“There are a huge number of wearable activity trackers on the market, but few of those are directed specifically at women, an issue Bellabeat aims to fix with the Leaf. The Leaf is an activity tracker that looks more like a stylish piece of jewelry than a standard tracker from a company like Jawbone, Nike, or Fitbit, disguising itself as a "leaf" accessory able to be worn on the wrist, collar, or neck.

Like many activity trackers on the market, the Leaf tracks steps taken and sleep quality, but when paired with the app using Bluetooth, it also tracks menstruation and ovulation and it offers guided breathing exercises to reduce stress when worn around the neck, setting it apart from other fitness offerings.”

-[Bellabeat Leaf Activity Tracker Review - MacRumors](https://www.macrumors.com/review/bellabeat-leaf-activity-tracker/)

**Introduction and Company Profile**

Hello, my name is Kevin, and for the purposes of this case study I’m a Junior Data Analyst in the marketing department for Bellabeat. Our company is a high-tech manufacturer of health-focused products for women that was founded in 2013 by Urska Srsen and Sando Mur.

**Summary of the Business Task and Stakeholders**

**The Business Task**

The aim of this case study is to analyze smart device usage data from a competitor to gain insight into how consumers use non-Bellabeat smart devices and how those insights could be applied to improve marketing strategies for comparable Bellabeat products.

Specifically, the questions being asked are:

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

**Key Stakeholders:**

* Urška Sršen: Bellabeat’s cofounder and Chief Creative Officer
* Sando Mur: Mathematician and Bellabeat’s cofounder; key member of the Bellabeat executive team
* Bellabeat marketing analytics team

**Examining the Data Profile**

**Data availability and License**

The dataset made available for this study can be accessed [here](https://www.kaggle.com/datasets/arashnic/fitbit/download?datasetVersionNumber=1) under this [license](https://creativecommons.org/publicdomain/zero/1.0/) and appears to be comprised of mainly quantitative, structured data in 18 tables with varying degrees of granularity regarding the period length that is under examination. The dataset is from Fitbit, who had a 22% market share on unit shipments during the timeframe of the study in 2016 outselling Apple and Garmin devices combined (according to this article from [CNET.com](https://www.cnet.com/tech/mobile/fitbit-sold-more-wearables-in-2016-than-apple-and-samsung-combined/)).

**Overview Description of The Dataset**

The available tables have data covering many aspects of health and activity. Areas of focus consist of activity levels and duration including step counting, distance tracking, and calorie usage. Also, there is data regarding heartrates, sleep tracking, and weight logging. Further calculations were made in the dataset using metrics to determine levels of activity intensity.

The table that is the most readily usable in its raw form for the purposes of this project is the daily\_activity table. There are narrowly focused subsequent tables that contain identical data that were possibly coalesced to form the component columns of the daily\_activity table. All the data that is contained in the daily\_calories, daily\_intensities, and daily\_steps tables is repeated from/to the daily\_activites table. The data includes metrics regarding ID, date, step counts, distances, activity intensity, and calories burned. There are other tables measuring similar fitness and activity metrics based on varying time scales ranging from hours on down to seconds.

**Limitations of The Dataset**

One of the features of the Bellabeat Leaf (an activity tracking device) and its accompanying app is that it has guided breathing exercises to control stress levels. After examining the provided dataset there is nothing in the data that can be used to analyze user stress.

At the time of this study’s publication, the LEAF doesn’t currently appear to make weight tracking a priority for the user experience. With that in mind, as well as viewing how sparse the data in the table is, analyzation of the weight\_log table for the sake of its fitness activity is of limited value unless Bellabeat plans on adding a weight-focused feature at a later date. However, given the sparsity of the provided weight\_log table it would be wise to collect or acquire a more comprehensive dataset that provides a higher level of data value on that topic.

Perhaps a better perspective on the lack of utilization of the weight\_log data by the user is just that. Why do the users by-and-large seem to find little-to-no value in that feature? Why might it be so underutilized? Unfortunately, there is insufficient data from which to gain insight on this matter utilizing this dataset alone.

The study period in the dataset is only one month. A longer timeframe would be beneficial to see if user’s trends and activity consistency hold up over time or show any signs of cyclicality that could be explained by changes such as seasonality or maintaining activity interest levels. Additionally, the overall data pool is not very large with a sample size of roughly 30 users for most tables.

Another issue is that we don’t have any information on the profiles of the individual users being tracked. Since Fitbit products and marketing strategies appear to be gender-neutral, it would be safest to assume that the Fitbit user pool is comprised of mixed genders which is potentially antithetical to the Bellabeat mission of providing health tracking devices specifically for women. If there exist any trends that are more specific to female users, those insights cannot be highlighted by this data in its current form. Therefore, the analyzation of the activity trends of the Fitbit users will be of a more general manner, and a forward assumption was made to treat the data as genderless.

