This is an R HTML document. When you click the **Knit HTML** button a web page will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
who <- read.csv("C:/Users/prera/Downloads/Life_Expectancy_Data.csv")
head(who)
```

```
##       Country Year     Status Life.expectancy Adult.Mortality infant.deaths
## 1 Afghanistan 2015 Developing            65.0             263            62
## 2 Afghanistan 2014 Developing            59.9             271            64
## 3 Afghanistan 2013 Developing            59.9             268            66
## 4 Afghanistan 2012 Developing            59.5             272            69
## 5 Afghanistan 2011 Developing            59.2             275            71
## 6 Afghanistan 2010 Developing            58.8             279            74
##   Alcohol percentage.expenditure Hepatitis.B Measles  BMI under.five.deaths
## 1    0.01              71.279624          65    1154 19.1                83
## 2    0.01              73.523582          62     492 18.6                86
## 3    0.01              73.219243          64     430 18.1                89
## 4    0.01              78.184215          67    2787 17.6                93
## 5    0.01               7.097109          68    3013 17.2                97
## 6    0.01              79.679367          66    1989 16.7               102
##   Polio Total.expenditure Diphtheria HIV.AIDS       GDP Population
## 1     6              8.16         65      0.1 584.25921   33736494
## 2    58              8.18         62      0.1 612.69651     327582
## 3    62              8.13         64      0.1 631.74498   31731688
## 4    67              8.52         67      0.1 669.95900    3696958
## 5    68              7.87         68      0.1  63.53723    2978599
## 6    66              9.20         66      0.1 553.32894    2883167
##   thinness..1.19.years thinness.5.9.years Income.composition.of.resources
## 1                 17.2               17.3                           0.479
## 2                 17.5               17.5                           0.476
## 3                 17.7               17.7                           0.470
## 4                 17.9               18.0                           0.463
## 5                 18.2               18.2                           0.454
## 6                 18.4               18.4                           0.448
##   Schooling
## 1      10.1
## 2      10.0
## 3       9.9
## 4       9.8
## 5       9.5
## 6       9.2
```

```
dim(who)
```

```
## [1] 2938   22
```

```
status.of.countries <- who[(who$Status %in% c("Developing") & who$Life.expectancy<55) | (who$Status %in% c("Developed") & who$Life.expectancy>80) ,]
dim(status.of.countries)
```

```
## [1] 509  22
```

```
#View(status.of.countries)

class(status.of.countries)
```

```
## [1] "data.frame"
```

```
head(status.of.countries)
```

```
##        Country Year     Status Life.expectancy Adult.Mortality infant.deaths
## 16 Afghanistan 2000 Developing            54.8             321            88
## 49      Angola 2015 Developing            52.4             335            66
## 50      Angola 2014 Developing            51.7             348            67
## 51      Angola 2013 Developing            51.1             355            69
## 53      Angola 2011 Developing            51.0             361            75
## 54      Angola 2010 Developing            49.6             365            78
##    Alcohol percentage.expenditure Hepatitis.B Measles  BMI under.five.deaths
## 16    0.01               10.42496          62    6532 12.2               122
## 49      NA                0.00000          64     118 23.3                98
## 50    8.33               23.96561          64   11699 22.7               101
## 51    8.10               35.95857          77    8523 22.1               105
## 53    8.06              239.89139          72    1449 21.0               115
## 54    7.80              191.65374          77    1190  2.4               121
##    Polio Total.expenditure Diphtheria HIV.AIDS       GDP Population
## 16    24              8.20         24      0.1  114.5600     293756
## 49     7                NA         64      1.9 3695.7937    2785935
## 50    68              3.31         64      2.0  479.3122    2692466
## 51    67              4.26         77      2.3  484.6169    2599834
## 53    73              3.38         71      2.5 4299.1289   24218565
## 54    81              3.39         77      2.5 3529.5348   23369131
##    thinness..1.19.years thinness.5.9.years Income.composition.of.resources
## 16                  2.3                2.5                           0.338
## 49                  8.3                8.2                           0.531
## 50                  8.5                8.3                           0.527
## 51                  8.6                8.5                           0.523
## 53                  8.9                8.8                           0.495
## 54                  9.1                9.0                           0.488
##    Schooling
## 16       5.5
## 49      11.4
## 50      11.4
## 51      11.4
## 53       9.4
## 54       9.0
```

