Analysis of FitBit Fitness Track Data for Bellabeat

Yaxin Guan

2022/3/3

Data Source

FitBit Fitness Track Data

Load necessary packages

```
library(tidyverse)
```

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

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.0     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(data.table)
```

```
##
## Attaching package: 'data.table'

## The following objects are masked from 'package:lubridate':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday, week,
##     yday, year
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last

## The following object is masked from 'package:purrr':
##
##     transpose
```

Loading CSV files

The data are from April 12th, 2016 to May 12th, 2016.

```
# Filepath <- "User/Capstone-Project/"
daily_activity <- read_csv(paste0(Filepath, "dailyActivity_merged.csv"))
```

```
## Rows: 940 Columns: 15

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): ActivityDate
## dbl (14): Id, TotalSteps, TotalDistance, TrackerDistance, LoggedActivitiesDi...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
sleep_day <- read_csv(paste0(Filepath, "sleepDay_merged.csv"))
```

```
## Rows: 413 Columns: 5

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): SleepDay
## dbl (4): Id, TotalSleepRecords, TotalMinutesAsleep, TotalTimeInBed

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
weight_log <- read_csv(paste0(Filepath, "weightLogInfo_merged.csv"))
```

```
## Rows: 67 Columns: 8

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Date
## dbl (6): Id, WeightKg, WeightPounds, Fat, BMI, LogId
## lgl (1): IsManualReport
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
hourly_intensities <- read_csv(paste0(Filepath, "hourlyIntensities_merged.csv"))
```

```
## Rows: 22099 Columns: 4
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): ActivityHour
## dbl (3): Id, TotalIntensity, AverageIntensity
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
hourly_calories <- read_csv(paste0(Filepath, "hourlyCalories_merged.csv"))
```

```
## Rows: 22099 Columns: 3
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): ActivityHour
## dbl (2): Id, Calories
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
hourly_steps <- read_csv(paste0(Filepath, "hourlySteps_merged.csv"))
```

```
## Rows: 22099 Columns: 3
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): ActivityHour
## dbl (2): Id, StepTotal
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Explore data

```
head(daily_activity)
```

```
## # A tibble: 6 x 15
##        Id ActivityDate TotalSteps TotalDistance TrackerDistance LoggedActivitie~
##     <dbl> <chr>              <dbl>         <dbl>           <dbl>            <dbl>
## 1  1.50e9 4/12/2016         13162          8.5             8.5                 0
## 2  1.50e9 4/13/2016         10735          6.97            6.97                0
## 3  1.50e9 4/14/2016         10460          6.74            6.74                0
## 4  1.50e9 4/15/2016          9762          6.28            6.28                0
## 5  1.50e9 4/16/2016         12669          8.16            8.16                0
## 6  1.50e9 4/17/2016          9705          6.48            6.48                0
## # ... with 9 more variables: VeryActiveDistance <dbl>,
## #   ModeratelyActiveDistance <dbl>, LightActiveDistance <dbl>,
## #   SedentaryActiveDistance <dbl>, VeryActiveMinutes <dbl>,
## #   FairlyActiveMinutes <dbl>, LightlyActiveMinutes <dbl>,
## #   SedentaryMinutes <dbl>, Calories <dbl>
```

```
colnames(daily_activity)
```

```
##  [1] "Id"                       "ActivityDate"
##  [3] "TotalSteps"               "TotalDistance"
##  [5] "TrackerDistance"          "LoggedActivitiesDistance"
##  [7] "VeryActiveDistance"       "ModeratelyActiveDistance"
##  [9] "LightActiveDistance"      "SedentaryActiveDistance"
## [11] "VeryActiveMinutes"        "FairlyActiveMinutes"
## [13] "LightlyActiveMinutes"     "SedentaryMinutes"
## [15] "Calories"
```

```
head(sleep_day)
```

```
## # A tibble: 6 x 5
##           Id SleepDay          TotalSleepRecor~ TotalMinutesAsle~ TotalTimeInBed
##        <dbl> <chr>                        <dbl>             <dbl>          <dbl>
## 1 1503960366 4/12/2016 12:00:~                1               327            346
## 2 1503960366 4/13/2016 12:00:~                2               384            407
## 3 1503960366 4/15/2016 12:00:~                1               412            442
## 4 1503960366 4/16/2016 12:00:~                2               340            367
## 5 1503960366 4/17/2016 12:00:~                1               700            712
## 6 1503960366 4/19/2016 12:00:~                1               304            320
```

```
colnames(sleep_day)
```

```
## [1] "Id"                 "SleepDay"           "TotalSleepRecords"
## [4] "TotalMinutesAsleep" "TotalTimeInBed"
```

```
head(weight_log)
```

```
## # A tibble: 6 x 8
##           Id Date      WeightKg WeightPounds   Fat   BMI IsManualReport    LogId
##        <dbl> <chr>        <dbl>        <dbl> <dbl> <dbl> <lgl>             <dbl>
## 1 1503960366 5/2/2016~     52.6         116.    22  22.6 TRUE            1.46e12
## 2 1503960366 5/3/2016~     52.6         116.    NA  22.6 TRUE            1.46e12
## 3 1927972279 4/13/201~    134.          294.    NA  47.5 FALSE           1.46e12
```

```
## 4 2873212765 4/21/201~     56.7         125.    NA  21.5 TRUE            1.46e12
## 5 2873212765 5/12/201~     57.3         126.    NA  21.7 TRUE            1.46e12
## 6 4319703577 4/17/201~     72.4         160.    25  27.5 TRUE            1.46e12
```

```
colnames(weight_log)
```

```
## [1] "Id"             "Date"           "WeightKg"       "WeightPounds"
## [5] "Fat"            "BMI"            "IsManualReport" "LogId"
```

```
head(hourly_intensities)
```

```
## # A tibble: 6 x 4
##           Id ActivityHour          TotalIntensity AverageIntensity
##        <dbl> <chr>                          <dbl>            <dbl>
## 1 1503960366 4/12/2016 12:00:00 AM             20            0.333
## 2 1503960366 4/12/2016 1:00:00 AM               8            0.133
## 3 1503960366 4/12/2016 2:00:00 AM               7            0.117
## 4 1503960366 4/12/2016 3:00:00 AM               0            0
## 5 1503960366 4/12/2016 4:00:00 AM               0            0
## 6 1503960366 4/12/2016 5:00:00 AM               0            0
```

```
colnames(hourly_intensities)
```

```
## [1] "Id"               "ActivityHour"     "TotalIntensity"   "AverageIntensity"
```

```
head(hourly_calories)
```

```
## # A tibble: 6 x 3
##           Id ActivityHour          Calories
##        <dbl> <chr>                    <dbl>
## 1 1503960366 4/12/2016 12:00:00 AM       81
## 2 1503960366 4/12/2016 1:00:00 AM        61
## 3 1503960366 4/12/2016 2:00:00 AM        59
## 4 1503960366 4/12/2016 3:00:00 AM        47
## 5 1503960366 4/12/2016 4:00:00 AM        48
## 6 1503960366 4/12/2016 5:00:00 AM        48
```

```
colnames(hourly_calories)
```

```
## [1] "Id"           "ActivityHour" "Calories"
```

```
head(hourly_steps)
```

```
## # A tibble: 6 x 3
##           Id ActivityHour          StepTotal
##        <dbl> <chr>                     <dbl>
## 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
```

```
colnames(hourly_steps)
```

```
## [1] "Id"           "ActivityHour" "StepTotal"
```

