Chapter 14 output

```
########Chapter 14, Cluster Analysis##########


#### Example: Birth and death rates
fn.data <- "http://statacumen.com/teach/ADA2/ADA2_notes_Ch14_birthdeath.dat"
bd <- read.table(fn.data, header = TRUE)
str(bd)
```

```
## 'data.frame':    74 obs. of  3 variables:
##  $ country: Factor w/ 74 levels "afghan","algeria",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ birth  : int  52 50 47 22 16 12 47 12 36 17 ...
##  $ death  : int  30 16 23 10 8 13 19 12 10 10 ...
```

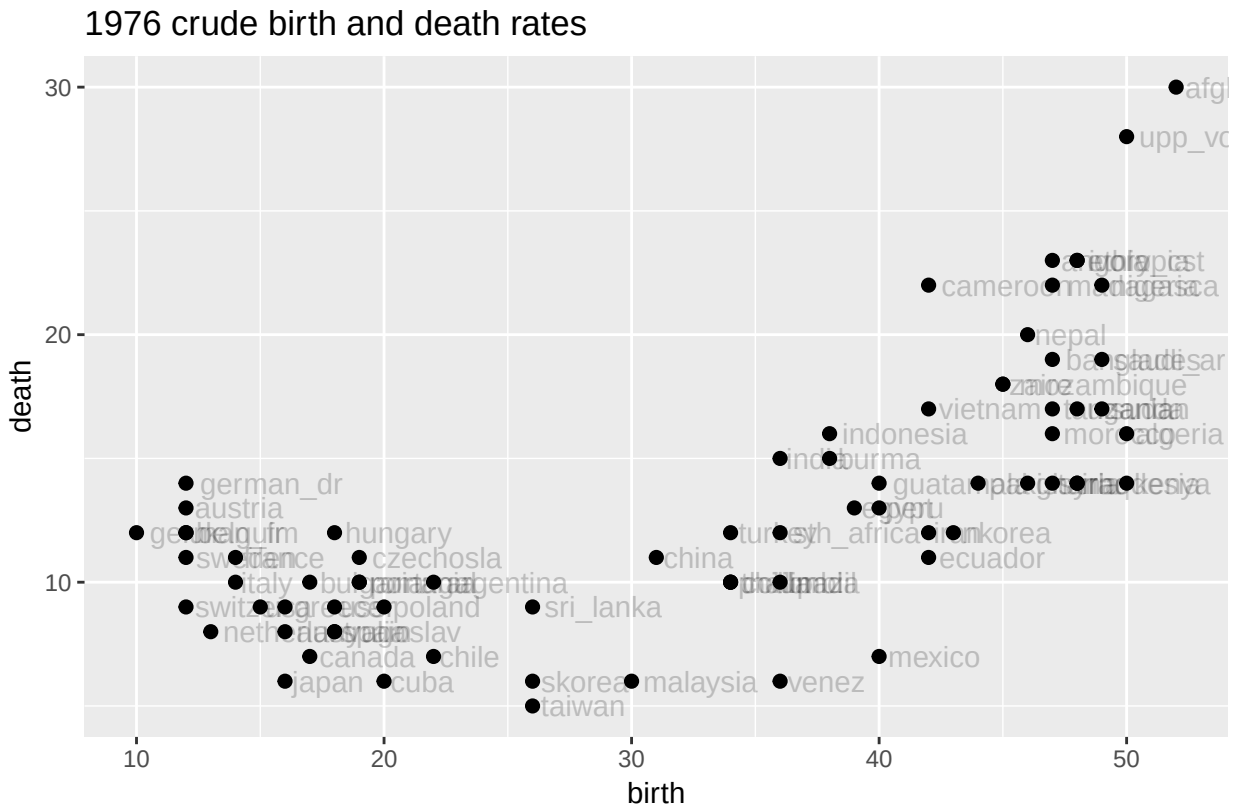
```
nrow(bd)  #74
```

```
## [1] 74
```

```
head(bd)
```

```
##     country birth death
## 1    afghan    52    30
## 2   algeria    50    16
## 3    angola    47    23
## 4 argentina    22    10
## 5 australia    16     8
## 6   austria    12    13
```

```
# plot original data
library(ggplot2)
p1 <- ggplot(bd, aes(x = birth, y = death))
p1 <- p1 + geom_point(size = 2)                                          # points
p1 <- p1 + geom_text(aes(label = country), hjust = -0.1, alpha = 0.2)    # labels
