# BellaBeat Fitness Company Case Study

# How Can a Wellness Technology Company Play It Smart?

**Background:**

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

**Aim for analyzing this case study**

1.Analyze smart device data to gain insight into how consumers are using their smart devices

2.Present the analysis to the Bellabeat executive team along with your high-level recommendations for Bellabeat's marketing strategy

**Bellabeat's Current Marketting Strategy**

1.Bellabeat products became available through a growing number of online retailers in addition to their own website.

2.The company has invested in traditional advertising media: radio, out-of-home billboards, print, and television.

3.Focused on digital marketing extensively.

4.Invests year-round in Google Search , maintaining active Facebook and Instagram pages, and consistently engages consumers on twitter.

5.Runs video ads on Youtube and display ads on the Google Display Network to support campaigns around key marketing dates.

**Google Data Analytics Steps:** Ask, Prepare, Process, Analyze, Share, Act

Let's go with the steps one by one

**-> Ask Phase : What do we need to know?**

What are some trends in smart device usage?

How could these trends apply to Bellabeat customers?

How could these trends help influence Bellabeat marketing strategy?

Some questions after looking through the datasets will be

What is the problem you are trying to solve?

How can your insights drive business decisions?

**-> Prepare Phase : What is our Data?**

**Data Collection:**

**What?:** This Kaggle public data set contains personal fitness tracker data from thirty FitBit users, including minute-level data for physical activity, heart rate, and sleep monitoring.

**Who?:** The dataset was collected by Amazon Mechanical Turk from consenting FitBit users in their survey.

**Why?:** The data was to collected with an inspiration to understand Human temporal routine behavior and pattern recognition.

**Data Brief:** The dataset contains 18 tables linked via the user IDs and timestamps. These tables contain information pertaining to the various intensities of physical activity,sedantary periods, calories burned (measured on daily basis, hourly basis and minute-wise basis)sleep duration(daily and minute-wise) and its frequency, weight logs, as well as heartrate.

**Data Limitations:**

The data has been collected with predefined datasets with limited data. This makes it outdated to use and get current data.

The sample size is of around 33 participants for most of the parameters and lesser for parmeters like heart rate per second (7) and weight (8). It is not sufficient to establish a good confidence level.

The data set does not provide any demographic information pertaining to the participants which makes it difficult to analyse the data with respect to Bellabeat's target of female customers.

# Findings to proceed with data

**Preparing Work Environment by loading Packages and Data**

```
# Loading the required Libraries

library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✔ ggplot2 3.3.6      ✔ purrr   0.3.5 
## ✔ tibble  3.1.8      ✔ dplyr   1.0.10
## ✔ tidyr   1.2.1      ✔ stringr 1.4.1 
## ✔ readr   2.1.3      ✔ forcats 0.5.2 
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(skimr)
```

```
## Warning: package 'skimr' was built under R version 4.2.2
```

```
library(janitor)
```

```
## Warning: package 'janitor' was built under R version 4.2.2
```

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

```
library(dplyr)
library(lubridate)
```

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

```
library(ggplot2)
library(here)
```

```
## Warning: package 'here' was built under R version 4.2.2
```

```
## here() starts at E:/Fitabase Data 4.12.16-5.12.16/files_used
```

```
# Loading the data

daily_activity <- read.csv("dailyActivity_merged.csv")
hourly_intensity <- read.csv("hourlyIntensities_merged.csv")
daily_sleep <- read.csv("sleepDay_merged.csv")
weight <- read.csv("weightLogInfo_merged.csv")
```

```
# Checking the datasets with either head()

head(daily_activity)
```

```
##           Id ActivityDate TotalSteps TotalDistance TrackerDistance
## 1 1503960366   04-12-2016      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
```

```
head(hourly_intensity)
```

```
##           Id          ActivityHour TotalIntensity AverageIntensity
## 1 1503960366 4/12/2016 12:00:00 AM             20         0.333333
## 2 1503960366  4/12/2016 1:00:00 AM              8         0.133333
## 3 1503960366  4/12/2016 2:00:00 AM              7         0.116667
## 4 1503960366  4/12/2016 3:00:00 AM              0         0.000000
## 5 1503960366  4/12/2016 4:00:00 AM              0         0.000000
## 6 1503960366  4/12/2016 5:00:00 AM              0         0.000000
```

```
head(daily_sleep)
```

```
##           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
```

```
head(weight)
```

```
##           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
```

**-> Process Phase: Data Cleaning**

For ensuring that each table is cleaned we need to perform the following:

Ensure the naming consistancy.

