Bellabeat Case Study: Smart Device Usage Analysis

Gonca Turgen

2025-06-20

Introduction

Purpose of the analysis

The purpose of this analysis is to explore how consumers use non-Bellabeat smart devices, particularly Fitbit, and apply these insights to generate marketing strategies for one of Bellabeat's products. This case study uses data from a public Fitbit dataset to conduct exploratory data analysis and generate actionable insights.

Selected product: Bellabeat Leaf

Bellabeat offers a range of wellness products including the Leaf, Time, and Spring. For the purpose of this analysis, I focus on the **Bellabeat Leaf** — a smart health tracker designed specifically for women. The Leaf tracks physical activity, sleep, reproductive health, and mindfulness through a stylish wearable device that can be worn as a necklace, bracelet, or clip.

By analyzing Fitbit user data, this study aims to extract behavioral patterns that can inform more targeted marketing strategies for the Bellabeat Leaf and increase user engagement.

Stakeholders and Business Context

This analysis has been conducted for Bellabeat, a high-tech company that manufactures health-focused smart devices primarily targeted at women. The key stakeholders for this case study include:

- Bellabeat's cofounder and Chief Creative Officer, who is responsible for product vision and creative direction.
- Bellabeat's cofounder and executive team member, focused on strategic leadership and business growth.
- The marketing analytics team, which will use the insights to inform data-driven marketing campaigns and user engagement strategies.

These stakeholders are interested in understanding how users interact with smart health devices — especially activity, sleep, and weight tracking behaviors — to design more targeted, personalized, and engaging marketing strategies for Bellabeat's product line.

Although Bellabeat offers several products, for the purpose of this analysis we focus on the **Bellabeat Leaf**, a wellness tracker designed specifically for women. The insights derived from the analysis of Fitbit user data are intended to help improve marketing approaches and user experience for this product.

Business Task

Key question

How do users of non-Bellabeat smart devices behave?

Objective

Apply insights to Bellabeat marketing strategy.

About the Data

Data source

The dataset used in this project is the "Fitbit Fitness Tracker Data" sourced from Kaggle. It includes information such as daily steps, calories burned, sleep duration, and activity levels collected from 30 Fitbit users in 2016.

ROCCC assessment

- Relevant: Yes — directly related to smart device usage.
- Original: No — not first-party Fitbit data.
- Comprehensive: Limited — small sample size (30 users).
- Current: No — data is from 2016.
- Cited: Minimally — sourced from public sharing, not a verified provider.

Limitations and Assumptions

Limitations

- **Small Sample Size:** The dataset includes only 30 users, which limits the generalizability of the findings. The behavior patterns observed may not fully represent the broader population of smart device users.

- **No Demographic Information:** Critical demographic variables such as gender, age, location, or occupation are missing. This is especially relevant since Bellabeat primarily targets women, but the dataset does not allow for gender-specific insights.

- **Outdated Data:** The data was collected in 2016. While it provides general behavioral patterns, it may not reflect current trends in smart health device usage or the evolving preferences of modern users.

- **Limited Data Provenance:** The dataset was shared publicly via Kaggle and was not officially published by Fitbit. It lacks formal documentation or verification, which may impact its reliability.

- **Inconsistent Logging:** Some users did not consistently track all metrics (e.g., sleep, weight), resulting in varying data completeness across users and variables.

Assumptions

- The users in the dataset are assumed to be general smart device users whose behaviors can offer exploratory insights relevant to Bellabeat's target market.
- Due to the lack of demographic data, it is assumed that the users reflect a lifestyle aligned with Bellabeat's target audience (e.g., adult women interested in wellness, activity tracking, and mindfulness).
- The analysis assumes that consistent tracking behavior reflects higher user engagement, which is used to define user consistency levels and inform marketing strategies.

Tools and environment

R Packages Used

The following R packages were used throughout the analysis:

- **tidyverse**: For data wrangling, transformation, and visualization
- **lubridate**: For parsing and handling date-time formats
- **janitor**: For cleaning column names and simplifying datasets
- **skimr**: For generating summary statistics
- **ggplot2**: For creating data visualizations

All packages are part of the CRAN ecosystem and were loaded using the `library()` function at the start of the analysis.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)
library(janitor)
```

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

```r
library(skimr)
library(ggplot2)
```

File Structure

The project directory includes the following structure:

- **Bellabeat_case_study.Rmd**: Main analysis and report document
- **/Bellabeat Dataset/**: Contains all original CSV files downloaded from Kaggle
- **/output/**: Folder for exported plots or visual summaries (optional)
- **Bellabeat_case_study.html / .pdf**: Rendered final report (created via Knit)

Reproducibility Notes

- The analysis was conducted in **RStudio** using **R version 2025.05.0+496**.
- All random or time-sensitive functions were avoided to ensure results are reproducible.
- The date fields across all datasets were reformatted to a consistent `Date` class using `lubridate::mdy_hms()` and `as_date()` functions.
- All code blocks are included in this report to ensure transparency and reproducibility.

Data Cleaning

Import and Inspect

```r
daily_activity <- read_csv("Bellabeat Dataset/Fitabase Data 3.12.16-4.11.16/dailyActivity_merged.csv")
```

```
## Rows: 457 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.
```

```r
sleep_minute <- read_csv("Bellabeat Dataset/Fitabase Data 3.12.16-4.11.16/minuteSleep_merged.csv")
```

```
## Rows: 198559 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): date
## dbl (3): Id, value, logId
##
## 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.
```

```r
heartrate_seconds <- read_csv("Bellabeat Dataset/Fitabase Data 3.12.16-4.11.16/heartrate_seconds_merged
```

```
## Rows: 1154681 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Time
## dbl (2): Id, Value
##
## 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.
```

```r
hourly_calories <- read_csv("Bellabeat Dataset/Fitabase Data 3.12.16-4.11.16/hourlyCalories_merged.csv")
```

```
## Rows: 24084 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): ActivityHour
## dbl (2): Id, Calories
##
## 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.
```

```r
hourly_intensities <- read_csv("Bellabeat Dataset/Fitabase Data 3.12.16-4.11.16/hourlyIntensities_merge
```

```
## Rows: 24084 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): ActivityHour
## dbl (3): Id, TotalIntensity, AverageIntensity
##
## 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.
```

