Bellabeat

Introduction

Bellabeat is a high-tech manufacturer of health-focused products for women. Collecting data on activity, sleep, stress, and reproductive health has allowed Bellabeat to empower women with knowledge about their own health and habits. Although Bellabeat is a successful small company, 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.

1. Ask

Business Task:

To identify potential opportunities for growth and provide recommendations for the Bellabeat marketing strategy improvement based on trends in smart device usage.

Key Stakeholders:

- Urška Sršen: Bellabeat's cofounder and Chief Creative Officer
- Sando Mur: Mathematician and Bellabeat's co-founder

Questions to explore for the analysis:

- What are some trends in smart device usage?
- How could these trends apply to Bellabeat customers?
- How could these trends help influence Bellabeat marketing strategy?

2. Prepare

This Kaggle data set contains personal fitness tracker from thirty fitbit users. Thirty eligible Fitbit users 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.

Loading Packages

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.4.4     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
```

```
## ── 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(lubridate)
library(dplyr)
library(ggplot2)
library(tidyr)
library(janitor)
```

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

3. Process

Importing the Datasets

```
# Read the dataframes
activity <- read_csv("C:/Users/karis/Downloads/input/Fitabase Data 4.12.16-
5.12.16/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.
```

```
calories <- read_csv("C:/Users/karis/Downloads/input/Fitabase Data 4.12.16-
5.12.16/dailyCalories_merged.csv")
```

```
## Rows: 940 Columns: 3
## ── Column specification
──────────────────────────────────────────────────────
## Delimiter: ","
## chr (1): ActivityDay
## 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.

intensities <- read_csv("C:/Users/karis/Downloads/input/Fitabase Data
4.12.16-5.12.16/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.

sleep <- read_csv("C:/Users/karis/Downloads/input/Fitabase Data 4.12.16-
5.12.16/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 <- read_csv("C:/Users/karis/Downloads/input/Fitabase Data 4.12.16-
5.12.16/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.
```

data

```
head(activity)

## # A tibble: 6 × 15
##           Id ActivityDate TotalSteps TotalDistance TrackerDistance
```

```
##        <dbl> <chr>             <dbl>         <dbl>           <dbl>
## 1 1503960366 4/12/2016         13162          8.5             8.5
## 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
## # ℹ 10 more variables: LoggedActivitiesDistance <dbl>,
## #   VeryActiveDistance <dbl>, ModeratelyActiveDistance <dbl>,
## #   LightActiveDistance <dbl>, SedentaryActiveDistance <dbl>,
## #   VeryActiveMinutes <dbl>, FairlyActiveMinutes <dbl>,
## #   LightlyActiveMinutes <dbl>, SedentaryMinutes <dbl>, Calories <dbl>
```

```
colnames(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(weight)
```

```
## # A tibble: 6 × 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 4/13/2016…    134.          294.    NA  47.5 FALSE
1.46e12
## 4 2873212765 4/21/2016…     56.7         125.    NA  21.5 TRUE
1.46e12
## 5 2873212765 5/12/2016…     57.3         126.    NA  21.7 TRUE
1.46e12
## 6 4319703577 4/17/2016…     72.4         160.    25  27.5 TRUE
1.46e12
```

```
colnames(weight)
```

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

Converting date time format

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

4. Analyze

Number of Participants in each category

```
n_distinct(activity$Id)

## [1] 33

n_distinct(calories$Id)

## [1] 33

n_distinct(intensities$Id)

## [1] 33

n_distinct(sleep$Id)

## [1] 24

n_distinct(weight$Id)

## [1] 8
```

To summarize the above data, there are 33 participants in the activity, calories, and intensities datasets, 24 in the sleep dataset, and only 8 in the weight dataset. The fact that there are only 8 participants in the weight dataset means that more data would be needed to make a strong reccomendation or conclusion.

checking for significant change in weight

```
weight%>%
 group_by(Id)%>%
  summarise(min(WeightKg),max(WeightKg))

## # A tibble: 8 × 3
##           Id `min(WeightKg)` `max(WeightKg)`
##        <dbl>           <dbl>           <dbl>
## 1 1503960366            52.6            52.6
```

```
## 2 1927972279           134.            134.
## 3 2873212765            56.7            57.3
## 4 4319703577            72.3            72.4
## 5 4558609924            69.1            70.3
## 6 5577150313            90.7            90.7
## 7 6962181067            61              62.5
## 8 8877689391            84              85.8
```

There is no significant changes in weight of 8 participants.

The summaries for the rest of the datasets:

```
# activity
activity %>%
  select(TotalSteps,
         TotalDistance,
         SedentaryMinutes, Calories) %>%
  summary()
```

```
##    TotalSteps    TotalDistance    SedentaryMinutes    Calories
##  Min.   :    0   Min.   : 0.000   Min.   :   0.0   Min.   :   0
##  1st Qu.: 3790   1st Qu.: 2.620   1st Qu.: 729.8   1st Qu.:1828
##  Median : 7406   Median : 5.245   Median :1057.5   Median :2134
##  Mean   : 7638   Mean   : 5.490   Mean   : 991.2   Mean   :2304
##  3rd Qu.:10727   3rd Qu.: 7.713   3rd Qu.:1229.5   3rd Qu.:2793
##  Max.   :36019   Max.   :28.030   Max.   :1440.0   Max.   :4900
```

```
# active minutes per category
activity %>%
  select(VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes) %>%
  summary()
```

```
##  VeryActiveMinutes FairlyActiveMinutes LightlyActiveMinutes
##  Min.   :  0.00    Min.   :  0.00      Min.   :  0.0
##  1st Qu.:  0.00    1st Qu.:  0.00      1st Qu.:127.0
##  Median :  4.00    Median :  6.00      Median :199.0
##  Mean   : 21.16    Mean   : 13.56      Mean   :192.8
##  3rd Qu.: 32.00    3rd Qu.: 19.00      3rd Qu.:264.0
##  Max.   :210.00    Max.   :143.00      Max.   :518.0
```

```
# calories
calories %>%
  select(Calories) %>%
  summary()
```

```
##     Calories
##  Min.   :   0
##  1st Qu.:1828
##  Median :2134
##  Mean   :2304
##  3rd Qu.:2793
##  Max.   :4900
```

```r
# sleep
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
```

```r
# weight
weight %>%
  select(WeightKg, BMI) %>%
  summary()
```

```
##     WeightKg           BMI
##  Min.   : 52.60   Min.   :21.45
##  1st Qu.: 61.40   1st Qu.:23.96
##  Median : 62.50   Median :24.39
##  Mean   : 72.04   Mean   :25.19
##  3rd Qu.: 85.05   3rd Qu.:25.56
##  Max.   :133.50   Max.   :47.54
```

Observations made from the above summaries:

- Sedetary minutes on average is 16.5 hours.
- The average number of steps per day is 7638. The CDC recommends people take 10,000 steps daily.
- The majority of the participants are lightly active.
- The average participant burns 97 calories per hour.
- On an average, participants sleep for 7 hours.

Merging Data

Merging two datasets Activity and Sleep on Columns Id and date.

```r
merged_data <- merge(sleep, activity, by = c('Id', 'date'))
head(merged_data)
```

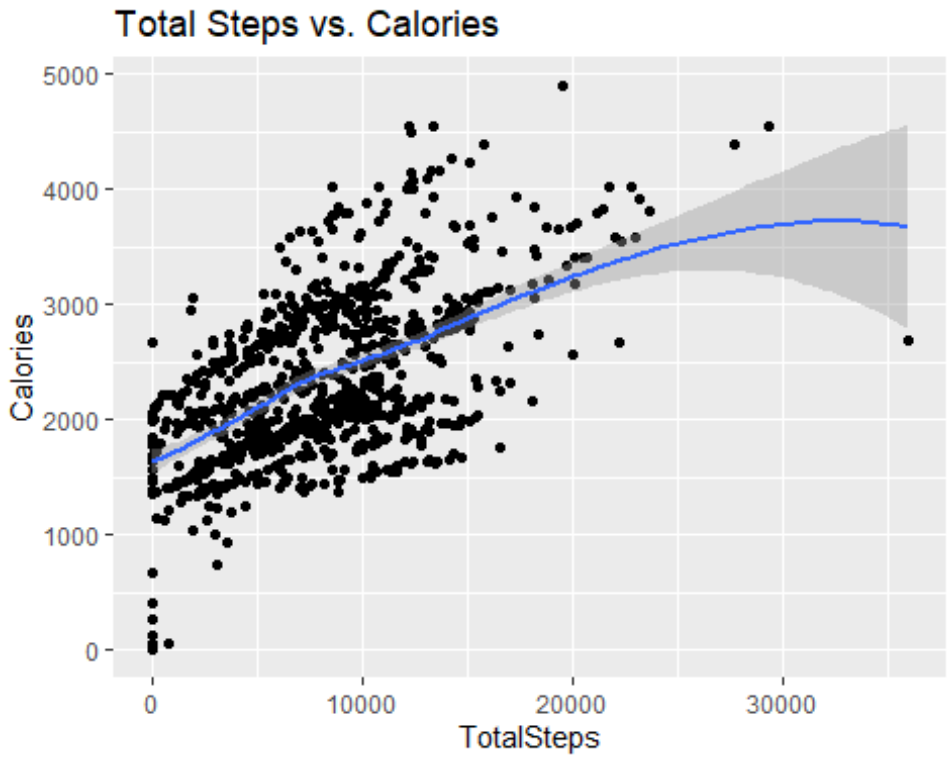
```
##           Id     date   SleepDay TotalSleepRecords TotalMinutesAsleep
## 1 1503960366 04/12/16 2016-04-12                 1                327
## 2 1503960366 04/13/16 2016-04-13                 2                384
## 3 1503960366 04/15/16 2016-04-15                 1                412
## 4 1503960366 04/16/16 2016-04-16                 2                340
## 5 1503960366 04/17/16 2016-04-17                 1                700
## 6 1503960366 04/19/16 2016-04-19                 1                304
##   TotalTimeInBed ActivityDate TotalSteps TotalDistance TrackerDistance
## 1            346   2016-04-12      13162          8.50            8.50
```

```
## 2            407   2016-04-13      10735          6.97             6.97
## 3            442   2016-04-15       9762          6.28             6.28
## 4            367   2016-04-16      12669          8.16             8.16
## 5            712   2016-04-17       9705          6.48             6.48
## 6            320   2016-04-19      15506          9.88             9.88
##   LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance
## 1                        0               1.88                     0.55
## 2                        0               1.57                     0.69
## 3                        0               2.14                     1.26
## 4                        0               2.71                     0.41
## 5                        0               3.19                     0.78
## 6                        0               3.53                     1.32
##   LightActiveDistance SedentaryActiveDistance VeryActiveMinutes
## 1                6.06                       0                25
## 2                4.71                       0                21
## 3                2.83                       0                29
## 4                5.04                       0                36
## 5                2.51                       0                38
## 6                5.03                       0                50
##   FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories
## 1                  13                  328              728     1985
## 2                  19                  217              776     1797
## 3                  34                  209              726     1745
## 4                  10                  221              773     1863
## 5                  20                  164              539     1728
## 6                  31                  264              775     2035
```

5. Share

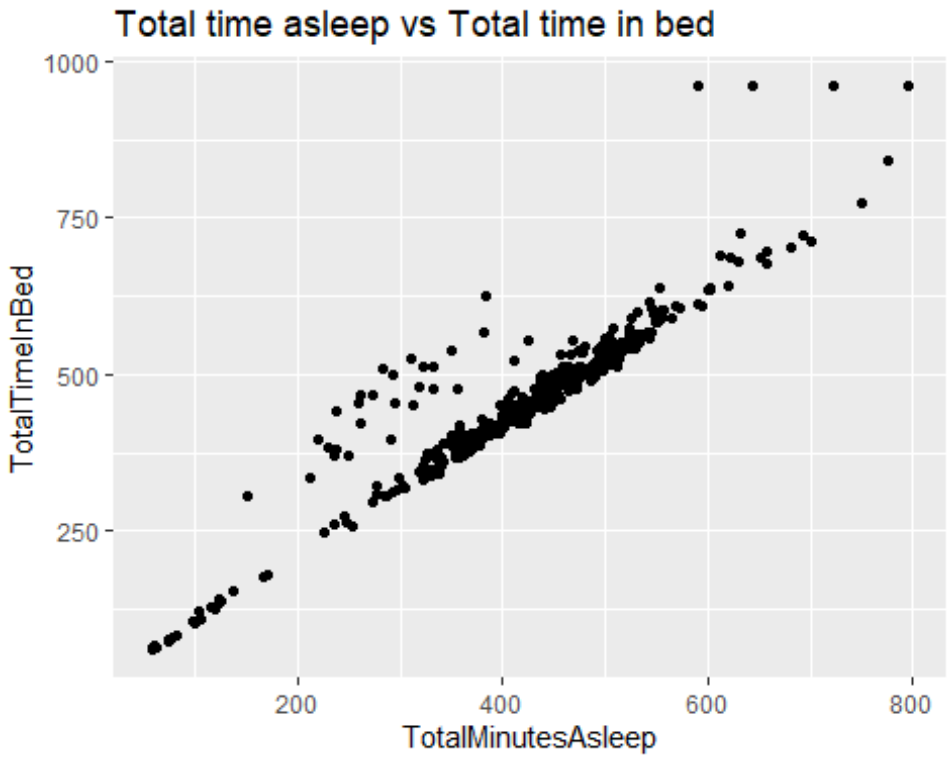
```
ggplot(data = activity, aes(x = TotalSteps, y = Calories)) + geom_point() +
geom_smooth() + labs(title = "Total Steps vs. Calories")

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



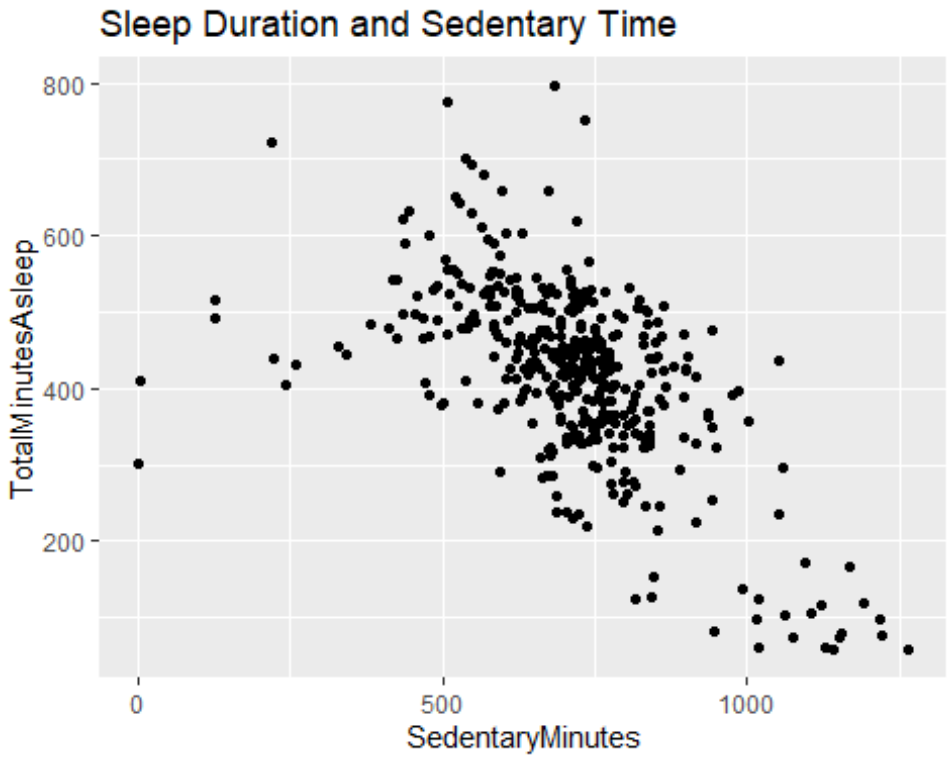
There is a correlation between total number of steps taken and calories burned. The more steps each participant takes, the more calories they burn.

```
ggplot(data = sleep, aes(x = TotalMinutesAsleep, y = TotalTimeInBed)) +
geom_point() + labs(title = "Total time asleep vs Total time in bed")
```



There is a positive correlation between total time asleep vs total time in bed. To improve sleep quality for its users, bellabeat should consider having a section where users can customize their sleep schedule to ensure consistency.

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




```
cor(merged_data$TotalMinutesAsleep,merged_data$SedentaryMinutes)

## [1] -0.599394
```

There is a negative correlation between SedentaryMinutes and TotalMinutesAsleep. This means that the less active a participant is, the less sleep they tend to get.

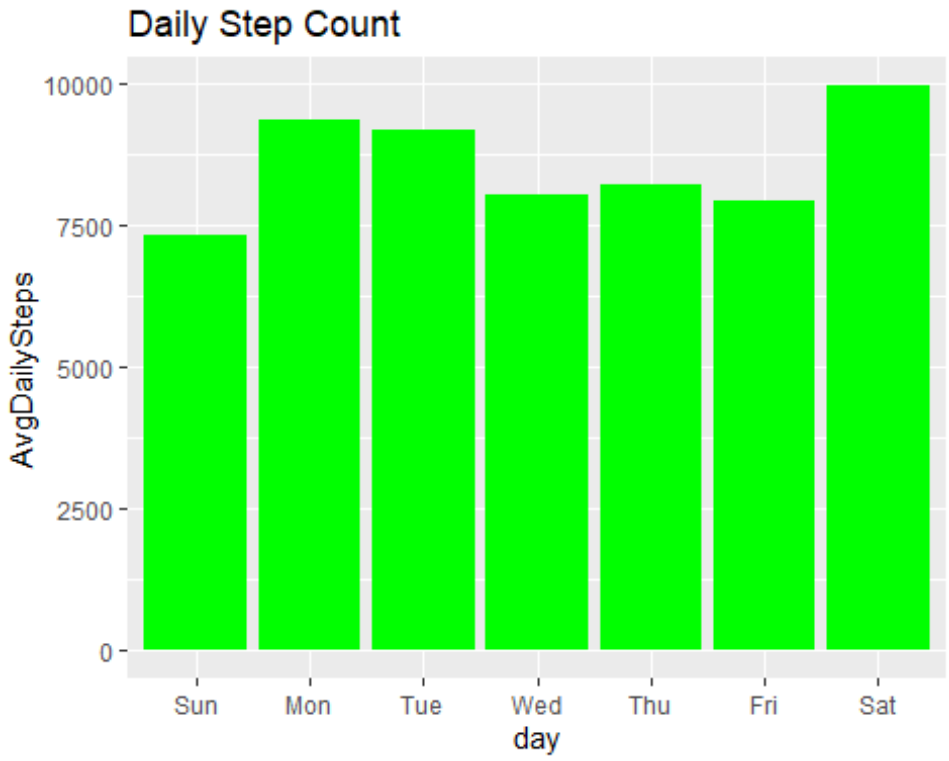
Whether the day of the week affects our activity levels and sleep.

```
# aggregate data by day of week to summarize averages
merged_data <- mutate(merged_data,day = wday(SleepDay, label = TRUE))
summarized_activity_sleep <- merged_data %>%
  group_by(day) %>%
  summarise(AvgDailySteps = mean(TotalSteps),
            AvgAsleepMinutes = mean(TotalMinutesAsleep),
            AvgAwakeTimeInBed = mean(TotalTimeInBed),
            AvgSedentaryMinutes = mean(SedentaryMinutes),
            AvgLightlyActiveMinutes = mean(LightlyActiveMinutes),
            AvgFairlyActiveMinutes = mean(FairlyActiveMinutes),
            AvgVeryActiveMinutes = mean(VeryActiveMinutes),
            AvgCalories = mean(Calories))
head(summarized_activity_sleep)

## # A tibble: 6 × 9
##   day   AvgDailySteps AvgAsleepMinutes AvgAwakeTimeInBed
AvgSedentaryMinutes
##   <ord>         <dbl>            <dbl>             <dbl>
<dbl>
```

```
## 1 Sun           7298.             453.              504.
688.
## 2 Mon           9340.             419.              456.
718.
## 3 Tue           9183.             405.              443.
740.
## 4 Wed           8023.             435.              470.
714.
## 5 Thu           8205.             402.              436.
701.
## 6 Fri           7901.             405.              445.
743.
## # ℹ 4 more variables: AvgLightlyActiveMinutes <dbl>,
## #   AvgFairlyActiveMinutes <dbl>, AvgVeryActiveMinutes <dbl>, AvgCalories
<dbl>
```

```
ggplot(data = summarized_activity_sleep, mapping = aes(x = day, y =
AvgDailySteps)) +
geom_col(fill = "green") + labs(title = "Daily Step Count")
```



The bar graph above shows us that participants are most active on saturdays and least active on sundays.

6. Act

After analyzing the FitBit Fitness Tracker data, I came up with some recommendations for Bellabeat marketing strategy based on trends in smart device usage.

- The majority of participants are lightly active. Bellabeat should offer a progression system in the app to encourage participants to become at least fairly active.
- If users want to improve the quality of their sleep, Bellabeat should consider using app notifications reminding users to get enough rest, as well as recommending reducing sedentary time.
- Participants are most active on Saturdays. Bellabeat can use this knowledge to remind users to go for a walk or a jog on that day. Participants seem to be the least active on Sundays. Bellabeat can use this to motivate users to go out and continue exercising on Sundays.