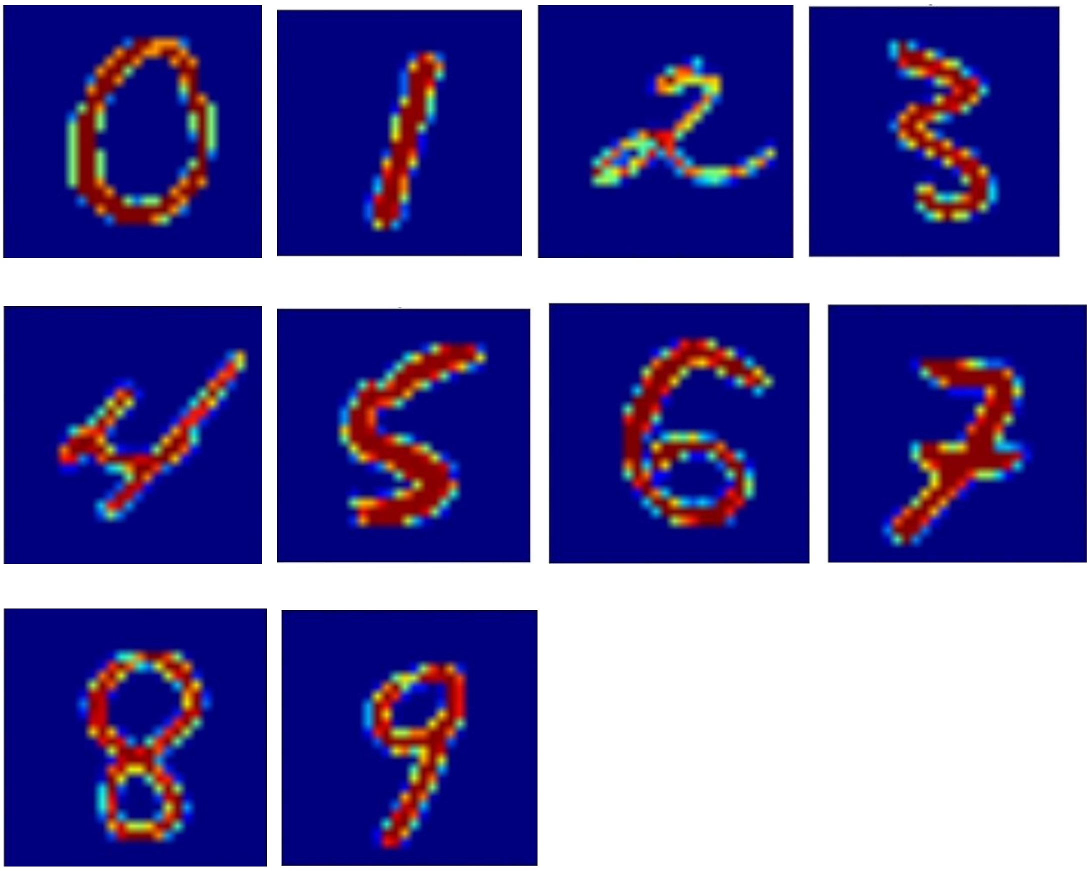
CS 573: Homework 5

Piyush Dugar (pdugar@purdue.edu)

Used 1 extra day

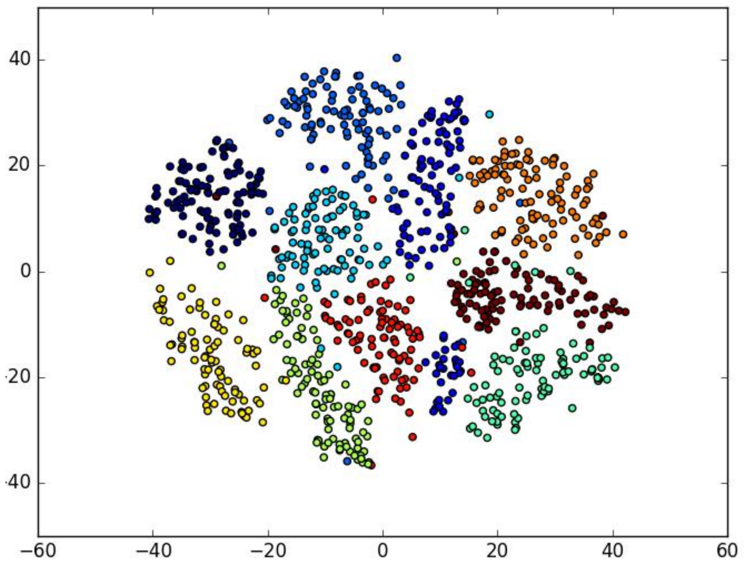
A. Exploration

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



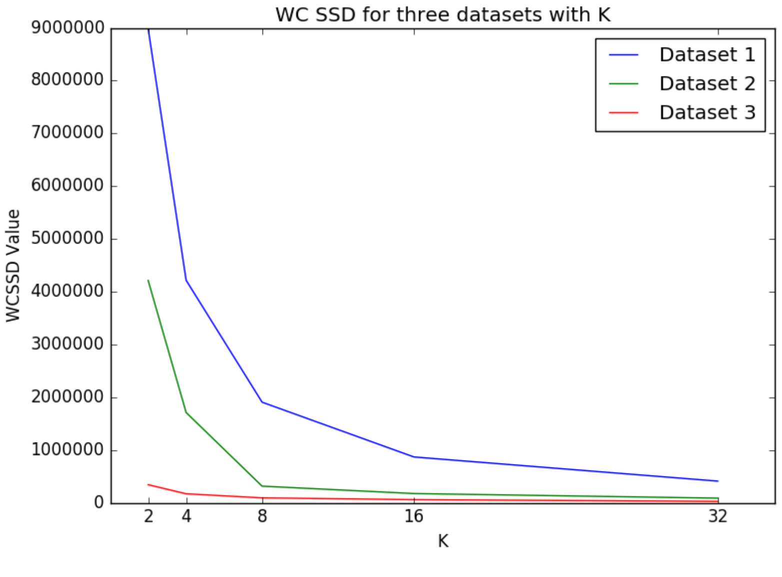
2. Visualize 1000 randomly selected examples in 2d, coloring the point to show their corresponding class label