```r
hourly_steps <- read.csv("Bellabeat Dataset/Fitabase Data 3.12.16-4.11.16/hourlySteps_merged.csv")
minute_calories <- read_csv("Bellabeat Dataset/Fitabase Data 3.12.16-4.11.16/minuteCaloriesNarrow_merge
```

```
## Rows: 1445040 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): ActivityMinute
## dbl (2): Id, Calories
##
## 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.
```

```r
minute_steps <- read_csv("Bellabeat Dataset/Fitabase Data 3.12.16-4.11.16/minuteStepsNarrow_merged.csv")
```

```
## Rows: 1445040 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
```

```
## chr (1): ActivityMinute
## dbl (2): Id, Steps
##
## 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.
```

```
weight_log_info <- read_csv("Bellabeat Dataset/Fitabase Data 3.12.16-4.11.16/weightLogInfo_merged.csv")
```

```
## Rows: 33 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Date
## dbl (6): Id, WeightKg, WeightPounds, Fat, BMI, LogId
## lgl (1): IsManualReport
##
## 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.
```

```
minute_METs <- read.csv("Bellabeat Dataset/Fitabase Data 3.12.16-4.11.16/minuteMETsNarrow_merged.csv")
minute_intensities <- read.csv("Bellabeat Dataset/Fitabase Data 3.12.16-4.11.16/minuteIntensitiesNarrow_
sleep_day <- read_csv("Bellabeat Dataset/Fitabase Data 3.12.16-4.11.16/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.
```

Datasets Used

Bellabeat's Leaf product is marketed as a holistic wellness tracker that supports users in staying mindful of their physical activity, rest, and meditation patterns. As their official messaging states:

> *"Whether you're sleeping, meditating, or moving, Leaf keeps up with you."*

With this in mind, the datasets used in this analysis were chosen based on how closely they align with the Leaf experience and the type of insights its users are most likely to value. Specifically:

- **`dailyActivity_merged.csv`** was used to assess **physical movement patterns** such as steps, calories burned, and time spent in various intensity levels.
- **`sleepDay_merged.csv`** was used to analyze **sleep behavior**, focusing on duration, consistency, and sleep efficiency.
- **`weightLogInfo_merged.csv`** was included optionally to explore **long-term wellness trends**, but not emphasized since Leaf does not track weight directly.

More granular datasets, such as minute- or second-level data (`minuteStepsNarrow.csv`, `heartrate_seconds_merged.csv`, etc.), were intentionally excluded. These files are better suited for real-time biometric tracking, which is more aligned with athletic-focused wearables than wellness-oriented devices like Leaf.

By focusing on **daily trends and behaviors**, this analysis remains faithful to the product's core value: helping users develop healthier routines without overwhelming them with technical data.

Exploratory Data Analysis

Activity Behavior

Business value: Identify target users for *fitness engagement* campaigns.

Physical activity + calories, per user per day

- How many steps/calories users average per day
- How active they are (using active minutes)
- How sedentary they are (sedentary minutes)
- Whether activity patterns are consistent or vary a lot

Activity Behavior Data Cleaning

Column cleaning and formatting

```
daily_activity_df <- daily_activity %>%
  select(Id, ActivityDate, TotalSteps, TotalDistance, Calories, VeryActiveMinutes, FairlyActiveMinutes,
  rename(id=Id, activity_date=ActivityDate,total_steps=TotalSteps, total_distance=TotalDistance, calory=
  mutate(active_minutes=VeryActiveMinutes+FairlyActiveMinutes+LightlyActiveMinutes) %>%
  select(-VeryActiveMinutes, -FairlyActiveMinutes, -LightlyActiveMinutes)
```

Date standardization

To ensure consistency across all datasets, I standardized the date columns by converting them into a unified Date format (YYYY-MM-DD). This was necessary because daily_activity dataset used only dates but were also stored as character strings. Using the lubridate package, I parsed all date-related columns and renamed columns to ensure consistent naming (activity_date) across datasets, which is essential for filtering and joining data accurately.

```
daily_activity_df <- daily_activity_df %>%
  mutate(activity_date = mdy(activity_date))
```

Activity Behavior Analysis

1- User-level activity profile Step 1: Group by user → Calculate daily averages

```
user_activity_summary <- daily_activity_df %>%
  group_by(id) %>%
  summarize(
    avg_steps = mean(total_steps, na.rm = TRUE),
    avg_calories = mean(calory, na.rm = TRUE),
    avg_active = mean(active_minutes, na.rm = TRUE),
    avg_sedentary = mean(sedentary_minutes, na.rm = TRUE),
    days_logged = n_distinct(activity_date)
  )
```

This gives me one row per user — their *daily fitness profile*.

Step 2: Add activity ratio

```
user_activity_summary <- user_activity_summary %>%
  mutate(activity_ratio = avg_active / (avg_active + avg_sedentary))
```

Step 3: Explore the results

The following user segmentation could enable Bellabeat to tailor marketing campaigns:

1- Top 5 most active users *by steps*:

```
user_activity_summary %>%
  arrange(desc(avg_steps)) %>%
  head(5)
```

```
## # A tibble: 5 x 7
##           id avg_steps avg_calories avg_active avg_sedentary days_logged
##        <dbl>     <dbl>        <dbl>      <dbl>         <dbl>       <int>
## 1 8877689391    17417.        3451.       323.         1046.          12
## 2 8053475328    14844.        2893.       251.         1114.          11
## 3 6962181067    12640.        2089.       323.          606.          14
## 4 7007744171    12260.        2627.       338.         1001.          12
## 5 2022484408    12175.        2475.       316.         1059.          12
## # i 1 more variable: activity_ratio <dbl>
```

- These users average 12,000–17,000 steps/day, which is well above the recommended 10,000.
- They also log 300+ active minutes/day, which is very high.
- Their activity ratios (active vs. total time) are low-to-moderate (around 18%–35%), meaning that while they are very active, they're still sedentary for many hours (e.g., 1000+ sedentary mins/day = ~17 hours).

2- Top 5 most sedentary users:

```
user_activity_summary %>%
  arrange(desc(avg_sedentary)) %>%
  head(5)
```

```
## # A tibble: 5 x 7
##           id avg_steps avg_calories avg_active avg_sedentary days_logged
##        <dbl>     <dbl>        <dbl>      <dbl>         <dbl>       <int>
## 1 4388161847        0         1805.        0           1384.           8
## 2 4057192912     1887.        1904.       49.8         1358.          32
## 3 8253242879     2390.        1463        69.8         1352.          12
## 4 6290855005     1618.        2166.       57.8         1290.          10
## 5 1624580081     4226.        1353.      122.          1278.          19
## # i 1 more variable: activity_ratio <dbl>
```

- These users log low to no steps or activity.
- Their sedentary time is extreme: 1270–1380 minutes/day → that's 21–23 hours/day, likely due to not wearing the device regularly or low activity overall.
- Activity ratios are near zero (0%–9%).

3- Top 5 users with highest activity ratio:

```r
user_activity_summary %>%
  arrange(desc(activity_ratio)) %>%
  head(5)
```

```
## # A tibble: 5 x 7
##           id avg_steps avg_calories avg_active avg_sedentary days_logged
##        <dbl>     <dbl>        <dbl>      <dbl>         <dbl>       <int>
## 1 6962181067    12640.        2089.       323.          606.          14
## 2 2347167796     9800.        2021.       288.          684.          15
## 3 5577150313     8608.        3300.       267.          666.          11
## 4 5553957443     8355.        1802.       224.          608.          12
## 5 4702921684     7943.        2821.       267.          727.          15
## # i 1 more variable: activity_ratio <dbl>
```

- These users don't necessarily take the most steps, but they have the best balance between movement and rest.
- Their activity ratios (~27–35%) suggest they stay consistently active, not just during intense bursts.

2- User consistency over time Are users active every day? Only on some days? Do they have patterns?

Step 1: Count number of active days per user

I will define an active day as one with more than 0 steps:

```r
user_consistency <- daily_activity_df %>%
  filter(total_steps > 0) %>%
  group_by(id) %>%
  summarize(
    active_days = n_distinct(activity_date)
  )
```

I will also calculate the total number of days they were tracked:

```r
user_total_days <- daily_activity_df %>%
  group_by(id) %>%
  summarize(
    total_days = n_distinct(activity_date)
  )
```

Then combine:

```r
consistency_summary <- left_join(user_consistency, user_total_days, by = "id") %>%
  mutate(
    consistency_rate = active_days / total_days
  )
```

Step 2: Merge consistency summary with activity summary

I will analyze consistency alongside steps, calories, etc.

```r
user_activity_summary <- left_join(user_activity_summary, consistency_summary, by = "id")
```

Step 3: Explore consistency patterns

1- Top 5 most consistent users:

```
user_activity_summary %>%
  arrange(desc(consistency_rate)) %>%
  head(5)
```

```
## # A tibble: 5 x 10
##           id avg_steps avg_calories avg_active avg_sedentary days_logged
##        <dbl>     <dbl>        <dbl>      <dbl>         <dbl>       <int>
## 1 1503960366    11641.        1796.       280.          810.          19
## 2 1624580081     4226.        1353.       122.         1278.          19
## 3 1644430081     9275.        2916.       286          1034.          10
## 4 1927972279     2181.        2254        113.          953.          12
## 5 2022484408    12175.        2475.       316.         1059.          12
## # i 4 more variables: activity_ratio <dbl>, active_days <int>,
## #   total_days <int>, consistency_rate <dbl>
```

- These users logged activity every single day they were tracked.
- Some were very active (e.g., ID 2022484408 had 12,175 steps/day), while others were lower (e.g., 1927972279 had just 2,181).
- Even lower-step users still had a consistent habit of using the device — that's key.

2- Top 5 least consistent users:

```
user_activity_summary %>%
  arrange(consistency_rate) %>%
  head(5)
```

```
## # A tibble: 5 x 10
##           id avg_steps avg_calories avg_active avg_sedentary days_logged
##        <dbl>     <dbl>        <dbl>      <dbl>         <dbl>       <int>
## 1 6290855005     1618.        2166.       57.8         1290.          10
## 2 2891001357      774.        2273.      251.          1100.           8
## 3 6391747486     1337.        1763.       40.1         1262.           9
## 4 8253242879     2390.        1463        69.8         1352.          12
## 5 4057192912     1887.        1904.       49.8         1358.          32
## # i 4 more variables: activity_ratio <dbl>, active_days <int>,
## #   total_days <int>, consistency_rate <dbl>
```

- Users like 6290855005 and 2891001357 only used the device on 20–25% of the days they were being tracked.
- Step counts are also extremely low (774–1,600/day).
- These users are both inconsistent and inactive.

Step 4: Create consistency segments I will group them into 3 categories:

- Highly Consistent → used device on 80% of tracked days
- Moderately Consistent → 50–79%
- Low Consistency → < 50%

```r
user_activity_summary <- user_activity_summary %>%
  mutate(consistency_level = case_when(
    consistency_rate >= 0.8 ~ "Highly Consistent",
    consistency_rate >= 0.5 ~ "Moderately Consistent",
    TRUE ~ "Low Consistency"
  ))
```

Then, I will compare avg_steps, avg_active minutes, and avg_sedentary time across these groups:

