1. WSTĘPNA ANALIZA DANYCH

ZADANIE 1.1 Wczytać zbiór *pima* znajdujący się w bibliotece *faraway*.

```
> install.packages("faraway")      # instalujemy bibliotekę faraway
> library(faraway)                 # uaktywniamy bibliotekę faraway
> data(pima)                       # uaktywniamy zbiór pima
```

(a) Obejrzeć opis zbioru *pima* by sprawdzić jakie informacje zawierają jego zmienne.

```
> ?pima
```

Wyznaczyć podstawowe miary liczbowe dla wszystkich tych zmiennych i przyjrzeć się czy wśród badanych danych nie ma błędów i rzeczy nietypowych.

```
> summary(pima)
```

W szczególnści zauważyć, że zmienna *test*, będąca zmienną jakościową, została zapisana jako zmienna ilościowa. Przyjrzeć się także wartościom obserwacji dla zmiennych *diastolic*, *glucose*, *triceps* i *bmi* i zwrócić uwagę, że pojawia się 0, które nie mają sensu (0 zostały wpisane w miejsca brakujących obserwacji). Wprowadzić stosowne poprawki.

```
> pima$test <- factor(pima$test)    # zmienna test ze zbioru danych pima zostaje
                                    # przerobiona na zmienną typu jakościowego (factor)
> levels(pima$test)<-c("brak objawow","sa objawy") # opisujemy poziomy zmiennej test
> pima$diastolic[pima$diastolic==0] <- NA    # w kolumnie diastolic w zbiorze pima
                                    # obserwacje, któr równają się 0 zostają zamienione
                                    # na obserwacje brakujące (NA - not available)
```

(b) Wyznaczyć średnią, medianę, dolny i górny kwartyl, 1-szy decyl, rozstęp, rozstęp międzywartylowy oraz odchylenie standardowe dla zmiennej *diastolic*.

(c) Wyznaczyć średnie rozkurczowe ciśnienie krwi oraz jego odchylenie standardowe dla kobiet, u których zaobserwowano objawy cukrzycy.

```
> pima$diastolic[pima$test=="sa objawy"]    # dla zmiennej diastolic ze zbioru pima
                                    # zostają wybrane jedynie te obserwacje, dla których
                                    # zmienna test przyjmuje wartość "sa objawy"
```

(d) Dla zmiennej *pregnant* sporządzić i opisać wykres skrzynkowy.

(e) Odczytać, u ilu spośród wszystkich badanych kobiet, stwierdzono objawy cukrzycy.

(f) Dla zmiennej *diastolic* sporządzić histogram częstości oraz narysować jądrowy estymator gęstości.

ZADANIE 1.2 Dane zawarte w pliku *gala_data.txt* zawierają informacje o kilkudziesięciu wyspach.

(a) Wczytać te dane. Uzyskać bezpośredni dostęp do zmiennych w tym zbiorze.

```
> dane.o.wyspach <- read.table(choose.files(),header=TRUE)
# wczytujemy dane z pliku, który, dzięki użyciu funkcji choose.files(), wskażemy klikając na
# ten plik (alternatywnie można podać pełną ścieżkę dostępu do pliku);
# w pierwszym wierszu wczytywanego pliku są umieszczone nazwy kolumn, więc argumentowi
# header musimy nadać wartość TRUE (domyśnie jest przypisana wartość FALSE)
> attach(dane.o.wyspach)    # uzyskujemy bezpośredni dostęp do zmiennych w zbiorze
                            # dane.o.wyspach; po skończeniu pracy z tym zbiorem piszemy:
> detach(dane.o.wyspach)
```

(b) Wyznaczyć podstawowe statystyki próbkowe (średnią, medianę, dolny i górny kwartyl, wartości ekstremalne, wariancję i odchylenie standardowe) dla danych opisujących liczbę gatunków żółwi występujących na badanych wyspach (zmienna *Species*).

(c) Narysować histogram o pięciu klasach dla danych opisujących powierzchnię badanych wysp (zmienna *Area*). Podpisać osie i umieścić nagłówek.

(d) Narysować i opisać wykres skrzynkowy dla danych przedstawiających liczbę gatunków żółwi na wyspach, których powierzchnia jest mniejsza od 25 (km^2).

ZADANIE 1.3 W niektórych źródłach można znaleźć komentarz, że współczynnik asymetrii A dostarcza informacji na temat symetrii rozkładu danych lub jej braku; w szczególności, że $A = 0$ świadczy o tym, iż rozkład danych jest symetryczny. Aby przekonać się, że stwierdzenie to nie jest słuszne, wyznaczyć współczynnik asymetrii dla następujących danych:

$$\frac{-1-\sqrt{10}}{4},\ -\frac{1}{4},\ -\frac{1}{4},\ \frac{\sqrt{10}-1}{4},\ 1.$$

Czy dane te są symetrycznie rozłożone względem średniej?

ZADANIE 1.4 Dla następujących danych:

$$5,\ 8,\ 9,\ 3,\ 8,\ 7$$

wyznaczyć ręcznie (tzn. bez użycia komputera) medianę oraz dolny i górny kwartyl. Uzyskane wyniki porównać z wartościami tych parametrów podawanymi przez R.