Kurs K1 (Wprowadzenie do eksploracji danych) – Laboratorium 5

Identyfikacja i usuwanie punktów oddalonych

Wczytujemy dane ze zbioru iris:

```
> path = "C:/R/dm"            ## scieżka dostępu do plików
> setwd(path)                 ## ustawienie ścieżki
> library(datasets)
> data(iris)
> View(iris)
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa

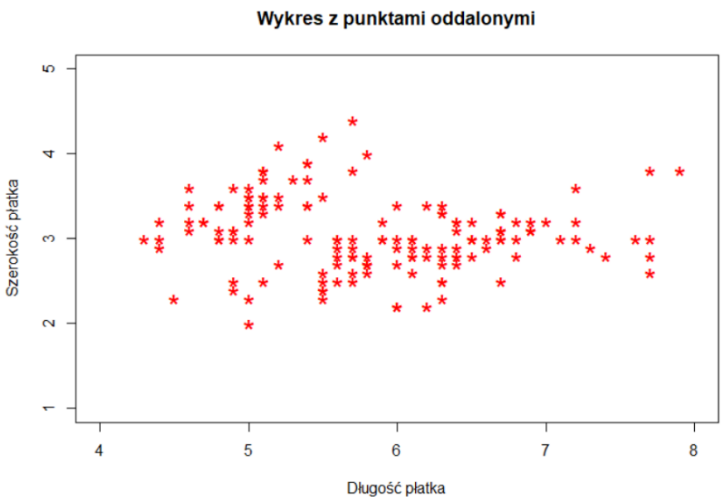
Szukamy obserwacji oddalonych (np. według zmiennych Sepal.Length i Sepal.Width)

Sprawdzamy zakresy zmiennych

```
> range(iris$Sepal.Length)
[1] 4.3 7.9
> range(iris$Sepal.Width)
[1] 2.0 4.4
```

Tworzymy wykres rozrzutu dla zmiennych "Sepal.Length" i "Sepal.Width", żeby sprawdzić punkty oddalone

```
> plot(iris$Sepal.Length, iris$Sepal.Width,
     xlim=c(4,8),      ## zakres wartości wyświetlanych na osi x
     ylim=c(1,5),      ## zakres wartości wyświetlanych na osi y
     main="Wykres z punktami oddalonymi",
     xlab="Długość płatka",
     ylab="Szerokość płatka",
     pch="*",          ## kształt punktu na wykresie
     col="red",        ## kolor punktu
     cex=2)            ## rozmiar punktu
```



Wykres wskazuje na brak ewidentnych punktów oddalonych dla rozważanych zmiennych.

Dla celów ilustracyjnych utworzymy zbiór "iris2" z wyraźnym punktem oddalonym.

Tworzymy tabelę z jednym punktem (o nietypowej wartości pod względem długości i szerokości płatka):

```
> iris_outliers<-data.frame(Sepal.Length=c(20), Sepal.Width=c(7),
  Petal.Length=c(1.5), Petal.Width=c(0.3), Species=("setosa"))
> View(iris_outliers)
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	20	7	1.5	0.3	setosa

Tworzymy tabelę łączącą tabele "iris" i "iris_outliers":

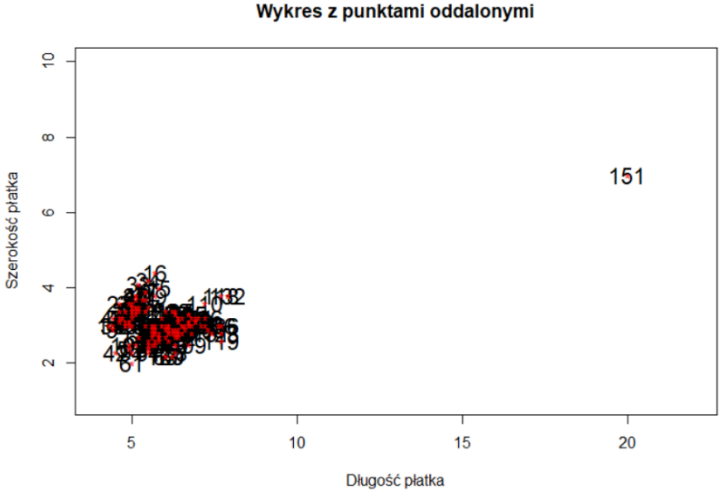
```
> iris2<-rbind(iris, iris_outliers)
> View(iris2)
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
147	6.3	2.5	5.0	1.9	virginica
148	6.5	3.0	5.2	2.0	virginica
149	6.2	3.4	5.4	2.3	virginica
150	5.9	3.0	5.1	1.8	virginica
151	20.0	7.0	1.5	0.3	setosa

Ponownie tworzymy wykres rozrzutu, żeby sprawdzić punkty oddalone - tym razem dla tabeli "iris2" (zauważ, że zakres wartości na osi x i y zostały zwiększony):

```
> plot(iris2$Sepal.Length, iris2$Sepal.Width,
     xlim=c(4,22),
     ylim=c(1,10),
     main="Wykres z punktami oddalonymi",
     xlab="Długość płatka",
     ylab="Szerokość płatka",
     pch="*",          ## kształt punktu na wykresie
     col="red",        ## kolor punktu
     cex=1,            ## rozmiar punktu
     text(iris2$Sepal.Length, iris2$Sepal.Width, labels=as.numeric(rownames
          (iris2)), cex= 1.5))
```

Na wykresie widać punkt oddalony (wiersz numer 151):



Usuniemy punkt oddalony (wiersz 151 tabeli "iris2") **metodą rozstępu międzykwartylowego** biorąc pod uwagę zmienną "Sepal.Length"
(*przypomnienie: metoda była omawiana na konwersatorium – zob. moodle, plik "_WED_K02_GS", slajd 25-26*)

Możemy sprawdzić kwantyle zmiennej:

```
> quantile(iris2$Sepal.Length)
  0%  25%  50%  75% 100%
 4.3  5.1  5.8  6.4 20.0
```

Kwantyle Q1 i Q3 możemy też otrzymać następująco:

```
> quantile(iris2$Sepal.Length, probs=c(.25, .75))
          25%           75%
-0.0006367493  1.0020519159
```

Możemy też obliczyć rozstęp międzykwartylowy,czyli IQR(iris2$Sepal.Length) = 6.4 - 5.1 = 1.3:

```
> IQR(iris2$Sepal.Length)
[1] 1.3
```

Usuwamy z tabeli iris2 punkty oddalone (uwzględniając zmienną"Sepal.Length")

1) Wyznaczamy Q1 and Q3:

```
> qnt = quantile(iris2$Sepal.Length, probs=c(.25, .75))
> qnt
25% 75%
5.1 6.4
```

2) Zastępujemy wartości oddalone zmiennej wartościami brakującymi ("NA"):

```
> iris2$Sepal.Length[iris2$Sepal.Length < qnt[1] - 1.5 *
  IQR(iris2$Sepal.Length)] <- NA
> iris2$Sepal.Length[iris2$Sepal.Length > qnt[2] + 1.5 *
  IQR(iris2$Sepal.Length)] <- NA
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species	Id
147	6.3	2.5	5.0	1.9	virginica	147
148	6.5	3.0	5.2	2.0	virginica	148
149	6.2	3.4	5.4	2.3	virginica	149
150	5.9	3.0	5.1	1.8	virginica	150
151	NA	7.0	1.5	0.3	setosa	151

3) Usuwamy z tabeli iris2 punkty z brakującymi wartościami (elementy NA):

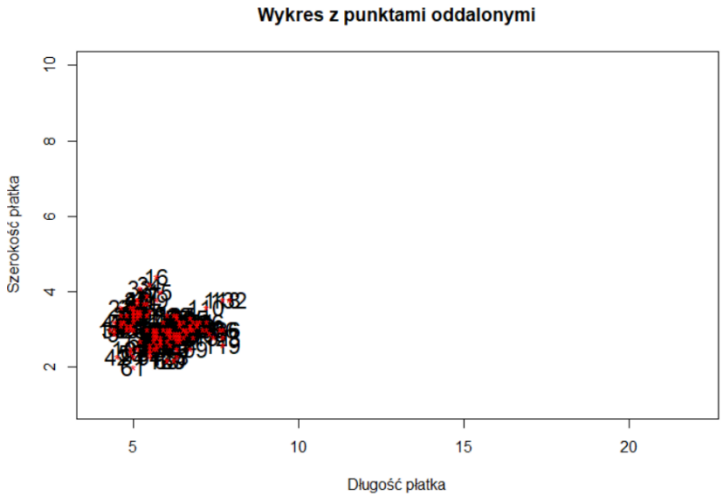
```
> iris2 <- iris2 [complete.cases(iris2),]
> View(iris2)
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species	Id
146	6.7	3.0	5.2	2.3	virginica	146
147	6.3	2.5	5.0	1.9	virginica	147
148	6.5	3.0	5.2	2.0	virginica	148
149	6.2	3.4	5.4	2.3	virginica	149
150	5.9	3.0	5.1	1.8	virginica	150

Showing 146 to 150 of 150 entries

Ponownie tworzymy wykres rozrzutu dla tabeli "iris2" - nie ma już punktu oddalonego:

```
> plot(iris2$Sepal.Length, iris2$Sepal.Width,
      xlim=c(4,22),
      ylim=c(1,10),
      main="Wykres z punktami oddalonymi",
      xlab="Długość płatka",
      ylab="Szerokość płatka",
      pch="*",          ## kształt punktu na wykresie
      col="red",        ## kolor punktu
      cex=1,            ## rozmiar punktu
      text(iris2$Sepal.Length, iris2$Sepal.Width, labels=iris2$Id, cex= 1.5))
```



Zadania do wykonania

1. Sprawdź na zbiorze „iris” działanie kodu R przedstawionego w niniejszym opracowaniu.
2. Następnie wykonaj poniższe instrukcje **dla zbioru danych *wine***. W pliku tekstowym przygotuj odpowiednie informacje (wklej polecenia i wyniki/rysunki):
   a) Sprawdź metodą rozstępu międzykwartylowego (dla wszystkich zmiennych numerycznych), czy w zbiorze są jakieś obserwacje oddalone (napisz, ile było punktów oddalonych dla każdej zmiennej); napisz pętlę.
   (Jeżeli w zbiorze nie ma punktów oddalonych, to dodaj „sztucznie” co najmniej jeden taki punkt.)
   b) Zilustruj punkty oddalone (dla odpowiednich zmiennych) na wykresie rozrzutu lub histogramie (w przypadku dużej liczby cech proszę zrobić to dla maksymalnie 4 zmiennych).
   c) Usuń punkty oddalone ze zbioru danych (w pętli). Podsumuj, ile punktów ostatecznie usunięto, a ile zostało.
   d) Sporządź analogiczne wykresy jak w punkcie b po usunięciu punktów oddalonych.
   e) Zapisz nowy zbiór danych (bez punktów oddalonych) do pliku. Załącz wynik polecenia *View*.

Przyślij mailem plik z odpowiedziami, a także plik wygenerowany w punkcie 2e.