AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

Wydział Zarządzania

Projekt ze statystycznej analizy danych

Porządkowanie liniowe oraz Analiza skupień

Autor: *Łukasz Pyrek*

Kierunek studiów: Informatyka i Ekonometria

Spis treści

1. Wstęp

W dzisiejszym złożonym świecie, analiza danych staje się kluczowym narzędziem do zrozumienia współczesnych społeczeństw i ich interakcji z otoczeniem. Niniejszy projekt ma na celu przeprowadzenie wszechstronnej analizy danych z 27 krajów UE, przy uwzględnieniu szóstki kluczowych wskaźników: PKB per capita, długość życia, wskaźnik HICP (Indeks Harmonizowany Cen Konsumpcyjnych), zadowolenie z życia, emisje gazów cieplarnianych według klasyfikacji NACE oraz procent bezrobocia.

2. Opis i wstępna analiza danych

Głównym celem tego projektu jest porównanie 27 krajów poprzez analizę wymienionych wskaźników:

1. **Procent osób z wykształceniem średnim**(education_prc.csv): Ten zestaw danych mierzy odsetek osób w wieku od 15 do 64 lat, które uzyskały wykształcenie co najmniej średnie w różnych krajach. Jest to wskaźnik poziomu edukacji w populacji, odzwierciedlający dostęp do edukacji wyższej oraz inwestycje w kapitał ludzki (stymulanta).

2. **PKB per capita** (GDP.csv): Dane te przedstawiają Produkt Krajowy Brutto na mieszkańca dla różnych krajów, wyrażony w euro. PKB per capita jest ważnym wskaźnikiem ekonomicznym, który pomaga ocenić poziom życia i dobrobyt ekonomiczny w różnych regionach, uwzględniając wartość wszystkich wyprodukowanych towarów i usług (destymulanta).

3. **Indeks Harmonizowanych Cen Konsumpcyjnych** (hicp_prc.csv): Ten zestaw danych zawiera informacje o zmianach cen konsumpcyjnych w różnych krajach. Jest to miernik inflacji, który pozwala na monitorowanie i porównywanie poziomu cen dóbr i usług w różnych regionach (destymulanta).

4. **Niespełnione potrzeby medyczne** (med_unmet_prc.csv): Dane te mierzą procent populacji, która zgłaszała niespełnione potrzeby medyczne w różnych krajach. Jest to wskaźnik dostępu do opieki zdrowotnej i poziomu zaspokajania potrzeb zdrowotnych społeczeństwa (destymulanta).

5. **Emisje gazów cieplarnianych na mieszkańca** (pollutionPerCapita.csv): Ten zestaw danych przedstawia emisję gazów cieplarnianych na mieszkańca w tonach dla różnych krajów. Jest to wskaźnik wpływu na środowisko i pokazuje, jak działalność gospodarcza i codzienne życie wpływają na zmiany klimatyczne (destymulanta).

6. **Zadowolenie z życia** (satisfaction.csv): Dane te pokazują ocenę zadowolenia z życia w skali od 1 do 10 w różnych krajach. Jest to subiektywny wskaźnik, który odzwierciedla ogólne postrzeganie jakości życia i szczęścia przez mieszkańców (destymulanta).

7. **Stopa bezrobocia** (unemployment_prc.csv): Ten zestaw danych zawiera informacje o stopie bezrobocia w różnych krajach, wyrażonej jako procent populacji aktywnej zawodowo, która jest bez pracy. Jest to kluczowy wskaźnik zdrowia gospodarki, wskazujący na dostępność miejsc pracy i stabilność ekonomiczną (destymulanta).

Tabela 1. Statystyki opisowe zmiennej education_prc

Średnia	Mediana	Minimalna	Maksymalna
78,744	80,500	60,400	88,700
Odch.stand.	**Wsp. zmienności**	**Skośność**	**Kurtoza**
7,8521	0,099716	-1,1246	0,50703
Percentyl 5%	**Percentyl 95%**	**Zakres Q3-Q1**	**Brakujące obs.**
60,720	88,300	6,9000	0

Z uwagi na fakt, że zmienna znajduje się na granicy dopuszczalności według współczynnika zmienności, a jej znaczenie dla naszego projektu jest istotne, zdecydowaliśmy o włączeniu jej do dalszej analizy.

