Sprawozdanie 4

Eksploracja danych

Kacper Szmigielski, 282255 i Mateusz Wizner, 277508

2025-06-23

Spis treści

1 Zadanie 1

1.1 a) Wybór i zapoznanie się z danymi

Opis zmiennych w zbiorze danych **Wine**

Kolumna	Nazwa zmiennej	Opis
V1	Alcohol	Zawartość alkoholu (%)
V2	Malic acid	Zawartość kwasu jabłkowego (g/l)
V3	Ash	Zawartość popiołu (g/l)
V4	Alcalinity of ash	Zasadowość popiołu (g/l)
V5	Magnesium	Zawartość magnezu (mg/l)
V6	Total phenols	Zawartość fenoli ogółem (g/l)
V7	Flavanoids	Zawartość flawonoidów (g/l)
V8	Nonflavanoid phenols	Zawartość fenoli nienależących do flawonoidów (g/l)
V9	Proanthocyanins	Zawartość proantocyjaninów (g/l)
V10	Color intensity	Intensywność koloru (od 0 do 13)
V11	Hue	Odcień barwy
V12	OD280/OD315 of diluted wines	Absorbancja przy długości fali 280 nm do 315 nm (rozcieńczone wino)
V13	Proline	Zawartość proliny (mg/l)

2 zadanie 2

2.1 a) Wybór i przygotowanie danych

Do analizy skupień wykorzystano zbiór danych **Glass Identification**, zawierający informacje chemiczne na temat różnych rodzajów szkła. Celem analizy jest identyfikacja naturalnych skupień w danych na podstawie składu chemicznego próbek, bez użycia etykiet klas. Zbiór ten jest często wykorzystywany w badaniach klasyfikacyjnych i klasteryzacyjnych jako benchmark

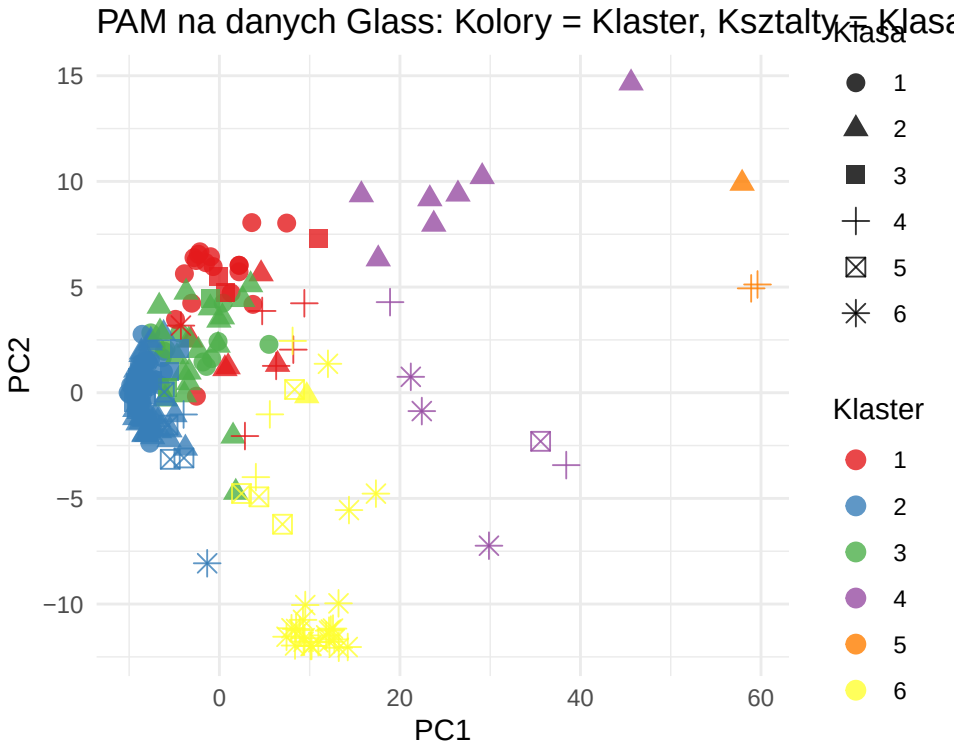
Pełny zbiór zawiera **214 próbek szkła** oraz **9 zmiennych numerycznych**, opisujących zawartość chemicznych pierwiastków (m.in. Na, Mg, Al, Si, Ca). Dodatkowo zawiera zmienną `Type`, określającą rzeczywisty typ szkła (klasa 1–7).

Zmienna `Type` zawiera informację o rodzaju szkła i pełni rolę etykiety klasowej. Ponieważ celem analizy skupień jest znalezienie naturalnych grup bez nadzoru (tzn. bez znajomości klas), zmienna ta została **usunięta przed procesem grupowania**.

