**Bellabeat Playing it Smart: Marketing Insights and Recommendations**

Laura Heffner, Junior Data Analyst

October 2024

**PROBLEM**

How should Bellabeat market their smart fitness products to female customers to improve sales among their smart devices and be able to take advantage of their position to compete against the larger players in the industry?

**BUSINESS TASK**

Analyze data about how people (specifically women) have and are using their non-Bellabeat smart fitness devices, identify the trends and insights from that data, and apply this to a Bellabeat product (the Leaf) to create high-level recommendations for a marketing strategy that will allow Bellabeat to compete with the larger players in the industry, and deliver these findings to the executive and marketing analytics teams.

**KEY STAKEHOLDERS**

* Urska Srsen (cofounder and CCO, executive team member)
* Sando Mur (cofounder, executive team member)
* Marketing analytics team

**ABOUT THE DATA**

* Fitbit Fitness Tracker Data (from Mobius, downloaded from Kaggle, Public dataset – CCO confirmed permission for public distribution and use).
* Data is now stored on my computer as well as uploaded to Rstudio in 18 CSV files. The majority of these files are in long format, with some reproduced in wide format as well.
* 30 Fitbit users agreed to have their data tracked – this is a relatively small sample that may not be representative of Bellabeat or Fitbit users. There is no demographic data about the users, so this may result in sampling bias.
* Bellabeat is aimed at women, which may cause results to differ when using data from trackers used by men and women – they may have different preferences on tracking, different exercise routines, and menstrual cycle tracking data is not included.
* Data was collected between 3/12/2016-5/12/2016. This is a fairly short data collection window, and may not reflect trends seen during other seasons of the year.
* Data includes activity values for calories, intensities, steps, heartrate, and sleep. Most of this data can be viewed by the day, hour, and minute.
* Reliability: fair, the data usability is rated at 8.75 on Kaggle, but comes from a third-party source
* Original: this data appears to be original, however, as third-party source, it is impossible to guarantee.
* Comprehensive: no, this data is only representative of 30 individuals and lacks any information about demographics of these individuals to make a determination on whether the data is an unbiased representation. Additionally, the data only shows two months of activity, leaving questions about most of the year and participant’s longer-term activity.
* Current: no, this data is from 2016, now 8 years old.
* Cited: yes, comes from Mobius as a third-party provider.
* Data integrity is questionable based on these findings, however, attempts to search for additional public datasets regarding this topic were fruitless. Future research and analysis are required for a more comprehensive look into fitness tracker use.
* Upon exploration of the data on Rstudio, the data appears complete and organized without missing or null values being found. I believe the data provided to be overall accurate and reliable.
* This data can help answer the business task, although additional data would have been helpful. Specifically, data pertaining to fitness tracker use by women, data with demographic information about participants, data with a larger population sample, and a more current dataset. Looking at the trends present may give insights into how Bellabeat can focus their marketing strategy, or where to focus further research.

**DATA EXPLORATION**

* Attempts to upload data to Excel and BigQuery resulted in errors due to the size of some of the files, and thus, data was processed for analysis in Rstudio.
* Initial exploration of the data showed limited participation by users in sleep, heartrate, and weight log information datasets, resulting in potentially problematic data biases. For this reason, I elected to not make them the focus of this analysis.
* To gain the greatest level of detail from which to draw insights, I focused on the minute data, specifically those datasets that focused on calories, intensities, and steps. Daily and hourly data largely had the same information summarized, and analysis of these was deemed unnecessary. I used the long format datasets for analysis.

