Laboratorium - Hierarchiczna analiza danych (HCA) – Techniki eksploracji danych wielowymiarowych

Maja Jarmoc, grupa 1, nr indeksu: 274263

Cel ćwiczenia: Analiza podobieństwa dla cech przy użyciu odległości dopełniającej.

Wykonanie analizy HCA (hierarchiczna analiza skupień) dla obiektów przy użyciu trzech rożnych metod:

- Metody Warda
- Single linkage
- Complete linkage

i trzech różnych odległości:

- Euklidesowej
- Manhattan
- Czebyszewa

Do wykonania ćwiczenia użyłam danych z przygotowanego zestawu: 31 rodzajów płatków śniadaniowych wraz z informacjami o kaloriach, ilości białka, tłuszczu, błonnika, węglowodanów, cukrów (w gramach) i sodu (w miligramach); (porcja: 30g.)

Dane zostały ówcześnie przygotowane: usunięcie danych odstających, transformacja (o ile tego wymagały).

Niezbędnym krokiem przed analizą HCA jest standaryzacja (autoskalowanie) danych. Celem autoskalowania danych jest uczynienie poszczególnych zmiennych współmiernymi (uczynienie współmiernymi wszystkich wymiarów w wielowymiarowej przestrzeni cech).

Wzór użyty do standaryzacji danych:

$$Z_{Ax} = \frac{x_{Ax} - m_x}{s_x}$$

gdzie:
- z_{Ax} - standaryzowana wartość cechy **X** dla obiektu **A**;
- x_{Ax} - oryginalna[13] wartość cechy **X** dla obiektu **A**;
- m_x - wartość średnia zmiennej **X**;
- s_x - odchylenie standardowe populacji zmiennej **X**.

Po standaryzacji możemy przejść do obliczenia odległości pomiędzy obiektami. Zaczęłam od odległości euklidesowej określonej wzorem:

$$D(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2)}$$

x, y – obiekty (tu: płatki)

i – cecha (tu: cecha płatków)

Wynikiem obliczania odległości pomiędzy obiektami (niezależnie jaką metodą) jest kwadratowa, symetryczna macierz odległości z zerami na diagonali.

	Bran Chex	Bran Flakes	Cap'n'Crunch	Cinnamon Toast Crunch	Clusters	Cocoa Puffs	Corn Chex	Corn Flakes
Bran Chex	0.00	1.86	5.06	5.11	3.33	4.24	4.36	3.85
Bran Flakes	1.86	0.00	6.00	6.24	3.82	5.16	4.96	4.37
Cap'n'Crunch	5.06	6.00	0.00	1.37	3.41	1.71	4.67	5.43
Cinnamon Toast Crunch	5.11	6.24	1.37	0.00	3.29	2.66	4.73	5.48
Clusters	3.33	3.82	3.41	3.29	0.00	3.33	4.42	4.87
Cocoa Puffs	4.24	5.16	1.71	2.66	3.33	0.00	4.52	5.10
Corn Chex	4.36	4.96	4.67	4.73	4.42	4.52	0.00	1.55
Corn Flakes	3.85	4.37	5.43	5.48	4.87	5.10	1.55	0.00

Tabela 1. Przykładowa macierz odległości policzona dla pierwszych 8 obiektów.

Po obliczeniu odległości pomiędzy obiektami możemy przejść do analizy HCA. Zaczęłam od metody Warda.

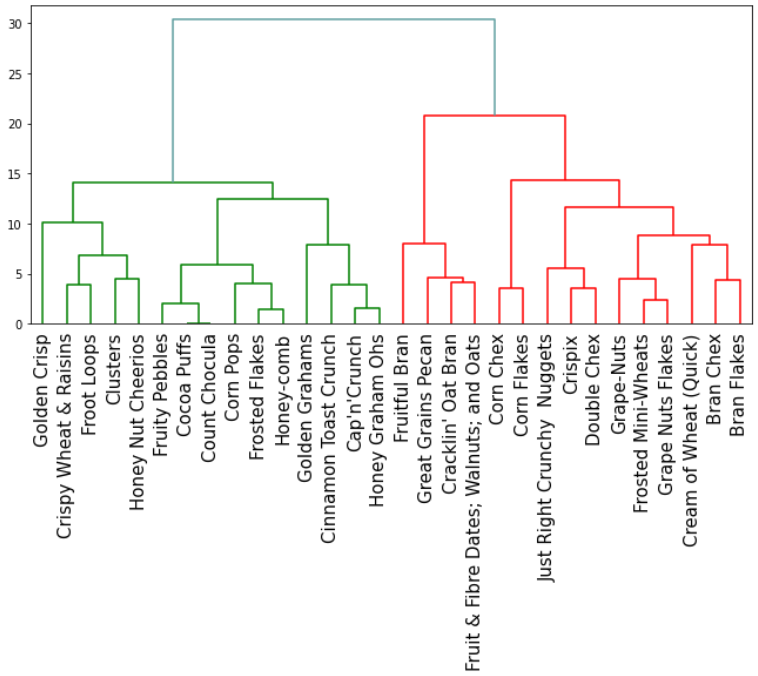
1) Odległość pomiędzy obiektami: euklidesowa

HCA – metoda Warda

Do oszacowania odległości między skupieniami wykorzystuje podejście analizy wariancji. Metoda ta zmierza do minimalizacji sumy kwadratów odchyleń dowolnych dwóch skupień, które mogą zostać uformowane na każdym etapie

W wyniku analizy HCA metodą Warda uzyskaliśmy następujący dendrogram:

Rysunek 1. Analiza HCA metodą Warda (Odległość: euklidesowa)

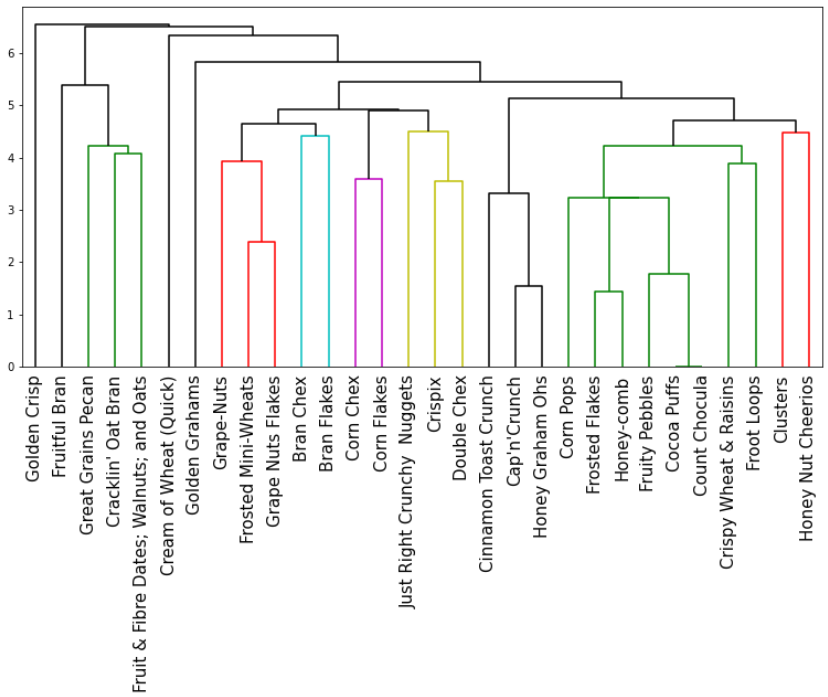


HCA – metoda single linkage

W metodzie tej odległość między dwoma skupieniami jest określona przez odległość między dwoma najbliższymi obiektami (najbliższymi sąsiadami) należącymi do różnych skupień. Zgodnie z tą zasadą obiekty formują skupienia łącząc się w ciągi, a wynikowe skupienia tworzą długie "łańcuchy".

