Capstone Project, Bellabeat Casestudy

Tamakloe Vivian Anuoluwa

2024-07-05

## **INTRODUCTION**

Welcome to my Google Data Analytics Capstone project, Bellabeat Case Study! This case study will answer a business task, which follows the steps of data analysis process: Ask, Prepare, Process, Analyze, Share, and Act.

### About Bellabeat

Bellabeat is a high-tech company that manufactures health-focused smart products which are based on 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. One can say Bellabeat is a tech-driven wellness company for women.

### Bellabeats Product

* Bellabeat App
* Leaf
* Time
* Spring
* Bellabeat Membership

### Scenario

I am a junior data analyst working on the marketing analyst team at Bellabeat. I 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. This insights will help guide marketing strategy for the company.

### Product

The focused product of this analysis is **BellaBeat App**

## **ASK PHASE**

In this phase, I tried to understand the problem the marketing team has targeted, which is how a trend in smart device usage can be applied to Bellabeat customers. I will analyze smart device usage data in order to gain insight which will reveal opportunities for Bellabeat’s growth. Bellabeat marketing team will be able to target efforts to products that meets the needs of their customers based on their fitness smart devices usage.

The Stakeholders of this case study are:

* Urška Sršen: Bellabeat’s co-founder and Chief Creative Officer
* Sando Mur: Mathematician and Bellabeat’s co-founder; key member of the Bellabeat executive team
* Bellabeat marketing analytics team

### Business Task

**Focusing on Bellabeat product and analyzing how people are already using their smart devices in order to gain insights that would reveal more opportunities for growth and recommendations for Bellabeat marketing team based on trends in smart device data usage.**

## **PREPARE PHASE**

In this phase, I will download and import my dataset. I will also install and load the necessary R packages

### Downloading my dataset

The stakeholders encouraged me to use public data that explores smart device users’ daily habits. FitBit Fitness Tracker Data (CC0: Public Domain, dataset made available through Mobius). This Kaggle data set contains personal fitness tracker from fitbit users who 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. The dataset

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

### Installing the necessary packages

I installed all the necessary packages to help in my analysis, using **install.packages()** function this packages includes:

* tidyverse
* skimr
* janitor
* lubridate
* ggplot2
* dplyr
* tidyr
* here

### Loading the necessary packages

Then I proceed to loading them

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── 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(skimr)  
library(janitor)

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

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

## here() starts at C:/Users/user/Documents/BellaBeat

### Importing the datasets

I proceeded to importing the necessary datasets needed for my analysis, then I VIEWED the data so as to be able to CLEAN,FORMAT and ORGNISE in the next phase.

* **dailyActivity\_merged.csv**

Activity <- read\_csv("C:/Users/user/Documents/BellaBeat/Datasets/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.

head(Activity)

## # A tibble: 6 × 15  
## Id ActivityDate TotalSteps TotalDistance TrackerDistance  
## <dbl> <chr> <dbl> <dbl> <dbl>  
## 1 1503960366 4/12/2016 13162 8.5 8.5   
## 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  
## # ℹ 10 more variables: LoggedActivitiesDistance <dbl>,  
## # VeryActiveDistance <dbl>, ModeratelyActiveDistance <dbl>,  
## # LightActiveDistance <dbl>, SedentaryActiveDistance <dbl>,  
## # VeryActiveMinutes <dbl>, FairlyActiveMinutes <dbl>,  
## # LightlyActiveMinutes <dbl>, SedentaryMinutes <dbl>, Calories <dbl>

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)

## spc\_tbl\_ [940 × 15] (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 ...  
## $ 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 ...  
## $ LoggedActivitiesDistance: num [1:940] 0 0 0 0 0 0 0 0 0 0 ...  
## $ VeryActiveDistance : num [1:940] 1.88 1.57 2.44 2.14 2.71 ...  
## $ ModeratelyActiveDistance: num [1:940] 0.55 0.69 0.4 1.26 0.41 ...  
## $ LightActiveDistance : num [1:940] 6.06 4.71 3.91 2.83 5.04 ...  
## $ SedentaryActiveDistance : num [1:940] 0 0 0 0 0 0 0 0 0 0 ...  
## $ 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 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. Id = col\_double(),  
## .. ActivityDate = col\_character(),  
## .. TotalSteps = col\_double(),  
## .. TotalDistance = col\_double(),  
## .. TrackerDistance = col\_double(),  
## .. LoggedActivitiesDistance = col\_double(),  
## .. VeryActiveDistance = col\_double(),  
## .. ModeratelyActiveDistance = col\_double(),  
## .. LightActiveDistance = col\_double(),  
## .. SedentaryActiveDistance = col\_double(),  
## .. VeryActiveMinutes = col\_double(),  
## .. FairlyActiveMinutes = col\_double(),  
## .. LightlyActiveMinutes = col\_double(),  
## .. SedentaryMinutes = col\_double(),  
## .. Calories = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

* **dailyCarlories\_merged.csv**

Calories <- read\_csv("C:/Users/user/Documents/BellaBeat/Datasets/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.

head(Calories)

## # A tibble: 6 × 3  
## Id ActivityDay Calories  
## <dbl> <chr> <dbl>  
## 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)

## 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: chr [1:940] "4/12/2016" "4/13/2016" "4/14/2016" "4/15/2016" ...  
## $ Calories : num [1:940] 1985 1797 1776 1745 1863 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. Id = col\_double(),  
## .. ActivityDay = col\_character(),  
## .. Calories = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

* **dailyintensities\_merged.csv**

Intensity <- read\_csv("C:/Users/user/Documents/BellaBeat/Datasets/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.

head(Intensity)

## # A tibble: 6 × 10  
## Id ActivityDay SedentaryMinutes LightlyActiveMinutes FairlyActiveMinutes  
## <dbl> <chr> <dbl> <dbl> <dbl>  
## 1 1.50e9 4/12/2016 728 328 13  
## 2 1.50e9 4/13/2016 776 217 19  
## 3 1.50e9 4/14/2016 1218 181 11  
## 4 1.50e9 4/15/2016 726 209 34  
## 5 1.50e9 4/16/2016 773 221 10  
## 6 1.50e9 4/17/2016 539 164 20  
## # ℹ 5 more variables: VeryActiveMinutes <dbl>, SedentaryActiveDistance <dbl>,  
## # LightActiveDistance <dbl>, ModeratelyActiveDistance <dbl>,  
## # VeryActiveDistance <dbl>

colnames(Intensity)

## [1] "Id" "ActivityDay"   
## [3] "SedentaryMinutes" "LightlyActiveMinutes"   
## [5] "FairlyActiveMinutes" "VeryActiveMinutes"   
## [7] "SedentaryActiveDistance" "LightActiveDistance"   
## [9] "ModeratelyActiveDistance" "VeryActiveDistance"

str(Intensity)

