Linear ordering and cluster analysis

Author: Mateusz Kucharz

Table of contents

1. Description of variables

The data I will be working on comes from the websites https://www.gapminder.org/data/, https://www.numbeo.com/, https://www.macrotrends.net/ and are presented for 25 countries, 5 from each continent.

Europe – Poland, Spain, Norway, Greece, Netherlands

North America – Cuba, United States, Canada, Mexico, Jamaica

South America – Peru, Brazil, Argentina, Venezuela, Chile

Asia – China, India, Japan, South Korea, Singapore

Africa – Namibia, Nigeria, South Africa, Egypt, Kenya

Variables:

1. Life_expectancy – the expected lifespan in the country.
2. Crime_rate – the level of crime in the country, data are presented on a 0-100 scale.
3. Cost_of_living – the cost of living index in the country, created in comparison to the price of a basket of goods in New York.
4. Household_income – the average monthly household income.
5. Emissions_of_co2 – the number of tons of carbon dioxide emissions by the country.

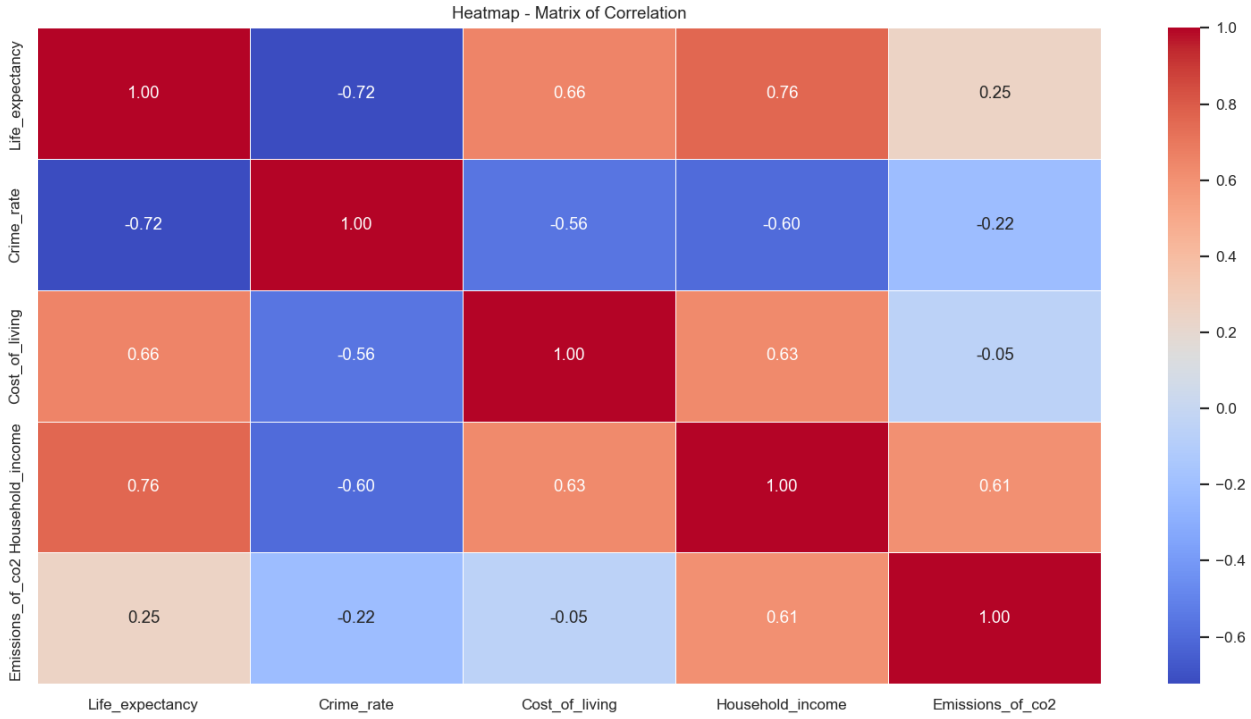
2. Descriptive statistics

	Life_expectancy	**Crime_rate**	**Cost_of_living**	**Household_income**	**Emissions_of_co2**
mean	76.33	47.93	51.08	11908.8	889,402.5
std	6.43	18.15	20.46	9424.21	2,289,944
25%	73	31.8	38.7	4430	46,579

median	77	45.2	43.1	8570	154,536
75%	81	63.8	67.6	17100	414,139
min	64.3	20.7	24.6	1100	3953
max	84.9	84.5	101.4	37300	10,944,690
cv	0.08	0.38	0.4	0.79	2.57

These statistics show that life in the countries studied varies significantly. The average life expectancy is about 76 years, but there are areas with longer and shorter lifespans. Crime rates also vary, as do living costs. Household incomes differ significantly, but most people earn less than the average. Carbon dioxide emissions are very diverse and have the highest coefficient of variation among the presented variables.

