Adam Zdanowski
3/29/2022

Google Data Analytics Certification
Case Study: Bellabeat

Overview

Bellabeat is a technology company specializing in fitness-related products for women. The business is looking to gain insights into how real users utilize similar products that are made by competing brands. Working as an analyst for Bellabeat, I will be analyzing a public data set about smartwatch usage for non-Bellabeat products. I will seek to answer several key questions posed by Bellabeat management and provide high-level recommendations about the possible directions the company can go in the future.

Case Study Roadmap - Ask

To begin my analysis, the first step will be to determine the business task and stakeholders of the project at hand.

Bellabeat is looking for information regarding the usage of similar smart devices made by competitors, specifically Fitbit. By reviewing and analyzing how consumers use this brand's products, Bellabeat will be able to institute some of the same practices and features into its products to further drive up sales and customer satisfaction by providing products consumers desire. Specific questions were asked about smart device usage trends, potential applications of discovered trends within Bellabeat's products, and potential implications the trends could have on the marketing strategy of Bellabeat and their product line as a whole.

Key stakeholders within the company include Urška Sršen (co-founder and CCO) and Sando Mur (co-founder and marketing executive). Also, the marketing team at Bellabeat that I am working with will also be a frequent contact throughout this project with a direct interest in the resulting takeaways. However, the most important stakeholder is likely to be found external to the company in the form of potential consumers. After all, Bellabeat is trying to learn about customer trends in an effort to better appeal to prospective consumers.

Case Study Roadmap - Prepare

The next step in the process will be to prepare the data for analysis. To start this, I must first explore the dataset I will be working with, looking for potential sources of bias and other limitations. This publicly available dataset is found at "FitBit Fitness Tracker Data" (https://www.kaggle.com/arashnic/fitbit). The data found here represent FitBit data from 30 users who agreed to participate in the study and contains information like calories burned, heart rate, steps, sleep data, and more.

The FitBit database I will be using does not have any issues with licensing or privacy as it is a public dataset available through and stored in Kaggle under the CC0: Public Domain License. Also, all participants in the study are told to have consented to take part in data collection.

The given dataset does include information greater than what I will need to answer the questions posed, so I will omit certain subsets of the data. Because the scope of all of the questions is focused on the *daily* habits of users, I will focus on the daily sets of Activity, Calories, Intensities, Steps, Sleep, and Weight Log.

To sort, organize, analyze, and ultimately present my findings I will be using Google Sheets, RStudio, and Tableau.

To get a quick overview, I begin by loading the required tables into RStudio and previewing them using the head() function.

```
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, LoggedActivitiesDist...

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.
> View(dailyActivity_merged)
> library(readr)

> dailyCalories_merged <- 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.*
> View(dailyCalories_merged)
> library(readr)

> dailyIntensities_merged <- 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, Very...

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.*
> View(dailyIntensities_merged)
> library(readr)

> dailySteps_merged <- 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.*
> View(dailySteps_merged)
> library(readr)

> minuteSleep_merged <- 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.*

```
> View(minuteSleep_merged)
Session restored from your saved work on 2022-Mar-29 13:37:53 UTC (52 minutes ago)
> library(readr)

 daily_sleep <- read_csv('sleepDay_merged.csv')
Rows: 413 Columns: 5
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (1): SleepDay
dbl (4): Id, TotalSleepRecords, TotalMinutesAsleep, TotalTimeInBed

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.
>
> weight_log <- read_csv('weightLogInfo_merged.csv')
Rows: 67 Columns: 8
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (1): Date
dbl (6): Id, WeightKg, WeightPounds, Fat, BMI, LogId
lgl (1): IsManualReport

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.
>

head(daily_sleep)
# A tibble: 6 × 5
        Id SleepDay              TotalSleepRecor… TotalMinutesAsl… TotalTimeInBed
     <dbl> <chr>                            <dbl>           <dbl>          <dbl>
1 1503960366 4/12/2016 12:00:00 AM             1             327            346
2 1503960366 4/13/2016 12:00:00 AM             2             384            407
3 1503960366 4/15/2016 12:00:00 AM             1             412            442
4 1503960366 4/16/2016 12:00:00 AM             2             340            367
5 1503960366 4/17/2016 12:00:00 AM             1             700            712
6 1503960366 4/19/2016 12:00:00 AM             1             304            320

> head(dailyActivity_merged)
# A tibble: 6 × 15
        Id ActivityDate TotalSteps TotalDistance TrackerDistance LoggedActivitiesDistan… VeryActiveDista…
ModeratelyActiv… LightActiveDist…
     <dbl> <chr>             <dbl>         <dbl>           <dbl>                   <dbl>            <dbl>
1 1503960366 4/12/2016      13162          8.5             8.5                     0                1.88
2 1503960366 4/13/2016      10735          6.97            6.97                    0                1.57
3 1503960366 4/14/2016      10460          6.74            6.74                    0                2.44
```

```
4 1503960366 4/15/2016       9762       6.28       6.28              0       2.14       1.26       2.83
5 1503960366 4/16/2016       12669       8.16       8.16              0       2.71       0.410       5.04
6 1503960366 4/17/2016       9705       6.48       6.48              0       3.19       0.780       2.51
# … with 6 more variables: SedentaryActiveDistance <dbl>, VeryActiveMinutes <dbl>, FairlyActiveMinutes <dbl>,
LightlyActiveMinutes <dbl>,
#   SedentaryMinutes <dbl>, Calories <dbl>

