Case Study 2: How Can Bellabeat- *A Wellness Technology Company* Play It Smart?

By: Oluwatosin Joy Isaiah

Ask Phase

Introduction

Since its establishment in 2013, Bellabeat has expanded quickly into a high-tech manufacturer of health-focused products for women, they empower women to reconnect with themselves, unleash their inner strengths and be what they were meant to be. The Co-founder and Chief Creative Officer believes that analyzing fitness data from smart devices could help the organization discover new growth prospects. I've been requested to focus on one of Bellabeat' s products and examine data from smart devices to learn more about how consumers are using their smart gadgets. The company's marketing technique will then be influenced by the insights I discover.

As a Junior Data Analyst, I will be looking at some of the data from the smart devices to assist in identifying the new growth prospective and present my analysis and suggestions to the executive team. Below are the products:

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.

Key Stakeholders

The stakeholders are;

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.

Questions considered for Analysis

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

Prepare Phase

Data Source: I would be working on the FitBit Fitness Tracker Data (Public Domain, dataset made available through Mobius). This dataset contains personal fitness tracker from thirty fitbit users.

Data Info: The dataset is made up of 18 csv files. The data is organized in a long format as each subject has data in multiple rows.

Checking my data Credibility using ROCCC

Reliable: This dataset contains personal fitness tracker from 30 Fitbit users. This sample size is definitely small compared to millions of users. This makes it not reliable.

Original: Since we were unable to find the original source, the data came from a third party. Therefore, it cannot be trusted.

Comprehensive: It is challenging to analyze the data with regard to Bellabeat' s clients because the dataset lacks demographic data about the participants. And this makes it bias.

Current: This data is not a current data. Data is from March 2016 to May 2016.

Cited: Not cited

Tools Used

R Programming was used for cleaning, analyzing and visualizing the data provided.

Installing and loading packages

```
library(tidyverse)
library(ggpubr)
library(ggplot2)
library(here)
library(skimr)
library(janitor)
library(lubridate)
library(scales)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✓ ggplot2 3.3.6     ✓ purrr   0.3.5
## ✓ tibble  3.1.8     ✓ dplyr   1.0.10
## ✓ tidyr   1.2.1     ✓ stringr 1.4.1
## ✓ readr   2.1.3     ✓ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## here() starts at /kaggle/working

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

```
##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
##    date, intersect, setdiff, union

##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##    discard

##
## The following object is masked from 'package:readr':
##
##    col_factor
```

Importing dataset files

I will upload the following datasets to ensure completion of the business task given.

- Daily Calories
- Daily Intensities
- Daily Steps
- Sleep

```
daily_calories <- read.csv("../input/fitbit/Fitabase Data 4.12.16-5.12.16/dailyCalories_merged.csv")
daily_intensities <- read.csv("../input/fitbit/Fitabase Data 4.12.16-5.12.16/dailyIntensities_merged.csv")
daily_steps <- read.csv("../input/fitbit/Fitabase Data 4.12.16-5.12.16/dailySteps_merged.csv")
sleep <- read.csv("../input/fitbit/Fitabase Data 4.12.16-5.12.16/sleepDay_merged.csv")
```

Validation and verification of datasets

Let's examine the basic structure of our dataset after importing the relevant datasets into our working environment.

- **Daily Calories**

```
head(daily_calories)
colnames(daily_calories)
str(daily_calories)
```

A data.frame: 6 × 3

	Id	ActivityDay	Calories
	<dbl>	<chr>	<int>
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

'Id''ActivityDay''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 ...

- **Daily Intensities**

