Meng Deng
Dr. Jennifer Neville
CS 573, hw 5
April 28, 2017

Clustering Analysis

Part A. Exploration

1. Visualize one randomly picked digit.

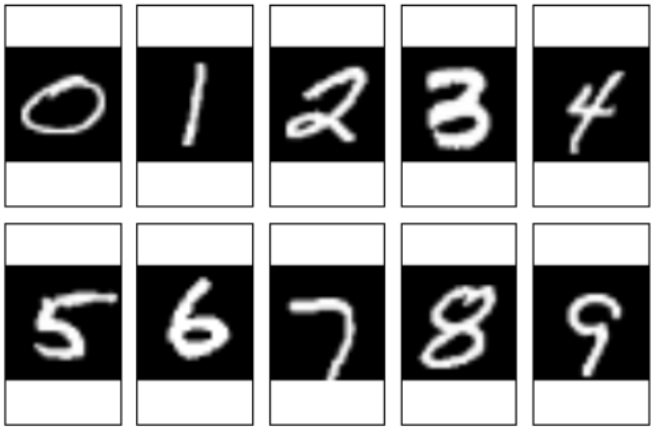


Figure 1: Image Visualization.

2. Visualize 1000 randomly select examples in 2D.

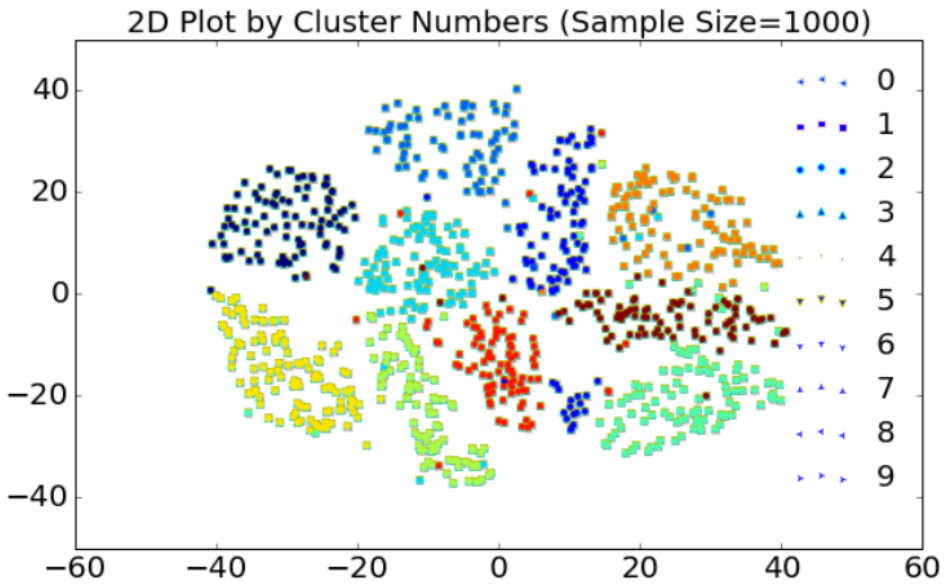
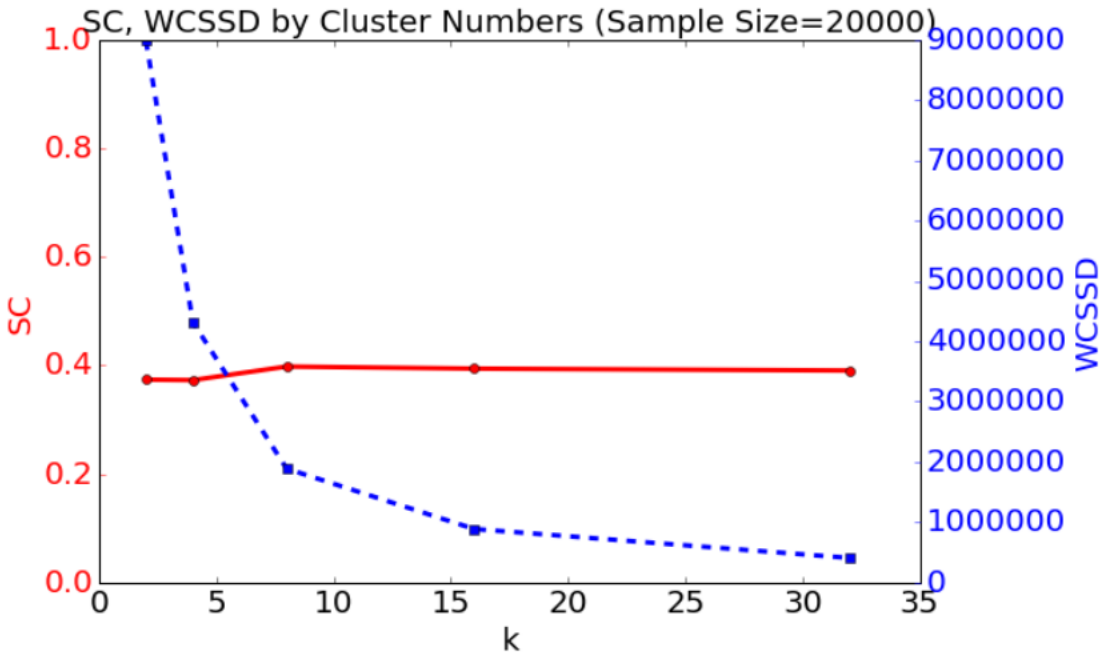


Figure 2: 2D Image Visualization.

Part B. Analysis of K-means.

