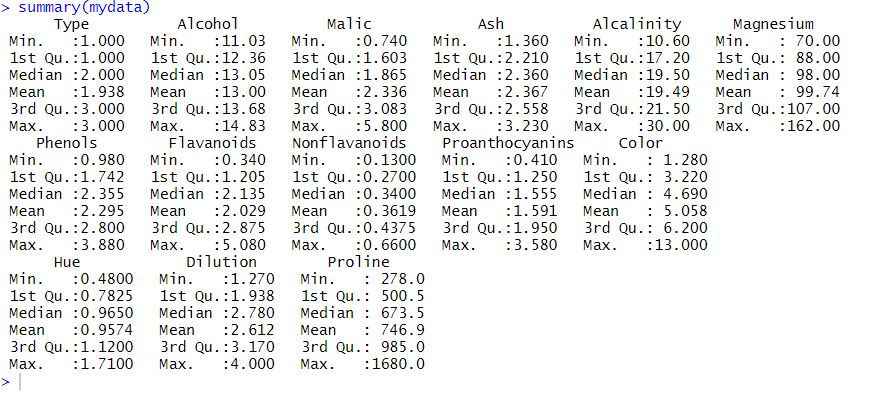
Dimension Reduction Assignment

Business Problem: Perform Principal component analysis and perform clustering using first 3 principal component scores

Objective: Perform hierarchical and k mean clustering and obtain Optimum number of clusters and check whether we have obtained same number of clusters with the original data.

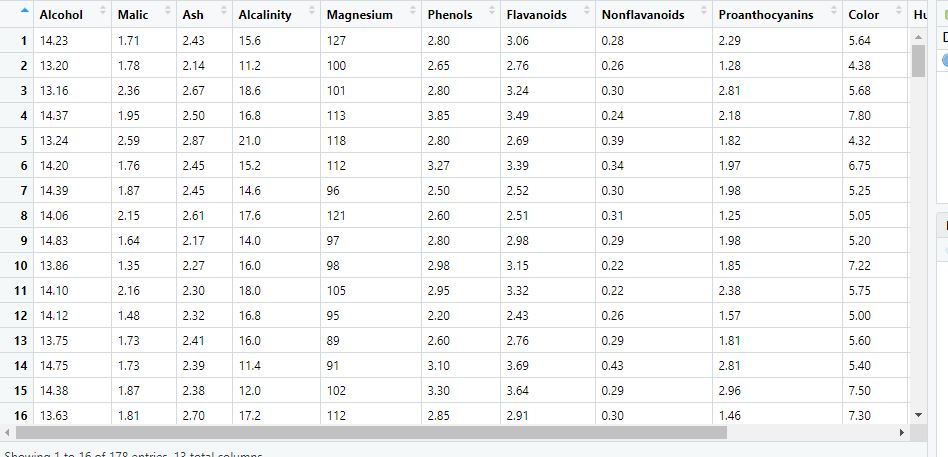
About the data :

Wines.csv file contains different types of wines and their ingredient content. The summary of the files can be given as:



Data Pre-processing:

The first column has the type of the wines. Removing the first column, considering only the numeric values for applying PCA.

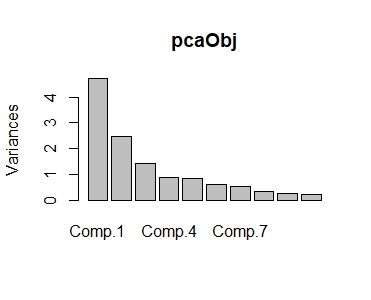


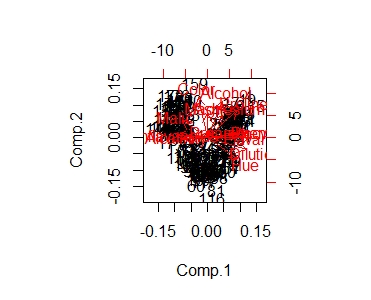
PCA suggested clustering with first three PCs.

PLOT DIAGRAM

Graph showing importance of principal components,

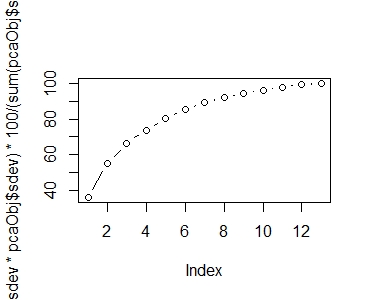
Comp.1 having highest importance (highest variance)