p1 <- p1 + coord_fixed(ratio = 1) # makes 1 unit equal length on x- and y-axis
p1 <- p1 + labs(title = "1976 crude birth and death rates")
print(p1)
```




```
dev.copy(jpeg,filename="~/Desktop/jenn/teaching/ADA2/lecture notes/plots/ch14plot1.jpg")
```

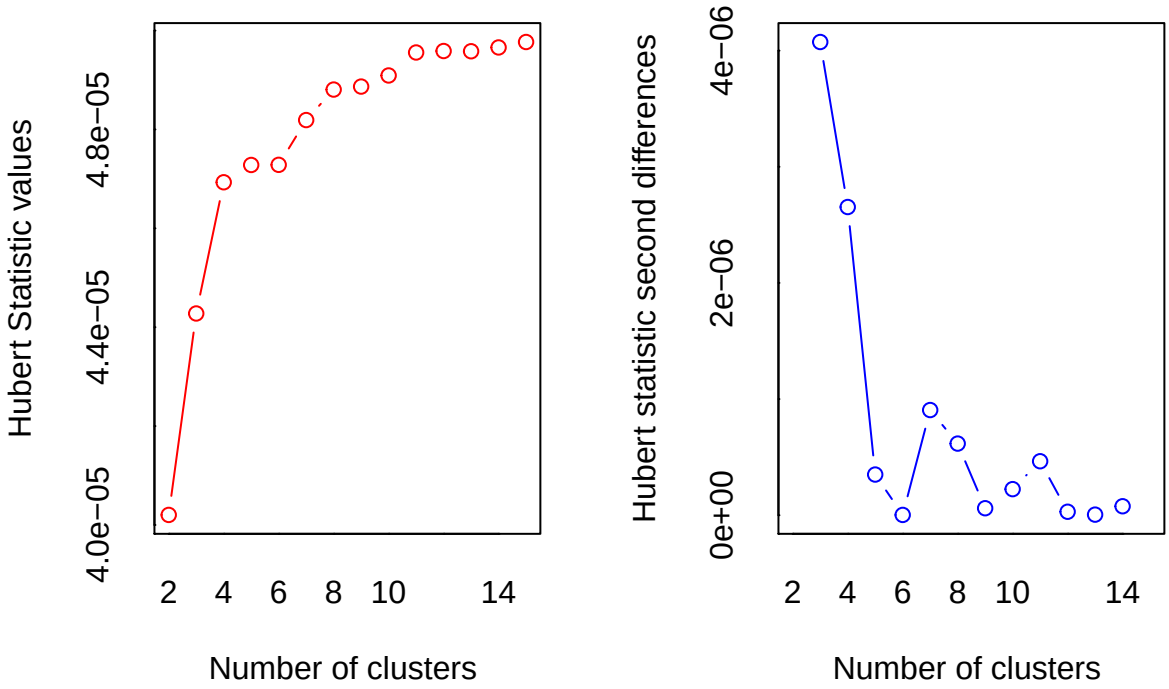
```
## jpeg
##    3
```

```
dev.off()
```

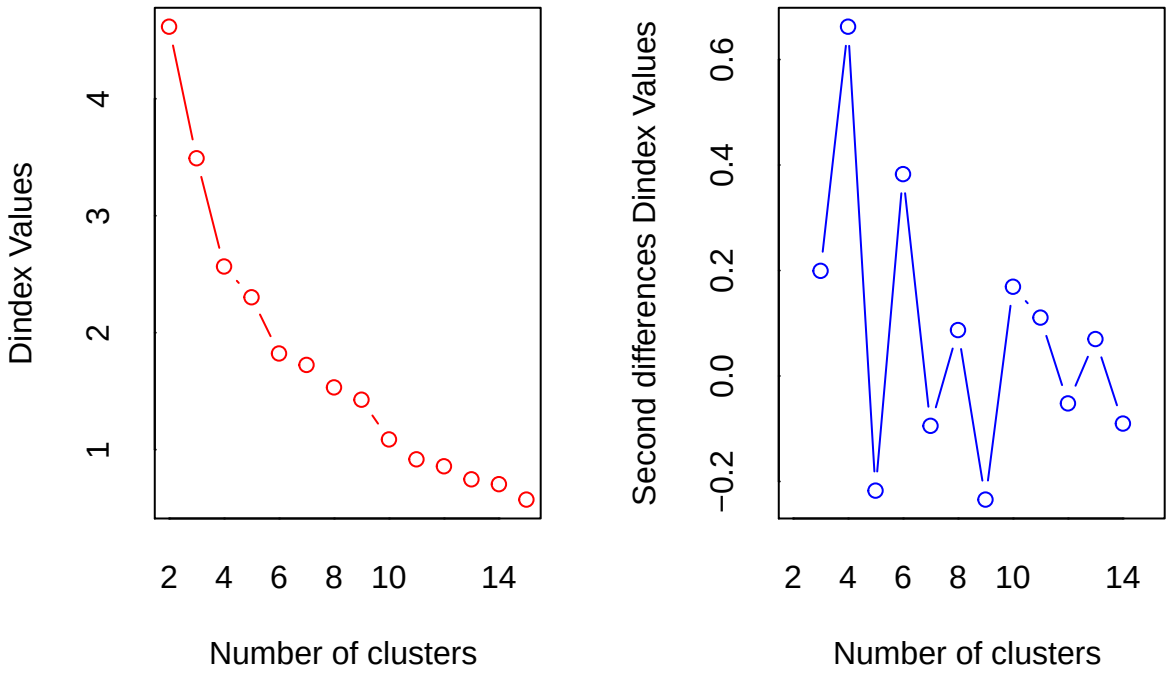
```
## pdf
##   2
```

```
library(NbClust)
# Change integer data type to numeric
bd.num <- as.numeric(as.matrix(bd[,-1]))
NC.out <- NbClust(bd.num, method = "complete", index = "all")
```

```
## Warning in max(DiffLev[, 5], na.rm = TRUE): no non-missing arguments to
## max; returning -Inf
```




```
## *** : The Hubert index is a graphical method of determining the number of clusters.
##                 In the plot of Hubert index, we seek a significant knee that corresponds to a
##                 significant increase of the value of the measure i.e the significant peak in Hubert
##                 index second differences plot.
##

