Good Data Analytics Case Study

John Pippins

9/9/2021

Project Intro

This is a case study where I am assuming the role of a data analyst for a fitness-oriented smartwear company called Bellabeat and I have been tasked to analyze a publicly available dataset available here. Using insights from this data, I will provide a recommendation to Bellabeat on how they might better market their brands.

Ask

In this phase, I took the stakeholder's desires and transformed them into a clear business task that I can use to guide my analysis.

Business task: Analyze publicly available smart device data to better market a Bellabeat product.

Prepare

The data I used was uploaded to Kaggle by a user named Mobius. He acquired the data by responders to a distributed survey via Amazon Mechanical Turk over the course of a two month period in 2016. It's organized into a number of .csv files. The respondents are only identified by an ID number and the distribution was random, so there are no issues with bias.

First I loaded a number of useful libraries to better analyze the data.

```
library("tidyverse")
library("dplyr")
library("janitor")
library("skimr")
library("here")
library("ggplot2")
library("patchwork")
```

Next added some code to ensure that scientific notation wouldn't be present in the document.

```
options(scipen = 100)
```

Then I loaded in some datasets that looked promising. Some of the other datasets were either incomplete or were subsets of these datasets. I wanted to see if there might be some trends in how often users wore their devices or if their might be some relationship between calories burned, steps taken, and amount of sleep.

```
daily_activity = read.csv("dailyActivity_merged.csv")
daily_calories = read.csv("dailyCalories_merged.csv")
daily_steps = read.csv("dailySteps_merged.csv")
sleep_info = read.csv("sleepDay_merged.csv")
```

Then I previewed the data to get a good idea of what it looked like and if there were any discrepancies.

```
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
```

```
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
```

```
head(sleep_info)
```

```
##           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
```

```
head(daily_steps)
```

```
##           Id ActivityDay StepTotal
## 1 1503960366   4/12/2016     13162
## 2 1503960366   4/13/2016     10735
## 3 1503960366   4/14/2016     10460
## 4 1503960366   4/15/2016      9762
## 5 1503960366   4/16/2016     12669
## 6 1503960366   4/17/2016      9705
```

I could see immediately that a number of the columns were the same, like “Id” and “Date,” but some of the datasets had their date format in different ways.

Process and Analyze

I did all of my data processing in analysis in R utilizing RStudio because R is capable of performing the analysis and also producing quality visualizations.

First I made sure that the “Date” column was the same in all datasets.

```
sleep_info = separate(sleep_info,SleepDay,c("Date","Time","AM/PM"), sep = "
")
daily_calories = rename(daily_calories, Date = ActivityDay)
daily_activity = rename(daily_activity, Date = ActivityDate)
daily_steps = rename(daily_steps, Date = ActivityDay)
colnames(daily_activity)
```

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

```
colnames(daily_calories)
```

```
## [1] "Id"       "Date"     "Calories"
```

```
colnames(sleep_info)
```

```
## [1] "Id"                   "Date"                 "Time"
## [4] "AM/PM"                "TotalSleepRecords"  "TotalMinutesAsleep"
## [7] "TotalTimeInBed"
```

```
colnames(daily_steps)
```

```
## [1] "Id"        "Date"      "StepTotal"
```

Next I wanted to join up some similar datasets for easier cleanup.

```
merged_steps_calories = full_join(daily_steps,daily_calories)
```

```
## Joining, by = c("Id", "Date")
```

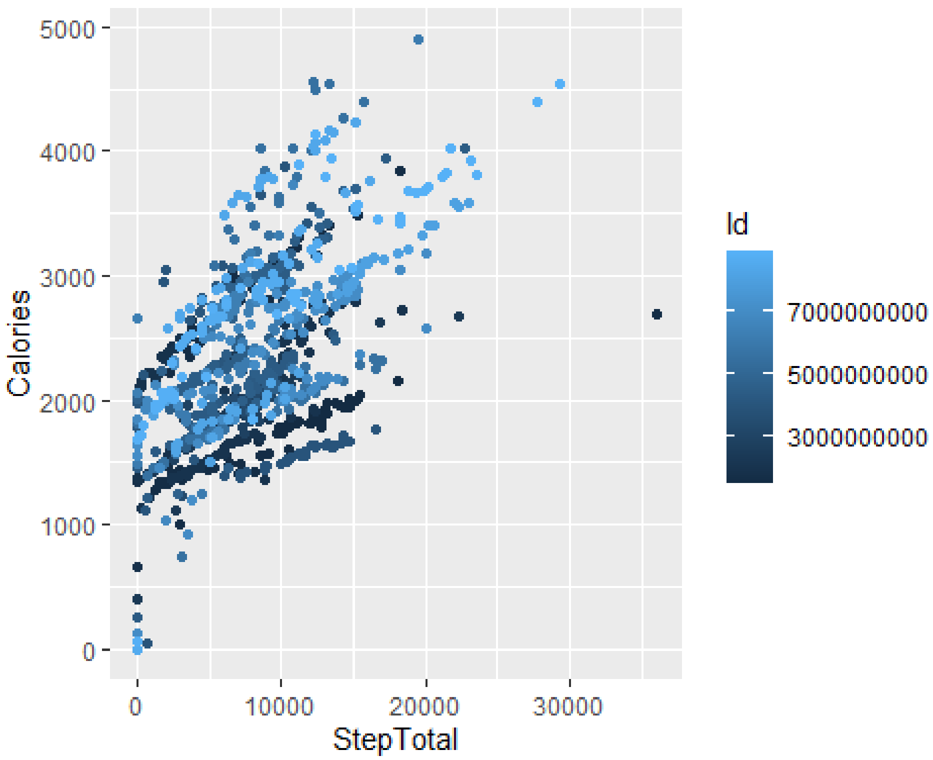
Now I wanted to check for any outliers.

```
merged_steps_calories%>%
  select(StepTotal,Calories)%>%
  summary()
```

```
##    StepTotal        Calories
##  Min.   :    0   Min.   :   0
##  1st Qu.: 3790   1st Qu.:1828
##  Median : 7406   Median :2134
##  Mean   : 7638   Mean   :2304
##  3rd Qu.:10727   3rd Qu.:2793
##  Max.   :36019   Max.   :4900
```

It does look initially like we have some outliers in both StepTotal and Calories. There are entries of 0, which means that either the user did not enter their information for that day or did not wear the device. Also, the max on steps is abnormally high. I also want to visualize if there are any outliers.

```
ggplot(data = merged_steps_calories)+
  geom_jitter(mapping = aes(x = StepTotal, y = Calories, color = Id))
```



There is certainly a strong correlation between total steps and calories burned. There are some data points where users who walked the same number of steps as other users burned less calories. This can be attributed to the fact that some users have a higher reported BMI than others and may be more overweight.

I then corrected for outliers to make analysis more accurate.