> head(dailyCalories_merged)
# A tibble: 6 × 3
        Id ActivityDay Calories
     <dbl> <chr>        <dbl>
1 1503960366 4/12/2016      1985
2 1503960366 4/13/2016      1797
3 1503960366 4/14/2016      1776
4 1503960366 4/15/2016      1745
5 1503960366 4/16/2016      1863
6 1503960366 4/17/2016      1728

> head(dailyIntensities_merged)
# A tibble: 6 × 10
       Id ActivityDay SedentaryMinutes LightlyActiveMi… FairlyActiveMin… VeryActiveMinut… SedentaryActive…
LightActiveDist… ModeratelyActiv…
    <dbl> <chr>              <dbl>        <dbl>        <dbl>        <dbl>        <dbl>        <dbl>        <dbl>
1   1.50e9 4/12/2016           728          328           13           25            0         6.06        0.550
2   1.50e9 4/13/2016           776          217           19           21            0         4.71        0.690
3   1.50e9 4/14/2016          1218          181           11           30            0         3.91        0.400
4   1.50e9 4/15/2016           726          209           34           29            0         2.83        1.26
5   1.50e9 4/16/2016           773          221           10           36            0         5.04        0.410
6   1.50e9 4/17/2016           539          164           20           38            0         2.51        0.780
# … with 1 more variable: VeryActiveDistance <dbl>