```r
consistency_groups <- user_activity_summary %>%
  group_by(consistency_level) %>%
  summarize(
    avg_steps = mean(avg_steps, na.rm = TRUE),
    avg_active = mean(avg_active, na.rm = TRUE),
    avg_sedentary = mean(avg_sedentary, na.rm = TRUE),
    avg_activity_ratio = mean(activity_ratio, na.rm = TRUE),
    user_count = n()
  )
```

1- Highly Consistent (27 users) * Most active and engaged: nearly 8,000 steps/day and 4 hours of movement. * High activity ratio shows they consistently break up sedentary time. * Likely already health-conscious and motivated.

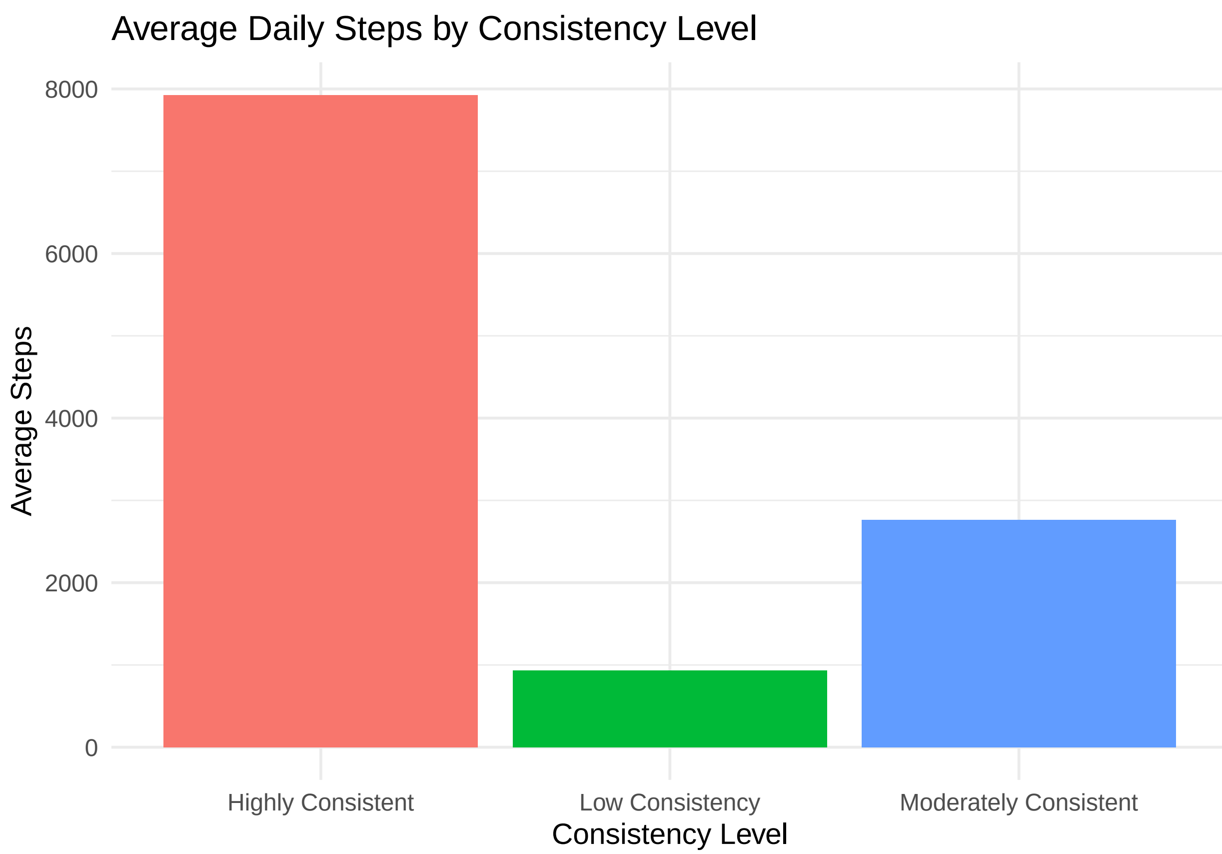
2- Moderately Consistent (4 users) * Around 2,700 steps/day; decent but irregular engagement. * Activity levels suggest potential that's not fully unlocked.

3- Low Consistency (4 users) * Very low activity, high sedentary time. * May have lost motivation or never fully onboarded.

Visual Plots Bar Plot of Average Steps by Consistency

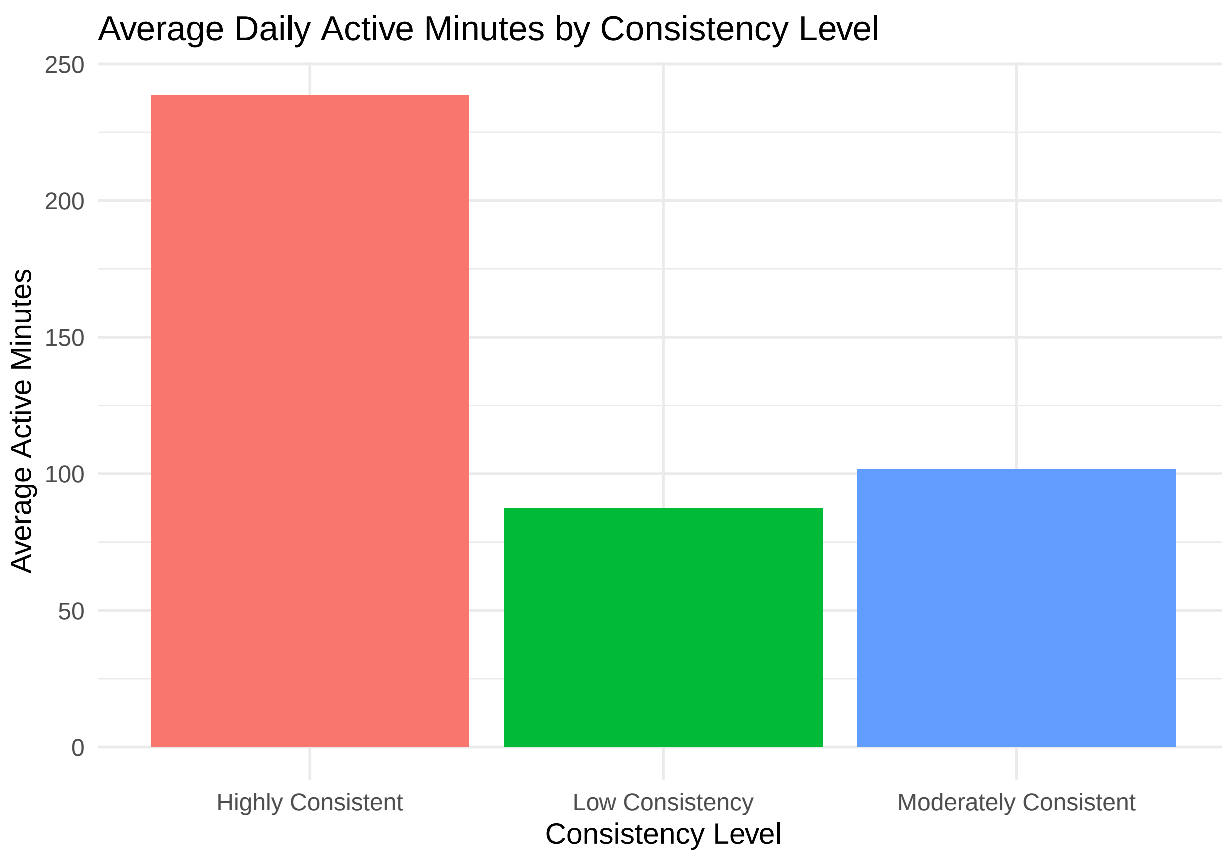
```r
library(ggplot2)

ggplot(consistency_groups, aes(x = consistency_level, y = avg_steps, fill = consistency_level)) +
  geom_col() +
  labs(
    title = "Average Daily Steps by Consistency Level",
    x = "Consistency Level",
    y = "Average Steps"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```



