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

**Soln->**

> inp<-read.csv(file.choose())

> dat<-inp[,-1]

> attach(inp)

The following objects are masked from inp (pos = 5):

Alcalinity, Alcohol, Ash, Color, Dilution, Flavanoids, Hue, Magnesium, Malic,

Nonflavanoids, Phenols, Proanthocyanins, Proline, Type

> x<-cor(inp)

>

> pcaObj<- princomp(dat,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 = dat, 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 Comp.6 Comp.7

Standard deviation 2.1692972 1.5801816 1.2025273 0.9586313 0.92370351 0.80103498 0.74231281

Proportion of Variance 0.3619885 0.1920749 0.1112363 0.0706903 0.06563294 0.04935823 0.04238679

Cumulative Proportion 0.3619885 0.5540634 0.6652997 0.7359900 0.80162293 0.85098116 0.89336795

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

Standard deviation 0.59033665 0.53747553 0.50090167 0.47517222 0.41081655 0.321524394

Proportion of Variance 0.02680749 0.02222153 0.01930019 0.01736836 0.01298233 0.007952149

Cumulative Proportion 0.92017544 0.94239698 0.96169717 0.97906553 0.99204785 1.000000000

> 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 = dat, cor = TRUE, scores = TRUE, covmat = NULL)

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

> loadings(pcaObj)

Loadings:

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

> pcaObj$loadings

Loadings:

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

>

> #following is done to resolve the error->Error in plot.new() : figure margins too large

> par("mar")

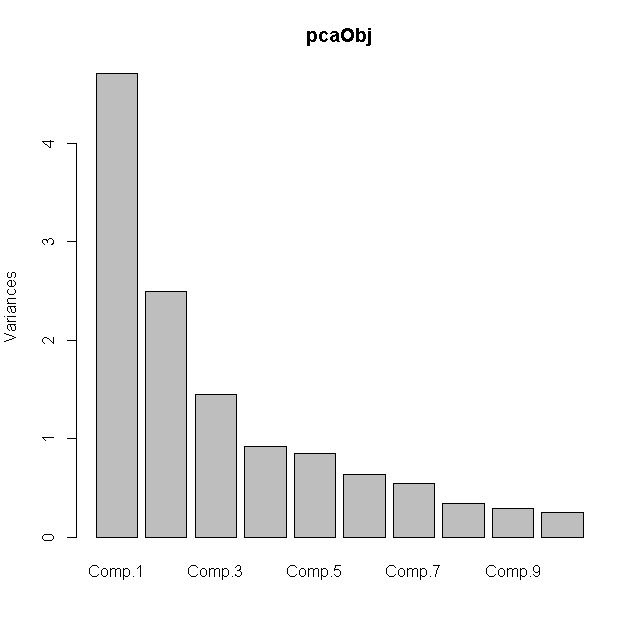
[1] 5.1 4.1 4.1 2.1

> par(mar=c(1,1,1,1))

>

> windows()

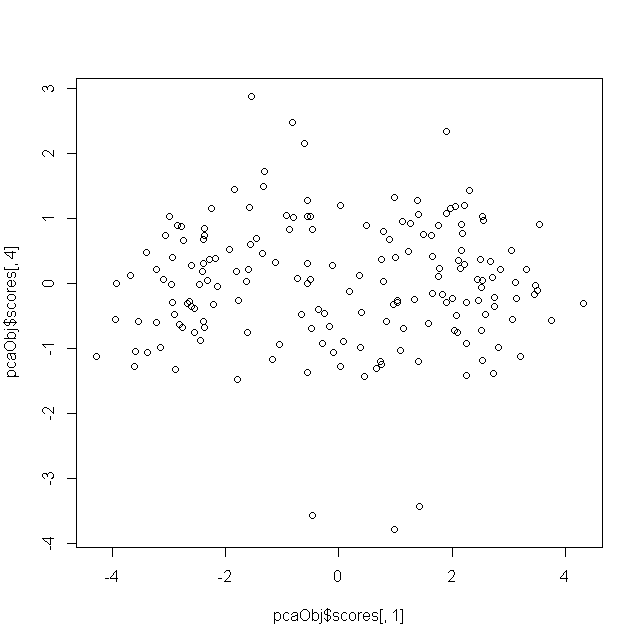
> plot(pcaObj)



