---

title: "\*\*Bellabeat Case Study\*\*"

author: "Mukesh Shirke"

date: "2025-05-03"

output:

html\_document:

theme: united

toc: true

toc\_float:

collapsed: true

---

```{r setup, include=FALSE}

knitr::opts\_chunk$set(echo = TRUE, message = FALSE, warning = FALSE)

```

# \*\*Introduction\*\*

Welcome to my Bellabeat data analysis case study! In this project, I took on the role of a junior data analyst at Bellabeat, a high-tech company revolutionizing women’s wellness through products like the Leaf tracker and Ivy+ wearable. I collaborated with cross-functional teams, analyzed real-world health data, and followed the data analysis process-ask, prepare, process, analyze, share, and act-to uncover actionable insights for improving user engagement and product design.

Using Bellabeat’s ecosystem of smart jewelry, hydration trackers, and app-based wellness tools as my focus, I structured my workflow around the Case Study Roadmap tables. These kept me aligned with key tasks, from defining business questions about user activity patterns to visualizing how menstrual cycle data could inform personalized health recommendations.

By the end of this project, I developed a portfolio-ready case study that demonstrates my ability to transform raw data into strategic insights. I’ve included detailed documentation of my process, which I now reference regularly to showcase my analytical skills to potential employers. This case study not only highlights my technical expertise but also my understanding of how data can empower women’s health-a mission central to Bellabeat’s work.

# \*\*Scenario\*\*

As a junior data analyst on Bellabeat’s marketing team, I took on a critical project to unlock growth opportunities for our health-focused smart devices. Bellabeat, a leader in women’s wellness tech, already had a strong foothold with products like the Leaf Urban (a stress-predicting smart necklace) and Spring (a hydration tracker), but Urška Sršen, our cofounder and Chief Creative Officer, challenged me to analyze consumer smart device data to identify untapped potential in the global market.

# \*\*About the company\*\*

Urška Sršen and Sando Mur founded Bellabeat, a high-tech company that manufactures health-focused smart products.

Sršen used her background as an artist to develop beautifully designed technology that informs and inspires women around

the world. Collecting data on activity, sleep, stress, and reproductive health has allowed Bellabeat to empower women with

knowledge about their own health and habits. Since it was founded in 2013, Bellabeat has grown rapidly and quickly positioned

itself as a tech-driven wellness company for women.

By 2016, Bellabeat had opened offices around the world and launched multiple products. Bellabeat products became available

through a growing number of online retailers in addition to their own e-commerce channel on their website. The company has

invested in traditional advertising media, such as radio, out-of-home billboards, print, and television, but focuses on digital

marketing extensively. Bellabeat invests year-round in Google Search, maintaining active Facebook and Instagram pages, and

consistently engages consumers on Twitter. Additionally, Bellabeat runs video ads on Youtube and display ads on the Google

Display Network to support campaigns around key marketing dates.

Sršen knows that an analysis of Bellabeat’s available consumer data would reveal more opportunities for growth. She has

asked the marketing analytics team to focus on a Bellabeat product and analyze smart device usage data in order to gain

insight into how people are already using their smart devices. Then, using this information, she would like high-level

recommendations for how these trends can inform Bellabeat marketing strategy.

# \*\*Step 1: ASK\*\*

### \*\*Business Task\*\*

The business task was to analyse smart device usage data in order to gain insights into how consumers use non-Bellabeat smart devices. Then selecting one Bellabeat product to apply these insights to in the presentation. The business questions are:

\*\*1. What are some trends in smart device usage?\*\*

\*\*2. How could these trends apply to Bellabeat customers?\*\*

\*\*3. How could these trends help influence Bellabeat marketing strategy?\*\*

### \*\*Stakeholders\*\*

• Urška Sršen: Bellabeat cofounder and Chief Creative Officer

• Sando Mur: Cofounder, Mathematician, Key member of the executive team

• marketing analytics team: Responsible for collecting, analysing and reporting data to guide marketing strategy. I’m a junior data analyst in the team.

### \*\*Products\*\*

• \*\*Bellabeat app\*\*: The app provides users with data about their activity, sleep, stress, menstrual cycle and mindfulness habits and helps people better understand their heath and habits. The app connects to their line of smart wellness products.

• \*\*Leaf\*\*: a wellness tracker can be worn as a bracelet, necklace and clip. Leaf can connect with Bellabeat app to track activity, sleep and stress.

