Sleep efficiency prediciton

Zuzanna Nasiłowska, Maria Nowacka

Spis treści:

1. Wprowadzenie

Sen odgrywa kluczową rolę w naszym życiu, wpływając na zdrowie fizyczne, kondycję psychiczną lub ogólną jakość życia. Jako studenci często spotykamy się z problemem niedostatecznego snu, co jest nie tylko wynikiem intensywnego trybu życia pod kątem nauki oraz pracy, ale również wpływu różnych czynników, takich jak stres, nawyki żywieniowe czy używki. W rozmowach z naszymi kolegami wielokrotnie pojawia się temat problemów z zasypianiem, niskiej jakości snu czy odczuwania zmęczenia mimo przespanych godzin. Zainsporowało to nas do spojrzenia na zadane zagadnienie z perspektywy statystyki.

CEL ANALIZY

Głównym celem analizy naszego zadanego problemu jest zbadanie czynników wpływających na jakość snu, mierzoną jako jej efektywność. Podejście ze strony statystycznej pozowli nam uzyskać ciekawe spostrzeżenia, które pomogą nam w odpowiedzi na pytania dotyczące tego, jakie zmienne mogą być kluczowe w poprawie jakości snu studentów, ale również ludzi w różnym przedziale wiekowym.

1.1 Pochodzenie danych

Użyty przez nas w raporcie zestaw danych pt: “Sleep Efficiency Prediction” jest dostępny na platformie Kaggle.

- **źródło**: Kaggle (udostępnione przez użytkownika o nazwie Ishhjain)
- **licencja**: Brak informacji na stronie (Unknown)

1.2 Opis zmiennych

1) **ID**: Unikalny identyfikator każdego wpisu, jednostki brak, możliwe wartości: liczby całkowite, statystyki opisowe:

- **średnia**: 309.5
- **wartość minimalna**: 1
- **wartość maksymalna**: 610
- **odchylenie standardowe**: 178.55

2) **Age**: Wiek, jednostka: lata, możliwe wartości liczbowe około od 1 do 100, statystyki opisowe:

    - **średnia**: 40.34
    - **wartość minimalna**: 9
    - **wartość maksymalna**: 69
    - **odchylenie standardowe**: 13.08

3) **Gender**: Płeć, jednostka: brak, możliwe wartości: Famale (kobieta), Male (Mężczyzna).

4) **Bedtime**: Godzina położenia się spać, format: data i czas, jednostka: godzina i minuty.

5) **Wakeup time**: Godzina obudzenia się, Format: data i czas, jednostka: godziny i minuty.

6) **Sleep duration**: Czas trwania snu, jednostka: godziny, możliwe wartości: od 0 do 24, statystyki opisowe:

    - **średnia**: 7.45
    - **wartość minimalna**: 5
    - **wartość maksymalna**: 10
    - **odchylenie standardowe**: 0.84

7) **Sleep efficiency**: efektywność snu, jednostki brak, możliwe wartości: z przedziału (0,1), statystyki opisowe:

    - **średnia**: 0.79
    - **wartość minimalna**: 0.5
    - **wartość maksymalna**: 0.99
    - **odchylenie standardowe**: 0.13

8) **REM sleep percentage**: Procent snu REM, jednostka: procenty, możliwe wartości: od 0 do 100, statystyki opisowe:

    - **średnia**: 22.57
    - **wartość minimalna**: 15
    - **wartość maksymalna**: 30
    - **odchylenie standardowe**: 3.55

9) **Deep sleep percentage**: Procent snu głębokiego, jednostka: procenty, możliwe wartości: 0 do 100, statystyki opisowe:

- **średnia**: 53.16
- **wartość minimalna**: 18
- **wartość maksymalna**: 75
- **odchylenie standardowe**: 15.50

10) **Light sleep percentage**: Procent snu lekkiego, jednostka: procenty, możliwe wartości: 0 do 100, statystyki opisowe:

    - **średnia**: 24.27
    - **wartość minimalna**: 7
    - **wartość maksymalna**: 63
    - **odchylenie standardowe**: 15.11

11) **Awakenings**: Przebudzenia podczas snu, jednostka: liczba całkowita, możliwe wartości: od 0 w górę, statystyki opisowe:

    - **średnia**: 1.68
    - **wartość minimalna**: 0
    - **wartość maksymalna**: 4
    - **odchylenie standardowe**: 1.34

12) **Caffeine consumption**: Spożycie kofeiny, jednostka: miligramy, możliwe wartości: od 0 w górę, statystyki opisowe:

    - **średnia**: 24.53
    - **wartość minimalna**: 0
    - **wartość maksymalna**: 200
    - **odchylenie standardowe**: 32.35

13) **Alcohol consumption**: Spożycie alkoholu, jednostka: unjce, możliwe wartości: od 0 w górę, statystyki opisowe:

    - **średnia**: 1.12
    - **wartość minimalna**: 0
    - **wartość maksymalna**: 5
    - **odchylenie standardowe**: 1.60

14) **Smoking status**: Status palenia, możliwe wartości: “Yes” (pali) lub “No” (nie pali)

15) **Exercise frequency**: Częstotliwość ćwiczeń w tygodniu, jednostka: liczba dni, możliwe wartości: od 0 do 7, statystyki opisowe:

    - **średnia**: 1.78
    - **wartość minimalna**: 0
    - **wartość maksymalna**: 5
    - **odchylenie standardowe**: 1.41

PYTANIA BADAWCZE

W celu realizacji tematu skonstruowałyśmy kilka pytań badawczych:

- Jakie czynniki mają wpływ na efektywność snu (alkohol, kofeina, sport)?
- Czy istnieje wiązek między długością snu a efektywnością i strukturą?
- Jak różne grupy demograficzne różnią się pod względem snu?
- Czy ilość przebudzeń w ciągu nocy wpływa na jakość snu?
- Czy czas pójścia spać ma znaczenie?

2. Wczytanie danych

Pierwszym krokiem wprowadzającym do analizy danych będzie ich wprawidłowe wczytanie. Następnie musimy zadbać o odpowednie typy danych.

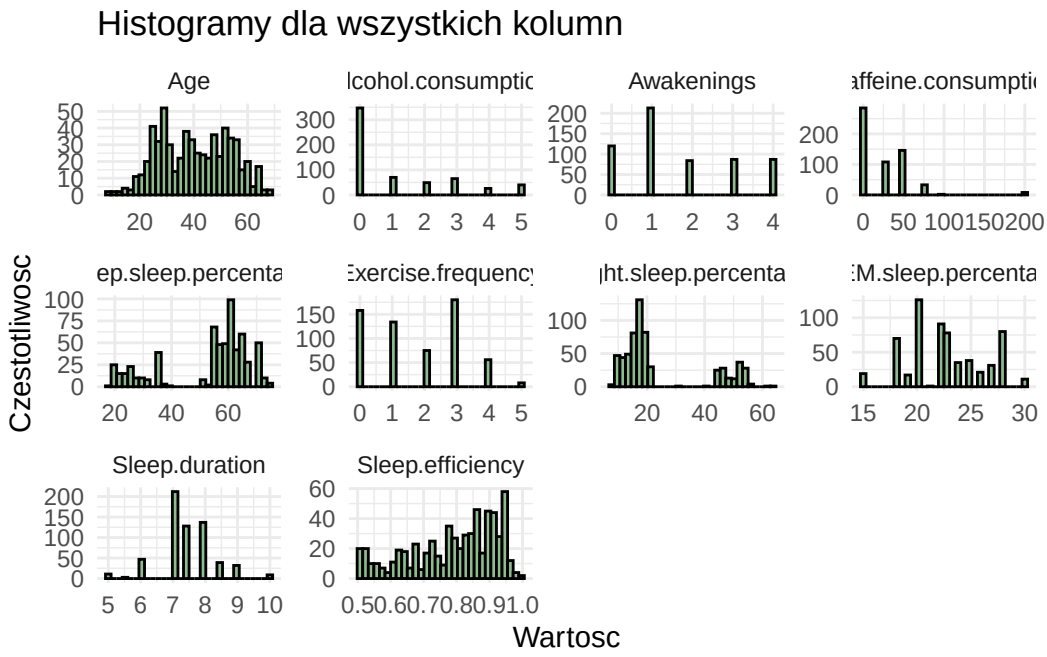
```
# A tibble: 4 x 2
  Column         DataType
  <chr>          <chr>
1 Gender         character
2 Bedtime        character
3 Wakeup.time    character
4 Smoking.status character
```

Możemy zauważyć, że wybrane kolumny mają typ zmiennych 'character'. Chcielibyśmy to zmienić, aby łatwiej się na nich pracowało. Po tej poprawce nowe typy danych są przedstawione w tabeli.

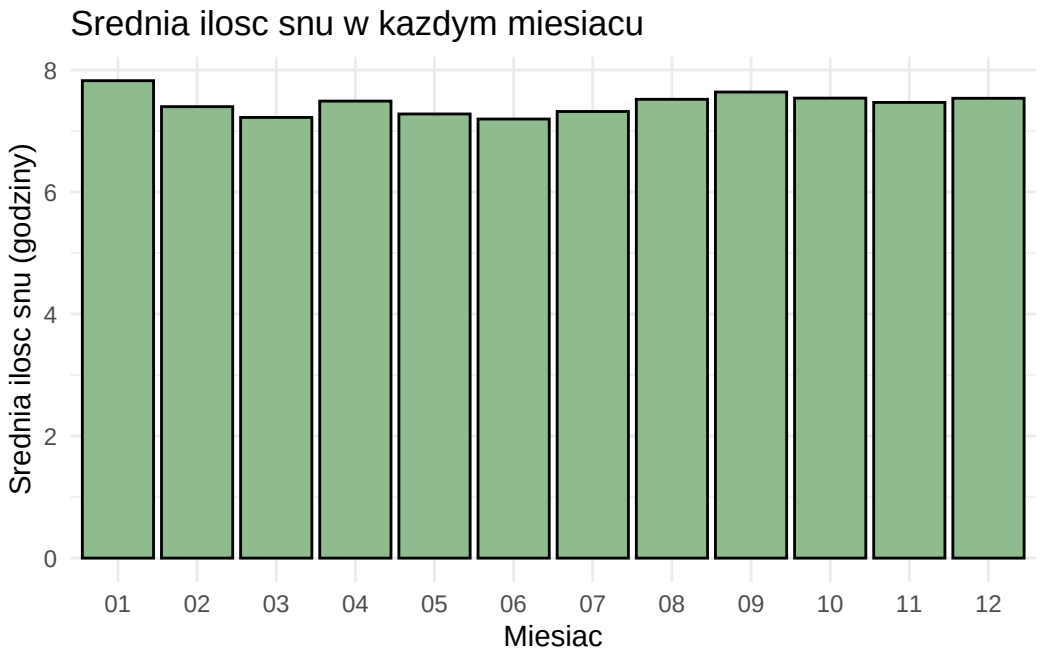
```
          Column DataType
1         Gender   factor
2        Bedtime  POSIXct
3    Wakeup.time  POSIXct
4 Smoking.status   factor
```

3. Analiza danych

Przyjrzyjmy się naszym danym, sprawdzając ich histogramy.

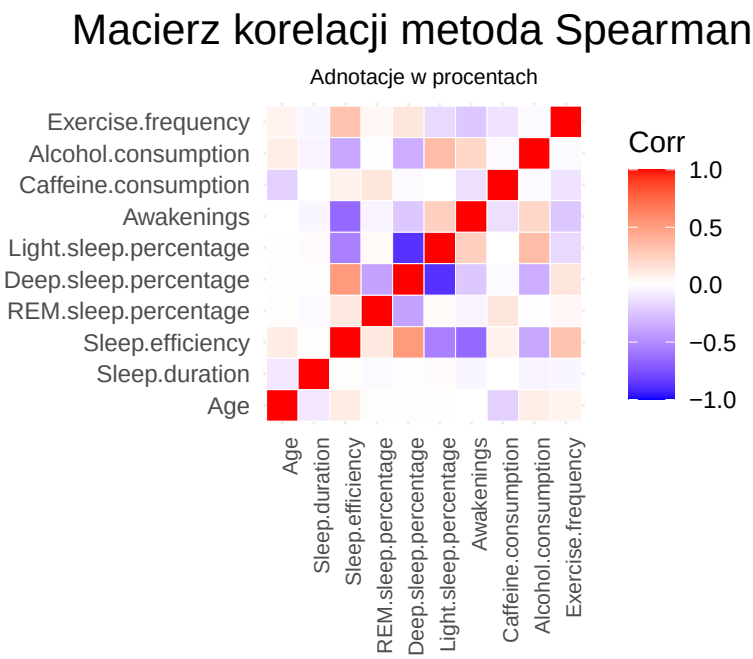


Dodatkowo sprawdzimy, czy pora roku ma związek z długością snu,

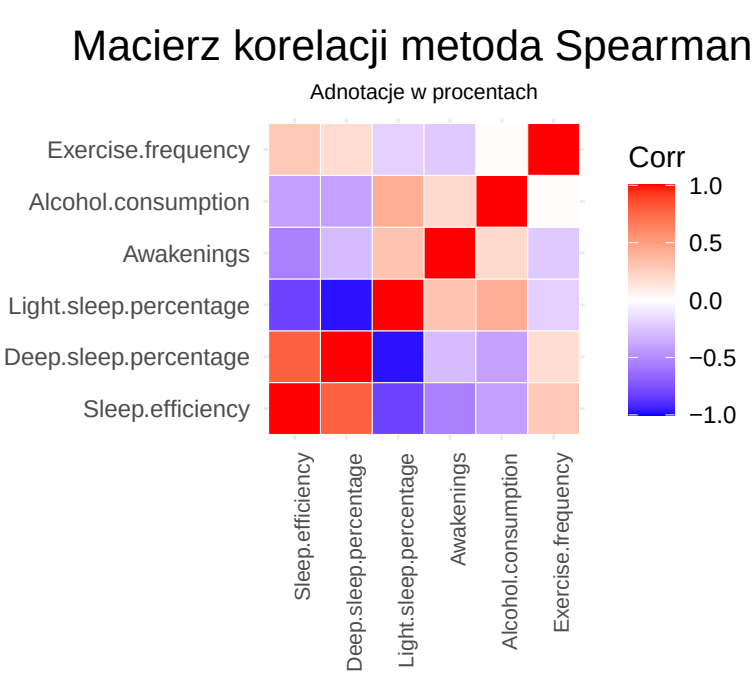


Choć mogłoby się wydawać, że jesienna i zimowa pogoda zachęcają do dłuższego snu, widoczne różnice są minimalne. Możemy jednak poszukać innych korelacji w naszych danych.

Posłużymy się w tym celu korelacją Spearmana.



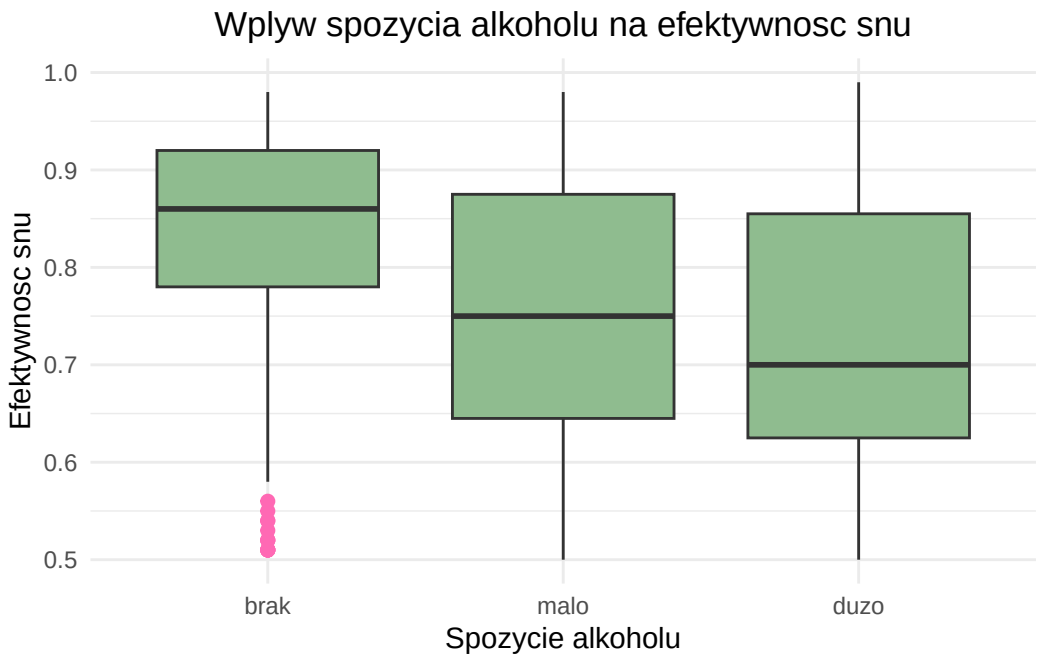
Wybierzemy teraz 6 najbardziej skorelowanych zmiennych i skupimy się na nich w dalszej analizie.



```r
data_better <- data_better[!is.na(data_better$Alcohol.consumption), ]

# Podział na kategorie
data_better$Alcohol.consumption <- cut(data_better$Alcohol.consumption,
                                       breaks = c(-Inf,0,2,Inf),
                                       labels = c("brak", "mało", "dużo"))

# Wizualizacja za pomocą ggplot2
library(ggplot2)
ggplot(data_better, aes(x = Alcohol.consumption, y = Sleep.efficiency)) +
  geom_boxplot(fill = "darkseagreen", outlier.color = "hotpink", outlier.size = 2) +
  labs(title = "Wpływ spożycia alkoholu na efektywność snu",
       x = "Spożycie alkoholu",
       y = "Efektywność snu") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5)
  )
```

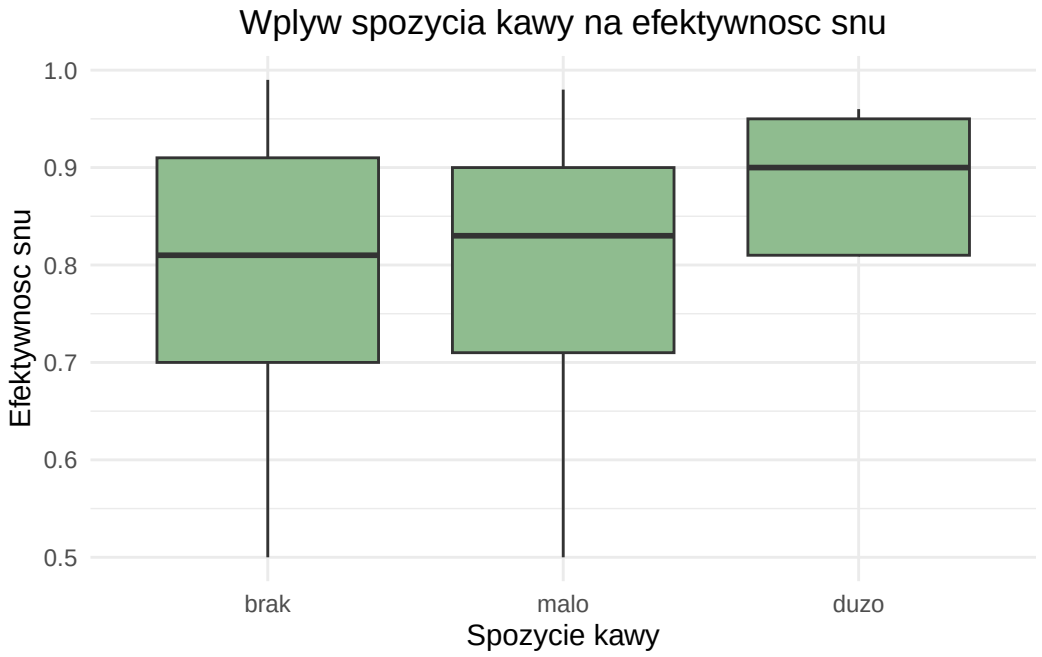



```r
data_better <- data_better[!is.na(data_better$Caffeine.consumption), ]

# Podział na kategorie
```

```r
data_better$Caffeine.consumption <- cut(data_better$Caffeine.consumption,
                                        breaks = c(-Inf,0,99,Inf),
                                        labels = c("brak", "mało", "dużo"))

# Wizualizacja za pomocą ggplot2
library(ggplot2)
ggplot(data_better, aes(x = Caffeine.consumption, y = Sleep.efficiency)) +
  geom_boxplot(fill = "darkseagreen", outlier.color = "hotpink", outlier.size = 2) +
  labs(title = "Wpływ spożycia kawy na efektywność snu",
       x = "Spożycie kawy",
       y = "Efektywność snu") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5)
  )
```

