#################################PCA#####################################

#Problem Statement: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 #clusterswith the original data

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

#Data : wine.csv

###########################################################################

wineData <- read.csv(file.choose()) #wine.csv

wineData <- na.omit(wineData)

head(wineData)

tail(wineData)

View(wineData)

str(wineData)

summary(wineData)

#Type Alcohol Malic Ash Alcalinity Magnesium

#Min. :1.000 Min. :11.03 Min. :0.740 Min. :1.360 Min. :10.60 Min. : 70.00

#1st Qu.:1.000 1st Qu.:12.36 1st Qu.:1.603 1st Qu.:2.210 1st Qu.:17.20 1st Qu.: 88.00

#Median :2.000 Median :13.05 Median :1.865 Median :2.360 Median :19.50 Median : 98.00

#Mean :1.938 Mean :13.00 Mean :2.336 Mean :2.367 Mean :19.49 Mean : 99.74

#3rd Qu.:3.000 3rd Qu.:13.68 3rd Qu.:3.083 3rd Qu.:2.558 3rd Qu.:21.50 3rd Qu.:107.00

#Max. :3.000 Max. :14.83 Max. :5.800 Max. :3.230 Max. :30.00 Max. :162.00

#Phenols Flavanoids Nonflavanoids Proanthocyanins Color Hue

#Min. :0.980 Min. :0.340 Min. :0.1300 Min. :0.410 Min. : 1.280 Min. :0.4800

#1st Qu.:1.742 1st Qu.:1.205 1st Qu.:0.2700 1st Qu.:1.250 1st Qu.: 3.220 1st Qu.:0.7825

#Median :2.355 Median :2.135 Median :0.3400 Median :1.555 Median : 4.690 Median :0.9650

#Mean :2.295 Mean :2.029 Mean :0.3619 Mean :1.591 Mean : 5.058 Mean :0.9574

#3rd Qu.:2.800 3rd Qu.:2.875 3rd Qu.:0.4375 3rd Qu.:1.950 3rd Qu.: 6.200 3rd Qu.:1.1200

#Max. :3.880 Max. :5.080 Max. :0.6600 Max. :3.580 Max. :13.000 Max. :1.7100

#Dilution Proline

#Min. :1.270 Min. : 278.0

#1st Qu.:1.938 1st Qu.: 500.5

#Median :2.780 Median : 673.5

#Mean :2.612 Mean : 746.9

#3rd Qu.:3.170 3rd Qu.: 985.0

#Max. :4.000 Max. :1680.0

install.packages("lattice")

require("lattice")

?princomp ## to understand the api for princomp

## the first column in mydata has type

View(wineData[-1])

pcaObj<-princomp(wineData[-1], cor = TRUE, scores = TRUE, covmat = NULL)

summary(pcaObj)

#Importance of components:

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

#Standard deviation 2.1692972 1.5801816 1.2025273 0.9586313 0.92370351 0.80103498 0.74231281 0.59033665

#Proportion of Variance 0.3619885 0.1920749 0.1112363 0.0706903 0.06563294 0.04935823 0.04238679 0.02680749

#Cumulative Proportion 0.3619885 0.5540634 0.6652997 0.7359900 0.80162293 0.85098116 0.89336795 0.92017544

# Comp.9 Comp.10 Comp.11 Comp.12 Comp.13

#Standard deviation 0.53747553 0.50090167 0.47517222 0.41081655 0.321524394

#Proportion of Variance 0.02222153 0.01930019 0.01736836 0.01298233 0.007952149

#Cumulative Proportion 0.94239698 0.96169717 0.97906553 0.99204785 1.000000000

**#As per the summary above (Importance of components); the first 3 variables contribute ~67% of the information required for the entire data.**

**#Hence the 13 components can be reduced to 3 for further analysis with 67% information.**

loadings(pcaObj)

# Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11 Comp.12

#Alcohol 0.144 0.484 0.207 0.266 0.214 0.396 0.509 0.212 0.226 0.266

#Malic -0.245 0.225 -0.537 0.537 -0.421 -0.309 -0.122

#Ash 0.316 -0.626 0.214 0.143 0.154 0.149 -0.170 -0.308 0.499

#Alcalinity -0.239 -0.612 -0.101 0.287 0.428 0.200 -0.479

#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 -0.320 -0.304 0.304

#Flavanoids 0.423 -0.151 -0.152 0.109 -0.187 -0.163

#Nonflavanoids -0.299 -0.170 0.203 0.501 -0.259 -0.595 -0.233 0.196 0.216 -0.117

#Proanthocyanins 0.313 -0.149 -0.399 -0.137 -0.534 -0.372 0.368 -0.209 0.134 0.237

