CS-C4100 Digital Health and Human Behaviour
Project report

Topic 4

1. Introduction

Consumer wearable devices refer to electronic devices located somewhere in our bodies or clothing. Different consumer wearable devices, such as activity trackers like Fitbit, Apple Watch, Polar, and Garmin, and smart rings like the Oura Ring, are used in monitoring sleep or daily activity such as steps, heart rate, and motion. Monitoring these metrics can provide valuable information about overall health and well-being. According to Walch et al. (2019), around 50-70 million individuals in the United States are impacted by inadequate sleep either in duration or quality. Sleep pattern and sleep quality monitoring helps ensure that an individual is getting enough restorative sleep, since poor sleep quality or insufficient sleep can lead to various health issues, such as cardiovascular diseases and a weakened immune system (Wang et al., 2017).

In this report, I will present observations of which kind of behavioural patterns affect sleep quality and sleeping patterns based on Apple Watch data. In addition, I will create a model prediction if an individual is either Asleep or Awake at a certain time. The data used in my project was collected in the study by Walch et al. (2019). The data consists of motion (in g), heart rate (in bpm), and steps from Apple Watch data, as well as labeled sleep scored from polysomnography (0-5, wake = 0, N1 = 1, N2 = 2, N3 = 3, REM = 5). Stages N1, N2, and N3 refer to non-rapid eye movement (NREM) sleep stages or the deeper sleep phases, and REM refers to rapid eye movement sleep.

After cleaning and preprocessing the data, I categorized the subjects as either ‘Awake’ or ‘Asleep’ at certain time points and used a logistic regression model to predict the sleeping type based on heart rate and motion in the x, y, and z directions. I used scikit-learn implementation of Logistic Regression, and Standard Scaler to scale the features. I also used stratified K-Fold cross-validation to assess the model’s performance. As evaluation metrics, I used accuracy, precision, recall, and F1 score. The model performed quite well with 89% accuracy, 88% precision, 99.9% recall, and 93% F1 score, especially in correctly identifying instances of being ‘Asleep’ (high recall) while maintaining a good balance with precision. Further fine-tuning will be needed to reduce false positives and obtain a more precise model. This is crucial for applications related to health since wrong impressions could cause misunderstandings in users.

I performed exploratory data analysis on the group level and the individual level to identify the factors that contribute to better sleep quality, using average steps, resting heart rate, the amount of deep sleep (stage N3=3), and the total amount of sleep as parameters. The results show that an active lifestyle (over or equal to 10 000 steps per day) results in a longer deep sleep phase than an inactive lifestyle (less than 10 000 steps per day). In addition, the higher average step count results in a lower

resting heart rate, which is associated with a lower risk of cardiovascular diseases (Cooney et al., 2010).

While these observations shed light on potential influences of sleep quality and a healthy lifestyle, they are solely based on the number of steps, heart rate, and the amount of deep sleep. For example, the subjects' age or gender was not mentioned in the data, which could potentially impact the interpretation of the results. Although the study ruled out subjects with diseases affecting the study, such as insomnia and restless legs syndrome, they did not consider the subjects' stress levels, which could have influenced the observations.

This report is divided into 6 parts. The next part will look more into the problem, and after that, the data set will be explained in more detail. Part 4 explains the methods and justifications for them. After that, the results are presented followed by a discussion about them.

2. Problem Formulation

The project aims to analyze data collected from Apple Watch, including motion (acceleration), heart rate, and step counts, alongside labeled sleep data obtained from polysomnography (PSG). The labeled sleep data categorizes sleep stages into wake (0), N1 (1), N2 (2), N3 (3), and REM (5). The objective is to derive insights and build a model to predict sleep states and explore factors influencing the quality of sleep.

The challenge I will be exploring is whether I will be able to train a model to predict the sleep stages solely on motion and heart rate and observe the reliability of consumer wearable devices in sleep analysis. In addition, defining the factors that contribute to better sleep quality is important for users to identify and evaluate their patterns, and make appropriate changes to them if needed. This could potentially reduce the risks of diseases related to poor sleep quality, such as cardiovascular diseases. Sleep quality is crucial for overall well-being, and understanding factors influencing deep sleep phases can contribute to improved health outcomes. Utilizing Apple Watch data for sleep analysis provides a non-intrusive and continuous monitoring solution, making it practical for long-term insights into sleep patterns.

While there is existing research on sleep analysis, the project's focus on using wearable technology data, specifically Apple Watch, for predicting sleep states without relying on labeled sleep values, is a less-explored aspect. The challenge lies in developing models that consider individual differences in sleep patterns and uncovering conditions that support longer deep sleep phases.

By addressing the research problem, this project aims to contribute to the field of sleep analysis by utilizing Apple Watch data to predict sleep states and understand conditions affecting deep sleep. The findings have the potential to enhance our understanding of individual sleep patterns and contribute to personalized interventions for improved sleep quality.

3. Dataset

The data was collected for a study by Walch et al. (2019) at the University of Michigan. The data collection included using Apple Watches and polysomnography (PSG), a sleep study in a laboratory. 39 subjects were recruited for the study, out of which 31 were chosen as final participants for the study. Restless legs syndrome, sleep-related breathing disorders, insomnia, parasomnias, central disorders of hypersomnolence, cardiovascular disease (including congenital heart disease, congestive heart failure, coronary artery disease, myocardial infarction, and cardiac arrhythmias), peripheral vascular disease, vision impairment not correctable by glasses or contacts, or any other conditions expected to result in significant neurological or psychiatric impairment were considered exclusion criteria. Individuals who had engaged in night shift work or transmeridian travel exceeding two time zones within the month before enrollment were also excluded. The Epworth Sleepiness Scale (ESS) was employed to eliminate participants with significant excessive daytime sleepiness, ensuring that their scores did not exceed 10, indicative of excessive daytime sleepiness.

The 31 subjects were provided with Apple Watches (Series 2 and 3, Apple Inc) applied to the wrists and a mobile application by OW that contained a sleeping diary. The watches were worn for 7-14 days before the sleep study (PSG), except during nights when the device was charged. This allowed the collection of the subjects' daily activity patterns. After the 7-14 day time period, the subjects underwent an 8-hour sleep monitored with PSG at the time of their natural bedtime. During the PSG, the subjects wore their Apple Watches to record heart rate and raw acceleration. The Apple Watch employs a triaxial MEMS accelerometer to gauge acceleration in the x, y, and z directions, quantified in units of g (9.8 m/s^2). The measurement of heart rate is facilitated by the Apple Watch using Photoplethysmography (PPG) on the dorsal aspect of the wrist. Acquisition of raw acceleration signals and heart rate data involves initiating a "Workout Session" on the device and utilizing functionalities inherent in the iOS WatchKit and HealthKit frameworks.

For each subject, there are four types of data available: motion, heart rate, steps, and labels. The date of the data was marked as seconds, where the date=0 denoted the start of PSG. Therefore, the date for heart rate, motion, and steps started from negative seconds.

The following types of data are provided:

- **heart rate (bpm):** Recorded from the Apple Watch and saved as txt files with the naming convention '[subject-id-number]_heartrate.txt'

  Each line in this file has the format: *date (in seconds since PSG start), heart rate (bpm)*

  Around 7000 data points for each subject

- **motion (acceleration)**: Recorded from the Apple Watch and saved as txt files with the naming convention '[subject-id-number]_acceleration.txt'

Each line in this file has the format: *date (in seconds since PSG start), x acceleration (in g), y acceleration, z acceleration*

Around a million data points for each subject

- **labeled sleep**: Recorded from polysomnography and saved in the format '[subject-id-number]_labeled_sleep.txt'

  Each line in this file has the format: *date (in seconds since PSG start) stage (0-5, wake = 0, N1 = 1, N2 = 2, N3 = 3, REM = 5)*

  Around 1000 data points for each subject.

- **steps (count):** Recorded from the Apple Watch and saved in the format '[subject-id-number]_steps.txt'

  Each line in this file has the format: *date (in seconds since PSG start), steps (total in bin from this timestamp to next timestamp)*

  Around 1200 data points for each subject

There were a few missing values in the Motion data. In addition, some subjects had cropped data from the night of the PSG. This was explained to be due to a failing battery on the Apple Watch. Moreover, for some subjects, there were labels in the Labels data equal to -1 which was not explained in the dataset.