> head(dailySteps_merged)
# A tibble: 6 × 3
        Id ActivityDay StepTotal
     <dbl> <chr>          <dbl>
1 1503960366 4/12/2016     13162
2 1503960366 4/13/2016     10735
3 1503960366 4/14/2016     10460
4 1503960366 4/15/2016      9762
5 1503960366 4/16/2016     12669
6 1503960366 4/17/2016      9705
```

> head(weight_log)
 # A tibble: 6 × 8

Id	Date	WeightKg	WeightPounds	Fat	BMI	IsManualReport	LogId
<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<lgl>	<dbl>
1 1503960366	5/2/2016 11:59:59 PM	52.6	116.	22	22.6	TRUE	1462233599000
2 1503960366	5/3/2016 11:59:59 PM	52.6	116.	NA	22.6	TRUE	1462319999000
3 1927972279	4/13/2016 1:08:52 AM	134.	294.	NA	47.5	FALSE	1460509732000
4 2873212765	4/21/2016 11:59:59 PM	56.7	125.	NA	21.5	TRUE	1461283199000
5 2873212765	5/12/2016 11:59:59 PM	57.3	126.	NA	21.7	TRUE	1463097599000
6 4319703577	4/17/2016 11:59:59 PM	72.4	160.	25	27.5	TRUE	1460937599000

Seeing as though values in the first column do not repeat, the data is organized in the wide format.

In order to determine the validity and strength of this dataset as a source, a ROCCC analysis will be conducted.

Reliable - This dataset was sourced using Amazon's Mechanical Turk and therefore not subject to official governmental or educational oversight. No standard collection methods were given so this data cannot be assumed to be 100% reliable.
Original - This data is not original as it is data that was sourced through external methods previously discussed (Kaggle).
Comprehensive - The actual fitness data is adequately comprehensive for the purpose of this study, the breadth of the data collected is surely enough to find the answers that Bellabeat is looking for. However, do keep in mind that the data is based only on 30 subjects (from unknown demographics), a number that we would ideally like to be greater in order to get a better overall picture.
Current - This dataset frequently includes the date of collection as a variable throughout the data gathering process. From this, we can see that the majority of this data was collected from 3/12/16-5/12/16 which is fairly recent and is expected to be updated annually. However, it is worth noting that in the technology space, 6 years is a long time and FitBit likely has improved its products since that time.
Cited - This dataset is also cited, all collection credits go to:
Furberg, Robert; Brinton, Julia; Keating, Michael ; Ortiz, Alexa
https://zenodo.org/record/53894#.YMoUpnVKiP9

Final Note: This data has been recorded by FitBit devices with no inherent collection bias. However, please keep in mind that the study was not subject to oversight, no demographic information was provided about subjects other than gender, and only 30 participants were included in this study. Please take these concerns into account when considering the results of this analysis.

Case Study Roadmap - Process

Parsing through this amount of data will certainly require the help of tools. For this assignment, I will be using a combination of Google Sheets, Tableau, and RStudio.

