Projekt z przedmiotu:

Modelowanie statystyczne i data mining w R

Michał Górnik, Michał Muzykant

24 kwietnia 2022 r.

Spis treści

Wprowadzenie

Celem naszego projektu jest analiza danych dotyczących 27 krajów Unii Europejskiej przy pomocy analizy składowych głównych i algorytmu grupowania w programie R Studio.

Opis Danych

Zmienne, które posłużyły nam do stworzenia projektu:

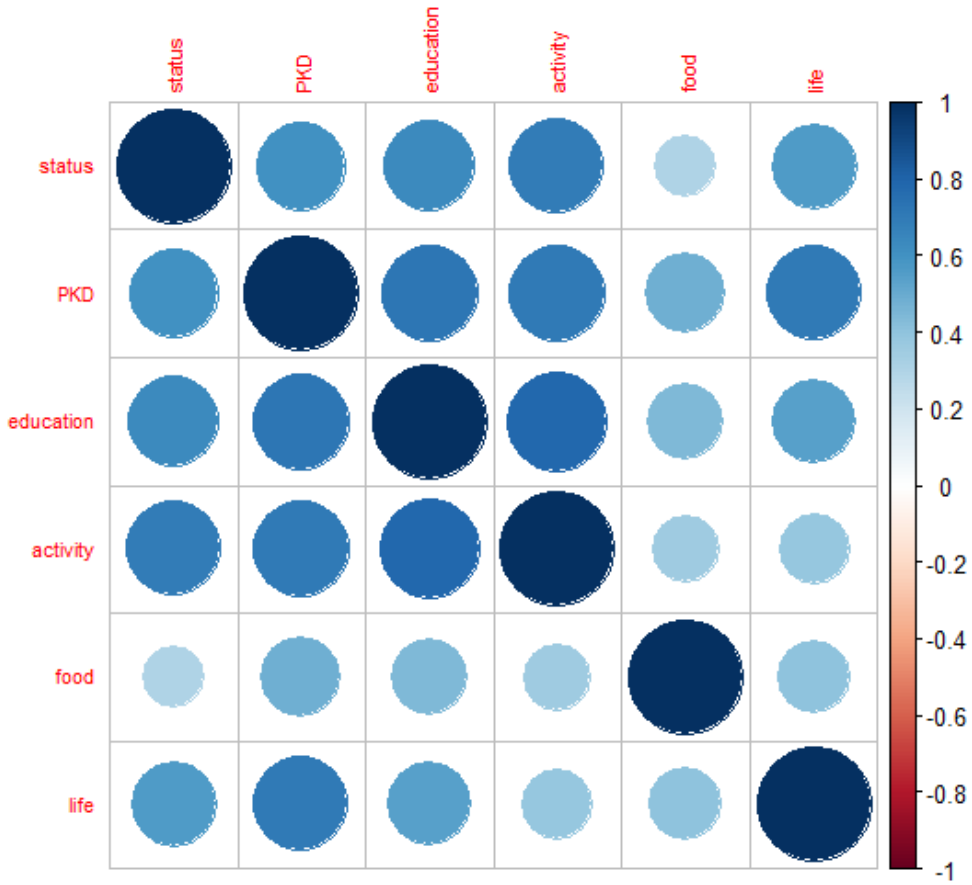
- Dzienne spożycie owoców i warzyw – zmienna „jedzenie” https://ec.europa.eu/eurostat/databrowser/view/HLTH_EHIS_FV3E__custom_2544306/default/table?lang=en&fbclid=IwAR1A0c06k3BWzDo0k_f1ojqFhm2Ip-7JaPgGHzNKc3WQa873TRtspm4RFNk
- Procentowy udział w tworzeniu PKB w stosunku do średniej krajów Unii Europejskiej – zmienna „PKD” https://ec.europa.eu/eurostat/databrowser/view/NAMA_10_PC__custom_2565194/default/table?lang=en
- Oczekiwana długość życia – zmienna „zycie” https://ec.europa.eu/eurostat/databrowser/view/DEMO_MLEXPEC__custom_2544020/default/table?lang=en&fbclid=IwAR33GHqA1EXUdd0z168fqLsFeFK7UyUq6-CrHQdR5udsMeCvDF3ODKqD1Zk
- Procentowy udział ludności danego kraju w nauce lub kształceniu się w ostatnich 4 tygodniach – zmienna „education” https://ec.europa.eu/eurostat/databrowser/view/TRNG_LFS_02__custom_2565159/default/table?lang=en
- Odsetek osób, które nie uprawiają sportu – zmienna „activity” https://ec.europa.eu/eurostat/databrowser/view/HLTH_EHIS_PE2E__custom_2539561/default/table?lang=en&fbclid=IwAR0Se5u3jmfJQtVeLVU96X9nVeusIAC38moFLsA0BFP3v8k432O8SvV-HO8
- Odsetek osób w danym kraju, które nie są w stanie zaspokoić swoich potrzeb materialnych – zmienna „status” https://ec.europa.eu/eurostat/databrowser/view/ILC_MDDD11__custom_2565067/default/table?lang=en&fbclid=IwAR1kWNA4ugdW_XDKXPXCuZUkJl81C_zaaHv0-ZrC4SnK_FBI9WlS_9aQM5o

Skróty, które zostały zastosowane w projekcie:

