Cluster Analysis: USArrests dataset

Nathacia Nathacia

2022-11-08

Load packages and data

```
library(cluster)
library(factoextra)
```

```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(car)
```

```
## Loading required package: carData
```

```
library(dbscan)

data(package='factoextra')
data(package='datasets')
data("USArrests")

View(USArrests)
```

Investigate data

```
summary(USArrests)
```

```
##      Murder          Assault         UrbanPop          Rape
##  Min.   : 0.800   Min.   : 45.0   Min.   :32.00   Min.   : 7.30
##  1st Qu.: 4.075   1st Qu.:109.0   1st Qu.:54.50   1st Qu.:15.07
##  Median : 7.250   Median :159.0   Median :66.00   Median :20.10
##  Mean   : 7.788   Mean   :170.8   Mean   :65.54   Mean   :21.23
##  3rd Qu.:11.250   3rd Qu.:249.0   3rd Qu.:77.75   3rd Qu.:26.18
##  Max.   :17.400   Max.   :337.0   Max.   :91.00   Max.   :46.00
```

```
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 ...
```

```
head(USArrests)
```

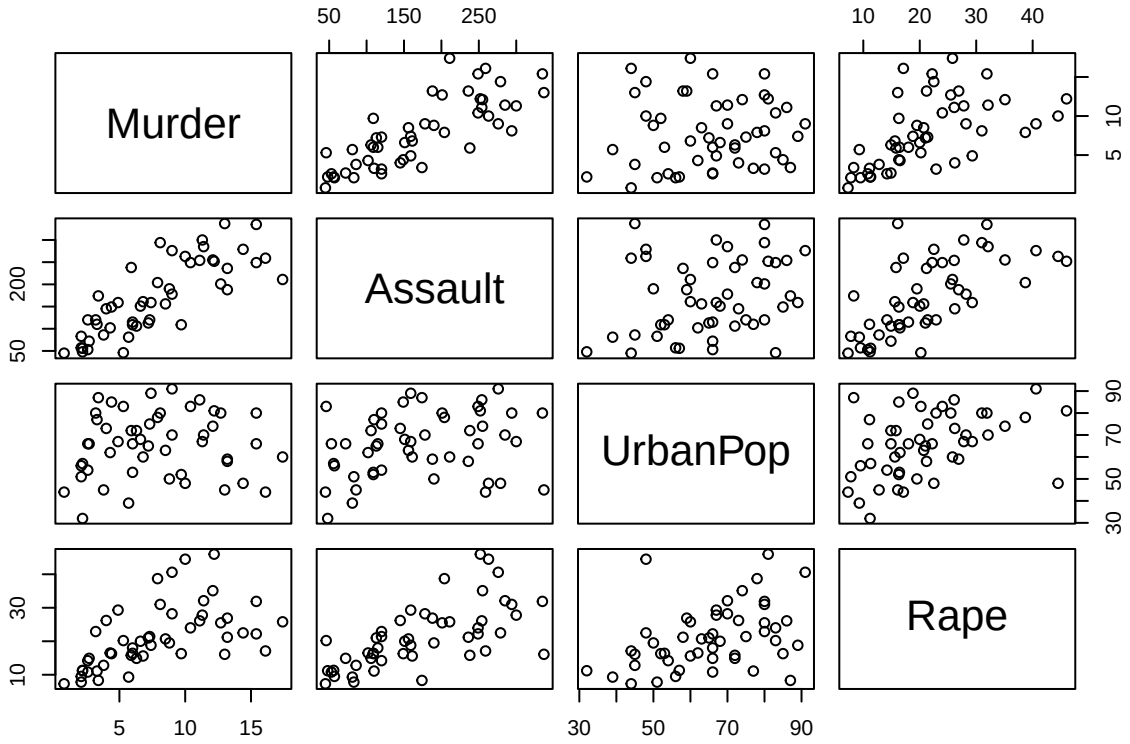
```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
```

```
## 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
```

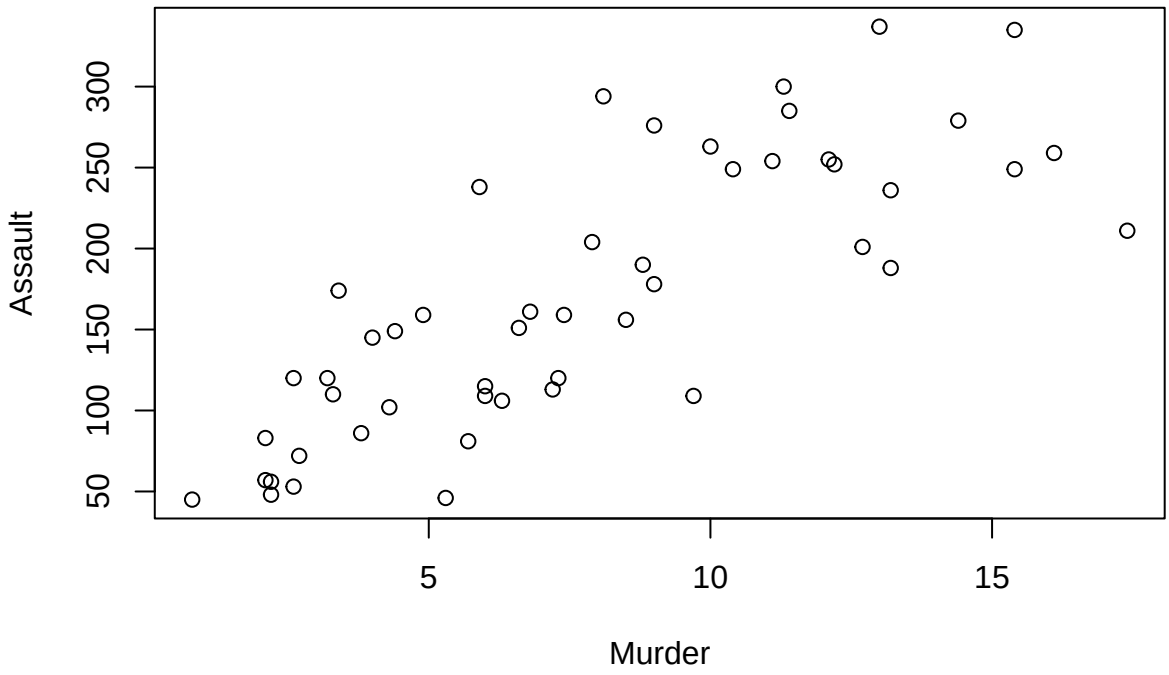
```
tail(USArrests)
```

```
##               Murder Assault UrbanPop Rape
## Vermont          2.2      48       32 11.2
## Virginia         8.5     156       63 20.7
## Washington       4.0     145       73 26.2
## West Virginia    5.7      81       39  9.3
## Wisconsin        2.6      53       66 10.8
## Wyoming          6.8     161       60 15.6
```

