Lauren Holstein
November 2023
Capstone Project

How a Wellness Company Can Play It Smart

Case Study

Business Overview

Company Profile	**Bellabeat** is a manufacturer of health-focused technology products for women looking to become a larger player in the global smart-device market.
Stakeholders	Cofounder and Chief Creative Officer: **Urška Sršen** Cofounder and key member of executive team: **Sando Mur** **Bellabeat marketing analytics team**
Products	**Bellabeat app**: health-data dashboard for users **Leaf**: classic wellness tracker (bracelet, necklace, or clip) **Time**: wellness watch and smart tracker **Spring**: smart water bottle **Bellabeat membership**: subscription-based wellness guidance

Project Overview

Business Task	Analyze smart-device usage data and apply insights on how consumers use non-Bellabeat devices to the Bellabeat marketing strategy.
Key Questions	What are the current trends in smart-device usage? How could these trends apply to Bellabeat customers? How could these trends influence the marketing strategy?
Deliverables	Summary of business task Description of all data sources Documentation of any cleaning or manipulation of data Summary of analysis

	Supporting visualizations and key findings High-level content recommendations based on analysis

Data Overview

Dataset	FitBit Fitness Tracker Data
Date	This dataset was generated by respondents to a distributed survey via Amazon Mechanical Turk between 03/12/2016 and 05/12/2016.
Description	Thirty eligible Fitbit users consented to the submission of personal tracker data including minute-level output for physical activity, heart rate, and sleep monitoring.
Permissions	Per the metadata, this dataset is in the public domain. Users are free to copy, modify, and distribute the content, even for commercial purposes.
Location	The dataset was available through Kaggle, shared by user Mobius.
Details	The dataset is stored across 18 CSV files (15 unique datasets plus 3 duplicates in wide format). Each file contains one of the user metrics listed below tracked across either day, hour, minute, or second.

Daily	**Hourly**	**By the Minute**	**By the Second**
Activity Calories burned Intensities Steps Sleep Weight log	Calories burned Intensities Steps	Calories burned Intensities METs (metabolic equivalent of task) Sleep Steps	Heartrate

Data Processing

Choosing Data for Analysis

When choosing tables to include in the analysis, I reasoned that daily user trends would provide the most useful insights for marketing recommendations (compared to the tables that were broken down by hours or minutes).

Lauren Holstein
November 2023
Capstone Project

The **daily activity** table included 33 unique user IDs across 31 days of records and tracked: user ID, activity date, total steps, total distance, tracker distance, logged activities distance, distance by activity level, minutes by activity level, and calories burned.

The **daily sleep** table included 24 unique user IDs across 31 days of records and tracked: user ID, sleep day, total number of sleep records, total minutes asleep, and total time in bed.

The **weight log** table included 8 unique user IDs across 31 days of records and tracked: user ID, date, weight (kg), weight (lbs), fat, BMI, whether the report was manual or automated, and log number.

I knew with the data provided, I should be able to gain insights on:

- How many days during the tracking period participants wore their trackers.
- How long participants wore their tracker each day.
- The amount of activity each participant engaged in per day.
- The impact of wearing a tracker on participants' weight.
- Any relationship between a participant's activity and sleep.

I downloaded my chosen tables in CSV format from Kaggle and uploaded them to Google Sheets.

Data Limitations

Even the largest dataset contained a fairly small sample size of 33 participants with an average of 28.5 records per participant. We don't know the characteristics of the population from which participants were recruited, so we don't know if they are representative of the general population. Furthermore, we do not know the gender of participants; considering Bellabeat is a woman-focused company, we would ideally work with a dataset of all women.

Data Cleaning

Spreadsheet and SQL

1. **Duplicates:** After concatenating the user ID and activity date columns, I applied conditional formatting to check for duplicate records. **There were no duplicates.**
    a. Formula, find duplicates: `=COUNTIF(C:C,C2)>1`
2. **Blank cells:** I used conditional formatting to highlight blank cells. **There were no blank cells.**
3. **Extra spaces:** All of the fields contained numeric data types so there was no need to trim extra spaces.

4. **Errors and outliers:** For each column, I found the MIN and MAX to ensure that all values fit within expected parameters. **Each range fit expectations.**

    a. Note: The maximum daily steps were 36,019 over approximately 7 hours of activity; while generally high, these numbers are realistic for a marathoner or other long-distance athlete.

5. **Removing 0-distance records:** Examining the dataset revealed that some records existed for days on which a user **did not wear their tracker.** These records showed a value of 0 distance tracked.

    I kept one copy of the table with 0-days included to calculate days worn/not worn. I created a duplicate table and removed records showing 0.0 distance (87 out of 940 records)—records that indicated the participant did not wear their tracker at all that day—to calculate activity while the tracker was worn.