Correlation



The variables are moderately or weakly correlated with each other; there is no very high correlation present anywhere. It is worth paying attention to, for example, the correlation between the variables Cost_of_living and Crime_rate, which may suggest that the higher the cost of living, the more developed the country is and the more effectively it deals with crime.

3. Linear ordering

The goal of conducting linear ordering will be to check how the ranking of countries will differ in terms of the variables presented earlier when using the method of standardized sums and the Hellwig method.

Input data:

	Life_expectancy	Crime_rate	Cost_of_living	Household_income	Emissions_of_co2
Venezuela	75.1	84.5	27.2	7520	72509
South Africa	64.3	77.5	42.9	4430	393242
Brazil	73.6	68.9	40.2	5980	414139
Peru	77.4	68.2	38.7	5810	46579
Namibia	65.4	67.2	43.1	3770	3953
Jamaica	76.2	66.0	57.8	4620	5836
Nigeria	64.5	63.8	31.0	1100	111978
Argentina	74.6	61.8	33.0	11900	154536
Kenya	66.4	61.7	40.2	2230	19447
Mexico	73.0	54.0	35.7	4310	383131
United States	77.0	47.2	71.1	25300	4320533
Egypt	71.0	46.9	29.5	2500	210752
Chile	79.1	45.2	43.6	8450	84828
India	70.8	43.3	24.6	2240	2200836
Greece	80.4	40.3	55.7	8570	51002
Canada	81.5	39.7	67.6	22100	516874
Norway	83.2	35.4	101.4	26700	36177
Spain	81.5	32.0	53.8	15000	202706
China	84.5	31.8	40.0	37300	10944686
Singapore	84.9	30.6	81.1	23000	43705
Poland	76.7	28.5	40.0	8890	279224
South Korea	82.8	28.0	78.2	17100	569682
Netherlands	81.0	27.6	73.7	19600	130315
Cuba	78.7	27.5	43.5	13700	24328
Japan	84.7	20.7	83.3	15600	1014065

Method of standardized sums

The first step is to identify stimulants, destimulants, and nominants. In this case, Life_expectancy and Household_income are considered as stimulants, Crime_rate, Cost_of_living, and Emissions_of_co2 as destimulants, and there are no nominants.

I convert destimulants into stimulants by multiplying their values by -1. I standardize the data according to the formula $zij = \frac{x_{ij} - \overline{x_j}}{s_{xj}}$. I calculate the row sums and substitute them into the formula $s2_i = \frac{s_i - \min(s_i)}{\max\{s_i - \min(s_i)\}}$ and then sort the data according to the $s2_i$ column.

The results are as follows. As expected, African countries occupy the lower positions along with a few countries from South America. The best country in terms of the discussed variables turned out to be Cuba, followed by Singapore. Europe and Asia dominate the upper part. In the middle, more countries from South America can be seen.

Cuba	1.000000
Singapore	0.989357
Spain	0.958193
Netherlands	0.911042
Canada	0.879247
Poland	0.866313
Japan	0.858106
South Korea	0.842487
Norway	0.805243
China	0.766352
Chile	0.754329
Greece	0.739355
Argentina	0.629758
Egypt	0.533180
Peru	0.499754
Mexico	0.489629
India	0.453009
United States	0.452816
Venezuela	0.414024
Brazil	0.361369
Jamaica	0.318549
Kenya	0.206028
Nigeria	0.186047
Namibia	0.135632
South Africa	0.000000

Hellwig method

The first steps are the same as in the case of the method of standardized sums. This involves converting variables into stimulants and standardization.

Then, I create a pattern object by selecting the largest values from each column and calculate the distances of the objects from the pattern using the formula $d_i = \sqrt{\sum_j (z_{ij} - d_j^+)^2}$. For each object, I create a possibly far distance $d_0 = \bar{d} + 2 * s(d_i)$ and check which objects are farthest from the possibly far distance $s_i = 1 - \frac{d_i}{d_0}$ and sort the table according to them.

Cuba remains in first place, but the rest of the positions have shifted somewhat. While the bottom of the table looks similar, meaning the African countries still rank the lowest, there have been changes at the top and in the middle. Spain occupies the 2nd position, and Singapore has dropped to 5th place.

