CAPSTONE_Bellabeat

Tracey Johnson

21 March 2022

Introduction

Bellabeat is a high-tech manufacturer of health-focused products for women. This is a case study designed to review ~30 smart device's user data for insights into how consumers are using smart devices for potential market growth opportunities.

Business Task

1. What are some trends in smart device usage?
2. How could these trends apply to Bellabeat customers?
3. How could these trends help influence Bellabeat marketing strategy?

Data Sources

Fitbit Fitness Tracker Data link Data source presented by Case Study: (CC0: Public Domain, dataset made available through Mobius): This Kaggle data set contains personal fitness tracker from thirty fitbit users. Thirty eligible Fitbit users consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. It includes information about daily activity, steps, and heart rate that can be used to explore users' habits.

Additional articles were referenced: Exercise.com "Reasons Why People Don't Exercise" link; UW School of Medicine and Public Health "Is 20 Minutes of Exercise Enough" link; and BBC.com "How 'survivorship bias' can cause you to make mistakes" link.

Installing Packages

Install tidyverse and janitor packages to review and clean data.

```
install.packages("tidyverse")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.1'
## (as 'lib' is unspecified)
```

```
library(tidyverse)
```

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

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

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

```
install.packages("janitor")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.1'
## (as 'lib' is unspecified)
```

```
library(janitor)
```

```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

Uploading Datasets

Download dataset provided by Bellabeat and upload into RStudio.

```
library(readr)
dailyActivity_merged <- read.csv("Fitabase Data 4.12.16-5.12.16/dailyActivity_merged.csv")
```

Create Dataframe

Dataset "dailyActivity_merged" is a long title when used in analysis. I created a new dataframe (df) with a simple title "x".

```
x <- dailyActivity_merged
```

Get to know the data

In this section I used various methods to review details of the dataset. At the same time, I'm reviewing data to see if there are specific elements I might need to clean later.

```
head(x)
```

```
##           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
```

```
str(x)
```

```
## 'data.frame':    940 obs. of  15 variables:
##  $ Id                       : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ ActivityDate             : chr  "4/12/2016" "4/13/2016" "4/14/2016" "4/15/2016" ...
##  $ TotalSteps               : int  13162 10735 10460 9762 12669 9705 13019 15506 10544 9819 ...
##  $ TotalDistance            : num  8.5 6.97 6.74 6.28 8.16 ...
##  $ TrackerDistance          : num  8.5 6.97 6.74 6.28 8.16 ...
##  $ LoggedActivitiesDistance: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ VeryActiveDistance       : num  1.88 1.57 2.44 2.14 2.71 ...
##  $ ModeratelyActiveDistance: num  0.55 0.69 0.4 1.26 0.41 ...
##  $ LightActiveDistance      : num  6.06 4.71 3.91 2.83 5.04 ...
##  $ SedentaryActiveDistance : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ VeryActiveMinutes        : int  25 21 30 29 36 38 42 50 28 19 ...
##  $ FairlyActiveMinutes      : int  13 19 11 34 10 20 16 31 12 8 ...
##  $ LightlyActiveMinutes     : int  328 217 181 209 221 164 233 264 205 211 ...
##  $ SedentaryMinutes         : int  728 776 1218 726 773 539 1149 775 818 838 ...
##  $ Calories                 : int  1985 1797 1776 1745 1863 1728 1921 2035 1786 1775 ...
```

