Uniwersytet Rzeszowski

KOLEGIUM NAUK PRZYRODNICZYCH

KIERUNEK: INFORMATYKA

Klasteryzacja danych GPW z wykorzystaniem języka R oraz biblioteki Shiny

Inż. Krzysztof Małysa

Rzeszów 2020

1. **Język R oraz R-Studio**

R (R Project for Statistical Computing) jest jednocześnie językiem programowania, środowiskiem obliczeniowym oraz graficznym. Celem twórców było stworzenie platformy do obliczeń statystycznych, służącej do prezentowania danych w nowy sposób, oraz tworzenia ciekawych wizualizacji np. w postaci wykresów 3D.

R jest zatem wykorzystywany w dziedzinach jak analiza danych i statystyka. W tych dziedzinach mocno rywalizuje z Pythonem, jednak nie zapowiada się, by go doścignął. Python to też język ogólnego zastosowania, dlatego jest częstszym wyborem programistów. R w głównej mierze ogranicza się do wąskiej dziedziny, jaką jest *data science*.

R może też być wykorzystywany do *machine learning*, jednak tu Python jest dużo częstszym wyborem.

R jest również podstawowym językiem programowania w bioinformatyce i biostatystyce.

RStudio jest narzędziem ułatwiającym pracę z R. Jest to edytor, manager wersji, narzędzie wspierające debugowanie, tworzenie pakietów, aplikacji czy raportów. Można żyć bez tej nakładki, ale co to za życie.

2. **Shiny**

Aplikacje Shiny to strony internetowe tworzone i zasilane przez działającą na serwerze aplikację R. Użytkownicy aplikacji mogą poprzez stronę HTML wybierać parametry przetwarzania, przetwarzanie danych i parametrów ma miejsce na serwerze a jego wynik jest przedstawiany ponownie na stronie internetowej.
Model budowy aplikacji z użyciem shiny jest zgodny z modelem *akcja-reakcja* (ang. *reactive programming*). Jest to popularny model budowy aplikacji wokół pracy z danymi, znany np. z arkuszy kalkulacyjnych w których część komórek arkusza może zależeć od innych komórek. Zmieniając stan komórek wejściowych automatycznie zmienia się stan komórek zależnych (mówimy, że wartości są ,,odświeżone"). Kontrolując, które komórki zależą od których, możemy ,,odświeżyć" (czyli przeliczyć na nowo wartości) tylko tych komórek, które mogą się zmienić.
Podobnie będzie w aplikacji zbudowanej z biblioteką shiny, tyle że struktury zależności pomiędzy wejściem a wyjściem nie musimy jawnie opisywać, będzie ona odtworzona przez samą bibliotekę. Takie ,,inteligentne odświeżanie" tylko tych wartości, które

mogą się zmienić pozwala na budowę szybkich i jednocześnie złożonych aplikacji. Wszystkie elementy aplikacji podzielone są na elementy wejściowe i wyjściowe. Model *akcja-reakcja* wymaga oprogramowania sposobu w jaki elementy wejściowe wpływają na wygląd/stan/wartość elementów wyjściowych. Jeżeli w trakcie pracy zmieniona zostanie wartość jakiegoś elementu wejściowego (kliknięty przycisk, wpisana wartość liczbowa, przesunięty suwak) to przeliczone zostaną odpowiednie elementy wyjściowe.

3. Dane

Dane wykorzystane do eksperymentu w aplikacji zostały pobrane ze strony https://stooq.pl/. Pochodzą one z giełdy, a konkretnie dotyczą Giełdy Papierów Wartościowych(GPW). W danych mamy takie pola jak: data, otwarcie, zamknięcie, najwyższy, najniższy oraz wolumen (jest to ilość akcji, która zmieniła właściciela w danym dniu od rozpoczęcia sesji). Zakres danych obejmuje daty 9.10.2019 – 10.01.2020.

	Data	Otwarcie	Najwyzszy	Najnizszy	Zamkniecie	Wolumen
1	2019-10-09	38	38.4	37.2	37.4	45561
2	2019-10-10	37.4	37.7	36.55	37	27235
3	2019-10-11	37.2	37.65	36.8	37	33118
4	2019-10-14	37.05	37.65	37	37.5	43695
5	2019-10-15	37.5	37.6	37.1	37.4	25372
6	2019-10-16	37.4	37.85	37.2	37.5	34812
7	2019-10-17	37.5	38	37.25	37.7	150984
8	2019-10-18	37.7	37.95	37.1	37.6	292758
9	2019-10-21	37.65	38	37.35	37.65	17619
10	2019-10-22	37.85	37.85	37.4	37.65	44348

Showing 1 to 10 of 60 entries Previous 1 2 3 4 5 6 Next

4. Klasteryzacja

Analiza skupień inaczej zwana również **analizą klasterową** (cluster analysis) ma na celu pogrupowanie badanych elementów w podobne do siebie grupy.

