Lab 5 - Clustering Assignments

Vũ Đặng Quỳnh Giang 17133015

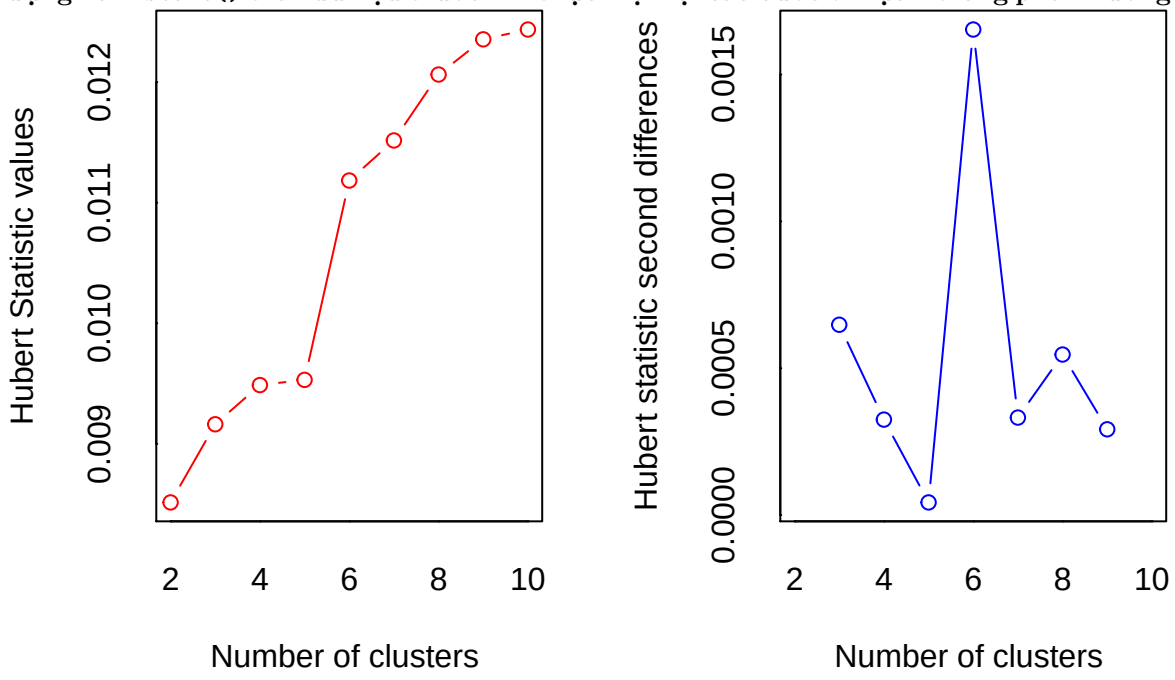
2020/12/20

Câu hỏi 1

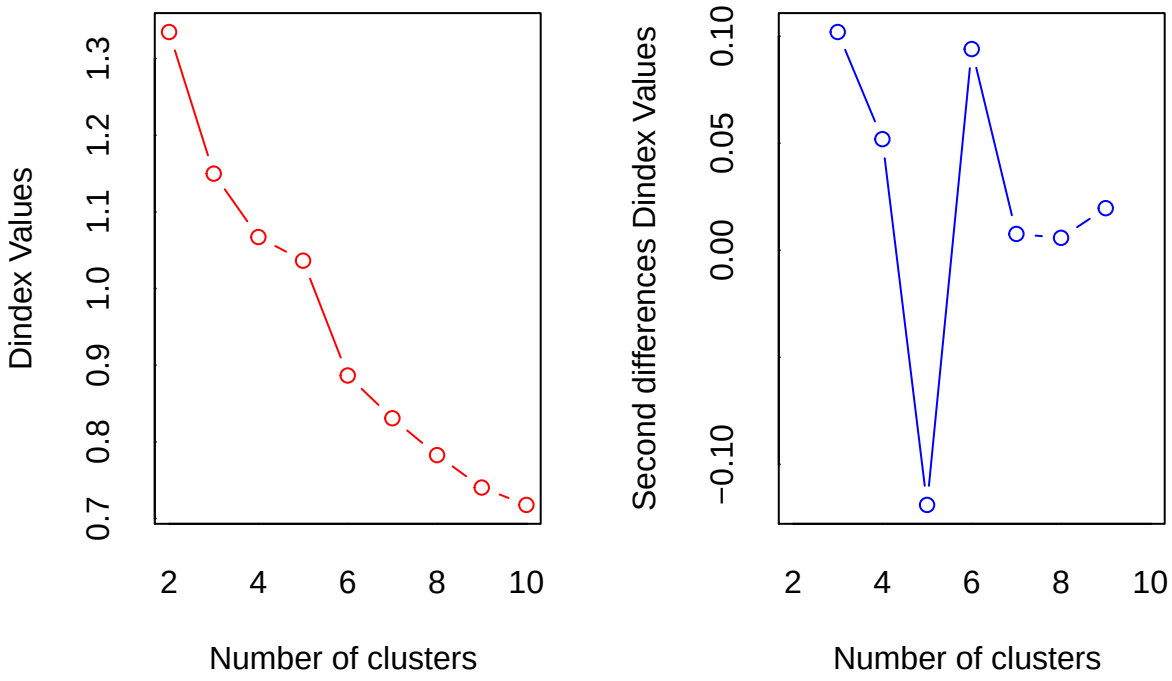
Ở phần hướng dẫn, ta thực hiện lại các thuật toán `kmeans` và `hclust` cho tập dữ liệu `USArrests`. Hãy thực hiện lại các bước đã mô tả ở trên nhưng áp dụng hàm `scale()` trên dữ liệu trước áp dụng thuật toán các `kmeans`, `hclust`

```
USA_scale <- scale(USArrests)
library(NbClust)
nb <- NbClust(USA_scale, diss=NULL, distance = "euclidean", min.nc=2, max.nc=10, method = "kmeans")
```

a. Áp dụng hàm `scale()` trên dữ liệu trước khi thực hiện lại các bước ở mục 2 trong phần hướng dẫn.




```
## *** : 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:
## * 11 proposed 2 as the best number of clusters
## * 2 proposed 3 as the best number of clusters
## * 1 proposed 4 as the best number of clusters
## * 1 proposed 5 as the best number of clusters
## * 7 proposed 6 as the best number of clusters
## * 1 proposed 9 as the best number of clusters
## * 1 proposed 10 as the best number of clusters
##
##                     ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  2
##
##
## *******************************************************************
```

```r
set.seed(17133015)
res <- kmeans(USA_scale, 2, nstart = 20)
print(res)
```

```
## K-means clustering with 2 clusters of sizes 20, 30
##
## Cluster means:
##      Murder    Assault   UrbanPop       Rape
## 1  1.004934  1.0138274  0.1975853  0.8469650
## 2 -0.669956 -0.6758849 -0.1317235 -0.5646433
##
```

```
## Clustering vector:
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              1              1              2              1
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              1              2              2              1              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              2              2              1              2              2
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              2              2              1              2              1
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              2              1              2              1              1
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              2              2              1              2              2
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              1              1              1              2              2
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              2              2              2              2              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              2              1              1              2              2
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              2              2              2              2              2
##
## Within cluster sum of squares by cluster:
## [1] 46.74796 56.11445
##  (between_SS / total_SS =  47.5 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
USArrestsWithCluster <- cbind(USA_scale, cluster = res$cluster)
USArrestsWithCluster
```

```
##                    Murder     Assault    UrbanPop         Rape cluster
## Alabama         1.24256408  0.78283935 -0.52090661 -0.003416473       1
## Alaska          0.50786248  1.10682252 -1.21176419  2.484202941       1
## Arizona         0.07163341  1.47880321  0.99898006  1.042878388       1
## Arkansas        0.23234938  0.23086801 -1.07359268 -0.184916602       2
## California      0.27826823  1.26281442  1.75892340  2.067820292       1
## Colorado        0.02571456  0.39885929  0.86080854  1.864967207       1
## Connecticut    -1.03041900 -0.72908214  0.79172279 -1.081740768       2
## Delaware       -0.43347395  0.80683810  0.44629400 -0.579946294       2
## Florida         1.74767144  1.97077766  0.99898006  1.138966691       1
## Georgia         2.20685994  0.48285493 -0.38273510  0.487701523       1
## Hawaii         -0.57123050 -1.49704226  1.20623733 -0.110181255       2
## Idaho          -1.19113497 -0.60908837 -0.79724965 -0.750769945       2
## Illinois        0.59970018  0.93883125  1.20623733  0.295524916       1
## Indiana        -0.13500142 -0.69308401 -0.03730631 -0.024769429       2
## Iowa           -1.28297267 -1.37704849 -0.58999237 -1.060387812       2
## Kansas         -0.41051452 -0.66908525  0.03177945 -0.345063775       2
## Kentucky        0.43898421 -0.74108152 -0.93542116 -0.526563903       2
## Louisiana       1.74767144  0.93883125  0.03177945  0.103348309       1
## Maine          -1.30593210 -1.05306531 -1.00450692 -1.434064548       2
## Maryland        0.80633501  1.55079947  0.10086521  0.701231086       1
```

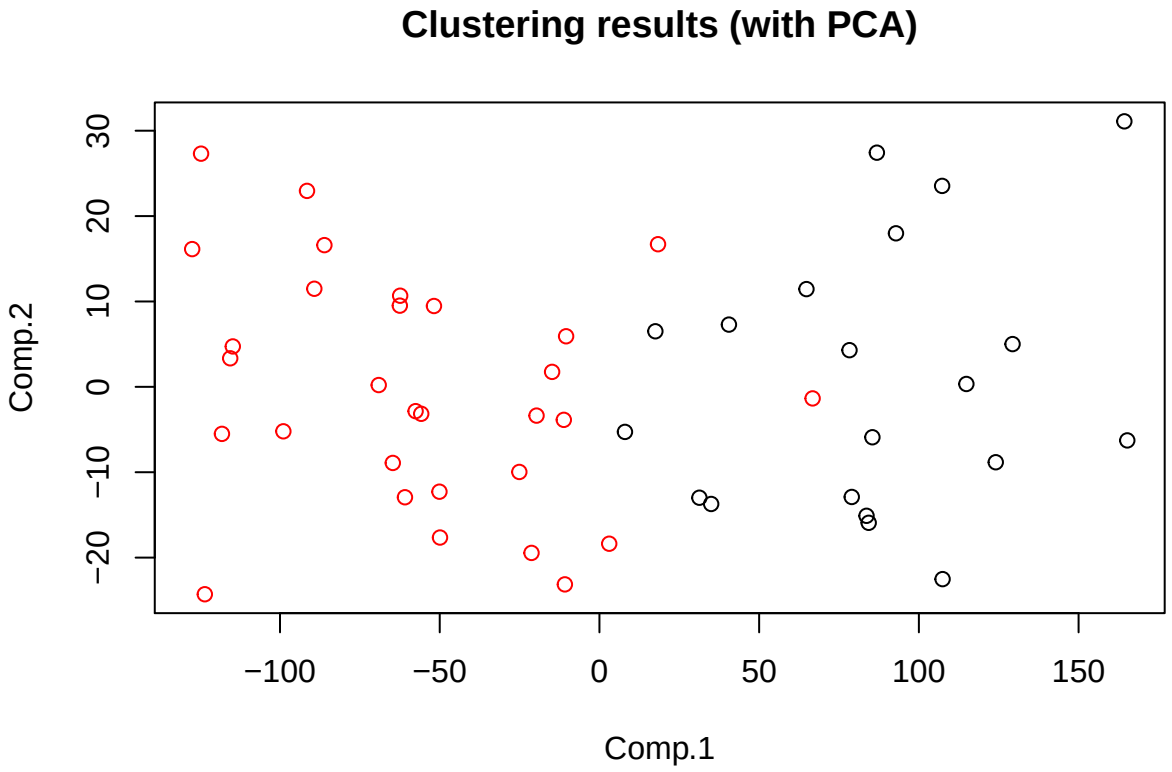
```
## Massachusetts  -0.77786532 -0.26110644  1.34440885 -0.526563903       2
## Michigan        0.99001041  1.01082751  0.58446551  1.480613993       1
## Minnesota      -1.16817555 -1.18505846  0.03177945 -0.676034598       2
## Mississippi     1.90838741  1.05882502 -1.48810723 -0.441152078       1
## Missouri        0.27826823  0.08687549  0.30812248  0.743936999       1
## Montana        -0.41051452 -0.74108152 -0.86633540 -0.515887425       2
## Nebraska       -0.80082475 -0.82507715 -0.24456358 -0.505210947       2
## Nevada          1.01296983  0.97482938  1.06806582  2.644350114       1
## New Hampshire  -1.30593210 -1.36504911 -0.65907813 -1.252564419       2
## New Jersey     -0.08908257 -0.14111267  1.62075188 -0.259651949       2
## New Mexico      0.82929443  1.37080881  0.30812248  1.160319648       1
## New York        0.76041616  0.99882813  1.41349461  0.519730957       1
## North Carolina  1.19664523  1.99477641 -1.41902147 -0.547916860       1
## North Dakota   -1.60440462 -1.50904164 -1.48810723 -1.487446939       2
## Ohio           -0.11204199 -0.60908837  0.65355127  0.017936483       2
## Oklahoma       -0.27275797 -0.23710769  0.16995096 -0.131534211       2
## Oregon         -0.66306820 -0.14111267  0.10086521  0.861378259       2
## Pennsylvania   -0.34163624 -0.77707965  0.44629400 -0.676034598       2
## Rhode Island   -1.00745957  0.03887798  1.48258036 -1.380682157       2
## South Carolina  1.51807718  1.29881255 -1.21176419  0.135377743       1
## South Dakota   -0.91562187 -1.01706718 -1.41902147 -0.900240639       2
## Tennessee       1.24256408  0.20686926 -0.45182086  0.605142783       1
## Texas           1.12776696  0.36286116  0.99898006  0.455672088       1
## Utah           -1.05337842 -0.60908837  0.99898006  0.178083656       2
## Vermont        -1.28297267 -1.47304350 -2.31713632 -1.071064290       2
## Virginia        0.16347111 -0.17711080 -0.17547783 -0.056798864       2
## Washington     -0.86970302 -0.30910395  0.51537975  0.530407436       2
## West Virginia  -0.47939280 -1.07706407 -1.83353601 -1.273917376       2
## Wisconsin      -1.19113497 -1.41304662  0.03177945 -1.113770203       2
## Wyoming        -0.22683912 -0.11711392 -0.38273510 -0.601299251       2
```

```
USA_PCA <- princomp(USArrests)
summary(USA_PCA)
```

b. Không áp dụng hàm `scale()` nhưng áp dụng PCA trên dữ liệu, sau đó thực hiện k-means với số cụm là 4 trên trên 2 thành phần chính đầu tiên và minh họa kết quả dùng hàm `plot`. Bình luận kết quả thu được.

```
## Importance of components:
##                             Comp.1      Comp.2      Comp.3       Comp.4
## Standard deviation     82.8908472 14.06956001 6.424204055 2.4578367034
## Proportion of Variance  0.9655342  0.02781734 0.005799535 0.0008489079
## Cumulative Proportion   0.9655342  0.99335156 0.999151092 1.0000000000
```

```
plot(USA_PCA$scores[,1:2], col=res$cluster, main="Clustering results (with PCA)")
```



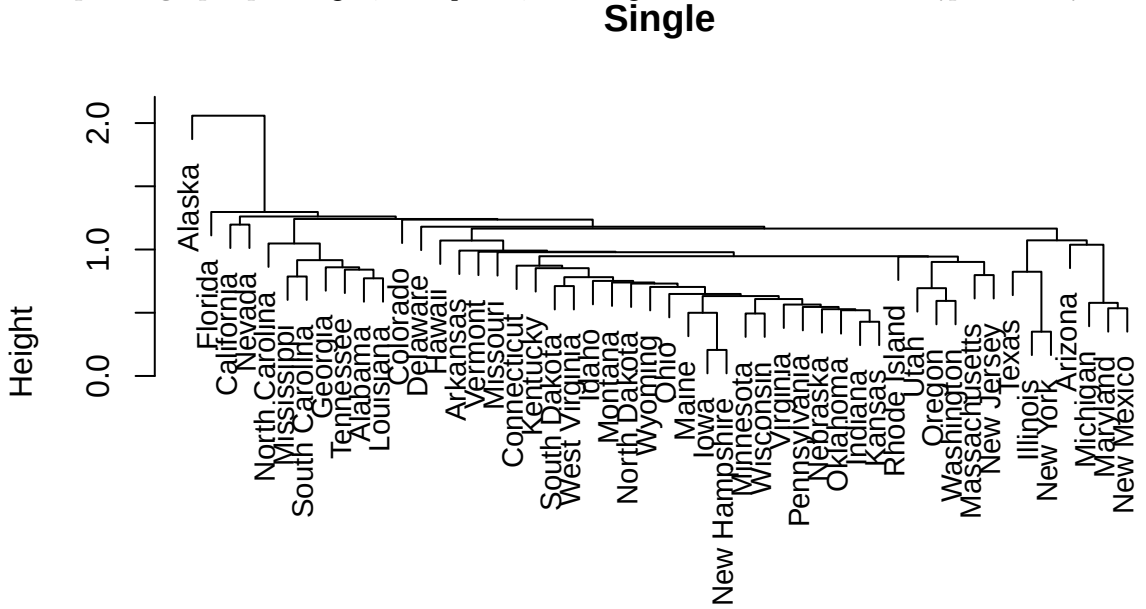

```
set.seed(17133015)
res <- kmeans(USArrests, 4, nstart = 20)
print(res)
```

```
## K-means clustering with 4 clusters of sizes 16, 10, 10, 14
##
## Cluster means:
##      Murder  Assault UrbanPop     Rape
## 1 11.812500 272.5625 68.31250 28.37500
## 2  2.950000  62.7000 53.90000 11.51000
## 3  5.590000 112.4000 65.60000 17.27000
## 4  8.214286 173.2857 70.64286 22.84286
##
## Clustering vector:
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              1              1              4              1
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              4              3              1              1              4
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              2              3              1              3              2
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              3              3              1              2              1
