Efekt 5, Analiza skupień

Mateusz Nowak, Damian Okoń, Robert Zamiar

# Analiza skupień – Clustering

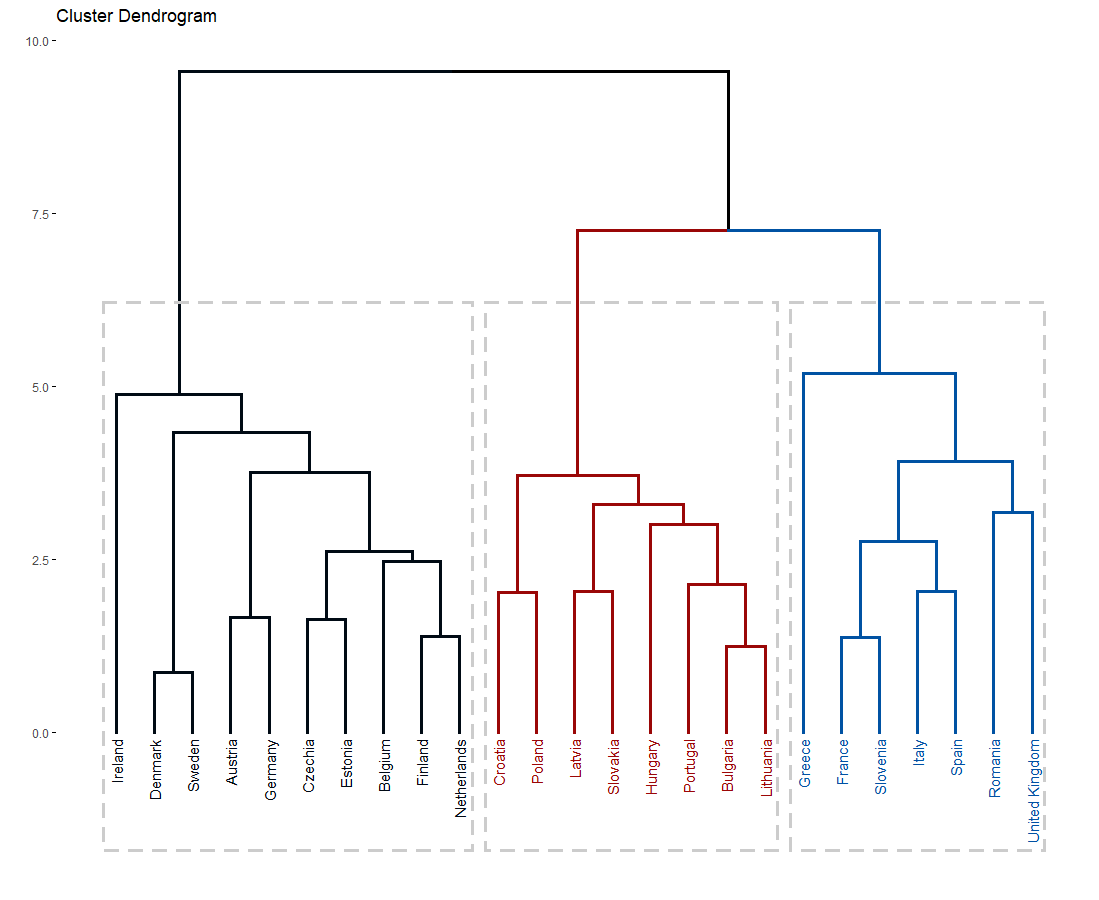
Drugą analizą wykonaną w ramach projektu jest analiza skupień. Przeprowadzono ją na dwa sposoby. Najpierw metodą grupowania hierarchicznego, a następnie metodą k-średnich.

# Pytanie badawcze

W jakie skupiska można połączyć kraje, przy użyciu dwóch różnych metod? Czy skupiska są takie same?

