Sleep Disorder Analysis: Unveiling the Interplay Between Lifestyle Health and Sleep Quality

Sarah Alshumayri, Reema Abdallah, Yehya Asseri

2023-12-04

Contents

Introduction

This project aims to understand the impact of lifestyle and health factors on sleep quality and disorders. This project compiles data including metrics related to sleep duration, sleep quality, physical activity, stress levels, BMI categories, blood pressure, heart rate, and daily steps. We aim to use data visualization and statistical analysis methods to investigate the variables that significantly affect sleep health. Preliminary analysis may reveal insights such as the correlation between physical activity, stress levels, and sleep quality, but the project will also explore more complex relationships and potential predictors of sleep disorders. The rest of the report is organized as follows: section 2 provides a background on the importance of sleep health and its relationship with lifestyle factors, section 3 presents the research question and problem statement that the report aims to answer, section 4 discusses the data used in this project, its sources, and provides a brief overview of the contents of each dataset, section 5 analyzes those datasets and offers a statistical view of the data, section 6 presents the findings of the project, section 7 discusses the implications of these findings and their potential applications, and section 8 concludes the report with a summary of the key insights and suggestions for future research.

Background

Sleep quality, a critical factor for health and well-being, is influenced by a multitude of factors, including occupational hazards, lifestyle choices, and individual behaviors. In certain professions, such as long-distance heavy goods vehicle (HGV) drivers, the combination of demanding work schedules and poor lifestyle choices leads to increased risks of chronic diseases and reduced life expectancy [2]. This is compounded by inadequate sleep, which is linked to an increased risk of accidents and comorbidities [2].

Sleep behavior is also influenced by demographic, occupational, and lifestyle factors. For instance, sleep efficiency and duration are known to decrease with age, and this is a significant concern in professions with an aging workforce [3]. Similarly, in athletes, optimal sleep is critical for performance, but factors such as training and competition times, travel, stress, and use of stimulants like caffeine can lead to substantial variation in sleep onset and offset times [1].

The most important factor influencing sleep efficiency is bedtime and low variability in sleep onset times [2]. Regular sleepers tend to exhibit consistent sleep onset and offset times compared to irregular sleepers. However, achieving this regularity can be challenging due to training schedules and other commitments [2].

For elite athletes, the biological bases of sleep, driven by homeostatic drive and the circadian clock, are relatively stable. However, sleep regularity can be significantly affected by external factors such as training schedules, psychological stress, and societal influences. These factors impact sleep regularity and highlight the importance of modifying behaviors that can lead to poor sleep quality and duration [4].

Research Question and Problem Statement

Can machine learning models effectively identify key lifestyle and health factors influencing sleep quality?

Understanding the intricate relationship between various lifestyle and health factors and their impact on sleep quality and disorders is essential for developing effective health interventions. Traditional analytical methods may not fully capture the complex interactions and nonlinear relationships between these factors. This research aims to leverage the capabilities of machine learning models to analyze a comprehensive dataset encompassing demographic, occupational, physical activity, stress levels, and health indicators. The objective is to determine how these factors collectively influence sleep duration, quality, and the presence of sleep disorders. By evaluating the performance of various machine learning models, this study seeks to pinpoint the most significant factors affecting sleep health. The insights gained could provide valuable guidance for healthcare professionals and policymakers in formulating strategies to enhance sleep quality and address sleep-related issues in the population.

Data

Unit of Observation

The unit of observation in this dataset is an individual person. Each row represents data for one individual, with various attributes related to their demographics, lifestyle, and health.

Outcome Variable

- The outcome variable appears to be related to sleep health, which can be represented by ‘Quality of Sleep’, or ‘Sleep Disorder’.

1. **Quality of Sleep:** Rated on a scale (exact scale not specified in the preview).
2. **Sleep Disorder:** Categorical variable indicating the type of sleep disorder, if any.

Source: Presumably collected from individuals health records https://data.world/andrewpyong/sleep-health. Distribution: This will be described using a graph and a table.

Predictor Variables: Predictor variables include ‘Gender’, ‘Age’, ‘Occupation’, ‘Physical Activity Level’, ‘Stress Level’, ‘BMI Category’, ‘Blood Pressure’, ‘Heart Rate’, and ‘Daily Steps’. These are measured as categorical (e.g., Gender, Occupation, BMI Category), ordinal (e.g., Quality of Sleep), or continuous variables (e.g., Age, Blood Pressure).

Source: Presumably collected from individuals health records https://data.world/andrewpyong/sleep-health. Distribution: This will be depicted using tables and graphs for each variable.

Potential Issues with the Data: When considering potential issues with the data, the primary concerns are the lack of variation in some categories and potential biases. For instance, certain occupations or BMI categories might have limited representation, which could affect the generalizability of the findings. Also, if the sample is not representative of the broader population (e.g., skewed towards a specific age group, or occupational category), it could introduce biases in the analysis.

To overcome or mitigate the issues of lack of variation and potential biases in your data analysis, a multifaceted approach can be adopted. Firstly, it’s essential to transparently report the limitations of the study due to these factors. Acknowledging the specific areas where the dataset may not perfectly represent the broader population or where certain categories are underrepresented adds to the credibility of the research and aids in the accurate interpretation of the results. Alongside this, the implementation of regression techniques serves as a robust method to address imbalances in the dataset. By using logistic or linear regression models, it becomes possible to control for confounding variables, allowing for a more accurate isolation of the effects of primary predictors. This approach not only helps in drawing more reliable conclusions but also enhances the overall integrity of the analysis by systematically adjusting for known dataset limitations.

Analysis

Methods/Tools Explored

In our project, we employed a variety of methods and tools to thoroughly analyze the “Sleep, Health, and Lifestyle” dataset. The primary tool used was R, renowned for its robust capabilities in data analysis, statistical computing, and graphical representation. This choice was driven by R’s comprehensive support for data manipulation, visualization, and advanced analytics.

Key R packages used included:

- dplyr and tidyr: For efficient data manipulation and tidying.
- ggplot2: For creating comprehensive and aesthetically pleasing visualizations.
- caret: For streamlined data preprocessing and machine learning.

The analysis included rigorous exploratory data analysis (EDA) to comprehend the data structure, identify missing values, and explore potential correlations among variables. Given the dataset's characteristics, Random Forest was chosen as the primary predictive modeling technique, valued for its effectiveness in handling numerous predictor variables and capturing complex data patterns.

Detailed Analysis Outline

The analysis followed these key stages:

1. **Data Preprocessing and Cleaning:**

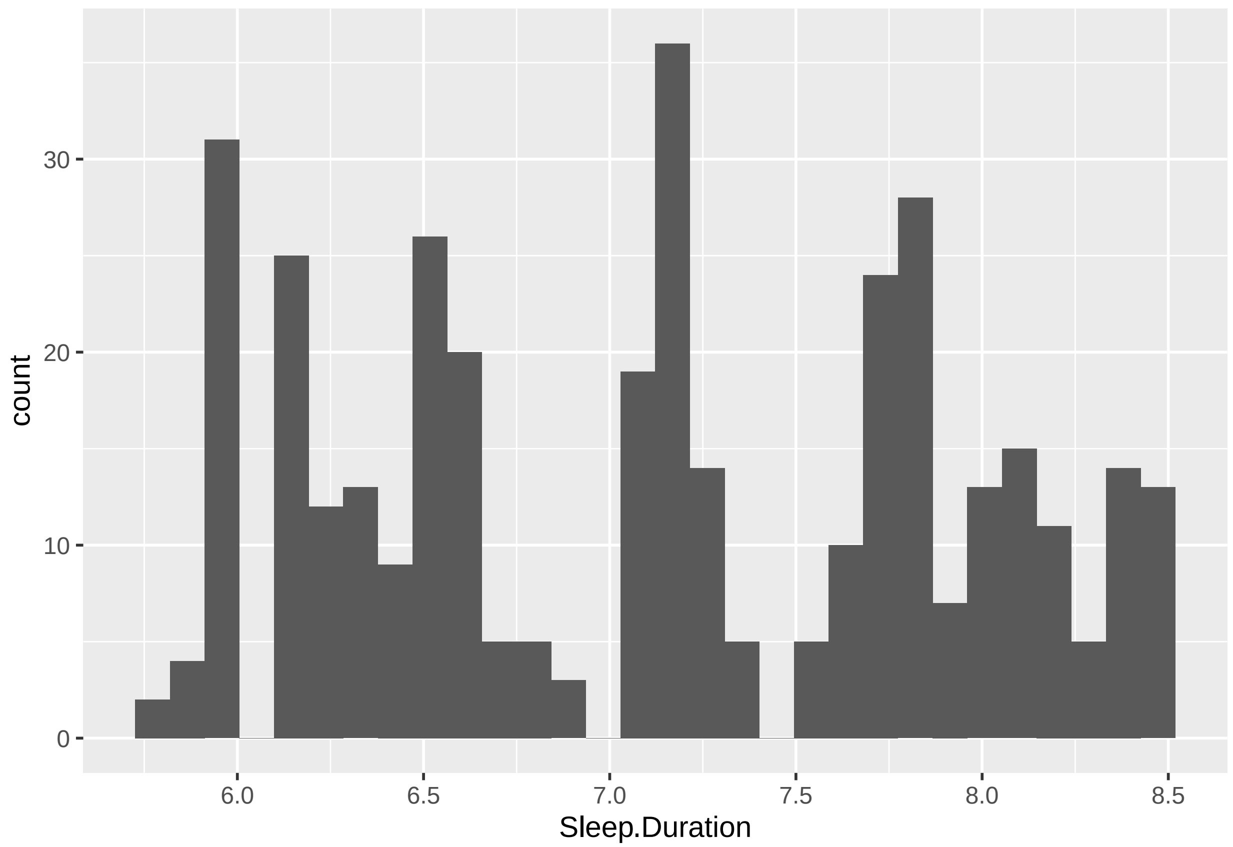
- Encoding categorical variables and normalizing numeric data.

