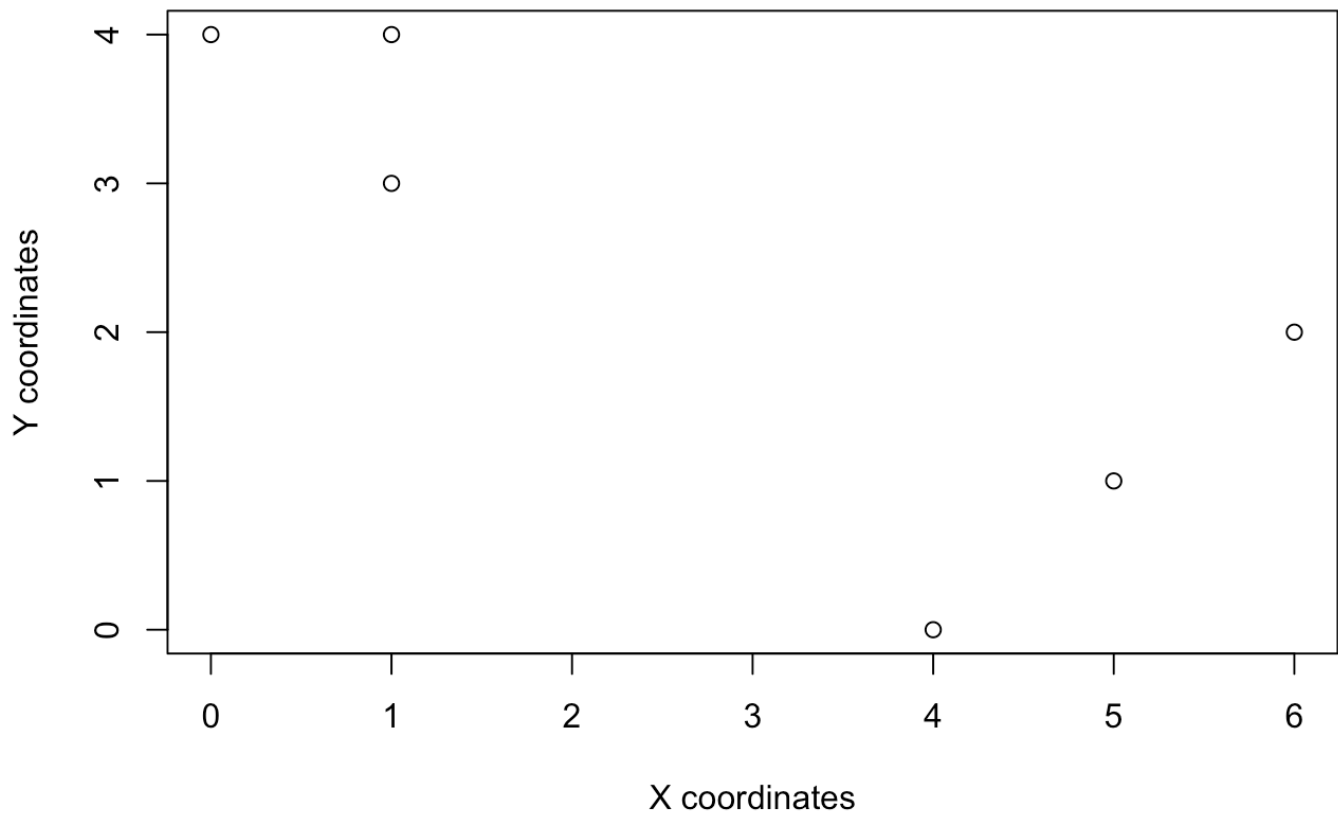# Misc

I accidentally did problem 3 from the textbook, so here's a scatter plot:
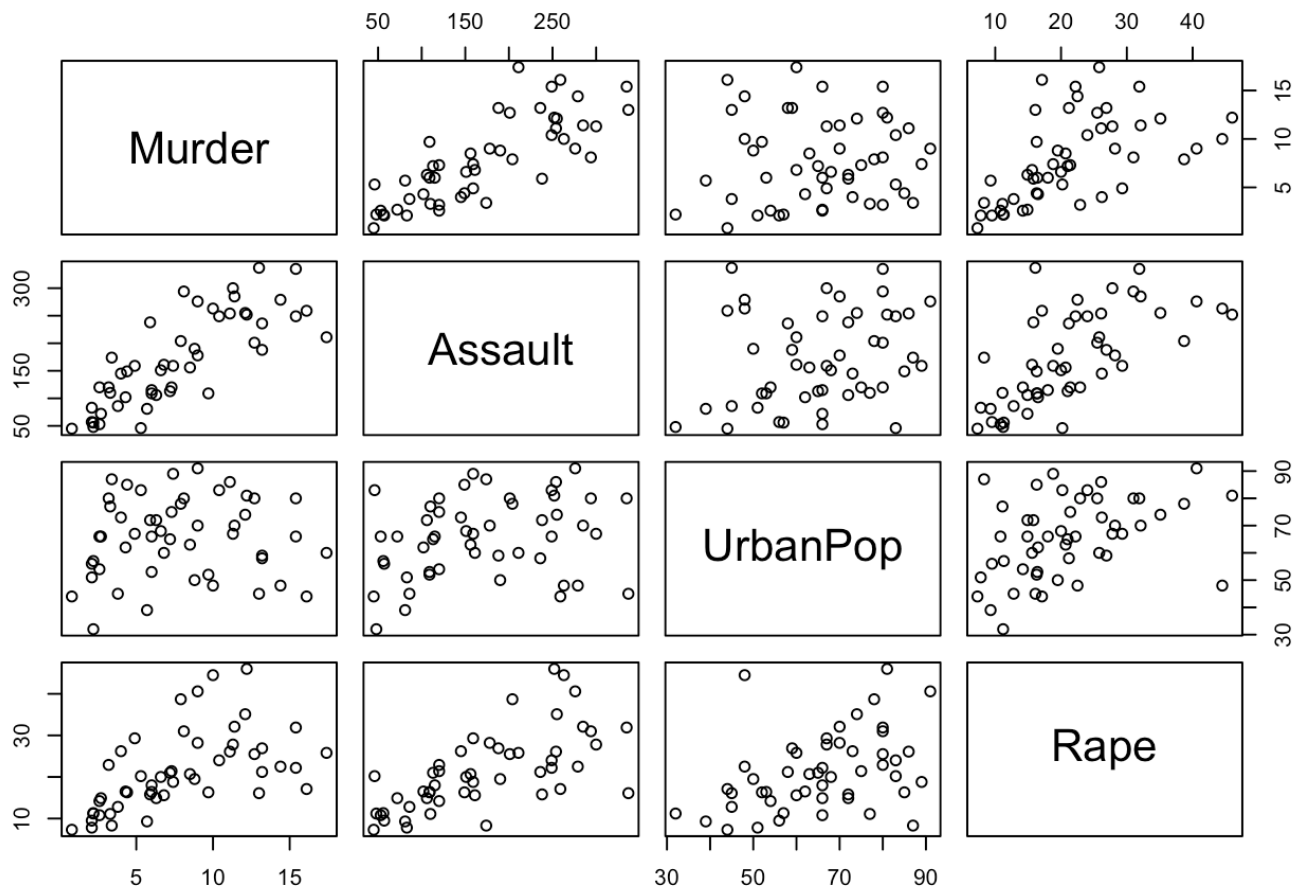
```
pts = matrix(