• \*\*Time\*\*: A wellness watch can track activity, sleep, and stress and connect with Bellabeat app. A watch provides you with insights about your daily wellness.

• \*\*Spring\*\*: A water bottle tracks daily water intake to ensure you are hydrated enough daily by applying the state-of-art smart technology. Spring can connect with the app and track your hydration level.

• \*\*Bellabeat membership\*\*: a subscription-based program for users 24/7 access to fully personalized guidance on nutrition, activity, sleep, health and beauty, and mindfulness based on their lifestyle and goals.

# \*\*Step 2: PREPARE\*\*

### \*\*Data Selection\*\*

The data for this analysis is the Fitbit Fitness Tracker Data stored on Kaggle and made available by Mobius. This dataset is under CC0: Public Domain license meaning the creator has waived his right to the work under copyright law.

[Visit fitbit dataset on Kaggle](https://www.kaggle.com/datasets/arashnic/fitbit) (CC0: Public Domain, dataset made available through Mobius)

This dataset contains personal health data from 30 fitbit users. All users consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. It includes data about daily activity, steps and heart rate that can be used to explore users’ habits. Totally, I downloaded 18 CSV files and stored them properly.

This data set contains a personal ﬁtness tracker from thirty ﬁtbit users. It includes information about daily activity, steps, and heart rate that can be used to explore users’ habits. It was generated via a survey through Amazon Mechanical Turk between 12.03.2016 and 12.05.2016. Thirty eligible Fitbit users consented to submit their tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. Variation between output represents different Fitbit trackers and individual tracking behaviors/preferences.

The dataset has the following 18 files in .csv format.

• dailyActivity\_merged.csv

• dailyCalories\_merged.csv

• dailyIntensities\_merged.csv

• dailySteps\_merged.csv

• heartrate\_seconds\_merged.csv

• hourlyCalories\_merged.csv

• hourlyIntensities\_merged.csv

• hourlySteps\_merged.csv

• minuteCaloriesNarrow\_merged.csv

• minuteCaloriesWide\_merged.csv

• minuteIntensitiesNarrow\_merged.csv

• minuteIntensitiesWide\_merged.csv

• minuteMETsNarrow\_merged.csv

• minuteSleep\_merged.csv

• minuteStepsNarrow\_merged.csv

• minuteStepsWide\_merged.csv

• sleepDay\_merged.csv

• weightLogInfo\_merged

### \*\*Limitations of this Data\*\*

• Sample size: the dataset collects the information from 30 individuals, which is a small sample size to determine behavior patterns of a population of people who use smart devices

• The absence of essential characteristics of the participants might have a significant impact on the data collected, such as gender, age, and lifestyle

• Third-party data collected using Amazon Mechanical Turk.

• The data was collected six years ago. It is not as current as it could be for the analysis.

### \*\*ROCCC analysis\*\*

• \*\*Reliable\*\*: low — the data collected from users without demographic information

• \*\*Originality\*\*: low — the data was collected from third-party Amazon Mechanical Turk

• \*\*Comprehensive\*\*: high — the data contained personal health data which allowed me to answer business questions

• \*\*Current\*\*: low — the respondents were generated during 03.12.2016–05.12.2016.

• \*\*Cited\*\*: high — the data source was well-documented.

### \*\*Environment\*\*

\*\*R Studio\*\* was used to complete this analysis because it has many excellent data processing and visualization features.

\*\*Install packages\*\*

```{r libs, message=FALSE}

# Package names

packages <- c(

"tidyverse",

"here",

"knitr", "rmarkdown",

"ggplot2", "dplyr", "tidyr", "janitor",

"lubridate",

"todor","lintr",

"DT", "kableExtra",

"roxygen2", "testthat",

"ggpubr"

)

#Install packages not yet installed

installed\_packages <- packages %in% rownames(installed.packages())

if (any(installed\_packages == FALSE)) {

install.packages(packages[!installed\_packages])

}

# Packages loading

lapply(packages, library, character.only = TRUE) |>

invisible()

```

\*\*Set Working Directory\*\*

```{r, results= FALSE}

getwd()

```

\*\*Import Data\*\*

For this analysis, I imported three key datasets:

• dailyActivity\_merged.csv

• sleepDay\_merged.csv

• hourlySteps\_merged.csv

```{r}

Daily\_Activity <- read.csv("dailyActivity\_merged.csv", header = TRUE, sep = ",")

Daily\_Sleep <- read.csv("sleepDay\_merged.csv", header = TRUE, sep = ",")

Hourly\_Steps <- read.csv("hourlySteps\_merged.csv", header = TRUE, sep = ",")