> **Google Sheets**: the given data is small enough to allow the use of spreadsheets rather than SQL, here I will be able to sort, organize, and clean data in a familiar format as well as make simple visualizations.
> **Tableau:** using this software will give me the greatest range of options when it comes to making data visualizations. Google Sheets or even RStudio would likely be sufficient given the requirements, though I feel Tableau gives me the most creative liberty.
> **RStudio:** I will use this to fill in the gaps between Sheets and Tableau. R offers the best of both worlds between coding/analyzing and visualizing and I will use it whenever that becomes necessary to have.

Before being able to analyze it, I will clean the data to get rid of anything I don't need, as well as ensure everything is in the same format as it should be. To start, I will get all the CSV files I need into one Google Slides file to keep them all in one spot. From here I will do some minor formatting to make the documents easier to read (Header clarification, column sizing, font formatting, etc) After this, I will be able to start cleaning the actual data. A log of changes made follows

Activity

- Standardize Decimal Places to 2 (Columns D, E, F, G, H, I, J)
- Used Conditional Formatting to check for non-numerical, null, and missing values
- Changed Column B to Date Data Type
- Ran Remove Duplicates Function (None Found)
- Validated all data collected fell into correct date ranges (3/12/16-5/12/16)
- Checked total number of entries for consistency (940)
- Used the CountUnique Function to determine the number of participants contributing to this table (33)

Sleep

- Changed Column B to Date/Time Data Type
- Added New Column “HoursAsleep” using MinutesAsleep/60 rounded to one decimal place
- Added New Column “HoursInBed” using TotalTimeInBed/60 rounded to one decimal place
- Reviewed columns for non-numerical, null, and missing values
- Ran Remove Duplicates Function (3 Found and Removed)
- Validated all data collected fell into correct date ranges (3/12/16-5/12/16)
- Checked total number of entries for consistency (411)
- Used the CountUnique Function to determine the number of participants contributing to this table (24)

Steps

- Changed Column B to Date Data Type
- Reviewed columns for non-numerical, null, and missing values
- Ran Remove Duplicates Function (None Found)
- Validated all data collected fell into correct date ranges (3/12/16-5/12/16)
- Checked total number of entries for consistency (940)
- Used the CountUnique Function to determine the number of participants contributing to this table (33)

Intensity

- Changed Column B to Date Data Type
- Standardize Decimal Places to 2 (Columns G, H, I, J)
- Ran Remove Duplicates Function (None Found)
- Reviewed columns for non-numerical, null, and missing values
- Validated all data collected fell into correct date ranges (3/12/16-5/12/16)
- Checked total number of entries for consistency (940)
- Used the CountUnique Function to determine the number of participants contributing to this table (33)

Calories

- Changed Column B to Date Data Type
- Reviewed columns for non-numerical, null, and missing values
- Ran Remove Duplicates Function (None Found)
- Validated all data collected fell into correct date ranges (3/12/16-5/12/16)
- Checked total number of entries for consistency (940)
- Used the CountUnique Function to determine the number of participants contributing to this table (33)

Weight

- Changed Column B to Date/Time Data Type
- Standardize Decimal Places to 2 (Columns C, D, F)
- Reviewed columns for non-numerical, null, and missing values
    - Left the "Fat" column alone, accepting only two responses
- Ran Remove Duplicates Function (None Found)
- Validated all data collected fell into correct date ranges (3/12/16-5/12/16)
- Checked total number of entries for consistency (67)
- Used the CountUnique Function to determine the number of participants contributing to this table (8)

Data Processing Takeaways

While the datasets did need some minor alterations, on the whole, they were largely clean and ready to use. There were a few issues that stakeholders should be aware of, however. First, for the Weight and Daily Sleep tables, there were far fewer total entries (67 and 410) than there were for the other tables (940). Similarly, far fewer participants reported data for these two tables as well, with 8 participants for weight and 24 for daily sleep as compared to the 33 recorded in the other tables. While this does raise small questions about the validity of the data, it also provides insights into the user's habits. Accordingly, I decided to not omit any data and leave these inconsistencies as they are, because I find them valuable. One final thing worth mentioning is that there were only 2 entries in the weight log that included a value for Fat out of the 67 total data points. Rather than delete the column altogether, I again decided to leave this information as it was because I believe it highlights another interesting point about the usage of the products. Overall, with the acceptance of these inconsistencies, comprehensive cleaning and validating, and minor formatting and data type changes, I am left with several datasets that are clean, valid, and ready to work with and analyze.

Case Study Roadmap - Analyze

**Note: while this section is typically reserved just for analysis, I am going to combine this phase alongside the Share phase. Here I will combine my thoughts during analysis as well as some findings and visualizations. Additionally, I will attach a more formal presentation that would be along the lines of what I would actually present for the Share phase.*

Now that the data has been appropriately cleaned, it is time to perform an analysis on it. For this portion, I decided to stick with the single Google Sheets Document with 6 pages for the datasets I am choosing to analyze (Activity, Calories, Intensity, Steps, Sleep, and Weight). Through this, RStudio, and Tableau, I will have more than enough tools to utilize to properly find answers to questions posed by my stakeholders. Speaking of questions, to revisit them, Bellabeat is interested in finding out:

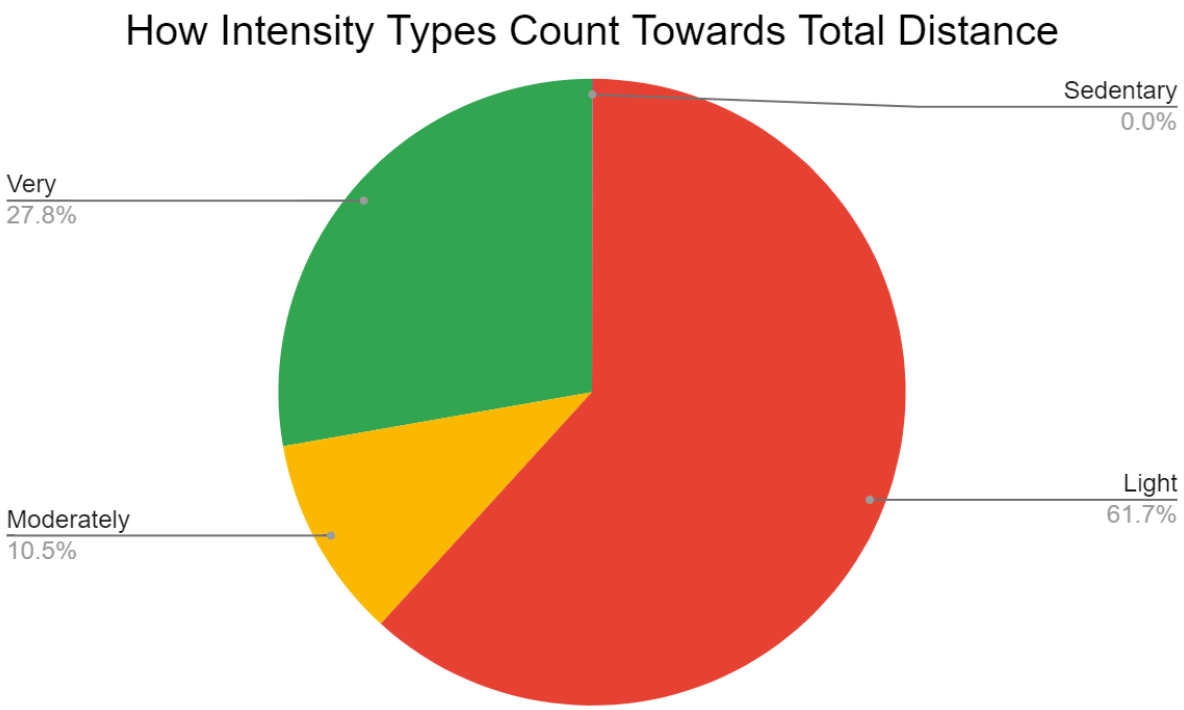
- What are some trends in smart device usage?
- How could these trends apply to Bellabeat customers?
- How could these trends help influence Bellabeat marketing strategy?

These questions are rather broad and give little indication of any specific directions so I attacked them and the data from all angles. To start, I calculated a few summary statistics for every quantitative data point including minimum, maximum, mean, median, and standard deviation. This is a quick way to allow me to understand the data within these tables at a high level. From here, I will have to get more specific. A few areas I will focus on are

- Various Intensities
- Sleep Function Usage
- Average Sleep vs Calories
- Average Sleep vs Steps

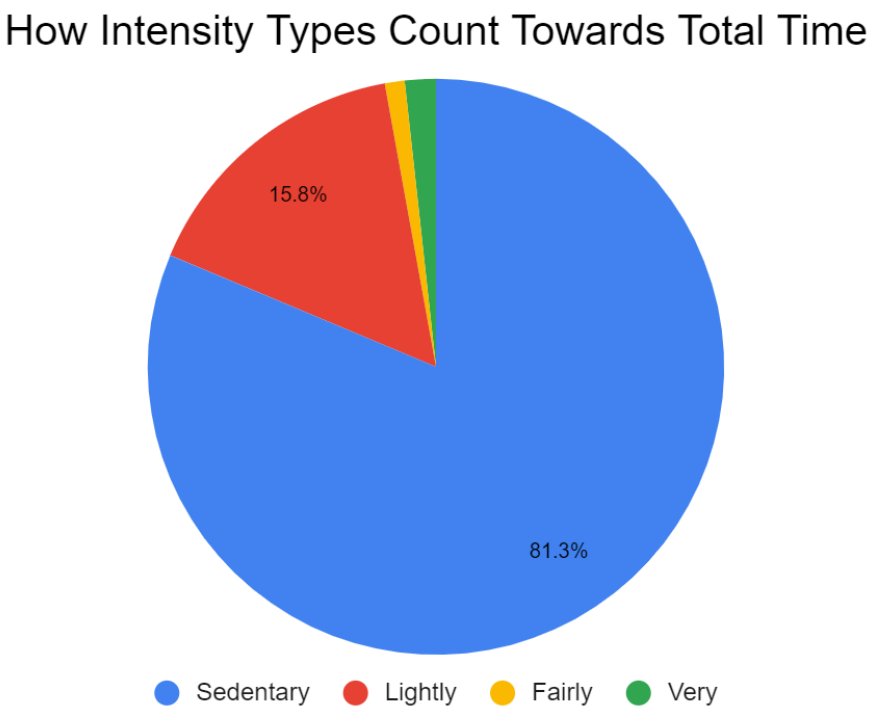
Intensity Breakdown

First off, I was interested in what types of activities were being recorded most frequently and which type accounted for most of the total. To look deeper into this issue, I will be using Pivot Charts and Google Sheets Charts. I hypothesized that with FitBits being among the lower-cost smart tracking options, it would be mainly used by everyday users who likely would not consider themselves to be fitness enthusiasts, and therefore "Sedentary" and "Lightly Active" would make up the majority of both distance and minutes. Upon looking at the chart titled "How Intensity Types Count Towards Total Distance" I was not surprised to find that Light Activity contributes the overwhelming majority (61.7%) of the total distance recorded by all FitBit users at this time. This category likely consists of most of the users' daily motion around the house, while at work, and even some light forms of cardio such as walking on the treadmill or leisurely bike rides. I was surprised to see how large of a portion the "Very Active Distance" made up of this chart at 27.8%. Upon further digging, I discovered that of the 1,412 miles of distance covered by the thirty-three participants while in the Very Active state, 563 of them were due to the efforts of just three users. In other words, 9% of the users accounted for nearly 40% of the active minutes. This tells me that while there is certainly a small percentage of outliers, the majority of the users of this product are 'casual' fitness users.



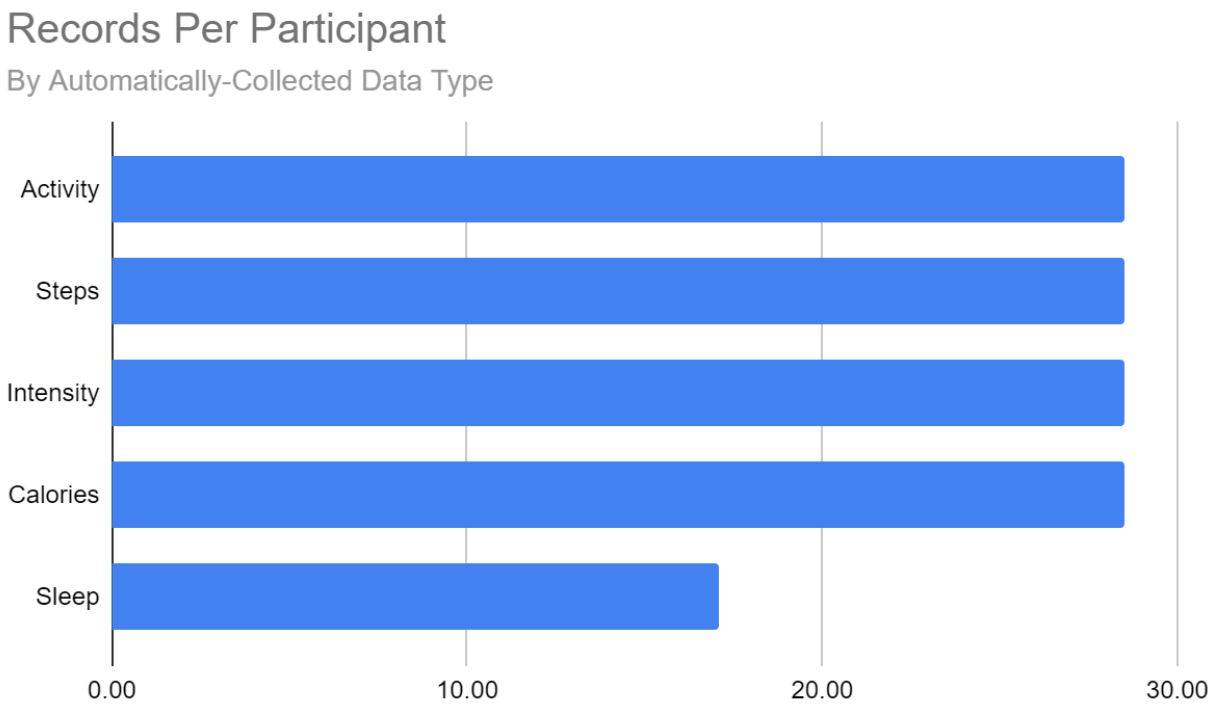
Top N	Very Active Distance	% of Total Very Active Distance
3	563.13	39.87%
5	737.96	52.24%
10	1056.84	74.82%
33	1412.52	100.00%

This takeaway was again confirmed by the same form of analysis on which types of intensity make up the total time spent by users, perhaps even more dramatically. Of all of the minutes spent using these smart devices, users were in the Sedentary state for 81.3 percent of them! When adding in the time spent in the Lightly Active state, the figure shoots up to 97.1%! While these may be ostentatious figures at a glance, they do make sense when you step back. These two states *should* account for the majority of the user's time, they include time spent working at a desk, driving a car, doing daily household activities, and, perhaps most importantly, sleeping (which will be covered in a later section). With Very Active and Fairly Active accounting for less than 3% of the total time that participants wore a smart device, it's clear that the majority of their users are everyday fitness goers.



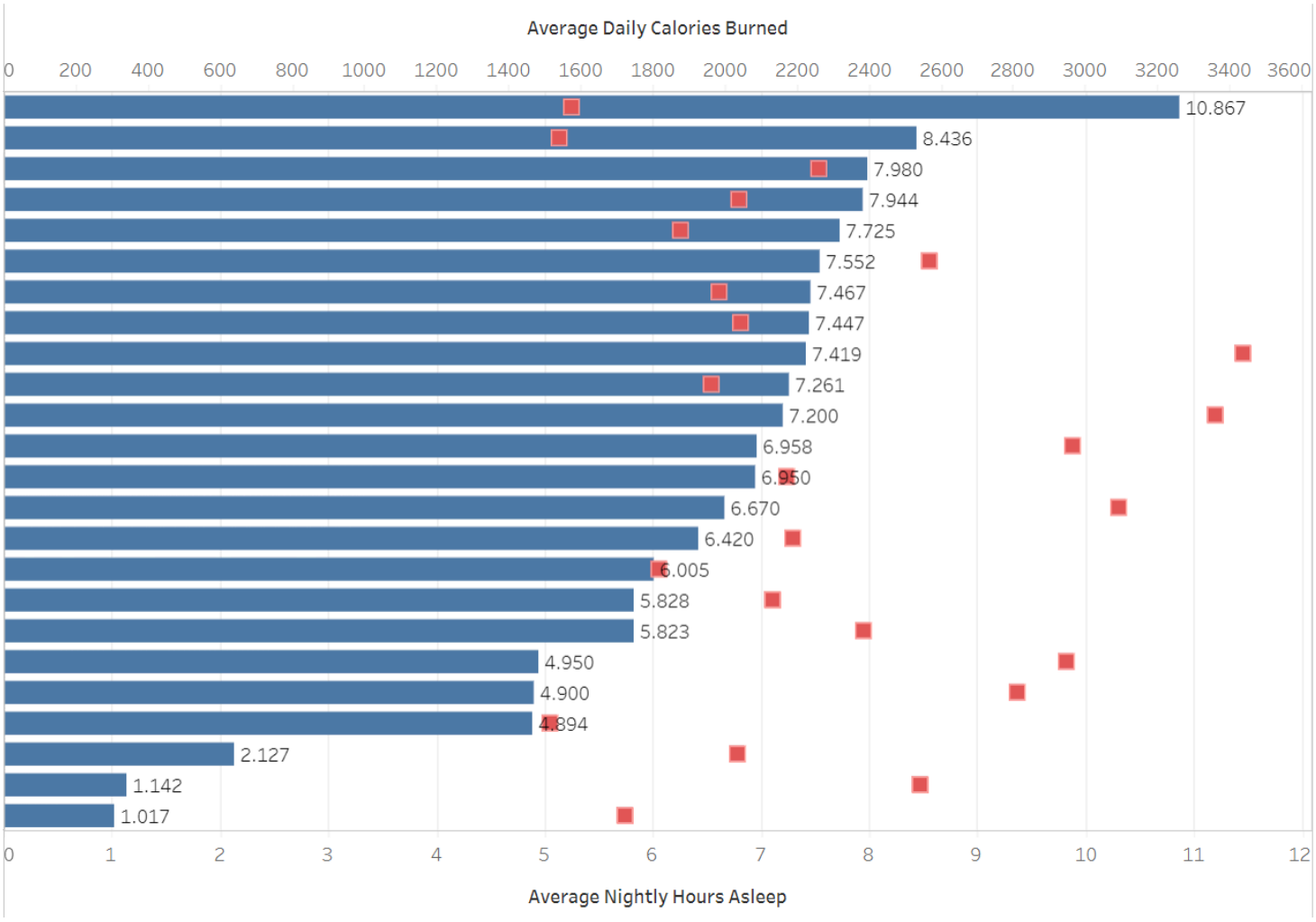
Sleep Function Usage Breakdown

Secondly, I wanted to consider the sleep trends of the participants in this study. When considering the habits of users of the products in the study in regard to sleep, the first main figure to consider here is the amount of data available. For activity records, there were a total of 940 days where data was reported across the 33 participants (28.5 records per participant). However, when looking at sleep records, there were only a total of 410 days where data was reported, or 12.4 records per participant. Even omitting the number of participants that did not report a sleep figure at all (9) -which I do not recommend as this figure is insightful in its own right- there are still only 17.1 records per user for sleep data. This clearly shows that regardless of the reason, users tracked their sleep data far less frequently than they tracked their daily activity data. When broken out into the functions that the device automatically tracks (meaning data types requiring extra input like weight are excluded), sleep is far and away the least utilized function amongst the data from the participants in this study.

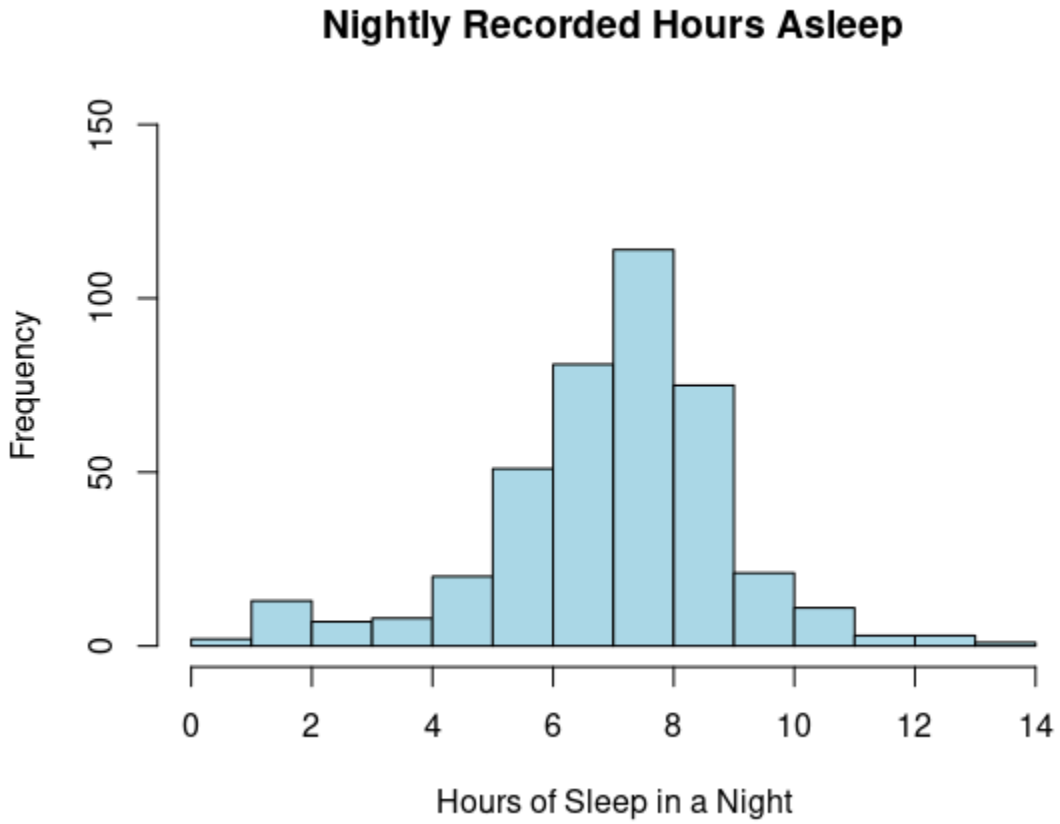


Average Sleep vs Calories Breakdown

When digging into the sleep data more, I was interested to find how the average amount of sleep users reported related to both their average daily calorie expenditure and their average number of steps per day. Generally, I predicted that more average time asleep would lead to more restful sleep and therefore a more active day with more steps taken and more calories burned. I was surprised to learn that this was not necessarily the case. As evident by the below chart, more sleep did not necessarily lead to more calories burned. In fact, the users that burned the 2nd and 3rd least amount of calories per day reported sleeping the most per night at 8.436 and 10.867 hours per night. What was more surprising about this to me was that the same trend did not hold true in the opposite way; as in, the users with the least sleep did not report the most amount of calories burned either, rather that distinction fell somewhere in the middle.



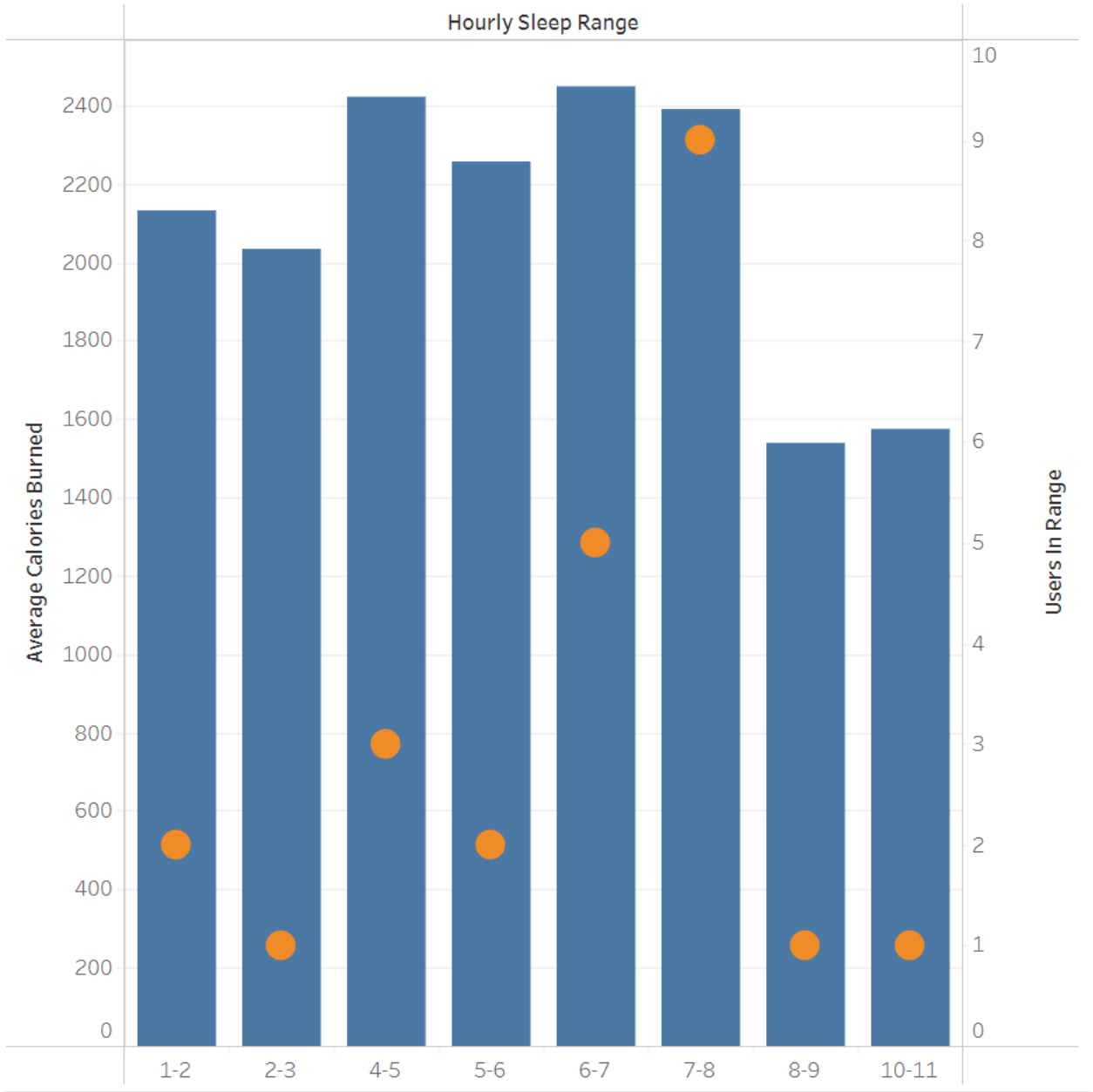
After conducting a histogram in RStudio, it is very apparent that the overwhelming majority of sleep records fall between 6-8 hours a night. Given what we know about sleep recommendations, this makes sense, but what is interesting is how these average sleep ranges impact the average amount of calories that that specific user burns.



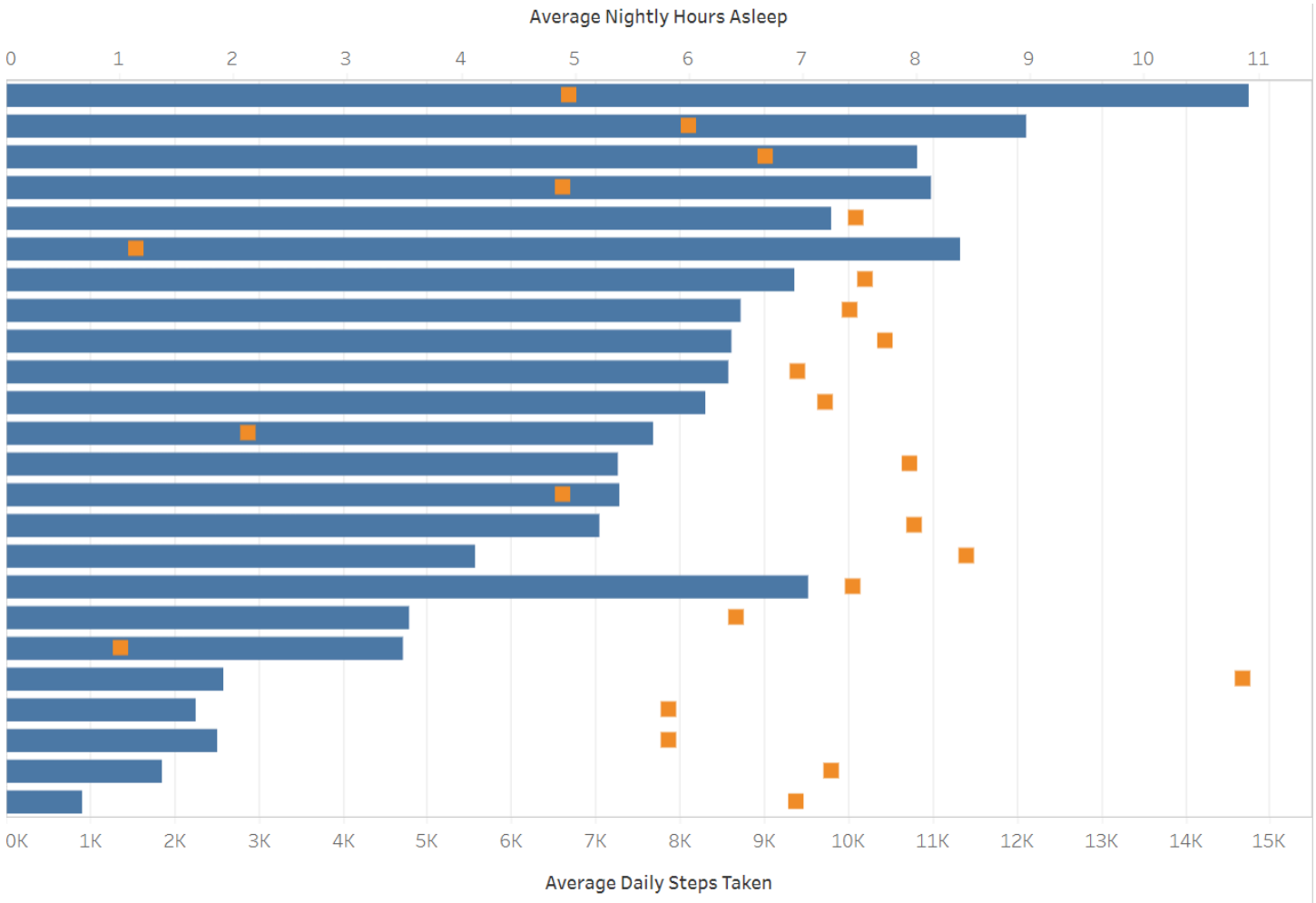

