Cluster analysis of Economic Freedom Index in European countries

Multivariate Statistics Project

Mariya Hristova

2020-01-20

1 Introduction

Country clustering has been explored as a technique for reducing the complexity and exploring relationships between countries. Rather than examining country level indicators in isolation, clustering offers the opportunity to determine which countries are similar and explore the relationships between variables driving cluster membership.

This project focuses on the clusters of european countries formed by the variables included in Economic Freedom Index (EFI). The dataset (https://www.heritage.org/index/explore?u=637130613549367845) is publuished each year from the Heritage Foundation. The data is also freely available on the website kaggle.com (https://www.kaggle.com/lewisduncan93/the-economic-freedom-index).

The Economic Freedom Index is a measure of the quality of economic instutiotions in each country. The economic institutions are the institutions which form the regulations, laws, and policies that affect economic incentives and thus the incentives to invest in technology, physical capital, and human capital. This possibility implies that institutions are a major fundamental cause of economic growth and cross-country differences in economic performance. Institutions are influenced by the historical development of a society, culture, geography, membership in political and financial unions and many other socio-economic factors. Freedom itself is an important value, and economic freedom is a vital engine for generating long-term wealth that makes possible the wide range of important economic and social achievements.

The goal of the project is to determine the clusters of european countries based on the variables included in the Economic Freedom Index and try to discover the connection between economic freedom and economic performance.

2 Data

The Economic Freedom Index, published in 2019 by Wall Street Journal and Heritage Foundation, measures economic freedom based on 12 variables graded on a scale from 0 to 100 and grouped into the following four broader categories:

i. Rule of law: 1. property rights, 2. government integrity, 3. judicial effectiveness;
ii. Government size: 4. tax burden, 5. government spending, and 6. fiscal health;
iii. Regulatory efficiency: 7. business freedom, 8. labor freedom, and 9. monetary freedom; and
iv. Market openness: 10. trade freedom, 11. investment freedom, and 12. financial freedom.

Macroeconomic variables on state level are also included in the dataset: Tariff Rate, Income.Tax.Rate, Corporate Tax Rate, Tax.Burden of GDP (Gross Domestic Product),Government Expenditure of GDP, Population Millions, GDP in Billions PPP, GDP Growth Rate, X5 Year GDP Growth Rate (last 5 years), GDP per Capita PPP, Unemployment rate, Inflation, FDI Inflow Millions (Foreign direct investments), Public Debt of GDP.

```
data = read.csv("D:/Lisbon/Multivariate statistics/project/economic_freedom_index2019
_data.csv",stringsAsFactors=FALSE)
```

The project focuses only on European countries, in order to capture the differences in economic intitutions across countries.

Standardization of the variables is necessary, because they are measured in different units and the variances differ from each other too much.

```
data_europe_normal = data_europe #making a copy of the original data
data_europe[,8:32] <- scale(data_europe[,8:32]) #only the numeric variables
```

Table of variables:

```
 [1] "CountryID"                     "Country.Name"
 [3] "WEBNAME"                       "Region"
 [5] "World.Rank"                    "Region.Rank"
 [7] "X2019.Score"                   "Property.Rights"
 [9] "Judical.Effectiveness"         "Government.Integrity"
[11] "Tax.Burden"                    "Gov.t.Spending"
[13] "Fiscal.Health"                 "Business.Freedom"
[15] "Labor.Freedom"                 "Monetary.Freedom"
[17] "Trade.Freedom"                 "Investment.Freedom"
[19] "Financial.Freedom"             "Income.Tax.Rate...."
[21] "Corporate.Tax.Rate...."        "Tax.Burden...of.GDP"
[23] "Gov.t.Expenditure...of.GDP"    "Population..Millions."
[25] "GDP..Billions..PPP."           "GDP.Growth.Rate...."
[27] "X5.Year.GDP.Growth.Rate...." "GDP.per.Capita..PPP."
[29] "Unemployment...."              "Inflation...."
[31] "FDI.Inflow..Millions."         "Public.Debt....of.GDP."
```

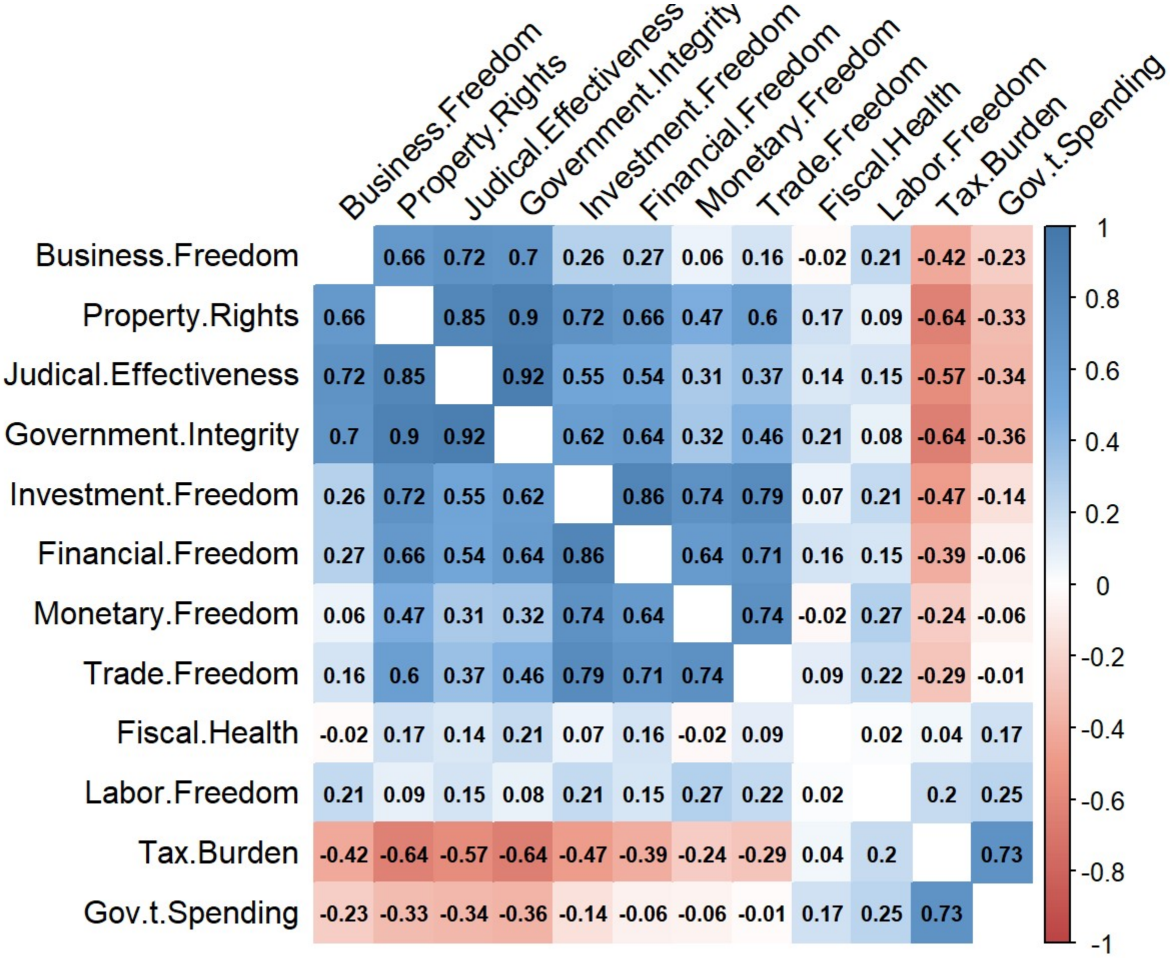
3 PCA

In this section I apply Principal Component Analysis (PCA) in order to reduce the data dimensionality, i.e., the number of variables. The original variables are correlated, but the principal components retained capture most of the total variability in the dataset of european countries.

