# Exploring Sleep Quality Affecting Factors and Temporal Patterns in Sleeping Data on a Subject-Level

## 1. Introduction

Humans spend roughly a third of their lifetime sleeping. Sleep is essential for the physical and mental well-being of humans, and sleep deprivation can have a significant negative effect on the attention, cognition, and mood of humans (Worley 2018, p. 758). Moreover, sleep deprivation can cause mood swings and increase the risk of getting ill (Pacheco & Singh, 2023). During sleep, our minds and bodies recover from the stresses of the previous day, which is why sleeping affects our performance on the next day (A.D.A.M. Medical Encyclopedia, 2022). The minimum amount of sleep that adults generally need has been shown to be around 7 hours per night (Pacheco & Singh, 2023; Worley, 2018, p. 759).

The awareness of the importance of sleep has slowly but steadily increased in recent years, as people are understanding the possible negative effects that sleep deprivation can have on their lives (Worley, 2018, p. 759). An interesting trend related to this are sleep tracking devices and applications. These devices and applications gather data which typically includes features, such as sleep duration, sleep quality, and sleep cycles. Furthermore, these tools often offer detailed analytics to the user about their sleeping habits.

While users often do not have direct access to the algorithms behind these analytics, many of these applications allow the user to export the raw sleeping data. This is a valuable feature, because it allows the user to perform their own analysis on the data. Furthermore, the exported data can be combined with data from other sources, such as activity data or location data, which can reveal interesting observations.

In this project, I perform data analysis on a dataset containing sleeping data from a single subject. Thus, this project serves as an example on how the data obtained with the described sleep tracking tools can be analyzed. After this introductory chapter, the structure of this report is as follows: Firstly, the second chapter describes the research questions that this project aims to answer. Following this, the third chapter provides a detailed description of the dataset that is used in this project and the operations that were performed on the dataset before starting the analysis. Then, the fourth chapter presents the methods used in the analysis, and the fifth chapter describes the findings that were obtained, as well as their meaning. The last chapter contains additional discussion on the topic and serves as a conclusion.

## 2. Problem formulation

The question that this project aims to answer is: How can the raw sleeping data obtained with sleep tracking tools be analyzed? As already mentioned, the dataset that is analyzed in this project contains sleeping data from a single subject. Hence, the main objective of this project is to make subject-level observations from the data, and preferably find causal connections between the different features of the data. More specifically, the two questions that this analysis aims to answer are:

- What kind of relationships are there between the different features in the data?
- Are there any temporal patterns in the data?

There are various methods that can be used to find answers to these questions. However, before utilizing these methods, the dataset must be carefully examined and preprocessed. The next chapter introduces the dataset that is used in this project, and explains the different stages of data exploration and preprocessing that were conducted before further analyzing the data.

## 3. Dataset

The dataset that is used in this project is a Kaggle dataset owned by Dana Diotte (Diotte, 2022). The owner has added two datasets to the site, the second of which is used in this analysis. The data of the dataset has been collected with the Sleep Cycle iOS mobile application between May 2019 and March 2022. In the original dataset, there are a total of 21 features and 920 data points. However, in this project, only a subset of the 21 features is used.

From the initial 21 features, a subset of 12 features was chosen for the purposes of this analysis. The choice was based mainly on the number of null values in the columns, and the data types of the columns. Generally, columns with fewer null values were preferred over columns with many null values, and columns with numerical values were preferred over columns with string values. Thus, textual features, such as ‘City’ and ’Weather type’, were dropped from the dataset. Based on these conditions, the chosen features were: timestamps for going to sleep and waking up, sleep quality rating, sleep regularity rating, time spent in bed, time asleep, time before sleep, movements per hour, activity data in the form of steps per day, air pressure, snoring time, and weather temperature. All the features except the regularity rating are quite self-explanatory. In the context of sleep, regularity refers to the degree at which a person goes to sleep and wakes up at consistent times (Sleep Cycle, 2023a).

