Aalto University

CS-C4100 Digital Health and Human Behavior - Course Project

Daily Activity Analysis with Fitbit Fitness Tracker Dataset

December 2023

1 Introduction

1.1 Context - Related work

In the contemporary pursuit of a health-conscious lifestyle, wearable devices have ascended to the forefront as indispensable tools, providing individuals with personalized insights and motivation to actively engage in physical activities. This assertion gains empirical support from the work of Ringeval et al. [2], whose exhaustive analysis of various studies reveals a demonstrable increase in walking behaviors through the seamless integration of *Fitbit* devices into daily routines. Beyond serving as mere accessories, these wearables function as constant companions, offering real-time feedback and fostering a symbiotic relationship with users, thereby substantiating the intrinsic relevance of wearable technologies in actively shaping positive lifestyle changes.

Embarking on a healthy lifestyle involves understanding individual capabilities and aligning efforts with evidence-based guidelines. The recent report by the World Health Organization (WHO) [3] plays a pivotal role in shaping these guidelines, recommending that adults engage in 150–300 minutes of moderate-intensity or 75–150 minutes of vigorous-intensity aerobic physical activity per week, or an equivalent combination. However, the report highlighted how almost 27.5% of adults do not meet these requirements. Crucially, the intensity levels of these previous mentioned activities are standardized through the utilization of the Metabolic Equivalent of

Task (MET)[7], a key indicator that facilitates the precise definition of exertion levels for a given physical activity.

Furthermore, the comprehensive significance of physical activity transcends mere step counts, as evidenced by Kilic et al.'s sophisticated exploration into sleep quality prediction using advanced algorithms like Random Forest and Convolutional Neural Network [5]. However, the subsequent chapters will elucidate the intricacies and challenges posed by limitations within our dataset, particularly the uniformity and insufficiency of *Sleep Efficiency* values (an indicator defined as ratio between time asleep and time spent in bed). Taking into consideration this latest metric, it is important to mention the work of Reed et al. [9], who discussed a potential extension of what is conventionally considered "time in bed," proposing a new type of denominator called *DSE* (duration of sleep episode). However, with the available data, it was not possible to delve into this metric.

1.2 Project objectives

In this project, the "*Fitbit Fitness Tracker Dataset*" was analyzed, containing data related to physical activity, sleep, and weight of various users, recorded through a Fitbit wearable device.

As the first task, an exploratory analysis was conducted to examine the extent and duration for which users effectively adhered to WHO guidelines. For simplicity, adherence to the *moderate intensity physical activity* requirement is denoted by the acronym ***MIPA***, while the *vigorous-intensity physical activity* is referred to as ***VIPA***. In the data analysis, only individuals classified as "Active" (refer to *ActivityType* definition in 3.2.1) were considered.

From this initial analysis, it emerged that, within the examined sample, users found it easier to meet the VIPA requirement rather than the MIPA requirement. This is attributed to the fact that the former demands less activity time, albeit at a higher intensity. Moreover, it was observed that as the number of considered weeks increased, the count of users compliant with at least one of the two guidelines decreased, as expected due to the smaller sample size. However, a concerning finding is evident: **nearly 70% of the considered active users do not meet the VIPA requirement over a 4-week period**, and additionally, **97%** of them are not compliant with either one of the two requirements over the same period. In fact, during this time frame, only **3%** and **30%** manage to meet the MIPA and VIPA requirements, respectively.

The second objective of the project was to investigate a relationship between sleep and physical activity. Initially, the idea was to support Kline's thesis [10],

which posits a bidirectional connection between sleep and physical activity. However, with the available data, **it was not possible to confirm this assumption**. Nevertheless, a moderate negative correlation between hours slept and minutes dedicated to sedentary activities has been noticed. Further details are provided in the *Results* section 5.

An additional facet of exploratory analysis is introduced, focusing on activity patterns contingent upon MET and temporal allocations to bed rest, both within the overarching cohort and on an individual basis. This investigation encompasses the entirety of the week, with a specific focus on weekdays and weekends as distinct temporal entities. It became evident that the discernment of patterns in Moderate activity was notably facilitated by treating weekends and weekdays as discrete temporal segments rather than amalgamating them across the entire week. Noteworthy findings surfaced, revealing that during weekdays, the preponderance of individuals tends to engage in activities of Moderate intensity around midday, whereas the weekend exhibits a proclivity for activity during the early morning hours (approximately 10:00 AM) or the late afternoon (5:00 PM). Subsequently, in addition to exploring the classical linear relationship among distance traveled, physical activity, and calories expended, it was observed that, despite the absence of a discernible correlation between physical activity and sleep quality, distinct patterns emerged for the group.

Regarding the weekly trends in sleep duration, it was still feasible to conduct an analysis across the entire group and among individual participants, including the most somnolent and those with reduced sleep duration. It transpired that, on average, **users manage to attain approximately 7.29 hours of sleep per night**. Remarkably, only one user, identified as *4020332650*, consistently sleeps less than 6 hours per week. Additionally, it has been observed that there are correlations between users who sleep more and those who sleep less. In fact, it was noted that on **Saturdays, bi-weekly, is the day when users sleep less**. This is likely due to their participation in social events or engaging in recreational activities until the early hours of the morning.

2 Problem Formulation

From the latest report on the Global Action Plan on Physical Activity [4], the World Health Organization (WHO) has set a target to reduce the level of physical inactivity, particularly among adolescents, by 15% by 2030. In this regard, guidelines were proposed in 2020 to establish recommended thresholds for physical activity for both adults and adolescents. According to the WHO, adults should undertake 150–300 minutes of moderate-intensity or 75–150 minutes of vigorous-intensity physical ac-

tivity per week, or an equivalent combination of both. For children and adolescents, engaging in 60 minutes per day of moderate-to-vigorous intensity aerobic activity is considered suitable to attain health benefits.

In this project, the aim was to assess how many users in the dataset adhere to these guidelines. Unfortunately, due to the lack of specified age information for users, the analysis focuses solely on the adult population. As mentioned in the introduction, the parameters investigated in this study are Moderate-Intensity Physical Activity (MIPA) and Vigorous-Intensity Physical Activity (VIPA). To accomplish this, it is crucial to assess how the number of compliant users may vary with an increasing number of weeks considered (and thus an increase in data points). Therefore, focusing on users engaging in at least 60 minutes of Moderate or Very Active physical activity, three types of metrics are compiled: users compliant with both requirements (MIPA and VIPA), users compliant only with MIPA, and users compliant only with VIPA. Results are reported in the dedicated section.

Following Kline's paper [10], the objective is to explore the relationship between physical activity and sleep. According to this study, poor sleep is associated with lower levels of physical activity, while simultaneously, physical activity can help address sleep disturbances and improve sleep quality. To achieve this, it is necessary to investigate the relationship between actual sleep duration and other parameters recorded by the wearable device, such as the number of steps taken, calories burned, and minutes spent in different activity intensities (light, moderate, or very active) and others more.

Finally, given the anonymized nature of user data within the dataset and the absence of supplementary indicators pertaining to age and country of origin, an intriguing avenue of inquiry involves a nuanced exploration of distinctive patterns characterizing dataset users at both aggregate and individual levels. To undertake this investigation, a meticulous examination of the temporal dynamics inherent in the MET indicator becomes imperative, shedding light on prevalent instances of physical activity. This analysis will systematically delineate patterns across various temporal dimensions, encompassing weekdays, weekends, and the entirety of the week.

In concluding this study, a discerning exploration of supplementary patterns pertaining to sleep quality, operationalized as "*sleep efficiency*" as explicated in subsection 3.2.2, is envisaged. Additionally, an inquiry into the actual duration of sleep aims to ascertain whether the sampled cohort aligns with a judicious range of sleep hours, specifically exceeding seven hours [11], thereby contributing to a comprehensive understanding of the dataset's characteristics.

3 Dataset

3.1 Description

In this project, the *Fitbit Fitness Tracker Dataset*[1] was utilized. This dataset, available on the *Zenodo* website, contains anonymized data from 30 users collected through the *Amazon Mechanical Turk program*. Users were monitored during the period between March 12, 2016, and May 12, 2016. Since the dataset was divided into two periods (March-April and April-May), this report primarily analyzes the first dataset, which was available on the well-known *Kaggle* website. The dataset consists of 18 CSV files that include various parameters recorded by the Fitbit wearable device, such as physical activity, heart rate, time spent in bed, Metabolic Equivalent of Task (MET), and more. The data is provided at different temporal resolutions, ranging from daily aggregated data to minute-by-minute tracking or multiples of 5 seconds (in the case of heart rate).

The content of the entire dataset is reported in table 2. A first brief analysis of the dataset revealed that there were 33 users instead of 30 and that not all users had the same number of tracked records. Moreover, as mentioned earlier, the dataset has been anonymized, so each user is associated with a unique ID. For the project, emphasis was predominantly placed on the examination of the following sub-datasets: `dailyActivity_merged.csv`, `sleepDay_merged.csv`, and `minuteMETsNarrow_merged.csv`.

The first dataset considered, analyzed in Table 3, aggregates the key information pertaining to the physical activity of each individual. Note that for clarity, the column *FairlyActiveMinutes* will be subsequently renamed as *ModeratelyActiveMinutes* to ensure consistency with the *ModeratelyActiveDistance* column. This renaming is undertaken as both these measures, in terms of time and space, since they are referred to the same type of activity. It is also important to note that the various types of activities are defined based on the MET indicator. According to [7], "*One metabolic equivalent (MET) is defined as the amount of oxygen consumed while sitting at rest and is equal to 3.5 ml O2 per kg body weight x min*". Thus, this indicator enables the definition of intensity levels for a given physical activity. Moreover, the correspondence between METs and physical activity type for Fitbit wearable devices is provided, taking in consideration Semanik et al. work [8], in Table 1.