Wartości cech w zbiorze różnią się skalą – np. zawartość sodu (Na) czy wapnia (Ca) występuje w innych zakresach niż zawartość żelaza (Fe) czy baru (Ba). Aby zapobiec dominacji zmiennych o większym rozrzucie w macierzy odległości zmienne zostały ustandaryzowane.

2.2 b) Grupowanie i wizualizacja

2.2.1 Grupowanie za pomocą metody PAM



Na podstawie analizy wykresu można stwierdzić, że uzyskane skupienia wykazują **umiarkowany poziom separacji** – **najlepiej odseparowany jest klaster nr 6**, natomiast pozostałe częściowo się nakładają. Sugeruje to, że niektóre obserwacje mogą być trudne do jednoznacznego przypisania do jednej grupy.

Pomimo częściowego pokrywania się skupień, wykazują one **dobrą zwartość** – obiekty należące do tego samego klastra są do siebie **stosunkowo podobne**, co świadczy o spójności wewnętrznej grup.

Z drugiej strony, zaobserwowano **niską jednorodność klas pod względem etykiet rzeczywistych** – obiekty należące do różnych klas (oznaczone różnymi kolorami) **mieszają się wewnątrz tych samych skupień**. Szczególnie wyraźne jest to w przypadku **niebieskiej, zielonej i czerwonej**.

```
## Direct agreement: 1 of 6 pairs
## Iterations for permutation matching: 120
## Cases in matched pairs: 42.06 %
```

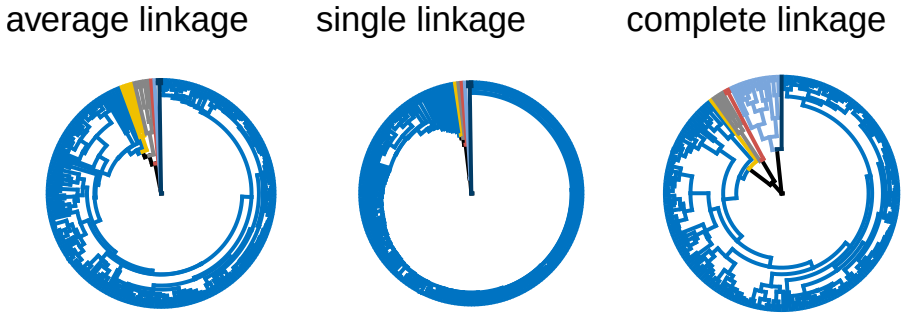
```
## Dokładność przypisania klastrów do klas: 42.06 %
```

W wyniku analizy zgodności przypisań klastrów do klas rzeczywistych, obliczono tzw. **wskaźnik zgodności (purity)**. Niestety, uzyskana wartość wyniosła jedynie **42.06%**, co należy uznać za **niski poziom dopasowania**.

Taki wynik wskazuje, że **grupowanie metodą PAM** nie odzwierciedla w sposób satysfak-

cjonujący rzeczywistej struktury klas w danych. **Zastosowanie tego rodzaju podejścia klasteryzacyjnego** do zbioru *Glass* **nie jest w tym przypadku uzasadnione**, ponieważ prowadzi do znacznego nakładania się klas i nie pozwala na ich skuteczne rozróżnienie.

2.2.2 Podział hierarchiczny



W przypadku metody **single linkage** zaobserwowano wystąpienie tzw. **efektu łańcuchowego** (*chaining effect*). Zjawisko to polega na tym, że kolejne obserwacje są stopniowo dołączane do jednego dużego skupienia na podstawie minimalnych odległości między pojedynczymi punktami, co prowadzi do **tworzenia wydłużonych, sztucznie połączonych struktur**, zamiast wyraźnych, zwartych klastrów.