```
head(daily_intensities)
colnames(daily_intensities)
str(daily_intensities)
```

A data.frame: 6 × 10

	Id	ActivityDay	SedentaryMinutes	LightlyActiveMinutes	FairlyActiveMinutes
	<dbl>	<chr>	<int>	<int>	<int>
1	1503960366	4/12/2016	728	328	13
2	1503960366	4/13/2016	776	217	19
3	1503960366	4/14/2016	1218	181	11
4	1503960366	4/15/2016	726	209	34
5	1503960366	4/16/2016	773	221	10
6	1503960366	4/17/2016	539	164	20

VeryActiveMinutes	SedentaryActiveDistance	LightActiveDistance	ModeratelyActiveDistance	VeryActiveDistance
<int>	<dbl>	<dbl>	<dbl>	<dbl>
25	0	6.06	0.55	1.88
21	0	4.71	0.69	1.57
30	0	3.91	0.40	2.44
29	0	2.83	1.26	2.14
36	0	5.04	0.41	2.71
38	0	2.51	0.78	3.19

'Id'' ActivityDay'' SedentaryMinutes'' LightlyActiveMinutes''

FairlyActiveMinutes'' VeryActiveMinutes'' SedentaryActiveDistance''

LightActiveDistance''ModeratelyActiveDistance''VeryActiveDistance'

'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" "4/15/2016" ...

$ SedentaryMinutes : int 728 776 1218 726 773 539 1149 775 818 838 ...

$ 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 ...

- **Daily Steps**

```
head(daily_steps)
colnames(daily_steps)
str(daily_steps)
```

A data.frame: 6 × 3

	Id	ActivityDay	StepTotal
	<dbl>	<chr>	<int>
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

'Id''ActivityDay''StepTotal'

'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 9819 ...

- **Sleep**

```
head(sleep)
colnames(sleep)
str(sleep)
```

A data.frame: 6 × 5

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

'Id''SleepDay''TotalSleepRecords''TotalMinutesAsleep''TotalTimeInBed'

'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 12:00:00 AM" "4/13/2016 12:00:00 AM" "4/15/2016

12:00:00 AM" "4/16/2016 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 ...

Process Phase

Here, I would be documenting my data cleaning or manipulation process.

Checking for duplicates.

```
n_distinct(daily_calories$Id)
n_distinct(daily_intensities$Id)
n_distinct(daily_steps$Id)
n_distinct(sleep$Id)
```

```
33
```

```
33
```

```
33
```

```
24
```

Only 33 people recorded their daily activity, daily calories, daily intensities and daily steps while 24 only recorded their sleep day.

Formatting the data

```
daily_intensities$ActivityDay <- as.POSIXct(daily_intensities$Activity
Day, format="%m/%d/%Y", tz=Sys.timezone())
daily_calories$ActivityDay <- as.POSIXct(daily_calories$ActivityDay, f
ormat="%m/%d/%Y", tz=Sys.timezone())
daily_steps$ActivityDay <- as.POSIXct(daily_steps$ActivityDay, format
="%m/%d/%Y", tz=Sys.timezone())
daily_intensities <- daily_intensities[,-(7:10)]
head(daily_steps)
head(daily_intensities)
head(daily_calories)

sleep$SleepDay <- as.POSIXct(sleep$SleepDay, format="%m/%d/%Y %I:%M:%S
%p", tz=Sys.timezone())
sleep <- sleep[,-3]
head(sleep)
```

A data.frame: 6 × 3

	Id	ActivityDay	StepTotal
	<dbl>	<dttm>	<int>
1	1503960366	2016-04-12	13162
2	1503960366	2016-04-13	10735
3	1503960366	2016-04-14	10460
4	1503960366	2016-04-15	9762
5	1503960366	2016-04-16	12669
6	1503960366	2016-04-17	9705

A data.frame: 6 × 6

	Id	ActivityDay	SedentaryMinutes	LightlyActiveMinutes	FairlyActiveMinutes	VeryActiveMinutes
	<dbl>	<dttm>	<int>	<int>	<int>	<int>
1	1503960366	2016-04-12	728	328	13	25
2	1503960366	2016-04-13	776	217	19	21
3	1503960366	2016-04-14	1218	181	11	30
4	1503960366	2016-04-15	726	209	34	29
5	1503960366	2016-04-16	773	221	10	36
6	1503960366	2016-04-17	539	164	20	38

A data.frame: 6 × 3

	Id	ActivityDay	Calories
	<dbl>	<dttm>	<int>
1	1503960366	2016-04-12	1985
2	1503960366	2016-04-13	1797
3	1503960366	2016-04-14	1776
4	1503960366	2016-04-15	1745
5	1503960366	2016-04-16	1863
6	1503960366	2016-04-17	1728

A data.frame: 6 × 4

	Id	SleepDay	TotalMinutesAsleep	TotalTimeInBed
	<dbl>	<dttm>	<int>	<int>
1	1503960366	2016-04-12	327	346
2	1503960366	2016-04-13	384	407
3	1503960366	2016-04-15	412	442
4	1503960366	2016-04-16	340	367
5	1503960366	2016-04-17	700	712
6	1503960366	2016-04-19	304	320

Merging similar data

Looking at the data, daily_steps, daily_calories and daily_intensities share similar data. So before visualizing, I would be merging the data by creating a new dataset called daily.

```
daily <- full_join(daily_steps, daily_calories, by = c("Id" = "Id", "A
ctivityDay" = "ActivityDay"))
daily <- full_join(daily, daily_intensities, by = c("Id" = "Id", "Acti
vityDay" = "ActivityDay"))
head(daily)
```

A data.frame: 6 × 8

	Id	ActivityDay	StepTotal	Calories	SedentaryMinutes	LightlyActiveMinutes	FairlyActiveMinutes	VeryActiveMinutes
	<dbl>	<dttm>	<int>	<int>	<int>	<int>	<int>	<int>
1	1503960366	2016-04-12	13162	1985	728	328	13	25
2	1503960366	2016-04-13	10735	1797	776	217	19	21
3	1503960366	2016-04-14	10460	1776	1218	181	11	30
4	1503960366	2016-04-15	9762	1745	726	209	34	29
5	1503960366	2016-04-16	12669	1863	773	221	10	36
6	1503960366	2016-04-17	9705	1728	539	164	20	38

Next, I would be using the sleep day table as the primary key. It has the least numbers of respondents. Here I would be semi-joining with daily which is the new table created before I lastly make use of the inner join.

```
daily_semijoin <- semi_join(sleep, daily, by = c("Id" = "Id", "SleepDa
y" = "ActivityDay"))
head(daily_semijoin)
```

A data.frame: 6 × 4

	Id	SleepDay	TotalMinutesAsleep	TotalTimeInBed
	<dbl>	<dttm>	<int>	<int>
1	1503960366	2016-04-12	327	346
2	1503960366	2016-04-13	384	407
3	1503960366	2016-04-15	412	442
4	1503960366	2016-04-16	340	367
5	1503960366	2016-04-17	700	712
6	1503960366	2016-04-19	304	320

```
daily_completed <- inner_join(daily, daily_semijoin, by = c("Id" = "I
d", "ActivityDay" = "SleepDay"))
n_distinct(daily_completed$Id)
n_distinct(daily_completed$ActivityDay)
head(daily_completed)
```

24

31

A data.frame: 6 × 10

	Id	ActivityDay	StepTotal	Calories	SedentaryMinutes	LightlyActiveMinutes
	<dbl>	<dttm>	<int>	<int>	<int>	<int>
1	1503960366	2016-04-12	13162	1985	728	328
2	1503960366	2016-04-13	10735	1797	776	217
3	1503960366	2016-04-15	9762	1745	726	209
4	1503960366	2016-04-16	12669	1863	773	221
5	1503960366	2016-04-17	9705	1728	539	164
6	1503960366	2016-04-19	15506	2035	775	264

FairlyActiveMinutes	VeryActiveMinutes	TotalMinutesAsleep	TotalTimeInBed
<int>	<int>	<int>	<int>
13	25	327	346
19	21	384	407
34	29	412	442
10	36	340	367
20	38	700	712
31	50	304	320

Analyze Phase

Here, I would be examining Fitbit user trends to see whether they may influence BellaBeat's marketing strategy. Below is a quick summary statistic about the data frame named daily completed.

```
summary(daily_completed)
```

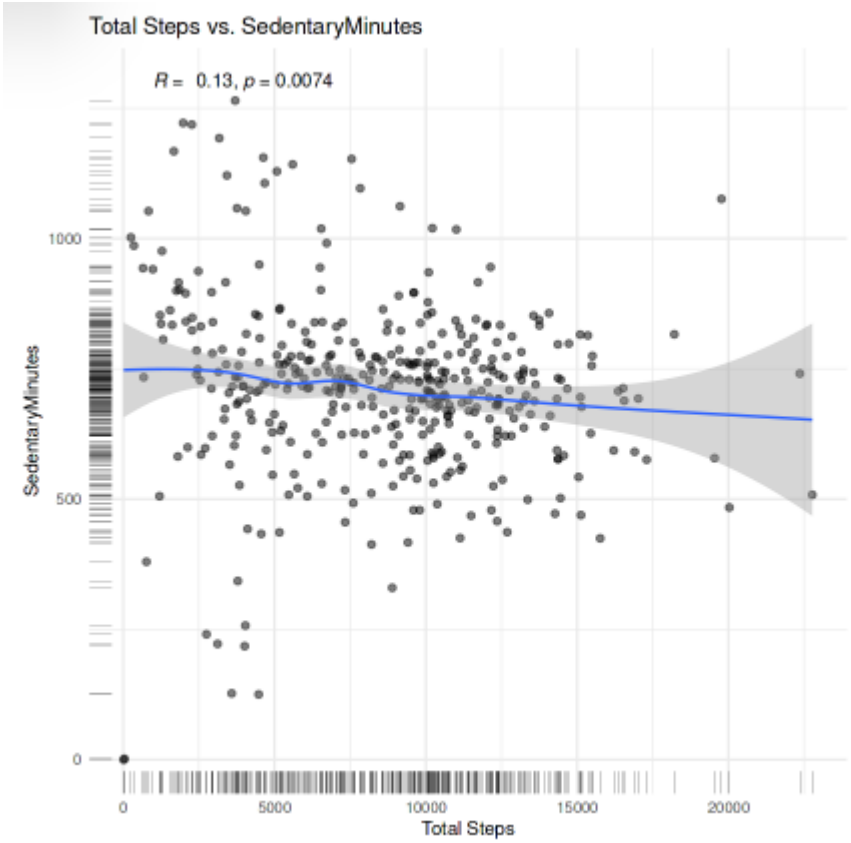
```
##       Id              ActivityDay                    StepTotal
##  Min.   :1.504e+09   Min.   :2016-04-12 00:00:00   Min.   :   17
##  1st Qu.:3.977e+09   1st Qu.:2016-04-19 00:00:00   1st Qu.: 5206
##  Median :4.703e+09   Median :2016-04-27 00:00:00   Median : 8925
##  Mean   :5.001e+09   Mean   :2016-04-26 12:40:05   Mean   : 8541
##  3rd Qu.:6.962e+09   3rd Qu.:2016-05-04 00:00:00   3rd Qu.:11393
##  Max.   :8.792e+09   Max.   :2016-05-12 00:00:00   Max.   :22770
## Calories    SedentaryMinutes LightlyActiveMinutes FairlyActiveMinutes ##  Min
.   : 257   Min.   :   0.0   Min.   :  2.0        Min.   :  0.00
##  1st Qu.:1850   1st Qu.: 631.0   1st Qu.:158.0        1st Qu.:  0.00
##  Median :2220   Median : 717.0   Median :208.0        Median : 11.00
##  Mean   :2398   Mean   : 712.2   Mean   :216.9        Mean   : 18.04
##  3rd Qu.:2926   3rd Qu.: 783.0   3rd Qu.:263.0        3rd Qu.: 27.00
##  Max.   :4900   Max.   :1265.0   Max.   :518.0        Max.   :143.00

##  VeryActiveMinutes TotalMinutesAsleep TotalTimeInBed
##  Min.   :  0.00    Min.   : 58.0      Min.   : 61.0
##  1st Qu.:  0.00    1st Qu.:361.0      1st Qu.:403.0
##  Median :  9.00    Median :433.0      Median :463.0
##  Mean   : 25.19    Mean   :419.5      Mean   :458.6
##  3rd Qu.: 38.00    3rd Qu.:490.0      3rd Qu.:526.0
##  Max.   :210.00    Max.   :796.0      Max.   :961.0
```

Share Phase

Here I would be creating my own data visualizations. Key observations and suggestions are clearly communicated through the visuals.

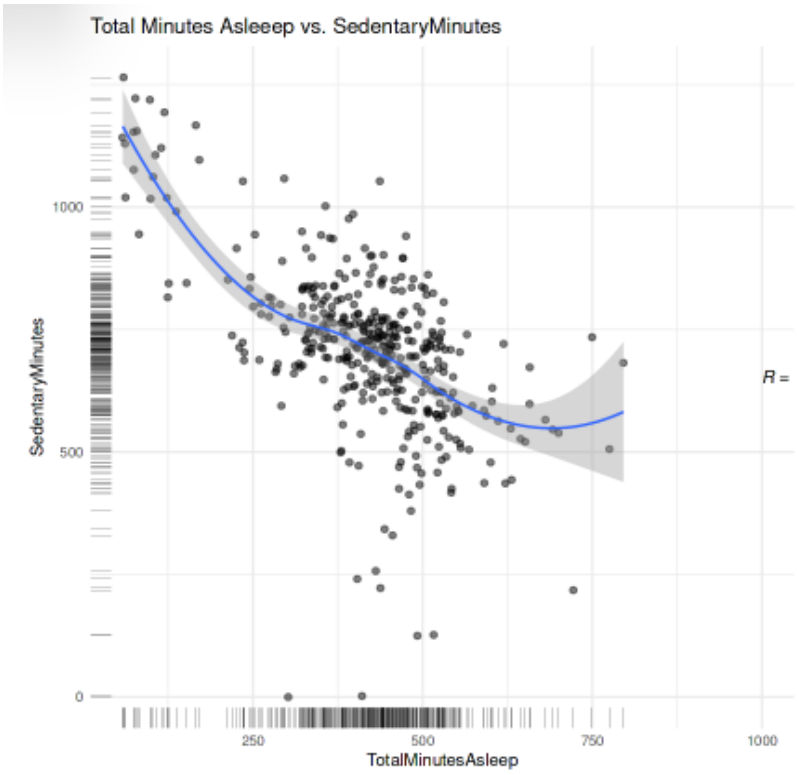
```
ggplot(daily_completed,aes(StepTotal,SedentaryMinutes))+geom_jitter(al
pha=.5)+
    geom_rug(position="jitter", size=.08)+
    geom_smooth(size =.6)+
    stat_cor(method = "pearson", label.x = 1000, label.y = 1300)+
    labs(title= "Total Steps vs. SedentaryMinutes", x= "Total Steps",
y="SedentaryMinutes")+
    theme_minimal()
```

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



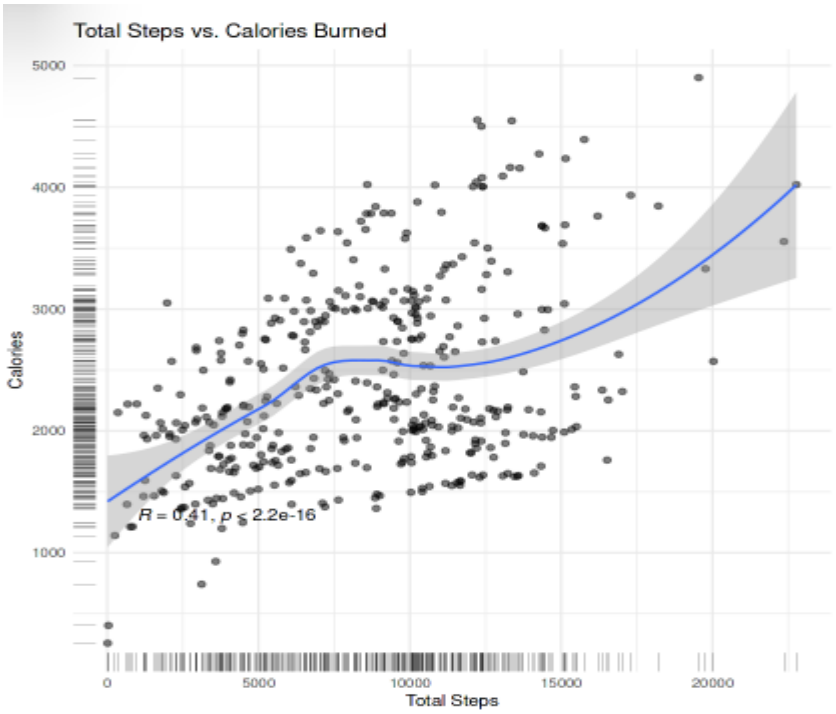

```
ggplot(daily_completed,aes(TotalMinutesAsleep,SedentaryMinutes))+geom_
jitter(alpha=.5)+
    geom_rug(position="jitter", size=.08)+
    geom_smooth(size =.6)+
    stat_cor(method = "pearson", label.x = 1000, label.y = 650)+
    labs(title= "Total Minutes Asleeep vs. SedentaryMinutes", y= "Sede
ntaryMinutes", x="TotalMinutesAsleep")+
    theme_minimal()
```

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



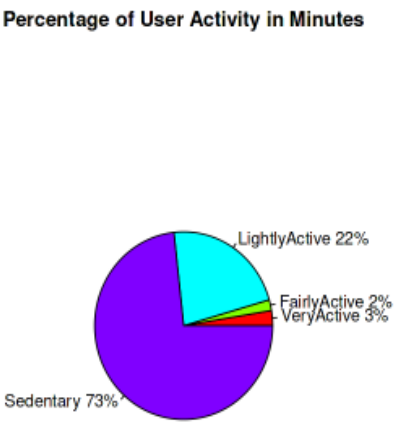

```
ggplot(daily_completed,aes(StepTotal,Calories))+geom_jitter(alpha=.5)+
    geom_rug(position="jitter", size=.08)+
    geom_smooth(size =.6)+
    stat_cor(method = "pearson", label.x = 1000, label.y = 1300)+
    labs(title= "Total Steps vs. Calories Burned", x= "Total Steps", y
="Calories")+
    theme_minimal()
```

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



There is a wide variety centered towards the lower quantities, but it is quite obvious that persons who moved around the most tended to burn the most calories. So, there is a positive correlation between total calories burned and total steps taken. However Sedentary minutes and overall sleep minutes are inversely correlated. This implies that the amount of time spent sleeping decreases as sedentary minutes rise.

```
VeryActive_Minutes <- sum(daily_completed$VeryActiveMinutes)
FairlyActive_Minutes <- sum(daily_completed$FairlyActiveMinutes)
LightlyActive_Minutes <- sum(daily_completed$LightlyActiveMinutes)
Sedentary_Minutes <- sum(daily_completed$SedentaryMinutes)
Total_Minutes <- VeryActive_Minutes + FairlyActive_Minutes + LightlyAc
tive_Minutes + Sedentary_Minutes

slices <- c(VeryActive_Minutes, FairlyActive_Minutes, LightlyActive_Mi
nutes, Sedentary_Minutes)
lbls <- c("VeryActive","FairlyActive","LightlyActive","Sedentary")
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls, pct)
lbls <- paste(lbls, "%", sep="")
pie(slices, labels = lbls, col = rainbow(length(lbls)), main = "Percen
tage of User Activity in Minutes")
```



The graph above depicts how much time an individual spent throughout the survey period being very active, lightly active, lightly active, and not energetic enough (sedentary).

Conclusion

- The highest number of steps users take in a day is 8925, which is less than the 10,000 that the Centers for Disease Control and Prevention (CDC) advise.
- Individuals have expended 2,398 calories each day on average. which is only little more than the maximum allowable daily calorie intake of 2,200. In other words, they should ideally be losing weight unless they are ingesting more calories than 2,400 per day.
- There is too much sedentary time on the average and unquestionably has to be decreased with an effective marketing plan.
- Also, the vast majority of individuals engage in light activity with a lot of sedentary/idle time(73%).
- Respondents sleep time is thought to be an estimate of 7 hours.

- In Conclusion, Users were not constantly logging their data, and some people who were consistently logging their data did not see weight loss or other results over the course of the data collection.

Recommendations

- Bellabeat can create a marketing plan that focuses on providing advice for leading a healthy lifestyle, which includes being active every day, spending less time doing sedentary activities, walking more each day, and getting enough sleep.
- Due to the relationship between the appropriate amount of sleep and daily activities, the company might add features to their devices, such as reminders to log weight information and notifications to go to sleep or wake up based on customized data of customers, to assure the ideal timings specifically.