3.1 Preprocessing the data

In my project, the missing values in the Motion data were dropped, since the number of missing data points which were insignificant. The labels defined as -1 were replaced with 0 since the positions for the -1 labels were mostly at the beginning or the end of the study.

The Pandas DataFrames were formed for each feature (heart rate, steps, labels, and motion) by looping through the 31 .txt files and matching the data with the subject's ID collected from the file name. Each data frame has columns Minutes, ID, and the measurement type corresponding to each data frame. Finally, the four data frames were merged into one based on 'ID' and 'Minutes'. I categorized each timestamp as either 'Asleep' (stages 1, 2, 3, 5) or 'Awake' (stage 0).

The data was very high-resolution, so I downsampled the data for easier comparison and to make the merging of data sets easier. I did that by binning the data into 10-minute intervals. For steps, motion, and heart rate, the value was obtained by taking the average of the values within that 10-minute interval. For labels, the value was obtained by taking the most prevalent sleep stage within that 10-minute interval. Each data frame has the same time intervals, starting at -10 070 min, which corresponds to 7 days, and ending at +500 min, which corresponds to around 8 hours, i.e. the end of the PSG. For subjects that do not have data recorded within these minutes, the values will be 0.

For creating a model to predict the sleep phases, I will mainly focus on heart rate, motion, and labels. The step data is used to identify whether an active lifestyle results in better sleep quality.

To exploratory data analysis and to obtain a clear view of the subjects' sleeping cycles and activity patterns, I created a new activity data frame that consists of the following columns: 'TotalTimeInBed', 'TotalSleepTime', 'DeepSleep' (Stage=3), 'AvgSteps', 'RestingHR', and 'Activity' for each subject. The values for these columns were the mean values from the original data set, and the resting heart rate was obtained as the average heart rate during the sleep study. Since we know that the sleep study lasted for 8 hours (Walch et al., 2019), every subject gets the value of 480 minutes to 'TotalTimeInBed'. I naively categorized the subjects as 'Active', if the 'AvgSteps' >= 10 000, and 'Inactive' if it was less.

The methods section should describe the analysis methods, for example, those covered in the course programming assignments, that are used in the project to obtain the results. These methods should include descriptive and data analysis methods as well as other techniques such as clustering, dimension reduction, and classification methods, depending on the task. The choice of the methods used should be justified in the report. The usage of the methods should be documented and attached to the report.

4. Methods

This project aims to analyze the sleep and activity patterns of the subjects on the group level and an individual level, as well as train a model to predict if the subject is asleep or awake at a certain time during the night. The data was very high-resolution, meaning that the timestamps were very specific and therefore difficult to use for analysis. Therefore, I used dimension reduction to bin the data into 10-minute intervals. The values for heart rate, steps, and motion in the x, y, and z directions were obtained by taking the average of the values within the 10-minute interval. The values for labels were obtained by taking the most prominent stage within the 10-minute interval.

The sleep stages were labelled from 0 to 5, where 0=wake, 1=N1, 2=N2, 3=N3, and 5=REM. Since the objective of my work was to train a model to identify whether the subject is asleep or awake, I categorized the 10-minute intervals to either 'Awake', if the Stage was 0, and 'Asleep', otherwise. First, I will explain the methods for training a logistic regression model.

4.1 Logistic Regression

In this project, I use a logistic regression model to predict whether the subject is asleep or awake at a certain 10-minute time. Logistic regression was chosen because the classification problem was binary, i.e. there were two possible outcomes: 'Asleep' or 'Awake'. For logistic regression, I used only the data from the sleep study, i.e., from minute 0 onwards. The predictors chosen to train the model were heart rate

and motion in the x, y, and z directions. The target was the type, either 'Asleep' or 'Awake', where 'Asleep' was the positive prediction.