## spc\_tbl\_ [940 × 10] (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 : chr [1:940] "4/12/2016" "4/13/2016" "4/14/2016" "4/15/2016" ...  
## $ 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 ...  
## $ SedentaryActiveDistance : num [1:940] 0 0 0 0 0 0 0 0 0 0 ...  
## $ LightActiveDistance : num [1:940] 6.06 4.71 3.91 2.83 5.04 ...  
## $ ModeratelyActiveDistance: num [1:940] 0.55 0.69 0.4 1.26 0.41 ...  
## $ VeryActiveDistance : num [1:940] 1.88 1.57 2.44 2.14 2.71 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. Id = col\_double(),  
## .. ActivityDay = col\_character(),  
## .. SedentaryMinutes = col\_double(),  
## .. LightlyActiveMinutes = col\_double(),  
## .. FairlyActiveMinutes = col\_double(),  
## .. VeryActiveMinutes = col\_double(),  
## .. SedentaryActiveDistance = col\_double(),  
## .. LightActiveDistance = col\_double(),  
## .. ModeratelyActiveDistance = col\_double(),  
## .. VeryActiveDistance = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

* **weightLogInfo\_merged.csv**

Weight <- read\_csv("C:/Users/user/Documents/BellaBeat/Datasets/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.

head(Weight)

## # A tibble: 6 × 8  
## Id Date WeightKg WeightPounds Fat BMI IsManualReport LogId  
## <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <lgl> <dbl>  
## 1 1503960366 5/2/2016 … 52.6 116. 22 22.6 TRUE 1.46e12  
## 2 1503960366 5/3/2016 … 52.6 116. NA 22.6 TRUE 1.46e12  
## 3 1927972279 4/13/2016… 134. 294. NA 47.5 FALSE 1.46e12  
## 4 2873212765 4/21/2016… 56.7 125. NA 21.5 TRUE 1.46e12  
## 5 2873212765 5/12/2016… 57.3 126. NA 21.7 TRUE 1.46e12  
## 6 4319703577 4/17/2016… 72.4 160. 25 27.5 TRUE 1.46e12

colnames(Weight)

## [1] "Id" "Date" "WeightKg" "WeightPounds"   
## [5] "Fat" "BMI" "IsManualReport" "LogId"

str(Weight)

## 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 : chr [1:67] "5/2/2016 11:59:59 PM" "5/3/2016 11:59:59 PM" "4/13/2016 1:08:52 AM" "4/21/2016 11:59:59 PM" ...  
## $ 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>

* **heartrate\_seconds\_merged.csv**

Heartrate <- read\_csv("C:/Users/user/Documents/BellaBeat/Datasets/Fitabase Data 4.12.16-5.12.16/heartrate\_seconds\_merged.csv")

## Rows: 2483658 Columns: 3  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): Time  
## dbl (2): Id, Value  
##   
## 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.

head(Heartrate)

## # A tibble: 6 × 3  
## Id Time Value  
## <dbl> <chr> <dbl>  
## 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(Heartrate)

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

str(Heartrate)

## spc\_tbl\_ [2,483,658 × 3] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ Id : num [1:2483658] 2.02e+09 2.02e+09 2.02e+09 2.02e+09 2.02e+09 ...  
## $ Time : chr [1:2483658] "4/12/2016 7:21:00 AM" "4/12/2016 7:21:05 AM" "4/12/2016 7:21:10 AM" "4/12/2016 7:21:20 AM" ...  
## $ Value: num [1:2483658] 97 102 105 103 101 95 91 93 94 93 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. Id = col\_double(),  
## .. Time = col\_character(),  
## .. Value = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

* **sleepDay\_merged.csv**

Sleep <- read\_csv("C:/Users/user/Documents/BellaBeat/Datasets/Fitabase Data 4.12.16-5.12.16/SleepDay\_merged.csv")

## Rows: 410 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.

head(Sleep)

## # A tibble: 6 × 5  
## Id SleepDay TotalSleepRecords TotalMinutesAsleep TotalTimeInBed  
## <dbl> <chr> <dbl> <dbl> <dbl>  
## 1 1503960366 4/12/2016 0:00 1 327 346  
## 2 1503960366 4/13/2016 0:00 2 384 407  
## 3 1503960366 4/15/2016 0:00 1 412 442  
## 4 1503960366 4/16/2016 0:00 2 340 367  
## 5 1503960366 4/17/2016 0:00 1 700 712  
## 6 1503960366 4/19/2016 0:00 1 304 320

colnames(Sleep)

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

str(Sleep)

## spc\_tbl\_ [410 × 5] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ Id : num [1:410] 1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...  
## $ SleepDay : chr [1:410] "4/12/2016 0:00" "4/13/2016 0:00" "4/15/2016 0:00" "4/16/2016 0:00" ...  
## $ TotalSleepRecords : num [1:410] 1 2 1 2 1 1 1 1 1 1 ...  
## $ TotalMinutesAsleep: num [1:410] 327 384 412 340 700 304 360 325 361 430 ...  
## $ TotalTimeInBed : num [1:410] 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>

## **PROCESS PHASE**

In this phase, I will process the data by cleaning,formatting and organizing the data using the following:

### glimpse() function to view each dataset

glimpse(Activity)

## Rows: 940  
## Columns: 15  
## $ Id <dbl> 1503960366, 1503960366, 1503960366, 150396036…  
## $ ActivityDate <chr> "4/12/2016", "4/13/2016", "4/14/2016", "4/15/…  
## $ TotalSteps <dbl> 13162, 10735, 10460, 9762, 12669, 9705, 13019…  
## $ TotalDistance <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9.8…  
## $ TrackerDistance <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9.8…  
## $ LoggedActivitiesDistance <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
## $ VeryActiveDistance <dbl> 1.88, 1.57, 2.44, 2.14, 2.71, 3.19, 3.25, 3.5…  
## $ ModeratelyActiveDistance <dbl> 0.55, 0.69, 0.40, 1.26, 0.41, 0.78, 0.64, 1.3…  
## $ LightActiveDistance <dbl> 6.06, 4.71, 3.91, 2.83, 5.04, 2.51, 4.71, 5.0…  
## $ SedentaryActiveDistance <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
## $ VeryActiveMinutes <dbl> 25, 21, 30, 29, 36, 38, 42, 50, 28, 19, 66, 4…  
## $ FairlyActiveMinutes <dbl> 13, 19, 11, 34, 10, 20, 16, 31, 12, 8, 27, 21…  
## $ LightlyActiveMinutes <dbl> 328, 217, 181, 209, 221, 164, 233, 264, 205, …  
## $ SedentaryMinutes <dbl> 728, 776, 1218, 726, 773, 539, 1149, 775, 818…  
## $ Calories <dbl> 1985, 1797, 1776, 1745, 1863, 1728, 1921, 203…

glimpse(Calories)

## Rows: 940  
## Columns: 3  
## $ Id <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 1503960366…  
## $ ActivityDay <chr> "4/12/2016", "4/13/2016", "4/14/2016", "4/15/2016", "4/16/…  
## $ Calories <dbl> 1985, 1797, 1776, 1745, 1863, 1728, 1921, 2035, 1786, 1775…

glimpse(Intensity)

