Decomposing and clustering World Economic Freedom

Abstract

Economic freedom is the fundamental right of every human to control his or her own labor and property. In an economically free society, individuals are free to work, produce, consume and invest in any way they please. To understand better the concept of economic freedom, we use data provided by the Heritage Foundation, consisting in 12 indexes related to economic freedom in more than 185 countries. After a descriptive analysis of the dataset, we use Principal Component Analysis to reduce the dimensionality of the data and gain some intuition about the performances of each country. After that, we run k-means algorithm and hierarchical clustering to group together countries characterized by similar level of economic freedom.

Data definition

Heritage Foundation provides official and reliable sources for measuring economic freedom. More specifically, the Foundation provides 12 indexes, grouped into four broad categories of economic freedom:

1. **Rule of Law**
    - *Property rights*: it assesses the extent to which a country's legal framework allows individuals to acquire, hold, and utilize private property, secured by clear laws that the government enforces effectively. The more effective this legal protection is, the higher a country's score will be.
    - *Government integrity*: it assesses the extent of corruption within government institutions and decision-making. The higher the corruption, the the smaller a country's score will be.
    - *Judicial effectiveness*: it assesses the judicial independence, the quality of judicial processes and eventual favoritism in obtaining judicial decisions. The higher the index, the higher the judicial effectiveness of the considered country.
2. **Government Size**
    - *Government spending*: it regards the government expenditures, which includes consumption by the state and all transfer payments related to various entitlement programs. Differently from what the name could suggest, the higher the index, the lower the state's expenditure.
    - *Tax burden*: it assesses the overall level of taxation. Once again, the name is confusing, as the higher the index, the lower the level of taxation of a country.
    - *Fiscal health*: it compares the level of debt and deficit of a country with its GDP. The largest the index, the better the fiscal situation of a country.
3. **Regulatory Efficiency**
    - *Business freedom*: it measures the extent to which the regulatory and infrastructure environments constrain the efficient operation of businesses. The higher the index, the freest the country.
    - *Labor freedom*: it regards the various aspects of the legal and regulatory framework of a country's labor market, including regulations concerning minimum wages, laws inhibiting layoffs, severance requirements and measurable regulatory restraints on hiring and hours worked. High values for that index mean low intrusiveness of labor rights in the labor market.
    - *Monetary freedom*: it measures how stable are prices and the extent to which the state intervenes in the economy. The larger the index, the higher the freedom, i.e. the lower the state intervention.

4. **Open Markets**
   - *Trade freedom*: it concerns the extent of tariff and non-tariff barriers that affect imports and exports of goods and services. The higher the index, the less obstacles to trade are present.
   - *Investment freedom*: it measures the extent to which the free flow of capital is restricted by the law. The higher the index, the lower the restrictions.
   - *Financial freedom*: it is an indicator of banking efficiency as well as a measure of independence from government control and interference in the financial sector. The higher the index, the lower the level of government interference.

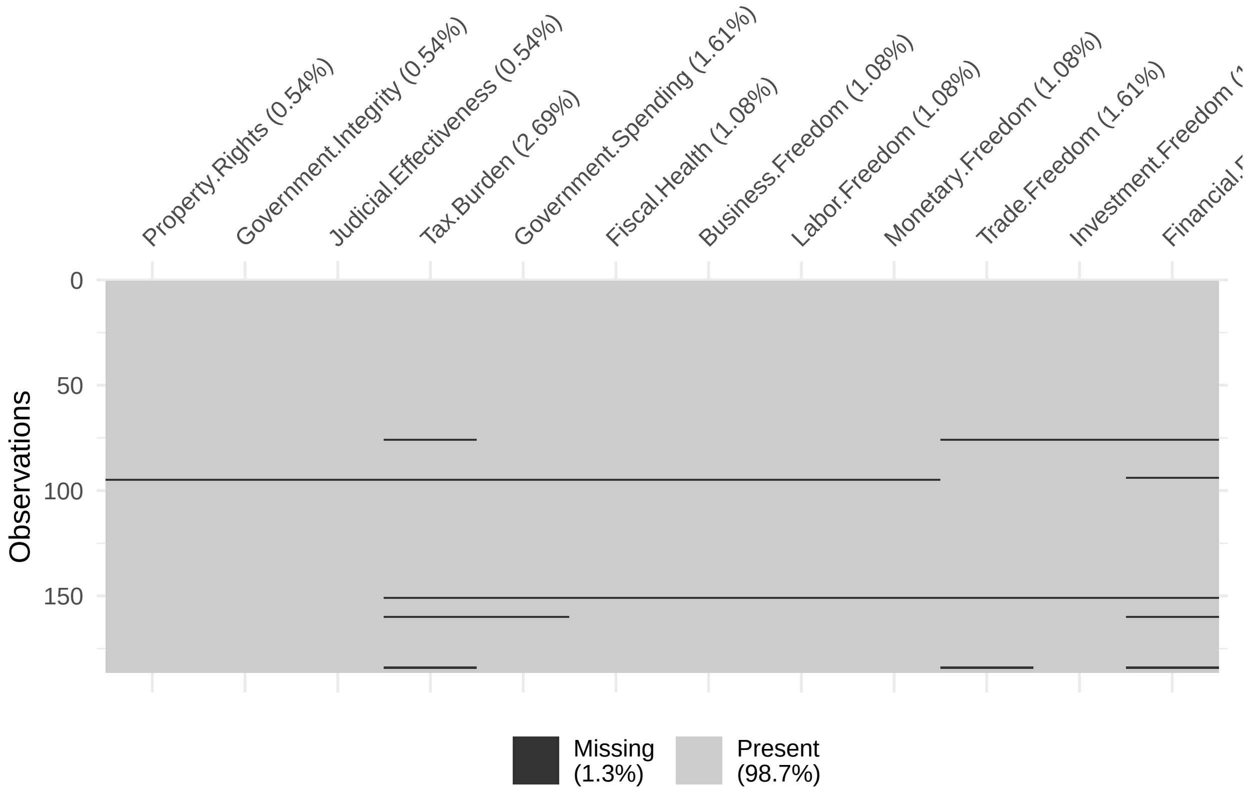
Each of the twelve economic freedom indexes is graded on a scale of 0 to 100, so there is no need to scale the variables before proceeding in the analysis.

NAs handling

Reassuringly, the number of NA values is negligible in our dataset.

```
## [1] 29
```

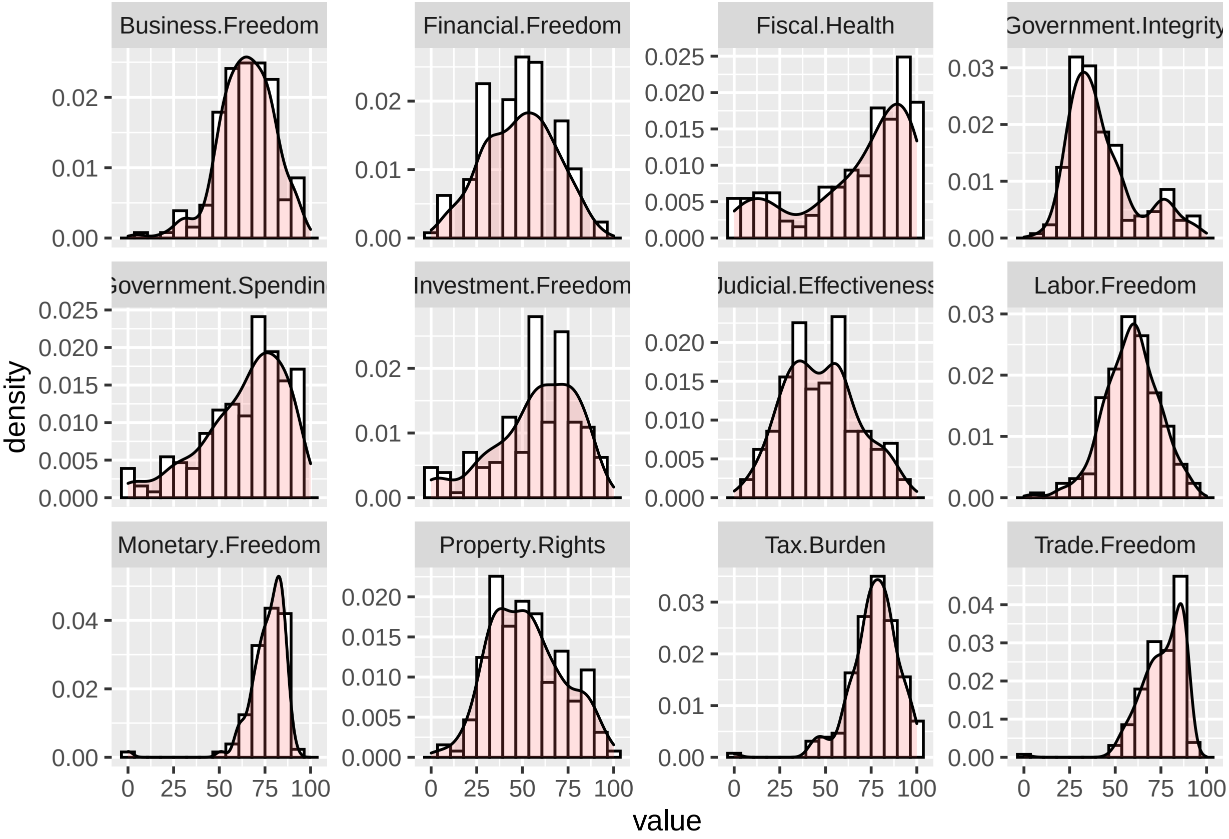
This plot shows the locations of the missing values in a given row and column.



