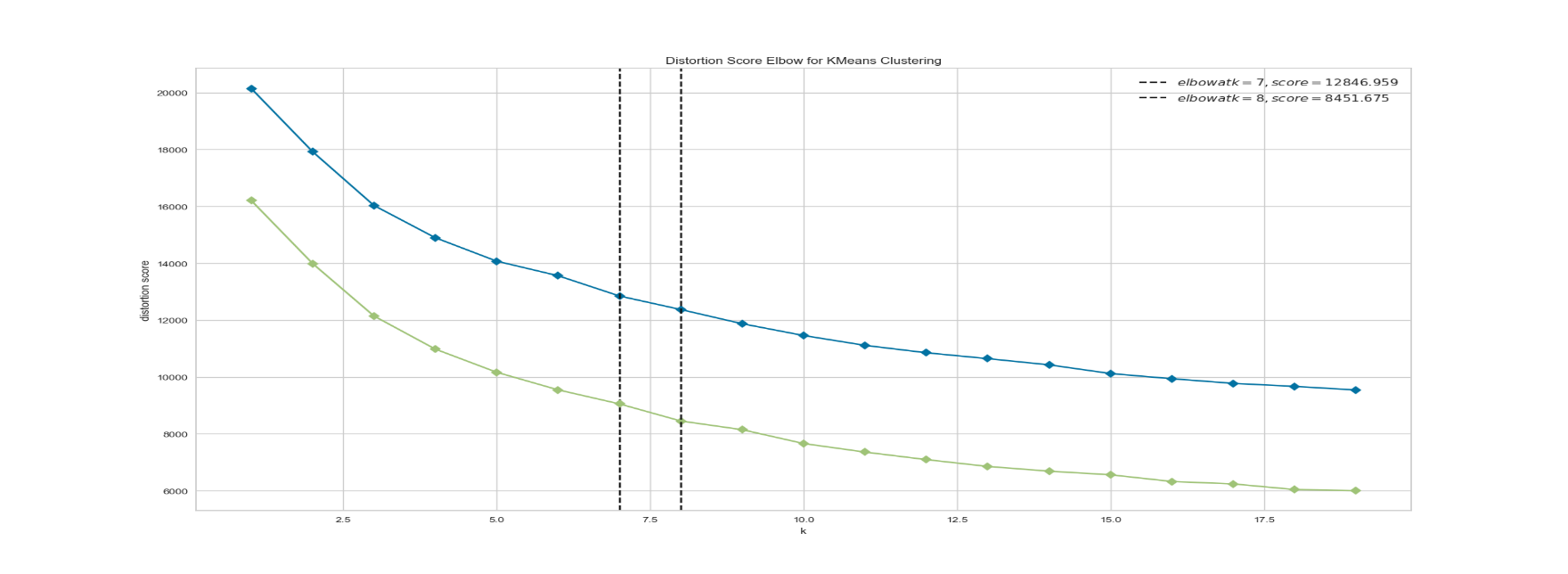
**Data Mining Assignment 2**

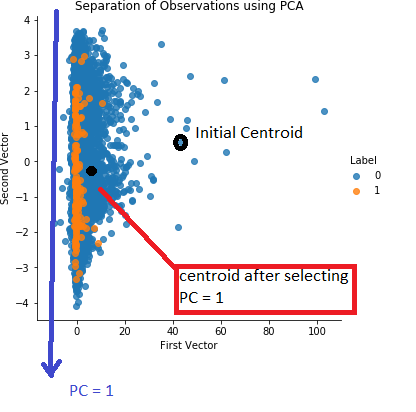
**Question 5:**

*PCA:*



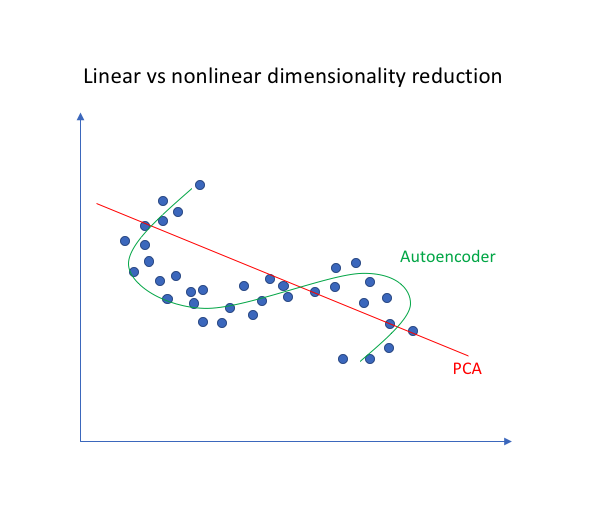
The reduction observed after principal component analysis was significant. Only 10 features out of 38 contain more than 80% of variance. As we have discarded the remaining clusters, the number of clusters are reduced by 1. This is because those removed dimensions contributed as outliers and/or stretched the centroids towards them. Though the pull was infinitesimal, the samples with stronger contributions from these non-principal components formed their own cluster, resulting in addition of one more cluster. As a result, the distortion within the clusters were also reduced to a great extent. As distortion is distance from centroid, each added feature increased the distance diagonally as compared to that of one without it.

Like the centroid shown in image below, centroids were pulled in the redundant principal component 2(not shown in image). As we removed non-principal component, the centroids more homogeneously represented the cluster.



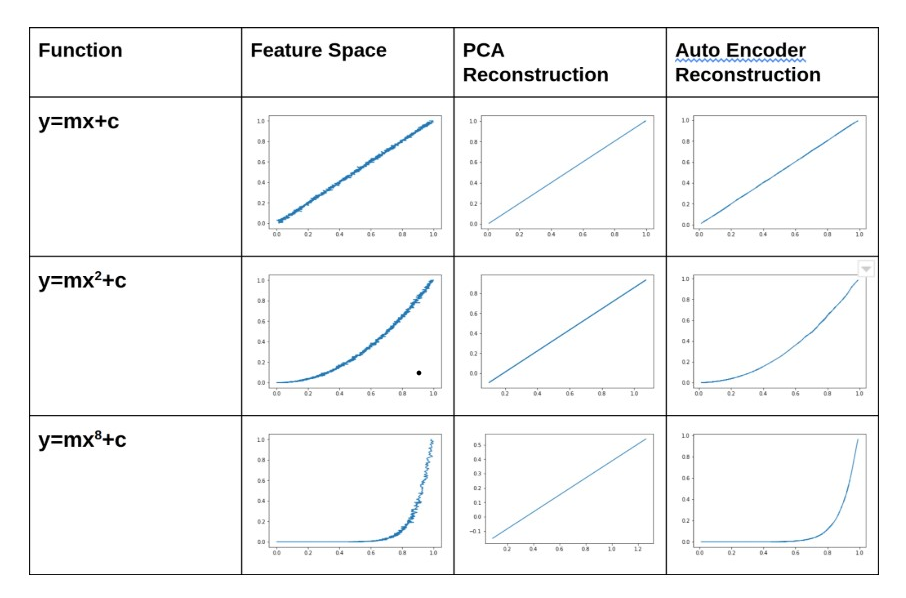
Also, it has been observed that for the same initial cluster centers, samples were classified into different clusters after PCA than clusters assigned to them in previous ones. This can be attributed to both facts that either they were part of cluster that was removed due to PCA and/or they were part of different cluster due to pull experienced from these non-principal component features. This can also be inferred as as decreased distortion among the clusters as we removed non-principal component, the centroids more homogeneously represented the cluster.

**Problem 6:**



PCA of given dataset tried to represent the dataset as a linear function of orthogonal axes. PCA revealed that only 10 features out of 38 are significant. This reduced the number of clusters required to classify the data. Though we recovered most of the data after projecting back the reduced dataset, the losses were significant as compared to Autoencoders.

Autoencoder, on the other hand, provided closer approximation of original data. But at the same time, resultant features of autoencoder, unlike PCA, showed correlation in-between them. Also, while the data represented from encoder was more complex, for many values of (n = encoded features) one or more features of output were 0. Reasons, however, remain unclear.



As PCA did not involve any neural networks, it was computationally cheaper and simple. But, AutoEncoders were both computationally complex and expensive. Encoders took more time to produce results than PCA.

Mean Squared Error of dataset after PCA transformation (2.85011739011994 for n = 10) increased significantly as number of PC’s increased. But, the MSE of dataset for same number of output feature dimension was much lower(0.4177522659301758 for n = 10) as compared to that of PCA.