Case_Study

Matt Colbert

2023-06-04

This is my case study project for my Google Data Analytics Certificate. (DONE)

The goal of this project is to help a company called Bellabeat, come up with a marketing strategy by using data that is previously collected from smart devices to see how users are currently using fitness tracking devices. The data set (https://www.kaggle.com/datasets/arashnic/fitbit) contains the data of 30 users, which includes their activity, steps, heart rate, and sleep.

Bellabeat has three main products, as well as an app for users to better understand their health and current habits, and a membership which gives them access to guidance on activity, nutrition, sleep, and health. The three main products are the Spring, a water bottle that tracks water intake; the Time, which contains the looks of a classic timepiece while still tracking the users sleep, activity, and stress; and the Leaf, which can be worn as a necklace, bracelet, or clip to track activity, sleep, and stress.

The data set contains a group of 30 FitBit users who were tracked between April 12, 2016 and May 12, 2016. The data collected contains information about their sleep habits, physical activity, and heart rates. The physical activity is tracked by minutes, hours, and days.

Through this analysis, I hope to see more about the exercise habits of FitBit users, see what features of FitBits are used the most, and which features are not present with the hopes of coming up with a marketing strategy for Bellabeat.

Data Cleaning with Steps (DONE)

This can be changed to where the folder is with the data and the rmd file.

This step identifies the number of users who inputted their weight, as well as the number of times that they did so.

```
weight <- read.csv("./Case_Study_Data/weightLogInfo_merged.csv")
weight_by_id <- weight %>% group_by(Id)
sum_weight <- summarize(weight_by_id, count = n())
sum_weight
```

```
## # A tibble: 8 x 2
##           Id count
##        <dbl> <int>
## 1 1503960366     2
## 2 1927972279     1
## 3 2873212765     2
## 4 4319703577     2
## 5 4558609924     5
## 6 5577150313     1
## 7 6962181067    30
## 8 8877689391    24
```

This creates a table that summarizes the sleep information. It gives: the id of the user, the number of sleeps tracked, the average time spent sleeping, the average time spent in bed, the average time spent not sleeping, and how far the average is from 8 hours It is then ordered by the number of sleeps tracked from least to most.

```
sleep <- read.csv("./Case_Study_Data/sleepDay_merged.csv")
sleep_by_id <- sleep %>% group_by(Id)
sleep_sum <- summarize(sleep_by_id, sleeps = n(), average_sleep = mean(TotalMinutesAsleep), average_bed
sleep_sum$not_sleeping = sleep_sum$average_bed - sleep_sum$average_sleep
sleep_sum$time_under_eight_hours = 480 - sleep_sum$average_sleep

ordered_sleep <- sleep_sum[order(sleep_sum$sleeps, decreasing = FALSE),]

ordered_sleep
```

```
## # A tibble: 24 x 6
##            Id sleeps average_sleep average_bed not_sleeping time_under_eight_ho~
##         <dbl>  <int>         <dbl>       <dbl>        <dbl>                <dbl>
##  1 2320127002      1          61          69           8                   419
##  2 7007744171      2          68.5        71.5         3                   412.
##  3 1844505072      3         652         961         309                  -172
##  4 6775888955      3         350.        369          19.3                 130.
##  5 8053475328      3         297         302.          4.67                183
##  6 1644430081      4         294         346          52                   186
##  7 1927972279      5         417         438.         20.8                  63
##  8 4558609924      5         128.        140          12.4                 352.
##  9 4020332650      8         349.        380.         30.4                 131.
## 10 2347167796     15         447.        491.         44.5                  33.2
## # ... with 14 more rows
```

This creates a summary table of the dailyActivity data that tracks: user id, days tracked, average number of sedentary minutes, average active minutes, average step count, average distance, average calories lost, and average time spent at light, medium, and very active.

```
dailyActivity <- read.csv("./Case_Study_Data/dailyActivity_merged.csv")
activity_by_id <- dailyActivity %>% group_by(Id)
activity_sum <- summarize(activity_by_id, count = n(), rest = mean(SedentaryMinutes), avg_steps = mean(
activity_sum$active <- 1440 - activity_sum$rest
activity_sum
```

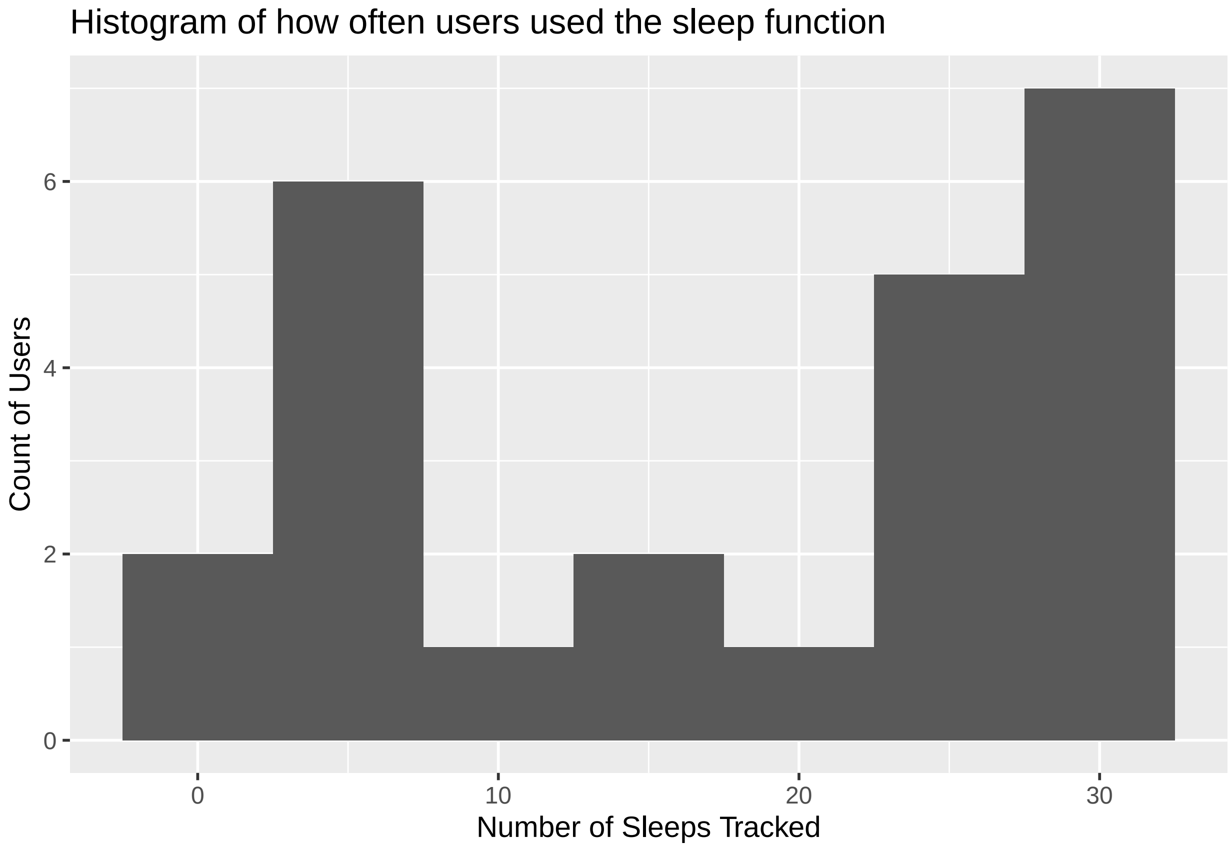
```
## # A tibble: 33 x 10
##          Id count  rest avg_steps avg_distance avg_calories avg_light avg_fairly
##       <dbl> <int> <dbl>     <dbl>        <dbl>        <dbl>     <dbl>      <dbl>
##  1   1.50e9    31  848.    12117.        7.81         1816.     220.      19.2
##  2   1.62e9    31 1258.     5744.        3.91         1483.     153.       5.81
##  3   1.64e9    30 1162.     7283.        5.30         2811.     178.      21.4
##  4   1.84e9    31 1207.     2580.        1.71         1573.     115.       1.29
##  5   1.93e9    31 1317.      916.        0.635        2173.      38.6      0.774
##  6   2.02e9    31 1113.    11371.        8.08         2510.     257.      19.4
##  7   2.03e9    31  689.     5567.        3.45         1541.     257.       0.258
##  8   2.32e9    31 1220.     4717.        3.19         1724.     198.       2.58
##  9   2.35e9    18  687.     9520.        6.36         2043.     252.      20.6
## 10   2.87e9    31 1097.     7556.        5.10         1917.     308        6.13
## # ... with 23 more rows, and 2 more variables: avg_very <dbl>, active <dbl>
```

Summary of Analysis

Most people did not track their weight. Only 26.6666667% of users tracked their weight. Most people who did so did not do it often. 67 was the total number of times that the weight tracking functionality was used.

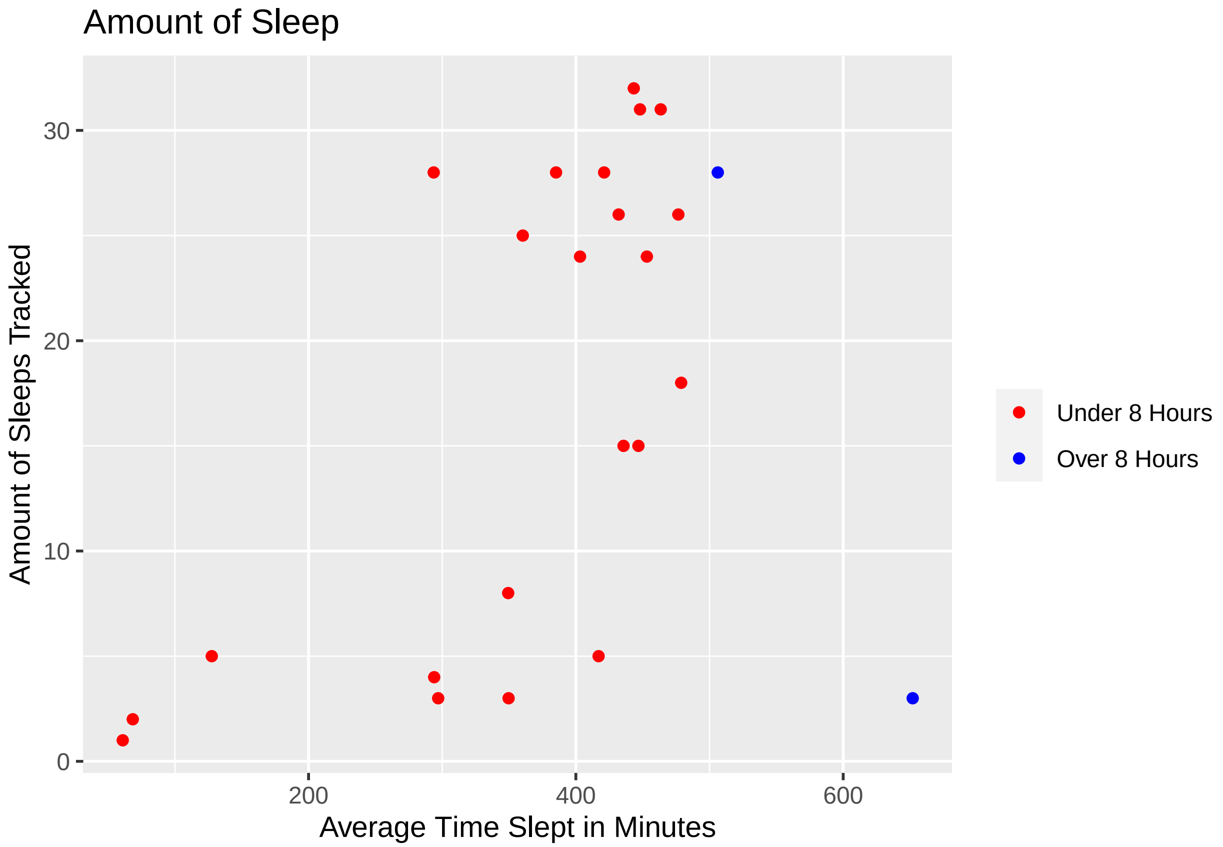
Exactly half of the users recorded 15 or more sleeps. While there are more users who recorded 0 sleeps, the histogram of times a user measured their sleep is a bimodal distrubtion. This means that there are two large categories of users: users who regularly use the sleep function, and users that only occasionaly use the sleep tracking feature.

```
library(ggplot2)
p <- ggplot(ordered_sleep, aes(x=sleeps)) + geom_histogram(binwidth = 5) + labs(title="Histogram of how
p
```



This is also an area that people could improve their health in, given that only 2 users average more that 8 hours (480 minutes) of sleep every night. We can see in the scatterplot below about the sleep habits of users with FitBits.

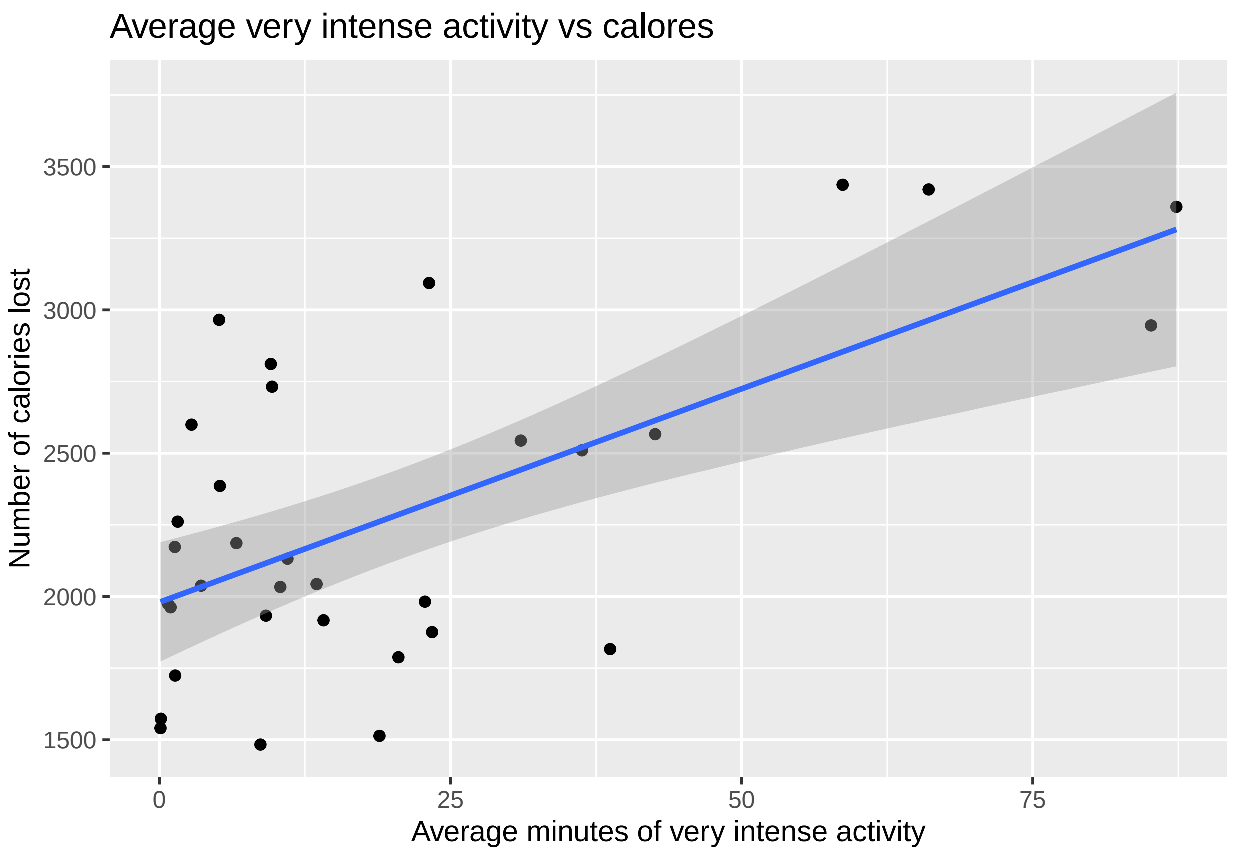
```
ggplot(sleep_sum, aes(x=average_sleep, y=sleeps)) + geom_point(aes(color = ifelse(average_sleep < 480,
```



There is a large chunk of users that are not very active people. Only about 25.8510638% of days have more than 30 minutes that are very active. On top of that, 20 of the 30 users averaged under 15 minutes of very active activity. We can see in the two scatterplots below that having longer times of high intensity does more for losing calories than does having long times of activity.

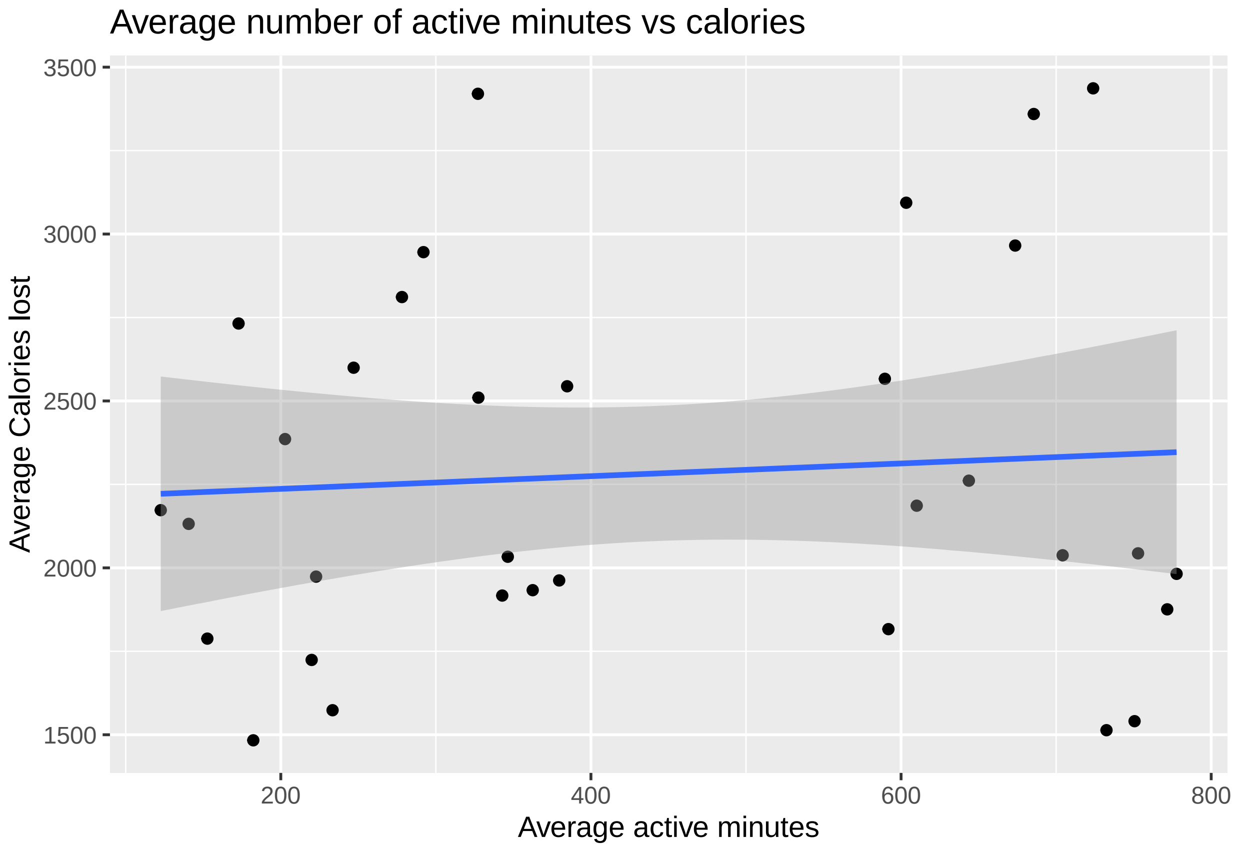
```
ggplot(activity_sum, aes(x=avg_very, y=avg_calories)) + geom_point() + geom_smooth(method=lm) + labs(ti
```

```
## `geom_smooth()` using formula 'y ~ x'
```



```
ggplot(activity_sum, aes(x=active, y=avg_calories)) + geom_point() +geom_smooth(method=lm) + labs(title=
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



We can also see that a higher average of intense exercise correlates with more calories lost, with a correlation coefficient, or how well two variables correlate on a scale of 1 (meaning high positive corrlation) and -1 (meaning high negative correlation), of 0.6299662. We see that this is not the case for high amounts of general activity as the correlation coefficient is 0.0769996

High-level recommendations (DONE)

These trends are important for Bellabeat as the company makes similar technologies to FitBit, mainly that they are used for tracking the users' health.

From the analysis, we can see that there are two groups of users: those who frequently have high intensity while wearing their fitbit, and those who are more casual users. There is the casual users take up a significantly greater portion, and products that target this group's health would help us with a large portion of the market share.

We can see that these customers are not losing as many calories as those who have periods of high intensity, and so providing them with more info about nutrition could help them to create a calorie deficit without an increase in intensity or time spent moving. This service is provided in the membership, which means that we would likely have to convince users to purchase the membership.

Hooking the customer with unique products that are not currently on the market and have great secondary functionality, such as the Spring and Time would get new casual users into our ecosystem. Using the products they buy, we can help improve their health, such as an increase in sleep, seeing that many users don't get enough sleep, or through more consumption of water, which may eventually lead to more customers that are interested in the membership.

We can compliment this by making the membership extremely competitive with other companies, so that those who frequently exercise would choose our platform as a great way to get all the information that they need and track their exercising.

This dataset generated by respondents to a distributed survey via Amazon Mechanical Turk between 03.12.2016-05.12.2016. Thirty eligible Fitbit users consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. Individual reports can be parsed by export session ID (column A) or timestamp (column B). Variation between output represents use of different types of Fitbit trackers and individual tracking behaviors / preferences.

You will produce a report with the following deliverables: 1. A clear summary of the business task 2. A description of all data sources used 3. Documentation of any cleaning or manipulation of data 4. A summary of your analysis 5. Supporting visualizations and key findings