BellabeatAnalysis

Shahenaz H

2024-01-01

## Bellabeat Google Analytics Capstone Project Case Study (Ask phase)

1. Background Information

Bellabeat is a small successful company that manufactures high tech fitness gadgets for women. The main goal is to provide health related data like sleep, stress, menstrual cycle, and mindfulness habits of user ensuring, inspiring and empowering women about their own health.

1. Characters and products

Characters

* Urška Sršen: Bellabeat’s cofounder and Chief Creative Officer
* Sando Mur: Mathematician and Bellabeat’s cofounder; key member of the Bellabeat executive team
* Bellabeat marketing analytics team: A team of data analysts responsible for collecting, analyzing, and reporting data that helps guide Bellabeat’s marketing strategy. You joined this team six months ago and have been busy learning about Bellabeat’s mission and business goals — as well as how you, as a junior data analyst, can help Bellabeat achieve them.

Products

* Bellabeat app: The Bellabeat app provides users with health data related to their activity, sleep, stress, menstrual cycle, and mindfulness habits. This data can help users better understand their current habits and make healthy decisions. The Bellabeat app connects to their line of smart wellness products.
* Leaf: Bellabeat’s classic wellness tracker can be worn as a bracelet, necklace, or clip. The Leaf tracker connects to the Bellabeat app to track activity, sleep, and stress.
* Time: This wellness watch combines the timeless look of a classic timepiece with smart technology to track user activity, sleep, and stress. The Time watch connects to the Bellabeat app to provide you with insights into your daily wellness.
* Spring: This is a water bottle that tracks daily water intake using smart technology to ensure that you are appropriately hydrated throughout the day. The Spring bottle connects to the Bellabeat app to track your hydration levels.
* Bellabeat membership: Bellabeat also offers a subscription-based membership program for users. Membership gives users 24/7 access to fully personalized guidance on nutrition, activity, sleep, health and beauty, and mindfulness based on their lifestyle and goals.

1. Business Task

The objective is to analyze the device usage and present the findings to the marketing team for making strategies for Bellabeat gadget growth across users and non-users.

1. About the Data

The data is FitBit Fitness Tracker Data which is Public Domain, dataset made available through Mobius. This Kaggle data set contains personal fitness tracker from thirty fitbit users. Thirty eligible Fitbit users consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. It includes information about daily activity, steps, and heart rate that can be used to explore users’ habits.

* Limitations

The data samples only 30 users.

The data is outdated dating back to 2016 and has been collected by third party.

The data does not include demographic information of users.

## Setting up environment (Prepare Phase)

Notes: Setting up environment by loading the packages

install.packages('tidyverse')

## 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)

install.packages('here')

## Installing package into '/cloud/lib/x86\_64-pc-linux-gnu-library/4.3'  
## (as 'lib' is unspecified)

install.packages('janitor')

## Installing package into '/cloud/lib/x86\_64-pc-linux-gnu-library/4.3'  
## (as 'lib' is unspecified)

install.packages('dplyr')

## Installing package into '/cloud/lib/x86\_64-pc-linux-gnu-library/4.3'  
## (as 'lib' is unspecified)

install.packages('rmarkdown')

## Installing package into '/cloud/lib/x86\_64-pc-linux-gnu-library/4.3'  
## (as 'lib' is unspecified)

library(tidyverse)

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

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(ggplot2)  
library(here)

## here() starts at /cloud/project

library(janitor)

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

library(dplyr)  
library('rmarkdown')

## Load CSV files (Prepare Phase)

daily\_activity <- read.csv("dailyActivity\_merged.csv")  
sleep\_day <- read.csv("sleepDay\_merged.csv")  
weight\_log\_info <- read.csv("weightLogInfo\_merged.csv")  
daily\_steps <- read.csv("dailySteps\_merged.csv")  
hourly\_steps <- read.csv("hourlySteps\_merged.csv")  
daily\_calories <- read.csv("dailyCalories\_merged.csv")

## Explore a few key tables (Prepare Phase)

head(daily\_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

colnames(daily\_activity)

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

head(sleep\_day)

