**Business Problem:**

Perform Principal component analysis and perform clustering using first 3 principal component scores (both heirarchial 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.

**Data:**

Data in the form of numerical data and it has 3 types of alcohol.

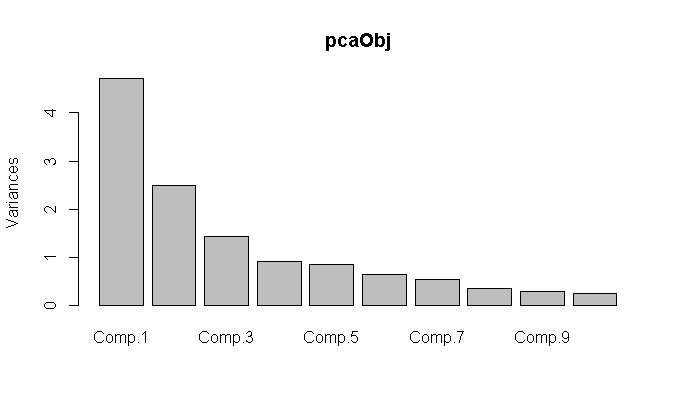
**Pre-processing Data:**

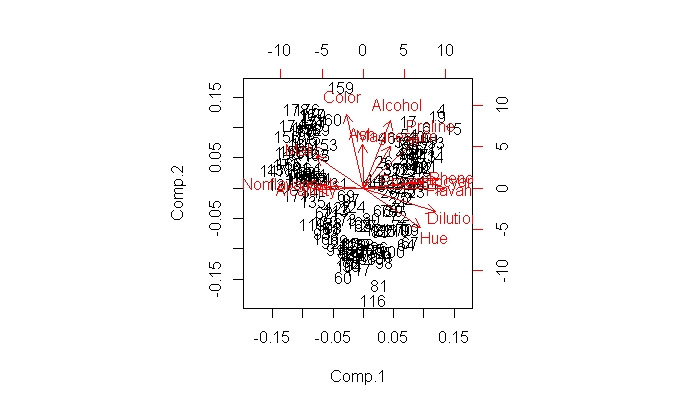
All the features are not in same scale. So first of all convert them all into single scale. There was no outlier and NA in the data. Delete the unused feature like TYPE from the processing.

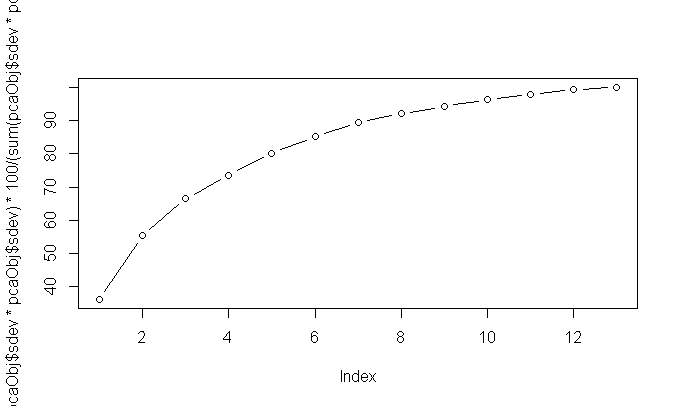
**Converting to PCA:**

The dataset contains the 12 features, It will be little tough handle all features. So we are going to processing data into Principle Component Analysis.

Look the below diagram and consider the how much percentage of data you want.







Data in the PCA component wise:

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7

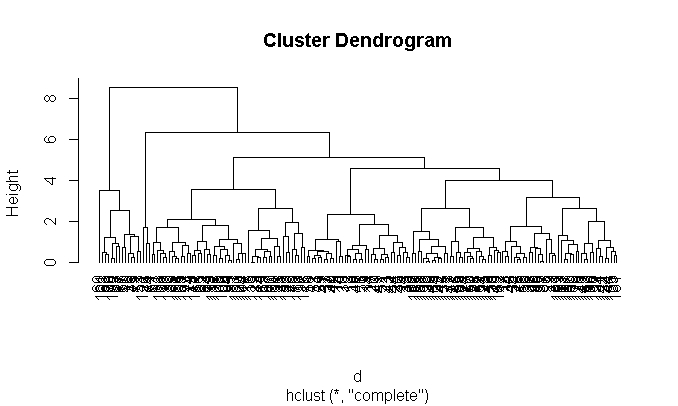
36.19885 55.40634 66.52997 73.59900 80.16229 85.09812 89.33680

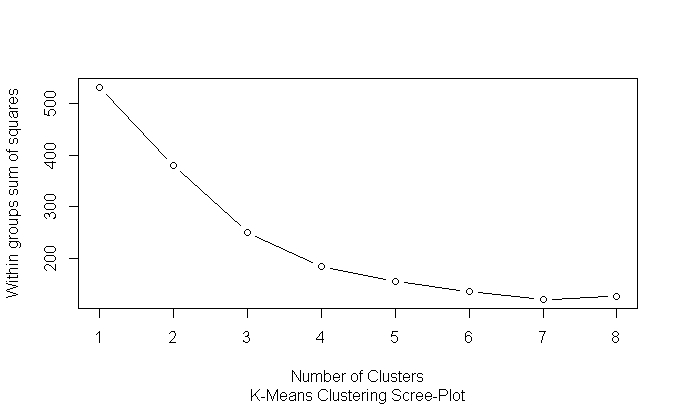
Comp.8 Comp.9 Comp.10 Comp.11 Comp.12 Comp.13

92.01754 94.23970 96.16972 97.90655 99.20479 100.00000

As per the requirement, I’m considering the 3 PCA data and it’s a 66.5% data of total data. We will find the ‘K’ for 3 PCA and will compare with Normal data clustering.

Build the model using Euclidean distance and **complete** linkage functions. Please find the Dendrogram.

**Elbow Curve:**

****

By looking the Elbow Curve, **3 will be good K** value as per my analysis.

**K-Selection:**

k <- kselection(normalized\_data, k\_threshold = 0.9, max\_centers=10)

It is giving the **k value as 3**.

**KMeans:**

Calculate the tot.withinss and betweenss. For a good model tot.withinss should be more and betweenss should be less. If there should be less variation in the difference then we can treat it as final K.

Trails :

K=2 : km <- kmeans(normalized\_data,2)

$ tot.withinss: num 398

$ betweenss : num 133

K=3 : km <- kmeans(normalized\_data,3)

$ tot.withinss: num 251

$ betweenss : num 288

K=4 : km <- kmeans(normalized\_data,4)

$ tot.withinss: num 29189

$ betweenss : num 14789

By looking the data**, 2 will be good k** value as per the KMEANS.

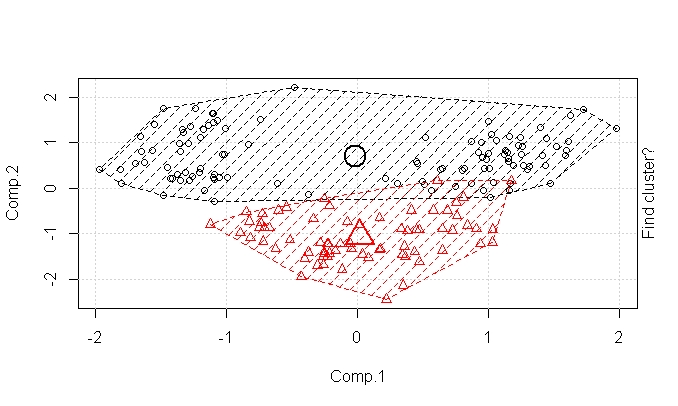
**KMEANS & Centres:**

**When K Equal to 2.**

Comp.1 Comp.2 Comp.3

[1,] -0.01266777 0.6968272 0.1365931

[2,] 0.01909086 -1.0501481 -0.2058516



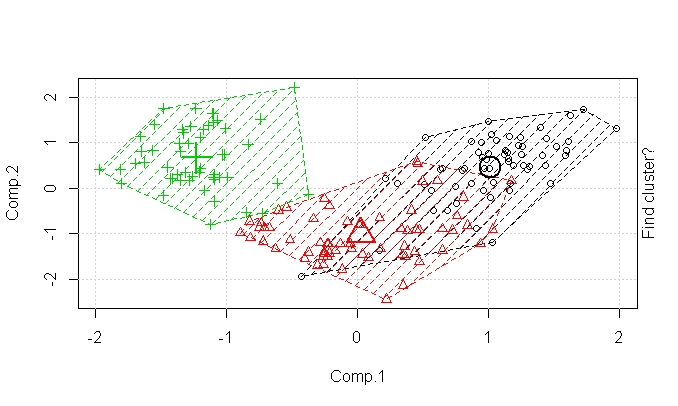
When K is equal to 3.

Comp.1 Comp.2 Comp.3

[1,] 1.01825082 0.4579227 0.3955366

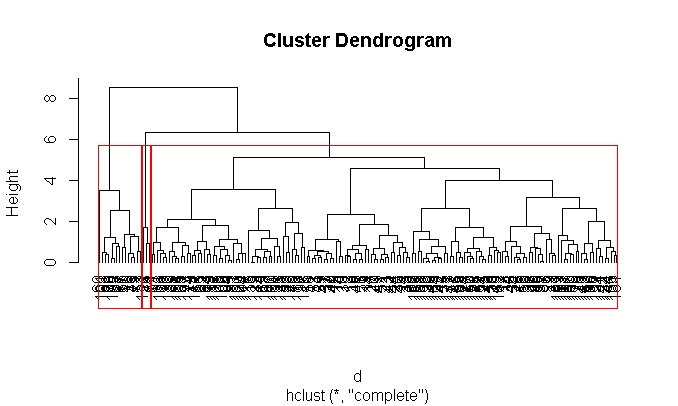
[2,] 0.02750377 -1.0146273 -0.5673448

[3,] -1.22385451 0.6703832 0.2116878



By looking the animation and k means centres, **2 will be good value for k**.

Hence, with all above data, I am taking the k value as 3 and going for clustering.



**3 Clustered Data:**

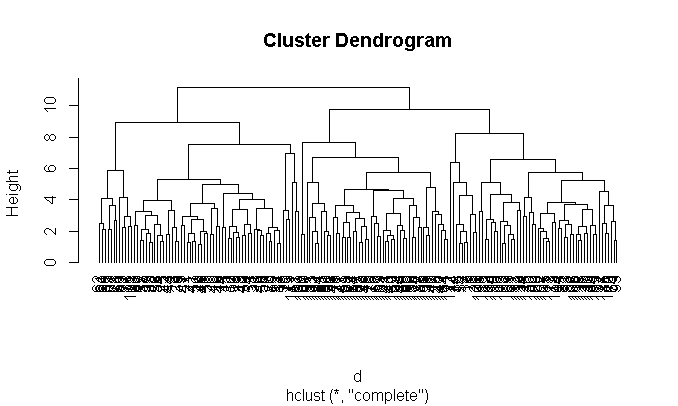
Cluster 1 Cluster 2 Cluster 3

160 15 3

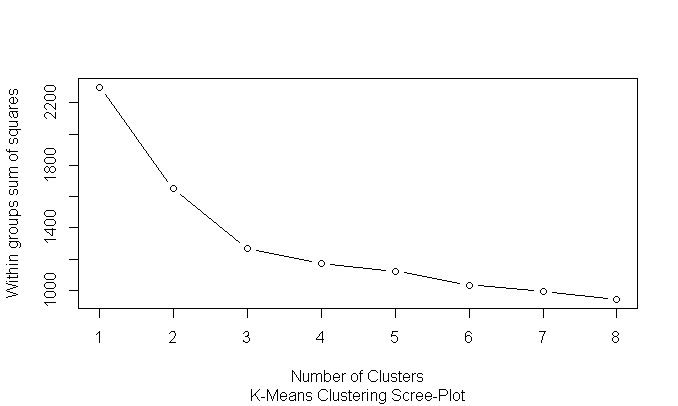
**Applying the Clustering on Normal Wines Data:**

All the features are not in same scale. So first of all convert them all into single scale. There was no outlier and NA in the data. Delete the unused feature like TYPE from the processing.

Build the model using Euclidean distance and complete linkage functions. Please find the Dendrogram.



**Elbow Curve:**



Elbow Curve value is subjective. As per my analysis **3 will be good value for K.**

**KSelection :**

k <- kselection(normalized\_data, k\_threshold = 0.9, max\_centers=10)

KSelection is giving 2.

**KMeans:**

Calculate the tot.withinss and betweenss. For a good model tot.withinss should be more and betweenss should be less. If there should be less variation in the difference then we can treat it as final K.

Trails :

K=2 : km <- kmeans(normalized\_data,2)

$ tot.withinss: num 1649

$ betweenss : num 652

K=3 : km <- kmeans(normalized\_data,3)

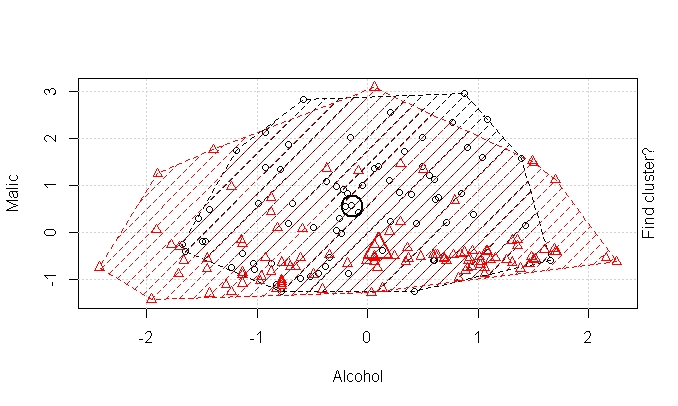
$ tot.withinss: num 1271

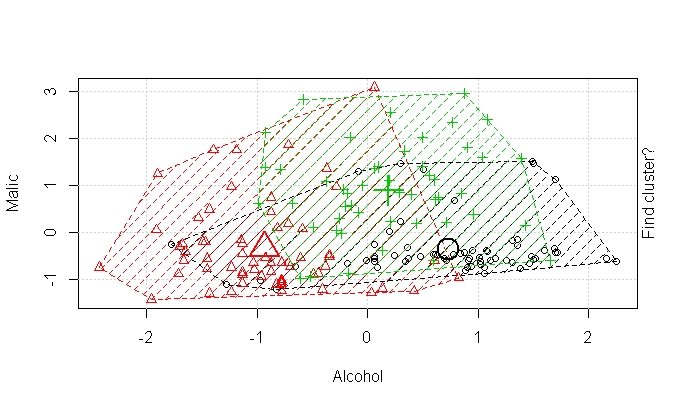
$ betweenss : num 1030

By looking the data , **3 will be good value for K**.

**KMeans Centers & Animation :**

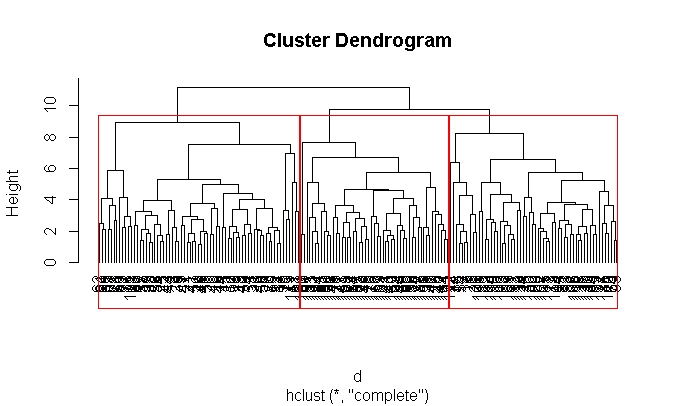
When K =2

 When K =3



By looking the both diagrams, 3 will be a good value for K.

So, Taking Hierarchical Clustering using k =3.



**3 Clustered Data:**

Cluster 1 Cluster 2 Cluster 3

69 58 51

**Conclusion:**

When we compared 3 PCA data with Normal data, we are not getting same number of observation in the clusters. One good thing is we are getting same number of clusters for both PCA and Non PCA data. There difference in the observation. And also 3 PCA data is covering only 66.6% data is also one reason and we have less number of observations.

PCA is basically applied for dimensionality reduction, while K-means is used to cluster your data. However, utilizing them \*together\* is not necessary, unless you have a proper reason.