1. Plot the WC SSD and Silhouette Coefficient by K = 2, 4, 8, 16, 32.



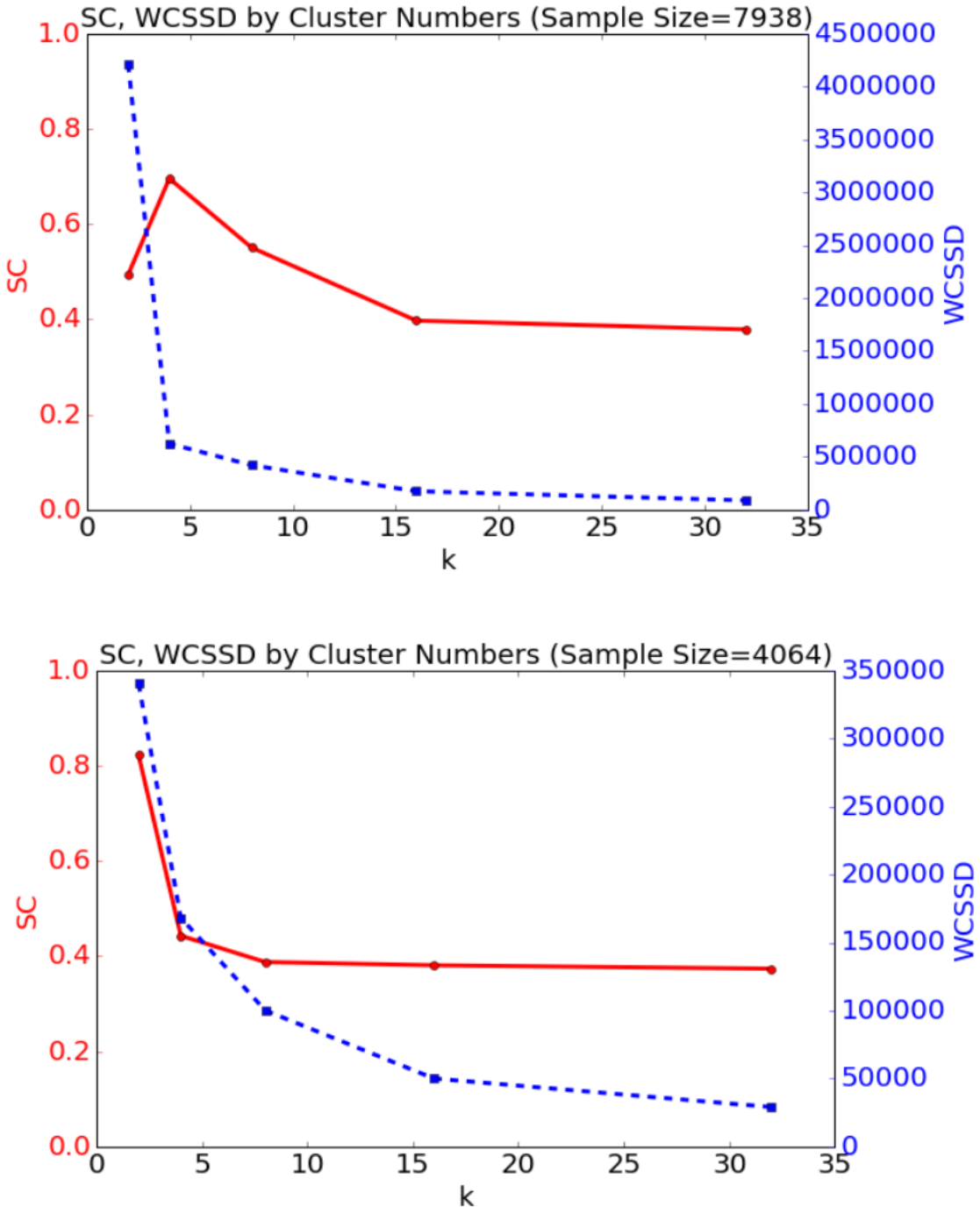
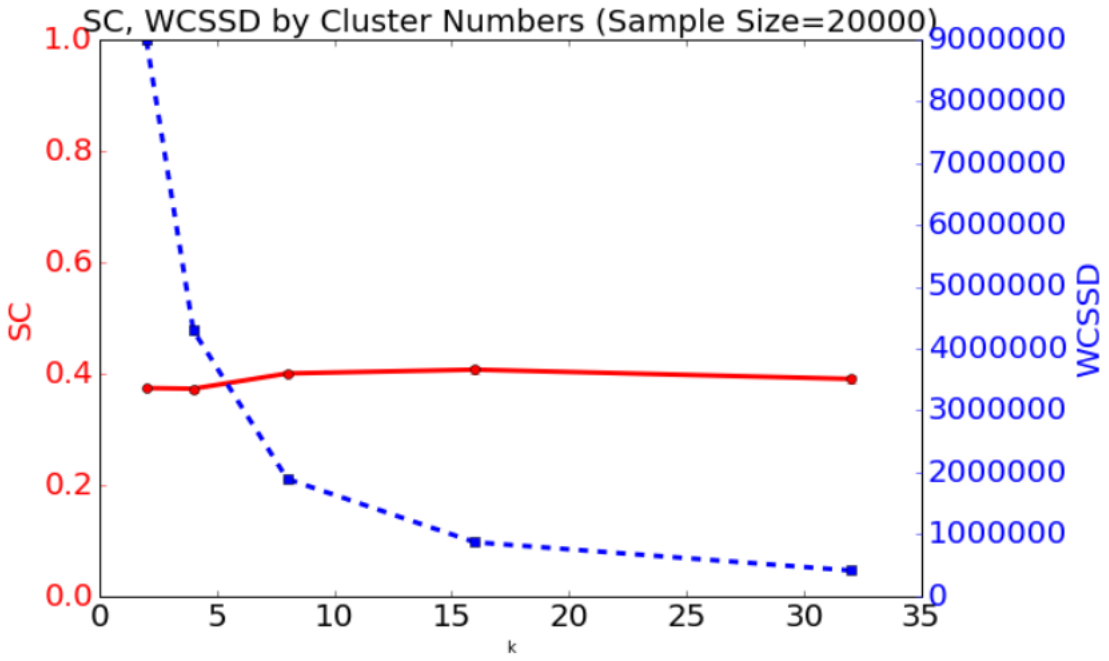
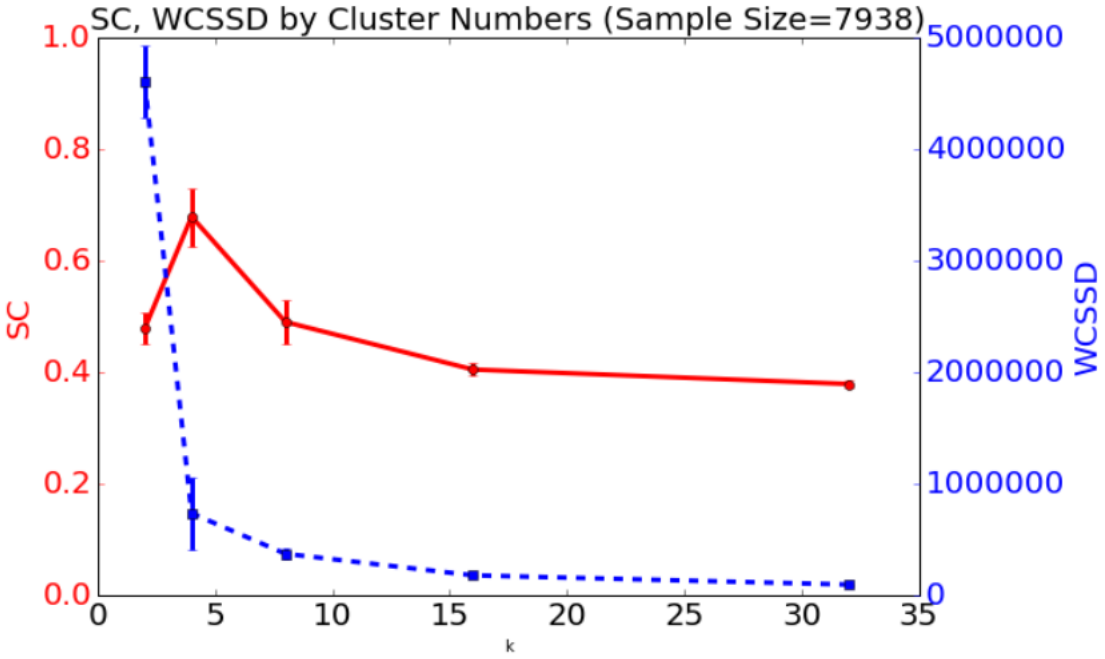


Figure 3: Within Cluster Sum Squared Distances, Silhouette Coefficient by K.

2. I chose K = 16 for dataset 1, K = 4 for dataset 2, and K = 4 for dataset 3. It is because we are trying to find K where Silhouette is relatively high while WC SSD is relatively low. As Figure 3 suggests, Silhouette either peaks or maintain in a relatively stable state around those specific K values; Meanwhile, the WC SSD dropped to a relatively low value around those picked K.

3. I observed increasing variance due to the randomization of the initial states as the dataset size decreases, especially when the K number is small. My choice of K remains the same as suggested in Figure 3.





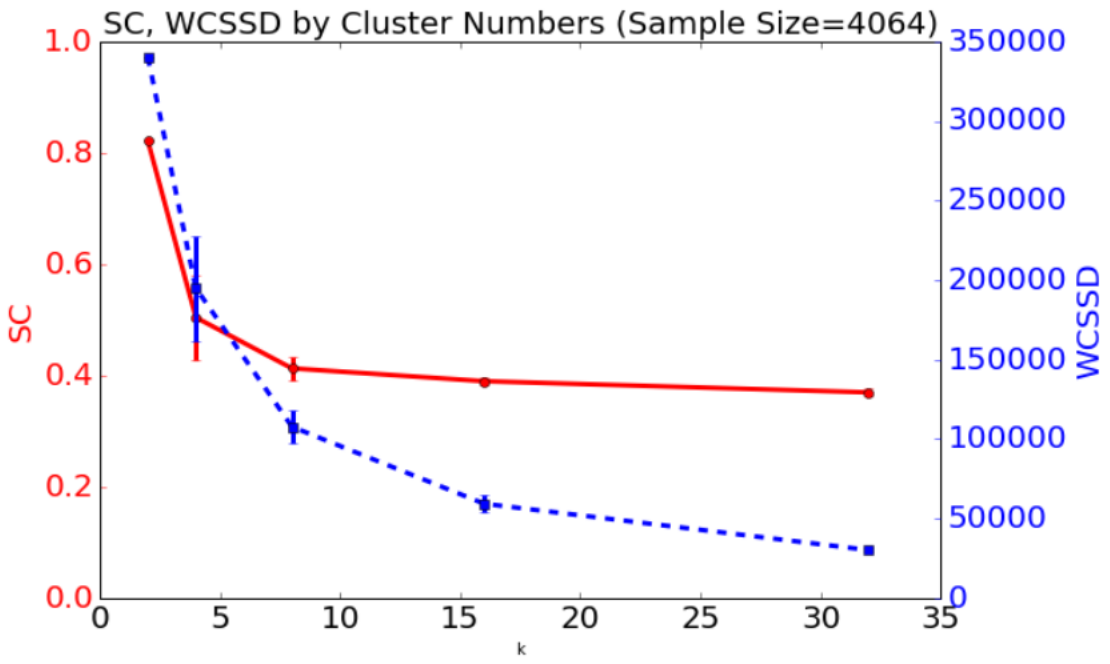
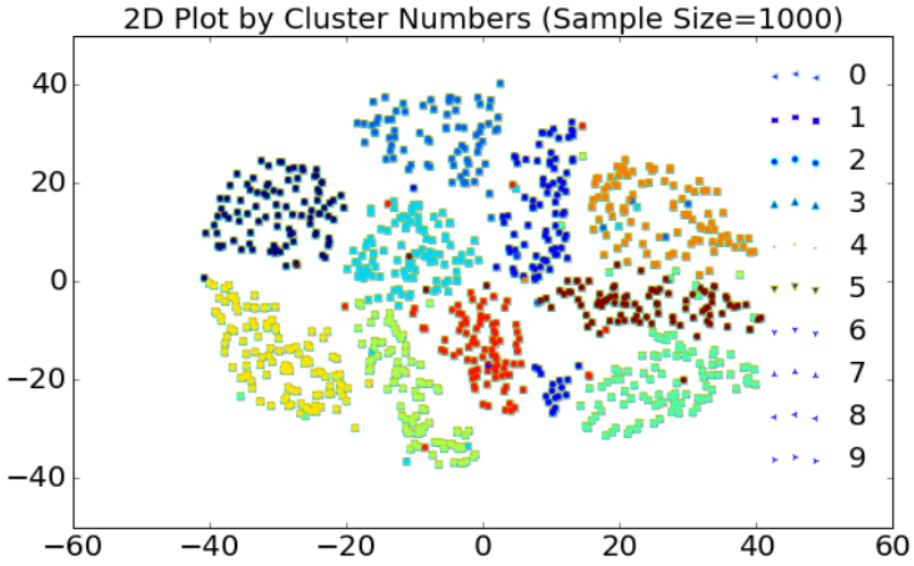


Figure 4: Mean and Std. Dev. of Within Cluster Sum Squared Distances, Silhouette Coefficient by K.

