Bellabeat

Suketha

17/02/2022

Introduction and Goals

This is the capstone project for the Google Data Analytics Certification. For this case study, we are tasked with assisting a wearable fitness technology company, Bellabeat, improve their marketing strategies for their products by investigating customer activity with other fitness trackers like FitBit

Our goals are to look at datasets to find out:

- How are customers using other fitness trackers, like fitbit, in their daily life?
- What particular features seem to be the most heavily used?
- What features do Bellabeat products already have that consumers want, and how do we focus marketing on those aspects?
- What features should Bellabeat products consider adding to entice more customers?

Uploading Data

What data are we using?

The first dataset we'll be looking at comes from here: https://www.kaggle.com/arashnic/fitbit

There are a number of different csv files that range from Daily activity, calories, steps; hourly calories, intensities, and steps; and heart rate, sleep data and weight logs.A few immediate things come to mind when simply looking at the types of data collected by these 30 fitbit users.No water intake data has been collectedThese data may not actually assist me, but that will come with exploration.Data will be stored in our documents folder which will serve as our working directory for the project

Exploring the FitBit Data

We'll be using the tidyverse package as well as the skimr, here, and janitor packages for help with this project.

We're also using the sqldf package, which will allow us to emulate SQL syntax when looking at data

Loading the CSV files

Here we'll create our data frames for review. The data frames I'll be working with in this review will be creating objects for:

- daily_activity
- daily calories
- daily sleep
- weight log info
- daily intensities

We'll follow typical naming conventions based on the csv file names.

```
daily_activity <- read.csv("dailyActivity_merged.csv")
daily_calories <- read.csv("dailyCalories_merged.csv")
sleep_day <- read.csv("sleepDay_merged.csv")
daily_intensities <- read.csv("dailyIntensities_merged.csv")
weight_log <- read.csv("weightLogInfo_merged.csv")
```

Exploring our Tables

Let's take a beat to investigate the tables. For each one we'll look at the first six values using the head() function, and the column names with the colnames() function

Since we're looking at 5 different data frames, it might be a bit overwhelming to look at all of them at once, but it's critical to get a look at each of the tables now

daily_activity

```
head(daily_activity)
```

```
##           Id ActivityDate TotalSteps TotalDistance TrackerDistance
## 1 1503960366    4/12/2016      13162          8.50            8.50
## 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
##   LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance
## 1                        0              1.88                     0.55
## 2                        0              1.57                     0.69
## 3                        0              2.44                     0.40
## 4                        0              2.14                     1.26
## 5                        0              2.71                     0.41
## 6                        0              3.19                     0.78
##   LightActiveDistance SedentaryActiveDistance VeryActiveMinutes
## 1                6.06                       0                25
## 2                4.71                       0                21
```

```
## 3               3.91                        0               30
## 4               2.83                        0               29
## 5               5.04                        0               36
## 6               2.51                        0               38
##   FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories
## 1                  13                  328              728     1985
## 2                  19                  217              776     1797
## 3                  11                  181             1218     1776
## 4                  34                  209              726     1745
## 5                  10                  221              773     1863
## 6                  20                  164              539     1728
```

daily_calories

```
head(daily_calories)
```

```
##           Id ActivityDay Calories
## 1 1503960366   4/12/2016     1985
## 2 1503960366   4/13/2016     1797
## 3 1503960366   4/14/2016     1776
## 4 1503960366   4/15/2016     1745
## 5 1503960366   4/16/2016     1863
## 6 1503960366   4/17/2016     1728
```

sleep_day

```
head(sleep_day)
```

```
##           Id              SleepDay TotalSleepRecords TotalMinutesAsleep
## 1 1503960366 4/12/2016 12:00:00 AM                 1                327
## 2 1503960366 4/13/2016 12:00:00 AM                 2                384
## 3 1503960366 4/15/2016 12:00:00 AM                 1                412
## 4 1503960366 4/16/2016 12:00:00 AM                 2                340
## 5 1503960366 4/17/2016 12:00:00 AM                 1                700
## 6 1503960366 4/19/2016 12:00:00 AM                 1                304
##   TotalTimeInBed
## 1            346
## 2            407
## 3            442
## 4            367
## 5            712
## 6            320
```

daily_intensities

```
head(daily_intensities)
```

```
##           Id ActivityDay SedentaryMinutes LightlyActiveMinutes
## 1 1503960366   4/12/2016              728                  328
## 2 1503960366   4/13/2016              776                  217
## 3 1503960366   4/14/2016             1218                  181
## 4 1503960366   4/15/2016              726                  209
## 5 1503960366   4/16/2016              773                  221
## 6 1503960366   4/17/2016              539                  164
##   FairlyActiveMinutes VeryActiveMinutes SedentaryActiveDistance
## 1                  13                25                       0
## 2                  19                21                       0
## 3                  11                30                       0
## 4                  34                29                       0
## 5                  10                36                       0
## 6                  20                38                       0
##   LightActiveDistance ModeratelyActiveDistance VeryActiveDistance
## 1                6.06                     0.55               1.88
## 2                4.71                     0.69               1.57
## 3                3.91                     0.40               2.44
## 4                2.83                     1.26               2.14
## 5                5.04                     0.41               2.71
## 6                2.51                     0.78               3.19
```

weight_log

```
head(weight_log)
```

```
##           Id                   Date WeightKg WeightPounds Fat   BMI
## 1 1503960366  5/2/2016 11:59:59 PM     52.6     115.9631  22 22.65
## 2 1503960366  5/3/2016 11:59:59 PM     52.6     115.9631  NA 22.65
## 3 1927972279  4/13/2016 1:08:52 AM    133.5     294.3171  NA 47.54
## 4 2873212765 4/21/2016 11:59:59 PM     56.7     125.0021  NA 21.45
## 5 2873212765 5/12/2016 11:59:59 PM     57.3     126.3249  NA 21.69
## 6 4319703577 4/17/2016 11:59:59 PM     72.4     159.6147  25 27.45
##   IsManualReport        LogId
## 1           True 1.462234e+12
## 2           True 1.462320e+12
## 3          False 1.460510e+12
## 4           True 1.461283e+12
## 5           True 1.463098e+12
## 6           True 1.460938e+12
```

At a Glance

All 5 data frames have the same 'ID' field, so we can merge the datasets if need be.

It looks like the daily_activity, daily_calories, and daily_intensities have the exact same number of observations.

Furthermore, it seems the daily_activity table might have a log of calories and intensities already, so we should confirm that the values actually match for any given 'ID' number.

Let's use SQL syntax to see if there are any values in daily_calories that are in daily_activity... however, this won't work if the two dataframes have a different number of columsn, so we'll need to create a temporary dataframe where we select the important columns from daily_activity. Let's just call it "daily_activity2"

```
daily_activity2 <- daily_activity %>%
  select(Id, ActivityDate, Calories)
head(daily_activity2)
```

```
##           Id ActivityDate Calories
## 1 1503960366    4/12/2016     1985
## 2 1503960366    4/13/2016     1797
## 3 1503960366    4/14/2016     1776
## 4 1503960366    4/15/2016     1745
## 5 1503960366    4/16/2016     1863
## 6 1503960366    4/17/2016     1728
```

Great, now let's see what's the same between the two data frames of daily_activity2 and daily_calories

```
sql_check1 <- sqldf('SELECT *
                     FROM daily_activity2
                     INTERSECT
                     SELECT *
                     FROM daily_calories')
head(sql_check1)
```

```
##           Id ActivityDate Calories
## 1 1503960366    4/12/2016     1985
## 2 1503960366    4/13/2016     1797
## 3 1503960366    4/14/2016     1776
## 4 1503960366    4/15/2016     1745
## 5 1503960366    4/16/2016     1863
## 6 1503960366    4/17/2016     1728
```

```
nrow(sql_check1)
```

```
## [1] 940
```

It looks like there were 940 observations from the sql query, so it's safe to assume that the values are the same between the dataframes.

This leads us to assume the same is true for daily intensities, so we can drop those two dataframes from analysis... but just for completion sake, lets do the same check again :)