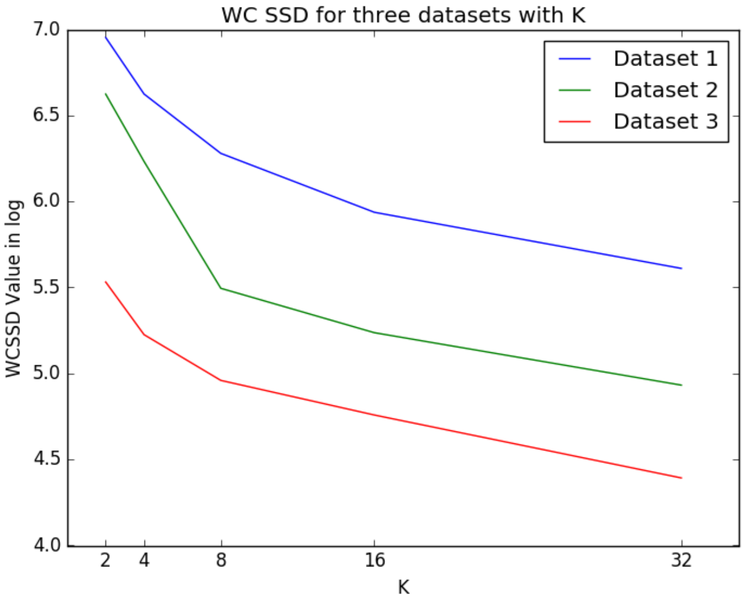
Visualization of the tsne 2d data with class labels

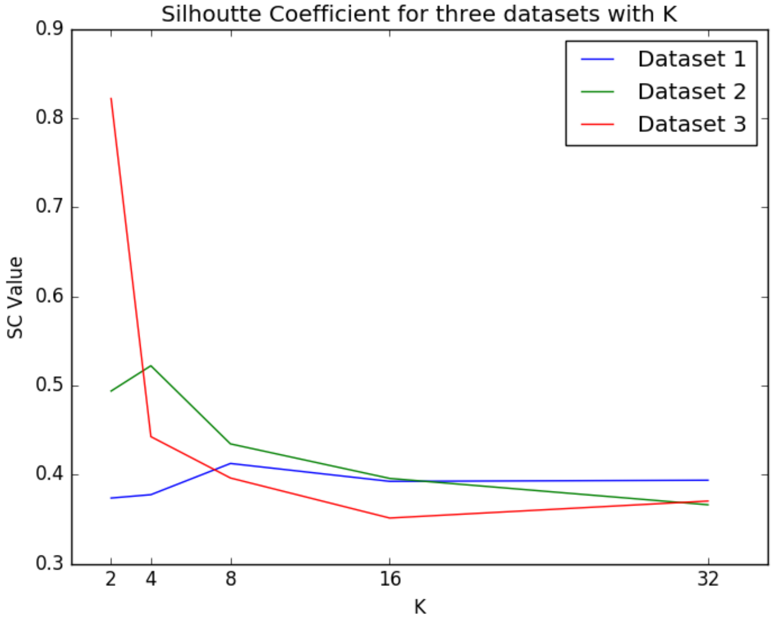


B. Analysis of k-means

1. Cluster the data with different values of K ∈ [2, 4, 8, 16, 32] and construct a plot showing the within-cluster sum of squared distances (WC SSD) and silhouette coefficient (SC) as a function of K







2. Using the results from B.1, choose an appropriate K for each dataset and argue why your choice of K is the best. Discuss how the results compare across the two scores and the three versions of the data.

Dataset 1 scores. Number of classes = 10

K	WCSSD	SC Value
2	8983899.99	0.3736
4	4215072.74	0.3773
8	1904477.13	0.4123
16	865680.73	0.3923
32	408297.04	0.3935

Dataset 2 scores. Number of classes = 4

K	WCSSD	SC Value
2	4211155.68	0.4936
4	1708759.43	0.5220
8	312860.81	0.4343
16	172693.32	0.3956
32	85586.42	0.3660

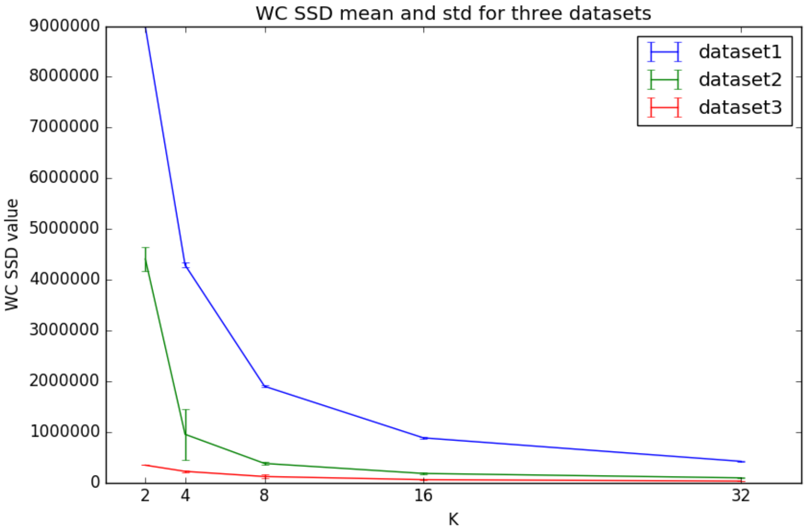
Dataset 3 scores. Number of classes = 2

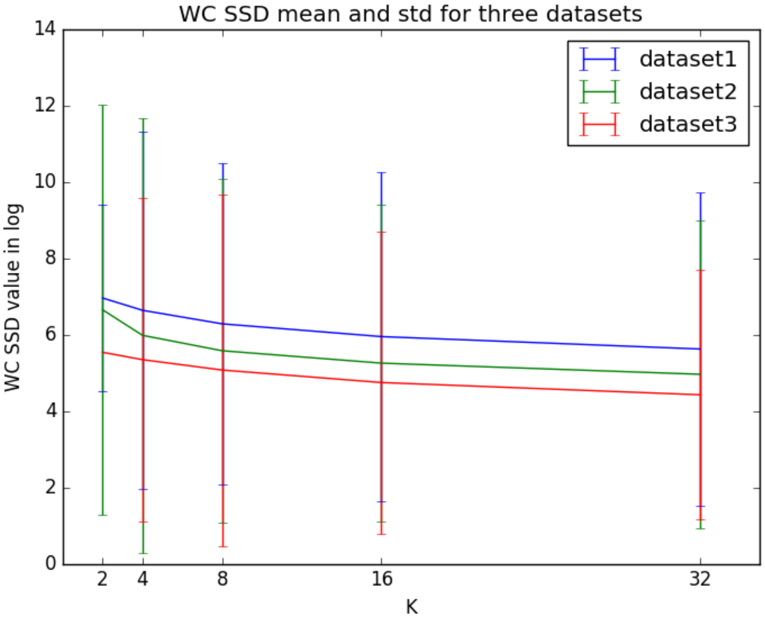
K	WCSSD	SC Value
2	340372.41	0.8218
4	168176.79	0.4424
8	91182.46	0.3960
16	57371.70	0.3511
32	24705.38	0.3702