#Color 0.530 0.137 -0.419 0.228 -0.291 -0.604

#Hue 0.297 -0.279 0.428 0.174 0.106 -0.232 0.437 -0.522 -0.259

#Dilution 0.376 -0.164 -0.166 -0.184 0.101 0.266 0.137 0.524 -0.601

#Proline 0.287 0.365 0.127 0.232 0.158 0.120 0.120 -0.576 0.162 -0.539

#Comp.13

#Alcohol

#Malic

#Ash -0.141

#Alcalinity

#Magnesium

#Phenols -0.464

#Flavanoids 0.832

#Nonflavanoids 0.114

#Proanthocyanins -0.117

#Color

#Hue

#Dilution -0.157

#Proline

# Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11 Comp.12

#SS loadings 1.000 1.000 1.000 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 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 0.769 0.846 0.923

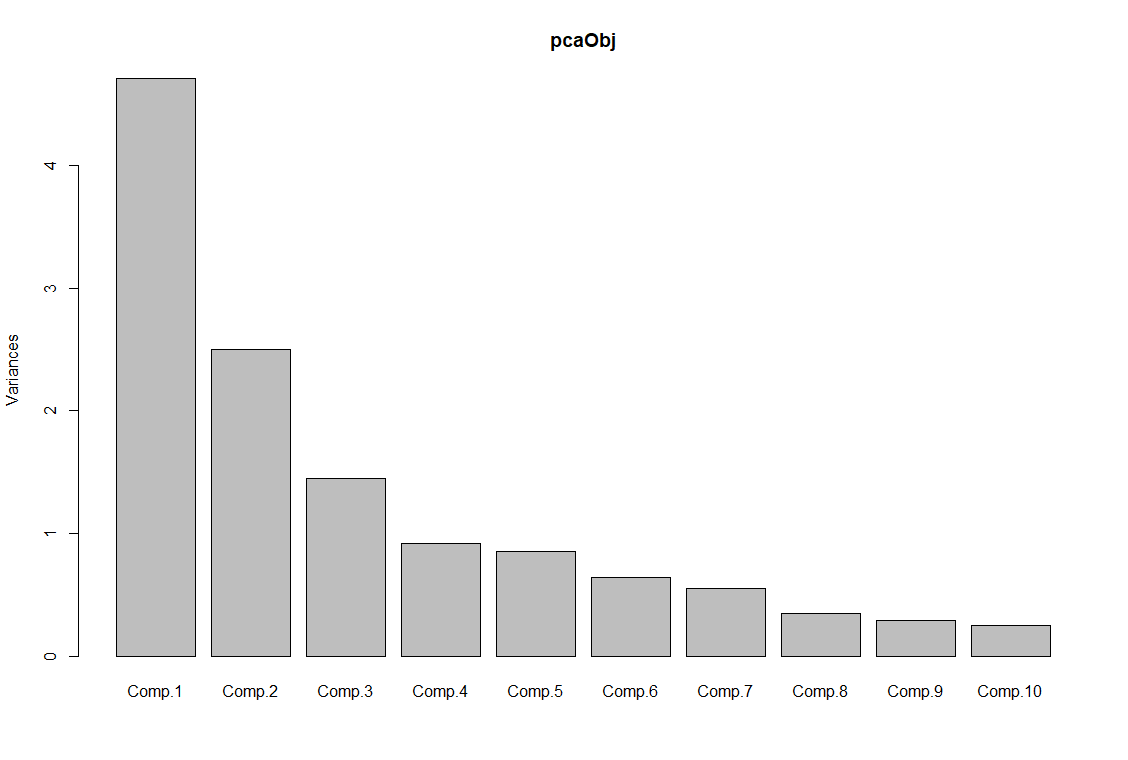
# Comp.13

#SS loadings 1.000

#Proportion Var 0.077

#Cumulative Var 1.000

plot(pcaObj) # graph showing importance of principal components



# Comp.1 having highest importance (highest variance)

pcaObj$scores #- this is to check the scores of your principal components

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

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

View(wineDataNScore)

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

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

View(clus\_data)

#############Performing Hierarchical Clustering on reduced dimension after applying pca########

