Nicole Maria Formenti DSE, 941481

STATISTICAL LEARNING PROJECT
unsupervised learning

ABSTRACT

The report focuses on a dataset regarding health indicators of different countries, containing continuous variables and no response variable. In order to find similarities among countries and to produce an intuitive graphical representation, unsupervised learning methods have been used. first, the data dimensionality has been reduced by using principal component analysis. Afterwards, clustering methods such as k-means clustering and hierarchical clustering are applied to obtain subgroups of similar countries. The models have been able to properly summarize 44 countries using 4 or 5 clusters, based on 15 indicators.

GOALS AND METHODS OF THE REPORT

This analysis has two main goals. The first goal is to find clusters of similar data and interpret them, while the second one is to provide a clear graphical representation in order to better understand them. The best method to tackle these unsupervised learning tasks is clustering. Both centroid-based clustering or k-means and hierarchical clustering will be used.
Since the dataset contains many variables, it would be impossible to plot the data. That's the reason why, before applying clustering, another technique called principal component analysis (PCA) is used. This allows to reduce data dimensionality, obtaining far less features, so that it is possible to represent them graphically.
The theory behinds the methods will also be presented.

DATASET DESCRIPTION

The dataset used for the analysis has been provided by Kaggle and it is called 'Country Health Indicator'. It has 180 observations and 70 features, either continuous or dates values. No response variable is included. The observations are countries and the variables are different health indicators, collected from different opensource sources. The categories of indicators are about Covid-19 cases and deaths (Kaggle), death causes (ourworldindata), food sources (FAO1), health care system (WHO), tuberculosis vaccine status (BCG1), school closures (UNESCO) and people/society facts (CIA).
Only a subset of countries and indicators has been used, in order to be able to provide a graphical representation of data, which wouldn't be possible with a high dimensional dataset.
The considered countries are:

- Argentina
- Australia

- Austria
- Belgium
- Bolivia
- Cambodia
- Cameroon
- Central African Republic
- China
- Cuba
- Cyprus
- Denmark
- Dominican Republic
- Finland
- France
- Germany
- Greece
- Haiti
- Iceland
- India
- Israel
- Italy
- Japan
- Kenya
- South Korea
- Malaysia
- Malta
- Mexico
- Morocco
- Netherlands
- New Zealand
- Russia
- Senegal
- South Africa
- Spain
- Sri Lanka
- Switzerland
- Thailand
- Turkey
- US
- United Arab Emirates
- Vietnam

While the indicators used are:

- Vegetables
- Vegetable oils
- Sugar and sweeteners
- Total fertility rate
- Birth rate
- Median age
- Nutritional deficiencies mortality rate

- Net migration rate
- Pneumonia mortality rate
- Cereals (excluding beer)
- Diabetes, blood and endocrine diseases mortality rate
- Tree nuts
- Fish and seafood
- Liver diseases mortality rate
- Animal products

DATA CLEANSING

The dataset Country Health Indicator has a high dimensionality, with 180 observations and 70 variables. Since one of the two aims of this project is to provide a visual representation, it is better to reduce the dimension of the dataset.
First of all, a subset of the 49 most relevant countries is selected. Afterwards, the columns about the covid-19 and school closures have been deleted since no information about their content is available.
The next step is to get rid of variables having a high percentage of missing values and thereafter the observations with at least one null value are dropped. The number of countries left out are 5, so that the cleansed version of the dataset contains 44 observations and 46 indices.
Afterwards, data are checked to assess the presence of outliers. In the above boxplot, it can be observed that many variables have one or more extreme data points far from the main bulk of data.

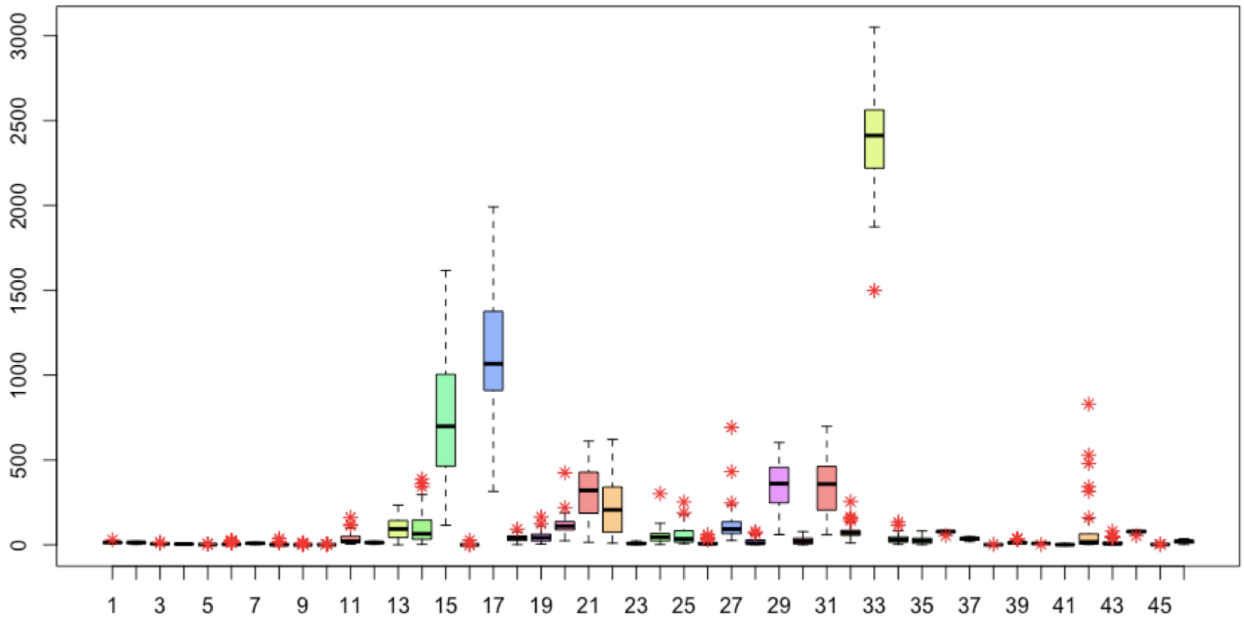

Figure 1: boxplot for all the dataset variables

The outliers have not been removed, due to the fact that the number of data points is small and each observation is rather significative. In order to avoid the negative effect of observations' extreme values on PCA, a robust correlation matrix will be used instead.

DATA ANALYSIS

The analysis will be presented in this section. It relies on the so-called unsupervised learning methods, which are used when the task is not that of predicting a response, but rather that of exploring data in order to find meaningful clues about them.
They are regarded as more difficult to apply with respect to supervised learning techniques, since there is no way to definitely assess how good the model produced are. This is because there is no true response for making comparisons.

PRINCIPAL COMPONENT ANALYSYS

Principal component analysis (PCA) is a useful method when dealing with high dimensional data, which are probably also correlated. It works by reducing the number of variables, summarizing the most relevant ones into principal components. Their direction corresponds to the direction in which the variables vary the most.
Starting from p variables and n data points, p principal components are found, but only the first few ones are significative in explaining data variability. To begin with, the first principal component is computed as a normalized linear combination of the original set of variables $X_1, X_2, \dots, X_p$, having the largest possible sample variance.

$$Z_1 = \emptyset_{11}X_1 + \emptyset_{21}X_2 + \cdots + \emptyset_{p1}X_p$$

Normalized means subject to the constraint

$$\sum_{j=1}^{p} \emptyset_{j1}^2 = 1$$

The principal component loadings vector is $\emptyset_1 = (\emptyset_{11}, \emptyset_{21}, \dots, \emptyset_{p1})'$, defining the direction along which the data have the highest variability
By projecting the n data points $x_1, x_2, \dots, x_n$ onto the loading vector $\emptyset_1$, the principal component scores $z_{11}, z_{21}, \dots, z_{n1}$ are found.
After the first principal component, the second one is calculated in the same way with the additional constraint of $\emptyset_2$ being orthonormal to $\emptyset_1$. This means that Z_2 is uncorrelated to Z_1.
In order to choose the optimal number of principal components to be used, the explained variance is used. If it is greater than one, it means that the principal component explains more variance than one original variable and vice versa. The components which have an explained variance lower than one should be dropped since they are not enough informative.
PCA is considered an unsupervised learning method due to the fact that it doesn't consider any response associated to the features.

Before applying PCA it is important to standardize the variables by subtracting their mean and dividing by their standard deviation. This is done in order to avoid that variables with different unity of measure provide a skewed variance. Moreover, to be able to plot the data points the number of informative principal components must be at most three. That's why instead of using all the features a random subset of 15 variables has been used.

The selected features are:

- Vegetables
- Vegetable oils
- Sugar and sweeteners
- Total fertility rate
- Birth rate
- Median age
- Nutritional deficiencies rate
- Net migration rate
- Pneumonia death rate
- Cereals (excluding beer)
- Diabetes, blood and endocrine diseases rate
- Tree nuts
- Fish and seafood
- Liver diseases rate
- Animal products

In order to compute PCA, the starting point is either the correlation or the covariance matrix. It is preferable to use the correlation matrix since it standardizes the variables, moreover it allows to obtain the variance explained by the principal components. In this case, a robust correlation matrix is used since the outliers have not been removed, so that to end up with a robust PCA. A robust method has the advantage of not being affected by outliers, which is especially useful with PCA, given it's quite sensitive to extreme values of data.
The next step is to calculate the eigenvectors and eigenvalues of the correlation matrix. The eigenvectors are the principal components, while the corresponding eigenvalues coincide with the explained variance. Their value is decreasing.
By looking at the eigenvalues only the first three are useful in explaining data, with a respective explained variance of 7.41, 2.24, 1.41. Altogether they represent about 74% of total data variance.

The explained variance and proportion of cumulating variance (PVE) can be shown graphically.