**DATA CLEANING**

* The complete cleaning process is documented on the R script for easy replication. Here is a summary of the steps taken:
* Installed packages and loaded libraries for tidyverse, skimr, readr, dplyr, tidyr, janitor, lubridate, and ggplot2.
* Used glimpse() to get a sense of each data frame (columns, variable types, number of observations, etc.).
* Changed Id and ActivityMinute data types (dbl and chr to numeric and datetime types, respectively).
  + minutecalories1$Id <- as.numeric(minutecalories1$Id)
  + minuteintensities1$Id <- as.numeric(minuteintensities1$Id)
  + minutesteps1$Id <- as.numeric(minutesteps1$Id)
  + minutecalories1$ActivityMinute <- mdy\_hms(minutecalories1$ActivityMinute)
  + minuteintensities1$ActivityMinute <- mdy\_hms(minuteintensities1$ActivityMinute)
  + minutesteps1$ActivityMinute <- mdy\_hms(minutesteps1$ActivityMinute)
* Confirmed this change with class().
* Ordered the data frames by Id and then ActivityMinute in ascending order using arrange().
  + minutecalories1 <- arrange(minutecalories1, Id, ActivityMinute)
  + minuteintensities1 <- arrange(minuteintensities1, Id, ActivityMinute)
  + minutesteps1 <- arrange(minutesteps1, Id, ActivityMinute)
* Used clean\_names() to ensure all columns are labeled in a uniform manner.
* Used sum(duplicated()) to determine if any data was duplicated. There were no duplicates found.
* Used filter(!complete.cases()) to see if there were any blanks in the data. There were no blanks found.
* Used unique() to determine every different Id in each data frame. Each had the same 30 different Id values. Given that there are the same number of unique Id values and the same number of observations in each data frame, I made the assumption that ActivityMinute values were also the same for each data frame (meaning each had a recording for every minute between 3/12/2016 and 5/12/2016).
* Used bind\_rows() to merge all three minute data frames to one, creating a new data frame with Id, ActivityMinute, Calories, Intensity, and Steps as columns. The data merged, but showed up as a data frame containing all three datasets, inserting “NA” where there was no data in previous data frames for any given variable. This data had to be condensed so there were not duplicate Id and ActivityMinute entries and by doing so, all “NA” values were removed. This was done with merge() using more complex code generated with the help of ChatGPT.
  + merged\_mindata <- bind\_rows(minutecalories1, minuteintensities1, minutesteps1)
  + merged\_mindata <- merge(minutecalories1, minuteintensities1, minutesteps1, by = "Id")
  + merged\_calint <- merge(minutecalories1, minuteintensities1, by = c("Id", "ActivityMinute"))
  + merged\_min <- merge(merged\_calint, minutesteps1, by = c("Id", "ActivityMinute"))
* View() was used to ensure the data frame was properly formatted and cleaned.
* Next, I used arrange() to order the data by Id, then ActivityMinute in ascending order.
  + merged\_min <- merged\_min %>%

arrange(Id, ActivityMinute)

* Used rename() to change “ActivityMinute” to “Activity Minute”. No other column names required editing. View() ensured this change as well as previous changes took place correctly.
  + merged\_min <- merged\_min %>%

rename(`Activity Minute` = ActivityMinute)

* Used sum(is.na()) to ensure all “NA” values had been removed and replaced with the appropriate data values. The resulting output indicated that this was the case.
* The data up to this point has been explored, datasets were merged, variables were formatted consistently, data types were corrected, confusing labels were corrected, null values and blanks were accounted for and corrected, duplicates were checked for and removed as necessary, incomplete data was checked for, data was organized and ordered, and the data was now ready for analysis.

**ANALYSIS**

* In my analysis, my first step was to determine what if any correlations existed between the different variables.
* I created a correlation matrix between calories, intensity, and steps using select() and cor(). Strong positive correlations were found between all three variables.
  + numeric\_cor\_data <- merged\_min %>%

select(Calories, Intensity, Steps)

* + num\_cor\_matrix <- cor(numeric\_cor\_data, use = "complete.obs")
  + print(num\_cor\_matrix)
* I installed the corrplot package, and created a visual of the correlation matrix from code provided by ChatGPT.
  + corrplot(num\_cor\_matrix, method = "circle", title = "Correlation Matrix of FitBit Tracking Data",

subtitle = "Relationship between Calories, Intensity, and Steps",

mar = c(0, 0, 3, 0))