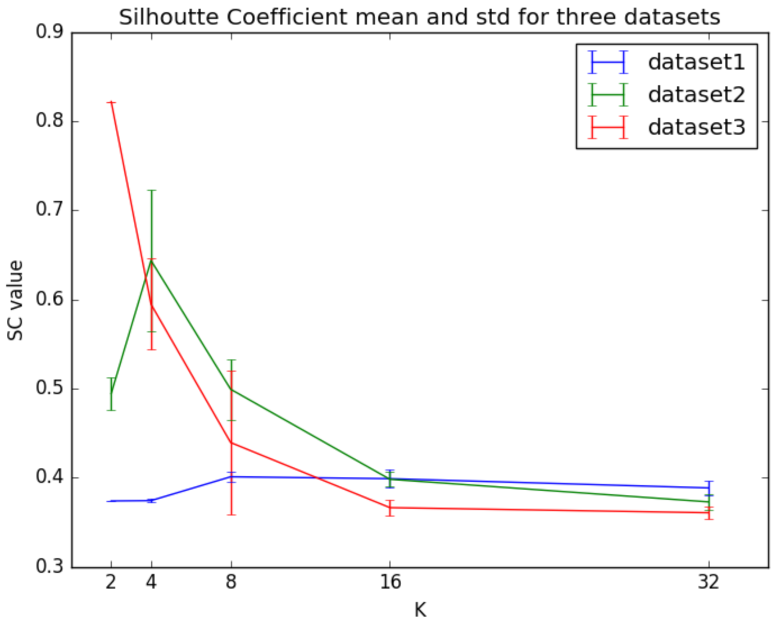
Following are the conclusions:

1. Best K values are selected where we observe the peak in SC plot.
2. We select the following values of K
    - K = 8 for dataset 1. Since the original dataset contains 10 classes, the value of K does make sense
    - K = 4 for dataset 1. Since the dataset 2 contains 4 classes (2, 4, 6 and 7) the value of K we got makes sense
    - K = 2 for dataset 3. Since the dataset 3 contains 2 classes (6 and 7), the value of k we got makes sense
3. WC SSD value for Dataset 1 > Wcssd for Dataset 2 > WC ssd for dataset. This is obvious as their sizes are in this order and as size increases , WC ssd increases.
4. Even from the Wcssd plots, we can see that the maximum slope difference is at the values of K obtained from the SC graph

3. Repeat the experiment from B.1 with 10 different random seeds. Measure and report the average and **STD** (for WC SSD and SC) for the different values of K. Discuss what the results show about k-means sensitivity to initial starting conditions.


WC SSD mean and std for three datasets
WC SSD value
9000000
8000000
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
2
4
8
16
32
K
dataset1
dataset2
dataset3


WC SSD mean and std for three datasets
WC SSD value in log
14
12
10
8
6
4
2
0
2
4
8
16
32
K
dataset1
dataset2
dataset3



Dataset 1 scores. Number of classes = 10

K	WCSSD Average	WCSSD STD	Silhouette Average	Silhouette STD
2	8983359.23	270.38	0.37372559	5.7098e-05
4	4287484.41	47418.03	0.37409602	2.0421e-03
8	1896424.51	16507.28	0.40081563	5.7617-03
16	879088.57	20832.05	0.39869576	1.0352e-02
32	415922.84	12905.08	0.38833567	7.6987e-03

Dataset 2 scores. Number of classes = 4

K	WCSSD Average	WCSSD STD	Silhouette Average	Silhouette STD
2	4407263.22	236063.97	0.49408886	0.018243
4	949333.54	497160.94	0.64348872	0.07952329
8	374621.54	32080.38	0.49903803	0.03395925
16	177430.94	14433.08	0.39802387	0.00799945
32	90727.80	10912.30	0.37268522	0.00905347

Dataset 3 scores. Number of classes = 2

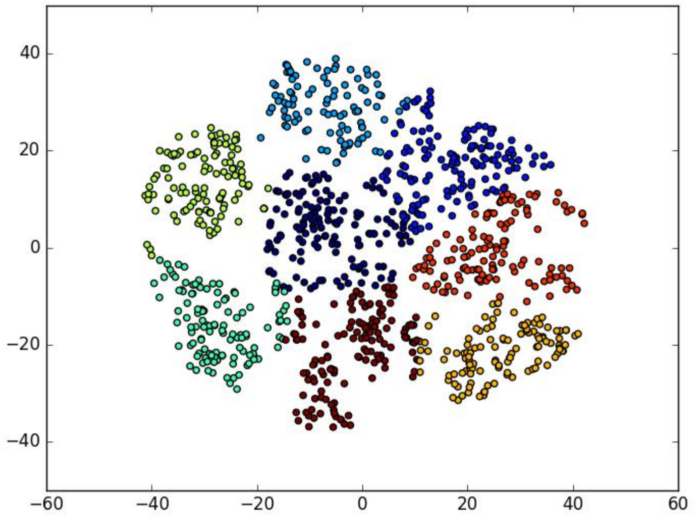
K	WCSSD Average	WCSSD STD	Silhouette Average	Silhouette STD
2	340372.41	0.0	0.82183305	0.0
4	218712.65	17179.83	0.59444393	0.05095082
8	117078.03	39978.66	0.43926272	0.08022748
16	55107.86	9102.69	0.36620968	0.00932584
32	26332.50	1834.00	0.36048141	0.00688781

Following are the conclusions on the dependence of Kmeans on the initial dataset:

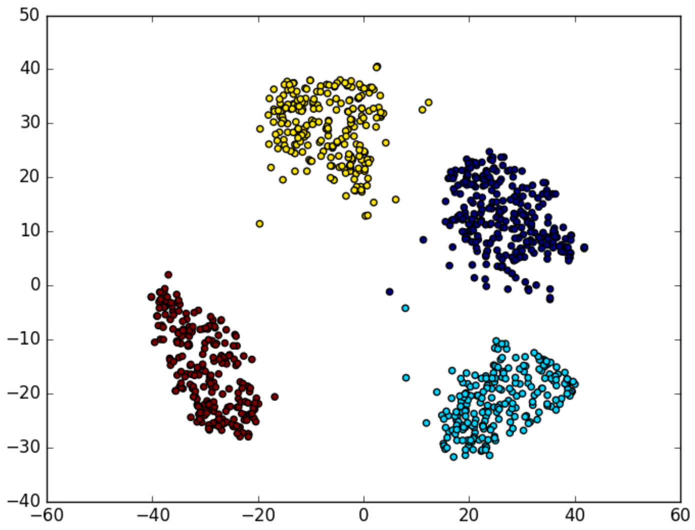
- From the error bar, we can see that the variance is very high. Just to get more insight on the variance, I plotted a log graph of the wcssd and its variance and its evident from that graph that it has a very high variance/std.
- Even in the SC graph, the variance is very high.
- So kmeans is highly depended on the initial starting point.