```

\*\*Preview and Understand Data Structure\*\*

```{r, results= FALSE}

head(Daily\_Activity)

str(Daily\_Activity)

head(Daily\_Sleep)

str(Daily\_Sleep)

head(Hourly\_Steps)

str(Hourly\_Steps)

```

### \*\*Problems Identified in the Data\*\*

• \*\*Duplicate and Missing Entries\*\*:

The datasets contain duplicate records and missing (NA) values, which can lead to inaccurate analysis if not addressed.

• \*\*Inconsistent Date and Time Formats\*\*:

Date and time columns appear in multiple formats across datasets, making it difficult to merge and compare records without standardization.

• \*\*Mismatched Data Types for Primary Keys\*\*:

Key columns used for merging datasets (such as id and date fields) have inconsistent data types (e.g., character vs. Date), which can cause issues during data integration.

These issues must be resolved during the data preparation phase to ensure the reliability and accuracy of subsequent analyses.

### \*\*Data Cleaning\*\*

\*\*Identify and Remove Duplicates & Missing Values\*\*

```{r}

sum(duplicated(Daily\_Activity))

sum(duplicated(Daily\_Sleep))

sum(duplicated(Hourly\_Steps))

```

```{r}

Daily\_Activity <- Daily\_Activity %>%

distinct() %>%

drop\_na()

Daily\_Sleep <- Daily\_Sleep %>%

distinct() %>%

drop\_na()

Hourly\_Steps <- Hourly\_Steps %>%

distinct() %>%

drop\_na()

```

\*\*Standardize Column Names\*\*

Convert all column names to lowercase for consistency and easier data manipulation.

```{r, results= FALSE}

clean\_names(Daily\_Activity)

Daily\_Activity <- rename\_with(Daily\_Activity, tolower)

clean\_names(Daily\_Sleep)

Daily\_Sleep <- rename\_with(Daily\_Sleep, tolower)

clean\_names(Hourly\_Steps)

Hourly\_Steps <- rename\_with(Hourly\_Steps, tolower)

```

\*\*Format Date and Time Columns\*\*

• Convert date columns to a standard format in each dataset.

• Parse and format time columns as required for later analysis.

```{r}

Daily\_Activity <- Daily\_Activity %>%

mutate(activitydate = format(parse\_date\_time(activitydate, orders = c("mdy", "mdY")), "%Y-%m-%d"))

```

```{r}

Daily\_Sleep <- Daily\_Sleep %>%

mutate(

sleepday = as.Date(

parse\_date\_time(sleepday, orders = c("mdy IMS p", "mdy HM"))

)

)

```

```{r}

Hourly\_Steps <- Hourly\_Steps %>%

mutate(

activityhour = as.POSIXct(activityhour, format = "%m/%d/%Y %I:%M:%S %p", tz=Sys.timezone()))

```

\*\*Prepare for Merging\*\*

\*\*• Rename columns as needed for consistency (e.g., 'sleepday' to 'activitydate').\*\*

```{r}

Daily\_Sleep <- Daily\_Sleep %>%

rename(activitydate = sleepday)

```

\*\*• Ensure key columns ('id' and 'activitydate') have the same data type in both datasets.\*\*

```{r}

str(Daily\_Activity)

str(Daily\_Sleep)

```

```{r}

Daily\_Activity$activitydate <- as.Date(Daily\_Activity$activitydate)

Daily\_Sleep$activitydate <- as.Date(Daily\_Sleep$activitydate)

```

\*\*Merge Datasets\*\*

Combine daily activity and sleep datasets into a single dataframe using 'id' and 'activitydate' as primary keys.

```{r}

Daily\_Activity\_Sleep <- merge(Daily\_Activity, Daily\_Sleep, by=c ("id", "activitydate"))

```

\*\*Create New Variables\*\*

```{r}

Daily\_Activity\_Sleep <- Daily\_Activity\_Sleep %>%

mutate(totalactiveminutes = veryactiveminutes + fairlyactiveminutes + lightlyactiveminutes + sedentaryminutes)

```

# \*\*Step 3: PROCESS\*\*

\*\*Further Data Transformation\*\*

• Separate date and time in the hourly steps dataset for time-based analysis.

