Wiktoria Arendarczyk, Cezary Moskal

Projekt 2 - Grupowanie

Spis treści

Wiktoria Arendarczyk, Cezary Moskal

Wstęp

W niniejszym projekcie podjęto próbę segmentacji zbioru danych Cereals – zawierającego informacje o różnych płatkach śniadaniowych i ich wartościach odżywczych – za pomocą trzech różnych podejść do grupowania (clusteringu):

- klasteryzacja partycjonalna:
  - algorytmy k-means oraz k-medoid (traktowane wspólnie jako podejście partycjonalne);
- klasteryzacja hierarchiczna:
  - aglomeracyjne grupowanie metodą Warda;
- klasteryzacja gęstościowa:
  - algorytm DBSCAN;

Celem jest nie tylko wyodrębnienie naturalnych segmentów wśród płatków (np. o niskiej zawartości cukru, wysokiej zawartości błonnika, itp.), ale także porównanie skuteczności i stabilności poszczególnych metod. Posłużono się m.in. miarami wewnętrznej jakości klastrów (np. silhouette score), oraz wizualizacjami w przestrzeni zredukowanej (MDS – Multidimensional Scaling, Skalowanie Wielowymiarowe), aby przeanalizować spójność i heterogeniczność uzyskanych grup. W przypadku klasteryzacji partycjonalnej posłużono się metodą łokcia oraz wskaźnikiem sylwetkowym w celu określenia kandydatów na optymalną liczbę grup.

Wiktoria Arendarczyk, Cezary Moskal

Exploratory Data Analysis (EDA)

Opis danych

Zbiór Cereals gromadzi informacje o popularnych płatkach śniadaniowych dostępnych na rynku amerykańskim, łącząc w sobie dane o wartościach odżywczych (kalorie, białko, tłuszcz, cukry, błonnik itp.) z ocenami konsumenckimi. Dzięki tak szerokiemu spektrum zmiennych możliwe jest wydzielenie grup produktów o podobnych profilach – na przykład niskokalorycznych, wysokobiałkowych czy bogatych w błonnik. Ponadto zmienne takie jak typ płatków czy pozycja na półce dają dodatkowy kontekst marketingowy, pozwalając na analizę zarówno z perspektywy zdrowotnej, jak i komercyjnej.

Taki zestaw danych stanowi doskonałą bazę do porównania trzech podejść do segmentacji: partycjonalnych (K-Means i K-Medoid), hierarchicznych (metoda Warda i najdalszego sąsiada) oraz gęstościowych (DBSCAN).

Struktura danych

Zbiór Cereals składa się z 77 obserwacji (każda to inny produkt płatków śniadaniowych) i 16 zmiennych opisujących zarówno cechy jakościowe, jak i ilościowe. Wśród nich wyróżniamy:

zmienne ilościowe:

- *calories* – ilość kcal w porcji,
- *protein* – ilość białka w porcji (g),
- *fat* – ilość tłuszczu w porcji (g),
- *sodium* – ilość soli w porcji (mg),
- *fiber* – ilość błonnika w porcji (g),
- *carbo* – ilość złożonych węglowodanów w porcji (g),
- *sugars* – ilość cukrów w porcji (g),
- *potass* – ilość potasu w porcji (mg),
- *vitamins* – typowy procent zaspokojenia potrzeby dla kluczowych witamin i minerałów ustalonej przez FDA,
- *weight* – waga porcji w uncjach,
- *cups* – objętość porcji w szklankach,
- *rating* – Ocena w skali 0-100 obliczona przez Consumer Reports;

zmienne kategoryczne:

- *mfr* – producent płatków z mlekiem (zmienna nominalna):
    - A – American Home Food Products,
    - G – General Mills,
    - K – Kellogs,

  - N – Nabisco,
  - P – Post,
  - Q – Quaker Oats,
  - R – Ralston Purina;
- *type* – czy podawane na zimno czy ciepło (zmienna nominalna):
  - C – zimno,
  - H – ciepło;
- *shelf* – na jakiej półce są wystawione (zmienna porządkowa):
  - 1 – najbliższa podłogi,
  - 2 – druga od podłogi,
  - 3 – trzecia od podłogi;

Analiza danych

Analiza eksploracyjna danych (EDA) rozpoczęła się od przeglądu statystyk zmiennych ilościowych (Tabela 1), aby uzyskać wstępne zrozumienie ogólnej struktury zbioru danych oraz wykryć ewentualne anomalie lub niespójności. Dostarczyło to cennych informacji o tendencjach centralnych, zmienności oraz możliwych nieprawidłowościach. Warto zauważyć, że przed przystąpieniem do analizy przeskalowano zmienne odnoszące się do wartości odżywczej, aby odnosiły się do uncji, a nie do porcji oraz usunięto zmienne weight i cups. Pozwoliło to na porównanie płatków niezależne od wielkości porcji, co uznano za bardziej wartościowe.

Tabela 1 Statystyki opisowe dla zmiennych ilościowych

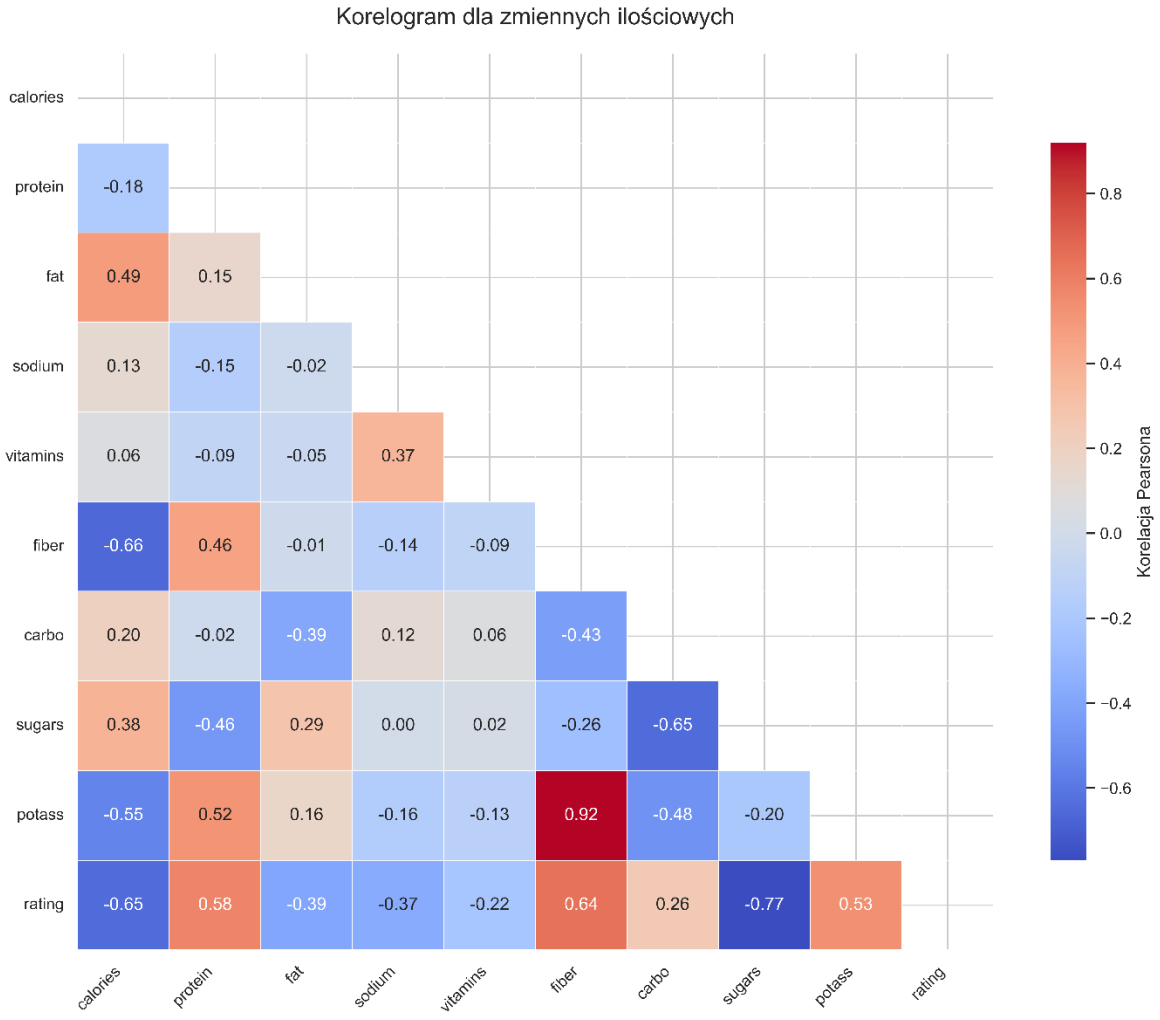
	Liczba	Średnia	Odchylenie	Min	25%	50%	75%	Max
calories	77.0	104	13.85	50.0	100.0	106.67	110.0	150.0
protein	77.0	2.49	1.08	1.0	2.0	2.26	3.0	6.0
fat	77.0	0.97	0.98	0.0	0.0	1.0	1.33	5.0
sodium	77.0	153.36	82.77	0.0	125.0	157.89	200.0	320.0
vitamins	77.0	26.87	20.51	0.0	25.0	25.0	25.0	100.0
fiber	77.0	2.05	2.30	0.0	1.0	1.54	3.0	14.0
carbo	76.0	14.63	4.23	5.0	11.46	14.0	17.0	26.0
sugars	76.0	6.65	4.09	0.0	3.0	6.01	10.0	15.0
potass	75.0	93.87	63.64	20.0	42.5	90.0	114.73	330.0
rating	77.0	42.67	14.05	18.04	33.17	40.40	50.83	93.70

