**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**

**Ans:-**

> library(corrplot)

> library(psych)

> wine\_data<-read.csv("D:\\Assignment\_\_Rstudio\\PCA\\wine.csv") ## use read.csv for csv files

> View(wine\_data)

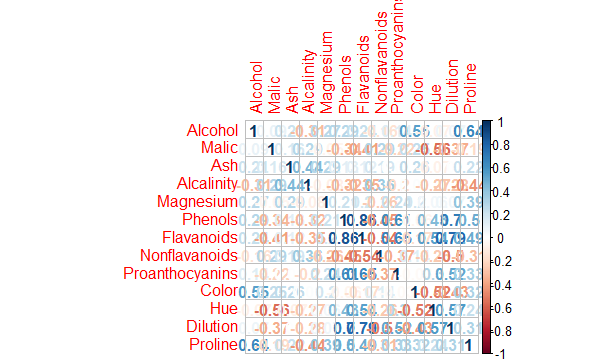
> help(princomp) ## to understand the api for princomp

>> # wine\_data[-1] -> Considering only numerical values for applying PCA

> wine <- wine\_data[,-1]

> attach(wine)

corrplot(cor(wine),method = "number")



# cor = TRUE use correlation matrix for getting PCA scores

?princomp

# cor = TRUE use correlation matrix for getting PCA scores

> ?princomp

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

> str(pcaObj)

List of 7

$ sdev : Named num [1:13] 2.169 1.58 1.203 0.959 0.924 ...

..- attr(\*, "names")= chr [1:13] "Comp.1" "Comp.2" "Comp.3" "Comp.4" ...

$ loadings: 'loadings' num [1:13, 1:13] 0.14433 -0.24519 -0.00205 -0.23932 0.14199 ...

..- attr(\*, "dimnames")=List of 2

.. ..$ : chr [1:13] "Alcohol" "Malic" "Ash" "Alcalinity" ...

.. ..$ : chr [1:13] "Comp.1" "Comp.2" "Comp.3" "Comp.4" ...

$ center : Named num [1:13] 13 2.34 2.37 19.49 99.74 ...

..- attr(\*, "names")= chr [1:13] "Alcohol" "Malic" "Ash" "Alcalinity" ...

$ scale : Named num [1:13] 0.81 1.114 0.274 3.33 14.242 ...

..- attr(\*, "names")= chr [1:13] "Alcohol" "Malic" "Ash" "Alcalinity" ...

$ n.obs : int 178

$ scores : num [1:178, 1:13] 3.32 2.21 2.52 3.76 1.01 ...

..- attr(\*, "dimnames")=List of 2

.. ..$ : NULL

.. ..$ : chr [1:13] "Comp.1" "Comp.2" "Comp.3" "Comp.4" ...

$ call : language princomp(x = wine, cor = TRUE, scores = TRUE, covmat = NULL)

- attr(\*, "class")= chr "princomp"

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

> summary(pcaObj)

Importance of components:

Comp.1 Comp.2 Comp.3 Comp.4 Comp.5

Standard deviation 2.1692972 1.5801816 1.2025273 0.9586313 0.92370351

Proportion of Variance 0.3619885 0.1920749 0.1112363 0.0706903 0.06563294

Cumulative Proportion 0.3619885 0.5540634 0.6652997 0.7359900 0.80162293

Comp.6 Comp.7 Comp.8 Comp.9 Comp.10

Standard deviation 0.80103498 0.74231281 0.59033665 0.53747553 0.50090167

Proportion of Variance 0.04935823 0.04238679 0.02680749 0.02222153 0.01930019

Cumulative Proportion 0.85098116 0.89336795 0.92017544 0.94239698 0.96169717

Comp.11 Comp.12 Comp.13

Standard deviation 0.47517222 0.41081655 0.321524394

Proportion of Variance 0.01736836 0.01298233 0.007952149

Cumulative Proportion 0.97906553 0.99204785 1.000000000

> loadings(pcaObj)

Loadings:

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

Alcohol 0.144 0.484 0.207 0.266 0.214 0.396 0.509

Malic -0.245 0.225 -0.537 0.537 -0.421

Ash 0.316 -0.626 0.214 0.143 0.154 0.149 -0.170 -0.308

Alcalinity -0.239 -0.612 -0.101 0.287 0.428 0.200

Magnesium 0.142 0.300 -0.131 0.352 -0.727 -0.323 -0.156 0.271

Phenols 0.395 -0.146 -0.198 0.149 -0.406 0.286

Flavanoids 0.423 -0.151 -0.152 0.109 -0.187

Nonflavanoids -0.299 -0.170 0.203 0.501 -0.259 -0.595 -0.233 0.196

Proanthocyanins 0.313 -0.149 -0.399 -0.137 -0.534 -0.372 0.368 -0.209

Color 0.530 0.137 -0.419 0.228

Hue 0.297 -0.279 0.428 0.174 0.106 -0.232 0.437

Dilution 0.376 -0.164 -0.166 -0.184 0.101 0.266 0.137

Proline 0.287 0.365 0.127 0.232 0.158 0.120 0.120 -0.576

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

Alcohol 0.212 0.226 0.266

Malic -0.309 -0.122

Ash 0.499 -0.141

Alcalinity -0.479

Magnesium

Phenols -0.320 -0.304 0.304 -0.464

Flavanoids -0.163 0.832

Nonflavanoids 0.216 -0.117 0.114

Proanthocyanins 0.134 0.237 -0.117

Color -0.291 -0.604

Hue -0.522 -0.259

Dilution 0.524 -0.601 -0.157

Proline 0.162 -0.539

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

SS loadings 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000 1.000

Proportion Var 0.077 0.077 0.077 0.077 0.077 0.077 0.077 0.077 0.077

Cumulative Var 0.077 0.154 0.231 0.308 0.385 0.462 0.538 0.615 0.692

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

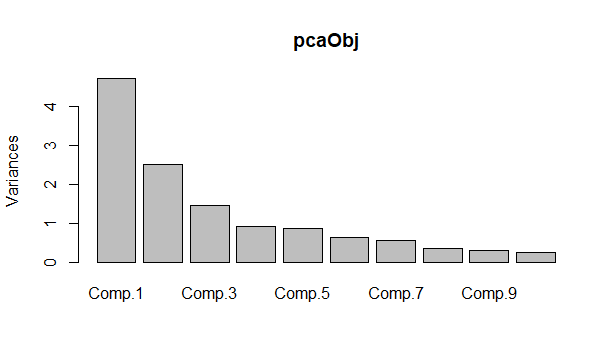
SS loadings 1.000 1.000 1.000 1.000

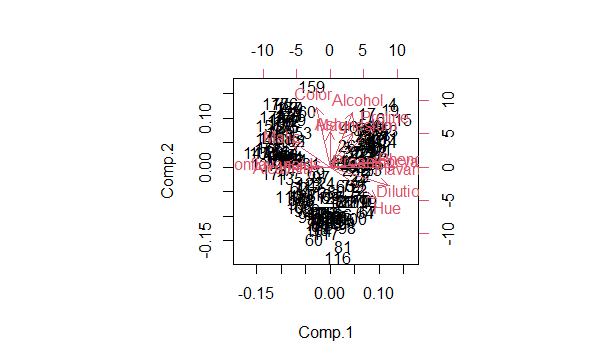
Proportion Var 0.077 0.077 0.077 0.077

Cumulative Var 0.769 0.846 0.923 1.000

plot(pcaObj) # graph showing importance of principal components

# Comp.1 having highest importance (highest variance)

biplot(pcaObj)



# cbind used to bind the data in column wise

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

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

> View(wine\_score)

>#cluster analysis

############### hierarchial clustering for principle components scores ###############

install.packages("factoextra")

library(factoextra)

library(NbClust)

library(dendextend)

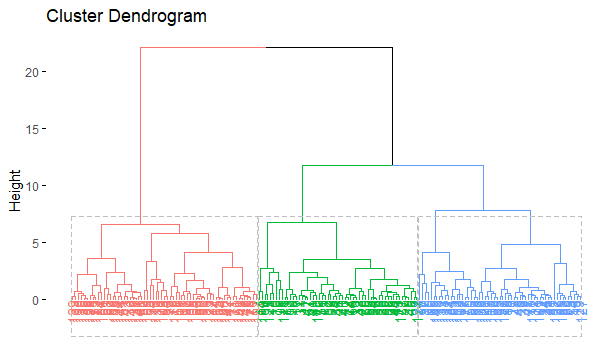
# preparing data for clustering (considering only pca scores as they represent the entire data)

clust\_data <-wine\_score[,8:10]

norm\_clust <- scale(clust\_data)

clust <- eclust(norm\_clust,"hclust",k=3,graph = FALSE)

fviz\_dend(clust,rect = TRUE)



groups <- cutree(clust,k=3)

finalhclust <- data.frame(groups,wine)

aggregate(finalhclust,by=list(clust$cluster),FUN=mean)

Group.1 groups Alcohol Malic Ash Alcalinity Magnesium Phenols

1 1 1 13.39482 2.041607 2.353393 17.50000 103.75000 2.881429

2 2 2 12.84105 2.065789 2.332456 19.50351 102.05263 2.263860

3 3 3 12.80092 2.827538 2.407692 21.20615 94.26154 1.817385

Flavanoids Nonflavanoids Proanthocyanins Color Hue Dilution Proline

1 3.000893 0.2691071 2.169286 5.089821 1.0494643 3.162500 946.9464

2 2.141228 0.3022807 1.434912 4.392982 0.9921053 2.710526 716.4386

3 1.094000 0.4940000 1.229385 5.614000 0.8477846 2.050462 601.2462

>> #\*\*\*\*\*\*\*\*\*\*\*\*\*\*kmeans clustering for principle components scores\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

> wss <- (nrow(norm\_clust)-1)\*sum(apply(norm\_clust,2,var))

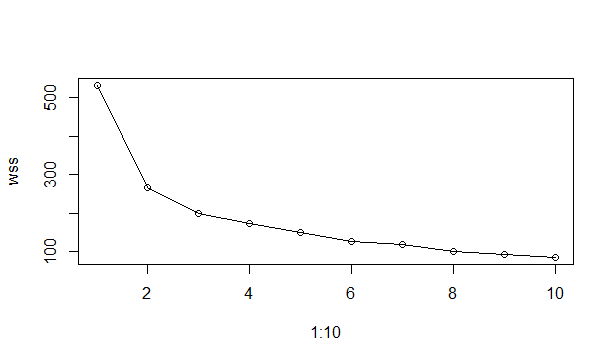
> for(i in 1:10){

+ wss[i]=sum(kmeans(norm\_clust,centers = i)$withinss)

+ }

> plot(1:10,wss, type = "o") #from scree plot no of clusters is 3

>

#alternative method

> noofclust <- NbClust(clust\_data,distance = "euclidean",method = "kmeans",min.nc = 2,max.nc = 10,index = "all")

\*\*\* : The Hubert index is a graphical method of determining the number of clusters.

In the plot of Hubert index, we seek a significant knee that corresponds to a

significant increase of the value of the measure i.e the significant peak in Hubert

index second differences plot.

\*\*\* : The D index is a graphical method of determining the number of clusters.

In the plot of D index, we seek a significant knee (the significant peak in Dindex

second differences plot) that corresponds to a significant increase of the value of

the measure.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* Among all indices:

\* 12 proposed 2 as the best number of clusters

\* 5 proposed 3 as the best number of clusters

\* 2 proposed 4 as the best number of clusters

\* 1 proposed 5 as the best number of clusters

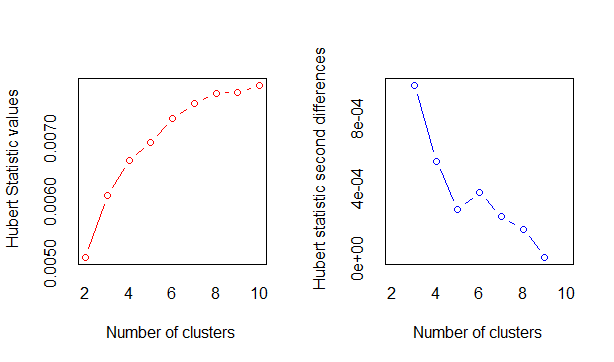
\* 1 proposed 8 as the best number of clusters

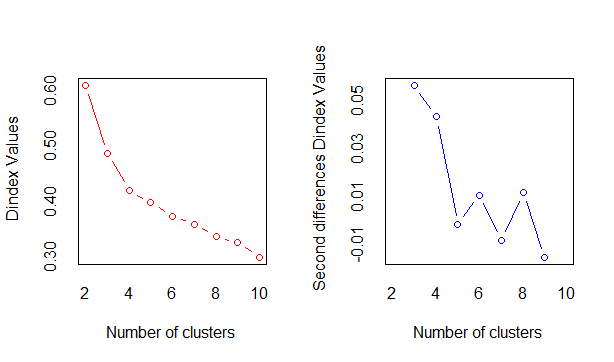
\* 1 proposed 9 as the best number of clusters

\* 2 proposed 10 as the best number of clusters

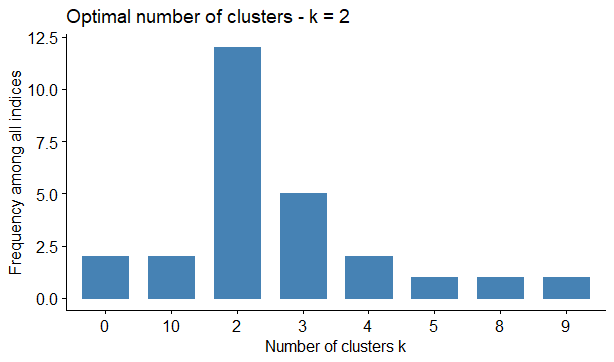
\*\*\*\*\* Conclusion \*\*\*\*\*

\* According to the majority rule, the best number of clusters is 2



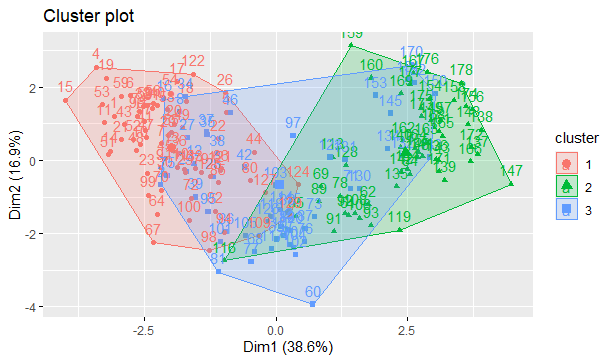


fviz\_nbclust(noofclust)



km <- kmeans(norm\_clust,3)

> fviz\_cluster(km,data = wine[-1])

>

final <- data.frame(km$cluster,wine)

> aggregate(final,by=list(km$cluster),FUN = mean)

Group.1 km.cluster Alcohol Malic Ash Alcalinity Magnesium Phenols

1 1 1 13.30913 2.024638 2.390290 18.11304 104.81159 2.844783

2 2 2 12.93093 3.006111 2.404815 20.93704 94.92593 1.747037

3 3 3 12.68200 2.069818 2.299091 19.81273 98.10909 2.143636

Flavanoids Nonflavanoids Proanthocyanins Color Hue Dilution Proline

1 2.9717391 0.2904348 2.117101 4.932609 1.0520290 3.135797 911.9710

2 0.9455556 0.5096296 1.193704 6.199074 0.8000000 1.915370 601.5741

3 1.9109091 0.3063636 1.320727 4.095273 0.9933818 2.637818 682.4727

>