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

colnames(sleep\_day)

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

head(weight\_log\_info)

## Id Date WeightKg WeightPounds Fat BMI  
## 1 1503960366 5/2/2016 11:59:59 PM 52.6 115.9631 22 22.65  
## 2 1503960366 5/3/2016 11:59:59 PM 52.6 115.9631 NA 22.65  
## 3 1927972279 4/13/2016 1:08:52 AM 133.5 294.3171 NA 47.54  
## 4 2873212765 4/21/2016 11:59:59 PM 56.7 125.0021 NA 21.45  
## 5 2873212765 5/12/2016 11:59:59 PM 57.3 126.3249 NA 21.69  
## 6 4319703577 4/17/2016 11:59:59 PM 72.4 159.6147 25 27.45  
## IsManualReport LogId  
## 1 True 1.462234e+12  
## 2 True 1.462320e+12  
## 3 False 1.460510e+12  
## 4 True 1.461283e+12  
## 5 True 1.463098e+12  
## 6 True 1.460938e+12

head(daily\_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(daily\_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(hourly\_steps)

## Id ActivityHour StepTotal  
## 1 1503960366 4/12/2016 12:00:00 AM 373  
## 2 1503960366 4/12/2016 1:00:00 AM 160  
## 3 1503960366 4/12/2016 2:00:00 AM 151  
## 4 1503960366 4/12/2016 3:00:00 AM 0  
## 5 1503960366 4/12/2016 4:00:00 AM 0  
## 6 1503960366 4/12/2016 5:00:00 AM 0

## Cleaning the data frames (Process Phase)

### 1.Removing LoggedActivitiesDistance and SedentaryAvtiveDistance as they are not used in analysis from daily\_activity dataframe

daily\_activity <- subset(daily\_activity,select = -c(LoggedActivitiesDistance, SedentaryActiveDistance))

colnames(daily\_activity)[2] = 'Date'

colnames(daily\_activity)

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

### 2. Separating ActivityHour column from hourly\_steps dataframe into Date and Hour column for analysis and check modified hourly\_steps

hourly\_steps <- hourly\_steps %>%   
 separate(ActivityHour, c("Date", "Hour"), sep = "^\\S\*\\K")  
head(hourly\_steps)

## Id Date Hour StepTotal  
## 1 1503960366 4/12/2016 12:00:00 AM 373  
## 2 1503960366 4/12/2016 1:00:00 AM 160  
## 3 1503960366 4/12/2016 2:00:00 AM 151  
## 4 1503960366 4/12/2016 3:00:00 AM 0  
## 5 1503960366 4/12/2016 4:00:00 AM 0  
## 6 1503960366 4/12/2016 5:00:00 AM 0

### 3.Removing SleepRecords and time as it is not used in analysis from sleep\_day dataframe and check modified dataframe

sleep\_day <- subset(sleep\_day,select = -c(TotalSleepRecords))  
sleep\_day <- sleep\_day %>%   
 separate(SleepDay, c("Date"), sep = "^\\S\*\\K")

## Warning: Expected 1 pieces. Additional pieces discarded in 413 rows [1, 2, 3, 4, 5, 6,  
## 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].

head(sleep\_day)

## Id Date TotalMinutesAsleep TotalTimeInBed  
## 1 1503960366 4/12/2016 327 346  
## 2 1503960366 4/13/2016 384 407  
## 3 1503960366 4/15/2016 412 442  
## 4 1503960366 4/16/2016 340 367  
## 5 1503960366 4/17/2016 700 712  
## 6 1503960366 4/19/2016 304 320

## Understanding some summary statistics (Analyze Phase)

### How many unique participants are there in each dataframe?

n\_distinct(daily\_activity$Id)

## [1] 33

n\_distinct(sleep\_day$Id)

## [1] 24

n\_distinct(daily\_steps$Id)

## [1] 33

n\_distinct(daily\_calories$Id)

## [1] 33

n\_distinct(weight\_log\_info$Id)

## [1] 8

Note: There are only 8 participants in weight\_log\_info dataframe which is very less to draw any calculation. So we will exclude it from our analysis.