## *** : The D index is a graphical method of determining the number of clusters.
##                 In the plot of D index, we seek a significant knee (the significant peak in Dindex
##                 second differences plot) that corresponds to a significant increase of the value of
##                 the measure.
##

## Warning in matrix(c(results), nrow = 2, ncol = 26): data length [51] is not
## a sub-multiple or multiple of the number of rows [2]

## Warning in matrix(c(results), nrow = 2, ncol = 26, dimnames =
## list(c("Number_clusters", : data length [51] is not a sub-multiple or
## multiple of the number of rows [2]
```




```
## ************************************************************************
## * Among all indices:
## * 2 proposed 2 as the best number of clusters
## * 1 proposed 4 as the best number of clusters
## * 1 proposed 5 as the best number of clusters
## * 1 proposed 6 as the best number of clusters
## * 1 proposed 15 as the best number of clusters
##
##                     ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  2
##
##
## ************************************************************************
```

```
dev.copy(jpeg,filename="~/Desktop/jenn/teaching/ADA2/lecture notes/plots/ch14plot2.jpg")
```

```
## jpeg
##    3
```

```
dev.off()
```

```
## pdf
##   2
```

```
# most of the methods suggest 2 to 6 clusters, as do the plots
NC.out$Best.nc
```

```
##                     KL       CH Hartigan     CCC   Scott  Marriot TrCovW
## Number_clusters 2.000   15.000   5.0000  2.0000  4.0000    6.000   -Inf
## Value_Index     3.333 1780.714 209.2456 20.7606 86.7855 9041.261      4
##                   TraceW Friedman     Rubin Cindex     DB Silhouette
## Number_clusters 854.6162  395.428 -131.4723 0.2254 0.4292     0.7468
## Value_Index      15.0000   13.000    2.0000 2.0000 2.0000     2.0000
##                   Duda PseudoT2  Beale Ratkowsky     Ball PtBiserial
```

```
## Number_clusters 0.2486 142.0413 0.9864    0.4628 5166.333     0.8512
## Value_Index     2.0000   2.0000 3.0000    3.0000    2.000     3.0000
##                   Frey McClain   Dunn Hubert SDindex Dindex   SDbw
## Number_clusters 3.5386  0.1705 0.3333      0  0.3167      0 0.0073
## Value_Index     2.0000 13.0000 0.0000      3  0.0000     15 2.0000
```

```
# create distance matrix between points
bd.dist <- dist(bd[,-1])

# number of clusters to identify with red boxes and ellipses
i.clus <- 3


# create dendrogram
par(mfrow=c(1,1))
bd.hc.complete  <- hclust(bd.dist, method = "complete")
plclust(bd.hc.complete, hang = -1
        , main = paste("Birth and death with complete linkage and", i.clus, "clusters")
        , labels = bd[,1])
```

```
## Warning: 'plclust' is deprecated.
## Use 'plot' instead.
## See help("Deprecated")
```

```
dev.copy(jpeg,filename="~/Desktop/jenn/teaching/ADA2/lecture notes/plots/ch14plot3_1.jpg")
```

```
## jpeg
##    3
```

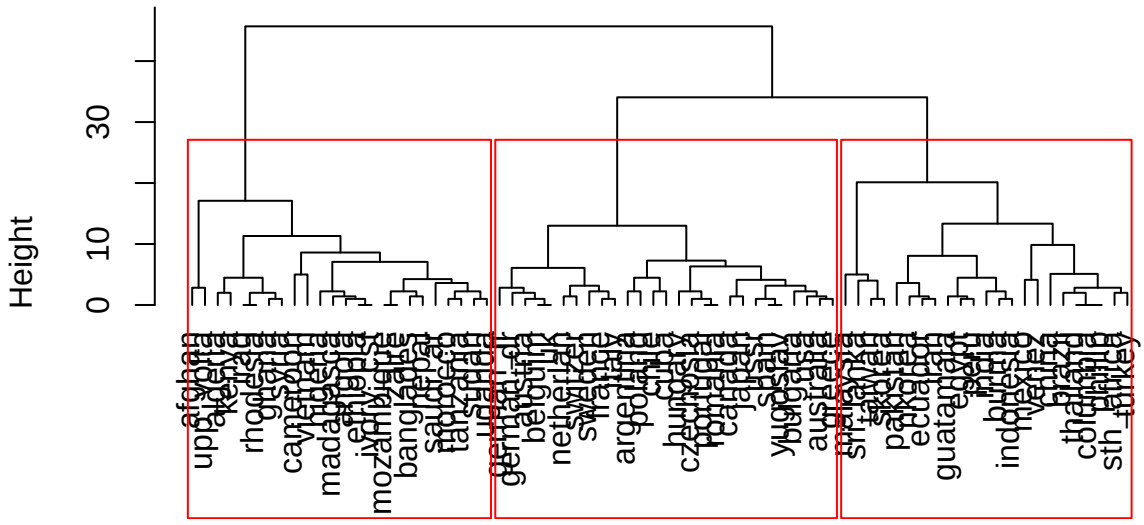
```
dev.off()
```

```
## pdf
##   2
```

```
# create dendrogram with the three clusters
bd.hc.complete  <- hclust(bd.dist, method = "complete")
plclust(bd.hc.complete, hang = -1
      , main = paste("Birth and death with complete linkage and", i.clus, "clusters")
      , labels = bd[,1])
```

