Niehierarchiczne metody analizy skupień

Marzena Nowakowska
Wydział Zarządzania i Modelowania Komputerowego
Politechnika Świętokrzyska

Istota metod niehierarchicznych

Metody niehierarchiczne polegają na podziale zbioru danych na pewną liczbę skupień, bez tworzenia struktury drzewiastej. Mogą działać poprzez iteracyjne optymalizowanie podziału ma określoną z góry liczbę skupień lub poprzez wykrywanie struktur w danych, takich jak obszary większej gęstości, bez konieczności podawania liczby klastrów.

Zalety

- Bardziej skalowalne (lepiej radzi sobie ze wzrostem wielkości danych, zarówno liczby obiektów, jak i liczby cech) niż metody hierarchiczne
- Prosta implementacja i interpretacja
- Możliwość łatwego ponawiania podziału z różnymi parametrami

Wady

- Część metod (np. metody partycjonujace) wymaga określenia liczby klastrów
- Wyniki zależą od wartości początkowych (mogą utknąć w minimach lokalnych)
- Najlepiej działają dla kulistych, podobnie licznych klastrów; słabe dla skupień o nieregularnych kształtach

Metody niehierarchiczne; klasyczne modelowe

To grupa algorytmów, w których zakłada się, że dane pochodzą z mieszaniny prostszych rozkładów (modeli), np. normalnych.

Zamiast szukać grup na podstawie odległości, te metody modelują nieznany rozkład danych, aby znaleźć skupienia.

Jest wymagane określenie liczby docelowych skupień.

Przykładowe metody modelowe

- Gaussian Mixture Models (GMM)
- EM (Expectation-Maximization)
- AutoClass
- Latent Class Analysis (LCA)

Metody niehierarchiczne; rozmytej analizy skupień

Elementy mogą być przydzielane do więcej niż jednego skupienia z uwzględnieniem prawdopodobieństwa przynależności. Ta częściowa przynależność wyrażona współczynnikiem $\mu \in [0,1]$.

Jest wymagane określenie liczby docelowych skupień.

Przykładowa metoda analizy rozmytej

- Fuzzy k-means

Metody niehierarchiczne; partycjonujące

Dzielą zbiór danych na **ustaloną liczbę klastrów (k)**, aby minimalizować wariancję wewnątrz każdej grupy; tworzą płaski podział, zakładają pewien kształt skupień (zwykle kulisty).

W tych metodach każdy klaster jest reprezentowany przez centroid – punkt będący średnią (lub inną „centralną” reprezentacją) wszystkich obiektów należących do danego skupienia. Algorytm bazuje na iteracyjnym:

- przypisywaniu obiektów do najbliższego centroidu,
- aktualizowaniu centroidów jako średniej obiektów w klastrach.

Zadaniem algorytmu jest minimalizacja sumy kwadratów odchyleń punktów x_j od środka m_i swojego klastra C_i:

$$J = \sum_{i=1}^{k} \sum_{x_j \in C_i} d(x_j, m_i)$$

Wykorzystuje się metryką euklidesową.

Cechy metod partycjonujących

Istotna rolę odgrywa centroid jako reprezentant klastra
Działają najlepiej, gdy klastry mają zbliżone zagęszczenie i kształt (najczęściej kulisty).

Najpopularniejsze metody centroidowe

- k-means – klasyczny algorytm wykorzystujący średnią arytmetyczną jako centroid.
- k-medians – wykorzystuje medianę zamiast średniej (bardziej odporna na wartości odstające).
- k-medoids (PAM, CLARA) – centroidem jest rzeczywisty obiekt (medoid), nie średnia; lepsze dla danych nienumerycznych i wrażliwych na wartości odstające
- k-modes – wersja dla danych kategorycznych (centroid to modalna wartość cech)

Algorytmy: k-means, k-medoids

1. Wybierz liczbę skupień K (z góry).
2. Wylosuj początkowe centroidy (średnie – punkty spoza danych lub medoidy – rzeczywiste obserwacje wybrane losowo z danych).
3. Przypisz kolejną obserwację do najbliższego centroidu (tworząc wstępne klastry).
4. Oblicz nowe centroidy (jako średnie arytmetyczne lub medoid punktów przypisanych do każdego klastra.
5. Powtarzaj kroki 3–4, aż położenie centroidów przestanie się zmieniać (układ klastrów nie zmienia się).