Energy expenditure formula is from: **HSS** Calories/minutes = 0.0175 x MET (of activity) x body weight (in kg) Calories will burn even when sitting quietly, so calories burn when the hourly step is 0.

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

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

```
n_distinct(sleep_day$Id)
```

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

```
n_distinct(weight_log$Id)
```

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

It looks like there may be more participants in the daily activity dataset than the sleep dataset.

Number of observations

```
nrow(daily_activity)
```

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

```
nrow(sleep_day)
```

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

```
nrow(weight_log)
```

```
## [1] 67
```

```
nrow(hourly_calories)
```

```
## [1] 22099
```

```
nrow(hourly_intensities)
```

```
## [1] 22099
```

```
nrow(hourly_steps)
```

```
## [1] 22099
```

Summaries

For the daily activity data frame:

```
daily_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 frame:

```
sleep_day %>%
  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
```

For the weight log data frame:

```
weight_log %>%
  select(WeightKg,WeightPounds, BMI) %>%
  summary()
```

```
##     WeightKg       WeightPounds        BMI
##  Min.   : 52.60   Min.   :116.0   Min.   :21.45
##  1st Qu.: 61.40   1st Qu.:135.4   1st Qu.:23.96
##  Median : 62.50   Median :137.8   Median :24.39
##  Mean   : 72.04   Mean   :158.8   Mean   :25.19
##  3rd Qu.: 85.05   3rd Qu.:187.5   3rd Qu.:25.56
##  Max.   :133.50   Max.   :294.3   Max.   :47.54
```

For the hourly intensities data frame:

```
hourly_intensities %>%
  select(TotalIntensity, 'AverageIntensity(in sec)' = AverageIntensity) %>%
  summary()
```

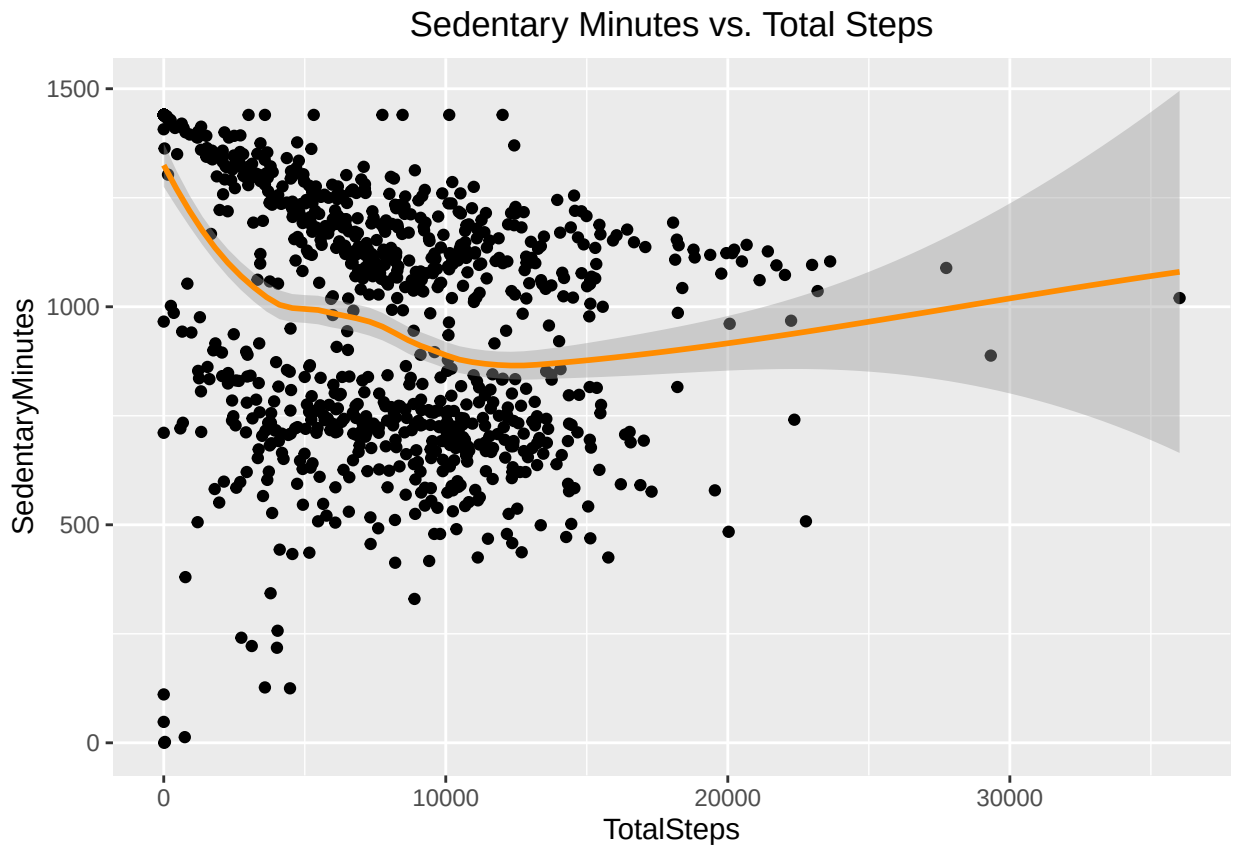
```
##  TotalIntensity   AverageIntensity(in sec)
##  Min.   :  0.00   Min.   :0.0000
##  1st Qu.:  0.00   1st Qu.:0.0000
##  Median :  3.00   Median :0.0500
##  Mean   : 12.04   Mean   :0.2006
##  3rd Qu.: 16.00   3rd Qu.:0.2667
##  Max.   :180.00   Max.   :3.0000
```

Data Visualization

```
ggplot(data=daily_activity, aes(x=TotalSteps, y=SedentaryMinutes)) + geom_point() + geom_smooth(color =
  theme(plot.title = element_text(hjust = 0.5))
```

Sedentary Minutes vs. Total Steps

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




```
# ggsave("sedentaryminutes_totalsteps.jpg")
```

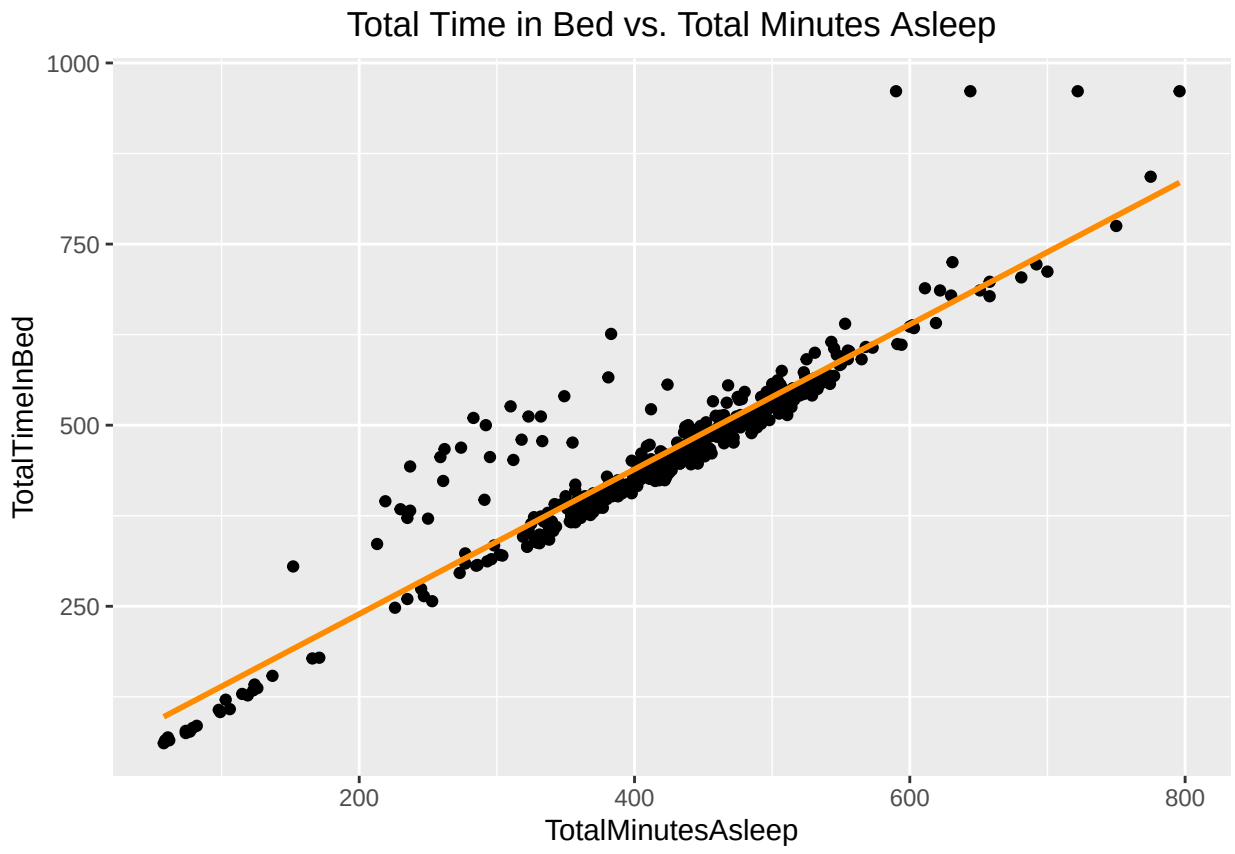
```
cor.test(daily_activity$TotalSteps, daily_activity$SedentaryMinutes,
         alternative = "less")
```

```
##
##  Pearson's product-moment correlation
##
## data:  daily_activity$TotalSteps and daily_activity$SedentaryMinutes
## t = -10.615, df = 938, p-value < 2.2e-16
## alternative hypothesis: true correlation is less than 0
## 95 percent confidence interval:
##  -1.0000000 -0.2786998
## sample estimates:
##        cor
## -0.3274835
```

Although the trend is not obvious in the graph, the correlation is negative.

```
ggplot(data=sleep_day, aes(x=TotalMinutesAsleep, y=TotalTimeInBed)) + geom_point() +
  geom_smooth(aes(group = 1),method = "lm", formula = y ~ x,se = FALSE, color = "darkorange") + labs(ti
  theme(plot.title = element_text(hjust = 0.5))
```

Total Time in Bed vs. Total Minutes Asleep




```
# ggsave("Totaltimeinbed_totalasleep.jpg")
```

```
cor.test(sleep_day$TotalMinutesAsleep, sleep_day$TotalTimeInBed,
         alternative = "greater")
```

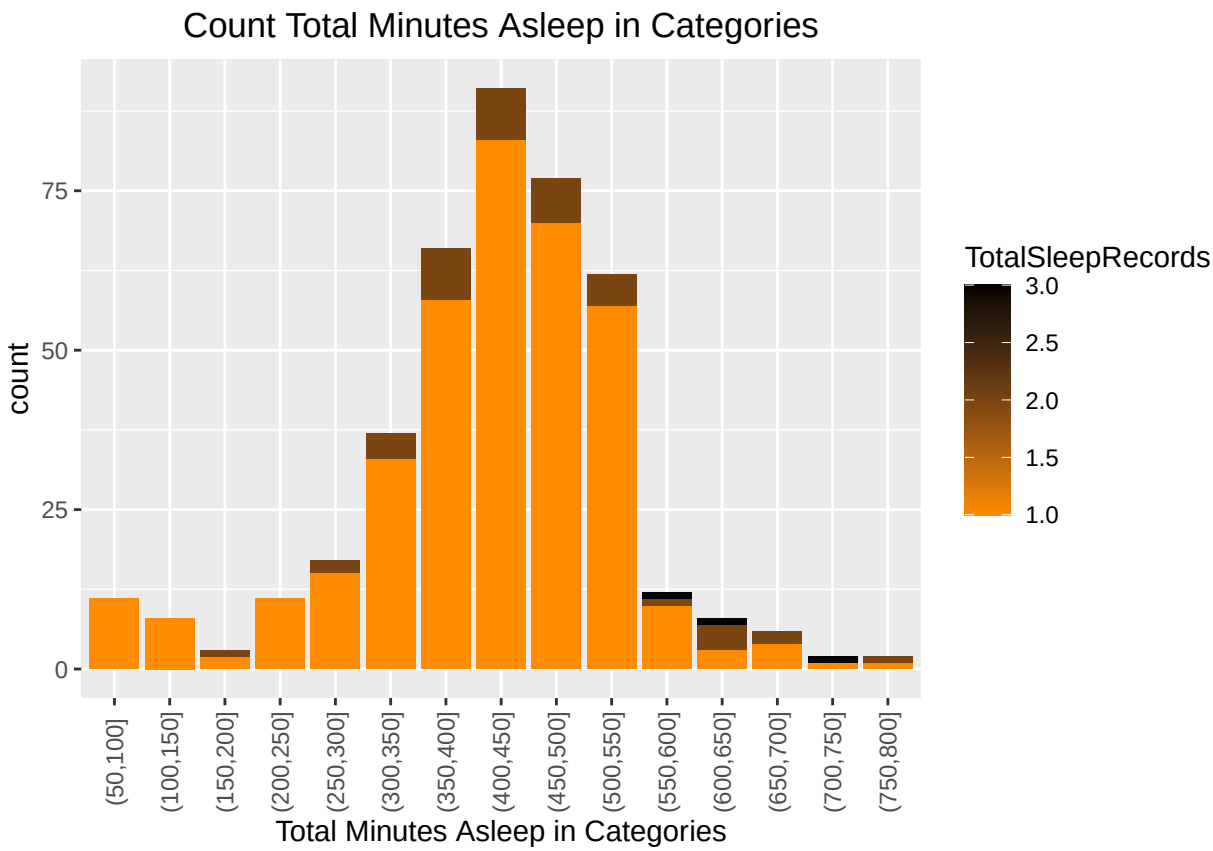
```
##
##  Pearson's product-moment correlation
##
## data:  sleep_day$TotalMinutesAsleep and sleep_day$TotalTimeInBed
## t = 51.483, df = 411, p-value < 2.2e-16
## alternative hypothesis: true correlation is greater than 0
## 95 percent confidence interval:
##  0.9186882 1.0000000
## sample estimates:
##       cor
## 0.9304575
```

Positive and strong correlation (close to 1) as expected.

```
sleep_day$asleep_categories <- cut(sleep_day$TotalMinutesAsleep, seq(from = 0, to = 800, by = 50))
sleep_day %>%
  group_by(asleep_categories,TotalSleepRecords) %>%
  summarise(count = n()) %>%
  ggplot(aes(x = asleep_categories, y = count, fill = TotalSleepRecords)) +
  geom_bar(position= "stack",stat="identity") +
  scale_fill_gradient(low = "darkorange", high = "black") +
  labs(x = "Total Minutes Asleep in Categories", title = "Count Total Minutes Asleep in Categories")+
  theme(plot.title = element_text(hjust = 0.5), axis.text.x = element_text(vjust = 0.5, angle = 90))
```

Time Asleep & Total Sleep Records

```
## `summarise()` has grouped output by 'asleep_categories'. You can override using the `.groups` argume
```




```
#ggsave("Count Total Minutes Asleep in Categories.jpg")
```

```
#combined_data <- merge(sleep_day, daily_activity, by = "Id", allow.cartesian=TRUE)
combined_data <- right_join(sleep_day, daily_activity, by = "Id")
```

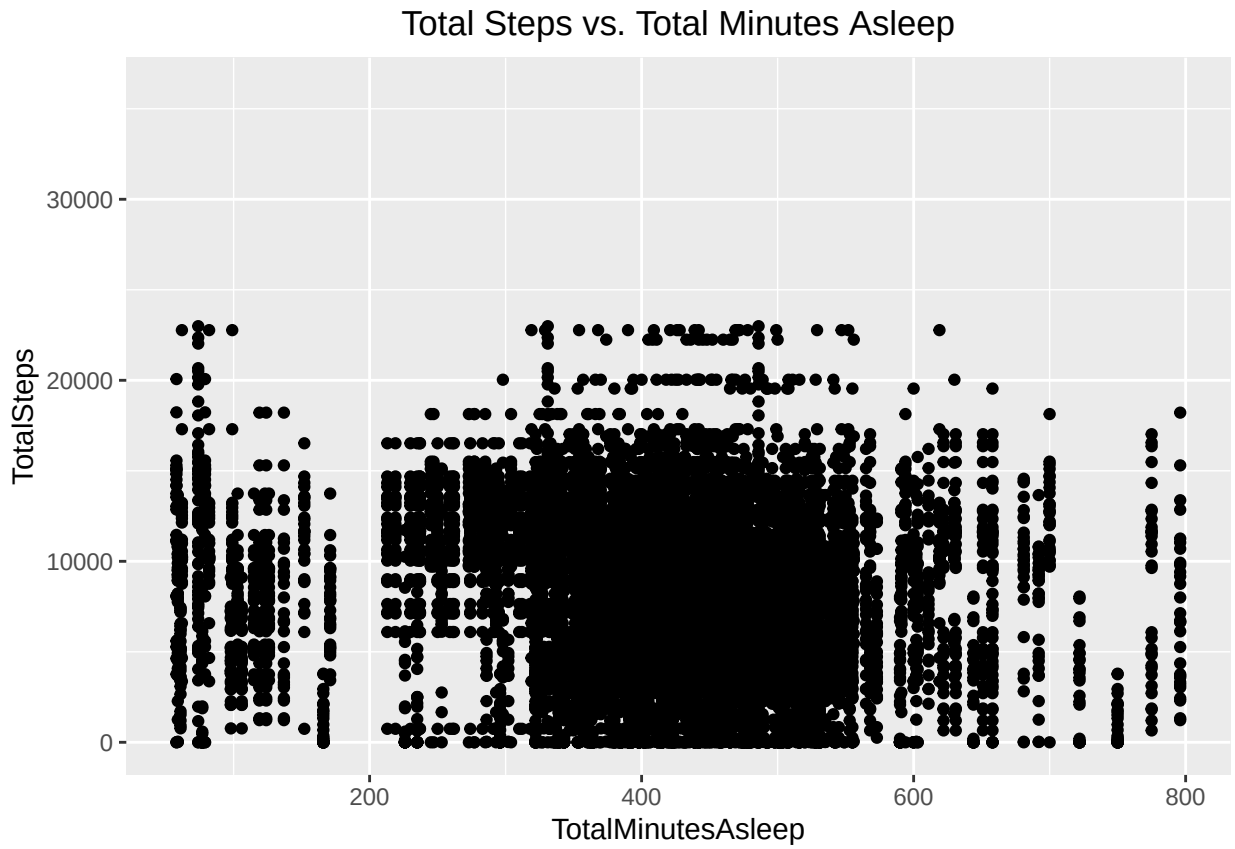
```
n_distinct(combined_data$Id)
```

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

```
ggplot(data = combined_data, aes(x = TotalMinutesAsleep, y = TotalSteps)) +
  geom_point() + labs(title = "Total Steps vs. Total Minutes Asleep") +
  theme(plot.title = element_text(hjust = 0.5))
```

Total Steps vs. Total Minutes Asleep

```
## Warning: Removed 227 rows containing missing values (geom_point).
```




```
cor.test(combined_data$TotalMinutesAsleep, combined_data$TotalSteps)
```

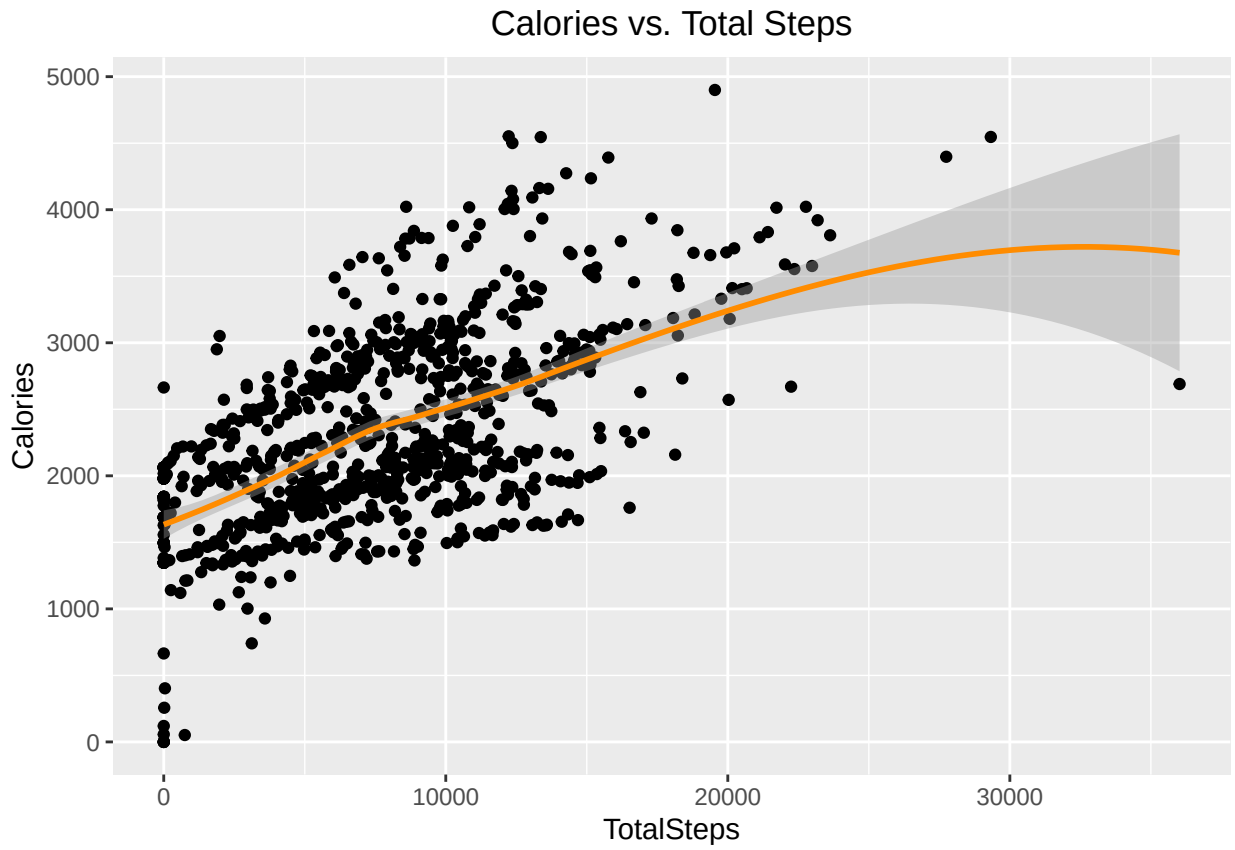
```
##
##  Pearson's product-moment correlation
##
## data:  combined_data$TotalMinutesAsleep and combined_data$TotalSteps
## t = -11.044, df = 12439, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.11591302 -0.08110962
## sample estimates:
##         cor
## -0.09854146
```

The correlation is negative, but correlation does not mean the causation. More data and investigation (increase of sample size, controlled study, etc.) are needed to show that total steps can keep people feel energetic and decrease the sleep times.

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

Calories vs. Total Steps

```
## `geom_smooth()` using formula 'y ~ x'
```




```
# ggsave("calories_totalsteps.jpg")
```

```
cor.test(daily_activity$TotalSteps, daily_activity$Calories,
         alternative = "greater")
```

```
##
##  Pearson's product-moment correlation
##
## data:  daily_activity$TotalSteps and daily_activity$Calories
## t = 22.472, df = 938, p-value < 2.2e-16
## alternative hypothesis: true correlation is greater than 0
## 95 percent confidence interval:
##  0.5555268 1.0000000
## sample estimates:
##       cor
## 0.5915681
```

The correlation is positive, which matches the trend in graph. The more time a person spends on walking, the more calories one will burn.

```
hourly_intensities$Date <- format(as.Date(hourly_intensities$ActivityHour,
                                  format = "%m/%d/%Y"), format = "%m/%d/%Y") # Date

hourly_intensities$ActivityHour <- mdy_hms(hourly_intensities$ActivityHour,
                                           tz = Sys.timezone())

hourly_intensities$Time <- format(hourly_intensities$ActivityHour,
                                  format = "%H:%M:%S")

hourly_intensities$day_of_week <- format(as.Date(hourly_intensities$ActivityHour), "%A")
```

```
hourly_intensities$day_of_week <- ordered(hourly_intensities$day_of_week, levels=c("Sunday", "Monday",
```

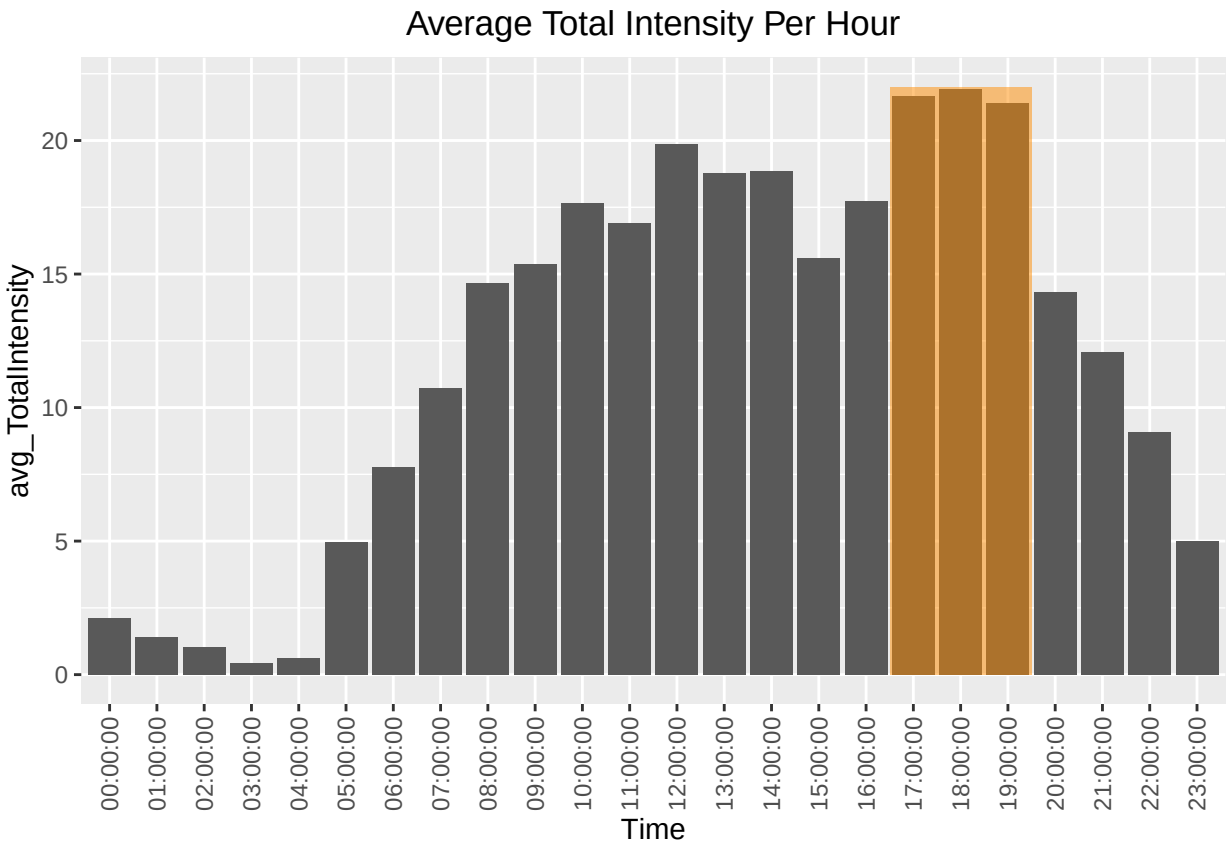
```
extract_data <- hourly_intensities[, c(3,6)]
plot_data <- extract_data %>%
  group_by(Time) %>%
  summarise(avg_TotalIntensity = mean(TotalIntensity))

extract_data2 <- hourly_intensities[, c(3,7)]
plot_data2 <- extract_data2 %>%
  group_by(day_of_week) %>%
  summarise(avg_TotalIntensity = mean(TotalIntensity))
```

Time of Intensities

```
ggplot(data = plot_data, aes(x = Time, y = avg_TotalIntensity)) +
  geom_bar(stat = "identity") + labs(title = "Average Total Intensity Per Hour") +
  annotate("rect", xmin = 17.5, ymin = 0, xmax = 20.5, ymax = 22,
           fill = "darkorange", alpha = 0.5)+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5),
        plot.title = element_text(hjust = 0.5))
```

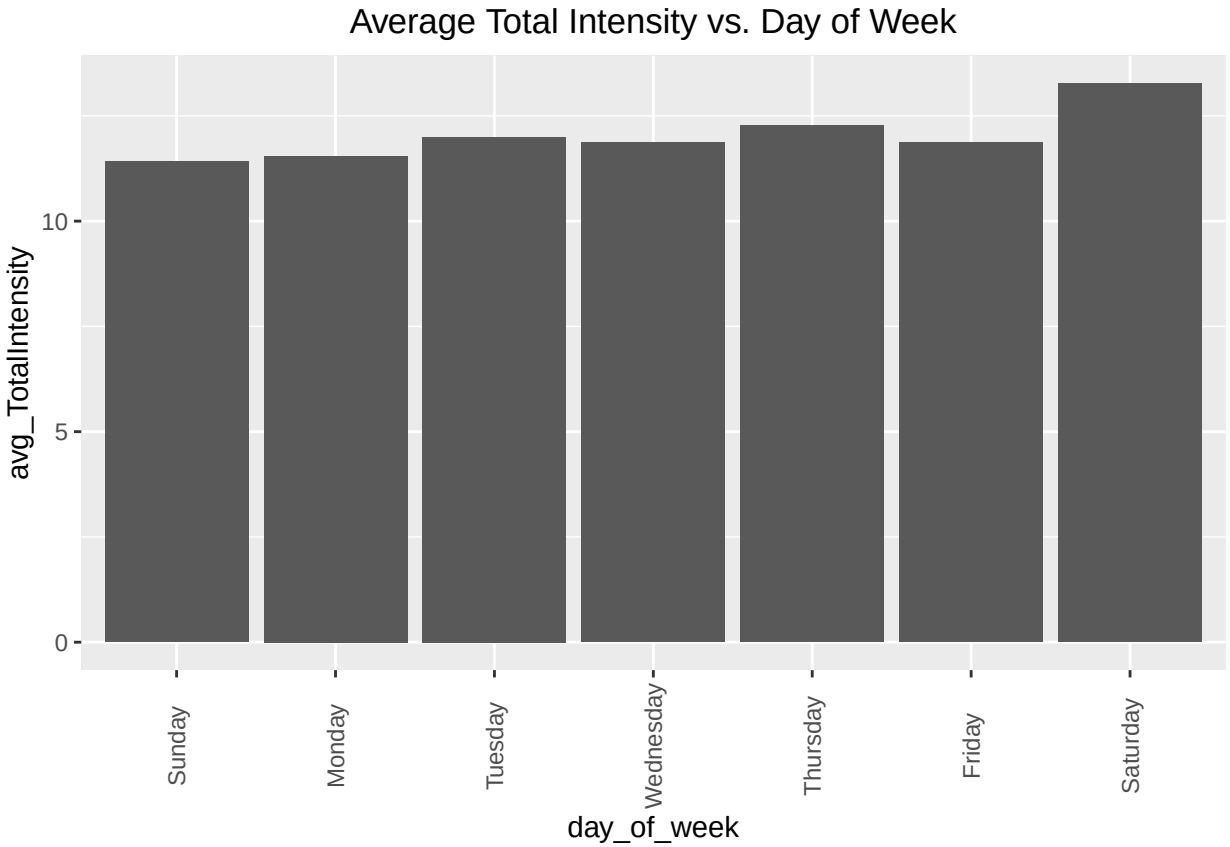
Average of Total Intensity vs. Time




```
#ggsave("time_total_intensity.jpg")
```

```
ggplot(data = plot_data2, aes(x = day_of_week, y = avg_TotalIntensity)) +
  geom_bar(stat = "identity") + labs(title = "Average Total Intensity vs. Day of Week") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5),
        plot.title = element_text(hjust = 0.5))
```

Average of Total Intensity vs. Days of Week



Conclusion

The total steps, calories burned, total minutes asleep, total time in bed, hourly intensity, and sedentary minutes are six key elements that people recorded with their smart devices and used in this analysis. The correlation between total steps and calories burned is positive (0.5915681). The correlation between total minutes asleep and total time in bed burned is strong and positive (0.9304575). However, the graph of count total minutes asleep shows that there are many sleep records of the participants are under 7 hours (420 minutes) each day. This needs to be improved because adults need 7 or more hours of sleep per night to maintain wellness based on CDC.

Also, the data shows that the mean of daily total step is 7,638 and the third quartile is 10,727. Based on Lifestyle Coach Facilitation Guide: Post-Core of CDC, the goal of daily total steps to maintain wellness is 10,000. Only about 25% of total steps in this data reaches the goal.

Moreover, the hourly intensity shows that the participants (8 unique Id in this case) have more intensities at 5:00 PM to 7:00 PM. It is reasonable because most people are off work at that time, yet increasing the sample size will help defining the trend better. The data from the dailyActivity_merged.csv file shows the mean sedentary minutes of the participant is 991.2 (16.52 hr).

Based on the trends above, the Leaf (classic wellness tracker) of Bellabeat is good products for keeping track of activity and sleep. Bellabeat Leaf tracker can keep track of activity like walking and calories burned. The Leaf tracker connects with the app has sleep goal to maintain a good sleep habit. The Leaf tracker is able to keep track of light sleep and deep sleep. It also has goal for steps and active hours for exercise to achieve wellness goal. Leaf has the inactivity alert feature, which it will remind the user to move more or less often, by vibrating consecutively when the user has been inactive. The user needs to do a certain number of steps to not have the Leaf tracker reminds. It will be better if Bellabeat can extend on this feature, such as reminding the user to stand up when ones sedentary time is long. Since there is job position requires the

person to stay in the same location most of time and cannot satisfy the requirement of steps, it will help the user to burn more calories even in the work environment when the user stands up instead of sitting.

These trends could help influence Bellabeat marketing strategy by allowing Bellbeat to advertise its products are capable to do the same as other fitness trackers and more. Bellabeat Leaf tracker connects to the Bellabeat app to track activity, sleep, and stress. Based on Office on Women's Health (OASH), stress and hormonal changes can cause insomnia for women. The Leaf keeps track of stress with Bellabeat app will be helpful to evaluate stress level. In addition, Bellabeat app offers meditation and period tracking.