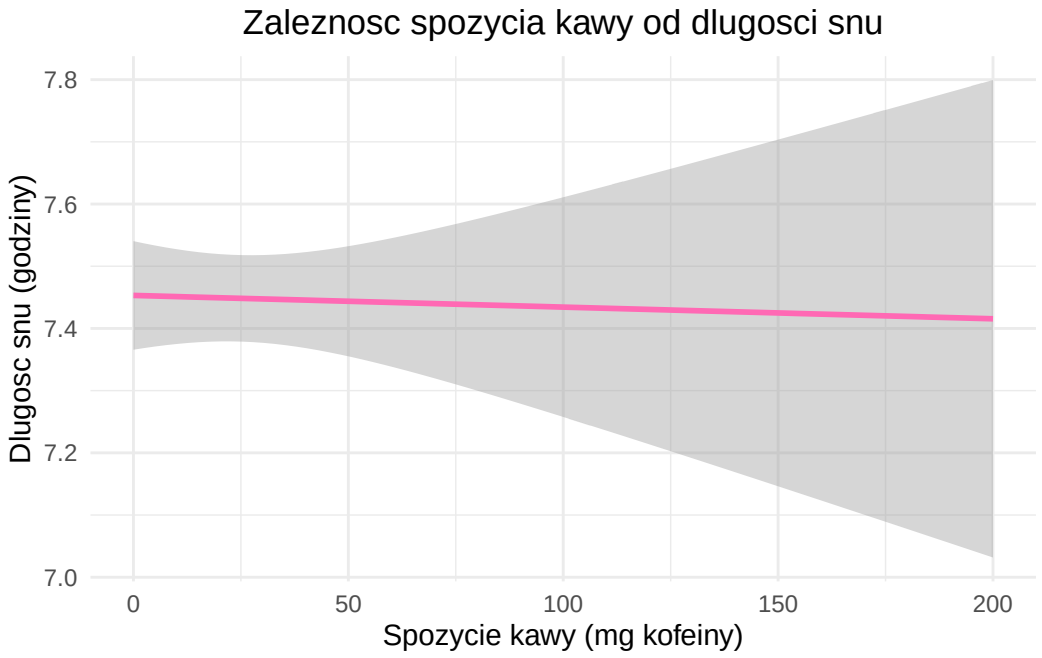



```r
data <- read.csv('sleep.csv')
data_better <- data %>%
  mutate(
    Caffeine.consumption = as.numeric(Caffeine.consumption)
  )
ggplot(data_better, aes(x = Caffeine.consumption, y = Sleep.duration)) +
```

```
  geom_smooth(method = "lm", color = "hotpink") +
  labs(title = "Zależność spożycia kawy od długości snu",
       x = "Spożycie kawy (mg kofeiny)",
       y = "Długość snu (godziny)") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5)
  )
```

```
`geom_smooth()` using formula = 'y ~ x'
```

