Uniwersytet Rzeszowski

Dokumentacja

Przedmiot: **Sztuczna Inteligencja**

Wykonał: **Tomasz Niemczyk**

Rzeszów 2020

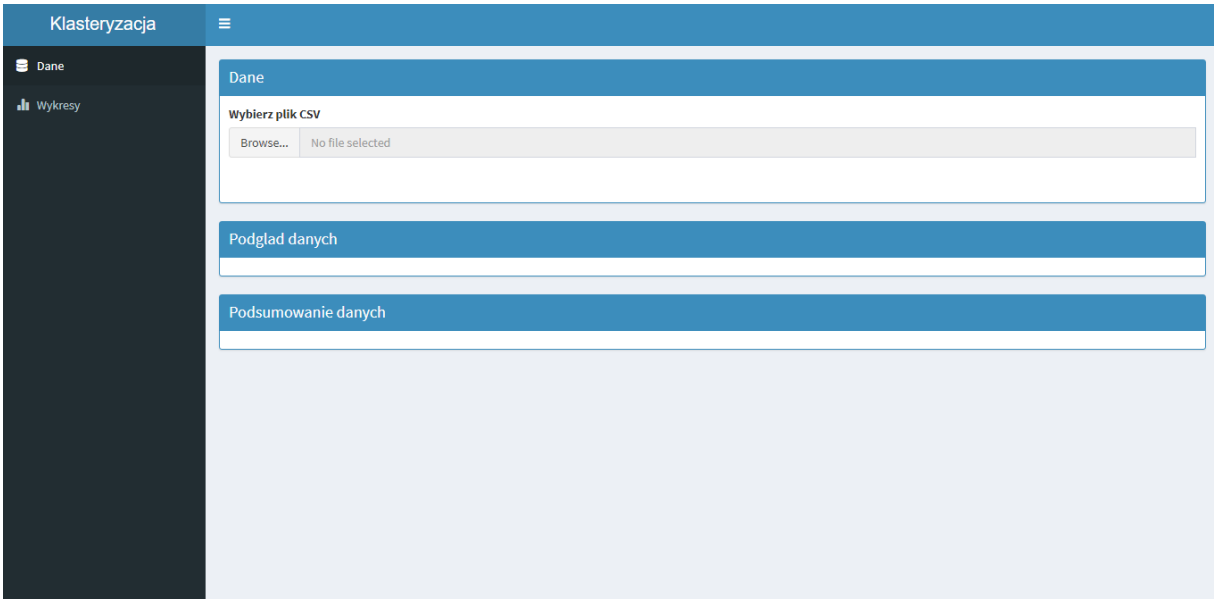
1. Opis projektu

W projekcie przedstawiony został problem klasteryzacji danych. Aplikacja webowa zaimplementowana została w języku R wraz z pakietem Shiny.

Klasteryzacja obejmuje takie metody jak:

- Hierarchiczna – zstępująca,
- Hierarchiczna – wstępująca,
- Metoda K-średnich,
- Separowalność,
- Sylwetka K-średnich.

Po wczytaniu danych z pliku CSV aplikacja w graficzny sposób wyświetla dane.



2. Wykorzystane narzędzia i technologie

Do budowy aplikacji wykorzystane zostały następujące narzędzia i technologie:

- **Język 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 Studio** - jest narzędziem ułatwiającym pracę z R. Jest to edytor, manager wersji, narzędzie wspierające debugowanie, tworzenie pakietów, aplikacji czy raportów.

- **Pakiet Shiny** - pozwala na tworzenie interaktywnych aplikacji web w prosty sposób. Wystarczy podstawowa znajomość R, aby tworzyć tabele, interaktywne wykresy i dashboardy. Dzieki R Shiny możemy eksplorować dane w zależności od poszczególnych zmiennych i parametrów oraz oglądać ich zmiany w czasie.

- **Shiny Dashboard** – jest to pakiet, który w prosty sposób pozwala tworzyć graficzny interfejs aplikacji internetowych

3. Przykładowe dane

W projekcie wykorzystane zostały dane krypto waluty BitCoin'a w okresie 1 – 31 październik 2019.

Dane w pliku „bitcoin.csv" oddzielone od siebie są średnikiem „;".

Dane przedstawiają:

- Datę,
- Symbol,
- Kurs otwarcia,
- Kurs najwyższy w danym dniu,
- Kurs najniższy w danym dniu,
- Kurs zamknięcia,
- Wolumen BTC i USD.

Przykład danych:

