Business Analytics - ETC3250 2018 - Lab 6 - Solution

Clustering

Souhaib Ben Taieb

10 April 2018

Exercice 1

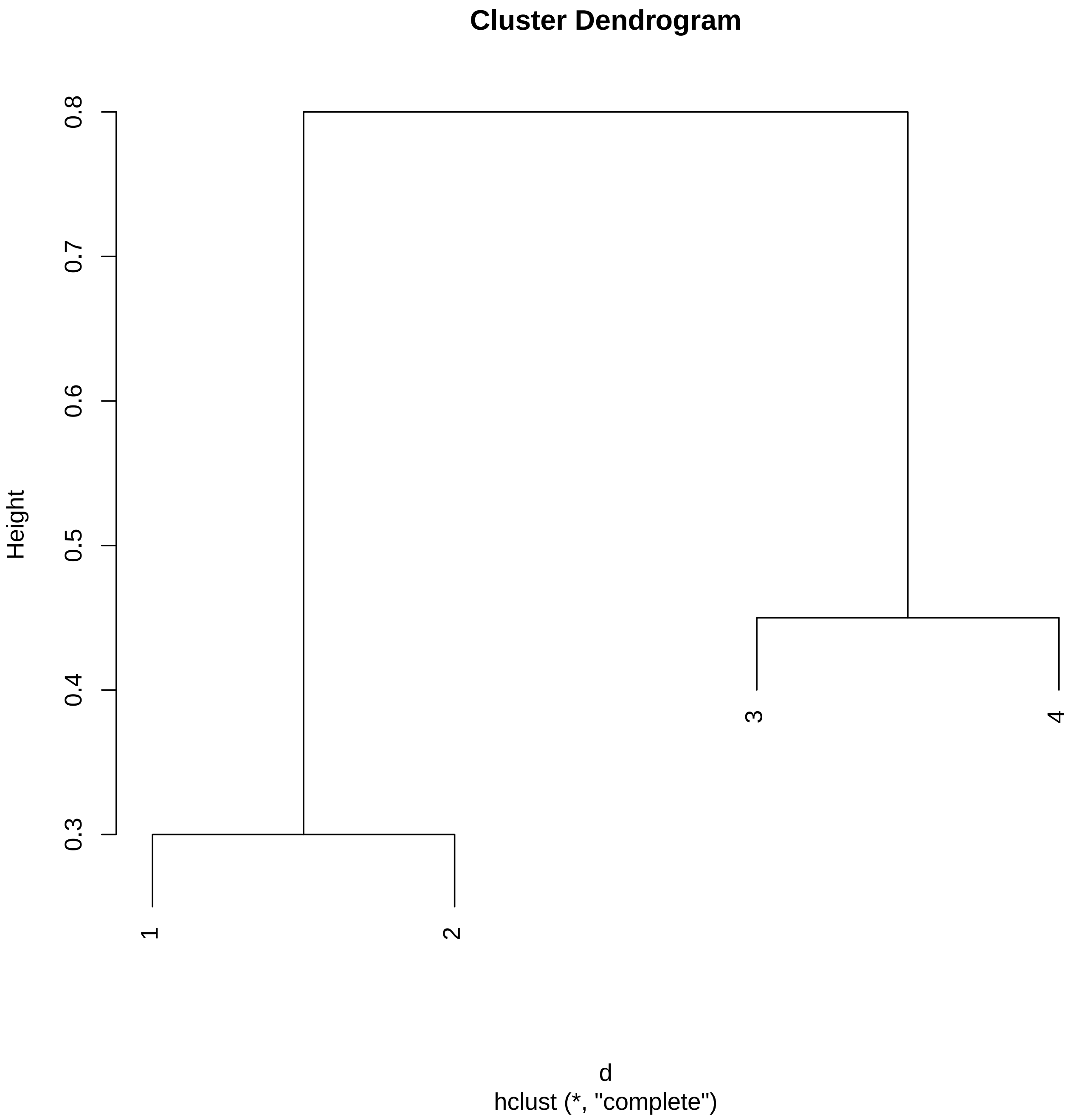
Read and run the code in Sections 10.5.1 and 10.5.2 of ISLR.