Cuba	0.747785
Spain	0.744245
Canada	0.740163
Netherlands	0.698382
Singapore	0.685361
Poland	0.646652
South Korea	0.637319
Chile	0.603874
Greece	0.591909
Japan	0.577574
United States	0.569709
Argentina	0.534239
Norway	0.505223
Mexico	0.403839
Peru	0.389056
Egypt	0.370216
Brazil	0.328999
Jamaica	0.300794
China	0.268450
Venezuela	0.240305
Kenya	0.143803
Namibia	0.099857
Nigeria	0.059117
South Africa	-0.017590

Summary

The rankings from both methods differ slightly. It can be seen that comparing distances from the pattern provides slightly different solutions than the non-pattern method. Cuba remains in first place in both methods, the same with the Republic of South Africa, which stays last. The biggest difference that can be noticed is for China, where in the method of standardized sums it is in 10th place and in the Hellwig method it drops to 20th. This indicates far distances from the pattern. Most countries fall or rise by 2 or 3 positions. In the upper part of the table in both cases, there are the most countries from Europe.

	rank_s	rank_h
Cuba	1	1
Singapore	2	5
Spain	3	2
Netherlands	4	4
Canada	5	3
Poland	6	6
Japan	7	10
South Korea	8	7
Norway	9	13
China	10	20
Chile	11	8
Greece	12	9
Argentina	13	12
Egypt	14	16
Peru	15	15
Mexico	16	14
India	17	17
United States	18	11
Venezuela	19	21
Brazil	20	18
Jamaica	21	19
Kenya	22	22
Nigeria	23	24
Namibia	24	23
South Africa	25	25

4. Cluster analysis

Through cluster analysis, I want to answer the question of what groups of countries can be distinguished in terms of characteristics: life expectancy, crime rate, cost of living, household income.

Input data:

	Life_expectancy	Crime_rate	Cost_of_living	Household_income
Venezuela	75.1	84.5	27.2	7520
South Africa	64.3	77.5	42.9	4430
Brazil	73.6	68.9	40.2	5980
Peru	77.4	68.2	38.7	5810
Namibia	65.4	67.2	43.1	3770
Jamaica	76.2	66.0	57.8	4620
Nigeria	64.5	63.8	31.0	1100
Argentina	74.6	61.8	33.0	11900
Kenya	66.4	61.7	40.2	2230
Mexico	73.0	54.0	35.7	4310
United States	77.0	47.2	71.1	25300
Egypt	71.0	46.9	29.5	2500
Chile	79.1	45.2	43.6	8450
India	70.8	43.3	24.6	2240
Greece	80.4	40.3	55.7	8570
Canada	81.5	39.7	67.6	22100
Norway	83.2	35.4	101.4	26700
Spain	81.5	32.0	53.8	15000
China	84.5	31.8	40.0	37300
Singapore	84.9	30.6	81.1	23000
Poland	76.7	28.5	40.0	8890
South Korea	82.8	28.0	78.2	17100
Netherlands	81.0	27.6	73.7	19600
Cuba	78.7	27.5	43.5	13700
Japan	84.7	20.7	83.3	15600