Tabela 2. Statystyki opisowe zmiennej GDP

Średnia	Mediana	Minimalna	Maksymalna
35131	27660	13270	118710
Odch.stand.	**Wsp. zmienności**	**Skośność**	**Kurtoza**
22168	0,63102	2,1709	5,7493
Percentyl 5%	**Percentyl 95%**	**Zakres Q3-Q1**	**Brakujące obs.**
13879	99719	27785	0

Tabela 3. Statystyki opisowe zmiennej med_unmet_prc

Średnia	Mediana	Minimalna	Maksymalna
2,5556	1,8000	0,10000	9,1000
Odch.stand.	**Wsp. zmienności**	**Skośność**	**Kurtoza**

2,4997	0,97816	1,3828	1,1985
Percentyl 5%	**Percentyl 95%**	**Zakres Q3-Q1**	**Brakujące obs.**
0,14000	9,0600	2,7000	0

Tabela 4. Statystyki opisowe zmiennej hicp

Średnia	**Mediana**	**Minimalna**	**Maksymalna**
126,55	124,00	111,93	153,14
Odch.stand.	**Wsp. zmienności**	**Skośność**	**Kurtoza**
11,223	0,088690	0,73271	-0,39248
Percentyl 5%	**Percentyl 95%**	**Zakres Q3-Q1**	**Brakujące obs.**
112,24	150,93	18,893	0

Z powodu niskiego współczynnika zmienności **odrzucamy** zmienną.

Tabela 5. Statystyki opisowe zmiennej pollutionPerCapita

Średnia	**Mediana**	**Minimalna**	**Maksymalna**
2,1398	2,0740	1,1070	3,8470
Odch.stand.	**Wsp. zmienności**	**Skośność**	**Kurtoza**
0,62064	0,29004	0,67228	0,48730
Percentyl 5%	**Percentyl 95%**	**Zakres Q3-Q1**	**Brakujące obs.**
1,1444	3,5684	0,74700	0

Tabela 6. Statystyki opisowe zmiennej satisfaction

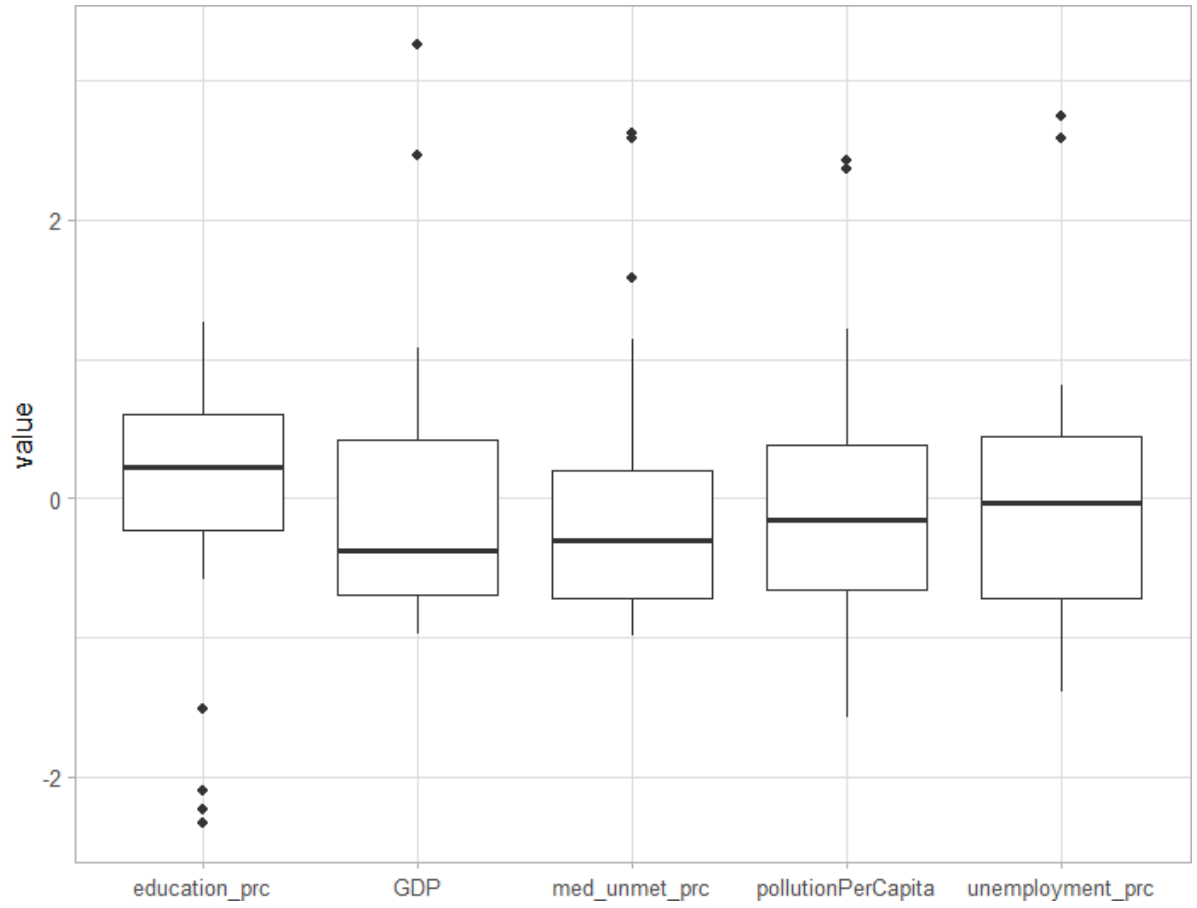
Średnia	**Mediana**	**Minimalna**	**Maksymalna**
7,1808	7,2000	5,6000	7,9000
Odch.stand.	**Wsp. zmienności**	**Skośność**	**Kurtoza**
0,47751	0,066498	-1,3083	2,7122
Percentyl 5%	**Percentyl 95%**	**Zakres Q3-Q1**	**Brakujące obs.**
5,9150	7,8300	0,55000	0

Z powodu niskiego współczynnika zmienności **odrzucamy** zmienną.