* Seeing the results made me think, the number of calories burned is the leading goal of many people with fitness trackers. Even when users buy a fitness tracker to measure how many steps they take or how high their heartrate is, they want to see this data because ultimately, they want to be fit and healthy, and in order to do this, users must burn calories. The number of steps or intensity of exercise correlates with how many calories are burned. I wanted to see if there was a significant difference in the calories burned while exercising at various levels of intensity versus at rest.
* Used unique() to determine how many different intensity values existed. 0,1,2, and 3 were the outputs. Quick manual observation of the data timed during the middle of the night showed mostly “0”, leading to the assumption that an intensity of 0 was for sedentary activity, and as the number increased so did the intensity.
* To view the calories burned at each intensity level, I first created Date and Time columns from the Activity Minute column using as.Date() and format(as.POSIXct()). View() ensured this code produced the desired result. The data frame now included Date and Time columns in additional to the original Id, Activity Minute, Calories, Intensity, and Steps columns.
  + merged\_min$Time <- format(as.POSIXct(merged\_min$`Activity Minute`), format = "%H:%M:%S")
  + merged\_min$Date <- as.Date(merged\_min$`Activity Minute`)
  + View(merged\_min)
* Used group\_by() and summarize() to determine the total number of calories burned at each intensity level for the entire dataset and the daily number of calories burned per intensity level, and created a bar chart with ggplot2 to showcase the total number of calories burned at each intensity level for the duration of the data collection.
  + daily\_calories\_by\_intensity <- merged\_min %>%

group\_by(Id, Date, Intensity) %>%

summarize(Total\_Calories = sum(Calories, na.rm = TRUE)) %>%

ungroup()

* + View(daily\_calories\_by\_intensity)
  + total\_calories\_per\_intensity <- daily\_calories\_by\_intensity %>%

group\_by(Intensity) %>%

summarize(Total\_Calories = sum(Total\_Calories))

* + View(total\_calories\_per\_intensity)
  + ggplot(total\_calories\_per\_intensity, aes(x = factor(Intensity), y = Total\_Calories, fill = factor(Intensity))) +

geom\_bar(stat = "identity") +

labs(

title = "Total Calories Burned at Each Intensity Level",

subtitle = "Fitbit Data From March 12, 2016 to May 12, 2016",

x = "Intensity Level",

y = "Total Calories Burned",

fill = "Intensity"

) +

scale\_x\_discrete(labels = c("0" = "Sedentary", "1" = "Light Intensity", "2" = "Medium Intensity", "3" = "High Intensity")) +

scale\_fill\_manual(

values = c("0" = "red", "1" = "orange", "2" = "yellow", "3" = "green"),

labels = c("0" = "Sedentary", "1" = "Light Intensity", "2" = "Medium Intensity", "3" = "High Intensity")

) +

theme\_minimal()+

theme(axis.text.x = element\_text(angle = 45))

* The bar chart shows clearly that sedentary intensity is where the most calories were burned, followed by light intensity, then high intensity and finally medium intensity. The majority of calories are burned while at rest. This has significant indications for an individual’s health status, and can be useful if tracked with the other health markers Bellabeat seeks to track for women. This is why it’s important for individuals to be able to wear their devices as much as possible. A comprehensive picture of an individual’s health can only truly be obtained if the user wears the device as much as possible. This is why wearing a versatile piece of jewelry as a fitness tracker is convenient for women. When they want to look nice, or are in a situation where a clunky fitness watch just isn’t appropriate to wear (work, weddings, dates, etc.), Bellabeat offers a unique answer to this problem that no other device on the market currently offers.
* Again, used group\_by() and summarize() to determine the daily average number of calories burned at each intensity level, then created a line plot using ggplot(). Findings show that the average number of calories burned per day were highest at high intensity, and decreased as intensity also decreased. This is predictable. It is worth noting that as the intensity increased, the consistency over time of calories burned decreased. This variation finding could be key when tracking health and fitness over time.
  + average\_calories\_by\_intensity <- merged\_min %>%

group\_by(Id, Date, Intensity) %>%

summarize(Average\_Calories = mean(Calories, na.rm = TRUE)) %>%

ungroup()

* + View(average\_calories\_by\_intensity)
  + average\_calories\_per\_intensity <- average\_calories\_by\_intensity %>%

group\_by(Intensity, Date) %>%

summarize(Daily\_Average\_Calories = mean(Average\_Calories)) %>%

ungroup()

