Impact of different factors on Sleep Duration and Quality

David Ruiz Cáceres with NIU:1672891

June 16, 2024

Abstract

This study explores the impact of different parameters on sleep duration and quality, utilizing a dataset obtained from Kaggle that includes variables such as sleep duration, sleep quality, physical activity level, stress level, BMI category, and the presence of sleep disorders. The dataset was modified to ensure consistency and relevance, including merging similar categories and removing unused columns.

A variety of statistical techniques were employed, including parametric and non-parametric bootstrap methods, chi-squared tests, and Pearson correlation tests. The results indicate that older individuals tend to sleep longer on average and that females generally report longer sleep duration than males, although this may be influenced by the age distribution within the dataset. A significant relationship was found between BMI categories and sleep disorders, with higher BMI categories associated with an increased prevalence of sleep disorders. Additionally, a strong negative correlation was observed between stress levels and sleep quality, suggesting that higher stress levels are associated with poorer sleep quality.

These findings highlight the need for age and gender-specific recommendations in sleep health practices. The study underscores the importance of considering demographic factors in sleep research.

1 Introduction

Sleep is a vital component of overall health and well-being, influencing various physiological and psychological processes. Adequate sleep is essential for cognitive function, emotional regulation, physical health, and overall quality of life. Despite its importance, many individuals experience sleep disturbances, which can be influenced by a multitude of factors, including demographic variables such as age and gender.

Numerous studies have explored the relationship between demographic factors and sleep patterns, consistently finding that both age and gender play significant roles in determining sleep duration and quality. For instance, older adults often experience changes in sleep architecture, leading to shorter sleep durations and more fragmented sleep. Gender differences in sleep patterns have also been observed, with women generally reporting longer sleep durations but higher incidences of sleep disturbances compared to men.

The primary objective of this project is to analyze the relationships between age, gender, and sleep patterns using a robust statistical framework. To achieve this, we employed a variety of statistical tests, including parametric and non-parametric bootstrap methods, chi-squared tests, and Pearson correlation tests. These techniques were chosen to provide a comprehensive understanding of the data and to account for potential variations and biases.

By leveraging a dataset sourced from Kaggle, which includes variables such as sleep duration, sleep quality, physical activity level, stress level, BMI category, and the presence of sleep disorders, this project aims to uncover significant patterns and relationships.

2 Dataset

2.1 Obtainance

Upon receiving this task, our professor presented several projects from previous years. One of these projects utilized a dataset extracted from the website Kaggle. Motivated by this example, I tried to find a dataset that would not only be useful for the project's requirements but also cover a subject of personal interest. While browsing the previosly mentioned website, I discovered a dataset discussing various factors affecting people's sleep. It particularly captured my attention and decided to use it for this project.

2.2 Description

- **Person ID:** An identifier for each individual.
- **Gender:** The gender of the person (Male/Female).
- **Age:** The age of the person in years.
- **Occupation:** The occupation or profession of the person.
- **Sleep Duration (hours):** The number of hours the person sleeps per day.
- **Quality of Sleep (scale: 1-10):** A subjective rating of the quality of sleep, ranging from 1 to 10.
- **Physical Activity Level (minutes/day):** The number of minutes the person engages in physical activity daily.
- **Stress Level (scale: 1-10):** A subjective rating of the stress level experienced by the person, ranging from 1 to 10.
- **BMI Category:** The BMI category of the person (e.g., Underweight, Normal, Overweight).
- **Blood Pressure (systolic/diastolic):** The blood pressure measurement of the person, indicated as systolic pressure over diastolic pressure.
- **Heart Rate (bpm):** The resting heart rate of the person in beats per minute.
- **Daily Steps:** The number of steps the person takes per day.
- **Sleep Disorder:** The presence or absence of a sleep disorder in the person (None, Insomnia, Sleep Apnea).

2.3 Modification

Although the dataset could be used directly as obtained from the website, I made several modifications to facilitate its manipulation and removed any unused information.