```
## Warning: 'plclust' is deprecated.
## Use 'plot' instead.
## See help("Deprecated")
```

```
rect.hclust(bd.hc.complete, k = i.clus)
```



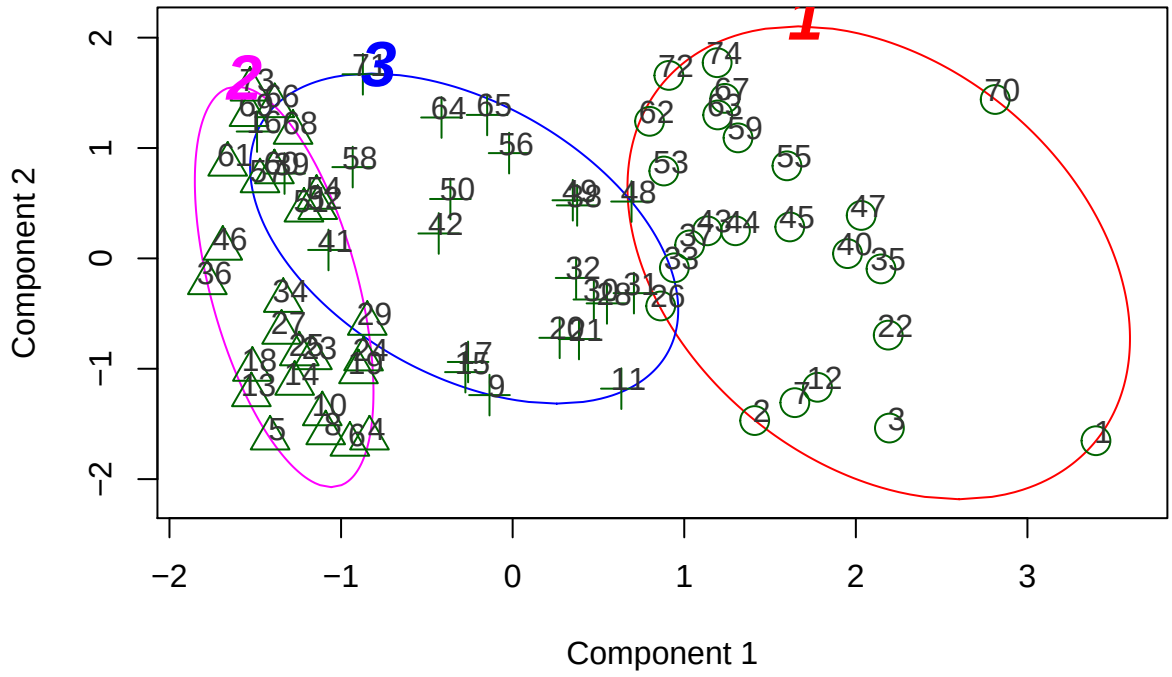

```
dev.copy(jpeg,filename="~/Desktop/jenn/teaching/ADA2/lecture notes/plots/ch14plot3.jpg")
```

```
## jpeg
##   3
```

```
dev.off()
```

```
## pdf
##   2
```

```
# create PCA scores plot with ellipses
library(cluster)
clusplot(bd, cutree(bd.hc.complete, k = i.clus)
        , color = TRUE, labels = 2, lines = 0
        , cex = 2, cex.txt = 1, col.txt = "gray20"
        , main = paste("Birth/Death PCA with complete linkage and", i.clus, "clusters"), sub = NULL)
```




```
dev.copy(jpeg,filename="~/Desktop/jenn/teaching/ADA2/lecture notes/plots/ch14plot4.jpg")
```

```
## jpeg
##   3
```

```
dev.off()
```

```
## pdf
##   2
```

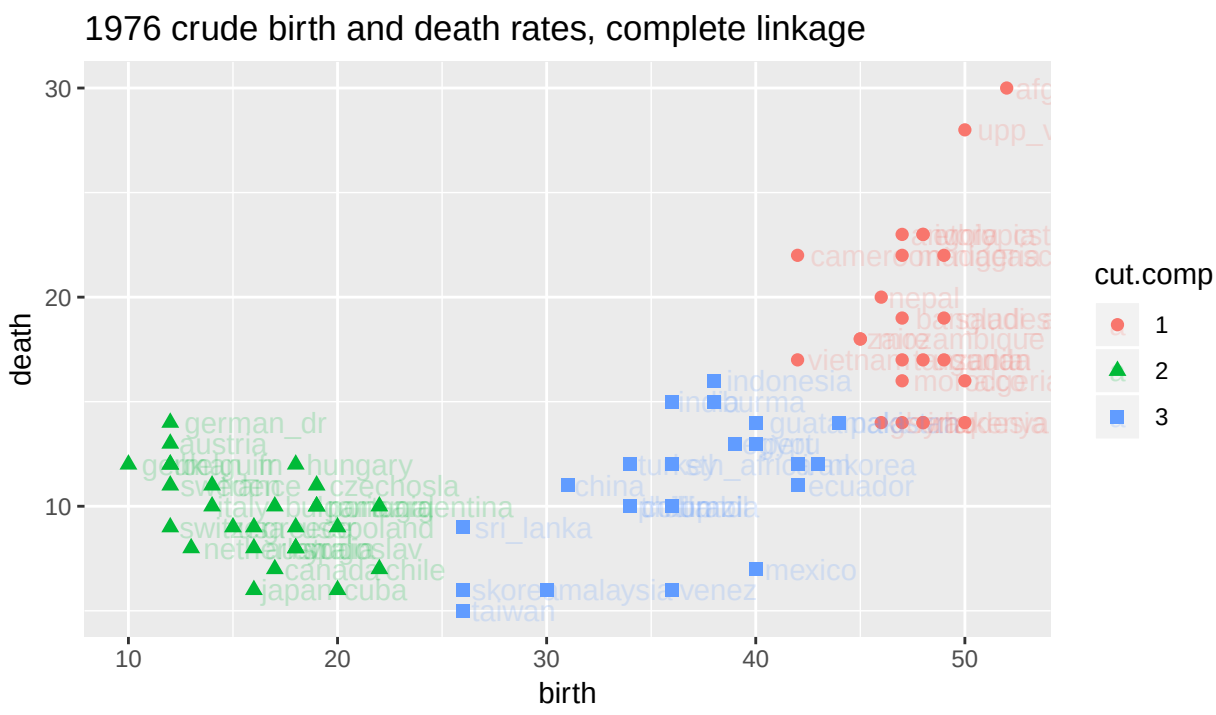
```
# create a column with group membership
bd$cut.comp <- factor(cutree(bd.hc.complete, k = i.clus))



# print the observations in each cluster
for (i.cut in 1:i.clus) {
  print(paste("Cluster", i.cut, " ------------------------------ "))
  print(bd[(cutree(bd.hc.complete, k = i.clus) == i.cut),])
}
```

```
## [1] "Cluster 1  ------------------------------ "
##       country birth death cut.comp
## 1      afghan    52    30        1
## 2     algeria    50    16        1
## 3      angola    47    23        1
## 7   banglades    47    19        1
## 12   cameroon    42    22        1
## 22   ethiopia    48    23        1
## 26      ghana    46    14        1
## 33       iraq    48    14        1
```

```
## 35  ivory_cst    48    23        1
## 37      kenya    50    14        1
## 40  madagasca    47    22        1
## 43    morocco    47    16        1
## 44 mozambique    45    18        1
## 45      nepal    46    20        1
## 47    nigeria    49    22        1
## 53   rhodesia    48    14        1
## 55   saudi_ar    