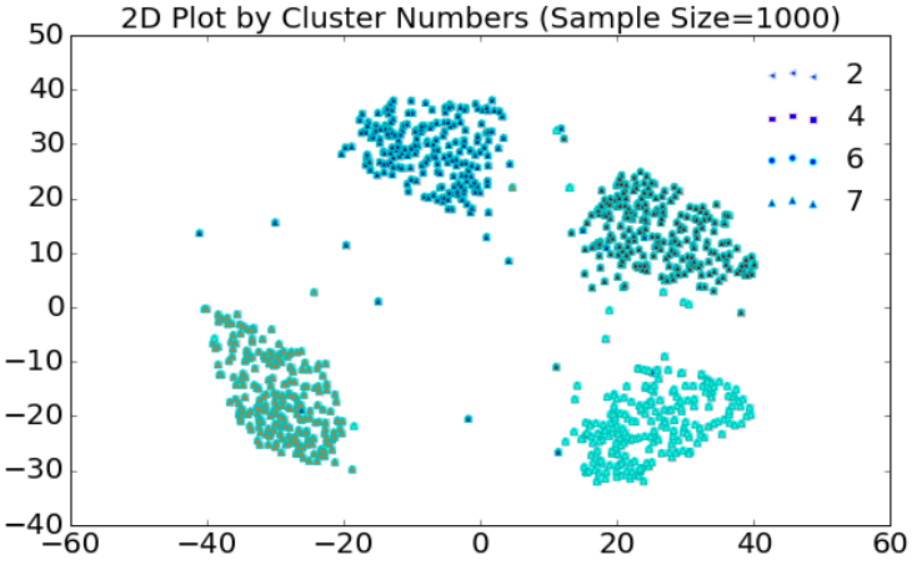
4. The NMI index output and plots for the three datasets are showing below.

	Dataset 1	Dataset 2	Dataset 3
K	16	4	4
NMI	0.37	0.45	0.33
Visualization	Generally good separation by the class numbers.	Clear four clusters	Clear two clusters. Rerun the clustering with K=2 returns a higher NMI.

NMI: 0.370607350229(sample size=20000)



NMI: 0.45465341281(sample size=7938)



NMI: 0.327320174638(sample size=4064)

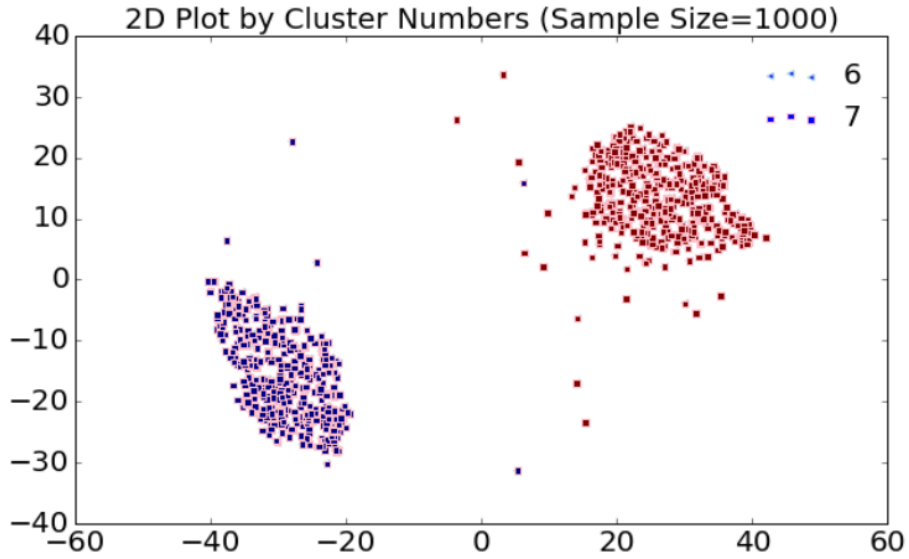


Figure 5: 2D Plot by Cluster Numbers and NMI index.

Part C. Hierarchical clustering comparison.

1. Dendrogram by single linkage.

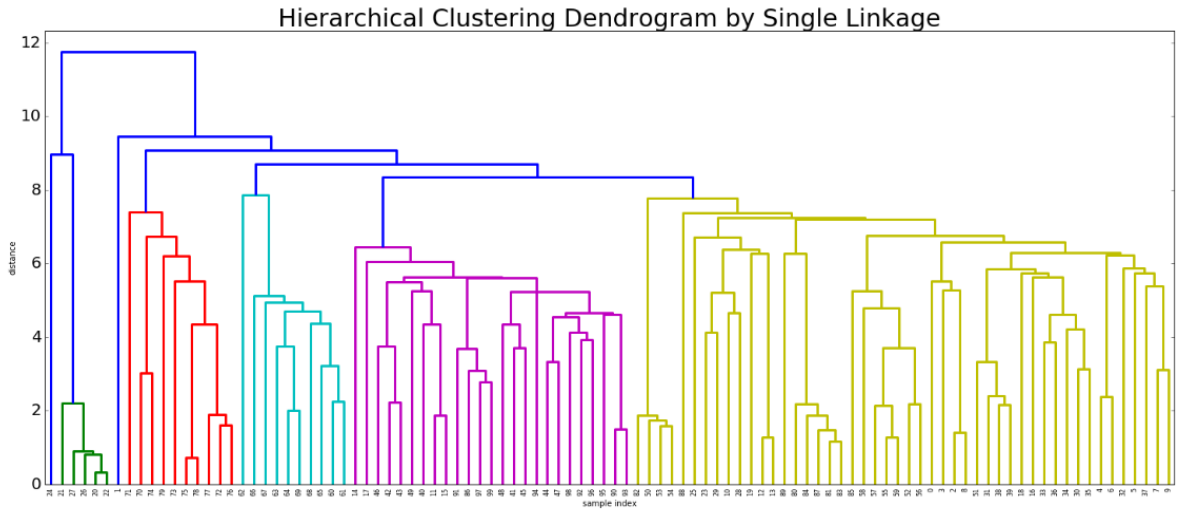


Figure 6: Hierarchical Clustering Dendrogram by Single Linkage.

2. Dendrogram by complete linkage and average linkage.

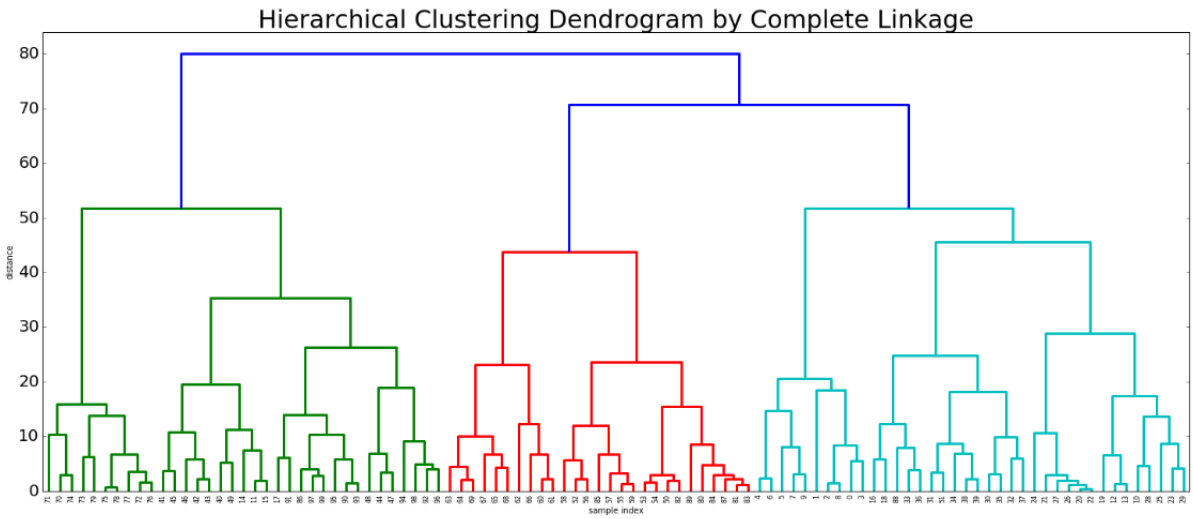


Figure 7: Hierarchical Clustering Dendrogram by Complete Linkage.