• Replace missing time values with a default (e.g., "00:00:00").

```{r}

Hourly\_Steps <- Hourly\_Steps %>%

separate(activityhour, into = c("date", "time"), sep = " ") %>%

mutate(date = ymd(date))

Hourly\_Steps <- Hourly\_Steps %>%

mutate(time = if\_else(is.na(time), "00:00:00", time))

```

\*\*Summarizing the Data\*\*

Determine min, max, and average values for the variables of Daily\_Activity\_Sleep data frame.

```{r}

Daily\_Activity\_Sleep %>%

select(totalsteps, totaldistance, veryactivedistance, moderatelyactivedistance, lightactivedistance, sedentaryactivedistance, calories, totalminutesasleep, totalactiveminutes) %>%

summary()

```

```{r}

DT::datatable(

Daily\_Activity\_Sleep,

options = list(scrollX = TRUE),

caption = "Dataset 1 : Merged Daily Activity and Sleep Data"

)

```

```{r}

DT::datatable(

Hourly\_Steps,

caption = "Table 2 : Hourly Steps (average)"

)

```

# \*\*Step 4: ANALYSE\*\*

### \*\*4.1 User Type Classification (calculating average steps, categorizing users)\*\*

• Calculate average daily steps for each user.

• Categorize users into 'very active', 'fairly active', 'lightly active', or 'sedentary' based on their average steps.

```{r}

Daily\_Average <- Daily\_Activity\_Sleep %>%

group\_by(id) %>%

summarise (mean\_daily\_steps = mean(totalsteps), mean\_daily\_calories = mean(calories), mean\_daily\_sleep = mean(totalminutesasleep))

Daily\_Average <- Daily\_Average %>%

mutate(

user\_type = case\_when(

mean\_daily\_steps > 10000 ~ "very active",

mean\_daily\_steps > 7500 ~ "fairly active",

mean\_daily\_steps > 5000 ~ "lightly active",

TRUE ~ "sedentary"

)

)

```

• Calculate the percentage of users in each activity category.

```{r}

user\_type\_percent <- Daily\_Average %>%

group\_by(user\_type) %>%

summarise(user\_count = n\_distinct(id)) %>%

mutate(percentage = paste0(round(100 \* user\_count / sum(user\_count), 2), " %"))

```

```{r}

DT::datatable(

head(user\_type\_percent),

caption = "Table 1 : User Type with Percentage",

width = "60%"

)

```

### \*\*4.2 Hourly Steps Throughout The Day (analyzing activity patterns)\*\*

```{r}

Hourly\_Steps\_summary <- Hourly\_Steps %>%

group\_by(time) %>%

summarize(average\_steps = mean(steptotal))

```

```{r}

htmltools::div(

style = "text-align:center;",

DT::datatable(

Hourly\_Steps\_summary,

caption = "Table 2 : Hourly Steps (average)",

width = "60%"

) %>%

DT::formatRound('average\_steps', 2) %>%

DT::formatStyle(

columns = 'average\_steps',

'text-align' = 'center'

)

)

```

### \*\*4.3 Correlation (examining relationships of daily steps with daily sleep and calories)\*\*

\*\*Daily steps vs daily sleep Correlation Value\*\*

```{r}

round(cor(Daily\_Activity\_Sleep$totalsteps, Daily\_Activity\_Sleep$totalminutesasleep, use = "complete.obs"), 2)

```

\*\*Daily steps vs calories Correlation Value\*\*

```{r}

round(cor(Daily\_Activity\_Sleep$totalsteps, Daily\_Activity\_Sleep$calories, use = "complete.obs"), 2)

```

### \*\*4.4 Correlation (examining relationships of Calories with daily steps, total distance and daily sleep)\*\*

\*\*Daily steps vs calories Correlation Value\*\*

```{r}

round(cor(Daily\_Activity\_Sleep$totalsteps, Daily\_Activity\_Sleep$calories, use = "complete.obs"), 2)

```

\*\*Total distance vs calories Correlation Value\*\*

```{r}

round(cor(Daily\_Activity\_Sleep$totaldistance, Daily\_Activity\_Sleep$calories, use = "complete.obs"), 2)

```

\*\*Daily sleep vs Calories sleep Correlation Value\*\*

```{r}

round(cor(Daily\_Activity\_Sleep$totalminutesasleep, Daily\_Activity\_Sleep$calories, use = "complete.obs"), 2)

