Google Data Analytics Professional Certificate Capstone Project

Caleb Koresh

2022-07-22

## Introduction

For my data analytics capstone project, I have selected to complete the Bellabeat product analysis case study. Bellabeat is a technology manufacturer of women’s health and fitness products. I will be placing myself in the shoes of a junior data analyst on their marketing analytics team. I will be following the data analytics process outlined in the course:

1. **Ask**
2. **Prepare**
3. **Process**
4. **Analyze**
5. **Share**
6. **Act**

## Background

The founders of Bellabeat, Srsen and Mur are motivated by their goal of providing quality products that empower and motivate women by providing information about their activity, sleep, stress, and reproductive health. Their products are available from various online retailers as well as their website. Their advertising strategy focuses on digital advertising such as Google and major social media networks.

In this hypothetical scenario, I have been instructed to analyze available data in order to find insights into how potential customers use their smart technology on a daily basis. Based on the analysis, I must deliver a high-level recommendation on how to adjust their marketing strategy to accommodate the new information and optimize their advertising for one of the current Bellabeat products.

## 1) Ask

#### Identify the Business Task:

Deliver an updated marketing strategy for one of the Bellabeat products based on the results of the data analysis on non-Bellabeat products.

#### Consider Key Stakeholders:

**Urska Srsen:** Cofounder and Chief Creative Officer of Bellabeat

**Sando Mur:** Cofounder and key member of the Bellabeat executive team

Srsen and Mur’s company has experienced rapid growth since its founding and has grown into a global company. They hope to use data insights to continue this growth and reach more customers.

#### Key Questions:

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

## 2) Prepare

My hypothetical supervisor has directed me to a publicly available data set that was linked below:

<https://www.kaggle.com/datasets/arashnic/fitbit>

The data was gathered from thirty fitbit users between 3/12/2016 and 5/12/2016 and contains various different statistics such as daily activity, steps, and heart rate. The data is slightly outdated but smart technology has not changed substantially since the data was collected. Another consideration that must be made before formulating any concrete conclusions is the relatively small data size and the survey method. The data was collected through Amazon Mechanical Turk and may not be representative of all smart technology users.

## 3) Process

#### Load Packages

library(tidyverse)  
library(ggplot2)  
library(dplyr)  
library(readr)  
library(lubridate)

#### Load Data

I chose to examine the daily and hourly data to find more general trends in everyday usage, in order to find potential value that Bellabeat could include in their marketing strategy. Unfortunately the weight data set only contained eight participants so no conclusions could be drawn and the data was not included.

daily\_activity <- read\_csv("Fitabase Data 4.12.16-5.12.16/dailyActivity\_merged.csv")  
hourly\_calories <- read\_csv("Fitabase Data 4.12.16-5.12.16/hourlyCalories\_merged.csv")  
hourly\_intensities <- read\_csv("Fitabase Data 4.12.16-5.12.16/hourlyIntensities\_merged.csv")  
hourly\_steps <- read\_csv("Fitabase Data 4.12.16-5.12.16/hourlySteps\_merged.csv")  
sleep <- read\_csv("Fitabase Data 4.12.16-5.12.16/sleepDay\_merged.csv")

#### Preview Data

head(daily\_activity)

## # A tibble: 6 × 15  
## Id ActivityDate TotalSteps TotalDistance TrackerDistance LoggedActivitie…  
## <dbl> <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 1.50e9 4/12/2016 13162 8.5 8.5 0  
## 2 1.50e9 4/13/2016 10735 6.97 6.97 0  
## 3 1.50e9 4/14/2016 10460 6.74 6.74 0  
## 4 1.50e9 4/15/2016 9762 6.28 6.28 0  
## 5 1.50e9 4/16/2016 12669 8.16 8.16 0  
## 6 1.50e9 4/17/2016 9705 6.48 6.48 0  
## # … with 9 more variables: VeryActiveDistance <dbl>,  
## # ModeratelyActiveDistance <dbl>, LightActiveDistance <dbl>,  
## # SedentaryActiveDistance <dbl>, VeryActiveMinutes <dbl>,  
## # FairlyActiveMinutes <dbl>, LightlyActiveMinutes <dbl>,  
## # SedentaryMinutes <dbl>, Calories <dbl>

head(hourly\_calories)

## # A tibble: 6 × 3  
## Id ActivityHour Calories  
## <dbl> <chr> <dbl>  
## 1 1503960366 4/12/2016 12:00:00 AM 81  
## 2 1503960366 4/12/2016 1:00:00 AM 61  
## 3 1503960366 4/12/2016 2:00:00 AM 59  
## 4 1503960366 4/12/2016 3:00:00 AM 47  
## 5 1503960366 4/12/2016 4:00:00 AM 48  
## 6 1503960366 4/12/2016 5:00:00 AM 48

head(hourly\_intensities)

## # A tibble: 6 × 4  
## Id ActivityHour TotalIntensity AverageIntensity  
## <dbl> <chr> <dbl> <dbl>  
## 1 1503960366 4/12/2016 12:00:00 AM 20 0.333  
## 2 1503960366 4/12/2016 1:00:00 AM 8 0.133  
## 3 1503960366 4/12/2016 2:00:00 AM 7 0.117  
## 4 1503960366 4/12/2016 3:00:00 AM 0 0   
## 5 1503960366 4/12/2016 4:00:00 AM 0 0   
## 6 1503960366 4/12/2016 5:00:00 AM 0 0

head(hourly\_steps)

## # A tibble: 6 × 3  
## Id ActivityHour StepTotal  
## <dbl> <chr> <dbl>  
## 1 1503960366 4/12/2016 12:00:00 AM 373  
## 2 1503960366 4/12/2016 1:00:00 AM 160  
## 3 1503960366 4/12/2016 2:00:00 AM 151  
## 4 1503960366 4/12/2016 3:00:00 AM 0  
## 5 1503960366 4/12/2016 4:00:00 AM 0  
## 6 1503960366 4/12/2016 5:00:00 AM 0

head(sleep)

## # A tibble: 6 × 5  
## Id SleepDay TotalSleepRecor… TotalMinutesAsl… TotalTimeInBed  
## <dbl> <chr> <dbl> <dbl> <dbl>  
## 1 1503960366 4/12/2016 12:00:0… 1 327 346  
## 2 1503960366 4/13/2016 12:00:0… 2 384 407  
## 3 1503960366 4/15/2016 12:00:0… 1 412 442  
## 4 1503960366 4/16/2016 12:00:0… 2 340 367  
## 5 1503960366 4/17/2016 12:00:0… 1 700 712  
## 6 1503960366 4/19/2016 12:00:0… 1 304 320

#### Adjust Date and Time Formats

Making the date and time formats consistent throughout the data will make the analysis much simpler.

daily\_activity$ActivityDate <- mdy(daily\_activity$ActivityDate)  
hourly\_calories$ActivityHour <- as\_datetime(hourly\_calories$ActivityHour, format = "%m/%d/%Y %I:%M:%S %p")  
hourly\_intensities$ActivityHour <- as\_datetime(hourly\_intensities$ActivityHour, format = "%m/%d/%Y %I:%M:%S %p")  
hourly\_steps$ActivityHour <- as\_datetime(hourly\_steps$ActivityHour, format = "%m/%d/%Y %I:%M:%S %p")  
sleep$SleepDay <- as\_datetime(sleep$SleepDay, format = "%m/%d/%Y %I:%M:%S %p")

#### Remove Duplicates and Missing Data

#Daily Activity  
daily\_activity <- daily\_activity %>%   
 distinct() %>%  
 drop\_na()  
  
#Hourly Calories  
hourly\_calories <- hourly\_calories %>%   
 distinct() %>%  
 drop\_na()  
  
#Hourly Intensities   
hourly\_intensities <- hourly\_intensities %>%   
 distinct() %>%  
 drop\_na()  
  
#Hourly Steps  
hourly\_steps <- hourly\_steps %>%   
 distinct() %>%  
 drop\_na()

#### Summarize Data

I began by looking at some basic statistics within the data in hopes of finding potential areas of value that Bellabeat could deliver to the customer.

##### Daily Activity

daily\_activity %>%   
 select(TotalSteps,   
 TotalDistance) %>%  
 summary()

## TotalSteps TotalDistance   
## Min. : 0 Min. : 0.000   
## 1st Qu.: 3790 1st Qu.: 2.620   
## Median : 7406 Median : 5.245   
## Mean : 7638 Mean : 5.490   
## 3rd Qu.:10727 3rd Qu.: 7.713   
## Max. :36019 Max. :28.030

daily\_activity %>%   
 select(VeryActiveDistance,   
 ModeratelyActiveDistance,  
 LightActiveDistance) %>%  
 summary()

## VeryActiveDistance ModeratelyActiveDistance LightActiveDistance  
## Min. : 0.000 Min. :0.0000 Min. : 0.000   
## 1st Qu.: 0.000 1st Qu.:0.0000 1st Qu.: 1.945   
## Median : 0.210 Median :0.2400 Median : 3.365   
## Mean : 1.503 Mean :0.5675 Mean : 3.341   
## 3rd Qu.: 2.053 3rd Qu.:0.8000 3rd Qu.: 4.782   
## Max. :21.920 Max. :6.4800 Max. :10.710

daily\_activity %>%  
 select(VeryActiveMinutes,  
 FairlyActiveMinutes,  
 LightlyActiveMinutes,  
 SedentaryMinutes)%>%  
 summary()

## VeryActiveMinutes FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes  
## Min. : 0.00 Min. : 0.00 Min. : 0.0 Min. : 0.0   
## 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.:127.0 1st Qu.: 729.8   
## Median : 4.00 Median : 6.00 Median :199.0 Median :1057.5   
## Mean : 21.16 Mean : 13.56 Mean :192.8 Mean : 991.2   
## 3rd Qu.: 32.00 3rd Qu.: 19.00 3rd Qu.:264.0 3rd Qu.:1229.5   
## Max. :210.00 Max. :143.00 Max. :518.0 Max. :1440.0

daily\_activity %>%  
 select(Calories)%>%  
 summary()

## Calories   
## Min. : 0   
## 1st Qu.:1828   
## Median :2134   
## Mean :2304   
## 3rd Qu.:2793   
## Max. :4900

##### Hourly Calories

hourly\_calories %>%  
 select(Calories)%>%  
 summary()

## Calories   
## Min. : 42.00   
## 1st Qu.: 63.00   
## Median : 83.00   
## Mean : 97.39   
## 3rd Qu.:108.00   
## Max. :948.00

##### Hourly Intensity

hourly\_intensities %>%  
 select(TotalIntensity,  
 AverageIntensity) %>%  
 summary()

## TotalIntensity AverageIntensity  
## Min. : 0.00 Min. :0.0000   
## 1st Qu.: 0.00 1st Qu.:0.0000   
## Median : 3.00 Median :0.0500   
## Mean : 12.04 Mean :0.2006   
## 3rd Qu.: 16.00 3rd Qu.:0.2667   
## Max. :180.00 Max. :3.0000

##### Hourly Steps

hourly\_steps %>%  
 select(StepTotal) %>%  
 summary()

## StepTotal   
## Min. : 0.0   
## 1st Qu.: 0.0   
## Median : 40.0   
## Mean : 320.2   
## 3rd Qu.: 357.0   
## Max. :10554.0

##### Sleep

sleep %>%  
 select(TotalMinutesAsleep,  
 TotalTimeInBed)%>%  
 summary()

## TotalMinutesAsleep TotalTimeInBed   
## Min. : 58.0 Min. : 61.0   
## 1st Qu.:361.0 1st Qu.:403.0   
## Median :433.0 Median :463.0   
## Mean :419.5 Mean :458.6   
## 3rd Qu.:490.0 3rd Qu.:526.0   
## Max. :796.0 Max. :961.0

I gathered two interesting insights from these summaries. I noticed a substantial difference between the median and mean time asleep compared to the time spent in bed. I also saw that the median and mean daily steps were well below the advisable amount for general health. Consequently, I will be analyzing these differences in the next section.

#### Merge Data into Daily and Hourly

Merging the data into two different tables will simplify the analysis process.

#Make compatible   
colnames(sleep)[2] <- "ActivityDate"   
  
  
#Merge Data  
daily <- merge(daily\_activity, sleep, by = c("Id", "ActivityDate"))  
hourly <- merge(hourly\_calories, hourly\_intensities, by = c("Id", "ActivityHour"))  
hourly <- merge(hourly, hourly\_steps, by = c("Id", "ActivityHour"))  
head(hourly)

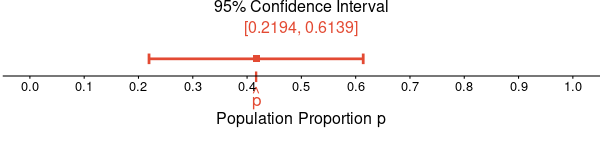
## 4) Analyze

### Sleep

In order to expand on the current marketing strategy I will attempt to find potential value that Bellabeat can offer customers. The reason I chose to analyze time in bed and time asleep was the potential for diagnosing insomnia or identifying poor sleeping habits. Healthy people should be able to fall asleep in 10-20 minutes and taking longer than 30 has shown to decrease sleep quality (<https://www.healthline.com/health/healthy-sleep/how-long-does-it-take-to-fall-asleep#normal-sleep>).

sleep\_data <- daily %>%  
 group\_by(Id) %>%  
 summarise(mean\_time\_in\_bed = mean(TotalTimeInBed), mean\_time\_asleep = mean(TotalMinutesAsleep))  
  
sleep\_data <- sleep\_data %>%  
 mutate(potential\_insomnia = case\_when(  
 (mean\_time\_in\_bed - mean\_time\_asleep) >= 30 ~ "Yes",  
 (mean\_time\_in\_bed - mean\_time\_asleep) < 30 ~ "No"  
 ))  
  
sleep\_data %>%  
 count(potential\_insomnia)

## # A tibble: 2 × 2  
## potential\_insomnia n  
## <chr> <int>  
## 1 No 14  
## 2 Yes 10



**Proportion of Users with Poor Sleeping Habits**

(<https://artofstat.com/web-apps>)

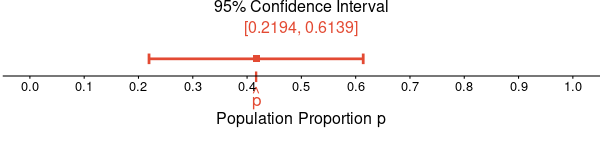
Simple statistical analysis through Art of Stat revealed that the 95% confidence interval of the population proportion of fitbit users who have poor sleeping habits is [0.2194, 0.6139]. In other words, there is a 95% chance that the population proportion is between 0.2194 and 0.6139. This indicates that a large proportion of consumers can benefit from Bellabeat sleep tracking data.

### Steps

The CDC recommends 10000 steps per day for maintaining general health so it is worth investigating if users are meeting these goals. If they are consistently failing to reach this goal then using Bellabeat’s step tracking data can empower users to improve their overall well-being. (<https://www.cdc.gov/diabetes/prevention/pdf/postcurriculum_session8.pdf>)

step\_data <- daily %>%  
 group\_by(Id) %>%  
 summarise(mean\_daily\_steps = mean(TotalSteps))  
  
step\_data <- step\_data %>%  
 mutate(adequate\_steps = case\_when(  
 mean\_daily\_steps >= 10000 ~ "Yes",  
 mean\_daily\_steps < 10000 ~ "No"  
 ))  
  
step\_data %>%  
 count(adequate\_steps)

## # A tibble: 2 × 2  
## adequate\_steps n  
## <chr> <int>  
## 1 No 19  
## 2 Yes 5



**Proportion of Users Who Average 10000 Steps Per Day**

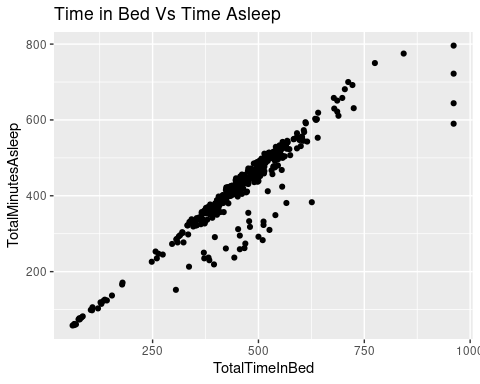
(<https://artofstat.com/web-apps>)

The 95% confidence interval for the population proportion of fitbit users who reach the reccomended 10000 steps per day is [0.04586, 0.3708]. This interval indicates that most users fail to reach this goal. These users could greatly benefit in overall well-being if Bellabeat were to provide them with the necessary data and reminders to reach this goal on a daily basis.

## 5) Share

The trend between time in bed and time asleep is clearly going to be linear but the outliers from the line can be insightful. A large gap away from the expected line can be an indication of insomnia or poor sleeping habits.

ggplot(data = daily, aes(x = TotalTimeInBed, y = TotalMinutesAsleep)) +  
 geom\_point() +  
 labs(title = "Time in Bed Vs Time Asleep")



## 6) Act

After analyzing the data, two key insights were revealed that could be used to adjust Bellabeats marketing strategy. The data set was not large enough to make concrete conclusions, but was more oriented towards practicing key data manipulation skills. Despite this, analyzing the data and using this information to find new ways to use smart technology can be beneficial to their marketing strategy. Bellabeat is focused on empowering women through data and these insights can be used to do just that.

1. A large portion of smart technology users may have poor sleeping habits and potentially even insomnia. Sleep quality diminishes when time to fall asleep exceeds thirty minutes so it is important to be aware of this data. Bellabeat can emphasize that their products keep track of this data and more features can be implemented to help educate and remind people about healthy sleeping habits, or even advise consulting a doctor about insomnia. This adds value to their products while differentiating themselves from other smart technology.
2. More than half of the users in the study were not meeting CDC recommendations regarding taking 10000 steps per day. Being sedentary is a serious detriment to ones health and Bellabeats products are designed to help people stay on top of their daily activity. Revolving most of their marketing around these concepts is crucial to demonstrating the value that these products have for improving personal health.