Google Data Analytics - Bellabeat Case Study

Jed Ofori

20/02/2022

Business problem

I am a junior data analyst, part of the marketing analytics team at Bellabeat, a high-tech manufacturer of health-orientated products for women. Bellabeat is a successful small company, but they have the potential to become a larger figure in the global smart-device market. This requires the company to gather better market understanding, along with more competitor and product intelligence. Therefore, the chances of finding opportunities to improve current products or emerge into new markets. I have been tasked to focus on one of Bellabeat's products and conduct analysis on smart device data to gather insights in how consumers use their smart devices. The insights I discover will be used to provide guidance for the company's marketing strategy, concerning the selected product. My analysis and recommendations will be presented to my stakeholders. The business problem can be summarised as, "How do our consumers utilise our smart-devices for keeping tabs of their daily activities?"

Stakeholders

*** Urška Sršen - Bellabeat co-founder and Chief Creative Officer.**

*** Sando Mur - Bellabeat co-founder and key member of the Bellabeat executive team.**

*** Bellabeat marketing analytics team - A team of data analysts responsible for collecting, analysing, and reporting data to help progress Bellabeat's marketing strategy.**

Data preparation

I was encouraged to use a public dataset which explores the daily habits of smart-device users. Sršen directly pointed me to FitBit Fitness Tracker Data (a public dataset that was made available thanks to Mobius). The dataset is publicly available on Kaggle, the dataset contains consensual personal tracker data of 30 FitBit users. The data included is: information about the user's daily activity, step count, heart rate, sleep monitoring and calories. The data is organised as a folder with 18 files. The data in each spreadsheet was formatted majorly as long data. I obtained the data from a third party source but after cross checking with the original source (cited by Mobius), the data is valid. The data source is comprehensive for the business task and up to date (for the sake of the case study). In order to use the data, sorting and fishing out relevant data for the business task had to be done. To find relatable trends I will focus on daily summarisations of certain variables like sleep for example.

Install packages

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

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

```
install.packages("ggplot2")
```

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

```
install.packages("scales")
```

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

Load packages

```
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()
```

```
library(ggplot2)
library(scales)
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

Importing packages

```
activity <- read.csv("dailyActivity_merged.csv")
calories <- read.csv("dailyCalories_merged.csv")
intensities <- read.csv("dailyIntensities_merged.csv")
steps <- read.csv("dailySteps_merged.csv")
sleep <- read.csv("sleepDay.csv")
weight <- read.csv("weightLogInfo_merged.csv")
```

Processing the data

```
head(activity)
```

I previewed the data in Excel, then to double-check everything had been imported correctly I used the view() and head() functions.

```
##           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(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(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
```

```
head(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
```

```
head(sleep)
```

```
##           Id              SleepDay TotalSleepRecords TotalMinutesAsleep
## 1 1503960366      04/12/2016 00:00                 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(weight)
```

```
##           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
```

Fixing the format

```
# activity
activity$ActivityDate=as.POSIXct(activity$ActivityDate,
                                 format="%m/%d/%Y", tz=Sys.timezone())
activity$date <- format(activity$ActivityDate, format = "%d/%m/%y")
# calories
calories$ActivityDay=
  as.POSIXct(calories$ActivityDay,
             format="%m/%d/%Y %I:%M:%S %p", tz=Sys.timezone())
calories$date <- format(calories$ActivityDay, format = "%d/%m/%Y")
# intensities
intensities$ActivityDay=
  as.POSIXct(intensities$ActivityDay,
             format="%m/%d/%Y %I:%M:%S %p", tz=Sys.timezone())
intensities$date <- format(intensities$ActivityDay, format = "%m/%d/%Y")
# sleep
sleep$SleepDay=as.POSIXct(sleep$SleepDay,
                          format="%m/%d/%Y %I:%M:%S %p", tz=Sys.timezone())
sleep$date <- format(sleep$SleepDay, format = "%d/%m/%y")
# steps
steps$ActivityDay=as.POSIXct(intensities$ActivityDay,
                             format="%m/%d/%Y", tz=Sys.timezone())
steps$date <- format(steps$ActivityDay, format = "%m/%d/%y")
# weight
weight$Date=as.POSIXct(weight$Date, format ="%m/%d/%Y", tz=Sys.timezone())
weight$date <- format(weight$Date, format = "%m/%d/%y")
```

There were some formatting issues with various sections of the data. The data was converted to a date time format and then separated as date and time.

Exploring data

```
n_distinct(activity$Id)
```

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

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

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

```
n_distinct(intensities$Id)
```

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

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

```
## [1] 24
```

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

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

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

```
## [1] 8
```

The information above shows that some participants did not provide data for some variables. There were 33 participants in this study, 24 provided sleep data and only 8 provided weight data. Based on the fact only 8 participants provided weight data, I decided it was best to not include the weight dataset. 8 participants is a small sample and not enough to provide conclusions or recommendations.

Merging and cleaning the data

```
merged_data <- merge(sleep, activity,  by = c("Id","date")) %>%
  drop_na() %>%
  select(-SleepDay, -TrackerDistance, -ActivityDate )
head(merged_data)
```

The activity dataset contains all the data of from the other imported datasets except the weight and sleep data. I merged the activity and sleep data via outer join. I believe the sleep data has a lot of possible insights as it is a big factor that impacts our livelihood.

```
##           Id     date TotalSleepRecords TotalMinutesAsleep TotalTimeInBed
## 1 1503960366 13/04/16                 2                384            407
## 2 1503960366 15/04/16                 1                412            442
## 3 1503960366 16/04/16                 2                340            367
## 4 1503960366 17/04/16                 1                700            712
## 5 1503960366 19/04/16                 1                304            320
## 6 1503960366 20/04/16                 1                360            377
##   TotalSteps TotalDistance LoggedActivitiesDistance VeryActiveDistance
## 1      10735          6.97                        0               1.57
## 2       9762          6.28                        0               2.14
## 3      12669          8.16                        0               2.71
## 4       9705          6.48                        0               3.19
## 5      15506          9.88                        0               3.53
## 6      10544          6.68                        0               1.96
##   ModeratelyActiveDistance LightActiveDistance SedentaryActiveDistance
## 1                     0.69                4.71                       0
## 2                     1.26                2.83                       0
## 3                     0.41                5.04                       0
## 4                     0.78                2.51                       0
## 5                     1.32                5.03                       0
## 6                     0.48                4.24                       0
##   VeryActiveMinutes FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes
## 1                21                  19                  217              776
## 2                29                  34                  209              726
## 3                36                  10                  221              773
## 4                38                  20                  164              539
## 5                50                  31                  264              775
## 6                28                  12                  205              818
##   Calories
## 1     1797
## 2     1745
## 3     1863
## 4     1728
## 5     2035
## 6     1786
```

Data summary

```
summary(merged_data)
```

```
##        Id                date           TotalSleepRecords TotalMinutesAsleep
##  Min.   :1.504e+09   Length:252         Min.   :1.000     Min.   : 59.0
##  1st Qu.:3.977e+09   Class :character   1st Qu.:1.000     1st Qu.:361.0
##  Median :4.703e+09   Mode  :character   Median :1.000     Median :429.0
##  Mean   :4.976e+09                      Mean   :1.127     Mean   :418.4
##  3rd Qu.:6.822e+09                      3rd Qu.:1.000     3rd Qu.:486.5
##  Max.   :8.792e+09                      Max.   :3.000     Max.   :775.0
##  TotalTimeInBed   TotalSteps    TotalDistance    LoggedActivitiesDistance
##  Min.   : 65    Min.   :   42   Min.   : 0.030   Min.   :0.0000
##  1st Qu.:405    1st Qu.: 5224   1st Qu.: 3.620   1st Qu.:0.0000
##  Median :461    Median : 9114   Median : 6.310   Median :0.0000
##  Mean   :457    Mean   : 8598   Mean   : 6.092   Mean   :0.1039
##  3rd Qu.:522    3rd Qu.:11396   3rd Qu.: 8.075   3rd Qu.:0.0000
##  Max.   :961    Max.   :22359   Max.   :17.190   Max.   :4.0817
##  VeryActiveDistance ModeratelyActiveDistance LightActiveDistance
##  Min.   : 0.000     Min.   :0.000            Min.   :0.030
##  1st Qu.: 0.000     1st Qu.:0.000            1st Qu.:2.538
##  Median : 0.565     Median :0.420            Median :3.710
##  Mean   : 1.515     Mean   :0.745            Mean   :3.786
##  3rd Qu.: 2.527     3rd Qu.:1.032            3rd Qu.:4.910
##  Max.   :12.540     Max.   :5.120            Max.   :9.480
##  SedentaryActiveDistance VeryActiveMinutes FairlyActiveMinutes
##  Min.   :0.0000000       Min.   :  0.00    Min.   : 0.00
##  1st Qu.:0.0000000       1st Qu.:  0.00    1st Qu.: 0.00
##  Median :0.0000000       Median :  9.50    Median :12.00
##  Mean   :0.0008333       Mean   : 26.44    Mean   :18.22
##  3rd Qu.:0.0000000       3rd Qu.: 36.50    3rd Qu.:28.00
##  Max.   :0.1100000       Max.   :210.00    Max.   :98.00
##  LightlyActiveMinutes SedentaryMinutes    Calories
##  Min.   :  4.0        Min.   :   2.0   Min.   : 403
##  1st Qu.:158.8        1st Qu.: 646.5   1st Qu.:1882
##  Median :206.0        Median : 720.5   Median :2202
##  Mean   :216.7        Mean   : 723.8   Mean   :2421
##  3rd Qu.:263.2        3rd Qu.: 781.2   3rd Qu.:2913
##  Max.   :518.0        Max.   :1265.0   Max.   :4900
```

Findings of merged data

***Average sedentary time is 724 minutes or 12 hours, this is half of the 24 hours each human has been gifted with, lifestyle changes should be made, however, this easier said than done!**

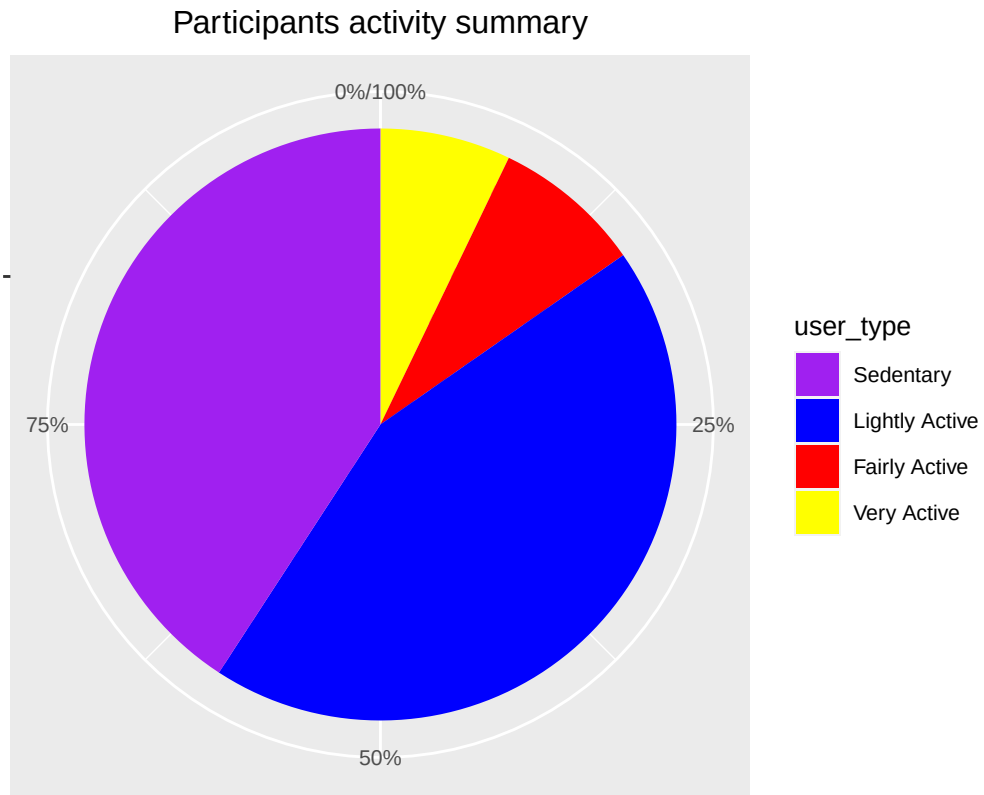
***Most of the participants are lightly active.**

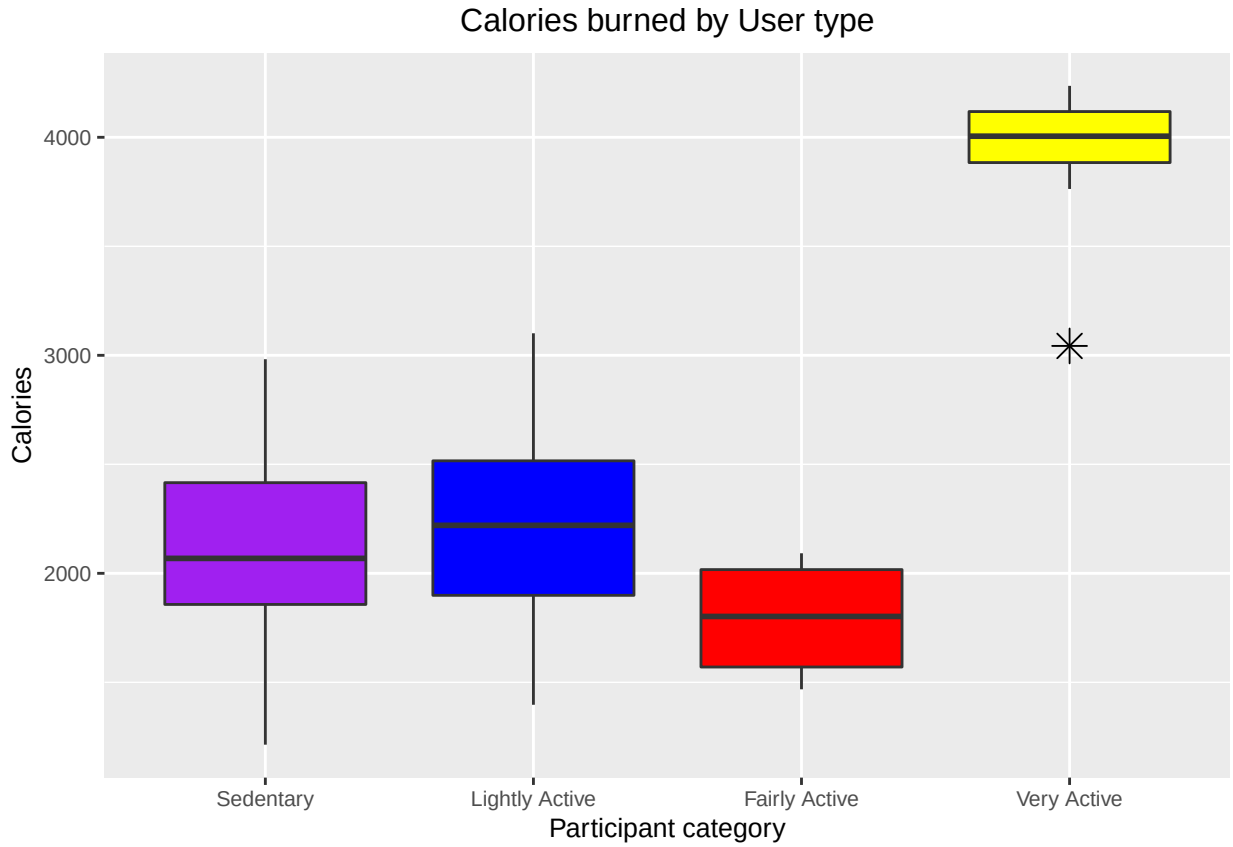
***On the average, participants sleep 1 time for 7 hours.**

***Average total steps per day are 8598 which is less than the recommended 10000 steps. According to CDC's research. 10000 steps is roughly 5 miles, which is the number said to help reduce certain health conditions, like heart disease and high blood pressure.**

In order to make some profound findings, it was neccessary to make a new table and group participants based on their intensity minutes.

```
usertype <- merged_data %>%
summarise(
user_type = factor(case_when(
    SedentaryMinutes > mean(SedentaryMinutes) & LightlyActiveMinutes
    < mean(LightlyActiveMinutes) &
      FairlyActiveMinutes < mean(FairlyActiveMinutes) &
      VeryActiveMinutes < mean(VeryActiveMinutes) ~ "Sedentary",
    SedentaryMinutes < mean(SedentaryMinutes) &
      LightlyActiveMinutes > mean(LightlyActiveMinutes) &
      FairlyActiveMinutes < mean(FairlyActiveMinutes) &
      VeryActiveMinutes < mean(VeryActiveMinutes) ~ "Lightly Active",
    SedentaryMinutes < mean(SedentaryMinutes) &
      LightlyActiveMinutes < mean(LightlyActiveMinutes) &
      FairlyActiveMinutes > mean(FairlyActiveMinutes) &
      VeryActiveMinutes < mean(VeryActiveMinutes) ~ "Fairly Active",
    SedentaryMinutes < mean(SedentaryMinutes) &
      LightlyActiveMinutes < mean(LightlyActiveMinutes) &
      FairlyActiveMinutes < mean(FairlyActiveMinutes) &
      VeryActiveMinutes > mean(VeryActiveMinutes) ~ "Very Active",
),levels=c("Sedentary", "Lightly Active", "Fairly Active", "Very Active")),
Calories, .group=Id) %>%
drop_na()
```






```
usertype %>%
group_by(user_type) %>%
summarise(group_total = n()) %>%
mutate(total = sum(group_total)) %>%
group_by(user_type)%>%
  summarise(total_percentage = group_total*100/total)
```

I struggled to place the percentages on to the pie chart in the way I wanted to. I decided to just place a relating table below to care of any questions.

```
## # A tibble: 4 x 2
##   user_type      total_percentage
##   <fct>                     <dbl>
## 1 Sedentary                 40.8
## 2 Lightly Active            43.9
## 3 Fairly Active              8.16
## 4 Very Active                7.14
```

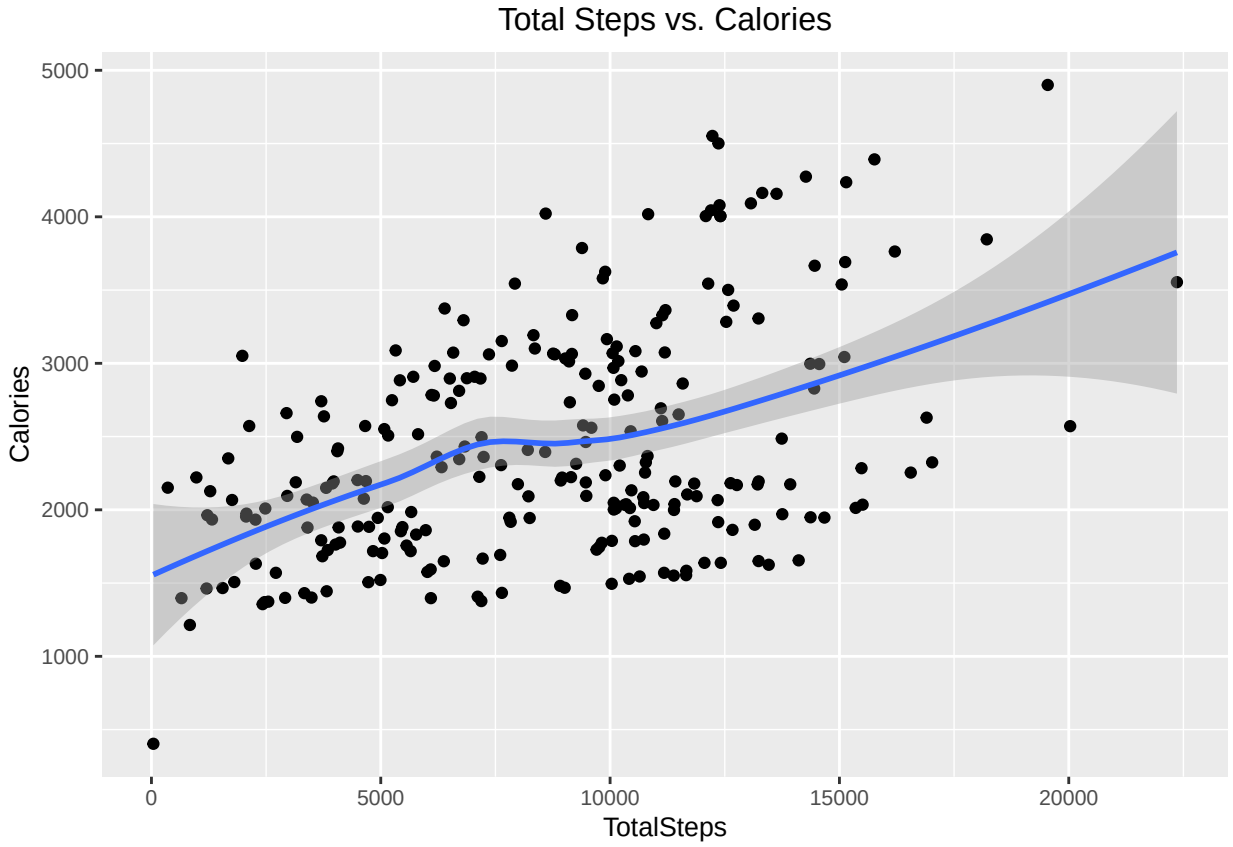
Analysis

It is not surprising to see that majority of the participants were classed as sendentary or lightly active. The world we live in geared towards sendentary lifestyles as it is very common to do jobs that inlcude a lot of seated activity (9-5 office jobs). I was not surprised by the smallest group being the very active. There could be a possiblity of Bellabeats formulating programmes for people of different groups to stay active and gradually improve to a higher group, or maintaoning activity that suits the individual's lifestyle. There is an outlier in terms

of the box plot visual, it was expected that the fairly active group would have burned more calories than the sedentary and light active groups. However, I believe this could be down to the minutes performing fairly active activity. Nonetheless, the remainder of the box plot visual reiterates the belief that a greater level of activity expends more energy, hence burning more calories. Additionally, notifications that notify users they have been sitting for an hour, this could help users be more mindful about the time they have spend in seated posistions.

```
ggplot(data= merged_data, aes(x=TotalSteps, y=Calories)) +
  geom_point() + geom_smooth() + labs(title="Total Steps vs. Calories") +
  theme(plot.title = element_text(hjust = 0.5), text = element_text(size = 10))
```

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



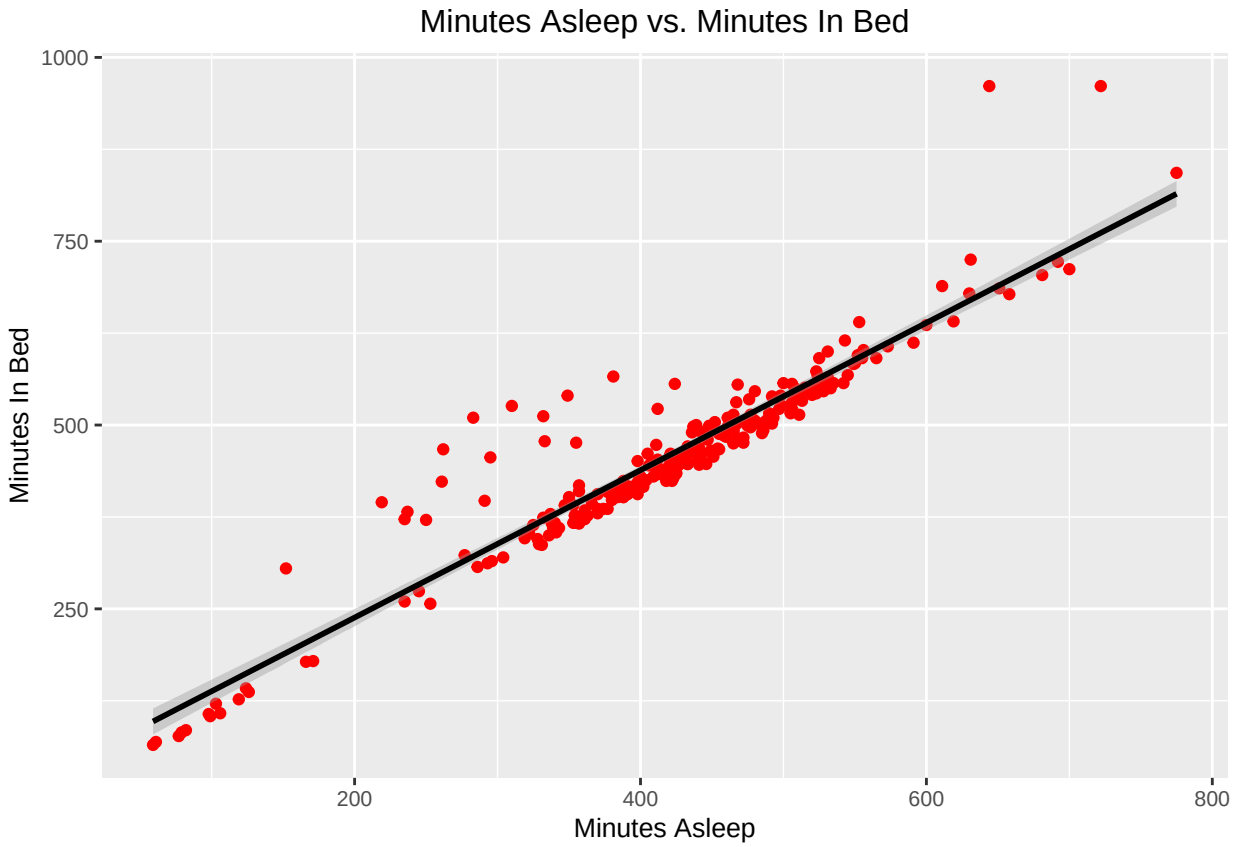
Analysis

Positive correlation is displayed by the above visualisation. The more active a person is the more likely they are to burn more calories. However, correlation does not always equate cause, there are several factors that determine a person's calorie output, such as weight, metabolism, age, hormone levels and diet. Bellabeats could utilise notifications that help individuals reach steps goals, with encouraging messages like "keep up the momentum!".

```
ggplot(data=merged_data, aes(x=TotalMinutesAsleep, y=TotalTimeInBed )) +
geom_point(color='red') + geom_smooth(color='black', method = "lm") +
  labs(title="Minutes Asleep vs. Minutes In Bed",
       x= "Minutes Asleep", y= "Minutes In Bed") +
```

```
  theme(plot.title = element_text(hjust = 0.5), text = element_text(size = 10))
```

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

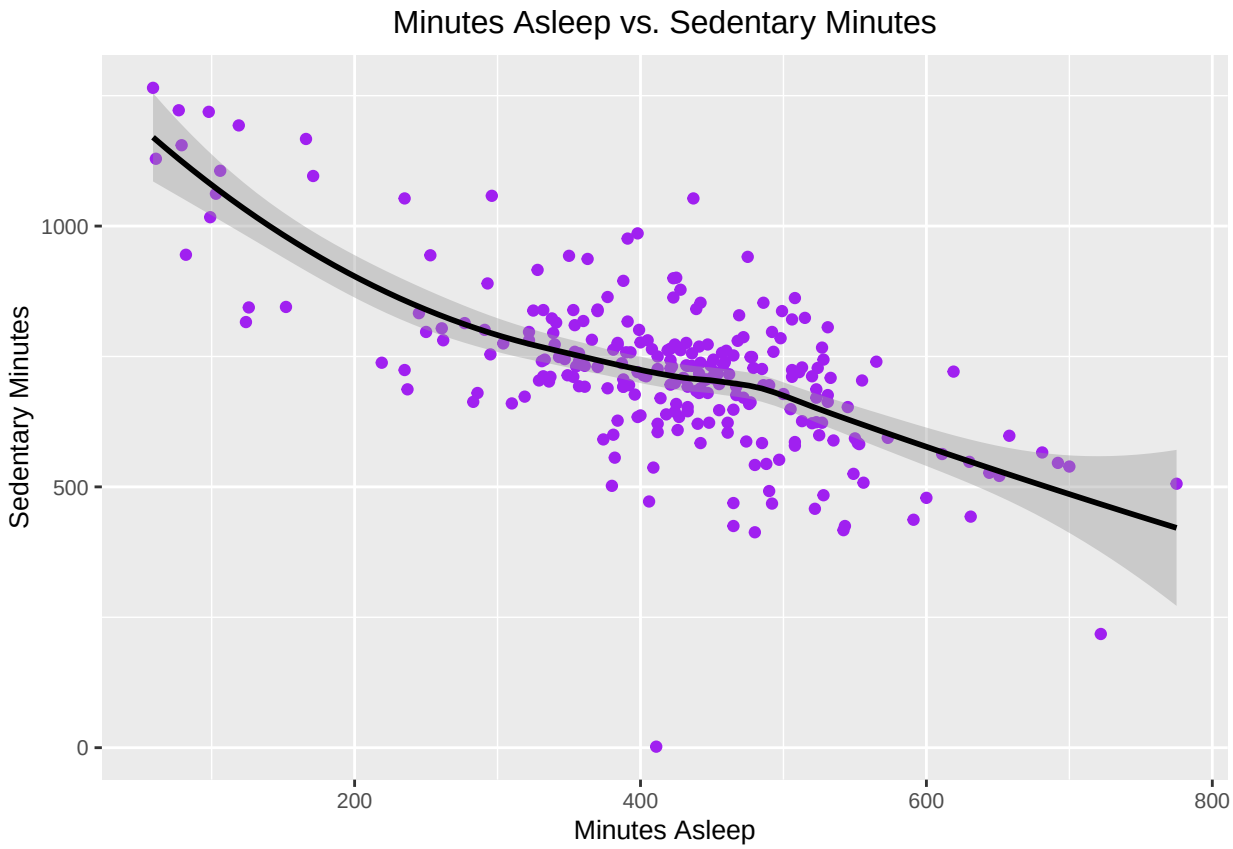


Analysis

The relationship between time in bed and time spent sleeping is strongly positive. Bellabeats could make a notifications prompting users to go to sleep. Users could be able set a bedtime schedule that helps users to be organised and have a routine.

```
ggplot(data=merged_data, aes(x=TotalMinutesAsleep, y=SedentaryMinutes)) +
geom_point(color='purple') + geom_smooth(color='black') +
  labs(title="Minutes Asleep vs. Sedentary Minutes",
       x= "Minutes Asleep", y= "Sedentary Minutes") +
  theme(plot.title = element_text(hjust = 0.5), text = element_text(size = 10))
```

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



Analysis

The visual above presents negative correlation between minutes spent asleep and time spent in sedentary positions. A reason for people with high amount of sedentary minutes can be down to being sleep deprived and requiring positions that expend less energy. Again this puts emphasis on the need for smart device functions that help progress healthier sleep habits. However, more data on other factors such as job occupation and commuting time need to be collected to allow a solid conclusion to be made.

Share and Act - Business recommendations

*** Product choice - Time, my product choice was between the Leaf and Time product because the Fitbit is similar to these two Bellabeat products. The other products such as the Bellabeat membership and Spring were not considered as the data-set provided was sufficient enough to spark marketing ideas for the two products. I believe changes done to the Time product should automatically incorporated in the Bellabeat app.**

*** Target audience - Women working full-time in office jobs (9-5) were the target audience based on intensity and sedentary time data. As aforementioned, the participants were assumed to be females for the sake of the case study.**

Main marketing message

*** Bellabeat should be centered around information and inspiration, allowing them to feel empowered and potentially positively impact people around them.**

Ideas

*** Step goals - Users would be allowed set their own personal step goals. The smartwatch should give prompts during the day. The prompts could be encouraging phrases or educative information like "10,000 steps regularly a day can reduce the chance of having heart related diseases".**

*** Motion tracking - The smartwatch would be able to prompt you to stand up each hour if you have been in a sedentary position 10 minutes before the next hour commences. This would help people be mindful of how long they stay seated for.**

*** Sleep notifications - A setting should be implemented, where users can set bedtime schedules and recieve prompts to go to bed. Notifications such as "time to turn off electronics close to your bed", could be beneficial to users.**

Recommendations

*** There was no data relating to stress, there may be a relationship between sleeping and stress levels that could be explored. However, research will have to be done to determine an accurate method for smart-devices to measure stress levels.**

***Data surrounding weight loss could be explored as there are several people struggling with losing weight or maintaining a desired weight.**

Thank you for your time!