Bellabeat Capstone

Austin Witthuhn

2025-02-05

Table of Contents

Synopsis

In this hypothetical case study, I assume the role of a junior data analyst working on the marketing analytics team at **Bellabeat**, a high-tech company that specializes in health-focused products for women. As part of my responsibilities, I have been tasked with analyzing smart device data to uncover valuable insights into consumer behavior and product usage. Specifically, I have been provided with a dataset of Fitbit users from 2016 and have been asked to conduct a trend analysis to understand how these users interact with their devices.

The goal of this analysis is to identify patterns that could inform Bellabeat's marketing strategy, helping the company better understand its customer base and optimize its product offerings. This case study will break down the data analysis process, from exploring and cleaning the data to generating actionable insights and making recommendations based on the findings.

Ask: Defining the Business Task

Problem Statement

The goal of this analysis is to identify behavioral trends among Fitbit users. These insights will help refine Bellabeat's marketing strategy, target the right customer segments, and optimize the design and features of Bellabeat's products to better meet user needs.

Key Stakeholders

- **Urška Sršen** — (Co-founder & Chief Creative Officer)
- **Sando Mur** — (Co-founder & Mathematician)
- **Bellabeat** — Marketing Analytics Team (including myself)

Prepare: Understanding the Data

Data Sources

The dataset I used for this analysis was sourced from **Kaggle**, containing Fitbit fitness tracker data.

Data Storage

The data files are stored in Posit Cloud. I uploaded all CSV files after extracting them from a ZIP file. I created backup data for all files and stored them in my personal Google Drive.

Data Organization

- The dataset is in **long format**, with each row representing a unique time point for each user.
- The data includes multiple levels: **daily**, **hourly**, and **minute-level** data, providing a detailed view of user activity.

Data Credibility & Bias Considerations

- **Credibility**: The data is credible, though it comes from a relatively small sample size of **33 users**, which is just above the minimum threshold of 30 users recommended by the Central Limit Theorem to ensure the sampling distribution of the mean approximates normality.
- **Potential Bias**: Given that the dataset only includes Fitbit users, it may not be representative of the broader population. There could be a self-selection bias, as individuals who choose to use fitness trackers may be more health-conscious than the general public.

Process: Cleaning and Transforming Data

Install and Load Packages

Below is my code for getting started.

```
# Install any packages that may be needed
install.packages("janitor")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
install.packages("skimr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
install.packages("here")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
# Load in the necessary libraries
library(tidyverse) # For data manipulation and visualization
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.4     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate) # For working with dates and times
library(janitor) # For cleaning data
```

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

```
library(skimr) # For summarizing data
library(here) # For path management
```

```
## here() starts at /cloud/project
```

```
library(knitr) # For dynamic report generation
library(scales) # For scaling data and adding commas to large numbers in plots
```

```
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
# set the working directory to the right folder for csv files (using here for better portability)
setwd(here::here("Fitbit_041216_to_051216/Fitbit_041216_to_051216"))

# Set variables for each CSV file I want to evaluate
daily_activity_may <- read_csv("dailyActivity_merged.csv")
```

```
## Rows: 940 Columns: 15
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (1): ActivityDate
## dbl (14): Id, TotalSteps, TotalDistance, TrackerDistance, LoggedActivitiesDi...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
sleep_day_may <- read_csv("sleepDay_merged.csv")
```

```
## Rows: 413 Columns: 5
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (1): SleepDay
## dbl (4): Id, TotalSleepRecords, TotalMinutesAsleep, TotalTimeInBed
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

View and Clean Data

Below is my code with comments on how I went about with this process.

```
# View the data to confirm the data type, column names, etc. - skimr is also very useful here
str(sleep_day_may)
```

```
## spc_tbl_ [413 × 5] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Id                : num [1:413] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ SleepDay          : chr [1:413] "4/12/2016 12:00:00 AM" "4/13/2016 12:00:00 AM" "4/15/2016 12:00:00 AM" "4/
16/2016 12:00:00 AM" ...
##  $ TotalSleepRecords : num [1:413] 1 2 1 2 1 1 1 1 1 1 ...
##  $ TotalMinutesAsleep: num [1:413] 327 384 412 340 700 304 360 325 361 430 ...
##  $ TotalTimeInBed    : num [1:413] 346 407 442 367 712 320 377 364 384 449 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Id = col_double(),
##   ..   SleepDay = col_character(),
##   ..   TotalSleepRecords = col_double(),
##   ..   TotalMinutesAsleep = col_double(),
##   ..   TotalTimeInBed = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(daily_activity_may)
```