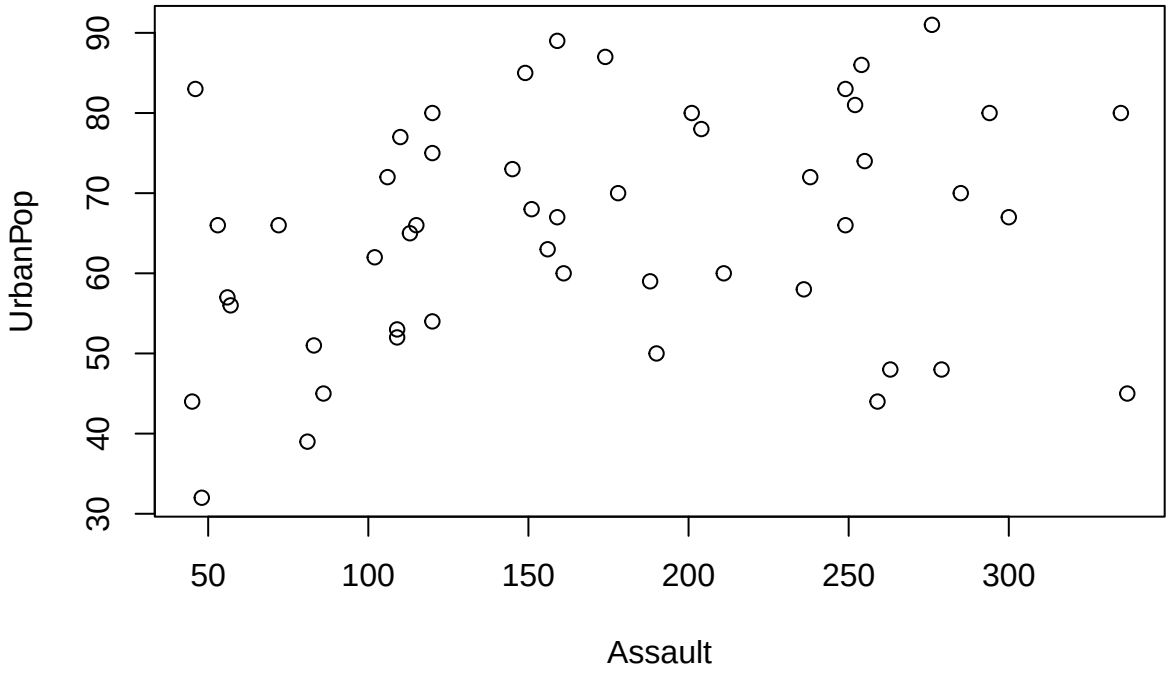
```
pairs(USArrests)
```




```
plot(USArrests[,c(1,2)])
```




```
plot(USArrests[,c(2,3)])
```

