Google Analytic Case Study: How Can a Wellness Technology Company Play It Smart?

Scenario

I am assigned as a junior data analyst working in the marketing analyst team at Bellabeat, a high-tech manufacturer of health-focused products for women. Bellabeat collects data on activity, sleep, stress, and reproductive health that allows Bellabeat to empower women with knowledge about their own health and habits. Bellabeat is a successful small company, but they have the potential to become a larger player in the global smart device market.

Background:

Urška Sršen, cofounder and Chief Creative Officer of Bellabeat, wants my team to focus on a Bellabeat product and analyze smart device usage data in order to gain insight into how people are already using their smart devices. Then, using this information, she would like high-level recommendations for how these trends can inform Bellabeat marketing strategy.

Stakeholders:

Urška Sršen: Bellabeat's cofounder and Chief Creative Officer

Secondary Stakeholders:

Bellabeat marketing analytics team: A team of data analysts responsible for collecting, analyzing, and reporting data that helps guide Bellabeat's marketing strategy.

Objective:

Examine the dataset collected on Fitbit, a non-Bellabeat product, and find out how Fitbit user's usage pattern can be used to study how users might use Bellabeat product. Then, apply gained insights on which area to target our marketing effort, thus expanding our clientele.

In doing so, I hope to help stakeholder better understands:

1. What are some trends in smart device usage?
2. How could these trends apply to Bellabeat customers?
3. How could these trends help influence Bellabeat marketing strategy?

Product to compare Fitbit to: Bellabeat Time – a wellness watch that combines the timeless look of a classic timepiece with smart technology to track user activity, sleep, and stress. The Time watch connects to the Bellabeat app to provide you with insights into your daily wellness.

This is because the functionality resembles the Fitbit the most.

Data Preparation:

The analysis will be done with the Fitbit Fitness Tracker Data Set (CC0: Public Domain, dataset made available through Mobius). 30 Fitbit users consented to the submission of the personal tracker data that includes information about daily activity, steps, and heart rate that can be used to explore users' habits.

I will be focusing on three tables total: "dailyCalories_merged", "sleepDay_merged', "dailyActivity_merged", and "weightLogInfo_merged". Now renamed to be "calories", "sleepday", "dailyactivity", and "weight".

Limitation and Assumptions:

- We can only assume this sample size is sufficient. Given the approximate number of Fitbits users is around 31 million people, the results derived from this analysis would be far from statistically significant to make any business decisions.
- There are no metadata available to identify the units of measurements or explanation of data
- Old data. Since the data is collected back in 2016, and Fitbits have pushed out new products with a bigger client base, we can only assume that this data is still relevant.
- We have no way to validate the credibility of the third-party data.
- A lot of records are missing. This could be due to technical issues (Fitbits out of battery), users not wearing their Fitbits, and other unforeseen reasons.

- There is no demographic information, such as age, gender, height, or initial weight about the users. We will need this data to be able to target our market effort better.

Data Cleaning

The datasets are duplicated, and the raw data set is stored in a separate folder. The duplicated set is then reviewed in Microsoft Excel:

1) Reviewed the spreadsheets to get a good idea on what I have to work with.
2) Sorting and filtering: Removing numbers that are too big, too small, or values that does not belong in a column.
    a. Reveal "null", or NA values in the columns using filters. By removing those rows, it would allow for smoother cleaning later.
    b. Check in excel that all id contains 10 characters. This will be essential for using JOINS in SQL later. This can be done easily by using the =LEN() function in excel to count the number of characters, then apply a filter to see the max length of the ID column. In our case, all the IDs are 10 characters long. Furthermore, we can also use this function to check the number of bytes we allocate to each column when we create the SQL tables with CHAR() or VARCHAR().
3) Performed the duplicate removal function on all data sets.

MySQL Date cleaning:

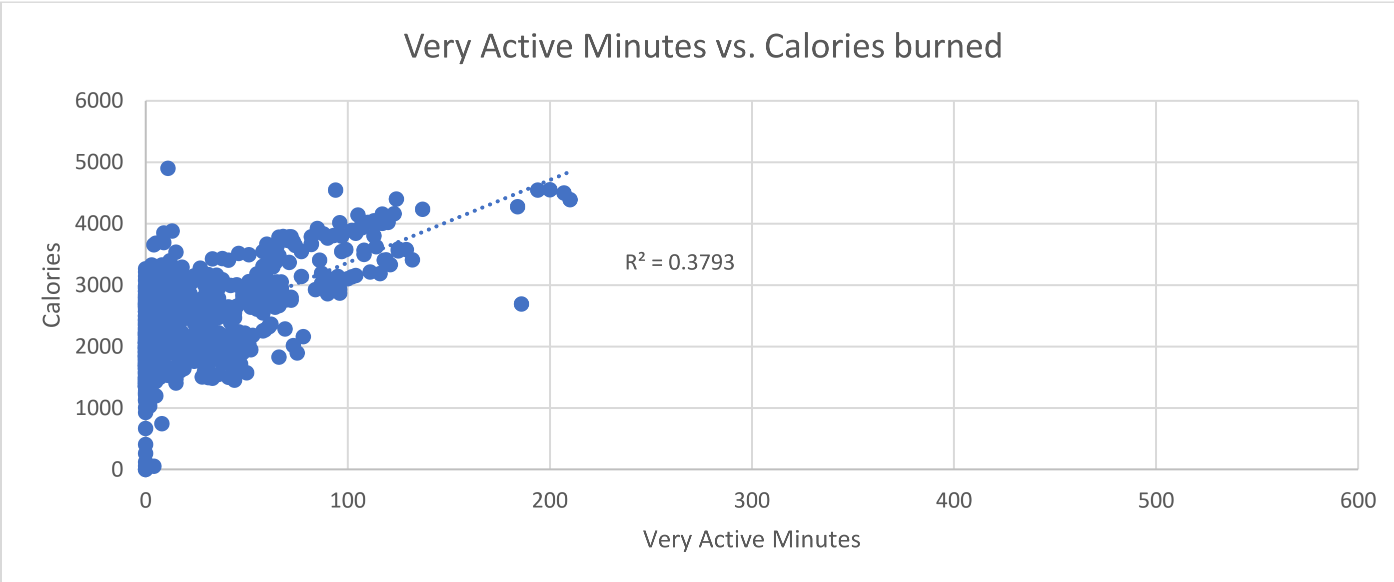
1) Since the CSV date format does not fit the ISO format used in MySQL, I will have to first update the dates into YYYY/MM/DD by using the "UPDATE" function. Alternatively, we can also create a new table with the days in ISO format using a STR_TO_DATE() function, drop the existing table, then alter the new table name to replace the dropped table.
2) In excel, we have checked for duplicated ID with the filter function. In MySQL, we can go a step further by checking to see if there are duplicated rows on the same day. Results shows three duplicated rows in the sleedaytable. This was eliminated by creating a new table using the DISTINCT (*) function.
    - This is beneficial because it can perform the clean much faster than going through each of the four excel separately. In MySQL, we can simply change the table name and reuse the old codes.
3) Assigned each field with their respective data types.

Area of focus:

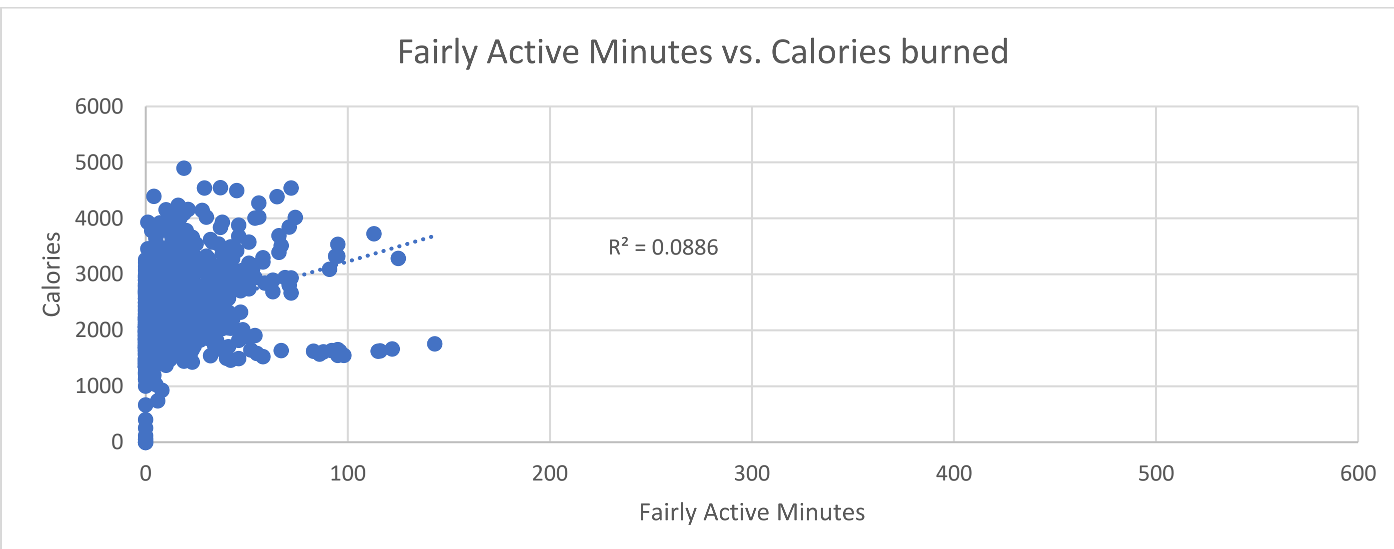
1) The minutes of different activeness should be based upon the number of calories burned.
2) Recording the activeness of each user would allow Bellabeats Time to better recommend fitness plans and programs.
3) People with higher average active minutes to sedentary minute ratio have shown weight loss.
4) Being more active should improve the user's sleep quality.