# Normalizing the data

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

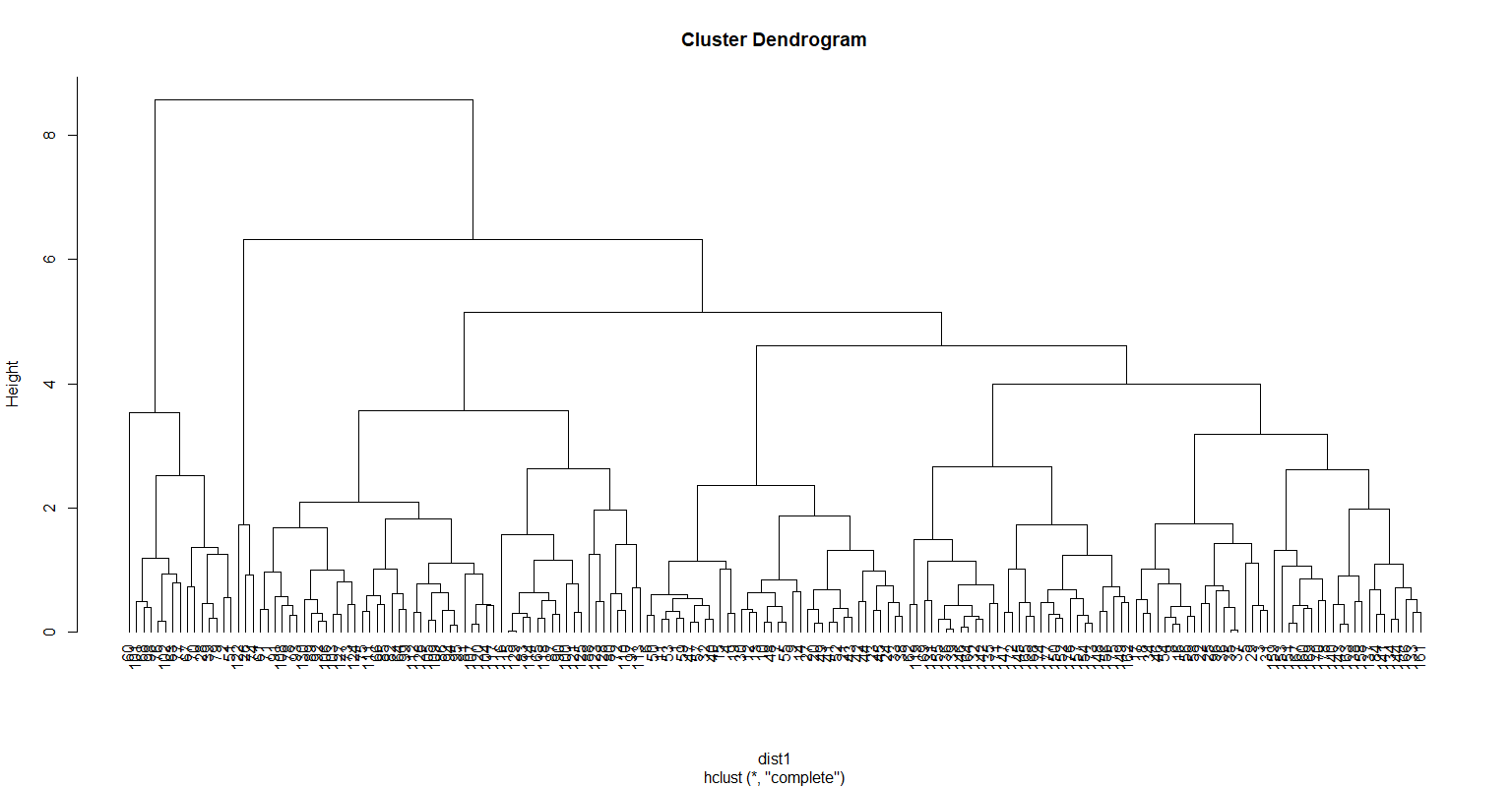
dist1<-dist(norm\_clus,method = "euclidean") # method for finding the distance

# here I am considering Euclidean distance

# Clustering the data using hclust function --> Hierarchical

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

plot(fit1, hang=-1)



# Displaying Dendrogram

groups<-cutree(fit1,7)

# Cutting the dendrogram for 7 clusters

membership\_1<-as.matrix(groups)

# cluster numbering

View(membership\_1)

final1<-cbind(membership\_1,wineData)

# binding column wise with orginal data

View(final1)

FinalGrp <-aggregate(final1[,-c(2,16:18)],by=list(membership\_1),FUN=mean)

FinalGrp

View(FinalGrp) # Inferences can be

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

#Group.1 membership\_1 Alcohol Malic Ash Alcalinity Magnesium Phenols Flavanoids Nonflavanoids

#1 1 13.89333 2.000556 2.373056 16.15000 104.91667 2.907500 3.0808333 0.2736111

#2 2 12.68933 1.392667 1.882000 15.66000 97.46667 2.192000 2.0773333 0.2680000

#3 3 13.35525 2.775750 2.602750 21.26750 106.00000 2.277250 1.8115000 0.3950000

#4 4 12.53333 1.923333 3.016667 27.83333 127.33333 3.036667 3.5500000 0.3833333

#5 5 12.24970 2.005758 2.210303 19.92121 90.27273 2.258788 2.0572727 0.3572727

#6 6 13.11452 3.159032 2.308387 19.71935 98.48387 1.610323 0.7864516 0.4493548

#7 7 12.05050 2.102000 2.496000 22.54500 93.05000 2.316000 2.1880000 0.3935000

# Proanthocyanins Color Hue Dilution Proline

#1 1.979722 5.747222 1.0397222 3.218333 1132.4167

#2 1.584667 3.884000 1.1460000 2.745333 710.3333

#3 1.527750 6.163000 0.9072500 2.368000 829.0500

#4 1.916667 4.310000 1.1233333 3.463333 760.0000

#5 1.625455 2.991515 0.9841212 2.830606 504.6667

#6 1.064194 7.247742 0.6745161 1.672581 639.4194

#7 1.732500 2.616500 1.1380000 2.873500 480.3500

#######Performing hierarchial clustering on original wines data ###########################