## How many observations are there in each dataframe? (Analyze Phase)

nrow(daily\_activity)

## [1] 940

nrow(sleep\_day)

## [1] 413

nrow(daily\_calories)

## [1] 940

nrow(daily\_steps)

## [1] 940

nrow(hourly\_steps)

## [1] 22099

##Quick Summary

daily\_activity %>%   
 select(TotalSteps,  
 TotalDistance,  
 SedentaryMinutes) %>%  
 summary()

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

sleep\_day %>%   
 select(TotalMinutesAsleep,  
 TotalTimeInBed) %>%  
 summary()

## TotalMinutesAsleep TotalTimeInBed   
## Min. : 58.0 Min. : 61.0   
## 1st Qu.:361.0 1st Qu.:403.0   
## Median :433.0 Median :463.0   
## Mean :419.5 Mean :458.6   
## 3rd Qu.:490.0 3rd Qu.:526.0   
## Max. :796.0 Max. :961.0

daily\_calories %>%  
 select(Id,  
 Calories) %>%  
 summary()

## Id Calories   
## Min. :1.504e+09 Min. : 0   
## 1st Qu.:2.320e+09 1st Qu.:1828   
## Median :4.445e+09 Median :2134   
## Mean :4.855e+09 Mean :2304   
## 3rd Qu.:6.962e+09 3rd Qu.:2793   
## Max. :8.878e+09 Max. :4900

daily\_steps %>%  
 select(Id,  
 StepTotal) %>%  
 summary()

## Id StepTotal   
## Min. :1.504e+09 Min. : 0   
## 1st Qu.:2.320e+09 1st Qu.: 3790   
## Median :4.445e+09 Median : 7406   
## Mean :4.855e+09 Mean : 7638   
## 3rd Qu.:6.962e+09 3rd Qu.:10727   
## Max. :8.878e+09 Max. :36019

hourly\_steps %>%  
 select(Id,  
 StepTotal) %>%  
 summary()

## Id StepTotal   
## Min. :1.504e+09 Min. : 0.0   
## 1st Qu.:2.320e+09 1st Qu.: 0.0   
## Median :4.445e+09 Median : 40.0   
## Mean :4.848e+09 Mean : 320.2   
## 3rd Qu.:6.962e+09 3rd Qu.: 357.0   
## Max. :8.878e+09 Max. :10554.0

###Finding average steps taken by each participant

mean\_steps <- daily\_activity %>%  
 group\_by(Id)%>%  
 summarise(mean\_steps = mean(TotalSteps)) %>%  
 select(Id, mean\_steps) %>%  
 arrange(mean\_steps) %>%  
 as.data.frame()  
head(mean\_steps)

## Id mean\_steps  
## 1 1927972279 916.129  
## 2 8792009665 1853.724  
## 3 4020332650 2267.226  
## 4 6775888955 2519.692  
## 5 1844505072 2580.065  
## 6 4057192912 3838.000

###Finding average sedentary minutes of each participant

mean\_sedentary\_minutes <- daily\_activity %>%  
 group\_by(Id)%>%  
 summarise(mean\_sedentary\_minutes = mean(SedentaryMinutes)) %>%  
 select(Id, mean\_sedentary\_minutes) %>%  
 arrange(mean\_sedentary\_minutes) %>%  
 as.data.frame()  
head(mean\_sedentary\_minutes)

## Id mean\_sedentary\_minutes  
## 1 6962181067 662.3226  
## 2 5553957443 668.3548  
## 3 2347167796 687.1667  
## 4 2026352035 689.4194  
## 5 3977333714 707.5333  
## 6 8378563200 716.1290

###Finding average sleep taken by each participant

mean\_sleep <- sleep\_day %>%  
 group\_by(Id)%>%  
 summarise(mean\_sleep = mean(TotalMinutesAsleep)) %>%  
 select(Id, mean\_sleep) %>%  
 arrange(mean\_sleep) %>%  
 as.data.frame()  
head(mean\_sleep)