Just 1.3% of the whole dataset is missing, which amounts to only five countries, namely Liechtenstein, Yemen, Iraq, Libya and Syria. They are missing because probably it is substantially difficult in these countries to collect the data due to internal conflict that has been going on for years or because they are too small to provide such data(i.e. Liechtenstein). We therefore decided to drop them because it is a negligible part of the whole dataset.

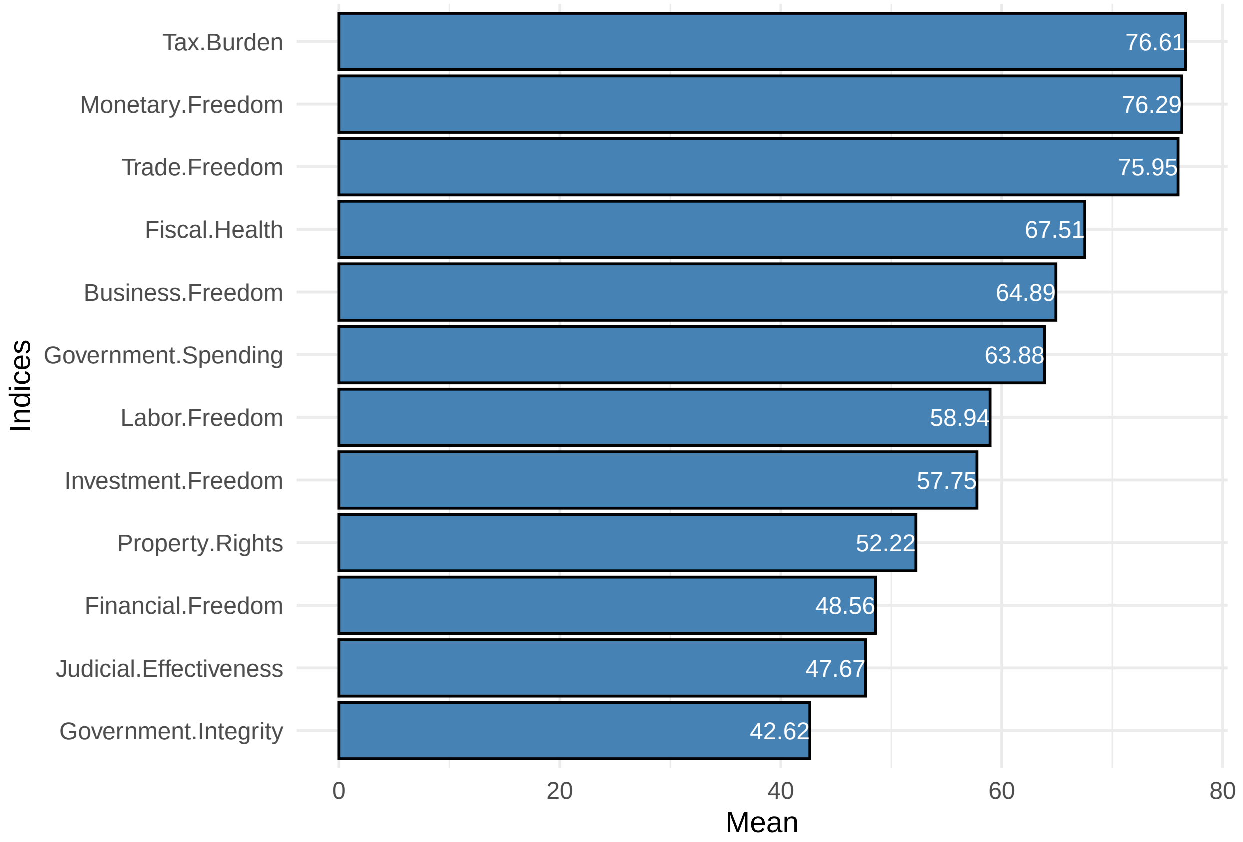
Descriptive statistics

In the following plot, we see the distribution of each index present in our dataset.



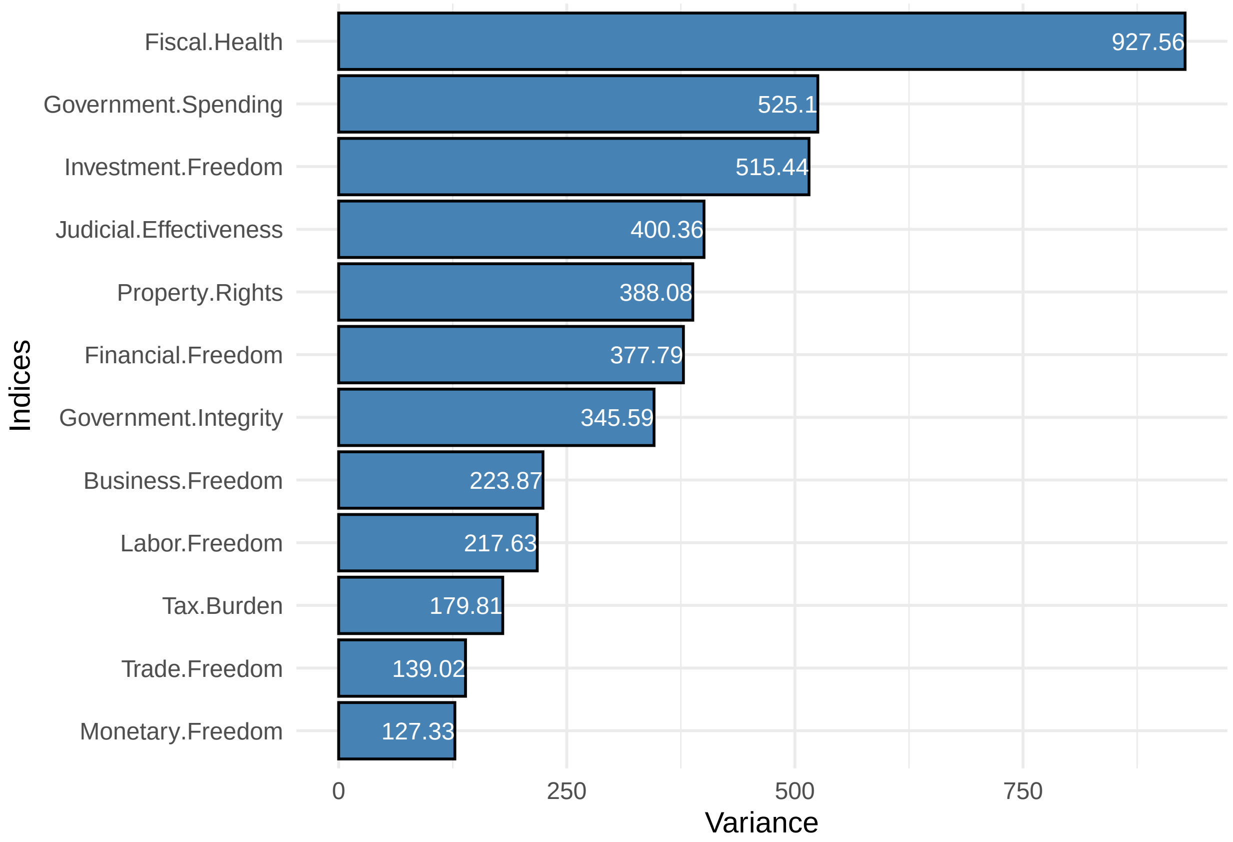
First of all, we can notice that most of the indexes are slightly skewed to the left, meaning that most of the countries have a good level of freedom overall, especially for what concerns `Trade.Freedom`, `Monetary.Freedom` and `Tax.Burden`. A similar, but less extreme, trend can be seen for `Government.Spending`, `Business.Freedom`, `Fiscal.Health` and `Investment.Freedom`. On the contrary, `Financial.Freedom` and `Labor.Freedom` shows a quasi-normal distribution. The two indexes that seem to be more uniformly distributed are `Judicial.Effectiveness` and `Property.Rights`. Finally, `Government.Integrity` is skewed to the right, which unfortunately means most of the countries show a low level of integrity.

The fact that most of the indexes are skewed to the left is confirmed by the plot below. Since all indexes vary between 0 and 100 we can directly confront them.