# Normalizing the data

normWineData<-scale(wineData[-1]) # Scale function is used to normalize data

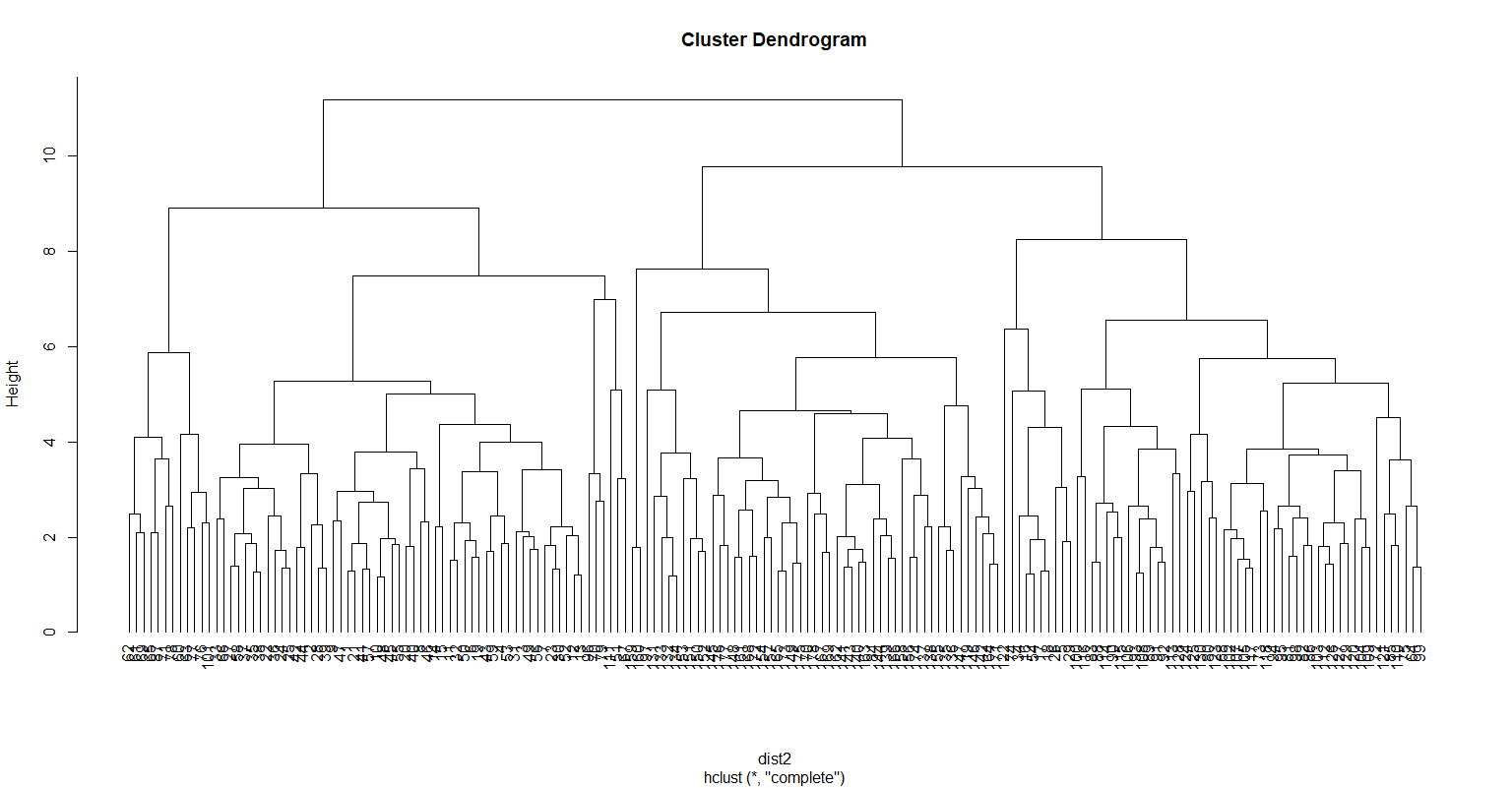
dist2<-dist(normWineData,method = "euclidean") # method for finding the distance

# here I am considering Euclidean distance

# Clustering the data using hclust function --> Hierarchical

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

plot(fit2, hang=-1)