## Rows: 940  
## Columns: 10  
## $ Id <dbl> 1503960366, 1503960366, 1503960366, 150396036…  
## $ ActivityDay <chr> "4/12/2016", "4/13/2016", "4/14/2016", "4/15/…  
## $ SedentaryMinutes <dbl> 728, 776, 1218, 726, 773, 539, 1149, 775, 818…  
## $ LightlyActiveMinutes <dbl> 328, 217, 181, 209, 221, 164, 233, 264, 205, …  
## $ FairlyActiveMinutes <dbl> 13, 19, 11, 34, 10, 20, 16, 31, 12, 8, 27, 21…  
## $ VeryActiveMinutes <dbl> 25, 21, 30, 29, 36, 38, 42, 50, 28, 19, 66, 4…  
## $ SedentaryActiveDistance <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
## $ LightActiveDistance <dbl> 6.06, 4.71, 3.91, 2.83, 5.04, 2.51, 4.71, 5.0…  
## $ ModeratelyActiveDistance <dbl> 0.55, 0.69, 0.40, 1.26, 0.41, 0.78, 0.64, 1.3…  
## $ VeryActiveDistance <dbl> 1.88, 1.57, 2.44, 2.14, 2.71, 3.19, 3.25, 3.5…

glimpse(Weight)

## Rows: 67  
## Columns: 8  
## $ Id <dbl> 1503960366, 1503960366, 1927972279, 2873212765, 2873212…  
## $ Date <chr> "5/2/2016 11:59:59 PM", "5/3/2016 11:59:59 PM", "4/13/2…  
## $ WeightKg <dbl> 52.6, 52.6, 133.5, 56.7, 57.3, 72.4, 72.3, 69.7, 70.3, …  
## $ WeightPounds <dbl> 115.9631, 115.9631, 294.3171, 125.0021, 126.3249, 159.6…  
## $ Fat <dbl> 22, NA, NA, NA, NA, 25, NA, NA, NA, NA, NA, NA, NA, NA,…  
## $ BMI <dbl> 22.65, 22.65, 47.54, 21.45, 21.69, 27.45, 27.38, 27.25,…  
## $ IsManualReport <lgl> TRUE, TRUE, FALSE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, …  
## $ LogId <dbl> 1.462234e+12, 1.462320e+12, 1.460510e+12, 1.461283e+12,…

glimpse(Heartrate)

## Rows: 2,483,658  
## Columns: 3  
## $ Id <dbl> 2022484408, 2022484408, 2022484408, 2022484408, 2022484408, 2022…  
## $ Time <chr> "4/12/2016 7:21:00 AM", "4/12/2016 7:21:05 AM", "4/12/2016 7:21:…  
## $ Value <dbl> 97, 102, 105, 103, 101, 95, 91, 93, 94, 93, 92, 89, 83, 61, 60, …

glimpse(Sleep)

## Rows: 410  
## Columns: 5  
## $ Id <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 150…  
## $ SleepDay <chr> "4/12/2016 0:00", "4/13/2016 0:00", "4/15/2016 0:00…  
## $ TotalSleepRecords <dbl> 1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …  
## $ TotalMinutesAsleep <dbl> 327, 384, 412, 340, 700, 304, 360, 325, 361, 430, 2…  
## $ TotalTimeInBed <dbl> 346, 407, 442, 367, 712, 320, 377, 364, 384, 449, 3…

### clean\_names() function to clean the names of each dataset

clean\_names(Activity)

## # A tibble: 940 × 15  
## id activity\_date total\_steps total\_distance tracker\_distance  
## <dbl> <chr> <dbl> <dbl> <dbl>  
## 1 1503960366 4/12/2016 13162 8.5 8.5   
## 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  
## 7 1503960366 4/18/2016 13019 8.59 8.59  
## 8 1503960366 4/19/2016 15506 9.88 9.88  
## 9 1503960366 4/20/2016 10544 6.68 6.68  
## 10 1503960366 4/21/2016 9819 6.34 6.34  
## # ℹ 930 more rows  
## # ℹ 10 more variables: logged\_activities\_distance <dbl>,  
## # very\_active\_distance <dbl>, moderately\_active\_distance <dbl>,  
## # light\_active\_distance <dbl>, sedentary\_active\_distance <dbl>,  
## # very\_active\_minutes <dbl>, fairly\_active\_minutes <dbl>,  
## # lightly\_active\_minutes <dbl>, sedentary\_minutes <dbl>, calories <dbl>

clean\_names(Calories)

## # A tibble: 940 × 3  
## id activity\_day calories  
## <dbl> <chr> <dbl>  
## 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  
## 7 1503960366 4/18/2016 1921  
## 8 1503960366 4/19/2016 2035  
## 9 1503960366 4/20/2016 1786  
## 10 1503960366 4/21/2016 1775  
## # ℹ 930 more rows

clean\_names(Intensity)

## # A tibble: 940 × 10  
## id activity\_day sedentary\_minutes lightly\_active\_minutes  
## <dbl> <chr> <dbl> <dbl>  
## 1 1503960366 4/12/2016 728 328  
## 2 1503960366 4/13/2016 776 217  
## 3 1503960366 4/14/2016 1218 181  
## 4 1503960366 4/15/2016 726 209  
## 5 1503960366 4/16/2016 773 221  
## 6 1503960366 4/17/2016 539 164  
## 7 1503960366 4/18/2016 1149 233  
## 8 1503960366 4/19/2016 775 264  
## 9 1503960366 4/20/2016 818 205  
## 10 1503960366 4/21/2016 838 211  
## # ℹ 930 more rows  
## # ℹ 6 more variables: fairly\_active\_minutes <dbl>, very\_active\_minutes <dbl>,  
## # sedentary\_active\_distance <dbl>, light\_active\_distance <dbl>,  
## # moderately\_active\_distance <dbl>, very\_active\_distance <dbl>

clean\_names(Weight)

## # A tibble: 67 × 8  
## id date weight\_kg weight\_pounds fat bmi is\_manual\_report log\_id  
## <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <lgl> <dbl>  
## 1 1503960366 5/2/… 52.6 116. 22 22.6 TRUE 1.46e12  
## 2 1503960366 5/3/… 52.6 116. NA 22.6 TRUE 1.46e12  
## 3 1927972279 4/13… 134. 294. NA 47.5 FALSE 1.46e12  
## 4 2873212765 4/21… 56.7 125. NA 21.5 TRUE 1.46e12  
## 5 2873212765 5/12… 57.3 126. NA 21.7 TRUE 1.46e12  
## 6 4319703577 4/17… 72.4 160. 25 27.5 TRUE 1.46e12  
## 7 4319703577 5/4/… 72.3 159. NA 27.4 TRUE 1.46e12  
## 8 4558609924 4/18… 69.7 154. NA 27.2 TRUE 1.46e12  
## 9 4558609924 4/25… 70.3 155. NA 27.5 TRUE 1.46e12  
## 10 4558609924 5/1/… 69.9 154. NA 27.3 TRUE 1.46e12  
## # ℹ 57 more rows

clean\_names(Heartrate)

## # A tibble: 2,483,658 × 3  
## id time value  
## <dbl> <chr> <dbl>  
## 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  
## 7 2022484408 4/12/2016 7:22:10 AM 91  
## 8 2022484408 4/12/2016 7:22:15 AM 93  
## 9 2022484408 4/12/2016 7:22:20 AM 94  
## 10 2022484408 4/12/2016 7:22:25 AM 93  
## # ℹ 2,483,648 more rows

clean\_names(Sleep)

## # A tibble: 410 × 5  
## id sleep\_day total\_sleep\_records total\_minutes\_asleep total\_time\_in\_bed  
## <dbl> <chr> <dbl> <dbl> <dbl>  
## 1 1.50e9 4/12/201… 1 327 346  
## 2 1.50e9 4/13/201… 2 384 407  
## 3 1.50e9 4/15/201… 1 412 442  
## 4 1.50e9 4/16/201… 2 340 367  
## 5 1.50e9 4/17/201… 1 700 712  
## 6 1.50e9 4/19/201… 1 304 320  
## 7 1.50e9 4/20/201… 1 360 377  
## 8 1.50e9 4/21/201… 1 325 364  
## 9 1.50e9 4/23/201… 1 361 384  
## 10 1.50e9 4/24/201… 1 430 449  
## # ℹ 400 more rows

### skim\_without\_chart() function to check for empty or extra space, missing cells

skim\_without\_charts(Activity)

Data summary

|  |  |
| --- | --- |
| Name | Activity |
| Number of rows | 940 |
| Number of columns | 15 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 1 |
| numeric | 14 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ActivityDate | 0 | 1 | 8 | 9 | 0 | 31 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Id | 0 | 1 | 4.855407e+09 | 2.424805e+09 | 1503960366 | 2.320127e+09 | 4.445115e+09 | 6.962181e+09 | 8.877689e+09 |
| TotalSteps | 0 | 1 | 7.637910e+03 | 5.087150e+03 | 0 | 3.789750e+03 | 7.405500e+03 | 1.072700e+04 | 3.601900e+04 |
| TotalDistance | 0 | 1 | 5.490000e+00 | 3.920000e+00 | 0 | 2.620000e+00 | 5.240000e+00 | 7.710000e+00 | 2.803000e+01 |
| TrackerDistance | 0 | 1 | 5.480000e+00 | 3.910000e+00 | 0 | 2.620000e+00 | 5.240000e+00 | 7.710000e+00 | 2.803000e+01 |
| LoggedActivitiesDistance | 0 | 1 | 1.100000e-01 | 6.200000e-01 | 0 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 | 4.940000e+00 |
| VeryActiveDistance | 0 | 1 | 1.500000e+00 | 2.660000e+00 | 0 | 0.000000e+00 | 2.100000e-01 | 2.050000e+00 | 2.192000e+01 |
| ModeratelyActiveDistance | 0 | 1 | 5.700000e-01 | 8.800000e-01 | 0 | 0.000000e+00 | 2.400000e-01 | 8.000000e-01 | 6.480000e+00 |
| LightActiveDistance | 0 | 1 | 3.340000e+00 | 2.040000e+00 | 0 | 1.950000e+00 | 3.360000e+00 | 4.780000e+00 | 1.071000e+01 |
| SedentaryActiveDistance | 0 | 1 | 0.000000e+00 | 1.000000e-02 | 0 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 | 1.100000e-01 |
| VeryActiveMinutes | 0 | 1 | 2.116000e+01 | 3.284000e+01 | 0 | 0.000000e+00 | 4.000000e+00 | 3.200000e+01 | 2.100000e+02 |
| FairlyActiveMinutes | 0 | 1 | 1.356000e+01 | 1.999000e+01 | 0 | 0.000000e+00 | 6.000000e+00 | 1.900000e+01 | 1.430000e+02 |
| LightlyActiveMinutes | 0 | 1 | 1.928100e+02 | 1.091700e+02 | 0 | 1.270000e+02 | 1.990000e+02 | 2.640000e+02 | 5.180000e+02 |
| SedentaryMinutes | 0 | 1 | 9.912100e+02 | 3.012700e+02 | 0 | 7.297500e+02 | 1.057500e+03 | 1.229500e+03 | 1.440000e+03 |
| Calories | 0 | 1 | 2.303610e+03 | 7.181700e+02 | 0 | 1.828500e+03 | 2.134000e+03 | 2.793250e+03 | 4.900000e+03 |

skim\_without\_charts(Calories)

Data summary

|  |  |
| --- | --- |
| Name | Calories |
| Number of rows | 940 |
| Number of columns | 3 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 1 |
| numeric | 2 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ActivityDay | 0 | 1 | 8 | 9 | 0 | 31 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Id | 0 | 1 | 4.855407e+09 | 2.424805e+09 | 1503960366 | 2320127002.0 | 4445114986 | 6.962181e+09 | 8877689391 |
| Calories | 0 | 1 | 2.303610e+03 | 7.181700e+02 | 0 | 1828.5 | 2134 | 2.793250e+03 | 4900 |

skim\_without\_charts(Intensity)

Data summary

|  |  |
| --- | --- |
| Name | Intensity |
| Number of rows | 940 |
| Number of columns | 10 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 1 |
| numeric | 9 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ActivityDay | 0 | 1 | 8 | 9 | 0 | 31 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Id | 0 | 1 | 4.855407e+09 | 2.424805e+09 | 1503960366 | 2.320127e+09 | 4.445115e+09 | 6.962181e+09 | 8.877689e+09 |
| SedentaryMinutes | 0 | 1 | 9.912100e+02 | 3.012700e+02 | 0 | 7.297500e+02 | 1.057500e+03 | 1.229500e+03 | 1.440000e+03 |
| LightlyActiveMinutes | 0 | 1 | 1.928100e+02 | 1.091700e+02 | 0 | 1.270000e+02 | 1.990000e+02 | 2.640000e+02 | 5.180000e+02 |
| FairlyActiveMinutes | 0 | 1 | 1.356000e+01 | 1.999000e+01 | 0 | 0.000000e+00 | 6.000000e+00 | 1.900000e+01 | 1.430000e+02 |
| VeryActiveMinutes | 0 | 1 | 2.116000e+01 | 3.284000e+01 | 0 | 0.000000e+00 | 4.000000e+00 | 3.200000e+01 | 2.100000e+02 |
| SedentaryActiveDistance | 0 | 1 | 0.000000e+00 | 1.000000e-02 | 0 | 0.000000e+00 | 0.000000e+00 | 0.000000e+00 | 1.100000e-01 |
| LightActiveDistance | 0 | 1 | 3.340000e+00 | 2.040000e+00 | 0 | 1.950000e+00 | 3.360000e+00 | 4.780000e+00 | 1.071000e+01 |
| ModeratelyActiveDistance | 0 | 1 | 5.700000e-01 | 8.800000e-01 | 0 | 0.000000e+00 | 2.400000e-01 | 8.000000e-01 | 6.480000e+00 |
| VeryActiveDistance | 0 | 1 | 1.500000e+00 | 2.660000e+00 | 0 | 0.000000e+00 | 2.100000e-01 | 2.050000e+00 | 2.192000e+01 |

skim\_without\_charts(Weight)

