Case Study 2:

How Can a Welness Technology Company Play It Smart?

Micael Alves

November 2023

Contents

Introduction

About the company

Bellabeat is a high-tech manufacturing company that specializes in women's health products. It was founded in 2013 by Urška Sršen and Sando Mur. Since then, Bellabeat has experienced remarkable growth, establishing itself as one of the most successful small tech companies in recent years.

Their product lineup includes Leaf (a wellness tracker), Time (a wellness watch), and Spring (a water bottle that tracks daily water intake). These devices monitor users' exercise regimens, sleep patterns, stress levels, menstrual cycle, and mindfulness habits.

Bellabeat's mission is to empower women with the knowledge to make informed decisions about their health and well-being. Through its personalized membership-based subscription service, Bellabeat provides clients with tailored guidance on nutrition, activity, sleep, health, beauty, and mindfulness, all aligned with their individual lifestyles and goals.

The company aspires to be a major player in the global market for smart devices. It has made significant investments in traditional advertising channels to ensure a strong brand presence, while also maintaining an active engagement with its customers and potential customers via social media platforms.

Methodology of the study

This study will follow Google Analytics' six-phase methodology: ask, prepare, process, analyze, share, and act.

Ask

Business Task

- Analyze non-Bellabeat smart device usage data to understand consumer preferences and trends.
- Identify key insights from the analyzed data to inform Bellabeat's marketing and product strategies.
- Select one Bellabeat product and apply these insights to improve its marketing positioning and messaging.
- Prepare a presentation summarizing the findings and recommendations for Bellabeat's decision-makers.

Key Stakeholders

- **Urška Sršen**: Bellabeat's cofounder and Chief Creative Officer
- **Sando Mur**: Mathematician and Bellabeat's co-founder; 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.

Prepare

The dataset to be used for analysis is public and is stored in Kaggle (https://www.kaggle.com/arashnic/fitbit). This dataset, consisting of 18 CSV files, includes minute-level physical activity, heart rate, and sleep monitoring data from 30 Fitbit users.

Limitations of the dataset

- Data may be outdated: the data was gathered in 2016, and users' routines may have changed since then, potentially affecting the relevance of the data for current analysis.
- Small sample size: A sample size of 30 Fitbit users is insufficient to represent the entire fitness market, potentially limiting the generalizability of the findings.
- We cannot guarantee the integrity or correctness of the data because it was acquired through a survey.
- No gender description: the absence of gender information fails to align with Bellabeat's primary focus on women. A dataset predominantly composed of female participants would have been more relevant for the analysis.

Is data ROCC (Reliable, Original, Comprehensive, Current, Cited)?

ROCCC	Assessment (Low-Medium-High)	Rationale
Reliable	Low	The small sample size of 30 respondents raises concerns about the reliability of the data for representing the broader fitness market.
Original	Low	The data was collected through a third-party provider, reducing its originality and potentially introducing biases.
Comprehensive	Medium	The data parameters align with the majority of metrics tracked by

		Bellabeat devices, providing some level of comprehensiveness.
Current	Low	The data was collected in 2016, raising concerns about its relevance and applicability to current user behavior patterns and fitness trends.
Cited	High	Amazon Mechanical Turk is a well-known source, and the data is well documented.

Process

I've chosen to work with six files to keep the project manageable. The sets I'll be working with are:

- dailyActivity_merged.csv
- dailyCalories_merged.csv
- dailyIntensities_merged.csv
- sleepDay_merged.csv
- dailySteps_merged.csv
- weightLogInfo_merged.csv

Before uploading the csv files to BigQuery tables, I cleaned the data/time zone in Excel, removing the AM/PM.

Then, I created separate tables in BigQuery, and counted the distinct Ids in each table to find out if all of them have the same amount of data.

```
SELECT
COUNT (DISTINCT Id)
FROM `gleaming-glass-383714.FitBit.DailyActivity`
```

dailyActivity_merged.csv: 33

```
SELECT
COUNT (DISTINCT Id)
FROM `gleaming-glass-383714.FitBit.Calories`
```

dailyCalories_merged.csv: 33

```
SELECT
COUNT (DISTINCT Id)
FROM `gleaming-glass-383714.FitBit.Intensities`
```

dailyIntensities_merged.csv: 33

```
SELECT
COUNT (DISTINCT Id)
FROM `gleaming-glass-383714.FitBit.SleepDay`
```

sleepDay_merged.csv: 24

```
SELECT
COUNT (DISTINCT Id)
```

```
FROM `gleaming-glass-383714.FitBit.Steps`
```

dailySteps_merged.csv: 33

```
SELECT
COUNT (DISTINCT Id)
FROM `gleaming-glass-383714.FitBit.WeightLog`
```

weightLogInfo_merged.csv: 8

The data is inconsistent; we expected to observe 30 unique IDs in each table.