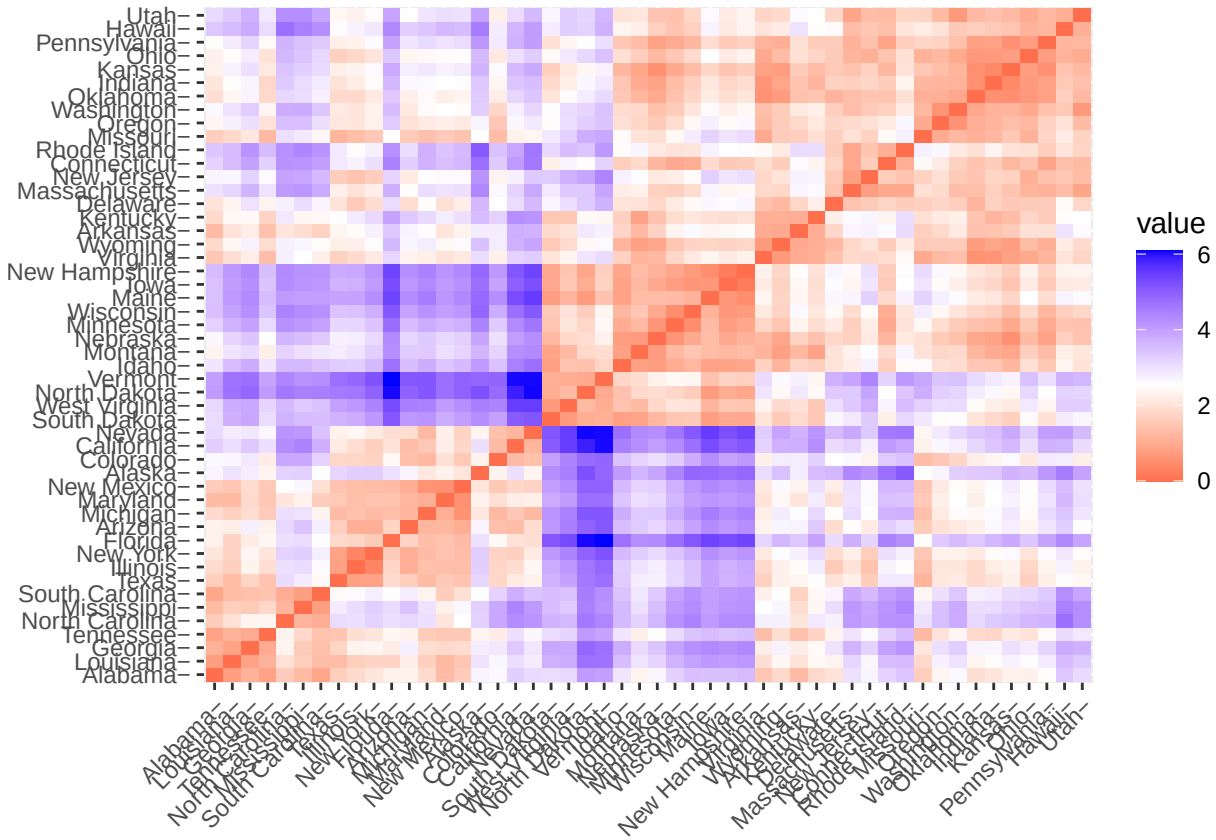


Preprocessing

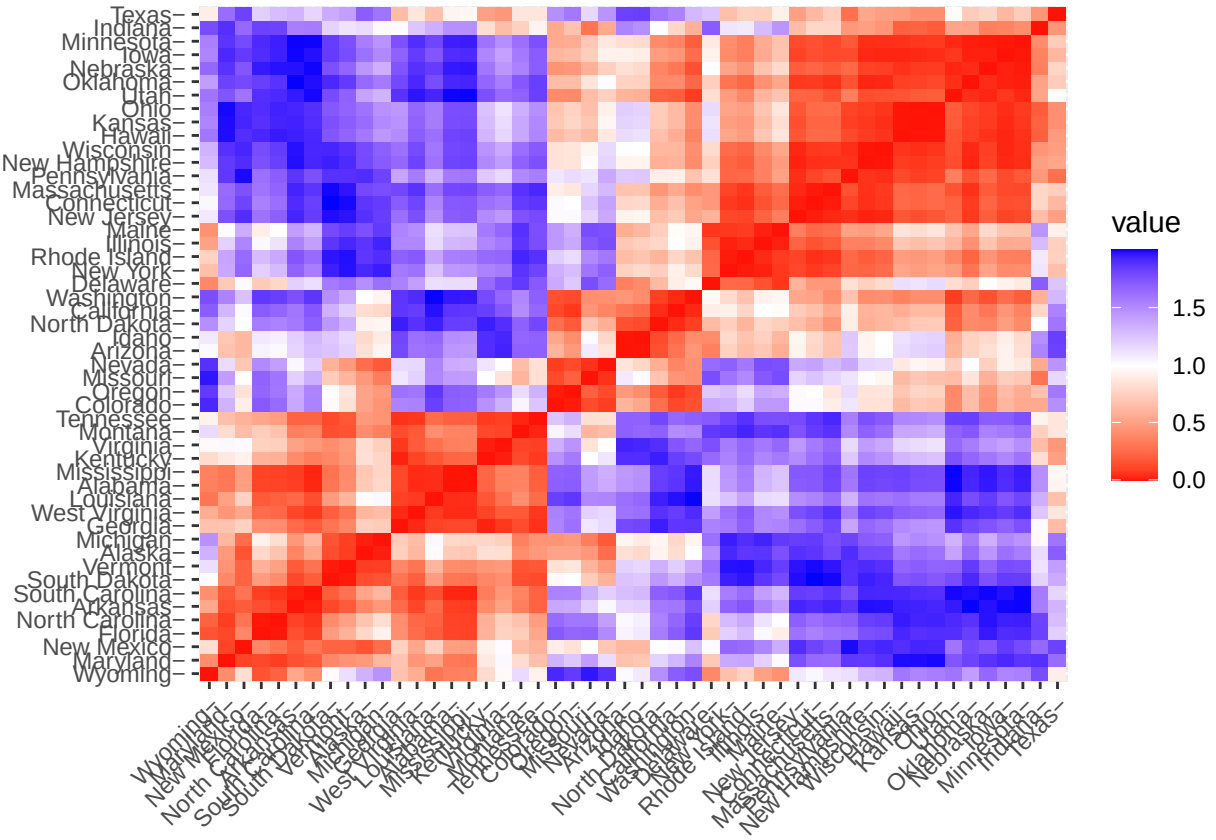
```
d1 <- na.omit(USArrests)
d1 <- scale(USArrests)
View(d1)
```

Distance plot

```
d1dist <- get_dist(d1, method = 'euclidian')
fviz_dist(d1dist)
```




```
d1dist <- get_dist(d1, method = 'pearson')
fviz_dist(d1dist)
```



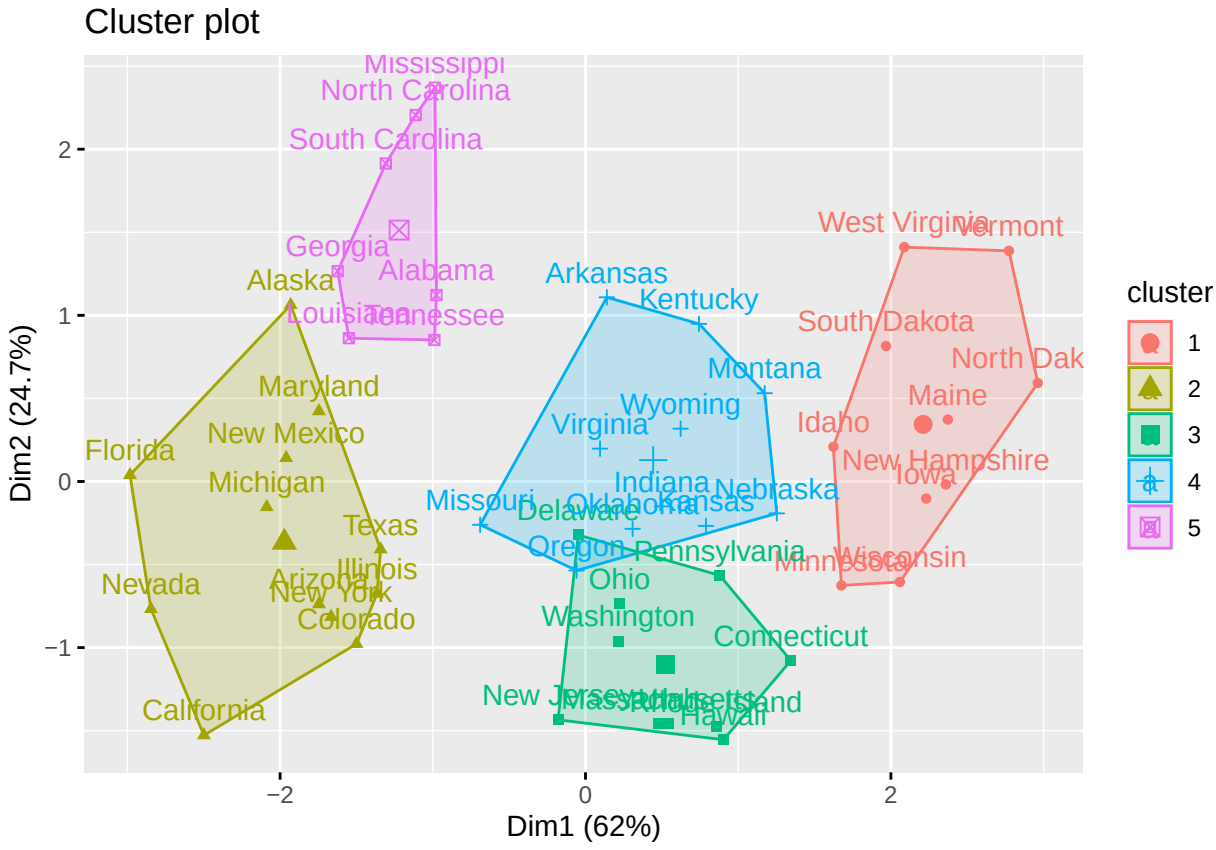
Kmeans clustering

```
set.seed(123)
k1 <- kmeans(d1, centers = 5, nstart = 25)
k1
```

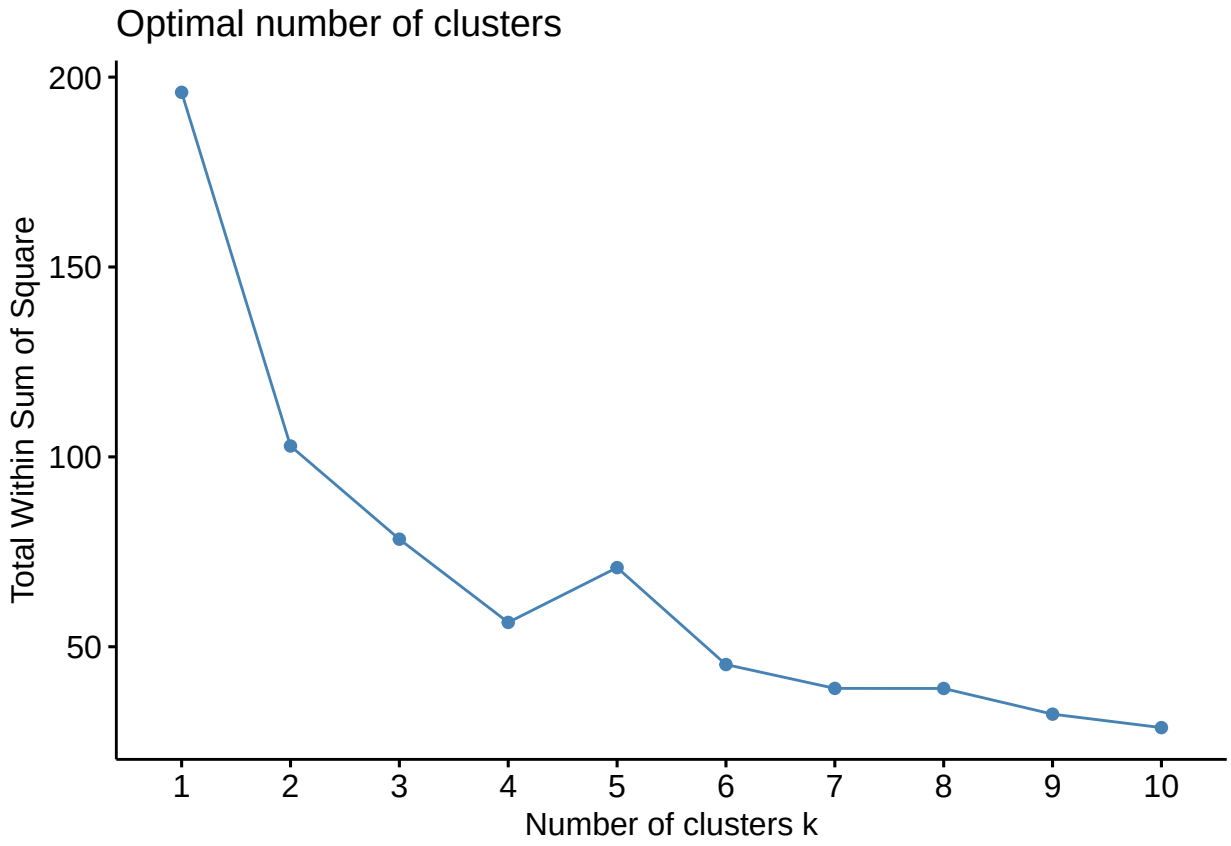
```
## K-means clustering with 5 clusters of sizes 10, 12, 10, 11, 7
##
## Cluster means:
##       Murder    Assault   UrbanPop        Rape
## 1 -1.1727674 -1.2078573 -1.0045069 -1.10202608
## 2  0.7298036  1.1188219  0.7571799  1.32135653
## 3 -0.6286291 -0.4086988  0.9506200 -0.38883734
## 4 -0.1642225 -0.3658283 -0.2822467 -0.11697538
## 5  1.5803956  0.9662584 -0.7775109  0.04844071
##
## Clustering vector:
##        Alabama         Alaska        Arizona       Arkansas     California
##              5              2              2              4              2
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              3              3              2              5
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              3              1              2              4              1
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              4              4              5              1              2
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              3              2              1              5              4
```

```
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              4              4              2              1              3
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              2              2              5              1              3
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              4              4              3              3              5
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              1              5              2              3              1
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              4              3              1              1              4
##
## Within cluster sum of squares by cluster:
## [1]  7.443899 18.257332  9.326266  7.788275  6.128432
##  (between_SS / total_SS =  75.0 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

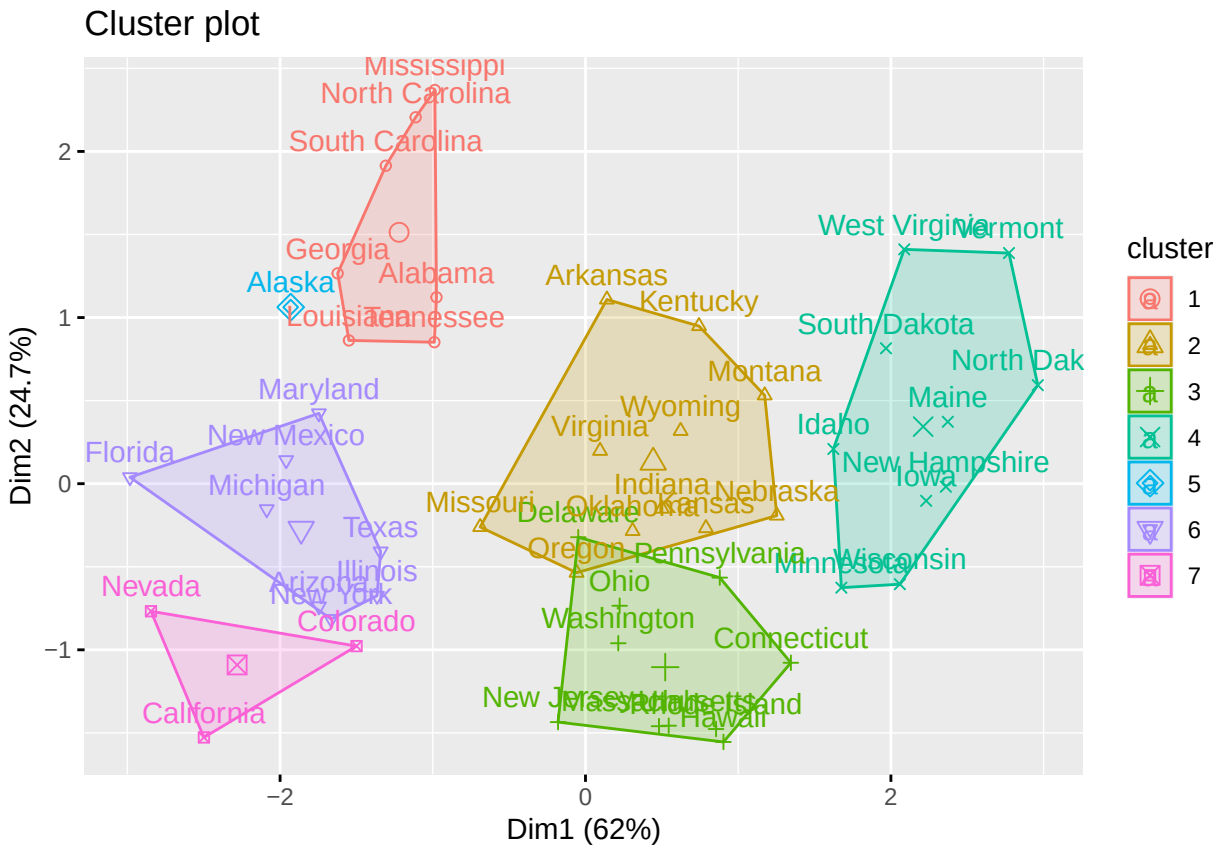
```
fviz_cluster(k1, data = d1)
```




