BÀI TẬP PHÂN TÍCH PHÂN LỚP, PHÂN BIỆT

Bùi Khánh Duy

2023-05-17

Bài 1: (6.3)

1. Nhập dữ liệu measure vào R.

```
data = read.csv("measure.csv", header=T)
```

2. Tính khoảng cách Euclide giữa các quan sát trên các biến "chest", "waist", "hips".

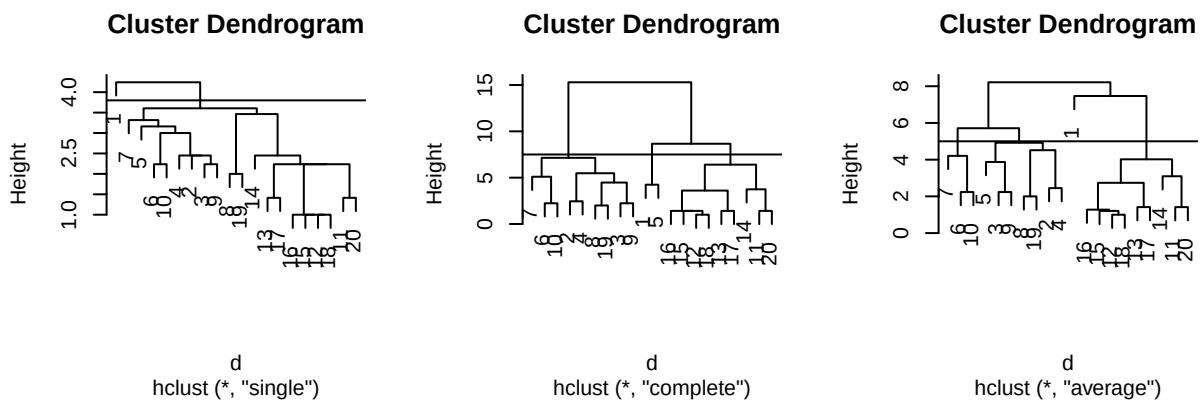
```
d = dist(data[, c("chest", "waist", "hips")])
d
```

```
##           1         2         3         4         5         6         7
## 2   6.164414
## 3   5.656854  2.449490
## 4   7.874008  2.449490  4.690416
## 5   4.242641  5.099020  3.162278  7.483315
## 6  11.000000  6.082763  5.744563  7.141428  7.681146
## 7  12.041595  5.916080  7.000000  5.000000 10.049876  5.099020
## 8   8.944272  3.741657  4.000000  3.741657  7.071068  5.744563  4.123106
## 9   7.810250  3.605551  2.236068  5.385165  4.582576  3.741657  5.830952
## 10 10.099505  4.472136  4.690416  5.099020  7.348469  2.236068  3.316625
## 11  7.000000  8.306624  6.403124  9.848858  5.744563 11.045361 12.083046
## 12  7.348469  7.071068  5.477226  8.246211  6.000000  9.949874 10.246951
## 13  7.810250  8.544004  7.280110  9.433981  7.549834 12.083046 11.916375
## 14  8.306624 11.180340  9.643651 12.449900  8.660254 14.696938 15.297059
## 15  7.483315  6.164414  4.898979  7.071068  6.164414  9.219544  9.000000
## 16  7.071068  6.000000  4.242641  7.348469  5.099020  8.544004  9.110434
## 17  7.810250  7.681146  6.708204  8.306624  7.549834 11.401754 10.770330
## 18  6.708204  6.082763  4.582576  7.280110  5.385165  9.273618  9.486833
## 19  9.165151  5.099020  4.472136  5.477226  7.071068  6.708204  5.744563
## 20  7.681146  9.433981  7.681146 10.816654  7.000000 12.409674 13.190906
##           8         9        10        11        12        13        14
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9   3.605551
```

```
## 10  3.741657  3.000000
## 11  8.062258  7.483315 10.246951
## 12  6.164414  6.403124  8.831761  2.236068
## 13  7.810250  8.485281 10.816654  2.828427  2.236068
## 14 11.180340 11.045361 13.747727  3.741657  5.196152  3.741657
## 15  4.898979  5.744563  7.874008  3.605551  1.414214  3.000000  6.403124
## 16  5.099020  5.000000  7.483315  3.000000  1.414214  3.605551  6.403124
## 17  6.708204  7.874008  9.949874  3.741657  2.236068  1.414214  5.099020
## 18  5.385165  5.656854  8.062258  2.828427  1.000000  2.828427  5.830952
## 19  2.000000  4.123106  5.099020  6.708204  4.690416  6.403124  9.848858
## 20  9.110434  8.831761 11.532563  1.414214  3.000000  2.449490  2.449490
##           15        16        17        18        19
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16  1.414214
## 17  2.236068  3.316625
## 18  1.000000  1.000000  2.449490
## 19  3.464102  3.741657  5.385165  4.123106
## 20  4.358899  4.123106  3.741657  3.741657  7.681146
```