* + View(average\_calories\_per\_intensity)
  + ggplot(average\_calories\_per\_intensity, aes(x = Date, y = Daily\_Average\_Calories, color = factor(Intensity))) +

geom\_line(size = 1) +

labs(

title = "Average Daily Calories Burned per Intensity Level",

subtitle = "Fitbit Data From March 12, 2016 to May 12, 2016",

x = "Date",

y = "Average Daily Calories Burned",

color = "Intensity"

) +

scale\_color\_manual(

values = c("0" = "yellow", "1" = "orange", "2" = "red", "3" = "purple"),

labels = c("0" = "Sedentary", "1" = "Light Intensity", "2" = "Medium Intensity", "3" = "High Intensity")

) +

theme\_minimal()

* This data shows that indications about an individual’s health can be significant while at rest and while exercising. A fitness tracker was designed for tracking exercise, but Bellabeat’s Leaf can do more than that. It can become a lifestyle tracker that completes the picture for those seeking ideal health and wellbeing.

**KEY FINDINGS**

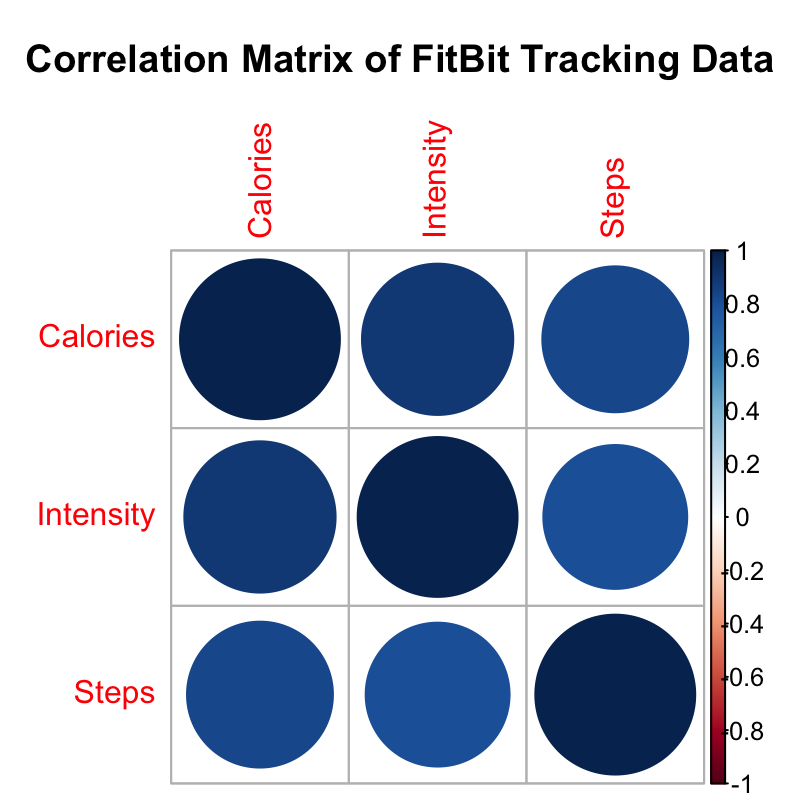


Fig. 1 Heatmap of the strong positive correlations between calories, intensity, and steps of FitBit users from March 12, 2016 to May 12, 2016 (Data Source: Mobius. 2016 May. Fitbit Fitness Tracker Data. V2.).

A strong positive correlation was found between calories burned, intensity of the activity, and steps taken, as shown in Fig, 1. What this means is that there is a very strong likelihood that as the calories increase, the intensity and steps have also increased. Likewise, when steps increase, the calories and the intensity also increase. Finally, when intensity increases, the calories and steps have also increased. This data suggests that the more users track that they are active and exercising, the more they can see how much their hard work is paying off. This is immensely beneficial for the fitness tracker industry, as it creates the positive feedback necessary to keep customers wearing and using the products. Note that for the purpose of this analysis, causation was not determined.

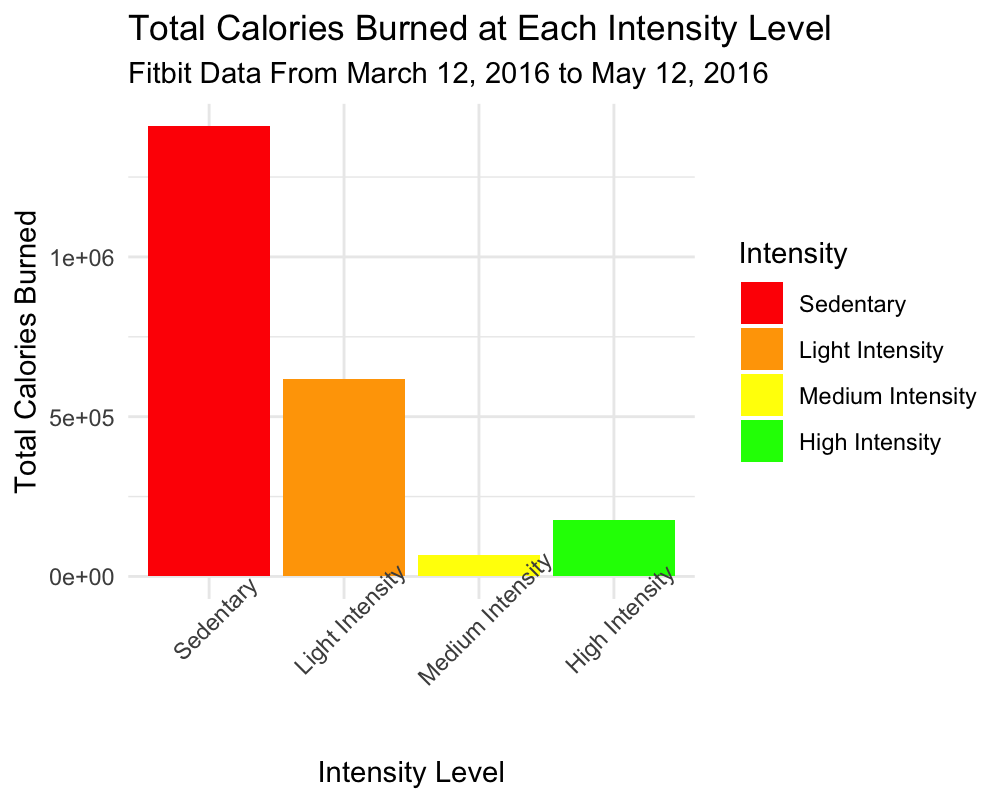


Fig. 2 Bar chart showing the total amount of calories burned at each intensity level by FitBit users from March 12, 2016 to May 12, 2016 (Data Source: Mobius. 2016 May. Fitbit Fitness Tracker Data. V2.).

In Fig. 2, the amount of total calories burned during this period were largely burned during sedentary activity. What this suggests is that over time, the majority of our calories are burned while at rest rather than while exercising. This has huge implications for fitness tracker users, as we can see that there is value to tracking activity even when individuals are not being particularly active. In order for individuals to gain a more complete picture of their health, it’s important to track this information at all times on a continuous basis.

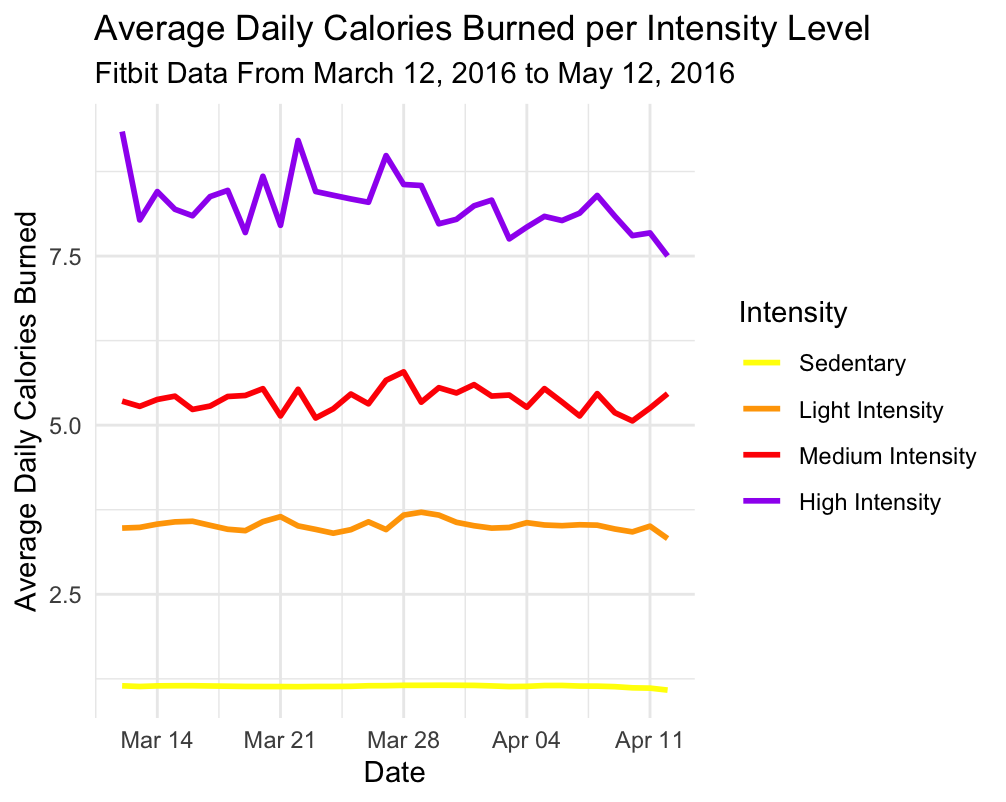


Fig. 3 Line graph showing the average daily calories burned per intensity level for Fitbit users from March 12, 2016 to May 12, 2016 (Data Source: Mobius. 2016 May. FitBit Fitness Tracker Data. V2.).

Fig. 3 shows that the average daily calories that were burned by participants were highest during high intensity activity, and decreased as intensity levels decreased. This confirms that it is very important to track activity data during exercise to further reinforce continued activity. Additionally, the data over time becomes more varied the higher the intensity becomes. This is another indication that it’s important to view health over time because it varies and changes with time. These fitness tracking devices can tell us so much more than how many steps we took or how many calories we burned in a day. They can provide a thorough picture of an individual’s entire health and well-being over time, something I imagine will become an important tool as we progress with technology to higher levels of ideal health and fitness.

**RECOMMENDATIONS**

My recommendation focuses on the Leaf, a fitness tracker that looks like of a piece of jewelry, and can be worn as a necklace, bracelet, or pin. Bellabeat should focus its efforts on what makes it unique from other fitness tracker device companies in the industry. The versatility of the Leaf should be used to our advantage based on the trends found in this analysis. This tracker can, of course, be worn to the gym, but it is also an appropriate accessory for a wedding, for work, or for a football game. This accessory can be worn in a variety of ways, and is tastefully disguised as jewelry, allowing for easy continuity of wear. Being appropriate to wear continuously, no matter the events of the day, the Leaf allows women to better track their health and well-being, and get the most comprehensive picture of their health. This is at the core of Bellabeat’s dedication and attention to the health and well-being of women. Bellabeat doesn’t just create fitness trackers. We create health trackers.

Based on the findings of this analysis, these features should be emphasized during marketing campaigns to show that wearing a health tracker works best when it’s worn at all times, and that the Bellabeat Leaf is the only product on the market that allows women to do just that and continue living their best healthy lives.

**FUTURE IMPLICATIONS**

Research and data analysis on more current data that provides a wider timeline of data, a larger population sample, demographic data about the participants, and data specifically about women wearing fitness trackers is needed to ensure no unintentional biases occurred during this analysis. No such data was found at this time.

**REFERENCES**

* Mobius. 2016 May. Fitbit Fitness Tracker Data: Pattern Recognition with Tracker Data: Improve Your Overall Health, V2. Retrieved 09/30/3024 from https://www.kaggle.com/datasets/arashnic/fitbit/data