(AT) Austria	**(EE)** Estonia	**(LV)** Łotwa	**(RO)** Rumunia
(BE) Belgia	**(FI)** Finlandia	**(LU)** Luksemburg	**(SK)** Słowacja
(BG) Bułgaria	**(FR)** Francja	**(MT)** Malta	**(SI)** Słowenia
(HR) Chorwacja	**(EL)** Grecja	**(NL)** Holandia	**(SE)** Szwecja
(CY) Cypr	**(ES)** Hiszpania	**(DE)** Niemcy	**(HU)** Węgry
(CZ) Czechy	**(IE)** Irlandia	**(PL)** Polska	**(IT)** Włochy
(DK) Dania	**(LT)** Litwa	**(PT)** Portugalia	

Zależności między zmiennymi

Korelacje zmiennych:



Rysunek 1 Wykres korelacji zmiennych

Zmienne status, PKD, education i activity są ze sobą mocno skorelowane dodatnio. Natomiast zmienna „food" jest słabo skorelowana z resztą zmiennych.

Test sferyczności Barletta

```
$chisq
[1] 87.769

$p.value
[1] 2.581145e-12

$df
[1] 15
```

P-value jest bliskie zero, więc można odrzucić hipotezę, że macierz korelacji jest macierzą jednostkową i można uznać, że pomiędzy zmiennymi występują pewne korelacje.

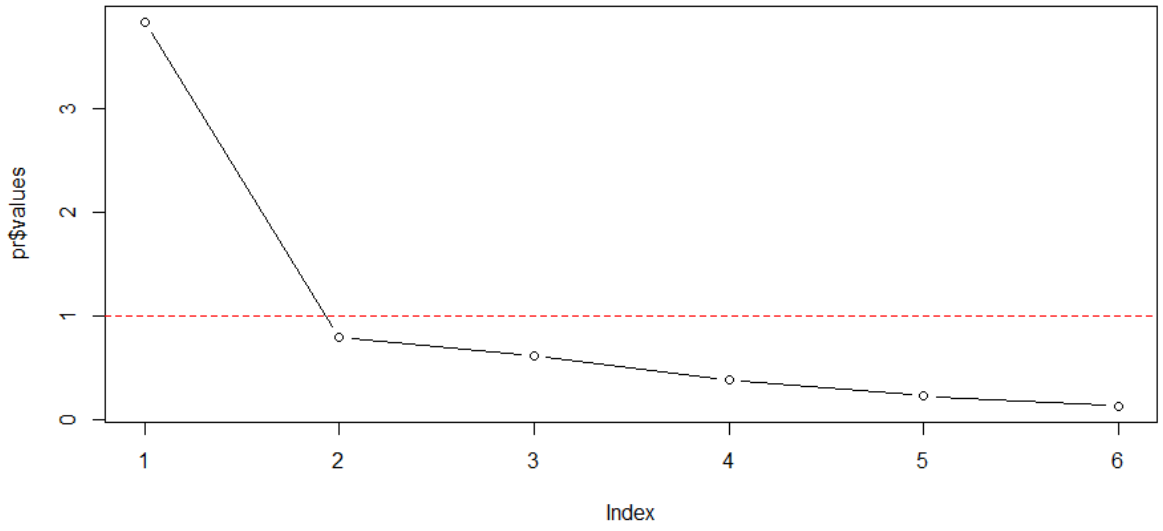
Kryterium Kaisera-Meyera-Olkina

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = cor(BZ[, 3:8]))
Overall MSA =  0.78
MSA for each item =
     food       PKD      life education  activity    status
     0.92      0.79      0.68      0.87      0.70      0.81
```

Kryterium KMO jest równe 0,78 co oznacza, że jest większe od 0,5. Analiza PCA jest dopuszczalna.

Analiza PCA

Wybór składowych



Rysunek 2 Wykres Osypiska – analiza PCA

Wykres osypiska wskazuje na wybranie dwóch składowych.

```
Loadings:
          PC1    PC2    PC3    PC4    PC5    PC6
food       0.589  0.701  0.361  0.178
PKD        0.893        -0.100 -0.283 -0.274 -0.179
life       0.750  0.258 -0.581                0.159
education  0.878 -0.147  0.177 -0.187  0.364
activity   0.839 -0.352  0.307        -0.135  0.237
status     0.806 -0.302 -0.126  0.482        -0.105

                 PC1   PC2   PC3   PC4   PC5   PC6
SS loadings    3.831 0.799 0.619 0.383 0.235 0.133
Proportion Var 0.639 0.133 0.103 0.064 0.039 0.022
Cumulative Var 0.639 0.772 0.875 0.939 0.978 1.000
```

Rysunek 3 pr$loadings (nfactors = 6)

Pierwsza składowa wyjaśnia zmienność w 64% a druga w 13%. Łącznie wyjaśniają zmienność w około 77%.

Interpretacja składowych

```
Loadings:
           PC1    PC2
food        0.589  0.701
PKD         0.893
life        0.750  0.258
education   0.878 -0.147
activity    0.839 -0.352
status      0.806 -0.302