   c(1, 1, 0, 5, 6, 4, 4, 3, 4, 1, 2, 0),
   nrow = 6,
   ncol = 2
)
colnames(pts) = c('X coordinates', 'Y coordinates')
plot(pts)
```
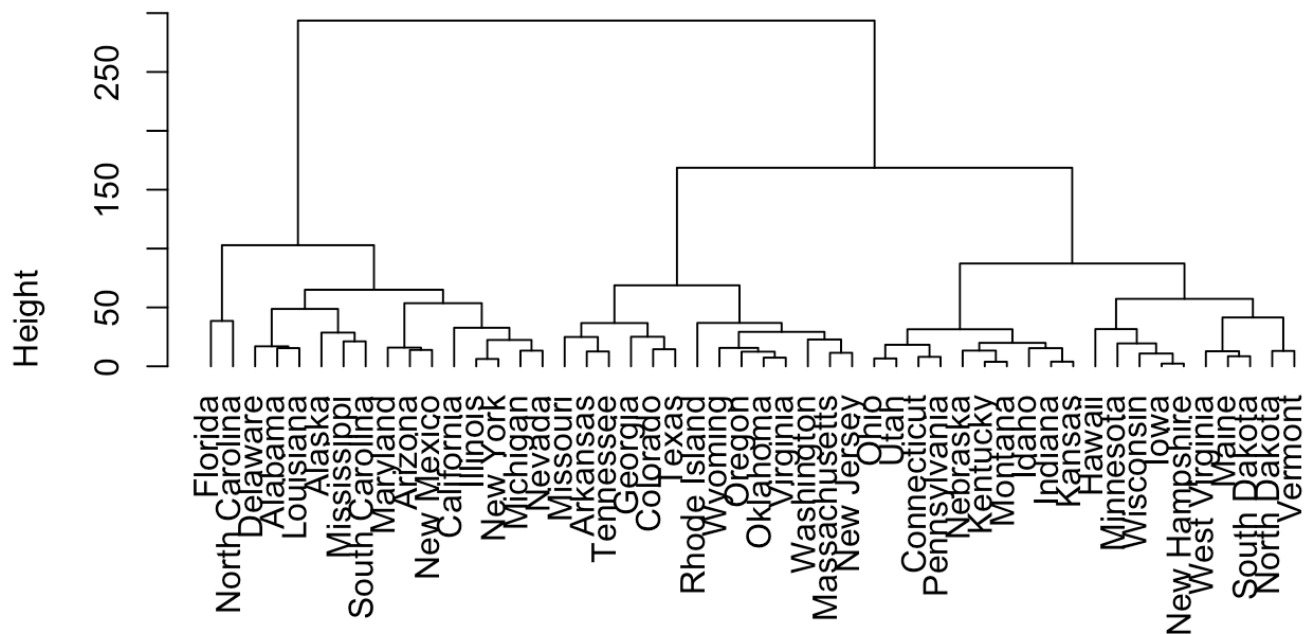

Problem 3

Part A

```
arrests=USArrests
plot(USArrests)
```

```
hc <- hclust(dist(USArrests), "complete")
plot(hc, hang = -1, main = 'Complete linkage + Euclidian distance dendogram')
```

Complete linkage + Euclidian distance dendogram

dist(USArrests)
hclust (*, "complete")

Part B

Note: I relied heavily on this website (https://www.biostars.org/p/86563/) to solve this problem.

```
# cutree returns a vector of cluster membership in the order of the original data
rows examine it
clusters <- cutree(hc, k=10)

## to grab a cluster:
# cluster1 <- USArrests[clusters == 1,]

# to add the cluster ID to your data:
all_clusters <- cbind(USArrests, clusterID=clusters)

# examine the data with cluster ids attached, ordered by the 'clusterID' column
all_clusters[order(all_clusters[['clusterID']]), ]
```

```
##                 Murder Assault UrbanPop Rape clusterID
## Alabama           13.2    236       58 21.2          1
## Delaware           5.9    238       72 15.8          1
## Louisiana         15.4    249       66 22.2          1
## Alaska            10.0    263       48 44.5          2
## Mississippi       16.1    259       44 17.1          2
## South Carolina    14.4    279       48 22.5          2
## Arizona            8.1    294       80 31.0          3
## Maryland          11.3    300       67 27.8          3
## New Mexico        11.4    285       70 32.1          3
## Arkansas           8.8    190       50 19.5          4
## Colorado           7.9    204       78 38.7          4
## Georgia           17.4    211       60 25.8          4
## Missouri           9.0    178       70 28.2          4
## Tennessee         13.2    188       59 26.9          4
## Texas             12.7    201       80 25.5          4
## California         9.0    276       91 40.6          5
## Illinois          10.4    249       83 24.0          5
## Michigan          12.1    255       74 35.1          5
## Nevada            12.2    252       81 46.0          5
## New York          11.1    254       86 26.1          5
## Connecticut        3.3    110       77 11.1          6
## Idaho              2.6    120       54 14.2          6
## Indiana            7.2    113       65 21.0          6
## Kansas             6.0    115       66 18.0          6
## Kentucky           9.7    109       52 16.3          6
## Montana            6.0    109       53 16.4          6
## Nebraska           4.3    102       62 16.5          6
## Ohio               7.3    120       75 21.4          6
## Pennsylvania       6.3    106       72 14.9          6
## Utah               3.2    120       80 22.9          6
## Florida           15.4    335       80 31.9          7
## North Carolina    13.0    337       45 16.1          7
## Hawaii             5.3     46       83 20.2          8
## Iowa               2.2     56       57 11.3          8
## Minnesota          2.7     72       66 14.9          8
## New Hampshire      2.1     57       56  9.5          8
## Wisconsin          2.6     53       66 10.8          8
## Maine              2.1     83       51  7.8          9
## North Dakota       0.8     45       44  7.3          9
## South Dakota       3.8     86       45 12.8          9
## Vermont            2.2     48       32 11.2          9
## West Virginia      5.7     81       39  9.3          9
## Massachusetts      4.4    149       85 16.3         10
## New Jersey         7.4    159       89 18.8         10
## Oklahoma           6.6    151       68 20.0         10
## Oregon             4.9    159       67 29.3         10
## Rhode Island       3.4    174       87  8.3         10
## Virginia           8.5    156       63 20.7         10
## Washington         4.0    145       73 26.2         10
```

```
## Wyoming                6.8      161        60 15.6          10
```

