# Bellabeat Starter

## Introduction and background

Health is a big issue around the world, and people are wanting to do their best to make sure that their health is the best to live a long and meaningful life. To do so, they are using smart technology, such as smart watches, phone apps, and other things to improve their daily habits. The following information will show how consumers use non-Bellabeat smart devices, including trends, how they can apply to Bellabeat customers, and help influence the Bellabeat marketing strategy.

## The data

The data is provided by FitBit, which is collected via Amazon Mechanical Turk and stored in CSV files. This information is in the public domain and can be accessed via Kaggle. Consent for this data was obtained, as well as being anonymized.

With the data provided, I will be specifically looking at the daily activity and sleep patterns.

## Potential issues

With the data provided, there are some potential issues and biases:

1. The dataset provided only shows users with FitBits, and thera are roughly 30 participants.
2. The data is third-party, as stated earlier, which cannot be 100% verified.
3. Different FitBit devices were used, so the tracking can be off per person.
4. The data provided hasn't been updated in several years.

Ideally, first-party data would be better as there would be verification regarding the data provided. Not only that, but being able to include other smart technology, such as data from Apple Watch and Oura Ring.

However, I will treat this dataset as complete for the purpose of this case study.

## Preparing data

Prior to being able to analyze any data whatsoever, I needed to load the relevant packages and import the datasets for daily activity and sleep patterns.

```r
# Installing the Tidyverse packages

library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2     -- 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 error
```

```r
library(lubridate)
library(dplyr)
library(ggplot2)
library(tidyr)

# Importing datasets

daily_activity <- read.csv("dailyActivity_merged.csv")
sleep_patterns <- read.csv("sleepDay_merged.csv")
```

Here, we have some data from the `daily_activity` file:

```r
head(daily_activity)
```

```
##           Id ActivityDate TotalSteps TotalDistance TrackerDistance
## 1 1503960366    4/12/2016      13162          8.50            8.50
## 2 1503960366    4/13/2016      10735          6.97            6.97
## 3 1503960366    4/14/2016      10460          6.74            6.74
## 4 1503960366    4/15/2016       9762          6.28            6.28
## 5 1503960366    4/16/2016      12669          8.16            8.16
## 6 1503960366    4/17/2016       9705          6.48            6.48
##   LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance
## 1                        0               1.88                     0.55
## 2                        0               1.57                     0.69
## 3                        0               2.44                     0.40
## 4                        0               2.14                     1.26
## 5                        0               2.71                     0.41
## 6                        0               3.19                     0.78
##   LightActiveDistance SedentaryActiveDistance VeryActiveMinutes
## 1                6.06                       0                25
## 2                4.71                       0                21
## 3                3.91                       0                30
## 4                2.83                       0                29
## 5                5.04                       0                36
## 6                2.51                       0                38
##   FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories
## 1                  13                  328              728     1985
## 2                  19                  217              776     1797
## 3                  11                  181             1218     1776
## 4                  34                  209              726     1745
## 5                  10                  221              773     1863
## 6                  20                  164              539     1728
```

Here, we have some data from the `sleep_patterns` file:

```r
head(sleep_patterns)
```

```
##           Id              SleepDay TotalSleepRecords TotalMinutesAsleep
## 1 1503960366 4/12/2016 12:00:00 AM                 1                327
## 2 1503960366 4/13/2016 12:00:00 AM                 2                384
## 3 1503960366 4/15/2016 12:00:00 AM                 1                412
## 4 1503960366 4/16/2016 12:00:00 AM                 2                340
## 5 1503960366 4/17/2016 12:00:00 AM                 1                700
## 6 1503960366 4/19/2016 12:00:00 AM                 1                304
##   TotalTimeInBed
## 1            346
```

```
## 2              407
## 3              442
## 4              367
## 5              712
## 6              320
```

### Understanding some statistics

The following information shows how many participates there are in `daily_activity` and `sleep_patterns`:

```r
# Participants in daily_activity

n_distinct(daily_activity$Id)
```

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

```r
# Participants in sleep_patterns

n_distinct(sleep_patterns$Id)
```

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

The following information shows how many observations there are in `daily_activity` and `sleep_patterns`:

```r
# Observations in daily_activity

nrow(daily_activity)
```

```
## [1] 940
```

```r
# Observations in sleep_patterns

nrow(sleep_patterns)
```

```
## [1] 413
```

The following information shows the total steps and total distances in `daily_activity`:

```r
daily_activity %>%
  select(TotalSteps,
         TotalDistance,
         SedentaryMinutes) %>%
  summary
```

```
##    TotalSteps    TotalDistance    SedentaryMinutes
##  Min.   :    0   Min.   : 0.000   Min.   :   0.0  
##  1st Qu.: 3790   1st Qu.: 2.620   1st Qu.: 729.8  
##  Median : 7406   Median : 5.245   Median :1057.5  
##  Mean   : 7638   Mean   : 5.490   Mean   : 991.2  
##  3rd Qu.:10727   3rd Qu.: 7.713   3rd Qu.:1229.5  
##  Max.   :36019   Max.   :28.030   Max.   :1440.0  
```

The above data shows that the more steps taken, the further the distance that they go, as well as the more time that they will be in the sedentary position. This is most likely due to them needing more rest after traveling a longer distance.

The following information shows the overall sleep records, meaning how many times they've slept, as well as the amount of time, in minutes, slept, as well as the total time each person was in bed, as provided in `sleep_patterns`:

```r
sleep_patterns %>%
  select(TotalSleepRecords,
         TotalMinutesAsleep,
         TotalTimeInBed) %>%
  summary()
```

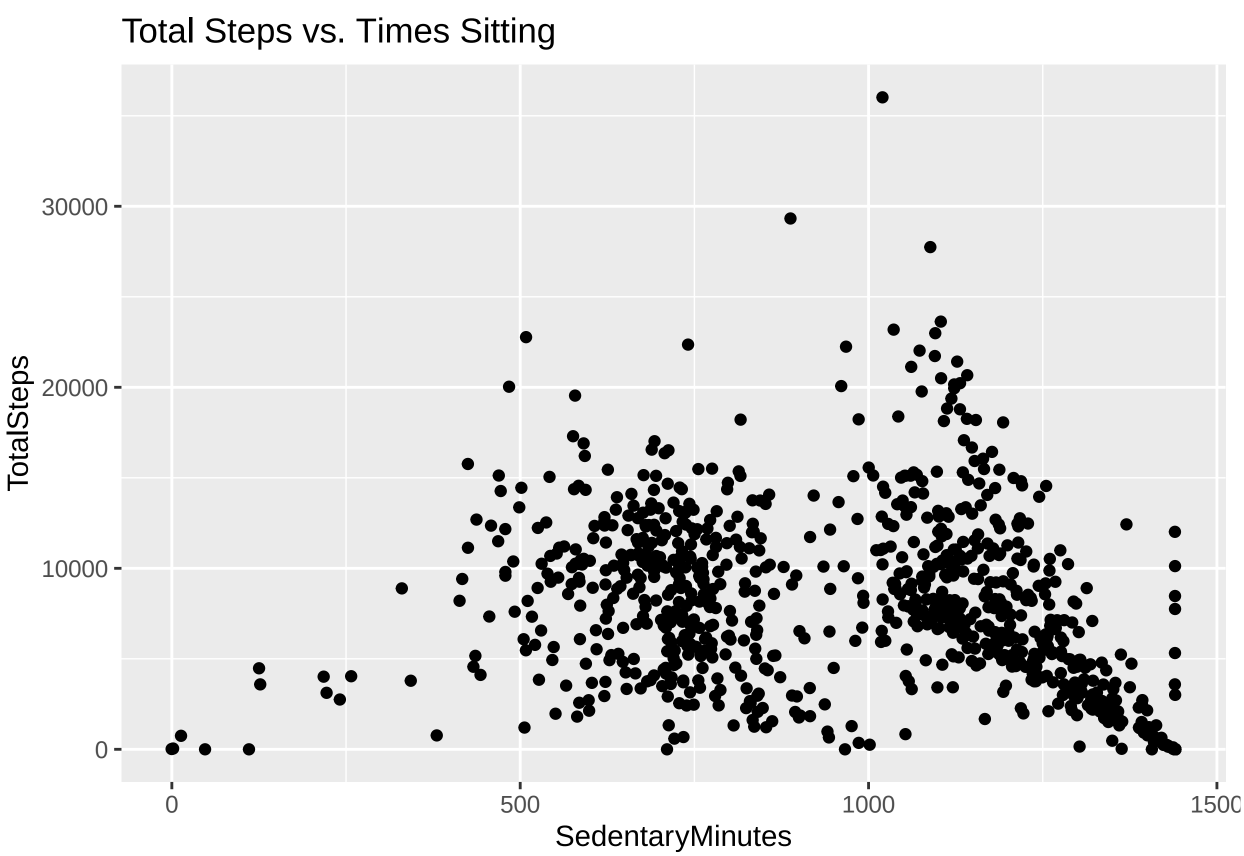
```
##  TotalSleepRecords TotalMinutesAsleep TotalTimeInBed
##  Min.   :1.000     Min.   : 58.0      Min.   : 61.0
##  1st Qu.:1.000     1st Qu.:361.0      1st Qu.:403.0
##  Median :1.000     Median :433.0      Median :463.0
##  Mean   :1.119     Mean   :419.5      Mean   :458.6
##  3rd Qu.:1.000     3rd Qu.:490.0      3rd Qu.:526.0
##  Max.   :3.000     Max.   :796.0      Max.   :961.0
```

### Analysis

Now that we have all the information needed, we can now move forward in regards to graph the above information.

This scatterplot shows the relationship between total steps and sedentary minutes, as provided by `daily_activity`:

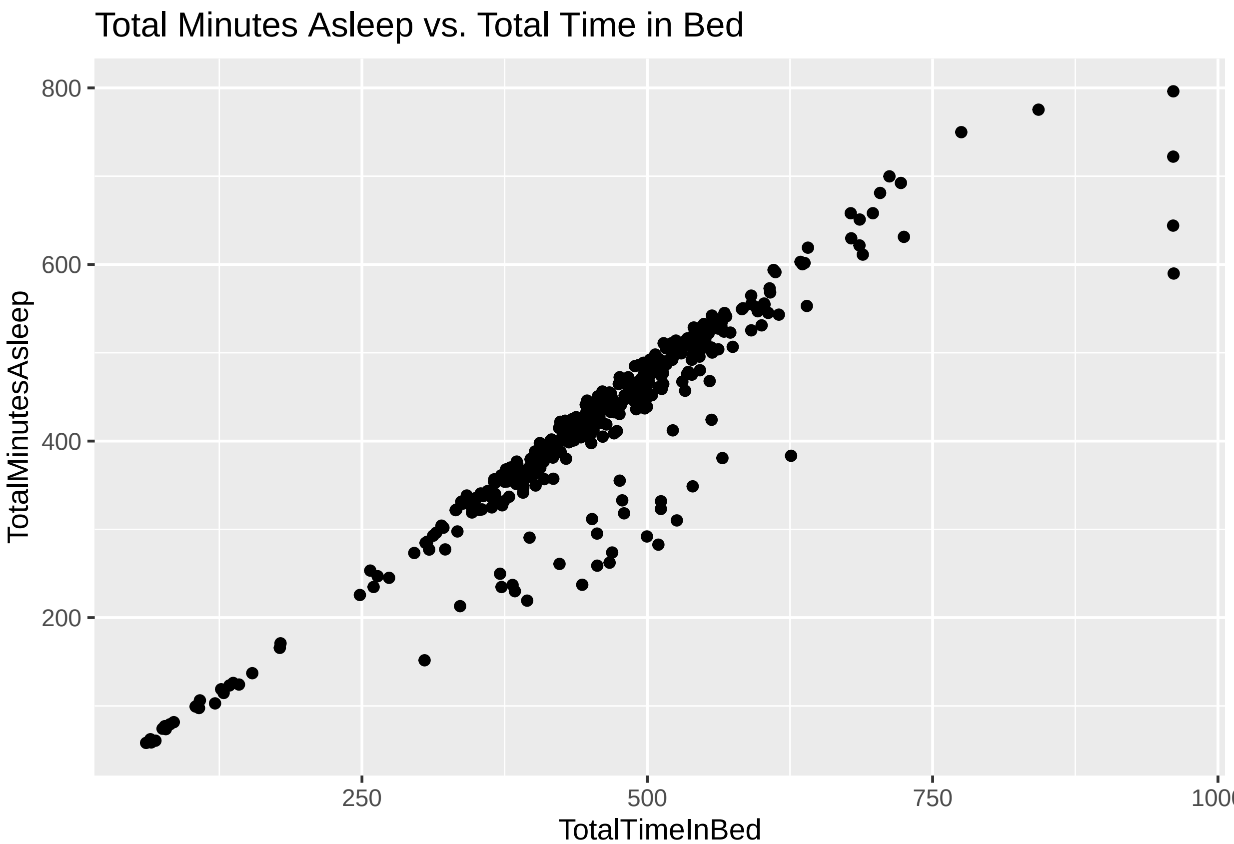
```r
ggplot(data=daily_activity, aes(x=SedentaryMinutes,y=TotalSteps)) + geom_jitter() +
  labs(title="Total Steps vs. Times Sitting")
```




The above scatterplot shows, in essence, the more time in the sednetary minutes, the less steps taken. This means that the more time taking steps, whether running or walking, or some sort of combination of the two, will allow for more total steps and less time sitting down. This will allow for a healthier lifestyle than those who have less steps and are more inclined to be in a sedentary position, unless it's in regards to other health issues.

This next scatterplot shows us the total time a person has been sleeping versus the amount of time to which they are in bed, as provided in `sleep_patterns`:

```
ggplot(data=sleep_patterns, aes(x=TotalTimeInBed,y=TotalMinutesAsleep)) + geom_jitter() +
  labs(title="Total Minutes Asleep vs. Total Time in Bed")
```




As we can see in the above scatterplot, we see that, in general, the more time spent asleep the more time a person stays in bed.

### Joining datasets together

We combine the two datasets together to show some more information:

```
combined_data <- merge(sleep_patterns,daily_activity, by="Id")
```

With the merging of both datasets, we can find how many participants are in this dataset:

```
n_distinct((combined_data$Id))
```

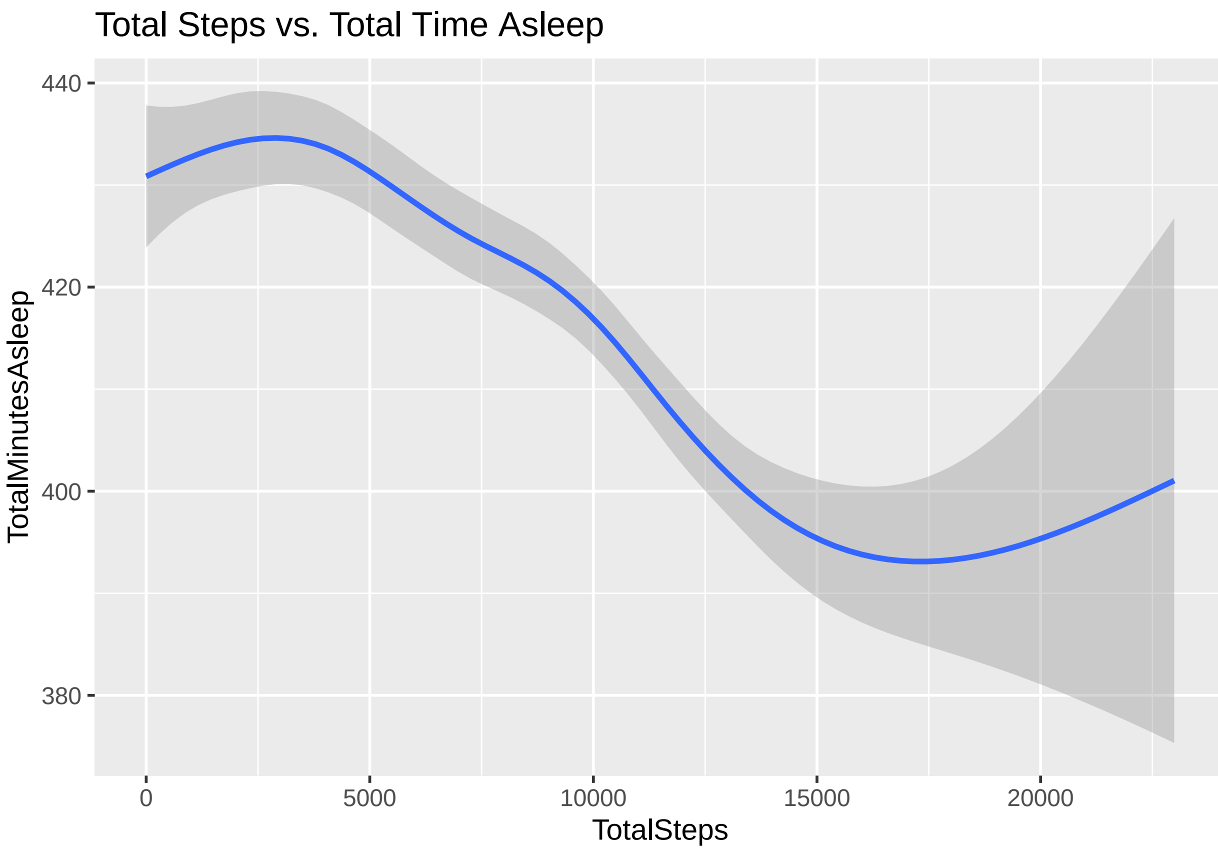
```
## [1] 24
```

As we can see, there are 24 distinct participants, meaning that these participants are all different people.

Now, let's take a look at the comparison between time asleep and total steps in a day:

```
ggplot(data=combined_data, aes(x=TotalSteps,y=TotalMinutesAsleep)) +
  geom_smooth() + labs(title="Total Steps vs. Total Time Asleep")
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```




Here we can see that the more time a person is asleep, there is a higher chance of them not taking more steps per day. This can cause other issues regarding health.

### Suggestions for Bellabeat

With the information provided above, I recommend Bellabeat implement the following:

1. It would be best to have some sort of campaign that shows that the more time being sedentary and not moving can cause issues for sleep. Even moving just some can improve sleeping and allow them to move more and more.
2. Enable notifications that lets the person know to stand up, and possibly even move for a couple of minutes to give their body some exercise.
3. Implement user reports on a weekly and monthly basis that would show the overall activity of the user and challenge the user to do better the next week or month.