```
hist(sleep$HoursAsleep,
     main="Nightly Recorded Hours Asleep",
     xlab="Hours of Sleep in a Night",
     ylim=c(0,150),
     col="lightblue"
 )
```

In the below chart, I split users into groups based on how many average hours of sleep their device recorded per night. When these groups are plotted against the average daily calories burned by the same groups, a few things jump out. First off, my original hypothesis that users who sleep more burn more calories could not be more incorrect. The two biggest ranges reporting data (8-9 and 10-11 hours per night) far and away burned the least calories of any of the groups. Secondly, it is not the groups who sleep the least either as the two shortest ranges (1-2 and 2-3) burned the 3rd and 4th lowest calories. In fact, it

was the middle ranges representing 4-8 hours of sleep per night that had the most active days in terms of calories burned. One consideration to keep in mind when viewing the chart is the number of participants reporting. The orange balls represent the number of participants that fell into each nightly sleep range. With this, we are able to tell that the two lowest-calorie reporters were two individual participants. I would be very interested to see how this same chart would react to a much larger sample size but that is outside of the scope of this assignment at this time.



Unsurprisingly, very similar trends exist between sleep and steps as they did for sleep and calories burned. The below chart is the same as before except it plots steps rather than calories. Again it's clear to see that the middle ranges around 6-8 hours of nightly sleep led to the most activity, in this case, steps.



Case Study Roadmap - Share

To view my presentation on this case study, please visit this link

Case Study Roadmap - Act

After extensively looking into the data provided, there are a few high-level recommendations that I have for the Bellabeat executives and stakeholders.

- Create a base model of your popular Leaf tracker at a more affordable pricepoint for more casual consumers that scales back on functionality.
- Create two separate marketing strategies.
    - One geared towards the high-level fitness enthusiasts with the premium tracker.
    - One geared towards the cheaper, less feature-rich option for the more casual, everyday customer.
- Sleep is essential to physical and mental well-being, target those looking to improve this facet of their lives, but may not know the importance.
    - Bellabeat's sleep features are far better than competitors, make sure customers are aware of this.
    - Also make sure customers are aware of how much better, higher quality sleep can move the needle towards physical results.

Next Steps

There are a few different directions the direct stakeholders of this analysis can move in. First off, I think it is important to decide which business Bellabeat would like to be in. Providing the best possible product for those looking to level up their health no matter the cost? Becoming a dominant player in the smart device fitness industry that users of all skill levels will go to? Or, a mix of both with more product lines? Ultimately this is a massive decision for the company so it's imperative to figure out the direction to go in before anything else. Secondly, I would suggest that no matter what choice is selected, increase marketing and customer knowledge about the benefits of sleeping as they relate to both health and Bellabeat's products. If at a crossroads, an additional analysis could be done in a similar manner on either a larger sample size or of a subset of users who use Bellabeat products already. Either one of these courses would help decide more definitively the future of Bellabeat.