Merit America
Google Analytics Certificate
Capstone Project : Case study 2
Bellabeats

Author: Jenise Chester
Date: June 22, 2024
Email: jenise83@gmail.com
LinkedIn
GitHub

Introduction:

In this case study I will be analyzing the data from Bellabeat. Bellabeat is a smart device company of health-focused products for women. I analyzed the data in order to answer key questions and give my analysis based on my findings.

Ask:

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 strategies?

Prepare:
I uploaded and analyzed data from the FitBit Fitness Tracker

- There were 18 datasets generated by respondents to a distributed survey via Amazon Mechanical Turk between the dates of 03/12/2016 - 05/12/2016.
- 30 FitBit users consent to the submission of the personal tracker data, which includes information about daily activity, steps, heart rate, and sleep monitoring that can be used to explore user habits.

Process:
Imported Data:

- `sleep <- read_csv("Downloads/bellafit_capstone_data_05_17_23/sleepDay_merged.csv")`
- `activity <- read_csv("Downloads/bellafit_capstone_data_05_17_23/dailyActivity_merged.csv")`

`head(activity)`

```
## # A tibble: 6 × 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 × 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
```

I checked the formatting,checked for unique name IDs, cleaned columnames, and overall checking to make sure the data was cleaning to begin the process of analyzing the data. I will give examples below on the programming language I used in this process.

n_distinct(activity$Id)

There were 33 unique name IDs for activity

n_distinct(sleep$Id)

There were 24 unique name IDs for sleep

clean_names(sleep)

```
## # A tibble: 410 × 5
##          id sleep_day total_sleep_records total_minutes_asleep total_time_in_bed
##       <dbl> <chr>                   <dbl>                <dbl>             <dbl>
##  1 1.50e9 4/12/201…                     1                  327               346
##  2 1.50e9 4/13/201…                     2                  384               407
##  3 1.50e9 4/15/201…                     1                  412               442
##  4 1.50e9 4/16/201…                     2                  340               367
##  5 1.50e9 4/17/201…                     1                  700               712
##  6 1.50e9 4/19/201…                     1                  304               320
##  7 1.50e9 4/20/201…                     1                  360               377
```

```
##  8   1.50e9 4/21/201…                1              325              364
##  9   1.50e9 4/23/201…                1              361              384
## 10   1.50e9 4/24/201…                1              430              449
## # i 400 more rows
```

clean_names(activity)

```
## # A tibble: 940 × 15
##           id activity_date total_steps total_distance tracker_distance
##        <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
##  7 1503960366 4/18/2016           13019           8.59             8.59
##  8 1503960366 4/19/2016           15506           9.88             9.88
##  9 1503960366 4/20/2016           10544           6.68             6.68
## 10 1503960366 4/21/2016            9819           6.34             6.34
## # i 930 more rows
## # i 10 more variables: logged_activities_distance <dbl>,
## #   very_active_distance <dbl>, moderately_active_distance <dbl>,
## #   light_active_distance <dbl>, sedentary_active_distance <dbl>,
## #   very_active_minutes <dbl>, fairly_active_minutes <dbl>,
## #   lightly_active_minutes <dbl>, sedentary_minutes <dbl>, calories <dbl>
```

I went through the name columns to make sure there were no duplicates, missing data, and data was parallel to one another.