>

> windows()

> plot(pcaObj$scores[,1],pcaObj$scores[,4])



> cor(pcaObj$scores)

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

Comp.1 1.000000e+00 1.084907e-16 5.424530e-16 -1.726336e-16 5.091231e-16 1.545839e-16

Comp.2 1.084907e-16 1.000000e+00 6.946481e-16 -2.049219e-16 -2.683303e-16 -5.179499e-16

Comp.3 5.424530e-16 6.946481e-16 1.000000e+00 1.113971e-16 -1.131520e-15 2.284362e-16

Comp.4 -1.726336e-16 -2.049219e-16 1.113971e-16 1.000000e+00 -3.646699e-16 2.763785e-17

Comp.5 5.091231e-16 -2.683303e-16 -1.131520e-15 -3.646699e-16 1.000000e+00 5.089835e-17

Comp.6 1.545839e-16 -5.179499e-16 2.284362e-16 2.763785e-17 5.089835e-17 1.000000e+00

Comp.7 -9.989661e-17 -2.196702e-16 2.574754e-16 -6.057405e-16 2.140808e-16 3.298568e-16

Comp.8 -2.194311e-16 -1.041895e-15 3.193601e-16 -4.991080e-16 -2.598673e-17 9.074127e-16

Comp.9 -4.919735e-16 -1.037072e-15 8.148581e-17 -2.829802e-16 -2.499345e-16 2.112365e-16

Comp.10 -5.648958e-17 -5.789146e-16 -1.852401e-16 -2.169106e-16 3.171417e-16 1.058383e-15

Comp.11 -1.792457e-15 -1.006925e-15 4.320010e-17 -1.123866e-15 2.557922e-16 -4.755083e-16

Comp.12 -1.763222e-15 -6.246918e-16 1.583423e-18 -6.137930e-17 -6.460898e-16 -8.416251e-16

Comp.13 6.787802e-17 -3.060702e-16 -9.612507e-16 -2.191582e-15 -2.595143e-16 4.205364e-16

Comp.7 Comp.8 Comp.9 Comp.10 Comp.11 Comp.12

Comp.1 -9.989661e-17 -2.194311e-16 -4.919735e-16 -5.648958e-17 -1.792457e-15 -1.763222e-15

Comp.2 -2.196702e-16 -1.041895e-15 -1.037072e-15 -5.789146e-16 -1.006925e-15 -6.246918e-16

Comp.3 2.574754e-16 3.193601e-16 8.148581e-17 -1.852401e-16 4.320010e-17 1.583423e-18

Comp.4 -6.057405e-16 -4.991080e-16 -2.829802e-16 -2.169106e-16 -1.123866e-15 -6.137930e-17

Comp.5 2.140808e-16 -2.598673e-17 -2.499345e-16 3.171417e-16 2.557922e-16 -6.460898e-16

Comp.6 3.298568e-16 9.074127e-16 2.112365e-16 1.058383e-15 -4.755083e-16 -8.416251e-16

Comp.7 1.000000e+00 1.410933e-15 -1.793242e-15 -5.764857e-16 2.386578e-15 -1.359376e-15

Comp.8 1.410933e-15 1.000000e+00 -1.168831e-15 -1.010880e-15 4.309294e-16 -6.996814e-16

Comp.9 -1.793242e-15 -1.168831e-15 1.000000e+00 -1.589646e-15 4.246101e-15 6.278407e-16

Comp.10 -5.764857e-16 -1.010880e-15 -1.589646e-15 1.000000e+00 -5.814421e-16 -2.247602e-15

Comp.11 2.386578e-15 4.309294e-16 4.246101e-15 -5.814421e-16 1.000000e+00 7.521650e-16

Comp.12 -1.359376e-15 -6.996814e-16 6.278407e-16 -2.247602e-15 7.521650e-16 1.000000e+00

Comp.13 -9.512075e-17 3.405128e-16 5.966436e-16 -1.184333e-16 2.442285e-15 -1.680463e-15

Comp.13

Comp.1 6.787802e-17

Comp.2 -3.060702e-16

Comp.3 -9.612507e-16

Comp.4 -2.191582e-15

Comp.5 -2.595143e-16

Comp.6 4.205364e-16

Comp.7 -9.512075e-17

Comp.8 3.405128e-16

Comp.9 5.966436e-16

Comp.10 -1.184333e-16

Comp.11 2.442285e-15

Comp.12 -1.680463e-15

Comp.13 1.000000e+00

>

>

> inp1<-cbind(dat,pcaObj$scores[,1:4])

> clus\_dat<-inp1[,14:17]

>

> norm\_clus<-scale(clus\_dat)

>

> #to determine clusters we will use the kselection

> library(kselection)

> k<-kselection(norm\_clus,parallel = TRUE,k\_threshold = 0.95,max\_centers = 20)

> k

f(k) finds 20 clusters>

> #lets use the graphical Method to determine the clusters better

>

> twss=NULL

> for (i in 1:30)

+ {

+ twss[i]=sum(kmeans(norm\_clus,i)$tot.withinss)

+

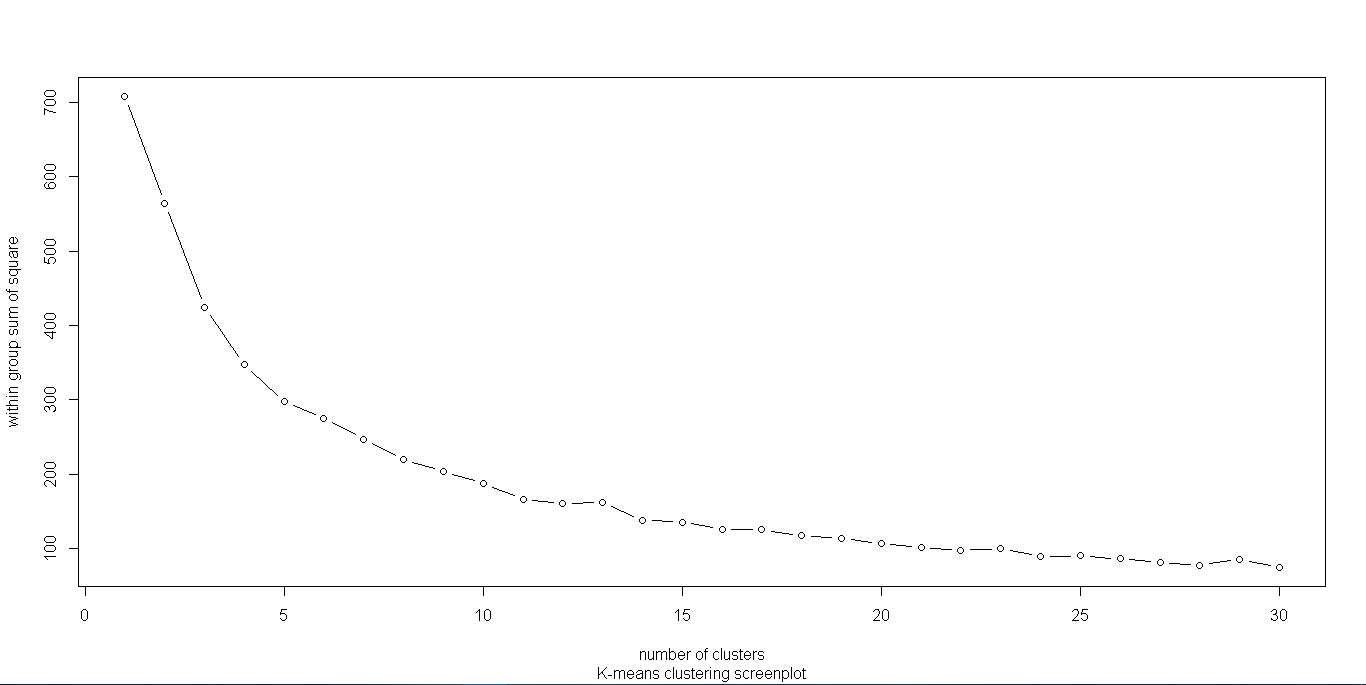
+ }

>

> windows()

> plot(1:30,twss,type = "b",xlab = "number of clusters",ylab = "within group sum of square")

> title(sub="K-means clustering screenplot")



> #the first elbow is at 5 and next is at 8

> #thus we will go for k =5

>

> d<-dist(norm\_clus,method="euclidean")

> fit<-hclust(d,method="complete")

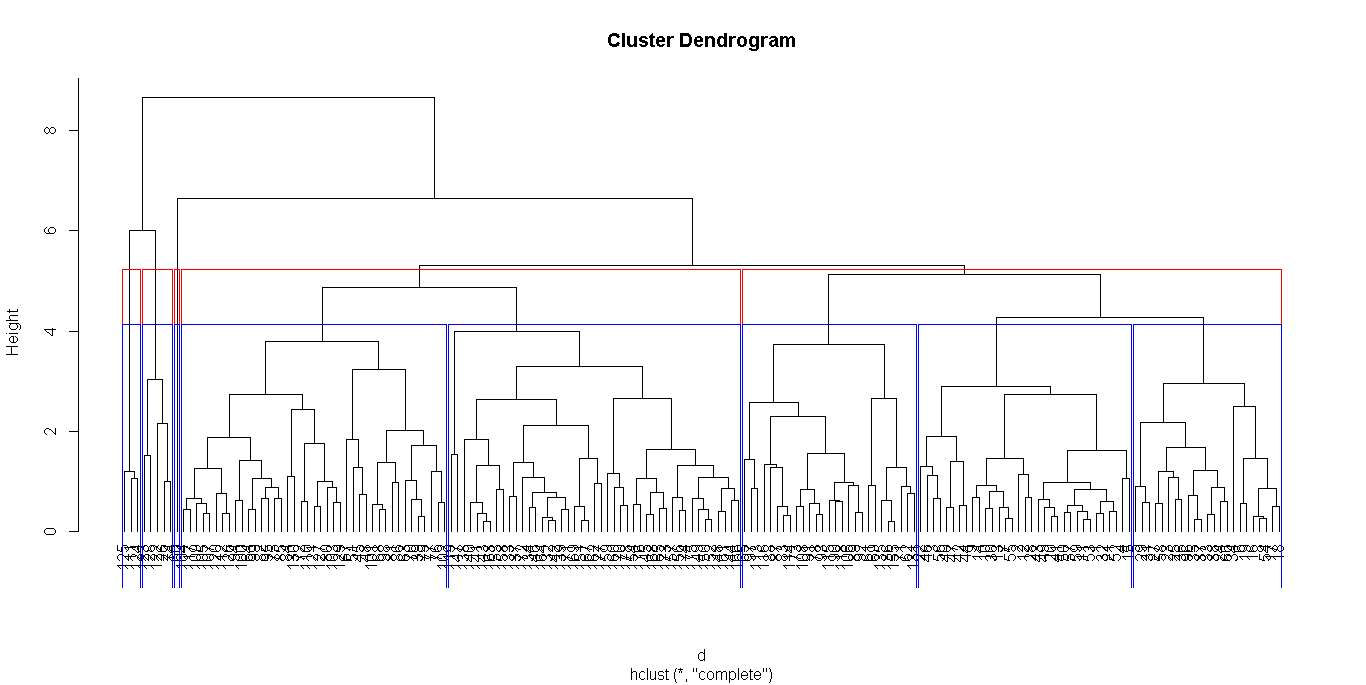
>

> windows()

> plot(fit,hang=-1)

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

> rect.hclust(fit,k=8,border="blue")