Oba algorytmy są wrażliwe na początkowy wybór centroidów

Algorytm k-means	**Algorytm k-medoids**
• Zakłada podobny rozmiar i gęstość skupień • Źle działa, gdy dane mają zmienne o różnej skali (wymaga standaryzacji zmiennych)	• Odporny na wartości odstające • Dostarcza bardziej stabilne wyniki niż k-średnich, ale wolniejszy • Mniej efektywny dla bardzo dużych zbiorów danych.

Metody niehierarchiczne; siatkowe

Technika grupowania wielowymiarowych danych poprzez podział przestrzeni danych na siatkę komórek. Porównuje się dane na poziomie tych komórek, a nie na poziomie pojedynczych punktów danych, co sprawia, że metody te są szybkie i mają niską złożoność obliczeniową

Przykłady

- STING
- CLIQUE

Metody niehierarchiczne; gęstościowe

Grupowanie gęstościowe polega na budowaniu grup o dużej gęstości punktów (obserwacji) w przestrzeni. Zagęszczenie punktów wyznacza się na podstawie odległości między nimi.

Metody gęstościowe bardzo dobrze sprawdzają się przy identyfikacji klastrów o nieregularnych (niewypukłych) kształtach. Cechy:

- Liczba grup nie jest znana z góry; można częściowo zmieniać poprzez zmianę parametrów modelu.
- Wymaga zdefiniowania minimalnej liczby obserwacji potrzebnych do zbudowania grupy.
- Wymaga zdefiniowania minimalnej odległości, która definiuje sąsiada.
- Nie wszystkie obserwacje muszą zostać przypisane do grup. Obserwacje niespełniające zakładanych kryteriów są uznawane za obserwacje odstające.

Przykłady

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
- OPTICS (Ordering Points To Identify the Clustering Structure)
- DENCLUE (DENsity-based CLUstEring)

Ocena jakości skupień

Rozważa się dwa główne aspekty:

- spójność (*cohesion*) – jak bardzo elementy w tym samym klastrze są do siebie podobne
- separację (*separation*) – jak bardzo różne są od siebie klastry (czy są dobrze od siebie oddzielone).

Dobry podział to taki, gdzie obiekty wewnątrz klastra są blisko siebie, a klastry są daleko od siebie.

Podstawowe miary ww. podejścia (idea)

WSS – *Within Sum of Squares* (*withinss*)	$\mathrm{WSS} = \sum_{k=1}^{K} \sum_{x_i \in C(k)} \sum_{x_j \in C(k)} d(x_i, x_j)$
BSS – *Between Sum of Squares* (*betweenss*)	$\mathrm{BSS} = \sum_{k=1}^{K} \sum_{x_i \in C(k)} \sum_{x_j \in C(k)} d(x_i, x_j)$
TSS – *Total Sum of Squares* (*totss*)	$\mathrm{TSS} = \sum_{i=1}^{n} \sum_{j=1}^{n} d(x_i, x_j)$

Inne miary oceny jakości skupień

- Wskaźniki sylwetki (*Silhouette coefficients*). Wyznaczany dla każdego punktu. Mierzy jak bardzo obiekt pasuje do swojego klastra w porównaniu z innymi. Pochodne: średni wskaźnik sylwetki dla każdego skupienia oraz dla całego modelu (średnia sylwetka wszystkich punktów). Interpretacja wskaźnika sylwetki :
  - [0,5; 1] → bardzo dobrze dopasowany model; dobra separacja wyrazista struktura skupień
  - [0,25; 0,5] → średnio dopasowany model; struktura skupień istnieje, ale nie jest bardzo wyraźna
  - [0; 0,25] → na granicy, słaba klasteryzacja; punkty w skupieniach nie są wyraźnie oddzielone od punktów w innych skupieniach, model może mieć za mało lub za dużo skupień
  - [-1, 0) → zła klasteryzacja; punkt jest bliżej innego skupienia niż swojego, może być za dużo skupień, źle dobrana miara odległości lub brak naturalnej struktury skupień
- Wskaźnik Calinskiego-Harabasza (*CH Index*)

  CH = międzyklastrowa_wariancja / wewnątrzklastrowa_wariancja

  Im większa wartość tym lepiej
- Wskaźnik Daviesa-Bouldina (DB Index) mierzy, jak bardzo klastry się nakładają. Im mniejsza wartość, tym lepiej.
- Wskaźniki wizualne
  - Wykres sylwetek
  - Mapy cieplne odległości między klastrami
  - Redukcja wymiarowości (PCA, t-SNE, UMAP) i wizualna ocena, czy klastry są sensownie oddzielone