                 PC1   PC2
SS loadings    3.831 0.799
Proportion Var 0.639 0.133
Cumulative Var 0.639 0.772
```

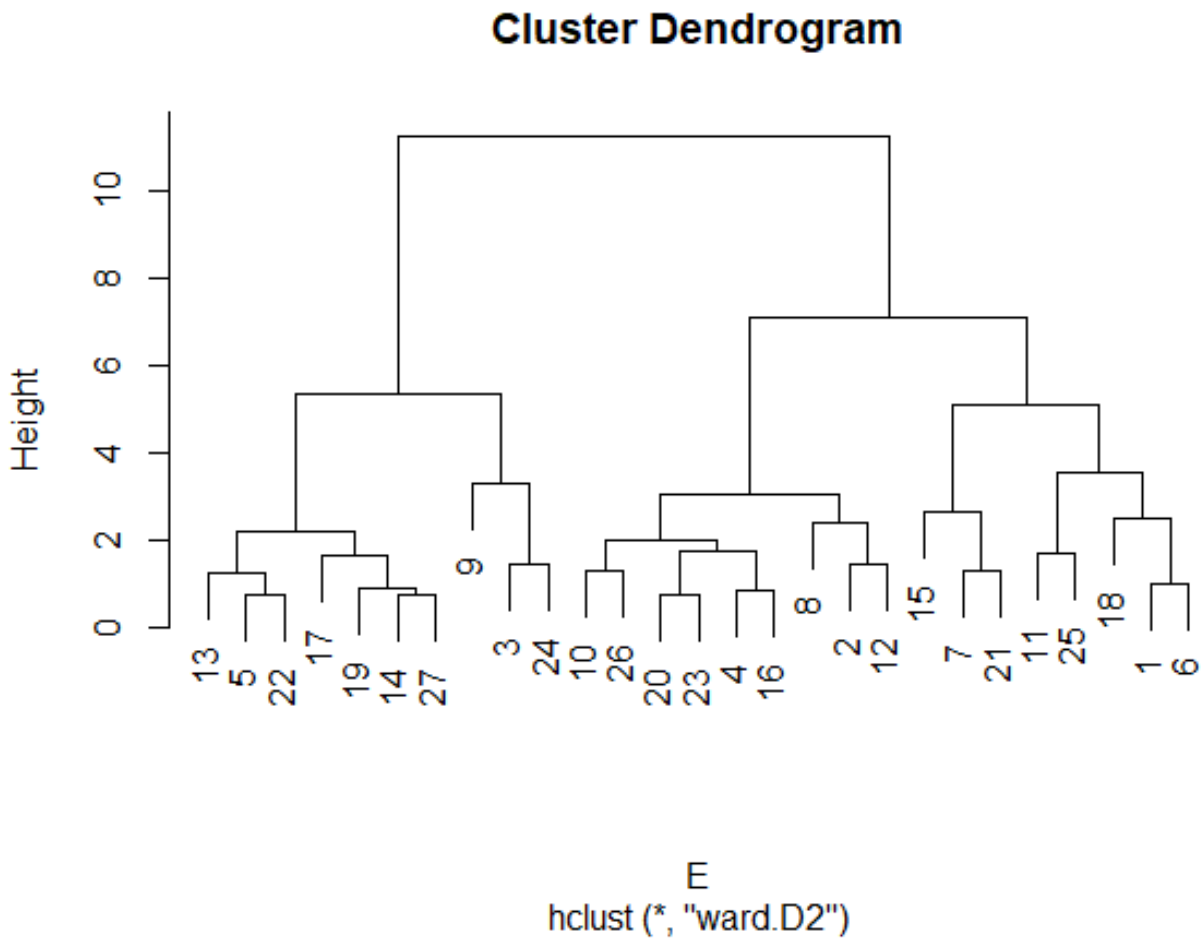
Rysunek 4 pr$loadings (nfactors = 2)

- Pierwsza składowa (PC1) – jest silnie skorelowana ze zmienną PKD, life, education, activity, status. Najmocniej odpowiada zmiennej PKD, czyli „Procentowy udział w tworzeniu PKB w stosunku do średniej krajów Unii Europejskiej”. Tą składową można nazwać jakość życia. Można wywnioskować, że w do tej grupy będą należeć kraje najlepiej rozwinięte na tle Unii Europejskiej.
- Druga składowa (PC2) – jest najsilniej skorelowana ze zmienną food. Z pozostałymi zmiennymi jest słabo skorelowana. Można powiedzieć, że jest przeciwieństwem pierwszej składowej. Tą składową można nazwać jakość posiłków. Do tej grupy będą należeć kraje najsłabiej rozwinięte w Unii Europejskiej.

Analiza skupień

Grupowanie hierarchiczne

Metoda warda:

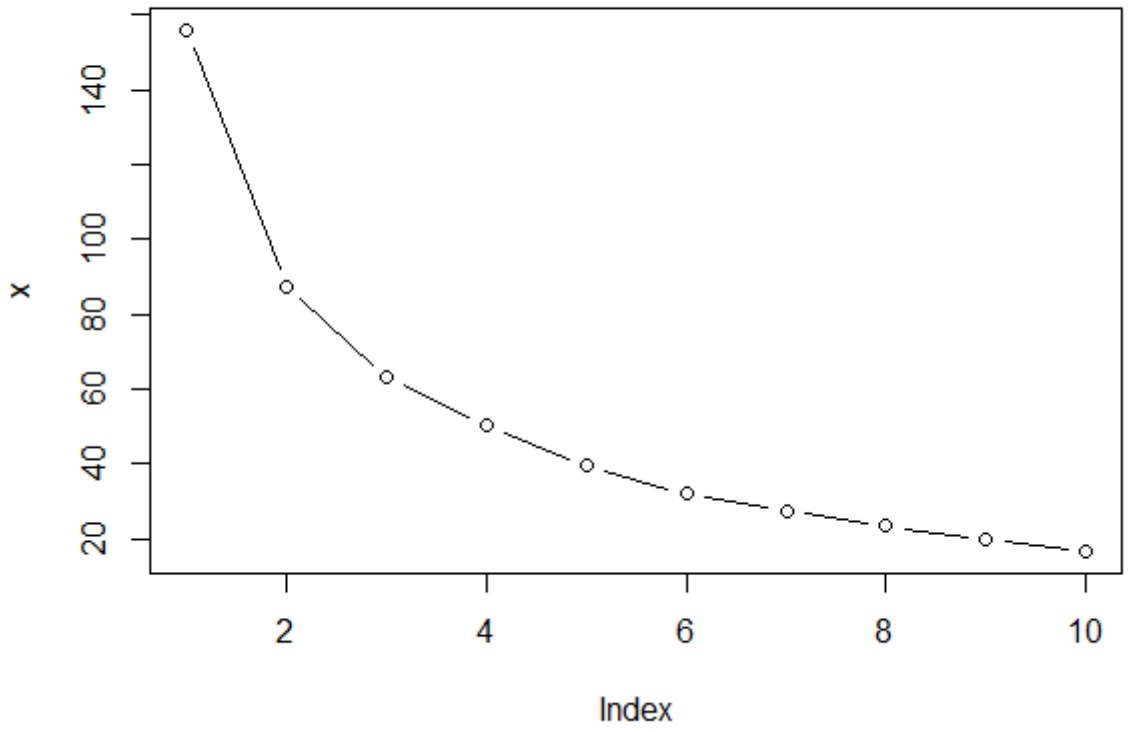


Rysunek 5 Dendrogram

Jak widać na załączonym dendrogramie, najrozsądniejszym rozwiązaniem jest podzielić obserwacje na dwie grupy.

Grupowanie k-średnich

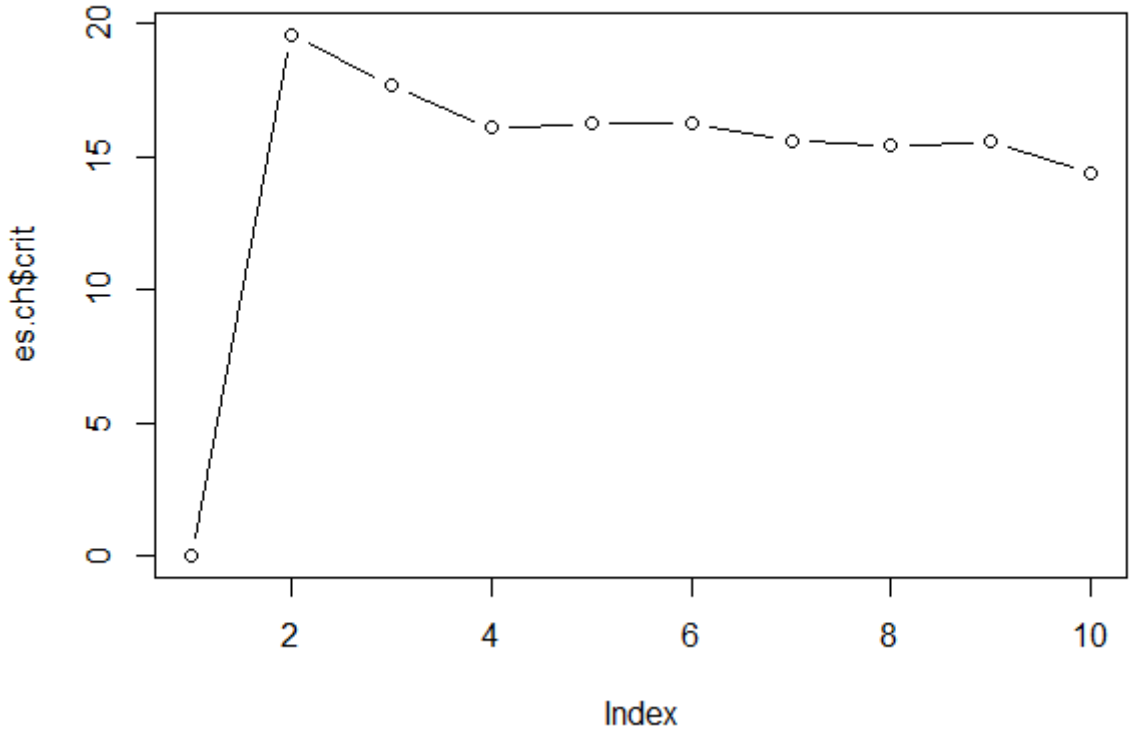
Wykres osypiska



Rysunek 6 Wykres osypiska – Metoda k-średnich

Z wykresu osypiska wnioskujemy, że należy podzielić obserwacje na dwie grupy.

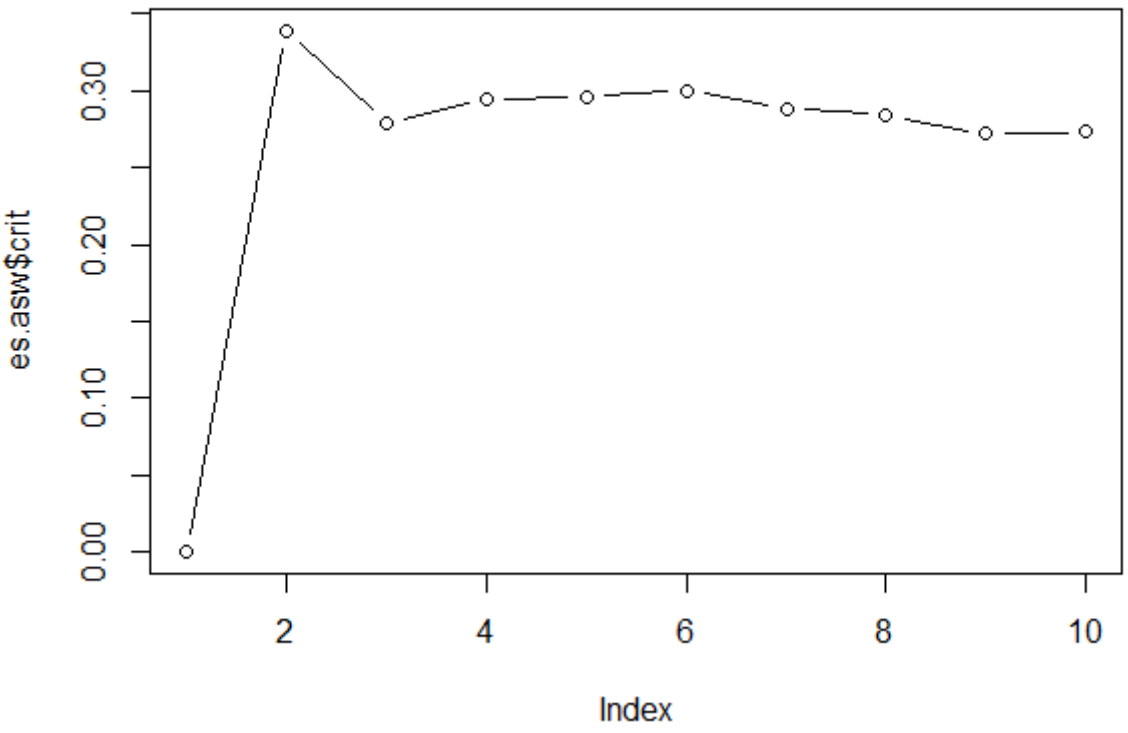
Kryterium Calińskiego-Harabasza



Rysunek 7 Wykres z kryterium Calińskiego-Harabasza

Według kryterium Calińskiego-Harabasza należy wybrać 2 grupy.

Kryterium średniej sylwetki



Rysunek 8 Wykres kryterium średniej sylwetki

Według kryterium średniej sylwetki należy wybrać podział na 2 grupy.

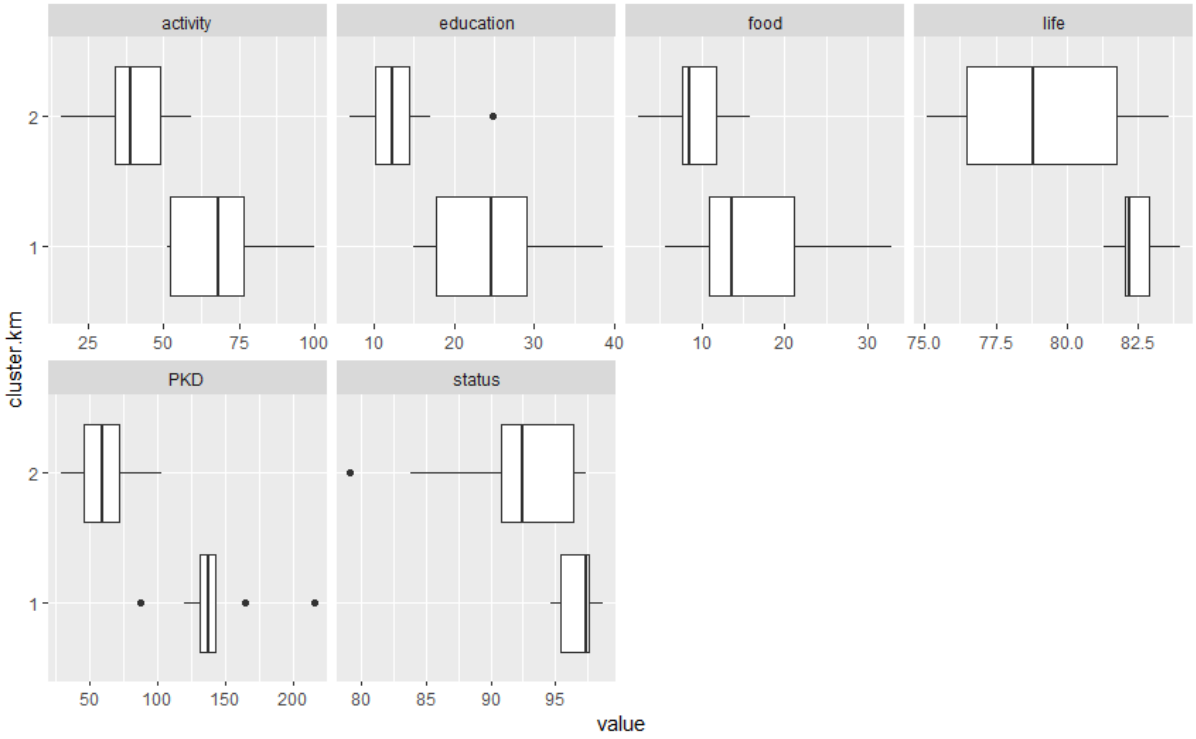
Wszystkie powyższe wykresy udowadniają racjonalność podziału na dwie grupy.

Porównanie grup

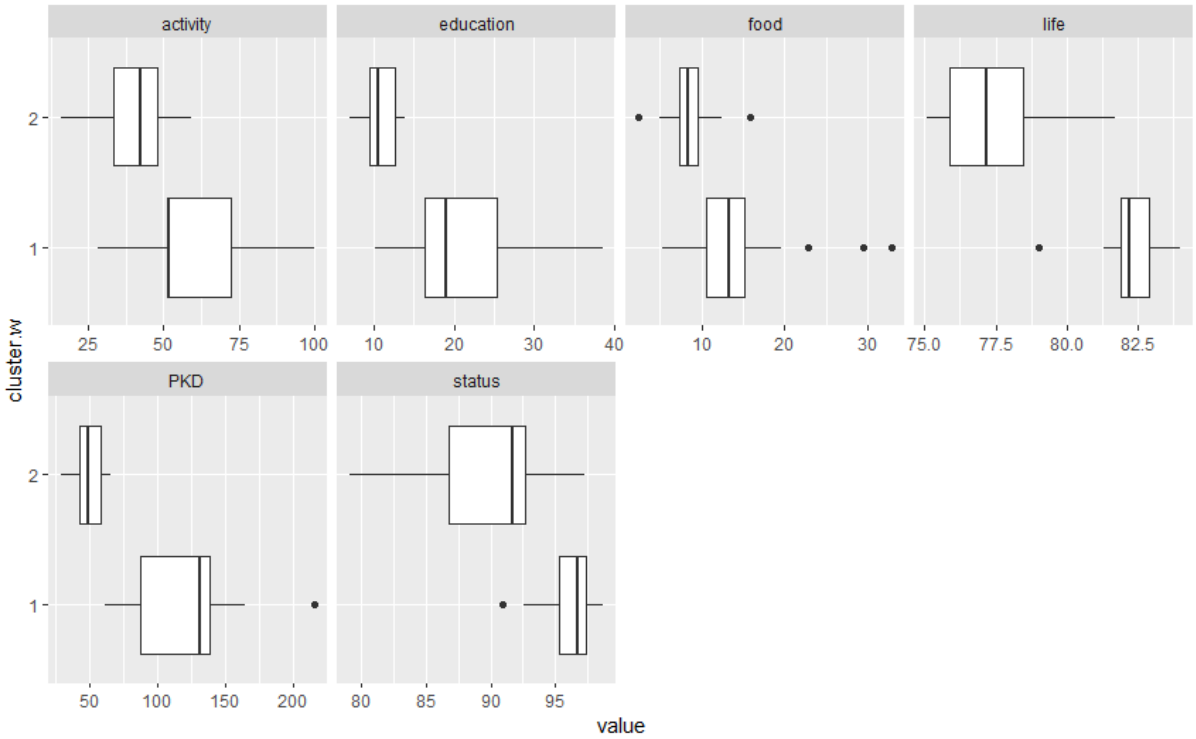
```
> table(es.ch$cluster, es.asw$cluster)