We will have to create another temporary data frame since daily_intensities only has 10 variables.

```
daily_activity3 <- daily_activity %>%
  select(Id, ActivityDate, SedentaryMinutes, LightlyActiveMinutes, FairlyActiveMinutes, VeryActiveMinut
         SedentaryActiveDistance, LightActiveDistance, ModeratelyActiveDistance, VeryActiveDistance)
head(daily_activity3)
```

```
##           Id ActivityDate SedentaryMinutes LightlyActiveMinutes
## 1 1503960366    4/12/2016              728                  328
## 2 1503960366    4/13/2016              776                  217
## 3 1503960366    4/14/2016             1218                  181
## 4 1503960366    4/15/2016              726                  209
```

```
## 5 1503960366    4/16/2016              773                  221
## 6 1503960366    4/17/2016              539                  164
##   FairlyActiveMinutes VeryActiveMinutes SedentaryActiveDistance
## 1                  13                25                       0
## 2                  19                21                       0
## 3                  11                30                       0
## 4                  34                29                       0
## 5                  10                36                       0
## 6                  20                38                       0
##   LightActiveDistance ModeratelyActiveDistance VeryActiveDistance
## 1                6.06                     0.55               1.88
## 2                4.71                     0.69               1.57
## 3                3.91                     0.40               2.44
## 4                2.83                     1.26               2.14
## 5                5.04                     0.41               2.71
## 6                2.51                     0.78               3.19
```

Great, now let's see what's the same between the two data frames of daily_activity2 and daily_calories

```
sql_check2 <- sqldf('SELECT * FROM daily_activity3 INTERSECT SELECT * FROM daily_intensities')
head(sql_check2)
```

```
##           Id ActivityDate SedentaryMinutes LightlyActiveMinutes
## 1 1503960366    4/12/2016              728                  328
## 2 1503960366    4/13/2016              776                  217
## 3 1503960366    4/14/2016             1218                  181
## 4 1503960366    4/15/2016              726                  209
## 5 1503960366    4/16/2016              773                  221
## 6 1503960366    4/17/2016              539                  164
##   FairlyActiveMinutes VeryActiveMinutes SedentaryActiveDistance
## 1                  13                25                       0
## 2                  19                21                       0
## 3                  11                30                       0
## 4                  34                29                       0
## 5                  10                36                       0
## 6                  20                38                       0
##   LightActiveDistance ModeratelyActiveDistance VeryActiveDistance
## 1                6.06                     0.55               1.88
## 2                4.71                     0.69               1.57
## 3                3.91                     0.40               2.44
## 4                2.83                     1.26               2.14
## 5                5.04                     0.41               2.71
## 6                2.51                     0.78               3.19
```

```
nrow(sql_check2)
```

```
## [1] 940
```

Looks like it's that magical 940 observations, so we can officially remove those two datasets from analysis Hooray for exploration!!

That leaves us with 3 data frames:

- daily_activity
- sleep_day
- weight_log

The Analysis

First, it looks like there may be more id's in the daily_activity dataframe over the sleep_day dataframe, and even more over the weight_log data frame, so let's use the n_distinct() function to find out

```
n_distinct(daily_activity$Id)
```

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

```
n_distinct(sleep_day$Id)
```

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

```
n_distinct(weight_log$Id)
```

```
## [1] 8
```

How many observations are there in each dataframe?

```
nrow(daily_activity)
```

```
## [1] 940
```

```
nrow(sleep_day)
```

```
## [1] 413
```

```
nrow(weight_log)
```

```
## [1] 67
```

What are some quick summary statistics we'd want to know about each data frame?

For the daily activity dataframe:

```
daily_activity %>%
  select(TotalSteps,
         TotalDistance,
         SedentaryMinutes,
         VeryActiveMinutes) %>%
  summary()
```

```
##    TotalSteps    TotalDistance    SedentaryMinutes VeryActiveMinutes
##  Min.   :    0   Min.   : 0.000   Min.   :   0.0   Min.   :  0.00
##  1st Qu.: 3790   1st Qu.: 2.620   1st Qu.: 729.8   1st Qu.:  0.00
##  Median : 7406   Median : 5.245   Median :1057.5   Median :  4.00
##  Mean   : 7638   Mean   : 5.490   Mean   : 991.2   Mean   : 21.16
##  3rd Qu.:10727   3rd Qu.: 7.713   3rd Qu.:1229.5   3rd Qu.: 32.00
##  Max.   :36019   Max.   :28.030   Max.   :1440.0   Max.   :210.00
```

For the sleep dataframe:

```
sleep_day %>%
  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
```

For the weight dataframe

```
weight_log %>%
  select(WeightPounds,
  BMI) %>%
  summary()
```

```
##   WeightPounds        BMI
##  Min.   :116.0   Min.   :21.45
##  1st Qu.:135.4   1st Qu.:23.96
##  Median :137.8   Median :24.39
##  Mean   :158.8   Mean   :25.19
##  3rd Qu.:187.5   3rd Qu.:25.56
##  Max.   :294.3   Max.   :47.54
```