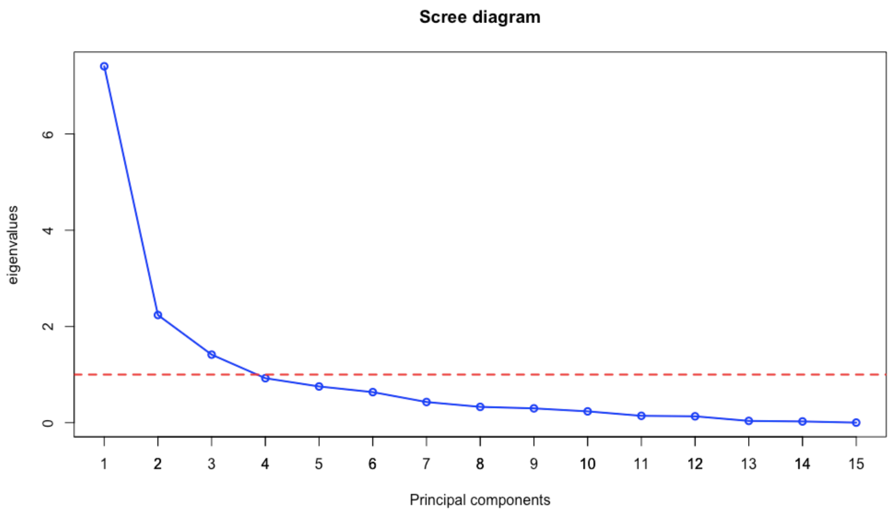

Figure 2: scree diagram showing the explained variance of each principal component

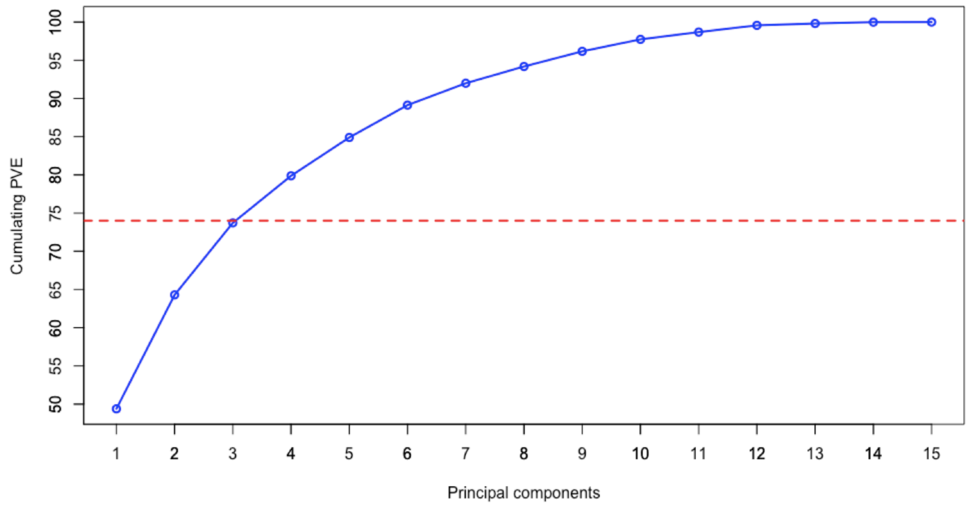

Figure 3: cumulating proportion of variance explained by principal components

The loadings are derived using the selected eigenvalues and the corresponding eigenvectors with the following formula

$$eigenvector \times \sqrt{eigenvalue}$$

Loadings are eigenvectors scaled by their eigenvalues.
Then, the component matrix is created by binding together the loadings of each component.
Another interesting metric is the communality, which is computed as the sum per row of the squares of the values in the component matrix. It represents the fraction of the variance of the original variable accounted for by the principal components.

	PC1	PC2	PC3	communality
vegetables	0.461	-0.137	-0.098	0.240894
vegetable_oils	0.547	0.184	0.765	0.918290
sugar_._sweeteners	0.720	-0.119	0.028	0.533345
total.fertility.rate	-0.907	0.202	0.049	0.865854
birth.rate	-0.949	0.133	0.079	0.924531
median.age	0.958	0.046	-0.097	0.929289
Nutritional.deficiencies %	-0.911	0.053	0.158	0.857694
net.migration.rate	0.668	0.514	-0.111	0.722741
pneumonia.death.rates	-0.896	0.170	-0.078	0.837800
cereals_._excluding_beer	-0.219	-0.750	0.286	0.692257
Diabetes, blood and endocrine.diseases %	0.208	-0.785	0.392	0.813153
treenuts	0.717	0.346	0.515	0.899030
fish._seafood	0.505	-0.206	-0.290	0.381561
Liver.disease %	0.225	-0.657	-0.344	0.600610
animal_products	0.862	0.202	-0.229	0.836289

It can be seen that it is greater than 60 for almost all the variables, excluding the *'sugar and sweeteners'* feature with a still acceptable value for communality (53.33%), plus *'vegetables'* and *'fish and seafood'* variables with a low communality value.

In order to better understand the relationship between the principal components and the original variables, the loadings can be plotted against the features.
Since there are three principal components, two plots are needed. One plot is for representing the first and second principal components and another one for the third principal component.

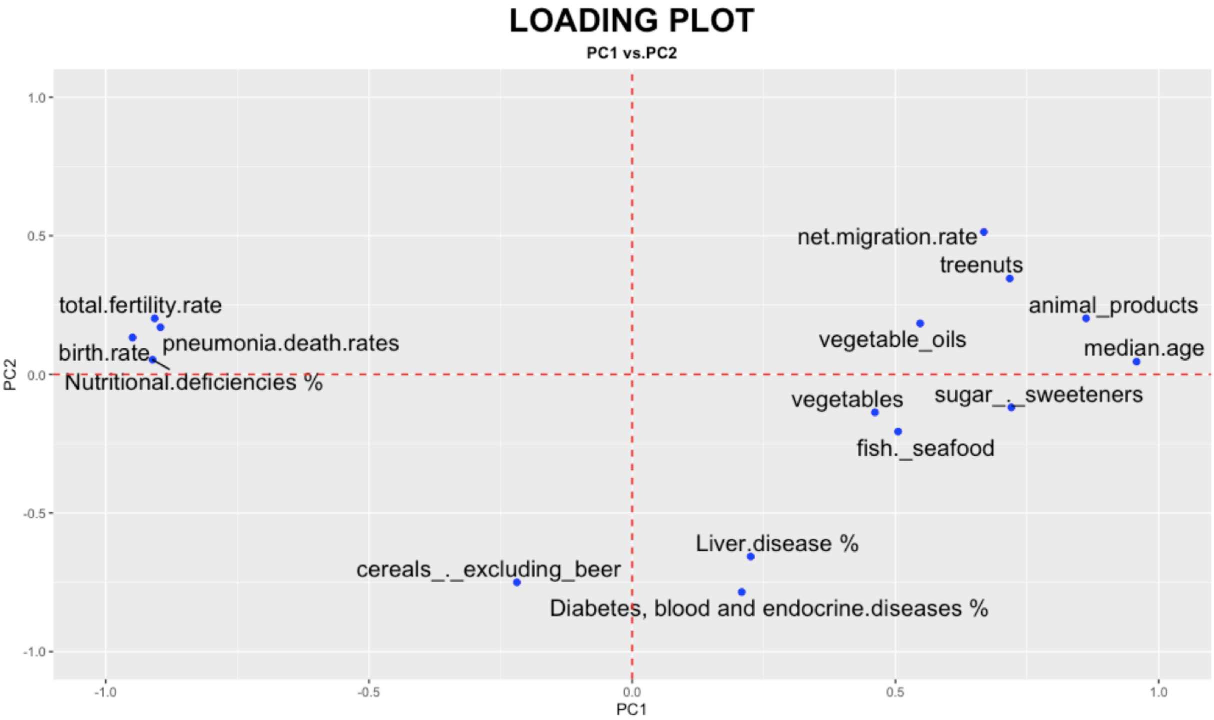

Figure 4: loadings plot for the first and second principal components

The relationship between the first two principal components and variables are as follows:

- *Total fertility rate*, *Pneumonia death rate, Birth rate, Nutritional deficiencies rate*: strong negative correlation with the first PC and slightly positive correlation with the second PC
- *Cereals (excluding beer):* slightly negative correlation with the first PC and strong negative correlation with the second PC
- *Liver disease mortality rate, Diabetes, blood and endocrine diseases mortality rate:* slightly positive correlation with the first PC and strong negative correlation with the second PC
- *Fish and seafood, Vegetables*: medium positive correlation with the first PC and slightly negative correlation with the second PC
- *Sugar and sweeteners:* strong positive correlation with the first PC and slightly negative correlation with the second PC
- *Vegetable oils:* medium positive correlation with the first PC and slightly positive correlation with the second PC
- *Median age, Animal products*: strong positive correlation with the first PC and slightly positive correlation with the second PC
- *Tree nuts, Net migration rate:* strong positive correlation with the first PC and medium positive correlation with the second PC

The first PC could be summarized as type of foods included in the diet, population growth rate (positively correlated to net migration rate and negatively to fertility and birth rate) and age, plus pneumonia death rate.

On the other hand, the second PC is primarily related to death due to pathologies (liver diseases and diabetes, blood and endocrine diseases) and cereals introduced in the diet. Moreover, it also partly explains the net migration rate

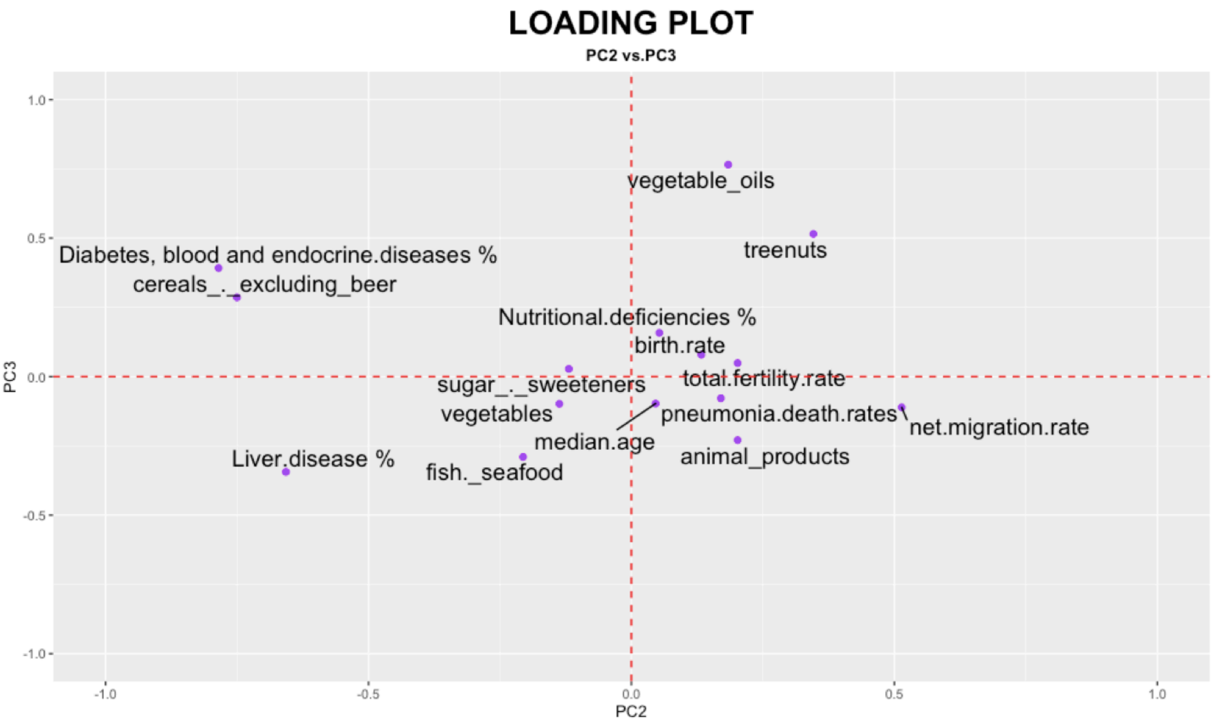

Figure 5: loadings plot for the second and third principal components

Next, the relationship between the third principal component and variables is the following:

- *Vegetable oils*: strongly positive correlation
- *Diabetes, blood and endocrine diseases rate, tree nuts*: medium positive correlation
- *Cereals (excluding beer), nutritional deficiencies rate, birth rate*: slightly positive correlation
- *Total fertility rate, sugar (sweeteners), Vegetables, Pneumonia death rates, Median age, Net migration rate:* nearly null correlation
- *Animal products, Fish seafood, Liver disease rate*: slightly negative correlation

The last principal component captures the effect of the inclusion of certain food in the diet (vegetable oils, tree nuts and cereals) and diabetes, blood and endocrine diseases.

The last step of PCA is to compute the scores. In order to do it, the original scaled data are transformed into the new feature space by using the projection matrix, which is the matrix built by binding together the eigenvectors corresponding to the chosen eigenvalues.
Then, the scores are calculated by dividing the values just found by the square root of the associated eigenvalues.
The scores can be graphically represented against the principal components, in order to better understand the relationship between the principal components and the observations in the dataset

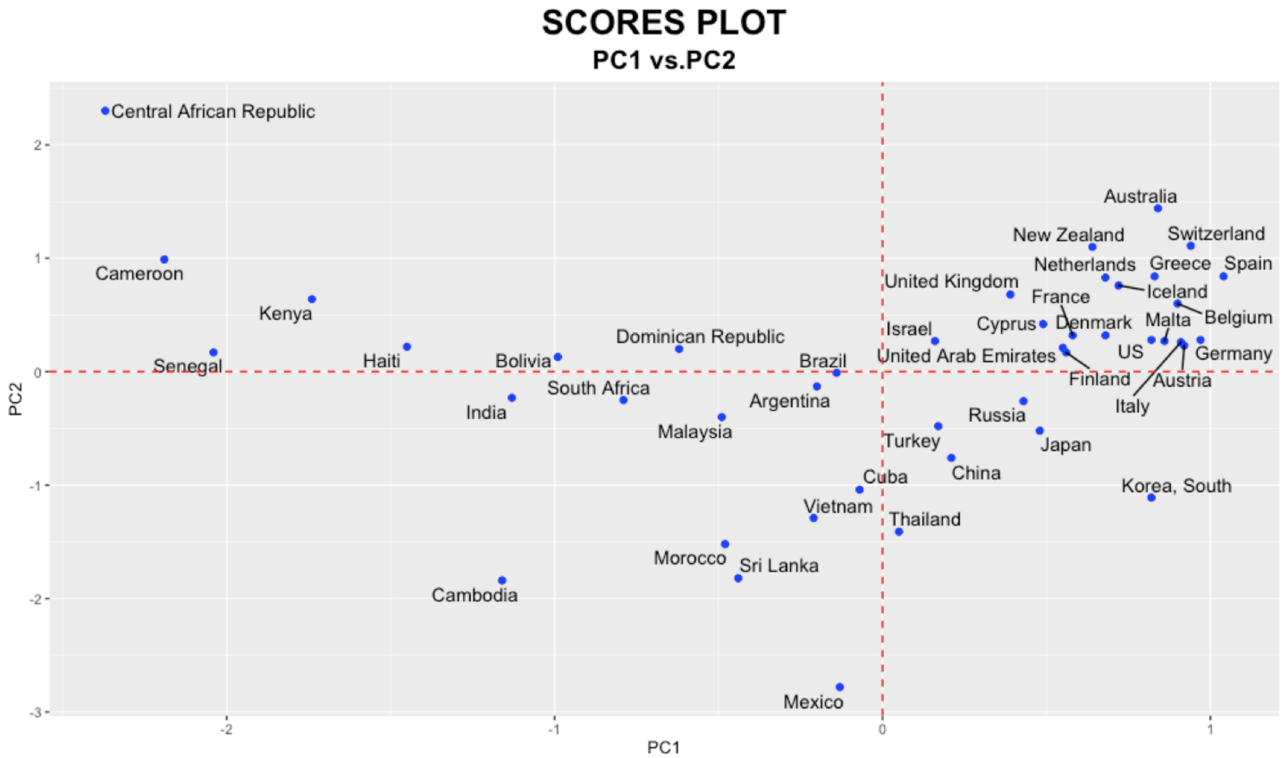

Figure 6: scores plot for the first and second principal components

To summarize, the first principal component is related to the observations as follows:

- *Central African Republic, Cameroon, Senegal, Kenya, Haiti, India, Cambodia, South Africa, Dominican Republic, Malaysia, Morocco:* strong negative relation
- *Brazil, Bolivia, Argentina, Brazil, Vietnam, Sri Lanka, Mexico, Cuba*: slighlty negative relation
- *Israel, Turkey, Thailand, China, Russia, United Kingdom, United Arab Emirates:* slightly positive relation
- *Japan, Cyprus, France, Denmark, Finland, US, Netherlands, New Zealand, South Korea, Italy, Austria, Malta, Iceland, Greece, Switzerland, Australia, Spain, Belgium, Germany:* strong positive relation

Then, the relationship with the second principal component is summarized as:

- *Central African Republic, Australia, Switzerland, New Zealand, Cameroon:* strong positive relation
- *Senegal, Kenya, Haiti, Bolivia, Dominican Republic, Netherlands, Greece, Spain, Iceland, United Kingdom, France, Israel, Cyprus, United Arab Emirates, US, Finland, Italy, Austria, Germany, Malta, Belgium, New Zealand, Australia, Switzerland:* slightly positive relation
- *Brazil*: no relation
- *India, South Africa, Malaysia, Argentina, Turkey, Russia, China, Japan:* slightly negative relation
- *Cambodia, Morocco, Sri Lanka, Mexico, Vietnam, Cuba, Thailand, South Korea:* strong negative relation

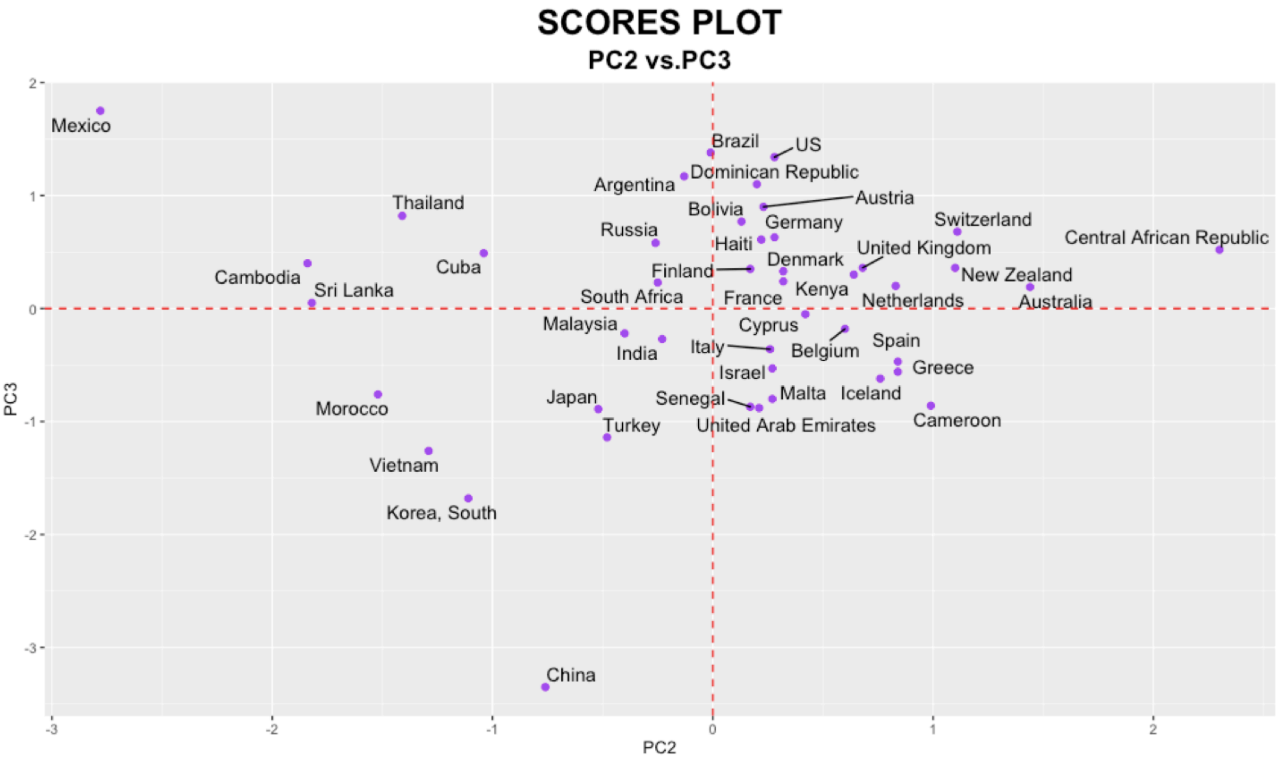

Figure 7: scores plot for the second and third principal components

Last, the third principal components can be summed up as follows:

- *Mexico, Brazil, US, Dominican Republic, Argentina:* strong positive relation
- *Cambodia, Thailand, Cuba, Russia, South Africa, Bolivia, Germany, Austria, Haiti, Denmark, Kenya, France, Finland, Netherlands, United Kingdom, Switzerland, New Zealand, Australia, Central African Republic*: slightly positive relation
- *Sri Lanka, Cyprus:* null relation
- *Vietnam, Turkey, South Korea, China*: strong negative relation
- *Morocco, Japan, Malaysia, India, Italy, Belgium, Spain, Greece, Iceland, Malta, United Arab Emirates, Cameroon, Senegal, Turkey:* slightly negative relation

Lastly, scores and loadings can be represented in the same plot against the principal components, so that it is possible to visualize loadings direction and how they relate with scores. This type of plot is known as biplot