4. For the value of K chosen in B.2, cluster the data again (a single time) and evaluate the resulting clusters using normalized mutual information gain (NMI). Calculate NMI with respect to the image class labels. Visualize 1000 randomly selected examples in 2d, colouring the points to show their corresponding cluster labels. Discuss how the both the NMI and visualization results compare across the three versions of the data.

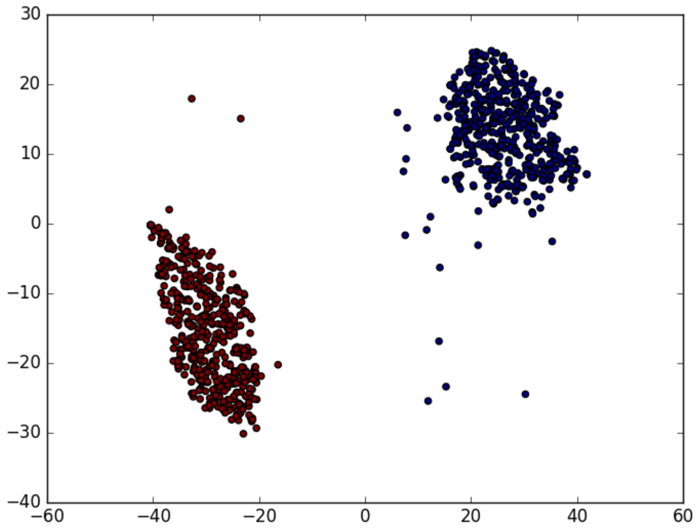
Dataset 1 with 8 clusters (k=8)



Dataset 2 with 4 clusters (k=4)



Dataset 3 with 2 clusters (k=2)



Dataset	Best K value	NMI Value
1	8	0.693056503504
2	4	0.90930682562
3	2	0.981421980408

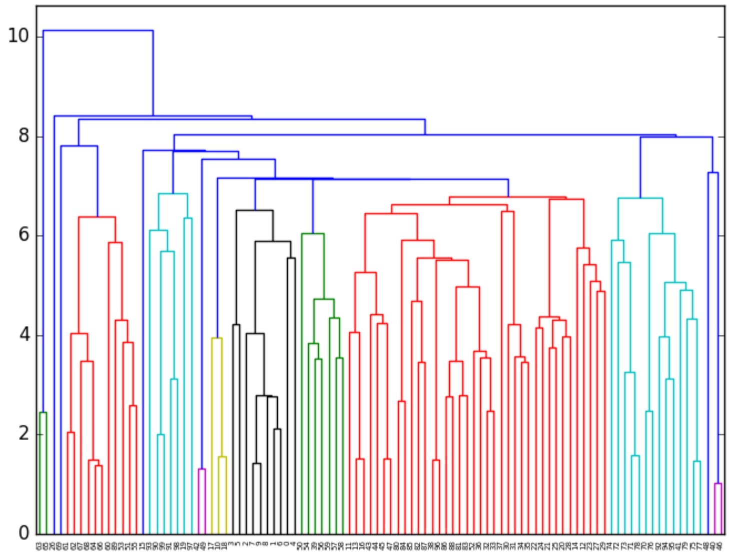
Following are the conclusions:

- We can see from the graph that, in dataset 1, we made 8 clusters, and they are nearby. In dataset 2, we made 4 clusters and they are farther than the distance between the 8 clusters in dataset 1. In the dataset 3, we made 2 clusters and they are very far.
- The same trend we can see in NMI values for the three dataset.
- NMI of dataset 3 > NMI of dataset 2 > NMI of dataset 1

C. Comparison to hierarchical clustering

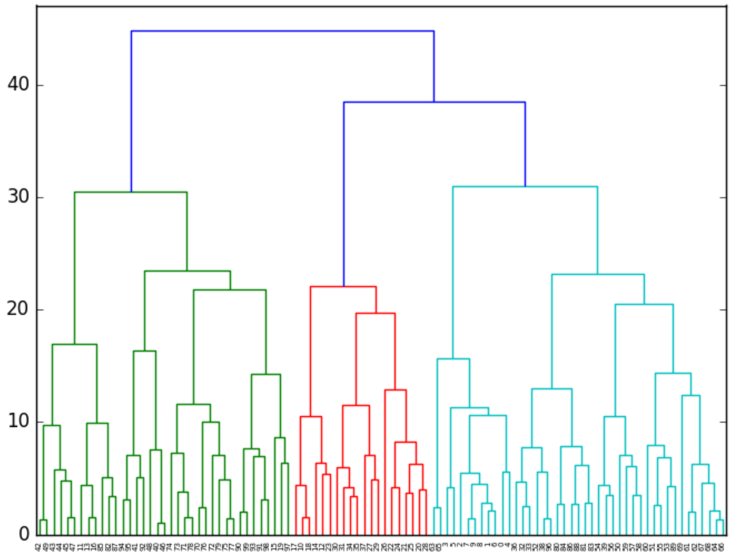
1. Create subsamples for each of the full dataset, by sampling 10 images at random from each digit group (i.e., 100 images). Use the scipy agglomerative clustering method to cluster the data using single linkage. Plot the dendrogram

