Bellabeat Project: How Can a Wellness Technology Company Play It Smart?

Introduction

This Capstone project is part of Google Data Analytics Course from Coursera. This project I am perform as junior data analyst who working on the marketing analyst team at Bellabeat, a high-tech manufacturer of health-focused products for women. Our task is analyze device fitness data to find new growth opportunities for the company. I have been asked to search for insight of how consumers are using their smart devices, and use that insights to help guide marketing strategy for the company. Then Present the analysis to the Bellabeat executive team along with high-level recommendations for Bellabeat's marketing strategy.

Ask

Business task

- Identify trends in smart device usage
- apply these trend to Bellabeat customer and help influence Bellabeat marketing strategy

Key Stakeholder

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

Prepare

Data source dataset from FitBit Fitness Tracker Data (CC0: Public Domain, dataset made available through Mobius. 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.

Limitation data came from about thirty users and for two month long. Record in some fields such as sleep and weight are come from a fewer users than others, So we have a fewer data to analyze trend in data visualize that relate with this data field. Moreover this data is about seven years old so it might not relevant with user behaviors today.

Process

1. Extract zip file, check all data table to choose data table we want to use in analytic.I chose table 'Daily_activity_merged.csv', 'sleep_day_merged.csv' and 'weight_log_info_merged.csv'

2. Change column name that record data about date in every table to 'date' and change data format for this column from 'general' to 'date' and save table to 'Daily_activity_cleaned.csv', 'sleep_day_cleaned.csv' and 'weight_log_info_cleaned.csv'

3. Open R and install packages

4. Import data into R

```
daily_activity <- read_csv("daily_activity_cleaned.csv")
```

```
## Rows: 940 Columns: 15
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): Date
## 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("sleep_day_cleaned.csv")
```

```
## Rows: 413 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Date
## 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.
```

5. Exploring each table

```
colnames(daily_activity)
```

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

```
head(daily_activity)
```

```
## # A tibble: 6 x 15
##         Id Date  TotalSteps TotalDistance TrackerDistance LoggedActivitiesDist~1
##      <dbl> <chr>      <dbl>         <dbl>           <dbl>                   <dbl>
## 1   1.50e9 4/12~      13162          8.5             8.5                        0
## 2   1.50e9 4/13~      10735          6.97            6.97                       0
## 3   1.50e9 4/14~      10460          6.74            6.74                       0
```

```
## 4   1.50e9 4/15~       9762          6.28           6.28                      0
## 5   1.50e9 4/16~      12669          8.16           8.16                      0
## 6   1.50e9 4/17~       9705          6.48           6.48                      0
## # i abbreviated name: 1: LoggedActivitiesDistance
## # i 9 more variables: VeryActiveDistance <dbl>, ModeratelyActiveDistance <dbl>,
## #   LightActiveDistance <dbl>, SedentaryActiveDistance <dbl>,
## #   VeryActiveMinutes <dbl>, FairlyActiveMinutes <dbl>,
## #   LightlyActiveMinutes <dbl>, SedentaryMinutes <dbl>, Calories <dbl>
```

```
str(daily_activity)
```

```
## spc_tbl_ [940 x 15] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Id                       : num [1:940] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ Date                     : chr [1:940] "4/12/2016" "4/13/2016" "4/14/2016" "4/15/2016" ...
##  $ TotalSteps               : num [1:940] 13162 10735 10460 9762 12669 ...
##  $ TotalDistance            : num [1:940] 8.5 6.97 6.74 6.28 8.16 ...
##  $ TrackerDistance          : num [1:940] 8.5 6.97 6.74 6.28 8.16 ...
##  $ LoggedActivitiesDistance: num [1:940] 0 0 0 0 0 0 0 0 0 0 ...
##  $ VeryActiveDistance       : num [1:940] 1.88 1.57 2.44 2.14 2.71 ...
##  $ ModeratelyActiveDistance: num [1:940] 0.55 0.69 0.4 1.26 0.41 ...
##  $ LightActiveDistance      : num [1:940] 6.06 4.71 3.91 2.83 5.04 ...
##  $ SedentaryActiveDistance : num [1:940] 0 0 0 0 0 0 0 0 0 0 ...
##  $ VeryActiveMinutes        : num [1:940] 25 21 30 29 36 38 42 50 28 19 ...
##  $ FairlyActiveMinutes      : num [1:940] 13 19 11 34 10 20 16 31 12 8 ...
##  $ LightlyActiveMinutes     : num [1:940] 328 217 181 209 221 164 233 264 205 211 ...
##  $ SedentaryMinutes         : num [1:940] 728 776 1218 726 773 ...
##  $ Calories                 : num [1:940] 1985 1797 1776 1745 1863 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Id = col_double(),
##   ..   Date = col_character(),
##   ..   TotalSteps = col_double(),
##   ..   TotalDistance = col_double(),
##   ..   TrackerDistance = col_double(),
##   ..   LoggedActivitiesDistance = col_double(),
##   ..   VeryActiveDistance = col_double(),
##   ..   ModeratelyActiveDistance = col_double(),
##   ..   LightActiveDistance = col_double(),
##   ..   SedentaryActiveDistance = col_double(),
##   ..   VeryActiveMinutes = col_double(),
##   ..   FairlyActiveMinutes = col_double(),
##   ..   LightlyActiveMinutes = col_double(),
##   ..   SedentaryMinutes = col_double(),
##   ..   Calories = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
colnames(sleep_day)
```