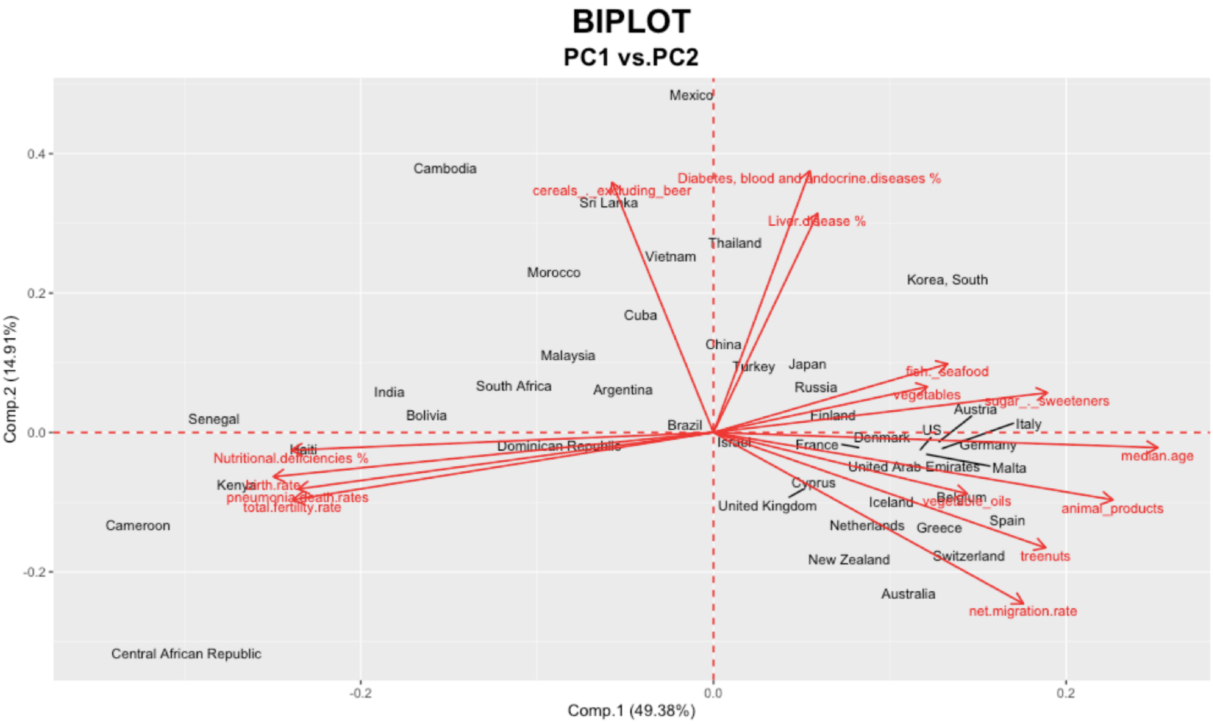

Figure 8: biplot for the first and second principal components

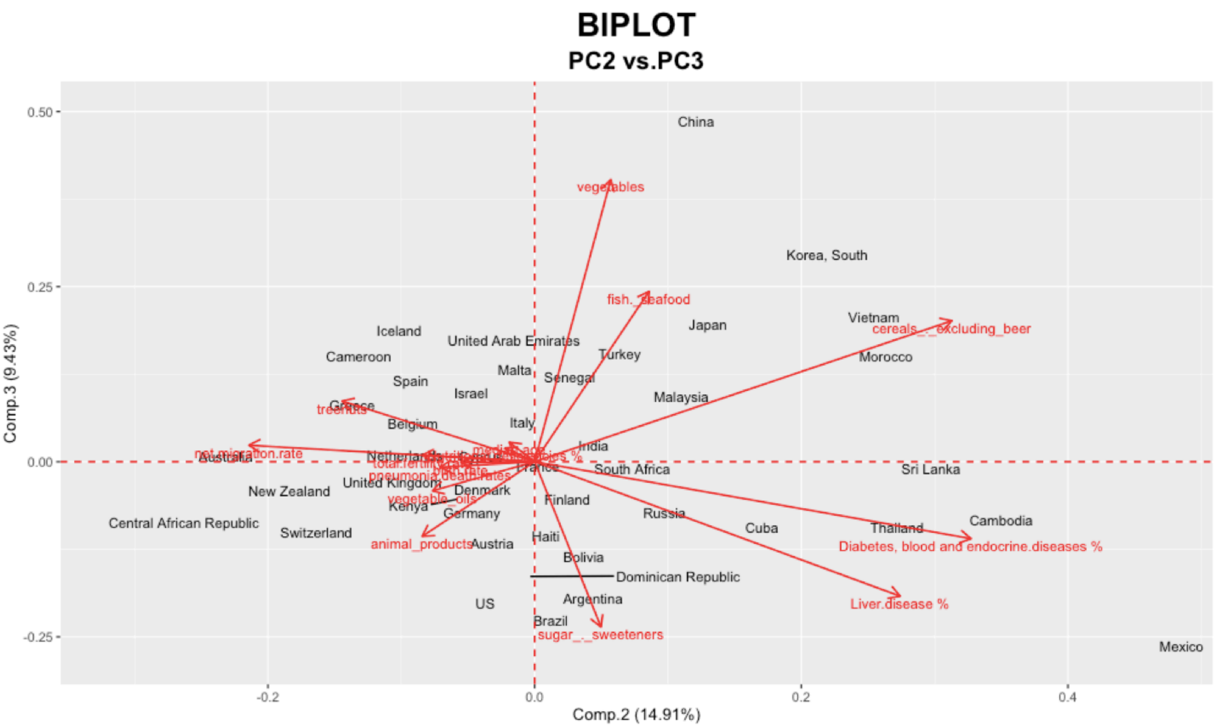

Figure 9: biplot for the second and third principal components

CLUSTERING

After having obtained a new version of the original dataset with reduced dimensionality, the next step is to inspect it in order to find interesting patterns in the data. The family of models best suited for this task is the clustering. In particular, the considered methods are the centroid-based clustering or k-means and the hierarchical clustering.
Clustering aim is to find homogeneous subgroups in the data where observations inside the same group are similar while being different enough from observations in other subgroups. It can be either performed on dataset observations based on features similarity or the other way round.

K-MEANS CLUSTERING

The k-means clustering, also known as centroid-based clustering, finds K non-overlapping different clusters, where K is the hyperparameter chosen a priori. The ideal number of clusters depends on the result obtained and its usefulness for interpreting the data. It is domain specific and it depends on the goal of the analysis.
The model aim is to minimize the within-cluster variation, which measures how much the observations in the same cluster are different from each other. This difference is computed using distances when the dataset contains only continuous variables and similarity (or dissimilarity) measures when there are only categorical variables. In the case of mixed type data, the solution usually adopted is that of using the Gower index, which makes continuous and categorical variables comparable.
Since the Country Health Indicator dataset variables are all continuous, distance is used. There are different measures for distance that can be applied. The most popular ones are the Manhattan distance or City Block and the Euclidian distance, which general form is the Minkowski distance.
Given two groups $u_i = [x_{i1}, x_{i2}, \cdots, x_{ip}]$ and $u_j = [x_{j1}, x_{j2}, \cdots, x_{jp}]$, the general distance formula is

$$d_{ij} = \left[\sum_{s=1}^{p} |x_{is} - x_{js}|^k \right]^{\frac{1}{k}}$$

Specifically, the Manhattan distance formula is

$$d_{ij} = \sum_{s=1}^{p} |x_{is} - x_{js}|$$

While the Euclidian is calculated as follows

$$d_{ij} = \sqrt{\sum_{s=1}^{p} (x_{is} - x_{js})^2}$$

For brevity, only the Euclidian distance is used. Moreover, Euclidian and Manhattan distance usually provide a similar result.

The optimization problem to be solved when applying the clustering is that of minimizing the sum of the pairwise squared Euclidian distances between the observations in the same cluster, divided by the total observations.

$$\min_{C_1,\cdots,C_K}\left\{\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}\left(x_{ij}-x_{i'j}\right)^2\right\}$$

Where $|C_k|$ is the number of observations in the k-th cluster.

The algorithm starts by randomly assigning each observation to a cluster. Then the centroid, which is the arithmetic mean of the points in a cluster, is computed for every cluster. Each observation is then reassigned to the closest centroid, according to the Euclidian distance, and the centroids are recalculated again.
The iterations continue up to the point where the clusters assignment isn't changing anymore.
Usually, the result of clustering is a local optimum rather than a global one, depending on the initial assignment. It is better, therefore, to perform clustering several times using different starting point and then to select the one with the lowest within-cluster sum of squares

In order to choose the optimal number of clusters before running the clustering model, it is useful to look at the within deviance value and how it decreases as more clusters are added.

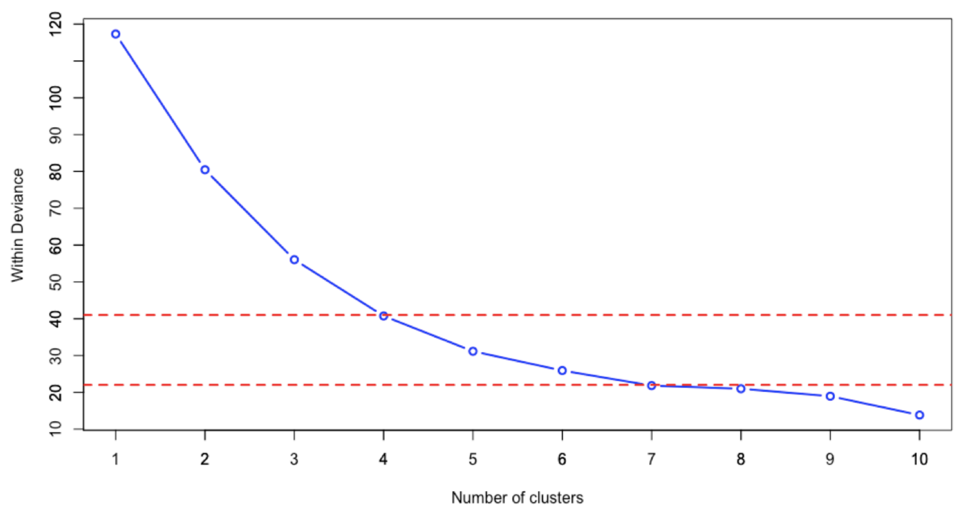

Figure 10: within deviace corresponding to the number of clusters

From the above plot, the optimal number of clusters is probably between four and seven. The first 4 clusters are responsible for the major part of the reduction in the within deviance, while after the 7th cluster, the decrease becomes quite small.
The two values chosen for clustering are four and seven, and the final models will be assessed based on their interpretability. A smaller model may be more appropriate for understanding the similarity between countries rather than a model with many clusters.

The models with four and seven clusters are fitted, each of them 50 times starting from different random starting points. The model having the smallest within variance is selected.

Afterwards, the points and the corresponding clusters are represented in a 3-dimensional plot

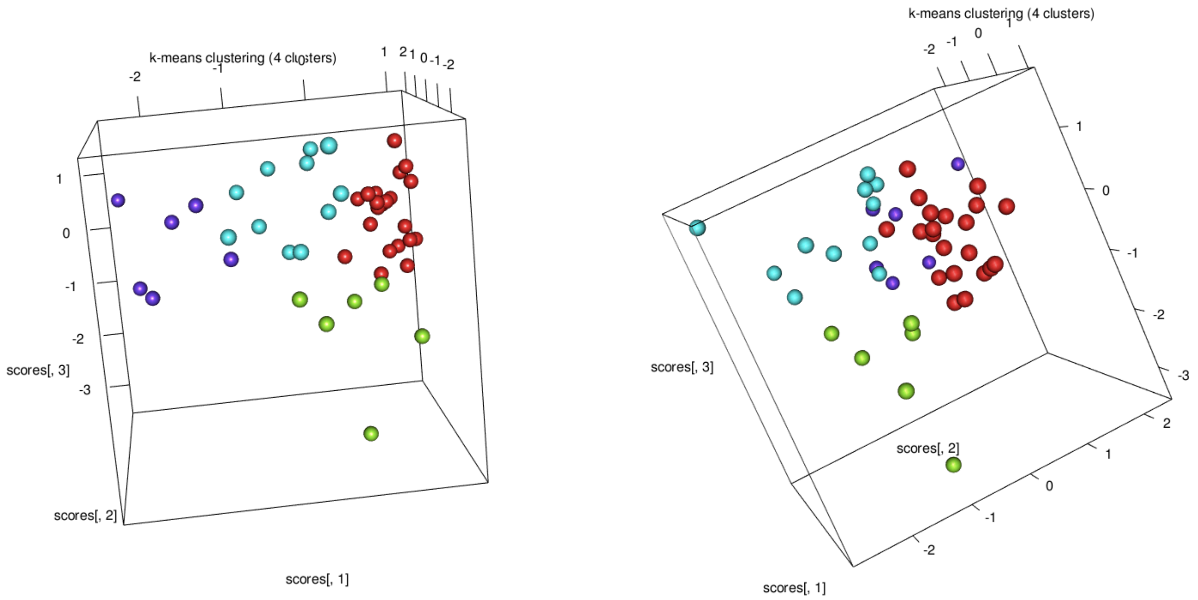

Figure 11: 3-dimensional plot of k-means clustering with 4 clusters

The observations seem properly separated in the clusters. There are no clusters made by a single outlier