```

# \*\*Step 5: SHARE\*\*

### \*\*Visualize Key Insights\*\*

\*\*5.1 User Type Classification\*\*

\*Pie chart for User Type Distribution\*

```{r}

user\_type\_percent %>%

ggplot(aes(x = "", y = percentage, fill = user\_type)) +

geom\_bar(stat = "identity", width = 1, color = "white") +

coord\_polar("y", start = 0) +

theme\_minimal(base\_size = 12) +

theme(

axis.title.x = element\_blank(),

axis.title.y = element\_blank(),

panel.border = element\_blank(),

panel.grid = element\_blank(),

axis.ticks = element\_blank(),

axis.text.x = element\_blank(),

plot.title = element\_text(hjust = 0.5, size = 14, face = "bold"),

legend.position = "right"

) +

scale\_fill\_manual(values = c("#BB3E00", "#657C6A", "#F7AD45", "#A2B9A7")) +

geom\_text(aes(label = percentage),

position = position\_stack(vjust = 0.5), size = 4, color = "white", fontface = "bold") +

labs(title = "User Type Distribution", fill = "User Type")

```

\*\*5.2 Hourly Steps Throughout The Day\*\*

\*Bar chart for Hourly Steps Throughout The Day\*

```{r}

Hourly\_Steps %>%

group\_by(time) %>%

summarize(average\_steps = mean(steptotal)) %>%

ggplot(aes(x = time, y = average\_steps, fill = average\_steps)) +

geom\_col(color = "white") +

labs(title = "Hourly Steps Throughout the Day", x = "", y = "") +

scale\_fill\_gradient(low = "#A2B9A7", high = "#BB3E00") +

theme\_minimal(base\_size = 12) +

theme(

axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust = 1),

axis.title.x = element\_blank(),

axis.title.y = element\_blank(),

panel.border = element\_blank(),

panel.grid = element\_blank(),

axis.ticks = element\_blank(),

plot.title = element\_text(hjust = 0.5, size = 14, face = "bold"),

legend.position = "right"

)

```

\*\*5.3 Comparative Analysis of daily steps with daily sleep and calories\*\*

```{r, fig.width=10, fig.height=6}

ggarrange(

ggplot(Daily\_Activity\_Sleep, aes(x = totalsteps, y = totalminutesasleep)) +

geom\_jitter(color = "#657C6A") +

geom\_smooth(color = "#BB3E00", fill = "#F7AD45") +

labs(

title = "Daily steps vs daily sleep",

x = "Daily steps", y = "daily sleep",

subtitle = paste("Correlation Value = ",

round(cor(Daily\_Activity\_Sleep$totalsteps, Daily\_Activity\_Sleep$totalminutesasleep, use = "complete.obs"), 2))

) +

theme(panel.background = element\_blank(), plot.title = element\_text(size = 14)

),

ggplot(Daily\_Activity\_Sleep, aes(x = totalsteps, y = calories)) +

geom\_jitter(color = "#657C6A") +

geom\_smooth(color = "#BB3E00", fill = "#F7AD45") +

labs(

title = "Daily steps vs Calories",

x = "Daily steps", y = "Calories",

subtitle = paste("Correlation Value = ",

round(cor(Daily\_Activity\_Sleep$totalsteps, Daily\_Activity\_Sleep$calories, use = "complete.obs"), 2))

) +

theme(panel.background = element\_blank(), plot.title = element\_text(size = 14)

)

)

```

\*\*5.4 Comparative Analysis of Calories with daily steps, total distance and daily sleep\*\*

```{r, fig.width=10, fig.height=12}

ggarrange(

ggplot(Daily\_Activity\_Sleep, aes(x = totalsteps, y = calories)) +

geom\_jitter(color = "#657C6A") +

geom\_smooth(color = "#BB3E00", fill = "#F7AD45") +

labs(

title = "Daily steps vs Calories",

x = "Daily steps", y = "Calories",

subtitle = paste("Correlation Value = ",

round(cor(Daily\_Activity\_Sleep$totalsteps, Daily\_Activity\_Sleep$calories, use = "complete.obs"), 2))

) +

theme(panel.background = element\_blank(), plot.title = element\_text(size = 14)

),

ggplot(Daily\_Activity\_Sleep, aes(x = totaldistance, y = calories)) +

geom\_jitter(color = "#657C6A") +

geom\_smooth(color = "#BB3E00", fill = "#F7AD45") +

labs(

title = "Daily distance vs Calories",

x = "Daily distance", y = "Calories",

subtitle = paste("Correlation Value = ",

round(cor(Daily\_Activity\_Sleep$totaldistance, Daily\_Activity\_Sleep$calories, use = "complete.obs"), 2))

) +

theme(panel.background = element\_blank(), plot.title = element\_text(size = 14)

),

ggplot(Daily\_Activity\_Sleep, aes(x = totalsteps, y = totalminutesasleep)) +

geom\_jitter(color = "#657C6A") +

geom\_smooth(color = "#BB3E00", fill = "#F7AD45") +

labs(

title = "Daily sleep vs Calories",

x = "Daily sleep", y = "Calories",

subtitle = paste("Correlation Value = ",

round(cor(Daily\_Activity\_Sleep$totalminutesasleep, Daily\_Activity\_Sleep$calories, use = "complete.obs"), 2))

) +

theme(panel.background = element\_blank(), plot.title = element\_text(size = 14)

)

)

