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.

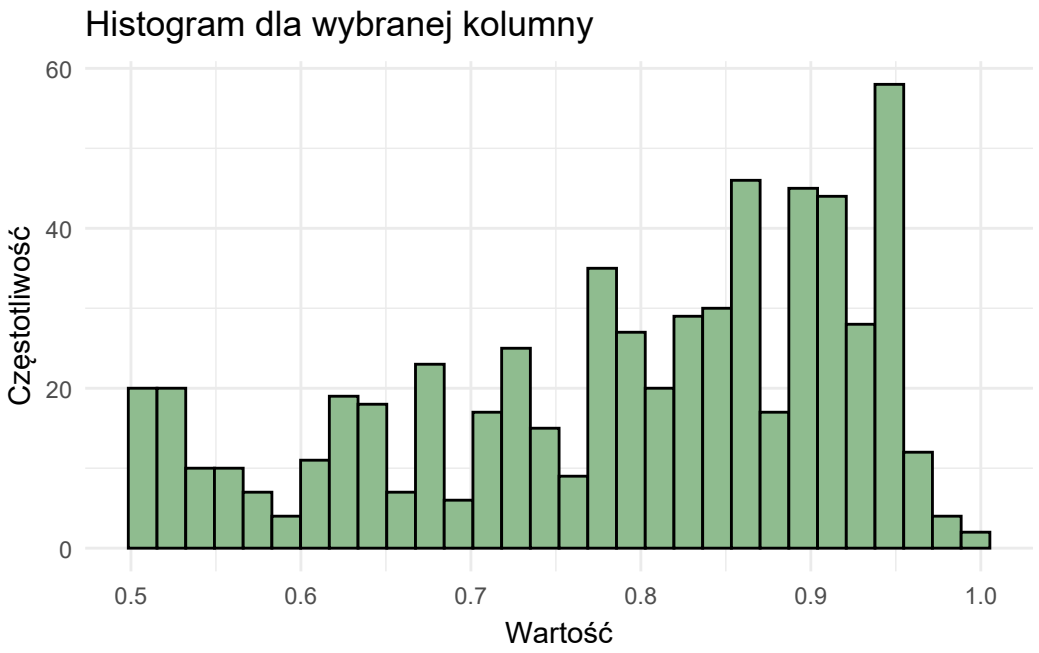
```
          Column  DataType
1         Gender character
2        Bedtime character
3    Wakeup.time character
4 Smoking.status character
```

Możemy zauważyć, że wybrane kolumny mają typ zmiennych 'character'. Chcielibyśmy to zmienić, aby łatwiej się na nich pracowało. Zmienione typy danych widzimy w tabeli.

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

3. Analiza danych

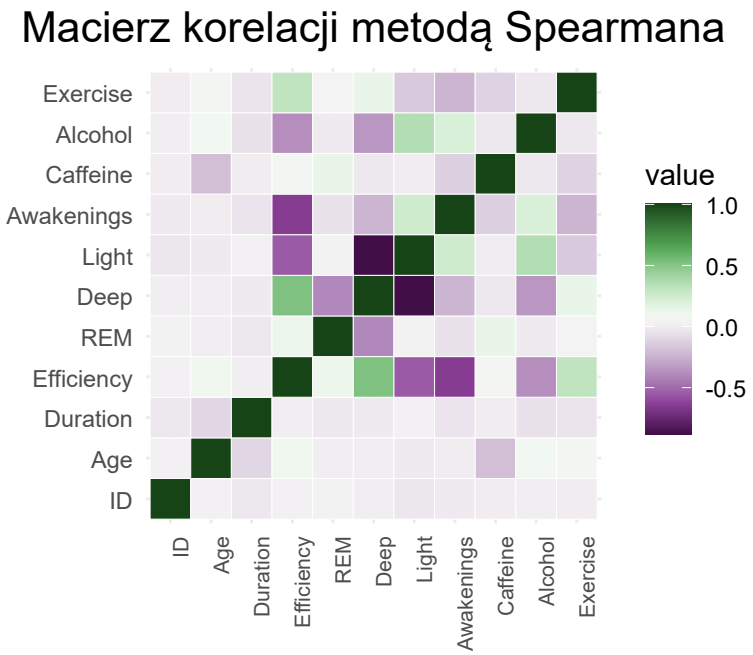
Na początek spójrzmy, jak rozkłada się się zmienna "Sleep.efficiency" wśród naszych danych.