W wyniku analizy HCA metodą Single linkage uzyskaliśmy następujący dendrogram:

Rysunek 2. Analiza HCA metodą Single linkage (Odległość: euklidesowa)

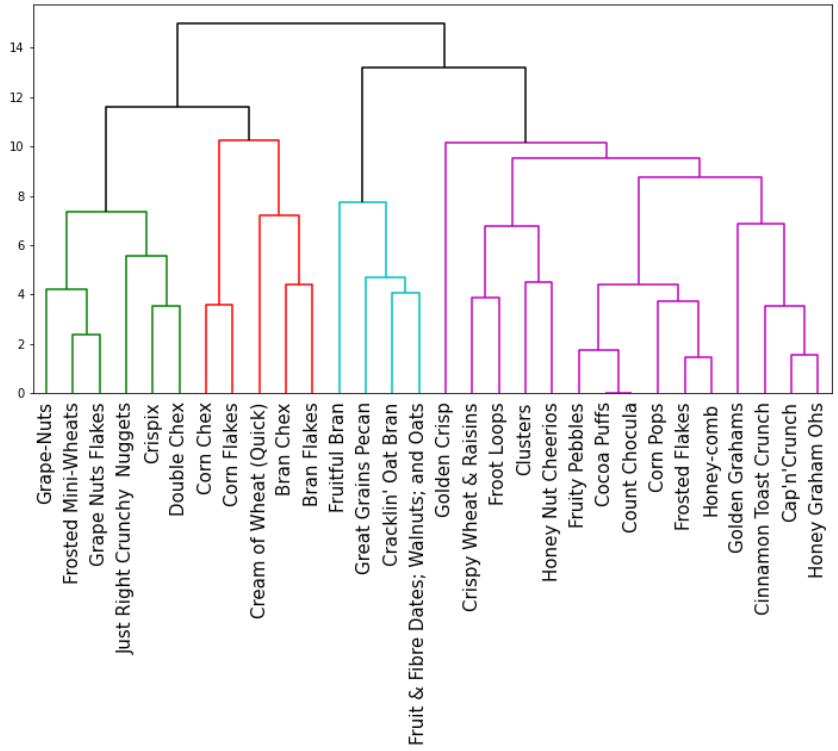


HCA – metoda complete linkage

W tej metodzie odległość między skupieniami jest zdeterminowana przez największą z odległości między dwoma dowolnymi obiektami należącymi do różnych skupień (tzn. "najdalszymi sąsiadami").

W wyniku analizy HCA Complete linkage uzyskaliśmy następujący dendrogram:

Rysunek 3. Analiza HCA metodą Complete linkage (Odległość: euklidesowa)



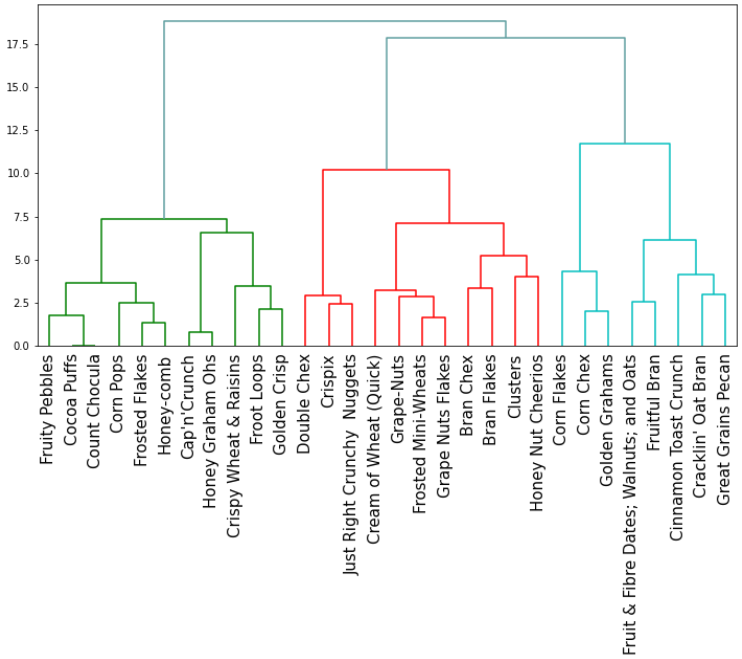
2) Odległość pomiędzy obiektami: Czebyszewa