2. **Exploratory Data Analysis (EDA):**

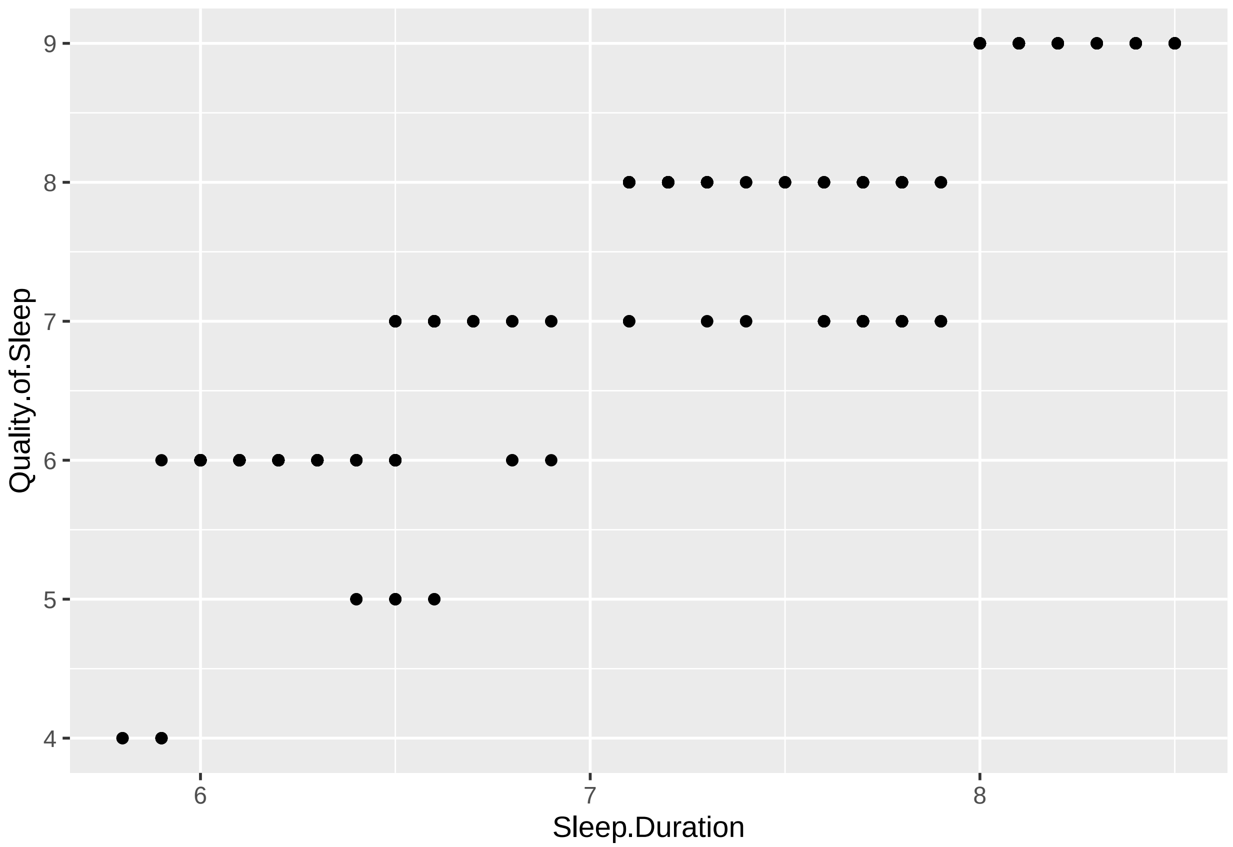
- Utilizing various visualization tools (like histograms, box plots, scatter plots) to understand variable distributions and relationships.

```
ggplot(Sleep_dataset, aes(x = `Sleep.Duration`)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
ggplot(Sleep_dataset, aes(x = `Sleep.Duration`, y = `Quality.of.Sleep`)) + geom_point()
```



