Clustering

## 1.Data structure

### 1.1 Data

"knitr"package는 표를 출력하기 위해 이용하였다.

library(knitr)

This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.

str(USArrests)

'data.frame': 50 obs. of 4 variables:  
 $ Murder : num 13.2 10 8.1 8.8 9 7.9 3.3 5.9 15.4 17.4 ...  
 $ Assault : int 236 263 294 190 276 204 110 238 335 211 ...  
 $ UrbanPop: int 58 48 80 50 91 78 77 72 80 60 ...  
 $ Rape : num 21.2 44.5 31 19.5 40.6 38.7 11.1 15.8 31.9 25.8 ...

kable(head(USArrests))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Murder | Assault | UrbanPop | Rape |
| Alabama | 13.2 | 236 | 58 | 21.2 |
| Alaska | 10.0 | 263 | 48 | 44.5 |
| Arizona | 8.1 | 294 | 80 | 31.0 |
| Arkansas | 8.8 | 190 | 50 | 19.5 |
| California | 9.0 | 276 | 91 | 40.6 |
| Colorado | 7.9 | 204 | 78 | 38.7 |

x <- USArrests[c(6, 18, 31, 33, 50),c(1,4)]  
kable(x)

|  |  |  |
| --- | --- | --- |
|  | Murder | Rape |
| Colorado | 7.9 | 38.7 |
| Louisiana | 15.4 | 22.2 |
| New Mexico | 11.4 | 32.1 |
| North Carolina | 13.0 | 16.1 |
| Wyoming | 6.8 | 15.6 |

## 2.Similaity measures

### 2.1 Euclidean distance

#### 2.1.1 Euclidean distance (using R)

d1 = round(dist(x,diag = TRUE),digits = 2)   
d1

Colorado Louisiana New Mexico North Carolina Wyoming  
Colorado 0.00   
Louisiana 18.12 0.00   
New Mexico 7.47 10.68 0.00   
North Carolina 23.17 6.56 16.08 0.00   
Wyoming 23.13 10.84 17.13 6.22 0.00

#### 2.1.2 Euclidean distance (calculate)

x[1,]

Murder Rape  
Colorado 7.9 38.7

x[2,]

Murder Rape  
Louisiana 15.4 22.2

x[1,]-x[2,]

Murder Rape  
Colorado -7.5 16.5

(x[1,]-x[2,])^2

Murder Rape  
Colorado 56.25 272.25

sqrt(sum((x[1,]-x[2,])^2))

[1] 18.12457

R로 계산한 것과 직접 계산한 것이 일치한다.

### 2.2 Maximum distance

#### 2.2.1 Maximum distance (using R)

각각 변수의 거리 차 중 큰 값을 택하는 방법

d2 = dist(x,method = "maximum",diag = TRUE)  
d2

Colorado Louisiana New Mexico North Carolina Wyoming  
Colorado 0.0   
Louisiana 16.5 0.0   
New Mexico 6.6 9.9 0.0   
North Carolina 22.6 6.1 16.0 0.0   
Wyoming 23.1 8.6 16.5 6.2 0.0

#### 2.2.2 Maximum distance (calculate)

abs( x[1,] - x[2,] )

Murder Rape  
Colorado 7.5 16.5

max(abs( x[1,] - x[2,] ))

[1] 16.5

abs( x[2,] - x[3,] )

Murder Rape  
Louisiana 4 9.9

R로 계산한 것과 직접 계산한 것이 일치한다.

### 2.3 Manhattan distance

#### 2.3.1 Manhattan distance (using R)

d3 = dist(x,method = "manhattan",diag = TRUE)  
d3

Colorado Louisiana New Mexico North Carolina Wyoming  
Colorado 0.0   
Louisiana 24.0 0.0   
New Mexico 10.1 13.9 0.0   
North Carolina 27.7 8.5 17.6 0.0   
Wyoming 24.2 15.2 21.1 6.7 0.0

R로 계산한 것과 직접 계산한 것이 일치한다.

