Bellabeat Case Study

Tsun Hei Tai

In this case study, I will be looking at wellness company Bellabeat, a high-tech manufacturer of health-focused products for women. Bellabeat Chief Creative Officer believes that analyzing smart device fitness data could help unlock new growth opportunities for the company.

The case study was completed as a part of the 'Google Data Analytics Certificate' online course on Coursera.

Business task

Key questions to answer:

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 used for this case study

Data used for this case study was sourced from this link.

These datasets were 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.

Data acknowledgement - Furberg, Robert; Brinton, Julia; Keating, Michael ; Ortiz, Alexa https://zenodo.org/record/53894#.YMoUpnVKiP9

Setup working environment in R

```
library(tidyverse)
```

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

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

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

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
## The following objects are masked from ’package:base’:
##
##     date, intersect, setdiff, union
```

```
library(dplyr)
library(ggplot2)
library(tidyr)
```

Importing data

```
activity <- read.csv("dailyActivity_merged.csv")
sleep <- read.csv("sleepDay_merged.csv")
hourly_steps <- read.csv("hourlySteps_merged.csv")
```

Previewing data

Taking a look at the activity data set

```
head(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
```

```
str(activity)
```

```
## ’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 ...
```

Take a look at the sleep and hourly steps data.

```
head(sleep)
```

```
##           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
```

```
str(sleep)
```

```
## ’data.frame’:    413 obs. of  5 variables:
##  $ Id                : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ SleepDay          : chr  "4/12/2016 12:00:00 AM" "4/13/2016 12:00:00 AM" "4/15/2016 12:00:00 AM"
##  $ TotalSleepRecords : int  1 2 1 2 1 1 1 1 1 1 ...
##  $ TotalMinutesAsleep: int  327 384 412 340 700 304 360 325 361 430 ...
##  $ TotalTimeInBed    : int  346 407 442 367 712 320 377 364 384 449 ...
```

```
head(hourly_steps)
```

```
##           Id          ActivityHour StepTotal
## 1 1503960366 4/12/2016 12:00:00 AM       373
## 2 1503960366  4/12/2016 1:00:00 AM       160
## 3 1503960366  4/12/2016 2:00:00 AM       151
## 4 1503960366  4/12/2016 3:00:00 AM         0
## 5 1503960366  4/12/2016 4:00:00 AM         0
## 6 1503960366  4/12/2016 5:00:00 AM         0
```

```
str(hourly_steps)
```

```
## ’data.frame’:    22099 obs. of  3 variables:
##  $ Id          : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ ActivityHour: chr  "4/12/2016 12:00:00 AM" "4/12/2016 1:00:00 AM" "4/12/2016 2:00:00 AM" "4/12/20
##  $ StepTotal   : int  373 160 151 0 0 0 0 0 250 1864 ...
```

Formatting data

```r
# activity
activity$ActivityDate=as.POSIXct(activity$ActivityDate, format="%m/%d/%Y", tz=Sys.timezone())
activity$date <- format(activity$ActivityDate, 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 = "%m/%d/%y")
# hourly steps
hourly_steps$ActivityHour=as.POSIXct(hourly_steps$ActivityHour, format="%m/%d/%Y %I:%M:%S %p", tz=Sys.ti
hourly_steps$time <- format(hourly_steps$ActivityHour, format = "%H:%M:%S")
hourly_steps$date <- format(hourly_steps$ActivityHour, format = "%m/%d/%y")
```

Check cleaned data sets

```r
head(activity)
```

```
##           Id ActivityDate TotalSteps TotalDistance TrackerDistance
## 1 1503960366   2016-04-12      13162          8.50            8.50
## 2 1503960366   2016-04-13      10735          6.97            6.97
## 3 1503960366   2016-04-14      10460          6.74            6.74
## 4 1503960366   2016-04-15       9762          6.28            6.28
## 5 1503960366   2016-04-16      12669          8.16            8.16
## 6 1503960366   2016-04-17       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     date
## 1                  13                  328              728     1985 04/12/16
## 2                  19                  217              776     1797 04/13/16
## 3                  11                  181             1218     1776 04/14/16
## 4                  34                  209              726     1745 04/15/16
## 5                  10                  221              773     1863 04/16/16
## 6                  20                  164              539     1728 04/17/16
```