I used scikit-learn implementation for Logistic Regression. To standardize the features, I used Standard Scaler. This ensures that the features contribute equally to the calculations in the logistic regression model. Finally, I used stratified K-Fold cross-validation to assess the model's performance. It ensures that each fold has the same proportion of observations with a given label and provides a more reliable measure of the model's performance than a single train-test split. (scikit-learn, Pedregosa, 2011)

To evaluate the performance of the binary classifier, I used different model evaluation metrics:

1. Accuracy
    - The accuracy represents the proportion of correctly classified instances out of the total instances, giving a good view of the model's performance.
2. Precision
    - The precision is the proportion of true positive predictions out of all positive predictions. It is used to measure the classifier's accuracy in identifying positive instances. A higher precision means a lower occurrence of false positives.
3. Recall
    - Recall (Sensitivity or True Positive Rate) is the proportion of actual positives that were predicted by the model out of all actual positives. It is used to measure the classifier's accuracy in capturing all relative instances. A higher recall means a minimal number of false negatives.
4. F1-score
    - The F1 score is the harmonic mean of precision and recall. A higher score means a good balance between precision and recall.

4.2 Exploratory Data Analysis

For observations on the group level, I used the activity data frame explained in the Dataset section. In the visualizations, I used the categorized activity ('Active'/'Inactive') types as colours to identify possible differences among them. The visualizations I used included a violin plot of the amount of deep sleep among the activity types, histograms of the activity measures, and joint plots of average steps and resting heart rate, as well as average steps and the amount of sleep. The visualizations and results of the group-level observations will be discussed in the next section.

For observations on an individual level, I chose one subject (ID = 6220552) and created an average data frame of all of the subjects for comparison. I used line plots to compare the number of steps among the individual and the group average, as well as the sleep stages.

5.Results

5.1 Logistic Regression

I created a logistic regression model to predict if an individual is asleep or awake. The predictors were heart rate and motion in the x, y, and z directions. The target was the type, either ‘Asleep’ or ‘Awake’, where ‘Asleep’ was treated as the positive label. After training and scaling the data, I calculated the accuracy, precision, recall, and F1 scores. The results were the following:

1. Accuracy (89.61%): The model is correct about the sleep state nearly 90% of the time. Accuracy is a measure of overall correctness and, in this case, suggests that the model performs well on classifying both 'Asleep' and 'Awake' instances.

2. Precision (88.63%): When the model predicts 'Asleep', it is correct about 89% of the time. Precision is the ratio of true positive predictions to the total predicted positives. In this context, it means that when the model predicts someone is 'Asleep', it is accurate about 89% of the time.

3. Recall (99.92%): The model captures nearly all instances that are actually 'Asleep'. Recall, also known as sensitivity or true positive rate, measures the model's ability to correctly identify all relevant instances (in this case, instances of being 'Asleep'). A recall of 99.92% indicates that the model is very good at capturing instances when the person is actually 'Asleep'.

4. F1 Score (93.93%): An F1 score of about 94% indicates a good balance between precision and recall. The F1 score is the harmonic mean of precision and recall, providing a single metric that balances both false positives and false negatives. In this case, a high F1 score suggests a good trade-off between precision and recall.

Figure 1. shows the confusion matrix for the type ‘Asleep’.