Ideą **analizy skupień** jest takie pogrupowanie badanych osób, aby wedle wyznaczonych kryteriów wyodrębnić podobne do siebie jednostki w oddzielne grupy. Stosowana jest tutaj zasada podobieństwa wewnętrznego i niepodobieństwa zewnętrznego. Innymi słowy, grupowanie polega na takim przyporządkowaniu obiektów do grup, aby wewnątrz każdej z wydzielonych grup jednostki w niej znajdujące się były podobne do siebie, ale różne wyodrębnione grupy były jak najmniej podobne do siebie.

Wybrane cele dokonywania grupowania są następujące:

- uzyskanie jednorodnych przedmiotów badania, ułatwiających wyodrębnienie ich zasadniczych cech,
- zredukowanie dużej liczby danych pierwotnych do kilku podstawowych kategorii, które mogą być traktowane jako przedmioty dalszej analizy,
- zmniejszenie nakładu pracy i czasu analiz, których przedmiotem będzie uzyskanie klasyfikacji obiektów typowych,
- odkrycie nieznanej struktury analizowanych danych,
- porównywanie obiektów wielocechowych.

5. Wybrane metody

Aplikacja webowa wykorzystuje następujące metody:

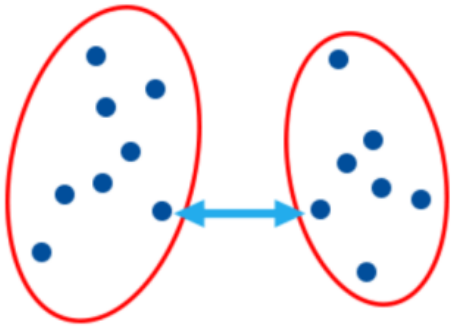
- Metody hierarchiczne:
  Używając klasteryzacji hierarchicznej nie zakładamy z góry ilości klastrów, na jakie chcemy podzielić dane. Wychodzimy od sytuacji, gdy mamy n klastrów, czyli każda obserwacja jest oddzielną grupą. W każdym kroku algorytmu łączymy 2 klastry, czyli zmniejszamy ich liczbę o jeden i tak aż do połączenia wszystkich obserwacji w jedną grupę. Wybór ilości klastrów opieramy na wykresie separowalności, która obliczana jest dla każdego kroku algorytmu.

  - Miara odległości euklidesowa

    $$\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$$

    Do wyliczania odległości euklidesowej została zastosowana metoda pojedynczego połączenia (Odległość między dwoma klastrami jest minimalną

odległością między obserwacją w jednym klastrze a obserwacją w innym klastrze):

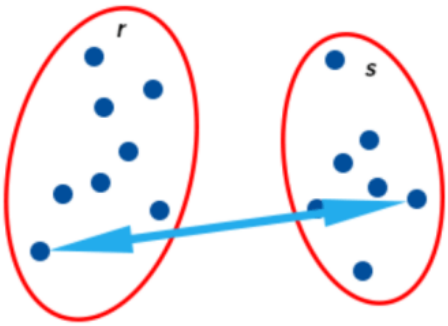


$L(r,s) = min(D(x_{ri}, x_{sj}))$

- Miara odległości Manhattan

$$\|a-b\|_2^2 = \sum_i (a_i - b_i)^2$$

Do wyliczania miary odległości Manhattan została użyta metoda kompletnego połączenia (Odległość między dwoma klastrami jest maksymalną odległością między obserwacją w jednym klastrze a obserwacją w innym klastrze):



