Google Data Analytics Capstone : How Can a Wellness Technology Company Play It Smart?

soon kien yuan

2022-08-18

Scenario

I'm a junior data analyst at Bellabeat, a high-tech maker of women's health goods. Bellabeat is a modest, successful startup with potential to grow in the global smart device industry. Bellabeat founders and CCO Urka Sren believes examining smart device fitness data might help the company develop. I have been requested to study smart device data for one of Bellabeat's products to learn how consumers use smart devices. The insights will shape the company's marketing approach. I will offer the research and recommendations to Bellabeat's management team.

Introduction

About the company

Bellabeat was formed by Urka Sren and Sando Mur. Sren used her artistic experience to make technology that encourages women worldwide. Bellabeat empowers women by collecting data on movement, sleep, stress, and reproductive health. Bellabeat was started in 2013 and has swiftly become a tech-driven women's wellness firm. Bellabeat had various offices and products by 2016. Bellabeat items are sold by a growing number of online shops in addition to their website. The corporation invests in radio, billboards, print, and TV, but focuses on internet marketing.

Products

1. **Bellabeat app**

   The Bellabeat app provides users with health data related to their activity, sleep, stress, menstrual cycle, and mindfulness habits. This data can help users better understand their current habits and make healthy decisions. The Bellabeat app connects to their line of smart wellness products.

2. **Leaf**

   Bellabeat's classic wellness tracker can be worn as a bracelet, necklace, or clip. The Leaf tracker connects to the Bellabeat app to track activity, sleep, and stress.

3. **Time**

   This wellness watch combines the timeless look of a classic timepiece with smart technology to track user activity, sleep, and stress. The Time watch connects to the Bellabeat app to provide you with insights into your daily wellness.

4. **Spring**

   This is a water bottle that tracks daily water intake using smart technology to ensure that you are appropriately hydrated throughout the day. The Spring bottle connects to the Bellabeat app to track your hydration levels.

5. **Bellabeat membership**

   Bellabeat also offers a subscription-based membership program for users. Membership gives users 24/7 access to fully personalized guidance on nutrition, activity, sleep, health and beauty, and mindfulness based on their lifestyle and goals.

Data Analysis Process

1.0 PHASE 1 :ASK

First up, the analysts needed to define what the project would look like and what would qualify as a successful result. So, to determine these things, they **asked** effective questions and collaborated with leaders and managers who were interested in the outcome of their people analysis.

- Ask effective questions
- Define the problem
- Use structured thinking
- Communicate with others
- What is the problem you are trying to solve?

1.1 Identify the business task

The business task will be associated with a non-Bellabeat smart devices and analyze the smart device usage data in order to gain insight into how people are already using their smart devices.Then, using the data, generate recommendations for the Bellabeat marketing strategy team to understand the trends. These recommendations and trend insights will be used to enhance the features, functionality, and service quality of the Bellabeat app. So, the business task can be summed up as follows:

1. Analyze the non-Bellabeat smart device usage data in order to gain insight to help Bellabeat marketing strategy team to understand the trends.

2. Utilizing the trends and insights provided by smart device usage data in order to improve the features, functionality, and overall service quality of the Bellabeat app

1.2 Consider key stakeholders

Urška Sršen: Bellabeat's cofounder and Chief Creative Officer

Sando Mur: Mathematician and Bellabeat's cofounder; key member of the Bellabeat executive team

Bellabeat marketing analytics team: A team of data analysts responsible for collecting, analyzing, and reporting data that helps guide Bellabeat's marketing strategy. You joined this team six months ago and have been busy learning about Bellabeat' 's mission and business goals — as well as how you, as a junior data analyst, can help Bellabeat achieve them.

2.0 PHASE 2 : PREPARE

- Understand how data is generated and collected
- Identify and use different data formats, types, and structures
- Make sure data is unbiased and credible
- Organize and protect data

2.1 Dataset overview

This case study utilises FitBit Fitness Tracker Data that is publicly accessible on **Kaggle.** The dataset was collected based on the responses of thirty individuals who participated in a distributed survey conducted by **Amazon Mechanical Turk** between **December 3 and December 5, 2016**.

The **thirty people** who participated in the survey are qualified **Fitbit users** who gave their approval to the submission of personal tracker data. This data included **minute-by-minute output for tracking of physical activity, heart rate, and sleep.**

FitBit Fitness Tracker Data consists of 18 csv file in long format. The data should then be converted to wide format in order to reduce data dimensionality for data analysis.

2.2 Data credibility

ROCC will be applied to detect if there are data credibility issues.

1. **Reliabile**

   It is **less reliable**. The dataset is comprised of 30 Fitbit users, which is insufficient to represent the eight million active users in 2016. This would result in a margin of error of 23.56% with a confidence level of 95%, which is less reliable based on margin error calculator .

   Central Limit Theorem (CLT) says that 30 is the smallest number of samples that can be used. The dataset is still valid based on CLT.

   However, it depends on the stakes. Larger samples are needed for reliable results.

2. **Original:**

   It is **not original**. The data set was produced by Amazon Mechanical Turk responders to a distributed survey, which is considered as third party data. It would have been preferable if FitBit had delivered the data directly.

3. **Comprehensive**:

   It does not cover every comprehensive aspect. The data are insufficient because they are lacking certain information (e.g., sex, age, genetic disease) that would assist in producing a more precise analysis. As a result, they cannot be considered comprehensive.

   And yet again, the information was gathered over a span of two months, which is insufficient. It is preferable to have data covering a period of at least a year.

   Last, How did they chose thirty people at random? Which strategy they implement? Does it come from a sample that was picked at random and from a place that was chosen at random as well?

4. **Current**

   It is **not current.**The data was collected between December 3 and December 5 of 2016, which is now a total of six years ago.

5. **Cited:**

   It is **Cited**. The datasets were generated by respondents to a distributed survey via Amazon Mechanical Turk.

2.3 How Data organized

2.3.1 Loading Packages

```
dailyActivity_merged <- read.csv("https://raw.githubusercontent.com/soonkienyuan/DataAnalytics_Capstone_

dailyCalories_merged  <- read.csv("https://raw.githubusercontent.com/soonkienyuan/DataAnalytics_Capstone

dailyIntensities_merged <- read.csv("https://raw.githubusercontent.com/soonkienyuan/DataAnalytics_Capst

dailySteps_merged <- read.csv("https://raw.githubusercontent.com/soonkienyuan/DataAnalytics_Capstone_ca

sleepDay_merged <- read.csv("https://raw.githubusercontent.com/soonkienyuan/DataAnalytics_Capstone_case_

weightLogInfo_merged <- read.csv("https://raw.githubusercontent.com/soonkienyuan/DataAnalytics_Capstone_
```

