Data Analytics Case Study with R - Sreemaee Akshathala

Introduction

This project is an unguided project as a part of the Google Data Analytics course's Capstone project. As part of the course, I will be performing a data analysis project for Bellabeat, a high-tech manufacturer of health-focused products for women. The objective of this project is to gain insights into how consumers are using Bellabeat's smart devices and provide recommendations for the company's marketing strategy.

To achieve this goal, I will be following the six stages of data analysis: ask, prepare, process, analyze, share, and act.

What are the six stages of Data Analysis?

The first stage, **Ask** involves asking relevant questions and identifying the key metrics that will help answer those questions. In the **prepare** stage, I will collect and organize the data required for the analysis. The **process** stage involves cleaning and transforming the data to ensure its accuracy and consistency. Next, in the **analyze** stage, I will use various data analysis techniques to uncover patterns and trends in the data. The insights gained from the analysis will then be **shared** with the Bellabeat executive team in the share stage, along with high-level recommendations for the company's marketing strategy. Finally, in the **act** stage, Bellabeat will use these insights to make informed business decisions that will enable them to achieve their growth objectives.

About the Company

Bellabeat is a high-tech company that creates health-focused smart products for women, founded by Urška Sršen and Sando Mur in 2013. They empower women with knowledge about their health and habits by collecting data on activity, sleep, stress, and reproductive health. Bellabeat has grown rapidly and positioned itself as a tech-driven wellness company for women with offices around the world and multiple products. They invest extensively in digital marketing, including Google Search, Facebook, Instagram, Twitter, Youtube, and the Google Display Network. Bellabeat uses beautifully designed technology to inform and inspire women around the world. They have a strong focus on data analysis to identify growth opportunities and improve their products. Bellabeat products are available through their own e-commerce channel and various online retailers. They have invested in traditional

advertising media, such as radio, out-of-home billboards, print, and television. Bellabeat is committed to promoting women's health and well-being through technology.

What is my role?

As a junior data analyst on the marketing analyst team, I have been tasked with focusing on one of Bellabeat's products and analyzing smart device data to gain insights into how consumers are using their smart devices. The objective of this analysis is to identify areas where Bellabeat can improve their product offerings, understand how users are interacting with their devices, and develop effective marketing strategies that will help the company grow.

The insights that will be gained from this analysis will be presented to the Bellabeat executive team, along with high-level recommendations for the company's marketing strategy. By leveraging smart device data, Bellabeat can better understand their customer's needs and preferences, which will enable the company to develop products that align with their customers' desires and position themselves for long-term growth in the global smart device market.

By following the six stages of data analysis, I aim to provide Bellabeat with a comprehensive understanding of how their customers are using their smart devices, which will help the company improve its product offerings, and develop effective marketing strategies to position themselves for long-term success in the global smart device market.

Key Objectives

1. Provide a concise overview of the business task at hand.
2. Identify and outline all the data sources utilized during the analysis.
3. Record any cleaning or data manipulation procedures that were performed.
4. Summarize the analysis performed, including any significant observations or insights.
5. Create visualizations to support the analysis and showcase key findings.
6. Develop high-level content recommendations based on the insights gleaned from the analysis.

Ask Phase

Business Task:

The main objective of this project is to analyze the Fitbit data and gain insights that will help guide Bellabeat's marketing strategy to become a major player in the global market.

Key Stakeholders

The executive team, including Urška Sršen and Sando Mur, are the key stakeholders interested in using these insights to identify opportunities for growth and improve their products.

Prepare Phase

During this phase, my goal is to obtain and import the dataset. Once I have done so, I will verify that the data is well-organized and trustworthy. Additionally, I will sort and filter the data as necessary to facilitate subsequent analysis.

Data sources used:

As an analyst for Bellabeat, Sršen has tasked me with exploring smart device users' daily habits using publicly available data. She has directed me to a specific data set called the **FitBit Fitness Tracker Data**, which is available on Kaggle and contains minute-level output for physical activity, heart rate, and sleep monitoring from thirty Fitbit users who consented to sharing their personal data. The data set includes information about daily activity, steps, and heart rate that I can use to gain insights into users' habits. Sršen has cautioned me that this data set may have some limitations, and she encourages me to consider adding another data set to help address these limitations as I delve deeper into the data analysis process.

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

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

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

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

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

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

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

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
install.packages("tidyr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

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

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

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

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

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

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

Loading the various libraries

```
install.packages("scales")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.1     ✔ 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_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
```

```
library(lubridate)
library(dplyr)
library(ggplot2)
library(tidyr)
library(here)
```

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

```
library(skimr)
library(janitor)
```

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

Importing the datasets

The Kaggle data set mentioned prior contains personal fitness tracker from thirty fitbitusers. Thirty eligible Fitbit users consented to the submission of personal tracker data,including minute-level output for physical activity,heart rate, and sleep monitoring. And Itincludes information about daily activity, steps, and heart rate that can be used toexplore users' habits.

About the dataset:

The dataset was generated by respondents to a distributed survey via AmazonMechanical Turk between 03.12.2016 and 05.12.2016. And include 18 CSV files.

Activity

```
Activity <- read.csv("/cloud/project/Fitabase Data 4.12.16-5.12.16/dailyActivity_merged
head(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(Activity)
```

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

```
str(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" "4/15/2016" ..
##  $ TotalSteps               : int  13162 10735 10460 9762 12669 9705 13019 15506 1054
##  $ 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 ...
##  $ 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 ...
##  $ 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
```

Calories

```
Calories <- read.csv("/cloud/project/Fitabase Data 4.12.16-5.12.16/dailyCalories_merged
head(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
```

```
colnames(Calories)
```

```
## [1] "Id"          "ActivityDay" "Calories"
```

```
str(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 ...
```

Sleep Minutes

```
sleep <- read.csv("/cloud/project/Fitabase Data 4.12.16-5.12.16/minuteSleep_merged.csv"
head(sleep)
```

```
##           Id                  date value       logId
## 1 1503960366 4/12/2016 2:47:30 AM     3 11380564589
## 2 1503960366 4/12/2016 2:48:30 AM     2 11380564589
## 3 1503960366 4/12/2016 2:49:30 AM     1 11380564589
## 4 1503960366 4/12/2016 2:50:30 AM     1 11380564589
## 5 1503960366 4/12/2016 2:51:30 AM     1 11380564589
## 6 1503960366 4/12/2016 2:52:30 AM     1 11380564589
```

```
colnames(sleep)
```

```
## [1] "Id"    "date"  "value" "logId"
```

```
str(sleep)
```

```
## 'data.frame':    188521 obs. of  4 variables:
##  $ Id   : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ date : chr  "4/12/2016 2:47:30 AM" "4/12/2016 2:48:30 AM" "4/12/2016 2:49:30 AM"
##  $ value: int  3 2 1 1 1 1 1 2 2 2 ...
##  $ logId: num  1.14e+10 1.14e+10 1.14e+10 1.14e+10 1.14e+10 ...
```

Steps

```
hourly_step <- read.csv("/cloud/project/Fitabase Data 4.12.16-5.12.16/dailySteps_merged
head(hourly_step)
```

```
##           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
```

Heartrate

```
Heart <- read.csv("/cloud/project/Fitabase Data 4.12.16-5.12.16/heartrate_seconds_merge
head(Heart)
```

```
##           Id                  Time Value
## 1 2022484408 4/12/2016 7:21:00 AM    97
## 2 2022484408 4/12/2016 7:21:05 AM   102
## 3 2022484408 4/12/2016 7:21:10 AM   105
## 4 2022484408 4/12/2016 7:21:20 AM   103
## 5 2022484408 4/12/2016 7:21:25 AM   101
## 6 2022484408 4/12/2016 7:22:05 AM    95
```

```
colnames(Heart)
```

```
## [1] "Id"    "Time"  "Value"
```

```
str(Heart)
```

```
## 'data.frame':    2483658 obs. of  3 variables:
##  $ Id   : num  2.02e+09 2.02e+09 2.02e+09 2.02e+09 2.02e+09 ...
##  $ Time : chr  "4/12/2016 7:21:00 AM" "4/12/2016 7:21:05 AM" "4/12/2016 7:21:10 AM"
##  $ Value: int  97 102 105 103 101 95 91 93 94 93 ...
```

Process Phase - Cleaning the Dataset

```
glimpse(Activity)
clean_names(Activity)
glimpse(Calories)
clean_names(Calories)
clean_names(sleep)
clean_names(Heart)
```

Formatting

```
sleep <- sleep[!duplicated(sleep), ]
sum(duplicated(sleep))
```

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

Analyze Phase - Summarising the Dataset

Some functions like the Pipe operator is unable to execute on the RStudio Cloud.

```
library(dplyr)

summary(select(Activity, TotalSteps, TotalDistance, SedentaryMinutes, Calories))
```

```
##    TotalSteps    TotalDistance    SedentaryMinutes    Calories
##  Min.   :    0   Min.   : 0.000   Min.   :   0.0   Min.   :   0
##  1st Qu.: 3790   1st Qu.: 2.620   1st Qu.: 729.8   1st Qu.:1828
##  Median : 7406   Median : 5.245   Median :1057.5   Median :2134
##  Mean   : 7638   Mean   : 5.490   Mean   : 991.2   Mean   :2304
##  3rd Qu.:10727   3rd Qu.: 7.713   3rd Qu.:1229.5   3rd Qu.:2793
##  Max.   :36019   Max.   :28.030   Max.   :1440.0   Max.   :4900
```

```
summary(select(Calories, Calories))
```

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

Key findings from this analysis :

1. Sedentary time is a concern: The average sedentary time among the participants is more than 16 hours, which is high and needs to be reduced with a good marketing strategy.
2. Light physical activity and high sedentary time: The majority of participants are lightly active, but they have a high sedentary time, which is not ideal for overall health.
3. Steps per day: The average total steps per day among participants is 7638, which is slightly lower than the recommended 8000 steps per day by the CDC. According to CDC research, taking 8000

steps per day was associated with a 51% lower risk for all-cause mortality, and taking 12000 steps per day was associated with a 65% lower risk compared to taking only 4000 steps.

Share and Act Phase - Data Visualization

We have a notion about how calories lost is correlated to the number of steps one takes in a day. We shall see how well that notion chalks out.

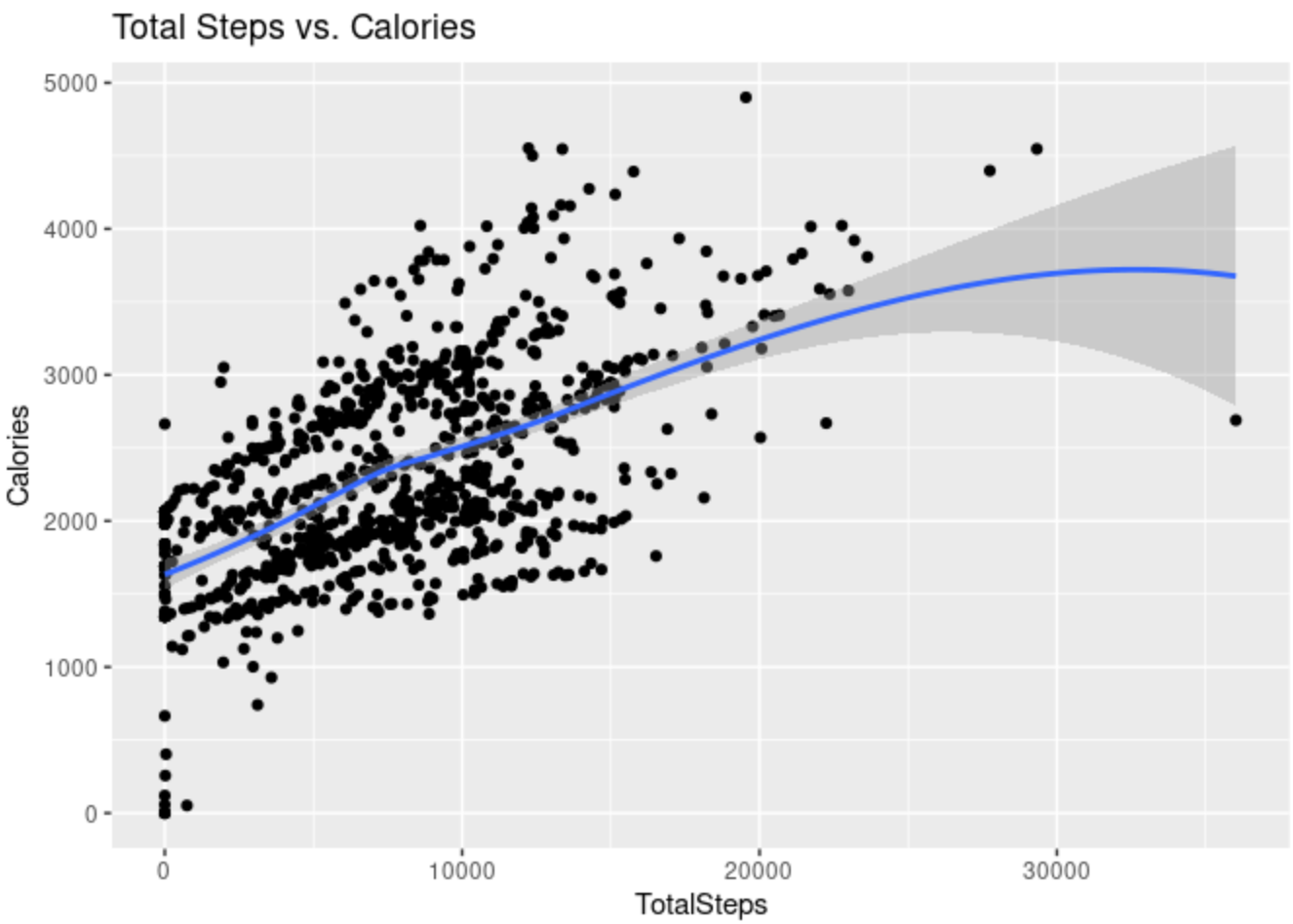
But again, It must be noted that the correlation between the number of steps taken and calories consumed depends on various factors such as age, weight, height, gender, and physical activity level. **In general, there is a positive correlation between the number of steps taken and calories burned**, as physical activity typically leads to an increase in energy expenditure.

However, the correlation may not be perfectly linear, as other factors such as diet, metabolism, and genetics can also influence the number of calories burned. Additionally, the relationship may be influenced by the type and intensity of physical activity, as some activities may burn more calories per step than others.

Calories versus steps plot

```
ggplot(data=Activity, aes(x=TotalSteps, y=Calories)) +
geom_point() + geom_smooth() + labs(title="Total Steps vs. Calories")
```

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



The above graph reinforces the positive correlation between steps and calories

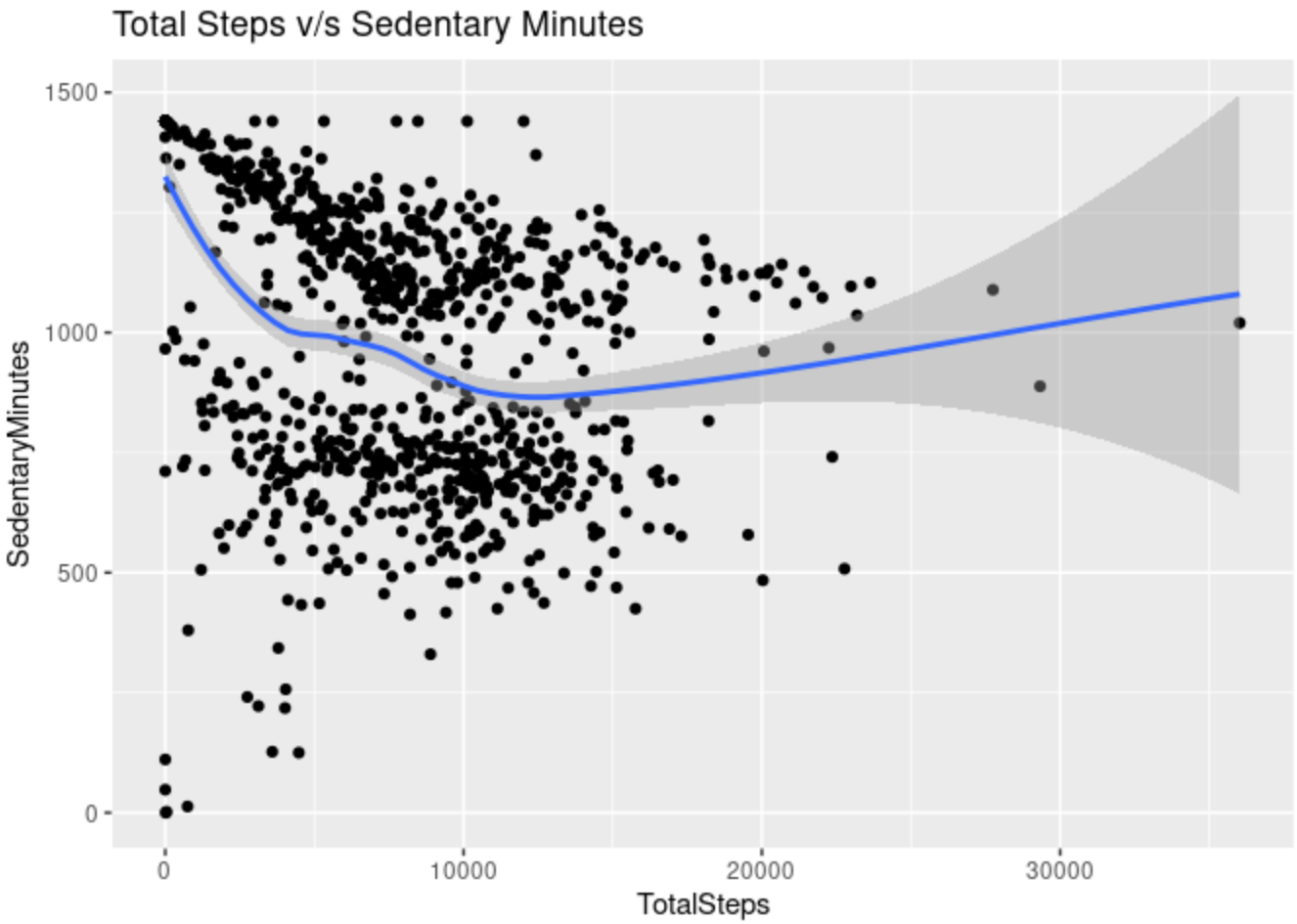
Total Steps v/s Sedentary minutes

This will give you a correlation coefficient between the two variables. A positive correlation coefficient indicates that as the total number of steps increases, the sedentary minutes decrease. A negative correlation coefficient would indicate the opposite. The human body is designed to move, and physical activity is crucial for maintaining good health. Sedentary behavior, on the other hand, involves very little physical movement and typically involves sitting or reclining.igned to move, and physical activity is crucial for maintaining good health. Sedentary behavior, on the other hand, involves very little physical movement and typically involves sitting or reclining.

Therefore, it's important to find a balance between physical activity and sedentary behavior. This is where tracking the number of steps taken and sedentary minutes can be helpful. By monitoring these two variables, individuals can gain insight into their daily activity levels and make adjustments to reduce their sedentary time and increase their physical activity.

```
library(ggplot2)
ggplot(data=Activity, aes(x=TotalSteps, y=SedentaryMinutes)) + geom_point() + geom_smoo
```

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



Insights and Takeaways from the analysis

Summary

Based on the analysis of FitBit Fitness Tracker Data, it is evident that collecting data on activity, sleep, stress, etc. can empower customers with knowledge about their health and daily habits. Bellabeat is a tech-driven wellness company that is rapidly growing and positioning itself as a prominent player in the industry. The analysis of the data revealed insights that can inform Bellabeat's marketing strategy.

Target audience:

Based on the analysis of FitBit Fitness Tracker Data, it is evident that collecting data on activity, sleep, stress, etc. can empower customers with knowledge about their health and daily habits. Bellabeat is a tech-driven wellness company that is rapidly growing and positioning itself as a prominent player in the industry. The analysis of the data revealed insights that can inform Bellabeat's marketing strategy.

Takeaways and reccomendations

Bellabeat has the potential to become a tech-driven wellness company that empowers its customers with knowledge about their own health and daily habits. Based on the analysis of the FitBit Fitness Tracker Data set, it is recommended that the company focuses on a target audience consisting of full-time working individuals who spend a lot of time sitting in front of a computer or in the office and require fitness and daily activities to stay in shape.

To improve, Bellabeat should focus on becoming a comprehensive wellness solution for its target audience, which includes people with sedentary jobs and a need to balance their personal and professional lives with healthy habits. The company should leverage the insights from the FitBit Fitness Tracker Data set to offer personalized and engaging fitness programs, stress-reduction techniques, sleep quality tracking, and healthy nutrition tips.

The app should also provide continuous motivation and support through gamification, challenges, social connections, and rewards. In addition, Bellabeat should invest in user research and feedback mechanisms to improve the user experience and address any pain points or barriers to adoption. Finally, the company should establish partnerships with other wellness brands, healthcare providers, and employers to expand its reach and impact.