## Id mean\_sleep  
## 1 2320127002 61.0000  
## 2 7007744171 68.5000  
## 3 4558609924 127.6000  
## 4 3977333714 293.6429  
## 5 1644430081 294.0000  
## 6 8053475328 297.0000

###Finding actual minutes of sleep in percentage by each participant

actual\_sleep <- sleep\_day %>%  
 group\_by(Id) %>%  
 mutate(percent\_sleep = (TotalMinutesAsleep/TotalTimeInBed)\*100) %>%  
 select(Id, percent\_sleep) %>%  
 summarize(actual\_sleep = mean(percent\_sleep)) %>%  
 arrange(actual\_sleep) %>%  
 mutate\_if(is.numeric, round, 2)  
 glimpse(actual\_sleep)

## Rows: 24  
## Columns: 2  
## $ Id <dbl> 3977333714, 1844505072, 1644430081, 2320127002, 455860992…  
## $ actual\_sleep <dbl> 63.37, 67.85, 88.20, 88.41, 90.71, 91.04, 91.52, 91.90, 9…

Note : All participants sleep almost 90% of time except four. The least recorded is 63%

### Merging these two datasets together to find additional insights

combined\_data <- merge(sleep\_day, daily\_activity)  
glimpse(combined\_data)

## Rows: 413  
## Columns: 15  
## $ Id <dbl> 1503960366, 1503960366, 1503960366, 150396036…  
## $ Date <chr> "4/12/2016", "4/13/2016", "4/15/2016", "4/16/…  
## $ TotalMinutesAsleep <int> 327, 384, 412, 340, 700, 304, 360, 325, 361, …  
## $ TotalTimeInBed <int> 346, 407, 442, 367, 712, 320, 377, 364, 384, …  
## $ TotalSteps <int> 13162, 10735, 9762, 12669, 9705, 15506, 10544…  
## $ TotalDistance <dbl> 8.50, 6.97, 6.28, 8.16, 6.48, 9.88, 6.68, 6.3…  
## $ TrackerDistance <dbl> 8.50, 6.97, 6.28, 8.16, 6.48, 9.88, 6.68, 6.3…  
## $ VeryActiveDistance <dbl> 1.88, 1.57, 2.14, 2.71, 3.19, 3.53, 1.96, 1.3…  
## $ ModeratelyActiveDistance <dbl> 0.55, 0.69, 1.26, 0.41, 0.78, 1.32, 0.48, 0.3…  
## $ LightActiveDistance <dbl> 6.06, 4.71, 2.83, 5.04, 2.51, 5.03, 4.24, 4.6…  
## $ VeryActiveMinutes <int> 25, 21, 29, 36, 38, 50, 28, 19, 41, 39, 73, 3…  
## $ FairlyActiveMinutes <int> 13, 19, 34, 10, 20, 31, 12, 8, 21, 5, 14, 23,…  
## $ LightlyActiveMinutes <int> 328, 217, 209, 221, 164, 264, 205, 211, 262, …  
## $ SedentaryMinutes <int> 728, 776, 726, 773, 539, 775, 818, 838, 732, …  
## $ Calories <int> 1985, 1797, 1745, 1863, 1728, 2035, 1786, 177…

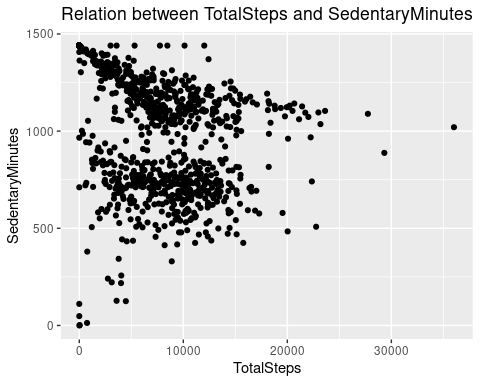
combined\_steps\_calories <- merge(daily\_steps, daily\_calories)  
head(combined\_steps\_calories)

## Id ActivityDay StepTotal Calories  
## 1 1503960366 4/12/2016 13162 1985  
## 2 1503960366 4/13/2016 10735 1797  
## 3 1503960366 4/14/2016 10460 1776  
## 4 1503960366 4/15/2016 9762 1745  
## 5 1503960366 4/16/2016 12669 1863  
## 6 1503960366 4/17/2016 9705 1728