#### 2.3.2 Manhattan distance (calculate)

x[1,]

Murder Rape  
Colorado 7.9 38.7

x[2,]

Murder Rape  
Louisiana 15.4 22.2

x[1,]-x[2,]

Murder Rape  
Colorado -7.5 16.5

sum(abs(x[1,]-x[2,]))

[1] 24

R로 계산한 것과 직접 계산한 것이 일치한다.

### 2.4 Canberra distance

#### 2.4.1 Canberra distance (using R)

d4 = dist(x,method = "canberra",diag = TRUE)  
d4

Colorado Louisiana New Mexico North Carolina Wyoming  
Colorado 0.0000000   
Louisiana 0.5928244 0.0000000   
New Mexico 0.2745675 0.3315742 0.0000000   
North Carolina 0.6564279 0.2437760 0.3975240 0.0000000   
Wyoming 0.5002443 0.5619906 0.5986592 0.3289042 0.0000000

#### 2.4.2 Canberra distance (calculate)

abs(x[1,]-x[2,])

Murder Rape  
Colorado 7.5 16.5

x[1,]+x[2,]

Murder Rape  
Colorado 23.3 60.9

sum(abs(x[1,]-x[2,])/(x[1,]+x[2,]))

[1] 0.5928244

### 2.5 Minkowski distance

#### 2.5.1 Minkowski distance (using R)

d6 = dist(x,method = "minkowski",diag =TRUE)  
d6

Colorado Louisiana New Mexico North Carolina Wyoming  
Colorado 0.000000   
Louisiana 18.124569 0.000000   
New Mexico 7.470609 10.677547 0.000000   
North Carolina 23.168297 6.555151 16.079801 0.000000   
Wyoming 23.126176 10.840664 17.129215 6.220129 0.000000

#### 2.5.2 Minkowski distance (calculate)

minkowski = function(one,two,m){  
 d = ( sum((abs(x[one,]-x[two,]))^m) )^(1/m)  
}  
value = minkowski(2,3,2)

## 3. Hierarchical clustering method

### 3.1 Single linkage : nearest neighbor

#### 3.1.1 Single linkage : nearest neighbor

hc1 = hclust(dist(x),method = "single")   
hc1

Call:  
hclust(d = dist(x), method = "single")  
  
Cluster method : single   
Distance : euclidean   
Number of objects: 5

rev(hc1)

$dist.method  
[1] "euclidean"  
  
$call  
hclust(d = dist(x), method = "single")  
  
$method  
[1] "single"  
  
$labels  
[1] "Colorado" "Louisiana" "New Mexico" "North Carolina"  
[5] "Wyoming"   
  
$order  
[1] 2 4 5 1 3  
  
$height  
[1] 6.220129 6.555151 7.470609 10.677547  
  
$merge  
 [,1] [,2]  
[1,] -4 -5  
[2,] -2 1  
[3,] -1 -3  
[4,] 2 3

#### 3.1.2 Single linkage : nearest neighbor

d1

Colorado Louisiana New Mexico North Carolina Wyoming  
Colorado 0.00   
Louisiana 18.12 0.00   
New Mexico 7.47 10.68 0.00   
North Carolina 23.17 6.56 16.08 0.00   
Wyoming 23.13 10.84 17.13 6.22 0.00

min(d1)

[1] 6.22

d1 = as.matrix(d1)  
d1 = lower.tri(d1, diag=TRUE)\*d1  
d1

Colorado Louisiana New Mexico North Carolina Wyoming  
Colorado 0.00 0.00 0.00 0.00 0  
Louisiana 18.12 0.00 0.00 0.00 0  
New Mexico 7.47 10.68 0.00 0.00 0  
North Carolina 23.17 6.56 16.08 0.00 0  
Wyoming 23.13 10.84 17.13 6.22 0

5번 개체와 4번 개체를 묶어 군집(45)으로 한다.

