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

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

# Loading wine data

mydata<-read.csv(file.choose()) ## use read.csv for csv file

View(mydata)

help(princomp) ## to understand the api for princomp

attach(mydata)

cor(mydata)

# cor = TRUE use correlation matrix for getting PCA scores

?princomp

pcaObj<-princomp(mydata, cor = TRUE, scores = TRUE, covmat = NULL)

str(pcaObj)

## princomp(mydata, cor = TRUE) not\_same\_as prcomp(mydata, scale=TRUE); similar, but different

summary(pcaObj)

str(pcaObj)

loadings(pcaObj)

plot(pcaObj) # graph showing importance of principal components

# Comp.1 having highest importance (highest variance)

biplot(pcaObj)

# Showing the increase of variance with considering principal components

# Which helps in choosing number of principal components

plot(cumsum(pcaObj$sdev\*pcaObj$sdev)\*100/(sum(pcaObj$sdev\*pcaObj$sdev)),type="b")

#pcaObj$loadings

pcaObj$scores[,1:3] # Top 3 PCA Scores which represents the whole data

# cbind used to bind the data in column wise

# Considering top 3 principal component scores and binding them with mydata

mydata<-cbind(mydata,pcaObj$scores[,1:3])

View(mydata)

# preparing data for clustering (considering only pca scores as they represent the entire data)

clus\_data<-mydata[,15:17]

# Normalizing the data

norm\_clus<-scale(clus\_data) # Scale function is used to normalize data

dist1<-dist(norm\_clus,method = "euclidean") # method for finding the distance

# here I am considering Euclidean distance

# Clustering the data using hclust function --> Hierarchical

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

fit2 <- kmeans(norm\_clus,3)

fit2

plot(fit1) # Displaying Dendrogram

groups<-cutree(fit1,3) # Cutting the dendrogram for 5 clusters

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

View(membership\_1)

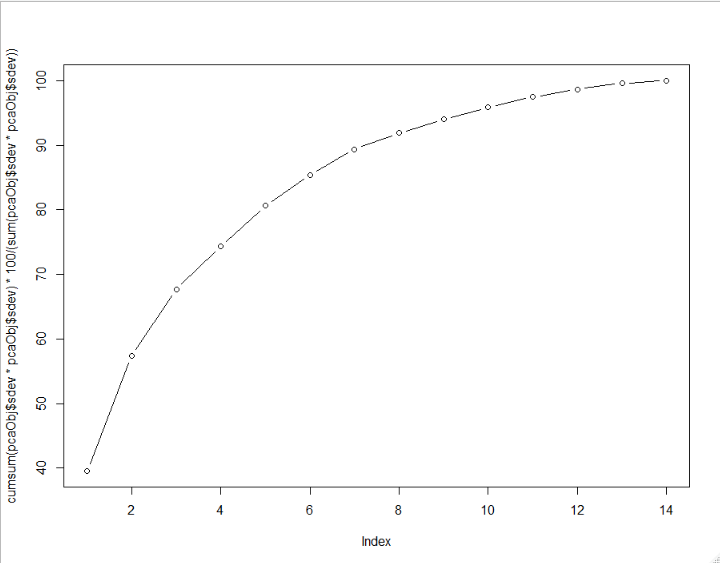
final1<-cbind(membership\_1,mydata) # binding column wise with orginal data

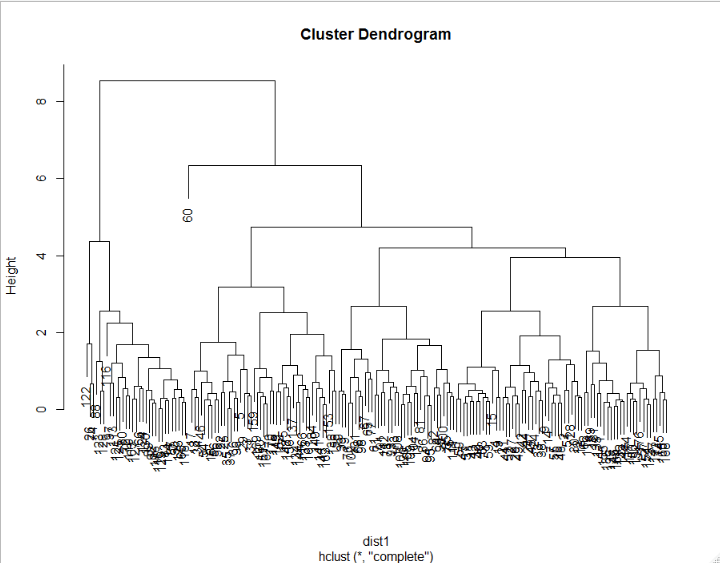
View(final1)

View(aggregate(final1[,-c(1,16:18)],by=list(membership\_1),FUN=mean)) # Inferences can be

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

Plots





Output from model

Importance of components:

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9

Standard deviation 2.3528595 1.5802140 1.2025283 0.96328492 0.93675104 0.8202307 0.74417670 0.59163686 0.54271718

Proportion of Variance 0.3954249 0.1783626 0.1032910 0.06627984 0.06267875 0.0480556 0.03955707 0.02500244 0.02103871

Cumulative Proportion 0.3954249 0.5737874 0.6770785 0.74335831 0.80603706 0.8540927 0.89364973 0.91865217 0.93969088

Comp.10 Comp.11 Comp.12 Comp.13 Comp.14

Standard deviation 0.51215828 0.47523512 0.41084877 0.359950245 0.240441928

Proportion of Variance 0.01873615 0.01613203 0.01205691 0.009254584 0.004129451

Cumulative Proportion 0.95842703 0.97455906 0.98661596 0.995870549 1.000000000

> str(pcaObj)