Projekt: Analiza skupień

Magdalena Smalarz

1. Wstęp

Celem projektu jest pogrupowanie ogólnodostępnych piw w Polsce biorąc pod uwagę cechy takie jak: procentowa zawartość alkoholu, cena, dostępność, rozpoznawalność i preferencje. Wykorzystano analizę skupień polegającą na grupowaniu obiektów danego zbioru na jednorodne klasy, czyli wyłonieniu grup obiektów, które są bardziej podobne do obiektów tworzących daną grupę niż do obiektów spoza tej grupy. Fundamentem większości algorytmów dotyczących analizy skupień jest podobieństwo między obserwacjami, które jest wyrażone przy pomocy funkcji podobieństwa. Można scharakteryzować dwie grupy algorytmów klasyfikowania: podziałowe i hierarchiczne. Różnica między nimi polega na tym, że dla grupowania podziałowego z góry znana jest liczba skupień, zaś w grupowaniu hierarchicznym szuka się momentu, w którym najlepiej jest przerwać dalsze grupowanie.

2. Wykorzystane narzędzia

Analizę skupień przeprowadzono w środowisku R z wykorzystaniem pakietu *clusterSim*. Badanie porównuje wyniki grupowania otrzymane z wykorzystaniem algorytmów grupowania podziałowego (k-średnich oraz k-medoidów) oraz hierarchicznego.

2.1. Grupowanie podziałowe

Algorytm k-średnich zaczyna się od przyporządkowania punktów startowych do z góry zadanej liczby skupień. Następnie buduje się centroidę powstałych skupień, a z kolei każdą obserwację przyporządkowuje się do skupienia, do którego centroidy jest najbliżej. W taki sposób budowana jest centroida. Algorytm kończy się gdy punkty nie zmieniają swojego przypisania. Ta metoda ma jednak wiele wad, m.in. początkowe zdefiniowanie liczby grup. Dodatkowo początkowe centroidy wybierane są w sposób losowy, podczas gdy ich wybór ma decydujący wpływ na jakość otrzymanego grupowania.

Algorytm k-medoidów jest odporniejszy na obserwacje odstające od wcześniej opisanej metody. W odróżnieniu od algorytmu k-średnich w kolejnych etapach algorytmu nowe prototypy grup wyznaczane są spośród obiektów należących do rozpatrywanego zbioru. Podobnie jak przy wyżej opisanym algorytmie - na wstępie należy zdefiniować liczbę klas.

2.2. Grupowanie hierarchiczne

W projekcie została wykorzystana metoda aglomeracyjna grupowania hierarchicznego, która rozpoczyna się od stworzenia hierarchii od podziału zbioru n obserwacji na n jednoelementowych grup, które w kolejnych krokach są ze sobą scalane, aż do uzyskania jednego skupienia zawierającego wszystkie obiekty. Metoda ta wykorzystuje funkcje odległości. Najbardziej popularnym sposobem reprezentacji wyników grupowania hierarchicznego jest graficzny zapis w postaci drzewa binarnego nazywanego dendrogramem.

Wykorzystane w projekcie miary odległości:

- **odległość euklidesowa** - klasyczna i najbardziej popularna odległość używana w grupowaniu. Można jej użyć pod warunkiem, że zmienne nie są wzajemnie skorelowane, w innym wypadku efekt grupowania może okazać się bezwartościowy $d(x,y) = \sqrt{\sum_i (x_i - y_i)^2}$,
- **odległość miejska** - jedna z bardziej neutralnych miar, mało czuła na wartości skrajne. Podobnie jak dla miary euklidesowej - użycie tylko w przypadku zmiennych nieskorelowanych $d(x,y) = \sum_i |x_i - y_i|$,
- **odległość Minkowskiego** - uogólniona miara wyżej wymienionych $d(x,y) = (\sum_i |x_i - y_i|^m)^{\frac{1}{n}}$.

Metody klasyfikacji użyte w badaniu: Warda, Centroidalna, Najbliższego i Najdalszego sąsiada.

Metoda Warda należy do grupy algorytmów, które na każdym etapie optymalizują podział otrzymany poprzez połączenie dwóch elementów, stosując kryterium minimalnego wzrostu łącznej wewnątrz-grupowej sumy kwadratów odchyleń wszystkich zmiennych dla każdego obiektu od ich średnich grupowych. Powstające skupienia mają podobną liczbę obiektów oraz nie wykazują tendencji do łączenia się w łańcuchy. Jest uważana za efektywną metodę grupowania.

Do określania odległości między punktami w **Metodzie Centroidalnej** wykorzystuje się punkty ciężkości (centroidy), którymi są punkty o współrzędnych równych średnim wartościom cech wszystkich obiektów w poszczególnych skupieniach. Posiada niekorzystną cechę - w kolejnych krokach występuje problem w łączeniu skupisk o różnej liczności, ponieważ centroid nowej grupy będzie bliski punktowi ciężkości większej z dwóch łączonych grup i może pozostawać w jej obrębie.

