Google Data Analytics Capstone Case Study

Kevin Kent Ventura

2022-03-16

Project: Bellabeat

How can a Wellness Technology Company Play It Smart?

#Introduction

This is part of Course 8: Google Data Analytics: Capstone in relation to Google Professional Data Analytics Certificate. In this case study, I am a junior data analyst working on the marketing analyst team at Bellabeat, a high-tech manufacturer of health-focused products for women. Bellabeat is a successful small company, but they have the potential to become a larger player in the global smart device market.

In this case-study, I will follow the 6 steps of Data Analysis process in which in this course tackle from Course 1 to 7. This includes: *Ask*, *Prepare*, *Process*, *Analyze*, *Share*, *Act.*

PHASE 1: ASK

1.0 Guiding Questions for this phase:

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

1.1 Business Task:

- Analyze Fitbit fitness tracker data to gain insights into how consumers are using the FitBit app and discover trends for Bellabeat marketing strategy.

1.2 Business Objectives

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

1.3 Deliverables:

- A clear summary of the business task
- A description of all data sources used

- Documentation of any cleaning or manipulation of data
- A summary of analysis
- Supporting visualizations and key findings
- High-level content recommendations based on the analysis

1.4 Key Stakeholders:

- Urška Sršen: Bellabeat's co-founder and Chief Creative Officer
- Sando Mur: Mathematician, Bellabeat's co-founder and key member of the Bellabeat executive team
- Bellabeat marketing analytics team: A team of data analysts guiding Bellabeat's marketing strategy.

PHASE2: PREPARE

2.1 About the Data Source and where it is stored:

- Data is publicly available on Kaggle: FitBit Fitness Tracker Data and stored in 18 csv files.
- Thirty eligible Fitbit users consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring.
- Data collected from April to May 2016
- Data collected includes information about daily activity, steps, and heart rate that can be used to explore users' habits.

2.2 How is the data organized? Is it in long or wide format?

- The provided from source is in a long format. Each file contains different type of activity recorded in a CSV format.

2.3 Are there issues with bias or credibility in this data?

- Reliable: The provided as not that reliable since it only has 30 respondents and data is in short timescale with only 1 month of records.
- Originality: Third part provider (Amazon Mechanical Turk)
- Comprehensive: Mostly match parameters with Bellabeat products.
- Current: The data collected was 6 years ago and it may be outdated for current trends
- Cited: Data collected from third part has limited information.

2.4 Are there any problems with the data?:

- A sample size of thirty FitBit users may not represent the entire fitness population and age.
- Data is 6 years old, relevance of the data may be an issue.

2.5 Data Selection

The following file is selected and copied for analysis.

- dailyActivity_merged.csv
- sleepDay_merged.csv

2.6 Tool

- We are going to use Excel for viewing

- R programming for data cleaning, transformation, and visualization

PHASE 3: PROCESS

- Explore and observe data
- Merging Data Sets
- Check for null values and missing values
- Transform data and format data type
- Perform preliminary statistical analysis

3.1 Importing Relevant Files for Data Analysis

#installing packages

#Remove”#” for actual code chunk

```
#install.packages("tidyverse")
#install.packages("dplyr")
#install.packages("tidyr")
#install.packages("ggplot2")
#library(tidyverse)
#library(dplyr)
#library(tidyr)
#library(ggplot2)
```

3.2 Importing Relevant Datasets

```
daily_activity <- read.csv("C:/Users/Kent/Desktop/Coursera/DATA ANALYTICS/COURSE 8_Google Data Analytic
sleepday_merged <- read.csv("C:/Users/Kent/Desktop/Coursera/DATA ANALYTICS/COURSE 8_Google Data Analyti
```

3.3 Checking the Data Structure

```
head(daily_activity)
```

```
##           Id ActivityDate TotalSteps TotalDistance TrackerDistance
## 1 1503960366   04/12/2016      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(sleepday_merged)
```

```
##           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(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"
```

```
colnames(sleepday_merged)
```

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

```
str(daily_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  "04/12/2016" "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(sleepday_merged)
```

```
## '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"
## $ 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 ...
```

3.4 Data Transformation

- Observe and Identify which columns is relevant for analysis

Creating new columns and transforming columns

For *daily_activity* table

#We will create a new columns for better understanding

- New column *Total_Active_Minutes* by adding [VeryActiveMinutes + FairlyActiveMinutes + LightlyActiveMinutes + SedentaryMinutes]

```
daily_activity$Total_Active_Minutes <- daily_activity$VeryActiveMinutes + daily_activity$FairlyActiveMi
                           daily_activity$LightlyActiveMinutes + daily_activity$SedentaryMinutes
```