Single Linkage Dendrogram

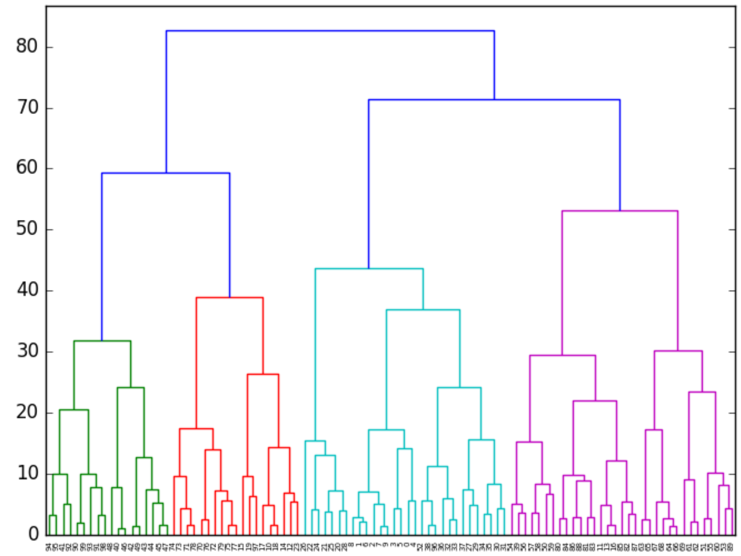


2. Cluster the data again, but this time using (i) complete linkage, and (ii) average linkage. Plot the associated dendrograms.

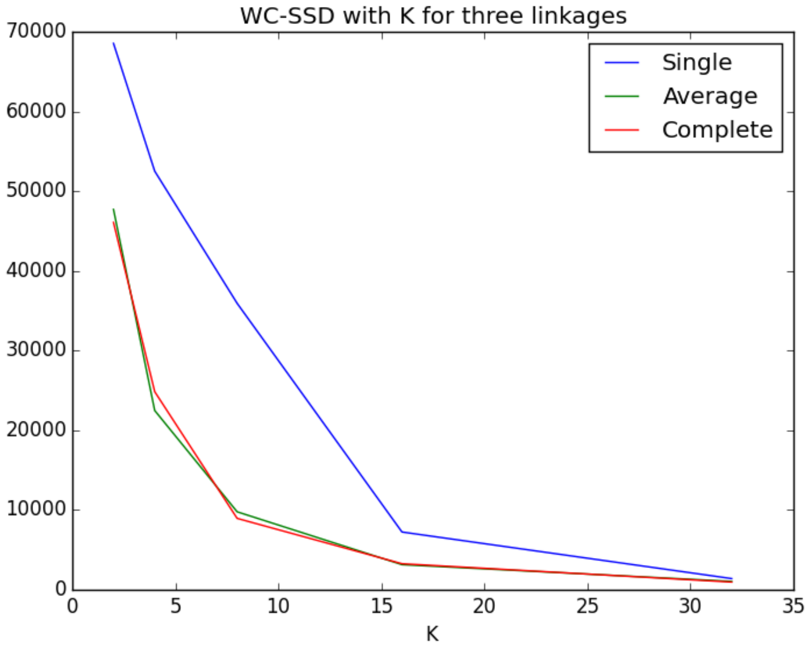
Average Linkage Dendrogram

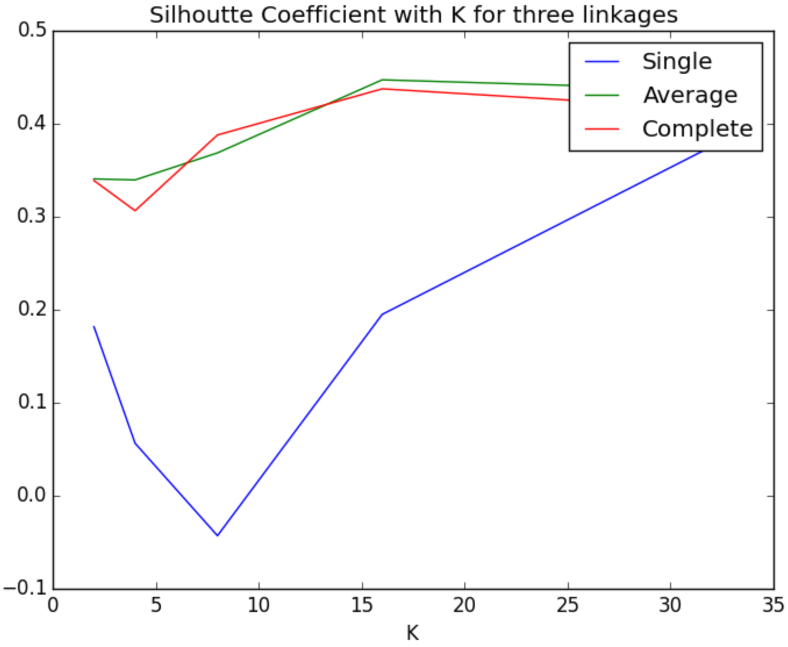


Complete Linkage Dendrogram



3. Consider cutting each of the dendrograms at successive levels of the hierarchy to produce partitions of different sizes (i.e., vary choice of K). Construct a plot showing the within-cluster sum of squared distances (WC SSD) and silhouette coefficient (SC) as a function of K.





4. Discuss what value you would choose for K and whether the results differ from your choice of K using k-means in part B.

   Following are the conclusions on the hierarchal clustering:

   1. We will select k= 32 for single linkage
   2. K = 16 for average linkage
   3. K = 16 for the complete linkage
   4. Yes, these choices differ as we selected k= 8 for the dataset comprising of all the digits using the k means
   5. The possible reasons we got higher values of k is that we took only 10 points from each digit. Those 10 points themselves would be quite far.
   6. So, in single linage, we find the minimum distance between 2 clusters and then we group them and since the points are far away as our dataset is very small, we got high value of K for single linkage
   7. Reason for average and complete linkage to get k=16 is that , average finds the total average distance and complete finds the maximum distance between two clusters. So even if the dataset is small, it won't affect much as it would to the single linkage.

