Capstone Project Bellabeat

Ebere James

2023-06-10

Introduction

Welcome ! After completing the google data analytics course, I have chosen this dataset for my capstone project. Here I will be doing real-world analysis in order to make data-driven decisions. I will be using the ask, prepare, process, analyze, share and act processes of data anlysis.

About Bellabeat

Urška Sršen and Sando Mur founded Bellabeat, a high-tech company that manufactures health-focused smart products. Sršen used her background as an artist to develop beautifully designed technology that informs and inspires women around the world. Collecting data on activity, sleep, stress, and reproductive health has allowed Bellabeat to empower women with knowledge about their own health and habits. Since it was founded in 2013, Bellabeat has grown rapidly and quickly positioned itself as a tech-driven wellness company for women.

Scenario of the Study

In this study, I will analyze the use of one of bellabeat's smart devices by consumers to establish and identify trends and patterns in the usage of this device. Based on those established trends and patterns, high level recommendations will be presented to the bellabeat marketing team.

The Ask Phase

In this Phase, I tried to understand the problem, I'm trying to solve and the data I'm working with, In orer for that to happen, I need to ask questions 1. What are the trends in smart device usage? We need to establish the trends in smart device usage and identify the needs if the consumer, then tailor the company's marketing strategy to meet the needs of the consumer. 2.Who are the main stakeholders ? The main stakeholders in this project are Urska Srsen, the co-founder and chief creative officer; Sando Mur who is a co-founder as well. The marketing team are also to be considered and carried along in this project ## Business task Identify and establosh trends in smart device usage data and provide recommendations based on the data to bellabeat marketing team to foster the growth of the company

Loading Packages

Now I'm going to load these packages. If you notice I used the 'warning=FALSE' and 'message= FALSE' to save space by preventing the generation of error and warning messages.

```
library(tidyverse)
library(lubridate)
library(dplyr)
library(ggplot2)
library(tidyr)
library(skimr)
library(here)
```

```
library(janitor)
library(readr)
```

Importing Datasets

For this project I will be using the fitbit dataset from kaggle.

```
Activity <- read_csv("Fitabase Data 4.12.16-5.12.16/dailyActivity_merged.csv")
Calories <- read_csv("Fitabase Data 4.12.16-5.12.16/hourlyCalories_merged.csv")
Intensity <- read_csv("Fitabase Data 4.12.16-5.12.16/hourlyIntensities_merged.csv")
Sleep <- read_csv("Fitabase Data 4.12.16-5.12.16/sleepDay_merged.csv")
Weight <- read_csv("Fitabase Data 4.12.16-5.12.16/weightLogInfo_merged.csv")
```

Now I'm going to take a glance at the dataset to be sure we are in order and it matches with what we have on the spreadsheet

```
head(Activity)
```

```
# A tibble: 6 x 15
         Id ActivityDate TotalSteps TotalDistance TrackerDistance
      <dbl> <chr>             <dbl>         <dbl>           <dbl>
1 1503960366 4/12/2016         13162          8.5             8.5
2 1503960366 4/13/2016         10735          6.97            6.97
3 1503960366 4/14/2016         10460          6.74            6.74
4 1503960366 4/15/2016          9762          6.28            6.28
5 1503960366 4/16/2016         12669          8.16            8.16
6 1503960366 4/17/2016          9705          6.48            6.48
# i 10 more variables: LoggedActivitiesDistance <dbl>,
#   VeryActiveDistance <dbl>, ModeratelyActiveDistance <dbl>,
#   LightActiveDistance <dbl>, SedentaryActiveDistance <dbl>,
#   VeryActiveMinutes <dbl>, FairlyActiveMinutes <dbl>,
#   LightlyActiveMinutes <dbl>, SedentaryMinutes <dbl>, Calories <dbl>
```

```
head(Sleep)
```

```
# A tibble: 6 x 5
         Id SleepDay        TotalSleepRecords TotalMinutesAsleep TotalTimeInBed
      <dbl> <chr>                       <dbl>              <dbl>          <dbl>
1 1503960366 4/12/2016 12:0~                1                327            346
2 1503960366 4/13/2016 12:0~                2                384            407
3 1503960366 4/15/2016 12:0~                1                412            442
4 1503960366 4/16/2016 12:0~                2                340            367
5 1503960366 4/17/2016 12:0~                1                700            712
6 1503960366 4/19/2016 12:0~                1                304            320
```

```
head(Weight)
```

```
# A tibble: 6 x 8
         Id Date       WeightKg WeightPounds   Fat   BMI IsManualReport   LogId
      <dbl> <chr>         <dbl>        <dbl> <dbl> <dbl> <lgl>            <dbl>
1 1503960366 5/2/2016 ~     52.6         116.    22  22.6 TRUE           1.46e12
2 1503960366 5/3/2016 ~     52.6         116.    NA  22.6 TRUE           1.46e12
3 1927972279 4/13/2016~    134.          294.    NA  47.5 FALSE          1.46e12
4 2873212765 4/21/2016~     56.7         125.    NA  21.5 TRUE           1.46e12
5 2873212765 5/12/2016~     57.3         126.    NA  21.7 TRUE           1.46e12
6 4319703577 4/17/2016~     72.4         160.    25  27.5 TRUE           1.46e12
```

```
head(Intensity)
```

```
# A tibble: 6 x 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(Calories)
```

```
# A tibble: 6 x 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
```

