Bellabeat

Breanna Sayre

08/07/2022

Phase 1: Ask

The Business Task: In order to unlock new growth opportunities for their business, Bellabeat has asked me to analyze their smart device data to see how customers utilize their devices, so they can use this information to guide a new marketing strategy.

Phase 2: Prepare

FitBit Fitness Tracker Data The company provided me with the FitBit Fitness Tracker Data made available by Mobius and found on Kaggle under public domain found here: link.

The data was collected from thirty fitbit users between April and May of 2016. To meet the business task, I plan on utilizng the data from the daily steps, caloric intake, weight log, and sleep logs spreadsheets.

Phase 3: Process

Excel then R Because the datasets aren't too large, I cleaned the data with Excel. For each of the four spreadsheets:

- The date columns were changed to the format MM/DD/YYYY, and their column name was changed to "Date" to make it more consistent.
- Duplicate rows were removed to prevent errors during analysis.

Phase 4: Analyze

Running some explatory calculations on the data First, I need to install the packages needed to analyze the data.

```
install.packages("tidyverse")
library(tidyverse)
library(dplyr)
library(ggplot2)
```

Then, I'm going to create dataframes for all the datasets.

```
calories <- read.csv("capstone_calories.csv")
steps <- read.csv("capstone_daily_steps.csv")
sleep <- read.csv("capstone_sleep.csv")
weight <- read.csv("capstone_weight_log.csv")
```

Let's see how many participants we have for each data set.

```
n_distinct(calories$Id)
```

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

```
n_distinct(steps$Id)
```

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

```
n_distinct(sleep$Id)
```

```
## [1] 26
```

```
n_distinct(weight$Id)
```

```
## [1] 9
```

So there are 33 participants for both the calories and steps dataset, but there are only 26 for the sleep log data, and 8 participants included in the weight log data. The decrease in participants for the latter two datasets may have to do with having to input this data manually.

Let"s create a tibble to look at the average, maximumum, and minimum values for each participant in the datasets while filtering out zero values.

The mean, maximum, and minimum calories burned daily by each participant.

```
calories %>%
  filter(Calories != 0) %>%
  group_by(Id) %>%
  summarize(mean(Calories),max(Calories), min(Calories))
```

```
## # A tibble: 33 x 4
##            Id `mean(Calories)` `max(Calories)` `min(Calories)`
##         <dbl>            <dbl>           <int>           <int>
##  1 1503960366            1877.            2159            1728
##  2 1624580081            1483.            2690            1002
##  3 1644430081            2811.            3846            1276
##  4 1844505072            1573.            2130             665
##  5 1927972279            2173.            2638            1383
##  6 2022484408            2510.            3158            1848
##  7 2026352035            1541.            1926            1141
##  8 2320127002            1724.            2124            1125
##  9 2347167796            2043.            2670             403
## 10 2873212765            1917.            2241            1431
## # ... with 23 more rows
## # i Use `print(n = ...)` to see more rows
```

The mean, maximum, and minimum daily steps of each participant.

```
steps %>%
  filter(StepTotal != 0) %>%
  group_by(Id) %>%
  summarize(mean(StepTotal),max(StepTotal), min(StepTotal))
```

```
## # A tibble: 33 x 4
##            Id `mean(StepTotal)` `max(StepTotal)` `min(StepTotal)`
##         <dbl>             <dbl>            <int>            <int>
##  1 1503960366            12521.            18134             9705
##  2 1624580081             5744.            36019             1510
##  3 1644430081             7283.            18213             1223
##  4 1844505072             3809.             8054                4
##  5 1927972279             1671.             3790              149
##  6 2022484408            11371.            18387             3292
##  7 2026352035             5567.            12357              254
```

```
##  8 2320127002             4717.            10725              772
##  9 2347167796             9520.            22244               42
## 10 2873212765             7556.             9685             2524
## # ... with 23 more rows
## # i Use `print(n = ...)` to see more rows
```

There are a lot of zeros in this dataset that I filtered out, so the true mean of each participant's steps could be found. These zeros likely mean that the person took off their device and didn't wear it that day.