However, as already stated, the data must be preprocessed before starting the analysis. The preprocessing operations that were executed in the beginning are explained next. Firstly, the ‘Start and ‘End’ columns were converted into datetime objects. Secondly, ’the sleep quality and regularity ratings, which are represented as percentages in the original dataset, were converted into a decimal representation. Thirdly, the columns with time values were converted into more representative units. For instance, the time asleep and time in bed features were changed to hours, because people typically think of sleeping time in hours and not in seconds or minutes, for example. Fourthly, all the zero values in the dataset that should have been NaN, were replaced with NaN values. This was done to prevent NaN values from disguising themselves as zero values. However, this couldn’t be done for all columns, because for some features, such as snoring time, zero values are quite possible. After this, two additional columns, ‘Weekday’ and ‘Date’ were created for the purposes of exploring the temporal patterns in the data. Furthermore, a third additional column ‘PercentageSnoring’ was created to more accurately illustrate the amount of snoring per night.

After these changes, the missing data in the dataset were explored. It was observed that the ‘AirPressure’ and ‘WeatherTemperature’ columns had quite a lot of missing data. However, despite this, these features were kept because they are later used in the correlation analysis. In addition to these two columns, a handful of columns had a marginal amount of rows with missing data. Due to their marginality, these rows were simply dropped from the dataframe. After all these preprocessing operations, the resulting number of datapoints in the dataset was 897, and the number of features was 15. The final column names and their data types are presented in table 1.

Table 1. The features of the dataset after data preprocessing.

| Column name | Data type | Number of NaN values |
|---|---|---|
| Date | object (string) | 0 |
| Weekday | object (string) | 0 |
| Start | datetime64[ns] | 0 |
| End | datetime64[ns] | 0 |
| HoursInBed | float64 | 0 |
| HoursAsleep | float64 | 0 |
| MinutesBeforeSleep | float64 | 0 |
| MovementsPerHour | float64 | 0 |
| SleepQuality | float64 | 0 |
| Regularity | float64 | 0 |
| Steps | float64 | 0 |
| HoursSnoring | float64 | 0 |
| PercentageSnoring | float64 | 0 |
| AirPressure | float64 | 553 |
| WeatherTemperature | float64 | 479 |

| | |
|---|---|
| **Total number of datapoints** | 897 |
| **Total number of features** | 15 |

## 4. Methods

Two different methods are used in the analysis with the aim to answer the previously presented research questions. The first method is correlation analysis, which is performed for the columns with numerical values. The Pandas correlation function is used to create a correlation matrix for the numerical features in the dataset. In addition to this, scatter plots are created for all the numerical feature pairs in order to visualize the data and to spot outliers in the data. After this, the outputs are carefully explored and the most interesting findings are explained in the results section of this report.

The second method that is used is temporal pattern analysis. More specifically, the aspect that the analysis aims to delve into is the variation in the sleeping data between different weekdays. This is achieved by plotting a Seaborn boxplot and swarmplot on top of each other, which effectively illustrates the differences between the subject's sleeping data on different weekdays. The plots are carefully examined and as with the results of the correlation analysis, the findings are explained in the results chapter.

## 5. Results

In this chapter, the most interesting findings obtained with the two methods presented earlier, and the visualizations related to them, are presented and explained. First, the results of the correlation analysis are described, after which the most interesting plots from the temporal pattern analysis are presented. To begin with, the pairwise correlations between the numerical features are visualized in figure 1, and the correlation coefficients are presented in table 2.