## Data Visualisation

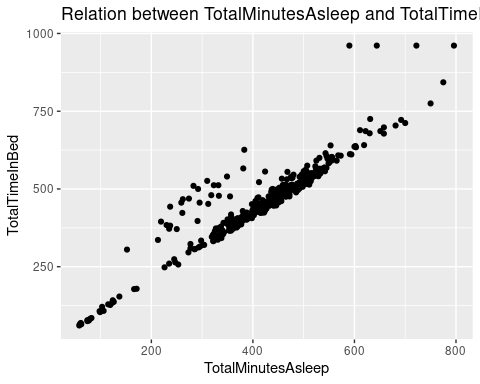
###Plotting a few explorations

ggplot(data=daily\_activity, aes(x=TotalSteps, y=SedentaryMinutes)) + geom\_point() +  
 labs(title = "Relation between TotalSteps and SedentaryMinutes ",  
 x = "TotalSteps", y = "SedentaryMinutes")



Note: The relation between SedentaryMinutes and TotalSteps taken by participant is non-linear. That means TotalSteps taken in a day doesnot depend on sedentary minutes.

ggplot(data=sleep\_day, aes(x=TotalMinutesAsleep, y=TotalTimeInBed)) + geom\_point() + labs(title = "Relation between TotalMinutesAsleep and TotalTimeInBed ",  
 x = "TotalMinutesAsleep", y = "TotalTimeInBed")



Note: The relation between TotalMinutesAsleep and TotalTimeInBed is linear.

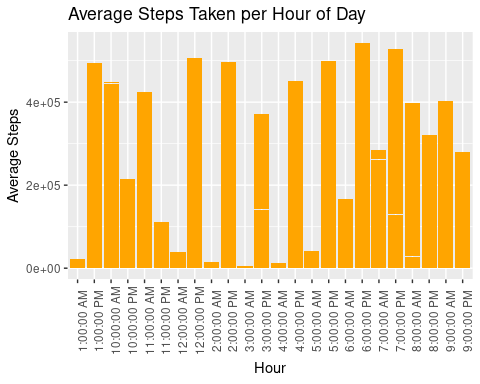
ggplot(data=combined\_steps\_calories, aes(x=StepTotal, y=Calories)) + geom\_point() + geom\_smooth() + labs(title = "Relation between TotalSteps and Calories ",  
 x = "StepTotal", y = "Calories")

