# PCA of wine dataset

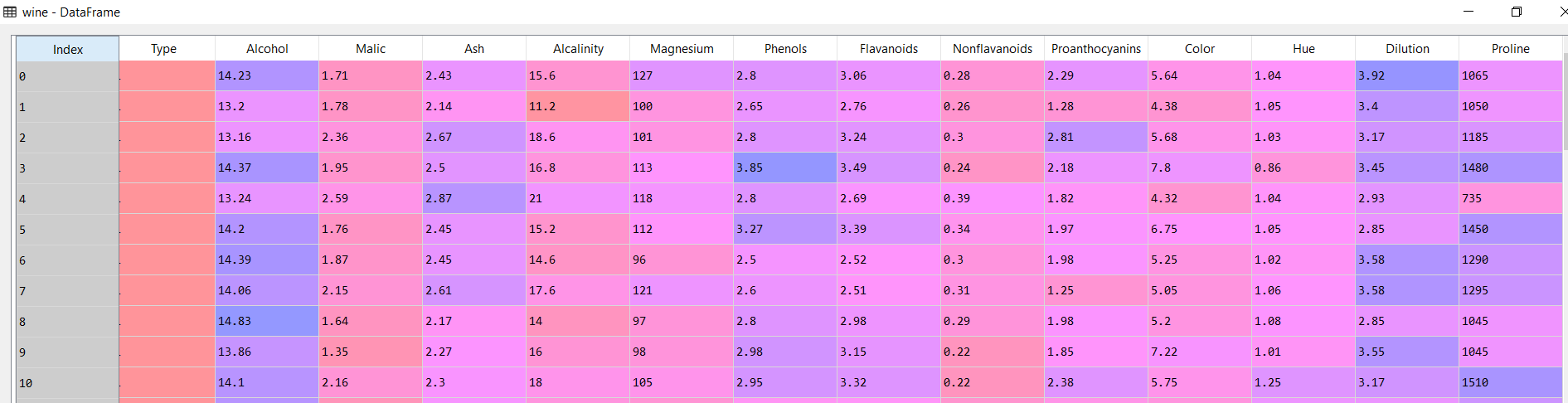
Problem statement:

Perform Principal component analysis and perform clustering using first 3 principal component scores (both hierarchical and k mean clustering(scree plot or elbow curve) and obtain optimum number of clusters and check whether we have obtained same number of clusters with the original data

(class column we have ignored at the beginning who shows it has 3 clusters)

Solution:

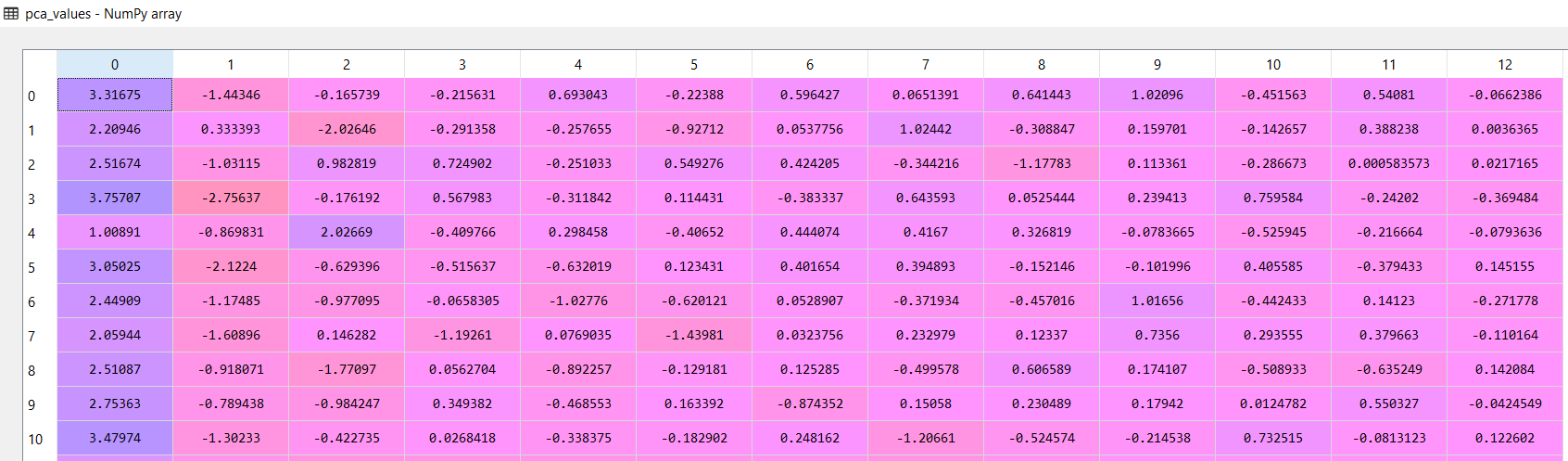
The following is the wine dataset



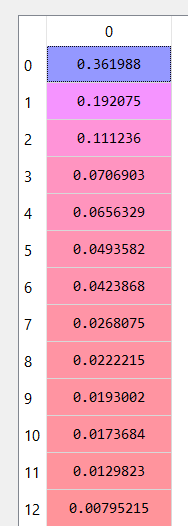
Normalizing the above dataset, we get the following:



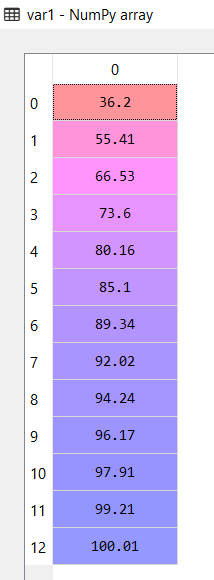
For the above normalized data we obtain the PCA values with 13 n components value:



The variance for the above dataset is as follows:

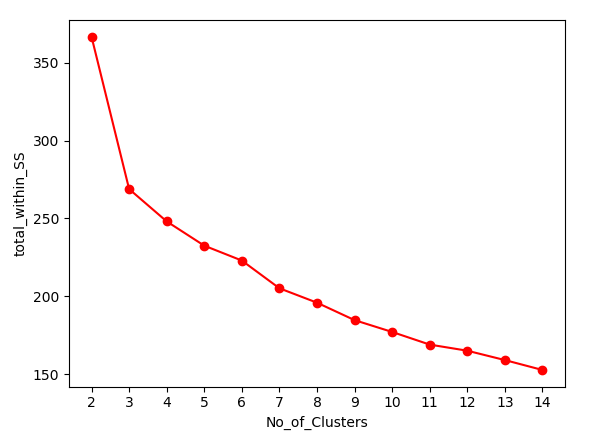


The cumulative variance for the dataset is as follows:



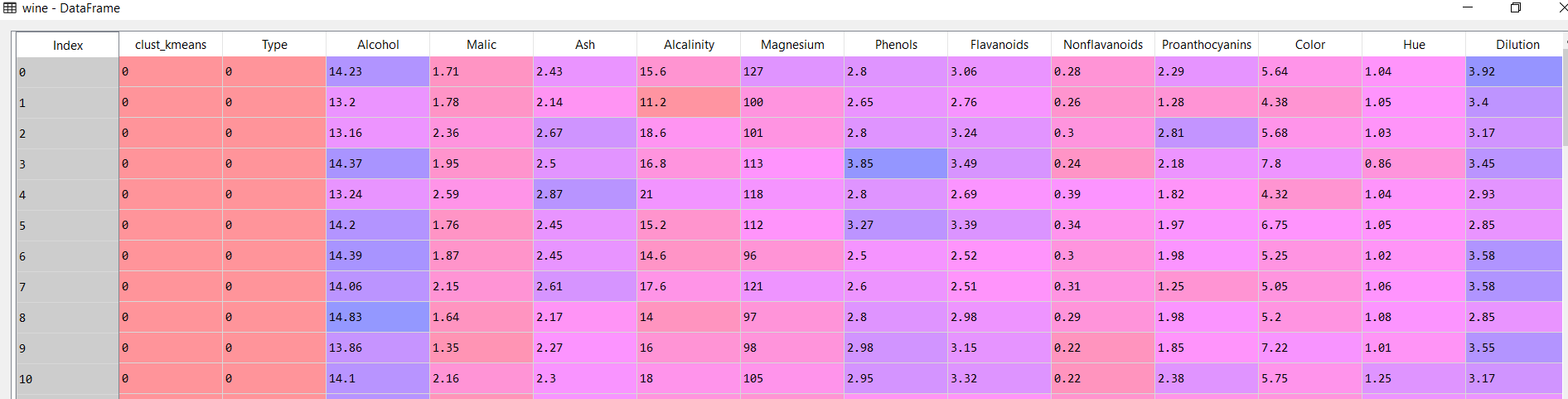
From the above dataset, if we consider the first 3 components, we only retain 73.6% of the information of the dataset.