Showing the increase of variance with considering principal components

Which helps in choosing number of principal components



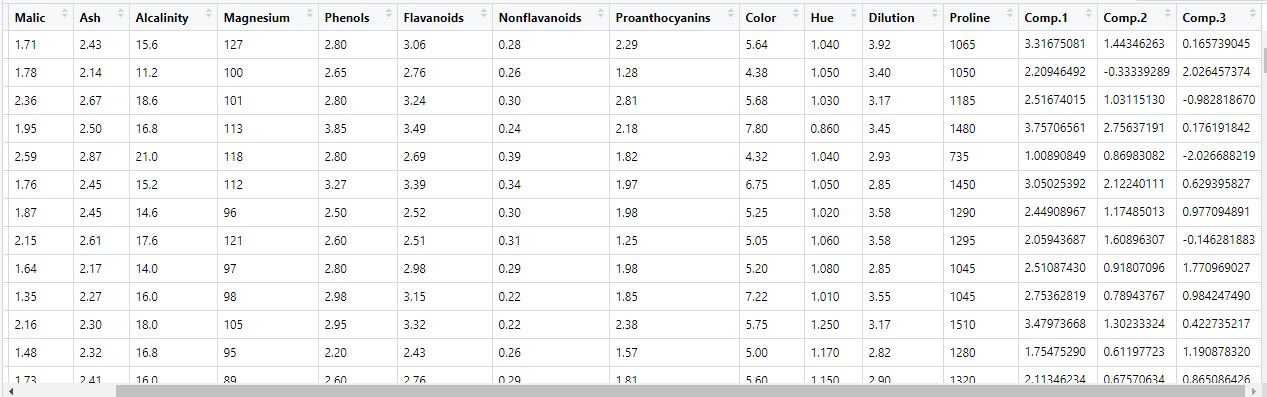
first 3 PCA Scores which we have to select

cbind used to bind the data in column wise

Considering first 3 principal component scores and binding them with wine data

winepc\_data<-cbind(wine\_data,pcaObj$scores[,1:3])

View(winepc\_data)

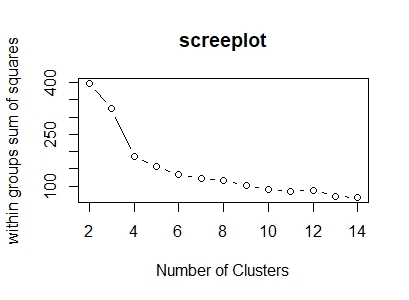


Considering only PCA scores as they represent the entire data, we preparing data for clustering

clus\_data<-winepc\_data[,14:16]

Normalizing the data by using the scale function and then to find the distance considering Euclidean distance.

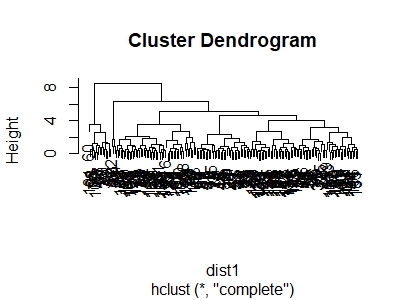
scree plot for selection of no.of clusters with suggested pcs

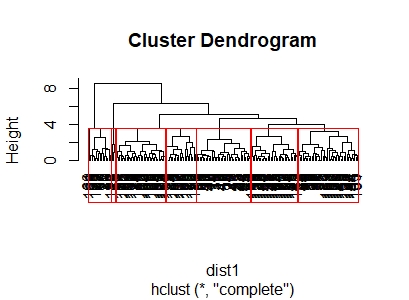


Clustering the data using hclust function (Hierarchical with PC suggested)

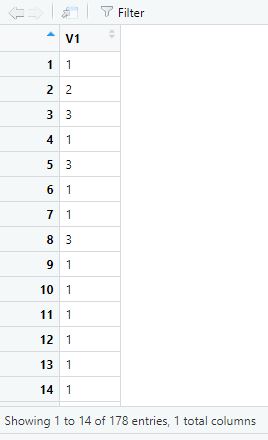
fit1<-hclust(dist1,method="complete") # method here is complete linkage

PLOT fit1

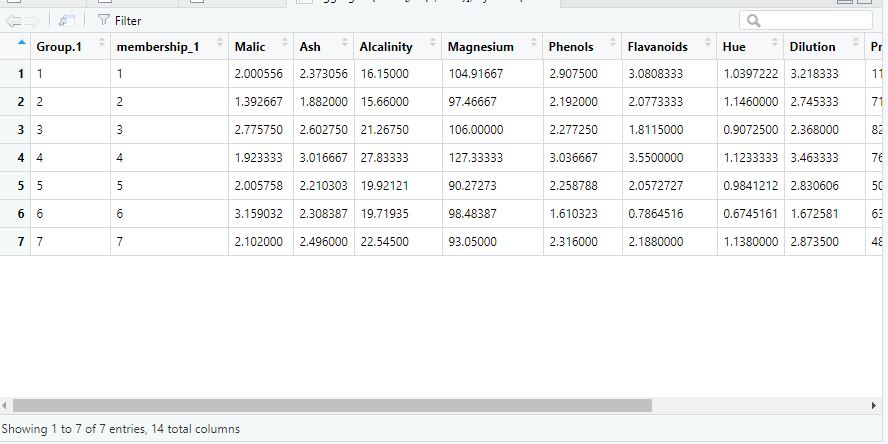




membership\_1<-as.matrix(groups) # cluster numbering

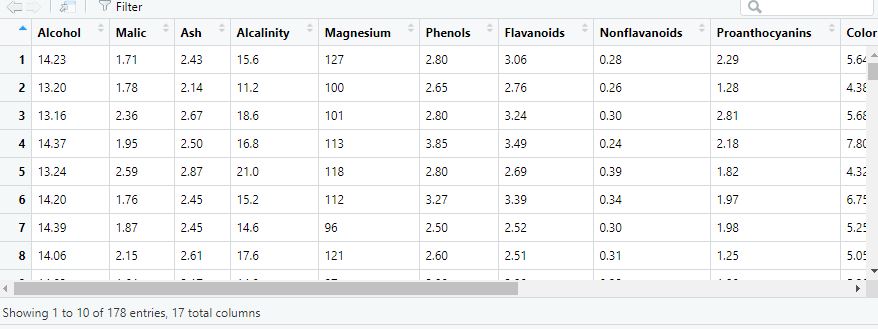


drawn from the aggregate of the wine data on membership\_1



##Kmeans clustering with PC

# choosing best clusters as 7



By forming 7 clusters value of total withiness is reduced and betweeness is increased but we will check that once by elbow turn at 7

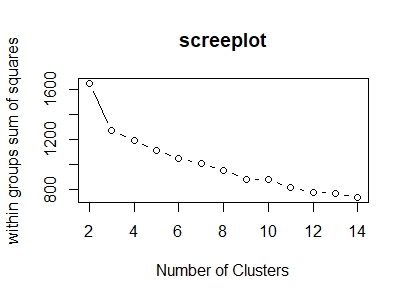
# considering 7 clusters is best

### All variable clustering ###

## Normalizing the data

Again normalizing the data by using the scale function and then to find the distance considering Euclidean distance.

SCREE PLOT



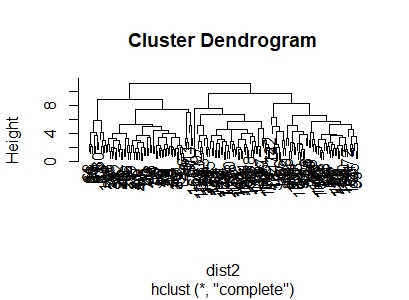
From scree plot first elbow bend is at 5 so we can say that we have to go with 5 no. of clusters.

Clustering

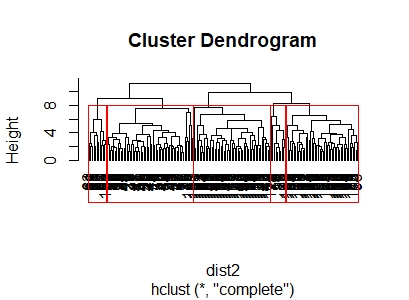
clustering the data using hclust function --> Hierarchical with all variables.

fit2<-hclust(dist2,method="complete") # method here is complete linkage

Plot (fit2)

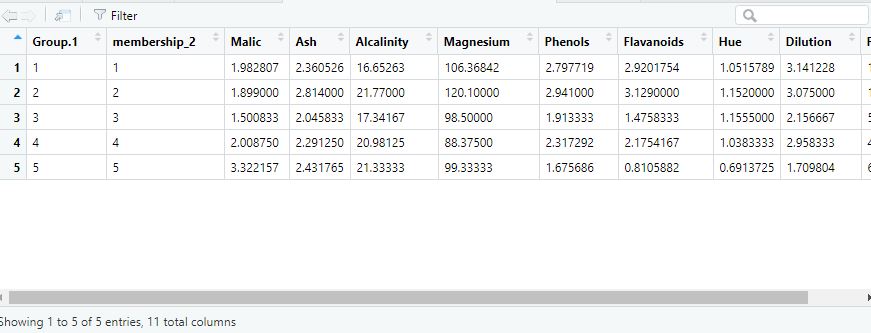


rect.hclust(fit2, k=5, border="red")



final2<-cbind(membership\_2,wine\_data) # binding column wise with orginal data

View(aggregate(final2[,-c(2,9:11)],by=list(membership\_2),FUN=mean))



# Inferences can be drawn from the aggregate of the wine data on membership\_2

##Kmeans clustering with all variables

# choosing best clusters as 5

k\_5all <- kmeans(wine\_norm,5)

k\_5all <- kmeans(wine\_norm,5)

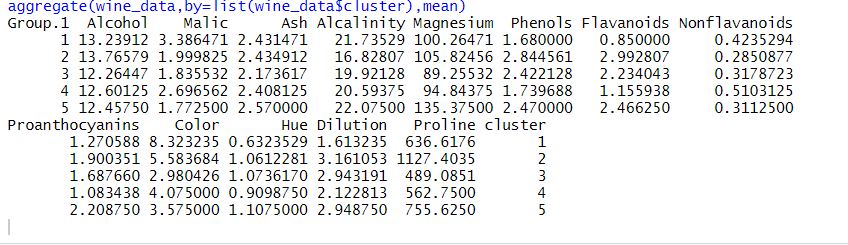
str(k\_5all)

View(k\_5all)

wine\_data$cluster <- as.matrix(k\_5all$cluster)

View(wine\_data)

aggregate(wine\_data,by=list(wine\_data$cluster),mean)



considering 5 clusters is best by taking all variables as that of pc because first three pc we have selected as it is given in problem

first 3 pc cover only 66% of information so this variation of cluster selection we are getting.

