**PCA**

**Q). 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**:

> mydata <- read.csv("wine.csv")

> head(mydata)

Type Alcohol Malic Ash Alcalinity Magnesium Phenols Flavanoids

1 1 14.23 1.71 2.43 15.6 127 2.80 3.06

2 1 13.20 1.78 2.14 11.2 100 2.65 2.76

3 1 13.16 2.36 2.67 18.6 101 2.80 3.24

4 1 14.37 1.95 2.50 16.8 113 3.85 3.49

5 1 13.24 2.59 2.87 21.0 118 2.80 2.69

6 1 14.20 1.76 2.45 15.2 112 3.27 3.39

Nonflavanoids Proanthocyanins Color Hue Dilution Proline

1 0.28 2.29 5.64 1.04 3.92 1065

2 0.26 1.28 4.38 1.05 3.40 1050

3 0.30 2.81 5.68 1.03 3.17 1185

4 0.24 2.18 7.80 0.86 3.45 1480

5 0.39 1.82 4.32 1.04 2.93 735

6 0.34 1.97 6.75 1.05 2.85 1450

**Excluding The class column as mentioned in the problem statement.**

> data <- mydata[,-1]

> head(data)

Alcohol Malic Ash Alcalinity Magnesium Phenols Flavanoids Nonflavanoids Proanthocyanins Color Hue

1 14.23 1.71 2.43 15.6 127 2.80 3.06 0.28 2.29 5.64 1.04

2 13.20 1.78 2.14 11.2 100 2.65 2.76 0.26 1.28 4.38 1.05

3 13.16 2.36 2.67 18.6 101 2.80 3.24 0.30 2.81 5.68 1.03

4 14.37 1.95 2.50 16.8 113 3.85 3.49 0.24 2.18 7.80 0.86

5 13.24 2.59 2.87 21.0 118 2.80 2.69 0.39 1.82 4.32 1.04

6 14.20 1.76 2.45 15.2 112 3.27 3.39 0.34 1.97 6.75 1.05

Dilution Proline

1 3.92 1065

2 3.40 1050

3 3.17 1185

4 3.45 1480

5 2.93 735

6 2.85 1450

>attach(data)

> cor(data)

Alcohol Malic Ash Alcalinity Magnesium Phenols

Alcohol 1.00000000 0.09439694 0.211544596 -0.31023514 0.27079823 0.28910112

Malic 0.09439694 1.00000000 0.164045470 0.28850040 -0.05457510 -0.33516700

Ash 0.21154460 0.16404547 1.000000000 0.44336719 0.28658669 0.12897954

Alcalinity -0.31023514 0.28850040 0.443367187 1.00000000 -0.08333309 -0.32111332

Magnesium 0.27079823 -0.05457510 0.286586691 -0.08333309 1.00000000 0.21440123

Phenols 0.28910112 -0.33516700 0.128979538 -0.32111332 0.21440123 1.00000000

Flavanoids 0.23681493 -0.41100659 0.115077279 -0.35136986 0.19578377 0.86456350

Nonflavanoids -0.15592947 0.29297713 0.186230446 0.36192172 -0.25629405 -0.44993530

Proanthocyanins 0.13669791 -0.22074619 0.009651935 -0.19732684 0.23644061 0.61241308

Color 0.54636420 0.24898534 0.258887259 0.01873198 0.19995001 -0.05513642

Hue -0.07174720 -0.56129569 -0.074666889 -0.27395522 0.05539820 0.43368134

Dilution 0.07234319 -0.36871043 0.003911231 -0.27676855 0.06600394 0.69994936

Proline 0.64372004 -0.19201056 0.223626264 -0.44059693 0.39335085 0.49811488

Flavanoids Nonflavanoids Proanthocyanins Color Hue Dilution

Alcohol 0.2368149 -0.1559295 0.136697912 0.54636420 -0.07174720 0.072343187

Malic -0.4110066 0.2929771 -0.220746187 0.24898534 -0.56129569 -0.368710428

Ash 0.1150773 0.1862304 0.009651935 0.25888726 -0.07466689 0.003911231

Alcalinity -0.3513699 0.3619217 -0.197326836 0.01873198 -0.27395522 -0.276768549

Magnesium 0.1957838 -0.2562940 0.236440610 0.19995001 0.05539820 0.066003936

Phenols 0.8645635 -0.4499353 0.612413084 -0.05513642 0.43368134 0.699949365

Flavanoids 1.0000000 -0.5378996 0.652691769 -0.17237940 0.54347857 0.787193902

Nonflavanoids -0.5378996 1.0000000 -0.365845099 0.13905701 -0.26263963 -0.503269596

