Google Data Analytics Capstone Project

Zainab

2022-06-06

Scenario

I am a junior data analyst working on the marketing analyst team at Bellabeat, a high-tech manufacturer of health-focused products for women. Bellabeat is a successful small company, but they have the potential to become a larger player in the global smart device market. Urška Sršen, cofounder and Chief Creative Officer of Bellabeat, believes that analyzing smart device fitness data could help unlock new growth opportunities for the company. I have been asked to focus on one of Bellabeat's products and analyze smart device data to gain insight into how consumers are using their smart devices. The insights i discover will then help guide marketing strategy for the company. I will present my analysis to the Bellabeat executive team along with my recommendations for Bellabeat's marketing strategy. To aid my analysis, I will go through the six phases of data analysis; **Ask**, **Prepare**, **Process**, **Analyze**, **Share** and **Act**.

Ask Phase

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?

Business Task:

Analyze smart device usage data in order to gain insight into how consumers use non-Bellabeat smart devices.

Stakeholders

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

Prepare

The data i woukd be using for this project is the FitBit fitness tracker data which is an open source public dataset containing personal fitness tracker from thirty fitbit users. It is organized in a long format which means each user has multiple rows of data. The dataset is reliable, original, comprehensive, cited but not current(It was updated a year ago) and it is open source so it is accessible. Since the dataset contains information on only 30 users, and the sampling technique was not cited, there could be issues of bias. I verified the integrity of the data using a pivot table in Google Sheets with an average of 30 columns. The data i would be importing include; * daily_activity - contains information about daily activity and total steps taken each day. * daily_intensity - contains information about sedentary levels * daily_sleep - contains information total sleep per day.

```
daily_activity <- read.csv("daily_activity.csv")
daily_intensity <- read.csv("daily_intensity.csv")
daily_sleep <- read.csv("daily_sleep.csv")
```

Process

I chose R for my analysis due to the volume of the data. This phase entails cleaning my data and making it ready for analysis. I will install the necessary packages for this purpose.

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

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library(tidyverse)
```

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

```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## 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()
```

Installing the necessary data cleaning packages

```
install.packages("here")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library("here")
```

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

```
install.packages("skimr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library("skimr")

install.packages("janitor")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library("janitor")
```

```
##
## Attaching package: 'janitor'
```

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

I would get the Summary of the data

```
head(daily_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
```

```
head(daily_intensity)
```

```
##           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(daily_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(daily_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 ...
```

```
str(daily_intensity)
```

```
## 'data.frame':    940 obs. of  10 variables:
##  $ Id                       : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ ActivityDay              : chr  "4/12/2016" "4/13/2016" "4/14/2016" "4/15/2016" ...
##  $ SedentaryMinutes         : int  728 776 1218 726 773 539 1149 775 818 838 ...
##  $ LightlyActiveMinutes     : int  328 217 181 209 221 164 233 264 205 211 ...
##  $ FairlyActiveMinutes      : int  13 19 11 34 10 20 16 31 12 8 ...
##  $ VeryActiveMinutes        : int  25 21 30 29 36 38 42 50 28 19 ...
##  $ SedentaryActiveDistance : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LightActiveDistance      : num  6.06 4.71 3.91 2.83 5.04 ...
##  $ ModeratelyActiveDistance: num  0.55 0.69 0.4 1.26 0.41 ...
##  $ VeryActiveDistance       : num  1.88 1.57 2.44 2.14 2.71 ...
```

```
str(daily_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 ...
```

Checking the data to ensure there are 30 unique participants

```
n_unique(daily_activity$Id)
```

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

```
n_unique(daily_intensity$Id)
```

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

```
n_unique(daily_sleep$Id)
```

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

Checking for duplicates

```
sum(duplicated(daily_activity))
```

```
## [1] 0
```

```
sum(duplicated(daily_intensity))
```

```
## [1] 0
```

```
sum(duplicated(daily_sleep))
```

```
## [1] 3
```

Now, removing duplicate data from the daily_sleep data

```
daily_sleep <- daily_sleep %>%
  distinct()
```

Confirming the duplicte has been removed

```
sum(duplicated(daily_sleep))
```

```
## [1] 0
```

Next, dropping any missing data

```
daily_activity <- daily_activity %>%
    drop_na()

daily_intensity <- daily_intensity %>%
    drop_na()

daily_sleep <- daily_sleep %>%
    drop_na()
```

Ensuring the consistency of the date and time columns

```
daily_activity <- daily_activity %>%
    mutate(ActivityDate = as.Date(ActivityDate, format = "%m/%d/%Y"))

daily_intensity <- daily_intensity %>%
    mutate(ActivityDay = as.Date(ActivityDay, format = "%m/%d/%Y"))

daily_sleep <- daily_sleep %>%
  mutate(date = as.Date(SleepDay,format ="%m/%d/%Y %I:%M:%S %p" , tz=Sys.timezone()))
```

I have ensured my data is clean. I will proceed to the Analyze phase

Analyze

Merging the first two dataframes together;

```
merged_activity <- merge(daily_activity, daily_sleep, by= "Id")
head(merged_activity)
```

```
##           Id ActivityDate TotalSteps TotalDistance TrackerDistance
## 1 1503960366   2016-05-07      11992          7.71            7.71
## 2 1503960366   2016-05-07      11992          7.71            7.71
## 3 1503960366   2016-05-07      11992          7.71            7.71
## 4 1503960366   2016-05-07      11992          7.71            7.71
## 5 1503960366   2016-05-07      11992          7.71            7.71
## 6 1503960366   2016-05-07      11992          7.71            7.71
##   LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance
## 1                        0              2.46                     2.12
## 2                        0              2.46                     2.12
## 3                        0              2.46                     2.12
## 4                        0              2.46                     2.12
## 5                        0              2.46                     2.12
## 6                        0              2.46                     2.12
##   LightActiveDistance SedentaryActiveDistance VeryActiveMinutes
## 1                3.13                       0                37
## 2                3.13                       0                37
## 3                3.13                       0                37
## 4                3.13                       0                37
## 5                3.13                       0                37
## 6                3.13                       0                37
##   FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories
## 1                  46                 175              833     1821
## 2                  46                 175              833     1821
## 3                  46                 175              833     1821
## 4                  46                 175              833     1821
## 5                  46                 175              833     1821
## 6                  46                 175              833     1821
##                SleepDay TotalSleepRecords TotalMinutesAsleep TotalTimeInBed
## 1 4/12/2016 12:00:00 AM                 1                327            346
## 2 4/13/2016 12:00:00 AM                 2                384            407
## 3 4/15/2016 12:00:00 AM                 1                412            442
## 4 4/16/2016 12:00:00 AM                 2                340            367
## 5 4/17/2016 12:00:00 AM                 1                700            712
## 6 4/19/2016 12:00:00 AM                 1                304            320
##         date
## 1 2016-04-12
## 2 2016-04-13
## 3 2016-04-15
## 4 2016-04-16
## 5 2016-04-17
## 6 2016-04-19
```

Summary Statistics to identify initial trends

```
merged_activity%>%
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.:144.0        1st Qu.: 659.0
##  Median :  8.00    Median : 10.00      Median :200.0        Median : 734.0
##  Mean   : 23.94    Mean   : 17.34      Mean   :199.8        Mean   : 799.4
##  3rd Qu.: 36.00    3rd Qu.: 24.00      3rd Qu.:258.0        3rd Qu.: 853.0
##  Max.   :210.00    Max.   :143.00      Max.   :518.0        Max.   :1440.0
```

From the summary statistics above, it shows that a lot of people spend most of their time sedentary rather than being active.

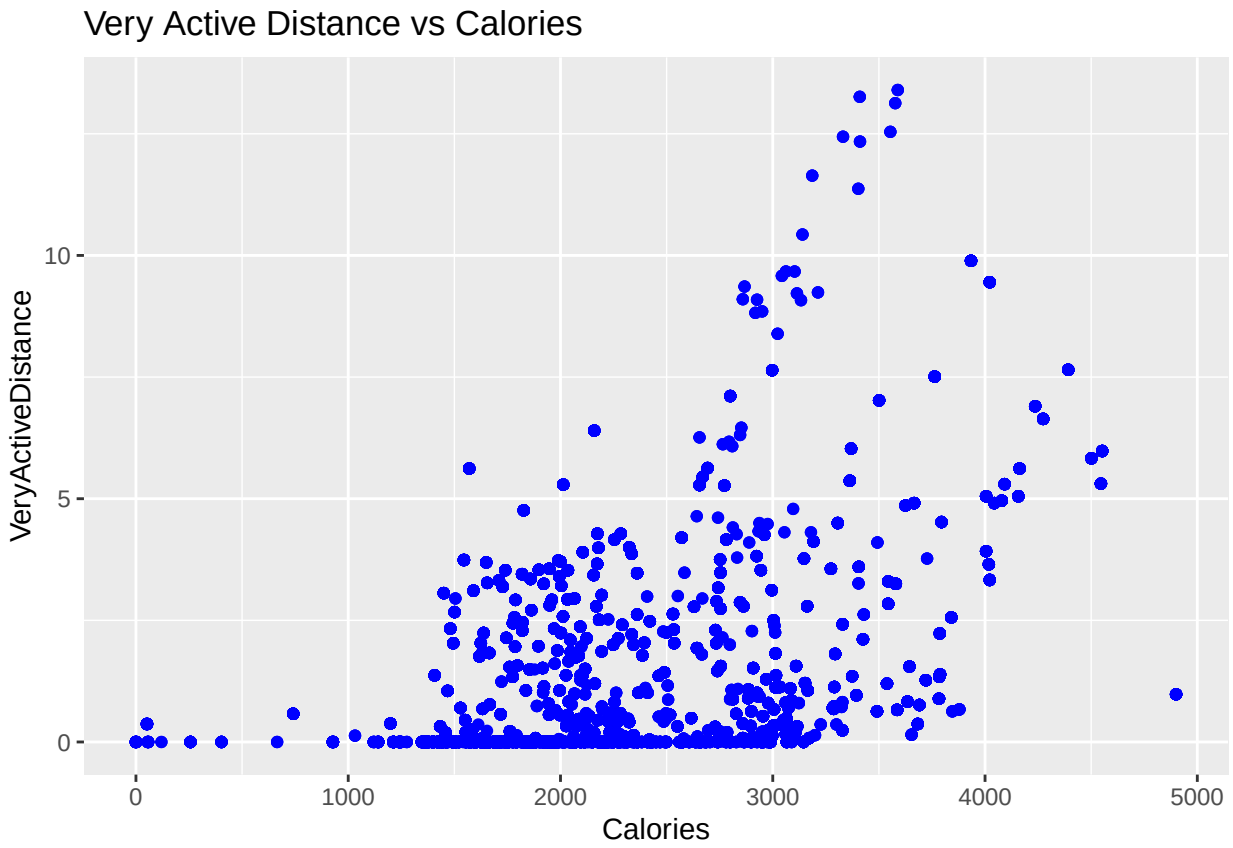
```
merged_activity%>%
select(VeryActiveDistance, ModeratelyActiveDistance, LightActiveDistance, SedentaryActiveDistance) %>%
  summary()
```

```
##  VeryActiveDistance ModeratelyActiveDistance LightActiveDistance
##  Min.   : 0.000     Min.   :0.0000           Min.   : 0.000
##  1st Qu.: 0.000     1st Qu.:0.0000           1st Qu.: 2.350
##  Median : 0.530     Median :0.4000           Median : 3.540
##  Mean   : 1.397     Mean   :0.7309           Mean   : 3.532
##  3rd Qu.: 2.310     3rd Qu.:1.0000           3rd Qu.: 4.830
##  Max.   :13.400     Max.   :6.4800           Max.   :10.300
##  SedentaryActiveDistance
##  Min.   :0.0000000
##  1st Qu.:0.0000000
##  Median :0.0000000
##  Mean   :0.0006795
##  3rd Qu.:0.0000000
##  Max.   :0.1100000
```

This also shows that most people go on light rather than very active walk.

Exploring an initial scatterplot;

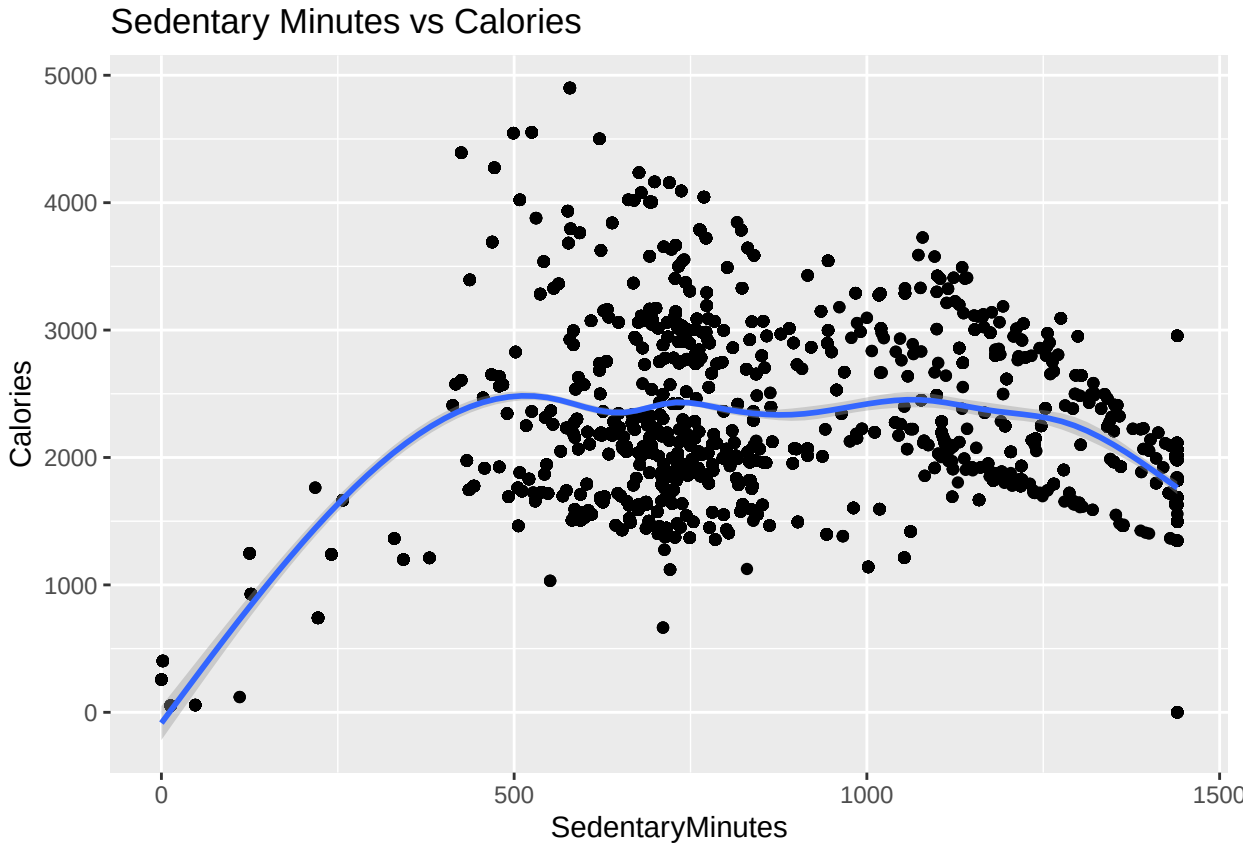
```
ggplot(data=merged_activity)+
  geom_point(mapping=aes(y=VeryActiveDistance, x= Calories), color='blue')+
  labs(title="Very Active Distance vs Calories")
```



People who take a high amount of calories do not tend to undergo very active distances.

```
ggplot(data=merged_activity,aes(x=SedentaryMinutes, y= Calories))+
  geom_point()+
  geom_smooth()+
  labs(title="Sedentary Minutes vs Calories")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Instead, people with a higher calorie intake tend to be more sedentary.

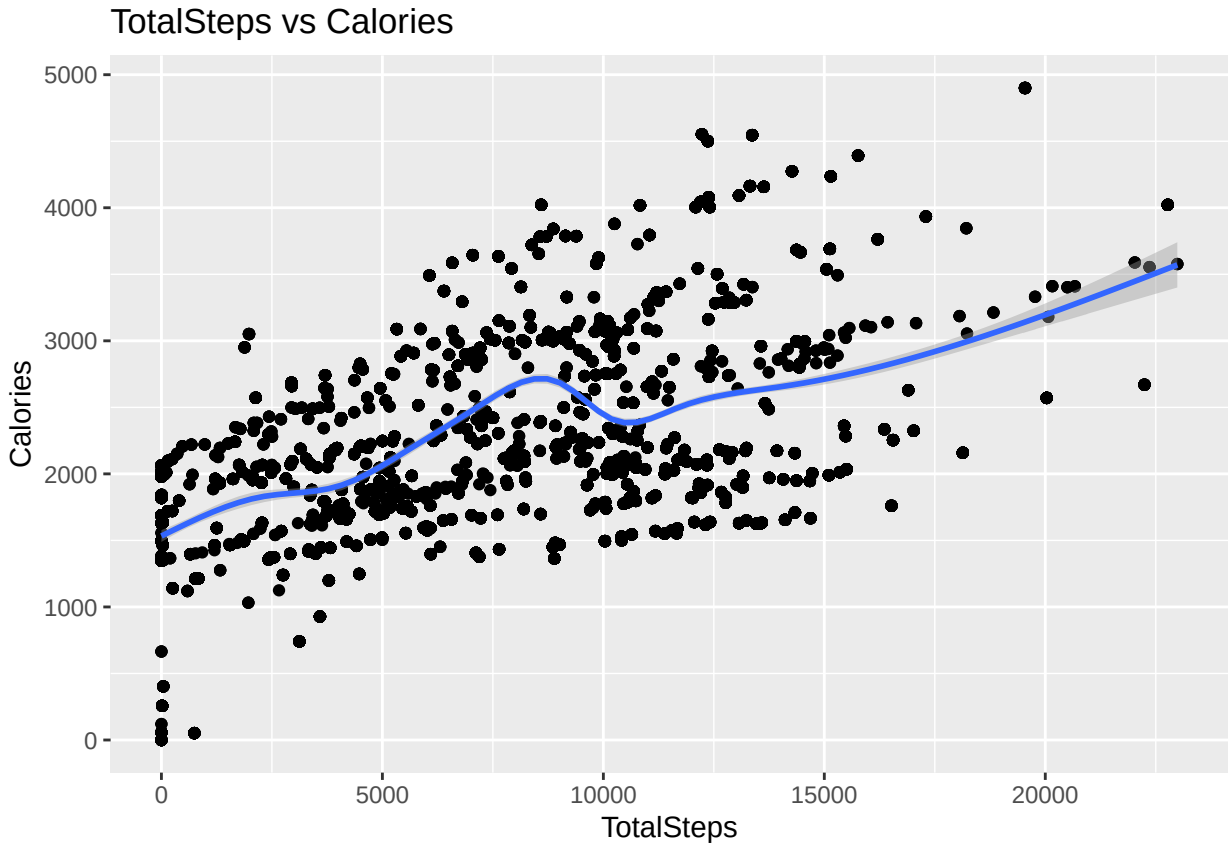
From the initial analysis; *A lot of people spend most of their time sedentary rather than being active.* Most people go on light rather than very active walk. *People who take a high amount of calories do not tend to undergo very active distances.* Instead, people with a higher calorie intake tend to be more sedentary.

Share

Here i will show more visualizations about the dataset. Relationship between Total Steps and Calories taken

```
ggplot(data=merged_activity,aes(x=TotalSteps, y= Calories))+
  geom_point()+
  geom_smooth()+
  labs(title="TotalSteps vs Calories")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

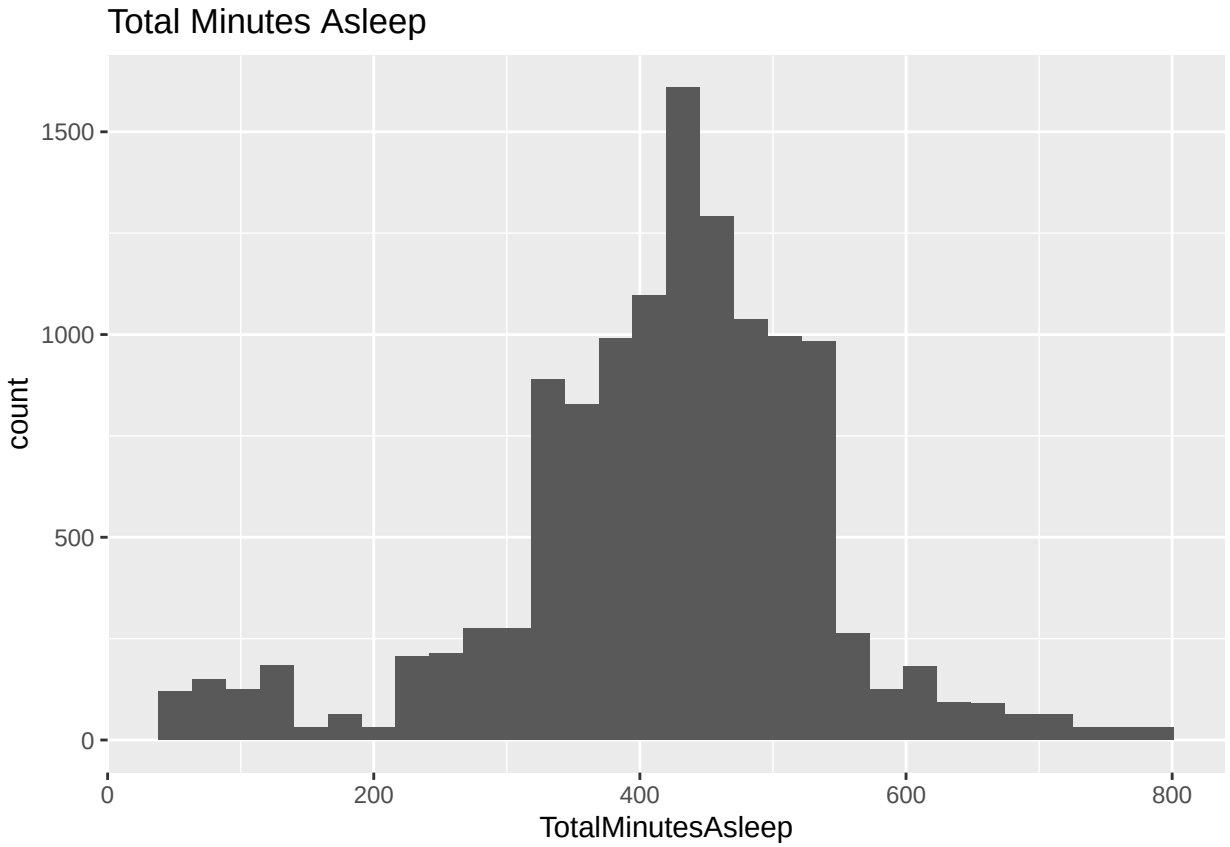


There is a positive correlation; More calories are burned with increasing steps.

I will explore the average minutes of sleep by the users using a histogram.

```
ggplot(data=merged_activity,aes(x= TotalMinutesAsleep))+
  geom_histogram()+
    labs(title="Total Minutes Asleep")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The average minutes of sleep is about 450 which is 8hrs 30mins.Though some people are yet to observe this milestone.

```
mean(merged_activity$TotalMinutesAsleep)
```

```
## [1] 419.1028
```

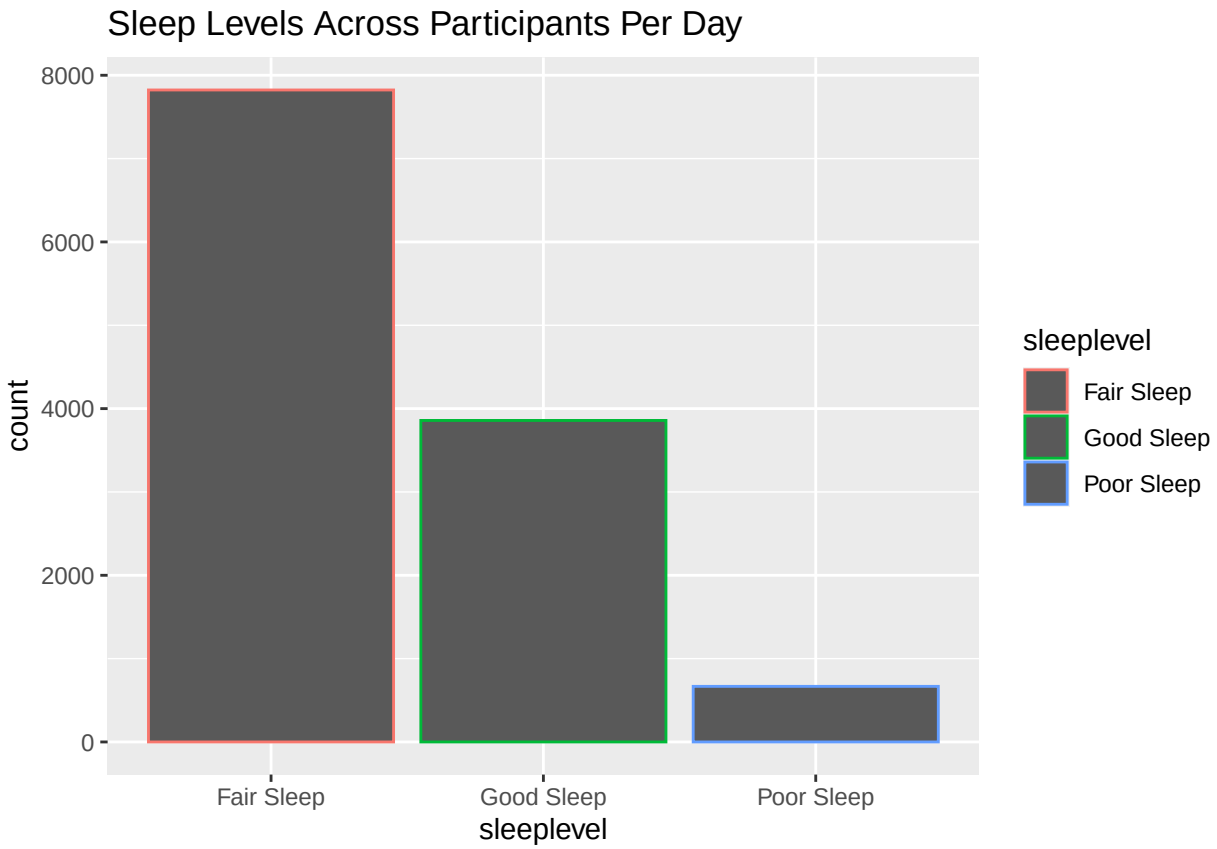
```
sd(merged_activity$TotalMinutesAsleep)
```

```
## [1] 118.9441
```

```
 sleep_activity <- merged_activity %>% mutate(sleeplevel = case_when(TotalMinutesAsleep<=200  ~ "Poor S
```

```
ggplot(data=sleep_activity, aes(x=sleeplevel))+
  geom_bar()+
  geom_bar(aes(color= sleeplevel, shape=sleeplevel))+
  labs(title = "Sleep Levels Across Participants Per Day")
```

```
## Warning: Ignoring unknown aesthetics: shape
```



Majority of people get about 6 hours of sleep each day while few people tend to get less than 3 hours of sleep daily.

Conclusion

- A lot of people spend most of their time sedentary rather than being active.
- Most people go on light rather than very active walk.
- People who take a high amount of calories do not tend to undergo very active distances.
- Instead, people with a higher calorie intake tend to be more sedentary.
- More calories are burned with increasing steps. *Majority of people get about 6 hours of sleep each day while few people tend to get less than 3 hours of sleep daily.

Recommendations/Act

Since the Bellabeat app provides usere with health data related to their activity, sleep, stress and mindfulness habits;

1. The app may improve sleep time for users by notifying them through phone alarms from their time in bed before they fall asleep.
2. The wellness watch can regulate activity levels by providing users their average activity level and the recommended health activity levels.
3. The Bellabeat Spring water bottle can also ensure very active users are properly hydrated.
4. The wellness watch can also provide insights on the amount of steps taken to encourage sedentary users to become more active.

This can help their users better understand their current habits and make healthier decisions.