# Displaying Dendrogram

groups<-cutree(fit2,7)

# Cutting the dendrogram for 7 clusters

membership\_2<-as.matrix(groups)

# cluster numbering

View(membership\_2)

final2<-cbind(membership\_2,wineData)

# binding column wise with orginal data

View(final2)

FinalGrp2 <-aggregate(final2[,-c(2,16:18)],by=list(membership\_2),FUN=mean)

FinalGrp2

View(FinalGrp2)

#Group.1 membership\_2 Alcohol Malic Ash Alcalinity Magnesium Phenols Flavanoids Nonflavanoids

# 1 1 13.89333 2.000556 2.373056 16.15000 104.91667 2.907500 3.0808333 0.2736111

# 2 2 12.68933 1.392667 1.882000 15.66000 97.46667 2.192000 2.0773333 0.2680000

# 3 3 13.35525 2.775750 2.602750 21.26750 106.00000 2.277250 1.8115000 0.3950000

# 4 4 12.53333 1.923333 3.016667 27.83333 127.33333 3.036667 3.5500000 0.3833333

# 5 5 12.24970 2.005758 2.210303 19.92121 90.27273 2.258788 2.0572727 0.3572727

# 6 6 13.11452 3.159032 2.308387 19.71935 98.48387 1.610323 0.7864516 0.4493548

# 7 7 12.05050 2.102000 2.496000 22.54500 93.05000 2.316000 2.1880000 0.3935000

#Proanthocyanins Color Hue Dilution Proline

#1 1.979722 5.747222 1.0397222 3.218333 1132.4167

#2 1.584667 3.884000 1.1460000 2.745333 710.3333

#3 1.527750 6.163000 0.9072500 2.368000 829.0500

#4 1.916667 4.310000 1.1233333 3.463333 760.0000

#5 1.625455 2.991515 0.9841212 2.830606 504.6667

#6 1.064194 7.247742 0.6745161 1.672581 639.4194

#7 1.732500 2.616500 1.1380000 2.873500 480.3500

#################################Performing Kmeans clustering on original wine data###

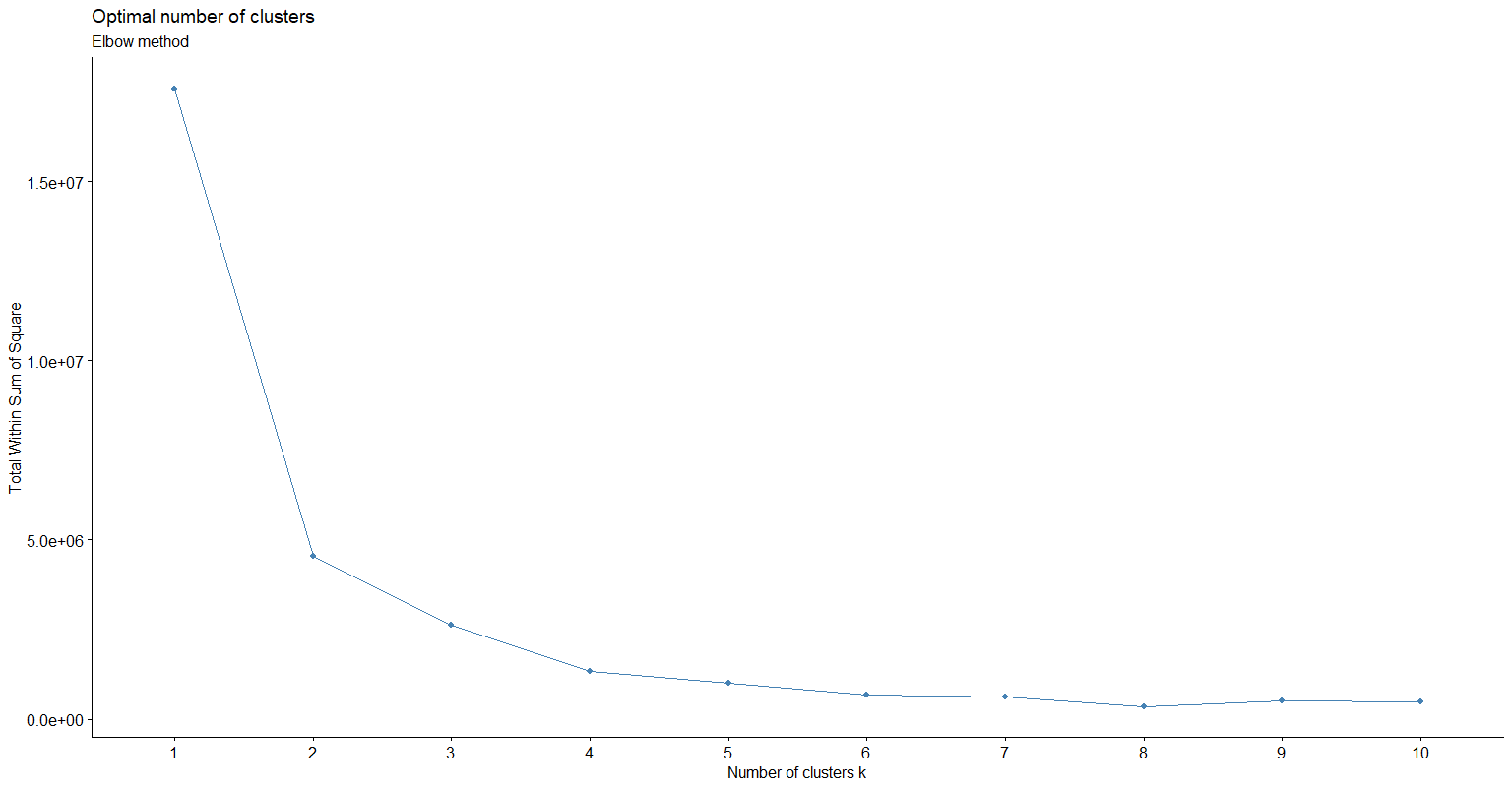
#elbow curve & k ~ sqrt(n/2) to decide the k value