> #from the dedogram we can see the clusters are better for value of 8 than 5

>

> #using Heirarchical method to find clusters

> groups<- cutree(fit,k=8)

>

> member<-as.matrix(groups) # groups or cluster numbers

> final <- data.frame(member,dat)

>

> View(final)

>

> write.csv(final, file="finale.csv",row.names = F)

> aggregate(final[,-1],by = list(final$member),mean)

Group.1 Alcohol Malic Ash Alcalinity Magnesium Phenols Flavanoids Nonflavanoids

1 1 13.89394 2.180000 2.374848 16.23636 105.51515 2.886667 3.0278788 0.2636364

2 2 13.52870 1.769130 2.595217 18.28696 110.00000 2.755652 2.8686957 0.3343478

3 3 12.36200 2.466000 2.912000 27.70000 115.20000 2.688000 3.0040000 0.4320000

4 4 12.42951 1.820488 2.129024 18.89512 92.82927 2.444390 2.3273171 0.3017073

5 5 12.37000 0.940000 1.360000 10.60000 88.00000 1.980000 0.5700000 0.2800000

6 6 12.24741 1.643333 2.353333 20.66296 94.59259 1.851481 1.4944444 0.4637037

7 7 12.12667 4.616667 2.113333 20.66667 91.66667 2.886667 2.7533333 0.2500000

8 8 13.19111 3.491333 2.446444 21.55556 98.73333 1.680000 0.7931111 0.4431111

Proanthocyanins Color Hue Dilution Proline

1 2.011515 5.748182 1.0263636 3.227273 1118.6970

2 1.768261 5.133043 1.1269565 3.011304 1115.4348

3 1.844000 3.602000 1.0520000 3.190000 622.2000

4 1.760732 3.371951 1.0536585 2.960732 554.0244

5 0.420000 1.950000 1.0500000 1.820000 520.0000

6 1.264074 3.404074 1.0461481 2.348519 545.0000

7 2.833333 2.766667 0.7433333 3.183333 440.6667

8 1.148222 7.426000 0.6811111 1.711111 622.0444

>

> #Non herachical clusters

>

> fit1<-kmeans(norm\_clus,8)

> final2<- data.frame(fit1$cluster,dat)

> write.csv(final2, file="final2.csv",row.names = F)

> aggregate(final[,-1],by = list(final$member),mean)

Group.1 Alcohol Malic Ash Alcalinity Magnesium Phenols Flavanoids Nonflavanoids

1 1 13.89394 2.180000 2.374848 16.23636 105.51515 2.886667 3.0278788 0.2636364

2 2 13.52870 1.769130 2.595217 18.28696 110.00000 2.755652 2.8686957 0.3343478

3 3 12.36200 2.466000 2.912000 27.70000 115.20000 2.688000 3.0040000 0.4320000

4 4 12.42951 1.820488 2.129024 18.89512 92.82927 2.444390 2.3273171 0.3017073

5 5 12.37000 0.940000 1.360000 10.60000 88.00000 1.980000 0.5700000 0.2800000

6 6 12.24741 1.643333 2.353333 20.66296 94.59259 1.851481 1.4944444 0.4637037

7 7 12.12667 4.616667 2.113333 20.66667 91.66667 2.886667 2.7533333 0.2500000

8 8 13.19111 3.491333 2.446444 21.55556 98.73333 1.680000 0.7931111 0.4431111

Proanthocyanins Color Hue Dilution Proline

1 2.011515 5.748182 1.0263636 3.227273 1118.6970

2 1.768261 5.133043 1.1269565 3.011304 1115.4348

3 1.844000 3.602000 1.0520000 3.190000 622.2000

4 1.760732 3.371951 1.0536585 2.960732 554.0244

5 0.420000 1.950000 1.0500000 1.820000 520.0000

6 1.264074 3.404074 1.0461481 2.348519 545.0000

7 2.833333 2.766667 0.7433333 3.183333 440.6667

8 1.148222 7.426000 0.6811111 1.711111 622.0444

**The Cluster details are given in the excel files which were saved in the directory and are shared accordingly**