```
## spc_tbl_ [940 × 15] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Id                       : num [1:940] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ ActivityDate             : chr [1:940] "4/12/2016" "4/13/2016" "4/14/2016" "4/15/2016" ...
##  $ TotalSteps               : num [1:940] 13162 10735 10460 9762 12669 ...
##  $ TotalDistance            : num [1:940] 8.5 6.97 6.74 6.28 8.16 ...
##  $ TrackerDistance          : num [1:940] 8.5 6.97 6.74 6.28 8.16 ...
##  $ LoggedActivitiesDistance: num [1:940] 0 0 0 0 0 0 0 0 0 0 ...
##  $ VeryActiveDistance       : num [1:940] 1.88 1.57 2.44 2.14 2.71 ...
##  $ ModeratelyActiveDistance: num [1:940] 0.55 0.69 0.4 1.26 0.41 ...
##  $ LightActiveDistance      : num [1:940] 6.06 4.71 3.91 2.83 5.04 ...
##  $ SedentaryActiveDistance : num [1:940] 0 0 0 0 0 0 0 0 0 0 ...
##  $ VeryActiveMinutes        : num [1:940] 25 21 30 29 36 38 42 50 28 19 ...
##  $ FairlyActiveMinutes      : num [1:940] 13 19 11 34 10 20 16 31 12 8 ...
##  $ LightlyActiveMinutes     : num [1:940] 328 217 181 209 221 164 233 264 205 211 ...
##  $ SedentaryMinutes         : num [1:940] 728 776 1218 726 773 ...
##  $ Calories                 : num [1:940] 1985 1797 1776 1745 1863 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Id = col_double(),
##   ..   ActivityDate = col_character(),
##   ..   TotalSteps = col_double(),
##   ..   TotalDistance = col_double(),
##   ..   TrackerDistance = col_double(),
##   ..   LoggedActivitiesDistance = col_double(),
##   ..   VeryActiveDistance = col_double(),
##   ..   ModeratelyActiveDistance = col_double(),
##   ..   LightActiveDistance = col_double(),
##   ..   SedentaryActiveDistance = col_double(),
##   ..   VeryActiveMinutes = col_double(),
##   ..   FairlyActiveMinutes = col_double(),
##   ..   LightlyActiveMinutes = col_double(),
##   ..   SedentaryMinutes = col_double(),
##   ..   Calories = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
# Verify number of distinct user IDs in both tables
n_distinct(sleep_day_may$Id) # 24 unique Ids - under 30 participants (Central Limit Theorem)
```

```
## [1] 24
```

```
n_distinct(daily_activity_may$Id) # 33 unique user Ids
```

```
## [1] 33
```

```
# Determine any duplicated rows
sum(duplicated(daily_activity_may))
```

```
## [1] 0
```

```
sum(duplicated(sleep_day_may))
```

```
## [1] 3
```

```
# Clean and preprocess the dataset: remove duplicates and missing values, standardize column names, rename specif
ic columns for consistency, reformat date columns to the correct date format, and trim any leading/trailing space
s from the 'id' column.
daily_activity_may <- daily_activity_may %>%
  distinct() %>%
  drop_na() %>%
  clean_names() %>%
  rename(date = activity_date) %>%
  mutate(date = as_date(date, format = "%m/%d/%Y")) %>%
  mutate(id = trimws(id))

sleep_day_may <- sleep_day_may %>%
  distinct() %>%
  drop_na() %>%
  clean_names() %>%
  rename(date = sleep_day) %>%
  mutate(date = as_date(date, format = "%m/%d/%Y %I:%M:%S %p")) %>%
  mutate(id = trimws(id))
```

Analyze: Finding Trends & Insights

Get Summaries of Tables

Below you will find my code where I got the summary statistics for two of the tables.