```
fviz_nbclust(d1, kmeans, method = 'wss')
```

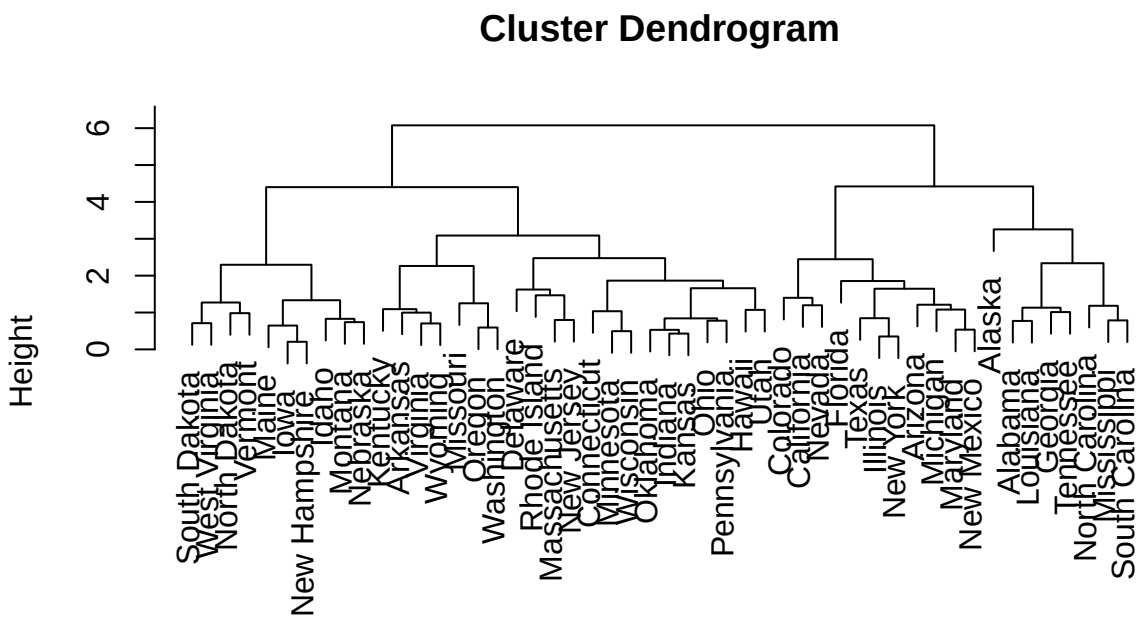



```
k2 <- kmeans(d1, centers = 7, nstart = 25)
fviz_cluster(k2, data = d1)
```

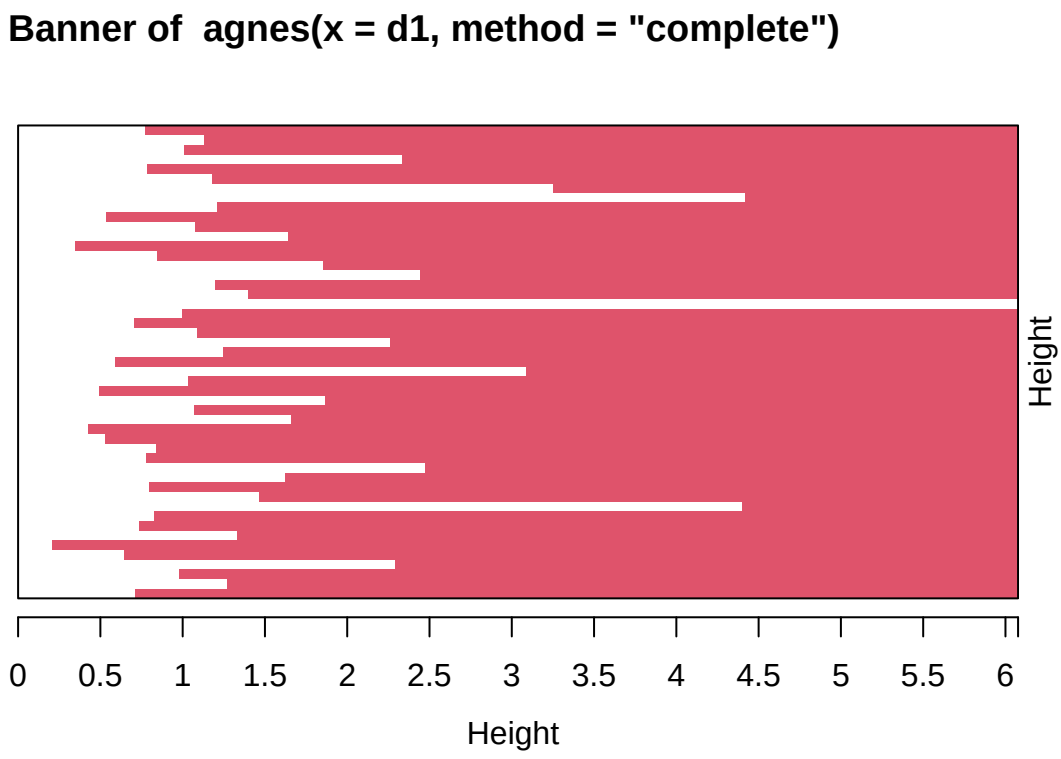


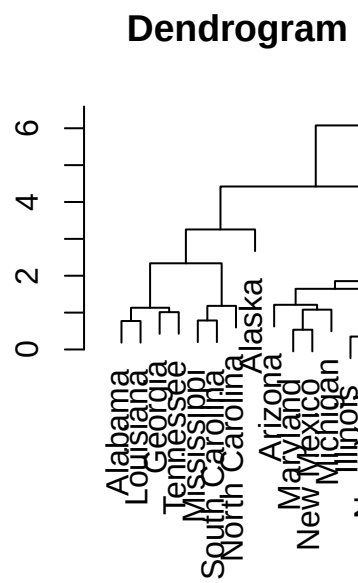
Hierarchical clustering

```
hdist <- get_dist(d1, method = 'euclidian')
h1 <- hclust(hdist, method = 'complete')
plot(h1)
```

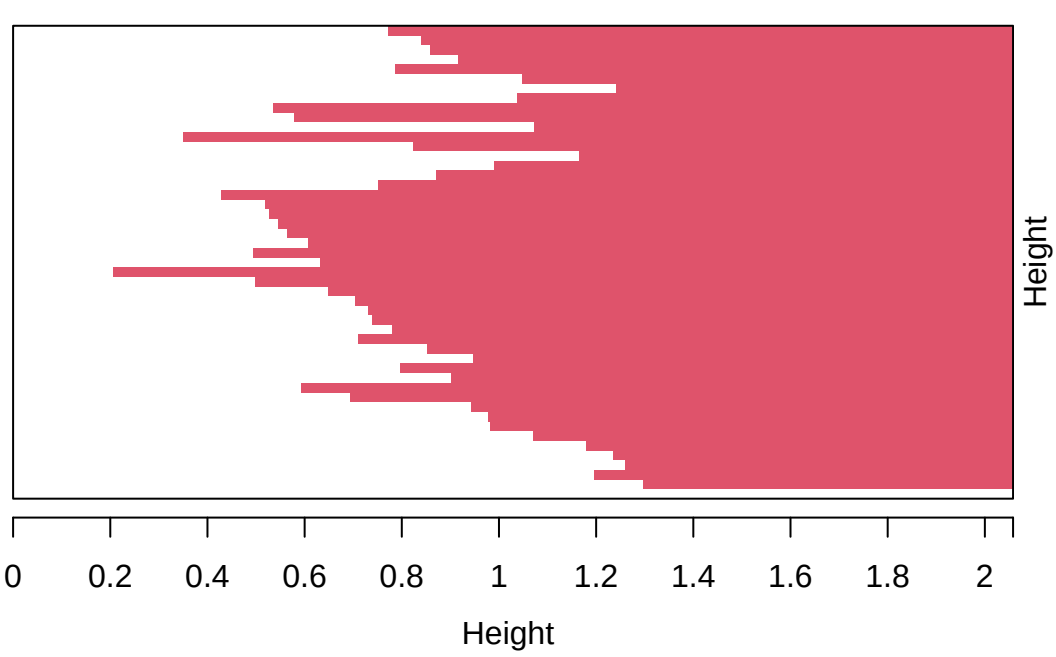


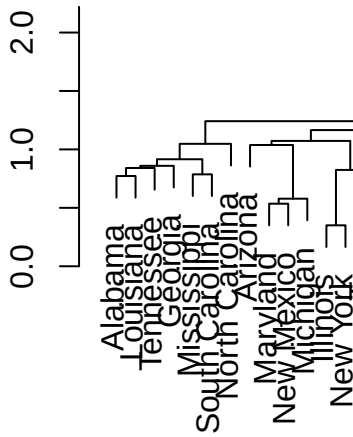

```
h2 <- agnes(d1, method = 'complete')
plot(h2)
```





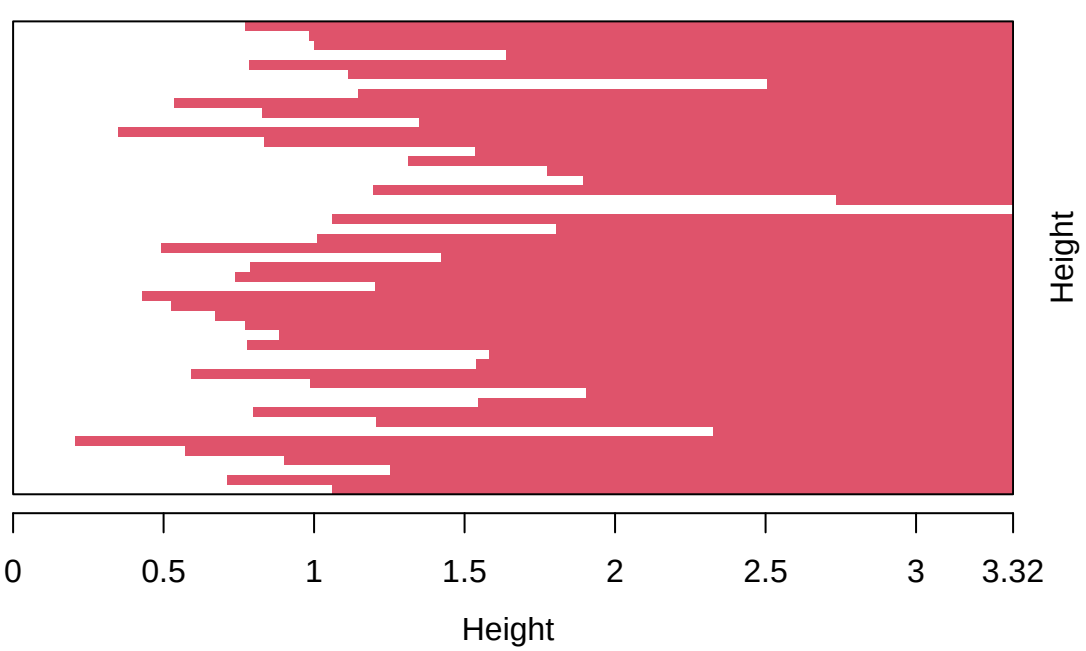
```
h3 <- agnes(d1, method = 'single')
plot(h3)
```

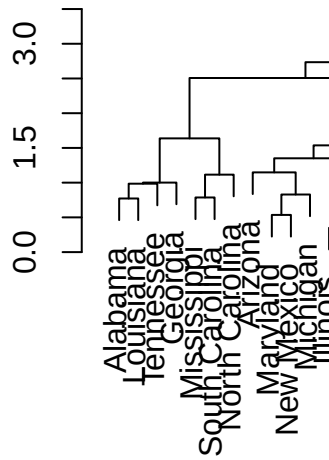




Agglomerative Coefficient = 0.63

```
h4 <- agnes(d1, method = 'average')
plot(h4)
```





Agglomerative Coefficient = 0.74