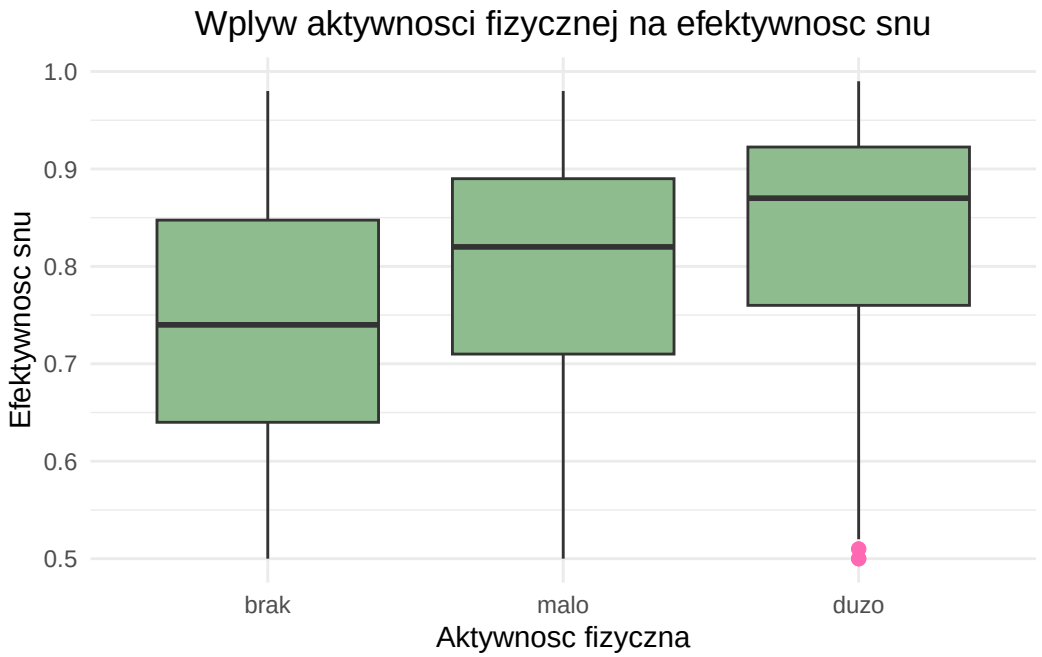
```
Warning: Removed 36 rows containing non-finite outside the scale range
(`stat_smooth()`).
```




```
data_better <- data_better[!is.na(data_better$Exercise.frequency), ]

# Podział na kategorie
data_better$Exercise.frequency <- cut(data_better$Exercise.frequency,
                                      breaks = c(-Inf,0,2,Inf),
                                      labels = c("brak", "mało", "dużo"))
```

```r
# Wizualizacja za pomocą ggplot2
library(ggplot2)
ggplot(data_better, aes(x = Exercise.frequency, y = Sleep.efficiency)) +
  geom_boxplot(fill = "darkseagreen", outlier.color = "hotpink", outlier.size = 2) +
  labs(title = "Wpływ aktywności fizycznej na efektywność snu",
       x = "Aktywność fizyczna",
       y = "Efektywność snu") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5)
  )
```




```r
# Usunięcie braków danych w kolumnie Smoking.status i Sleep.efficiency (jeśli są)
data_better <- data_better[!is.na(data_better$Smoking.status) & !is.na(data_better$Sleep.effi

# Tworzenie boxplota
ggplot(data_better, aes(x = Smoking.status, y = Sleep.efficiency)) +
  geom_boxplot(fill = "darkseagreen", outlier.color = "hotpink", outlier.size = 2) +
  labs(title = "Wpływ palenia na efektywność snu",
       x = "Status palenia (yes = pali, no = nie pali)",
       y = "Efektywność snu") +
  theme_minimal() +
```

```
  theme(
    plot.title = element_text(hjust = 0.5)
  )
```