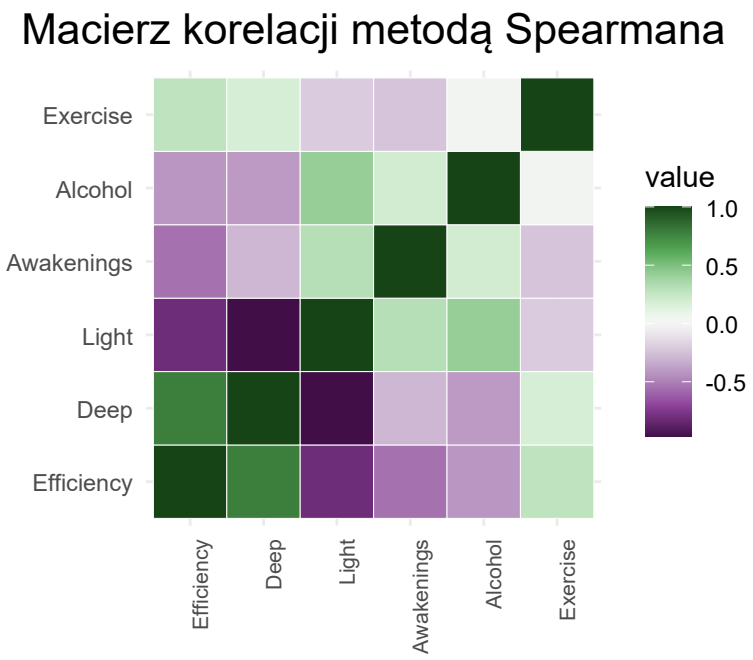
Z histogramów możemy odczytać, że większość osób dosyć efektywnie się wysypia. Co jednak wpływa na niższą efektywność snu u innych osób? Użyjemy narzędzi statystycznych, aby odpowiedzieć na postawione wcześniej pytania badawcze.

Korelacja

Sprawdźmy, które zmienne są ze sobą skorelowane. Wykorzystujemy metodę Spearmana.



Widzimy, że niektóre zmienne są od siebie ściśle zależne, a inne nie wpływają na siebie nawzajem. Wybierzemy teraz 6 najbardziej skorelowanych zmiennych i skupimy się na nich w dalszej analizie.



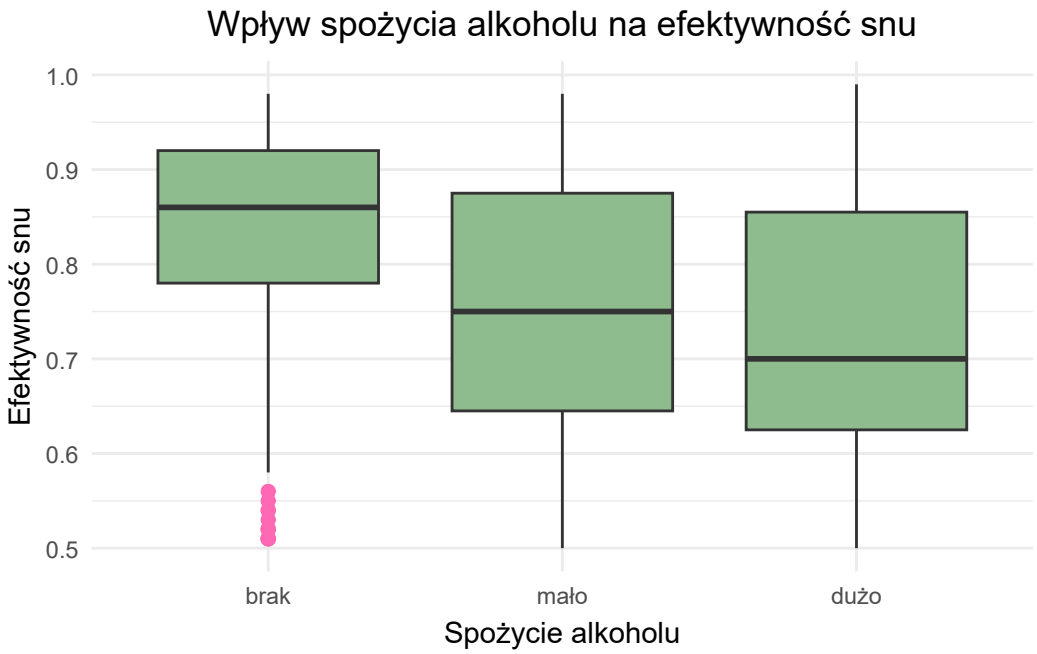
Efektywność snu, definiowana jako stosunek czasu spędzonego w fazie snu do całkowitego

czasu spędzonego w łóżku, jest istotnym wskaźnikiem jakości wypoczynku. Szczególnie w przypadku studentów istotna może okazać się analiza czynników, które posiadają zasadniczy wpływ w tym zakresie. Korzystając z danych zawartych w naszym zestawi, możemy postawić tezę, że czynniki takie jak spożycie alkoholu i kawy, palenie papierosów czy uprawianie sportu będą posiadały takowy wpływ.

- **Konsumpcja alkoholu** – alkohol często wpływa na strukturę snu, w szczególności fazę REM, co może obniżać jego efektywność.
- **Spożycie kawy** – kofeina jest znanym stymulantem, który może utrudniać zasypianie lub zmieniać długość faz snu.
- **Palenie tytoniu** – nikotyna, jako substancja psychoaktywna, również może wpływać na czas zasypiania oraz jakość snu.
- **Aktywność fizyczna** – regularne ćwiczenia mogą przyczyniać się do lepszej jakości snu.

Przeprowadzimy analizę z wykorzystaniem boxplotów, które pozwolą nam zobrazować rozkład danych oraz wychwycić różnice i potencjalne zależności pomiędzy podanymi parametrami. Boxploty umożliwią również identyfikację ewentualnych wartości odstających, które mogą dostarczyć dodatkowych informacji na temat analizowanych zmiennych.

Najpierw skupimy się na omówieniu wpływu spożywania alkoholu w ciągu dnia. W dostępnych danych spożycie alkoholu zostało wyrażone w uncjach (oz). Jedna uncja odpowiada około 28,35 gramów, co pozwala łatwo przeliczyć wartości na bardziej znane jednostki, takie jak miligramy. Wartości w danych mieszczą się w zakresie od 0 do 5. Na potrzeby analizy podzielimy je na cztery kategorie: brak spożywania (0), małą ilość (1-2), dużą ilość (3-5) . Wykres pudełkowy prezentuje się następująco:



Opis wykresu

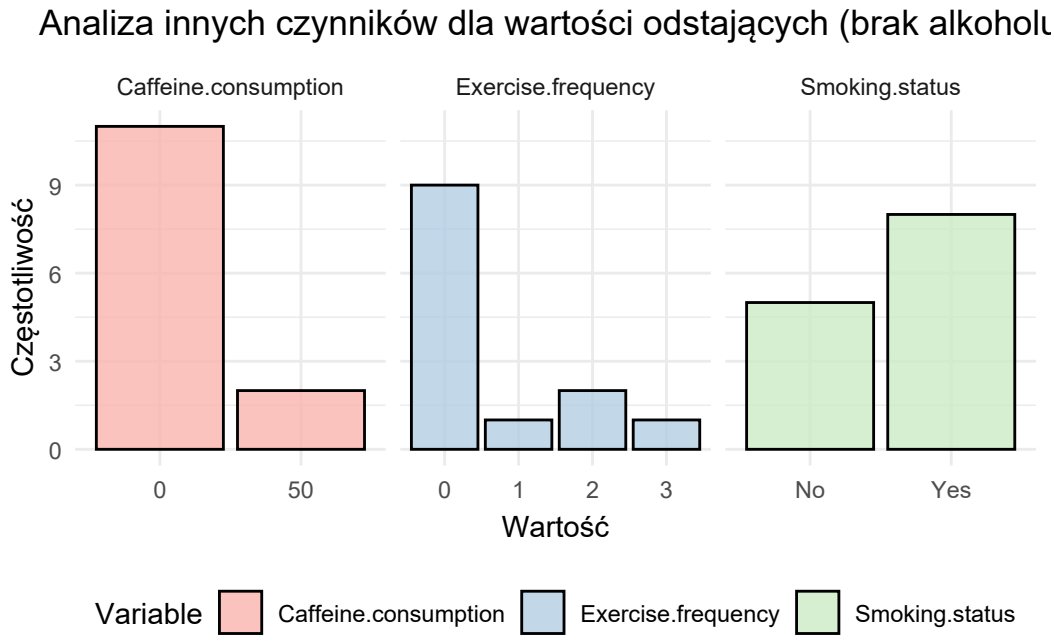
- ***K*ategoria brak:**
- Efektywność w tej grupie jest najwyższa, z medianą około 0,85.
- Rozkład jest dość zwarty, co oznacza mniejsze zróżnicowanie efektywności snu w tej grupie.
- Wartości odstające wskazują pewne przypadki wyjątkowo niskiej efektywności.
- **Kategoria mało:**
- Rozstęp międzykwartylowy jest większy, co sugeruje większe zróżnicowanie w efektywności snu w tej kategorii.
- Mediana efektywności snu w tej grupie wynosi 0,75 i jest mniejsza w porównaniu z kategorią "brak".
- **Kategoria dużo:**
- Efektywność snu w tej grupie jest najniższa (z medianą 0,7), a rozstęp międzykwartylowy jest nieco większy niż w kategorii "mało".

Na podstawie przedstawionych boxplotów dochodzimy do wniosku, że alkohol jest jednym z istotnych czynników wpływających na jakość i efektywność naszego snu. Im więcej alkoholu spożyjemy w ciągu dnia, tym ta efektywność będzie niższa. Jeśli przyjrzymy się rożnicy w medianie dla kategorii "brak" oraz "dużo", możemy zauważyć, że wynosi ona około 0,15.

Warto jednak ponownie zwrócić uwagę na odstające wartości w grupie "brak", które sugerują nam, że na niską efektywność spory wpływ muszą mieć jeszcze jakies czynniki. Możemy łatwo sprawdzić, które dokładnie aspekty wpłynęły aż tak na obniżenie jakości snu naszych wartości odstających. Wyniki prezentują się następująco:

```
 [1] "ID"                    "Age"                   "Gender"
 [4] "Bedtime"               "Wakeup.time"           "Sleep.duration"
 [7] "Sleep.efficiency"      "REM.sleep.percentage"  "Deep.sleep.percentage"
[10] "Light.sleep.percentage" "Awakenings"            "Caffeine.consumption"
[13] "Alcohol.consumption"   "Smoking.status"        "Exercise.frequency"
[16] "Month"
```