Exercice 2

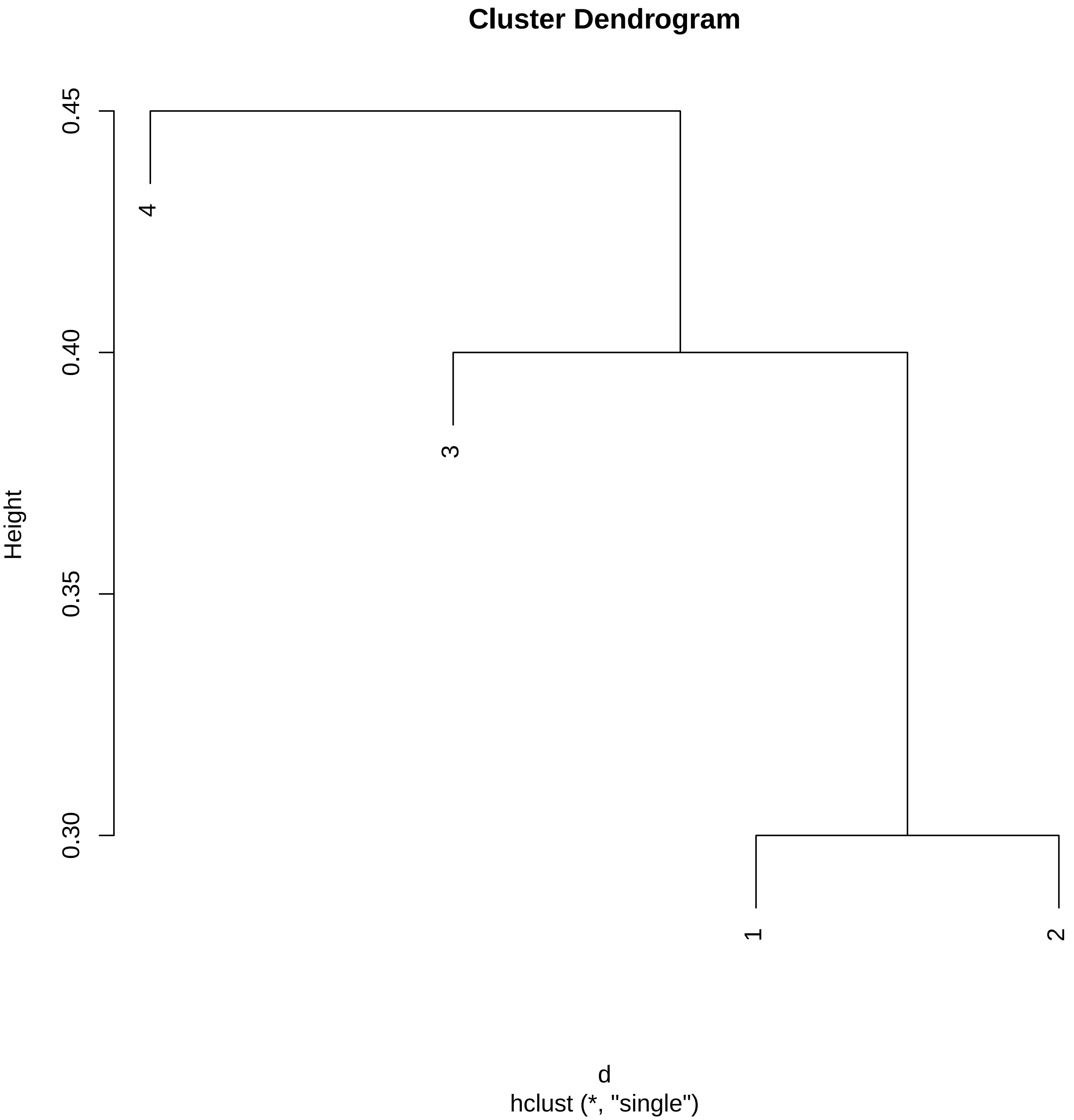
ISLR Section 10.7, exercises 2(a), 2(b), 2(c) and 2(d).

(a) and (b)

```
d <- as.dist(matrix(c(0, 0.3, 0.4, 0.7,
                     0.3, 0, 0.5, 0.8,
                     0.4, 0.5, 0.0, 0.45,
                     0.7, 0.8, 0.45, 0.0), nrow = 4))
plot(hclust(d, method="complete"))
```



```
plot(hclust(d, method="single"))
```



(c)

(1,2) and (3,4)

(d)

(4) and (1, 2, 3)

Exercice 3

1. Let $\{x_1, \ldots, x_n\}$ be a set of points where $x_i \in \mathbb{R}^d$, and let x be any point in $\mathbb{R}^d$. Prove that

$$\sum_{i=1}^{n} ||x_i - x||^2 = \sum_i ||x_i - \bar{x}||^2 + n \; ||\bar{x} - x||^2, \quad (1) \tag{1}$$

where $|| \cdot ||$ is the L_2 norm, and $\bar{x}$ is the centroid of the set of points, i.e. $\bar{x} = \frac{1}{n} \sum_i x_i$. (Hint: add and substract $\bar{x}$ in the left-hand side of the previous expression).

$$\begin{aligned} & \sum_{i=1}^{n} ||x_i - x||^2 & (2) \\ & = \sum_{i=1}^{n} ||x_i - \bar{x} + \bar{x} - x||^2 & (3) \\ & = \sum_{i=1}^{n} ||x_i - \bar{x}||^2 + ||\bar{x} - x||^2 + 2(x_i - \bar{x})^{'}(\bar{x} - x) & (4) \\ & = \sum_{i=1}^{n} ||x_i - \bar{x}||^2 + n \; ||\bar{x} - x||^2, & (5) \end{aligned}$$

since $\sum_i (x_i - \bar{x}) = 0$ (centroid definition).

2. Which value of x minimizes $\sum_{i=1}^{n} ||x_i - x||^2$? Prove it.

In (1) we can see that $x = \bar{x}$ is the minimizer. Suppose that $x = \tilde{x} \neq \bar{x}$ is the minimizer. Then we have

$$\begin{aligned} & \sum_{i=1}^{n} ||x_i - \tilde{x}||^2 < \sum_{i=1}^{n} ||x_i - \bar{x}||^2 & (6) \\ \implies & \sum_i ||x_i - \bar{x}||^2 + n \; ||\bar{x} - \tilde{x}||^2 \leq \sum_i ||x_i - \bar{x}||^2 + n \; ||\bar{x} - \bar{x}||^2 \leq \sum_i ||x_i - \bar{x}||^2 & (7) \end{aligned}$$

which is a contradiction since we assumed $\tilde{x} \neq \bar{x}$. So, the minimizer is $x = \tilde{x} = \bar{x}$.

3. Using expression (1), prove that

$$\sum_{i,j} ||x_i - x_j||^2 = 2n \sum_i ||x_i - \bar{x}||^2 \tag{8}$$

$$\begin{aligned} & \sum_{i,j} ||x_i - x_j||^2 & (9) \\ & = \sum_j (\sum_i ||x_i - \bar{x}||^2 + n \; ||\bar{x} - x_j||^2) & (10) \\ & = 2n \sum_i ||x_i - \bar{x}||^2 & (11) \end{aligned}$$

Data

The crime dataset (https://github.com/bsouhaib/BA2017/blob/master/data/crimes2008.csv) contains FBI crime rate statistics. These are the indices for 9 different types of crimes reported by the states of the USA, for 2008: violent, property, murder, rape, robbery, assault, burglary, ltheft (larceny theft), vtheft (vehicle theft). The values have been population adjusted so that the numers are per million people.

```
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:reshape':
##
##     rename

## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following object is masked from 'package:GGally':
##
##     nasa

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
crime <-  as.data.frame(read_csv(url("https://github.com/bsouhaib/BA2018/raw/master/data/crimes2008.csv
```

```
## Parsed with column specification:
## cols(
##   Year = col_integer(),
##   Population = col_integer(),
##   State = col_character(),
##   index = col_double(),
##   violent = col_double(),
##   property = col_double(),
##   murder = col_double(),
##   rape = col_double(),
##   robbery = col_double(),
##   assault = col_double(),
##   burglary = col_double(),
##   ltheft = col_double(),
##   vtheft = col_double()
## )
```

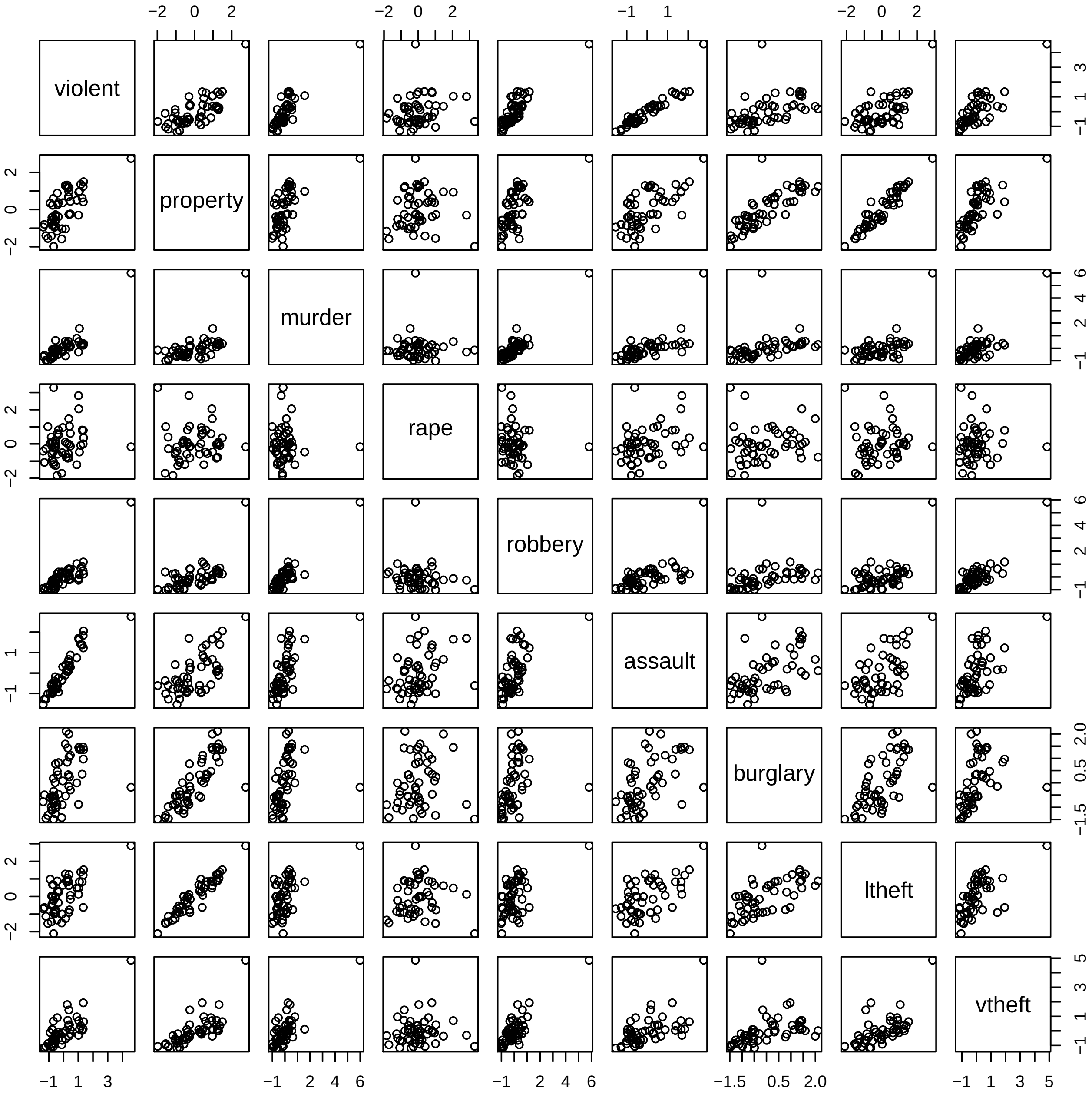
```
dim(crime)
crime <- crime[,-c(1,2,4)]
head(crime)
#crime[,-1] <- scale(crime[,-1])
```

```r
base::row.names(crime) <- crime[,1]
crime[, 2:10] <- scale(crime[,-1])
```