Dane do analiz – ocena jakości wody

- **ph** – określa kwasowość/zasadowość wody. Zbyt niski lub zbyt wysoki pH może być szkodliwy dla zdrowia i wpływa na reakcje chemiczne w wodzie.
- **Hardness** (twardość) – wysoka twardość nie jest toksyczna, ale wpływa na smak wody i osadzanie się kamienia. Mniej krytyczna dla jakości zdrowotnej.
- **Solids** (rozpuszczone substancje stałe) – wysoki poziom wskazuje na dużą zawartość minerałów i zanieczyszczeń. To istotny wskaźnik jakości wody.
- **Chloramines** – używane do dezynfekcji wody; ich nadmiar może być szkodliwy. Ważny parametr zdrowotny.
- **Sulfate** (siarczany) – w nadmiarze mogą powodować smakową uciążliwość i działanie przeczyszczające. Ważne, ale raczej pomocniczo.
- **Conductivity** (przewodność) – informuje o ogólnej zawartości jonów w wodzie, czyli pośrednio o jakości. Często używane jako wskaźnik sumaryczny.
- **Organic_carbon** (węgiel organiczny) – wskaźnik obecności związków organicznych, które mogą prowadzić do tworzenia się szkodliwych produktów ubocznych po chlorowaniu. Bardzo istotny.
- **Trihalomethanes** – produkty uboczne chlorowania wody, które są rakotwórcze. Bardzo ważny parametr zdrowotny.
- **Turbidity** (mętność) – wysoka mętność może wskazywać na obecność mikroorganizmów i zanieczyszczeń fizycznych. Kluczowy wskaźnik jakości mikrobiologicznej
- **Potability** – wskaźnik „zdatności do spożycia” (0 = „nie nadaje się”, 1 = „nadaje się”). Zmienna wykluczona z analizy skupień.

https://huggingface.co/datasets/DarkNeuronAI/water-potability-3k

Etapy badań: przygotowanie danych

Etap 1. Pobranie danych - program w Pythonie

Pobranie danych z portalu Kaggle i zapisanie go do pliku *water.csv*

Etap 2. Preprocessing danych – program w Pythonie

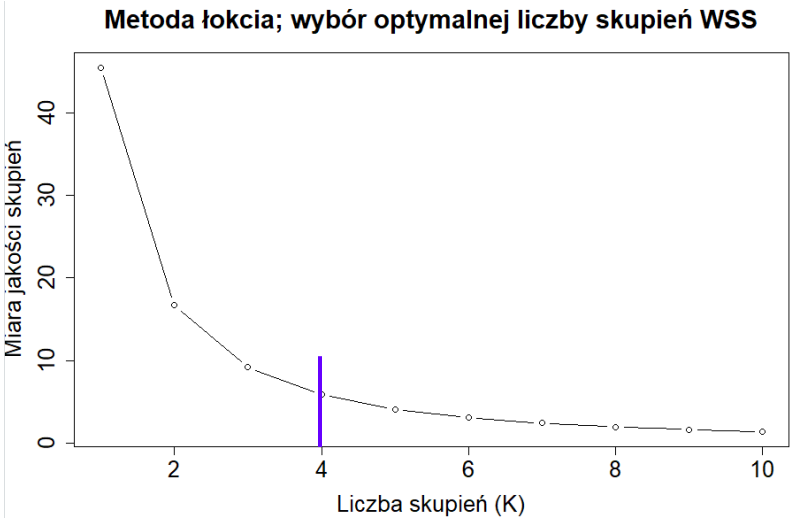
1. Wczytanie pliku *water.csv* do ramki danych
2. Wyprowadzanie informacji o pliku
3. Imputacja; średnia
4. Transformacja zmiennych (bez *Potability*); normalizacja min-max do dziedziny [0, 1]. Dopisanie znormalizowanych zmiennych do ramki z danymi oryginalnymi
5. Zapis przygotowanych danych do pliku csv o nazwie *water_norm.csv*

```
RangeIndex: 3276 entries, 0 to 3275
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ph               2785 non-null   float64
 1   Hardness         3276 non-null   float64
 2   Solids           3276 non-null   float64
 3   Chloramines      3276 non-null   float64
 4   Sulfate          2495 non-null   float64
 5   Conductivity     3276 non-null   float64
 6   Organic_carbon   3276 non-null   float64
 7   Trihalomethanes  3114 non-null   float64
 8   Turbidity        3276 non-null   float64
 9   Potability       3276 non-null   int64
```