$L(r,s) = max(D(x_{ri}, x_{sj}))$

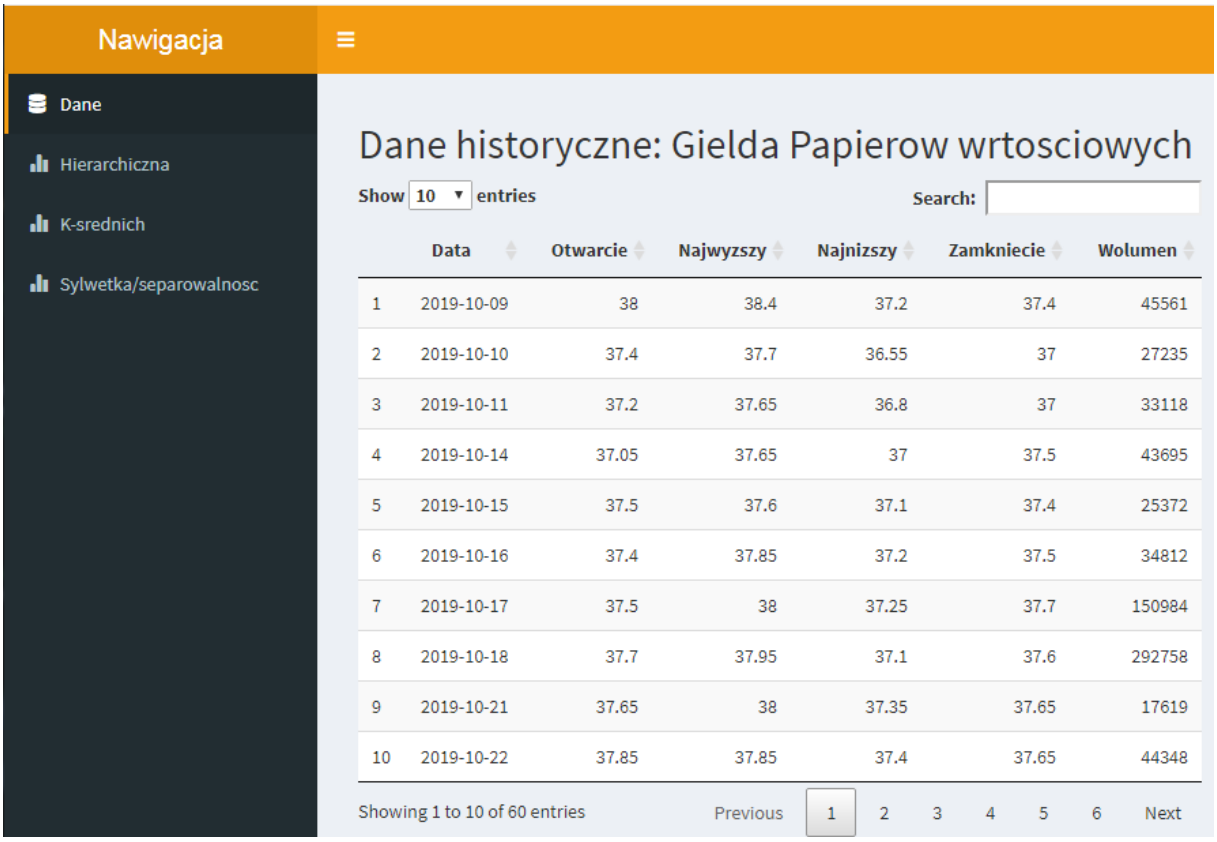
- Metoda k-średnich - wpisująca się w katalog algorytmów niehierarchicznych, w której grupowanie polega na wstępnym podzieleniu populacji na z góry założoną liczbę klas (tzw. skupień). Następnie uzyskany podział jest poprawiany w ten sposób, że niektóre elementy są przenoszone do innych klas, tak, aby uzyskać minimalną wariancję wewnątrz każdej z nich - dąży się do zapewnienia jak

największego podobieństwa elementów w ramach każdego ze skupień, przy jednoczesnej maksymalnej różnicy pomiędzy samymi klasami (skupieniami).

