Statistical Learning

Report

E. Naples, D. Foronda

2023-01-20

Introduction

Members of the group

Members	Background	NIA
Emilie Naples	Mathematics	100469154
Daniel Foronda	Mathematics & Computer engineering	100482712

Dataset

For this study we chose the “Country Statistics - UNData” dataset on the different countries of the world with several indicators that cover four main fields: general information, economy, sociology and environment/infrastructure indicators.

The original file can be found at https://www.kaggle.com/datasets/sudalairajkumar/undata-country-profiles (https://www.kaggle.com/datasets/sudalairajkumar/undata-country-profiles)

Preliminary analysis

Variable description and data preparation

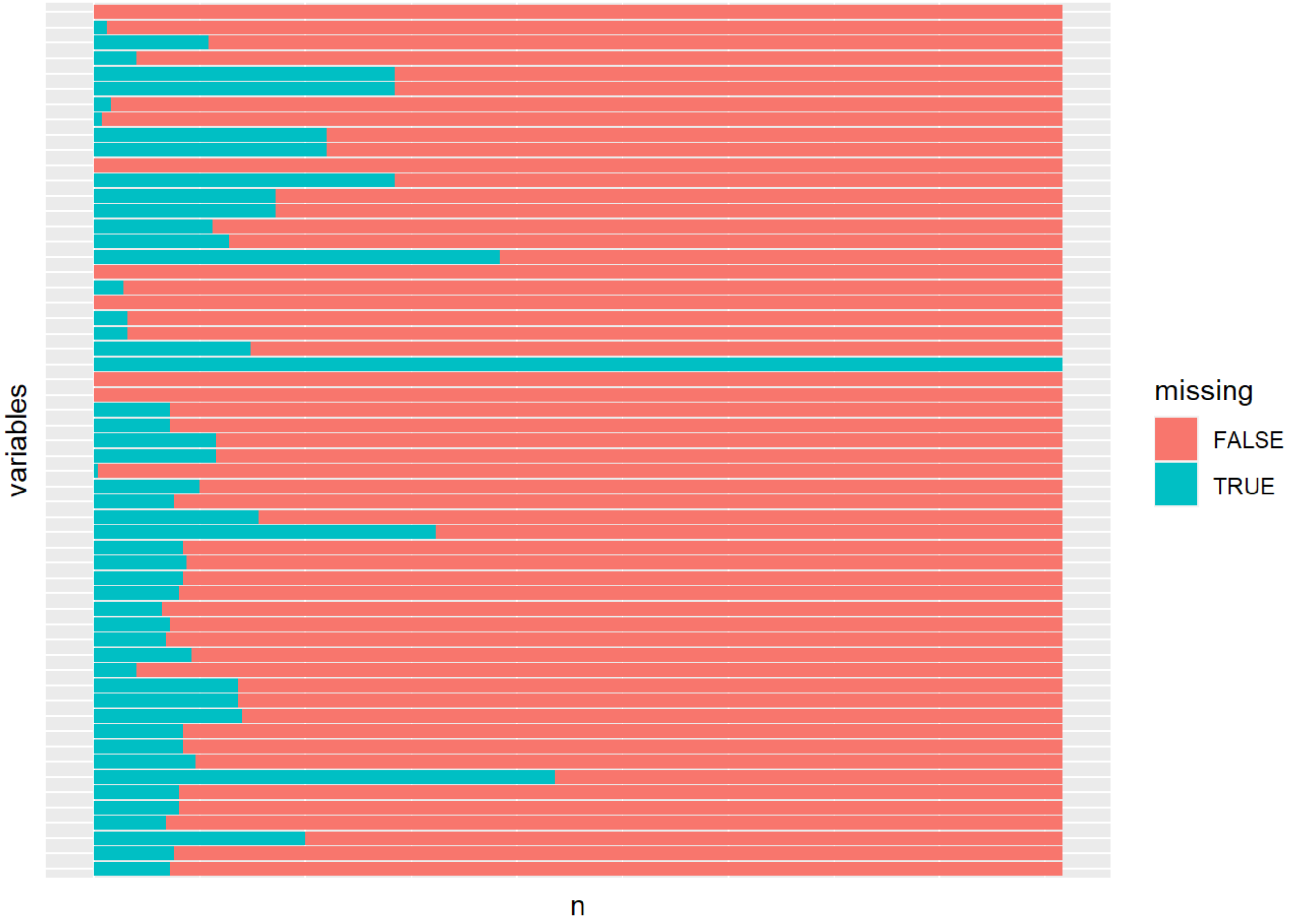
The dataset is composed of 229 rows, one for each country, and 50 columns from which the first one is the country name (unique). Apart from that we have one categorical variable (Region) and 48 numeric variables as it is shown below. We can observe that the units are included in the column names (RStudio is not showing the full name).

```
## 'data.frame':    229 obs. of  50 variables:
##  $ country                                                   : chr  "Afghanistan" "Albani
a" "Algeria" "American Samoa" ...
##  $ Region                                                    : chr  "SouthernAsia" "Southe
rnEurope" "NorthernAfrica" "Polynesia" ...
##  $ Surface.area..km2.                                        : chr  "652864" "28748" "2381
741" "199" ...
##  $ Population.in.thousands..2017.                            : int  35530 2930 41318 56 77
29784 15 102 44271 2930 ...
##  $ Population.density..per.km2..2017.                        : num  54.4 106.9 17.3 278.2
163.8 ...
##  $ Sex.ratio..m.per.100.f..2017.                             : num  106 102 102 104 102
...
##  $ GDP..Gross.domestic.product..million.current.US..         : int  20270 11541 164779 -99
2812 117955 320 1356 632343 10529 ...
##  $ GDP.growth.rate..annual....const..2005.prices.            : chr  "-2.4" "2.6" "3.8" "-9
9" ...
##  $ GDP.per.capita..current.US..                              : num  623 3984 4154 -99 3989
6 ...
##  $ Economy..Agriculture....of.GVA.                           : chr  "23.3" "22.4" "12.2"
"-99" ...
##  $ Economy..Industry....of.GVA.                              : num  23.3 26 37.3 -99 10.8
51.2 15.7 18.3 27.8 28.3 ...
##  $ Economy..Services.and.other.activity....of.GVA.           : num  53.3 51.7 50.5 -99 88.
6 42 82 79.8 66.2 52.8 ...
##  $ Employment..Agriculture....of.employed.                   : chr  "61.6" "41.4" "10.8"
"..." ...
##  $ Employment..Industry....of.employed.                      : chr  "10.0" "18.3" "34.5"
"..." ...
##  $ Employment..Services....of.employed.                      : chr  "28.5" "40.3" "54.7"
"..." ...
##  $ Unemployment....of.labour.force.                          : chr  "8.6" "15.8" "11.4"
"..." ...
##  $ Labour.force.participation..female.male.pop....           : chr  "19.3/83.6" "40.2/61.
0" "17.0/70.7" ".../..." ...
##  $ Agricultural.production.index..2004.2006.100.             : int  125 134 160 112 -99 17
5 -99 88 119 135 ...
##  $ Food.production.index..2004.2006.100.                     : int  125 134 161 112 -99 17
6 -99 88 119 135 ...
##  $ International.trade..Exports..million.US..                : chr  "1458" "1962" "29992"
"-99" ...
##  $ International.trade..Imports..million.US..                : chr  "3568" "4669" "47091"
"-99" ...
##  $ International.trade..Balance..million.US..                : chr  "-2110" "-2707" "-1709
9" "-99" ...
##  $ Balance.of.payments..current.account..million.US..        : chr  "-5121" "-1222" "-2722
9" "-99" ...
##  $ Population.growth.rate..average.annual...                 : chr  "3.2" "-0.1" "2.0" "-~
0.0" ...
##  $ Urban.population....of.total.population.                  : num  26.7 57.4 70.7 87.2 8
5.1 44.1 100 23.8 91.8 62.7 ...
##  $ Urban.population.growth.rate..average.annual...           : chr  "4.0" "2.2" "2.8" "-0.
1" ...
##  $ Fertility.rate..total..live.births.per.woman.             : chr  "5.3" "1.7" "3.0" "2.
6" ...
```

```
##  $ Life.expectancy.at.birth..females.males..years.           : chr  "63.5/61.0" "79.9/75.
6" "76.5/74.1" "77.8/71.1" ...
##  $ Population.age.distribution..0.14...60..years....         : chr  "43.2/4.1" "17.4/19.0"
"29.3/9.4" "33.3/9.0" ...
##  $ International.migrant.stock..000...of.total.pop..         : chr  "382.4/1.2" "57.6/2.0"
"242.4/0.6" "23.2/41.8" ...
##  $ Refugees.and.others.of.concern.to.UNHCR..in.thousands.    : chr  "1513.1" "8.8" "99.8"
"-99" ...
##  $ Infant.mortality.rate..per.1000.live.births               : chr  "68.6" "14.6" "27.7"
"9.6" ...
##  $ Health..Total.expenditure....of.GDP.                      : num  8.2 5.9 7.2 -99 8.1 3.
3 -99 5.5 4.8 4.5 ...
##  $ Health..Physicians..per.1000.pop..                        : chr  "0.3" "1.3" "..." "-9
9" ...
##  $ Education..Government.expenditure....of.GDP.              : chr  "3.3" "3.5" "..." "-9
9" ...
##  $ Education..Primary.gross.enrol..ratio..f.m.per.100.pop..  : chr  "91.1/131.6" "111.7/11
5.5" "112.7/119.5" "-99" ...
##  $ Education..Secondary.gross.enrol..ratio..f.m.per.100.pop..: chr  "39.7/70.7" "92.5/98.
8" "101.7/98.1" "-99" ...
##  $ Education..Tertiary.gross.enrol..ratio..f.m.per.100.pop.. : chr  "3.7/13.3" "68.1/48.7"
"45.1/28.9" "-99" ...
##  $ Seats.held.by.women.in.national.parliaments..             : num  27.7 22.9 31.6 -99 32.
1 38.2 -99 11.1 38.9 9.9 ...
##  $ Mobile.cellular.subscriptions..per.100.inhabitants.       : chr  "61.6" "106.4" "113.0"
"..." ...
##  $ Mobile.cellular.subscriptions..per.100.inhabitants..1     : chr  "8.3" "63.3" "38.2" "-
99" ...
##  $ Individuals.using.the.Internet..per.100.inhabitants.      : int  42 130 135 92 13 146 5
2 55 256 114 ...
##  $ Threatened.species..number.                               : chr  "2.1" "28.2" "0.8" "8
7.9" ...
##  $ Forested.area....of.land.area.                            : chr  "9.8/0.3" "5.7/2.0" "1
45.4/3.7" "-99" ...
##  $ CO2.emission.estimates..million.tons.tons.per.capita.     : int  63 84 5900 -99 1 3902
0 -99 3167 48 ...
##  $ Energy.production..primary..Petajoules.                   : int  5 36 55 -99 119 25 136
84 85 46 ...
##  $ Energy.supply.per.capita..Gigajoules.                     : chr  "78.2/47.0" "94.9/95.
2" "84.3/81.8" "100.0/100.0" ...
##  $ Pop..using.improved.drinking.water..urban.rural....       : chr  "45.1/27.0" "95.5/90.
2" "89.8/82.2" "62.5/62.5" ...
##  $ Pop..using.improved.sanitation.facilities..urban.rural....: chr  "21.43" "2.96" "0.05"
"-99" ...
##  $ Net.Official.Development.Assist..received....of.GNI.      : int  -99 -99 -99 -99 -99 -9
9 -99 -99 -99 -99 ...
```

The first step in dealing with the dataset is to fix some problems. For example the **data type** appears as char for many quantitative variables. There are also **bivalued columns** (they contain information from two different fields) so in these cases we must divide them into two. We also observe cells with value **-99**, which we must change to NA. Also we drop the field Energy supply since in the documentation we cannot find the meaning of its values.

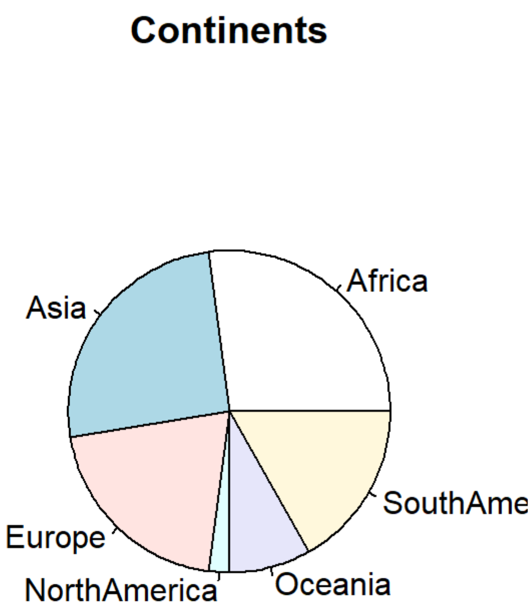
We also have a considerable number of NA's:

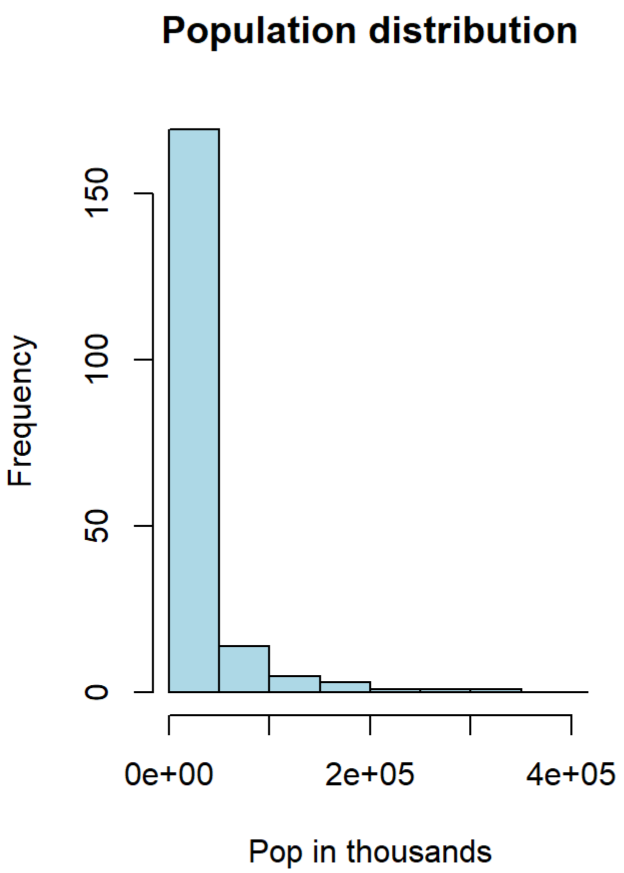


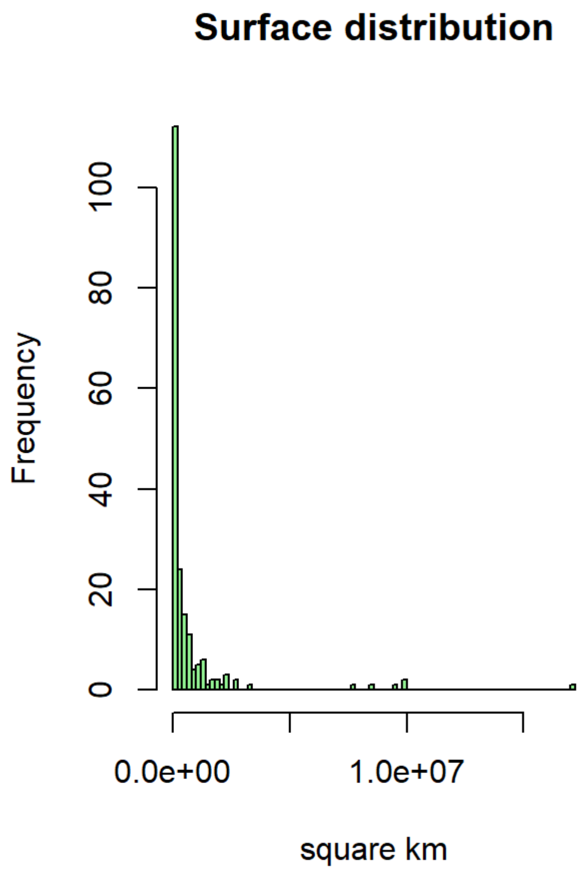
Therefore we drop rows and columns with many NA’s and also some columns that are clearly a linear combination of others (for example we have surface, population and population density when the latter could be obtained from the previous ones). However later, as we will see, we will have to remove more columns since we have many very high correlations. After dropping these columns we create the column *continent* which will be our main categorical variable.

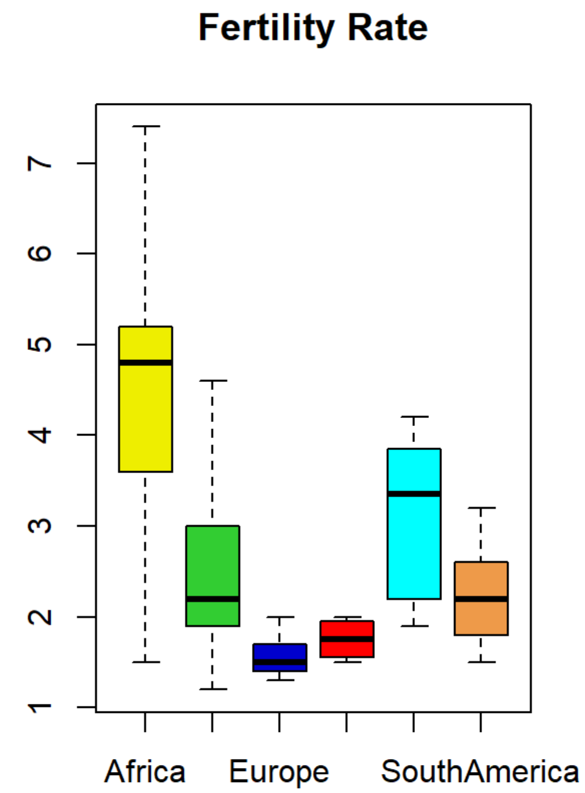
EDA

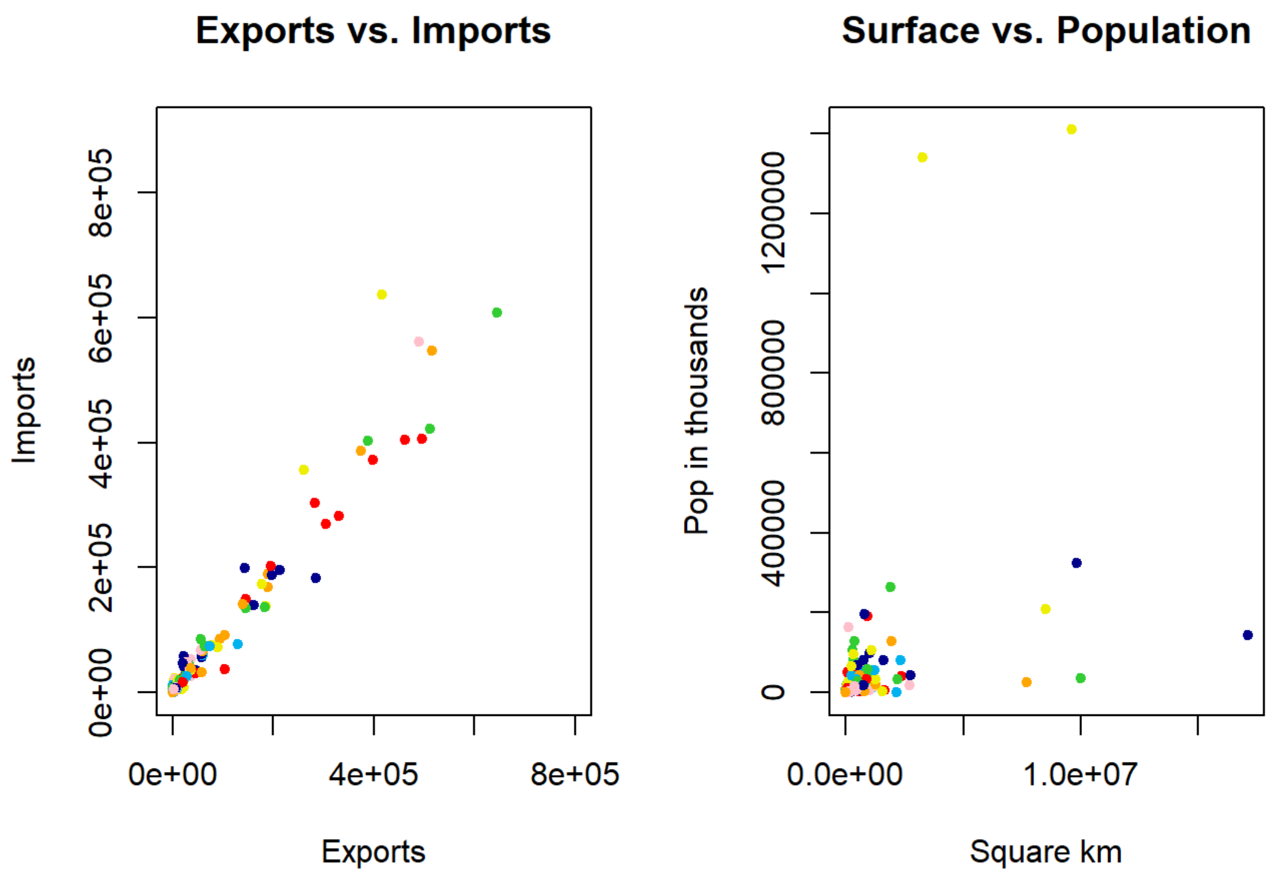
Now we will explore some characteristics of our dataset, the distribution of some variables, and in general, some information to begin to get familiar with it and draw some initial conclusions.


Continents
Africa
Asia
SouthAme
Europe
NorthAmerica
Oceania

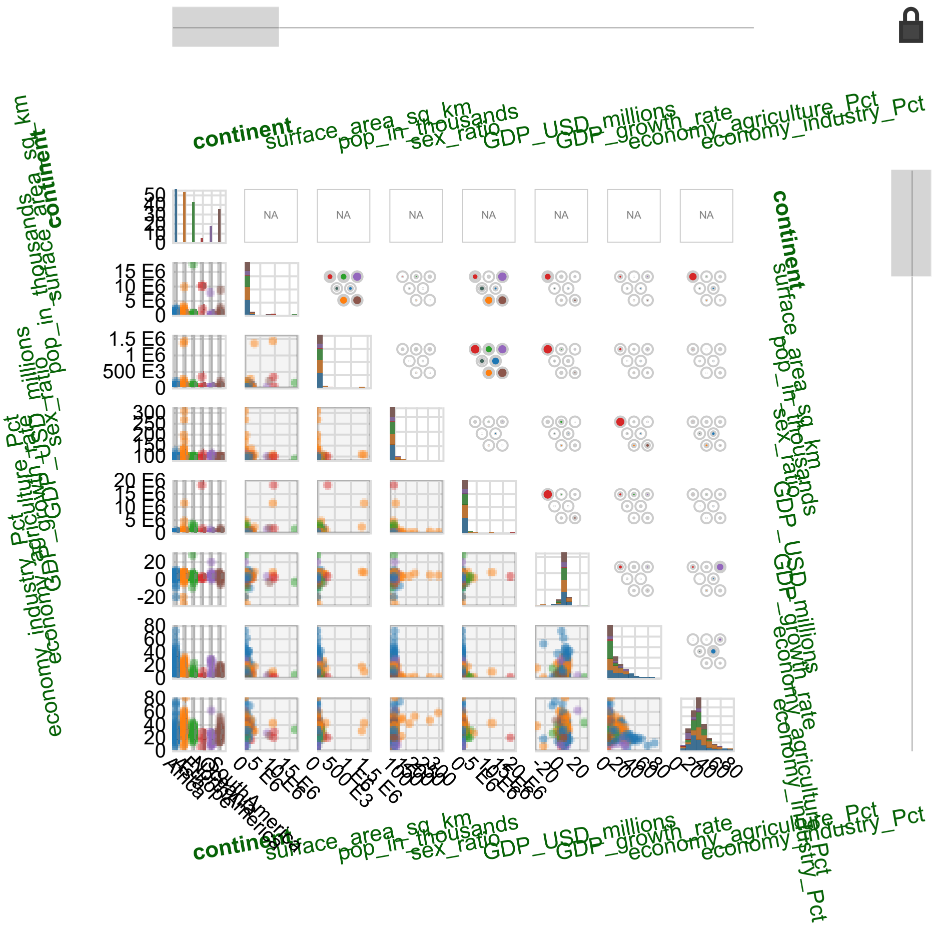

Population distribution
Frequency
0
50
100
150
0e+00
2e+05
4e+05
Pop in thousands


Surface distribution
Frequency
0
20
40
60
80
100
0.0e+00
1.0e+07
square km


Fertility Rate
1
2
3
4
5
6
7
Africa
Europe
SouthAmerica

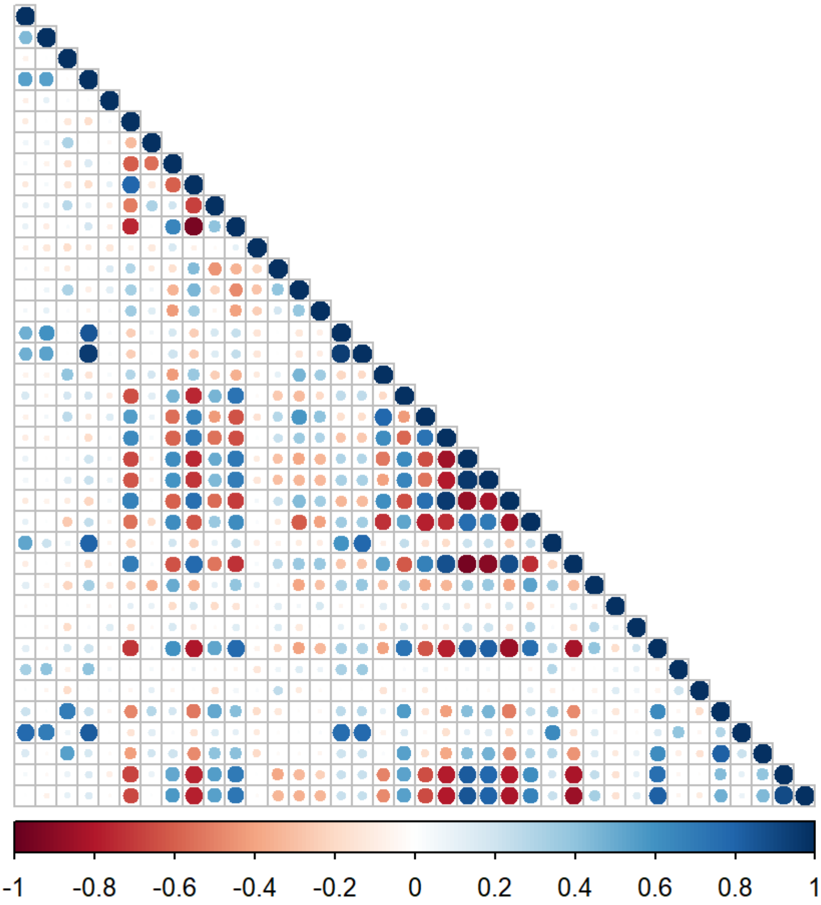


From these graphs we can draw some conclusions such as that there are still variables with **high correlations** (i.e. exports-imports), in certain variables there are very clear **outliers**, other variables are strongly **skewed.** The dataset is also **partially imbalanced** with respect to a continent: North America. Finally we can appreciate that there are some variables such as fertility rate that seem to be, on average, remarkably different for some continents.