49    19        1
## 59      sudan    49    17        1
## 62      syria    47    14        1
## 63   tanzania    47    17        1
## 67     uganda    48    17        1
## 70  upp_volta    50    28        1
## 72    vietnam    42    17        1
## 74      zaire    45    18        1
## [1] "Cluster 2  ----------------------------- "
##      country birth death cut.comp
## 4  argentina    22    10        2
## 5  australia    16     8        2
## 6    austria    12    13        2
## 8    belguim    12    12        2
## 10  bulgaria    17    10        2
## 13    canada    17     7        2
## 14     chile    22     7        2
## 18      cuba    20     6        2
## 19 czechosla    19    11        2
## 23    france    14    11        2
## 24 german_dr    12    14        2
## 25 german_fr    10    12        2
## 27    greece    16     9        2
## 29   hungary    18    12        2
## 34     italy    14    10        2
## 36     japan    16     6        2
## 46 netherlan    13     8        2
## 51    poland    20     9        2
## 52  portugal    19    10        2
## 54   romania    19    10        2
## 57     spain    18     8        2
## 60    sweden    12    11        2
## 61   switzer    12     9        2
## 66      ussr    18     9        2
## 68        uk    12    12        2
## 69       usa    15     9        2
## 73  yugoslav    18     8        2
## [1] "Cluster 3  ----------------------------- "
##       country birth death cut.comp
## 9      brazil    36    10        3
## 11      burma    38    15        3
## 15      china    31    11        3
## 16     taiwan    26     5        3
## 17   columbia    34    10        3
## 20    ecuador    42    11        3
## 21      egypt    39    13        3
```

```
