CS-C4100 Digital Health and Human Behavior

Course Project: FitBit Data Analysis

1 Introduction

It is more and more common to use some kind of smart device for self-monitoring in everyday life, according to a survey conducted by Sitra [1]. The survey revealed that 51% of the respondents use smart devices for self-measurement, mostly mobile phone applications or smart watches. These kinds of devices can for example monitor physical activity, sleep, and other health-related behaviors. Although self-tracking is becoming very popular, physical inactivity is still one of the leading causes of ill-health in industrialized countries [2]. However, it seems that utilization of wearable activity trackers might increase physical activity participation in the long term and one's interest in one's health [2].

In this report I will introduce a project where I explore data collected by FitBit activity-tracking wristwatches to find some interesting correlations for example between sleep, and activity level, and also to identify different patterns in the participants' lifestyles. I also want to explore, how background information of the participants, such as body mass and body mass index, affect their health behavior. Moreover, the aim of this project is to make conclusions both on individual and community level lifestyles and habits and compare the results to the recommended values estimated by scientific papers.

In this report, I will be making conclusions of the participant's current activity state based on the recommended daily step count of at least 7000 [3]. According to the study, participants taking at least 7000 steps daily had 50-70% lower risk of mortality among middle-aged adults [3]. Moreover, I will compare the sleep data to the recommended 7 to 9 hours of sleep for adults [4]. Although the appropriate sleep duration may vary between individuals, sleep

durations far outside from the recommended range are rare, and not habitually not getting enough sleep can cause serious health-problems [4].

When discussing consumer-based wearable trackers, it is also worthy to take reliability and validity into consideration. A study examined the validity properties of FitBit trackers and it was found out that step measurement data has higher validity and in contrast, sleep and energy expenditure measurements data have lower validity [5]. Generally, FitBit has good measurement reliability between different devices [5]. Therefore, we can assume that the data used for this project is generally reliable but not extremely precise. I am also going to use features with higher validity more, such as step count, in this data analysis.

The report begins with problem formulation, where I will explain more in detail what kind of questions I want to answer with this data analysis. After that, I will describe the data used in this project and introduce data preprocessing steps in section 3. Then, I will introduce all the data processing and analysis methods used in this project and give some justifications for using them. In the fifth section, I will introduce and explain the results of the data analysis part and finally, give some conclusions and discussion in the last part of the report.

The data analysis indicates that on average, the participants are generally active, but too many of them are severely inactive, which might contribute to the fact that a lot of the participants are not getting sufficient amount of sleep and many of them are overweight. Moreover, it seems that physical inactivity negatively affects sleep duration. However, too high intensity exercise might also negatively affect sleep. As there was very little data on the background information of the participants, such as BMI, it is hard to draw any conclusions based on the participants lifestyles, but overall, it seems that many of them would benefit from more physical activity.

2 Problem formulation

In this project, my goal is to find interesting correlations between participant's activity state, sleep quality, and overall health state. I am also aiming to find out interesting lifestyle patterns of modern humans, as physical inactivity seems to be one of the leading causes of health problems in industrialized countries. Firstly, I want to investigate the participant's current health state based on their daily step count because it is the measurement with highest accuracy. I want to explore both individual level and community level activity state.

One of the most important factors contributing to one's health is sleep. Sleep quality is associated with sleep duration, and therefore I want to investigate the participant's sleep durations and how physical activity levels affect it. Lastly, I want to explore if participant's background information has any effect on their health-related behavior. The only background information provided were weight, body mass index and fat percentage. In addition, I would have wanted to explore the relationship between participant's background and sleep quality, but there was too little data to do that, unfortunately.

3 Dataset description

3.1 Activity dataset

The data was collected with FitBit activity-tracking wristwatches from 30 participants for 31 consecutive days, during 2016. The data was collected by a survey generated by Amazon Mechanical Turk. All data is obtained from Kaggle [6]. The dataset consists of information about user id, date, total steps, total distance, tracker distance, logged activities distance, distances for different activity levels, minutes spent in different activity levels, and the amount of burned calories. This dataset will be called "activity" and there are in total of 940 data points and no missing data.

However, after exploring the dataset I noticed that some of the features seem to add no value in the context if this project, so I decided to drop them. For example, it seems that TotalDistance and TrackerDistance are almost identical and perfectly correlated, so I decided to drop TrackerDistance. Additionally, I noticed that 908 of the 940 datapoints in the column LoggedActivitiesDistance are equal to 0 so I decided to drop that column, too. Therefore, the final features in the activity dataset are Id, ActivityDate, TotalSteps, TotalDistance, VeryActiveDistance, ModeratelyActiveDistance, LightActiveDistance, SedentaryActiveDistance, VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes, SedentaryMinutes, and Calories. Additionally, ActiveMinutes was created based on the sum of VeryActiveMinutes and FairlyActiveMinutes.

The data types are mostly in float64 and int64 format, but the ActivityDay column's data type is an object, so I decided to change it to datetime64 format and renamed it to Date to make data processing easier. Also, it seems that there are 33 different user ids in the activity dataset.

3.2 Sleep dataset

Additionally, I decided to investigate some sleep data collected from the participants during this experiment. This dataset consists of information about user id, date, total sleep records, total minutes asleep, and total time spent in bed. There are in total of 413 datapoints, with

zero null values. However, there are only 24 unique user ids, so not all participants logged their sleep data. I decided to drop total sleep records column as it does not provide any useful information in the scope of this project. Therefore, the final chosen features are Id, Date, TotalMinutesAsleep, TotalTimeInBed, and TotalHoursAsleep, which was created based on TotalMinutesAsleep.

Like previously with the activity dataset, date data was converted into datetime-format, and the name was changed from SleepDate to Date to make data processing and merging easier.

3.3 Weight dataset

Moreover, I decided to use a dataset containing some background information about the participants, because it might provide some interesting insights. This data set contains features such as user id, date, weight in kilograms and pounds, fat percentage, body mass index (BMI), log id, and whether the data was logged manually or not. I decided to drop most of the unnecessary features, such as WeightPounds, IsManualReport, and LogId. Also, the column containing information on fat mass was dropped because 65 of its 67 values are null. Therefore, the final features chosen are Id, Date, WeightKg, and BMI. The date data was also converted into datetime-format.

Unfortunately, I noticed that there are only 8 unique ids in the weight dataset and 95% of the 67 datapoints are logged by only two user ids. Thus, the data is very small and biased, and this will make it very hard to draw any conclusion from the data.

4 Methods

4.1 Overall physical activity level

To investigate the overall physical activity of the participants, I used daily step count data from the activity dataset. First, the data points with 0 daily steps were excluded because it is very unlikely that any of the participants would not have taken any steps a day, and it is more likely that the participant just has not used the tracker that day. After that, the data was grouped by participant id, so that the average daily step count of every id is calculated. Finally, the data was plotted to a bar plot per user id, to better visualize the data. This way it is possible to visualize both individual and community level physical activity at the same time.

4.2 Overall sleep duration

To draw conclusions on the participant's overall sleep duration levels, I used TotalMinutesAsleep data from the sleep data set. The data was again grouped by the user id, so that the average sleep duration was calculated for each id. Based on that data, a histogram count plot was created with kernel density estimation line, to get better insights on the overall sleep of the participants. The count plot was divided into 12 bins.

4.3 Physical activity level's effect on sleep

To explore the relationship between physical activity level and sleep, the activity dataset and sleep dataset were merged so that the user ids and date information are matched with each datapoint. As a result, we now have a dataset with 413 datapoints containing information of the participant's daily activity and sleep.

Next, I decided to label each datapoint based on the ActiveMinutes (sum of VeryActiveMinutes and FairlyActiveMinutes) column so that if the participant's daily ActiveMinutes exceeds 60 minutes, this datapoint is labeled as "Active", and vice versa as "Inactive" if not. After that, a similar histogram count plot of sleep durations is created, but the data is divided into "Active" and "Inactive", to better compare the two activity groups. The limitations of this plotting method are that it emphasizes the individual level too much,

as the data is not grouped by user id. Therefore, if an individual with certain lifestyle pattern has been more active with logging their information, it distorts the data.

Additionally, a scatterplot describing the relationship between SedentaryMinutes and TotalHoursAsleep was created, to get another insight on physical activity's effect on sleep.

4.4 Body mass index's effect on physical activity

Lastly, to investigate the relationship between BMI and physical activity, the activity dataset and weight dataset were merged so that the user ids and date information match. As a result, a dataset with 67 datapoints was created. Moreover, the datapoints were labeled based on the BMI so that if the BMI exceeds 25, the datapoint is labeled as "Overweight", and if the BMI is under 18 it is labeled as "Underweight". All the datapoints falling between these two categories are labeled as "Normal".

Next, the data was grouped by the labels, so that the average step count of each label was calculated. Finally, a bar plot comparing the average step count of each BMI label was created to get insights on the health behavior of the participants. The limitations of this plotting style, however, are that it totally ignores the individual level findings.

5 Results

5.1 Overall physical activity level

The overall average step count among all participants is 7922 which exceeds the recommended 7000 daily step count. This indicates that the overall activity of all participants is in a sufficient level. However, 39% of the participants are below that recommended level on average, as can be seen in Figure 1. In Figure 1, we can also see that some of the individuals are very active and on the contrary, others are alarmingly inactive.

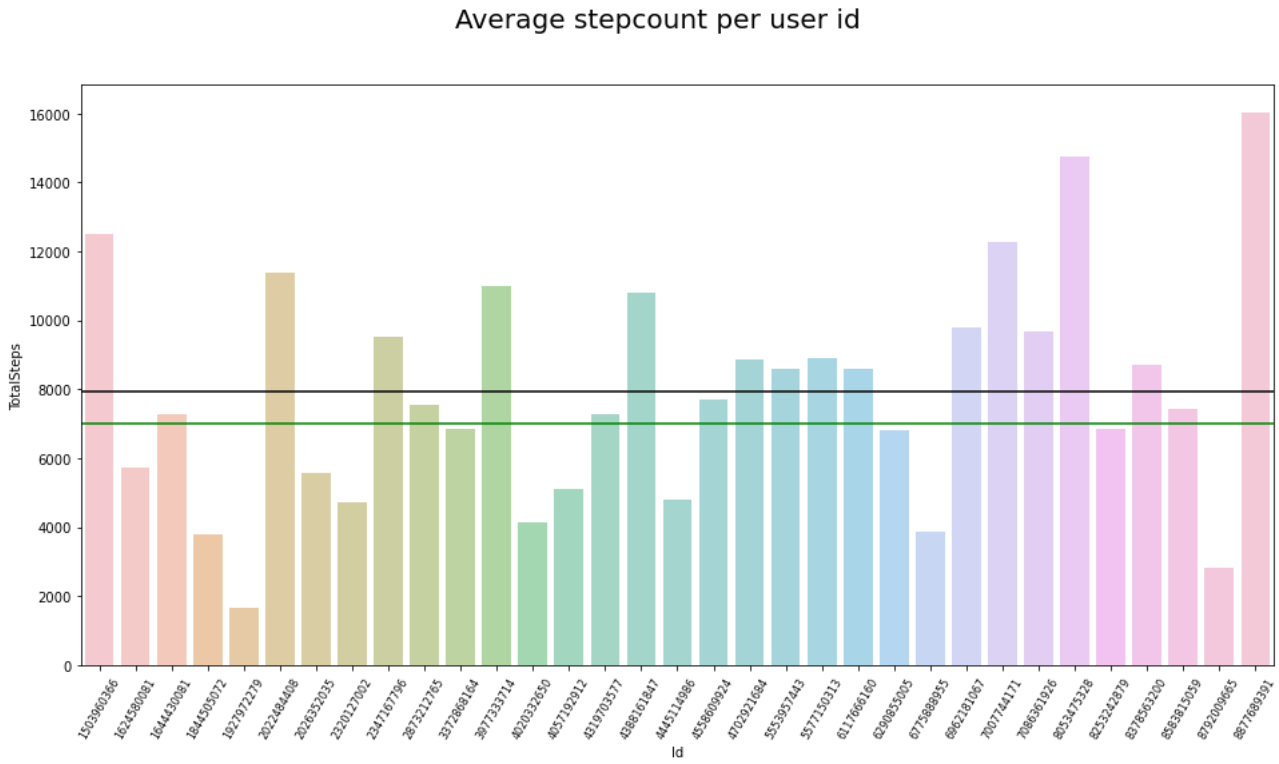

Figure 1: Daily average stepcount per user id. Black horizontal line describes the average stepcount among all participants and green line describes the recommended stepcount.