군집(45)와 나머지 개체 1,2,3과의 거리를 계산한다.

d{(45)1} = min{d(41),d(51)} = min{23.17, 23.13} = 23.13

d{(45)2} = min{d(42),d(52)} = min{ 6.56, 10.84} = 6.56

d{(45)3} = min{d(43),d(53)} = min{16.08, 17.13} = 16.08

d45 = c(23.13,6.56,16.08)  
mat2 = cbind(d45,d1[1:3,1:3])  
mat2

d45 Colorado Louisiana New Mexico  
Colorado 23.13 0.00 0.00 0  
Louisiana 6.56 18.12 0.00 0  
New Mexico 16.08 7.47 10.68 0

45번 개체와 2번 개체가 가까우므로 군집 45(2)로 계산한다.

d{(452)1} = min{d(45)1, d21} = min{ 23.13, 18.12} = 18.12

d{(452)3} = min{d(45)3, d23} = min{ 16.08, 10.68} = 10.68

d{1 3} = 7.47

d45.2 = c(0,18.12,10.68)  
d1 = c(0,0,7.47)  
d3 = c(0,0,0)  
mat3 = cbind(d45.2, d1, d3)  
mat3

d45.2 d1 d3  
[1,] 0.00 0.00 0  
[2,] 18.12 0.00 0  
[3,] 10.68 7.47 0

1번 개체와 3번 개체가 가장 가까우므로 군집으로 묶는다.

(45)123 -> (452)13 -> (452)(13)

### 3.2 Complete linkage : farthest neighbor

#### 3.2.1 Complete linkage : farthest neighbor

hc2 = hclust(dist(x),method = "complete")   
hc2

Call:  
hclust(d = dist(x), method = "complete")  
  
Cluster method : complete   
Distance : euclidean   
Number of objects: 5

rev(hc2)

$dist.method  
[1] "euclidean"  
  
$call  
hclust(d = dist(x), method = "complete")  
  
$method  
[1] "complete"  
  
$labels  
[1] "Colorado" "Louisiana" "New Mexico" "North Carolina"  
[5] "Wyoming"   
  
$order  
[1] 1 3 2 4 5  
  
$height  
[1] 6.220129 7.470609 10.840664 23.168297  
  
$merge  
 [,1] [,2]  
[1,] -4 -5  
[2,] -1 -3  
[3,] -2 1  
[4,] 2 3

#### 3.2.2 Complete linkage : farthest neighbor

가장 거리가 가까운 쌍을 찾는다.

d1 = round(dist(x,diag = TRUE),digits = 2)   
min(d1)

[1] 6.22

d1 = as.matrix(d1)  
d1 = lower.tri(d1, diag=TRUE)\*d1  
d1

Colorado Louisiana New Mexico North Carolina Wyoming  
Colorado 0.00 0.00 0.00 0.00 0  
Louisiana 18.12 0.00 0.00 0.00 0  
New Mexico 7.47 10.68 0.00 0.00 0  
North Carolina 23.17 6.56 16.08 0.00 0  
Wyoming 23.13 10.84 17.13 6.22 0

4번 개체와 5번 개체를 묶어 군집(45)으로 한다. 가장 거리가 먼 거리를 계산한다.

d{(45)1} = max{d(41),d(51)} = max{23.17, 23.13} = 23.17

d{(45)2} = max{d(42),d(52)} = max{ 6.56, 10.84} = 10.84

d{(45)3} = max{d(43),d(53)} = max{16.08, 17.13} = 17.13

계산 결과를 매트릭스로 나타내면 아래와 같다.

d45 = c(23.17,10.84, 17.13)  
mat = cbind(d45,d1[1:3,1:3])  
mat

d45 Colorado Louisiana New Mexico  
Colorado 23.17 0.00 0.00 0  
Louisiana 10.84 18.12 0.00 0  
New Mexico 17.13 7.47 10.68 0

1번 개체와 3번 개체가 가장 가까우므로 1번 개체와 3번 개체를 묶는다.

d{(45)(13)} = max{ d(45)1, d(45)3} = max{ 23.17, 17.13} = 23.17

d{(45)2} = 10.84

d{(13)2} = max{ d12, d32} = max{ 18.12, 10.68} = 18.12

d45 = c(0,23.17,10.84)  
d13 = c(0,0,18.12)  
d2 = c(0,0,0)  
mat = cbind(d45,d13,d2)  
mat

d45 d13 d2  
[1,] 0.00 0.00 0  
[2,] 23.17 0.00 0  
[3,] 10.84 18.12 0

여기서 가장 가까운 거리는 10.84로 개체 45 와 2가 가까우므로 군집으로 묶는다.

(45)123 -> (45)(13)2 -> (452)(13)

### 3.3 Ward method

#### 3.3.1 Ward method

hc3 = hclust(dist(x),method = "ward")

The "ward" method has been renamed to "ward.D"; note new "ward.D2"

hc3

Call:  
hclust(d = dist(x), method = "ward")  
  
Cluster method : ward.D   
Distance : euclidean   
Number of objects: 5

rev(hc3)

$dist.method  
[1] "euclidean"  
  
$call  
hclust(d = dist(x), method = "ward")  
  
$method  
[1] "ward.D"  
  
$labels  
[1] "Colorado" "Louisiana" "New Mexico" "North Carolina"  
[5] "Wyoming"   
  
$order  
[1] 1 3 2 4 5  
  
$height  
[1] 6.220129 7.470609 9.523834 32.542291  
  
$merge  
 [,1] [,2]  
[1,] -4 -5  
[2,] -1 -3  
[3,] -2 1  
[4,] 2 3

### 3.4 Average linkage

#### 3.4.1 Average linkage

hc4 = hclust(dist(x),method = "average")   
hc4

Call:  
hclust(d = dist(x), method = "average")  
  
Cluster method : average   
Distance : euclidean   
Number of objects: 5

rev(hc4)

$dist.method  
[1] "euclidean"  
  
$call  
hclust(d = dist(x), method = "average")  
  
$method  
[1] "average"  
  
$labels  
[1] "Colorado" "Louisiana" "New Mexico" "North Carolina"  
[5] "Wyoming"   
  
$order  
[1] 1 3 2 4 5  
  
$height  
[1] 6.220129 7.470609 8.697907 18.050934  
  
$merge  
 [,1] [,2]  
[1,] -4 -5  
[2,] -1 -3  
[3,] -2 1  
[4,] 2 3

#### 3.4.2 Average linkage

d1 = round(dist(x,diag = TRUE),digits = 2)   
d1

Colorado Louisiana New Mexico North Carolina Wyoming  
Colorado 0.00   
Louisiana 18.12 0.00   
New Mexico 7.47 10.68 0.00   
North Carolina 23.17 6.56 16.08 0.00   
Wyoming 23.13 10.84 17.13 6.22 0.00

min(d1)

[1] 6.22

d1 = as.matrix(d1)  
d1 = lower.tri(d1, diag=TRUE)\*d1  
d1

Colorado Louisiana New Mexico North Carolina Wyoming  
Colorado 0.00 0.00 0.00 0.00 0  
Louisiana 18.12 0.00 0.00 0.00 0  
New Mexico 7.47 10.68 0.00 0.00 0  
North Carolina 23.17 6.56 16.08 0.00 0  
Wyoming 23.13 10.84 17.13 6.22 0

4번 개체와 5번 개체를 묶어 군집(45)으로 한다.

d{(45)1} = 1/(2\*1) x (d41 + d51) = 1/2 x (23.17+ 23.13) = 23.15

d{(45)2} = 1/(2\*1) x (d42 + d52) = 1/2 x (6.56 + 10.84) = 8.7

d{(45)3} = 1/(2\*1) x (d43 + d53) = 1/2 x (16.08 + 17.13) = 16.605

d45 = c(23.15, 8.7, 16.605)  
mat = cbind(d45,d1[1:3,1:3])  
mat

d45 Colorado Louisiana New Mexico  
Colorado 23.150 0.00 0.00 0  
Louisiana 8.700 18.12 0.00 0  
New Mexico 16.605 7.47 10.68 0

7.47로 개체 1과 개체 3이 가까우므로 군집으로 묶는다.

d{(45)(13)} = 1/(2\*2) x (d41 + d43 + d51 + d53) = 1/4 x 79.51 = 19.88

d{(45)2} = 1/(2\*1) x (d42 + d52) = 1/2 x (6.56 + 10.84) = 8.7

d{(13)2} = 1/(2\*1) x (d12 + d32) = 1/2 x (18.12 + 10.68) = 14.4

d45 = c(0, 19.88, 8.7)  
d13 = c(0,0,14.4)  
d2 = c(0,0,0)  
mat = cbind(d45, d13, d2)  
mat

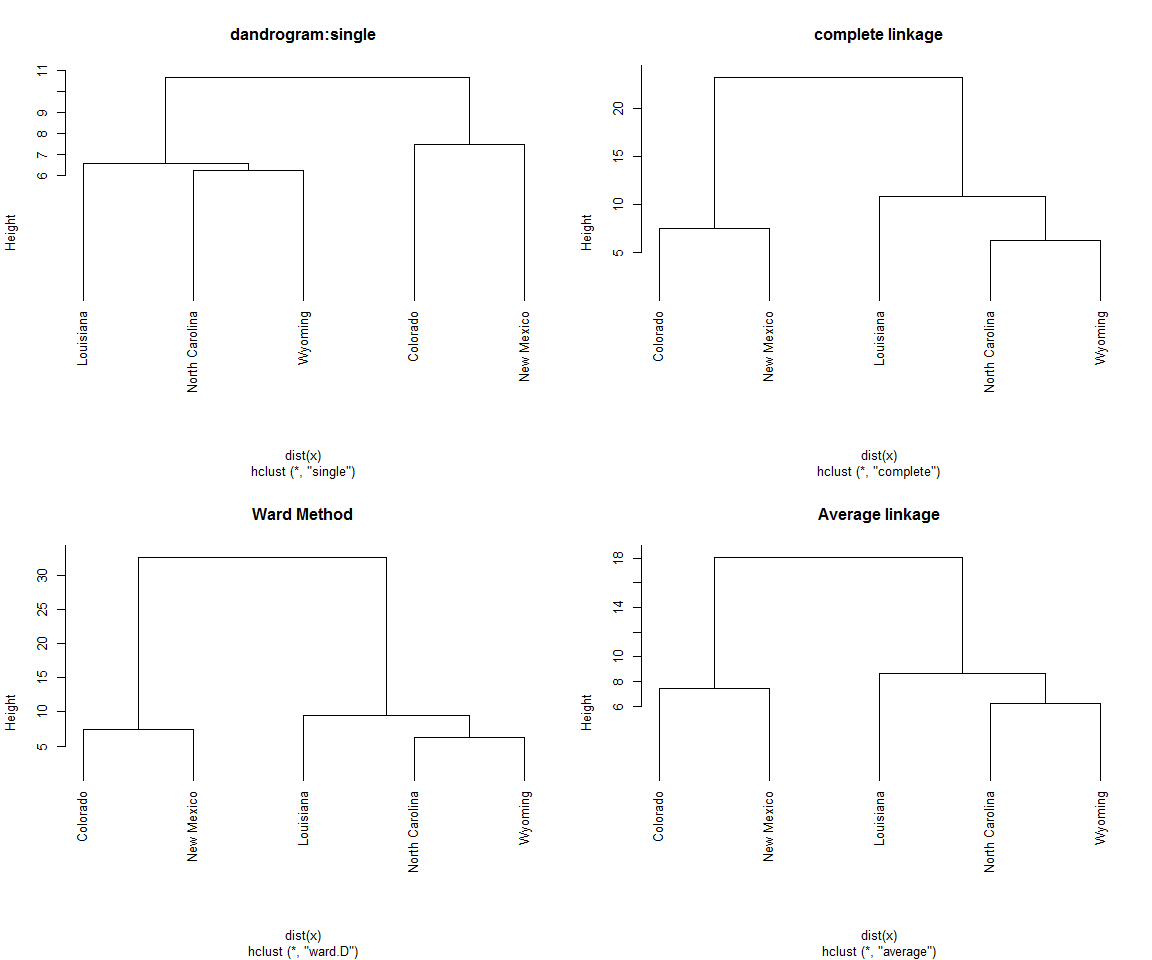
d45 d13 d2  
[1,] 0.00 0.0 0  
[2,] 19.88 0.0 0  
[3,] 8.70 14.4 0

개체 45와 개체 2가 가장 가까우므로 개체 452로 군집한다.

(45)123 -> (45)(13)2 -> (452)(13)

### 3.5 Dendrogram

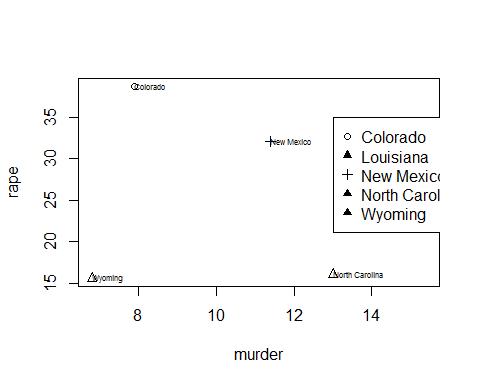
par(mfrow = c(2,2))  
plot(hc1,labels =rownames(x),hang=-1,main="dandrogram:single")  
abline(h=18,col="red", lty = 3,lwd = 2)  
plot(hc2,labels =rownames(x),hang=-1,main="complete linkage")  
plot(hc3,labels =rownames(x),hang=-1,main="Ward Method")  
plot(hc4,labels = rownames(x),hang=-1,main="Average linkage")



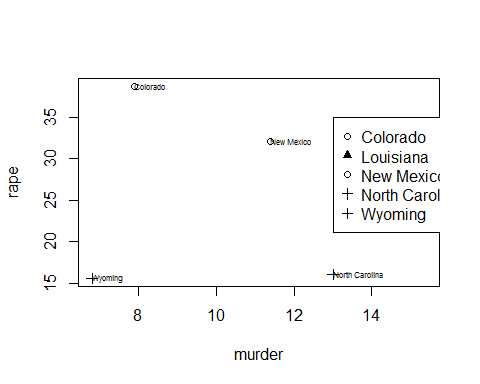
### 3.6 Cutree function

cutree 함수를 이용하여 군집을 나눌 수 있다.

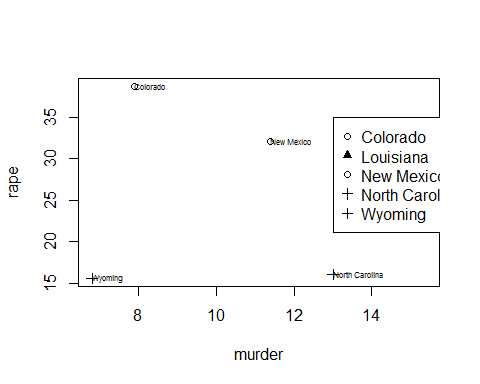
cl.num = 3  
colnames(x) = c("murder","rape")  
  
hc1.result = cutree(hc1,k=cl.num)  
plot(x,pch=hc1.result)  
text(x,labels = rownames(x), adj=0,cex=0.5, main ="single")  
legend(x= 13, y= 35,  
 c('Colorado','Louisiana','New Mexico','North Carolina', 'Wyoming'),  
 pch = c(1,17,3,17,17))



hc2.result = cutree(hc2,k=cl.num)  
plot(x,pch=hc2.result)  
text(x,labels = rownames(x), adj=0,cex=0.5, main ="single")  
legend(x= 13, y= 35,  
 c('Colorado','Louisiana','New Mexico','North Carolina', 'Wyoming'),  
 pch = c(1,17,1,3,3))



hc3.result = cutree(hc2,k=cl.num)  
plot(x,pch=hc3.result)  
text(x,labels = rownames(x), adj=0,cex=0.5, main ="single")  
legend(x= 13, y= 35,  
 c('Colorado','Louisiana','New Mexico','North Carolina', 'Wyoming'),  
 pch = c(1,17,1,3,3))



plot : 기호 모양 pch, 크기 cex, 선 유형 lty, 선 두께 lwd

## 4. Nonhierarchical clustering method

### 4.1 K-means clustering

kmeans 함수를 사용한다.

str(x)

'data.frame': 5 obs. of 2 variables:  
 $ murder: num 7.9 15.4 11.4 13 6.8  
 $ rape : num 38.7 22.2 32.1 16.1 15.6

crime\_k = kmeans(x,centers=3)  
attributes(crime\_k)

$names  
[1] "cluster" "centers" "totss" "withinss"   
[5] "tot.withinss" "betweenss" "size" "iter"   
[9] "ifault"   
  
$class  
[1] "kmeans"

crime\_k$cluster

Colorado Louisiana New Mexico North Carolina Wyoming   
 2 1 2 3 3

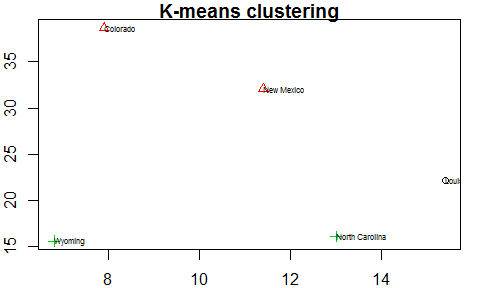
kc = table(crime\_k$cluster) ## number of each cluster  
kc

1 2 3   
1 2 2

par("mar")

[1] 5.1 4.1 4.1 2.1

par(mar=c(2,2,1,1))  
par(mfrow=c(1,1))  
plot(x, pch = crime\_k$cluster, col = crime\_k$cluster, main= "K-means clustering")  
text(x,labels=rownames(x),adj=0,cex=0.5)



### 4.2 Model-based clustering

모형기반 군집 방법에서는 "mclust"package를 사용한다. package 에는 아래와 같이 묘사되어 있다.

"The optimal model according to BIC for EM initialized by hierarchical clustering for parameterized Gaussian mixture models."

library(mclust)

Warning: package 'mclust' was built under R version 3.3.2

Package 'mclust' version 5.2.1

Type 'citation("mclust")' for citing this R package in publications.

군집 개수를 지정한 2~5개 사이에서 자동으로 적절한 군집을 선택해준다.

crime\_mc = Mclust(x,2:5)  
crime\_mc

'Mclust' model object:  
 best model: ellipsoidal, equal shape (VEV) with 2 components

attributes(crime\_mc)

$names  
 [1] "call" "data" "modelName" "n"   
 [5] "d" "G" "BIC" "bic"   
 [9] "loglik" "df" "hypvol" "parameters"   
[13] "z" "classification" "uncertainty"   
  
$class  
[1] "Mclust"

mc = table(crime\_mc$classification)  
mc

1 2   
3 2

plot(crime\_mc, what = "classification")

