Identifying clusters on planet Earth: an unsupervised approach to country analysis

Picascia Sergio
Data Science and Economics
University of Milan

Abstract

Is the USA still the first world power? Are the African states catching up with other economies? In which direction are developing the BRICS countries in the latest years? These and many other questions come to our mind if we reflect about the social and economic situation of different regions around the globe. Keeping track of the evolutions each society undergoes is a complex task, due to the amount of countries and indicators to monitor. In order to simplify it, I adopt an unsupervised approach that allows to group similar realities together: k-means and hierarchical clustering are ideal methods to aggregate individuals according to their distance, while PCA shrinks the dimensions considered so that a clearer interpretation can be given. The data is gathered from UNdata, a web based data service that provides statistical profiles for each area of the world; the information collected are various and concern economic, social, environmental and infrastructural indicators. It will surprise us that not only China has reached USA as first world power (if not surpassed it), but they are more similar, in some aspect, than we could imagine, despite the differences in political view. The disparity between African countries and the rest of the world are still very evident, while another interesting cluster is composed by states of the Arabian peninsula, which base their economy on oil.

Introduction

The amount of countries and the data we can collect about them these days is huge. Understanding the ones who resemble each other the most and what distinguishes them from other clusters is the aim of this analysis. Unsupervised methods come in handy in this situation, since we can identify groups of individuals, based on how much each of them is similar to the others, and interpret the results establishing what are the attributes that characterise a cluster rather than another one, instead of look for each feature individually.

The dataset used is available online at Kaggle and it is, in turn, collected from UNData.org: in particular, it contains the profiles of countries from all over the world updated on 2017. The indicators observed regard three main

area, such as economics, society, and environment and infrastructure, together with other general information like population and surface. The original dataset was composed of 229 observation and 49 variables, but lots of missing values were present, so I left out columns with more than 40 NAs and countries which still missed some observation. A couple of attributes needed to be split, because they represented different variables grouped in the same column (for example: percentage of women/men, or rural/urban areas). Finally, I recoded the factors of the region variable, grouping countries by the six major areas of the world: this helped noticing if the split generated by the analysis followed in some way the geographical allocation of the individuals.

Here follows the complete list of attributes, divided by area of interest:

General information:

- Region
- Surface area (km^2)
- Population in thousands
- Population density (per km^2)
- Sex ratio (m per 100 f)

Economic indicators:

- GDP: Gross domestic product (million current US$)
- GDP growth rate (annual %, const. 2005 prices)
- GDP per capita (current US$)
- Economy: Agriculture (% of GVA)
- Economy: Industry (% of GVA)
- Economy: Services and other activity (% of GVA)
- Employment: Agriculture (% of employed)
- Employment: Industry (% of employed)
- Employment: Services (% of employed)
- Unemployment (% of labour force)
- Labour force participation (female pop. %)
- Labour force participation (male pop. %)
- Agricultural production index (2004-2006=100)
- Food production index (2004-2006=100)
- International trade: Exports (million US$)
- International trade: Imports (million US$)
- International trade: Balance (million US$)

Social indicators:

- Population growth rate (average annual %)
- Urban population (% of total population)
- Urban population growth rate (average annual %)
- Fertility rate, total (live births per woman)
- Life expectancy at birth (females, years)

- Life expectancy at birth (males, years)
- Population age distribution (0-14, %)
- Population age distribution (60+ years, %)
- International migrant stock (x1000)
- International migrant stock (% of total pop.)
- Infant mortality rate (per 1000 live births)
- Health: Total expenditure (% of GDP)
- Seats held by women in national parliaments

Environment and infrastructure indicators:

- % Mobile-cellular subscriptions (per 100 inhabitants)
- Individuals using the Internet (per 100 inhabitants)
- Threatened species (number) Forested area (% of land area)
- CO_2 emission estimates (million tons)
- CO_2 emission estimates (tons per capita)
- Energy production, primary (Petajoules)
- Energy supply per capita (Gigajoules)
- Pop. using improved drinking water (urban, %)
- Pop. using improved drinking water (rural, %)
- Pop. using improved sanitation facilities (urban, %)
- Pop. using improved sanitation facilities (rural, %)

Analysis

Principal Component Analysis

The first step of my analysis consists in performing PCA, Principal Component Analysis, which is very useful considered the number of variables involved; in fact, it is quite difficult to look at 46 different attributes altogether, even impossible to think about plotting them. This approach shrinks the number of dimensions, identifying the so called principal components, which are normalised linear combinations of the features. In particular, the first principal component is the one that capture the largest amount of sample variance subject to the constraint that sum of the squared loadings is equal to 1; these loadings form the loading vector, that position itself in the direction along which the data varies the most. The subsequent principal components follow the same rules as the first one, with the additional constraint of being orthogonal to earlier ones. Since we are dealing with variables that have complete different ranges, scaling the variables before performing the analysis is a must, otherwise we will end up with unbalanced loadings. The crucial result of this approach is that, considering just a few principal components, we are able to plot them together and finally visualise our data in two dimensions.

In order to decide the number of principal components to examine, I decided to look at the scree plot (Figure 1), whose elbow corresponds to the fourth one,

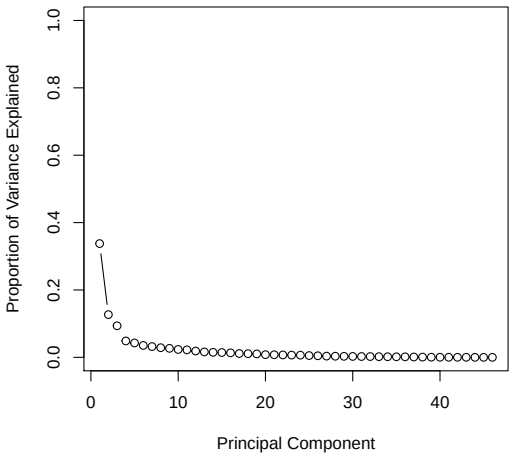


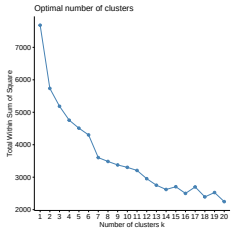
Figure 1: Scree plot.

at which point the amount of variance explained is approximately 61%. The first plot immediately shows a couple of interesting findings: USA and China far away from any other country, having high scores, especially for the second principal component; it is also possible to identify two clusters that resemble the corresponding geographical areas, Africa and Europe. While the majority of African countries are located at the left of the plot, with really low scores for the first principal component, European countries place themselves at the opposite side. Due to the massive amount of information on screen, a further analysis, whose results are discussed later, can be performed looking at the loadings and the scores.

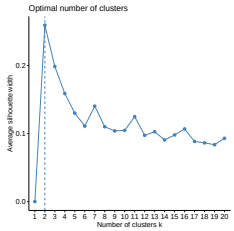
K-means

A completely distinct approach, even though it still belongs to the unsupervised theory, is the identification of clusters among individuals. I performed this different analysis by making use of two methods: k-means and hierarchical clustering. The main difference between the two is that, in the former, it is need an initial definition of the number of clusters, k, to be obtained. There are different methodologies that can be used in order to decide the ideal k (Figure 2): the within-cluster sum of squares, the average silhouette width, the gap statistic. In my case, the suggested k for the last two methods is 2, while for the first one, even though the elbow is not so well defined, 6 seems to be the ideal choice; after that, I performed the analysis with both values.

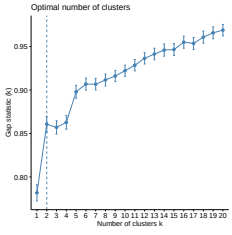
When k-means is run, k random observation are chosen to be the initial centroids: at each iteration, all the points are assigned to the closest centroid



(a) Within-cluster sum of squares.



(b) Average silhouette width.



(c) Gap statistic

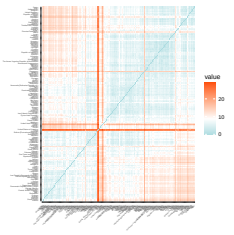
Figure 2: Number of clusters k: decision methods.

based on the Euclidean distance; the algorithm then computes the new k centroids for each new group and repeats the process. After several reiteration, the final groups are returned. In order to visualise the resulting clusters, since we are in a high dimension context, a PCA can be performed and the groups are plotted on the first two principal components.

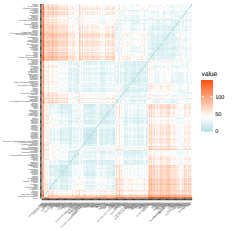
Hierarchical Clustering

Hierarchical clustering has the advantage of not being bound to an initial constraint regarding the number of cluster to consider. In fact, considering the agglomerative type, it starts from a number of clusters equal to the number of observation and aggregates them based on their distance, until there is only one group which contains all the individuals; on the other hand, the divisive hierarchical clustering, works the other way around: it starts from one common group and then splits it up.

Two are the things that have to be taken in consideration: the distance and the linkage methods. For the former, the most used are the Euclidean and the Manhattan (Figure 3), while for the latter, the complete method or the average method are the ones that are preferred. In my case, I opted for the Manhattan distance and Ward method, combination that lead to the highest agglomerative coefficient. Looking at the dendogram, it is possible to identify who are the most similar individuals, at which point they converge and in how many group they can be split.



(a) Euclidean distance.



(b) Manhattan distance.

Figure 3: Distances.

Conclusion

It is possible to draw several conclusions from the analysis we performed, especially from the principal component analysis results.

The first principal component (Figure 4) can be seen as a 'development index': countries placed on the right are the ones with high life expectancy, easy access to internet, improved sanitation facilities and drinking water, economy based on services and industries rather than agriculture, high GDP per capita and a more aged population which lives mostly in urban areas; low fertility rate goes hand in hand with the small percentage of young individuals and the low population growth rate, that are certainly not caused by the infant mortality rate, which is very low. It comes not as a surprise that the highest scores are observed in more developed countries, like USA, Japan, Australia, Canada, as well as the major European states. On the other hand, it is possible to recognise low values for the majority of African countries and some Asian ones, reflecting the poor conditions in which these regions live.

The second principal component (Figure 4) can be addressed as the 'economic power': high values for GDP in millions, international trades (both import and export), energy production and CO_2 emissions; countries with high scores are also the ones with the greatest surface area, population and number of immigrants. Quite surprisingly, we observe an high value also for threatened species that, together with significant quantities of pollutants, indicates that these countries do not care at all about the environment. USA and China have the highest scores, far away from any other country, making them the first two world powers; among the countries that follows, the ones that stand out

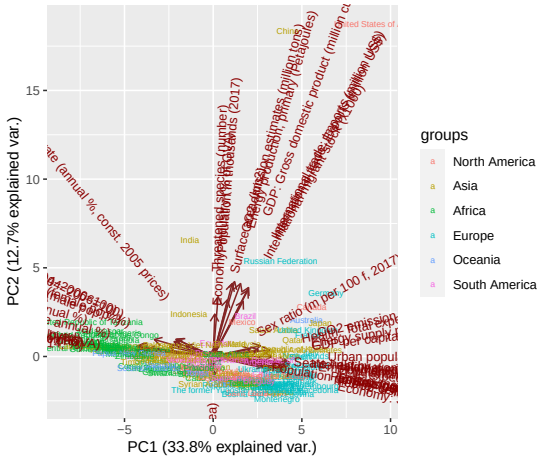


Figure 4: First and second principal components.

are Russia, India and Brazil which, together with China, represent the BRIC group.

The third principal component (Figure 5) is very particular and I defined it 'imported manpower': the population ratio m/f is totally in favour of men, lots of immigrant and high population growth rate, huge CO_2 emissions and energy supply, economy based on industries. There is a specific group of countries with a high score on this component, the ones coming from the Arabian peninsula: Qatar, Kuwait, UAE, Oman, Bahrain and Saudi Arabia. This phenomenon can be addressed to an historic event: the raise of oil price in 1970s. That episode provoked a massive movement of men workers, who were not allowed to bring their respective wife and children, from other Asian region to the peninsula. From that moment on, these states undertook a path that made them distinguishable from any other region.

Finally, looking at the fourth principal component (Figure 5), it becomes clear that less variance is being explained, because there is no more a marked distinction in both loadings and scores. The high values for female employment, also in political positions, immigrants and health expenditure, might lead to the idea of interpreting this fourth dimension as a 'solidarity index'. Once again it is possible to observe high scores for USA, UK, Singapore and Luxembourg, together with some African countries, like Burundi and Uganda, while, on the other side, China and India can be found, together with some Middle-East countries, known for not being so supportive for women.

The k-means analysis brought two different subdivisions: one with 2 clusters and the other with 6. In the first case (Figure 6), it is quite impossible to take any kind of conclusion, because the split is too much generalised: the only

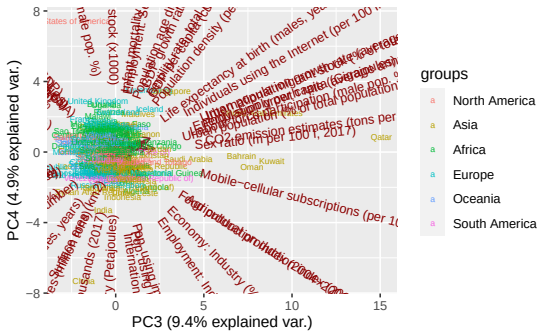


Figure 5: Third and fourth principal components.

statement it is possible to make regards the left group, mainly composed by African countries and some poor Asian regions; the right group just represents the rest of the world.

A more interesting study can be performed on the 6-cluster split (Figure 7); just looking at the plot, we might recognise some patterns already seen in PCA: the first group, on the right of the first principal component, is formed by the most developed country, in particular the major European ones; USA and China are the only presence in the second cluster, they can not be included in any other group, due to the huge distance that separates them from any other country; in group number four, the one on the opposite side of the first, we find African nations, while the countries from the Arabian peninsula form a standalone cluster, the fifth one.

The two formations I left out represents the cluster number three and six, two new faces I didn't discuss before. In order to give a clearer definition of these groups, we can look at the their means: the third group does not distinguish itself for a particular characteristic, it just represents the middle group for every attribute; the sixth group, instead, is formed by small countries, with low population, whose demographic and economic growth is very slow and with a high unemployment rate. Trying to locate this group on the world map, it can be noticed that there are mostly countries from Eastern Europe and South/Central America.

Finally, we look at the dendogram from the hierarchical clustering analysis (Figure 8). The combination of the Manhattan distance and Ward linkage method returns a plot that shows the most interpretable result among the trees I obtained. In the other cases, in fact, countries like USA, China, Singapore and many more, formed a standalone distinct cluster, bringing the need to increase the total number of groups in order to notice substantial differences. The study

Figure 6: K-means with 2 clusters.

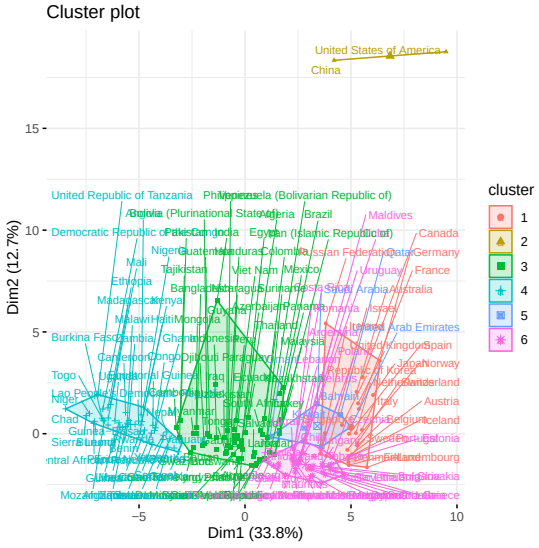


Figure 7: K-means with 6 clusters.

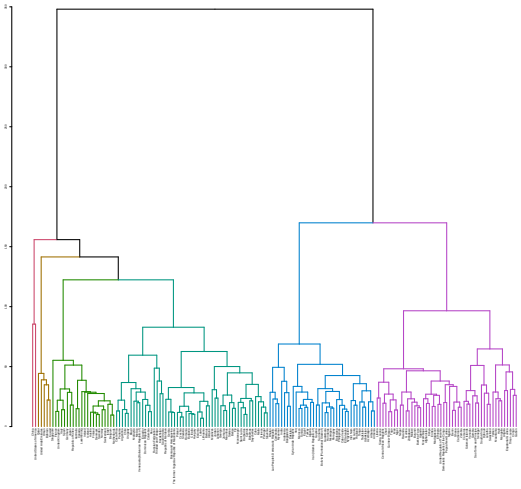


Figure 8: Dendogram with 6 groups.

of this dendogram does nothing but confirm the observation we previously made with the other methods, without adding any other considerable findings.

Several evidences have been found, some of them were quite obvious, while some others were very interesting. Starting from here, a deeper analysis could be performed, considering a even larger number of indicators or trying to retrieve missing information for the individuals that have been left out, even though, the majority of them were very small countries or special municipalities.

R code

```
# Setup
library(tidyverse)
library(devtools)
install_github("vqv/ggbiplot")
library(ggbiplot)
library(factoextra)
library(cluster)
library(dendextend)
set.seed(22)

# Import data
df <- read_delim('/Users/sergiopicascia/Desktop/country_
    profile.csv', ';', na = c('-99', '.../...', '...'))
```

```
str(df)
summary(df)
sapply(df, function(x) sum(is.na(x)))

# Removing 'Region' and columns/rows with too many NA
countries <- subset(df, select = (colSums(is.na(df)) <
    40))
countries <- countries %>% drop_na()

# Separating columns with two values
countries <- separate(countries, col = `Labour force
    participation (female/male pop. %)`, sep = '/', into =
     c('Labour force participation (female pop. %)', '
    Labour force participation (male pop. %)'))
countries <- separate(countries, col = `Life expectancy
    at birth (females/males, years)`, sep = '/', into = c(
    'Life expectancy at birth (females, years)', 'Life
    expectancy at birth (males, years)'))
countries <- separate(countries, col = `Population age
    distribution (0-14 / 60+ years, %)`, sep = '/', into =
     c('Population age distribution (0-14, %)', '
    Population age distribution (60+ years, %)'))
countries <- separate(countries, col = `International
    migrant stock (000/% of total pop.)`, sep = '/', into
    = c('International migrant stock (x1000)', '
    International migrant stock (% of total pop.)'))
countries <- separate(countries, col = `CO2 emission
    estimates (million tons/tons per capita)`, sep = '/',
    into = c('CO2 emission estimates (million tons)', 'CO2
     emission estimates (tons per capita)'))
countries <- separate(countries, col = `Pop. using
    improved drinking water (urban/rural, %)`, sep = '/',
    into = c('Pop. using improved drinking water (urban,
    %)', 'Pop. using improved drinking water (rural, %)'))
countries <- separate(countries, col = `Pop. using
    improved sanitation facilities (urban/rural, %)`, sep
    = '/', into = c('Pop. using improved sanitation
    facilities (urban, %)', 'Pop. using improved
    sanitation facilities (rural, %)'))

# Transform each variable to numeric
countries[countries == c('~0', '~0.0')] <- '0'
countries[, -1:-2] <- mutate_all(countries[, -1:-2],
    function(x) as.numeric(x))
countries[is.na(countries)] <- 0
```

```
# Converting countries to row names, region to
    categorical
countries <- column_to_rownames(countries, 'country')
countries$Region <- as.factor(countries$Region)
countries$Region <- recode(countries$Region, '
    CentralAmerica' = 'North America', 'NorthernAmerica' =
     'North America', 'Caribbean' = 'North America', '
    SouthAmerica' = 'South America', 'EasternAfrica' = '
    Africa', 'MiddleAfrica' = 'Africa', 'NorthernAfrica' =
     'Africa', 'SouthernAfrica' = 'Africa', 'WesternAfrica
    ' = 'Africa', 'CentralAsia' = 'Asia', 'EasternAsia' =
    'Asia', 'South-easternAsia' = 'Asia', 'SouthernAsia' =
     'Asia', 'WesternAsia' = 'Asia', 'EasternEurope' = '
    Europe', 'NorthernEurope' = 'Europe', 'SouthernEurope'
     = 'Europe', 'WesternEurope' = 'Europe', 'Melanesia' =
     'Oceania', 'Polynesia' = 'Oceania')

summary(countries)

### Principal Component Analysis
countries_pca <- prcomp(countries[, -1], center = T,
    scale. = T)
summary(countries_pca)
str(countries_pca)

# Visualisation
countries_pca$sdev
pca_var <- countries_pca$sdev^2
pve <- pca_var / sum(pca_var)

plot(pve, xlab = "Principal Component", ylab = "
    Proportion of Variance Explained", ylim = c(0, 1),
    type = "b")

ggbiplot(countries_pca, obs.scale = 1, var.scale = 1,
    groups = countries$Region)

ggbiplot(countries_pca, labels = row.names(countries),
    obs.scale = 1, labels.size = 2, var.scale = 1, groups
    = countries$Region)

ggbiplot(countries_pca, obs.scale = 1, var.scale = 1,
    groups = countries$Region, choices = c(3, 4))

ggbiplot(countries_pca, labels = row.names(countries),
    obs.scale = 1, labels.size = 2, var.scale = 1, groups
```

```
    = countries$Region, choices = c(3, 4))

# Loadings
loadings_1 <- countries_pca$rotation[, 1]
loadings_1[order(loadings_1)]
loadings_2 <- countries_pca$rotation[, 2]
loadings_2[order(loadings_2)]
loadings_3 <- countries_pca$rotation[, 3]
loadings_3[order(loadings_3)]
loadings_4 <- countries_pca$rotation[, 4]
loadings_4[order(loadings_4)]

# Scores
cov(countries_pca$x)
scores_1 <- countries_pca$x[, 1]
scores_1[order(scores_1)]
scores_2 <- countries_pca$x[, 2]
scores_2[order(scores_2)]
scores_3 <- countries_pca$x[, 3]
scores_3[order(scores_3)]
scores_4 <- countries_pca$x[, 4]
scores_4[order(scores_4)]

### K-means
countries_scaled <- scale(countries[, -1])

# N of clusters
fviz_nbclust(countries_scaled, FUNcluster = kmeans,
    method = 'wss', k.max = 20)
fviz_nbclust(countries_scaled, FUNcluster = kmeans,
    method = 'silhouette', k.max = 20)
gap_stat <- clusGap(countries_scaled, FUN = kmeans,
    nstart = 25, K.max = 20, B = 50)
fviz_gap_stat(gap_stat)

# k = 2
countries_km_2 <- kmeans(countries_scaled, 2, nstart =
    25)
countries_km_2
fviz_cluster(countries_km_2, data = countries_scaled,
    stand = T, repel = T, pointsize = 1, labelsize = 8,
    ggtheme = ggplot2::theme_minimal())

# k = 6
countries_km_6 <- kmeans(countries_scaled, 6, nstart =
    25)
```

```
countries_km_6
fviz_cluster(countries_km_6, data = countries_scaled,
    stand = T, repel = T, pointsize = 1, labelsize = 8,
    ggtheme = ggplot2::theme_minimal())

aggregate(countries[, -1], by = list(cluster = countries_
    km_6$cluster), mean)

### Hierarchical clustering
fviz_nbclust(countries_scaled, FUNcluster = hcut, method
    = 'wss', k.max = 20)
fviz_nbclust(countries_scaled, FUNcluster = hcut, method
    = 'silhouette', k.max = 20)
gap_stat2 <- clusGap(countries_scaled, FUN = hcut, nstart
     = 25, K.max = 20, B = 50)
fviz_gap_stat(gap_stat2)

# Distancies
dist_euc <- get_dist(countries_scaled, method = '
    euclidean')
dist_man <- get_dist(countries_scaled, method = '
    manhattan')
fviz_dist(dist_euc, gradient = list(low = "#00AFBB", mid
    = "white", high = "#FC4E07"), lab_size = 5)
fviz_dist(dist_man, gradient = list(low = "#00AFBB", mid
    = "white", high = "#FC4E07"), lab_size = 5)

# Linkage methods
hclust_avg <- hclust(dist_euc, method = 'average')
plot(hclust_avg, cex = 0.5, hang = -1)
hclust_sing <- hclust(dist_euc, method = 'single')
plot(hclust_sing, cex = 0.5, hang = -1)
hclust_cent <- hclust(dist_euc, method = 'centroid')
plot(hclust_cent, cex = 0.5, hang = -1)
hclust_comp <- hclust(dist_euc, method = 'complete')
plot(hclust_comp, cex = 0.5, hang = -1)
hclust_comp2 <- hclust(dist_man, method = 'complete')
plot(hclust_comp2, cex = 0.5, hang = -1)
hclust_ward <- hclust(dist_euc, method = 'ward.D2')
plot(hclust_ward, cex = 0.5, hang = -1)
hclust_ward2 <- hclust(dist_man, method = 'ward.D2')
plot(hclust_ward2, cex = 0.5, hang = -1)

# Agglomerative coefficient
coef.hclust(hclust_comp)
coef.hclust(hclust_comp2)
```

```
coef.hclust(hclust_ward)
coef.hclust(hclust_ward2)

plot(hclust_ward2, cex = 0.5, hang = -1)
rect.hclust(hclust_ward2, k = 6, border = 2:6)

# Dendogram
dend_ward2 <- as.dendrogram(hclust_ward2)
par(cex = 0.5, mar = c(20,4,4,2) + 0.1)
plot(color_branches(dend_ward2, k = 6))

# Divisive HC
div_hclust <- diana(countries_scaled)
div_hclust$dc
pltree(div_hclust, cex = 0.5, hang = -1)
rect.hclust(div_hclust, k = 5, border = 2:6)
rect.hclust(div_hclust, k = 13, border = 2:6)
```