1 Laboratorium 1. Wstępna analiza danych.

Wczytaj do pakietu dane z pliku 'dane do lab1.txt'.

1. Na wykresie kołowym (pie chart) przedstaw rozkład zmiennej *typ*. Umieść informacje o procentowym udziale w próbie każdego z zawodów. Dodaj również legendę umożliwiającą identyfikację każdego z tych zawodów.

2. Utwórz dwie nowe zmienne:

   - zmienną *wykszt*, która przyjmuje wartości 1, 2 i 3 odpowiadające wykształceniu podstawowemu, średniemu i wyższemu, tzn.

   $$wykszt = \begin{cases} 1, & \text{gdy} \quad edu <= 8, \\ 2, & \text{gdy} \quad 8 < edu <= 12, \\ 3, & \text{gdy} \quad edu > 12. \end{cases}$$

   - zmienną *pensja*, która opisuje miesięczne zarobki w tysiącach złotych (przyjmij, że miesiąc ma 22 ośmiogodzinne dni robocze i uwzględnij aktualny kurs dolara wobec złotówki).

3. Dla zmiennej *pensja* wyznacz następujące parametry charakteryzujące jej rozkład w próbie: średnią, medianę, pierwszy i trzeci kwartyl, rozstęp międzykwartylowy, wariancję, odchylenie standardowe, wartość najmniejszą i wartość największą.

4. Polecenia z punktu 3. wykonaj dla każdej z trzech prób, generowanych przez wartości zmiennej *wykszt*. Wyniki przedstaw w postaci **zbiorczej** tabelki, której kolejne wiersze zawierają wartości parametrów dla osób z wykształceniem podstawowym, średnim i wyższym. Następnie umieść **na jednym** rysunku box-ploty ilustrujące rozkład zmiennej *pensja*, odpowiadające tym próbom. Czy w którejś z prób pojawiają się obserwacje odstające? Jaki wpływ ma wykształcenie na pensję?

5. Narysuj histogram dla zmiennej *pensja*. Liczbę klas dobierz eksperymentalnie, tak aby rysunek wyglądał "dobrze". Skonstruuj także histogram z liczbą klas wybraną za pomocą **reguły Freedmana-Diaconisa**:

   $$\text{liczba klas} = \left\lceil \frac{x_{(n)} - x_{(1)}}{h} \right\rceil, \quad \text{gdzie } h = 2 \cdot \text{IQR} \cdot n^{-1/3}.$$

   Uwaga. IQR to rozstęp międzykwartylowy w próbie, $x_{(1)}$ i $x_{(n)}$ to najmniejsza i największa wartość w próbie, a symbol $\lceil x \rceil$ oznacza najmniejszą liczbę całkowitą większą lub równą x.

   (a) Czy rozkład w próbie zmiennej *pensja* jest symetryczny, czy też prawostronnie albo lewostronnie skośny?

   (b) Jaki rozkład ciągły dobrze przybliża rozkład zmiennej *pensja*?. Aby odpowiedzieć na to pytania porównaj kształt histogramu z wykresami kilku znanych Ci gęstości. Spośród tych gęstości wybierz tę, która „najlepiej" pasuje do histogramu. Jakie są parametry tej gęstości?

6. Zbuduj tablicę dwudzielczą dla zmiennych *rasa* i wykszt. Wykonaj dwie wersje tej tabeli: jedną tylko z liczebnościami i drugą zawierającą także udziały procentowe (percentages of total count).