```
h2$ac
```

```
## [1] 0.8531583
```

```
h3$ac
```

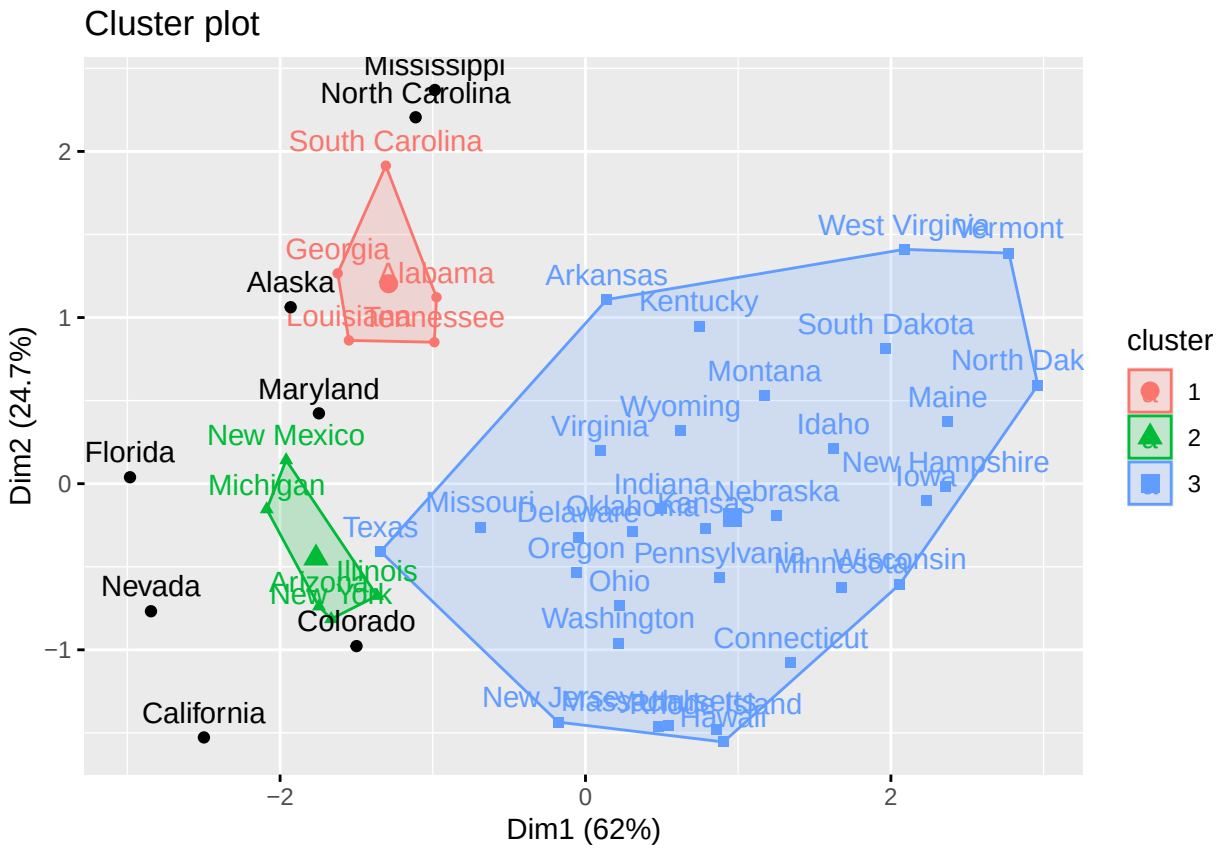
```
## [1] 0.6276128
```

```
h4$ac
```

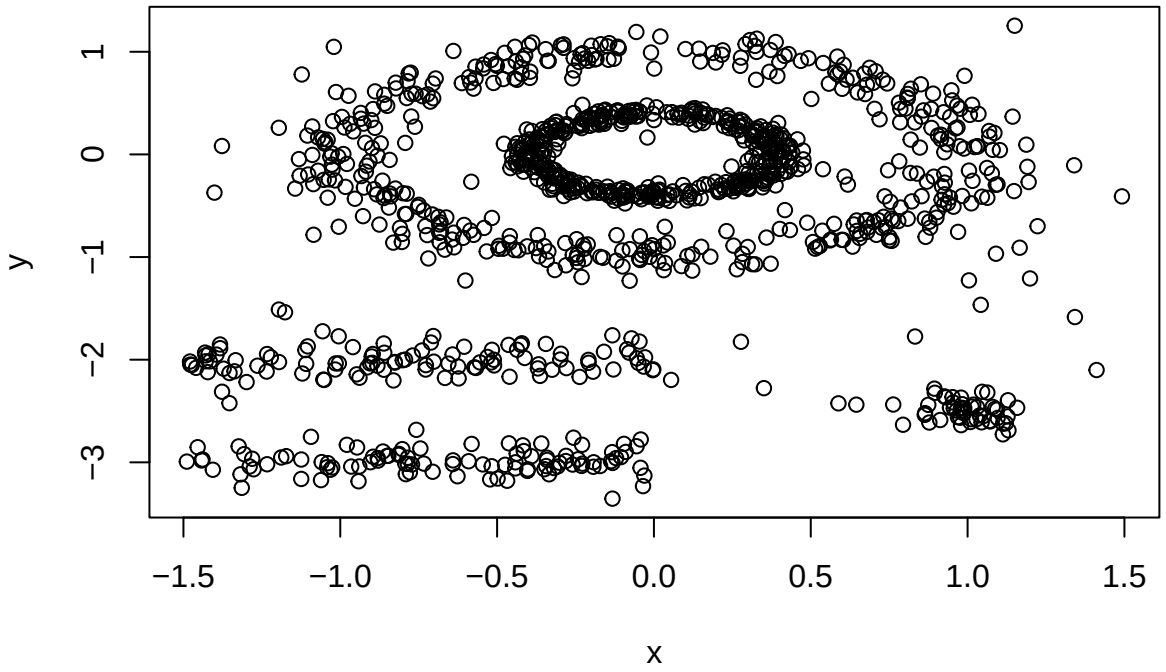
```
## [1] 0.7379371
```

Density based clustering

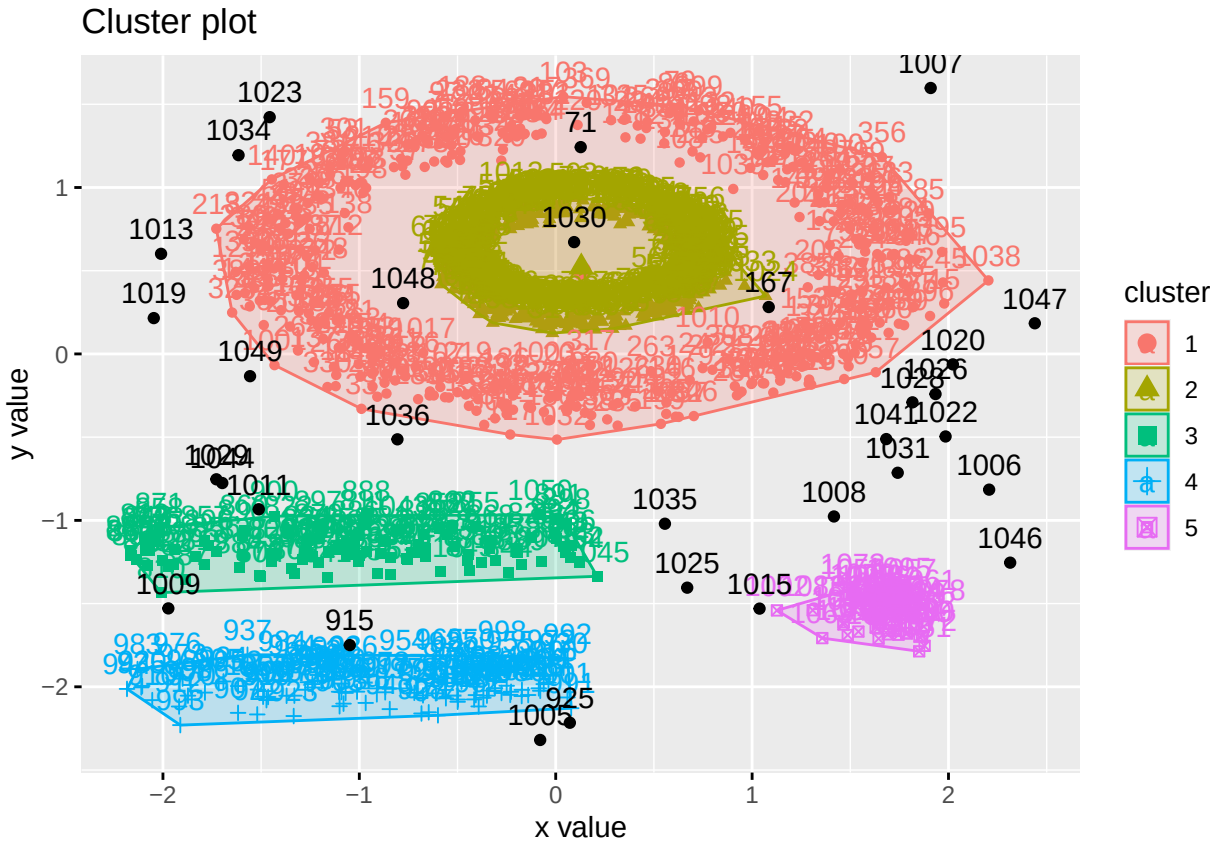
```
db <- dbscan::dbscan(d1, eps = 1.2, minPts = 5)
fviz_cluster(db, data = d1)
```



