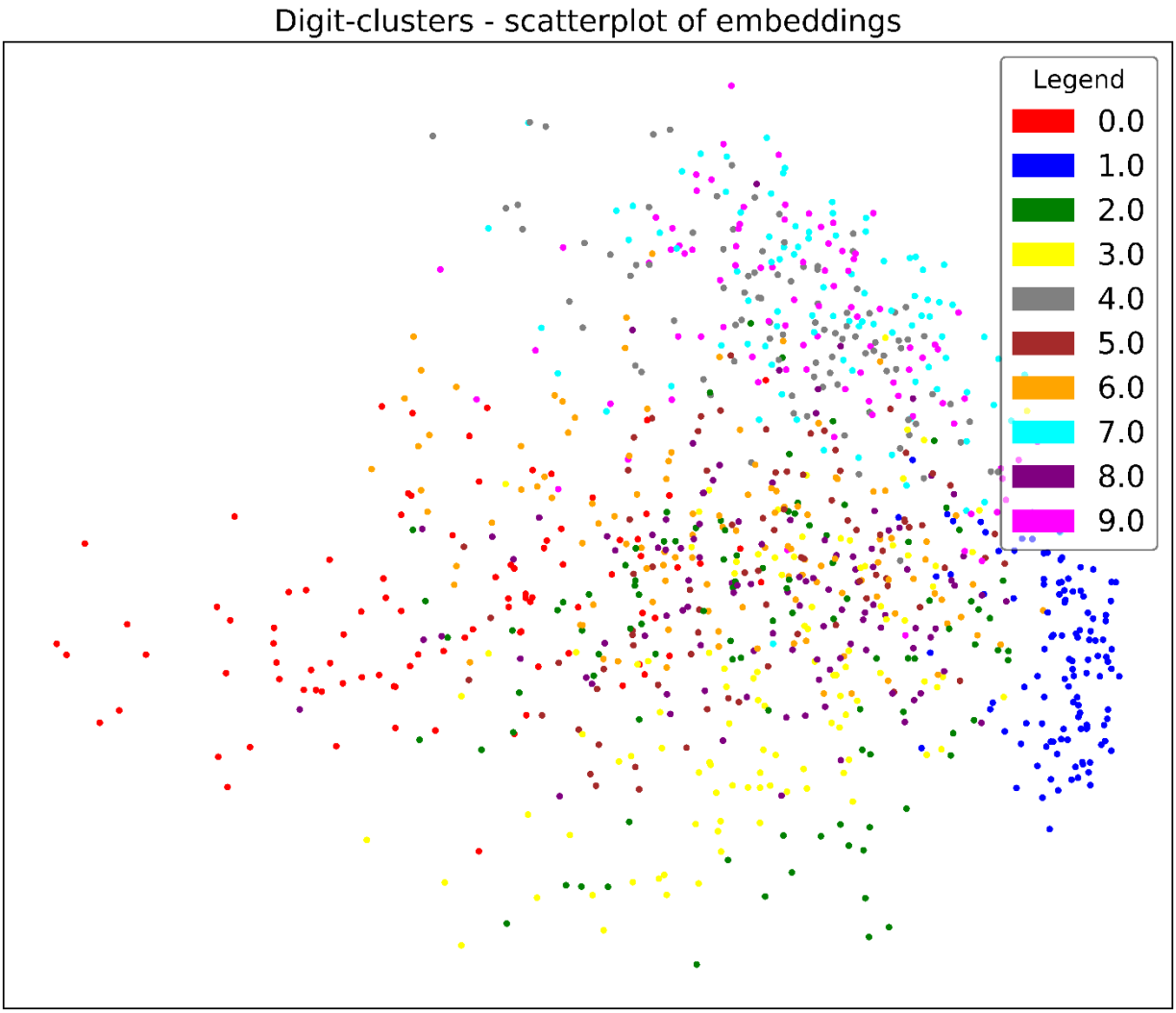
Bonus – PCA:

2. Eigenvectors Grayscale Plots:

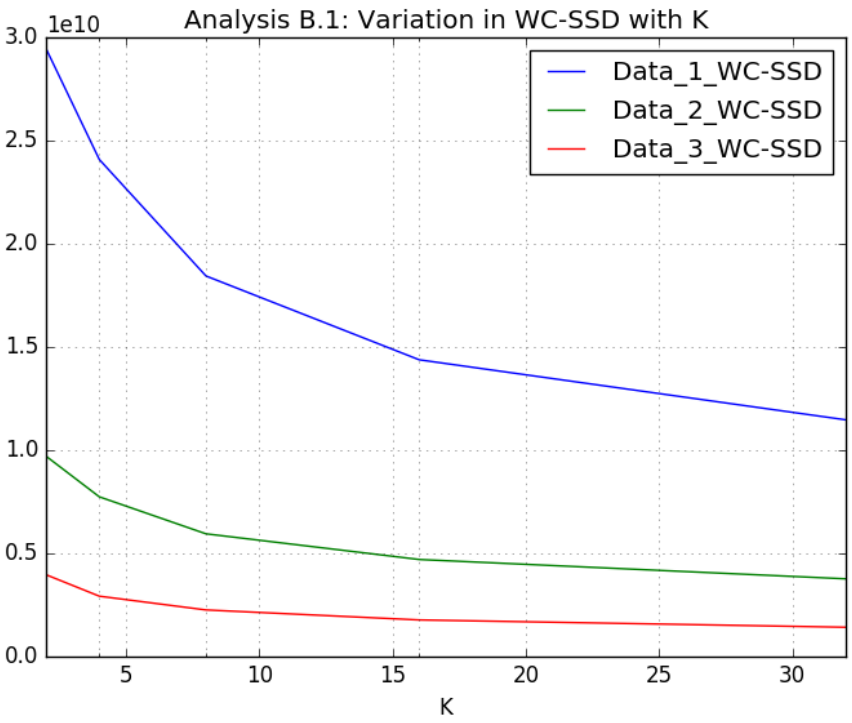


Plot 15: Top 10 Eigenvectors as grayscale matrices

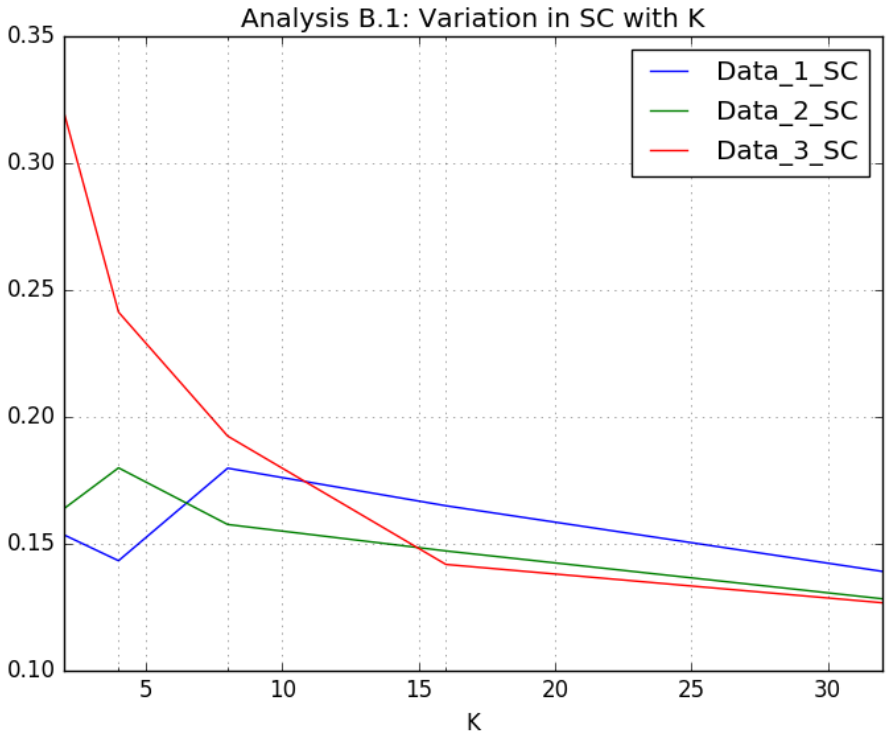
3: Visualization Scatterplot:



Plot 16: Scatterplot of First to Principle Components



Plot 17: WC-SSD vs K with PCA



Plot 18: SC vs K

B2. With PCA: Choice of K for each dataset:

Argument for Best K:

To choose the appropriate K from **WC-SSD plot, we look at the "Knee"** point in the WC-SSD curve, as it is the point after which, the curve starts flattening – showing only the marginal decrease in WC-SSD for increase in K and thus causing too many clusters.