2.3.2 Importing Data Sets

2.4 Dataset structure

```
# first view of the data
head(dailyActivity_merged)
```

2.4.1 `dailyActivity_merged`

```
##           Id ActivityDate TotalSteps TotalDistance TrackerDistance
## 1 1503960366    4/12/2016      13162          8.50            8.50
## 2 1503960366    4/13/2016      10735          6.97            6.97
## 3 1503960366    4/14/2016      10460          6.74            6.74
## 4 1503960366    4/15/2016       9762          6.28            6.28
## 5 1503960366    4/16/2016      12669          8.16            8.16
## 6 1503960366    4/17/2016       9705          6.48            6.48
##   LoggedActivitiesDistance VeryActiveDistance ModeratelyActiveDistance
## 1                        0               1.88                     0.55
## 2                        0               1.57                     0.69
## 3                        0               2.44                     0.40
## 4                        0               2.14                     1.26
## 5                        0               2.71                     0.41
## 6                        0               3.19                     0.78
##   LightActiveDistance SedentaryActiveDistance VeryActiveMinutes
## 1                6.06                       0                25
## 2                4.71                       0                21
## 3                3.91                       0                30
## 4                2.83                       0                29
## 5                5.04                       0                36
## 6                2.51                       0                38
##   FairlyActiveMinutes LightlyActiveMinutes SedentaryMinutes Calories
```

```
## 1                     13                   328              728      1985
## 2                     19                   217              776      1797
## 3                     11                   181             1218      1776
## 4                     34                   209              726      1745
## 5                     10                   221              773      1863
## 6                     20                   164              539      1728
```

```
#structure of dataset
str(dailyActivity_merged)
```

```
## 'data.frame':    940 obs. of  15 variables:
##  $ Id                      : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ ActivityDate            : chr  "4/12/2016" "4/13/2016" "4/14/2016" "4/15/2016" ...
##  $ TotalSteps              : int  13162 10735 10460 9762 12669 9705 13019 15506 10544 9819 ...
##  $ TotalDistance           : num  8.5 6.97 6.74 6.28 8.16 ...
##  $ TrackerDistance         : num  8.5 6.97 6.74 6.28 8.16 ...
##  $ LoggedActivitiesDistance: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ VeryActiveDistance      : num  1.88 1.57 2.44 2.14 2.71 ...
##  $ ModeratelyActiveDistance: num  0.55 0.69 0.4 1.26 0.41 ...
##  $ LightActiveDistance     : num  6.06 4.71 3.91 2.83 5.04 ...
##  $ SedentaryActiveDistance : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ VeryActiveMinutes       : int  25 21 30 29 36 38 42 50 28 19 ...
##  $ FairlyActiveMinutes     : int  13 19 11 34 10 20 16 31 12 8 ...
##  $ LightlyActiveMinutes    : int  328 217 181 209 221 164 233 264 205 211 ...
##  $ SedentaryMinutes        : int  728 776 1218 726 773 539 1149 775 818 838 ...
##  $ Calories                : int  1985 1797 1776 1745 1863 1728 1921 2035 1786 1775 ...
```

```
#skimming

skim_without_charts(dailyActivity_merged)
```

Table 1: Data summary

Name	dailyActivity_merged
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	1.503960e+09	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

```
#how many participant records exist in the datasets?

as.data.frame(table(dailyActivity_merged$Id)) %>% rename(Id =Var1 )
```

```
##            Id Freq
## 1  1503960366   31
## 2  1624580081   31
## 3  1644430081   30
## 4  1844505072   31
## 5  1927972279   31
## 6  2022484408   31
## 7  2026352035   31
## 8  2320127002   31
## 9  2347167796   18
## 10 2873212765   31
## 11 3372868164   20
## 12 3977333714   30
## 13 4020332650   31
## 14 4057192912    4
## 15 4319703577   31
## 16 4388161847   31
## 17 4445114986   31
## 18 4558609924   31
## 19 4702921684   31
## 20 5553957443   31
## 21 5577150313   30
## 22 6117666160   28
## 23 6290855005   29
## 24 6775888955   26
## 25 6962181067   31
## 26 7007744171   26
## 27 7086361926   31
```

```
## 28 8053475328   31
## 29 8253242879   19
## 30 8378563200   31
## 31 8583815059   31
## 32 8792009665   29
## 33 8877689391   31
```

```
#How many partcipants in the datasets

length(table(dailyActivity_merged$Id))
```

```
## [1] 33
```

```
#or
n_distinct(dailyActivity_merged$Id)
```

```
## [1] 33
```

```


# first view of the data
head(dailyCalories_merged)
```

2.4.2 dailyCalories_merged

```
##           Id ActivityDay Calories
## 1 1503960366   4/12/2016     1985
## 2 1503960366   4/13/2016     1797
## 3 1503960366   4/14/2016     1776
## 4 1503960366   4/15/2016     1745
## 5 1503960366   4/16/2016     1863
## 6 1503960366   4/17/2016     1728
```

```
#structure of dataset
str(dailyCalories_merged)
```

```
## 'data.frame':    940 obs. of  3 variables:
##  $ Id         : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ ActivityDay: chr  "4/12/2016" "4/13/2016" "4/14/2016" "4/15/2016" ...
##  $ Calories   : int  1985 1797 1776 1745 1863 1728 1921 2035 1786 1775 ...
```

```
#skimming

skim_without_charts(dailyCalories_merged)
```

Table 4: Data summary

Name	dailyCalories_merged
Number of rows	940

Table 4: Data summary

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	4445114986	6.962181e+09	8877689391
Calories	0	1	2.303610e+03	7.181700e+02	0	1828.5	2134	2.793250e+03	4900

```
#how many participant records exist in the datasets?

as.data.frame(table(dailyCalories_merged$Id)) %>% rename(Id =Var1 )
```

```
##            Id Freq
## 1  1503960366   31
## 2  1624580081   31
## 3  1644430081   30
## 4  1844505072   31
## 5  1927972279   31
## 6  2022484408   31
## 7  2026352035   31
## 8  2320127002   31
## 9  2347167796   18
## 10 2873212765   31
## 11 3372868164   20
## 12 3977333714   30
## 13 4020332650   31
## 14 4057192912    4
## 15 4319703577   31
## 16 4388161847   31
## 17 4445114986   31
## 18 4558609924   31
## 19 4702921684   31
## 20 5553957443   31
## 21 5577150313   30
## 22 6117666160   28
## 23 6290855005   29
## 24 6775888955   26
```

```
## 25 6962181067   31
## 26 7007744171   26
## 27 7086361926   31
## 28 8053475328   31
## 29 8253242879   19
## 30 8378563200   31
## 31 8583815059   31
## 32 8792009665   29
## 33 8877689391   31
```

```
#How many partcipants in the datasets

length(table(dailyCalories_merged$Id))
```

```
## [1] 33
```

```
#or
n_distinct(dailyCalories_merged$Id)
```

```
## [1] 33
```

```


# first view of the data
head(dailyIntensities_merged)
```

2.4.3 dailyIntensities_merged

```
##           Id ActivityDay SedentaryMinutes LightlyActiveMinutes
## 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
##   FairlyActiveMinutes VeryActiveMinutes SedentaryActiveDistance
## 1                  13                25                       0
## 2                  19                21                       0
## 3                  11                30                       0
## 4                  34                29                       0
## 5                  10                36                       0
## 6                  20                38                       0
##   LightActiveDistance ModeratelyActiveDistance VeryActiveDistance
## 1                6.06                     0.55               1.88
## 2                4.71                     0.69               1.57
## 3                3.91                     0.40               2.44
## 4                2.83                     1.26               2.14
## 5                5.04                     0.41               2.71
## 6                2.51                     0.78               3.19
```

```r
#structure of dataset
str(dailyIntensities_merged)
```

```
## 'data.frame':    940 obs. of  10 variables:
##  $ Id                      : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ ActivityDay             : chr  "4/12/2016" "4/13/2016" "4/14/2016" "4/15/2016" ...
##  $ SedentaryMinutes        : int  728 776 1218 726 773 539 1149 775 818 838 ...
##  $ LightlyActiveMinutes    : int  328 217 181 209 221 164 233 264 205 211 ...
##  $ FairlyActiveMinutes     : int  13 19 11 34 10 20 16 31 12 8 ...
##  $ VeryActiveMinutes       : int  25 21 30 29 36 38 42 50 28 19 ...
##  $ SedentaryActiveDistance : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LightActiveDistance     : num  6.06 4.71 3.91 2.83 5.04 ...
##  $ ModeratelyActiveDistance: num  0.55 0.69 0.4 1.26 0.41 ...
##  $ VeryActiveDistance      : num  1.88 1.57 2.44 2.14 2.71 ...
```

```r
#skimming

skim_without_charts(dailyIntensities_merged)
```

Table 7: Data summary

Name	dailyIntensities_merged
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

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
VeryActiveDistance	0	1	1.500000e+00	2.660000e+00	0	0.000000e+00	2.100000e-01	2.050000e+00	2.192000e+01

```
#how many participant records exist in the datasets?

as.data.frame(table(dailyIntensities_merged$Id)) %>% rename(Id =Var1 )
```

```
##             Id Freq
## 1  1503960366   31
## 2  1624580081   31
## 3  1644430081   30
## 4  1844505072   31
## 5  1927972279   31
## 6  2022484408   31
## 7  2026352035   31
## 8  2320127002   31
## 9  2347167796   18
## 10 2873212765   31
## 11 3372868164   20
## 12 3977333714   30
## 13 4020332650   31
## 14 4057192912    4
## 15 4319703577   31
## 16 4388161847   31
## 17 4445114986   31
## 18 4558609924   31
## 19 4702921684   31
## 20 5553957443   31
## 21 5577150313   30
## 22 6117666160   28
## 23 6290855005   29
## 24 6775888955   26
## 25 6962181067   31
## 26 7007744171   26
## 27 7086361926   31
## 28 8053475328   31
## 29 8253242879   19
## 30 8378563200   31
## 31 8583815059   31
## 32 8792009665   29
## 33 8877689391   31
```

```
#How many partcipants in the datasets

length(table(dailyIntensities_merged$Id))
```

```
## [1] 33
```

```
#or
n_distinct(dailyIntensities_merged$Id)
```

```
## [1] 33
```

```r
# first view of the data
head(dailySteps_merged)
```

2.4.4 dailySteps_merged

```
##           Id ActivityDay StepTotal
## 1 1503960366   4/12/2016     13162
## 2 1503960366   4/13/2016     10735
## 3 1503960366   4/14/2016     10460
## 4 1503960366   4/15/2016      9762
## 5 1503960366   4/16/2016     12669
## 6 1503960366   4/17/2016      9705
```

```r
#structure of dataset
str(dailySteps_merged)
```

```
## 'data.frame':    940 obs. of  3 variables:
##  $ Id         : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ ActivityDay: chr  "4/12/2016" "4/13/2016" "4/14/2016" "4/15/2016" ...
##  $ StepTotal  : int  13162 10735 10460 9762 12669 9705 13019 15506 10544 9819 ...
```

```r
#skimming

skim_without_charts(dailySteps_merged)
```

Table 10: Data summary

Name	dailySteps_merged
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	2.320127e+09	4445114986	6962181067	8877689391

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
StepTotal	0	1	7.637910e+03	5.087150e+03	0	3.789750e+03	7405.5	10727	36019

```
#how many participant records exist in the datasets?

as.data.frame(table(dailySteps_merged$Id)) %>% rename(Id =Var1 )
```

```
##            Id Freq
## 1  1503960366   31
## 2  1624580081   31
## 3  1644430081   30
## 4  1844505072   31
## 5  1927972279   31
## 6  2022484408   31
## 7  2026352035   31
## 8  2320127002   31
## 9  2347167796   18
## 10 2873212765   31
## 11 3372868164   20
## 12 3977333714   30
## 13 4020332650   31
## 14 4057192912    4
## 15 4319703577   31
## 16 4388161847   31
## 17 4445114986   31
## 18 4558609924   31
## 19 4702921684   31
## 20 5553957443   31
## 21 5577150313   30
## 22 6117666160   28
## 23 6290855005   29
## 24 6775888955   26
## 25 6962181067   31
## 26 7007744171   26
## 27 7086361926   31
## 28 8053475328   31
## 29 8253242879   19
## 30 8378563200   31
## 31 8583815059   31
## 32 8792009665   29
## 33 8877689391   31
```

```
#How many partcipants in the datasets

length(table(dailySteps_merged$Id))
```

```
## [1] 33
```

```
#or
n_distinct(dailySteps_merged$Id)
```

```
## [1] 33
```

```r
# first view of the data
head(sleepDay_merged)
```

2.4.5 sleepDay_merged

```
##           Id              SleepDay TotalSleepRecords TotalMinutesAsleep
## 1 1503960366 4/12/2016 12:00:00 AM                 1                327
## 2 1503960366 4/13/2016 12:00:00 AM                 2                384
## 3 1503960366 4/15/2016 12:00:00 AM                 1                412
## 4 1503960366 4/16/2016 12:00:00 AM                 2                340
## 5 1503960366 4/17/2016 12:00:00 AM                 1                700
## 6 1503960366 4/19/2016 12:00:00 AM                 1                304
##   TotalTimeInBed
## 1            346
## 2            407
## 3            442
## 4            367
## 5            712
## 6            320
```

```r
#structure of dataset
str(sleepDay_merged)
```

```
## 'data.frame':    413 obs. of  5 variables:
##  $ Id                : num  1.5e+09 1.5e+09 1.5e+09 1.5e+09 1.5e+09 ...
##  $ SleepDay          : chr  "4/12/2016 12:00:00 AM" "4/13/2016 12:00:00 AM" "4/15/2016 12:00:00 AM"
##  $ TotalSleepRecords : int  1 2 1 2 1 1 1 1 1 1 ...
##  $ TotalMinutesAsleep: int  327 384 412 340 700 304 360 325 361 430 ...
##  $ TotalTimeInBed    : int  346 407 442 367 712 320 377 364 384 449 ...
```

```r
#skimming

skim_without_charts(sleepDay_merged)
```

Table 13: Data summary

Name	sleepDay_merged
Number of rows	413
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	20	21	0	31	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
Id	0	1	5.000979e+09	2.06036e+09	1503960366	3977333714	4702921684	6962181067	8792009665
TotalSleepRecords	0	1	1.120000e+00	3.50000e-01	1	1	1	1	3
TotalMinutesAsleep	0	1	4.194700e+02	1.18340e+02	58	361	433	490	796
TotalTimeInBed	0	1	4.586400e+02	1.27100e+02	61	403	463	526	961

```r
#how many participant records exist in the datasets?

as.data.frame(table(sleepDay_merged$Id)) %>% rename(Id =Var1 )
```

```
##            Id Freq
## 1  1503960366   25
## 2  1644430081    4
## 3  1844505072    3
## 4  1927972279    5
## 5  2026352035   28
## 6  2320127002    1
## 7  2347167796   15
## 8  3977333714   28
## 9  4020332650    8
## 10 4319703577   26
## 11 4388161847   24
## 12 4445114986   28
## 13 4558609924    5
## 14 4702921684   28
## 15 5553957443   31
## 16 5577150313   26
## 17 6117666160   18
## 18 6775888955    3
## 19 6962181067   31
## 20 7007744171    2
## 21 7086361926   24
## 22 8053475328    3
## 23 8378563200   32
## 24 8792009665   15
```

```r
#How many partcipants in the datasets

length(table(sleepDay_merged$Id))
```

```
## [1] 24
```

```r
#or
n_distinct(sleepDay_merged$Id)
```

```
## [1] 24
```

```r
# first view of the data
head(weightLogInfo_merged)
```

2.4.6 weightLogInfo_merged

```
##           Id                  Date WeightKg WeightPounds Fat   BMI
## 1 1503960366  5/2/2016 11:59:59 PM     52.6     115.9631  22 22.65
## 2 1503960366  5/3/2016 11:59:59 PM     52.6     115.9631  NA 22.65
## 3 1927972279  4/13/2016 1:08:52 AM    133.5     294.3171  NA 47.54
## 4 2873212765 4/21/2016 11:59:59 PM     56.7     125.0021  NA 21.45
## 5 2873212765 5/12/2016 11:59:59 PM     57.3     126.3249  NA 21.69
## 6 4319703577 4/17/2016 11:59:59 PM     72.4     159.6147  25 27.45
##   IsManualReport        LogId
## 1           True 1.462234e+12
## 2           True 1.462320e+12
## 3          False 1.460510e+12
## 4           True 1.461283e+12
## 5           True 1.463098e+12
## 6           True 1.460938e+12
```

```r
#structure of dataset
str(weightLogInfo_merged)
```

```
## 'data.frame':    67 obs. of  8 variables:
##  $ Id            : num  1.50e+09 1.50e+09 1.93e+09 2.87e+09 2.87e+09 ...
##  $ Date          : chr  "5/2/2016 11:59:59 PM" "5/3/2016 11:59:59 PM" "4/13/2016 1:08:52 AM" "4/21/20
##  $ WeightKg      : num  52.6 52.6 133.5 56.7 57.3 ...
##  $ WeightPounds  : num  116 116 294 125 126 ...
##  $ Fat           : int  22 NA NA NA NA 25 NA NA NA NA ...
##  $ BMI           : num  22.6 22.6 47.5 21.5 21.7 ...
##  $ IsManualReport: chr  "True" "True" "False" "True" ...
##  $ LogId         : num  1.46e+12 1.46e+12 1.46e+12 1.46e+12 1.46e+12 ...
```

```r
#skimming

skim_without_charts(weightLogInfo_merged)
```

Table 16: Data summary

Name	weightLogInfo_merged
Number of rows	67
Number of columns	8

Table 16: Data summary

Column type frequency:	
character	2
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
IsManualReport	0	1	4	5	0	2	0

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

```
#how many participant records exist in the datasets?

as.data.frame(table(weightLogInfo_merged$Id)) %>% rename(Id =Var1 )
```

```
##           Id Freq
## 1 1503960366    2
## 2 1927972279    1
## 3 2873212765    2
## 4 4319703577    2
## 5 4558609924    5
## 6 5577150313    1
## 7 6962181067   30
## 8 8877689391   24
```

```
#How many partcipants in the datasets

length(table(weightLogInfo_merged$Id))
```

```
## [1] 8
```

```
#or
n_distinct(weightLogInfo_merged$Id)
```

```
## [1] 8
```

3.0 PHASE 3 : PROCESS

Noticed that he dailyActivity_merged datasets are just a composite of dailyCalories_merged dataset, dailyIntensities_merged dataset, and dailySteps_merged dataset.

Hence, dataset will used in this study would be `dailyActivity_merged`, `sleepDay_merged` and `weightLogInfo_merged`

3.1 Loading Packages

3.2 Checking for duplicte data

```
dailyActivity_merged %>% get_dupes()
```

```
## No variable names specified - using all columns.

## No duplicate combinations found of: Id, ActivityDate, TotalSteps, TotalDistance, TrackerDistance, Log

##  [1] Id                       ActivityDate             TotalSteps
##  [4] TotalDistance            TrackerDistance          LoggedActivitiesDistance
##  [7] VeryActiveDistance       ModeratelyActiveDistance LightActiveDistance
## [10] SedentaryActiveDistance  VeryActiveMinutes        FairlyActiveMinutes
## [13] LightlyActiveMinutes     SedentaryMinutes         Calories
## [16] dupe_count
## <0 rows> (or 0-length row.names)
```

```
sleepDay_merged %>% get_dupes()
```

```
## No variable names specified - using all columns.

##           Id              SleepDay TotalSleepRecords TotalMinutesAsleep
## 1 4388161847  5/5/2016 12:00:00 AM                 1                471
## 2 4388161847  5/5/2016 12:00:00 AM                 1                471
## 3 4702921684  5/7/2016 12:00:00 AM                 1                520
## 4 4702921684  5/7/2016 12:00:00 AM                 1                520
## 5 8378563200 4/25/2016 12:00:00 AM                 1                388
## 6 8378563200 4/25/2016 12:00:00 AM                 1                388
##   TotalTimeInBed dupe_count
## 1            495          2
## 2            495          2
## 3            543          2
## 4            543          2
## 5            402          2
## 6            402          2
```

```
weightLogInfo_merged %>% get_dupes()
```

```
## No variable names specified - using all columns.

## No duplicate combinations found of: Id, Date, WeightKg, WeightPounds, Fat, BMI, IsManualReport, LogI
```

```
## [1] Id             Date           WeightKg       WeightPounds   Fat
## [6] BMI            IsManualReport LogId          dupe_count
## <0 rows> (or 0-length row.names)
```

I discovered three duplicate data in `sleepDay_merged` by using the `getdupes()` function in janitor package.

By running `skim without charts` from the `skimr` package,found that `weightLogInfo_merged` fat variable contains 65 NA values.

Hence, I decided to drop fat column in `weightLogInfo_merged` and duplicate data in `sleepDay_merged`

3.3 Remove duplicates and NA

```
sleepDay_merged <- sleepDay_merged %>% distinct()
```

```
weightLogInfo_merged <- subset(weightLogInfo_merged, select = -c(Fat))
```

3.4 Tranforming data

The time and date are recorded together in the same column of both the sleep and weight databases, which I found interesting. If I do decide to use dates to analyse the data across the three datasets, it will be most helpful to divide them into `Date` and `Time` columns.

```
weightLogInfo_merged <- weightLogInfo_merged %>% separate(Date, into=c("Date", "Time"), sep=" ")
```

```
## Warning: Expected 2 pieces. Additional pieces discarded in 67 rows [1, 2, 3, 4,
## 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

```
sleepDay_merged <- sleepDay_merged %>% separate(SleepDay, into=c("Date", "Time"), sep=" ")
```

```
## Warning: Expected 2 pieces. Additional pieces discarded in 410 rows [1, 2, 3, 4,
## 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

4.0 PHASE 4 :Analyze

4.1 Loading Packages

4.2 Quick summary statistics

For the `dailyActivity_merged`

```
dailyActivity_merged %>%
  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
```

For the `sleepDay_merged`

```
sleepDay_merged %>%
  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
```

For the `weightLogInfo_merged`

```
weightLogInfo_merged %>%
  select(WeightKg,BMI ) %>%
  summary()
```

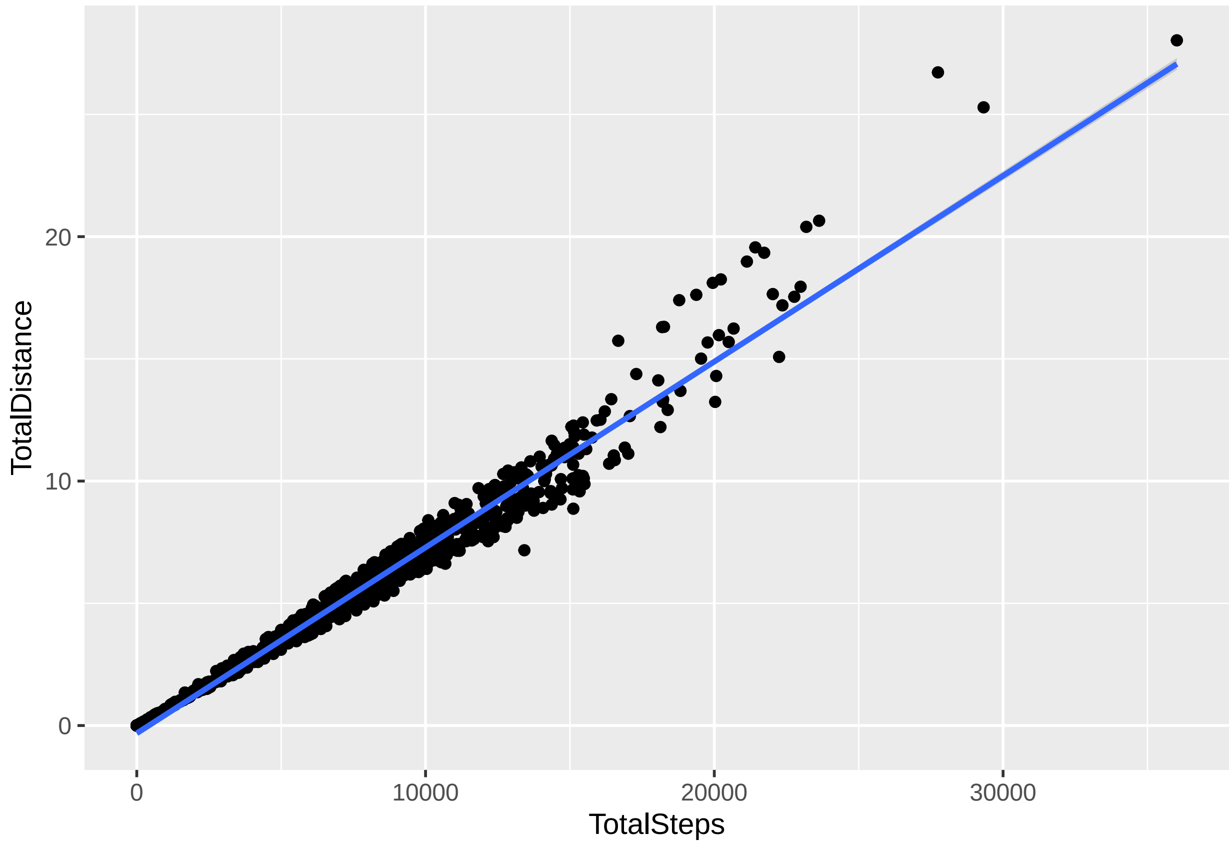
```
##     WeightKg           BMI
##  Min.   : 52.60   Min.   :21.45
##  1st Qu.: 61.40   1st Qu.:23.96
##  Median : 62.50   Median :24.39
##  Mean   : 72.04   Mean   :25.19
##  3rd Qu.: 85.05   3rd Qu.:25.56
##  Max.   :133.50   Max.   :47.54
```

4.3 Data Exploration

```
ggplot(data=dailyActivity_merged, aes(x=TotalSteps, y=TotalDistance)) + geom_point() +geom_smooth(metho
```

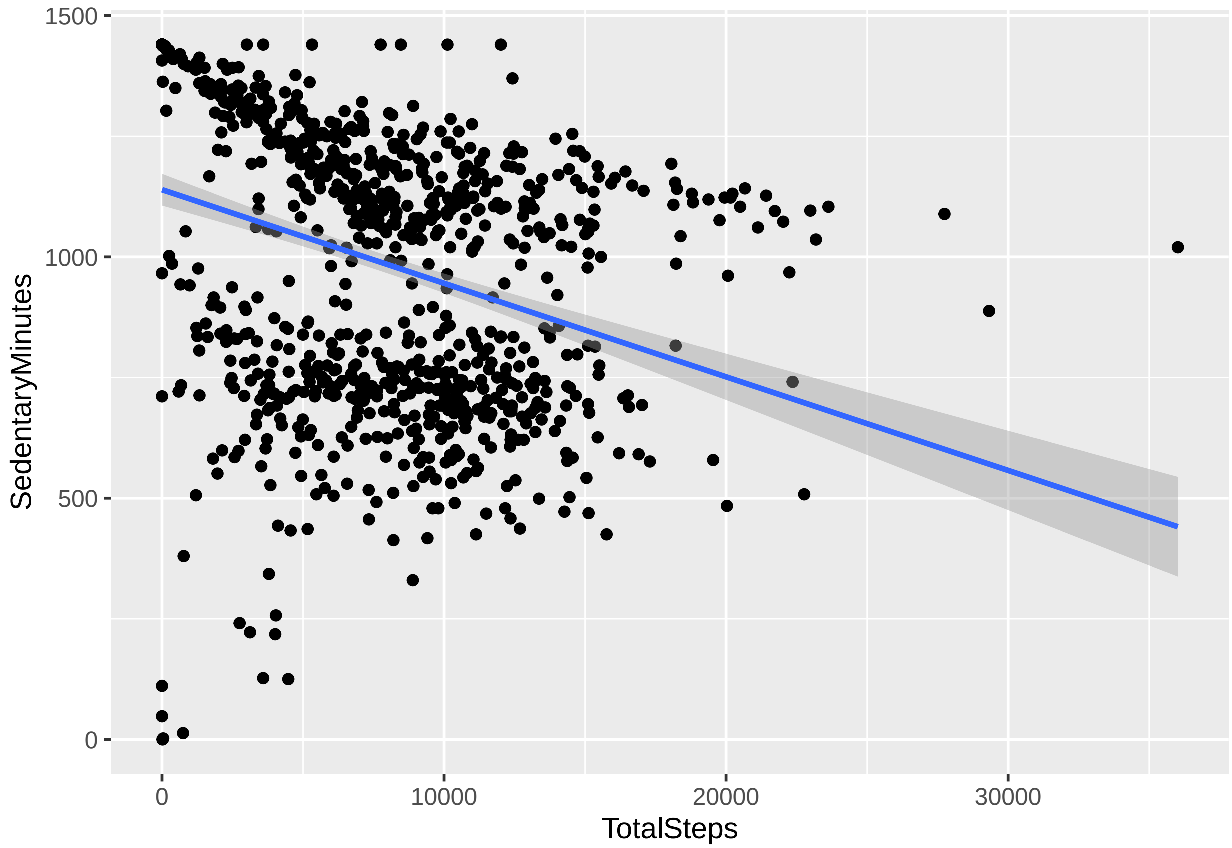
4.3.1 dailyActivity_merged

```
## `geom_smooth()` using formula 'y ~ x'
```



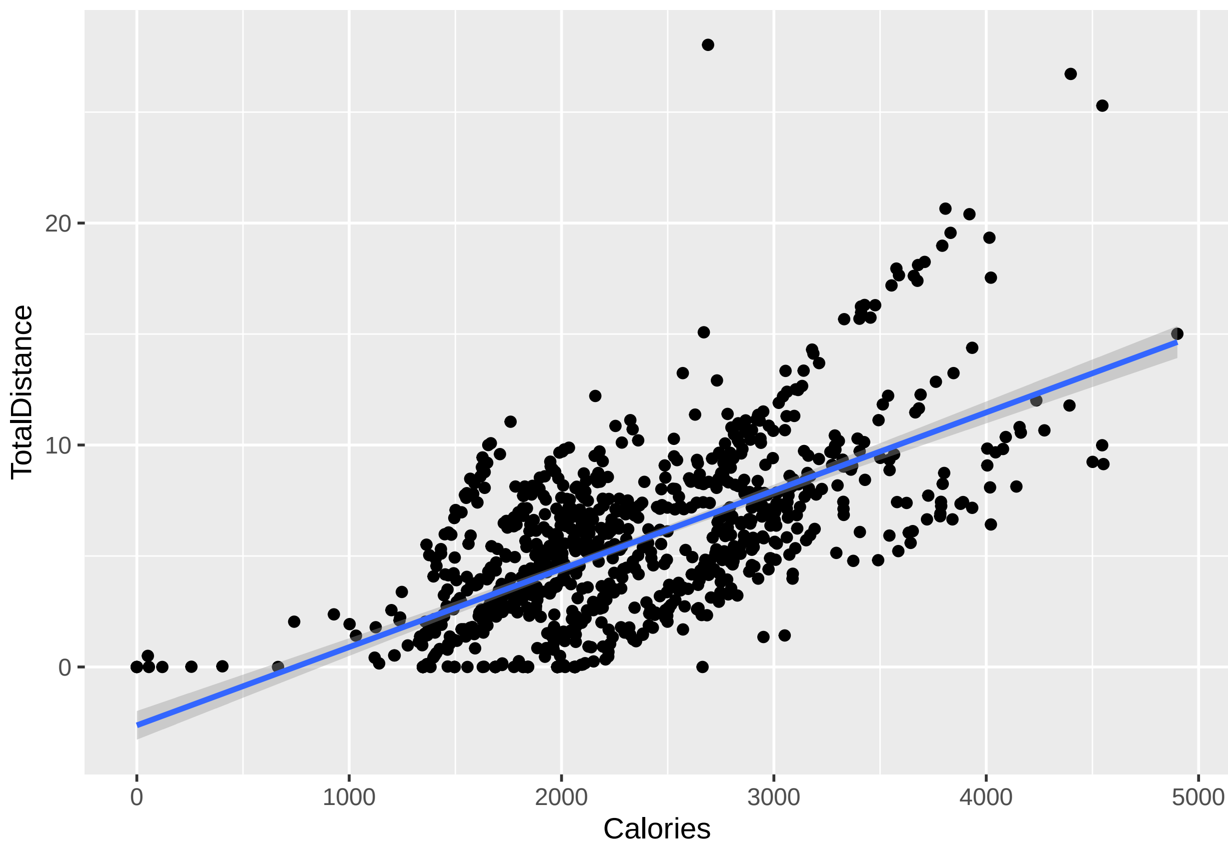

```
ggplot(data=dailyActivity_merged, aes(x=TotalSteps, y=SedentaryMinutes)) + geom_point() + geom_smooth(me
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
ggplot(data=dailyActivity_merged, aes(x=Calories, y=TotalDistance)) + geom_point() + geom_smooth(method=
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Trends for dailyActivity_merged

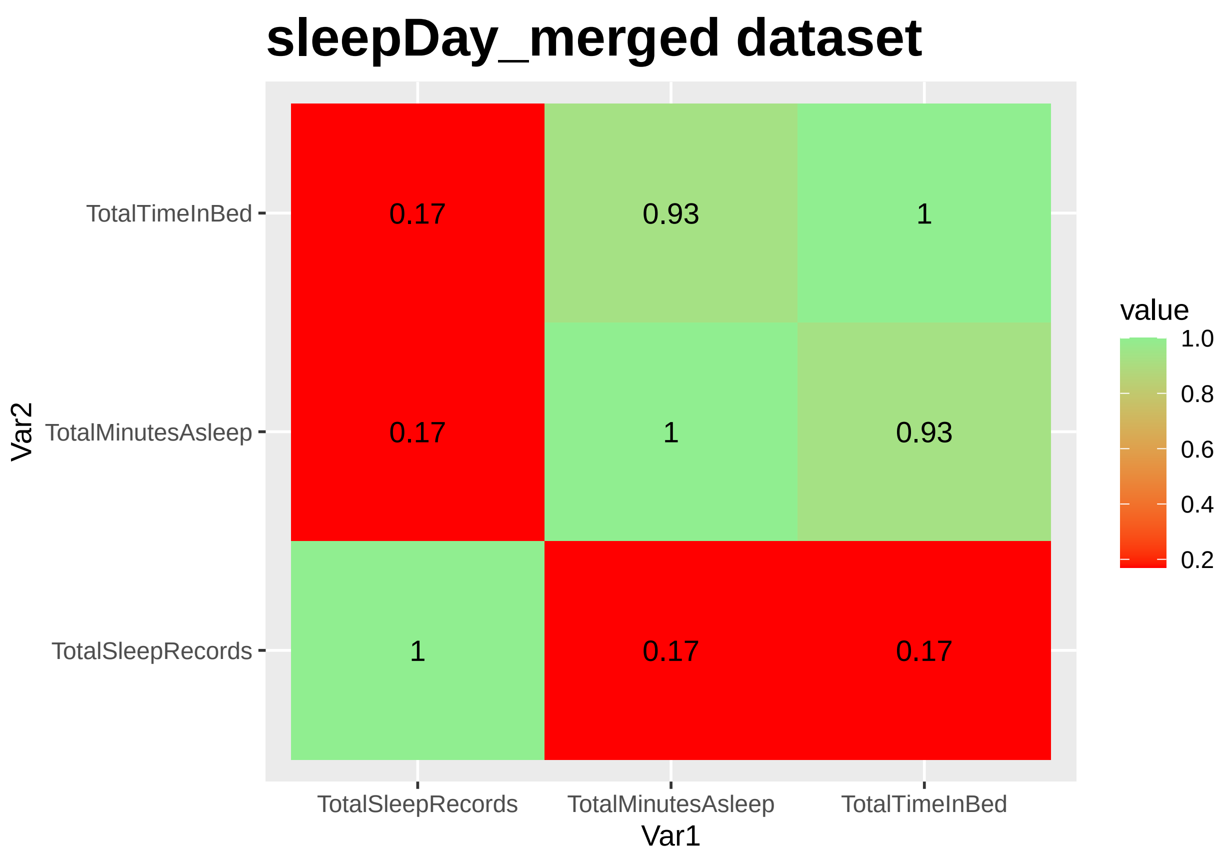
- The correlation coefficient between `TotalSteps` and `TotalDistance` is extremely high. This means that the more steps the user walks, the farther they travel.
- The scatter plot between `TotalSteps` and `SedentaryMinutes` allows for the estimation of the following conclusion: the longer the `SedentaryMinutes`, the more likely the `totalsteps` have decreased. This means the longer Sedentary time of user, the more likely the users walk less.
- In other words, the longer a user spends on Sedentary activities, the less probable it is that they would engage in physical activity.
- The scatter plot between `Calories` and `TotalDistance` allows for the estimation of the following conclusion: the higher the TotalDistance, the more likely the `Calories` burn have increase. The greater the `TotalDistance`, the greater the likelihood that the Calories burned will increase.
- In other words, the longer a user walks, the greater the chance that they will have a caloric deficit.

4.3.2 `sleepDay_merged` heatmap

```
# corelation matrix
cormat <- sleepDay_merged %>%
  select(TotalSleepRecords,
         TotalMinutesAsleep,
         TotalTimeInBed) %>%
  cor() %>% round(2)
```

```r
#reshape the data for heatmap
melted_cormat <- melt(cormat)

#data visualization
ggplot(data=melted_cormat,aes(x = Var1, y = Var2, fill = value))+
  geom_tile()+scale_fill_gradient(high = "green", low = "red")+
  ggtitle("sleepDay_merged dataset")+
  theme(plot.title = element_text(size = 20, face = "bold"))+
  geom_text(aes(label = round(value, 3)))+
  scale_fill_continuous(low = "red", high = "lightgreen")
```

```
## Scale for 'fill' is already present. Adding another scale for 'fill', which
## will replace the existing scale.
```



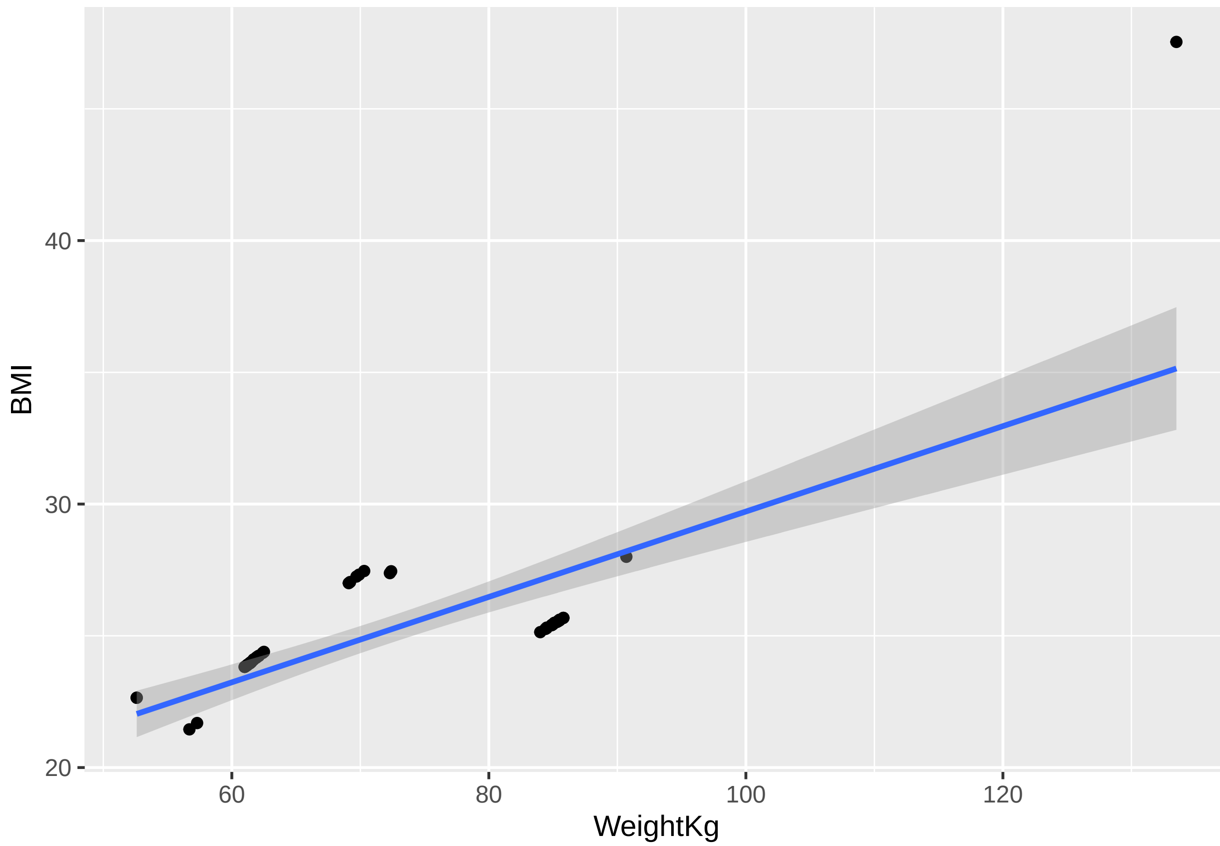
Trends for dailyActivity_merged

- The correlation coefficient between `TotalTimeInBed` and `TotalMinutesAsleep` is very high and positive based on the heat map.
- other words, when sleep duration increases, so does total bedtime.

