

1. WSTĘPNA ANALIZA DANYCH

ZADANIE 1.1 Poniższe dane zawierają wagę (w kg) oraz wzrost (w metrach) 6 mężczyzn:

waga: 60, 72, 57, 90, 95, 72
wzrost: 1,75; 1,80; 1,65; 1,90; 1,74; 1,91

(a) Wpisać powyższe dane jako **(a1)** dwa wektory, **(a2)** ramkę danych.

(b) Obliczyć $bmi = waga/(wzrost^2)$ dla każdej z osób. Wiedząc, że dla mężczyzn wskaźnik ten interpretuje się następująco:

- $bmi < 20,7 \implies$ niedowaga
- $20,7 \leq bmi < 26,5 \implies$ waga prawidłowa
- $bmi \geq 26,5 \implies$ nadwaga

utworzyć nową zmienną wyjaśniającą czy dana osoba ma niedowagę, wagę prawidłową czy nadwagę.

(c) Dla zmiennej *waga* wyznaczyć średnią, medianę, wariancję i odchylenie standardowe.

(d) Wyznaczyć współczynnik korelacji pomiędzy wagą i wzrostem.

(e) Sporządzić histogram oraz wykres skrzynkowy dla zmiennej *waga*.

(f) Narysować wykres rozrzutu wagi względem wzrostu oraz wagi względem kwadratu wzrostu. Umieścić oba te wykresy w jednym oknie i podpisać osie.

ZADANIE 1.2 Zbiór *barley.txt* zawiera dane dotyczące wielkości zbiorów (*yield*) różnych gatunków (*variety*) w różnych miejscach eksperymentalnych upraw (*site*) w latach 1931 i 1932 (*year*).

(a) Wczytać ten zbiór. Uzyskać bezpośredni dostęp do zmiennych w tym zbiorze.

(b) Wyznaczyć podstawowe statystyki próbkowe dla danych opisujących wielkość zbiorów w 1931 roku. Obliczyć kwantyl rzędu 0,9 dla tych danych.

(c) Wypisać te wielkości zbiorów w roku 1931, które były wyższe od kwantyla wyznaczonego w punkcie (b).

(d) Narysować histogram dla danych opisujących wielkość zbiorów w 1931 roku w miejscowości Grand Rapids.

(e) Sporządzić tablicę kontyngencji opisującą ile pomiarów zebrano w każdym z miejsc eksperymentalnych w poszczególnych latach.

ZADANIE 1.3 Dane zawarte w pliku *gala_data.txt* zawierają informacje o kilkudziesięciu wyspach.

(a) Wczytać te dane.

(b) Wyznaczyć podstawowe statystyki próbkowe (średnią, medianę, dolny i górny kwantyl, 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 nagłówki.

(d) W jednym oknie sporządzić wykres rozrzutu liczby gatunków żółwi względem powierzchni wyspy oraz wykres rozrzutu liczby gatunków żółwi względem logarytmu naturalnego (\log) z powierzchni wyspy. Na którym wykresie widać silniejszą zależność liniową?

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

ZADANIE 1.4

(a) Wygenerować po 500 liczb z następujących rozkładów:

- (a1) normalnego o średniej = 20 i odchyleniu standardowym = 5,
- (a2) jednostajnego na przedziale $(-1, 1)$,
- (a3) wykładniczego o średniej = 5,
- (a4) Poissona o średniej = 3.

WSKAZÓWKA: (a1) `rnorm` (a2) `runif` (a3) `rexp` (a4) `rpois`

(b) Dla każdej wygenerowanej próbki sporządzić wykres normalności i wykresy te wyświetlić w jednym oknie. Przeanalizować ich kształt.

(c) Dla danych wygenerowanych w pkt. (a1) i (a3) przeprowadzić, na poziomie istotności 0,05, test normalności Shapiro-Wilka.