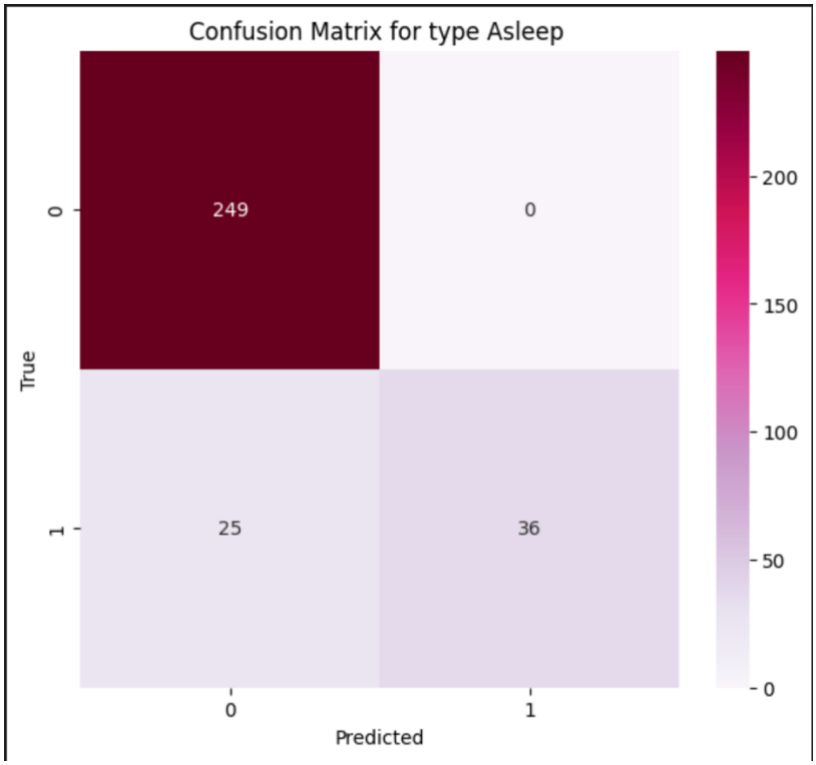

FIGURE 1. CONFUSION MATRIX

In summary, the model seems to perform well, especially in correctly identifying instances of being ‘Asleep’ (high recall) while maintaining a good balance with precision. However, if false positives (predicting 'Asleep' when the person is 'Awake') have significant consequences, the model might need to be fine-tuned further to reduce false positives, even if it means sacrificing some recall.

5.2 Exploratory Data Analysis and Visualizations

I categorized the subjects as either ‘Active’ or ‘Inactive’ based on their daily average steps. The subject was labelled as ‘Active’ if the average steps were over 10 000, and ‘Inactive’ otherwise. Out of the 31 subjects, there were 18 Inactive and 13 Active subjects. The distribution of average steps taken as Inactive and Active is seen in Figure 2. The distributions seem to be quite binomially distributed among the activity types.

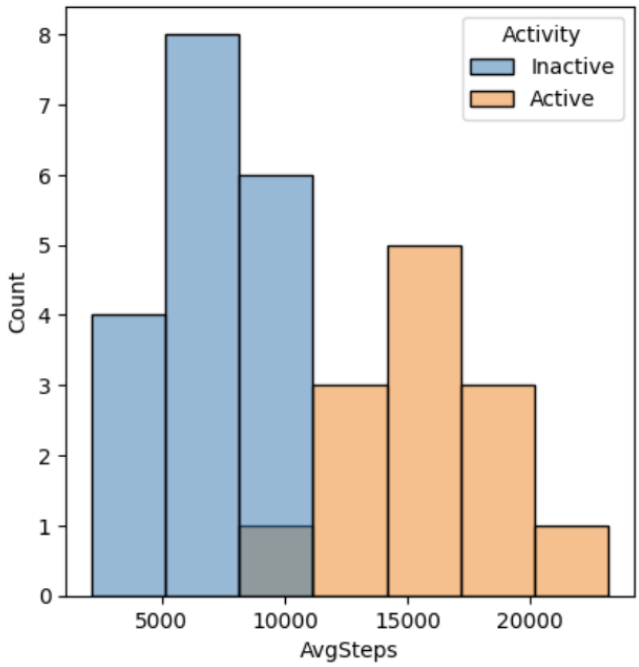

FIGURE 2. DISTRIBUTION OF AVERAGE STEPS

One objective of this study was to identify what contributes to a longer deep sleep phase. The deep sleep phase is defined as the N3 stage, which equals stage number 3 in this data. I calculated the amount of deep sleep for each subject and made a violin plot to visualize the results for both Activity types.