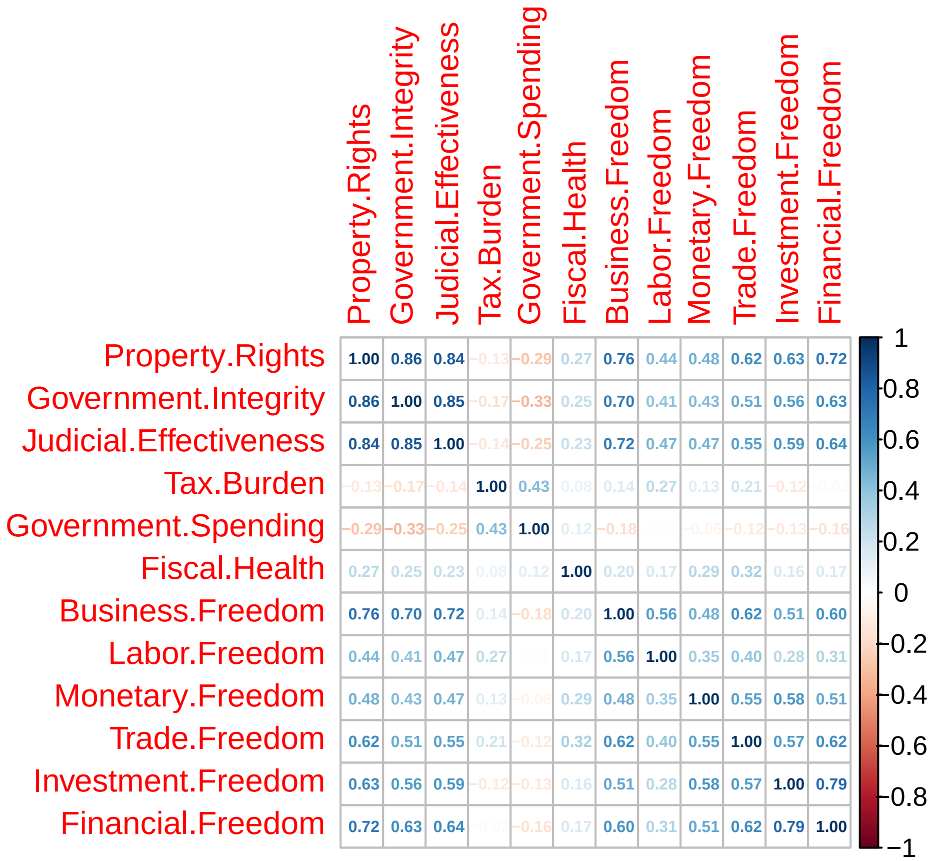
The largest means are for `Tax.Burden`, `Monetary.Freedom` and `Trade.Freedom`, meaning that, on average, countries display low taxation level, intervene little in the economy and impose few obstacles to free trade. As discussed above, `Government.Integrity` shows the smallest average value.

We can also take a look at the variability of each index across countries by plotting the associated variance.



Notice that the indexes that were on top of the previous graph, are now on the bottom: this reflects the fact that most of the countries share the same position about that three particular aspects of economic freedom. `Fiscal.Health` is by far the first index in terms of variance: this means that the fiscal situation varies significantly across the world.

Finally, we take a look at eventual multicollinearity among the indexes.



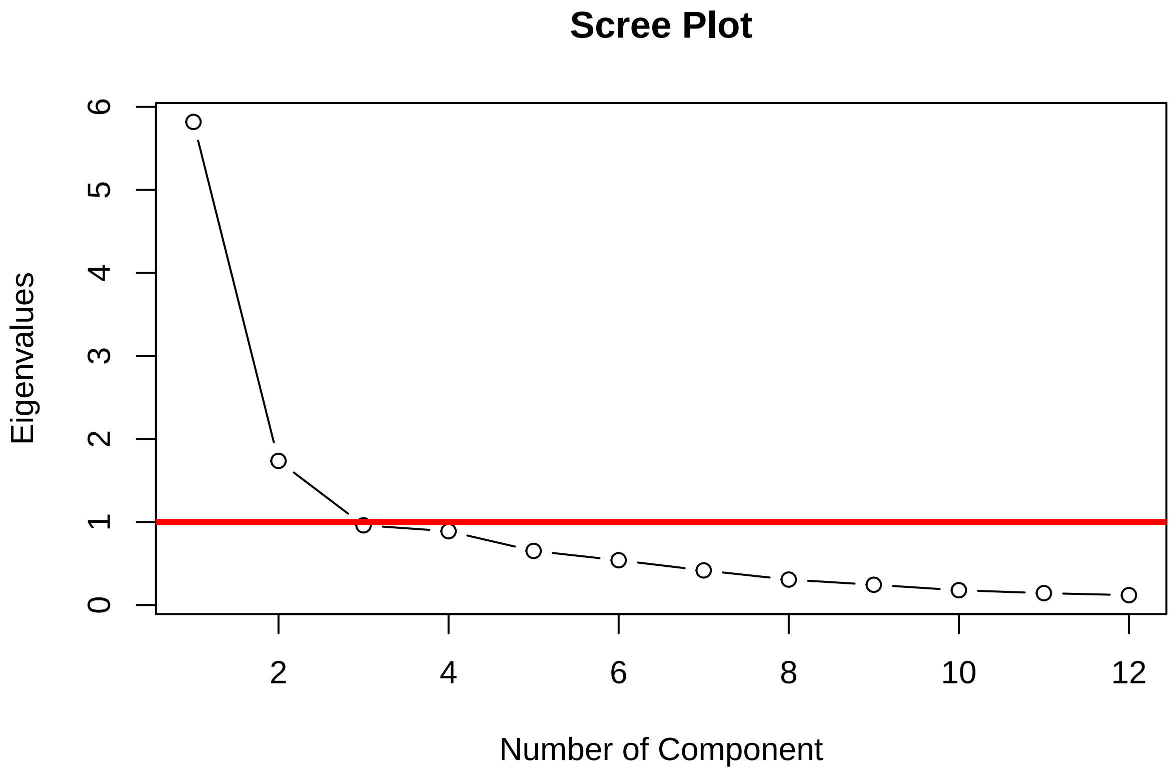
	Property.Rights	Government.Integrity	Judicial.Effectiveness	Tax.Burden	Government.Spending	Fiscal.Health	Business.Freedom	Labor.Freedom	Monetary.Freedom	Trade.Freedom	Investment.Freedom	Financial.Freedom
Property.Rights	1.00	0.86	0.84	−0.13	−0.29	0.27	0.76	0.44	0.48	0.62	0.63	0.72
Government.Integrity	0.86	1.00	0.85	−0.17	−0.33	0.25	0.70	0.41	0.43	0.51	0.56	0.63
Judicial.Effectiveness	0.84	0.85	1.00	−0.14	−0.25	0.23	0.72	0.47	0.47	0.55	0.59	0.64
Tax.Burden	−0.13	−0.17	−0.14	1.00	0.43	0.08	0.14	0.27	0.13	0.21	−0.12	−0.02
Government.Spending	−0.29	−0.33	−0.25	0.43	1.00	0.12	−0.18	0.01	−0.06	−0.12	−0.13	−0.16
Fiscal.Health	0.27	0.25	0.23	0.08	0.12	1.00	0.20	0.17	0.29	0.32	0.16	0.17
Business.Freedom	0.76	0.70	0.72	0.14	−0.18	0.20	1.00	0.56	0.48	0.62	0.51	0.60
Labor.Freedom	0.44	0.41	0.47	0.27	0.01	0.17	0.56	1.00	0.35	0.40	0.28	0.31
Monetary.Freedom	0.48	0.43	0.47	0.13	−0.06	0.29	0.48	0.35	1.00	0.55	0.58	0.51
Trade.Freedom	0.62	0.51	0.55	0.21	−0.12	0.32	0.62	0.40	0.55	1.00	0.57	0.62
Investment.Freedom	0.63	0.56	0.59	−0.12	−0.13	0.16	0.51	0.28	0.58	0.57	1.00	0.79
Financial.Freedom	0.72	0.63	0.64	−0.02	−0.16	0.17	0.60	0.31	0.51	0.62	0.79	1.00

We have a few variables showing multicollinearity caused by the very structure of the dataset. For example, it is highly intuitive the positive relationship between government integrity,property rights and judicial effectiveness since government integrity is a sort of guaarantee for such institutions. They are highly related with economic variables as well since such institutions are essetial for having economic freedom related to business, investment and finance. We therefore expect to see such variables' loading vectors closely projected to each other. In particular, government integrity, property rights and judicial effectiveness should set out themselves relatively apart from others in the biplot.

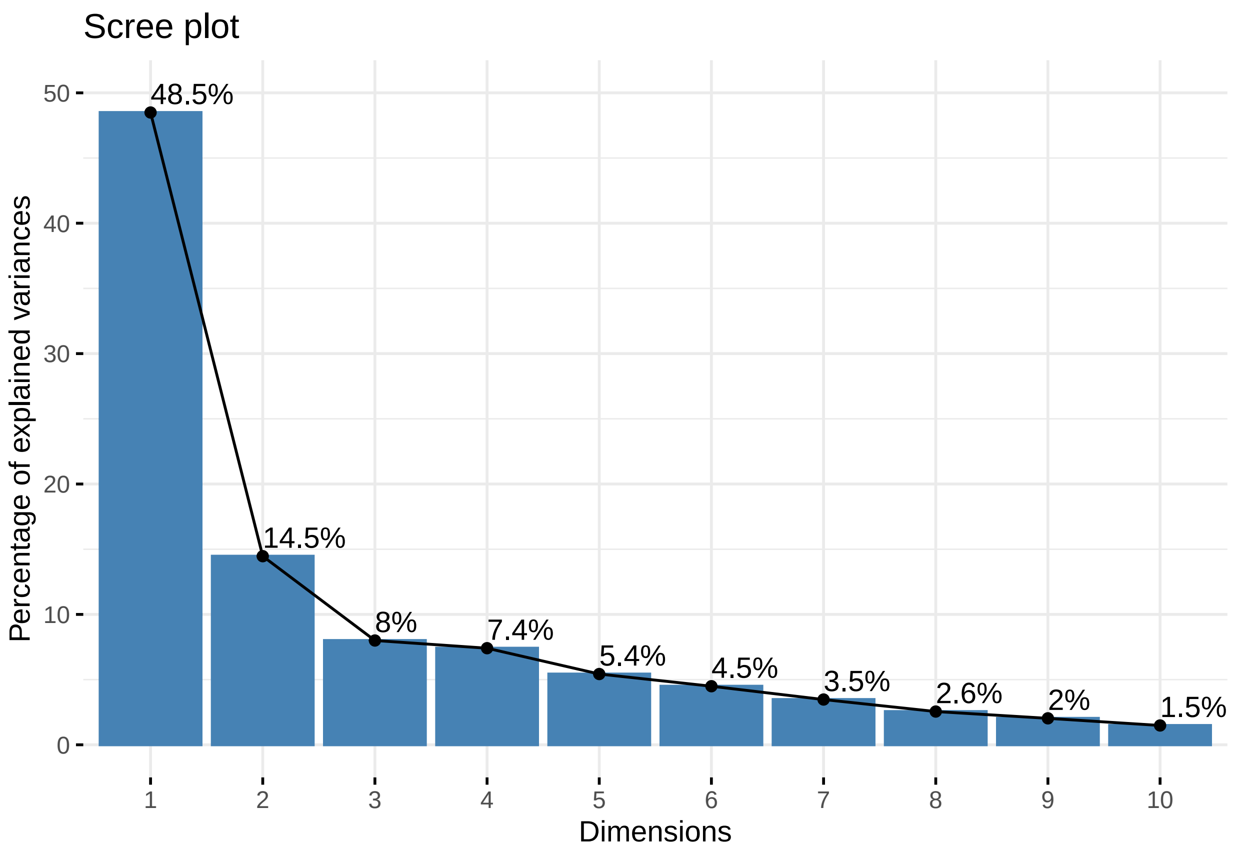
PCA Analysis

Here we want to figure out amount of information carried by each country namely its score, collecting all the correlated information contained in whole dataset and shrink it. This is done through the PCA analysis that respons to such aim by reducing the dataset dimension to a few components that are keeping most of the information contained in it. In particular, two is a desirable number of dimensions for graphical applications.

The principal components are extracted from the singular value decomposition of the correlation matrix of the variables. In this case, each principal component is associated with an eigenvalue of that matrix, and principal components associated with eigenvalues larger than 1 are able to explain more variance than just one variable. That is why, as a rule of thumb, we retain only components whose eigenvalue is > 1. We have three eigenvalues satisfying this conditions however the third one is right at the border, as we can see from the scree plot below.

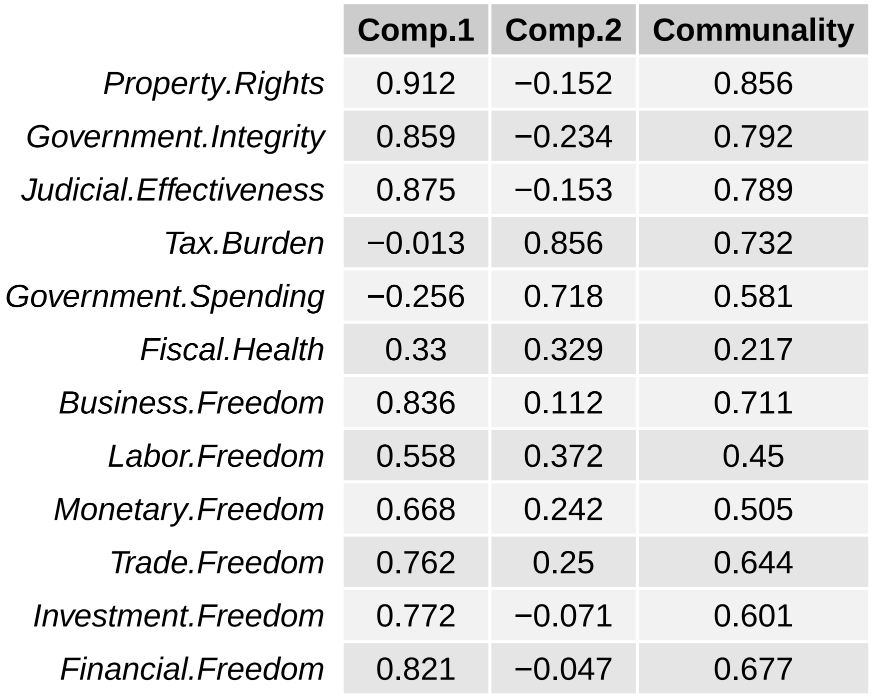


With only two eigenvalues, we are able to explain around 64% of the variablity contained in the whole dataset, as we can see from the graph below. That is good enough for our exploratory purpose. Moreover, the third component would add only 8% of variance if included in the analysis.



Results

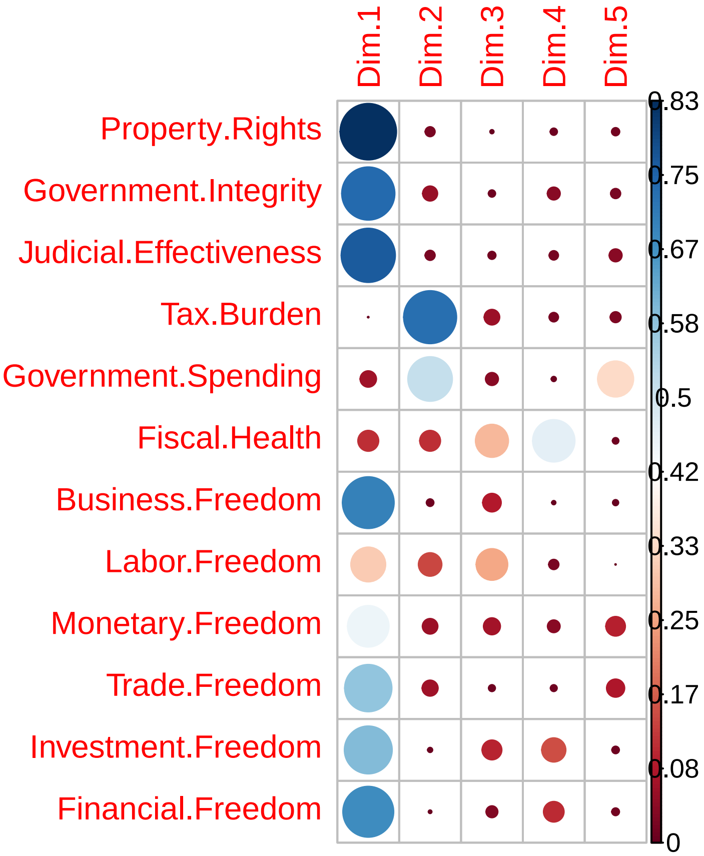
The table displayed below shows the level of correlation and communality of the two components we have just selected.



	Comp.1	Comp.2	Communality
Property.Rights	0.912	−0.152	0.856
Government.Integrity	0.859	−0.234	0.792
Judicial.Effectiveness	0.875	−0.153	0.789
Tax.Burden	−0.013	0.856	0.732
Government.Spending	−0.256	0.718	0.581
Fiscal.Health	0.33	0.329	0.217
Business.Freedom	0.836	0.112	0.711
Labor.Freedom	0.558	0.372	0.45
Monetary.Freedom	0.668	0.242	0.505
Trade.Freedom	0.762	0.25	0.644
Investment.Freedom	0.772	−0.071	0.601
Financial.Freedom	0.821	−0.047	0.677

With the dimensionality reduction we retain most of the information carried by the first three indexes, namely `Property.Rights`, `Government.Integrity` and `Judicial.Effectiveness`. The two components are able to capture more than half of the variability associated with each variable, and only `Monetary.Freedom` and `Labor.Freedom` show a communality lower or equal to 50%. The only variable for which the reduction performs really bad is `Fiscal.Health`, but this is quite inevitable when moving from 12 dimension to just 2.

