# Bellabeat Case Analysis

Leo Pham

2022-10-15

# BELLABEAT CASE ANALYSIS

## INTRODUCTION

### 1. Overview

Bellabeat, a high-tech manufacturer of health-focused smart products for women. Since 2013, Bellabeat has inspired women around the world to gain knowledge in improving one's health. Although Bellabeat has grown rapidly and quickly positioned itself as a large player in the global smart device market, Urška Sršen believes that analyzing smart fitness data can unlock new opportunities and insights for the company's business outlook.

### 2. Products

- Bellabeat App: The Bellabeat app provides users with health data related to their activity, sleep, stress, menstrual cycle, and mindfulness habits. This data can help users better understand their current habits and make healthy decisions. The Bellabeat app connects to their line of smart wellness products.
- Leaf: Bellabeat's classic wellness tracker can be worn as a bracelet, necklace, or clip. The Leaf tracker connects to the Bellabeat app to track activity, sleep, and stress.

## ASK

### 1. Business task

This project analyzes smart device usage data to gain insight into how people are already using their smart devices and give recommendations on how these trends can inform Bellabeat marketing strategy.

### 2. Stakeholders

- Urška Sršen: Bellabeat's co-founder and Chief Creative Officer.
- Sando Mur: Mathematician and Bellabeat's cofounder, a key member of the Bellabeat executive team.
- Bellabeat marketing analytics team: A team of data analysts responsible for collecting, analyzing, and reporting data that helps guide Bellabeat's marketing strategy. I am a junior data analyst who can help Bellabeat achieve its missions and goals.

## PREPARE

### 1. Data description

- Data source: FitBit Fitness Tracker Data Link (https://www.kaggle.com/datasets/arashnic/fitbit).
- Data organization: 18 CSV files organized in a long data format.
- Sample size: 33 observations.
- Data duration: 2016-03-12 to 2016-05-12.
- Data credibility: 33 Fitbit users consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. Since the data was collected by a third party, it

is difficult to verify the reliability of it. Additionally, the data contains information from only 33 current Fitbit users, this could create a sampling bias.
- Data license: CC0 Public Domain.

### 2. Previewing data

Before importing the data, I used Excel to preview all the datasets and realized that those tables covered a data scale from days to minutes.

In this project, I am focusing on analyzing daily data. Therefore, our research is conducted on dailyActivity, dailyCalories, dailyIntensities, dailySteps, sleepDay, and weightLogInfo.

Additionally, dailyCalories, dailyIntensities, and dailySteps are the three subsets of dailyActivity so it is possible to remove these three tables from my database. As a result, I perform data cleaning on three datasets dailyActivity, sleepDay, and weightLogInfo.

# PROCESS

I decide to use SQL for data cleaning because dailyActivity, sleepDay, and weightLogInfo are relational databases. The process will be starting with importing the CSV files to SQL Management Studio, and ending up with 2 cleaned data files: Weight.csv and SleepAndActivity.csv.

Document of the cleaning process: Cleaning Documentation.txt.

# ANALYZE

Data analyzing and data visualization are carried out on Rstudio.

Let's install and load some packages: tidyverse, tidyr, dplyr, and ggplot2.

```
library("tidyverse")
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✓ ggplot2 3.4.0      ✓ purrr   0.3.5
## ✓ tibble  3.1.8      ✓ dplyr   1.0.10
## ✓ tidyr   1.2.1      ✓ stringr 1.4.1
## ✓ readr   2.1.3      ✓ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
```