Finalnie, zmienną vitamins postanowiono usunąć ze względu na zbyt niską zmienność.

Dodatkowo, zdecydowano zbadać korelacje między zmiennymi ilościowymi – co prawda, współliniowość do pewnego stopnia nie stoi na przeszkodzie grupowania, ale warto poznać strukturę danych. Stworzono korelogram (Wykres 1), który w sposób

przystępny ukazuje zależności między zmiennymi – wynika z niego, że bardzo silna korelacja istnieje wyłącznie pomiędzy potasem oraz błonnikiem.

Jeśli chodzi o wartości odstające to stwierdzono, że występują one dla niemalże każdej zmiennej. Nie można jednak powiedzieć, że wynikają one z błędów w danych bądź błędów pomiaru, a raczej z charakterystyki. Toteż, zdecydowano się nie przeprowadzać żadnej korekcji, ale należy mieć na uwadze fakt, że metody takie jak grupowanie K-Średnich są na to dość wrażliwe.



Wykres 1 Korelogram dla zmiennych numerycznych

W przypadku zmiennych kategorycznych zdecydowano się zbadać proporcje kategorii w zmiennych. Zdecydowano się na następujące zmiany:

- W przypadku zmiennej mfr pozostawiono kategorie „K” oraz „G”, zaś resztę zgrupowano w kategorię „Other”. Pozwoliło to uniknąć pracy z kategoriami bardzo rzadkimi (Tabela 2).

- W przypadku zmiennej type posiadającej wyłącznie dwie kategorie okazało się, że druga kategoria jest znacznie niedoreprezentowana, co spowodowało jej wykluczenie z dalszych analiz (Tabela 3).

Tabela 2 Proporcje klas dla zmiennej mfr

	Proporcja
K	0.299
G	0.286
P	0.117
Q	0.104
R	0.104
N	0.078
A	0.013

Tabela 3 Proporcje klas dla zmiennej type

	Proporcja
C	0.961
H	0.039

Dla pozostawionych zmiennych kategorycznych policzono współczynnik V-Cramera i przedstawiono go w tabeli poniżej (Tabela 4). Wartość nie była na tyle wysoka, żeby podejrzewać, że zmienne reprezentują to samo.

Tabela 4 Współczynnik V-Cramera dla zmiennych kategorycznych

	mfr	**shelf**
mfr	1.0	0.111
shelf	0.111	1.0

Wiktoria Arendarczyk, Cezary Moskal

Metodologia

Wstęp

W trakcie analizy wstępnej zbioru danych zidentyfikowano 4 brakujące wartości. W celu zapewnienia kompletności danych i rzetelności dalszych analiz, podjęto decyzję o ich uzupełnieniu za pomocą imputacji. Wybraną metodą jest algorytm **K najbliższych sąsiadów (KNN)**. Decyzja ta wynika z charakterystyki braków danych. Nie można założyć, że są to braki całkowicie losowe (MCAR - *Missing Completely at Random*), gdzie ich wystąpienie byłoby zupełnie niezależne od jakichkolwiek innych zmiennych. Znacznie bardziej prawdopodobnym scenariuszem jest, że są to braki losowe (MAR - *Missing at Random*). Oznacza to, że prawdopodobieństwo braku danej wartości jest powiązane z innymi obserwowanymi zmiennymi w zbiorze. Właśnie w takich warunkach algorytm KNN sprawdza się doskonale.

Wybrane metody

W projekcie zastosowano cztery różne algorytmy grupujące, które reprezentują różne podejścia metodologiczne: partycjonujące, hierarchiczne oraz gęstościowe. Porównanie wyników uzyskanych z tych metod pozwoliło na głębsze zrozumienie charakterystyki danych i wybór optymalnego podziału.

Toteż użyto, podsumowując:

1. **K-Means (K-Średnich)** – Jej działanie polega na podziale zbioru danych na z góry określoną liczbę klastrów (K). Algorytm działa iteracyjnie, dążąc do minimalizacji sumy kwadratów odległości euklidesowych każdego punktu od **centroidy** swojego klastra. Centroida jest punktem reprezentującym środek danego skupienia (obliczonym jako średnia arytmetyczna wszystkich punktów w klastrze). K-Means jest metodą szybką i efektywną, jednak najlepiej sprawdza się dla klastrów o sferycznym kształcie i jest wrażliwa na wartości odstające. Kluczowym elementem jest inicjalizacja położeń centroid – w praktyce najczęściej stosuje się metodę K-Means++, która dobiera punkty startowe w sposób minimalizujący ryzyko złej zbieżności. Dokładniej, pierwsza centroida wybierania jest całkowicie losowo, ale następne są losowane z prawdopodobieństwem proporcjonalnym do kwadratu odległości euklidesowej od najbliższej już wylosowanej centroidy.

2. **K-Medoids (K-Medoid)** – Stanowi wariant metody K-Średnich, zaprojektowany w celu zwiększenia odporności na obserwacje nietypowe. **Główna różnica polega na tym, że centra klastrów nie są centroidami, lecz medoidami – czyli rzeczywistymi punktami ze zbioru danych (takimi o najmniejszej sumarycznej odległości od innych punktów w klastrze).** Dzięki temu algorytm jest znacznie mniej podatny na wpływ wartości skrajnych (outlierów), co często prowadzi do bardziej trafnego grupowania w zaszumionych danych. Analogicznie do K-Means,

zastosowano inicjalizację za pomocą K-Medoids++, która również dobiera punkty startowe w sposób inteligentny. Tutaj prawdopodobieństwa są proporcjonalne do odległości taksówkowej. **W przypadku metody K-Medoids zastosowano odległości taksówkowe.**

3. **Grupowanie hierarchiczne (z łączeniem Warda i najdalszego sąsiada)** – Jest to aglomeracyjna (czyli "oddolna") metoda hierarchiczna. Na początku każda obserwacja stanowi osobny, jednoelementowy klaster. Następnie, w kolejnych krokach, algorytm łączy ze sobą pary najbliższych klastrów, aż wszystkie obserwacje znajdą się w jednym, wielkim skupieniu.

    1. Metoda Warda jako kryterium łączenia przyjmuje minimalizację przyrostu sumy kwadratów odległości euklidesowych od centroidy wewnątrz klastrów. W praktyce oznacza to, że na każdym etapie łączone są te dwa klastry, których fuzja w najmniejszym stopniu zwiększa całkowitą wariancję. Prowadzi to do tworzenia kompaktowych i w miarę równolicznych skupień. W przypadku tego łączenia zastosowano odległość euklidesową (nie jest nawet możliwe wykorzystanie innej).

    2. Metoda najdalszego sąsiada, wyznacza odległość między dwoma klastrami jako maksymalną odległość pomiędzy dowolnymi punktami każdego z nich. Oznacza to, że w każdym kroku aglomeracji łączone są te dwa skupienia, których fuzja minimalizuje tę maksymalną wartość – innymi słowy, poszukuje się pary klastrów o najmniejszą odległością między najbardziej oddalonymi od siebie punktami. W przypadku tego łączenia zastosowano odległość taksówkową.

4. **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)** – To algorytm oparty na gęstości, który definiuje klastry jako gęsto zaludnione obszary w przestrzeni, oddzielone od siebie obszarami o mniejszej gęstości. Jego działanie opiera się na dwóch parametrach: promieniu sąsiedztwa (ε) i minimalnej liczbie punktów. Algorytm DBSCAN najpierw wyszukuje tzw. punkty rdzenne – są to obserwacje, wokół których w zadanym promieniu ε znajduje się przynajmniej określona liczba sąsiadów (min_pts). Każdy taki punkt traktowany jest jako zalążek klastra, a następnie klaster jest „rozrastany" przez dołączanie wszystkich punktów mieszczących się w zadanym promieniu od punktu rdzennego; jeżeli wśród tych nowych punktów również znajdą się rdzenne, ich sąsiedztwo jest rekurencyjnie dołączane do klastra, aż wyczerpie się możliwość dalszego rozrostu. Obserwacje, które nie należą ani do żadnych klastrów, klasyfikowane są jako szum (outliery). Jego ogromną zaletą jest zdolność do odkrywania klastrów o dowolnych kształtach (np. wklęsłych, podłużnych) oraz automatyczne klasyfikowanie punktów nienależących do żadnego gęstego obszaru jako szum (wartości odstające). **W**

przypadku metody DBSCAN zastosowano odległości taksówkowe i euklidesowe.

Wybieranie liczby grup

Kluczowym wyzwaniem w analizie skupień jest wybór właściwej liczby grup, która najlepiej odzwierciedla naturalną strukturę danych. W zależności od zastosowanej metody, podejście do tego zagadnienia było zróżnicowane.

Dla metod partycjonujących, czyli K-Means i K-Medoids, zastosowano dwutorowe podejście analityczne. Po pierwsze, wykorzystano **metodę łokcia (Elbow Method)**, analizując wykres sumy kwadratów odległości od centroid wewnątrz klastrów (SSE/WCSS) w zależności od liczby grup (K) w przypadku K-Means, oraz po prostu sumy odległości od medoid (WCSD) w przypadku K-Medoid. Poszukiwano punktu „złamania" wykresu, po którym dalsze zwiększanie K nie przynosiło już znaczącego zysku informacyjnego. W celu potwierdzenia wyboru, posłużono się również **wskaźnikiem sylwetki (Silhouette Score)**. Mierzy on stopień spójności i separacji klastrów, a jego wartość dla pojedynczej obserwacji i jest definiowana wzorem

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$

gdzie $a(i)$ to średnia odległość od punktu i do pozostałych punktów w tym samym klastrze (spójność), a $b(i)$ to średnia odległość do punktów w najbliższym sąsiednim klastrze (separacja). Optymalna liczba grup to ta, dla której średni wskaźnik sylwetki dla całego zbioru jest najwyższy.

W przypadku grupowania hierarchicznego, decyzję o liczbie klastrów podjęto na podstawie wizualnej inspekcji dendrogramu. Analiza wysokości, na jakiej poszczególne gałęzie się łączą, pozwoliła na subiektywną, ale uzasadnioną ocenę i „cięcie" drzewa w miejscu, które najlepiej oddzielało logicznie uformowane skupienia.

Natomiast algorytm DBSCAN wyróżnia się tym, że nie wymaga wcześniejszego zdefiniowania liczby grup. Liczba klastrów jest dobierana automatycznie i stanowi bezpośredni wynik działania algorytmu, zależny od parametrów gęstości (ε i min_pts). W związku z tym dla tej metody nie prowadzono odrębnej procedury wyboru liczby skupień.

Dobieranie hiperparametrów

Strojenie hiperparametrów to ważny etap w dopracowywaniu modeli uczenia maszynowego. Polega na szukaniu takiej kombinacji ustawień modelu, która zapewni możliwie najlepsze wyniki – większą dokładność, mniejsze ryzyko nadmiernego dopasowania i lepsze działanie na nowych danych. W przypadku zastosowanych metod, sensowne jest przeprowadzanie takiego strojenia dla algorytmu DBSCAN.

Wiktoria Arendarczyk, Cezary Moskal

Wstępne analizy i modyfikacje zbioru danych

Imputacja brakujących wyników

Zaobserwowano cztery braki danych w trzech obserwacjach (Tabela 5). **Imputacja została wykonana za pomocą KKNN (Kernel K-Nearest Neighbors) – gdzie wynik był ważony odwrotnością odległości sąsiada (Tabela 6).** Zaimplementowane KKNN polega na tym, że najpierw oddziela się kolumnę celu od danych, a pozostałe cechy ilościowe poddaje się standaryzacji (StandardScaler), a kategoryczne koduje za pomocą OneHotEncoder (kolumna shelf została zostawiona taka jaka jest). Następnie dobiera się liczbę sąsiadów jako pierwiastek z liczby obserwacji i trenuje standardowy KNNImputer. Na koniec odwraca wszystkie transformacje, by zwrócić wyniki w oryginalnej skali.

Tabela 5 Dane z brakami (zaznaczone na czerwono)

	mfr	calories	protein	fat	sodium	fiber	carbo	sugars	potass	shelf	rating
Almond Delight	Other	110.0	2.0	2.0	200.0	1.0	14.0	8.0	**NaN**	3	34.38
Cream of Wheat (Quick)	Other	100.0	3.0	0.0	80.0	1.0	21.0	0.0	**NaN**	2	64.53
Quaker Oatmeal	Other	100.0	5.0	2.0	0.0	2.7	**NaN**	**NaN**	110.0	1	50.83

Tabela 6 Dane po imputacji KNN (k = 5)

	mfr	calories	protein	fat	sodium	fiber	carbo	sugars	potass	shelf	rating
Almond Delight	Other	110.0	2.0	2.0	200.0	1.0	14.0	8.0	**82.30**	3	34.38
Cream of Wheat (Quick)	Other	100.0	3.0	0.0	80.0	1.0	21.0	0.0	98.02	2	64.53
Quaker Oatmeal	Other	100.0	5.0	2.0	0.0	2.7	**16.03**	**3.49**	110.0	1	50.83

Warto wspomnieć, że oryginalnie dane mają wartości zaokrąglone do jedności, ale występowanie liczb po przecinku nie będzie tutaj problemem.

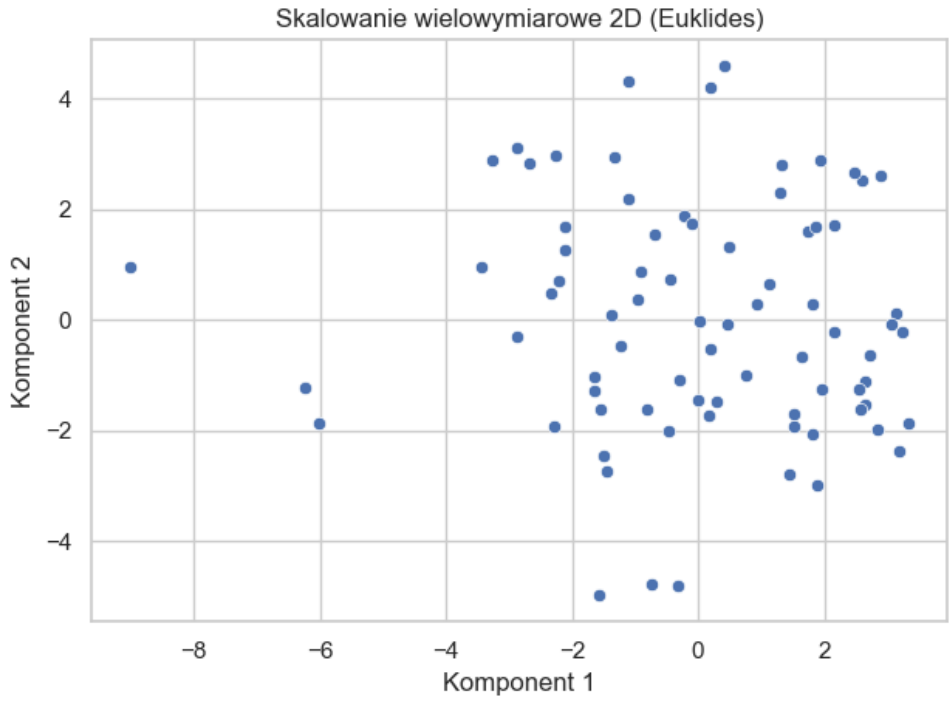
Transformacja danych

Najpierw zastosowano klasyczną standaryzację – centrowanie względem wartości średniej i skalowanie względem odchylenia standardowego - co pozwala zrównoważyć wpływ różnych jednostek pomiaru. **Klasyczna standaryzacja została użyta w przypadku odległości euklidesowych. Następnie, przeprowadzono centrowanie względem mediany i skalowanie według rozstępu międzykwartylowego (IQR), co zostało użyte w przypadku odległości taksówkowych.**

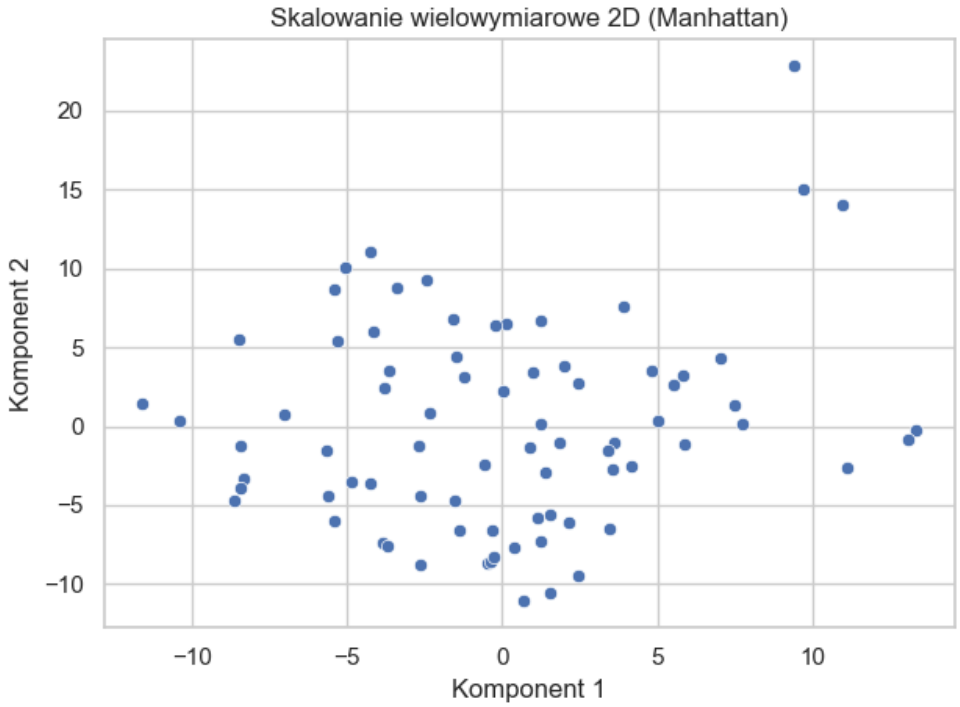
Zmienne kategoryczne podzielono na nominalne i porządkowe, przy czym dla wartości nominalnych wykorzystano kodowanie OneHotEncoder, tworząc zmienne binarne odzwierciedlające przynależność do poszczególnych kategorii. Z kolei zmiennej porządkowej nie trzeba było kodować, gdyż już była w formacie 1-3.

Wizualizacja danych w przestrzeni 2D

Dla każdego typu odległości (Euklides – standaryzacja [Wykres 2], Manhattan – IQR [Wykres 3]) zastosowano MDS w celu zobrazowania danych w przestrzeni 2D. MDS to technika analizy danych służąca odwzorowaniu obiektów z przestrzeni wysokowymiarowej w przestrzeń o mniejszej liczbie wymiarów (np. 2D lub 3D) tak, aby zachować możliwie jak najwierniej odległości (lub podobieństwa) między nimi.



Wykres 2 MDS (Euklides)



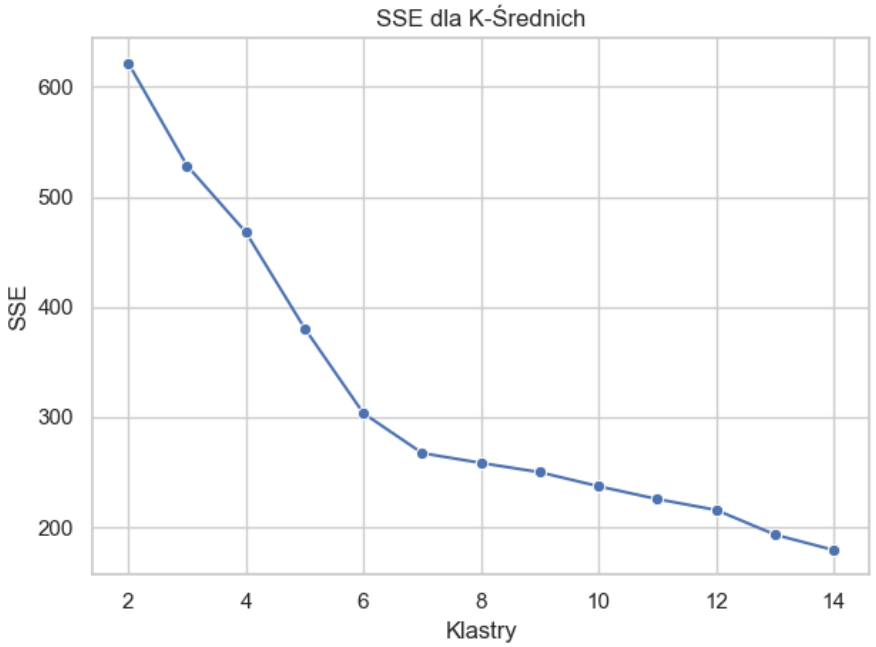
Wykres 3 MDS (Manhattan - Taksówka)

Na pierwszy rzut oka można zauważyć, że istnieć będą na pewno przynajmniej dwie grupy – centralna, i ta zawierająca outliery. Aczkolwiek, teraz należy zobaczyć jak widzieć to będą algorytmy.

Wyniki doboru klastrów

K-Means

W przypadku metody K-Means wstępnie należy ustalić, ile klastrów będzie wyborem optymalnym – taką decyzję można podjąć na podstawie wykresu łokciowego (Wykres 4), oraz wskaźnika sylwetki (Wykres 5).



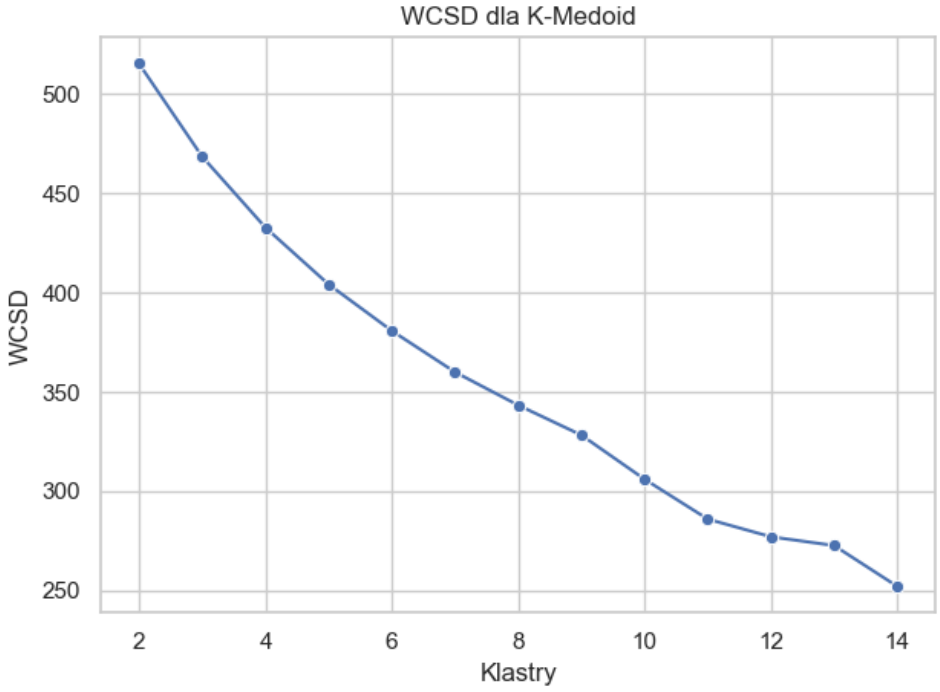
Wykres 4 Wykres łokciowy dla K-Means

Wykres 5 Wskaźnik sylwetki dla K-Means

Można zauważyć, że oba wykresy ewidentnie wskazują na liczbę klastrów równą 6 – łokieć w tym miejscu się załamuje, a wskaźnik sylwetki osiąga maksimum.

K-Medoids

W przypadku K-Medoids sytuacja jest analogiczna do K-Means. Przedstawiono wykres łokciowy oraz wskaźnik sylwetki.



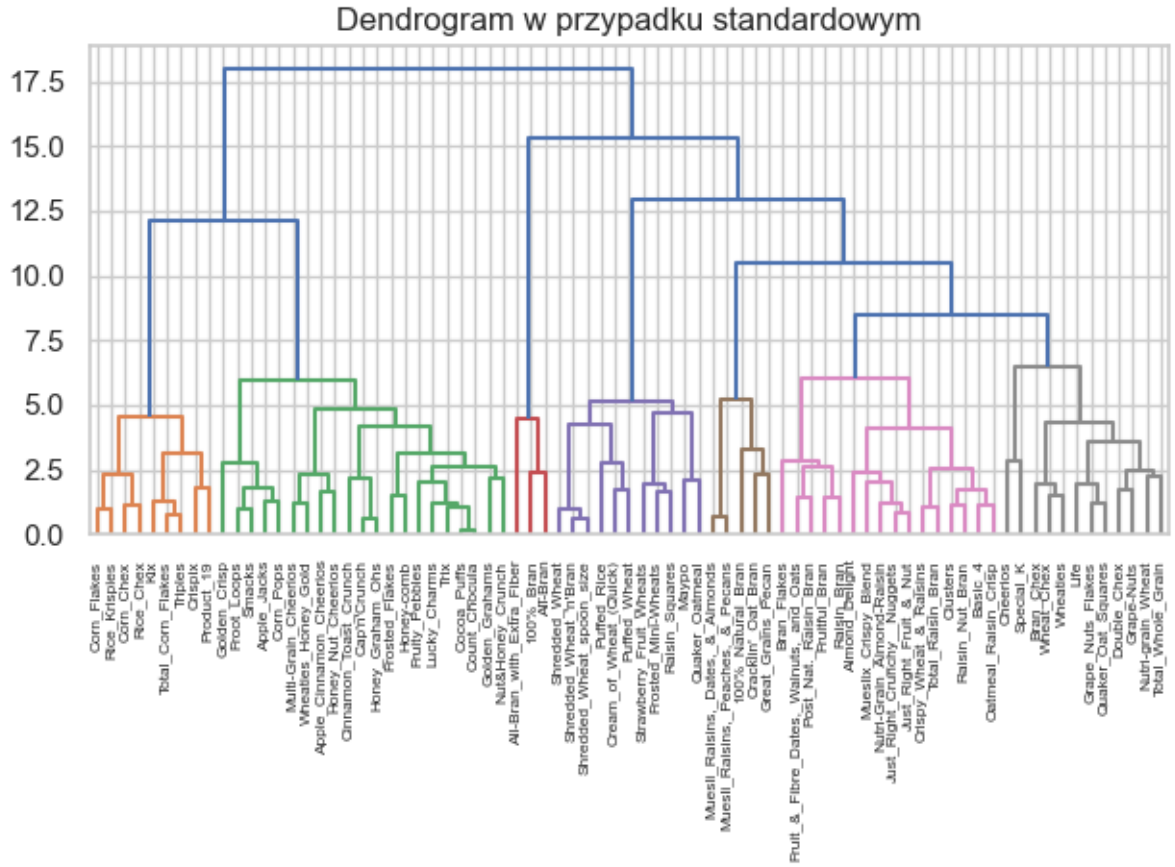
Wykres 6 Wykres łokciowy dla K-Medoids

Wykres 7 Wskaźnik sylwetki dla K-Medoids

Z wykresów widać, że ta metoda nie działa dobrze dla tego zbioru danych – zdecydowano wybrać się dwa klastry, ale zauważyć należy, że łokieć tak nie ma widocznego załamania, a wskaźnik sylwetkowy również wskazuje kilka sensownych – z jego punktu widzenia - kandydatów (w tym 2 klastry).

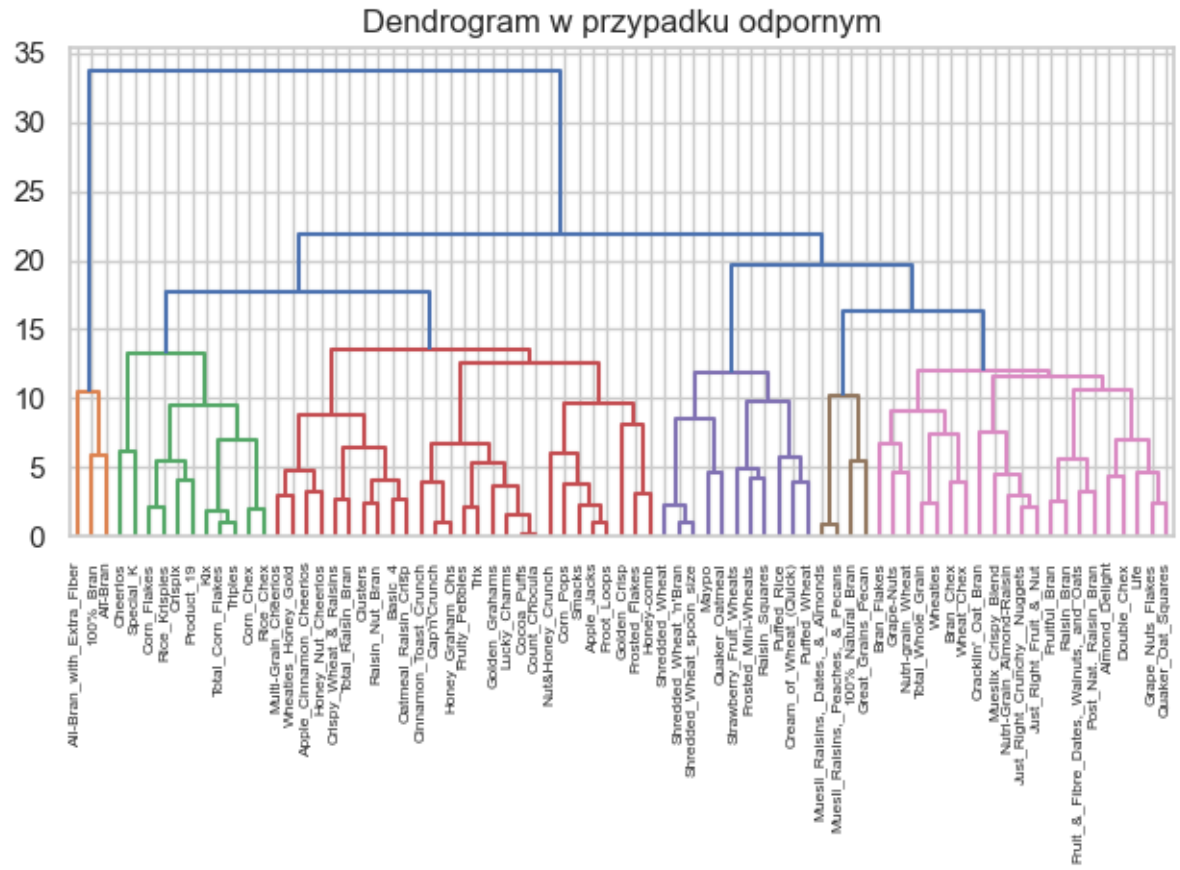
Grupowanie hierarchiczne

W przypadku grupowania hierarchicznego należy przyciąć dendrogram w „poprawnym” miejscu, ale tak naprawdę nie ma żadnego wskaźnika, który by w tym pomógł. W przypadku standardowym (łączenie Warda, odległość euklidesowa [Wykres 8]) dendrogram zdecydowano się przyciąć na wysokości 7.



Wykres 8 Dendrogram dla łączenia Warda z odległością euklidesową

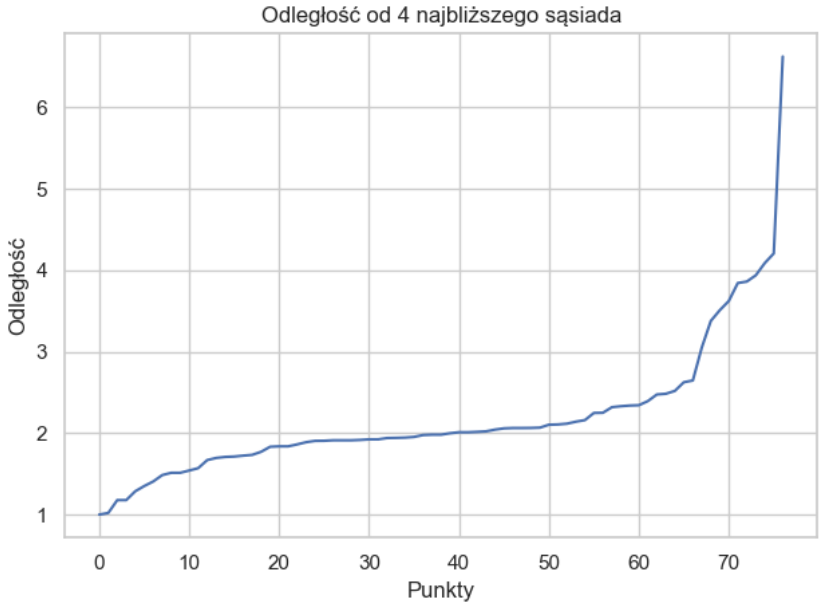
W przypadku łączenia najdalszego sąsiada z odległością taksówkową (Wykres 9) zdecydowano przyciąć dendrogram na wysokości 14.



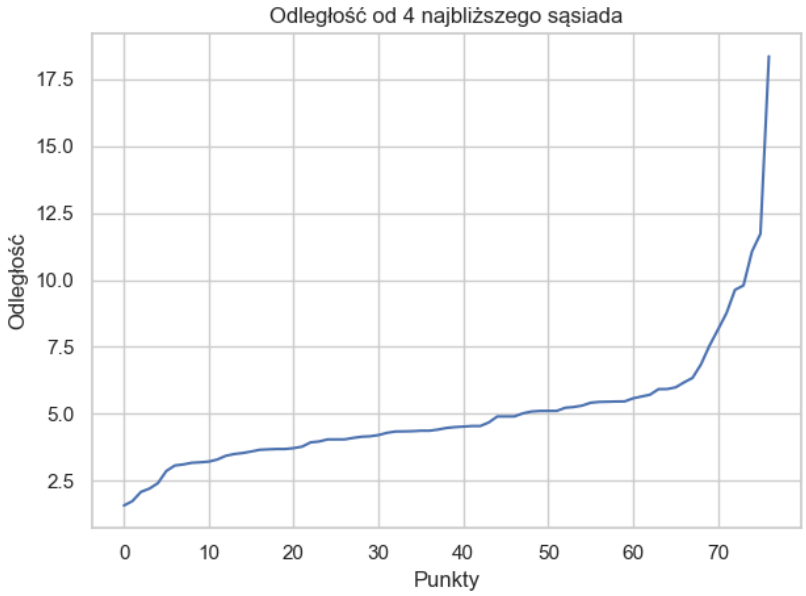
Wykres 9 Dendrogram dla łączenia najdalszego sąsiada z odległością taksówkową

DBSCAN

W przypadku algorytmu klastrującego DBSCAN nie wybiera się liczby klastrów, ale dostosowuje się hiperparametry, a algorytm sam na ich podstawie określa odpowiednią liczbę grup. Parametr min_pts został przyjęty jako 4, gdyż danych jest mało i pożądane jest aby algorytm nie był zbyt agresywny w odrzucaniu wartości odstających. ε został dobrany na podstawie dystansu od 4-tego najbliższego sąsiada w zestawie danych (Euklides - Wykres 10, Taksówka - Wykres 11).



Wykres 10 Odległość od 4 najbliższego sąsiada (Euklides)



Wykres 11 Odległość od 4 najbliższego sąsiada (Taksówka)

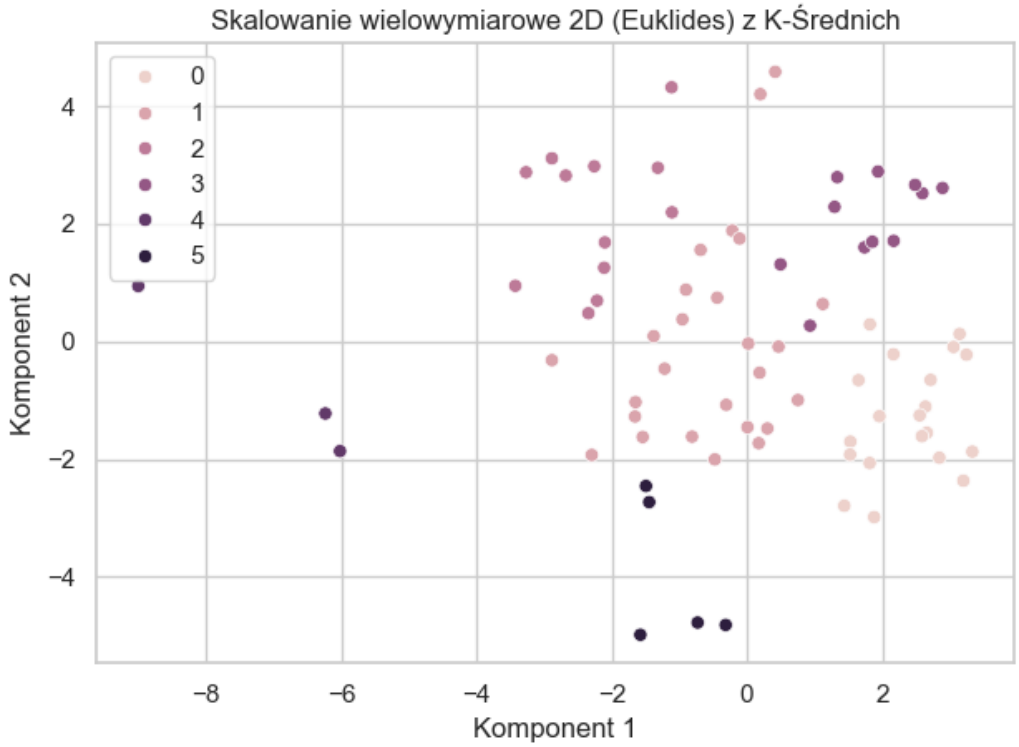
W przypadku odległości euklidesowej zdecydowano się przyjąć wartość 2.6, zaś w przypadku odległości taksówkowej 6.25.

Wyniki grupowania

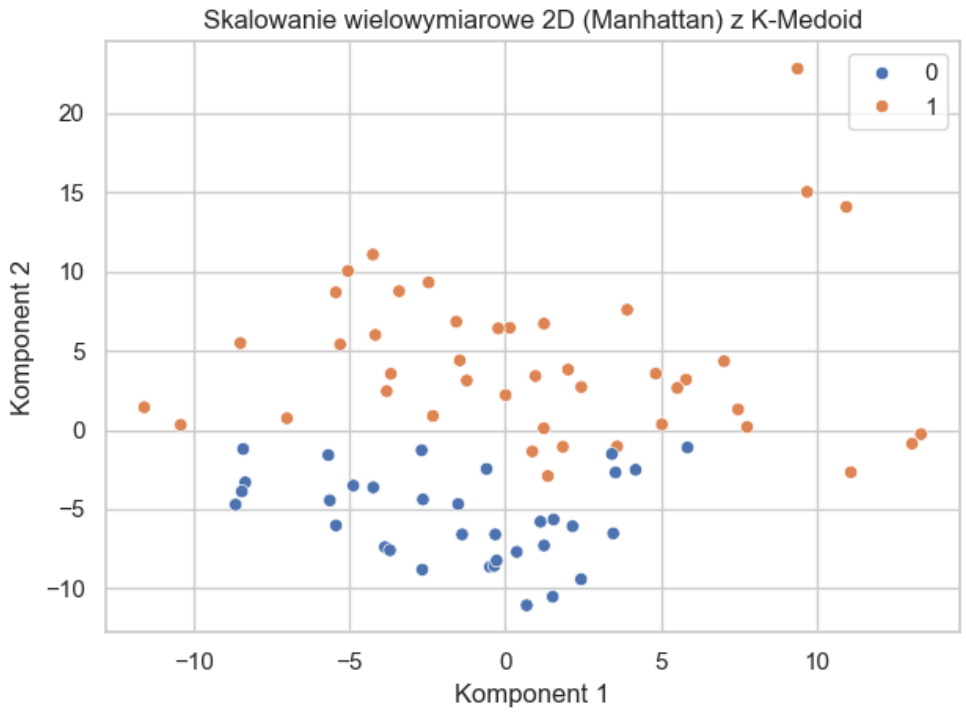
Wizualizacja grupowania za pomocą MDS

Wyniki grupowania dla każdej metody zostały zwizualizowane za pomocą MDS w celu sprawdzenia, czy wyglądają one sensownie. Podsumowując, zwizualizowano:

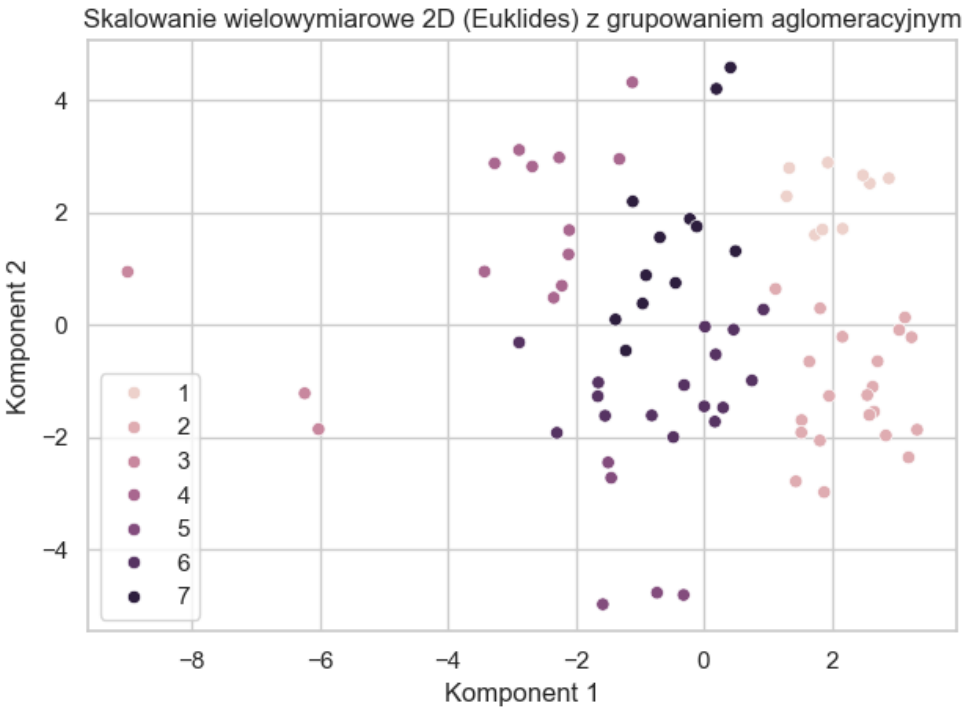
- K-Means z 6 klastrami (Euklides) [Wykres 12],
- K-Medoids z 2 klastrami (Manhattan/Taksówka) [Wykres 13],
- Grupowanie hierarchiczne z 7 klastrami (Ward – Euklides) [Wykres 14],
- Grupowanie hierarchiczne z 6 klastrami (łączenie najdalszego sąsiada – Manhattan/Taksówka) [Wykres 15],
- DBSCAN z 1 klastrem (Euklides) [Wykres 16],
- DBSCAN z 1 klastrem (Manhattan/Taksówka) [Wykres 17].



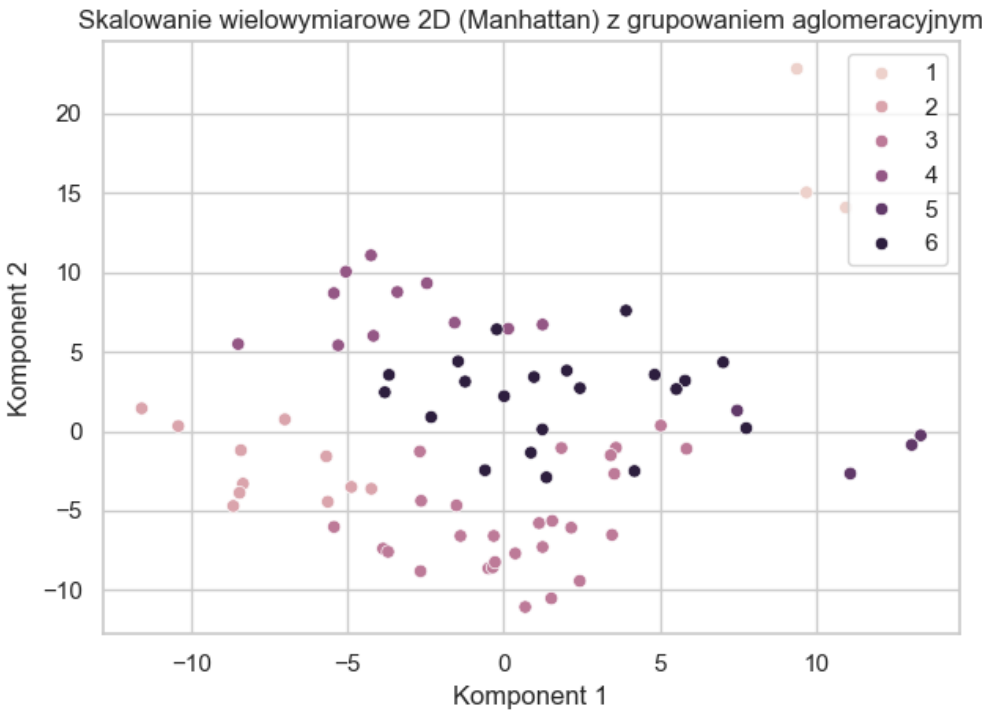
Wykres 12 MDS dla K-Means z 6 klastrami (Euklides)



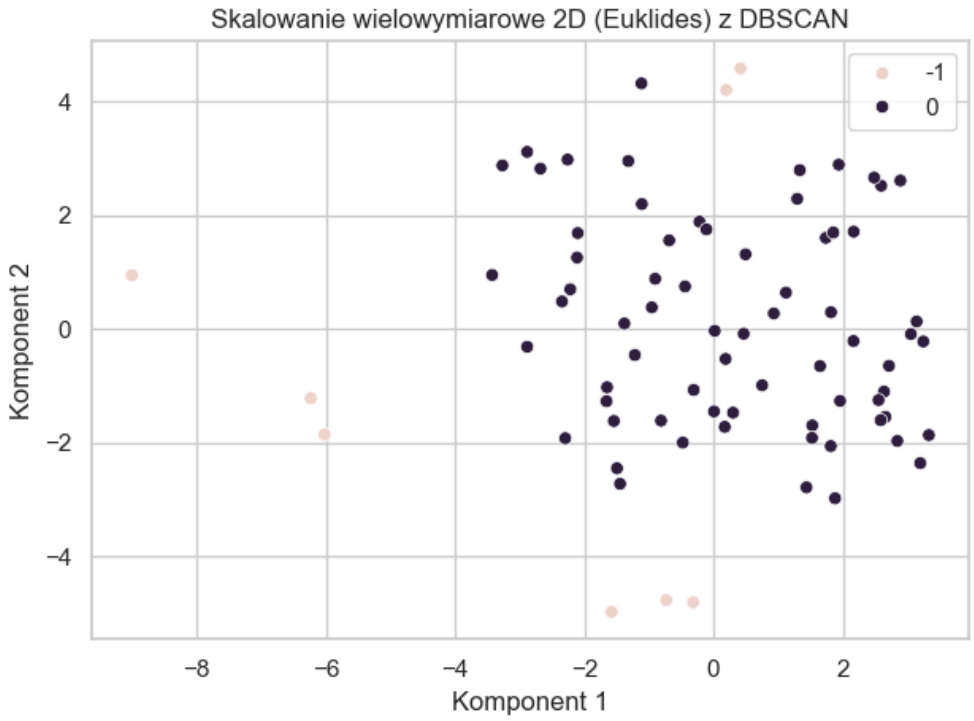
Wykres 13 MDS dla K-Medoids z 2 klastrami (Taksówka)



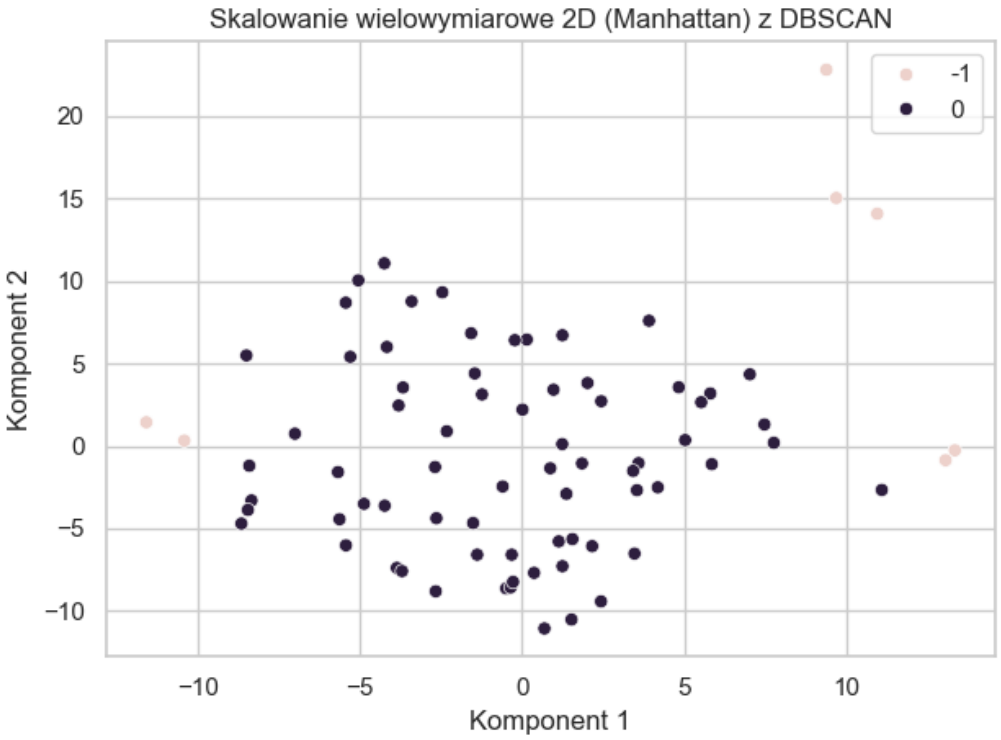
Wykres 14 MDS dla grupowania hierarchicznego z 7 klastrami (Ward – Euklides)



Wykres 15 MDS dla grupowania hierarchicznego z 6 klastrami (łączenie najdalszego sąsiada – Taksówka)



Wykres 16 DBSCAN z 1 klastrem (Euklides)



Wykres 17 DBSCAN z 1 klastrem (Taksówka)

Tabela 7 Wskaźniki sylwetki dla wszystkich metod

	Euklides	**Manhattan/Taksówka**
K-Means/K-Medoids	0.257	0.226
Linkage	0.245	0.249
DBSCAN	0.354	0.375

W przypadku DBSCAN otrzymano największe wartości wskaźnika sylwetki, natomiast w tym przypadku nie usunięto outlierów – toteż, indeks ten jest dodatkowo zaniżony. Dodatkowo, należy zauważyć, że po usunięciu outlierów posiada on tylko jeden klaster, więc w sumie nic nie wymyślił - nie znalazł sensownych grup. Sugeruje to, że mogą nie istnieć istotnie różne klastry.

Tak czy inaczej, w tym projekcie finalnie spróbowano opisać klastry dwóch metod – K-Means oraz grupowania hierarchicznego dla łączenia najdalszego sąsiada.

Wiktoria Arendarczyk, Cezary Moskal

Próba opisania klastrów

K-Means

Tabela 8 Porównanie zmiennych numerycznych dla klastrów w przypadku K-Means

Klaster	calories	protein	fat	sodium	fiber	carbo	sugars	potass	rating
0	111.5	1.5	1	170	0.5	12.5	11.55	43.75	28.29
1	100.44	2.86	1.1	170.33	2.6	13.39	6.3	115.7	42.62
2	95.53	3.03	0.25	22.08	2.28	18.28	2.21	99.79	61.85
3	107.27	2	0.36	246.36	0.45	20.64	3.27	43.18	41.48
4	63.33	4	0.67	176.67	11	6.67	3.67	310	73.84
5	130	3.4	3.4	95	3	12.6	8.2	147	38.3

Tabela 9 Porównanie zmiennej shelf dla klastrów w przypadku K-Means

Klaster	shelf		
	1	2	3
0	0.3	0.7	0
1	0.23	0.08	0.69
2	0.33	0.33	0.33
3	0.36	0.09	0.55
4	0	0	1
5	0	0	1

Tabela 10 Porównanie zmiennej mfr dla klastrów w przypadku K-Means

Klaster	mfr		
	G	K	Other
0	0.45	0.3	0.25
1	0.38	0.23	0.38
2	0	0.25	0.75
3	0.27	0.45	0.27
4	0	0.67	0.33
5	0	0.2	0.8

Tabela 11 Porównanie liczebności każdego klastra w przypadku K-Means

Klaster	Liczba
0	20
1	26
2	12
3	11
4	3
5	5

Grupowanie hierarchiczne (Complete – Taksówka)

Tabela 12 Porównanie zmiennych numerycznych dla klastrów w przypadku grupowania hierarchicznego (łączenie najdalszego sąsiada – Taksówka)

Klaster	calories	protein	fat	sodium	fiber	carbo	sugars	potass	rating
1	63.33	4	0.67	176.67	11	6.67	3.67	310	73.84
2	108.18	2.73	0.45	260.91	0.55	20.45	2.64	45	43.58
3	108.71	1.73	1.1	165.21	0.88	12.36	10.57	61.28	30.3
4	96.04	3.04	0.27	8.64	2.21	18.3	2.23	100.68	62.05
5	135	3.5	3.5	83.75	2.75	13.25	8.5	143.75	37.77
6	99.82	2.65	0.99	167.52	2.85	14.02	5.97	117.08	44.41

Tabela 13 Porównanie zmiennej shelf dla klastrów w przypadku grupowania hierarchicznego (łączenie najdalszego sąsiada – Taksówka)

Klaster	shelf		
	1	2	3
1	0	0	1
2	0.55	0.09	0.36
3	0.26	0.52	0.22
4	0.36	0.36	0.27
5	0	0	1
6	0.14	0.1	0.76

Tabela 14 Porównanie zmiennej mfr dla klastrów w przypadku grupowania hierarchicznego (łączenie najdalszego sąsiada – Taksówka)

Klaster	mfr		
	G	K	Other
1	0	0.67	0.33
2	0.36	0.45	0.18
3	0.59	0.22	0.19
4	0	0.18	0.82
5	0	0	1
6	0.1	0.38	0.52

Tabela 15 Porównanie liczebności każdego klastra w przypadku grupowania hierarchicznego (łączenie najdalszego sąsiada – Taksówka)

Klaster	Liczba
1	3
2	11
3	27
4	11
5	4
6	21

Próbując nadać temu sens, to można zobaczyć, że metody wyodrębniają grupy takie jak:

- **Niskokaloryczne, bogate w błonnik („fit”),**
    - K-Means: 4,
    - Hierarchiczne: 1.
- **Wysokobiałkowe, niskocukrowe, średniokaloryczne („fit gym”),**
    - K-Means: 2,
    - Hierarchiczne: 4.
- **Bardzo kaloryczne i tłuste („treściwe”),**
    - K-Means: 5,
    - Hierarchiczne: 5.
- **Słodkie, niskobłonnikowe („śniadaniowe słodycze”),**
    - K-Means: 0,
    - Hierarchiczne: 3.

Pozostałe grupy można uznać za „przeciętne” i autorzy nie posiadają wystarczającej wiedzy o płatkach i żywieniu, aby móc je sensownie zinterpretować.

Wnioski

Na podstawie przeprowadzonych analiz i wizualizacji (szczególnie MDS) można wyciągnąć pewne wnioski. Po pierwsze, projekcje uzyskane za pomocą MDS jasno wskazują, że obserwacje – czyli poszczególne rodzaje płatków – układają się na płaszczyźnie w sposób ciągły, bez wyraźnych obszarów koncentracji czy szczelin oddzielających odrębne gęstości poza wartościami odstającymi.

Algorytm DBSCAN, który identyfikuje gęste regiony i uważa za anomalie te punkty, które nie należą do żadnego z takich „rdzeni", w praktyce oznajmił istnienie jednego skupiska obejmującego wszystkie obiekty bez outlierów. To zachowanie jest w pełni zgodne z obserwowaną, słabą rozróżnialnością w zbiorze.

Pomimo braku wyraźnych granic, zarówno metoda K-Means, jak i grupowanie hierarchiczne pozwoliły wyodrębnić kilka sensownych segmentów. Te wyniki wskazują, że chociaż granice pomiędzy klastrami są rozmyte, to pewne profile żywieniowe można wydzielić i opisać w sposób spójny.

Wydaje się, że zestaw ten średnio nadaje się do klasycznych metod klasteryzacji, zwłaszcza takich, które zakładają istnienie wyraźnych, gęstych skupisk w przestrzeni cech.

Podsumowując, mimo że poszczególne algorytmy „znalazły" pewne segmenty – co dostarczyło wartościowych opisów profili żywieniowych – to ogólna struktura danych nie sprzyja wyjściowej idei naturalnego, wyraźnego podziału na grupy. W praktycznych zastosowaniach rekomendowane jest rozważenie metod nadzorowanych czy analiz wielowymiarowych pozwalających pracować z ciągłym charakterem zmienności tego typu produktów.

Wykaz tabel

Wykaz wykresów