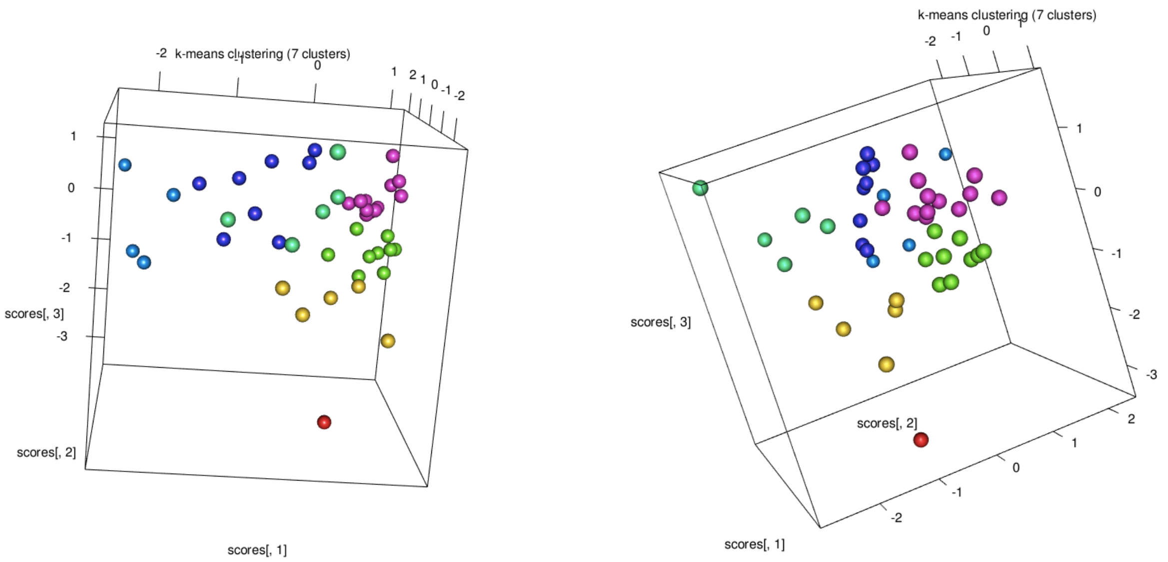

Figure 12: 3-dimensional plot of k-means clustering with 7 clusters

Even in the case of seven clusters the observations are clearly separated, however one cluster corresponds to an outlier.

The observations assigned to each cluster will be now compared.
Regarding the case with four clusters, the data points has been assigned as follows:

- *Cluster 1*: Australia, Austria, Belgium, Cyprus, Denmark, Finland, France, Germany, Greece, Iceland, Israel, Italy, Malta, Netherlands, New Zealand, Russia, Spain, Switzerland, US, United Arab Emirates, UK
- *Cluster 2*: China, Japan, South Korea, Morocco, Turkey, Vietnam
- *Cluster 3*: Argentina, Bolivia, Brazil, Cambodia, Cuba, Dominican Republic, Malaysia, Mexico, South Africa, Sri Lanka, Thailand
- *Cluster 4:* Cameroon, Central African Republic, Haiti, India, Kenya, Senegal

As for the k-mean clustering with seven clusters, the result is:

- *Cluster 1*: China
- *Cluster 2*: Japan, South Korea, Morocco, Turkey, Vietnam
- *Cluster 3*: Belgium, Cyprus, Greece, Iceland, Israel, Italy, Malta, Spain, United Arab Emirates
- *Cluster 4*: Cambodia, Cuba, Mexico, Sri Lanka, Thailand
- *Cluster 5*: Cameroon, Central African Republic, Kenya, Senegal
- *Cluster 6*: Argentina, Bolivia, Brazil, Dominican Republic, Haiti, India, Malaysia, South Africa
- *Cluster 7*: Austria, Denmark, Finland, France, Germany, Netherlands, New Zealand, Russia, Switzerland, US, UK

The clustering assignment in the second model seems more difficult to interpret than the one of the first one. In the four clusters model, the subgroups can be summarized as:

- *Cluster 1*: wealthy Western countries, Middle Eastern country (United Arab Emirates, Israel) and Russia
- *Cluster 2*: wealthy Asian countries expect for Vietnam and middle-Eastern countries (Morocco, Turkey)
- *Cluster 3:* Latin American countries, non-wealthy Asian countries (Cambodia, Malaysia, Sri Lanka, Thailand), plus one African country (South Africa)
- *Cluster 4:* African countries, plus one Asian country (India) and one Latin American country (Haiti)

It is noteworthy that Western countries and wealthy Asian countries, which number is small, are similar enough to be included in their own cluster. On the other hand, there is intergroup dissimilarity between Middle Eastern countries, non-wealthy Asian countries, African countries and Latin American countries. This has implied their assignment to different clusters.

For completeness, the mean of each cluster for every principal component can be plotted in order to compare them. If they are different enough, the clusters have been properly separated by the algorithm.

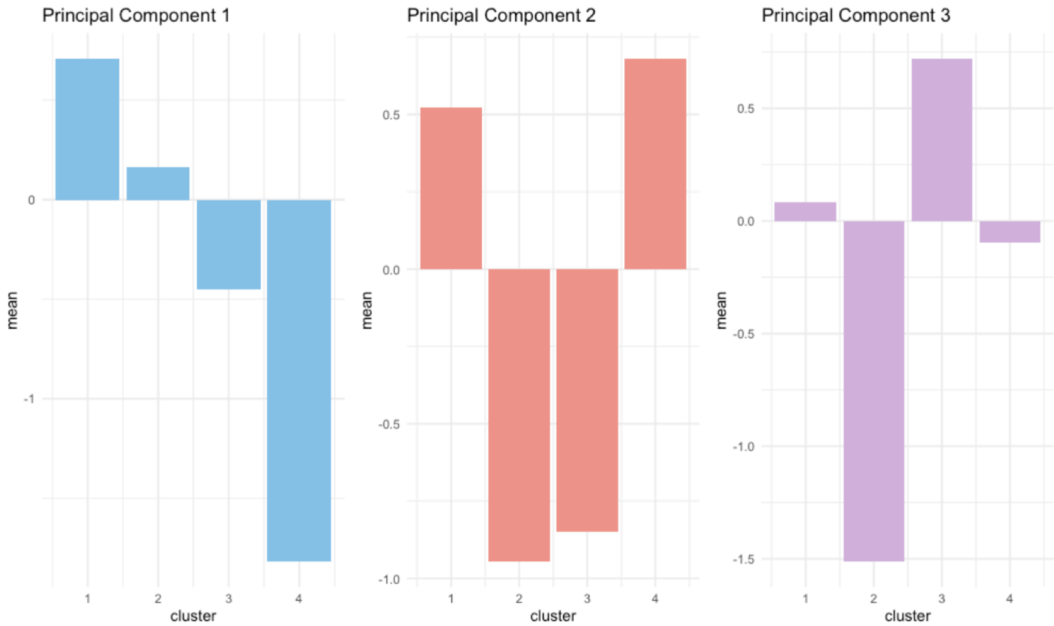

Figure 13: mean of clusters for every principal component for k-means clustering

Each cluster presents a different enough mean for the different principal components, meaning they have been properly selected.

The main drawback of k-means clustering is that the number of clusters must be specified in advance, which is not an easy task.

HIERARCHICAL CLUSTERING

Hierarchical clustering is a clustering method overcoming the problem of choosing the number of clusters before running the algorithm. The optimal number of clusters can be decided based on the so-called dendogram, which is a tree-based graphical representation showing how are the observations aggregated in clusters.
There are two different types of hierarchical clustering. The first one is the agglomerative clustering, which starts from the single observations and combines them recursively until there is only one big cluster including all of them. It is a bottom-up approach. The other method is called divisive clustering and has a top-down approach, meaning that it begins with a single cluster containing the whole dataset and then splits it until each data point belongs to a singleton cluster.

The agglomerative clustering is the most common type of hierarchical clustering, so it’s the one that will be used for this analysis.
This type of clustering faces an additional issue with respect to k-mean clustering when dealing with the concept of distance or dissimilarity. In the case of k-mean clustering the dissimilarity is calculated between pairs of observations, while as for hierarchical clustering, the dissimilarity measure must be extended to the comparison of clusters containing multiple observations. That’s because after having chosen a dissimilarity measure, in this case the Euclidian distance, the method through which the algorithm groups clusters have to be defined. This is done by the so-called Linkage methods. There are five main Linkage methods, which are the Single Linkage, Complete Linkage, Average Linkage, Centroid Linkage and Ward’s Linkage.
The Single Linkage defines the distance as the minimal intercluster dissimilarity, meaning that it merges clusters which members have the minimum possible distance. Its main

drawback is that of creating unbalanced clusters by adding one observation at a time. On the other hand, the Complete Linkage uses the maximal intercluster dissimilarity, which merges clusters with the closest maximum distance. It provides more compact clusters, since it favors intercluster's homogeneity.
Another approach is taken by the Average Linkage, which is based on the mean intercluster dissimilarity. The method computes the pairwise distances between all objects in the two different clusters and merges clusters with the lowest average distance.
Similarly, the Centroid Linkage, as suggested by its name, calculates the centroid for each cluster and groups the ones with the lowest centroid distances.
Finally, the Ward's Linkage takes a different approach, by merging clusters which increase the least their error sum of square when merged.
All the methods will be applied and compared

The next crucial step is interpreting the dendogram. Each leaf in the bottom part corresponds to an observation. As they move upwards, they start merging into branches, corresponding to clusters with many data points. This means that a cluster can merge with a single observation or with another cluster. The height at which observations are joined indicates how different they are, meaning that if they merge at the bottom of the tree they are quite similar, vice versa if they fuse at the top of it.
This fact implies that it is the vertical axis defining whether the observations are similar or not rather than the horizontal one
The dendogram can then be horizontally cut at the height corresponding to the number of clusters that seems more appropriate. However, it is not necessarily clear what is the optimal number of clusters, which depends on a subjective choice made by the analyst.

Before applying the hierarchical clustering, it is better to standardize the data points as with k-means clustering. Afterward, the clustering is applied using the different Linkage methods.
The Linkage that performed worse, meaning that they weren't able to properly discriminate clusters, were Single and Centroid Linkage. Their dendograms and plots are displayed below.

First, those of Single Linkage are showed.