Proanthocyanins 0.6526918 -0.3658451 1.000000000 -0.02524993 0.29554425 0.519067096

Color -0.1723794 0.1390570 -0.025249931 1.00000000 -0.52181319 -0.428814942

Hue 0.5434786 -0.2626396 0.295544253 -0.52181319 1.00000000 0.565468293

Dilution 0.7871939 -0.5032696 0.519067096 -0.42881494 0.56546829 1.000000000

Proline 0.4941931 -0.3113852 0.330416700 0.31610011 0.23618345 0.312761075

Proline

Alcohol 0.6437200

Malic -0.1920106

Ash 0.2236263

Alcalinity -0.4405969

Magnesium 0.3933508

Phenols 0.4981149

Flavanoids 0.4941931

Nonflavanoids -0.3113852

Proanthocyanins 0.3304167

Color 0.3161001

Hue 0.2361834

Dilution 0.3127611

Proline 1.0000000

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

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

> 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

> 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

Alcohol 0.144 0.484 0.207 0.266 0.214 0.396 0.509 0.212 0.226

Malic -0.245 0.225 -0.537 0.537 -0.421 -0.309

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

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

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

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

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

Comp.12 Comp.13

Alcohol 0.266

Malic -0.122

Ash -0.141

Alcalinity

Magnesium

Phenols 0.304 -0.464

Flavanoids 0.832

Nonflavanoids 0.114

Proanthocyanins -0.117

Color -0.604

Hue -0.259

Dilution -0.601 -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

SS loadings 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

Cumulative Var 0.077 0.154 0.231 0.308 0.385 0.462 0.538 0.615 0.692 0.769 0.846

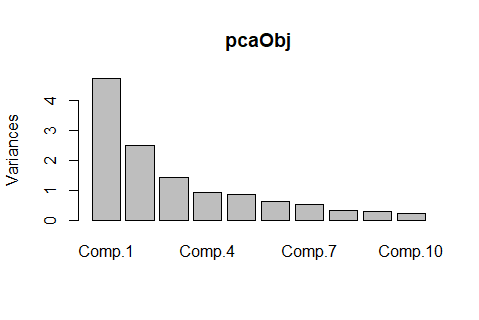
Comp.12 Comp.13

SS loadings 1.000 1.000

Proportion Var 0.077 0.077

Cumulative Var 0.923 1.000

> plot(pcaObj)



**As shown in the Plot First 3 Principal Component are carrying highest importance(Highest Variance).**

**Comp.1 having highest importance.**

**So selecting first 3 PC.**

> scores <- pcaObj$scores[,1:3]

> head(scores)

Comp.1 Comp.2 Comp.3

[1,] 3.316751 1.4434626 0.1657390

[2,] 2.209465 -0.3333929 2.0264574

[3,] 2.516740 1.0311513 -0.9828187

[4,] 3.757066 2.7563719 0.1761918

[5,] 1.008908 0.8698308 -2.0266882

[6,] 3.050254 2.1224011 0.6293958

**Binding PC columns to the mydata df.**

> mydata<-cbind(mydata,scores)

> head(mydata)

Type Alcohol Malic Ash Alcalinity Magnesium Phenols Flavanoids Nonflavanoids

1 1 14.23 1.71 2.43 15.6 127 2.80 3.06 0.28

2 1 13.20 1.78 2.14 11.2 100 2.65 2.76 0.26

3 1 13.16 2.36 2.67 18.6 101 2.80 3.24 0.30

4 1 14.37 1.95 2.50 16.8 113 3.85 3.49 0.24

5 1 13.24 2.59 2.87 21.0 118 2.80 2.69 0.39

6 1 14.20 1.76 2.45 15.2 112 3.27 3.39 0.34

Proanthocyanins Color Hue Dilution Proline Comp.1 Comp.2 Comp.3

1 2.29 5.64 1.04 3.92 1065 3.316751 1.4434626 0.1657390

2 1.28 4.38 1.05 3.40 1050 2.209465 -0.3333929 2.0264574

3 2.81 5.68 1.03 3.17 1185 2.516740 1.0311513 -0.9828187

4 2.18 7.80 0.86 3.45 1480 3.757066 2.7563719 0.1761918

5 1.82 4.32 1.04 2.93 735 1.008908 0.8698308 -2.0266882

6 1.97 6.75 1.05 2.85 1450 3.050254 2.1224011 0.6293958

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

> head(clus\_data)