Keeping the gender assumption of the dataset in mind, one of the features of the LEAF is that it helps with tracking user menstruation and ovulation. There is nothing in the provided dataset that can be used to mimic or validate the function of that feature in even a rudimentary sense.

An additional issue to consider with the user profiles is that there is no assumption for age or level of physical conditioning. It may or may not be safe to assume that a condition of inclusion in the user test pool is that all users are of at least legal adult age. Also, without some metric regarding individual user level of physical conditioning, some potential desirable insights would lack context and therefore be of little or no value to this study. For example, someone who is young and has a physical condition capable of running a marathon would have a different activity intensity reaction level to walking up a flight of stairs in comparison to an elderly obese man that maintains a predominantly sedentary lifestyle. With that in mind, there is no clear definition for how the intensity values were calculated throughout the various timeframes. Additionally, at initial glance, the METs figures in the minute-related timeframe do not appear to corroborate the corresponding intensity values.

According to the [database dictionary](https://www.fitabase.com/media/1546/fitabasedatadictionary.pdf) provided by Fitbit (Fitabase):

Intensity

Description: Time spent in one of four intensity categories.

Note: The cut points for intensity classifications and METs are not determined by Fitabase, but by proprietary algorithms from Fitbit.

Upon examination, the minute-related Intensity values are either “0” or “1” regardless of the METs value, giving only two intensity-level classification possibilities.

Given that information, it was determined that the Intensity or METs data will not be useful for the purposes of this study.

**Cleaning and Manipulation of Data**

**Overview of the Cleaning Process**

The tools used during the cleaning process and data manipulation for analysis was VSCode accessing a locally installed Microsoft SQL Server Management Studio Express Server utilizing the SQL database language.

**Steps Taken During the Cleaning Process**

* Renamed all tables to maintain consistency and standardization
* Explored database tables to determine which tables would be most useful
* Checked each table schema
  + Updated and converted column datatypes from default import as varchar(50)
* Created minute\_activity table from joining all minute-timeframe tables
* Created hourly\_activity table from joining all hour-timeframe tables
* Established metadata info
  + Count of users for each table
  + Earliest date in each table
  + Latest date in each table
  + Used date math to calculate difference in start/end dates
  + Counted each distinct date record for comparison with date math
* Checked for record completion/saturation using two tables to determine level of user participation across device functionality
* Checked for duplicate records in each table being used for the study
  + No duplicates found in minute\_activity table
  + No duplicates found in hourly\_activity table
  + No duplicates found in daily\_activity
  + 3 duplicates found in daily\_sleep
    - Created new table with DISTINCT info from daily\_sleep
    - Deleted original daily\_sleep table
    - Renamed new table to replace daily\_sleep after deletion
  + No duplicates found in hourly\_intensity
  + 543 duplicates found in minute\_sleep
    - Created new table with DISTINCT info from minute\_sleep
    - Deleted original minute\_sleep table
    - Renamed new table to replace minute\_sleep after deletion
  + No ndluplicates found in weight\_log table
* Checked for NULL values in all tables - none were found
* Checked the daily\_activity table for various issues/inconsistencies
  + There were no NULL values found in the daily\_activity table
  + Deleted records from daily\_activity where sedentary\_minutes = 1440 (24hours)
  + Deleted records from daily\_activity where TotalSteps > 0, but all distance values = 0
  + Deleted records from daily\_activity where TotalSteps = 0
* Checked the character length of the Id column in all tables to be used
  + Id column in all tables matched with 10 characters
* Checked for outliers in all tables
  + Both the MIN and the MAX for all tables appear to be legitimate or explainable
* Validated the dataset
  + Double checked for accuracy due to any changes made
  + Double checked that all data was formatted correctly

**Analysis and Discussion**

**Overview of the Analysis Process**

How to determine how often the users are wearing their devices? Day vs sleep?

The initial question posed regarding trends in smart device usage led to an examination of the measurable features available from the Fitbit data and how they appear to be used by the members of the test pool. A series of metrics can be extrapolated from those features for the purposes of objective analyzation for marketing while keeping subjective metrics in mind for the customer experience and which available features might be most likely to have influence by leading to a potential change in activity or behavior by the individual user.

Generally, summary statistics were calculated using different options for each metric based on various available time frames.

**Daily Active Users and Churn Rate**

**How many participants are interacting with, or utilizing, their fitness trackers daily? Does the data show any indication of churn rate?**

Out of the 33 total participants in the study, 21 of them (or 67%) wore their devices in some fashion every day over the study period. Out of that group, only 3 wore their devices every night of the study while they slept. And further still, 1 participant measured their weight for 30 days of the study with another participant coming in second place with 24 days.

Unfortunately, since we aren’t furnished with the dates that participants first started using their fitness trackers, as well as the duration of the study period being simply too short, it isn’t possible to determine whether the frequency of participants that did not wear their fitness trackers consistently is a marker of churn rate (a measure of the rate at which users stop using their devices). A dataset of a longer duration and preferably one that included participant feedback would need to be acquired before a definitive answer on that topic was possible.

**Feature Usage and Device Interoperability**

**Which features of the wearable devices are most frequently used by participants? Are participants integrating their wearable devices with other smart devices or platforms (e.g., fitness apps, smart scales, smartwatches)?**

By and large, most of the data that was collected was done so in a passive manner for the participants, meaning all they had to do was to wear their fitness tracker and the data was collected automatically for them and then synced with the accompanying app.

**Session Duration**

**What's the average duration of each session users spend interacting with their devices?**

This metric provides insights into user engagement levels.

**What day of the week are users most/least active?**

**What hour of the day are users most/least active?**

**How much sleep do the users get and how long are they awake in bed?**

What is the overall AVG Hours Asleep for the user group (6.99 hours)?

What is the overall AVG Mins Awake for the user group (39.31 mins)?

**What day of the week do the users get the most/least sleep?**

**What day of the week are users most likely to weigh themselves?**

**How many users gained/lost weight during the study period?**

**Average Steps per day and Duration of Time Being Active**

The most common methods by which the health community uses to categorize levels of fitness are usually based on the average number of steps taken over a specified time duration or the average amount of time that an individual spends at various activity levels.

The CDC and the NIH state that the average number of daily steps taken by a US citizen is 3000-4000, which is approximately equal to 1.5-2 miles.

A peer-reviewed [study](https://ijbnpa.biomedcentral.com/articles/10.1186/1479-5868-8-79) from 2011 published through [BMC](https://www.biomedcentral.com/about) cited that a correlation was observed between lower BMI levels and higher average daily step counts. Simply stated, people who have higher average step counts are generally healthier.

[Viz showing percentages in activity groups by STEPS]

[Viz showing percentages in activity groups by TIME]

<><><> I’m trying to figure out a way to show what percentage of the user group is achieving active time “goals” as recommended by the CDC and the NIH

Roughly 30 minutes per day of moderate activity

What are the most important metrics? CDC Recommendations?

AVG daily steps grouped by user type

AVG daily active minutes grouped by user type

Cdc reco is AVG 150mins/wk = 21.5mins/day

AVG daily steps

AVG daily active minutes

AVG daily calories burned

AVG daily sleep time

There are 1440 minutes in 1 day

**Correlations Between Metrics**

Steps vs sleep

Time active vs sleep

Time active vs steps

Calories vs steps

Calories vs time

The “low-hanging fruit” of this dataset is to look specifically at the various activity levels of the user group. However, from a marketing standpoint it’s potentially more useful to examine the manner in which the user group members interacted with their fitness tracking devices.

**Top Recommendations based on Analysis**

Add goal tracking functionality option to the Bellbeat app, such as weight, or some other fitness related metric (couch to 5k?)

App could offer option of “morning briefing report” that showed the previous days stats to help with daily and long-term goal setting/tracking.

As previously stated in the analysis summary, weight logging doesn’t seem to be a high priority for the Fitbit users in the test pool. If this was a feature that Bellabeat would consider adding to in the future, it would be prudent to acquire additional data beyond what is provided by this study due to the anemic representation of the weightLogInfo table. However, the fact that the user test pool displayed such limited participation in this study could indicate that the feature is not something that the Bellabeat users would embrace. The ROI for the effort may not be worth the undertaking, but further study would be required to properly answer the question. If Bellabeat were potentially interested in the answer, a follow-up question with a more limited scope might be how users would respond to the weight logging feature if they were issued a reminder through the device app. Perhaps the participation numbers were so low because the other data is collected automatically, and weight logging requires the user to remember and take action.

Create a new product as a connected smart scale so that users would be more likely to log their weight with a reminder prompt from the app.

OR partner with a company with an existing smart-enabled product that could be made compatible with the Bellbeat app.

The Bellbeat app could have active reminders based on activity tracking, such as alarms for consistent sleep habits, or notifications after either a set or customizable duration of recoded sedentary activity.

Launch education campaign about health-related topics, such as benefits of reaching daily step goals.

Push the message that Bellabeat products have 6-month battery life and don’t require daily charging.

Also, that they are elegant style and fashion statements that can be worn for almost any occasion.

Also, that Bellbeat products can potentially be worn more comfortably

Create a health/fitness leaderboard where users can connect with and encourage each other.

AVG of total dailyActivity by day of the week

AVG total distance in km per user overall

AVG total distance in km overall for test pool per day

AVG total distance in km per day of the week – Handled in the big AVG query

AVG total distance in km per day of the week per user

AVG Calorie burn per day of the week

AVG Calorie burn per day of the week per user

AVG Active distance by (corresponding) Active minutes (and the converse)

* The way that I wrote this query it eliminates any field that has a 0. It treats each 0 as an outlier. Comparing the results of including or eliminating the 0’s, the averages are affected by almost half if the 0’s are allowed to remain.

What day had the highest activity? Lowest?