- Summary statistics to capture central tendencies and dispersion in key variables.

```
#Summary Statistics for Key Variables
Sleep_dataset %>%
  summarise(across(c(Sleep.Duration, Quality.of.Sleep, Physical.Activity.Level,
                     Stress.Level, Blood.Pressure.1, Blood.Pressure.2,
                     Heart.Rate, Daily.Steps, BMI.Levels),
                   list(mean = ~mean(., na.rm = TRUE),
                        sd = ~sd(., na.rm = TRUE),
                        median = ~median(., na.rm = TRUE))))
```

```
##   Sleep.Duration_mean Sleep.Duration_sd Sleep.Duration_median
## 1            7.132086         0.7956567                   7.2
##   Quality.of.Sleep_mean Quality.of.Sleep_sd Quality.of.Sleep_median
## 1              7.312834            1.196956                       7
##   Physical.Activity.Level_mean Physical.Activity.Level_sd
## 1                     59.17112                    20.8308
##   Physical.Activity.Level_median Stress.Level_mean Stress.Level_sd
## 1                             60          5.385027        1.774526
##   Stress.Level_median Blood.Pressure.1_mean Blood.Pressure.1_sd
## 1                   5              128.5535            7.748118
##   Blood.Pressure.1_median Blood.Pressure.2_mean Blood.Pressure.2_sd
## 1                     130              84.64973            6.161611
##   Blood.Pressure.2_median Heart.Rate_mean Heart.Rate_sd Heart.Rate_median
## 1                      85        70.16578      4.135676                70
##   Daily.Steps_mean Daily.Steps_sd Daily.Steps_median BMI.Levels_mean
## 1         6816.845       1617.916               7000        1.449198
##   BMI.Levels_sd BMI.Levels_median
## 1     0.5492739                 1
```

3. **Feature Engineering and Selection:**

- Identifying crucial predictor variables via correlation analysis and initial model insights.
- Crafting new features that could enhance model performance and interpretability.

4- **Predictive Modeling with Random Forest:** - Building and training the Random Forest model. - Tuning hyperparameters to optimize the model's performance. - Validating the model using cross-validation techniques.

5. **Model Interpretation and Evaluation:**

- Interpreting the model using feature importance scores and visualization tools like Partial Dependence Plots (PDP).
- Evaluating the model's performance through metrics like accuracy, recall, precision, and the Area Under the Curve (AUC) for ROC analysis.

6. **Validation and Testing:** -Assessing model robustness on a separate test set.

- Using a range of performance metrics to ensure reliability and accuracy.

The approach was crafted to be accessible to readers with basic knowledge of R and machine learning, explaining each step with clarity and its rationale based on the dataset's nature and the research objectives. The methodology was selected to provide a comprehensive understanding of the dataset and to ensure the predictive modeling was both robust and interpretable.

Result

References

[1] S. L. Halson et al., "Sleep Regularity and Predictors of Sleep Efficiency and Sleep Duration in Elite Team Sport Athletes," Sports Medicine - Open, vol. 8, no. 79, 2022.

[2] R. Smith et al., "Sleep Patterns and Disorders Among Long-Distance HGV Drivers: A Concern for Road Safety," Transportation Research Part F: Traffic Psychology and Behaviour, vol. 77, pp. 1-14, 2021.

[3] R. Smith et al., "Occupational Factors Affecting Sleep Health of HGV Drivers," Safety and Health at Work, vol. 12, no. 4, pp. 500-507, 2021.

[4] S. L. Halson et al., "The Impact of Training and Competition on Sleep Patterns of Elite Athletes," Sports Medicine - Open, vol. 8, no. 79, 2022.