Comp.1 Comp.2 Comp.3

1 3.316751 1.4434626 0.1657390

2 2.209465 -0.3333929 2.0264574

3 2.516740 1.0311513 -0.9828187

4 3.757066 2.7563719 0.1761918

5 1.008908 0.8698308 -2.0266882

6 3.050254 2.1224011 0.6293958

**Normalizing the data**

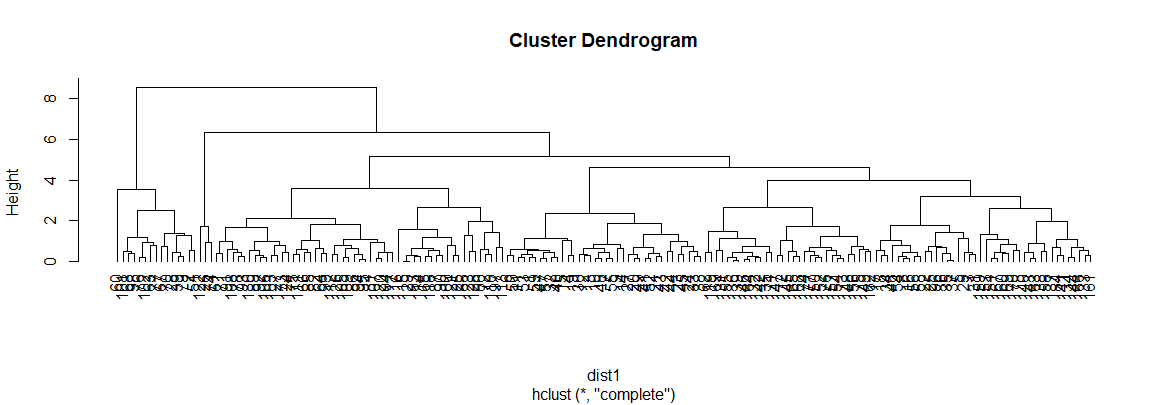
>norm\_clus<-scale(clus\_data)

>dist1<-dist(norm\_clus,method = "euclidean")

**Clustering the data using hclust function i.e. Hierarchical Clustering**

>fit1<-hclust(dist1,method="complete")

>plot(fit1, hang = -1)



> groups<-cutree(fit1,5)

**Numbering the Clusters and binding to the df.**

> membership\_1<-as.matrix(groups)

> final1<-cbind(membership\_1,mydata)

> head(final1)

membership\_1 Type Alcohol Malic Ash Alcalinity Magnesium Phenols Flavanoids

1 1 1 14.23 1.71 2.43 15.6 127 2.80 3.06

2 2 1 13.20 1.78 2.14 11.2 100 2.65 2.76

3 3 1 13.16 2.36 2.67 18.6 101 2.80 3.24

4 1 1 14.37 1.95 2.50 16.8 113 3.85 3.49

5 3 1 13.24 2.59 2.87 21.0 118 2.80 2.69

6 1 1 14.20 1.76 2.45 15.2 112 3.27 3.39

Nonflavanoids Proanthocyanins Color Hue Dilution Proline Comp.1 Comp.2 Comp.3

1 0.28 2.29 5.64 1.04 3.92 1065 3.316751 1.4434626 0.1657390

2 0.26 1.28 4.38 1.05 3.40 1050 2.209465 -0.3333929 2.0264574

3 0.30 2.81 5.68 1.03 3.17 1185 2.516740 1.0311513 -0.9828187

4 0.24 2.18 7.80 0.86 3.45 1480 3.757066 2.7563719 0.1761918

5 0.39 1.82 4.32 1.04 2.93 735 1.008908 0.8698308 -2.0266882

6 0.34 1.97 6.75 1.05 2.85 1450 3.050254 2.1224011 0.6293958

> aggregate(final1[,-c(2,13:15)],by=list(membership\_1),FUN=mean)

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

1 1 1 13.89333 2.000556 2.373056 16.15000 104.91667 2.907500 3.080833

2 2 2 12.68933 1.392667 1.882000 15.66000 97.46667 2.192000 2.077333

3 3 3 13.25014 2.943099 2.474225 20.59155 102.71831 1.986056 1.363944

4 4 4 12.53333 1.923333 3.016667 27.83333 127.33333 3.036667 3.550000

5 5 5 12.17453 2.042075 2.318113 20.91132 91.32075 2.280377 2.106604

Nonflavanoids Proanthocyanins Color Comp.1 Comp.2 Comp.3

1 0.2736111 1.979722 5.747222 2.5396039 1.0234310 0.54817457

2 0.2680000 1.584667 3.884000 0.9143818 -1.5787626 1.79152127

3 0.4187324 1.325352 6.636620 -1.4491126 1.0944081 -0.01691091

4 0.3833333 1.916667 4.310000 1.6295427 0.4847343 -4.18023005

5 0.3709434 1.665849 2.850000 -0.1347755 -1.7418727 -0.62010828