3.1 Testing for adequacy of data for performing PCA

Principal component analysis is based on the existence of statistical association between the original variables, that means that the variables are highly correlated between each other, thus an analysis of the correlation matrix is conducted:

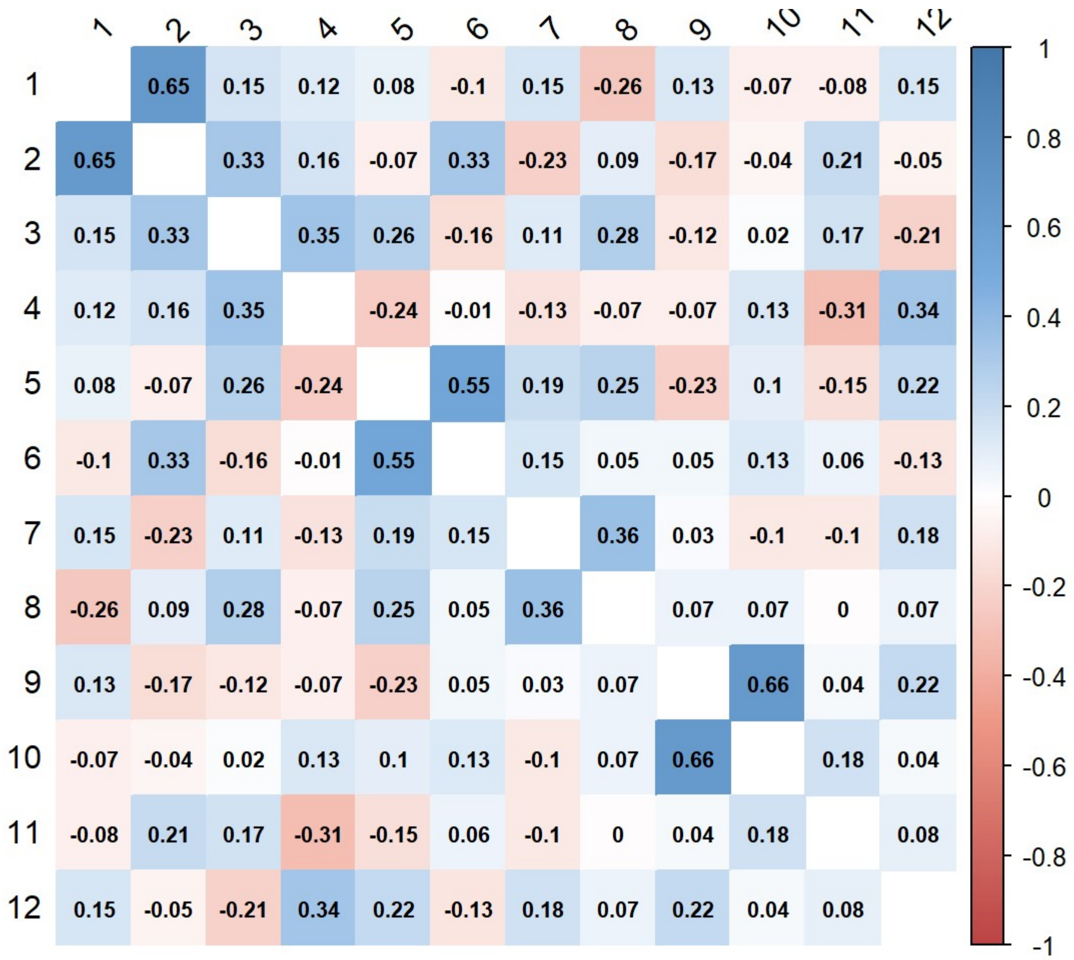
```
corrplot 0.84 loaded
```



	Business.Freedom	Property.Rights	Judical.Effectiveness	Government.Integrity	Investment.Freedom	Financial.Freedom	Monetary.Freedom	Trade.Freedom	Fiscal.Health	Labor.Freedom	Tax.Burden	Gov.t.Spending
Business.Freedom		0.66	0.72	0.7	0.26	0.27	0.06	0.16	-0.02	0.21	-0.42	-0.23
Property.Rights	0.66		0.85	0.9	0.72	0.66	0.47	0.6	0.17	0.09	-0.64	-0.33
Judical.Effectiveness	0.72	0.85		0.92	0.55	0.54	0.31	0.37	0.14	0.15	-0.57	-0.34
Government.Integrity	0.7	0.9	0.92		0.62	0.64	0.32	0.46	0.21	0.08	-0.64	-0.36
Investment.Freedom	0.26	0.72	0.55	0.62		0.86	0.74	0.79	0.07	0.21	-0.47	-0.14
Financial.Freedom	0.27	0.66	0.54	0.64	0.86		0.64	0.71	0.16	0.15	-0.39	-0.06
Monetary.Freedom	0.06	0.47	0.31	0.32	0.74	0.64		0.74	-0.02	0.27	-0.24	-0.06
Trade.Freedom	0.16	0.6	0.37	0.46	0.79	0.71	0.74		0.09	0.22	-0.29	-0.01
Fiscal.Health	-0.02	0.17	0.14	0.21	0.07	0.16	-0.02	0.09		0.02	0.04	0.17
Labor.Freedom	0.21	0.09	0.15	0.08	0.21	0.15	0.27	0.22	0.02		0.2	0.25
Tax.Burden	-0.42	-0.64	-0.57	-0.64	-0.47	-0.39	-0.24	-0.29	0.04	0.2		0.73
Gov.t.Spending	-0.23	-0.33	-0.34	-0.36	-0.14	-0.06	-0.06	-0.01	0.17	0.25	0.73	

High bivariate correlation exists between the original variables, but it is not sufficient to make conclustions and a further analysis of the partial correlation shoud be conducted. Partial correlations are measures of bivariate linear association, removing the effect of the remaining variables. The smaller the partial correlation, the more appropriate is the data for principal component analysis.

Parial correlation matrix:



	1	2	3	4	5	6	7	8	9	10	11	12
1		0.65	0.15	0.12	0.08	-0.1	0.15	-0.26	0.13	-0.07	-0.08	0.15
2	0.65		0.33	0.16	-0.07	0.33	-0.23	0.09	-0.17	-0.04	0.21	-0.05
3	0.15	0.33		0.35	0.26	-0.16	0.11	0.28	-0.12	0.02	0.17	-0.21
4	0.12	0.16	0.35		-0.24	-0.01	-0.13	-0.07	-0.07	0.13	-0.31	0.34
5	0.08	-0.07	0.26	-0.24		0.55	0.19	0.25	-0.23	0.1	-0.15	0.22
6	-0.1	0.33	-0.16	-0.01	0.55		0.15	0.05	0.05	0.13	0.06	-0.13
7	0.15	-0.23	0.11	-0.13	0.19	0.15		0.36	0.03	-0.1	-0.1	0.18
8	-0.26	0.09	0.28	-0.07	0.25	0.05	0.36		0.07	0.07	0	0.07
9	0.13	-0.17	-0.12	-0.07	-0.23	0.05	0.03	0.07		0.66	0.04	0.22
10	-0.07	-0.04	0.02	0.13	0.1	0.13	-0.1	0.07	0.66		0.18	0.04
11	-0.08	0.21	0.17	-0.31	-0.15	0.06	-0.1	0	0.04	0.18		0.08
12	0.15	-0.05	-0.21	0.34	0.22	-0.13	0.18	0.07	0.22	0.04	0.08	

The values of the partial correlation matrix are relatively small, so it is concluded that the data is suitable of PCA.

