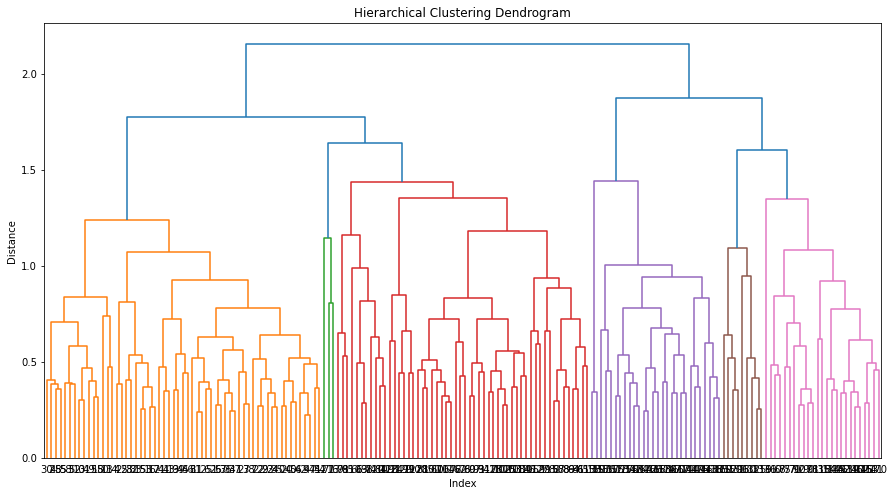
* **Hierarchical clustering without PCA.**
* Perfroming Hclust without reducing its dimensions gives a dendrogram which is difficult to inference.
* Using complete linkage as the criteria and distance using Euclidean distance.

****

* Using Agglomerative with cluster size of 3 .
* wine.groupby('Belonging Cluster').mean()

Type Alcohol Malic Ash Alcalinity Magnesium Phenols Flavanoids Nonflavanoids Proanthocyanins Color Hue Dilution Proline

Belonging Cluster

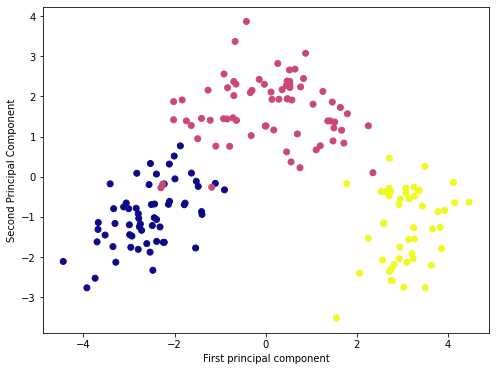
0 1.491379 13.009828 1.975000 2.358276 18.824138 100.396552 2.622845 2.634655 0.315000 1.826293 4.298534 1.067241 3.058534 819.732759

1 2.588235 12.730588 2.420882 2.345882 20.102941 100.470588 1.645588 1.025000 0.420294 1.048529 4.933824 0.830176 1.867059 579.941176

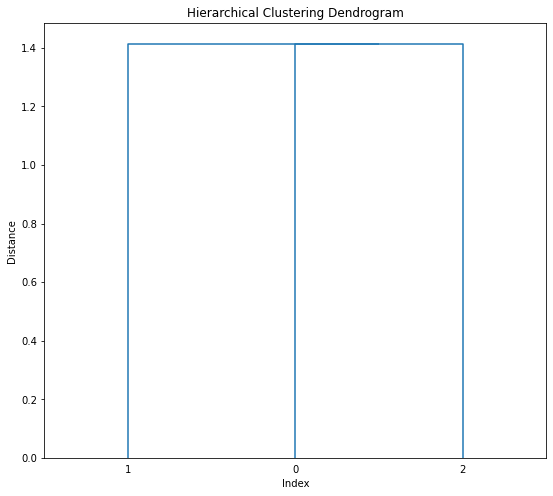
2 3.000000 13.290357 3.730714 2.425714 21.535714 96.142857 1.726071 0.740714 0.485000 1.274286 8.355714 0.657143 1.664643 647.857143

After using PCA :

* scaled\_data.shape🡪 178 \* 14
* x\_pca 🡪 178\* 3
* Drawing a scatter plot to see its correlation. Of x-axis Pca1 and Y-axis Pca2 .

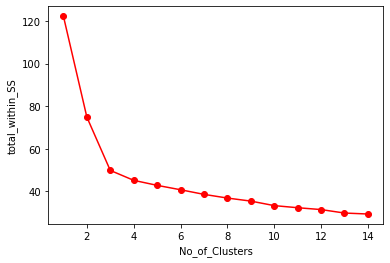


**After applying PCA:**

* All PCA components are uncorrelated and hence its variance varies from high to low (PCA1 with high variance and decreases by further PC)

**Kmeans without PCA:**

* Cluster k=3 , when the elbow curve is found on scree plot.



* model.cluster\_centers\_

array([[0.00833333, 0.71109649, 0.25 , 0.58716578, 0.34295533,

0.4009058 , 0.64408046, 0.55713783, 0.29937107, 0.4701367 ,

0.35938567, 0.47655827, 0.69358974, 0.59595816],

[0.98979592, 0.55832438, 0.5147213 , 0.57470261, 0.55869977,

0.31543922, 0.24074595, 0.09661586, 0.60531382, 0.23755875,

0.51734345, 0.16741331, 0.15392091, 0.24932312],

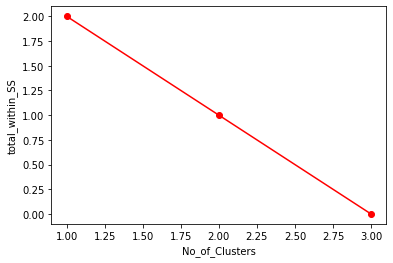
[0.5 , 0.32295957, 0.23093888, 0.46981322, 0.4878231 ,

0.26134216, 0.43883058, 0.36629365, 0.43833744, 0.38348649,

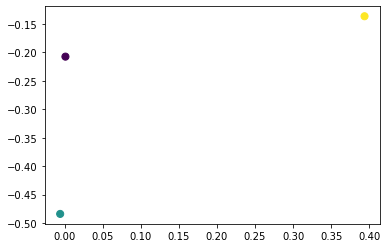
0.15170401, 0.46808059, 0.55539629, 0.16749364]])

**Kmeans with PCA:**

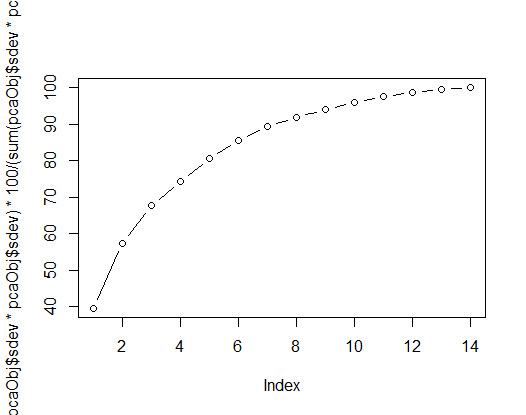
* Here the elbow curve is not present since it s components are reduced to 3 .



* The data is compressed and only PCA (3 rows ).
* Data is uncorrelated for column [Type] and [Alcohol] in wine compressed data.



PCA is limited to 3 because ,it has high variance and then it detoriates if compressed even further ,Thus losing crucial information.



Cluster Plot of Component 1 and Component 2 using R Kmeans