```
glimpse(x)
```

```
## Rows: 940
## Columns: 15
## $ Id                       <dbl> 1503960366, 1503960366, 1503960366, 150396036~
## $ ActivityDate             <chr> "4/12/2016", "4/13/2016", "4/14/2016", "4/15/~
## $ TotalSteps               <int> 13162, 10735, 10460, 9762, 12669, 9705, 13019~
## $ TotalDistance            <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9.8~
## $ TrackerDistance          <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9.8~
## $ LoggedActivitiesDistance <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ VeryActiveDistance       <dbl> 1.88, 1.57, 2.44, 2.14, 2.71, 3.19, 3.25, 3.5~
## $ ModeratelyActiveDistance <dbl> 0.55, 0.69, 0.40, 1.26, 0.41, 0.78, 0.64, 1.3~
## $ LightActiveDistance      <dbl> 6.06, 4.71, 3.91, 2.83, 5.04, 2.51, 4.71, 5.0~
## $ SedentaryActiveDistance  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ VeryActiveMinutes        <int> 25, 21, 30, 29, 36, 38, 42, 50, 28, 19, 66, 4~
## $ FairlyActiveMinutes      <int> 13, 19, 11, 34, 10, 20, 16, 31, 12, 8, 27, 21~
## $ LightlyActiveMinutes     <int> 328, 217, 181, 209, 221, 164, 233, 264, 205, ~
## $ SedentaryMinutes         <int> 728, 776, 1218, 726, 773, 539, 1149, 775, 818~
## $ Calories                 <int> 1985, 1797, 1776, 1745, 1863, 1728, 1921, 203~
```

Clean Data

First, ensure the names of observations are clean.

```
x <- clean_names(x)
```

Next, remove any columns or rows that are entirely empty.

```
x <- remove_empty(x, which = c("rows", "cols"))
```

Load dplyr for the %>% pipe feature.

```
library(dplyr)
```

I used the janitor tabyl feature to review data IDs. This told me two things, it told me the number of individual users in the dataset and the frequency of the activity entries by individual users over the duration of the dataset collected.

```
x %>%
  tabyl(id) %>%
  adorn_pct_formatting(digits = 0, affix_sign = TRUE)
```

```
##          id  n percent
##  1503960366 31      3%
##  1624580081 31      3%
##  1644430081 30      3%
##  1844505072 31      3%
##  1927972279 31      3%
##  2022484408 31      3%
##  2026352035 31      3%
##  2320127002 31      3%
##  2347167796 18      2%
##  2873212765 31      3%
##  3372868164 20      2%
##  3977333714 30      3%
##  4020332650 31      3%
##  4057192912  4      0%
##  4319703577 31      3%
##  4388161847 31      3%
##  4445114986 31      3%
##  4558609924 31      3%
##  4702921684 31      3%
##  5553957443 31      3%
##  5577150313 30      3%
##  6117666160 28      3%
##  6290855005 29      3%
##  6775888955 26      3%
##  6962181067 31      3%
##  7007744171 26      3%
##  7086361926 31      3%
##  8053475328 31      3%
##  8253242879 19      2%
##  8378563200 31      3%
##  8583815059 31      3%
##  8792009665 29      3%
##  8877689391 31      3%
```

Lastly for cleaning, lets check for any duplicate entries.I already know there are multiple entries for each *id* (or user), but they should all have no more than one entry per *activity_date*. So in this case I'm using the janitor package "get_dupes" for columns *id* and *activity_date* to verify.

```
x %>% get_dupes(id, activity_date)
```

```
## No duplicate combinations found of: id, activity_date
```

```
##  [1] id                       activity_date
##  [3] dupe_count               total_steps
##  [5] total_distance           tracker_distance
```

```
##  [7] logged_activities_distance very_active_distance
##  [9] moderately_active_distance light_active_distance
## [11] sedentary_active_distance  very_active_minutes
## [13] fairly_active_minutes      lightly_active_minutes
## [15] sedentary_minutes          calories
## <0 rows> (or 0-length row.names)
```

Analyze and Visualize Data

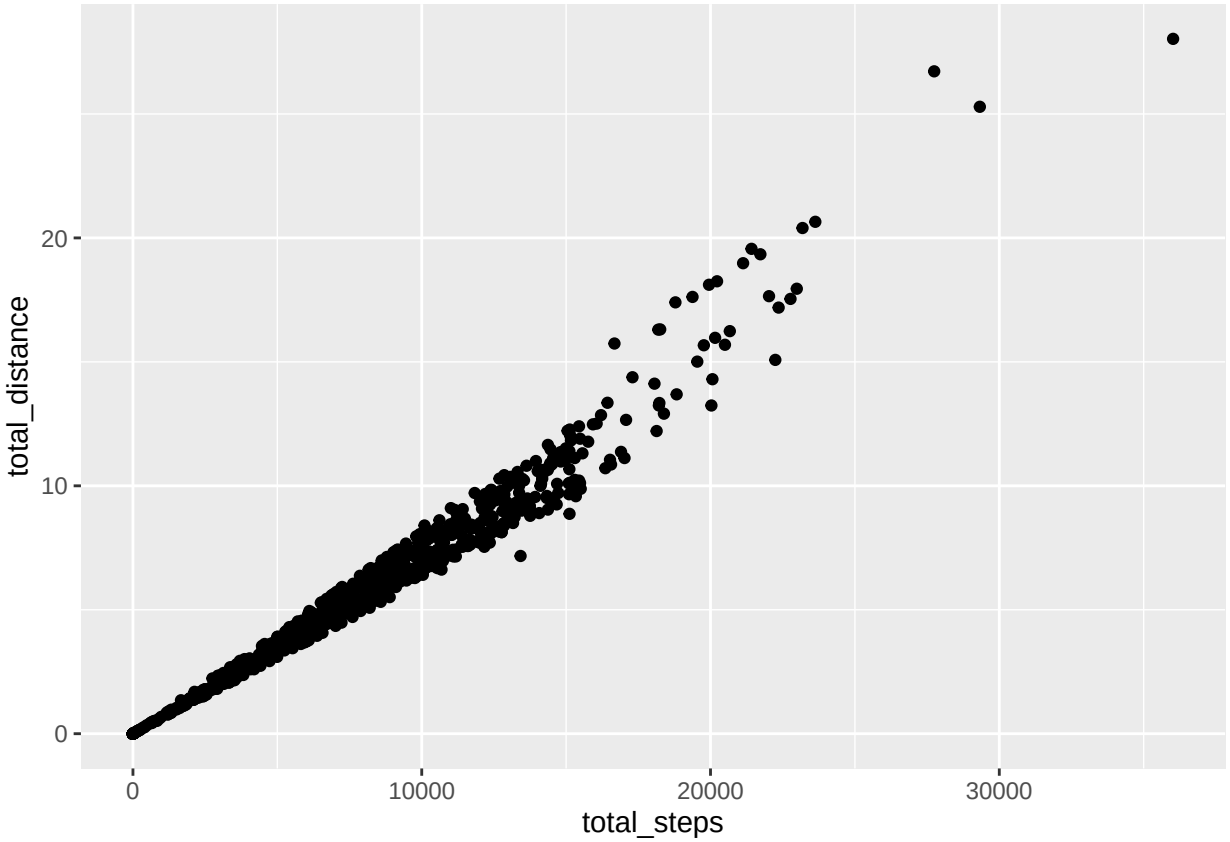
Now that we've cleaned our data, it's time to analyze. I'm a visual person. By changing the point of view in how we look at the data, we can see different aspects and hopefully, in this instance, help us identify some potential areas of growth for Bellabeat.

Let's use the ggplot2 package.

```
library(ggplot2)
```

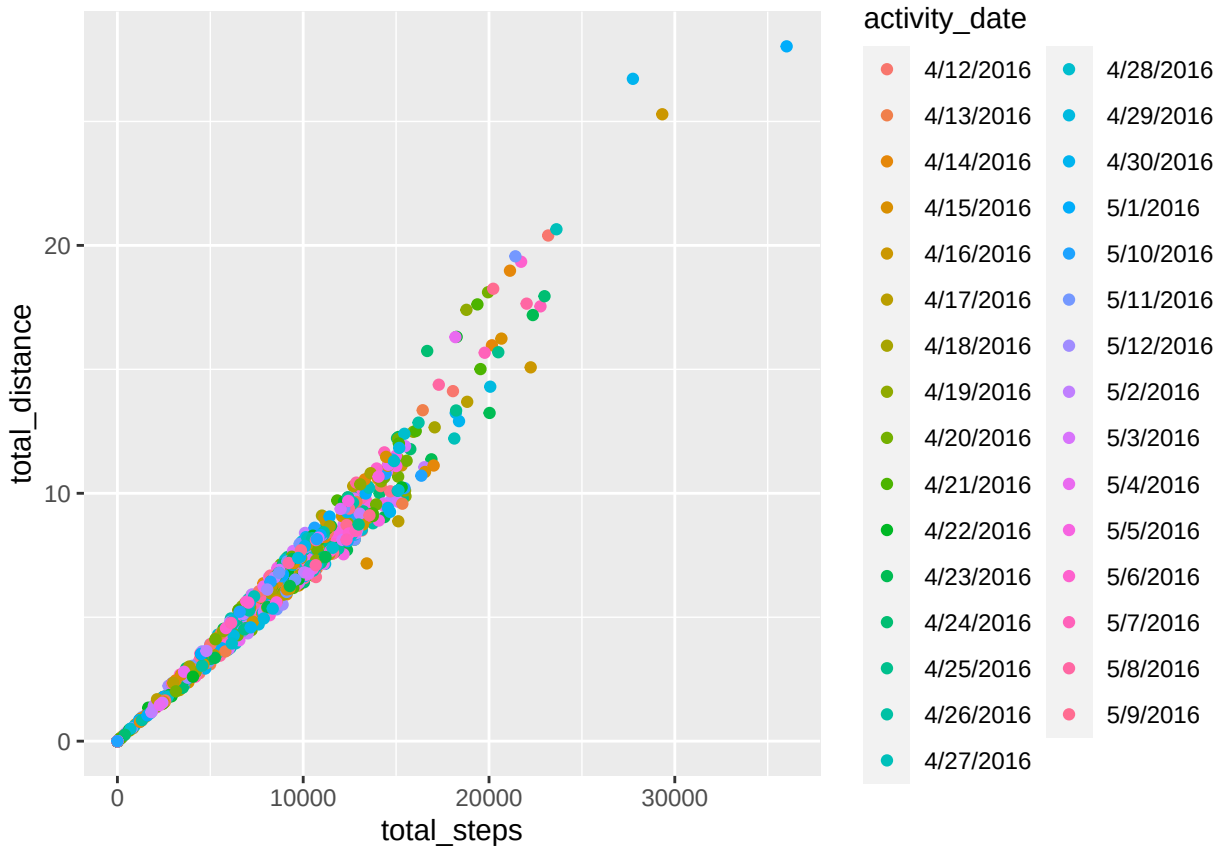
Next, I want to see the relationship between total steps and total distance.

```
ggplot(data = x) + geom_point(mapping = aes(x = total_steps, y = total_distance))
```



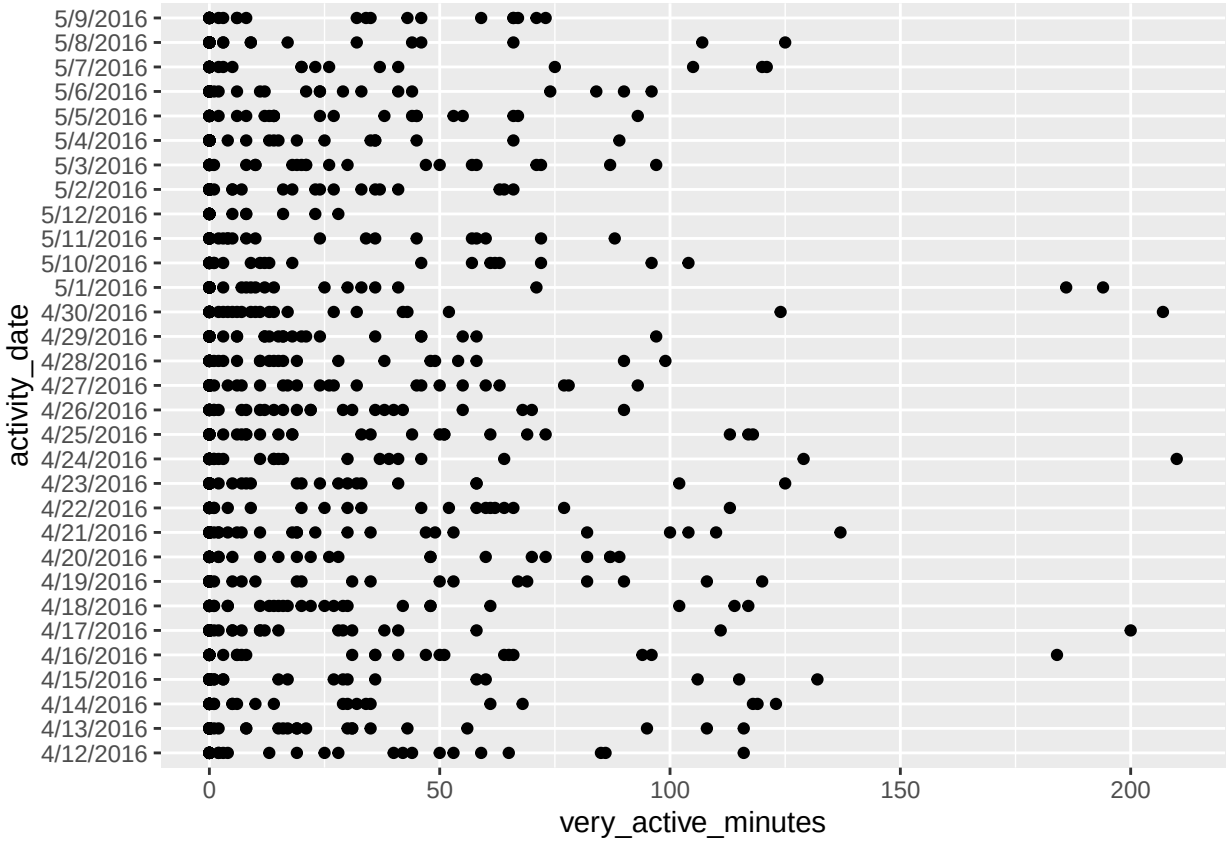
This initial graph shows that there is a positive relationship between the total_steps and total_distance. In other words, the more steps the greater the distance. I'd also like to view this by the time of the month. Are certain periods more active than others? I'm adding a color by month, to see if that identifies more active periods.

```
ggplot(data = x) + geom_point(mapping = aes(x = total_steps, y = total_distance, color = activity_date))
```



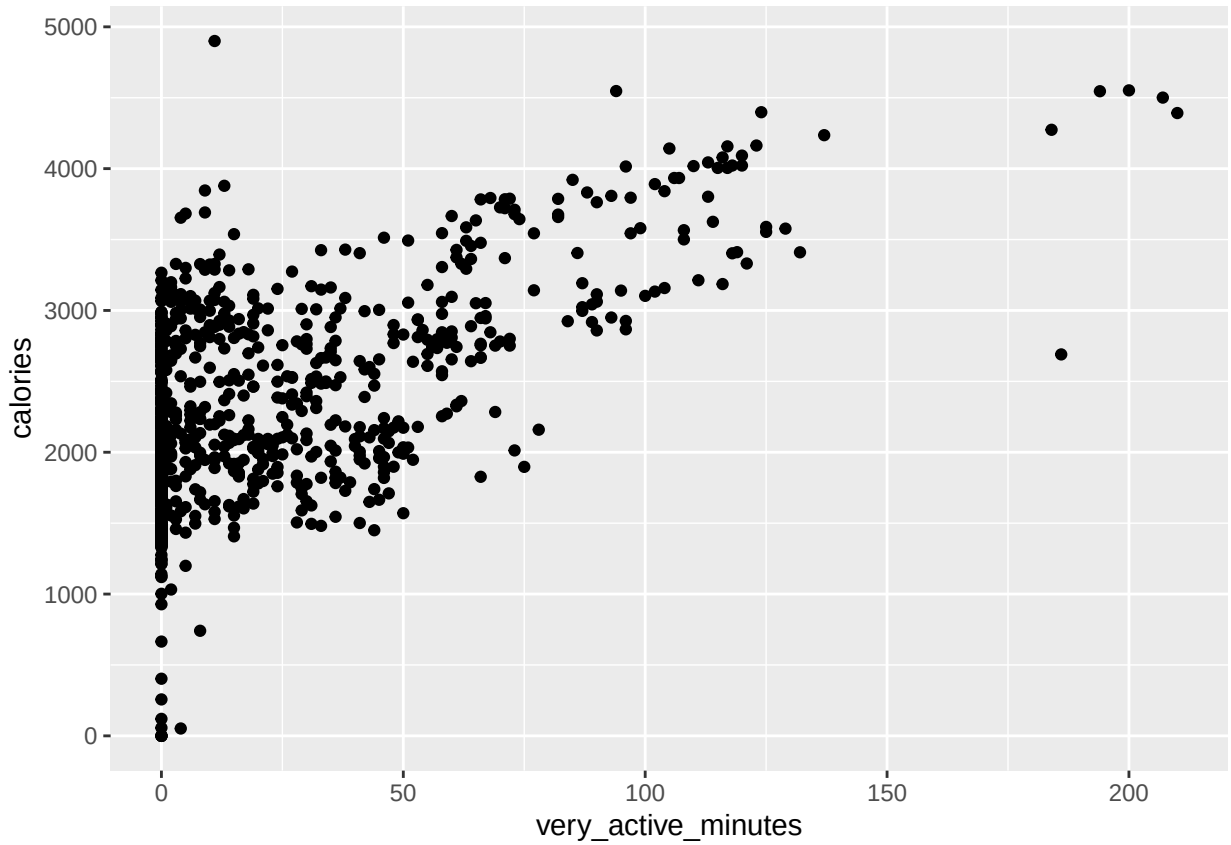
While very pretty, it doesn't clearly indicate any relationship or trends based on dates other than what we already know. The trend shows that people are initially very diligent with their logging of data and meeting their initial goals, but as time passes their initiative wanes. Let's look at activity_date and very_active_minutes.

```
ggplot(data = x) + geom_point(mapping = aes(x = very_active_minutes, y = activity_date))
```



This shows similar outputs, very active in initial stages, slowly ebbing as the active minutes increases. Another view was to look at *very_active_minutes* by *calories*.

```
ggplot(data = x, mapping = aes(x = very_active_minutes, y = calories)) + geom_point()
```



I next wanted to find the average of *very_active_minutes* to see how many minutes users were most active. Looking back to our "Get to know the data" section we see that *very_active_minutes* is an integer observation and mean functions only work with numerics, so in order to calculate average, I first need to convert data from integer to numeric and then average that data. I didn't want to permanently change in my dataset, so I just worked with one column and renamed to X1_numeric.

```
X1_numeric <- as.numeric(as.character(x$very_active_minutes))
```

```
mean(X1_numeric)
```

```
## [1] 21.16489
```

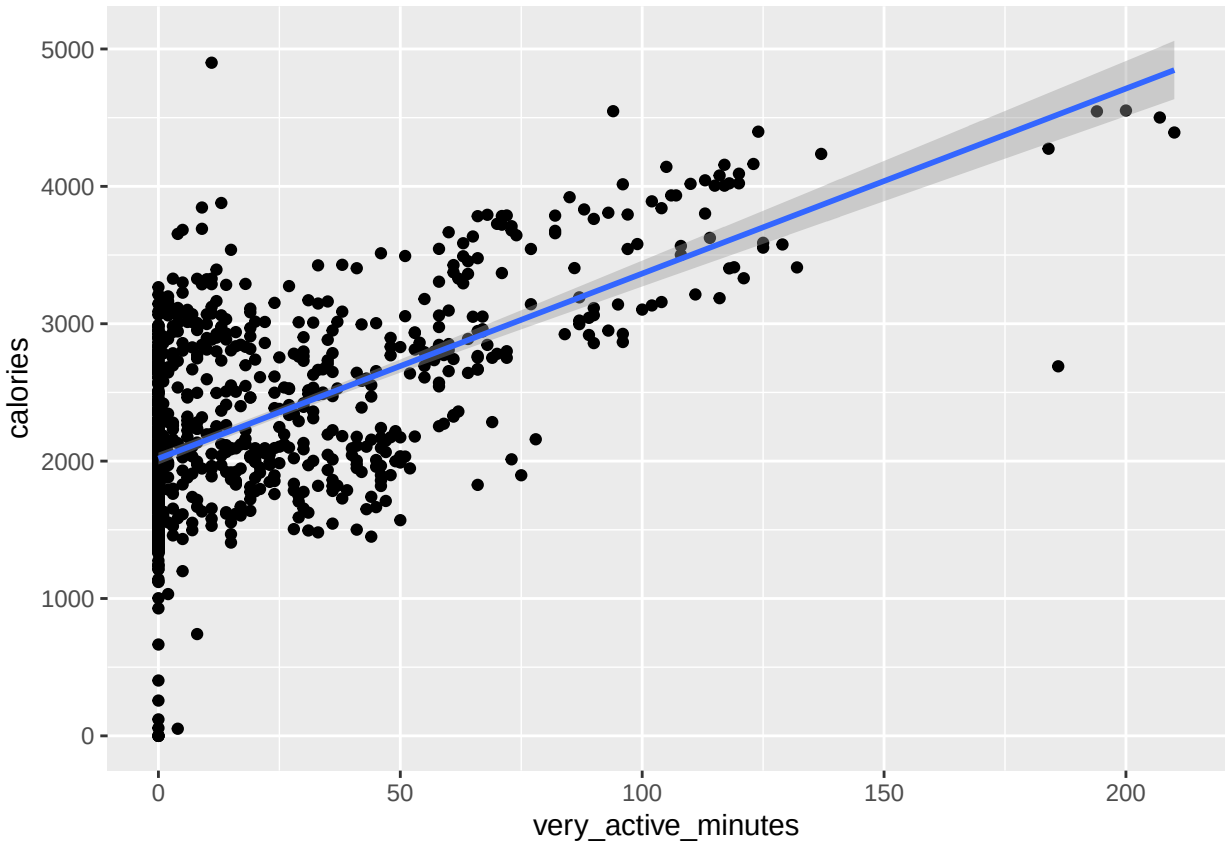
So this says that the average *very_active_minutes* was a little over 20 minutes. According to exercise.com there are three main reasons why people do not exercise:

"The leading reasons include a lack of motivation, lack of funds, and lack of time"

This initial insight, supports growth opportunities in marketing towards time requirements and motivation. Dataset showed an average of 20 minutes will return positive results. According to the University of Wisconsin School of Medicine and Public Health, *"Yes, 20 minutes of exercise is better than nothing. Any and every bout of physical activity/exercise contributes to a fitter, healthier - and, very likely, happier - you!"* https://www.uwhealth.org/news/is-20-minutes-of-exercise-enough

To follow-up on this potential trend, let's look to see if there is a correlation between *very_active_minutes* and *calories* burned, and include a trend line.

```
ggplot(x, aes(very_active_minutes, calories)) + geom_point() + geom_smooth(formula = y ~ x, method = "lm
```



The analysis so far was reviewing data over the total 31 days of the dataset. However we did notice a significant portion of those that did not complete the full 31 days, another potential correlation to the "lack of time" or "lack of motivation" identified by the exercise.com research article.

One caveat to this analysis is that as we're only looking at current smart device users, we potentially could be committing survivorship or selection bias, by only selecting or analyzing the data 'survivors', or those that completed the devise usage. I think there is additional opportunities for growth if we were to further investigate non-smart device users or focused analysis on those that didn't complete 31 days in the study.

Summary

Back to the Business Tasks

1. What are some trends in smart device usage?

Review of the data showed a positive correlation between smart device usage and exercise factors of time and motivation. Research into these factors identified that time and motivation were two leading causes of why people don't exercise, thus negatively impacting general health and personal happiness.

2. How could these trends apply to Bellabeat customers?

Based on the data analysis executed, the analysis team anticipates similar results in a review of Bellabeat's current customer's use of Bellabeat products, with particular focus on *Bellabeat Leaf* and *Bellabeat app*. The more time it takes to execute tasks reduces the overall likeliness of customer completion of personal goals, while also further reducing their motivation to execute tasks.

3. How could these trends help influence Bellabeat marketing strategy?

Marketing should focus on:

- improving **customer motivation** to complete and achieve health and wellness goals
- a minimum **time investment** of 20 minutes can significantly improve overall health and wellness