Bellabeat Project

Rogelio C. Alonso

2023-03-06

Bellabeat

This report consists in four parts, and it used the techniques and phases that Google Data Analyst Course recommend, as follow:

- **Introduction:** A brief description about the company, and summarize the report's purposes, scenario, and background.
- **Problems:** After knowing the concerns of the stakeholders, problems are defined, then, thing the best way of approach and analyze them through Ask, Prepare, Process and Analyze Data Phases.
- **Solutions:** After finding answers through the thinking process, I'll give some advice using the Data Analyst Phases, Share and Act.
- **Conclusion:** Include key takeaways, solutions, and expectations for this project.

1. Introduction

1.1 Objective.

This report tries to figure out trends, patterns, and insights, from the usage of smart devices in women users, and how Bellabeat Company, may use this finding to improve their marketing strategy to keep and reach more customers.

1.2 About the company.

Bellabeat, is a high-tech company that manufactures health-focused smart products. It's a successful small company, but they have the potential to become a larger player in the global smart device market. One of the cofounder and Chief Creative Officer of Bellabeat, believes that analyzing smart device fitness data could help unlock new growth opportunities for the company. Bellabeat has grown rapidly and quickly positioned itself as a tech-driven wellness company for women.

Bellabeat want to focus on a Bellabeat product and analyze smart device usage data in order to gain insight into how people are already using their smart devices. Then, using this information, They would like high-level recommendations for how these trends can inform Bellabeat marketing strategy.

2. Deliverables and Business Tasks.

Deliverable:

1. A clear summary of the business tasks.
2. A description of all data sources used.
3. Documentation of any cleaning or manipulation of data.
4. A summary of your analysis.
5. A supporting visualizations and key findings.
6. Your top high-level content recommendations based on your analysis.

Business tasks:

1. What are some trends in smart device usage?
2. How could these trends apply to Bellabeat customer?
3. How could these trends help influence Bellabeat marketing strategy?

3. The Data.

This dataset generated by respondents to a distributed survey via Amazon Mechanical Turk between march and may, both months of 2016. Thirty eligible Fitbit users consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. (The dataset used for this case, is publicly available at https://www.kaggle.com/arashnic/fitbit.)

Observations:

We only have 30 user's record, that's indicate the limitation of data.

Also, the data was recorded in 2016, which means the data is outdated, that's why this analysis need to be considered before applying into these days.

4. The Process.

As already mentioned in previous paragraphs, the data is publicity available, at https://www.kaggle.com/arashnic/fitbit, after downloaded it and unzipped the file, it contained 18 csv files.

First, I checked all the csv files in Excel and saw that the other files were basically giving me the same information, but just structured differently, then I cleaned it, remove duplicates, checked for null cells, eliminated empty columns and in two files (H_Calories and H_Intensities)I created two columns, Date and Time.

Then, I used R program, it is good to say that this project accomplishes in R programming.

To access the data, first I loaded the necessary packages used in this project (tidyverse, tidyr, dplyr, lubridate and ggplot2).

```
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.2.3

## Warning: package 'ggplot2' was built under R version 4.2.3

## Warning: package 'tibble' was built under R version 4.2.3

## Warning: package 'dplyr' was built under R version 4.2.3

## Warning: package 'lubridate' was built under R version 4.2.3

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.
0.0 ──
## ✓ dplyr     1.1.2     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.2     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.1
## ── Conflicts ──────────────────────────────────────── tidyverse_conflict
s() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force
all conflicts to become errors

library(lubridate)
library(dplyr)
library(ggplot2)
library(tidyr)
library(hms)

## Warning: package 'hms' was built under R version 4.2.3

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

Now, after loading the packages, I will import the Fitbit Fitness Tracker Data, into R program to work with those files.

```
D_Activity <- read.csv("C:\\Users\\Drac_\\OneDrive\\Desktop\\Bellabeat_Used_F
iles\\Files_for_Analysis\\dailyActivity_merged.csv")
D_Calories <- read.csv("C:\\Users\\Drac_\\OneDrive\\Desktop\\Bellabeat_Used_F
iles\\Files_for_Analysis\\dailyCalories_merged.csv")
```

```
D_Intensities <- read.csv("C:\\Users\\Drac_\\OneDrive\\Desktop\\Bellabeat_Use
d_Files\\Files_for_Analysis\\dailyIntensities_merged.csv")
D_Steps <- read.csv("C:\\Users\\Drac_\\OneDrive\\Desktop\\Bellabeat_Used_File
s\\Files_for_Analysis\\dailySteps_merged.csv")
H_Calories <- read.csv("C:\\Users\\Drac_\\OneDrive\\Desktop\\Bellabeat_Used_F
iles\\Files_for_Analysis\\hourlyCalories_cleaned.csv")
H_Intensities <- read.csv("C:\\Users\\Drac_\\OneDrive\\Desktop\\Bellabeat_Use
d_Files\\Files_for_Analysis\\hourlyIntensities_cleaned.csv")
Sleep_Day <- read.csv("C:\\Users\\Drac_\\OneDrive\\Desktop\\Bellabeat_Used_Fi
les\\Files_for_Analysis\\sleepDay_merged.csv")
```

First, I am focusing on view the data, using head() and str() function to explore and analyze the data.

