Case Study: How Can a Wellness Technology Company Play It Smart?

Saulius Macys

11/09/2021

About the company

Urška Sršen and Sando Mur founded Bellabeat, a high-tech company that manufactures health-focused smart products. Sršen used her background as an artist to develop beautifully designed technology that informs and inspires women around the world. Collecting data on activity, sleep, stress, and reproductive health has allowed Bellabeat to empower women with knowledge about their own health and habits. Since it was founded in 2013, Bellabeat has grown rapidly and quickly positioned itself as a tech-driven wellness company for women.

Phase 1: Ask

Business Task

Analyze Fitbit's smart device data to gain insights(trends) into how consumers are using their smart devices and present high-level recommendations for Bellabeat's marketing strategy.

Key Stakeholders

- Urška Sršen: Bellabeat's cofounder and Chief Creative Officer
- Sando Mur: Mathematician and Bellabeat's cofounder
- Bellabeat marketing analytics team

Phase 2: Prepare

Installing packages which I will need to use

```r
install.packages('tidyverse')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.1'
## (as 'lib' is unspecified)
```

```r
install.packages('lubridate')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.1'
## (as 'lib' is unspecified)
```

```r
install.packages('skimr')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.1'
## (as 'lib' is unspecified)
```

```r
install.packages('janitor')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.1'
## (as 'lib' is unspecified)
```

```r
install.packages('ggplot2')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.1'
## (as 'lib' is unspecified)
```

```r
install.packages('readr')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.1'
## (as 'lib' is unspecified)
```

Loading libraries of installed packages

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

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

```
##
## Attaching package: 'janitor'

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

```r
library(ggplot2)
library(readr)
```

Importing my CSV files

```r
DailyActivity<- read_csv("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.
```

```
SleepDay<- read_csv("sleepDay_merged.csv")
```

```
## Rows: 413 Columns: 5
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): SleepDay
## dbl (4): Id, TotalSleepRecords, TotalMinutesAsleep, TotalTimeInBed
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
WeightLogInfo<- read_csv("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.
```

I chose to import DailyActivity dataset since it already has merged data for Daily Calories, Intensities and Steps. I added Sleep and Weight Log datasets as well.

Phase 3 and 4: Process and Analyze

Exploring a few key tables

```
head(DailyActivity)
```

```
## # A tibble: 6 x 15
##           Id ActivityDate TotalSteps TotalDistance TrackerDistance LoggedActivitie~
##        <dbl> <chr>             <dbl>         <dbl>           <dbl>            <dbl>
## 1 1503960366 4/12/2016         13162          8.5             8.5                 0
## 2 1503960366 4/13/2016         10735          6.97            6.97                0
## 3 1503960366 4/14/2016         10460          6.74            6.74                0
## 4 1503960366 4/15/2016          9762          6.28            6.28                0
## 5 1503960366 4/16/2016         12669          8.16            8.16                0
## 6 1503960366 4/17/2016          9705          6.48            6.48                0
## # ... with 9 more variables: VeryActiveDistance <dbl>,
## #   ModeratelyActiveDistance <dbl>, LightActiveDistance <dbl>,
## #   SedentaryActiveDistance <dbl>, VeryActiveMinutes <dbl>,
## #   FairlyActiveMinutes <dbl>, LightlyActiveMinutes <dbl>,
## #   SedentaryMinutes <dbl>, Calories <dbl>
```

```
head(SleepDay)
```

```
## # A tibble: 6 x 5
##           Id SleepDay              TotalSleepRecor~ TotalMinutesAsl~ TotalTimeInBed
##        <dbl> <chr>                            <dbl>            <dbl>          <dbl>
## 1 1503960366 4/12/2016 12:00:00 AM                1              327            346
## 2 1503960366 4/13/2016 12:00:00 AM                2              384            407
## 3 1503960366 4/15/2016 12:00:00 AM                1              412            442
## 4 1503960366 4/16/2016 12:00:00 AM                2              340            367
## 5 1503960366 4/17/2016 12:00:00 AM                1              700            712
## 6 1503960366 4/19/2016 12:00:00 AM                1              304            320
```