Analysis:

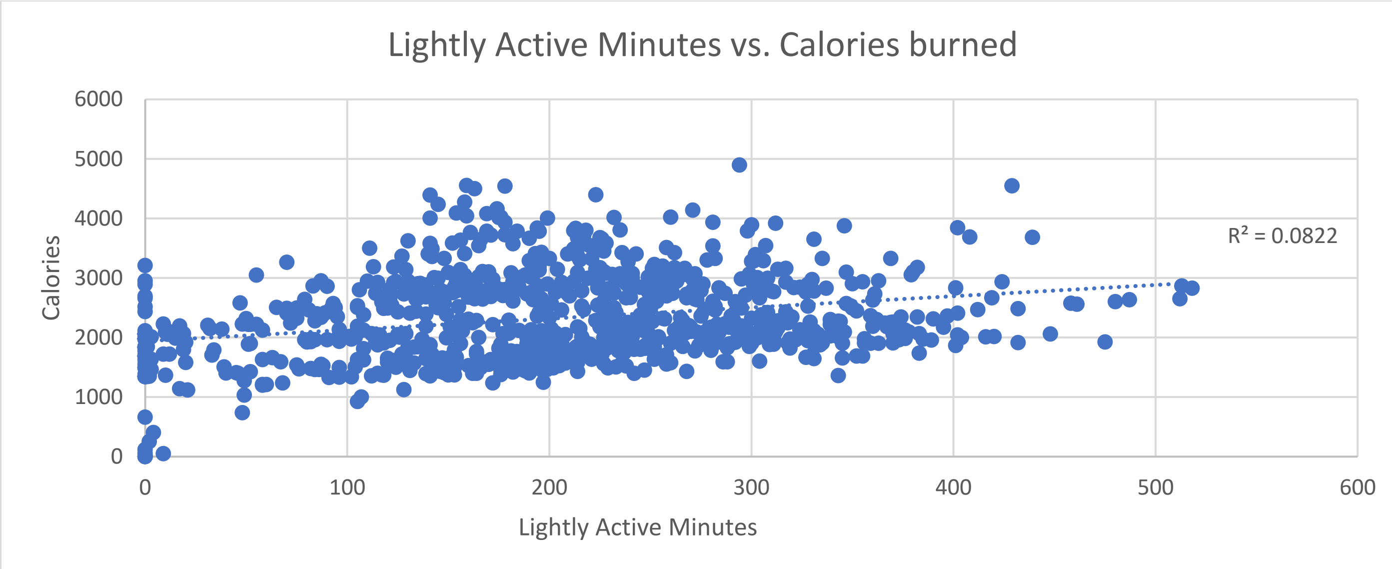
<u>Active minutes vs. Calories burned</u>



Minutes where users are very active shows the strongest relationship with amount of calorie burnt



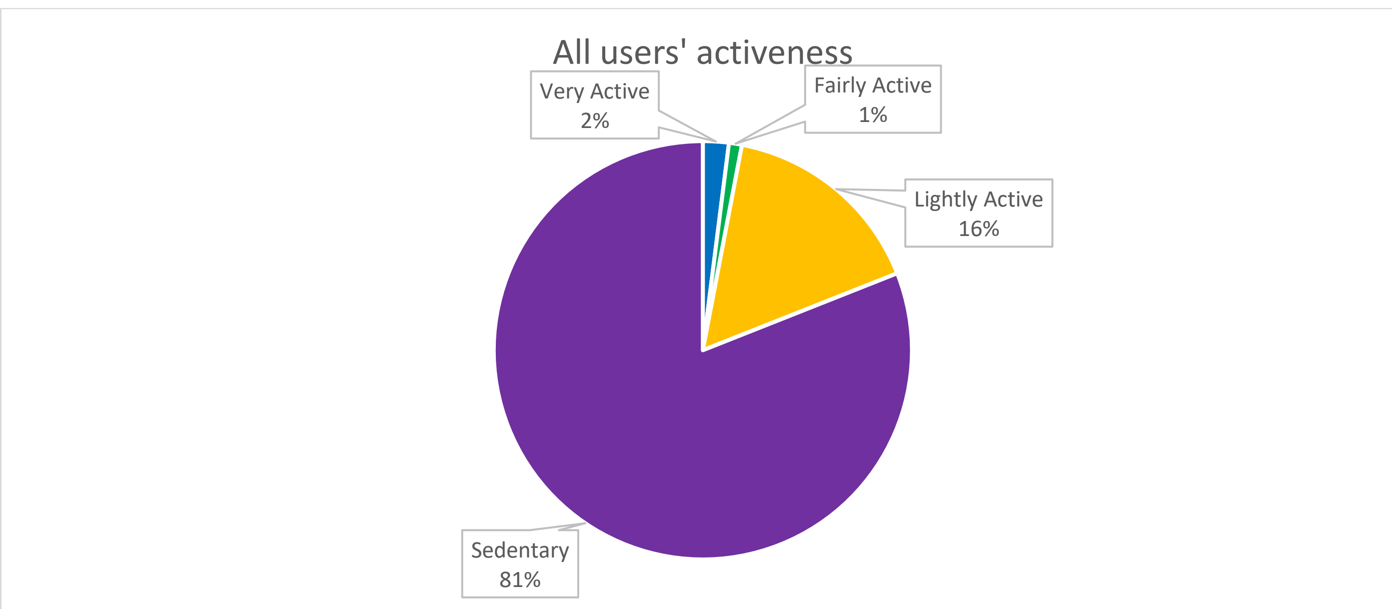
Consequentially, the fairly active minutes has a weaker correlation to calorie burned



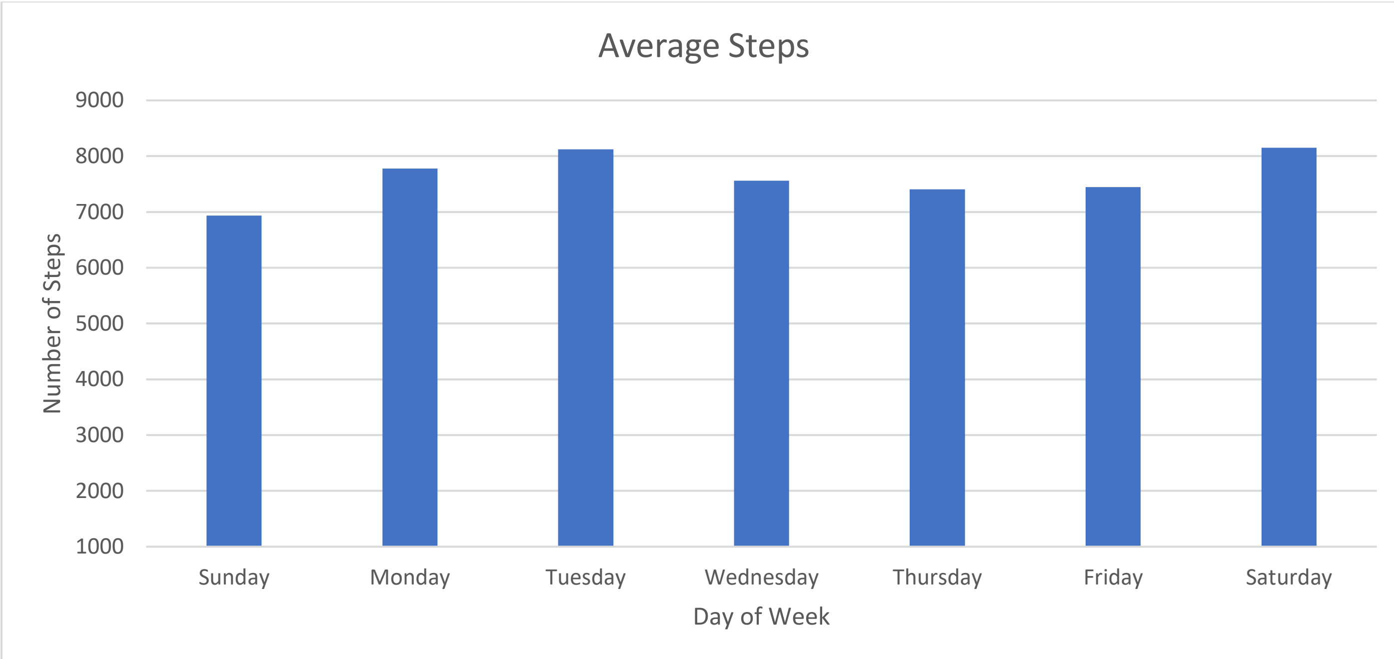
The lightly active minutes to caloric expenditure relationship is consistent with the previous notes. The correlation is decreasing.

Note: The analysis did not include sedentary minutes because that does is not considered as being “active” according to the [US Centre for Disease Control and Prevention (CDC)].

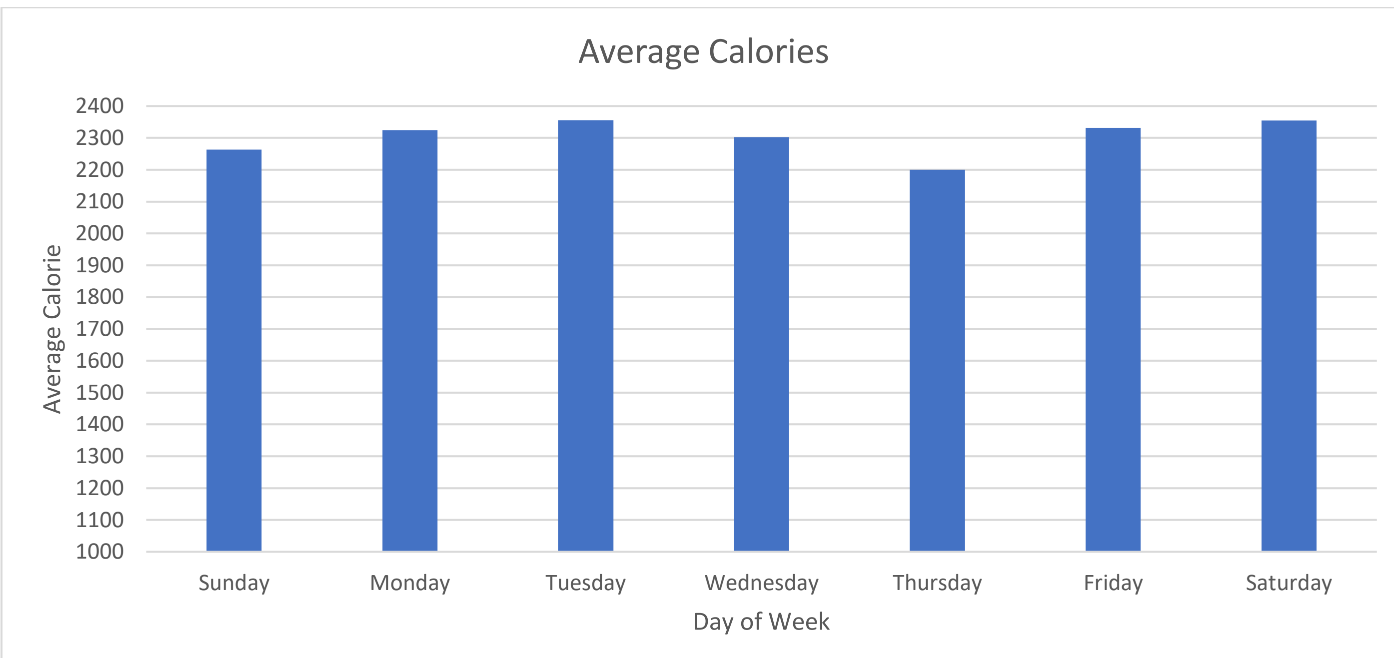
Users’ Activeness vs. Recommended amount



