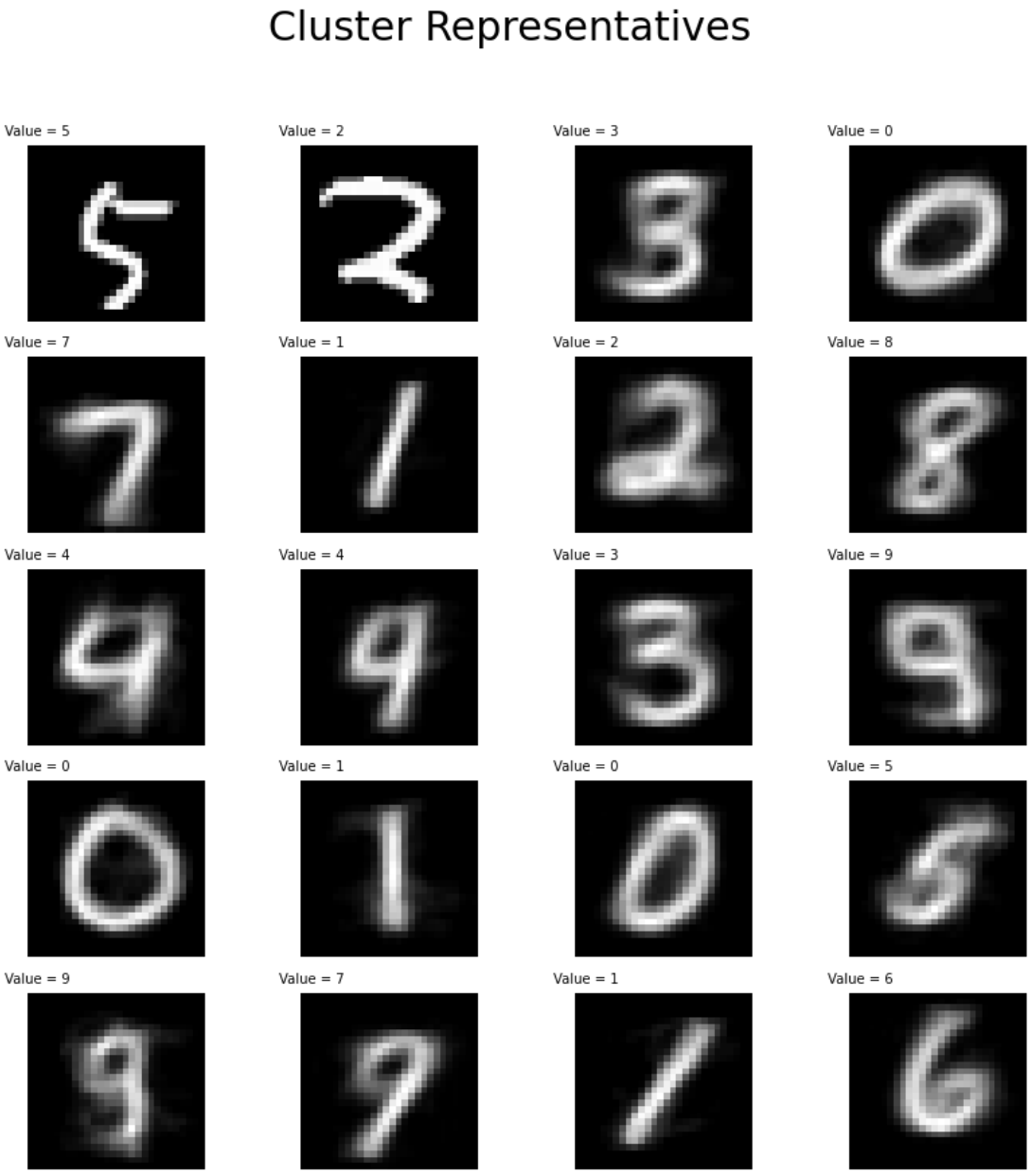
IMAGE CLUSTERING

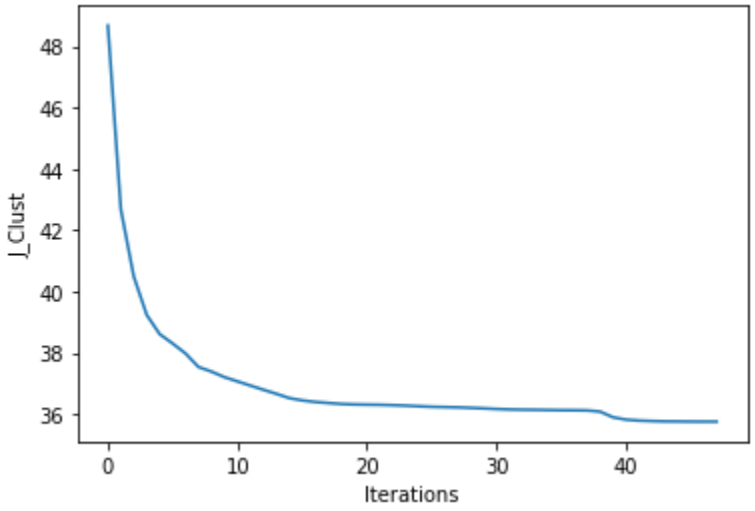
Link to Colab notebook - https://colab.research.google.com/drive/1XWU-TTOYKLKdrbXDDFAlRwvgmXUTAsJT?usp=sharing
As we are selecting 100 images of each digit we have 1000 training samples in all, so N=1000. Each image is 28*28 pixels which is 784 features. Thus