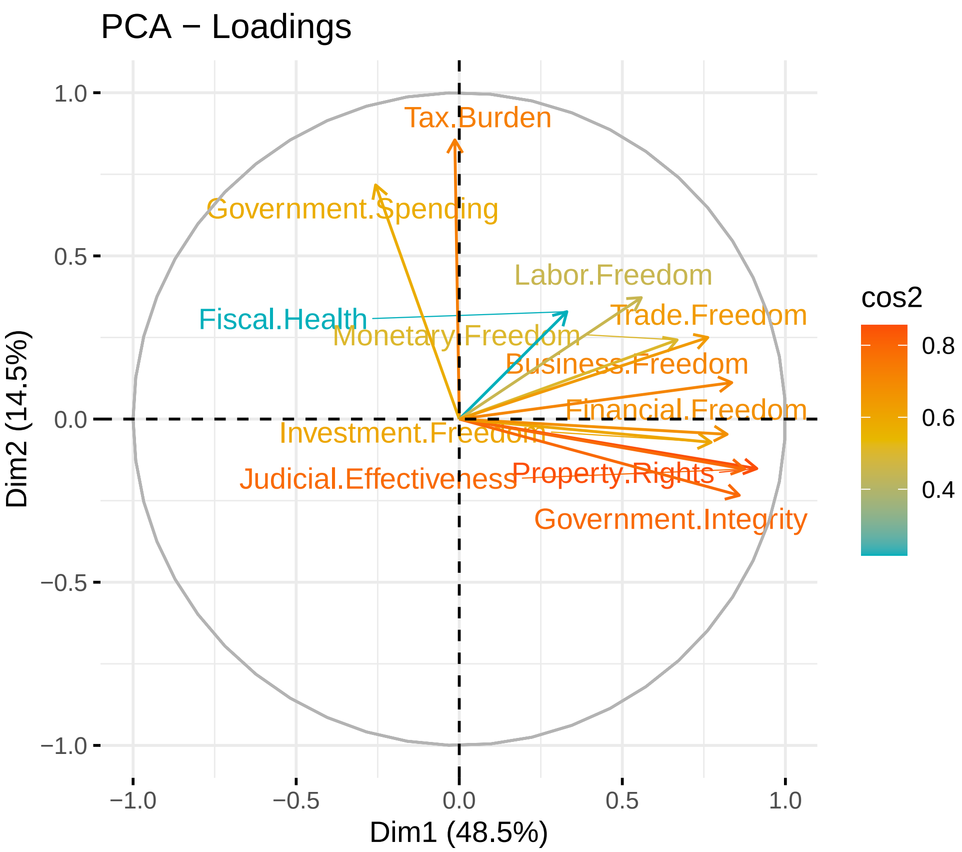
Below we can see how well each component represents each initial variable. Large values stand for good representation, while values close to 0 means a bad representation.



First of all, we can notice that after the second components the strength of the relation with the variables decreases to non significant level, with the exception of `Fiscal.Health` and `Labor.Freedom`, which in fact were the variables about which we lost the greater part of information, as highlighted in the communality table. Nonetheless, we can give the following interpretation of the components:

- The first component is strongly correlated with all the indexes with the exception of `Tax.Burden` and `Government.Spending`. Varying from `Judicial.Effectiveness` to `Financial.Freedom`, we can say this component represents the extent to which a country is developed, both from a democratic and an economic point of view.
- The second component captures the variance that the first component missed, and it is mainly related to `Tax.Burden` and `Government.Spending`. Overall, we can say this component represents the extent to which the state intervene directly in the economy, not through policies but through the injection (`Government.Spending`) or collection (`Tax.Burden`) of money.

To give the correct interpretation of the scores' value, we need to plot the loadings in the components' space.

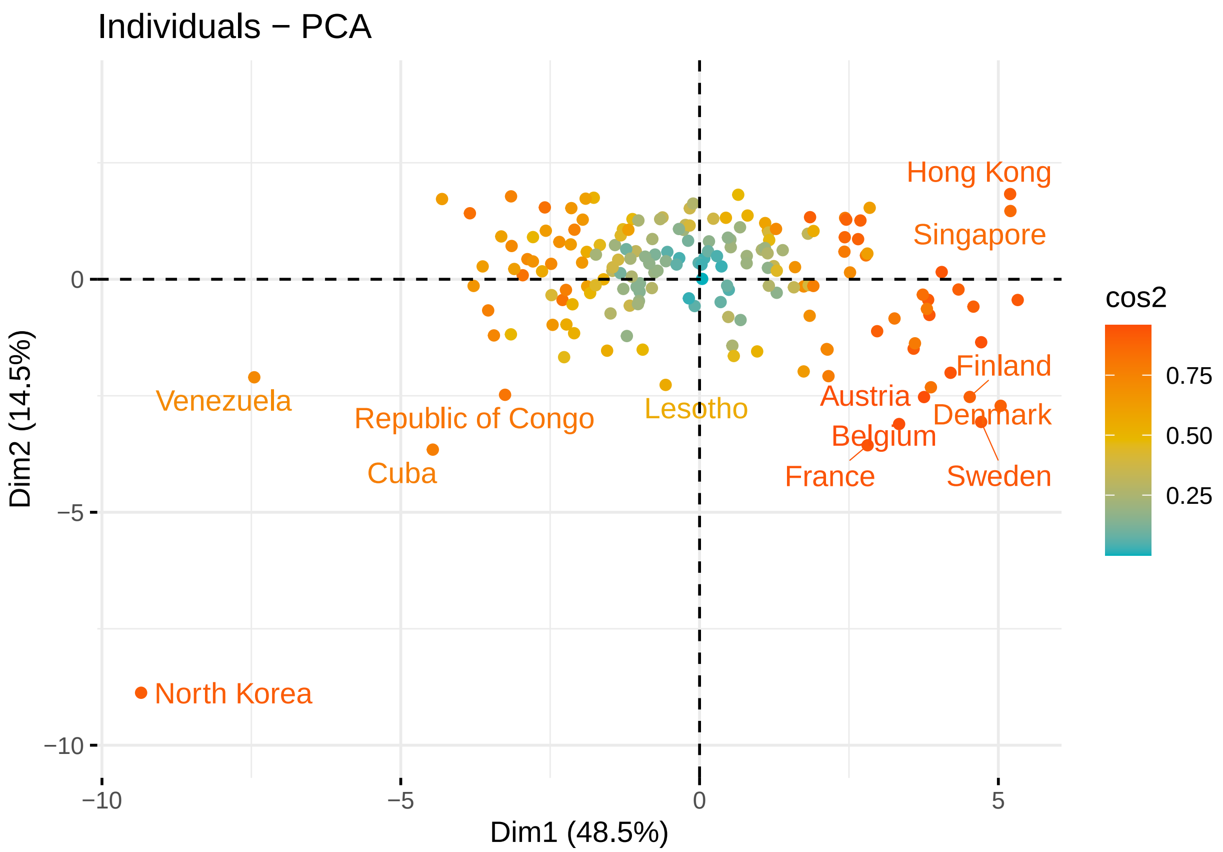


As we expected from our multicollinearity analysis, `Government.Integrity`, `Judicial.Effectiveness` and `Property.Rights` show and lie over almost exact the same direction having the strongest relation with the first PC. `Fiscal.Health` and `Labor.Freedom` are not well represented by none of the components, and this is evident also geometrically, as they point in a direction different from the one of each component. Finally, `Tax.Burden` and `Government.Spending` show an orthogonal direction with respect to the other variables and are mainly related to the second component. We can therefore give the following interpretation of the PC space:

- Countries with a large firs PC are characterized by a developed economy and democracy, while as we move to the left we should encounter developing and underdeveloped countries.
- Due to the counterintuitive meaning of `Tax.Burden` and `Government.Spending`, countries with large second component are characterized by low taxes and low government expenditure, while as we move downwards, we should encounter countries characterized by an increasing level of intervention of the state in the economy.