Part C

```
scaled_arrests = scale(USArrests)
scaled_arrests
```

```
##                   Murder     Assault    UrbanPop          Rape
## Alabama         1.24256408  0.78283935 -0.52090661 -0.003416473
## Alaska          0.50786248  1.10682252 -1.21176419  2.484202941
## Arizona         0.07163341  1.47880321  0.99898006  1.042878388
## Arkansas        0.23234938  0.23086801 -1.07359268 -0.184916602
## California      0.27826823  1.26281442  1.75892340  2.067820292
## Colorado        0.02571456  0.39885929  0.86080854  1.864967207
## Connecticut    -1.03041900 -0.72908214  0.79172279 -1.081740768
## Delaware       -0.43347395  0.80683810  0.44629400 -0.579946294
## Florida         1.74767144  1.97077766  0.99898006  1.138966691
## Georgia         2.20685994  0.48285493 -0.38273510  0.487701523
## Hawaii         -0.57123050 -1.49704226  1.20623733 -0.110181255
## Idaho          -1.19113497 -0.60908837 -0.79724965 -0.750769945
## Illinois        0.59970018  0.93883125  1.20623733  0.295524916
## Indiana        -0.13500142 -0.69308401 -0.03730631 -0.024769429
## Iowa           -1.28297267 -1.37704849 -0.58999237 -1.060387812
## Kansas         -0.41051452 -0.66908525  0.03177945 -0.345063775
## Kentucky        0.43898421 -0.74108152 -0.93542116 -0.526563903
## Louisiana       1.74767144  0.93883125  0.03177945  0.103348309
## Maine          -1.30593210 -1.05306531 -1.00450692 -1.434064548
## Maryland        0.80633501  1.55079947  0.10086521  0.701231086
## Massachusetts  -0.77786532 -0.26110644  1.34440885 -0.526563903
## Michigan        0.99001041  1.01082751  0.58446551  1.480613993
## Minnesota      -1.16817555 -1.18505846  0.03177945 -0.676034598
## Mississippi     1.90838741  1.05882502 -1.48810723 -0.441152078
## Missouri        0.27826823  0.08687549  0.30812248  0.743936999
## Montana        -0.41051452 -0.74108152 -0.86633540 -0.515887425
## Nebraska       -0.80082475 -0.82507715 -0.24456358 -0.505210947
## Nevada          1.01296983  0.97482938  1.06806582  2.644350114
## New Hampshire  -1.30593210 -1.36504911 -0.65907813 -1.252564419
## New Jersey     -0.08908257 -0.14111267  1.62075188 -0.259651949
## New Mexico      0.82929443  1.37080881  0.30812248  1.160319648
## New York        0.76041616  0.99882813  1.41349461  0.519730957
## North Carolina  1.19664523  1.99477641 -1.41902147 -0.547916860
## North Dakota   -1.60440462 -1.50904164 -1.48810723 -1.487446939
## Ohio           -0.11204199 -0.60908837  0.65355127  0.017936483
## Oklahoma       -0.27275797 -0.23710769  0.16995096 -0.131534211
## Oregon         -0.66306820 -0.14111267  0.10086521  0.861378259
## Pennsylvania   -0.34163624 -0.77707965  0.44629400 -0.676034598
## Rhode Island   -1.00745957  0.03887798  1.48258036 -1.380682157
## South Carolina  1.51807718  1.29881255 -1.21176419  0.135377743
## South Dakota   -0.91562187 -1.01706718 -1.41902147 -0.900240639
## Tennessee       1.24256408  0.20686926 -0.45182086  0.605142783
## Texas           1.12776696  0.36286116  0.99898006  0.455672088
## Utah           -1.05337842 -0.60908837  0.99898006  0.178083656
## Vermont        -1.28297267 -1.47304350 -2.31713632 -1.071064290
## Virginia        0.16347111 -0.17711080 -0.17547783 -0.056798864
## Washington     -0.86970302 -0.30910395  0.51537975  0.530407436
## West Virginia  -0.47939280 -1.07706407 -1.83353601 -1.273917376
## Wisconsin      -1.19113497 -1.41304662  0.03177945 -1.113770203
```

```
## Wyoming          -0.22683912 -0.11711392 -0.38273510 -0.601299251
## attr(,"scaled:center")
##   Murder  Assault UrbanPop      Rape
##    7.788  170.760   65.540   21.232
## attr(,"scaled:scale")
##    Murder   Assault  UrbanPop       Rape
##  4.355510 83.337661 14.474763  9.366385
```

```
hc <- hclust(dist(scaled_arrests), "complete")
plot(hc, hang = -1, main = 'Complete linkage + Euclidian distance dendogram, scale
d')
```