5.2 Overall sleep duration

The average sleep duration among all participants was 6.99 hours, which is barely sufficient amount. But as can be seen from Figure 2, there are a few logs of only one to two hours of sleep which might distort the data and thus the average sleep duration. However, we do not know whether these logs are just errors in the measurement or real results. Figure 2 shows

that the data is somewhat normally distributed, and most participants get approximately 7 hours of sleep a night on average, which is barely sufficient.

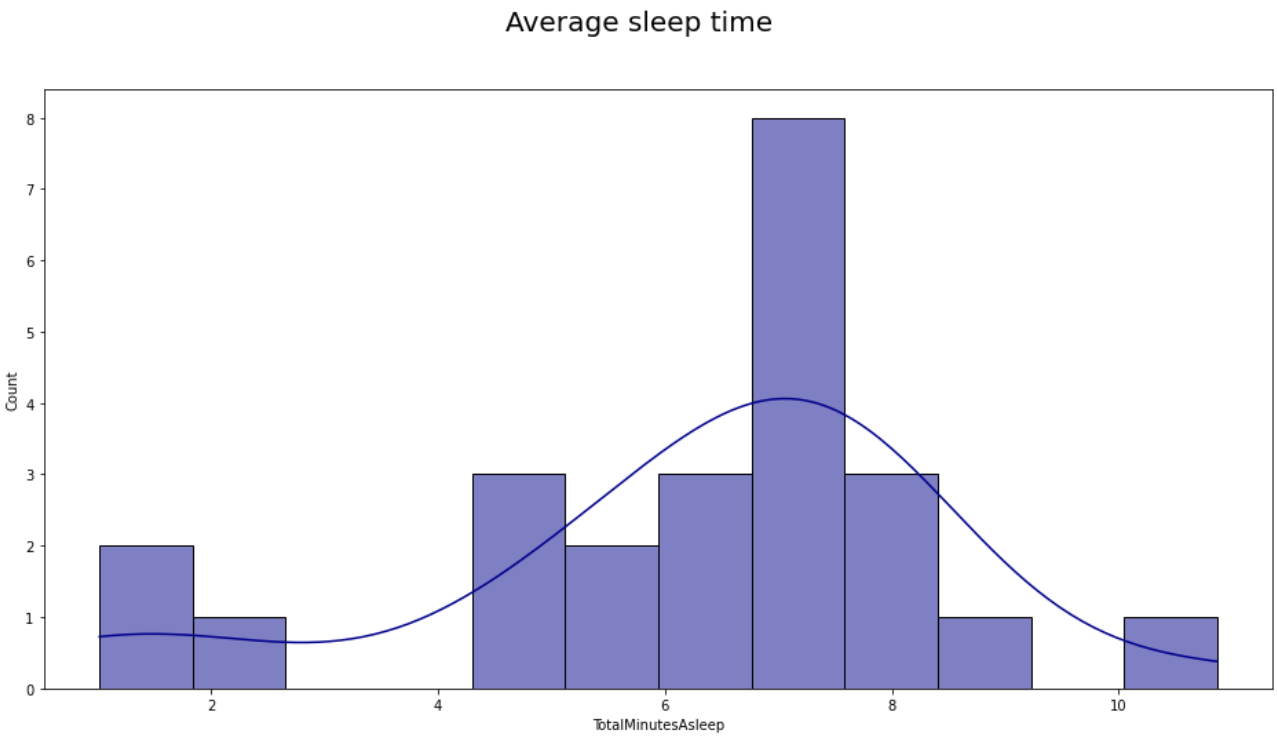

Figure 2: Average sleeping hours among all participants. Blue line describes kernel density estimation.

5.3 Physical activity level's effect on sleep

According to Figure 3, it seems that participants belonging to the "Inactive" category get even slightly more sleep on average than the participants belonging to the "Active" category. This is actually the opposite of what I would have initially expected. However, the data visualized in this picture is strongly influenced by individuals who have been more active with using the FitBit tracker. It is also possible that the more active participants are exercising with too high intensity, which might negatively influence sleep.

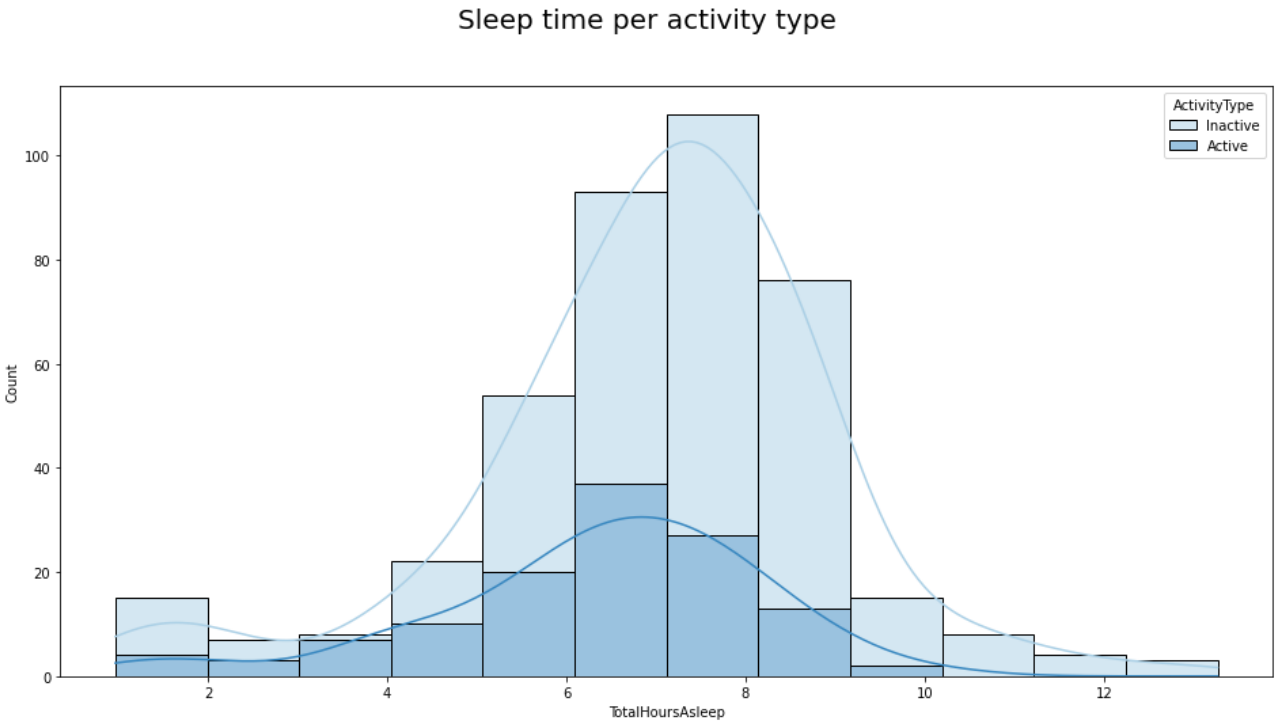

Figure 3: Daily sleep time grouped by daily activity level.

In Figure 4, however, we can see the relationship between the number of sedentary minutes and sleep duration among all datapoints. These two features seem to have negative linear correlation between them. Although, the correlation does not seem to be very strong. To conclude, it seems that light and sufficient physical activity positively influences sleep.