Check and Remove duplicates and errors if any.

Store time and date in different columns.

**Removing duplicates and ensuting naming conventions**

```
daily_activity<- daily_activity%>%
clean_names()%>%
unique()%>%
glimpse()
```

```
## Rows: 940
## Columns: 15
## $ id                         <dbl> 1503960366, 1503960366, 1503960366, 1503960…
## $ activity_date              <chr> "04-12-2016", "4/13/2016", "4/14/2016", "4/…
## $ total_steps                <int> 13162, 10735, 10460, 9762, 12669, 9705, 130…
## $ total_distance             <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9…
## $ tracker_distance           <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9…
## $ logged_activities_distance <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ very_active_distance       <dbl> 1.88, 1.57, 2.44, 2.14, 2.71, 3.19, 3.25, 3…
## $ moderately_active_distance <dbl> 0.55, 0.69, 0.40, 1.26, 0.41, 0.78, 0.64, 1…
## $ light_active_distance      <dbl> 6.06, 4.71, 3.91, 2.83, 5.04, 2.51, 4.71, 5…
## $ sedentary_active_distance  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ very_active_minutes        <int> 25, 21, 30, 29, 36, 38, 42, 50, 28, 19, 66,…
## $ fairly_active_minutes      <int> 13, 19, 11, 34, 10, 20, 16, 31, 12, 8, 27, …
## $ lightly_active_minutes     <int> 328, 217, 181, 209, 221, 164, 233, 264, 205…
## $ sedentary_minutes          <int> 728, 776, 1218, 726, 773, 539, 1149, 775, 8…
## $ calories                   <int> 1985, 1797, 1776, 1745, 1863, 1728, 1921, 2…
```

```
hourly_intensity<- hourly_intensity%>%
clean_names()%>%
unique()%>%
glimpse()
```

```
## Rows: 22,099
## Columns: 4
## $ id                <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 1503…
## $ activity_hour     <chr> "4/12/2016 12:00:00 AM", "4/12/2016 1:00:00 AM", "4/…
## $ total_intensity   <int> 20, 8, 7, 0, 0, 0, 0, 0, 13, 30, 29, 12, 11, 6, 36, …
## $ average_intensity <dbl> 0.333333, 0.133333, 0.116667, 0.000000, 0.000000, 0.…
```

```
daily_sleep<-daily_sleep%>%
clean_names()%>%
unique()%>%
glimpse()
```

```
## Rows: 410
## Columns: 5
## $ id                   <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 1…
## $ sleep_day            <chr> "4/12/2016 12:00:00 AM", "4/13/2016 12:00:00 AM",…
## $ total_sleep_records  <int> 1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ total_minutes_asleep <int> 327, 384, 412, 340, 700, 304, 360, 325, 361, 430,…
## $ total_time_in_bed    <int> 346, 407, 442, 367, 712, 320, 377, 364, 384, 449,…
```

```
weight<- weight%>%
clean_names()%>%
unique()%>%
glimpse()
```

```
## Rows: 67
## Columns: 8
## $ id               <dbl> 1503960366, 1503960366, 1927972279, 2873212765, 28732…
## $ date             <chr> "5/2/2016 11:59:59 PM", "5/3/2016 11:59:59 PM", "4/13…
## $ weight_kg        <dbl> 52.6, 52.6, 133.5, 56.7, 57.3, 72.4, 72.3, 69.7, 70.3…
## $ weight_pounds    <dbl> 115.9631, 115.9631, 294.3171, 125.0021, 126.3249, 159…
## $ fat              <int> 22, NA, NA, NA, NA, 25, NA, NA, NA, NA, NA, NA, NA, N…
## $ bmi              <dbl> 22.65, 22.65, 47.54, 21.45, 21.69, 27.45, 27.38, 27.2…
## $ is_manual_report <chr> "True", "True", "False", "True", "True", "True", "Tru…
## $ log_id           <dbl> 1.462234e+12, 1.462320e+12, 1.460510e+12, 1.461283e+1…
```

### Making date formats consistent and Store time and date in different columns

```
# daily_activity table
daily_activity$date<-mdy(daily_activity$activity_date)
#adding a new column "day" for finding out the day-wise trends
daily_activity$day<- format(daily_activity$date, "%A")

daily_activity<- subset(daily_activity, select= -c(activity_date))
head(daily_activity)
```

```
##           id total_steps total_distance tracker_distance
## 1 1503960366       13162           8.50             8.50
## 2 1503960366       10735           6.97             6.97
## 3 1503960366       10460           6.74             6.74
## 4 1503960366        9762           6.28             6.28
## 5 1503960366       12669           8.16             8.16
## 6 1503960366        9705           6.48             6.48
##   logged_activities_distance very_active_distance moderately_active_distance
## 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
##   light_active_distance sedentary_active_distance very_active_minutes
## 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
##   fairly_active_minutes lightly_active_minutes sedentary_minutes 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
##         date       day
## 1 2016-04-12   Tuesday
## 2 2016-04-13 Wednesday
## 3 2016-04-14  Thursday
## 4 2016-04-15    Friday
## 5 2016-04-16  Saturday
## 6 2016-04-17    Sunday
```

```
# hourly_intensity table
hourly_intensity$date<- mdy_hms(hourly_intensity$activity_hour)
hourly_intensity$intensity_hour<-format(hourly_intensity$date,format= "%H:%M")
```

```
hourly_intensity$intensity_day<-format(hourly_intensity$date, format= "%A")
hourly_intensity<- subset(hourly_intensity, select= -activity_hour)
head(hourly_intensity)
```

```
##           id total_intensity average_intensity                date
## 1 1503960366              20          0.333333 2016-04-12 00:00:00
## 2 1503960366               8          0.133333 2016-04-12 01:00:00
## 3 1503960366               7          0.116667 2016-04-12 02:00:00
## 4 1503960366               0          0.000000 2016-04-12 03:00:00
## 5 1503960366               0          0.000000 2016-04-12 04:00:00
## 6 1503960366               0          0.000000 2016-04-12 05:00:00
##   intensity_hour intensity_day
## 1          00:00       Tuesday
## 2          01:00       Tuesday
## 3          02:00       Tuesday
## 4          03:00       Tuesday
## 5          04:00       Tuesday
## 6          05:00       Tuesday
```

```
#sleep log table
daily_sleep$date<- mdy_hms(daily_sleep$sleep_day)
daily_sleep$day<- format(daily_sleep$date, "%A")
# adding a column to determine the time participants lie awake in bed
new_sleep<-daily_sleep%>%
mutate(total_time_awake_in_bed= ( total_time_in_bed- total_minutes_asleep) )%>%
glimpse()
```

```
## Rows: 410
## Columns: 8
## $ id                      <dbl> 1503960366, 1503960366, 1503960366, 1503960366…
## $ sleep_day               <chr> "4/12/2016 12:00:00 AM", "4/13/2016 12:00:00 A…
## $ total_sleep_records     <int> 1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ total_minutes_asleep    <int> 327, 384, 412, 340, 700, 304, 360, 325, 361, 4…
## $ total_time_in_bed       <int> 346, 407, 442, 367, 712, 320, 377, 364, 384, 4…
## $ date                    <dttm> 2016-04-12, 2016-04-13, 2016-04-15, 2016-04-1…
## $ day                     <chr> "Tuesday", "Wednesday", "Friday", "Saturday", …
## $ total_time_awake_in_bed <int> 19, 23, 30, 27, 12, 16, 17, 39, 23, 19, 46, 29…
```

```
new_sleep<- subset(new_sleep, select= -sleep_day)
head(new_sleep)
```

```
##           id total_sleep_records total_minutes_asleep total_time_in_bed
## 1 1503960366                   1                  327               346
## 2 1503960366                   2                  384               407
## 3 1503960366                   1                  412               442
## 4 1503960366                   2                  340               367
## 5 1503960366                   1                  700               712
## 6 1503960366                   1                  304               320
##         date       day total_time_awake_in_bed
## 1 2016-04-12   Tuesday                      19
## 2 2016-04-13 Wednesday                      23
## 3 2016-04-15    Friday                      30
## 4 2016-04-16  Saturday                      27
## 5 2016-04-17    Sunday                      12
## 6 2016-04-19   Tuesday                      16
```

```
#weight table
weight$date<- mdy_hms(weight$date)
weight$weight_day<-format(weight$date, "%A")
head(weight)
```

```
##           id                date weight_kg weight_pounds fat   bmi
## 1 1503960366 2016-05-02 23:59:59      52.6      115.9631  22 22.65
## 2 1503960366 2016-05-03 23:59:59      52.6      115.9631  NA 22.65
## 3 1927972279 2016-04-13 01:08:52     133.5      294.3171  NA 47.54
## 4 2873212765 2016-04-21 23:59:59      56.7      125.0021  NA 21.45
## 5 2873212765 2016-05-12 23:59:59      57.3      126.3249  NA 21.69
## 6 4319703577 2016-04-17 23:59:59      72.4      159.6147  25 27.45
##   is_manual_report       log_id weight_day
## 1             True 1.462234e+12     Monday
## 2             True 1.462320e+12    Tuesday
## 3            False 1.460510e+12  Wednesday
## 4             True 1.461283e+12   Thursday
## 5             True 1.463098e+12   Thursday
## 6             True 1.460938e+12     Sunday
```