3.2 Mauchley’s test for sphericity:

This procedure allows to test for the sphericity of the data and thus to conclude on the suitability of conducting a PCA.

```
[1] "U star: 422.336"
```

```
Crtical value of  Chi-square distibution with 5% significance level:
422.3363
```

```
[1] "P-vaue: 0"
```

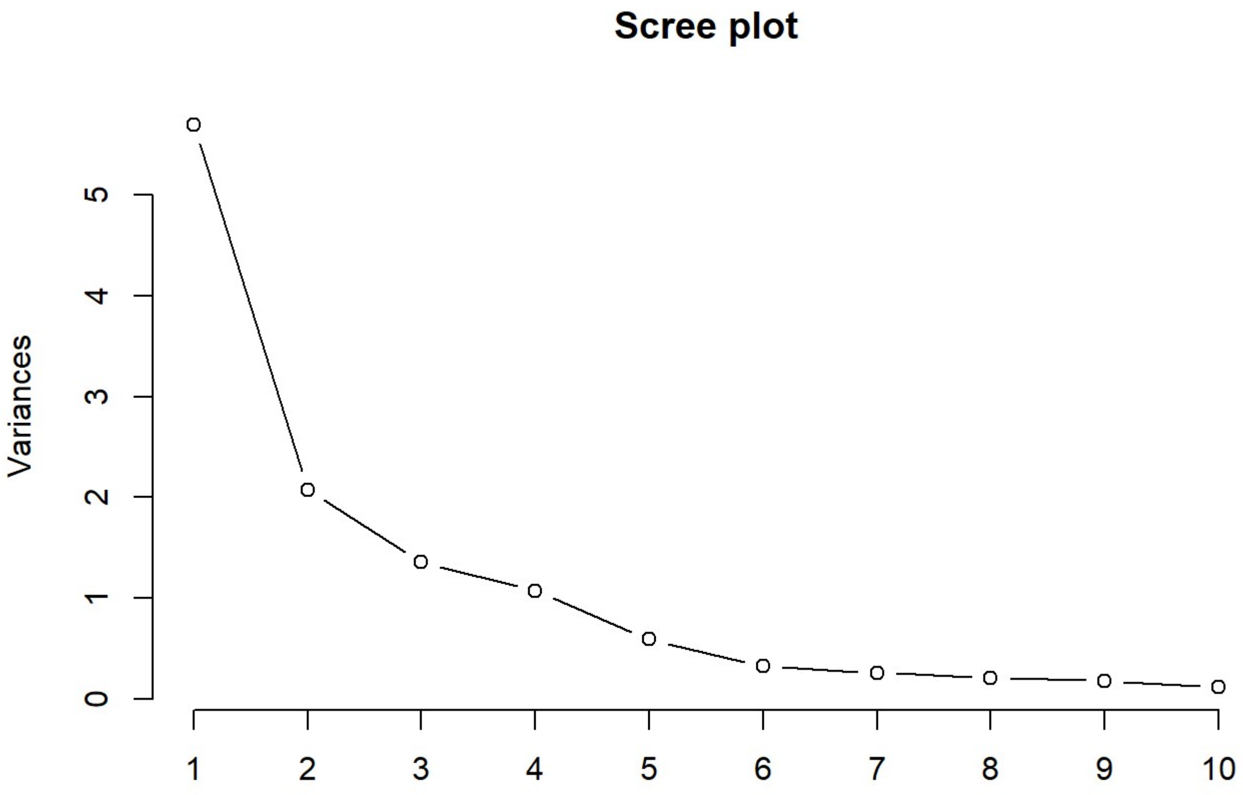
In this case we reject the null hypothesis for 5% significance level , e.g the data is appropriate for PCA.

3.3 Performing Principal Component Analysis

```
### PCA analysis
library(stats)
pca<-prcomp(x = data_europe[,8:19],scale. = TRUE) # standartizing the data
summary(pca)
```

```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7
Standard deviation     2.3864 1.4413 1.1656 1.03496 0.76810 0.56787 0.50814
Proportion of Variance 0.4746 0.1731 0.1132 0.08926 0.04916 0.02687 0.02152
Cumulative Proportion  0.4746 0.6477 0.7609 0.85017 0.89933 0.92620 0.94772
                          PC8     PC9    PC10    PC11    PC12
Standard deviation     0.4530 0.42204 0.33655 0.28525 0.22226
Proportion of Variance 0.0171 0.01484 0.00944 0.00678 0.00412
Cumulative Proportion  0.9648 0.97966 0.98910 0.99588 1.00000
```

3.4 Choosing the number of principal components



```
 [1] 5.69482464 2.07739958 1.35862474 1.07114723 0.58997439 0.32247719
 [7] 0.25820870 0.20520025 0.17811572 0.11326330 0.08136575 0.04939850
```

The scree plot shows the number of the ordered eigenvalues from largest to smallest on the x axis and the magnitude of the eigenvalue on the y axis. There exists a so-called “elbow” at the fourth principal component.

Kaiser’s criterion is another method for choosing the optimal number of principal components. For standartized variables the eigenvalues should be above 1, e.g the first four components. Further cluster analysis would be based on four principal compoents, instead of the original variables. They retain approximately 85% of the cumulative proportion of the variance.

The relationship between the initial variables and the principal components are given by:

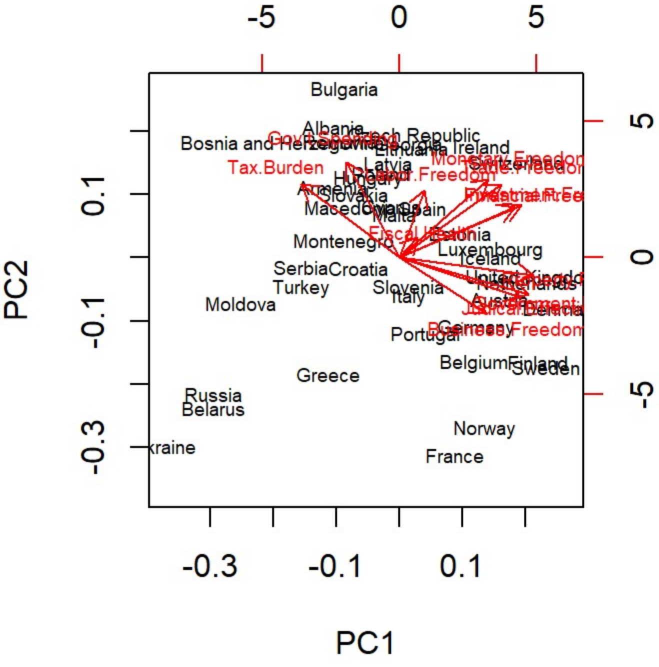
```
                         PC1    PC2    PC3    PC4
Property.Rights        0.392 -0.084  0.101 -0.056
Judical.Effectiveness  0.355 -0.190  0.263  0.040
Government.Integrity   0.377 -0.177  0.202 -0.078
Tax.Burden            -0.287  0.355  0.290  0.044
Gov.t.Spending        -0.153  0.461  0.383 -0.072
Fiscal.Health          0.054  0.101  0.382 -0.775
Business.Freedom       0.255 -0.270  0.424  0.289
Labor.Freedom          0.072  0.327  0.409  0.515
Monetary.Freedom       0.262  0.381 -0.285  0.116
Trade.Freedom          0.298  0.357 -0.187 -0.035
Investment.Freedom     0.356  0.252 -0.176 -0.016
Financial.Freedom      0.338  0.246 -0.095 -0.133
```