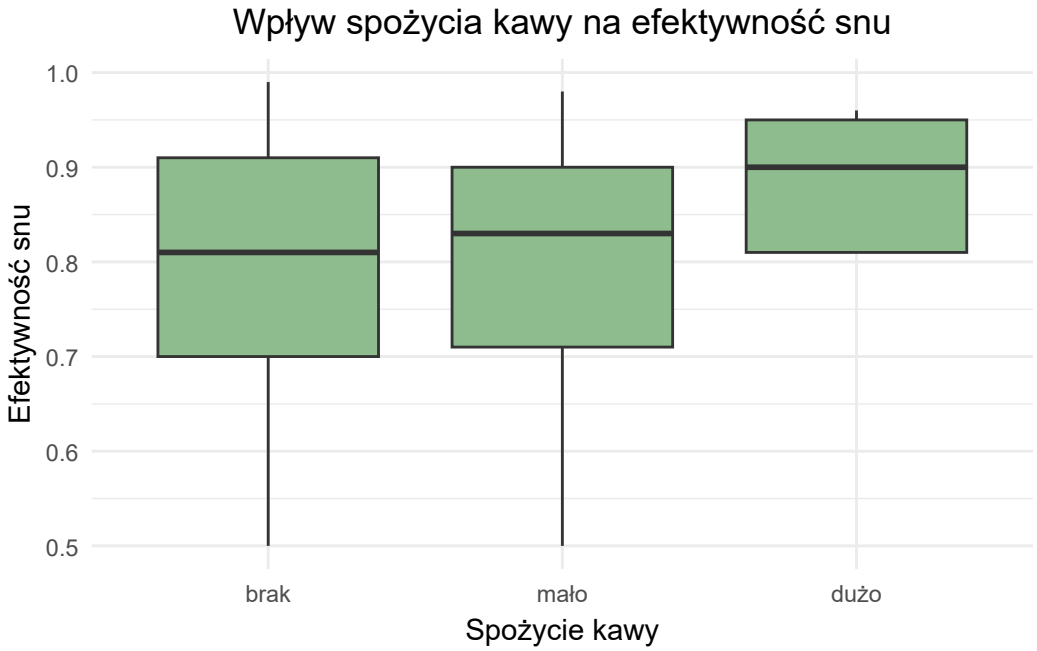
```
Warning: attributes are not identical across measure variables; they will be
dropped
```



Na podstawie powyższego wykresu widać, że większość osób, mimo nie picia alkoholu, pali papierosy lub wykazuje ograniczoną aktywność fizyczną. Interesujący jest aspekt kawy, ponieważ większość nie spożywa jej w ogóle lub w małych ilościach. Może wskazywać to na fakt, iż picie kawy nie odgrywa tak dużej roli pod kątem jakości snu, na tyle innych czynników.

W związku z powyższymi obserwacjami warto dokładniej przeanalizować parametry, takie jak picie kawy, aktywność fizyczna i palenie papierosów, aby upewnić się, czy wyciągnięte wnioski są słuszne i znajdują potwierdzenie w danych.

Najpierw skupimy się na aspekcie picia kawy w ciągu dnia. Według powyższych wniosków, nie powinna mieć ona, aż tak dużego wpływu na efektywność snu naszych badanych, co stanowi ciekawą obserwacje. Kofeina, znana ze swoich właściwości pobudzających, często kojarzona jest z potencjalnym zakłócaniem snu, zwłaszcza w przypadku studentów, którzy piją duże ilości kawy. Może to prowadzić do błędnego przekonania, że to właśnie kawa ma największy wpływ na obniżoną efektywność snu, co wymaga dokładniejszej analizy, aby zweryfikować te założenia. W tym celu również utworzymy boxplota i przeanalizujemy wyniki:



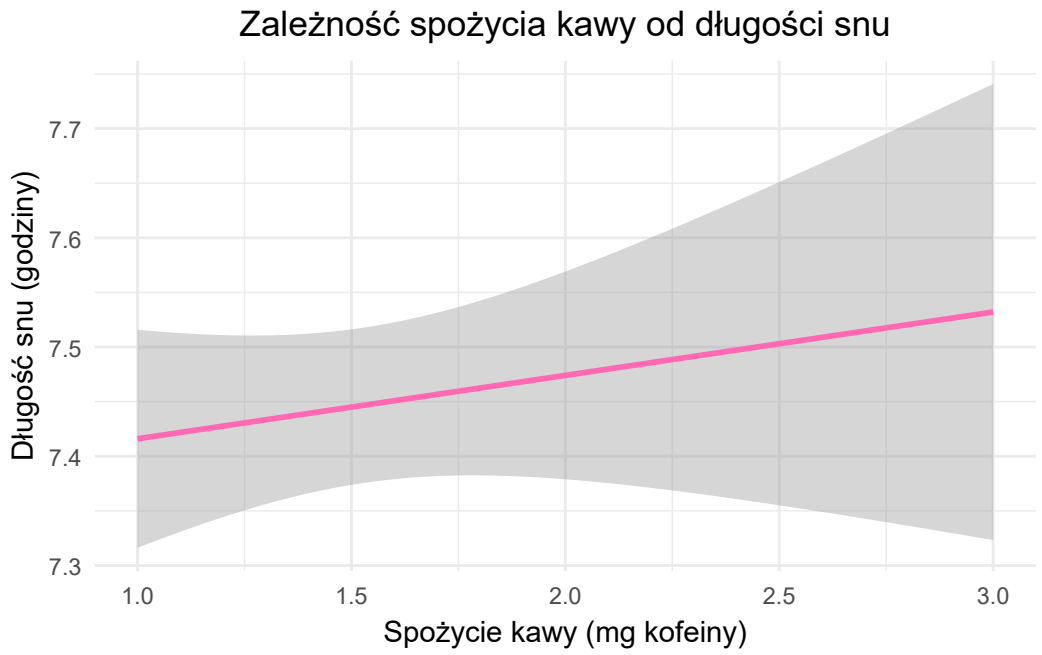
Wnioski:

- Brak jednoznacznego negatywnego wpływu kawy na efektywność snu.
- Niewielkie różnice między grupami:
- Między grupami "brak" i "mało" różnica jest niewielka, co sugeruje, że umiarkowane spożycie kawy nie wpływa wyraźnie na efektywność snu
- Zmienność w grupie "brak":
- Wzkazywać to może na znaczący wpływ innych czynników (alkohol, palenie, aktywność fizyczna

```
Okazuje się, że relacja między spożywa
```