The data reflected that all users are sedentary most of the record time.

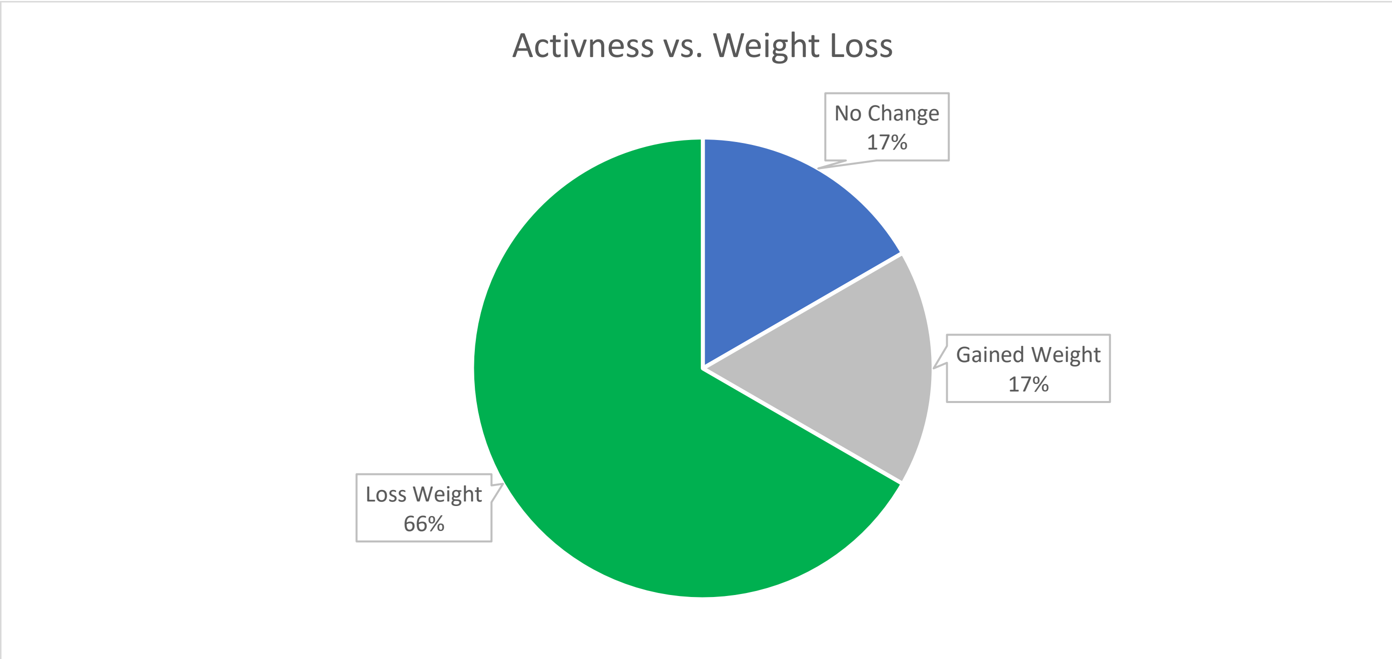


The highest difference of steps is between Saturday and Sunday. However, none of them met the [recommended 10,000 steps] everyday.



Similarly, there is no significant differences between the calorie burned each day despite the different level of activeness. The biggest difference in caloric expenditure is 156 calories, which is roughly a teaspoon of olive oil.

Activeness vs. Weight loss



Only 66% of the users have recorded weight loss at the end of the recorded period.