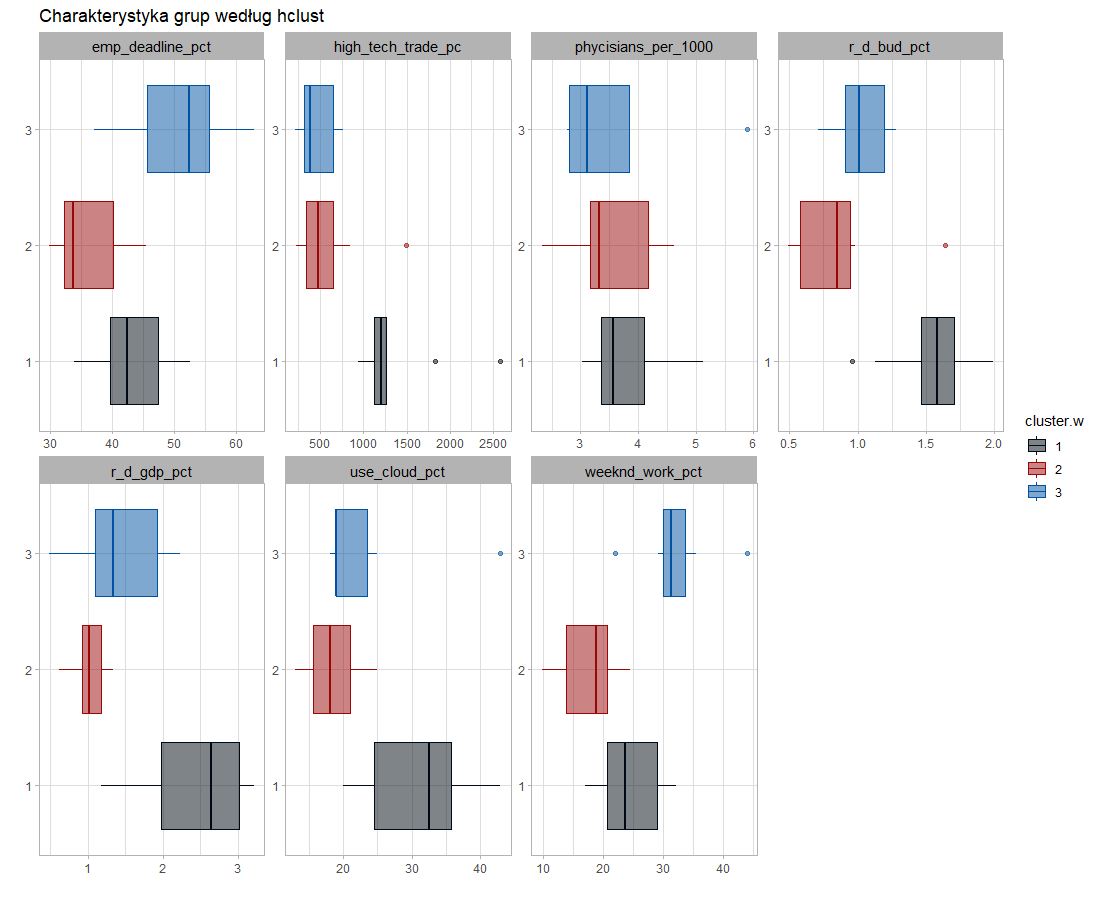
# Grupowanie hierarchiczne

W celu wykonania grupowania niezbędna jest macierz dystansu. Do jej utworzenia zastosowano metodę euklidesową przy pomocy języka R. Potem skorzystano z grupowania hclust, przy użyciu metody „ward.D2”. Wynik grupowania można przedstawić w postaci dendrogramu.



Rysunek 1. Grupowanie hierarchiczne - dendrogram

Zgodnie z powyższym wykresem pogrupowanie danych na trzy części wydaje się być najodpowiedniejsze. Można stwierdzić, że grupy całkiem dobrze pokrywają się z podziałem wynikającym z analizy PCA. Prawie wszystkie kraje grupy pierwszej to państwa, należące do pierwszych dwóch kwadrantów analizy składowych głównych. Jedynym wyjątkiem jest Wielka Brytania, która trafiła do grupy z krajami z kwadrantu czwartego. Do grupy drugiej z kolei trafiły głównie kraje postkomunistyczne, znajdujące się w trzecim kwadrancie analizy PCA. Do ostatniej grupy należą państwa czwartego kwadrantu, a także wcześniej wymieniona Wielka Brytania oraz Słowenia, która w analizie składowych głównych znajdowała się w kwadrancie trzecim.



Rysunek 2. Charakterystyka grup - grupowanie hierarchiczne

Przechodząc do charakterystyki grup, od razu można zobaczyć, że pierwsza grupa to kraje najlepiej rozwinięte oraz przeznaczające najwięcej środków na rozwój i nowe technologie. Charakteryzują się najwyższą wartością mediany handlu wysokimi technologiami oraz wydatkami na rozwój. Nawet bez wstępnego sprawdzenia członków danych grup można wysnuć wniosek, że to głównie kraje Europy zachodniej i północnej. Druga grupa to z kolei państwa przeznaczające najmniej funduszy na nowe technologie. W kategoriach, gdzie pierwsza grupa przodowała, druga grupa notuje najczęściej najsłabszy wynik. Można podejrzewać, że głównie są to państwa byłego bloku wschodniego. Ostatnia grupa charakteryzuje się wysokim odsetkiem ludności pracującej w ścisłych terminach oraz populacji pracującej w weekendy. Jest to najprawdopodobniej grupa, w skład której wchodzą kraje śródziemnomorskie.