install.packages("factoextra")

library(factoextra)

fviz\_nbclust(wineData[-1],kmeans,method="wss")+labs(subtitle = "Elbow method")

#Seeing the elbow chart we can say 7 is the optimal k value.



# k clustering for large dataset - Clustering Large Applications (Clara)

install.packages("cluster")

library(cluster)

?clara

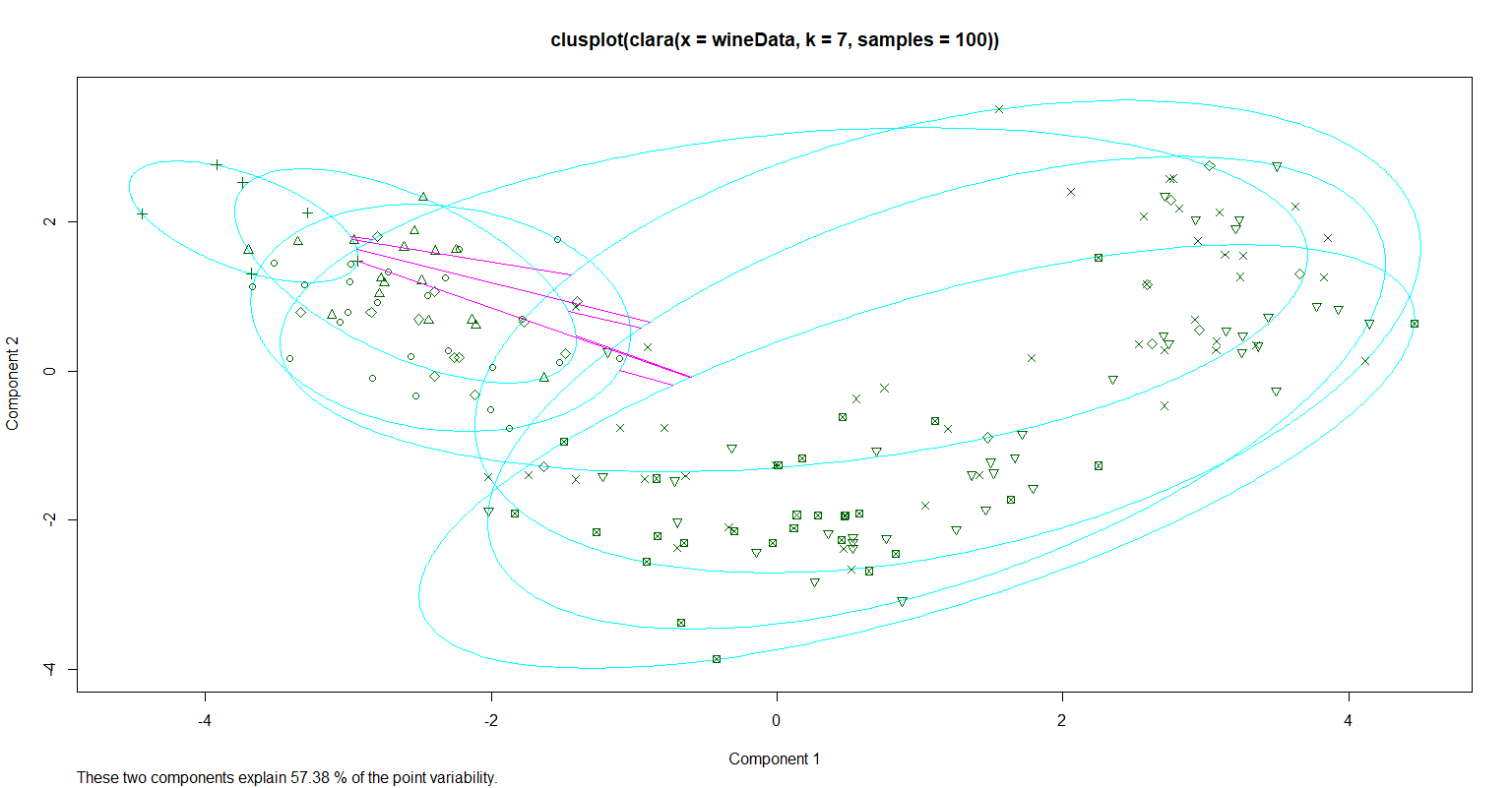
xcl <- clara(wineData, 7, sample = 100)

