# Project Bellabeat

Juan Pablo Montilla

2025-07-11

## Contents



### 0.1 1. Introduction

For this project, the stakeholder at Bellabeat wants to analyze trends in health-focused smart devices. Using this analysis, we aim to understand how the insights derived from Fitbit data can be applied to Bellabeat's product portfolio and how they may help shape Bellabeat's future marketing strategies. ## 2. Data Loading

We are using a public dataset: Fitbit Fitness Tracker Data (CC0: Public Domain, made available via Mobius). This dataset contains minute-level data on physical activity, sleep, calories and heart rate. It was used in the Google Data Analytics Capstone as a proxy to analyze user behavior for Bellabeat.

The files were downloaded, unzipped and loaded into R using `read_csv()`:

```
library(readr)

dailyActivity_merged <- read_csv("Fitabase Data 4.12.16-5.12.16/dailyActivity_merged.csv")
```

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

```
daily_Calories <- read_csv("Fitabase Data 4.12.16-5.12.16/dailyCalories_merged.csv")
```

```
## Rows: 940 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
```

```
## chr (1): ActivityDay
## 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.
```

```
daily_Intensities <- read_csv("Fitabase Data 4.12.16-5.12.16/dailyIntensities_merged.csv")
```

```
## Rows: 940 Columns: 10
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): ActivityDay
## dbl (9): Id, SedentaryMinutes, LightlyActiveMinutes, FairlyActiveMinutes, Ve...
##
## 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.
```

```
daily_Steps <- read_csv("Fitabase Data 4.12.16-5.12.16/dailySteps_merged.csv")
```

```
## Rows: 940 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): ActivityDay
## dbl (2): Id, StepTotal
##
## 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.
```

```
hourly_Calories <- read_csv("Fitabase Data 4.12.16-5.12.16/hourlyCalories_merged.csv")
```

```
## Rows: 22099 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.
```

```
minute_Sleep <- read_csv("Fitabase Data 4.12.16-5.12.16/minuteSleep_merged.csv")
```

```
## Rows: 188521 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.
```

## 0.2 3. Cleaning and exploration

in this section, I cleaned and explored the `dailyActivity_merged` dataset, which includes steps, calories, distances and minutes of activity.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
```

```
## v dplyr     1.1.4     v purrr     1.0.4
## 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
## -- 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)
```

### 0.3 3.1 Data Preparation

# 1 Load dataset

```
daily <- read_csv("Fitabase Data 4.12.16-5.12.16/dailyActivity_merged.csv")
```

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

# 2 Fix column names and format date

```
daily <- daily %>%
  mutate(ActivityDate = as.Date(ActivityDate, format = "%m/%d/%Y")) %>%
  rename(
    user_id = Id,
    date = ActivityDate,
    total_steps = TotalSteps,
    total_distance = TotalDistance,
    tracker_distance = TrackerDistance,
    logged_active_distance = LoggedActivitiesDistance,
    very_active_minutes = VeryActiveMinutes,
    fairly_active_minutes = FairlyActiveMinutes,
    lightly_active_minutes = LightlyActiveMinutes,
    sedentary_minutes = SedentaryMinutes,
    calories = Calories
  )
```

### 2.1 3.2 Summary statistics

```
summary(daily)
```

```
##     user_id              date             total_steps    total_distance
##  Min.   :1.504e+09   Min.   :2016-04-12   Min.   :    0   Min.   : 0.000
##  1st Qu.:2.320e+09   1st Qu.:2016-04-19   1st Qu.: 3790   1st Qu.: 2.620
##  Median :4.445e+09   Median :2016-04-26   Median : 7406   Median : 5.245
```