## 28  guatamala    40    14        3
## 30      india    36    15        3
## 31  indonesia    38    16        3
## 32       iran    42    12        3
## 38     nkorea    43    12        3
## 39     skorea    26     6        3
## 41   malaysia    30     6        3
## 42     mexico    40     7        3
## 48   pakistan    44    14        3
## 49       peru    40    13        3
## 50    phillip    34    10        3
## 56 sth_africa    36    12        3
## 58  sri_lanka    26     9        3
## 64   thailand    34    10        3
## 65     turkey    34    12        3
## 71      venez    36     6        3
```

```
# plot original data with cluster information
library(ggplot2)
p1 <- ggplot(bd, aes(x = birth, y = death, colour = cut.comp, shape = cut.comp))
p1 <- p1 + geom_point(size = 2)                                          # points
p1 <- p1 + geom_text(aes(label = country), hjust = -0.1, alpha = 0.2)  # labels
p1 <- p1 + coord_fixed(ratio = 1) # makes 1 unit equal length on x- and y-axis
p1 <- p1 + labs(title = "1976 crude birth and death rates, complete linkage")
print(p1)
```




```
dev.copy(jpeg,filename="~/Desktop/jenn/teaching/ADA2/lecture notes/plots/ch14plot5.jpg")
```

```
## jpeg
##    3
```

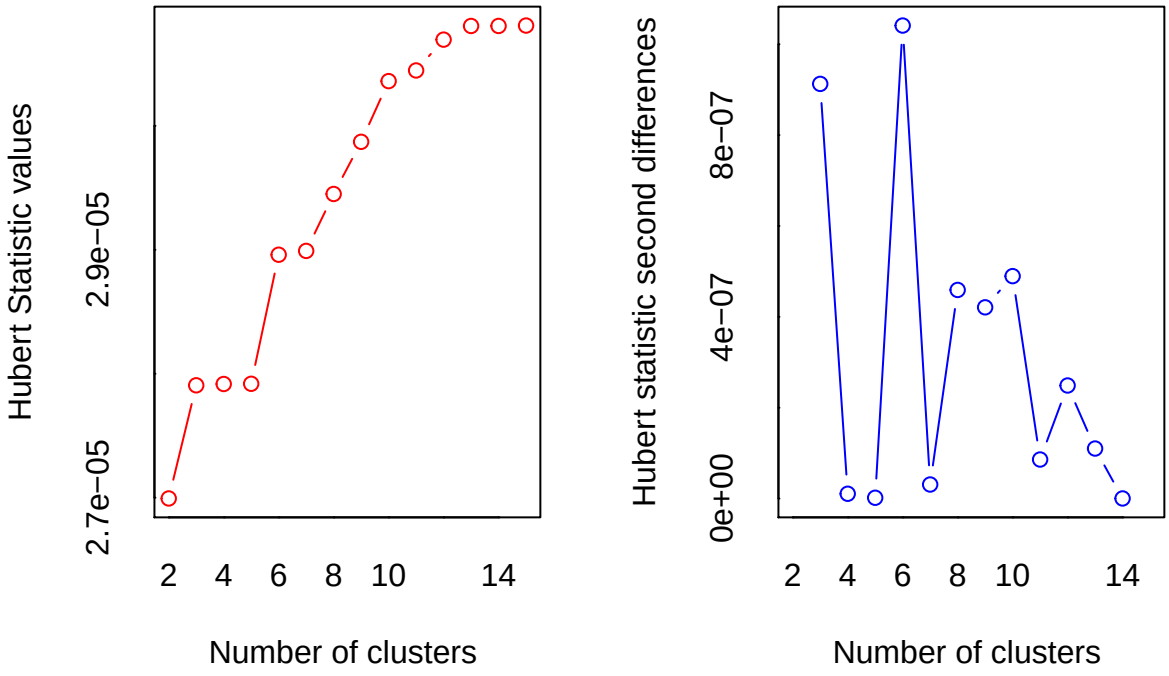
```
dev.off()
```

```
## pdf
```

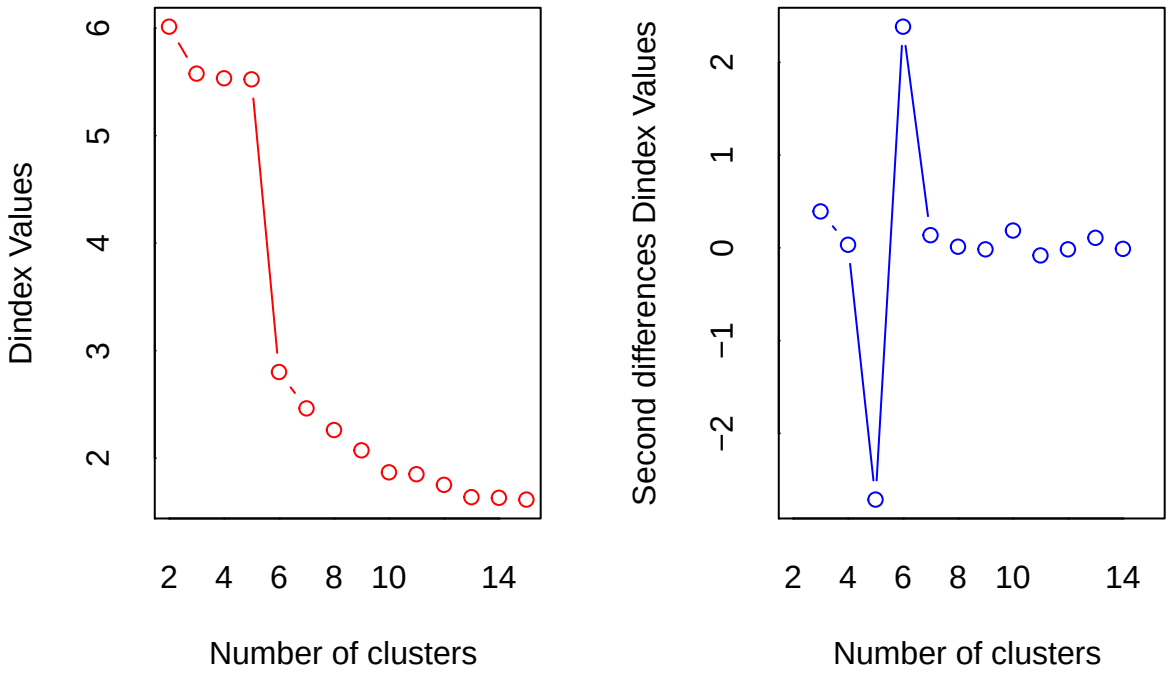
```
##   2
```

```
#################use method ="single", you may also try method="average"####################

library(NbClust)
# Change integer data type to numeric
bd.num <- as.numeric(as.matrix(bd[,-1]))
NC.out <- NbClust(bd.num, method = "single", index = "all")
```

```
## Warning in max(DiffLev[, 5], na.rm = TRUE): no non-missing arguments to
## max; returning -Inf
```




```
## *** : The Hubert index is a graphical method of determining the number of clusters.
##                 In the plot of Hubert index, we seek a significant knee that corresponds to a
##                 significant increase of the value of the measure i.e the significant peak in Hubert
##                 index second differences plot.
##