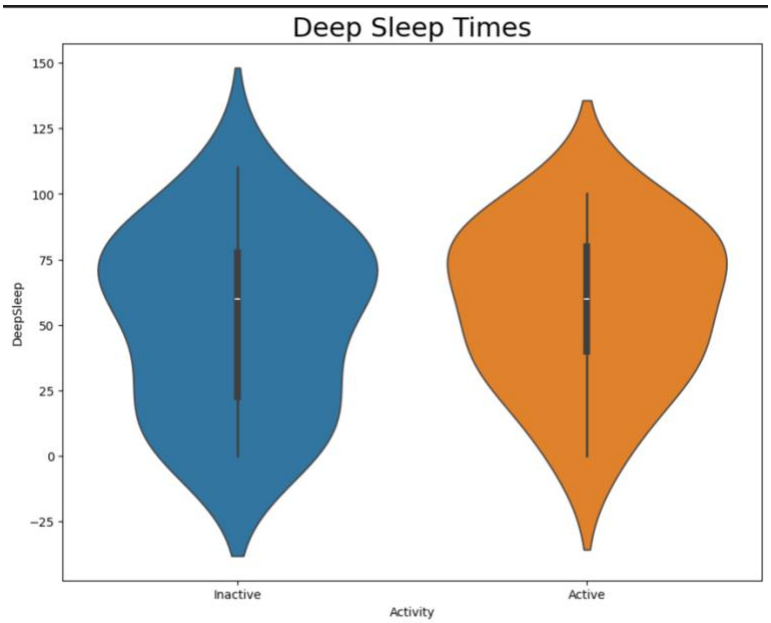

FIGURE 3. VIOLIN PLOT OF DEEP SLEEP TIMES

As we can see from Figure 3., for the ‘Active’ type the deep sleep phase is more likely, whereas for the ‘Inactive’ type there are more lower deep sleep values. This could imply that an active lifestyle (average steps >= 10 000) could result in a longer deep

sleep phase. Figure 4. shows a scatter plot, with average steps on the x-axis and deep sleep minutes on the y-axis to see if this is the case.

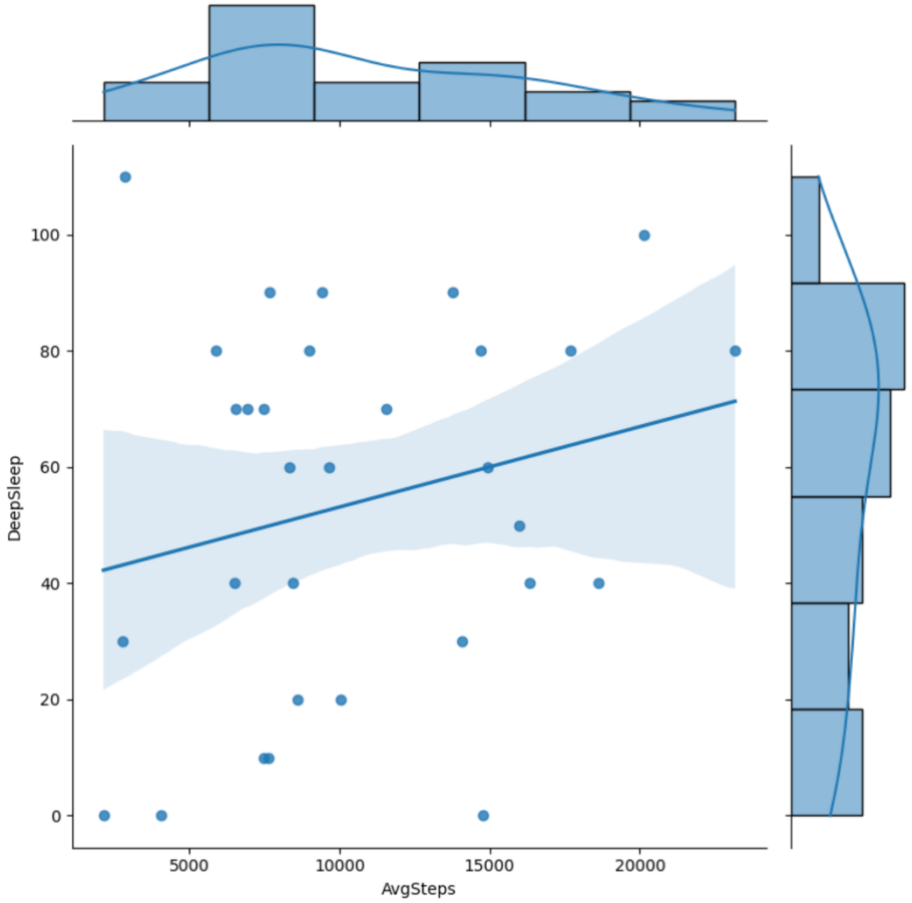

FIGURE 4. AVERAGE STEPS AND DEEP SLEEP

According to these visualizations, the subjects whose average steps were higher got longer deep sleep. Therefore, it could be deduced that the more active lifestyle, the better the sleep quality. Next, I will observe if an active lifestyle affects the subjects' resting heart rate.