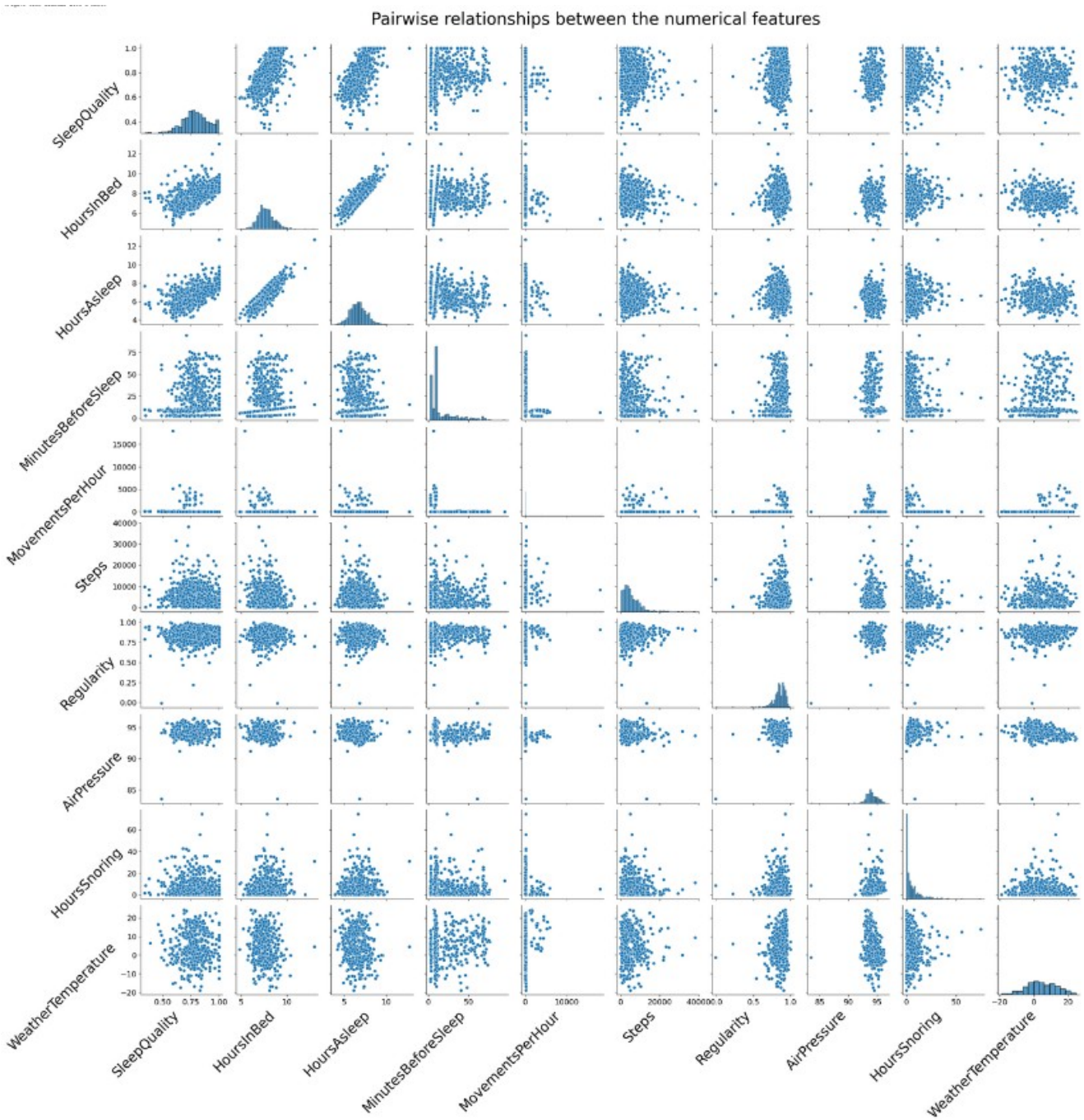



Figure 1. Pairwise correlations between the numerical fieatures of the dataset.

Table 2. Correlation coefficients between the numerical features of the data.

| | SleepQuality | HoursInBed | HoursAsleep | MinutesBeforeSleep | MovementsPerHour | Steps |
|---|---|---|---|---|---|---|
| **SleepQuality** | 1.000000 | 0.551149 | 0.587696 | 0.086660 | -0.110412 | -0.071938 |
| **HoursInBed** | 0.551149 | 1.000000 | 0.868568 | 0.039700 | -0.143035 | -0.095680 |
| **HoursAsleep** | 0.587696 | 0.868568 | 1.000000 | -0.154946 | -0.085515 | -0.084692 |
| **MinutesBeforeSleep** | 0.086660 | 0.039700 | -0.154946 | 1.000000 | -0.086922 | -0.029433 |
| **MovementsPerHour** | -0.110412 | -0.143035 | -0.085515 | -0.086922 | 1.000000 | 0.107182 |
| **Steps** | -0.071938 | -0.095680 | -0.084692 | -0.029433 | 0.107182 | 1.000000 |
| **Regularity** | 0.086574 | -0.019963 | -0.006531 | -0.005697 | 0.035974 | -0.003782 |
| **AirPressure** | 0.038477 | -0.082965 | -0.024518 | -0.134863 | -0.023421 | -0.191748 |
| **HoursSnoring** | -0.060884 | 0.088460 | 0.028969 | 0.111613 | -0.022602 | 0.021221 |
| **PercentageSnoring** | -0.116516 | -0.014800 | -0.088037 | 0.122366 | -0.010854 | 0.035269 |
| **WeatherTemperature** | 0.044708 | -0.120339 | -0.127961 | 0.211621 | 0.237775 | 0.038868 |

| Regularity | AirPressure | HoursSnoring | PercentageSnoring | WeatherTemperature |
|---|---|---|---|---|
| 0.086574 | 0.038477 | -0.060884 | -0.116516 | 0.044708 |
| -0.019963 | -0.082965 | 0.088460 | -0.014800 | -0.120339 |
| -0.006531 | -0.024518 | 0.028969 | -0.088037 | -0.127961 |
| -0.005697 | -0.134863 | 0.111613 | 0.122366 | 0.211621 |
| 0.035974 | -0.023421 | -0.022602 | -0.010854 | 0.237775 |
| -0.003782 | -0.191748 | 0.021221 | 0.035269 | 0.038868 |
| 1.000000 | 0.154560 | -0.049673 | -0.037822 | 0.161539 |
| 0.154560 | 1.000000 | 0.117409 | 0.133169 | -0.235269 |
| -0.049673 | 0.117409 | 1.000000 | 0.983788 | -0.040595 |
| -0.037822 | 0.133169 | 0.983788 | 1.000000 | -0.017912 |
| 0.161539 | -0.235269 | -0.040595 | -0.017912 | 1.000000 |

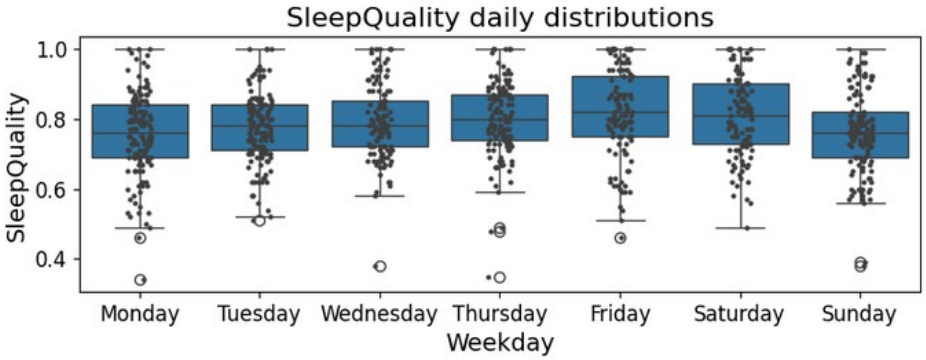
In both the correlation plot and the corresponding correlation matrix, it is evident that the sleep quality rating shows a fairly strong positive correlation with both the 'HoursInBed' and 'HoursAspleep' features. The sleep quality rating is a value that has been obtained from the Sleep Cycle application, and in the website of the app, time in bed is mentioned as one of the measurements affecting the sleep quality rating (Sleep Cycle, 2023b). This relationship is clearly evident in the data.

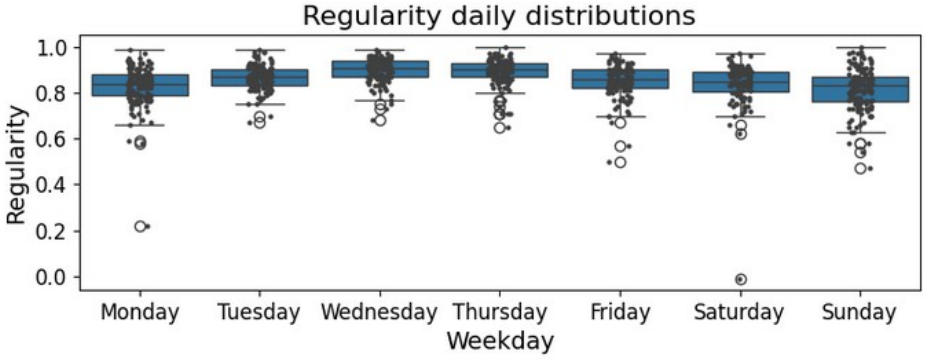
Another observation from the correlation analysis is that the average sleep regularity rating of the subject is fairly high. However, it seems that this does not affect the sleep quality rating of the subject, meaning that regularity does not necessarily guarantee better sleep for the subject. The subject also seems to be quite regular with his falling asleep times, since the ’MinutesBeforeSleep’ feature has two clear spikes indicating that the subject often falls asleep after those times. The two spikes could mean that the subject usually falls asleep either very quickly (the first spike) or if he is not that tired, it takes him a while longer to fall asleep (the second spike).

