BELLABEAT A DATA ANALYSIS CASE STUDY

Overview

The Google Data Analytics Professional Certificate, the course I am currently working on, requires me to do a case study. In this Case Study, a fictional company called Bellabeat has hired me as their junior data analyst. Bellabeat is a successful small company, but they have the potential to become a larger player in the global smart device market. The co-founder and Chief Creative Officer of Bellabeat believes that analyzing smart device fitness data could help unlock new growth opportunities for the company. I am tasked to focus on one of Bellabeat's products and analyze smart device data to gain insight into how consumers are using their smart devices.

Dataset

The dataset used is https://www.kaggle.com/arashnic/fitbit, The Fitbit Fitness Tracker Dataset. These datasets were generated by respondents to a distributed survey via Amazon Mechanical Turk between 03.12.2016 and 05.12.2016.

Features/Data frames

Sleep data -> 'Id', 'SleepDay', 'TotalSleepRecords', 'TotalMinutesAsleep', 'TotalTimeInBed'
Weight data -> 'Id', 'Date', 'WeightKg', 'WeightPounds', 'Fat', 'BMI', 'IsManualReport', 'LogId'
Daily Intensities -> 'Id', 'ActivityDay', 'SedentaryMinutes', 'LightlyActiveMinutes', 'FairlyActiveMinutes', 'VeryActiveMinutes', 'SedentaryActiveDistance', 'LightActiveDistance', 'ModeratelyActiveDistance', 'VeryActiveDistance'
Daily Steps -> 'Id', 'ActivityDay', 'StepTotal'
Daily Activity -> 'Id', 'ActivityDate', 'TotalSteps', 'TotalDistance', 'TrackerDistance', 'LoggedActivitiesDistance', 'VeryActiveDistance', 'ModeratelyActiveDistance', 'LightActiveDistance', 'SedentaryActiveDistance', 'VeryActiveMinutes', 'FairlyActiveMinutes', 'LightlyActiveMinutes', 'SedentaryMinutes', 'Calories'

Cleaning Process and Analysis

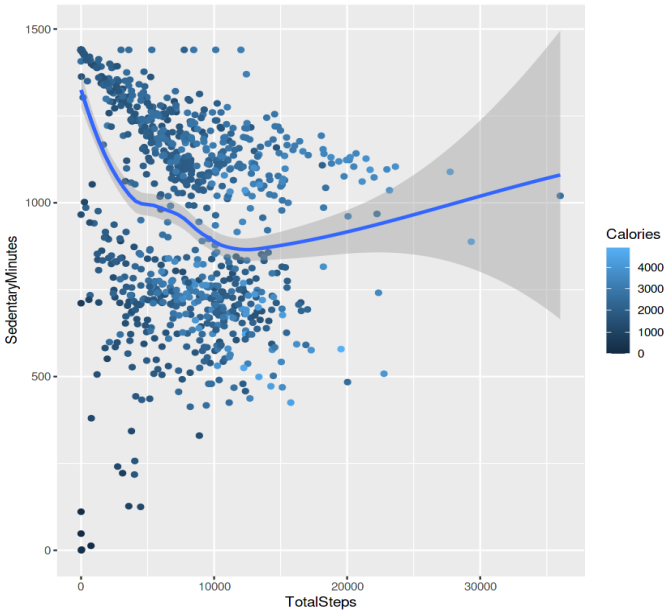
Cleaning

For the cleaning and the visualization process, I used R studio. With all the different data frames to explore, I went through the data frames one by one. The data frames Data Activity and Daily intensities had pretty much the same features. So, I wanted to merge them into one dataset, but before that, I had to check if they actually had the same data. So, I used the all.equal() function. Checked the fairness and then merged the data frames into one data frame.

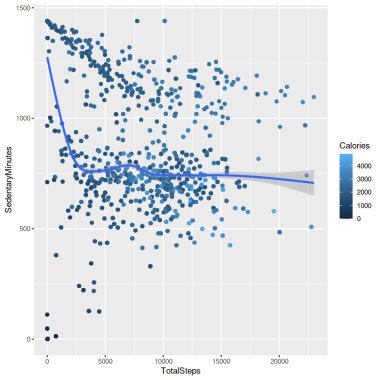
Analysis

My initial thought process was the same, as when everyone first looks at the data. Since there were TotalSteps and Calories as features in the data frame, I came to the conclusion that they must have a positive correlation. There was also Sedentary Minutes as a feature, Sedentary Minutes must be inversely proportional to TotalSteps. So, can there be a relation between Sedentary Minutes and Calories? To figure that out I plotted a point graph.

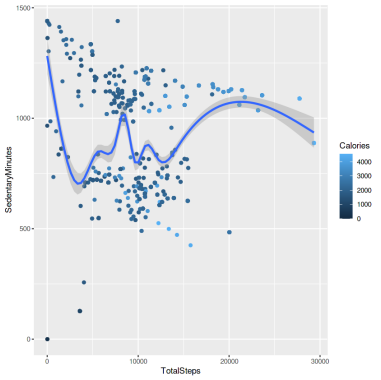
From the graph, it is clear that Calories and total steps have a positive correlation and Total steps and Sedentary Minutes are inversely proportional. But, surprisingly there is no correlation between sedentary minutes and Calories.



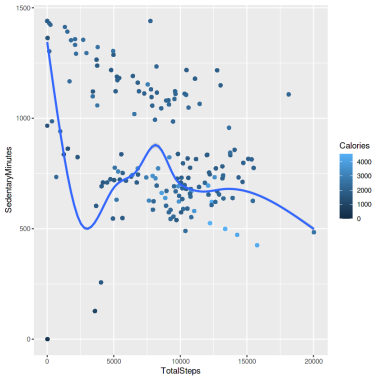
With the datasets about the sleep logs and weight logs, it became clear that many people did not in fact keep track of their sleep and weight logs. Everyone kept track of their Calories and Total Steps but very little kept track of their sleep and weight logs. To see how Calories and Total Steps varied from people who took sleep logs and weight logs. I merged the data grouped by Id. So, I had 3 different data frames now with data of people who took only sleep logs and data of people who took only weight logs, and people who took both sleep and weight logs.



People who took only sleep logs.

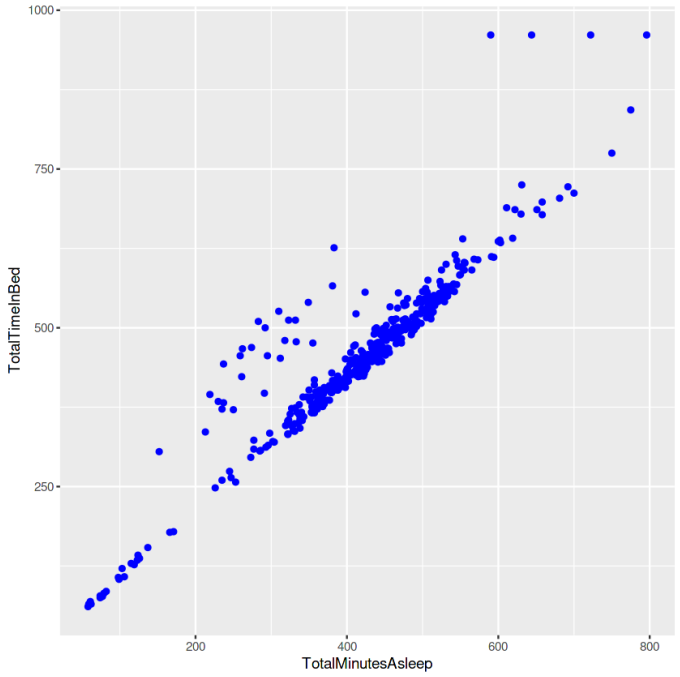


People who took only weight logs.



People who took both sleep & weight logs.

I found a fun fact with the analysis of the sleep logs data set.



With this graph it is is clear that many people who are in bed aren't necessarily asleep.

Even with the ambiguous and very little data on weight and sleep, it is clear that keep logs have a positive effect to get the expected result of Calorie loss.

Conclusion

- Very few people log their sleep and weight. Many log only their steps taken and Total Calories.
- Total Steps have a positive correlation along with weight log and sleep log. So, logging these information will have a good effect.
- Factors such as Sedentary Minutes has a negative correlation with Calories. So, using these features will not provide any results.

Notebook:- https://www.kaggle.com/adhiau/bellabeat