- New column *Total_Active_Hours*

```
daily_activity$Total_Active_Hours <- round(daily_activity$Total_Active_Minutes/60)
```

- And lastly, new column *Dates*, and change the format to (*M/D/Y*)

```
daily_activity$Dates <- as.Date(daily_activity$ActivityDate, "%m/%d/%Y")
```

- Renaming columns in *daily_activity*

```
names(daily_activity) <- c("Id", "Activity_Date", "Total_Steps", "Total_Distance", "Tracker_Distance",
                           "Very_Active_Distance", "Moderately_Active_Distance", "Light_Active_Distance
                           "Very_Active_Minutes", "Fairly_Active_Minutes", "Lightly_Active_Minutes", "S
                           "Total_Active_Hours", "Total_Active_Mintues", "Dates")
```

For sleepDay_merged table

#We will create a new columns for better understanding

- New column *Total_Hours_Asleep*

```
sleepday_merged$Total_Hours_Asleep <- round(sleepday_merged$TotalMinutesAsleep/60)
```

- New column *Dates*

```
sleepday_merged$Dates <- as.Date(sleepday_merged$SleepDay, "%m/%d/%Y")
```

- Renaming columns in *sleepDay_merged* and change the format to (*M/D/Y*)

```
names(sleepday_merged) <- c("Id", "Sleep_Day", "Total_Sleep_Records", "Total_Minutes_Asleep", "Total_Ti
                            "Dates")
```

```
str(sleepday_merged)
```

```
## 'data.frame':    413 obs. of  7 variables:
##  $ Id                  : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ Sleep_Day           : chr  "4/12/2016 12:00:00 AM" "4/13/2016 12:00:00 AM" "4/15/2016 12:00:00 AM
##  $ Total_Sleep_Records : int  1 2 1 2 1 1 1 1 1 1 ...
##  $ Total_Minutes_Asleep: int  327 384 412 340 700 304 360 325 361 430 ...
##  $ Total_Time_In_Bed   : int  346 407 442 367 712 320 377 364 384 449 ...
##  $ Total_Hours_Asleep  : num  5 6 7 6 12 5 6 5 6 7 ...
##  $ Dates               : Date, format: "2016-04-12" "2016-04-13" ...
```

Creating new table and adding relevant columns

- We will create new table for *daily_activity* as *daily_activity_new* and adding relevant columns [Id, Dates, Total_Steps, Total_Distance, Total_Active_Hours, Calories]

#Remove "#" for actual code chunk

```
#daily_activity_new <- daily_activity %>%
  #select(Id, Dates, Total_Steps, Total_Distance, Total_Active_Hours, Calories)
```

```
summary(daily_activity_new)
```

```
##        Id                 Dates            Total_Steps    Total_Distance
##  Min.   :1.504e+09   Length:940         Min.   :    0   Min.   : 0.000
##  1st Qu.:2.320e+09   Class :character   1st Qu.: 3790   1st Qu.: 2.620
##  Median :4.445e+09   Mode  :character   Median : 7406   Median : 5.245
##  Mean   :4.855e+09                      Mean   : 7638   Mean   : 5.490
##  3rd Qu.:6.962e+09                      3rd Qu.:10727   3rd Qu.: 7.713
##  Max.   :8.878e+09                      Max.   :36019   Max.   :28.030
##  Total_Active_Hours    Calories
##  Min.   :   2.0     Min.   :   0
##  1st Qu.: 989.8     1st Qu.:1828
##  Median :1440.0     Median :2134
##  Mean   :1218.8     Mean   :2304
##  3rd Qu.:1440.0     3rd Qu.:2793
##  Max.   :1440.0     Max.   :4900
```

- We will also create new table for *sleepDay_merged* as *sleepDay_merged_new* and adding relevant columns [Id, Dates,Total_Hours_Asleep]