Data summary

|  |  |
| --- | --- |
| Name | Weight |
| Number of rows | 67 |
| Number of columns | 8 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 1 |
| logical | 1 |
| numeric | 6 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Date | 0 | 1 | 19 | 21 | 0 | 56 | 0 |

**Variable type: logical**

| skim\_variable | n\_missing | complete\_rate | mean | count |
| --- | --- | --- | --- | --- |
| IsManualReport | 0 | 1 | 0.61 | TRU: 41, FAL: 26 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Id | 0 | 1.00 | 7.009282e+09 | 1.950322e+09 | 1.503960e+09 | 6.962181e+09 | 6.962181e+09 | 8.877689e+09 | 8.877689e+09 |
| WeightKg | 0 | 1.00 | 7.204000e+01 | 1.392000e+01 | 5.260000e+01 | 6.140000e+01 | 6.250000e+01 | 8.505000e+01 | 1.335000e+02 |
| WeightPounds | 0 | 1.00 | 1.588100e+02 | 3.070000e+01 | 1.159600e+02 | 1.353600e+02 | 1.377900e+02 | 1.875000e+02 | 2.943200e+02 |
| Fat | 65 | 0.03 | 2.350000e+01 | 2.120000e+00 | 2.200000e+01 | 2.275000e+01 | 2.350000e+01 | 2.425000e+01 | 2.500000e+01 |
| BMI | 0 | 1.00 | 2.519000e+01 | 3.070000e+00 | 2.145000e+01 | 2.396000e+01 | 2.439000e+01 | 2.556000e+01 | 4.754000e+01 |
| LogId | 0 | 1.00 | 1.461772e+12 | 7.829948e+08 | 1.460444e+12 | 1.461079e+12 | 1.461802e+12 | 1.462375e+12 | 1.463098e+12 |

skim\_without\_charts(Heartrate)

Data summary

|  |  |
| --- | --- |
| Name | Heartrate |
| Number of rows | 2483658 |
| Number of columns | 3 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 1 |
| numeric | 2 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Time | 0 | 1 | 19 | 21 | 0 | 961274 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Id | 0 | 1 | 5.513765e+09 | 1950223761.0 | 2022484408 | 4388161847 | 5553957443 | 6962181067 | 8877689391 |
| Value | 0 | 1 | 7.733000e+01 | 19.4 | 36 | 63 | 73 | 88 | 203 |

skim\_without\_charts(Sleep)

Data summary

|  |  |
| --- | --- |
| Name | Sleep |
| Number of rows | 410 |
| Number of columns | 5 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 1 |
| numeric | 4 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| SleepDay | 0 | 1 | 13 | 14 | 0 | 31 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Id | 0 | 1 | 4.994963e+09 | 2.060863e+09 | 1503960366 | 3.977334e+09 | 4702921684.0 | 6962181067 | 8792009665 |
| TotalSleepRecords | 0 | 1 | 1.120000e+00 | 3.500000e-01 | 1 | 1.000000e+00 | 1.0 | 1 | 3 |
| TotalMinutesAsleep | 0 | 1 | 4.191700e+02 | 1.186400e+02 | 58 | 3.610000e+02 | 432.5 | 490 | 796 |
| TotalTimeInBed | 0 | 1 | 4.584800e+02 | 1.274600e+02 | 61 | 4.037500e+02 | 463.0 | 526 | 961 |

**Due to the numerous missing values in Weight dataset, column Fat was removed**

Weight <- Weight[, - which(names(Weight) == "Fat")]  
head(Weight)

## # A tibble: 6 × 7  
## Id Date WeightKg WeightPounds BMI IsManualReport LogId  
## <dbl> <chr> <dbl> <dbl> <dbl> <lgl> <dbl>  
## 1 1503960366 5/2/2016 11:59:… 52.6 116. 22.6 TRUE 1.46e12  
## 2 1503960366 5/3/2016 11:59:… 52.6 116. 22.6 TRUE 1.46e12  
## 3 1927972279 4/13/2016 1:08:… 134. 294. 47.5 FALSE 1.46e12  
## 4 2873212765 4/21/2016 11:59… 56.7 125. 21.5 TRUE 1.46e12  
## 5 2873212765 5/12/2016 11:59… 57.3 126. 21.7 TRUE 1.46e12  
## 6 4319703577 4/17/2016 11:59… 72.4 160. 27.5 TRUE 1.46e12

### distinct\_all() to remove duplicate rows

distinct\_all(Activity)

## # A tibble: 940 × 15  
## Id ActivityDate TotalSteps TotalDistance TrackerDistance  
## <dbl> <chr> <dbl> <dbl> <dbl>  
## 1 1503960366 4/12/2016 13162 8.5 8.5   
## 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  
## 7 1503960366 4/18/2016 13019 8.59 8.59  
## 8 1503960366 4/19/2016 15506 9.88 9.88  
## 9 1503960366 4/20/2016 10544 6.68 6.68  
## 10 1503960366 4/21/2016 9819 6.34 6.34  
## # ℹ 930 more rows  
## # ℹ 10 more variables: LoggedActivitiesDistance <dbl>,  
## # VeryActiveDistance <dbl>, ModeratelyActiveDistance <dbl>,  
## # LightActiveDistance <dbl>, SedentaryActiveDistance <dbl>,  
## # VeryActiveMinutes <dbl>, FairlyActiveMinutes <dbl>,  
## # LightlyActiveMinutes <dbl>, SedentaryMinutes <dbl>, Calories <dbl>

distinct\_all(Calories)

## # A tibble: 940 × 3  
## Id ActivityDay Calories  
## <dbl> <chr> <dbl>  
## 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  
## 7 1503960366 4/18/2016 1921  
## 8 1503960366 4/19/2016 2035  
## 9 1503960366 4/20/2016 1786  
## 10 1503960366 4/21/2016 1775  
## # ℹ 930 more rows

distinct\_all(Intensity)

## # A tibble: 940 × 10  
## Id ActivityDay SedentaryMinutes LightlyActiveMinutes FairlyActiveMinutes  
## <dbl> <chr> <dbl> <dbl> <dbl>  
## 1 1.50e9 4/12/2016 728 328 13  
## 2 1.50e9 4/13/2016 776 217 19  
## 3 1.50e9 4/14/2016 1218 181 11  
## 4 1.50e9 4/15/2016 726 209 34  
## 5 1.50e9 4/16/2016 773 221 10  
## 6 1.50e9 4/17/2016 539 164 20  
## 7 1.50e9 4/18/2016 1149 233 16  
## 8 1.50e9 4/19/2016 775 264 31  
## 9 1.50e9 4/20/2016 818 205 12  
## 10 1.50e9 4/21/2016 838 211 8  
## # ℹ 930 more rows  
## # ℹ 5 more variables: VeryActiveMinutes <dbl>, SedentaryActiveDistance <dbl>,  
## # LightActiveDistance <dbl>, ModeratelyActiveDistance <dbl>,  
## # VeryActiveDistance <dbl>

distinct\_all(Weight)

## # A tibble: 67 × 7  
## Id Date WeightKg WeightPounds BMI IsManualReport LogId  
## <dbl> <chr> <dbl> <dbl> <dbl> <lgl> <dbl>  
## 1 1503960366 5/2/2016 11:59… 52.6 116. 22.6 TRUE 1.46e12  
## 2 1503960366 5/3/2016 11:59… 52.6 116. 22.6 TRUE 1.46e12  
## 3 1927972279 4/13/2016 1:08… 134. 294. 47.5 FALSE 1.46e12  
## 4 2873212765 4/21/2016 11:5… 56.7 125. 21.5 TRUE 1.46e12  
## 5 2873212765 5/12/2016 11:5… 57.3 126. 21.7 TRUE 1.46e12  
## 6 4319703577 4/17/2016 11:5… 72.4 160. 27.5 TRUE 1.46e12  
## 7 4319703577 5/4/2016 11:59… 72.3 159. 27.4 TRUE 1.46e12  
## 8 4558609924 4/18/2016 11:5… 69.7 154. 27.2 TRUE 1.46e12  
## 9 4558609924 4/25/2016 11:5… 70.3 155. 27.5 TRUE 1.46e12  
## 10 4558609924 5/1/2016 11:59… 69.9 154. 27.3 TRUE 1.46e12  
## # ℹ 57 more rows

distinct\_all(Heartrate)

## # A tibble: 2,483,658 × 3  
## Id Time Value  
## <dbl> <chr> <dbl>  
## 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  
## 7 2022484408 4/12/2016 7:22:10 AM 91  
## 8 2022484408 4/12/2016 7:22:15 AM 93  
## 9 2022484408 4/12/2016 7:22:20 AM 94  
## 10 2022484408 4/12/2016 7:22:25 AM 93  
## # ℹ 2,483,648 more rows

distinct\_all(Sleep)

## # A tibble: 410 × 5  
## Id SleepDay TotalSleepRecords TotalMinutesAsleep TotalTimeInBed  
## <dbl> <chr> <dbl> <dbl> <dbl>  
## 1 1503960366 4/12/2016 0:00 1 327 346  
## 2 1503960366 4/13/2016 0:00 2 384 407  
## 3 1503960366 4/15/2016 0:00 1 412 442  
## 4 1503960366 4/16/2016 0:00 2 340 367  
## 5 1503960366 4/17/2016 0:00 1 700 712  
## 6 1503960366 4/19/2016 0:00 1 304 320  
## 7 1503960366 4/20/2016 0:00 1 360 377  
## 8 1503960366 4/21/2016 0:00 1 325 364  
## 9 1503960366 4/23/2016 0:00 1 361 384  
## 10 1503960366 4/24/2016 0:00 1 430 449  
## # ℹ 400 more rows

**Three duplicate rows were removed from Sleep dataset**

### Formatting Datatype in Datasets

**Activity**

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

## spc\_tbl\_ [940 × 15] (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 ...  
## $ ActivityDate : Date[1:940], format: "2016-04-12" "2016-04-13" ...  
## $ 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 ...  
## $ LoggedActivitiesDistance: num [1:940] 0 0 0 0 0 0 0 0 0 0 ...  
## $ VeryActiveDistance : num [1:940] 1.88 1.57 2.44 2.14 2.71 ...  
## $ ModeratelyActiveDistance: num [1:940] 0.55 0.69 0.4 1.26 0.41 ...  
## $ LightActiveDistance : num [1:940] 6.06 4.71 3.91 2.83 5.04 ...  
## $ SedentaryActiveDistance : num [1:940] 0 0 0 0 0 0 0 0 0 0 ...  
## $ 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 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. Id = col\_double(),  
## .. ActivityDate = col\_character(),  
## .. TotalSteps = col\_double(),  
## .. TotalDistance = col\_double(),  
## .. TrackerDistance = col\_double(),  
## .. LoggedActivitiesDistance = col\_double(),  
## .. VeryActiveDistance = col\_double(),  
## .. ModeratelyActiveDistance = col\_double(),  
## .. LightActiveDistance = col\_double(),  
## .. SedentaryActiveDistance = col\_double(),  
## .. VeryActiveMinutes = col\_double(),  
## .. FairlyActiveMinutes = col\_double(),  
## .. LightlyActiveMinutes = col\_double(),  
## .. SedentaryMinutes = col\_double(),  
## .. Calories = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

**Calories**

Calories$ActivityDay=as.Date(Calories$ActivityDay, format="%m/%d/%Y", tz=Sys.timezone())  
str(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: "2016-04-12" "2016-04-13" ...  
## $ Calories : num [1:940] 1985 1797 1776 1745 1863 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. Id = col\_double(),  
## .. ActivityDay = col\_character(),  
## .. Calories = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

**Intensity**

Intensity$ActivityDay=as.Date(Intensity$ActivityDay, format="%m/%d/%Y", tz=Sys.timezone())  
str(Intensity)

## spc\_tbl\_ [940 × 10] (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: "2016-04-12" "2016-04-13" ...  
## $ 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 ...  
## $ SedentaryActiveDistance : num [1:940] 0 0 0 0 0 0 0 0 0 0 ...  
## $ LightActiveDistance : num [1:940] 6.06 4.71 3.91 2.83 5.04 ...  
## $ ModeratelyActiveDistance: num [1:940] 0.55 0.69 0.4 1.26 0.41 ...  
## $ VeryActiveDistance : num [1:940] 1.88 1.57 2.44 2.14 2.71 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. Id = col\_double(),  
## .. ActivityDay = col\_character(),  
## .. SedentaryMinutes = col\_double(),  
## .. LightlyActiveMinutes = col\_double(),  
## .. FairlyActiveMinutes = col\_double(),  
## .. VeryActiveMinutes = col\_double(),  
## .. SedentaryActiveDistance = col\_double(),  
## .. LightActiveDistance = col\_double(),  
## .. ModeratelyActiveDistance = col\_double(),  
## .. VeryActiveDistance = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

## **ANALYSIS PHASE**

### Number of user in each dataset using smart device.

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

## # A tibble: 1 × 1  
## Activity\_participants  
## <int>  
## 1 33

n\_distinct(Calories$Id)

## [1] 33

n\_distinct(Intensity$Id)

## [1] 33

n\_distinct(Weight$Id)

## [1] 8

n\_distinct(Heartrate$Id)

## [1] 14

n\_distinct(Sleep$Id)

## [1] 24

There are 33 users in Activity, Calories and Intensity datasets, 24 in Sleep dataset, 14 in Heartrate dataset and 8 in weight dataset. In order not to be bias and to be able to make analysis on a good sample size, further analysis will be done with Activity dataset along with Calories,Intensity and Sleep datasets.