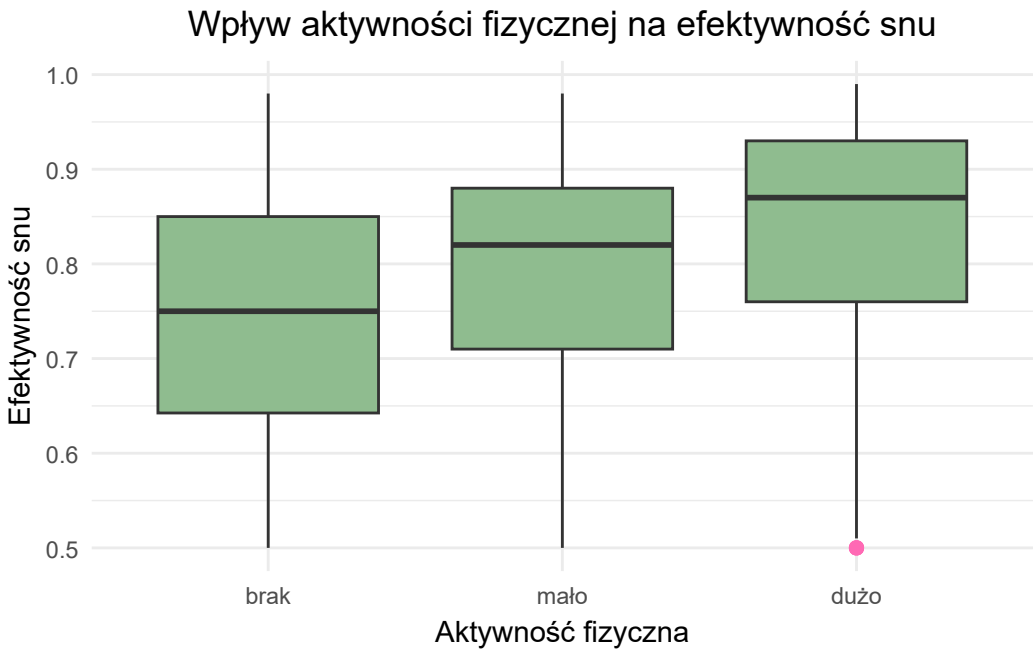
```
Call:
```

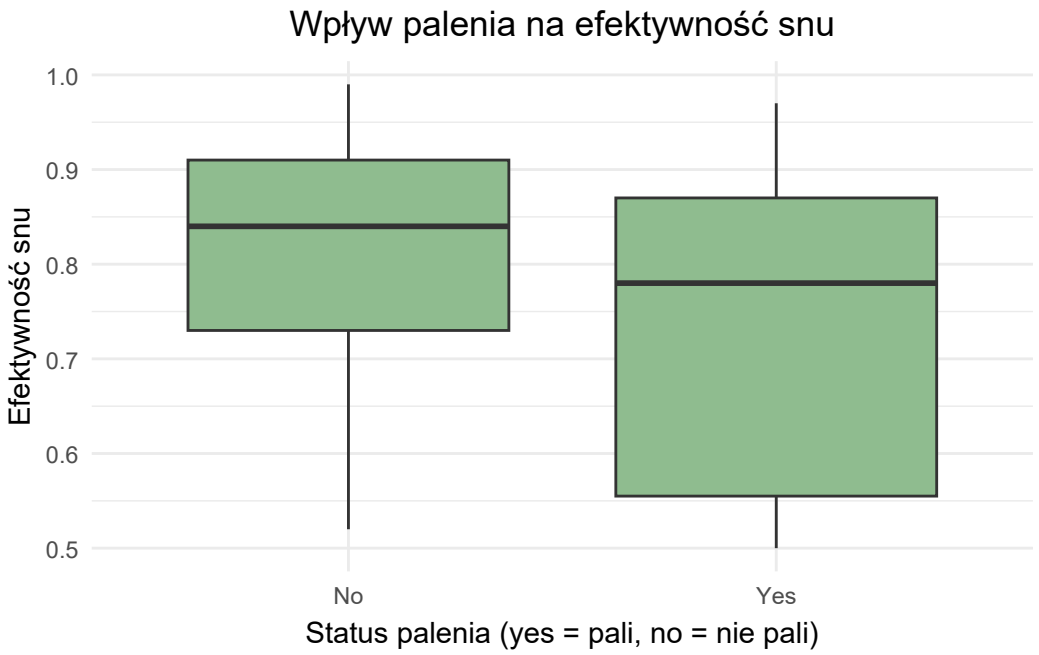
```
lm(formula = Sleep.duration ~ Caffeine.consumption, data = data_better)

Residuals:
     Min       1Q   Median       3Q      Max
-2.47403 -0.41602  0.02597  0.52597  2.58398

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          7.35800    0.10961  67.131   <2e-16 ***
Caffeine.consumption 0.05802    0.06780   0.856    0.393
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8573 on 559 degrees of freedom
Multiple R-squared:  0.001308,  Adjusted R-squared:  -0.0004785
F-statistic: 0.7322 on 1 and 559 DF,  p-value: 0.3926


Nachylenie prostej (slope): 0.0580165

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



Wykres na pierwszy rzut oka nie daje nam wielu informacji, ale w celu dokładniejszej analizy możemy użyć funkcji *summary*, która w R jest używana do generowania szczegółowego podsumowania obiektów. Przy pomocy *lm()* wyliczymy model liniowy, a wymieniona wcześniej funkcja zwróci nam konkretne informacje:

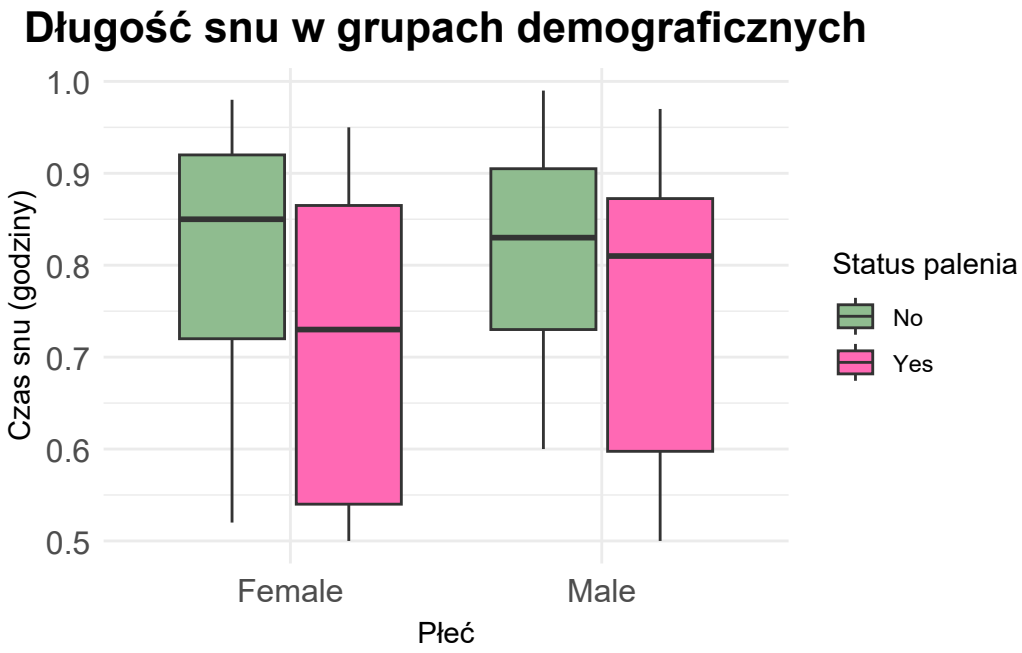
- **Residuals**: statystyki dotyczące reszt.
- **Coefficients**: współczynniki regresji:
- Slope (nachylenie): wpływ zmiennej niezależnej na zmienną zależną.
- Intercept





Na powyższym boxplocie widać, że palenie nie wpływa na efektywność snu. Ale czy na pewno? Sprawdźmy tę zależnoćś, dzieląc dodatkowo grupy palących i niepalących ze względu na płeć.