We want to understand how this data correlates with Bellabeat users. To do this, we can calculate average activity values for each participant to facilitate comparisons.

<u>So, what is the average level of activity among the FitBit users?</u>

```
 SELECT
  Id,
  AVG(FairlyActiveMinutes + VeryActiveMinutes) AS avg_user_activity
FROM
  `gleaming-glass-383714.FitBit.DailyActivity`
 GROUP BY
  Id
ORDER BY
  avg_user_activity;
```

Twenty users are getting at least 20 minutes of fairly active minutes. From those, six users are getting more than 1 hour of activity (on average).

Id	avg_user_activity
2026352035	0.35483870967741932
1844505072	1.4193548387096775
1927972279	2.0967741935483875
4057192912	2.25
6117666160	3.6071428571428577
2320127002	3.9354838709677415
8792009665	5.0000000000000009
6290855005	6.5517241379310347
4445114986	8.35483870967742
4020332650	10.548387096774192
3372868164	13.249999999999998
1624580081	14.483870967741936
4319703577	15.903225806451614
2873212765	20.225806451612897
4558609924	24.096774193548388
6775888955	25.807692307692303
1644430081	30.933333333333326
4702921684	31.161290322580644
8583815059	31.870967741935488
2347167796	34.055555555555557
8253242879	34.8421052631579
5553957443	36.41935483870968
6962181067	41.322580645161295
4388161847	43.516129032258057
7007744171	47.307692307692307
2022484408	55.645161290322584
1503960366	57.87096774193548
7086361926	67.93548387096773
8378563200	68.93548387096773
8877689391	76.0
3977333714	80.166666666666671
8053475328	94.741935483870961
5577150313	117.16666666666667

Now, we want to determine the average amount of sleep that users typically get

```
SELECT Id,
AVG(TotalMinutesAsleep)/60 AS avg_user_sleep
FROM `gleaming-glass-383714.FitBit.SleepDay`
GROUP BY Id
ORDER BY (avg_user_sleep)
```

Id	avg_user_sleep
2320127002	1.0166666666666666
7007744171	1.1416666666666666
4558609924	2.1266666666666665
3977333714	4.89404761904762
1644430081	4.9
8053475328	4.95
4020332650	5.822916666666667
6775888955	5.8277777777777784
1503960366	6.0046666666666662
4445114986	6.4196428571428577
4388161847	6.7187499999999991
1927972279	6.95
4702921684	7.0190476190476181
5577150313	7.2
8792009665	7.261111111111112
8378563200	7.3890625
2347167796	7.4466666666666672
6962181067	7.4666666666666668
7086361926	7.552083333333333
5553957443	7.7247311827956979
4319703577	7.9442307692307708
6117666160	7.9796296296296285
2026352035	8.4363095238095234
1844505072	10.866666666666667

12 users recorded an average of 7 hours of sleep. Another 12 users recorded sleep durations of less than 7 hours, with some users indicating an average sleep of 1 hour, due to incomplete data recording throughout the month. To investigate further, we can explore additional relationships between users' sleep patterns and their activity levels.

Then, we transitioned to **R Studio**, where we conducted further data manipulation and analysis.

We started by installing and loading the packages:

```
install.packages("tidyverse")
install.packages("lubridate")
install.packages("dplyr")
install.packages("ggplot2")
install.packages("tidyr")

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

Then, we imported the data:

```
> daily_activity = read.csv("dailyActivity_merged.csv")
> daily_calories = read.csv("dailyCalories_merged.csv")
> daily_intensity = read.csv("dailyIntensities_merged.csv")
> daily_steps = read.csv("dailySteps_merged.csv")
> daily_sleep = read.csv("sleepDay_merged.csv")
> weight_log = read.csv("weightLogInfo_merged.csv")
```

Afterwards, we checked if there are any missing entries in the data, and verified the unique Ids:

```
head(daily_activity)
missing_activity <-colSums(is.na(daily_activity))
missing_activity

> distinct_ids <- daily_activity %>%
+     summarise(distinct_ids = n_distinct(Id))
> print(distinct_ids$distinct_ids)
[1] 33
```

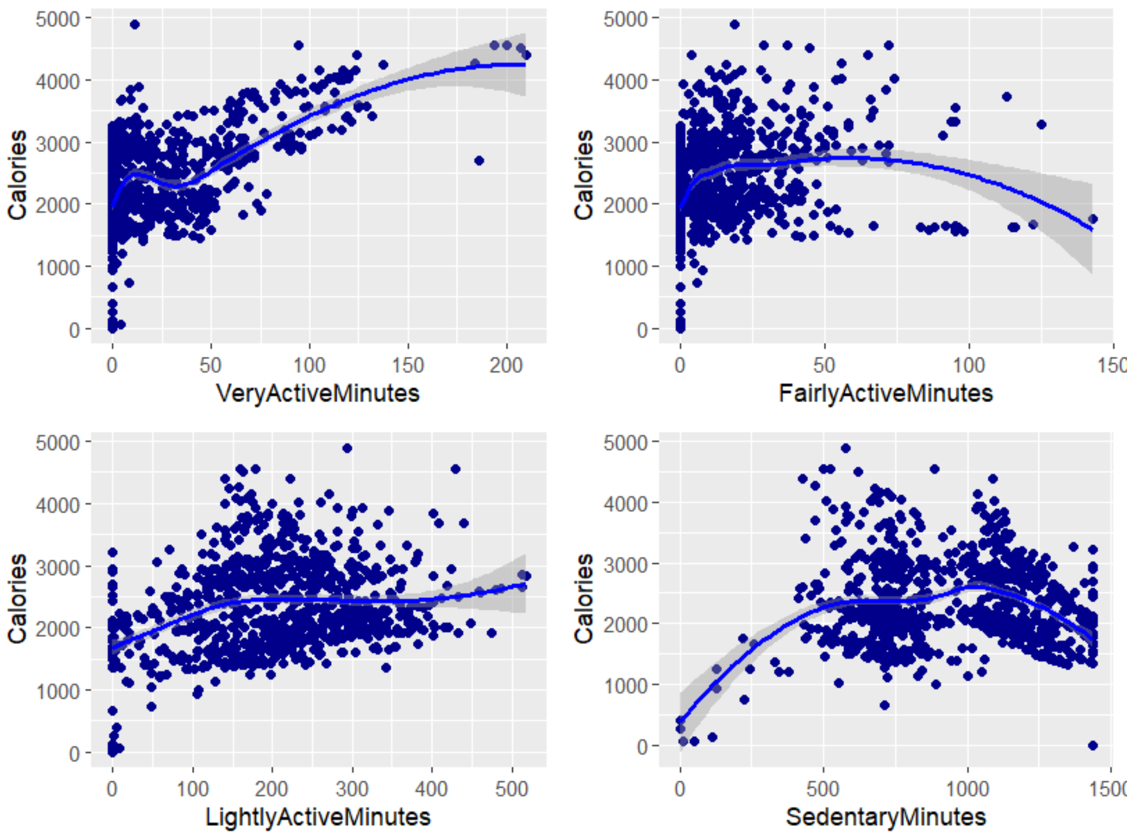
Visualize

We then repeated the code for the other CSV files. In the end, we found the data to be consistent with our SQL analysis and are now prepared to summarize and visualize it.

```
plot1 <- ggplot(data=daily_activity)+ geom_point(mapping=aes(x=VeryActiveMinutes,
y=Calories), color="darkblue") +
  geom_smooth(mapping=aes(x=VeryActiveMinutes, y=Calories),color="blue")

plot2 <-ggplot(data=daily_activity)+ geom_point(mapping=aes(x=FairlyActiveMinutes,
y=Calories), color="darkblue") +
  geom_smooth(mapping=aes(x=FairlyActiveMinutes, y=Calories),color="blue")

plot3 <-ggplot(data=daily_activity)+ geom_point(mapping=aes(x=LightlyActiveMinutes,
y=Calories), color="darkblue") +
  geom_smooth(mapping=aes(x=LightlyActiveMinutes, y=Calories),color="blue")

plot4 <-ggplot(data=daily_activity)+ geom_point(mapping=aes(x=SedentaryMinutes,
y=Calories), color="darkblue") +
  geom_smooth(mapping=aes(x=SedentaryMinutes, y=Calories),color="blue")
```

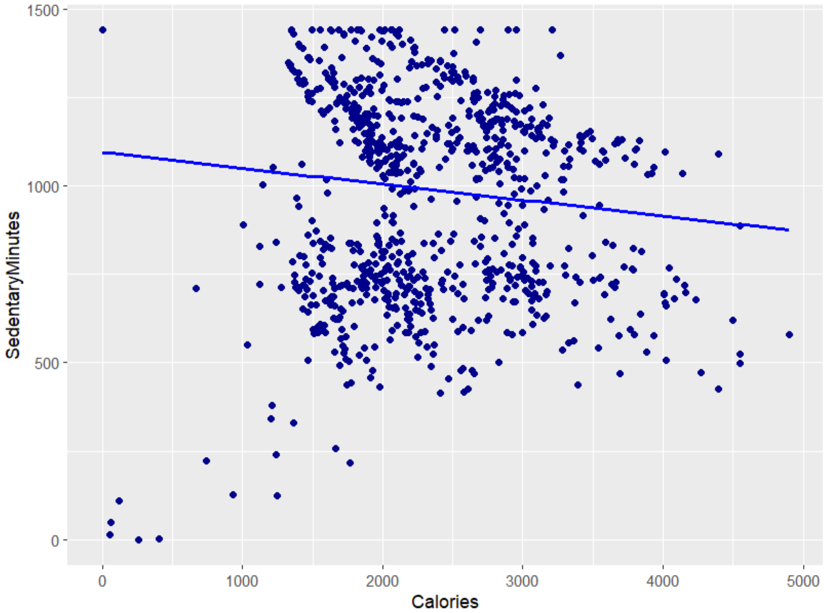
Then, to show them in a grid, we used gridExtra:

```
install.packages("gridExtra")
library(gridExtra)

grid.arrange(plot1, plot2, plot3, plot4, ncol = 2)
```



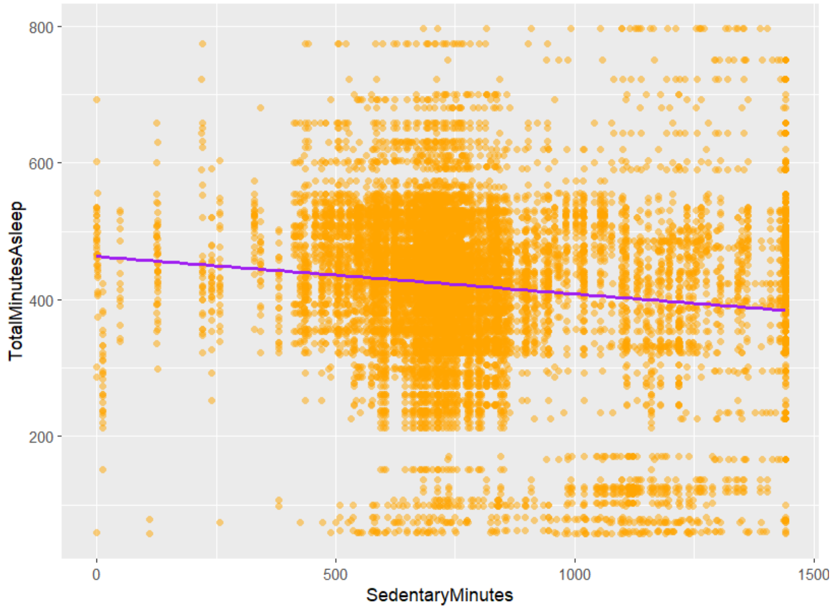
Active minutes have a positive correlation with the total calories burned, indicating a direct relationship. Contrariwise, there is an inverse relationship between calories burned and sedentary minutes.

```
daily_activity$calories <- daily_calories$calories
head(daily_activity)
ggplot(daily_activity, aes(x = Calories, y = SedentaryMinutes)) +
  geom_point(color = "darkblue", shape = 16) +
  geom_smooth(method = "lm", se = FALSE, color = "blue")
```



Likewise, we can determine whether there exists a correlation between sleep and physical activity.

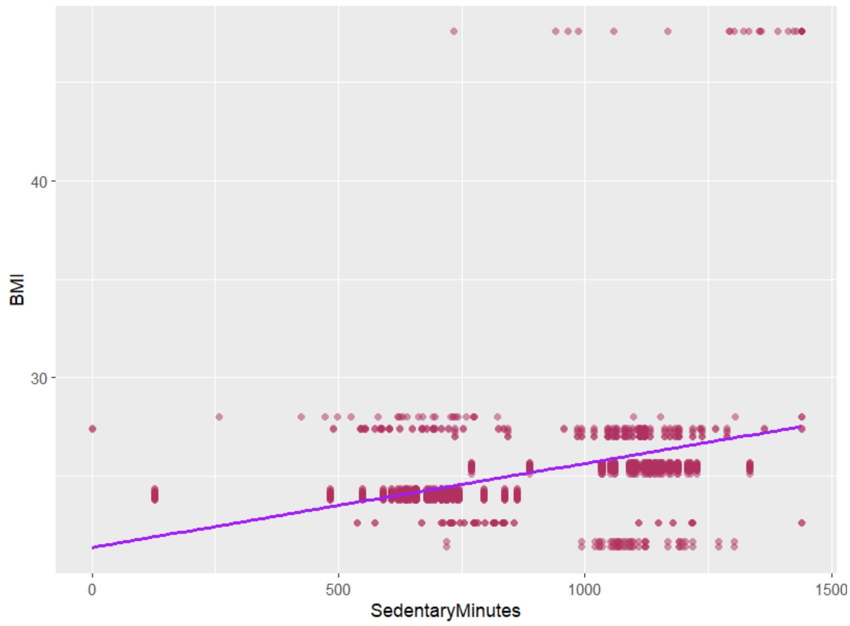
```
merged_data <- merge(daily_activity, daily_sleep[, c("Id", "TotalMinutesAsleep")], by = "Id", all.x = TRUE)
head(merged_data)
ggplot(merged_data, aes(x = SedentaryMinutes, y = TotalMinutesAsleep)) +
 geom_point(color = "orange", shape = 16, alpha = 0.5) +
 geom_smooth(method = "lm", se = FALSE, color = "purple")
cor(daily_activity$TotalMinutesAsleep, daily_activity$SedentaryMinutes)
```



The chart indicates an inverse relationship between sleep quality and activity levels. Individuals who spend more time in sedentary activities tend to experience fewer hours of sleep.

We also decided to compare BMI with Activity levels:

```
merged_data <- merge(daily_activity, weight_log[, c("Id", "BMI")], by = "Id", all.x = TRUE)
head(merged_data)
ggplot(merged_data, aes(x = SedentaryMinutes, y = BMI)) +
  geom_point(color = "maroon", shape = 16, alpha = 0.5)  +
  geom_smooth(method = "lm", se = FALSE, color = "purple")
```



The chart reveals a correlation between higher sedentary behavior and higher BMI.

Findings

Based on the analysis, a significant number of users either did not record data or, on average, are predominantly sedentary.

An increase in active minutes corresponds to higher calorie burn, whereas sedentary minutes correlate with increased calorie intake. Individuals with sedentary lifestyles tend to exhibit poor sleep patterns.

Moreover, a person with minimal physical activity is more likely to have a higher BMI (and vice-versa).

Recommendations for Bellabeat:

Sleep Reminder or Tracker Feature:

Introduce a sleep reminder or tracker feature since half of the users are getting less than 7 hours of sleep.

Engage Sedentary Users in High Activity Minutes:

Users who are sedentary are also experiencing sleep debt. Consider sending alerts encouraging them to engage in high activity minutes right before their downtime.

Nutritional Resources for High BMI Users:

Provide recipe cards, leverage social media marketing, or offer nutritional resources to support users struggling with a high BMI, which constitutes over half of the user base.

Set Limits on Sedentary Minutes:

Implement a daily limit for sedentary minutes to motivate users to incorporate more movement into their routines, aiding in calorie loss.