Analysis on Health Data for Bellabeat

Nadia Shureih

2024-07-16

About Bellabeat

Bellabeat is a self described "pioneer in the fem-tech realm." Focused on finding an intersection between women's wellness and women's fashion, the company works to aid women in tracking a myriad of personal health data. Although a small company, founder Urška Sršen believes that analyzing health fitness data could unlock new areas of growth for the company.

Business Task

Analyze trends in health tracking smart device usage and provide recommendations and opportunity for growth

Downloading Packages

Importing Libraries

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

##
## Attaching package: 'janitor'

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

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.5
## v ggplot2   3.5.1     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
##
## Attaching package: 'kableExtra'
```

```
##
##
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

Cleaning Data

```r
sleep <- read.csv('sleepDay_merged.csv')
weight <- read.csv('weightLogInfo_merged.csv')
daily_activity <-read.csv('dailyActivity_merged.csv')
```

```r
head(sleep)
```

Taking a Look at the Data

```
##           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
```

```r
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
```

```r
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
```

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

Cleaning Duplicates

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

```r
sum(duplicated(sleep))
```

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

```r
sum(duplicated(weight))
```

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

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

sleep <- sleep %>%
  distinct() %>%
  drop_na()
```

```r
sum(duplicated(sleep))
```

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

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

sleep <- sleep %>%
```

```r
  rename(Date = SleepDay) %>%
  mutate(Date = as.Date(Date, format = "%m/%d/%y"))

weight <- weight %>%
  select(-LogId) %>%
  mutate(Date=as.Date(Date, format = "%m/%d/%y")) %>%
  mutate(IsManualReport = as.factor(IsManualReport))
```

Creating Cohesive Date Formatting

```r
clean_names(daily_activity)
clean_names(sleep)
clean_names(weight)
```

```r
daily_activity <- rename_with(daily_activity, tolower)
sleep <- rename_with(sleep, tolower)
weight <- rename_with(weight, tolower)
```

Cleaning Column Names

Summarizing the Datasets

How many participants are in each dataset? I wanted to get an understanding of how big each dataset is before developing my analysis further and focusing in on the specifics.

```r
n_distinct(daily_activity$id)
```

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

```r
n_distinct(sleep$id)
```

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

```r
n_distinct(weight$id)
```

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

There are 35 participants in the activity data, 24 in the sleep data, and 8 in weight data.

Getting a Summary of the Datasets

```r
daily_activity %>%
  select(totalsteps, veryactiveminutes, veryactivedistance, calories) %>%
  summary()
```

Daily Activity Summaries

```
##    totalsteps    veryactiveminutes veryactivedistance    calories
##  Min.   :    0   Min.   :  0.00    Min.   : 0.000     Min.   :   0
##  1st Qu.: 3790   1st Qu.:  0.00    1st Qu.: 0.000     1st Qu.:1828
##  Median : 7406   Median :  4.00    Median : 0.210     Median :2134
##  Mean   : 7638   Mean   : 21.16    Mean   : 1.503     Mean   :2304
##  3rd Qu.:10727   3rd Qu.: 32.00    3rd Qu.: 2.053     3rd Qu.:2793
##  Max.   :36019   Max.   :210.00    Max.   :21.920     Max.   :4900
```

```r
daily_activity %>%
  select(sedentaryminutes, sedentaryactivedistance, fairlyactiveminutes, moderatelyactivedistance, ligh
  summary()
```

```
##  sedentaryminutes sedentaryactivedistance fairlyactiveminutes
##  Min.   :   0.0   Min.   :0.000000        Min.   :  0.00
##  1st Qu.: 729.8   1st Qu.:0.000000        1st Qu.:  0.00
##  Median :1057.5   Median :0.000000        Median :  6.00
##  Mean   : 991.2   Mean   :0.001606        Mean   : 13.56
##  3rd Qu.:1229.5   3rd Qu.:0.000000        3rd Qu.: 19.00
##  Max.   :1440.0   Max.   :0.110000        Max.   :143.00
##  moderatelyactivedistance lightlyactiveminutes lightactivedistance
##  Min.   :0.0000           Min.   :  0.0        Min.   : 0.000
##  1st Qu.:0.0000           1st Qu.:127.0        1st Qu.: 1.945
##  Median :0.2400           Median :199.0        Median : 3.365
##  Mean   :0.5675           Mean   :192.8        Mean   : 3.341
##  3rd Qu.:0.8000           3rd Qu.:264.0        3rd Qu.: 4.782
##  Max.   :6.4800           Max.   :518.0        Max.   :10.710
```

Some Conclusions from the Above Data The average steps is 6547 (compare with the recommended 10,000 steps per day). The average calories burned is 2189 (though this is wholy dependent on age and gender).

Average amount of very active minutes is about 16.5, and very active distance is about 1.18m. Average amount of sedentary minutes is about 995.3 Average amount of fairly active minutes is about 13.07, and moderately active distance is about 0.48m. Average amount of lightly active minutes is about 170.1, and very active distance is about 2.89m.

People spend the most amount of time sedentary, and are able to complete more light activity next. People who complete light active are able to go longer distances For people who are active, they will probably spend more time in 'very active' than 'fairly active' meaning maybe they are working harder in those workouts.

```r
sleep %>%
  select(totalsleeprecords, totalminutesasleep, totaltimeinbed) %>%
  summary()
```

Sleep Summaries

```
##  totalsleeprecords totalminutesasleep totaltimeinbed
##  Min.   :1.00      Min.   : 58.0      Min.   : 61.0
##  1st Qu.:1.00      1st Qu.:361.0      1st Qu.:403.8
##  Median :1.00      Median :432.5      Median :463.0
##  Mean   :1.12      Mean   :419.2      Mean   :458.5
##  3rd Qu.:1.00      3rd Qu.:490.0      3rd Qu.:526.0
##  Max.   :3.00      Max.   :796.0      Max.   :961.0
```

The average time asleep is 419.2 minutes or about 6.9 hours the average time in bed is 458.5 minutes or about 7.5 hours.

```r
weight %>%
  select(weightpounds, fat, bmi) %>%
  summary()
```

Weight Summaries

```
##   weightpounds        fat             bmi
##  Min.   :116.0   Min.   :22.00   Min.   :21.45
##  1st Qu.:135.4   1st Qu.:22.75   1st Qu.:23.96
##  Median :137.8   Median :23.50   Median :24.39
##  Mean   :158.8   Mean   :23.50   Mean   :25.19
##  3rd Qu.:187.5   3rd Qu.:24.25   3rd Qu.:25.56
##  Max.   :294.3   Max.   :25.00   Max.   :47.54
##                  NA's   :65
```

The average weight is about 158.8lbs, average bmi is 25.19 (overweight), which is understandable as someone most likely to be tracking their weight would be someone who is overweight and trying to lose weight.

```r
merged_data <- merge(merge(daily_activity, sleep, by= c('id','date'), all= TRUE), weight, by = c('id',

merged_activity_sleep <- merge(daily_activity, sleep, by=c('id','date'))
```

Merging Data Sets

```r
merged_data <- merged_data %>%
  select(-c(trackerdistance, totalsleeprecords, weightkg, ismanualreport))

merged_activity_sleep <- merged_activity_sleep %>%
  select(-c(trackerdistance, totalsleeprecords))
```

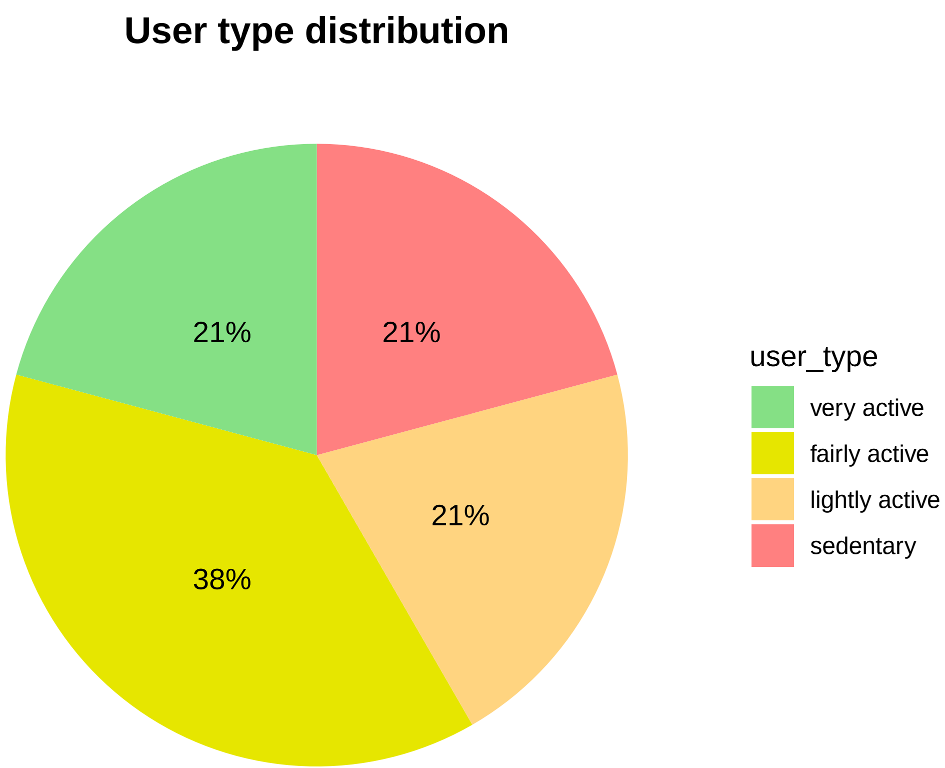
Removing Extra Variables Creating a data set which shows the averages of the users steps,calories, and sleep.

```r
daily_average <- merged_activity_sleep %>%
  group_by(id) %>%
  summarise(average_daily_steps = mean(totalsteps), average_daily_calories = mean(calories), average_da

average_steps <- daily_activity %>%
  group_by(id)
```

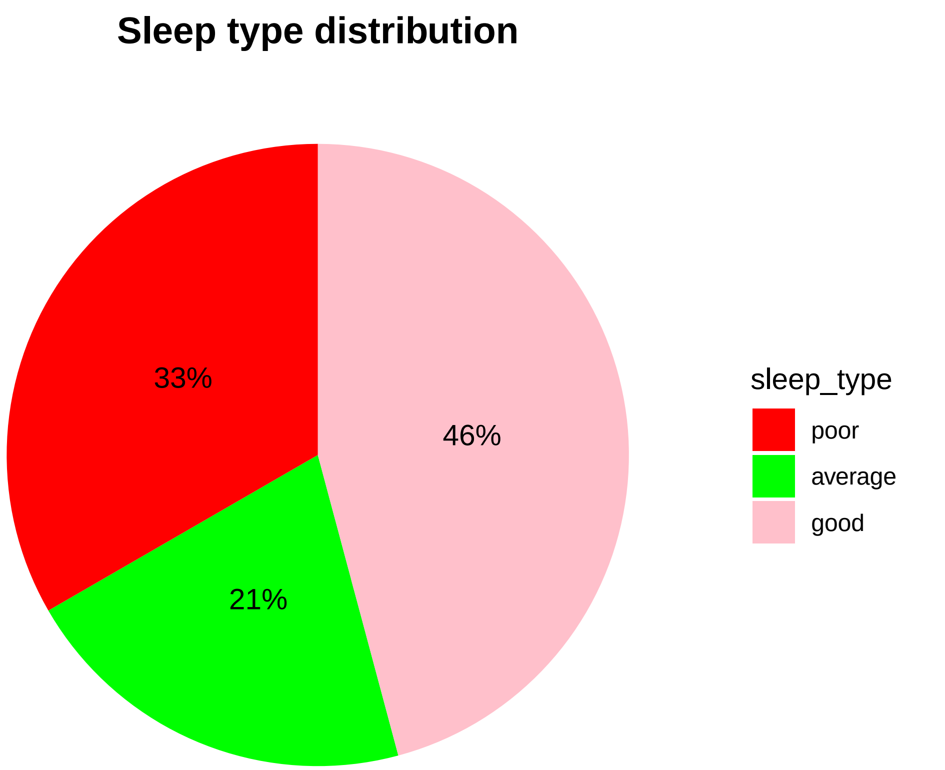
Visualizations

Understanding the User Data Base This pie chart then shows the different types of users based on activity:



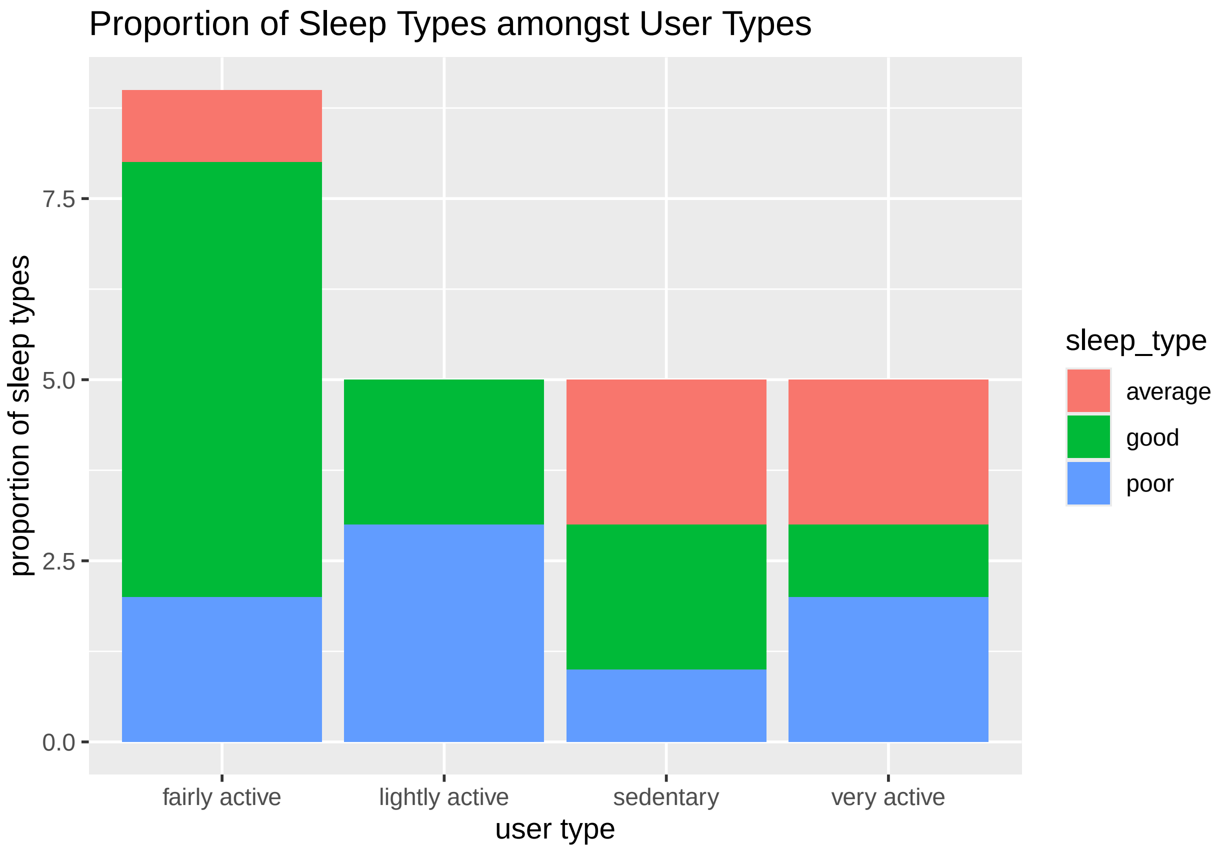
Here we can see that a majority of users are actually within the fairly active category, with roughly equal parts of the other types of activity. **Bellabeat should aim to market towards people who are already quite active, as well as market towards people inclining towards improving their activity levels.**

The following pie chart differentiates users based on their sleep type.



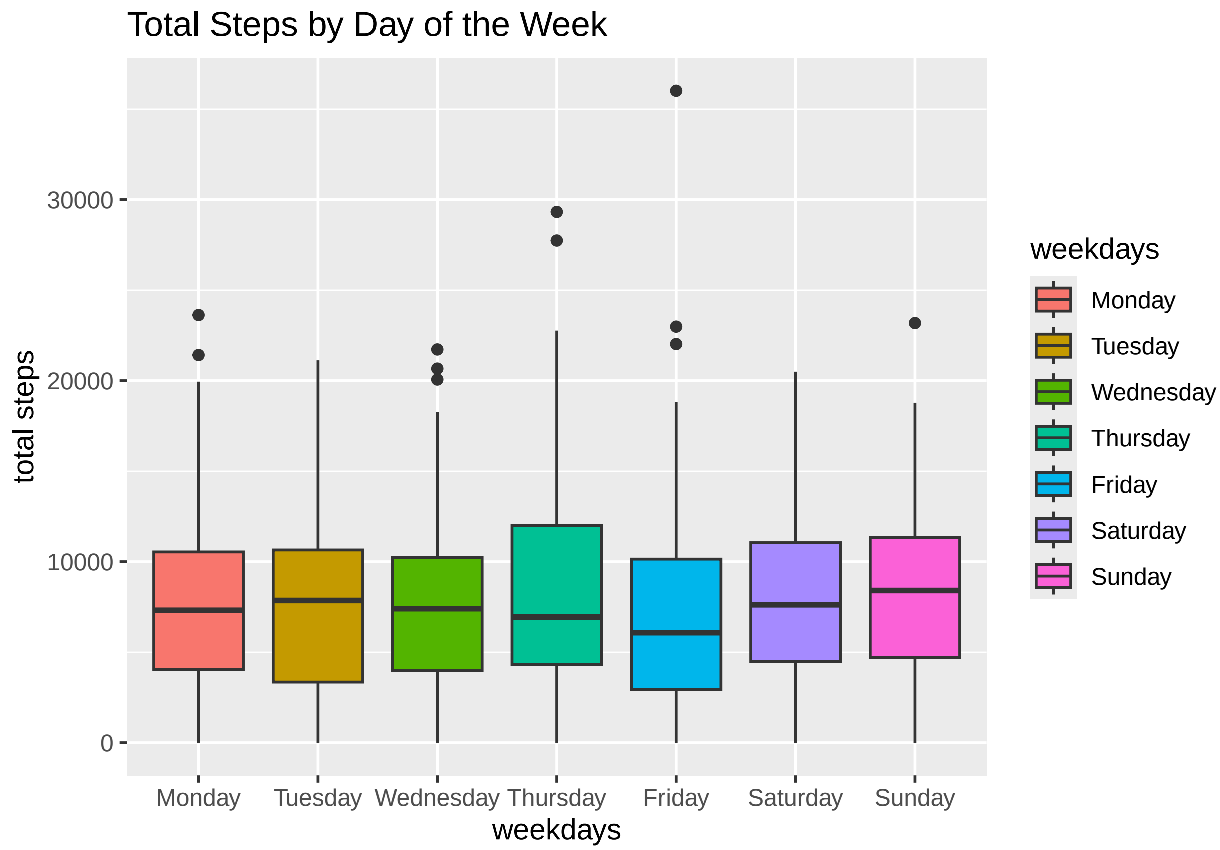
From here we see that the average user gets good sleep, although a large portion of users get poor sleep. **Bellabeat can encourage better sleep health for its users, perhaps by notifying users when it is time to 'wind down' or by giving users sleep tips to build better sleep habits.**

The following bar graph shows the proportions of sleep types in the different user types.



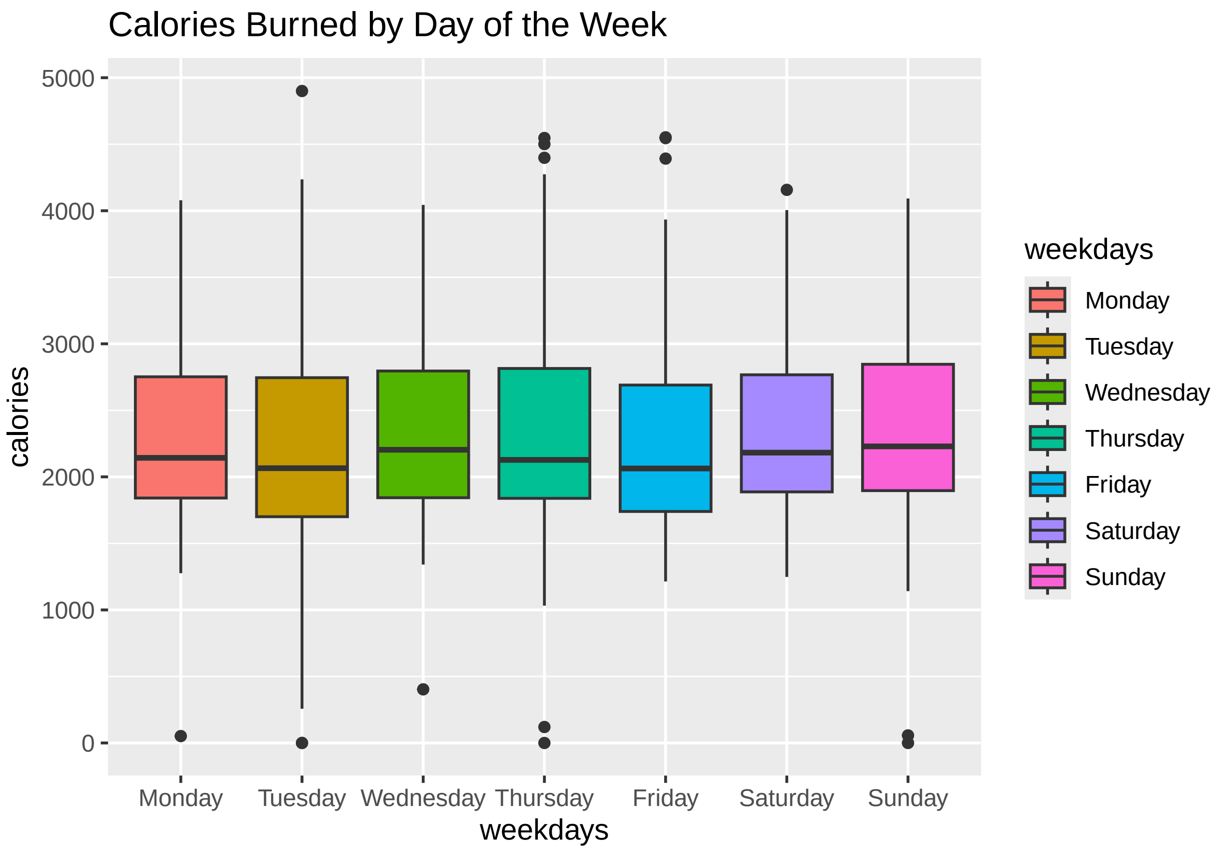
Here we see that fairly active users make up a majority of people with 'good' sleep, whereas poor sleep is relatively consistent amongst activity levels. Perhaps ***encouraging users to increase their activity levels will help their sleep trends***, although other factors may be at play such as time of activity, caffeine intake, caloric intake, etc. ***Bellabeat should consider taking holistic data of their users to provide better individualized health tips.***

Activity during the week Total Steps by day of the week



This box plot shows us the amount of activity one might get during the week, and specifically tells us that people in this data set are consistent throughout the week but more sedentary on Sundays. **Bellabeat can encourage user activity on these days, and encourage rest days on Sundays to provide its users with moderation and time to recuperate.**

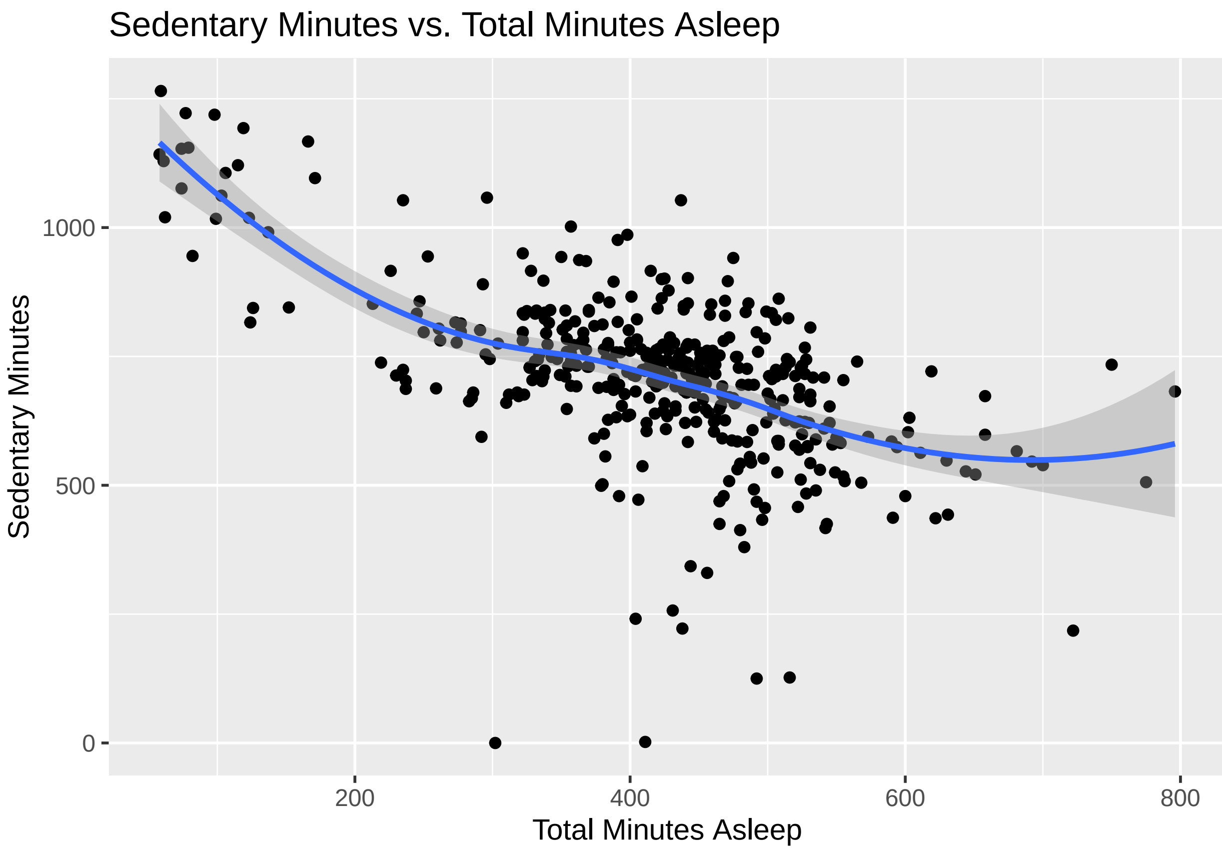
Calories burned by day of the week.



Again, this box plot reiterates that people are less inclined to workout on a Sunday, and more-so throughout the week.

Sleep Sleep vs. Sedentary Minutes

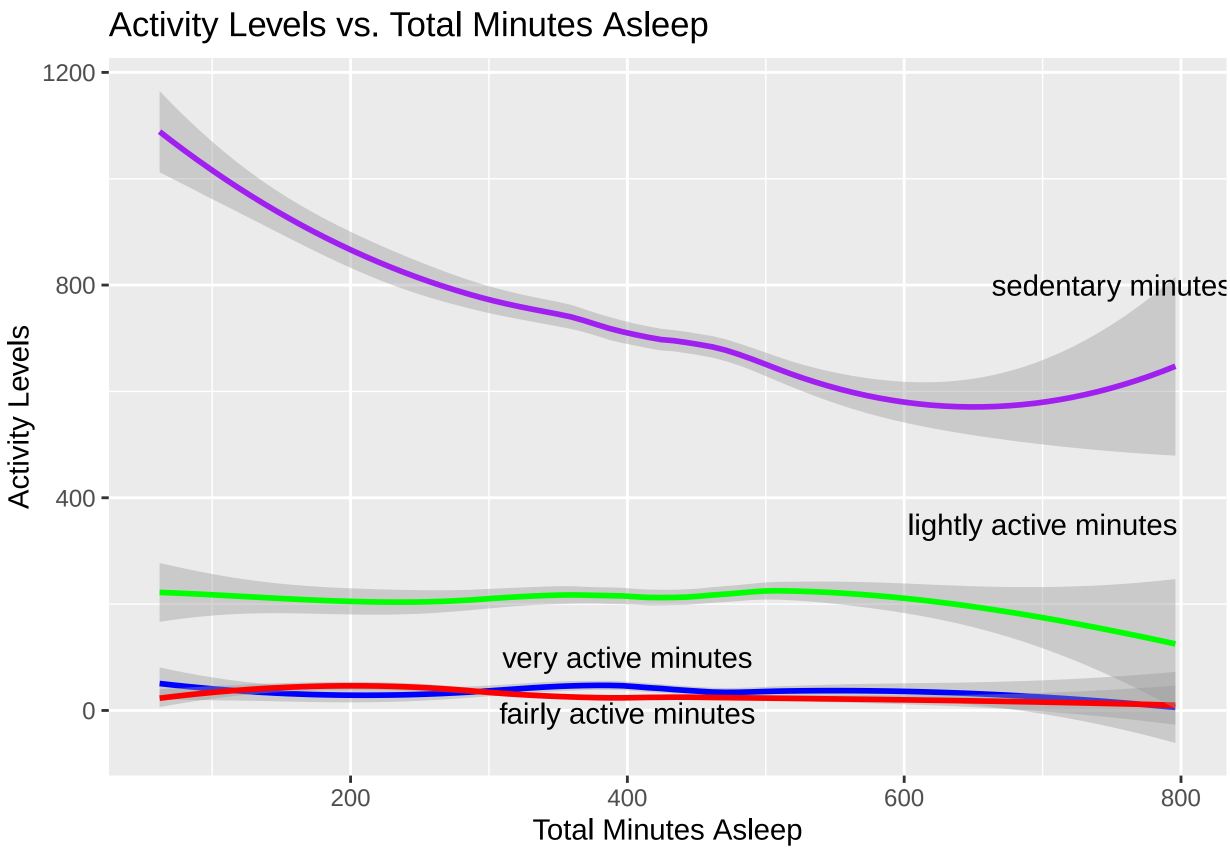
```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



The above graph shows us that more sedentary minutes may roughly correlate to less sleep. Taking our prior analysis from the user type distribution graphs, **Bellabeat should consider promoting and encouraging users with poor sleep to walk more or attain more daily physical activity.**

lets compare with other activity types:

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



Here we see that sedentary minutes are truly the biggest factor in total sleep.

Summarizing Recommendations

A majority of fitness device users are already quite active. Bellabeat can market to this demographic by creating devices which would benefit those who are fairly active, **possibly by making lightweight fitness devices that will individuals track their health data without getting in the way of their workouts.** Bellabeat could also attempt to open up their market to those who are not currently active, by **creating advertisements which promote 'getting started' or 'jumping back into' their health and fitness journey.**

Among this demographic, **Bellabeat might curate their device to make it on the users schedule.** By assessing the above data on weekly steps and caloric intake, **Bellabeat can encourage exercise throughout the week, and encourage rest or cheat days on Weekends (particularly Sunday).** By doing so, Bellabeat can market themselves as a self care app, **focused on creating a schedule that works for the individual.**

Bellabeat can also work towards improving user's sleep trends **by encouraging daily walking or activity through notifications**, letting the user know how many steps they have left for the day or encouraging a workout when their device notices them being sedentary for an extended period of time.

Bellabeat can also use their app to post health recommendations or fitness blogs, which can encourage more activity or better sleep habits.