Date;Symbol;Open;High;Low;Close;Volume BTC;Volume USD

2019-10-31;BTCUSD;9203.53;9401.84;8981.31;9195;4.947;45171.46

2019-10-30;BTCUSD;9449.75;9449.75;9010.58;9203.53;4.774;43867.35

2019-10-29;BTCUSD;9237.9;9539.26;9129.21;9449.75;13.66;127758.35

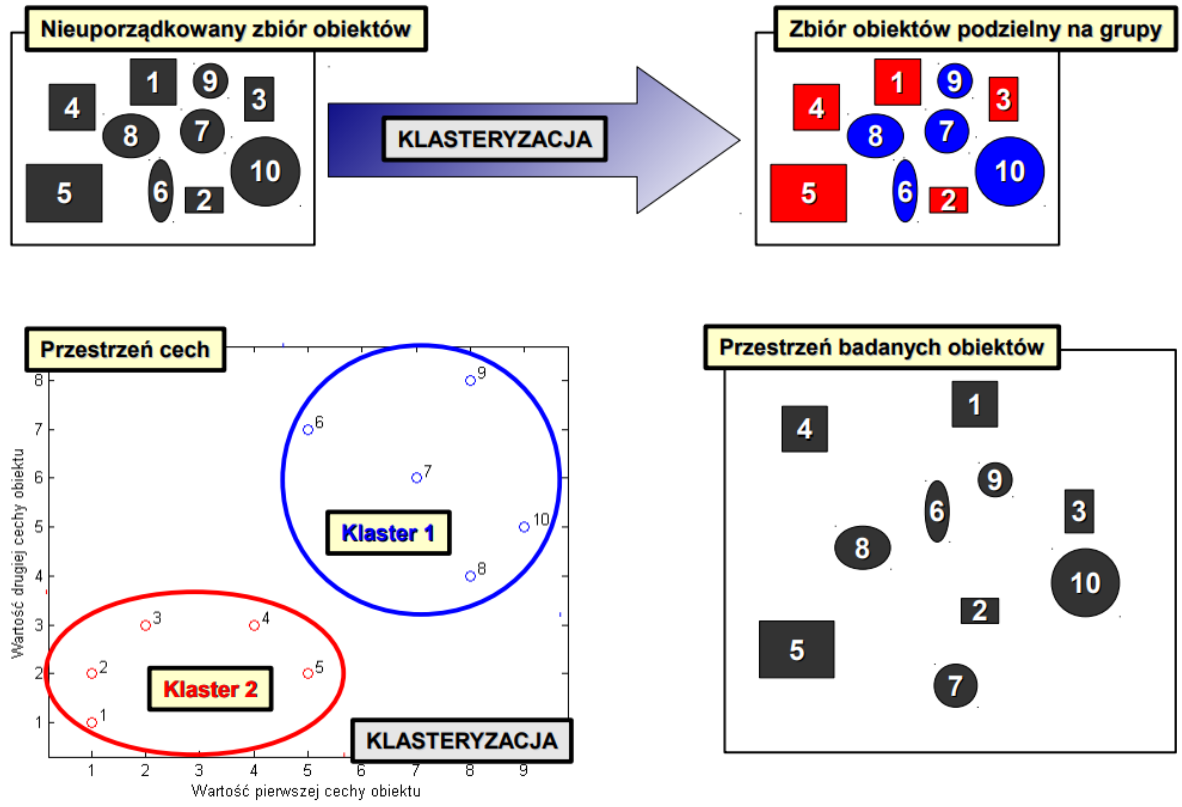
2019-10-28;BTCUSD;9489.84;9868.05;9214.93;9237.9;18.18;170792.7

2019-10-27;BTCUSD;9174.45;9790;9128.37;9489.84;10.75;101258.66

4. Klasteryzacja i podejścia

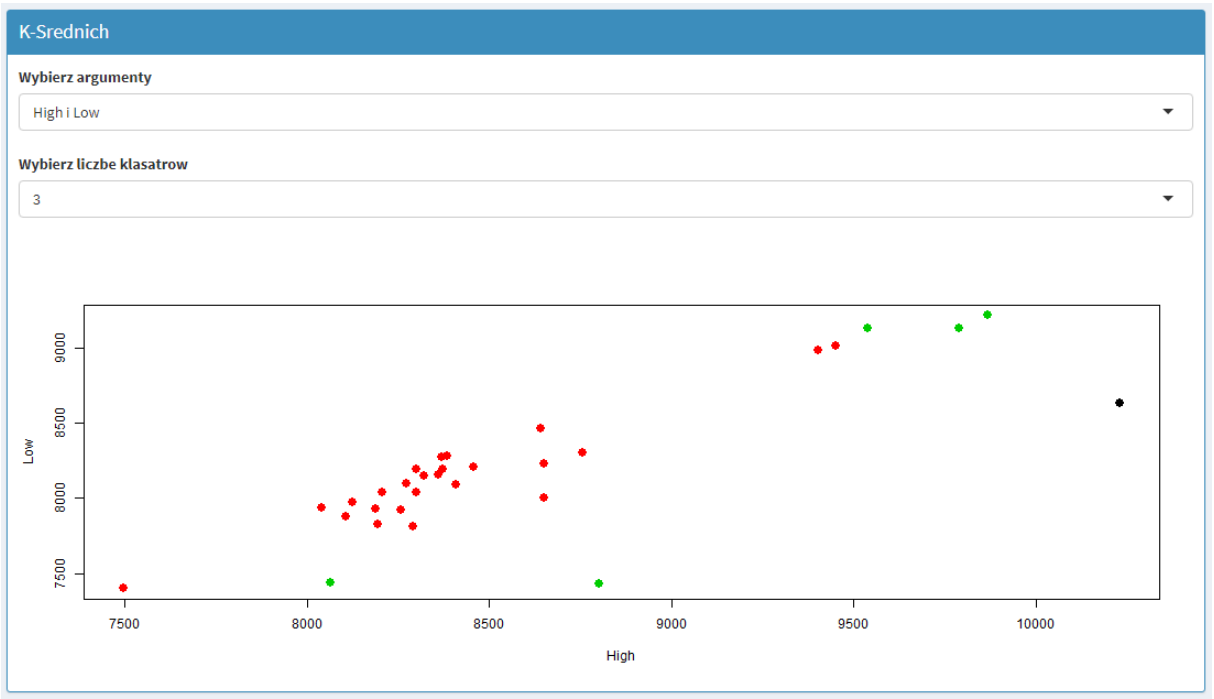
Projekt opiera się na klasteryzacji wybranych danych.