    1  2
  1 16  0
  2  0 11
> table(es.ch$cluster, ec1$cluster)

    1  2
  1  0 16
  2 11  0
>
```

Rysunek 9 Liczebność grup

według kryterium Calinskiego-Harabasza i k średnich podział liczby zmiennych na grupy jest taki sam



Rysunek 10 Opis zmiennych według k-means

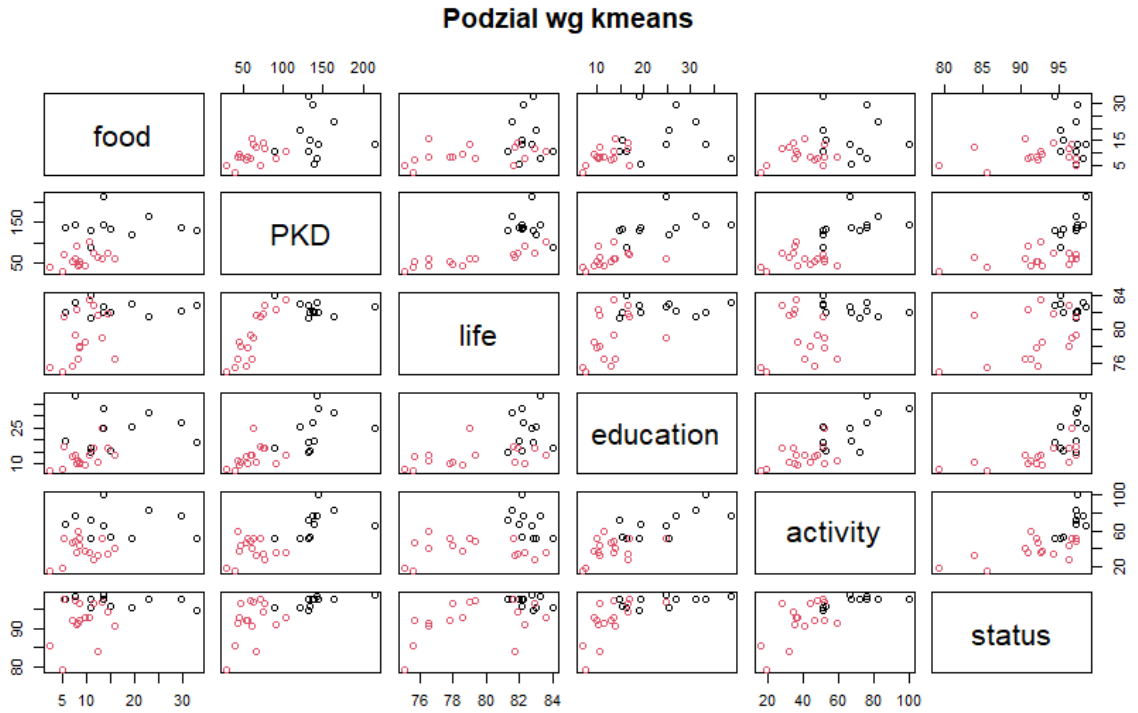


Rysunek 11 Opis zmiennych według hclust

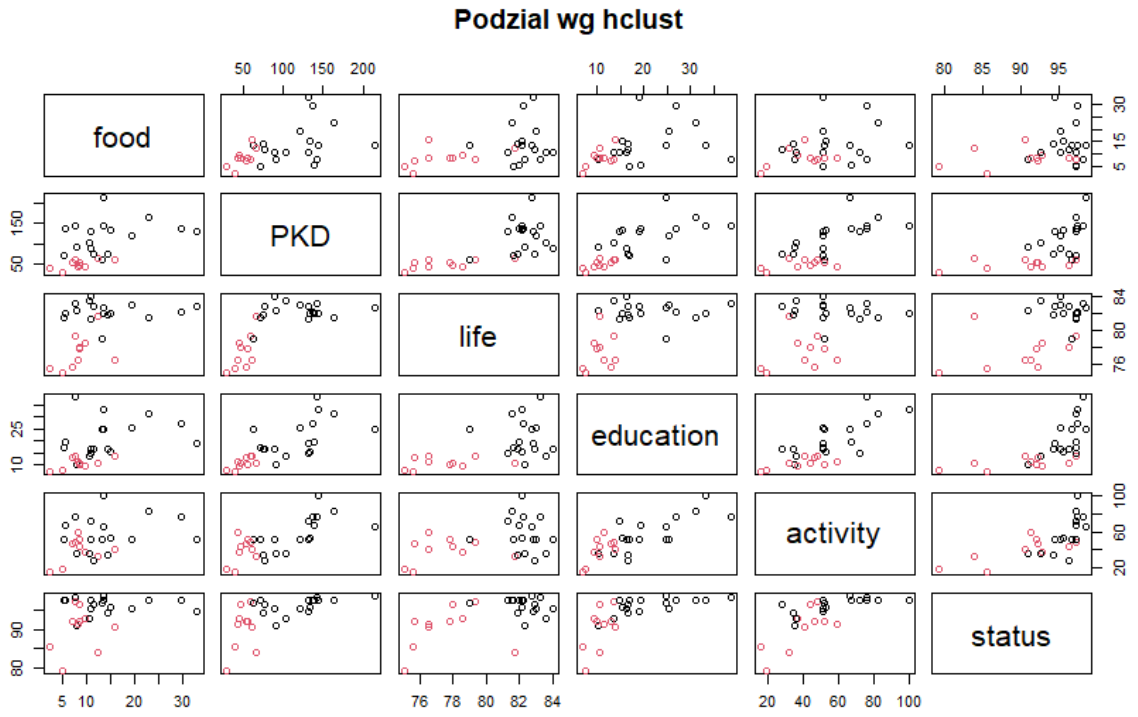
W obu metodach grupowania pierwsza grupa wypada wyraźnie lepiej we wszystkich zmiennych.