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

After that, I import the Weight and SleepAndActivity datasets.

```
sleep_and_activity <- read.csv("E:/Downloads/SleepAndActivity.csv")
weight <- read.csv("E:/Downloads/Weight.csv")
```

```
#Previewing SleepAndActivity
head(sleep_and_activity)
```

```
##           id      date total_steps total_distance tracker_distance
## 1 1503960366 4/12/2016       13162           8.50             8.50
## 2 1503960366 4/13/2016       10735           6.97             6.97
## 3 1503960366 4/15/2016        9762           6.28             6.28
## 4 1503960366 4/16/2016       12669           8.16             8.16
## 5 1503960366 4/17/2016        9705           6.48             6.48
## 6 1503960366 4/19/2016       15506           9.88             9.88
##   logged_activities_distance very_active_distance moderately_active_distance
## 1                          0                1.88                       0.55
## 2                          0                1.57                       0.69
## 3                          0                2.14                       1.26
## 4                          0                2.71                       0.41
## 5                          0                3.19                       0.78
## 6                          0                3.53                       1.32
##   light_active_distance sedentary_active_distance very_active_minutes
## 1                  6.06                         0                  25
## 2                  4.71                         0                  21
## 3                  2.83                         0                  29
## 4                  5.04                         0                  36
## 5                  2.51                         0                  38
## 6                  5.03                         0                  50
##   fairly_active_minutes lightly_active_minutes sedentary_minutes calories
## 1                    13                   328               728     1985
## 2                    19                   217               776     1797
## 3                    34                   209               726     1745
## 4                    10                   221               773     1863
## 5                    20                   164               539     1728
## 6                    31                   264               775     2035
##   total_sleep_records total_minutes_asleep total_time_in_bed sleep_quality
## 1                   1                  327               346  Insufficient
## 2                   2                  384               407  Insufficient
## 3                   1                  412               442  Insufficient
## 4                   2                  340               367  Insufficient
## 5                   1                  700               712  Oversleeping
## 6                   1                  304               320  Insufficient
##         user_type
## 1   Highly active
## 2          Active
## 3 Somewhat active
## 4   Highly active
## 5 Somewhat active
## 6   Highly active
```

```
str(sleep_and_activity)
```

```
## 'data.frame':    410 obs. of  20 variables:
##  $ id                       : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ date                     : chr  "4/12/2016" "4/13/2016" "4/15/2016" "4/16/2016" ...
##  $ total_steps              : int  13162 10735 9762 12669 9705 15506 10544 9819 14371 10039
...
##  $ total_distance           : num  8.5 6.97 6.28 8.16 6.48 ...
##  $ tracker_distance         : num  8.5 6.97 6.28 8.16 6.48 ...
##  $ logged_activities_distance: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ very_active_distance     : num  1.88 1.57 2.14 2.71 3.19 ...
##  $ moderately_active_distance: num  0.55 0.69 1.26 0.41 0.78 ...
##  $ light_active_distance    : num  6.06 4.71 2.83 5.04 2.51 ...
##  $ sedentary_active_distance : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ very_active_minutes      : int  25 21 29 36 38 50 28 19 41 39 ...
##  $ fairly_active_minutes    : int  13 19 34 10 20 31 12 8 21 5 ...
##  $ lightly_active_minutes   : int  328 217 209 221 164 264 205 211 262 238 ...
##  $ sedentary_minutes        : int  728 776 726 773 539 775 818 838 732 709 ...
##  $ calories                 : int  1985 1797 1745 1863 1728 2035 1786 1775 1949 1788 ...
##  $ total_sleep_records      : int  1 2 1 2 1 1 1 1 1 1 ...
##  $ total_minutes_asleep     : int  327 384 412 340 700 304 360 325 361 430 ...
##  $ total_time_in_bed        : int  346 407 442 367 712 320 377 364 384 449 ...
##  $ sleep_quality            : chr  "Insufficient" "Insufficient" "Insufficient" "Insufficien
t" ...
##  $ user_type                : chr  "Highly active" "Active" "Somewhat active" "Highly activ
e" ...
```

```
#Previewing Weight dataset
head(weight)
```

```
##           id      date weight_kg weight_pounds  fat   bmi is_manual_report
## 1 1503960366  5/2/2016      52.6      115.9631   22 22.65                1
## 2 1503960366  5/3/2016      52.6      115.9631 NULL 22.65                1
## 3 1927972279 4/13/2016     133.5      294.3171 NULL 47.54                0
## 4 2873212765 4/21/2016      56.7      125.0021 NULL 21.45                1
## 5 2873212765 5/12/2016      57.3      126.3249 NULL 21.69                1
## 6 4319703577 4/17/2016      72.4      159.6147   25 27.45                1
##        log_id
## 1 1.46223e+12
## 2 1.46232e+12
## 3 1.46051e+12
## 4 1.46128e+12
## 5 1.46310e+12
## 6 1.46094e+12
```

```
str(weight)
```

```
## 'data.frame':    67 obs. of  8 variables:
##  $ id              : num  1.50e+09 1.50e+09 1.93e+09 2.87e+09 2.87e+09 ...
##  $ date            : chr  "5/2/2016" "5/3/2016" "4/13/2016" "4/21/2016" ...
##  $ weight_kg       : num  52.6 52.6 133.5 56.7 57.3 ...
##  $ weight_pounds   : num  116 116 294 125 126 ...
##  $ fat             : chr  "22" "NULL" "NULL" "NULL" ...
##  $ bmi             : num  22.6 22.6 47.5 21.5 21.7 ...
##  $ is_manual_report: int  1 1 0 1 1 1 1 1 1 1 ...
##  $ log_id          : num  1.46e+12 1.46e+12 1.46e+12 1.46e+12 1.46e+12 ...
```

Summary() function is written down to gain statistical information about the datasets. I primarily use MEAN values to identify the findings.

```
#Summarizing step, distance, and calories data
df_1 <- sleep_and_activity[,c(3,4,15)]
summary(df_1)
```

```
##   total_steps    total_distance      calories
##  Min.   :   17   Min.   : 0.010   Min.   : 257
##  1st Qu.: 5189   1st Qu.: 3.592   1st Qu.:1841
##  Median : 8913   Median : 6.270   Median :2207
##  Mean   : 8515   Mean   : 6.012   Mean   :2389
##  3rd Qu.:11370   3rd Qu.: 8.005   3rd Qu.:2920
##  Max.   :22770   Max.   :17.540   Max.   :4900
```

Findings: The average step count is 8515 (somewhat active), according to Medical News Today, taking 8,000 steps per day reduces 51% lower risk of dying by any cause. Individuals who walk up to 12,000 steps per day are associated with a 65% lower risk compared with those taking 4,000 steps:

- Sedentary: Less than 5,000 steps daily.
- Low active: About 5,000 to 7,499 steps daily.
- Somewhat active: About 7,500 to 9,999 steps daily.
- Active: More than 10,000 steps daily.
- Highly active: More than 12,500 steps daily.

```
#Summarizing sleep data
df_2 <- sleep_and_activity[,c(16,17,18)]
summary(df_2)
```

```
##  total_sleep_records total_minutes_asleep total_time_in_bed
##  Min.   :1.00        Min.   : 58.0        Min.   : 61.0
##  1st Qu.:1.00        1st Qu.:361.0        1st Qu.:403.8
##  Median :1.00        Median :432.5        Median :463.0
##  Mean   :1.12        Mean   :419.2        Mean   :458.5
##  3rd Qu.:1.00        3rd Qu.:490.0        3rd Qu.:526.0
##  Max.   :3.00        Max.   :796.0        Max.   :961.0
```

Findings: The average user sleeps about 6.99 hours per day, which nearly falls into the recommended range of sleep hours for adults (7-9 hours) by Sleepfoundation. However the average time in bed is higher than the average sleeping time per day around 40 minutes, which shows that participants have the tendency of staying on bed after they wake up.

```
#Summarizing active minutes per category
df_3 <- sleep_and_activity[,c(11,12,13,14)]
summary(df_3)
```

```
##  very_active_minutes fairly_active_minutes lightly_active_minutes
##  Min.   :  0.00      Min.   :  0.00        Min.   :  2.0
##  1st Qu.:  0.00      1st Qu.:  0.00        1st Qu.:158.0
##  Median :  9.00      Median : 11.00        Median :208.0
##  Mean   : 25.05      Mean   : 17.92        Mean   :216.5
##  3rd Qu.: 38.00      3rd Qu.: 26.75        3rd Qu.:263.0
##  Max.   :210.00      Max.   :143.00        Max.   :518.0
##  sedentary_minutes
##  Min.   :   0.0
##  1st Qu.: 631.2
##  Median : 717.0
##  Mean   : 712.1
##  3rd Qu.: 782.8
##  Max.   :1265.0
```

Findings:

- The average sedentary minutes is 712 minutes, which equates to 11.87 hours, which should be reduced.
- Among lightly_active_minutes, fairly_active_minutes, and very_active_minutes; lightly_active_minutes has the highest mean of 216.5 minutes suggesting that most participants engage in light activities.

```
df_4 <- weight[,c(3,6)]
summary(df_4)
```

```
##    weight_kg           bmi
##  Min.   : 52.60   Min.   :21.45
##  1st Qu.: 61.40   1st Qu.:23.96
##  Median : 62.50   Median :24.39
##  Mean   : 72.04   Mean   :25.19
##  3rd Qu.: 85.05   3rd Qu.:25.56
##  Max.   :133.50   Max.   :47.54
```

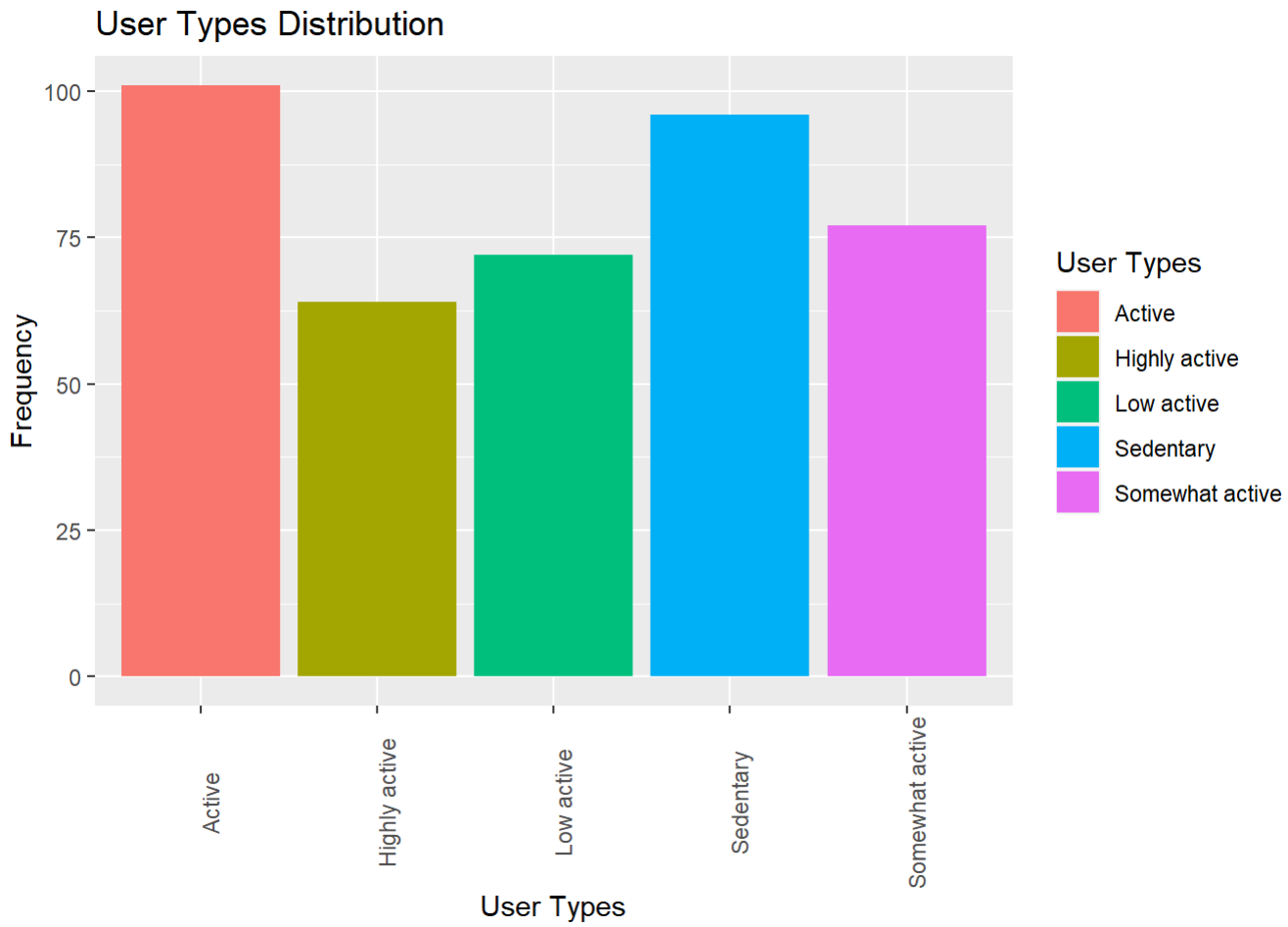
Findings: The average BMI is 25.19, which is considered in the overweight category according to the standards from CDC:

- If your BMI is less than 18.5, it falls within the underweight range.
- If your BMI is 18.5 to <25, it falls within the healthy weight range.
- If your BMI is 25.0 to <30, it falls within the overweight range.
- If your BMI is 30.0 or higher, it falls within the obesity range.

# SHARE

Because there are only 8 distinct observations in the weight table, which is not significant enough to generate recommendations for our business goals. In this stage, we will mainly be constructing visualizations for the SleepAndActivity dataset.

```
# User Types distribution
ggplot(data = sleep_and_activity) + geom_bar(mapping = aes(x = user_type, fill = user_type)) +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(title = "User Types Distribution", x= "User Types", y = "Frequency", fill = "User Types")
```




Insights: More than 50% of the users are active or sedentary.

```
ggplot(data = sleep_and_activity) + geom_bar(mapping = aes(x = sleep_quality, fill = sleep_quali
ty)) +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(title = "Sleep Quality Distribution", x= "Sleep Quality", y = "Frequency", fill = "Sleep
 Quality")
```


Insights: More than 50% of the users are oversleeping or having insufficient sleep.

```
# Sleep quality and calories burned for each user type
ggplot(data = sleep_and_activity) +
  geom_histogram(mapping = aes(x = calories, fill = sleep_quality, position = "identity")) +
  facet_wrap(~user_type) +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(title = "Sleep Quality and Calories Burned", y = "Frequency", x = "Calories Burned", fill
= "Sleep Quality", subtitle = "For 5 different types of user")
```

```
## Warning in geom_histogram(mapping = aes(x = calories, fill = sleep_quality, :
## Ignoring unknown aesthetics: position
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


Insights: Sedentary people with insufficient sleep tend to burn less calories than other user types.

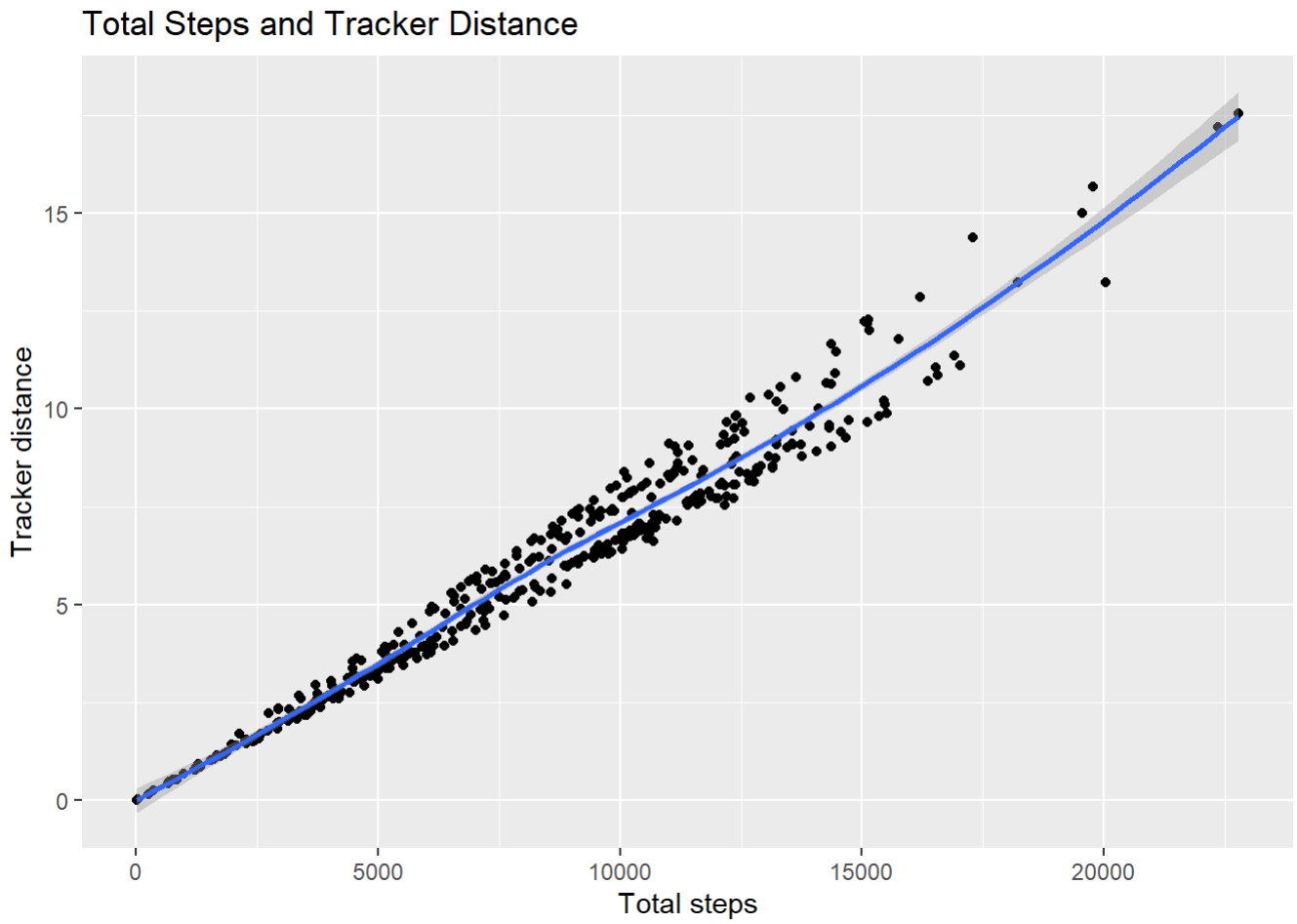
```
# Total steps and Calories
ggplot(data = sleep_and_activity) +
  geom_point(mapping = aes(x=total_steps, y=calories)) +
  geom_smooth(mapping = aes(x=total_steps, y=calories)) +
  labs(title="Total Steps and Calories Burned", x= "Total Steps", y = "Calories Burned")
```

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


Insights: A positive correlation between the total steps and calories burned. This means that the more active the user is, the higher calories they need every day.