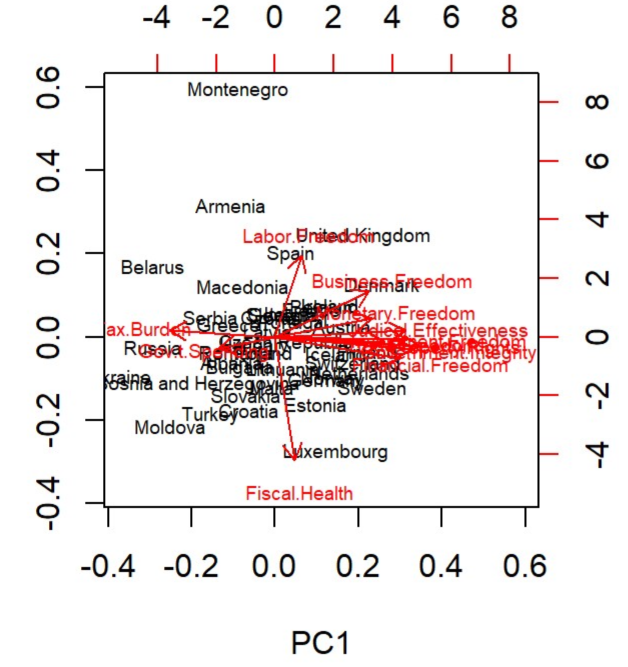
The first principal component represents most the variables, which describe the rule of law in each country. The highest coefficients are of property rights, government integrity, investment freedom and judical effectiveness.

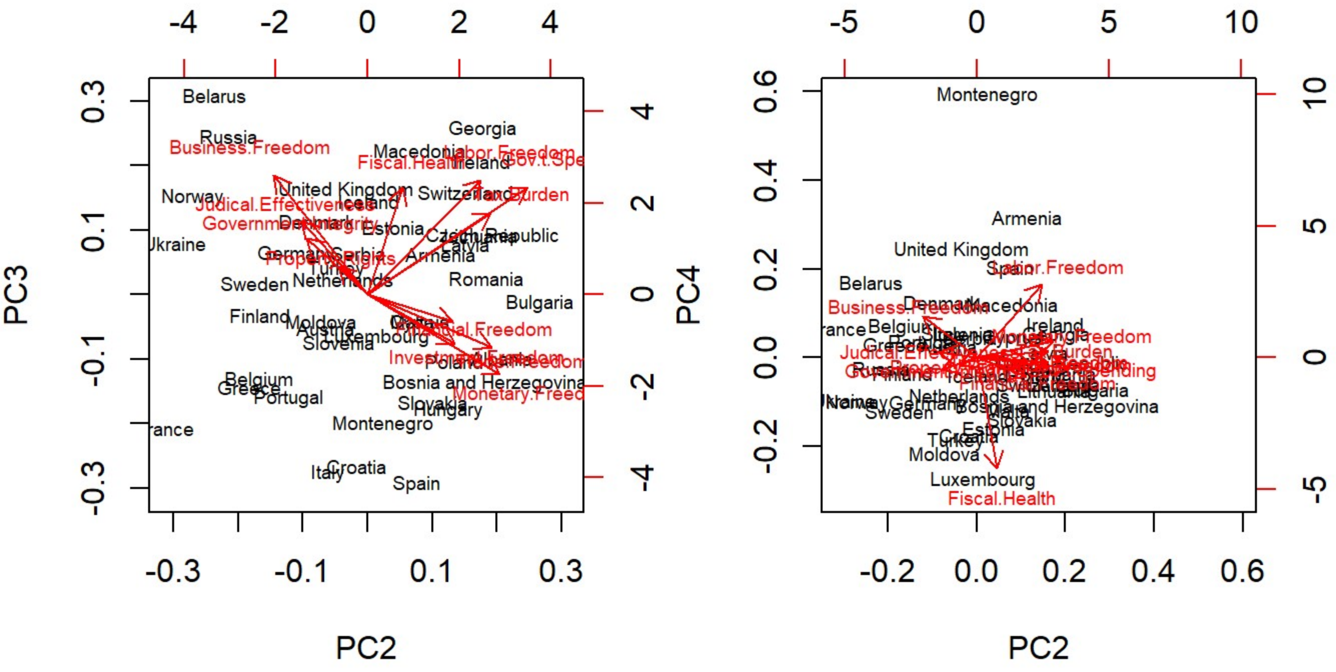
The second and third principal component respresent each original variable with no noticable group of variables, which prevails. The forth principal component is dominated by the fiscal health variable.

3.5 Biplots

A biplot, which includes both the position of each country in terms of the principal componets and it also shows how the initial variables map onto this. It allows us to visualize how the observations differ from each other based, if there are any groups formed and how different principal components discriminate the observations. It simultaneously reveals how each original variable contribites to the PC based on the length of the vectors.


-5
0
5
PC2
0.1
-0.1
-0.3
5
0
-5
-0.3
-0.1
0.1
PC1
Bulgaria
Albania
Czech Republic
Bosnia and Herzegovina
Georgia
Ireland
Lithuania
Gov't.Spending
Tax.Burden
Latvia
Monetary.Freedom
Trade.Freedom
Hungary
Labor.Freedom
Armenia
Slovakia
Macedonia
Cyprus
Spain
Malta
Fiscal.Health
Estonia
Montenegro
Luxembourg
Iceland
Serbia
Croatia
Turkey
Slovenia
Italy
United Kingdom
Netherlands
Moldova
Austria
Denmark
Judicial.Effectiveness
Germany
Business.Freedom
Portugal
Belgium
Finland
Sweden
Greece
Russia
Belarus
Norway
Ukraine
France


-4
-2
0
2
4
6
8
PC4
0.6
0.4
0.2
0.0
-0.2
-0.4
8
6
4
2
0
-2
-4
-0.4
-0.2
0.0
0.2
0.4
0.6
PC1
Montenegro
Armenia
Labor.Freedom
United Kingdom
Spain
Belarus
Business.Freedom
Denmark
Macedonia
Serbia
Greece
Monetary.Freedom
Tax.Burden
Judicial.Effectiveness
Russia
Gov't.Spending
Czech Republic
Austria
Iceland
Property.Rights
Government.Integrity
Financial.Freedom
Albania
Bulgaria
Switzerland
Netherlands
Ukraine
Bosnia and Herzegovina
Malta
Germany
Sweden
Slovakia
Croatia
Estonia
Turkey
Moldova
Luxembourg
Fiscal.Health