The information regarding the other two datasets are illustrated Table 4 and 5, respectively. Even though METs dataset and SleepDay dataset have the same time format, METs has a minute resolution, while SleepDay just a daily one.

METs range	Activity Type
≤ 1.5	Sedentary
1.5 – 3	Light
3 – 6	Moderate / Fairly
> 6	Very active

Table 1: Mapping between Fitbit activity type and required MET indicator

Dataset Name	**Rows**	**Columns**
dailyActivity_merged.csv	940	15
dailyCalories_merged.csv	940	3
dailyIntensities_merged.csv	940	10
dailySteps_merged.csv	940	3
heartrate_seconds_merged.csv	2483658	3
hourlyCalories_merged.csv	22099	3
hourlyIntensities_merged.csv	22099	4
hourlySteps_merged.csv	22099	3
minuteCaloriesNarrow_merged.csv	1325580	3
minuteCaloriesWide_merged.csv	21645	62
minuteIntensitiesNarrow_merged.csv	1325580	3
minuteIntensitiesWide_merged.csv	21645	62
minuteMETsNarrow_merged.csv	1325580	3
minuteSleep_merged.csv	188521	4
minuteStepsNarrow_merged.csv	1325580	3
minuteStepsWide_merged.csv	21645	62
sleepDay_merged.csv	413	5
weightLogInfo_merged.csv	67	8

Table 2: Dataset files and their sizes

Column name	Data Type	Description
Id	int64	Unique identifier of each participant
ActivityDate	date	Date value in MM/DD/YYYY format
TotalSteps	int64	Total number of steps taken
TotalDistance	float64	Total amount of kilometers reached
TrackerDistance	float64	Kilometers tracked by the device
LoggedActivitiesDistance	float64	Kilometers recorded through intentional user actions, such as logging a workout session.
VeryActiveDistance	float64	Kilometers reached during very active activities
ModeratelyActiveDistance	float64	Kilometers reached during moderately active activities
LightActiveDistance	float64	Kilometers reached during light active activities
SedentaryActiveDistance	float64	Kilometers reached during sedentary activities
VeryActiveMinutes	int64	Minutes spent during very active activities
FairlyActiveMinutes/ModeratelyActiveMinutes	int64	Minutes spent during fairly/moderately active activities
LightlyActiveMinutes	int64	Minutes spent during light active activities
SedentaryMinutes	int64	Minutes spent during sedentary activities
Calories	int64	Burned calories

Table 3: dailyActivity_merged.csv description

Column name	Data Type	Description
Id	int64	Unique identifier of each participant
SleepDay	date	Date value in MM/DD/YYYY H/M/S format with daily resolution
TotalMinutesAsleep	int64	Amount of minutes effectively spent sleeping
TotalTimeInBed	int64	Amount of minutes spent in bed

Table 4: sleepDay_merged.csv description

Column name	Data Type	Description
Id	int64	Unique identifier of each participant
ActivityMinute	date	Date value in MM/DD/YYYY H/M/S format with minute resolution
METs	int64	Metabolic Equivalent of Task

Table 5: minuteMETsNarrow_merged.cs description

3.2 Data analysis and preprocessing

As previously mentioned, an initial analysis of the dataset revealed that 33 users contributed to its development, surpassing the number declared in the Kaggle description (30). Nevertheless, some dataframes exhibit a lower number of users, specifically 12, 24, and 8 for the datasets related to heart rate, sleep, and weight logging, respectively. No datasets exhibited any missing values. Concerning duplicates, they were only identified in the datasets pertaining to sleep and weight logging. Furthermore, the weight logging dataset appeared notably inconsistent and small when compared to the other dataframes; therefore, it will not be taken into account in the upcoming sections.

A first analysis is proposed separately for `dailyActivity_merged.csv` and `sleepDay_merged.csv` datasets. This distinct analysis is undertaken due to a substantial reduction in the dataset's size in a subsequent phase, following the merging of both datasets. **The size of the considered dataset has been significantly diminished, plummeting from 925 to 353 records.** As a result, this alteration has influenced the data distribution and average values. However, a more detailed explanation of this observation will be provided in the following sections.

3.2.1 Activity dataset

During an initial data analysis, it was observed that less than 2% of the records had different values between the variables TrackedDistance and TotalDistance. Consequently, these outliers were removed to ensure that both columns could represent the same quantity. The result is depicted in Figure 1. Furthermore, only in 2% of the records did users intentionally log their workout sessions, effectively recording the distance covered. For this reason, the column "*LoggedActivitiesDistance*" has been removed.

Analyzing the entire user group, we can observe that 81.5% of the time spent during the day is dedicated to sedentary activities (including sitting, driving, or sleeping). In contrast, 15.7% of the time is allocated to light-intensity activities, and less than 2% is spent on truly active and moderately active activities. Regarding the distance covered, it is evident that the majority of the distance falls into the light activity category. For example, commuting to work or engaging in daily activities easily falls within this category. It is not surprising that the distance associated with inactive activities accounts for only 0.03%. Inactive activities, by definition, involve users remaining stationary or covering relatively short distances (on the order of a few hundred meters). Lastly, we can observe that the Very Active distance covers almost 30% of the total distance traveled during a day. This is because during

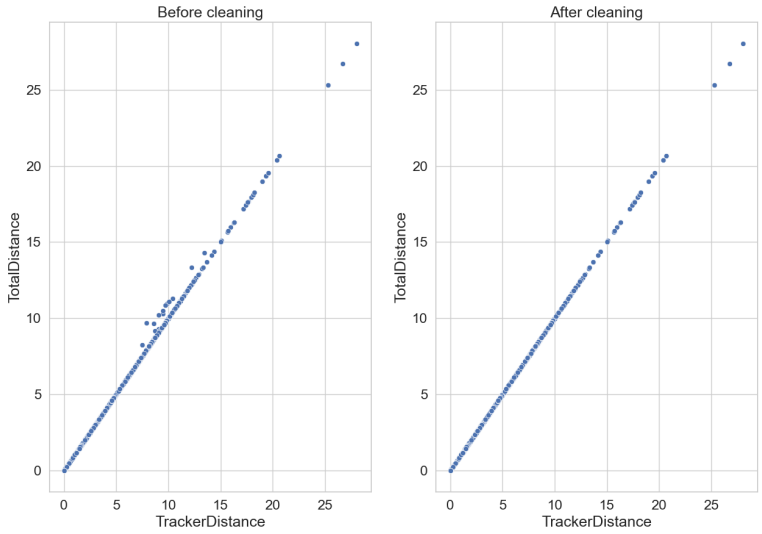


Figure 1: Relationship between Totaldistance and TrackedDistance before and after the cleaning

high-intensity activities such as running, it is easy to cover long distances in a short amount of time. These findings are summarized in the visualization presented in Figure 2.

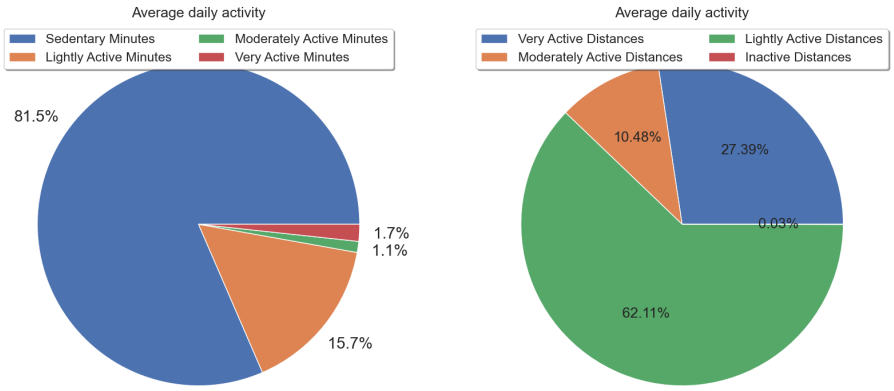


Figure 2: Average daily activity comparing type of time and activity performed

Finally, in order to extract additional insights from the data, the following features were introduced:

- *TotalActiveMinutes*, defined as the sum of *VeryActiveMinutes* and *ModeratelyActiveMinutes*,
- *TotalTrackedMinutes*, calculated as the sum of *TotalActiveMinutes*, *LightlyActiveMinutes*, and *SedentaryMinutes*,
- *ActivityType* categorized as "Active" for individuals with TotalActiveMinutes ≥ 60, otherwise labeled as "Inactive."

Considering these new variables, it can be observed from the countplot in Figure 3 that the majority of available records categorize individuals as inactive. Furthermore, only 6 out of 31 people have been active for 15 days (non-consecutive).

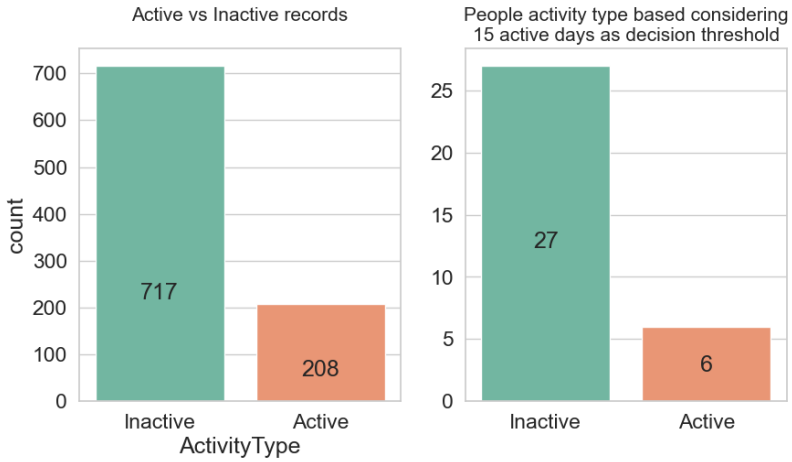


Figure 3: Dataset records and users counts categorised by their activity type

3.2.2 Sleep dataset

To preprocess the sleep dataset, a minimum threshold of 7 records per user was established. Consequently, users who recorded their sleep data for less than a week over the entire month were excluded from the dataframe. Through this thresholding process, the considered sample size decreased from 24 to 17 users. In Figure 4, a bar plot displays the number of records for each user.

Subsequently, the following features were created:

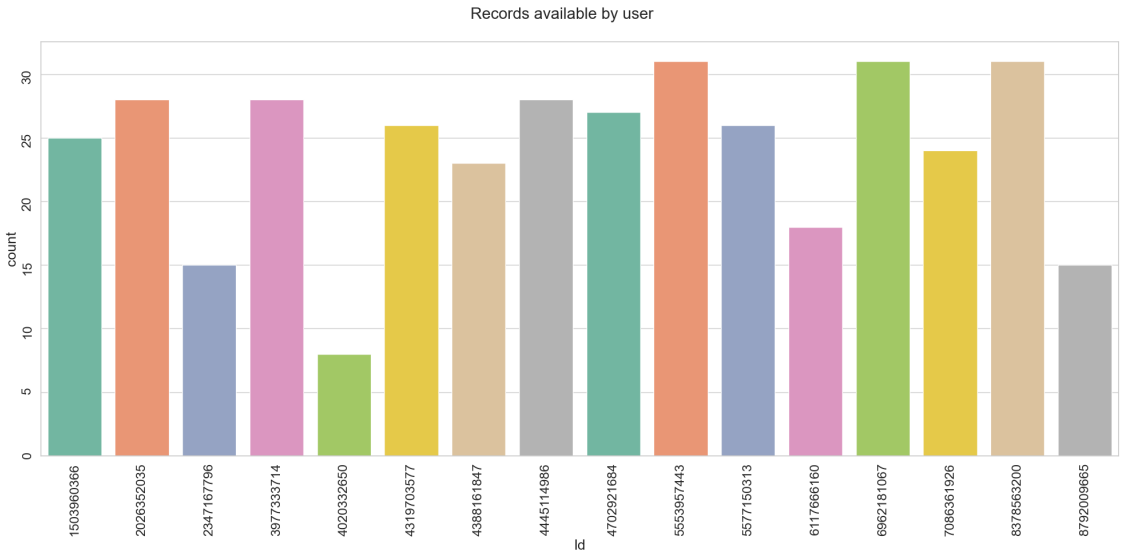


Figure 4: Sleep records available for each user over 31 days

- *HoursSlept*: *TotalMinutesAsleep* in hours
- *SleepEfficiency*: Defined as the ratio between HoursSlept and TotalTimeInBed.
- *Sleep type*: Including the categories *Oversleep*, *Adequate*, and *Inadequate* based on the number of hours slept. Specifically, more than 9 hours for the first category, between 7 and 9 hours for the second, and less than 7 hours for the third (According to [11] and [12]).
- *SleepScore*: $SleepEfficiency \cdot HoursSlept$

Then, analyzing the datapoints, the contribution of outliers has been noticed. For this reason a filtering of the records considering the *SleepEfficiency* has been applied based on the interquartile range (IQR), defined as the difference between the third quartile and the first quartile (IQR = Q3 − Q1). The result can be observed in figure 5.
Regrettably, during both the pre-processing and post-processing stages, a limited degree of variance has been observed in the values of Sleep Efficiency. Considering the relatively modest number of data points (356), the forthcoming sections will explicate the challenges encountered in discerning associations between sleep quality and physical activity.

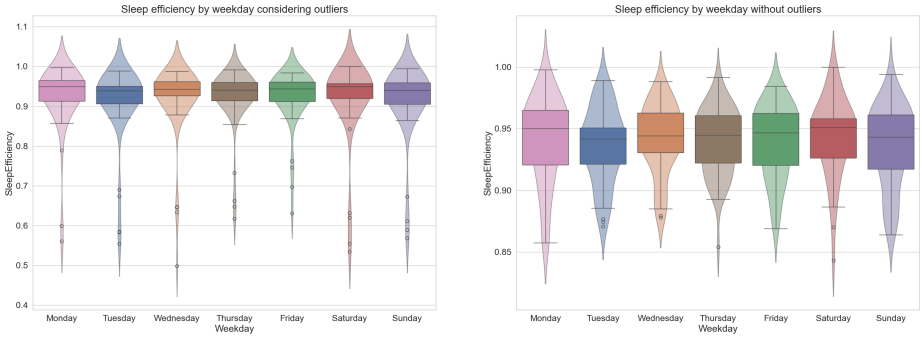


Figure 5: Sleep efficiency distribution by weekday before and after thresholding

The overall sleep distribution can be observed in Figure 6, since we performed a thresholding most of the people manifest an adequate sleep routine.

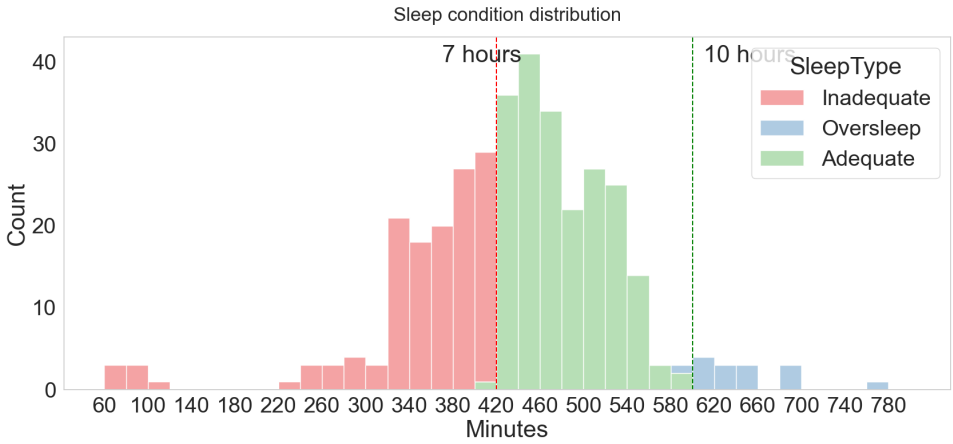


Figure 6: Sleep distribution according to SleepType

3.2.3 Merged dataset (activity + sleep)

By merging the two datasets, it can be observed that a substantial portion of records is lost, as the total user sample decreases from 33 to 17 users. Additionally, it is noteworthy that the mean values of key activity indicators undergo a significant change. This outcome is depicted in Figure 7. In addition, a histogram for each primary feature of the merged dataset is presented in Figure 8.