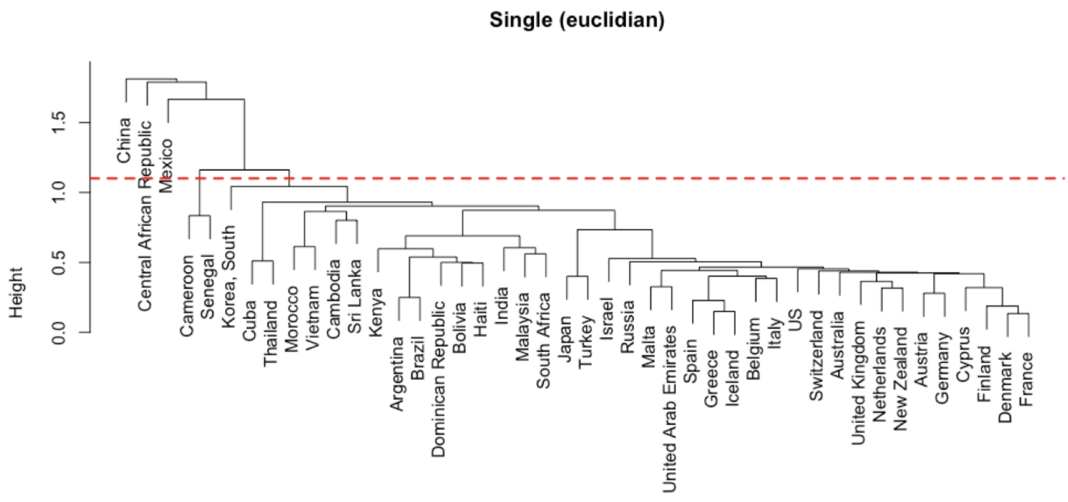

Figure 14: dendogram for Single Linkage method (5 clusters)

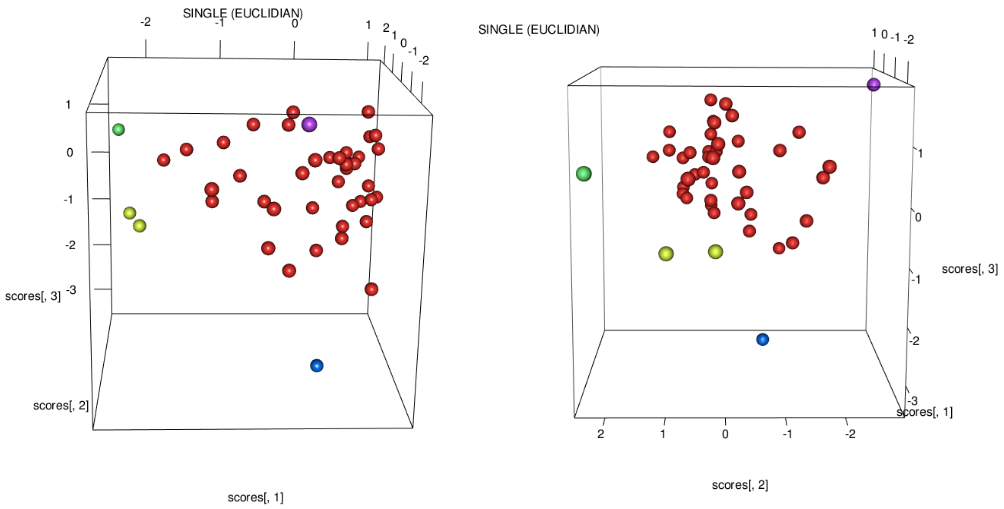

Figure 15: 3-dimensional plot of hierarchical clustering with Single Linkage (5 clusters)

Then, there are the dendogram and plot of Centroid Linkage

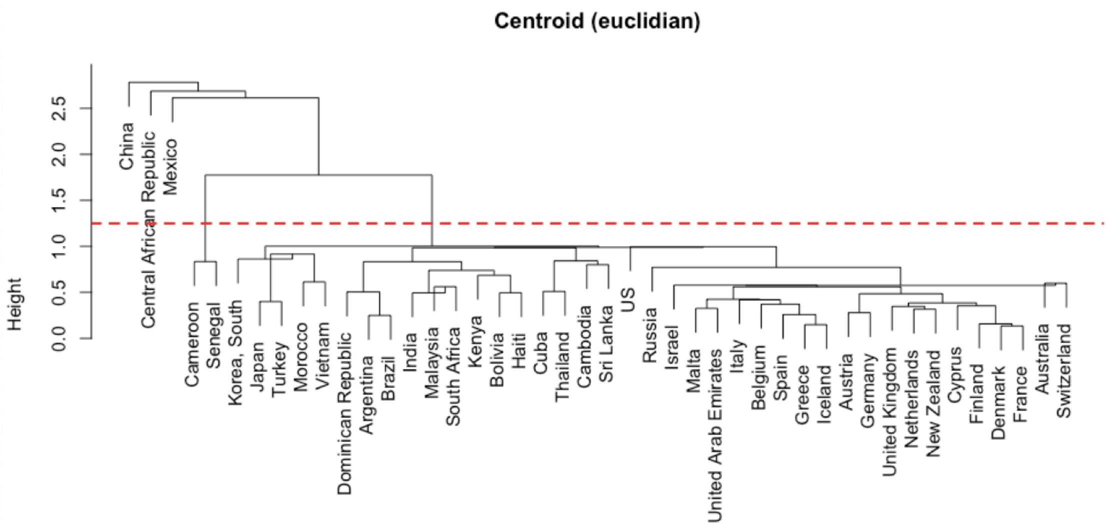

Figure 16: dendogram for Centroid Linkage method (5 clusters)

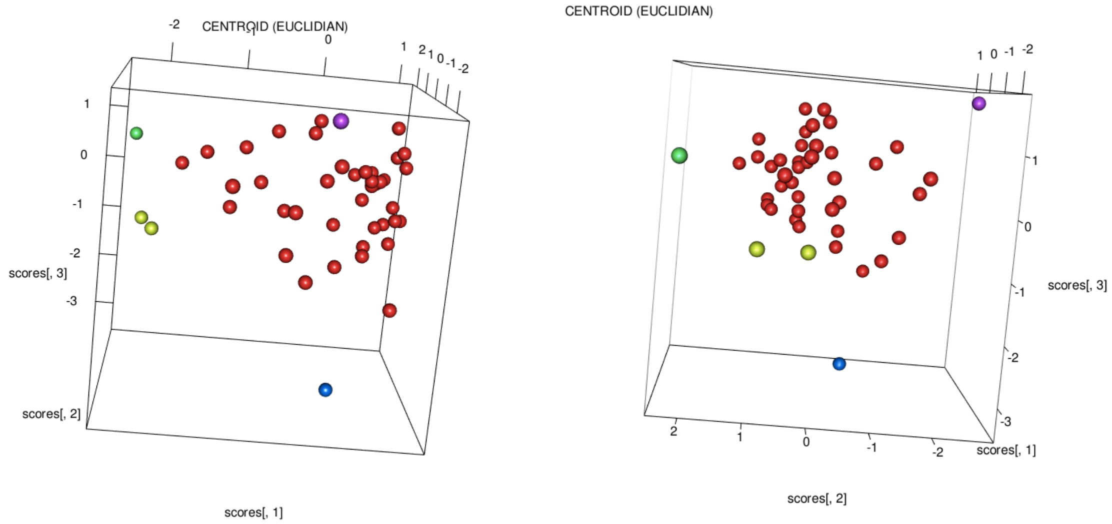

Figure 17: 3-dimensional plot of hierarchical clustering with Centroid Linkage (5 clusters)

In both cases five clusters are used and the two methods provide the same exact result. The model isn't able to properly separate clusters since three out of five clusters are outliers. They correspond to China, Mexico and Central African Republic. By looking at the dendogram, it is clear that even by increasing the number of clusters the result is not improving due to the presence of subgroups containing single observations.

The next method presented is the Average Linkage which results in a slightly clearer clusters separation, but still providing subgroups containing single outliers.
In this case two different models, one with four and the other with seven clusters are fitted. Their dendogram and plots are the following

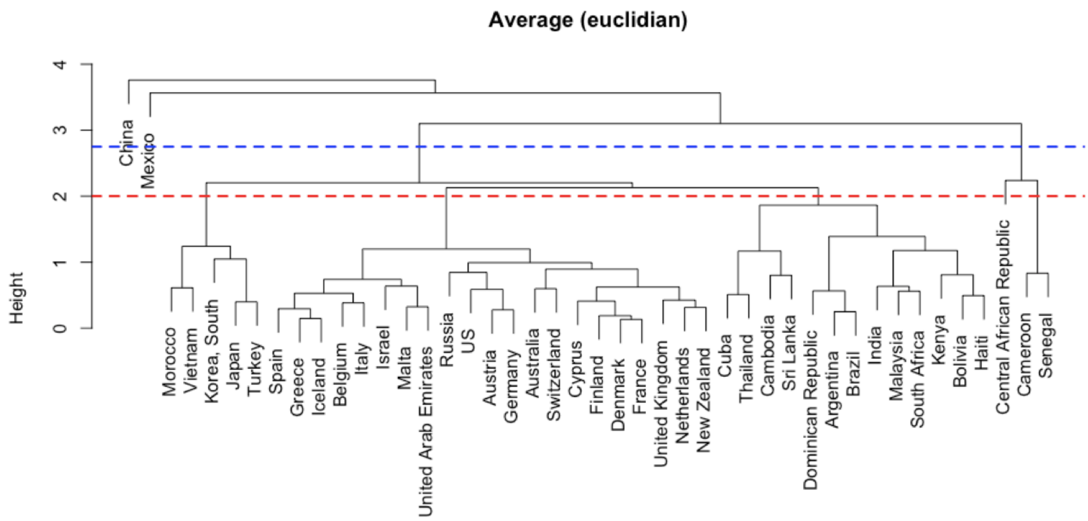

Figure 18: dendogram for Average Linkage method (4 and 7 clusters)

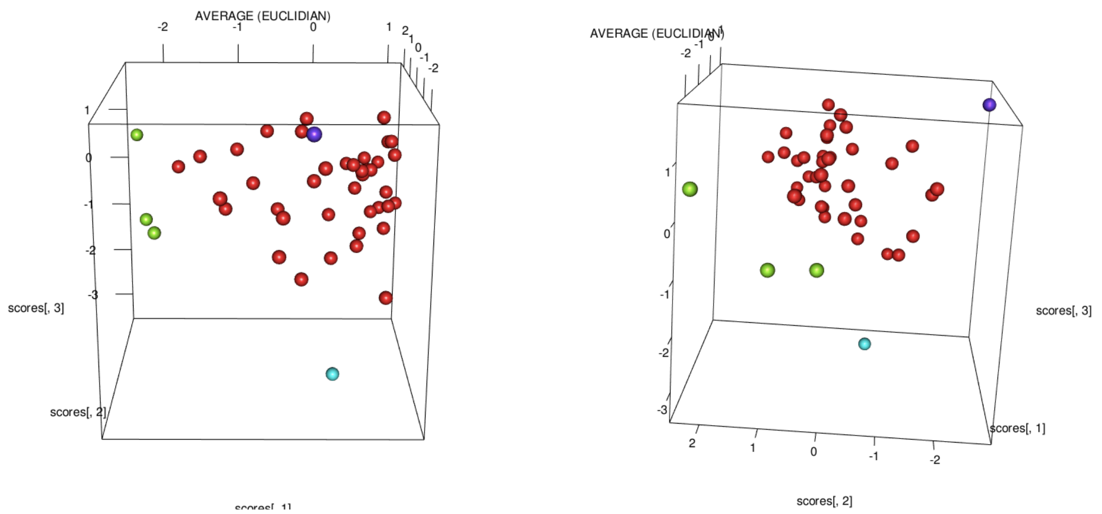

Figure 19: 3-dimensional plot of hierarchical clustering with Average Linkage (4 clusters)

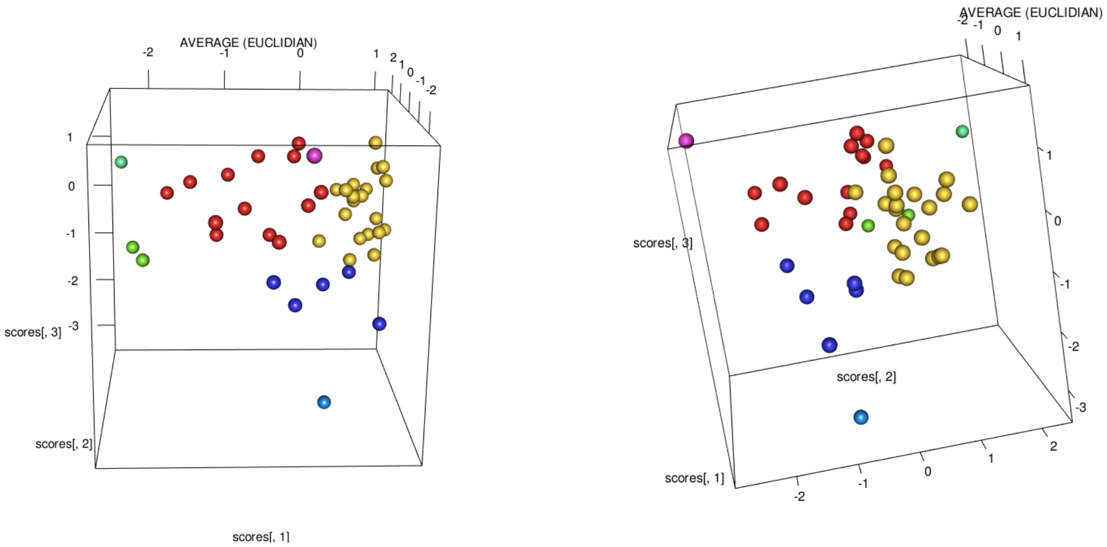

Figure 20: 3-dimensional plot of hierarchical clustering with Average Linkage (7 clusters)

The model with four clusters fails in clearly separating data, while the model with seven clusters do a better job. The problem with the Average Linkage is that it captures outliers, as the Centroid and Single Linkage did.
Since these first three methods are not suited for the dataset, their result won't be analyzed in depth.