```
head(D_Activity)

##           Id ActivityDate TotalSteps TotalDistance TrackerDistance
## 1 1503960366    4/12/2016      13162          8.50            8.50
## 2 1503960366    4/13/2016      10735          6.97            6.97
## 3 1503960366    4/14/2016      10460          6.74            6.74
## 4 1503960366    4/15/2016       9762          6.28            6.28
## 5 1503960366    4/16/2016      12669          8.16            8.16
## 6 1503960366    4/17/2016       9705          6.48            6.48
##   LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance
## 1                        0               1.88                     0.55
## 2                        0               1.57                     0.69
## 3                        0               2.44                     0.40
## 4                        0               2.14                     1.26
## 5                        0               2.71                     0.41
## 6                        0               3.19                     0.78
##   LightActiveDistance SedentaryActiveDistance VeryActiveMinutes
## 1                6.06                       0                25
## 2                4.71                       0                21
## 3                3.91                       0                30
## 4                2.83                       0                29
## 5                5.04                       0                36
## 6                2.51                       0                38
##   FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories
## 1                  13                  328              728     1985
## 2                  19                  217              776     1797
## 3                  11                  181             1218     1776
## 4                  34                  209              726     1745
## 5                  10                  221              773     1863
## 6                  20                  164              539     1728

head(D_Calories)
```

```
##           Id ActivityDay Calories
## 1 1503960366   4/12/2016     1985
## 2 1503960366   4/13/2016     1797
## 3 1503960366   4/14/2016     1776
## 4 1503960366   4/15/2016     1745
## 5 1503960366   4/16/2016     1863
## 6 1503960366   4/17/2016     1728
```

```
head(D_Intensities)
```

```
##           Id ActivityDay SedentaryMinutes LightlyActiveMinutes
## 1 1503960366   4/12/2016              728                  328
## 2 1503960366   4/13/2016              776                  217
## 3 1503960366   4/14/2016             1218                  181
## 4 1503960366   4/15/2016              726                  209
## 5 1503960366   4/16/2016              773                  221
## 6 1503960366   4/17/2016              539                  164
##   FairlyActiveMinutes VeryActiveMinutes SedentaryActiveDistance
## 1                  13                25                       0
## 2                  19                21                       0
## 3                  11                30                       0
## 4                  34                29                       0
## 5                  10                36                       0
## 6                  20                38                       0
##   LightActiveDistance ModeratelyActiveDistance VeryActiveDistance
## 1                6.06                     0.55               1.88
## 2                4.71                     0.69               1.57
## 3                3.91                     0.40               2.44
## 4                2.83                     1.26               2.14
## 5                5.04                     0.41               2.71
## 6                2.51                     0.78               3.19
```

```
head(D_Steps)
```

```
##           Id ActivityDay StepTotal
## 1 1503960366   4/12/2016     13162
## 2 1503960366   4/13/2016     10735
## 3 1503960366   4/14/2016     10460
## 4 1503960366   4/15/2016      9762
## 5 1503960366   4/16/2016     12669
## 6 1503960366   4/17/2016      9705
```

```
head(H_Calories)
```

```
##           Id   ActivityHour Calories      Date    Time
## 1 1503960366 4/12/2016 0:00       81 4/12/2016 0:00:00
## 2 1503960366 4/12/2016 1:00       61 4/12/2016 1:00:00
## 3 1503960366 4/12/2016 2:00       59 4/12/2016 2:00:00
## 4 1503960366 4/12/2016 3:00       47 4/12/2016 3:00:00
## 5 1503960366 4/12/2016 4:00       48 4/12/2016 4:00:00
## 6 1503960366 4/12/2016 5:00       48 4/12/2016 5:00:00
```

```
head(H_Intensities)
```

```
##            Id   ActivityHour TotalIntensity AverageIntensity      Date    T
ime
## 1 1503960366 4/12/2016 0:00             20         0.333333 4/12/2016 0:00
:00
## 2 1503960366 4/12/2016 1:00              8         0.133333 4/12/2016 1:00
:00
## 3 1503960366 4/12/2016 2:00              7         0.116667 4/12/2016 2:00
:00
## 4 1503960366 4/12/2016 3:00              0         0.000000 4/12/2016 3:00
:00
## 5 1503960366 4/12/2016 4:00              0         0.000000 4/12/2016 4:00
:00
## 6 1503960366 4/12/2016 5:00              0         0.000000 4/12/2016 5:00
:00
```

```
head(Sleep_Day)
```

```
##            Id       SleepDay TotalSleepRecords TotalMinutesAsleep TotalTime
InBed
## 1 1503960366 4/12/2016 0:00                 1                327
346
## 2 1503960366 4/13/2016 0:00                 2                384
407
## 3 1503960366 4/15/2016 0:00                 1                412
442
## 4 1503960366 4/16/2016 0:00                 2                340
367
## 5 1503960366 4/17/2016 0:00                 1                700
712
## 6 1503960366 4/19/2016 0:00                 1                304
320
```

```
str(D_Activity)
```

```
## '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" "4/1
5/2016" ...
##  $ 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 83
8 ...
##  $ Calories                : int  1985 1797 1776 1745 1863 1728 1921 2035
1786 1775 ...
```

```
str(D_Calories)
```

```
## 'data.frame':    940 obs. of  3 variables:
##  $ Id         : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ ActivityDay: chr  "4/12/2016" "4/13/2016" "4/14/2016" "4/15/2016" ...
##  $ Calories   : int  1985 1797 1776 1745 1863 1728 1921 2035 1786 1775 ...
```

```
str(D_Intensities)
```

```
## 'data.frame':    940 obs. of  10 variables:
##  $ Id                      : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09
...
##  $ ActivityDay             : chr  "4/12/2016" "4/13/2016" "4/14/2016" "4/1
5/2016" ...
##  $ SedentaryMinutes        : int  728 776 1218 726 773 539 1149 775 818 83
8 ...
##  $ LightlyActiveMinutes    : int  328 217 181 209 221 164 233 264 205 211
...
##  $ FairlyActiveMinutes     : int  13 19 11 34 10 20 16 31 12 8 ...
##  $ VeryActiveMinutes       : int  25 21 30 29 36 38 42 50 28 19 ...
##  $ SedentaryActiveDistance : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LightActiveDistance     : num  6.06 4.71 3.91 2.83 5.04 ...
##  $ ModeratelyActiveDistance: num  0.55 0.69 0.4 1.26 0.41 ...
##  $ VeryActiveDistance      : num  1.88 1.57 2.44 2.14 2.71 ...
```

```
str(D_Steps)
```

```
## 'data.frame':    940 obs. of  3 variables:
##  $ Id         : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ ActivityDay: chr  "4/12/2016" "4/13/2016" "4/14/2016" "4/15/2016" ...
##  $ StepTotal  : int  13162 10735 10460 9762 12669 9705 13019 15506 10544 9
819 ...
```

```
str(H_Calories)
```

```
## 'data.frame':    22099 obs. of  5 variables:
##  $ Id          : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ ActivityHour: chr  "4/12/2016 0:00" "4/12/2016 1:00" "4/12/2016 2:00" "
4/12/2016 3:00" ...
##  $ Calories    : int  81 61 59 47 48 48 48 47 68 141 ...
##  $ Date        : chr  "4/12/2016" "4/12/2016" "4/12/2016" "4/12/2016" ...
##  $ Time        : chr  "0:00:00" "1:00:00" "2:00:00" "3:00:00" ...
```

```
str(H_Intensities)
```

```
## 'data.frame':    22099 obs. of  6 variables:
##  $ Id              : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ ActivityHour    : chr  "4/12/2016 0:00" "4/12/2016 1:00" "4/12/2016 2:0
0" "4/12/2016 3:00" ...
##  $ TotalIntensity  : int  20 8 7 0 0 0 0 0 13 30 ...
##  $ AverageIntensity: num  0.333 0.133 0.117 0 0 ...
##  $ Date            : chr  "4/12/2016" "4/12/2016" "4/12/2016" "4/12/2016"
...
##  $ Time            : chr  "0:00:00" "1:00:00" "2:00:00" "3:00:00" ...
```

```
str(Sleep_Day)
```

```
## 'data.frame':    413 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 0:00" "4/13/2016 0:00" "4/15/2016 0
:00" "4/16/2016 0:00" ...
##  $ 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 ...
```

With str function, we can see that those files have the Activity column in chr format, and we need to change those columns to date format in each csv file as a Date for R can read it.

```
D_Activity$ActivityDate <- mdy(D_Activity$ActivityDate)
D_Calories$ActivityDay <- mdy(D_Calories$ActivityDay)
D_Intensities$ActivityDay <- mdy(D_Intensities$ActivityDay)
D_Steps$ActivityDay <- mdy(D_Steps$ActivityDay)
Sleep_Day$Date <- as.Date(Sleep_Day$SleepDay, format="%m/%d/%Y")
```

In the csv files we have the Id column, and this represents the identification number of one user, but this column has a lot of rows and for this reason, I’ll use the distinct function and see how many users recorded their activity:

```
n_distinct (D_Activity$Id)

## [1] 33

n_distinct (D_Calories$Id)

## [1] 33

n_distinct (D_Intensities$Id)

## [1] 33
```

```r
n_distinct (D_Steps$Id)

## [1] 33

n_distinct (H_Calories$Id)

## [1] 33

n_distinct (H_Intensities$Id)

## [1] 33

n_distinct (Sleep_Day$Id)

## [1] 24
```

Now that we know the number of users, I'll use the same function to find out how many days the users logged their activity, as follow:

```r
n_distinct(D_Activity$ActivityDate)

## [1] 31

n_distinct(D_Calories$ActivityDay)

## [1] 31

n_distinct(D_Intensities$ActivityDay)

## [1] 31

n_distinct(D_Steps$ActivityDay)

## [1] 31

n_distinct(H_Calories$Date)

## [1] 31

n_distinct(H_Intensities$Date)

## [1] 31

n_distinct(Sleep_Day$SleepDay)

## [1] 31
```

From now, we know the number of users, and the number of days that users logged their activity, next, I´ll check for duplicates in the daily files, and if I have duplicates, I going to eliminate.

```
sum(duplicated(D_Activity))

## [1] 0

sum(duplicated(D_Calories))

## [1] 0

sum(duplicated(D_Intensities))

## [1] 0

sum(duplicated(D_Steps))

## [1] 0

sum(duplicated(Sleep_Day))

## [1] 3
```

In the Sleep_Day data, the duplicated function found 3 duplicates, for that reason, I'll use the unique function to eliminate those 3 duplicated in the Sleep_Day data, and then start finding patterns and trends in our analysis.

5. The Analysis

For this phase, first I want to check for the min, max, average, etc., of the main columns of the files, and get a better idea of the whole data and make this analysis easier to understand and comprehend, as follow:

```
#D_Activity
D_Activity %>%
  select (TotalDistance, TrackerDistance, TotalSteps, Calories) %>%
  summary()

##  TotalDistance    TrackerDistance    TotalSteps       Calories
##  Min.   : 0.000   Min.   : 0.000   Min.   :    0   Min.   :   0
##  1st Qu.: 2.620   1st Qu.: 2.620   1st Qu.: 3790   1st Qu.:1828
##  Median : 5.245   Median : 5.245   Median : 7406   Median :2134
##  Mean   : 5.490   Mean   : 5.475   Mean   : 7638   Mean   :2304
##  3rd Qu.: 7.713   3rd Qu.: 7.710   3rd Qu.:10727   3rd Qu.:2793
##  Max.   :28.030   Max.   :28.030   Max.   :36019   Max.   :4900
```

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

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

```
#D_Intensities
D_Intensities %>%
  select(SedentaryMinutes, LightlyActiveMinutes, FairlyActiveMinutes, VeryAct
iveMinutes, SedentaryActiveDistance, LightActiveDistance,
  ModeratelyActiveDistance, VeryActiveDistance) %>%
  summary()
```

```
##  SedentaryMinutes LightlyActiveMinutes FairlyActiveMinutes VeryActiveMinut
es
##  Min.   :   0.0   Min.   :  0.0        Min.   :  0.00      Min.   :  0.00
##  1st Qu.: 729.8   1st Qu.:127.0        1st Qu.:  0.00      1st Qu.:  0.00
##  Median :1057.5   Median :199.0        Median :  6.00      Median :  4.00
##  Mean   : 991.2   Mean   :192.8        Mean   : 13.56      Mean   : 21.16
##  3rd Qu.:1229.5   3rd Qu.:264.0        3rd Qu.: 19.00      3rd Qu.: 32.00
##  Max.   :1440.0   Max.   :518.0        Max.   :143.00      Max.   :210.00
##  SedentaryActiveDistance LightActiveDistance ModeratelyActiveDistance
##  Min.   :0.000000        Min.   : 0.000      Min.   :0.0000
##  1st Qu.:0.000000        1st Qu.: 1.945      1st Qu.:0.0000
##  Median :0.000000        Median : 3.365      Median :0.2400
##  Mean   :0.001606        Mean   : 3.341      Mean   :0.5675
##  3rd Qu.:0.000000        3rd Qu.: 4.782      3rd Qu.:0.8000
##  Max.   :0.110000        Max.   :10.710      Max.   :6.4800
##  VeryActiveDistance
##  Min.   : 0.000
##  1st Qu.: 0.000
##  Median : 0.210
##  Mean   : 1.503
##  3rd Qu.: 2.053
##  Max.   :21.920
```

```
#D_Steps
D_Steps %>%
  select(StepTotal) %>%
  summary()
```

```
##    StepTotal
##  Min.   :    0
##  1st Qu.: 3790
##  Median : 7406
##  Mean   : 7638
##  3rd Qu.:10727
##  Max.   :36019

#Sleep_Day
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
```

When we are looking in these data, show us useful information about user’s activity, like there are 33 users and 31 days that the users logged their activity, algo we can see burned calories, step walked, distance traveled, etc.

I think the D_Activity dataset has a lot of information, so, I will create a column named weekday, and I will call it W_Activity.

```
W_Activity <- D_Activity
W_Activity$Weekday <- weekdays(W_Activity$ActivityDate)
```

I will use colnames() function to watch the columns of W_Activity.

```
colnames(W_Activity)

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

Observations:

a. Average user sleep time is 419.2 minutes (6.9 hours), let’s round it up to 7 hours, Centers for Disease Control and Prevention (CDC) and Sleep Foundation A OneCare Media Company, recommended for adults between the ages 18 and 60 years, 7 or more hours of sleep per night. This number is barely ideal, because hit the lowest range of the ideal range.

b. Looking in the step total in day, the average people that logged their steps, is 7638, which is very good, because they almost achieve the 8000 steps, and according to National Institute of Health (NIH), adults who took 8,000 or more steps a day had a reduce risk of death over the following decade than those who only walked 4,000 steps a day.

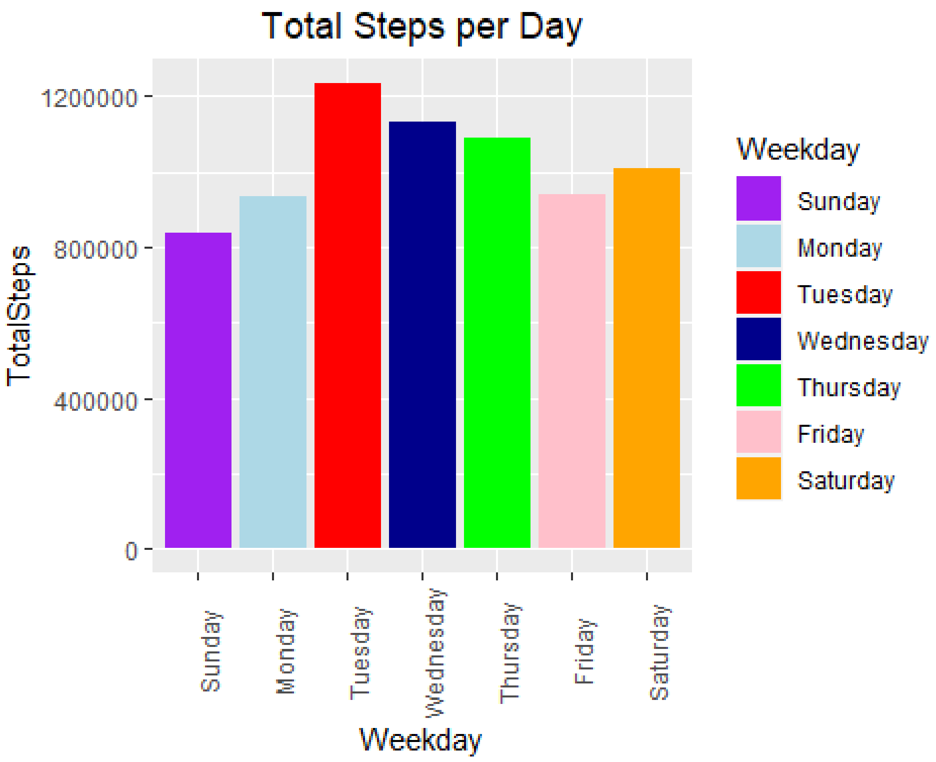
c. Another statistic interesting, it’s the average sedentary minutes, which is 991.2, that means 16.5 hours, this is incredible high, because the time that the users were sitting without doing a physic activity, within the 24 hours that a day has.

Now that we have an idea of the data, I’ll start creating some plots and look for trends, patterns, and correlations between the data.

First we have to order the days of the week:

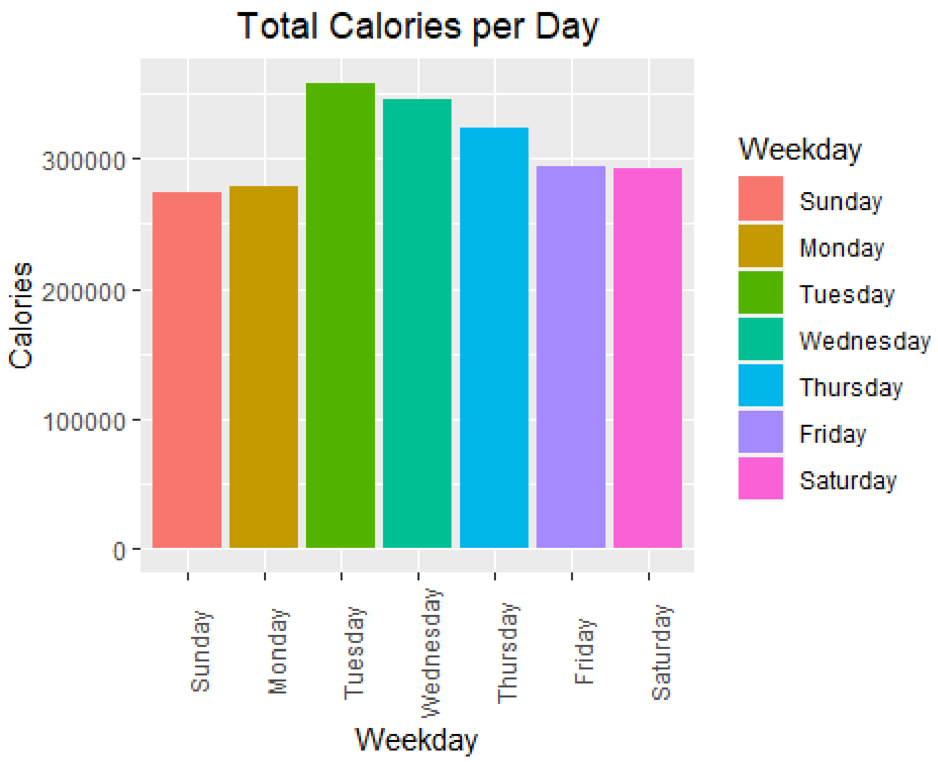
```
W_Activity$Weekday <- factor(W_Activity$Weekday, levels= c(“Sunday”, “Monday”
,
            “Tuesday”, “Wednesday”, “Thursday”, “Friday”, “Saturday”))

ggplot(data = W_Activity, aes (x=Weekday, y = TotalSteps, fill = Weekday)) +
  geom_col() +
  scale_fill_manual(values=c(“purple”, “lightblue”, “red”, “darkblue”, “green
”, “pink”, “orange”)) +
labs(title=”Total Steps per Day”) +
theme(plot.title = element_text(hjust = 0.5)) +
theme(axis.text.x = element_text(angle = 90))
```



This shows us, that the users are most active in the middle of the week, and the activity levels decline as the end of the week, starting to level up again at the beginning of the week. This could be from a busy work schedule.

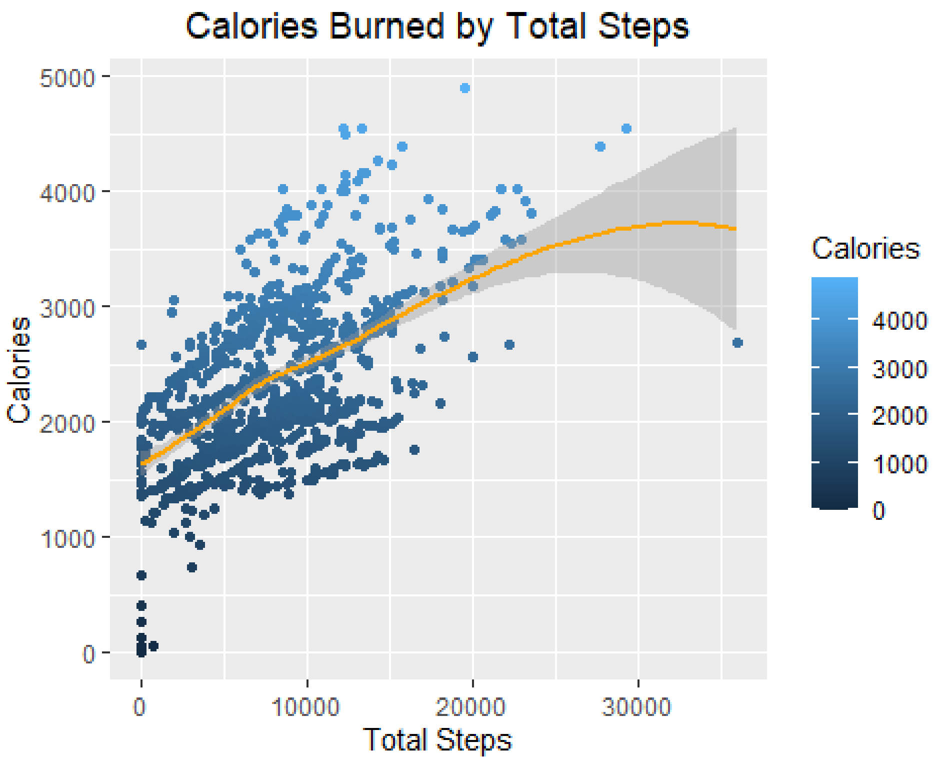
```
options(scipen = 999)
ggplot(data = W_Activity, aes (x=Weekday, y = Calories, fill = Weekday))+
  geom_col()+
  labs(title="Total Calories per Day") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(axis.text.x = element_text(angle = 90))
```



Here it shows us almost the same result as the first graph, this means that when you are more active, you burn more calories.

```
ggplot(data=D_Activity, aes(x=TotalSteps, y=Calories)) +
  geom_point(aes(colour = Calories)) +
  geom_smooth (color ="orange") +
  labs(title ="Calories Burned by Total Steps",
       x = "Total Steps",
       y = "Calories") +
  theme(plot.title = element_text(hjust = 0.5))

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

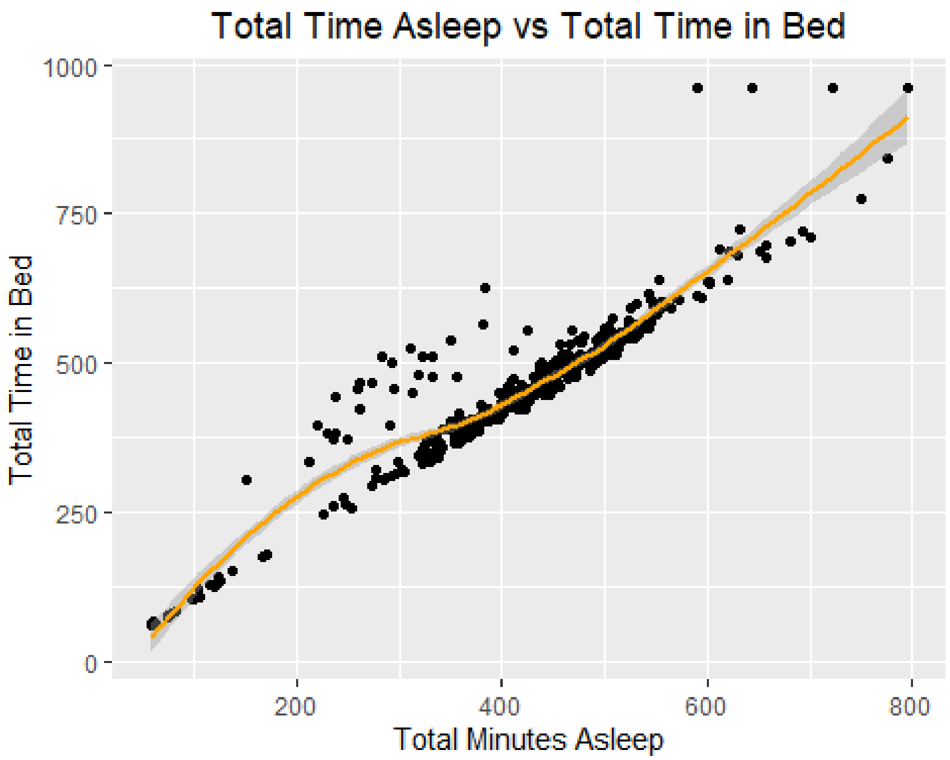


From these plots, we can see a strong correlation between the number of steps, and the calories burned. When users take more steps, more calories are burned.

A positive trend appears on these graphs, but there are also users that burned a lot of calories taking fewer steps, this may be due to different reasons, perhaps they do yoga, bodybuilding or another activity that does not involve walking or running.

```
ggplot(data=Sleep_Day, aes(x=TotalMinutesAsleep, y=TotalTimeInBed)) +
  geom_point() +
  geom_smooth (color ="orange") +
  labs(title ="Total Time Asleep vs Total Time in Bed",
       x = "Total Minutes Asleep",
       y = "Total Time in Bed") +
  theme(plot.title = element_text(hjust = 0.5))

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



There is a strong correlation between these two variables, it seems that if users go to bed earlier, they can sleep earlier.

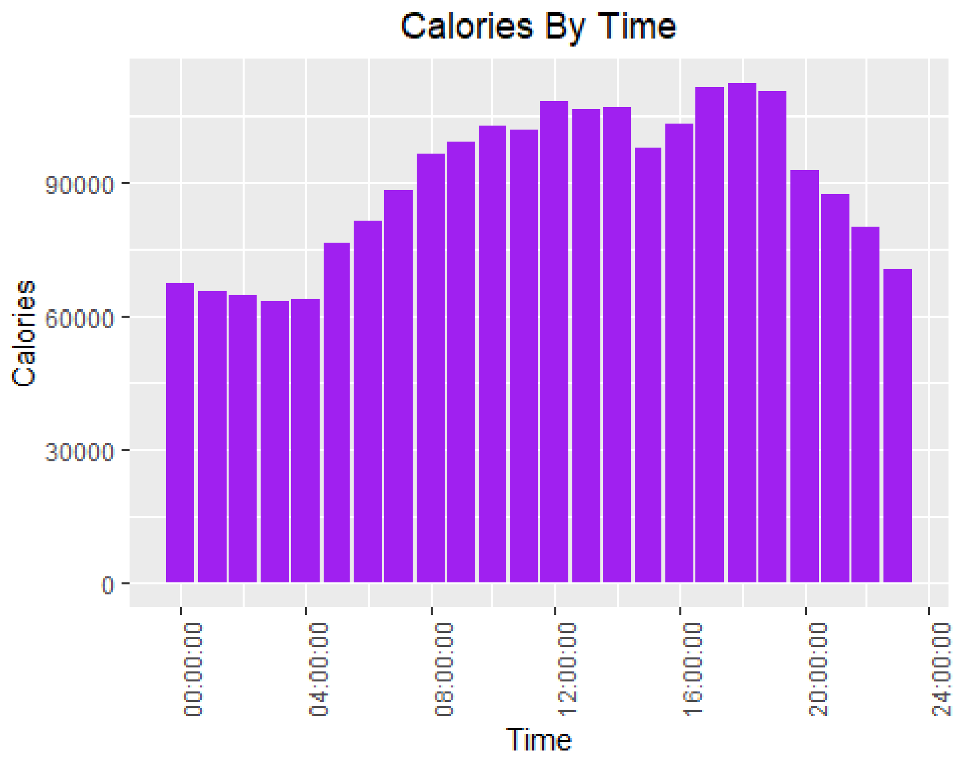
It is well known that nowadays, most of the people use their cellphones to watch some trend videos or spend time in social medias when they are in bed, this may be mean because users fall asleep late at night, which causes them to rest less.

The time columns in H_Calories and H_Intensities are character type, so I'll transform those columns into time format using the hms function, as follows:

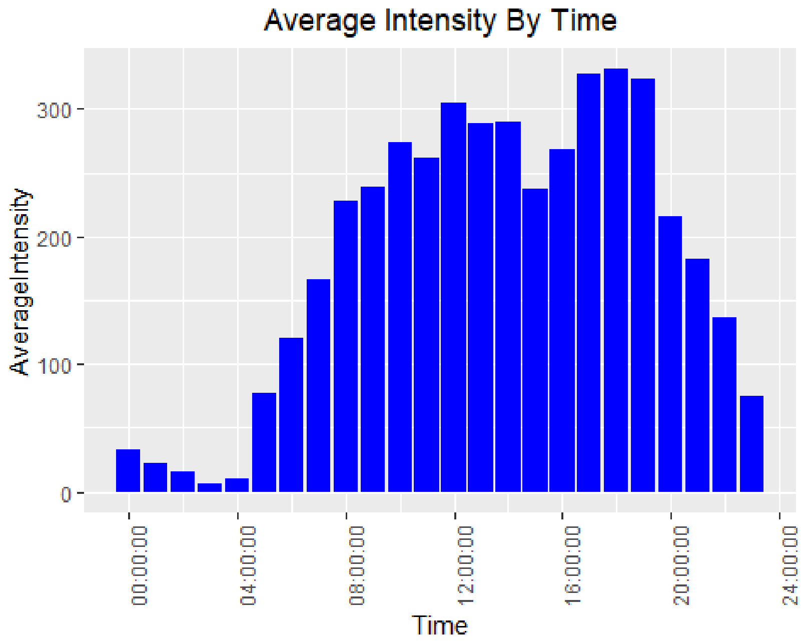
```
H_Intensities$Time <- as_hms(H_Intensities$Time)
H_Calories$Time <- as_hms(H_Calories$Time)
```

Now we can make some plots:

```
ggplot (data = H_Calories, aes (x=Time,y=Calories)) +
  geom_bar(stat = "identity", fill=('purple')) +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(title= "Calories By Time") +
  theme(plot.title = element_text(hjust = 0.5))
```



```r
ggplot (data = H_Intensities, aes (x=Time,y=AverageIntensity)) +
  geom_bar(stat = "identity", fill=('blue')) +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(title= "Average Intensity By Time") +
  theme(plot.title = element_text(hjust = 0.5))
```

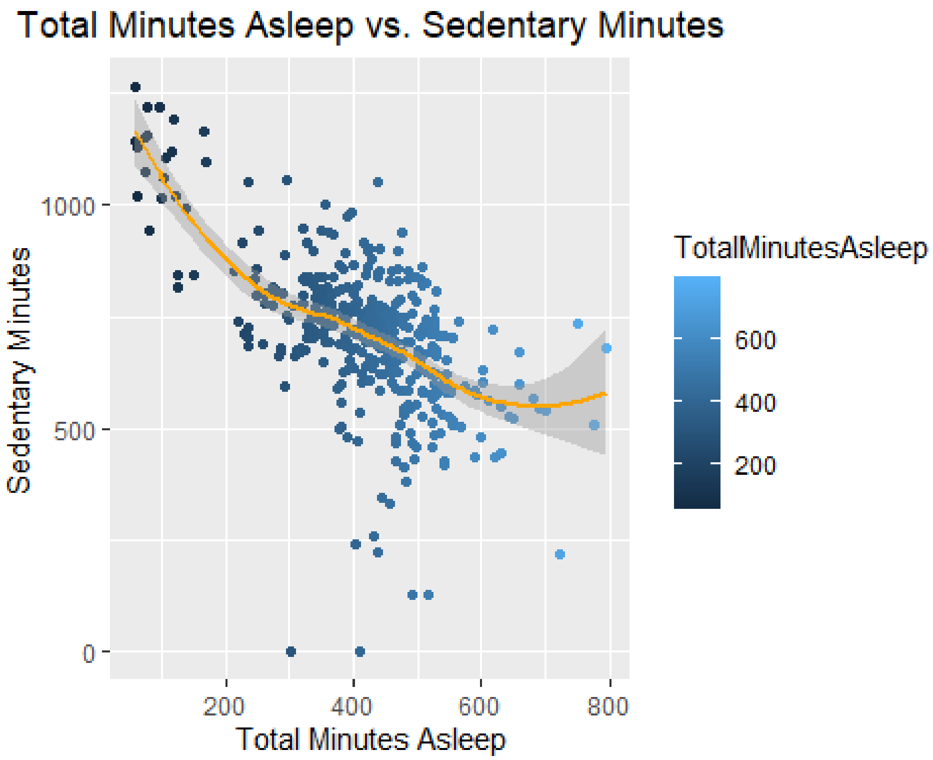


In both chart, the highest user's activity was during 12 pm and 5 pm to 7 pm. This lets us know that the users focus on being active either in between their work breaks and after their job.

```
Intensities_Sleep_merged <- merge(D_Intensities, Sleep_Day,     by.x=c("Id",
"ActivityDay"), by.y=c("Id", "Date"))

ggplot(data=Intensities_Sleep_merged, aes(x=TotalMinutesAsleep, y=SedentaryMi
nutes)) +
  geom_point(aes(colour = TotalMinutesAsleep)) +
  geom_smooth (color ="orange") +
  labs(title ="Total Minutes Asleep vs. Sedentary Minutes",
       x = "Total Minutes Asleep",
       y = "Sedentary Minutes") +
  theme (plot.title = element_text(hjust = 0.5))

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



There is a negative relationship between these two variables, since if someone spends too much time sitting or is not active, they may have problems wanting to sleep, and the consequence will be that they will sleep less.

Bellabeat products have to send notifications or alerts about being active when the user spends too much time sitting.

6. Analysis Results.

The dataset of this study is outdated because is already 6 years old, main reason that Bellabeat will need to gather more actual information, another key points for improvement would be an increase in the number of participants and capture of demographic data such as age, gender, height, location, weight, etc. Another important key point, it would be that Bellabeat covered a larger period of gathered data, including two or more seasons,

Due to the lack of variety of information, the analysis can result inaccurate, in addition of the small number of users, and the short period of time that the information cover.

It is well known that when a person walks more, they burn more calories, a situation that is demonstrated in this analysis, the users who took more steps or were more active, burned more calories, also, we could see that the hours where the users presented more activity was at 12 noon and between 5 and 7 hours.

Another important cause to highlight is that the more minutes of sedentary lifestyle you have, the less you sleep.

In order to devise new marketing ideas, we have to take into account the facts we already have, for better focus and new ways to reach more users and keep old users interested in Bellabeat products. The recommendations I would provide to help solve this business-related scenario is shown below.

7. Marketing Strategies.

- First Sample data was dated in 2016 and lacks information such as demographic, gender and age. Bellabeat could collect new data, with a larger sample size, ideally having demographic in line with its target customers, e.g. women working in office, university students, and see if the usage pattern still in line with the assumption above.

- People with healthy habits, should know about weight, calorie, diet, etc., Bellabeat's social media should post content about healthy activities, nutrient content, benefits of being active, and the last news about healthy lifestyle, including exercise.

- Bellabeat's team can create content about light routines of exercises to recommend for users because they may not have enough time to go to the gym, or another activity like yoga.

- Device with a lot of customization and notification settings, e.g., set the minimum steps a user wants (National Institute of Health recommends 8000 steps per day), the user sets their number of steps depending on their own goals.
- Bellabeat’s products must have a step counter, including a reminder to be active when the device detects a prolonged period of sedentary time.
- Create content on it home pages or social media, about awareness of spending less time in bed, since that implies resting less, in addition to setting it as an alarm and adding it as a reminder.
- Best time for notifications, depending on the user’s settings, to send warnings if the sedentary time reaches the unhealthy rate.
- Two types of subscription, free and paid, where the free one contains basic information and the other one that contains, in addition to more complete information, other types of recommendations such as low-calorie meals, etc.

8. Conclusion.

After completing the data analysis processes, such as Ask, Prepare, Process, Analyze, and Sharing, some patterns and trends could be found among the few registered users in this dataset.

We were able to find the days when users burned the most calories, the days they took the most steps, data about their sleeping time, etc., now if we could do this with so few users as well as information, now to think what could be done With a data set that contains more and updated information, we can find many other things and therefore make more marketing recommendations.

These findings, which were already demonstrated in this analysis, we were able to recommend some marketing strategies for the Bellabeat company.