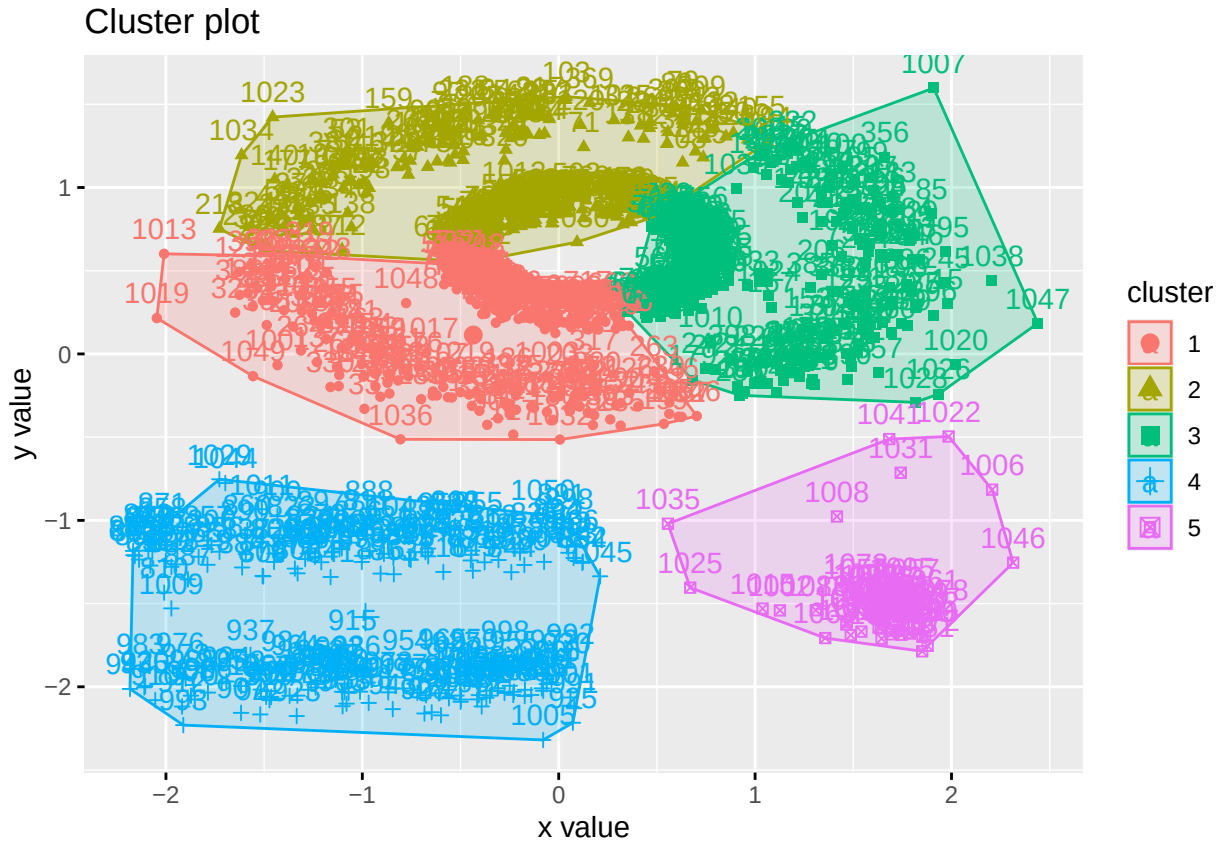

```
#multishapes[,1:2]
plot(multishapes[,1:2])
```




```
db1 <- dbscan(multishapes[,1:2], eps = .15, minPts = 5)
fviz_cluster(db1, data = multishapes[,1:2])
```




```
kdense <- kmeans(multishapes[,1:2], centers = 5, nstart = 25)
fviz_cluster(kdense, data = multishapes[,1:2])
```




```
kNNdistplot(multishapes[,1:2], 5)
abline(h=.15, col='red')
```