```
#View(status.of.countries)
WHONew<-status.of.countries
#resting the index values
row.names(WHONew) <- NULL
#View(WHONew)
dim(WHONew)
```

```
## [1] 509  22
```

```
for(i in 1:347)
{
  if(is.na(WHONew$Alcohol[i]))
  {
    WHONew$Alcohol[i] <- with(WHONew, mean(WHONew$Alcohol[Country == WHONew$Country[i]], na.rm = TRUE))
  }
}
for(i in 1:347)
{
  if(is.na(WHONew$Hepatitis.B[i]))
  {
    WHONew$Hepatitis.B[i] <- with(WHONew, mean(WHONew$Hepatitis.B[Country == WHONew$Country[i]], na.rm = TRUE))
  }
}
for(i in 1:347)
{
  if(is.na(WHONew$Total.expenditure[i]))
  {
    WHONew$Total.expenditure[i] <- with(WHONew, mean(WHONew$Total.expenditure[Country == WHONew$Country[i]], na.rm = TRUE))
  }
}
dim(WHONew)
```

```
## [1] 509  22
```

```
# Deleting the Empty rows where there is no data present.
new.life<- na.omit(WHONew)
dim(new.life)
```

```
## [1] 285  22
```

```
#View(new.life)

install.packages("cluster", lib="/Library/Frameworks/R.framework/Versions/3.5/Resources/library")
```

```
## Warning in install.packages("cluster", lib = "/Library/Frameworks/R.framework/
## Versions/3.5/Resources/library"): 'lib = "/Library/Frameworks/R.framework/
## Versions/3.5/Resources/library"' is not writable
```

```
## Error in install.packages("cluster", lib = "/Library/Frameworks/R.framework/Versions/3.5/Resources/library"): unable to install packages
```

```
library(cluster)
```

```
## Warning: package 'cluster' was built under R version 3.6.3
```

```
bar <- subset.data.frame(new.life, Year == '2015')
dim(bar)
```

```
## [1] 16 22
```

```
rownames(bar)<-1:16
dim(bar)
```

```
## [1] 16 22
```

```
new.life123<- na.omit(bar)
dim(new.life123)
```

```
## [1] 16 22
```

```
#cluster.life <- bar[ c(1:10),c(4,5,7,11,16,17,19,20,21,22)]
```

You can also embed plots, for example:

```r
install.packages("factoextra")
```

```
## Installing package into 'C:/Users/prera/OneDrive/Documents/R/win-library/3.6'
## (as 'lib' is unspecified)
```

```
## Error in contrib.url(repos, "source"): trying to use CRAN without setting a mirror
```

```r
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 3.6.3
```

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

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

```r
library(dplyr)
```

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

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

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

```r
#For plotting the scatter density plots
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from   
##   +.gg   ggplot2
```

```
## 
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:dplyr':
## 
##     nasa
```

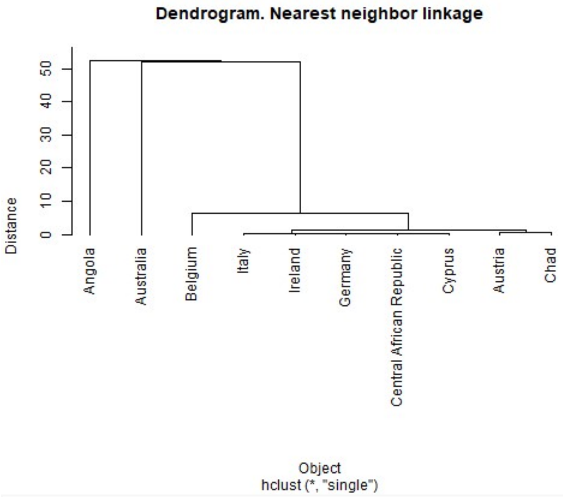
```r
cluster.life <- read.csv("C:/Users/prera/Downloads/life1.csv",row.names=1, fill = TRUE)

View(cluster.life)

dist.mat5 <- as.dist(cluster.life)
dim(cluster.life)
```

```
## [1] 10 10
```

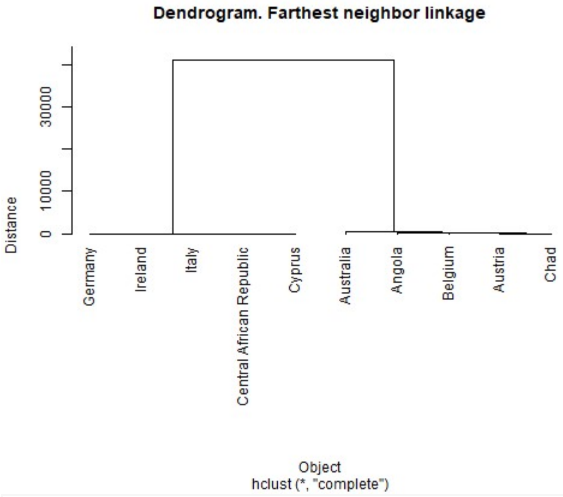
```r
dist.mat5
```

```
##                                Angola    Australia      Austria      Belgium
## Australia                8.280000e+01                                       
## Austria                  8.150000e+01 6.500000e+01                          
## Belgium                  8.110000e+01 7.400000e+01 1.114000e+01             
## Central African Republic 5.250000e+01 3.970000e+02 1.318571e+00 2.270000e+01
## Chad                     5.310000e+01 3.560000e+02 4.071429e-01 1.910000e+01
## Cyprus                   8.500000e+01 5.200000e+01 4.525000e+00 6.300000e+00
## Germany                  8.100000e+01 6.800000e+01 1.111000e+01 6.230000e+01
## Ireland                  8.140000e+01 6.400000e+01 1.126600e+01 6.280000e+01
## Italy                    8.270000e+01 5.600000e+01 7.816364e+00 6.360000e+01
##                          Central African Republic         Chad       Cyprus
## Australia                                                                  
## Austria                                                                    
## Belgium                                                                    
## Central African Republic                                                   
## Chad                                 2.800000e+00                          
## Cyprus                               1.000000e-01 2.375113e+03             
## Germany                              1.000000e-01 4.117688e+04 1.100000e+00
## Ireland                              1.000000e-01 6.664144e+03 3.000000e-01
## Italy                                1.000000e-01 3.491476e+02 6.000000e-01
##                               Germany      Ireland
## Australia                                         
## Austria                                           
## Belgium                                           
## Central African Republic                          
## Chad                                              
## Cyprus                                            
## Germany                                           
## Ireland                  2.000000e-01
## Italy                    6.000000e-01 8.810000e-01
```

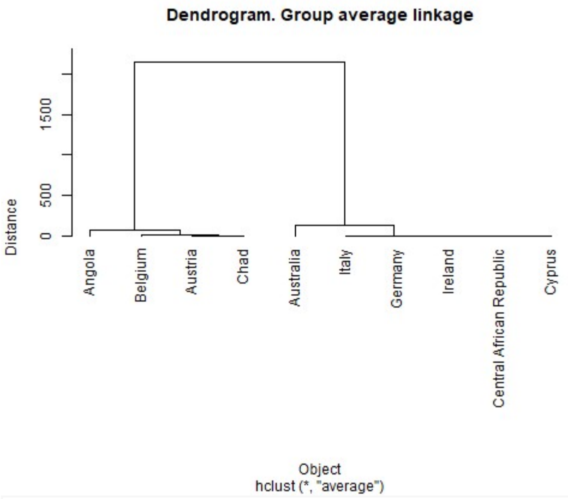
```
#index(cluster.life, data.frame(1="app"))

#Input dataset is a matrix where each row is a sample, and each column is
#a variable.Clustering is performed on a square matrix (sample x sample)
#that provides the distance between samples. It can be computed using the
#dist() or the cor() function depending on the question.The hclust()
#function is used to perform the hierarchical clustering.

#Single
mat5.nn <- hclust(dist.mat5, method = "single")
plot(mat5.nn, hang=-1,xlab="Object",ylab="Distance",
     main="Dendrogram. Nearest neighbor linkage")
```



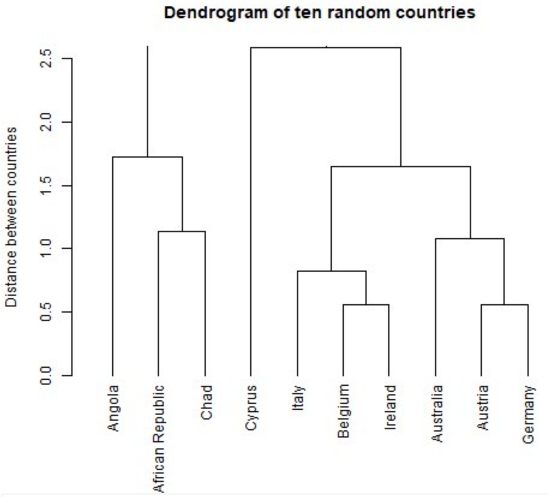
```
#Default - Complete
mat5.fn <- hclust(dist.mat5)
plot(mat5.fn,hang=-1,xlab="Object",ylab="Distance",
     main="Dendrogram. Farthest neighbor linkage")
```



```
#Average
mat5.avl <- hclust(dist.mat5,method="average")
plot(mat5.avl,hang=-1,xlab="Object",ylab="Distance",
     main="Dendrogram. Group average linkage")
```



```
# Canines
matstd.can <- scale(cluster.life)

# Creating a (Euclidean) distance matrix of the standardized data
dist.canine <- dist(matstd.can, method="euclidean")

# Invoking hclust command (cluster analysis by single linkage method)
cluscanine.nn <- hclust(dist.canine, method = "single")

# Plotting vertical dendrogram
# create extra margin room in the dendrogram, on the bottom
par(mar=c(6, 4, 4, 2) + 0.1)
plot(as.dendrogram(cluscanine.nn),ylab="Distance between countries",ylim=c(0,2.5),main="Dendrogram of ten random countries")
```

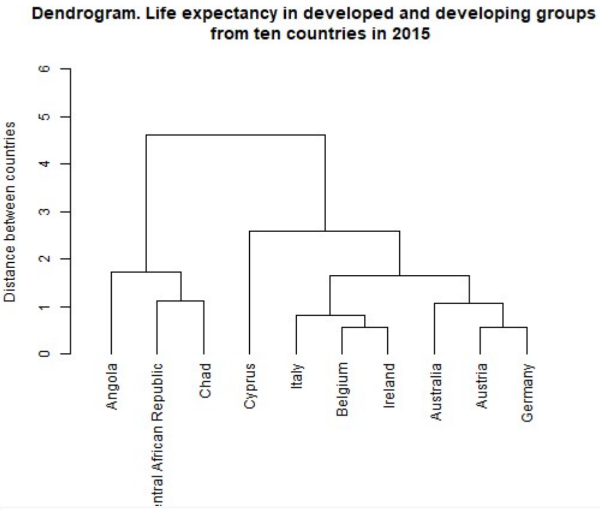


```
##Here in the above graph we can see that the countries "Angola","Central African Republic",
#"Chad" are grouped together as they are developing countries and the rest are grouped together
#as they are developed countries.We can also observe that the values of chad and
#Central African Republic are similar hence they come under a single branch while
#there values add upto Angola which can be visualised in the graph

# New Example

employ <- cluster.life
attach(employ)
dim(employ)
```

```
## [1] 10 10
```

```
# Hirerarchic cluster analysis, Nearest-neighbor
#1)Take distances between objects.
#2)Seek the smallest distance between 2 objects.
#3)Aggregate the 2 objects in a cluster.
#4)Replace them with their barycenter. ??? Again until having only one cluster
#containing every points.There are several ways to calculate the distance
#between 2 clusters ( using the max between 2 points of the clusters, or
#the mean, or the min, or ward (default) )

# Standardizing the data with scale()
```

```
matstd.employ <- scale(employ[,1:10])
View(matstd.employ)
# Creating a (Euclidean) distance matrix of the standardized data
dist.employ <- dist(matstd.employ, method="euclidean")
# Invoking hclust command (cluster analysis by single linkage method)
clusemploy.nn <- hclust(dist.employ, method = "single")


#Plotting

# Create extra margin room in the dendrogram, on the bottom (Countries labels)
par(mar=c(8, 4, 4, 2) + 0.1)
# Object "clusemploy.nn" is converted into a object of class "dendrogram"
# in order to allow better flexibility in the (vertical) dendrogram plotting.
plot(as.dendrogram(clusemploy.nn),ylab="Distance between countries",ylim=c(0,6),
     main="Dendrogram. Life expectancy in developed and developing groups  \n from ten countries in 2015")
```



```
#Horizontal Dendrogram

dev.new()
par(mar=c(5, 4, 4, 7) +0.1)
plot(as.dendrogram(clusemploy.nn), xlab= "Distance between countries", xlim=c(6,0),
     horiz = TRUE,main="Dendrogram. Life expectancy in developed and developing groups  \n from ten countries in 2015")

#K-Means Clustering
#The purpose of clustering analysis is to identify patterns in your data and create groups
#according to those patterns. Therefore, if two points have similar characteristics, that
#means they have the same patternand consequently, they belong to the same group. By
#doing clustering analysis we should be able to check what #features usually appear
#together and see what characterizes a group.
# Standardizing the data with scale()
matstd.employ <- scale(cluster.life)
# K-means, k=2, 3, 4, 5, 6
# Centers (k's) are numbers thus, 10 random sets are chosen
# Computing the percentage of variation accounted for. Two clusters
(kmeans2.employ <- kmeans(matstd.employ,2,nstart = 10))
```

```
## K-means clustering with 2 clusters of sizes 7, 3
##
## Cluster means:
##   Life.expectancy Adult.Mortality    Alcohol        BMI   HIV.AIDS        GDP
## 1       0.6189501      -0.6172773  0.5045037  0.4169728 -0.5723464  0.2813017
## 2      -1.4442168       1.4403136 -1.1771754 -0.9729366  1.3354749 -0.6563706
##   thinness..1.19.years thinness.5.9.years Income.composition.of.resources
## 1           -0.6166803         -0.6151248                       0.6064828
## 2            1.4389207          1.4352911                      -1.4151266
##    Schooling
## 1  0.5588619
## 2 -1.3040111
##
## Clustering vector:
##                   Angola                Australia                  Austria
##                        2                        1                        1
##                  Belgium Central African Republic                     Chad
##                        1                        2                        2
##                   Cyprus                  Germany                  Ireland
##                        1                        1                        1
##                    Italy
##                        1
##
## Within cluster sum of squares by cluster:
## [1] 15.915934  3.196165
##  (between_SS / total_SS =  78.8 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