Tabela 7. Statystyki opisowe zmiennej unemployment_prc

Średnia	**Mediana**	**Minimalna**	**Maksymalna**
5,8577	5,8000	2,2000	12,900
Odch.stand.	**Wsp. zmienności**	**Skośność**	**Kurtoza**
2,6222	0,44766	1,1067	1,3663
Percentyl 5%	**Percentyl 95%**	**Zakres Q3-Q1**	**Brakujące obs.**
2,4450	12,760	3,3000	0

Wykres 1. Wykres pudełkowy zmiennych standaryzowanych



Na wykresie można zauważyć że zbiór danych posiada wiele wartości odstających. Może to spowodować trudności w zastosowaniu analizy skupień, dlatego decydujemy się zastosować metodę K-medoidów, która jest odporna na wartości odstające.

Tabela 1. Tablica korelacji zmiennych

education_prc	GDP	med_unmet_prc	pollutionPerCapita	unemployment_prc	
	Corr: -0.102	Corr: 0.178	Corr: 0.284	Corr: -0.367.	education_prc
Corr: -0.102		Corr: -0.239	Corr: 0.647***	Corr: -0.171	GDP
Corr: 0.178	Corr: -0.239		Corr: -0.021	Corr: 0.379.	med_unmet_prc
Corr: 0.284	Corr: 0.647***	Corr: -0.021		Corr: -0.328.	ollutionPerCapit
Corr: -0.367.	Corr: -0.171	Corr: 0.379.	Corr: -0.328.		nemployment_pr

Tabela również informuje o wyniku testu na istotność statystyczną korelacji z następującym zestawem hipotez:

$$H_0: corr = 0$$
$$H_1: corr \neq 0$$

Gwiazdki przy wartości korelacji oznaczają przedział w jakim znajduje się p-value policzone dla testu dla danej korelacji:

*** p-value < 0.001
** p-value < 0.01
* p-value < 0.05

Jedynie korelacja pomiędzy zmiennymi *pollutionPerCapita* i *GDP* jest istotnie różna od zera oraz wynosi ona 0.647. Nie jest to na tyle wysoka wartość, aby komplikowała ona zastosowanie użytych metod i interpretacje wyników.

3. Porządkowanie liniowe

Do porządkowania liniowego zastosujemy dwie techniki: Hellwiga i TOPSIS.

Metoda Hellwiga			Metoda TOPSIS		
rank	country	score	rank	country	score
1	SE	0,464	1	LU	0,736
2	AT	0,463	2	NL	0,665
3	NL	0,431	3	IE	0,663
4	DE	0,403	4	DE	0,642
5	BE	0,379	5	AT	0,641
6	SI	0,357	6	MT	0,624
7	DK	0,343	7	DK	0,620
8	FR	0,335	8	SE	0,617
9	CZ	0,318	9	BE	0,603
10	IE	0,315	10	CZ	0,588
11	HU	0,310	11	CY	0,560
12	LU	0,304	12	HU	0,542
13	MT	0,303	13	BG	0,530
14	SK	0,291	14	IT	0,522
15	CY	0,287	15	HR	0,520
16	HR	0,281	16	SI	0,507
17	FI	0,271	17	FR	0,506
18	LT	0,256	18	PL	0,505
19	BG	0,246	19	SK	0,484
20	LV	0,232	20	PT	0,482
21	PL	0,219	21	ES	0,479
22	RO	0,208	22	LT	0,467
23	IT	0,132	23	RO	0,405
24	PT	0,104	24	LV	0,389
25	EE	0,079	25	FI	0,388
26	ES	-0,022	26	EE	0,287
27	EL	-0,092	27	EL	0,183

Tabela 82. Porządkowanie liniowe metodą Hellwiga i TOPSIS

Metoda Hellwiga

Wykorzystując Metodę Hellwiga, kraje są oceniane na podstawie miary użyteczności wybranych cech, gdzie każdy kraj otrzymuje wynik odzwierciedlający jego pozycję względem modelu idealnego. W naszym projekcie Szwecja (SE) przoduje z wynikiem 0,464, sugerując najwyższą zgodność z modelem idealnym. Austria (AT) i Holandia (NL) plasują się tuż za liderem. Warto zauważyć, że w tej metodzie niektóre kraje mogą otrzymać ujemne wyniki, co wskazuje na dalekie oddalenie od idealnego modelu, jak to jest w przypadku Grecji (EL) z wynikiem -0,092.