Etapy badań: ustalenie zbioru i uruchomienie k-means

Etap 3. Wybranie losowo 25% obserwacji ze zbioru – program w R

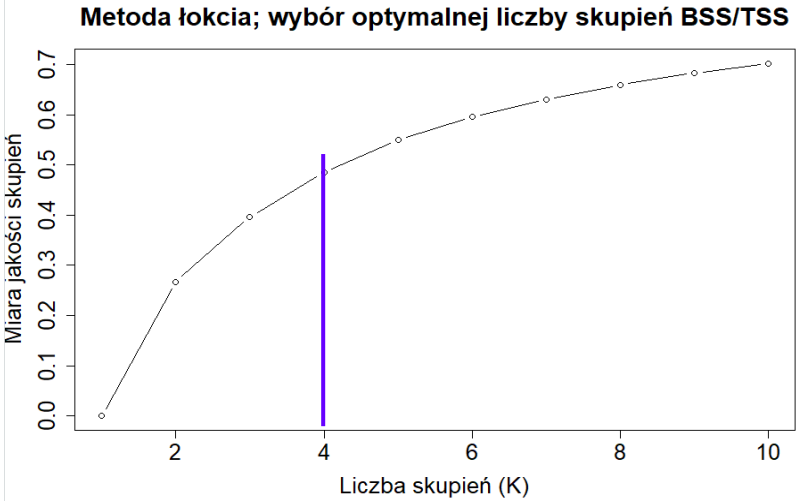
- Losowanie równomiernie (prosta próba losowa z każdej warstwy) – pobranie 25% obserwacji z każdej warstwy (0 i 1) zmiennej „Potability”
- Zmieszanie rekordów z obu warstw
- Zapis wyniku do pliku water_25.csv



Etap 4. Uruchomienie analizy skupień dla K=4 – program w R

Studenci sprawdzą znaczenie parametrów i zwracany wynik funkcji *kmeans*

kmeans(x, centers, iter.max = 10, nstart = 1, algorithm = c("Hartigan-Wong", "Lloyd", "Forgy", "MacQueen"))



Analiza skupień k-means - wyniki

km	list [9] (S3: kmeans)	List of length 9
cluster	integer [820]	3 4 4 2 3 2 ...
centers	double [4 x 3]	0.568 0.592 0.584 0.369 0.307 0.541 0.594 0.475 0.556 0.410 0.637 0.521 ...
totss	double [1]	45.49621
withinss	double [4]	6.67 6.11 5.67 4.98
tot.withinss	double [1]	23.43039
betweenss	double [1]	22.06581
size	integer [4]	235 205 198 182
iter	integer [1]	4
ifault	integer [1]	0

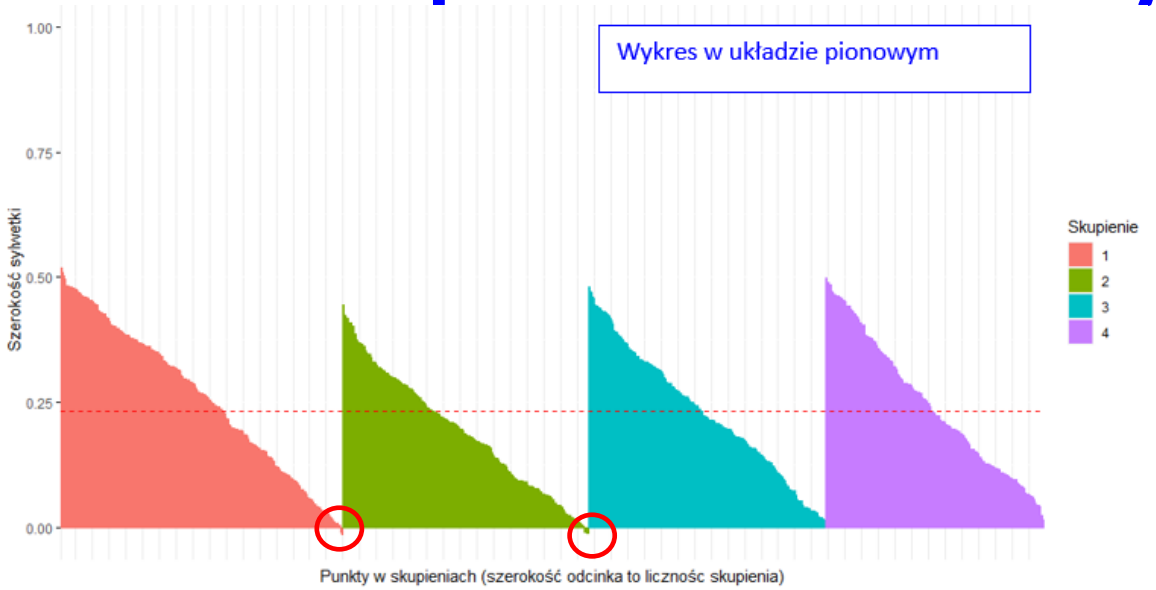
Numer skupienia dla obserwacji (wektor), centra (średnie) skupień (macierz 4x3 – View(km$centers), TSS, WSS (wektor), tot.WSS, BSS, rozmiary (liczebności) skupień

