Project Big Data - Group 22

Fitbit Analysis Report

Yildiz, Damla	Akaçık, Onder	Giaj Levra, Federico
2755457	2757404	2674188

3 July 2022

Contents

0.1 Introduction

In this project, we are using Fitbit Dataset which is a generated by 30 participant's personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring [3][4]. We will try to answer our research questions which covers a wide range, from sleeping habits of the participants to hours participants are most active, and thus helps us analyze the dataset in a systematic way. In the end, we will try to draw conclusions from these analysis and use them to fit regression models to our dataset.

0.2 Research questions

0.2.1 Activity of Users

The data of daily activity and daily sleep were used from the Kaggle: FitBit Fitness Tracker Data Set.

- What is the relationship between the number of steps taken per day and the amount of calories burned? What else can be obtained other than that relationship, given the data?
- What is the relationship between the time spent and the distance taken for each type of activity? What else can be obtained other than that relationship, given the data?
- Is it possible to cluster users into two groups as less/ more sedentary using the time spent sedentarily in a day? If yes, then do the sedentary behavior of those two groups change depending on it's weekdays or weekends, and do those two groups differ in the total time that they sleep per day?

0.2.2 Sleep

The Kaggle data for Fitbit includes sleep data for the users which can be used to answer the following questions:

- What are the sleep habits of the fitbit users?
- What are the activity habits of those who sleep the best and the worst?
- Is there an actionable way of improving sleep quality of those who sleep poorly?

0.2.3 Heart Rate

The data 'heart rate seconds merged' includes recorded heart rate of the users.

- Are there any irregularities in the heart rate data of the users that may indicate a health problem? What else we can conclude using this data?

0.2.4 Activity Days

- Which days users are more active during the week?
- Which hours users are more active during the day?

0.3 Analysis

0.3.1 Activity of Users

In that part of our analysis, we will investigate three questions that are related to the activity of users.

The first question is "What is the relationship between the number of steps taken per day and the amount of calories burned? What else can be obtained other than that relationship, given the data?"

The investigation of the question was started by plotting a regression plot of data points on the dimensions TotalSteps and Calories. As expected, the linear relationship between the TotalSteps and Calories was observed. What is more valuable than observing that relationship is it was noticed that Calories value when TotalSteps equals zero corresponds to Basal Metabolic Rate, shortly BMR. Also, data points with 0 TotalSteps and very low Calories values were observed. Then, it was decided to investigate BMR value for the users and looking more closely to the data points with low Calories values because they might be outliers.

Figure 1: Regression plot of total step taken and calories burned per a day

According to Holland Barett [5], the total calories burned for a day being less than 1000 for an adult is not realistic. Therefore, the data points with Calories value being less than 1000 were selected. It was observed that most of the data points are comes from a day the smartwatch was not worn the whole day. However, there were 4 data points that 0 calories were burned but the smartwatch was on all day. Those 4 data points were removed because the smartwatch was not worn actually.

The next step was investigating the BMR value of the users. Since a BMR value is for a day, the data points that the smartwatch was used all day and 0 steps taken were selected. When the selected data frame was described, it was noticed that there are outlier points with very active minutes spent even though 0 steps were taken. It was thought that it might be because of an activity that does not require taking steps. Since we are investigating Basal Metabolic Rate, these data points were removed as well. Then we ended up with 68 data points coming from 12 different users, with Calories values corresponding to the BMR ranging from 1347 to 2064 with a mean value

of 1804,5 and standard deviation of 248,2. According to Holland Barett [5], BMR value for men ranges from 1600 to 1800, while for women it's around 1550. Therefore, it can be deduced that the data points we have are coming from mostly men users. However, we should also note down the fact that these values might vary depending on the individual's physical characteristics.

The second question is "What is the relationship between the time spent and the distance taken for each type of activity? What else can be obtained other than that relationship, given the data?"

The investigation of the question was started by plotting a regression plot of data points on the dimensions of time spent and distance taken for each type of activity. In the regression plots, the slope of regression lines corresponds to the average speed for each type of activity. In the regression plots, it was observed that the regression line is steeper for a more active type of activity. To make sure numerically, a linear regression model was fitted for each category and the slopes of the linear regression lines were compared. As expected, the slope for VeryActiveDistance/VeryActiveMinutes is 0,067, the slope for ModeratelyActiveDistance/FairlyActiveMinutes is 0,042, the slope for LightActiveDistance/LightlyActiveMinutes is 0,017 and the slope for SedentaryActiveDistance/SedentaryMinutes is 0,000.

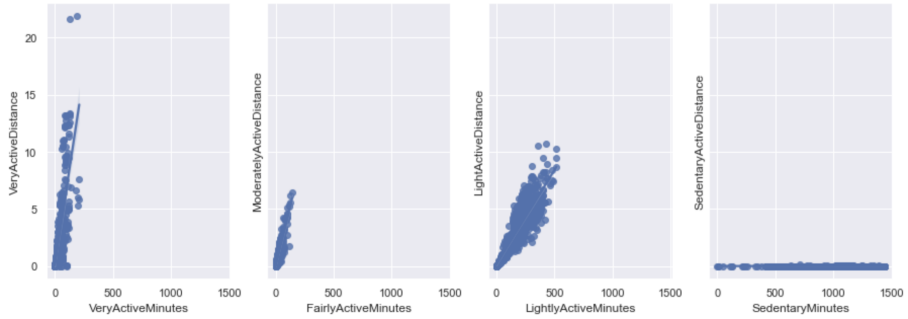


Figure 2: Regression plot of data points on the dimensions of time spent and distance taken for each type of activity

Therefore, there is a correlation between the type of activity and the average speed for that category. If we had proven the opposite, we would have supported one of our previous claims: 'intensity of activities might not be directly proportional to the number of steps taken' because obviously number of steps taken and distance taken are correlated.

The third question is "Is it possible to cluster users into two groups as less/ more sedentary using the time spent sedentarily in a day? If yes, then do the sedentary behavior of those two groups change depending on it's weekdays or weekends, and do those two groups differ in the total time that they sleep per day?"

The investigation of the question was started by plotting the distribution of sedentary minutes. It was noticed that for the density of the distribution there are two peaks, one around 600 SedentaryMinutes and the other around 1250 SedentaryMinutes. In addition, the distribution of average sedentary minutes per user was plotted and a jump after the user with Id: '7086361926' was observed. Lastly, the average sedentary minutes of the data were calculated and it was observed that the discrimination of two peaks of the first distribution and the jump of the second distribution correspond to the average value calculated. Therefore, the users can be clustered into two groups using the average as a criterion. Then a new column that indicates whether being more sedentary or not was added.

Figure 3: Distribution of sedentary minutes using histogram and kernel density estimator

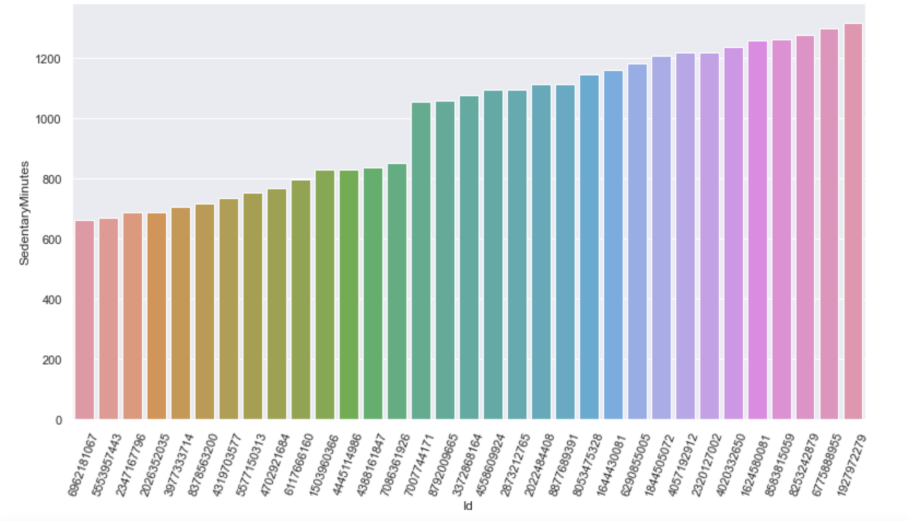


Figure 4: Distribution of of average sedentary minutes per user

Then, the behavior of those two groups change depending on their weekdays or weekends was plotted using boxplots. It was observed that for the less sedentary group, users tend to be more sedentary on the weekdays. While the difference between weekdays and weekends for the more sedentary group is not that remarkable.

