Capstone Project for Google Finance Data Analyst Professional Certificate - Using FitBit dataset

Angela Truong

2023-03-01

Bellabeat Company

Website URL:

https://www.bellabeat.com

Scenario

Bellabeat is a successful small company that manufacturers health-focused smart products and have a target market of women. It has four products:

- Bellabeat app: App provides users with health data related to their activity, sleep, stress, mindfulness habits, and menstrual cycle.
- Leaf wellness tracker: It can be worn as a bracelet, necklace, or clip. The tracker connects to the Bellabeats app to track activity, sleep, and stress.
- Time wellness watch: The watch connects to the Bellabeats app to track activity, sleep, and stress. Additionally, it provides the user with insights into daily wellness.
- Spring water bottle: The water bottle connects to the Bellabeats app and tracks daily water intake to ensure users are appropriately hydrated throughout the day.

Business Task

I am a junior data analyst on the the Bellabeat marketing team and have been asked by Urska Srsen (co-founder and Chief Creative Office) to analyze smart device usage data in order to gain insight into how consumers use non-Bellabeat smart devices. She then wants me to select one Bellabeat product to apply these insights to in my presentation.

My analysis will be used to assist the Bellabeat marketing team to develop a marketing strategy specific to Bellabeats to help unlock new growth opportunities for the company. The goal of the new marketing strategy is to convince women to increase their activity levels by tracking various health-related metrices to improve health and track additional health-related metrics.

Description of Data Sources

I used public survey data that was downloaded from the Kaggle website (https://www.kaggle.com). The dataset is called "FitBit Fitness Tracker Data" (dataset made available through Mobius). The data set contains personal fitness tracker data from 33 Fitbit users who consented to the submission and use of their personal data. The time period of the data is April 12, 2016 to May 11, 2016 (one month). The data set has a total of 16 tables.

- 5 tables have Daily data related to Sleep, Activity level, Steps taken, and Calories burned. The tables have between 3 and 15 columns. The sleep table has 413 rows and the Activity, Steps, and Calories tables each have 940 rows.
  These were the tables that I used for my analysis.
- 3 tables have Hourly data related to Activity level, Steps taken, and Calories burned. The tables have 3 or 4 columns and each have 22,099 rows.
- 4 tables have Minutes data related to Activity level, Sleep, Steps taken (wide table and narrow table), Calories burned (wide table). The wide tables have 62 columns and 21,646 rows. The narrow tables have 3-5 columns and number of rows ranging between 21,646 and 1,325,580.
- 1 table has Heart Rate data by second. It has 3 columns and 2,483,658 rows.
- 1 table has METs data by minute (METs shows the ratio of working metabolic data vs. resting metabolic data). It has 3 columns and 1,325,580 rows.
- 1 table has Weight Log data for the FitBit users who manually logged weight data. Table has 8 columns and 67 rows.

In terms of licensing, privacy, security, and accessibility of the survey data, the manufacturer of FitBits and database administrator of the Kaggle website are responsible. I am responsible for the security and accessibility of the data on my personal computer. My personal computer is locked unless the password is entered.

Limitations of the Data

- Only 33 people submitted FitBit data, which is not representative of the entire population of people who own and use a FitBit.
- Only one month of data was collected - April 12, 2016 to May 11, 2016. Potentially different insights would have been determined from a longer period of time.
- Demographics data was not collected so it is not known what tracking data was submitted by women only. Bellabeat products are targeted to women so only FitBit data submitted by women is relevant.
- For the Weight Log data, only 8 people submitted it. Because of the lack of data, it was not used for my analysis.
- The tables for minutes and hourly data have a large number of rows. Since I felt that the tables showing daily information had the level of data that I needed, I did not use the minutes and hourly data tables.

Prepare Data and R (Statistical Programming Language) for Analysis

Install the needed R packages. After installing the R packages, I added functions eval=FALSE and echo=FALSE to code chunk so the code will not show, copy, or evaluate after the first time of running.

Run library code chunk that included message=FALSE so error messages will not show in report.

```
library("tidyverse")
library("here")
library("skimr")
library("janitor")
library("dplyr")
library("readr")
```

Since the downloaded tables are csv file, run the below code chunk so data frames are created for the tables used in the analysis.

```
sleepDay_merged <- read_csv("sleepDay_merged.csv")
minuteSleep_merged <- read_csv("minuteSleep_merged.csv")
hourlyCalories_merged <- read_csv("hourlyCalories_merged.csv")
dailySteps_merged <- read_csv("dailySteps_merged.csv")
dailyActivity_merged <- read_csv("dailyActivity_merged.csv")
dailyCalories_merged <- read_csv("dailyCalories_merged.csv")
dailyIntensities_merged <- read_csv("dailyIntensities_merged.csv")
```

Dataframes and Tools Used for Analysis

I used 4 dataframes for the analysis, insights, and recommendations. These dataframes show daily data for Sleep, Activity Level, and Calories burned.

For tools utilized, I used only R Statistical Programming Language.
I considered using Google Sheets or Microsoft Excel to create pivot tables, but I decided that visualizations showed the data in the best format for stakeholders.

Analysis of Sleep Data

Per the below website URL, adults 18-64 years old should get 7-9 hours of sleep per night and adults 65+ years old should get 7-8 hours per night.

https://www.sleepfoundation.org/how-sleep-works/how-much-sleep-do-we-really-need

For the sleepDay_merged table, do Group by function for the Id column then calculate the mean for TotalMinutesAsleep and TotalTimeInBed.

```
avg_sleep_hours_df <- sleepDay_merged %>%
  mutate(
    TotalHoursAsleep = TotalMinutesAsleep / 60,
    TotalHoursInBed = TotalTimeInBed / 60
  ) %>%
  group_by(Id) %>%
  summarize(
    avg_sleep_hours = mean(TotalHoursAsleep),
    avg_bed_hours = mean(TotalHoursInBed)
  )
```

Do Head function to see 6 rows of data and calculations for average number of hours asleep and average number of hours in bed.

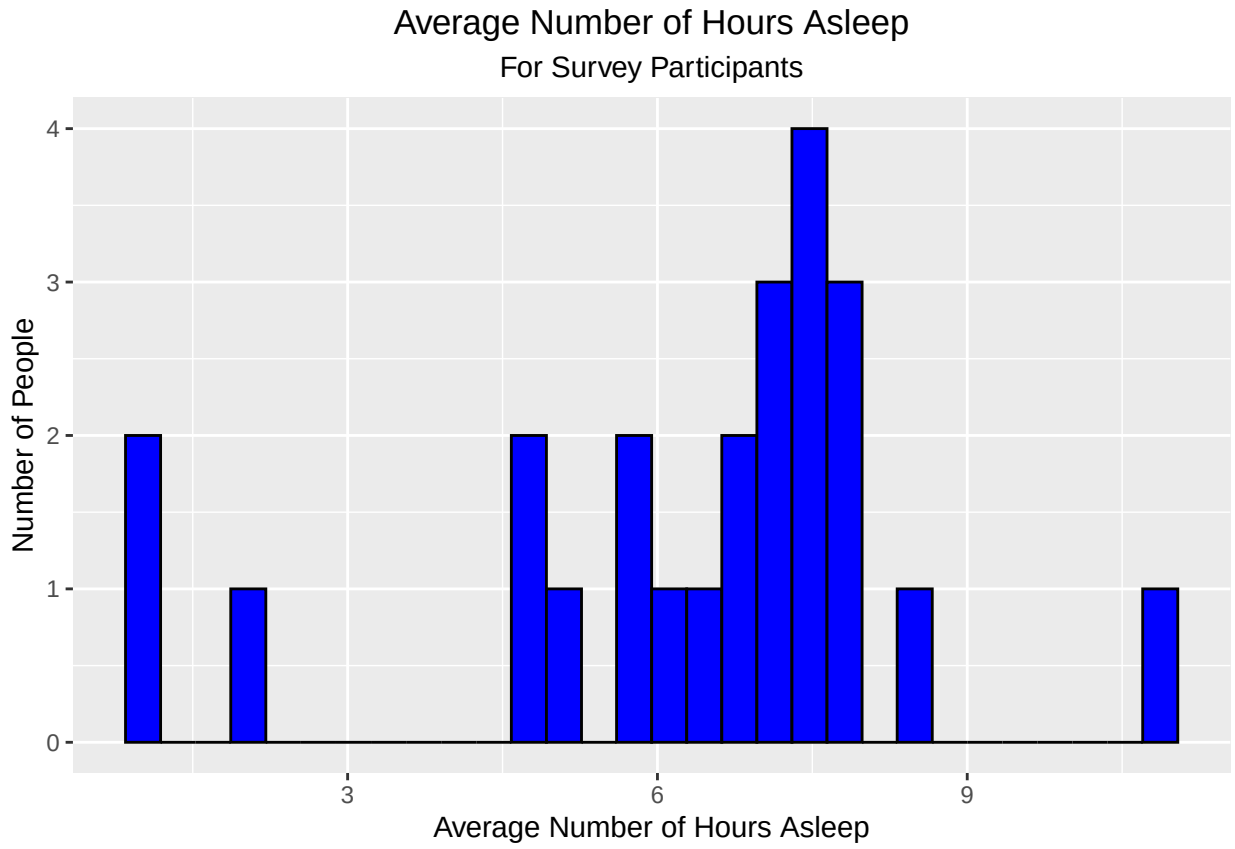
```
head(avg_sleep_hours_df)
```

```
## # A tibble: 6 x 3
##           Id avg_sleep_hours avg_bed_hours
##        <dbl>           <dbl>         <dbl>
## 1 1503960366            6.00          6.39
## 2 1644430081            4.9           5.77
## 3 1844505072           10.9          16.0
## 4 1927972279            6.95          7.30
## 5 2026352035            8.44          8.96
## 6 2320127002            1.02          1.15
```

After analyzing the above table, I decided that people probably want to track the number of hours asleep. Knowing the number of hours in bed is not useful to change sleep behavior.

