Project Big Data
An investigation on FitBit data

Group 9:
Mia Jacobsen (2755163)
Le Duy Manh (2694215)
Luu Thu Trang (2695303)

July 3, 2022

Contents

1 Introduction

The dataset available for analysis in this report is the Fitbit data. The data was gathered through the Amazon Mechanical Turk, and it consists of data from 30 Fitbit users collected over two months in 2016. This report will start with a brief data exploration section and some statistics, to answer a series of interesting questions pertaining to the activity and sleep patterns of the Fitbit users. Next, modelling will be applied to yield valuable insights, which will be used to answer 2 assumptions pertaining to whether increased activity levels have a beneficial effect on sleep. Finally, the report ends with a discussion of the data itself and suggestions for future research.

2 Data Exploration and Statistics

2.1 Dataset selected:

Within the package data from Fitbit, there were several interesting dataset, however some has to be eliminated due to the low quality. More specifically, the datasets selected for the analyses in this report are dailyActivity, dailyCalories, dailyIntensities, dailySteps, hourlySteps and sleepDay. The remaining datasets are irrelevant to answering the proposed questions, consisting of very few data points and/or contains inconsistencies in the data.

2.2 Proposed questions

Upon inspecting the obtained datasets, some interesting research question was proposed, and they are:
1. At which hour during the day are people generally most active?
2. At which day of the week are people generally most active?
3. Does having more steps lead to a higher calorie burn?
4. Does longer active duration increase calorie burn?
5. Does longer exercise duration increase sleep duration?
6. Does longer exercise duration decrease time spent on bed without sleeping?
All these questions will be investigated with the help of statistics in this section of the report, and in addition to that question 5 and question 6 will be further investigated with the help of modelling in latter parts of this report.

2.3 At which hours of the day are people generally most active?

Before answering this question, some other interesting statistics can be observed. Firstly, it is interesting to look at people's activity during the most active hour. The most active hour based on the total steps taken is at the 18th hour of the day, or informally at 6pm. The activity of all users during this hour is shown in figure 1.

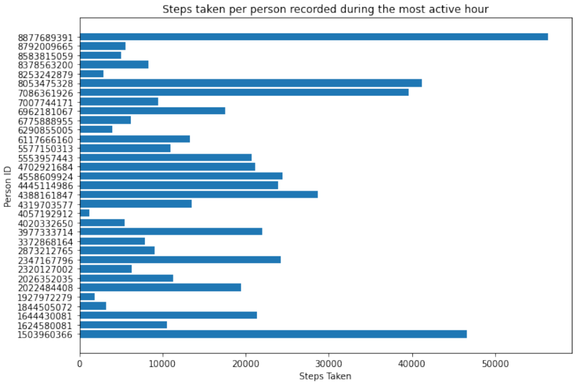

Figure 1: Most active hour's activity per user

It can be extracted from figure 1 that there are 4 users who are very active compared to the remaining users, however it can be said that most users were as expected active during this hour. Contrastingly, it can also be interesting to look at the activity of users during the least active hour, which is the 3rd hour of the day, or informally at 3 am. Figure 2 shows this.

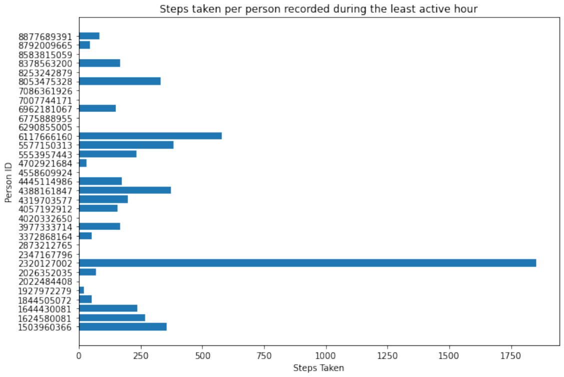

Figure 2: Least active hour's activity per user

As expected, majority of users are incredibly inactive during this hour, besides 1 user, however the steps taken by this user at this hour is still significantly lower compared to the average steps taken during the most active hour, which is just 181.69 compared to 16449.94. Another interesting observation is to look at the activity of the most active and the least active person. Figure 3 and figure 4 shows this data.

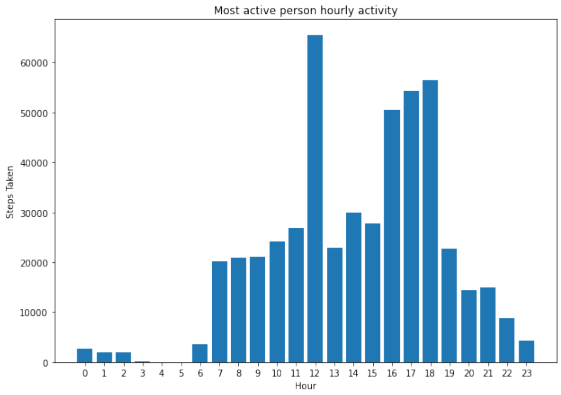

Figure 3: Most active user's hourly activity

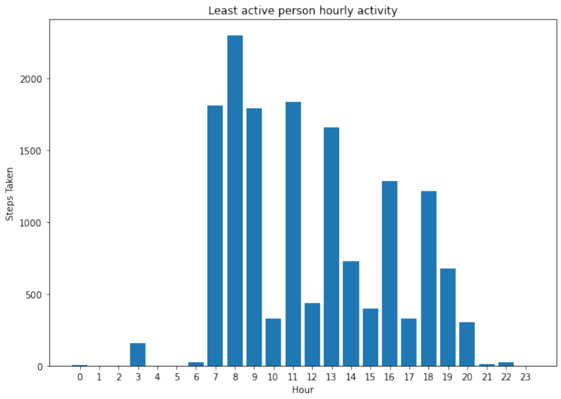

Figure 4: Least active user's hourly activity

The maximum step taken by the most active person, with ID numbered 8877689391, can be seen at the 12th hour, and more specifically it is 65340 steps. This is by far higher than the maximum steps taken by the least active person, with ID numbered 4057192912, at the 9th hour, of 2295. In fact, the lowest number of steps taken by the most active person, with the condition that the most active person has at least taken 100 steps in order to only compare when they are active, is 1887 steps, which is only slightly lower than the least active person maximum steps. Finally, to answer the main question of this section, figure 5 shows the graph of the hourly activity based on the steps taken.

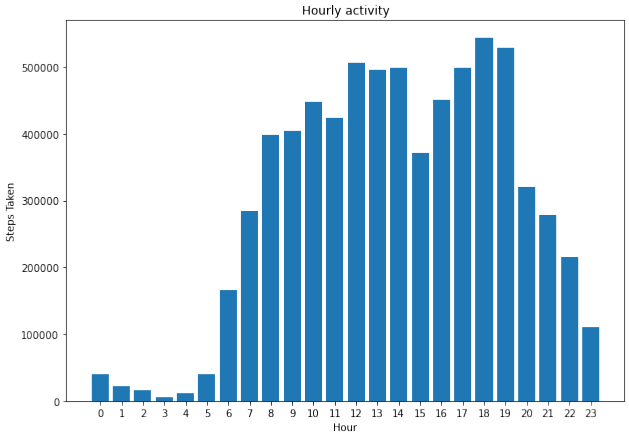

Figure 5: Hourly activity

From this graph, it can be said that people are generally active from the 8th hour to the 20th hour of the day.

2.4 At which day of the week are people generally most active?

Figure 6 shows the data of the steps taken per weekday.

The above figure shows that there is not a single day where people are most active, as it can be observed that the activeness spread quite evenly throughout the week. The most active day is Tuesday, while the

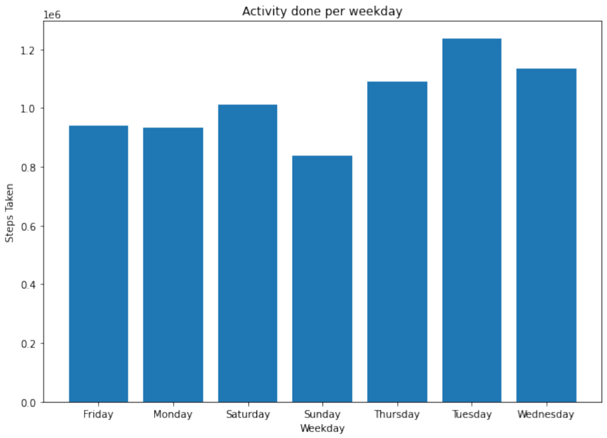

Figure 6: Weekday activity

least active day is Sunday, however the differences are minimal.