```
# Total steps and tracker distance
ggplot(data = sleep_and_activity) +
  geom_point(mapping = aes(x=total_steps, y=tracker_distance)) +
  geom_smooth(mapping = aes(x=total_steps, y=tracker_distance)) +
  labs(title="Total Steps and Tracker Distance", x= "Total steps", y = "Tracker distance")
```

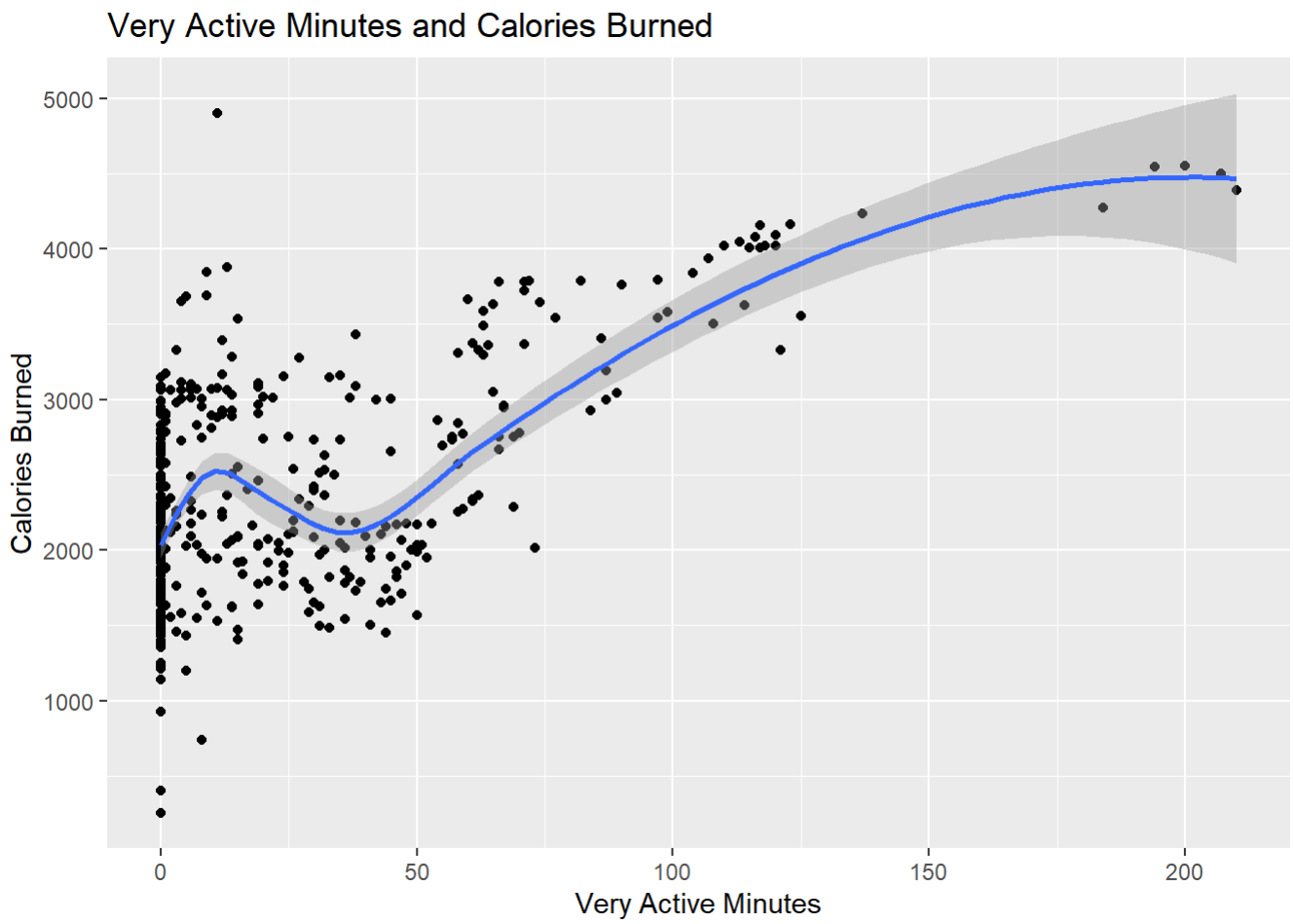
```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```




Insights: There is a strong positve correlation between the total steps and the tracker distance.