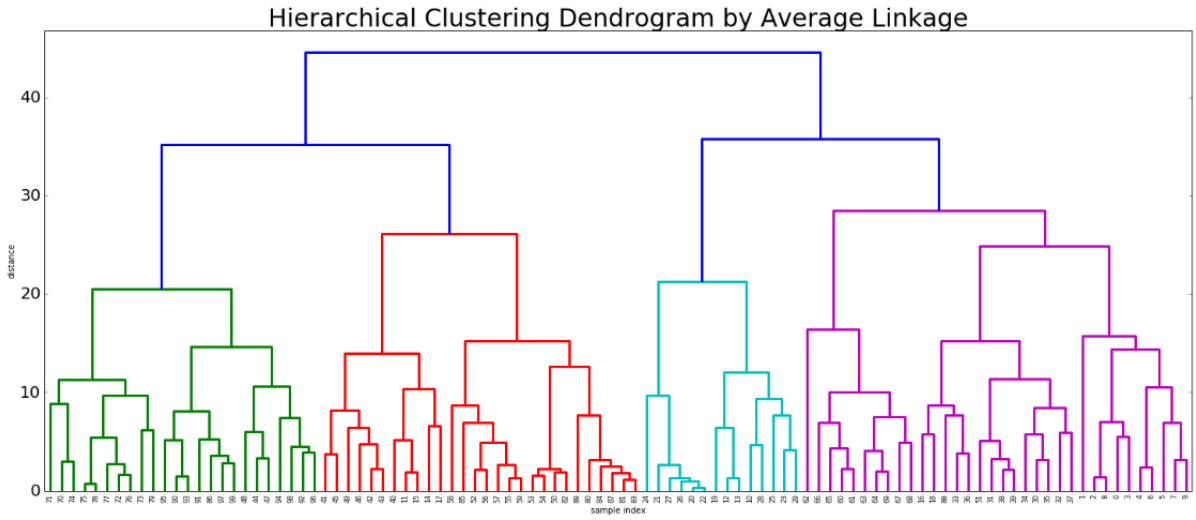


Figure 8: Hierarchical Clustering Dendrogram by Average Linkage.

3. For single linkage, I picked K = 2, 3, 5, 7, 8, 15, 27; for complete linkage, K = 2, 3, 4, 5, 7, 8, 16, 30; for average linkage, K = 2, 4, 6, 8, 9, 16, 30.

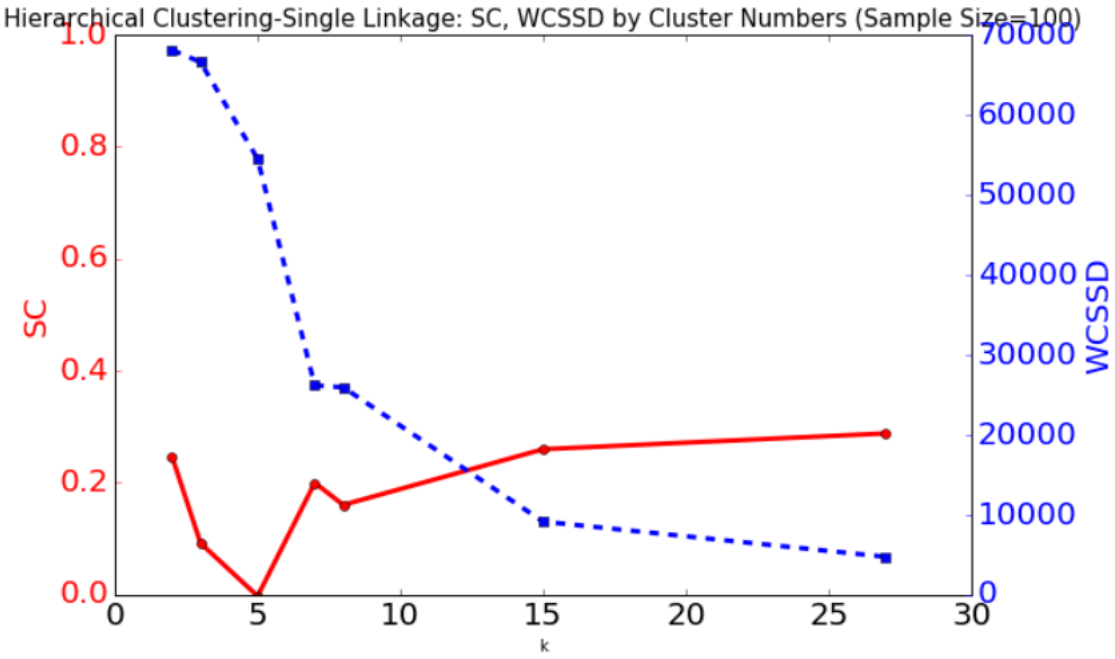


Figure 9: Hierarchical Clustering-Single Linkage: SC, WCSSD by Cluster Numbers.

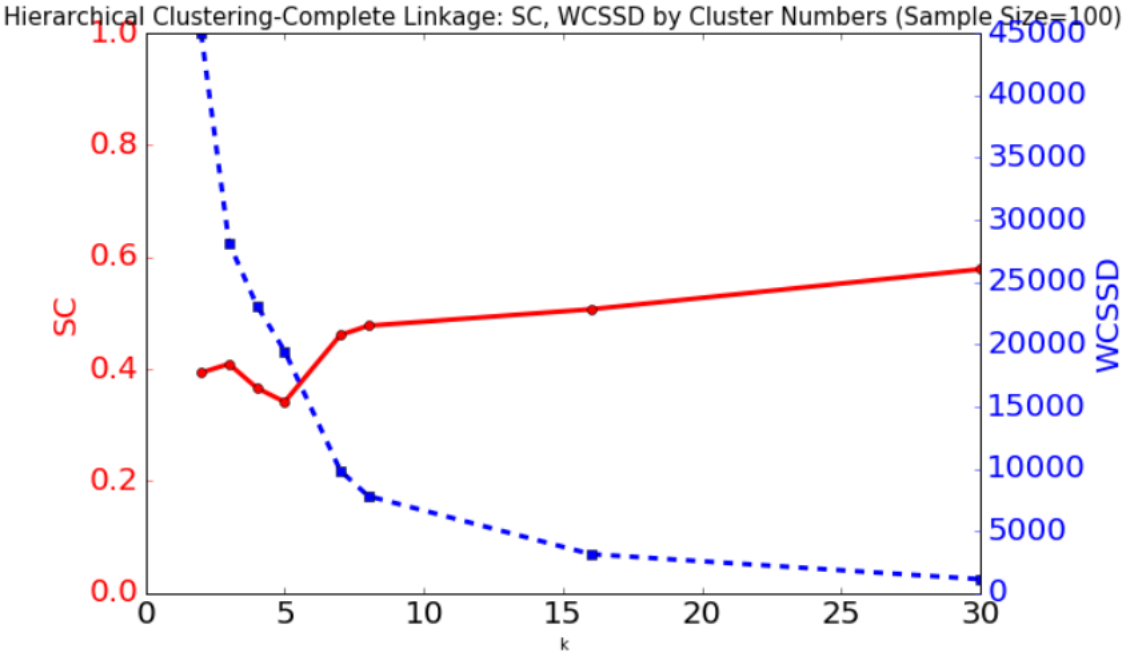


Figure 10: Hierarchical Clustering-Complete Linkage: SC, WCSSD by Cluster Numbers.

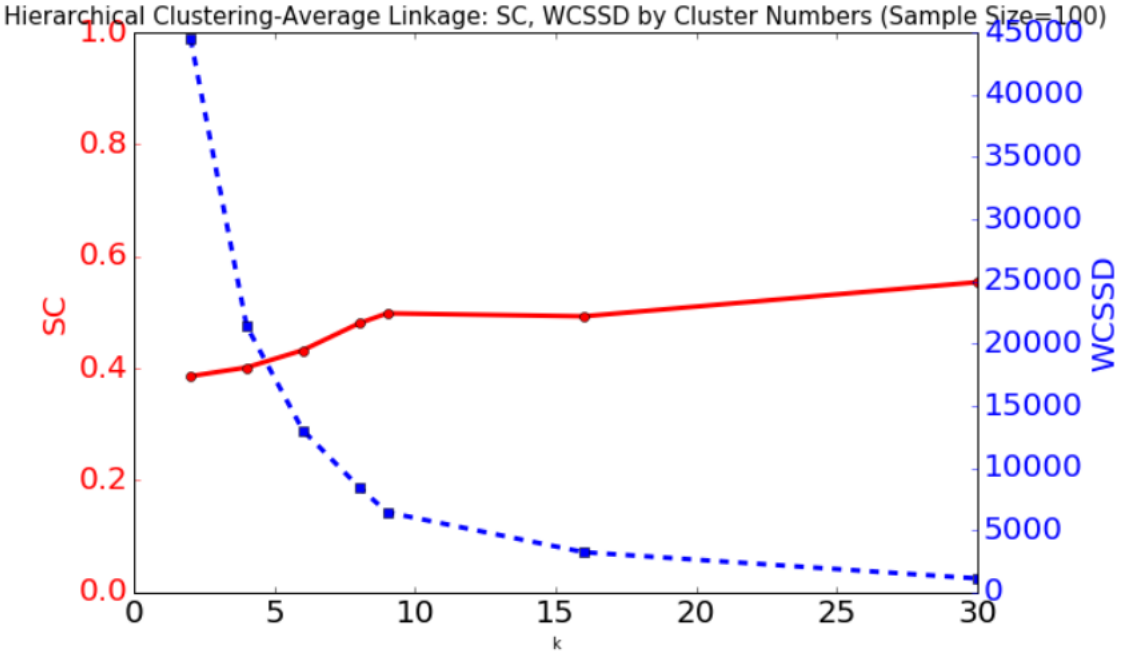


Figure 11: Hierarchical Clustering-Average Linkage: SC, WCSSD by Cluster Numbers.

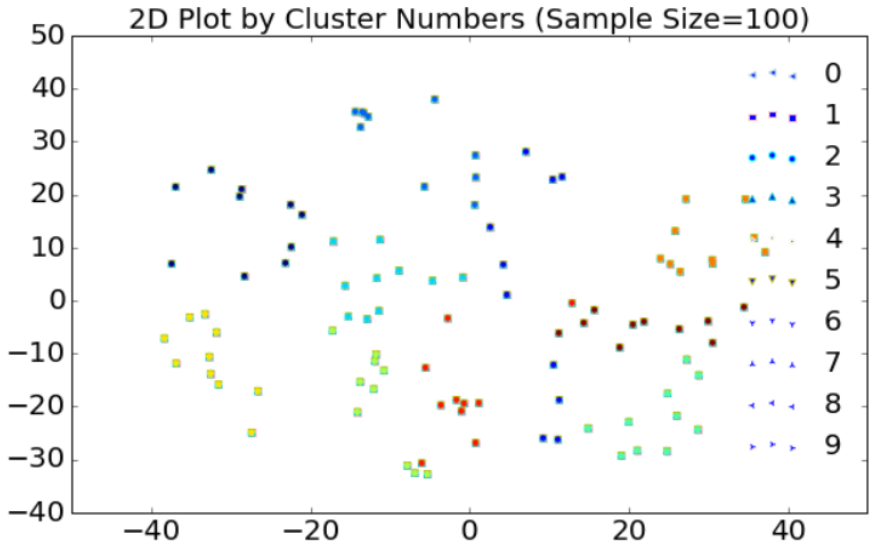
4. I would pick K = 15 for single linkage, K = 8 for complete and average linkages. This is similar as the K chosen when running k-means on the full and subset datasets.

5. The table below shows the comparison of the NMI across the three distance measures. The complete linkage with 8 clusters returns the highest NMI index, followed by average and single linkage. We can probably see the pattern from the dendrogram as single linkage tends to lead to unbalanced clusters.

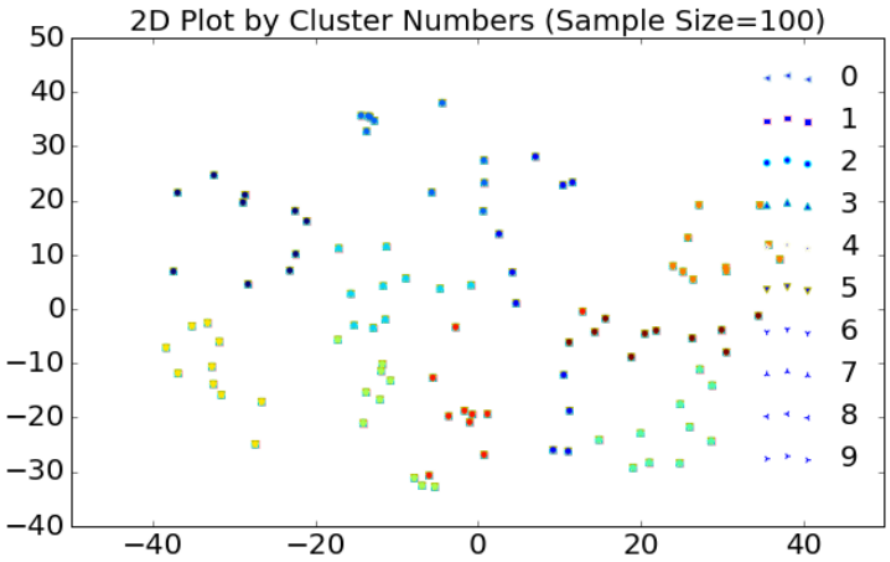
Compared to the k-means methods, I don't find any obvious difference in terms of the NMI as well as the number of clusters concluded.

	Single	Complete	Average
K	15	8	8
NMI	0.37	0.41	0.39
Visualization	Generally good separation by the class numbers.	Generally good separation by the class numbers.	Generally good separation by the class numbers.

```
Hierarchical Clustering-Single Linkage NMI: 0.368504509103(sample size=100)
```



```
Hierarchical Clustering-Complete Linkage NMI: 0.406396982639(sample size=100)
```




```
Hierarchical Clustering-Average Linkage NMI: 0.387099101576(sample size=100)
```

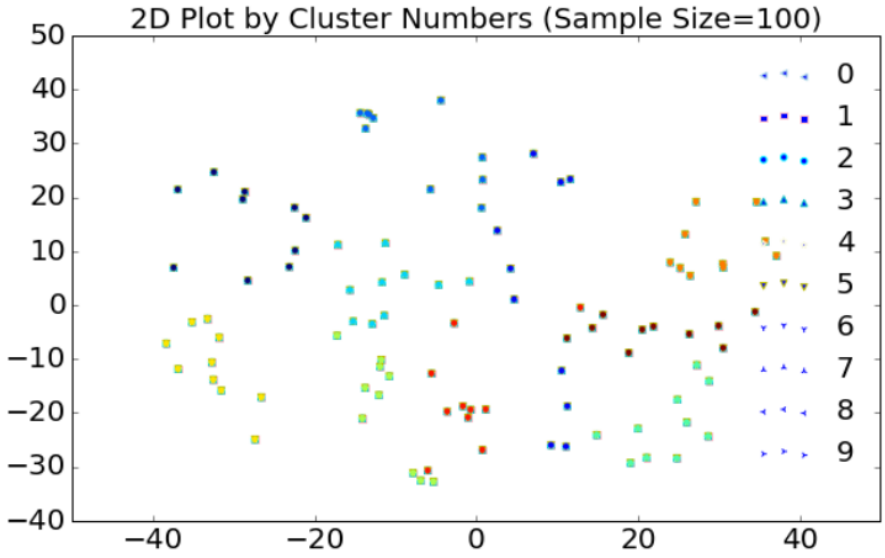


Figure 12: 2D Plot by Cluster Numbers and NMI index of Single, Complete, and Average.

Part D. Bonus PCA.

1-2. The top two PCs contributes to only 49% of the total variance. Therefore, we can tell from the image of using only the first two PCs.

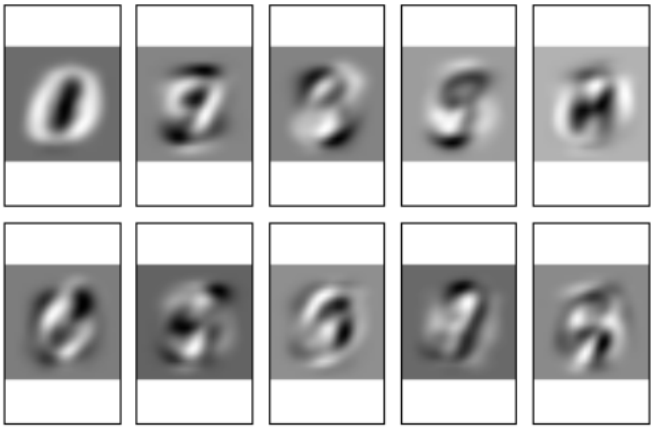


Figure 13: Image Visualization after Dimension Reduction.

3. The pattern is less clear compared to the TSNE embedding in part B.

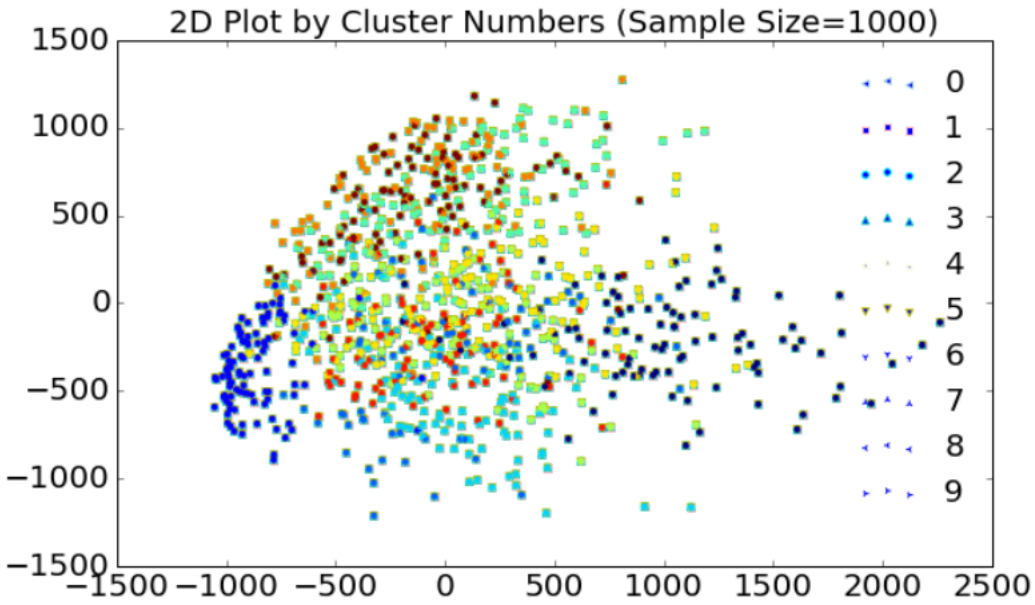


Figure 14: 2D Plot by Cluster Numbers after Dimension Reduction.

4. I picked K = 16 according to Figure 15. The result is consistent as the K picked when applying the full TSNE embedding dataset. Using the chosen K, I calculated the NMI=.26, which is much lower than the NMI using the TSNE embeding (.37). The visualization also suggests a less clear clustering compared to the TSNE.

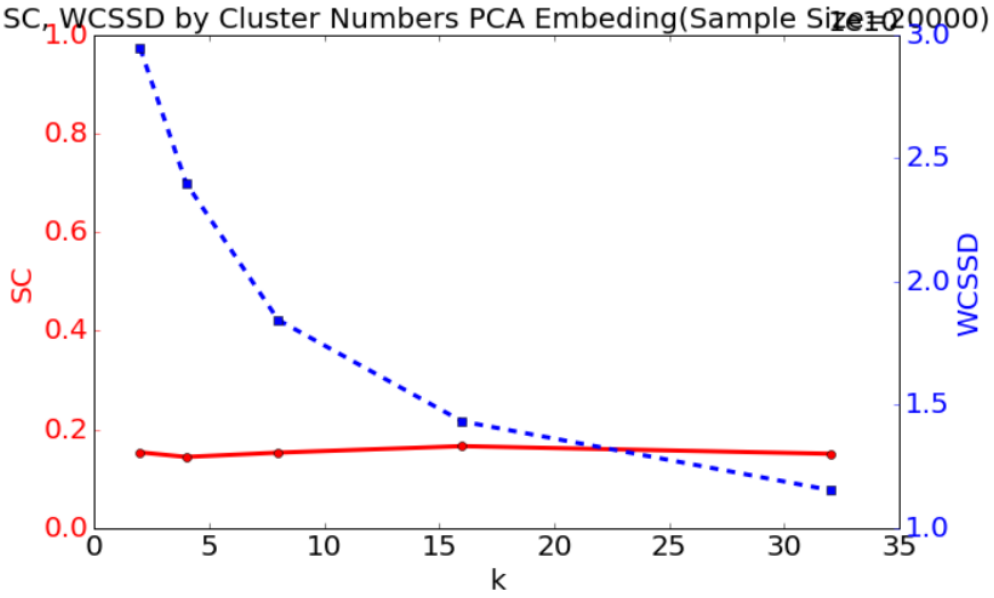


Figure 15: SC, WCSSD by Cluster Numbers PCA Embedding.

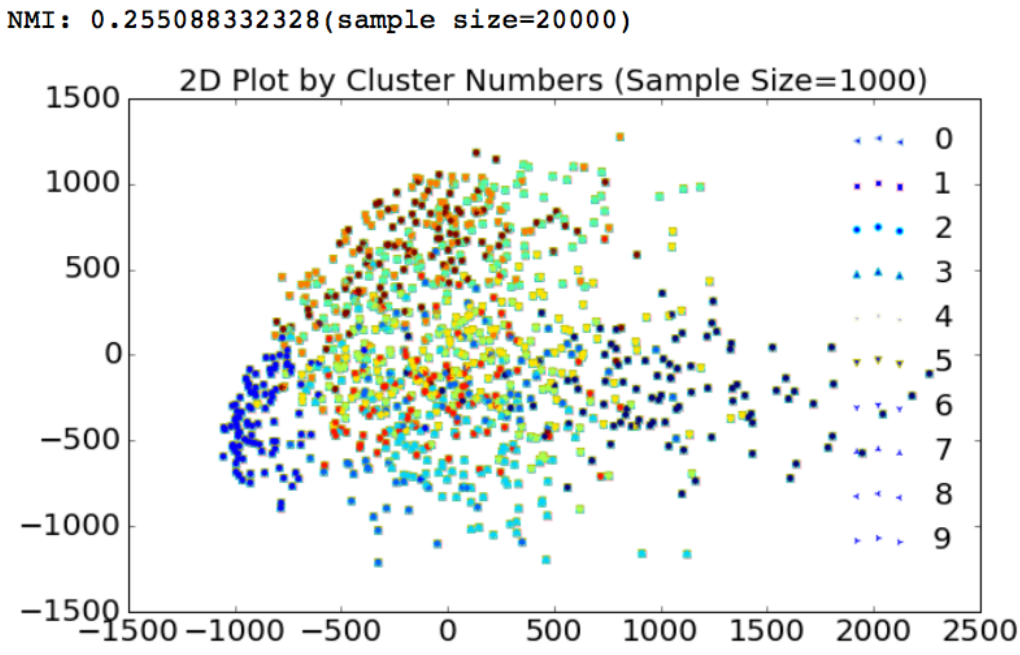


Figure 16: 2D Plot by Cluster Numbers PCA Embedding.