$$D_{\text{Chebyshev}}(x, y) := \max_i(|x_i - y_i|).$$

Odległość między dwoma obiektami jest największą z ich różnic wartości poszczególnych cech.

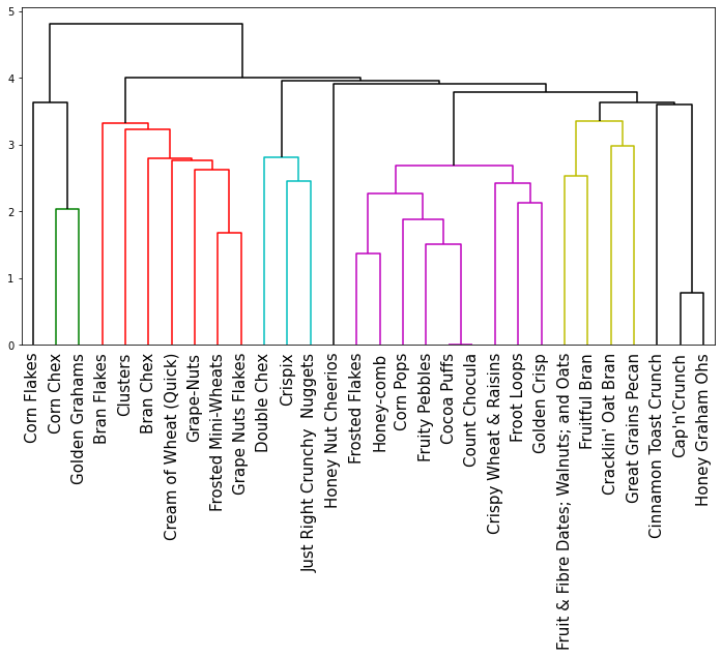
HCA – metoda Warda

Rysunek 4. Analiza HCA metodą Warda (Odległość: Czebyszewa)



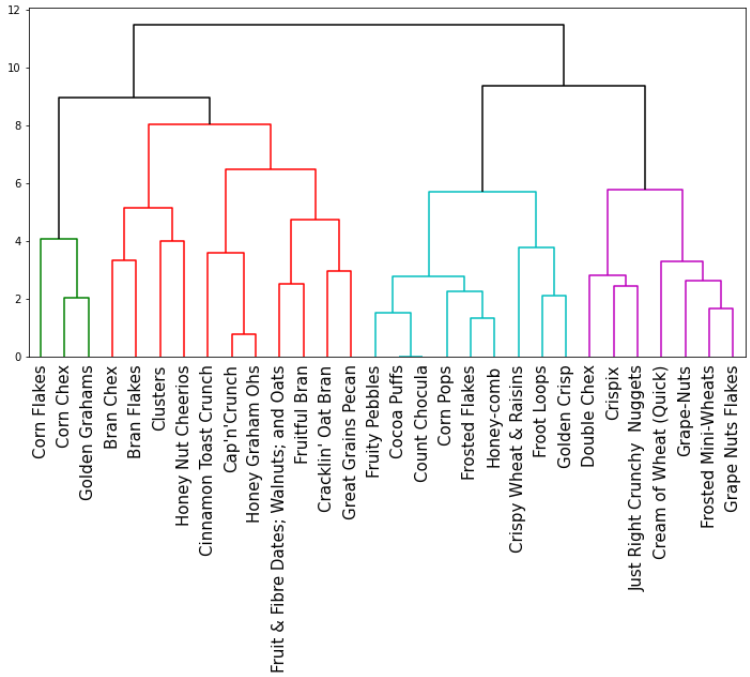
HCA – metoda single linkage

Rysunek 5. Analiza HCA metodą Single linkage (Odległość: Czebyszewa)



