Bellabeat Case Study

Bishop Sally

2023-06-01

# **Introduction**

This is my breakdown of my Bellabeat Case study capstone project, which is part of my Google Data Analytics Certificate. This breakdown will give general insights to the analyst team at Bellabeat to make informed business decisions that can help the company in the future The method I used to analyze Bellabeat was breaking everything down by using:

ASK PREPARE PROCESS ANALYZE SHARE ACT

## **About Bellabeat**

[Bellabeat](https://bellabeat.com/?gclid=CjwKCAjwg-GjBhBnEiwAMUvNW8dU3gg_7xVeMp_WV-1Acp6kc_57Kj7kPTRdLJCLFy2Cq3PRWln7cRoCep0QAvD_BwE) is 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.

# **Ask**

In the *Ask* section we should ask some key questions to find the problem and how we can solve them. The key questions that need to be asked to gain further insight are: **1.** What are the trends of smart device usuage? **2.** How can these trends apply to Bellabeat Customers? **3.** How can these trends help influence Bellabeat Marketing Strategy

**Business Task:** The questions that need to be asked will lead us to analyze smart device usage to gain insight into how consumers use non-Bellabeat smart devices. After we gain the insights we need we will then select one Bellabeat product that we can apply to these insights.

**How I will report the information:** 1. Give a clear summary of the business task 2. Adding a description of all data sources used 3. Documenting any cleaning or manipulation of data 4. Providing a summary of my analysis 5. Supporting visualizations and key findings 6. Your top high-level content recommendations based on your analysis

**Identifying Key Stakeholders:** ***Primary StakeHolders:*** *-* ***Urška Sršen:*** *Bellabeat’s cofounder and Chief Creative Officer -* ***Sando Mur:*** *Mathematician and Bellabeat’s cofounder; key member of the Bellabeat executive team Secondary Stakeholders:* \_\_Secondary Stakeholders:\_\_ - **Bellabeat marketing analytics team:** A team of data analysts responsible for collecting, analyzing, and reporting data that helps guide Bellabeat’s marketing strategy

# **Prepare**

In the prepare section we find, download, and store the data in a secured folder.

* **What data was gathered:** [\*Kaggle Dataset](https://www.kaggle.com/datasets/arashnic/fitbit)
  + This data set was generated by respondents to a distributed survey via [Amazon Mechanical Turk](https://zenodo.org/record/53894#.YMoUpnVKiP9) between 03.12.2016-05.12.2016. 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 Summary**:
  + 33 Fitbit users agreed to submit personal data while using their Fitbit Trackers.The data tracked contained different factors for each data set
  + There is a total of 18 csv files that were used for the [FitBit Fitness Tracker Data](https://www.kaggle.com/datasets/arashnic/fitbit). Out of the 18, I used 8 out of the 18 data sets to help me find a solution. These data sets included: *daily\_activity*, *daily\_calories*, *daily\_intensities*, *daily\_steps*, *hourly\_calories*, *hourly\_steps*,*sleep\_day*, and *weight\_info*.
  + The data was generated between March 12 2016 through May 12 2016
  + Source: Furberg, R., Brinton, J., Keating, M., & Ortiz, A. (2016). Crowd-sourced Fitbit data sets 03.12.2016-05.12.2016 [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.53894>
  + [Creative Commons Atrribution 4.0 International](https://creativecommons.org/licenses/by/4.0/legalcode)

## **Data Reliability**

* 33 participants is considered a sufficient sample size, but a larger sample size would be more efficient based on the market size of the fitness tracker market as a whole.
* The participants are also anonymous so it does not include specific demographic data to avoid any biases.
* The data tracked activity and correlation but it does not state any specific features being utilized and tracked during this study.
* The data was generated between March and May 2016, so this data can be considered old and not being up to date with newer trends and features of Fitness Trackers

## **Conclusion:** I think overall this data is sufficient enough to get a general consensus of the Fitness tracker market, but a larger and more current dataset would be a good alternative in the future.

# **Process**

After securing the data now we can load, view, and clean the data to make sure the data is valid enough to complete the remaining steps

# Installing necessary packages  
options(repos = "https://cran.rstudio.com/") # Set the CRAN mirror  
  
  
install.packages("tidyverse")

## Installing package into 'C:/Users/bsall/AppData/Local/R/win-library/4.3'  
## (as 'lib' is unspecified)

## package 'tidyverse' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\bsall\AppData\Local\Temp\RtmpOEIZ53\downloaded\_packages

install.packages("ggplot2")

## Installing package into 'C:/Users/bsall/AppData/Local/R/win-library/4.3'  
## (as 'lib' is unspecified)

## package 'ggplot2' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\bsall\AppData\Local\Temp\RtmpOEIZ53\downloaded\_packages

install.packages("janitor")

## Installing package into 'C:/Users/bsall/AppData/Local/R/win-library/4.3'  
## (as 'lib' is unspecified)

## package 'janitor' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\bsall\AppData\Local\Temp\RtmpOEIZ53\downloaded\_packages

#Loading necessary packages   
  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.2 ✔ 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 to become errors

library(ggplot2)  
library(janitor)

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

library(readr)

## **Importing Data**

#Loading csv files into data sets   
  
daily\_activity <- read\_csv("C:/Users/bsall/OneDrive/Documents/Fitabase Data 4.12.16-5.12.16/dailyActivity\_merged.csv")

## Rows: 940 Columns: 15  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): ActivityDate  
## dbl (14): Id, TotalSteps, TotalDistance, TrackerDistance, LoggedActivitiesDi...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

View(daily\_activity)  
  
daily\_calories <- read\_csv("C:/Users/bsall/OneDrive/Documents/Fitabase Data 4.12.16-5.12.16/dailyCalories\_merged.csv")

## Rows: 940 Columns: 3  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): ActivityDay  
## dbl (2): Id, Calories  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

View(daily\_calories)  
  
daily\_intensities <- read\_csv("C:/Users/bsall/OneDrive/Documents/Fitabase Data 4.12.16-5.12.16/dailyIntensities\_merged.csv")

## Rows: 940 Columns: 10  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): ActivityDay  
## dbl (9): Id, SedentaryMinutes, LightlyActiveMinutes, FairlyActiveMinutes, Ve...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

View(daily\_intensities)  
  
daily\_steps <- read\_csv("C:/Users/bsall/OneDrive/Documents/Fitabase Data 4.12.16-5.12.16/dailySteps\_merged.csv")

## Rows: 940 Columns: 3  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): ActivityDay  
## dbl (2): Id, StepTotal  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

View(daily\_steps)  
  
sleep\_day <- read\_csv("C:/Users/bsall/OneDrive/Documents/Fitabase Data 4.12.16-5.12.16/sleepDay\_merged.csv")

## Rows: 413 Columns: 5  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): SleepDay  
## dbl (4): Id, TotalSleepRecords, TotalMinutesAsleep, TotalTimeInBed  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

View(sleep\_day)  
  
weight\_info <- read\_csv("C:/Users/bsall/OneDrive/Documents/Fitabase Data 4.12.16-5.12.16/weightLogInfo\_merged.csv")

## Rows: 67 Columns: 8  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): Date  
## dbl (6): Id, WeightKg, WeightPounds, Fat, BMI, LogId  
## lgl (1): IsManualReport  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

View(weight\_info)  
  
hourly\_steps <- read\_csv("C:/Users/bsall/OneDrive/Documents/Fitabase Data 4.12.16-5.12.16/hourlySteps\_merged.csv")

## Rows: 22099 Columns: 3  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): ActivityHour  
## dbl (2): Id, StepTotal  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

View(hourly\_steps)  
  
hourly\_Calories <- read\_csv("C:/Users/bsall/OneDrive/Documents/Fitabase Data 4.12.16-5.12.16/hourlyCalories\_merged.csv")

## Rows: 22099 Columns: 3  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): ActivityHour  
## dbl (2): Id, Calories  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

View(hourly\_Calories)

# Removing unnecessary columns  
  
daily\_activity = subset(daily\_activity,select = -c(LoggedActivitiesDistance, VeryActiveDistance, ModeratelyActiveDistance, LightActiveDistance, SedentaryActiveDistance))  
  
daily\_intensities = subset(daily\_intensities,select = -c(SedentaryActiveDistance, LightActiveDistance, ModeratelyActiveDistance, VeryActiveDistance))

# Finding null or missing values in my datasets  
  
is.null(daily\_activity)

## [1] FALSE

is.null(daily\_calories)

## [1] FALSE

is.null(daily\_intensities)

## [1] FALSE

is.null(daily\_steps)

## [1] FALSE

is.null (sleep\_day)

## [1] FALSE

is.null(weight\_info)

## [1] FALSE

is.null(hourly\_Calories)

## [1] FALSE

is.null(hourly\_steps)

## [1] FALSE

#Viewing the columns to check for any inconsistent data   
  
head(daily\_activity)

## # A tibble: 6 × 10  
## Id ActivityDate TotalSteps TotalDistance TrackerDistance VeryActiveMinutes  
## <dbl> <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 1.50e9 4/12/2016 13162 8.5 8.5 25  
## 2 1.50e9 4/13/2016 10735 6.97 6.97 21  
## 3 1.50e9 4/14/2016 10460 6.74 6.74 30  
## 4 1.50e9 4/15/2016 9762 6.28 6.28 29  
## 5 1.50e9 4/16/2016 12669 8.16 8.16 36  
## 6 1.50e9 4/17/2016 9705 6.48 6.48 38  
## # ℹ 4 more variables: FairlyActiveMinutes <dbl>, LightlyActiveMinutes <dbl>,  
## # SedentaryMinutes <dbl>, Calories <dbl>

str(daily\_activity)

## tibble [940 × 10] (S3: tbl\_df/tbl/data.frame)  
## $ Id : num [1:940] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...  
## $ ActivityDate : chr [1:940] "4/12/2016" "4/13/2016" "4/14/2016" "4/15/2016" ...  
## $ TotalSteps : num [1:940] 13162 10735 10460 9762 12669 ...  
## $ TotalDistance : num [1:940] 8.5 6.97 6.74 6.28 8.16 ...  
## $ TrackerDistance : num [1:940] 8.5 6.97 6.74 6.28 8.16 ...  
## $ VeryActiveMinutes : num [1:940] 25 21 30 29 36 38 42 50 28 19 ...  
## $ FairlyActiveMinutes : num [1:940] 13 19 11 34 10 20 16 31 12 8 ...  
## $ LightlyActiveMinutes: num [1:940] 328 217 181 209 221 164 233 264 205 211 ...  
## $ SedentaryMinutes : num [1:940] 728 776 1218 726 773 ...  
## $ Calories : num [1:940] 1985 1797 1776 1745 1863 ...

n\_distinct(sleep\_day$Id)

## [1] 24

n\_distinct(daily\_activity$Id)

## [1] 33

is\_empty(daily\_activity)

## [1] FALSE

is\_empty(daily\_calories)

## [1] FALSE

is\_empty(daily\_intensities)

## [1] FALSE

is\_empty(daily\_steps)

## [1] FALSE

is\_empty(sleep\_day)

## [1] FALSE

is\_empty(weight\_info)

## [1] FALSE

is\_empty(hourly\_Calories)

## [1] FALSE

is\_empty(hourly\_steps)

## [1] FALSE

# Updating and reformatting the date columns to get a more desired format   
  
daily\_activity$ActivityDate <- as.Date(daily\_activity$ActivityDate, "%m/%d/%y")  
daily\_calories$ActivityDay <- as.Date(daily\_calories$ActivityDay, "%m/%d/%y")  
daily\_intensities$ActivityDay <- as.Date(daily\_intensities$ActivityDay, "%m/%d/%y")  
daily\_steps$ActivityDay <- as.Date(daily\_steps$ActivityDay, "%m/%d/%y")  
sleep\_day$SleepDay <- as.Date(strptime(sleep\_day$SleepDay, "%m/%d/%Y"))  
hourly\_Calories$ActivityHour <- as.Date(hourly\_Calories$ActivityHour, "%m/%d/%y")  
hourly\_steps$ActivityHour <- as.Date(hourly\_steps$ActivityHour, "%m/%d/%y")  
weight\_info$Date <- as.Date(weight\_info$Date, "%m/%d/%y")

# Checking for the new date format   
  
str(daily\_activity)

## tibble [940 × 10] (S3: tbl\_df/tbl/data.frame)  
## $ Id : num [1:940] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...  
## $ ActivityDate : Date[1:940], format: "2020-04-12" "2020-04-13" ...  
## $ TotalSteps : num [1:940] 13162 10735 10460 9762 12669 ...  
## $ TotalDistance : num [1:940] 8.5 6.97 6.74 6.28 8.16 ...  
## $ TrackerDistance : num [1:940] 8.5 6.97 6.74 6.28 8.16 ...  
## $ VeryActiveMinutes : num [1:940] 25 21 30 29 36 38 42 50 28 19 ...  
## $ FairlyActiveMinutes : num [1:940] 13 19 11 34 10 20 16 31 12 8 ...  
## $ LightlyActiveMinutes: num [1:940] 328 217 181 209 221 164 233 264 205 211 ...  
## $ SedentaryMinutes : num [1:940] 728 776 1218 726 773 ...  
## $ Calories : num [1:940] 1985 1797 1776 1745 1863 ...

str(daily\_calories)

## spc\_tbl\_ [940 × 3] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ Id : num [1:940] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...  
## $ ActivityDay: Date[1:940], format: "2020-04-12" "2020-04-13" ...  
## $ Calories : num [1:940] 1985 1797 1776 1745 1863 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. Id = col\_double(),  
## .. ActivityDay = col\_character(),  
## .. Calories = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

str(daily\_steps)

## spc\_tbl\_ [940 × 3] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ Id : num [1:940] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...  
## $ ActivityDay: Date[1:940], format: "2020-04-12" "2020-04-13" ...  
## $ StepTotal : num [1:940] 13162 10735 10460 9762 12669 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. Id = col\_double(),  
## .. ActivityDay = col\_character(),  
## .. StepTotal = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

str(sleep\_day)

## spc\_tbl\_ [413 × 5] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ Id : num [1:413] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...  
## $ SleepDay : Date[1:413], format: "2016-04-12" "2016-04-13" ...  
## $ TotalSleepRecords : num [1:413] 1 2 1 2 1 1 1 1 1 1 ...  
## $ TotalMinutesAsleep: num [1:413] 327 384 412 340 700 304 360 325 361 430 ...  
## $ TotalTimeInBed : num [1:413] 346 407 442 367 712 320 377 364 384 449 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. Id = col\_double(),  
## .. SleepDay = col\_character(),  
## .. TotalSleepRecords = col\_double(),  
## .. TotalMinutesAsleep = col\_double(),  
## .. TotalTimeInBed = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

str(daily\_intensities)

## tibble [940 × 6] (S3: tbl\_df/tbl/data.frame)  
## $ Id : num [1:940] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...  
## $ ActivityDay : Date[1:940], format: "2020-04-12" "2020-04-13" ...  
## $ SedentaryMinutes : num [1:940] 728 776 1218 726 773 ...  
## $ LightlyActiveMinutes: num [1:940] 328 217 181 209 221 164 233 264 205 211 ...  
## $ FairlyActiveMinutes : num [1:940] 13 19 11 34 10 20 16 31 12 8 ...  
## $ VeryActiveMinutes : num [1:940] 25 21 30 29 36 38 42 50 28 19 ...

str(hourly\_Calories)

## spc\_tbl\_ [22,099 × 3] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ Id : num [1:22099] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...  
## $ ActivityHour: Date[1:22099], format: "2020-04-12" "2020-04-12" ...  
## $ Calories : num [1:22099] 81 61 59 47 48 48 48 47 68 141 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. Id = col\_double(),  
## .. ActivityHour = col\_character(),  
## .. Calories = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

str(hourly\_steps)

## spc\_tbl\_ [22,099 × 3] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ Id : num [1:22099] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...  
## $ ActivityHour: Date[1:22099], format: "2020-04-12" "2020-04-12" ...  
## $ StepTotal : num [1:22099] 373 160 151 0 0 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. Id = col\_double(),  
## .. ActivityHour = col\_character(),  
## .. StepTotal = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

str(weight\_info)

## spc\_tbl\_ [67 × 8] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ Id : num [1:67] 1.50e+09 1.50e+09 1.93e+09 2.87e+09 2.87e+09 ...  
## $ Date : Date[1:67], format: "2020-05-02" "2020-05-03" ...  
## $ WeightKg : num [1:67] 52.6 52.6 133.5 56.7 57.3 ...  
## $ WeightPounds : num [1:67] 116 116 294 125 126 ...  
## $ Fat : num [1:67] 22 NA NA NA NA 25 NA NA NA NA ...  
## $ BMI : num [1:67] 22.6 22.6 47.5 21.5 21.7 ...  
## $ IsManualReport: logi [1:67] TRUE TRUE FALSE TRUE TRUE TRUE ...  
## $ LogId : num [1:67] 1.46e+12 1.46e+12 1.46e+12 1.46e+12 1.46e+12 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. Id = col\_double(),  
## .. Date = col\_character(),  
## .. WeightKg = col\_double(),  
## .. WeightPounds = col\_double(),  
## .. Fat = col\_double(),  
## .. BMI = col\_double(),  
## .. IsManualReport = col\_logical(),  
## .. LogId = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

# merging data sets: sleep\_day and hourly\_steps to create \*combined\_bars\*  
  
combined\_bars <- merge(sleep\_day, hourly\_steps, by = c('Id'))  
  
# viewing the newly merged data   
  
view(combined\_bars)  
head(combined\_bars)