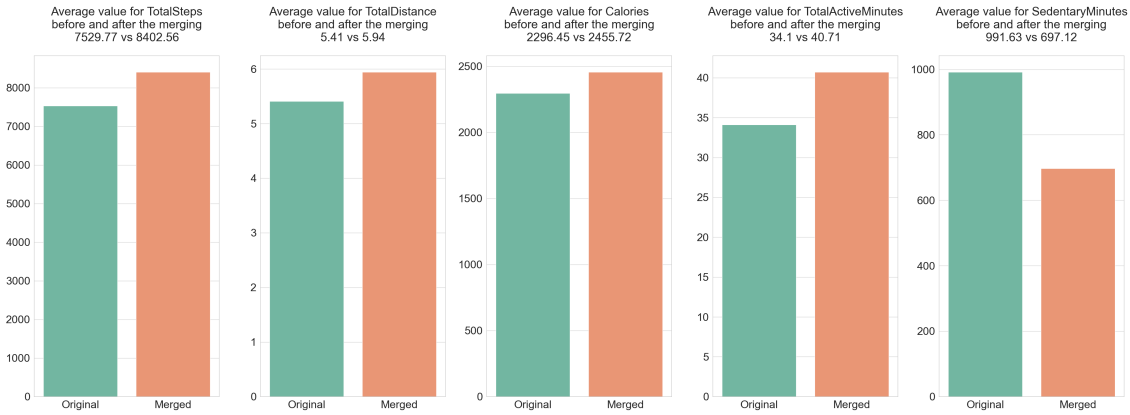


Figure 7: Average values of main dataset features before and after the merging

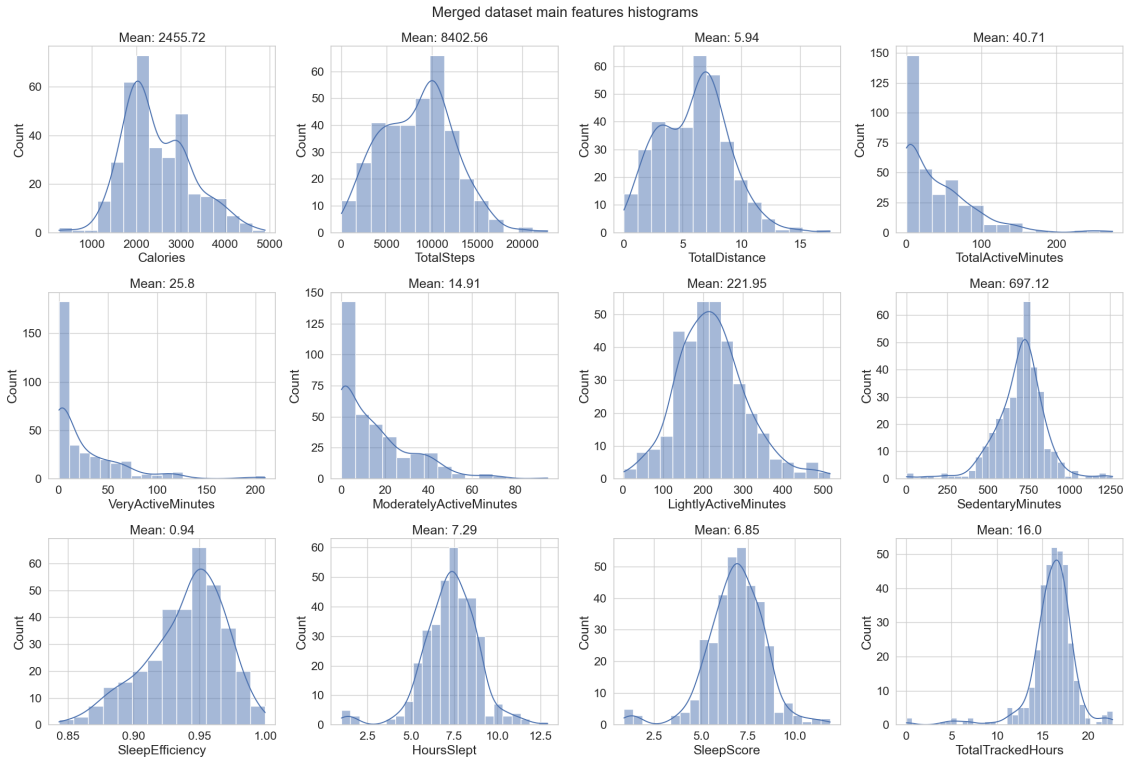


Figure 8: Merged dataset features histogram

3.2.4 METs dataset

In the analysis of both community and individual user patterns, the dataset in which the MET indicator is tracked minute by minute was utilized separately. To maintain consistency with the previously mentioned table (1), the MET value for each entry was divided by 10, as Fitbit wearable devices store this value by multiplying it by 10.

4 Methods

To work with the available data, the following set of *Python* libraries was employed: *Pandas*, *Matplotlib*, *Plotly*, *Seaborn*, and *Numpy*. Following data loading, a preprocessing operation was executed as detailed in the preceding section (3.2). Subsequently, a descriptive analysis of the available data pertaining to physical activity was conducted. These findings are documented and discussed in the section dedicated to results and conclusions (5, 6).

After separately analyzing the other two datasets (sleep dataset and METs dataset), a merging operation occurred between the sleep dataset and the physical activity dataset. As previously mentioned, this resulted in the loss of numerous records, as only 17 users demonstrated a sufficient number of data points in the sleep dataset.

Finally, to explore a potential relationship between sleep and physical activity, the ***Spearman*** correlation coefficient was employed. This statistical tool was chosen due to its independence from assumptions regarding the distributions of the variables involved and its reduced sensitivity to outliers when compared to the Pearson coefficient. Additionally, the Spearman coefficient facilitates the measurement of a monotonic relationship between variables, signifying that as one variable increases, the other may increase or decrease, but not necessarily at a constant rate.

5 Results

5.1 WHO guidelines

For this initial analysis, pie charts were utilized to visually represent the percentages of "*Active*" users on a weekly basis under three scenarios: users compliant with both guidelines, users compliant only with the MIPA guideline, and users compliant only with the VIPA guideline. As elucidated in the introductory chapter, the MIPA guideline entails 150 to 300 minutes per week of moderate-intensity physical activity, while the VIPA guideline necessitates 75 to 150 minutes per week of high-intensity

physical activity, in our case corresponding to *Moderately Active Minutes* and *Very Active Minutes*.

From the graphs presented in Figure 9, it is evident, firstly, that with an increasing number of weeks considered, there is a general trend of decreasing percentages of individuals compliant with either of the guidelines. Naturally, this percentage diminishes even more rapidly when both recommendations are taken into account. Specifically, the proportion of individuals compliant with both requirements sharply declines from 33% to 3% when considering only one week of activity and comparing it with a four-week period. This suggests the difficulty for the majority of users to consistently adhere to these requirements.

Secondly, it is noticeable that, generally, it is easier to meet the VIPA requirement compared to the MIPA one. In fact, considering a four-week observation period, almost 30% of active users manage to fulfill the VIPA requirement, while only 3% meet the MIPA one. A potential explanation for this phenomenon may be found by considering that the VIPA requirement calls for a lower number of active minutes compared to MIPA. Therefore, it is attainable with a lower number of workout sessions per week, which could probably require less commitment for the final users compared to other types of medium-intensity activities.

In stark contrast to the projections posited by the World Health Organization (WHO), wherein an estimated 27.5% of the adult populace is purported to contravene either of the two guidelines, the extant dataset divulges a disconcerting revelation: approximately **70% of the ascribed active users fall short of meeting the VIPA requirement**. Moreover, a staggering **97% of the scrutinized sample fails to align with either guideline over a 4-week temporal span**. It is plausible that this percentage could be even higher should inactive users be incorporated into the analysis.

Upon contemplation of these statistics, it becomes discernible that individuals within our dataset may not manifest the anticipated levels of activity, and notably, they grapple with achieving the minimal benchmarks proffered by the WHO. It is imperative to underscore that the demarcated sample, comprising 33 users, lacks comparability to the WHO's expansive statistical analyses. As elucidated in the introduction, the absence of supplementary user information, encompassing age and country of origin, further constrains the interpretability and generalizability of these findings.

5.2 Relationship between sleep and physical activity

In the examination of the interplay between sleep patterns and physical activity at the group level, an initial exploration involved the construction of a heatmap displaying Spearman correlation coefficients among various variables within the dataset, as presented in Figure 11. Notably, a robust correlation was observed among features sharing common objectives, such as tracking distance traveled and step count, as well as between minutes spent actively and the corresponding physical activity.

However, to delve into the nuanced influence of sleep duration and time spent in bed, specific visualizations were created (refer to 12 13). Although these graphs employ color-coded data points clustered by *SleepType*, the analysis reveals a lack of substantive correlation between physical activity and sleep. **Most correlation coefficients obtained remain below 0.27** (in absolute value). The solitary notable correlation pertains to the linkage between hours slept, and minutes spent in sedentary activities, demonstrating a **coefficient of correlation at -0.55**. This implies a moderate correlation between these variables. However, the ambiguity persists regarding whether SedentaryMinutes encapsulate time spent in bed, given their daily tracking nature.

Moreover, the distribution of sleep efficiency values, as illustrated in Figure 10, indicates a predominant prevalence of values exceeding 0.85, **suggesting an absence of users exhibiting poor sleep quality**. While data processing operations have certainly contributed to altering these outcomes and reducing their variance, an analysis conducted during the pre-processing phase revealed a deficiency of values below 0.85, with only one such instance.

Consequently, the evidentiary foundation for affirming a substantive relationship between physical exercise and sleep quantity/quality is deemed insufficient. Nevertheless, a subtle trend intimates that **as sleep duration increases, the duration of sedentary activities tends to decrease**, hinting at a marginal increase in overall activity. However, it is imperative to clarify that **this observation does not specifically pertain to physical exercise**.

In the absence of identified correlation between physical activity and sleep duration, an alternative visualization (Figure 14) is proffered. From this representation, it is observable that individuals with inadequate sleep quantity (< 7 hours) generally exhibit higher "*VeryActiveMinutes*" compared to those with adequate sleep, except on Thursday. This observation may stem from individuals seeking more strenuous physical activities in the hope of inducing fatigue and subsequently achieving better sleep quality. Alternatively, it could be postulated that more active individuals tend to sleep less, although this contradicts the assertion put forth by Kline [10]. Furthermore, in assessing calorie expenditure, it is noticeable that, excluding Tuesdays

and Fridays, we can observe a trend where individuals with inadequate sleep tend to burn more calories on Mondays and Wednesdays, while the opposite holds true on the remaining days.

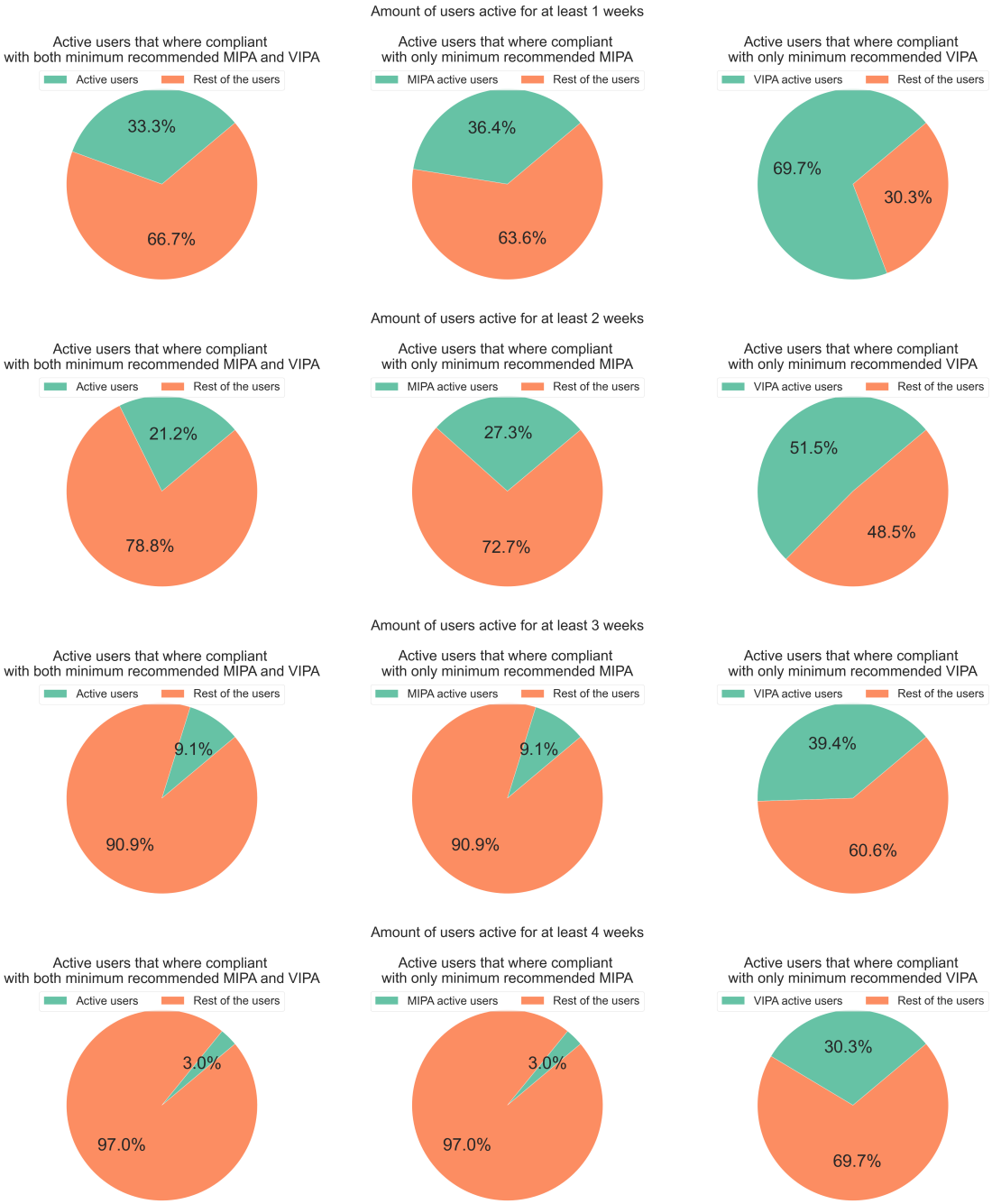


Figure 9: Pie charts of active users following WHO weekly requirements on physical activity
considering different timespans

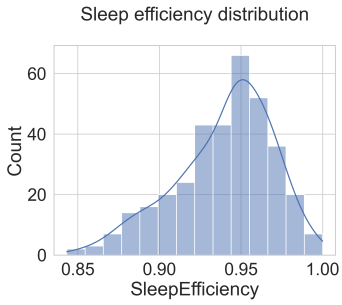


Figure 10: Sleep efficiency distribution

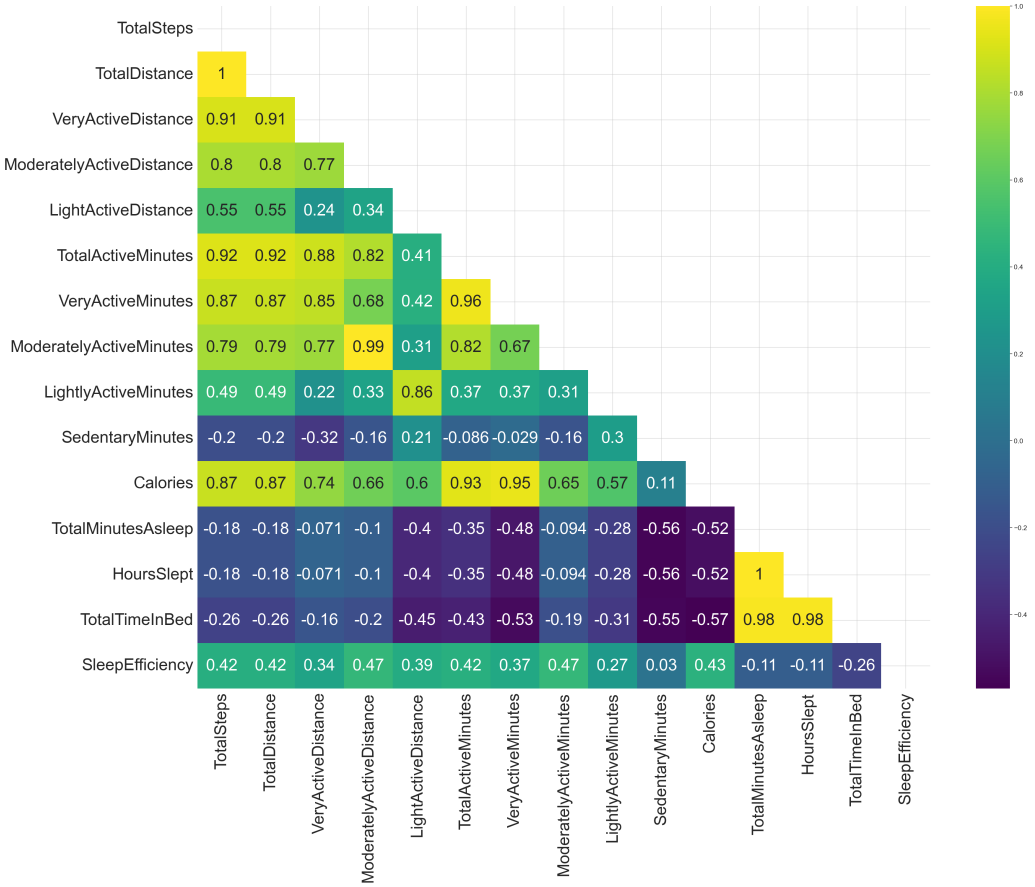


Figure 11: Spearman correlation matrix between activity and sleep indicators od the merged dataset

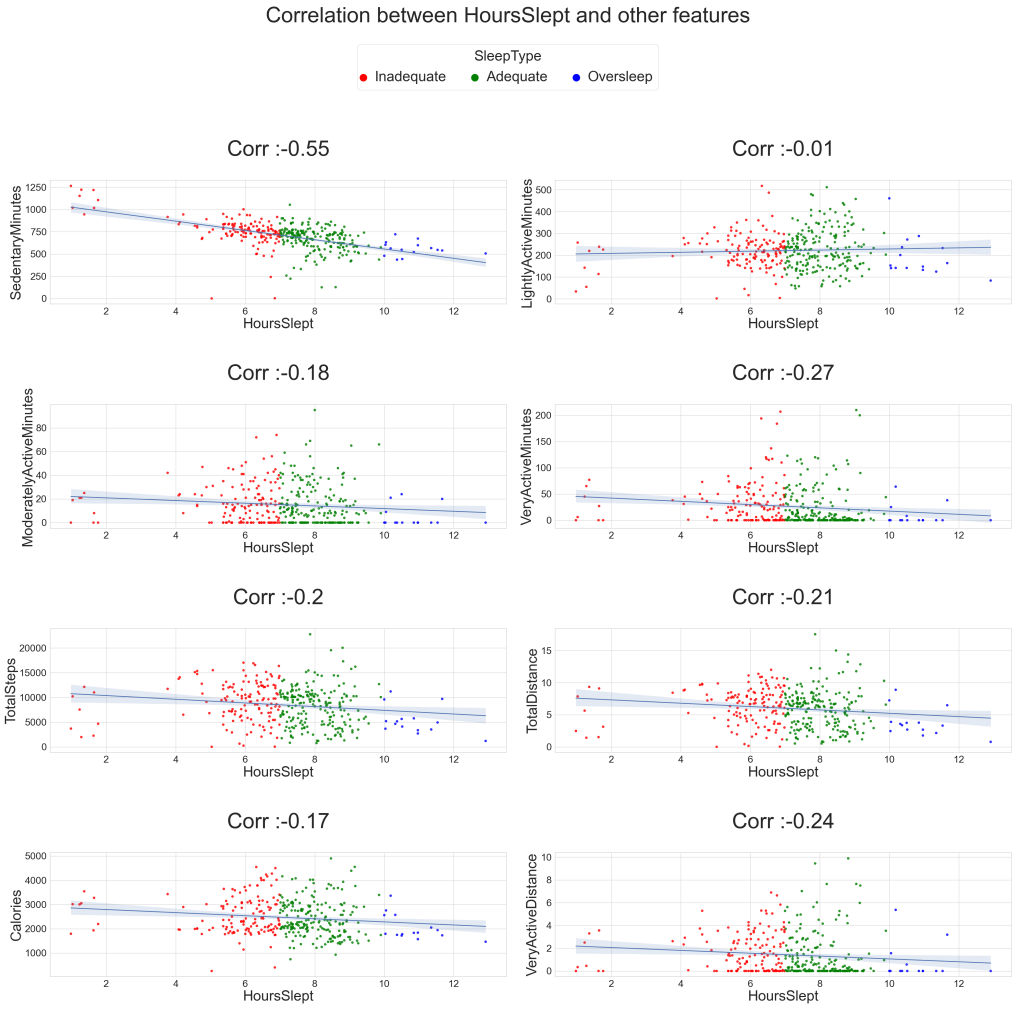


Figure 12: Correlation between Hours Slept and other activity indicators

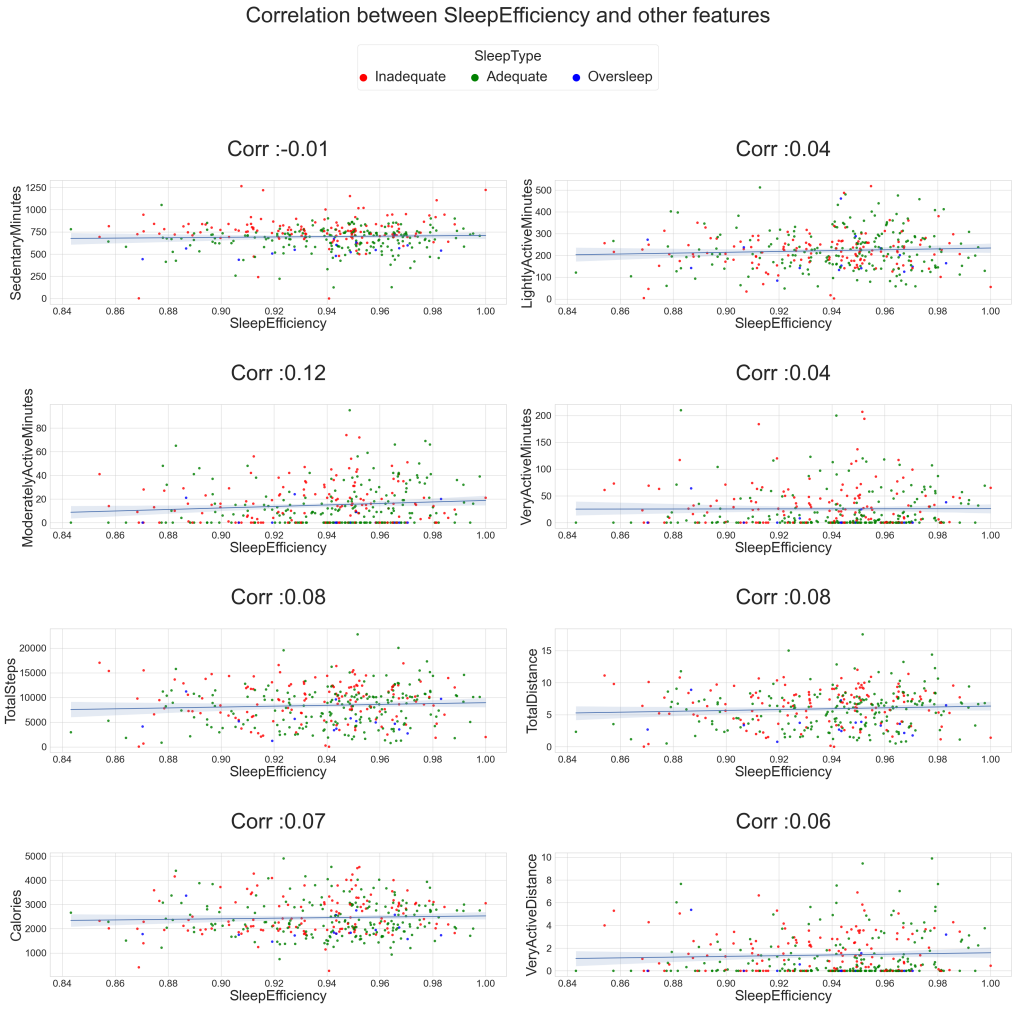


Figure 13: Correlation between Sleep Efficiency and other activity indicators

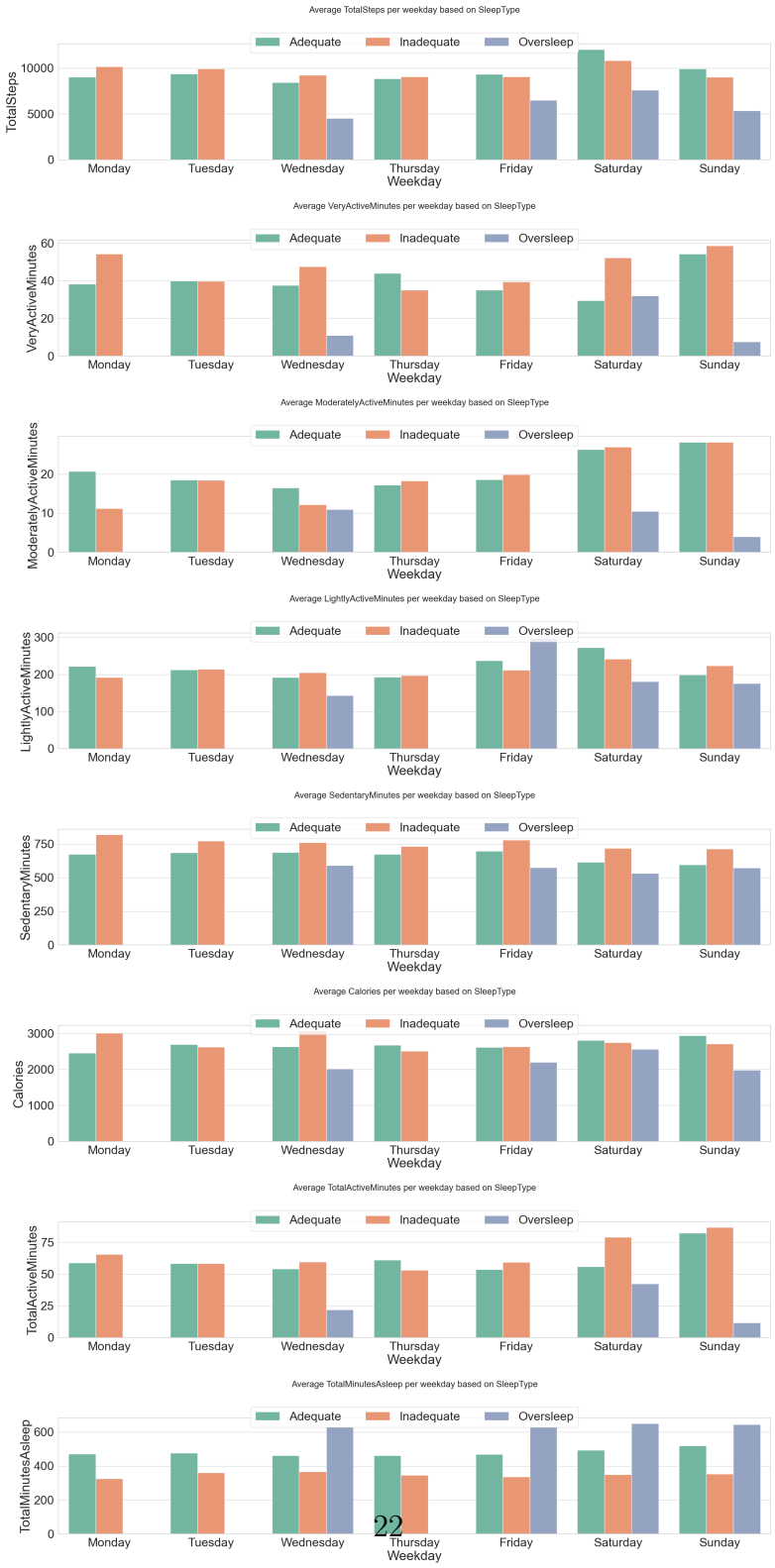


Figure 14: Average values for each feature based on Sleep Type

5.3 Physical activity analysis

Commencing with the most conspicuous observations (from Figure 11), it is noteworthy within this dataset that a linear relationship is evident among variables such as burned calories, distance covered, and total minutes of physical activity. A specific emphasis is placed on "VeryActiveMinutes," as these minutes substantially contribute to the overall energy expenditure. Figure 15 depicts the linear relationship between these main activity indicator and calories consumption.

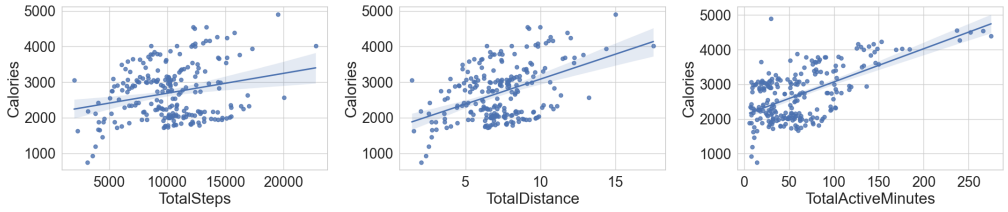


Figure 15: Relationship between Calories consumption, TotalSteps, TotalDistance and TotalActiveMinutes

Subsequently, the Physical Activity Intensity (MET) indicator, temporally averaged, was considered to identify users with the most significant patterns. In this context, MET values throughout the day were analyzed across three scenarios: the entire week, only weekdays, and only the weekend. From these analyses, the selected graph is presented in Figure 16, distinctly showcasing intriguing and consistent physical activity patterns for user *8378563200* (green plot, second column, fifth row). The chosen graph pertains specifically to weekdays. Upon comparing it with the weekend-specific graph, it became evident that there were no substantial and continuous activities during the weekend. This observation is likely attributed to users generally taking a rest during this period. Moreover, continuing to observe Figure 16, it can be noticed that during working days, the majority of individuals tend to exercise around noon. In contrast, on weekends, activities are more spread out, occurring either around 10 in the morning or around 17 in the afternoon.

In Figure 17, recurrent patterns for user 8378563200 are discernible. Specifically, during weekdays, a consistent physical activity occurs in the time window between 05:00 and 06:00. Notably, it is observed that only on Mondays and Tuesdays does the user engage in high-intensity activity during this period, while on Wednesdays, Thursdays, and Fridays, the activity reaches only moderate intensity. This pattern suggests that the user may participate in high-intensity activities like running on specific weekdays. Additionally, between 8:00 and 9:00, the user engages in another

form of moderate-intensity activity, potentially indicative of commuting to work using a bicycle, thereby burning more calories compared to using the car. Finally, during the weekend, the user predominantly engages in light-intensity activities, suggesting a tendency to be less active on weekends, possibly emphasizing rest.

5.4 Sleep activity analysis

The sleep analysis was conducted with a focused examination of three distinct users, with particular attention directed towards the user exhibiting the highest *sleepScore* (*6962181067*) and the two users manifesting the lowest *sleepScores* (*2347167796* and *4020332650*). The deliberate inclusion of multiple users in this investigation serves to underscore the potential influence of data quantity on the ultimate analytical outcomes. Observing the entire sample of users, it has been noted that, on average, they have an adequate daily quantity of sleep, amounting to **7.29 hours**.

The user boasting a sleepScore of 6962181067 demonstrated a cumulative sleepScore of 222.55, corresponding to an average nightly sleep duration of 7.47 hours, predicated on a dataset comprising 31 distinct datapoints. Illustrated in Figure 18, this user consistently maintained a sleep duration falling within the range deemed sufficient for nearly the entire month under scrutiny. It is noteworthy that on two consecutive days (Friday and Saturday, April 16 and 17, 2016), the user recorded approximately 6 hours of sleep each, and the day featuring the least amount of sleep was Saturday, May 7, 2016, with a total of **5 hours**.

Subsequent to the analysis of the user with the highest sleepScore, focus shifted towards the users with the lowest sleepScores, namely users 2347167796 and 4020332650. A comparative examination of the graphical representations for these two users accentuates the pivotal role played by the quantity of available datapoints in elucidating discernible sleep patterns (Figure 19). User 2347167796, characterized by a more extensive dataset (15 records), reveals instances where sleep falls below the recommended 7-hour threshold, notably on Fridays 22nd, Saturdays 23rd, Thursdays 28th, and Fridays 29th. In contrast, the dataset for user 4020332650, comprising only 8 records, lacks sufficient granularity to identify potential sleep patterns comprehensively. From the available data, it is discernible that the days meeting the minimum sleep threshold for this user were confined to Tuesdays and Wednesdays.

From this introductory exploratory analysis, we can observe that it is likely that the first user, with a **bi-weekly** frequency, does not sleep an adequate number of hours, especially on Saturdays. This could be attributed to the possibility that during the weekend, this individual may engage in social activities or recreational pursuits, leading to consistently low sleep durations.

While, for user 4020332650, we cannot draw substantial conclusions due to the limited number of datapoints (just one more than the threshold defined during preprocessing), we can instead note that user 2347167796 exhibits a sleep pattern similar to someone who sleeps more. Therefore, despite the lower sleep score, the average number of hours slept is comparable between these two users. Consequently, it can be inferred that **both follow a similar routine, choosing Saturdays, with a bi-weekly frequency, for their leisure activities.**

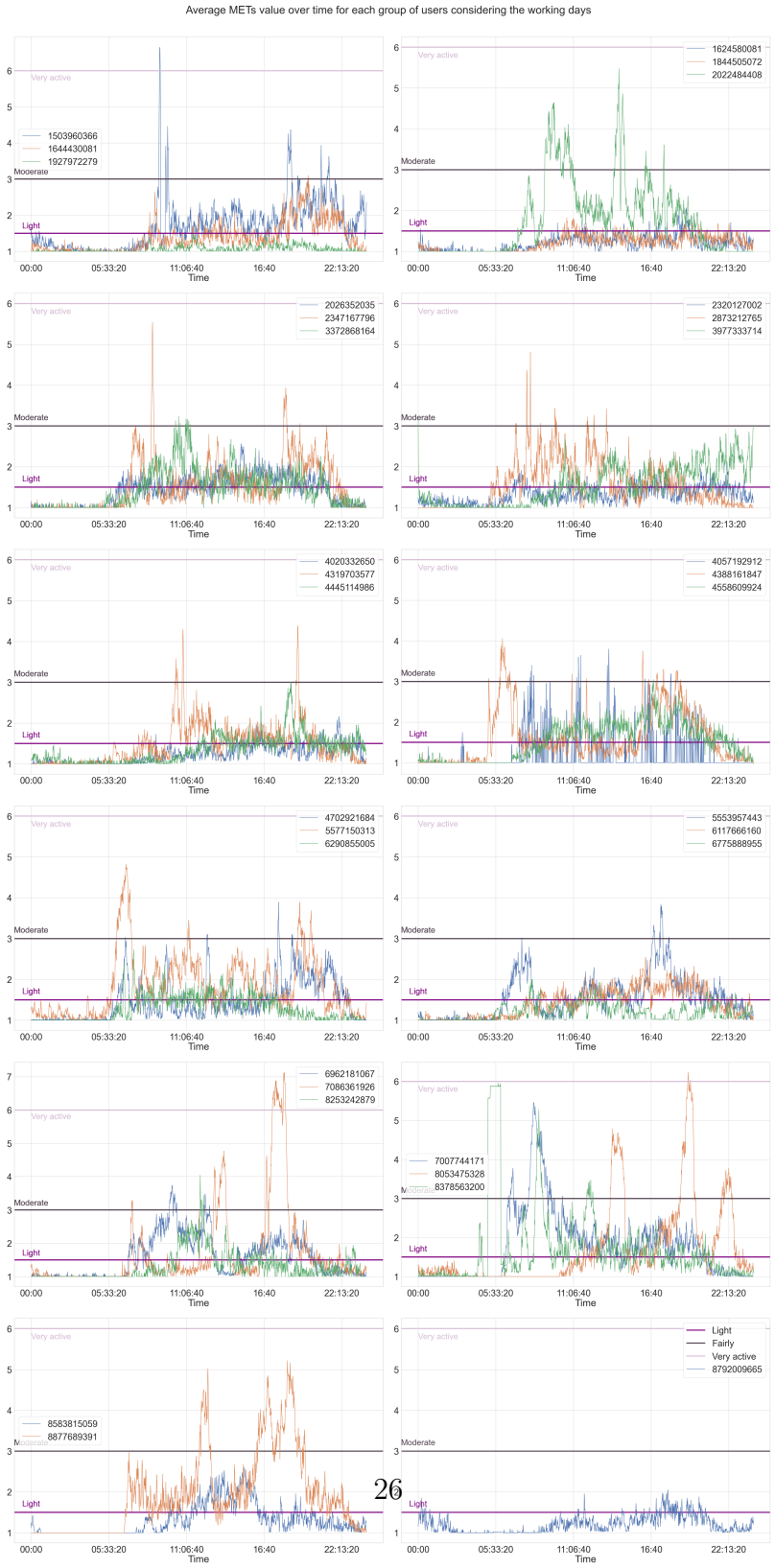


Figure 16: Average MET values during the day for each user considering only working days

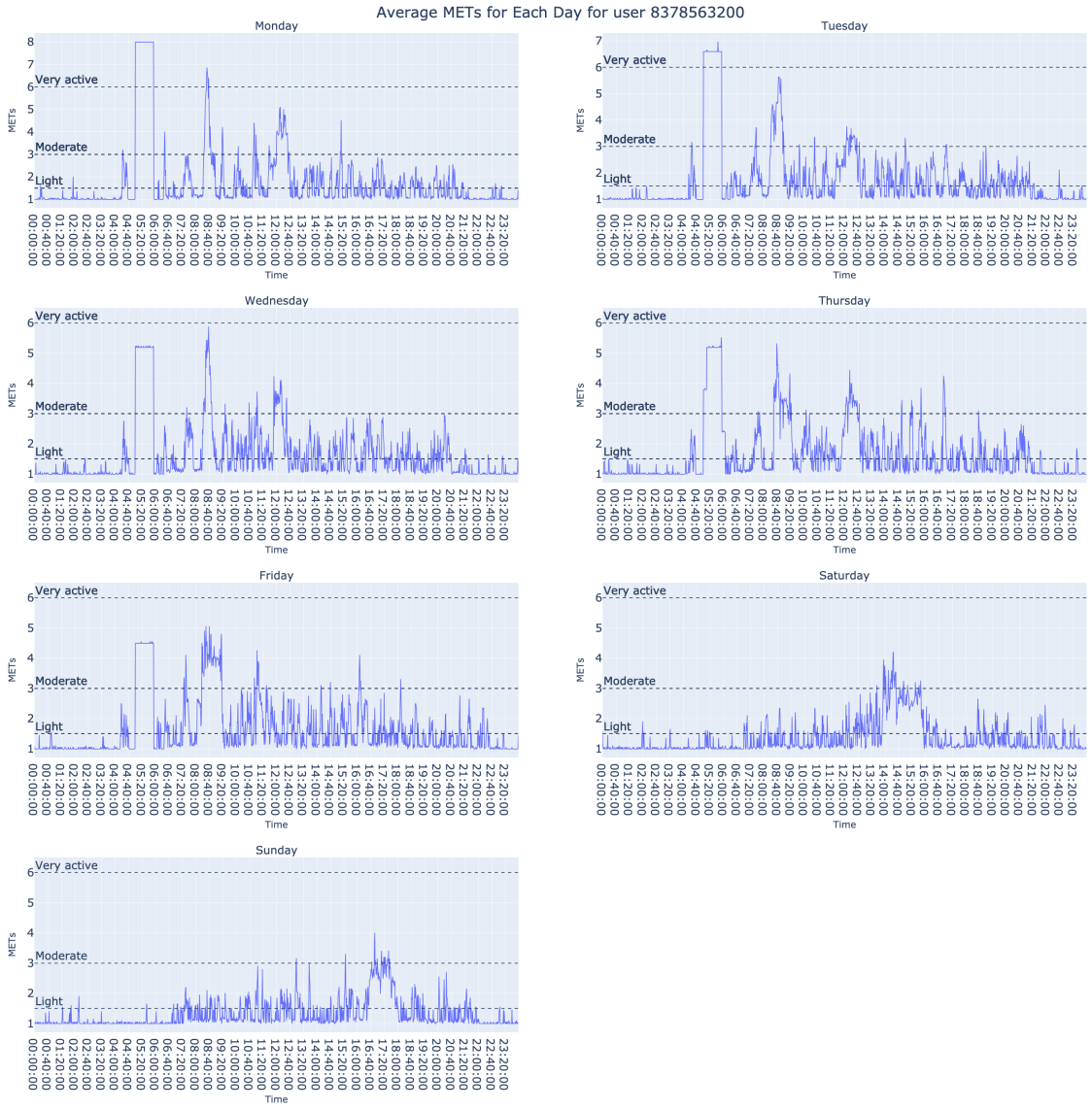


Figure 17: Average daily MET value for user 8378563200

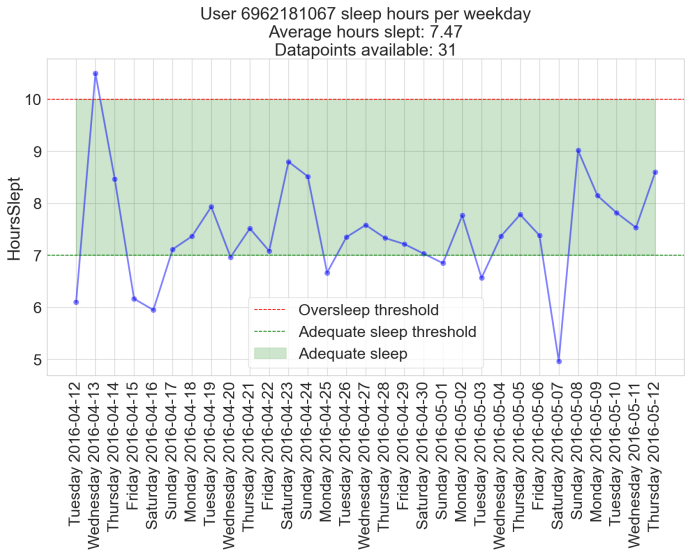
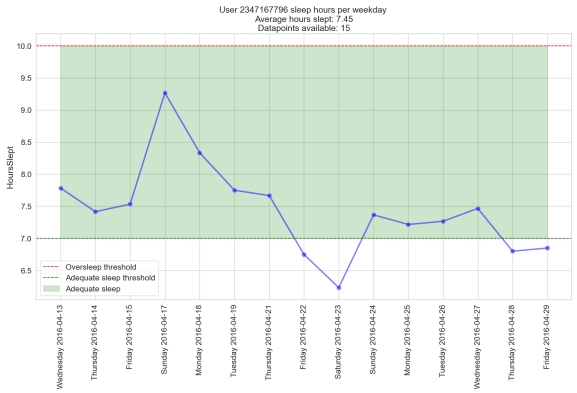
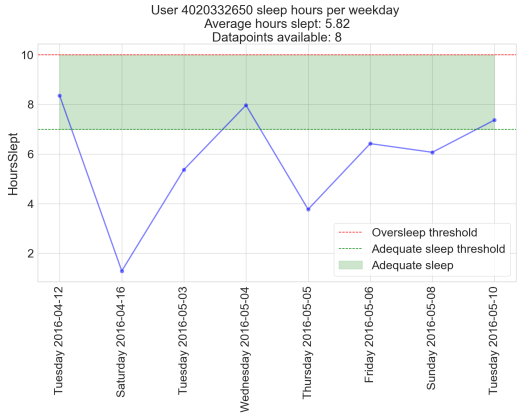


Figure 18: Sleep hours per weekday for user with highest SleepScore



(a) Sleep hours per weekday for user with second-last lowest SleepScore



(b) Sleep hours per weekday for user with the lowest SleepScore

Figure 19: Comparison of sleep patterns for users with low SleepScores

6 Conclusion & Discussion

In conclusion, the preliminary exploratory analysis revealed notable discrepancies in the adherence to high-intensity physical activity guidelines, particularly within the subset of users classified as "Active". **Approximately 70% of these individuals appear to fall short of meeting the World Health Organization's (WHO) prescribed standards for high-intensity physical activity, and an overwhelming 97% deviate from both WHO guidelines.** It is imperative, however, to contextualize this observation within the inherent limitations of the dataset. With a modest cohort size of 33 users and a mere one-month observational window, any definitive comparison with WHO statistics remains elusive.

The dataset's inadequacies notwithstanding, an introspective consideration emerges regarding users' self-perception of their physical activity levels. The discrepancy between subjective perceptions and objective measurements underscores the nuanced nature of physical activity, challenging preconceived notions about individuals' presumed activity levels.

Contrary to expectations and in apparent contradiction to extant literature such as the Kline's work [10], **the analysis failed to establish a significant correlation between sleep parameters (quality and quantity) and individuals' physical activity levels**. This unexpected result prompts scrutiny of the dataset's integrity, given its reliance on a limited subset of valid entries, with only 17 out of 33 users possessing a requisite minimum of 7 records or 31 days.

While a discernible negative correlation surfaced between physical activity and sedentary minutes, this alone does not substantiate the postulated bidirectional link between sleep and physical activity. The intricacies of these interrelationships necessitate more nuanced investigations and underscore the need for larger, more comprehensive datasets.

In further stratifying user activity, the utilization of the MET indicator dataset yielded insightful patterns, notably exemplified by **consistent morning exercise** routine and **probable sustainable commuting practices during weekdays** of user 8378563200. However, the scope of this analysis is inevitably constrained by the **limitations** inherent in the dataset, including a dearth of demographic information and a lack of granularity in the temporal dimension.

In light of these considerations, it is imperative to acknowledge the dataset's **insufficiency**, encompassing a mere 900 entries, and the considerable loss of data (almost 40%) incurred during the amalgamation of physical activity and sleep datasets. Future research endeavors should aim to access more extensive datasets featuring non-anonymized user attributes, enabling robust pattern recognition and meaning-

ful cluster analyses. This strategic approach, incorporating variables such as age, gender, geographical location, and physiological parameters, would undoubtedly enhance the depth and relevance of subsequent investigations into the intricate interplay between sleep, physical activity, and individual characteristics. Ultimately, moving forward, there is potential for the incorporation of additional details concerning patients' sleep-related activities, including the specific timing of their sleep onset and the frequency of nocturnal awakenings.

7 References

[1] Furberg, R., Brinton, J., Keating, M., & Ortiz, A. (2016). Crowd-sourced Fitbit datasets 03.12.2016-05.12.2016 [Data set]. Zenodo. https://doi.org/10.5281/zenodo.53894

[2] Ringeval, M., Wagner, G., Denford, J., Paré, G., & Kitsiou, S. (2020). Fitbit-based interventions for healthy lifestyle outcomes: systematic review and meta-analysis. Journal of medical Internet research, 22(10), e23954.

[3] Bull, F. C., Al-Ansari, S. S., Biddle, S., Borodulin, K., Buman, M. P., Cardon, G., Carty, C., Chaput, J. P., Chastin, S., Chou, R., Dempsey, P. C., DiPietro, L., Ekelund, U., Firth, J., Friedenreich, C. M., Garcia, L., Gichu, M., Jago, R., Katzmarzyk, P. T., Lambert, E., Willumsen, J. F. (2020). World Health Organization 2020 guidelines on physical activity and sedentary behaviour. British journal of sports medicine, 54(24), 1451–1462. https://doi.org/10.1136/bjsports-2020-102955

[4] World Health Organization. Global recommendations on physical activity for health. Geneva: World Health Organization, 2010.

[5] Kilic, O., Saylam, B., & Durmaz Incel, O. (2023, March). Sleep Quality Prediction from Wearables using Convolution Neural Networks and Ensemble Learning. In Proceedings of the 2023 8th International Conference on Machine Learning Technologies (pp. 116-120).

[6] NetHealth Fitbit Dataset, https://sites.nd.edu/nethealth/

[7] Jetté, M., Sidney, K., & Blümchen, G. (1990). Metabolic equivalents (METS) in exercise testing, exercise prescription, and evaluation of functional capacity. Clinical cardiology, 13(8), 555–565. https://doi.org/10.1002/clc.4960130809

[8] Semanik, P., Lee, J., Pellegrini, C. A., Song, J., Dunlop, D. D., & Chang, R. W. (2020). Comparison of Physical Activity Measures Derived From the Fitbit Flex and the ActiGraph GT3X+ in an Employee Population With Chronic Knee Symptoms. ACR open rheumatology, 2(1), 48–52. https://doi.org/10.1002/acr2.11099

[9] Reed, D. L., & Sacco, W. P. (2016). Measuring Sleep Efficiency: What Should the Denominator Be?. Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine, 12(2), 263–266. https://doi.org/10.5664/jcsm.5498

[10] Kline C. E. (2014). The bidirectional relationship between exercise and sleep: Implications for exercise adherence and sleep improvement. American journal of lifestyle medicine, 8(6), 375–379. https://doi.org/10.1177/1559827614544437

[11] https://www.cdc.gov/sleep/about_sleep/how_much_sleep.html

[12] Léger, D., Beck, F., Richard, J. B., Sauvet, F., & Faraut, B. (2014). The risks of sleeping "too much". Survey of a National Representative Sample of 24671 adults (INPES health barometer). PloS one, 9(9), e106950. https://doi.org/10.1371/journal.pone.0106950