```
# Option One (summary() function):
# daily_activity_may %>%
#   select(totalsteps, # Average = 7638 steps
#          totaldistance, # Average = 5.49 (unsure if KM or Miles)
#          sedentaryminutes, # Average = 991.2 (16.52 hours)
#          calories) %>%  # Average = 2304 calories
#   summary()
#
# sleep_day_may %>%
#   select(totalminutesasleep, # Average = 419.2 minutes (~7 hours)
#          totaltimeinbed) %>% # Average = 458.5 minutes (~7.5 hours)
#   summary()

# Option Two (skim() function)
skim(daily_activity_may)
```

Data summary

Name	daily_activity_may
Number of rows	940
Number of columns	15
_____	
Column type frequency:	
character	1
Date	1
numeric	13

Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
id	0	1	10	10	0	33	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
date	0	1	2016-04-12	2016-05-12	2016-04-26	31

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
total_steps	0	1	7637.91	5087.15	0	3789.75	7405.50	10727.00	36019.00	
total_distance	0	1	5.49	3.92	0	2.62	5.24	7.71	28.03	
tracker_distance	0	1	5.48	3.91	0	2.62	5.24	7.71	28.03	
logged_activities_distance	0	1	0.11	0.62	0	0.00	0.00	0.00	4.94	
very_active_distance	0	1	1.50	2.66	0	0.00	0.21	2.05	21.92	
moderately_active_distance	0	1	0.57	0.88	0	0.00	0.24	0.80	6.48	
light_active_distance	0	1	3.34	2.04	0	1.95	3.36	4.78	10.71	
sedentary_active_distance	0	1	0.00	0.01	0	0.00	0.00	0.00	0.11	
very_active_minutes	0	1	21.16	32.84	0	0.00	4.00	32.00	210.00	
fairly_active_minutes	0	1	13.56	19.99	0	0.00	6.00	19.00	143.00	
lightly_active_minutes	0	1	192.81	109.17	0	127.00	199.00	264.00	518.00	
sedentary_minutes	0	1	991.21	301.27	0	729.75	1057.50	1229.50	1440.00	
calories	0	1	2303.61	718.17	0	1828.50	2134.00	2793.25	4900.00	

```
skim(sleep_day_may)
```

Data summary

Name	sleep_day_may
Number of rows	410
Number of columns	5

Column type frequency:	
character	1
Date	1
numeric	3

Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
id	0	1	10	10	0	24	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
date	0	1	2016-04-12	2016-05-12	2016-04-27	31

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
total_sleep_records	0	1	1.12	0.35	1	1.00	1.0	1	3	▇▁▁▁▁
total_minutes_asleep	0	1	419.17	118.64	58	361.00	432.5	490	796	▁▃▇▅▁
total_time_in_bed	0	1	458.48	127.46	61	403.75	463.0	526	961	▁▃▇▂▁

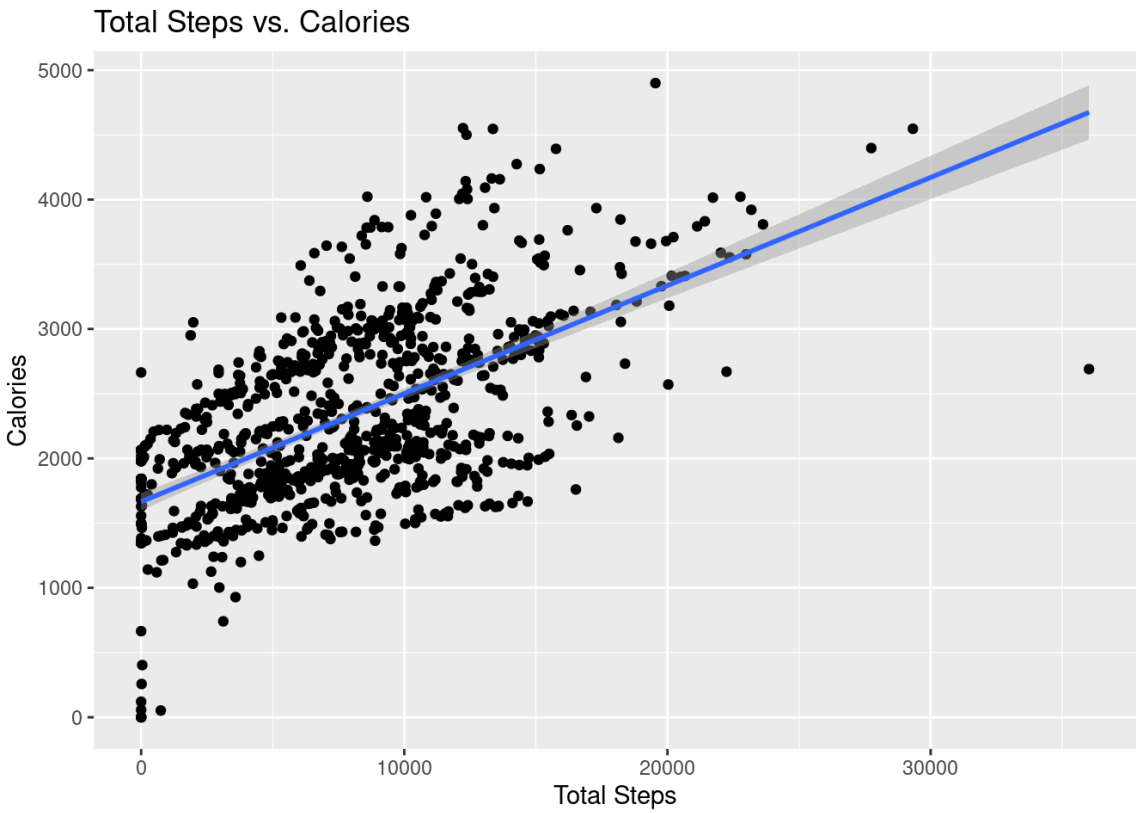
Key Findings

1. **Average Step Count and Comparison to U.S. Average**
    - The average Fitbit user in this study took 7,638 steps a day, which is double the amount of steps the average American walks per day (3,000–4,000 steps) according to the Mayo Clinic (https://tinyurl.com/5n85pppj (https://tinyurl.com/5n85pppj)).
2. **Market Opportunity**
    - Since Fitbit users are already significantly more active than the general population, Bellabeat could position their products as tools for those who are already committed to fitness, allowing for more targeted marketing toward health-conscious individuals.

Plot Variables & Determine Correlation

