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 przed snem, 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 przed snem, jednostka: miligramy, 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

```
# Przekształć dane
data_better <- data %>%
  mutate(
    Smoking.status = as.factor(Smoking.status),
    Gender = as.factor(Gender),
    Bedtime = as.POSIXct(Bedtime, format = "%d/%m/%Y %H:%M"),
    Wakeup.time = as.POSIXct(Wakeup.time, format = "%d/%m/%Y %H:%M"),
    #Sleep.duration = difftime(Wakeup.time, Bedtime, units = "hours"),
    Month = format(Bedtime, "%m")
  )
```

3. Analiza danych

```
# Przefiltruj dane, aby usunąć ujemne lub błędne wartości
#data_clean <- data_better %>%
#  filter(Sleep.duration > 0)  # Usuwa rekordy z ujemnymi wartościami lub zerami

# Oblicz średnią długość snu dla każdego miesiąca i wyświetl dane
average_sleep <- data_better %>%
  group_by(Month) %>%
  summarise(Average.sleep = mean(Sleep.duration, na.rm = TRUE))

# Sprawdź obliczone średnie (opcjonalne)
print(average_sleep)
```

```
# A tibble: 12 x 2
   Month Average.sleep
   <chr>         <dbl>
```

```
 1 01              7.82
 2 02              7.40
 3 03              7.22
 4 04              7.49
 5 05              7.28
 6 06              7.20
 7 07              7.32
 8 08              7.52
 9 09              7.64
10 10              7.54
11 11              7.47
12 12              7.54
```

```r
# Narysuj wykres słupkowy
average_sleep %>%
  ggplot(aes(x = Month, y = Average.sleep)) +
  geom_col(fill = "skyblue") +  # Wykres słupkowy
  labs(
    title = "Średnia ilość snu w każdym miesiącu",
    x = "Miesiąc",
    y = "Średnia ilość snu (godziny)"
  ) +
  theme_minimal()
```

```r
library(dplyr)
```

```r
# Zmiana wartości w kolumnie Gender (Male -> 1, Female -> 0)
#dataframe <- data %>%
#  mutate(Gender = ifelse(Gender == "Male", 1, 0))

# Zmiana wartości w kolumnie Smoking status (Yes -> 1, No -> 0)
#dataframe <- dataframe %>%
#  mutate(Smoking.status = ifelse(Smoking.status == "Yes", 1, 0))

# Konwersja Bedtime i Wakeup time na format daty i czasu
#dataframe <- dataframe %>%
#  mutate(
#    Bedtime = as.POSIXct(Bedtime, format = "%d/%m/%Y %H:%M"),
#    Wakeup.time = as.POSIXct(Wakeup.time, format = "%d/%m/%Y %H:%M")
#  )

# Konwersja Bedtime i Wakeup time na liczbę sekund od epoki (UNIX timestamp)
#dataframe <- dataframe %>%
#  mutate(
#    Bedtime = as.numeric(as.POSIXct(Bedtime)),
#    Wakeup.time = as.numeric(as.POSIXct(Wakeup.time))
#  )
```