5. Now rerun the dimension reduction on the full dataset, subset with image 2,4,6,7, and 6,7 respectively.

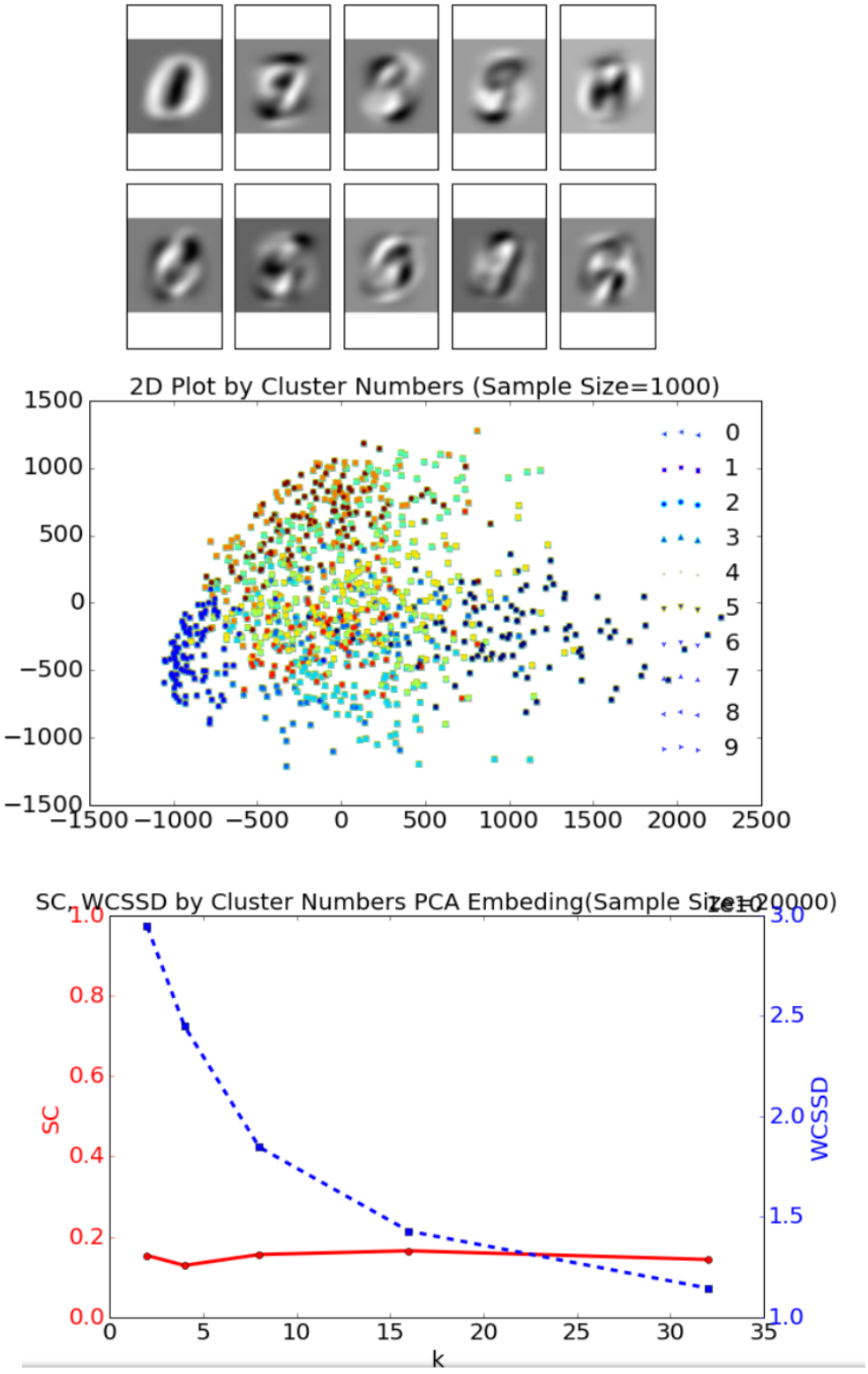
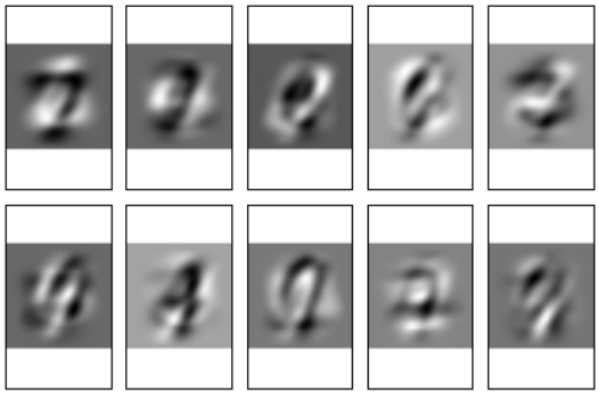
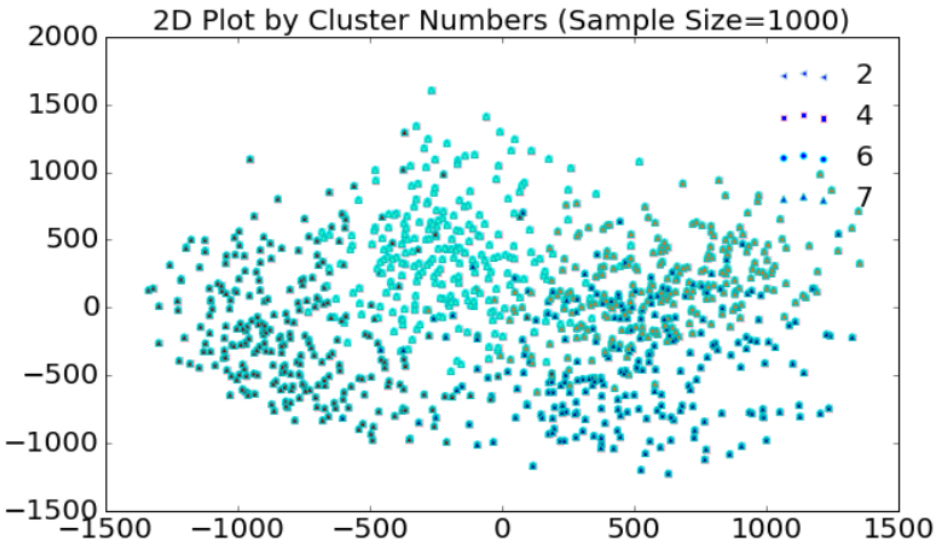


Figure 17: PCA Embedding Plots (Sample Size=20000).





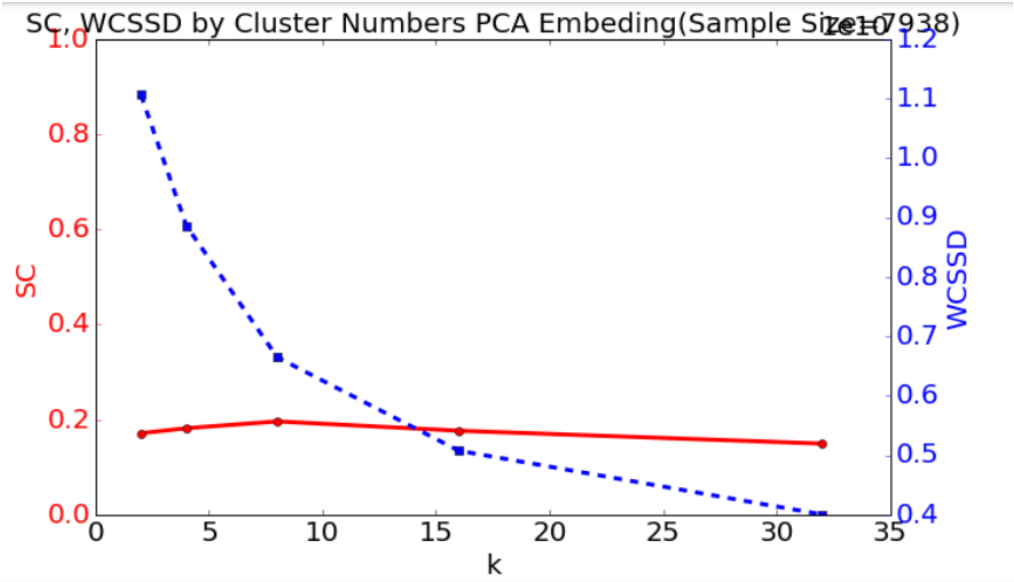
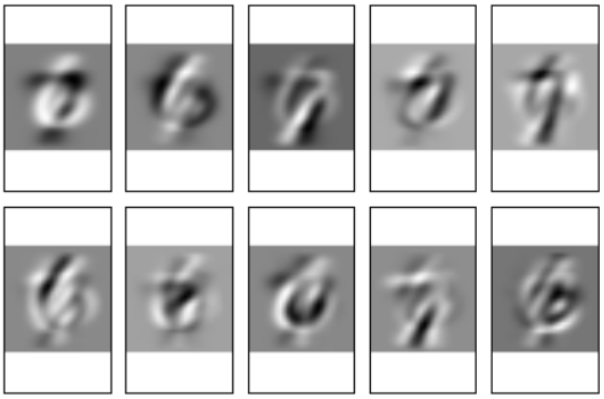
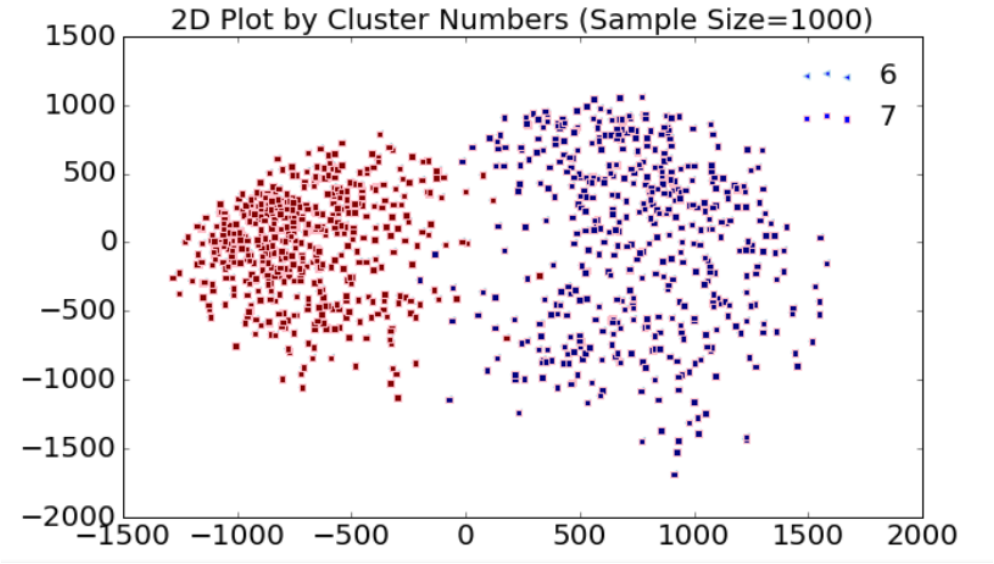


Figure 18: PCA Embedding Plots (Sample Size=7938).





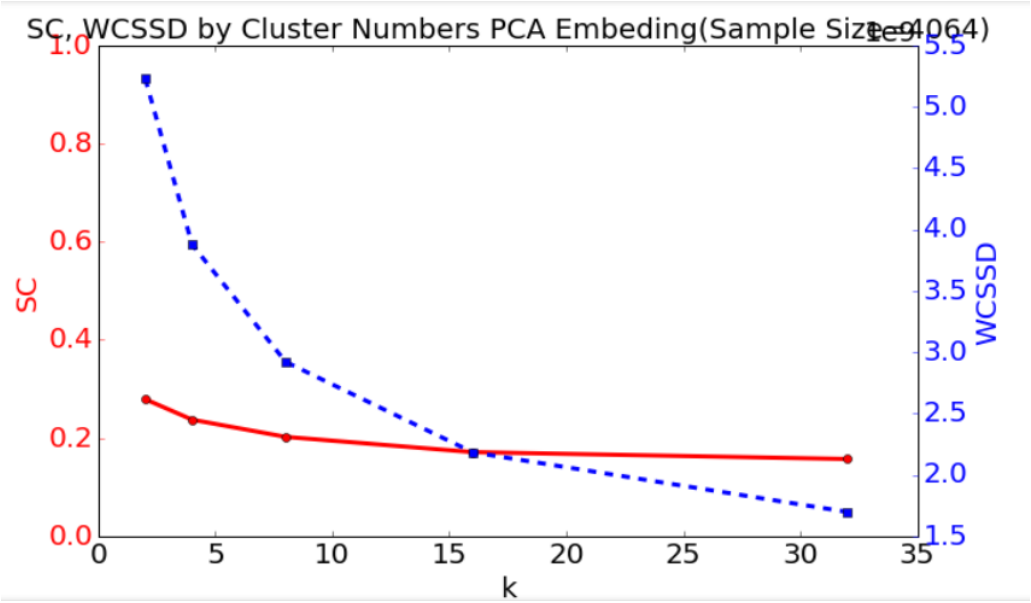
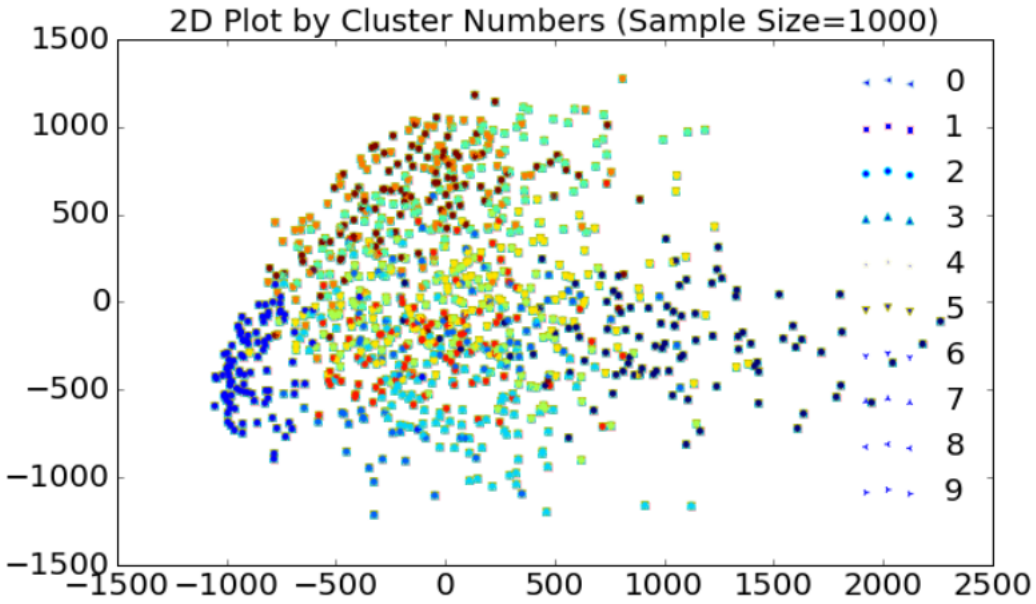


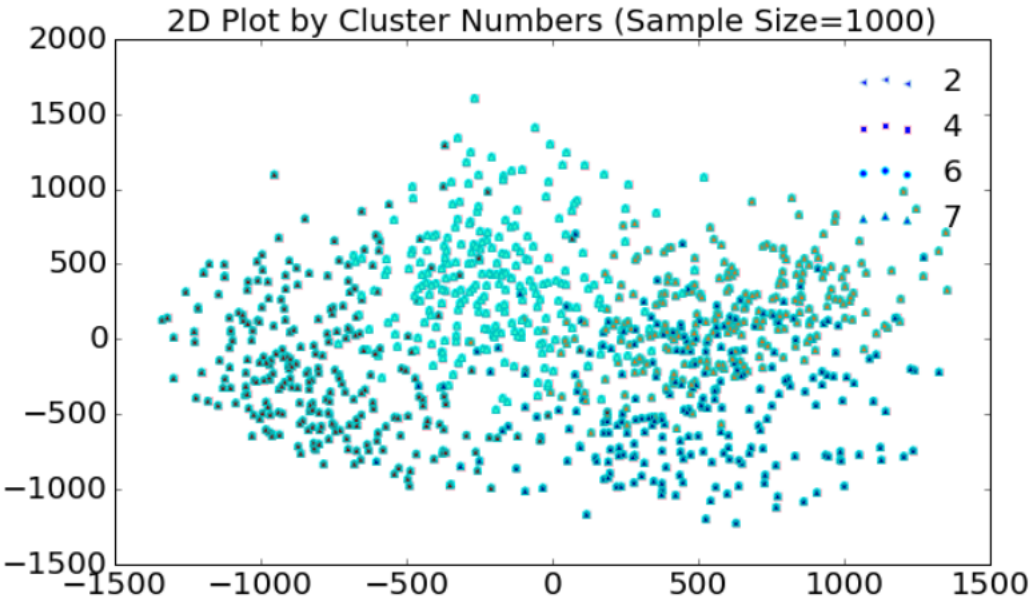
Figure 19: PCA Embedding Plots (Sample Size=4064).

Based on the output, I selected k = 16, 8, 8 for the full dataset and the two subsets respectively. From figure 20, we can tell that dataset 2 and k = 8 returns the highest NMI value. However, the NMI drops from all the three datasets compared to those on the three datasets using TNSE embedding datasets.

```
NMI: 0.251508538461(sample size=20000)
```




```
NMI: 0.294626814202(sample size=7938)
```



NMI: 0.242233772737(sample size=4064)

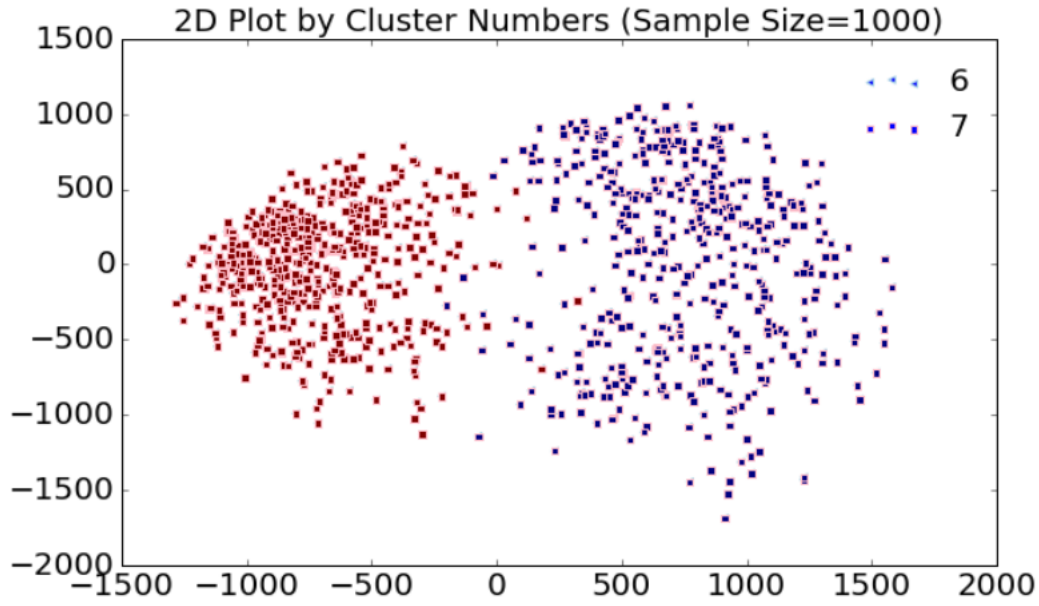


Figure 20: PCA Embedding 2-D Image and NMI Plots.

In sum, the PCA embedding methods using the top 10 eigenvectors produced results that are not as good as the TNSE embedding methods.