HCA – metoda complete linkage

Rysunek 6. Analiza HCA metodą Complete linkage (Odległość: Czebyszewa)



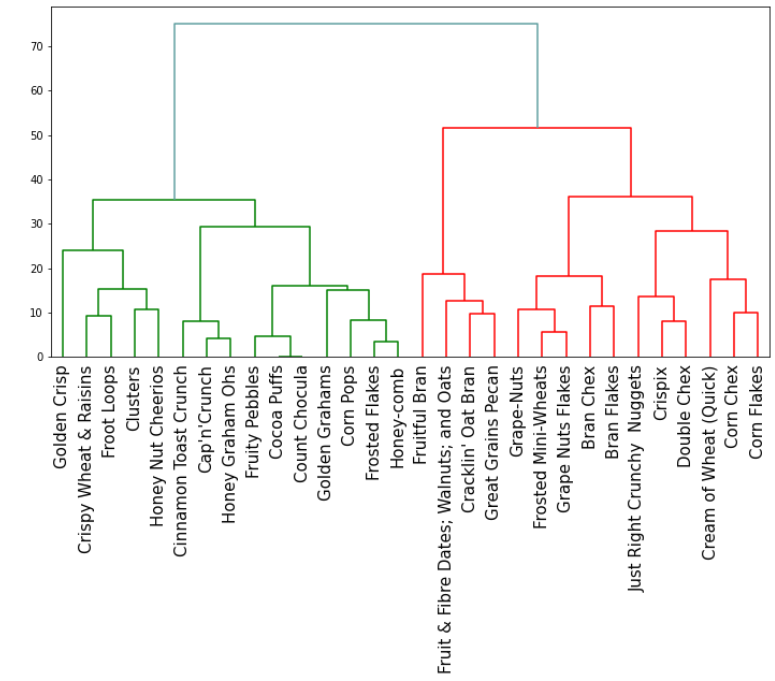
1) Odległość między obiektami: Manhattan

$$Manhattan\ Distance = \sum_{i=1}^{n} |p_i - q_i|$$

Odległość dwóch punktów to suma wartości bezwzględnych różnic ich współrzędnych.

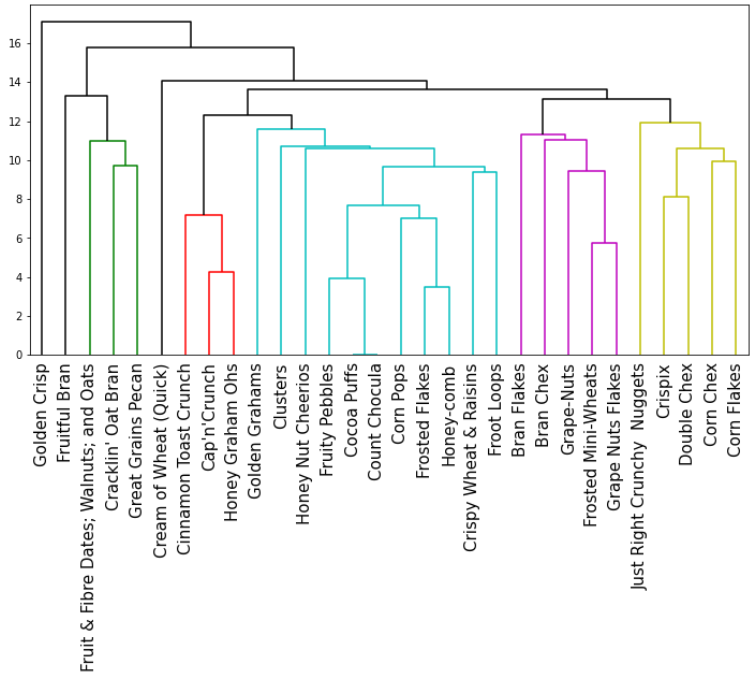
HCA – metoda Warda

Rysunek 7. Analiza HCA metodą Warda (Odległość: Manhattan)



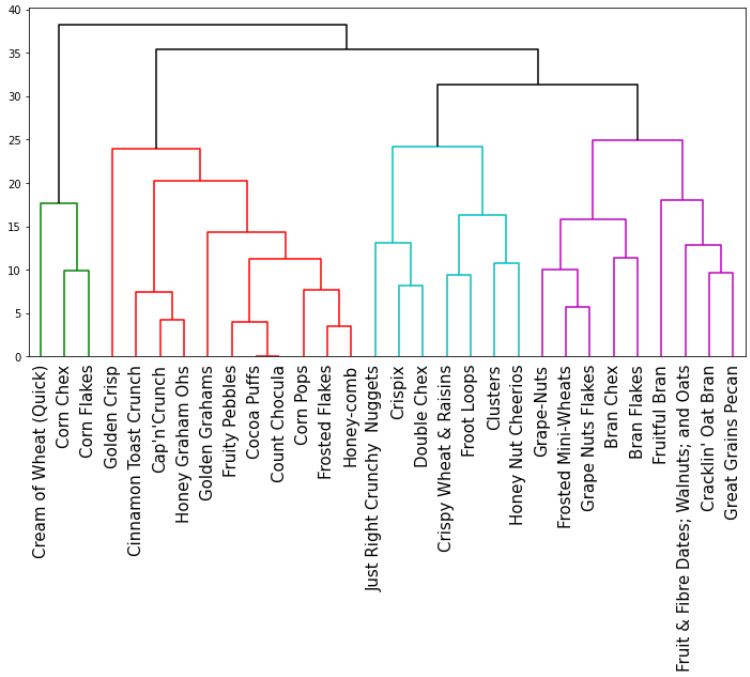
HCA – metoda single linkage

Rysunek 8. Analiza HCA metodą Single linkage (Odległość: Manhattan)



HCA – metoda complete linkage

Rysunek 9. Analiza HCA metodą Complete linkage (Odległość: Manhattan)



Podsumowanie analizy dla obiektów:

Najlepszą procedurą do analizy danych okazała się metoda Complete linkage. Niezależnie od użytej odległości, powstały dzięki niej dendrogram jest czytelny, a powstałe 4 duże skupiska wskazują jasno na zależności pomiędzy płatkami.

Skupisko nr. 1 (kolor zielony) – zawartość węglowodanów na poziomie 21 lub 22, brak tłuszczu. Bardzo niska zawartość cukrów.

Skupisko nr. 2 (kolor czerwony) – znacznie mniejsza zawartość węglowodanów: 11 do 15 g. Niska zawartość białka (~1g), niska zawartość błonnika (~0g)

Skupisko nr. 3 (kolor niebieski) – płatki zawierające podobną ilość węglowodanów co płatki czerwone, jednak procent zawartości cukrów prostych w węglowodanach znacznie niższy niż w płatkach czerwonych.

Skupisko nr. 4 (kolor fioletowy) – błonnik na znacznie wyższym poziomie niż reszta płatków (od 3 do 5 g), 3 gramy białka.

Etapy:

1. W macierzy odległości (n x n) poszukujemy wartości najmniejszej i sprawdzamy pomiędzy którymi zmiennymi ona występuje. (Nie bierzemy pod uwagę zer na diagonali, pamiętamy też, że macierz jest trójkątna)
2. Zmienne te łączymy w klaster, a następnie poszukujemy odległości stworzonego skupiska od reszty pozostałych zmiennych: w metodzie Complete linkage odległością **zmiennej** od naszego nowo utworzonego skupiska jest dłuższa wartość odległości pojedynczych zmiennych (tworzących teraz skupisko) od **zmiennej**.
3. Następnie ponownie tworzymy macierz odległości przypisując naszemu klastrowi etykietę i usuwając jedną ze zmiennych znajdujących się teraz w skupisku. Wielkość naszej macierzy to teraz n-1 x n-1. Powtarzamy procedurę, aż do uzyskania jednego dużego skupiska.

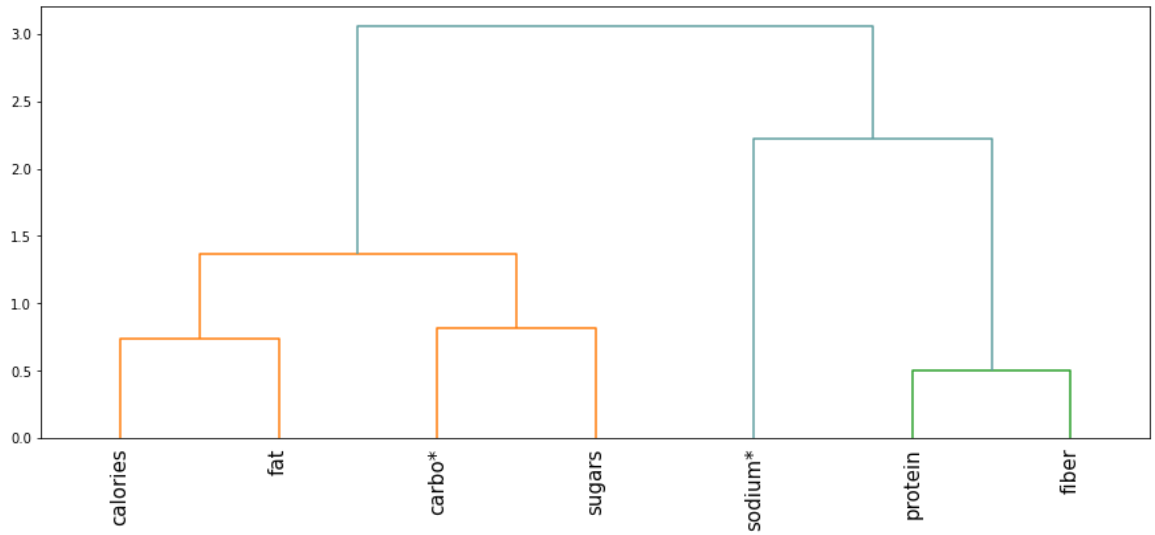
Analiza HCA dla cech obiektów.

Analizę HCA dla cech przeprowadziłam używając odległości dopełniającej:

$$|r - 1|$$

gdzie r – współczynnik korelacji Pearsona

$$r_{XY} = \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y}.$$



Rysunek 10. Analiza HCA metodą Warda bądź Complete linkage (wygląd dendrogramu identyczny w przypadku użycia dwóch metod) (Odległość dopełniająca)

Na dendrogramie widzimy, iż kalorie i tłuszcz tworzą jedno skupienie, ich współczynnik korelacji wynosi 0.51. Dzieje się tak, gdyż wraz ze wzrostem tłuszczu znacząco wzrasta ilość kalorii (bardziej niż przy wzroście węglowodanów i białka - w jednym gramie węglowodanów znajdują się 4 kcal, w jednym gramie białka również 4, zaś tłuszcz w 1 g posiada, aż 9 kcal.

Węglowodany i cukry to kolejne skupienie – w produktach cukry zawierają się w węglowodanach.

Białka i błonnik mogą tworzyć skupienie ze względu na fakt, iż płatki o wyższej zawartości białka uważane są za „zdrowsze”, a błonnik jest niejako wyznacznikiem dobrego, „zdrowego składu produktu. Lepsze płatki - więcej błonnika i białka.