### Statistical Summary of each Dataset

**Note:** Sedentary means not moving, being still **Activity**

Activity %>%   
 select(TotalSteps,  
 TotalDistance,  
 SedentaryMinutes, Calories) %>%  
 summary()

## 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

**Calories**

Calories %>%  
 select(Calories) %>%  
 summary()

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

**Activity Minutes in Intensity Dataset**

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

## VeryActiveMinutes FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes  
## Min. : 0.00 Min. : 0.00 Min. : 0.0 Min. : 0.0   
## 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.:127.0 1st Qu.: 729.8   
## Median : 4.00 Median : 6.00 Median :199.0 Median :1057.5   
## Mean : 21.16 Mean : 13.56 Mean :192.8 Mean : 991.2   
## 3rd Qu.: 32.00 3rd Qu.: 19.00 3rd Qu.:264.0 3rd Qu.:1229.5   
## Max. :210.00 Max. :143.00 Max. :518.0 Max. :1440.0

**Activity Distance in Intensity Dataset**

Intensity %>%  
 select(SedentaryActiveDistance,LightActiveDistance,ModeratelyActiveDistance,VeryActiveDistance) %>%  
 summary()

## SedentaryActiveDistance LightActiveDistance ModeratelyActiveDistance  
## Min. :0.000000 Min. : 0.000 Min. :0.0000   
## 1st Qu.:0.000000 1st Qu.: 1.945 1st Qu.:0.0000   
## Median :0.000000 Median : 3.365 Median :0.2400   
## Mean :0.001606 Mean : 3.341 Mean :0.5675   
## 3rd Qu.:0.000000 3rd Qu.: 4.782 3rd Qu.:0.8000   
## Max. :0.110000 Max. :10.710 Max. :6.4800   
## VeryActiveDistance  
## Min. : 0.000   
## 1st Qu.: 0.000   
## Median : 0.210   
## Mean : 1.503   
## 3rd Qu.: 2.053   
## Max. :21.920

**Sleep**

Sleep %>%  
 select(TotalSleepRecords, TotalMinutesAsleep, TotalTimeInBed) %>%  
 summary()

## TotalSleepRecords TotalMinutesAsleep TotalTimeInBed   
## Min. :1.00 Min. : 58.0 Min. : 61.0   
## 1st Qu.:1.00 1st Qu.:361.0 1st Qu.:403.8   
## Median :1.00 Median :432.5 Median :463.0   
## Mean :1.12 Mean :419.2 Mean :458.5   
## 3rd Qu.:1.00 3rd Qu.:490.0 3rd Qu.:526.0   
## Max. :3.00 Max. :796.0 Max. :961.0

### **Observations from Analysis**

The statistical summary aided me to observe negative points in correspondence with the user health. Strategies has to be put in place to work on these observation so as to better the user health

**Activity dataset**

* The Maximum Sedentary time is high 1440mins(24hrs)
* The Average Sedentary time is high 991.2mins(16hours and 52seconds)
* The mean Total Distance is low 5.490
* The average Total steps is low 7630

**Intensity dataset**

* Majority of these Users are lightly active
* Majority of these Users have a high sedentary time
* Majority of these Users posses lightly-active-distance

**Sleep dataset**

* Average sleep time of almost 7hours is observed
* A User slept for 796 mins(over 13hours) as recorded by Maximum Total Minutes Asleep
* A User was in bed for 961 mins(over 16hours) as recorded by Maximum Total Time in bed

### Merging Datasets

Activity dataset has columns that can be found in Calories and Intensity dataset thereby relating them. But columns in Sleep Dataset can’t be found in Activity dataset. So I will merge Activity and Sleep datasets together using OUTER JOIN because Activity dataset has 33 Users and Sleep dataset has 24 Users so as to have details of all 33 Users.

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

## [1] 33

head(Combined\_data\_outer)

## Id SleepDay TotalSleepRecords TotalMinutesAsleep TotalTimeInBed  
## 1 1503960366 4/12/2016 0:00 1 327 346  
## 2 1503960366 4/12/2016 0:00 1 327 346  
## 3 1503960366 4/12/2016 0:00 1 327 346  
## 4 1503960366 4/12/2016 0:00 1 327 346  
## 5 1503960366 4/12/2016 0:00 1 327 346  
## 6 1503960366 4/12/2016 0:00 1 327 346  
## ActivityDate TotalSteps TotalDistance TrackerDistance  
## 1 2016-05-07 11992 7.71 7.71  
## 2 2016-05-06 12159 8.03 8.03  
## 3 2016-05-01 10602 6.81 6.81  
## 4 2016-04-30 14673 9.25 9.25  
## 5 2016-04-12 13162 8.50 8.50  
## 6 2016-04-13 10735 6.97 6.97  
## LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance  
## 1 0 2.46 2.12  
## 2 0 1.97 0.25  
## 3 0 2.29 1.60  
## 4 0 3.56 1.42  
## 5 0 1.88 0.55  
## 6 0 1.57 0.69  
## LightActiveDistance SedentaryActiveDistance VeryActiveMinutes  
## 1 3.13 0 37  
## 2 5.81 0 24  
## 3 2.92 0 33  
## 4 4.27 0 52  
## 5 6.06 0 25  
## 6 4.71 0 21  
## FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories  
## 1 46 175 833 1821  
## 2 6 289 754 1896  
## 3 35 246 730 1820  
## 4 34 217 712 1947  
## 5 13 328 728 1985  
## 6 19 217 776 1797

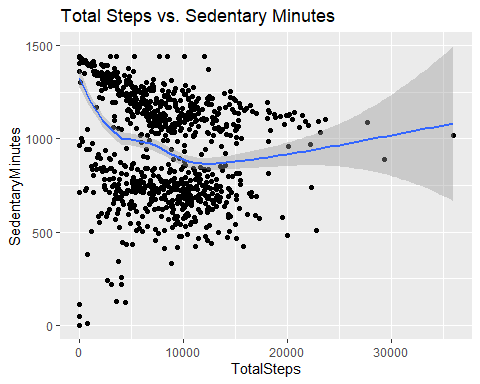
## **SHARE PHASE**

In this phase, i will visualize and show relationship between the data based on my analysis so far using the package “ggplot2” which has being loaded at the beginning of the project.