str(activity)

```
## spc_tbl_ [940 × 15] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Id                       : num [1:940] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ ActivityDate             : chr [1:940] "4/12/2016" "4/13/2016" "4/14/2016" "4/15/2016" ...
```

```
##  $ TotalSteps              : num [1:940] 13162 10735 10460 9762 12669 ...
##  $ TotalDistance           : num [1:940] 8.5 6.97 6.74 6.28 8.16 ...
##  $ TrackerDistance         : num [1:940] 8.5 6.97 6.74 6.28 8.16 ...
##  $ LoggedActivitiesDistance: num [1:940] 0 0 0 0 0 0 0 0 0 0 ...
##  $ VeryActiveDistance      : num [1:940] 1.88 1.57 2.44 2.14 2.71 ...
##  $ ModeratelyActiveDistance: num [1:940] 0.55 0.69 0.4 1.26 0.41 ...
##  $ LightActiveDistance     : num [1:940] 6.06 4.71 3.91 2.83 5.04 ...
##  $ SedentaryActiveDistance : num [1:940] 0 0 0 0 0 0 0 0 0 0 ...
##  $ VeryActiveMinutes       : num [1:940] 25 21 30 29 36 38 42 50 28 19 ...
##  $ FairlyActiveMinutes     : num [1:940] 13 19 11 34 10 20 16 31 12 8 ...
##  $ LightlyActiveMinutes    : num [1:940] 328 217 181 209 221 164 233 264 205 211 ...
##  $ SedentaryMinutes        : num [1:940] 728 776 1218 726 773 ...
##  $ Calories                : num [1:940] 1985 1797 1776 1745 1863 ...
##  - attr(*, "spec")=
##   .. 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()
##   .. )
```

```
##  - attr(*, "problems")=<externalptr>
```

```
str(sleep)
```

```
## tibble [410 × 5] (S3: tbl_df/tbl/data.frame)
##  $ Id                : num [1:410] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ SleepDay          : chr [1:410] "4/12/2016 12:00:00 AM" "4/13/2016 12:00:00 AM" "4/15/2016 12:00:00 AM" "4/16/2016 12:00:00
AM" ...
##  $ TotalSleepRecords : num [1:410] 1 2 1 2 1 1 1 1 1 1 ...
##  $ TotalMinutesAsleep: num [1:410] 327 384 412 340 700 304 360 325 361 430 ...
##  $ TotalTimeInBed    : num [1:410] 346 407 442 367 712 320 377 364 384 449 …
```

I see that the datetime format for ActivityDate and SleepDay needs to be converted. I did that as well.

```
## tibble [410 × 5] (S3: tbl_df/tbl/data.frame)

##  $ Id                : num [1:410] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...

##  $ SleepDay          : chr [1:410] "4/12/2016 12:00:00 AM" "4/13/2016 12:00:00 AM" "4/15/2016 12:00:00 AM" "4/16/2016 12:00:00
AM" ...

##  $ TotalSleepRecords : num [1:410] 1 2 1 2 1 1 1 1 1 1 ...

##  $ TotalMinutesAsleep: num [1:410] 327 384 412 340 700 304 360 325 361 430 ...

##  $ TotalTimeInBed    : num [1:410] 346 407 442 367 712 320 377 364 384 449 …
```

Analyze:

```
activity %>%

  select(TotalSteps, TotalDistance, Calories) %>%

  summary()
```

```
##    TotalSteps    TotalDistance      Calories

##  Min.   :    0   Min.   : 0.000   Min.   :   0

##  1st Qu.: 3790   1st Qu.: 2.620   1st Qu.:1828

##  Median : 7406   Median : 5.245   Median :2134

##  Mean   : 7638   Mean   : 5.490   Mean   :2304

##  3rd Qu.:10727   3rd Qu.: 7.713   3rd Qu.:2793

##  Max.   :36019   Max.   :28.030   Max.   :4900
```

We see here that the average steps per day under “TotalSteps” are 7,638 steps and the average calories burned per day under “Calories” are 2,304. The recommended steps per day by the

CDC are 10,000, the CDC states that there are still benefits to walking less that 10,000 steps per day as shown above.

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

```
##  TotalSleepRecords TotalTimeInBed  TotalMinutesAsleep
##  Min.   :1.00      Min.   : 61.0   Min.   : 58.0
##  1st Qu.:1.00      1st Qu.:403.8   1st Qu.:361.0
##  Median :1.00      Median :463.0   Median :432.5
##  Mean   :1.12      Mean   :458.5   Mean   :419.2
##  3rd Qu.:1.00      3rd Qu.:526.0   3rd Qu.:490.0
##  Max.   :3.00      Max.   :961.0   Max.   :796.0
```

The average time users spend in bed is 458.5 minutes, which is 7.6 hrs., but are only asleep for 419.2 minutes which is 7 hrs. Now based on the CDC Sleep Recommendation 7 hrs or more is the recommended average hours of sleep per night.

The user is getting the minimal average sleep that is recommended however, a suggested sleep schedule including a wind down period reminder could be great to incorporate to get an increase on the sleep average.

```
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
```

There are 4 categories of activity minutes:

Categories	Averages Minutes
Very active	21.16
Fairly active	13.56
Lightly active	192.8
sedentary	991.2

It appears that the users spend an average of 991.2 (16.5 hrs) sitting/inactive which is a lot of idle time compared to the other active times combined. This type of time not being utilized can increase the chances of health issues such as cardiovascular disease. Check out MayoClinic
7 benefits of regular physical activity.

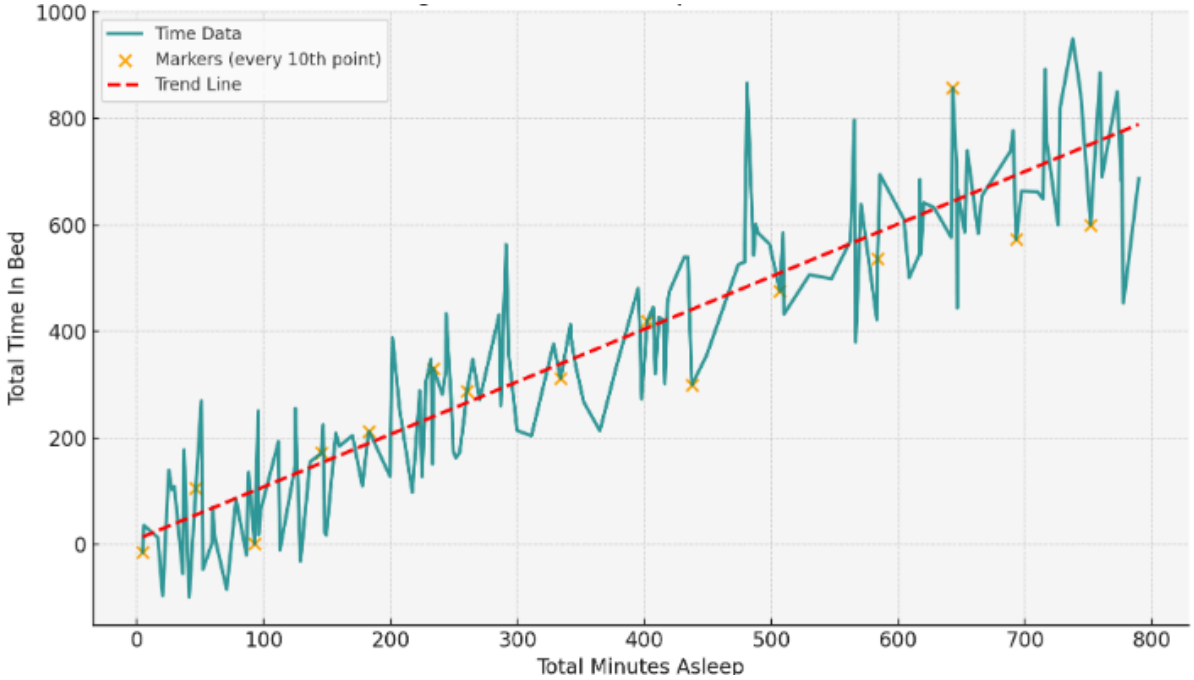
Incorporating mini 3-5 minute activity goals along with a weekly workout plan that start the users at lightly active with the goal of becoming very active could be a good asset in increasing the active average minutes.

```
activity_sleep <- merge(activity, sleep, by= c("Id","date"), all.x = TRUE)
```

I'm merging data so that I can display a better visualization

Share:

```
ggplot(data=sleep, aes(x=TotalMinutesAsleep, y=TotalTimeInBed)) +
  geom_ribbon(aes(ymin=0, ymax=TotalTimeInBed), alpha=0.2, fill="blue", color="blue") +
  geom_point(color="red") +
  labs(title="Total Time Asleep vs. Total Time in Bed")
```



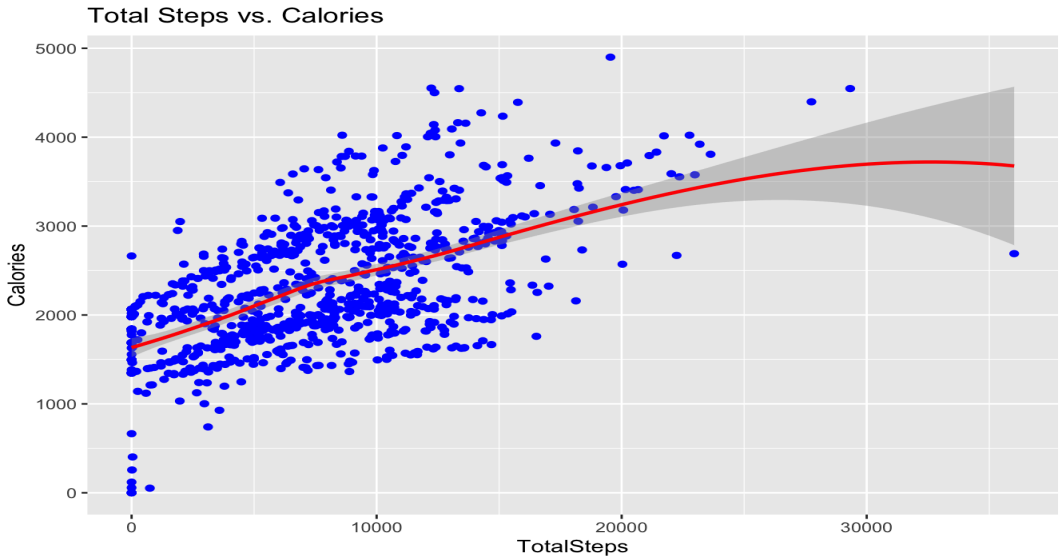
Total Time Asleep vs Total Time In Bed

This gives you a better visual of the time that is spent being in bed not getting sleep vs getting sleep which the time awake is more than the time sleeping giving us the average of 7 hrs that is the minimum required by CDC for sleep.

This reiterates the suggestion made to incorporate a sleep schedule including a wind down period reminder could be great to get an increase on the sleep average.

```
ggplot(data=activity, aes(x=TotalSteps, y=Calories)) +
  geom_point(color="blue") +
  geom_smooth(color="red") +
  labs(title="Total Steps vs. Calories")
```

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



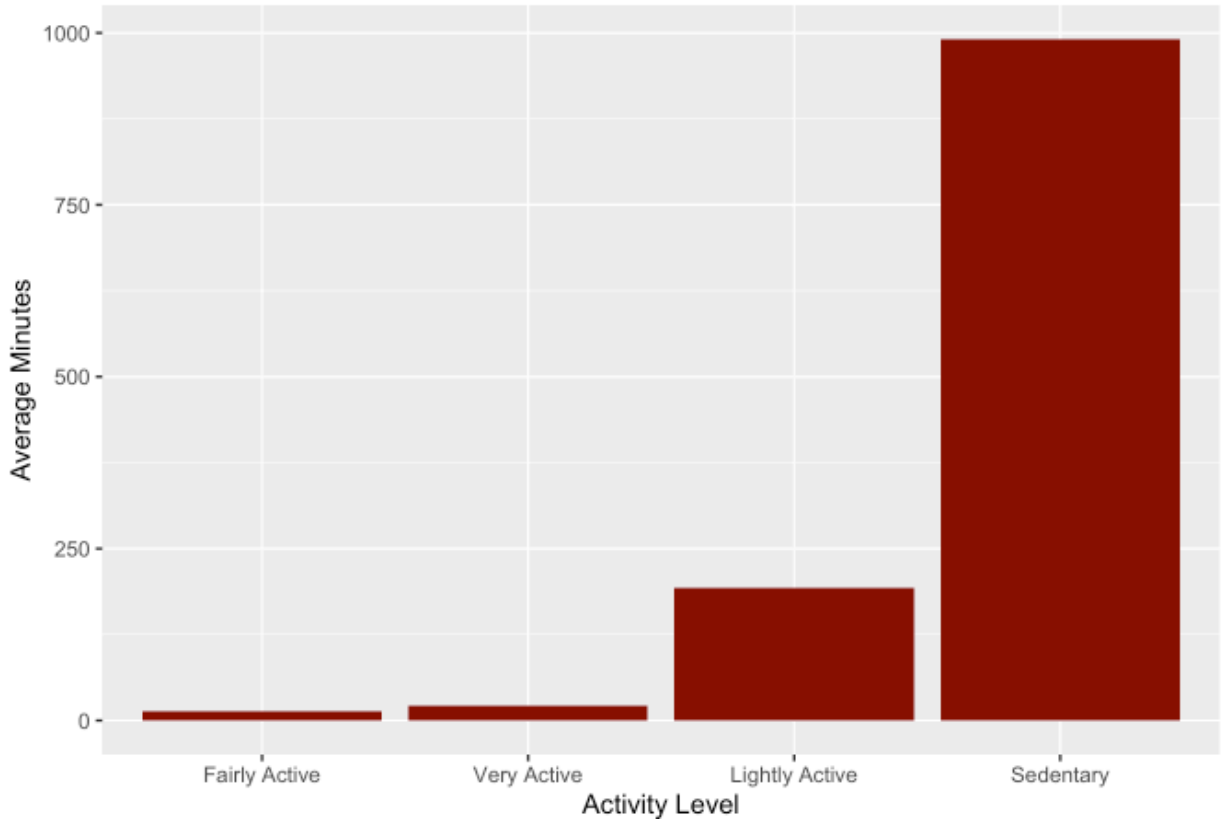
Total Steps vs Calories

This shows how taking more steps daily will cause more calories to burn which aligns with the CDC article and also implementing the daily mini 3-5 minute workout goals for users to get them to become more active on a regular basis to increase calorie burn.

Creating daily goals for users starting with lightly active activities like stretching, 15-20 minute walks, or yoga. To finally become very active like running, weight lifting, or swimming laps making a more active and healthier lifestyle.

```
averages <- data.frame(
  ActivityLevel = c("Very Active", "Fairly Active", "Lightly Active", "Sedentary"),
  AverageMinutes = c(21.2, 13.6, 193, 991))
```

```
ggplot(data = averages, aes(x = reorder(ActivityLevel, AverageMinutes), y = AverageMinutes)) +
    geom_col(fill = "darkred") +
    labs(title = "Average Minutes per Activity Level",
         x = "Activity Level",
         y = "Average Minutes")
```



Average Minutes per Activity Level

This chart also shows that the users are not as active as they can be. They are spending more time sitting/inactive and barely being lightly active which can be normal walking like to car, in the home, or grocery shopping etc.

This further shows that we need to implement more activities to get the users active more.

Act:

These are my findings:

1. The user's average sleep time is 7 hours which is the minimum time required according to the CDC. They spend 30-40 minutes in bed prior to sleeping as well.
2. The users are burning calories but not taking advantage of doing more steps daily in order to burn more calories.
3. The users are spending most of their time sedentary, not being active at all or doing minimal light activity.

Recommendations:

- Utilize free time to create goals in making a more active lifestyle by making workout times, workout dates with friends and little things around the house. I suggest that the app should create reminders and calendar inserts to help users organizer a plan to be actively successful.
- Creating an atmosphere for users to become a part of an community in the app the helps them feel more comfortable with making time to be active with success in making a healthier lifestyle and welcoming healthier habits.

Conclusion:

While there are so many ways we can explore to incorporate a more healthier active lifestyle for our users we have to have a starting point and my recommendations can help with that.

Thank You

This was an amazing journey and I Thank You for being along with it with me. There will be more to come soon