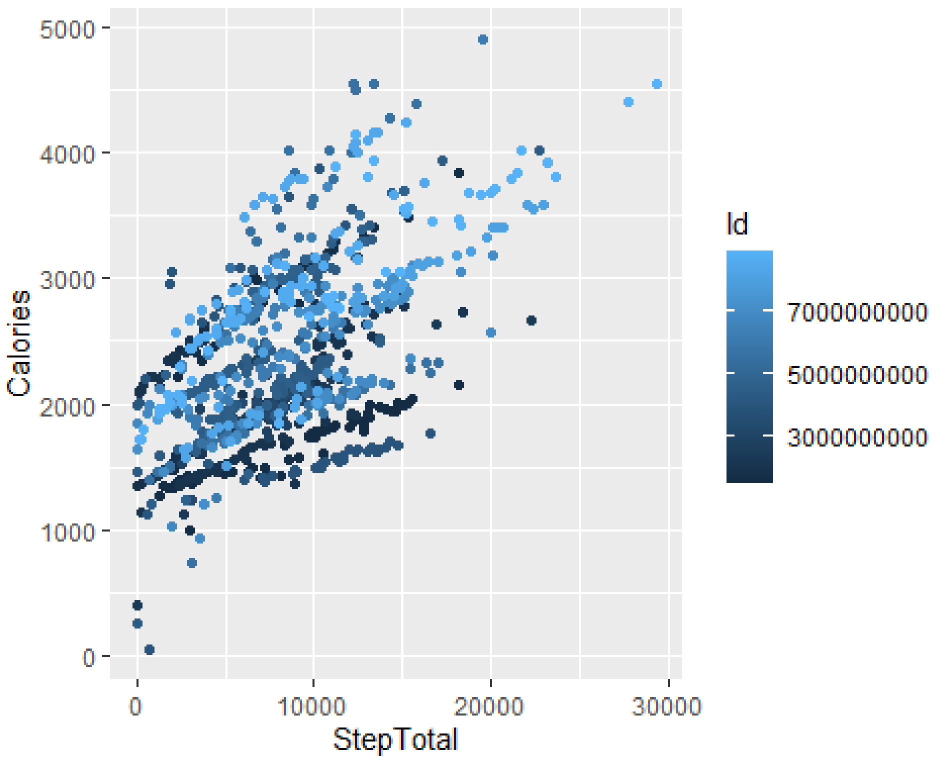
```
corrected_data = subset(merged_steps_calories, StepTotal < 30000 & StepTotal
> 0 & Calories > 0)

corrected_data%>%
  select(StepTotal,Calories)%>%
  summary()

##    StepTotal        Calories   
##  Min.   :    4   Min.   :  52  
##  1st Qu.: 4922   1st Qu.:1855  
##  Median : 8027   Median :2220  
##  Mean   : 8287   Mean   :2361  
##  3rd Qu.:11075   3rd Qu.:2832  
##  Max.   :29326   Max.   :4900
```

Then I plot the data without outliers.

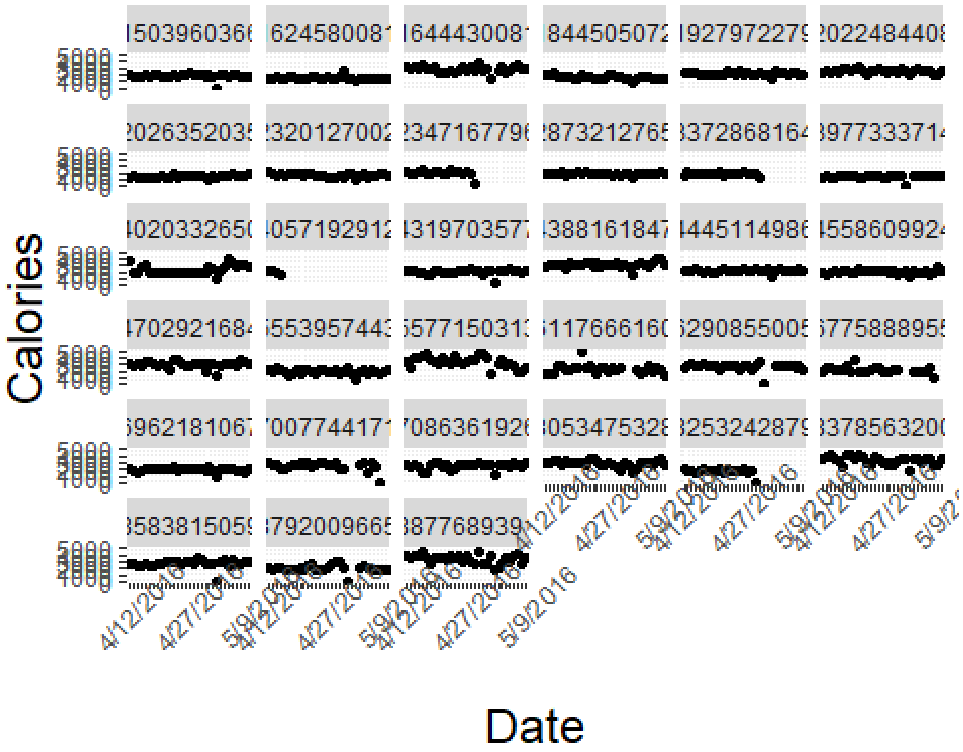
```
ggplot(data = corrected_data)+
  geom_jitter(mapping = aes(x = StepTotal, y = Calories, color = Id))
```