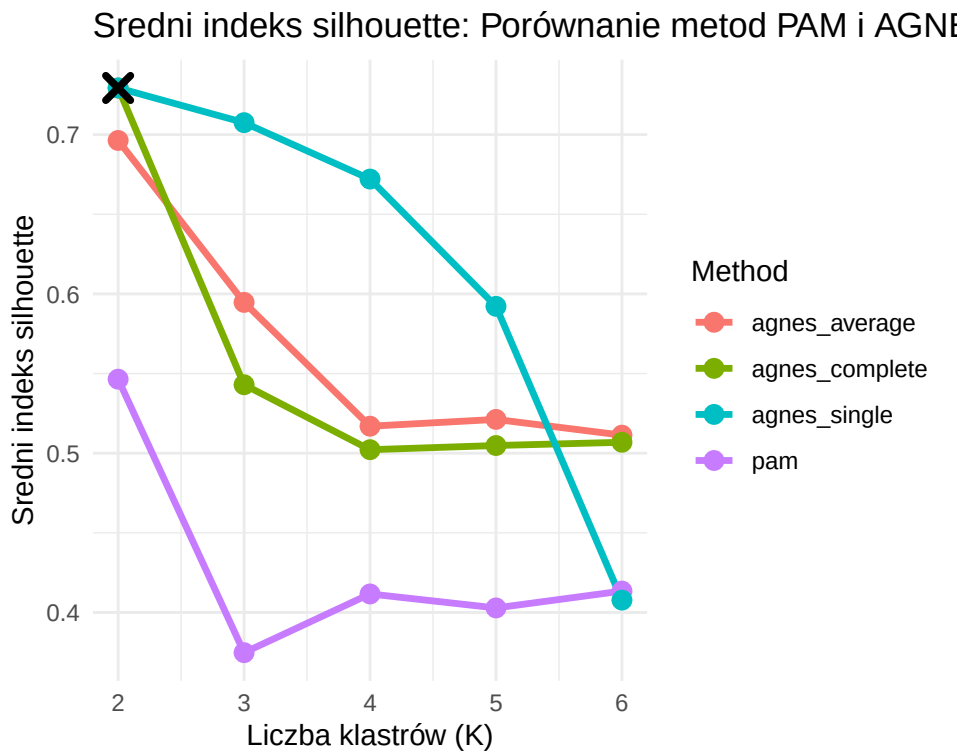
Przyczyną wystąpienia tego efektu w analizowanych danych jest **duży rozrzut obserwacji** oraz **obecność wartości odstających**. Te same czynniki wpłynęły również negatywnie (leczw o wiele mniejszym stopniu) na wyniki uzyskane za pomocą metody **average linkage**, w której efekt łańcuchowy również jest widoczny, choć w nieco łagodniejszej formie.

W przypadku metody **complete linkage**, zjawisko łańcuchowe **również występuje**, jednak jego **intensywność jest znacząco mniejsza**. Klastery są **bardziej zwarte i lepiej odseparowane**, co przekłada się na **większą równowagę w podziale danych** oraz **lepszą zgodność z rzeczywistym podziałem klas**.

2.3 c) Ocena jakości grupowania. Wybór optymalnej liczby skupień i porównanie metod.

2.3.1 Wskaźniki wewnętrzne

W celu dokładniejszego porównania działania poszczególnych algorytmów, ocena została przeprowadzona na oryginalnych (niestandaryzowanych) danych.



Pomimo że zbiór danych Glass zawiera aż 6 rzeczywistych klas, najwyższa średnia wartość współczynnika silhouette została uzyskana dla podziału na 2 klastry. Wskazuje to, że dane te posiadają wyraźniejszą, dwugrupową strukturę wewnętrzną, niezależną od etykiet klas przypisanych z góry. **Współczynnik silhouette** mierzy spójność wewnętrzną klastrów oraz ich separację względem siebie, dlatego może preferować mniejszą liczbę skupień, jeśli podział taki lepiej odzwierciedla naturalne różnice między obserwacjami.

2.3.2 Wskaźniki zewnętrzne

Funkcja `matchClasses()` (z pakietu **e1071**) zakłada, że liczba klastrów w obu porównywanych partycjach (czyli przewidywanych i rzeczywistych etykietach) **jest taka sama.**

Dlatego pomimo uzyskania najlepszego współczynnika sillhouse dla 2 klastróq, wskaźniki zewnętrzne będziemy porównywać dla 7 klastrów

```
## Direct agreement: 0 of 6 pairs
## Iterations for permutation matching: 720
## Cases in matched pairs: 37.85 %
```

```
## Direct agreement: 1 of 6 pairs
## Iterations for permutation matching: 120
## Cases in matched pairs: 36.45 %
```

```
## Direct agreement: 1 of 6 pairs
## Iterations for permutation matching: 120
## Cases in matched pairs: 40.65 %
```

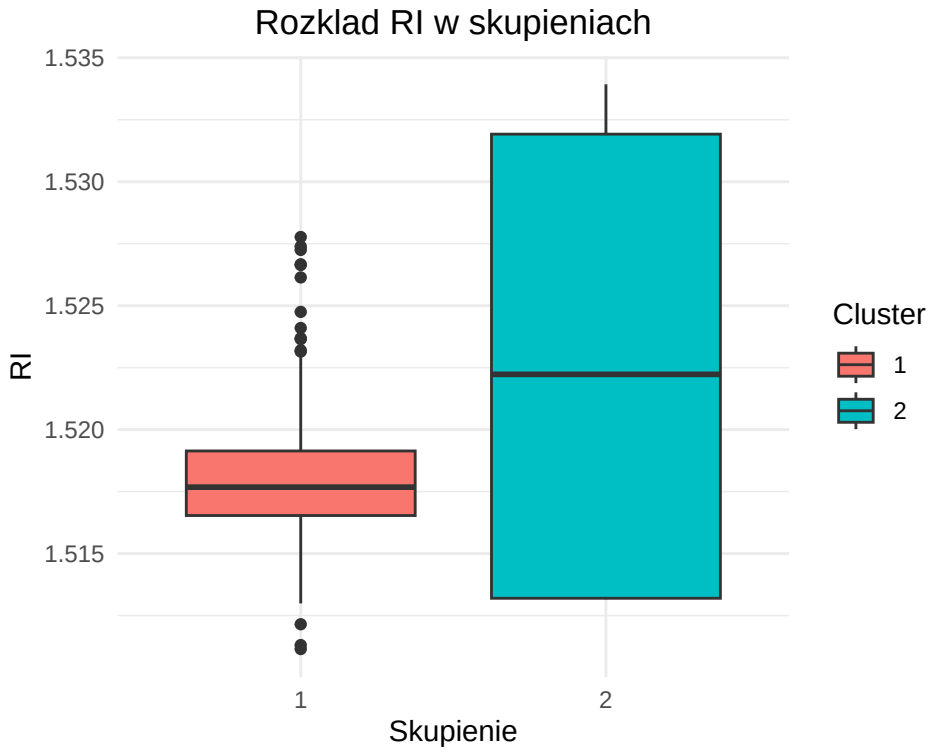
```
## Direct agreement: 1 of 6 pairs
## Iterations for permutation matching: 120
## Cases in matched pairs: 42.06 %
```

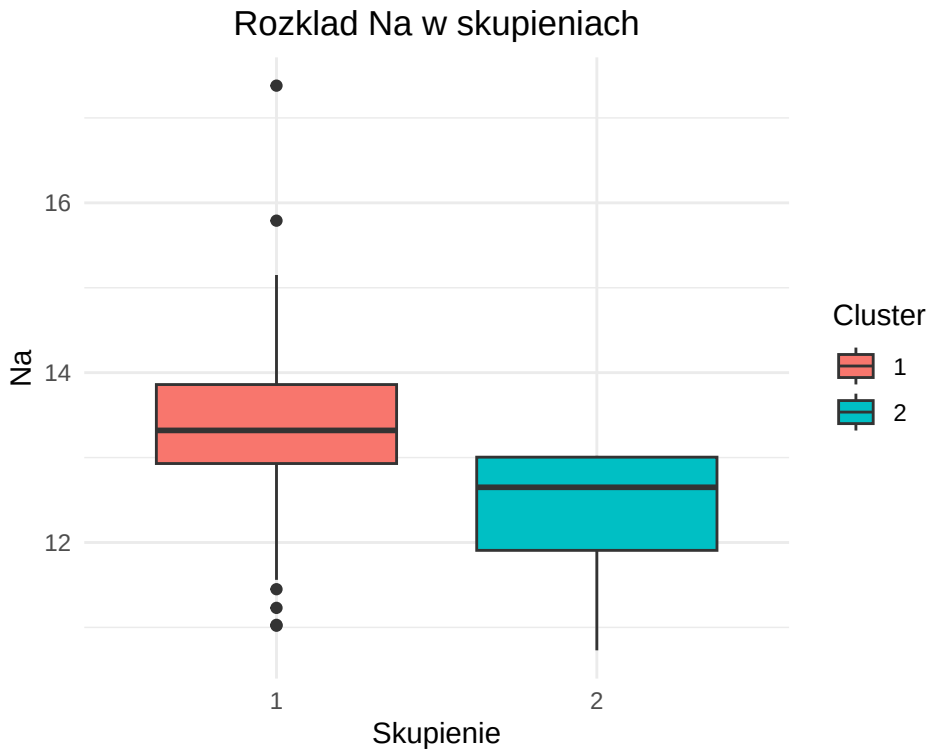
2.4 d) Interpretacja wyników grupowania – charakterystyki skupień