```r
head(sleep)
```

```
##           Id   SleepDay TotalSleepRecords TotalMinutesAsleep TotalTimeInBed
## 1 1503960366 2016-04-12                 1                327            346
## 2 1503960366 2016-04-13                 2                384            407
## 3 1503960366 2016-04-15                 1                412            442
## 4 1503960366 2016-04-16                 2                340            367
## 5 1503960366 2016-04-17                 1                700            712
## 6 1503960366 2016-04-19                 1                304            320
##       date
## 1 04/12/16
## 2 04/13/16
## 3 04/15/16
## 4 04/16/16
## 5 04/17/16
## 6 04/19/16
```

```
head(hourly_steps)
```

```
##           Id        ActivityHour StepTotal     time     date
## 1 1503960366 2016-04-12 00:00:00       373 00:00:00 04/12/16
## 2 1503960366 2016-04-12 01:00:00       160 01:00:00 04/12/16
## 3 1503960366 2016-04-12 02:00:00       151 02:00:00 04/12/16
## 4 1503960366 2016-04-12 03:00:00         0 03:00:00 04/12/16
## 5 1503960366 2016-04-12 04:00:00         0 04:00:00 04/12/16
## 6 1503960366 2016-04-12 05:00:00         0 05:00:00 04/12/16
```

Summary statistics

Let's have a look at some summary statistics of our data sets.

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

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

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

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

```
n_distinct(hourly_steps$Id)
```

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

There are 33 distinct participants in our activity data set, 24 in the sleep data set and 33 in the hourly steps data set.

Let's have a look at some summary statistics for each data set.

For the activity data set:

```
activity %>%
  select(TotalSteps,
         TotalDistance,
         SedentaryMinutes) %>%
  summary()
```

```
##    TotalSteps    TotalDistance    SedentaryMinutes
##  Min.   :    0   Min.   : 0.000   Min.   :   0.0  
##  1st Qu.: 3790   1st Qu.: 2.620   1st Qu.: 729.8  
##  Median : 7406   Median : 5.245   Median :1057.5  
##  Mean   : 7638   Mean   : 5.490   Mean   : 991.2  
##  3rd Qu.:10727   3rd Qu.: 7.713   3rd Qu.:1229.5  
##  Max.   :36019   Max.   :28.030   Max.   :1440.0  
```

For the sleep data set:

```
sleep %>%
  select(TotalSleepRecords,
  TotalMinutesAsleep,
  TotalTimeInBed) %>%
  summary()
```

```
##  TotalSleepRecords TotalMinutesAsleep TotalTimeInBed
##  Min.   :1.000     Min.   : 58.0      Min.   : 61.0
##  1st Qu.:1.000     1st Qu.:361.0      1st Qu.:403.0
##  Median :1.000     Median :433.0      Median :463.0
##  Mean   :1.119     Mean   :419.5      Mean   :458.6
##  3rd Qu.:1.000     3rd Qu.:490.0      3rd Qu.:526.0
##  Max.   :3.000     Max.   :796.0      Max.   :961.0
```

- Average sedentary time is 991 minutes or 16.5 hours. That's quite a lot of time spent being sedentary!
- Average total steps per day was 7638, which sits in between 7000 - 8000 steps recommended by the CDC.
- Average minutes asleep is 419.5 minutes, just under 7 hours. The National Sleep Foundation recommends 7–9 hours of sleep a night for adults up to the age of 65, and 7–8 hours for those over 65. The data shows that a person is just getting the recommended amound of sleep each night on average.

Merging these two datasets together

Let's complete an inner join for these two data sets.

```
combined_data <- merge(sleep, activity, by=c('Id','date'))
```

Take a look at how many participants are in this data set and a quick look at this combined data set.

```
n_distinct(combined_data$Id)
```

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

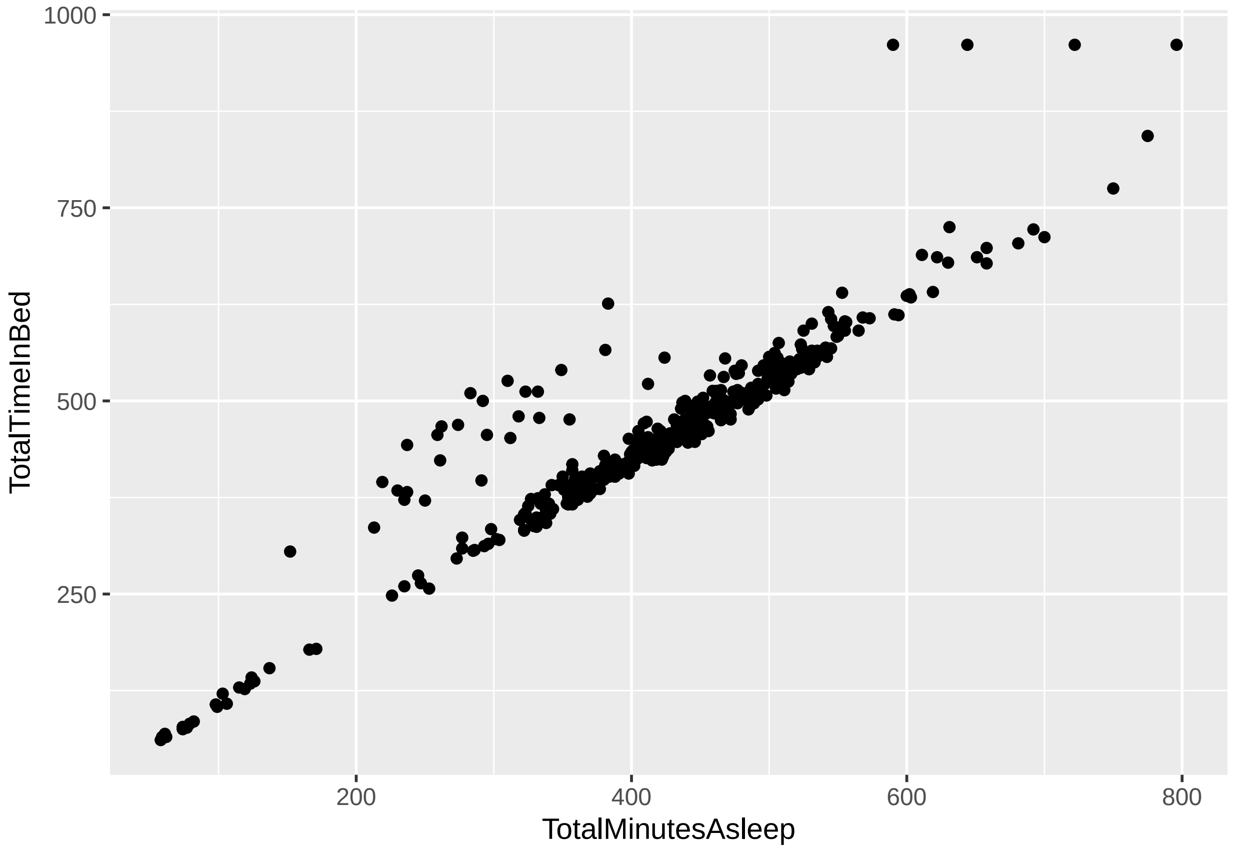
```
head(combined_data)
```

```
##           Id     date   SleepDay TotalSleepRecords TotalMinutesAsleep
## 1 1503960366 04/12/16 2016-04-12                 1                327
## 2 1503960366 04/13/16 2016-04-13                 2                384
## 3 1503960366 04/15/16 2016-04-15                 1                412
## 4 1503960366 04/16/16 2016-04-16                 2                340
## 5 1503960366 04/17/16 2016-04-17                 1                700
## 6 1503960366 04/19/16 2016-04-19                 1                304
##   TotalTimeInBed ActivityDate TotalSteps TotalDistance TrackerDistance
## 1            346   2016-04-12      13162          8.50            8.50
## 2            407   2016-04-13      10735          6.97            6.97
## 3            442   2016-04-15       9762          6.28            6.28
## 4            367   2016-04-16      12669          8.16            8.16
## 5            712   2016-04-17       9705          6.48            6.48
## 6            320   2016-04-19      15506          9.88            9.88
##   LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance
## 1                        0               1.88                     0.55
## 2                        0               1.57                     0.69
## 3                        0               2.14                     1.26
## 4                        0               2.71                     0.41
## 5                        0               3.19                     0.78
## 6                        0               3.53                     1.32
##   LightActiveDistance SedentaryActiveDistance VeryActiveMinutes
## 1                6.06                       0                25
## 2                4.71                       0                21
## 3            2.83                    0             29
## 4            5.04                    0             36
## 5            2.51                    0             38
## 6            5.03                    0             50
##   FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories
## 1                  13                  328              728     1985
## 2                  19                  217              776     1797
## 3                  34                  209              726     1745
## 4                  10                  221              773     1863
## 5                  20                  164              539     1728
## 6                  31                  264              775     2035
```

Plotting a few explorations

What's the relationship between minutes asleep and time in bed?

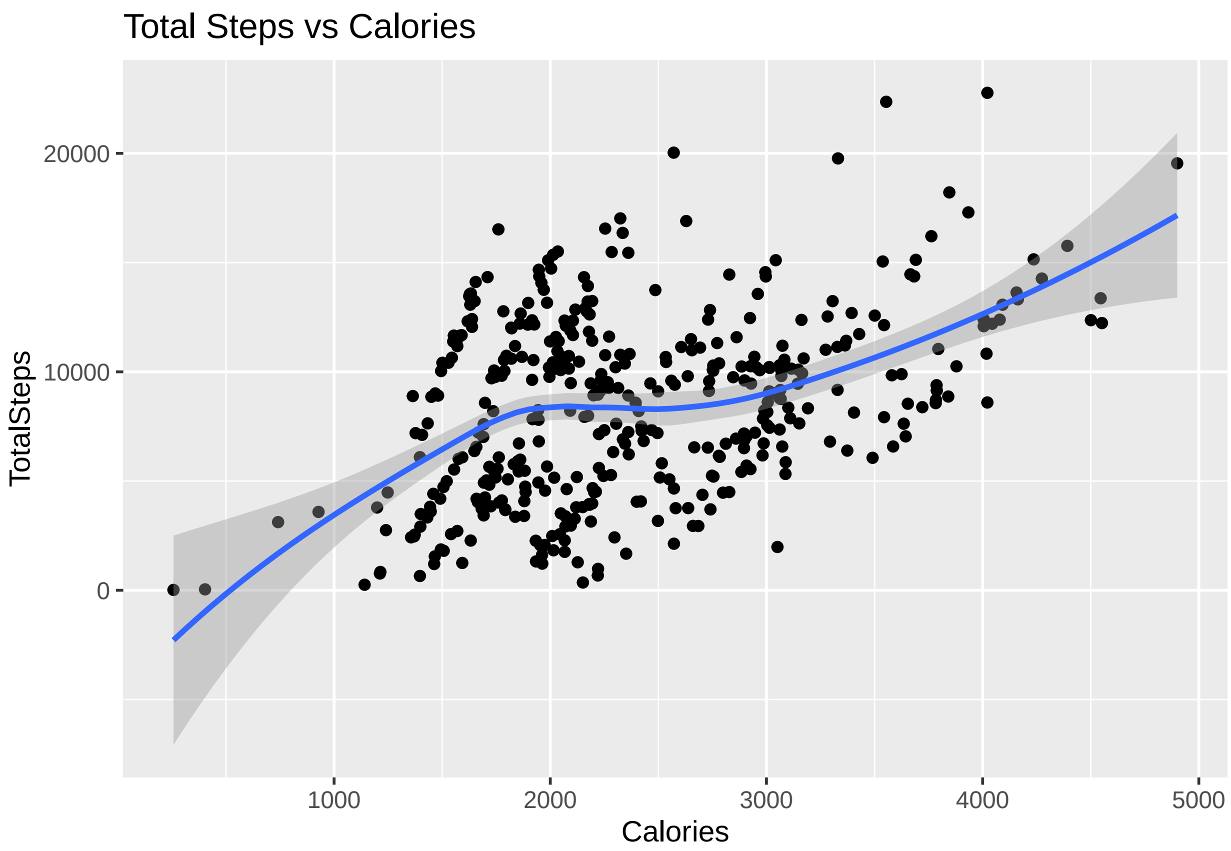
```
ggplot(data=combined_data, aes(x=TotalMinutesAsleep, y=TotalTimeInBed)) + geom_point()
```



No real surprises here as a linear relationship was expected.

```
ggplot(data=combined_data, mapping = aes(x = Calories, y = TotalSteps)) +
  geom_point() + geom_smooth() + labs(title = "Total Steps vs Calories")
```

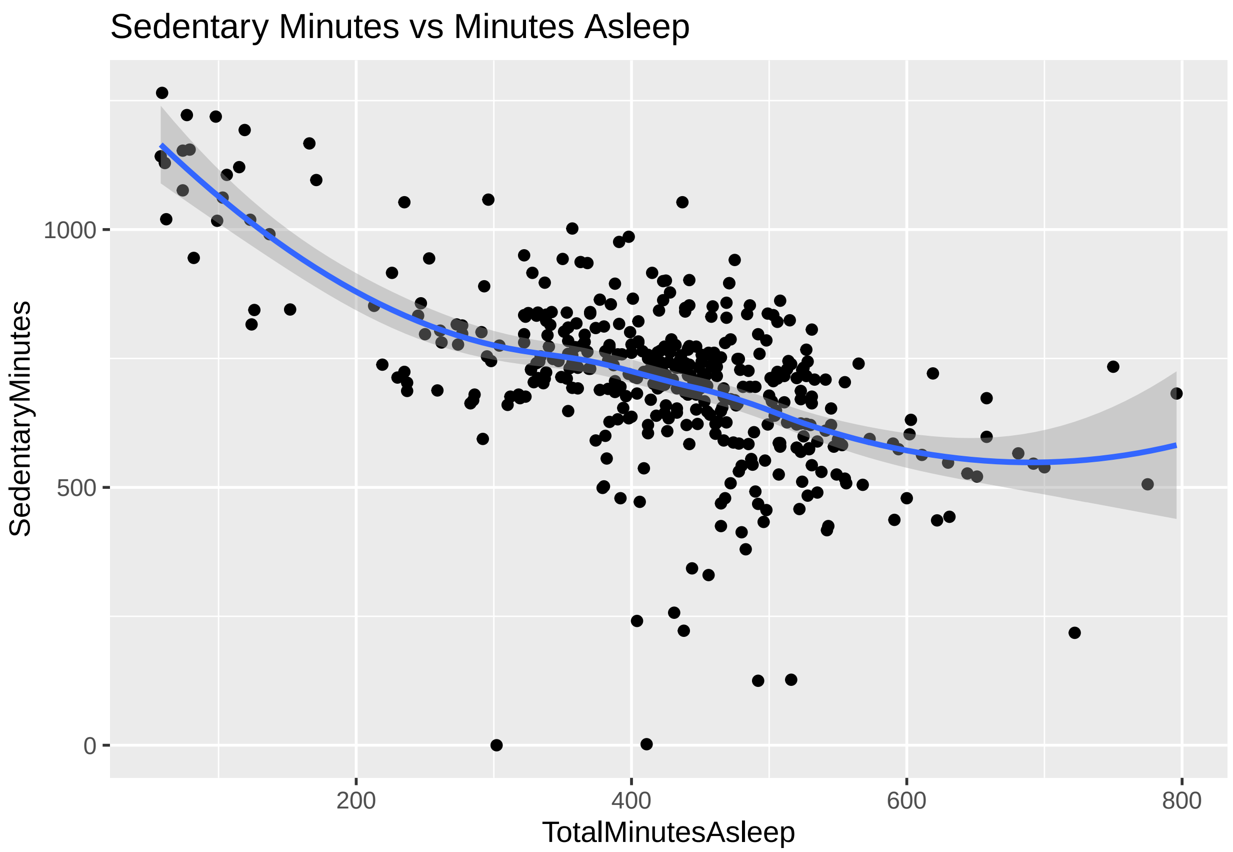
```
## ‘geom_smooth()‘ using method = ’loess’ and formula ’y ~ x’
```



No surprises that there's a positive correlation seen here. More steps taken would most likely lead to more calories burned.

```
ggplot(data=combined_data, mapping = aes(x = TotalMinutesAsleep, y = SedentaryMinutes)) +
  geom_point() + geom_smooth() + labs(title = "Sedentary Minutes vs Minutes Asleep")
```

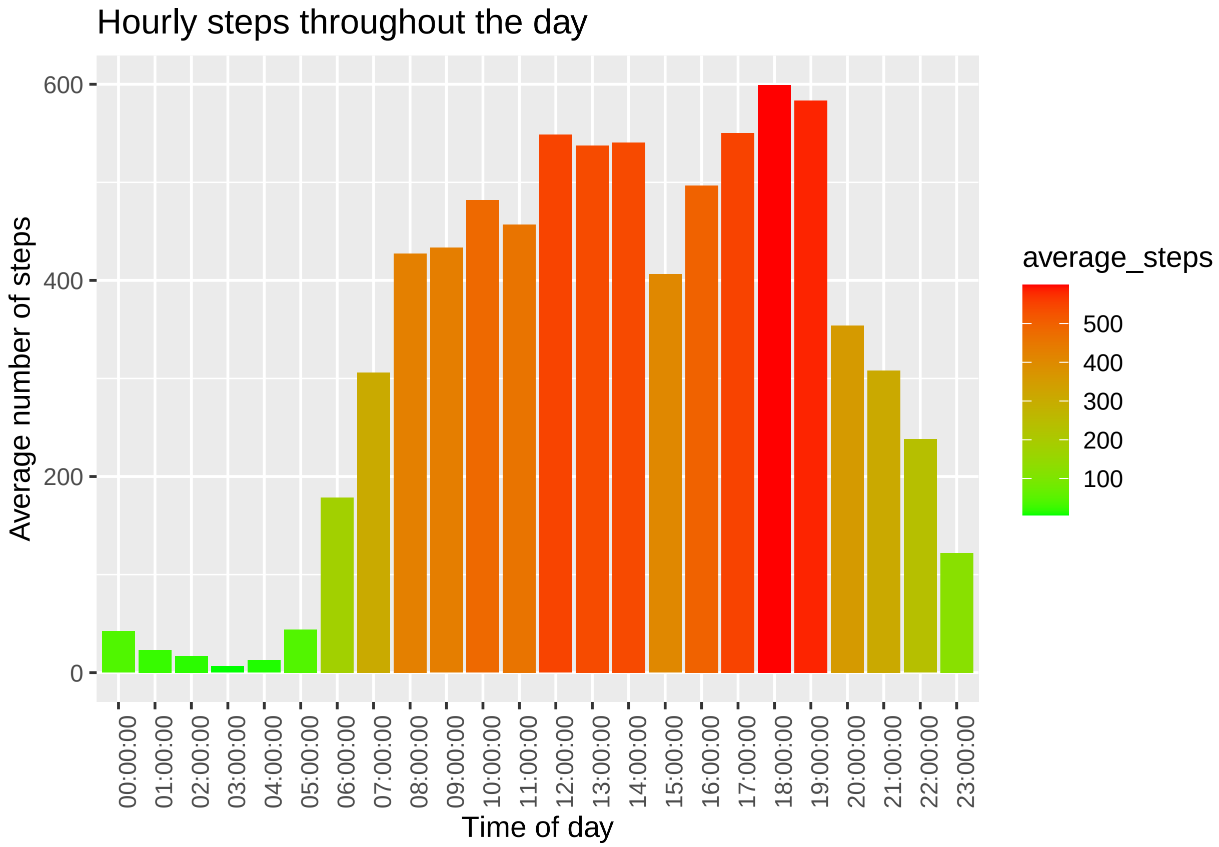
```
## 'geom_smooth()' using method = 'loess' and formula 'y ~ x'
```



There is a slight negative correlation between sedentary minutes vs total minutes asleep which is interesting. Will need more data analysis from another data set to see if fewer sedentary minutes is a cause of increased sleep time.

Let's look at the number of steps taken throughout the day

```
hourly_steps %>%
  group_by(time) %>%
  summarize(average_steps = mean(StepTotal)) %>%
  ggplot() +
  geom_col(mapping = aes(x=time, y = average_steps, fill = average_steps)) +
  labs(title = "Hourly steps throughout the day", x="Time of day", y="Average number of steps") +
  scale_fill_gradient(low = "green", high = "red")+
  theme(axis.text.x = element_text(angle = 90))
```



We can see that users are more active between 8am and 7pm. Walking more steps during lunch time from 12pm to 2pm and finishing work from 5pm and 7pm.

Recommendations for Bellabeat

Based of the analysis conducted, I would recommend the following:

1. The average number of steps per day taken by the users was 7638, which is between the 7000 to 8000 steps recommended by the CDC. **Bellabeat can send users a notification if the daily number of steps has not been reached**. CDC research findings show that more steps taken decreases the mortality rate. For more reading of the CDC research click this link.
2. The average sedentary time from the data analysed was around 16.5 hours. **Notifications could be set up on the device to remind users to decrease their sedentary time.**
3. Users had an average sleep time of less than 7 hours a day. **A notification could be sent to users showing their sleep time from the previous day/week. Alarms can be set up by users 30 minutes or an hour before the users' desired sleep time.**