```

### \*\*Insights & Key Findings\*\*

\*\*1. User Activity Levels \*\*

\*\*• Distribution\*\*:

Users were categorized into four groups based on average daily steps: very active, fairly active, lightly active, and sedentary.

\*\*• Key Finding\*\*:

The majority of users fall into the "fairly active" and "lightly active" categories, while a smaller proportion are "very active" or "sedentary".

\*\*• Implication\*\*:

There is an opportunity to encourage lightly active and fairly active users to increase their activity and move into higher activity categories.

\*\*2. Hourly Activity Patterns\*\*

\*\*• Observation\*\*:

The bar chart of hourly steps shows that users are least active during late night and early morning hours, with activity increasing after 7 AM, peaking in the late afternoon and early evening, and then declining at night.

\*\*• Key Finding\*\*:

Most steps are taken during the day, especially between 9 AM and 7 PM.

\*\*• Implication\*\*:

Wellness programs or app notifications could be scheduled to motivate users during periods of low activity, such as mid-morning or early afternoon.

\*\*3. Comparative Analysis of daily steps with daily sleep and calories\*\*

\*\*• Observation\*\*:

The relationship between steps and calories (0.41) is substantially stronger than between steps and sleep (-0.19).

\*\*• Key Finding\*\*:

Physical activity has a more direct and predictable impact on calorie expenditure than on sleep patterns.

\*\*• Implication\*\*:

Wellness recommendations should emphasize the reliable connection between activity and calorie burn, while acknowledging that sleep optimization may require consideration of additional factors beyond step count.

\*\*4. Comparative Analysis of Calories with daily steps, total distance and daily sleep\*\*

\*\*• Observation\*\*:

Daily distance vs Calories shows the strongest positive correlation (0.52), indicating that distance covered is the best predictor of calories burned among the three metrics.

Daily steps vs Calories also demonstrates a moderate positive correlation (0.41), meaning more steps generally lead to more calories burned, but the relationship is not as strong as with distance.

Daily sleep vs Calories displays almost no correlation (-0.03), suggesting that sleep duration has virtually no impact on calorie expenditure in this dataset.

\*\*• Key Finding\*\*:

Both activity-based metrics (steps and distance) are positively and significantly related to calorie burn, with distance being the more reliable predictor.

Sleep duration does not meaningfully influence calories burned, as shown by the flat trend and near-zero correlation.

\*\*• Implication\*\*:

For users and product design: Focusing on increasing daily distance and step count can effectively help users boost their calorie burn. Features or challenges that track and encourage these activities are likely to be most impactful for fitness and weight management goals.

For wellness recommendations: Calorie management strategies should be based on activity metrics rather than sleep duration. Sleep guidance should focus on other health outcomes, as it does not directly affect calorie expenditure in this population.

\*\*5. Data Quality and Limitations\*\*

\*\*• Data Quality\*\*:

The dataset required cleaning for duplicates, missing values, and inconsistent formats, which was addressed in the preparation steps.

\*\*• Limitations\*\*:

The data comes from a small sample (30 users), lacks demographic diversity, and was collected several years ago. Insights should be interpreted with caution for broader populations.

# \*\*Step 6: ACT\*\*

### \*\*Recommendations for Bellabeat\*\*

\*\*1. Target "Fairly Active" and "Lightly Active" Users\*\*

\*\*• Insight\*\*: The majority of users are "fairly active" or "lightly active," with fewer in the "very active" or "sedentary" categories.