Create a histogram to see the average number of hours of asleep for each person who tracked this information via the FitBit watch.

```
ggplot(avg_sleep_hours_df, aes(x=avg_sleep_hours)) + geom_histogram(bins=30, color="black", fill="blue")
  labs(title = "Average Number of Hours Asleep", subtitle = "For Survey Participants", x="Average Numbe
  theme(plot.title = element_text(hjust = 0.5), plot.subtitle = element_text(hjust = 0.5))
```



Findings of Analysis: For the 24 people who tracked sleep, 4 people averaged either less than 3 hours or more than 9 hours. These are anomalies in the data. The other 21 people averaged 6-9 hours of sleep.

Analysis of Daily Activity Data

Create scatter plot of "TotalSteps" and "Calories" columns in dailyActivity_merged dataframe.

```
ggplot(data=dailyActivity_merged, aes(x=TotalSteps, y=Calories)) +
  geom_point(color="blue") +
      labs(title = "Total Steps vs. Calories Burned", subtitle = "For Survey Participants", x="Total St
  theme(plot.title = element_text(hjust = 0.5), plot.subtitle = element_text(hjust = 0.5))
```

Create scatter plot of "TotalDistance" and "Calories" columns in dailyActivity_merged dataframe.

```
ggplot(data=dailyActivity_merged, aes(x=TotalDistance, y=Calories)) +
  geom_point(color="blue") +
    labs(title = "Total Distance vs. Calories Burned", subtitle = "For Survey Participants", x="Total Di
  theme(plot.title = element_text(hjust = 0.5), plot.subtitle = element_text(hjust = 0.5))
```

Findings of Analysis: The scatter plots of Total Steps vs. Calories Burned and Total Distance vs. Calories Burned show a positive correlation.

Documentation of Data Cleaning and Manipulation

All of the data used for the analysis is included in this report, but a few data cleaning activities should be done before presenting high-level analysis information and recommendations to stakeholders.

- For the Daily Sleep data, in the histogram remove the anomalies for less than 6 hours and more than 10 hours of average sleep for the survey participants.
- For the Daily Activity data, in the scatter plots remove the data points for Steps of 0 and Calories Burned of less than 1000.

Trends in Smart-Device Usage for Fitness Tracking:

- Per https://www.pewresearch.org a 2020 study found that 21% (about one-in-five) American adults use a smart watch or fitness tracker. Women are more likely to use fitness trackers than men.
- Users monitor the data by utilizing the corresponding app and can consider lifestyle changes to improve physical and mental health.
- Scatter plots created during the analysis showed a positive correlation between 1) Total Steps and Total Calories Burned and 2) Total Distance and Total Calories Burned.
- FitBit and Bellabeat products both track number of steps taken, distance traveled, calories burned, weight, heart rate, and sleep. Bellabeat also tracks stress, menstrual cycle, and mindfulness. These three additional tracking features set Bellabeat apart from FitBit and should be highlighted in the marketing strategy as a reason that women should buy Bellabeat products instead of FitBit.

Summary of Analysis and Recommendations

Sleep Data

```
Insight:  Based on the analysis for people who tracked sleep, about 90% averaged 6-9 hours  asleep whic
```

```
Recommendation:  In the marketing strategy, include the guidance of getting 7-9 hours of sleep and the
```

Daily Activity Data

```
Insights:  Based on the analysis of Total Steps vs. Calories Burned and Total Distance vs. Calories Bur
```

```
Recommendations:  In the marketing strategy, highlight the importance of getting daily exercise (even a
```

Other Tracking Features in Bellabeat products

```
Insight:  As mentioned above in the Trends in Smart-Device Usage, Bellabeat has additional tracking fea
```

```
Recommendation:  In the marketing strategy, list all of the tracking features offered and especially hi
* Stress management and mindfulness tracking - Studies have found that concentrating on managing stress
* Menstrual cycle tracking - Some women may find benefits in tracking of their menstrual cycle.
```

Website about focusing on stress management and mindfulness:

https://www.mindful.org/how-to-manage-stress-with-mindfulness-and-meditation/

Website about tracking menstrual cycle.

https://helloclue.com/articles/cycle-a-z/5-reasons-you-should-pay-attention-to-your-period