For finding best K using **SC plot, we look for the "peak"** point since as the SC approaches 1, the quality of cluster improves. On the left side of peak are the K values which represent too few clusters whereas on the right side of the peak are the K values that show too many clusters.

Based on this reasoning, following are the "best" K values across 3 data sets and 2 scores:

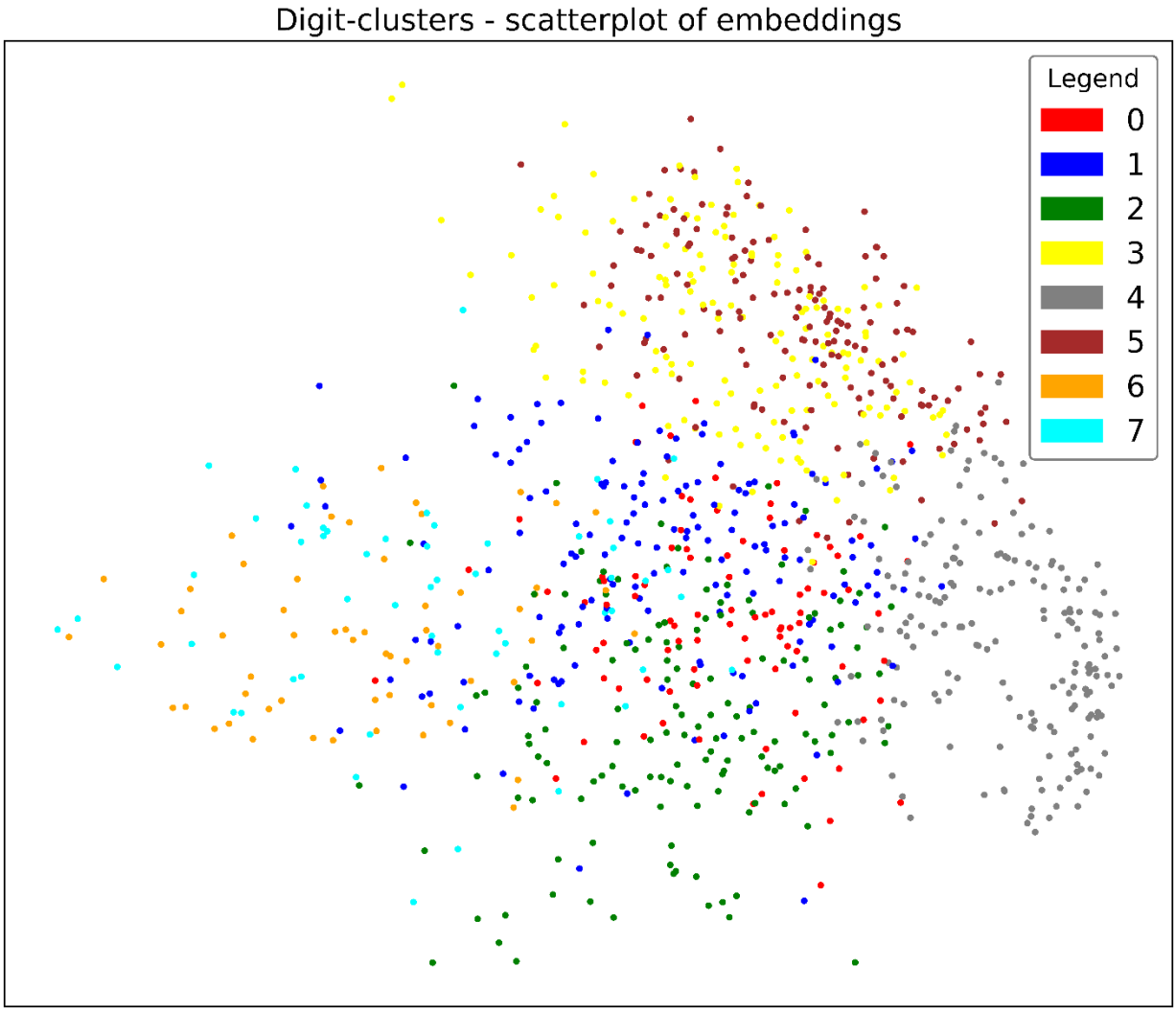
However, we observe that the differences in SC values are much more clear in Part B than in PCA part.

Data	Best K (WC-SSD)	Best K (SC)
Data_1	8	8
Data_2	4	4
Data_3	2	2

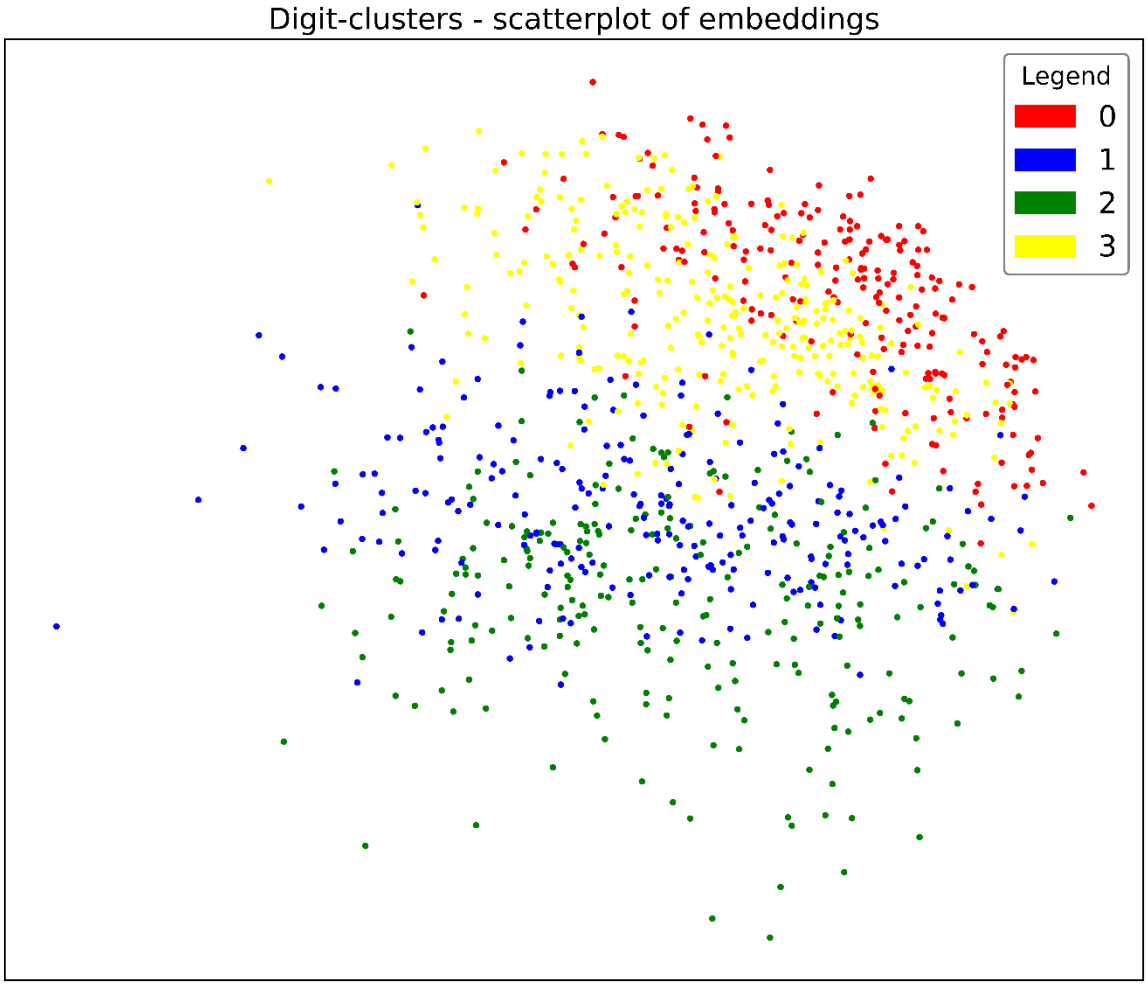
Table 3

We observe that both the scores agree on suggested best K value for all three datasets.
The suggested number of clusters i.e. K varies across the three datasets in alignment with the original number of image labels in the given datasets, which is an indication of good clustering.

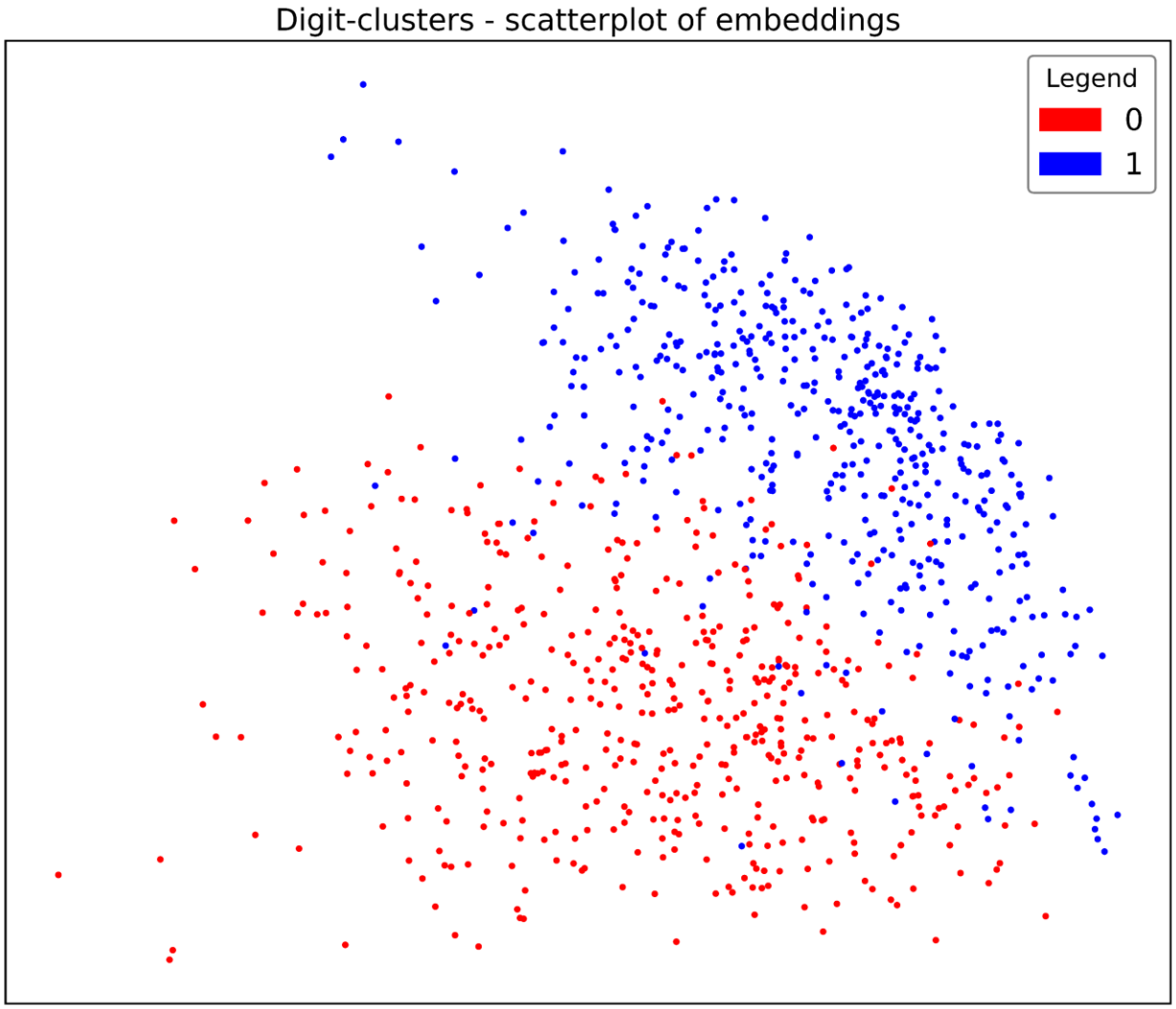
B4 with PCA:



Plot 19: Scatterplot with PCA (Data_1, K=8)

