Google Data Analytics Capstone:

How Can a Wellness Technology Company Play It Smart?

Clarissa McCarthy
March 18, 2024

Table of Contents

Background Information

Bellabeat is a high-tech female wellness company that manufactures smart products with the goal of empowering women to take control of their health. They offer a range of products and are looking to analyze smart device data to grow them further. Bellabeat has requested analysis of trends in how smart device consumers use non-Bellabeat products. These trends will guide Bellabeat in their future marketing strategies.

Phase One: Ask

Guiding Questions

What is the problem I'm trying to solve?

Bellabeat wants to analyze trends in how smart device consumers use non-Bellabeat products to guide their marketing strategies.

How can my insights drive business decisions?

Recognizing trends in which uses of smart devices are most popular can guide Bellabeat to market and further develop those features. Likewise, recognizing which features are the least popular could enable Bellabeat to ask further questions about why these features are less popular (lack of knowledge of their existence, barriers of entry, unnecessary, etc.).

Looking at trends within specific features could also guide Bellabeat in what consumers may want to achieve with their products. For example, if consumers who use their smart devices to track their sleep patterns on average do not get enough sleep, Bellabeat could advertise how their products can help them achieve healthy sleeping habits.

Analyzing survey data on what consumers want to achieve with their products and what barriers may be hindering their access to these products or their success with them is also likely to provide useful insights into future marketing. Bellabeat products can implement and advertise methods to help users overcome the most common barriers. If such data contains demographics, such as gender or age, Bellabeat can more accurately address the wants and needs of more specific demographics.

Key Tasks

1. **Identify the business task.**

The business task is: Analyze trends in smart device usage to guide the marketing of one of Bellabeat's products.

2. **Consider key stakeholders.**

Urška Sršen – Bellabeat co-founder and Chief Creative Officer.
As the Chief Creative Officer, Sršen will likely be interested in the marketing strategies that arise from my analysis.
As a member of the executive team, she will likely be more interested in high-level thought processes and visual data rather than highly detailed and complex explanations.

Sando Mur – Bellabeat co-founder and Mathematician; key member of the Bellabeat executive team.
As a Mathematician, Mur will likely be interested in seeing how the data supports my analysis.
As a member of the executive team, Mur will likely be more interested in high-level thought processes and visual data rather than highly detailed and complex explanations.

Bellabeat customers
The customers, or a subset of them, will be the target audience of the marketing campaigns that arise from my analysis. It is important to eliminate any biases in my data, as this will directly affect the customers.

Deliverables

1. **A clear statement of the business task.**

The business task is: Analyze trends in smart device usage to guide the marketing of one of Bellabeat's products.

Phase Two: Prepare

Guiding Questions

FitBit Fitness Tracker Data by Mobius

1. **Where is your data stored?**

This data is stored on my personal laptop, in the R Studio database, and in Google's BigQuery database.

2. **How is the data organized? Is it in long or wide format?**

This data is a folder that contains CSV files that are mostly in long format. The rows are organized by ID number and the date the data was taken on. The columns include data about steps taken, distance travelled, calories burnt, sleep time, weight, and daily time/distance partaking in 4 levels of activity intensity. Some files contain the same collection of data, but presented in wide format.

3. **Are there issues with bias or credibility in this data? Does your data ROCCC?**

It is difficult to say for sure whether this data is biased, since it does not contain any demographic information about the 30 consumers contributing their data to this data set. Due to this absence of information, I cannot know for certain if certain groups and identities are underrepresented in this data set. However, the sample size of 30 consumers is small, meaning that it is likely that underrepresented communities are not represented in this data.

This data set is, however, credible. It is published to Zenodo, which displays the DOI badge for the data set. This DOI badge proves this data is original and not stolen from elsewhere. Zenodo also verifies the contributors to this data set and tracks its version history.

When analyzing this data against the ROCCC methodology, I can say that:

- It is **reliable**. The majority of data points are non-empty and fall within the expected values and data types of the data it represents.
- It is **original**. This data set has a DOI identifier, which validates its originality.
- It is **comprehensive**. The data set contains many csv files covering a wide range of health data. Within each csv file is an extensive amount of data, providing me with a large quantity of valuable data to analyze.
- It is **not current**. This data was collected and published in 2016, which is 8 years ago as of writing. Preferably, I would like to analyze more current data, such as data

collected within the past 4-5 years, so I can be more confident that the trends I analyze are still true for today's smart device consumers.
- It is **cited**. Kaggle cites Mobius as the user who uploaded the data set and Mobius acknowledges Robert Furberg, Julia Brinton, Michael Keatin, and Alexa Ortiz as contributors. They are also cited on the data set's Zenodo page as being the authors of the data set.

While the data is not current, it does meet the other criteria of the ROCCC methodology. Thus, I believe it is good data to work with for this analysis.

4. **How are you addressing licensing, privacy, security, and accessibility?**

This data is protected under the Creative Commons license CC0, which means it is under public domain and I may use it for my analysis without explicit permission from the authors. The data set does not contain any identifiable information regarding the 30 Fitbit users who submitted their data. In this way, their identifiable information has been kept private. I will keep this data secure by storing multiple copies of it in multiple locations (i.e., my personal computer drive, the R databases, etc.), thus any modifications to the data can be identified. I will also limit the accessibility of the data such that only I can modify it. However, I will link the original data in my work so anyone can view it.

5. **How did you verify the data's integrity?**

I verified the data's integrity by skimming through it to see how much data was missing or potentially erroneous (extremely high/low outliers, etc.). I also checked to see how many entries were duplicates. To do this, I uploaded the data set to Google Cloud's BigQuery website and used SQL to count the duplicate rows in each of the CSV files.

I found that the Minute Sleep file has 543 duplicate entries, and the Sleep Day file has 3. With another SQL query, I returned the duplicate rows for both files. The Minute Sleep file had a duplicate entry for each of the entries on May 6th and 7th for one user ID. The Sleep Day file had a duplicate entry for 3 different user IDs.

Compared to the total number of entries for each of these files, which I also gathered from an SQL query, these duplicates make up 0.6% and 1.5% of the entries in their files, respectively. Considering this low percentage of duplicates and the absence of duplicates in the other files in the data set, the data set has integrity and should produce trustworthy results.

6. **How does it help you answer your question?**

This data set will help me find trends in smart device usage by giving me a variety of uses to analyze. One use of this data will be to consider which FitBit features are the most/least

popular. This will help Bellabeat create features in their products that cater to the more popular uses. Another use would be to analyze which intensities of activity (Very Active, Moderately Active, Light Active, and Sedentary) are most common across users. This can help Bellabeat create features that are marketed towards these intensities. Furthermore, if Light Active or Sedentary activity are the most common, Bellabeat could create features that encourage users to increase the intensity of their activity.

7. **Are there any problems with the data?**

One glaring problem with this data is the lack of demographic data. Since Bellabeat is a female wellness company, it would be highly beneficial to analyze trends that pertain to women. However, there is no indication of sex or gender in the data provided, meaning it could even be solely collected from men. To address this problem, I will be using another data set that contains demographical data, such as gender, to analyze trends specifically regarding women.

As mentioned previously, some of these data files contain duplicate data. While the duplicates make up a small portion of the total data in their respective files, I will need to remove them such that only one of each of the entries is accounted for when I begin my analysis.

Despite the large amount of data in this set, the data only comes from 30 consumers. This small sample size may not accurately reflect the larger population of smart device users and is likely to exclude underrepresented communities. To account for this small sample size, I will also analyze data from other sources to increase the total amount of consumer data analyzed.

Another problem, or rather a missing consideration, is the lack of survey data from the consumers. Analyzing user feedback about their experiences with smart devices would certainly help Bellabeat know what features consumers like, dislike, or want to see more support for. Since this type of data is incredibly valuable, I will be using another data set of survey data for this purpose.

Fitness Consumer Survey Data by Harshita Aswani

1. **Where is your data stored?**

This data is stored on my personal laptop, in the R Studio database, and in Google's BigQuery database.

2. **How is the data organized? Is it in long or wide format?**

This data is a CSV file in wide format. The rows are organized by datetime, and the columns include demographic data (age, gender, education, occupation), frequency of activity/smart device usage, and how the device has impacted their life.

3. **Are there issues with bias or credibility in this data? Does your data ROCCC?**

Like the FitBit Fitness Tracker data, this data set contains responses from 30 smart device users. This is a small sample size, and thus is likely to not account for typically underrepresented communities. Therefore, the analyses gleaned from this data may not accurately represent smart device users at large.

Unlike the FitBit Fitness Tracker data, though, this data contains some demographic data about the respondents. With a quick SQL query, I learned that 15 respondents are female, 13 are male, and 2 prefer not to disclose their gender. This is a good ratio of males to females, though the data does not contain responses from other genders. Since I will be focusing on the trends for female respondents, this is not a major issue.

I ran similar queries for the age, education, and occupation of the respondents. Most respondents are between the ages of 18 and 24, and few respondents are 35 or older. There are a relatively even number of responses across the categories for highest education, though only 1 response from someone whose highest education is less than high school. The occupations of the recipients are also mostly even, with most respondents being students. However, there are only 2 responses each from people who are unemployed or retired. Through this quick analysis I can see that there is little data from people aged 35 and older, whose highest level of education is less than high school, and/or are unemployed or retired. This may reflect the demographics of smart device users at large, but it does also show biases in this data.

The data is relatively credible since the owner of the data set, Harshita Aswani, claims to have collected this data firsthand themselves. Since there is no DOI for this data, it is difficult to confirm whether this is true. However, a quick Google search does not turn up this data anywhere besides the Kaggle link where I got it. Therefore, I can be relatively confident that this is, in fact, Aswani's data.

When analyzing this data against the ROCCC methodology, I can say that:

- It is **reliable**. There is no missing data and all the data falls within the expected range for its respective column.
- It is **original**. As previously discussed, I could not find any copy of this data besides the original Kaggle link where I got it.

- It is **not comprehensive**. While there are only 30 responses in this data set, the questions asked cover a variety of smart device and health topics. The responses to each question came from a limited selection of choices, which may lose some of the nuance and specificity in how the respondent would have answered if they could type their responses themselves. However, the limited selection of responses makes it easier to categorize them and thus analyze trends in the feedback.
- It is **current**. This data set was posted to Kaggle on April 22, 2023, and the responses were from March 30, 2023 to April 7, 2023. This means this data is more likely to reflect current trends in smart device usage compared to older data.
- It is **cited**. Harshita Aswani uploaded this data set to Kaggle and collected the survey responses themselves. While there is no DOI confirming that Aswani is the true owner of the data, the difficulty of finding the data elsewhere on the internet indicates that they likely are the true owner.

While the data could be more comprehensive and include more respondents, it does meet the other criteria of the ROCCC methodology. Thus, I believe it is good data to work with for this analysis.

4. **How are you addressing licensing, privacy, security, and accessibility?**

This data is protected under the Creative Commons license CC0, which means it is under public domain and I may use it for my analysis without explicit permission from the author. Since this data contains demographic information about the respondents, I should be mindful of maintaining their privacy. If any of the respondents contact me, requesting that their data not be included in my analysis, I will comply and remove their data. I will keep this data secure by storing multiple copies of it in multiple locations (i.e., my personal computer drive, the R databases, etc.), thus any modifications to the data can be identified. I will also limit the accessibility of the data such that only I can modify it. However, I will link the original data in my work so anyone can view it.

5. **How did you verify the data's integrity?**

Since this is a small dataset, I first verified its integrity by manually checking for missing data using Excel. No missing data was found, which increases its integrity. Next, I checked for any duplicate entries using BigQuery. The query resulted in no duplicate entries, further increasing the data's integrity.

6. **How does it help you answer your question?**

This data set will help me find trends in smart device usage by analyzing the sentiments of users regarding how effective their smart device is in various aspects of their life. If, for

example, many users say their smart device does not improve their sleep patterns, Bellabeat might want to try to cater to this market by creating features in their devices that better improve users' sleep patterns. Similarly, if many respondents say their device doesn't motivate them to exercise or makes exercising enjoyable, Bellabeat can create new features that better encourage users to exercise.

7. **Are there any problems with the data?**

One issue with this data set is the sample size. 30 is a small sample of data, and it is even smaller (by half) when only analyzing the data from female respondents. To account for this shortcoming, I will use an additional survey dataset that contains more responses from female smart device users.

Another shortcoming of this data set is that users could only choose from pre-defined responses. As I mentioned earlier, this makes it easier to categorize and analyze this data, but it comes at the cost of missing out on more detailed responses, which may provide reasoning for the respondents' choices.

Fitness Analysis Data by Nithilaa

1. **Where is your data stored?**

This data is stored on my personal laptop, in the R Studio database, and in Google's BigQuery database.

2. **How is the data organized? Is it in long or wide format?**

This data is a CSV file in wide format. The rows are organized by datetime and the name of the respondent. The columns include demographic data (name, gender, age) and questions regarding how often respondents exercise, what motivates them to exercise/stay healthy, and what barriers they face in exercising more or maintaining a healthier lifestyle.

3. **Are there issues with bias or credibility in this data? Does your data ROCCC?**

This data set contains data from over 500 respondents, which is an excellent sample size for my analysis. The large sample size decreases the chance of minorities being underrepresented.

Examining the demographic data using BigQuery, I learned that 302 respondents are female and 243 are male. This is a good ratio of male to female and indicates no bias between genders polled. Most respondents are between the ages of 15 and 25 and few respondents (20) are between the ages of 26 and 30. Overall, however, there is a good

number of responses from all age groups, which I believe does not indicate significant biases regarding age.

When examining the names of the respondents, I noticed that most names are Indian. This may indicate a bias in the location of the respondents polled. As a result, the trends analyzed in this data may not be true for people from other locations. This bias is further evident by examining the description of the data set on Kaggle, where the owner says that most of the data is collected from their friends and family. Since we are likely to befriend people who share demographics, personalities, and communities as ourselves, this may indicate that the data collected underrepresents certain communities.

The credibility of this data is also less than ideal. Kaggle informs me that the user Nithilaa owns this data set but it's unclear whether this is their real name or simply a username. Their credibility would increase if their username included their first and last name. The data also has no DOI, so there is a possibility that this data does not truly belong to Nithilaa. After a quick Google search, I could not find the data elsewhere online, which adds credibility to Nithilaa's ownership of the data.

When analyzing this data against the ROCCC methodology, I can say that:

- It is **reliable**. There is no missing data and all the data falls within the expected range for its respective column.
- It is **original**. As previously discussed, I could not find any copy of this data besides the original Kaggle link where I got it.
- It is **comprehensive**. The sample size of this data is quite large, at over 500 responses. There is also a good variety of questions asked in the survey. These questions include the respondent's motivations for exercising, barriers they encounter, and whether they've purchased exercise equipment. The responses to each question came from a limited selection of choices, which may lose some of the nuance and specificity in how the respondent would have answered if they could type their responses themselves. However, the limited selection of responses makes it easier to categorize them and thus analyze trends in the feedback.
- It is **current**. The responses are from July 3, 2019 to July 21, 2019, which is a little less than 5 years ago as of writing. I consider data current when it falls within at most 5 years of being collected. This data falls just within this timeframe, so I consider it current. However, I would prefer to work with data that is *more* current, especially considering this data was collected right before the COVID-19 pandemic. As a result of the pandemic, people's exercise habits may have changed, and such a change would not be reflected in this data set.

- It is **not cited**. While I can identify Nithilaa as the owner of this data on Kaggle, this is likely not their full legal name. Furthermore, the data has no DOI to identify who the authors of the data truly are. Therefore, it is difficult to know for certain that Nithilaa is the true author of the data or to verify that they are a credible source.

While the data is not properly cited, it does meet the other criteria of the ROCCC methodology. Thus, I believe it is good data to work with for this analysis.

4. **How are you addressing licensing, privacy, security, and accessibility?**

According to Kaggle, the licensing of this data is unknown. Considering Nithilaa uploaded this data set to Kaggle, which is known for providing open data, and Nithilaa did not include any restrictions for working with this data, I will assume it is open for anyone to work with. If, after uploading my analysis with this data, Nithilaa contacts me requesting that I do not include their data in my analysis, I will comply and remove it from my analysis.

Since this data contains demographic information about the respondents, especially the names of the respondents, I should be mindful of maintaining the respondents' privacy. If any of the respondents contact me, requesting that their data not be included in my analysis, I will comply and remove their data.

I will keep this data secure by storing multiple copies of it in multiple locations (i.e., my personal computer drive, the R databases, etc.), thus any modifications to the data can be identified. I will also limit the accessibility of the data such that only I can modify it. However, I will link the original data in my work so anyone can view it.

5. **How did you verify the data's integrity?**

To verify the data's integrity, I first used BigQuery to check for any missing data. After finding no missing data, I then queried for any duplicate entries, of which none were found. Since this data set contains the names of the respondents, I then decided to check if there were any duplicate names, which might indicate the same respondent submitted the form multiple times with different answers. This query revealed that 23 names appeared in more than one entry.

To determine whether the duplicate entries could possibly be from the same person, I viewed the entries with duplicate names. Examining these entries, I found that many of the entries that shared a name submitted their responses on the same day, selected the same gender and age range, and had similar responses to the questions. Due to these similarities, I concluded that those entries were likely from the same person and wanted to update their responses. This would be considered duplicate data, which I must remove when cleaning the data. Some entries, however, had the same name but very different

responses. I concluded that these entries were from different people who happened to share the same name. These entries would not be considered duplicate data and therefore will not be removed during cleaning.

While I concluded that the data does contain some duplicate entries, these entries represent a small subset of the data. Therefore, since the majority of the data does not contain duplicates, this data set has integrity.

6. **How does it help you answer your question?**

This data will help me find trends in smart device usage by analyzing the exercise habits and sentiments regarding health and exercise of the respondents. While the data does not specify whether the respondents use exercise smart devices, I believe analyzing the responses to barriers to exercise and what motivates respondents to exercise can help Bellabeat create features that better encourage their users to exercise more. Analyzing which forms of exercise are most common for female respondents would also be beneficial for Bellabeat since they could use this knowledge to develop features that cater to these types of exercise, which may increase their user base.

7. **Are there any problems with the data?**

The main problem with this data is that it is likely biased to respondents located in India, since most names in the data set are Indian. To account for this issue, I will be analyzing the trends in another survey-based data set that may contain responses from other countries. However, since this data set is significantly larger than the other survey data I will analyze, my analyses will likely still be biased towards Indian respondents.

This data also has some issues with credibility. The full name of the owner of the data is not given and there is no DOI to verify that this data is in fact theirs. For the purposes of this analysis, I will assume that Nithilaa is the owner of the data and is a trustworthy source. If the results from my analysis of this data is drastically different from the results from analyzing the other data sets, more research will need to be done to verify Nithilaa's credibility, and another data source will likely need to be used.

Another issue is that the data is not as current as I would like it to be. As I mentioned earlier, this data was collected prior to the COVID-19 pandemic. Therefore, it is unlikely to contain current trends that resulted from the pandemic. For my analysis, I will assume that the trends in this data are still applicable today and there are no major trends today that are not represented in this data. However, more research should be conducted to verify this assumption.

Key Tasks

1. **Download data and store it appropriately.**

The three data sets I will work with during my analysis are stored:

- On my personal laptop
- In R Studio's database
- In BigQuery's database

2. **Identify how it's organized.**

These data sets are CSV files. The FitBit Fitness Tracker data is a collection of CSV files, which contain both long and wide data. The Fitness Consumer Survey data and the Fitness Analysis data are both wide data.

3. **Sort and filter data.**

I used BigQuery to filter the data for missing and duplicate data. Only the FitBit Fitness Tracker data contains some null values. However, the number of null values is extremely small compared to the amount of non-null data. The FitBit Fitness Tracker data and the Fitness Analysis data contain duplicate data. Again, these duplicate data comprise only a small portion of the total data.

4. **Determine the credibility of the data.**

The FitBit Fitness Tracker data is credible. There is a DOI badge for this data, which validates its originality and the authors of the data. The main challenges to the data's credibility are the lack of current data and the small sample size of smart device users.

The Fitness Consumer Survey data is also credible. While there is no DOI badge for it, it is uploaded under what is presumably the owner's full legal name. I also did not find this data elsewhere on the internet, which increases the likelihood that this data does belong to the uploader. The main challenges to the data's credibility are the potential biases towards certain age groups and the small sample size.

The Fitness Analysis data is credible enough for the purposes of my analysis. While the data has issues with citation and potential location biases, I believe these are not significant enough issues for my analysis. However, if my findings from analyzing this data are drastically different from my findings from the other data sets I will analyze, then I must reevaluate the credibility of this data and likely include another data source in my final analysis.

Deliverables

1. **A description of all data sources used.**

The FitBit Fitness Tracker data contains 18 CSV files with numerical data from 30 FitBit users' devices. This data includes daily, hourly, and minutely records of the users' calories burnt, exercise intensity, and step count. It also includes data about the users' heart rate, weight, and sleep patterns. This data will be useful to analyze what FitBit features are most used and what types of exercise intensity are most common.

The Fitness Consumer Survey data contains survey responses from 30 fitness wearable users. It includes demographic data about the respondents, such as gender, age, and education, and answers to various questions regarding the impact of their fitness wearable on their lifestyle. This data will be useful to analyze what aspects of one's lifestyle fitness wearables are most likely to improve, and what aspects they should be better designed to improve.

The Fitness Analysis data contains survey responses from over 500 people. It includes demographic data about the respondents, such as name, gender, and age, and answers to questions regarding how healthy the respondent's lifestyle is and what barriers they face when trying to be healthier. This data will be useful to analyze trends in barriers to exercising and eating healthy and what motivates people to maintain healthy habits.

Phase Three: Process

Guiding Questions

1. **What tools are you choosing and why?**

I will use SQL and R to analyze my data. Since many of the files I will be working with contain hundreds of rows of data, it would be most beneficial for me to use tools that are designed to handle data of this magnitude. I used BigQuery to perform initial SQL queries on the data to check it for missing values and duplicate entries. This helps me understand some of the tasks I must complete when cleaning my data. The actual cleaning and analysis of my data will be done in R Studio. With R Studio, I can easily record my cleaning process, which allows others to see how I've transformed the data prior to my analysis. R Studio also lets me easily rerun code I've written in case I need to reexamine or remind

myself of trends I noticed earlier in my analysis. The data visualization tools in R also let me quickly notice trends in my data, which gives me a good way to begin my analysis.

2. **Have you ensured your data's integrity?**

Yes, all three of my data sources have integrity. The steps I took to ensure this are outlined in Phase 2 under each respective data set.

3. **What steps have you taken to ensure that your data is clean?**

First, I ran SQL queries in BigQuery to identify any missing or duplicate data. The specific steps I took for each data set are outlined in Phase 2. These queries highlighted what data I needed to remove during cleaning.

Next, I uploaded my data to R Studio to begin cleaning it. Since the FitBit Fitness Tracker data contains many files, I chose to only upload the files that I believed would be most useful for my analysis. Ultimately, I uploaded the files containing daily data about activity, calories, intensity, steps, and sleep, as well as the heartrate and weight data.

Then, I identified any missing data and decided how to handle it. From my initial SQL queries in Phase 2, I already knew that the only files with missing data were from the FitBit Fitness Tracker data set. I verified this in R by summing the number of null values in each of the files I uploaded. Through this, I discovered that only the daily exercise file had null values. I then summed the number of null values in each of the columns in the exercise file and learned that all the null values were in the "Fat" column. Since this column is not vital to my analysis, I decided to remove it.

Despite already knowing that the Fitness Consumer Survey and Fitness Analysis data sets did not contain any missing values, I checked this again in R for completeness in the documentation of my cleaning process.

After handling the null values, I handled the duplicate data. From my initial SQL queries in Phase 2, I already knew which files contained duplicate data. The FitBit Fitness Tracker data set contained two files with duplicates, of which I am only using one in my analysis – the Sleep Day data. In R, I retrieved the duplicated rows to ensure that they were duplicates and to make sure my findings here aligned with those in my SQL queries. Then, I removed the duplicates using the dplyr library's `distinct()` function and made sure that no duplicate entries remained.

I then moved on to cleaning the duplicates from the Fitness Analysis data set. From my SQL queries, I knew that there were no duplicate rows, but there *were* duplicate names. So, I first created a data frame containing all rows with a duplicate name. From here, I manually compared each row with its duplicate entry to determine whether the responses were likely

from the same or different people. If the duplicate entries had the same gender and age and had similar responses to most questions, I deduced they were from the same person and removed the oldest entries. If these columns did not match, I deduced they were from different people and kept both entries in the data set. To confirm that the correct number of rows were removed, I compared the difference of the number of rows in the original data set and the number of rows removed from the duplicate data to the number of rows in the cleaned data set. Since these numbers matched, the duplicate data was successfully removed.

Aftering cleaning the duplicate data, I removed the columns of data that I felt were unnecessary for my analysis. In the FitBit Fitness Tracker data, the columns I chose mostly consisted of all 0's or all 1's, meaning there were likely no interesting trends or correlations with those variables. In the Fitness Analysis data, I removed columns that contained information about diet since I wanted to focus on exercise in my analysis.

Then, I renamed the columns in the Fitness Analysis data set to make them easier to read.

Finally, I checked the minimum and maximum values of the numerical data to catch any outliers. To do so, I used the minimum and maximum functions in R. I also sorted each column from lowest to highest and highest to lowest to compare the minimum and maximum values with the next lowest/highest values. The only outlier I found was in the weight data from the FitBit Fitness Tracker data, but the weight and BMI were still within a reasonable range, so I kept this data.

4. **How can you verify that your data is clean and ready to analyze?**

Each time I removed missing or duplicated data, I verified that the data contained the expected rows. When I removed the "Fat" column from the FitBit Fitness Tracker weight data, I viewed the resulting data frame to confirm that that column was the only one removed. After removing the duplicate data from the FitBit Fitness Tracker sleep data, I counted the number of duplicate rows to ensure that no duplicate data remained. Finally, after removing the duplicate responses from the Fitness Analysis data, I compared the number of rows in the original data frame and the number of rows of duplicate responses to the number of rows in the clean data frame. Similarly, when I removed unnecessary columns of data, I checked the resulting data frame to ensure that only those columns were removed. After ensuring that no missing or duplicate data remained and the expected data remained in its data frames, I knew my data was clean and ready to analyze.

5. **Have you documented your cleaning process so you can review and share those results?**

Yes. The full documentation of my cleaning process can be found in the file cleaning_doc, which is available as an HTML, PDF, and DOCX file.

Key Tasks

1. **Check the data for errors.**

I checked my data for missing, duplicate, and outlier data. Missing values were found in the weight data, duplicate data was found in the sleep and fitness analysis survey data, and an outlier was found in the weight data.

2. **Choose your tools.**

I used R Studio to clean my data, as well as SQL for my initial queries on the data when evaluating their integrity.

3. **Transform the data so you can work with it effectively.**

The transformations I made include:

- Removing missing and duplicate data
- Removing unnecessary columns
- Renaming columns

4. **Document the cleaning process.**

Documentation of my cleaning process can be found in any of the cleaning_doc files. A complete collection of the code used to clean my data can be found in the cleaning_data.R file.

Deliverables

1. **Documentation of any cleaning or manipulation of data.**

Documentation of my cleaning process can be found in any of the cleaning_doc files. A complete collection of the code used to clean my data can be found in the cleaning_data.R file.

Phase Four: Analyze

Guiding Questions

1. **How should you organize your data to perform analysis on it?**

I organized my data in wide format for most of my analyses but converted some data frames to long format when graphing. I also already removed columns I knew I wouldn't use to make it easier to find and work with the data I needed. Similarly, I also renamed most of the columns I knew I would work with. When I made new columns, I renamed those as well.

2. **Has your data been properly formatted?**

Yes, my data was formatted correctly, and new columns and data frames followed the correct format of the other data. When data was not in the correct format, I converted it to the correct format.

3. **What surprises did you discover in the data?**

I was surprised that most respondents in the fitness device survey data owned their device for less than 6 months. This caused me to also account for this in my analysis. I paid close attention to trends in users who've owned their device for over 6 months so I could find what may lead someone to use their device more.

I was also very surprised that the only people who said they strongly disagreed that their device helped them stay motivated were people who owned their device for 6 to 12 months. I would have assumed that the longer someone has the device, the more effective it is at keeping them motivated.

4. **What trends or relationships did you find in the data?**

I found that the most popular reasons why people don't exercise are because they can't find the time and/or motivation for it. When examining the fitness wearable survey data, most respondents said that their fitness wearable improves their motivation to exercise. Therefore, there's a correlation between owning a fitness wearable and being more motivated to exercise.

Additionally, the FitBit data revealed that, on average, users get less than the recommended amount of sleep. However, according to the fitness wearable survey data, 80% of fitness device users say their wearable improved their sleep patterns. Since less FitBit users use the sleep feature than some of the other features, this seeming

contradiction might be due to ineffective advertising for the feature or problems with the feature itself.

Another fascinating trend I recognized was that, despite weight loss and body image being the most popular motivations for exercising, very few FitBit users use the weight feature on their device. Like the sleep feature, I assume this is either due to ineffective advertising of this feature or issues with the feature itself.

I also found that most respondents from both the exercise and fitness wearable surveys were between the ages of 18 and 25. While this could indicate a bias in the people surveyed, it might also indicate the target demographic for Bellabeat to advertise their fitness devices to.

Regarding the frequency of fitness device usage, I found that, on average, the longer someone has owned their device, the more often they use it.

5. **How will these insights help answer your business questions?**

These insights highlight two key features Bellabeat should invest resources in for their products: sleep and weight tracking.

Many fitness device users report their device improving their sleep patterns, yet FitBit users, on average, still don't get the recommended amount of sleep. This presents an opportunity for Bellabeat to do further research into how fitness wearables can help users improve their sleep. If done successfully, Bellabeat may be able to capture that demographic of fitness users who want a device to help them sleep better.

As previously mentioned, few FitBit users use the weight feature despite weight loss and body image being the leading motivators for exercise. Bellabeat can take advantage of this by implementing a weight tracking feature in their devices.

To address the top barriers to exercise, lack of time and motivation, Bellabeat could implement schedule and goal tracking features to their devices. A schedule feature would allow users to set up reminders that help them work exercise into their schedule. A goal tracking feature would help users stay motivated to continue exercising. When setting goals, the user could initially be presented with a list of the most common motivators for exercising, which were revealed during this analysis. The goal tracking feature could also be paired with the schedule feature to use their motivators to encourage them to exercise during the days and times they initially scheduled.

Key Tasks

1. **Aggregate your data so it's useful and accessible.**

My data has been separated into cleaned data frames that contain only the information most important to my analysis.

2. **Organize and format your data.**

My data has been organized primarily in wide format since this is the easiest format for most of my work to use. For the graphing functions that work best with long data, I transposed the data frame prior to using these functions. These two different formats of the same data were saved as separate data frames, so neither overwrote the other.

3. **Perform calculations.**

Most of my calculations involved calculating the percentage make-up of the responses for the two surveys. I also calculated the average and standard deviation for the sleep data to reveal the average number of hours users slept. When analyzing the responses for exercise motivation and sleep improvement, I separated each category into a separate column and indicated whether each user included that category in their response. Then, I summed the number of responses for each category.

4. **Identify trends and relationships.**

As mentioned earlier, two main trends I discovered were that (1) many fitness device users say their device improves their sleep, yet FitBit users, on average, still don't get enough sleep, and (2) two key motivators to exercise are weight loss and body image, yet very few FitBit users take advantage of the weight feature on their device. These two trends highlight features Bellabeat can implement and fine-tune in their devices to cater to these markets.