5. For your choice of K (for each of single, complete, and average linkage), compute the NMI with respect to the image class labels. Discuss how the results compare across distance measures and how they compare to the results from k-means in part B

Linkage	Best K value	NMI Value
Single	32	0.7491
Average	16	0. 7681
Complete	16	0.7635

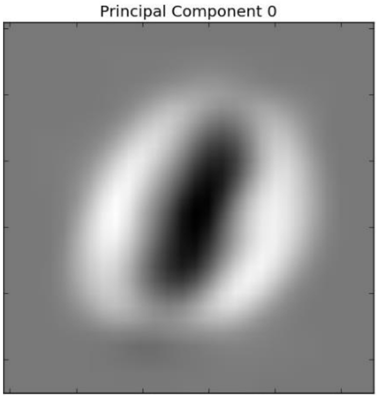
   Following are the conclusions:

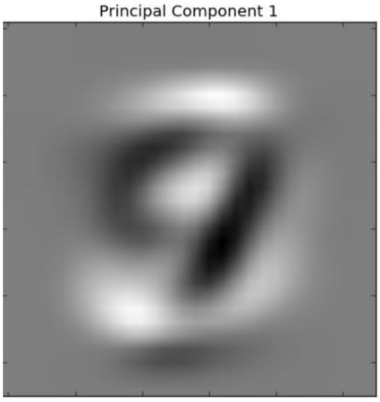
   - NMI values form Hierarchal clustering are higher than NMI values from K means

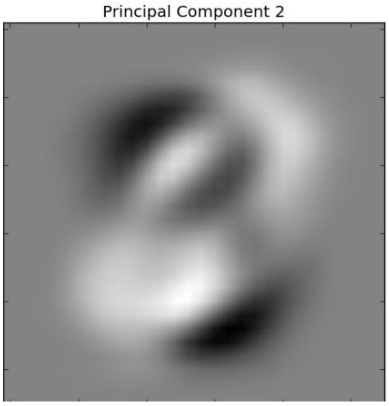
- The NMI value of average and complete linkage is nearly equal but both are higher than the single linkage
- Also, the SC values for Average and Complete Hierarchal clustering are higher than that of K means
- **So, we can say that, at least on small dataset, Hierarchal clustering is better than Kmeans**.
- Since we didn't apply Hierarchal clustering on Full dataset (20,000), we cannot say how will it perform on that.

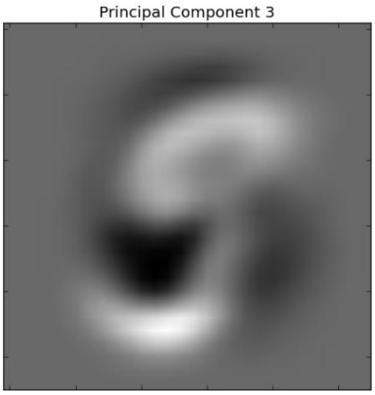
D. Bonus

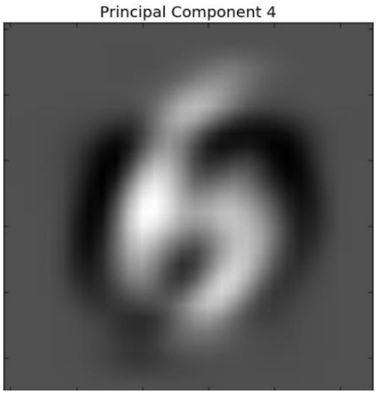
1. Implement PCA. Apply it to the digits-raw.csv to reduce the dimensionality of the digits data from 784 to 10.
2. For each of the 10 principal components, plot the eigenvectors (reshaped) as 28 × 28 grayscale matrices.

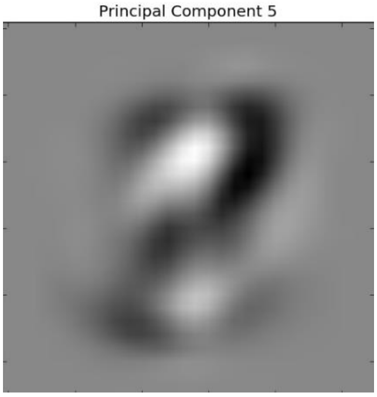


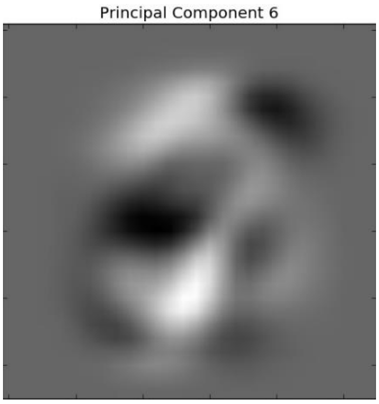


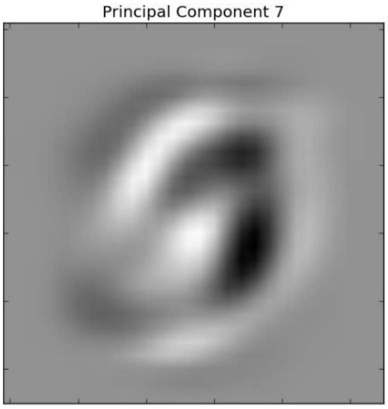


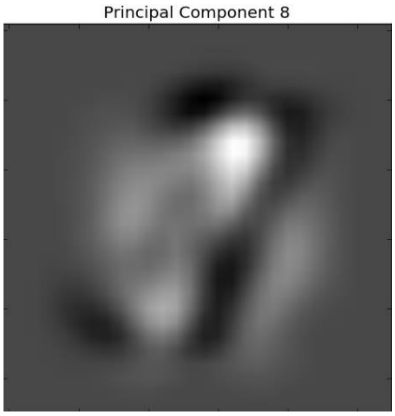


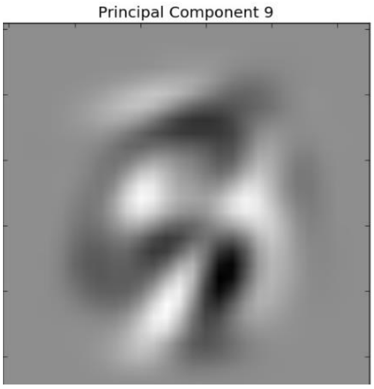






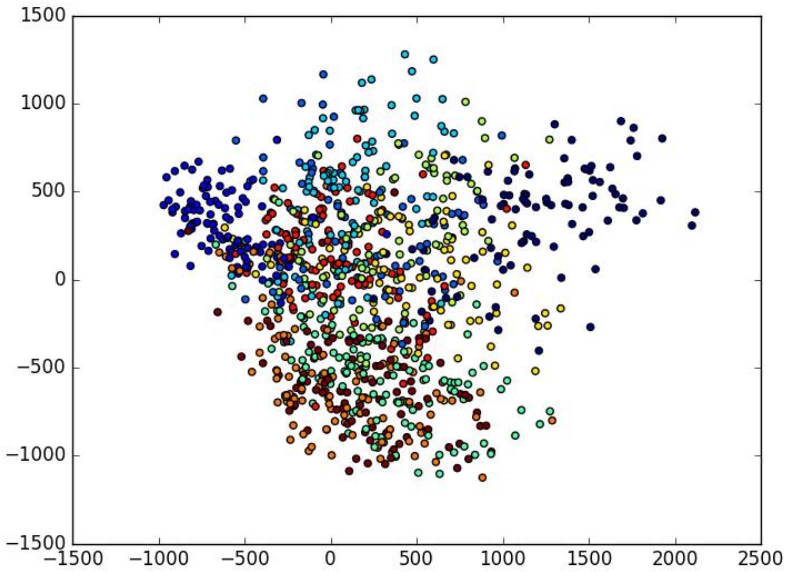






3. Visualize 1000 randomly selected examples using the first two principle components, coloring the points to show their corresponding class labels. Discuss how the results compare to the tSNE embedding.

