Bellabeat Capstone project

Feng-Chiu Tsai-Goss

2024-07-19

Introduction

Bellabeat is a high-tech company with the focus on health products for women, and plans to expand in the global smart device market. Bellabeat designs technology that informs, inspires, and collectas data on activity, sleep, stress and reproductive health to empower women with knowledge about their own health and habits. The company has invested two marketing strategies : traditional media (such as radio, out-of-home billboards, print, and television) and digital marketing (such as Google search and Dispaly Network, active Facebook and Instagram pages, video aids on YouTube, Twitter). The cofounder and Chief Creative Office of Bellabeat, Urska Srsen, asks the marketing analytic team to focus on a Bellabeat product and analyze smart device usage data to gain insight into how people are already using their smart devices. She is seeking the opportunity to growing into a bigger market and would like high-level recommendations for how these trends can inform and guide marketing strategy for the company. Below is my analysis based on the data analysis process: Ask, Prepare, Process, Analyze, Share, Act.

Ask

The project is aiming to identify tends that non-Bellabeat customers use smart fitness devices and further to provide insights to inform Bellabeat's marketing strategy. Here are questions that guided the analysis. a. What are some trends in smart device usage?
b. How could these trends apply to Bellabeat customers?
c. How could these trends help influence Bellabeat marketing strategy?

Our team will provide and present the findings and recommendations to the stakeholders:

Urška Sršen: Bellabeat's cofounder and Chief Creative Officer

Sando Mur: key member of the Bellabeat executive team

Bellabeat marketing analytics team

Prepare

1.Data sources and organization

The data used for this project was collected through a survey via Amazon Mechanical Turk between 03.12.2016-05.12.2016. The data shares thirty Fitbit users who consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. Thirteen sets of data was organized in long format and the rest was in wide format.https://www.kaggle.com/datasets/arashnic/fitbit

2.Accessibility and privacy of data

The data is open and publicly shared by Mobius on Kaggle. The data is licensed by CCO: Public Domain by waiving all his or her rights to the work worldwide under copyright law. One can copy, modify, distribute, and perform data without asking permission. In the data, all the participants were assigned an ID number without listing their credentials in the files.

3.ROCCC analysis

To verify the credibility of data, the ROCCC process was conducted.

(1) Reliable: Low. The dataset includes only 30 participants without knowing their genders, ages, and nationality so it might have concern of the sample selection bias to reflect the overall population.

(2) Original: Low. The dataset was collected by the third-party Amazon Mechanical Turk and published by Mobius on Kaggle.

(3) Comprehensive: Medium. The dataset provides information on daily activity intensity, calories used, daily steps taken, daily sleep time, and distance travelled in various levels of activities.

(4) Current: Low. The data set was collected in 2016, which has been 8 years old. The smart device trackers have been updated for the last 10 years. Functionalities and service might be different between now and then; therefore, it is needed to collect more information to help know better the fitness activities and health life habits.

(5) Cited: High. The dataset is cited with sources being well documented..

4.Data integrity

Dataset was sorted, organized, and sorted in 18 .cvs files. I firstly open the datasets in Excel and realized that the dataset from 03.12.2016-04.11.2016 is incomplete. Some data frames are in a minute-level output. Thus, several data frames will not be used for the analysis. This project will focus on the data set from 04.12.2016-05.12.2016, with the analysis on dailyActivity_merged and sleepDay_merged to identify trends of people using smart devices as daily habits for fitness purposes.

Process

First, I downloaded the dataset and stored it in OneDrive in Microsoft 365. Then I checked the file names that were already named in a way that we can easily recognize and separate from other files in the same folder.

This project will utilize the R programming in R Studio to clean, analyze, and create visualizations of the data.

1.Install and load the following packages in R Studio, and then open libraries.

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

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```
install.packages("lubridate")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```
install.packages("readr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```
install.packages("ggplot2")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```r
install.packages("cowplot")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

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

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```r
install.packages("janitor")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

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

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```r
install.packages("dplyr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
```

```
## -- 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
```

```r
library(lubridate)
library(readr)
library(ggplot2)
library(cowplot)
```

```
##
## Attaching package: 'cowplot'
##
## The following object is masked from 'package:lubridate':
##
##     stamp
```

```r
library(here)
```

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

```r
library(janitor)
```

```
##
```

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

```r
library(skimr)
library(dplyr)
```

2.Import datasets into RStudio

The following datasets are used for this project. * DailyActivity_merged: Daily Activities over 31 days of 33 IDs. Tracking daily steps, distance, intensity, active time, and calories. * sleepDay_merged: Daily sleeping time over 31 days of 24 IDs. Recording daily sleeping time and time in bed.

```r
DailyActivity <- read.csv("/cloud/project/Bellabeat/dailyActivity_merged.csv")
Sleep_Activity <- read.csv("/cloud/project/Bellabeat/sleepDay_merged.csv")
```

3.Preview data

After importing data sets, I firstly checked how many rows and columns are in each data frame, and then veiwed the column names. There are 15 columns and 940 rows in the data frame of daily activity, and 5 columns and 413 rows in the data frame of sleeping activity.

```r
glimpse(DailyActivity)
```

```
## Rows: 940
## Columns: 15
## $ Id                       <dbl> 1503960366, 1503960366, 1503960366, 150396036~
## $ ActivityDate             <chr> "4/12/2016", "4/13/2016", "4/14/2016", "4/15/~
## $ TotalSteps               <int> 13162, 10735, 10460, 9762, 12669, 9705, 13019~
## $ TotalDistance            <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9.8~
## $ TrackerDistance          <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9.8~
## $ LoggedActivitiesDistance <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ VeryActiveDistance       <dbl> 1.88, 1.57, 2.44, 2.14, 2.71, 3.19, 3.25, 3.5~
## $ ModeratelyActiveDistance <dbl> 0.55, 0.69, 0.40, 1.26, 0.41, 0.78, 0.64, 1.3~
## $ LightActiveDistance      <dbl> 6.06, 4.71, 3.91, 2.83, 5.04, 2.51, 4.71, 5.0~
## $ SedentaryActiveDistance  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ VeryActiveMinutes        <int> 25, 21, 30, 29, 36, 38, 42, 50, 28, 19, 66, 4~
## $ FairlyActiveMinutes      <int> 13, 19, 11, 34, 10, 20, 16, 31, 12, 8, 27, 21~
## $ LightlyActiveMinutes     <int> 328, 217, 181, 209, 221, 164, 233, 264, 205, ~
## $ SedentaryMinutes         <int> 728, 776, 1218, 726, 773, 539, 1149, 775, 818~
## $ Calories                 <int> 1985, 1797, 1776, 1745, 1863, 1728, 1921, 203~
```

```r
glimpse(Sleep_Activity)
```

```
## Rows: 413
## Columns: 5
## $ Id                 <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 150~
## $ SleepDay           <chr> "4/12/2016 12:00:00 AM", "4/13/2016 12:00:00 AM", "~
## $ TotalSleepRecords  <int> 1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ TotalMinutesAsleep <int> 327, 384, 412, 340, 700, 304, 360, 325, 361, 430, 2~
## $ TotalTimeInBed     <int> 346, 407, 442, 367, 712, 320, 377, 364, 384, 449, 3~
```

4.Cleaning data

Now inspect the data to see if there are any duplicates, null values, formatting errors or any inconsistencies.

```r
sum(duplicated(DailyActivity))
```

(1) check and remove duplicates and N/A values

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

```r
sum(duplicated(Sleep_Activity))
```

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

```r
n_unique(DailyActivity$Id)
```

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

```r
n_unique(Sleep_Activity$Id)
```

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

```r
sum(!complete.cases(DailyActivity))
```

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

```r
sum(!complete.cases(Sleep_Activity))
```

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

```r
cleaned_sleep <- Sleep_Activity %>%
  distinct( )
sum(duplicated(cleaned_sleep))
```

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

(2) clean and rename columns names Then I want to ensure that all the columns names are unique and consistent, including numbers, letter, underscores in the name only.

```r
DailyActivity_cleaned <- DailyActivity %>%
  distinct( ) %>%
  drop_na

head(DailyActivity_cleaned)
```

```
##           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
```

```
cleaned_sleep <- Sleep_Activity %>%
  distinct( )
sum(duplicated(cleaned_sleep))
```

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

(3) check errors of formatting Then I ensure column names are in the consistent format across data frames for the later merging.

```
str(DailyActivity_cleaned)
```

```
## '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(cleaned_sleep)
```

```
## 'data.frame':    410 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 ...
```

I noticed that I need to correct the date format in the frames that should be a date form.

```
DailyActivity_cleaned$ActivityDate <- as.Date(DailyActivity_cleaned$ActivityDate, '%m/%d/%y')

cleaned_sleep$SleepDay <- as.Date(cleaned_sleep$SleepDay, '%m/%d/%y')

str(DailyActivity)
```

```
## '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 ...
```

```r
str(cleaned_sleep)
```

```
## 'data.frame':    410 obs. of  5 variables:
##  $ Id                : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ SleepDay          : Date, format: "2020-04-12" "2020-04-13" ...
##  $ 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 ...
```

Then I separated date and time in the SleepDay column into two different columns so that I can combine merge the frames next.

(4) Adding a new column I also add a new column in the data frame of sleeping activity that finds the difference between the total time in bed and total minute asleep by the code below. That might help me to study the relation to other factors in activities.

```r
New_Sleep <- cleaned_sleep %>%
  mutate(diff = TotalTimeInBed - TotalMinutesAsleep)
view(New_Sleep)
```

Before merging, I rename columns regarding dates and identifications to be consistent between these two data sets so that I can combine them into one data frame to further investigate the relationship between daily activities and sleeping pattern.

```r
New_Sleep_cleaned <- New_Sleep %>%
  rename(
      Date = SleepDay,
      Id = Id
  )

view(New_Sleep_cleaned)

Activity_cleaned <- DailyActivity_cleaned %>%
  rename (
    Date = ActivityDate,
    Id = Id
  )

view(Activity_cleaned)
```

```r
DailyActivity_Sleep_merged <- merge(Activity_cleaned, New_Sleep_cleaned, by= c ("Id", "Date"))

head(DailyActivity_Sleep_merged)
```

```
##           Id       Date TotalSteps TotalDistance TrackerDistance
## 1 1503960366 2020-04-12      13162          8.50            8.50
## 2 1503960366 2020-04-13      10735          6.97            6.97
## 3 1503960366 2020-04-15       9762          6.28            6.28
## 4 1503960366 2020-04-16      12669          8.16            8.16
## 5 1503960366 2020-04-17       9705          6.48            6.48
## 6 1503960366 2020-04-19      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 diff
## 1                 1                327            346   19
## 2                 2                384            407   23
## 3                 1                412            442   30
## 4                 2                340            367   27
## 5                 1                700            712   12
## 6                 1                304            320   16
```

After merging, there are 24 participants that are on the final list for this project since they are the overlap between these two data sets.

Analyze

Atthe analyze phase, I will analyze the patterns and trends of these thirties FitBit users by providing statistical calculation information.

1.Summarize and explore data sets

```r
DailyActivity_Sleep_merged %>%

  select(TotalSteps,
```

```
        TotalDistance,

        TotalMinutesAsleep,

        TotalTimeInBed,

        diff,

        Calories) %>%

  summary()
```

```
##    TotalSteps    TotalDistance    TotalMinutesAsleep TotalTimeInBed
##  Min.   :   17   Min.   : 0.010   Min.   : 58.0      Min.   : 61.0
##  1st Qu.: 5189   1st Qu.: 3.592   1st Qu.:361.0      1st Qu.:403.8
##  Median : 8913   Median : 6.270   Median :432.5      Median :463.0
##  Mean   : 8515   Mean   : 6.012   Mean   :419.2      Mean   :458.5
##  3rd Qu.:11370   3rd Qu.: 8.005   3rd Qu.:490.0      3rd Qu.:526.0
##  Max.   :22770   Max.   :17.540   Max.   :796.0      Max.   :961.0
##       diff           Calories
##  Min.   :  0.00   Min.   : 257
##  1st Qu.: 17.00   1st Qu.:1841
##  Median : 25.50   Median :2207
##  Mean   : 39.31   Mean   :2389
##  3rd Qu.: 40.00   3rd Qu.:2920
##  Max.   :371.00   Max.   :4900
```

Findings:

*The average of daily total steps is 8515. It is recommended that an adult should aim to walk 10,000 steps in a day for health benefits.

*The average daily total distance is 6.012 kilometers and the median is 6.27 kilometers. That means, half of the participants moved less than 6.27 miles in a day.

*The average daily total asleep time is 419.2 minutes, which approximately equals to 7 hours a day. Also, average time that people stay on bed while not asleep is 39.31 minutes. It is recommended that staying for 15-30 minutes on bed after waking up should be enough for most people. Using few minutes to stretch body before getting out of your bed will benefit making your day better.

*The average calories burn is 2389 per day.

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

```
##  VeryActiveDistance ModeratelyActiveDistance LightActiveDistance
##  Min.   : 0.000     Min.   :0.0000           Min.   :0.010
##  1st Qu.: 0.000     1st Qu.:0.0000           1st Qu.:2.540
##  Median : 0.570     Median :0.4200           Median :3.665
##  Mean   : 1.446     Mean   :0.7439           Mean   :3.791
##  3rd Qu.: 2.360     3rd Qu.:1.0375           3rd Qu.:4.918
##  Max.   :12.540     Max.   :6.4800           Max.   :9.480
```

```
##  SedentaryActiveDistance
##  Min.   :0.0000000
##  1st Qu.:0.0000000
##  Median :0.0000000
##  Mean   :0.0009268
##  3rd Qu.:0.0000000
##  Max.   :0.1100000
```

*The average sedentary active distance is almost close to 0 miles per day whereas the average moderately active distance is 0.74 miles. This needs to be improved.

```
DailyActivity_Sleep_merged %>%
  select(VeryActiveMinutes,
         FairlyActiveMinutes,
         LightlyActiveMinutes,
         SedentaryMinutes) %>%
  summary()
```

```
##  VeryActiveMinutes FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes
##  Min.   :  0.00    Min.   :  0.00      Min.   :  2.0        Min.   :   0.0
##  1st Qu.:  0.00    1st Qu.:  0.00      1st Qu.:158.0        1st Qu.: 631.2
##  Median :  9.00    Median : 11.00      Median :208.0        Median : 717.0
##  Mean   : 25.05    Mean   : 17.92      Mean   :216.5        Mean   : 712.1
##  3rd Qu.: 38.00    3rd Qu.: 26.75      3rd Qu.:263.0        3rd Qu.: 782.8
##  Max.   :210.00    Max.   :143.00      Max.   :518.0        Max.   :1265.0
```

*The average active minutes is close to the average fairly active minutes, but much shorter than the average lightly active time and sedentary minutes. This needs to be improved. However, there is a need to investigate whether the sedentary minutes includes the measurement of asleep time.

2.Summarize weekday distance and activity time

I will find a summary of total distance and active time from the datasets and then display data in bar graphs.

```
DailyActivity_Sleep_merged <- DailyActivity_Sleep_merged %>%
  mutate(day = format(ymd(Date), format = '%a'))
head(DailyActivity_Sleep_merged)
```

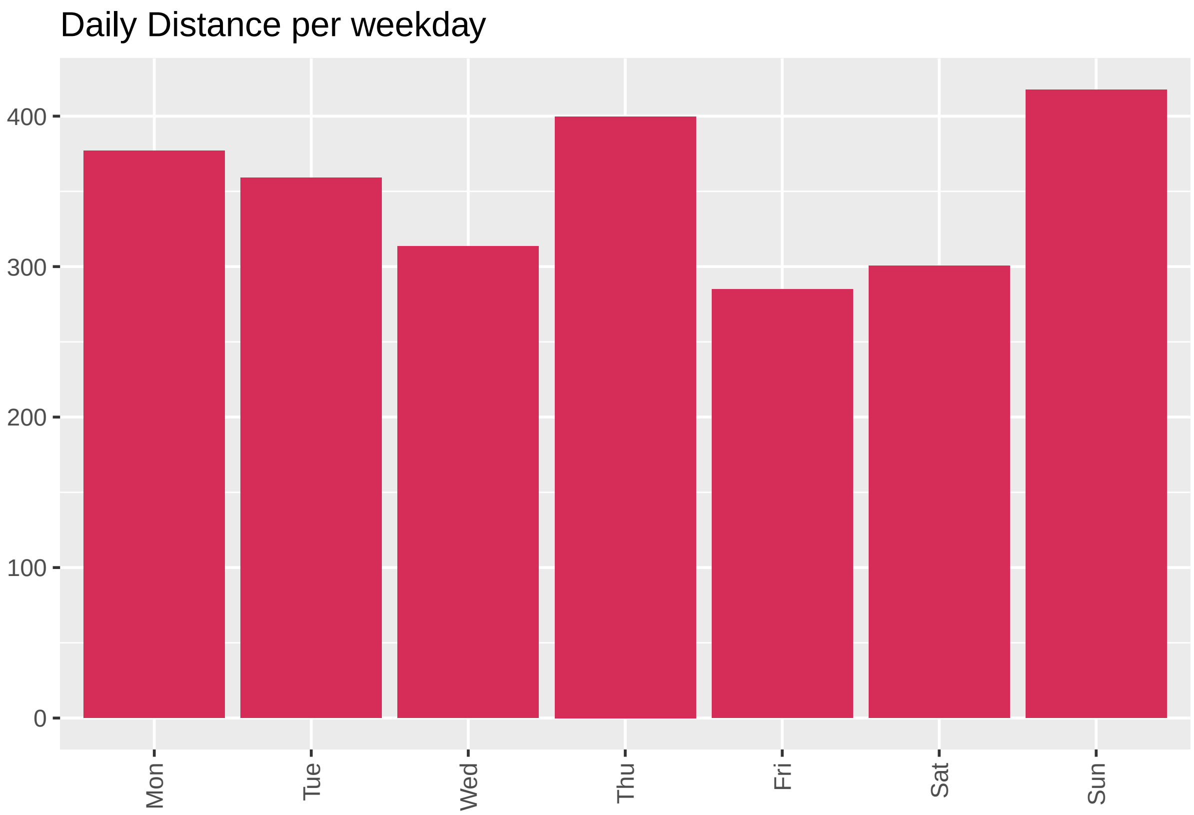
```
##           Id       Date TotalSteps TotalDistance TrackerDistance
## 1 1503960366 2020-04-12      13162          8.50            8.50
## 2 1503960366 2020-04-13      10735          6.97            6.97
## 3 1503960366 2020-04-15       9762          6.28            6.28
## 4 1503960366 2020-04-16      12669          8.16            8.16
## 5 1503960366 2020-04-17       9705          6.48            6.48
## 6 1503960366 2020-04-19      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 diff day
## 1                 1                327            346   19 Sun
## 2                 2                384            407   23 Mon
## 3                 1                412            442   30 Wed
## 4                 2                340            367   27 Thu
## 5                 1                700            712   12 Fri
## 6                 1                304            320   16 Sun
```

```
activitydistance_daily <- DailyActivity_Sleep_merged %>%
  group_by(day) %>%
  drop_na() %>%
  summarise(VeryActiveDistance = sum(VeryActiveDistance),
            ModeratelyActiveDistance = sum(ModeratelyActiveDistance),
            LightActiveDistance = sum(LightActiveDistance),
            SedentaryActiveDistance = sum(SedentaryActiveDistance))

weekday_steps <- activitydistance_daily %>%
  mutate(weekday = day)
weekday_steps$weekday <- ordered(weekday_steps$weekday, levels = c("Mon", "Tue", "Wed", "Thu", "Fri", "S
weekday_steps <- weekday_steps %>%
 group_by(weekday) %>%
 summarize(daily_distance = sum(VeryActiveDistance, ModeratelyActiveDistance, LightActiveDistance, Seden

ggplot(weekday_steps, aes(x = weekday, y = daily_distance)) +
 geom_col(fill = "#d62d58") +
 labs(title = "Daily Distance per weekday", x ="", y = "") +
 theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



Findings:

*In the observations, Thursday and Sunday are the days during the week that people tend to walk more, but less on Friday and Saturday. It is interesting to further investigate why the Thursdays cumulates longer walk distance than the rest of weekdays.

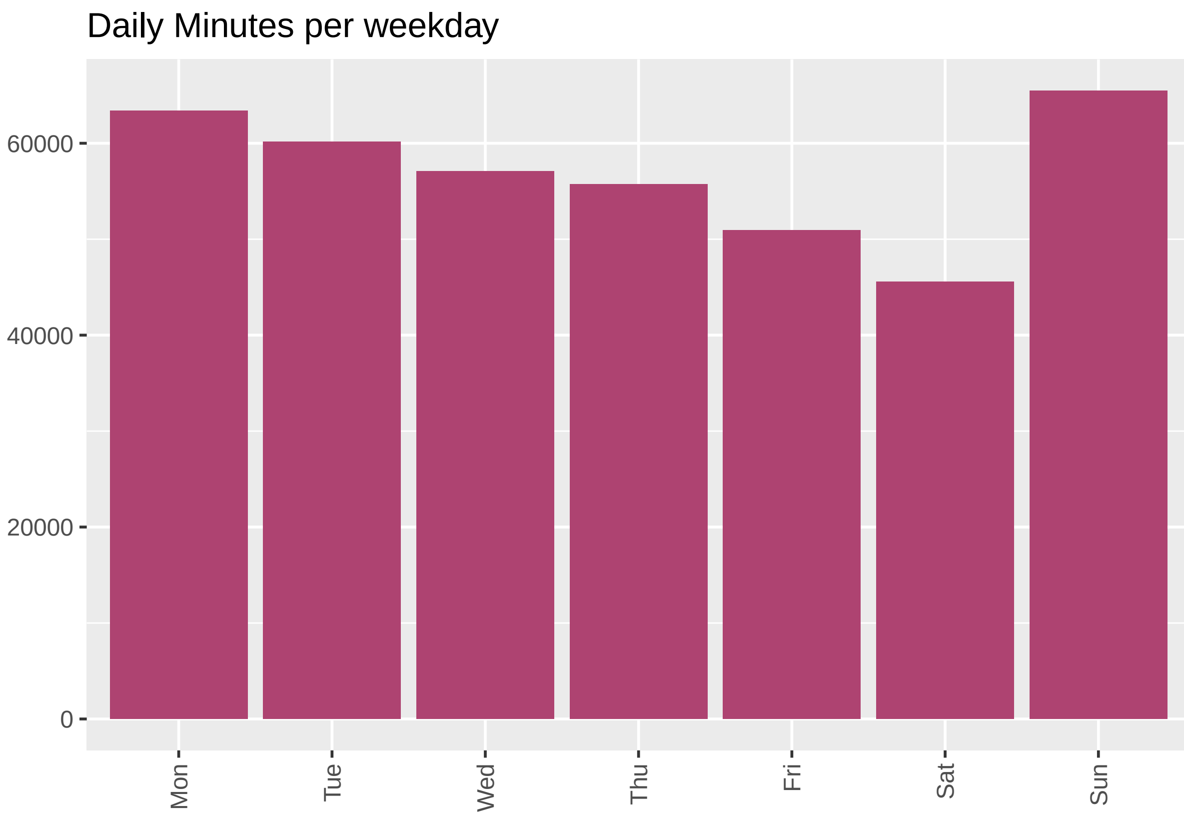
```
dailyactivity_time<- DailyActivity_Sleep_merged %>%
  group_by(day) %>%
  drop_na() %>%
  summarise(VeryActiveMinutes = sum(VeryActiveMinutes),
            FairlyActiveMinutes = sum(FairlyActiveMinutes),
            LightlyActiveMinutes = sum(LightlyActiveMinutes),
            SedentaryMinutes = sum(SedentaryMinutes))

weekday_minutes <- dailyactivity_time %>%
 mutate(weekday = day)
 weekday_minutes$weekday <- ordered(weekday_minutes$weekday, levels = c("Mon", "Tue", "Wed", "Thu", "Fr
weekday_minutes <- weekday_minutes %>%
 group_by(weekday) %>%
 summarize(daily_minutes = sum(VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes, SedentaryMi
head(weekday_minutes)
```

```
## # A tibble: 6 x 2
##   weekday daily_minutes
##   <ord>           <int>
## 1 Mon             63393
## 2 Tue             60162
## 3 Wed             57086
## 4 Thu             55717
```

```
## 5 Fri              50962
## 6 Sat              45567
```

```
ggplot(weekday_minutes, aes(x = weekday, y = daily_minutes)) +
  geom_col(fill = "#AE4371") +
  labs(title = "Daily Minutes per weekday", x ="", y = "") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



*The graph shows that, during the observation, people spend more time being active on Sunday and Monday. It displays a declining trend from Tuesday to Saturday being active.

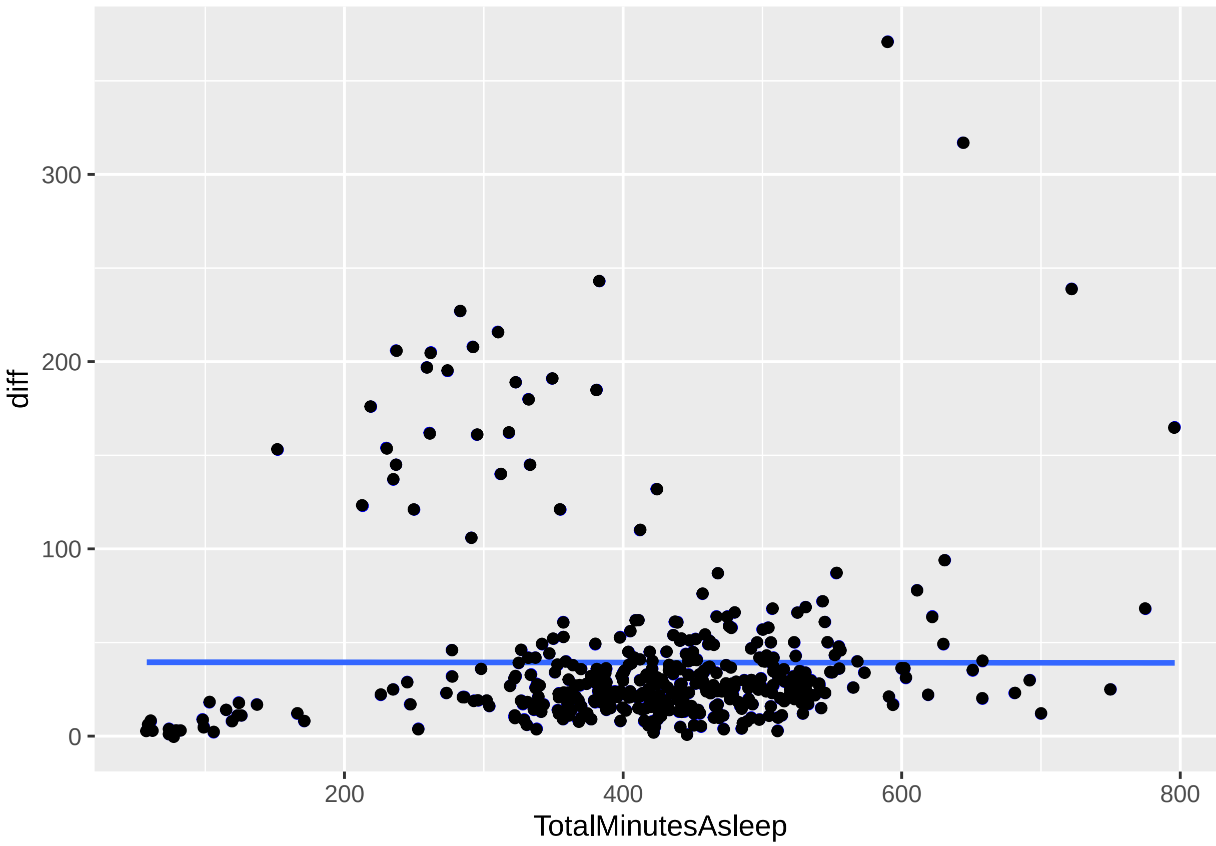
Share

More visualizations are created to share our findings.

1. Association between sleeping time, calories, and intensity distance

```
ggplot(data = DailyActivity_Sleep_merged, aes(TotalMinutesAsleep, diff)) +
 geom_point(color = "blue") + geom_smooth(method = "lm", se = FALSE) + geom_jitter()
```

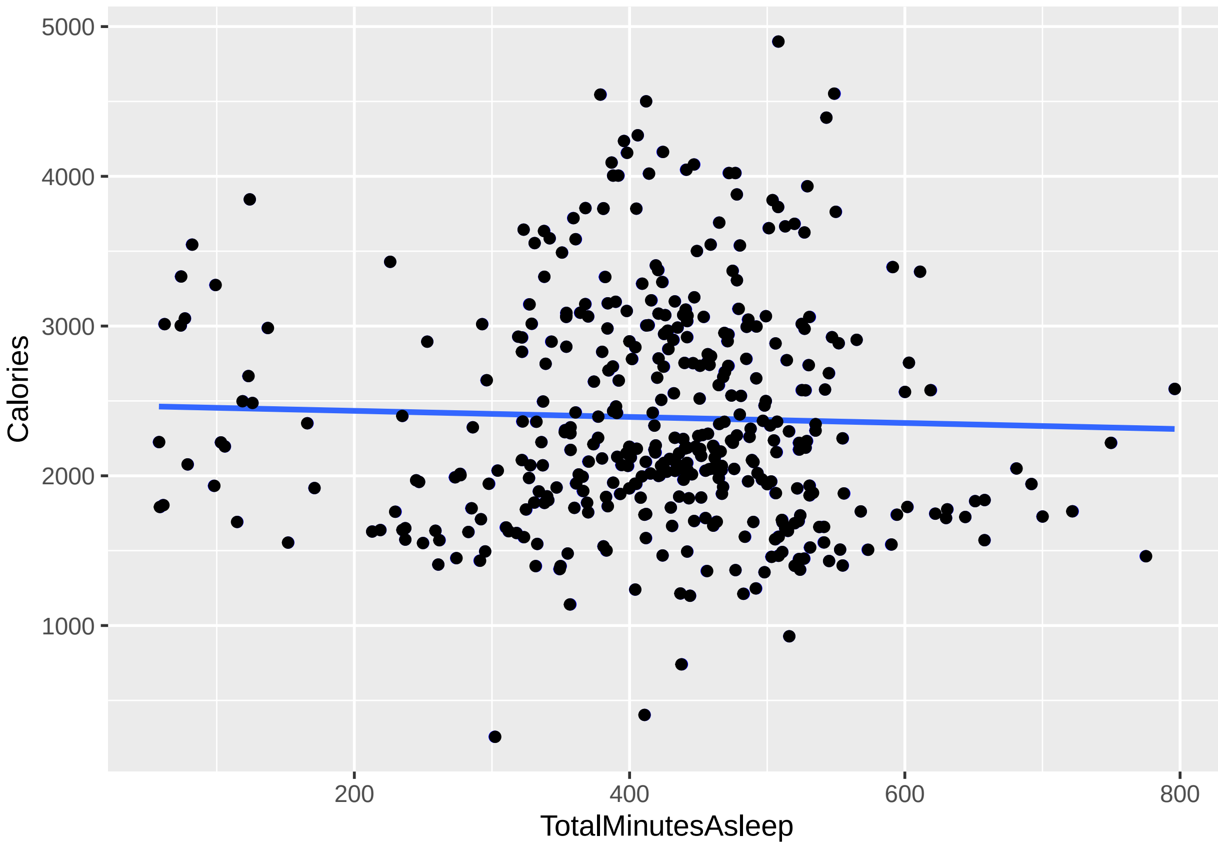
```
## `geom_smooth()` using formula = 'y ~ x'
```



It tends to have a negative weak association between daily calories burned and lounging time on bed.

```
ggplot(data = DailyActivity_Sleep_merged, aes(TotalMinutesAsleep, Calories)) +
 geom_point(color = "blue") + geom_smooth(method = "lm", se = FALSE) + geom_jitter()
```

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



It tends to have a week negative relation between daily sleeping duration and daily calories burned. The TotalAsleepTime data will be divided into groups to help further investigate the relation between sleep pattern, calories burned, and intensity.

2.Sleeping type

We will assign data into the following groups based on the length of total asleep time in a day. This categorization is based on CDC recommendations for health benefits.

Below Recommended type: less than 7 hours Recommended type: 7 ~ 9 hours Above Recommended type: more than 9 hours. Below provides more analysis that might help us understand the different trends underlying each sleep type.

```r
sleep_type <-DailyActivity_Sleep_merged %>%
 group_by(Id) %>%
 summarize(minutes_sleep = round(mean(TotalMinutesAsleep))) %>%
 mutate(sleep_type = case_when(
    minutes_sleep >=0 & minutes_sleep < 420 ~ "Below Recommended",
    minutes_sleep >=420 & minutes_sleep < 540 ~ "Recommended",
    minutes_sleep >=540 & minutes_sleep < 1000 ~ "Above Recommended"
  ))
head(sleep_type)
```

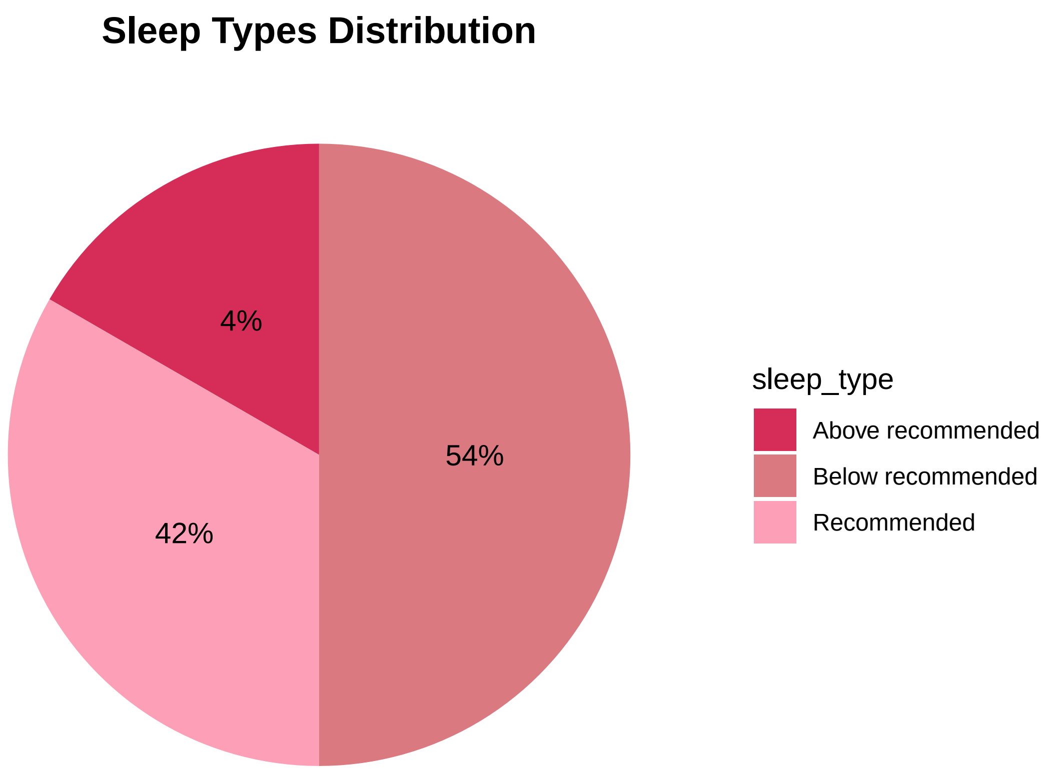
```
## # A tibble: 6 x 3
##           Id minutes_sleep sleep_type
##        <dbl>         <dbl> <chr>
## 1 1503960366           360 Below Recommended
## 2 1644430081           294 Below Recommended
## 3 1844505072           652 Above Recommended
```

```
## 4 1927972279           417 Below Recommended
## 5 2026352035           506 Recommended
## 6 2320127002            61 Below Recommended
```

```
table_sleep_type <-DailyActivity_Sleep_merged %>%
  left_join(sleep_type, by = "Id") %>%
  group_by(sleep_type) %>%
  summarise(participants = n_distinct(Id)) %>%
  mutate(perc = participants/sum(participants)) %>%
  arrange(perc) %>%
  mutate(perc = scales::percent(perc))
head(table_sleep_type)
```

```
## # A tibble: 3 x 3
##   sleep_type        participants perc
##   <chr>                    <int> <chr>
## 1 Above Recommended            1 4%
## 2 Recommended                 10 42%
## 3 Below Recommended           13 54%
```

```
table_sleep_type %>%
  ggplot(aes(x = "", y = perc, fill = sleep_type,)) +
  geom_bar(stat = "identity", width = 2) +
  coord_polar("y", start = 0) +
  theme_minimal() +
  theme(axis.title.x = element_blank(),
        axis.title.y = element_blank(),
        panel.border = element_blank(),
        panel.grid = element_blank(),
        axis.ticks = element_blank(),
        axis.text.x = element_blank(),
        plot.title = element_text(hjust = 0.5, size = 14, face = "bold")) +
  geom_text(aes(label = perc),
          position = position_stack(vjust = 0.5)) +
  scale_fill_manual(values = c("#d62d58", "#db7980", "#fc9fb7"),
                    labels = c("Above recommended",
                               "Below recommended",
                               "Recommended"))+
  labs(title = "Sleep Types Distribution")
```



Findings:

*4% of participants usually sleep MORE than 7 hours a night.

*42% of them usually sleep between 7-9 hours a night.

*54% of them usually sleep Less than 7 hours recommended.

More than half of the participants sleep less than 7 hours a night. From health benefits, it is recommended an adult should sleep between 7-9 hours a night. Adults who sleep less than 7 hours a night may have more health issues than those who sleep 7 or more hours a night. If you regularly need more than 9 hours of sleep per night to feel rested, it might be a sign of a sleep or medical problem.

3.Average daily distance, day, and sleep type

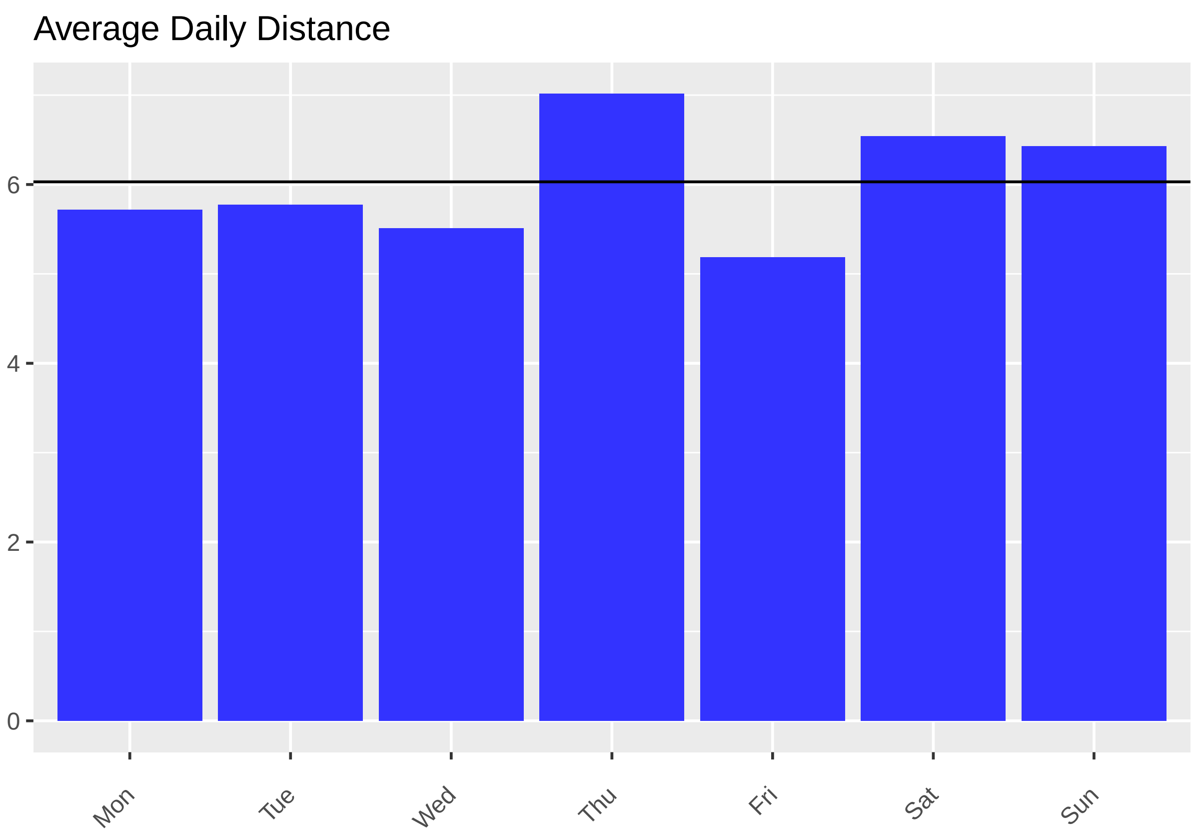
3.1 Average daily distance by day To know what day during the week participants are active the most.

```
weekday_average_distance <- DailyActivity_Sleep_merged %>%
  mutate (day)
 weekday_average_distance$day <- ordered(weekday_average_distance$day, levels = c("Mon", "Tue", "Wed","'
 weekday_average_distance <- weekday_average_distance %>%
  group_by(day) %>%
  summarize(average_daily_distance = mean(TotalDistance))
head(weekday_average_distance)
```

```
## # A tibble: 6 x 2
##   day   average_daily_distance
##   <ord>                  <dbl>
## 1 Mon                     5.72
## 2 Tue                     5.77
## 3 Wed                     5.51
## 4 Thu                     7.02
## 5 Fri                     5.18
```

```
## 6 Sat                     6.54
```

```
ggplot(weekday_average_distance, aes(day,average_daily_distance)) +
  geom_col(fill = "#3333FF") +
  geom_hline(yintercept = 6.03) +
  labs(title = "Average Daily Distance", x = "", y ="") +
  theme (axis.text.x = element_text(angle = 45, vjust = 0.5, hjust = 1))
```



Findings:

*The highest average daily activity distance is Thursday, followed by Saturday and Sunday. People might have more time to enjoy activities with friends and family and become more active.

*The lowest average daily activity distance is Friday. It might be the end of the working week and most of us just want to go home and rest.

*It will be interesting to analyze the reasons behind a very active Thursday.

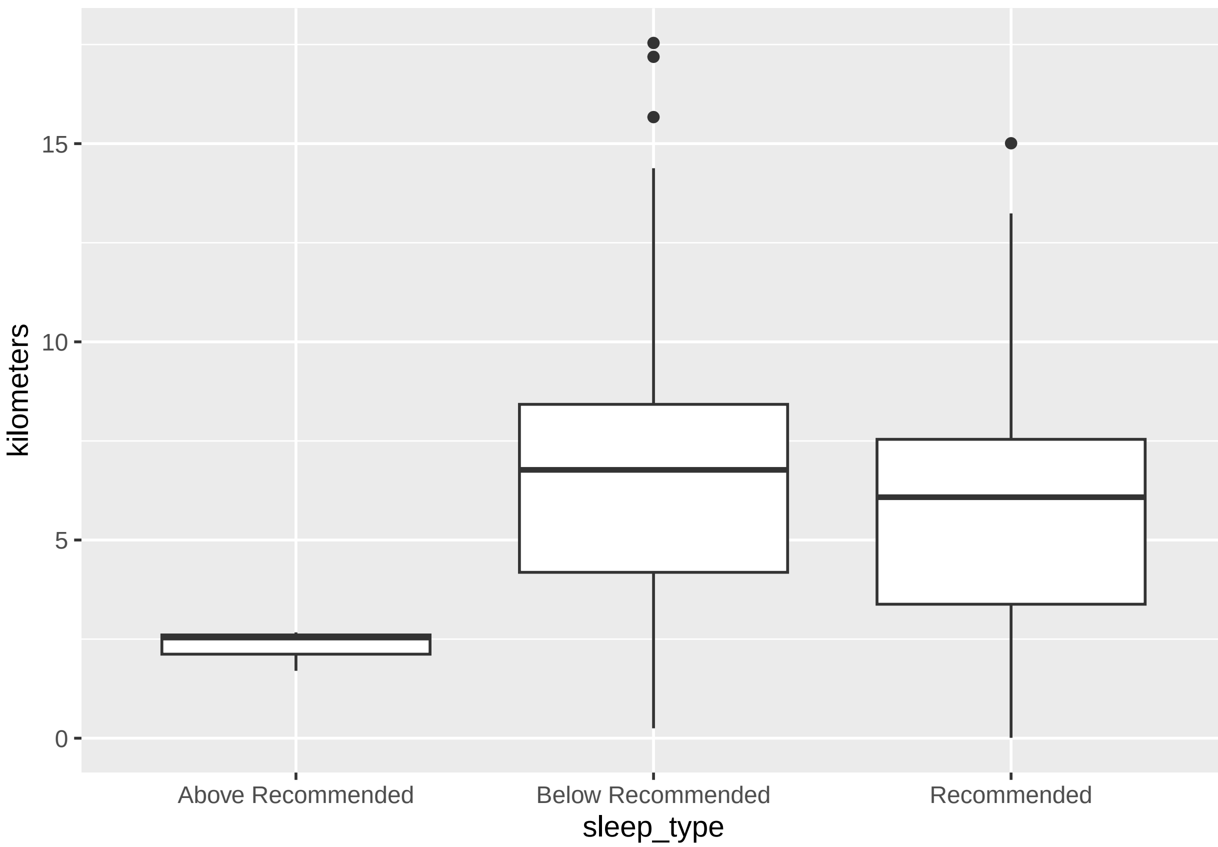
3.2 Daily distance by sleep type To know what sleep type of people travelled the most during activities.

```
daily_distance_sleep <- DailyActivity_Sleep_merged %>%
  left_join(sleep_type, by = 'Id') %>%
  group_by(day, sleep_type) %>%
  select(sleep_type, TotalDistance, day) %>%
  mutate(day = factor(day, levels = c("Mon", "Tue", "Wed", "Thu", "Fri", "Sat", "Sun")))
head(daily_distance_sleep)
```

```
## # A tibble: 6 x 3
## # Groups:   day, sleep_type [5]
##   sleep_type        TotalDistance day
##   <chr>                     <dbl> <fct>
```

```
## 1 Below Recommended          8.5  Sun
## 2 Below Recommended          6.97 Mon
## 3 Below Recommended          6.28 Wed
## 4 Below Recommended          8.16 Thu
## 5 Below Recommended          6.48 Fri
## 6 Below Recommended          9.88 Sun
```

```r
p <- ggplot(daily_distance_sleep, aes(x=sleep_type, y=TotalDistance))+
      geom_boxplot() + labs(y="kilometers")+
      theme(legend.position="none")
plot(p)
```



Findings:

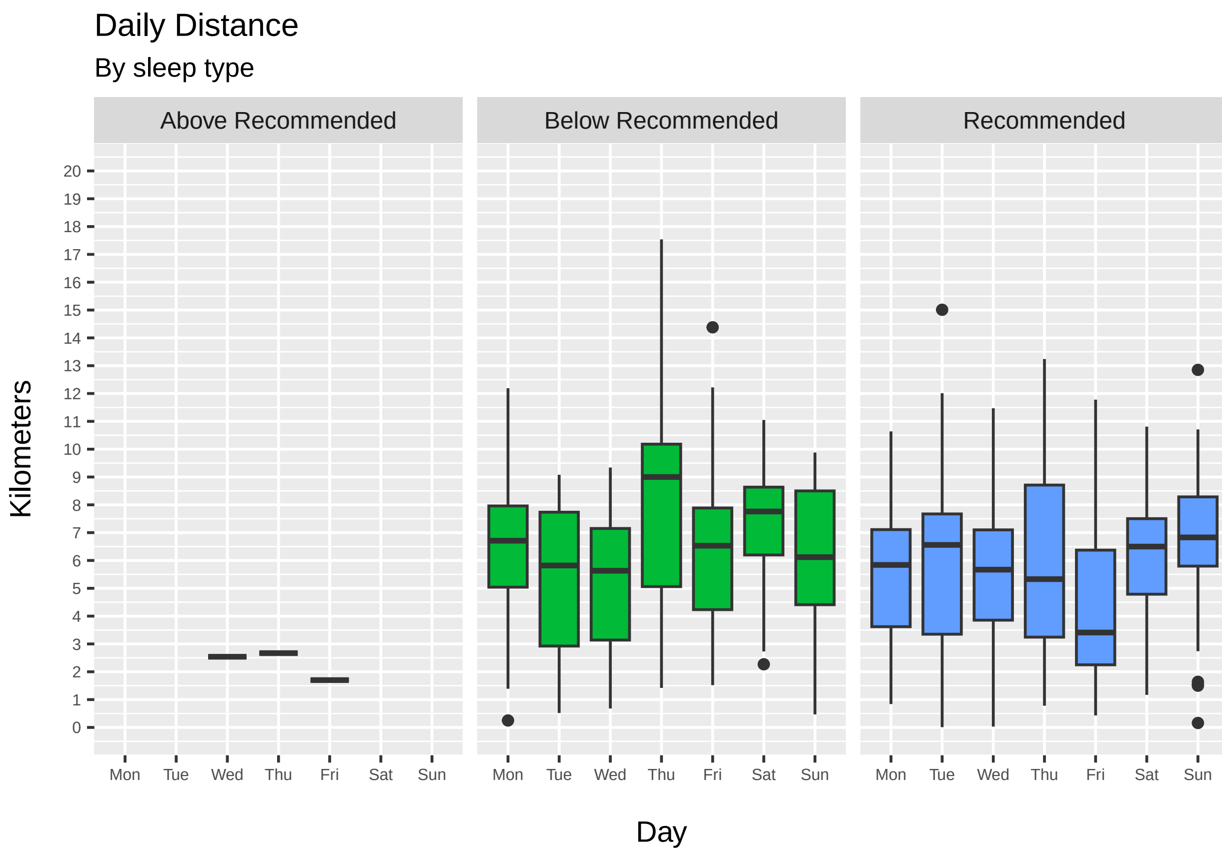
*The group of sleeping hours above recommended has a much lower median than the other two groups. This suggests that people who sleep longer than 9 hours a night tends to be less active than the ones who sleep less than 9 hours.

*The distribution for the group of sleeping hours by recommendation is similar to the group of sleeping hours below recommendation. The data of the group of the below recommendation is more spread out than the group of recommended.

3.3 Daily Distance by day, sleep type To make visualization to compare sleeping groups in terms of daily distance across a week.

```r
p1 <- ggplot(daily_distance_sleep, aes(x=day, y=TotalDistance, fill=sleep_type)) +
      geom_boxplot() +
      scale_y_continuous(breaks = seq(0, 20, by = 1), limits = c(0, 20)) +
      theme(legend.position="none", plot.title=element_text(size=11)) +
```

```
      ggtitle("A boxplot with jitter") +
      xlab("") +
      labs(title=("Daily Distance"), subtitle=("By sleep type"), x="Day", y="Kilometers") +
      theme(plot.title=element_text(size = 12,hjust = 0))+
      theme(plot.subtitle=element_text(size = 10,hjust = 0))+
      theme(axis.text.y=element_text(size=6)) +
      theme(axis.text.x=element_text(size=6,hjust= 0.5))+
      theme(axis.title.x = element_text(margin = margin(t = 14, r = 0, b = 0, l = 0)))+
      theme(axis.title.y = element_text(margin = margin(t = 0, r = 10, b = 0, l = 0)))+
      theme(legend.title=element_text(size=10))+
      theme(legend.text=element_text(size=8))+
      facet_grid(~sleep_type)
plot(p1)
```



Findings:

*It tends to have a trailing off the week from Monday to Friday among the three groups.

*The group below recommended tends to have a slightly higher median across the weekday than the rest two groups. That means, people sleeping below 7 hours a night might walk longer distance than the rest two groups.

*Also, it seems there is more spread out on Thursday on both groups of below recommended and recommended.

*Among all three groups, people seem to be involved in less active activities on Friday.

4.Daily calories, day, and sleep type

4.1 Average daily calories by day To know what day during a week participants burned calories the most.

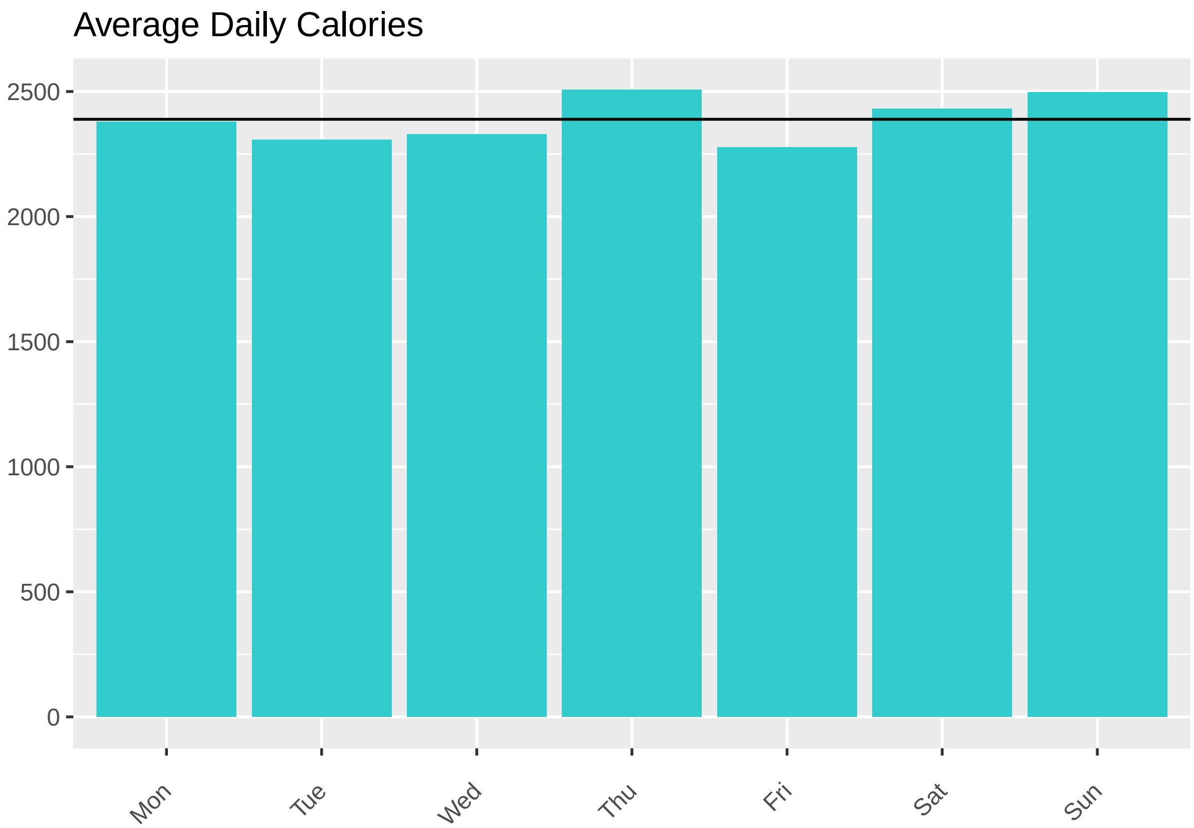
```
average_daily_calories <- DailyActivity_Sleep_merged %>%
  mutate(day)

average_daily_calories$day <- ordered(average_daily_calories$day, levels = c("Mon", "Tue", "Wed", "Thu"

average_daily_calories <- average_daily_calories %>%
  group_by(day) %>%
  summarize(weekday_calories = mean(Calories))
head(average_daily_calories)
```

```
## # A tibble: 6 x 2
##   day   weekday_calories
##   <ord>            <dbl>
## 1 Mon              2378.
## 2 Tue              2307.
## 3 Wed              2330.
## 4 Thu              2507.
## 5 Fri              2277.
## 6 Sat              2432.
```

```
ggplot(average_daily_calories, aes(day, weekday_calories)) +
  geom_col(fill = "#33CCCC") +
  geom_hline(yintercept = 2389) +
  labs(title = "Average Daily Calories", x = "", y = "") +
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5, hjust = 1))
```



Findings:

*The graph shows that participants burned calories on Thursday the most on average, followed by Sunday and Saturday. The least on Friday. This is consistent with the finding that participants are the least active on Friday.

*The weekdays that have average daily calories burned less than the mean are on Tuesday, Wednesday, and Friday.

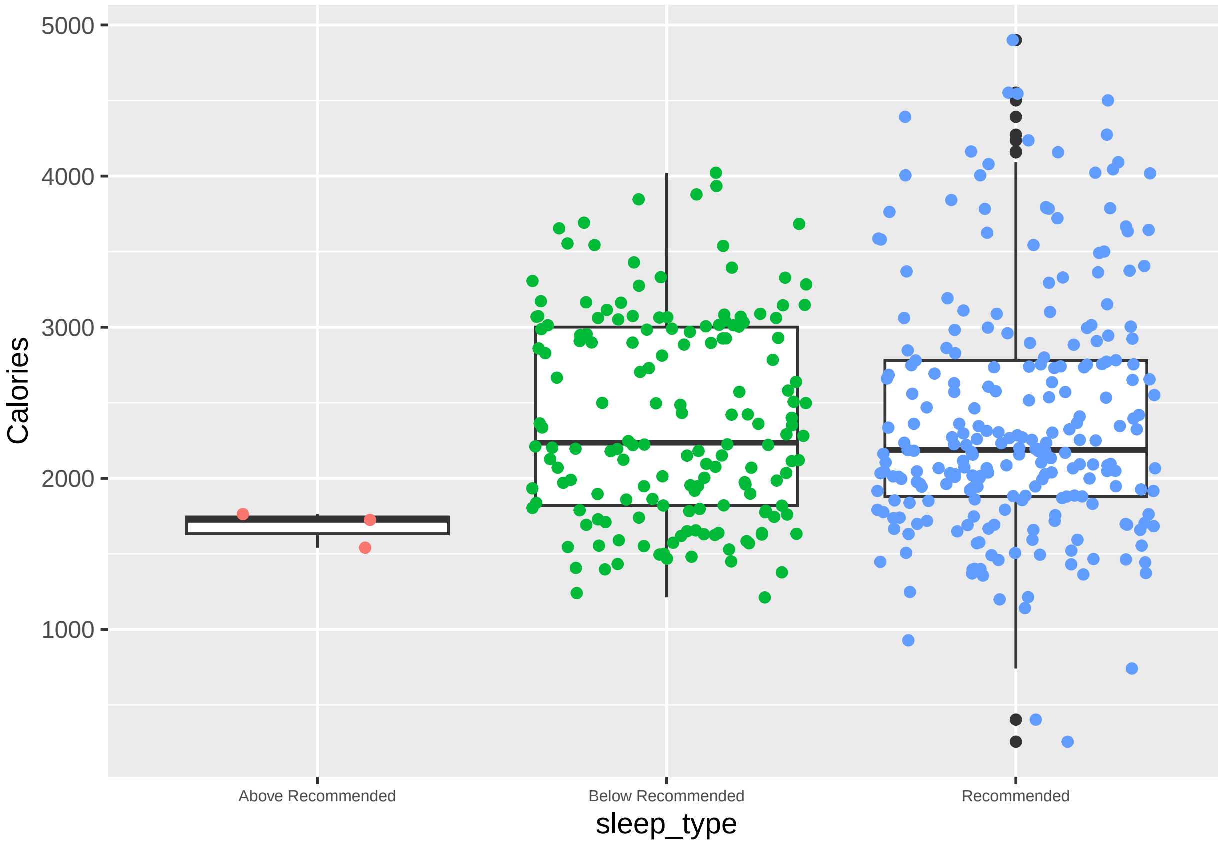
4.2 Daily calories by sleep type

To know what sleep type of people burned calories the most.

```
daily_calories_sleep <- DailyActivity_Sleep_merged %>%
  left_join(sleep_type, by = 'Id') %>%
  group_by(day, sleep_type) %>%
  select(sleep_type, Calories, day) %>%
  mutate(day = factor(day, levels = c("Mon", "Tue", "Wed", "Thu", "Fri", "Sat", "Sun")))
head(daily_calories_sleep)
```

```
## # A tibble: 6 x 3
## # Groups:   day, sleep_type [5]
##   sleep_type        Calories day
##   <chr>                <int> <fct>
## 1 Below Recommended     1985 Sun
## 2 Below Recommended     1797 Mon
## 3 Below Recommended     1745 Wed
## 4 Below Recommended     1863 Thu
## 5 Below Recommended     1728 Fri
## 6 Below Recommended     2035 Sun
```

```r
p2 <- ggplot(daily_calories_sleep, aes(x=sleep_type, y=Calories)) +
   geom_boxplot() + labs(y = "Calories") +
   geom_jitter(aes(color=sleep_type)) +
   theme(legend.position = "none") +
   theme(axis.text.x = element_text(size=6, hjust = 0.5))
plot(p2)
```



Findings:

*The group sleeping below recommended hours has a slightly higher median than the group sleeping 7-9 hours a night. That means, the group of people who sleep below 7 hours a night might take part in more activities, thus burning more calories.

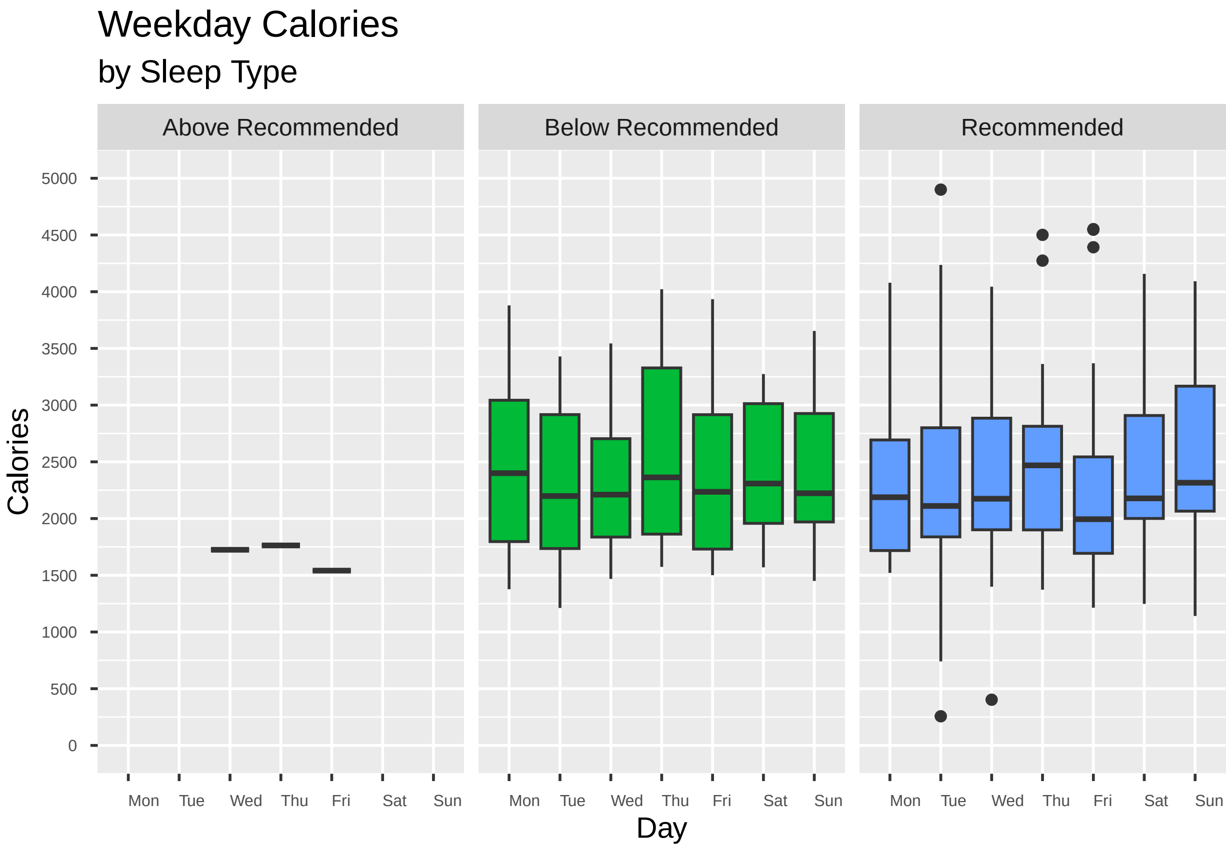
*The data of the group sleeping 7-9 hours a night more spread out than the rest two groups. Several of them are even exceeding the maximum of the group sleeping below 7 hours a night.

*The group sleeping over 9 hours a night seems to burn less calories with lowest median and less variability.

4.3 Daily calories by day, sleep type

```r
p3 <- ggplot(daily_calories_sleep, aes(x = day, y = Calories, fill = sleep_type)) +
  geom_boxplot() +
  scale_y_continuous(breaks = seq(0, 5000, by = 500), limits = c(0, 5000)) +
  theme(legend.position = "none", plot.title = element_text(size = 9)) +
  ggtitle("A boxplot with jitter") +
  xlab("") +
  labs(title = ("Weekday Calories"), subtitle = ("by Sleep Type"), x = "Day", y = "Calories") +
  theme(plot.title = element_text(size = 14, hjust = 0)) +
```

```
  theme(plot.subtitle = element_text(size = 12, hjust = 0)) +
  theme(axis.text.x = element_text(size = 6, hjust = 0)) +
  theme(axis.text.y = element_text(size = 6)) +
  theme(axis.text.x = element_text(margin = margin(t = 5, r = 1, b = 0, l = 0))) +
  theme(axis.text.y = element_text(margin = margin(t = 0, r = 5, b = 0, l = 0))) +
  theme(legend.title = element_text(size = 6)) +
  theme(legend.text = element_text(size = 6)) +
  facet_grid(~sleep_type)
plot(p3)
```



Findings:

*Both groups of sleeping below 7 hours a night and between 7-9 hours a night share a similar distribution that the median tends to go down from Monday to Friday, except Thursday, and then slightly increase over the weekend. That indicates that participants who sleep less than 9 hours a night tend to burn more calories over the weekend and may be more active.

*However, all three groups showed burning more calories on Thursday than the rest weekdays. Further study is needed to discuss the reason behind this.

5.Total active time, day and sleep type

5.1 Total active time by day Before this analysis, I add a new column that shows the total active time from four categories of active times.

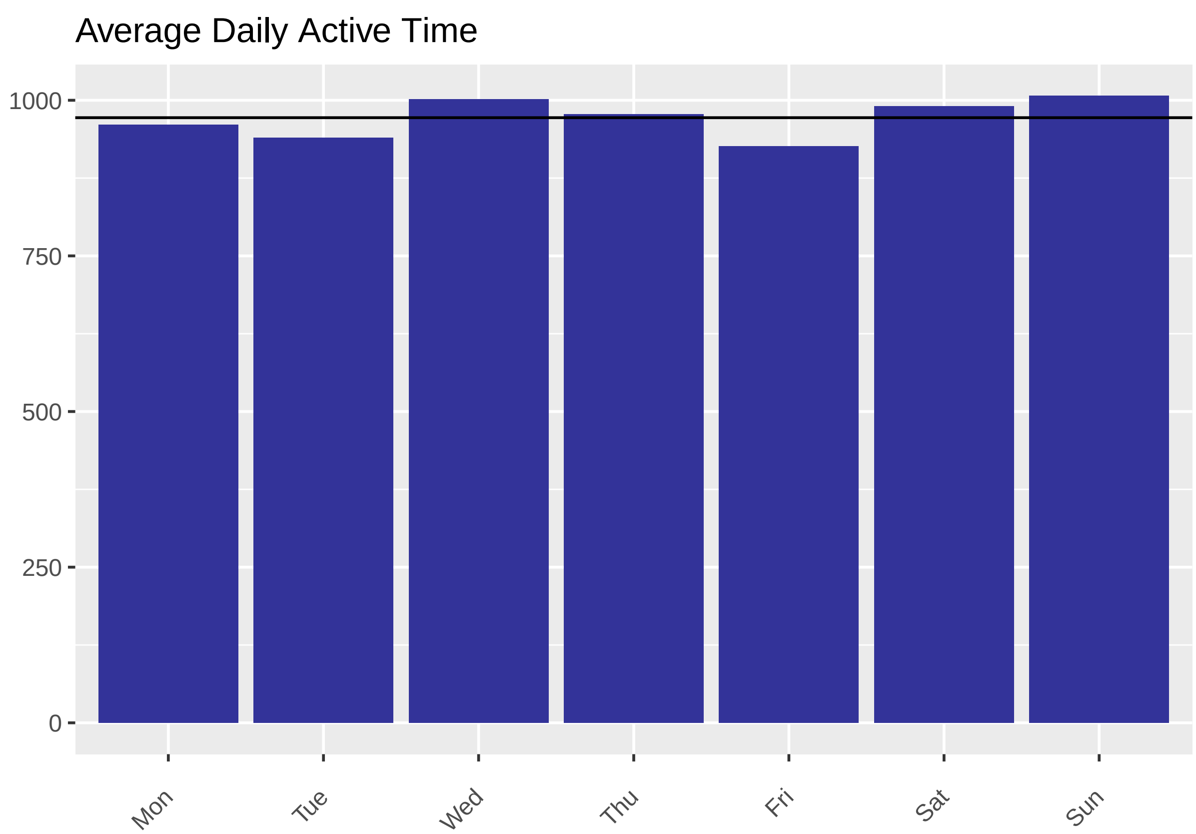
```
DailyActivity_Sleep_merged <- DailyActivity_Sleep_merged %>%
  mutate(TotalActiveTime = select(., VeryActiveMinutes:SedentaryMinutes) %>%
                                    rowSums(na.rm = TRUE))
```

```r
average_daily_time <- DailyActivity_Sleep_merged %>%
  mutate(day)
average_daily_time$day <- ordered(average_daily_time$day, levels = c("Mon", "Tue",
 "Wed", "Thu", "Fri", "Sat", "Sun"))
average_daily_time <- average_daily_time %>%
  group_by(day) %>%
  summarize(active_time = mean(TotalActiveTime))

head(average_daily_time)
```

```
## # A tibble: 6 x 2
##   day   active_time
##   <ord>       <dbl>
## 1 Mon          960.
## 2 Tue          940.
## 3 Wed         1002.
## 4 Thu          977.
## 5 Fri          927.
## 6 Sat          991.
```

```r
ggplot(average_daily_time, aes(day, active_time)) +
  geom_col(fill = "#333399") +
  geom_hline(yintercept = 972) +
  labs(title = "Average Daily Active Time", x = "", y = "") +
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5, hjust = 1))
```



Findings:

*Wednesday, Saturday, and Sunday are the days with the highest average daily active time. Participants spent at least 16.2 hours on Wednesday, Saturday, and Sunday being active on various tensity of activities.

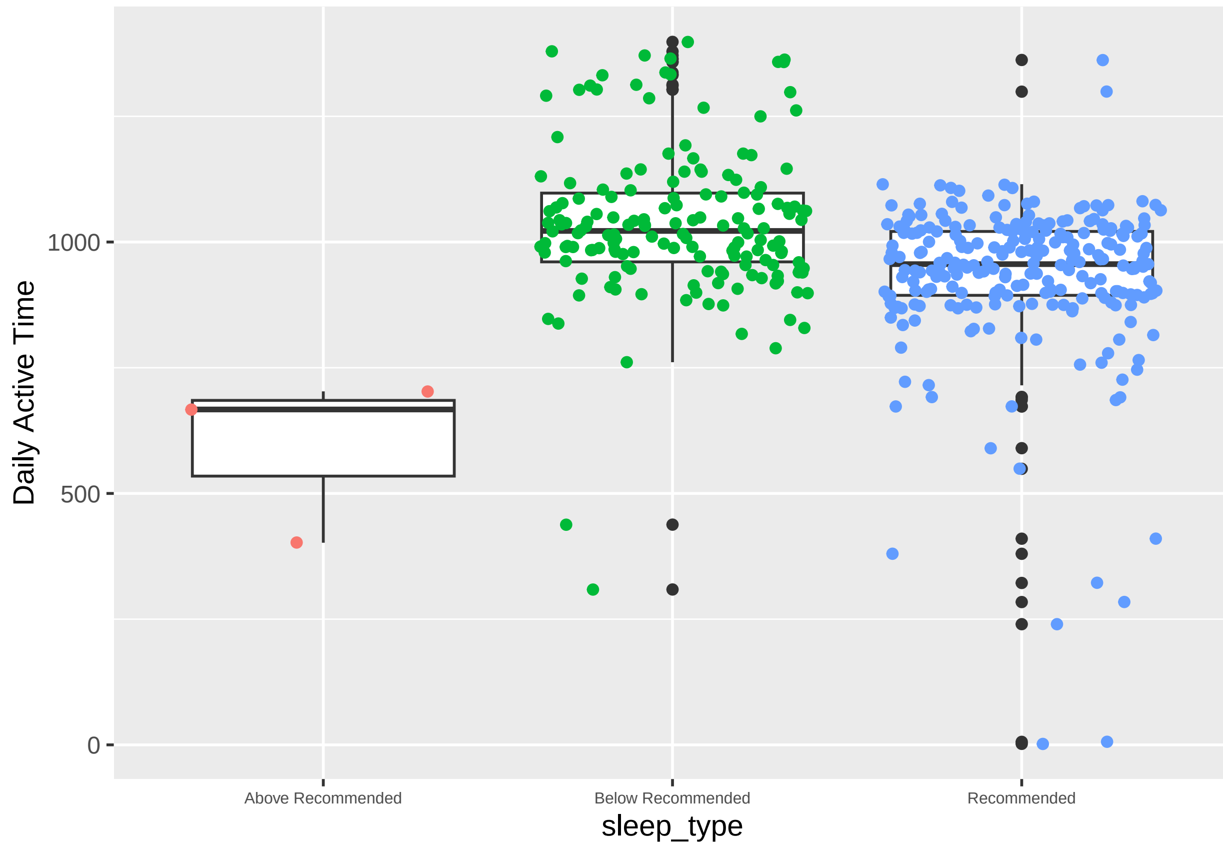
*The lowest is on Tuesday and Friday. Participants on average spent 15 hours on four different intensity levels of activities.

```r
daily_time_sleep <- DailyActivity_Sleep_merged %>%
  left_join(sleep_type, by = "Id") %>%
  group_by(day, sleep_type) %>%
  select(sleep_type, TotalActiveTime, day) %>%
  mutate(day = factor(day, levels = c("Mon", "Tue", "Wed", "Thu", "Fri", "Sat", "Sun")))
head(daily_time_sleep)
```

5.2 Total active time by sleep type

```
## # A tibble: 6 x 3
## # Groups:   day, sleep_type [5]
##   sleep_type        TotalActiveTime day
##   <chr>                       <dbl> <fct>
## 1 Below Recommended            1094 Sun
## 2 Below Recommended            1033 Mon
## 3 Below Recommended             998 Wed
## 4 Below Recommended            1040 Thu
## 5 Below Recommended             761 Fri
## 6 Below Recommended            1120 Sun
```

```r
p4 <- ggplot(daily_time_sleep, aes(x = sleep_type, y = TotalActiveTime)) +
  geom_boxplot() + labs(y = "Daily Active Time") +
  geom_jitter(aes(color = sleep_type)) +
  theme(legend.position = "none") +
  theme(axis.text.x = element_text(size = 6, hjust = 0.5))
plot(p4)
```



Findings:

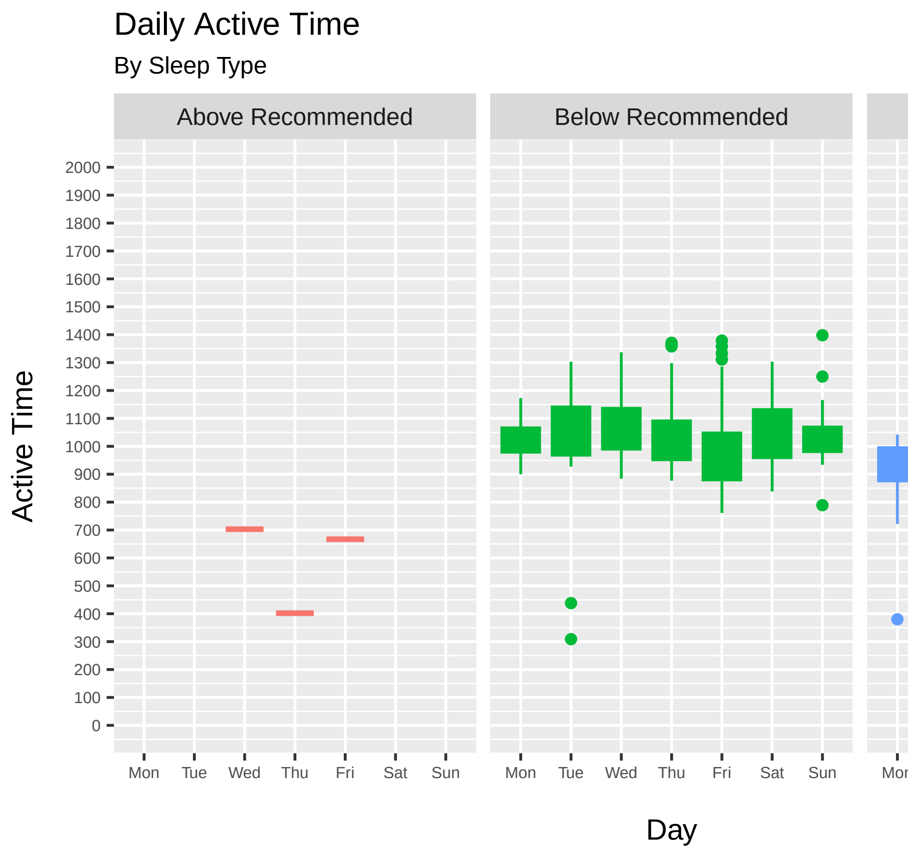
*The group sleeping less than 7 hours a night has the highest median than the other two groups. That means half of the group spend at least 1000 minutes (about 16 and a half hours) a day engaging in activities.

*However, the group sleeping 7-9 hours a night has a greater variability. Participants in this group spent time on activities in very various spectrums.

*The group sleeping more than 9 hours has the lowest median and the smallest range. This provides a indicator that participants in this group spent the least amount of time in a day on partaking activities.

```
p5 <- ggplot(daily_time_sleep, aes(x = day, y = TotalActiveTime, fill = sleep_type, colour = sleep_type)
  geom_boxplot() +
  scale_y_continuous(breaks = seq(0, 2000, by = 100), limits = c(0, 2000)) +
  theme(legend.position = "none", plot.title = element_text(size = 9)) +
  ggtitle("A boxplot with jitter") +
  xlab("") +
  labs(title = ("Daily Active Time"), subtitle = ("By Sleep Type"), x= "Day", y = "Active Time") +
  theme(plot.title = element_text(size = 12, hjust = 0)) +
  theme(plot.subtitle = element_text(size = 9, hjust = 0)) +
  theme(axis.text.x = element_text(size = 6, hjust = 0.5)) +
  theme(axis.text.y = element_text(size = 6)) +
  theme(axis.title.x = element_text(margin = margin(t= 14, r = 0, b = 0, l = 0))) +
  theme(axis.title.y = element_text(margin = margin(t = 0, r = 10, b = 0, l = 0))) +
  theme(legend.title = element_text(size = 12)) +
  theme(legend.text = element_text(size = 8)) +
  facet_grid(~sleep_type)
```

```r
plot(p5)
```



5.3 Total active time by day, sleep type

Findings:

*Overall, the group sleeping less than 7 hours a night has a higher median throughout the week than the other two groups as well as a greater variability. This suggests that people who sleep less than 7 hours a night spend more time daily of a week joining in different levels of intensity activities.

*Also, all three groups show a trend that daily active time goes down from Monday to Friday and slightly move back up over the weekend. This might not be surprising. It's a resting day and people intend to go out to hang out with friends and/or family, taking care of chores.

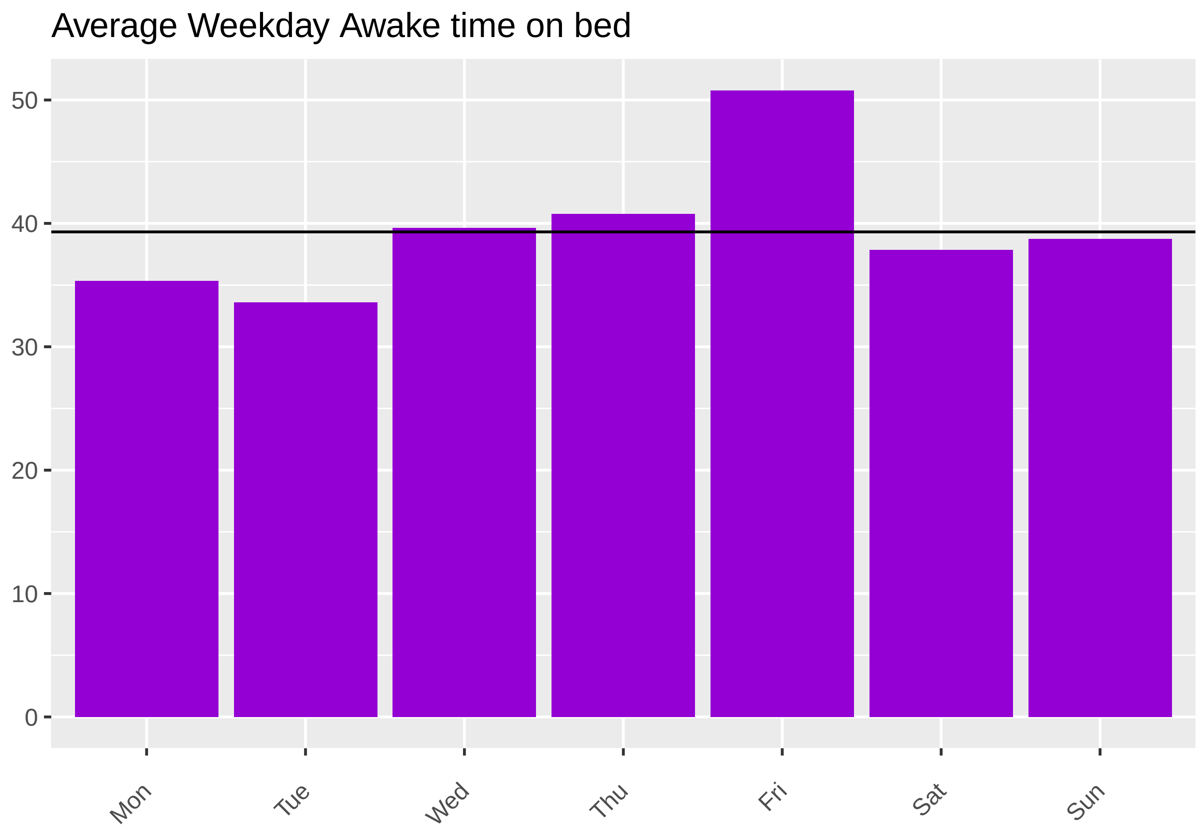
6.Awake time, day, and sleep type

Awake time is the difference between the total asleep time and total time in bed. We will investigate how sleep type relates to awake time on bed across the week.

6.1 Awake time by day To know how participants stay in bed over a week after sleeping.

```r
average_awake <- DailyActivity_Sleep_merged %>%
  mutate(day)
average_awake$day <- ordered(average_awake$day, levels = c("Mon", "Tue", "Wed","Thu", "Fri", "Sat", "Su
average_awake <- average_awake %>%
  group_by(day) %>%
  summarise(dailyawaketime = mean(diff))
ggplot(average_awake, aes(day, dailyawaketime))+
  geom_col(fill = "#9400D3") +
  geom_hline(yintercept = 39.31) +
```

```
  labs(title = "Average Weekday Awake time on bed", x = "", y = "") +
  theme(axis.text.x = element_text(angle =45, vjust = 0.5, hjust = 1))
```



Findings:

*The weekday that participants lounged on bed the longest after waking up from sleeping is on Friday.

*Participants stay awake in bed on Monday, Tuesday, and Saturday shorter than the rest weekdays.

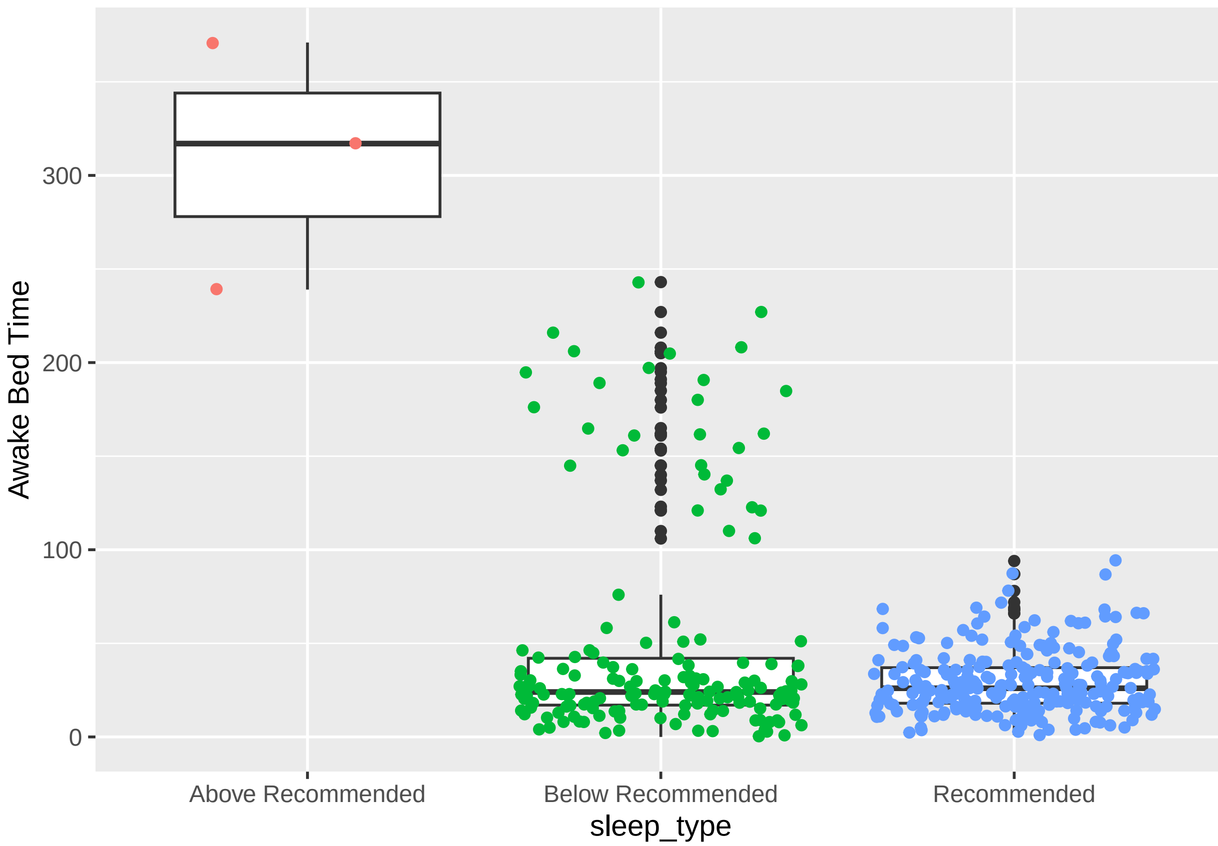
```
awake <- DailyActivity_Sleep_merged %>%
  left_join(sleep_type, by ='Id') %>%
  group_by(day, sleep_type) %>%
  select(sleep_type, diff, day) %>%
  mutate(day = factor(day, levels = c("Mon", "Tue", "Wed", "Thu", "Fri", "Sat", "Sun")))
head(awake)
```

6.2 Awake time by sleep types

```
## # A tibble: 6 x 3
## # Groups:   day, sleep_type [5]
##   sleep_type         diff day
##   <chr>             <int> <fct>
## 1 Below Recommended    19 Sun
## 2 Below Recommended    23 Mon
## 3 Below Recommended    30 Wed
## 4 Below Recommended    27 Thu
## 5 Below Recommended    12 Fri
## 6 Below Recommended    16 Sun
```

```
p6 <- ggplot(awake, aes(x = sleep_type, y = diff)) +
  geom_boxplot() + labs (y = "Awake Bed Time") +
  geom_jitter(aes(color = sleep_type)) +
  theme(legend.position = "none") +
  theme(asix.text.x = element_text(size = 6, hjust = 0.5))
plot(p6)
```

```
## Warning in plot_theme(plot): The `asix.text.x` theme element is not defined in
## the element hierarchy.
```



Findings:

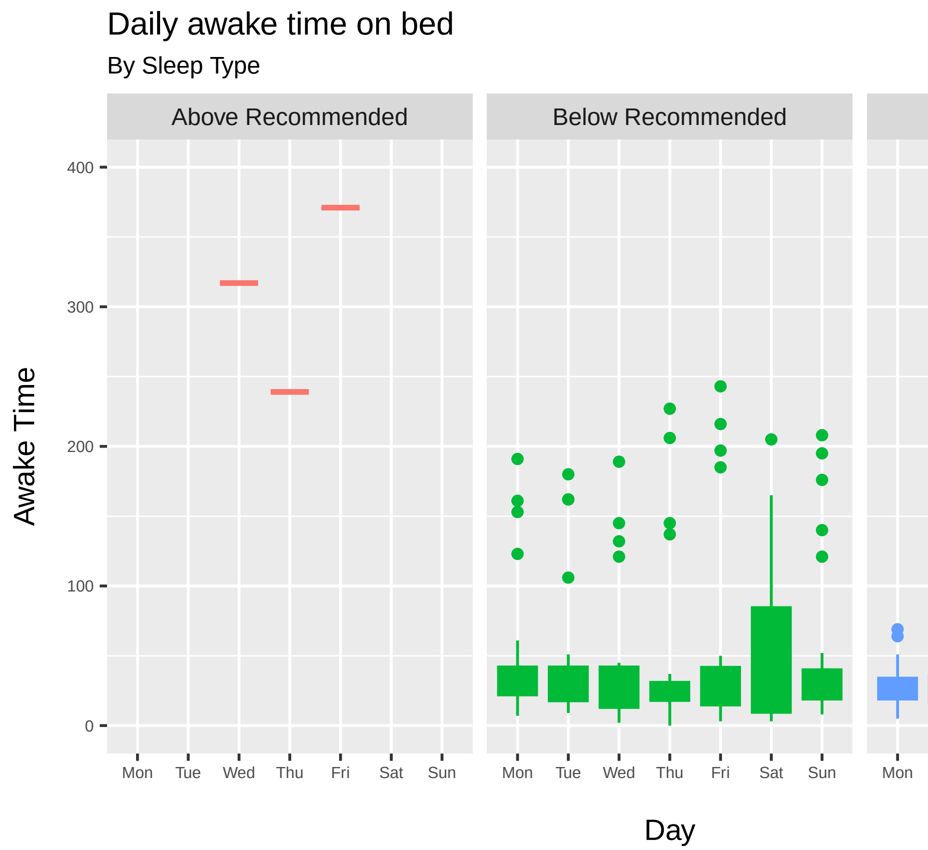
*The group sleeping more than 9 hours a night shows a greater median, which means they stay in bed longer after waking up from sleeping than the other two groups.

*The majority of groups sleep 7-9 hours or below a night is below 100 minutes. The means that those participants will stay in bed about one hour and half when not asleep.

*Some people who sleep below recommended might stay in bed longer than 100 minutes, but less than 250 minutes (about 4 hours).

```
p7 <- ggplot(awake, aes(x = day, y = diff, fill = sleep_type, colour = sleep_type))+
     geom_boxplot() +
     scale_y_continuous(breaks = seq(0, 400, by = 100), limits = c(0, 400)) +
     theme(legend.position = "none", plot.title = element_text(size = 9)) +
     ggtitle("A boxplot with jitter") +
     xlab("") +
```

```
    labs(title = ("Daily awake time on bed"), subtitle = ("By Sleep Type"), x= "Day", y = "Awake Time")
    theme(plot.title = element_text(size = 12, hjust = 0)) +
    theme(plot.subtitle = element_text(size = 9, hjust = 0)) +
    theme(axis.text.x = element_text(size = 6, hjust = 0.5)) +
    theme(axis.text.y = element_text(size = 6)) +
    theme(axis.title.x = element_text(margin = margin(t= 14, r = 0, b = 0, l = 0))) +
    theme(axis.title.y = element_text(margin = margin(t = 0, r = 10, b = 0, l = 0))) +
    theme(legend.title = element_text(size = 12)) +
    theme(legend.text = element_text(size = 8)) +
    facet_grid(~sleep_type)
plot(p7)
```



6.3 Awake time by day, by sleep types

Findings:

*The distributions of groups sleeping 7-9 hours a night and sleeping below 7 hours a night is close to each other. Their daily median seems to be lower than 50 minutes through the week.

*But the group sleeping below 7 hours a night seems to have more outlier on daily awake time than the group sleeping 7-9 hours a night. This implies it is more consistent that people who sleep 7-9 hours a night would not stay in bed long after they have a overnight sleeping and might get up to take part in activities that require more energy expenditures.

Act

Conclusion

*Bellabeat is a high-tech company that focus on health products for woman and collect health data to empower them with knowledge about their health and habits. The goal of this project is to analyze data from Fitbit smart device users and, ultimately, to use the results to inform the marketing strategies for the company's next global expansion. Below are the conclusions from this project:

*Overall, participants would travel (walk or move around) more distance and spend more time on activities on Monday and during the weekend. But the trend is trailing off from Tuesday to Friday.

*54% of participants sleep less than 7 hours a night, 42% of them between 7- 9 hours, and 4% is above 9 hours. Those who sleep less than 7 hours a night tends to travel more distance in kilometers daily, but those who sleep between 7-9 hours a night tends to burn more calories daily. This implies people who sleep between 7-9 hours a night engage in more active activities than lightly active or sedentary.

*For participants who sleep less than 7 hours a night, the pattern of daily active time spent is like the one from participants who sleep between 7 and 9 hours a night.

*The participants who sleep more than 9 hours a night tends to have a pattern of lowest daily distance travelled, daily calories burned, and shortest daily active time. This might correspond to the finding that this group seems to have daily longest awake time in bed. This might indicate that those participants sleeping more than 9 hours a night join mostly lightly active or sedentary distance and minutes. In addition, some of participants who sleep less than 7 hours a night would choose to lounge longer in bed than those sleep by recommended.

*Limitation is noted in this project. We have a small sample size, and datasets can be biased since demographic information is lacking.

Recommendations

*The Bellabeat company aims to empower woman with knowledge about their own health and habits. Regardless of ages, getting enough sleep and getting active have been recommended to stay healthy for woman. Here are my recommendations for the marketing strategy to expand globally.

*Bellabeat app. should put out more notifications or reminders to encourage more engagement in active activities during the weekdays, emphasizing consistently getting active throughout a week could get a better result to build a healthy habit and lifestyle.

*Bellabeat app. should allow users to flexibly personalize their settings on their smart devices based on their health information (sleeping status, diet restriction, illness, and etc.) and needs (such as stay healthy, lose weight, or others), and provide real-time feedback to users based on their personalization.

*The company should consider cooperating with devices that have AI functions (such as ALEXA) that users can set up alarm for or schedule a lightly workout or stretch in bed before or after sleeping since people tends to stay awake in bed almost 40 minutes daily.

*Bellabeat company should build a communication system between smart devices and their users. This system can be delivered by email, text, or other means so that a daily or weekly summary will be sent to users in terms of activities engaged, calories burned, sleeping status, and etc. Suggestions or tips will be followed at the end of the summary to remind users to maintain a healthy habit or behaviors.

*Marketing could deliver that Bellabeat smart devices are more than just a fitness tracker. It provides a means to depict the quality of a woman's life and to motivate and educate them to become better and healthier themselves.

*More further studies are needed with datasets from Bellabeat smart device users. The current project is analyzed with Fitbit datasets that target audiences are more than just woman. With Bellabeat's very own datasets, we can analyze patterns of womans' usage on smart devices for wellness purposes and compare them with Fitbit users to explore similarities and differences. From that, we can make suggestions and improve Bellabest smart devices so that we can more holistically take care of woman's wellness.

Resource:

https://www.cdc.gov/healthequity/features/nwhw/index.html https://www.medicalnewstoday.com/articles/how-many-steps-should-you-take-a-day#for-general-health

https://www.nytimes.com/2024/02/17/well/bed-rotting-hurkle-durkle.html

https://www.nbcnews.com/better/pop-culture/make-your-day-better-stay-bed-longer-really-ncna837176

https://www.nhlbi.nih.gov/health/sleep/how-much-sleep#:~:text=Experts%20recommend%20that%20adults%20sleep,or%20more%20hours%20a%20night

A special thank you to the following authors that have completed their Bellabeat capstone project and share their work. Their work provides a light to guide me through the tunel of navigating data analysis as a first-time R programming user.

https://www.kaggle.com/code/zulkhaireesulaiman/bellabeat-capstone-project-in-r/notebook

https://medium.com/@shogbaikeadeola/google-data-analytics-capstone-project-bellabeat-case-study-48431571702