To verify our expectations, we plot the scores associated to each country in the PC space.

```
## Warning: ggrepel: 167 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

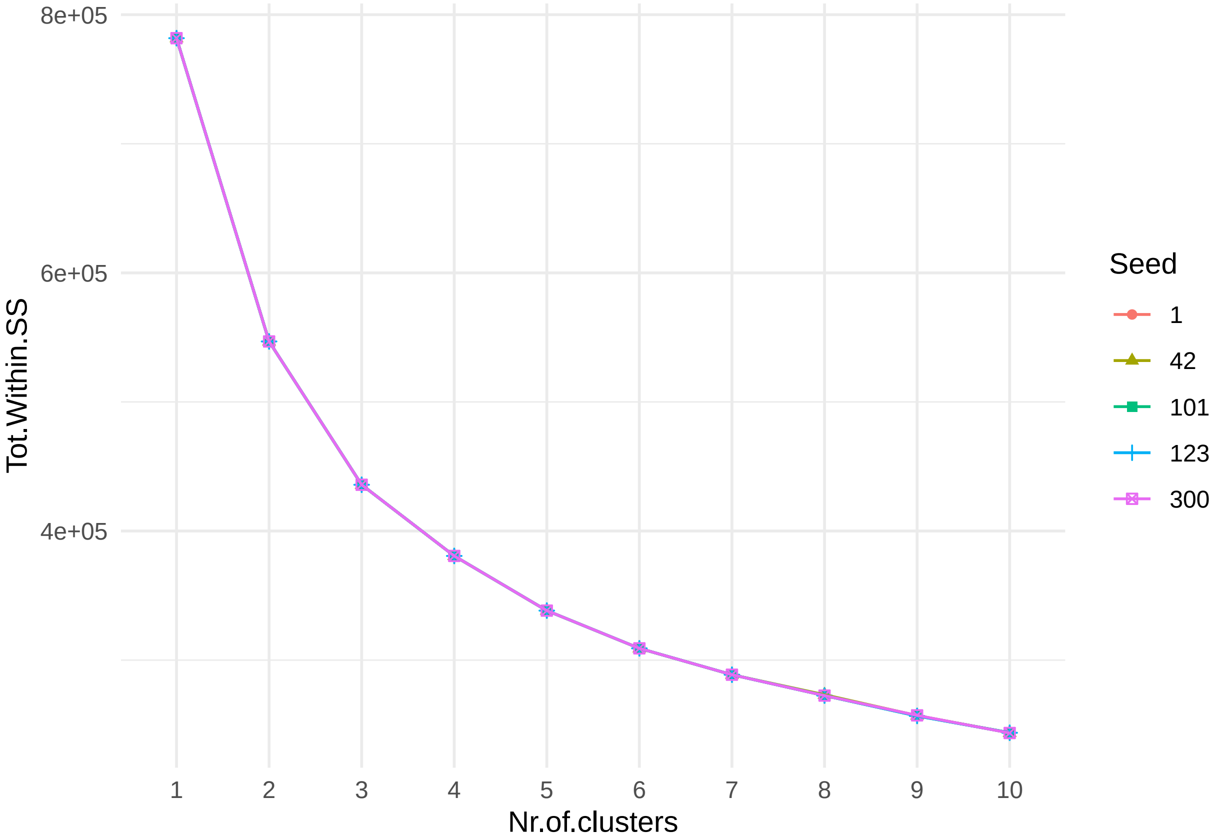


As expected, most (if not all) developed countries are in the fourth quadrant, associated with a well-established level of democratic and economic freedom. However, the second component takes negative value, meaning that in this countries we have a relevant level of taxation and government expenditure, probably due to the need to sustain welfare policies. Indeed, Ireland, known to be one of the tax haven of Europe, is placed in the first quadrant, i.e. it displays similar level of freedom but with much lower taxation. The two most-right countries are Singapore and Hong Kong, that are therefore the countries where one could enjoy at the best all the aspects of economic freedom. As far as we can see in that representation, most of the countries in the left part of the graph are really poor and are associated with low economic freedom; however, countries that are in the second quadrant are expected to have low taxation level and low government expenditure. Finally, in the third quadrant there are three "outliers", that are Venezuela, Cuba and North Korea. All these three countries are socialist states, therefore associated with large public expenditure, and in fact they display very low values for component 2.

K-Means

K-Means is a clustering algorithm in which the number of clusters is set a priori by the user. Given the number of clusters, the algorithm seeks for the best partition of individuals which minimizes the within cluster variance: observations within a cluster are expected to be similar, and so they should show a low level of variability. In our specific case, the within-cluster variation of a cluster is defined as the mean pairwise squared Euclidean distance between the observations part of the same cluster.

The first step of k-means algorithm consists in randomly assigning all the observations to a certain cluster, and then at each step we reduce the within-cluster variation by changing the assignment. As one could guess, the output of k-means critically depends on that initial starting random assignment. That is why we perform that random operation with different seeds, and see how the within cluster variation changes as we enlarge the number of clusters k.

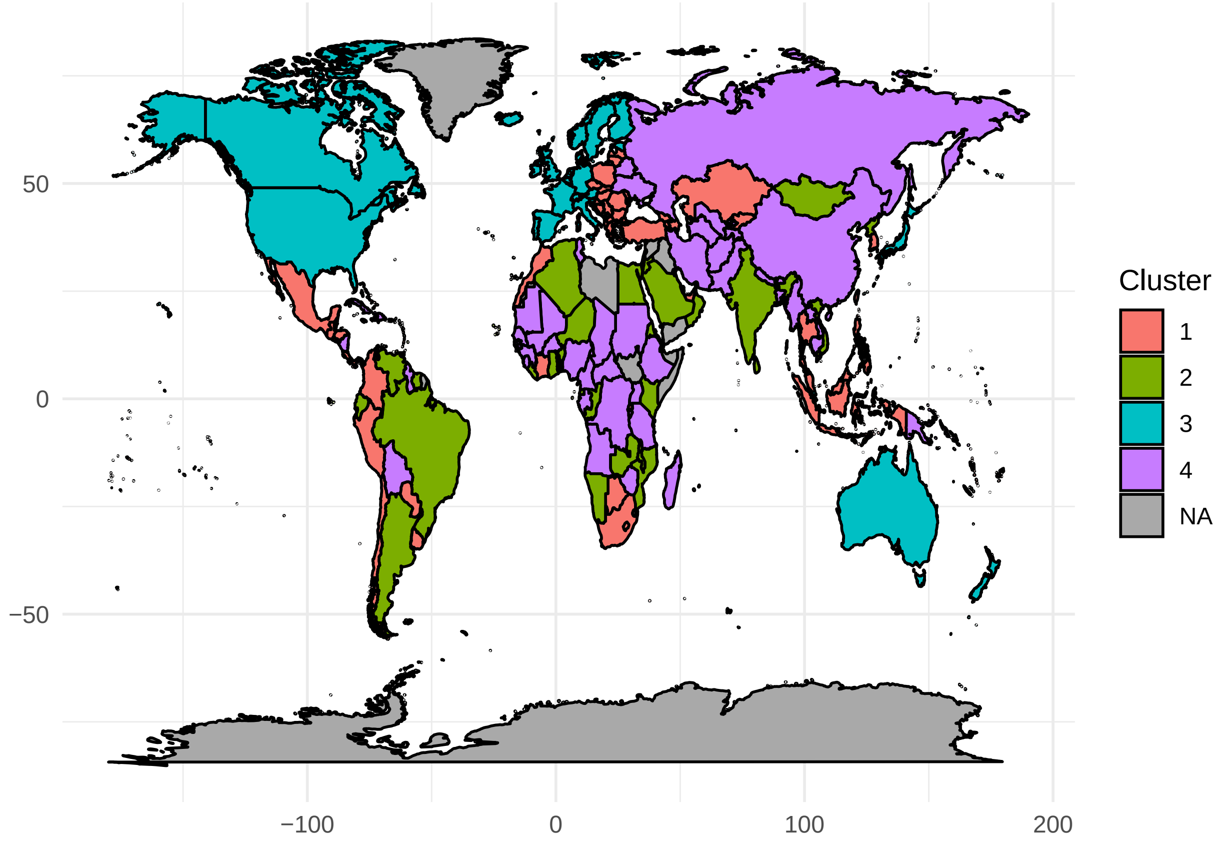


In our case, the seed does not influence much the final result. Even if all the points seem to be at the same level for each number of clusters, there is a slight variation among them. Nonetheless, we can see that after 4 clusters the within-cluster variability has halved with respect to the case in which all observations were grouped together. Moreover, the reduction in variability after that number of cluster is less relevant, so we would perform the analysis using k = 4. Below, we show the mean value of each index for each of the cluster.