Complete linkage + Euclidian distance dendogram, scaled
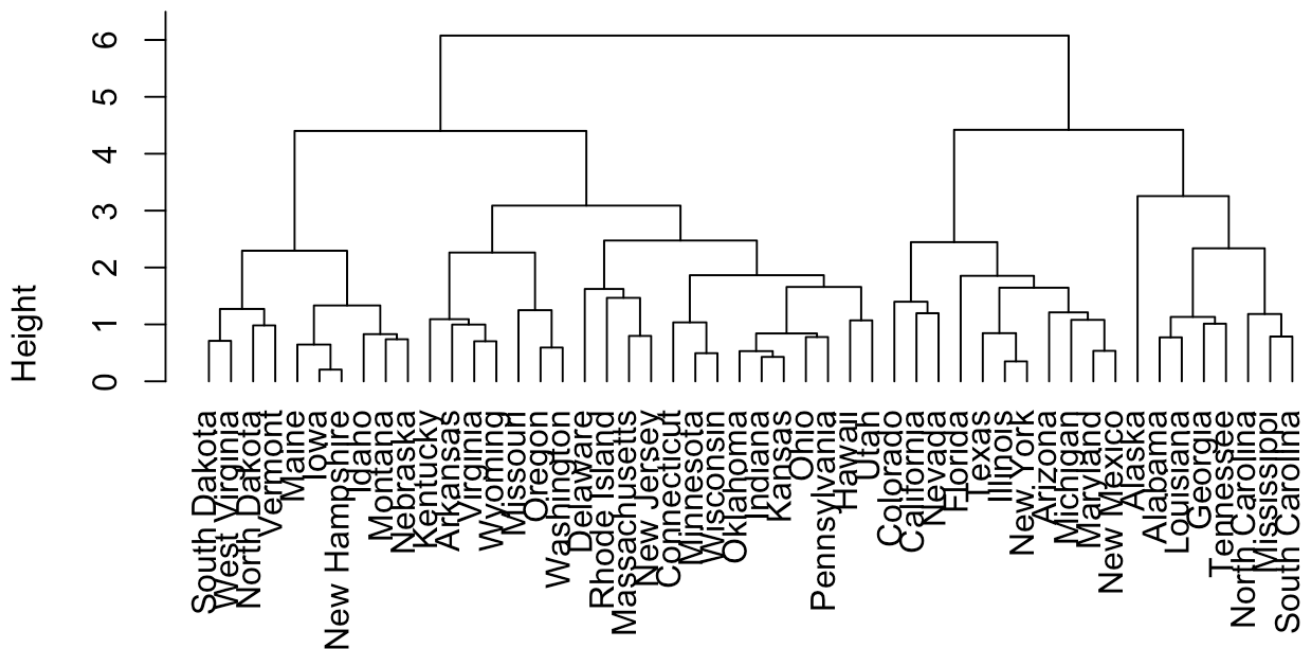
dist(scaled_arrests)
hclust (*, "complete")