```
# Very active minutes and calories burned
ggplot(data = sleep_and_activity) +
  geom_point(mapping = aes(x=very_active_minutes, y=calories)) +
  geom_smooth(mapping = aes(x=very_active_minutes, y=calories)) +
  labs(title="Very Active Minutes and Calories Burned", x= "Very Active Minutes", y = "Calories
 Burned")
```

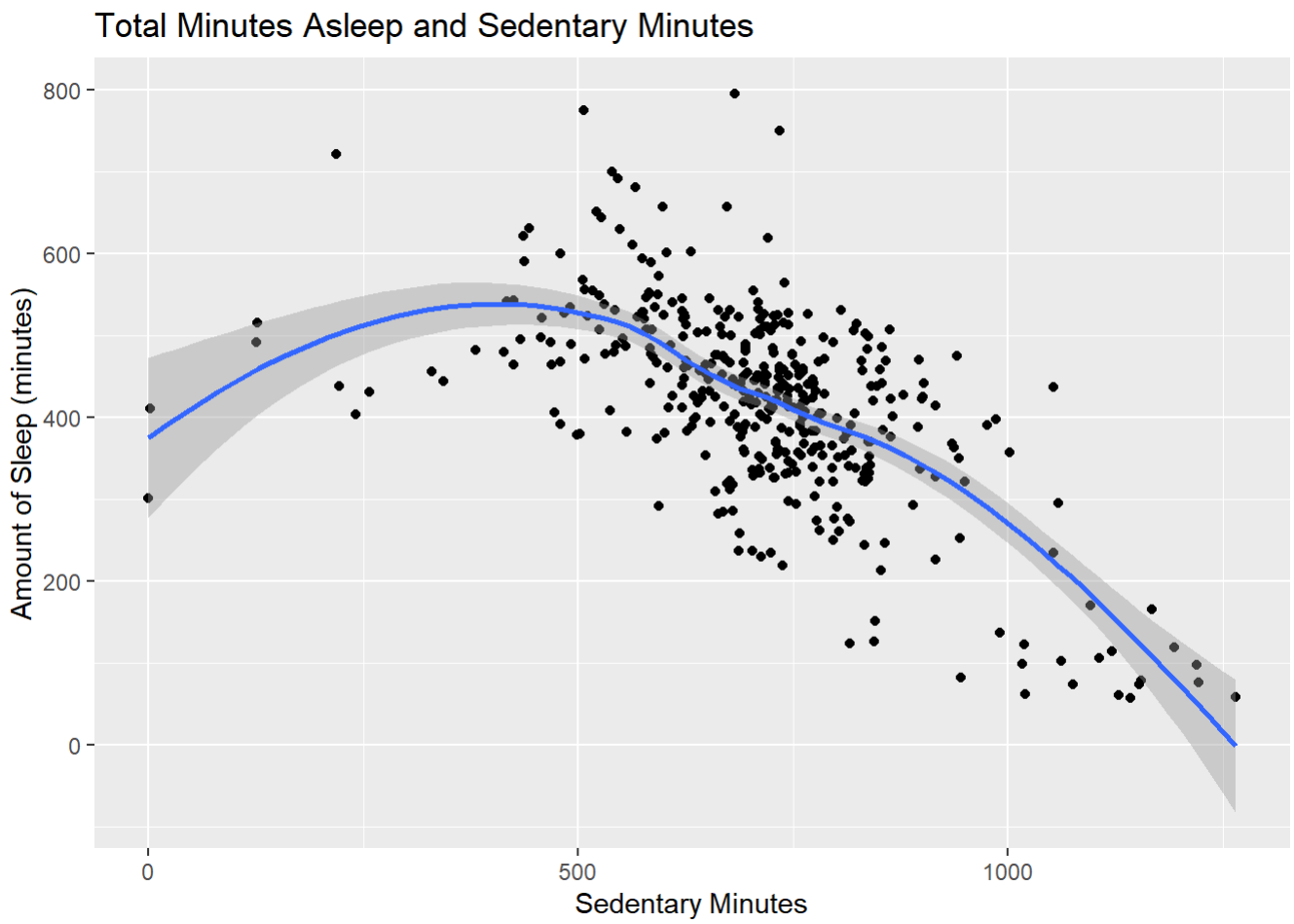
```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```




Insights: The higher active minutes, the more calories burned.

```
# Total minutes asleep and sedentary minutes
ggplot(data = sleep_and_activity) +
  geom_point(mapping = aes(x=sedentary_minutes, y=total_minutes_asleep)) +
  geom_smooth(mapping = aes(x=sedentary_minutes, y=total_minutes_asleep)) +
  labs(title = "Total Minutes Asleep and Sedentary Minutes", x= "Sedentary Minutes", y = "Amount
of Sleep (minutes)")
```

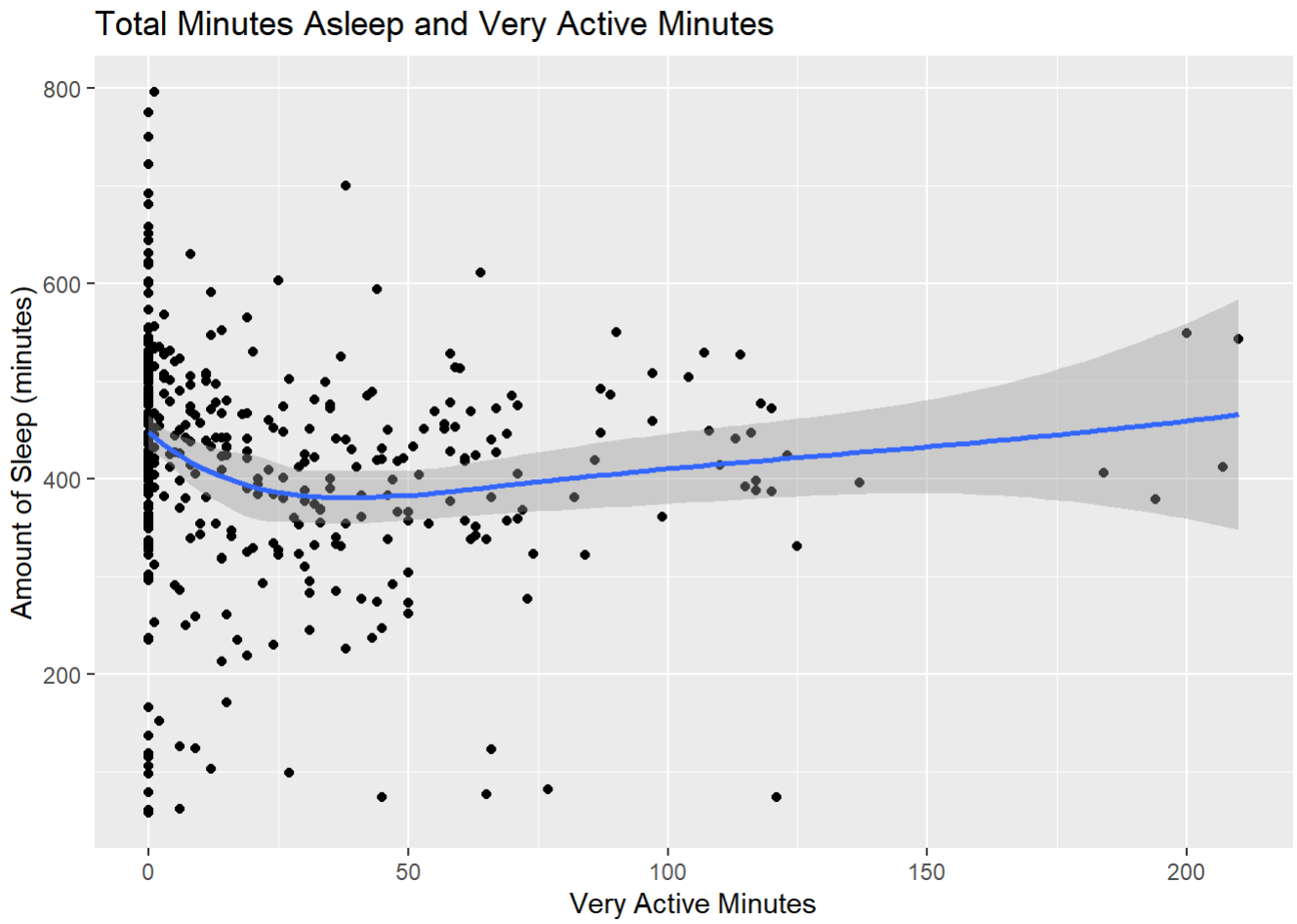
```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```