Metoda Najbliższego Sąsiada polega na przeliczaniu odległości między obiektami jednego skupienia a obiektami innego skupienia według kryterium najmniejszej odległości. Metoda skupia się na szukaniu wzajemnie izolowanych skupisk jednak nie zwraca uwagi na ich wewnętrzną spójność. Jest metodą zmniejszającą odległości i może zaburzać grupowanie. W praktyce nie jest używaną metodą.

W **Metodzie Najdalszego Sąsiada** nowe odległości powstają jako największa odległość między dowolnymi obiektami należącymi do różnych klas. W przeciwieństwie do Metody Najbliższego Sąsiada skupia się na wewnętrznej spójności grup, jest metodą zwiększającą odległości oraz czułą na obserwacje odstające. Nie jest używana w praktyce.

3. Dane

Wykorzystano dane piwo.csv zawierające informacje o procentowej zawartości alkoholu (*zawartosc.alk*), cenie (*cena*), dostępności (*dostepnosc*), znajomości (*znajomosc*) i preferencji (*preferencje*) 20 marek piw dostępnych w Polsce.

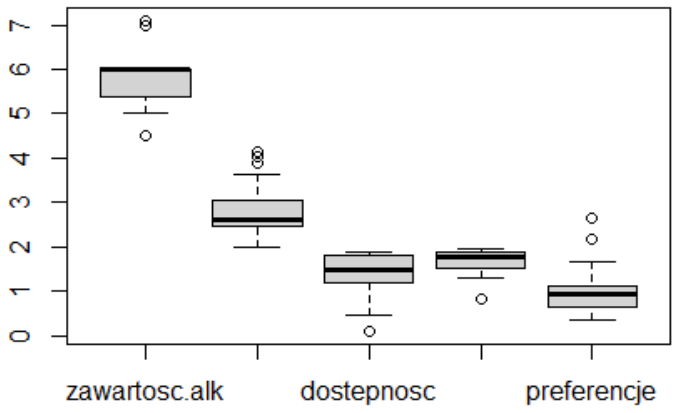
```
##             zawartosc.alk cena dostepnosc znajomosc preferencje
## Zywiec                5.6 2.90       1.90      1.95        2.65
## Desperados            6.0 4.15       1.85      1.77        0.89
## Kasztelan             5.7 2.47       1.80      1.84        2.17
## Wojak                 5.0 2.10       0.45      1.41        0.38
## Tyskie                5.6 2.53       1.90      1.93        1.48
## Heineken              5.0 3.64       1.60      1.90        1.14
## Warka                 5.7 2.59       1.20      1.92        1.06
## Łomża                 6.0 2.60       1.10      1.66        0.80
## Lech                  5.2 2.82       1.80      1.90        1.04
## Perła                 6.0 2.56       1.55      1.79        1.65
## Specjal               6.0 2.36       0.10      0.84        0.43
```

```
## Żubr              6.0 2.49       1.50      1.84         0.83
## Redds             4.5 3.87       1.40      1.69         0.98
## Carlsberg         6.0 3.17       1.85      1.87         0.82
## Somersby          4.5 4.04       1.30      1.59         1.00
## Tatra_Pils        6.0 2.00       1.45      1.30         0.55
## Harnas            6.0 1.99       1.65      1.66         1.02
## Tatra_mocne       7.0 2.62       1.10      1.44         0.35
## Okocim_mocne      7.1 2.71       1.30      1.77         0.76
## Debowe_mocne      7.0 2.89       1.15      1.44         0.34
```

Statystyki opisowe danych:

- *zawartosc.alk* – minimalna zawartość procentowa alkoholu wynosi 4.5%, a największa 7.1%. Przeciętna zawartość alkoholu wynosi 5.79%. Niewielka dodatnia skośność sugeruje, że większość piw ma alkoholu niewiele mniej niż średnia wielkość. Dodatnia wartość kurtozy wskazuje na istnienie wielu wartości bliskiej średniej. Połowa piw miała co najwyżej 6%.
- *cena* – przeciętna cena wynosi 2.825 jednostek i większość cen oscyluje blisko tej wartości. Najdroższe piwo można kupić za 4.150 jednostek, zaś najtańsze za 1.99 jednostek. Większość piw jest w cenie niższej niż 2.825 jednostek, a połowa piw kosztowała co najmniej 2.61 jednostek.
- *dostepnosc* – średnia dostępność piw na poziomie 1.397, maksimum wynosi 1.9, a minimum 0.1. Dostępność większości piw utrzymała się na poziomie niższym niż średnia. Połowa piw posiada dostępność na poziomie co najwyżej 1.475.
- *znajomosc* – średnia znajomość piw na poziomie 1.675, znajomość piw utrzymywała się na poziomie większym niż przeciętny.
- *preferencje* – przeciętnie na poziomie 1.017, a dodatnia wartość kurtozy wskazuje na istnienie wielu wartości bliskiej średniej. większość piw posiada tę wartość na poziomie niższym niż średnia.