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              4              1              2              1              4
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              3              3              1              2              4
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              1              1              1              2              3
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              4              4              3              4              1
```

```
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              2              4              4              3              2
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              4              4              2              2              4
##
## Within cluster sum of squares by cluster:
## [1] 19563.863  4547.914  1480.210  9136.643
##  (between_SS / total_SS =  90.2 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

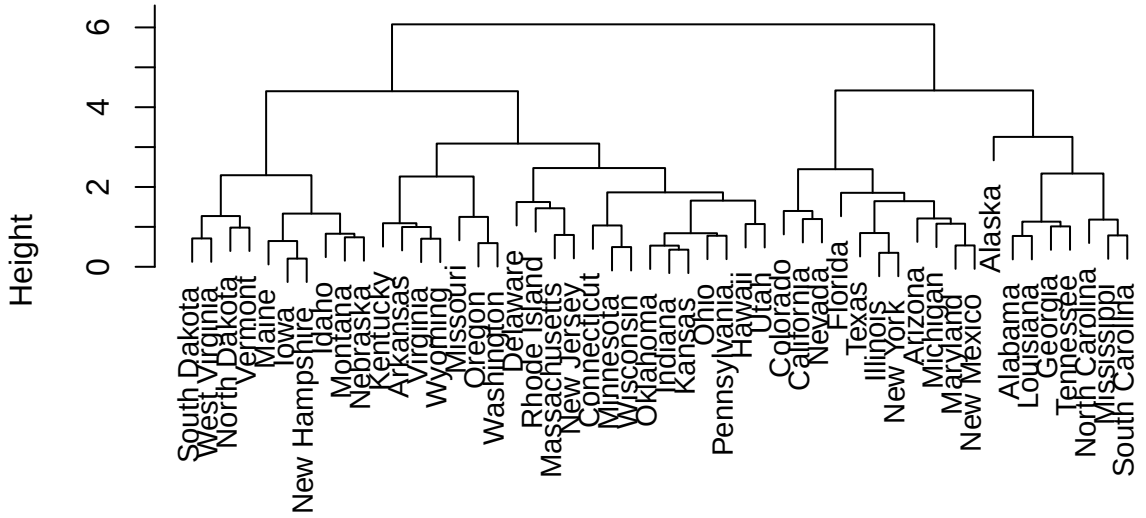
```
hc.single <- hclust(dist(USA_scale), "single")
plot(hc.single, main="Single", xlab="", sub ="", cex =.9)
```

c. Áp dụng hàm scale() trên dữ liệu trước khi thực hiện gom cụm phân cấp dùng các phương pháp single, complete, average và median trên tập dữ liệu USArrests.



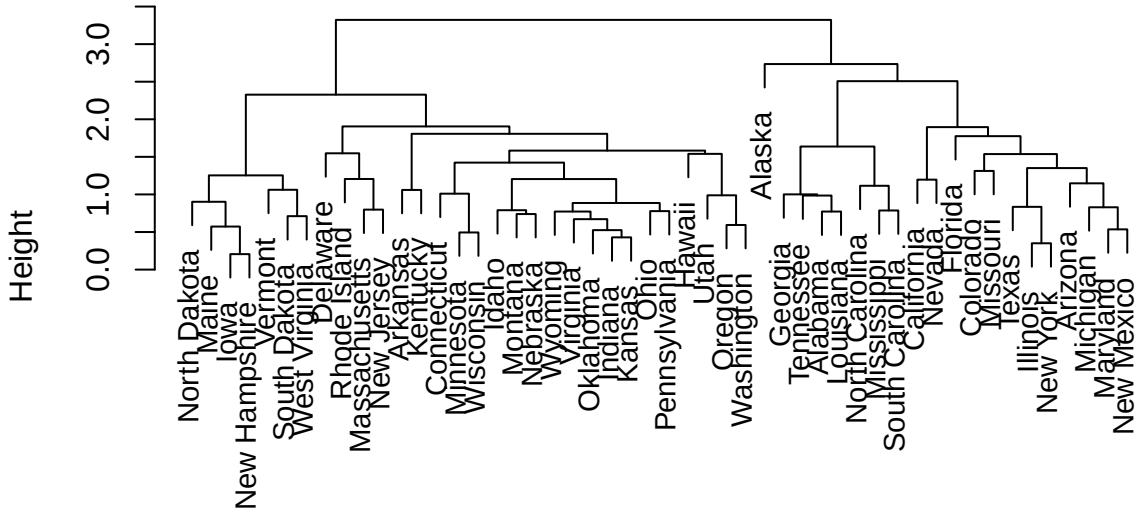

```
hc.complete <- hclust(dist(USA_scale), "complete")
plot(hc.complete, main="Complete", xlab="", sub ="", cex =.9)
```

Complete

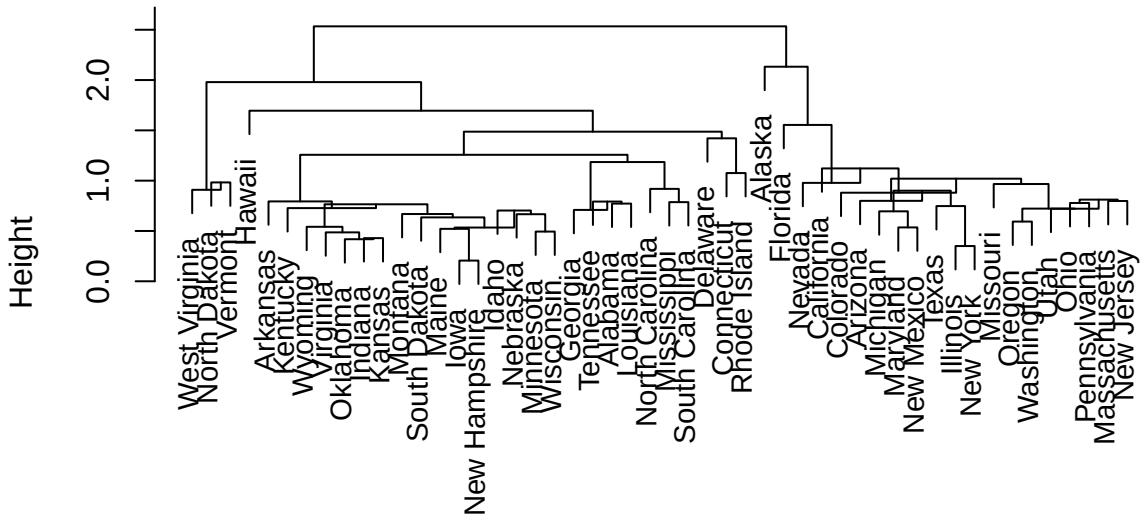