Insights: There is a negative correlation between the total minutes asleep and sedentary minutes. It seems like users with higher sedentary minutes have lower sleep quality.

```
# Total minutes asleep and very active minutes
ggplot(data = sleep_and_activity) +
  geom_point(mapping = aes(x=very_active_minutes, y=total_minutes_asleep)) +
  geom_smooth(mapping = aes(x=very_active_minutes, y=total_minutes_asleep)) +
  labs(title = "Total Minutes Asleep and Very Active Minutes", x= "Very Active Minutes", y = "Am
ount of Sleep (minutes)")
```

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




Insights: There is a positive correlation between the total minutes asleep and very active minutes.It seems like users with higher active minutes have better sleep quality.

# ACT

After analyzing the Fitbit Fitness Tracker Data, I have a few recommendations and insights that can guide Bellabeat's marketing strategies:

1. Target customer: Both men and women primarily work in sedentary environment such as office.
2. Product innovation: Bellabeat App:
    - Step tracker: According to Medical News Today taking 8,000 steps per day reduces 51% lower risk of dying by any cause. Therefore, we should add notification feature to remind users of the benefits of taking more steps which can boost their vitality. For example, it can be a 5-minute walk after 30-45 minutes of sitting.
    - Exercise reminder: This feature estimates the burned calories of some different exercises or physical activities, as well as set up a daily, weekly, and monthly schedule to achieve customers' fitness goals.
    - Sleep quality improver: Designs sleep time for specific age, gender, and user type, and reminds users to stay away from screens which help them not get distracted before bed and focus on getting quality bed.
    - Goal achiever: Gives compliments to users whenever they follow the tracker to adjust their daily routine. After a week or a month of maintaining a balanced lifestyle, based on users' performance, they can get points which can be transferred to product discounts from Bellabeat's partner.

3. Promotion: One-week free trial of the Bellabeat Leaf and Bellabeat App; Discounts for friend referral, and One-year free Bellabeat App subscription fee for a new Leaf purchase.

4. Project constraints: The Fitbit Fitness Tracker Dataset have a small sample size (33 observations) which cannot affect the reliability of the analysis results because it leads to a higher variability, which may lead to data bias. Therefore, after launching the new product to the market, Bellabeat can gather their own customer data, which is beneficial in further analysis to improve the product performance.

# REFERENCES

1. Bellabeat case study with R. Faustina Offiaeli. (n.d). Retrieved October 10, 2022, from https://www.kaggle.com/code/faustinaoffiaeli/bellabeat-case-study-with-r#Analyze-phase (https://www.kaggle.com/code/faustinaoffiaeli/bellabeat-case-study-with-r#Analyze-phase)

2. Defining Adult Overweight & Obesity. CDC. (n.d). Retrieved October 10, 2022, from https://www.cdc.gov/obesity/basics/adult-defining.html#:~:text=If%20your%20BMI%20is%2018.5,falls%20within%20the%20obesity%20range (https://www.cdc.gov/obesity/basics/adult-defining.html#:~:text=If%20your%20BMI%20is%2018.5,falls%20within%20the%20obesity%20range)

3. FitBit Fitness Tracker Data. Mobius. (n.d). Retrieved October 10, 2022, from https://www.kaggle.com/datasets/arashnic/fitbit (https://www.kaggle.com/datasets/arashnic/fitbit)

4. How Many Steps a Day Is Considered Active? Pallavi Suyog Uttekar. (July 15, 2021). Retrieved October 10, 2022, from https://www.medicinenet.com/how_many_steps_a_day_is_considered_active/article.htm (https://www.medicinenet.com/how_many_steps_a_day_is_considered_active/article.htm)

5. How many steps should people take per day? Medical News Today. (n.d). Retrieved October 10, 2022, from https://www.medicalnewstoday.com/articles/how-many-steps-should-you-take-a-day#for-weight-loss (https://www.medicalnewstoday.com/articles/how-many-steps-should-you-take-a-day#for-weight-loss)

6. How Much Sleep Do We Really Need? Eric Suni. (August 29, 2022). Retrieved October 10, 2022, from https://www.sleepfoundation.org/how-sleep-works/how-much-sleep-do-we-really-need (https://www.sleepfoundation.org/how-sleep-works/how-much-sleep-do-we-really-need)