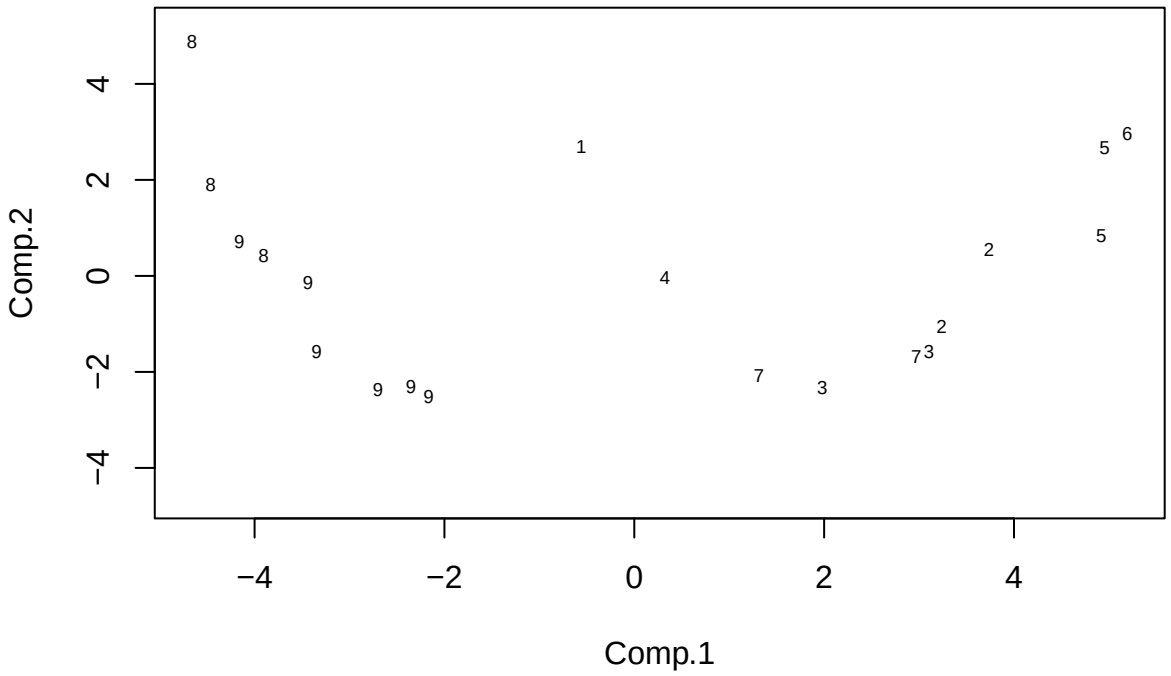
3. Tạo khung vẽ đồ thị gồm 2 hàng và 3 cột. Vẽ 3 biểu đồ dendrogram ở hàng trên và 3 đồ thị về hai thành phần chính đầu tiên tương ứng với 3 phương pháp "single", "complete", "average". Vẽ đường thẳng cắt ngang y = 3.8, y = 7.5 và y = 5.5 tương ứng trong 3 biểu đồ dendrogram.

```
par(mfrow = c(2,3))
plot(cs <- hclust(d, method = "single"))
abline(h = 3.8)
plot(cs <- hclust(d, method = "complete"))
abline(h=7.5)
plot(cs <- hclust(d, method = "average"))
abline(h=5)
```