Exercice 4

Make a scatterplot matrix of the crime indices, with and without Washingto DC. Write a paragraph describing the relationships between the statistics, and about any observations about cluster patterns in the data.

```r
pairs(crime[,2:10])
```



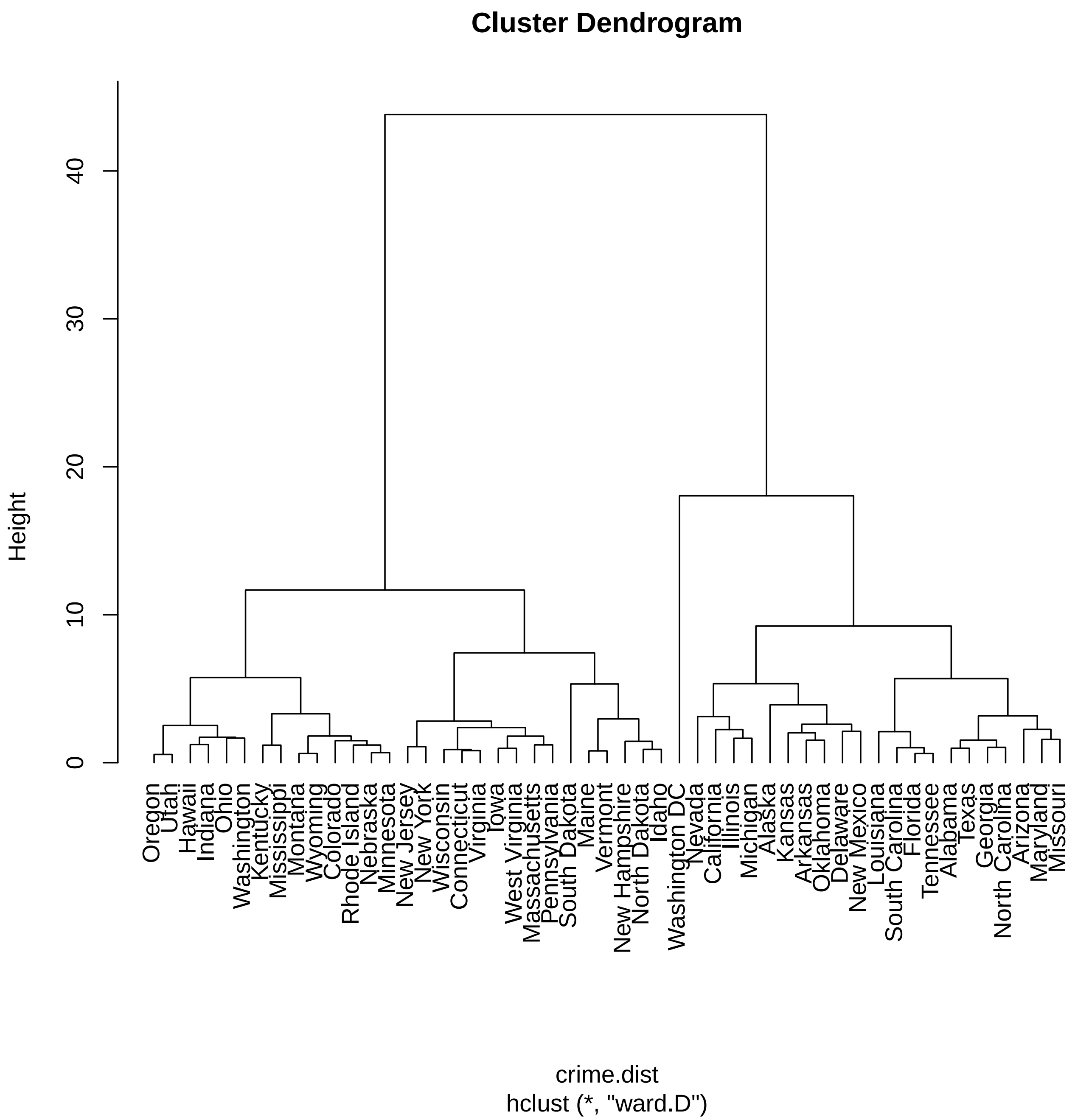
The pairwise relationships between the crime statistics reveal some outliers, and some positive association. One state has a very high rate of violent crimes, murder, robbery, larceny theft and vehicle theft. It also has the highest, albeit not by much, rate of property crime. (This is

Washington, DC.) Removing this case makes it easier to read the associations. Most crime statistics show a positive association. The strongest relationships are between property and larceny theft, assault and violent crime. Rape has a different relationship with other crime rates. It has no association with murder, burglary and theft crimes, and a slightly negative association with robbery! There area few states that have high vehicle theft but relatively other types of crimes.

Exercice 5

Cluster the states using hierarchical clustering, with Euclidean distance and wards linkage. Plot the dendrogram. How many clusters would be suggested by the dendrogram?

```r
crime.dist <- dist(crime[,-1])
crime.hc <- hclust(crime.dist, method="ward.D")
plot(crime.hc, hang=-1)
```



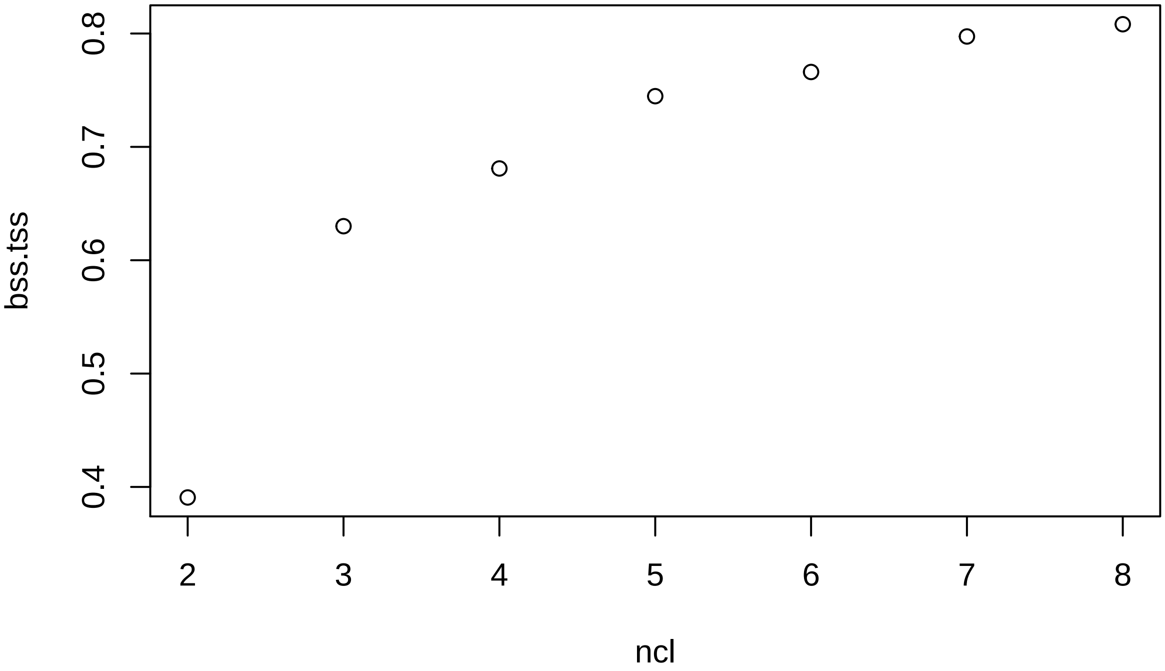
2 or 3, mostly. It might be interesting to look at 4, 5, 6 or more clusters, too.

Exercice 6

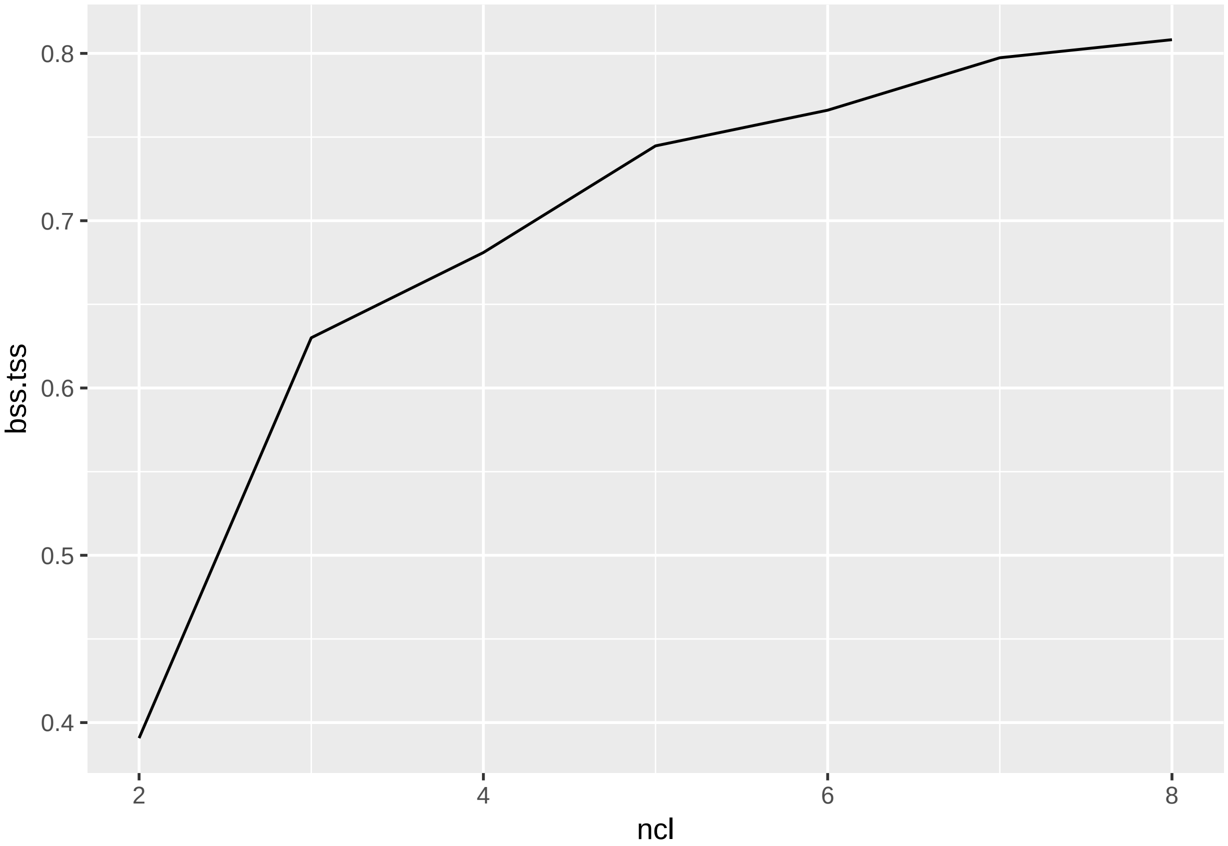
Use k-means clustering with k set to several different values, say 2-8. Calculate the ratio of between Sum of Squares (SS) to total SS for each value of k. Tabulate this. What is between SS? total SS? What happens to this value as k ranges from 2 to 8? Why is this? Also, what happens if you change the random seed, which changes the initialization of k-means?

```r
set.seed(407)
DT <- crime[,-1]
```

```r
results <- data.frame(ncl=2:8, bss.tss = NA)

for(k in 2:8){
  res_kmeans <- kmeans(DT, k)
  results[k - 1, 2] <- res_kmeans$betweenss/res_kmeans$totss
}

plot(results)
```



```r
qplot(ncl, bss.tss, data=results, geom="line")
```



It should increase. As more clusters are added the between cluster SS will be closer and closer to the total SS. Changing the initialization will change the results of the clustering.

Exercice 7

Use the *fpc* package in R, and the function *cluster.stats* to produce the statistic *wb.ratio* to examine the within group distances to the between group distances for each hierarchical cluster solution. How many clusters would be chosen by this approach?

(The *wb.ratio* statistic reports the ratio between two quantities comparing within to between distances. The average of the distances between points that are in the same cluster, ie within. And the distances between points that are not in the same cluster, ie between. The smaller the value of this the better the result describes clustering as explaining the variation in the data.)

```
for(k in 3:9){
  print(cluster.stats(crime.dist, clustering=cutree(crime.hc, k))$wb.ratio)
}
```

```
## [1] 0.5237476
## [1] 0.5596829
## [1] 0.5194983
## [1] 0.5002843
## [1] 0.5039737
## [1] 0.4938994
## [1] 0.449124
```

The result using 3 clusters is better than 4, but 5, 6, 7, 8, 9 get sequentially lower values. 6, 7, 8 are all very similar so probably 5 is best from this group. The k-means with 5 clusters

beats the hierarchical with 5 clusters.

Exercice 8

Decide on an appropriate number of clusters, and report the results. Tabulate the cluster means, standard deviation, and number of points in each cluster. Plot the cluster means using a parallel coordinate plot. List the states in each cluster. Write a paragraph describing the characteristics of each cluster, e.g. cluster 3 is characterized by low larceny and vehicle theft.

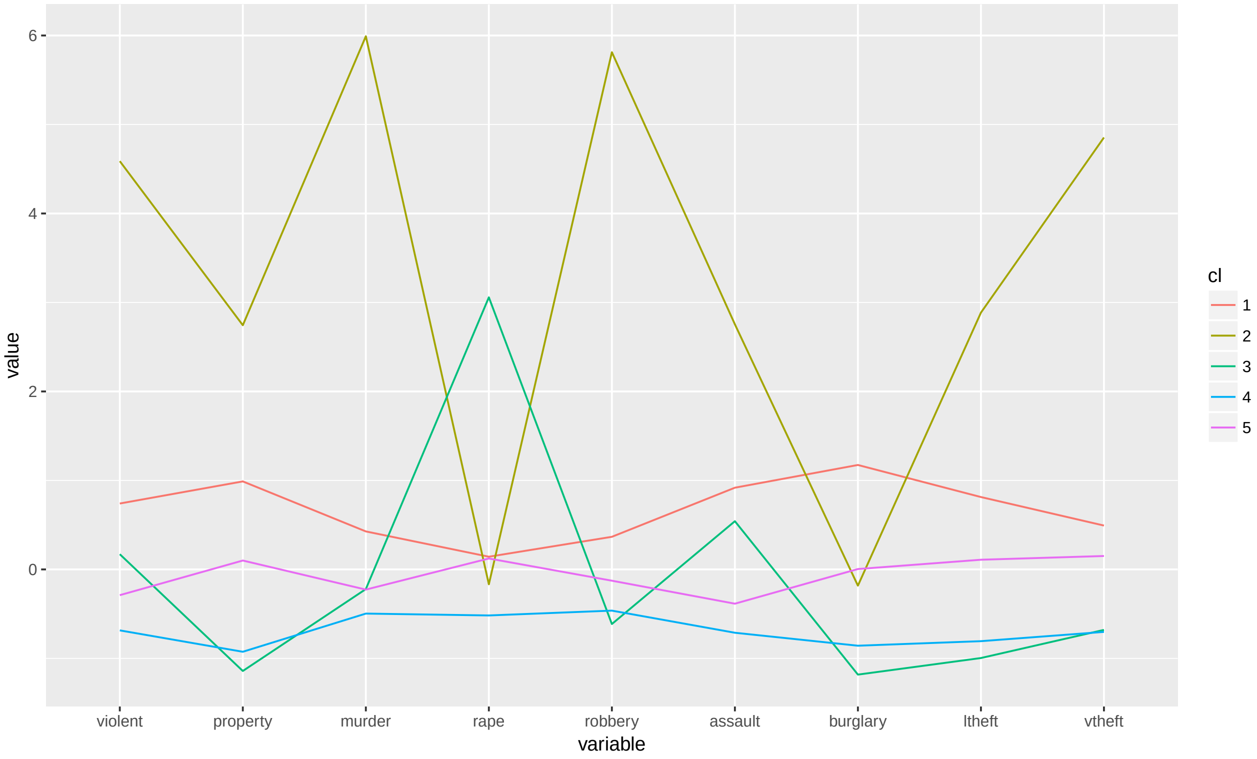
```
res_kmeans <- kmeans(DT, 5)

res_kmeans$centers
```

```
##      violent    property     murder       rape    robbery    assault
## 1  0.7405402  0.98904408  0.4270204  0.1426699  0.3664575  0.9182316
## 2  4.5877209  2.74408978  5.9913967 -0.1671835  5.8122397  2.7542238
## 3  0.1708152 -1.14149865 -0.2210892  3.0577328 -0.6136905  0.5418111
## 4 -0.6851087 -0.92550594 -0.4964364 -0.5172863 -0.4631076 -0.7117160
## 5 -0.2893023  0.09990886 -0.2268656  0.1228799 -0.1268564 -0.3851498
##       burglary     ltheft     vtheft
## 1  1.173729481  0.8132411  0.4929179
## 2 -0.182415612  2.8856250  4.8535616
## 3 -1.181967655 -0.9956927 -0.6795526
## 4 -0.857350457 -0.8064014 -0.7021328
## 5  0.004333686  0.1088870  0.1507216
```

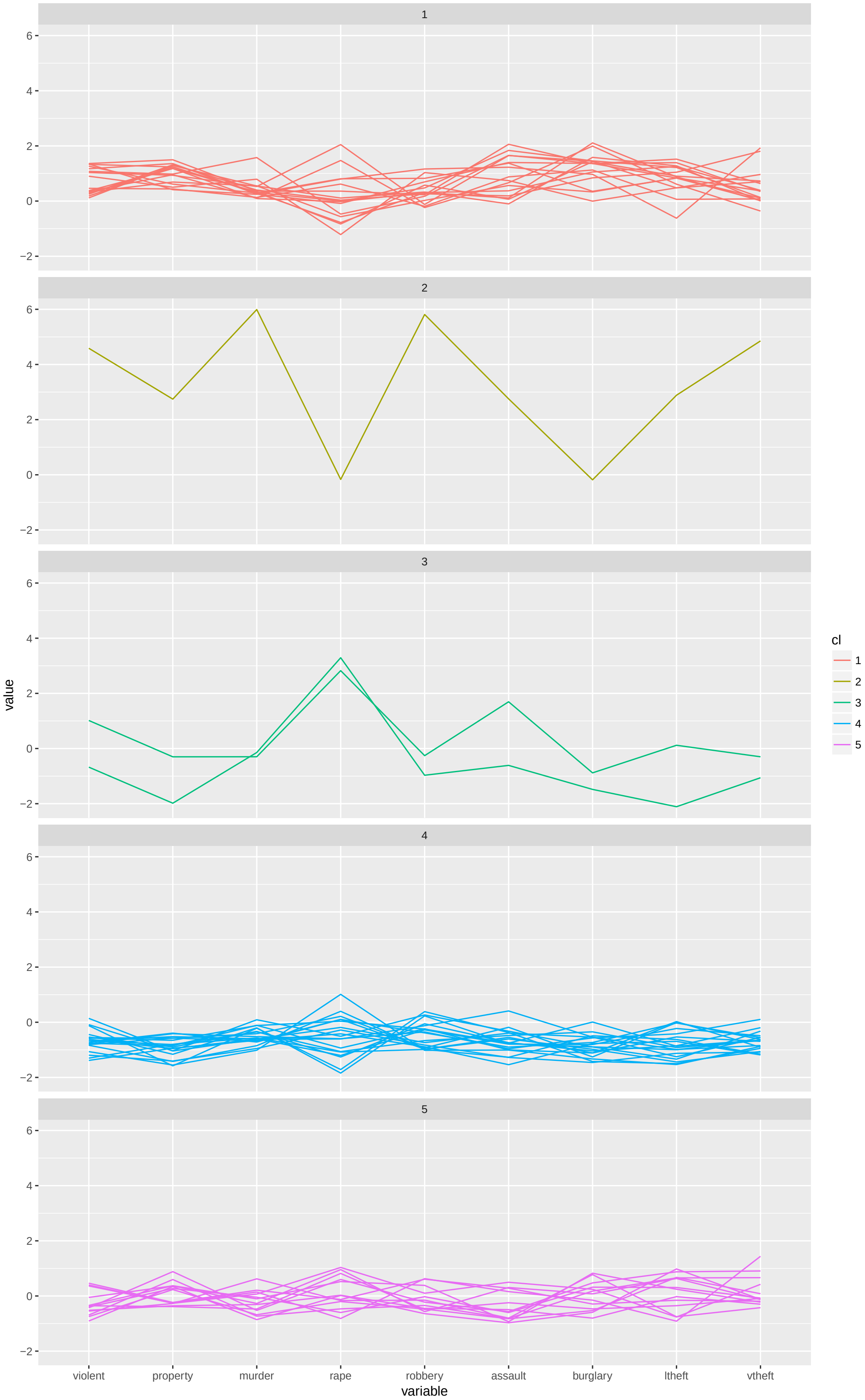
```
crime.km.centers <- ddply(DT, .(res_kmeans$cluster), colMeans)
crime.km.centers
```

```
##   res_kmeans$cluster    violent    property     murder       rape
## 1                  1  0.7405402  0.98904408  0.4270204  0.1426699
## 2                  2  4.5877209  2.74408978  5.9913967 -0.1671835
## 3                  3  0.1708152 -1.14149865 -0.2210892  3.0577328
## 4                  4 -0.6851087 -0.92550594 -0.4964364 -0.5172863
## 5                  5 -0.2893023  0.09990886 -0.2268656  0.1228799
##      robbery    assault     burglary     ltheft     vtheft
## 1  0.3664575  0.9182316  1.173729481  0.8132411  0.4929179
## 2  5.8122397  2.7542238 -0.182415612  2.8856250  4.8535616
## 3 -0.6136905  0.5418111 -1.181967655 -0.9956927 -0.6795526
## 4 -0.4631076 -0.7117160 -0.857350457 -0.8064014 -0.7021328
## 5 -0.1268564 -0.3851498  0.004333686  0.1088870  0.1507216
```

```
colnames(crime.km.centers)[1] <- "cl"
crime.km.centers$cl <- factor(crime.km.centers$cl)
ggparcoord(crime.km.centers, columns=2:10, groupColumn=1, scale="globalminmax")
```



```
crime$cl <- res_kmeans$cluster
crime$cl <- factor(crime$cl)
crime.m <- melt(crime, id.vars = c("State","cl"))
ggplot(crime.m, aes(x=variable, y=value, group=State, colour=cl)) + geom_line() + facet_wrap(~cl, ncol=
```


1
2
3
4
5
6
4
2
0
−2
value
violent
property
murder
rape
robbery
assault
burglary
ltheft
vtheft
variable
cl
1
2
3
4
5

```r
for(i in 1:5){
  print(as.vector(crime[crime$cl == i,1]))
  print("--------")
}
```

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

*** Five clusters is really enough to summarize the cities. If you look at the 7 cluster solution it is hard to characterize all the clusters as different from each other. Clusters 1, 2, 3 are generally consistent across all variables, and are lowest, medium, highest crime, respectively. Cluster 4 is Washingto DC, and it has high crime on all factors except rape and burglary. Cluster 5 (Alaska and South Dakota) is distinguished by having abnormally high rape statistics. The clusters we get reflects, to a large extent that we used overall counts, so large states will appear together, and small states together. If we had first calculated crime per 1000 people, the results would change.***