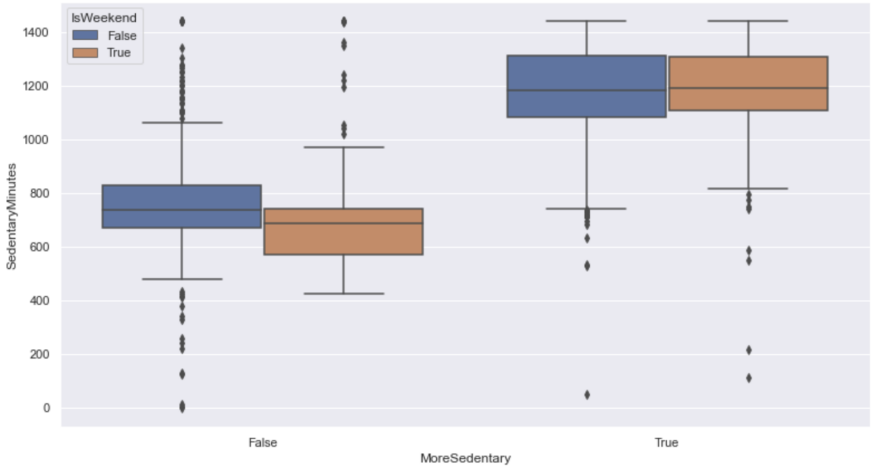


Figure 5: Boxplot of sedentary minutes depending on the two groups and whether it is weekend or not

Lastly, whether two groups differ in the total time that they sleep per day or not was investigated by plotting boxplots of two groups for total minutes asleep. It was observed that the average total sleeping time for the less sedentary group is greater than the average total sleeping time for the more sedentary group. In addition, it was observed that the more sedentary group has a wider range of data points for total minutes asleep. That might be because some more sedentary people tend to sleep less because they are not getting that tired during the daytime and some more sedentary people are sleeping a lot because they live a lazy lifestyle.

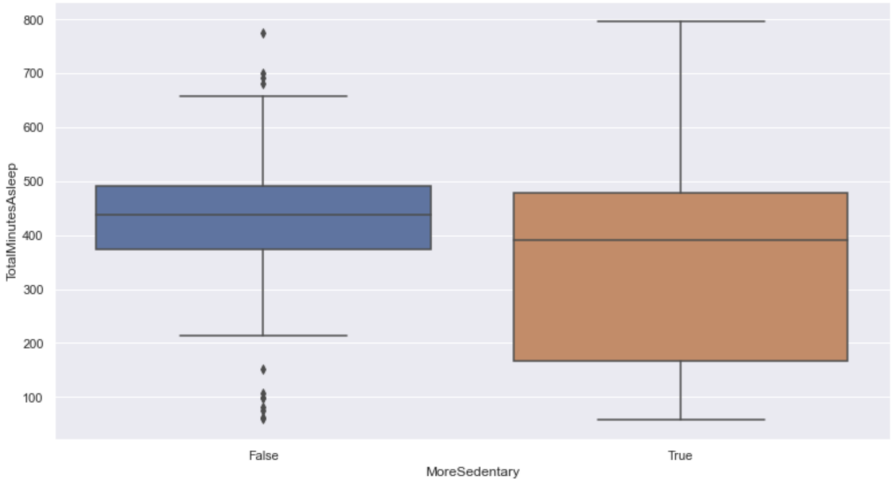


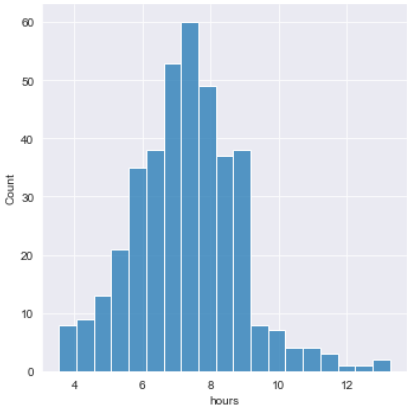
Figure 6: Boxplot of total minutes asleep per day for the two groups

0.3.2 Sleep

What are the sleep habits of the fitbit users?

The Kaggle dataset contained information about sleep which we decided to look into. We started by assigning each record to a group based on the total time spent sleeping.
We examined the dataset and discovered that there existed a certain amount of records which indicated less than 3 hours of sleep. Although this could happen sometimes, it is definetly not the norm, and therefore we decided to exclude the records from the dataset. This left us with a cleaner distribution.
We then identified each record based on the amount of sleep. We used three bins which identified three distinct ranges. The ranges were assigned based on the suggestions from the Sleep Foundation [2]. We defined the records which got between 7-9 hours of sleep as 'good', those which got less as 'deprived' and lastly those which got more than 9 hours as 'excess'.
This allowed us to deepen our analysis by verifying how many nights of good sleep each ID had recorded. In the dataset we counted 192 nights of good sleep, 160 of deprived and 38 of excess. Bare in mind that those values are shared among 21 users, therefore we will have to verify which users are getting good sleep and which are not.
We calculated a ratio of good nights over total recorded nights for each user. We then selected a cutoff point for the previously calculated ratio, which was used to define those which are 'good sleepers'. To be defined as such, the ID must have at least 65% of the recorded nights as 'good'. Among the 21 users, 5 were labelled as 'good sleepers'.

Figure 7: Sleep distribution.



What are the activity habits of those who sleep the best and the worst?

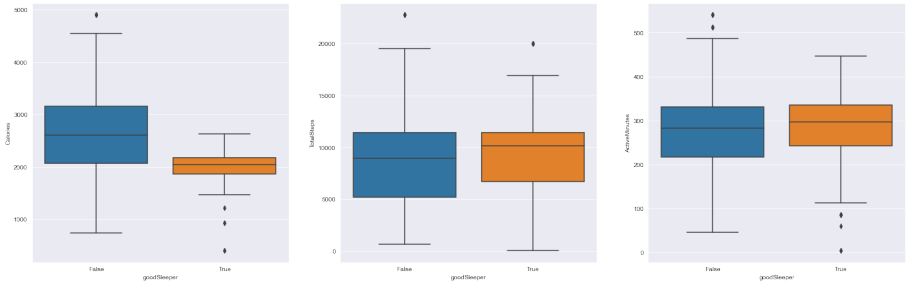


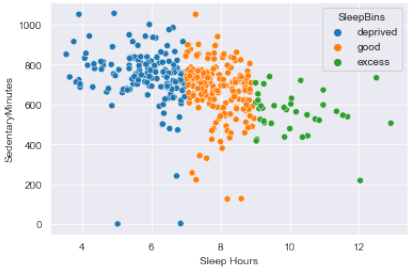
Figure 8: Different activities of good and bad sleepers

To answer the second question we introduced another dataset which included information about the activity of each user.
We plotted the amount of minutes an individual has passed being sedentary against the amount of hours of sleep. The results are shown in plot 9, which show that there is some correlation between

the amount of activity and more sleep. Moreover, in line with these findings, we discovered that those who sleep well on average have higher heart rate during the day. The results are nothing new, as we would expect those which are less sedentary and thus more active to sleep more, in order to rest and recover from the activity.

We then looked at the distributions of burnt calories, total steps and active minutes among the two groups of good sleepers vs bad sleepers. Plot 9 shows us that there are some significant differences among the users which sleep enough and those who do not. In fact we can see that regarding calories consumption, the better sleeping group burns much less calories than the opposite side. Secondly, on average good sleepers walk a little more and the interquartile range is more packed, while bad sleepers appear to be moving much more but with much more variability in the amount of steps. Lastly, as we stated before good sleepers are more active on average, and again the iqr is more packed. In both good and bad sleepers there are outliers in opposite directions, and good sleepers have tighter min-max values

Figure 9: Sedentary vs Sleep



These findings show an interesting version of the facts. Those which sleep better do not burn more calories than those who sleep badly. Although they are more active, they do not perform activities which burn excessive amounts of calories.

Is there an actionable way of improving sleep quality of those who sleep poorly?

Our previous analysis shows that users can improve their sleep through fairly simple actions. In fact, by exercising at a low intensity, one could adopt an activity habit of a good sleeper. As opposed to the high calorie burning activities, which bad sleeper perform in bursts.
A small amount of constant activity will make a user closer to the activity levels of good sleepers.

0.3.3 Heart Rate

In this section, we are looking at the heart rate data of the users however we have to note that this dataset does not contain gender, age, stress levels or whether the user is a professional athlete or not.

We first plot the average heart rates of the users during the week (Figure 11) and throughout the duration of our data which covers a whole month (Figure 10). From both of these plots, we can see that there are some users with average heart beats under 50 so we decided to take a closer look at their heart rate. We will plot this participant's heart rates that are lower than 45bpm.

Because for most of the adults, between 60 and 100 beats per minute (bpm) is normal. The rate can be affected by factors like stress, anxiety, hormones, medication, and how physically active you are. An athlete or more active person may have a resting heart rate as low as 40 beats per minute [1]. And using the Fitbit data, the only way we can identify whether a user has a health risk or not using the data we have is through their heart rates. Extremely low or extremely high heart rates may be an indication of a possible health problem.

But since we don't have any more data about the user it wouldn't be safe to say this user may be experiencing health problems. For some cases, user being an athlete may explain relatively low heart rates like this throughout the day. And since gender, age and health status are unknown,

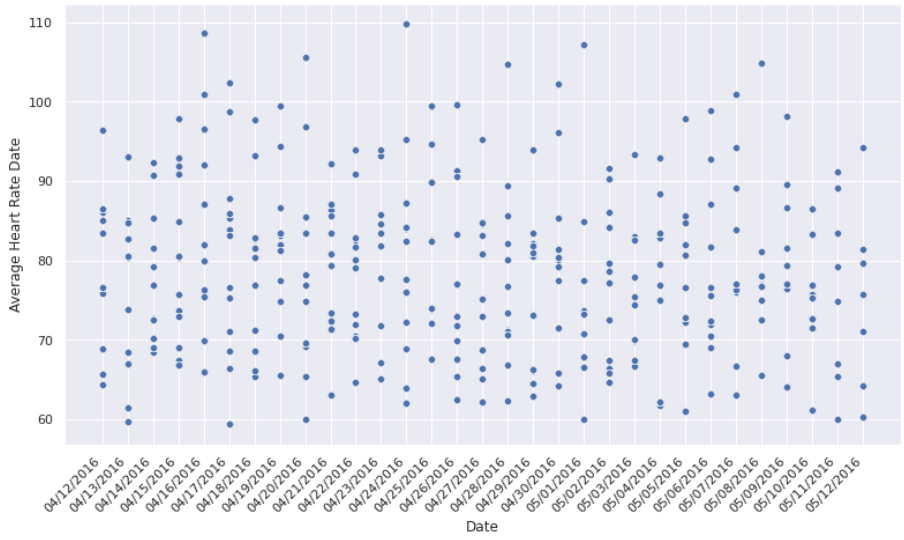


Figure 10: Average Heart Rate Throughout the Dates

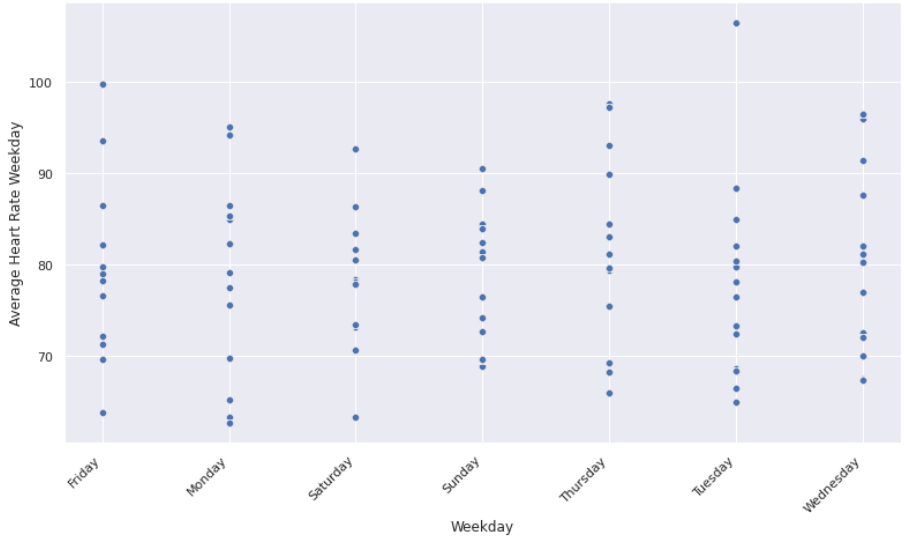


Figure 11: Average Heart Rate During the Week

we can not determine the health risks of the user only from this fitbit data. However, it seems the frequency and hours with low heart rate is irregular for this user throughout the month.

In the end, we decided not to continue with the heart rate data of the Fitbit dataset to make predictions about user's possible health risks.

0.3.4 Days/Hours Users are Most Active

To be able to see if there is a pattern on the days our hours during the users are more active compared to others, we used 3 main features: total steps, average heart rate, and calories.

We will first investigate which hours users are most active during the day.

From Figure 12, we can see that users start to become more active between 6 and 8 am. The peak hours seem to occur between 5 and 7 pm, this is probably because people choose to work out after their work/school. And after 8pm, activity levels start decreasing, probably because people go to bed at that time.

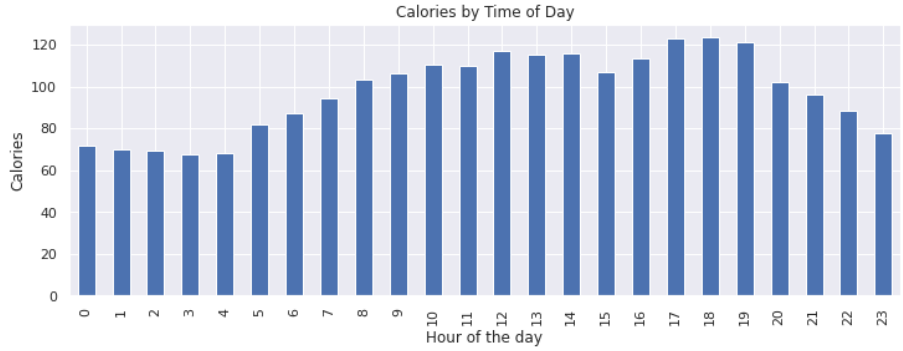


Figure 12: Calories by Time of the Day

We burn around 50 calories an hour [2] while we sleep and can see that in Figure 12. As the users wake up and start exercising, the number of calories burned increases, and peaks around the same time that the users are most active during the day.

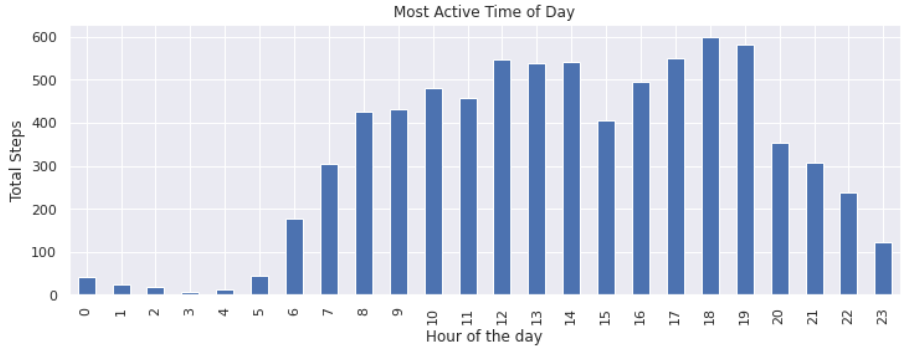


Figure 13: Steps by Time of the Day

Looking at Figure 14, we can safely say that most of the exercise is done between 16:30-18:30 during the day which aligns with our previous findings. And users are mostly inactive during evening between 23:00-05:00 which also aligns with our previous deductions.

Now we will investigate which days users are most active during the week, we will use calories burned and the total distance data. From Figure 15 and Figure 16, we can see that users seem a little less motivated to work out on Sundays and Thursdays. But overall activity level and amount of activity don't change much with different days of the week.

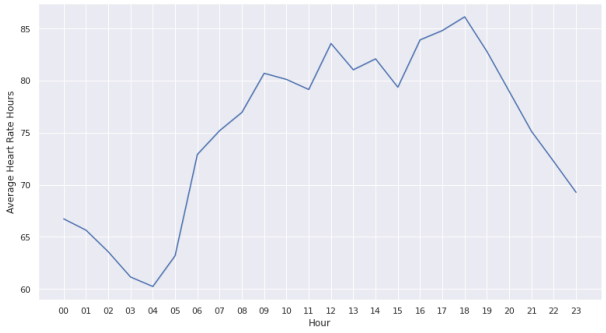


Figure 14: Average Heart Rate During the Day

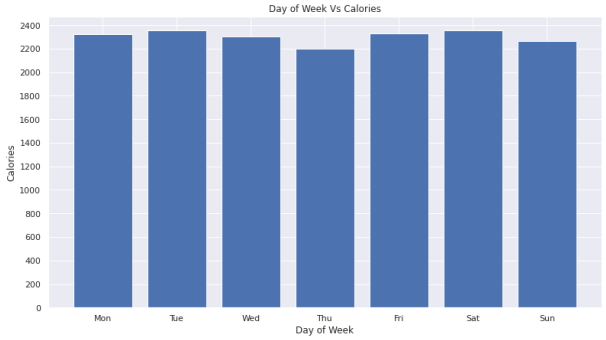


Figure 15: Calories by Day of the Week

In conclusion, we couldn't find a pattern for the days people choose to exercise more, it seems like amount of activity is equally divided among the days of the week. On the other hand, it is clear that users have a specific preferences when it comes to when they want to exercise during the day.

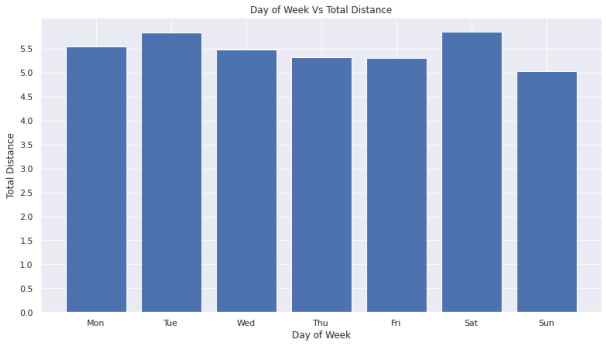


Figure 16: Distance by by Day of the Week

0.4 Models

0.4.1 Step count

Our goal for this section was to predict number of steps. To decide which features we are going to use in our prediction, we first want to see the relationship between several features in the dataset. For this, we can take a look at the correlation between the predictors. From Figure 15, we can see that there is a strong relationship between number of steps and the the other features.

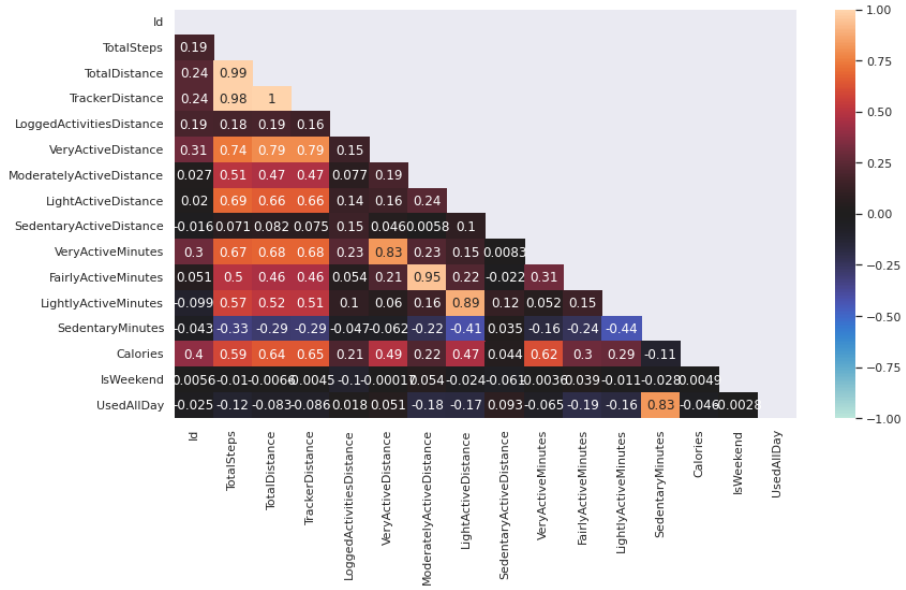


Figure 17: Correlation matrix

Here are few interesting findings not related to the number of steps.

- Different activity levels don't seem to correlate with each other: we can not say, for example, highly active users are more or less likely to engage in modarete or light leveled activites.
- Sedentary active minutes are most negatively correlated to light active minutes which means people who exercise hard are more likely to rest more during the rest of the day.
- As expected, calories burned are highly correlated to the total distance and total steps, however it seems like light activity distance correlate to calories more strongly than the moderate activity level.

In order to predict the steps count, we split our data into train(80%) and test(20%) datasets. The train data has the following predictors:

- TotalDistance
- VeryActiveDistance
- ModeratelyActiveDistance

- LightActiveDistance
- SedentaryActiveDistance
- VeryActiveMinutes
- FairlyActiveMinutes
- LightlyActiveMinutes
- SedentaryMinutes
- Calories
- ActiveMinutes

We fit 3 models to our training dataset: linear regression, ridge regression and random forest regression models. Ridge regression seems to be well suited for the nature of this prediction problem since it has a lower value of RMSE.

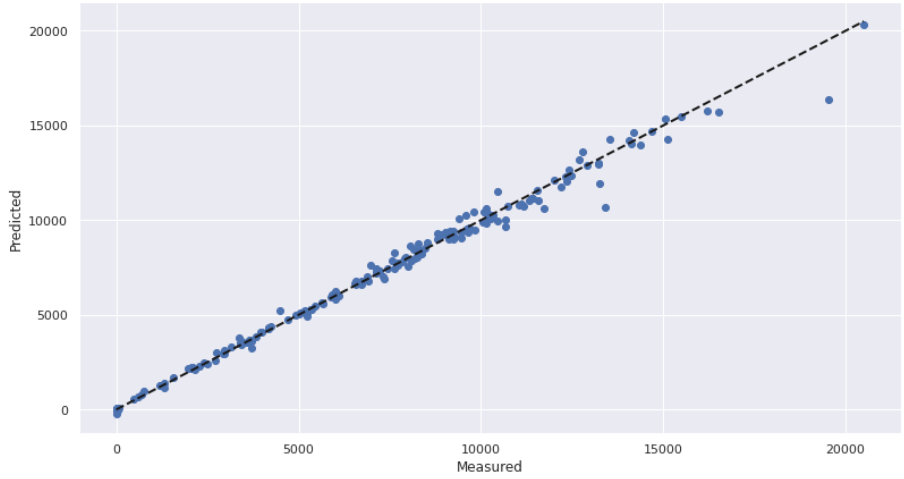


Figure 18: Ridge Regression Model Predictions

Linear Regression Model RMSE = 421.70349787933515
Ridge Regression Model RMSE = 400.7780914821483
Random Forest Regression Model RMSE = 501.7115752527074

Considering the nature of our data, we have strong linear relationships between most of the features. And for this model, it seems like regularization with Ridge (shrinking the least important coefficients) t improves the predictive power of the model. The poor results of Random Forest Regressor can be explained with the nature of the algorithm being a nonlinear algorithm. In our case, the data is very linear and because of this we need a lot more branches per tree to get a good approximation.

0.5 Conclusion

Different conclusions for each question investigated were drawn. Those conclusions will be summed up in this part.

In the first part, activity of users was investigated by three main questions. For the first question, it was shown that there is a linear relationship between the number of steps taken and calories burned. Also, BMR value for the proper data points was investigated. The data was coming from 12 different users and according to the mean, standard deviation, and range of the BMR values, it was deduced that those 12 people might be mostly men. For the second question, it was concluded that there is a correlation between the type of activity and the average speed for that category. More active the type of activity, the greater the average speed. Investigating the third question, users were clustered into two groups, more sedentary or not. Also, it was concluded that for the less sedentary group, users tend to be more sedentary on weekdays. Lastly, it was observed that the average total sleeping time for the less sedentary group is greater than the average total sleeping time for the more sedentary group.

In the second part, we were able to inspect the sleep and activity habits of fitbit users. We discovered that many users only 5 out of 21 users are getting an amount of sleep which is recommended by the sleep foundation. Moreover, good sleepers are slightly more active than those who don't sleep well. But the greatest difference lays in the kind of activity performed. In fact, good sleepers appear to be more constant in their activity, as shown by the packed interquartile ranges, and also less prone to burn a large amount of calories. This analyses ended with the suggestion that those which would like to improve their sleep habits might try to adopt an activity profile as the good sleepers, by being slightly more active during the day but without large energy efforts.

In the last part, we first wanted to see whether we can use the heart rate data to identify health problems among the users but quickly realized it would not be reliable due to lack of data for the age, gender, stress levels of the participants. Then we wanted to find out if there are any patterns on the days or hours people prefers to exercise and saw that there is no correlation between the day in a week and amount of activity done on that day. On the other hand, it was clear that participants preferred to exercise on certain times of the day particularly they were most active between 16:00-18:00. Lastly, we investigated the relationship between different features to select the ones we will use in steps prediction. We tried 3 different models but in the end using ridge regularization gave us the best result.

Bibliography

[1] American heart association. https://www.heart.org/en/healthy-living/fitness/fitness-basics/target-heart-rates, Mar 2021.

[2] How much sleep do we really need? https://www.sleepfoundation.org/how-sleep-works/how-much-sleep-do-we-really-need#:~:text=National%20Sleep%20Foundation%20guidelines1,to%208%20hours%20per%20night., Apr 2022.

[3] Robert Furberg, Julia Brinton, Michael Keating, and Alexa Ortiz. Crowd-sourced Fitbit datasets 03.12.2016-05.12.2016. https://doi.org/10.5281/zenodo.53894, May 2016.

[4] MÖBIUS. Fitbit fitness tracker data. https://www.kaggle.com/datasets/arashnic/fitbit, 2020.

[5] Bhupesh Panchal. What is basal metabolic rate (BMR)? https://www.hollandandbarrett.com/the-health-hub/weight-management/fitness/exercise/what-is-bmr/, 2021.