Metoda TOPSIS

Metoda TOPSIS ocenia kraje na podstawie ich bliskości do idealnej i najgorszej wartości w zestawie danych. Luksemburg (LU) osiągnął najwyższy wynik 0,736, co wskazuje na najbliższą bliskość do idealnego rozwiązania według zastosowanych wskaźników. Po nim, z niewielką różnicą, znajdują się Holandia (NL) i Irlandia (IE). W tej metodzie wszystkie kraje otrzymują wyniki dodatnie, a pozycja kraju jest bezpośrednio proporcjonalna do jego wyniku, z Grecją (EL) na ostatniej pozycji z wynikiem 0,183.

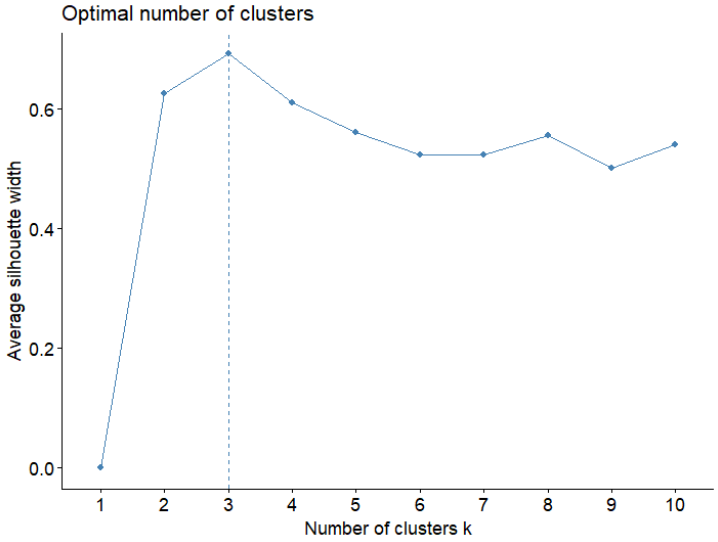
Interpretacja Wyników

Różnice w rankingu między metodami są znaczące i odzwierciedlają unikalne aspekty każdej techniki analitycznej. Wyniki Metody Hellwiga skupiają się na maksymalizacji miary użyteczności, podczas gdy TOPSIS koncentruje się na bliskości do idealnego rozwiązania. Zauważalna jest zmienność pozycji niektórych krajów między metodami, może to wynikać z dużego podobieństwa obiektów, wartości nie wiele się różnią.

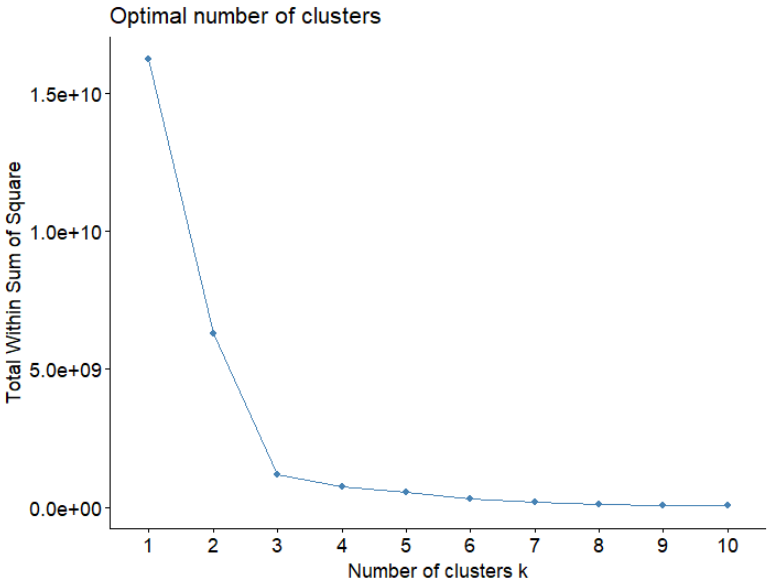
4. Analiza skupień

4.1 Metoda K-medoidów

Z uwagi na obecność licznych wartości odstających w naszym zbiorze danych, wybraliśmy metodę k-medoidów jako narzędzie klasterowania. Pierwszym krokiem w procesie jest ustalenie odpowiedniej liczby klastrów. Aby to osiągnąć, dokonamy przeglądu wykresów sumy kwadratów odległości wewnątrz klastrów (WSS) oraz analizy średniej wartości sylwetki (AWS), które posłużą jako wskaźniki w procesie decyzyjnym dotyczącym ilości grup.