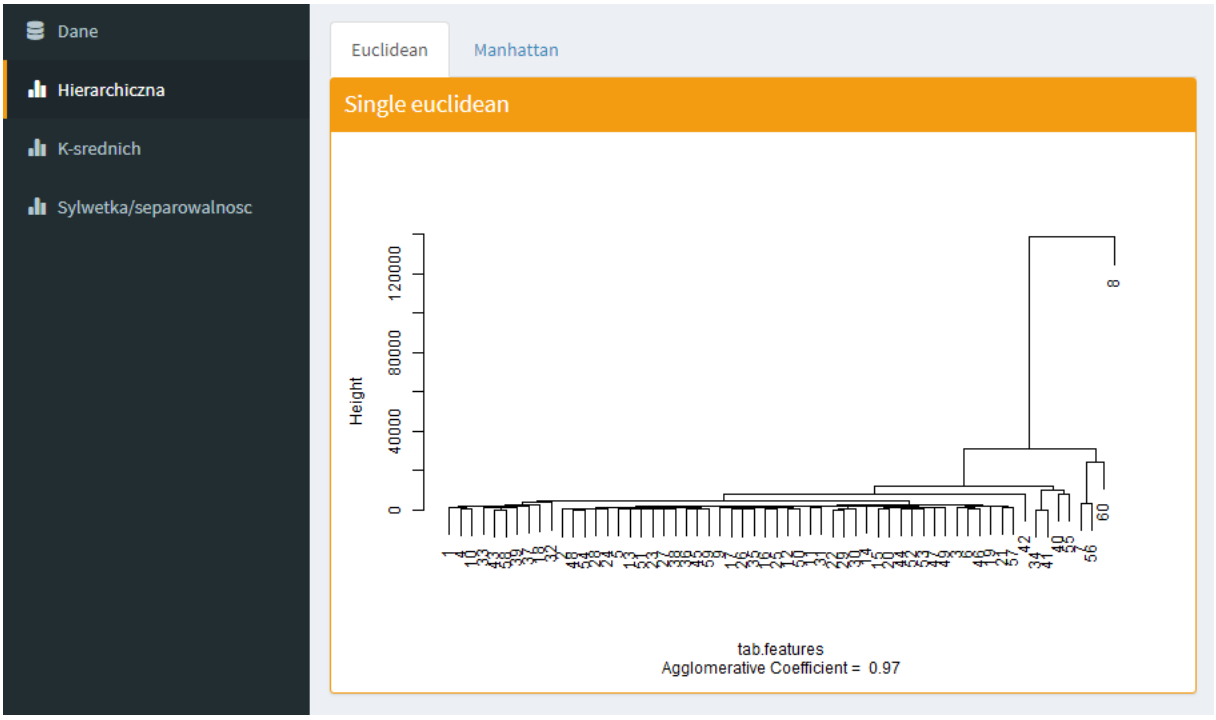
- Sylwetka – Jest syntetycznym wskaźnikiem jakości grupowania. Na jednej osi przedstawioną mamy ilość grup, na drugiej natomiast średnią miarę sylwetki, tzn. podobieństwo do pozostałych obserwacji w grupie oraz różnica od obserwacji innych grup.
- Separowalność – metoda podpowiada nam, ile klastrów maksymalnie jest sens tworzyć, co widać na wykresie. W momencie, gdy wykres przestaje rosnąć, to większa ilość klastrów nie ma sensu.

6. Aplikacja

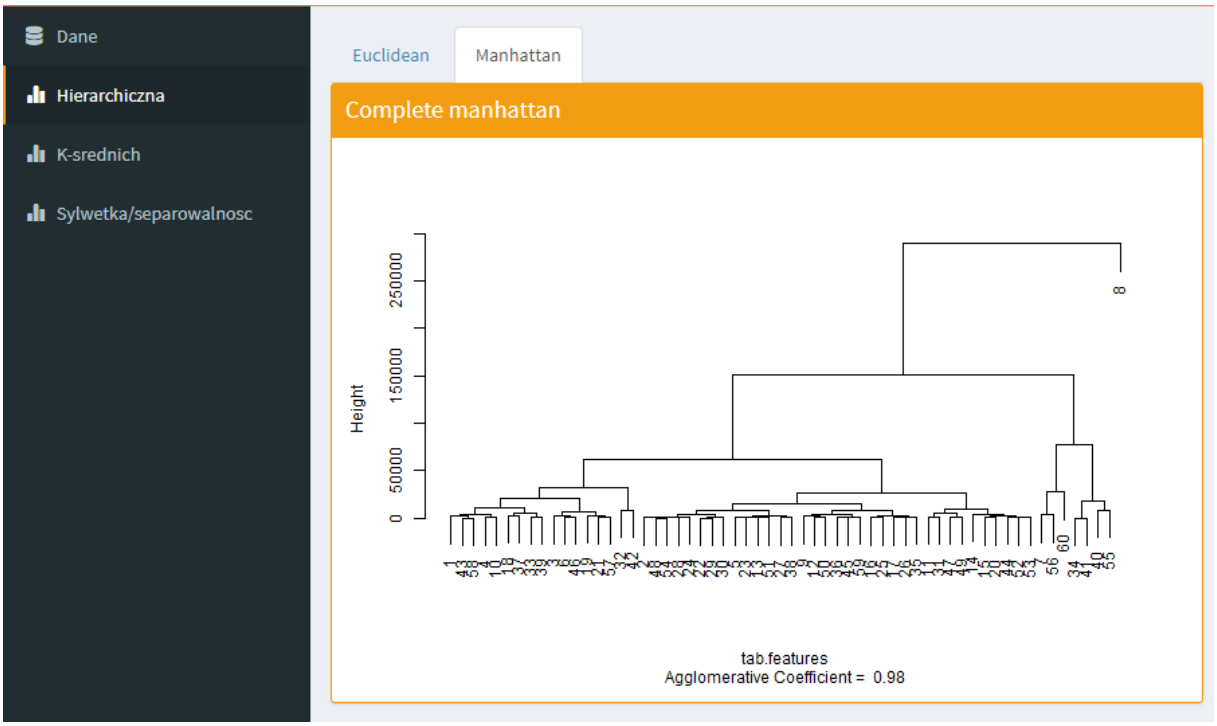
Po uruchomieniu aplikacji ukazuje się widok taki jak na poniższym screenie. Przedstawia on tabelę z danymi wczytanymi na potrzeby użycia metod klasteryzacji. Po lewej stronie znajduje się nawigacja po zakładkach aplikacji.