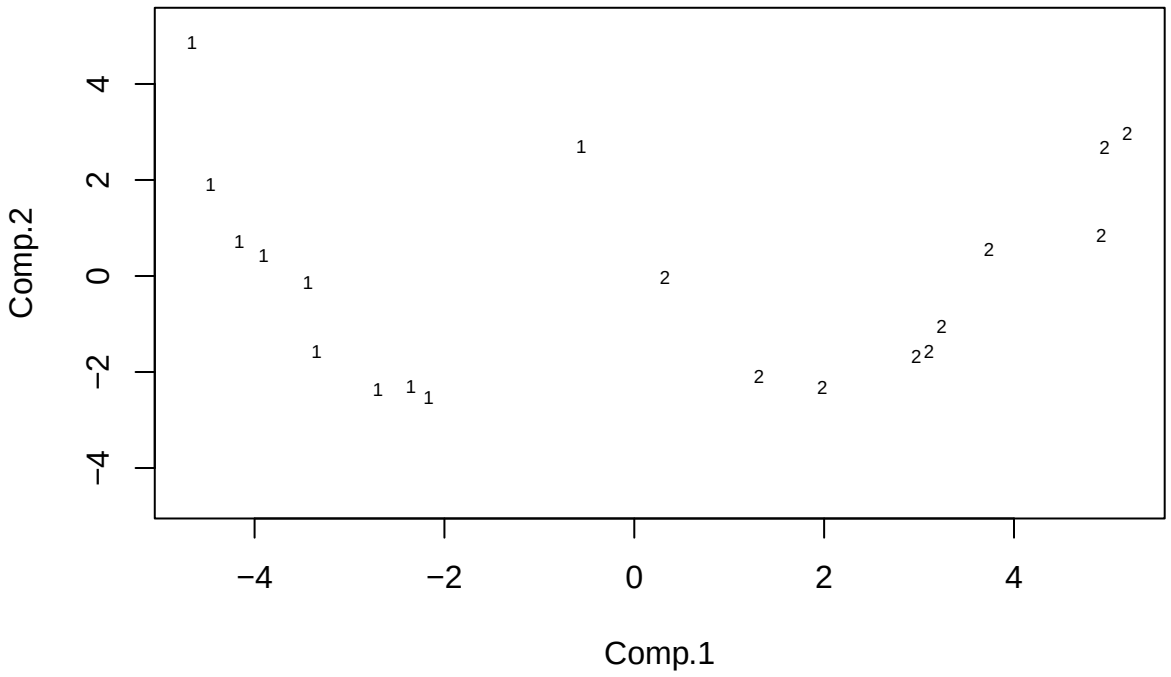


4. Với các ngưỡng khoảng cách được tạo ra ở trên, trong mỗi biểu đồ, các quan sát ban đầu được chia thành mấy nhóm? Mỗi nhóm gồm những quan sát nào?