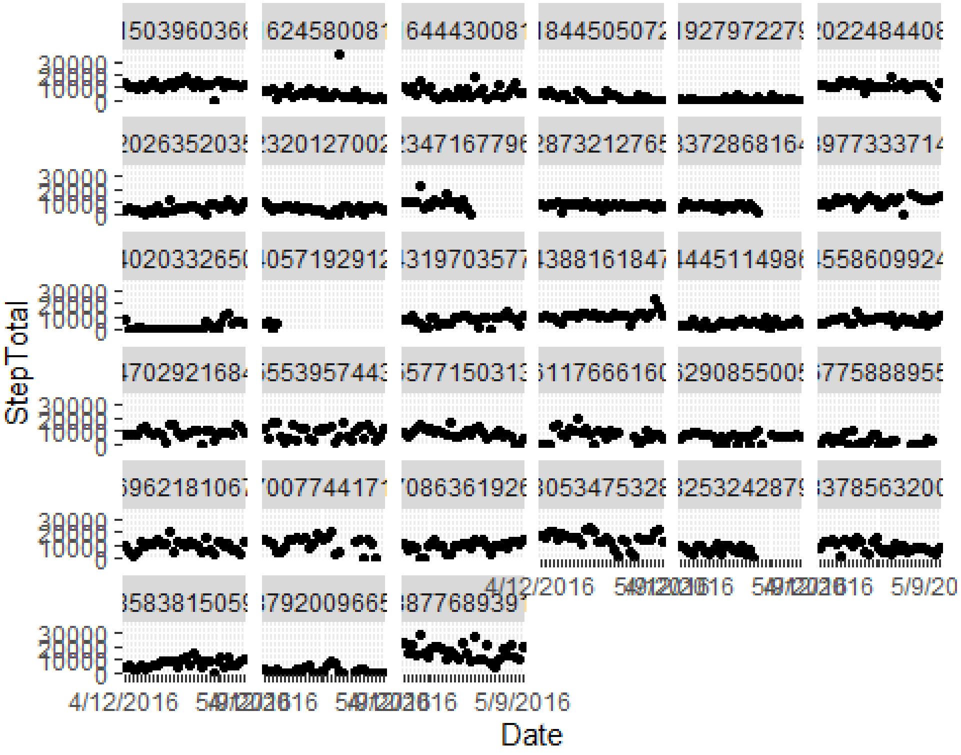
This implies a strong correlation between total number of steps and calories.

Next I wanted to see if any users did not participate in the full length of the survey or if there might be some general trends in daily wear.

```
ggplot(data = daily_activity)+
  geom_point(mapping = aes(x=Date,y=Calories))+
  facet_wrap(~Id)+
  theme(
    axis.title.y=element_text(size=18),
    axis.title.x=element_text(size=18),
    axis.text.x=element_text(angle=45, size=10)
  )+
  scale_x_discrete(guide = guide_axis(check.overlap = TRUE))
```



```
ggplot(data = daily_steps)+
  geom_point(mapping = aes(x=Date,y=StepTotal))+
  facet_wrap(~Id)+
  scale_x_discrete(guide = guide_axis(check.overlap = TRUE))
```



A couple of users did not complete the entire survey. For the ones that did finish, calories burned and steps taken are similarly shaped for the same Id's implying a positive correlations. Next, I wanted to calculate if there was a positive correlation

```
cor(merged_steps_calories$StepTotal,merged_steps_calories$Calories)

## [1] 0.5915681
```

This told me that there was a positive correlation between the two variables.

Recommendations

Based on this data, there is a connection between steps taken and calories burned. Bellabeat can use this connection to better market its products.

- Bellabeat should market the benefits of daily steps towards burning calories.
- The Bellabeat Leaf is suited for this, as it can track a user's steps over a period of time, and even give them data on how many steps/calories that they have burned.