As per the problem statement, we only consider the first 3 components for a kmeans classification. We obtain the elbow plot as follows:



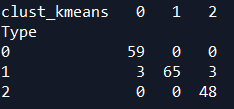
According to the above plot the optimum number of clusters would be 3 since there is a lot of steepness at 4 number of clusters and beyond.

Using kmeans algorithm we classify the different types of wines as the following data:



By using confusion matrix we can compare the already existing classified column “Type” with the newly classified “clust\_kmeans” which was created using kmeans algorithm.

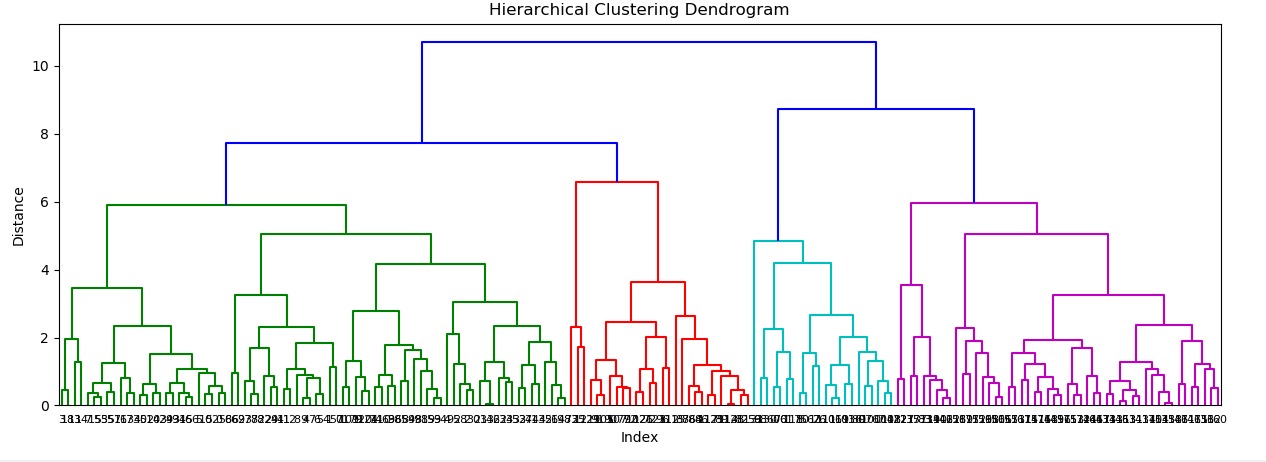
The following is the confusion matrix:



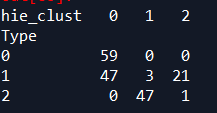
The accuracy of the above confusion matrix is 96.6 %

Now, we will use Hierarchical clustering technique.

The dendrogram for the dataset is as follows:



From the above diagram, we will consider 3 clusters as we have considered the same optimum number for the other techniques, the results are as follows:

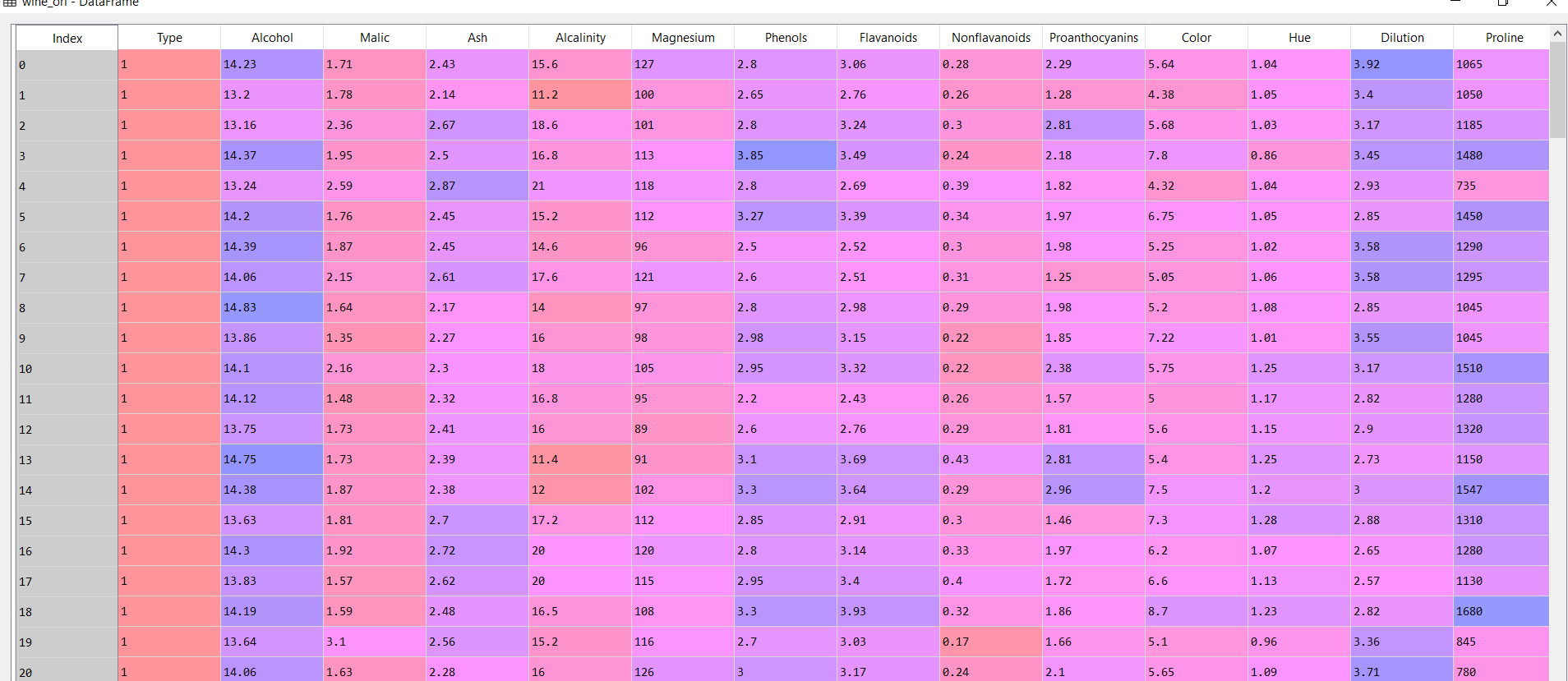


The accuracy of the above confusion matrix is 37.6 %

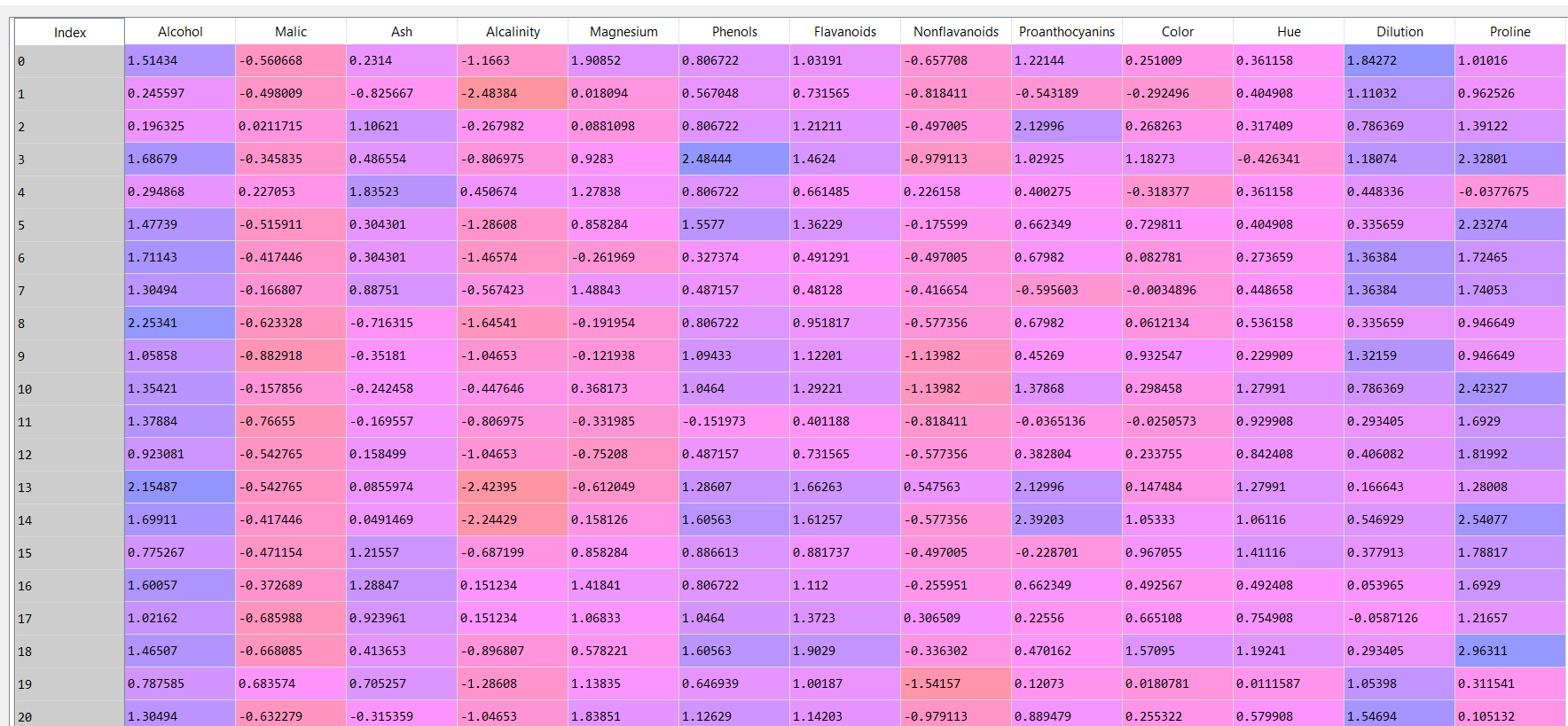
The kmeans clustering provides a greater accuracy then Hierarchical clustering after implementing the PCA technique.

From PCA technique we have obtained 3 clusters, now we will use the hierarchical clustering algorithm on the original data.

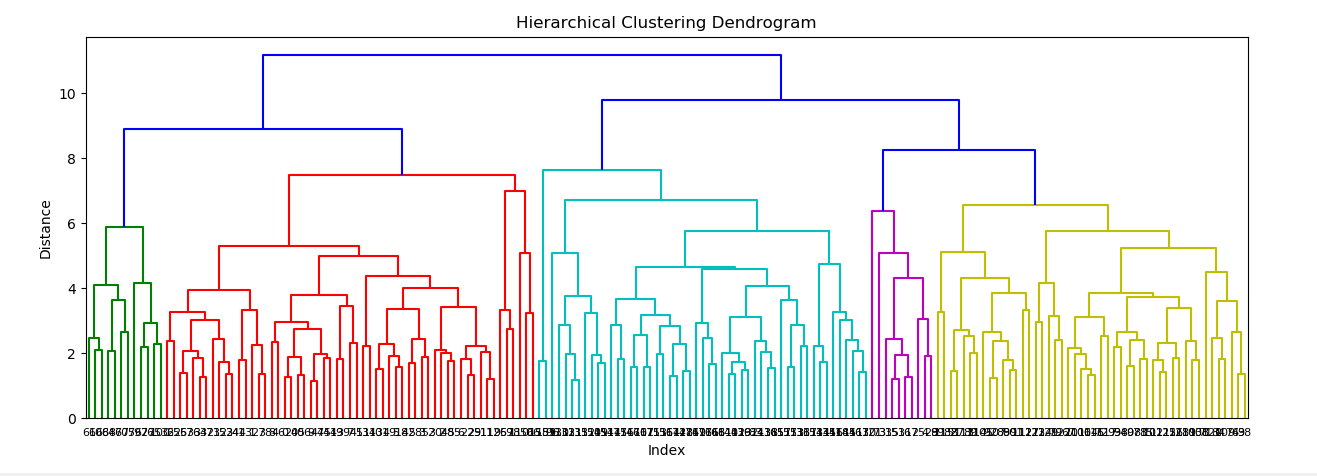
The following is the original data:



Since the difference in the scales are significantly large, we will normalize the above data:



Using complete linkage method with Euclidean distance calculation we get the following dendrogram:



In the above dendrogram we cat the lines at about 9.5 distance where there is a formation of 3 clusters.

From the original data and by implementing PCA, we have achieved the same number of clusters. Hence, we have established that with minimum loss of data by transforming into Principle Components, we still retain majority of the necessary information to make accurate classification.