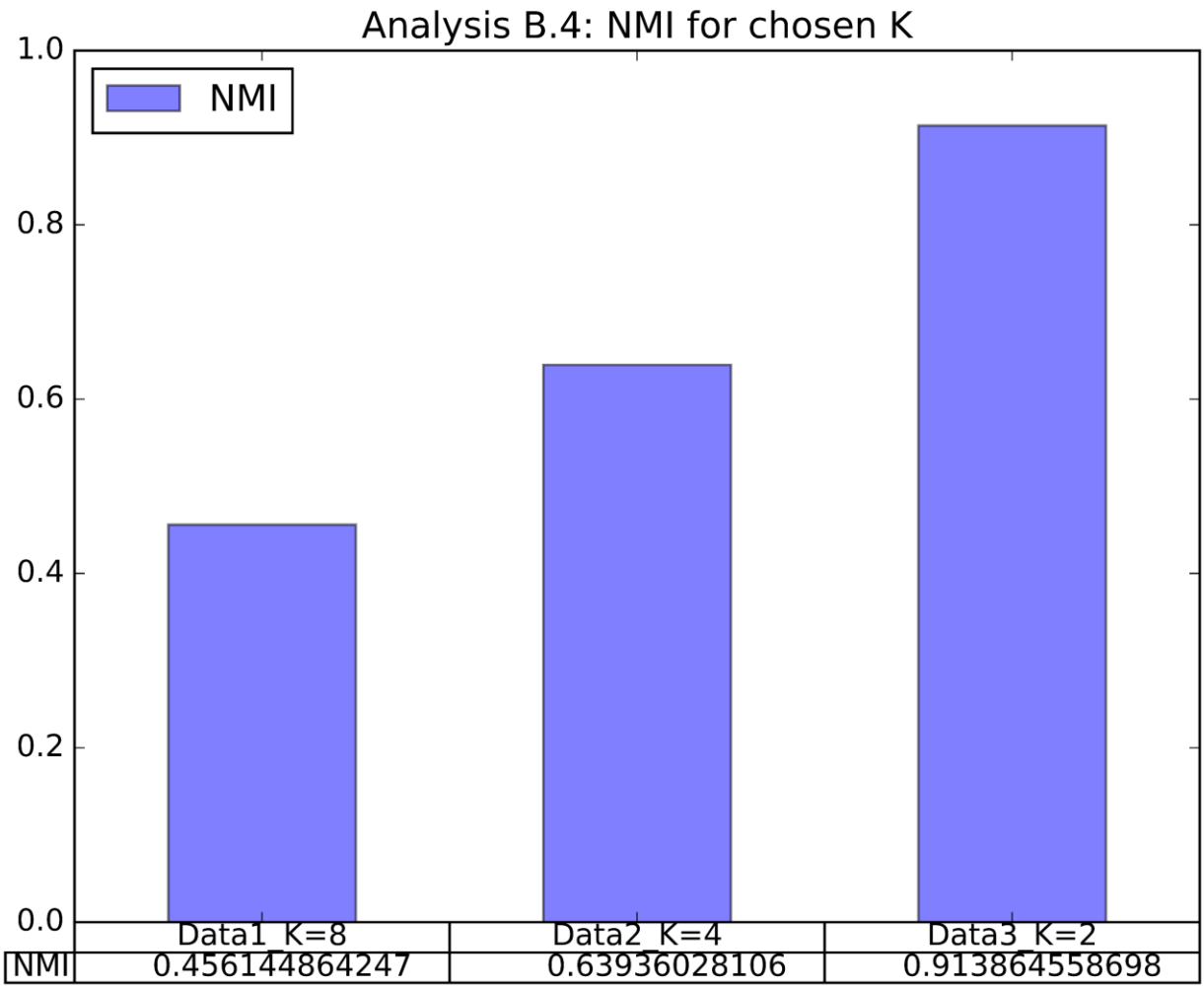


Plot 20: Scatterplot with PCA (Data_2, K=4)



Plot 19: Scatterplot with PCA (Data_3, K=2)

NMI:



Plot 20: NMI with PCA for 3 Datasets and Chosen K

Discussion: - Comparison of PCA outcomes with tSNE:

1. tSNE and PCA both parts gave same K-values for resp. data subsets.
2. From the visualization scatterplots, we observe that K means clustering visualization with tSNE for given K=8, 4, 2 values is much superior to clustering visualization results provided by PCA.
3. tSNE clustering shows much better inter-cluster separation at all values of K.
4. In the case of PCA however, the inter-cluster separation is poor for all values of K.
5. This is due to the dimensionality reduction applied. Additionally, we are selecting only the top 2 principal components causing further loss of information about the features contributing towards the weak clustering output

6. NMI is also lower for PCA than tSNE for all Data subsets.