```
##                                1        2        3      4
## Property.Rights         58.66935 41.65610 84.07778 35.698
## Government.Integrity    44.15645 34.25122 75.63704 29.764
## Judicial.Effectiveness 51.75645 38.62927 78.62963 33.306
## Tax.Burden              81.16935 77.00732 63.54444 77.690
## Government.Spending     66.95161 60.31707 42.17407 74.726
## Fiscal.Health           84.26613 18.81951 78.78889 80.580
## Business.Freedom        69.55323 59.12927 82.70370 54.226
## Labor.Freedom           61.51452 54.49024 66.03333 55.558
## Monetary.Freedom        79.58065 69.93171 84.63333 72.932
## Trade.Freedom           81.06935 69.15610 86.96296 69.230
## Investment.Freedom      67.82258 50.73171 83.14815 37.300
## Financial.Freedom       56.12903 42.92683 73.70370 30.200
```

First of all, we notice that cluster 3 almost always displays the higher value for all the indexes, with the exception of `Tax.Burde`, `Government.Spending` and `Fiscal.Health`. These are probably the most developed countries in the world, displaying a moderate level of taxation and public expenditure. In all the indexes where cluster 3 gets the largest value, cluster 1 takes the second largest value: these are probably developed or developing countries, whose performances in terms of economic freedom is only a little worse than the fully developed countries of cluster 3. Notably, cluster 1 takes the largest value in `Fiscal.Health`, meaning that those states (and not the fully developed ones) have the most serene financial situation. Cluster 1 has also the largest value of `Tax.Burden`, suggesting that a low level of taxation may be the driver behind these emerging economies. The differences between cluster 2 and cluster 4 are minimal, the only (strongly) significant difference we observe is about `Fiscal.Health`: countries in cluster 2 show a dramatic situation regarding their public finances. Despite that fact, countries in cluster 2 often show a higher level of freedom than countries in cluster 4, especially for what concerns `Investment.Freedom` and `Financial.Freedom`. All leads to suggest that countries in cluster 4 are probably the worst performing in terms of economic freedom.

In order to test our expectation, we plot a map explicitly showing to which cluster each country belongs to.



Cluster 1 contains countries from Eastern Europe, together with most of South East Asia and some other sparse countries like Mexico, South Africa, Perù and Chile. As expected, those countries are very similar to fully developed countries, and are just below their level of economic freedom.

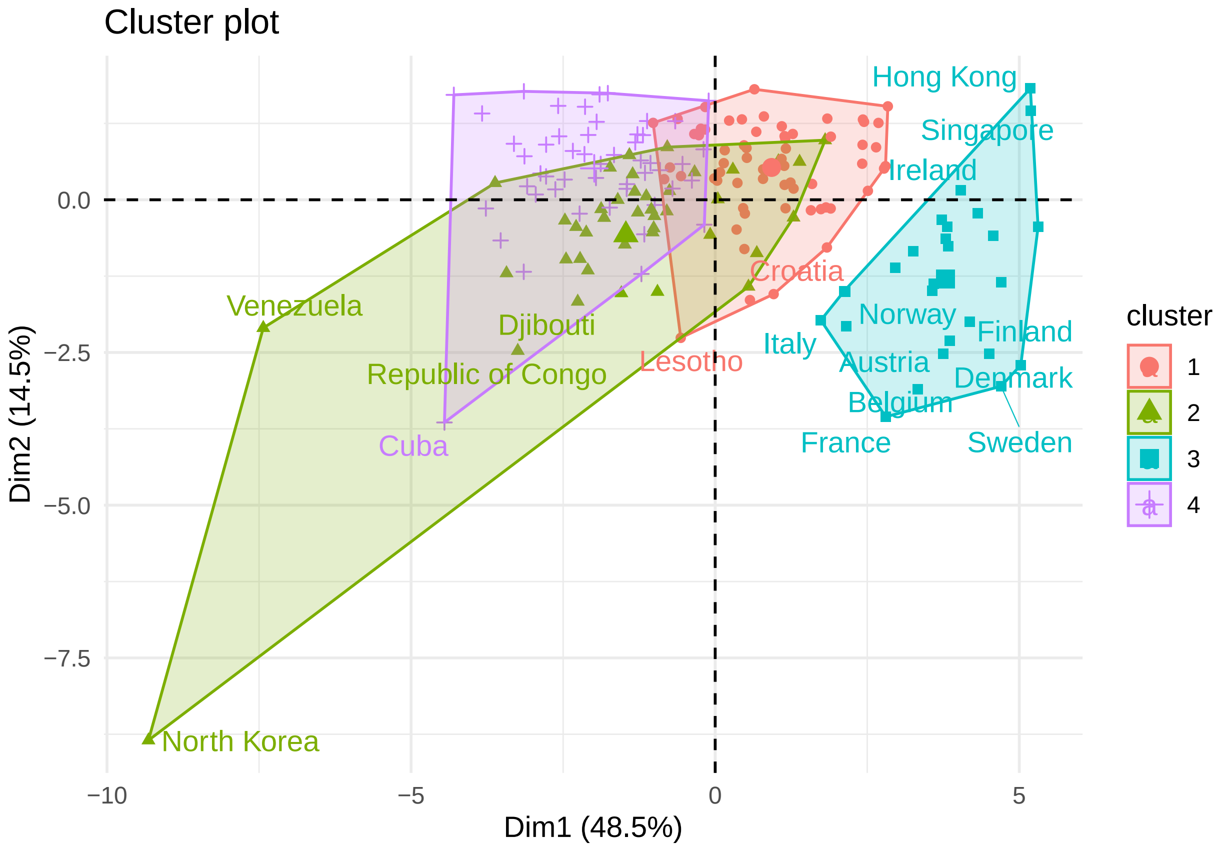
Cluster 2 is very sparse around the globe. It contains countries from South America (Argentina, Brazil, Venezuela) and North Africa (Algeria, Egypt, Saudi Arabia), together with few Asian countries (India and Mongolia). In line with our expectations, this cluster regards developing countries, whose performances are very diverse but probably worse than the ones in cluster 1.

Cluster 3 is the easiest to interpret, just like shown previously. It collects all Western Europe and North America, together with Australia, New Zealand and Japan. These are the countries with the highest economic freedom.

Cluster 4 is mainly related to Africa, but it also contains Russia, China and Cuba. These countries are not so different with respect to the ones in cluster 2, but the fact that most authoritarian countries are placed here seem to suggest that this cluster is related to the worst performances in economic freedom.

Below, we take a look at the states in the reduced principal components space.

```
## Warning: ggrepel: 162 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



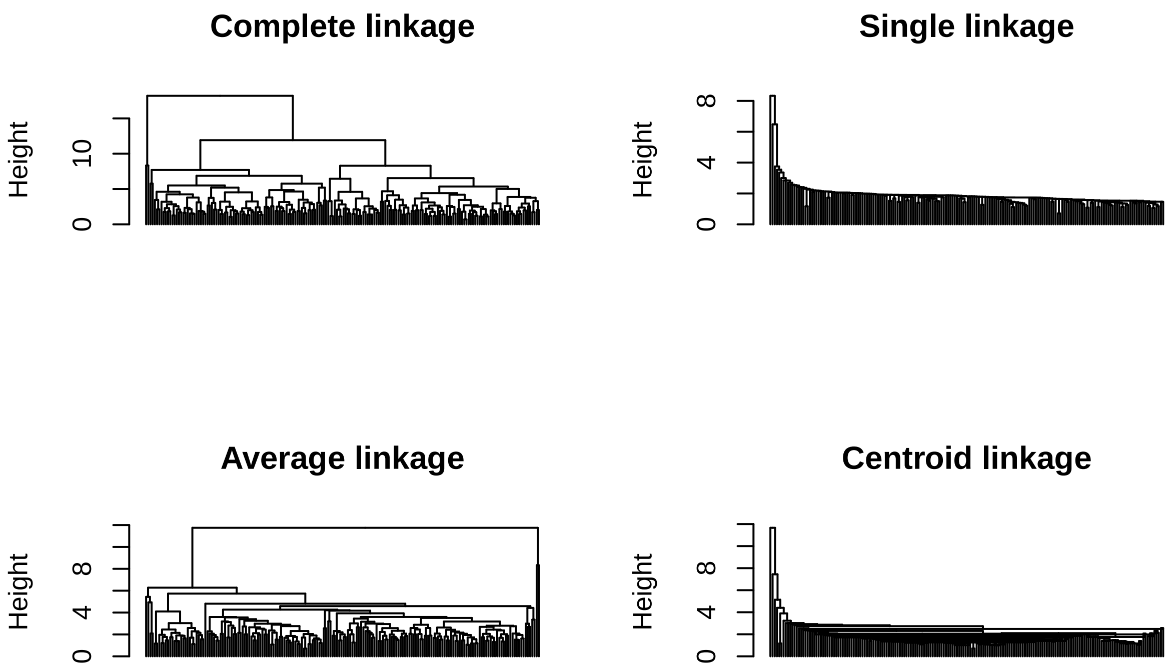
At first glance, we notice cluster 3 is clearly separated from the rest, while the others overlap a little bit. However, cluster 1 is on the same side of the graph with respect to cluster 3, so it is more similar to it than to the others. As already pointed out in the previous analysis, cluster 2 and 4 are the most similar, but (on average) cluster 4 displays higher levels for PC1, i.e. higher levels of economic and democratic freedoms, while two of the three outliers (performing very bad in terms of economic freedom) belongs to cluster 2.

Hierarchical clustering

In contrast to K-means, hierarchical clustering methods does not depend on the initial definition of the number of clusters to be searched or a starting configuration assignment. Instead, they require the user to specify a measure of dissimilarity between groups of observations. The greatest advantage of the hierarchical algorithms is to allow a tree-like visual representation of the observations, called a dendrogram. A dendrogram provides a highly interpretable complete description of the hierarchical clustering in a graphical format.