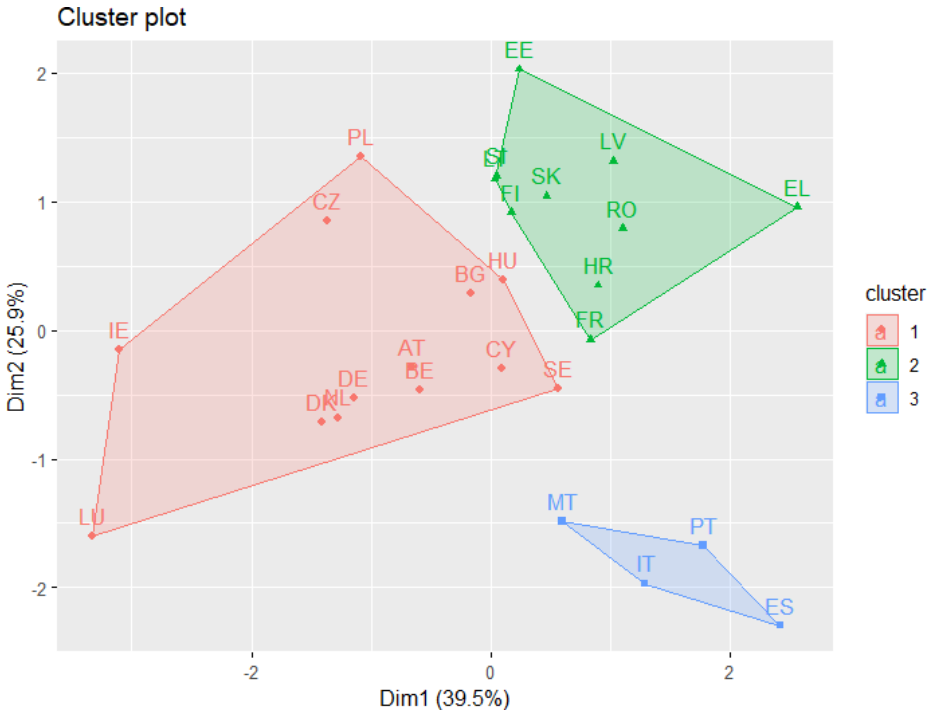


Wykres 2. Wykres AWS



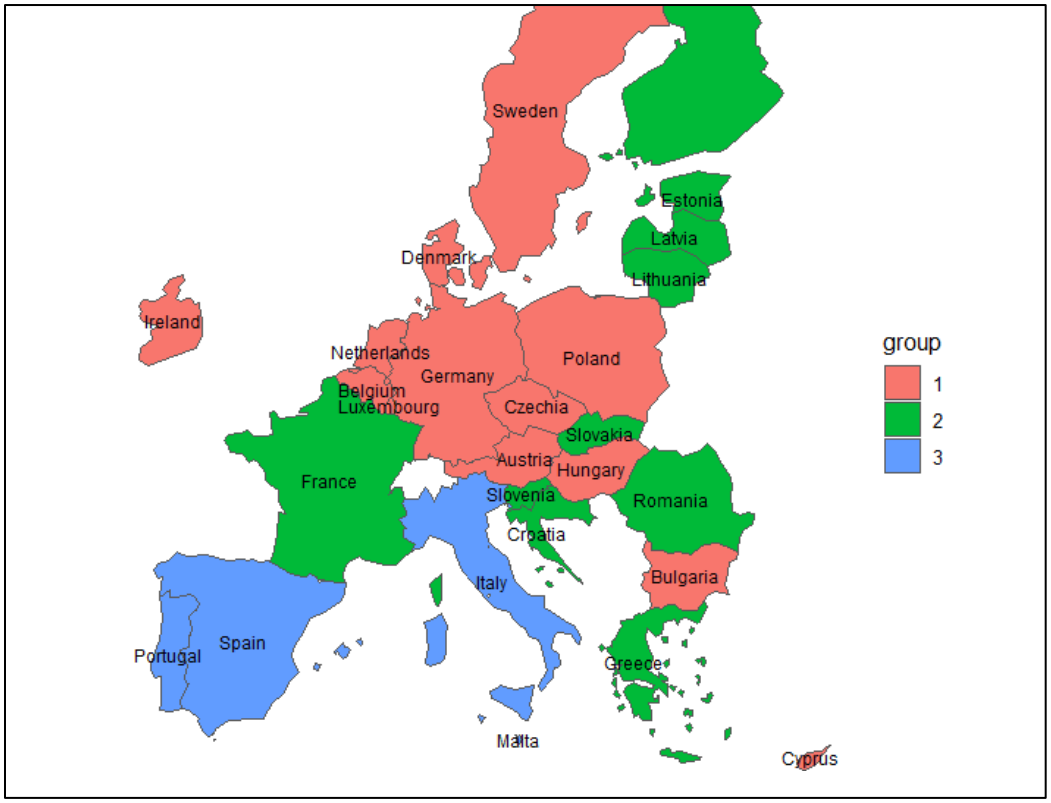
Wykres 3. Wykres WSS2

Oba wykresy jednoznacznie wskazują na liczbę klastrów równą 3. Wybierając liczbę klastrów chcemy maksymalizować średnia szerokość "silhouette", natomiast analizując wykres WSS wybieramy taką liczbę klastrów dla której suma kwadratów odległości przestaje gwałtownie maleć.



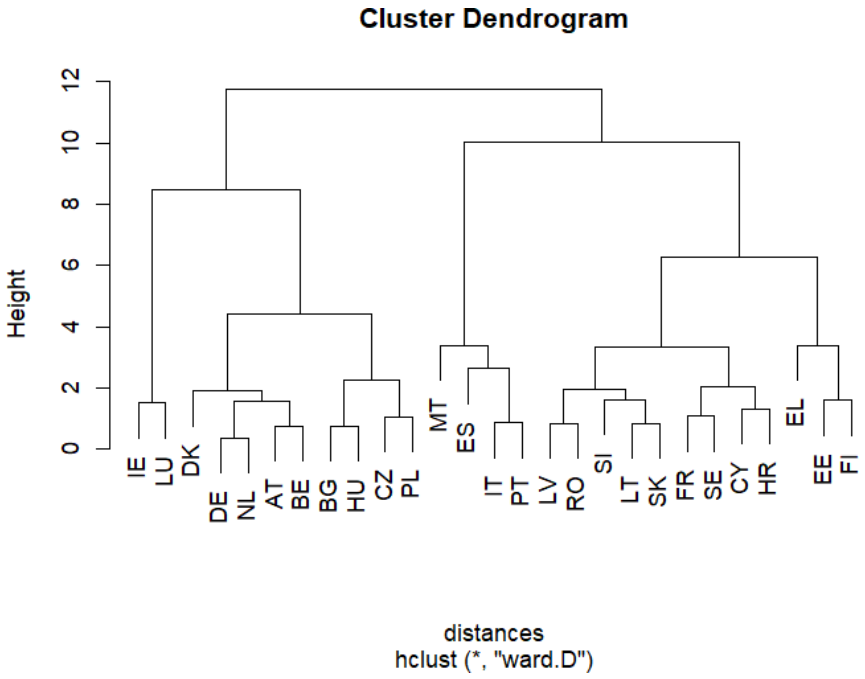
Wykres 4. Wykres klastrowy

Stosując Analizę Głównych Składowych (PCA) możemy narysować zmienne wielowymiarowe na płaszczyźnie dwuwymiarowej. Po zastosowaniu grupowania k-medoid klastry będą wyglądały następująco (Wykres 4).



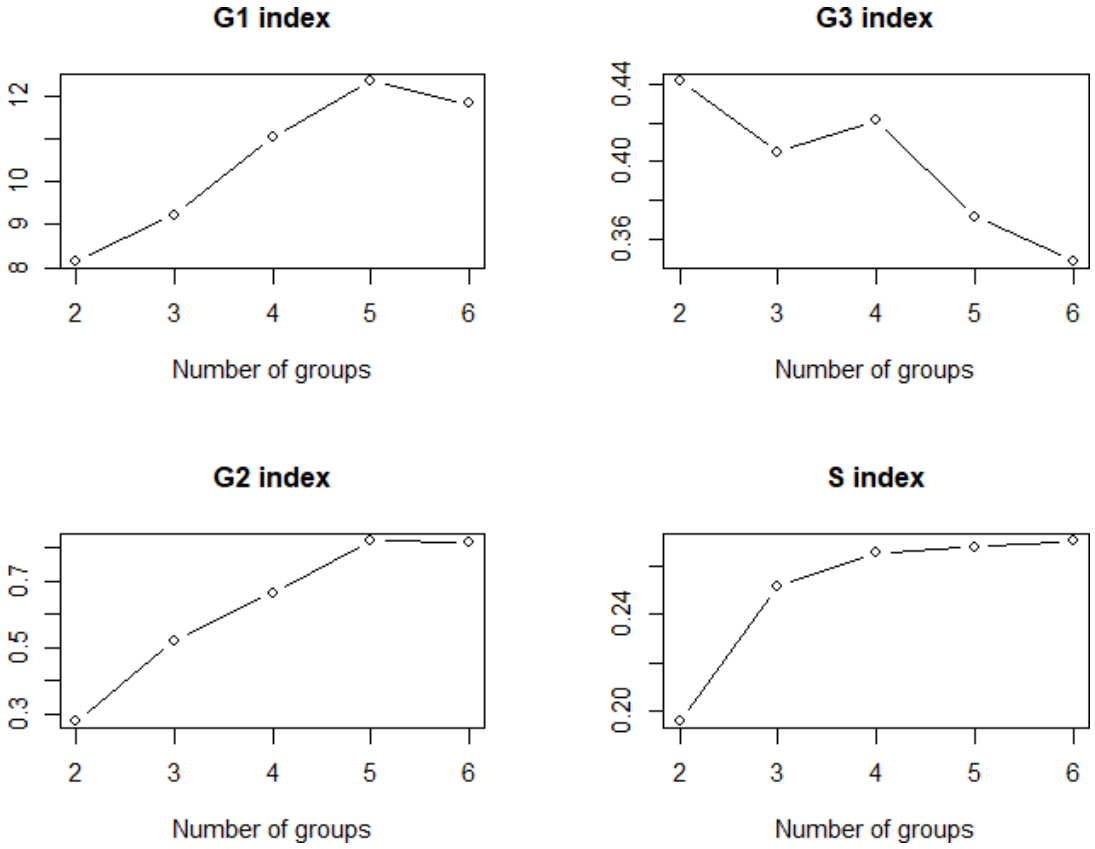
Wykres 5. Wykres podziału Państw na grupy metodą k-medoid

4.2 Metoda Warda



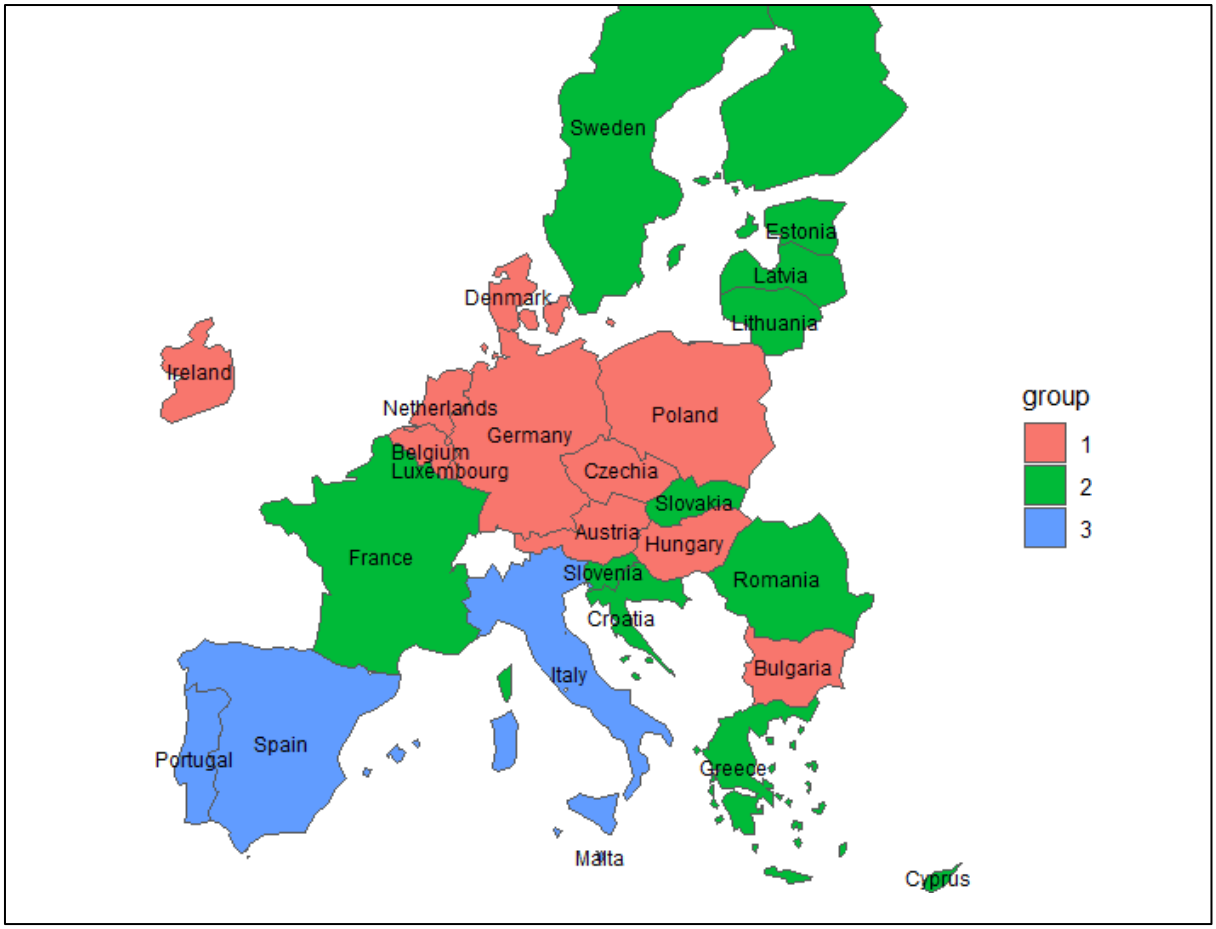
Wykres 6. Dendogram

Stosując metodę hierarchiczną również musimy wybrać liczbę grup, a dokładniej w którym miejscu uciąć dendrogram. Dla porównania analizy wyników wybieramy liczbę grup równą 3 tak jak w przypadku metody k-medoid. Jednak dla „upewnienia się" przeanalizujemy wartości indeksów G1, G2, G3 oraz S, które pomagają wybrać optymalną liczbę grup.



Wykres 7. Indeksy G1 G2 G3 oraz S

Dla indeksów G1, G2, S chcemy osiągnąć jak największą wartość, natomiast dla indeksu G3, jak najmniejszą. Indeksy sugerują 5 grup, jednakże założyliśmy a-priori liczbę grup równą 3, dla porównania wyników z metodą k-medoid.



Wykres 8. Wykres podziału Państw na grupy metodą Warda

4.3 Interpretacja wyników

Obie metody dają bardzo podobne wyniki. Jedyną różnicą jest inne przyporządkowanie Cypru oraz Szwecji. W metodzie k-medoid zostały przyporządkowane one do grupy 1, a w metodzie Warda do grupy 2.

Subiektywnie, lepszy wydaje się podział grup uzyskany metodą k-medoid. W przypadku grupowania Warda dziwnym wydaje się przyporządkowanie Grecji oraz Szwecji do jednej grupy. Wyniki porządkowania liniowego ustanowiły Grecje na ostatnim miejscu zarówno przy użyciu metody Helwiga jak i TOPSIS. Natomiast Szwecja jest najbliższa wzorcowi idealnemu, uzyskując pierwsze miejsce w metodzie Helwiga, dlatego przeanalizujemy grupy uzyskane metodą k-medoid uważając podział za ten właściwy.

GRUPA 1									
vars	**n**	**mean**	**sd**	**median**	**min**	**max**	**range**	**skew**	**kurtosis**
education_prc	13	80,3	4,000625	80,5	74,2	87,7	13,5	0,402633	-0,83461
GDP	13	49028,46	31376,94	47430	13270	118710	105440	0,843811	-0,30827
med_unmet_prc	13	1,076923	0,887087	1	0,1	2,7	2,6	0,475698	-1,37093
pollutionPerCapita	13	2,586308	0,735824	2,448	1,214	3,89	2,676	0,245888	-0,54483
unemployment_prc	13	4,476923	1,580693	4,2	2,2	7,6	5,4	0,492154	-0,86187
GRUPA 2									
vars	**n**	**mean**	**sd**	**median**	**min**	**max**	**range**	**skew**	**kurtosis**
education_prc	10	83,14	3,819308	83,5	77,3	88,7	11,4	-0,1253	-1,5423
GDP	10	25732	10352,12	22205	15010	48340	33330	1,017372	-0,30476
med_unmet_prc	10	4,88	2,644827	4,3	1,3	9,1	7,8	0,434842	-1,32321
pollutionPerCapita	10	1,991	0,40334	1,8225	1,551	2,794	1,243	0,700529	-1,00337
unemployment_prc	10	6,76	2,307572	6,65	3,4	12,5	9,1	1,198915	1,296924
GRUPA 3									
vars	**n**	**mean**	**sd**	**median**	**min**	**max**	**range**	**skew**	**kurtosis**
education_prc	4	62,7	2,906315	61,75	60,4	66,9	6,5	0,595572	-1,79376
GDP	4	29427,5	4506,406	30580	23530	33020	9490	-0,32539	-2,07853
med_unmet_prc	4	1,55	1,090871	1,5	0,3	2,9	2,6	0,09244	-1,97927
pollutionPerCapita	4	1,4985	0,301793	1,531	1,107	1,825	0,718	-0,2186	-1,9454
unemployment_prc	4	7,675	4,107209	7,45	2,9	12,9	10	0,120793	-1,88832

Tabela 8. Statystyki opisowe grup przyporządkowanych metodą k-medoid.

Grupa 1 charakteryzuje kraje o wysokim PKB per capita. Średnia wartość jest prawie 2 razy większa niż w innych grupach, jednakże odchylenie standardowe jest bardzo wysokie. Powodem może być występowanie w tej grupie krajów z największym oraz najmniejszym PKB per capita, czyli kolejno Luksemburg oraz Bułgaria. Grupa 1 posiada również najmniejszy % bezrobocia, % niespełnionych potrzeb medycznych oraz stosunkowo wysoki poziom edukacji. Jeżeli chodzi o zanieczyszczenia na mieszkańca jest ona średnio największa z grup, tak samo jak PKB. Zgadza się to ze współczynnikiem korelacji, który był największy właśnie pomiędzy tymi zmiennymi.

Grupa 1 wydaje się być najbardziej rozwinięta pod względem ekonomicznym, podczas gdy Grupa 3 ma niższy poziom edukacji i wyższe bezrobocie. Grupa 2 zajmuje pozycję pośrednią pod względem większości analizowanych zmiennych.