Bellabeat Case Study: Analyzing Smart Device Usage to Inform Product and Marketing Strategy

Greg Montalvo

2025-05-27

1. Ask — Business Task Summary

Bellabeat aims to gain insights into how consumers use non-Bellabeat smart devices, particularly Fitbit, to better understand users behavior and preferences. The primary goals are to:

Identify trends in smart device usage, such as activity levels and calorie burn patterns.

Understand how these trends might apply to Bellabeat's customer base, primarily women interested in health and wellness.

Use these insights to influence Bellabeat's product development and marketing strategies, tailoring features and campaigns to better meet customer needs.

2. Prepare — Data Sources Description

The primary data source is the Fitbit Fitness Tracker Dataset (public domain, available on Kaggle via Mobius), which includes personal fitness tracking data from 30 Fitbit users who consented to share minute-level activity, heart rate, and sleep data. The dataset spans March to May 2016 and includes daily steps, calories burned, and active minutes.

Two CSV files containing overlapping periods were merged for a comprehensive view of activity patterns. While rich, the data lacks demographic variables such as gender, which limits the ability to segment analysis by target customer groups.

3. Process — Data Cleaning and Preparation

The data processing workflow included:

Merging two datasets (daily_activity.csv and daily_activity_v2.csv) into a single dataframe.

Converting ActivityDate fields to Date type for accurate time-based analysis.

Checking for missing values, duplicates, and inconsistent entries (none significant).

Creating a new variable for the month extracted from ActivityDate for monthly aggregation.

Calculating average daily steps and calories burned per month to identify temporal trends.

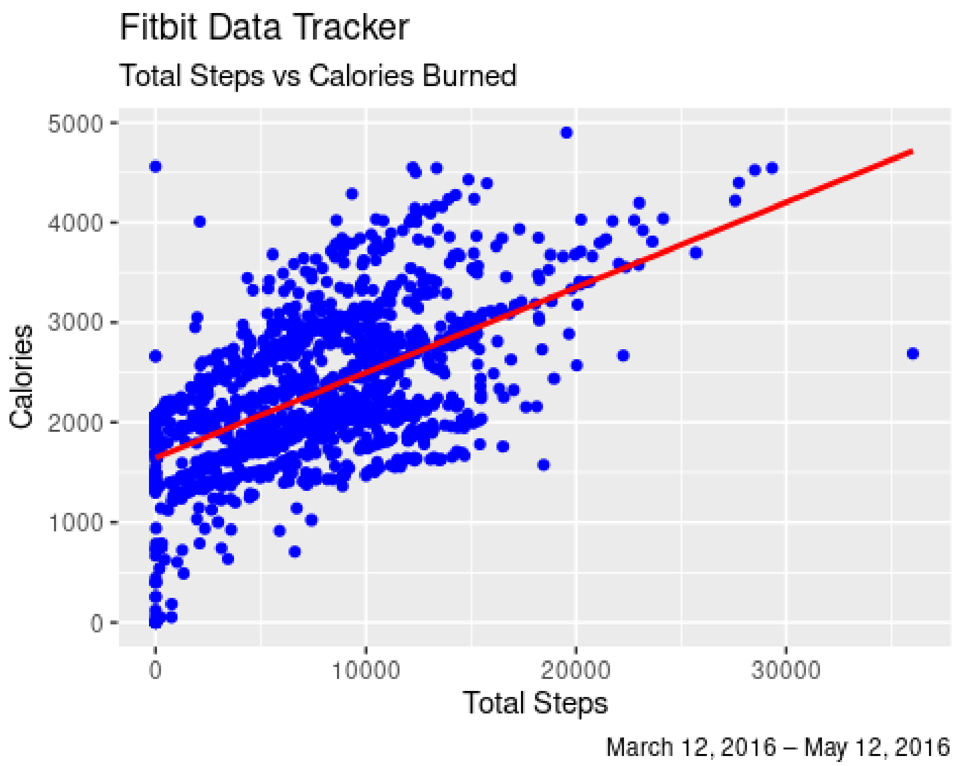
Using R and packages (ggplot2, dplyr, tidyr, janitor) to manipulate and visualize the data, ensuring reproducibility and clarity in the analysis.

4. Results and Analysis

Total Steps vs Calories

```
library(ggplot2)
library(tidyr)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(janitor)

##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test

daily_activity <- read.csv("dailyActivity_merged.csv")

daily_activity_v2 <- read.csv("dailyActivity_merged_v2.csv")

# Convert ActivityDate to Date type for both datasets
daily_activity$ActivityDate <- as.Date(daily_activity$ActivityDate, format =
"%m/%d/%Y")
daily_activity_v2$ActivityDate <- as.Date(daily_activity_v2$ActivityDate,
format = "%m/%d/%Y")

# Combine the datasets
daily_activity_merged <- bind_rows(daily_activity, daily_activity_v2)

# Plot: Total Steps vs Calories Burned
ggplot(daily_activity_merged, aes(x = TotalSteps, y = Calories)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(
    title = "Fitbit Data Tracker",
    subtitle = "Total Steps vs Calories Burned",
    caption = "March 12, 2016 – May 12, 2016",
    x = "Total Steps",
    y = "Calories"
  )
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Results

The red line is a linear regression trend line, showing a positive correlation between the number of steps taken and calories burned.

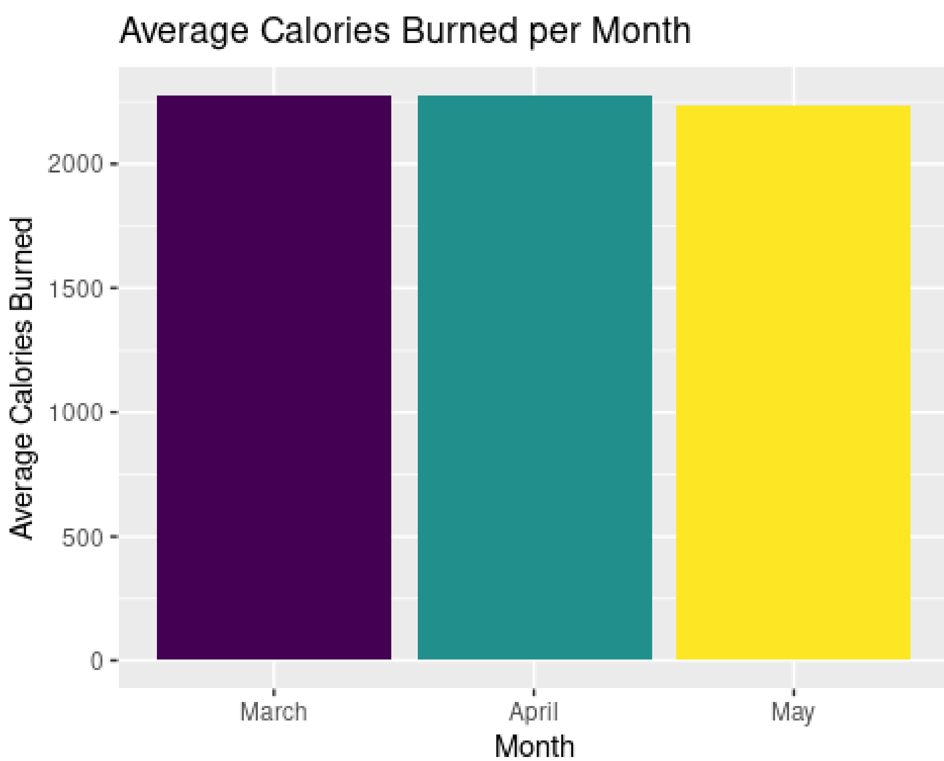
Higher step counts generally lead to higher calorie burn, indicating strong engagement with physical activity tracking.

However, some users logged calories burned without taking steps possibly due to basal metabolic rate (BMR) or other non-step activities.

Average Calories burnt by Months

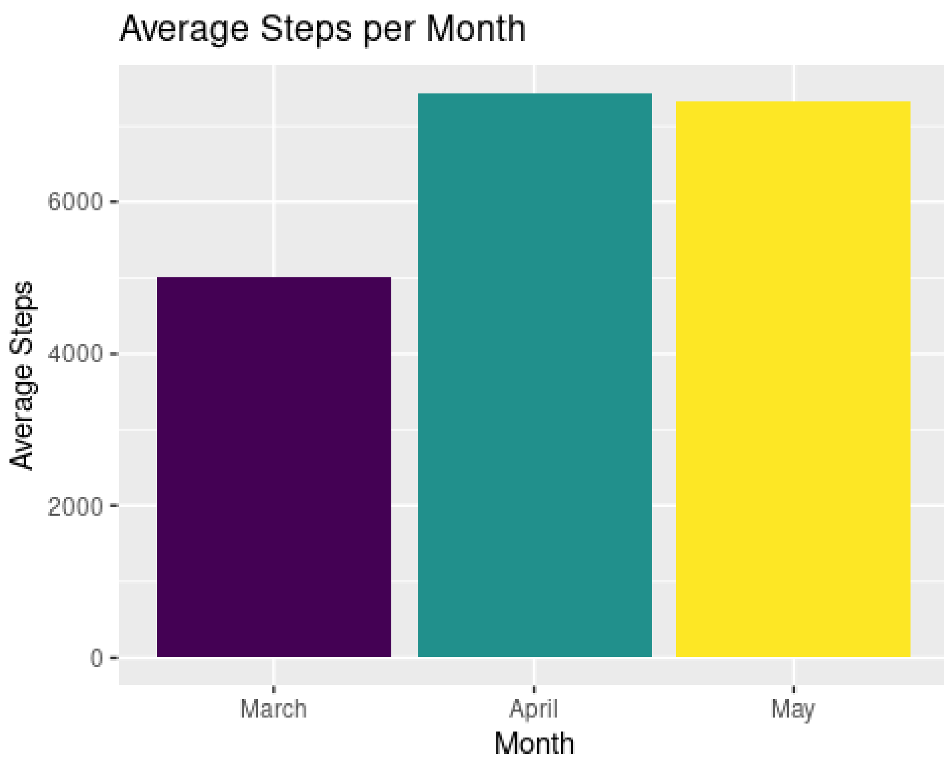
```
daily_activity_merged$month <- format(daily_activity_merged$ActivityDate,
"%B")

month_order <- c("March", "April", "May")

monthly_summary <- daily_activity_merged %>%
     group_by(month) %>%
     summarise(
         avg_steps = mean(TotalSteps),
         avg_calories = mean(Calories)
     )

monthly_summary$month <- factor(monthly_summary$month, levels = month_order,
```

```
ordered = TRUE)


ggplot(monthly_summary, aes(x = month, y = avg_calories, fill = month)) +
  geom_col(show.legend = FALSE) +
  labs(
    title = "Average Calories Burned per Month",
    x = "Month",
    y = "Average Calories Burned"
  )
```



Average Steps by Month

```
ggplot(monthly_summary, aes(x = month, y = avg_steps, fill = month)) +
    geom_col(show.legend = FALSE) +
    labs(
        title = "Average Steps per Month",
        x = "Month",
        y = "Average Steps"
    )
```



Results

March and April had the highest average number of steps, yet did not correspondingly yield the highest calories burned for the month of March.

May had fewer average steps than March, but users burned more calories on average in this month.

This discrepancy suggests that not all steps are equal in terms of energy expenditure. Possible reasons include: Non-step activities like biking, weightlifting, or swimming, which don't significantly increase step counts but do burn substantial calories.

Passive calorie burn from saunas or thermogenic effects.

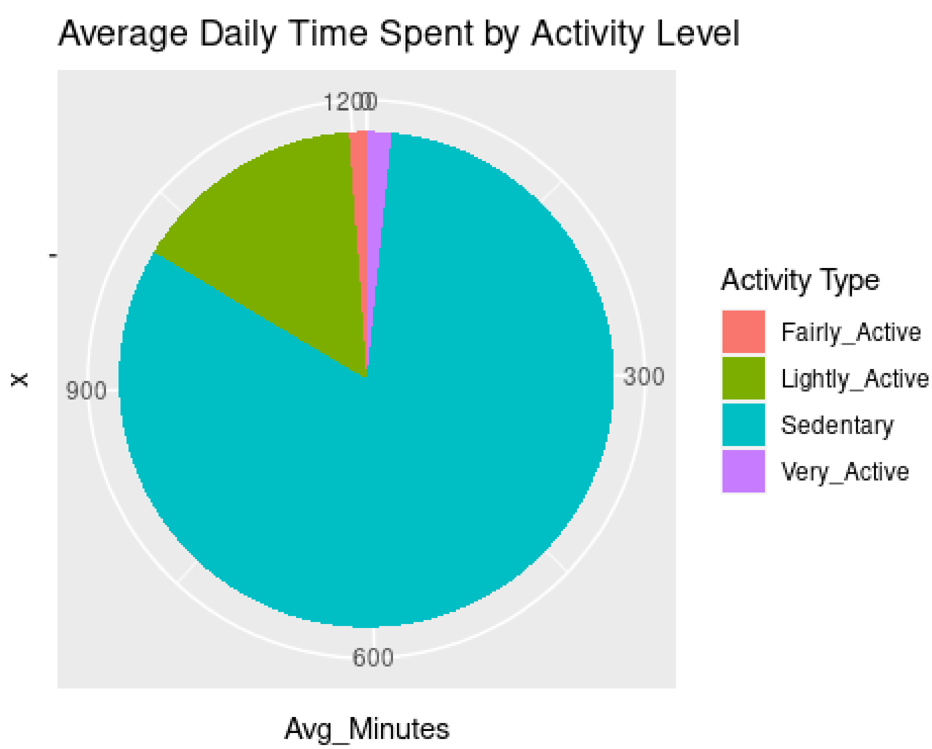
Increased exercise intensity, meaning fewer steps but more effort.

Potential variation in device wear time or activity tracking behavior across months.

Active Minutes Distribution

```
activity_summary <- daily_activity_merged %>%
  summarise(
    Very_Active = mean(VeryActiveMinutes),
    Fairly_Active = mean(FairlyActiveMinutes),
    Lightly_Active = mean(LightlyActiveMinutes),
    Sedentary = mean(SedentaryMinutes)
  ) %>%
  pivot_longer(cols = everything(), names_to = "Activity_Type", values_to =
```

```r
"Avg_Minutes")

ggplot(activity_summary, aes(x = "", y = Avg_Minutes, fill = Activity_Type))
+
    geom_bar(stat = "identity", width = 1) +
    coord_polar(theta = "y") +
    labs(
        title = "Average Daily Time Spent by Activity Level",
        fill = "Activity Type"
    )
```



Results

On average, users spend the majority of their daily minutes in the Sedentary category, followed by Lightly Active, Fairly Active, and then Very Active minutes. This distribution indicates that while users may be wearing their devices consistently, most are not engaging in high-intensity physical activity throughout the day.

For Bellabeat, this insight suggests an opportunity to design features and campaigns that encourage more active lifestyles among users. Strategies such as daily movement reminders, motivational badges for activity streaks, and in-app challenges could help shift users toward more active behaviors.

Limitation: Missing Gender Data

One key limitation of the available Fitbit dataset is the absence of demographic information, particularly gender. Given that Bellabeat primarily targets a female audience, the lack of gender data prevents a more granular analysis of user behavior based on biological or behavioral differences between men and women.

This information could have provided valuable insight into how women, specifically, engage with fitness technology in terms of physical activity, sleep patterns, and calorie expenditure. Future data collection efforts should prioritize the inclusion of gender and other demographic variables such as age, occupation, location, to better align product features and marketing strategies with Bellabeat's target market.

Final Recommendations

Encourage Bellabeat users to log additional activities beyond steps, like cycling or strength training.

Incorporate features that motivate consistent movement throughout the day, like reminders and challenges.

Encourage for more users to log sleep and weight, as these are important for holistic health monitoring.

Collect more detailed demographic data in future to better tailor marketing and product development