Bellabeat Smart Device Usage Analysis

Nate Mulder

2023-03-30

Introduction

Bellabeat is a wellness company for women interested in utilizing high-tech health focused devices to monitor and encourage healthy lifestyles. The company has been successful in its early years, and is looking to increase its market share position itself as a competitor to other wellness device brands like Fitbit.

Bellabeat offers five products:

- **Leaf**: This wellness tracker can be worn as a bracelet, necklace, or clip. It connects to the Bellabeat app to track activity, sleep, and stress.
- **Time**: This wellness watch is equipped with smart technology to track user activity, sleep, and stress. Paired with the watch, it provides additional insights into users' daily wellness.
- **Spring**: This is water bottle tracks daily water intake using smart technology to ensure optimal hydration throughout the day.
- **App**: The Bellabeat app provides users with health data related to their activity, sleep, stress, menstrual cycle, and mindfulness habits, and is fully integratied with the Leaf, Time, and Spring devices.
- **Membership**: Bellabeat's subcription based membership program gives users 24/7 access to fully personalized guidance on nutrition, activity, sleep, health and beauty, and mindfulness based on their lifestyle and goals.

ASK

Business Task

Help marketing executives at Bellabeat to understand trends in smart-device uses and where Bellabeat users fall amidst those trends. Provide useful, clear insights to drive data-driven marketing decisions for Bellabeat's continued growth.

Stakeholders

- Urška Sršen: Bellabeat's cofounder and Chief Creative Officer
- Sando Mur: Mathematician and Bellabeat's cofounder; key member of the Bellabeat executive team

Core Questions

As a junior data scientist working my team has been tasked with answering the following three questions:

- **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?**

Understanding the trends in smart device usage and how they apply to Bellabeat customers will allow the company to see where it falls ahead and behind compared to competitors in the same market segment. It can help to identify whether owners of one product might be likely to find value in other Bellabeat products, and tailor the app to better meet the needs of different segments of their user population. Ultimately, if Bellabeat understands how different segments of their users utilize the app, they can tailor marketing campaigns that are most appealing to each demographic range.

PREPARE

Loading Packages

The following packages were used to organize, process, and analyze the dataset.

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.1     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
to become errors
```

```
library(readr)
library(here)
```

```
## here() starts at /cloud/project
```

```
library(janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(skimr)
library(dplyr)
library(ggplot2)
```

Importing Data

Data Structure: The data comes from 30 Fitbit users who submitted their personal user data to an online survey.

License: The data was imported from Kaggle under a public license provided by Mobius (CC0:Public Domain) https://www.kaggle.com/datasets/arashnic/fitbit?resource=download (https://www.kaggle.com/datasets/arashnic/fitbit?resource=download)

The data was downloaded as a zipped file, stored in a finder folder, and each csv was imported into RStudio using the read.csv command. I opted to analyze the daily activity, calories, intensities and steps as well as the houly calories, intensities, and steps, and the weight and sleep data. Having daily and hourly data along with the correlations to weight and sleep should allow me to derive useful trend information.

```
daily_activity <- read.csv("dailyActivity_merged.csv")
daily_calories <- read.csv("dailyCalories_merged.csv")
hourly_calories <- read.csv("hourlyCalories_merged.csv")
daily_intensities <- read.csv("dailyIntensities_merged.csv")
hourly_intensities <- read.csv("hourlyIntensities_merged.csv")
daily_steps <- read.csv("dailySteps_merged.csv")
hourly_steps <- read.csv("hourlySteps_merged.csv")
weight <- read.csv("weightLogInfo_merged.csv")
sleep <- read.csv("sleepDay_merged.csv")
```

ROCCC ANALYSIS

Reliable: **LOW** The data seems at first to be reliable because it includes raw data from working 30 Fitbit trackers. However, there are discrepensies in the number of user ids meaning that

Original: **LOW** The data is from a third party survey conducted by Amazon Mechanical Turk. It comes from thirty uses who consented to share their fitbit data with this online survey.

Comprehensive: **MEDIUM** The data is relatively comprehensive. It includes 60 days of health data including steps taken, hours slept, weight data, calories used and intensity. However it includes only thirty users, and does not include demographic data like age, gender, or previous known health markers.

Current: **LOW** The data is seven years old, so it is no longer current. However, the use of smart devices like Fitbits has likely not changed dramatically in those years, so the usage trends may still be useful.

Cited: **LOW** This datasets primary use has been for training data analysts like myself. It does not appear to have been cited by other businesses or agencies.

Important Limitations

- There are discrepancies in the number of user ids across the data. There were 30 people who completed the survey, but some datasets have as many as 33 or as few as 8 distinct user ids. This means that some users created multiple ids and some users declined to track or share all of their health data.

- The users who provided their data are not a representative sample of Americans, and are likely not a perfect match for Bellabeat's users. The users were on average 50 lbs lighter than the average American, and Bellabeat is a wellness-technology brand oriented towards women, but the gender and other demographic data of the users in this dataset were not provided.

PROCESS

The data was initially processed using the head(), skim(), colnames(), glimpse(), and skim_without_charts() functions.

```
glimpse(daily_activity)
```

```
## Rows: 940
## Columns: 15
## $ Id                       <dbl> 1503960366, 1503960366, 1503960366, 150396036…
## $ ActivityDate             <chr> "4/12/2016", "4/13/2016", "4/14/2016", "4/15/…
## $ TotalSteps               <int> 13162, 10735, 10460, 9762, 12669, 9705, 13019…
## $ TotalDistance            <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9.8…
## $ TrackerDistance          <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9.8…
## $ LoggedActivitiesDistance <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
## $ VeryActiveDistance       <dbl> 1.88, 1.57, 2.44, 2.14, 2.71, 3.19, 3.25, 3.5…
## $ ModeratelyActiveDistance <dbl> 0.55, 0.69, 0.40, 1.26, 0.41, 0.78, 0.64, 1.3…
## $ LightActiveDistance      <dbl> 6.06, 4.71, 3.91, 2.83, 5.04, 2.51, 4.71, 5.0…
## $ SedentaryActiveDistance  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
## $ VeryActiveMinutes        <int> 25, 21, 30, 29, 36, 38, 42, 50, 28, 19, 66, 4…
## $ FairlyActiveMinutes      <int> 13, 19, 11, 34, 10, 20, 16, 31, 12, 8, 27, 21…
## $ LightlyActiveMinutes     <int> 328, 217, 181, 209, 221, 164, 233, 264, 205, …
## $ SedentaryMinutes         <int> 728, 776, 1218, 726, 773, 539, 1149, 775, 818…
## $ Calories                 <int> 1985, 1797, 1776, 1745, 1863, 1728, 1921, 203…
```

The three most useful data tables were checked to ensure that each had the expected 30 unique user ids.

```
n_distinct(daily_activity$Id)
```

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

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

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

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

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

Surprisingly, there were 33 unique user ids for daily_activity, only 24 for sleep, and merely 8 for weight. This suggests that some users created multiple ids for their fitbit trakers, and that some users declined to share their sleep and weight data. This makes the data less reliable because it some users will be over-represented in each category, and a sample side of 8 is not large enough to reliably base user trends off of.

I realized that the three dataset I wanted to merge, daily_activity, sleep, and weight, had their date and time columns named and formatted differently.

Fixing Formatting

```
#weight
weight$Date=as.POSIXct(weight$Date, format="%m/%d/%Y %I:%M:%S %p", tz=Sys.timezone())
weight$time <- format(weight$Date, format = "%H:%M:%S")
weight$date <- format(weight$Date, format = "%m/%d/%y")
# activity
daily_activity$ActivityDate=as.POSIXct(daily_activity$ActivityDate, format="%m/%d/%Y", tz=Sys.timezone())
daily_activity$date <- format(daily_activity$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")
#intensities
hourly_intensities$ActivityHour=as.POSIXct(hourly_intensities$ActivityHour, format="%m/%d/%Y %I:%M:%S %p", tz=Sys.timezone())
hourly_intensities$time <- format(hourly_intensities$ActivityHour, format = "%H:%M:%S")
hourly_intensities$date <- format(hourly_intensities$ActivityHour, format = "%m/%d/%y")
```

I merged the three datasets, daily_activity, sleep, and weight, by id and date.

```
merged_data1 <- merge(sleep, daily_activity, by=c('Id', 'date'))
merged_data <- merge(merged_data1, weight, by=c('Id', 'date'))
glimpse(merged_data)
```

```
## Rows: 35
## Columns: 28
## $ Id                       <dbl> 1503960366, 1503960366, 1927972279, 455860992…
## $ date                     <chr> "05/02/16", "05/03/16", "04/13/16", "05/01/16…
## $ SleepDay                 <dttm> 2016-05-02, 2016-05-03, 2016-04-13, 2016-05-…
## $ TotalSleepRecords        <int> 1, 1, 1, 1, 1, 1, 3, 2, 1, 1, 1, 1, 1, 1, 1, …
## $ TotalMinutesAsleep       <int> 277, 273, 398, 115, 549, 366, 630, 508, 370, …
## $ TotalTimeInBed           <int> 309, 296, 422, 129, 583, 387, 679, 535, 386, …
## $ ActivityDate             <dttm> 2016-05-02, 2016-05-03, 2016-04-13, 2016-05-…
## $ TotalSteps               <int> 14727, 15103, 356, 3428, 12231, 10199, 5652, …
## $ TotalDistance            <dbl> 9.71, 9.66, 0.25, 2.27, 9.14, 6.74, 3.74, 1.0…
## $ TrackerDistance          <dbl> 9.71, 9.66, 0.25, 2.27, 9.14, 6.74, 3.74, 1.0…
## $ LoggedActivitiesDistance <dbl> 0.000000, 0.000000, 0.000000, 0.000000, 0.000…
## $ VeryActiveDistance       <dbl> 3.21, 3.73, 0.00, 0.00, 5.98, 3.40, 0.57, 0.0…
## $ ModeratelyActiveDistance <dbl> 0.57, 1.05, 0.00, 0.00, 0.83, 0.83, 1.21, 0.0…
## $ LightActiveDistance      <dbl> 5.92, 4.88, 0.25, 2.27, 2.32, 2.51, 1.96, 1.0…
## $ SedentaryActiveDistance  <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.0…
## $ VeryActiveMinutes        <int> 41, 50, 0, 0, 200, 50, 8, 0, 0, 50, 5, 13, 35…
## $ FairlyActiveMinutes      <int> 15, 24, 0, 0, 37, 14, 24, 0, 0, 3, 13, 42, 41…
## $ LightlyActiveMinutes     <int> 277, 254, 32, 190, 159, 189, 142, 86, 217, 28…
## $ SedentaryMinutes         <int> 798, 816, 986, 1121, 525, 796, 548, 862, 837,…
## $ Calories                 <int> 2004, 1990, 2151, 1692, 4552, 1994, 1718, 146…
## $ Date                     <dttm> 2016-05-02 23:59:59, 2016-05-03 23:59:59, 20…
## $ WeightKg                 <dbl> 52.6, 52.6, 133.5, 69.9, 90.7, 62.5, 62.1, 61…
## $ WeightPounds             <dbl> 115.9631, 115.9631, 294.3171, 154.1031, 199.9…
## $ Fat                      <int> 22, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ BMI                      <dbl> 22.65, 22.65, 47.54, 27.32, 28.00, 24.39, 24.…
## $ IsManualReport           <chr> "True", "True", "False", "True", "False", "Tr…
## $ LogId                    <dbl> 1.462234e+12, 1.462320e+12, 1.460510e+12, 1.4…
## $ time                     <chr> "23:59:59", "23:59:59", "01:08:52", "23:59:59…
```

ANALYZE

In starting my analysis I was interested in examining:

- **How do the surveyed users compare with national norms?**
- **What times of day are users most and least active?**
- **Are greater amounts of daily activity directly correlated with greater amounts of sleep?**

I started by examining the most relevant columns from my merged dataset and looking for outliers and averages using the summary() function.

```
merged_data %>%
  dplyr::select(
         date,
         TotalSteps,
         TotalDistance,
         VeryActiveMinutes,
         FairlyActiveMinutes,
         LightlyActiveMinutes,
         SedentaryMinutes,
         Calories,
         TotalMinutesAsleep,
         TotalTimeInBed,
         WeightPounds,
         BMI
         ) %>%
  summary()
```

```
##      date             TotalSteps    TotalDistance    VeryActiveMinutes
##  Length:35          Min.   :  356   Min.   : 0.250   Min.   :  0.00   
##  Class :character   1st Qu.: 5780   1st Qu.: 3.825   1st Qu.:  0.00   
##  Mode  :character   Median :10524   Median : 6.960   Median : 18.00   
##                     Mean   : 9687   Mean   : 6.523   Mean   : 27.49   
##                     3rd Qu.:12484   3rd Qu.: 8.730   3rd Qu.: 42.00   
##                     Max.   :20031   Max.   :13.240   Max.   :200.00   
##  FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes    Calories   
##  Min.   : 0.00       Min.   : 32.0        Min.   : 127.0   Min.   : 928  
##  1st Qu.: 3.50       1st Qu.:197.0        1st Qu.: 635.5   1st Qu.:1852  
##  Median :15.00       Median :240.0        Median : 689.0   Median :2039  
##  Mean   :18.37       Mean   :236.5        Mean   : 688.5   Mean   :2052  
##  3rd Qu.:33.50       3rd Qu.:286.0        3rd Qu.: 736.0   3rd Qu.:2168  
##  Max.   :42.00       Max.   :369.0        Max.   :1121.0   Max.   :4552  
##  TotalMinutesAsleep TotalTimeInBed   WeightPounds        BMI       
##  Min.   :115.0      Min.   :129.0   Min.   :116.0   Min.   :22.65  
##  1st Qu.:399.0      1st Qu.:420.0   1st Qu.:134.9   1st Qu.:23.89  
##  Median :442.0      Median :455.0   Median :135.6   Median :24.00  
##  Mean   :430.3      Mean   :449.8   Mean   :141.5   Mean   :24.83  
##  3rd Qu.:472.5      3rd Qu.:494.0   3rd Qu.:136.5   3rd Qu.:24.17  
##  Max.   :630.0      Max.   :679.0   Max.   :294.3   Max.   :47.54  
```

I did not see any minimum or maximum values that appeared to be erroneous by either being much too high or much too low. I compared the average weight, BMI, and minutes asleep against national averages and USDA reccomendations and notived a few important findings:

- **The average BMI for surveyed group (24.83) was lower than the average American BMI (26.5) by 1.65 points.**
- **The average weight for the surveyed group (141) was 50 pounds lower than the average weight for an American! (191 pounds)**https://www.cdc.gov/nchs/fastats/body-measurements.htm (https://www.cdc.gov/nchs/fastats/body-measurements.htm)
- **The total time asleep for the surveyed group (7 hours 10 minutes) was just slightly over the CDC recommendation for adults to receive at least 7 hours of sleep each night**

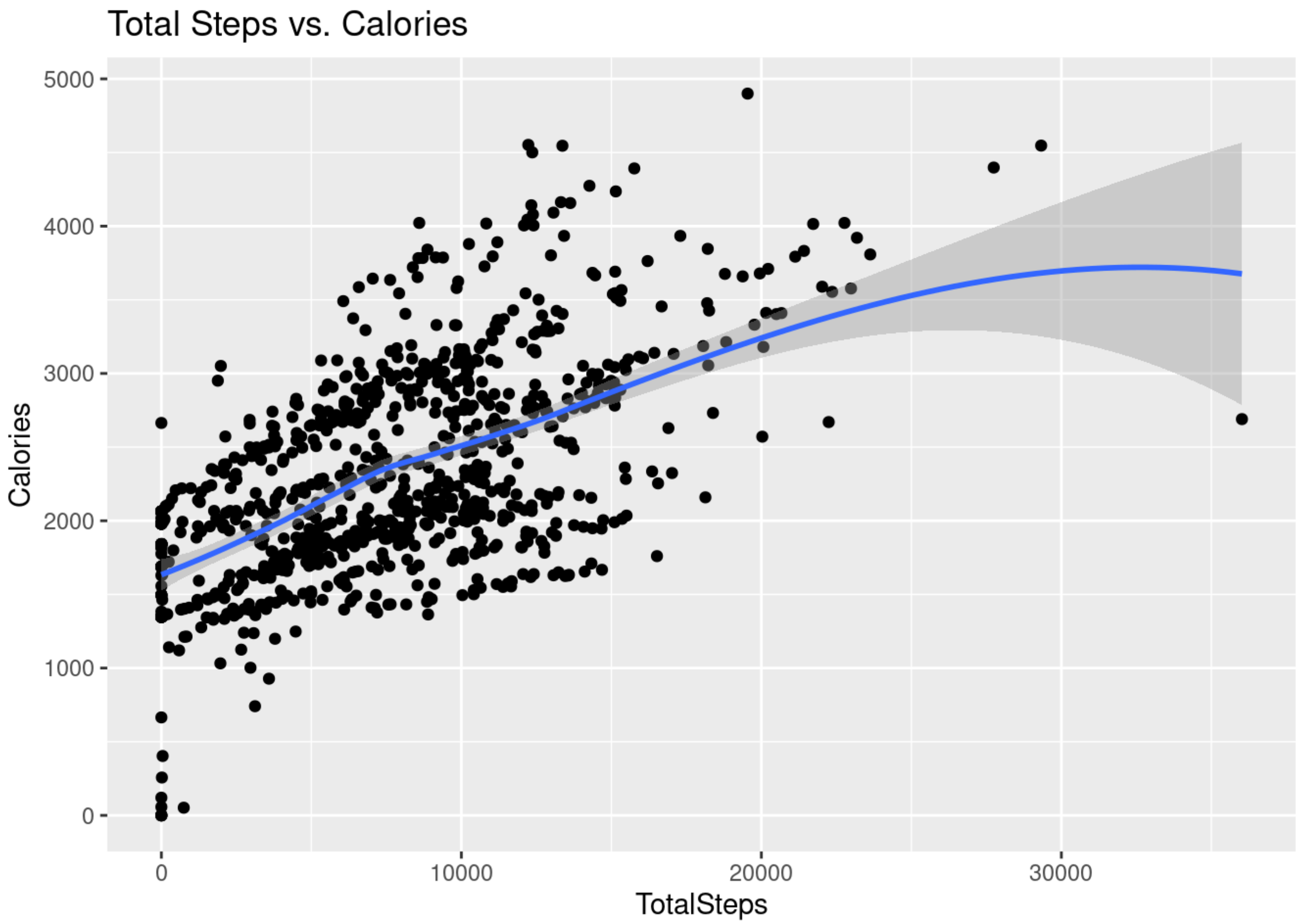
Visualizing the Data

I knew that I would need to create several visualizations to better understand the data.

I started by plotting total steps vs calories.

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

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



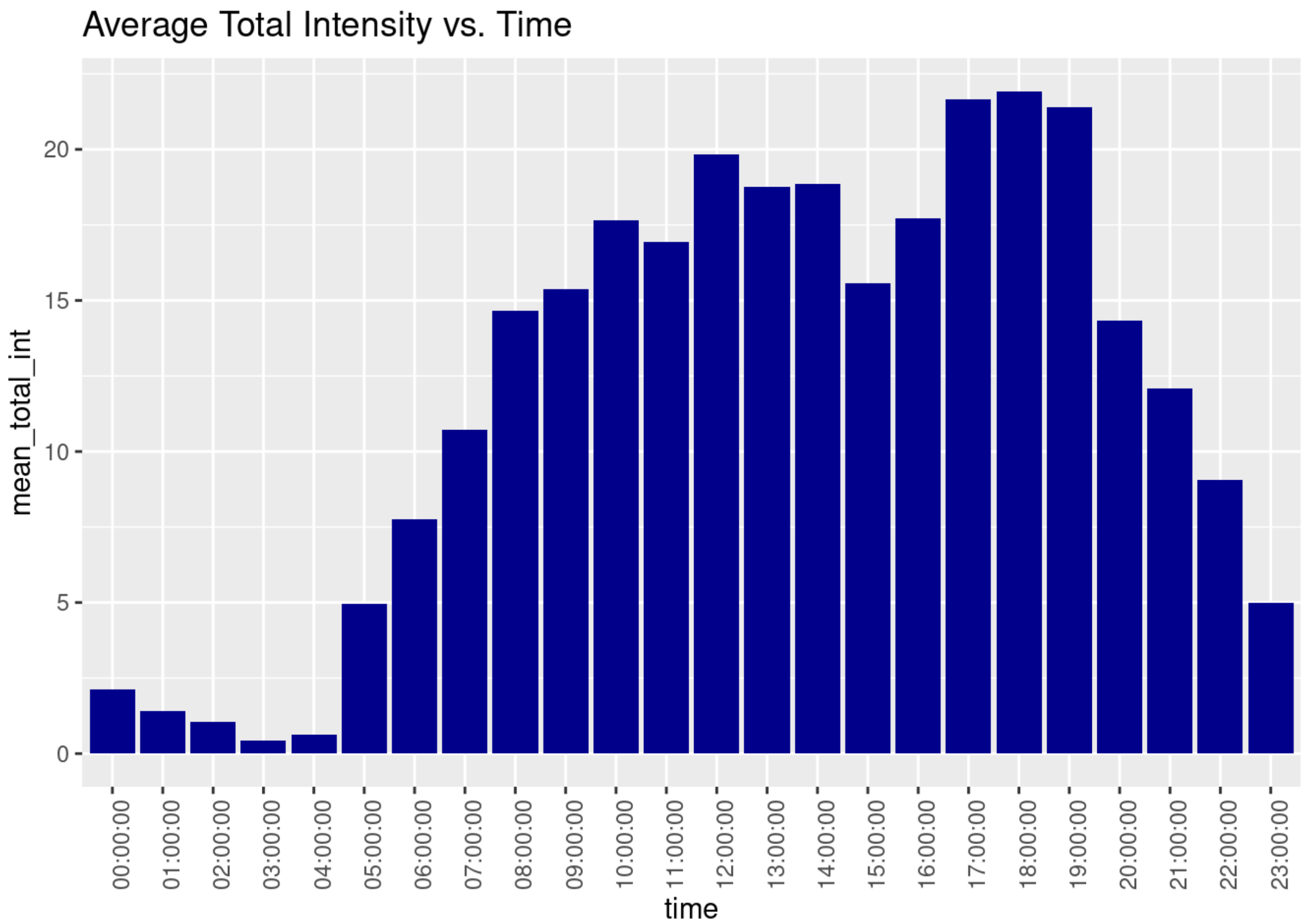
This positive correlation shows the expected result that increased activity seems to cause an increase in calories burned.

Knowing this, I wanted to see when users were most likely to be active. I created the following visualization plotting the intesity of users' activity against the time of day.

```
int_new <- hourly_intensities %>%
  group_by(time) %>%
  drop_na() %>%
  summarise(mean_total_int = mean(TotalIntensity))

ggplot(data=int_new, aes(x=time, y=mean_total_int)) + geom_histogram(stat = "identity",
fill='darkblue') +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(title="Average Total Intensity vs. Time")
```

```
## Warning in geom_histogram(stat = "identity", fill = "darkblue"): Ignoring
## unknown parameters: `binwidth`, `bins`, and `pad`
```

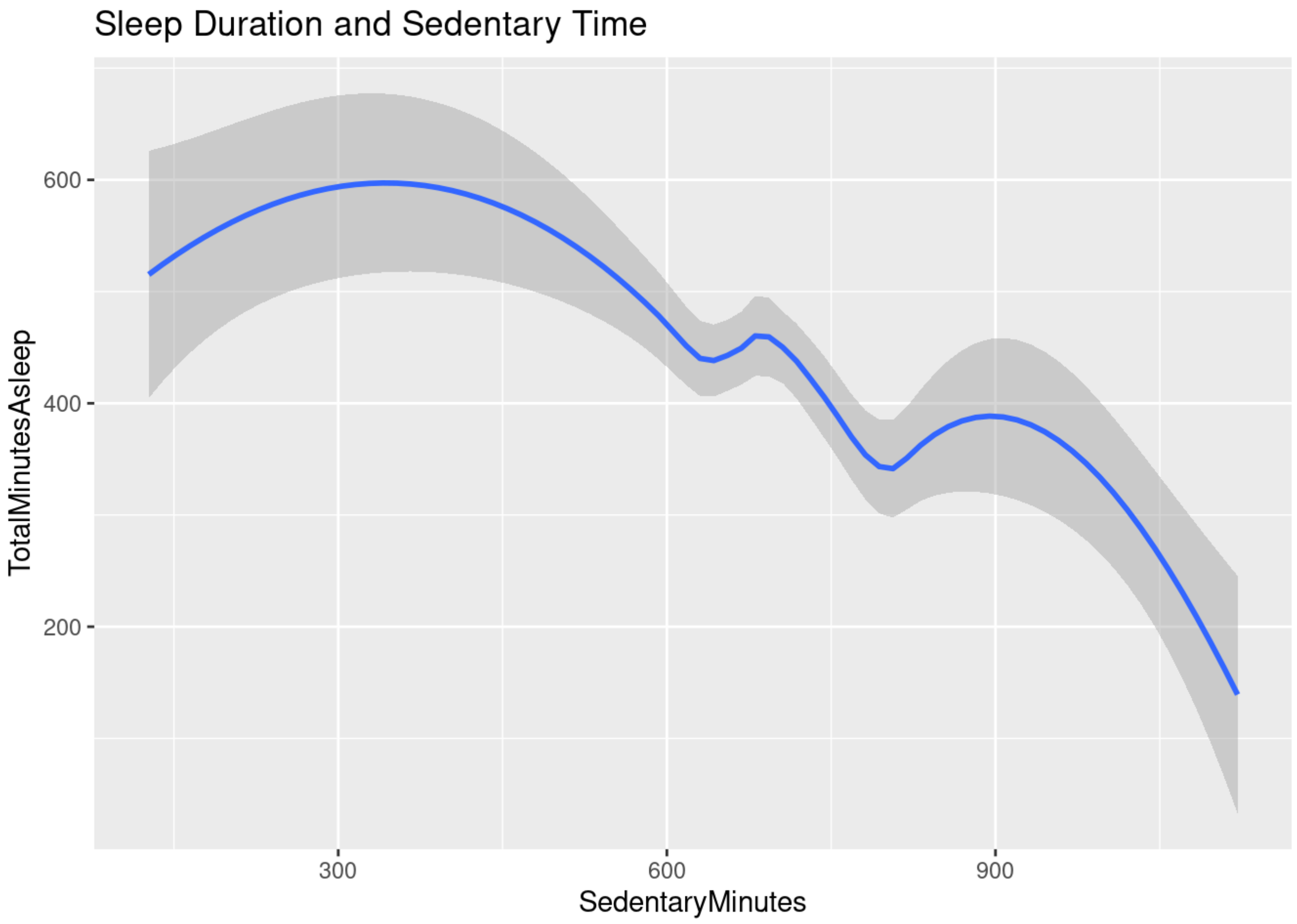


Based on this visualization, users are most likely to be active between 5:00 and 7:00 pm. Most likely this means that users are exercizing, visiting gyms, running, or doing other strenuous activity in these evening hours. Bellabeat can time its reminders and suggestions for the beginning of this time frame to remind users to stay active each day.

I was curious about how staying inactive would affect users, so I created a visualization plotting minutes asleep against minutes spent sedentary.

```
ggplot(data = merged_data, mapping = aes(x = SedentaryMinutes, y = TotalMinutesAsleep))
+
  geom_smooth() + labs(title= "Sleep Duration and Sedentary Time")
```

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



Based on this plot, we can see a clear negative correlation between total sedentery minutes and sleep time. This means that the Bellabeat could prompt users to hit a milestone of active minutes each day, in order to improve their sleep.

SHARE

CONCLUSIONS/SUMMARY OF INSIGHTS

There were strong correlations between steps taken and calories burned. Bellabeat can use its apps to encourage its users to increase their steps taken each day to meet their desired health outcomes. Users were most likely to be active from 5:00 to 7:00 each day, so that would likely be the highest impact time for Bellabeat to send reminders about workouts or health suggestions. Additionally, there is a strong correlation between the time that users spent sedentary and the amount of sleep they were able to achieve each night. This would be great information to share with users, and Bellabeat could benefit from a messaging campaign based around healthy days leading to healthy nights of sleep.

RECCOMENDATIONS

- **First, Bellabeat ought to seek out additional data of smart-health device users. The Fitbit dataset was not comprehensive enough to reliably analyze health trends.**

- **Second, Bellabeat can utilize the data that show that the 5:00 to 7:00 window is the most common time that users will be active. As users finish up work, Bellabeat can send targeted suggestions about workouts, reminders about users' goals, and encouragement to stay active each day.**
- **Finally, Bellabeat should communicate that its products can help users to improve their sleep. An ad campaing focused on how "healthy days lead to restful nights" might be an effective way to communicate the health and sleep benefit found in consistently utilizing Bellabeat's line of smart-wellness products**