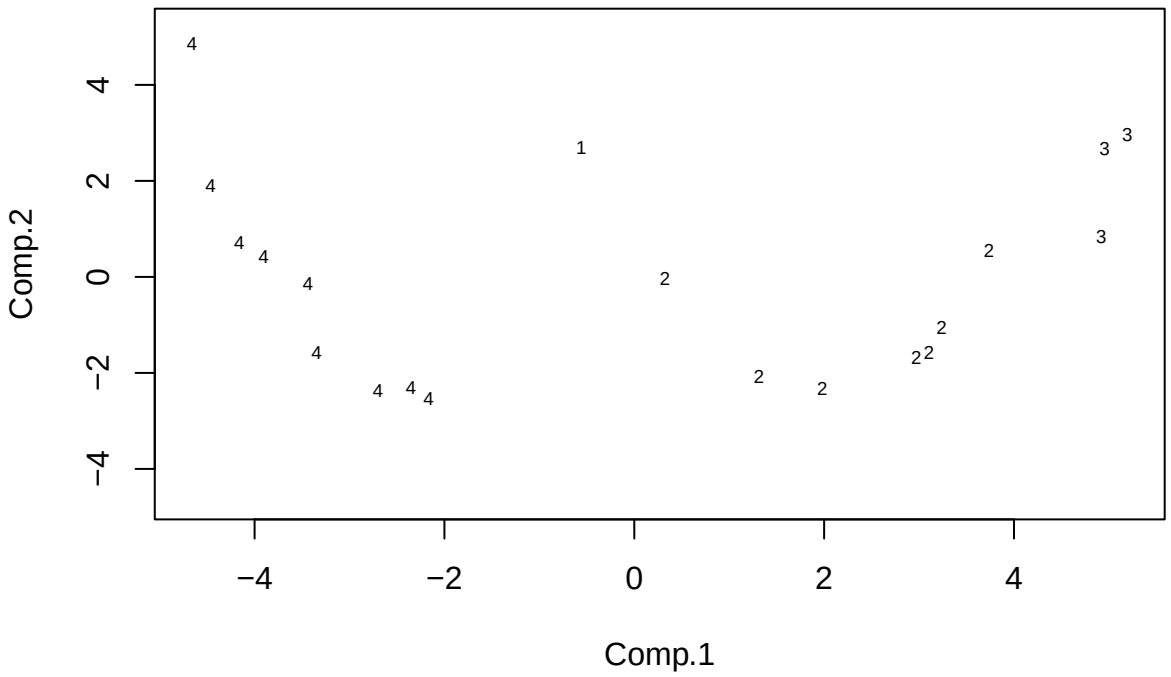
```
body_pc = princomp(d, cor = TRUE)
xlim = range(body_pc$scores[,1])
plot(body_pc$scores[,1:2], type = "n", xlim = xlim, ylim = xlim)
lab = cutree(cs, h = 3.8)
text(body_pc$scores[,1:2], labels = lab, cex = 0.6)
```




```
plot(body_pc$scores[,1:2], type = "n", xlim = xlim, ylim = xlim)
lab = cutree(cs, h = 7.5)
text(body_pc$scores[,1:2], labels = lab, cex = 0.6)
```




```
plot(body_pc$scores[,1:2], type = "n", xlim = xlim, ylim = xlim)
lab = cutree(cs, h = 5.5)
text(body_pc$scores[,1:2], labels = lab, cex = 0.6)
```



Bài 2: (6.3.1)

1. Nhập dữ liệu jet vào R.

```
dev.off()
```

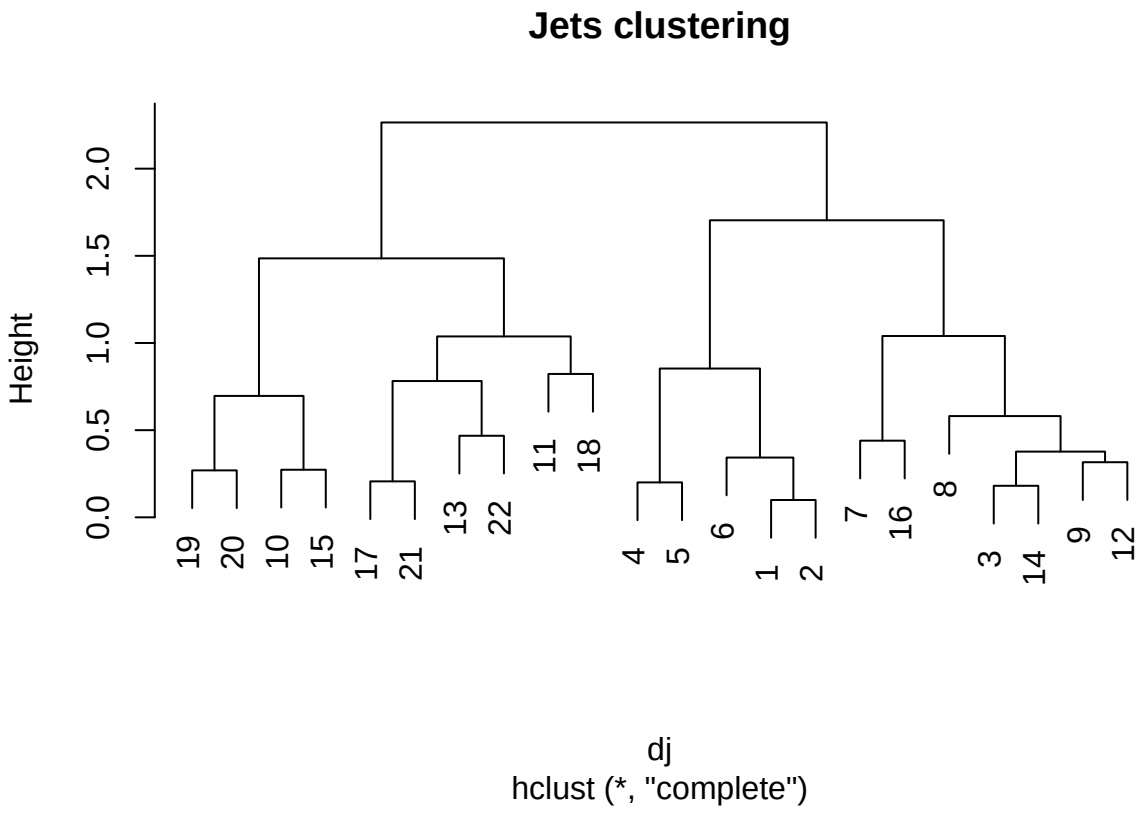
```
## null device
##           1
```

```
data2 = read.csv("jet.csv", header = T)
```