clusplot(xcl)

#Partitioning around medoids

xpm <- pam(wineData, 7)

clusplot(xpm)



#################################Performing Kmeans clustering after applying pca dimension reduction###

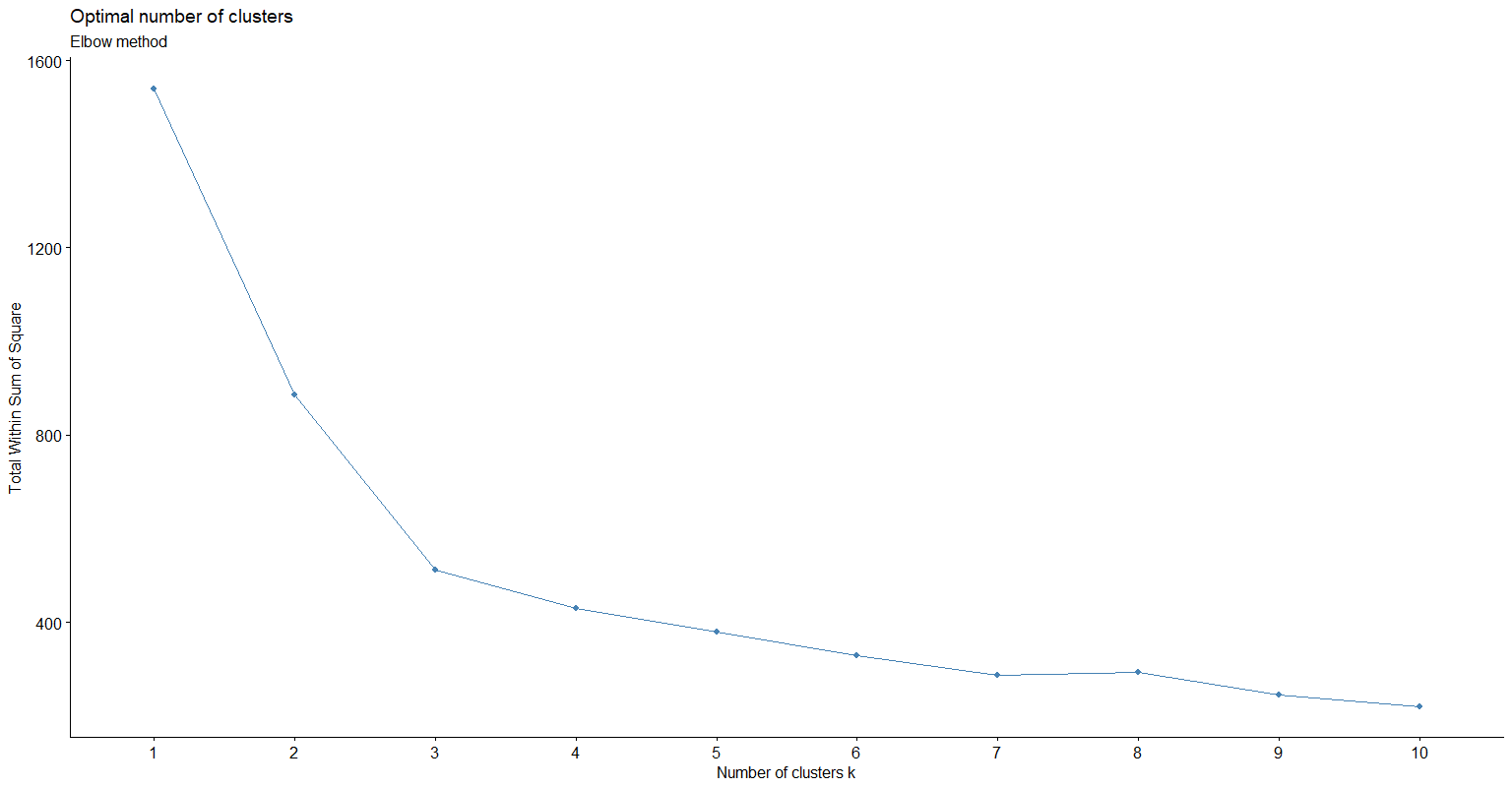
#elbow curve & k ~ sqrt(n/2) to decide the k value

install.packages("factoextra")

library(factoextra)

fviz\_nbclust(clus\_data,kmeans,method="wss")+labs(subtitle = "Elbow method")

#Seeing the elbow chart we can say 7 is the optimal k value.



