# Fitbit Tracker Data Analysis Report

## **1. Analytics Team**

*Team members:* Anh Quan Hua (Analytics Team Leader)

Anh Linh Le

Bowen Li

Xiaolu Li

Gouri Nandan Reddy Gangavaram

## **2. Introduction**

*Problem statement:* With Redback Operations aiming to create smart, high-tech environments that incentivize exercising, it is essential that we need to be able draw meaningful insights from users’ data collected through our future IOT devices. In this analysis, our aim is to analyze a comprehensive dataset collected by Fitbit trackers to draw out patterns in users’ behaviors and provide recommendations.

*Questions to be answered:* What can we observe from the typical user’s daily activities? What factors are closely related to users’ calories burnt and steps taken? Is it true that there is a particular time in the day where most users can get effective exercises done? How can the information obtained from analysis help Redback Operations? Can we provide recommendations that corresponds to the problems observed?

*Methodologies:* For this analysis, our team is going to conduct our analysis using Python. We are going to first prepare the datasets, outlining the notable CSV files, cleaning our data of choice then proceed to processing them. Since we have multiple datasets that shared a common attribute (User ID), it is essential for us to transform our data frames and manipulate them so that we can produce a comprehensive main data table. Lastly, we are going to analyze and visualize our desired findings.

## **3. Initial Observation**

*Data Overview:*

* This data set is taken from Kaggle and credited to the following URL: <https://zenodo.org/record/53894#.YnR5PL1BxhG>.
* Data collected from 30 eligible Fitbit users' personal tracker data (with consent). Variation between output represents use of different types of Fitbit trackers and individual tracking behaviors / preferences.
* All of the data presented were approved through the consent of their donors.
* In total, there are 18 CSV files with a combining total of 338 MB:
  + dailyActivity\_merged: A comprehensive file containing information about users' activity in different days.
  + dailyCalories\_merged: A file containing the amount of calories burnt for each day.
  + dailyIntensities\_merged: A file detailing the various user's states during the day.
  + dailySteps\_merged: Contains info about the amount of steps taken per day.
  + heartrate\_seconds\_merged: Containing the heart rate monitored during different periods of the day.
  + hourlyCalories\_merged: Describing the amount of calories burnt by hour.
  + hourlyIntensities\_merged: A file with hourly intensities info.
  + hourlySteps\_merged: File recordings the amount of steps taken per hour.
  + minuteCaloriesNarrow\_merged: A file with the overall amount of calories burnt per minute.
  + minuteCaloriesWide\_merged: File with details on WHAT KIND of calories burnt.
  + minuteMETsNarrow\_merged: Data about the Metabolic equivalents of user per minute.
  + minuteSleep\_merged: Unclear.
  + minuteStepsNarrow\_merged: Overall steps per minute.
  + minuteStepsWide\_merged: Details the step for each second of a minute.
  + sleepDay\_merged: Sleeping details for each day.
  + weightLogInfo\_merged: Daily weigh log.

*Limitations:*

* The dataset, although sizeable, is quite limited in terms of sample pool. As previously mentioned, we are only dealing with data from 30 “eligible” users.
* Data were collected across multiple devices with differing sensors.
* Unit of measurements are not specified, thus making it hard to quantify the meaningfulness of info.
* Being almost 6 years from released, data might not correctly reflect the current fitness trends.
* Unspecified targeted audience – it is undisclosed regarding the users age and sex.

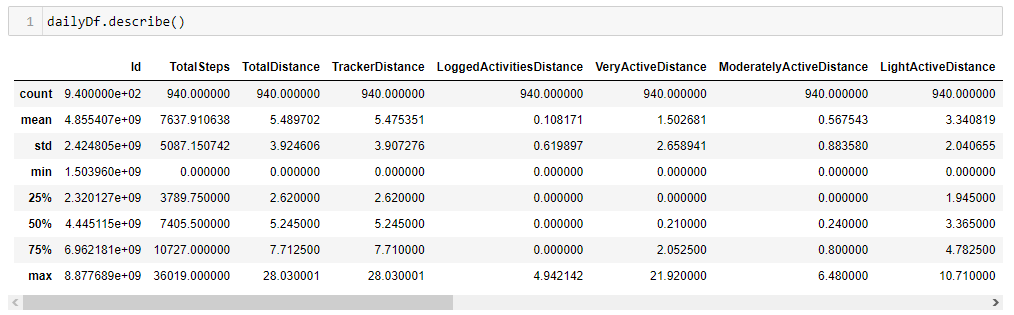
## **4. Data analysis**

*Statistical summary:*

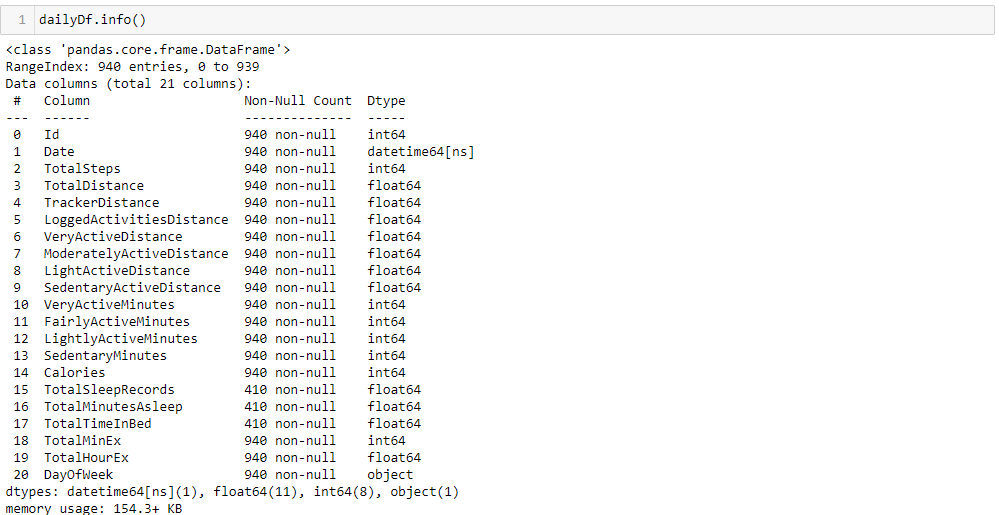
* First, we must prioritize the apparently more valuable CSV files. In this case, our data of choice includes the following: dailyACtivity\_merged, dailyCalories\_merged, dailyIntensities\_merged, dailySteps\_merged, sleepDay\_merged, weightLogInfo\_merged.
* We then prepare our data frames of choice, obtained by passing them through panda’s read\_csv function, for the merge of our dailyDF – a data frame containing all the most practical statistics about the users’ daily routine and activities.



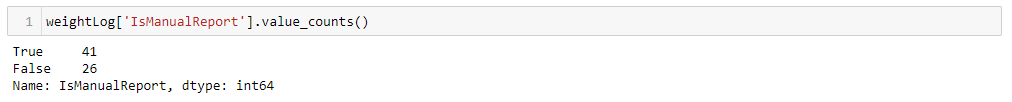
* After generating our dailyDf, we are taking a look at every columns’ min, max, count, meadian, etc.
  + Insights derived from summary:
    - The average number of steps per day is only recorded at 7,637, which is lower than the 10000 steps recommended by the CDC [1].
    - Users typically burn around 2,303 calories per day. Assuming that we have a relatively equal amount of both males and females, this number should be ideal for most adult to maintain weight. This is because according to the UK NHS, the recommended calories intake for men and women is 2,500 and 2,000 respectively [2].
    - The average sleeping time per day is around 419 minutes – which equates to roughly 7 hours. This aligns to CDC’s recommendation for adults [3] daily sleep time.



