Introduction

I am junior data analyst working on the marketing analyst team at Bellabeat. Bellabeat is a high-tech manufacturer that focuses on health products for women. The goal of this project is to “focus on one of Bellabeat’s products” utilizing smart device data to gain insight into how smart device users use their devices. Bellabeat would like to use this analysis to determine new growth opportunities for the company.

Stakeholders and Products

* **Stakeholders**

**○ Urška Sršen:** cofounder and Chief Creative Officer

○ **Sando Mur:** Mathematician and Bellabeat’s cofounder

**○ Bellabeat marketing analytics team:** Bellabeats data analyst team – responsible for providing insights to grow the Bellabeats company.

* **Products**

**○ Bellabeat app:** An app that provides users with health data related to their activity, sleep, stress, menstrual cycle, and mindfulness habits.

**○ Leaf:** wellness tracker that can be worn as a bracelet, necklace, or clip. The Leaf tracker connects to the Bellabeat app to track activity, sleep, and stress.

**○ Time:** Watch worn fitness tracker.

**○Spring:** Water bottle that tracks hydration levels.

**○ Bellabeat membership:** subscription membership program for users, that provides guidance on nutrition, activity, sleep, health and beauty, and mindfulness based on their lifestyle and goals.

Ask

For my analysis, I will be focusing on the Bellabeats Leaf product.

I will be looking at daily intensities, calories burned, and steps taken to provide a recommendation to the cofounders and marketing analytics team.

Prepare

Source and Credibility

The data set utilized in my analysis was provided by the cofounder and Chief Creative Officer, Sršen. This public dataset: [FitBit Fitness Tracker Data](https://www.kaggle.com/datasets/arashnic/fitbit) (CC0: Public Domain, dataset made available through [Mobius](https://www.kaggle.com/arashnic))provides insight into a small group of 33 individuals smart device users’ daily habits over a one-month period between April 12, 2016 and May 12, 2016.

As mentioned previously, this dataset is part of a public domain. This dataset has also been downloaded for research hundreds of times and is expected to be updated annually based on the domain. This leads me to believe this dataset is credible and does not present inherent bias.

The dataset was downloaded by me and stored on my private drive. Any updates made to or processing of the data has been saved in an alternate file so that the original data formats are not compromised.

I also verified the integrity of this data by reviewing the IDs provided across all datasets. These IDs are consistent across all of the provided files, and also do not appear to have any personal identifying information that could tie any participant to the provided IDs.

Description of Data

The dataset provides FitBit data for 33 individuals. This data included, activity levels, sleep data, calorie data, and step tracking for each individual.

Limitations of the Data

The largest issue with this data set is the age. Seven years, and the occurrence of a global pandemic is certainly enough, in my opinion, to warrant updated smart device user testing and data collection. The habits of these users could be different than the habits captured seven years ago.

While reviewing files in the provided data set, I found that some data sets were incomplete – some data sets did not include tracking data for all 33 participants.

This data was also collected in 2016, over 7 years ago. This may be a limitation because people may not track the same types of data exclusively and may be focusing more on other types of tracking data than data that was measured and tracked in this study. This could be due to new emerging tech or any other host of reasons – the Covid-19 pandemic for example.

Intensity levels were also not explicitly defined. So, for the purpose of this analysis, I understand sedentary to be the body at rest or not performing physical activity. Lightly active, fairly active, and very active intensities, I would assume, pertain to a person’s heart rate during activity and can be described as slightly elevated, moderately elevated, and the highest elevation when compared to the heart rate at sedentary levels. Respiration rate data could have also been beneficial to this category.

Workbooks Provided in Fitabase Dataset:

All datasets were provided in PDF format and included various components of data.

|  |  |
| --- | --- |
| dailyActivity\_merged.csv | dailyCalories\_merged.csv |
| dailyIntensities\_merged.csv | dailySteps\_merged.csv |
| heartrate\_seconds\_merged.csv | hourlyCalories\_merged.csv |
| hourlyIntensities\_merged.csv | hourlySteps\_merged.csv |
| minuteCaloriesNarrow\_merged.csv | minuteCaloriesWide\_merged.csv |
| minuteIntensitiesNarrow\_merged.csv | minuteIntensitiesWide\_merged.csv |
| minuteMETsNarrow\_merged.csv | minuteSleep\_merged.csv |
| minuteStepsNarrow\_merged.csv | minuteStepsWide\_merged.csv |
| sleepDay\_merged.csv | weightLogInfo\_merged.csv |

Workbooks Used in Analysis:

dailyActivity : This dataset was used to determine the total number of participants to ensure all other datasets analyzed would include full sets of data for participants.

dailyIntensities : This dataset depicts the sedentary, lightly active, fairly active, and very active intensities in minutes by day for each of the participants.

dailySteps : This dataset shows daily steps for each participant across the month long research period.

dailyCalories : This dataset shows daily calories for each participant across the month long research period.

Workbooks Excluded from Analysis:

Many workbooks were excluded from the analysis process for a myriad of reasons. Workbooks not utilized were excluded because they either did not contain complete data sets (i.e. data was not provided for all 33 participants), or they did not pertain to the question the founders requested analysis on. Three examples are listed below:

* weightLogInfo was excluded because there was only data or 8 of 33 participants. I do not believe any insights provided for this data set would be accurate or meaningful, since it does not provide data on all participants. Using this data set could have led to biased data and could have called into question the validity of the analysis derived from this set of data.
* minuteIntensitiesNarrow was excluded because it showed data by the second for intensity of data for each participant. Based on what was asked this data would be better shown across other datasets that show total active and sedentary minutes for each participant. This would be a fuller type of dataset to use than viewing activity intensities by seconds.
* sleepDay was almost used to provide analysis but did not provide a full picture in my opinion. This data would have been valuable if it included the sleep patterns for all 33 participants. Unfortunately, data was only provided for 24 of the 33 participants (72%). I believe using this dataset would lead to extrapolation and may lead to biased data. For this reason, this dataset was also excluded.

Process

During this phase of the Analysis process, I opened and scanned through each of the provided workbook csv files. This is how I determined which files would be key in providing insights as requested by the stakeholders, and how I determined which workbooks contained full data for all participants so that my analysis would be as thorough as possible.

During the cleaning phase, I saved all of my changes into new workbooks, with the same naming conventions as the downloaded workbooks, added “\_Cleaned” at the end of each workbook.

List of Cleaned Workbooks:

|  |
| --- |
| dailySteps\_merged\_Cleaned.csv |
| dailyCalories\_merged\_Cleaned.csv |
| dailyIntensities\_merged\_Cleaned.csv |

Cleaning Process of Each Dataset:

1. Cleaning of dailyIntensities\_merged.csv
   1. Used the UNIQUE formula to double check the total number of participants was equivalent to 33. Copied and pasted the data as values. Column is called “UniqueIDCheck”
   2. Used the SUM formula to determine the total intensity minutes per day by combing the sedentary, lightly, fairly, and very active minutes. This column is called “TotalMinuteIntensities”.
2. Cleaning of dailyCalories\_merged.csv
   1. Used the UNIQUE formula to double check the total number of participants was equivalent to 33. Copied and pasted the data as values. Column is called “UniqueIDCheckCalories”.
   2. Removed 4 data points of “0” Calories. A value of zero would indicate the fitness tracker was not worn during that particular day. So, this data was excluded.
   3. Created “AverageCalories” column to determine the average number of calories burned per day by each participant. Utilized the AVERAGE formula. Copied the output and pasted as values.
   4. Created “MedianCalories” column to determine the median amount of calories burned per day by each participant. I thought this would better depict the actual amount of calories burned, taking into account low or high outliers. Copied and pasted the output as values. I also dragged this value all the way down for each data point.
   5. Created “CaloriesEqualOrAboveAverage” to determine which data points were equal to or above the average value. The output produced either “True” or “False”. I multiplied these values by 1 to produce either a “1” or “0”. “1” equals “True”, and “0” equals “False”. These values were then copied and pasted as values. Example: =(C2>=E2)\*1.
   6. Created “CaloriesEqualOrAboveMedian” to determine which data points were equal to or above the average value. The output produced either “True” or “False”. I multiplied these values by 1 to produce either a “1” or “0”. “1” equals “True”, and “0” equals “False”. These values were then copied and pasted as values. Example: =(C2>=F2)\*1.
3. Cleaning of dailySteps\_merged. Csv
   1. Used the UNIQUE formula to double check the total number of participants was equivalent to 33. Copied and pasted the data as values. Column is called “UniqueIDCheckSteps”.
   2. Determined the first and third quartiles, IQR, and lower and upper limits to find any outliers in daily step count. First and third quartiles were determined by by the QUARTILE.INC formula. IQR was determined by taking the difference between the third and first quartile. Lower and upper limits were calculated by taking the first quartile for the lower limit and third quartile for the upper limit – 1.5 \* IQR.
   3. Utilized the OR function to determine outliers. The formula looked like this =OR(C2<L$6,C2>L$7). I also multiplied this value by 1 to provide results in numerical format. Outliers were indicated as “0”. There was a total of 95 outliers found – the majority of which were indicated as “0” steps. These 95 data points were removed.

Reviewing Cleaned Data

I compared both the raw files and cleaned files for each of the 3 files I cleaned to ensure the raw files were not altered in any way.

I also went through to ensure there were no N/A’s in any of the provided formulas I added. I also double check that there were no blank rows remaining from data that was removed from any of the three files listed above.

Analyze

I opted to analyze my data in R.

I double checked that packages needed to analyze the data I cleaned were installed.

The packages I double checked included:

* tidyverse
* tidyr
* dplyr
* lubridate
* ggplot2

The above packages were loaded and installed as follows:

library(tidyverse)

library(tidyr)

library(dplyr)

library(lubridate)

library(ggplot2)

install.packages("tidyverse")

install.packages("tidyr")

install.packages("dplyr")

install.packages("lubridate")

install.packages("ggplot2")

A screen shot of a computer program

Description automatically generated

Once I verified all packages were downloaded and installed, I proceeded to load my three cleaned data files:

|  |
| --- |
| dailySteps\_merged\_Cleaned.csv |
| dailyCalories\_merged\_Cleaned.csv |
| dailyIntensities\_merged\_Cleaned.csv |

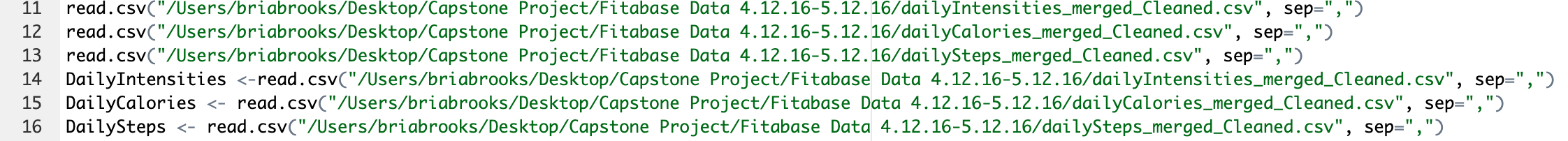
A white paper with black numbers

Description automatically generated

I loaded these files by using the read.csv function.

Once loaded, I realized that I did not name my data frames.

I then named my data frames as follows:



Analysis I. Daily Steps

I first ran summary(“DailySteps”) to get a summary of this data set.

The analysis showed that the average number of steps taken per day overall during this study was 8,495 steps. This output also provided the first and third quartile values of 5,057, and 11,140, respectively.

We also see that there were two extreme outliers: a minimum value of 356, and a maximum value of 36,019. Since these values are so far out of range of the mean, and 1st and 3rd quartile values, we can determine that most participants took between 5,057 and 11,140 steps daily during the duration of this study.

A screenshot of a computer

Description automatically generated

Analysis II. Daily Calories

For the second set of data, I also ran summary(DailyCalories)

The output for this workbook showed that, much like what was determined in the cleaning phase, the average daily calories burned was 2,313 calories.

We also see that the majority of daily calories burned fell in the range of 1,834 to 2,794 calories. The outliers of 52 calories, and 4900 calories do not appear to skew this dataset much higher or lower than the upper and lower quartile ranges.

A screenshot of a computer

Description automatically generated

Analysis III. Daily Intensities

This dataset appeared to be the most robust after running it through R.

I began by running summary(DailyIntensities).

The output for this dataset provided reporting for sedentary, lightly active, fairly active, and very active minutes across participants.

The summary shows that the maximum sedentary minutes had the largest number of minutes across all activity levels, and the fairly active minutes attributed to the least amount of maximum minutes.

I was surprised to find that very active minutes – the most intense activity level- had more maximum minutes than fairly active maximum minutes. This is shocking since I believed that the most intense activity level would entail the least amount of maximum minutes.

A screenshot of a computer

Description automatically generated

Determining if a relationship exists between calorie burn and activity intensity

Based on my analysis, I believe there is a direct relationship between calories burned, and activity intensity. Since more calories are burned by expending more energy, I am inclined to hypothesis that the more intense a participant’s activity level, the more calories will be burned.

To test this theory, I merged the DailyIntensity and DailyCalories datasets by naming this new data frame DailyCalorieIntensities using the ggplot function, as shown below:



When graphed, there is a clear positive correlation between the number of calories burned and very active minutes of physical activity. As the number of very active minutes increase, so does calorie burn, as shown below:

A graph with purple dots

Description automatically generated

I also input all three data sets into Tableau to determine if the participant step count over the course of the study led to greater calories burned.

Looking at the chart below, we can see that Step count alone is not the sole predictor in how many calories are tracked and burned.

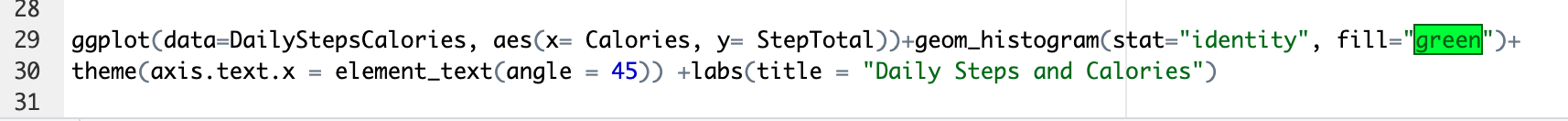
A screenshot of a graph

Description automatically generated

Share

A graph with green lines

Description automatically generated



When we compare this chart showing Step Total and Calories to the Chart listed in the Analysis section, we see that intensity level has a greater correlation with calories burned than does step total.

We can see that while step total does in fact burn the average amount calories, we saw depicted in the Analyze phase overall, there does not appear to be as strong of a relationship as activity intensity and calorie burn.

It is clear that while smart device users are using their fitness trackers for counting steps, time sleeping, and even weight tracking, the most beneficial pieces of data to track appear to be physical activity intensity as it relates to calorie burn.

Act

Findings from this analysis suggest that in order to burn the highest number of calories, users should focus on workout intensity instead of step count.

While step count may be meaningful to some users, Bellabeat would attract the most users by marketing the Leaf product as a device that can not only be worn multiple ways, but a device that can track workout intensity so that users are burning the maximum number of calories possible.

Bellabeat should investigate and implement software on the Leaf device (and other devices and the Bellabeat app) that can most accurately track respiratory and heart rates. This would give users the most accurate workout intensity reading, thereby providing accurate calorie burn.