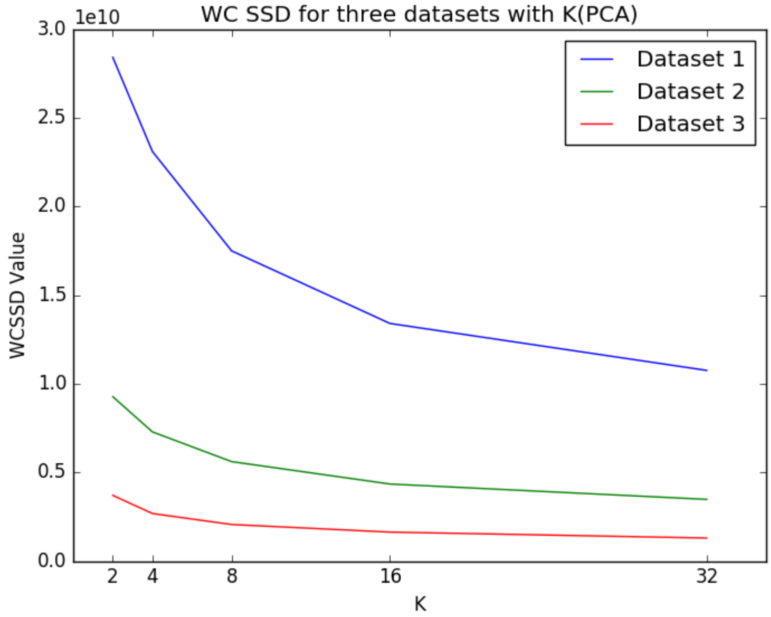
Visualization of the PCA 2d data with class labels (1000 points)

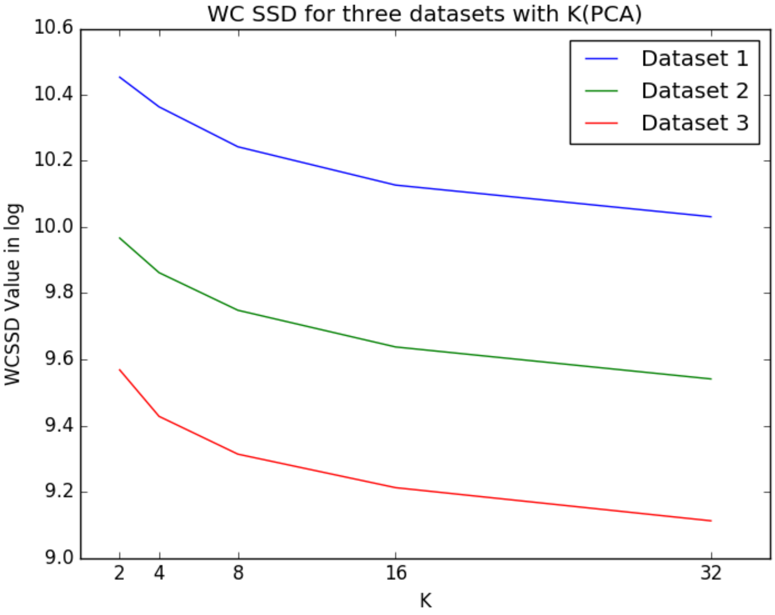


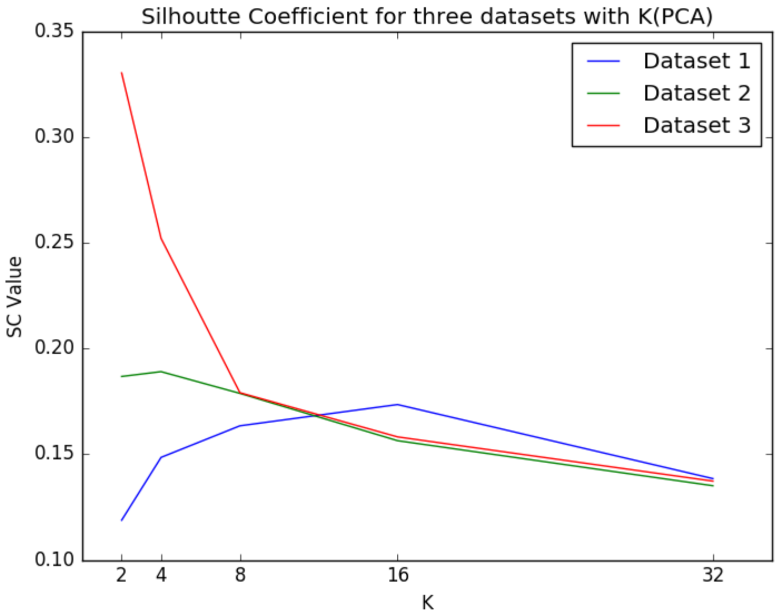
Following are the inferences from the PCA plot:

a. The clusters formed in PCA are very vague. They are spread all over the space.
b. TSNE does some extra calculation to preserve the cluster features and reduce it to two dimensions.
c. PCA on the other hand, just takes the most variant dimensions and reduces it accordingly

4. Using the PCA embedding of the data, repeat experiments B.1, B.2, and B.4. Discuss how the results compare to the clusters found with the Tsne embedding.
5. Repeat parts 1 and 4 using the same data subsets used above (i.e., first digits 2, 4, 6, 7, and then only digits 6, 7). Discuss how the results compare to what you found with Tsne.
   Answers of the part 4 and 5 are below:







Dataset 1 scores. Number of classes = 10

K	WCSSD (PCA)	SC Value(PCA)
2	28397264920.22	0.1187
4	23094758215.14	0.1484
8	17490151157.18	0.1633
16	13395255941.02	0.1734
32	10748293753.20	0.1384

Dataset 2 scores. Number of classes = 4

K	WCSSD(PCA)	SC Value(PCA)
2	9266146285.14	0.1867
4	7281136814.44	0.1889
8	5604404400.29	0.1787
16	4341121484.21	0.1563
32	3474731549.66	0.1350

Dataset 3 scores. Number of classes = 2

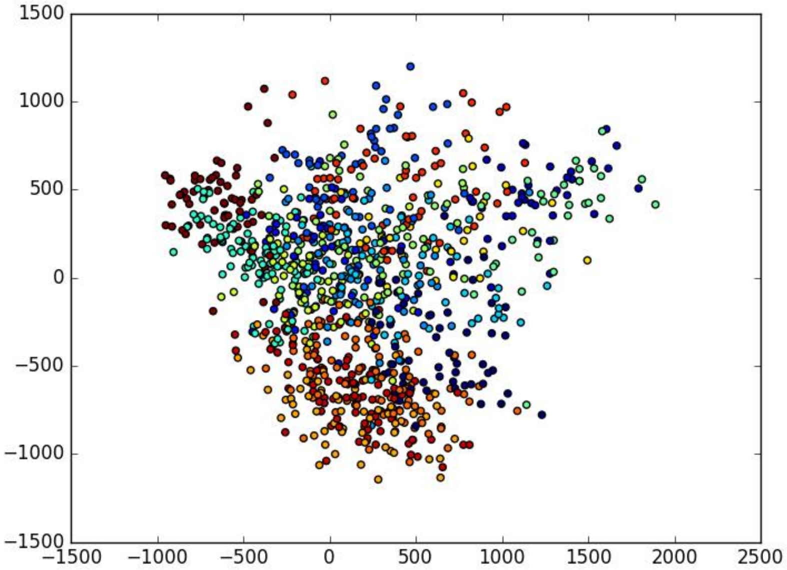
K	WCSSD(PCA)	SC Value(PCA)
2	3701815311.70	0.3302
4	2677681927.15	0.2519
8	2058184847.94	0.1790
16	1630077419.47	0.1581
32	1294259043.13	0.1372

Following are the conclusions on the PCA set:

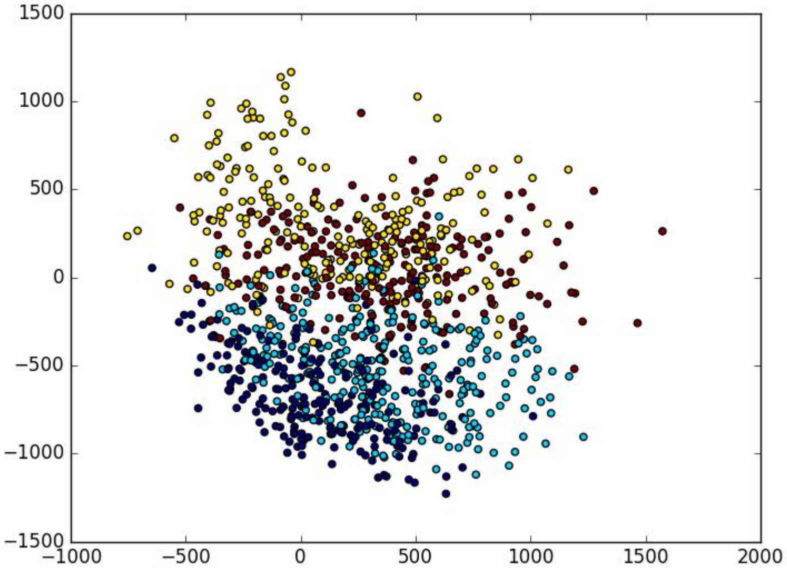
1. The WCSSD values are very high as compared to the Tsne. The reason is, we took 10 dimensions for PCA as compared to 2 dimensions of the Tsne.
2. The silhouette coefficients for the PCA are far less than the ones for Tsne. The reason is, Tsne had preserved the well-defined clusters and correspondingly kmeans clusters them well.
3. PCA dimensional reduction could not preserve that information as much as Tsne did and so the values are far less
4. Best value of K for PCA are (Based on SC score) :
    - Dataset 1 (Number of classes = 10): **k = 16** as compared to (k=8) for Tsne
    - Dataset 2 (Number of classes = 4): **K = 4**
    - Dataset 3 (Number of classes = 2): **k = 2**
5. For the dataset 2 and dataset 3, we got the same k value, only on dataset 1, we got k=16 for pca.

NMI Values and Visualization

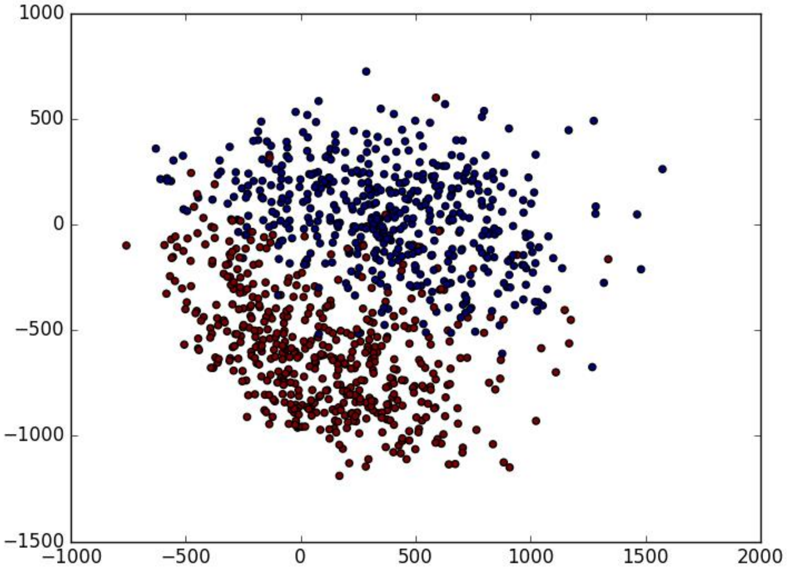
Dataset 1 with 16 clusters (k=16)



Dataset 2 with 4 clusters (k=4)



Dataset 3 with 2 clusters (k=2)



Dataset	Best K value	NMI Value
1	16	0.508812279478
2	4	0.642499461336
3	2	0.910772736134

Following are the conclusions:

1. The NMI values for PCA are also much less as compared to the values from Tsne.
2. **We conclude that Tsne is a better algorithm for dimensional reduction**