* Finally, we are going to view all the columns’ name, dtype and count.
  + We can see that 3 particular columns have way less members than the others. The three are TotalSleepRecords, TotalMinuteAsleep and TotalTimeInBed – this can be understood as the result of merging our sleeping records and daily activities data frames. The sleeping records simply were not logged often enough, thus result in the missing of data in over 500 rows.
  + For this analysis, we do not have to address this problem due to the nature of the missing attributes. From the sleeping records, we are still able to derive users sleeping habits and details.

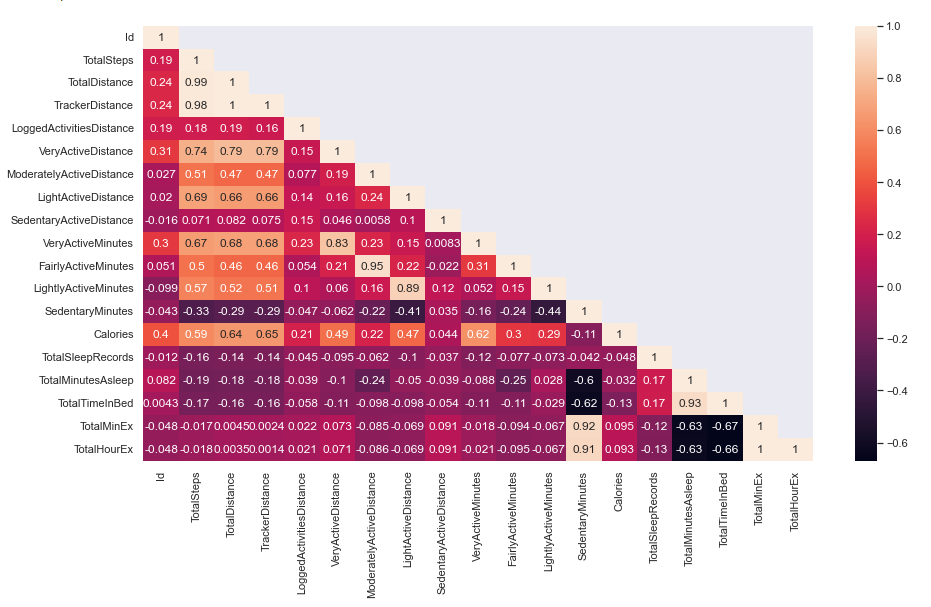


* Briefing the weight log.
  + - Insufficient amount of data with only 67 rows.
    - 60% of the input is manual and 40% is automatic. This might be due to some Fitbit devices being limited in terms of weighting capabilities, users’ unawareness or lack of interest.



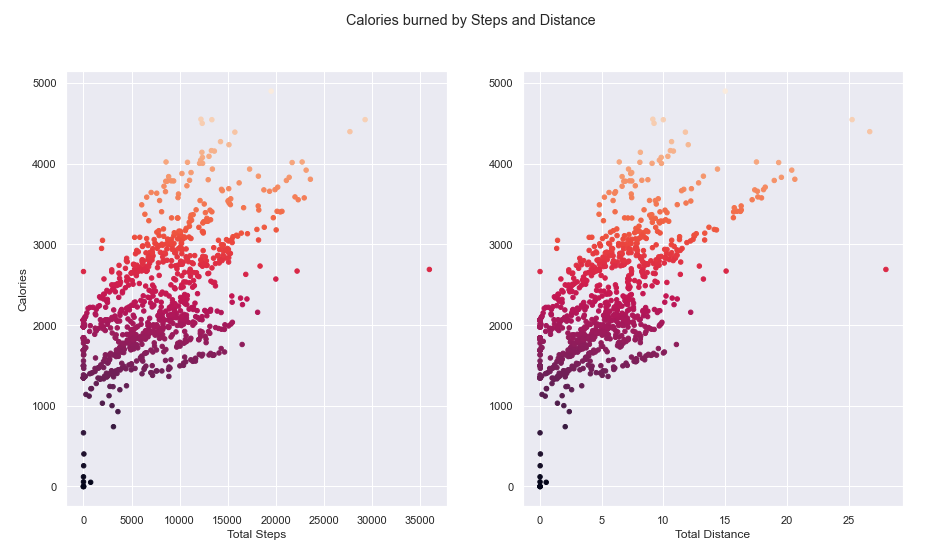
*Correlation matrix:*

* We started by generating a correlation matrix using Pearson’s formula to observe any linear relationships between our attributes:
  + Obviously, we see strong correlations between TotalStep and all the different kind of distance measurements. This just means that more time user spent walking/running/jogging, the more steps they are getting – which is to be expected.
  + Some correlations that also suggest a strong linear relationship FairlyActiveMinutes and LightlyActiveMinutes with their corresponding distance attributes. Once again, this is to be expected as user naturally gain more distance the more they exercise.
  + Interestingly, we also can observe extremely high correlations between TotalMinEx/TotalHourEx with SedentaryMinutes (0.92 and 0.91 each) indicating that the majority of users’ activeness logged by the device is considered “Sedentary”.
  + Overall, the only two noteworthy correlations are between TotalSteps and Calories, as well as TotalDistance and Calories.



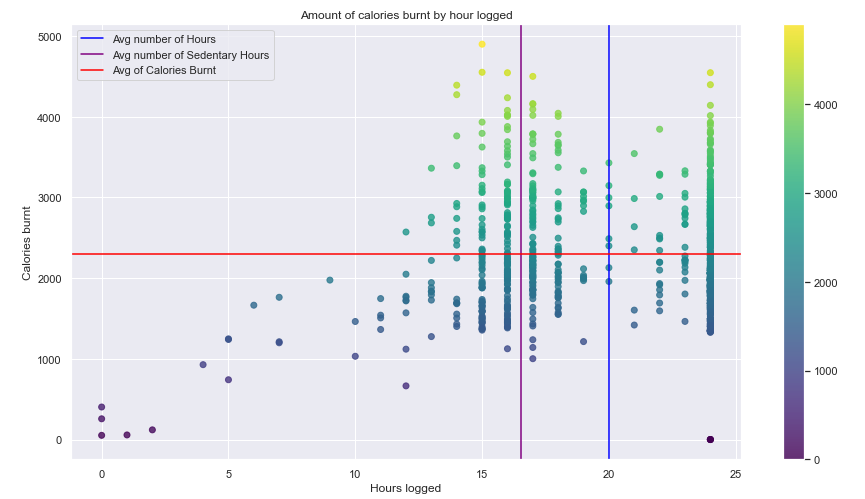
*Scatter plot for linear correlations:*

* After obtaining the correlation matrix, we began visualizing distributions of the attributes of interest:
  + - From the scatter plots, we can see that both strongly suggests the existence of a linear relationships between Calories and TotalSteps/TotalDistance.
    - Both graphs are showing a similar shape/distribution of data, even the outliers are observed to be identical.
    - This means that it is undeniable that steps taken and distance traveled is the simplest way to predict and calculate the amount of calories burned by users.