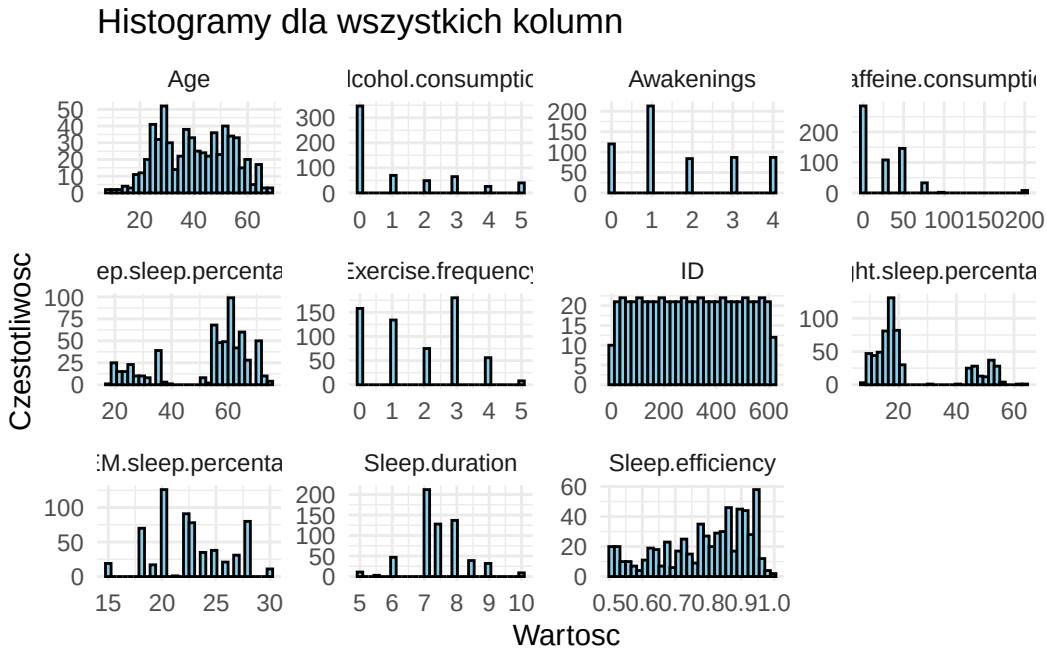
```r
# Przeniesienie kolumny "Sleep efficiency" na koniec
#cols <- colnames(data_better)
#data_better <- data_better %>%
#  select(all_of(setdiff(cols, "Sleep.efficiency")), "Sleep.efficiency")

# Usunięcie kolumny ID
```

```r
# Rysowanie histogramów dla wszystkich kolumn
library(ggplot2)
# Wyciąganie tylko zmiennych numerycznych
data_numeric <- data_better %>%
  select_if(is.numeric)

# Wyświetlenie pierwszych kilku wierszy z danych numerycznych
head(data_numeric)
```

```
  ID Age Sleep.duration Sleep.efficiency REM.sleep.percentage
1  1  65            6.0              0.88                   18
2  2  69            7.0              0.66                   19
3  3  40            8.0              0.89                   20
4  4  40            6.0              0.51                   23
5  5  57            8.0              0.76                   27
6  6  36            7.5              0.90                   23
  Deep.sleep.percentage Light.sleep.percentage Awakenings Caffeine.consumption
1                    70                     12          0                    0
2                    28                     53          3                    0
3                    70                     10          1                    0
4                    25                     52          3                   50
5                    55                     18          3                    0
6                    60                     17          0                   NA
  Alcohol.consumption Exercise.frequency
1                   0                  3
2                   3                  3
3                   0                  3
4                   5                  1
5                   3                  3
6                   0                  1
```

```r
data_numeric %>%
  gather(key = "Variable", value = "Value") %>%
  ggplot(aes(x = Value)) +
  facet_wrap(~Variable, scales = "free") +
  geom_histogram(bins = 30, fill = "skyblue", color = "black") +
  theme_minimal() +
  labs(title = "Histogramy dla wszystkich kolumn", x = "Wartość", y = "Częstotliwość")
```

```
Warning: Removed 91 rows containing non-finite outside the scale range
(`stat_bin()`).
```

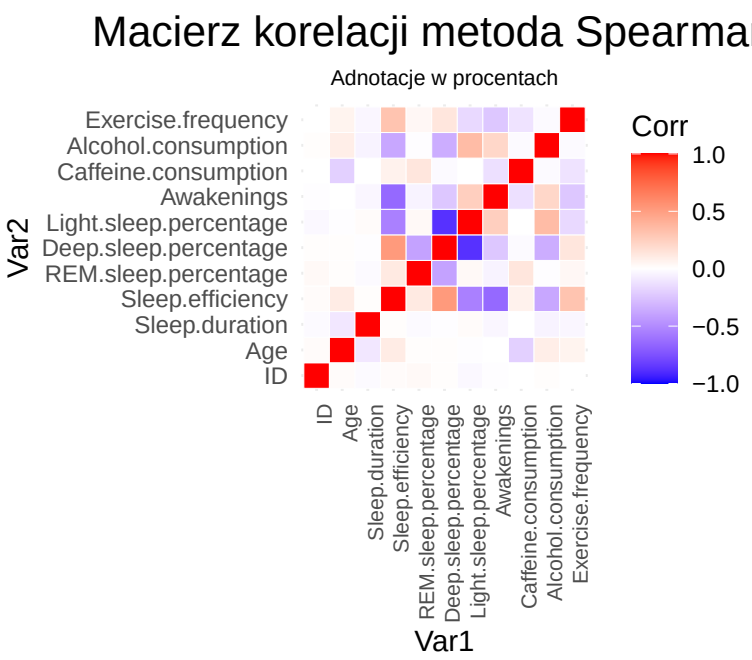



```
library(ggplot2)
library(ggcorrplot)
```

```
Warning: pakiet 'ggcorrplot' został zbudowany w wersji R 4.4.2
```

```
# Obliczenie macierzy korelacji metodą Spearmana
correlation_matrix <- cor(data_numeric, method = "spearman", use = "complete.obs")

# Rysowanie mapy cieplnej z adnotacjami
ggcorrplot(
  correlation_matrix,
  lab = FALSE,                # nie chcemy Wyświetlenie wartości korelacji na wykresie
  outline.color = "white"      # Białe obramowanie komórek
) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 16),
    plot.subtitle = element_text(hjust = 0.5, size=8),
    axis.text.x = element_text(angle = 90, hjust = 1, size = 8)  # Obrót podpisów osi X

  ) +
  labs(
```

```
    title = "Macierz korelacji metodą Spearmana",
    subtitle = "Adnotacje w procentach"
  )
```