2.5 Does having more steps lead to a higher calorie burn?

Figure 7 shows the scatter plot of calories burned against the steps taken

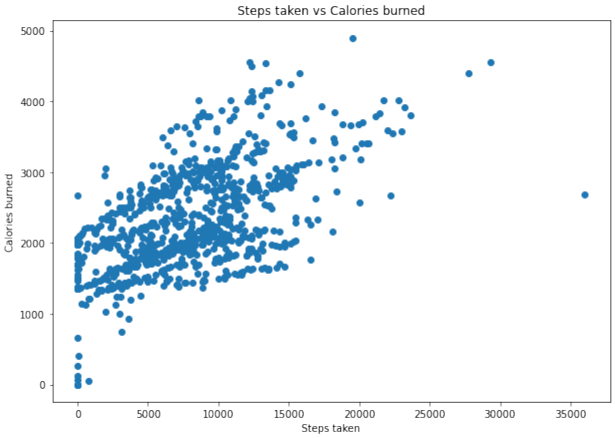

Figure 7: Scatter plot of the calories against steps taken

The scatter plot shows a positive linear relationship between steps taken and calories burned, which confirms the theory that more steps does indeed lead to higher calories burn. Further proof comes from the Pearson and Kendall correlation coefficients, which are 0.591 and 0.3974 respectively. Both coefficients are positive and quite large, within the range of -1 to 1, therefore it can be concluded that doing more steps does indeed lead to higher calorie burns. In addition, the p-value of the Pearson correlation test is 8.2e-90 and Kendall correlation test is 7.52e-74, and at a common test level of 0.05 it shows that the abovementioned correlation coefficient are statistically significant.

2.6 Does longer active duration increase calorie burn?

Figure 8 shows the scatter plot of the total active minutes against calories burned.

It can be seen from the figure that there is a linear relationship, however it is only visible from 0 to approximately 800 calories. From 800 calories onwards, there does not seem to be a systematic pattern

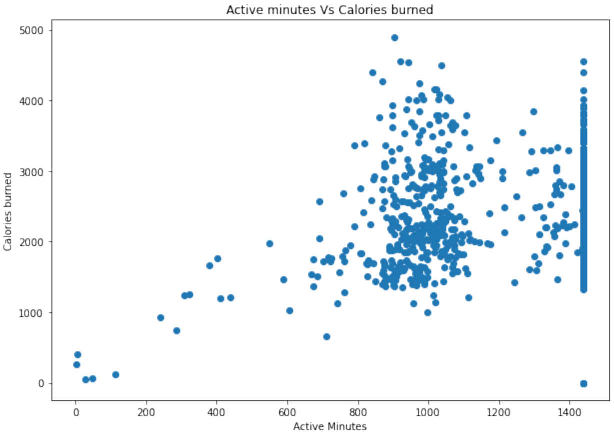

Figure 8: Scatter plot of the total active minutes against calories burned

to indicate whether higher active minutes result in higher calories burn. To further check this statement, Pearson and Kendall correlation coefficient was employed. The Pearson correlation coefficient came out to be 0.0945 and Kendall correlation coefficient is 0.0252. Both coefficients are positive, which does indicate a positive linear relationship, however the values are quite small, indicating that there are potentially other factors that stronger influences calories burn. This explains the linear relationship from 0 to approximately 800 calories, while beyond that there are no systematic pattern. The resulting p-value from the above Pearson and Kendall correlation tests came out to be 0.0035 and 0.285, meaning at test level 0.05 only the Pearson correlation coefficient is statistically significant. Since both coefficients are positive and similarly small, the conclusion does not change. All in all, it can be concluded that from 0 to approximately 800 calories, total active minutes does influence calories burn, however beyond this point other factors are to be considered as well.

2.7 Does longer exercise duration increase sleep duration?

Figure 9 shows the scatter plot of the total active minutes against sleeping time.

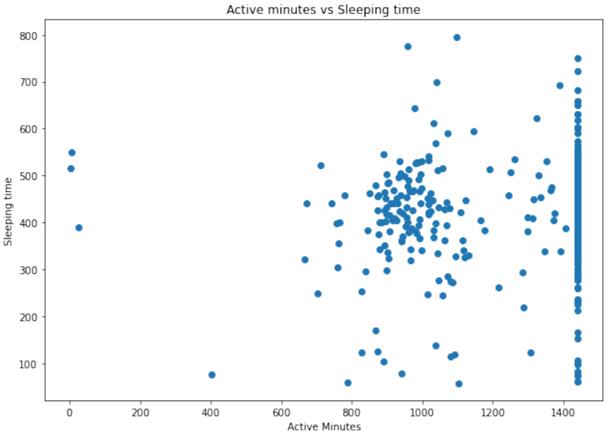

Figure 9: Scatter plot of the total active minutes against sleeping time

Figure 9 shows that there is not any visible systematic pattern. Therefore, with just the figure alone, there will be inconclusive results. To yield an appropriate conclusion, once again the use of Pearson and Kendall correlation coefficients will be utilized. The Pearson correlation and Kendall correlation are 0.08959 and

0.0768 respectively. Both coefficients are positive which indicates a positive linear relationship, although the values are quite small. The p-values are 0.0689 for Pearson correlation test and 0.03477 for Kendall correlation test, which means only Kendall correlation coefficient is statistically significant at test level 0.05. Since both coefficients are positive and small, the conclusion remains the same. All in all, it can be concluded that longer exercise duration does somewhat increase sleep duration, possibly also with the addition of other factors.

2.8 Does longer exercise duration decrease time spent on bed without sleeping?

Figure 10 shows the scatter plot of the total active minutes against the time spent on bed without sleeping.

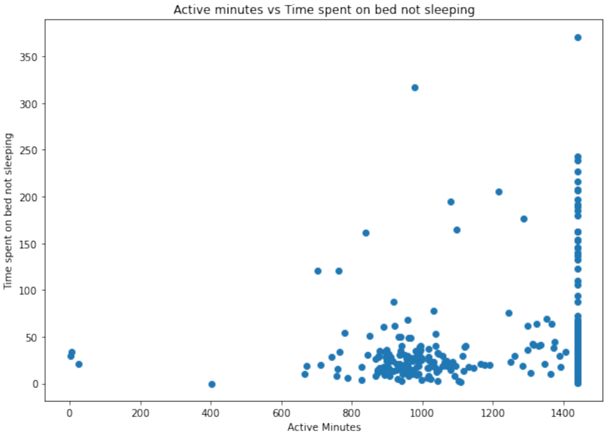

Figure 10: Scatter plot of the total active minutes against the time spent on bed without sleeping

Once again, the scatter plot seems to yield inconclusive results. Therefore, Pearson and Kendall correlation will be used. Pearson correlation and Kendall correlation are 0.123 and 0.118, both are positive. This indicates a positive linear relationship, which means that longer exercises duration does not decrease time spent on bed not sleeping. The p-value of Pearson and Kendall correlation tests are 0.012 and 0.0012, therefore both coefficients are statistically significant. This is a very interesting finding, as it is expected that more exercise will enable people to go to sleep faster. It is possible that this is happening due to the small set of data, however the result is interesting and should be investigated further in future research.

3 Regression Models

3.1 Active duration and sleep duration

In this section, we look into how being active would affect participants' sleep duration. This is achieved by investigating the 2 files daily activity and sleep day. As indicators for how active a participant is on a certain day, it is plausible to use the number of steps, active minutes, and calories burned. In particular, we will answer the question: Does more active people have longer time asleep?

3.1.1 Pre-processing

First, as we are not interested in the distance features, they are dropped from the daily activity data. The daily activity data and the daily sleep data were merged into one dataframe of 413 rows and 12 columns and consists of 24 unique participants. It is good to note that there are some days when one only sleep around a hour. However, to keep the variation of the data (every day of each individual is independent), and given that having fewer data points would reduce the accuracy of the model, we decided to keep the current merged dataframe.

3.1.2 Choosing predictors and target variables

Within the current merged data, we have many variables: `TotalSteps, VeryActiveMinutes, FairlyActiveMinutes,LightlyActiveMinutes, SedentaryMinutes, Calories, TotalSleepRecords, TotalMinutesAsleep, TotalTimeInBed`. To build the regression model, we will examine which feature would be relevant and significant predictor. One solution for this is to look at the pair graph of all existing feature. From Figure 11, clearly, there is almost a perfect linear relation between `TotalMinutesAsleep` and `TotalTimeInBed`. Therefore, we use `TotalMinutesAsleep` as our target variable for sleep duration. It seems that `VeryActiveMinutes` and `FairlyActiveMinutes` have the same effect on the target variable. To further analyse this, we will look at the correlation coefficient between each variable and the t-test result so see if they are significant to the model.

Var	Correlation
TotalSteps	-0.186866
VeryActiveMinutes	-0.090436
FairlyActiveMinutes	-0.244535
LightlyActiveMinutes	0.032914
SedentaryMinutes	-0.599394
Calories	-0.028526

Table 1: Correlation between TotalMinutesAsleep and features

	t-test p-value
TotalSteps	0.224
VeryActiveMinutes	0.010
FairlyActiveMinutes	0.000
LightlyActiveMinutes	0.002
SedentaryMinutes	0.000
Calories	0.000

Table 2: T-test p-values for predictors

Surprisingly, although `TotalSteps` have quite significant correlation coefficient with sleep duration, its p-value is not significant. Whereas, `LightlyActiveMinutes` and `Calories` are significant regardless of the fact that they are not as strongly correlated to `TotalMinutesAsleep`. The reason could be because of the small dataset that makes t-test unreliable, multicollinearity, lifestyle of participants, or other numerical problems. Contradict to what is normally exptected: `VeryActiveMinutes` should have great influences on `TotalMinutesAsleep`, the two are not so strongly correlated. This is because of the variance between individual and statistically speaking, it is less likely to be active for a long period of time day after day. We decided to choose the following features as our predictors: `TotalSteps, FairlyActiveMinutes, SedentaryMinutes, Calories`.

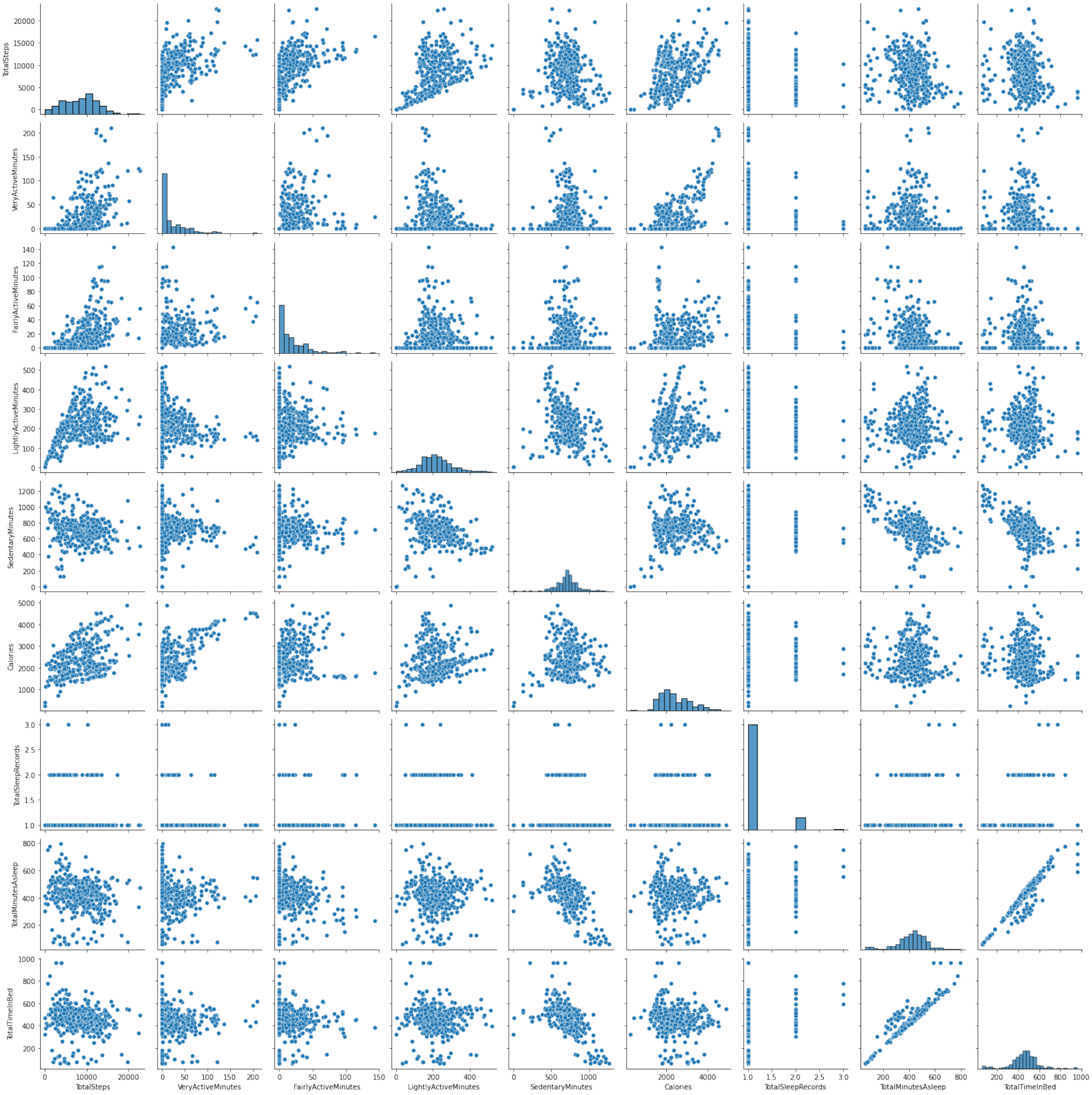

Figure 11: Pair graphs of all features

3.1.3 Model

Finally we build a regression model for the chosen predictors and target variables. To increase the accuracy of the model we adopted PolynomialFeatures with degree 2. Predictors are scaled before fitting with the help of StandardScaler. Then the hyperparameters are tuned using 10-folds cross validation and GridSearch was deployed to find the best parameter. We shall observe the performance of the following regressors in the next section: LinearRegression, SGDRegressor, Lasso, and Ridge.

There are many useful metrics to evaluate regression models, in this research we will only closely look at 2 of them namely coefficient of determination (R^2) and the root mean squared error (RMSE) in minutes. Table 3 shows the two metrics of each regression model evaluated at the best parameters.

	LinearRegression	SGDRegressor	Lasso	Ridge
Parameters	none	$\alpha = 0.2861$ penalty = ’l1’	$\alpha = 0.286$ selection = ’cyclic’ $\max_i ter = 1000000$	$\alpha = 0.4291$ solver = ’sag’ $\max_i ter = 1000000$
R^2	0.717	0.708	0.708	0.711
RMSE	56.390	57.301	57.296	57.001

Table 3: R^2 and RSME for different models

All results are quite similar. However, LinearRegression would not make a robust model as it does not have parameter tuning. To conclude, the regression model using Ridge regressor seems to be the best model for the given data. However given the abnormal relation between predictors and target, further research and data cleaning could be done to improve results. It is recommended to have larger dataset and to group users based on lifestyle to have a more reliable prediction.

3.2 Steps and time participants spend in bed awake

For this part of the report on the Fitbit dataset, we wish to investigate how much time participants spend falling asleep and how that might be related to their activity habits. Based on the data from the ‘daily activity’ file and the ‘sleep day’ file, we infer activity habits from the number of steps participants take on a given day. Time spent falling asleep is inferred from the difference between time spent in bed and time spent asleep – essentially how long they spend in bed being awake.

Specifically, we ask: Do participants who take more steps spend less time awake in bed? We hypothesize that increased activity levels decrease the time spend falling asleep. The analysis draws inspiration from the case study done by Julien Aranguren on Kaggle (https://www.kaggle.com/code/julenaranguren/bellabeat-case-study).

3.2.1 Pre-processing

Before the analysis itself could be conducted, some initial pre-processing of the data had to be done. Again, the daily activity data and the daily sleep data were merged into a single data frame on ID and date, leaving 413 rows and 20 columns of data from 24 unique participants. To further assess and ensure the quality of the data, a column with the total number of active minutes participants wore their Fitbit on a given day was created. Days where a participant wore their Fitbit for less than two hours were excluded on the assumption that it meant the data from that day would be inaccurate. Additionally, the same assumption was applied to days where a participant spent less than two hours or more than 13 hours in bed. This left the final data frame with 393 rows of data from 21 unique participants.

3.2.2 User types

Inspired by the analysis done by Julien Aranguren, to aid visualization and make the data more interpretable, the data was divided into different user types. For steps, three categories were created: Less than 6000 steps, between 6000 and 10000 steps, and more than 10000 steps. 163 rows were sorted as above 10000 steps, 121 as between 6000 and 10000 steps, and 110 as below 6000 – the proportion of each of these categories can be seen on figure 12. Lastly, the time participants spent awake in bed was also divided into three categories: Fast sleepers, normal people, and insomniacs. The fast sleepers where days were a person spent less than 20 minutes awake in bed, normal people were in between 20 minutes and one hour, and the insomniacs spent over an hour awake in bed on a given day. 219 rows were sorted as normal people, 129 as fast sleepers, and 46 as insomniacs – likewise, the proportions can be seen on figure 13.

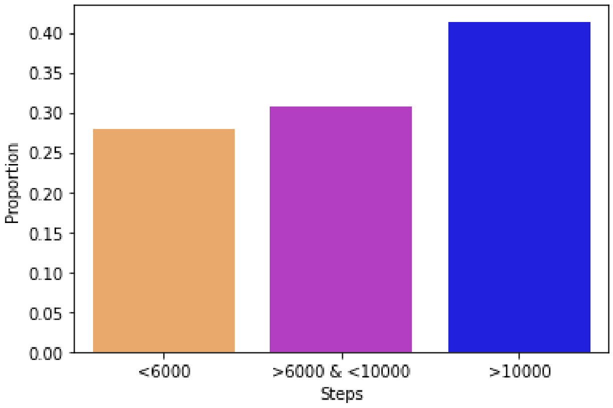

Figure 12

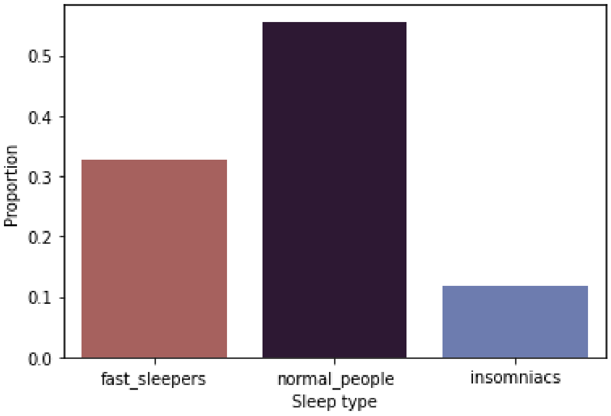

Figure 13

3.2.3 Model

To answer the questions ‘are people who take more steps spending less time falling asleep?’, an initial data visualization was conducted. The scatterplot on figure 14 shows the relationship between the number of steps taken on a given day and how long the participants spent falling asleep. From the figure, it is clear most people spend less than an hour falling asleep, however, on the days where participants do spend over 100 minutes awake in bed, there is an overrepresentation of the > 10000 steps group. This is corroborated by figure 15, where the > 10000 group, on average, spend the most time in bed awake, and the < 6000 group spend the least time – in direct contradiction to what we hypothesized.

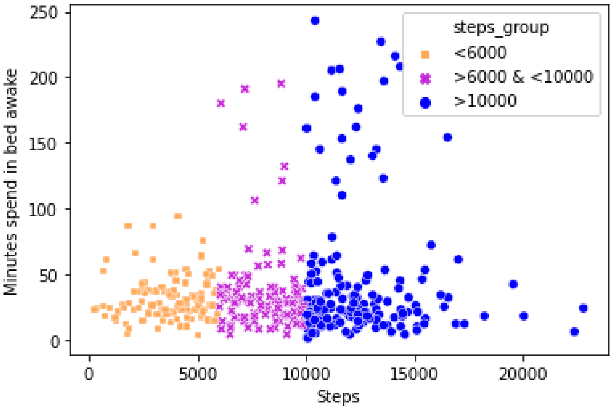

Figure 14

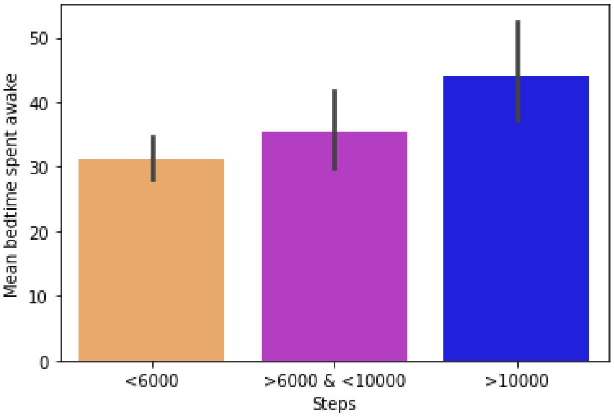

Figure 15

When looking at the proportion of each sleep user type in each of the step groups – which can be seen on figure 16 – the results become more nuanced. From the plot, the > 10000 steps group both has the largest proportion of ‘insomniacs’ and ‘fast sleepers’ out of all three groups – around 15% and 40% respectively. Additionally, the < 6000 steps group has the smallest proportion of ‘insomniacs’ and ‘fast sleepers’ – around 10% and 30% respectively. This could indicate that either there is some ‘sweet spot’ for number of steps taken to minimize the time one spends in bed awake, or that due to the small sample size, individual differences have a larger effect on the results and therefore confound the results.

To statistically test whether steps can predict how long it takes to fall asleep, we fitted a linear mixed-effects model with minutes spent in bed awake as the outcome variable and total number of steps from a given participants on a given day as fixed effect. Our model had random intercepts for participant to account for the lack of independence of data points. Number of steps was scaled using the StandardScaler() function

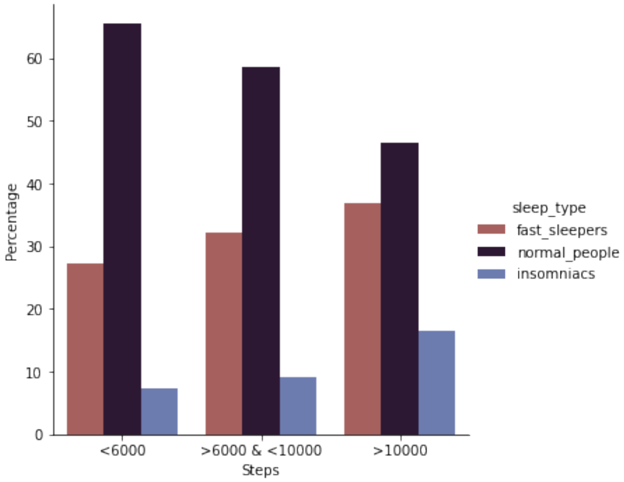

Figure 16

before fitting the model. Time spend in bed is not significantly modulated by the number of steps taken (β = -1.496, z = -1.486, p = 0.14). However, had the relationship been significant, the coefficient would indicate a negative relation between number of steps taken and time it takes to fall asleep when accounting for individual differences with random intercepts, meaning more steps predict less time spent in bed awake.

To answer whether people who take more steps fall asleep faster, we first visualized the data and then performed a linear mixed effects model to statistically test the relationship. The data visualization initially contradicted our hypothesis, as it looked as though more steps led to more time spent in bed awake. However, when individual variance was accounted for in the statistical model, a non-significant negative relationship was revealed. The results therefore remain inconclusive. Future research should focus on replicating the analysis with more data. With a sample size of just 21 unique participants, the statistical power of the analysis is likely insufficient.

4 Discussion

As mentioned throughout the report, the quality and consistency of the Fitbit data varied somewhat. This of course influences the accuracy of the different analyses performed, and therefore future research might want to look into how improvements could be implemented to obtain better results. A few of such improvements are listed below:

- A larger sample size: Modelling is a process that requires a large sample of data input, so when the data set provided contains very few unique data points, for example number of users, it may cause the analysis of the data to be somewhat inaccurate. Even though the Fitbit data includes a great amount of data for each participants (sometimes down to second-level collection), the fact that only 30 unique participants are included – and that there are large inconsistencies in the number of participants among the different data-files - can create confounds in the analyses.
- A more consistent data set: Several data points were missing in the Fitbit data, for example one user may have daily data, however another user may only have less than a full month's worth of data. Additionally, not all participants were represented in all files. This causes a lot of inconsistencies within the analysis, as one has to work around the constraints posed by the messy data.
- Experimental manipulations and greater control: Since participants merely used their Fitbits as they normally would, very little can be said about causation between the different variables. To cement a causal link between any of the variables - such as activity level and sleep - there should be a larger degree of control in data collection (so that some participants do not e.g., only wear the Fitbit for 30 minutes a given day). Additionally, at least one experimental condition could be implemented with advantage.