```
# A tibble: 4 x 4
  Gender Smoking.status Avg_Sleep_Duration Avg_Sleep_Efficiency
  <fct>  <fct>                      <dbl>                <dbl>
1 Female No                          7.48                0.816
2 Female Yes                         7.49                0.712
3 Male   No                          7.39                0.818
4 Male   Yes                         7.41                0.755
```



Po podziale obu grup ze względu na płeć widzimy, że palenie nie ma większego udziału w efektywności snu u mężczyzn, za to widocznie wpływa na kobiety. Jakie inne zależności nam umknęły ze względu na brak podziału naszeg grupy badawczej na kobiety i mężczyzn?

coś nie działa Z podanych boxplotów możemy zauważyć, że północ jest pewną granicą, po przekroczeniu której mamy widoczne zmiany w strukturze snu.

- Dla osób kładących się spać przed północą boxploty są bardziej zwarte.
- Dla pozostałych osób widzimy większy rozstrzał wartości. Oznacza to, że kładąc się później mamy statystycznie większe szanse na niewyspanie się.
- Dodatkowo możemy zauważyć, że faza REM jest najdłuższa (procentowo) w grupie osób, które kładą się przed 22.

_ Skupmy się jeszcze na oberwacjach odstających przy boxplocie długości snu głębokiego (Deep) dla osób kładących się spać przed północą. Dlaczego te osoby mają tak krótką tą fazę snu?

coś nie działa v2

Badając tą grupę ludzi możemy zauważyć, że większość tych osób budzi się w ciągu nocy. Takie pobudki przeszkadzają w zapadnięciu w sen głęboki.

W tym miejscu postaramy się odpowiedzieć na pytanie “Czy liczba przebudzeń w strakcie snu wpływa na jego efektywność?”

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

```
Warning: Removed 26 rows containing non-finite outside the scale range
(`stat_smooth()`).
```

```
Warning: Removed 26 rows containing missing values or values outside the scale range
(`geom_point()`).
```