Bellabeat Capstone Case Study

Allie Meristem

2022-10-12

Ask

- Business task: Bellabeat is a small company that produces high-tech, health-focused products for women, looking to become a larger player in the global smart device market. *Analyze usage data from a smart device similar to Bellabeat to find and apply insights to apply to a Bellabeat device for future marketing. Different devices collect data on activity, sleep, stress, and reproductive health.
- Key stakeholders: Urška Sršen and Sando Mur (co-founders and executives), Bellabeat marketing analytics team

Prepare

- Description of data sources used: FitBit Fitness Tracker Data (CC0: Public Domain, dataset made available through Mobius). Kaggle data set contains personal fitness tracker from thirty fitbit users who 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.
- Data organization and verification: Data appears to be in a wide format, with a new line of data for each occurance, separated by user ID. Data appears to be consistent, collected and recorded by a device.

Process

- Data credibility and integrity:
  - The original data is sited through Kaggle user Mobius: https://zenodo.org/record/53894#.Y0XLaHbMKUl (citation: Furberg, R., Brinton, J., Keating, M., & Ortiz, A. (2016). Crowd-sourced Fitbit datasets 03.12.2016-05.12.2016 [Data set]. Zenodo. https://doi.org/10.5281/zenodo.53894).
  - Data is crowd-sourced and open-access, where 30 eligible users consented to the submission of their personal tracker data.
  - There is no individual user data (like age, gender, location), so sampling bias is unclear.
  - Data appears to be complete, within the date range of 3/12 - 5/12, without any blank or NA fields; formatted and arranged in a clear, organized, and consistent way.
  - Data is from 2016, so it isn't very current and could miss capturing some key insights.
- I chose to use r to analyze this data, as the data sets are large, and I want to explore visualizations to share with the stakeholders
- Out of the available datasets to analyze, I used the "dailyActivity_merged.csv", "sleepDay_merged.csv", and "heartrate_seconds_merged.csv".
- Install packages and open libraries

```
install.packages("tidyverse")
install.packages("skimr")
install.packages("janitor")
library(tidyverse)
library(ggplot2)
```

```
library(readr)
library(tidyr)
library(dplyr)
library(skimr)
library(janitor)
```

- Import datasets, create data frames

```
heartrate <- read.csv("heartrate_seconds_merged.csv")
data <- read.csv("dailyActivity_merged.csv")
sleep <- read.csv("sleepDay_merged.csv")
```

- Preview datasets

```
head(data)
```

```
##           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(heartrate)
```

```
##           Id                 Time Value
## 1 2022484408 4/12/2016 7:21:00 AM    97
## 2 2022484408 4/12/2016 7:21:05 AM   102
## 3 2022484408 4/12/2016 7:21:10 AM   105
## 4 2022484408 4/12/2016 7:21:20 AM   103
## 5 2022484408 4/12/2016 7:21:25 AM   101
## 6 2022484408 4/12/2016 7:22:05 AM    95
```

```
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
```

- Steps to clean and format the data:
- 1. Identify how many unique Ids there are in each data frame. I notice that the heartrate data frame only has 14 unique Ids, which isn't enough for a comprehensive analysis; therefore I'll only use the dailyActivity_merged and sleepDay_merged data for this analysis.

```
n_distinct(data$Id)
```

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

```
n_distinct(heartrate$Id)
```

```
## [1] 14
```

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

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

- 2. How many observations are there in each data frame? Make sure there is a sufficient amount of data.

```
nrow(data)
```

```
## [1] 940
```

```
nrow(sleep)
```

```
## [1] 413
```

- 3. Rename columns to be all lowercase, clean names of columns to make sure they are all consistent:

```
data <- data %>%
  rename_with(tolower) %>%
  clean_names
sleep <- sleep %>%
  rename_with(tolower) %>%
  clean_names
```

- 4. Remove rows with 0 total steps from the dailyActivity data frame (they could indicate days that they didn't activate their device, for analysis I want to focus on tracked activity only); recheck the unique Ids and number of observances to see how they changed:

```
data <- filter(data, totalsteps !="0")
n_distinct(data$id)
```

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

```
nrow(data)
```

```
## [1] 863
```

- 4. Verification of cleaned data

```
colnames(data)
```

```
##  [1] "id"                       "activitydate"
##  [3] "totalsteps"               "totaldistance"
##  [5] "trackerdistance"          "loggedactivitiesdistance"
##  [7] "veryactivedistance"       "moderatelyactivedistance"
##  [9] "lightactivedistance"      "sedentaryactivedistance"
## [11] "veryactiveminutes"        "fairlyactiveminutes"
## [13] "lightlyactiveminutes"     "sedentaryminutes"
## [15] "calories"
```

```
colnames(sleep)
```

```
## [1] "id"                 "sleepday"           "totalsleeprecords"
## [4] "totalminutesasleep" "totaltimeinbed"
```

- Data summaries to gain insights:

```
data %>%
  select(totalsteps, totaldistance, trackerdistance, loggedactivitiesdistance, calories) %>%
  summary()
```

```
##    totalsteps    totaldistance   trackerdistance  loggedactivitiesdistance
##  Min.   :    4   Min.   : 0.00   Min.   : 0.000   Min.   :0.0000
##  1st Qu.: 4923   1st Qu.: 3.37   1st Qu.: 3.370   1st Qu.:0.0000
##  Median : 8053   Median : 5.59   Median : 5.590   Median :0.0000
##  Mean   : 8319   Mean   : 5.98   Mean   : 5.964   Mean   :0.1178
##  3rd Qu.:11092   3rd Qu.: 7.90   3rd Qu.: 7.880   3rd Qu.:0.0000
##  Max.   :36019   Max.   :28.03   Max.   :28.030   Max.   :4.9421
##     calories
##  Min.   :  52
##  1st Qu.:1856
##  Median :2220
##  Mean   :2361
##  3rd Qu.:2832
##  Max.   :4900
```

```
data %>%
  select(veryactivedistance, moderatelyactivedistance, lightactivedistance) %>%
  summary ()
```

```
##  veryactivedistance moderatelyactivedistance lightactivedistance
##  Min.   : 0.000     Min.   :0.0000           Min.   : 0.000
##  1st Qu.: 0.000     1st Qu.:0.0000           1st Qu.: 2.345
##  Median : 0.410     Median :0.3100           Median : 3.580
##  Mean   : 1.637     Mean   :0.6182           Mean   : 3.639
##  3rd Qu.: 2.275     3rd Qu.:0.8650           3rd Qu.: 4.895
##  Max.   :21.920     Max.   :6.4800           Max.   :10.710
```

```
data %>%
  select(veryactiveminutes, fairlyactiveminutes, lightlyactiveminutes, sedentaryminutes) %>%
```

```
  summary()
```

```
##  veryactiveminutes fairlyactiveminutes lightlyactiveminutes sedentaryminutes
##  Min.   :  0.00    Min.   :  0.00      Min.   :  0.0        Min.   :   0.0
##  1st Qu.:  0.00    1st Qu.:  0.00      1st Qu.:146.5        1st Qu.: 721.5
##  Median :  7.00    Median :  8.00      Median :208.0        Median :1021.0
##  Mean   : 23.02    Mean   : 14.78      Mean   :210.0        Mean   : 955.8
##  3rd Qu.: 35.00    3rd Qu.: 21.00      3rd Qu.:272.0        3rd Qu.:1189.0
##  Max.   :210.00    Max.   :143.00      Max.   :518.0        Max.   :1440.0
```

```
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
```

- Prepare to merge the two data frames: on "sleep" data frame, split "sleepday" column into "date" and "time" column

```
sleep <- separate(sleep, sleepday,into=c('date','time'), sep=' ')
```

```
## Warning: Expected 2 pieces. Additional pieces discarded in 413 rows [1, 2, 3, 4,
## 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

- Prepare to merge the two data frames: on "data" data frame, rename "activitydate" to "date"

```
data <- rename(data, date=activitydate)
```

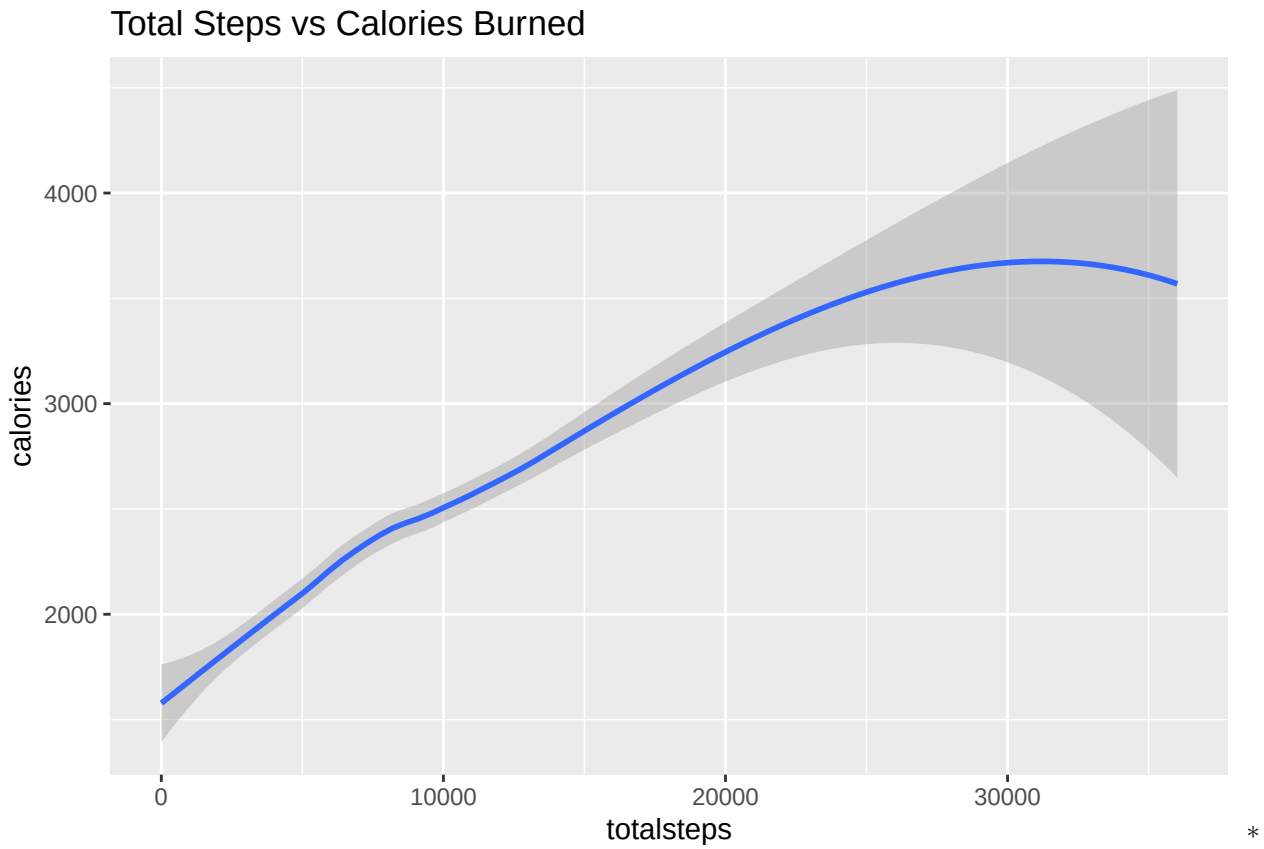
- Merge the two data frames by Id and date, then verify merge was successful

```
combined <- merge(data, sleep, by = c("id","date"))
```

Analyze

- Observations when looking at data summaries:
  - Comparing total steps: majority of steps are between 0 and 15,000 steps; average is 7,638
  - Regarding logged activities: looks like barely anyone is logging activities
  - Comparing different levels of activity distance: the majority is lightly active
  - Comparing different levels of activity minutes: mean of 991 sedentary minutes (very high), with lightly active minutes being the highest of the different active levels
  - Comparing total minutes asleep, and total time spent in bed: average is 7 hours of sleep, being in bed 7.6 hours
- According to this article (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3197470/), a reasonable, minimal amount of "moderate-to-vigorous physical acvivity (MVPA)" is 7,000-11,000 steps a day; the very low end of that corresponds to the average number of steps reflected from this data set.
- Plotting to explore observations:
- 1. Total steps vs. Calories burned; as predicted, the more steps taken corresponds to more calories burned.

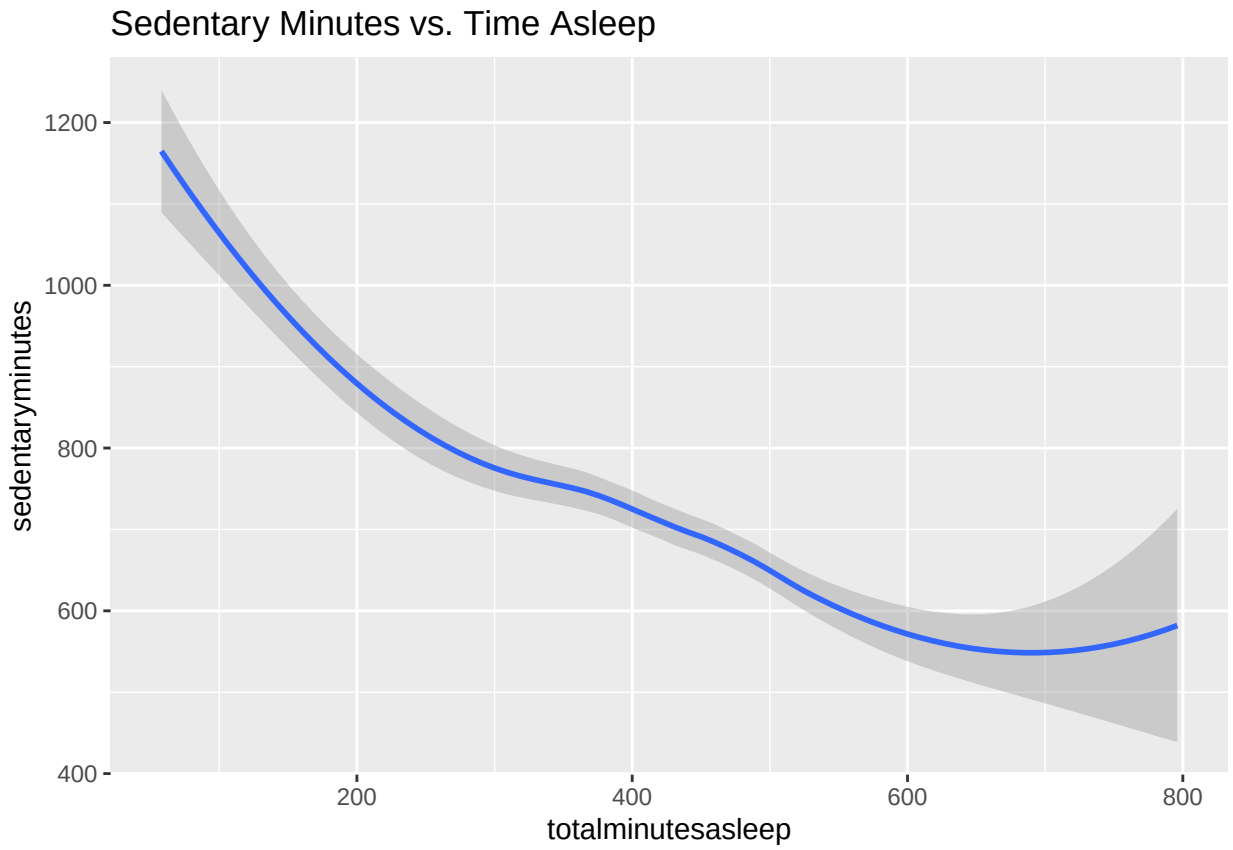
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



2. How total minutes asleep corresponds to amount of sedentary minutes; there is a clear correlation between the amount of sedentary minutes with decreased total minutes asleep.

```
ggplot(data=combined) +
  geom_smooth(mapping=aes(x=totalminutesasleep, y=sedentaryminutes)) +
  labs(title = "Sedentary Minutes vs. Time Asleep")
```

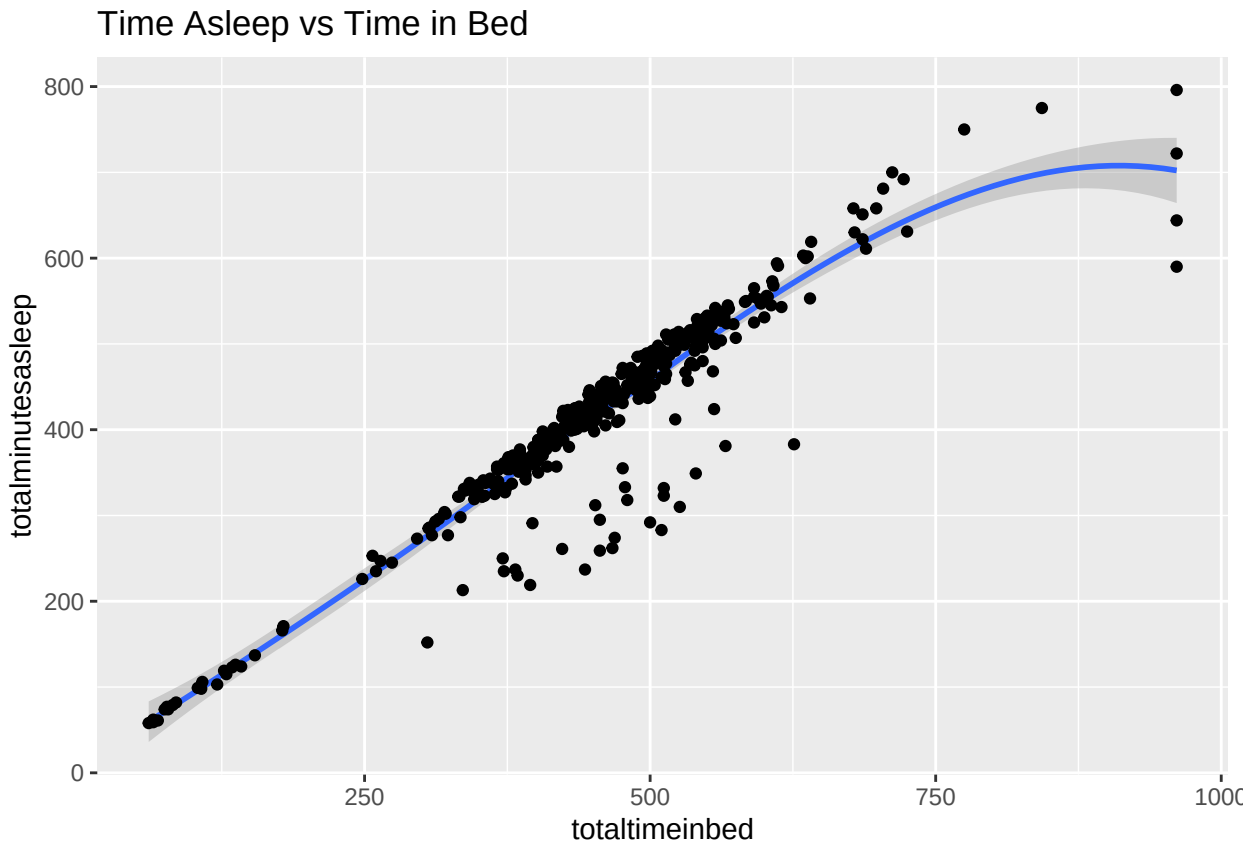
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



- 3. Total time asleep vs. total time in bed

```
ggplot(data=combined, aes(x=totaltimeinbed, y=totalminutesasleep)) +
  geom_smooth() + geom_point() + labs(title = "Time Asleep vs Time in Bed")
```

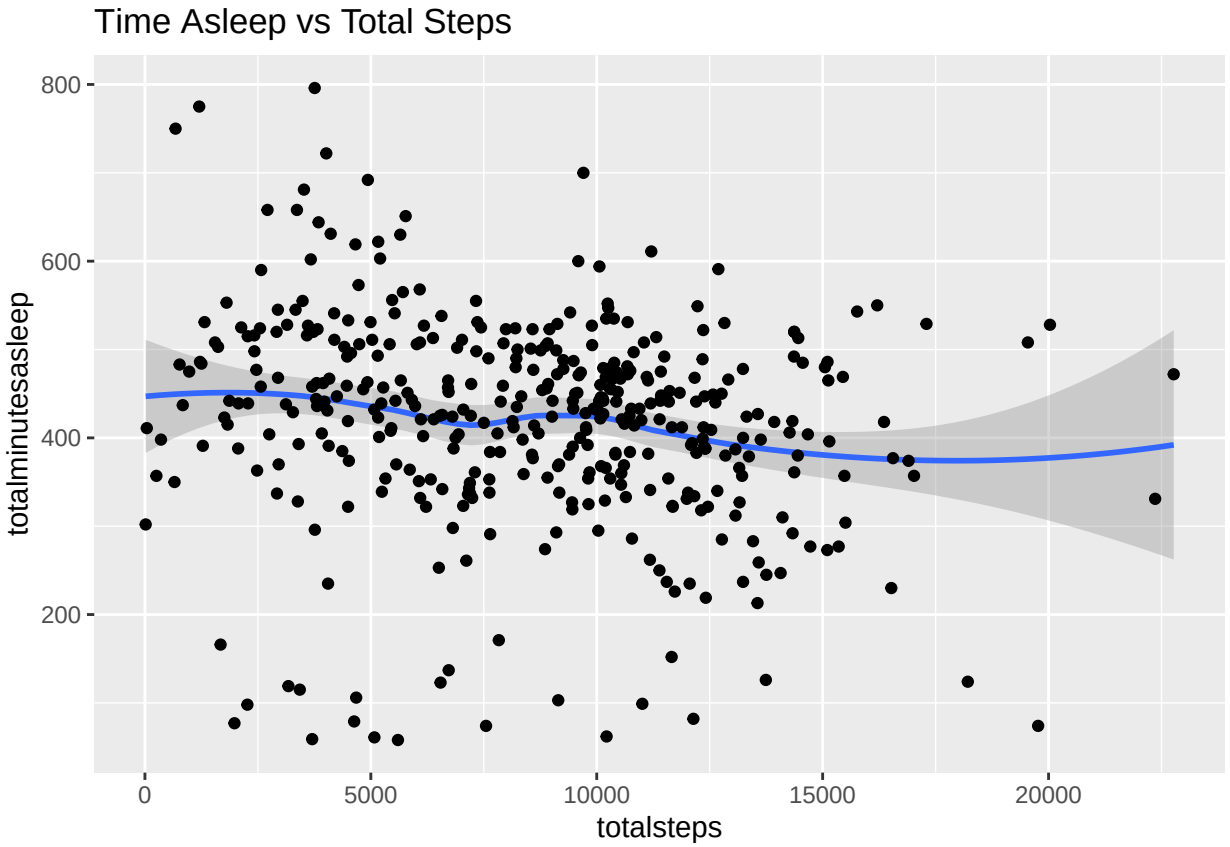
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



4. Total steps vs. Total time asleep; no correlation

```
ggplot(data=combined, aes(x=totalsteps, y=totalminutesasleep)) +
  geom_smooth() + geom_point() + labs(title = "Time Asleep vs Total Steps")
```

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



Share **Recommendations based off of analysis**

The average steps recorded are at the very low end of the recommended range, and the majority of activity tracked is sedentary. The data clearly shows that users aren't utilizing the "log activities" function, so we could encourage users to log more of their activities and share them to their social media through the app to increase the minutes and intensities of their exercise. There could be fun visualizations or rewards through the app to incentivize.

There is a clear correlation between total steps taken and calories burned, however, there isn't a clear relationship between the total amount of sleep one gets and the amount of steps taken. I interpret this to mean that it's not the number of steps taken, but the intensity of activity that influences the amount of sleep one gets. This is supported by seeing how a high amount of sedentary minutes affects the amount of sleep in the participant's data. Users can be encouraged to increase their step intensities to improve their sleep quality.