Bellabeat case study

Ashwini Shanbhogue

08/23/2021

The case study is presented in six parts, which are the milestones of this analysis, namely- Ask, prepare, process, analyze, share, and act.

1. Ask: A statement of the business task

The business task is to-

- analyze smart device usage data from non-Bellabeat smart devices to identify trends in device usage,
- select one Bellabeat product to apply these insights to and to come up with high-level recommendations for how these trends can inform Bellabeat marketing strategy and
- present the findings to the stakeholders- **Urška Sršen**: Bellabeat's cofounder and Chief Creative Officer, **Sando Mur**: Mathematician and Bellabeat's cofounder; key member of the Bellabeat executive team, and the **Bellabeat marketing analytics team**: A team of data analysts responsible for collecting, analyzing, and reporting data that helps guide Bellabeat's marketing strategy.

2. Prepare: A description of all data sources used

The data source is a public dataset, **FitBit Fitness Tracker Data** (CC0: Public Domain, made available through Mobius). This easily accesssible Kaggle data set contains personal fitness tracker data from thirty Fitbit users organized in 18 files with data in both long and wide formats. Thirty eligible Fitbit users consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. It includes information about daily activity, steps, and heart rate that can be used to explore users' habits. Individual entries are identified only by an ID number to protect the privacy of the users.

The dataset was downloaded and saved on a personal computer in a folder named 'Bellabeat case study' along with the file containing the brief for the analysis. The folder containing the 18 files in the dataset was renamed to '4-12-16-to-5-12-16_Fitabase-Data' to make it easier to organize and read by machines and humans.

ROCCC analysis of the data:

- Reliable: The data is not reliable because it is incomplete (weight and sleep data is not available for all IDs) and suffers from sample bias (very small dataset containing user information for only a month)
- Original: The data is not original. The data was collected via a survey by Amazon MTurk (second party), probably on the request of a client and shared by a Kaggle user, Mobius (third party)
- Comprehensive: The data is not comprehensive (very small dataset collected over a limited period of time. The survey collected data between 03.12.2016 and 05.12.2016 but the Kaggle dataset made available contains data from between 04.12.2016 and 05.12.2016 only)
- Current: The data is not current (it was created and last updated in 2016).
- Cited: It has been cited once (Torre, I., Sanchez, O., Koceva, F. *et al.* Supporting users to take informed decisions on privacy settings of personal devices. *Pers Ubiquit Comput* **22,** 345–364 (2018). https://doi.org/10.1007/s00779-017-1068-3).

Overall, the data does not ROCCC.

Although the dataset does not ROCCC, the kind of data available within it, i.e **time series data** should help answer the question being asked- 'what are the trends contained within fitness tracker usage data?'

3. Process: Documentation of any cleaning or manipulation of data

   R will be the data tool of choice for processing and analyzing data for its ability to both process large datasets quickly and create attractive visualizations.

   # Loading the R packages that will be used to process the dataset

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.4     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(skimr)
library(janitor)
```

```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(ggpubr)
```

Creating dataframes to be analyzed

```
Activity <- read_csv("dailyActivity_merged.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   Id = col_double(),
##   ActivityDate = col_character(),
##   TotalSteps = col_double(),
##   TotalDistance = col_double(),
##   TrackerDistance = col_double(),
##   LoggedActivitiesDistance = col_double(),
##   VeryActiveDistance = col_double(),
##   ModeratelyActiveDistance = col_double(),
##   LightActiveDistance = col_double(),
##   SedentaryActiveDistance = col_double(),
##   VeryActiveMinutes = col_double(),
##   FairlyActiveMinutes = col_double(),
##   LightlyActiveMinutes = col_double(),
##   SedentaryMinutes = col_double(),
##   Calories = col_double()
## )
```

```
Sleep <- read_csv("sleepDay_merged.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   Id = col_double(),
##   SleepDay = col_character(),
##   TotalSleepRecords = col_double(),
##   TotalMinutesAsleep = col_double(),
##   TotalTimeInBed = col_double()
## )
```

```
Weight <- read_csv("weightLogInfo_merged.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   Id = col_double(),
```

```
##   Date = col_character(),
##   WeightKg = col_double(),
##   WeightPounds = col_double(),
##   Fat = col_double(),
##   BMI = col_double(),
##   IsManualReport = col_logical(),
##   LogId = col_double()
## )
```

Checking for duplicate or missing values in all the dataframes

```
anyDuplicated(Activity)
```

```
## [1] 0
```

```
anyDuplicated(Sleep)
```

```
## [1] 162
```

```
anyDuplicated(Weight)
```

```
## [1] 0
```

```
any(is.na(Activity))
```

```
## [1] FALSE
```

```
any(is.na(Sleep))
```

```
## [1] FALSE
```

```
any(is.na(Weight))
```

```
## [1] TRUE
```

None of the dataframes have any duplicate values except ‘Sleep’ which has 162 and none of the dataframes have missing values except Weight.

Eliminating duplicate records from ‘Sleep’ and missing values from Weight and checking that they were indeed eliminated.

```
Sleep <- distinct(Sleep)
anyDuplicated(Sleep)
```

```
## [1] 0
```

```
Weight <- na.omit(Weight)
any(is.na(Weight))
```

```
## [1] FALSE
```

All duplicate and missing values have been eliminated.

Exploring the Activity, Sleep and Weight dataframes

```
skim_without_charts(Activity)
```

Data summary

Name	Activity
Number of rows	940

Number of columns	15
Column type frequency:	
character	1
numeric	14
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
ActivityDate	0	1	8	9	0	31	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
Id	0	1	4.855407e+09	2.424805e+09	1503960366	2.320127e+09	4.445115e+09	6.962181e+09	8.877689e+09
TotalSteps	0	1	7.637910e+03	5.087150e+03	0	3.789750e+03	7.405500e+03	1.072700e+04	3.601900e+04
TotalDistance	0	1	5.490000e+00	3.920000e+00	0	2.620000e+00	5.240000e+00	7.710000e+00	2.803000e+01
TrackerDistance	0	1	5.480000e+00	3.910000e+00	0	2.620000e+00	5.240000e+00	7.710000e+00	2.803000e+01
LoggedActivitiesDistance	0	1	1.100000e-01	6.200000e-01	0	0.000000e+00	0.000000e+00	0.000000e+00	4.940000e+00
VeryActiveDistance	0	1	1.500000e+00	2.660000e+00	0	0.000000e+00	2.100000e-01	2.050000e+00	2.192000e+01
ModeratelyActiveDistance	0	1	5.700000e-01	8.800000e-01	0	0.000000e+00	2.400000e-01	8.000000e-01	6.480000e+00
LightActiveDistance	0	1	3.340000e+00	2.040000e+00	0	1.950000e+00	3.360000e+00	4.780000e+00	1.071000e+01
SedentaryActiveDistance	0	1	0.000000e+00	1.000000e-02	0	0.000000e+00	0.000000e+00	0.000000e+00	1.100000e-01
VeryActiveMinutes	0	1	2.116000e+01	3.284000e+01	0	0.000000e+00	4.000000e+00	3.200000e+01	2.100000e+02
FairlyActiveMinutes	0	1	1.356000e+01	1.999000e+01	0	0.000000e+00	6.000000e+00	1.900000e+01	1.430000e+02
LightlyActiveMinutes	0	1	1.928100e+02	1.091700e+02	0	1.270000e+02	1.990000e+02	2.640000e+02	5.180000e+02
SedentaryMinutes	0	1	9.912100e+02	3.012700e+02	0	7.297500e+02	1.057500e+03	1.229500e+03	1.440000e+03
Calories	0	1	2.303610e+03	7.181700e+02	0	1.828500e+03	2.134000e+03	2.793250e+03	4.900000e+03

```
skim_without_charts(Sleep)
```

Data summary

Name	Sleep
Number of rows	410
Number of columns	5
Column type frequency:	
character	1
numeric	4
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
SleepDay	0	1	20	21	0	31	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
Id	0	1	4.994963e+09	2.060863e+09	1503960366	3.977334e+09	4702921684.0	6962181067879200	9665
TotalSleepRecords	0	1	1.120000e+00	3.500000e-01	1	1.000000e+00	1.0	1	3
TotalMinutesAsleep	0	1	4.191700e+02	1.186400e+02	58	3.610000e+02	432.5	490	796
TotalTimeInBed	0	1	4.584800e+02	1.274600e+02	61	4.037500e+02	463.0	526	961

```
skim_without_charts(Weight)
```

Data summary

Name	Weight
Number of rows	2
Number of columns	8
Column type frequency:	
character	1
logical	1
numeric	6

Group variables None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Date	0	1	20	21	0	2	0

Variable type: logical

skim_variable	n_missing	complete_rate	mean	count
IsManualReport	0	1	1	TRU: 2

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
Id	0	1	2.911832e+09	1.991031e+09	1.503960e+09	2.207896e+09	2.911832e+09	3.615768e+09	4.319704e+09
WeightKg	0	1	6.250000e+01	1.400000e+01	5.260000e+01	5.755000e+01	6.250000e+01	6.745000e+01	7.240000e+01
WeightPounds	0	1	1.377900e+02	3.087000e+01	1.159600e+02	1.268800e+02	1.377900e+02	1.487000e+02	1.596100e+02
Fat	0	1	2.350000e+01	2.120000e+00	2.200000e+01	2.275000e+01	2.350000e+01	2.425000e+01	2.500000e+01
BMI	0	1	2.505000e+01	3.390000e+00	2.265000e+01	2.385000e+01	2.505000e+01	2.625000e+01	2.745000e+01
LogId	0	1	1.461586e+12	9.164104e+08	1.460938e+12	1.461262e+12	1.461586e+12	1.461910e+12	1.462234e+12

Manipulating the Activity dataframe by-

dropping three columns (TotalDistance, TrackerDistance, and LoggedActivitiesDistance) and creating four new columns, VeryActive, Moderate, LightlyActive and Sedentary from existing columns. Summary function is used to check the summary statistics of all columns.

```
Activity <- Activity %>% select(-TotalDistance, -TrackerDistance, -LoggedActivitiesDistance) %>% mutate (Ver
yActive= mean(VeryActiveMinutes), Moderate= mean(FairlyActiveMinutes), LightlyActive= mean(LightlyActiveMinu
tes), Sedentary= mean(SedentaryMinutes))

summary(Activity)
```

```
##        Id            ActivityDate         TotalSteps    VeryActiveDistance
##  Min.   :1.504e+09   Length:940         Min.   :    0   Min.   : 0.000
##  1st Qu.:2.320e+09   Class :character   1st Qu.: 3790   1st Qu.: 0.000
##  Median :4.445e+09   Mode  :character   Median : 7406   Median : 0.210
##  Mean   :4.855e+09                      Mean   : 7638   Mean   : 1.503
##  3rd Qu.:6.962e+09                      3rd Qu.:10727   3rd Qu.: 2.053
##  Max.   :8.878e+09                      Max.   :36019   Max.   :21.920
##  ModeratelyActiveDistance LightActiveDistance SedentaryActiveDistance
##  Min.   :0.0000           Min.   : 0.000      Min.   :0.000000
##  1st Qu.:0.0000           1st Qu.: 1.945      1st Qu.:0.000000
##  Median :0.2400           Median : 3.365      Median :0.000000
##  Mean   :0.5675           Mean   : 3.341      Mean   :0.001606
##  3rd Qu.:0.8000           3rd Qu.: 4.782      3rd Qu.:0.000000
##  Max.   :6.4800           Max.   :10.710      Max.   :0.110000
##  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
##     Calories      VeryActive       Moderate     LightlyActive     Sedentary
##  Min.   :   0   Min.   :21.16   Min.   :13.56   Min.   :192.8   Min.   :991.2
##  1st Qu.:1828   1st Qu.:21.16   1st Qu.:13.56   1st Qu.:192.8   1st Qu.:991.2
##  Median :2134   Median :21.16   Median :13.56   Median :192.8   Median :991.2
##  Mean   :2304   Mean   :21.16   Mean   :13.56   Mean   :192.8   Mean   :991.2
##  3rd Qu.:2793   3rd Qu.:21.16   3rd Qu.:13.56   3rd Qu.:192.8   3rd Qu.:991.2
##  Max.   :4900   Max.   :21.16   Max.   :13.56   Max.   :192.8   Max.   :991.2
```

Combining the Activity and Sleep dataframes

```
Merged1 <- merge(Activity, Sleep, by="Id")
```

4. Analyze: A summary of the data analysis

How many unique users have documented their daily activities, sleep and weight?

Calculating the number of unique participants in each dataframe

```
n_distinct(Activity$Id)
```

```
## [1] 33
```

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

```
## [1] 24
```

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

```
## [1] 2
```

The Activity, Sleep and Weight dataframes have 33, 24 and 2 unique participants respectively. Fewer users have documented their sleep and weight than those that have documented their daily activity.

Are fewer users documenting their sleep because they take their device off their person while sleeping? How can the users be encouraged to keep it on?

Are fewer users documenting their weight because it needs to be done manually and they may not remember to do so every day? How can the users be encouraged to document their weight every day?

How active is the average Bellabeat user?

Calculating the activity level of the average Bellabeat user

```
w <- (Activity$VeryActive/60)*100/24
x <- (Activity$Moderate/60)*100/24
y <- (Activity$LightlyActive/60)*100/24
z <- (Activity$Sedentary/60)*100/24
a <- (100-(w+x+y+z))

paste("The average Bellabeat user spends", round(w[1:1], digits=1), "% of their day being very active, ", ro
und(x[1:1], digits=1), "% of their day being moderately active, ", round(y[1:1], digits=1), "% of their day
being lightly active, and", round(z[1:1], digits=1), "% of their day being sedentary. Activity levels are un
known for", round(a[1:1], digits=1), "% of the day.")
```

```
## [1] "The average Bellabeat user spends 1.5 % of their day being very active,  0.9 % of their day being mo
derately active,  13.4 % of their day being lightly active, and 68.8 % of their day being sedentary. Activit
y levels are unknown for 15.4 % of the day."
```

The average Bellabeat user is sedentary for most of the day and is only lightly active, when active. How can the average user be encouraged to be more active through the day?

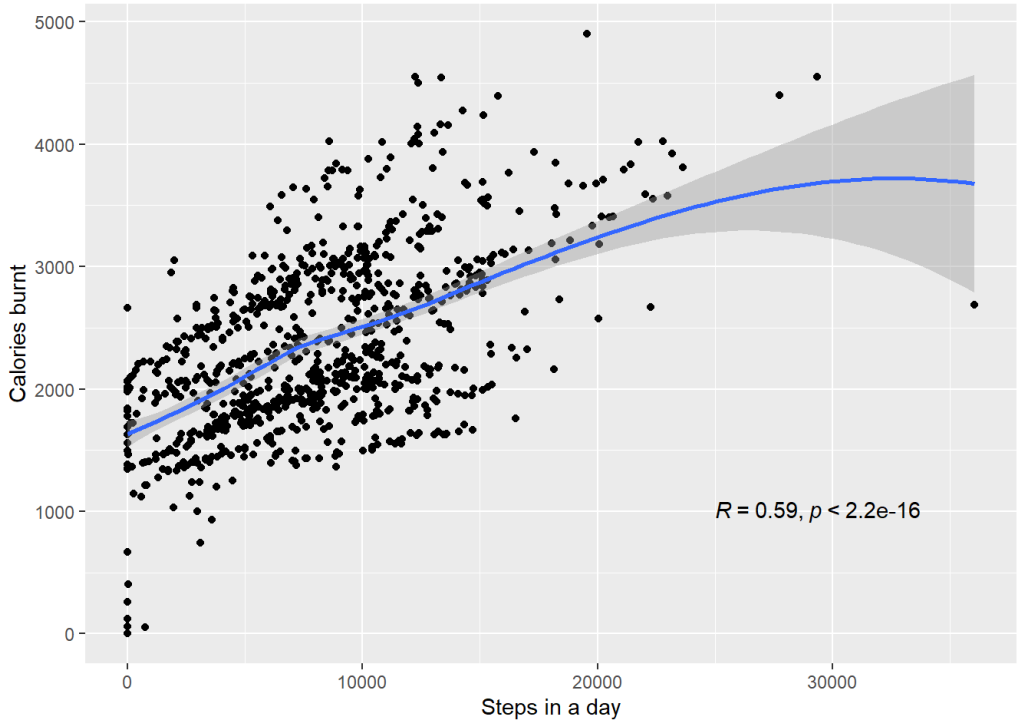
Activity levels for the average user are unknown for about 15% of their day. This is likely due to the user taking their Bellabeat device off their person for that duration. How can the user be encouraged to keep the device on longer?

Does taking more steps in a day burn more calories?

Creating a scatter plot of total steps taken in a day against the amount of calories burnt

```
ggplot(data=Activity, mapping = aes(x=TotalSteps, y=Calories)) + geom_point() + geom_smooth() + labs(x = "St
eps in a day", y = "Calories burnt") + stat_cor(method = "pearson", label.x = 25000, label.y = 1000)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

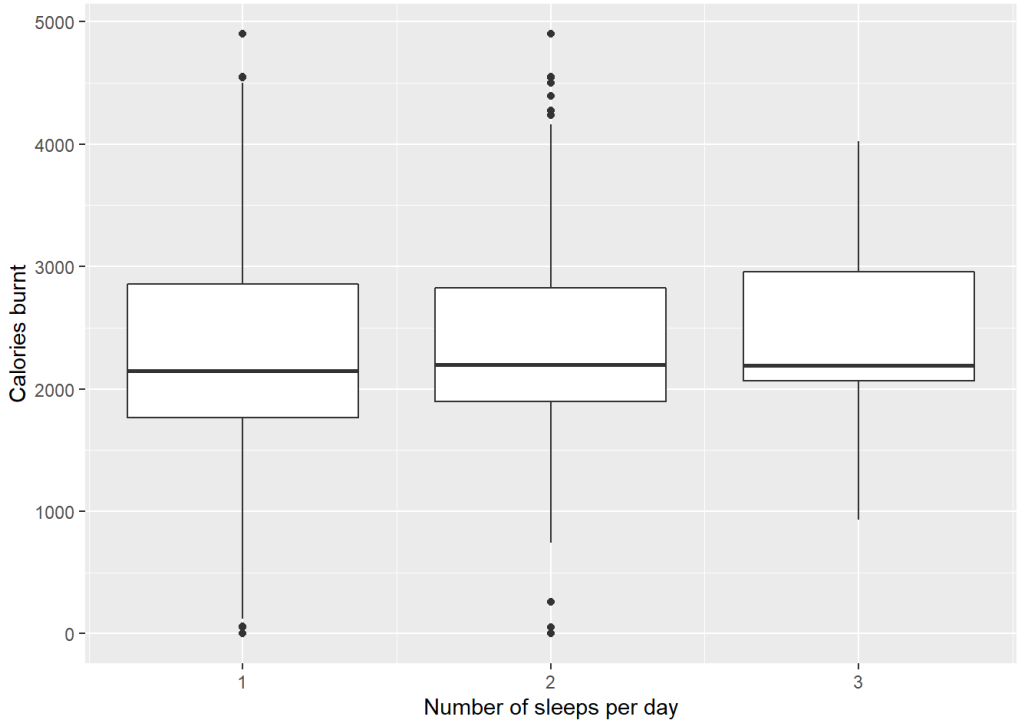


There is a positive correlation between steps taken in a day and calories burnt, ie. as the number is steps taken in a day increase, amount of calories burnt also increase, as evidenced by the upward trending regression line and the positive 'R' value (correlation coefficient). The dip in the regression line towards the end is because of an outlier.

Does number of times you sleep in a day affect the amount of calories burnt?

Creating a box plot of total sleep records per day against the amount of calories burnt

```
ggplot(data=Merged1, mapping = aes(group=TotalSleepRecords, x=TotalSleepRecords, y=Calories)) + geom_boxplot
() + labs(x = "Number of sleeps per day", y = "Calories burnt")
```



It appears that the number of times the user sleeps per day does not have any effect on calories burnt.

5. Share: Supporting visualizations and key findings
    1. Fewer users are documenting their sleep and weight compared to those that are documenting their daily activity.

2. The average Bellabeat user is sedentary for most of the day and is only lightly active, when active.
3. Activity levels for the average user are unknown for about 15% of their day.
4. The amount of calories burnt by the user in a day increases proportionally with the number of steps they take in a day.

6. Act: Top high-level insights based on the analysis

    1. Make Bellabeat wearables more comfortable to encourage users to keep them on while sleeping.
    2. Use the Bellabeat app and wearables to send reminders to document weight and to stay active through the day.
    3. Use the Bellabeat app and wearables to ‘celebrate’ with a special audio tone, visuals or vibration when the user hits a certain ‘steps’ goal.

Marketing strategy: Bellabeat wearables are that trusted friend that will stay with you throughout the day, gently encourage you to be your healthiest self and will celebrate with you when you make progress in the right direction.