```
> table(BZ$cluster.w, cluster.km =BZ$cluster.km )
   cluster.km
     1  2
  1 11  6
  2  0 10
> table(BZ$cluster.w)

 1  2
17 10
> table(BZ$cluster.km)

 1  2
11 16
>
```

Rysunek 12 Liczebność grup według obu metod



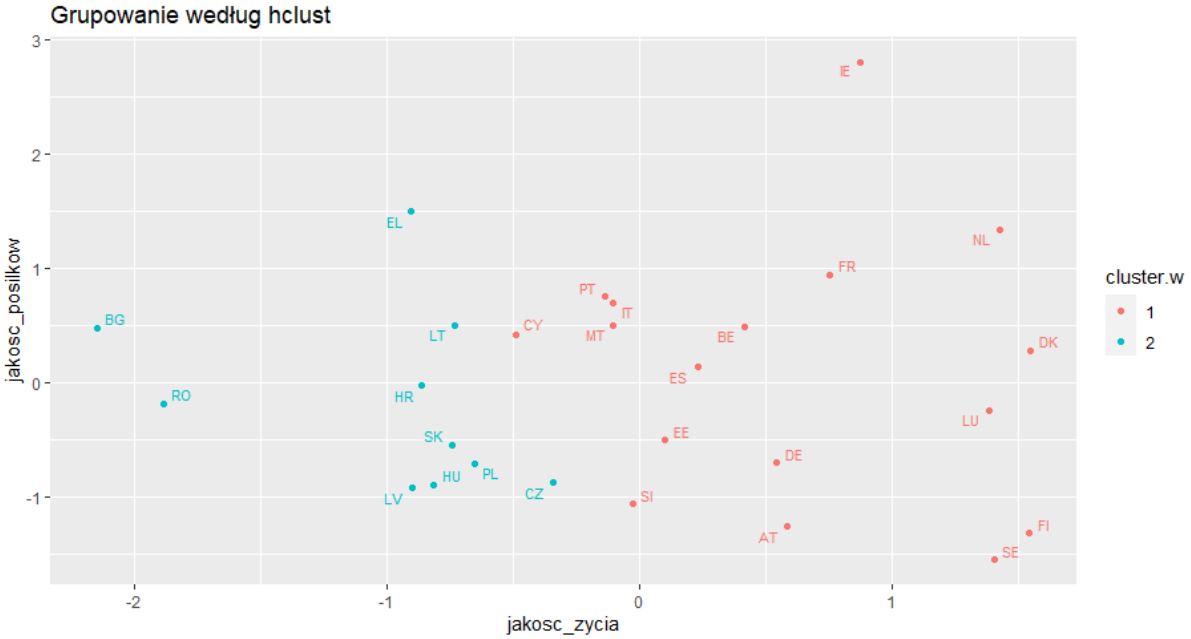
Rysunek 13 Podział grupowania k-means według zmiennych



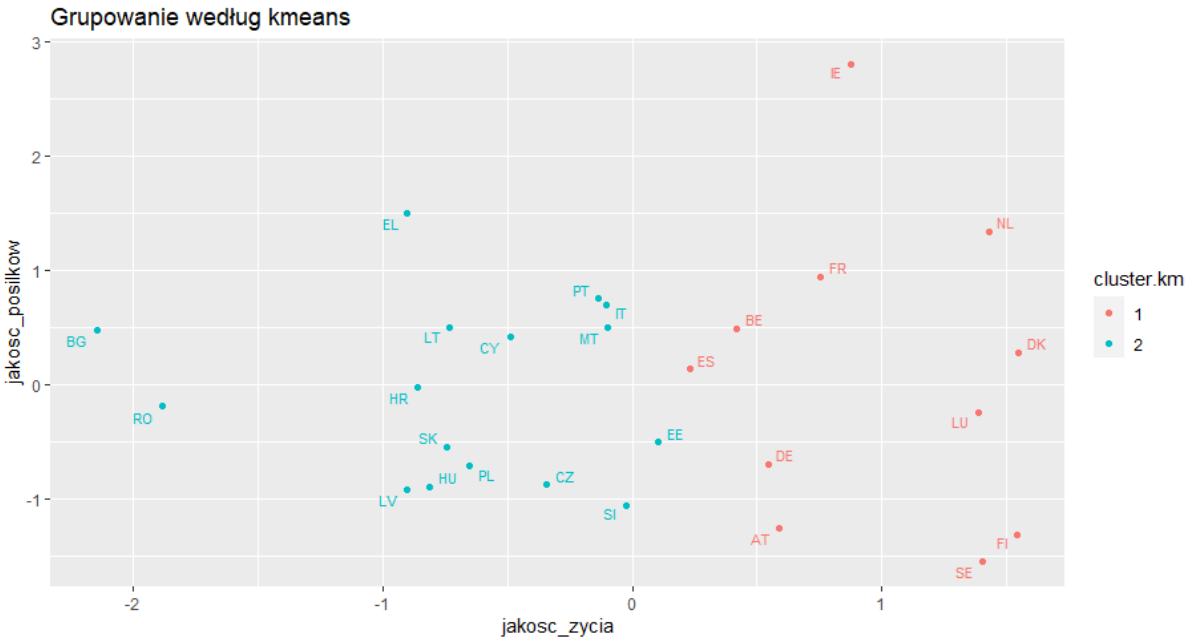
Rysunek 14 Podział grupowania hclust według zmiennych

Różnice w grupowaniach najbardziej widać przy zmiennych activity i life (jeśli kraje wypadają w tych zmiennych słabo to według metody hierarchicznej trafiają do gorszej grupy, natomiast przy metodzie k-średnich kraje gorzej radzące sobie w tych parametrach mogą trafić do pierwszej jak i drugiej grupy.

Podsumowanie



Rysunek 15 Grupowanie według hclust



Rysunek 16 Grupowanie według kmeans

- Na tle krajów Unii Europejskiej wyróżnia się Irlandia, której obywatele jedzą dużo owoców i warzyw oraz są krajem stosunkowo wysoko rozwiniętym.
- Szwecja i Finlandia cechuje się wysoką jakością życia a jej mieszkańcy jedzą najmniej warzyw i owoców w całej Unii Europejskiej.
- Najsłabszymi przedstawicielami Unii Europejskiej są zdecydowanie Bułgaria oraz Rumunia, gdyż ich średnie roczne PKB jest znacznie mniejsze od reszty krajów Unii Europejskiej.
- Polska na tle reszty krajów UE wypada bardzo przeciętnie pod względem obu zmiennych.

Spis rysunków: