Bellabeat

Project Background

Bellabeat, founded in 2013, is a high-tech company that manufactures health-focused smart products. By collecting data on activity, sleep, stress, and reproductive health, Bellabeat empowers women with knowledge about their health and habits.

Business Task

This project analyzes smart device usage data to gain insight into how consumers use non-Bellabeat smart devices. The aim is to provide key actionable recommendations to improve the marketing strategy and become a larger player in the global smart device market. The focus will be on the Bellabeat watch “Time”.

Insights and recommendations are provided on the following key areas:

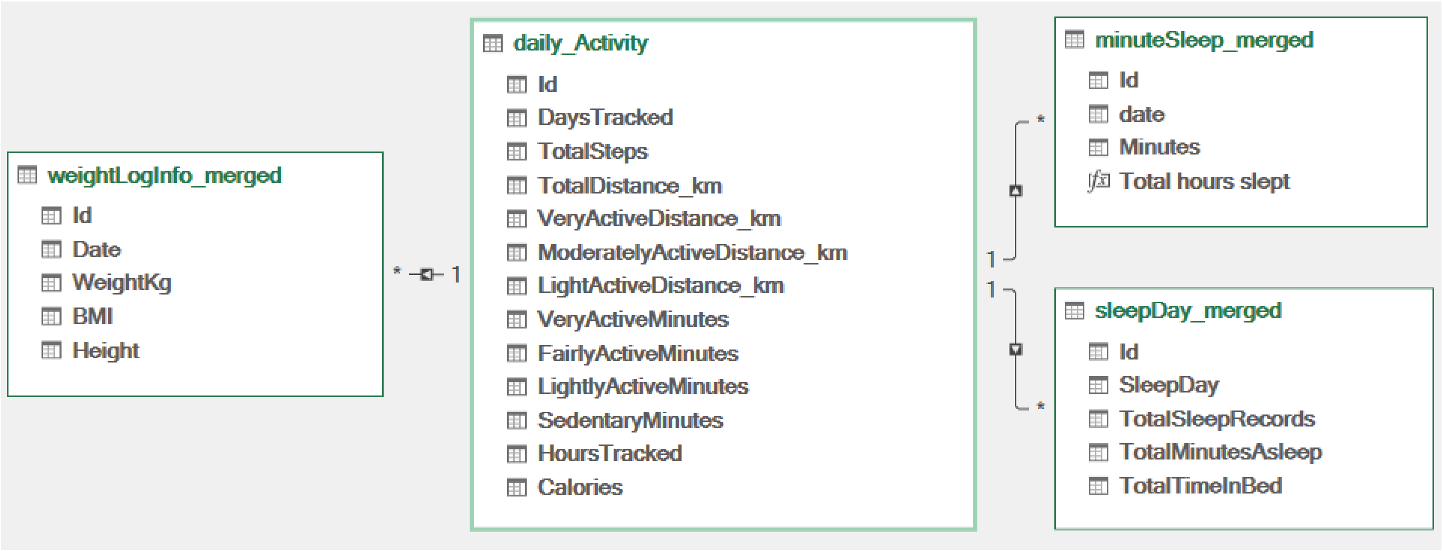
- **Smart device features usage**: An overview of the different features offered by smart devices and how likely consumers are to use them.
- **Participant segmentation**: An assessment of the participants of the study. Their health statistics and overall routine.
- **Relationship analysis**: An analysis of the relationship between the measured variables to determine the impact they have on overall health.

The SQL queries used to aggregate data for further analysis can be found **here**.

The R code used to create the relationship visualizations and data aggregation can be found **here**.

Data Structure

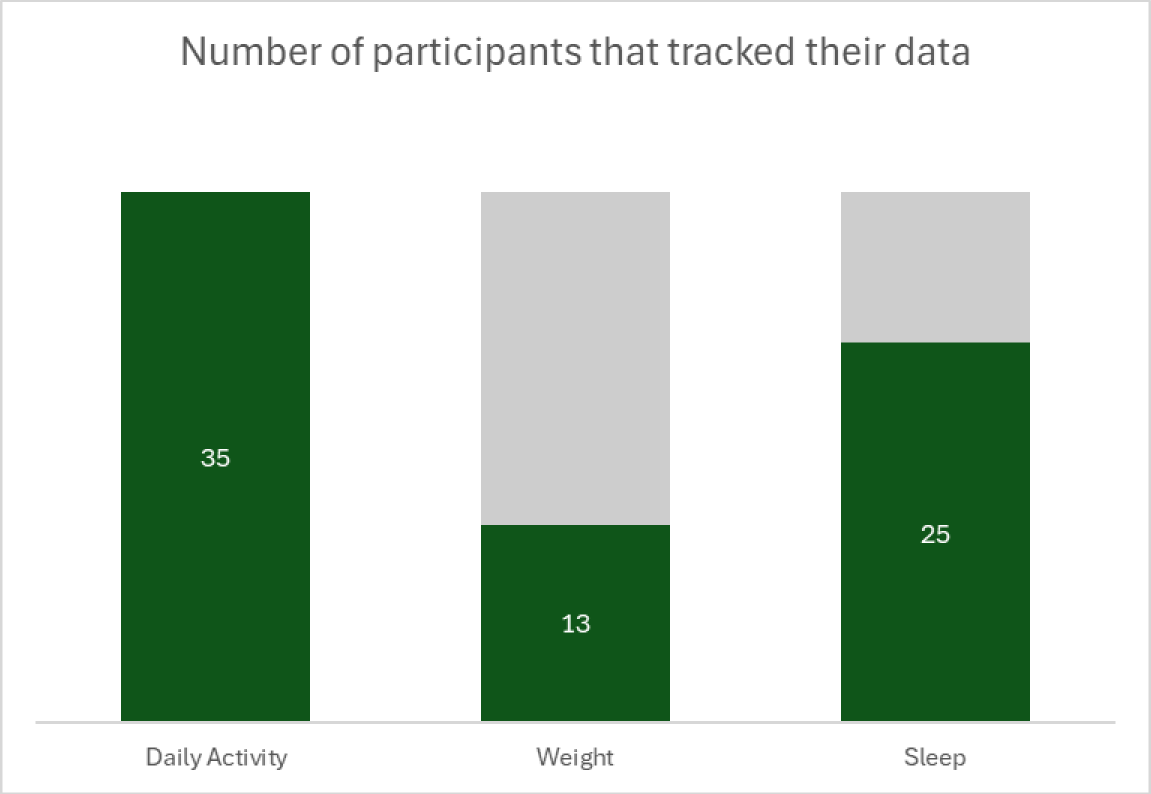
Bellabeat’s data collection consists of a total of 29 CSV files collected across 2 months. The files considered for this analysis were the following: dailyActivity, weightLogInfo, minuteSleep, and SleepDay.



Before the start of the analysis, a variety of checks were conducted for data familiarization. Each CSV file was evaluated in terms of data structure, data types, and possible connections to other data. Data aggregation was used in both SQL and R to merge datasets as they existed in 2 separate files due to the time constraint.

Executive Summary

Analysis of smart device usage patterns reveals that users consistently engage with features that track daily routines like step count, calories burned, active minutes, or distance traveled. In contrast, features like weight logging and sleep tracking are underutilized and lagging behind, respectively. These findings suggest an opportunity for smart health products to focus on seamless, all-day tracking of activity and sleep, the two most consistently used features. Enhancing usability and integration of these features could better align with user behavior, increase adoption, and long-term engagement.



Data Sources

- Fitabase Data 3.12.16-4.11.16: dailyActivity_merged.csv
- Fitabase Data 3.12.16-4.11.16: minuteSleep_merged.csv
- Fitabase Data 3.12.16-4.11.16: weightLogInfo_merged.csv

- Fitabase Data 4.12.16-5.12.16: dailyActivity_merged.csv
- Fitabase Data 4.12.16-5.12.16: minuteSleep_merged.csv
- Fitabase Data 4.12.16-5.12.16: sleepDay_merged.csv
- Fitabase Data 4.12.16-5.12.16: weightLogInfo_merged.csv

- National Heart, Lung, and Blood Institute, 2025: BMI Categories chart
- Examining Factors of Engagement With Digital Interventions for Weight Management, JMIR Res Protoc. 2017
- National Heart, Lung, and Blood Institute, 2024: How Much Sleep Is Enough?
- International Agency for Research on Cancer, 2015: Energy Balance and Obesity

Data Manipulation

Excel:

Used to merge the 3 tables of: dailyActivity, Sleep, and Weight through Power Query and Excel's Data Model to get a visual representation of the variables in each table and to create any extra visuals via connected Pivot Tables.

SQL:

Merged and selected only the relevant data from both of the "dailyActivity_merged.csv" files to later download it as a new CSV file and use it in Excel to get further insights from the aggregated data.

The following queries were used:

```
INSERT INTO `sql-practice-460602.BellaBeat.dailyActivity_March_April`
SELECT * FROM `sql-practice-460602.BellaBeat.dailyActivity_April_May`
;

select
  Id,
  count(ActivityDate) as DaysTracked,
  sum(TotalSteps) as TotalSteps,
  cast(avg(TotalSteps)as int64) as AverageStepsDay,
  cast(sum(round(TotalDistance,2))as int64) as TotalDistance_km,
  round(avg(TotalDistance),2) as AverageDistanceDay_km,
  cast(avg(VeryActiveMinutes)as int64) as AverageVeryActive_minutes,
  sum(calories) as TotalCalories,
  cast(avg(calories)as int64) as AverageCalories,
  sum(cast(round(((VeryActiveMinutes + FairlyActiveMinutes + LightlyActiveMinutes +
SedentaryMinutes)/60)) as int64)) as HoursTracked
from `sql-practice-460602.BellaBeat.dailyActivity_March_April`
group by Id
```

R:

A total of 5 CSV files were loaded:

- 2 daily activity files.
- 2 weight log files.
- 1 sleep day file.

Several dataframes were merged/united:

- Both daily activity dataframes were united with each other.
- Both weight dataframes were united with each other.
- The new daily activity dataframe was merged with the new weight dataframe.
- The new daily activity dataframe was merged with the sleep day dataframe.

This created a total of 3 dataframes from which we analyzed relationships between variables. The code used to generate all the visualizations is the following:

```
library(tidyverse)
library(ggplot2)


# Daily activity

dailyActivity <- read.csv('Fitabase Data 3.12.16-4.11.16/dailyActivity_merged.csv')
dailyActivity2 <- read.csv('Fitabase Data 4.12.16-5.12.16/dailyActivity_merged.csv')

str(dailyActivity)
str(dailyActivity2)

# Combine both dailyActivity datasheets
combined_daily_data <- union(dailyActivity, dailyActivity2)

str(combined_daily_data)

# Graph: Calories & Total Steps
ggplot(combined_daily_data, aes(x=Calories, y=TotalSteps)) +
  geom_point() +
  geom_smooth() +
  theme(panel.background = element_blank()) +
  labs(title = "Relationship between Calories & Total Steps") +
  annotate("text", y = 21000, x = 4700,
           label = as.character(round(cor(combined_daily_data$Calories,
                                          combined_daily_data$TotalSteps), 3)),
           color = "red")
```

```
# Graph: Calories & Total Distance traveled
ggplot(combined_daily_data, aes(x=Calories, y=TotalDistance)) +
  geom_point() +
  geom_smooth() +
  theme(panel.background = element_blank()) +
  labs(title = "Relationship between Calories & Total Distance traveled") +
  annotate("text", y = 17, x = 4700,
           label = as.character(round(cor(combined_daily_data$Calories,
                                          combined_daily_data$TotalDistance), 3)),
           color = "red")


# Graph: Calories & Very Active Minutes
ggplot(combined_daily_data, aes(x=Calories, y=VeryActiveMinutes)) +
  geom_point() +
  geom_smooth() +
  theme(panel.background = element_blank()) +
  labs(title = "Relationship between Calories & Very Active Minutes") +
  annotate("text", y = 168, x = 4700,
           label = as.character(round(cor(combined_daily_data$Calories,
                                          combined_daily_data$VeryActiveMinutes), 3)),
           color = "red")


# Graph: Calories & Very Active Distance traveled
ggplot(combined_daily_data, aes(x=Calories, y=VeryActiveDistance)) +
  geom_point() +
  geom_smooth() +
  theme(panel.background = element_blank()) +
  labs(title = "Relationship between Calories & Very Active Distance traveled") +
  annotate("text", y = 10.5, x = 4700,
           label = as.character(round(cor(combined_daily_data$Calories,
                                          combined_daily_data$VeryActiveDistance), 3)),
           color = "red")


# Graph: Calories & Lightly Active Minutes
ggplot(combined_daily_data, aes(x=Calories, y=LightlyActiveMinutes)) +
  geom_point() +
  geom_smooth() +
  theme(panel.background = element_blank()) +
  labs(title = "Relationship between Calories & Lightly Active Minutes") +
  annotate("text", y = 285, x = 4700,
           label = as.character(round(cor(combined_daily_data$Calories,
                                          combined_daily_data$LightlyActiveMinutes), 3)),
           color = "red")
```

```r
# Graph: Calories & Light Active Distance traveled
ggplot(combined_daily_data, aes(x=Calories, y=LightActiveDistance)) +
  geom_point() +
  geom_smooth() +
  theme(panel.background = element_blank()) +
  labs(title = "Relationship between Calories & Light Active Distance traveled") +
  annotate("text", y = 6, x = 4700,
           label = as.character(round(cor(combined_daily_data$Calories,
                                          combined_daily_data$LightActiveDistance), 3)),
           color = "red")



# Sleep

minuteSleep <- read.csv('Fitabase Data 3.12.16-4.11.16/minuteSleep_merged.csv')
minuteSleep2 <- read.csv('Fitabase Data 4.12.16-5.12.16/minuteSleep_merged.csv')

str(minuteSleep)
str(minuteSleep2)

# Combine both minuteSleep datasheets
combined_sleep_data <- union(minuteSleep, minuteSleep2)

combined_sleep_data %>%
  group_by('Id') %>%
  select(Id, logId) %>%
  summary(x = n_distinct(logId))

sleep_sessions <- combined_sleep_data %>%
  group_by(Id, logId) %>%
  summarise(
    session_minutes = n(),
    session_start = min(date),
    session_end = max(date),
    .groups = "drop"
  )

sleep_sessions

total_sleep_per_person <- sleep_sessions %>%
  group_by(Id) %>%
  summarise(
    total_sleep_minutes = sum(session_minutes),
    total_sleep_hours = total_sleep_minutes / 60
  )
```

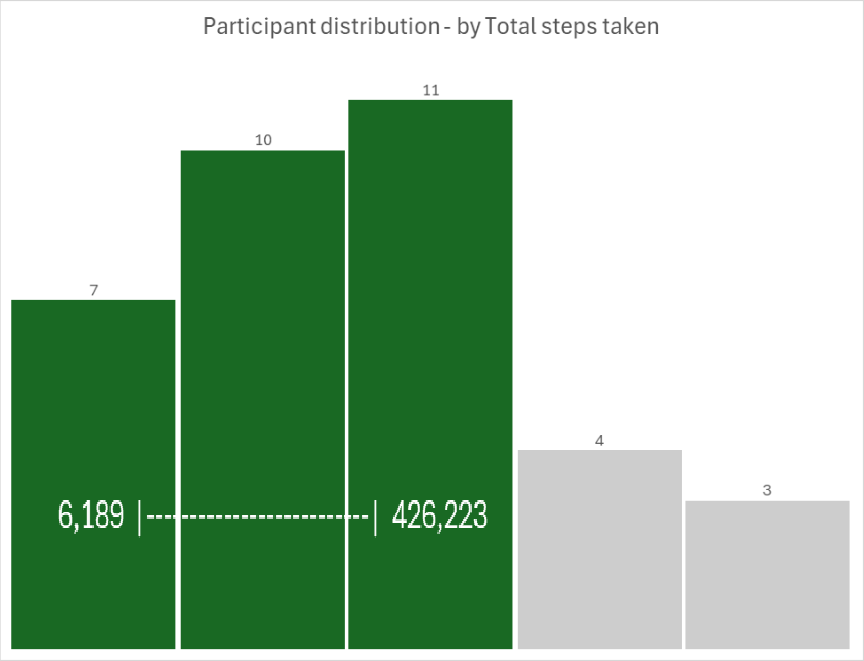
```
total_sleep_per_person

write.csv(sleep_sessions,file='sleep_sessions.csv', row.names = FALSE)
write.csv(total_sleep_per_person,file='total_sleep_per_person.csv', row.names = FALSE)


# Sleep day and daily activity

sleepDay <- read.csv("Fitabase Data 4.12.16-5.12.16/sleepDay_merged.csv")

str(sleepDay)

# Combine combined_daily_data and sleepDay datasheets

sleepDay$Date <- as.Date(sleepDay$SleepDay, format = "%m/%d/%Y %H:%M")
colnames(combined_daily_data)[colnames(combined_daily_data) == "ActivityDate"] <- "Date"
combined_daily_data$Date <- as.Date(combined_daily_data$Date, format = "%m/%d/%Y")


combined_daily_sleep_data <- merge(sleepDay, combined_daily_data, by = c("Id", "Date"), all.x = TRUE)


#Graph: Sedentary Minutes & Total Minutes Asleep
ggplot(combined_daily_sleep_data, aes(x=SedentaryMinutes, y=TotalMinutesAsleep)) +
  geom_point() +
  geom_smooth() +
  theme(panel.background = element_blank()) +
  labs(title = "Relationship between Sedentary Minutes & Total Minutes Asleep") +
  annotate("text", y = 350, x = 1200,
           label = as.character(round(cor(combined_daily_sleep_data$SedentaryMinutes,
                                          combined_daily_sleep_data$TotalMinutesAsleep), 3)),
           color = "red")


#Graph: Total Steps & Total Minutes Asleep
ggplot(combined_daily_sleep_data, aes(x=TotalSteps, y=TotalMinutesAsleep)) +
  geom_point() +
  geom_smooth() +
  theme(panel.background = element_blank()) +
  labs(title = "Relationship between Total Steps taken & Total Minutes Asleep") +
  annotate("text", y = 470, x = 19500,
           label = as.character(round(cor(combined_daily_sleep_data$TotalSteps,
                                          combined_daily_sleep_data$TotalMinutesAsleep), 3)),
           color = "red")
```

```
# Daily activity & Weight Log

weightLog <- read.csv("Fitabase Data 3.12.16-4.11.16/weightLogInfo_merged.csv")
weightLog2 <- read.csv("Fitabase Data 4.12.16-5.12.16/weightLogInfo_merged.csv")

str(weightLog)
str(weightLog2)

rm(combined_daily_weightLog_data)

# Combine both Weight Log datasheets
combined_weightLog_data <- union_all(weightLog, weightLog2)
str(combined_weightLog_data)

combined_weightLog_data$Date <- as.Date(combined_weightLog_data$Date, format = "%m/%d/%Y
%H:%M:%S")

# Combine both Weight Log datasheets and daily Activity

combined_daily_weightLog_data <- merge(combined_weightLog_data, combined_daily_data, by = c("Id",
"Date"))


# Graph: Weight & Total Steps
ggplot(combined_daily_weightLog_data, aes(x=WeightKg, y=TotalSteps)) +
  geom_point() +
  geom_smooth() +
  theme(panel.background = element_blank()) +
  labs(title = "Relationship between Weight & Total Steps") +
  annotate("text", y = 15000, x = 120,
           label = as.character(round(cor(combined_daily_weightLog_data$WeightKg,
                                          combined_daily_weightLog_data$TotalSteps), 3)),
           color = "red")


# Graph: Weight & Sedentary Minutes
ggplot(combined_daily_weightLog_data, aes(x=WeightKg, y=SedentaryMinutes)) +
  geom_point() +
  geom_smooth() +
  theme(panel.background = element_blank()) +
  labs(title = "Relationship between Weight & Sedentary Minutes") +
  annotate("text", y = 1100, x = 120,
           label = as.character(round(cor(combined_daily_weightLog_data$WeightKg,
                                          combined_daily_weightLog_data$SedentaryMinutes), 3)),
           color = "red")
```

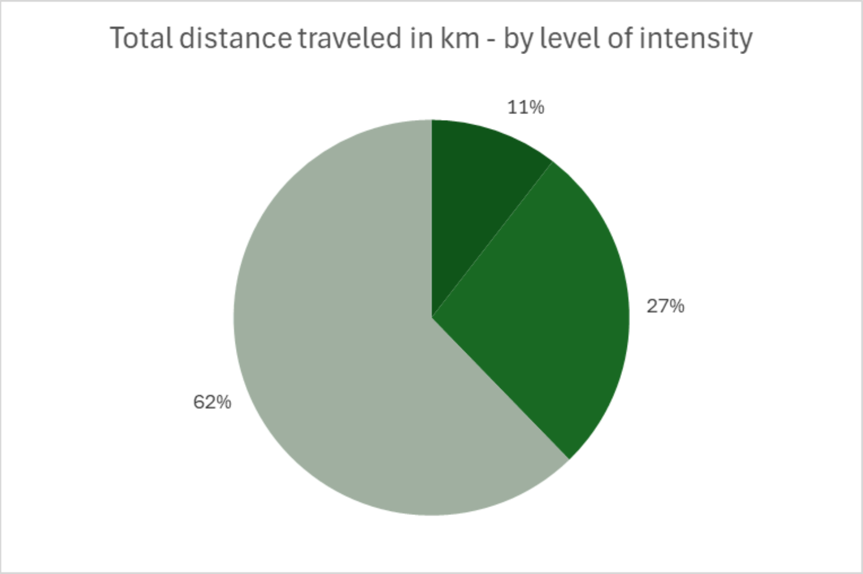
Findings

Daily Activity:

- A total of 35 participants agreed to this 2-month study.
- On average, participants tracked 40 days of data and walked an average distance of 5 km per day.
- The first 80% of the participants walked an average of 5,607 steps per day. While the remaining 20% were slightly higher on the steps scale.



- Furthermore, 62% of the total distance traveled was covered walking, 27% running, and 11% jogging.

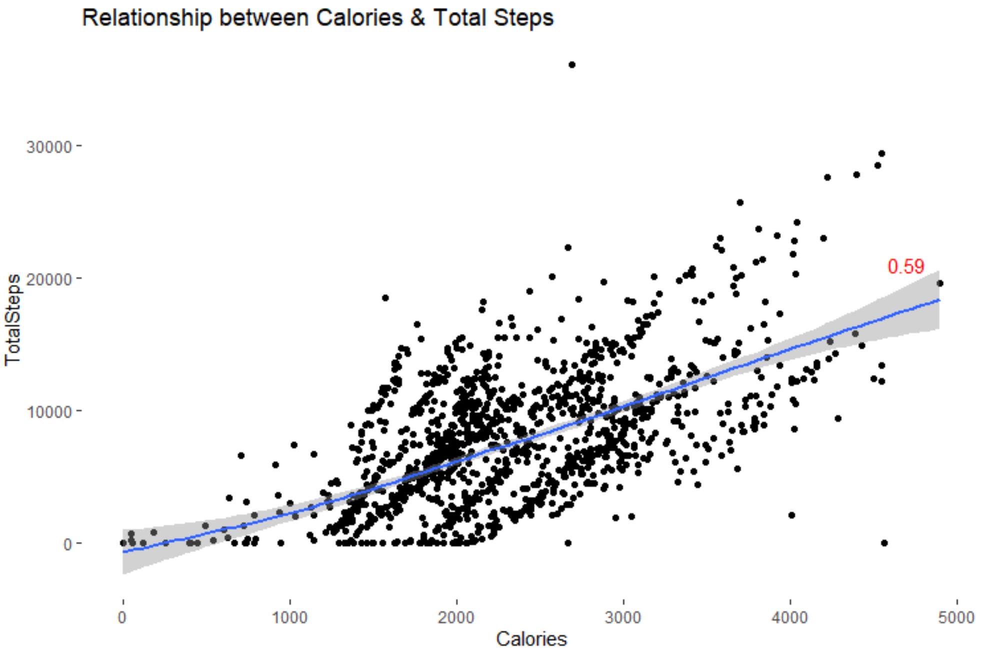


- It is important to note that participants burned on average 2,249 calories per day.

Daily Activity - Relationships:

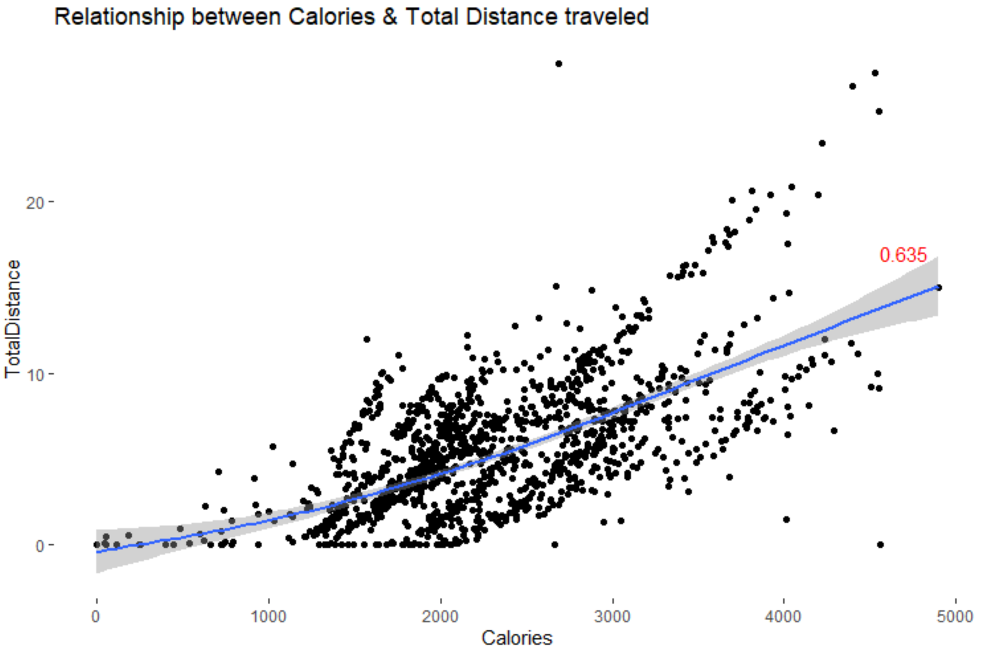
- Calories & Total Steps

    - The correlation coefficient shows 0.59, indicating a weak positive relationship between the amount of calories burned and the total steps taken that day.



- Calories & Total Distance

    - The correlation coefficient shows 0.635, indicating a weak positive relationship between the amount of calories burned and the total distance traveled that day.

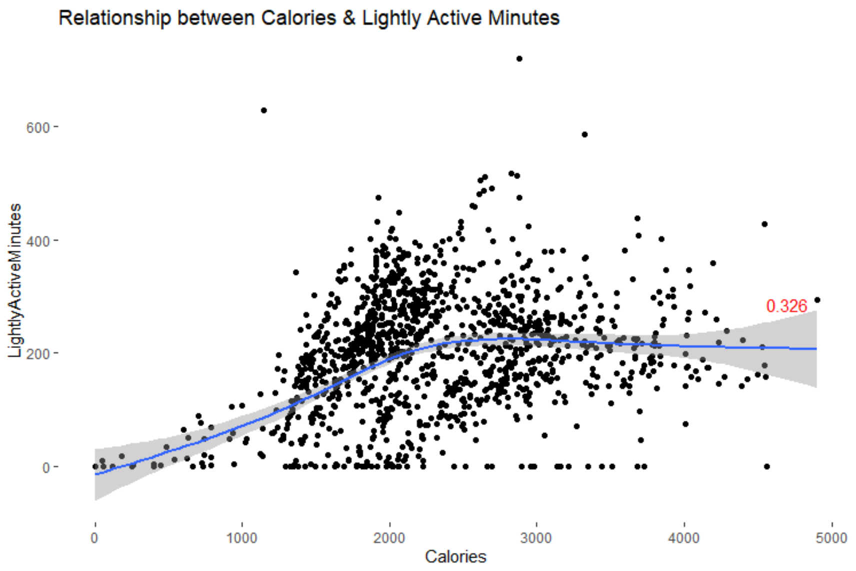


These 2 correlation graphs should be no surprise to anyone. It is expected that the more active a person is, the more calories they will burn, and these metrics are very similar: total steps and total distance.

What could be surprising is the following insight:

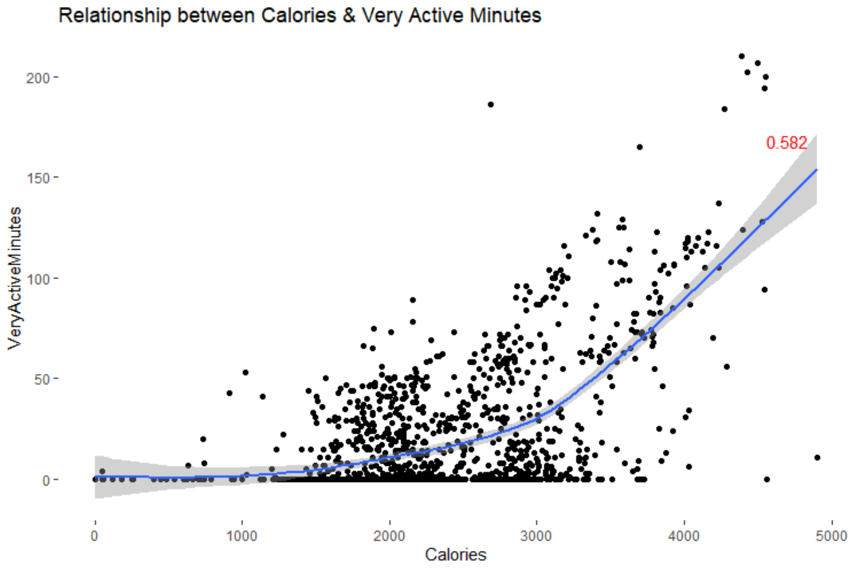
- Calories & Lightly Active Minutes

    - The correlation coefficient shows 0.326, indicating a very weak positive to almost no correlation between the amount of calories burned and the amount of minutes spent doing a "Light Activity" that day.



- Calories & Very Active Minutes

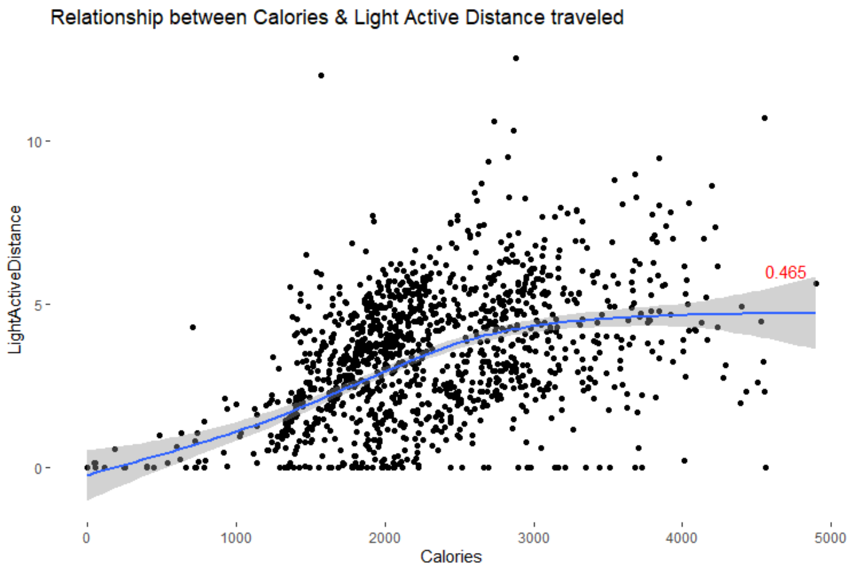
    - On the other hand, shows a correlation coefficient of 0.582, indicating a weak positive correlation between the amount of calories burned and the amount of minutes spent doing a "Very Active" activity that day.



This comparison alone also seems a bit expected. The more minutes you spend doing a demanding activity, the more calories you will burn. It becomes interesting when you add the following 2 graphs comparing calories burned and distance traveled in the same categories: “Light Activity” and “Very Active” activity.

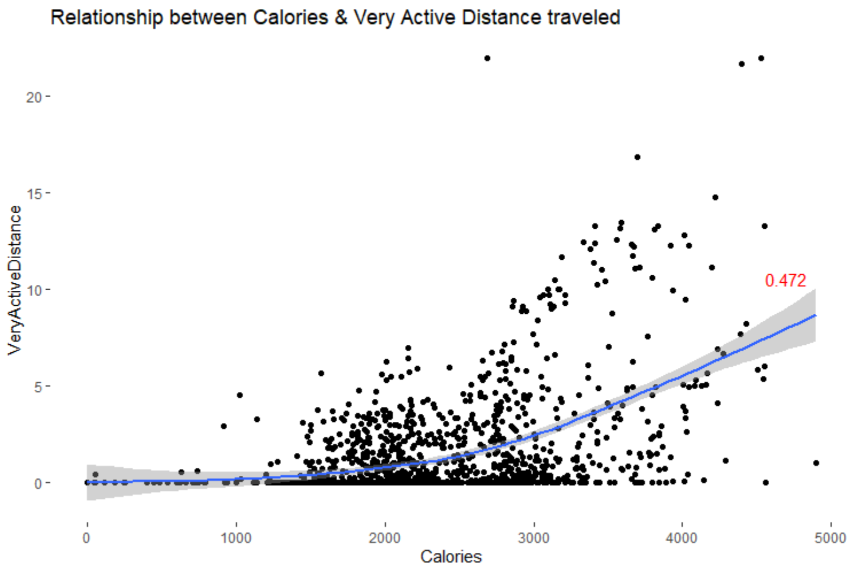
- Calories & Light Active Distance

  - The correlation coefficient of 0.465 indicates a weak positive correlation between the amount of calories burned and the amount of distance traveled doing a “Light Activity”.



- Calories & Very Active Distance

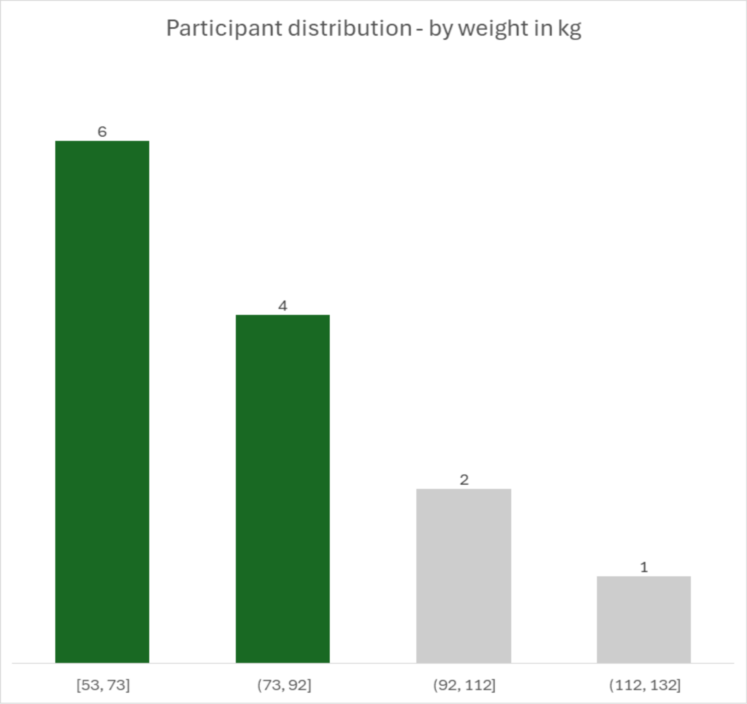
  - The correlation coefficient of 0.472 indicates a weak positive correlation between the amount of calories burned and the amount of “Very Active” distance traveled.



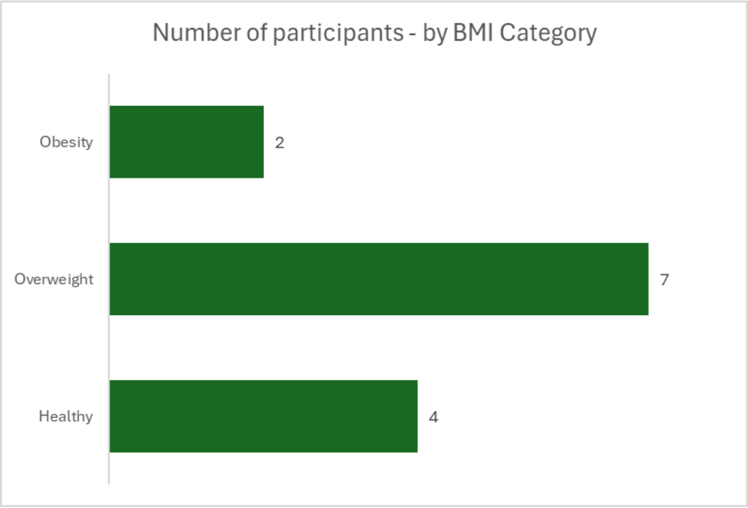
With this new insight, we can conclude that when it comes to burning calories, it matters not the distance you travel at a specified intensity, but rather, how much time you spend in that state of intensity. Marketing the "Time" from Bellabeat with this insight in mind is then highly recommended.

Weight Log:

- Out of the 35 total participants, only 13 used the weight log feature.
- Only 2 participants consistently tracked their weight on the app for more than 7 days. This means that less than 6% of participants tracked their weight for more than 1 week.
- 77% of the participants who used the weight log feature weigh less than 92 kg.



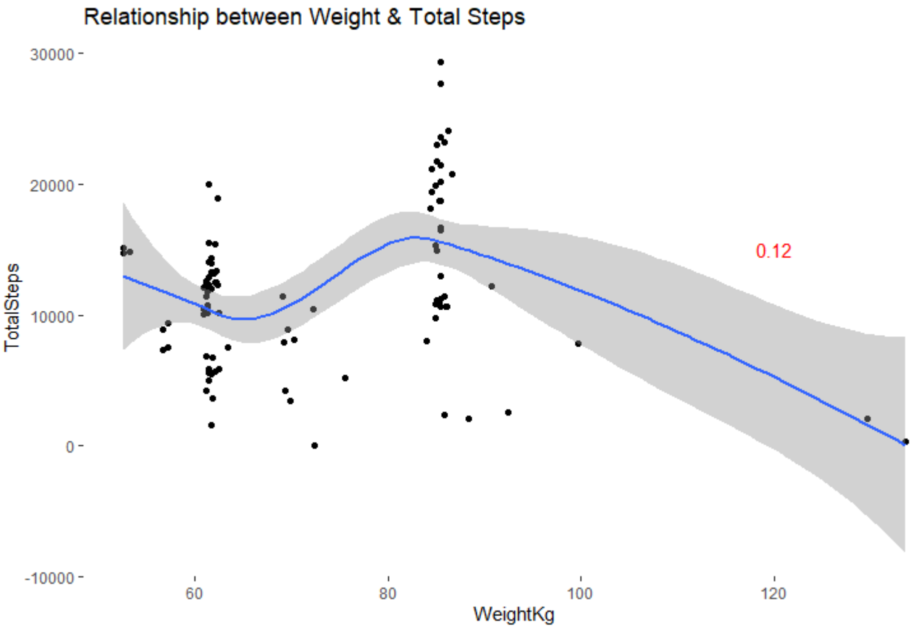
- BMI Categories by the National Heart, Lung, and Blood Institute (NHLBI) also indicate that only 31% of the 13 participants are considered healthy, with the remaining 69% being overweight or obese.



Weight Log - Relationships:

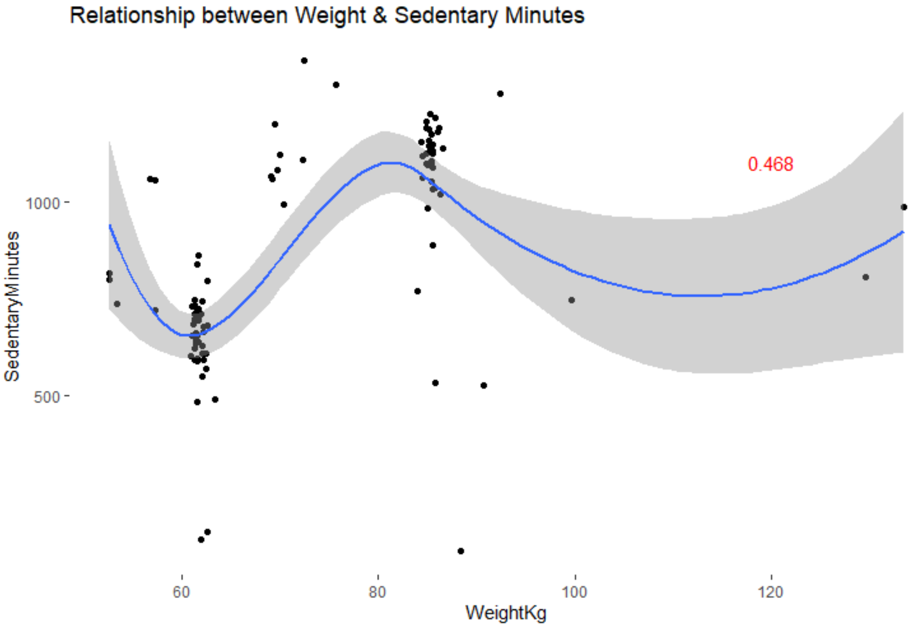
- Weight & Total Steps

    - The correlation coefficient shows 0.12, indicating no relationship between the weight of a participant and the total steps they took in a given day.



- Weight & Sedentary Minutes

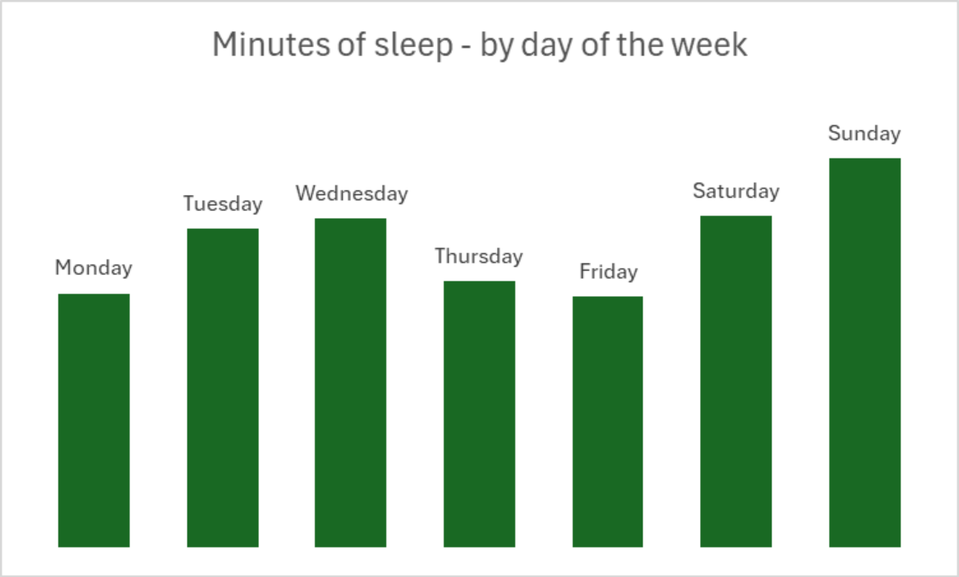
    - The correlation coefficient shows 0.468, indicating a weak positive correlation between the weight of a participant and the amount of time they spent in a sedentary state in a day.



There aren't a lot of data points in the weight category, as not many participants used this feature on their smart devices. However, the correlation found can be backed up by other studies. "Prolonged time spent sedentary decreases energy expenditure and displaces light-intensity physical activities, potentially leading to weight gain over time." (Energy Balance and Obesity, IARC, 2015). This insight also helps in the marketing strategy of our selected product, since clients would greatly benefit from reminders to start some form of activity that breaks sedentarism.

Sleep Tracker:

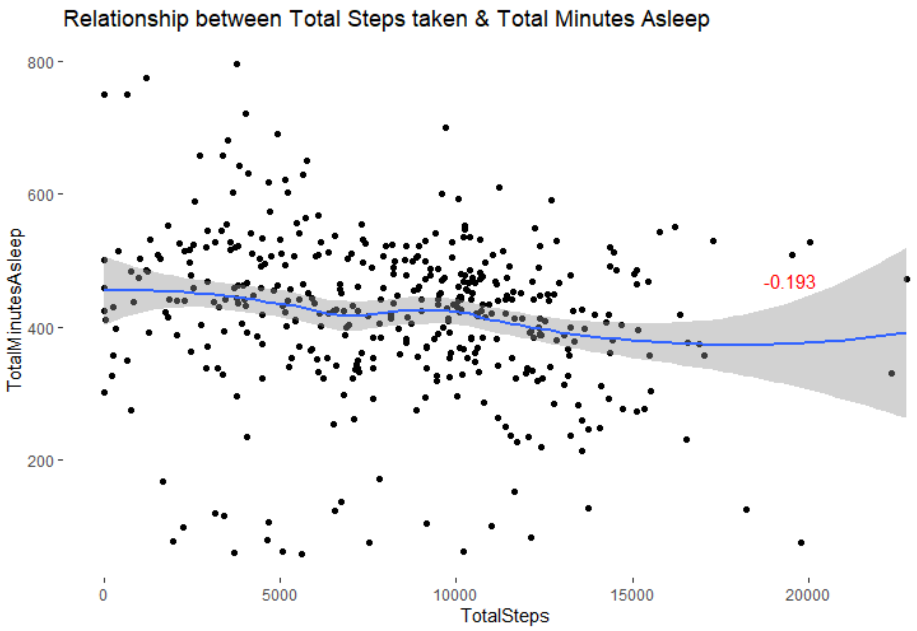
- 25 participants tracked their sleep data during the study.
- 20 participants consistently tracked their sleep on the app for more than 7 days. This means that more than half (57%) of the participants tracked their sleep data for over a week.
- On average, participants slept more during the weekends, though surprisingly, Tuesday and Wednesday were not so far off.



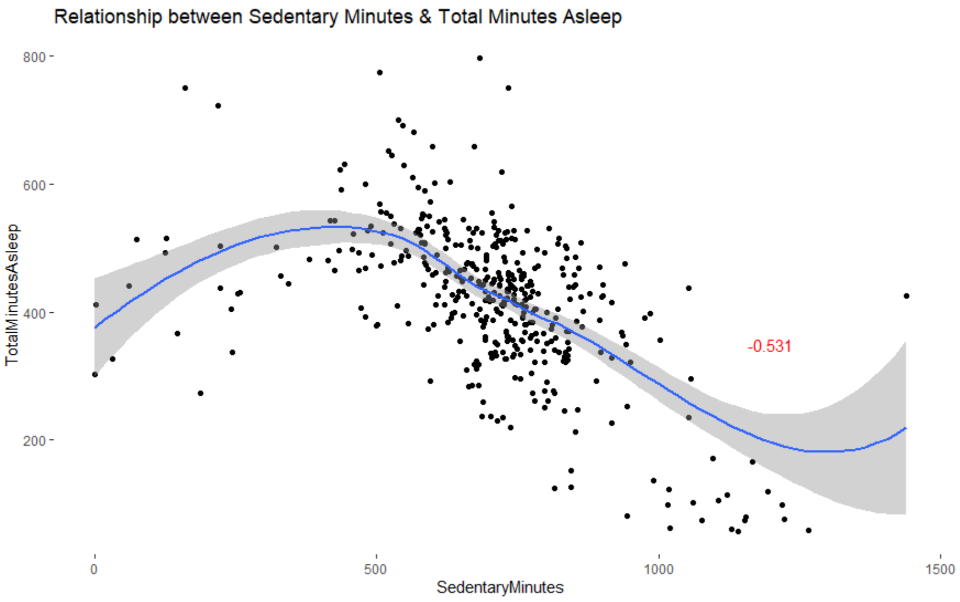
- The average sleep per session was 366 minutes or 6.1 hours of sleep a day. According to the NHLBI, adults who sleep less than 7 hours a night may have more health issues than those who sleep 7 or more hours a night.

Sleep Tracker - Relationships:

- Sleep & Total Steps
    - The correlation coefficient shows -0.193, indicating no relationship between the time slept by a participant and the total steps they took in a given day.



- Sleep & Sedentary minutes
    - The correlation coefficient shows -0.531, indicating a weak negative relationship between the time spent being sedentary and the amount of sleep a participant had in a day.



This means that more exercise doesn't necessarily equate to more sleep. That said, the second graph shows a weak inverse relationship between time slept and time spent being sedentary, meaning the participants who spent more time in sedentary activities slept less time on average. This is also another potential vantage point for the "Time" marketing strategy.

Recommendations

- The analysis showed that participants are more likely to utilize smart device features that provide insight into their day-to-day lives. This means customers value statistics like daily steps, distance traveled, activity time, calories burned, sleep time, and sleep quality, among many others. It is then recommended to market the “Time” product as an everyday companion that will always be there with your daily stats, recommendations, and reminders.

- The study showed the underutilization of the weight log feature. Whether this was due to a lack of knowledge of its existence among the participants or an ease-of-use situation, it stands that a very small percentage of subjects used this feature. A 2017 study found that the initial motivation to download and use digital weight management interventions came from the perceptions of one’s physical attractiveness and wanting to improve overall health. Retention, on the other hand, came from: personalization, social support, feedback, ease of set-up and use, etc. The recommendation would then be to market the “Time” as a goal-achieving facilitator. With features like being fully customizable, seamless tracking, auto-logging, a community through the Bellabeat app, feedback, and much more.

- The last recommendation is to market the “Time” as a “habit optimizer”. Sleep is extremely important for overall health, and the participants of the study slept on average 6.1 hours per day, when the recommended sleep from experts at the National Institute of Health and the American Academy of Sleep Medicine is from 7-9 hours per day. Building a habit can be challenging without the right tools and discipline, but with the “Time” always on your wrist, providing reminders and suggestions throughout the day, it becomes a much simpler task.