```
head(WeightLogInfo)
```

```
## # A tibble: 6 x 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 4/13/201~    134.          294.    NA  47.5 FALSE           1.46e12
## 4 2873212765 4/21/201~     56.7         125.    NA  21.5 TRUE            1.46e12
## 5 2873212765 5/12/201~     57.3         126.    NA  21.7 TRUE            1.46e12
## 6 4319703577 4/17/201~     72.4         160.    25  27.5 TRUE            1.46e12
```

Identifying all the columns:

```
colnames(DailyActivity)
```

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

```
colnames(SleepDay)
```

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

```
colnames(WeightLogInfo)
```

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

Understanding some summary statistics

How many unique participants are there in each dataframe?

```
n_distinct(DailyActivity$Id)
```

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

```
n_distinct(SleepDay$Id)
```

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

```
n_distinct(WeightLogInfo$Id)
```

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

Daily Activity dataset has the most participants (33), while Sleep dataframe and Weight dataframe have much less. This could be attributed to smartware devices not being charged at night and participants not measuring their weight as often as other data.

How many observations are there in each dataframe?

```
nrow(DailyActivity)
```

```
## [1] 940
```

```
nrow(SleepDay)
```

```
## [1] 413
```

```
nrow(WeightLogInfo)
```

```
## [1] 67
```

Some quick summary statistics about each data frame:

For the daily activity dataframe I would like to analyze corelation between Steps, Distance and calories:

```
DailyActivity %>%
  select(TotalSteps,
         TotalDistance,
         Calories) %>%
  summary()
```

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

For the Sleep dataframe I will compare total minutes asleep with total minutes in bed:

```
SleepDay %>%
  select(TotalMinutesAsleep,
         TotalTimeInBed) %>%
  summary()
```

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

For the Weight dataframe I will compare weight and BMI relationship:

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

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

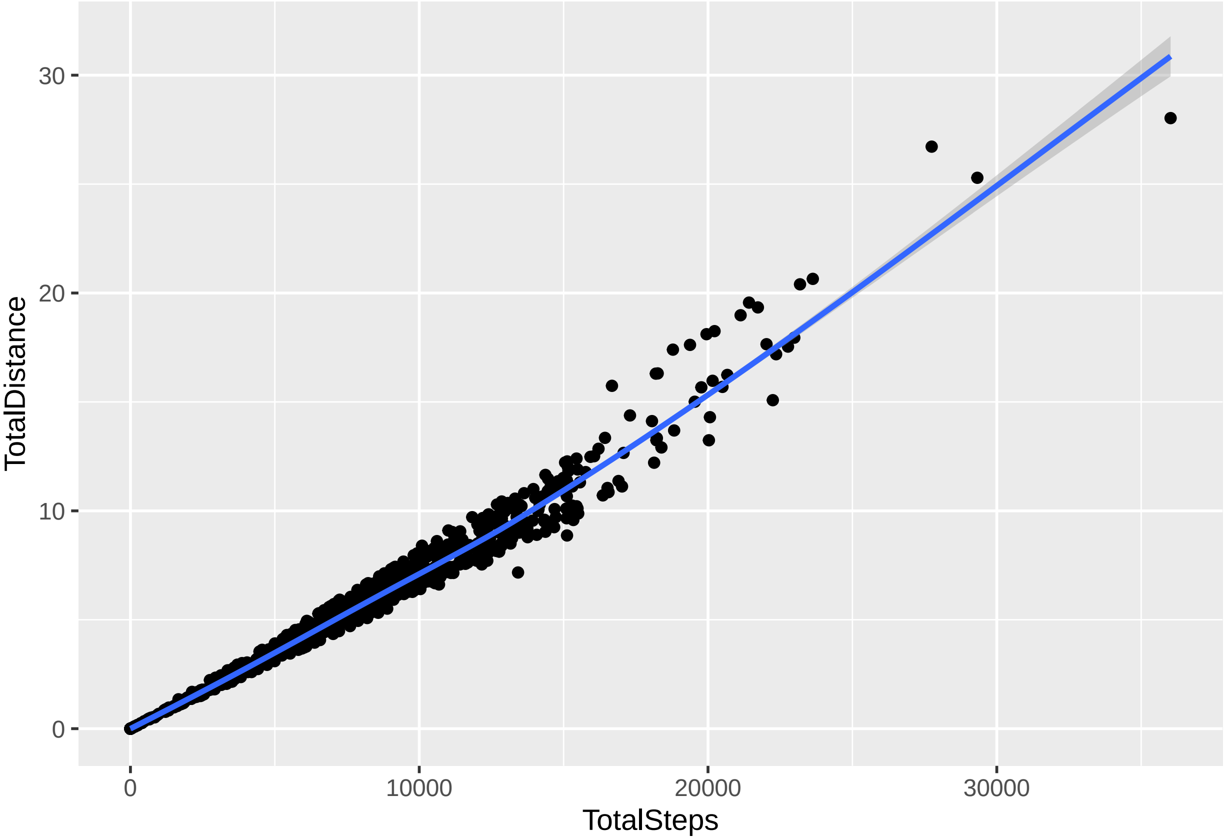
Phase 5: Share

Plotting a few explorations

On the following 3 plots showing a clear tendency between Total Steps, Total Distance and Calories used. The observation is that higher activity leads to more steps completed and more calories burnt.

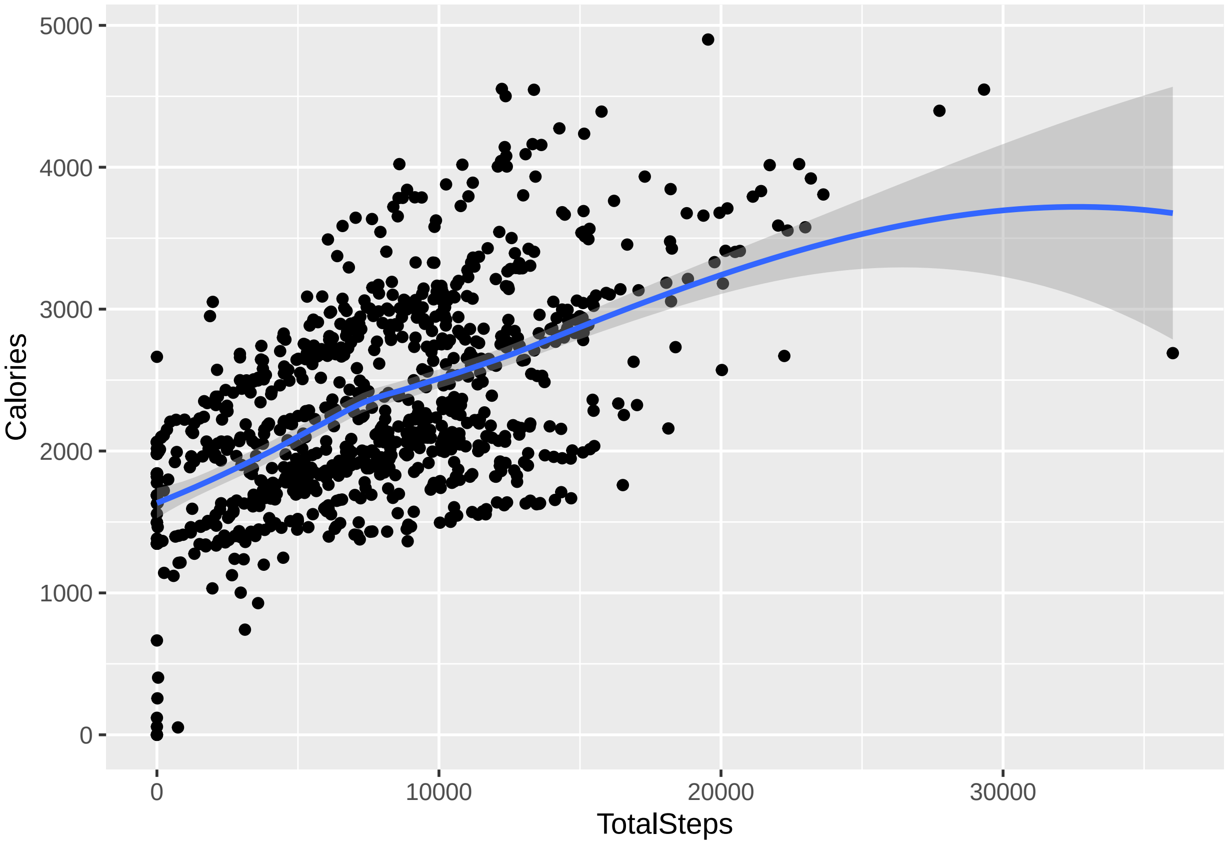
```
ggplot(data=DailyActivity, aes(x=TotalSteps, y=TotalDistance)) +
  geom_point() +
  geom_smooth()
```

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



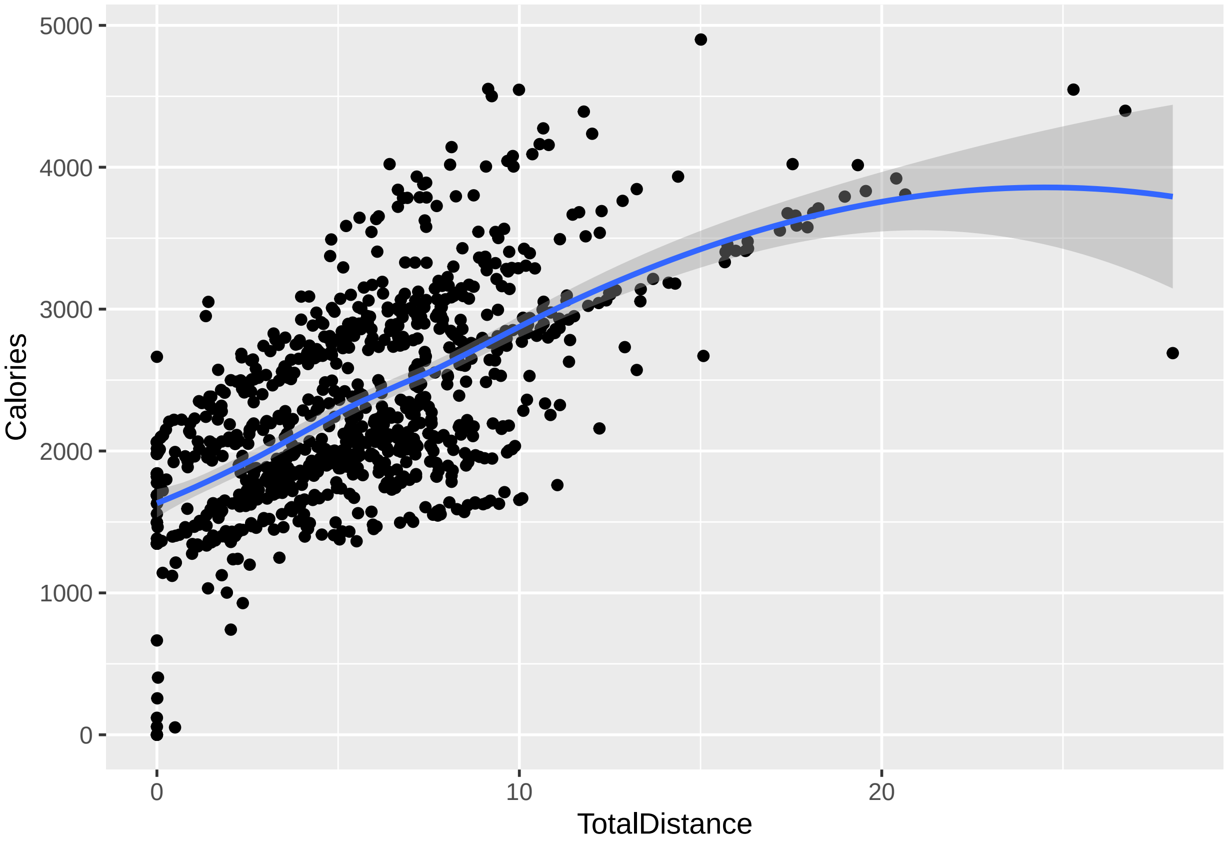
```
ggplot(data=DailyActivity, aes(x=TotalSteps, y=Calories)) +
   geom_point() +
  geom_smooth()
```

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



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

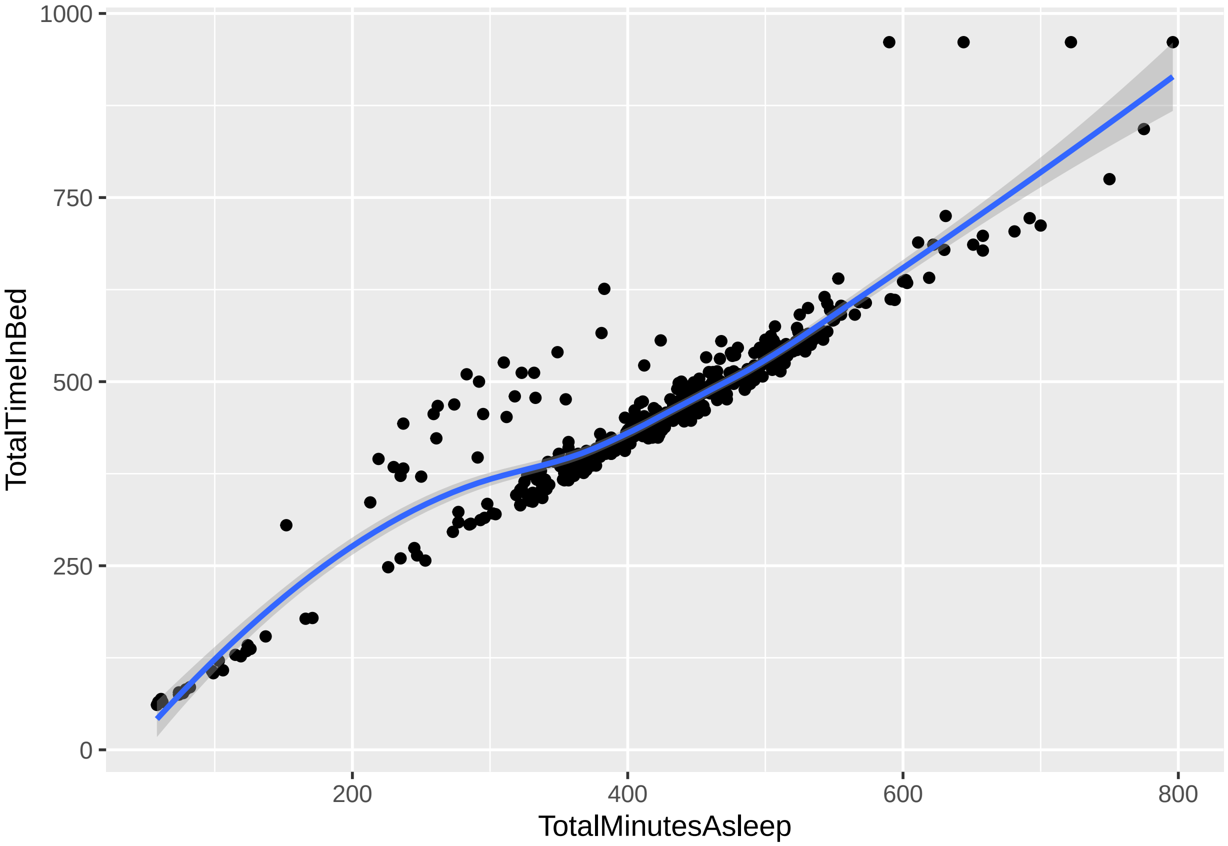
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



The following Sleep plot clearly shows that total minutes asleep are directly proportional to total time spent in bed.

```
ggplot(data=SleepDay, aes(x=TotalMinutesAsleep, y=TotalTimeInBed)) +
   geom_point() +
  geom_smooth()
```

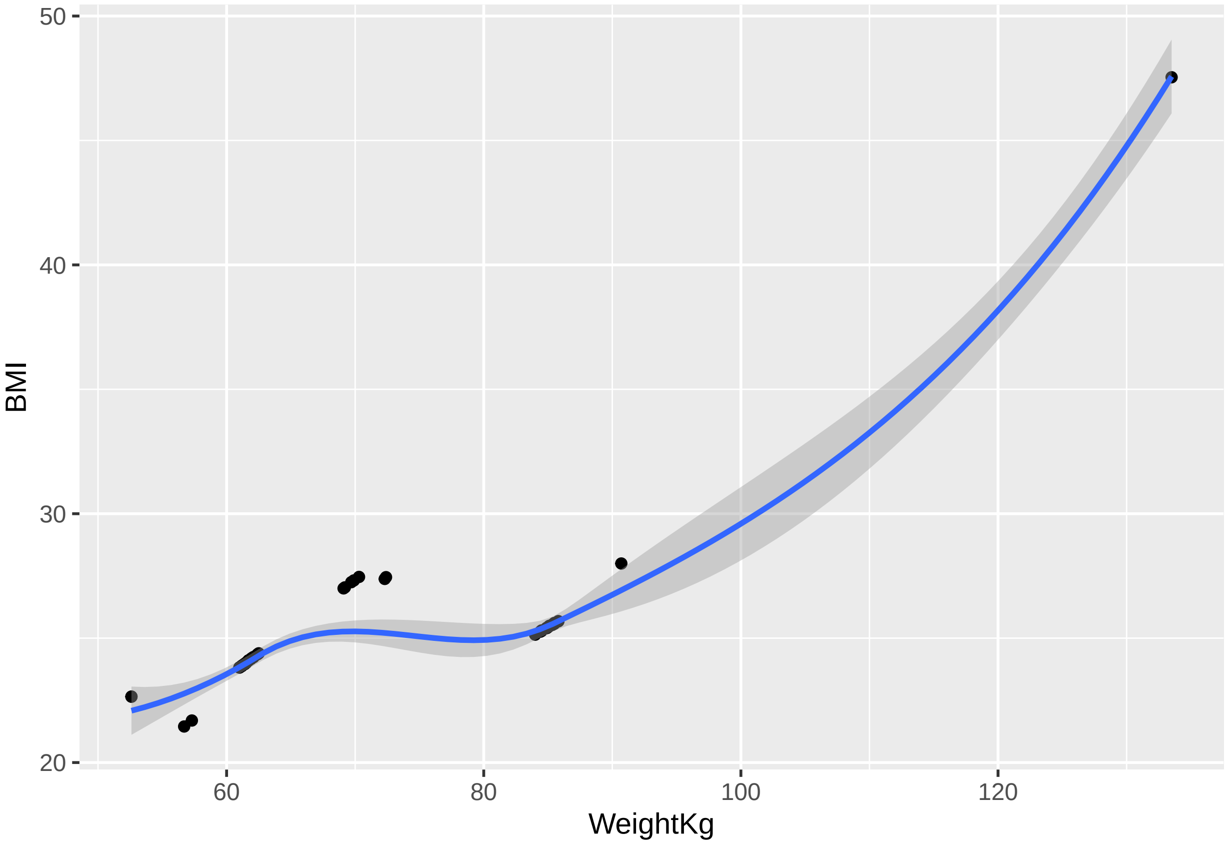
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



The following plot shows that BMI is very related to total weight and the higher weight it is - the more likely that BMI will be higher as well.

```
ggplot(data=WeightLogInfo, aes(x=WeightKg, y=BMI)) +
   geom_point() +
  geom_smooth()
```

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



Phase 6: Act

Based on my analysis I found that users of smart devices are mainly using them for day-to-day activities like steps completed, distance walked and calories burnt. These functions are a must in current market for releasing new smart devices for Bellabeat.

Sleeping data can be very valuable as well but it requires a device which would have a significant battery capacity to ensure that accurate data could be gathered throughout the night.

Weight measurement option is not being used that much in smart devices at the moment, because information needs to be added manually or synced with a digital scale and for this reason I would recommend to exclude it in order to invest resources in other aspects which are more in demand by customers.