## *** : The D index is a graphical method of determining the number of clusters.
##                 In the plot of D index, we seek a significant knee (the significant peak in Dindex
##                 second differences plot) that corresponds to a significant increase of the value of
##                 the measure.
##

## Warning in matrix(c(results), nrow = 2, ncol = 26): data length [51] is not
## a sub-multiple or multiple of the number of rows [2]

## Warning in matrix(c(results), nrow = 2, ncol = 26): data length [51] is not
## a sub-multiple or multiple of the number of rows [2]
```




```
## *****************************************************************
## * Among all indices:
## * 1 proposed 2 as the best number of clusters
## * 1 proposed 5 as the best number of clusters
## * 2 proposed 6 as the best number of clusters
## * 1 proposed 7 as the best number of clusters
## * 1 proposed 11 as the best number of clusters
##
##                     ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  6
##
##
## *****************************************************************
```

```
# most of the methods suggest 4 to 11 clusters, as do the plots
NC.out$Best.nc
```

```
##                     KL       CH Hartigan     CCC    Scott  Marriot TrCovW
## Number_clusters 7.0000  11.0000   5.0000  2.0000   6.0000      6.0   -Inf
## Value_Index     8.5944 342.3491 473.8986 13.7816 221.8442 115129.2      6
##                   TraceW Friedman    Rubin Cindex    DB Silhouette   Duda
## Number_clusters 4990.876  20.3754 -12.0601  0.189 0.341     0.7364 0.4667
## Value_Index        6.000   6.0000   7.0000 15.000 2.000     2.0000 2.0000
##                  PseudoT2 Beale Ratkowsky     Ball PtBiserial   Frey
## Number_clusters  53.7092 0.373    0.3521 9161.833     0.8462 4.0846
## Value_Index       2.0000 7.000    3.0000    2.000     2.0000 2.0000
##                  McClain   Dunn Hubert SDindex Dindex   SDbw
## Number_clusters  0.1235 0.1364      0  0.5134      0 0.0442
## Value_Index      3.0000 0.0000      3  0.0000     15 7.0000
```

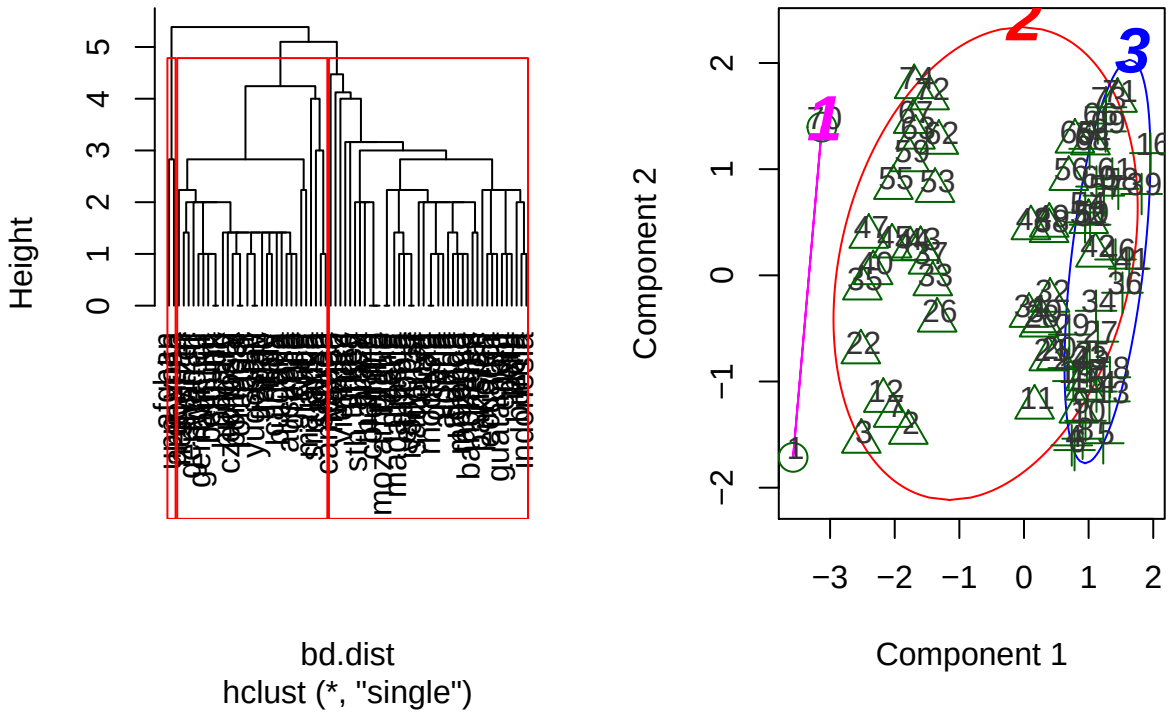
```
# create distance matrix between points
bd.dist <- dist(bd[,-1])
```

```
# number of clusters to identify with red boxes and ellipses
i.clus <- 3

# create dendrogram
bd.hc.single  <- hclust(bd.dist, method = "single")
plclust(bd.hc.single, hang = -1
      , main = paste("Birth and death with single linkage and", i.clus, "clusters")
      , labels = bd[,1])
```

```
## Warning: 'plclust' is deprecated.
## Use 'plot' instead.
## See help("Deprecated")
```

```
rect.hclust(bd.hc.single, k = i.clus)

# create PCA scores plot with ellipses
clusplot(bd, cutree(bd.hc.single, k = i.clus)
        , color = TRUE, labels = 2, lines = 0
        , cex = 2, cex.txt = 1, col.txt = "gray20"
        , main = paste("Birth/Death PCA with single linkage and", i.clus, "clusters")
        , sub = NULL)
```

and death with single linkage and 3 Death PCA with single linkage and



bd.dist
hclust (*, "single")

```
dev.copy(jpeg,filename="~/Desktop/jenn/teaching/ADA2/lecture notes/plots/ch14plot6.jpg")
```

```
## jpeg
##    3
```

```
dev.off()
```

```
## pdf
```

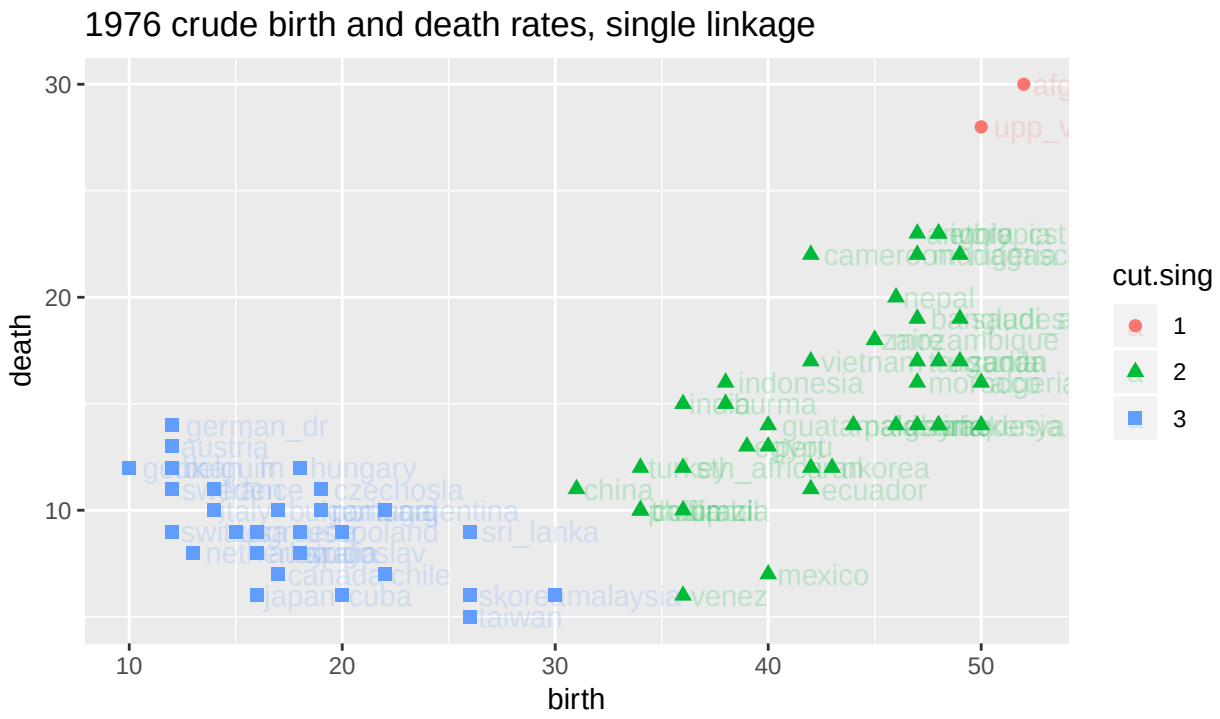
```
##   2
```

```
# create a column with group membership
bd$cut.sing <- factor(cutree(bd.hc.single, k = i.clus))



# print the observations in each cluster
for (i.cut in 1:i.clus) {
  print(paste("Cluster", i.cut, " ------------------------------ "))
  print(bd[(cutree(bd.hc.single, k = i.clus) == i.cut),])
}
```

```