#Remove “#” for actual code chunk

```
#sleepday_merged_new <- sleepday_merged %>%
  #select(Id, Dates, Total_Hours_Asleep)
```

```
summary(sleepday_merged_new)
```

```
##        Id                Dates           Total_Hours_Asleep
##  Min.   :1.504e+09   Length:413         Min.   : 1.000
##  1st Qu.:3.977e+09   Class :character   1st Qu.: 6.000
##  Median :4.703e+09   Mode  :character   Median : 7.000
##  Mean   :5.001e+09                      Mean   : 6.995
##  3rd Qu.:6.962e+09                      3rd Qu.: 8.000
##  Max.   :8.792e+09                      Max.   :13.000
```

- Merging the new tables

#Remove “#” for actual code chunk

```
#merged_data <- daily_activity_new %>%  #left_join(sleepday_merged_new)
```

- Lets use *summarize ()* & str() to analyze the new merged_data specifics

```
summary(merged_data)
```

```
##        Id                Dates            Total_Steps    Total_Distance
##  Min.   :1.504e+09   Length:410         Min.   :   17   Min.   : 0.010
##  1st Qu.:3.977e+09   Class :character   1st Qu.: 5189   1st Qu.: 3.592
##  Median :4.703e+09   Mode  :character   Median : 8913   Median : 6.270
##  Mean   :4.995e+09                      Mean   : 8515   Mean   : 6.012
##  3rd Qu.:6.962e+09                      3rd Qu.:11370   3rd Qu.: 8.005
##  Max.   :8.792e+09                      Max.   :22770   Max.   :17.540
##  Total_Active_Hours    Calories     Total_Hours_Asleep
##  Min.   :   2.0     Min.   : 257   Min.   : 1.00
##  1st Qu.: 906.2     1st Qu.:1841   1st Qu.: 6.00
##  Median : 983.0     Median :2207   Median : 7.00
##  Mean   : 971.6     Mean   :2389   Mean   : 6.99
##  3rd Qu.:1042.0     3rd Qu.:2920   3rd Qu.: 8.00
##  Max.   :1398.0     Max.   :4900   Max.   :13.00
```

```
str(merged_data)
```

```
## 'data.frame':    410 obs. of  7 variables:
##  $ Id                : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ Dates             : chr  "2016-04-12" "2016-04-13" "2016-04-15" "2016-04-16" ...
##  $ Total_Steps       : int  13162 10735 9762 12669 9705 15506 10544 9819 14371 10039 ...
##  $ Total_Distance    : num  8.5 6.97 6.28 8.16 6.48 ...
##  $ Total_Active_Hours: int  1094 1033 998 1040 761 1120 1063 1076 1056 991 ...
##  $ Calories          : int  1985 1797 1745 1863 1728 2035 1786 1775 1949 1788 ...
##  $ Total_Hours_Asleep: int  5 6 7 6 12 5 6 5 6 7 ...
```

- Removing *duplicates* and *NA*

#Remove "#" for the actual code chunk

```
#merged_data <- distinct(merged_data)
#remove any duplicates

#merged_data <- drop_na(merged_data)
#Remove missing data
```

- Analyzing the data

#Remove "#" for the actual code chunk

```
#merged_data %>%
  #select(Total_Steps, Total_Active_Hours, Total_Distance, #Total_Hours_Asleep, Calories) %>%
  summary(merged_data)
```

```
##        Id                Dates            Total_Steps    Total_Distance
##  Min.   :1.504e+09   Length:410         Min.   :   17   Min.   : 0.010
##  1st Qu.:3.977e+09   Class :character   1st Qu.: 5189   1st Qu.: 3.592
##  Median :4.703e+09   Mode  :character   Median : 8913   Median : 6.270
##  Mean   :4.995e+09                      Mean   : 8515   Mean   : 6.012
##  3rd Qu.:6.962e+09                      3rd Qu.:11370   3rd Qu.: 8.005
##  Max.   :8.792e+09                      Max.   :22770   Max.   :17.540
##  Total_Active_Hours    Calories    Total_Hours_Asleep
##  Min.   :   2.0     Min.   : 257   Min.   : 1.00
##  1st Qu.: 906.2     1st Qu.:1841   1st Qu.: 6.00
##  Median : 983.0     Median :2207   Median : 7.00
##  Mean   : 971.6     Mean   :2389   Mean   : 6.99
##  3rd Qu.:1042.0     3rd Qu.:2920   3rd Qu.: 8.00
##  Max.   :1398.0     Max.   :4900   Max.   :13.00
```

PHASE4: ANALYZE

4.1 Performing Calculations

Will analyze the statistics of the data that we manipulated

- The average total steps the users logged 8,514 steps or 6.01 km which is not far from the recommended. According to CDC,For general fitness, most adults should aim for 10,000 steps per day. This figure may rise or fall depending on a person's age, current fitness level, and health goals. Source: Medical News Today
- The more walks taken by each user the more calories are burned.
- the more active a user, the more calories they are going to burn.

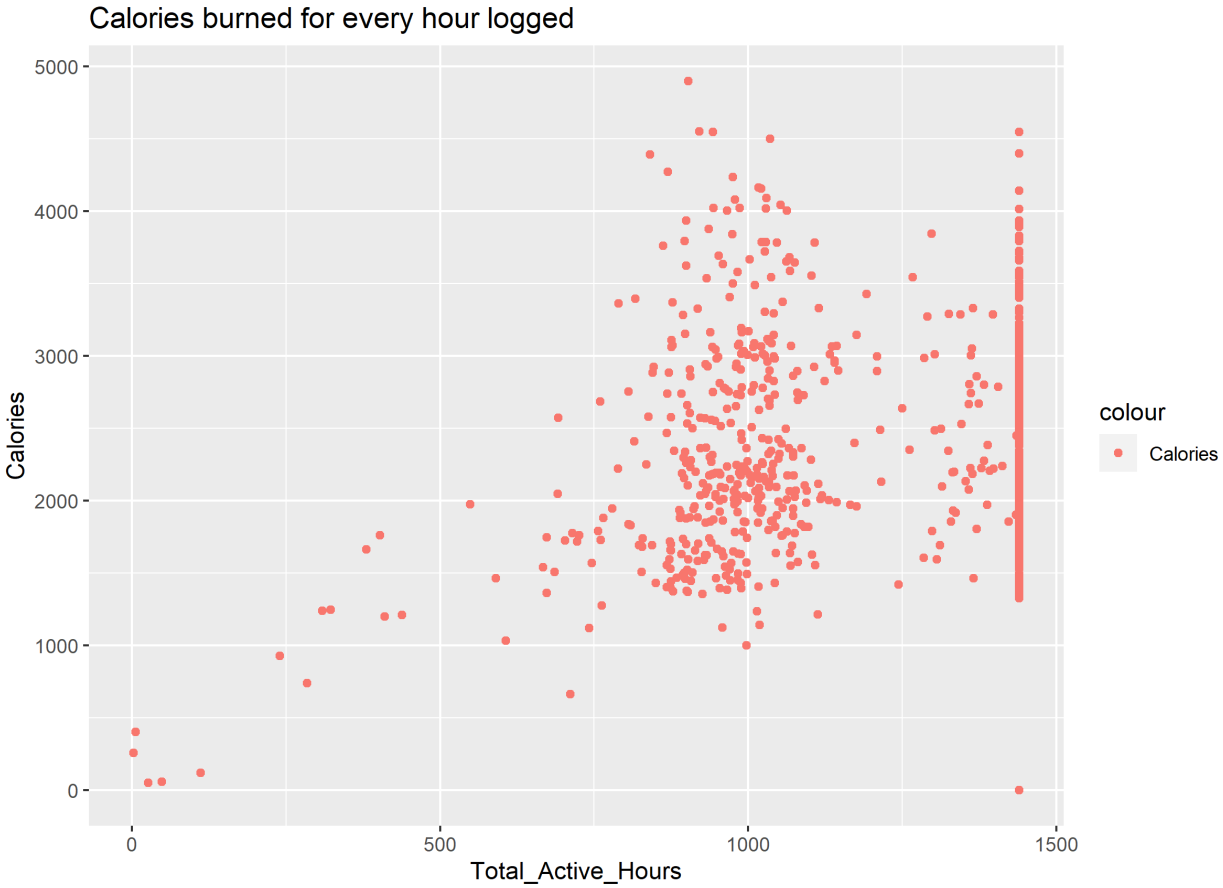
PHASE5: SHARE

- We will create a visualizations from the data and we will use the data to visualize the valuable insights to our stake holders

5.1 Calories Burned for every hour logged time

#Remove "#" for the actual code

```
#ggplot(data = merged_data) +
  #geom_point(mapping = aes(x = Total_Active_Hours, y = Calories, color = "Calories")) +
  #labs(title = "Calories burned for every hour logged")
```



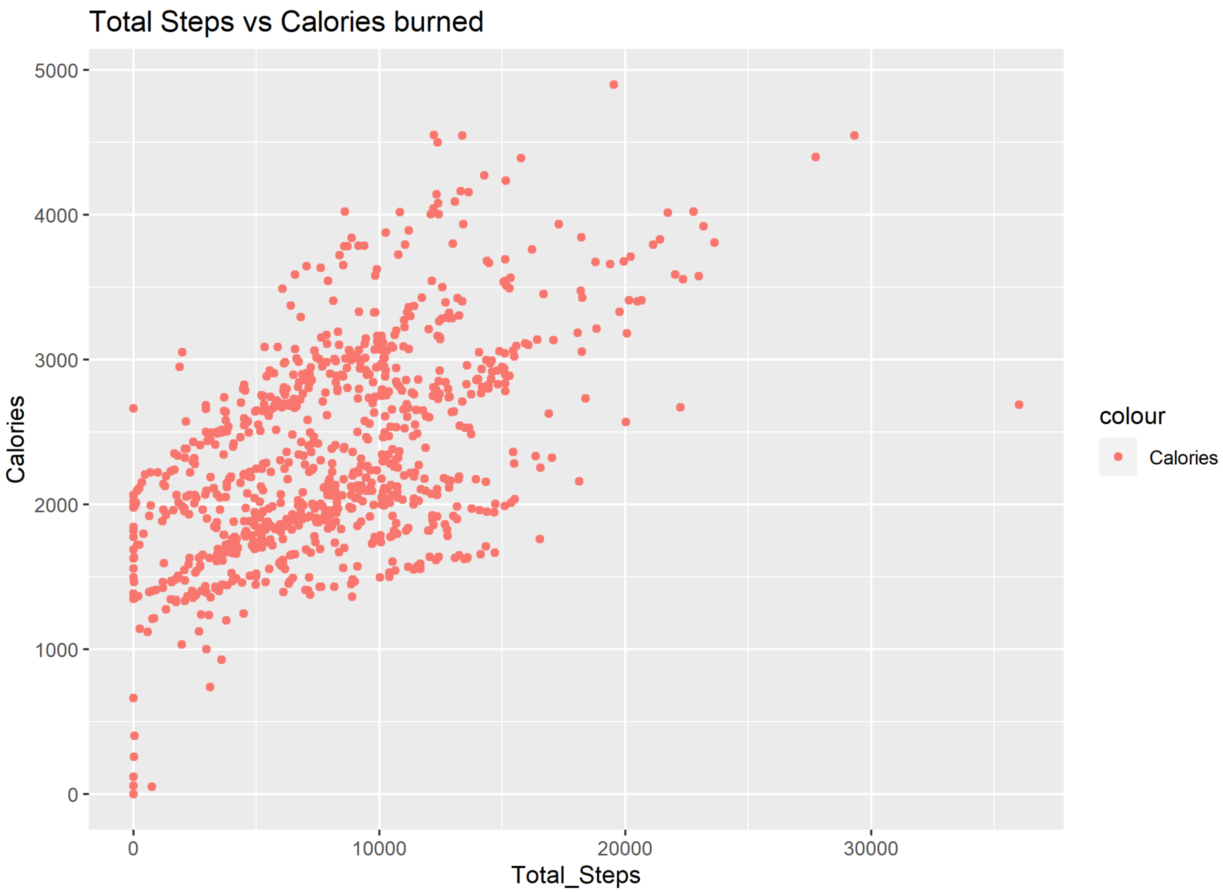
By analyzing the scatter plot using *GGplot*

- By looking at the graph, if logged hours are increasing then the number of calories burnt is also increased. This is mainly due to sedentary minutes.
- We can see the positive (weak) correlation here.
- An uncommon dot near "zero" in the Y axis means their are zero calories burned at 24 hours, which could be due to certain causes.

5.2 Number of Steps vs Calories Burned

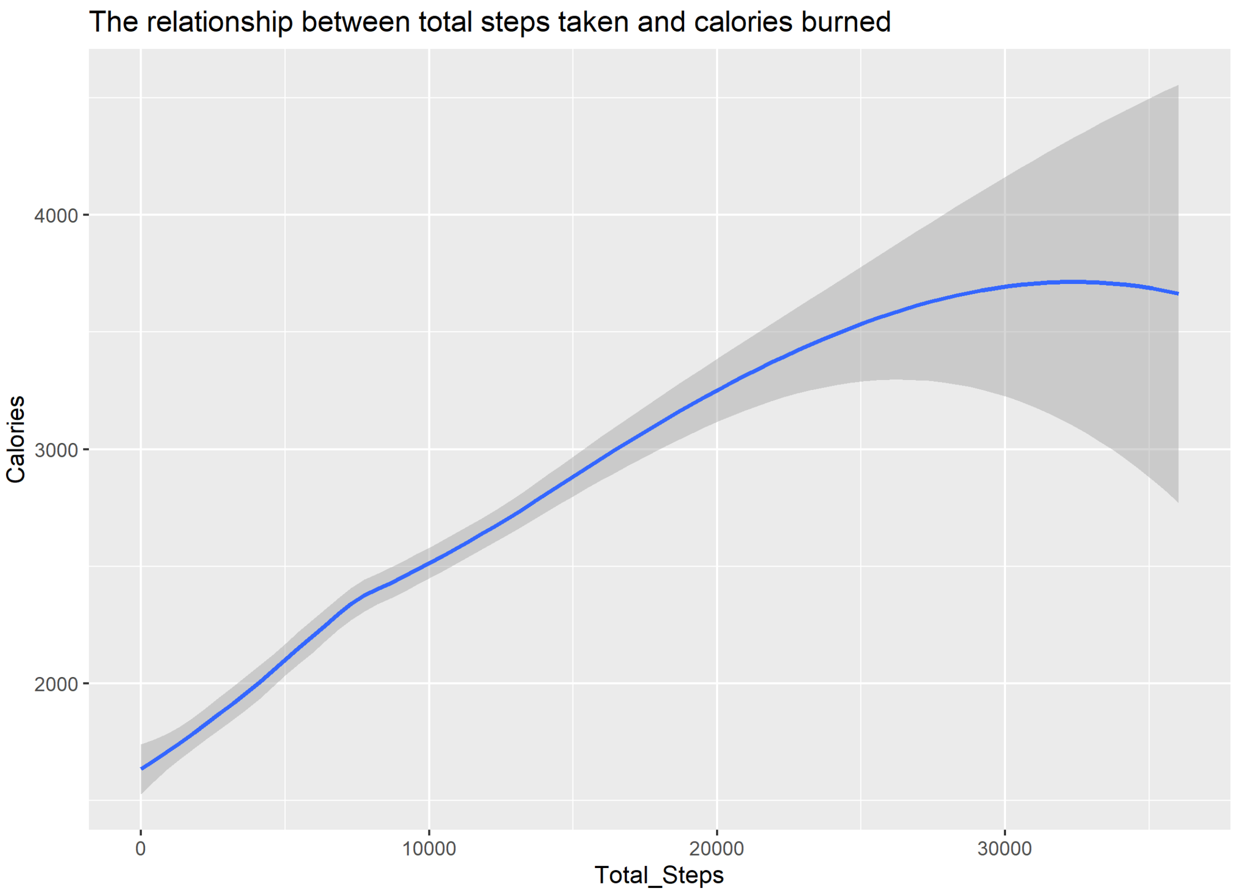
#Remove "#" for the actual code

```
#ggplot(data = merged_data) +
 #geom_point(mapping = aes(x = Total_Steps, y = Calories, color = "Calories")) +
 #labs(title = "Total Steps vs Calories burned")
```



#Remove "#" for the actual code

```
#ggplot(data = merged_data) +
  #geom_smooth(mapping = aes(x = Total_Steps, y = Calories)) +
  #labs(title = "The relationship between total steps taken and calories burned")
```



- By the looks of the graph we can say that it is positive correlations with limitations (*Scatterplot*)
- We can clearly see and identify the outliers that for 0 steps there is still data of calories burned.
- The longer they walk the higher the calories burned.

PHASE6: ACT

Based on our analysis backed by data and visualizations,

We will share with you the following actions with highly business recommendations.

1. The products on Bellabeat app should all collect data on calories burned per activity. It is important to include all the activity it can perform and record. also aggregate data for all calories combined together.
2. It has been observed that on average the users sleep 437.5 minutes per night. Bellabeat may recommend the users to set a target to 8 hours a day (480 minutes), and in case of meeting these targets share some motivation message, quote, or video with them.
3. On weekends, Bellabeat app can also prompt notifications to encourage users to exercise.
4. In order to encourage its users to adopt healthy sleeping, walking and workout practices and routines, Bellabeat can send periodic motivational quotes, videos and reminders to the users. Furthermore, the users can be encouraged to share their activities and results with friends and public to enable healthy competition in the user community.