Eventually, the methods that turned out to be the best fitted ones for these data are Complete and Ward's Linkage.
First, Complete Linkage dendogram and the plot are presented

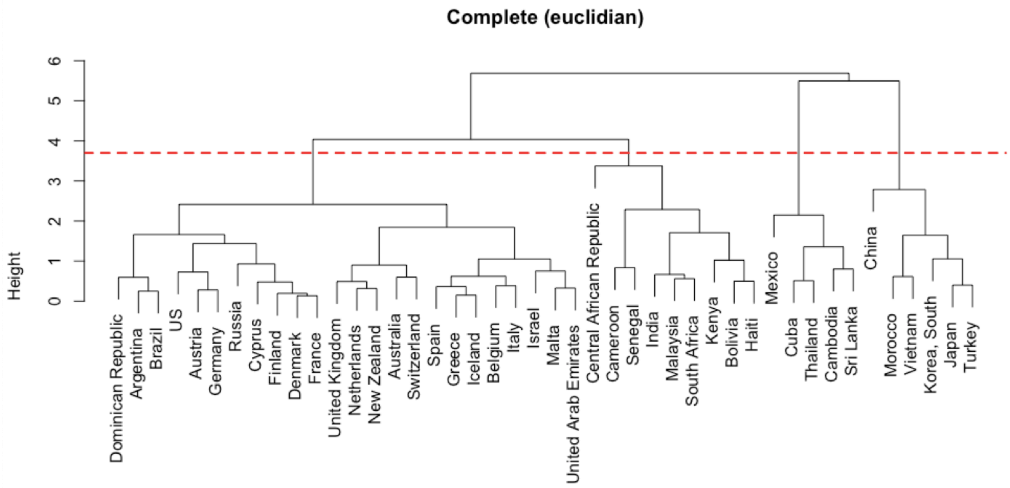

Figure 21: 3-dimensional plot of hierarchical clustering with Complete Linkage (4 clusters)

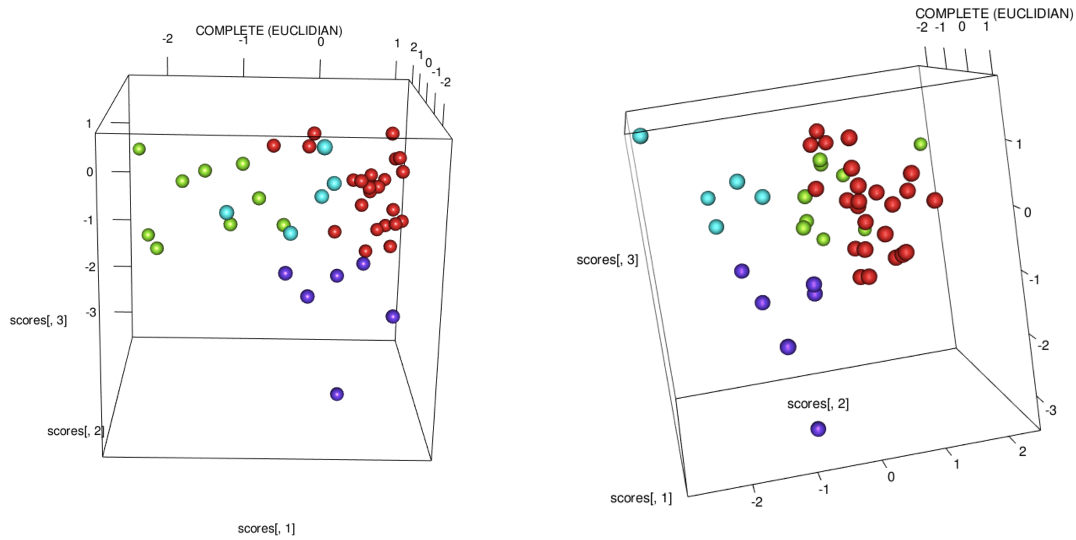

Figure 22: 3-dimensional plot of hierarchical clustering with Complete Linkage (4 clusters)

The clusters seem well-separated, presenting no outliers. They can be summarized as follows:

- *Cluster 1*: Western wealthy countries, Middle Eastern countries (Israel, the United Arab Emirates), Latin American countries (Dominican Republic, Argentina, Brazil) and Russia
- *Cluster 2*: African countries, Latin American countries (Bolivia, Haiti) and Asian countries (India, Malaysia). All of them are not wealthy countries
- *Cluster 3*: non-wealthy Asian countries (Thailand, Cambodia, Sri Lanka) and Latin American countries (Mexico, Cuba)
- *Cluster 4:* wealthiest Asian countries plus Vietnam and Middle Eastern countries (Morocco, Turkey)

The Western countries, African Countries and wealthy Asian countries are nicely divided into different clusters, whereas Latin American, Middle Eastern and other Asian countries differ much more between them. In fact, it can be seen they are spread in different clusters.

Increasing the number of clusters is not recommended in this case since some subgroups would contain single outliers.

The mean of each cluster for every principal component is plotted to understand whether they are different enough

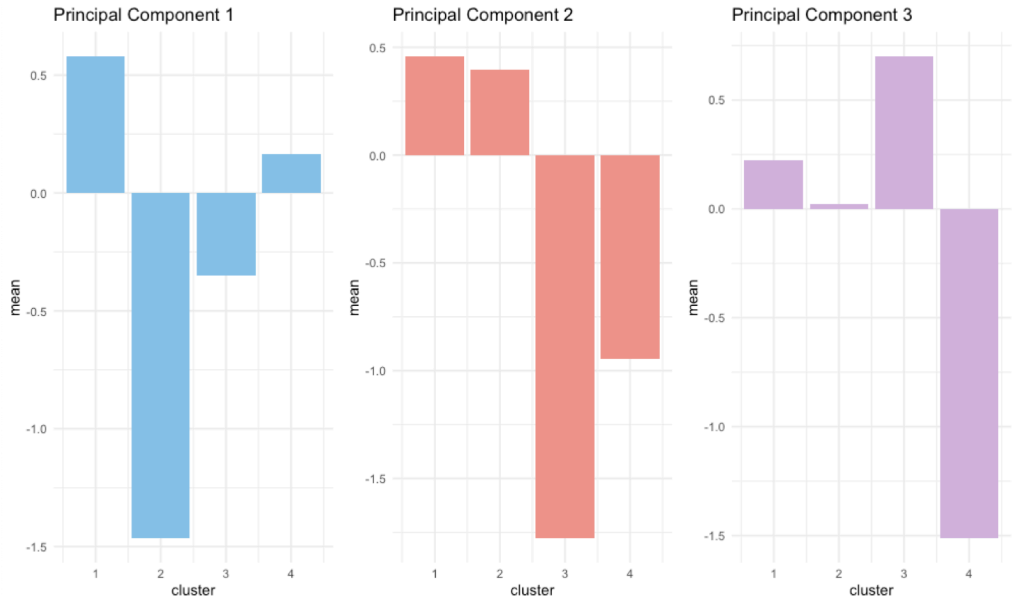

Figure 23: mean of clusters for every principal component for hierarchical clustering with complete linkage

The means of each cluster seem pretty different, so the method was able to properly separate them.

Afterwards, the dendogram and plot for Ward's Linkage is displayed below

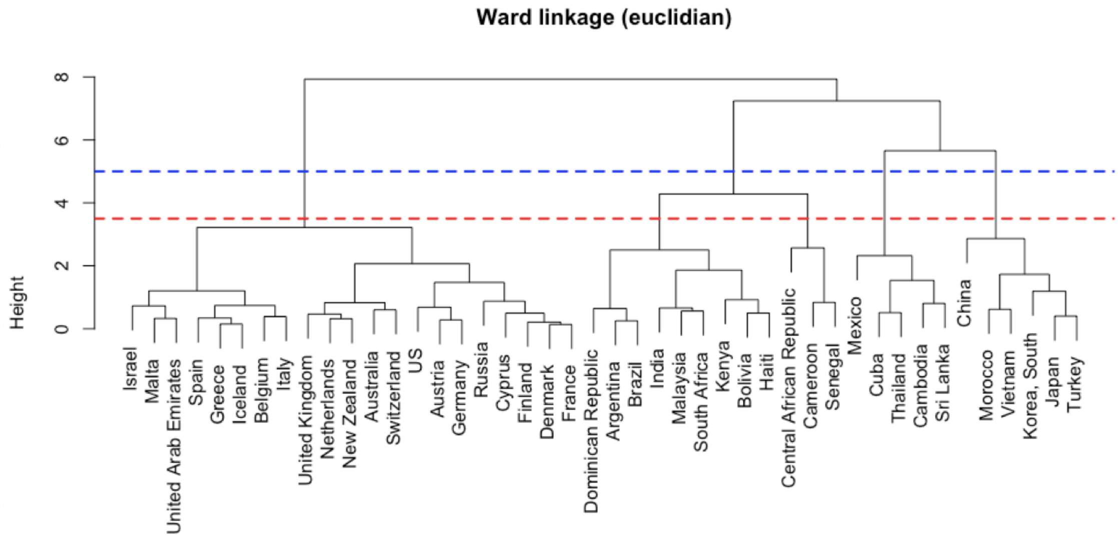

Figure 24: 3-dimensional plot of hierarchical clustering with Ward's Linkage (4 and 5 clusters)

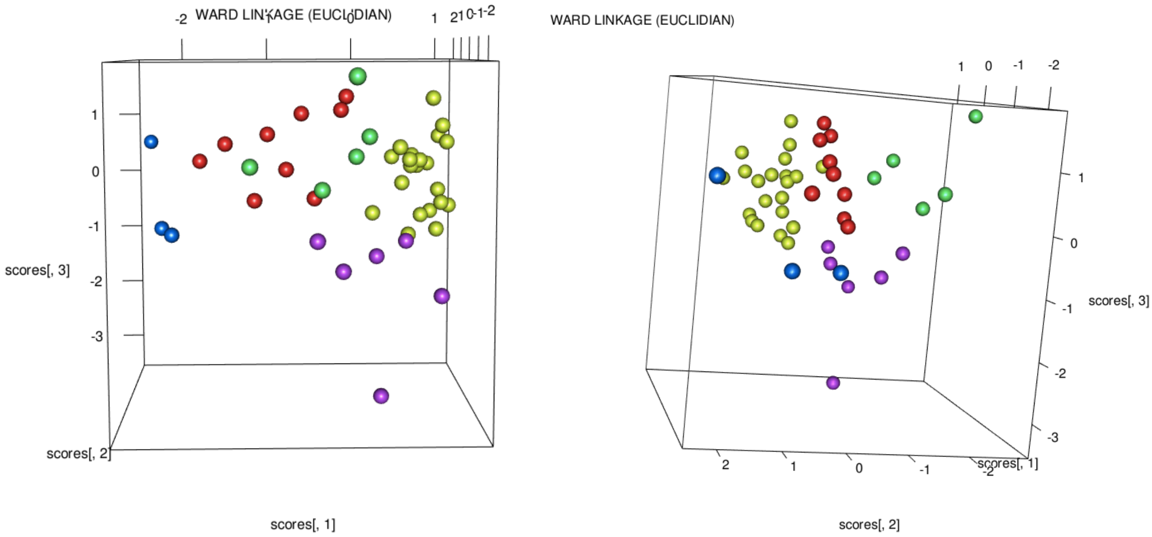

Figure 25: 3-dimensional plot of hierarchical clustering with Ward's Linkage (5 clusters)

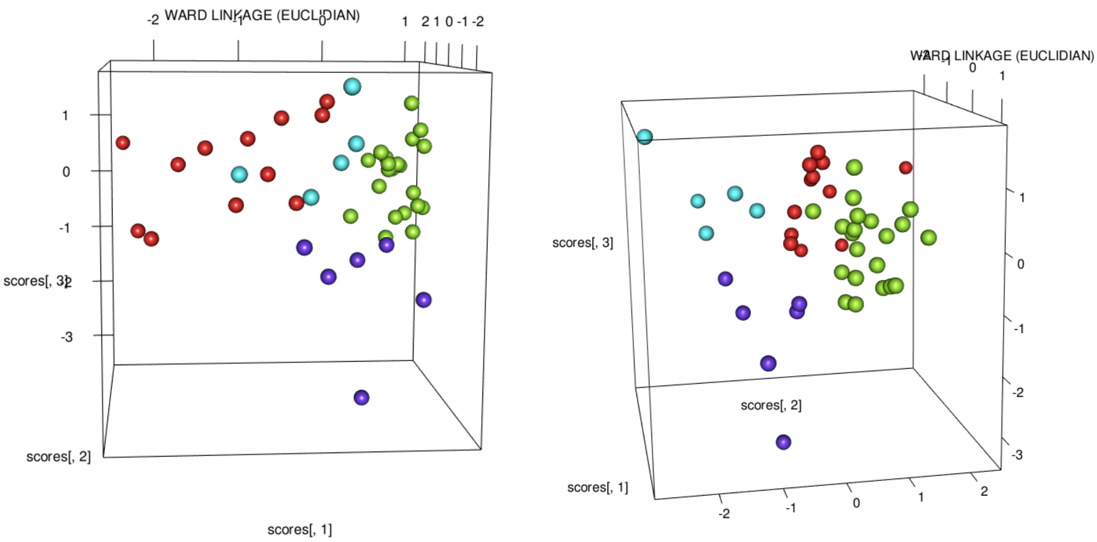

Figure 26: 3-dimensional plot of hierarchical clustering with Ward's Linkage (4 clusters)

By looking at the Ward's Linkage dendogram, it seems that the optimal number of clusters could be four or five. The clusters are properly separated here too, moreover the final result is similar. The five clusters are the following:

- *Cluster 1:* Western countries, Middle Eastern countries (United Arab Emirates, Israel) and Russia
- *Cluster 2:* Latin American countries (Dominican Republic, Argentina, Brazil, Haiti, Bolivia), non-wealthy Asian countries (India, Malaysia) and African countries (Kenya, South Africa)
- *Cluster 3:* African countries (Central African Republic, Cameroon, Senegal)
- *Cluster 4:* non-wealthy Asian countries (Thailand, Cambodia, Sri Lanka) and Latin American countries (Mexico, Cuba)
- *Cluster 5:* wealthiest Asian countries plus Vietnam and Middle Eastern countries (Morocco, Turkey)

If four clusters are used instead, cluster 2 and 3 are merged. Using more than five clusters would provide clusters containing single observations.

Once again, the mean of the clusters for the different principal component are presented graphically. first, the four clusters case is considered and then the five case one.

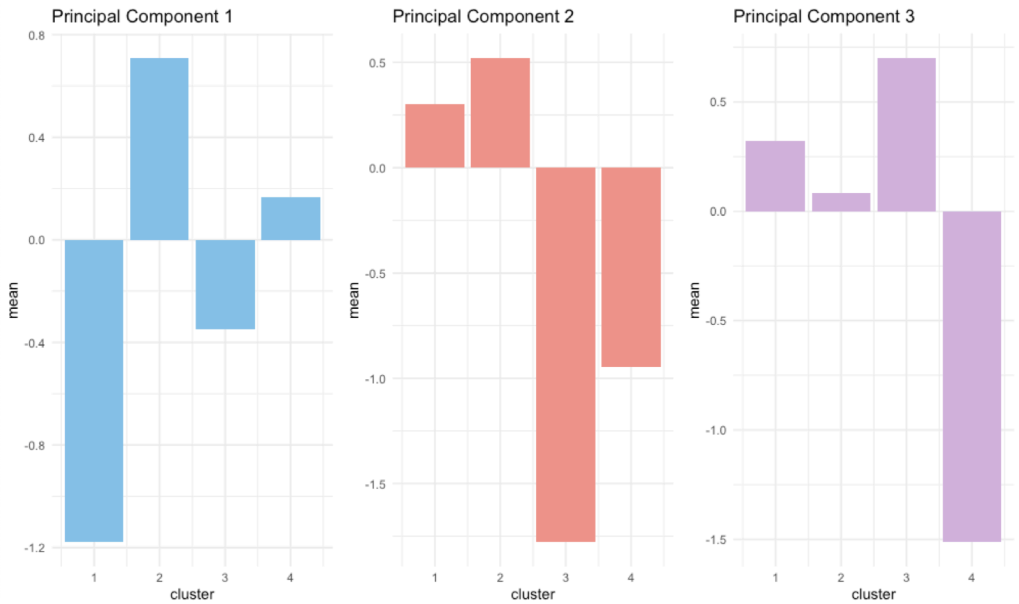

Figure 27: mean of clusters for every principal component for hierarchical clustering with Ward's linkage (4 clusters)

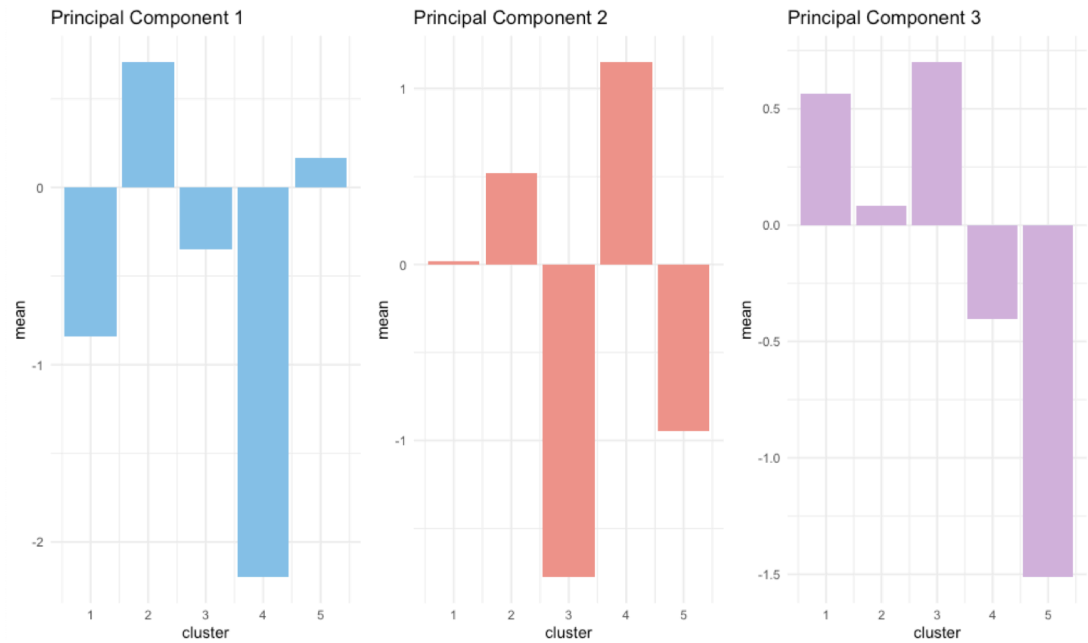

Figure 28: mean of clusters for every principal component for hierarchical clustering with Ward's linkage (5 clusters)

The clusters seem different enough by looking at their mean in both cases. The five clusters case is able to better discriminate the differences of the clusters for the second principal component.

The two methods' two last clusters are exactly the same. Further, Western countries are all in the same subgroup, whereas Latin American countries, Middle Eastern countries, African countries and non-wealthy Asian countries present greater dissimilarities.
The main difference between the cluster's assignment made by Complete and Ward's Linkage is that in the former all the African countries belongs to the same cluster, whereas in the latter they are split in two clusters. This depends only on the number of chosen clusters, since by cutting the Ward's Linkage dendogram to obtain only four clusters, the African countries are included in the same cluster.

The result of the hierarchical clustering is similar to the one obtained using the k-means clustering. They differ by the fact that k-means clustering assigns African countries to different subgroups in the four clusters model. Moreover, Latin American countries are separated so that most of them belongs to the same cluster and only one to a different one, differently from the hierarchical clustering.

CONCLUSION

The conclusion drawn from the above analysis is that the unsupervised learning techniques applied have been able to properly synthetize data.
Starting from 15 continuous variables, principal component analysis allowed to summarize them using only 3 variables which explain about 74% of total data variance. This step is fundamental in order to represent data points in a 3-dimensional space.
After having obtained the necessary number of variables, clustering has been used to find homogenous countries subgroups. First k-means clustering has been applied and then hierarchical clustering. Regarding the former, the optimal number of clusters found is 4. Concerning hierarchical clustering, five different Linkage methods have been tried and the only ones that gave satisfactory results were Complete and Ward's Linkage. Here too, 4 clusters seem to be able to properly separate data points while maintaining interpretability. Further, Ward's Linage delivers a useful result also when 5 clusters are used.

Since the outcome of both k-means clustering and hierarchical clustering using the Complete and Ward's Linkage are similar and moreover, by looking at their scatterplot the clusters are well separated with no outliers, the models are all equally valid. They are all able to separate data points so that the resulting subgroups are quite easily interpretable.

The most relevant findings can be summed up as:

1. Principal component analysis was able to synthetize 15 variables using 3 features, which explain 74% of total data variance. This means that the dataset dimensionality has been reduced to one fifth of the original dimension;
2. The clustering methods that delivered the best results in terms of clear data points separation and interpretability are k-means clustering and hierarchical clustering using Complete and Ward's Linkage. Further, their result is quite similar;
3. 44 countries have been properly summarized in 4 clusters with no outliers while retaining an easy interpretability;
4. Western and wealthy Eastern countries have a higher intergroup similarity based on heath indicators, according to all the considered clustering methods, than other Eastern countries, Middle Eastern, Latin American and African countries.