The mean, maximum, and minimum amount of daily sleep for each participant.

```
sleep %>%
  filter(TotalMinutesAsleep != 0) %>%
  group_by(Id) %>%
  summarize(mean(TotalMinutesAsleep),max(TotalMinutesAsleep), min(TotalMinutesAsleep))
```

```
## # A tibble: 24 x 4
##            Id `mean(TotalMinutesAsleep)` `max(TotalMinutesAsleep)` min(TotalMi~1
##         <dbl>                      <dbl>                     <int>         <int>
##  1 1503960366                       360.                       700           245
##  2 1644430081                       294                        796           119
##  3 1844505072                       652                        722           590
##  4 1927972279                       417                        750           166
##  5 2026352035                       506.                       573           357
##  6 2320127002                        61                         61            61
##  7 2347167796                       447.                       556           374
##  8 3977333714                       294.                       424           152
##  9 4020332650                       349.                       501            77
## 10 4319703577                       477.                       692            59
## # ... with 14 more rows, and abbreviated variable name
## #   1: `min(TotalMinutesAsleep)`
## # i Use `print(n = ...)` to see more rows
```

The mean, maximum, and minimum amount of daily weight for each participant.

```
weight %>%
  drop_na(WeightPounds) %>%
  group_by(Id) %>%
  summarize(mean(WeightPounds),max(WeightPounds), min(WeightPounds))
```

```
## # A tibble: 8 x 4
##           Id `mean(WeightPounds)` `max(WeightPounds)` `min(WeightPounds)`
##        <dbl>                <dbl>               <dbl>               <dbl>
## 1 1503960366                116.                116.                116.
## 2 1927972279                294.                294.                294.
## 3 2873212765                126.                126.                125.
## 4 4319703577                160.                160.                159.
## 5 4558609924                154.                155.                152.
## 6 5577150313                200.                200.                200.
## 7 6962181067                136.                138.                134.
## 8 8877689391                188.                189.                185.
```

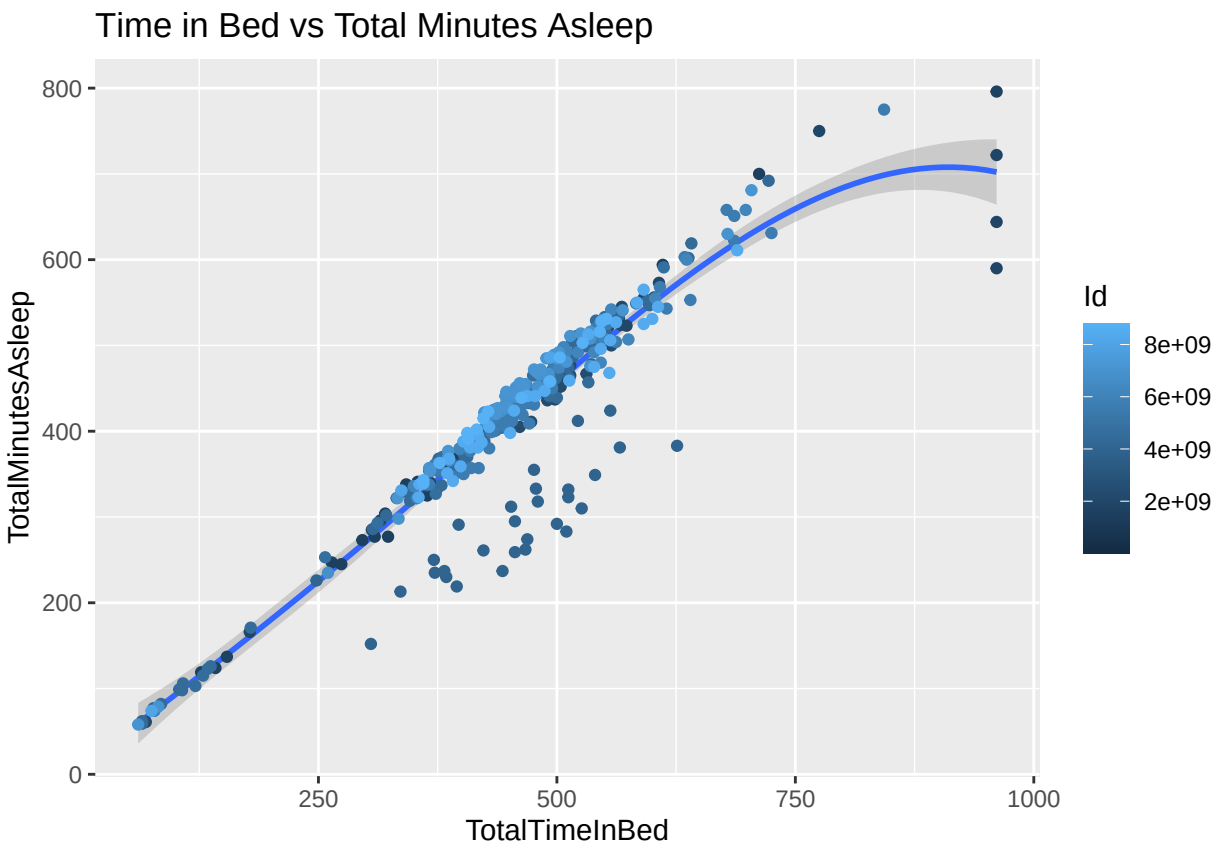
As can be seen, each participant's weight stayed pretty consistent; however, these summaries don't show the full picture. Participant 1503960366 only entered their weight twice, and only a day had passed, so their weight obviously would remain the same. In addition, participants 1927972279 and 5577150313 only entered their weight once. So their data was consistent only because they didn't enter their weight over any time. Other participants either waited a few weeks to enter their weight in again, or they tracked it semi-regularly

across multiple days. With this dataset, that really only leaves 5 participants with analyzable data, so I decided against analyzing this data to prevent creating false conclusions.

Phase 5: Share

Now let's get familiar with our data and discover new relationships between variables. How many minutes do people spend in bed compared to how much they actually sleep?

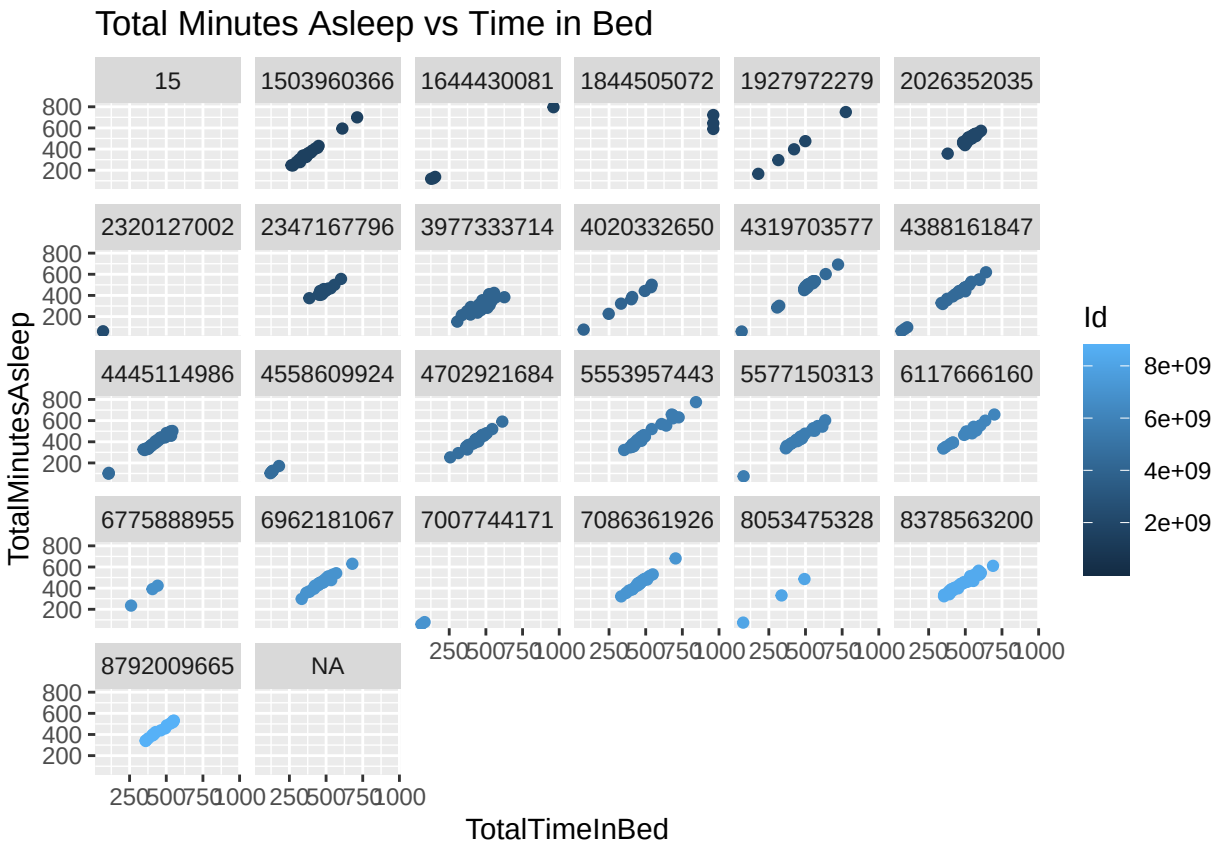
```
ggplot(data=sleep)+geom_smooth(mapping=aes(x=TotalTimeInBed, y=TotalMinutesAsleep))+geom_point(mapping=
```



As assumed, there is a strong positive correlation between the amount of time people spent in bed compared to the amount of time people were actually asleep, but there are are some points were it's obvious people were either having a difficult time falling asleep, or they may have been just laying down in bed on their phones or reading a book.

In addition, it seems the same people tend to have a similar sleeping pattern each night, but to see it more clearly, let's create more scatter plots.

```
ggplot(data=sleep)+geom_point(mapping=aes(x=TotalTimeInBed, y=TotalMinutesAsleep, color=Id))+facet_wrap
```



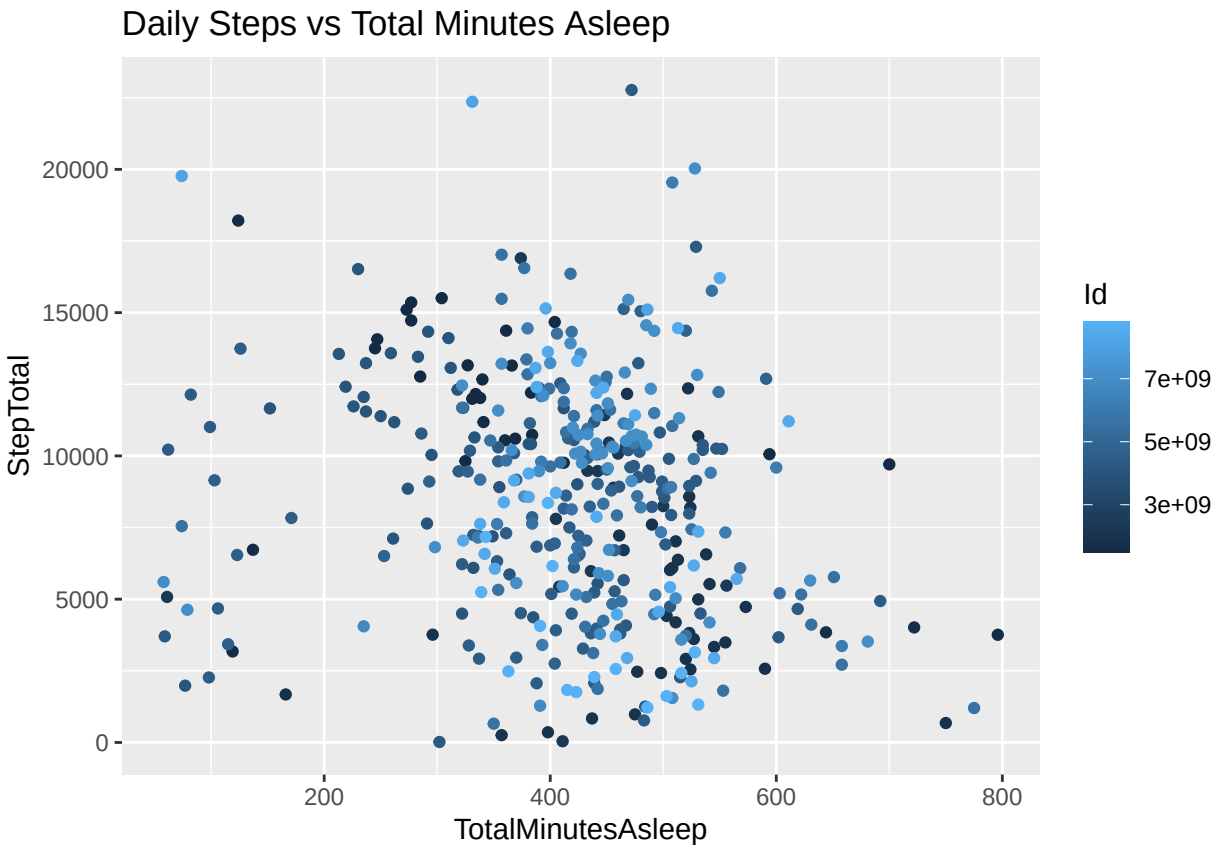
When looking closer at the data, it seems that some people have a more regular time in bed to sleeping time pattern than others.

Now I want to merge my data into larger tables in order to make calculations across the data.

```
merge_1 <- merge(calories, steps, by.Id = c("Date"))
merge_2 <- merge(sleep, steps, by.Id = c("Date"))
merge_3 <- merge(calories, sleep, by.Id = c("Date"))
```

Let's see the correlation between the amount of sleep the participants get per night and if that affects their daily steps.

```
ggplot(data=merge_2)+geom_point(mapping=aes(x=TotalMinutesAsleep, y=StepTotal, color=Id))+labs(title="Da
```

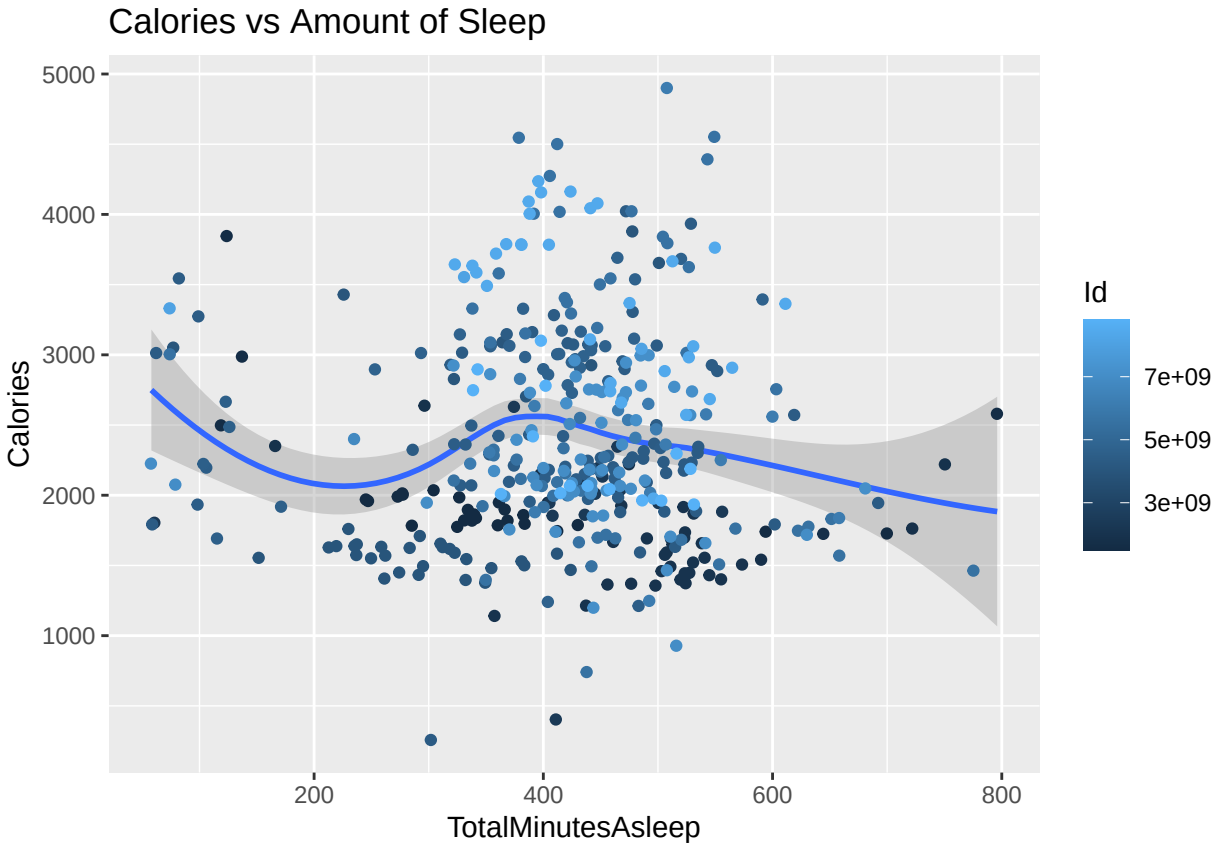


One would think that if a person received less sleep, they'd be more inclined to be less active, which would cause a decrease in daily steps. However, based on the graph, there is no correlation between how much sleep a person gets compared to how many steps they take in a day.

Now let's see if the amount of sleep a person gets a night has an affect on the number of calories they burn in a day.

```
ggplot(data=merge_3)+geom_smooth(mapping=aes(x=TotalMinutesAsleep, y=Calories)) +geom_jitter(mapping=aes
```

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

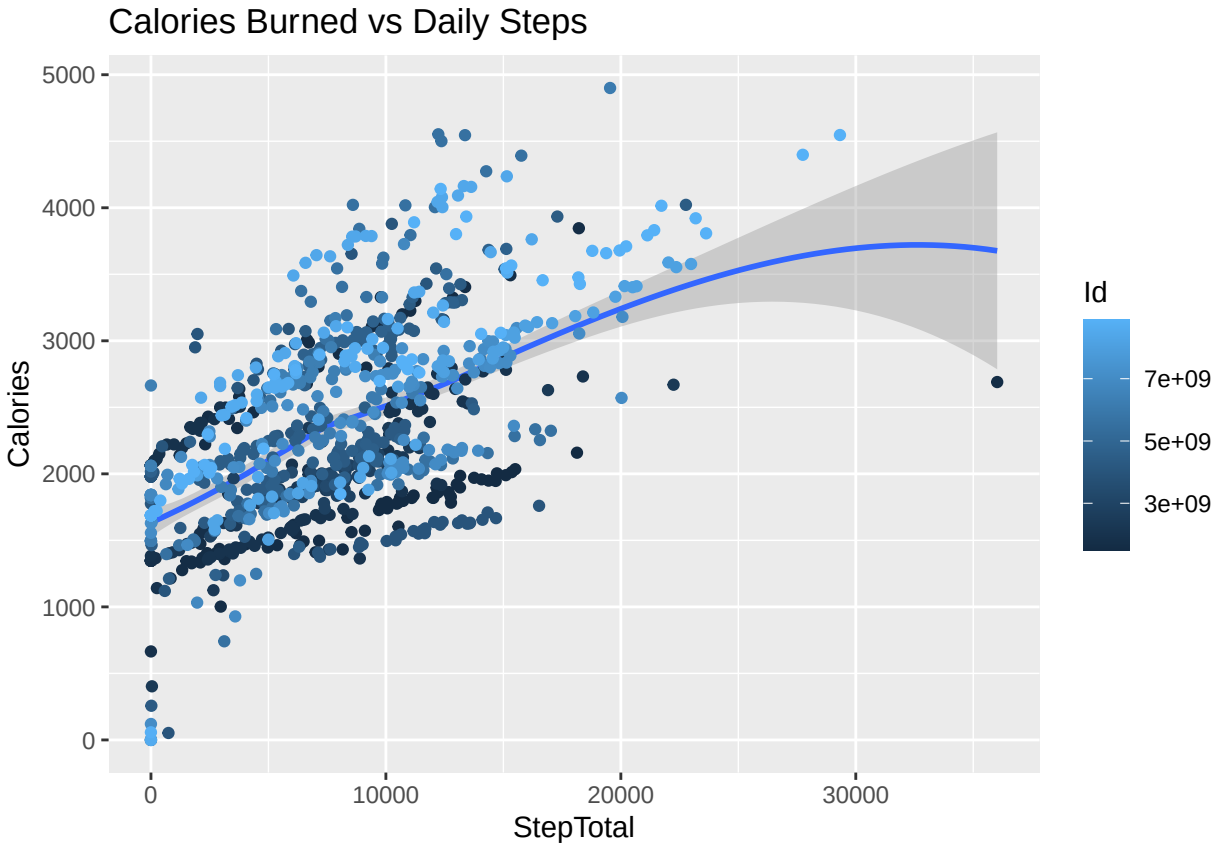


There is no correlation for a person's sleep changing the amount of calories they burn in a day, but it does seem like those who get between 6 to 9 hours of sleep burn the most calories.

Finally, let's see the relationship between the number of steps a person took, and the amount of calories they burned.

```
ggplot(data=merge_1)+geom_smooth(mapping=aes(x=StepTotal, y=Calories))+geom_jitter(mapping=aes(x=StepTo
```

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



As expected, there is a pretty strong correlation bewteen the number of steps a person takes in a day and the number of calories they burn. Some people probably burn more calories with less steps most likely due to the fact that they're running or jogging instead of walking. As stated previously, the zeros included in this data most likely show that the person wasn't wearing their device that day.

Phase 6: Act

My plan for a new marketing strategy. Overall, there are many ways that people utilize their smart devices. These devices can track heart rate, number of daily steps, and calories burned without the user having to lift a finger. As long as they're wearing their smart device, this data is automatically collected for them to view. However, as seen in the data, some users took their devices off, and they forgot to put them back on for days at a time.

In addition, there is some data that users can manually enter, such as their weight or sleep. However, these functions seem less popular as fewer people included in this dataset seemed to keep track of this data. This can be due to multiple reasons including they didn't know it was a function their device was capable of, they didn't know how to use it, or they just weren't interested in tracking this information.

Based on these findings, I think that Bellabeat should market their device as lightweight and waterproof. As some participants may have taken it off when showering or swimming without knowing it is protected from water. In addition, marketing should focus on the features you have to manually enter. A sizable number of people in the dataset utilized the feature that kept track of their sleep, so it would be interesting to add a dream journal or mood feature to see how a sleep schedule affects these things, and to see if that increases the number of women who use these features. In order to increase the usage of the weight loss and sleep tracking feature, Bellabeat could add a notification to the device at night or early in the morning, so a person is reminded to input this data for the day.

Also, I think a feature that would be nice to have is a caloric intake tracker. With this feature, you'd input what you ate for the day in your phone, and then you can compare how many calories were taken in to how

many calories you burned based on activity and steps! Given their market is women's health, there should also be an option to track your menstrual cycle with their device, so women can always be prepared for that time of the month.