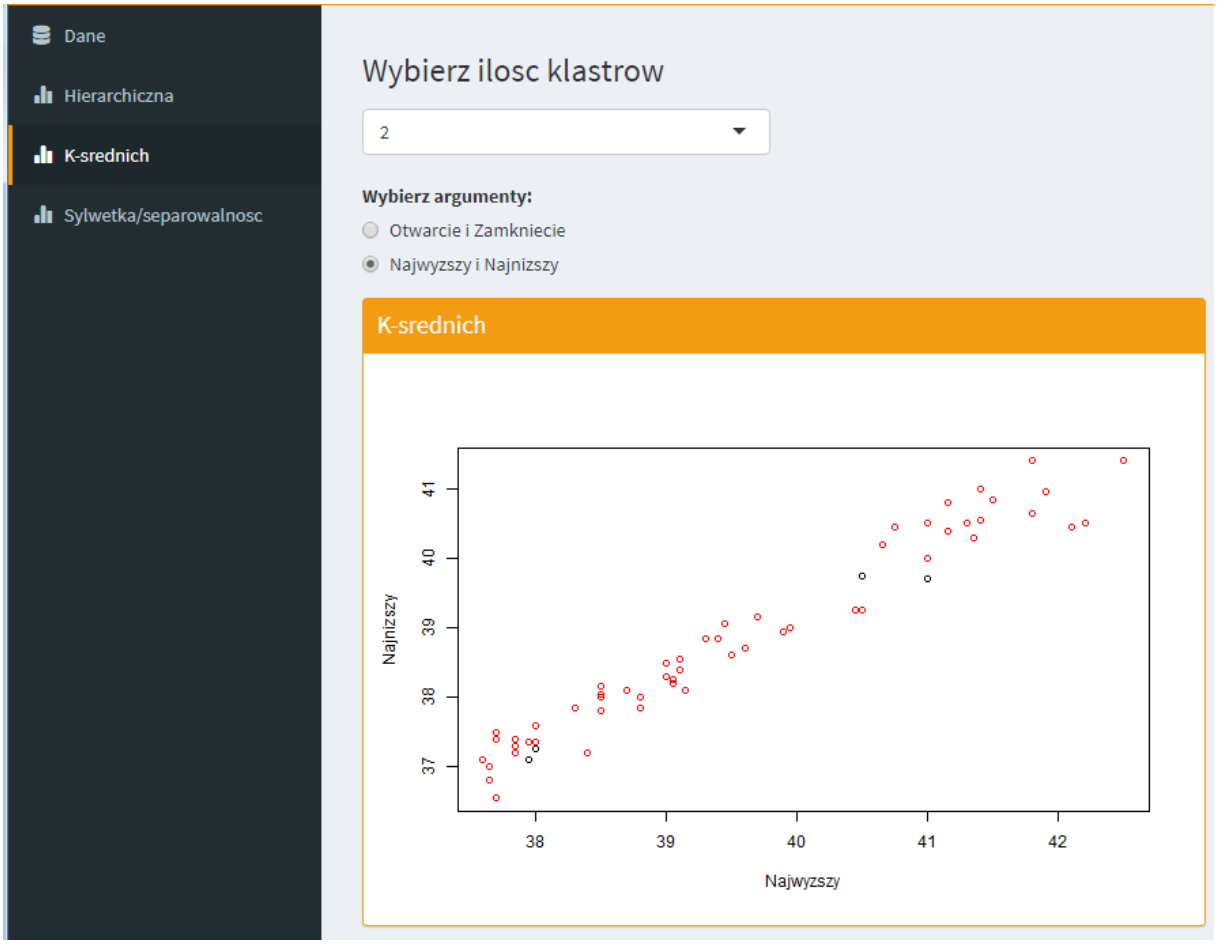
Po wybraniu opcji „hierarchiczna”, mamy do wyboru 2 zakładki: „Euclidean” oraz „Manhattan”. W pierwszej mamy wykres „Euclidean Manhattan”.



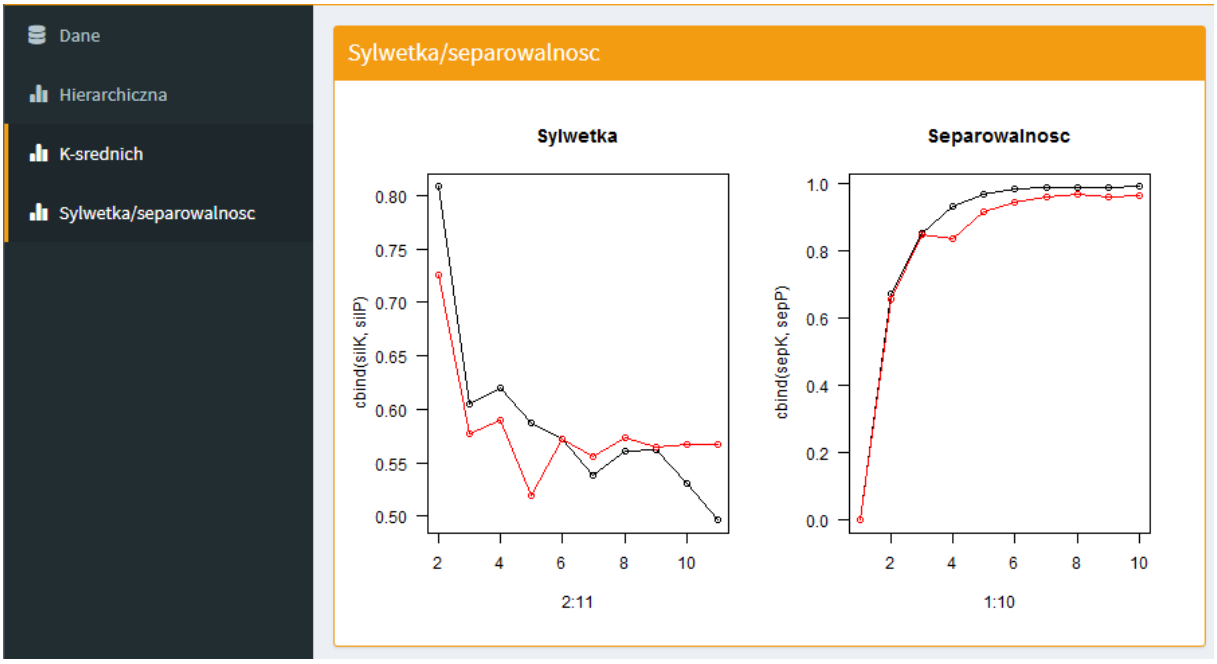
Druga zakładka w grupie metod hierarchicznych zawiera wykres wynikowy miary „Manhattan complete”.



Kolejna metoda „k-średnich” – mamy tutaj możliwość wyboru ilości klastrów (1,2 lub 3) oraz wyboru argumentów tabeli (otwarcie i zamknięcie lub najwyższy i najniższy). Na podstawie wybranych danych zostaje wygenerowany wykres z wizualnym podziałem na klastry.



Ostatni element to metody sylwetki i separowalności. Z wykresu „separowalność” widać, że maksymalna ilość klastrów dla użytych danych wynosi 5, ponieważ później wykres już nie rośnie więc traci to sens. Natomiast wykres „sylwetka” na jednej osi przedstawia ilość grup, na drugiej natomiast średnią miarę sylwetki, tzn. podobieństwo do pozostałych obserwacji w grupie oraz różnica od obserwacji innych grup.



7. Bibliografia

1. https://www.r-bloggers.com/how-to-perform-hierarchical-clustering-using-r/
2. https://pl.wikipedia.org/wiki/Grupowanie_hierarchiczne
3. https://pl.wikipedia.org/wiki/Analiza_skupie%C5%84
4. http://mst.mimuw.edu.pl/lecture.php?lecture=st2&part=Ch6#S2

Spis treści