Plotting a few explorations

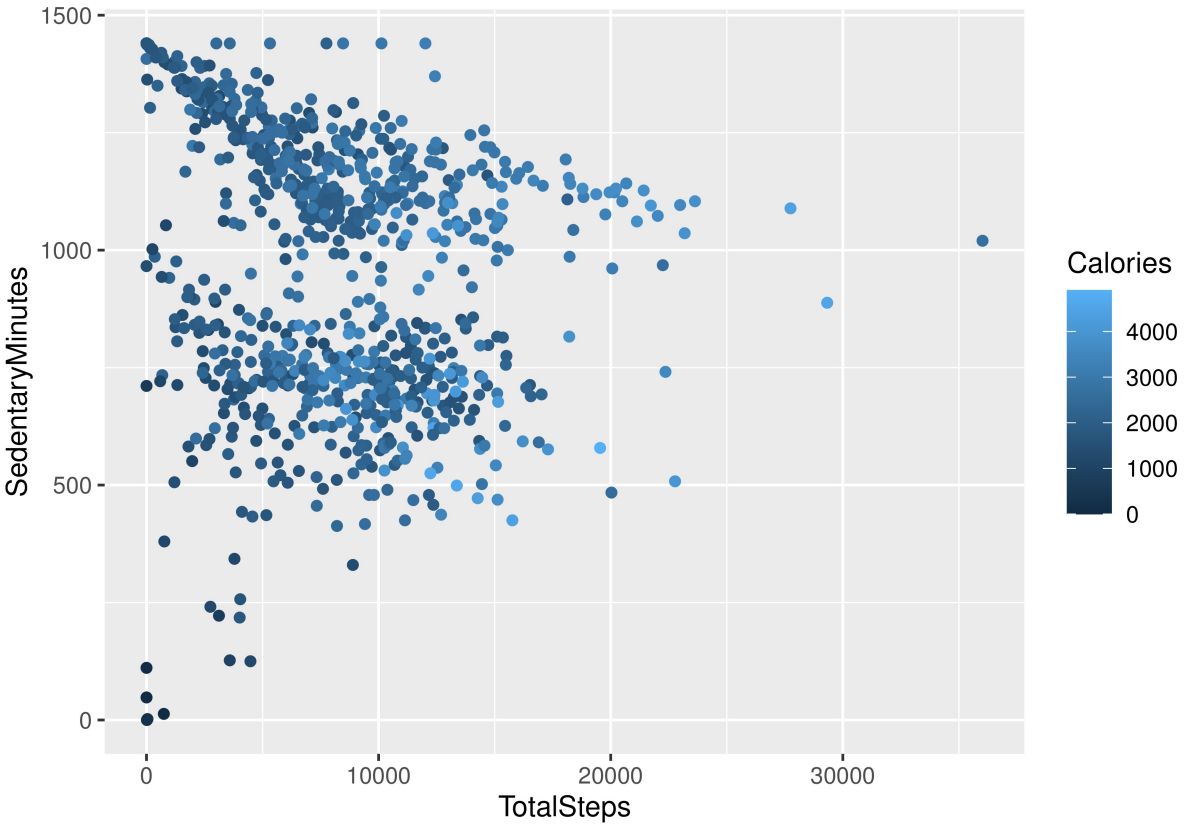
What's the relationship between steps taken in a day and sedentary minutes? It seems that we have a negative relationship between total steps taken and the minutes someone has remained sedentary. We also see that calories generally trend positively with total steps taking.

This shows that the data seem fairly accurate when it comes to recording steps and sedentary minutes. We could easily market this to consumers by telling them smart-devices could help them start their journey by measuring how much they're already moving!

You could also market the devices as a way to let people know how sedentary they actually are, and how active they could be. - not sure if I find that ethical, but hey, fear sells!

We can also note that sedentary time is not necessarily related to calories burned.

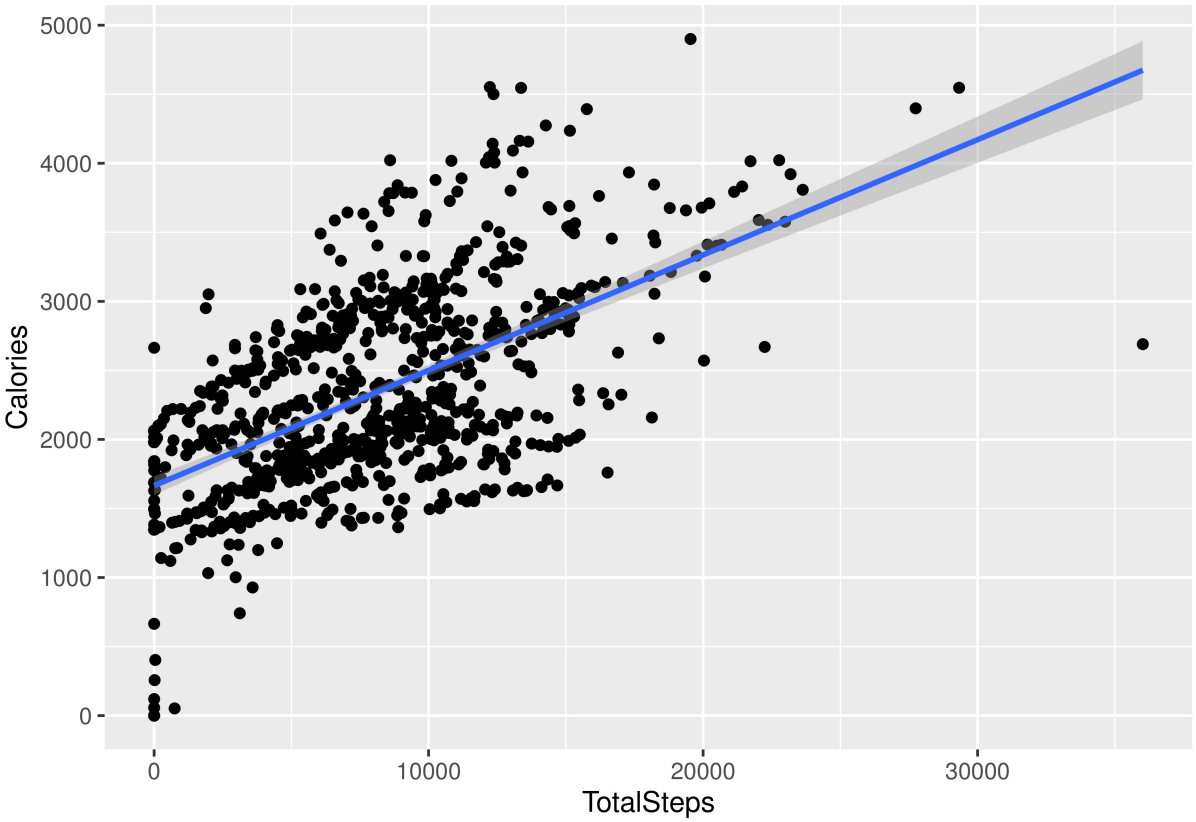
```
ggplot(data=daily_activity, aes(x=TotalSteps, y=SedentaryMinutes, color = Calories)) + geom_point()
```



Let's plot that really quickly and get rid of some of this noise.

```
ggplot(data=daily_activity, aes(x=TotalSteps, y = Calories))+ geom_point() + stat_smooth(method=lm)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



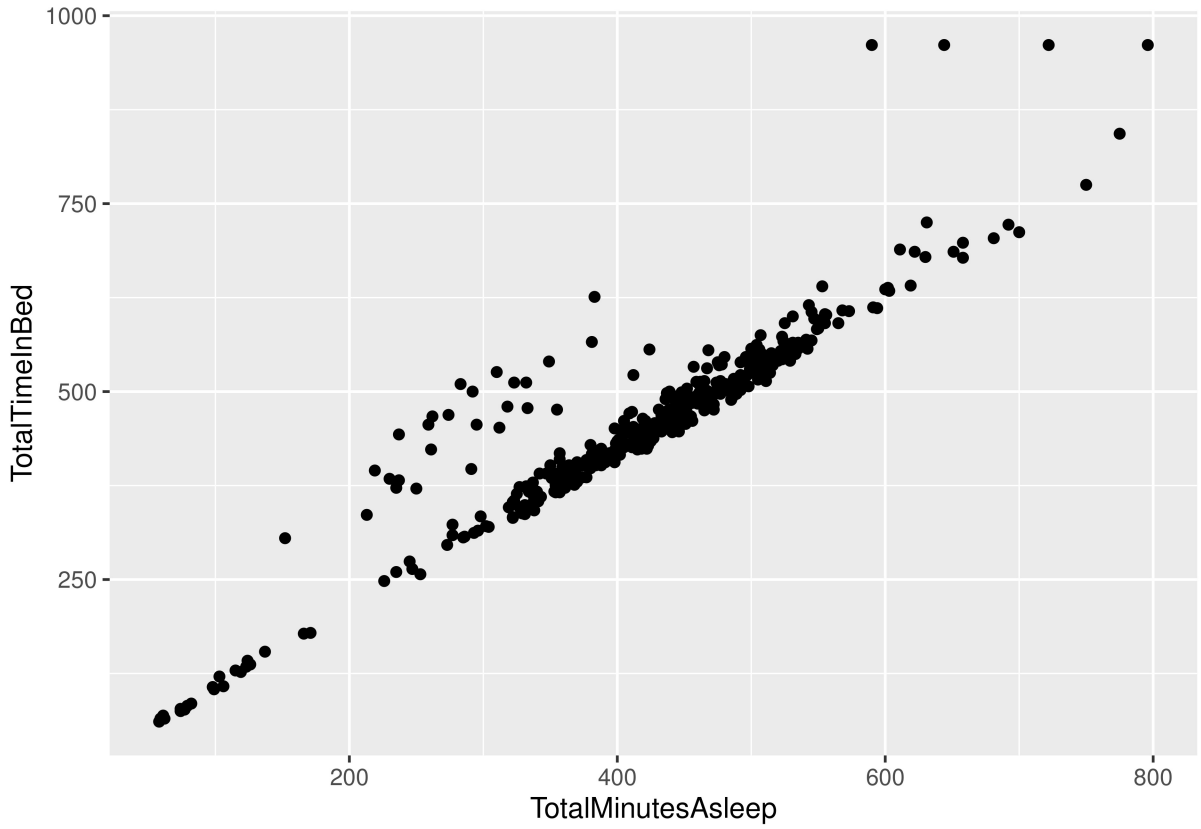
A potential strategy Here you could easily market that in order to burn calories, you don't have to do high-intensity work outs you can just get out there and start walking!

This would be such a relief as a consumer, because this proves to them that you can have results without starting a gym membership, or by starting a large workout routine. You can burn calories, simply by walking.

**Sleep and Time in Bed

Let's look at our sleep data, we should see a practically 1:1 trend from the amount of time slept and the total time someone spends in bed.

```
ggplot(data=sleep_day, aes(x=TotalMinutesAsleep, y=TotalTimeInBed)) + geom_point()
```



As we can see, there are some outliers here! some data points that spent a lot of time in bed, but didn't actually sleep, and then a small batch that slept a whole bunch and spent time in bed (I can relate)

We could definitely market to consumers to monitor their time in bed with the watch against their sleep time.

I wonder how this relates to the sedentary minutes data in the last dataset??

Merging these two datasets together

```
combined_sleep_day_data <- merge(sleep_day, daily_activity, by="Id")
head(combined_sleep_day_data)
```

```
##           Id              SleepDay TotalSleepRecords TotalMinutesAsleep
## 1 1503960366 4/12/2016 12:00:00 AM                 1                327
## 2 1503960366 4/12/2016 12:00:00 AM                 1                327
## 3 1503960366 4/12/2016 12:00:00 AM                 1                327
## 4 1503960366 4/12/2016 12:00:00 AM                 1                327
## 5 1503960366 4/12/2016 12:00:00 AM                 1                327
## 6 1503960366 4/12/2016 12:00:00 AM                 1                327
##   TotalTimeInBed ActivityDate TotalSteps TotalDistance TrackerDistance
## 1            346     5/7/2016      11992          7.71            7.71
## 2            346     5/6/2016      12159          8.03            8.03
## 3            346     5/1/2016      10602          6.81            6.81
## 4            346    4/30/2016      14673          9.25            9.25
```

```
## 5            346    4/12/2016      13162          8.50            8.50
## 6            346    4/13/2016      10735          6.97            6.97
##   LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance
## 1                        0              2.46                     2.12
## 2                        0              1.97                     0.25
## 3                        0              2.29                     1.60
## 4                        0              3.56                     1.42
## 5                        0              1.88                     0.55
## 6                        0              1.57                     0.69
##   LightActiveDistance SedentaryActiveDistance VeryActiveMinutes
## 1                3.13                       0                37
## 2                5.81                       0                24
## 3                2.92                       0                33
## 4                4.27                       0                52
## 5                6.06                       0                25
## 6                4.71                       0                21
##   FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories
## 1                  46                  175              833     1821
## 2                   6                  289              754     1896
## 3                  35                  246              730     1820
## 4                  34                  217              712     1947
## 5                  13                  328              728     1985
## 6                  19                  217              776     1797
```

We could perform an outer join to include all of the fitbit users in the dataset, but theoretically, their sleep data would be empty (either null or N/A). Let's try it!

```
combined_sleep_day_data2 <- merge(sleep_day, daily_activity, by="Id", all = TRUE)
head(combined_sleep_day_data2)
```

```
##           Id              SleepDay TotalSleepRecords TotalMinutesAsleep
## 1 1503960366 4/12/2016 12:00:00 AM                 1                327
## 2 1503960366 4/12/2016 12:00:00 AM                 1                327
## 3 1503960366 4/12/2016 12:00:00 AM                 1                327
## 4 1503960366 4/12/2016 12:00:00 AM                 1                327
## 5 1503960366 4/12/2016 12:00:00 AM                 1                327
## 6 1503960366 4/12/2016 12:00:00 AM                 1                327
##   TotalTimeInBed ActivityDate TotalSteps TotalDistance TrackerDistance
## 1            346     5/7/2016      11992          7.71            7.71
## 2            346     5/6/2016      12159          8.03            8.03
## 3            346     5/1/2016      10602          6.81            6.81
## 4            346    4/30/2016      14673          9.25            9.25
## 5            346    4/12/2016      13162          8.50            8.50
## 6            346    4/13/2016      10735          6.97            6.97
##   LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance
## 1                        0              2.46                     2.12
## 2                        0              1.97                     0.25
## 3                        0              2.29                     1.60
## 4                        0              3.56                     1.42
## 5                        0              1.88                     0.55
## 6                        0              1.57                     0.69
##   LightActiveDistance SedentaryActiveDistance VeryActiveMinutes
## 1                3.13                       0                37
```

```
## 2                5.81                      0                24
## 3                2.92                      0                33
## 4                4.27                      0                52
## 5                6.06                      0                25
## 6                4.71                      0                21
##   FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories
## 1                  46                  175              833     1821
## 2                   6                  289              754     1896
## 3                  35                  246              730     1820
## 4                  34                  217              712     1947
## 5                  13                  328              728     1985
## 6                  19                  217              776     1797
```

```
n_distinct(combined_sleep_day_data2$Id)
```

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

Excellent! as we can see all 33 values are available. Let's plot that sedentary time and time in bed data!

Sedentary Time vs Time In Bed

For this first plot we'll try it out with only the 24 unique IDs that have actually logged sleep data

Let's run a correlation to see what the correlation coefficient coefficient would be for a linear regression:

```
sedentary.lm <- lm(SedentaryMinutes ~ TotalTimeInBed, data = combined_sleep_day_data)
sedentary.lm
```

```
##
## Call:
## lm(formula = SedentaryMinutes ~ TotalTimeInBed, data = combined_sleep_day_data)
##
## Coefficients:
##    (Intercept)  TotalTimeInBed
##       921.9598         -0.2678
```

And now a pearson correlation coefficient:

```
cor(combined_sleep_day_data$TotalTimeInBed,combined_sleep_day_data$SedentaryMinutes, method = "pearson"
```

```
## [1] -0.128011
```

It looks like these two things are not related much at all. Which is an interesting finding. As time in bed goes up, sedentary minutes actually go down, but not to a statistically significant degree.

Parting Thoughts

We looked at this dataset of fitbit users pretty intensively to get an idea on what features are being used, and how we can market our items.

Takeaway 1

Fitbit does not collect hydration data, that puts Bellabeat way above the competition!

Takeaway 2

We showed that more people log their calories, steps taken, etc, and fewer users log their sleep data, and only a select few are logging their weight

Takeaway 3

To market this, we initially thought that simply being active and taking steps would help with people on their journey start to burn calories. While this may be true, we see that the correlation is beyond small, and maybe we shouldn't market it that way.