```
hc.average <- hclust(dist(USA_scale), "average")
plot(hc.average, main="Average", xlab="", sub ="", cex =.9)
```

Average




```
hc.median <- hclust(dist(USA_scale), "median")
plot(hc.median, main="Median", xlab="", sub ="", cex =.9)
```



####

d. Cắt dendrogram để thu được 2, 3, 4 cụm. Cho biết kết quả gom cụm tương ứng.

```
cutree(hc.single, 2)
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              2              1              1              1
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              1              1              1              1              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              1              1              1              1              1
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              1              1              1              1              1
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              1              1              1              1              1
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              1              1              1              1              1
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              1              1              1              1              1
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              1              1              1              1              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              1              1              1              1              1
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              1              1              1              1              1
```

```
cutree(hc.single, 3)
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              2              1              1              1
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              1              1              1              3              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              1              1              1              1              1
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              1              1              1              1              1
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
```

```
##             1             1             1             1             1
##        Montana      Nebraska        Nevada New Hampshire    New Jersey
##             1             1             1             1             1
##     New Mexico      New York North Carolina  North Dakota          Ohio
##             1             1             1             1             1
##       Oklahoma        Oregon  Pennsylvania  Rhode Island South Carolina
##             1             1             1             1             1
##   South Dakota     Tennessee         Texas          Utah       Vermont
##             1             1             1             1             1
##       Virginia    Washington West Virginia     Wisconsin       Wyoming
##             1             1             1             1             1
```

```
cutree(hc.single, 4)
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              2              1              1              3
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              1              1              1              4              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              1              1              1              1              1
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              1              1              1              1              1
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              1              1              1              1              1
##        Montana       Nebraska         Nevada New Hampshire     New Jersey
##              1              1              3              1              1
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              1              1              1              1              1
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              1              1              1              1              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              1              1              1              1              1
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              1              1              1              1              1
```

```
cutree(hc.complete, 2)
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              1              1              2              1
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              1              2              2              1              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              2              2              1              2              2
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              2              2              1              2              1
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              2              1              2              1              2
##        Montana       Nebraska         Nevada New Hampshire     New Jersey
##              2              2              1              2              2
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              1              1              1              2              2
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              2              2              2              2              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              2              1              1              2              2
```

```
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              2              2              2              2              2
```

```
cutree(hc.complete, 3)
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              1              2              3              2
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              3              3              2              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              3              3              2              3              3
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              3              3              1              3              2
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              3              2              3              1              3
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              3              3              2              3              3
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              2              2              1              3              3
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              3              3              3              3              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              3              1              2              3              3
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              3              3              3              3              3
```

```
cutree(hc.complete, 4)
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              1              2              3              2
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              3              3              2              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              3              4              2              3              4
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              3              3              1              4              2
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              3              2              3              1              3
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              4              4              2              4              3
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              2              2              1              4              3
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              3              3              3              3              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              4              1              2              3              4
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              3              3              4              3              3
```

```
cutree(hc.average, 2)
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              1              1              2              1
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              1              2              2              1              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
```

```
##             2             2             1             2             2
##        Kansas      Kentucky     Louisiana         Maine      Maryland
##             2             2             1             2             1
##  Massachusetts      Michigan     Minnesota   Mississippi      Missouri
##             2             1             2             1             1
##       Montana      Nebraska        Nevada New Hampshire    New Jersey
##             2             2             1             2             2
##    New Mexico      New York North Carolina  North Dakota          Ohio
##             1             1             1             2             2
##      Oklahoma        Oregon  Pennsylvania  Rhode Island South Carolina
##             2             2             2             2             1
##  South Dakota     Tennessee         Texas          Utah       Vermont
##             2             1             1             2             2
##      Virginia    Washington West Virginia     Wisconsin       Wyoming
##             2             2             2             2             2
```

```
cutree(hc.average, 3)
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              2              1              3              1
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              1              3              3              1              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              3              3              1              3              3
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              3              3              1              3              1
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              3              1              3              1              1
##       Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              3              3              1              3              3
##    New Mexico       New York North Carolina   North Dakota           Ohio
##              1              1              1              3              3
##      Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              3              3              3              3              1
##  South Dakota      Tennessee          Texas           Utah        Vermont
##              3              1              1              3              3
##      Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              3              3              3              3              3
```

```
cutree(hc.average, 4)
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              2              3              4              3
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              3              4              4              3              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              4              4              3              4              4
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              4              4              1              4              3
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              4              3              4              1              3
##       Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              4              4              3              4              4
##    New Mexico       New York North Carolina   North Dakota           Ohio
##              3              3              1              4              4
```

```
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              4              4              4              4              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              4              1              3              4              4
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              4              4              4              4              4
```

```
cutree(hc.median, 2)
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              2              2              1              2
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              1              1              2              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              1              1              2              1              1
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              1              1              1              1              2
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              2              2              1              1              2
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              1              1              2              1              2
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              2              2              1              1              2
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              1              2              2              1              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              1              1              2              2              1
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              1              2              1              1              1
```

```
cutree(hc.median, 3)
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              2              3              1              3
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              3              1              1              3              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              1              1              3              1              1
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              1              1              1              1              3
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              3              3              1              1              3
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              1              1              3              1              3
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              3              3              1              1              3
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              1              3              3              1              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              1              1              3              3              1
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              1              3              1              1              1
```

```
cutree(hc.median, 4)
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
```

```
##             1             2             3             1             3
##      Colorado   Connecticut      Delaware       Florida       Georgia
##             3             1             1             3             1
##        Hawaii         Idaho      Illinois       Indiana          Iowa
##             1             1             3             1             1
##        Kansas      Kentucky     Louisiana         Maine      Maryland
##             1             1             1             1             3
##  Massachusetts      Michigan     Minnesota   Mississippi      Missouri
##             3             3             1             1             3
##       Montana      Nebraska        Nevada New Hampshire    New Jersey
##             1             1             3             1             3
##    New Mexico      New York North Carolina  North Dakota          Ohio
##             3             3             1             4             3
##      Oklahoma        Oregon  Pennsylvania  Rhode Island South Carolina
##             1             3             3             1             1
##  South Dakota     Tennessee         Texas          Utah       Vermont
##             1             1             3             3             4
##      Virginia    Washington West Virginia     Wisconsin       Wyoming
##             1             3             4             1             1
```

e. Hãy cho biết ảnh hưởng của việc scaling đối với các kết quả thu được? Ta có nên thực hiện scaling trước khi áp dụng các thuật toán? Hãy chứng minh câu trả lời của bạn.

```
library(clValid)
```

f. So sánh Dunn index của các kết quả gom cụm khi áp dụng các thuật toán trên.

```
## Loading required package: cluster
```

```
d <- dist(USA_scale, method="euclidean")
hc.single <- hclust(d, "single")
hc.single.cluster <- cutree(hc.single, k = 2)
dunn(d, hc.single.cluster)
```

```
## [1] 0.3386885
```

```
dunn(d, res$cluster)
```

```
## [1] 0.1176163
```

```
d <- dist(USA_scale, method="euclidean")
hc.single <- hclust(d, "centroid")
hc.single.cluster <- cutree(hc.single, k = 2)
cl1 <- silhouette(hc.single.cluster, d)
mean(cl1[, 3])
```

g. So sánh chỉ số Silhouette của các kết quả gom cụm khi áp dụng các thuật toán trên.

```
## [1] 0.2101346
```

```
cl2 <- silhouette(res$cluster, d)
mean(cl2[, 3])
```

```
## [1] 0.09018768
```

Câu hỏi 2

Thực hiện thuật toán `kmeans` và `hclust` và `dbscan` cho tập dữ liệu `iris`.

```
irisN = iris[iris$Species=="setosa",1:4]
set.seed(17133015)
iris.cluster <- kmeans(irisN, center = 3, nstart = 20)
print(iris.cluster)
```

a. Áp dụng `kmeans` với k = 3 trên tập dữ liệu `iris` sau khi loại bỏ nhãn (thuộc tính `Species`) khỏi tập dữ liệu.

```
## K-means clustering with 3 clusters of sizes 8, 23, 19
##
## Cluster means:
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1     5.512500    4.000000     1.475000    0.275000
## 2     5.100000    3.513043     1.526087    0.273913
## 3     4.678947    3.084211     1.378947    0.200000
##
## Clustering vector:
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
##  2  3  3  3  2  1  3  2  3  3  1  2  3  3  1  1  1  2  1  2  2  2  3  2  2  3
## 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
##  2  2  2  3  3  2  1  1  3  3  2  2  3  2  2  3  3  2  2  3  2  3  2  2
##
## Within cluster sum of squares by cluster:
## [1] 0.958750 2.094783 2.488421
##  (between_SS / total_SS =  63.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
library(dendextend)
```

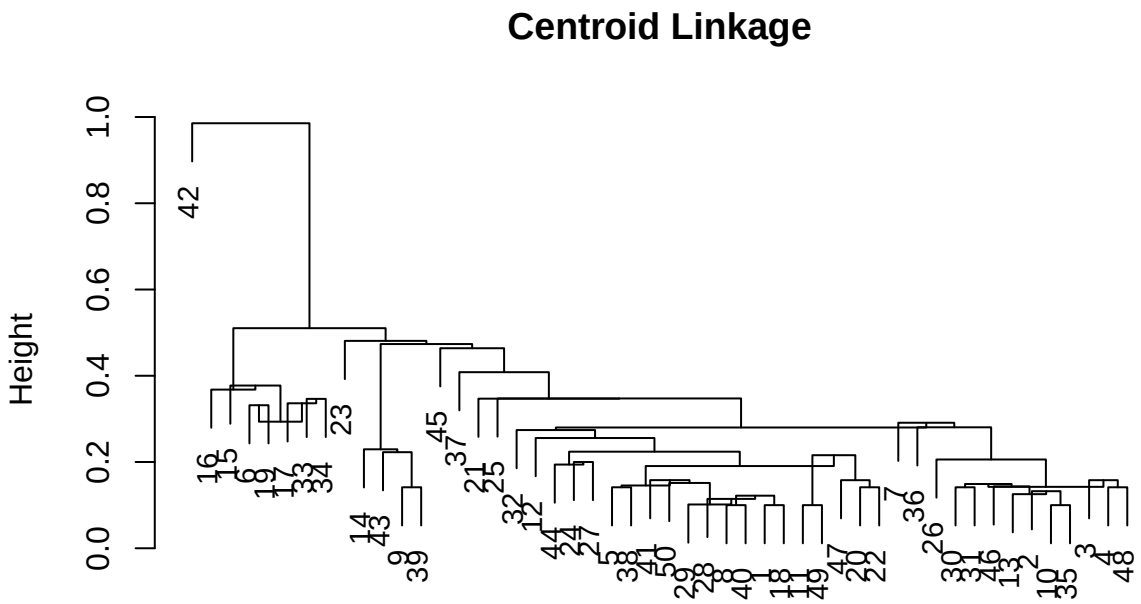
b. Áp dụng `hclust` và cắt dendrogram với k = 3 trên tập dữ liệu `iris` sau khi loại bỏ nhãn (thuộc tính `Species`) khỏi tập dữ liệu.

```
##
## ---------------------
## Welcome to dendextend version 1.14.0
## Type citation('dendextend') for how to cite the package.
##
## Type browseVignettes(package = 'dendextend') for the package vignette.
## The github page is: https://github.com/talgalili/dendextend/
##
## Suggestions and bug-reports can be submitted at: https://github.com/talgalili/dendextend/issues
## Or contact: <tal.galili@gmail.com>
##
##  To suppress this message use:  suppressPackageStartupMessages(library(dendextend))
## ---------------------
```

```
##
## Attaching package: 'dendextend'
```

```
## The following object is masked from 'package:stats':
##
##     cutree
```

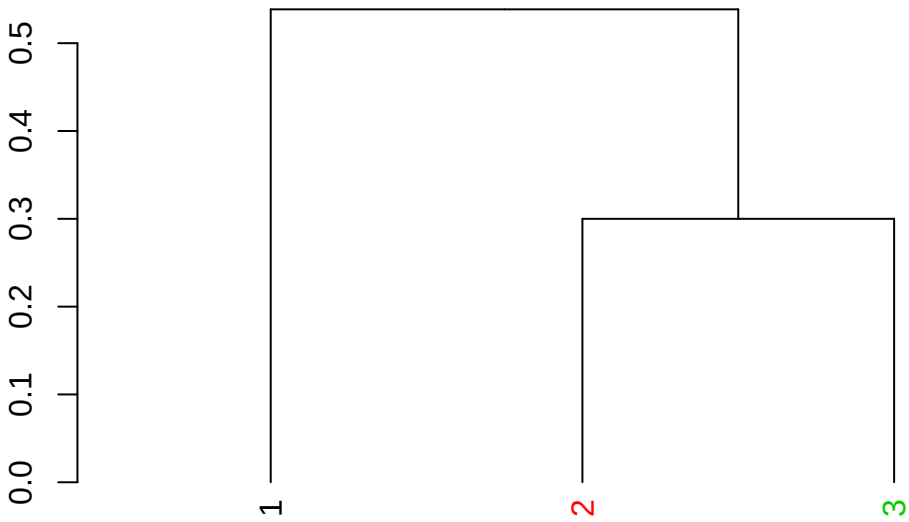
```
hc.centroid <- hclust(dist(irisN), "centroid")
plot(hc.centroid, main="Centroid Linkage", xlab="", sub ="", cex =.9)
```




```
den <- as.dendrogram(hclust(dist(irisN[1:3,])))
labels_colors(den) <- 1:3
labels_colors(den)
```

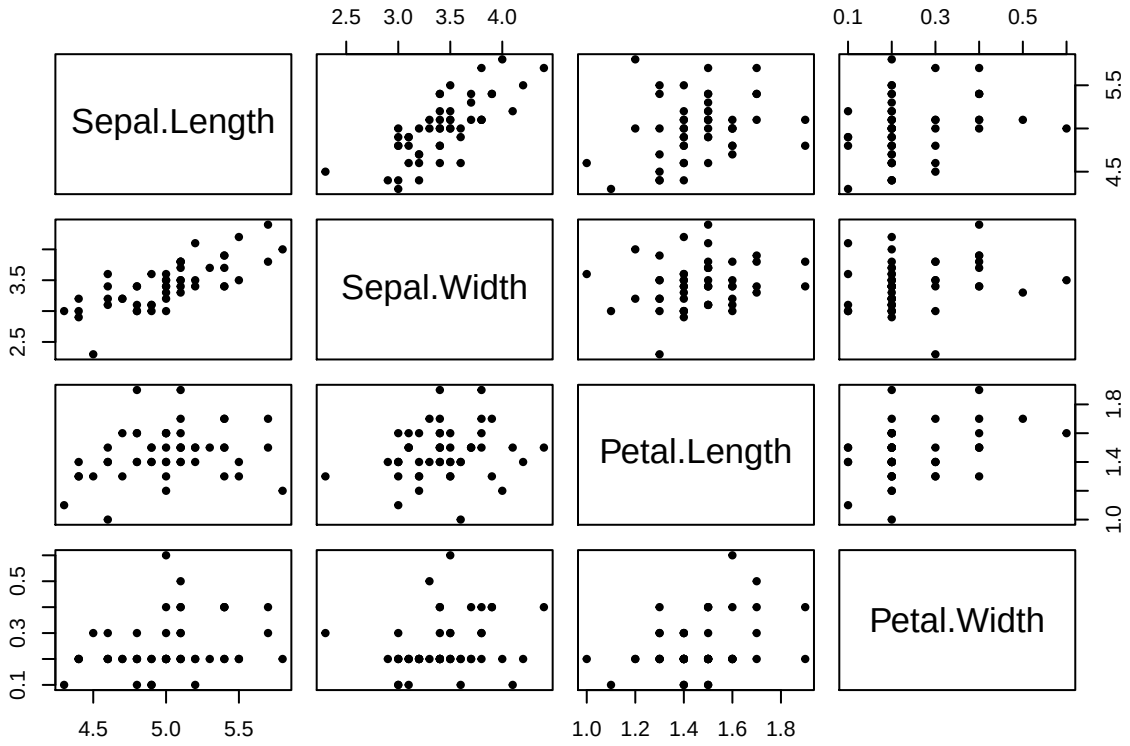
```
## 1 2 3
## 1 2 3
```

```
plot(den, main = "dendrogram")
```



c. Áp dụng `dbscan` trên tập dữ liệu `iris` sau khi loại bỏ nhãn (thuộc tính Species) khỏi tập dữ liệu. Hãy thử nghiệm với các tham số `eps` và `minPts` khác nhau và chọn các tham số bạn cho là tốt nhất. Bạn có chiến lược nào để chọn các tham số này không?