```
##  Mean   :4.855e+09   Mean   :2016-04-26   Mean   : 7638   Mean   : 5.490
##  3rd Qu.:6.962e+09   3rd Qu.:2016-05-04   3rd Qu.:10727   3rd Qu.: 7.713
##  Max.   :8.878e+09   Max.   :2016-05-12   Max.   :36019   Max.   :28.030
##  tracker_distance logged_active_distance VeryActiveDistance
##  Min.   : 0.000   Min.   :0.0000         Min.   : 0.000
##  1st Qu.: 2.620   1st Qu.:0.0000         1st Qu.: 0.000
##  Median : 5.245   Median :0.0000         Median : 0.210
##  Mean   : 5.475   Mean   :0.1082         Mean   : 1.503
##  3rd Qu.: 7.710   3rd Qu.:0.0000         3rd Qu.: 2.053
##  Max.   :28.030   Max.   :4.9421         Max.   :21.920
##  ModeratelyActiveDistance LightActiveDistance SedentaryActiveDistance
##  Min.   :0.0000           Min.   : 0.000      Min.   :0.000000
##  1st Qu.:0.0000           1st Qu.: 1.945      1st Qu.:0.000000
##  Median :0.2400           Median : 3.365      Median :0.000000
##  Mean   :0.5675           Mean   : 3.341      Mean   :0.001606
##  3rd Qu.:0.8000           3rd Qu.: 4.782      3rd Qu.:0.000000
##  Max.   :6.4800           Max.   :10.710      Max.   :0.110000
##  very_active_minutes fairly_active_minutes lightly_active_minutes
##  Min.   :  0.00      Min.   :  0.00        Min.   :  0.0
##  1st Qu.:  0.00      1st Qu.:  0.00        1st Qu.:127.0
##  Median :  4.00      Median :  6.00        Median :199.0
##  Mean   : 21.16      Mean   : 13.56        Mean   :192.8
##  3rd Qu.: 32.00      3rd Qu.: 19.00        3rd Qu.:264.0
##  Max.   :210.00      Max.   :143.00        Max.   :518.0
##  sedentary_minutes    calories
##  Min.   :   0.0    Min.   :   0
##  1st Qu.: 729.8    1st Qu.:1828
##  Median :1057.5    Median :2134
##  Mean   : 991.2    Mean   :2304
##  3rd Qu.:1229.5    3rd Qu.:2793
##  Max.   :1440.0    Max.   :4900
```

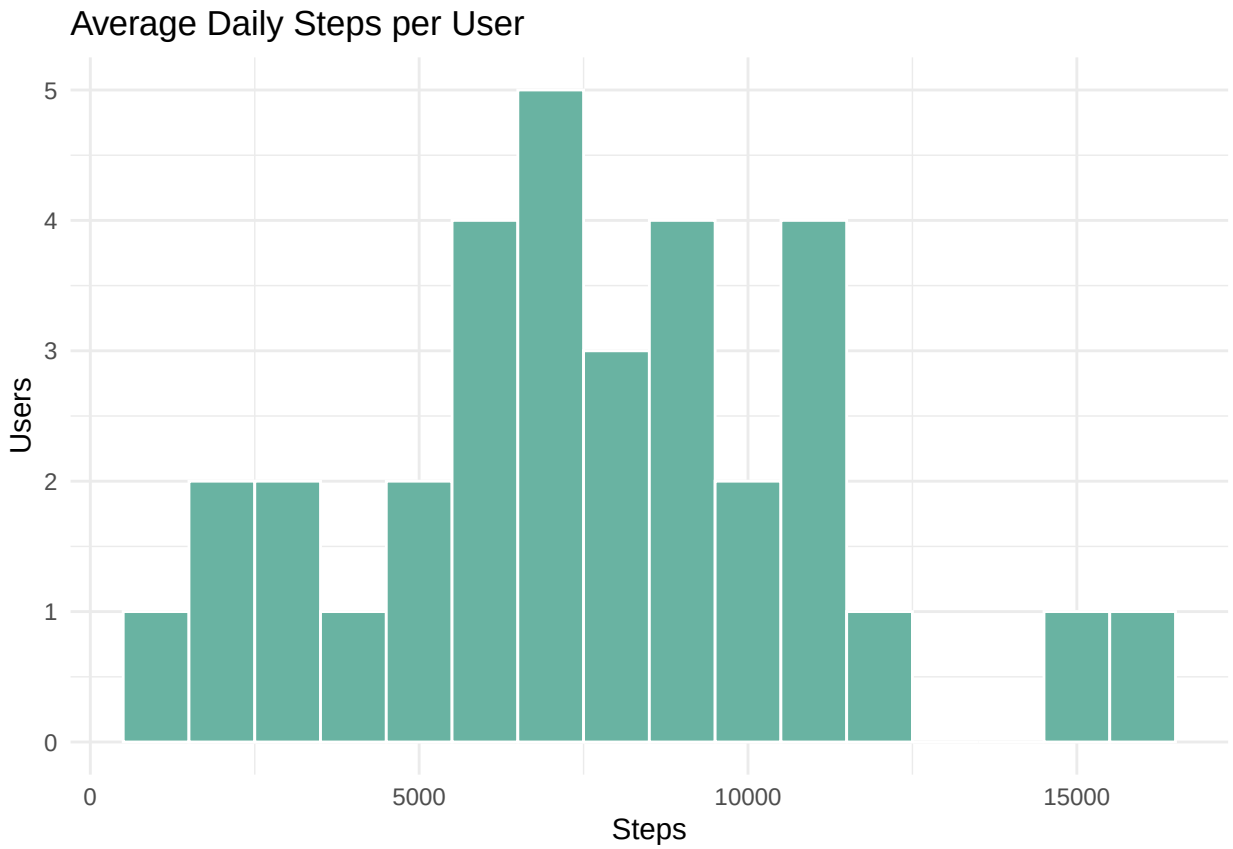
## 2.2 4. Behavior Analysis

Now I want to understand the behavior of each user in terms of activity and calories burned. This helps identify different user types and how Bellabeat could tailor its marketing accordingly.

```
user_summary <- daily %>%
  group_by(user_id) %>%
  summarise(
    days_tracked = n(),
    avg_steps = mean(total_steps),
    avg_calories = mean(calories),
    avg_sedentary_min = mean(sedentary_minutes),
    avg_active_min = mean(very_active_minutes + fairly_active_minutes + lightly_active_minutes)
  )
```

## 2.3 4.1 Histograms: Step distribution

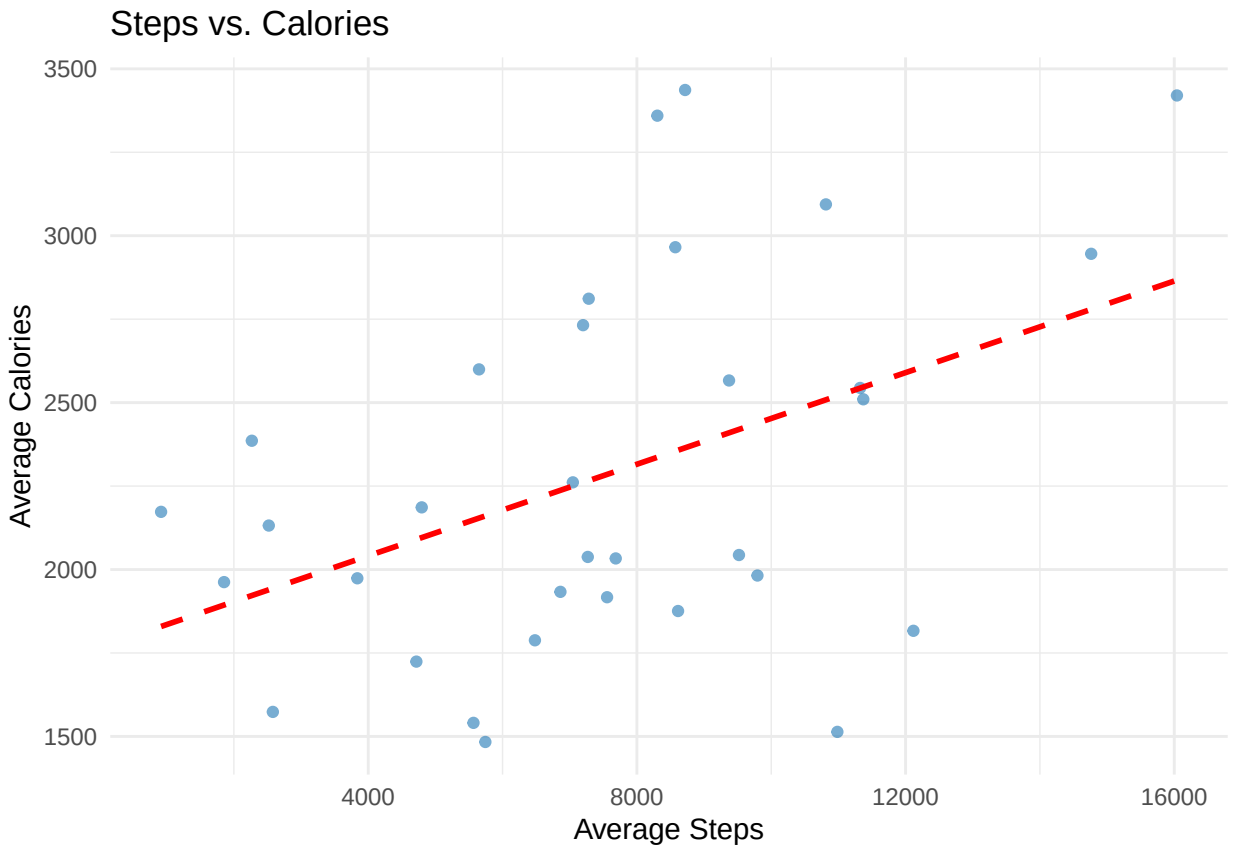
```
ggplot(user_summary, aes(x = avg_steps)) +
  geom_histogram(binwidth = 1000, fill = "#69b3a2", color = "white") +
  labs(title = "Average Daily Steps per User", x = "Steps", y = "Users") +
  theme_minimal()
```




## 4.2 Step VS Calories

```
ggplot(user_summary, aes(x = avg_steps, y = avg_calories)) +
  geom_point(alpha = 0.6, color = "#1f77b4") +
  geom_smooth(method = "lm", se = FALSE, linetype = "dashed", color = "red") +
  labs(title = "Steps vs. Calories", x = "Average Steps", y = "Average Calories") +
  theme_minimal()
```

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

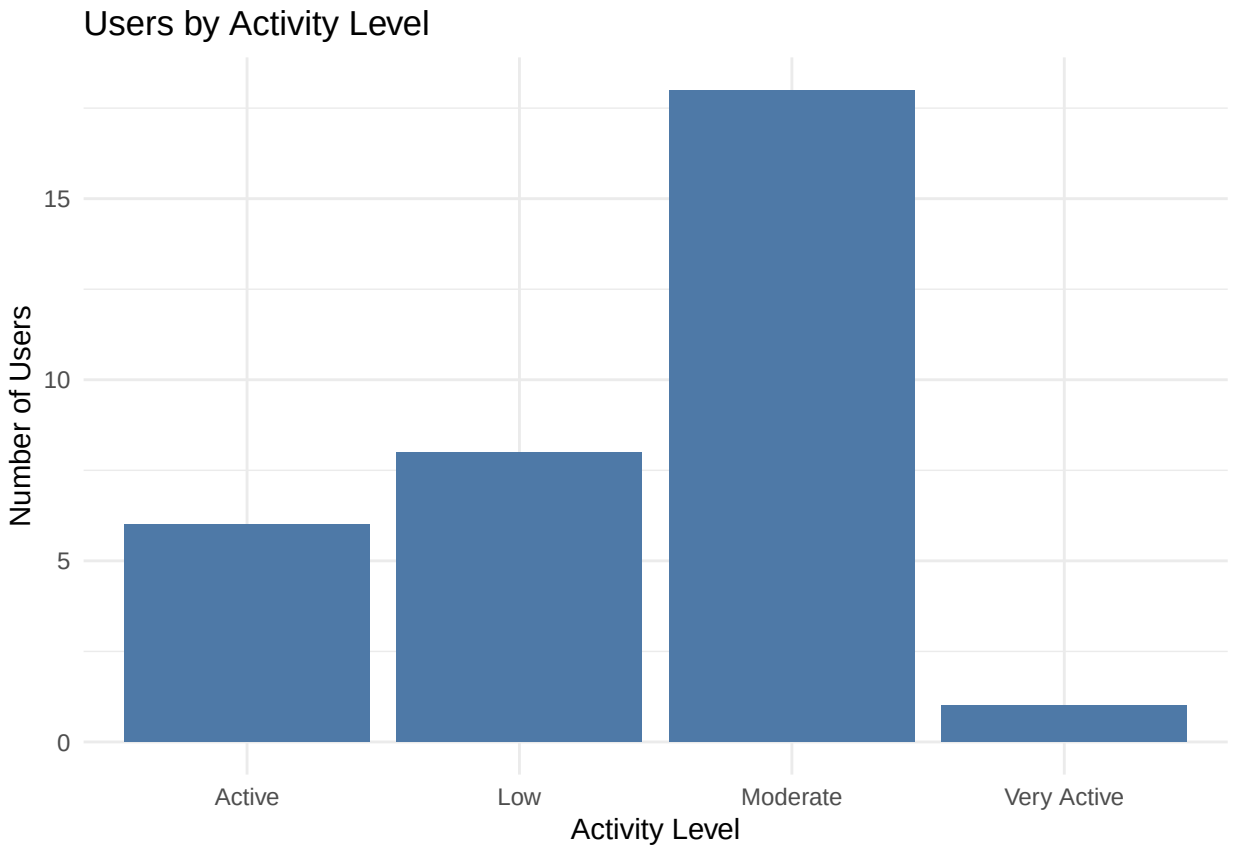



## 5. User segmentation To improve the marketing strategy, I created four activity segments based on users' average daily steps.

```
user_summary <- user_summary %>%
  mutate(activity_level = case_when(
    avg_steps < 5000 ~ "Low",
    avg_steps < 10000 ~ "Moderate",
    avg_steps < 15000 ~ "Active",
    TRUE ~ "Very Active"
  ))
```

### 2.4 5.1 Bar chart: Activity levels

```
ggplot(user_summary, aes(x = activity_level)) +
  geom_bar(fill = "#4e79a7") +
  labs(title = "Users by Activity Level", x = "Activity Level", y = "Number of Users") +
  theme_minimal()
```




## 2.5 6. Insights and recommendations

- Over 70% of users record less than 10,000 steps per day → opportunity for motivational marketing campaigns
- Clear correlation between steps and calories burned → can guide personalized feedback
- Sedentary time is high even in active users → opportunity to promote micro-movements or stretch reminders
- Bellabeat can use this segmentation to:

  -Design custom challenges per user group

  -Increase product engagement via targeted in-app notifications

  -Tailor its wellness content and product design to real user behavior

## 2.6 7. Author

Juan Pablo Montilla Montaña Portafolio: