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) df

Business problem:

To Perform Principal component analysis and perform clustering using first 3 principal component scores.

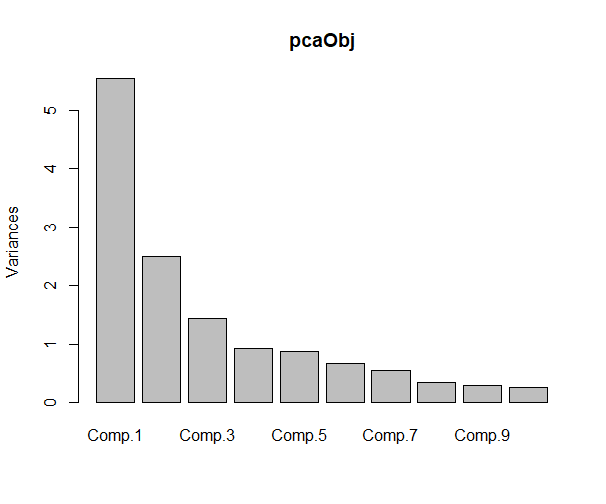
Pre-processing:

Access only numerical data by dropping the unwanted columns.

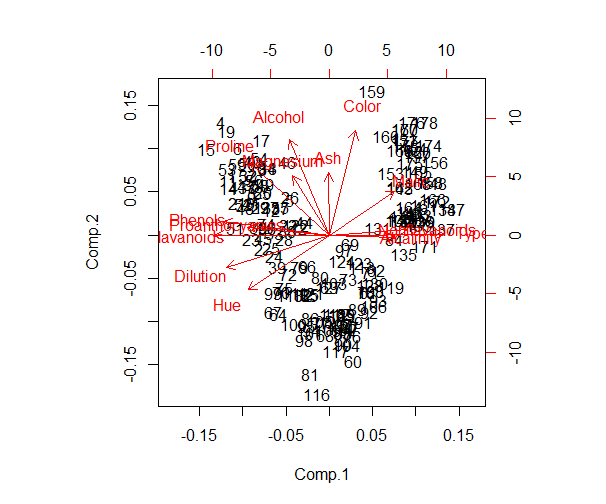
Look at the summary.

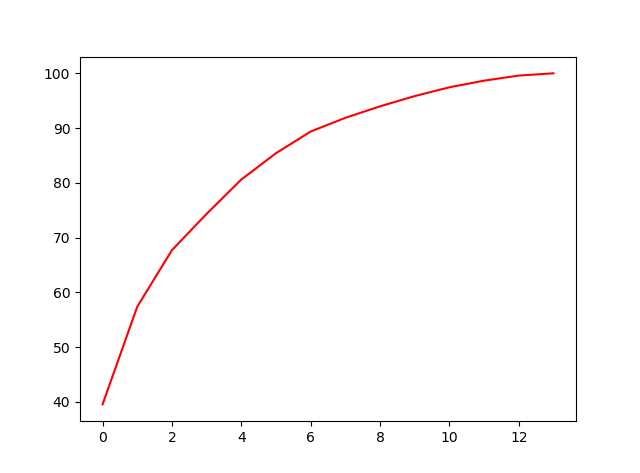
Standardize the values.

Plots:



The above plot describes about the importance of the principle components.





We will get three principle components at last based on the high variance values.

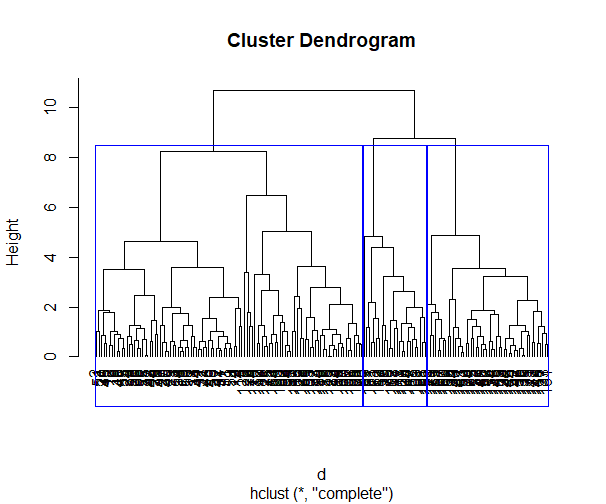
Hierarchical clustering:

Perform hierarchical clustering using 3 principal component scores.

=>load the required packages.

=>load three principle component scores.

Plot:



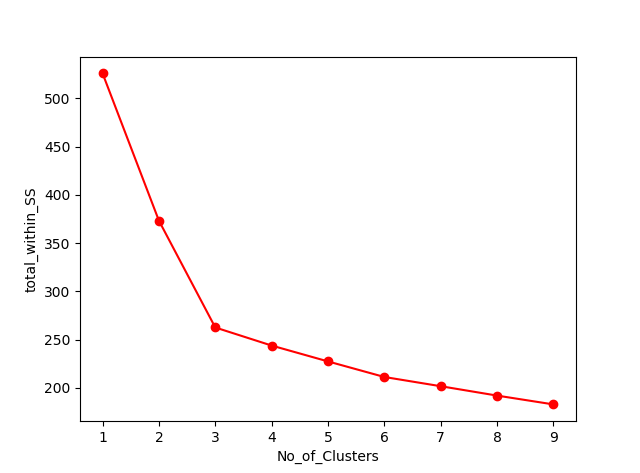
Based on the dendrogram I have taken the cluster value as 3 to form three clusters.

K-Means clustering:

k <- kselection(data[,-5], parallel = TRUE, k\_threshold = 0.9, )

k

by using kselection f(k) finds 3 clusters for the threshold value of 0.9.



Based on the elbow curve we can take the k value as 3 will best fit for the model.

fit <- kmeans (data, 3) # 3 cluster solution

str(fit)

$ totss : num 1687

$ withinss : num [1:3] 169.3 224.1 97.9

$ tot. withinss: num 491

$ betweenss : num 1196

$ size : int [1:3] 62 67 49

Fit

K-means clustering with 3 clusters of sizes 62, 67, 49

Cluster means:

Comp.1 Comp.2 Comp.3

1 -2.5488589 0.9366105 0.001631632

2 0.1217114 -1.7503689 0.128333665

3 3.0586650 1.2082626 -0.177541159

Within cluster sum of squares by cluster:

[1] 169.2931 224.0527 97.8848

(between\_SS / total\_SS = 70.9 %)

Based on the k means taking k as 3 is best for building the model.