Google Data Analytics Capstone FV

Balen Younis

8/9/2021

Notice from Author:

As I am starting my journey as a data analyst, this is my first case study, and I've read most other analyses on this case, and basically, they are too close in the result and assumption. There is not much gap left to work on it, and I did not try to be creative and come up with something new. But specific points drew my attention, and I'd like to mention some of them here. In this case study, we are dealing with a sample, but we don't know the stakeholder's aim regarding the confidence level, customer population size, standard deviation, and margin of error. So, we cannot determine what the right sample size is? In real life, we must handle this issue to come up with the best conclusion that may help stakeholders to make the right decision.

The Goal of This Case Study:

The aim of this project to help Bellabeat wellness tracker company to improve their marketing strategy by collecting data from non-Bellabeat consumer such as Fitbit users of smart device data, this is to analyze their usage and gain insights on to how these costumers are using their devices, what features do they like the most These insights will be applied on promoting Bellbeat devices. Then accordingly we will provide recommendations to Bellabeat marketing team based on noticed trends.

Import FitBit Data

By using R studio cloud, you will have no access to importing data from your PC directly, so click on the tab file and then upload the file first, then import it; out of 18 CSV files.

- dailyActivity_merged.csv
- dailyCalories_merged.csv
- dailyIntensities_merged.csv
- sleepDay_merged.csv
- weightLogInfo_merged.csv

all these data were checked as a raw material (spreadsheet) for cleanup, consistency, null, missing data. I also tested using the data and separated them to data and time, but they made no impact on the result, so I keep using the current format data, time, second.

Installed, loaded, and Used Packages and Functions:

- tidyverse
- head() , here() , colnames() , glimpse()
- str(), n-distinct , nrow()
- Summary(), Select(), merge()
- ggplot(), geom_point(),geom_smooth(), plot()

Followed Steps:

- Using read.csv to load five flies to start to work on

- Since this is the 1st project with R stoduio I prefered to use Head and colnames for all files and see the result.
- How many unique participants are there in each dataframe?
- To check number of observations in each dataframe

Summray of Code Chunks:

```
library("tidyverse")
```

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

```
## v ggplot2 3.3.5     v purrr   0.3.4
## 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
```

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

Read and Load files:

```
daily_activity <- read.csv("dailyActivity_merged.csv")
daily_calories <- read.csv("dailyCalories_merged.csv")
sleep_day <- read.csv("sleepDay_merged.csv")
daily_intensities <- read.csv("dailyIntensities_merged.csv")
weight_info <- read.csv("weightLogInfo_merged.csv")
```

Head and Colnames

```
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
```

```
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"
```

```
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
```

```
colnames(daily_calories)
```

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

```
head(daily_intensities)
```

```
##           Id ActivityDay SedentaryMinutes LightlyActiveMinutes
## 1 1503960366   4/12/2016              728                  328
## 2 1503960366   4/13/2016              776                  217
## 3 1503960366   4/14/2016             1218                  181
## 4 1503960366   4/15/2016              726                  209
## 5 1503960366   4/16/2016              773                  221
## 6 1503960366   4/17/2016              539                  164
##   FairlyActiveMinutes VeryActiveMinutes SedentaryActiveDistance
## 1                  13                25                       0
## 2                  19                21                       0
## 3                  11                30                       0
## 4                  34                29                       0
## 5                  10                36                       0
## 6                  20                38                       0
##   LightActiveDistance ModeratelyActiveDistance VeryActiveDistance
## 1                6.06                     0.55               1.88
## 2                4.71                     0.69               1.57
## 3                3.91                     0.40               2.44
## 4                2.83                     1.26               2.14
## 5                5.04                     0.41               2.71
## 6                2.51                     0.78               3.19
```

```
colnames(daily_intensities)
```

```
##  [1] "Id"                       "ActivityDay"
##  [3] "SedentaryMinutes"         "LightlyActiveMinutes"
##  [5] "FairlyActiveMinutes"      "VeryActiveMinutes"
##  [7] "SedentaryActiveDistance"  "LightActiveDistance"
##  [9] "ModeratelyActiveDistance" "VeryActiveDistance"
```

```
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
```

```
colnames(sleep_day)
```

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

```
head(weight_info)
```

```
##           Id                   Date WeightKg WeightPounds Fat   BMI
## 1 1503960366  5/2/2016 11:59:59 PM     52.6     115.9631  22 22.65
## 2 1503960366  5/3/2016 11:59:59 PM     52.6     115.9631  NA 22.65
## 3 1927972279  4/13/2016 1:08:52 AM    133.5     294.3171  NA 47.54
## 4 2873212765 4/21/2016 11:59:59 PM     56.7     125.0021  NA 21.45
## 5 2873212765 5/12/2016 11:59:59 PM     57.3     126.3249  NA 21.69
## 6 4319703577 4/17/2016 11:59:59 PM     72.4     159.6147  25 27.45
##   IsManualReport        LogId
## 1           True 1.462234e+12
## 2           True 1.462320e+12
## 3          False 1.460510e+12
## 4           True 1.461283e+12
## 5           True 1.463098e+12
## 6           True 1.460938e+12
```

```
colnames(weight_info)
```

```
## [1] "Id"             "Date"           "WeightKg"       "WeightPounds"
## [5] "Fat"            "BMI"            "IsManualReport" "LogId"
```

Combine Sleep_day and daily_activity

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

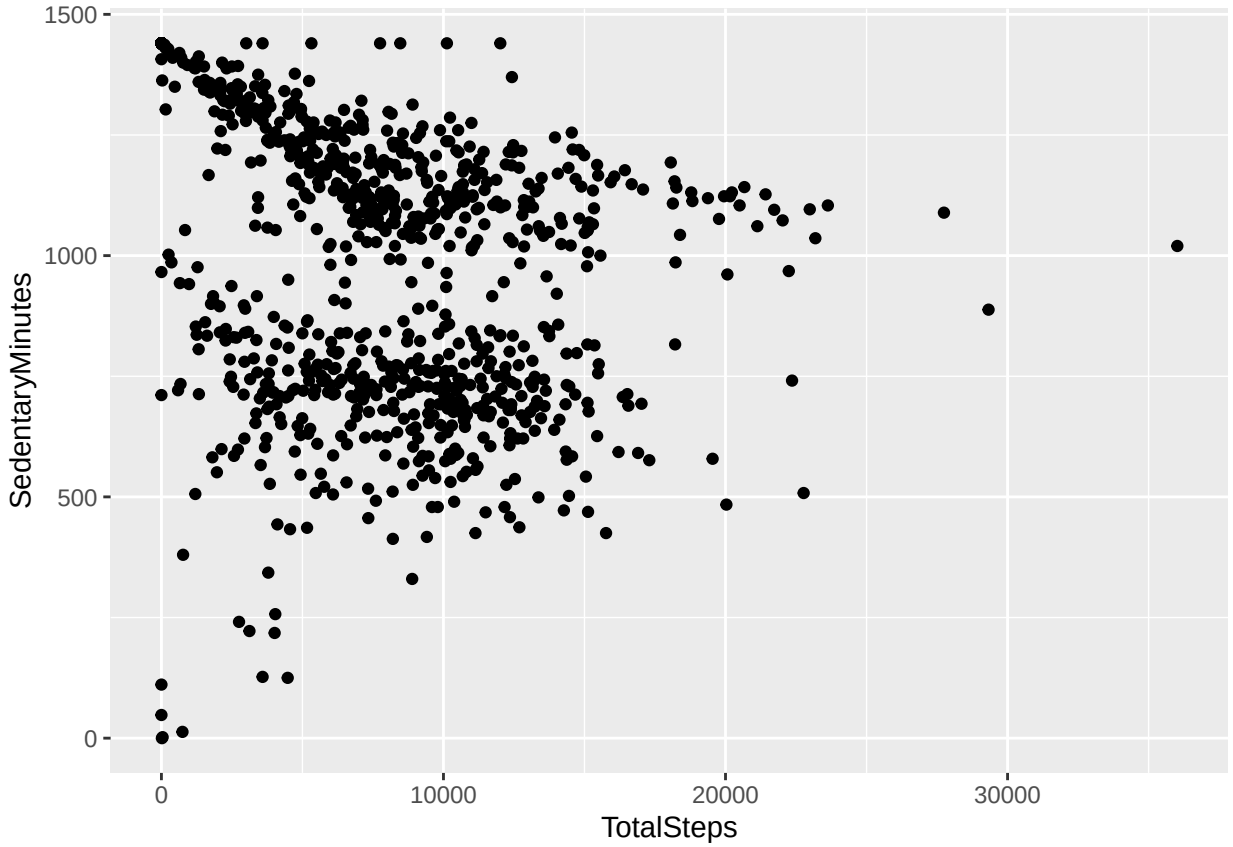
```
distance_weight <- merge(daily_activity,weight_info, by="Id", all=TRUE)
n_distinct(distance_weight$Id)
```

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

Plots and Visualization

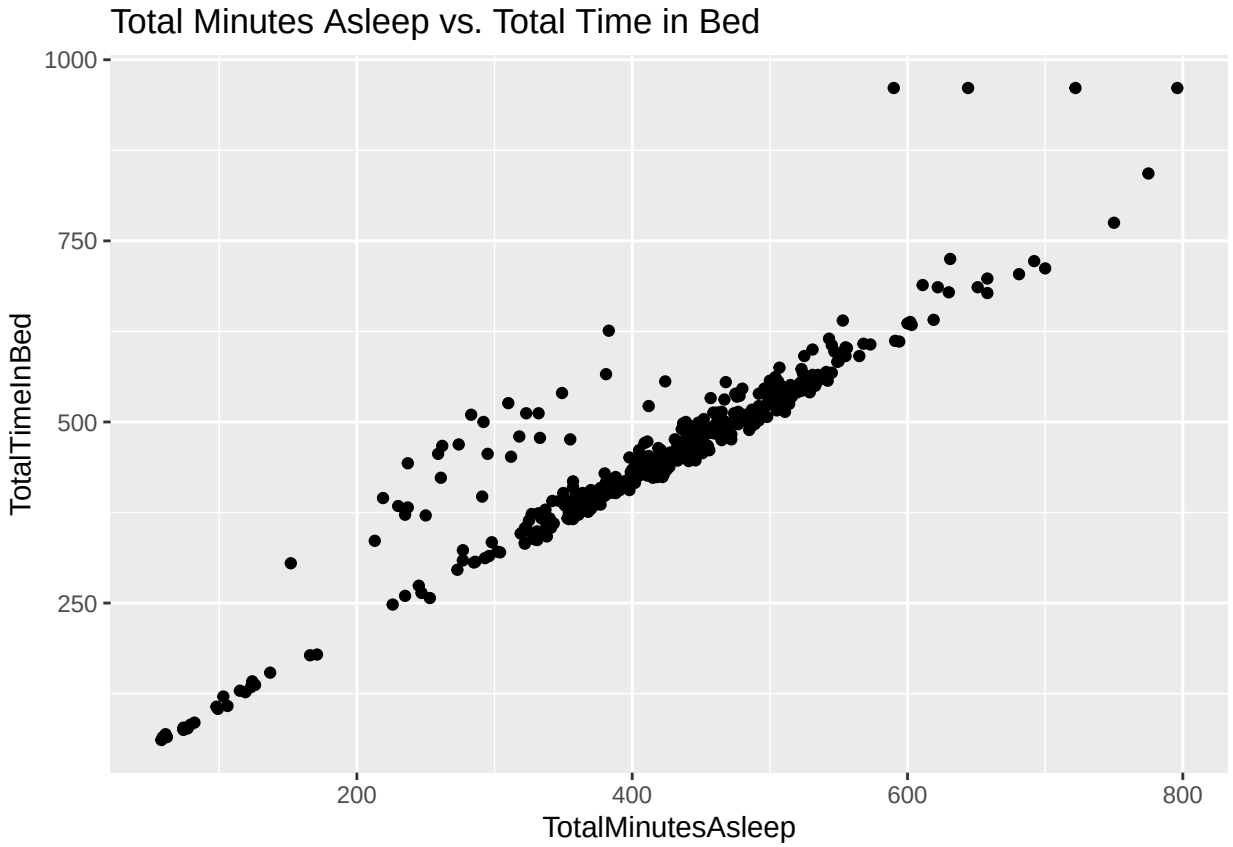
Showing the relation between total steps and sedentary minutes is a positive relation, as it's revealed the more steps walking, the less sedentary time we get.

```
ggplot(data=daily_activity)+
  geom_point(mapping=aes(x=TotalSteps, y=SedentaryMinutes))
```



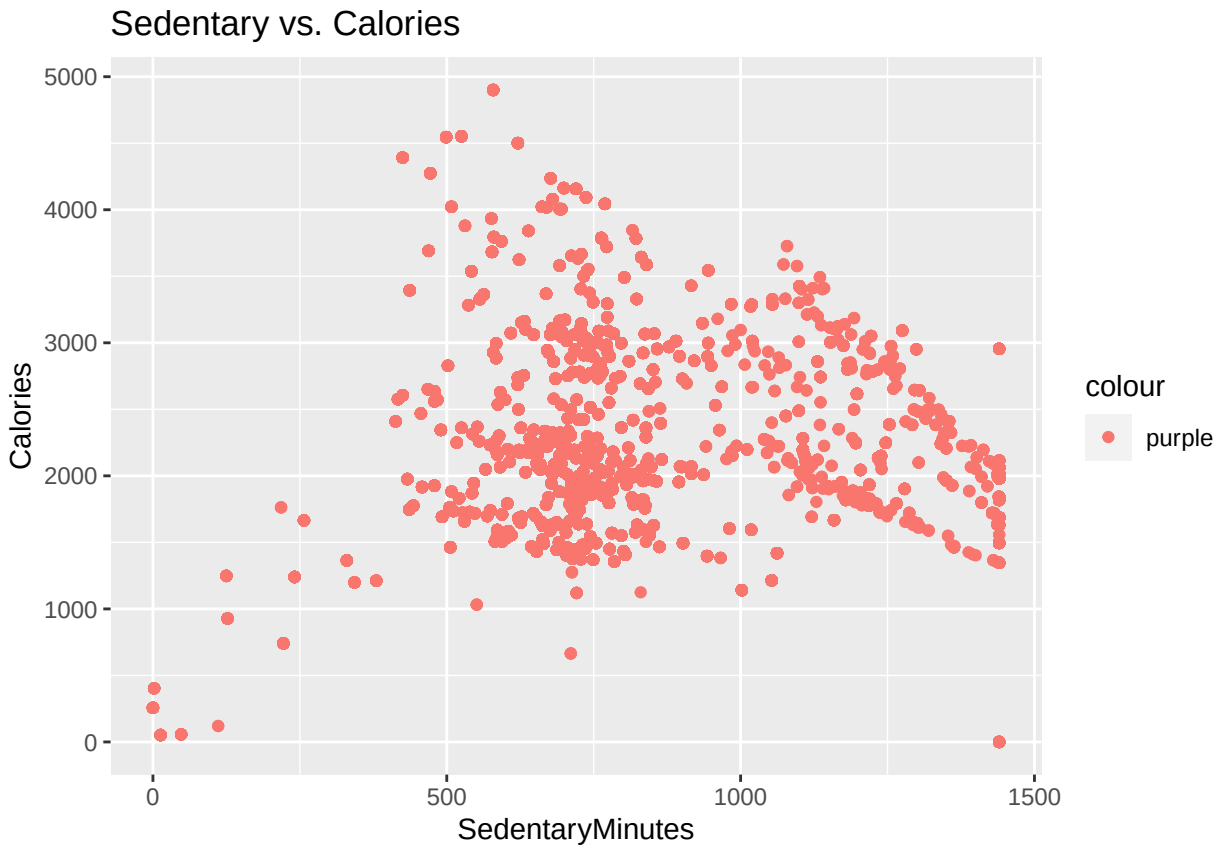
Positive relation more sleep more time in bed, that it to check we are having right data, obviously the outcome must be positive.

```
ggplot(data=sleep_day, aes(x=TotalMinutesAsleep, y=TotalTimeInBed)) +
  geom_point()+ labs(title="Total Minutes Asleep vs. Total Time in Bed")
```



Sedentary time and calories are not showing positive or negative relations; the smart device data keeps recording burning calories even when customers are in sedentary time. Many research shows that you can burn calories even at idle times. Metabolize processes are different from one person to another. The human body can record a random level of calorie burning.

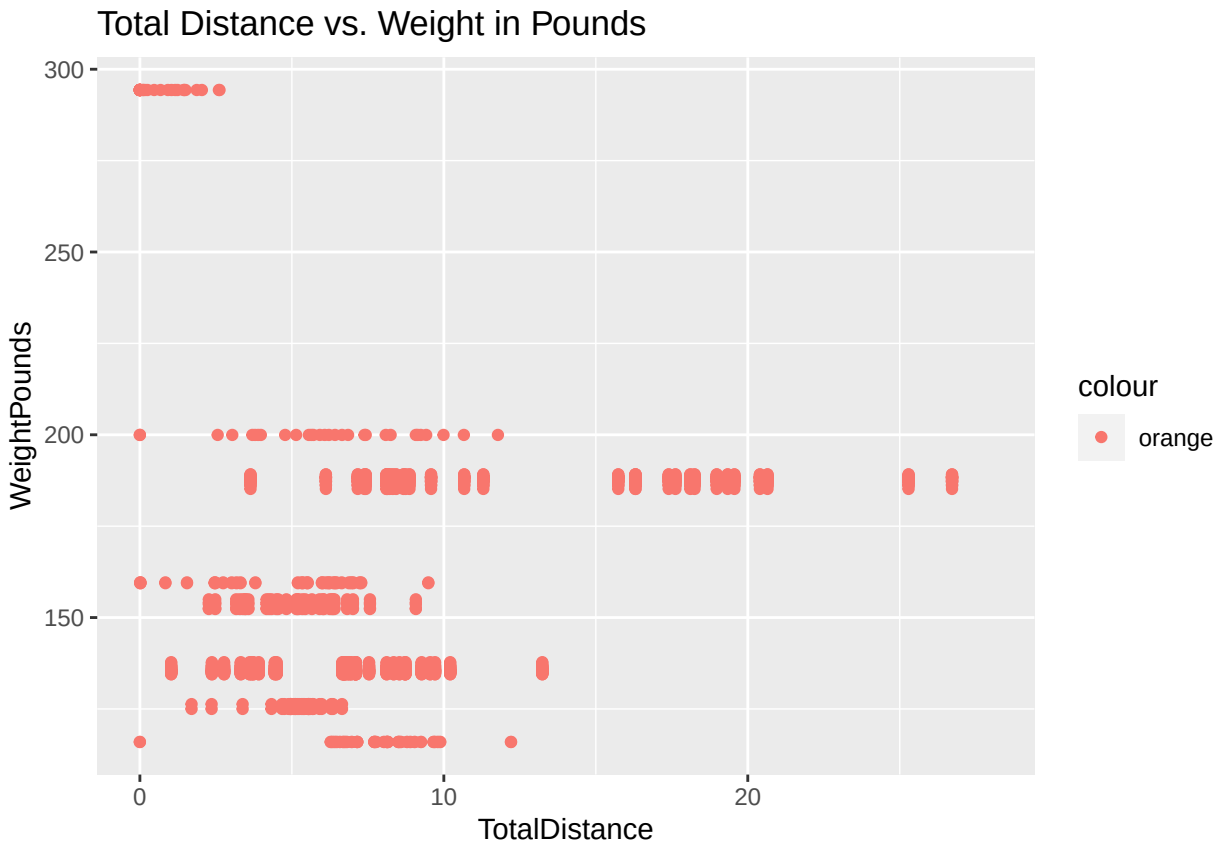
```
ggplot(data=combined_data, aes(x=SedentaryMinutes, y=Calories, color ="purple")) +
  geom_point()+ labs(title="Sedentary vs. Calories")
```



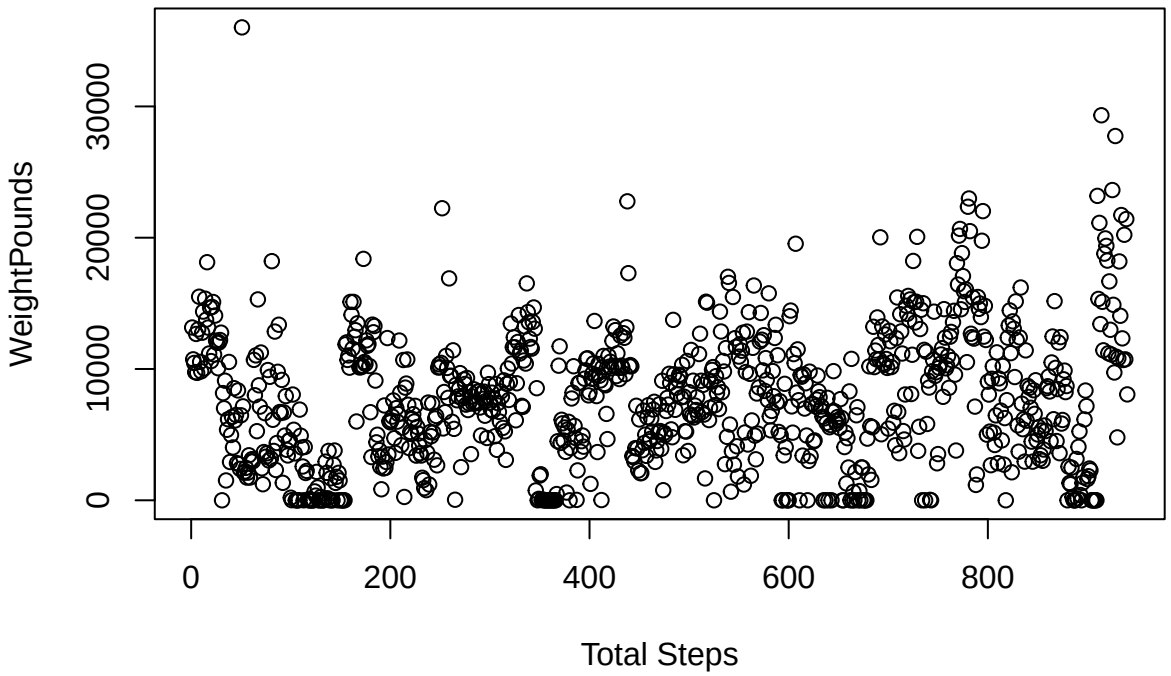
The relation between weight and distance is not clear. The general trend is that less weight can go the further distance, but since there are third factors, the walking speed significantly contributes to the result. This factor is missing in this correlation.

```
ggplot(data=distance_weight, aes(x=TotalDistance, y=WeightPounds, color ='orange')) +
  geom_point()+ labs(title="Total Distance vs. Weight in Pounds")
```

```
## Warning: Removed 693 rows containing missing values (geom_point).
```



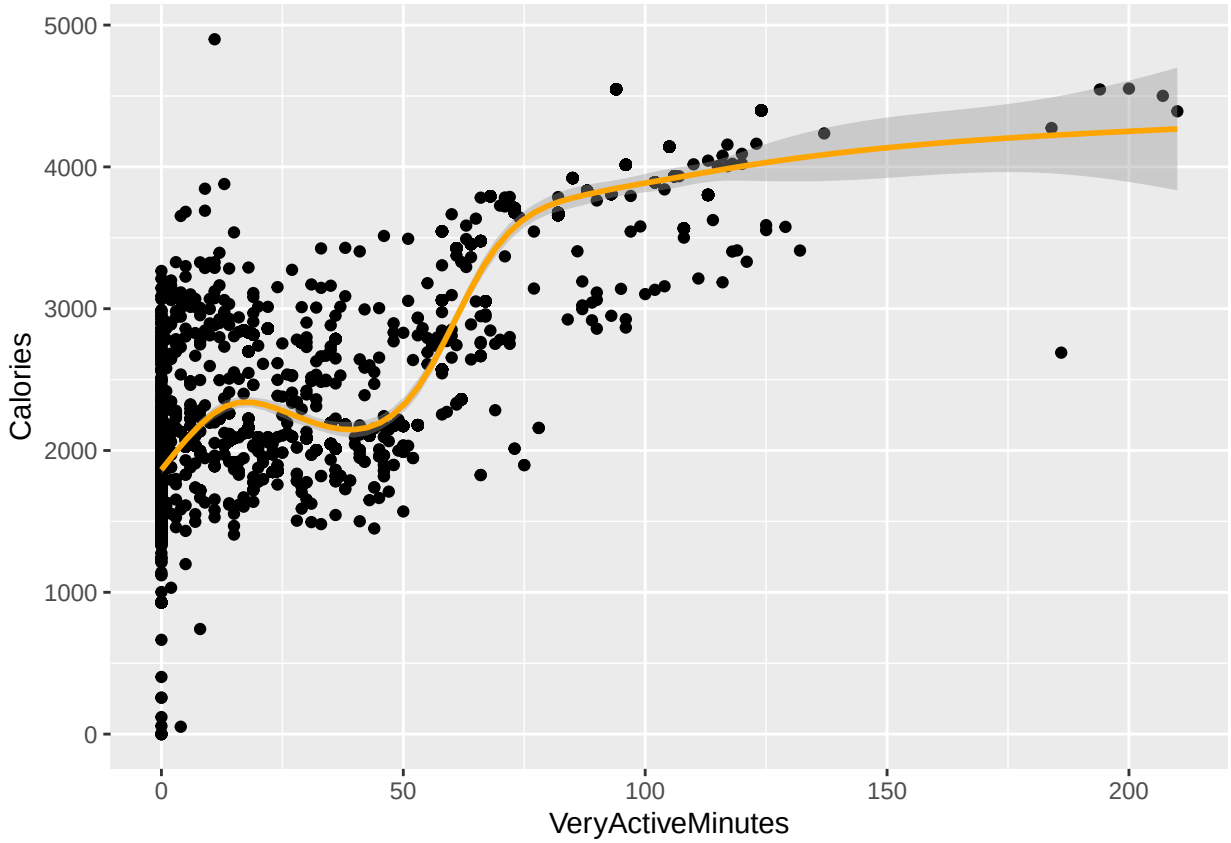

```
plot(daily_activity$TotalSteps, ylab="WeightPounds",
     xlab = "Total Steps" )
```



The longer duration you keep active and doing physical activity, the more calories you will burn

```
ggplot(data = distance_weight, aes(x=VeryActiveMinutes, y=Calories)) + geom_point() + geom_smooth(color=
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Recommendations for Bellabeat Business:

- Since they are targeting female customers, and most parts of their businesses are online. They have to make sure customers understand why Bellabeat product is different from other brands and focus on the healthy side. For instance, they can highlight the fact that:
- walking 20-30 minutes per mile lower the risk of diabetes.
- The high amount of sedentary time has been associated with a greater risk of diabetes.
- More time in bed and sedentary leads to high cholesterol, weight gain, memory loss, and depression.
- We recommend that their smart device collect more categories of data; for instance, for walking, we can have the pace of walking; the steep of road, the higher altitude, the more calories burnt.
- Targeting the right audience in the marketing campaign is essential for online business.
- Use customer feedback on their product to improve their product's good side and avoid what they do not like.