## `geom\_smooth()` using method = 'loess' and formula = 'y ~ x'

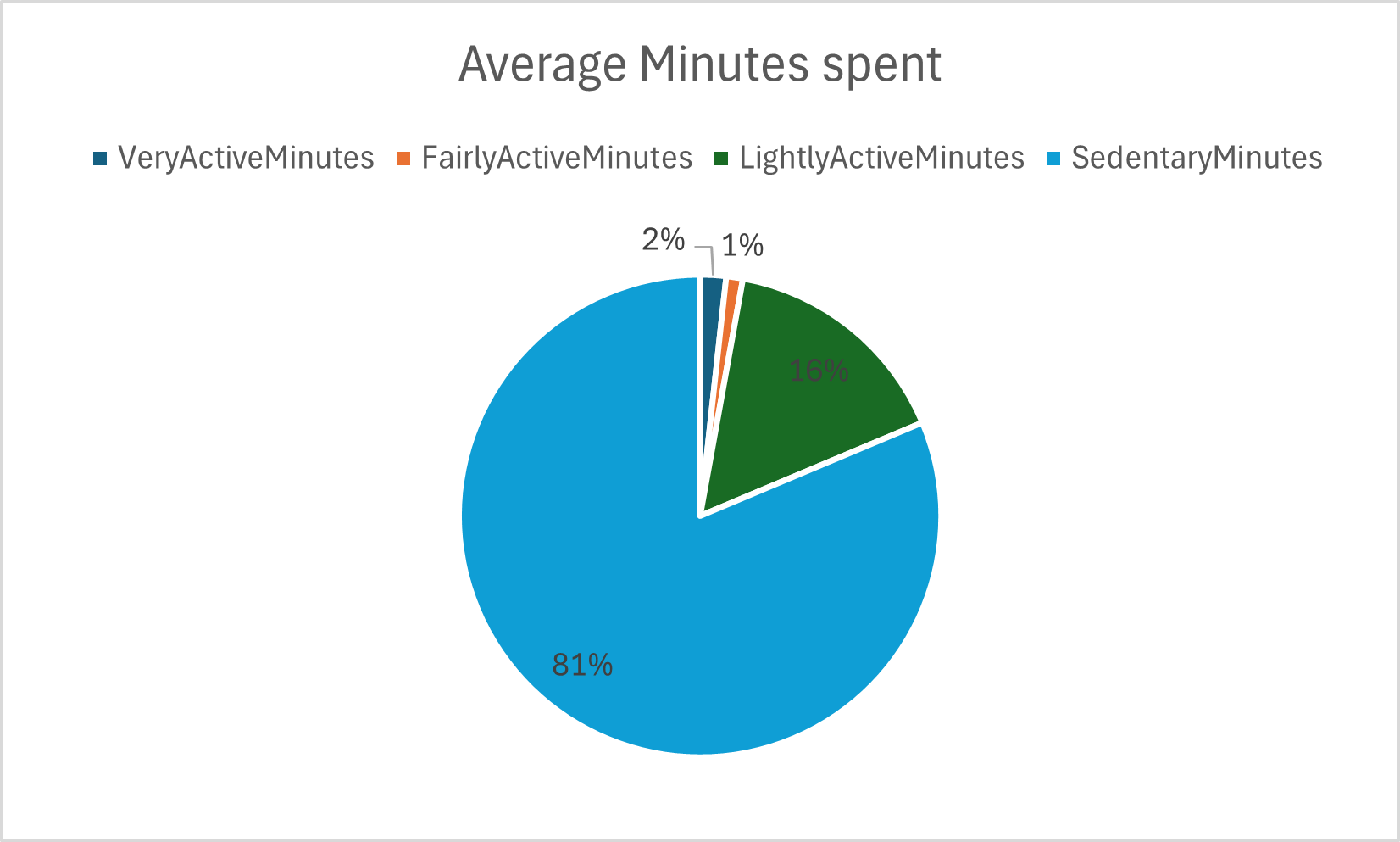


Note: The relation between StepTotal and Calories is increasing gradually. That means the more steps you take, the more calories you burn.