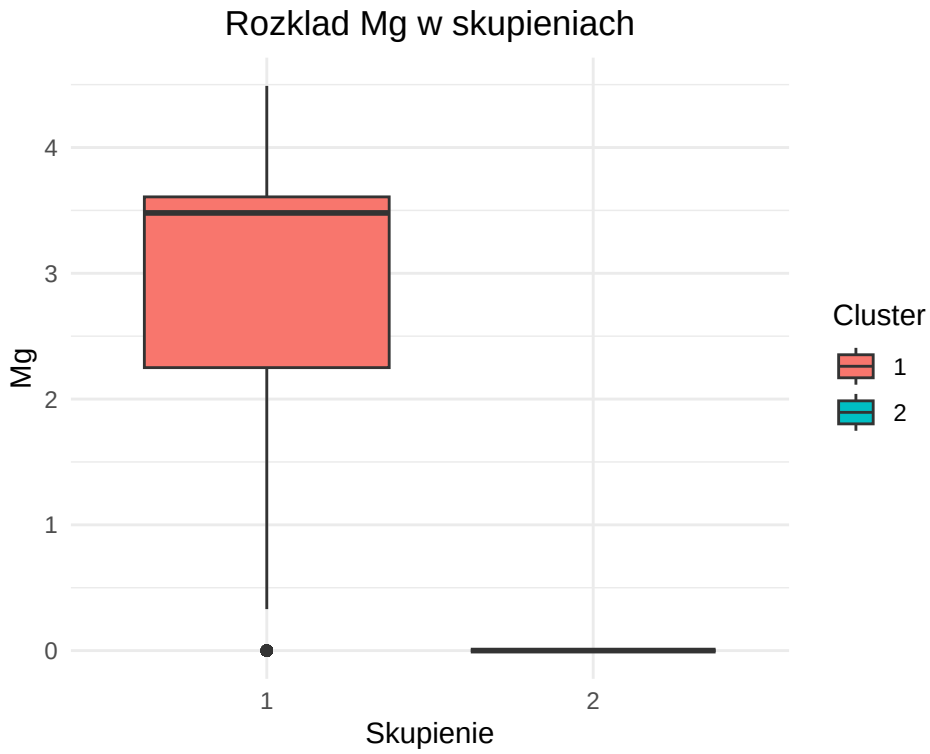
Na podstawie przeprowadzonych analiz, takich jak **współczynnik silhouette** oraz **dokładność dopasowania**, ustalono, że optymalna liczba skupień wynosi **K=2**. Aby lepiej zrozumieć charakterystykę poszczególnych skupień, przeprowadzono porównanie **średnich wartości cech** oraz analizę ich rozkładów za pomocą **wykresów pudełkowych** dla wybranych zmiennych.

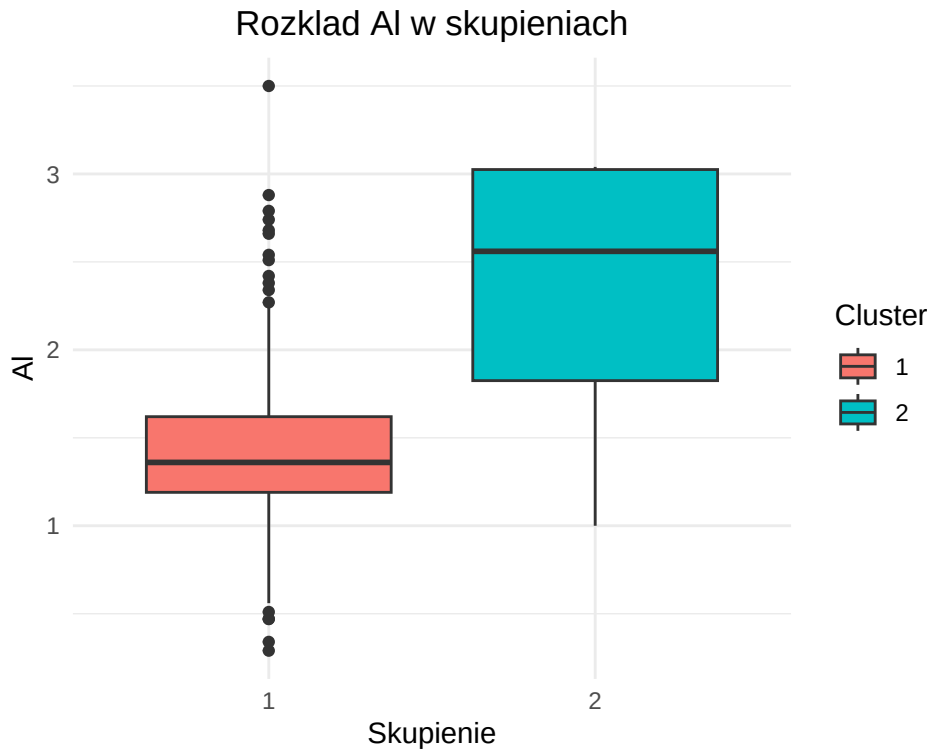
Tabela 2: Średnie wartości cech w skupieniach