ID	Total Active Minutes	Percentile Among Peers
8877689391	9632	6.25%
4558609924	9581	9.38%
6962181067	8901	12.50%
4319703577	7585	43.75%

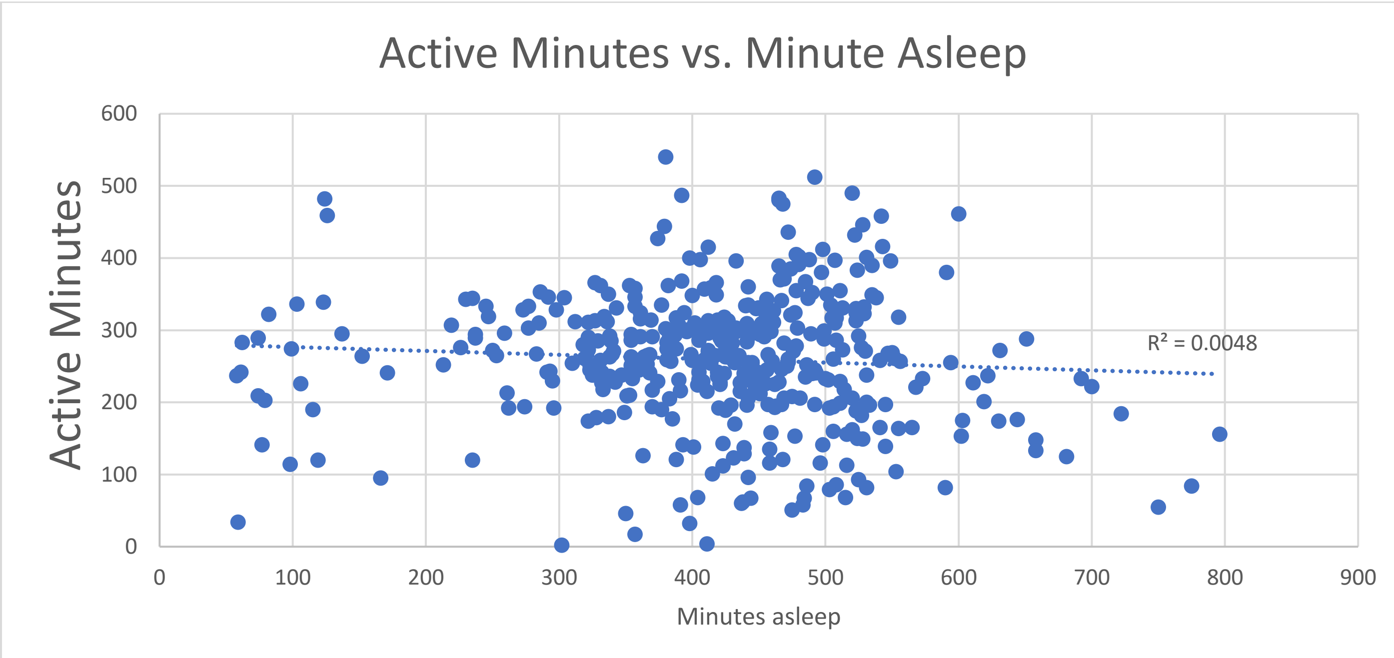
The total activeness of the recorded users falls within the first two quartiles.

ID	Total Steps	Percentile Among Peers
8877689391	23186	0.00%
6962181067	10199	28.13%
4319703577	7753	56.25%
4558609924	5135	71.88%

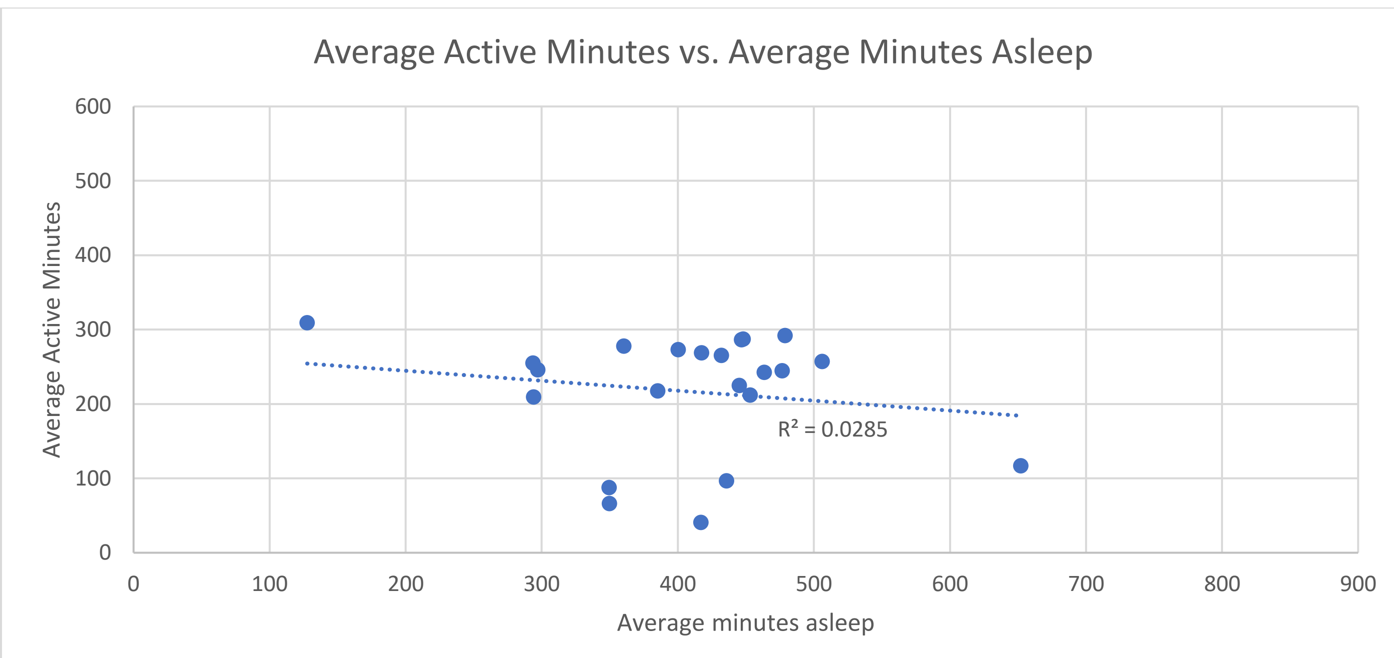
ID	Total Distances	Percentile Among Peers
8877689391	20.39999962	0.00%
6962181067	6.739999771	37.50%
4319703577	5.199999809	62.50%
4558609924	3.390000105	78.13%