6. **Importing to BigQuery:** I planned to join the daily activity spreadsheet with the daily sleep spreadsheet, so I uploaded them both into BigQuery as new tables.

7. **Data type:** I checked the data type of each field. **Each field had the appropriate data type and did not need conversion.**

8. **Standardizing data format:** To join the daily activity and daily sleep spreadsheets, I had to ensure that they both had the same format for the date column.

    The activity spreadsheet contained the date in yyyy-mm-dd format, while the sleep spreadsheet contained yyyy-mm-dd hh:mm:ss. To standardize the two, I needed to isolate the sleep spreadsheet's date and remove the timestamp (which read 00:00:00 for each record, so it could be deleted without meaningful data loss).

    I used the SQL DATE function to isolate only the date from the SleepDay column. I then saved the results as a new table.

```
SELECT

  Id,

  DATE(SleepDay) AS date,

  TotalSleepRecords,

  TotalMinutesAsleep,

  TotalTimeInBed

FROM `case-study-bellabeat-404404.fitbit_tracker.sleep`
```

9. **Joining two tables using SQL:** Before joining the two tables, I ensured that all 24 user IDs in the sleep table existed in the daily activity table; they did.

Once I had standardized the data between my two tables, I used the SQL JOIN function to combine the daily activity and sleep tables on the shared keys of user ID and date.

```
SELECT *
FROM `case-study-bellabeat-404404.fitbit_tracker.daily_activity`
JOIN
  `case-study-bellabeat-404404.fitbit_tracker.sleep`
ON `case-study-bellabeat-404404.fitbit_tracker.daily_activity`.id =
`case-study-bellabeat-404404.fitbit_tracker.sleep`.id
  AND

`case-study-bellabeat-404404.fitbit_tracker.daily_activity`.ActivityDate
= `case-study-bellabeat-404404.fitbit_tracker.sleep`.date
```

10. **Uploading to Google Sheets:** After joining the two tables using SQL, I downloaded the new table as a CSV file and uploaded it to Google Sheets.
11. **Uploading to RStudio:** I uploaded the weight log spreadsheet to RStudio for graphical analysis.

Data Analysis

Number of Days Each Participant Wore Their Tracker

Spreadsheet

First, I wanted to understand **how consistently an average user could be expected to wear their tracker**.

For this dataset, the description indicated that “variation between output represents use of different types of Fitbit trackers and individual tracking behaviors / preferences.” Indeed, not every participant had the same number of records.

To confirm that the number of records for each participant covered their entire tracked time period with one record per day and no missing records, I used the MIN and MAX functions with a nested FILTER function to determine when each user’s tracking period began and ended.

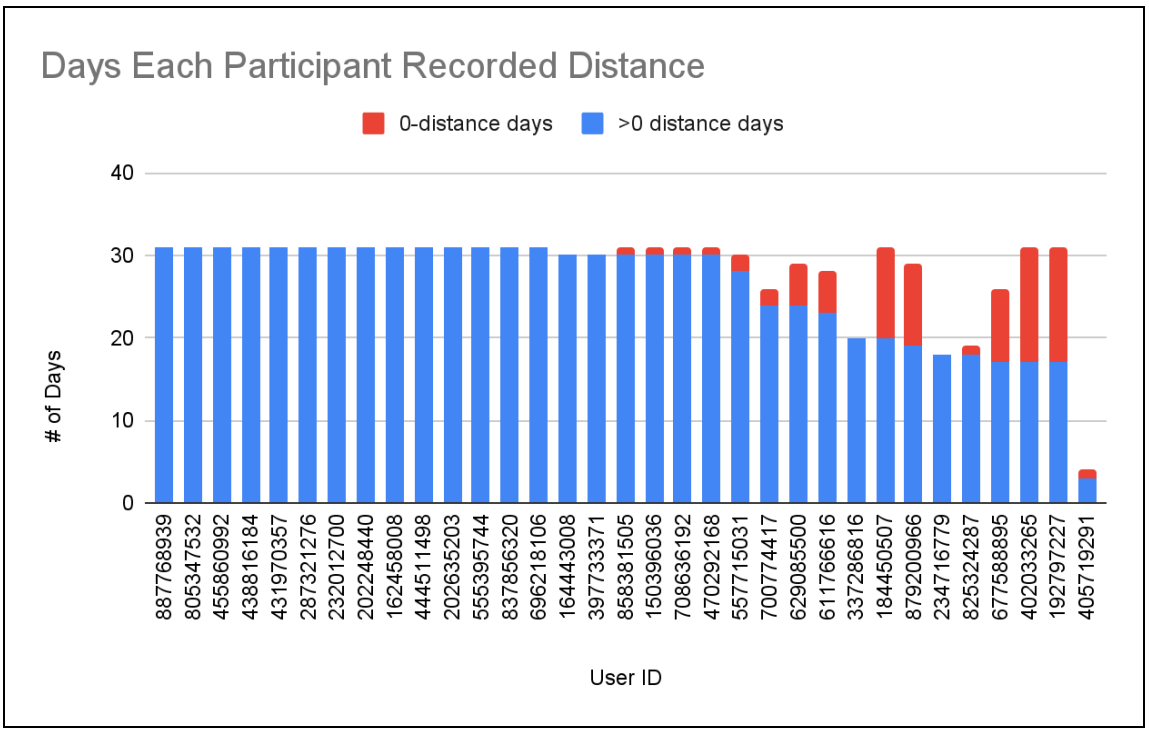
- Formula, start date: `=MIN(FILTER(dailyActivitywNull!B:B,dailyActivitywNull!A:A=A2))`
- Formula, end date: `=MAX(FILTER(dailyActivitywNull!B:B,dailyActivitywNull!A:A=A2))`