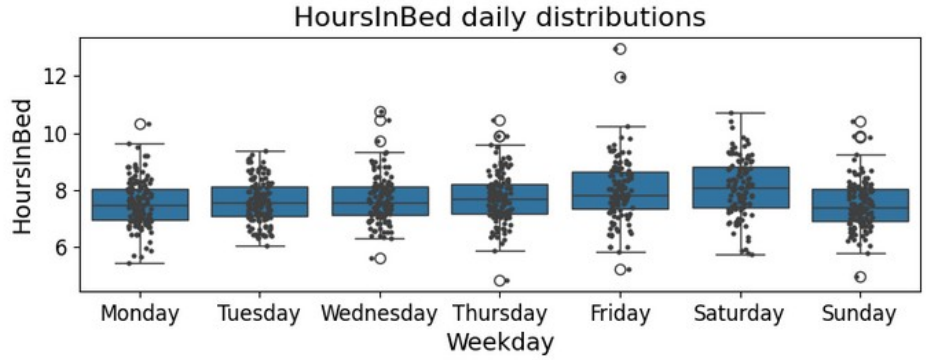
A third observation from the correlation coefficients is that the weather temperature has a weak positive correlation with movements per hour and minutes before sleep. This could mean that high temperatures are negatively affecting the subject’s sleep. However, the correlation between the sleep quality rating and the weather temperature does not have a correlation coeafficient that would provide any further insight into this topic. Thus, understanding the real effect of weather temperature would require further research with a larger amount of weather data.

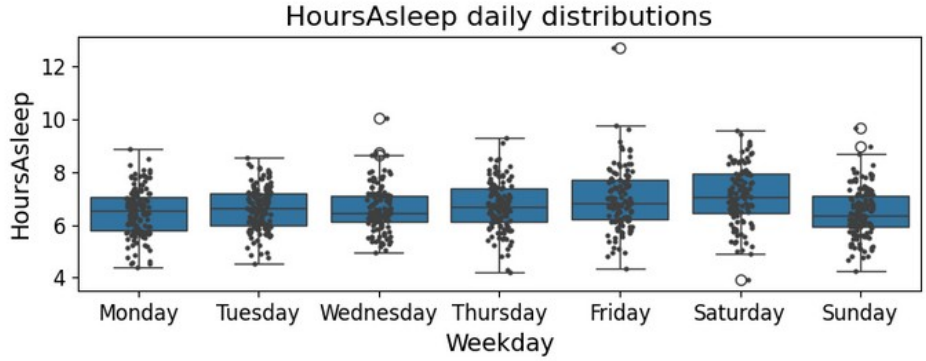
There are some odd patterns in the pairwise correlations as well. For example, the ’MovementsPerHour’ feature has a very low value in a big portion of the data points. However, simultaneously it has a cluster above the smaller values. One possible cause for this could be that the measurements are inaccurate. For example, if the phone that is meaasuring the movements is placed too far from the subject, it might not register all the movements, resulting in low values in the data.

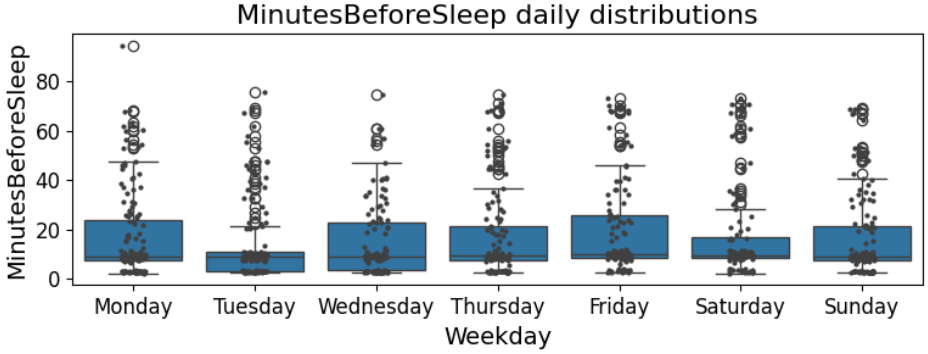
Moving on to the temporal pattern analysis, figure 2 illustrates the daily distribution of the directly sleep related features. The non-sleep related features, such as the weather temperature, the air pressure, and the steps count, were dropped out because their daily distribution does not provide any additional sleep related information.


SleepQuality daily distributions
SleepQuality
1.0
0.8
0.6
0.4
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Weekday


Regularity daily distributions
Regularity
1.0
0.8
0.6
0.4
0.2
0.0
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Weekday


HoursInBed daily distributions
HoursInBed
12
10
8
6
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Weekday


HoursAsleep daily distributions
HoursAsleep
12
10
8
6
4
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Weekday


MinutesBeforeSleep daily distributions
MinutesBeforeSleep
80
60
40
20
0
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Weekday

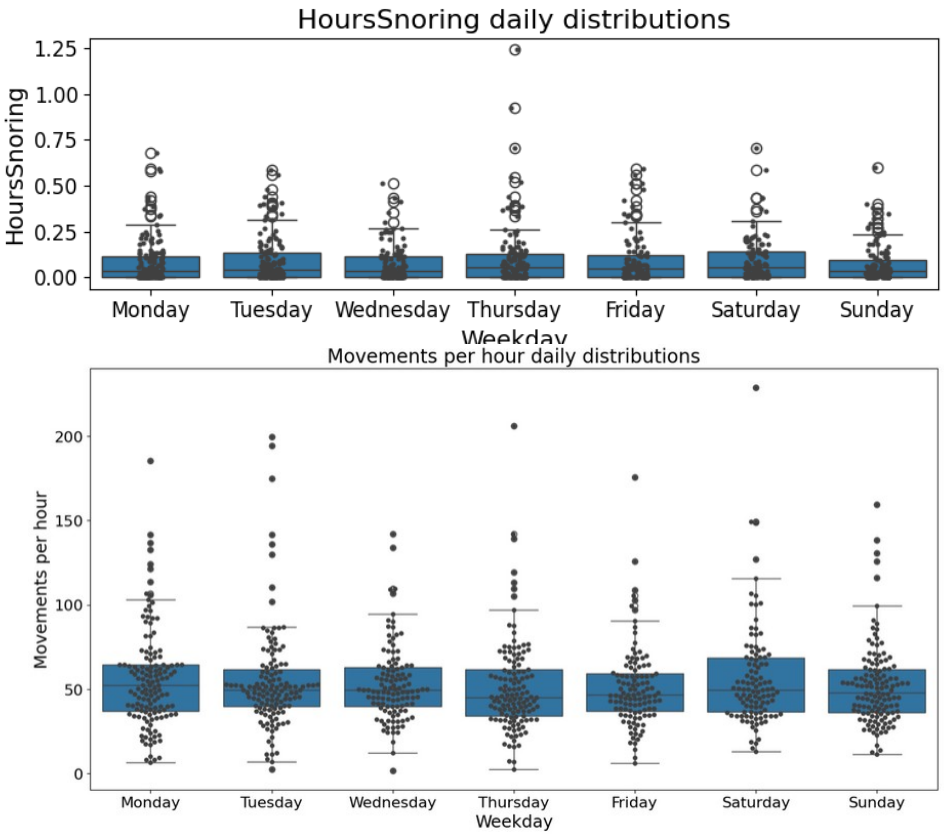



Figure 2. The daily distributions of the sleep related features.