## Id SleepDay TotalSleepRecords TotalMinutesAsleep TotalTimeInBed  
## 1 1503960366 2016-04-12 1 327 346  
## 2 1503960366 2016-04-12 1 327 346  
## 3 1503960366 2016-04-12 1 327 346  
## 4 1503960366 2016-04-12 1 327 346  
## 5 1503960366 2016-04-12 1 327 346  
## 6 1503960366 2016-04-12 1 327 346  
## ActivityHour StepTotal  
## 1 2020-04-12 1864  
## 2 2020-04-12 1166  
## 3 2020-04-12 676  
## 4 2020-04-14 0  
## 5 2020-04-12 344  
## 6 2020-04-12 250

# Merging the data sets: daily\_activity and daily\_steps to create \*record\_activity\*  
  
record\_activity <- merge(daily\_activity, daily\_steps, by = c('Id'))  
  
#Viewing the newly merged data set  
  
view(record\_activity)  
head(record\_activity)

## Id ActivityDate TotalSteps TotalDistance TrackerDistance  
## 1 1503960366 2020-04-12 13162 8.5 8.5  
## 2 1503960366 2020-04-12 13162 8.5 8.5  
## 3 1503960366 2020-04-12 13162 8.5 8.5  
## 4 1503960366 2020-04-12 13162 8.5 8.5  
## 5 1503960366 2020-04-12 13162 8.5 8.5  
## 6 1503960366 2020-04-12 13162 8.5 8.5  
## VeryActiveMinutes FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes  
## 1 25 13 328 728  
## 2 25 13 328 728  
## 3 25 13 328 728  
## 4 25 13 328 728  
## 5 25 13 328 728  
## 6 25 13 328 728  
## Calories ActivityDay StepTotal  
## 1 1985 2020-04-12 13162  
## 2 1985 2020-04-13 10735  
## 3 1985 2020-04-14 10460  
## 4 1985 2020-04-15 9762  
## 5 1985 2020-04-16 12669  
## 6 1985 2020-04-17 9705

# Mutating all the time based columns to create a total minutes column that combines the times together   
  
total\_intensities <- daily\_intensities %>%  
 mutate(daily\_intensities = rowSums(select(., SedentaryMinutes, LightlyActiveMinutes, FairlyActiveMinutes, VeryActiveMinutes), na.rm = TRUE))

#This string adds the calories column from the daily\_calories data set into the total\_intensities data set   
  
total\_intensities$Calories <- daily\_calories$Calories  
  
#This changes the colname of daily\_intensities into \*TotalMinutes\* to help organize the colnames when entering the data  
  
names(total\_intensities)[names(total\_intensities) == "daily\_intensities"] <- "TotalMinutes"

# Removes all falsely reported data sets that were presented in the IsManualReport column   
weight\_info <- subset(weight\_info, IsManualReport == TRUE)

#Summarizing weight of the selected columns and merged the data into the new variable \*weight\_summary\*   
  
weight\_summary <- weight\_info %>%  
 group\_by(Date) %>%  
 summarise(AverageWeight = mean(WeightPounds))  
  
glimpse(weight\_summary$AverageWeight)

## num [1:30] 138 137 136 136 137 ...

glimpse(weight\_info$WeightPounds)

## num [1:41] 116 116 125 126 160 ...

glimpse(weight\_info$Id)

## num [1:41] 1.50e+09 1.50e+09 2.87e+09 2.87e+09 4.32e+09 ...

# **Analyze**

After the data is cleaned now we can analyze the data so we can get an idea of how we are going to create and share the data

# Getting an overall summary of the data present in the \*daily\_activity\* data set before I begin visualizing   
daily\_activity %>%  
 select(TotalSteps,  
 TotalDistance, TrackerDistance, VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes, SedentaryMinutes, Calories) %>%  
 summary()

## TotalSteps TotalDistance TrackerDistance VeryActiveMinutes  
## Min. : 0 Min. : 0.000 Min. : 0.000 Min. : 0.00   
## 1st Qu.: 3790 1st Qu.: 2.620 1st Qu.: 2.620 1st Qu.: 0.00   
## Median : 7406 Median : 5.245 Median : 5.245 Median : 4.00   
## Mean : 7638 Mean : 5.490 Mean : 5.475 Mean : 21.16   
## 3rd Qu.:10727 3rd Qu.: 7.713 3rd Qu.: 7.710 3rd Qu.: 32.00   
## Max. :36019 Max. :28.030 Max. :28.030 Max. :210.00   
## FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories   
## Min. : 0.00 Min. : 0.0 Min. : 0.0 Min. : 0   
## 1st Qu.: 0.00 1st Qu.:127.0 1st Qu.: 729.8 1st Qu.:1828   
## Median : 6.00 Median :199.0 Median :1057.5 Median :2134   
## Mean : 13.56 Mean :192.8 Mean : 991.2 Mean :2304   
## 3rd Qu.: 19.00 3rd Qu.:264.0 3rd Qu.:1229.5 3rd Qu.:2793   
## Max. :143.00 Max. :518.0 Max. :1440.0 Max. :4900

#Summarizing \*sleep\* and \*Time\* based columns in the \*\*sleep\_day\*\* variable   
sleep\_day %>%  
 select(TotalSleepRecords, TotalMinutesAsleep, TotalTimeInBed) %>%  
 summary()

## TotalSleepRecords TotalMinutesAsleep TotalTimeInBed   
## Min. :1.000 Min. : 58.0 Min. : 61.0   
## 1st Qu.:1.000 1st Qu.:361.0 1st Qu.:403.0   
## Median :1.000 Median :433.0 Median :463.0   
## Mean :1.119 Mean :419.5 Mean :458.6   
## 3rd Qu.:1.000 3rd Qu.:490.0 3rd Qu.:526.0   
## Max. :3.000 Max. :796.0 Max. :961.0

#Summarizing the \*ActivityDate\* and \*StepTotal\* columns in the \*\*record\_activity\*\* variable   
  
record\_activity %>%   
 select(ActivityDate, StepTotal) %>%  
 summary()

## ActivityDate StepTotal   
## Min. :2020-04-12 Min. : 0   
## 1st Qu.:2020-04-19 1st Qu.: 3761   
## Median :2020-04-26 Median : 7443   
## Mean :2020-04-26 Mean : 7673   
## 3rd Qu.:2020-05-04 3rd Qu.:10771   
## Max. :2020-05-12 Max. :36019

#Summarizing the \*TotalMinutes\* and \*Calories\* columns in the \*\*total\_intensities\*\* variable   
  
total\_intensities %>%  
 select(TotalMinutes,Calories) %>%  
 summary()

## TotalMinutes 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

#Summarizing the \*ActivityDate\* and \*StepTotal\* columns in the \*\*record\_activity\*\* variable   
  
record\_activity %>%   
 select(ActivityDate, StepTotal) %>%  
 summary()

## ActivityDate StepTotal   
## Min. :2020-04-12 Min. : 0   
## 1st Qu.:2020-04-19 1st Qu.: 3761   
## Median :2020-04-26 Median : 7443   
## Mean :2020-04-26 Mean : 7673   
## 3rd Qu.:2020-05-04 3rd Qu.:10771   
## Max. :2020-05-12 Max. :36019

#Summarizing the \*Date\* and \*AverageWeight\* columns in the \*\*weight\_summary\*\* variable   
  
weight\_summary %>%  
 select(Date,AverageWeight) %>%  
 summary()

## Date AverageWeight   
## Min. :2020-04-12 Min. :125.2   
## 1st Qu.:2020-04-19 1st Qu.:134.9   
## Median :2020-04-27 Median :135.6   
## Mean :2020-04-27 Mean :137.0   
## 3rd Qu.:2020-05-04 3rd Qu.:136.9   
## Max. :2020-05-12 Max. :147.5

# **Share**

Now that we analyzed the different data sets now we can share the insights gathered to bring visuals to the important data shown

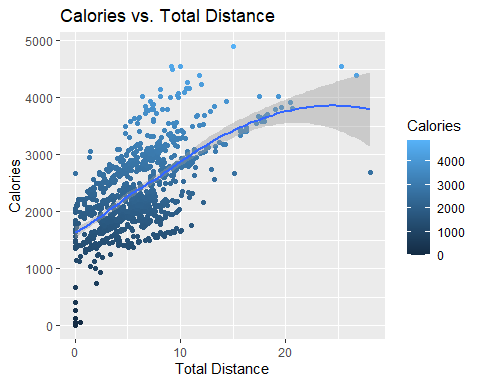
#This visual shows the relationship between Calories burned and Total Steps tracked   
  
ggplot(data = daily\_activity)+  
 geom\_point(mapping = aes(x = TotalSteps, y = Calories, color = Calories)) +  
 geom\_smooth(mapping = aes(x = TotalSteps, y = Calories))+   
 ggtitle("Calories vs. Total Steps")+  
 labs(title = "Calories vs. Total Steps",  
 x = "Total Steps", y = "Calories")

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

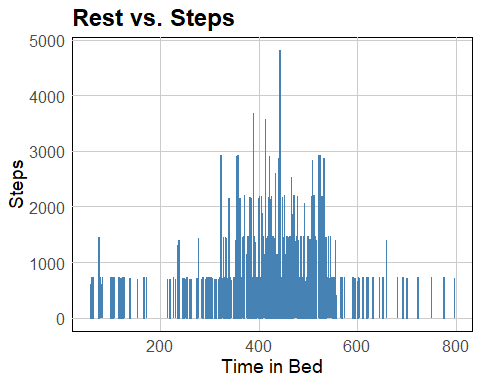


# Shows the relationship between Calories burned and Total Distance of the exercise activities   
  
ggplot(data = daily\_activity)+  
 geom\_point(mapping = aes(x = TotalDistance, y = Calories, color = Calories))+  
 geom\_smooth(mapping = aes(x = TotalDistance, y = Calories))+  
 ggtitle("Calories vs. Total Distance")+  
 labs(title = "Calories vs. Total Distance",  
 x = "Total Distance", y = "Calories")

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

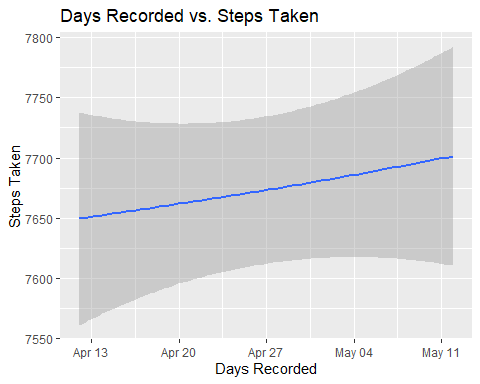


# Shows the relationship between the users \*Time Spent in Bed\* vs. \*Total Steps\*   
  
ggplot(data = combined\_bars) +  
 geom\_bar(mapping = aes(x = TotalMinutesAsleep), stat = "count", color = "steelblue", fill = "white") +  
 labs(x = "Time in Bed", y = "Steps", title = "Rest vs. Steps") +  
 theme\_minimal() +  
 theme(plot.title = element\_text(size = 18, face = "bold"),  
 axis.title = element\_text(size = 14),  
 axis.text = element\_text(size = 12),  
 panel.grid.major = element\_line(color = "gray80"),  
 panel.grid.minor = element\_blank(),  
 panel.background = element\_rect(fill = "white"))



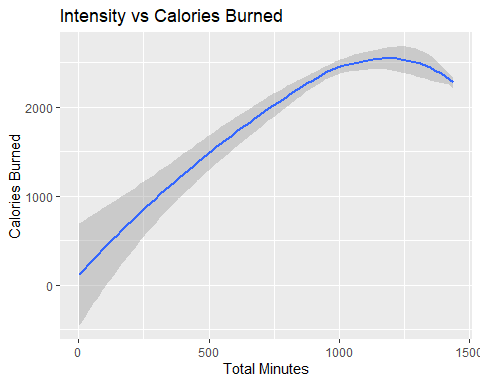
# This graph shows the \*\*rise\*\* of the users steps over time   
  
ggplot(record\_activity, aes(x = ActivityDate, y = StepTotal)) +  
 geom\_smooth() +  
 labs(x = "Days Recorded", y = "Steps Taken", title = "Days Recorded vs. Steps Taken")

## `geom\_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'



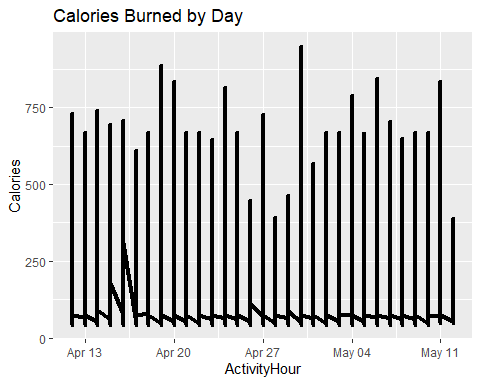
# This graph shows the correlation between \*longer exercise\* leading to \*more calories burned\*   
  
ggplot(total\_intensities, aes(x = TotalMinutes, y = Calories))+  
 geom\_smooth()+  
 labs(x = "Total Minutes", y = "Calories Burned", title = "Intensity vs Calories Burned")

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

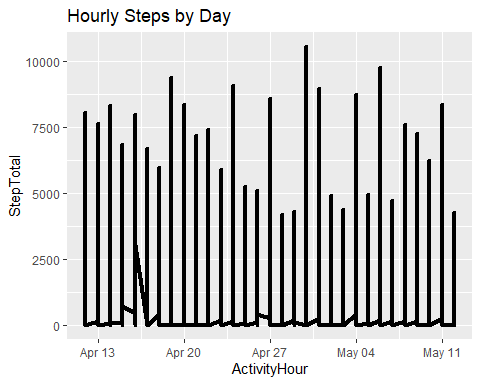


# This graph shows the amount of \*calories burned hourly each day\* between all the participants   
  
ggplot(data = hourly\_Calories, aes(x = ActivityHour, y = Calories))+  
 geom\_line(size = 1.5)+  
 labs(title = "Calories Burned by Day")

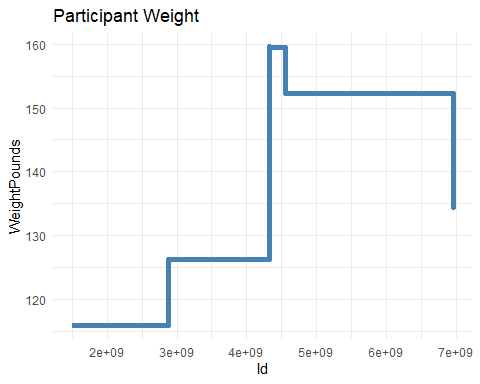
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.



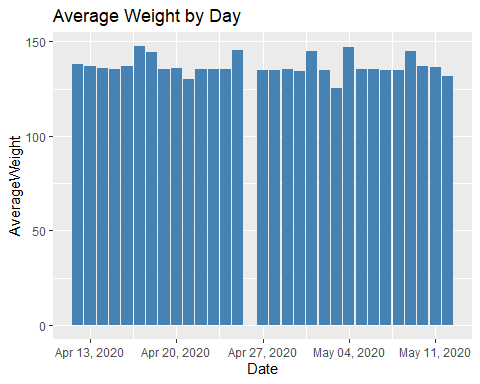
# Shows the amount of \*steps taken hourly each day\* between all the participants   
  
ggplot(data = hourly\_steps, aes(x = ActivityHour, y = StepTotal))+  
 geom\_line(size = 1.5)+  
 labs(title = "Hourly Steps by Day")



# This graph shows the \*participants weight from all of the values that were TRUE\*  
  
ggplot(data = weight\_info, aes(x = Id, y = WeightPounds)) +  
 geom\_step(size = 2, color = "steelblue") +  
 labs(title = "Participant Weight") +  
 theme\_minimal()



# Visualizes the \*weight averages for each day recorded\*   
  
ggplot(data = weight\_summary, aes(x = Date, y = AverageWeight)) +  
 geom\_col(fill = "steelblue") +  
 labs(title = "Average Weight by Day") +   
 scale\_x\_date(date\_labels = "%b %d, %Y", date\_breaks = "1 week")



# **Act**

## **Product focus: Bellabeat app**

* By analyzing the data above I came to the conclusion that as time goes on there is more *activity* and *usage*, but towards end there seems to be a downward slope of each graph, which are both shown in the **Calories vs. Total Distance** and **Calories vs. Total Steps**. This means after a month there should be more incentives and interact-able features to keep user retention on the app
* *Sleep* and *activity* seem to be more consistent in the middle levels of the chart which is shown in the **Rest vs. Steps** chart, so I think its important for users to have at least **400-500 minutes** *6-8 hours* of sleep each day. Data like this shows the opportunity to implement systems such as *sleep trackers*, *meal plans*, *fitness guides*, *interactive profiles* and many other features to keep the users returning to the app on a consistent basis
* Also, reviewing the *weight* of each user, we can see the users have a vast difference in weight between each other as shown in the **Participant Weight** and **Average Weight by Day** graphs. This means that each user is going to have unique goals so it’s important to cater to all kinds of users by considering weight and other attributes to retain usability of the smart devices and app.

## **Conclusion**

Overall, I think has a great opportunity to innovate and add helpful and informative features to the app to help retain user traffic on the app so they can help reach their goals, as well as learning and helping others along the way by including an interactive interface to the app. With the Fitness tracker market contentiously growing with large growth rate potential in the future as well, there are many ways *Bellabeat* can take advantage of this opportunity to innovate by being unique compared to other competitors. As *Bellabeat* focuses on innovation they can also focus heavily on marketing to reach a larger audience of consumers that highlights the importance using the *app* to help reach personal goals and to help others along the way.