Comparing the days in each range with the number of records **confirmed that each user had one record per one day for their entire tracking period.**

Next, I looked at how many days each participant in the dataset recorded any tracker distance. I analyzed the dataset before removing null distance values; **we can assume that on 0-distance days, the tracker was not worn.**

Finally, I calculated the number of 0-distance records out of a user's total number of records as the **percentage of days a user wore the device**.

1. I calculated how many records existed per user, how many days each user logged 0 distance, and how many days each user logged >0 distance.
    a. Formula, list unique user IDs: `=UNIQUE(dailyActivity!A2:A941)`
    b. Formula, # records: `=COUNTIF(dailyActivity!$A$2:$A$941,A2)`
    c. Formula, 0-distance days:
    `=COUNTIFS(dailyActivity!$A$2:$A$941,A2,dailyActivity!$D$2:$D$941,"0")`
    d. Formula, >0 distance days:
    `=COUNTIFS(dailyActivity!$A$2:$A$941,A2,dailyActivity!$D$2:$D$941,">0")`



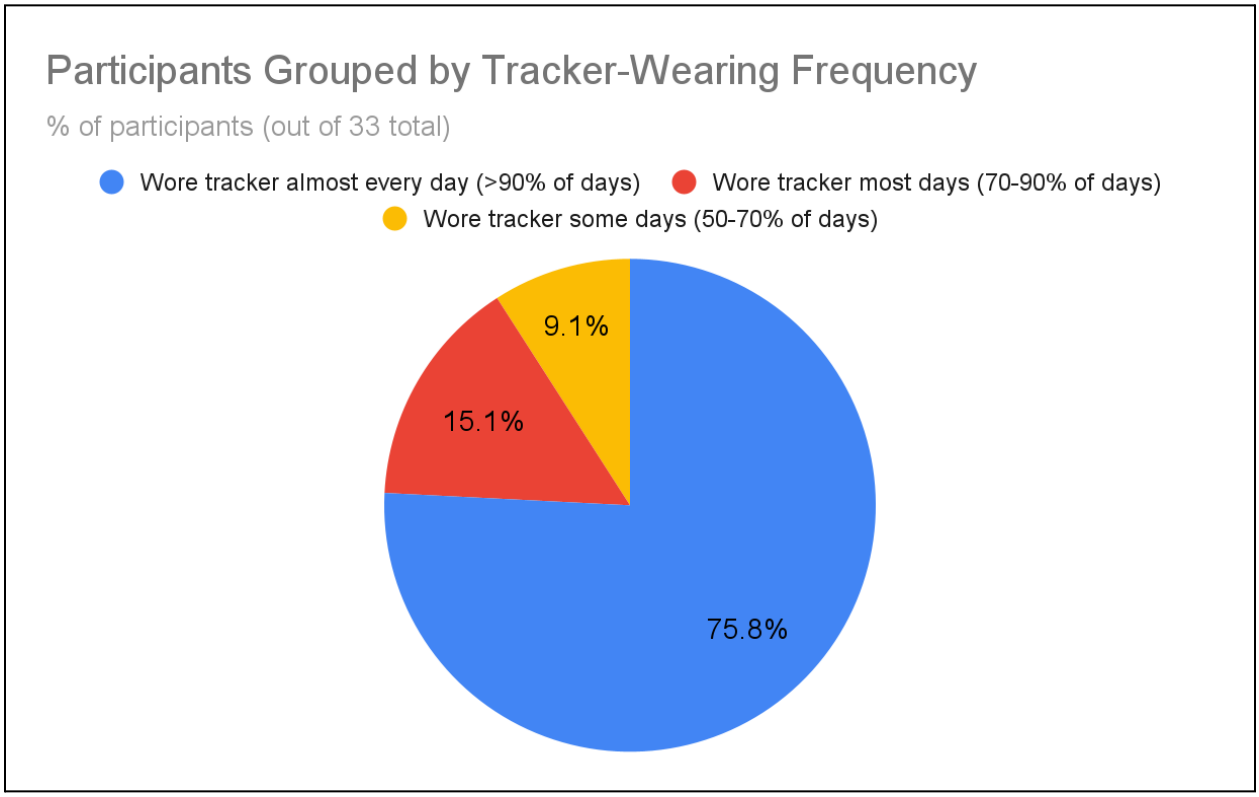
2. Grouped by days worn, the analysis revealed the following:

a. **25 of 33 participants (75.8%) wore their tracker almost every day** (at least 90% of their recorded days).

b. **3 of 33 participants (9.1%) wore their tracker most days** (70-90% of their recorded days).

c. **5 of 33 participants (15.1%) wore their tracker some days** (50-70% of their recorded days).

d. **No participants wore their tracker fewer than half of the days.**



Key takeaway

Three-out-of-four fitness-tracker users wore their device nearly every day.

Time Spent Wearing Tracker

Spreadsheet

Most participants wore their tracker nearly every day, but **how long did they wear it each day?**

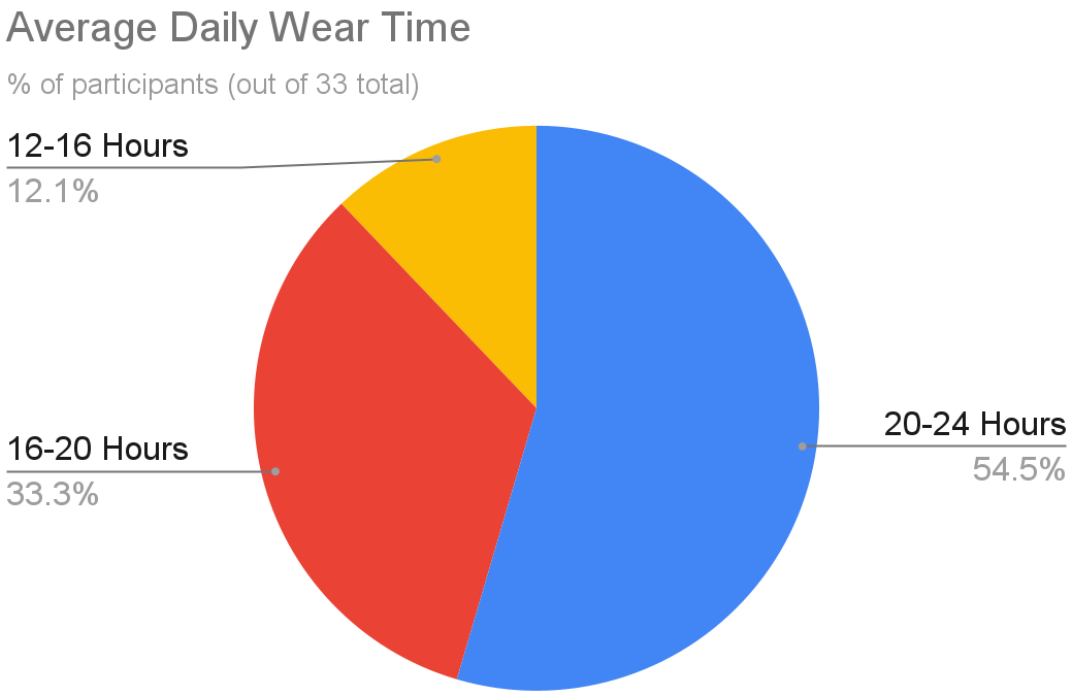
Lauren Holstein
November 2023
Capstone Project

To answer this question, I examined **how much time each participant spent in different activity levels (very active, fairly active, lightly active, and sedentary)** while wearing the tracker.

Since I only wanted data from days when the users were wearing their trackers, I used my table from Step 4 of the data-cleaning process described above, which had null-distance records removed.

1. To analyze the **total duration each user wore their smart device per day**, I created a new column titled **TotalTimeMinutes** calculating the total of the four activity categories.
    a. Formula: =SUM(K2:N2)
2. On days when participants wore their trackers, they wore it for an average of 1,204 minutes, or **20 hours and 4 minutes.**
3. Grouped by wear time, the analysis showed the following:
    a. **18 of 33 (54.5%) participants wore their tracker almost all day** (an average of 20-24 hours per day).
    b. **11 of 33 (33.3%) participants wore their tracker most of the day** (an average of 16-20 hours per day).
    c. **4 of 33 (12.1%) participants wore their tracker about half of the day** (an average of 12-16 hours per day)**.**
    d. **No participants had an average daily wear time of less than half the day.**

Lauren Holstein
November 2023
Capstone Project



Key Takeaway

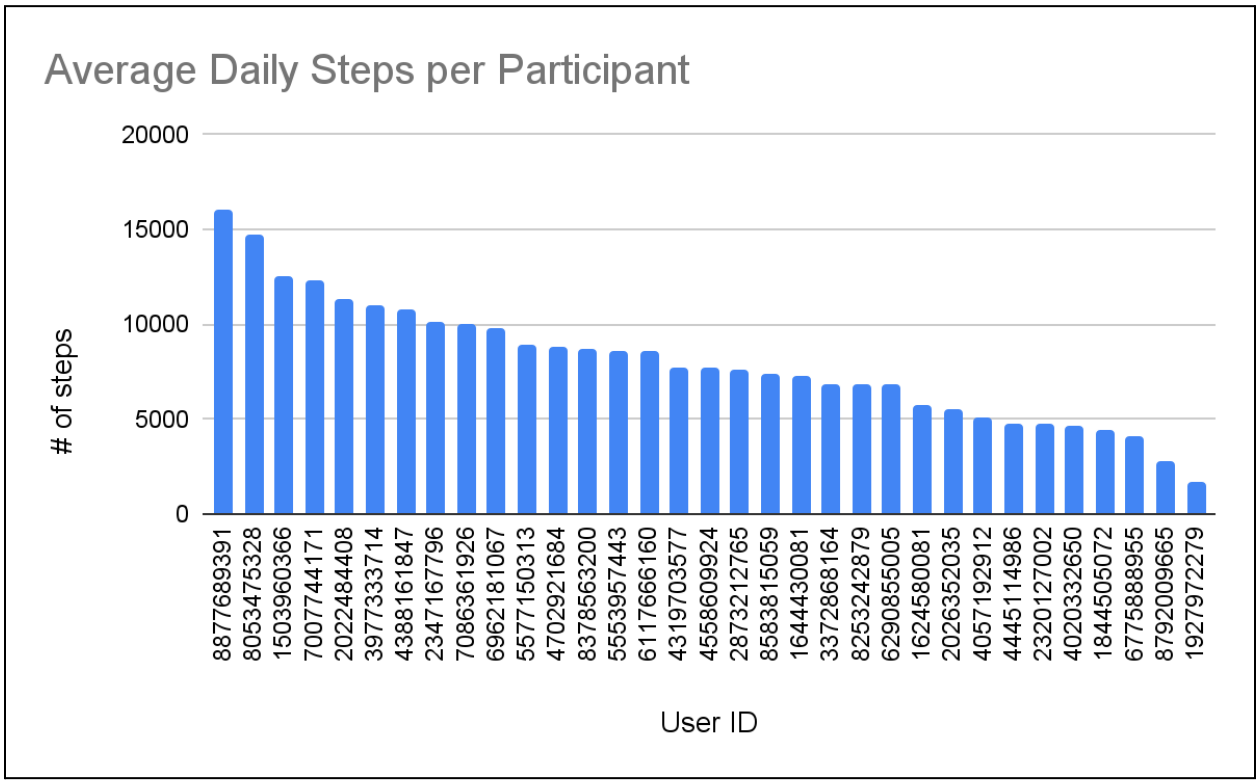
Over half of the fitness-tracker users wore their device nearly all day. All participants wore their tracker at least half the day.

Steps Taken Each Day

Spreadsheet

Next, I wanted to answer the question: on average, **how many steps did each participant take per day?**

1. I calculated the average daily steps per participant using the FILTER function to isolate each unique user ID.
    a. Formula, average steps:
    =AVERAGE(FILTER(dailyActivity_daysWorn!C:C,dailyActivity_daysWorn!A:A=ByParticipant!A2))



- **The dataset shows a fairly even distribution of average daily steps among tracker users ranging from low (1,670 steps per day) to high (16,040 steps per day).**

Key Takeaway

People of all activity levels use fitness trackers.

Weight Change

R Programming

I used RStudio (under Posit Cloud) to analyze the weight log table.

1. First, I set up my RStudio environment by installing the tidyverse package.
    a. `> install.packages('tidyverse')`
    b. `> library(tidyverse)`

2. I uploaded my weight log CSV file to RStudio and loaded it into a table using the read.csv function.

    a. `> weightlog <- read.csv('weightLogInfo_merged.csv')`

3. I explored the dataset by using the colnames and tibble functions.

    a. `> colnames(weightlog)`

```
[1] "Id"             "Date"           "WeightKg"       "WeightPounds"

[5] "Fat"            "BMI"            "IsManualReport" "LogId"
```

    b. `> tibble(weightlog)`

```
# A tibble: 67 × 8

           Id Date           WeightKg WeightPounds   Fat   BMI
IsManualReport   LogId

        <dbl> <chr>             <dbl>        <dbl> <int> <dbl> <chr>
<dbl>

 1 1503960366 5/2/2016 11:5…     52.6         116.    22  22.6 True
1.46e12

 2 1503960366 5/3/2016 11:5…     52.6         116.    NA  22.6 True
1.46e12

 3 1927972279 4/13/2016 1:0…    134.          294.    NA  47.5 False
1.46e12

 4 2873212765 4/21/2016 11:…     56.7         125.    NA  21.5 True
1.46e12

 5 2873212765 5/12/2016 11:…     57.3         126.    NA  21.7 True
1.46e12

 6 4319703577 4/17/2016 11:…     72.4         160.    25  27.5 True
1.46e12

 7 4319703577 5/4/2016 11:5…     72.3         159.    NA  27.4 True
1.46e12

 8 4558609924 4/18/2016 11:…     69.7         154.    NA  27.2 True
1.46e12

 9 4558609924 4/25/2016 11:…     70.3         155.    NA  27.5 True
1.46e12
```

```
10 4558609924 5/1/2016 11:5...     69.9         154.     NA  27.3 True
1.46e12

# ℹ 57 more rows

# ℹ Use `print(n = ...)` to see more rows
```

4. To use the date as the X-axis for my chart, I wanted to split the datetime into two separate columns.

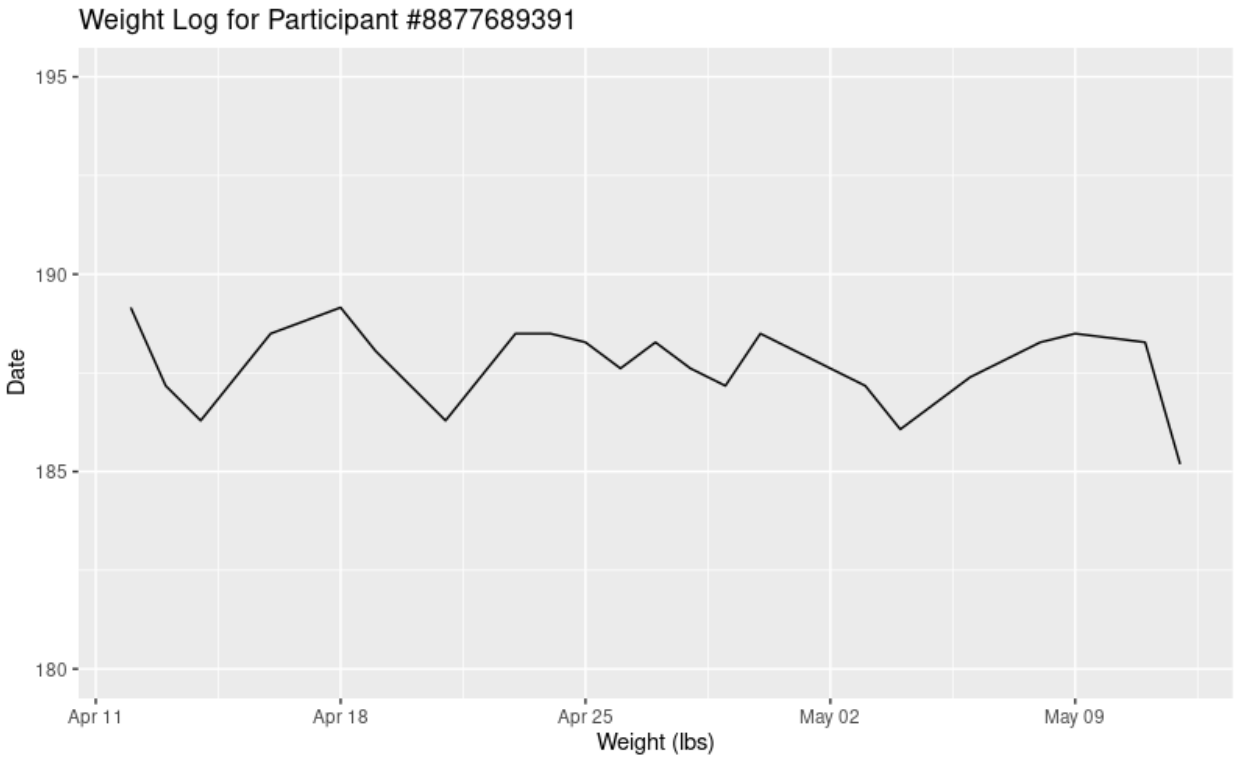
    a.
```
weightlog_datetime <- weightlog %>%
    mutate(
        Date = parse_date_time(
            Date, "m/d/Y I:M:S p"
        )
    ) %>%
    mutate(
        onlyDate = as.Date(Date),
        Time = format(Date, format = "%H:%M:%S")
    )
```

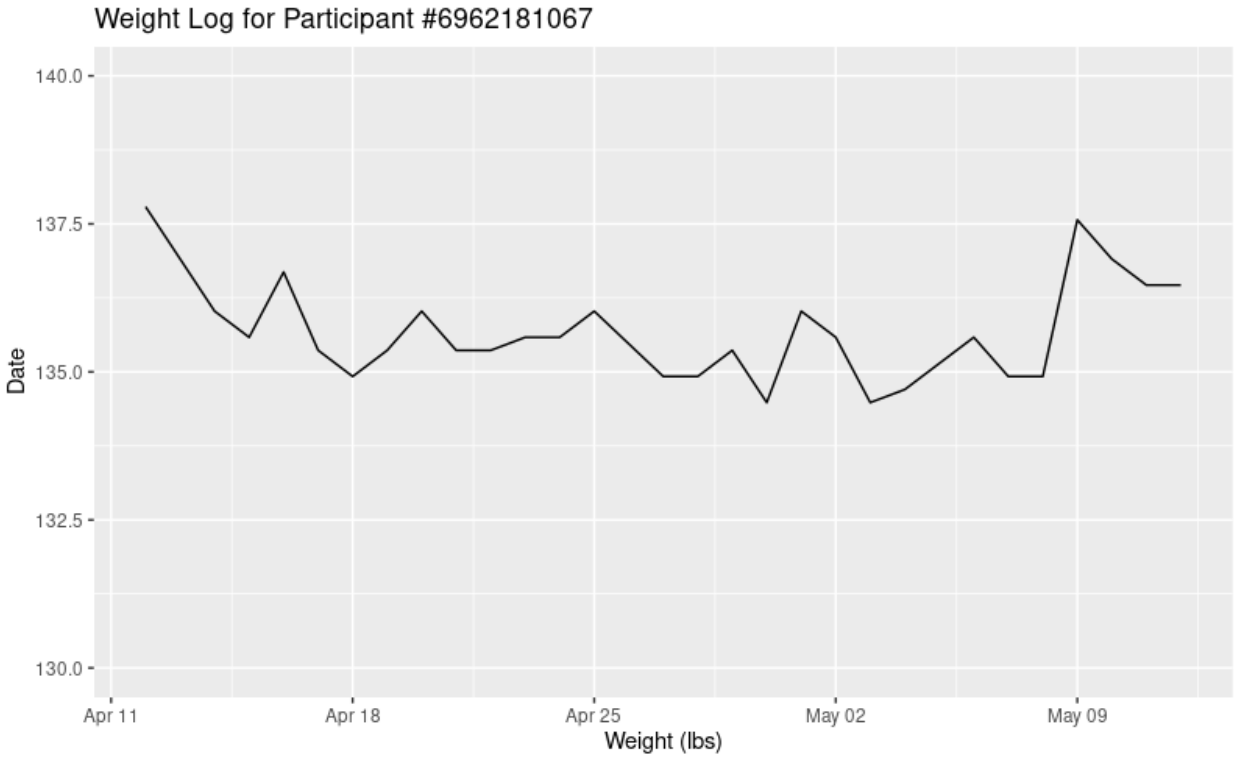
5. Unfortunately, upon exploring the data, I found it too limited to serve as the basis of any marketing recommendations. Of the eight participants, the number of records per person counted as follows (**User ID: # of daily records)**:

6962181067: 30	**8877689391:** 24
4558609924: 5	**4319703577:** 2
2873212765: 2	**1503960366:** 2
1927972279: 1	**5577150313:** 1

I charted the weight logs of the two participants with the most records using the ggplot2 package; however, only two sets of observations are not enough to identify trends.

Lauren Holstein
November 2023
Capstone Project





a.
```
ggplot(subset(weightlog_datetime,Id=="8877689391"), aes(x=onlyDate,
y=WeightPounds, group=Id)) + geom_line() + ylim(180, 195) +
ggtitle("Weight Log for Participant #8877689391") +
    xlab("Weight (lbs)") + ylab("Date")
```

   i. “Subset” indicates the unique user ID I chose to plot.

   ii. “Aes” sets the x and y axis as the date and weight in pounds.

   iii. “Geom_line” indicates a line graph.

   iv. “Ylim” sets the limits of the y axis.

   v. “Ggtitle” adds the chart’s title.

   vi. “Xlab” and “ylab” label the chart’s axes.