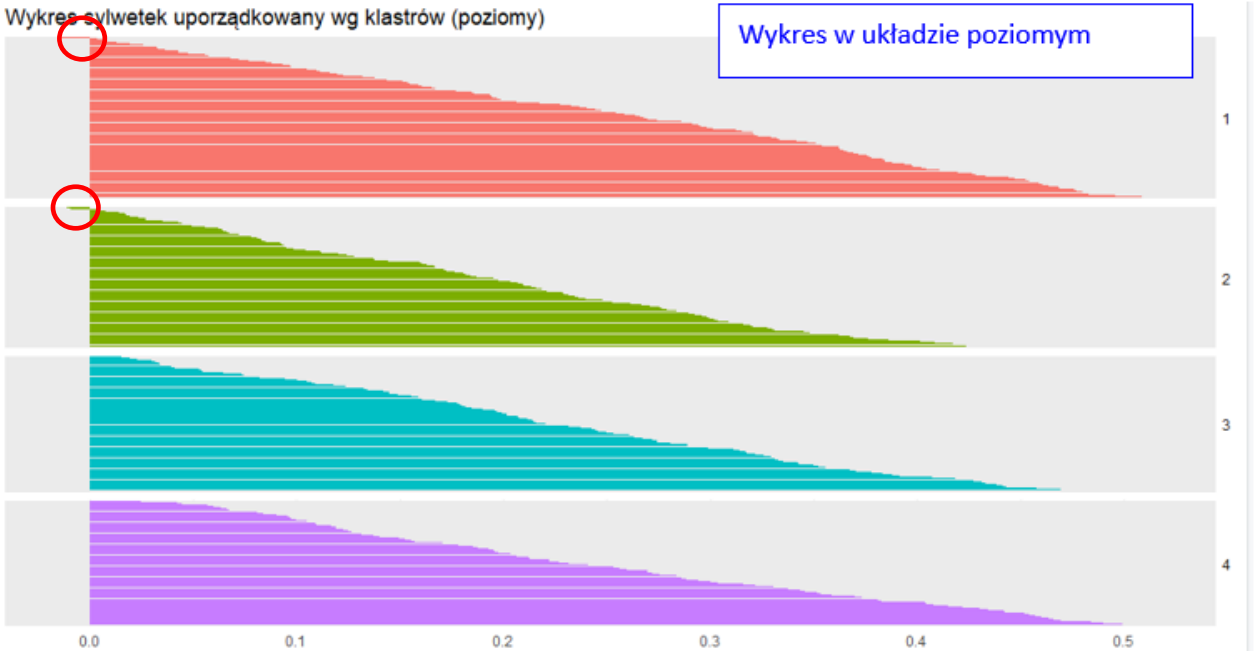
Sylwetki

- Średnia dla całości: 0,232
- Średnie dla poszczególnych skupień

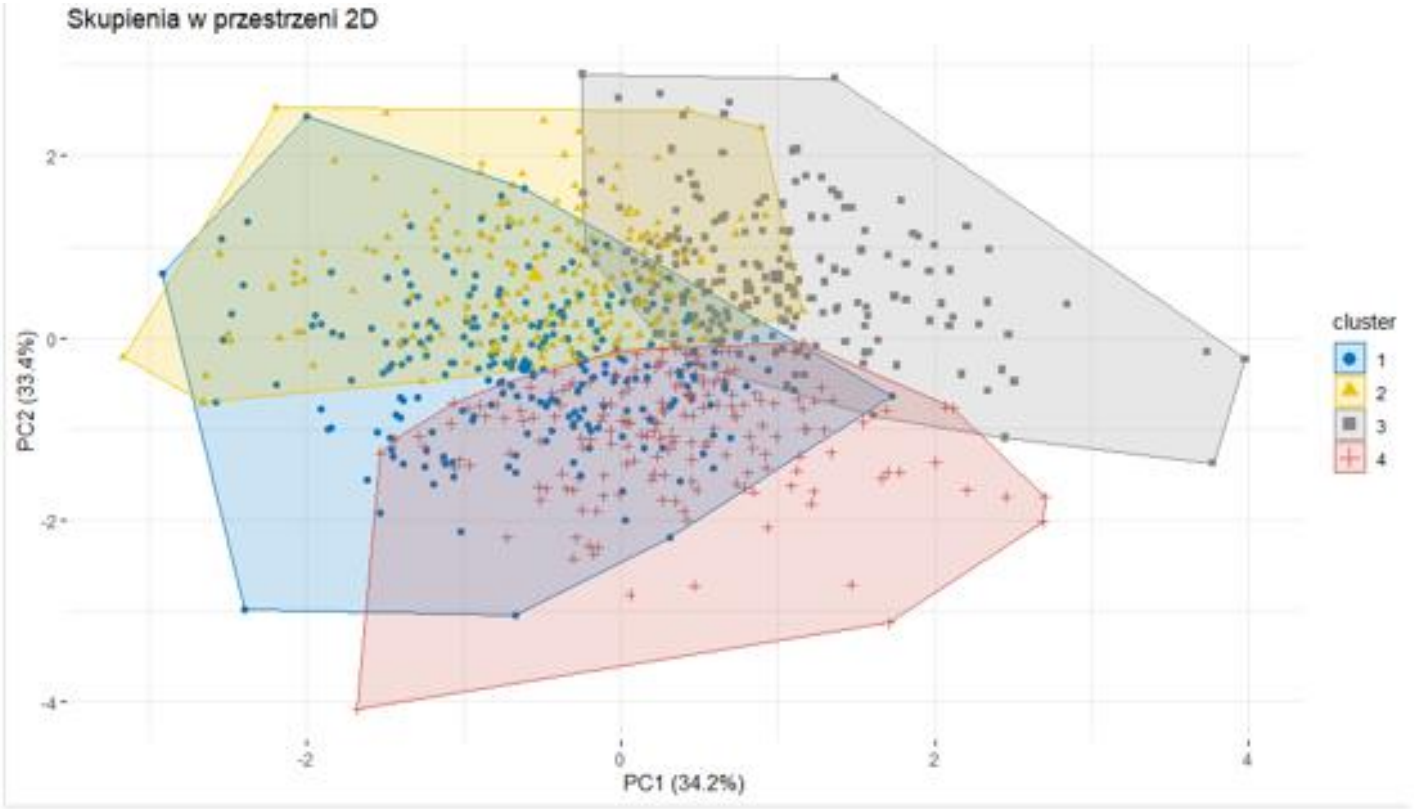
	cluster	x
1	1	0.2583680
2	2	0.1903075
3	3	0.2300519
4	4	0.2458576

Analiza skupień k-means – wykresy sylwetek





Analiza skupień k-means – skupienia w 2D



Wartości procentowe w nawiasach podają procent wariancji wyjaśniony przez pierwsze dwie składowe główne.

Studenci samodzielnie „dopieszczą” wykres

Charakterystyka skupień wg miar statystycznych zmiennych

Statystyka	Trihalomethanes	Turbidity	Chloramines
Skupienie 1			
Średnia	70.80536	3.074343	7.460539
Mediana	69.03511	3.096164	7.404435
Pozostałe miary			
Skupienie 2			
Średnia	73.66444	4.310412	5.588609
Mediana	70.77841	4.247240	5.766780
Pozostałe miary			
Skupienie 3			
Średnia	72.66786	4.590662	8.488176
Mediana	71.98023	4.555205	8.189837
Pozostałe miary			
Skupienie 4			
Średnia	46.28268	3.960373	7.008304
Mediana	48.28535	3.949861	7.046165
Pozostałe miary			

Studenci dopiszą dalsze instrukcje dla wyznaczenia miar statystycznych pozwalających scharakteryzować skupienia.

Rozkład cechy *Potability* w skupieniach

	cluster	Potability	n
	<int>	<int>	<int>
1	1	0	441
2	1	1	273
3	2	0	540
4	2	1	315
5	3	0	582
6	3	1	344
7	4	0	435
8	4	1	346

Studenci zrobią ładny wykres słupkowy (np. w Excelu, R, Pythonie)

Na zakończenie badań

Studenci wykonają analizę wszystkich wyników i wyciągną wnioski