2022



PROFILE:

<https://www.linkedin.com/in/anas-aljarrah/>

Data analytics project

Anas al jarrah

Bellabeat | Coursera - Google

Data Analytics Case Study with R – Case 2 Project

# ***Case Study 2:***

# **How Can a Wellness Technology Company Play It Smart?**



**Company**

Urška Sršen and Sando Mur founded Bellabeat, a high-tech company that manufactures health-focused smart products. Sršen used her background as an artist to develop beautifully designed technology that informs and inspires women around the world. Collecting data on activity, sleep, stress, and reproductive health has allowed Bellabeat to empower women with knowledge about their own health and habits. Since it was founded in 2013, Bellabeat has grown rapidly and quickly positioned itself as a tech-driven wellness company for women. By 2016, Bellabeat had opened offices around the world and launched multiple products. Bellabeat products became available through a growing number of online retailers in addition to their own e-commerce channel on their website. The company has invested in traditional advertising media, such as radio, out-of-home billboards, print, and television, but focuses on digital marketing extensively. Bellabeat invests year-round in Google Search, maintaining active Facebook and Instagram pages, and consistently engages consumers on Twitter. Additionally, Bellabeat runs video ads on YouTube and display ads on the Google Display Network to support campaigns around key marketing dates. Sršen knows that an analysis of Bellabeat’s available consumer data would reveal more opportunities for growth. She has asked the marketing analytics team 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, she would like high-level recommendations for how these trends can inform Bellabeat marketing strategy

**Scenario**

You are 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. Urška Sršen, cofounder and Chief Creative Officer of Bellabeat, believes that analyzing smart device fitness data could help unlock new growth opportunities for the company. You have been asked to focus on one of Bellabeat’s products and analyze smart device data to gain insight into how consumers are using their smart devices. The insights you discover will then help guide marketing strategy for the company. You will present your analysis to the Bellabeat executive team along with your high-level recommendations for Bellabeat’s marketing strategy.

# Data Analytics Process Stages

# Process Steps

## **Ask**

## **Prepare**

## **Process**

## **Share**

## **Analyze**

## **Act**

Recommendation

* *Click on the Step Process to Reference the Page.*

## Ask – Phase

**Questions**

1. *What is the problem you are trying to solve?*

* Consumers usage date and trends with other products and how can data be applied to Bellabeat products which helps inform the company’s marketing strategy.

1. *How can your insights drive business decisions?*

* The data insights can help better in driven decisions on company’s products with marketing strategies and target audiences.

**Key Tasks**

1. Task is to figure and analyze the data usages from other products and compare the trends in the market with Bellabeat products.

**Stakeholders**

* **Urška Sršen**: Bellabeat’s cofounder and Chief Creative Officer.
* **Sando Mur:** Mathematician and Bellabeat’s cofounder; key member of the Bellabeat executive team.

**Deliverable**

Analyze and define the trends of smart devices usage data on the market and equivalate to Bellabeat products with insights found on the analysis.

**The following deliverables:**

1. A clear summary of the business task

2. A description of all data sources used

3. Documentation of any cleaning or manipulation of data

4. A summary of your analysis

5. Supporting visualizations and key findings

6. Your top high-level content recommendations based on your analysis

## Prepare – Phase

Sršen encourages you to use public data that explores smart device users’ daily habits. She points you to a specific data set:

● FitBit Fitness Tracker Data (CC0: 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. Sršen tells you that this data set might have some limitations, and encourages you to consider adding another data to help address those limitations as you begin to work more with this data.

**Questions**

1. Where is your data stored?

* FitBit Fitness Tracker Data: <https://www.kaggle.com/arashnic/fitbit>.

1. Are there issues with bias or credibility in this data?

* I will download and import the dataset. Then make sure all the data is organized and credible. And I will sort and filter the data.

1. How are going to prepare the data and cleaned to process?

* Bellabeat encourages me to use public data that explores smart device users’ daily habits from FitBit Fitness Tracker Data. FitBit Fitness Tracker Data (CC0: Public Domain, dataset made available through Mobius).

**Key tasks**

1. Download data and store it appropriately.

2. Identify how it’s organized.

3. Sort and filter the data.

4. Determine the credibility of the data.

**Deliverable**

1. A description of all data sources used.

**Downloading Data**

Bellabeat encourages me to use open data that investigates the everyday routines of smart device users based on FitBit fitness tracker data. Data from FitBit fitness trackers (CC0: Public Domain, dataset made available through Mobius).

The Fitbit customers' data using the company’s fitness trackers are included in this Kaggle data collection. 30 qualified Fitbit users accepted the input of their own tracker data, including their physical activity statistics at the minute level, monitoring of sleep and heart rate.

Here is the URL where you can download the dataset:

Data from the FitBit fitness tracker can be found at *https://www.kaggle.com/arashnic/fitbit*.

18 CSV files are also included in my file after downloading it for the process phase.

**Packages**

Now, I’m going to Install some R packages that will help me in my process data analysis as well data cleaning packages (last 3).

install.packages("tidyverse")  
install.packages("lubridate")  
install.packages("dplyr")  
install.packages("ggplot2")  
install.packages("tidyr")  
install.packages("here")  
install.packages("skimr")  
install.packages("janitor")

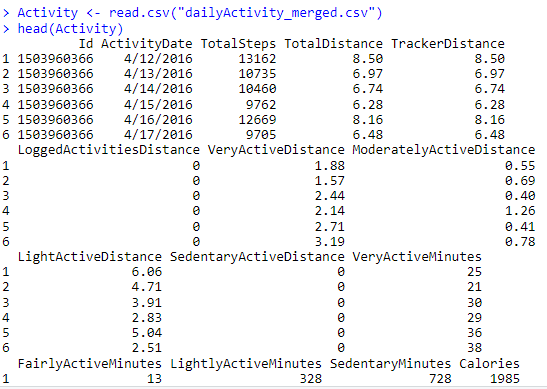
Loading packages on the R Script.

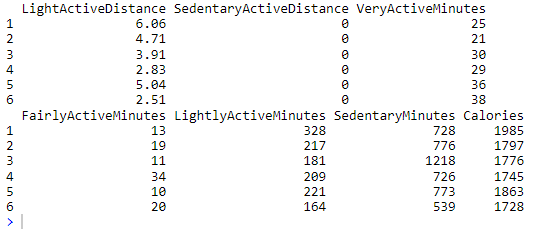
library(tidyverse)  
library(lubridate)  
library(dplyr)  
library(ggplot2)  
library(tidyr)  
library(here)  
library(skimr)  
library(janitor)

**Importing Dataset**

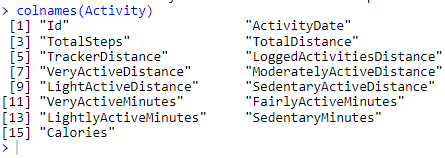
I will now import every dataset. Data should then be VIEWED, CLEANED, FORMATTED, and ORGANIZED. After looking over the entire dataset, I made the following assumptions and chose to focus exclusively on these data for my analysis:

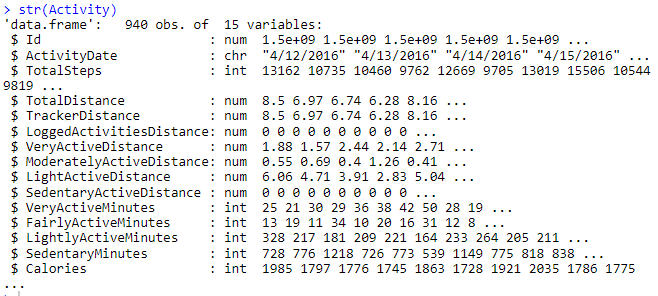
Activity <- read.csv("dailyActivity\_merged.csv")  
head (Activity)





> Colnames(Activity)

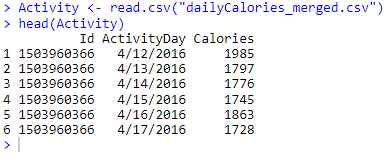




> str(Activity)

* **dailyCalories\_merged**

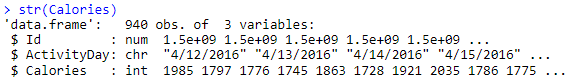
> Calories <- read.csv("dailyCalories\_merged.csv")  
head (Calories)



> colnames(Calories)

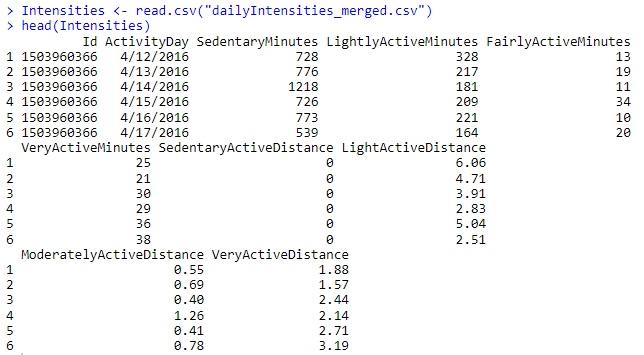


> str(Calories)

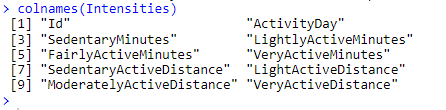


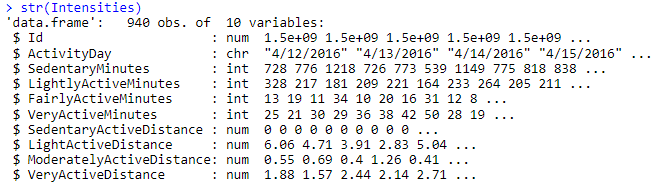
* **dailyIntensities\_merged.csv**

> Intensities <- read.csv("dailyIntensities\_merged.csv")  
head (Intensities)



> colnames(Intensities)

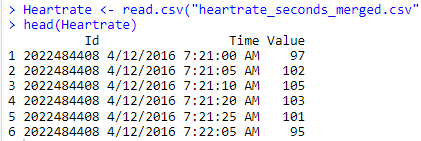




> str(Intensities)

• **heartrate\_seconds\_merged.csv**

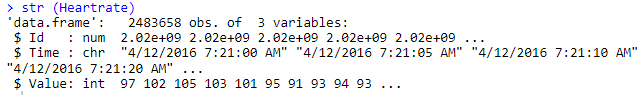
> Heartrate <- read.csv("heartrate\_seconds\_merged.csv")  
head (Heartrate)

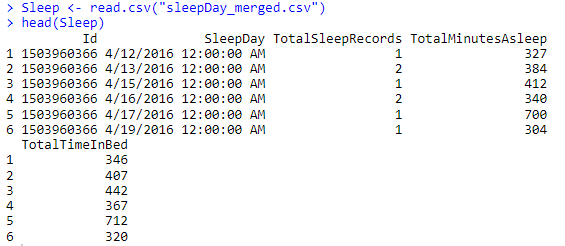


> colnames(Heartrate)



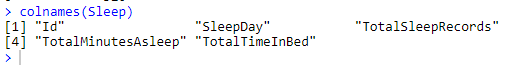
> str(Heartrate)



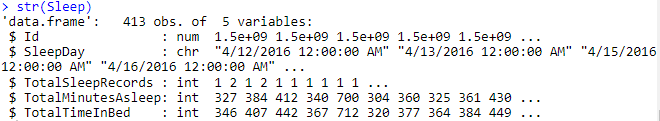
• **sleepDay\_merged.csv**

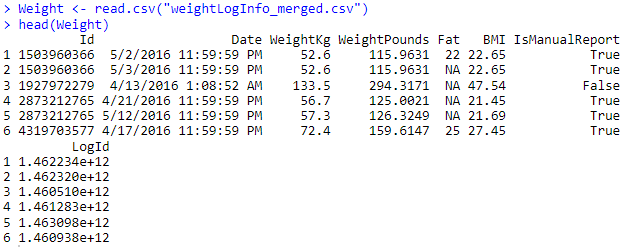
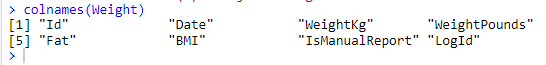
> Sleep <- read.csv("sleepDay\_merged.csv")  
head (Sleep)

> colnames(Sleep)



> str(Sleep)

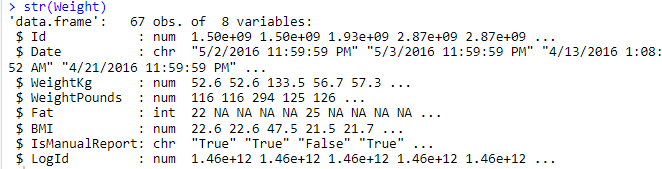


• **weightLogInfo\_merged.csv**

> colnames(Weight)

> Weight <- read.csv("weightLogInfo\_merged.csv")  
head (Weight)

> str(Weight)



Now that everything has been correctly imported, we can start the process phase in which we clean the data to organize the dataset.

## Process – Phase

Cleaning the dataset (Process Phase)

Basics cleaning:

Now, I’m going to Process, Clean and Organize the dataset for analysis. I used functions like glimpse(),

Skim without charts to quickly review the data. I also clean the names of the data using clean\_names().

**Questions**

1. What tools are you choosing and why?

* R Programming

1. How can you verify that your data is clean and ready to analyze?

• For Dataset (Activity, Calories and Intensities): For the data cleaning steps, I applied codes on R programming to clean this data (Spelling errors, Misfield values, Missing values, Extra and blank space, duplicated).

1. Have you documented your cleaning process so you can review and share those results?

* Yes, I will now process, clean, and organize the dataset in preparation for analysis. To swiftly review the data, I utilized routines like glimpse() and skim without charts. Additionally, I use clean names to purge the data's names ().

**Key tasks**

1. Check the data for errors.

2. Choose your tools.

3. Transform the data so you can work with it effectively.

4. Document the cleaning process.

**Deliverable**

1. Documentation of any cleaning or manipulation of data

**Basic cleaning:**

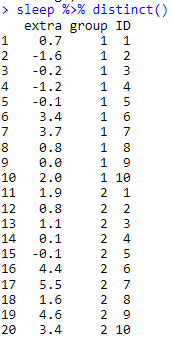
Then we process, clean, and organize the dataset for analysis by inspecting and investigating

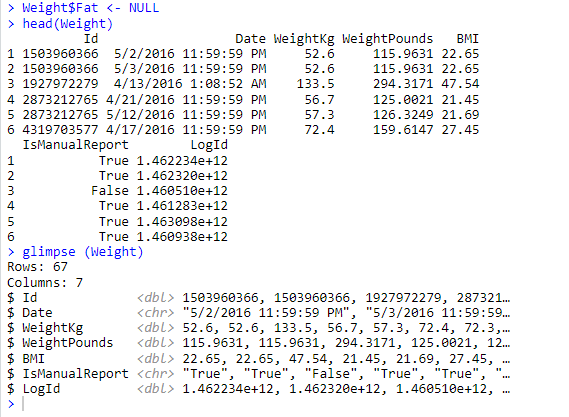
the data. And here are some cleaning steps I did with the data:

• For datasets (Activity, Calories, Intensity): No data cleaning steps found

Data (spelling errors, mistyped values, missing values, extra and empty spaces, no duplicates found).

• For sleep data: 3 duplicates were found and removed by using distinct().



• For weight data: There are too many missing values (NA) in $Fat - Column​​ and I decided to remove it.

**Formatting:**

We used explicit formatting. For data types, some dates are converted to numbers and dates

* **dailyActivity\_merged.csv**

> *Activity*

Activity$ActivityDate=as.POSIXct(Activity$ActivityDate, format="%m/%d/%Y", tz=Sys.timezone())

Activity$date <- format(Activity$ActivityDate, format = "%m/%d/%y")

Activity$ActivityDate=as.Date(Activity$ActivityDate, format="%m/%d/%Y", tz=Sys.timezone())

Activity$date=as.Date(Activity$date, format="%m/%d/%Y")

* **dailyIntensities\_merged.csv**

> *Intensities*Intensities$ActivityDay=as.Date(Intensities$ActivityDay, format="%m/%d/%Y", tz=Sys.timezone())

* **sleepDay\_merged.csv**

> *Sleep*

Sleep$SleepDay=as.POSIXct(Sleep$SleepDay, format="%m/%d/%Y %I:%M:%S %p"

, tz=Sys.timezone())

Sleep$date <- format(Sleep$SleepDay, format = "%m/%d/%y")

Sleep$date=as.Date(Sleep$date, "% m/% d/% y")

Now that the whole thing is ready, I can start discovering and analyzing the data sets.

## Analyze – Phase

**Summarizing**

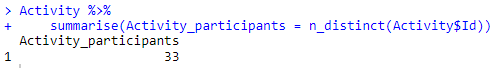
I can start using the data now that it has been properly organized and stored for analysis.

Let's examine the overall activity of the participant in each data set:

* What’s total number of recorded users in the data sets?

Activity %>%

summarise(Activity\_participants = n\_distinct(Activity$Id))



n\_distinct(Calories$Id)



n\_distinct(Intensities$Id)



n\_distinct(Heartrate$Id)



n\_distinct(Sleep$Id)



n\_distinct(Weight$Id)



* There are 33 participant the data sets for activity, calories, and intensities.
* 24 individuals made up the Sleep data.
* 14 individuals in the Heartrate study, and only 8 people in the weight study.

To draw any recommendations or conclusions from these datasets, the participation of 8 and 14 people is not important.

Therefore, I will concentrate my analysis on the following datasets:

**Activity, Calories, Intensities**, and **Sleep**.

The brief summary of the analysis for each data frame are provided below.

For the *Activity* data set:

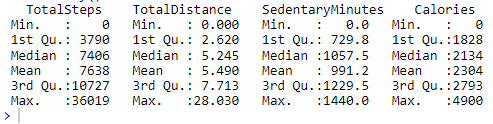
Activity %>%

select(TotalSteps,

TotalDistance,

SedentaryMinutes, Calories) %>%

summary()

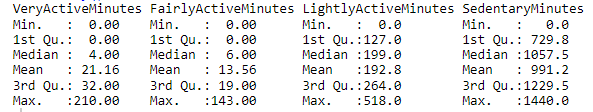


For the *Intense active participants* data set:

Intensities %>%

select(VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes, SedentaryMinutes) %>%

summary()

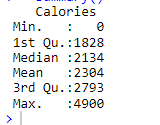


For the *Calories* data set:

Calories %>%

select(Calories) %>%

summary()

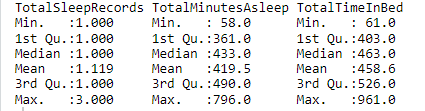


For the *Sleep* data set:

Sleep %>%

select(TotalSleepRecords, TotalMinutesAsleep, TotalTimeInBed) %>%

summary()



**Key Analysis Conclusions:**

• Too much time is spent sitting on average (more than 16 hours). and unquestionably has to be decreased with an effective marketing plan.

• Most of the participants engage in light physical activity and has a lot of inactive time.

* Participants slept for an average of seven hours every night.

• The average daily step count (7638) is somewhat below the CDC's suggested amount. The CDC found that walking 8,000 steps per day was linked to a 51% lower risk of death from all causes (or death from all causes). Furthermore, compared to 4,000 steps, completing 12,000 steps per day was connected with a 65% decreased risk.

**Merging Data Sets:**

I'm going to combine/join two data sets: **Activity** and **Sleep** data on columns Id, before starting to create the visualization data. It should be noted that the Activity dataset has more participant Ids than the Sleep dataset. I will therefore have the total number of participants from the Sleep data set if I apply the merging option <inner joint>.

Combined\_data\_inner <- merge(Sleep, Activity, by="Id")  
n\_distinct(Combined\_data\_inner$Id)



Because of this, I will think about utilizing "outer join" to keep all participants in the dataset for analysis. And I can accomplish that by including the additional option all=TRUE in my code section.

Combined\_data\_outer <- merge(Sleep, Activity, by="Id", all = TRUE)  
n\_distinct(Combined\_data\_outer$Id)



## Share – Phase

**Data Visualization:**

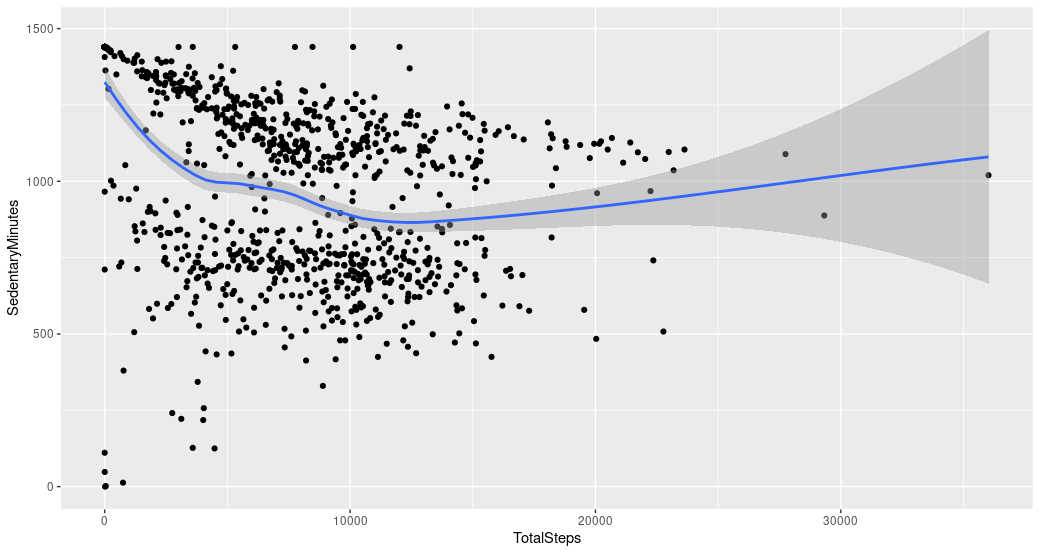
After combining the data sets above, we can now start visualization some important analysis searches and relationships between the data sets.

**Relationship between Steps and Sedentary time:**

* What’s the relationship between steps taken in a day and sedentary minutes?

ggplot(data=Activity, aes(x=TotalSteps, y=SedentaryMinutes)) + geom\_point() + geom\_smooth()

+ labs(title)

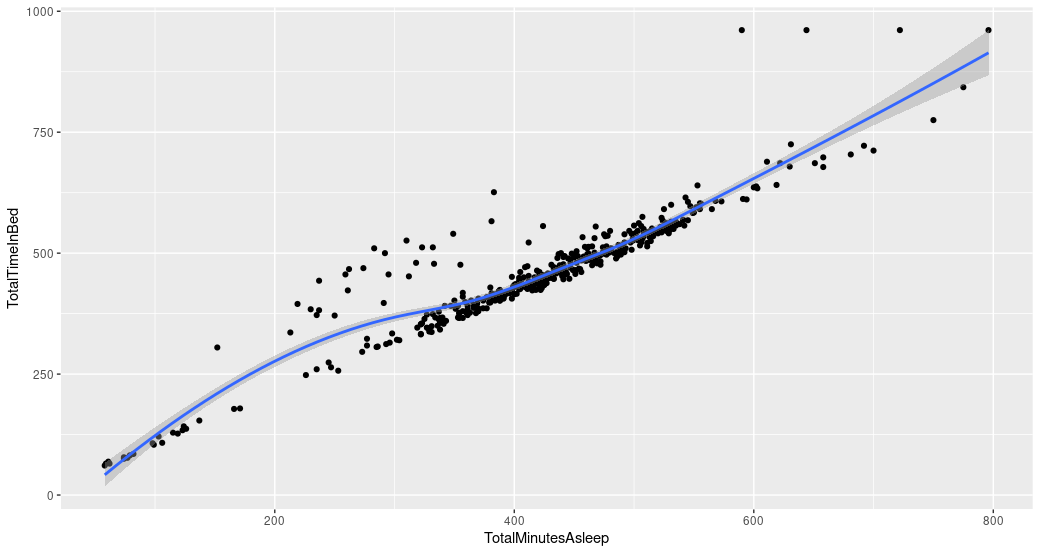
Total Steps vs. Sedentary Minutes 

As shown here, I observed a negative relationship between steps and sedentary time in this data set. Less steps are taken throughout the day as participants spend more time sitting down. This data information demonstrates the need for the business to promote to client groups that spend a lot of time sitting down. And to accomplish that, the business must identify means of finding clients to increase their walking and tracking their daily steps & activities.

**Relationship between Minutes Asleep and Time in Bed:**

* What’s the relationship between minutes asleep and time in bed?

ggplot(data=Sleep, aes(x=TotalMinutesAsleep, y=TotalTimeInBed)) + geom\_point()+ geom\_smooth() + labs(title)

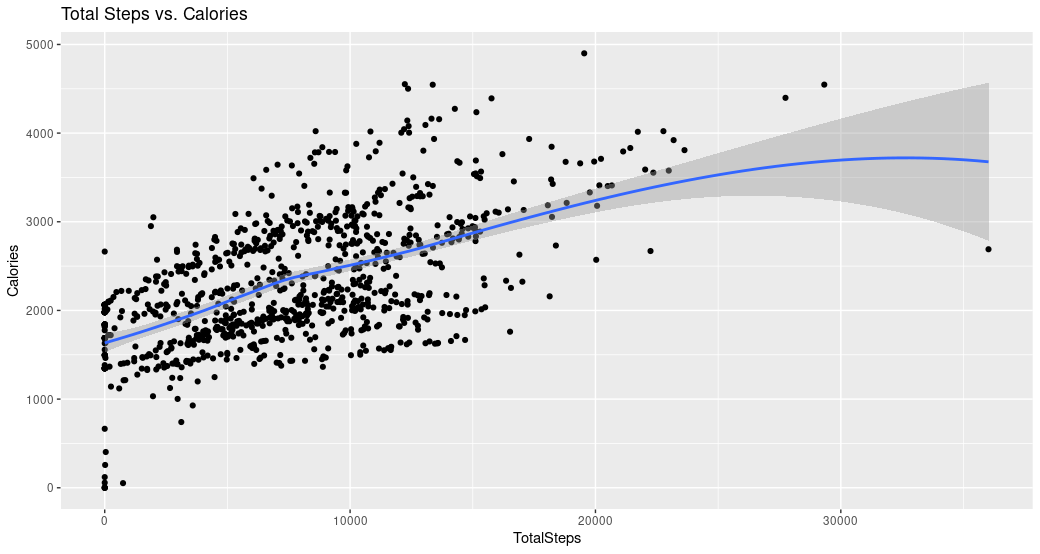
Minutes Asleep vs. Time in Bed Minutes  
****  
As shown, the relationship between Minutes Asleep and Time in Bed is virtually linear in this chart. The business should create a personal notification to go to sleep at specific time or help users get better sleep time.

**Relationship between Steps and Calories:**

* What’s the relationship between steps taken and Calories?

ggplot(data=Activity, aes(x=TotalSteps, y=Calories))

+geom\_point() + geom\_smooth() + labs(title="Total Steps vs. Calories")

Total Steps vs. Calories

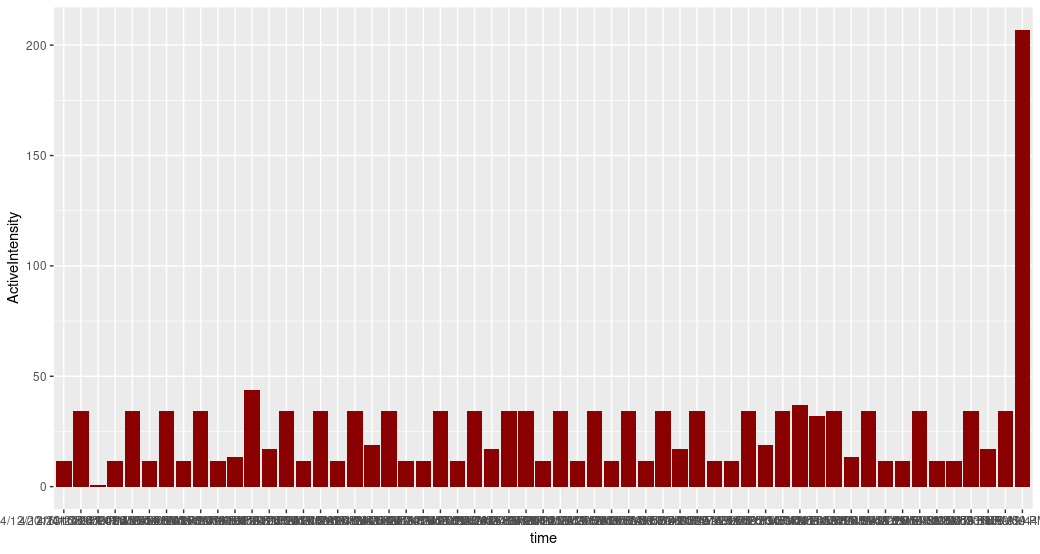
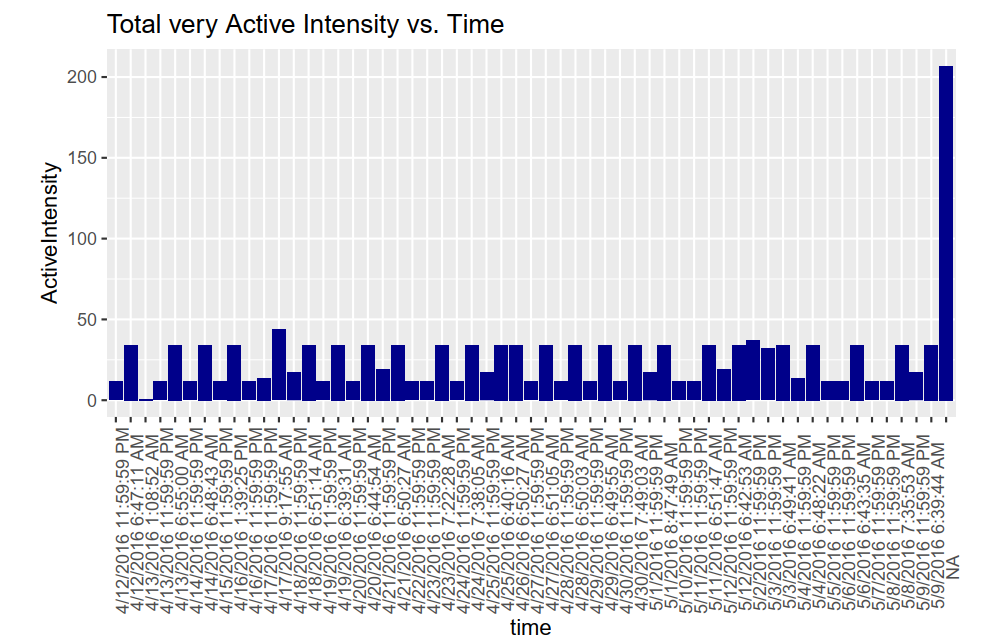
Here we can see that there is a positive correlation between total steps and calories. the more we act, the more calories burned.

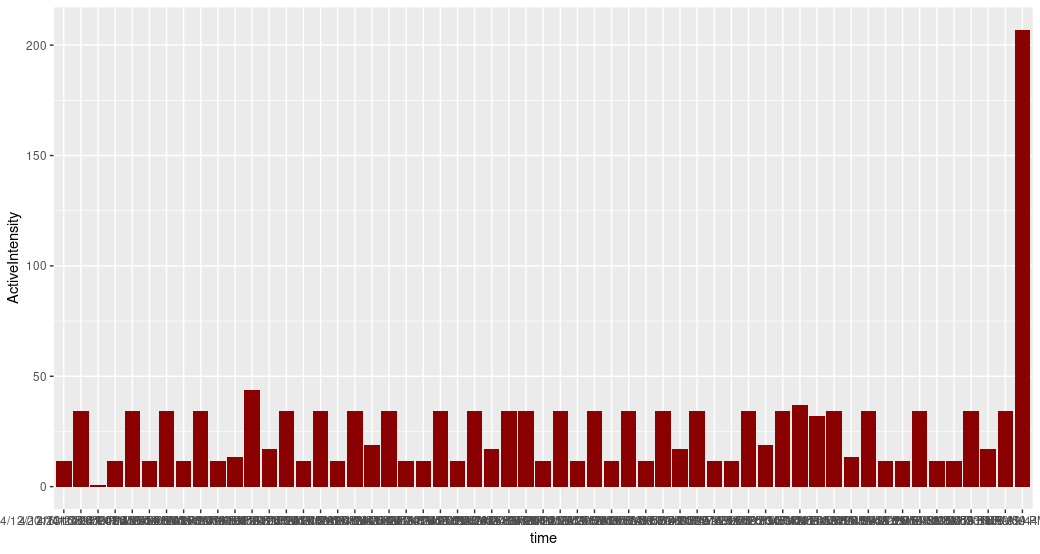
**Intensities data**Let's now examine some data on intensities over time.

Intensities$ActiveIntensity <- (Intensities$VeryActiveMinutes)/60

Combined\_data <- merge(Weight, Intensities, by="Id", all=TRUE)  
Combined\_data$time <- format(Combined\_data$Date, format = "%H:%M:%S")

ggplot(data=Combined\_data, aes(x=time, y=ActiveIntensity)) + geom\_histogram(stat = "identity", fill='dark red')theme(axis.text.x = element\_text(angle = 90)) + labs(title="Total very Active Intensity vs. Time ")

Active Intensity vs. Time



As shown, by analyzing some time-series data on intensity. The business will have a clear understanding of how customers use their product throughout the day. I guess the majority of users are active before and after work. In the Bellabeat app, the business can utilize this time to remind and encourage customers to go for a run or an activity.

## Act – Phase

**Conclusions & Recommendations for the Business:**

In conclusion, after gathering data insights on activity, sleep, stress, and other factors will enable Bellabeat to provide its customers with knowledge about their own health and daily routines. Bellabeat is a rapidly expanding business that has successfully positioned itself as a tech-driven wellness provider for its clients.

Here are my insights that would help direct Bellabeat marketing decisions and strategy:

* **Target Audience:**

People with full-time employment who spend a lot of time working at computers and in offices require regular exercise and activities to stay in shape. The users engage in some mild exercise to maintain their health (according to the activity type analysis). As result, more active hours is required to increase their health benefits, as well the company might require to provide some knowledge about creating healthy habits and finding a initiative to maintain them.

* **Company’s Suggestion**

The Bellabeat app must be a personal and adaptive fitness app. By becoming as companion guide to its users and customers and assist them manage their personal and professional lives with good healthy reminders and strong good habits.

# Recommendations

1. The average amount of time spent inactive is too long for app users (more than 16 hours). and unquestionably has to be decreased with an effective marketing plan. Data suggests that the business should target customers who spend a lot of time sitting down more. To do this, the business must discover strategies for encouraging clients to begin walking more by tracking their daily steps notifications.
2. Participants slept for an average of seven hours every night. Bellabeat should think about setting reminders & alarms notifications as a way to get users to sleep better. Additionally, the Bellabeat app can advise users to spend less time sitting down and inactive.
3. The average daily step count (7638) is somewhat below the CDC's suggested amount, Bellabeat app should encourage their users to set target goals to achieve minimum active steps.
4. As displayed in data, the Intensity over time. The business will have a clear understanding of how its users interact with its app throughout the day. In the Bellabeat app, the business can utilize this time to remind and encourage customers to go for a run or an activity.
5. Limiting daily calorie intake may be a smart move for clients who wish to reduce their weight. Bellabeat can offer some suggestions for calorie-free, nutritious foods (for lunch and dinner).

Data Analytics Case Study with R – Case 2 Project



Thank you very much for your interest in My Bellabeat Case Study!

This project is great example of showcasing my skills in data analytics and R programming which I learned and practiced in my google certificate that I obtained from Google.

Please check out my projects on my web blogs and stay tuned for more complex and challenging projects ahead on my profile on **LinkedIn & GitHub**.

Data analytics project

Anas al jarrah