Deliverables

1. **A summary of your analysis.**

A full summary of my analysis can be found in the file "Bellabeat-Analysis.pdf".

Phase Five: Share

Guiding Questions

1. **Were you able to answer the business questions?**

Yes, by analyzing popular and unpopular FitBit features, as well as user sentiments towards exercising and their devices, I was able to suggest features for Bellabeat to implement and market.

2. **What story does your data tell?**

My data shows the journey of discovering which FitBit features are popular and unpopular, analyzing how Bellabeat may be able to improve the unpopular features, and pinpointing which fitness device features Bellabeat should implement to meet consumer needs.

3. **How do your findings relate to your original question?**

I found that FitBit's weight feature is its least popular feature, yet the leading reason why people exercise is to lose weight. This highlights the potential for Bellabeat to develop a weight feature that serves consumers better than other leading fitness tech companies. I also saw that, depending on the quality of sleep Bellabeat users get, Bellabeat may be able to market their sleep feature as being better than FitBit's. Finally, I found that there's potential demand for scheduling and motivational reminder features. These would help users overcome the two most popular barriers to exercising: time and motivation. Helping users overcome these barriers would increase sales and customer satisfaction.

4. **Who is your audience? What is the best way to communicate with them?**

My audience is Bellabeat's cofounders and Bellabeat's marketing analytics team. The best way to communicate with them would be through high-level summaries of my findings with easily digestible visualizations. I should also focus on how my recommendations will improve Bellabeat's marketing and increase sales.

5. **Can data visualization help you share your findings?**

Yes, data visualizations will be crucial in communicating my findings to my audience. I will use bar charts to visualize how many users responded each way in survey questions. I will also create bar charts that show any potential differences between how different demographics, such as male vs. female respondents, answered each question.

6. **Is your presentation accessible to your audience?**

Yes. I chose colorblind accessible colors in my graphics. I also added alternate text to the images on my analysis documentation. This makes my graphs accessible for those with vision impairments and screen-readers.

Key Tasks

1. **Determine the best way to share your findings.**

The best way to share my findings is through a PowerPoint presentation with high-level summaries of my findings. The slides would include key takeaways, digestible visualizations, and actionable steps.

2. **Create effective data visualizations.**

My data visualizations include bar charts showing the distribution of responses for key survey questions. I made sure to use colorblind accessible color palettes and add alternate text for screen readers.

3. **Present your findings.**

I made a PowerPoint presentation and did a mock presentation of it.

4. **Ensure your work is accessible.**

I made my work accessible through colorblind-friendly color palettes and screen-reader friendly captions.

Deliverables

1. **Supporting visualizations and key findings.**

My visualizations done in R can be found in the folder “RCode/Plots”. My visualizations done in Tableau can be found in the folder “Tableau”. My final presentation can be found in the folder “Presentation”.

Phase Six: Act

Guiding Questions

1. **What is your final conclusion based on your analysis?**

My final conclusion is that Bellabeat should further analyze (1) why FitBit's weight feature is unpopular and what consumers want in this type of feature, and (2) how Bellabeat's sleep feature compares to that of FitBit's. I also suggest that Bellabeat implement features for scheduling workouts and setting up personalized reminders. These would help users find time for exercising and stay motivated to keep exercising.

2. **How could your team and business apply your insights?**

Bellabeat could apply my insights by starting additional data analyst projects to answer the two questions above. They could also apply my insights by implementing features for users to schedule their workouts and set personalized reminders to make exercising a habit.

3. **What next steps would you or your stakeholders take based on your findings?**

Bellabeat should start additional data analyst projects to answer the two questions above. They should also implement features for users to schedule their workouts and set personalized reminders to make exercising a habit.

4. **Is there additional data you could use to expand on your findings?**

For the additional research I suggested, I would need data on how effective Bellabeat's device features and user opinions on them. Specifically, I would need metrics on how Bellabeat's sleep feature compares to FitBit's sleep feature. I would also need survey data about FitBit users' opinions on FitBit's weight feature and survey data about what smart fitness device users want in such a feature.

Key Tasks

1. **Create your portfolio.**

I have portfolios on:

- My website
- GitHub
- Kaggle
- Tableau

2. **Add your case study.**

This case study has been added to each of my portfolio locations.

3. **Practice presenting your case study to a friend or family member.**

Deliverables

1. **Your top high-level insights based on your analysis.**

My final insights for Bellabeat are to:

- Analyze why FitBit's weight feature is unpopular and what users want in such a feature.
- Analyze how Bellabeat's sleep feature compares to that of FitBit's.
    - If it's better than FitBit's, highlight that in marketing!
    - If it's not better, research how it can be improved.
- Implement scheduling and personalized reminders to help users overcome the barriers of time and motivation.