The total steps and distances can also be used to supplement the total active minutes table. However, it cannot be relied on solely as some activities cannot be measured by distances (upper body exercises, for example).

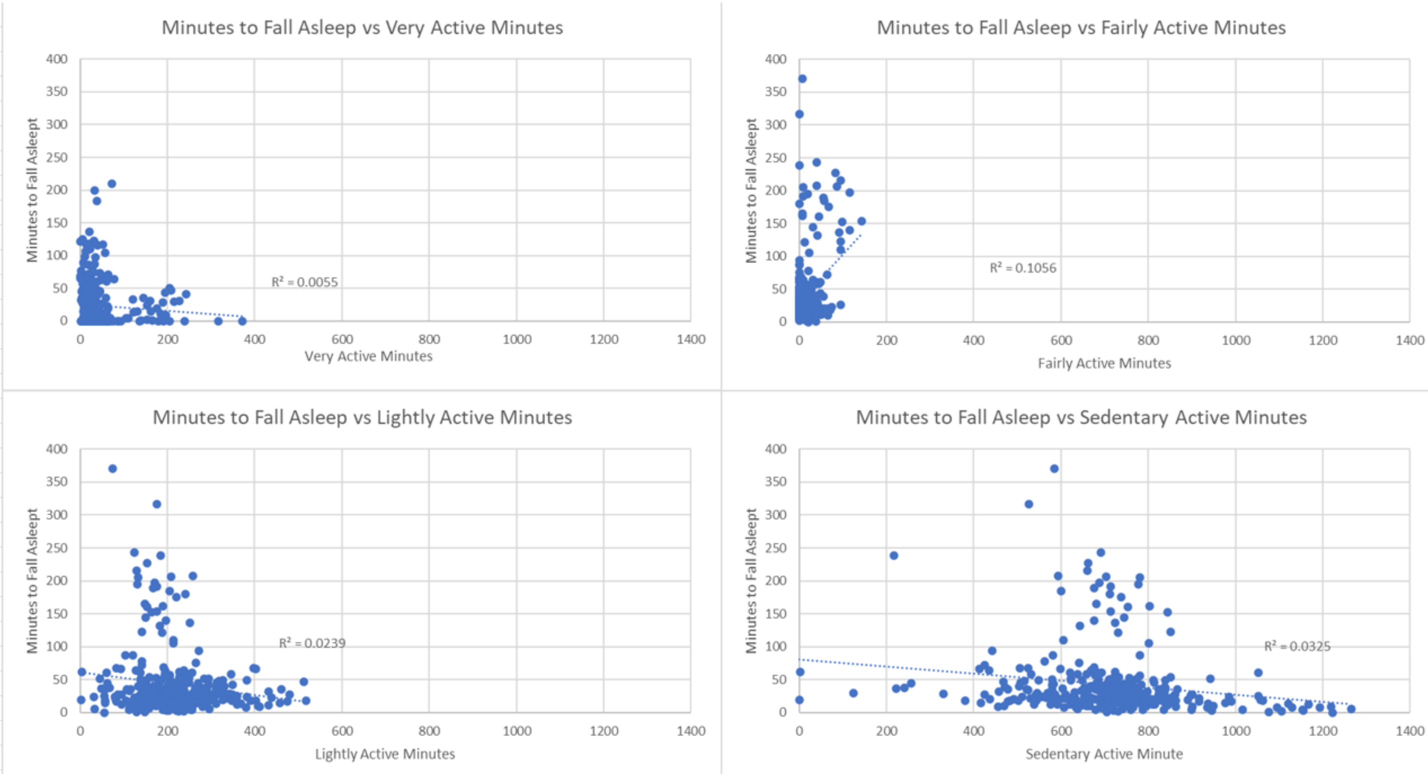
Activeness VS. Sleep



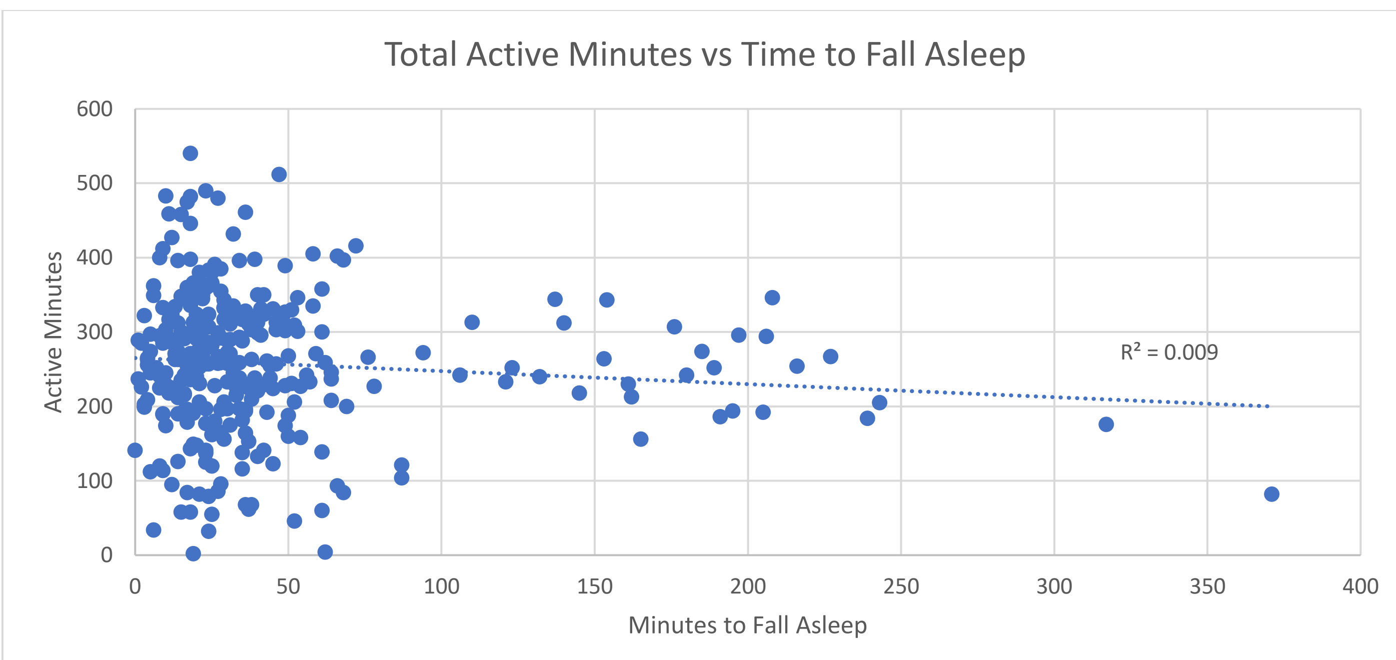
There shows no significant correlation between activeness and minutes you stay asleep without interruption.



To avoid certain days where the users did not record their sleep, the 30 days average minutes are used. Although the correlation is stronger, it is still not a significant.



Being fairly active has the highest correlation to taking longer to fall asleep. Whereas being very active has the weakest correlation with taking longer to fall asleep.



In general, any level of activeness shows weak correlation with time to fall asleep.

Findings

Active minutes vs. Calories burned

- The lack of a strong correlation between activeness and calories burnt suggest that the minutes of different activeness is not based upon the number of calories burned.
- This could be because people are already burning the number of calories recorded through daily activities and not through exercise. Therefore, having the extra bit of exercise time has little correlation to the amount of calories burnt. Only 1 person out of the 33 users burn the average calories (between 1600 – 2200 Kcal) everyday of the month.

Users' Activeness vs. Recommended amount

- All the users are sedentary 81% of the recorded time. This is a cautiously high amount of time, as lack of exercise could lead to numerous negative impact on your health.
- Data indicates that there are no substantial differences of calories burned between the days of the week. The highest difference is 156 Kcal, which is roughly a teaspoon of olive oil.

Activeness vs. Weight loss

- Of the six users that recorded sufficient weight information, four shown to have successfully lost weight and has a active to sedentary ratio of above 30%.
- The finding for this particular direction would've been much more significant if it has more user inputs. As of now, we can only rely on our previous assumption that this sample size is sufficient.

Activeness VS. Sleep

- No strong relationship between minute asleep and minutes being active. With the data on hand we cannot determine if being active has other benefits to sleep time, such as having more active minutes will leads to longer uninterrupted sleep.
- Being very actives has the weakest correlation to taking longer to fall asleep. Where being fairly active has the strongest correlation to taking a long time to fall asleep.
- Being active and users's ability to fall asleep or stay asleep would be difficult to track because:
    - Being active or not is not likely to be a determining factor of users' ability to fall asleep faster
    - There is no reference point, or a "normal" time on when an individual will fall asleep

Recommendation:

1) Bellabeat Time could target users with lower than recommended activeness/calorie expenditure and encourage the users to set their own goals. There are three standards the users can set:

    - User customized: Bellabeat Time can monitor the user's active to sedentary ratio and alert the user when they are not meeting their own goal.
    - Government Recommended: Bellabeat Time can partner with established health care organization, and alert the users when they falling behind the recommended amount of exercise.
    - Peer leaderboard: Bellabeat can host a platform for all Bellabeat Time users to share their progress in comparison to other Bellabeat Time users. In this platform, users can see what other users have achieved and set their goal similar to theirs.

2) If the user's activeness has shown weight loss in the process, Bellabeat can encourage the user to continue on their behavior. If the user's process is falling behind, Bellabeat Time can offer programs, products, or experts advise that matches their work style.
3) To break away from the many limitations due to the lack of data, we should postpone any data driven decision making until enough information has been collected for an analysis of significance.