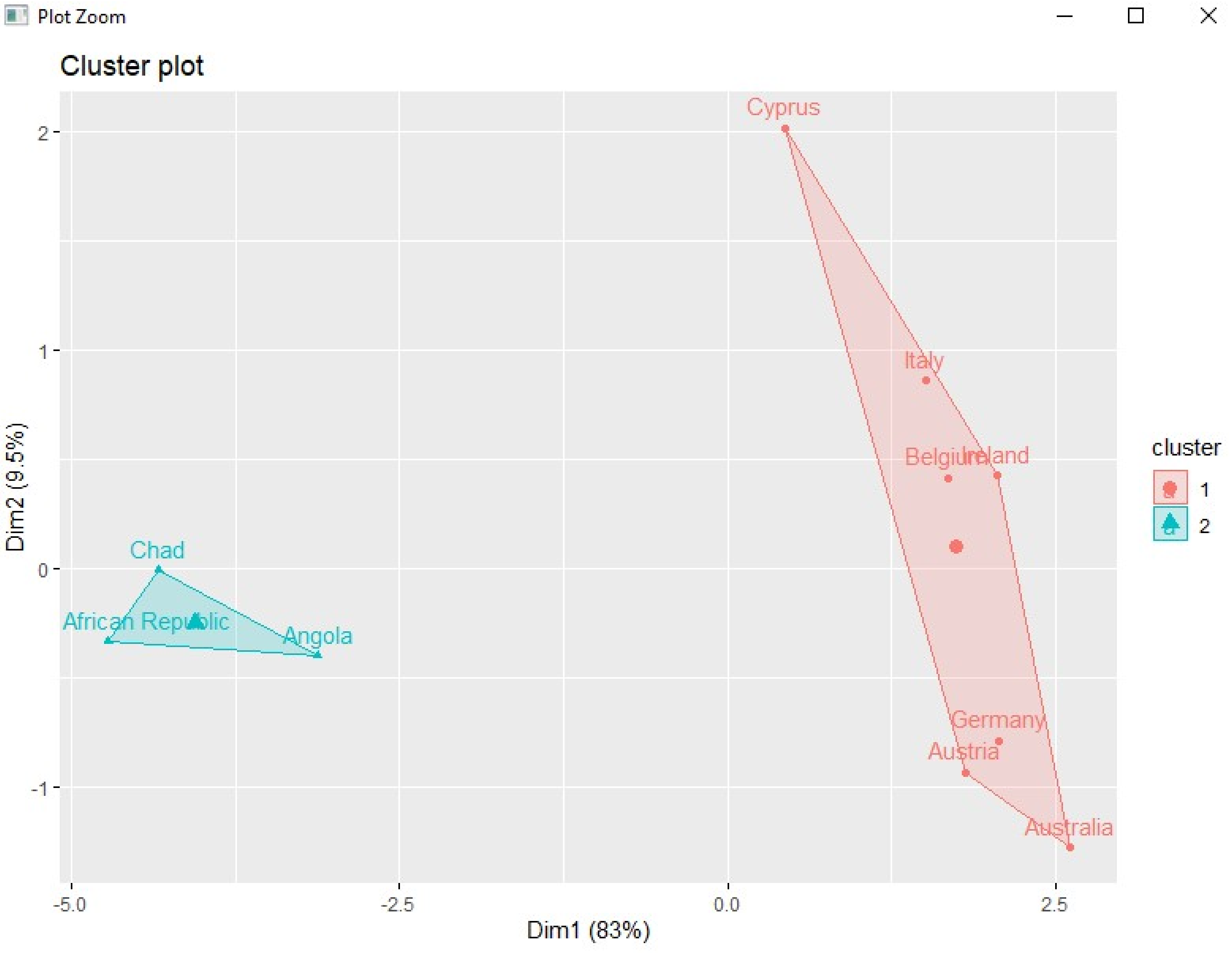
```
perc.var.2 <- round(100*(1 - kmeans2.employ$betweenss/kmeans2.employ$totss),1)
names(perc.var.2) <- "Perc. 2 clus"
perc.var.2
```

```
## Perc. 2 clus
##         21.2
```

```
# Saving two k-means clusters in a list
clus1 <- matrix(names(kmeans2.employ$cluster[kmeans2.employ$cluster == 1]),
                ncol=1, nrow=length(kmeans2.employ$cluster[kmeans2.employ$cluster == 1]))
colnames(clus1) <- "Cluster 1"
clus2 <- matrix(names(kmeans2.employ$cluster[kmeans2.employ$cluster == 2]),
                ncol=1, nrow=length(kmeans2.employ$cluster[kmeans2.employ$cluster == 2]))
colnames(clus2) <- "Cluster 2"

list(clus1,clus2)
```

```
## [[1]]
##      Cluster 1
## [1,] "Australia"
## [2,] "Austria"
## [3,] "Belgium"
## [4,] "Cyprus"
## [5,] "Germany"
## [6,] "Ireland"
## [7,] "Italy"
##
## [[2]]
##      Cluster 2
## [1,] "Angola"
## [2,] "Central African Republic"
## [3,] "Chad"
```

```
# graph for cluster
#From the graph below results we can observe that our groupings resulted in
#2 cluster sizes of 7 and 3. The 7 group is taken from the developed countries
#and 3 group is taken from developing countries. We observe that cluster centers
#or (means) for the two groups across the ten different variables (Life expectancy,
#Adult Mortality, Alcohol, BMI, HIV AIDS, GDP, Thinness 1-19 years, Thinness 5-19
#years, Income composition of resources and Schooling). We also get the cluster
#assignment for each observation (i.e. Australia, Austria, Belgium, Cyprus, Germany,
#Ireland and Italy was assigned to cluster 1 and Angola, Central African Republic
#and Chad was assigned to cluster 1).

fviz_cluster(kmeans2.employ, data= cluster.life)
```


Plot Zoom
Cluster plot
Cyprus
Italy
Belgium
Ireland
Germany
Austria
Australia
Chad
African Republic
Angola
cluster
1
2
Dim2 (9.5%)
2
1
0
-1
Dim1 (83%)
-5.0
-2.5
0.0
2.5