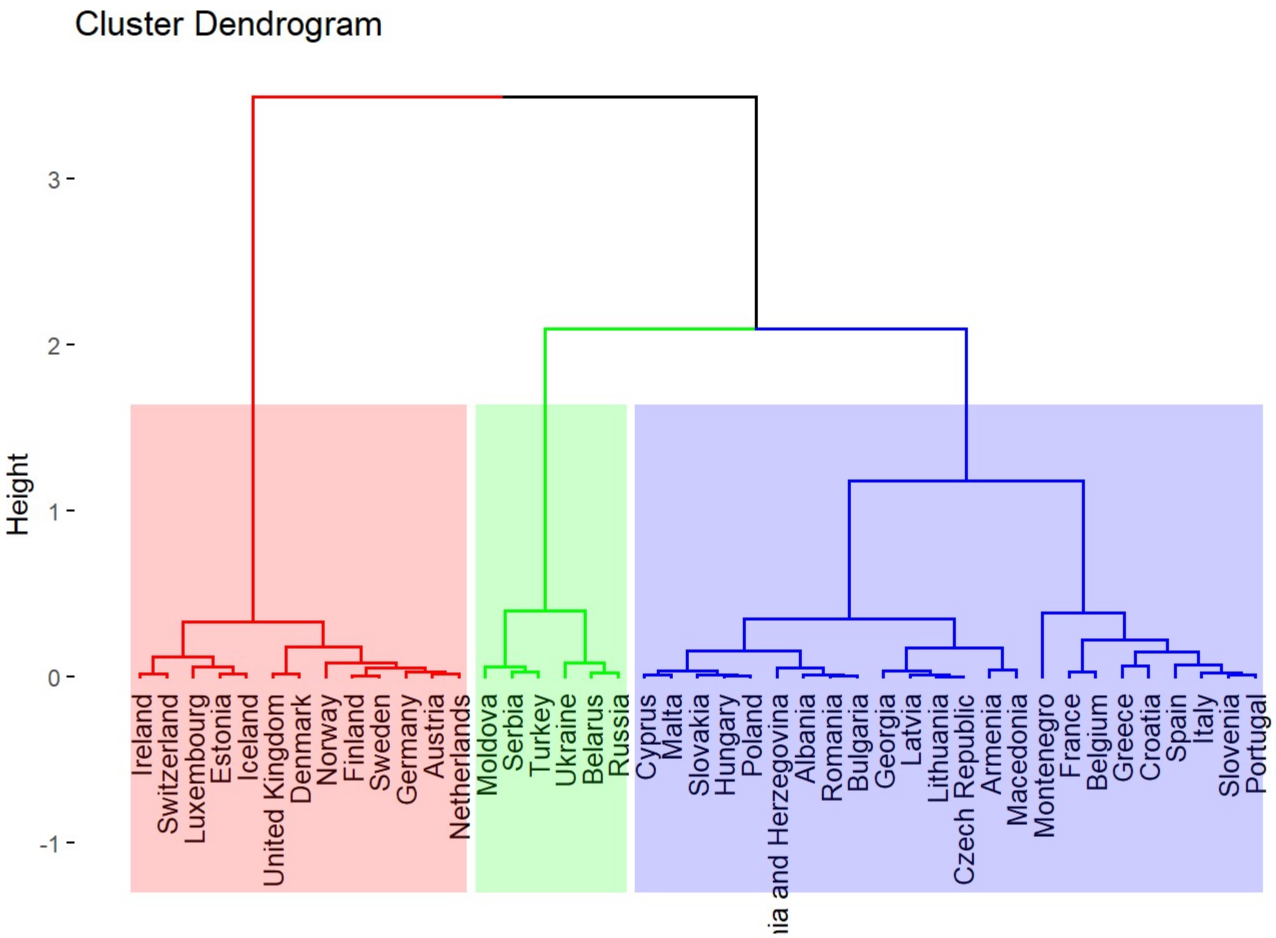


4 Hierarchical clustering

Cluster analysis is used to find groups which contain data with similar patterns, without existing previous information on the clusters. The hierarchical clustering consists of building a hierarchical structure, visualized by a dendrogram. I use the Ward minimization distance algorithm, based on the decomposition of the total variance in intra-class variance and inter-class variance. The more homogeneous the clusters, the lower the intra-class variance and the higher the inter-class variance.

4.1 Dendrogram:

In hierarchical clustering, the objects are categorized into a hierarchical structure similar to a tree-like diagram called a dendrogram. The distance of split or merge between the different countries is shown on the y-axis of the dendrogram below:



4.2 Results

The largest difference between gropus appear, when the number of clusters is 3. The three clusters are as follows:

Cluster 1: Ireland, Switzerland, Luxembourg, Estonia, Iceland, United Kingdom, Denmark, Norway, Finland, Sweden, Germany, Austria, Netherlands

Cluster 2: Moldova, Serbia, Turkey, Ukraine, Belarus, Russia

Cluster 3: Cyprus, Malta, Slovakia, Hungary, Poland, ,Bosnia and Herzegovina, Albania, Romania, Bulgaria, Georgia, Latvia, Armenia, Northern Macedonia, Montenegro, France, Belgium, Greecem Croatia, Spain, Italy, Slovenia, Portugal

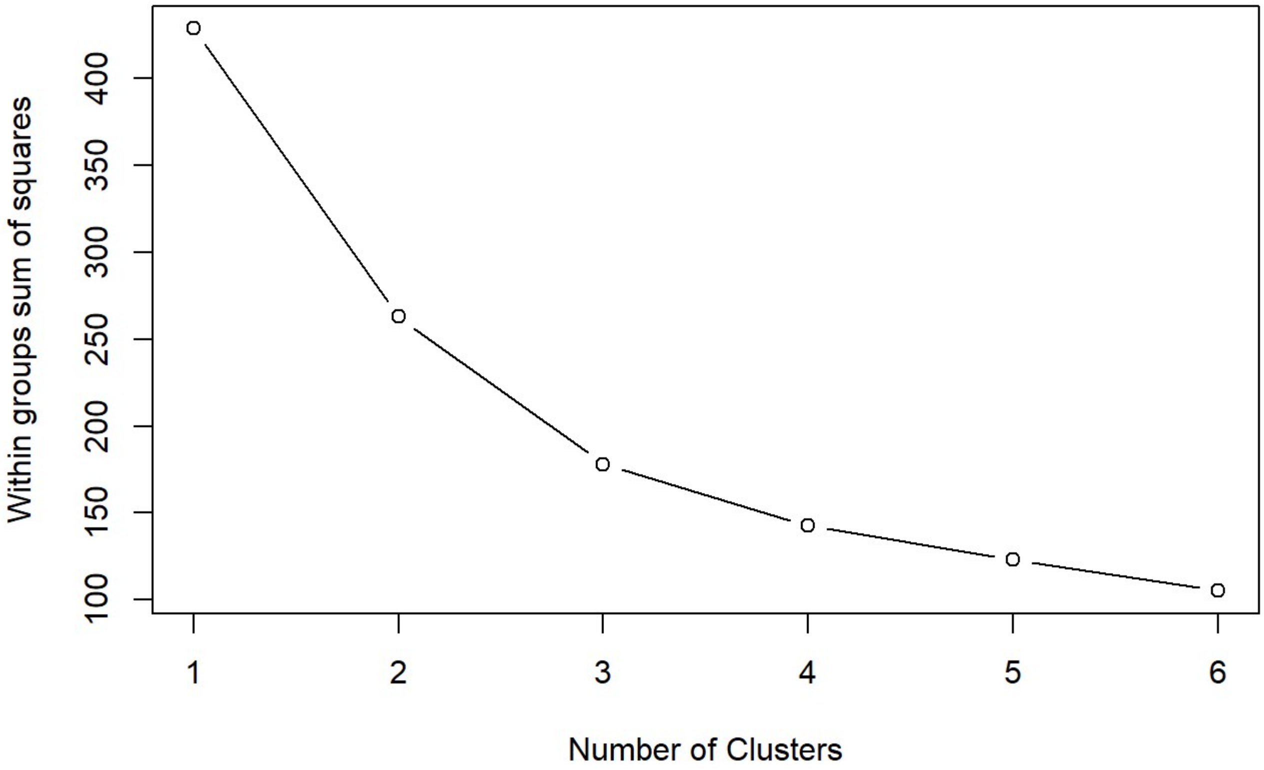
The cluster with the most observations is the third one, which is composed mostly of countries in Eastern and Central Europe, which are members of the European union. The countries in the first cluster are developed countries with high standard of living in Western and Northern Europe. The smallest cluster consists of Eastern European countries, which are not members of the European union.

When we look at the dendogram more closely, we notice that the countries which are the closest to each other share geograpgical, economic and political similarities.

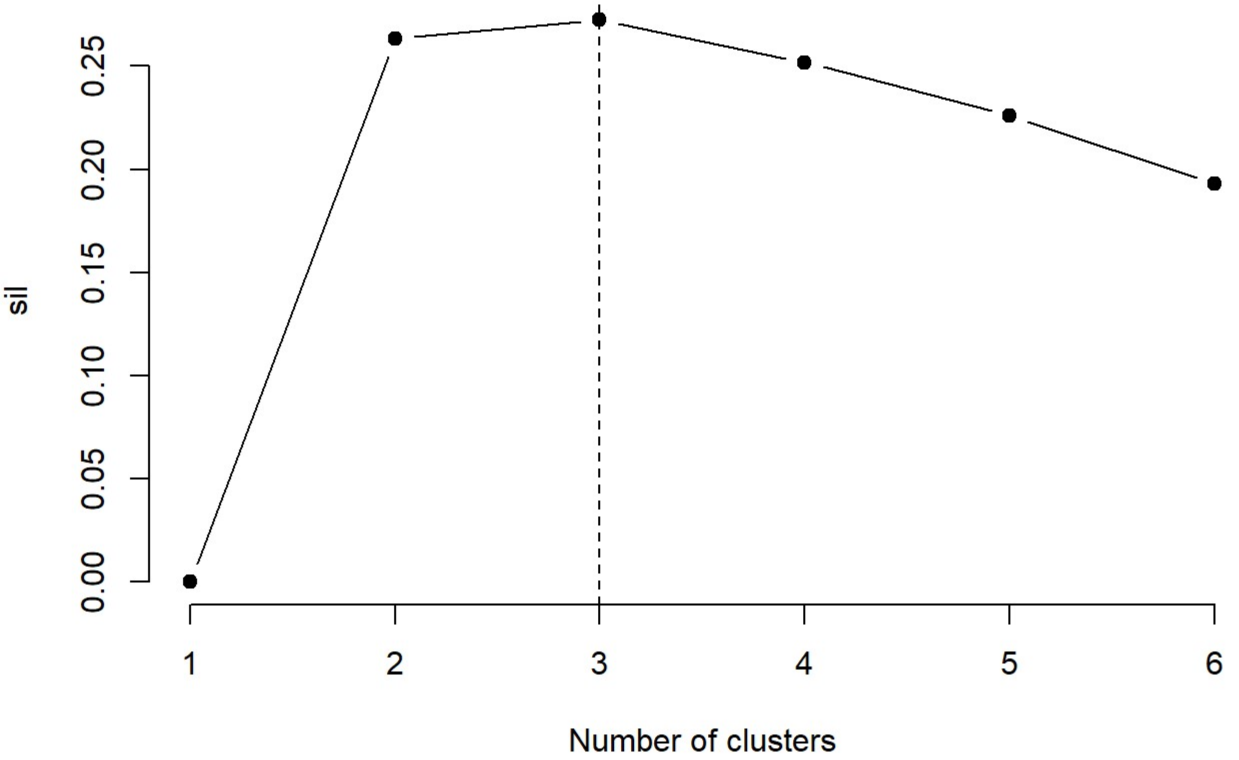
5 K-means algorithm

K-means is a centroid-based clustering algorithm. The clusters are represented by a central vector or a centroid, which is a point at the center of each cluster. The similarity between observations and the membership to a cluster is derived by how close a data point is to the centroid of the cluster.

Determining the optimal number of clusters based on the weighted sum of squares and the Silhouette score:



In this method the form of groups is based on minimizing the heterogeneity within the clusters. Thus, the clusters grouped are those that allow the minimization of the intra-group variability, e.g. Within Sum of Squares Clusters with a small sum of squares ire more compact than clusters that have a large sum of squares. In this graph, there is an "elbow" point at k=3.



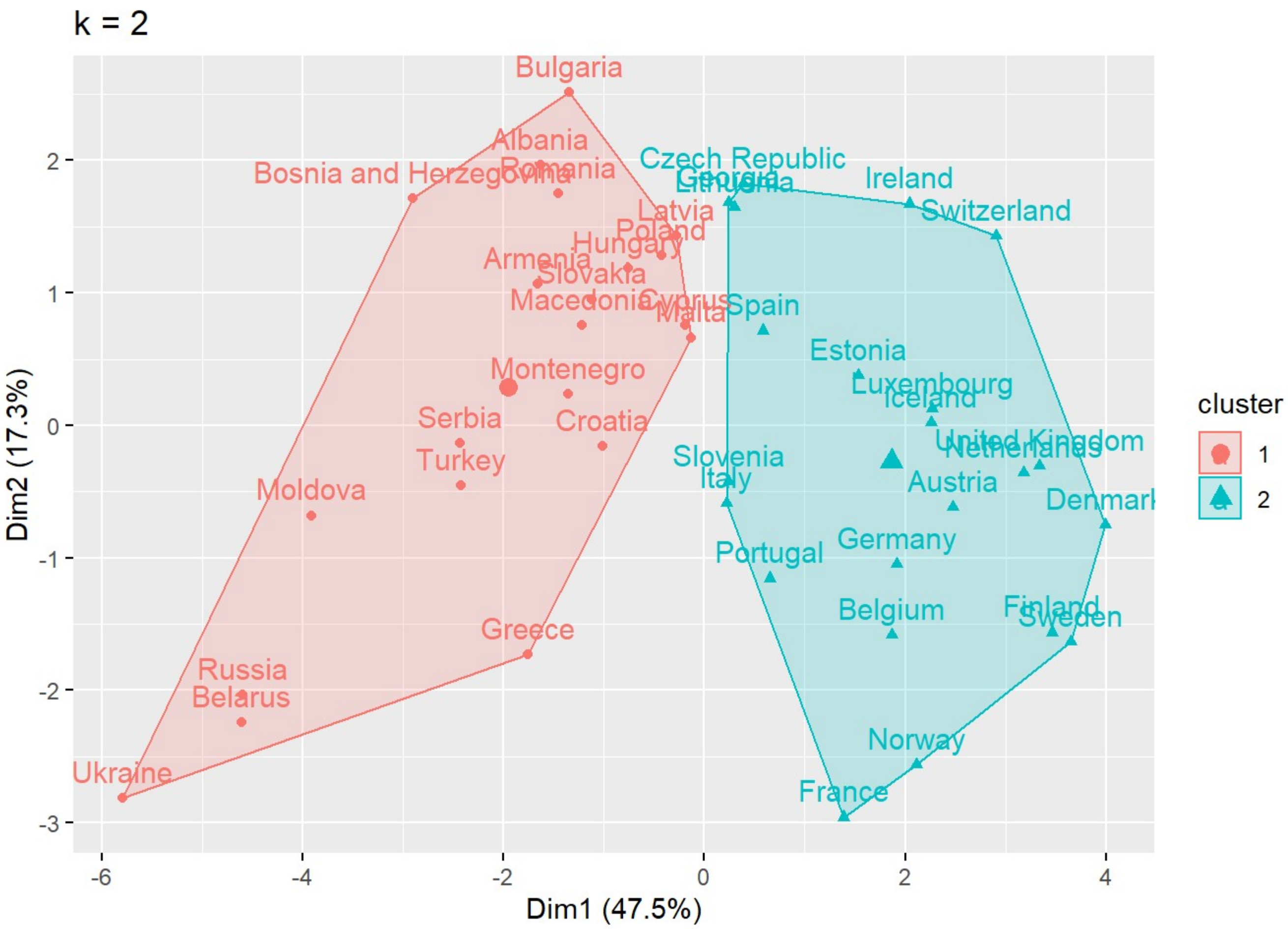
The average silhouette approach determines how well each country lies within its cluster. The silhouette score for the clustering is a value ranging from -1 to 1. The higher value, the more the distance between clusters there is and it indicates a good seperation of clusters.

Average silhouette method computes the average silhouette score of observations for different values of k, in this case from to 2 to 6. The optimal number of clusters k is the one that maximizes the average silhouette score over the range of possible values of k. The vertical line shows that the maximum average silhouette score is at k=3.

5.1 Clusters based on k-means clustering algorithm

5.1.1 Number of clusters k=2

On the next graph, we can see a clear difference between the cluster of Eastern and Central Europe and the cluster of Western Europe. The first cluster includes only post-socialist countries, Turkey, Greece and Malta.
The “freedom gap” between the two groups shows that the reforms to transition from socialist economy to market economy are still taking place and have a influence on the current economic institutions.



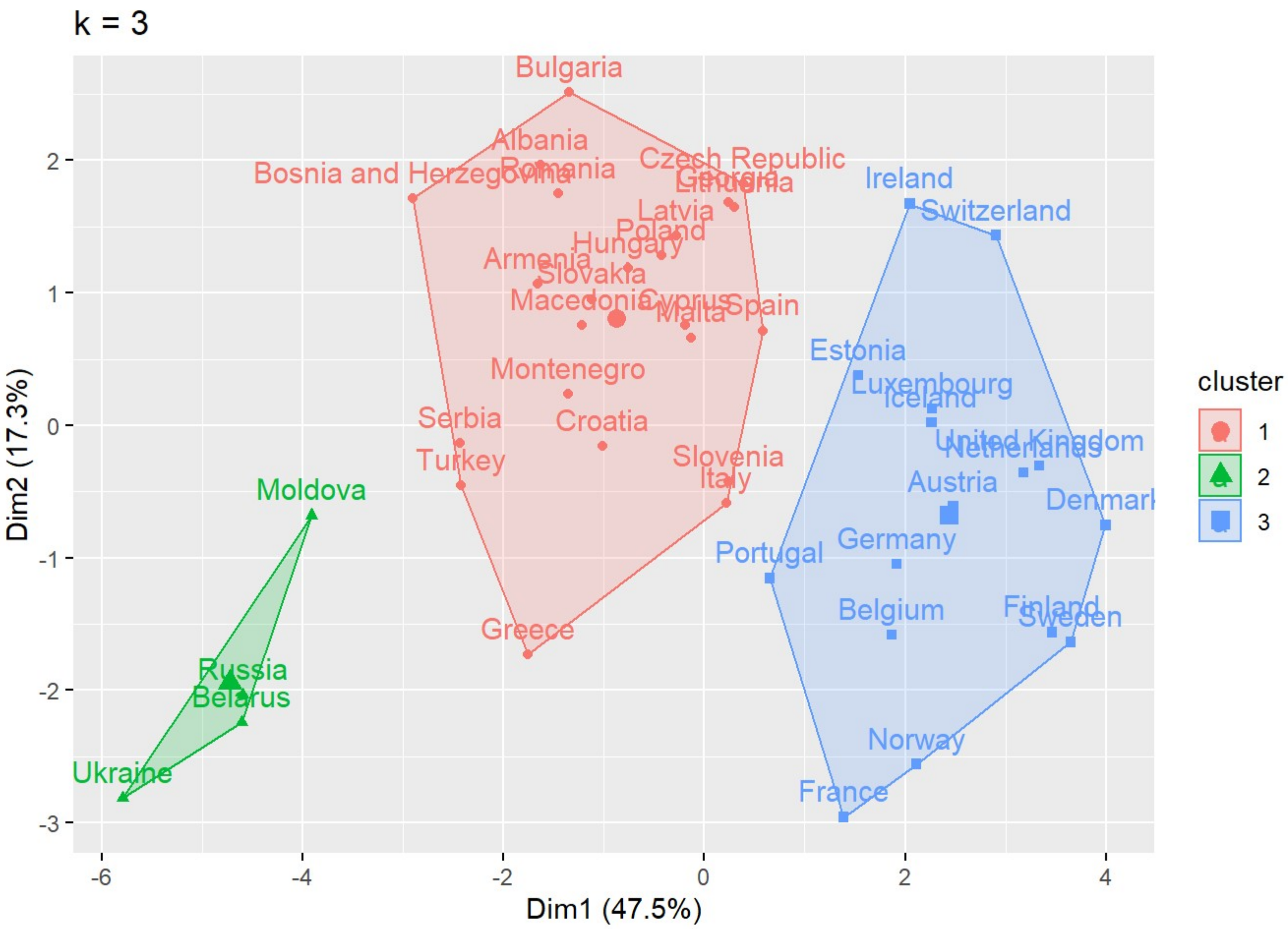
5.1.2 Number of clusters k=3

The next graph shows the countries divided into three clusters, which is the optimal number of clusters in this case. Here the clusters include the following countries:

Cluster 1 : Albania, Armenia, Bulgaria, Czech Republic, Latvia, Lithuania, Poland, Hungary, Bosnia and Herzegovina, Slovakia, Cyprus, Spain, Montenegro, Croatia, Malta, Italy

Cluster 2 : Ireland, Switzerland, Estonia, Luxembourg, Portugal, Germany, Iceland, Denmark, Sweden, Finland, United Kingdom, Netherlands, Belgium, France, Serbia, Turkey, Greece

Cluster 3 : Moldova, Ukraine, Russia, Belarus



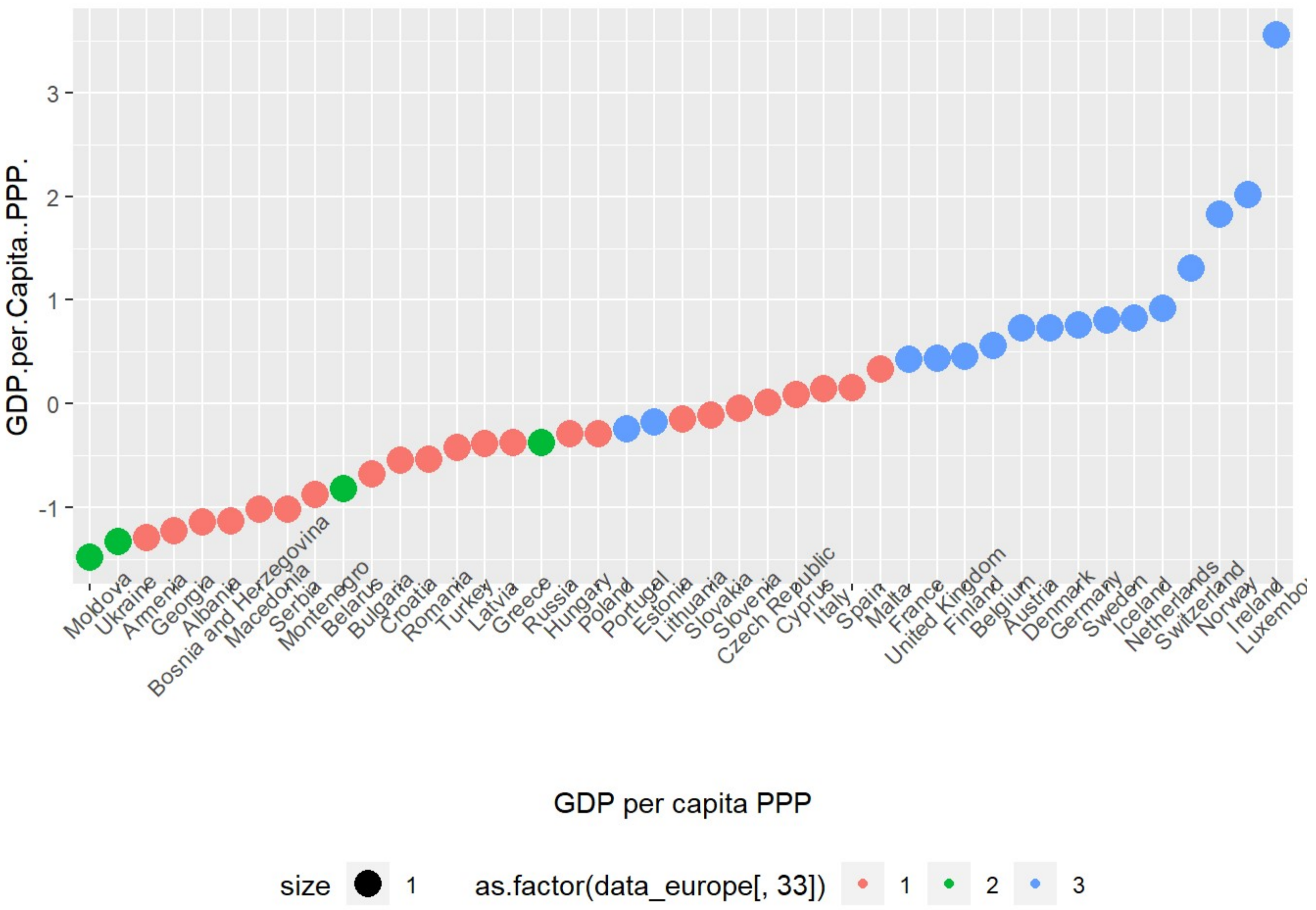
A new cluster forms with Eastern European countries, non-members of the European union, which could be explained by the different institutions, rules and laws, which influence the economic freedom. The members of the European Union have share common laws and institutions, which converge with time.

6 The relationship between economic freedom and economic performance

Comparing the clusters based on the Economic Freedom Index and the measure of economic performance Gross Domestic Product (GDP) per purchasing power parity (PPP), in current international dollars for 2019.

GDP itself is the primary measure of a country’s economic productivity that shows the market value of all goods and services produced during a certain time period. It is a commonly used measure for economic performance.

GDP per capita shows how much economic production value can be attributed to each individual citizen. Alternatively, this translates to a measure of national wealth since GDP value per person serves as a prosperity measure.



As depicted on the graph, it is clear that greater economic freedom is strongly correlated with higher GDP per capita (PPP).

Counties in cluster 3, which are associated with greater economic freedom have as well a higher GDP per capita. The exceptions are Portugal and Poland, which do not perform as economically well as expected based on their cluster membership.

Greece and Montenegro are the other exeptions in cluster 2, which have a better economic performance as expected, one possible explanation could be that both economies are highly dependent on the tourism industry.

6.1 Centroids of Economic Freedom Index variables

The following table shows the mean values of all the variables for each cluster, included in the Economic Freedom Index.

```
                      Cluster_1 Cluster_2 Cluster_3
Property.Rights              64        52        85
Judical.Effectiveness        48        39        74
Government.Integrity         43        32        81
Tax.Burden                   77        86        60
Gov.t.Spending               52        52        36
Fiscal.Health                82        87        88
Business.Freedom             70        72        82
Labor.Freedom                63        53        60
Monetary.Freedom             81        66        82
Trade.Freedom                85        77        86
Investment.Freedom           75        38        85
Financial.Freedom            62        30        74
```

6.2 Economic variables of different clusters

The following table shows the mean values of all economic variables for each cluster, included in the dataset.

```
                            Cluster_1 Cluster_2 Cluster_3
Income.Tax.Rate....                25        16        44
Corporate.Tax.Rate....             18        17        22
Tax.Burden...of.GDP                32        28        38
Gov.t.Expenditure...of.GDP         39        40        46
Population..Millions.              14        50        19
GDP..Billions..PPP.               427      1144       909
GDP.Growth.Rate....                 4         3         3
X5.Year.GDP.Growth.Rate....         3         1         2
GDP.per.Capita..PPP.            25896     15285     53952
Unemployment....                   11         5         6
Inflation....                       2         8         2
FDI.Inflow..Millions.            4213      7244     16097
Public.Debt....of.GDP.             64        45        61
```

Clear differences, which can help us differentiate between the clusters, occur in GDP per capita, the growth rate in the last 5 years, unemployment, inflation and foreign direct investments inflow (FDI.Inflow..Millions.).

7 Conclusions

Economic freedom matters, today’s more developed economies in Europe have adopted economic policies which make them score high in the Economic Freedom index. A clear gap between Eastern and Western Europe exists, which could be explained by the institutional differences. Membership in the European Union is associated with more economic freedom and better economic achievement.

The cluster analysis based on principal components in this project could be used for further research on the relationship between economic freedom and economic performance.

8 References

Acemoglu, D. (2009). Introduction to modern economic growth. Princton University Press Miller T. , Kim A., Roberts J. (2019), 2019 Index of economic freedom, The Heritage Foundation Kinnunen, J., Georgescu, I., & Tamminen, L. Do economic freedoms create national wealth?. In Real Option Workshop (p. 13). Johnson, R. and Wichern, D. W. (2007), Applied Multivariate Statistical Analysis, 6th Edition, Prentice Hall, New Jersey Georgescu, I., Androniceanu, A., & Kinnunen, J. A Computational Analysis Of Economic Freedom Indicators and GDP in EU states.