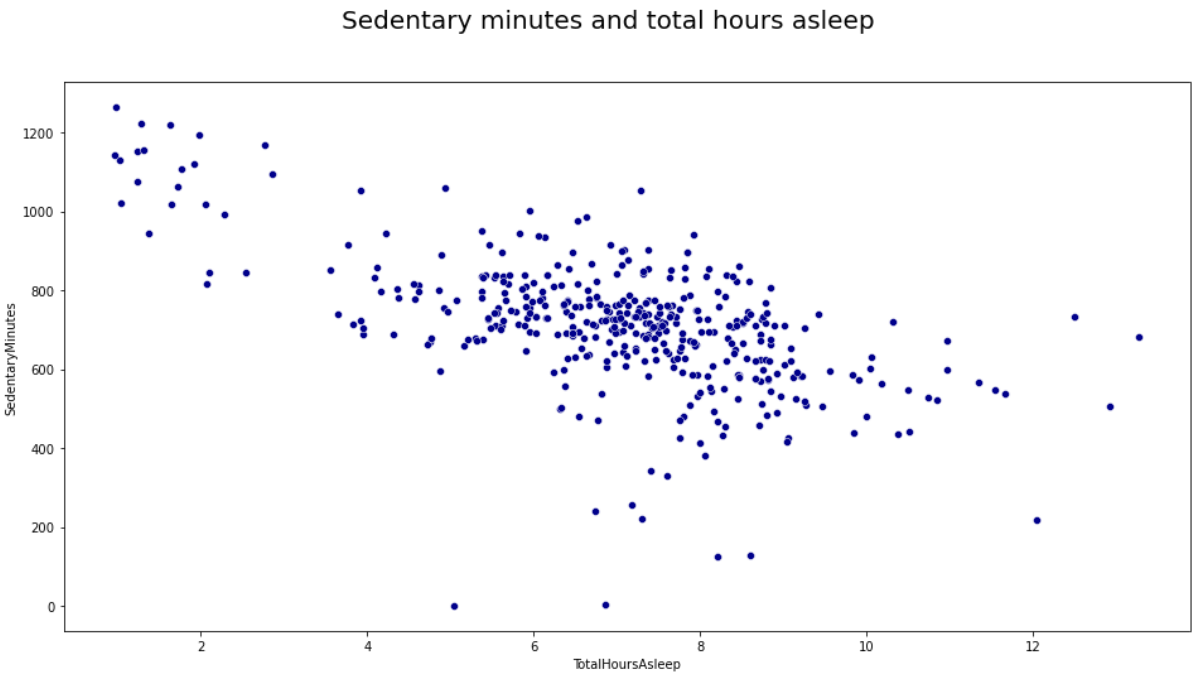

Figure 4: Sedentary minutes affecting total hours asleep.

5.4 Body mass index's effect on physical activity

In Figure 5, it seems that participants belonging to the "Normal" BMI category have higher daily step count on average. Also, it seems that there are no participants belonging to the "Underweight" category. This plot is not very informative on the individual level but more on the communal level. To conclude, it seems that participants with normal BMI have healthier lifestyles and that might have a positive effect on their health. However, the cause-and-effect relationship is not clear. Also, there was very little data on participant's BMI values, so it is hard to draw any conclusions based on this data.

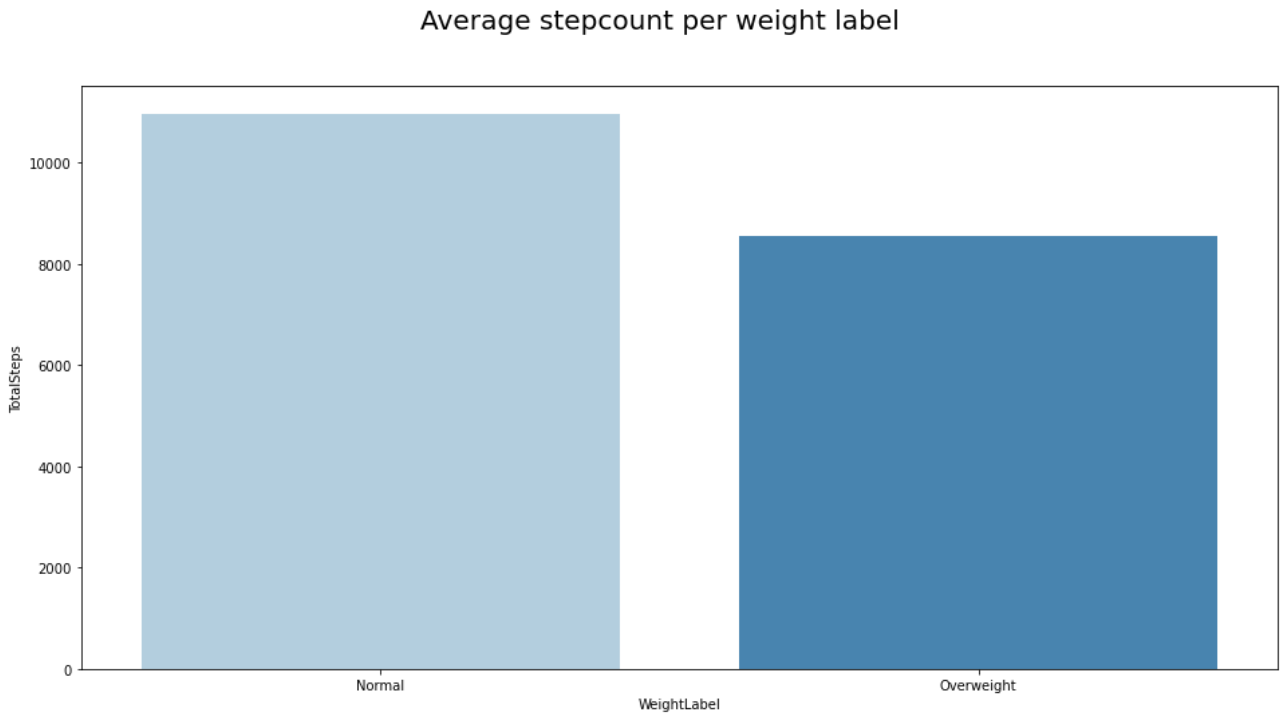

Figure 5: Daily average step count grouped by weight label.

6 Conclusion and Discussion

In conclusion, it seems that the participants are on a good physical activity level on average, but too many participants are alarmingly inactive. Physical inactivity might contribute to the fact that a lot of the participants are overweight and are barely getting enough sleep. Sleep is essential for both mental and physical health, as well as for cognitive and physical performance.

Based on this data analysis, physical inactivity does negatively affect sleep duration. Although, when dividing participants into "Active" and "Inactive" based on their daily high intensity activity time, it seems that participants in the "Active" category are getting less sleep. This indicates that too high intensity exercise might also negatively affect sleep duration.

Moreover, it seems that participants with higher BMI are taking fewer steps a day, which might negatively affect their health. As discussed before, this might also contribute to their sleep duration, which again negatively affects their health. This analysis highlights the fact that modern humans are suffering from physical inactivity and this kind of lifestyle can cause severe health issues, such as obesity and heart diseases. On a good note, physical activity is an easy habit to start because even a light walk goes a long way.

However, this data analysis has its limitations, as there was very little data on the backgrounds of the participants and thus, it is hard to draw any conclusion on the participant's lifestyles based on this data only. Therefore, for future perspectives, this analysis would highly benefit from having a lot more data from a longer period of time, for example.

References

[1] "Well-being data measurement offers opportunities for creating new innovative services," *Sitra*. https://www.sitra.fi/en/articles/well-being-data-measurement-offers-opportunities-for-creating-new-innovative-services/ (accessed Nov. 15, 2022).

[2] K.-J. Brickwood, G. Watson, J. O'Brien, and A. D. Williams, "Consumer-Based Wearable Activity Trackers Increase Physical Activity Participation: Systematic Review and Meta-Analysis," *JMIR mHealth and uHealth*, vol. 7, no. 4, p. e11819, Apr. 2019, doi: 10.2196/11819.

[3] A. E. Paluch *et al.*, "Steps per Day and All-Cause Mortality in Middle-aged Adults in the Coronary Artery Risk Development in Young Adults Study," *JAMA Network Open*, vol. 4, no. 9, p. e2124516, Sep. 2021, doi: 10.1001/jamanetworkopen.2021.24516.

[4] "National Sleep Foundation's sleep time duration recommendations: methodology and results summary - PubMed." https://pubmed.ncbi.nlm.nih.gov/29073412/ (accessed Dec. 19, 2022).

[5] K. R. Evenson, M. M. Goto, and R. D. Furberg, "Systematic review of the validity and reliability of consumer-wearable activity trackers," *International Journal of Behavioral Nutrition and Physical Activity*, vol. 12, no. 1, p. 159, Dec. 2015, doi: 10.1186/s12966-015-0314-1.

[6] "FitBit Fitness Tracker Data." https://www.kaggle.com/datasets/arashnic/fitbit (accessed Dec. 13, 2022).