The first modification involved the first line of the file, which contains the column names. For each column name that consisted of multiple words, I replaced the spaces with underscores.

Furthermore, I observed that the BMI Categories column contained four different values, two of which were **Normal** and **Normal Weight**. To synthesize these values, I merged them by converting all instances of **Normal Weight** to **Normal**.

At the end of the project, ones I knew all the tests I wanted to do, I erased the columns that were not used by creating a tiny Python code that would erase it automatically for me. Although it was not hard to create the code, a problem was found in the way the dataset was created. The last column had **None** instances, causing Python to skip those, and leaving the last column with blanks on it. To avoid this, I converted all instances from **None** to **Healthy**. Once the Python executable did his job, the remaining features are:

- **Person ID**
- **Gender**
- **Age**
- **Sleep Duration**
- **Quality of Sleep**
- **Physical Activity Level**
- **Stress Level**
- **BMI Category**
- **Sleep Disorder**

3 Techniques Used for Analysis

3.1 Parametric Bootstrap

The parametric bootstrap method consists of the following steps:

- Assume a specific parametric distribution (e.g., normal distribution) for the population based on the data.
- Estimate the parameters of the assumed distribution (e.g., mean and standard deviation for a normal distribution) using the sample data.
- Simulate new bootstrap samples from the estimated parametric distribution.
- For each bootstrap sample, calculate the sample statistic of interest (e.g., mean, difference in means, regression coefficients).
- Repeat the process for a large number of bootstrap samples to build the distribution of the sample statistic.
- Use the distribution of the bootstrap sample statistics to estimate confidence intervals or perform hypothesis tests.

3.2 Non-Parametric Bootstrap

The non-parametric bootstrap method consists of the following steps:

- Using the frequency distribution of the n data values as our best guess of the population or probability distribution.
- Simulating the sampling from the population distribution obtaining the new bootstrap samples.
- For each sampling, calculating the sample statistic of interest.

3.3 Chi-squared Test

This test is used to examine the association between categorical variables from a random sample to determine if there is a significant relationship between them. The procedure consists of the following steps:

- Create a contingency table of the observed frequencies for each category.
- Calculate the expected frequencies for each category based on the null hypothesis of independence.
- Use the formula to calculate the chi-square statistic:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

  where O_i represents the observed frequency and E_i represents the expected frequency.

- Determine the degrees of freedom, which is (number of rows-1)$\times$(number of columns-1).
- Find the critical chi-square value using a chi-square distribution table or statistical software.
- Compare the calculated chi-square statistic to the critical value:
  - If the chi-square statistic is greater than the critical value, reject the null hypothesis.
  - If the chi-square statistic is less than or equal to the critical value, do not reject the null hypothesis.

3.4 Pearson Correlation Test

This test is used to measure the strength and direction of the linear relationship between two continuous variables. The procedure consists of the following steps:

- Collect a paired data set of the two continuous variables you wish to examine.
- Calculate the means of both variables.
- Use the formula to calculate the Pearson correlation coefficient (r):

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

  where X_i and Y_i are the individual data points, and $\bar{X}$ and $\bar{Y}$ are the means of the variables X and Y, respectively.

- Determine the degrees of freedom, which is $n - 2$, where n is the number of paired observations.
- Find the critical r value using a correlation table or statistical software, based on the degrees of freedom and the desired significance level (e.g., 0.05).
- Compare the calculated Pearson correlation coefficient to the critical value:
  - If the absolute value of r is greater than the critical value, reject the null hypothesis of no correlation.
  - If the absolute value of r is less than or equal to the critical value, do not reject the null hypothesis.

4 Results of the Analysis

4.1 Average Sleep Duration

The first that was done is getting with parametric and non-parametric bootstrap a mean value of the sleep duration. After that, I calculated a coefficient interval of the 95%.