### Relationship between Total Steps and Sedentary Mintues

ggplot(data=Activity) + geom\_point(mapping = aes(x=TotalSteps, y=SedentaryMinutes)) + geom\_smooth(mapping = aes(x=TotalSteps, y=SedentaryMinutes)) + labs(title="Total Steps vs. Sedentary Minutes")

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



ggsave('Total\_Steps\_vs\_Sedentary\_Mintues\_Chart.png')

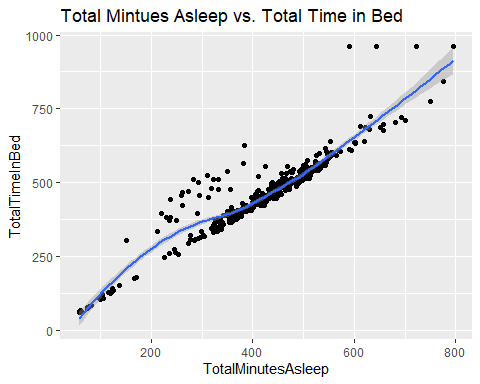
## Saving 5 x 4 in image  
## `geom\_smooth()` using method = 'loess' and formula = 'y ~ x'

The chart above shows a negative relationship between Total Steps and Sedentary Minutes, i.e the higher the Sedentary Minutes, the lower the Total steps. So therefore, to increase the Total Step of Bellabeat Users, sedentary time must be reduced

### Relationship between Total Minutes Asleep and Total time in Bed

ggplot(data=Sleep) + geom\_point(mapping = aes(x=TotalMinutesAsleep, y=TotalTimeInBed)) + geom\_smooth(mapping = aes(x=TotalMinutesAsleep, y=TotalTimeInBed)) + labs(title="Total Mintues Asleep vs. Total Time in Bed")

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



ggsave('Total\_Minutes\_Asleep\_vs\_Total\_time\_in\_Bed.png')

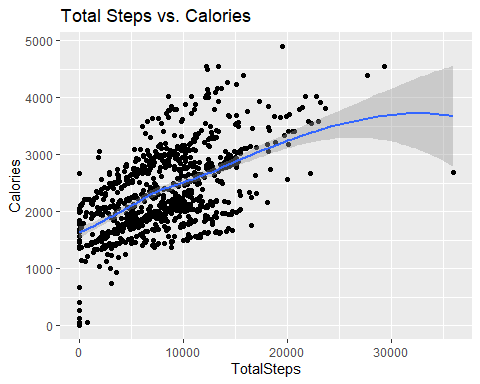
## Saving 5 x 4 in image  
## `geom\_smooth()` using method = 'loess' and formula = 'y ~ x'

The chart above shows a positive linear relationship between Total Minutes Asleep and Total Time in Bed, i.e the higher the Total Minutes Asleep, the higher the Total Time in Bed.

### Relationship between Total Steps and Calories

ggplot(data=Activity) + geom\_point(mapping = aes(x=TotalSteps, y=Calories)) + geom\_smooth(mapping = aes(x=TotalSteps, y=Calories)) + labs(title="Total Steps vs. Calories")

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



ggsave('Total\_Steps\_vs\_Calories.png')

## Saving 5 x 4 in image  
## `geom\_smooth()` using method = 'loess' and formula = 'y ~ x'

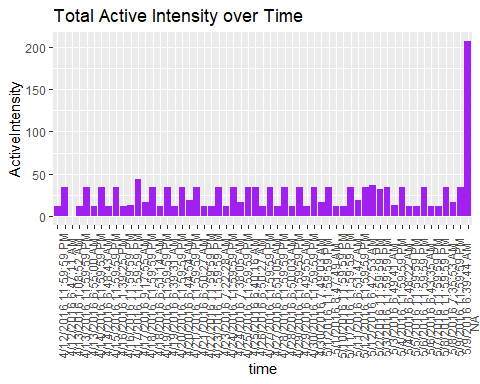
The chart above shows a positive relationship between Total Step and Calories, i.e the more steps taken, the more Calories are burnt.

### Relationship showing Active Intensity over Time

To visualize this, I will have to merge Intensity dataset with Weight dataset as a new dataframe “Intensity\_over\_time”using OUTER JOIN so as to access the time in Weight dataset

Intensity$ActiveIntensity <- (Intensity$VeryActiveMinutes)/60  
  
Intensity\_over\_time <- merge(Weight, Intensity, by="Id", all=TRUE)  
Intensity\_over\_time$time <- format(Intensity\_over\_time$Date, format = "%H:%M:%S")  
  
ggplot(data=Intensity\_over\_time, aes(x=time, y=ActiveIntensity)) + geom\_histogram(stat = "identity", fill='purple') +  
 theme(axis.text.x = element\_text(angle = 90)) +  
 labs(title="Total Active Intensity over Time ")

## Warning in geom\_histogram(stat = "identity", fill = "purple"): Ignoring unknown  
## parameters: `binwidth`, `bins`, and `pad`



ggsave('Intensity\_over\_time.png')

## Saving 5 x 4 in image

The chart above shows that most customers are more active during morning hours at around 6:30am - 8am and during evening hours at around 11:40pm -11:59pm

## **ACT PHASE**

In this phase, I will state my top insights based on my analysis and visualization

### Target Audience

The target audience of this analysis business tasks are women, who are busy with their daily activities and need to strike a balance between their work life and health using BellaBeat products.

### Recommendation to Marketing Team

1. **BellaBeat App should be redesigned to have an alarm/notification function which should be able to notify users periodically during the day to take a walk, thereby encouraging activeness and burning of calories**

Recommended steps according to Howdy Health,steps 7,000-10,000steps per day for adults aged 18-59years. 6,000-8,000steps per day for adults aged 60+years. 7,500steps per day for adults aged 62-101years.These steps could be achieved in bits during the day, say 4 notification per day, each notification aiming to acheive 2,000steps

* **BellaBeat App should be redesigned to have an alarm/notification function which should be able to notify users when to sleep, calculate the recommended sleep hours and also notify Users when to wake up so as to encourage a healthy sleeping habit.**

According to the National Sleep Foundation’s recommendations, adult aged 18 to 64years require 7-9hrs of sleep a night, while adults older than 64years require 7-8hrs of sleep.

1. **BellaBeat App should be redesigned to calculate and give feedback to User on their progress.**

These feedback should be weekly, which will encourage User to do better in the new week. These feedback should be in percentages. Total Recommended Steps per week should be 52,500steps so 26,250steps will be a 50% progress. Total Recommended Hours of sleep per week should be 56hrs so 45hrs should be 80% progress.Total Recommended Calories that should be burned by women per week 12,425 (as recommended by WebMD Fitness and Exercise is 1600-1950 per day, averagely 1775 ) so 8,000 should be 64%

1. **BellaBeat App should be redesigned to have an information section.**

This information section should be able to inform users on the benefits of healthy lifestyle, which includes taking walks, sleeping adequately, eating healthy which we help reduce calories in take.

1. **Intensity-over-time chart should be created so as to monitors users activity and constantly come up with strategy that will help keep BellaBeat in trend, unique and up-to-date.**

### References

This analysis was done with help of sample r codes and guidelines from the following site

* Coursera notes
* Case-study guidelines
* Case-study roadmap
* <https://rmarkdown.rstudio.com>
* <https://www.geeksforgeeks.org>
* <https://cran.r-project.org>