6. Both participants showed negligible weight change throughout the tracking period. Again, only two sets of observations are not sufficient to identify trends. However, if the dataset contained more records showing the same pattern, we might conclude that participants were not using the fitness tracker—at least not successfully—as a tool for weight change.

Key Takeaway

Dataset did not include enough observations to identify weight-related trends.

Minutes Slept Each Night

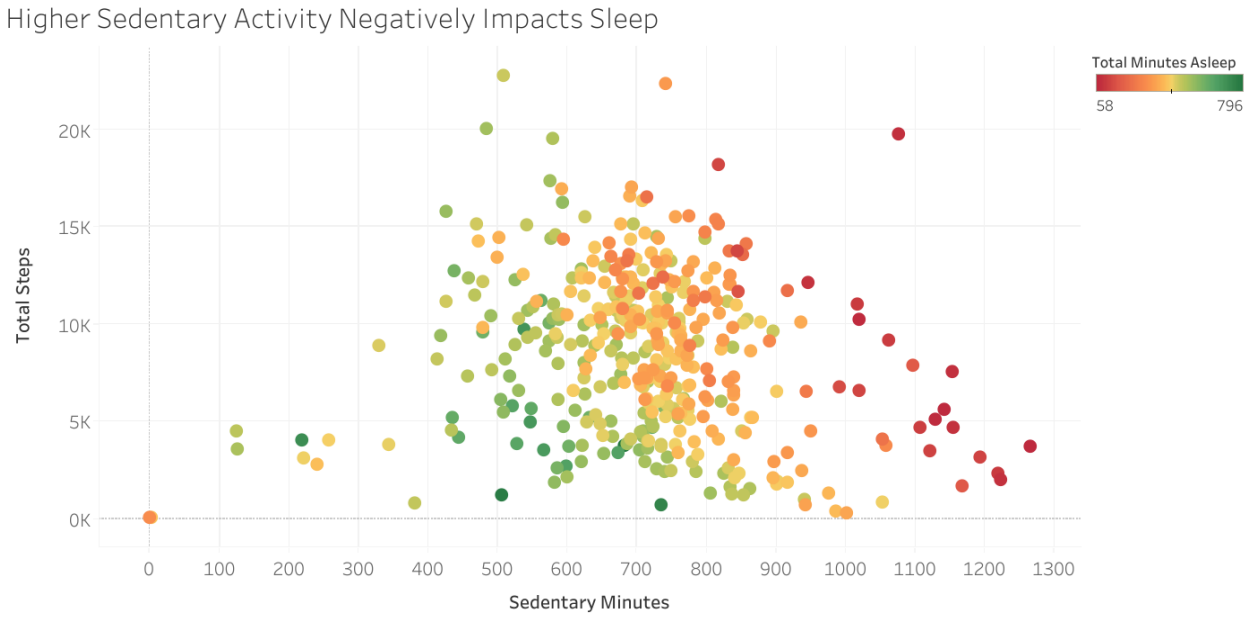
Tableau

As a next analysis step, I wanted to identify any correlations between activity level and the minutes each participant slept during the tracking period.

Total steps walked in a day did not show a strong correlation with sleep. However, **a participant’s daily sedentary minutes strongly correlated with fewer minutes of sleep.**

I identified this trend by creating a scatterplot chart in Tableau.

Lauren Holstein
November 2023
Capstone Project



As you can see in the chart above, as participants' sedentary time increased, their total minutes of sleep decreased.

Key Takeaway

People who were more sedentary slept less.

Recommendations

Everyday Wear Comfort

Most fitness-tracker users wear their tracker almost every day for an average of 20 hours and 4 minutes. For this reason, I recommend emphasizing the all-day wear comfort of the device. It is unobtrusive and comfortable to wear during all daily activities.

Aimed at People of All Activity Levels

Users spanned a broad range of activity levels, from being mostly sedentary to highly active. I recommend messaging around how tracking physical activity is useful for people of all fitness levels. Whether you're an athlete or just beginning your fitness journey, a wearable tracker can help you reach your goals.

Lauren Holstein
November 2023
Capstone Project

Not Weight Focused

The dataset yielded no recommendations regarding weight due to an insufficient number of participant observations. More data collection and analysis is recommended before investing in a weight-related marketing strategy.

A Tool to Reduce Sedentary Time and Increase Sleep

The data showed a strong negative correlation between sedentary time and amount slept. More sedentary participants slept less, while less sedentary participants slept more. This suggests that fitness-tracker wearers can improve their sleep by making sure they're less sedentary throughout the day.

Part of the marketing strategy could focus on helping consumers get a better night's rest by cutting down on sedentary time, perhaps through a reminder to get up and move in the Bellabeat app.