Based on figure 2, the features can be divided into three groups. In the first group, the features do not have a clear variation between different weekdays. This group includes the ’MovementsPerHour’ and ’HoursSnoring’ features. In the second group, the values on Saturdays and Fridays are higher than on the other weekdays. The features in this group are ’HoursInBed’, ’SleepQuality’, and ’HoursAsleep’. The third and fourth groups both have only one feature. The third group has random variability between weekdays and the feature in it is ’MinutesBeforeSleep’. The last group includes only the regularity rating feature, and it has the highest values in the middle of the week. Thus, the subject seems to have the most regular sleeping habits in the middle of the week, which is an interesting observation.

From the features of the second group it can be observed that the subject has a fairly consistent sleeping time distribution from Sunday to Thursday, while on Fridays and Saturdays the distributions differ from the other weekdays. On Fridays, the data seems to be the most spread out, and on Saturdays (the night between Saturday and Sunday) the subject seems to sleep more than on the other days. A possible interpretation for this is that the subject might tend to go out more on Fridays. On the other hand, if he does not go out, he might be so tired after the work week that he tends to sleep longer. These two interpretations could be some of the reasons causing the irregularities in the time spent in bed on the nights between Friday and Saturday.

## 6. Discussion & conclusion

Based on the findings presented in the previous chapter, it appears that no single factor strongly influences the subject's sleep quality. Instead, the subject's sleep quality seems to be the result of multiple smaller factors rather than one significant factor. However, the results must be treated with caution, because the origin of the sleep quality rating is not entirely clear. The website of the Sleep Cycle application simply states that the sleep quality rating is based on four measurements: time in bed, time in deep sleep, times being fully awake during the night, and the frequency and intensity of movement. However, the site does not mention the exact formula that is used to obtain the rating (Sleep Cycle, 2023b). Despite this, however, the fairly strong correlation between the hours in bed and hours asleep features and the sleep quality rating indicates that the sleep quality ratings are at least fairly accurate and illustrative.

Another aspect that must be pointed out is that the results of this analysis are not generalizable, because they have been obtained from a dataset containing data from a single subject. In fact, the effects of sleep loss on individuals can vary significantly (Worley, 2018, p. 759). Therefore, it is obvious that no universal conclusions can be drawn from this research. Nonetheless, it would be interesting to see the results of similar studies conducted with data from multiple subjects. For example, the companies offering sleep tracking tools have access to limitless sleeping data, which could be utilized to perform such studies. However, this data is often left unexplored because the companies are not willing to share the data. This is unfortunate, because many interesting observations could possibly be made if researchers had access to such extremely large datasets.

Overall, I am quite pleased with the results that I was able to obtain with the analysis. However, due to time constraints, the analysis remained somewhat shallow. Additionally, the utilized methods were a bit simpler than I had originally planned. If the analysis were to be continued further, using machine learning methods could have been the next step. Nonetheless, as I said, I am pleased with the results that I did manage to obtain with the simpler methods. The results that were obtained revealed interesting subject-level observations from the data, which was the aim of this project. I now have the required skills to perform simple analysis on my own sleeping data, if I were to start gathering it some day. Furthermore, I feel like I did learn a lot about data analysis while working on this project, which will surely prove to be useful in the future.

## References

A.D.A.M. Medical Encyclopedia. (2022). *Sleep and your health*. MedlinePlus. Retrieved December 6, 2023, from https://medlineplus.gov/ency/patientinstructions/000871.htm

Diotte, D. (2022). *Sleep Data*. Kaggle. Retrieved December 12, 2023, from https://www.kaggle.com/datasets/danagerous/sleep-data/data

Pacheco, D., & Singh, A. (2023). *Why Do We Need Sleep?* Sleep Foundation. Retrieved December 6, 2023, from https://www.sleepfoundation.org/how-sleep-works/why-do-we-need-sleep

Sleep Cycle. (2023a). *Sleep Training Guide: Article 5. Regularity and sleep*. Sleep Cycle. Retrieved December 12, 2023, from https://www.sleepcycle.com/sleeping-resources/sleep-training-guide/5-regularity-and-sleep/

Sleep Cycle. (2023b). *How does the app calculate Sleep Quality?* Sleep Cycle. Retrieved December 14, 2023, from https://support.sleepcycle.com/hc/en-us/articles/206704659-How-does-the-app-calculate-Sleep-Quality-

Worley, S. L. (2018). The Extraordinary Importance of Sleep. *Pharmacy and Therapeutics*, *43*(12), 758-763.