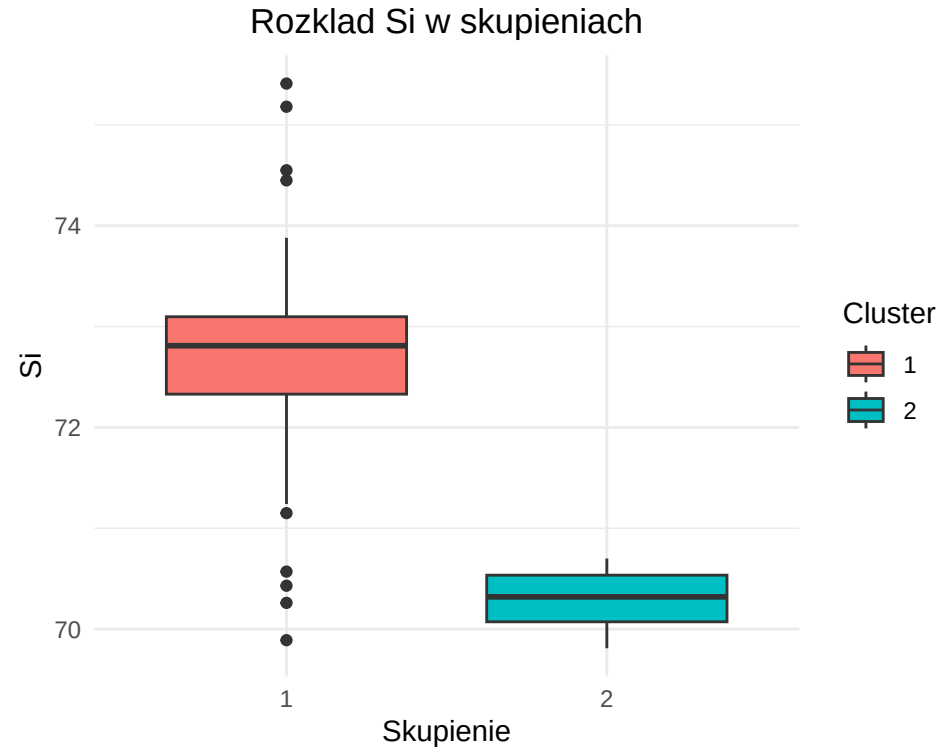
Cluster	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
1	1.52	13.43	2.74	1.43	72.70	0.44	8.92	0.16	0.06
2	1.52	12.26	0.00	2.29	70.29	3.28	10.85	0.79	0.13

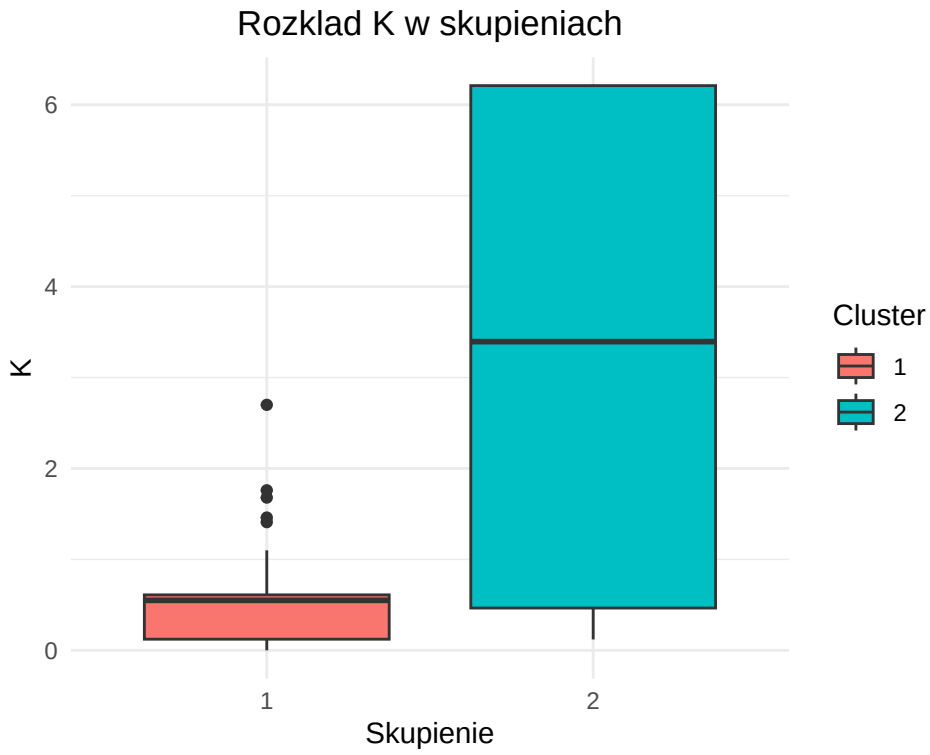


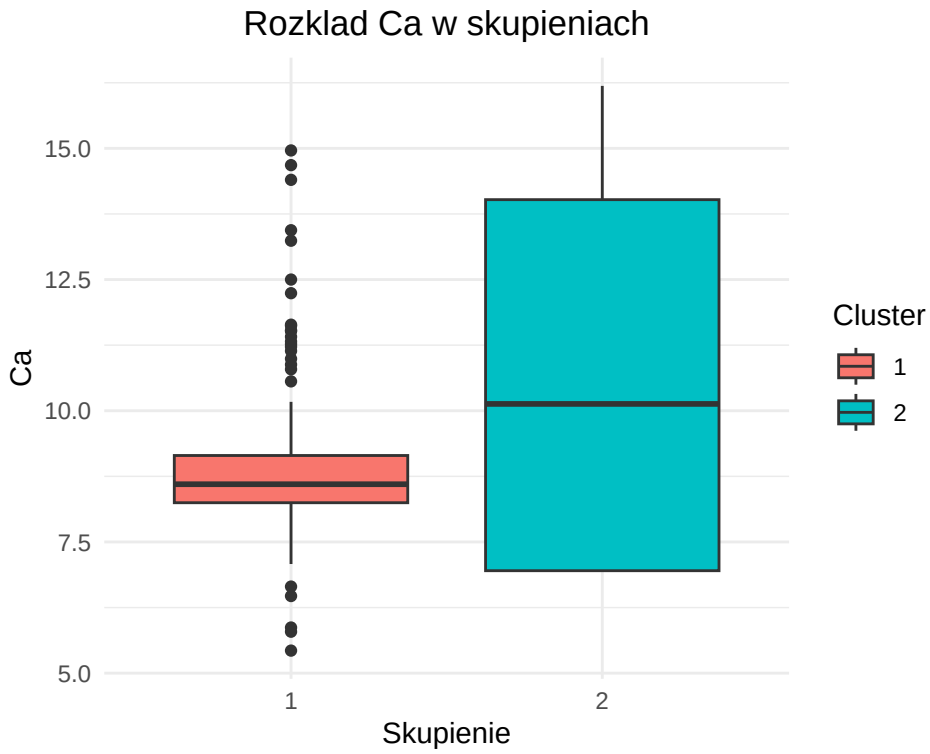

Rozklad Na w skupieniach
Na
16
14
12
1
2
Skupienie
Cluster
1
2

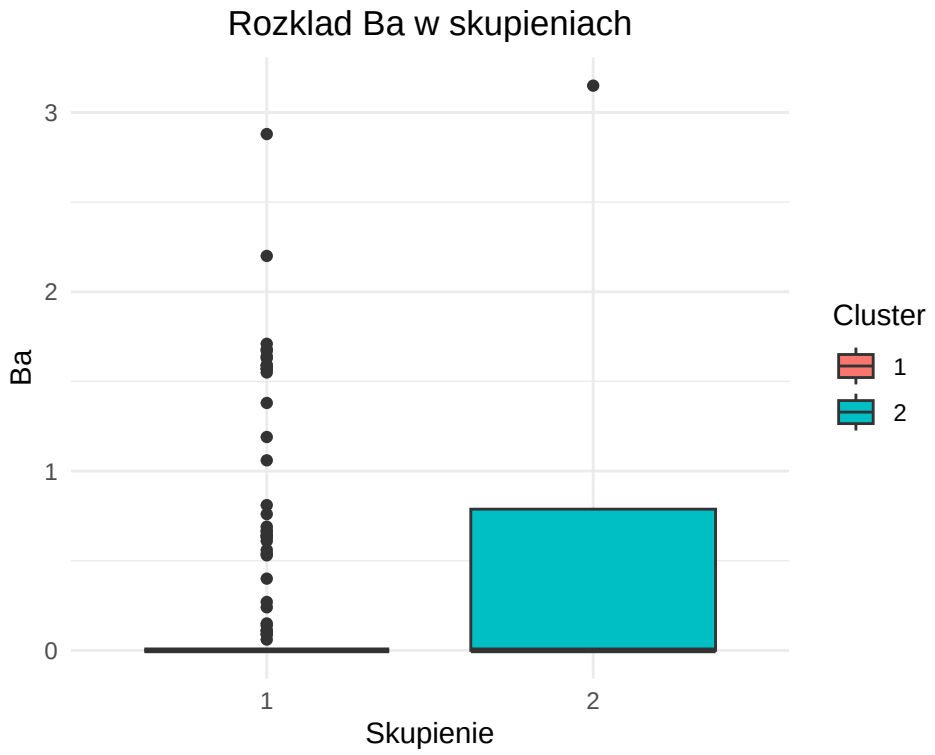

Rozklad Mg w skupieniach
Mg
4
3
2
1
0
1
2
Skupienie
Cluster
1
2

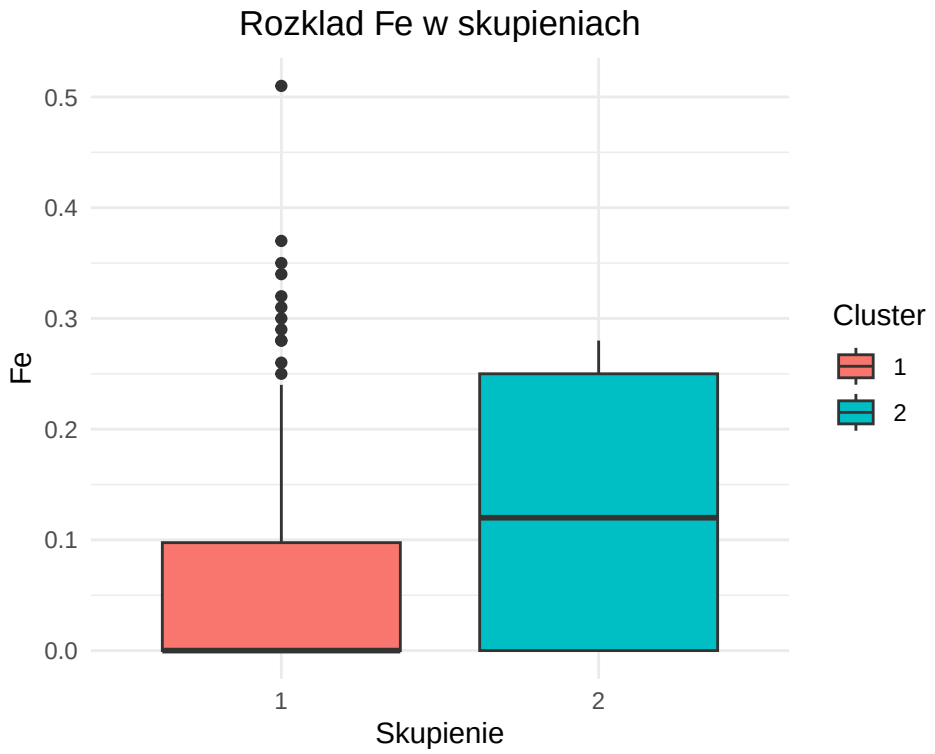

Rozklad Al w skupieniach
Al
3
2
1
1
2
Skupienie
Cluster
1
2


Rozklad Si w skupieniach
Si
74
72
70
1
2
Skupienie
Cluster
1
2


Rozklad K w skupieniach
6
4
2
0
K
1
2
Skupienie
Cluster
1
2


Rozklad Ca w skupieniach
15.0
12.5
10.0
7.5
5.0
Ca
1
2
Skupienie
Cluster
1
2






```
## numeric(0)
```

Jak pokazują **boxploty** oraz **tabela średnich wartości** (na danych bez standaryzacji, dla zachowania ich interpretowalności), **największe różnice między klastrami** dotyczą zmiennych **Ba (bar)** oraz **Mg (magnez)**. W szczególności w **klastrze 1** wartości obu

tych cech są wyraźnie wyższe, przy czym dla **Mg** obserwuje się również istotne **wartości odstające**.

Co ciekawe, **mediana Ba** w **klastrze 2** przewyższa tę z klastra 1, co prowadzi do **prawie idealnej separacji grup** w wymiarze tej zmiennej. Sugeruje to, że **Ba i Mg są kluczowymi czynnikami różnicującymi strukturę klastrów**.

Zbliżone różnice obserwujemy również dla zmiennych **K**, **Ri** i **Al**, gdzie **klaster 2** cechuje się wyższymi wartościami średnimi.

Taki rozkład jest spójny z oczekiwaniami – w kontekście klasyfikacji typu szkła, **zawartość baru i magnezu** to jedne z najistotniejszych parametrów różnicujących próbki, co potwierdzają zarówno analizy statystyczne, jak i wizualne.

Dodatkowo warto przyjrzeć się **medoidom** wyłonionym metodą **PAM (Partitioning Around Medoids)**, by zrozumieć, które obserwacje najlepiej reprezentują klastry oraz jakie cechy je wyróżniają na tle pozostałych. Pozwoli to lepiej uchwycić **typowe profile obserwacji** w każdej z grup.

Tabela 3: Analiza meoidów dla metody PAM

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe
36	1.51567	13.29	3.45	1.21	72.74	0.56	8.57	0.00	0
212	1.52065	14.36	0.00	2.02	73.42	0.00	8.44	1.64	0

Medoid pierwszego skupienia (rekord nr 36) charakteryzuje się wyraźnie podwyższonymi stężeniami magnezu i potasu, przy jednoczesnym obniżeniu poziomów baru i glinu. Natomiast medoid drugiego skupienia wykazuje odwrotną tendencję – wartości magnezu i potasu są niższe, natomiast stężenia baru i glinu wyższe, przy zachowaniu porównywalnych poziomów pozostałych pierwiastków.

Warto podkreślić, że średnie stężenie żelaza (Fe) w obu medoidach wynosi 0. Wskazuje to, że pierwiastek ten najprawdopodobniej występuje jedynie w śladowych ilościach. Nieliczne wyższe wartości można uznać za obserwacje odstające lub wynikające z przypadkowego zanieczyszczenia próbek.