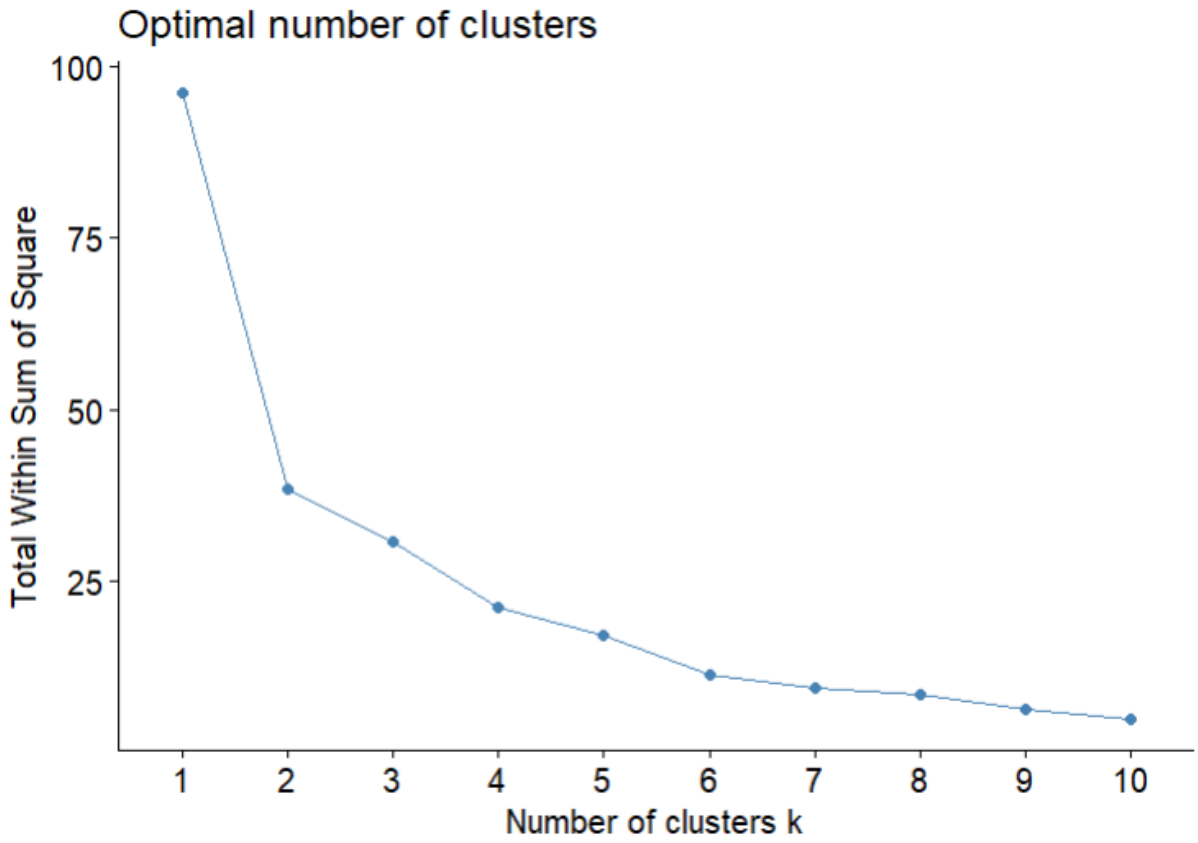
K-medoid

I am using the k-medoids method due to its greater resistance to outliers. After introducing the variables, I standardize them and check for the possible presence of outliers using the 3-sigma method.

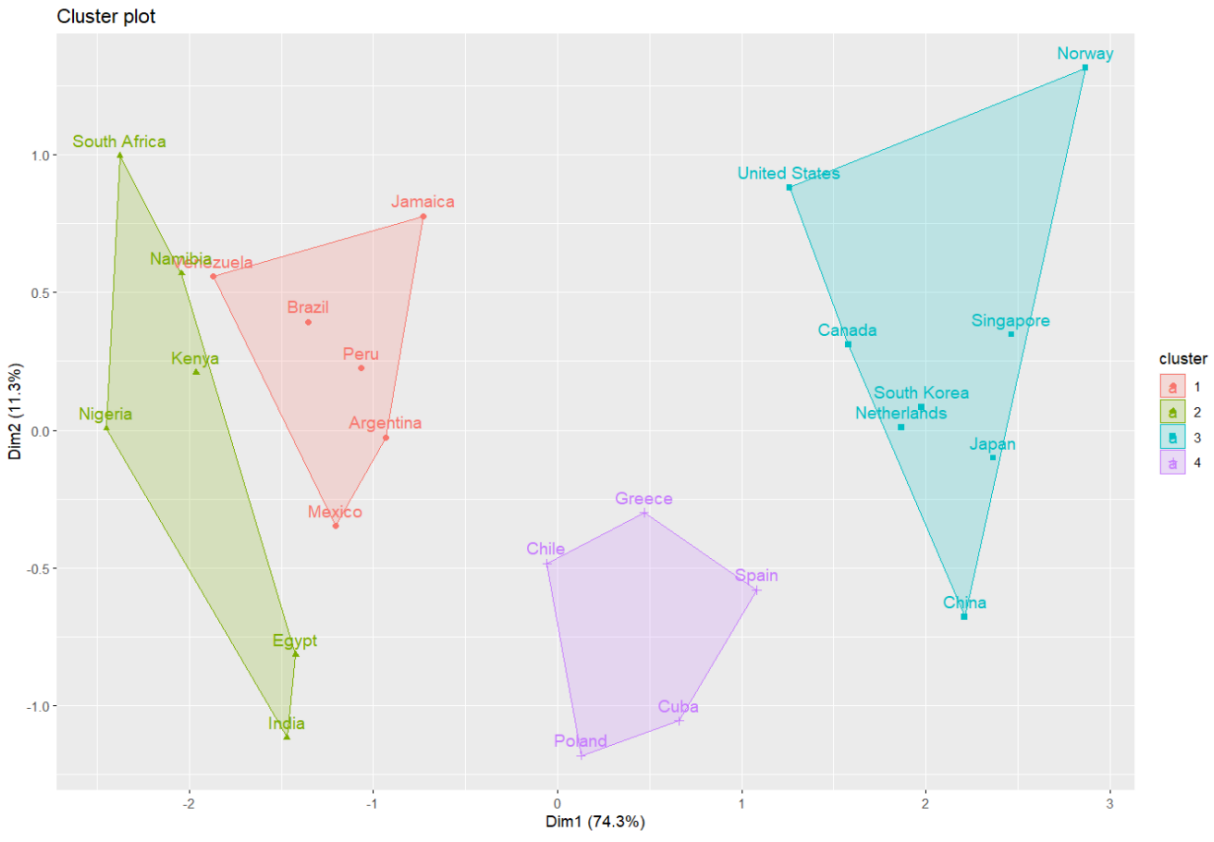
Life_expectancy	Crime_rate	Cost_of_living	Household_income
Min. :-1.8725	Min. :-1.5004	Min. :-1.2939	Min. :-1.1469
1st Qu.:-0.5185	1st Qu.:-0.8889	1st Qu.:-0.6048	1st Qu.:-0.7936
Median : 0.1040	Median :-0.1505	Median :-0.3898	Median :-0.3543
Mean : 0.0000	Mean : 0.0000	Mean : 0.0000	Mean : 0.0000
3rd Qu.: 0.8043	3rd Qu.: 0.8743	3rd Qu.: 0.8075	3rd Qu.: 0.5508
Max. : 1.3334	Max. : 2.0148	Max. : 2.4593	Max. : 2.6943

All values are less than 3 in absolute terms, so I determine that there are no outliers present.

I use the Elbow method to determine the optimal number of clusters.



It can be observed that a break occurs at 4, so I choose 4 as the number of clusters. Then, I use the k-medoids method and create a cluster plot.



According to the graph, it can be determined that the groups have been divided in terms of the degree of development of the countries. The axes of the graph resulted from the reduction of dimensionality to two and illustrate some variability of the input variables.

Cluster 1: Namibia, Kenya, Nigeria, RPA, Egypt, India

In this group of countries, one can expect a lower expected life expectancy and household income, with a higher crime rate and average/low cost of living. This cluster includes developing countries.

Country	Life_expectancy	Crime_rate	Cost_of_living	Household_income
Namibia	65.4	67.2	43.1	3770
Kenya	66.4	61.7	40.2	2230
Nigeria	64.5	63.8	31.0	1100
RPA	64.3	77.5	42.9	4430
Egypt	71.0	46.9	29.5	2500
India	70.8	43.3	24.6	2240

Cluster 2: Venezuela, Brasil, Peru, Jamaica, Argentina, Mexico

Countries should be characterized by a longer life expectancy, higher/average criminal activity, and similar sizes of living costs and household income. The countries that fall into this group can be considered developing or middle-income countries.

Country	Life_expectancy	Crime_rate	Cost_of_living	Household_income
Venezuela	75.1	84.5	27.2	7520
Brasil	73.6	68.9	40.2	5980
Peru	77.4	68.2	38.7	5810
Jamaica	76.2	66.0	57.8	4620
Argentina	74.6	61.8	33.0	11900
Mexico	73.0	54.0	35.7	4310

Cluster 3: Chile, Poland, Cuba, Spain, Greece

Here, one can observe a high life expectancy, low/average crime activity, higher/average living costs, and household incomes around the average. We are dealing with middle-income countries.

Country	Life_expectancy	Crime_rate	Cost_of_living	Household_income
Chile	79.4	45.2	43.6	8450
Greece	80.4	40.3	55.7	8570
Spain	81.5	32.0	53.8	15000
Poland	76.7	28.5	40.0	8890
Cuba	78.7	27.5	43.5	13700

Cluster 4: Norway, USA, Canada, Singapore, South Korea, Netherlands, Japan, China

High life expectancy, average/low crime rate, higher living costs, high household income. The magnitudes of these variables indicate developed countries.

Country	Life_expectancy	Crime_rate	Cost_of_living	Household_income
Norway	83.2	35.4	101.4	26700
USA	77.0	47.2	71.1	25300
Canada	81.5	39.7	67.6	22100
Singapore	84.9	30.6	81.1	23000
South Korea	82.8	28.0	78.2	17100
Netherlands	81.0	27.6	73.7	19600
Japan	84.7	20.7	83.3	15600
China	84.5	31.8	40.0	37300