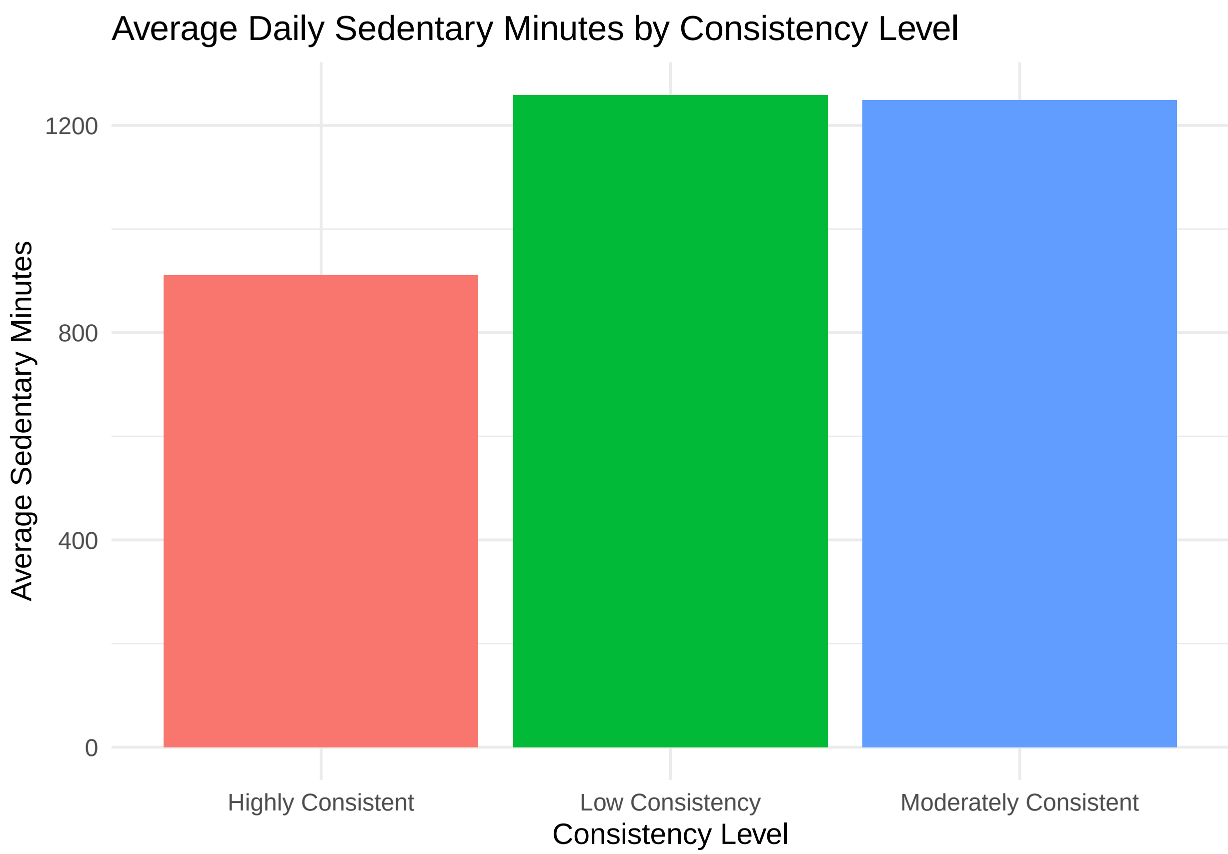
Bar Plot of Active Minutes

```
ggplot(consistency_groups, aes(x = consistency_level, y = avg_active, fill = consistency_level)) +
  geom_col() +
  labs(
    title = "Average Daily Active Minutes by Consistency Level",
    x = "Consistency Level",
    y = "Average Active Minutes"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```



Bar Plot of Sedentary Time

```r
ggplot(consistency_groups, aes(x = consistency_level, y = avg_sedentary, fill = consistency_level)) +
  geom_col() +
  labs(
    title = "Average Daily Sedentary Minutes by Consistency Level",
    x = "Consistency Level",
    y = "Average Sedentary Minutes"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```

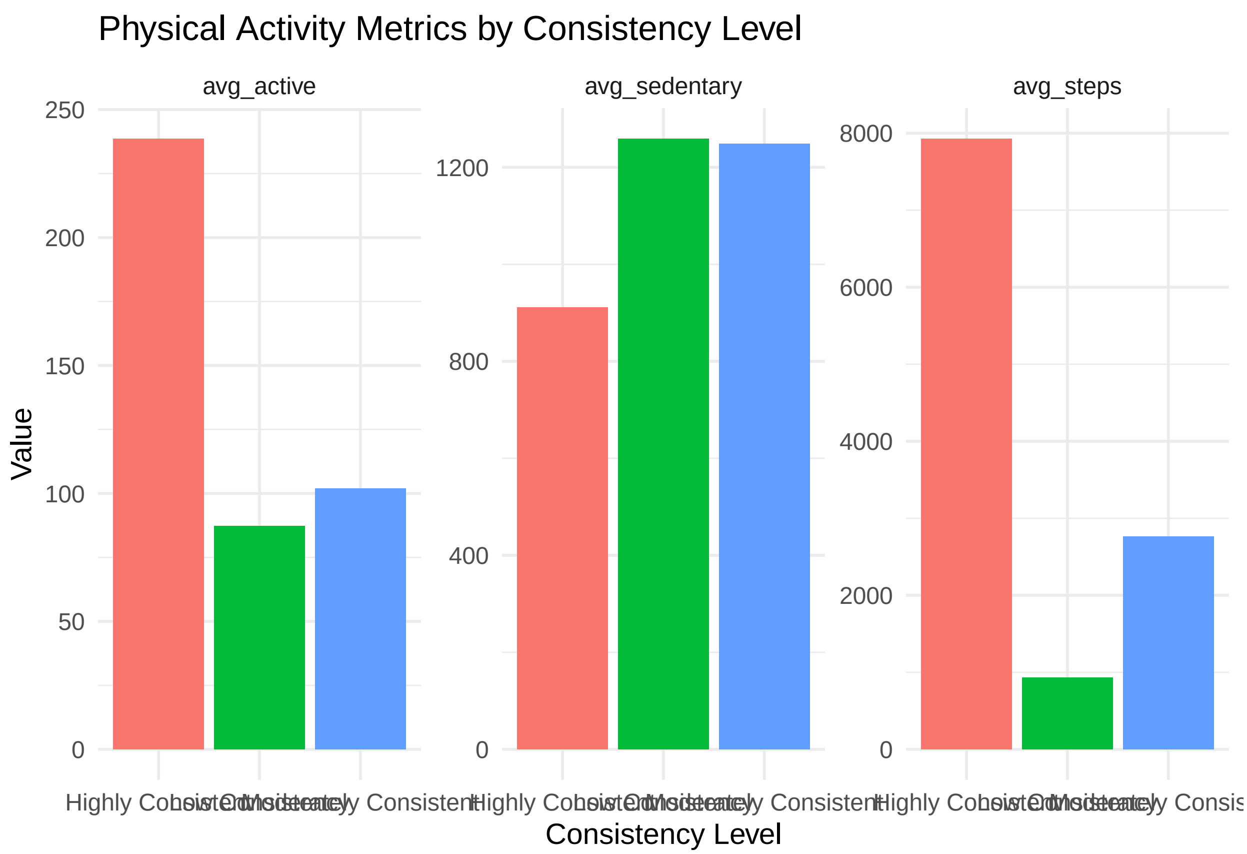


Combine all in one faceted chart

```
library(tidyr)

# Reshape data
consistency_long <- consistency_groups %>%
  select(consistency_level, avg_steps, avg_active, avg_sedentary) %>%
  pivot_longer(cols = c(avg_steps, avg_active, avg_sedentary),
               names_to = "metric", values_to = "value")

# Plot
ggplot(consistency_long, aes(x = consistency_level, y = value, fill = consistency_level)) +
  geom_col() +
  facet_wrap(~metric, scales = "free_y") +
  labs(
    title = "Physical Activity Metrics by Consistency Level",
    x = "Consistency Level",
    y = "Value"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```



These plots show:

- How much more active consistent users are
- The steep drop for low-consistency users
- Where Bellabeat can focus marketing efforts

Sleep Behavior

Sleep Behavior Data Cleaning

Column cleaning and formatting

```
sleep_day_df <- sleep_day %>%
  select(Id, SleepDay, TotalMinutesAsleep, TotalTimeInBed) %>%
  rename(id=Id, sleep_date=SleepDay, minutes_asleep=TotalMinutesAsleep, minutes_in_bed=TotalTimeInBed)
```

Date standardization

The dataset sleep_day includes both date and time values in character format. Using the lubridate package, I parsed all date-related columns and removed the time component using as_date(). I also renamed columns to ensure consistent naming (activity_date) across datasets, which is essential for filtering and joining data accurately.

```r
sleep_day_df <- sleep_day_df %>%
  mutate(activity_date = mdy_hms(sleep_date)) %>%
  mutate(activity_date = as_date(activity_date)) %>%
  select(-sleep_date)
```

Sleep behavior analysis

Per-user summary I will calculate average sleep stats per user:

```r
sleep_summary <- sleep_day_df %>%
  group_by(id) %>%
  summarize(
    avg_minutes_asleep = mean(minutes_asleep, na.rm = TRUE),
    avg_minutes_in_bed = mean(minutes_in_bed, na.rm = TRUE),
    sleep_efficiency = avg_minutes_asleep / avg_minutes_in_bed,
    days_logged = n()
  )
```

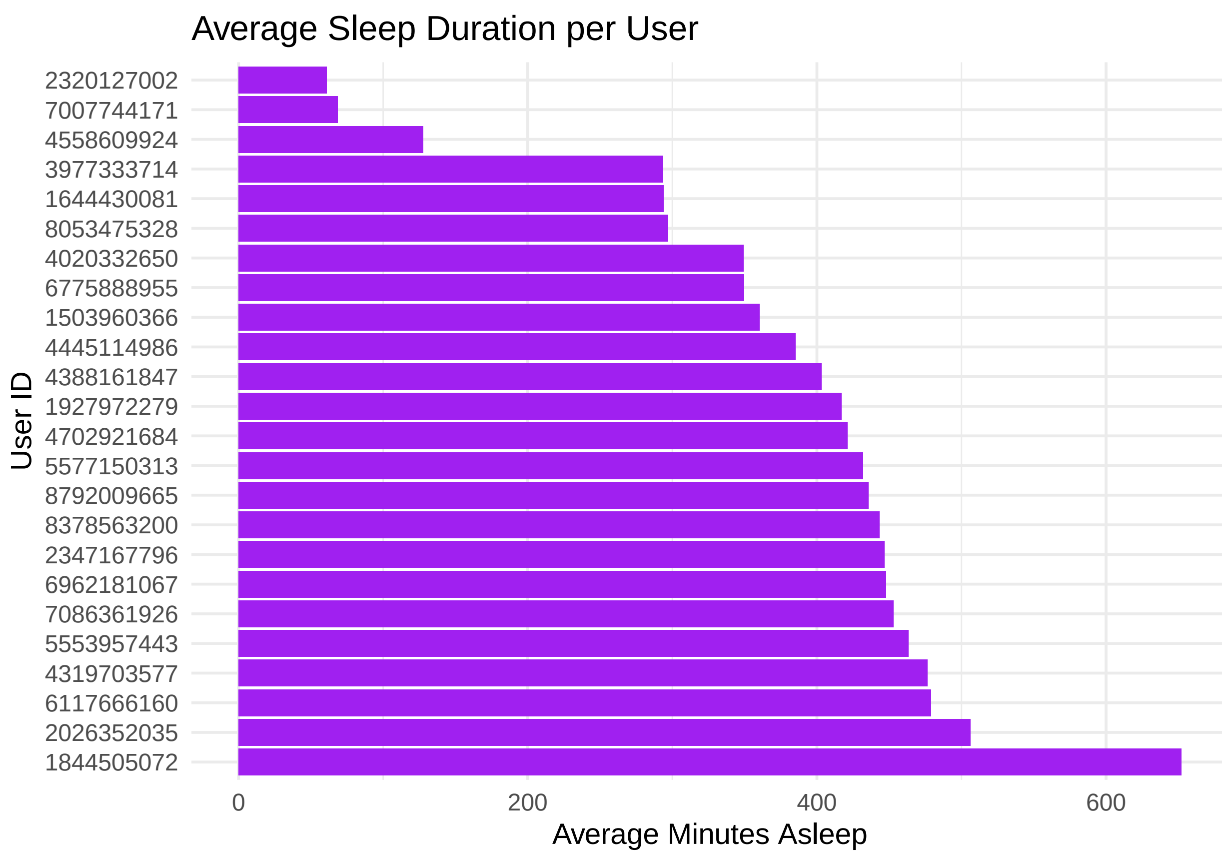
I will preview and then see the top/lowest sleepers:

```r
sleep_summary %>%
  arrange(desc(avg_minutes_asleep)) %>%
  head(5)
```

```
## # A tibble: 5 x 5
##           id avg_minutes_asleep avg_minutes_in_bed sleep_efficiency days_logged
##        <dbl>              <dbl>              <dbl>            <dbl>       <int>
## 1 1844505072               652                961             0.678           3
## 2 2026352035               506.               538.            0.941          28
## 3 6117666160               479.               510.            0.938          18
## 4 4319703577               477.               502.            0.950          26
## 5 5553957443               463.               506.            0.916          31
```

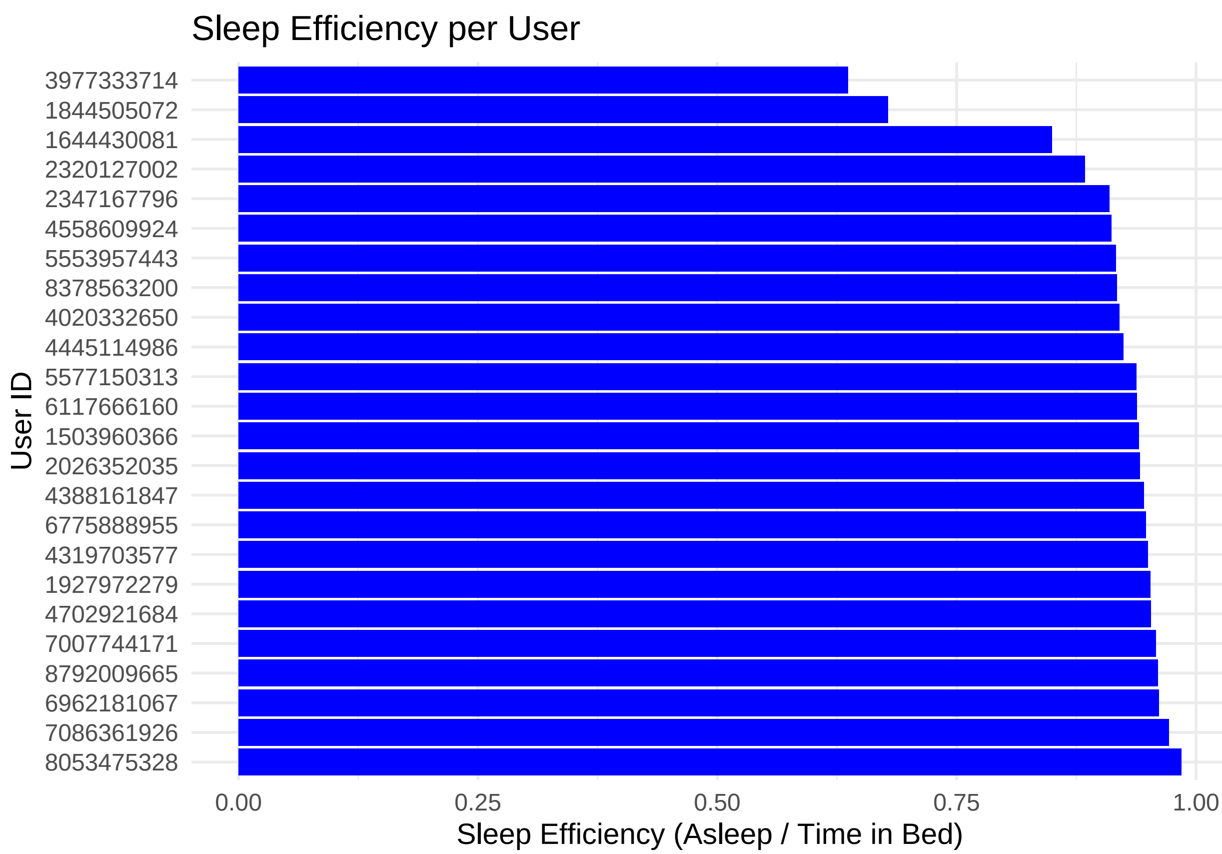
Sleep duration per user plot

```r
ggplot(sleep_summary, aes(x = reorder(as.factor(id), -avg_minutes_asleep), y = avg_minutes_asleep)) +
  geom_col(fill = "purple") +
  labs(
    title = "Average Sleep Duration per User",
    x = "User ID",
    y = "Average Minutes Asleep"
  ) +
  coord_flip() +
  theme_minimal()
```



Sleep efficiency per user plot

```
ggplot(sleep_summary, aes(x = reorder(as.factor(id), -sleep_efficiency), y = sleep_efficiency)) +
  geom_col(fill = "blue") +
  labs(
    title = "Sleep Efficiency per User",
    x = "User ID",
    y = "Sleep Efficiency (Asleep / Time in Bed)"
  ) +
  coord_flip() +
  theme_minimal()
```



The sleep behavior analysis reveals that most users average between 7–8.5 hours of sleep, with a few exceptions. While several users show high sleep efficiency (90%), some users log long time in bed with low sleep duration, indicating lower sleep quality or interrupted rest.

Comparison: Activity vs. sleep

I will check whether more physically active users tend to:

- Sleep longer
- Have better sleep efficiency

Combine the datasets I will join your 'user_activity_summary' and 'sleep_summary' by 'id'.