\*\*• Action\*\*:

• Develop personalized challenges, motivational messages, or rewards within the Bellabeat app to encourage these users to increase their daily step count and transition to higher activity categories.

• Use push notifications or in-app reminders to nudge lightly active users during times of low activity.

\*\*2. Leverage Hourly Activity Patterns\*\*

\*\*• Insight\*\*: Step counts are lowest during late night and early morning, peaking in the late afternoon and evening.

\*\*• Action\*\*:

• Schedule app notifications, wellness tips, or mini-challenges during mid-morning or early afternoon to boost activity during low-movement periods.

• Consider offering guided walks or stretch breaks through the app during these times.

\*\*3. Capitalize on Step-Calories Relationship\*\*

\*\*• Insight\*\*: There is a moderate positive correlation between steps and calories burned.

\*\*• Action\*\*:

• Emphasize calorie-burn tracking and goal-setting features in marketing and within the app.

• Provide users with feedback on how increasing their daily steps can help them achieve calorie or weight-management goals.

\*\*4. Enhance Sleep Recommendations\*\*

\*\*• Insight\*\*: There is only a weak negative correlation between steps and sleep duration.

\*\*• Action\*\*:

• When providing sleep advice, incorporate additional factors beyond daily activity (e.g., stress, hydration, bedtime routines).

• Use Bellabeat’s holistic wellness features to offer more comprehensive sleep health guidance.

\*\*5. Acknowledge Data Limitations\*\*

\*\*• Insight\*\*: The dataset is small, not demographically diverse, and a few years old.

\*\*• Action\*\*:

• Supplement these findings with ongoing data collection from Bellabeat’s own user base for more representative and up-to-date insights.

• Use these initial findings as a starting point for further research and product development.

\*\*6. Marketing and Engagement Strategies\*\*

\*\*• Action\*\*:

• Highlight the app’s ability to track and improve activity and calorie burn in marketing campaigns.

• Share user success stories and testimonials to motivate community engagement.

• Consider partnerships or challenges that encourage users to be active during typically low-activity hours.

# \*\*My Take\*\*

### What I Learned from the Bellabeat Case Study

Working on the Bellabeat case study was a valuable experience that deepened both my technical and business analytics skills. Here’s what I learned:

\*\*1. The Power of Structured Analysis\*\*

I experienced firsthand the importance of following a systematic data analysis process-Ask, Prepare, Process, Analyze, Share, and Act. This framework kept my work organized and ensured that every step, from defining business questions to delivering actionable recommendations, was purposeful and aligned with stakeholder needs.

\*\*2. Data Cleaning is Critical\*\*

Real-world data is rarely perfect. I encountered issues like duplicates, missing values, inconsistent formats, and mismatched data types. Addressing these challenges taught me that thorough data cleaning is foundational for any reliable analysis and that attention to detail at this stage saves time and prevents errors later.

\*\*3. Transforming Data into Insights\*\*

I learned how to merge and transform multiple datasets to create a unified view of user activity, sleep, and steps. Summarizing and visualizing this data-such as classifying users by activity level or analyzing hourly step patterns-showed me how raw numbers can be turned into business-relevant insights.

\*\*4. The Value of Visualization\*\*

Creating clear, compelling visuals (like pie charts for user types and bar charts for hourly activity) made it easier to communicate findings to non-technical stakeholders. I saw how effective visualizations can drive understanding and support decision-making.

\*\*5. Understanding Business Impact\*\*

Beyond technical analysis, I learned to connect insights to business actions. For example, identifying when users are least active led to recommendations for targeted notifications, and understanding user activity levels informed ideas for product engagement strategies.

\*\*6. Recognizing Data Limitations\*\*

This project highlighted the importance of critically evaluating data sources. The Fitbit dataset’s small sample size and lack of demographic diversity reminded me that all analyses have limitations, and that transparency about these is essential for responsible data science.

\*\*7. Portfolio and Communication Skills\*\*

Documenting my process in detail not only helped me reflect on my work but also created a portfolio-ready case study I can share with potential employers. It reinforced the importance of clear communication-both in code and in business reporting.

\*\*In summary:\*\*

This case study strengthened my technical skills in R and data wrangling, my ability to extract actionable insights from complex data, and my understanding of how analytics can directly support business goals in the health tech industry. Most importantly, it showed me how data can be used to empower better health outcomes for real people-a mission I’m excited to pursue in my future career.