## [1] "Cluster 1  ------------------------------ "
##      country birth death cut.comp cut.sing
## 1     afghan    52    30        1        1
## 70 upp_volta    50    28        1        1
## [1] "Cluster 2  ------------------------------ "
##       country birth death cut.comp cut.sing
## 2     algeria    50    16        1        2
## 3      angola    47    23        1        2
## 7   banglades    47    19        1        2
## 9      brazil    36    10        3        2
## 11      burma    38    15        3        2
## 12   cameroon    42    22        1        2
## 15      china    31    11        3        2
## 17   columbia    34    10        3        2
## 20    ecuador    42    11        3        2
## 21      egypt    39    13        3        2
## 22   ethiopia    48    23        1        2
## 26      ghana    46    14        1        2
## 28  guatamala    40    14        3        2
## 30      india    36    15        3        2
## 31  indonesia    38    16        3        2
## 32       iran    42    12        3        2
## 33       iraq    48    14        1        2
## 35  ivory_cst    48    23        1        2
## 37      kenya    50    14        1        2
## 38     nkorea    43    12        3        2
## 40  madagasca    47    22        1        2
## 42     mexico    40     7        3        2
## 43    morocco    47    16        1        2
## 44 mozambique    45    18        1        2
## 45      nepal    46    20        1        2
## 47    nigeria    49    22        1        2
## 48   pakistan    44    14        3        2
## 49       peru    40    13        3        2
## 50    phillip    34    10        3        2
## 53   rhodesia    48    14        1        2
## 55   saudi_ar    49    19        1        2
## 56 sth_africa    36    12        3        2
## 59      sudan    49    17        1        2
## 62      syria    47    14        1        2
## 63   tanzania    47    17        1        2
## 64   thailand    34    10        3        2
```

```
## 65     turkey    34    12        3        2
## 67     uganda    48    17        1        2
## 71      venez    36     6        3        2
## 72    vietnam    42    17        1        2
## 74      zaire    45    18        1        2
## [1] "Cluster 3  ---------------------------- "
##      country birth death cut.comp cut.sing
## 4  argentina    22    10        2        3
## 5  australia    16     8        2        3
## 6    austria    12    13        2        3
## 8    belguim    12    12        2        3
## 10  bulgaria    17    10        2        3
## 13    canada    17     7        2        3
## 14     chile    22     7        2        3
## 16    taiwan    26     5        3        3
## 18      cuba    20     6        2        3
## 19 czechosla    19    11        2        3
## 23    france    14    11        2        3
## 24 german_dr    12    14        2        3
## 25 german_fr    10    12        2        3
## 27    greece    16     9        2        3
## 29   hungary    18    12        2        3
## 34     italy    14    10        2        3
## 36     japan    16     6        2        3
## 39    skorea    26     6        3        3
## 41  malaysia    30     6        3        3
## 46 netherlan    13     8        2        3
## 51    poland    20     9        2        3
## 52  portugal    19    10        2        3
## 54   romania    19    10        2        3
## 57     spain    18     8        2        3
## 58 sri_lanka    26     9        3        3
## 60    sweden    12    11        2        3
## 61   switzer    12     9        2        3
## 66      ussr    18     9        2        3
## 68        uk    12    12        2        3
## 69       usa    15     9        2        3
## 73  yugoslav    18     8        2        3
```

```r
# plot original data
library(ggplot2)
p1 <- ggplot(bd, aes(x = birth, y = death, colour = cut.sing, shape = cut.sing))
p1 <- p1 + geom_point(size = 2)                                    # points
p1 <- p1 + geom_text(aes(label = country), hjust = -0.1, alpha = 0.2)   # labels
p1 <- p1 + coord_fixed(ratio = 1) # makes 1 unit equal length on x- and y-axis
p1 <- p1 + labs(title = "1976 crude birth and death rates, single linkage")
print(p1)
```




```
dev.copy(jpeg,filename="~/Desktop/jenn/teaching/ADA2/lecture notes/plots/ch14plot7.jpg")
```

```
## jpeg
##    3
```

```
dev.off()
```

```
## pdf
##   2
```