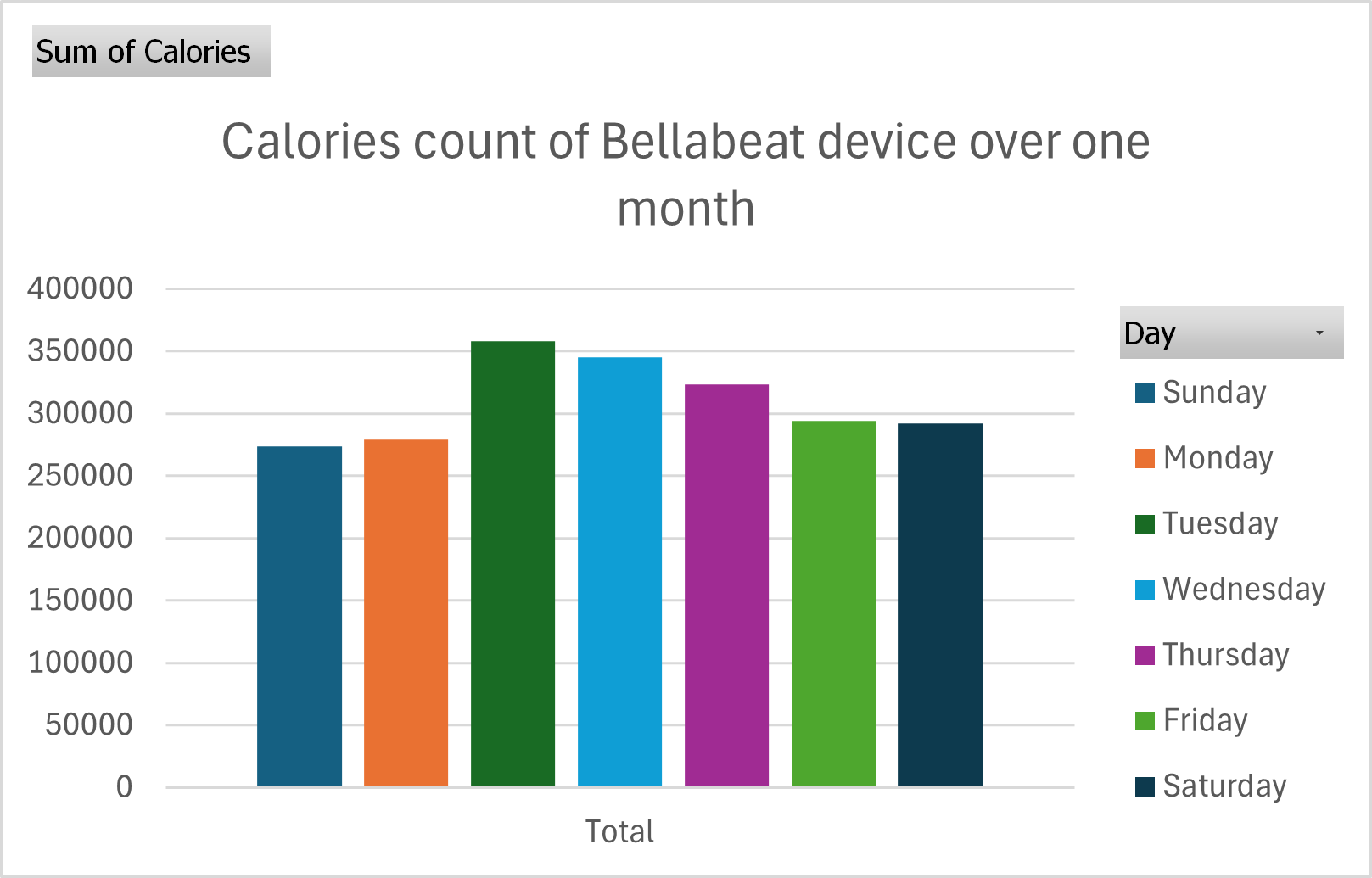
ggplot(data=hourly\_steps, mapping = aes(x=Hour, y=StepTotal)) +geom\_col(fill = 'orange') + theme(axis.text.x = element\_text(angle = 90)) +  
 labs(title = "Average Steps Taken per Hour of Day",  
 x = "Hour", y = "Average Steps")



Note: Evening 6 P.M, 7 P.M, 10 P.M steps taken are more. whereas morning 10 A.M and afternoon 12 P.M, 2 P.M steps taken are more.



The above pie chart shows average minutes spent on device by users. It is clear that 81% time users are idle that is not using the device. While 16% time users are lightly active, 2% very active and just 1% time fairly active.



The above graph tells about device usage day wise over one month in terms of calories count. During weekdays like Tuesday, Wednesday and Thursday users are burning more calories compare to weekends. Sunday counts minimum calories.

1. Recommendations:

* From the above analysis, it is drawn that average steps taken by users on daily basis are 7000 while recommended is 10000. So most of the users fall into light active category. We can recommend the notification from Bellabeat app that encourages participants to walk at least 10000 steps a day.
* On weekends, we can held competitions among users within app to complete challenges like walking, running or even eating healthy food. On completing the challenge, they will get discount for annual subscription. This will not only motivate users to stay active on weekends, but also increase annual subscription of Bellabeat app.
* There are only 8 participants recording their weight. So, weight and BMI tracking features need to be improved. So, more users are encouraged to track their weight and understand BMI to live a healthy lifestyle.
* Within app, control needs to be provided to users that allow them to record their sleep pattern and provide suggestions to improve their sleep at night.
* Occasional tips about healthy eating, meditation and healthy habits can be provided within app as a general guidance.