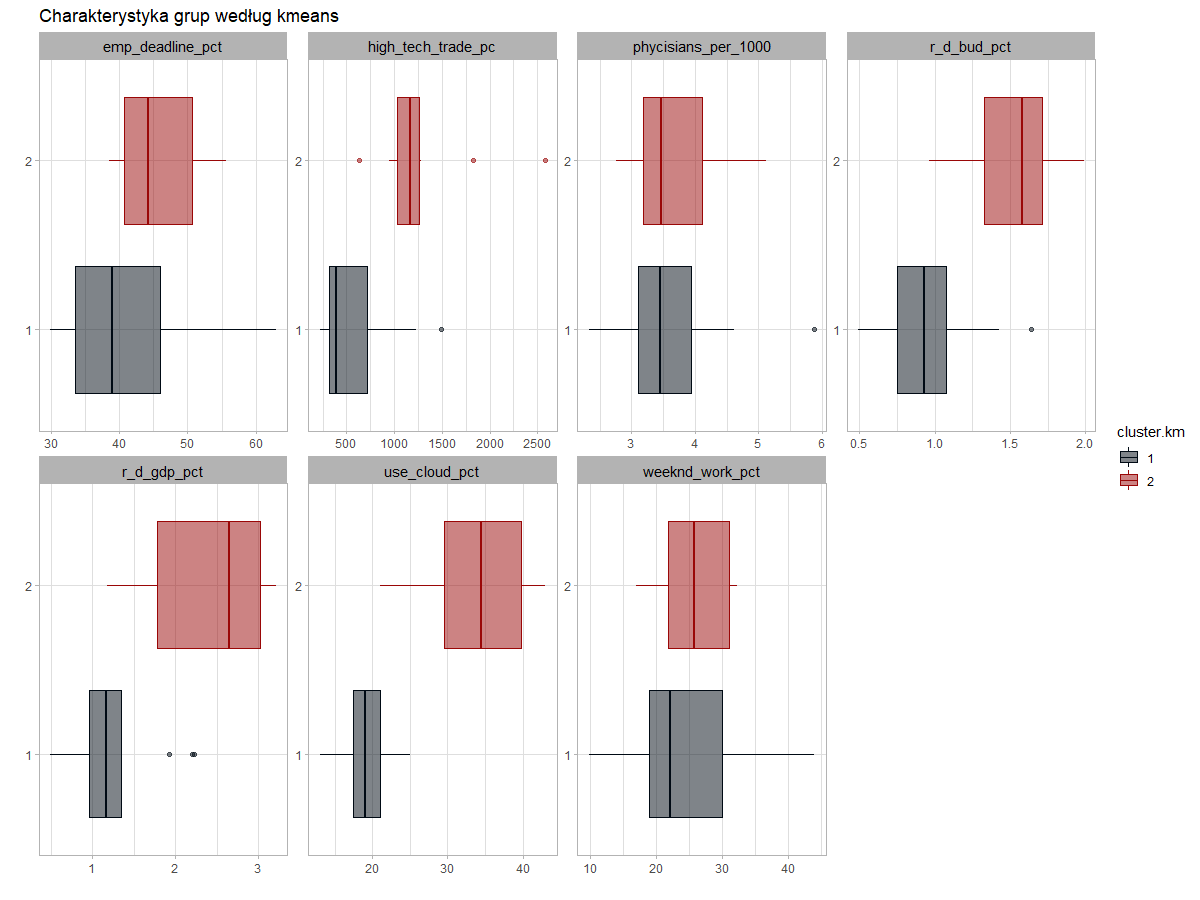
# Grupowanie k-means

Następnie przeprowadzono grupowanie metodą k-means (k-średnich) również przy pomocy języka R. Chcąc ustalić liczbę grupowań, jaką należało wykonać, kierowano się wykresem osypiska, a także kryteriami Calińskiego-Harabasza oraz Average Silhouette.

# 

Rysunek 3. Wykres osypiska, kryteria Average Silhouette oraz Calińskiego-Harabasza – metoda k-means

Wykres osypiska wskazuje na dwie lub trzy grupy, natomiast pozostałe kryteria sugerują grupowanie na dwa. Przeprowadzono zatem analizę skupień na dwóch grupach metodą k-means, a następnie skonfrontowano wyniki z metodą grupowania hierarchicznego.

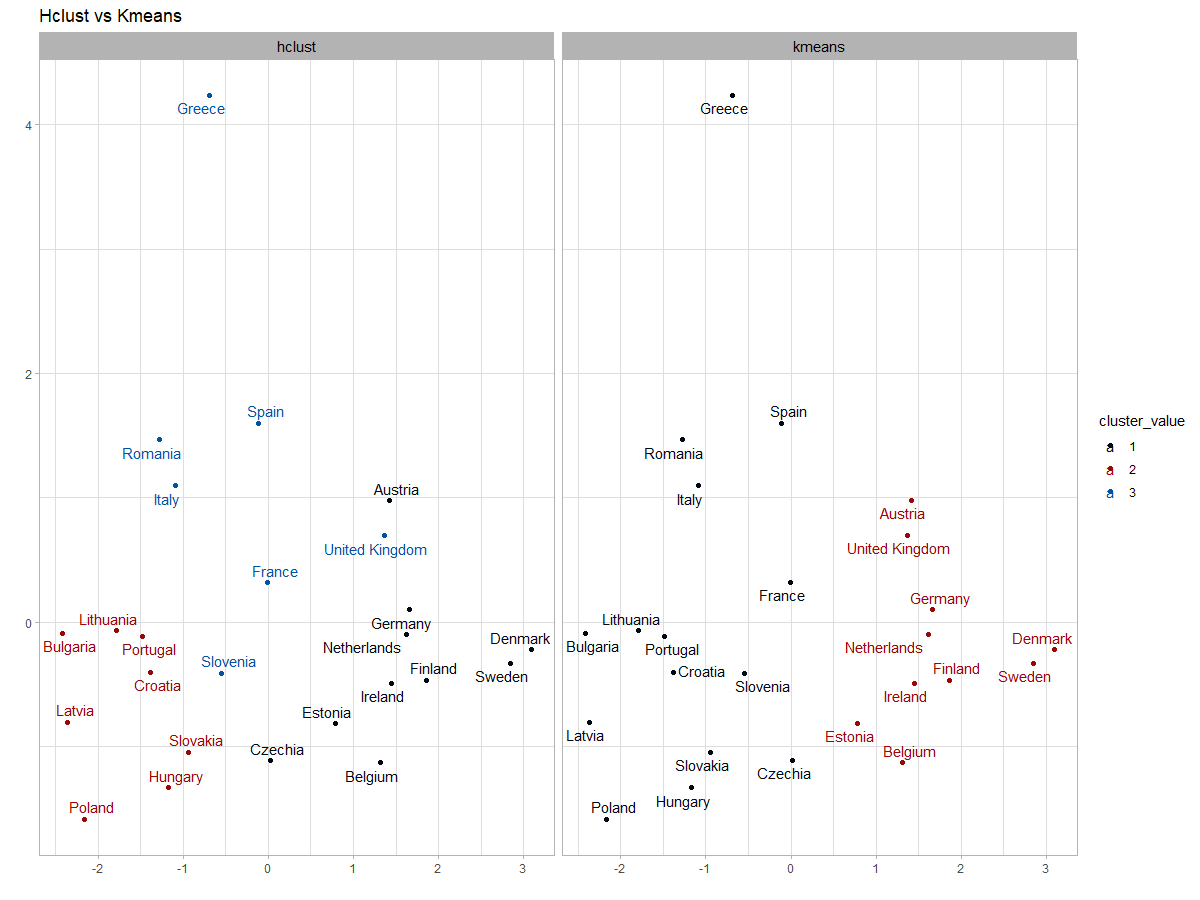


Rysunek 4. Charakterystyka grup - metoda k-means

Charakterystyka grup powstałych w wyniku grupowania k-means jest bardzo prosta. Członkowie grupy drugiej charakteryzują się przeciętnie wyższymi wartościami w każdej kategorii. Można zatem domniemać, że druga grupa to szeroko pojęta Europa Zachodnia, a grupa pierwsza to były blok komunistyczny.

# Porównanie grupowań

Ogólny podział krajów na grupy według metody k-średnich i grupowania hierarchicznego wygląda następująco.



Rysunek 5. Podział krajów - metoda grupowania hierarchicznego i k-średnich

Zgodnie z wcześniejszymi przypuszczeniami do grupy pierwszej w przypadku grupowania k-means w znacznej większości należą państwa postsowieckie. Dodatkowo w tej grupie znajdują się państwa śródziemnomorskie takie jak Grecja, Hiszpania czy Włochy. Graniczną wartością, decydującą o przynależności poszczególnego kraju wydaje się być zero na osi X (wartość średnia zestandaryzowanego wymiaru pierwszego w analizie PCA).

Grupowanie hierarchiczne z kolei bardziej odpowiada poszczególnym kwadrantom analizy PCA, co dokładniej opisano w części poświęconej grupowaniu hierarchicznemu. Ponadto podział ten zdaje się lepiej odzwierciedlać geograficzne położenie państw względem siebie.

# Odpowiedź na pytanie badawcze

Wykonując analizę skupień metodą grupowania hierarchicznego otrzymano trzy grupy, natomiast przy pomocy metody k-średnich dwie. Pierwsza metoda pozwoliła na uzyskanie grup w dużej mierze zbieżnych z kwadrantami wcześniej wykonanej analizy PCA, druga z kolei dzieli kraje starego kontynentu na państwa bogatego zachodu i północy oraz biedniejszego wschodu i południa.