```
plot(irisN, pch=20)
```

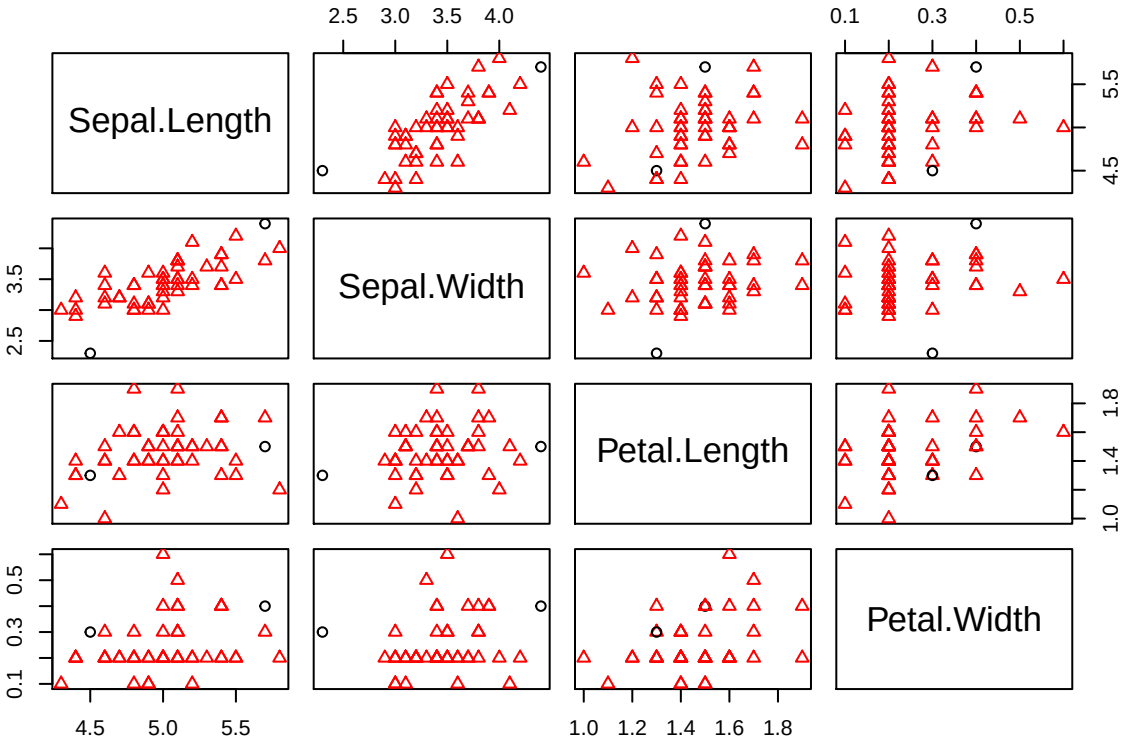



```
library(dbscan)
scan <- dbscan(irisN, eps = 0.5, minPts = 8)
scan
```

```
## DBSCAN clustering for 50 objects.
## Parameters: eps = 0.5, minPts = 8
```

```
## The clustering contains 1 cluster(s) and 2 noise points.
##
##  0  1
##  2 48
##
## Available fields: cluster, eps, minPts
```

```
plot(irisN, col = scan$cluster + 1L, pch = scan$cluster + 1L)
```



####

d. Sử dụng thuộc tính `Species` làm nhãn cụm thật sự, hãy tính và so sánh Precison, Recall, và F-measure của kết quả gom cụm khi dùng `kmeans`, `hclust` và `dbscan`.

```
cutree(hc.median, 4)
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              2              3              1              3
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              3              1              1              3              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              1              1              3              1              1
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              1              1              1              1              3
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              3              3              1              1              3
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              1              1              3              1              3
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              3              3              1              4              3
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              1              3              3              1              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              1              1              3              3              4
```

Table 1: Contingency table of clustering results

Clusters/Species	T_1	T_2	$\cdots$	T_p
C_1	n_{11}	n_{12}	$\cdots$	n_{1p}
C_2	n_{21}	n_{22}	$\cdots$	n_{2p}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
C_k	n_{k1}	n_{k2}	$\cdots$	n_{kp}

```
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              1              3              4              1              1
```

Giả sử tập dữ liệu D có n phần tử x_i được phân hoạch thành p nhóm (ở đây ứng với số loài). Gọi $y_i \in \{1, 2, \cdots, p\}$ là nhóm thật sự (ground-truth labels) cho mỗi phần tử. Ground-truth clustering được cho bởi $T = \{T_1, T_2, \cdots, T_p\}$, với T_j bao gồm tất cả các phần tử có nhãn j, nghĩa là, $T_j = \{x_i \in D | y_i = j\}$. Mặt khác, gọi $C = \{C_1, C_2, \cdots, C_k\}$ là một kết quả gom cụm của D thành k cụm (cluster), qua một thuật toán gom cụm nào đó, và $\hat{y_i} \in \{1, 2, \cdots, k\}$ là cluster label cho x_i. Ta sẽ xem T là một phân hoạch chuẩn (ground-truth partitioning) và mỗi T_i là một phân vùng (partition). Ta gọi C là một kết quả gom cụm (clustering), với mỗi C_i là một cụm (cluster). Giả sử ground truth là biết trước, một thuật toán gom cụm sẽ thực hiện gom cụm trên D với số cụm chính xác, tức với $k = p$. Tuy nhiên, để giữ tính tổng quát, ta cho phép $k \neq p$.

Các độ đo đánh giá kết quả gom cụm cố gắng nắm bắt mức độ mà các phần tử từ cùng một phân vùng (partition) xuất hiện trong cùng một cụm (cluster) và mức độ mà các phần tử từ các phân vùng (partition) khác nhau được nhóm thành các cụm (cluster) khác nhau. Những độ đo này dựa trên $k \times p$ contingency table N (xem Table 1) được thành lập dựa vào một kết quả gom cụm (clustering) C và một phân hoạch chuẩn (ground-truth partitioning) T, được định nghĩa như sau:

$$N(i, j) = n_{ij} = |C_i \cap T_j|$$

- *Recall* là tỷ lệ đối tượng cùng loài được gán cùng cụm.
- *Precision* là tỷ lệ đối tượng được gán cùng cụm thuộc cùng loài.
- *F-measure* là một độ đo cân bằng giữa *Precision* và *Recall* và được tính bằng trung bình điều hòa giữa *Precision* và *Recall*. Đây là một độ đo thường được sử dụng để so sánh các thuật toán gom cụm với nhau.

Các độ đo *Precision*, *Recall*, và *F-measure* được tính từ Table 1 dùng các công thức sau:

$$precision = \frac{\sum_{i=1}^{k} \max_{j \in \{1, \ldots p\}} \{n_{ij}\}}{\sum_{i=1}^{k} \sum_{j=1}^{p} n_{ij}} \tag{1}$$

$$recall = \frac{\sum_{j=1}^{p} \max_{i \in \{1, \ldots, k\}} \{n_{ij}\}}{\sum_{i=1}^{k} \sum_{j=1}^{p} n_{ij}} \tag{2}$$

$$F\text{-}measure = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{3}$$