Process Phase

The data seems pretty clean, there aren't any spelling errors or misfield values, however there are some problems with the data type as regards date and time in the activity,intensity, calories and sleep tables. The 'fat' column in the weight table had too many null values so I removed the entire column.

```
# Weight
Weight$Fat <- NULL

# Activity
Activity$ActivityDate= as.POSIXct(Activity$ActivityDate, format = "%m/%d/%y", tz =Sys.timezone())
Activity$date <- format(Activity$ActivityDate, "%m/%d/%y")

# Intensity
Intensity$ActivityHour= as.POSIXct(Intensity$ActivityHour, format="%m/%d/%y %p", tz = Sys.timezone())
Intensity$time <- format(Intensity$ActivityHour, format= "%H:%M:%S")
Intensity$date <- format(Intensity$ActivityHour, format = "%m/%d/%y")

# Calories
Calories$ActivityHour = as.POSIXct(Calories$ActivityHour,format ="%m/%d/%y %p", tz =Sys.timezone())
Calories$time <- format (Calories$ActivityHour, format = "%H:%M:%S")
Calories$date <- format (Calories$ActivityHour, format = "%m/%d/%y")

# Sleep
Sleep$Sleepday  = as.POSIXct(Sleep$SleepDay, format = "%m/%d/%y I:%M:%S %p", tz = Sys.timezone())
Sleep$date <- format(Sleep$SleepDay, format = '%m/%d/%y')
```

Analyze Phase

Now that the data is clean we can go ahead and summarize the data, first we need an idea of how many participants we have in this study using the dplyr function.

```
library(dplyr)
n_distinct(Activity$Id)
```

```
[1] 33
```

```
n_distinct(Calories$Id)
```

```
[1] 33
```

```
n_distinct(Weight$Id)
```

```
[1] 8
```

```
n_distinct(Intensity$Id)
```

```
[1] 33
```

```
n_distinct(Sleep$Id)
```

```
[1] 24
```

From the above we know there are 33 participants in the activity dataset, 33 in the Calories dataset, 8 in the weight, while 33 and 24 people participated in the Intensity and sleep surveys respectively. However due to the small number of participants in the weight survey we will not be making recommendations based off them. because it will lead to inconclusive results. We will now summarize the rest.

```
# Activity
Activity %>%
  select(TotalSteps,
         TotalDistance,
         SedentaryMinutes, Calories) %>%
  summary()
```

```
   TotalSteps    TotalDistance    SedentaryMinutes    Calories
 Min.   :    0   Min.   : 0.000   Min.   :   0.0   Min.   :   0
 1st Qu.: 3790   1st Qu.: 2.620   1st Qu.: 729.8   1st Qu.:1828
 Median : 7406   Median : 5.245   Median :1057.5   Median :2134
 Mean   : 7638   Mean   : 5.490   Mean   : 991.2   Mean   :2304
 3rd Qu.:10727   3rd Qu.: 7.713   3rd Qu.:1229.5   3rd Qu.:2793
 Max.   :36019   Max.   :28.030   Max.   :1440.0   Max.   :4900
```

```
# Explore the Variety of these activities
Activity %>%
  select(VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes) %>%
  summary()
```

```
 VeryActiveMinutes FairlyActiveMinutes LightlyActiveMinutes
 Min.   :  0.00    Min.   :  0.00      Min.   :  0.0
 1st Qu.:  0.00    1st Qu.:  0.00      1st Qu.:127.0
 Median :  4.00    Median :  6.00      Median :199.0
 Mean   : 21.16    Mean   : 13.56      Mean   :192.8
 3rd Qu.: 32.00    3rd Qu.: 19.00      3rd Qu.:264.0
 Max.   :210.00    Max.   :143.00      Max.   :518.0
```

```
# Intensity
Intensity %>%
  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
```

```
# Calories
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
```

```
# Sleep
Sleep %>%
  select(TotalSleepRecords, TotalMinutesAsleep, TotalTimeInBed) %>%
  summary()
```

```
 TotalSleepRecords TotalMinutesAsleep TotalTimeInBed
 Min.   :1.000     Min.   : 58.0      Min.   : 61.0
 1st Qu.:1.000     1st Qu.:361.0      1st Qu.:403.0
 Median :1.000     Median :433.0      Median :463.0
 Mean   :1.119     Mean   :419.5      Mean   :458.6
 3rd Qu.:1.000     3rd Qu.:490.0      3rd Qu.:526.0
 Max.   :3.000     Max.   :796.0      Max.   :961.0
```

```
# Weight
Weight %>%
  select(WeightKg, BMI) %>%
  summary()
```

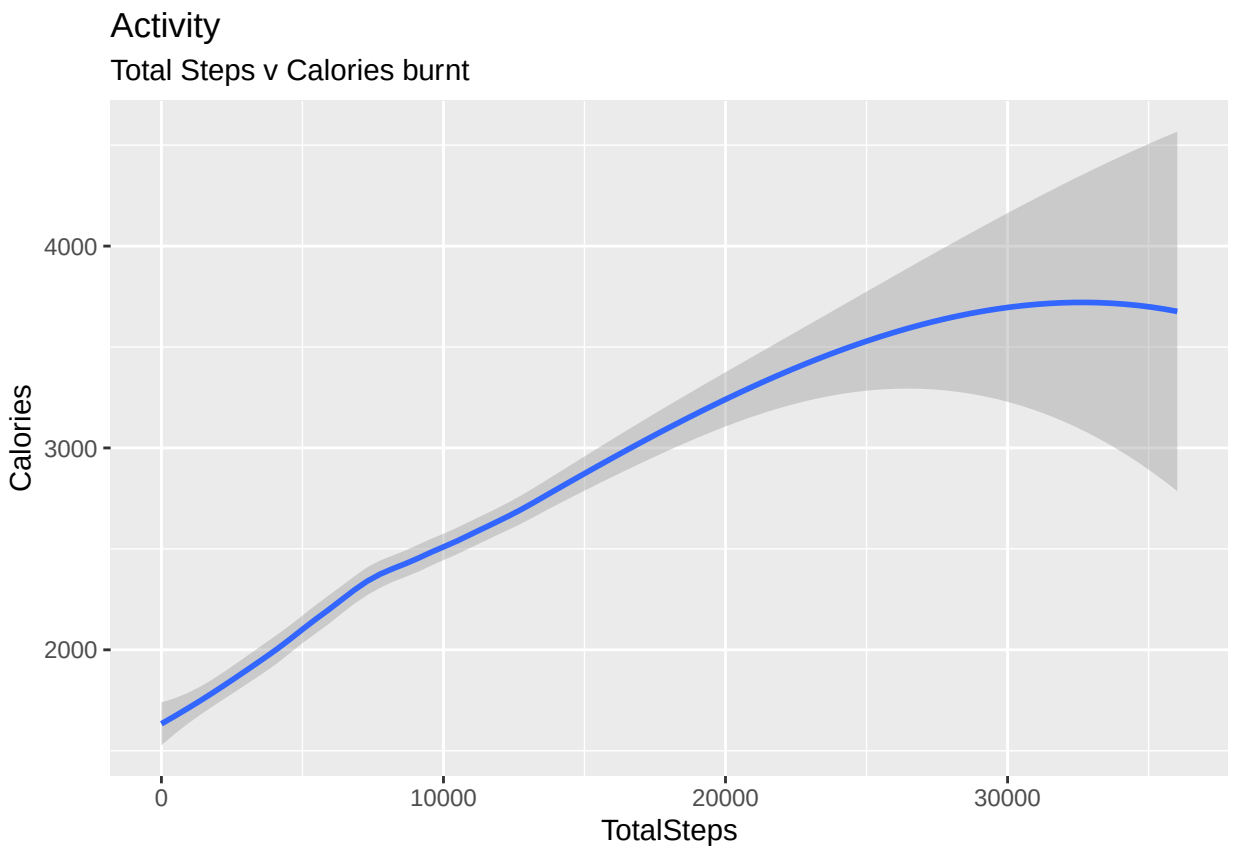
```
    WeightKg           BMI
 Min.   : 52.60   Min.   :21.45
 1st Qu.: 61.40   1st Qu.:23.96
 Median : 62.50   Median :24.39
 Mean   : 72.04   Mean   :25.19
 3rd Qu.: 85.05   3rd Qu.:25.56
 Max.   :133.50   Max.   :47.54
```

From the summary of this data we can say draw the following; *Most of the participants are lightly active, it reflects on the average number of calories burnt.* The mean sedentary minutes at 662 needs to be reduced *The mean Total Intensity is quite low.* The mean hours of sleep is in line withe CDC recommendation of 7 hours and above. * Mean number of steps however at 7638 is below the CDC recommended number of 10000 steps per day.

Share Phase(Data Visualization)

Now we are going to visualize our data to help us better understand it using ggplot2 package. Firstly we are going to see how much activity correlates with calories burnt

```
library(ggplot2)
ggplot(data=Activity, aes(x=TotalSteps, y=Calories))+geom_smooth()+ labs(title="Activity", subtitle="To
```



As expected there is apositive relationship between Total Steps and Calories burnt.

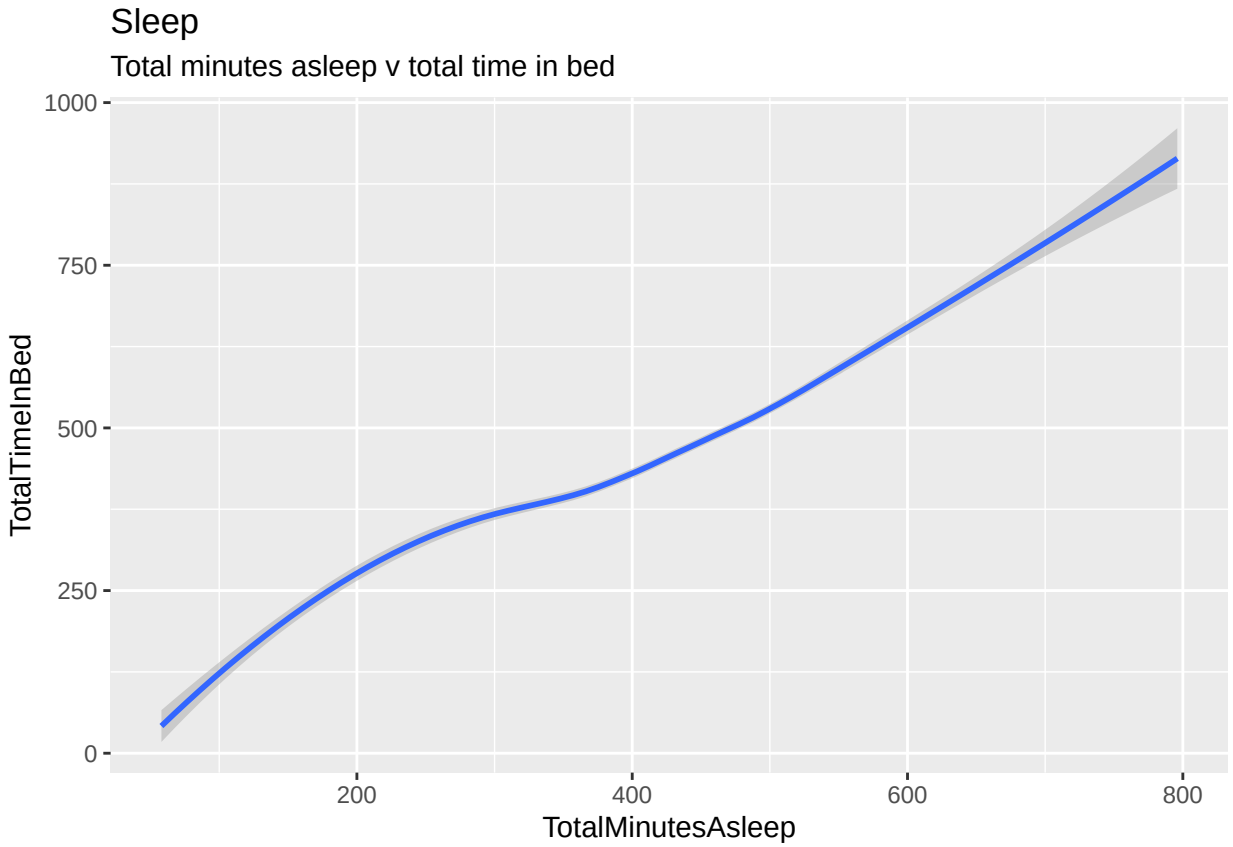
Now we will look at the relationship between Sedentary minutes and calories burnt

```
ggplot(data= Activity, aes(x = Calories, y =SedentaryMinutes))+geom_smooth()+ labs(title= 'Activity', su
```

I can see there's a negative relationship between Calories burnt and sedentary minutes, the marketing team should target reducing sedentary Minutes.

Let's also look at the relationship between Sleep and Total minutes in bed

```
ggplot(data=Sleep, aes(x=TotalMinutesAsleep, y = TotalTimeInBed))+ geom_smooth()+ labs(title = 'Sleep',
```



As we can see there's a clear correlation between minutes asleep and time spent in bed. I recommend that Bellabeat could remind its users to go bed early using notifications on the app as reminders.

Act Phase(Conclusion and Recommendations)

By collecting data on activity, health, sleep e.t.c the company bellabeat seeks to empower its users with information about their daily life and habits and how it affects their health. The firm has been doing this since 2013 and has grown rapidly since then. After analysibg this data, I have found some trends and have some recommendations for the bellabeat marketing team.

Target Audience Women who work full-time jobs (according to the hourly intensity data) and spend a lot of time at the computer/in a meeting/ focused on work they are doing (according to the sedentary time data).

They engage in some light activity to stay healthy (according to the activity type analysis). Although they still need to improve their everyday activity to really enjoy the benefits. Some motivation and expert assistance is needed, and bellabeat can offer that.

As there is no gender information about the participants, I assumed that all genders were presented and balanced in this data set.

Recommendation Bellabeat can be a guide to women who are in need of motivation and advice on how to maintain a healthy living in the midst of their professional lives and other hectic daily activities.

Ideas for bellabeat marketing team

- Average total steps per day are 7638 which a lower than the recommended 10000 according to the CDC research. They found that taking 8,000 steps per day was associated with a 51% lower risk for all-cause mortality (or death from all causes). Taking 12,000 steps per day was associated with a 65% lower risk compared with taking 4,000 steps. Bellabeat can set a target for it customers to take at least the recommended 10000 steps per day and send reminders at interval hours of the day.

- Regular notifications can be sent to users to take walks in between work hours to increase calories burnt and reduce sedentary time as the numbers are on the high side.
- There is a positive relationship between time spent in bed and minutes asleep, bedtime notifications can be sent to remind users of bedtime at a chosen time. This will help maintain consistency, increase hours and quality of sleep generally.
- Meal suggestions on low calorie food for users interested in weight loss can also be implemented.

That brings us to the end,thank you for your interest in my Bellabeat Case Study!

This is my first project using R. I would appreciate any comments and recommendations for improvement!