First I did the parametric one asuming normality in the data and the results obtained are the following: **7.133564** with a CI: **7.060663 - 7.210722**

Then I realised that the data might not be really adjusted to a Normal, so decided to try the non-parametric bootstrap and obtained the following results: **7.130822** with a CI: **7.053195 - 7.210468**

In order to get more insight, I thought that would be interesting to check if there was a significant difference in this statistic between males and females. That's why I did a non-parametric bootstrap method, getting the following results: **-0.1961605** with a CI: **-0.381575 - -0.02384005** so the mean value for women are slightly above than for men.

This result surprised me a lot, because I thought would be totally the opposite (based on personal experience) so I decided to make a plot of the dataset differentiating the man are women values.

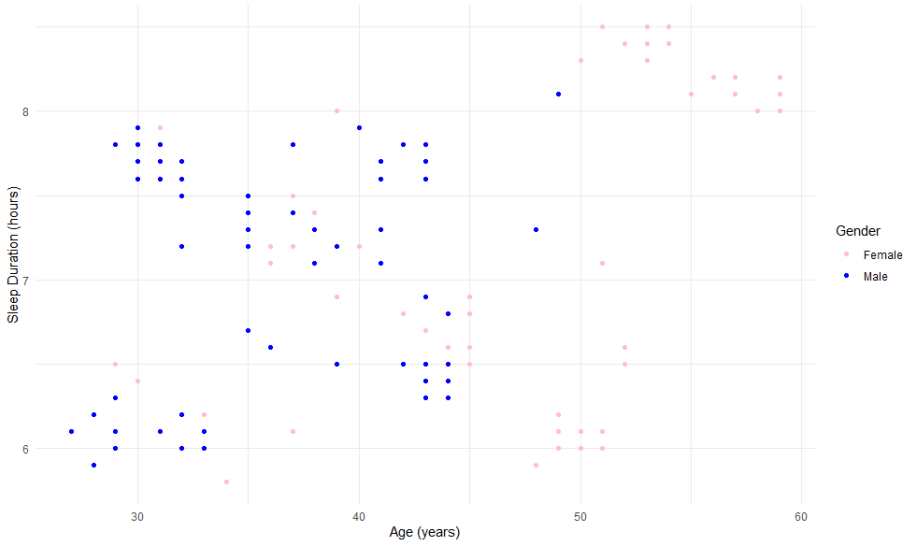

Figure 1: Plot of the points by Age and Sleep Duration by gender

As it can be seen, the dataset I chose concentrates all the male values between the 20's and the 50's but some women can be found as older than 50's.

As soon as I saw this, I decided to perform a test to see if the value of the Sleep Duration is affected by the Age parameter by creating a linear model that fits the dataset and perform a plot for the values of the Sleep Duration depending of the previous mentioned statistic. The result is the following:

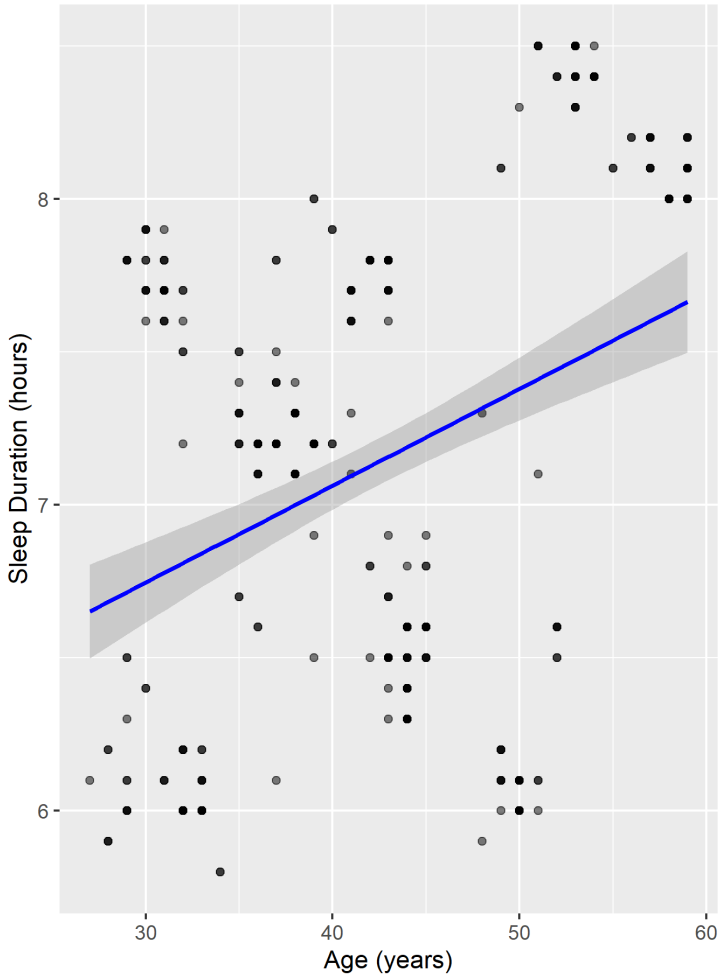

Figure 2: Plot of the Sleep Duration depending on the Age

Being the blue line the mean value approximation, it can be seen that the average sleep duration increases at the same time age does. Having seen this, seems easier to explain the difference found between genders.

4.2 Relation between the BMI category and sleep disorders

For this section I performed a Chi-Squared test to see if exists a relation between the BMI categories and sleep disorders. The Chi-Squared test is already programmed in R so it was just needed to perform it.

First I created a table that groups the cases of each combination of sleep disorders and weight category, which is the following one:

	Healthy	**Insomnia**	**Sleep Apnea**
Normal	200	9	7
Obese	0	4	6
Overweight	19	64	65

Table 1: Sleep Disorders by Weight Category

Then, after performing the test, I checked it's p-value and was lower than 0.05, meaning that there is a relation between them.

4.3 Quality of sleep depending on the stress level

In this last section, I performed a Pearson test for the correlation in order to see if there is a significant dependence between the quality of sleep and the stress level.

With the test, I obtained a p-value lower than 0.05, concluding that there is a correlation between them. The Pearson correlation coefficient obtained is -0.898752, which means that for every augment of 1 stress level, the quality of sleep is lowered by 0.898752. To see the progress I provide a plot:

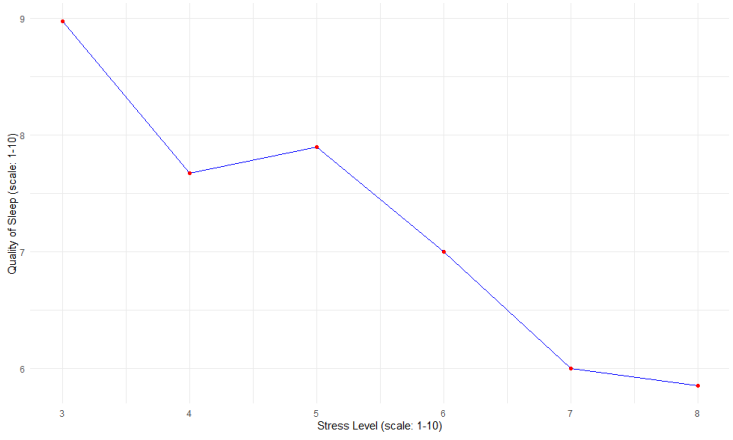

Figure 3: Quality of sleep depending on the stress level

5 Conclusions

- The analysis indicates a positive correlation between age and sleep duration.
- Older individuals tend to sleep longer on average.
- Gender differences in sleep duration are evident with females generally report longer sleep duration (although might be biased but how distributed is the dataset by the age).
- People that live with more stress usually sleeps worse than those that live more relaxed.

These findings suggest the need for age and gender-specific recommendations in sleep health practices, with also some recommendations for those with high-stress levels and a high BMI coefficient, as they are in risk of having a bad sleep quality.