Rozkład poszczególnych danych zaprezentowany został na wykresie pudełkowym poniżej.



4. Przebieg badania

4.1. Wybór zmiennych do badania

Poniżej została przedstawiona macierz korelacji wszystkich zmiennych. Żadna z wartości nie jest większa niż |0.9|, zatem można przyjąć, że w zbiorze danych nie występują powtórzenia oraz, że do badania można użyć wszystkich zmiennych.

```
##                zawartosc.alk   cena dostepnosc znajomosc preferencje
## zawartosc.alk          1.000 -0.394     -0.084    -0.192      -0.274
## cena                  -0.394  1.000      0.292     0.285       0.077
## dostepnosc            -0.084  0.292      1.000     0.816       0.612
## znajomosc             -0.192  0.285      0.816     1.000       0.616
## preferencje           -0.274  0.077      0.612     0.616       1.000
```

Ponadto zbadano współczynnik zmienności dla każdej ze zmniennych. Żadna wartość nie jest mniejsza niż 10% - wszystkie zmienne są statystycznie istotne i można wykorzystać je w badaniu.

```
##         zmienna współczynnik zmienności
## 1 zawartosc.alk                0.1252495
## 2          cena                0.2267073
## 3    dostepnosc                0.3387212
## 4     znajomosc                0.1649971
## 5   preferencje                0.5841203
```

4.2. Normalizacja zmiennych

Jako, że zmienne występują w różnych jednostkach należy je ujednolicić aby móc je ze sobą porównywać i dzielić na grupy. W tym celu korzysta się z normalizacji danych, w przypadku badania - standaryzacji zgodnie ze wzorem: $Z_{ij} = \frac{x_{ij} - \overline{x}_{ij}}{s_j}$. Zestandaryzowane dane prezentują się w poniższy sposób.

```
##              zawartosc.alk   cena dostepnosc znajomosc preferencje
## Zywiec              -0.269  0.117      1.062     0.993       2.749
## Desperados           0.282  2.069      0.956     0.342      -0.214
## Kasztelan           -0.131 -0.554      0.850     0.595       1.941
## Wojak               -1.095 -1.132     -2.002    -0.960      -1.072
## Tyskie              -0.269 -0.461      1.062     0.921       0.779
## Heineken            -1.095  1.273      0.428     0.812       0.207
## Warka               -0.131 -0.367     -0.417     0.884       0.072
## Łomża                0.282 -0.351     -0.628    -0.056      -0.365
## Lech                -0.820 -0.008      0.850     0.812       0.039
## Perła                0.282 -0.414      0.322     0.414       1.066
## Specjal              0.282 -0.726     -2.741    -3.022      -0.988
## Żubr                 0.282 -0.523      0.217     0.595      -0.315
## Redds               -1.784  1.632      0.005     0.052      -0.062
## Carlsberg            0.282  0.539      0.956     0.704      -0.332
## Somersby            -1.784  1.897     -0.206    -0.309      -0.029
## Tatra_Pils           0.282 -1.288      0.111    -1.358      -0.786
## Harnas               0.282 -1.304      0.533    -0.056       0.005
```

```
## Tatra_mocne          1.660 -0.320     -0.628    -0.852      -1.123
## Okocim_mocne         1.798 -0.180     -0.206     0.342      -0.433
## Debowe_mocne         1.660  0.101     -0.523    -0.852      -1.140
```

Algorytmy analizy skupień są czułe na istnienie elementów odstających i należy skontrolować ich występowanie w zbiorze danych. Korzystając z reguły trzech sigm ($P(|X| > 3\sigma) = 0.0025$) można uznać obiekt a outlier, gdy jedna z jego zmiennych po standaryzacji przyjmuje wartość większą od |3|. Wynik operacji sugeruje występowanie jednej takiej obserwacji odstającej dla zmiennej *znajomosc* w wierszu *Specjal*, jednak ze względu na jej nieznaną istotność w badaniu oraz możliwość zaburzenia klasyfikacji postanowiono ją zatrzymać.

4.3. Klasyfikacja z wykorzystaniem grupowania podziałowego

Dla obu metod algorytm przeprowadzono na zestandaryzowanych danych, przyjęto podział na 4 grupy.

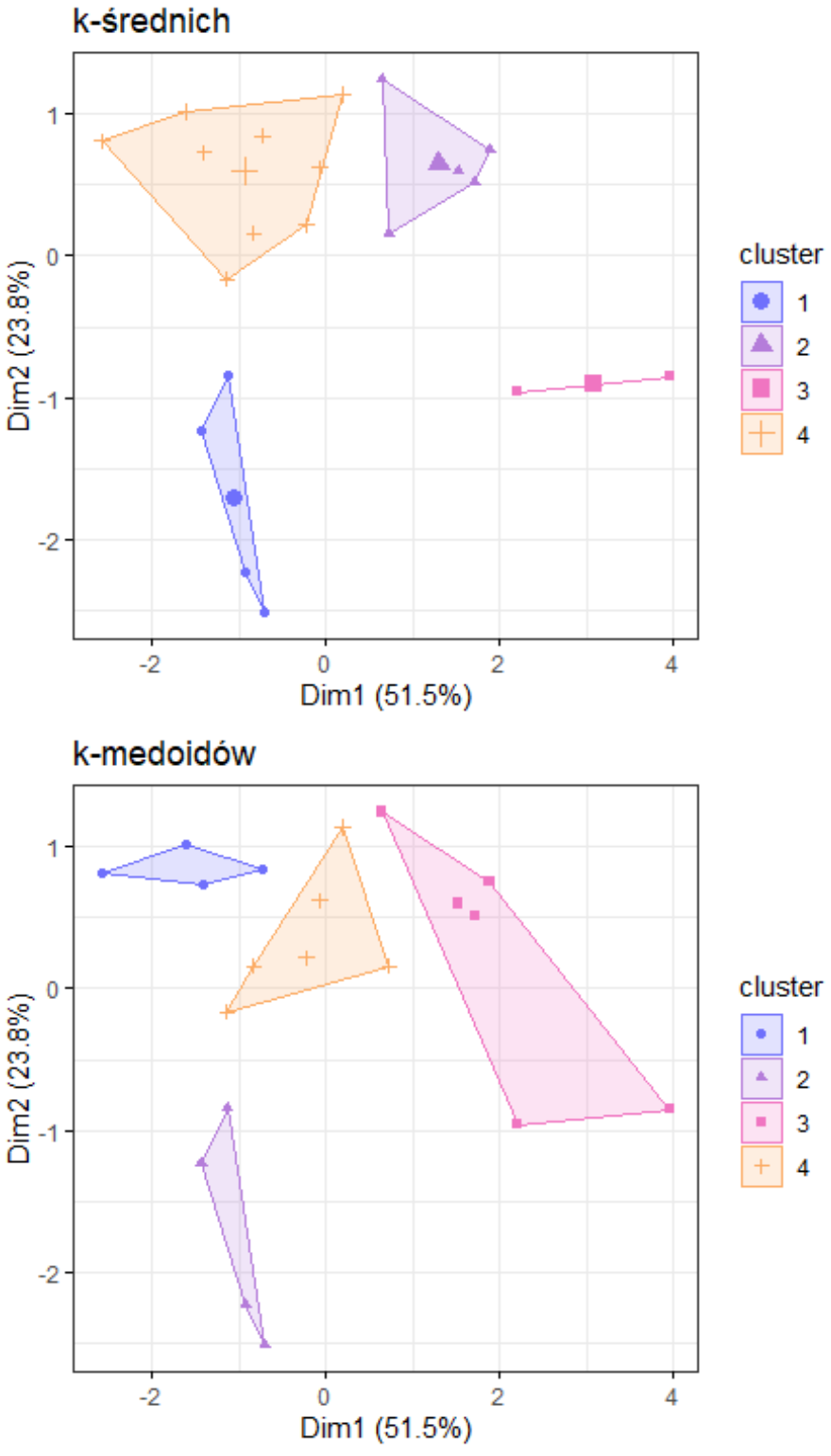
Metoda k-średnich - wykorzystano funkcję *kmeans()* w środowisku R.

Metoda k-medoidów - wykorzystano funkcję *pam()* w środowisku R.

Otrzymane wyniki zebrano w tabeli.

```
##              k-średnich k-medoidów
## Zywiec                4          1
## Desperados            1          2
## Kasztelan             4          1
## Wojak                 3          3
## Tyskie                4          1
## Heineken              1          2
## Warka                 4          4
## Łomża                 2          4
## Lech                  4          4
## Perła                 4          1
## Specjal               3          3
## Żubr                  4          4
## Redds                 1          2
## Carlsberg             4          4
## Somersby              1          2
## Tatra_Pils            2          3
## Harnas                4          4
## Tatra_mocne           2          3
## Okocim_mocne          2          3
## Debowe_mocne          2          3
```

Obserwując powyższą tabelę, można zauważyć, że algorytmy dokonały rozbieżnych klasyfikacji. Przykładowo według algorytmu k-średnich istnieje najmniejsza jedna grupa dwuelementowa z piwami Wojak i Specjal, zaś według algorytmu k-medoidów najmniejsza grupa liczy 4 elementy. Rozbieżność między licznościami grup dobrze obrazują poniższe wykresy, sporządzone z wykorzystaniem pakietów *factoextra* i *ggpubr*.



Poniżej zaprezentowane zostało zestawienie średnich w poszczególnych klasach dla algorytmu k-średnich i k-medoidów. Wartości średnich wokół których tworzone są skupienia w poszczególnych metodach są zestandaryzowane i ciężko poddać je szczegółowej interpretacji. Widać natomiast, że średnie są rozbieżne, a każda z metod pogrupowała piwa w inny sposób.

```
##   zawartosc.alk       cena dostepnosc  znajomosc preferencje
## 1   -1.09531173  1.7175476  0.2957562  0.2242699 -0.02440871
## 2    1.13664424 -0.4075272 -0.3749766 -0.5552488 -0.76929509
## 3   -0.40643643 -0.9290371 -2.3713311 -1.9912996 -1.03021575
## 4   -0.05434461 -0.3304978  0.6038356  0.6513071  0.66717131

##   zawartosc.alk       cena   dostepnosc   znajomosc preferencje
## 1    -0.1308863 -0.5542995  0.850299125  0.59503867  1.94091300
## 2    -1.7841870  1.6316702  0.005281361  0.05245022 -0.06228429
## 3     1.6601895 -0.3200884 -0.628481962 -0.85186387 -1.12280050
## 4     0.2824389 -0.5230713  0.216535802  0.59503867 -0.31478814
```

4.4. Klasyfikacja z wykorzystaniem grupowania hierarchicznego

Aby sklasyfikować piwa według grupowania hierarchicznego należy wybrać jedną z miar odległości oraz zdecydować się na jedną z metod klasyfikacji. W badaniu przeprowadzono analizę kilku rodzajów miar oraz metod w celu wybrania tej najlepszej kombinacji, którą poddano dalszej analizie. Ponadto sprawdzono również jak na grupowanie hierarchiczne wpływa rodzaj normalizacji.

4.4.1. Wyznaczenie miar odległości

W projekcie rozpatrzono trzy miary odległości: euklidesową, miejską i Minkowskiego. Do ich wyznaczenia wykorzystano funkcję *dist()* z pakietu *clusterSim* przypisując jako argument odpowiednią metodę.

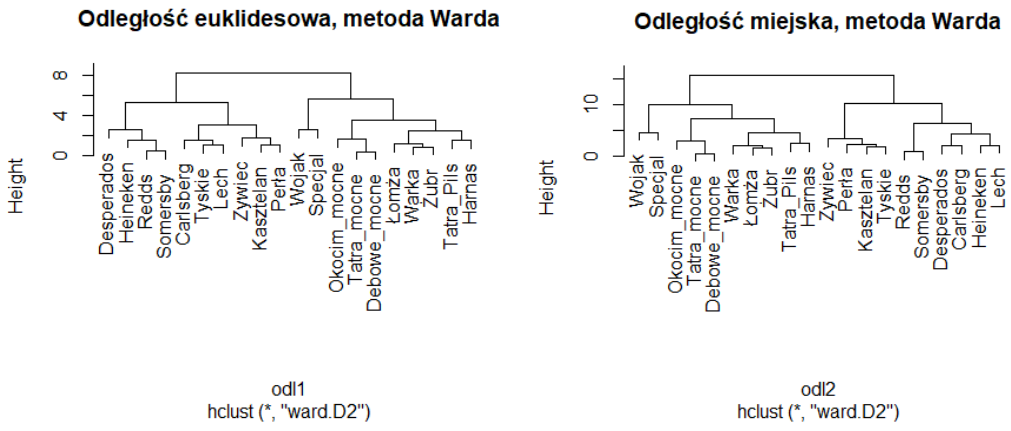
4.4.2. Metody klasyfikacji

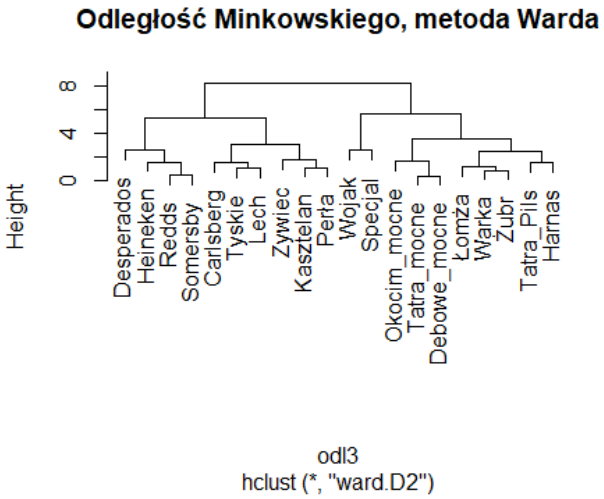
W badaniu postawiono na cztery metody: Warda, Centroidalną, Najbliższego i Najdalszego Sąsiada. Do ich wyznaczenia wykorzystano funkcję *hclust()* z pakietu *clusterSim* przypisując jako argument odpowiednią metodę.

4.4.3. Sporządzenie dendogramów

Posiadając obliczone odległości oraz wybrane metody klasyfikacji można przejść do sporządzenia dendogramów.

Zestawienie dendogramów dla metody klasyfikowania Warda przy zmianie funkcji odległości.

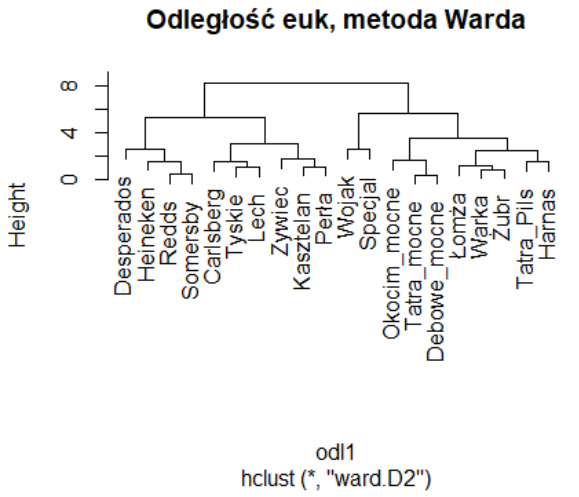


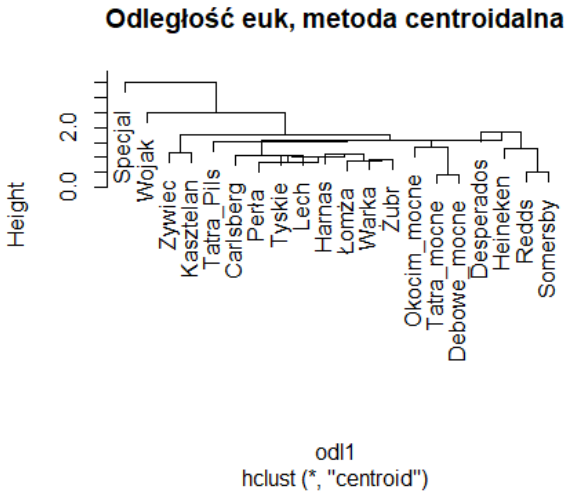


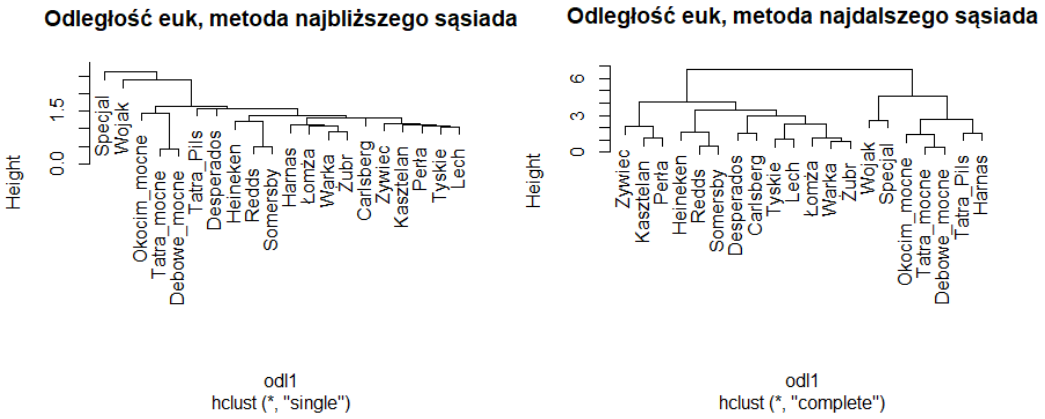
Dendogramy wykorzystujące odległość euklidesową i Minkowskiego niewiele się różnią. Przy odległości miejskiej widać, że drzewo jest wyższe, a kolejność tworzenia się grup inna niż na dwóch pozostałych.

Ze względu na popularność stosowania odległości euklidesowej, wykorzystano ją do sporządzenia finalnego grupowania hierarchicznego i interpretacji.

Zestawienie dendogramów dla euklidesowej miary odległości przy zmianie metody klasyfikowania.







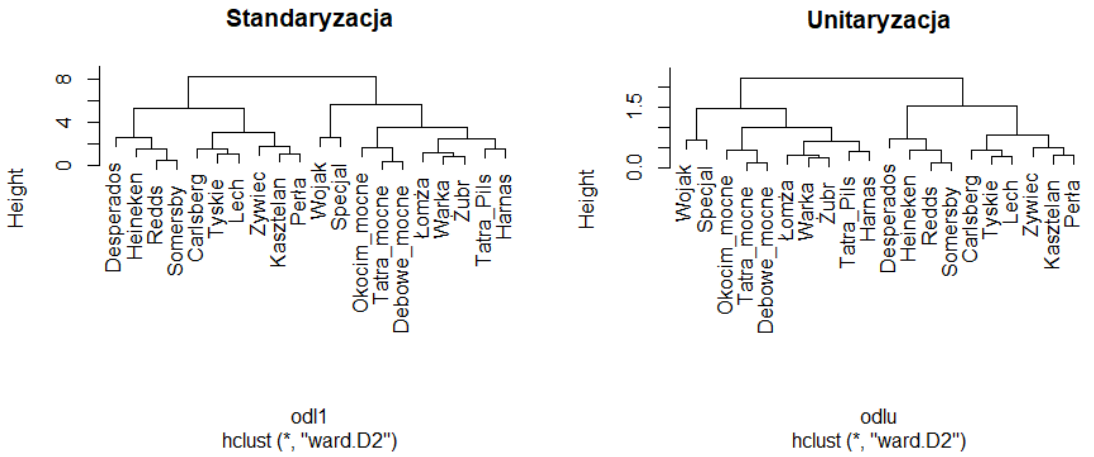
Dendrogramy otrzymane metodą Warda i Najdalszego Sąsiada są podobne. Różna konfiguracja polega na innej kolejności łączenia skupień. Na tych dendogramach odległości między kolejnymi podziałami są największe, co jest pożądanym efektem, ponieważ im wyższa wysokość połączenia, tym mniej podobne są do siebie obiekty.

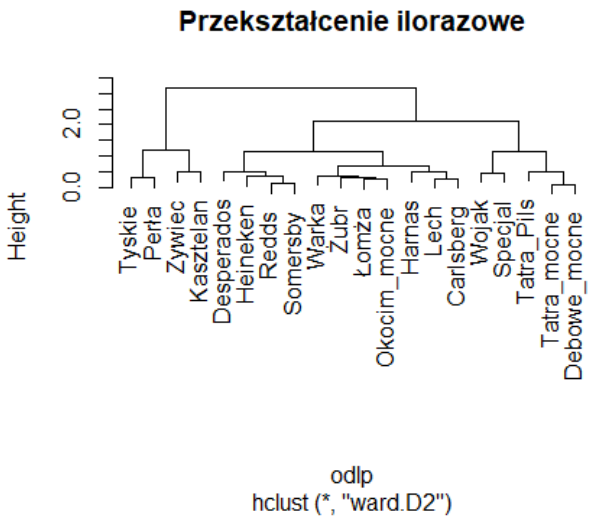
Dendogram metody centroidalnej widocznie różni się od innych. Można zauważyć, że powstało wiele, małolicznych grup. Często zdarza się, że w grupie znajduje się tylko jedno piwo. Występuje mała wysokość połączeń.

Drzewko połączeń dla metody najbliższego sąsiada jest najniższe ze wszystkich, ponieważ jest to metoda zmniejszająca odległości i ze względu na pomijanie wewnętrznej spójności obiektów może łączyć ze sobą obiekty zupełnie niepodobne, co fałszuje grupowanie. Na przeciwnej zasadzie działa Metoda Najdalszego Sąsiada, która zwiększa odległości. Obie metody mogą przekłamywać wynik (zniekształcać odległości). Z tego względu metody te nie będą poddawane dalszej analizie.

Podsumowując, Metoda Warda wskazała najbardziej optymalne grupowanie. Metoda ta jest uważana za najbardziej efektywną dlatego wybrano ją do sporządzenia finalnej klasyfikacji hierarchicznej i interpretacji.

Zastawienie dendogramów dla metody klasyfikowania Warda oraz euklidesowej funkcji odległości przy zmianie metody normalizacji.





Wykorzystanie unitaryzacji i przekształcenia ilorazowego do transformacji danych skraca drzewo - występują mniejsze wysokości połączeń. Można zauważyć, że dendogram z użyciem unitaryzacji i standaryzacji jest bardzo podobny, a różni je kolejność łączenia skupień.

4.4.4. Określenie liczby skupień

Optymalna kombinacja (euklidesowa funkcja odległości, metoda klasyfikacji Warda) została poddana dalszej analizie. W grupowaniu hierarchicznym należy wskazać liczbę klas przy której najlepiej jest przerwać dalsze grupowanie. W tym celu wykorzystano funkcję *index.S* pakietu *clusterSim*, która wyznacza indeks silhouette. Maksymalna wartość indeksu wskazuje optymalną liczbę skupień dla danego zbioru. Ze względu na występowanie w danych 20 rodzajów piw, zdecydowano się obliczyć indeks dla maksymalnie 8 klas.

```
##   Wskaźnik silhouette
## 1                 NaN
## 2           0.2719817
## 3           0.2600778
## 4           0.3005263
## 5           0.2893764
## 6           0.2558589
## 7           0.2360858
## 8           0.2401425
```

Największą wartość otrzymano dla liczby 4 - jest to optymalna liczba grup na jakie można podzielić posiadany zbiór danych.

4.4.5. Klasyfikacja i interpretacja wyników

Finalnie przeprowadzono grupowanie hierarchiczne zestandaryzowanych danych na 4 klasy z wykorzystaniem euklidesowej funkcji odległości oraz metody klasyfikacji Warda. W tym celu wykorzystano funkcję *cutree*. Otrzymano następujące wyniki.

```
##              klasa
## Zywiec           1
## Desperados       2
## Kasztelan        1
## Wojak            3
## Tyskie           1
## Heineken         2
## Warka            4
## Łomża            4
## Lech             1
## Perła            1
## Specjal          3
## Żubr             4
## Redds            2
## Carlsberg        1
## Somersby         2
## Tatra_Pils       4
## Harnas           4
## Tatra_mocne      4
## Okocim_mocne     4
## Debowe_mocne     4
```

Poniżej zostały zaprezentowane kolejno średnie oraz odchylenia standardowe argumentów w poszczególnych klasach przy użyciu funkcji *cluster.Description*.

```
##   zawartosc.alk cena     dostepnosc znajomosc preferencje
## 1 "5.6833"      "2.7417" "1.8"      "1.88"    "1.635"
## 2 "5"           "3.925"  "1.5375"   "1.7375"  "1.0025"
## 3 "5.5"         "2.23"   "0.275"    "1.125"   "0.405"
## 4 "6.35"        "2.4862" "1.3062"   "1.6287"  "0.7138"
```

```
##   zawartosc.alk cena     dostepnosc znajomosc preferencje
## 1 "0.2994"      "0.2707" "0.1304"   "0.0593"  "0.6869"
## 2 "0.7071"      "0.2222" "0.2428"   "0.131"   "0.1034"
## 3 "0.7071"      "0.1838" "0.2475"   "0.4031"  "0.0354"
## 4 "0.5757"      "0.3246" "0.206"    "0.2174"  "0.2766"
```

W pierwszej klasie znajdują się piwa o zawartości alkoholu ok. 5.6833% z odchyleniem standardowym 0.2994. Przeciętnie ich cena oscyluje wokół wartości 2.7417 jednostki. Dostępność na poziomie 1.8, znajomość 1.88, preferencje 1.635. Do tej kategorii należy pięć piw: Żywiec, Kasztelan, Tyskie, Lech, Perła oraz Carlsberg. Piwa te łączy wysoka dostępność i rozpoznawalność.

Grupa druga skupia piwa najdroższe, o średniej cenie 3.925 jednostek. Są nimi: Desperados, Heineken, Redds i Somersby.

Trzecia grupa skupia jedynie dwa piwa, które łączy najmniejsza dostępność na poziomie około 0.275. Są nimi Wojak i Specjal.

Najliczniejsza jest grupa czwarta, składająca się z ośmiu piw (Dębowe mocne, Okocim mocne, Tatra mocne, Harnaś, Tatra Pils, Żubr, Łomża i Warka), w której zawartość alkoholu skupia się wokół 6.35% z odchyleniem 0.5757% i średnią ceną 2.4863 jednostki. Piwa te łączy wysoka zawartość alkoholu.

5. Podsumowanie

```
##              hierarchiczna k-średnich k-medoidów
## Zywiec                   1          4          1
## Desperados               2          1          2
## Kasztelan                1          4          1
## Wojak                    3          3          3
## Tyskie                   1          4          1
## Heineken                 2          1          2
## Warka                    4          4          4
## Łomża                    4          2          4
## Lech                     1          4          4
## Perła                    1          4          1
## Specjal                  3          3          3
## Żubr                     4          4          4
## Redds                    2          1          2
## Carlsberg                1          4          4
## Somersby                 2          1          2
## Tatra_Pils               4          2          3
## Harnas                   4          4          4
## Tatra_mocne              4          2          3
## Okocim_mocne             4          2          3
## Debowe_mocne             4          2          3
```

Każda z metod zwróciła inne pogrupowanie, jednak można zauważyć pewne podobieństwa. Piwa: Desperados, Heineken, Redds i Somersby są do siebie podobne, ponieważ zostały złączone w jedną grupę. Piwa Żywiec, Kasztelan, Tyskie i Perła zawsze znajdują się w jednej grupie, a zmienia się jedynie ich towarzystwo w grupie. Podobnie jest z piwami Wojak i Specjal. Grupowanie hierarchiczne i k-średnich mimo różnej numeracji, zwróciły bardzo podobne grupowanie z różnicą przy piwie Warka. Otrzymanie rozbieżnych wyników w grupowaniu podziałowym wynika z wad jakie posiadają te algorytmy - są dość losowe oraz liczba skupień, która jest nadana "sztucznie" może nie być jednakowa dla obu algorytmów.

Literatura

1. A. Balicki, *„Statystyczna analiza wielowymiarowa"*,
2. A. Nowak-Brzezińska, *„Analiza skupień. Konspekt do zajęć: Statystyczne metody analizy danych"*.