n=784. Convergence criteria is that successive J_Clust values should differ by less than 1e-6.

Case (i) Random initialization of cluster representatives

a. Converged after 48 iterations. Cluster representatives -



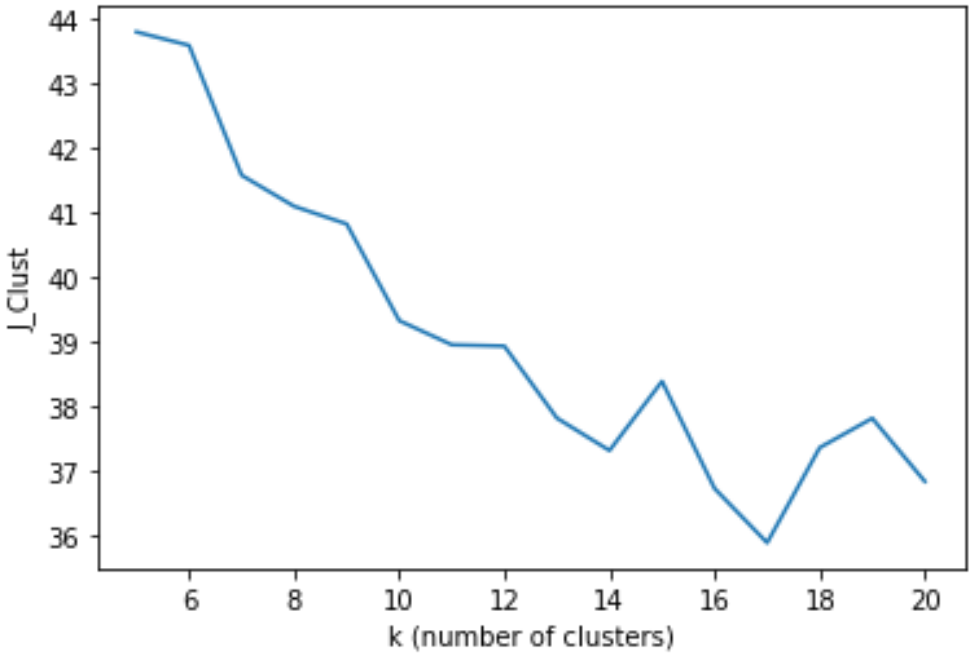
b. Variation of J_Clust with iterations -



c. Accuracy is 60%

d. Variation of J_Clust with k (number of clusters)

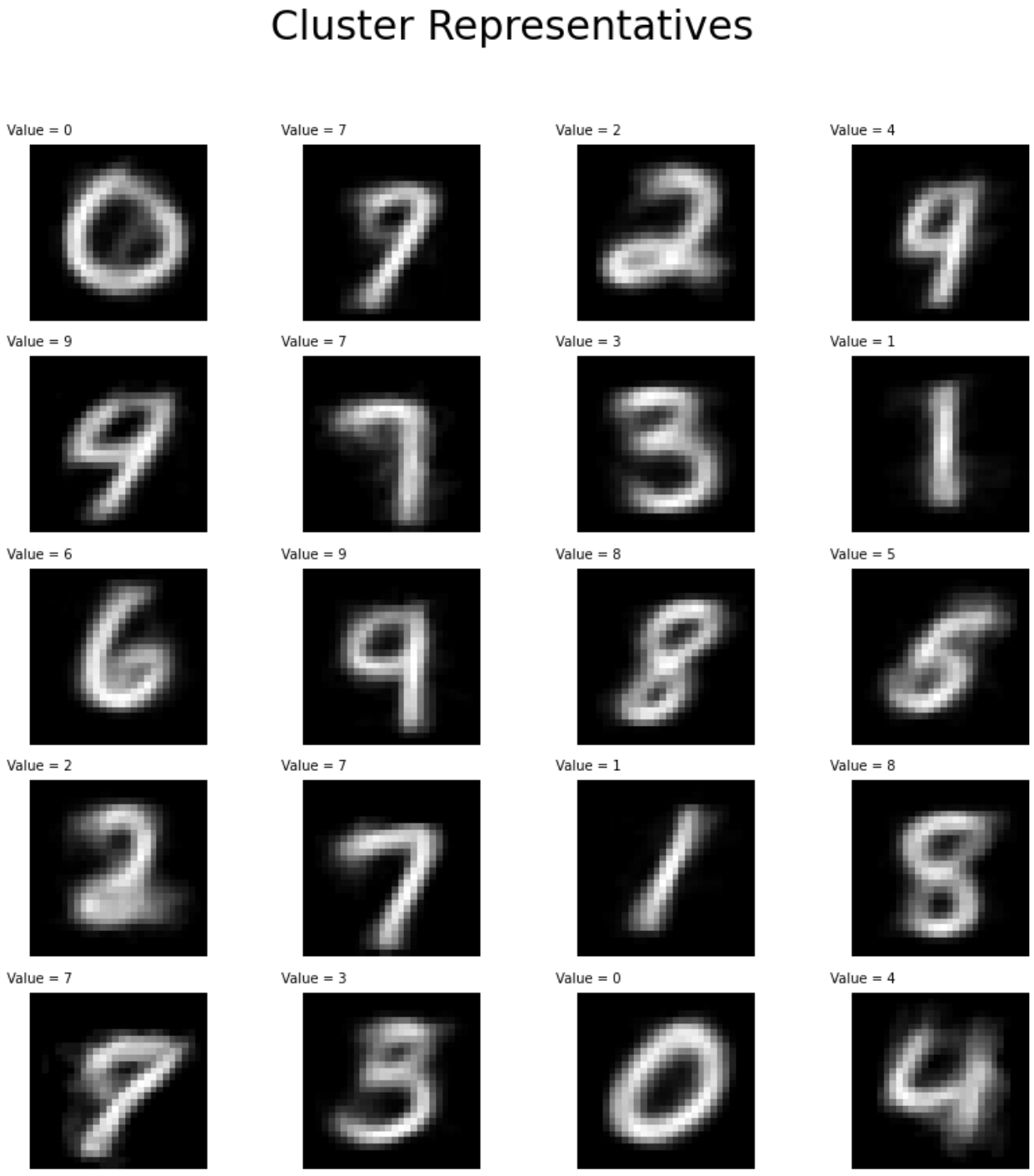
k = 5, J_clust = 43.802450387941015
k = 6, J_clust = 43.595079886232554
k = 7, J_clust = 41.58166536549617
k = 8, J_clust = 41.099918447969294
k = 9, J_clust = 40.82609244129326
k = 10, J_clust = 39.3276166827809
k = 11, J_clust = 38.95324663181483
k = 12, J_clust = 38.9330958354011
k = 13, J_clust = 37.81671033540342
k = 14, J_clust = 37.31486250752685
k = 15, J_clust = 38.386742913843705
k = 16, J_clust = 36.72514555912689
k = 17, J_clust = 35.88472834778296
k = 18, J_clust = 37.35899180521812
k = 19, J_clust = 37.8140183208526
k = 20, J_clust = 36.83572862223984



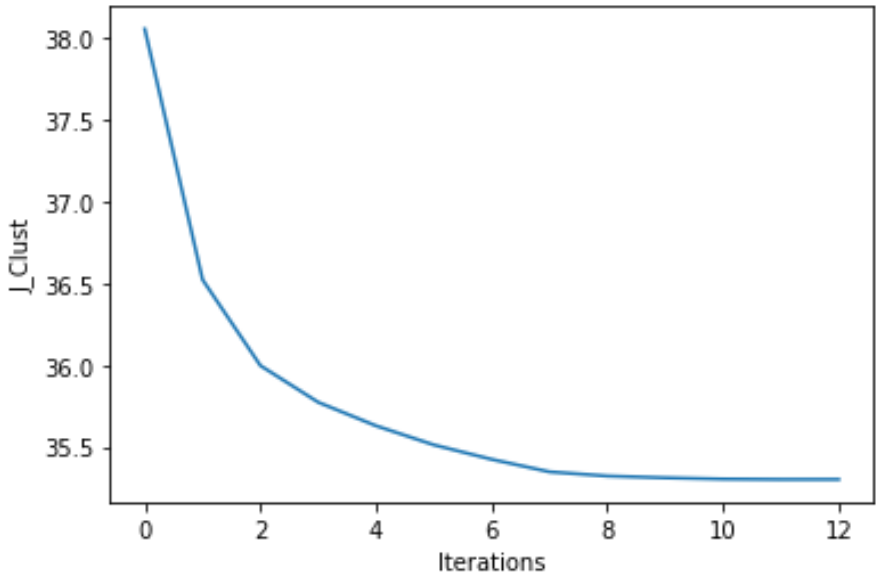
Min J_clust = 35.88472834778296 for k = 17

Case (ii) Choose cluster representatives from the given data set

a. Converged after 13 iterations. Cluster representatives -



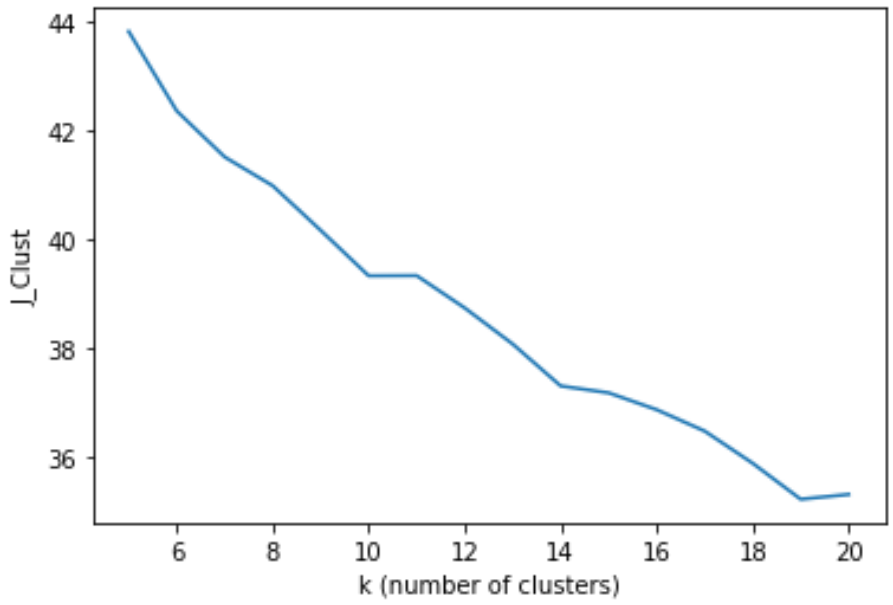
b. Variation of J_Clust with iterations -



c. Accuracy is 74%

d. Variation of J_Clust with k (number of clusters)

k = 5, J_clust = 43.808204561431744
k = 6, J_clust = 42.348363960656634
k = 7, J_clust = 41.5011426371537
k = 8, J_clust = 40.9732311343331
k = 9, J_clust = 40.153366224455766
k = 10, J_clust = 39.31797254375949
k = 11, J_clust = 39.321498187212576
k = 12, J_clust = 38.72795929132935
k = 13, J_clust = 38.06439649526521
k = 14, J_clust = 37.29399198467167
k = 15, J_clust = 37.16765845754992
k = 16, J_clust = 36.85785489720997
k = 17, J_clust = 36.464863233677114
k = 18, J_clust = 35.87367478202825
k = 19, J_clust = 35.21304702456804
k = 20, J_clust = 35.30053014463548



Min J_clust = 35.21304702456804 for k = 19

In random initialisation the optimal value of k is 17 and for dataset initialisation the optimal value of k is 19. Different styles of writing numbers leads to more clusters providing better accuracy. Each different style of writing a number goes into a different cluster.

The choice of the initial condition primarily affects the number of iterations required for convergence. With the initial cluster representatives from the dataset we observe a much faster convergence as compared to random initialisation. Moreover, we observe a better accuracy and slightly lower J_Clust value in case of initialisation from the dataset.

Also, in case of initialisation from the dataset the visualisation shows the final cluster representatives are much smoother and closer to the actual numbers compared to the random initialisation.