6 Bibliography

- Dataset "Sleep Health and Lifestyle Dataset":
  https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset
- Parametric Bootstrap presentation from UAB
- Non-Parametric Bootstrap presentation from UAB
- Chi-squared test information:
  https://datatab.es/tutorial/chi-square-test
- Pearson correlation test information:
  https://www.scribbr.com/statistics/pearson-correlation-coefficient/

7 Appendix

7.1 R Scripts

7.1.1 Parametric Bootstrap for mean value of Sleep Duration

```
path <- "C:/Users/David/Desktop/ADC_Final_Project/Cleaned_Dataset.csv"
data <- read.csv(path)

parametric_bootstrap_mean = function(x, n_bootstrap = 1000) {
  # Estimate parameters of the normal distribution
  sample_mean = mean(x, na.rm = TRUE)
  sample_sd = sd(x, na.rm = TRUE)

  bootstrap_means = numeric(n_bootstrap)
  set.seed(123)  # For reproducibility
  for (i in 1:n_bootstrap) {
    bootstrap_sample = rnorm(length(x), mean = sample_mean, sd = sample_sd)
    bootstrap_means[i] = mean(bootstrap_sample)
  }

  return(bootstrap_means)
}

stats_mean = parametric_bootstrap_mean(data$Sleep_Duration)

final_mean = mean(stats_mean)
final_mean
sd_mean = sd(stats_mean)
sd_mean

CImean = quantile(stats_mean, probs = c(0.025, 0.975))
CImean

cat("Bootstrap mean:", final_mean, "\n")
cat("Bootstrap standard deviation:", sd_mean, "\n")
cat("95% confidence interval for the mean:", CImean, "\n")
```

7.1.2 Non-Parametric Bootstrap for mean value of Sleep Duration

```
path <- "C:/Users/David/Desktop/ADC_Final_Project/Cleaned_Dataset.csv"
data <- read.csv(path)

non_param__mean = function(x){
  x = mean(sample(x, size = length(x), replace = TRUE))
  return (x)
}
```

```
set.seed(123)
stats_mean = replicate(1000, non_param__mean(data$Sleep_Duration))
stats_mean
final_mean = mean(stats_mean)
final_mean
sd_mean = sd(stats_mean)
sd_mean
CImean = quantile(stats_mean, probs = c(0.025, 0.975))
CImean
```

7.1.3 Non-Parametric Bootstrap for mean value of difference in Sleep Duration by gender

```
path <- "C:/Users/David/Desktop/ADC_Final_Project/Cleaned_Dataset.csv"
data <- read.csv(path)

non_param_mean_diff <- function(data){
  male_mean <- mean(sample(data$Sleep_Duration[data$Gender == "Male"],
                           size = sum(data$Gender == "Male"), replace = TRUE))
  female_mean <- mean(sample(data$Sleep_Duration[data$Gender == "Female"],
                             size = sum(data$Gender == "Female"), replace = TRUE))
  return (male_mean - female_mean)
}
set.seed(123)

stats_mean_diff <- replicate(1000, non_param_mean_diff(data))

final_mean_diff <- mean(stats_mean_diff)

sd_mean_diff <- sd(stats_mean_diff)

CI_mean_diff <- quantile(stats_mean_diff, probs = c(0.025, 0.975))

cat("Bootstrap Mean Difference: ", final_mean_diff, "\n")
cat("Bootstrap Standard Deviation: ", sd_mean_diff, "\n")
cat("95% Confidence Interval: ", CI_mean_diff, "\n")
```

7.1.4 Plot for points depending on gender

```
library(ggplot2)

path <- "C:/Users/David/Desktop/ADC_Final_Project/Cleaned_Dataset.csv"
dataset <- read.csv(path)

ggplot(dataset, aes(x = Age, y = Sleep_Duration, color = Gender)) +
  geom_point() +
```

```
  scale_color_manual(values = c("Female" = "pink", "Male" = "blue")) +
  labs(title = "Scatter Plot of Age vs. Hours of Sleep",
       x = "Age (years)",
       y = "Sleep Duration (hours)",
       color = "Gender") +
  theme_minimal()
```