```r
ggplot(data=weightLogInfo_merged, aes(x=WeightKg, y=BMI)) + geom_point() + geom_smooth(method=lm)
```

4.3.3 weightLogInfo_merged

```
## `geom_smooth()` using formula 'y ~ x'
```



Trends for dailyActivity_merged

- Nothing interesting found.
- BMI is the result measure that links body weight to height.

4.4 Data merging

I decide to merge the data between `dailyActivity_merged` and `sleepDay_merged`.

The `weightLogInfo_merged` will not be combined because there are only 8 users in this dataset, while the total number of users in all other datasets is 33.

In contrast to other datasets, which have 33 users contributing information, the `weightLogInfo_merged` datasets only have 8 users contributing data, which makes the information irrelevant and unsuitable for usage.

Hence, the weightLogInfo_merged

```r
#put all data frames into list
data_list <- list(dailyActivity_merged, sleepDay_merged)

#merge all data frames in list
combined_data<- data_list %>% reduce(full_join, by='Id')
```

```
ncol(combined_data)
```

```
## [1] 20
```

```
n_distinct(combined_data$Id)
```

```
## [1] 33
```

5.0 PRASE 5 : SHARE

5.1 Tableau Dashboard

```
write.csv(combined_data,"C:\\Users\\soonk\\Documents\\GitHub\\DataAnalytics_Capstone_case_study2\\datas
```

Tableau will be used to finish part of this PHASE. Refer to the Tableau file (share.twbx).

I created a dashbaord to make it easier to spot trends and obtain insights from the data.

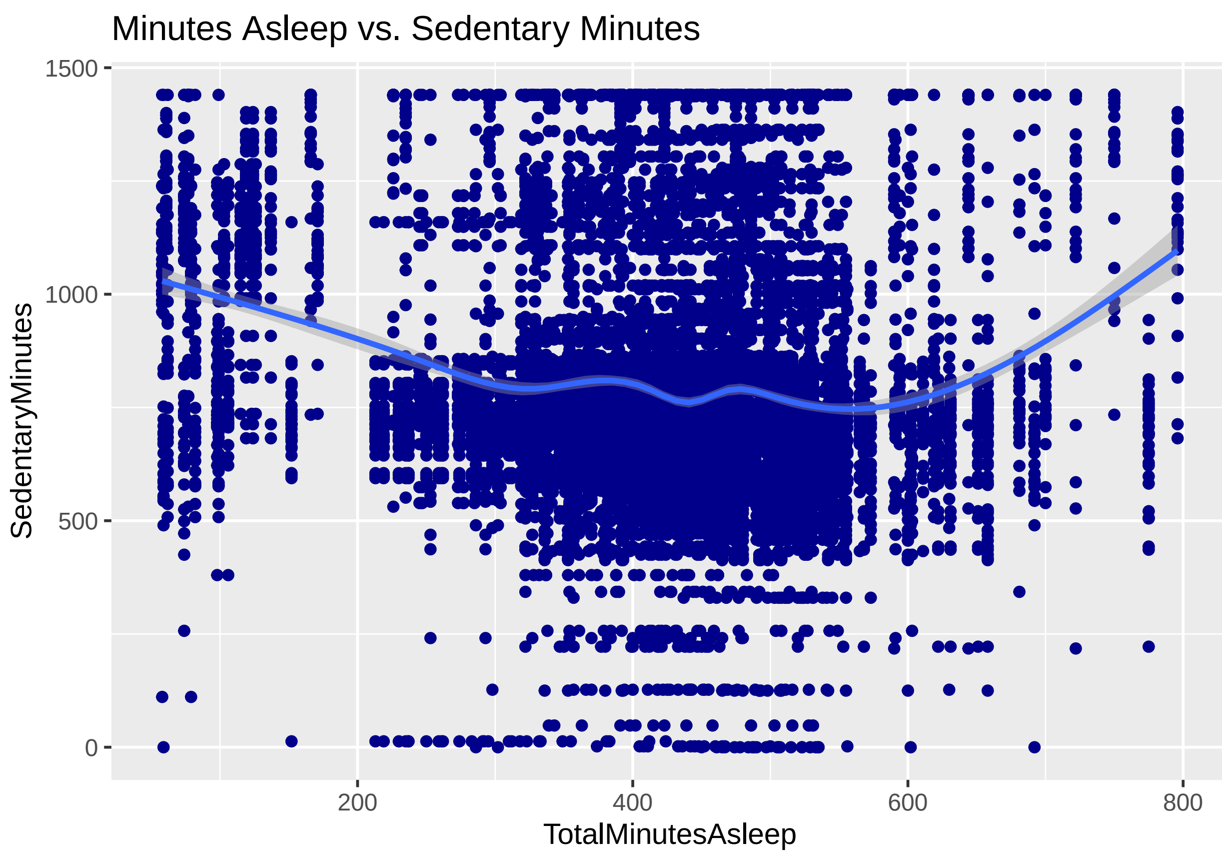
5.2 Minutes Asleep vs. Sedentary Minutes

```
ggplot(data=combined_data, aes(x=TotalMinutesAsleep, y=SedentaryMinutes)) +
geom_point(color='darkblue') + geom_smooth() +
  labs(title="Minutes Asleep vs. Sedentary Minutes")
```

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

```
## Warning: Removed 227 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 227 rows containing missing values (geom_point).
```



Interpretation

- Nothing special trends or findings found

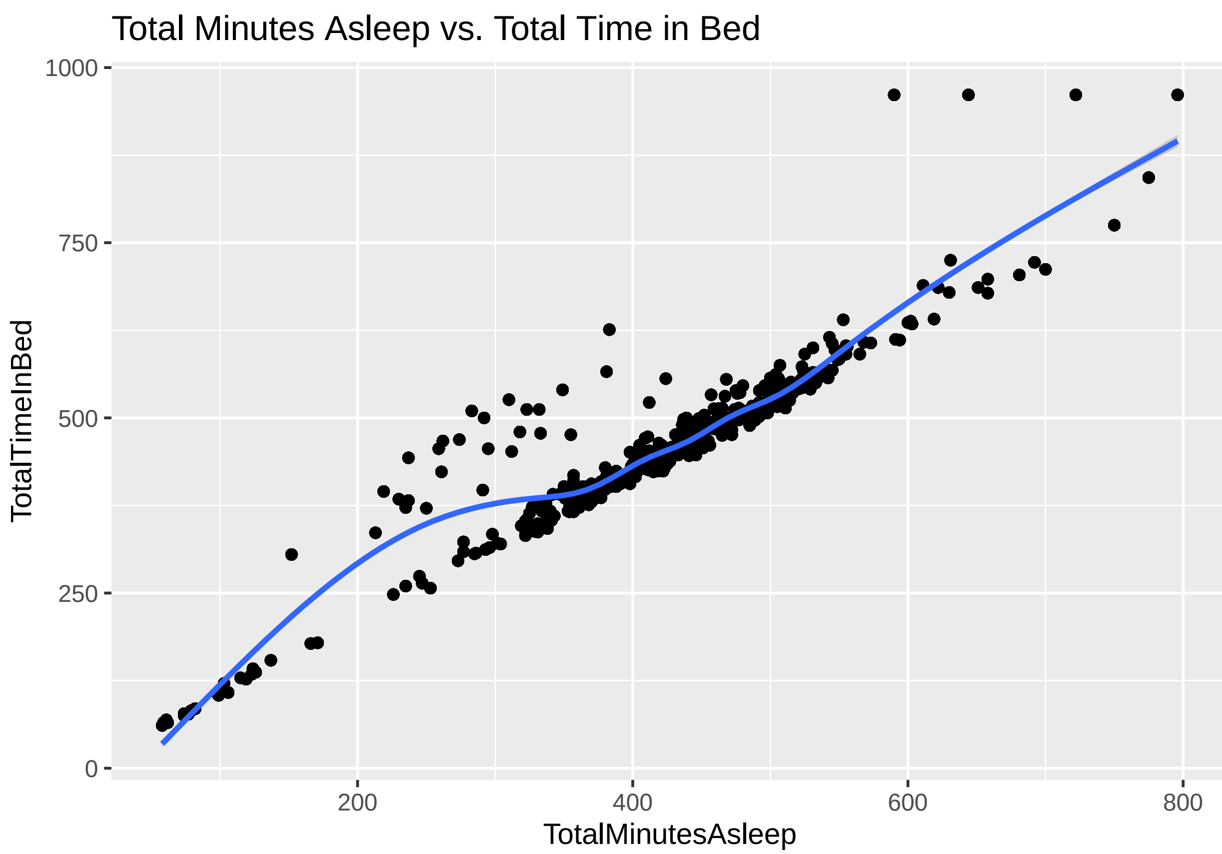
5.3 Total Minutes Asleep vs. Total Time in Bed