Mean vector, covariance and correlation matrices

In this section we calculated the mean vector and the covariance and correlation matrices of the whole dataset and also for each of the continents. The most interesting result is perhaps the correlation matrix, where we can observe high correlations between variables and between groups of variables.



As a result, to synthesize part of the information obtained, we can plot the table that shows for each continent the Euclidean distance between its mean vector and the mean vector of the total set of countries.

Continent	Difference
North America	6640265.6
Africa	380348.9
Oceania	338416.6
South America	235534.0
Europe	178636.8
Asia	168924.3

North America is the 'most different' continent and Asia the most similar to the planet's average. These data are logically influenced by the number of countries that make up each one, since Asia significantly influences the total mean of the dataset while North America hardly does so, since we are weighting the mean by number of countries and not by inhabitants.

Imputation

In this section, the high level of collinearity between the different variables has begun to cause serious problems. We tried to carry out the imputation of missing values with the mice library and with others, the process not working in any case (system computationally singular). For this reason we had to progressively eliminate the columns with the highest correlation indices until reaching a threshold that would allow imputation and subsequent processes to be executed. This threshold has been $r = 0.75$, with which a total of 16 columns have been eliminated, which have been the following:

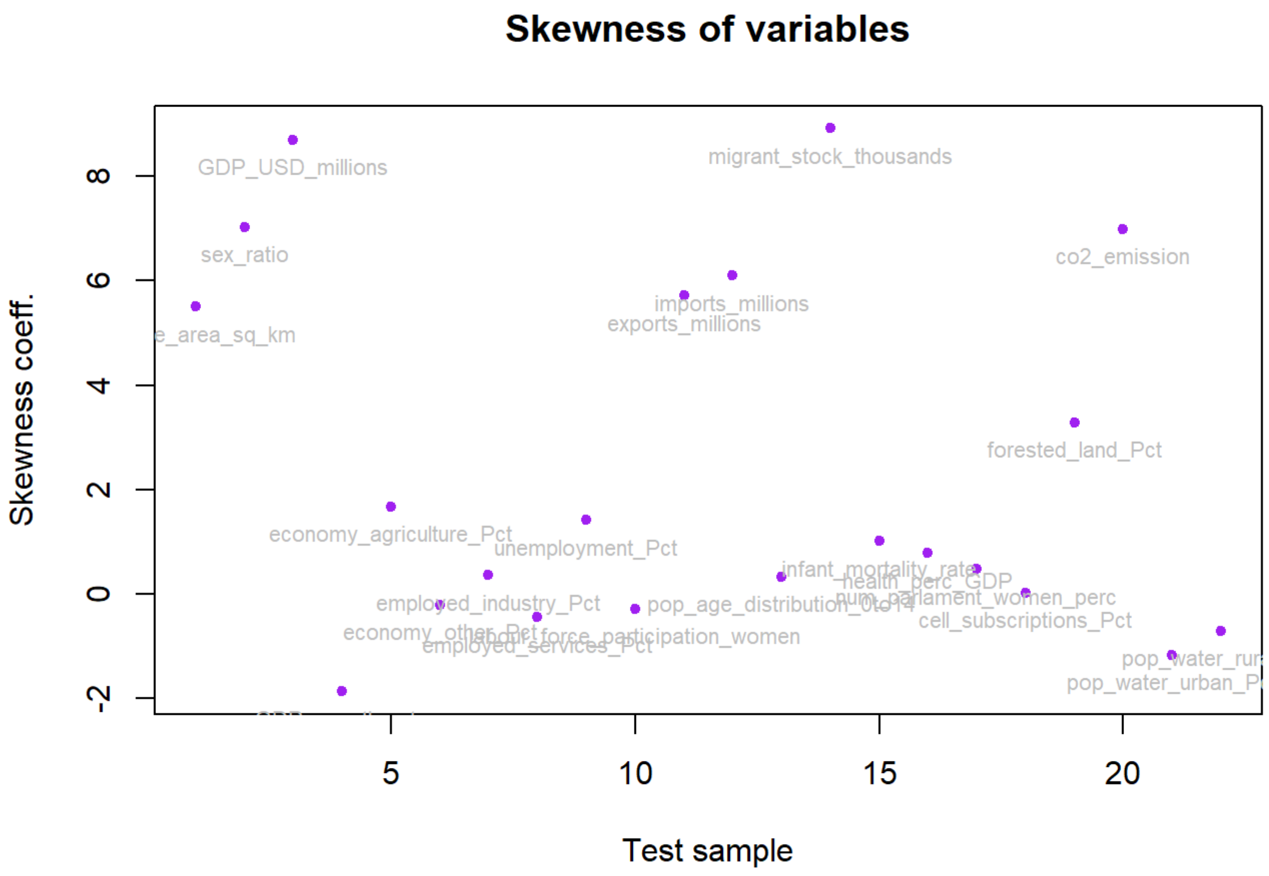
```
##  [1] "life_expectancy_women"         "primary_enroll_ratio_Pct"
##  [3] "urban_pop_growth_rate"         "economy_industry_Pct"
##  [5] "pop_age_distribution_+60"      "fertility_rate"
##  [7] "energy_prod_petajoules"        "urban_pop_Pct"
##  [9] "life_expectancy_men"           "employed_agriculture_Pct"
## [11] "pop_growth_rate"               "internet_users_Pct"
## [13] "agricultural_prod_index"       "labour_force_participation_men"
## [15] "pop_in_thousands"              "threatened_species"
```

After that we have finally been able to do imputation after the following code.

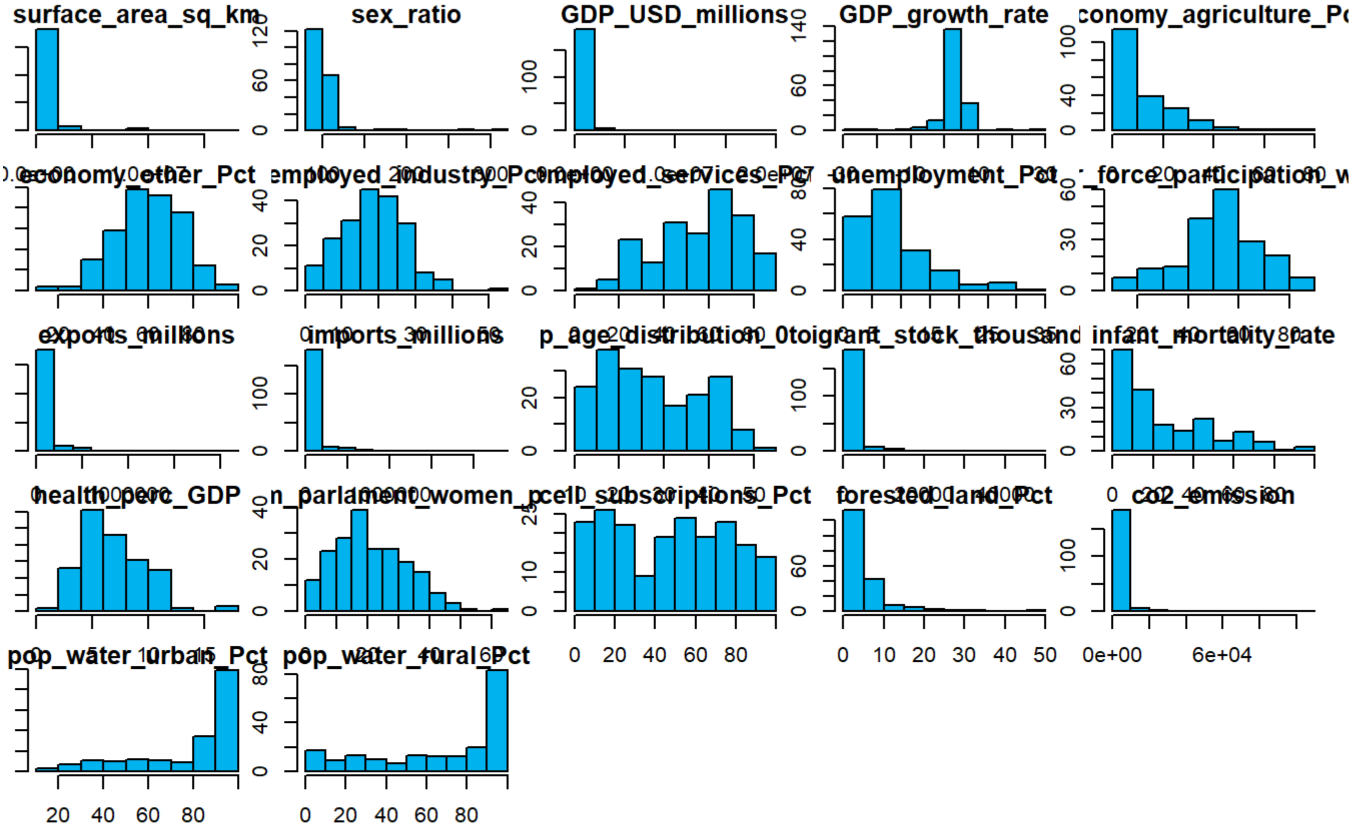
```
# We remove columns linearly related
library(caret)  # for findCorrelation
countries <- dplyr::select(countries,-c(findCorrelation(corrMatrix,cutoff = 0.75)))
covMatrix <- cov(countries[,-c(1,2)], use="complete.obs")
corrMatrix <- cor(countries[,-c(1,2)],use="complete.obs")

library(mice)
countries_numVars_imp <- mice(countries,m=5,method="pmm")
countries <- complete(countries_numVars_imp)
```

Variable Transformations

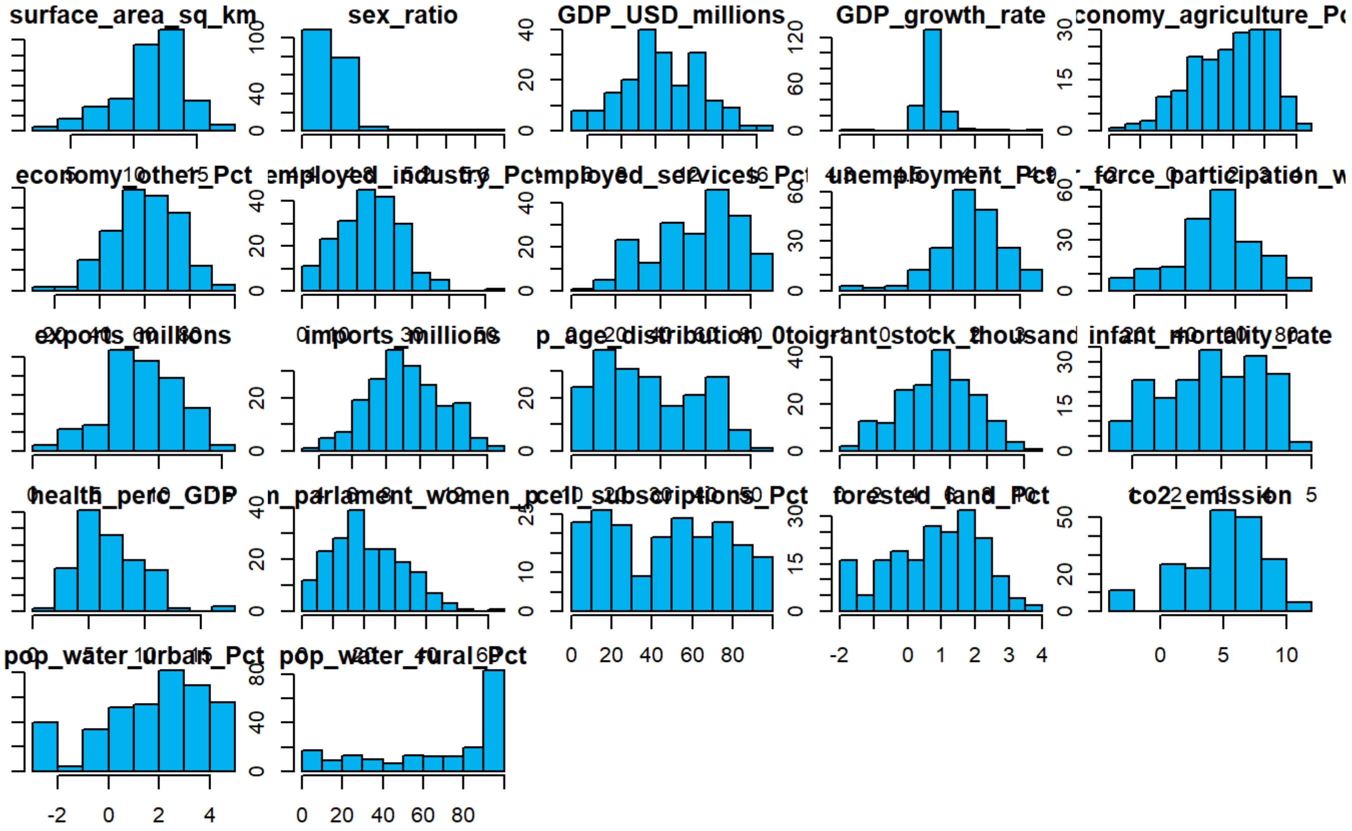
As we were able to observe in the EDA we have variables with strong skewness. We can reflect the skewness coefficient on a graph.



We can also see the histograms of the variables.

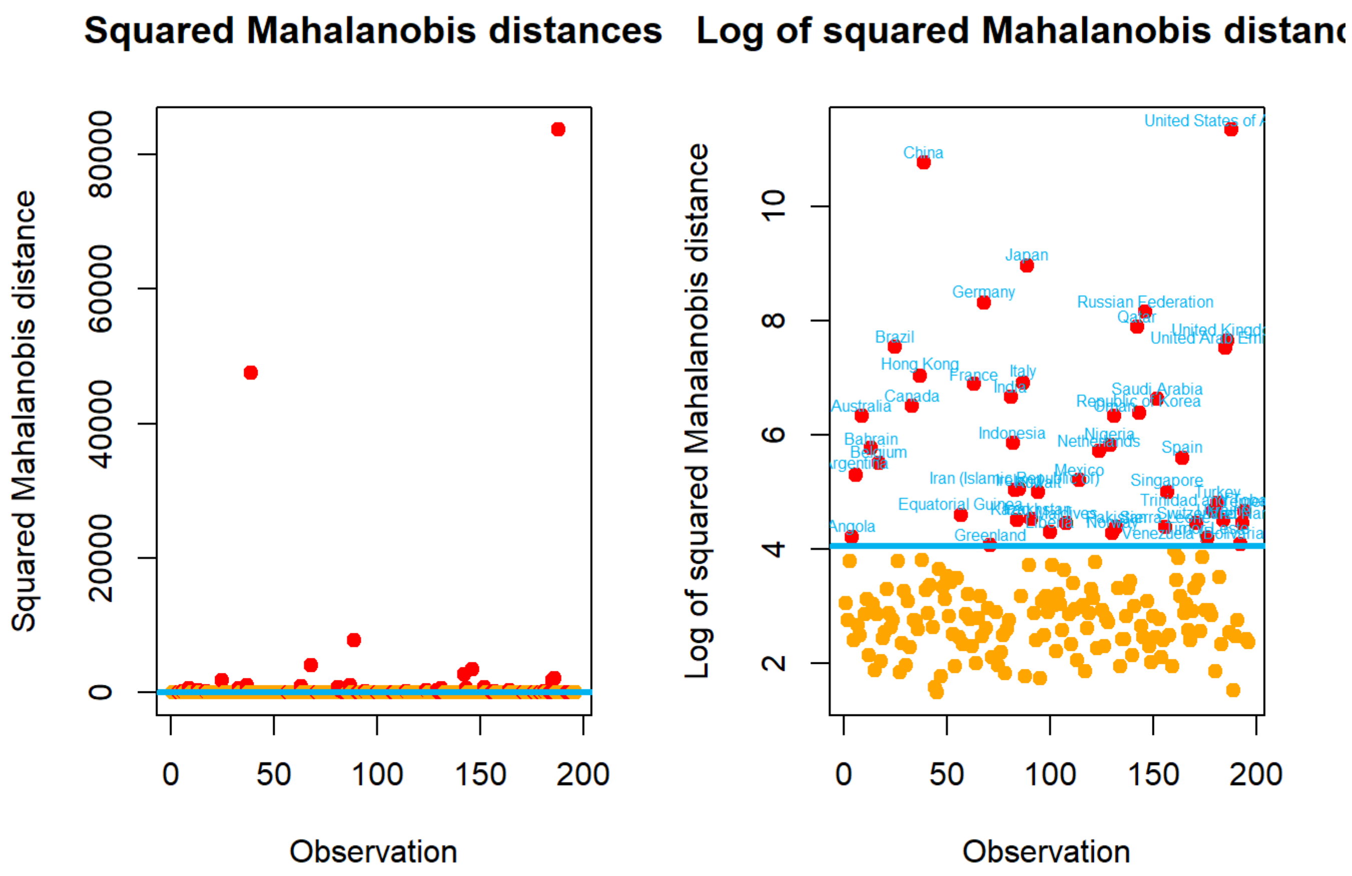


To carry out other processes later, it is convenient that the variables are as symmetric as possible, so we will transform those with higher skewness coefficients through logarithmic transformations. Specifically, we are going to transform the variables with a coefficient greater than 0.9 in absolute value. After that, the new variables remain as shown in the graph below. Note that the last variable: pop_water_rural_Pct, although it seems quite skewed, we have not transformed it since its skewness is 0.65.



Outliers

To identify outliers, the first thing we did was finding robust estimators of the mean and covariance using the MCD. Next, using the Mahalanobis distance, we made a graph of the outliers and separated them.



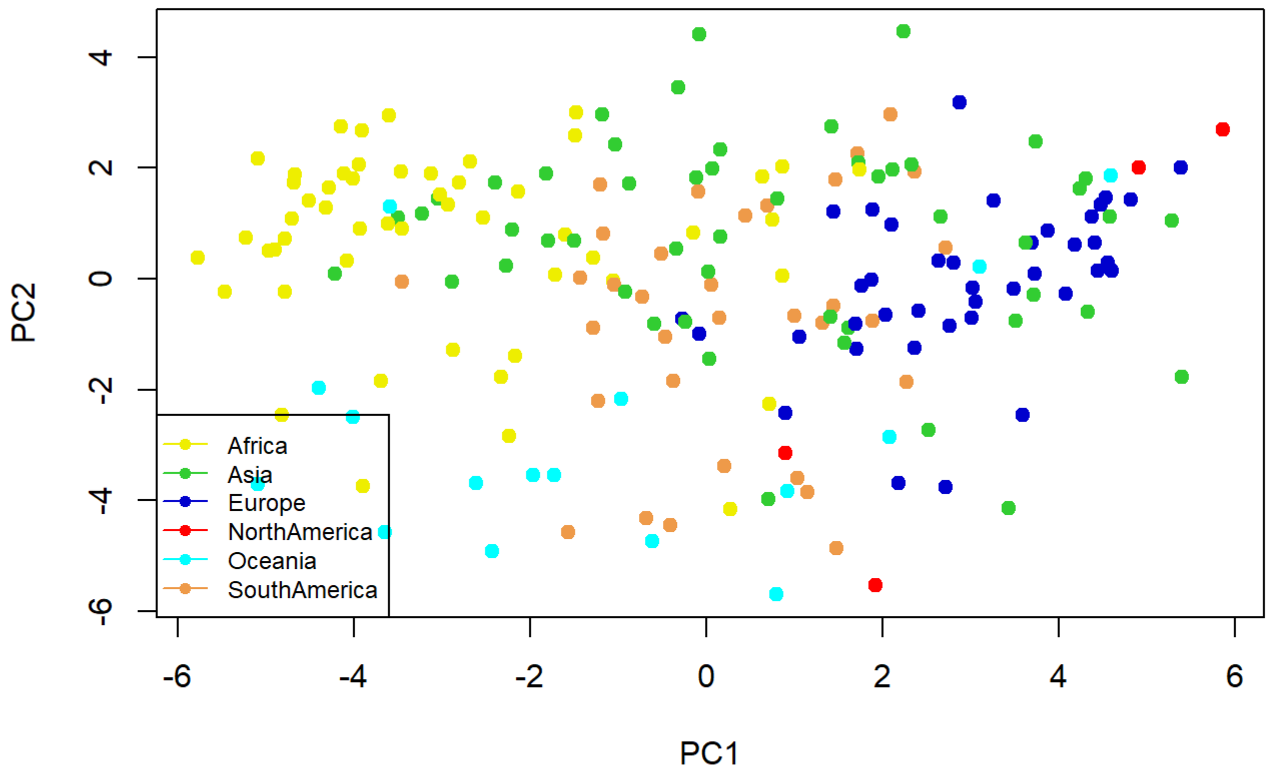
However, after seeing the graph we realized that the potential outliers are very important countries such as USA, Germany, Russia, UK… so we finally decided not to remove them from the dataset.

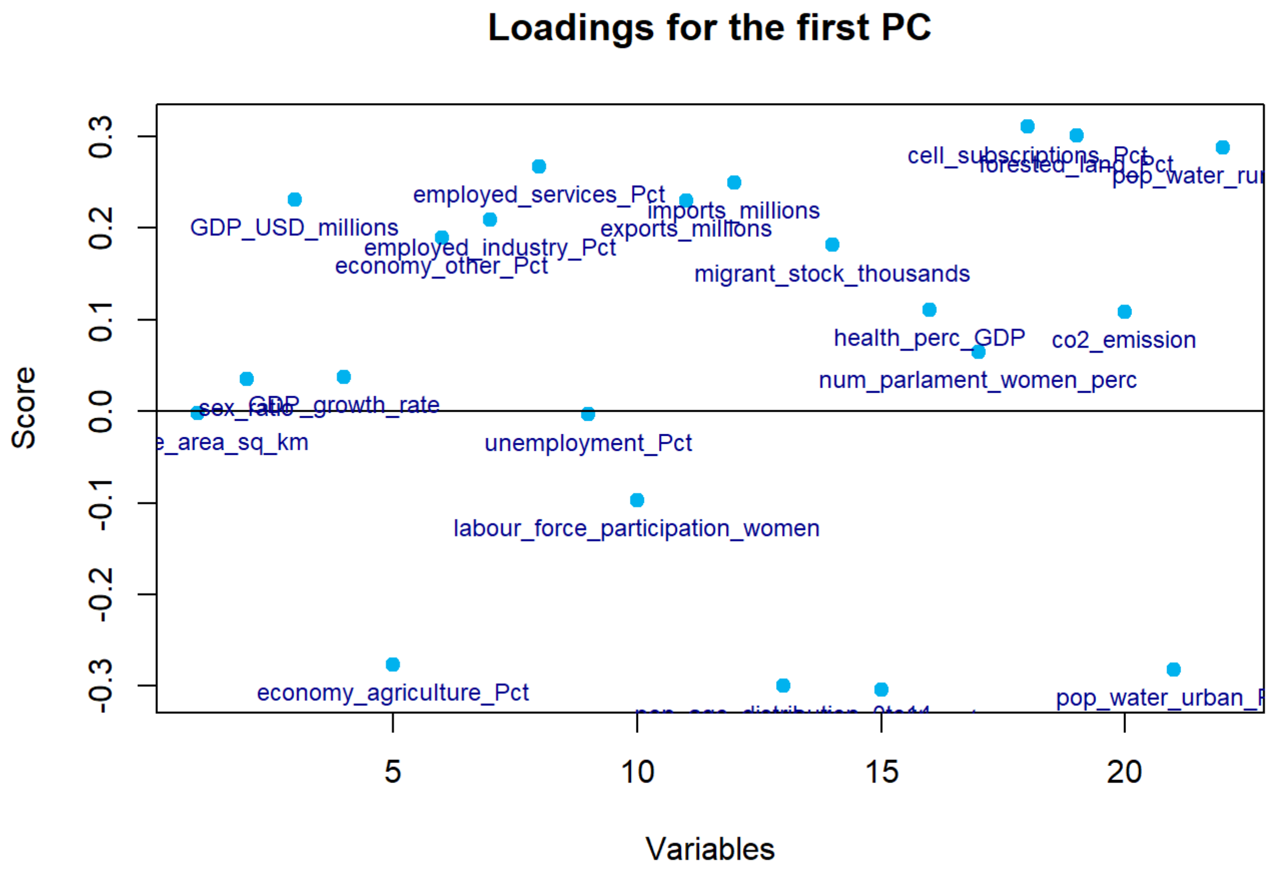
Principal Components Analysis (PCA)

We carried out the analysis with principal components and the most important results are shown below.

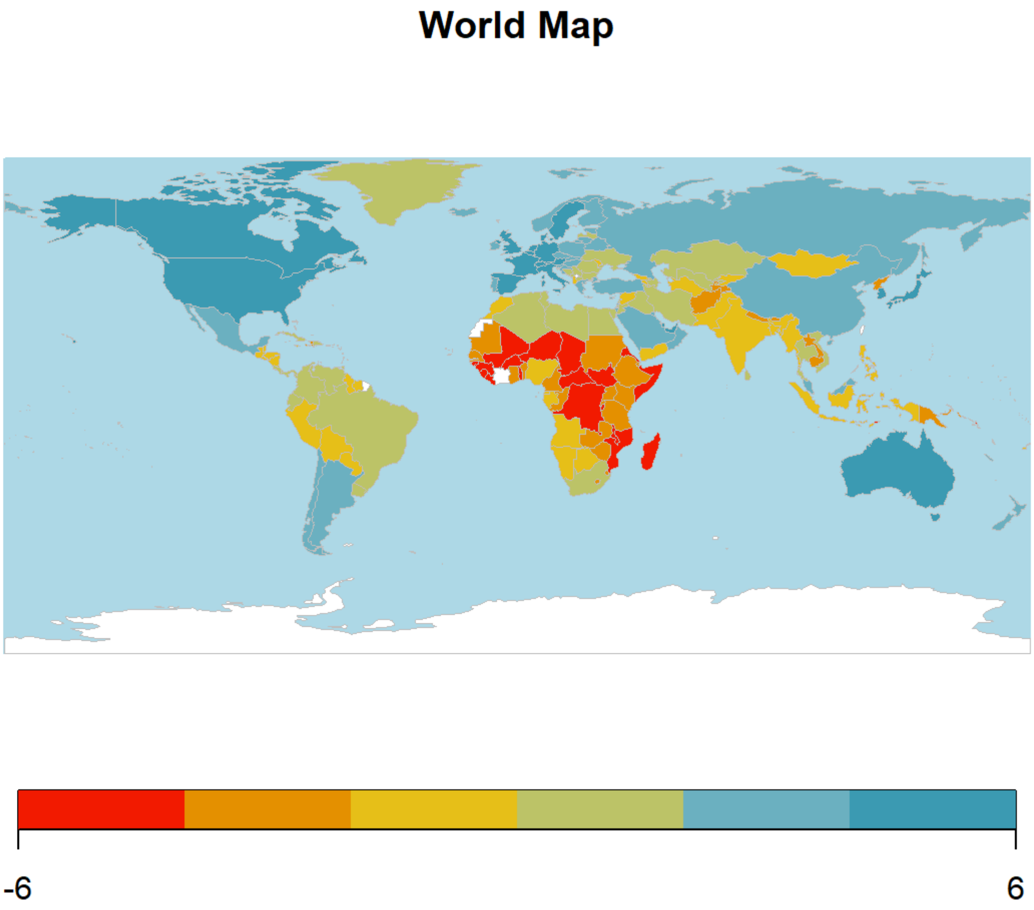
First PC

For the first PC we found the next loadings:



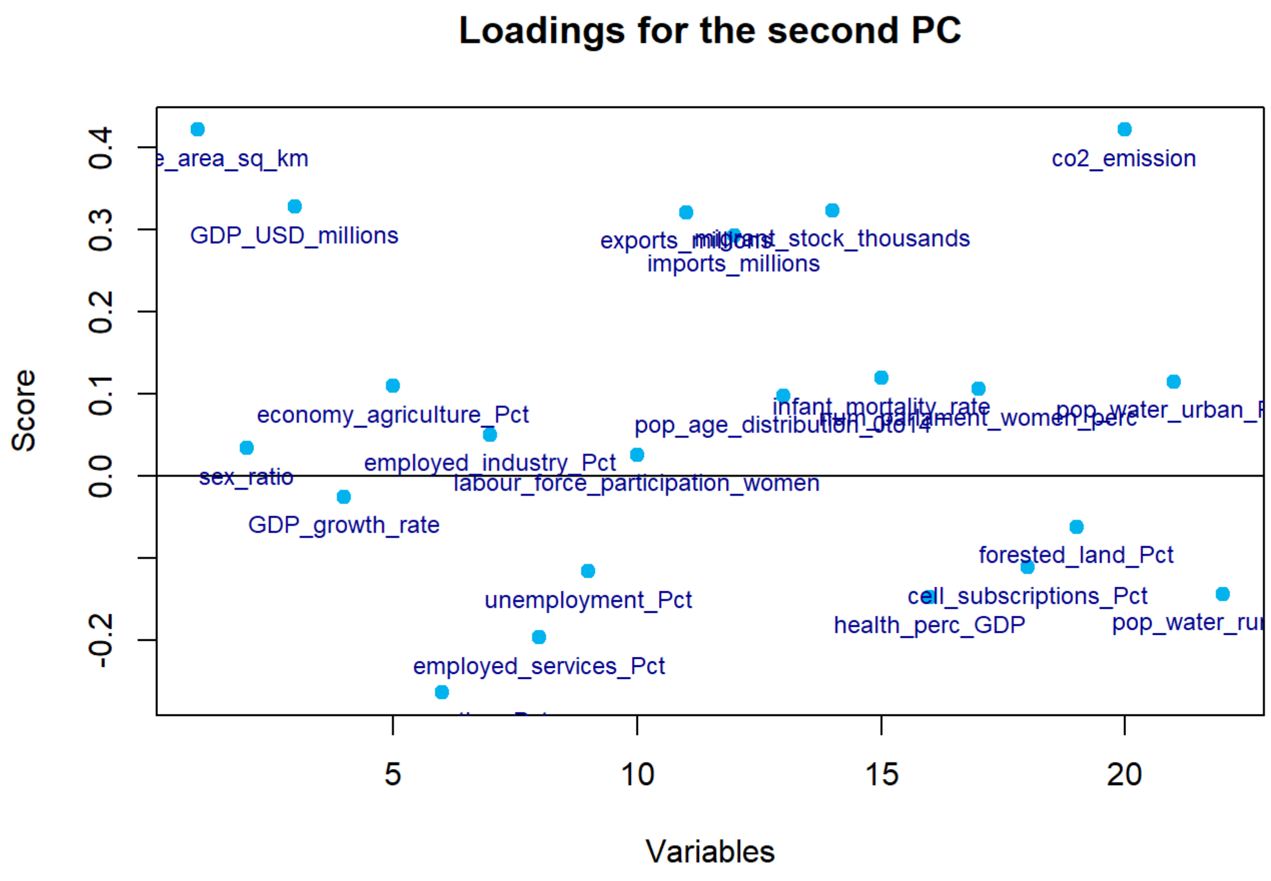


The interpretation we can give to the 1st PC is **underdevelopment** since they are countries with high percentage of people working in agriculture, high infant mortality rate and few imports/exports, GDP, industy. We can plot all the countries in a map according to this first component.

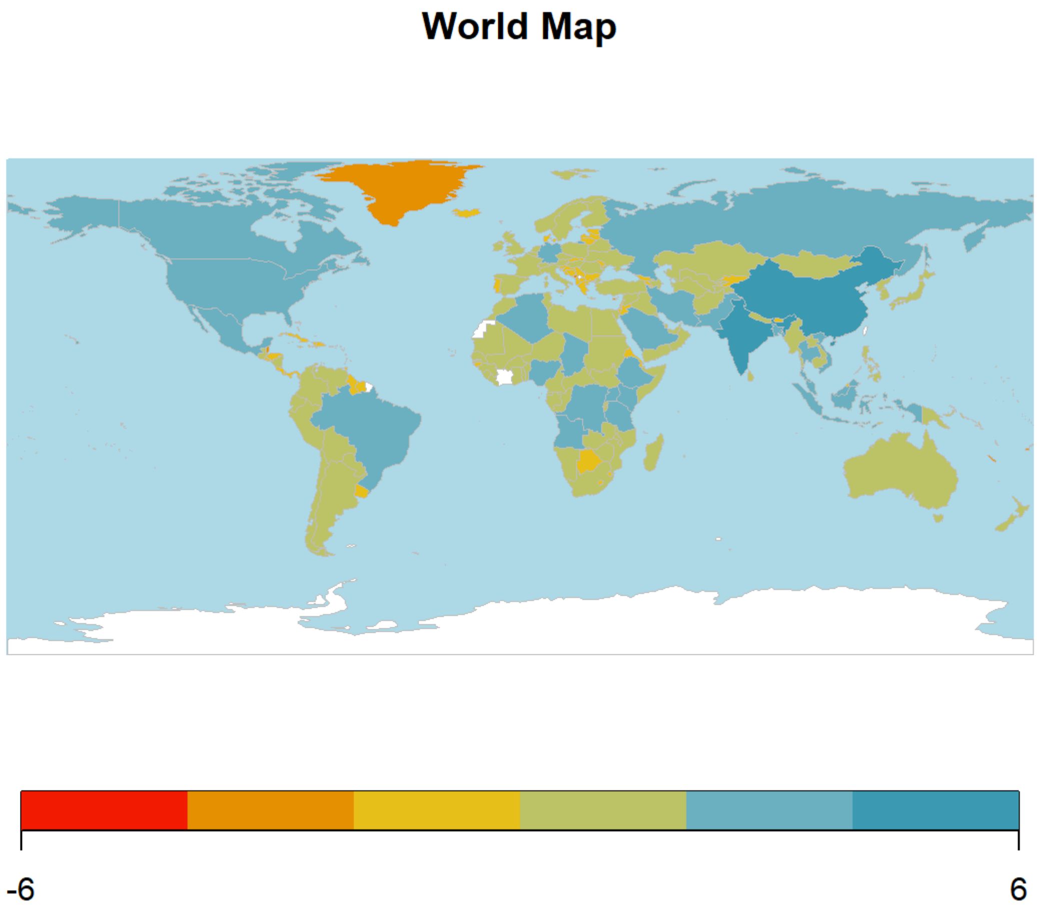


Second PC

For the second PC we found the next loadings:

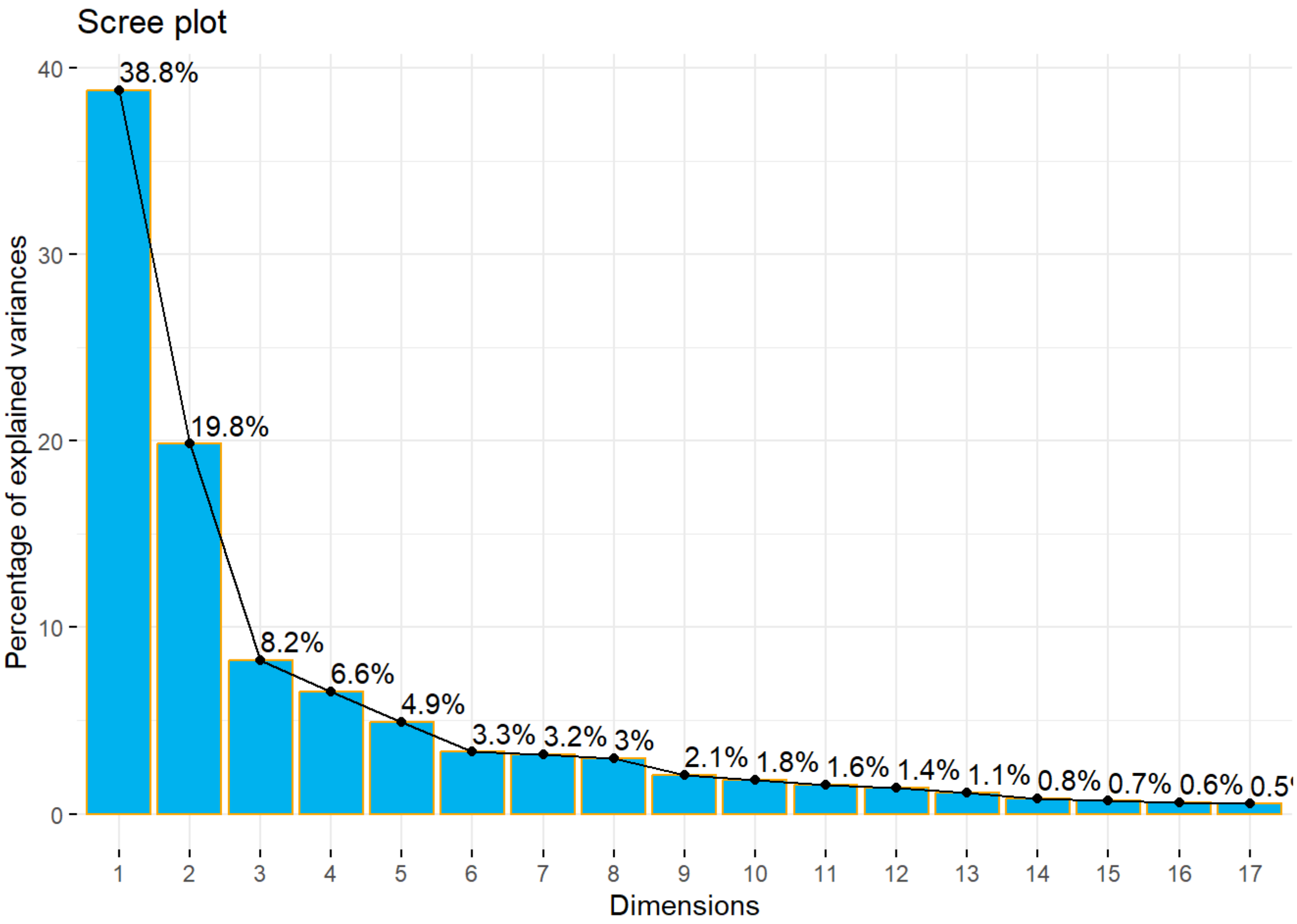


The interpretation we can give to the 2nd PC is **size of economy/population** since they are big countries with high exports/imports, GDP values and high CO2 emissions. We can plot all the countries in a map according to this first component.

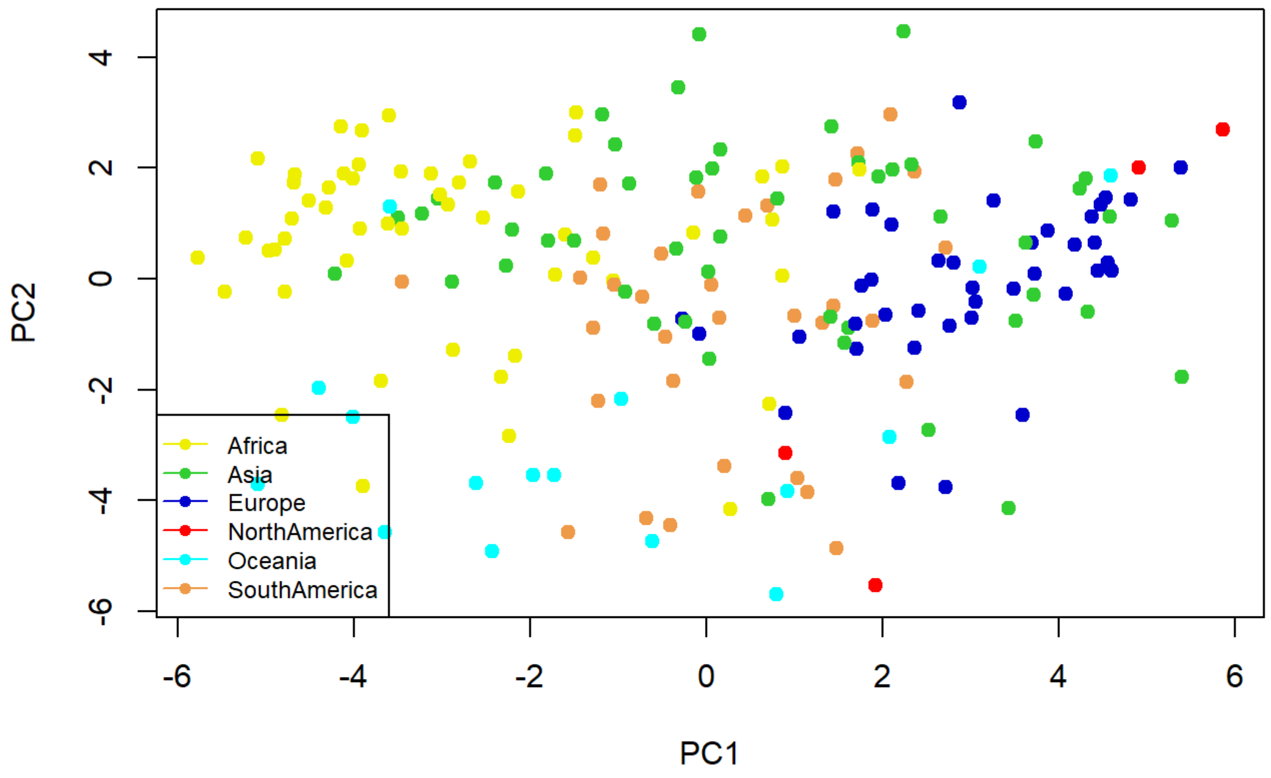


Other PCA results

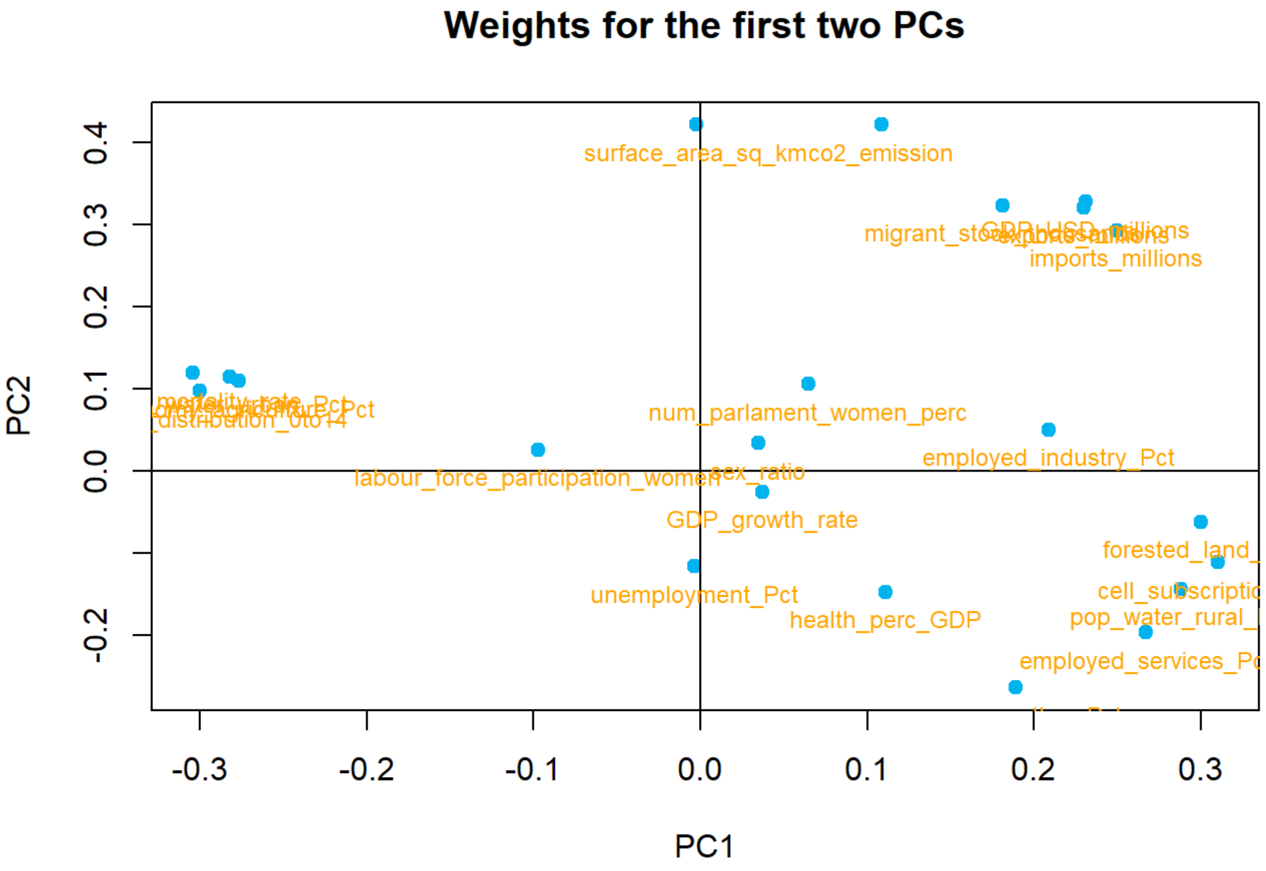
Clearly the first two PC’s are the most important ones:

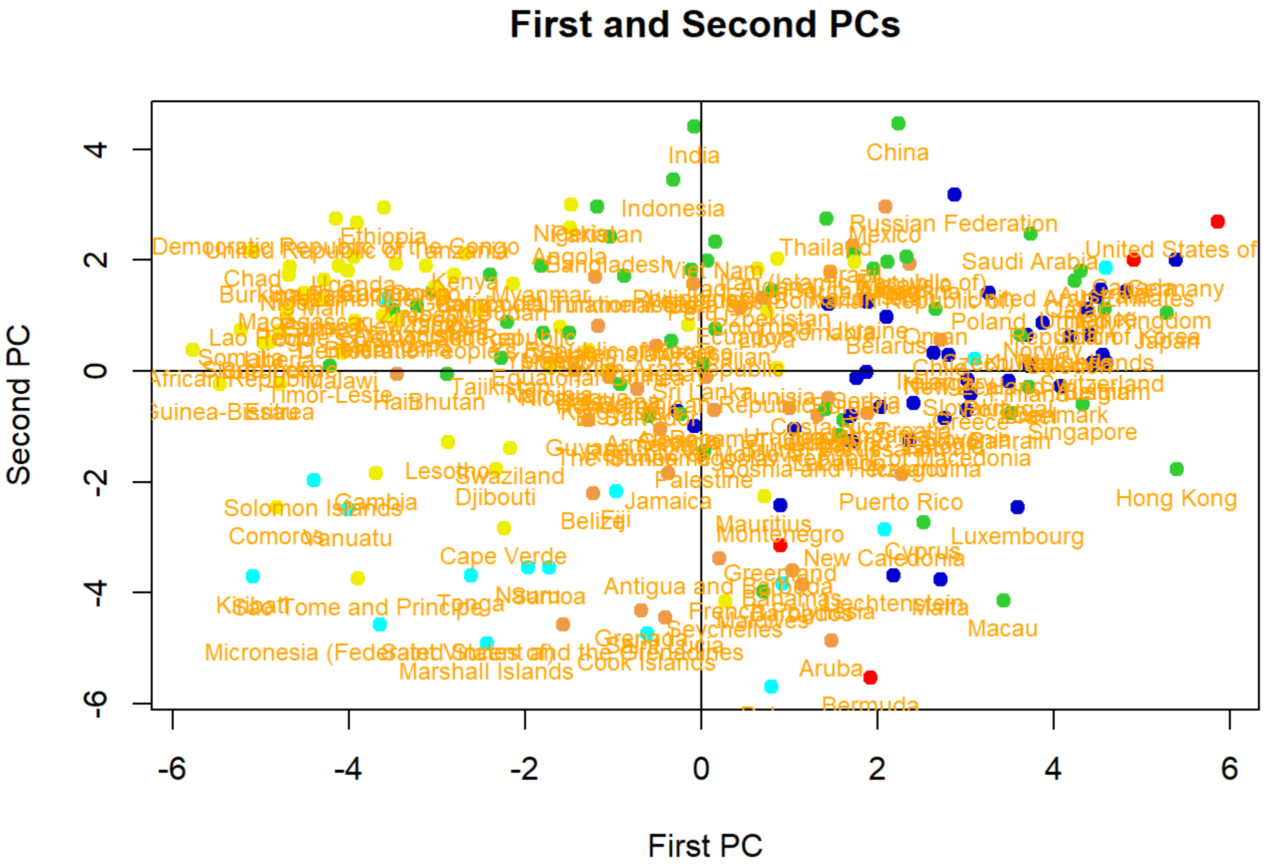


We can see the plot of countries according to the first 2 PC’s. As we see there’s not a clear division in the plot by continents though we can observe some tendencies like African countries (in yellow) being mostly on the right (underdevelopment) and European countries (dark blue) mostly on the left (developed). Anyway there is not a clear division in groups.



Also we can classify countries and also variables according to the first two PCs reminding that right = underdevelopment and up = big size.

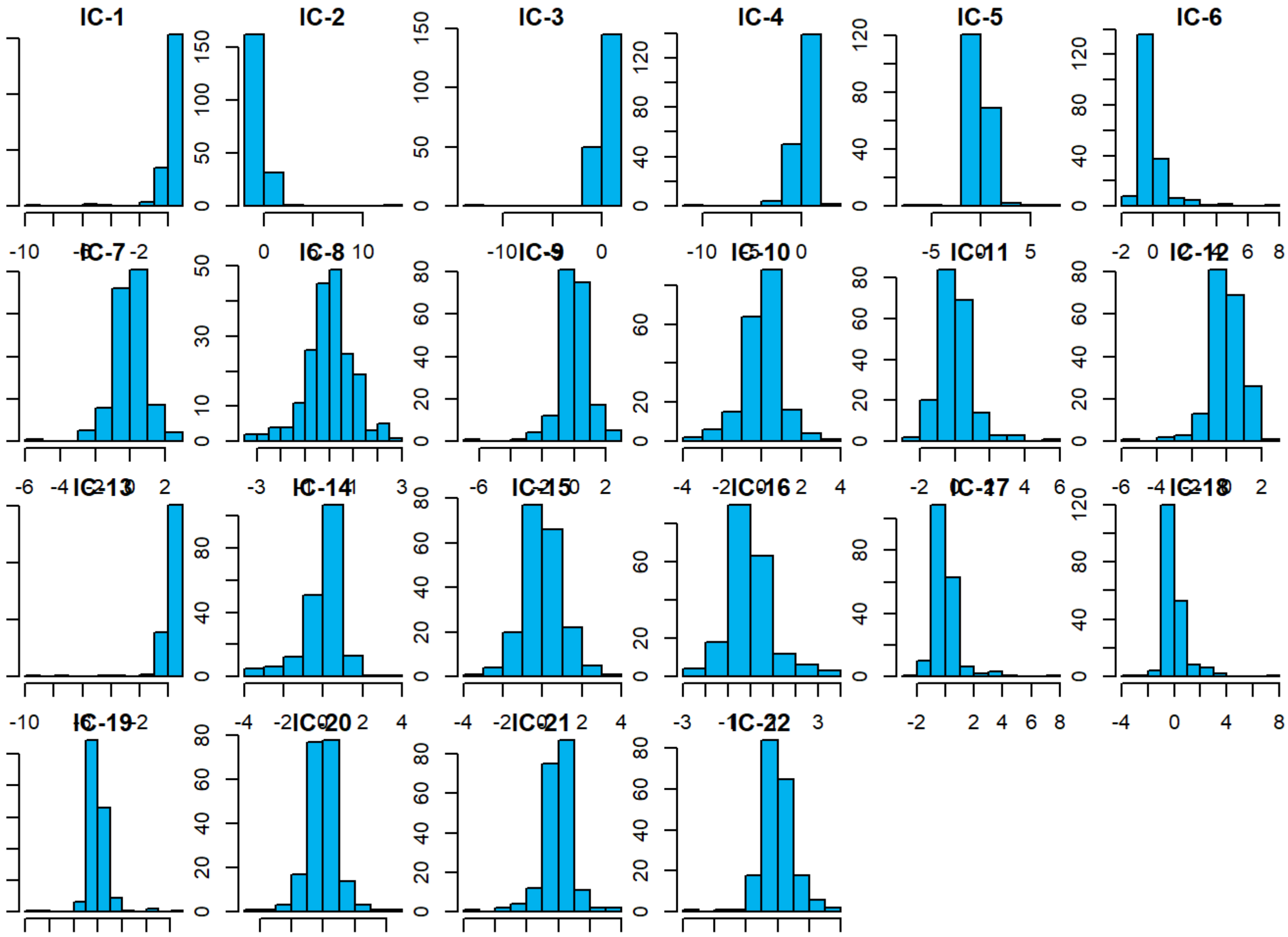




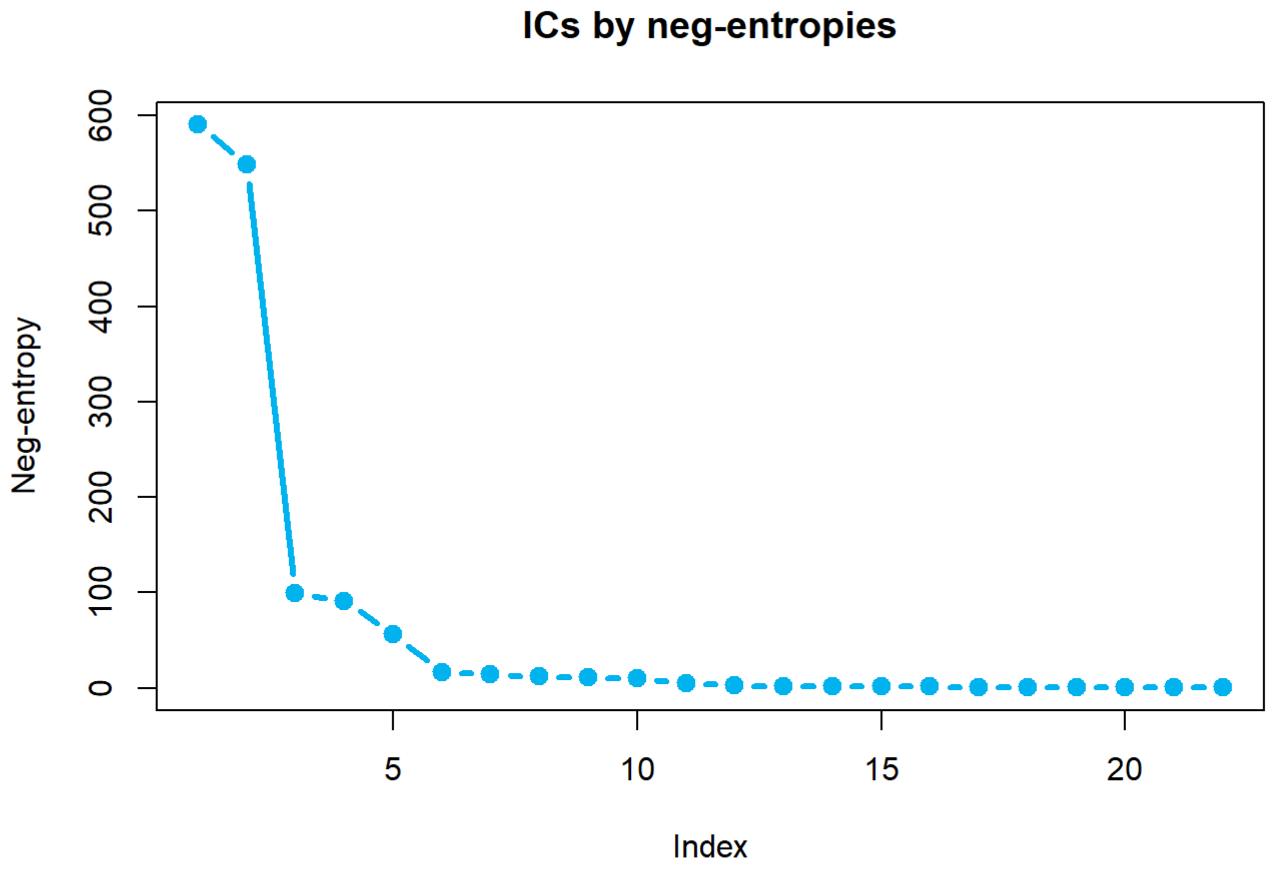
As conclusions it’s clear that the variable “continent” is not related to the principal components of the countries dataset. Rather, there are other categories like development or economy size that are taken into account.

Independent Component Analysis (ICA)

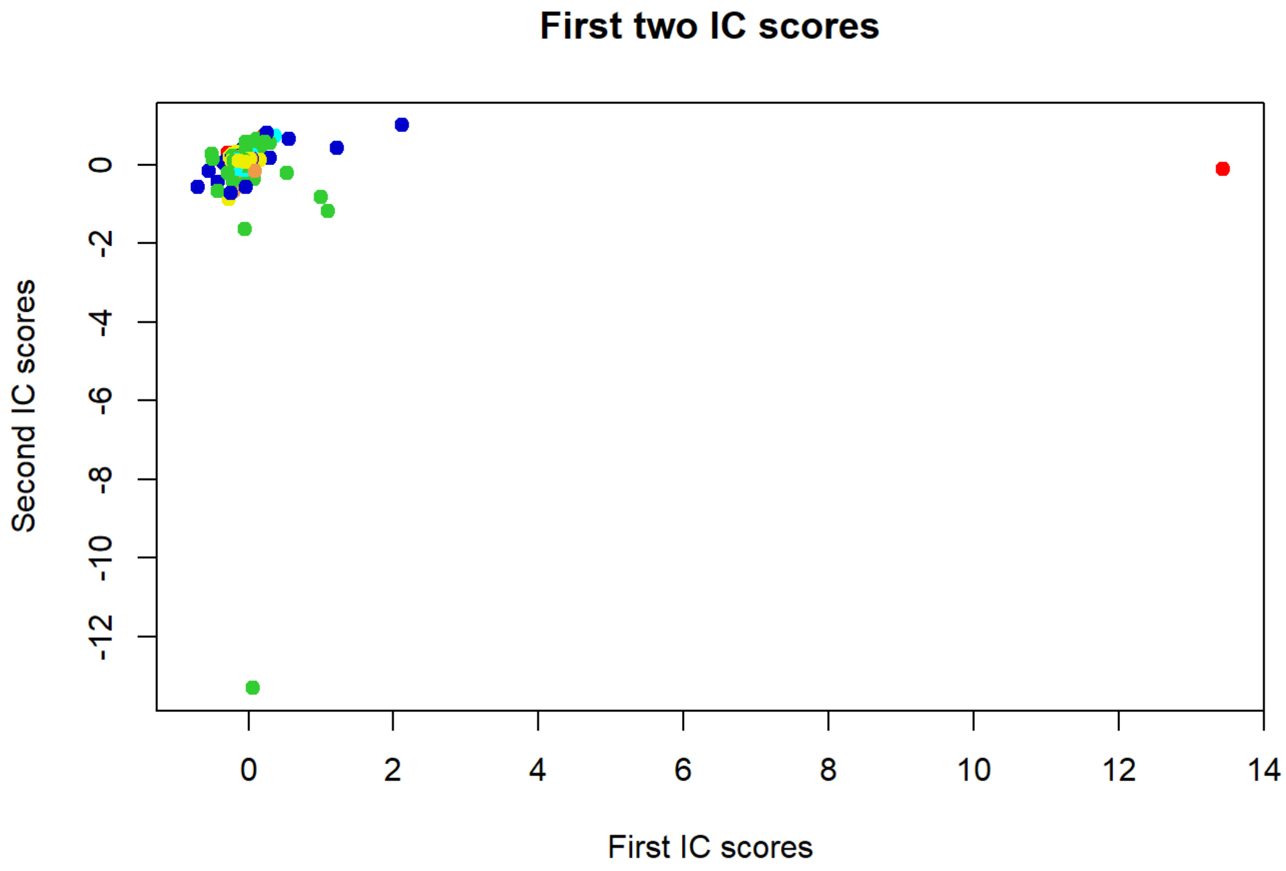
In the ICA, components are separated with the aim of making them independent, resulting in the first components being strongly non-Gaussian. In our case, after performing ICA, we obtain the following distributions.



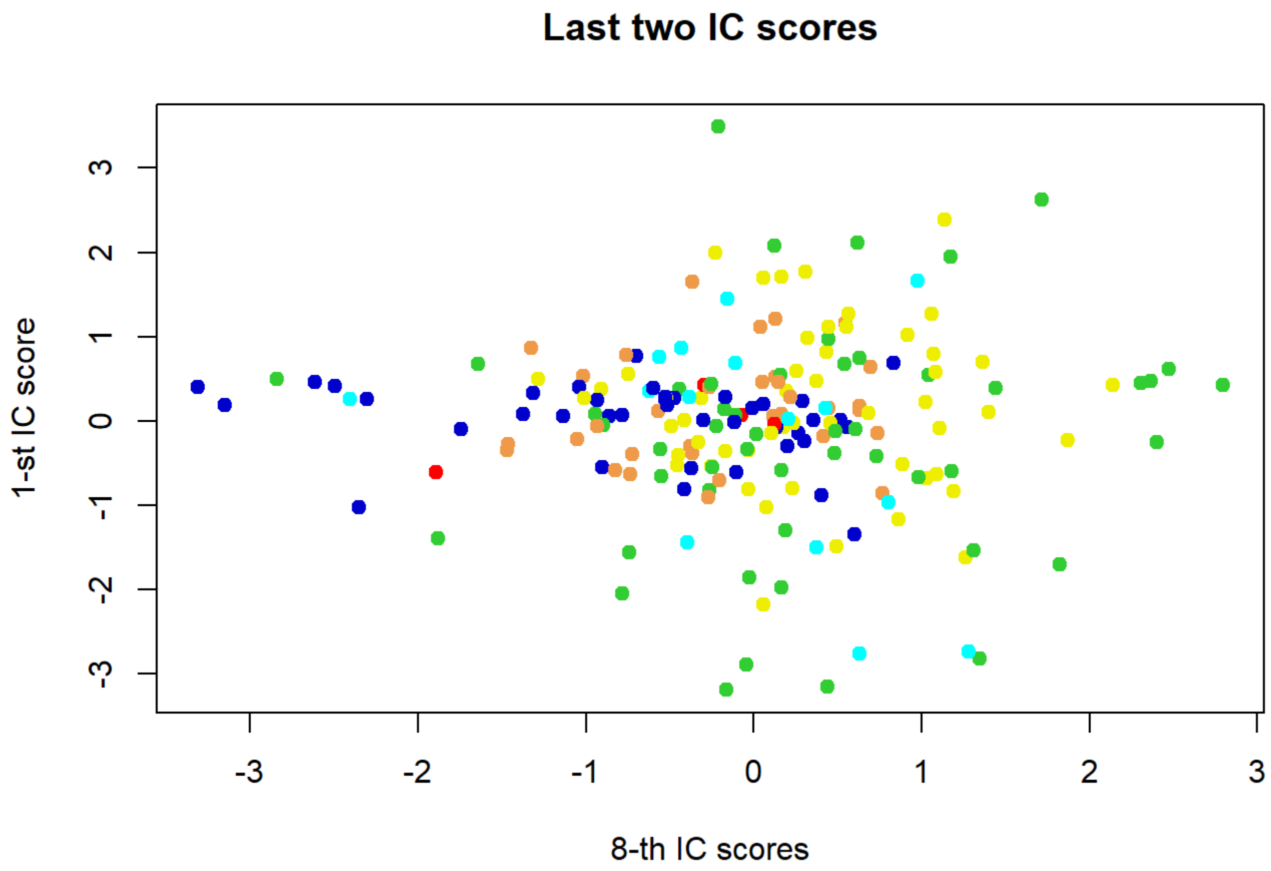
As wee see the first two ICs explain much more variance than the rest:



The plot of the first two ICs helps us to identify outliers. In this case we can observe two clear outliers: China (green) and USA (red).

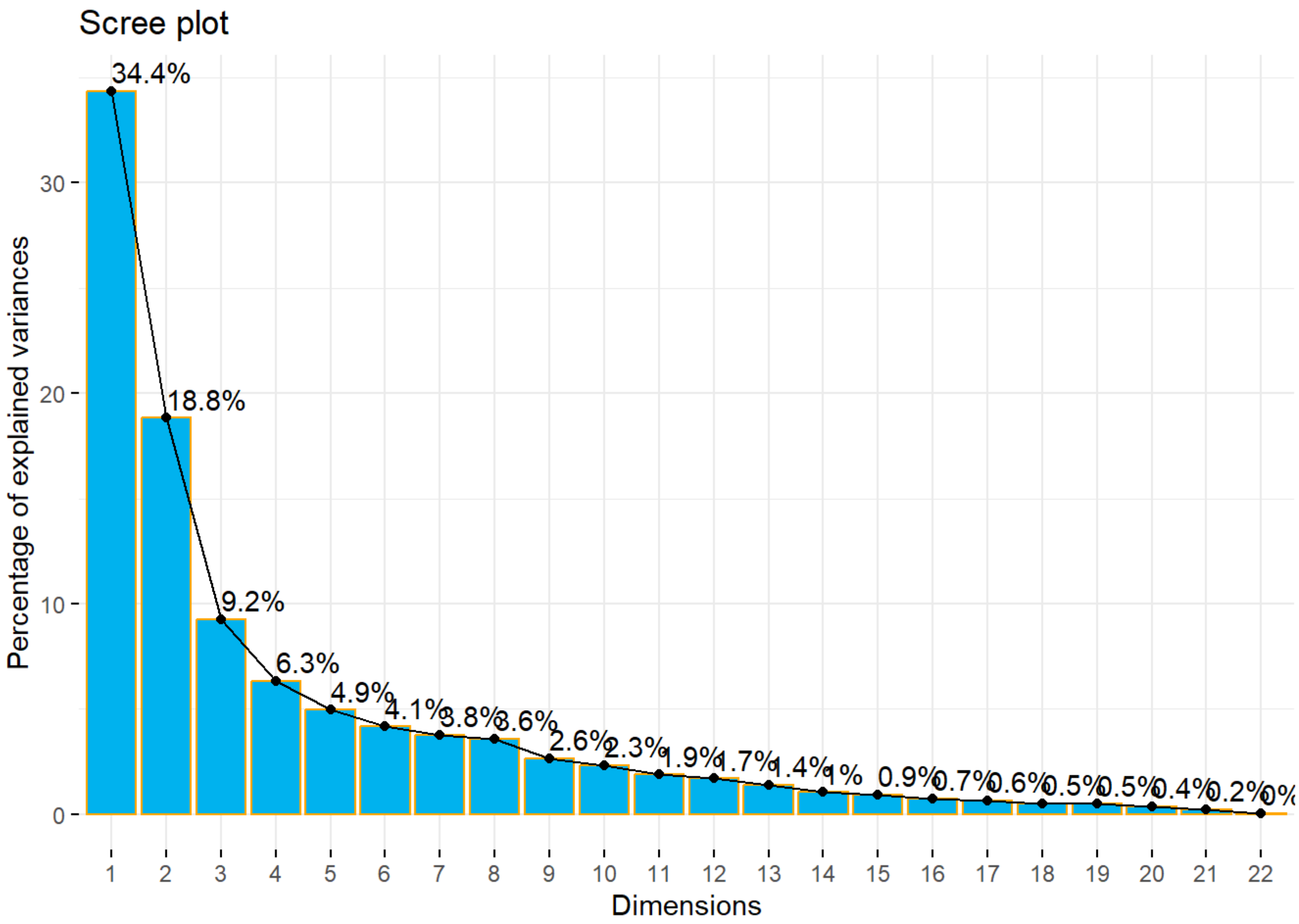


The last two ICs usually help us to identify groups in the dataset but in this case it doesn’t happen so and we see all the continents are mixed.



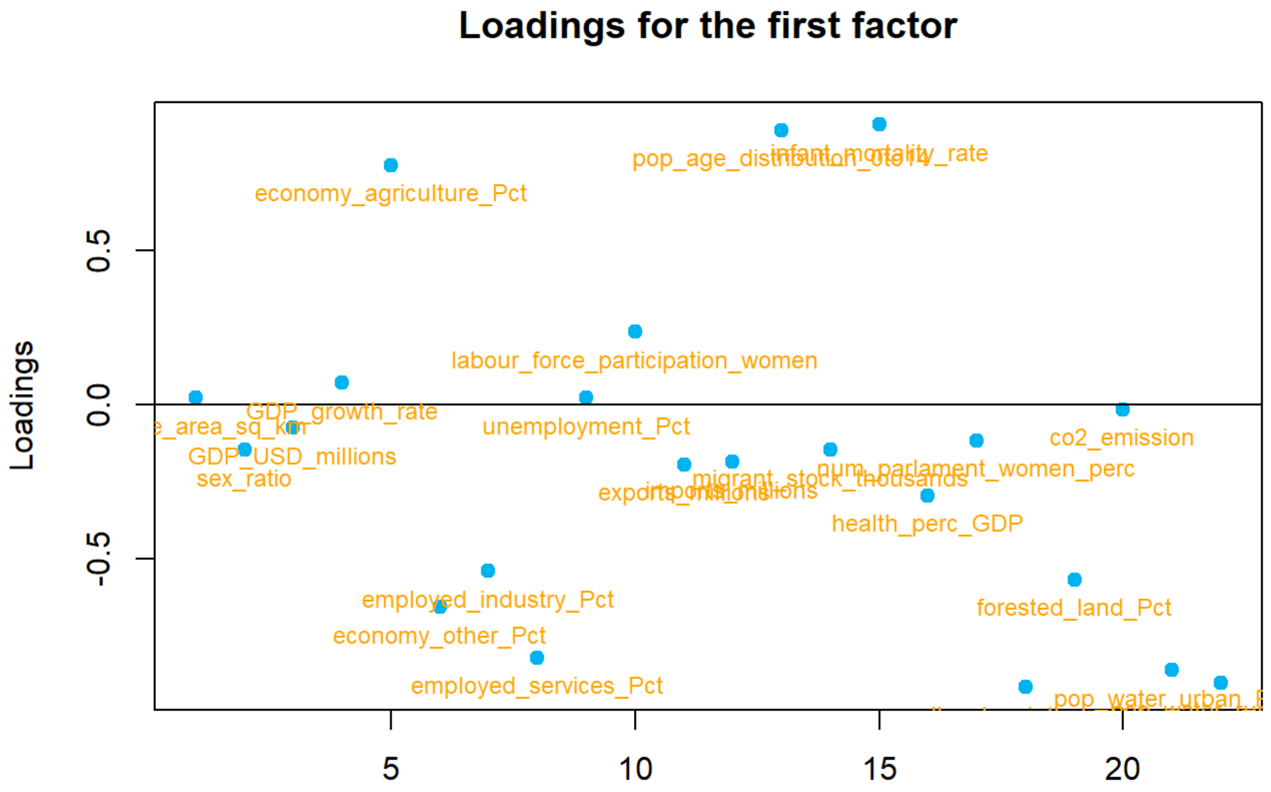
Factor analysis

Now, using Factor Analysis, we will try to identify factors not necessarily present in the dataset as variables that are behind the variability of the data. For this we see that the first two factors, on which we are going to focus, are the most important.

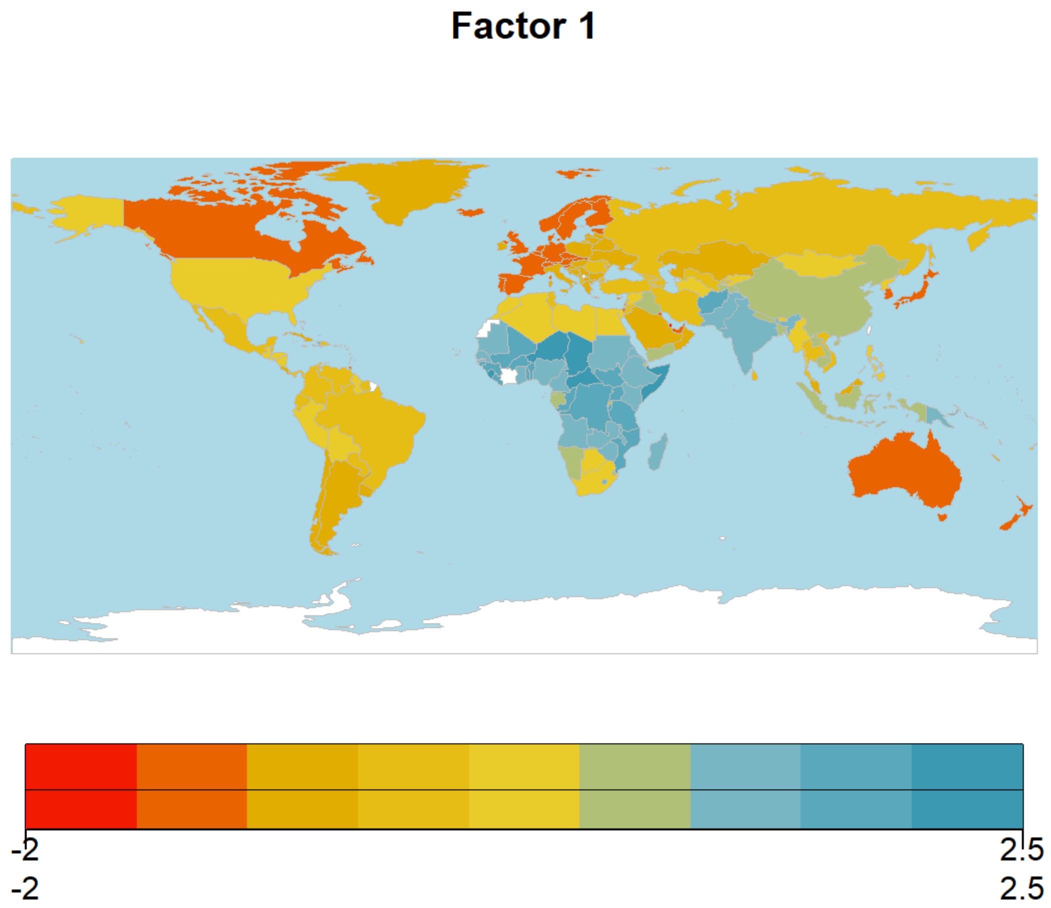


First factor

Using the four first PCs (the most relevant ones) we compute the factors and for the first factor we obtain the following loadings:



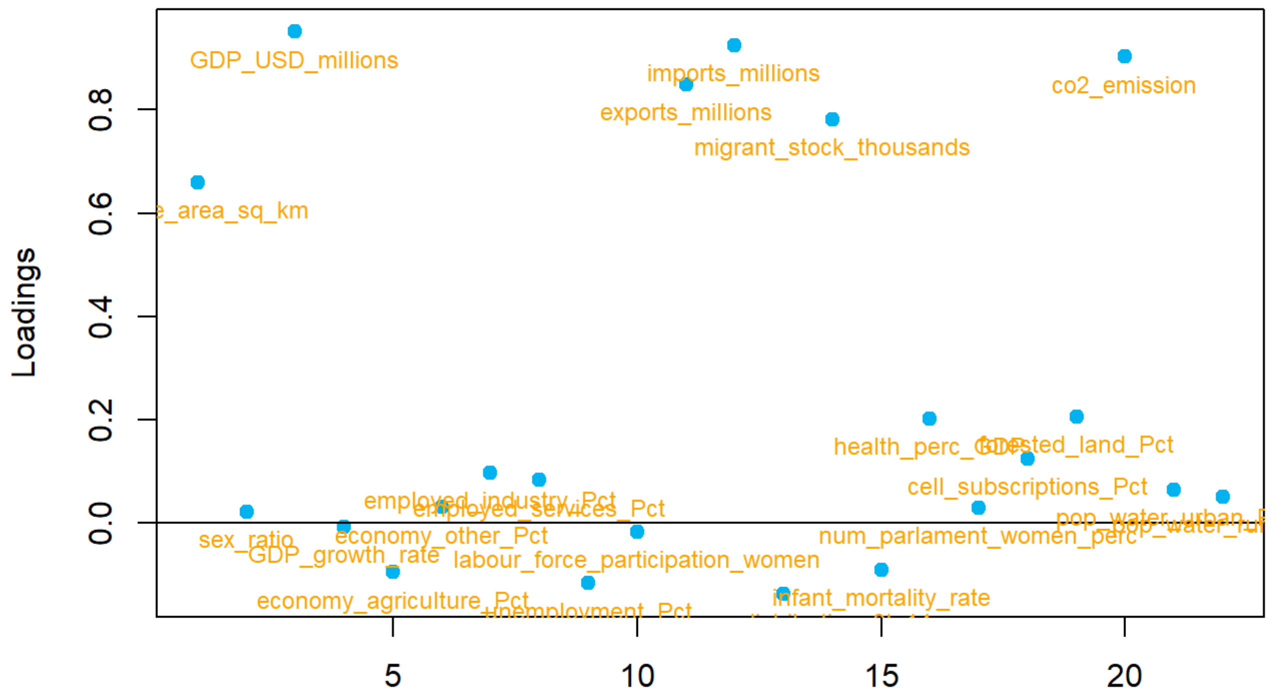
As we see it’s very related with the first PC, in fact we can give the same interpretation of it: an index of underdevelopment. We can see on the map its representation.



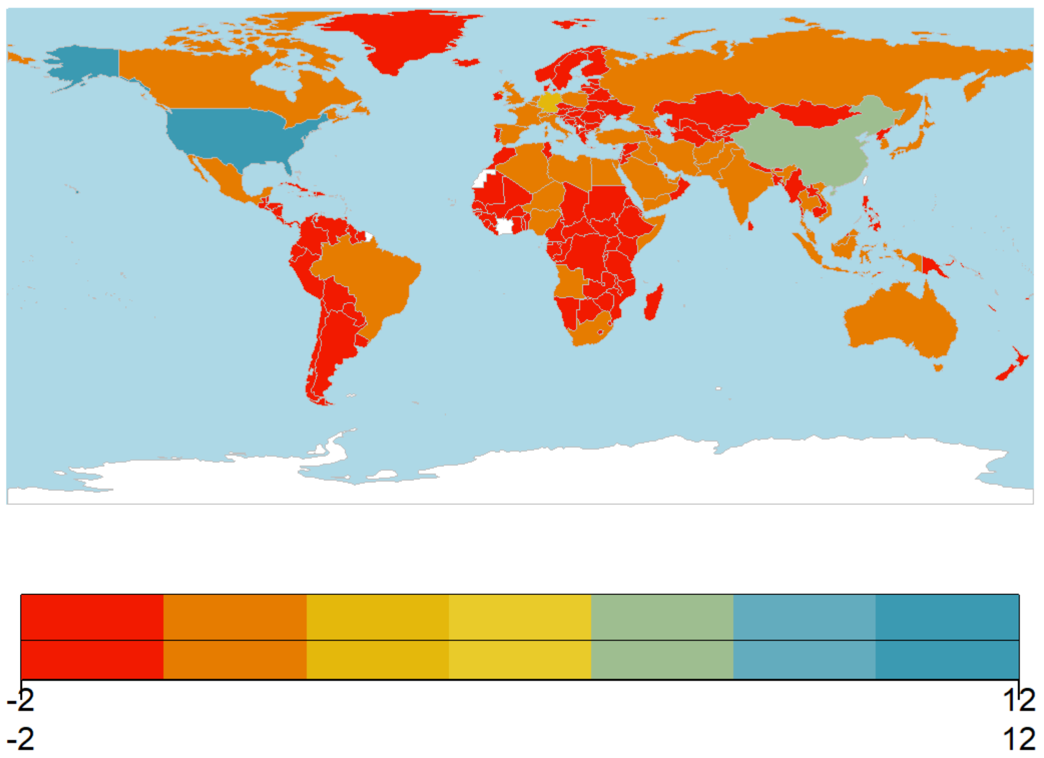
Second factor

For the second factor, similarly, we can draw the loadings plot. In this case we can interpret this factor as the economic activity size in absolute numbers.

Loadings for the second factor

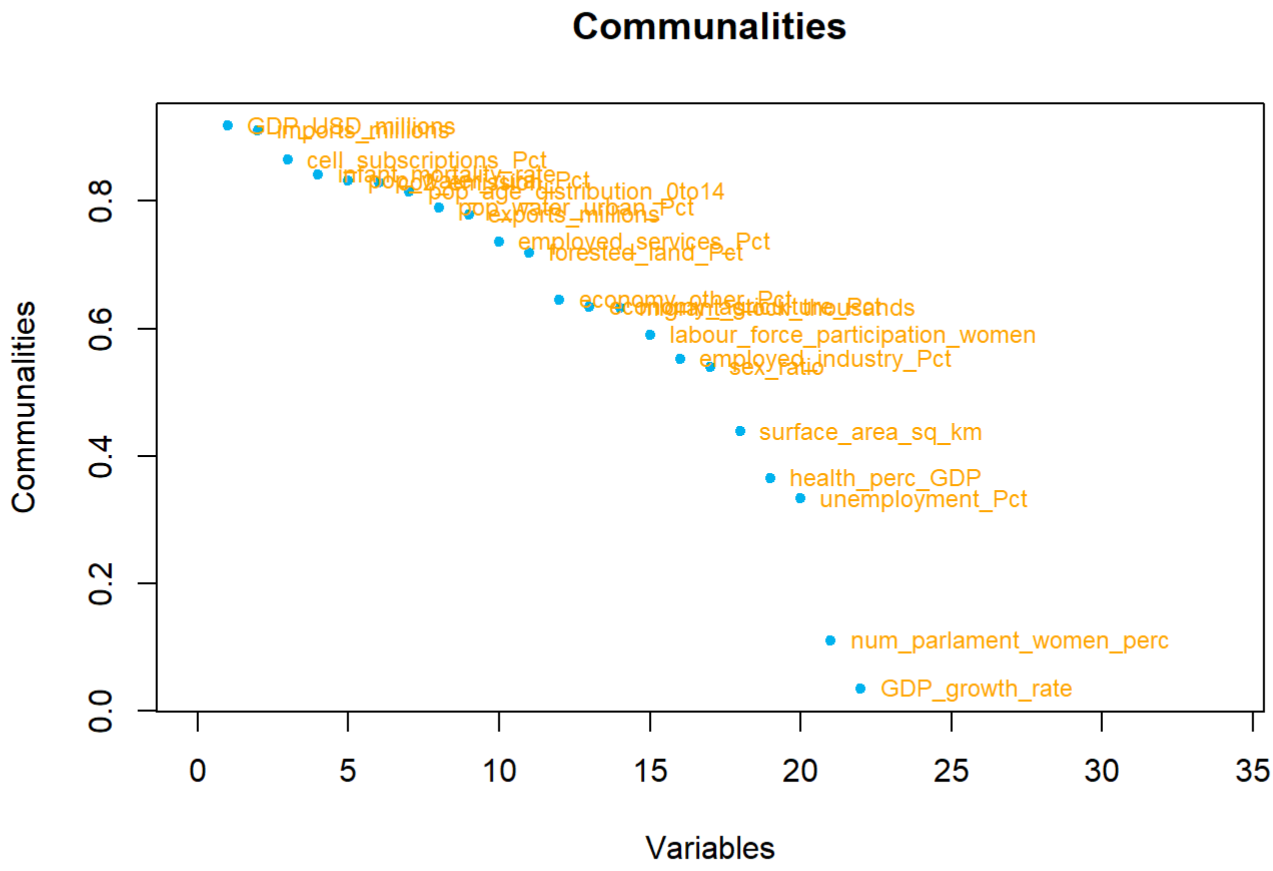


Factor 2



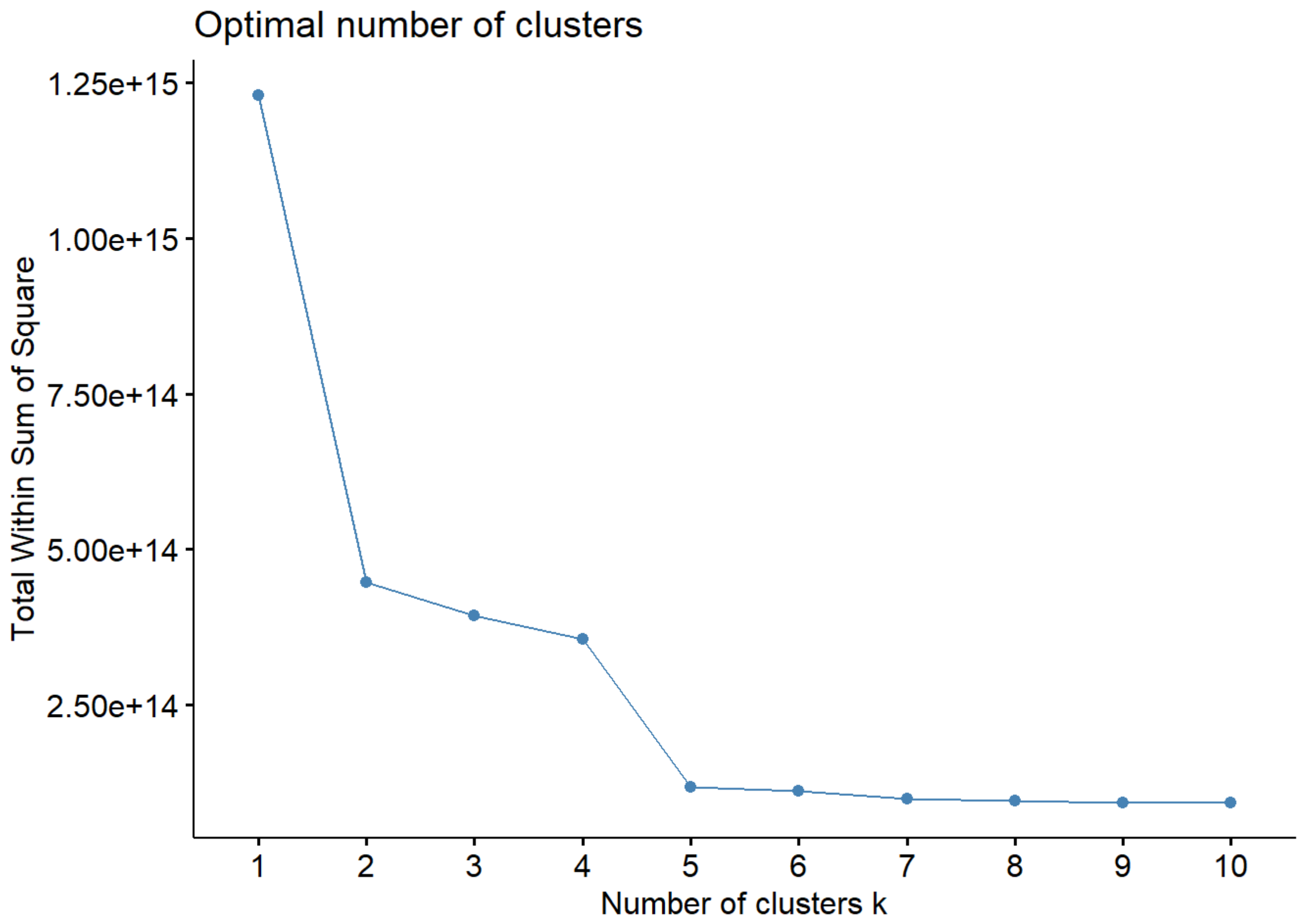
Conclusions of factor analysis

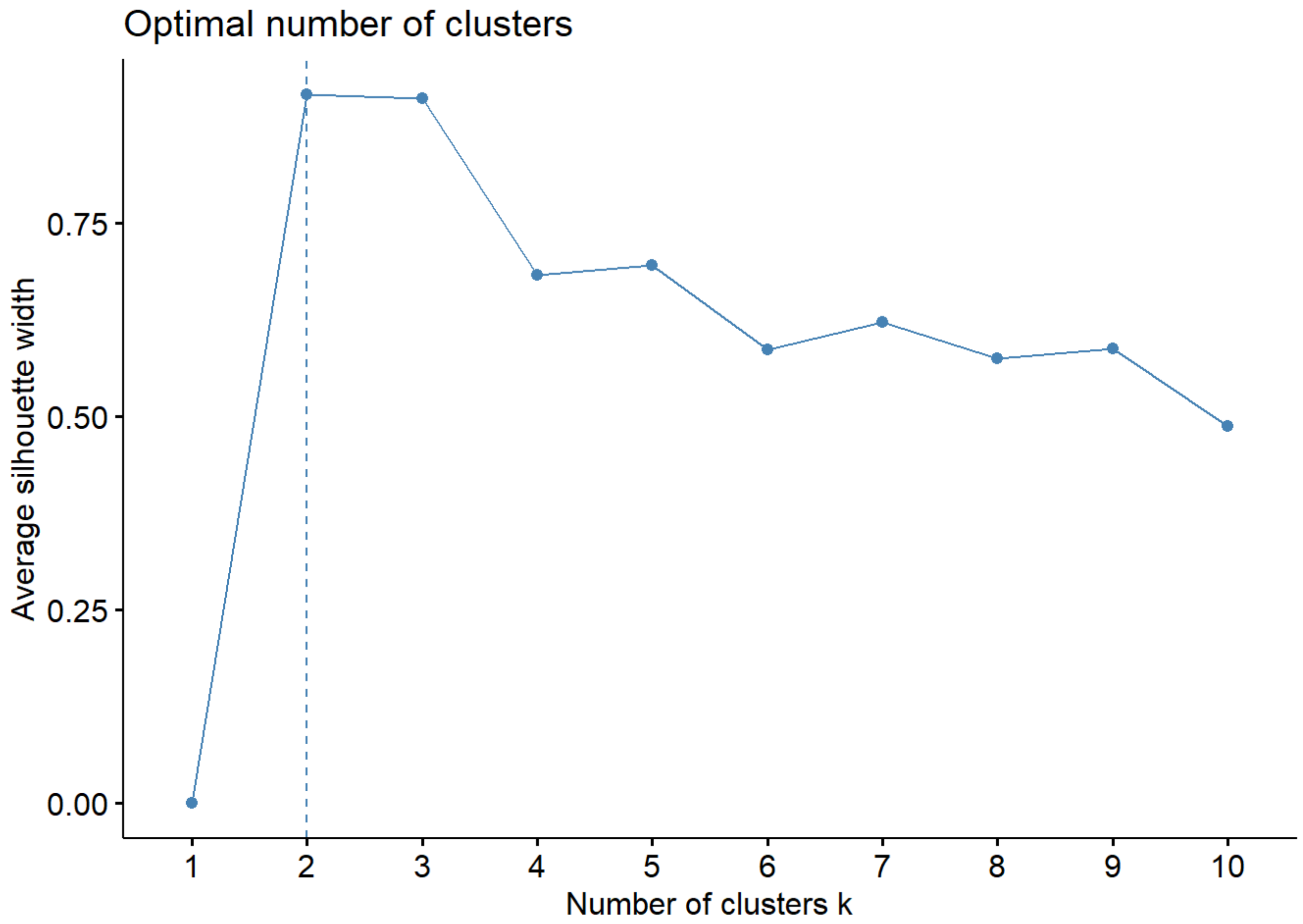
As we have been able to appreciate, the analysis of the factors has yielded similar results to the PCA, with the first two factors and the first two principal components being quite similar. To understand which variables are better explained by factor analysis we can plot the communalities.

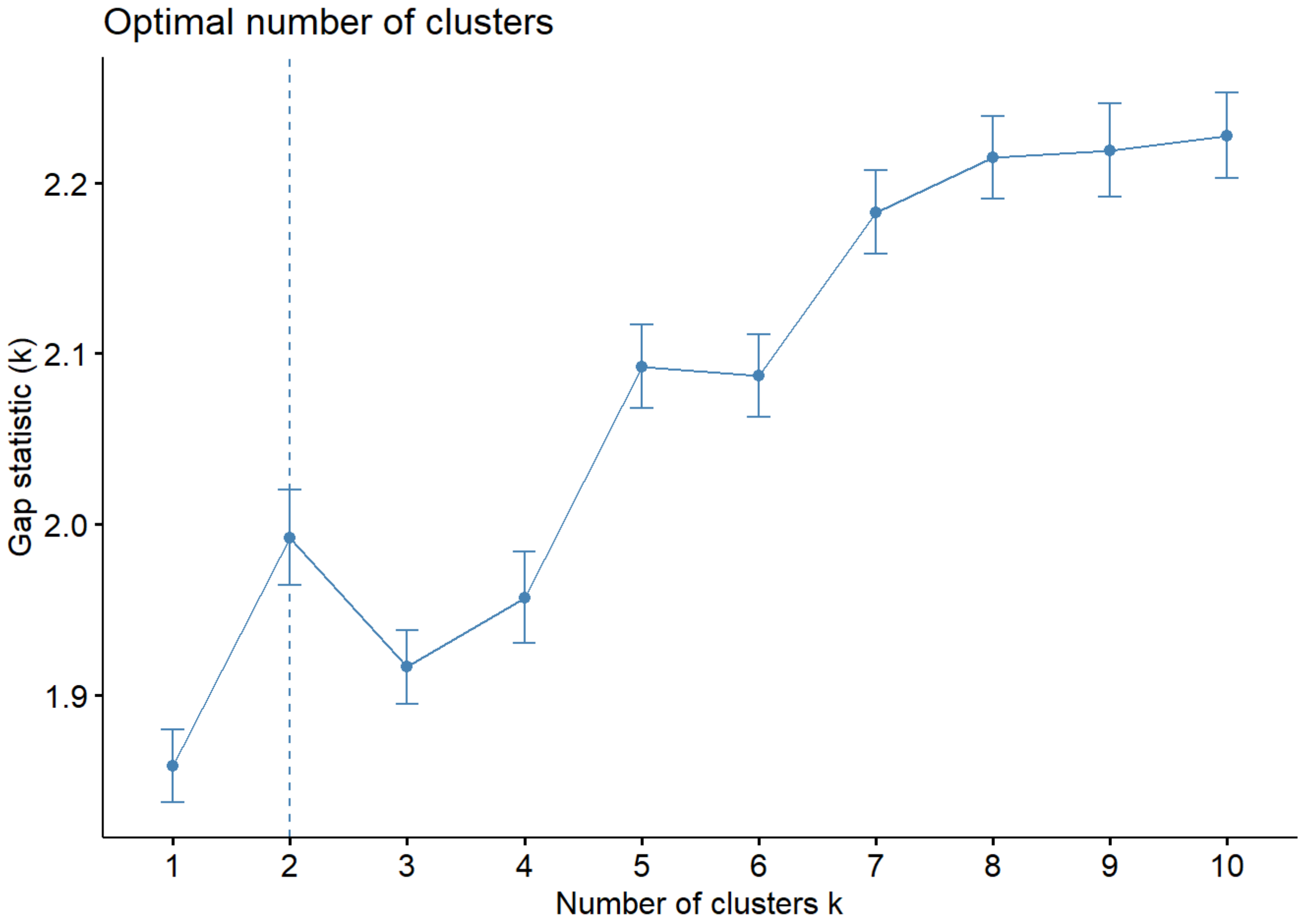


K-means

We are now going to study what grouping some unsupervised clustering methods perform to see if they have a certain relationship with the continents. First of all, to carry out K-means we need to find the optimal k for what we make some tests using different methods (WSS, silhouette and GAP statistic). These are the results.

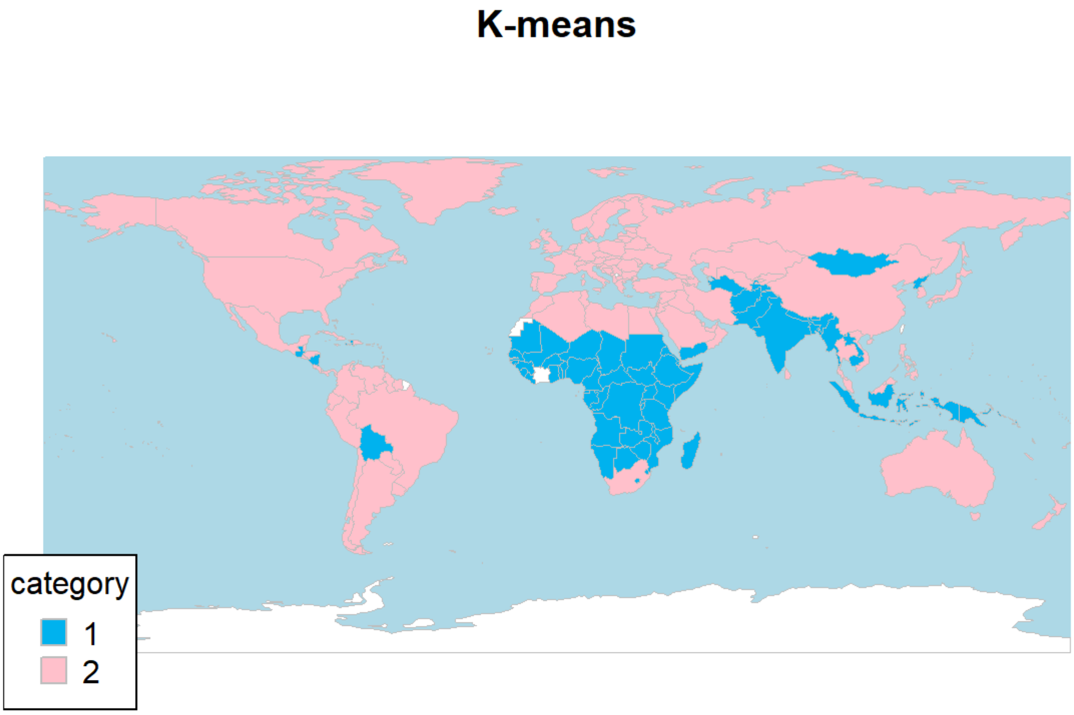

Optimal number of clusters
Total Within Sum of Square
1.25e+15
1.00e+15
7.50e+14
5.00e+14
2.50e+14
1
2
3
4
5
6
7
8
9
10
Number of clusters k


Optimal number of clusters
Average silhouette width
0.75
0.50
0.25
0.00
1
2
3
4
5
6
7
8
9
10
Number of clusters k

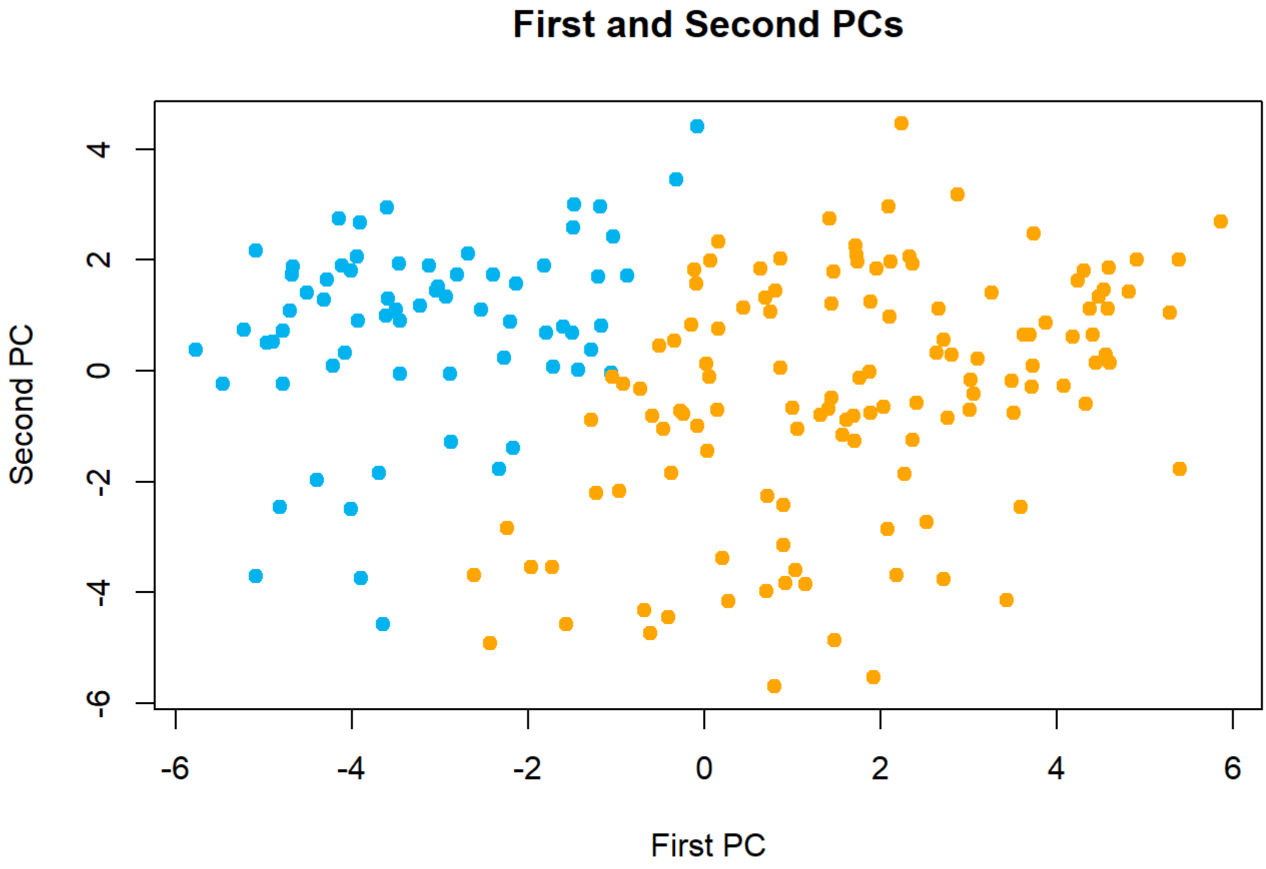


We can observe that the optimal k is 2. If we carry out k-means with k=2 we obtain the following two clusters.

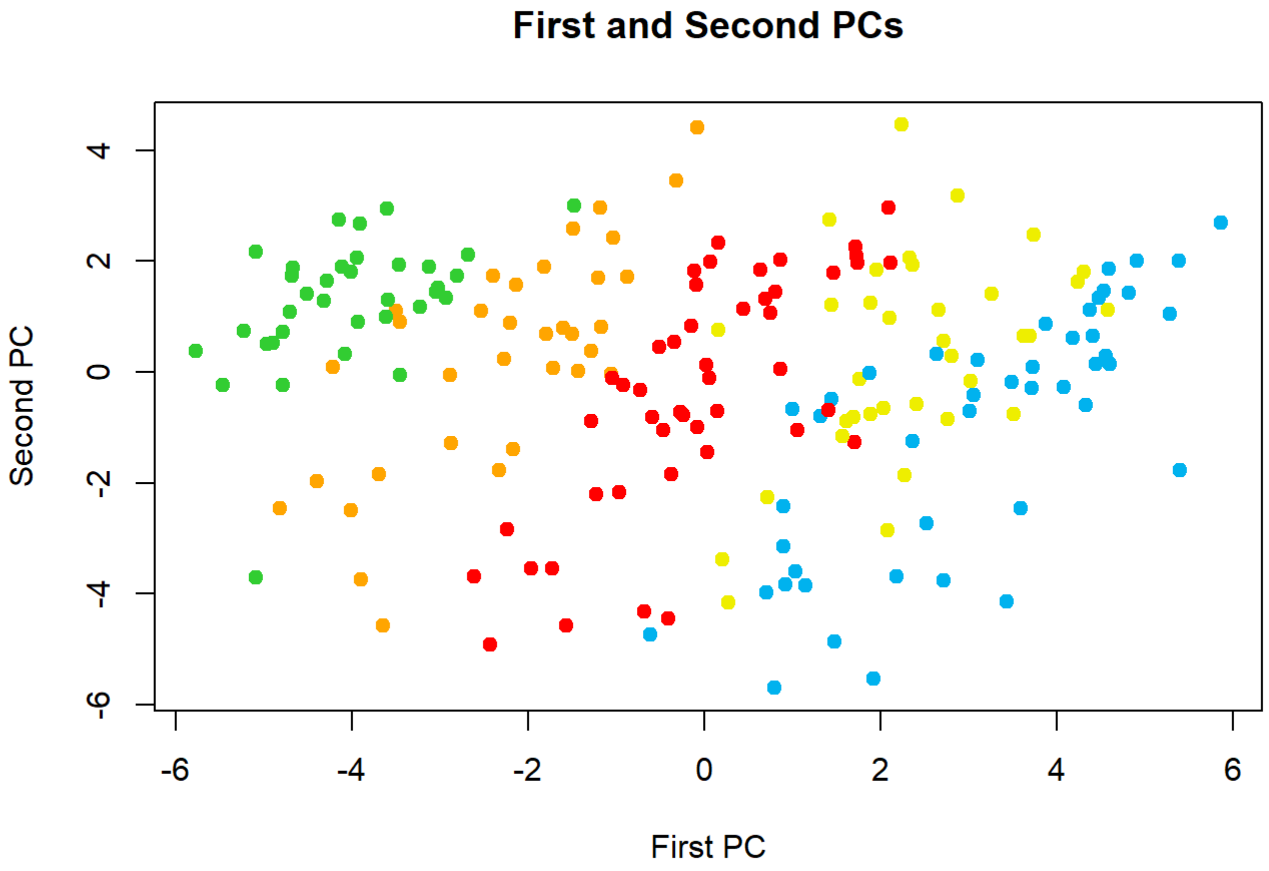
```
## Error in image.default(iy, ix, t(iz), xaxt = "n", yaxt = "n", xlab = "", : must have one m
ore break than colour
```



As we see the algorythm is dividing between poor countries and other countries, following in general terms the first component as we can appreciate in this plot: the division between clusters is done according (aproximately) to the first component axis.

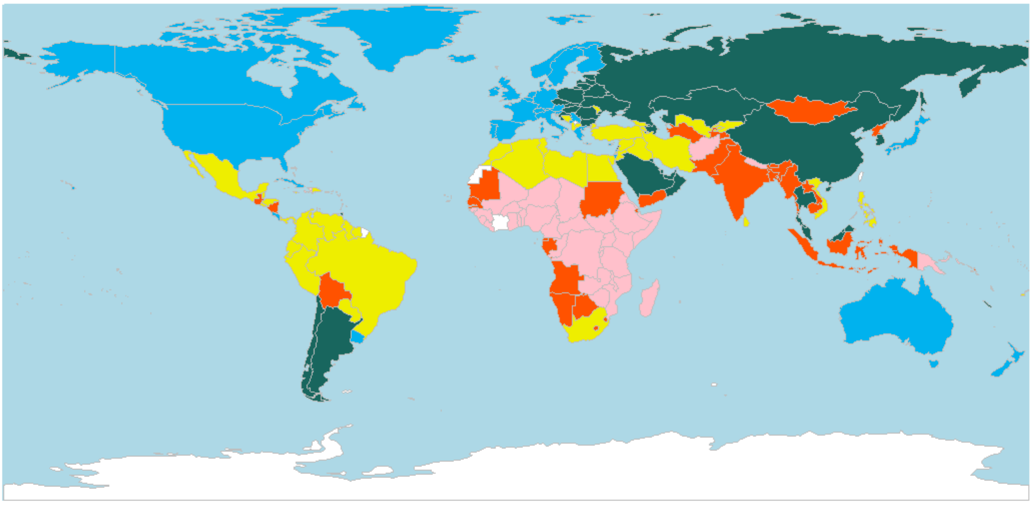


That division is interesting since it remarks the conclusions of PCA and Factor analysis but if we want to try to see a division between continents it’s not so usefull. For that reason we try also K-means with k=5 (k=5 was the second most interesting k in the graphs above)



```
## Error in image.default(iy, ix, t(iz), xaxt = "n", yaxt = "n", xlab = "", : must have one m
ore break than colour
```

K-means



Unsupervised Classification

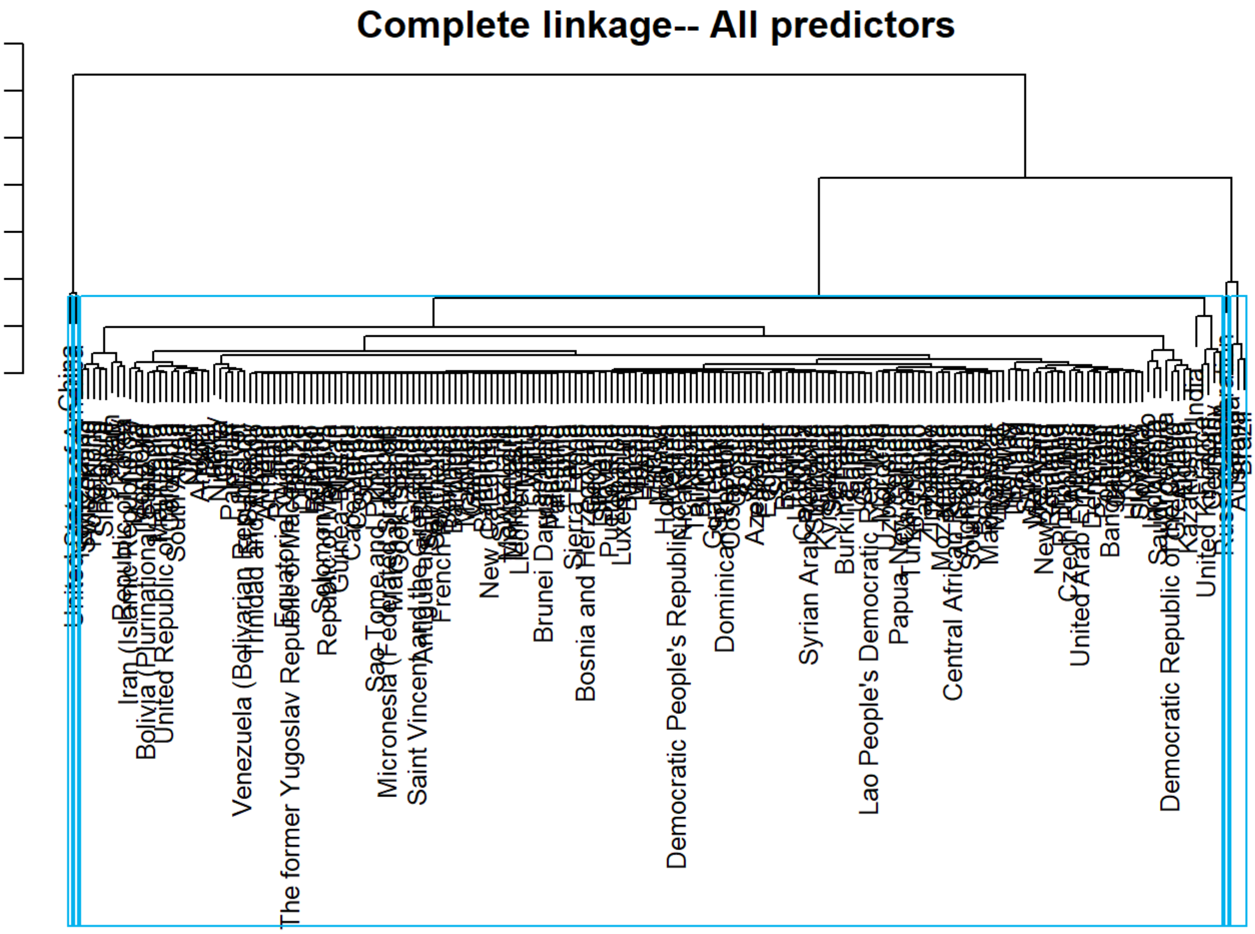
Hierarchical Clustering

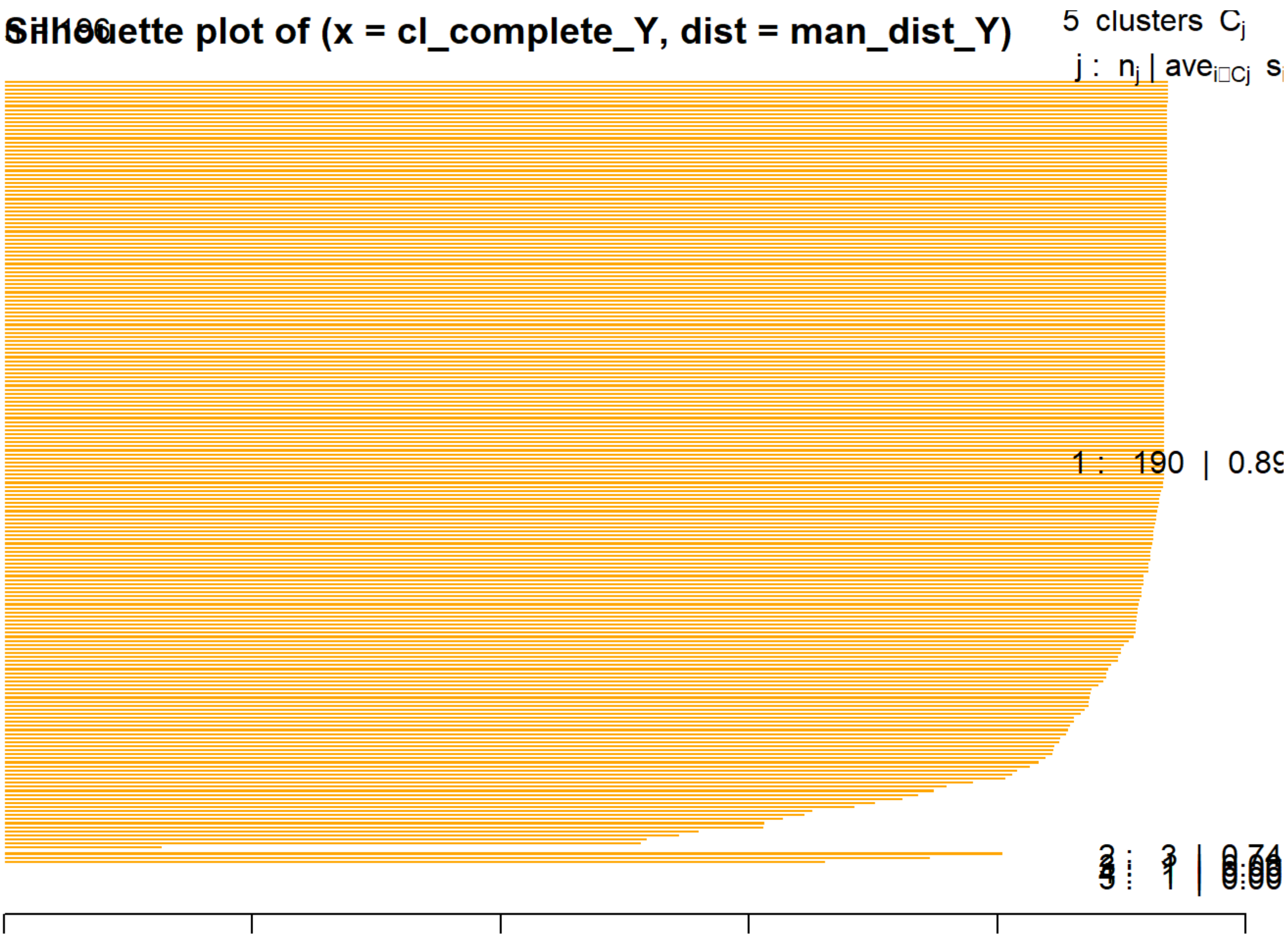
For the hierarchical clustering methods, we use all of the predictors since our data is a non-singular matrix (n>p)). Calculations for the first 4 PCs were done for some methods out of curiosity, but will not be shown here (they were shown previously in the presentation).

Hierarchical clustering did not produce solutions consistent with the groups (continents) in the data. Instead, other combinations between countries and continents were produced based on other criteria (sociological, economical, other…). Most of the time, classifications for hierarchical clustering did not seem to hold any significant meaning.

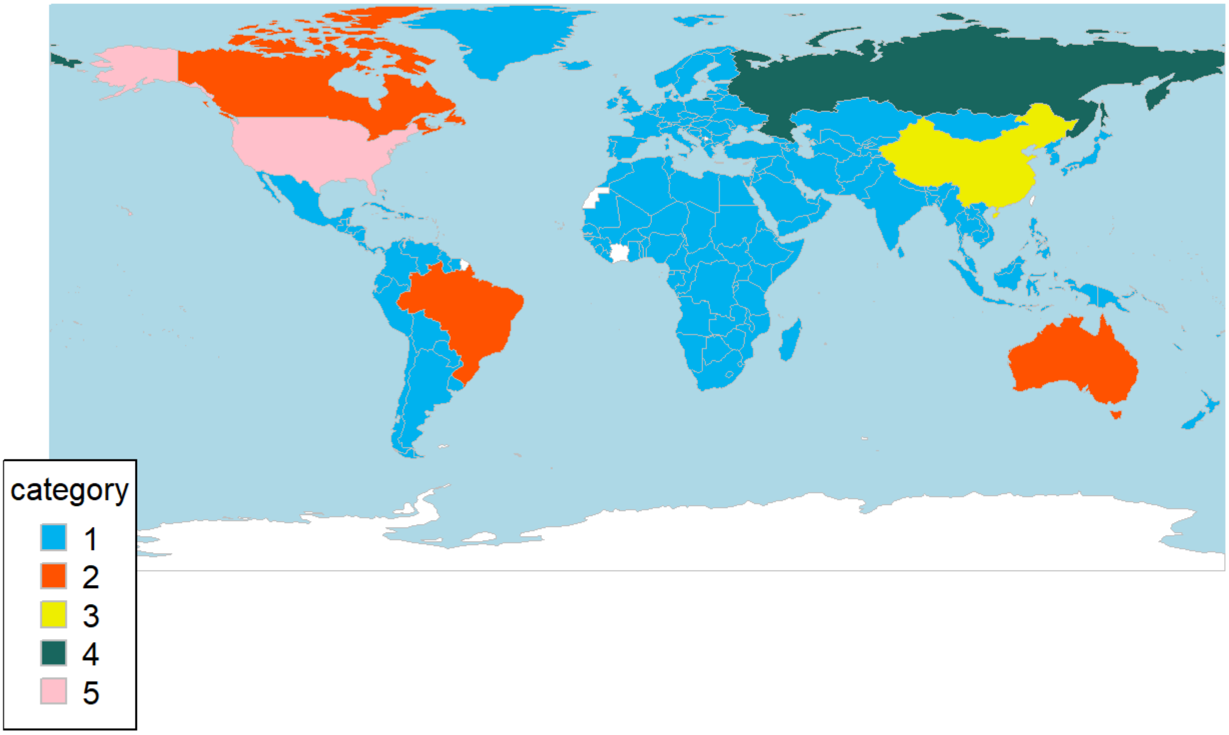
Complete Linkage

Five groups were produced but it is not completely clear what they represent. We show the results below for all of the predictors.





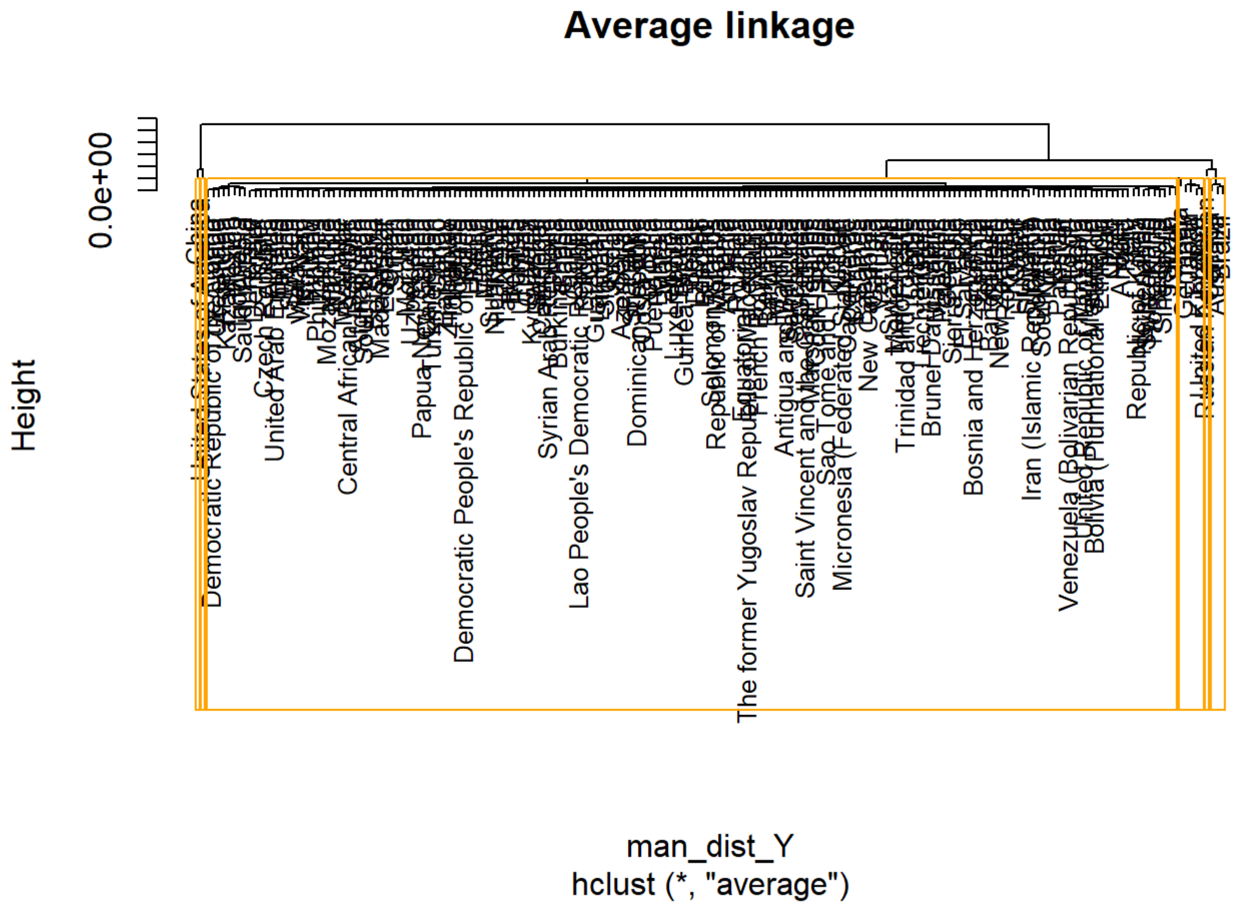
```
## Error in image.default(iy, ix, t(iz), xaxt = "n", yaxt = "n", xlab = "", : must have one m
ore break than colour
```

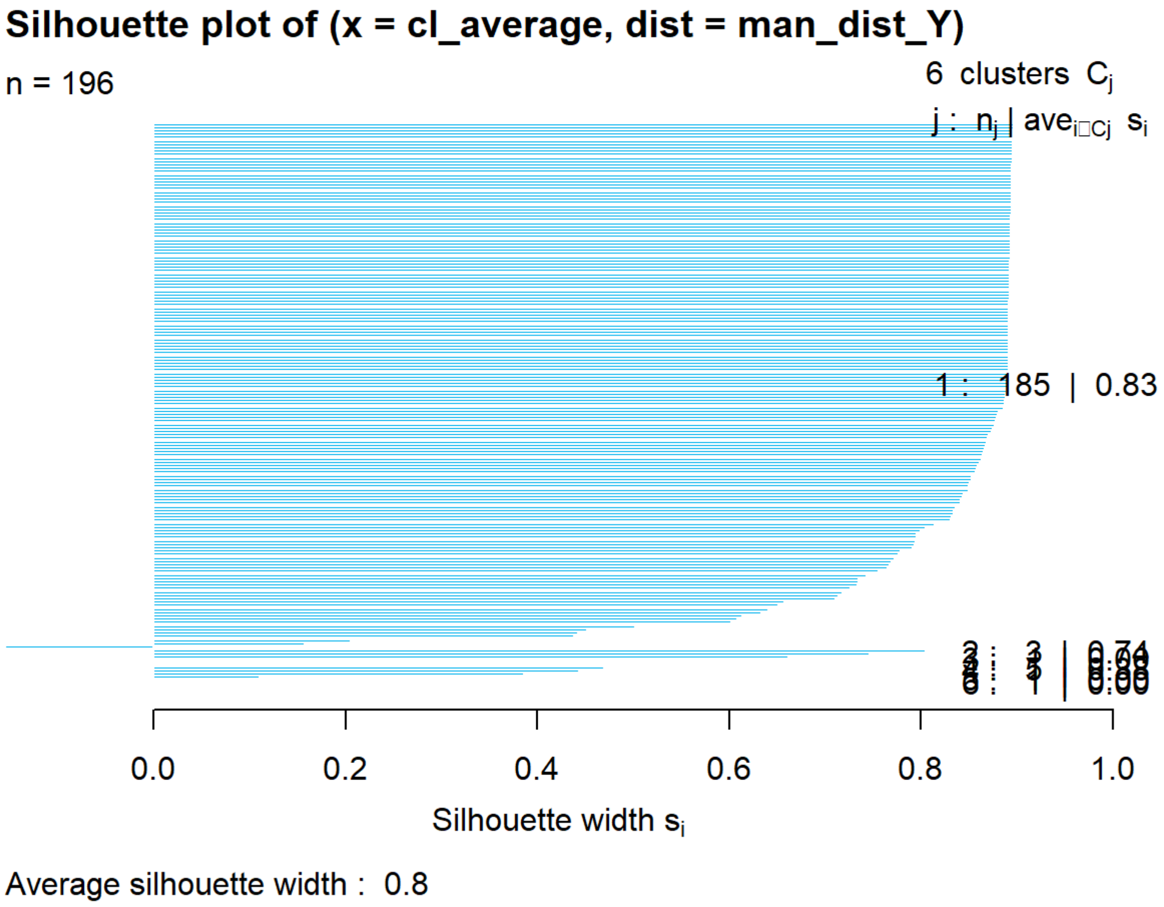


Hierarcical Clustering: Complete Linkage -- ALL PREDICTORS

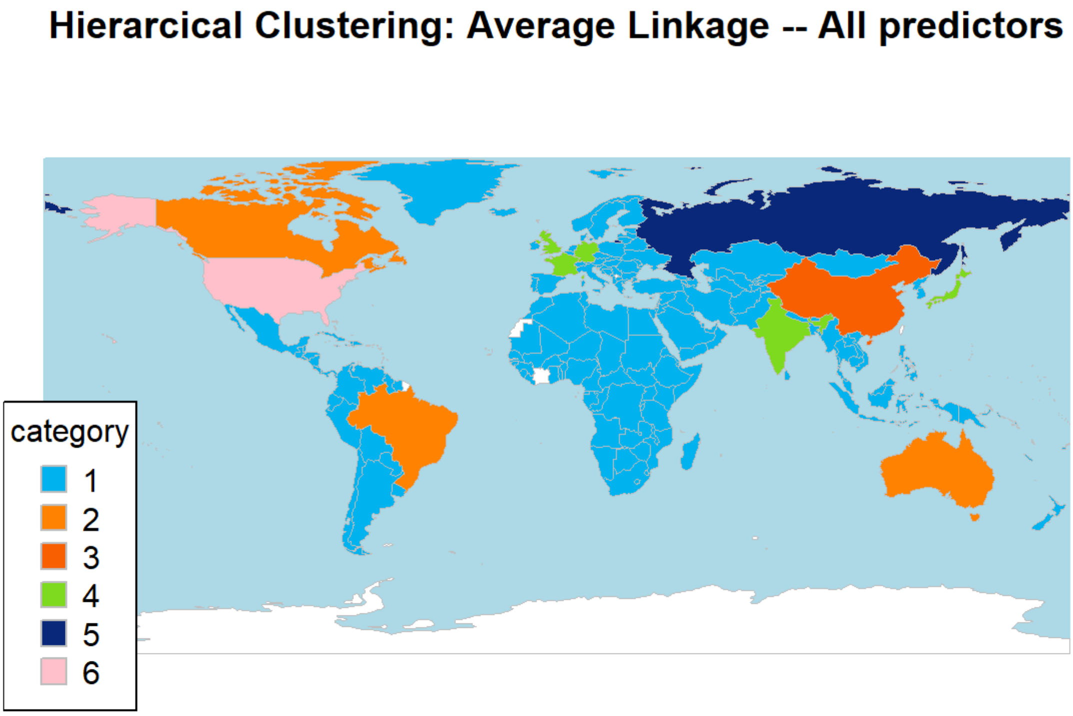
Average Linkage

Average linkage was performed on all of the predictors as well. Six very uneven groups were produced. The solution overall does not appear too interesting or insightful.



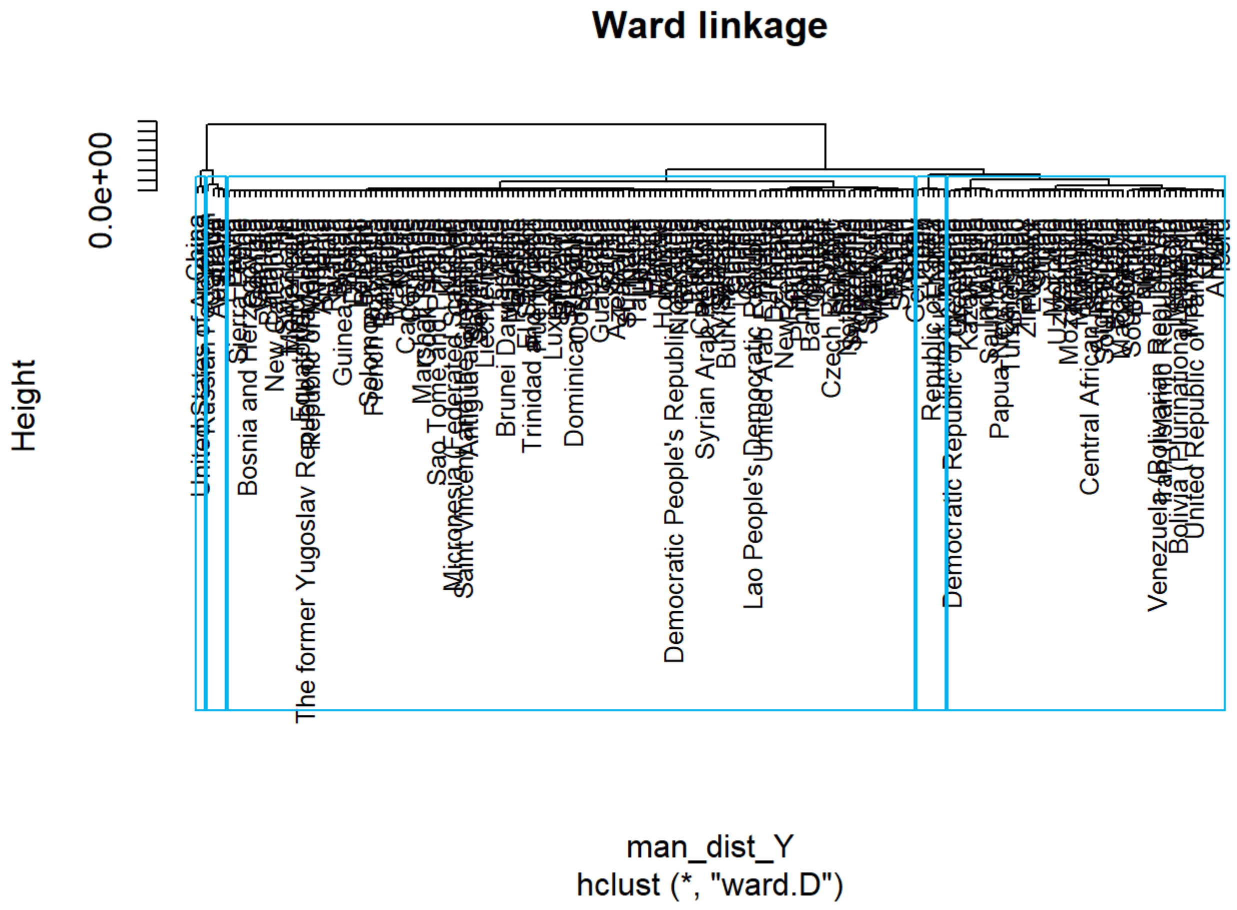


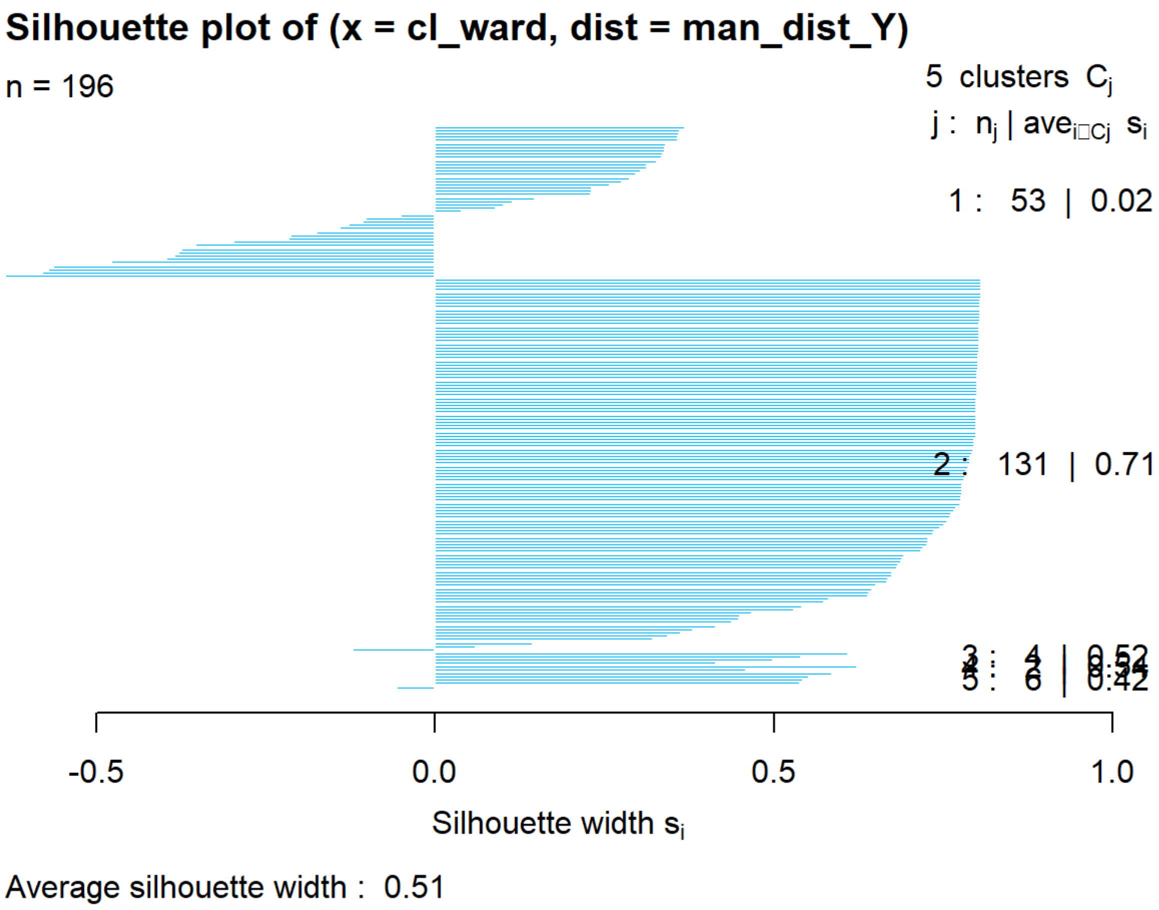
```
## Error in image.default(iy, ix, t(iz), xaxt = "n", yaxt = "n", xlab = "", : must have one m
ore break than colour
```


Hierarcical Clustering: Average Linkage -- All predictors

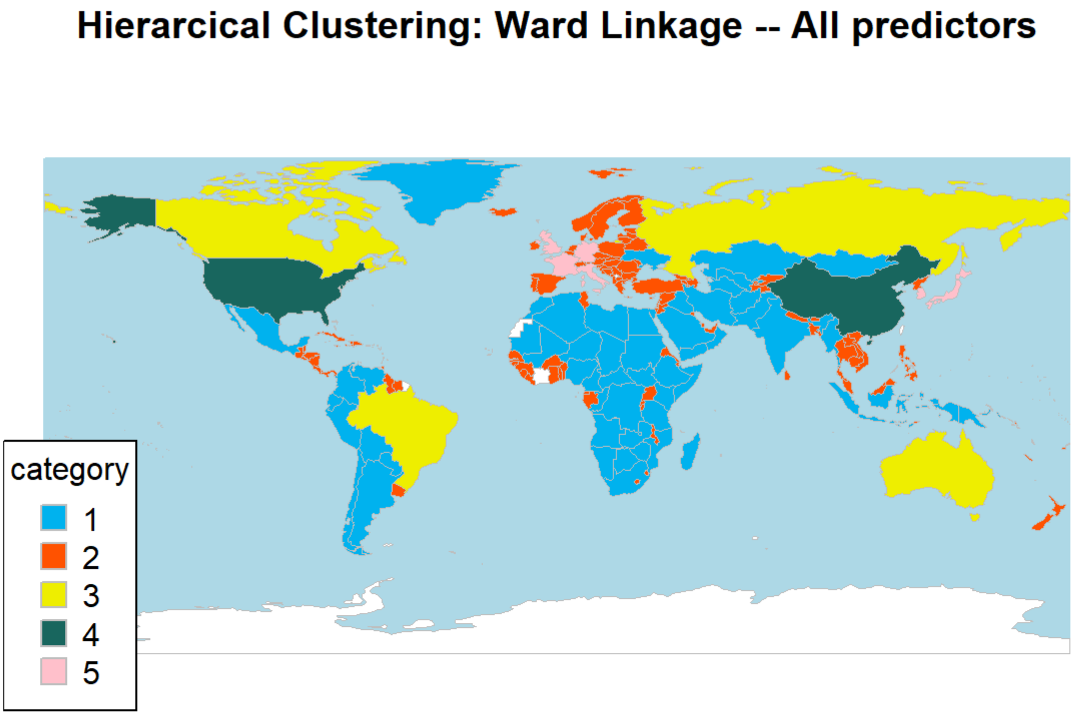
Ward Linkage

Ward linkage was performed on all predictors. The only interesting observation from this solution is that (1) the United States and China are together (green), (2) very poor countries (blue) are together, (3) countries with somewhat strong economies (yellow) are together, and (4) other countries with somewhat influential economies are together (pink). Perhaps the division of groups could represent a ranking.



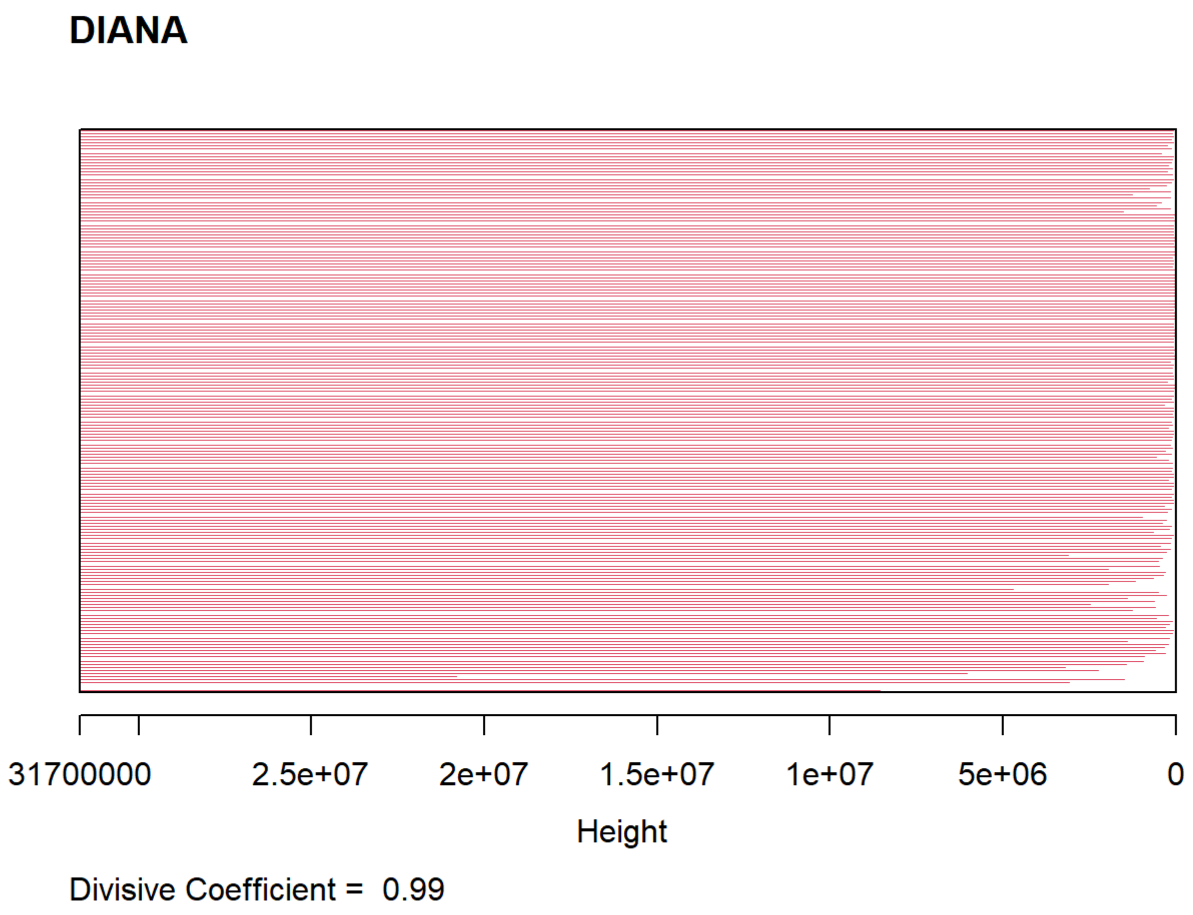


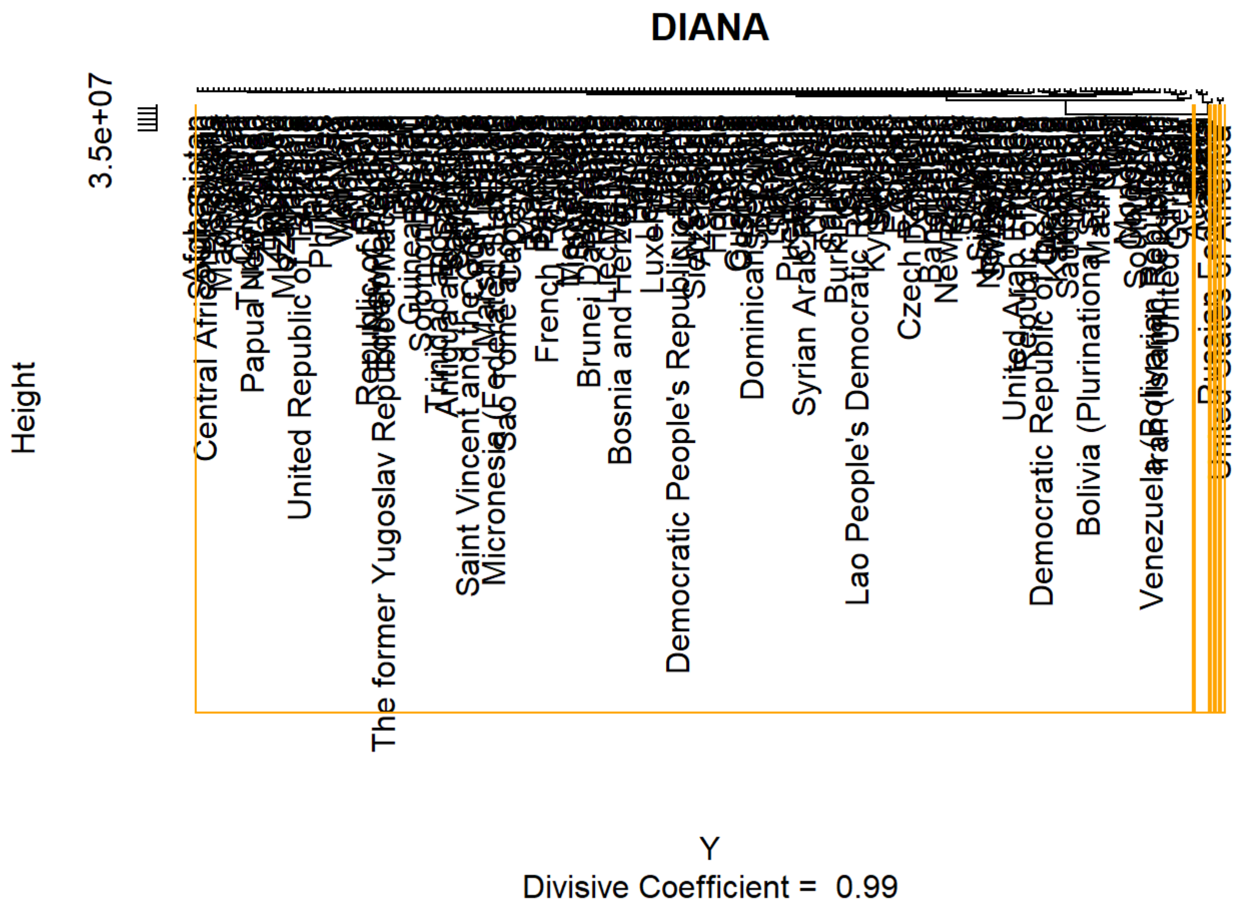
```
## Error in image.default(iy, ix, t(iz), xaxt = "n", yaxt = "n", xlab = "", : must have one m
ore break than colour
```

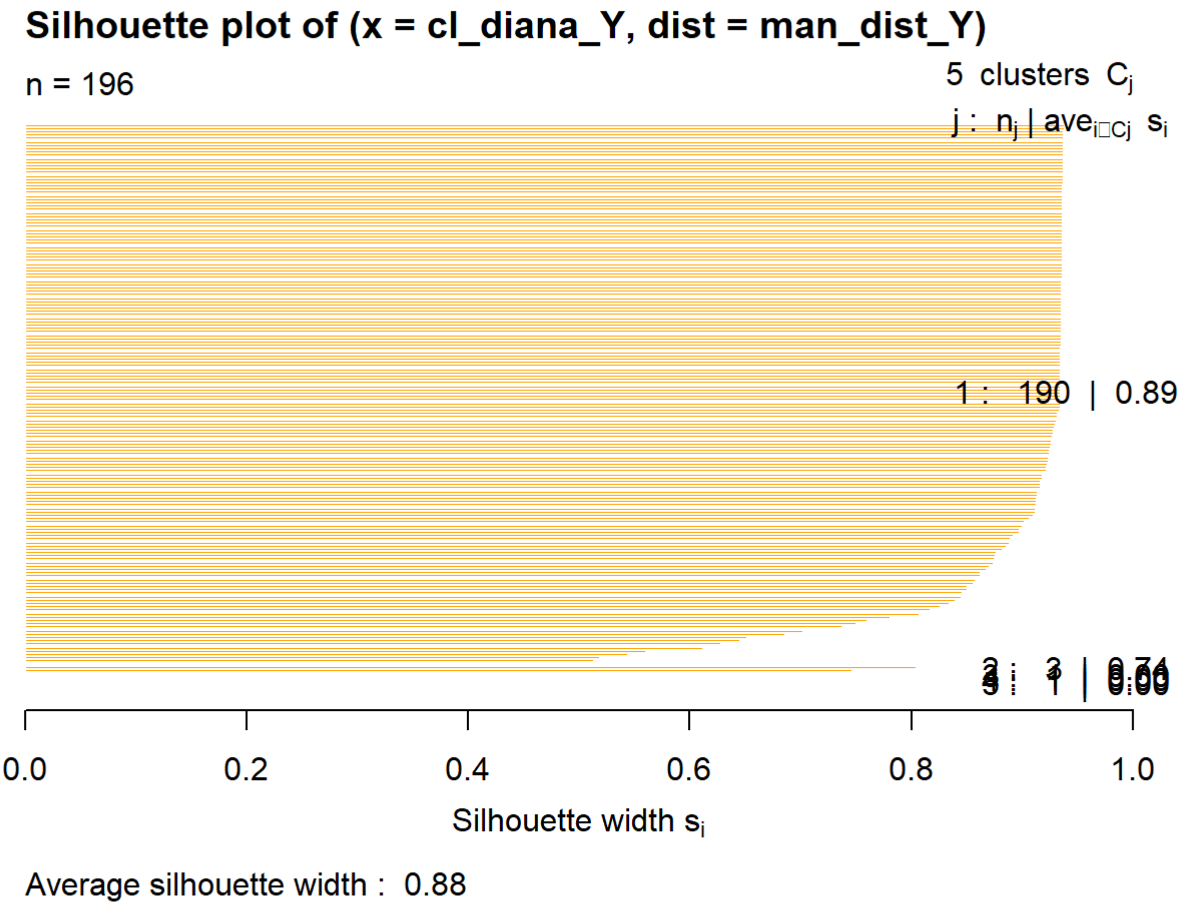


DIANA (Divisive Hierarchical Clustering)

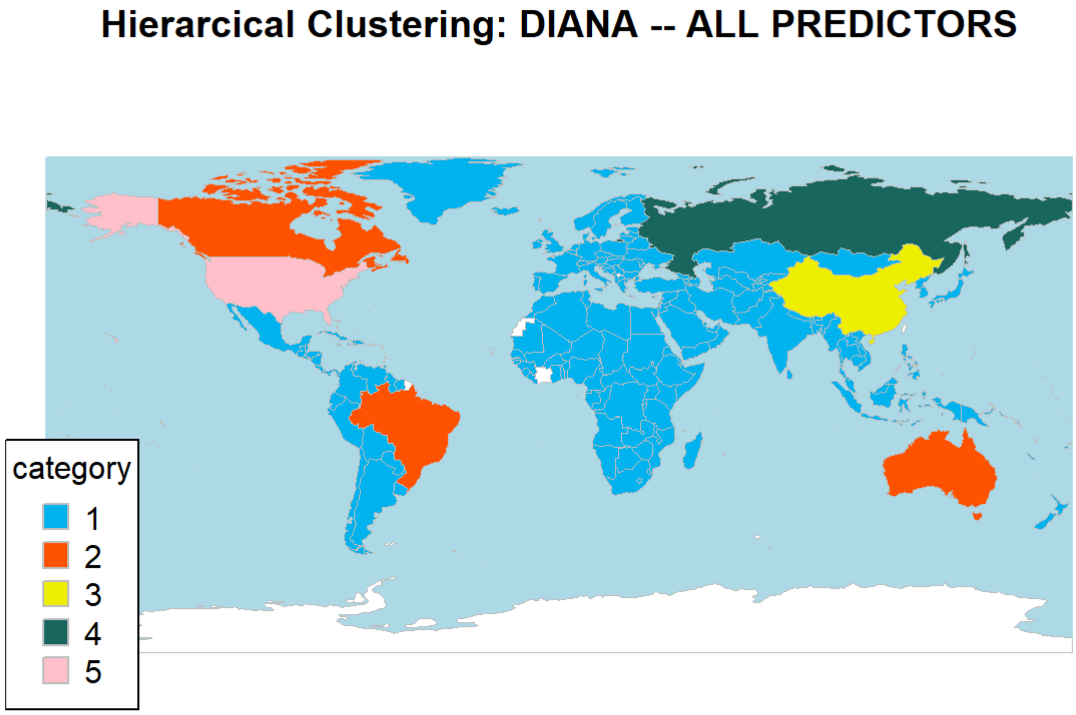
DIANA yielded 5 extremely uneven groups on all of the predictors. The solution was quite terrible with only 1 country in most groups.


DIANA
31700000
2.5e+07
2e+07
1.5e+07
1e+07
5e+06
0
Height
Divisive Coefficient = 0.99


DIANA
3.5e+07
Height
Afghanistan
Central Afri
Papua N
United Republic of
The former Yugoslav Rep
Republic of
Saint Vincent and the
Micronesia (Federated
Sao Tome and
French
Brunei Da
Bosnia and Herze
Luxe
Democratic People's Republic
Dominica
Syrian Arab
Lao People's Democratic
Czech
United Arab
Democratic Republic of
Bolivia (Plurination
Venezuela (Bolivarian
Russian Fed
United States of America
Y
Divisive Coefficient = 0.99

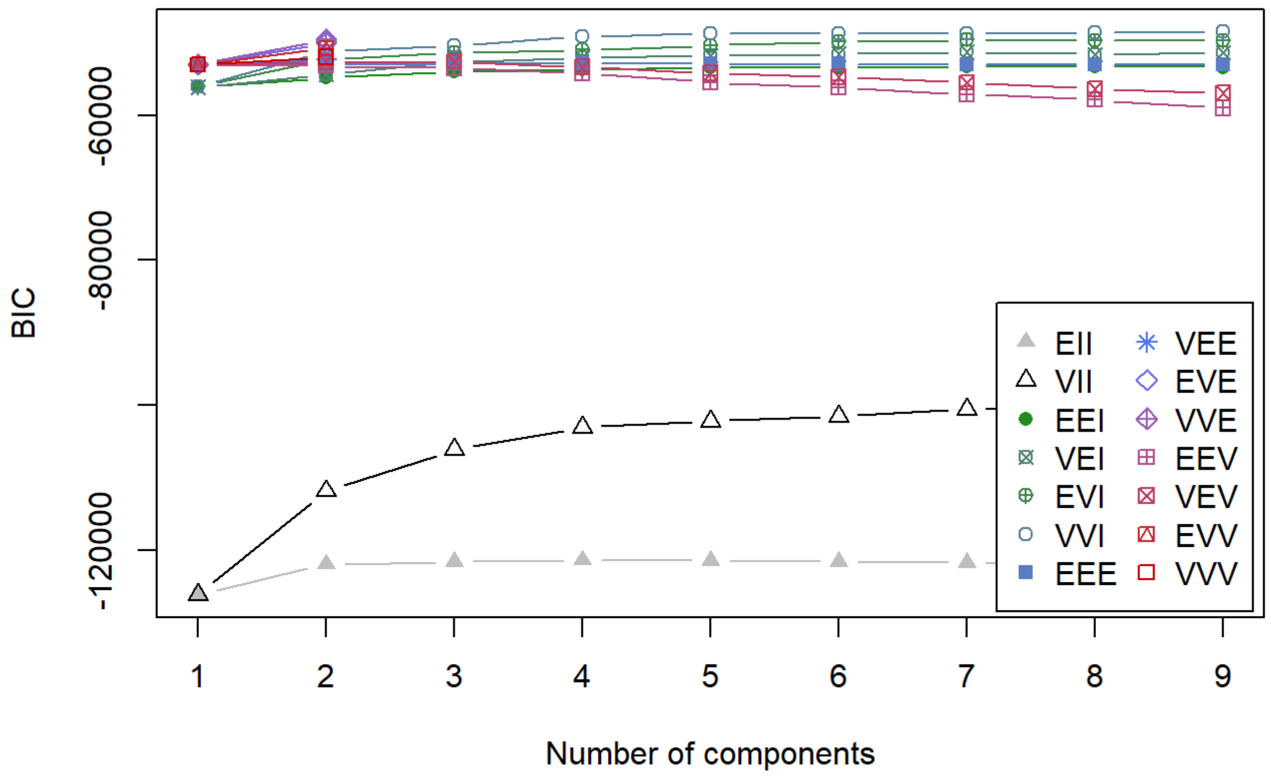


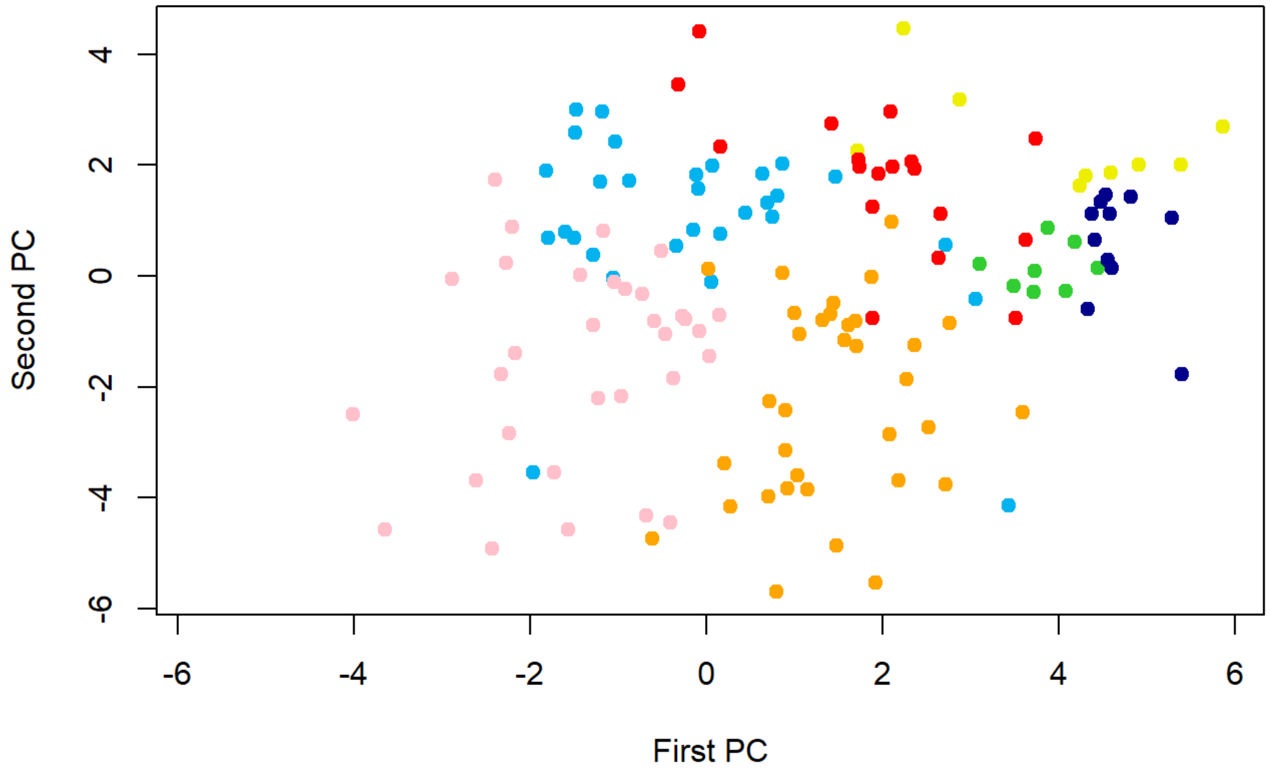
```
## Error in image.default(iy, ix, t(iz), xaxt = "n", yaxt = "n", xlab = "", : must have one m
ore break than colour
```



Model-Based Clustering: MCLUST

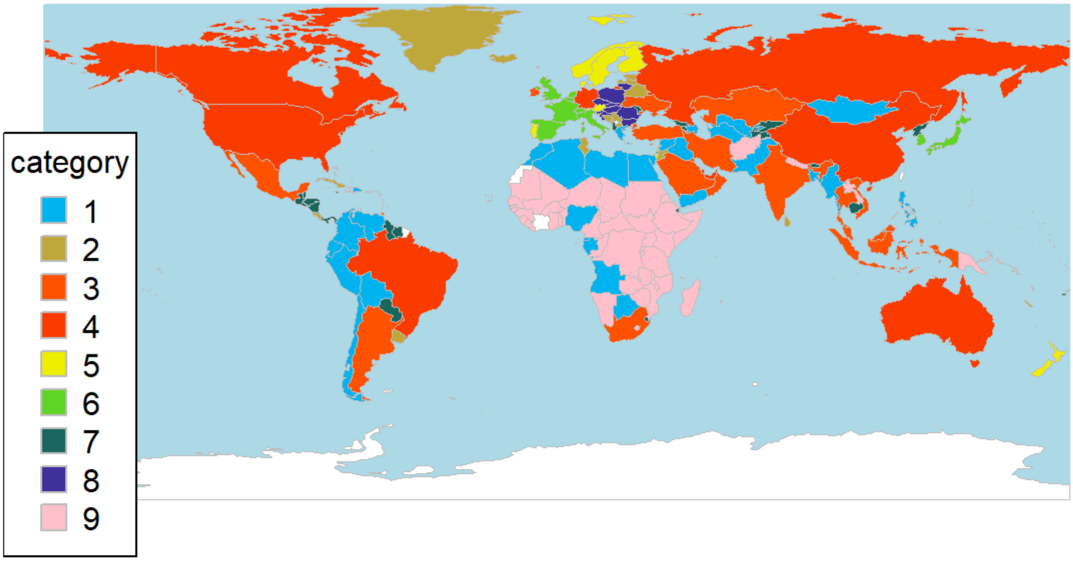
The MCLUST algorithm was performed on all of the predictors. Overall, it did not yield an interesting solution besides the fact that groups appear to be (somewhat) classified by economic activity (again). Just as in ward linkage, the groups could be a ranking by economy size. The best performing model for our dataset was diagonal with varying volume and shape.




First and Second PCs

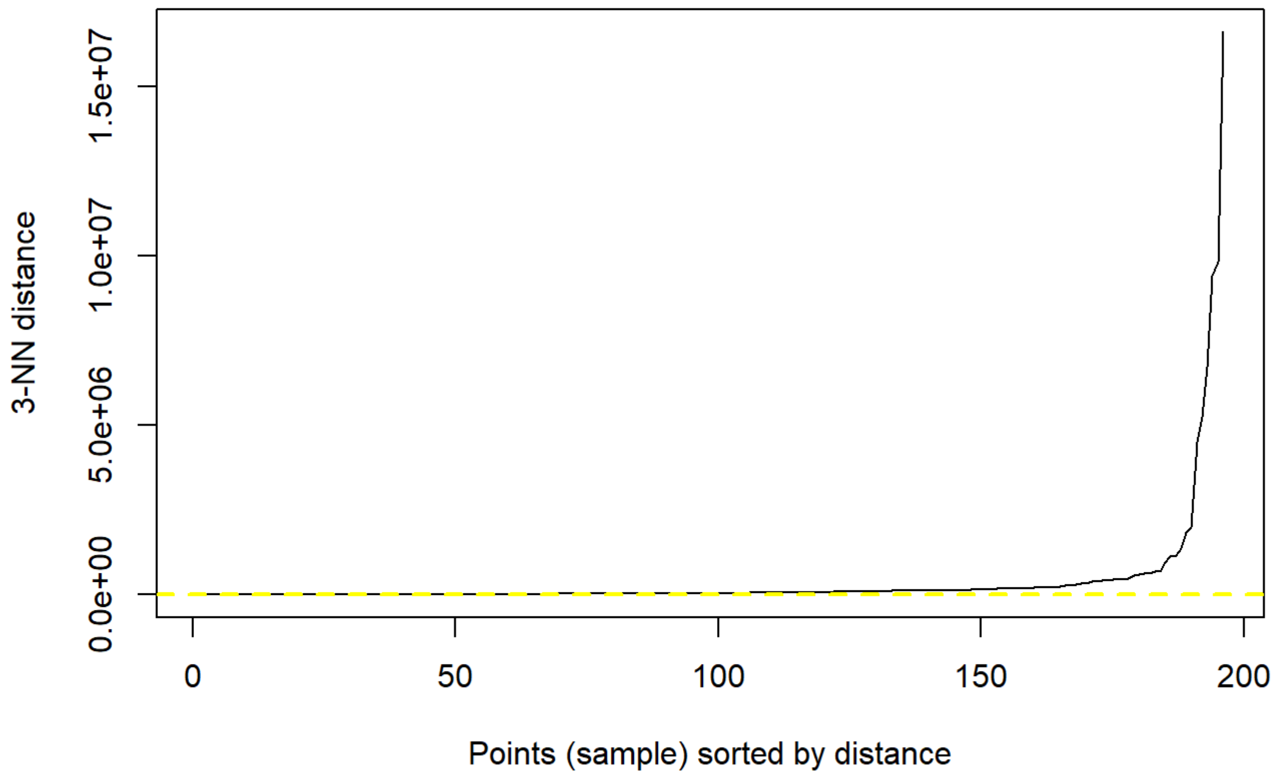
```
## Error in image.default(iy, ix, t(iz), xaxt = "n", yaxt = "n", xlab = "", : must have one m
ore break than colour
```

MClust groups-- all predictors

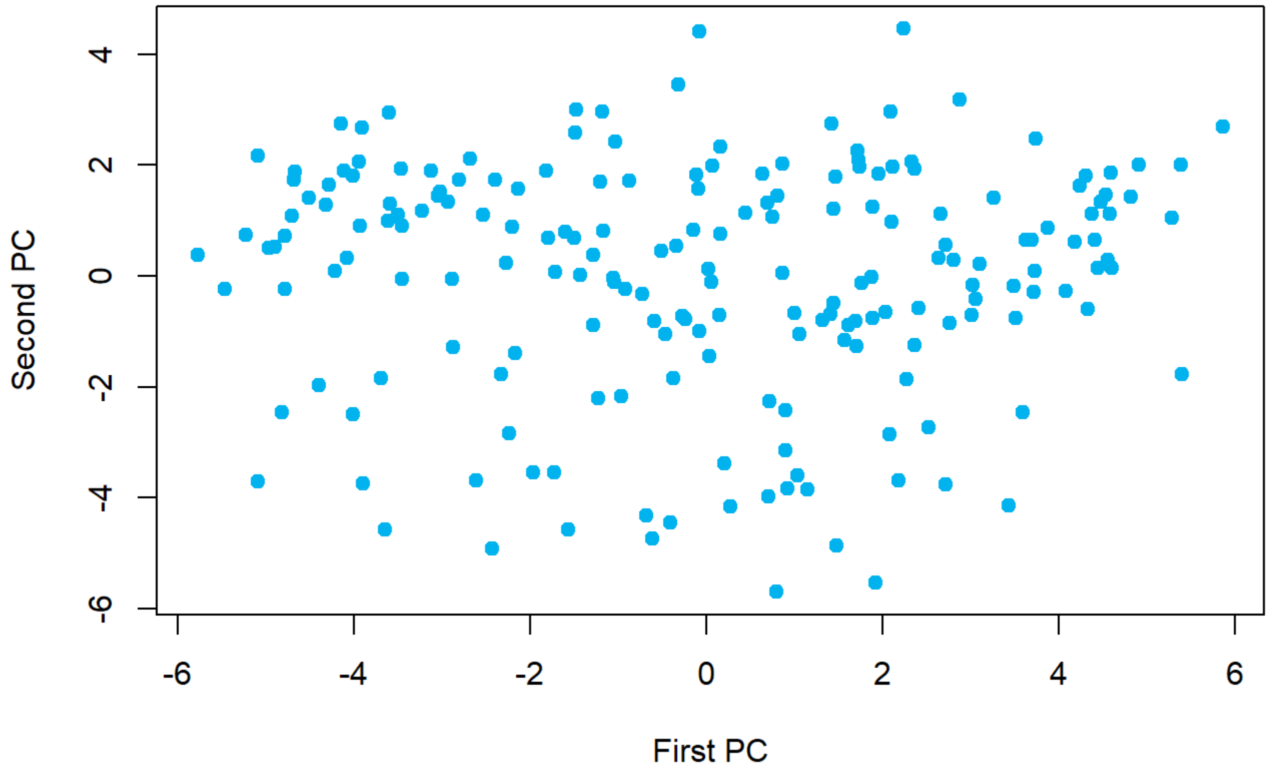


Density-Based Clustering: DBSCAN

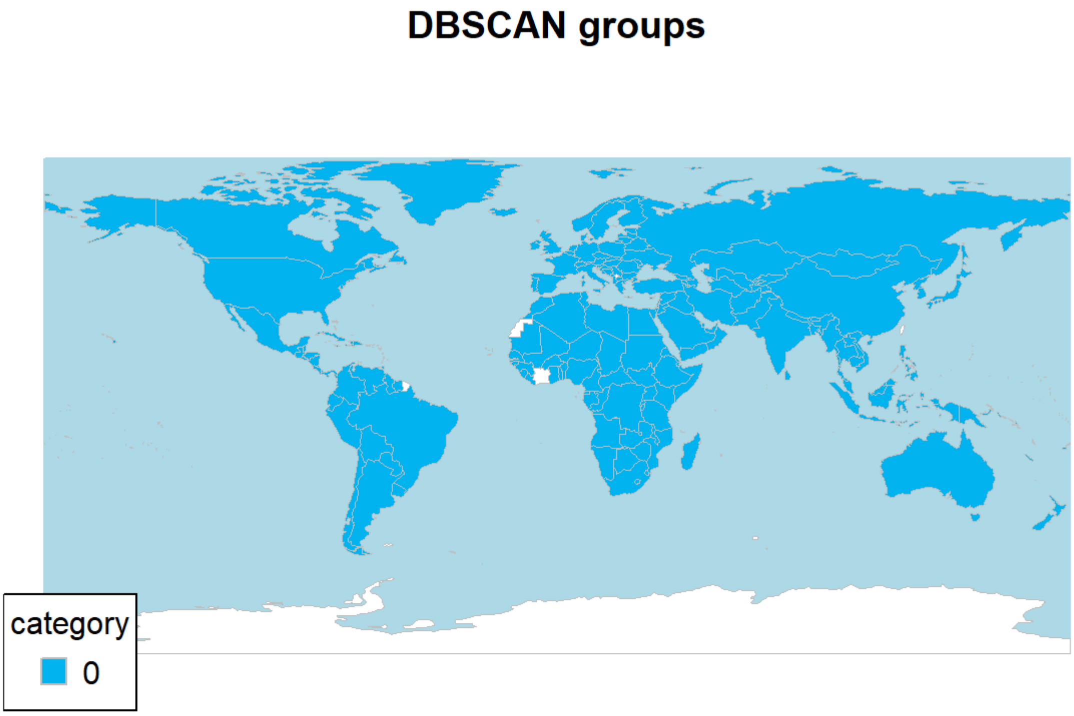
The DBSCAN algorithm was applied to all of the predictors with 4 minimum points and epsilon equal to 20. The first time running this algorithm, North America, Australia, Europe, Japan, New Zealand, and South Korea were found grouped together. Very strangely, the second time running the algorithm yielded only one group.



First and Second PCs



```
## Error in image.default(iy, ix, t(iz), xaxt = "n", yaxt = "n", xlab = "", : must have one m
ore break than colour
```

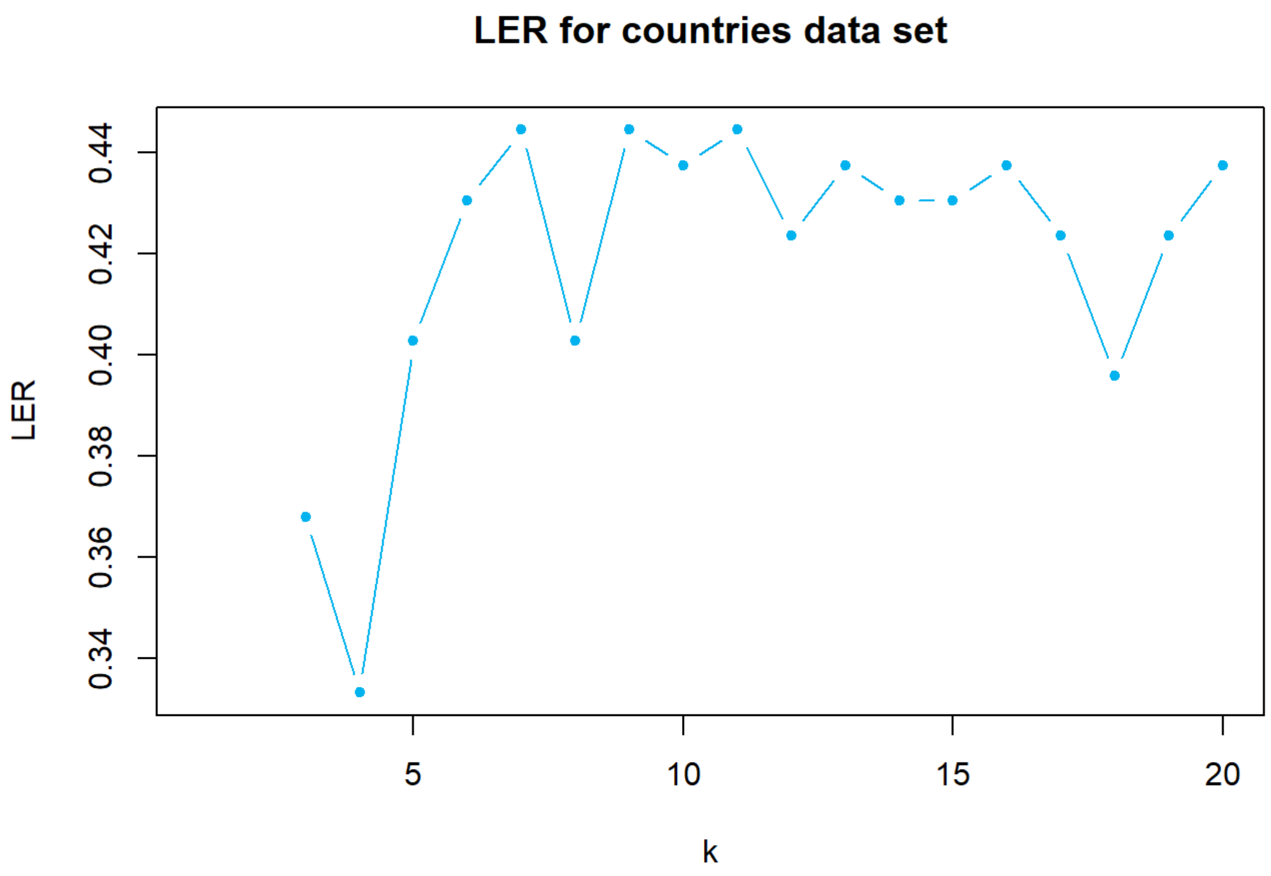


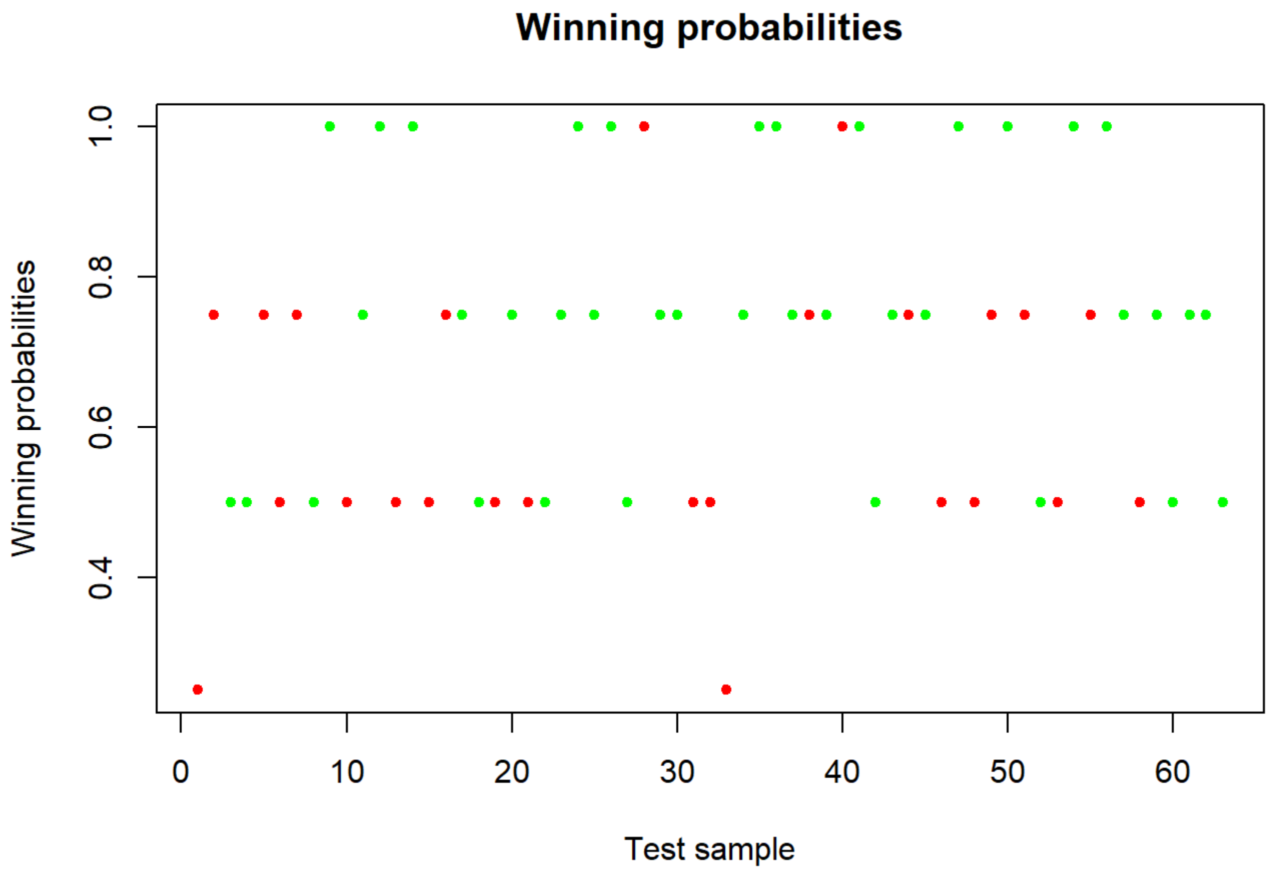
Supervised Classification

For supervised learning, we found the problem of North America having only a few countries, therefore, we were dealing with a partially unbalanced dataset. To solve that problem, we performed oversampling with the racog function from the “imbalance” package until North America had 15 countries (almost the same as Oceania).

KNN

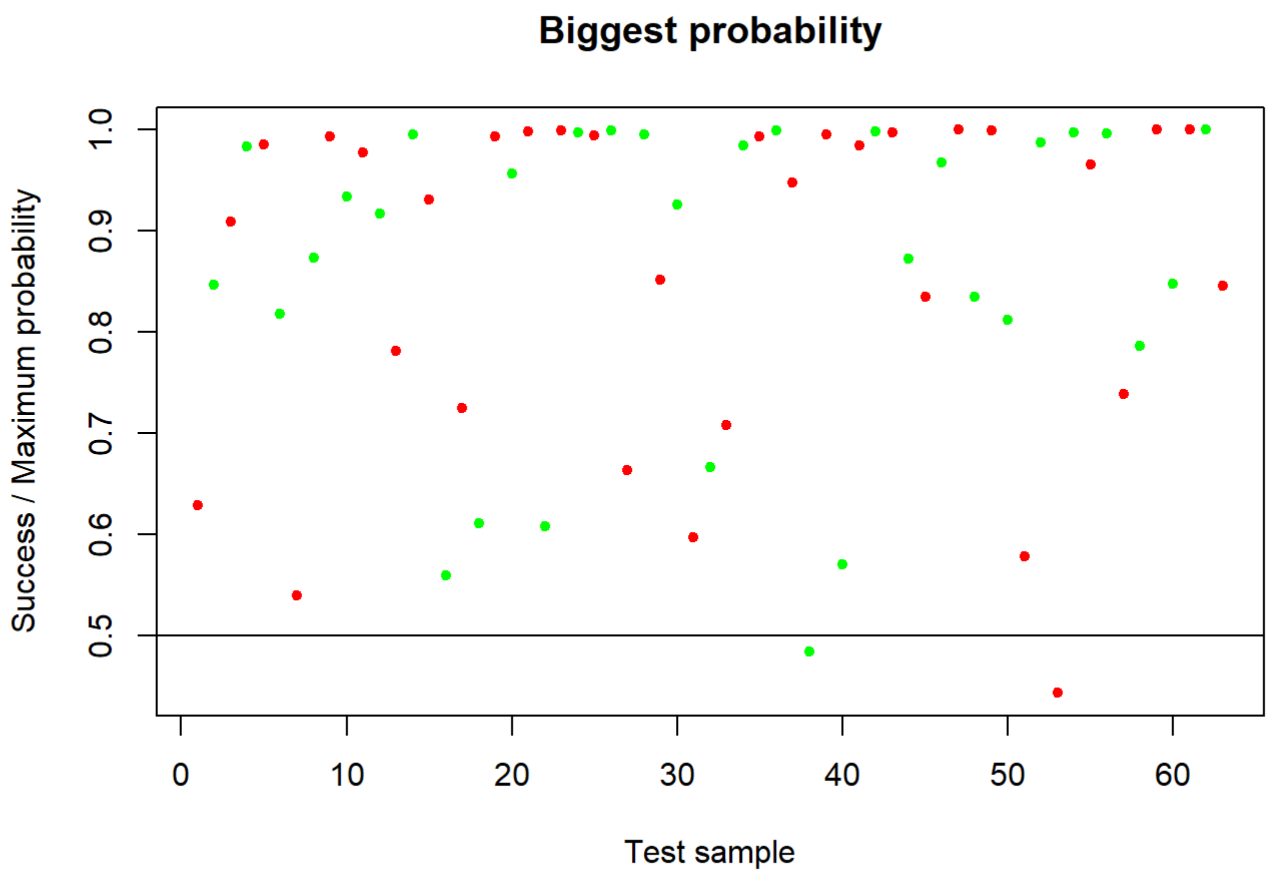
After splitting into train and test sets, we found that the optimal k was 6. After running the algorithm a few more times, we ended up with an optimal k=5. We ended up with a test error rate of 0.4285 the first time. This changed on later executions.


LER for countries data set
LER
0.34
0.36
0.38
0.40
0.42
0.44
5
10
15
20
k


Winning probabilities
Winning probabilities
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
Test sample

Linear Discriminant Analysis

LDA yielded a test error rate of 0.2539, making it almost twice as effective as KNN for this dataset. This changed on later executions.



Quadratic Discriminant Analysis

We do not include a QDA analysis here, as the following line:

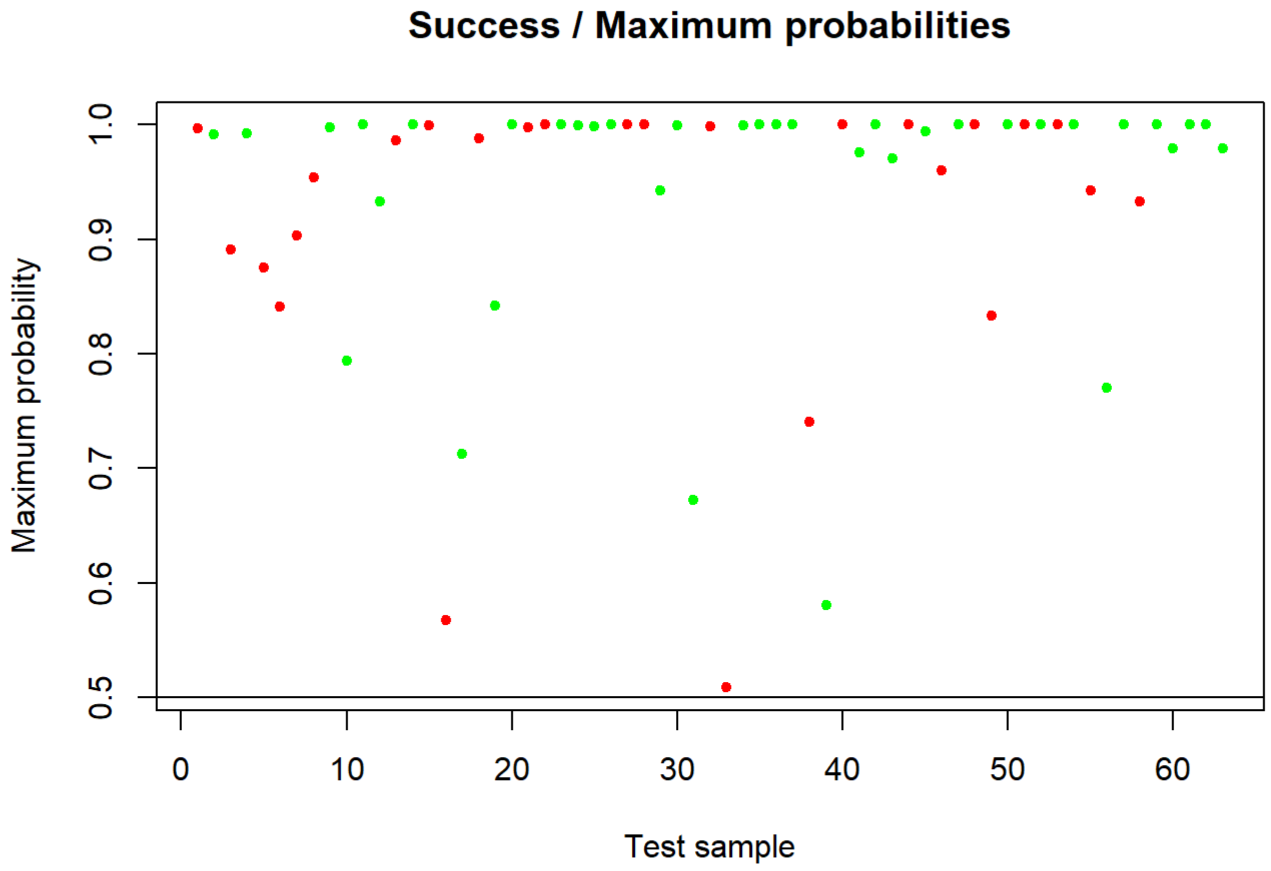
qda_train <- qda(Y_train~.,data=as.data.frame(X_train))

gives the following error: Error in qda.default(x, grouping, …) : some group is too small for ‘qda’

We do not perform QDA because at least one group in our dataset is too small, even after oversampling.

Naive Bayes

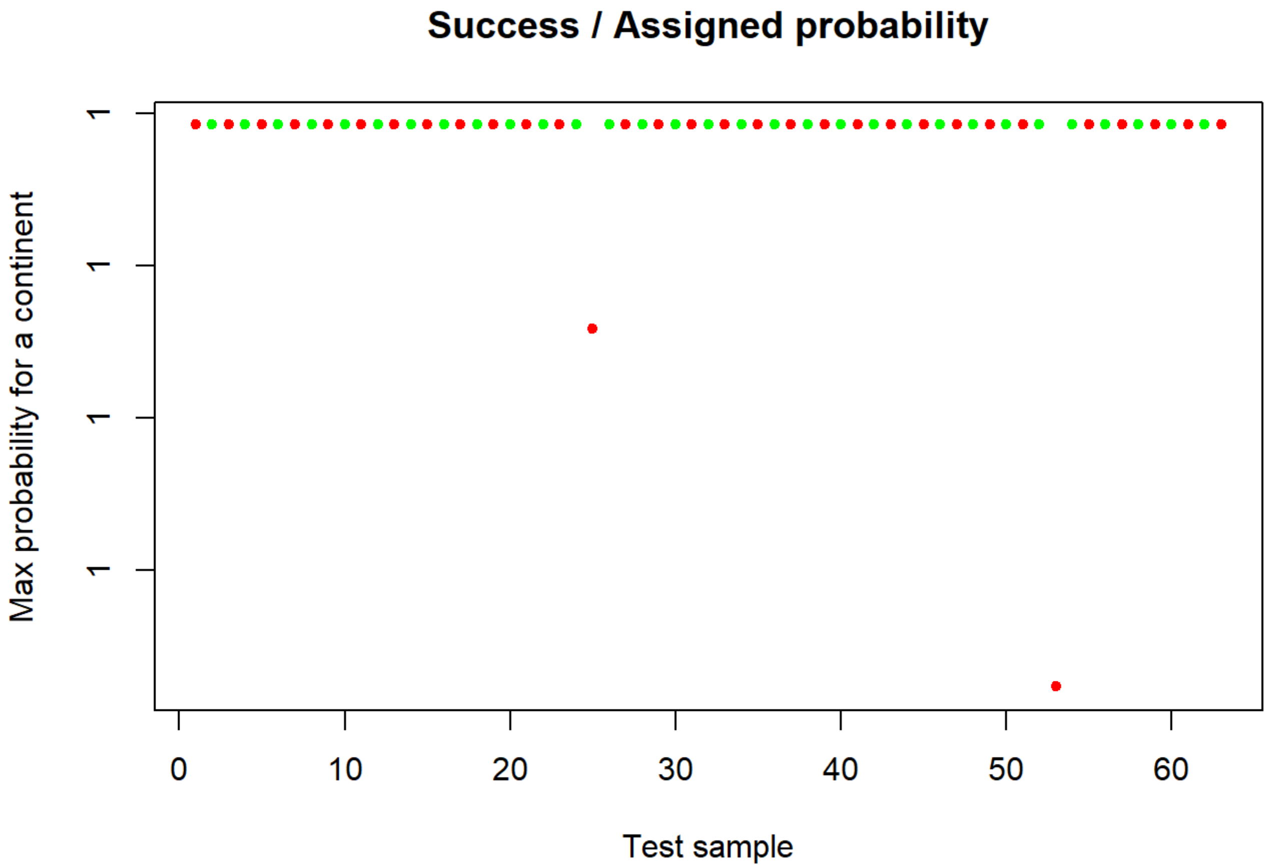
Naive Bayes yielded a test error rate of 0.4285. This changed on later executions. Below in the code, a contingency table with good and bad classifications and conditional probabilities for classifications made with the test sample can be seen.



So far, Naive Bayes (test error rate: 0.4285) performs better than KNN (test error rate: 0.4285) and worse than LDA (test error rate: 0.2539).

Logistic Regression

Logistic regression was performed using the multinom library for fitting multinominal log-linear models via neural networks. A model selection was then done using MASS::stepAIC, which appeared effective in eliminating predictors that did not contribute to the accuracy of the model. It’s test error rate was 0.4761. This changed on later executions.



Supervised Classification: Results

Method	Test_Error_Rate
KNN	0.3968254
QDA	0.3174603
Naive Bayes	0.4126984
Logistic Regression	0.3650794

Final Conclusions

The main conclusion of this study is that there are not clear groups in the dataset. In some other examples during the course we saw how there was a mostly clear separation between observations (i.e. cancer cells) defining groups. In our case, we have a continuous set of countries with the presence of few important outliers. In addition, if we try to group the countries in an unsupervised way, the groups produced are not related to geographical criteria such as the continent, but rather to economic and sociological factors.

As a result of this, it’s not easy to predict the continent of a new country, although, the TER when applying LDA was not extremely high (0.25). Also, despite grouping by continent not appear to be a natural classifier for countries, we can get some new insights on this topic. As we just said, a better way to divide countries in groups is attending to economic and sociological indicators. As we also saw in PCA, factor analysis, and KNN, there seem to be two more characteristics to perform this clustering:

- index of development
- size of the country’s economy

It is also important to point out that many of the variables we studied were highly skewed (i.e. GDP, surface area in square km,…), which indicates that these features are not normally distributed among countries. On the contrary, in the case of surface area for example, most of the countries are small, but there are a few that are quite large. For this reason, we think that some groups are consistent throughout different methods such as the USA, Canada, Europe, and Australia (for results calculated on the first 4 PCs, particularly in hierarchical clustering).

We would also like to point out that some of the features of this dataset were not very well chosen/put together since we encountered serious problems of dependency and multicollinearity that we had to overcome by selecting fewer of the features given in the original dataset.

We hope you enjoyed looking over this analysis thank you for reading.