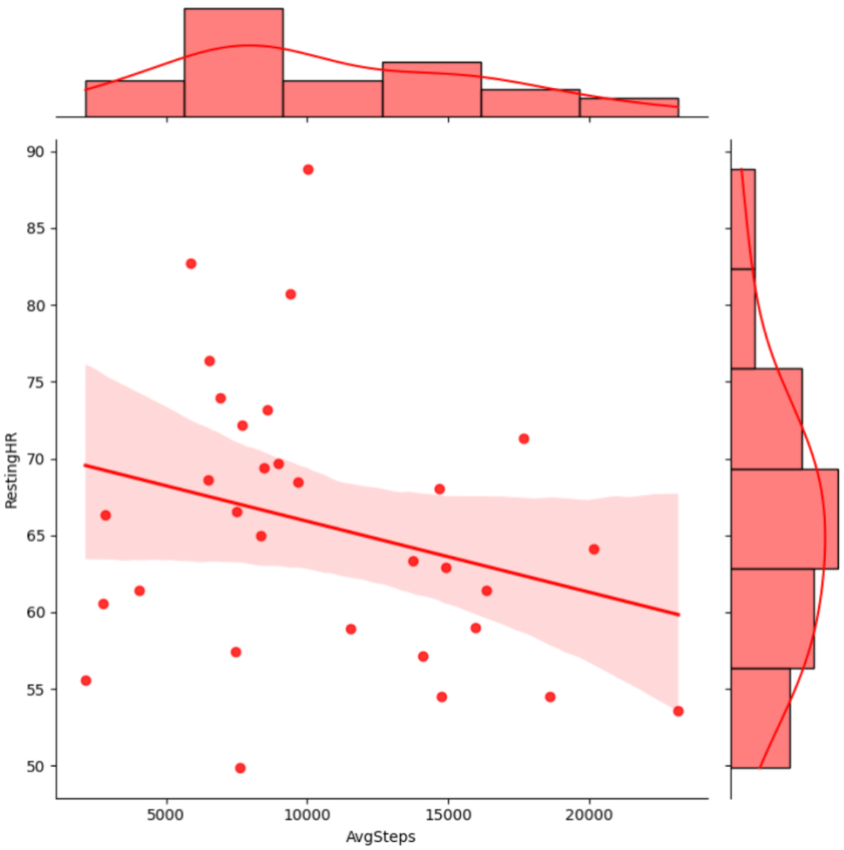

FIGURE 5. AVERAGE STEPS AND RESTING HEART RATE

Figure 5. visualizes how the higher average steps result in a lower resting heart rate (i.e., heart rate during sleep).

To alleviate these results, I compared an individual's values to the average values of the group. The individual chosen for this has an 'Active' activity type with around 20 000 steps per day. Table 1. presents the results and Figure 6. Shows the comparison of the average steps and resting heart rate of the individual versus the average of the group.

	Individual	**Group Average**
Average daily steps	20150	10552
Total Sleep Time (min)	470	401
Deep Sleep Time (min)	100	54
Average resting heart rate (bpm)	64	66

Table 1. Subject vs. average

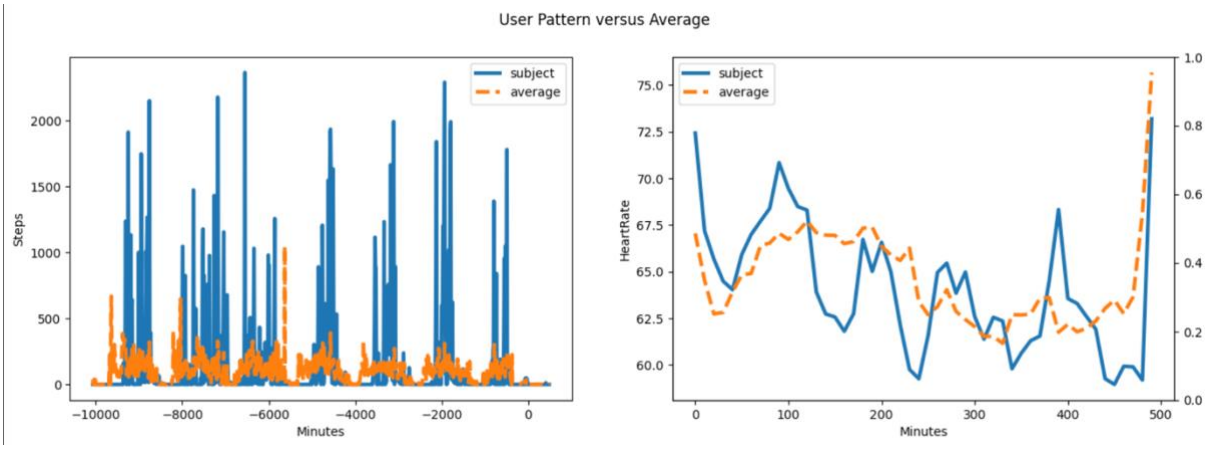

FIGURE 6. USER PATTERNS VS. AVERAGE

The individual's average steps are a lot higher than the group average. In addition, the amount of total sleep time as well as the deep sleep time is a lot higher. According to the results obtained from the previous group-level observations, the higher amount of deep sleep could be explained by an active lifestyle the individual seems to have.

On the contrary, the average resting heart rate during the night is not significantly lower compared to the group average as it could be deduced to be according to the significant difference in the activity between the individual and the group average, and the results from the group-level observation (Figure 5.). However, as seen in the right-hand side plot in Figure 6., there is more variety in the individual's heart rate compared to the average. This could be explained by the different sleep stages, that affect the heart rate. For example, the heart rate could drop during non-rapid eye movement (NREM), i.e., the deeper sleep phases, and increase during rapid eye movement (REM) sleep (Boudreau et al., 2013). There is also a significant increase in the heart rate at the end of the study, which is a sign of waking up and therefore heart rate increasing.

6.Conclusion and Discussion

All in all, from the group-level observations as well as the individual observations, it can be deduced that a higher average step count and an active lifestyle have an impact on better quality sleep since the deep sleep phase is more likely to be longer for people with an active lifestyle. In addition to the longer deep sleep phase, an active lifestyle results in a lower resting heart rate, which decreases the risk of cardiovascular diseases (Cooney et al., 2010).

Some further steps for this project could include fine-tuning the logistic regression to obtain fewer false positives. In addition, it could be fruitful to look at more classifications, such as Wake/N1+N2/N3/REM or Wake/N1/N2/N3/REM. In this project, I only used heart rate and motion to predict the sleep stages. In detecting for example sleep apnea, it could be useful to detect sound waves to notice snoring. Also, knowing the precise times of the measurements could be elaborate since the data set only presented seconds since the start of the PSG experiment. Therefore, it was not possible to create a precise sleep cycle for the subjects, for their bedtimes were unknown.

7. References

Arora, A., Chakraborty, P., Bhatia, M. P. S. (2020). Analysis of Data from Wearable Sensors for Sleep Quality Estimation and Prediction Using Deep Learning. Computer Engineering and Computer Science, 45, 10793–10812. https://doi.org/10.1007/s13369-020-04877-w

Boudreau, P., Yeh, W.-H., Dumont, G. A., & Boivin, D. B. (2013). Circadian Variation of Heart Rate Variability Across Sleep Stages. *Sleep, 36*(12), 1919–1928. https://doi.org/10.5665/sleep.3230

Cooney, M. T., Vartiainen, E., Laakitainen, T., Juolevi, A., Dudina, A., & Graham, I. M. (2010). Elevated resting heart rate is an independent risk factor for cardiovascular disease in healthy men and women. *American Heart Journal, 159*(4), 612-619.e3. https://doi.org/10.1016/j.ahj.2009.12.029

Liang, Z. & Chapa-Martell, M. A. (2021). A Multi-Level Classification Approach for Sleep Stage Prediction With Processed Data Derived From Consumer Wearable Activity Trackers. Frontiers in Digital Health, Sec. Health Informatics, 3, Article 665946. https://doi.org/10.3389/fdgth.2021.665946

Pandas Development Team. (2022). Pandas Documentation: Release 1.3.3. Retrieved from https://pandas.pydata.org/pandas-docs/stable/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830.

Walch, O., Huang, Y., Forger, D., & Goldstein, C. (2019). Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device. Sleep, 42(12), zsz180. https://doi.org/10.1093/sleep/zsz180

Wang, C., Hao, G., Bo, J., & Li, W. (2017). Correlations between sleep patterns and cardiovascular diseases in a Chinese middle-aged population. *Chronobiology International, 34*(5), 601-608. https://doi.org/10.1080/07420528.2017.1285785