FitBit Analysis

The Problem

The problem or the general question of this analysis is to figure out the latest trends in smart device usage in general. Then figure out how these relate the company and how Bellabeat, a smart device company could use these insights to direct it's marketing strategy.

Key Stakeholders

- Urška Sršen the Chief Creative Executive.
- Sando Mur Co-founder and Mathematician

Preparation and Proccessing of the datasets

- Download the dataset from Kaggle (https://www.kaggle.com/arashnic/fitbit) to my local storage and start working on it.
- The dataset contains small files which can be viewed from spreadsheets and large files for which a spreadsheet simply can't open.
- I'll clean the small files in Excel and then move over to work on R for everything.

- In the daily activity file, there are days where there is no movement or whatsoever, meaning, sedentary activity was 1440.
- As these mean the device was not used. A quick filter and Count shows there are 79 of these instances.
- I removed all of them as well as some other values that did not any activity or low activity. Same for the daily steps spreadsheet, removed 0-9 steps taken on that day.
- I then checked the sleep Day file, no data warranted removal there, although I was intrigued by the 1 hour sleep days.
- Overall, the data only has 30 participants.
- The sample is too small to make grand generalizations from.

Analysis

Introduction

This is the analysis of the data set using R. Below will be the thought process and the code to execute and answer the question and gain insight.

Start by loading all the packages I might need in the processing of the data set.

```
library(ggplot2)
library(tidyverse)
```

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

```
## v tibble  3.1.3     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.0     v forcats 0.5.1
## v purrr   0.3.4
```

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

```
library(plyr)
```

```
## ------------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------------
```

```
##
## Attaching package: 'plyr'
```

```
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following object is masked from 'package:purrr':
##
##     compact
```

```
library(dplyr)
```

Datasets

The next thing will be to import the data sets I will be working with. First one being the daily activity

```
daily_activity <- read.csv("dailyActivity_merged.csv")
```

Followed by the sleep data.

```
sleep_day <- read.csv("sleepDay_merged.csv")
```

Next, explore the 2 files a bit, see how they look like.

```
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
```

Get the column names..

```
colnames(daily_activity)
```

```
##  [1] "Id"                       "ActivityDate"
##  [3] "TotalSteps"               "TotalDistance"
##  [5] "TrackerDistance"          "LoggedActivitiesDistance"
##  [7] "VeryActiveDistance"       "ModeratelyActiveDistance"
##  [9] "LightActiveDistance"      "SedentaryActiveDistance"
## [11] "VeryActiveMinutes"        "FairlyActiveMinutes"
## [13] "LightlyActiveMinutes"     "SedentaryMinutes"
## [15] "Calories"
```

. . . then sleep data;

```
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
```

and the column names;

```
colnames(sleep_day)
```

```
## [1] "Id"                 "SleepDay"           "TotalSleepRecords"
## [4] "TotalMinutesAsleep" "TotalTimeInBed"
```

Diving into the data

First figure out how many participants are the in the datasets;

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

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

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

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

The results show we have 9 more participants in the daily activity dataset than in the sleep day data set.

Then we get the number of observations/rows for both data sets;

```
nrow(daily_activity)
```

```
## [1] 856
```

```
nrow(sleep_day)
```

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

The daily activity data set has more than 2 times the number of observations as the sleep day data set.

Check the basic stats from the datasets, starting with the daily activity dataset;

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

```
##    TotalSteps    TotalDistance    SedentaryMinutes
##  Min.   :    0   Min.   : 0.000   Min.   : 125.0
##  1st Qu.: 4924   1st Qu.: 3.370   1st Qu.: 721.8
##  Median : 8054   Median : 5.590   Median :1020.0
##  Mean   : 8328   Mean   : 5.983   Mean   : 955.4
##  3rd Qu.:11100   3rd Qu.: 7.912   3rd Qu.:1188.0
##  Max.   :36019   Max.   :28.030   Max.   :1439.0
```

Already from this summary, there is one thing that is clear, there is a huge gap between the average number of steps(mean) and the maximum number of steps, same goes for the total distance.

Then to the sleep day data set;

```
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
```

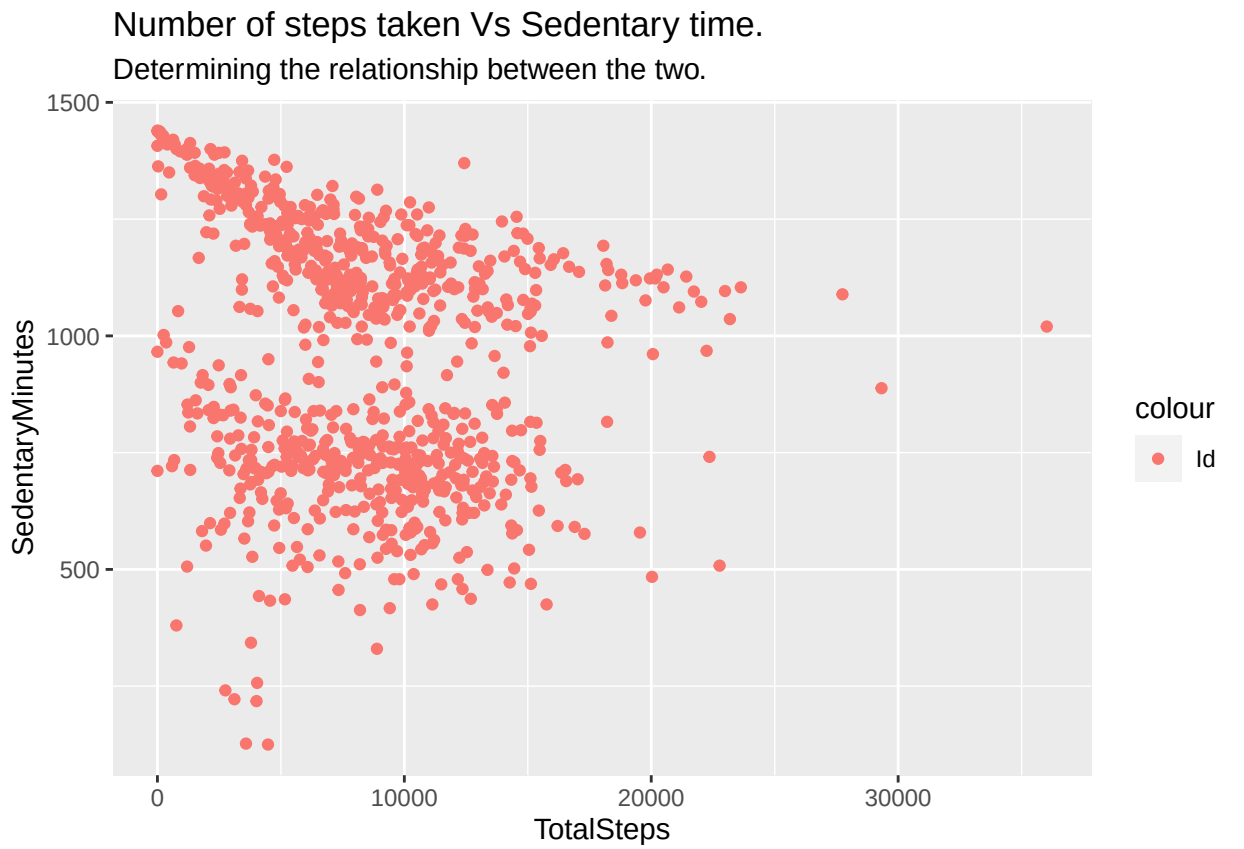
There are sleep hours which are questionable, like the 796 minutes, do they indicate sick days, should there be a system to notify a parent, spouse, doctor or a caretaker if there is ever such an unusual activity.

To get a better insight, we need to plot some graphs.

Plotting some graphs

Lets first check the relationship between the number of steps taken and the sedentary minutes. Hypothesis, the more steps you are taking the less sedentary minutes.

```
ggplot(data=daily_activity, aes(x=TotalSteps, y=SedentaryMinutes, color = "Id")) + geom_point() +labs(t
```

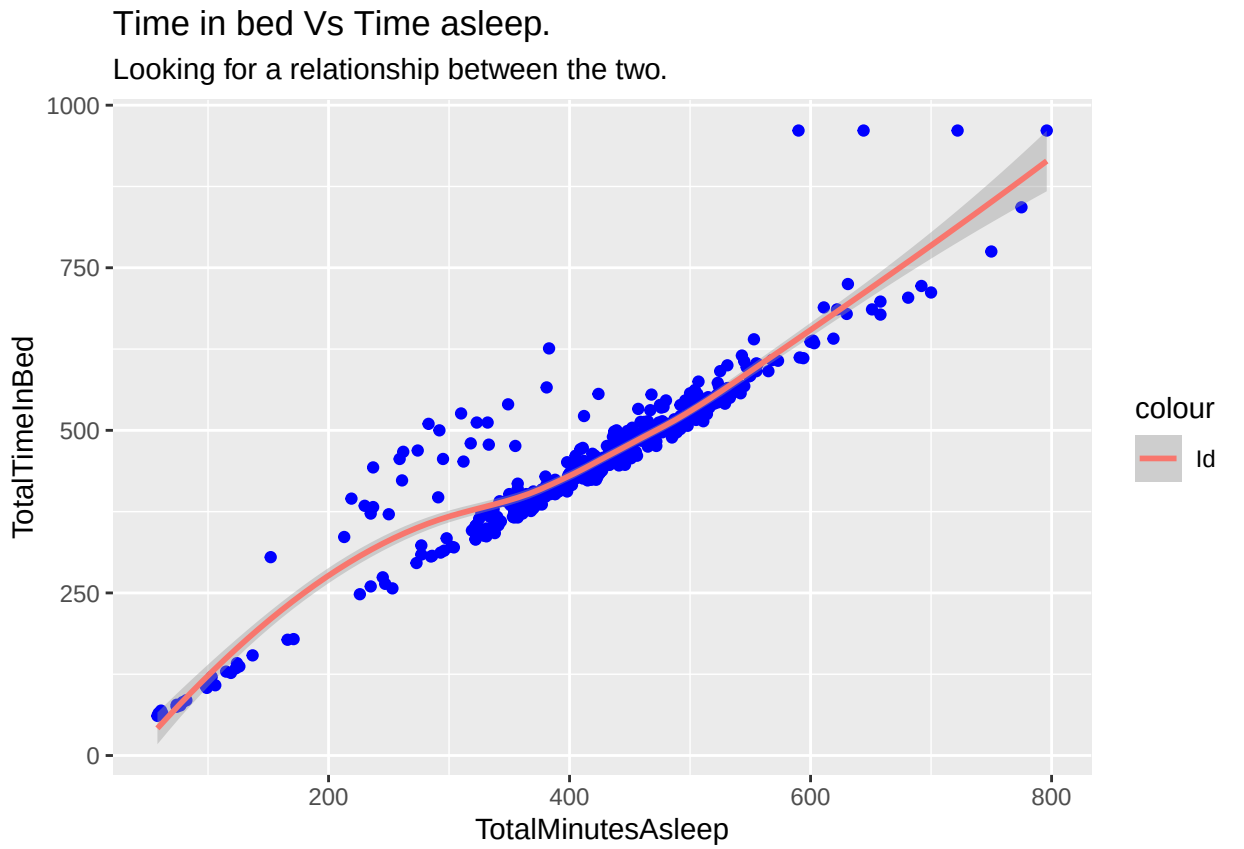


Take away from the graph... The two are negatively correlated as one might guess, if someone is spending more time being inactive, you can expect they will have fewer steps than someone who is more active. How does this help us? Are the other activities either than taking that should be tracked.

Next, minutes in bed vs the total minutes asleep

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

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




```
cor(sleep_day$TotalMinutesAsleep, sleep_day$TotalMinutesAsleep ,method = "pearson")
```

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

Almost perfectly linear relationship, some outliers though, hard to ignore, what could it mean? It might be that they struggle to sleep, are they having enough activity? Users like these, if they have made their information available to the company, could be offered program in the membership subscription to help with their inability to sleep or underlying problem.

What could these trends tell you about how to help market this product? Or areas where you might want to explore further?

Merge the two datasets

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

Check the number of participants, if not 24, there will be a need to adjust the code

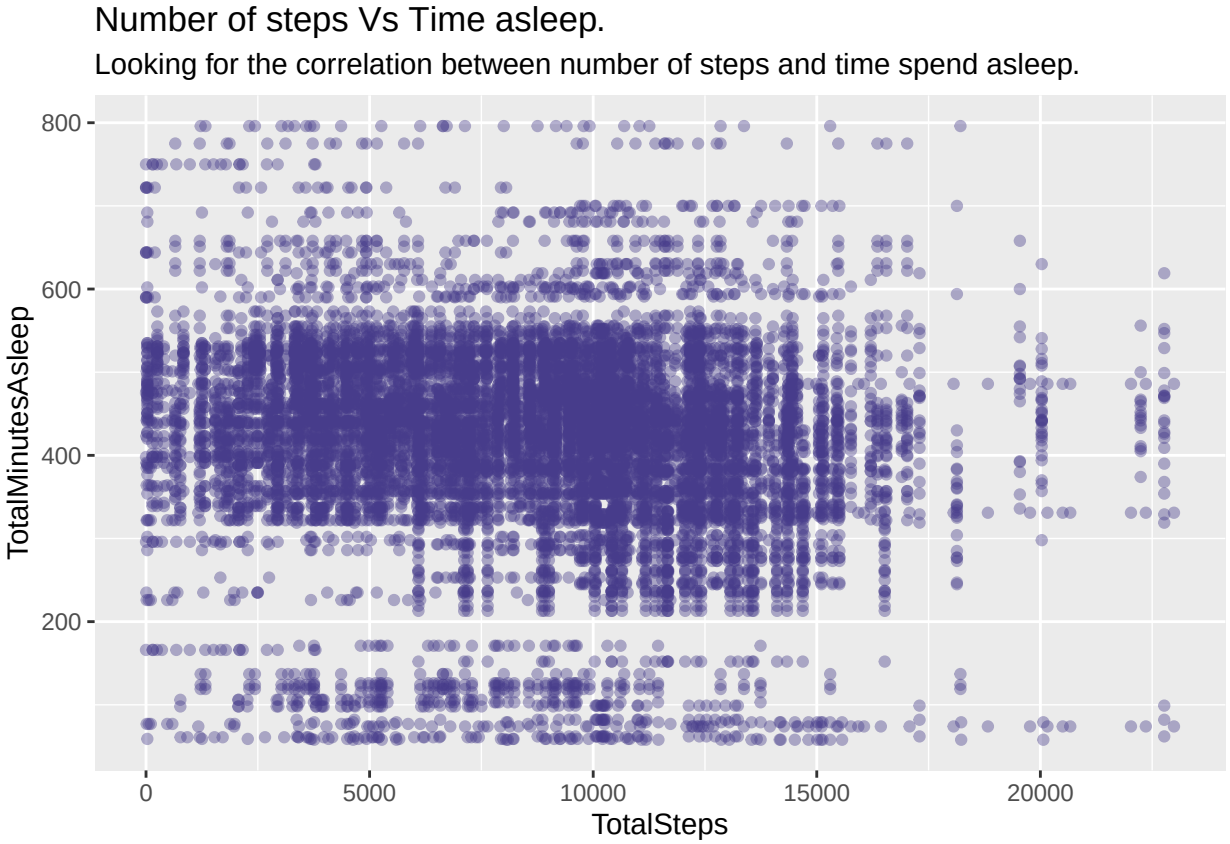
```
n_distinct(combined_data$Id)
```

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

There are equal number of Ids.

Next will be to look for relationships in this data, first will be the number of steps compared to the sleep time on a given day. I am expecting that when one takes more steps they will get better/more sleep, let's see if this is the case.

```
ggplot(data=combined_data, aes(x=TotalSteps, y=TotalMinutesAsleep)) + geom_point(color="slateblue4",alp
```



Most people seem to have about the same amount of sleep time regardless of the number of steps taken, that is the dominant trend.

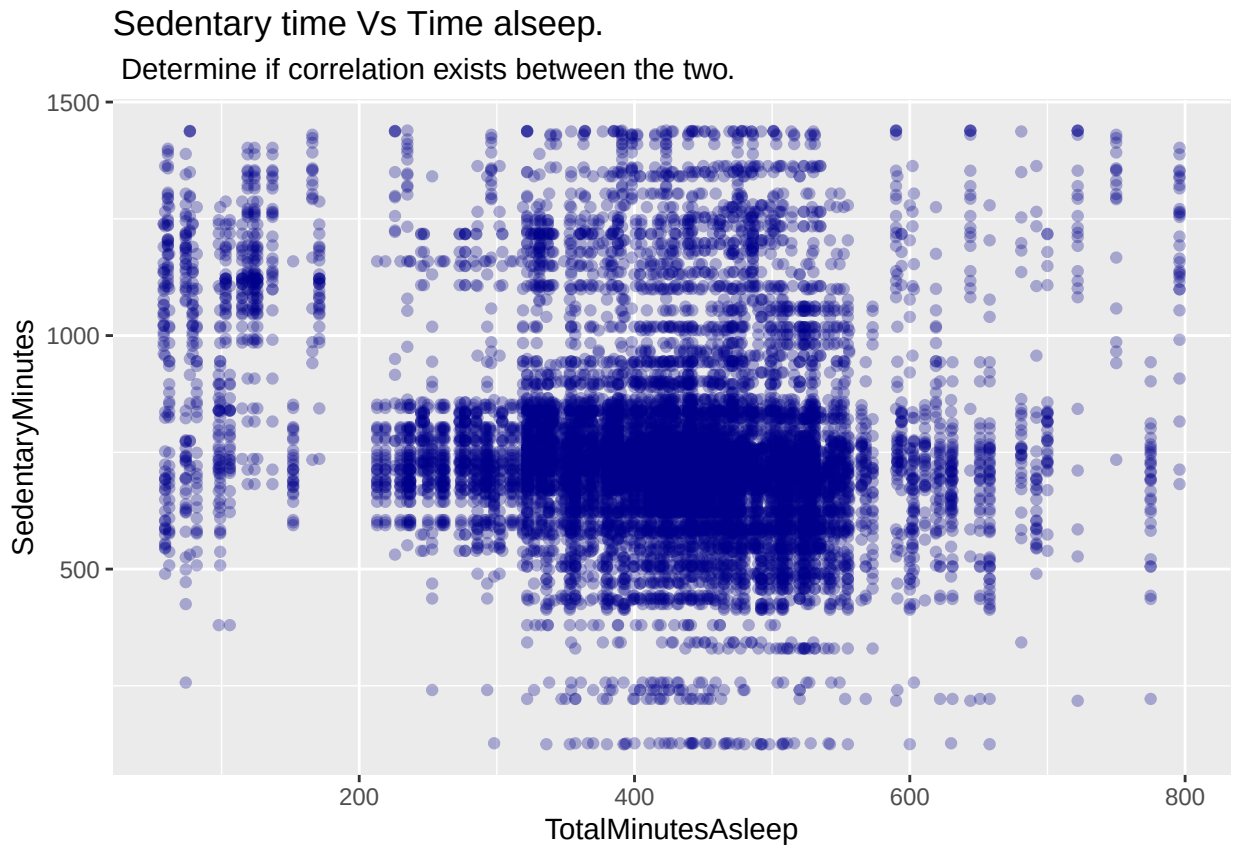
```
cor(combined_data$TotalMinutesAsleep, combined_data$TotalSteps ,method = "pearson")
```

```
## [1] -0.1106412
```

A weak negative correlation exists, bottom line, most people sleep for 6-8 hour regardless of activity on a given day.

I am now curious about sedentary minutes vs sleep time.

```
ggplot(data=combined_data, aes(x=TotalMinutesAsleep, y=SedentaryMinutes)) + geom_point(color = "darkblu
```




```
cor(combined_data$SedentaryMinutes, combined_data$TotalSteps ,method = "pearson")
```

```
## [1] -0.2024329
```

Similar results as before, a very weak negative correlation.

Import more files

At this point, one is shooting darts, seeing if there's anything that catches my eye.

```
daily_calories <- read.csv("dailyCalories_merged.csv")
```

```
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
```

```
n_distinct(daily_calories$Id)
```

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

```
combined_with_calories <- merge(combined_data, daily_calories, by="Id")
```

```
n_distinct(combined_with_calories$Id)
```

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

```
head(combined_with_calories)
```

```
##           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    4/22/2016      12764          8.13            8.13
## 2            346    4/22/2016      12764          8.13            8.13
## 3            346    4/22/2016      12764          8.13            8.13
## 4            346    4/22/2016      12764          8.13            8.13
## 5            346    4/22/2016      12764          8.13            8.13
## 6            346    4/22/2016      12764          8.13            8.13
##   LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance
## 1                        0              4.76                     1.12
## 2                        0              4.76                     1.12
## 3                        0              4.76                     1.12
## 4                        0              4.76                     1.12
## 5                        0              4.76                     1.12
## 6                        0              4.76                     1.12
##   LightActiveDistance SedentaryActiveDistance VeryActiveMinutes
## 1                2.24                       0                66
## 2                2.24                       0                66
## 3                2.24                       0                66
## 4                2.24                       0                66
## 5                2.24                       0                66
## 6                2.24                       0                66
##   FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories.x
## 1                  27                  130             1217       1827
## 2                  27                  130             1217       1827
## 3                  27                  130             1217       1827
## 4                  27                  130             1217       1827
## 5                  27                  130             1217       1827
## 6                  27                  130             1217       1827
##   ActivityDay Calories.y
## 1    5/7/2016       1821
## 2    5/6/2016       1896
## 3    5/1/2016       1820
## 4   4/30/2016       1947
## 5   4/12/2016       1985
## 6   4/13/2016       1797
```

The hypothesis here is that the more steps the more calories burnt.

```
ggplot(data= daily_activity, aes(x=Calories, y=TotalSteps)) + geom_point(color = "darkblue", alpha=0.5)
```

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


```
cor(daily_activity$Calories, daily_activity$TotalSteps ,method = "pearson")
```

```
## [1] 0.5585717
```

A positive correlation, moderately strong.

Conclusion and Recommendations

One can continue to import a lot more other datasets to analyse, but it is clear as day what people use their smart devices for and there is no insight to be gained in this regard from further digging, given our main objective/question.

- People use their smart devices to better take care of their health, all these variables we have been measuring here are of utmost importance to owners of smart devices so they can make healthy life choices.
- The other apparent trend is that in a given group, there is a huge difference between the activity of participants, some are more active while others aren't.

Now, to the most important question, how does this help and inform the markerting stratergy of Bellabeat?

- The first and most obvious is that Bellabeat is missing a product for a subset of their target that is above average active, commonly refered to as "sporty", this means they need to design a product that is going to cater for this group, both in the design aesthetics and functionality. This is obviously long term and will require more time.

An easier to implement program to better promote their existing product line up is as follows;

- Introduce reward programs for people achieving their health goal.

- If someone wants to have reach a certain number of steps per day, wants to achieve a certain level of weight and all the other metrics, they can then be rewarded with points for hitting the targets, and these point, after they reach a certain number, can be used to grant them access to the membership subscription for a limited period.
- This will have an effect where there is a consistent use of the three products and a much better chance that people become subscribers after the trial period, a win win.