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

```r
head(sleep_day)
```

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

```r
str(sleep_day)
```

```
## spc_tbl_ [413 x 5] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Id                : num [1:413] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ Date              : chr [1:413] "4/12/2016" "4/13/2016" "4/15/2016" "4/16/2016" ...
##  $ TotalSleepRecords : num [1:413] 1 2 1 2 1 1 1 1 1 1 ...
##  $ TotalMinutesAsleep: num [1:413] 327 384 412 340 700 304 360 325 361 430 ...
##  $ TotalTimeInBed    : num [1:413] 346 407 442 367 712 320 377 364 384 449 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Id = col_double(),
##   ..   Date = col_character(),
##   ..   TotalSleepRecords = col_double(),
##   ..   TotalMinutesAsleep = col_double(),
##   ..   TotalTimeInBed = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

6.check and clean data

6.1 Change column name to lowercase

```r
names(daily_activity) <- tolower(names(daily_activity))
names(sleep_day) <- tolower(names(sleep_day))
```

6.2 Check number of unique participants

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

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

```r
n_distinct(sleep_day$id)
```

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

6.3 Check missing value: we found 65 null value which all located in “fat” column in weight_log table. I desire to leave that cause I did not use this column to analyze anything.

```r
which(is.na(daily_activity))
```

```
## integer(0)
```

```r
which(is.na(sleep_day))
```

```
## integer(0)
```

6.4 Change date format to be consistent

```r
daily_activity <- daily_activity %>%
  mutate(date = as_date(date, format = "%m/%d/%Y"))

sleep_day <- sleep_day %>%
  mutate(date = as_date(date, format = "%m/%d/%Y"))
```

6.5 Merge 2 table using 'id' and 'date' to be identical key

```r
activity_sleep_daily <- merge(daily_activity, sleep_day, by=c('id', 'date'))
  head(activity_sleep_daily)
```

```
##           id       date totalsteps totaldistance trackerdistance
## 1 1503960366 2016-04-12      13162          8.50            8.50
## 2 1503960366 2016-04-13      10735          6.97            6.97
## 3 1503960366 2016-04-15       9762          6.28            6.28
## 4 1503960366 2016-04-16      12669          8.16            8.16
## 5 1503960366 2016-04-17       9705          6.48            6.48
## 6 1503960366 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
##   totalsleeprecords totalminutesasleep totaltimeinbed
## 1                 1                327            346
```

```
## 2                 2                 384               407
## 3                 1                 412               442
## 4                 2                 340               367
## 5                 1                 700               712
## 6                 1                 304               320
```

Analyze and Share

Summary daily_activity table

```
daily_activity %>%
  select(totalsteps,
  sedentaryminutes,
  calories) %>%
summary()
```

```
##    totalsteps    sedentaryminutes    calories
##  Min.   :    0   Min.   :   0.0   Min.   :   0
##  1st Qu.: 3790   1st Qu.: 729.8   1st Qu.:1828
##  Median : 7406   Median :1057.5   Median :2134
##  Mean   : 7638   Mean   : 991.2   Mean   :2304
##  3rd Qu.:10727   3rd Qu.:1229.5   3rd Qu.:2793
##  Max.   :36019   Max.   :1440.0   Max.   :4900
```

Out put shown that users burn calories about 2304 kcal, walk about 7638 total steps, and spent about 991 sedentary minutes.

Summary sleep_day table

```
sleep_day %>%
  select(totalsleeprecords,
  totalminutesasleep,
  totaltimeinbed) %>%
summary()
```

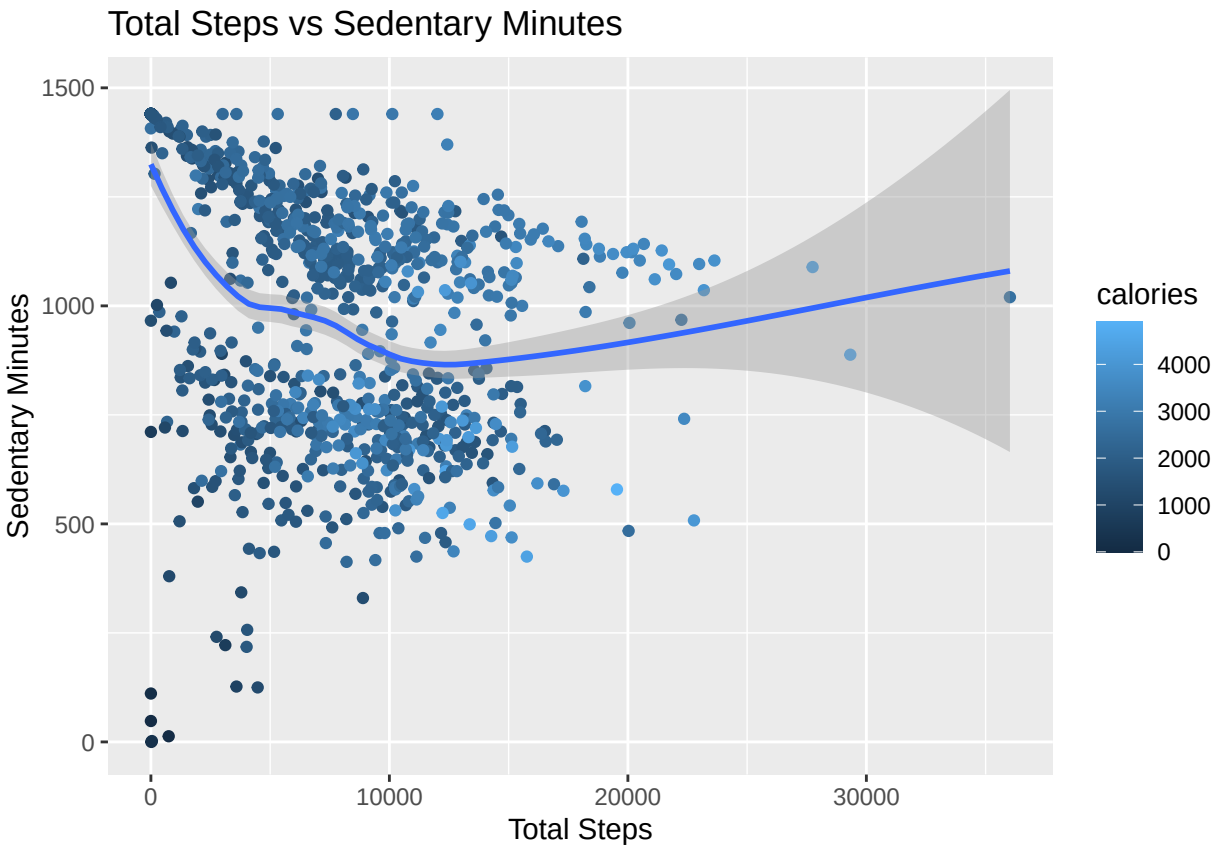
```
##  totalsleeprecords totalminutesasleep totaltimeinbed
##  Min.   :1.000     Min.   : 58.0      Min.   : 61.0
##  1st Qu.:1.000     1st Qu.:361.0      1st Qu.:403.0
##  Median :1.000     Median :433.0      Median :463.0
##  Mean   :1.119     Mean   :419.5      Mean   :458.6
##  3rd Qu.:1.000     3rd Qu.:490.0      3rd Qu.:526.0
##  Max.   :3.000     Max.   :796.0      Max.   :961.0
```

users slept about 420 minutes or 7 hours.

Plot Total Steps vs Sedentary Minutes

```
ggplot(data = daily_activity)+
  geom_point(mapping = aes(x=totalsteps,y=sedentaryminutes, color=calories))+
  geom_smooth(mapping = aes(x=totalsteps,y=sedentaryminutes))+
labs(x="Total Steps", y="Sedentary Minutes", title = "Total Steps vs Sedentary Minutes")
```

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



The scatter plot shows that sedentary minutes decreases with increasing of total steps. Which mean they are inverse relation between 0 - 10000 total steps (seem a little positive correlation when reaching about 15000 total steps, since there are little data scatter in that area so we do not confident about that). Moreover calories burn seem higher with increase of total steps, mean that it is in positive correlation.

Classify into weekday classify date into week day

```
activity_sleep_weekday <- activity_sleep_daily %>%
  mutate(weekday = weekdays(date))
activity_sleep_weekday$weekday <- ordered(activity_sleep_weekday$weekday,
  levels = c("Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday", "Sunday"))
```

summarize mean steps in each day

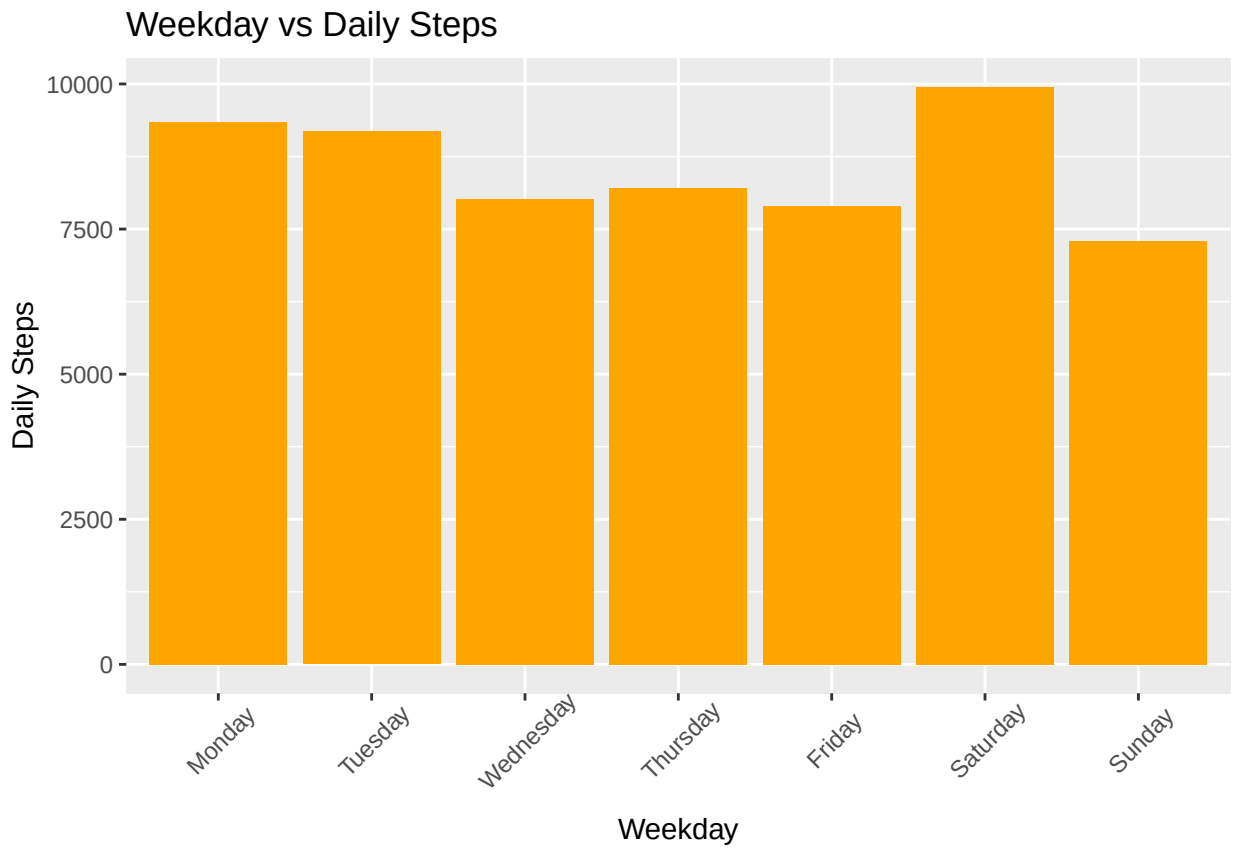
```
activity_weekday <- activity_sleep_weekday %>%
  group_by(weekday) %>%
  summarize (daily_steps = mean(totalsteps), daily_sleep = mean(totalminutesasleep), daily_calories=mean
head(activity_weekday)
```

```
## # A tibble: 6 x 4
##   weekday   daily_steps daily_sleep daily_calories
##   <ord>           <dbl>       <dbl>          <dbl>
```

```
## 1 Monday          9340.        419.          2465.
## 2 Tuesday         9183.        405.          2496.
## 3 Wednesday       8023.        435.          2378.
## 4 Thursday        8205.        402.          2316.
## 5 Friday          7901.        405.          2330.
## 6 Saturday        9949.        421.          2527.
```

summarize mean steps in each day

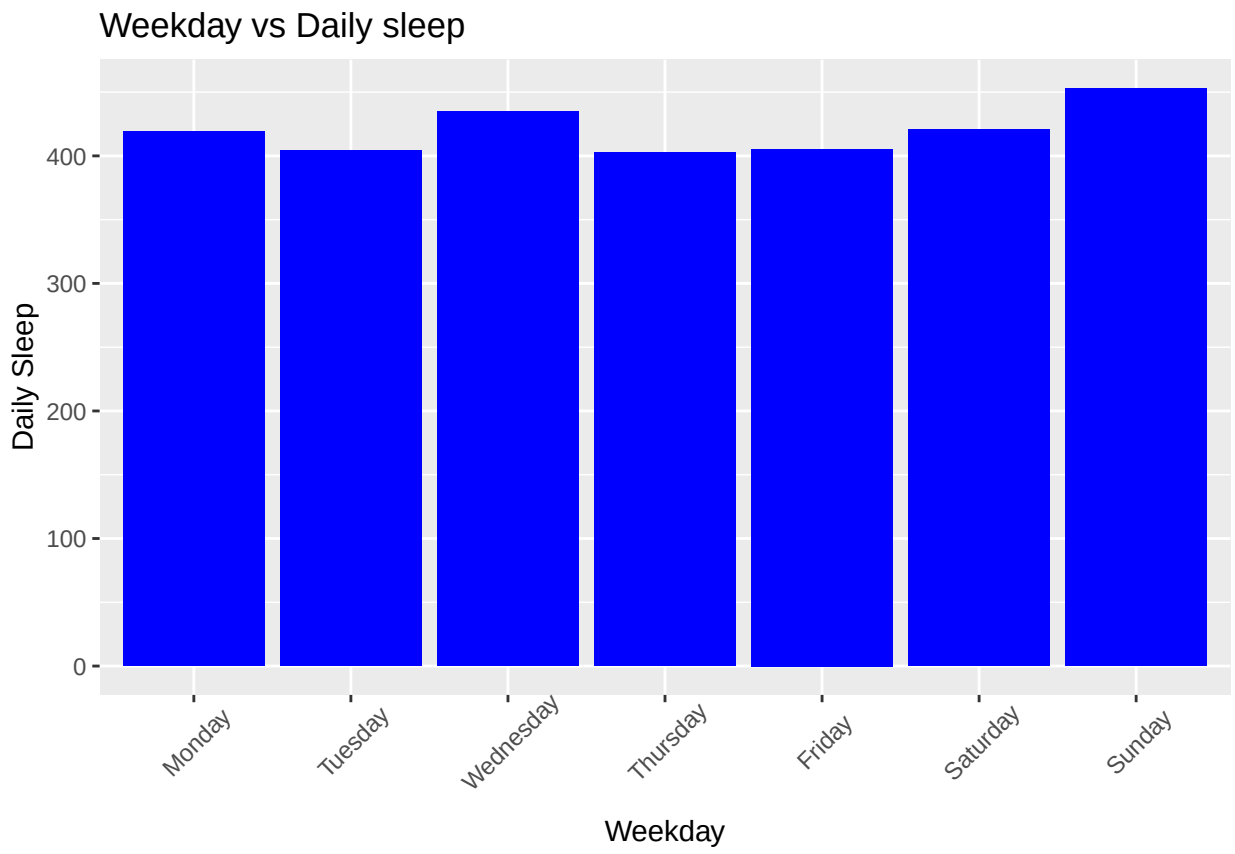
```
ggplot(data = activity_weekday)+
  geom_col(mapping = aes(weekday, daily_steps), fill="orange") +
labs(x="Weekday", y="Daily Steps", title = "Weekday vs Daily Steps")+
theme(axis.text.x = element_text(angle = 45,vjust = 0.7, hjust = 0.5))
```



Article "Counting Your Steps" suggest that for healthy adults the step goal of 10,000 is the recommended daily step target. Compare with our chart shows that users average step is about 7,500 steps per day, which is lower than step suggestion from the article. Only on Saturday that user have highest steps at about 9,948 steps which almost reach 10,000 target.

summarize mean sleep in each day

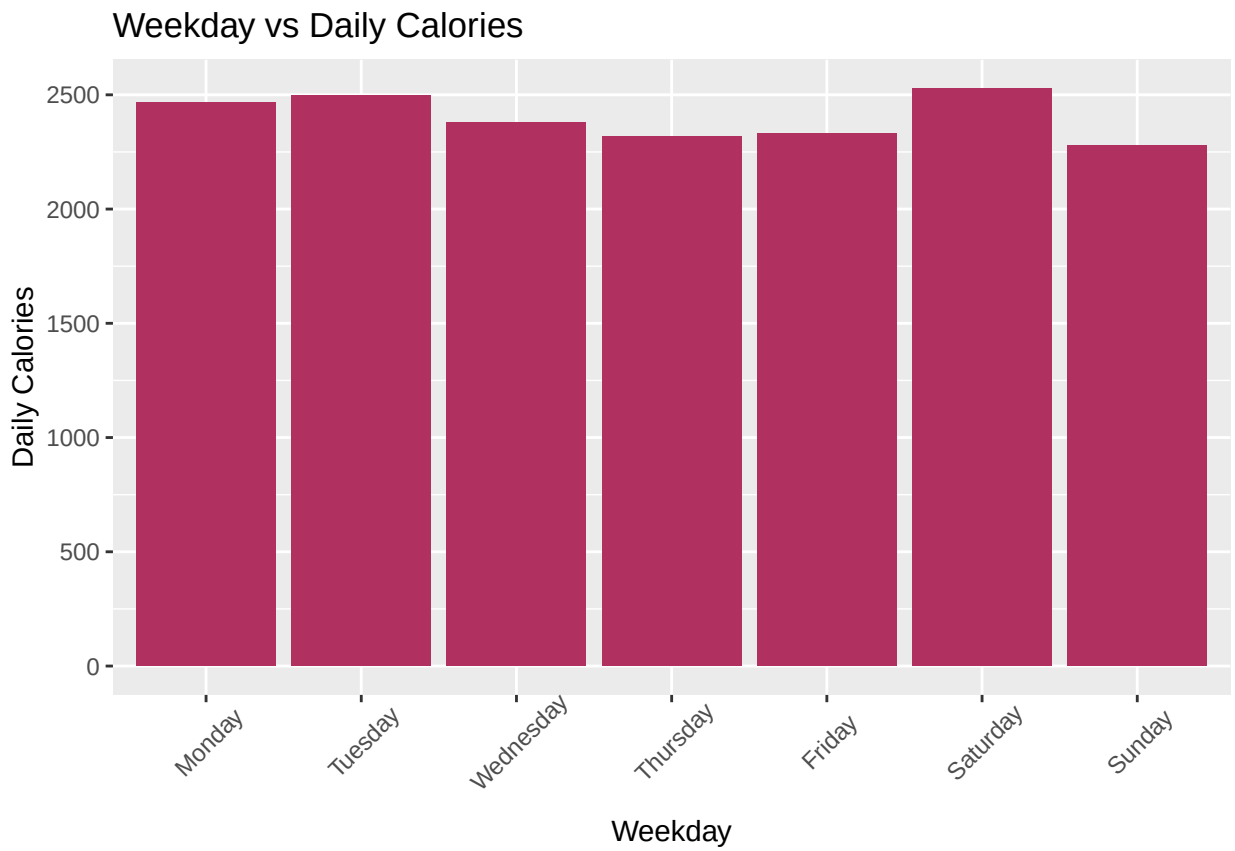
```
ggplot(data = activity_weekday)+
  geom_col(mapping = aes(weekday, daily_sleep), fill="blue") +
labs(x="Weekday", y="Daily Sleep", title = "Weekday vs Daily sleep")+
theme(axis.text.x = element_text(angle = 45,vjust = 0.7, hjust = 0.5))
```



All users are not reach recommendation sleep hours (8 hours), most users spent about about 6-7 to sleep. highest sleep minutes occur on Saturday which is reasonable compare with "weekly vs daily steps bar chart".

summarize mean calories in each day

```
ggplot(data = activity_weekday)+
  geom_col(mapping = aes(weekday, daily_calories), fill="maroon") +
labs(x="Weekday", y="Daily Calories", title = "Weekday vs Daily Calories")+
theme(axis.text.x = element_text(angle = 45,vjust = 0.7, hjust = 0.5))
```



Bar chart shows that users burned calories higher in Monday, Tuesday and Saturday. According to the Dietary Guidelines for Americans 2020–2025, Adult estimated calorie needs range from 1,600 to 2,400 calories per day for females and 2,000 to 3,000 calories per day for males. Use this range of calories need to represent range of calories burned (calories in = calories out), average calories at 1800 to 2700 will be key estimate to describe user's burned calories. From there it shows that all users are use calories within estimate range.

Act

1. For the lower of users sleep data tracks
    - most user are not collect their sleep data. It might because wearing those collecting tool while asleep is not feel comfortable, we need to survey and collect more data to confirm this hypothesis
    - most users are not reach recommendation sleep hours

We might use these 2 point to create marketing campaign that encourage people to arrange their sleep time to reach 8 recommendation hours along with showing our products such as "leaf" to tracking this data using the "comfortable feeling" campaign

2. For daily activity data:
    - From total steps daily data shown that users total steps per day is about 7500 steps which is not reaching recommendation steps per day, and when compare with totals steps categorize into weekday we found that those average 7,500 steps are occur on Wednesday, Thursday, Friday and Sunday
    - From calories data we found that it move in correlation with total steps data
    - From sedentary minutes we found that it move inverse with total steps

From these insight We might create campaign that encourage people to use health and activity tracking devices to know their own activity. Showing our products that have beauty and variety and when combining with our bellabeat app it would be useful for user. For example we know that users is have lower steps and burn lower calories on Wed, Thu, Fri and Sat, So we will send alert to our user that user steps is not reach recommendation steps or might create campaign on those day to encourage user to do more activities.

Recommendations for further improve

1. Since you have the "spring" product that tracks daily water intake, it would be good if you tie in this product with in other product's marketing campaign. It might bring new interested insights.
2. For weight log table, since weight data is depend on user responsible for manually measure and record it. If you plan to produce product that collect weight you might have to think about how you can remind user to usually measure their and record their own data.