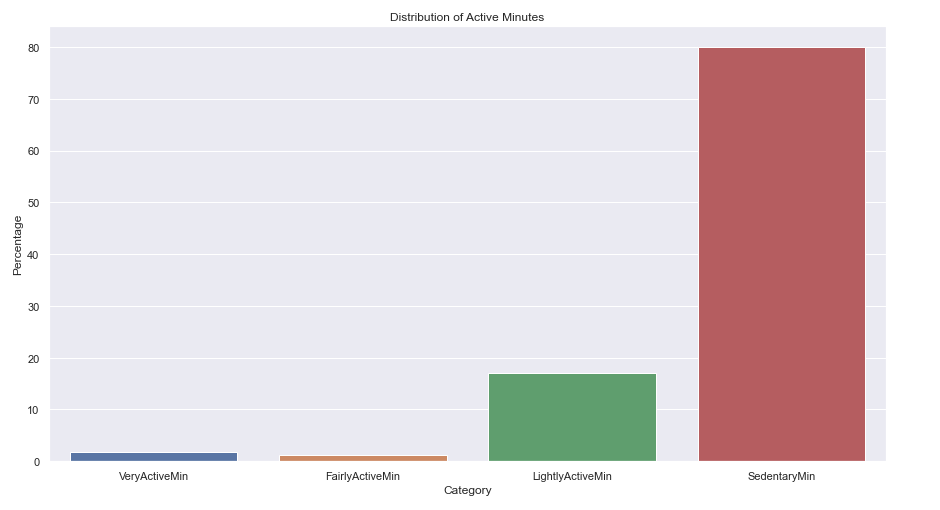
*Users’ exercise habit by hours of the day:*

* In this part, we are looking at how well calories are burned compared to the total amount of time (in hours) logged:
  + - Most users have their activities logged for around 12-17 hours per day or the full 24 hours. On average, 20 hours are logged per day by each user. This along with the lack of users’ sleeping records suggest that a great portion of user might have taken off their Fitbit trackers during certain activities of the day. This can either be the result of comfortability, sense of privacy or sensory inconsistency.
    - The average number of Sedentary Hours falls around 16, suggesting that on average, a substantial portion of users’ time during the day is spent being sedentary.



*Calories burned by Activeness Type:*

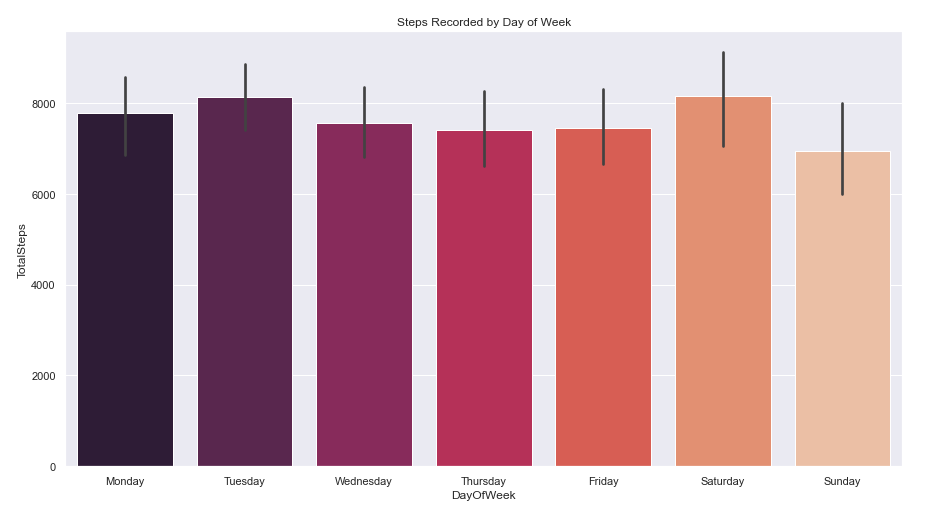
* In this section, we are attempting to take a look at how calories are burnt for time spent per activeness type:
  + - First, there is the average distribution of VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes and SedentaryMinutes. This is obtained by calculating all users’ mean of percentage for each category of activeness. We can see that most recorded time belongs to being sedentary followed by being lightly active, at 80% and roughly 18% respectively. Surprisingly, users tend to stay longer at the highly active state rather than fairly active state.
    - There are no clear relationships between any of the exercise intensity versus the calories, this might be directly due to the fact that there are no dedicated calories count for each type of activeness. We might want to include this feature in our future applications to better analyze how effective users are exercising and to deliver personalized advices for our customers.