Klasteryzacja (analiza skupień, grupowanie) oznacza podzielenie zbioru obiektów na pewną liczbę rozłącznych klastrów (skupisk, grup), w taki sposób, aby każdy klaster zawierał obiekty możliwie do siebie podobne (wedle ustalonego kryterium podobieństwa), przy jednoczesnym zachowaniu możliwie dużego niepodobieństwa wobec obiektów z pozostałych grup.



Źródło: http://www.ire.pw.edu.pl/~trubel/mpb/files/MPB_08.pdf

Metoda k-średnich - Metoda k-średnich jest metodą należącą do grupy algorytmów analizy skupień tj. analizy polegającej na szukaniu i wyodrębnianiu grup obiektów podobnych (skupień) . Reprezentuje ona grupę algorytmów niehierarchicznych. Główną różnicą pomiędzy niehierarchicznymi i hierarchicznymi algorytmami jest konieczność wcześniejszego podania ilości skupień. Przy pomocy metody k-średnich zostanie utworzonych k różnych możliwie odmiennych skupień. Algorytm ten polega na przenoszeniu obiektów ze skupienia do skupienia tak długo aż zostaną zoptymalizowane zmienności ewnątrz skupień oraz pomiędzy skupieniami.



Klasteryzacja hierarchiczna

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.

Algorytm klasteryzacji hierarchicznej

$C = \{1\}, \{2\}, \ldots, \{n\}$

for (l **in** 1:(n-1))

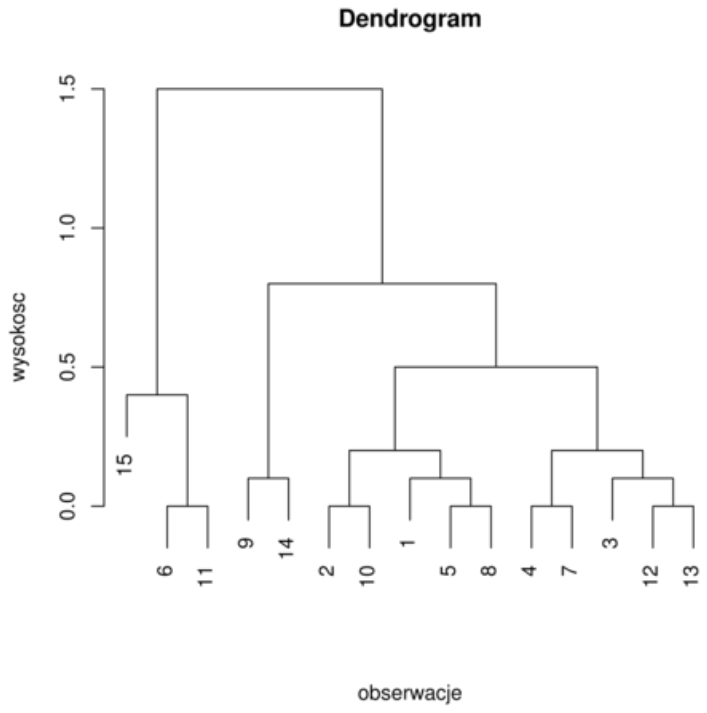
połącz najbliższe dwa klastry:

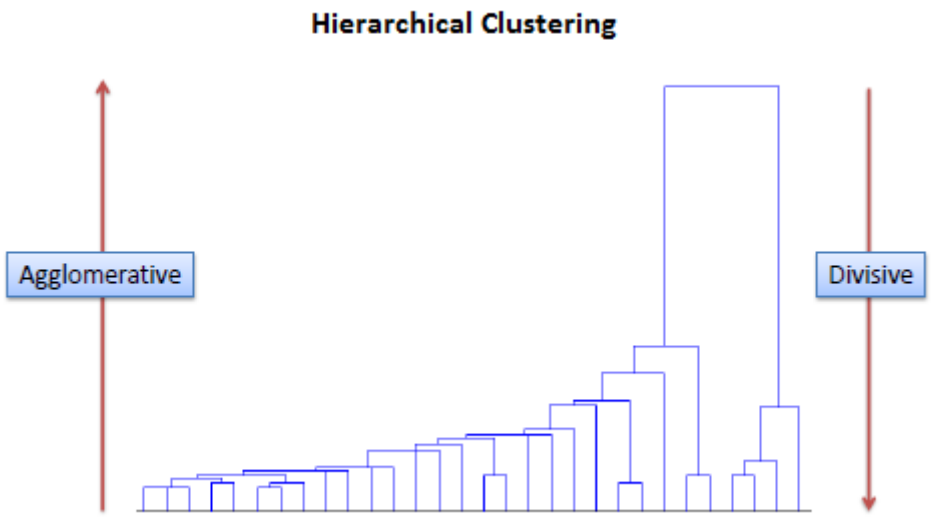
$(i_*, j_*) = \operatorname{argmin}_{i,j:i<j} d_{ij}$

klastry i_* oraz j_* zastąp przez 0

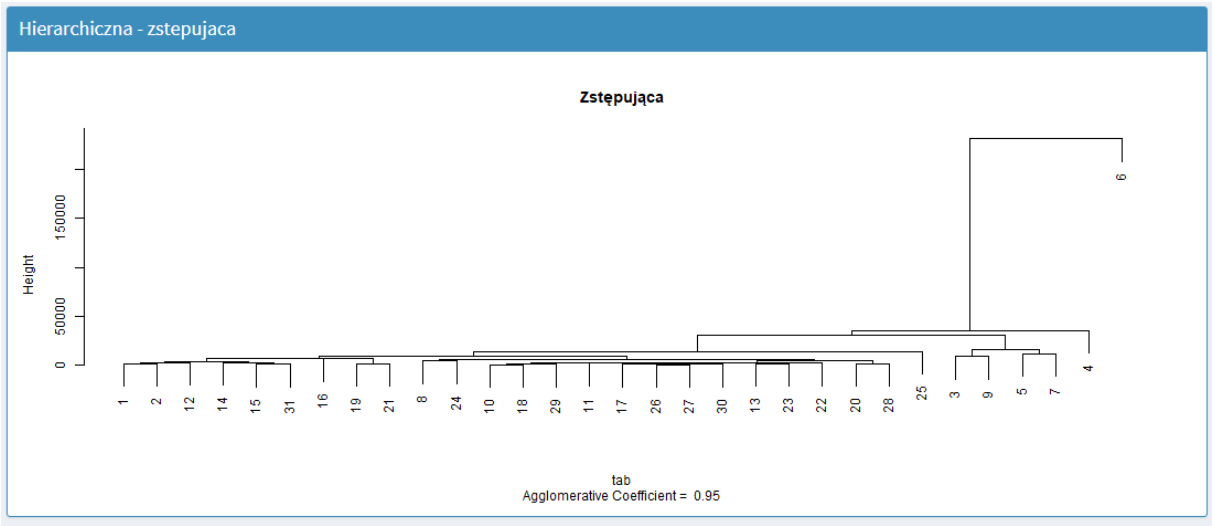
odnów macierz odległości $d_{0,k} = \min(d_{i_*k}, d_{j_*k})$

Dendrogram jest metodą ilustracji wyników klasteryzacji hierarchicznej. Możemy obserwować od dołu dendrogramu jak kolejne klastry się łączą i dla jakiej wysokości (odległości klastrów) to zachodzi.

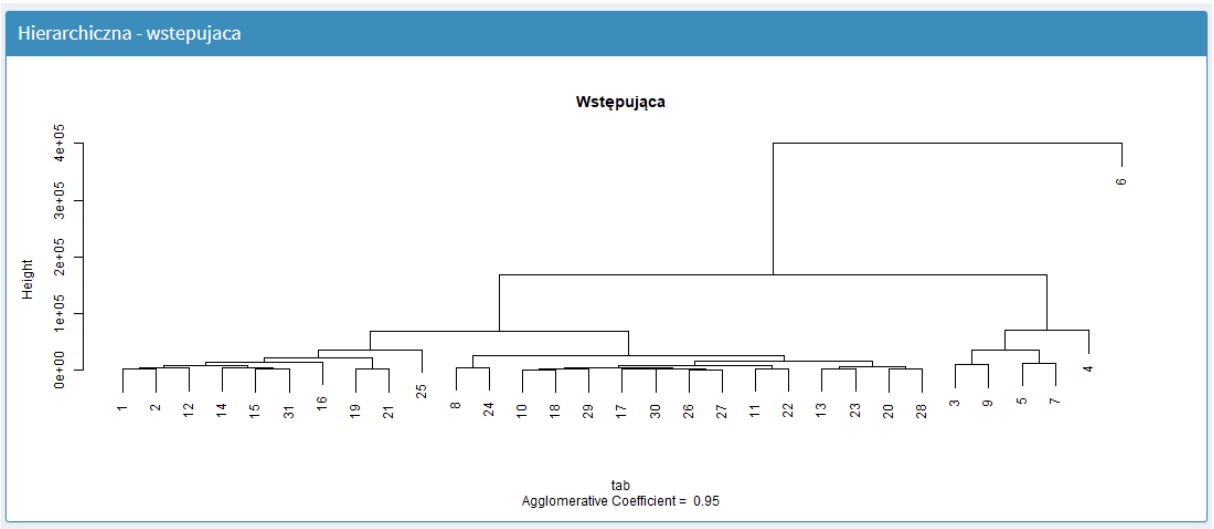




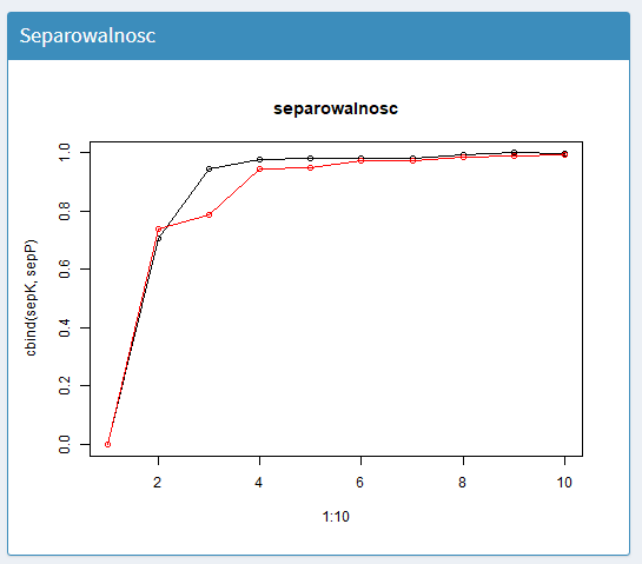
Metoda zstępująca (divisive method) - rozpoczyna działanie, gdy wszystkie obserwacje znajdują się w jednej grupie. W kolejnych iteracjach działania algorytmu grupy są dzielone mając na uwadze te same kryteria, co w przypadku algorytmów aglomeracyjnych: wariancję i miary odległości pomiędzy grupami. Tego typu algorytmy sprawdzają się, gdy zależy nam na znalezieniu dużych grup (podejście „od ogółu do szczegółu").



Metoda wstępująca (agglomerative method) - w tym podejściu algorytm zaczyna swe działanie w momencie, gdy liczba grup równa się liczbie obserwacji – każda obserwacja stanowi odrębną grupę. Następnie w sposób iteracyjny grupy są scalane w taki sposób, by wariancja wewnątrz nich była możliwie najmniejsza, a pomiędzy grupami możliwie duża. Algorytmy te reprezentują podejście „od szczegółu do ogółu”. Podejście zaimplementowane w algorytmach aglomeracyjnych sprawdza się, gdy decydujemy się na poszukiwanie małych grup (po kilku iteracjach proces może zostać przerwany).



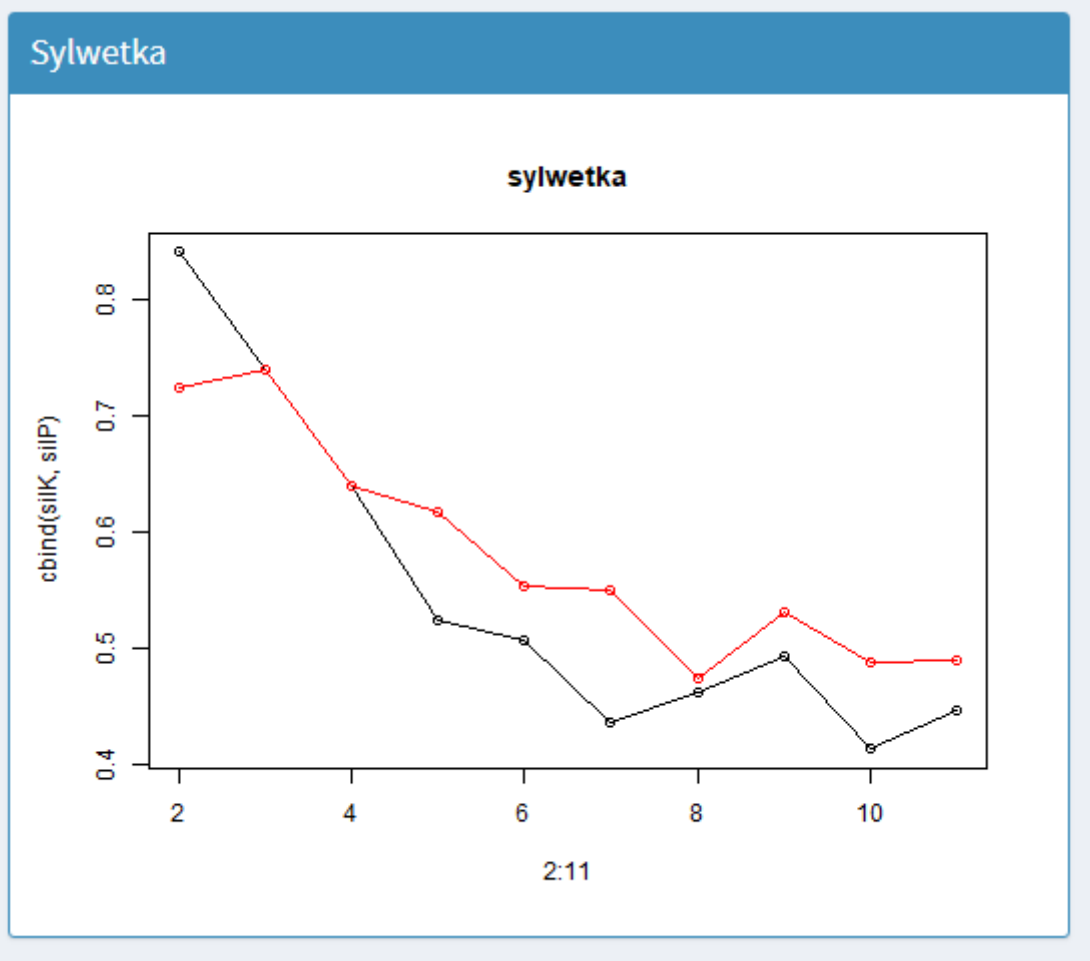
Separowalność dla klasteryzacji hierarchicznej - separowalność dla klasteryzacji K-średnich jest niemalejącą funkcją k, liczby klastrów. Na podstawie separowalności podejmuje się decyzję dotyczącą optymalnej ilości klastrów. Chcemy znaleźć taką niewielką liczbę klastrów, żeby zysk mierzony separowalnością przy łączeniu klastrów w danym kroku był duży, a dalsze sklejanie grup nie dawało już takich korzyści.