```
activity_sleep_df <- user_activity_summary %>%
  inner_join(sleep_summary, by = "id")
```

```
cor(activity_sleep_df$avg_active, activity_sleep_df$avg_minutes_asleep, use = "complete.obs")
```

Correlation between active minutes and sleep duration

```
## [1] -0.1456271
```

Correlation: -0.1456

There is a very weak negative correlation between physical activity and sleep duration: *users who are more active don't necessarily sleep more, and might even sleep slightly less.*

```
cor(activity_sleep_df$avg_active, activity_sleep_df$sleep_efficiency, use = "complete.obs")
```

Correlation between activity and sleep efficiency

```
## [1] 0.01599204
```

Correlation: +0.0159

This is a near-zero correlation between average active minutes and sleep efficiency: *more physical activity does not appear to improve sleep efficiency in this dataset.*

Weight Log Analysis

I aim to understand how users engage with weight logging, and whether there are meaningful patterns between weight and activity levels.

Weight Log Data Cleaning

Column cleaning and formatting

```
weight_log_df <- weight_log_info %>%
  select(Id, Date, WeightKg) %>%
  rename(id=Id, weight_kg=WeightKg, log_date=Date)
```

Date standardization

The dataset weight_log_info includes both date and time values in character format. I parsed all date-related columns and removed the time component using as_date(). I also renamed columns to ensure consistent naming (activity_date) across datasets.

```
weight_log_df <- weight_log_df %>%
  mutate(activity_date = mdy_hms(log_date)) %>%
  mutate(activity_date = as_date(activity_date)) %>%
  select(-log_date)
```

Weight tracking per user I will summarize weight tracking per user:

```
weight_summary <- weight_log_df %>%
  group_by(id) %>%
  summarize(
    weigh_ins = n(),
    avg_weight = mean(weight_kg, na.rm = TRUE),
    first_log = min(activity_date),
    last_log = max(activity_date)) %>%
  arrange(desc(weigh_ins))
```

Summary of Insights

Key Findings on Activity Behaviors

The analysis shows that physical activity behavior varies widely among users:

- Users varied significantly in their average daily steps, active minutes, and sedentary time.
- Some are highly active and consistent, averaging over 7,900 steps and 4 hours of movement per day.
- Others exhibit low consistency and minimal engagement with their devices.
- Consistency strongly correlates with higher step counts and lower sedentary time.

Key Findings on Sleep Behavior

1- Sleep duration

- Only a few users average above 7 hours/night (420 mins).
- The highest average was ~10.9 hours, but that user only logged 3 days — not reliable.
- Most consistent users sleep ~7.5–8.5 hours/night.

2- Sleep efficiency (asleep/in bed)

- A healthy efficiency is usually 0.85 (85%).
- Several users had ~94–95% efficiency, indicating quality rest.
- One user had only 67% efficiency, meaning nearly 5 hours in bed without sleeping — may signal restlessness, poor sleep quality, or device misclassification.

3- Sleep duration vs. sleep efficiency vs. physical activity

- Average sleep duration and efficiency ranged widely across users.
- Some users consistently slept well, while others had low sleep efficiency despite longer sleep duration.
- There was no meaningful correlation between physical activity and sleep quality.

Key Findings on Weight Log

Out of all Fitbit users in the dataset, *only 11 users* logged their weight during the 30-day collection period.

- Most users logged weight only once
- Only two users (IDs: 6962181067 and 8877689391) logged weight multiple times (14 and 9 entries respectively)
- Weight logging was concentrated in short time frames, e.g., 1–2 days or a single week

This clearly shows that weight tracking was not a widely adopted behavior among users in this dataset.

Marketing Recommendations

The behavioral insights gathered from Fitbit user data — including patterns in physical activity, sleep, and engagement consistency — suggest clear opportunities for Bellabeat to enhance user engagement by focusing on *consistency, personalization, and habit-building.*

1. Promote Daily Activity Routines * Target users with low consistency and sedentary behavior * Use reminders and motivational nudges to encourage regular movement * Consider campaigns like "10-day Active Streak" or "Afternoon Stretch Alerts"

2. Support Better Sleep Hygiene * Integrate bedtime reminder features or relaxation content in the app * Encourage users to monitor not just sleep duration, but also sleep efficiency * Educate users on the impact of routines, not just movement

3. Encourage Consistent Weight Logging * Gamify weight tracking with badges or progress visuals * Make weight input easier with smart scale integrations * Send weekly reminders for weigh-ins with supportive messaging

These insights indicate that Bellabeat can develop targeted marketing strategies for the **Leaf** product that:

- Emphasize habit-building features for users with inconsistent activity or sleep routines
- Promote the aesthetic and multifunctional appeal of the Leaf to encourage daily wear and logging
- Highlight the integrated wellness tracking (mindfulness, reproductive health) as a unique value for users already familiar with basic activity/sleep tracking

These recommendations are tailored to align with the Leaf's target audience and can help increase both acquisition and retention through personalized, data-informed marketing efforts.

Conclusion

Limitations

This analysis offers exploratory insights based on smart device usage data but is subject to several important limitations:

- **Sample Size & Representation:** The dataset includes only 30 users, which may not adequately represent the broader population of wellness-focused consumers, especially Bellabeat's female target audience.
- **Demographic Gaps:** No gender, age, or location information is available, limiting the ability to tailor recommendations to specific user segments.
- **Outdated Data:** The dataset was collected in 2016, and user behavior or market trends may have shifted significantly since then.
- **Data Provenance:** The data is third-party and not officially published by Fitbit, and thus lacks full verification and documentation.

Future Directions

To strengthen the reliability and business impact of future analyses, the following steps are recommended:

- **Access to Updated and Verified Data:** Partnering with Fitbit or conducting primary research would allow access to more recent and demographically rich datasets.

- **Incorporating User Segmentation:** Including gender, age, and lifestyle variables could enable the design of more targeted campaigns and personalized recommendations.
- **Monitoring Behavioral Trends Over Time:** Longitudinal data would allow Bellabeat to identify seasonal or habit-forming patterns that could inform both product development and marketing strategies.
- **Cross-Device Comparison:** Gathering data from users of Leaf and other competing wellness devices could highlight unique selling points and user engagement differences.

Closing Remarks

This case study explored how consumers use smart wellness devices and applied these insights to Bellabeat's Leaf product. Through analysis of physical activity, sleep patterns, and user consistency, the study identifies opportunities to engage target users with personalized wellness messaging.

While limited in scope, the findings highlight the importance of sustained engagement, balanced activity and rest, and the potential for Bellabeat to further refine its marketing through user-centered data.

With a focus on **holistic well-being** and user habits — rather than raw metrics — Bellabeat is well-positioned to inspire its users to lead healthier, more mindful lives.