```
# Plot and find correlation between different variables
ggplot(data=daily_activity_may, aes(x=total_steps, y=calories)) +
  geom_point() +
  ggtitle("Total Steps vs. Calories") +
  xlab("Total Steps") +
  ylab("Calories") +
  geom_smooth(method = "lm")
```

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



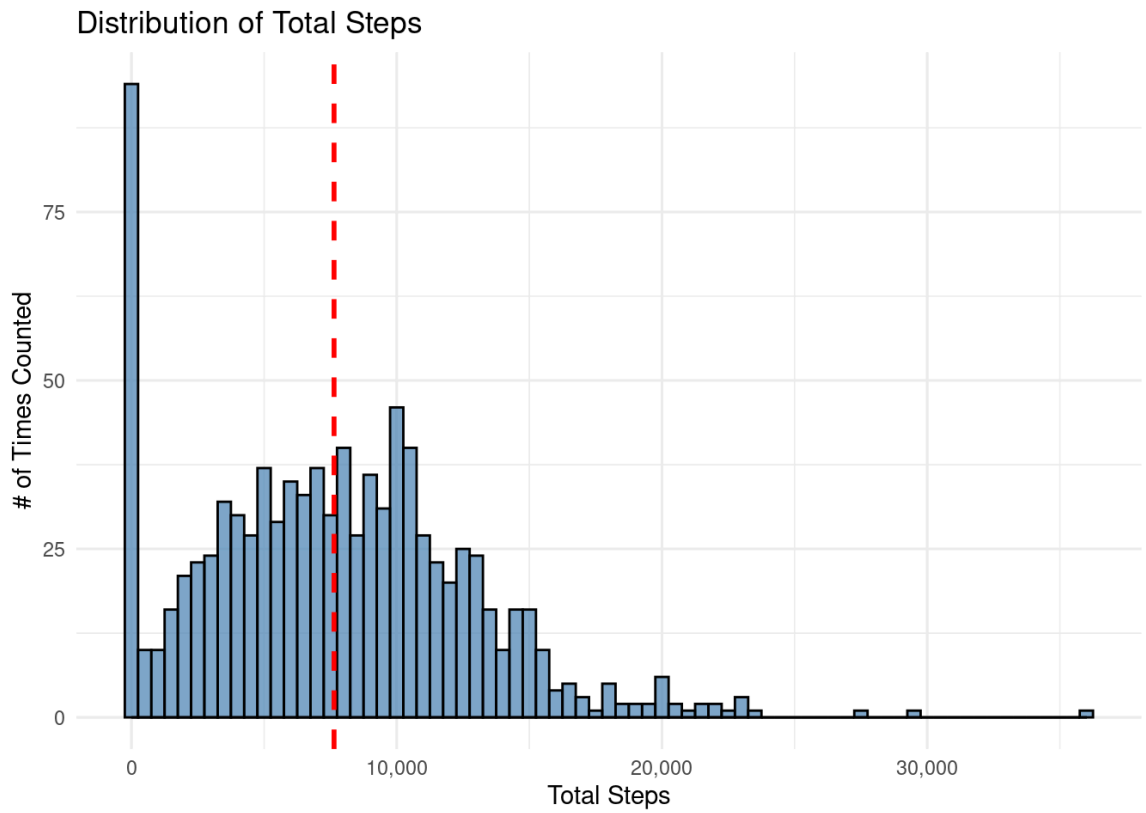

```
# cor(daily_activity_may$steps, daily_activity_may$calories)  # Moderate correlation (0.59)
```

Key Findings

1. **Correlation Between Steps and Calories**
   - The more steps each Fitbit user took, the more calories they generally burned. This suggests a positive relationship between physical activity (as measured by steps) and energy expenditure (as measured by calories burned).
2. **Activity Levels and Health Impact**
   - The relationship between steps and calories reinforces the idea that increased physical activity directly contributes to health outcomes like weight loss and improved fitness.

```
ggplot(daily_activity_may, aes(x=total_steps)) +
  geom_histogram(binwidth=500, fill="steelblue", color="black", alpha=0.7) +  # Adjusted binwidth
  geom_vline(aes(xintercept=mean(total_steps)), color="red", linetype="dashed", linewidth=1) +  # Vertical line f
or mean (using linewidth)
  geom_density(aes(y = ..density..), fill="orange", alpha=0.2) +  # Density curve
  scale_x_continuous(labels = scales::comma) +  # Fix: specify scales::comma for formatting x-axis labels
  ggtitle("Distribution of Total Steps") +
  xlab("Total Steps") +
  ylab("# of Times Counted") +
  theme_minimal()  # Clean theme
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## ℹ Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Key Findings

1. **Outlier Detection**
    - I included this distribution to demonstrate the presence of outlier data. As shown, there are nearly 100 instances where users recorded 0 steps. This creates a significant skew in the distribution, and it's crucial to address this anomaly.
2. **Possible Causes**
    - **Device Malfunction**: Could the Fitbit devices have malfunctioned on certain days, leading to missing data or failure to track activity accurately?
    - **Inactive Users**: Is it possible that these users were genuinely sedentary for the entire day, failing to achieve even 500 steps? This could suggest that a subset of users may not be engaging with the device as actively as intended.
    - **Device Wear Issues**: Another plausible explanation is that users are wearing their devices but not using them effectively, or the devices may not be worn at all on certain days (e.g., left at home).
3. **Impact on Analysis**
    - These zero-step occurrences can distort the overall analysis, and this warrants a deeper dive to understand whether these are legitimate data points or just noise. This investigation will help refine data quality for more accurate conclusions and decision-making.

Share: Communicate Findings

1. **User Engagement**
    - The average Fitbit user takes 7,638 steps per day — well above the U.S. average of 3,000–4,000 steps. This indicates that Fitbit users, and likely other people who utilize fitness wearables, are likely more health-conscious and engaged in physical activity.
2. **Correlation Between Activity and Calories**
    - There is a moderate correlation between the number of steps and the calories burned. The more active users are, the more calories they burn, which is expected but important to reinforce when creating fitness-related marketing materials or campaigns.
3. **Analysis of Outlier Data**
    - Depending on the data from the histogram, analysis needs to be done on the outlier data. The 0 steps taken may indicate that people will wear “fitness” devices for things other than fitness, allowing Bellabeat to market their products as doing

more than just fitness, as one example.

Act: Recommendations and Next Steps

1. **Target Market Expansion**
   - **Recommendation**: Based on the analysis showing higher activity levels among users of fitness devices, Bellabeat could broaden its target market to include health-conscious individuals who value overall wellness, not just fitness tracking.
   - **Strategic Fit**: Bellabeat's current products — **Leaf** (wellness tracker) and **Time** (wellness watch) — are versatile and stylish, designed to appeal to users who want a blend of functionality and aesthetic. By positioning these devices as holistic wellness tools that track not just steps but also sleep, stress, and mindfulness, Bellabeat can attract individuals who may be interested in monitoring a broader range of health metrics.
   - **Action**: Develop marketing campaigns that highlight the multifunctional nature of the **Leaf** and **Time**, showcasing how they help users track their physical activity, sleep patterns, stress levels, and mindfulness. This approach would resonate with individuals looking for a comprehensive health companion.
2. **Product Optimization**
   - **Recommendation**: Since there is a clear connection between activity levels and calories burned, Bellabeat should emphasize this messaging across all their products, particularly the **Leaf** and **Time**, positioning them as tools to help users stay active and achieve weight loss or maintenance goals.
   - **Action**: Update the app's messaging to include personalized recommendations based on the user's activity levels. For example, if a user consistently takes fewer than 5,000 steps, the app could suggest incremental activity goals and encourage users to increase their movement. This could be combined with tips for healthy habits, emphasizing how consistent activity (tracked through **Leaf** and **Time**) can lead to improved health outcomes, including weight loss.
3. **Market Fit**
   - **Recommendation**: Given the presence of nearly 100 occurrences of 0 steps, it's important to understand why users may not be using their devices as intended. Instead of dismissing these data points as errors, they could reveal valuable insights into user behavior, potentially indicating that the devices are being used for purposes other than fitness tracking.
   - **Market Opportunity**: The **Leaf** and **Time** are designed to be worn as accessories, which may appeal to users who are not strictly focused on fitness but still want to monitor their overall wellness. For example, some users might be using the devices primarily for sleep tracking or stress monitoring, not just step counting. By understanding these behaviors, Bellabeat could explore ways to better market its products as all-in-one wellness tools.
   - **Broader Product Positioning**: Bellabeat should consider rebranding or enhancing their messaging to reflect the holistic nature of the devices. Instead of just advertising the **Leaf** and **Time** as fitness trackers, they should position them as wellness companions that support users' health goals, such as improving sleep quality, reducing stress, and tracking mindfulness.
   - **Next Steps**: Bellabeat could run a survey targeting users who log frequent zero-step days to better understand how they are using the product. This feedback will help identify whether these users are using the devices for non-fitness purposes and allow Bellabeat to refine its marketing strategy to appeal to a wider audience interested in a broad range of health metrics. For example, an online campaign highlighting how **Time** tracks not only steps but also mindfulness, stress, and sleep could attract individuals interested in a more comprehensive health tool.

Summary of Recommendations:

1. **Expand Target Market**: Position **Leaf** and **Time** as holistic wellness devices that appeal to users interested in monitoring their overall health, not just fitness.
2. **Enhance Product Messaging**: Highlight the connection between activity levels, sleep, and stress management, emphasizing the comprehensive benefits of using Bellabeat products.
3. **Investigate User Behavior**: Use data insights to explore why users log 0 steps and understand if they are using the device for non-fitness purposes.

Conclusion

This analysis looked at a small portion of the available data to dig into user behavior and offer actionable insights for Bellabeat's marketing strategy. Key findings like the impact of activity levels on health outcomes and the potential to target a broader wellness market suggest that Bellabeat could broaden its product positioning. Instead of focusing only on fitness enthusiasts, they could market

their devices as all-encompassing wellness tools, which would appeal to a wider audience interested in overall health, like sleep tracking and mindfulness.

Notes and Considerations

While this analysis focused on a small subset of the available data, it's important to note that there's much more to explore within the full dataset. A deeper dive into the nearly 100 instances of zero-step days could shed light on user behavior and device usage patterns, while examining the CSV files that document user activity down to the minute could uncover even more actionable insights. These deeper analyses might help refine recommendations, uncover potential device issues, or give a broader understanding of health trends.

That being said, the dataset used here is far from ideal for a company looking to expand its operations. Fitbit data that's nine years old and only represents around 30 users isn't the best foundation for strategic decision-making. Though the analysis provided some useful insights, a more recent, larger dataset would offer a much clearer picture of current user behavior, particularly in a fast-changing market.

Given the scope of this capstone and the time constraints, the analysis was intentionally narrowed to key areas to provide focused insights. There's plenty of room for further exploration, and I'm looking forward to applying the skills learned here to new challenges and diving deeper into other areas of data analysis.

Resources and Acknowledgments

While I dedicated significant effort to this project, the final product wouldn't have been as polished without the support of the following resources:

- **ChatGPT**: For assistance with data analysis techniques, insights, and overall guidance throughout the project.
- **Bellabeat Capstone Project Template**: From the Google Data Analytics Professional Certificate course, which provided the initial framework for the analysis.
- **Adeola Shogbaike's Capstone on Medium.com**: For offering additional insights and inspiration on approaching the Bellabeat project and refining analysis techniques.

These resources were extremely valuable in ensuring a thorough and structured analysis, ultimately contributing to the actionable recommendations presented.