Sylwetka k-średnich – obrazuje on na jednej osi „x” liczbę grup. Oś y to średnia miara sylwetki (profilu) wszystkich obserwacji zbioru. Wskaźnik sylwetki mówi o tym, jak poszczególna obserwacja:

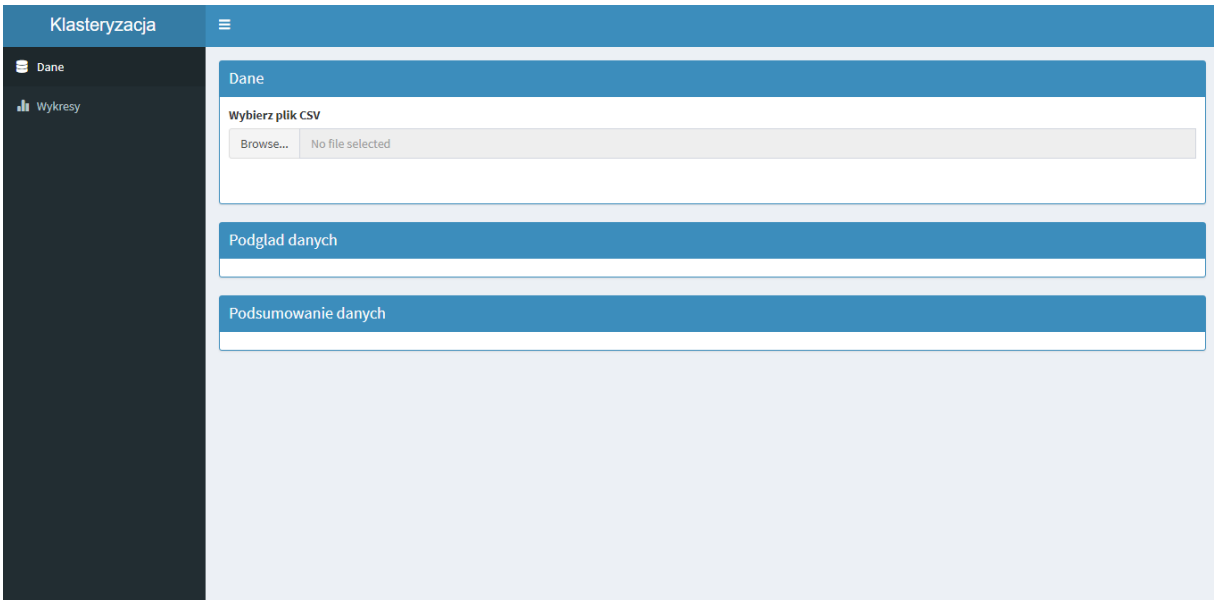
- jest podobna do pozostałych obserwacji w grupie,
- jest różna od obserwacji w pozostałych grupach.

Wybrać należy liczbę grup, dla której średnia wartość sylwetki jest największa.

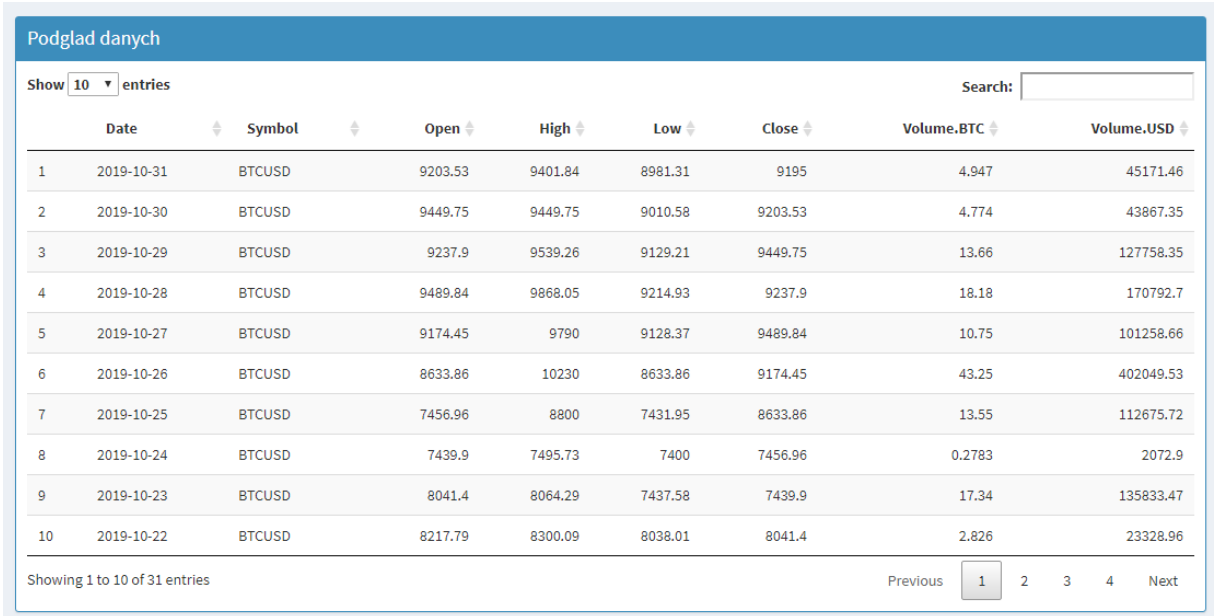


5. Implementacja i zrzuty ekranu

Okno główne:



Okno „Podgląd danych” po wczytaniu danych:



Podglad danych

Show 10 entries Search:

	Date	Symbol	Open	High	Low	Close	Volume.BTC	Volume.USD
1	2019-10-31	BTCUSD	9203.53	9401.84	8981.31	9195	4.947	45171.46
2	2019-10-30	BTCUSD	9449.75	9449.75	9010.58	9203.53	4.774	43867.35
3	2019-10-29	BTCUSD	9237.9	9539.26	9129.21	9449.75	13.66	127758.35
4	2019-10-28	BTCUSD	9489.84	9868.05	9214.93	9237.9	18.18	170792.7
5	2019-10-27	BTCUSD	9174.45	9790	9128.37	9489.84	10.75	101258.66
6	2019-10-26	BTCUSD	8633.86	10230	8633.86	9174.45	43.25	402049.53
7	2019-10-25	BTCUSD	7456.96	8800	7431.95	8633.86	13.55	112675.72
8	2019-10-24	BTCUSD	7439.9	7495.73	7400	7456.96	0.2783	2072.9
9	2019-10-23	BTCUSD	8041.4	8064.29	7437.58	7439.9	17.34	135833.47
10	2019-10-22	BTCUSD	8217.79	8300.09	8038.01	8041.4	2.826	23328.96

Showing 1 to 10 of 31 entries Previous 1 2 3 4 Next

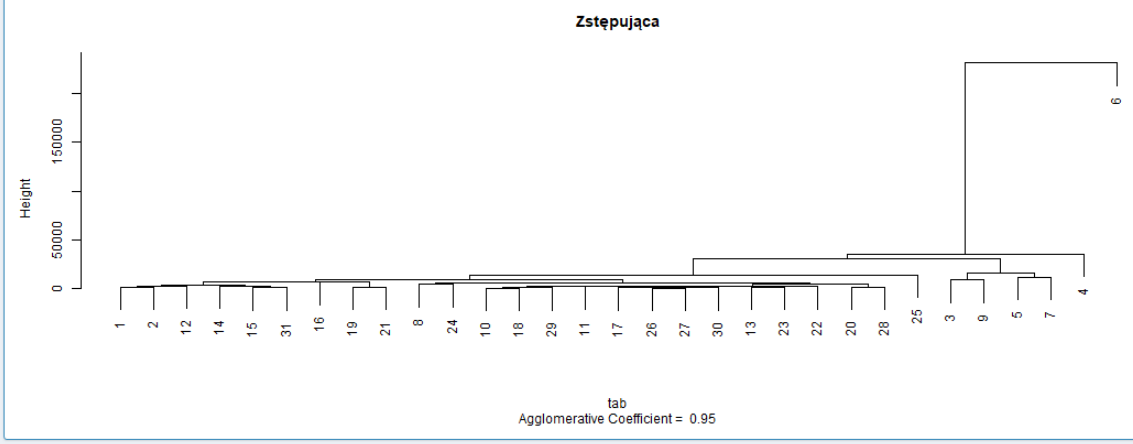
Okno „Podsumowanie danych”:

Podsumowanie danych

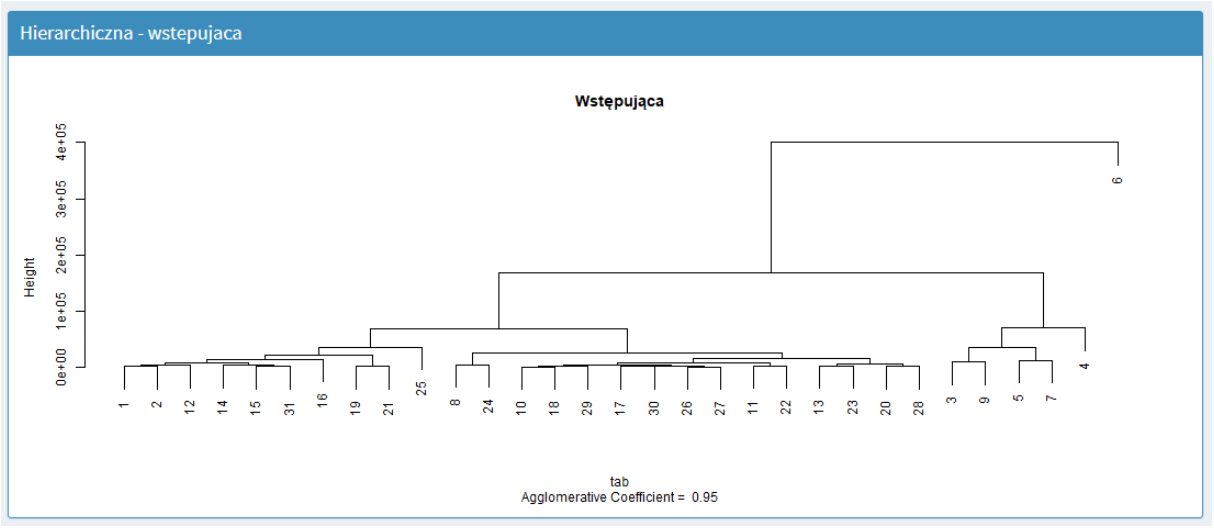
```
-- Data Summary ------------------------
                           Values
Name                       datas
Number of rows             31
Number of columns          8
_______________________
Column type frequency:
  factor                   2
  numeric                  6
_______________________
Group variables            None

-- Variable type: factor ---------------------------------------------------------
  skim_variable n_missing complete_rate ordered n_unique top_counts
1 Date                  0             1 FALSE         31 201: 1, 201: 1, 201: 1, 201: 1
2 Symbol                0             1 FALSE          1 BTC: 31

-- Variable type: numeric --------------------------------------------------------
  skim_variable n_missing complete_rate     mean       sd       p0      p25      p50      p75     p100 hist
1 Open                  0             1  8357.     502.   7440.     8120.    8273.    8473.     9490.  <U+2582><U+2587><U+2587><U+2581><U+2583>
2 High                  0             1  8590.     622.   7496.     8232.    8370.    8703.    10230   <U+2581><U+2587><U+2582><U+2581><U+2581>
3 Low                   0             1  8205.     479.   7400      7931.    8151.    8291.     9215.  <U+2582><U+2587><U+2587><U+2581><U+2583>
4 Close                 0             1  8386.     524.   7440.     8120.    8273.    8584.     9490.  <U+2582><U+2587><U+2587><U+2581><U+2585>
5 Volume.BTC            0             1     6.71     8.23    0.278     2.47     4.27     6.73     43.2 <U+2587><U+2581><U+2581><U+2581><U+2581>
6 Volume.USD            0             1 58327.   76002.   2073.    20153.   35825.   54168.   402050.  <U+2587><U+2581><U+2581><U+2581><U+2581>
```

Metoda „Hierarchiczna – zstępująca”:

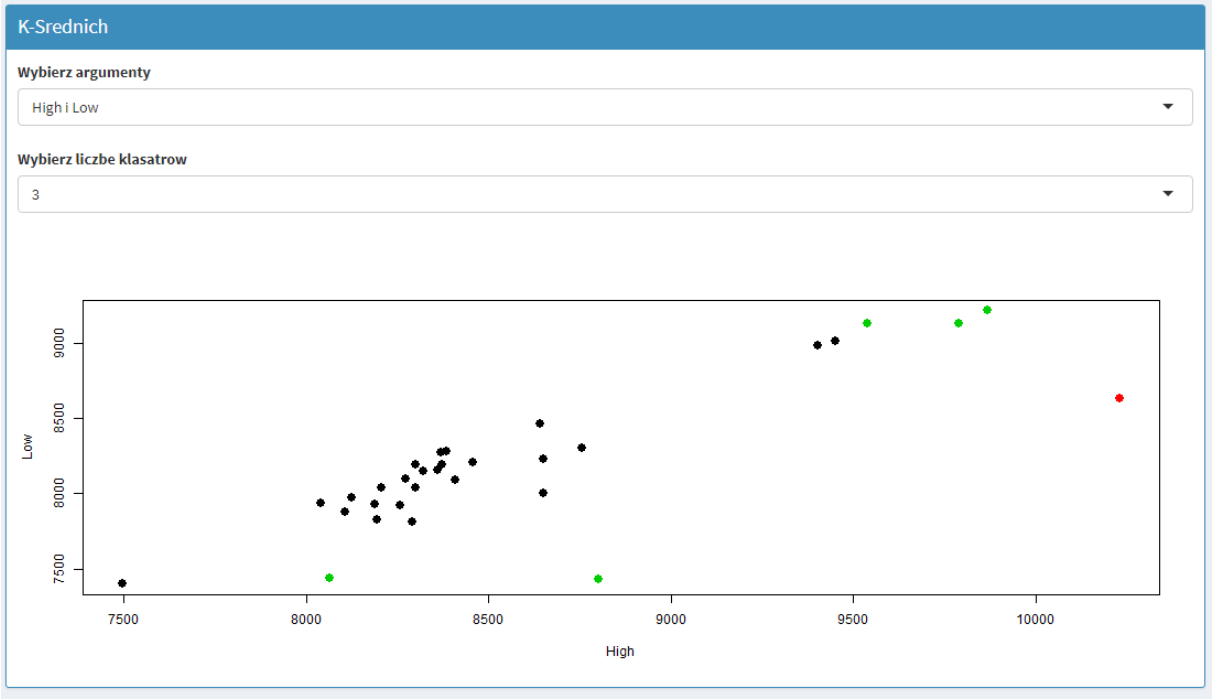
Hierarchiczna - zstepujaca



Metoda „Hierarchiczna – wstępująca”:



Metoda „K-średnich”:



Okno „sylwetka i separowalność”: