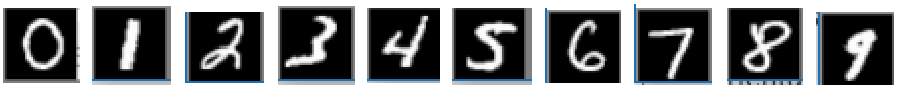
CS57300: Homework 5

A. Exploration (5 pts)

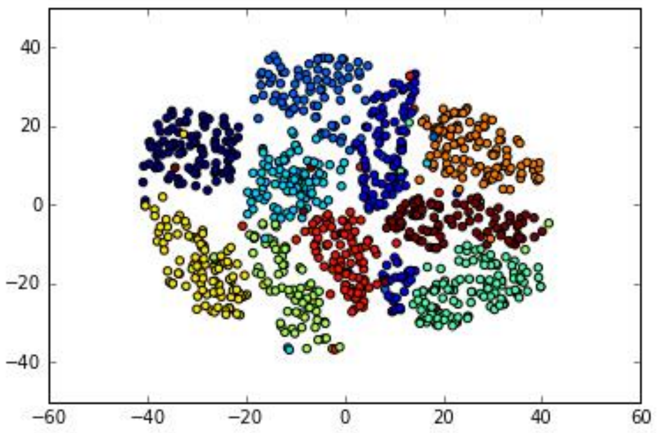
1. **Randomly pick one digit from each class in digits-raw.csv and visualize its image as a 28×28 grayscale matrix.**

Ans.



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

Ans.



B. Analysis of k-means (20 pts)

Consider three versions of the data for each of the questions below:

(i) use the full dataset digits-embedding.csv – **[A]**

(ii) use only the subset of the data consisting of the digits 2, 4, 6 and 7 – **[B]**

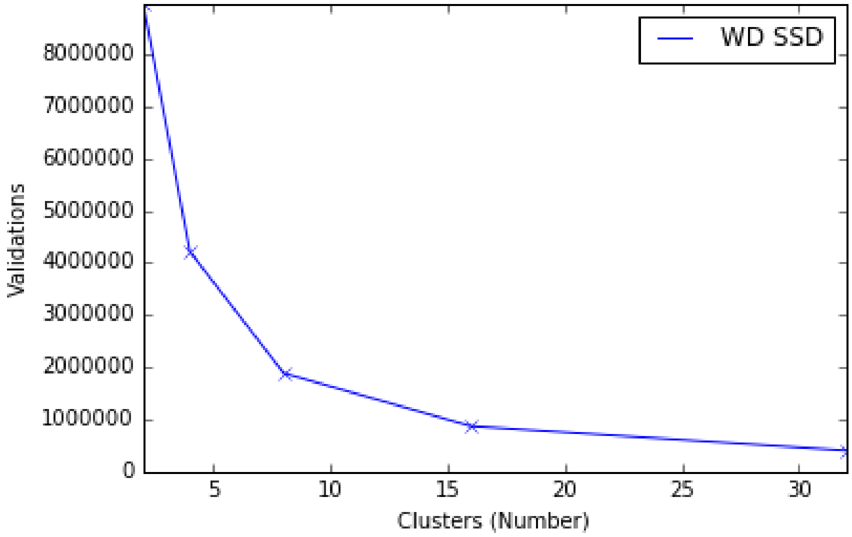
(iii) use only the subset of the data consisting of the digits 6 and 7 – **[C]**

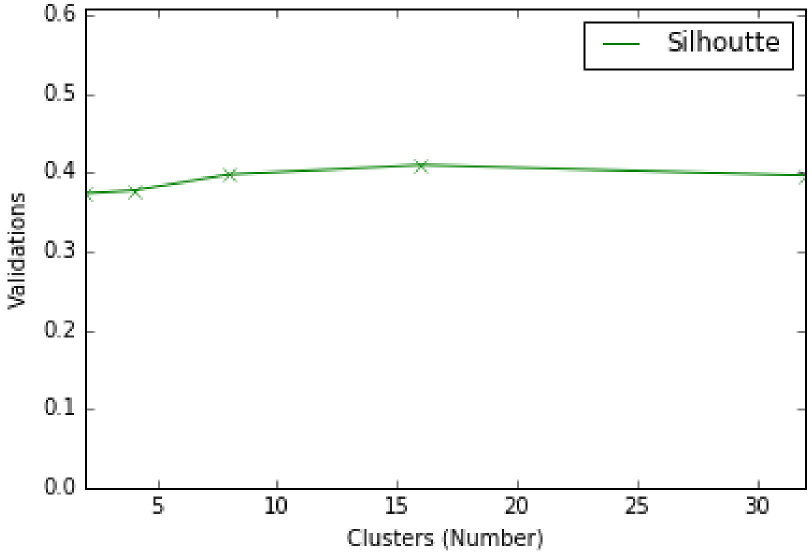
********Note the coding [A], [B], and [C]********

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*.**

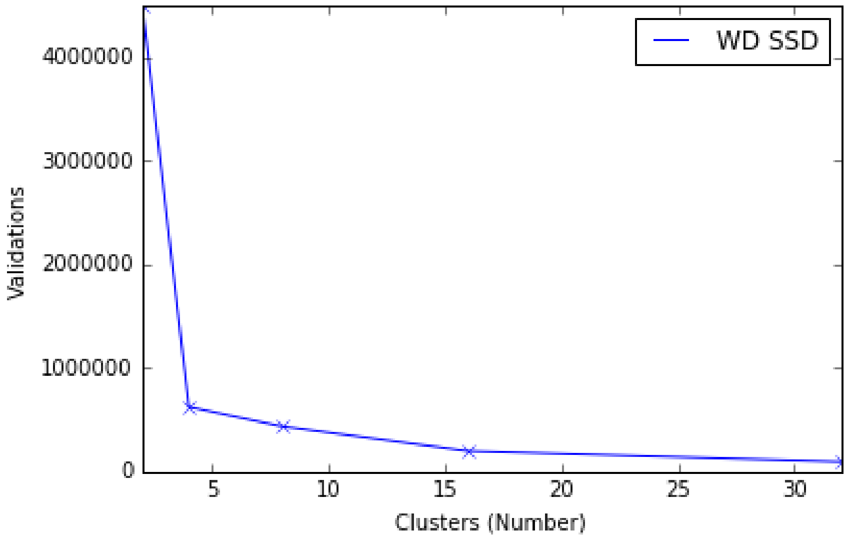
Ans.

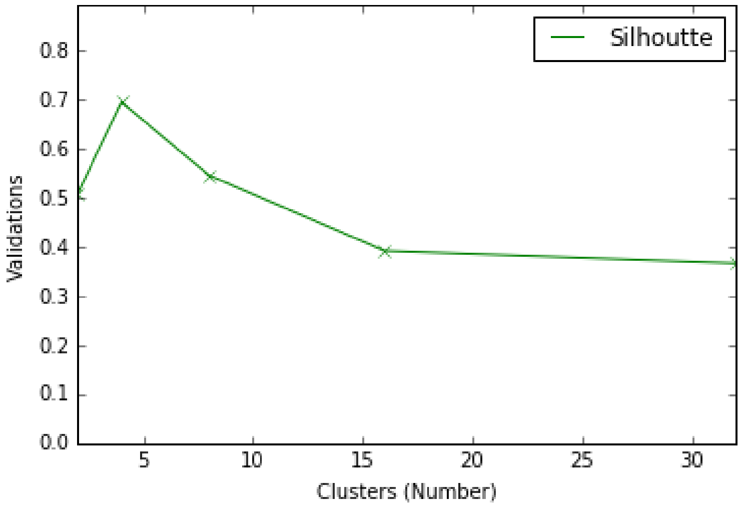
[A]



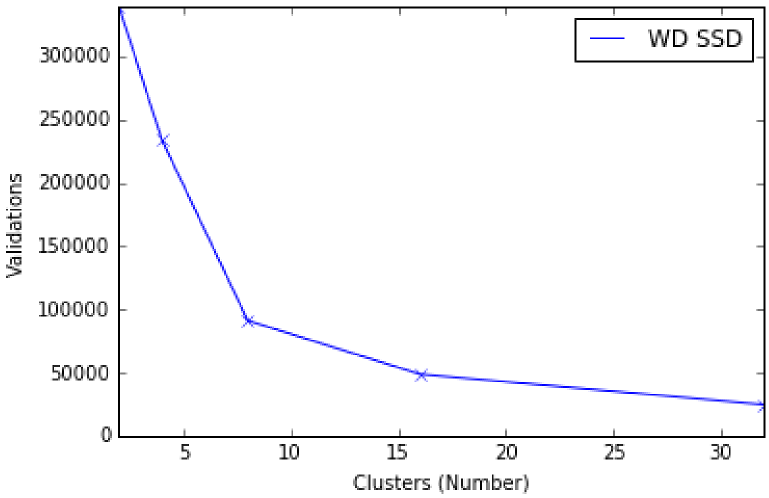


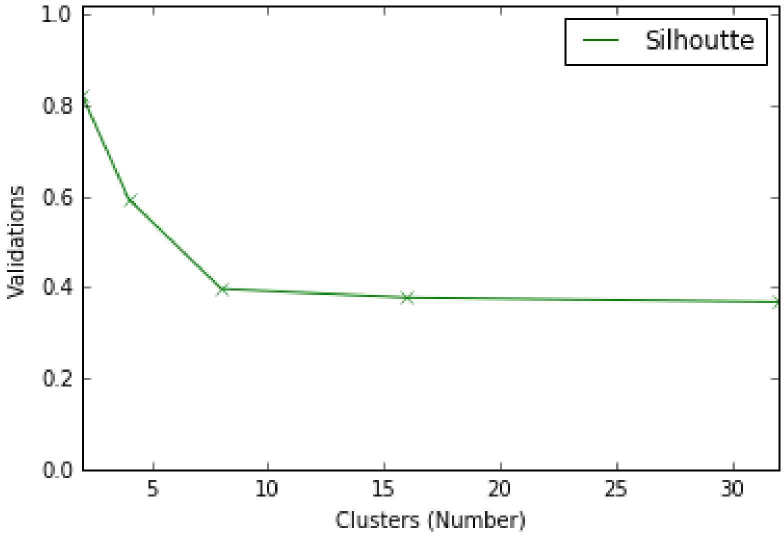
[B]


WD SSD
Validations
4000000
3000000
2000000
1000000
0
5
10
15
20
25
30
Clusters (Number)


Silhoutte
Validations
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
5
10
15
20
25
30
Clusters (Number)

[C]


WD SSD
Validations
300000
250000
200000
150000
100000
50000
0
5
10
15
20
25
30
Clusters (Number)



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.**

Ans.

[A] 16 Clusters seems to be the best K value.

We can conclude this as the maximum Silhouette Coefficient is at the value of 16.

The WD_SSD also show a sudden drop for this value of K.

This follows intuition as 2,4 and 8 clusters would be too short to be able to correctly classify 10 different classes. 16 would be over the top, but it would be better to have more labels than less.

[B] 4 Clusters seems the be the best K value.

We can conclude this per the maximum Silhouette Coefficient value at 4 clusters.

The WD_SSD also show a sudden drop for this value of K.

We notice that there are only 4 data class labels in the data, and obviously 4 clusters would be the best match for 4 class labels intuitively.

[C] 2 Clusters seems to be the best K value.

The maximum Silhouette Coefficient value is at 2 K value.

The WD_SSD also show a sudden drop for this value of K.

Intuitively, we know that 2 clusters would be the best fit for 2 class label data.

Overall, we notice that with the maximum Silhouette Coefficient, and the sudden drop in the WD SSD, we can sense the optimality of a model. We see that, for each of the data sets, the

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

Ans.

[A]

Means:

K Value	WC SSD	Silhouette Coefficient
2	8982316.73929	0.373755983844
4	4265862.28185	0.375159648086
8	1902025.28495	0.406787244531
16	869294.54744	0.403820289995
32	419283.685425	0.389117669444

Variances:

K Value	WC SSD	Silhouette Coefficient
2	122764.896251	3.06502945144e-09
4	2658249643.96	4.82387062041e-06
8	196267148.053	4.8330496684e-05
16	177085289.234	3.06805057732e-05
32	440287912.906	5.41101625825e-05

From the variance for the full data set of class values from 0 to 9, we see that the variance of the WC SSD is quite low at a value of k=2 compared to the other K values. This might be because with such a small number of clusters, there is more efficient clustering compared to others since there are only a binary number of clusters we can make. But, we also notice that the variance decreases with increase in K. This makes sense since with increases in the number of clusters, the clusters should be more compact, leading to lower means and lower variance of the WC SSD.

For the Silhouette Coefficient, we notice that the maximum value of mean is at the most optimal k value, as we have discussed earlier in the report. The variance decreases with increase in the value of K.

[B]

Means:

K Value	WC SSD	Silhouette Coefficient
2	4469406.68284	0.479832052764
4	840471.138657	0.66109099033
8	403820.752104	0.520468849056
16	188536.087275	0.415377559145
32	89129.3717474	0.37576923207

Variances:

K Value	WC SSD	Silhouette Coefficient
2	116923547139.0	0.000598052473942
4	188214872781.0	0.00477700260884
8	2637674840.04	0.00158146248015
16	1263729411.05	0.000399686877553
32	19939115.57	3.32556740998e-05

Here, we see that the WC SSD's variance decreases monotonically with increase in the value of K. This is because as we increase the number of clusters, since the data points are fixed, automatically the clusters are closer together, resulting in a direct decrease in the WC SSD.

As per the Silhouette Coefficient's variance, we notice that it first increases and then decreases. The Silhouette Coefficient is most stable at the value of k=32. This is because at a K value of 32, there is not much space for improvement.

[C]

Means:

K Value	WC SSD	Silhouette Coefficient
2	340179.993565	0.821887695158
4	215338.893144	0.57373372282
8	99377.9571163	0.393232966306
16	48861.6836739	0.374562061405
32	25690.7261231	0.364435239983

Variances:

K Value	WC SSD	Silhouette Coefficient
2	11986.4560498	1.11650961315e-09
4	584164252.647	0.00434829579407
8	98062722.9338	8.12599300766e-05
16	7362227.29257	5.59623381445e-05
32	753890.967877	3.36053158944e-05

We see that the variance of the WD SSD seems to increase at first and then decrease rapidly with increase in the value of K. The variance for Silhouette Coefficient increases with increase in the value of K.

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, coloring 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.**

Ans.

[A] 0.74731732541735041 – is the NMI obtained for the entire dataset, with k=16

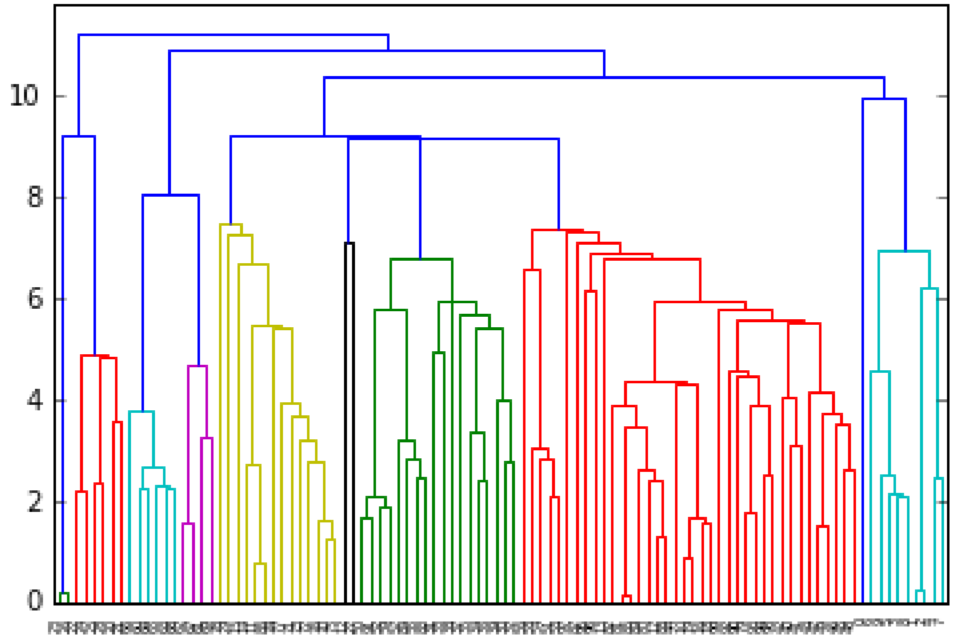
[B] 0.94890858479722984 - is the NMI obtained for the [2,4,6,7] dataset, with k=4

[C] 0.9653947541474529 - is the NMI obtained for the [6,7] dataset, with k=2

C. Comparison to hierarchical clustering (15 pts)

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.**

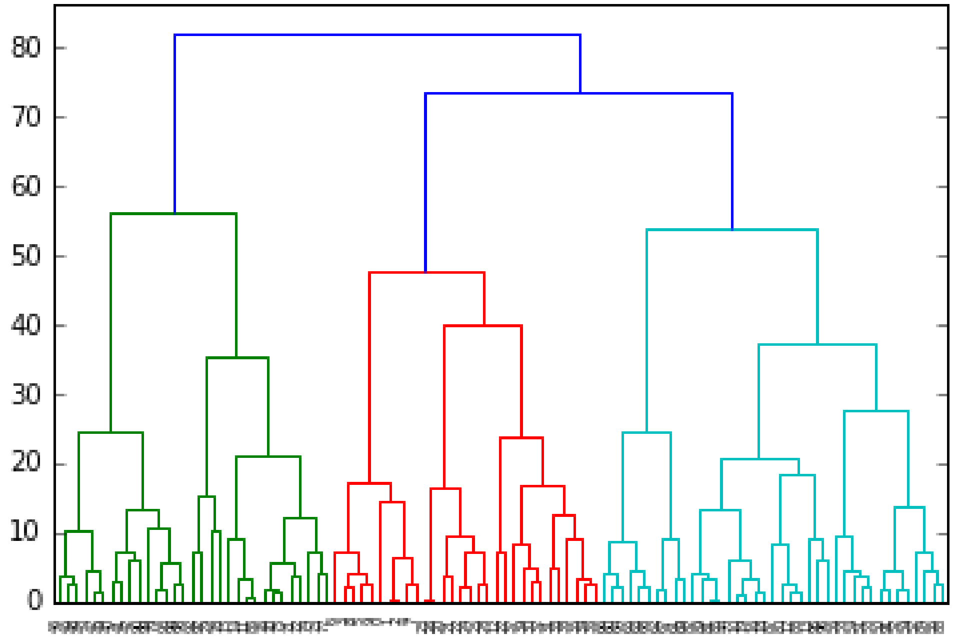
Ans.



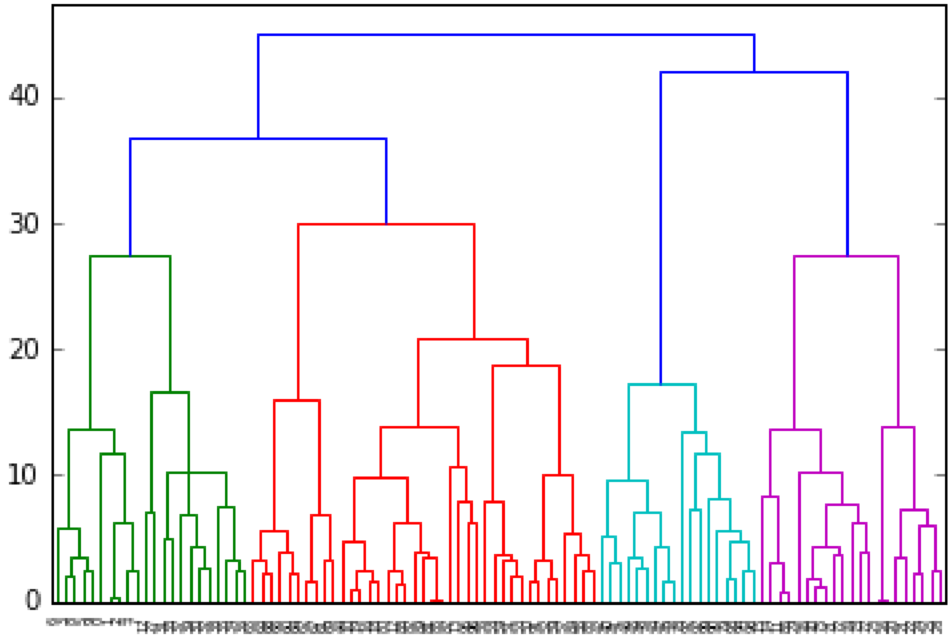
2. **Cluster the data again, but this time using (i) complete linkage, and (ii) average linkage. Plot the associated dendrograms.**

Ans.

(i)

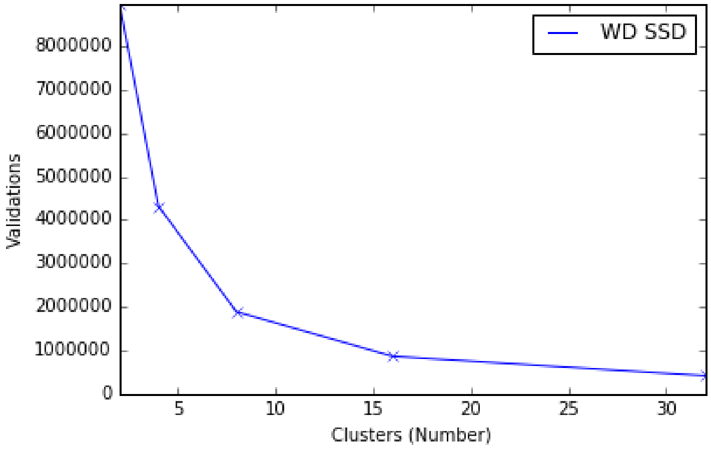


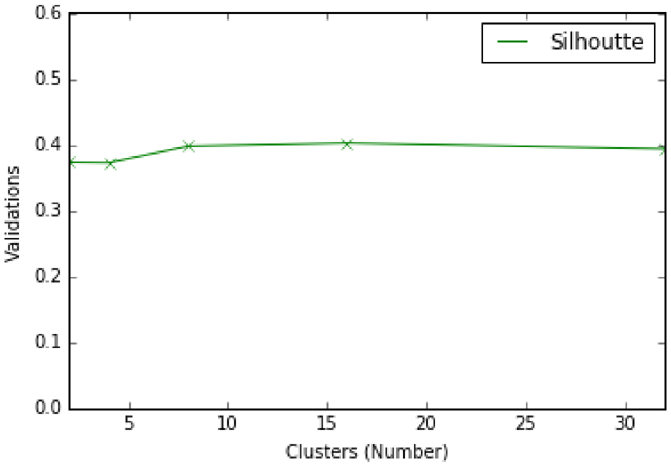
(ii)



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*.**

Ans.





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.**

Ans. We choose a K value of 8, since the Silhouette Coefficient for this value leads to be the maximum, indicating a most optimal scenario. For the case of K-means, we had chosen the value of K as 16, since there are more number of clusters than class labels as compared to a value of K of 8.

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.**

Ans.

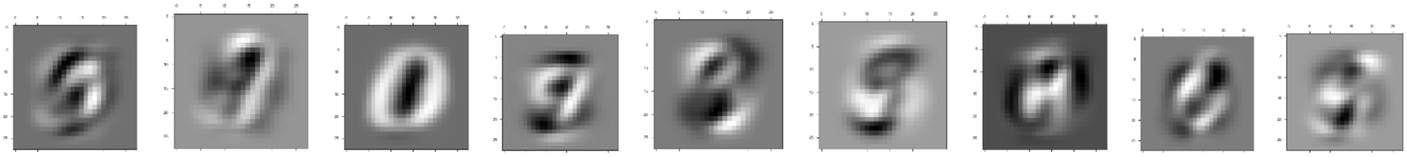
The value of NMI for the value of K = 8 is 0.608945393037.

Bonus (15 pts)

1. **Implement PCA. Apply it to the digits-raw.csv to reduce the dimensionality of the digits data from 784 to 10.**

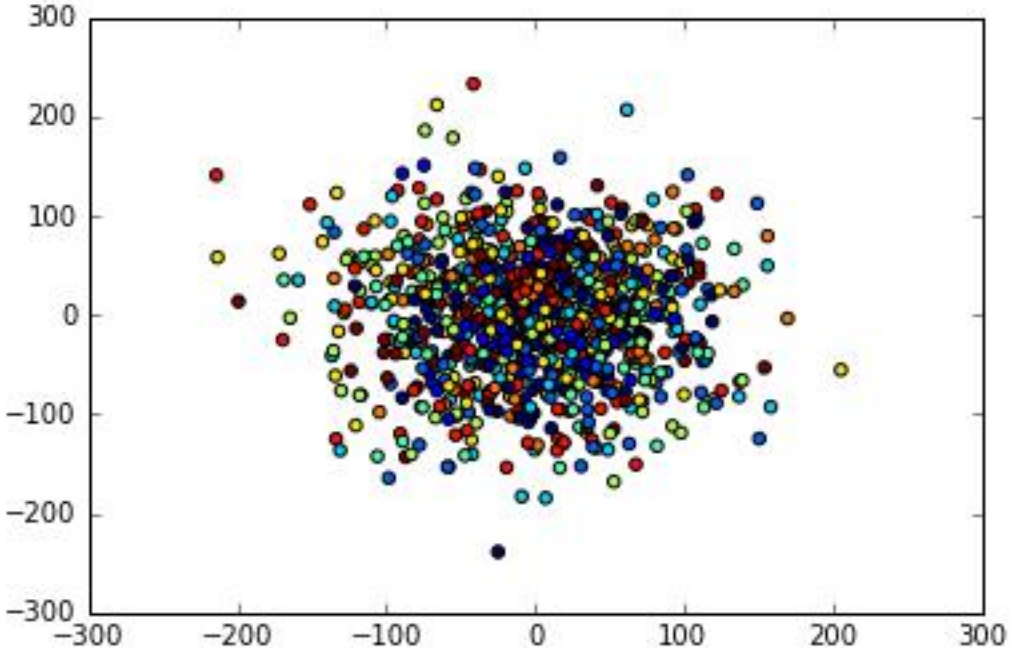
2. **For each of the 10 principal components, plot the eigenvectors (reshaped) as 28 × 28 grayscale matrices.**

Ans.



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.**

Ans.



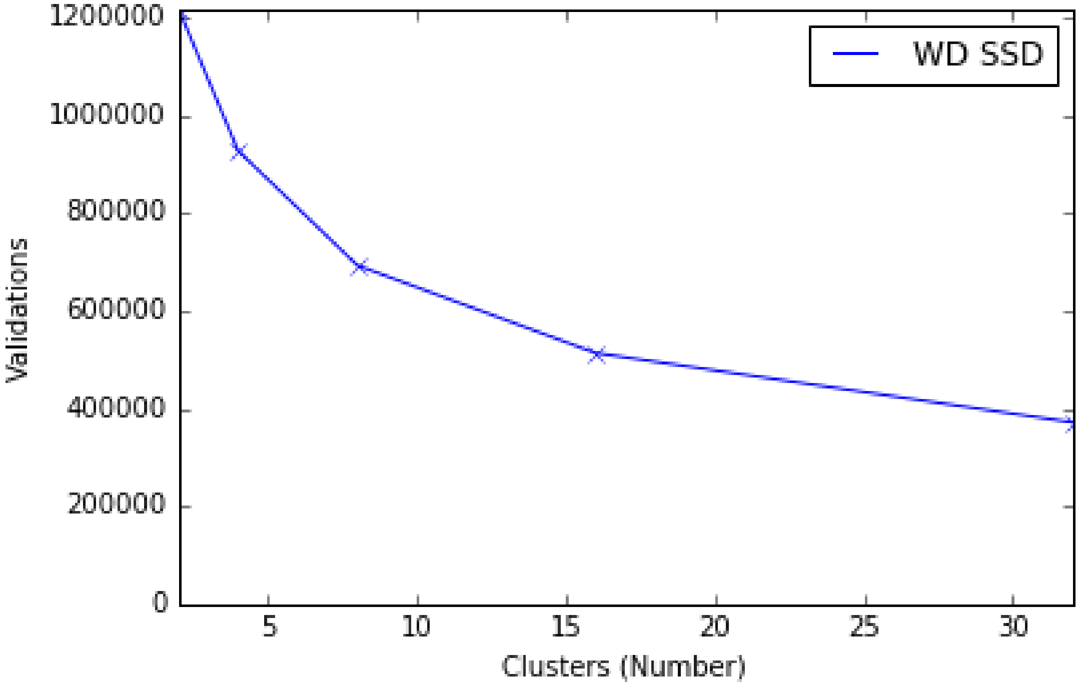
We see that all the points have been normalized and hence follow a trend as shown above. The PCA resulted in all the class labels being distributed around their own centers, and these centers coinciding at a global center of (0,0). Due to this, we hypothesize that we would not be able to improve the clustering optimality by changing the number of clusters.

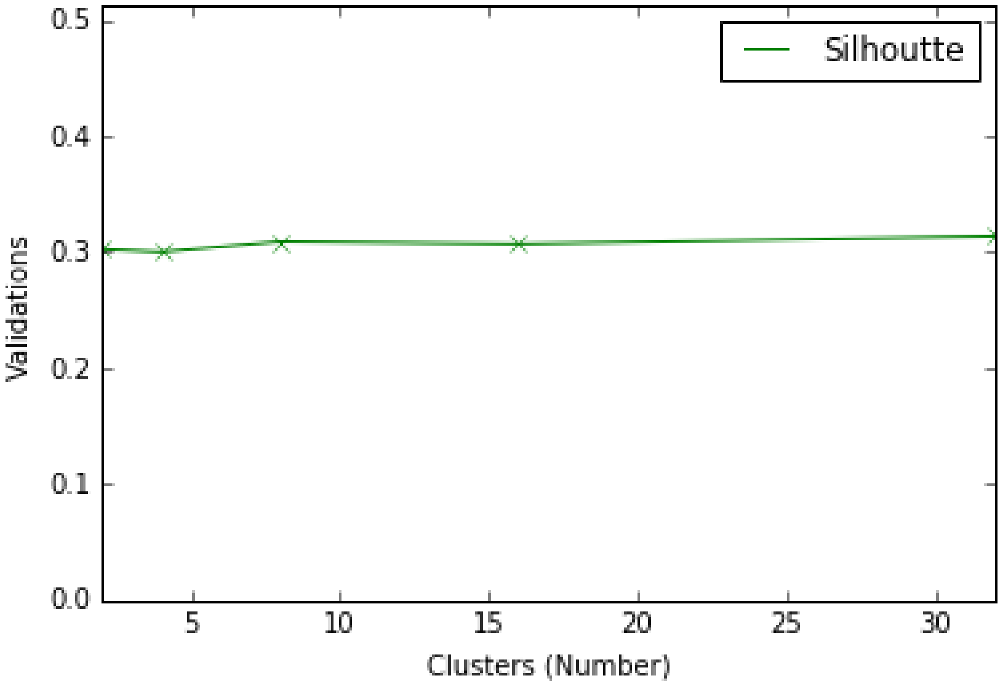
When compared to the tSNE embedding, we notice that this embedding results in an almost symmetric distribution around the (0,0) value, thereby making it much harder to note the different class labels. This may be because due to the dimensionality reduction of such massive scale, (from 784 to 2), there is a large loss of information, resulting in the above distribution.

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.**

Ans.

[A]





K Value	WC SSD	Silhouette Coefficient	NML
2	1.219022726790470537e+06	0.302710068691	0.000854344744456
4	9.272784470429604407e+05	0.305595350284	0.00626215163868
8	6.939793799195582978e+05	0.308634302424	0.0140309421158
16	5.129802522103801602e+05	0.309605152408	0.0178875620681
32	3.733914758842986776e+05	0.314868640744	0.019476978752

Choosing Best Value of K:

From previous analysis, we can see that the optimal number of clusters should be at the maximum value of the Silhouette Coefficient, making the optimal value of K to be 32. The WC SSD is at the lowest value at this position.

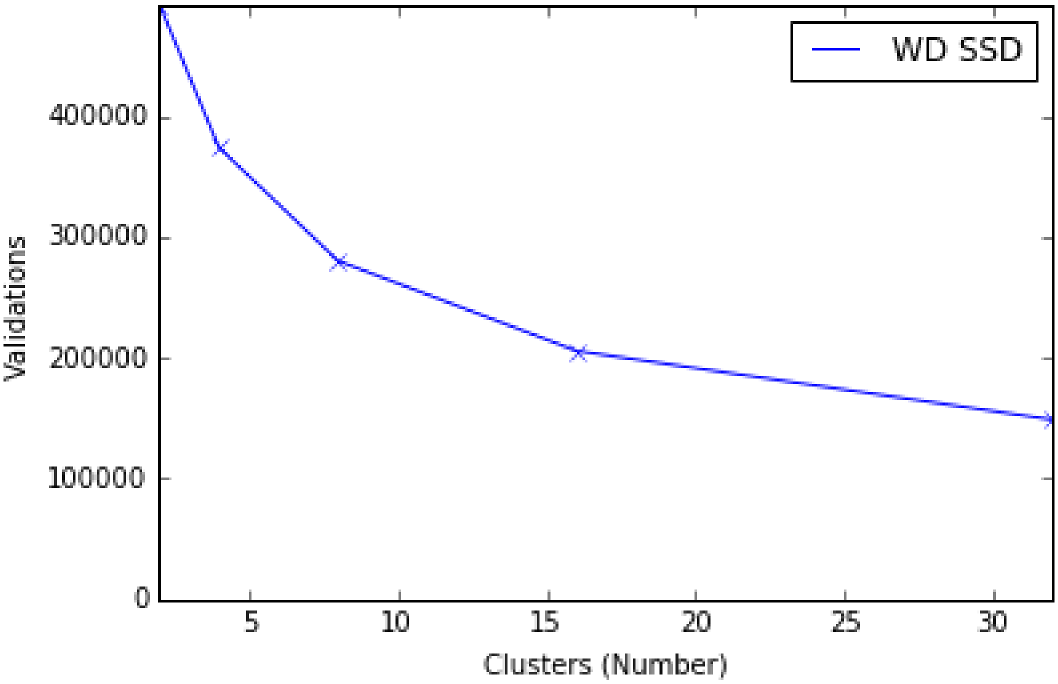
For the optimal K value of 32, the NML is 0.194769.

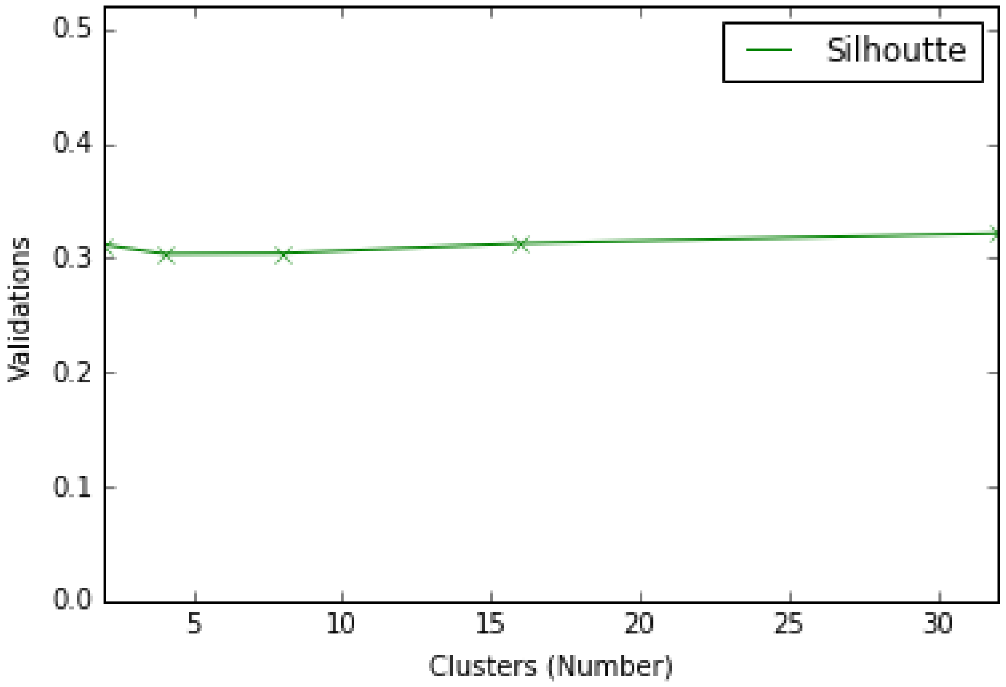
Comparison: In the tSNE embedding, the best value of K for this data set was 16. That was because the data had 10 class labels, and at least 10 were required to get the optimal clustering. Though 10 would have been optimal, since that was not available, 16 was taken to be the optimal value of K.

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.**

Ans.

[B]





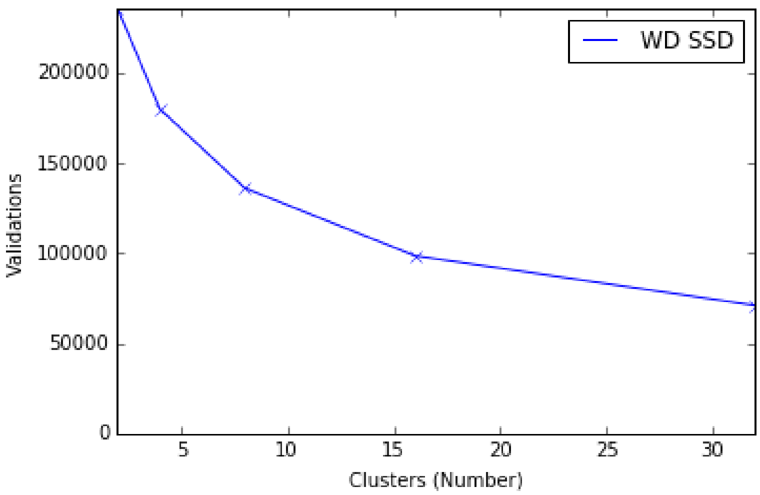
K Value	WC SSD	Silhouette Coefficient	NML
2	4.927158318893964170e+05	0.311258682715	0.803572089636
4	3.703635702950813575e+05	0.305507132282	0.671477072074
8	2.791919715672929888e+05	0.306250019017	0.579310118068
16	2.058585106102078571e+05	0.318749169929	0.510108686099
32	1.492964758105781511e+05	0.315130488891	0.454406284408

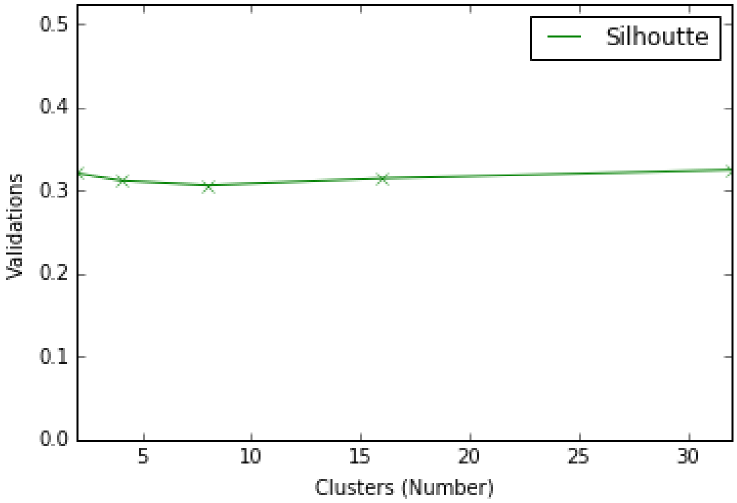
As done above, the maximum value of Silhouette Coefficient is obtained at a K value of 16. This indicates that K=16 tends to be the optimal number of clusters. The WC SSD is the least here. This confirms that 16 is indeed the optimal number of clusters for this dataset done by PCA.

The NML value for the optimal model with K = 16 is 0.51010.

Comparison: In the tSNE embedding, the best value of K for this data set was 4. That was because the data had 4 class labels, and 4 clusters were required to get the optimal clustering.

[C]





K Value	WC SSD	Silhouette Coefficient	NML
2	2.360103444751697825e+05	0.320828314875	0.965394754147
4	1.793637117247006972e+05	0.310504372023	0.947178175103
8	1.368118048936901032e+05	0.301652546452	0.928308610123
16	9.793246781543995894e+04	0.317919485093	0.908750308348
32	7.065070022350146610e+04	0.327751491714	0.888464860602

As done above, the maximum value of Silhouette Coefficient is obtained at a K value of 32. This indicates that K=16 tends to be the optimal number of clusters. The WC SSD is the least here. This confirms that 16 is indeed the optimal number of clusters for this dataset done by PCA.

The NML with the optimal value of K, at K=32 is 0.8884.

Comparison: In the tSNE embedding, the best value of K for this data set was 2. That was because the data had 4 class labels, and 4 clusters were required to get the optimal clustering.