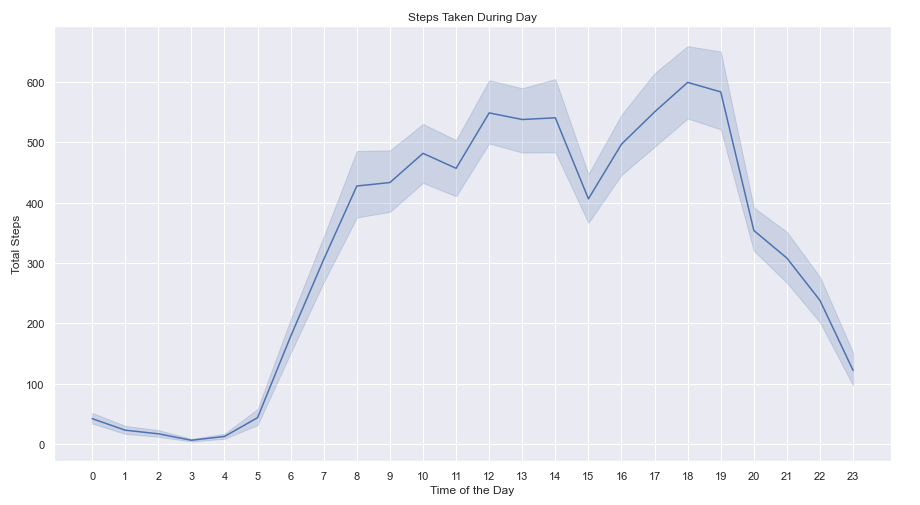
*Steps Recorded by Week Day:*

* For this visualization, we want to see if there is any preference when it comes to choosing a day for exercising:
  + - With our given data, there seems to be a minor decrease in the average of total steps recorded during Wednesday, Thursday and Friday though the difference might be due to limited sample size.
    - Notably, Sunday seems to be the day of choice for relaxing as we see the most significant fall in the number of steps taken.



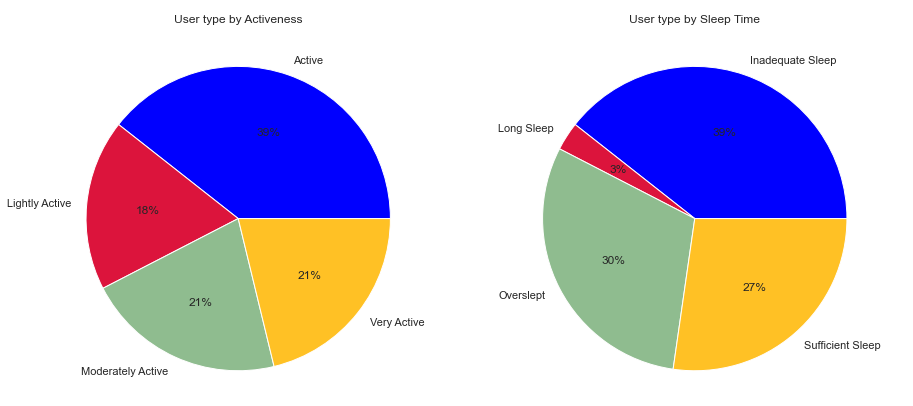
*Steps Recorded by Hour During Day:*

* After taking a look at how steps are distributed throughout the week, it is interesting to see what time during the day users usually exercise:
  + - The period from 7:00 AM to 8:00 PM is where most of our steps are recorded.
    - The two most active periods are 12:00 PM to 2:00 PM and 4:00 PM to 7:00 PM. This is not surprisingly as the two periods tend to be outside of work hours.
    - The first period suggests that users usually get the chance to exercise, in this case most likely walking, during lunch break while the second implies that users tend to exercise right after work.



*Categorizing users by Sleep Time and Steps Taken:*

* For our last visualizations, we are going to see how users are structured when being categorized by their steps taken and sleeping time:
  + - With our first pie chart, users are categorized as followed:
      * + Lightly Active: Taking less than 4,000 steps on average for each day.
        + Active: Taking 4,000 to 8,000 steps per day.
        + Moderately Active: Taking 8,000 to 10,000 steps per day.
        + Very Active: Taking over 10,000 steps per day.
    - For our second pie chart, users are categorized as followed:
      * + Inadequate Sleep: Taking less than 7 hours of sleep on average for each day.
        + Sufficient Sleep: Taking 7 to 8 hours of sleep per day.
        + Long Sleep: Taking 8 to 9 hours of sleep per day.
        + Overslept: More than 9 hours of sleep on average.
    - In terms of step taken, our users base this time are quite active – only 18% of which falls into the ‘Lightly Active’ category while 39% can be considered “Active”. Interestingly, 21% of users average more than 10,000 steps per day, highlighting a very active lifestyle.
    - On the other hand, 39% of our users does not have an adequate amount of sleep daily. Sufficient sleep and long sleep only occurs in 30% of our sample pool – which happens to be equal to the number of users who oversleep every day.



## **5. Summary, Suggestions and Recommendations**

*Summary:*

On average, users are taking a good amount of exercise daily, burning through the recommended calories intake and getting adequate sleep. However, when examining in further details, we see that a sizable portion still needs to get more sleep while others might want to lessen their time in bed. Additionally, users should be taking more steps and burn through more calories to help with reducing fat and losing weight.

Users are not keen on keeping track with their weight, this either is a result of not having the ability of registering weight automatically or just personal choices.

A strong linear relationship between the amount of calories burned and total steps/distance is definitely present. We could not see any other relevant correlations between attributes without heavy speculations – risking the integrity of our data.

On average, the trackers log data for around 20 hours per day – suggesting that many users do not always have their devices on. This might be due to personal habits, choices or comfortability/compatibility issues.

When exercising, users tend to be light active and on the extreme end, very active. This indicates that a minority of users had short bursts during their workout where they try to be as active as possible. Additionally, a substantial portion of time daily for the majority of users was spent being sedentary.

Users did not show any sign of preference for workout days but Sunday was the choice for lesser exercises.

Most steps are taken during lunch break and right after evening’s normal work hours, suggesting that users want to make the most of their time during their daily commutes. Our sample pool also prefers not going for exercises before 6:00 AM or after 8:00 PM.

*Suggestion and Recommendations:*

For future developments, it is recommended that a system should be created to incentivize users to be more active and lessen their sedentary time:

* + Allowing users to set up goals to easily keep track.
  + Adding an extra calorie burned attribute for the intensity records to better produce personalized advice on how users should conduct their workout.
  + Suggesting for reminders during break of after 5:00 PM to take a light walk/run/jog.
  + Letting users connect and socialize with friends, family, and exercise groups to generate friendly and productive competition.
  + Focus more on getting the users to work out during weekdays or personalized time to align with their schedule.
  + Thoroughly testing and prototyping wearable devices for comfortability as well as being durable for daily usage.
  + Alerting users to get enough sleep or recommending short naps for a healthier lifestyle.
  + Prompting for daily weight logs in order to keep track of BMI changes frequently.

## **6. Reference**

[1] Centre for Disease Control and Prevention (n.d), *Session 8: Take Charge of What’s Around You,* CDC – Centre for Disease Control and Prevention, accessed 26th April 2022, <https://www.cdc.gov/diabetes/prevention/pdf/postcurriculum_session8.pdf>

[2] Nation Health Service UK (2019), *What should my daily intake of calories be*, NHS UK, accessed 25th April 2022, <https://www.nhs.uk/common-health-questions/food-and-diet/what-should-my-daily-intake-of-calories-be/>

[3] Centre for Disease Control and Prevention (2017), *How Much Sleep Do I Need?*, CDC – Centre for Disease Control and Prevention, accessed 26th April 2022, <https://www.cdc.gov/sleep/about_sleep/how_much_sleep.html>