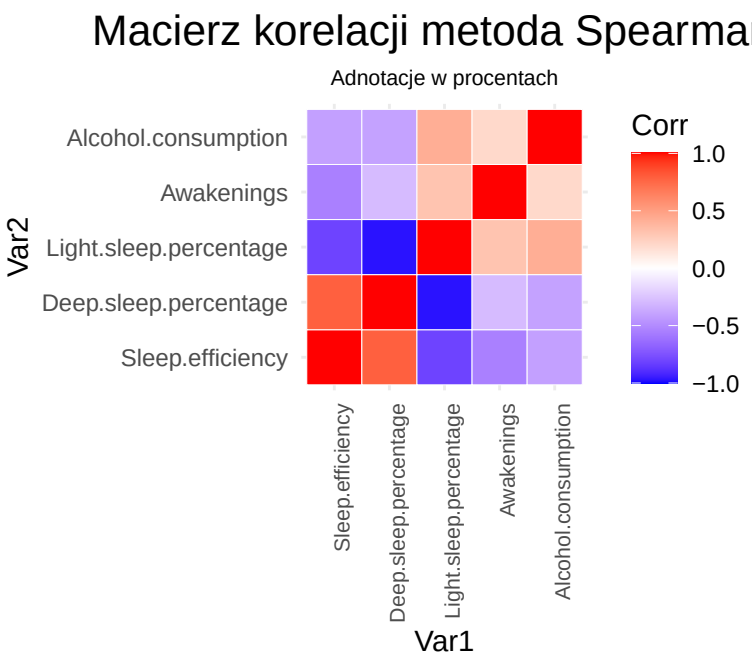



```
library(dplyr)

# 1. Obliczenie pełnej macierzy korelacji dla zmiennych numerycznych
cor_matrix <- cor(data_better %>% select_if(is.numeric), use = "complete.obs")

# 2. Przemieszczenie górnej części macierzy korelacji, aby sprawdzić sumy korelacji dla każdej
cor_sum <- apply(cor_matrix, 2, function(x) sum(abs(x), na.rm = TRUE))

# 3. Wyciągnięcie 5 kolumn o najwyższej sumie korelacji
top_5_cols <- names(sort(cor_sum, decreasing = TRUE)[1:5])

# 4. Tworzenie nowej macierzy korelacji tylko dla tych 5 kolumn
cor_matrix_top_5 <- cor_matrix[top_5_cols, top_5_cols]

# Wyświetlenie nowej macierzy korelacji
cor_matrix_top_5
```

```
                    Sleep.efficiency Deep.sleep.percentage
```

```
Sleep.efficiency              1.0000000             0.7785051
Deep.sleep.percentage         0.7785051             1.0000000
Light.sleep.percentage       -0.8095696            -0.9749998
Awakenings                   -0.5510386            -0.2920058
Alcohol.consumption          -0.4084527            -0.4035995
                      Light.sleep.percentage Awakenings Alcohol.consumption
Sleep.efficiency                  -0.8095696 -0.5510386          -0.4084527
Deep.sleep.percentage             -0.9749998 -0.2920058          -0.4035995
Light.sleep.percentage             1.0000000  0.3093659           0.4162357
Awakenings                         0.3093659  1.0000000           0.1995814
Alcohol.consumption                0.4162357  0.1995814           1.0000000
```

```r
# Rysowanie mapy cieplnej z adnotacjami
ggcorrplot(
  cor_matrix_top_5,
  lab = FALSE,                # nie chcemy Wyświetlenie wartości korelacji na wykresie
  outline.color = "white"      # Białe obramowanie komórek
) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 16),
    plot.subtitle = element_text(hjust = 0.5, size=8),
    axis.text.x = element_text(angle = 90, hjust = 1, size = 8)  # Obrót podpisów osi X

  ) +
  labs(
    title = "Macierz korelacji metodą Spearmana",
    subtitle = "Adnotacje w procentach"
  )
```


Macierz korelacji metoda Spearmana
Adnotacje w procentach
Alcohol.consumption
Awakenings
Light.sleep.percentage
Deep.sleep.percentage
Sleep.efficiency
Var2
Sleep.efficiency
Deep.sleep.percentage
Light.sleep.percentage
Awakenings
Alcohol.consumption
Var1
Corr
1.0
0.5
0.0
−0.5
−1.0