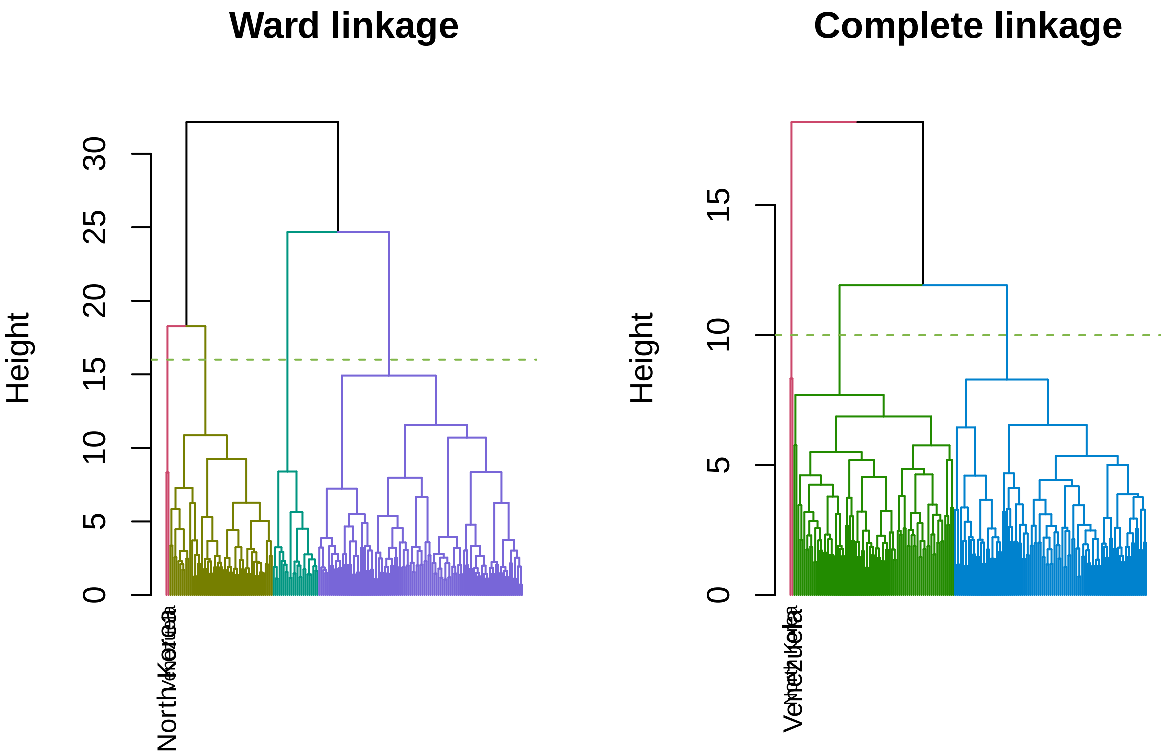
In agglomerative approach, the clustering starts at the bottom where each observation represents a separate cluster and it recursively combines the two nearest clusters together, until all the observations are grouped in a single cluster.

In order to perform clustering, we first compute the distance matrix with distances for each pair of observations. Secondly, the algorithm need us to specify the measure of the distance between clusters in order to decide the rules for clustering. We perform hierarchical clustering of the observations using four different distance measures and compare the results by plotting the corresponding dendrograms.



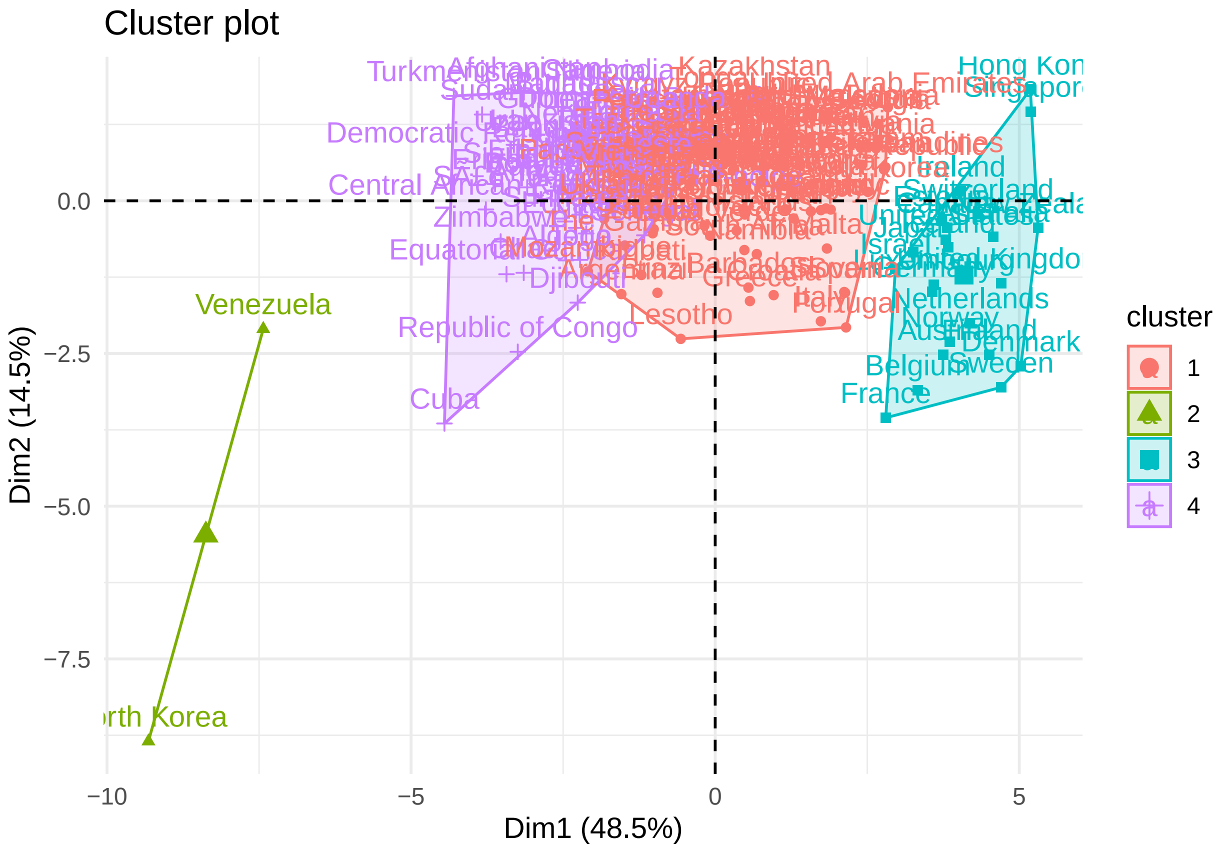
Selecting the agglomerative method

One can immediately notice that the dendrograms obtained with single linkage and centroid linkage, are strongly unbalanced and seem to be less helpful in identifying clusters. The **Complete** approach, on the other hand, looks much **more interpretable**. It seems to work better for this data, identifying well-separated groups. Nevertheless, aknowledging the **presence of outliers** among our observation we also try **Ward** agglomerative method which could be more appropriate in our case.



Both Ward and Complete agglomerative methods insist on treating North Korea and Venezuela as outliers, allocating then to a separate cluster. We will proceed with 4 clusters as suggested by Ward clustering and describe the main characteristics of the resulting groups.

Clustering results



- Cluster 3 is located on the right side of the map, detached from the others and represents the sample of countries with the highest level of freedom in all senses: these are the developed countries;

- Cluster 1 is located in the origin and is the hugest and the most heterogenuous one, displaying average freedom scores;

- Again, we see the overlap between cluster 1 and 4 : the main difference between the two clusters consists in their attitude towards economic interactions with the rest of the world. Cluster 4 is less open to economic and financial collaboration with other countries and has lower scores with respect to rule of law than the cluster 1.

- Cluster 4 consists of Asia, great part of Africa and some south American countries; Cluster 4 and cluster 2 are located on the left side of the map with negative scores in the first component. Indicating lower than the average scores of economic freedom.

- North Korea and Venezuela stay far apart from the others with the lowest freedom scores in all the dimentions. Notably, these and also countries like Cuba and the most of African countries have low scores with respect to second latent component. This indicates that in these states the fiscal policy is implemented on large scale, with the government as the main player in the economy of the country.

```
## [1] "cluster size:"
```

```
## segments
##   1   2   3   4
## 103   2  23  52
```

```
## # A tibble: 12 x 5
##    Clusters                 `1`   `2`   `3`   `4`
##    <chr>                  <dbl> <dbl> <dbl> <dbl>
##  1 Business.Freedom        68.4  20.2  84.1  51.3
##  2 Financial.Freedom       51.7   5    76.5  31.5
##  3 Fiscal.Health           66.0   9.2  83.1  65.9
##  4 Government.Integrity    42.2  16.4  80.1  28.0
##  5 Government.Spending     63.7  28.6  43.9  74.4
##  6 Investment.Freedom      61.0   0    84.1  41.9
##  7 Judicial.Effectiveness  50.6   9.4  81.4  28.4
##  8 Labor.Freedom           61.1  14.6  68.2  52.3
##  9 Monetary.Freedom        77.6   0    84.2  73.2
## 10 Property.Rights         55.5  17.5  86.1  32.0
## 11 Tax.Burden              80.1  36.2  64.3  76.6
## 12 Trade.Freedom           79.0  29.4  87.0  66.9
```

Comparison between Kmeans and Hierarchical clustering

```
##               kmeans
## ward.clusters  1  2  3  4
##             1 62 26  4 11
##             2  0  2  0  0
##             3  0  0 23  0
##             4  0 13  0 39
```

- Both kmeans and hierarchical clustering mostly agree on the 1 and 3 clusters, however:
- The kmeans tends to create more balanced groups as opposed to hierarchical clustering.