2. Tính khoảng cách Euclide sau khi chuẩn hóa giữa các quan sát trên các biến "FFD", "SPR", "RGF", "PLF", "SLF".

```
X = scale(data2[, c("SPR", "RGF", "PLF", "SLF")], center = F, scale = T)

dj = dist(X)
plot(cc <- hclust(dj), main = "Jets clustering")
```

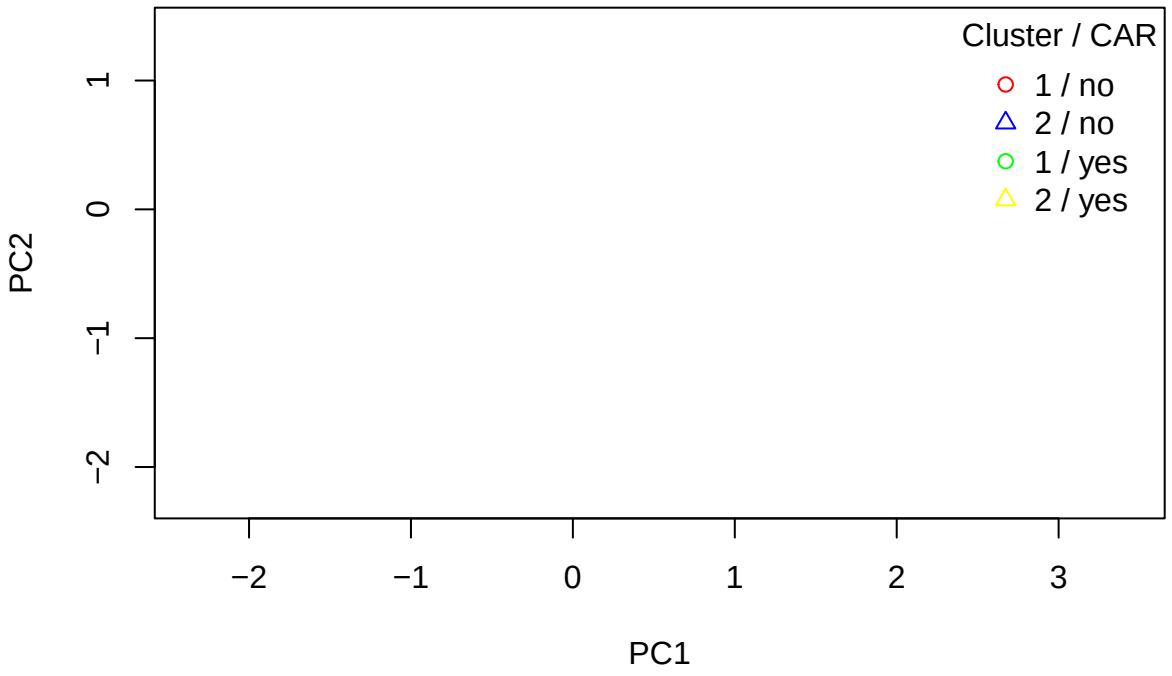



```
cc
```

```
##
## Call:
## hclust(d = dj)
##
## Cluster method   : complete
## Distance         : euclidean
## Number of objects: 22
```

3. Phân cụm phân cấp dựa trên cơ sở liên kết đầy đủ. Nếu khoảng cách giữa hai quan sát < 0.9 thì hai quan sát được gọi là cùng nhóm, nếu khoảng cách đó $>= 0.9$ thì hai quan sát được gọi là khác nhóm. Khi đó, các quan sát ban đầu được chia thành mấy nhóm? Mỗi nhóm gồm những quan sát nào?

```
pr = prcomp(dj)$x[, 1:2]
plot(pr, pch = (1:2)[cutree(cc, k = 2)], col = c("black", "darkgray")[data2$CAR], xlim = range(pr) * c(
legend("topright", col = c("red", "blue", "green", "yellow"), legend = c("1 / no", "2 / no", "1 / yes",
```



4. Vẽ và giải thích hình 6.6 (trang 174).

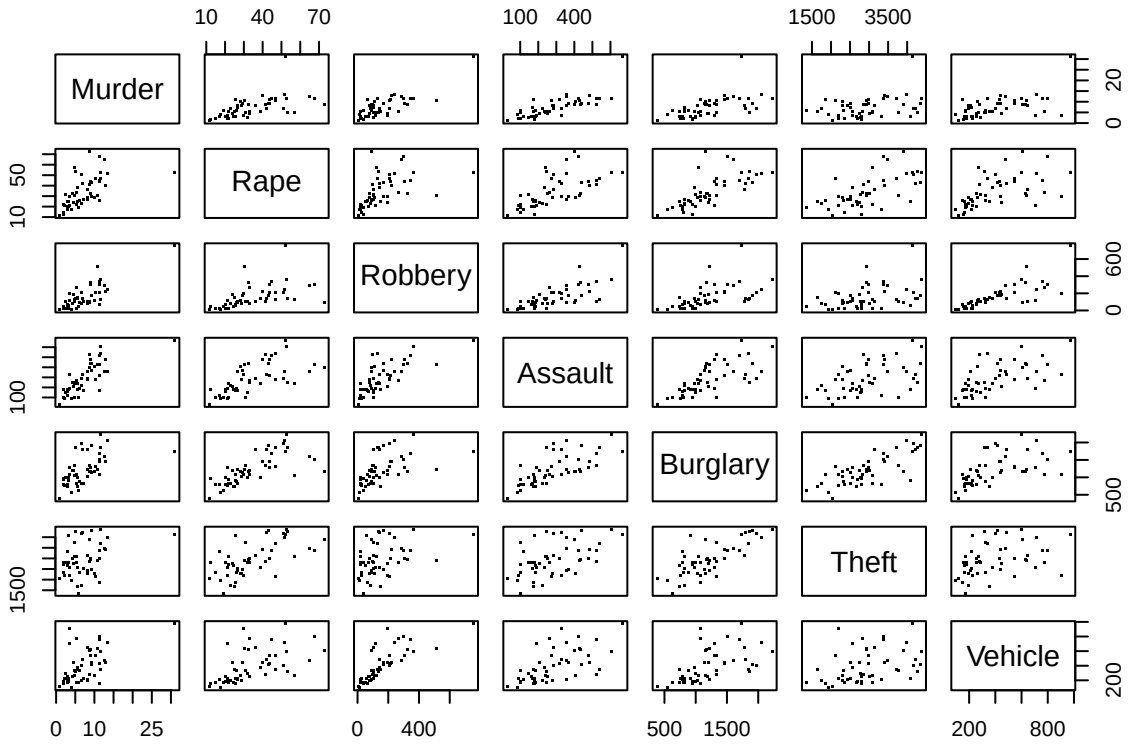
Bài 3: (6.4)

1. Nhập dữ liệu crime vào R.

```
data3 = read.csv("crime.csv", header = T)
```

2. Vẽ ma trận biểu đồ tán xạ. Nhận xét về bang khác biệt so với các quan sát còn lại.

```
pairs(data3[, -1], pch=".", cex=1.5)
```

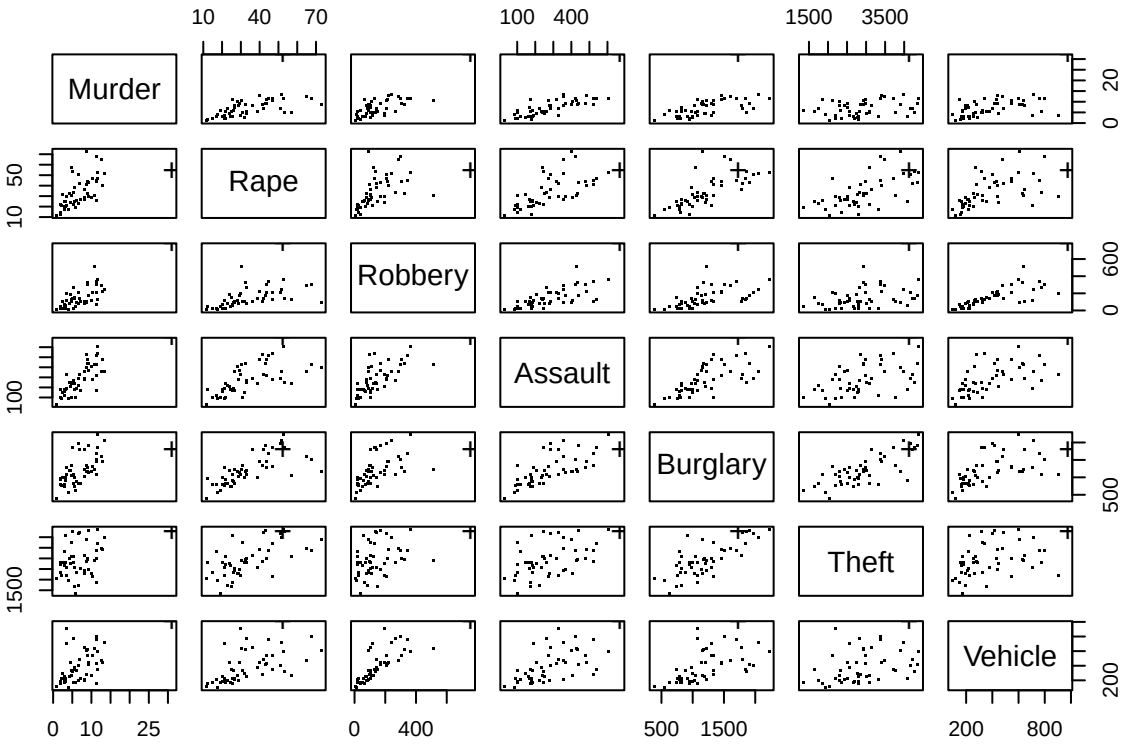


3. Vẽ biểu đồ tán xạ thể hiện rõ sự khác biệt của bang đó. Nhận xét về tỷ lệ các tội phạm các loại ở bang đó.

```
subset(data3, Murder > 15)
```

```
##    State Murder Rape Robbery Assault Burglary Theft Vehicle
## 24    DC     31 52.4     754     668     1728  4131     975
```

```
pch_vec = rep(".", nrow(data3[,-1]))
pch_vec[24] = "+"
pairs(data3[,-1], pch = pch_vec, cex = 1.5)
```



4. Tính phương sai và phương sai sau khi chuẩn hóa của các biến.

```
var = cov(data3[,-1])
sapply(data3[,-1], var)
```

```
##       Murder         Rape      Robbery      Assault     Burglary        Theft
##     23.20215    212.31228  18993.37020  22004.31294 177912.83373 582812.83843
##      Vehicle
##  50007.37490
```

```
rge = sapply(data3[,-1], function(x) diff(range(x)))
crime_s = sweep(data3[,-1], 2, rge, FUN = "/")
sapply(crime_s, var)
```

```
##     Murder       Rape    Robbery    Assault   Burglary      Theft    Vehicle
## 0.02578017 0.05687124 0.03403775 0.05439933 0.05277909 0.06411424 0.06516672
```

5. Thực hiện phân cụm K-means với k = 2. Đưa ra tâm (giá trị trung bình) của 2 nhóm. Bang ND và bang SD được xếp vào nhóm nào?

```
kmeans(crime_s, centers = 2)$centers * rge
```

```
##      Murder     Rape  Robbery    Assault  Burglary    Theft   Vehicle
## 1 10.359091 567.6133 628.0876 562.400515  52.25841  726.112 1991.5395
## 2  9.965621 259.7625 311.3398   8.893949 378.99966 1560.437  253.6552
```

Bài 4: (6.4.2)

```r
n = nrow(crime_s)
wss = rep(0, 6)
wss[1] = (n - 1) * sum(sapply(crime_s, var))
for (i in 2:6) {
  wss[i] = sum(kmeans(crime_s, centers = i)$withinss)
}
plot(1:6, wss, type = "b", xlab = "Number of groups", ylab = "Within groups sum of squares")
```