# k clustering for large dataset - Clustering Large Applications (Clara)

install.packages("cluster")

library(cluster)

?clara

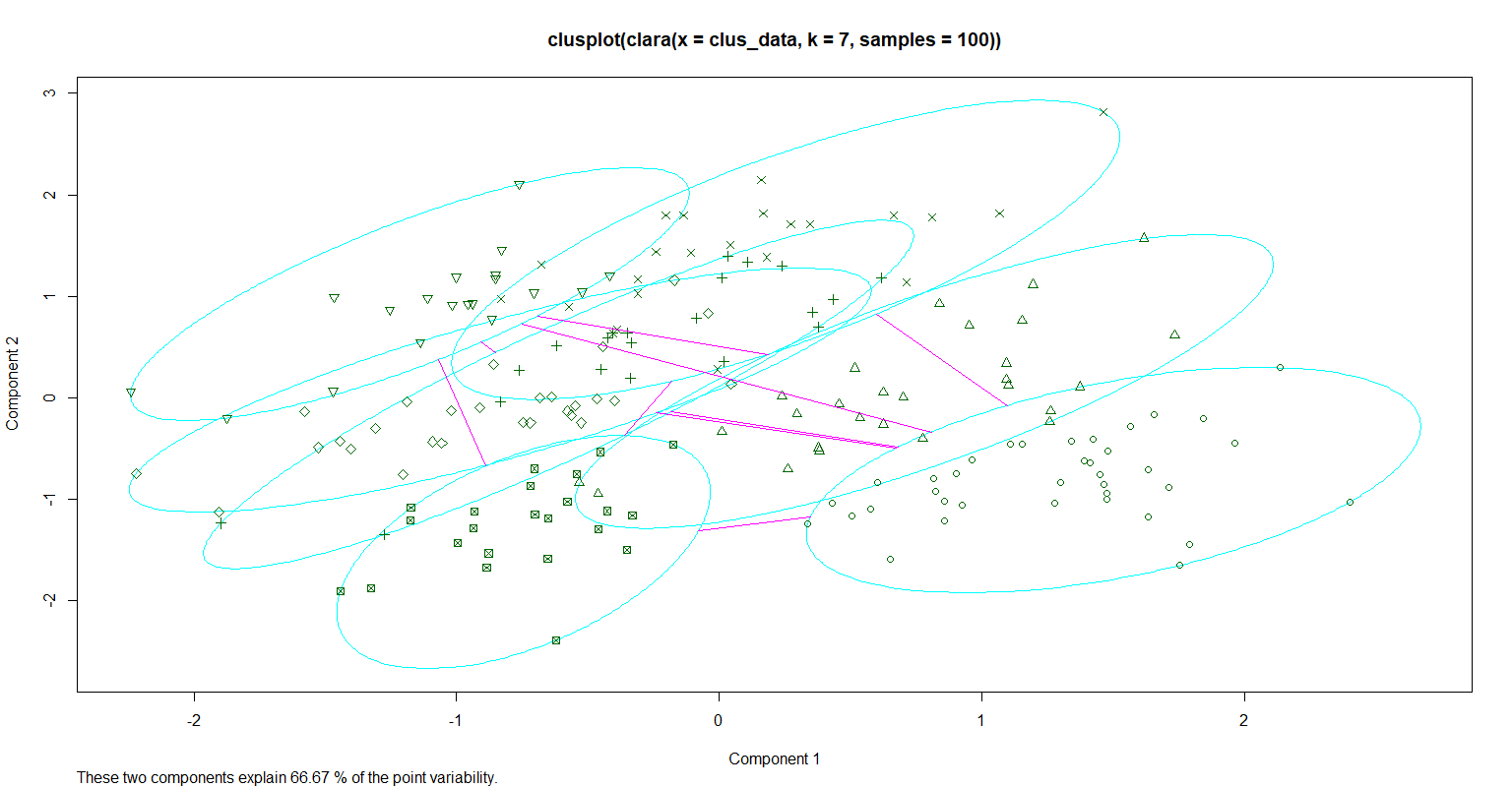
xcl <- clara(clus\_data, 7, sample = 100)

clusplot(xcl)

#Partitioning around medoids

xpm <- pam(clus\_data, 7)

clusplot(xpm)



**###Conclusion: When PCA was applied on the entire set of variables (13); PCA suggested that 67% of the information can be inferred from the first 3 variables. We then plotted dendrogram for both 13 variables and 3 variables data and found that the number of clustered required are 7 and the dendrogram seem identical.**