7.1.5 Plot for average sleep duration depending on age

```
path <- "C:/Users/David/Desktop/ADC_Final_Project/Cleaned_Dataset.csv"
dataset <- read.csv(path)

library(ggplot2)

ggplot(dataset, aes(x = Age, y = `Sleep_Duration`)) +
  geom_point(alpha = 0.5) +  # Scatter plot points with some transparency
  geom_smooth(method = "lm", col = "blue") +  # Add a linear trend line
  labs(title = "Average Sleep Duration by Age",
       x = "Age (years)",
       y = "Sleep Duration (hours)")

ggsave("Average_Sleep_Duration_by_Age.png")
```

7.1.6 Test for correlation between IBM categories and Sleep disorders

```
path <- "C:/Users/David/Desktop/ADC_Final_Project/Cleaned_Dataset.csv"
dataset <- read.csv(path)

contingency_table <- table(dataset$BMI_Category, dataset$Sleep_Disorder)

print(contingency_table)

chi_square_test <- chisq.test(contingency_table)

print(chi_square_test)

if (chi_square_test$p.value < 0.05) {
  cat("There is a significant association between BMI categories and sleep disorders (p-valu
} else {
  cat("There is no significant association between BMI categories and sleep disorders (p-val
}
```

7.1.7 Plot for Quality of sleep depending on Stress level

```
library(ggplot2)
```

```
path <- "C:/Users/David/Desktop/ADC_Final_Project/Cleaned_Dataset.csv"
dataset <- read.csv(path)

summary_data <- dataset %>%
  group_by(Stress_Level) %>%
  summarize(Average_Quality_of_Sleep = mean(Quality_of_Sleep, na.rm = TRUE))

ggplot(summary_data, aes(x = Stress_Level, y = Average_Quality_of_Sleep)) +
  geom_line(color = "blue") +
  geom_point(color = "red") +
  labs(title = "Average Quality of Sleep by Stress Level",
       x = "Stress Level (scale: 1-10)",
       y = "Quality of Sleep (scale: 1-10)") +
  theme_minimal()
```

7.1.8 Pearson test

```
path <- "C:/Users/David/Desktop/ADC_Final_Project/Cleaned_Dataset.csv"
data <- read.csv(path)

correlation <- cor(data$Quality_of_Sleep, data$Stress_Level,
use = "complete.obs")

correlation_test <- cor.test(data$Quality_of_Sleep, data$Stress_Level,
method = "pearson", use = "complete.obs")

cat("Pearson correlation coefficient:", correlation, "\n")
print(correlation_test)
```

7.1.9 Plot for Quality of sleep depending on Stress level

```
library(ggplot2)

path <- "C:/Users/David/Desktop/ADC_Final_Project/Cleaned_Dataset.csv"
dataset <- read.csv(path)

summary_data <- dataset %>%
  group_by(Stress_Level) %>%
  summarize(Average_Quality_of_Sleep = mean(Quality_of_Sleep, na.rm = TRUE))

ggplot(summary_data, aes(x = Stress_Level, y = Average_Quality_of_Sleep)) +
  geom_line(color = "blue") +
  geom_point(color = "red") +
  labs(title = "Average Quality of Sleep by Stress Level",
```

```
      x = "Stress Level (scale: 1-10)",
      y = "Quality of Sleep (scale: 1-10)") +
  theme_minimal()
```

7.2 Python code for Dataset Cleaning

```
import pandas as pd

df = pd.read_csv('C:/Users/David/Desktop/ADC_Final_Project/Dataset.csv')

print(df.columns)

columnas_a_eliminar = ['Occupation', 'Blood_Pressure','Heart_Rate','Daily_Steps']
df = df.drop(columns=columnas_a_eliminar)

df.to_csv('C:/Users/David/Desktop/ADC_Final_Project/Cleaned_Dataset.csv', index=False)
```