Part D

A major effect of scaling the variables is that the cluster sizes are more evenly sized. I prefer it with more evenly distributed clusters, because with uneven group sizes you get some really large groups that can contain more variance than they suggest. For instance, before scaling the variables the sheer magnitude (rather than per-capita) of crime committed in each state likely played a role in determining the clusters. As a result, the groups tended to indicate the approximate population of the states. For instance, cluster 6 in

part b tended to contain states with a middle- to low-level population, the main outliers being PA and OH. State population isn't what we're trying to measure here, so it confounds our groups. Rather, we'd rather cluster states by their crime rates relative to each other.

```
clusters <- cutree(hc, k=10)
all_clusters <- cbind(USArrests, clusterID=clusters)
all_clusters[order(all_clusters[['clusterID']]), ]
```

```
##                Murder Assault UrbanPop Rape clusterID
## Alabama           13.2     236       58 21.2          1
## Georgia           17.4     211       60 25.8          1
## Louisiana         15.4     249       66 22.2          1
## Tennessee         13.2     188       59 26.9          1
## Alaska            10.0     263       48 44.5          2
## Arizona            8.1     294       80 31.0          3
## Florida           15.4     335       80 31.9          3
## Illinois          10.4     249       83 24.0          3
## Maryland          11.3     300       67 27.8          3
## Michigan          12.1     255       74 35.1          3
## New Mexico        11.4     285       70 32.1          3
## New York          11.1     254       86 26.1          3
## Texas             12.7     201       80 25.5          3
## Arkansas           8.8     190       50 19.5          4
## Kentucky           9.7     109       52 16.3          4
## Missouri           9.0     178       70 28.2          4
## Oregon             4.9     159       67 29.3          4
## Virginia           8.5     156       63 20.7          4
## Washington         4.0     145       73 26.2          4
## Wyoming            6.8     161       60 15.6          4
## California         9.0     276       91 40.6          5
## Colorado           7.9     204       78 38.7          5
## Nevada            12.2     252       81 46.0          5
## Connecticut        3.3     110       77 11.1          6
## Hawaii             5.3      46       83 20.2          6
## Indiana            7.2     113       65 21.0          6
## Kansas             6.0     115       66 18.0          6
## Minnesota          2.7      72       66 14.9          6
## Ohio               7.3     120       75 21.4          6
## Oklahoma           6.6     151       68 20.0          6
## Pennsylvania       6.3     106       72 14.9          6
## Utah               3.2     120       80 22.9          6
## Wisconsin          2.6      53       66 10.8          6
## Delaware           5.9     238       72 15.8          7
## Massachusetts      4.4     149       85 16.3          7
## New Jersey         7.4     159       89 18.8          7
## Rhode Island       3.4     174       87  8.3          7
## Idaho              2.6     120       54 14.2          8
## Iowa               2.2      56       57 11.3          8
## Maine              2.1      83       51  7.8          8
## Montana            6.0     109       53 16.4          8
## Nebraska           4.3     102       62 16.5          8
## New Hampshire      2.1      57       56  9.5          8
## Mississippi       16.1     259       44 17.1          9
## North Carolina    13.0     337       45 16.1          9
## South Carolina    14.4     279       48 22.5          9
## North Dakota       0.8      45       44  7.3         10
## South Dakota       3.8      86       45 12.8         10
## Vermont            2.2      48       32 11.2         10
```

```
## West Virginia        5.7        81        39  9.3        10
```