```
ggplot(data=combined_data, aes(x=TotalMinutesAsleep, y=TotalTimeInBed)) +
  geom_point()+ labs(title="Total Minutes Asleep vs. Total Time in Bed")+geom_smooth()
```

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

```
## Warning: Removed 227 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 227 rows containing missing values (geom_point).
```



Interpretation

- Total time in bed seems positively related to total time asleep. So, to help Bellabeat users get better sleep and increase overall customer experiance, the Bellabeat should think about implementing a notification system that prompts them to lie down and relax.

5.3 Total Steps vs. Calories

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

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

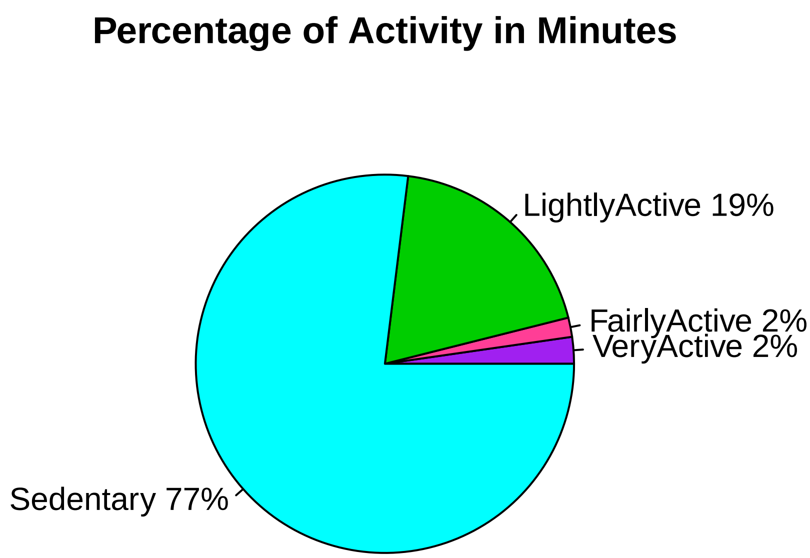
Interpretation

- The scatter plot between `Calories` and `TotalStep` allows for the estimation of the following conclusion: the higher the Total Step, the more likely the `Calories` burn have increase.
- In other words, the longer a user walks, the greater the chance that they will have a caloric deficit.
- Consequently, a reward system should be considered to motivate users to walk and exercise more. Consider a comprehensive reward system with targeted tasks. For example, if your total steps exceed 5,000, you will receive 500 points. The points earned can be redeemed for rewards such as Bellabeat merchandise or discounts on other Bellabeat products, allowing the company to expand and promote the product line while increasing revenue and attracting potential customers.

5.4

```
very_active_min <- sum(combined_data$VeryActiveMinutes)
fairly_active_min <- sum(combined_data$FairlyActiveMinutes)
lightly_activemin <- sum(combined_data$LightlyActiveMinutes)
sedentary_min <- sum(combined_data$SedentaryMinutes)
total_min <- very_active_min + fairly_active_min + lightly_activemin + sedentary_min


min_list <- c(very_active_min,fairly_active_min,lightly_activemin,sedentary_min)

lbls <- c("VeryActive","FairlyActive","LightlyActive","Sedentary")
```

```
pct <- round(min_list/total_min*100)

lbls <- paste(lbls, pct)
lbls <- paste(lbls, "%", sep="")

pie(min_list, labels = lbls, col = c("purple", "violetred1", "green3","cyan"), main = "Percentage of Ac
```



Interpretation

-Average user spends 79% of a time for Sedentary,Very active and fairly active only make up 2% of the entire time. Lightly active make up of 19% of the total time.

_ This is not recommended and not an ideal result for fitness tracking data.

5.0 PRASE 6 : ACT

5.1 Revisiting Business Task

The business task will be associated with a non-Bellabeat smart devices and analyze the smart device usage data in order to gain insight into how people are already using their smart devices.Then, using the data, generate recommendations for the Bellabeat marketing strategy team to understand the trends. These recommendations and trend insights will be used to enhance the features, functionality, and service quality of the Bellabeat app. So, the business task can be summed up as follows:

1. Analyze the non-Bellabeat smart device usage data in order to gain insight to help Bellabeat marketing strategy team to understand the trends.
2. Utilizing the trends and insights provided by smart device usage data in order to improve the features, functionality, and overall service quality of the Bellabeat app

5.2 Trends Identified

1. The longer a user spends on Sedentary activities, the less probable it is that they would engage in physical activity.
2. The longer a user walks, the greater the chance that they will have a caloric deficit.
3. Total time in bed seems positively related to total time asleep.
4. Average user spends 79% of a time for Sedentary,Very active and fairly active only make up 2% of the entire time. Lightly active make up of 19% of the total time.
5. The participants averaged 25.19 BMI, which is overweight.
6. On average, participants slept less than 8 hours.

5.3 Recommendations

5.3.1 Complete Reward System

- A system of rewards should be devised to recognise people who have done well.in an effort to entice a client
- Reward should be given to those who attain various levels depending on the number of daily steps taken.
- In order to advance to the next level, the user must maintain a certain level of activity for a certain amount of time (a month or a week). For each level, user would get a certain number of points redeemable for Bellabeat items or discounts on other Bellabeat products.
- This strategy will indirectly promote and raise sales of the other Bellabeat product line while also enhancing its reputation in the marketplace.

5.3.2 Social Media Contest

- In exchange for prizes and incentives, a social media contest encourages interaction, followers, leads, or brand exposure.
- Reward may give to the followers for liking, commenting, and sharing Bellabeat product content (facebook page, instagram, tiktok, Youtube and so on). This increases your brand's reach and buzz.
- Like/share/comment to win a Bellabeat product
- Creative video contests or Photo contest to win a Bellabeat product

5.3.3 In-App competition and rankings

- Bellabeat could enable in-app tournaments against friends or users in the same city/state to encourage physical activities engagement.
- Make an animated and creative total steps ranking page for users in the same city or state to get them to be more active and spend less time sitting (sedentary time).

5.3.4 Notification and Sleep time

- Participants slept less than 8 hours on average, therefore Notification and Sleep Time should be a major issue.
- Total time in bed is positively correlated with total time asleep, according to our results. They could set a bedtime and receive a reminder minutes before. Breathing advice, podcasts with relaxing music, and sleep techniques can help customers sleep.
- Bellabeat should therefore create a system that includes sleep guidance, heart rate monitoring while sleeping, and a sleep reminder in order to improve the user's quality of sleep.

5.4 Future Works

- Larger, more representative sample size (with 95% confidence and 5% margin of error).
- Random with no prejudice in selection.
- At least 1 year should be spent collecting data.
- More about the person's age, sex, height, etc.
- More recent and current information or anything from the previous year is suggested.
- Have an original (First Party Data) data source, or verify primary/secondary data for integrity and trustworthiness.