### Ensuring that the individual distances sums up to the total distance

```
l<-daily_activity%>%
mutate(new_sum= light_active_distance+moderately_active_distance+very_active_distance)%>%
subset(select=c(id, total_distance,new_sum, light_active_distance,moderately_active_distance,ve
ry_active_distance))
head(l)
```

```
##           id total_distance new_sum light_active_distance
## 1 1503960366           8.50    8.49                  6.06
## 2 1503960366           6.97    6.97                  4.71
## 3 1503960366           6.74    6.75                  3.91
## 4 1503960366           6.28    6.23                  2.83
## 5 1503960366           8.16    8.16                  5.04
## 6 1503960366           6.48    6.48                  2.51
##   moderately_active_distance very_active_distance
## 1                       0.55                 1.88
## 2                       0.69                 1.57
## 3                       0.40                 2.44
## 4                       1.26                 2.14
## 5                       0.41                 2.71
## 6                       0.78                 3.19
```

As evident from above:

1.The sum of induvidual distances(light_active_distance,moderately_active_distance,very_active_distance) does not add upto the total distance specified. This makes any analysis based upon **the sum of these values incorrect**. Hence we will not use the sum of these value in any part of our analysis.

2.Another noticable fact is that **the most users are majorly lightly active throughout the total distance they cover**. Hence, the need to promote some high intensity activites, so that the users may see evident changes.

Now we will be **merging** the cleaned daily_activity and daily_sleep tables for ease of analysis and visualisation

```
#merging daily_activity and new_sleep tables on date, day and userIDs.
#Since daily activity table has more number of participants we will use"all.x=TRUE"
#to ensure that all the non-matching cases of x are appended to the result as well.
```

```
daily<- merge(daily_activity, new_sleep, by= c("id", "date", "day"), all.x= TRUE)
daily<- distinct(daily)
glimpse(daily)
```

```
## Rows: 940
## Columns: 20
## $ id                         <dbl> 1503960366, 1503960366, 1503960366, 1503960…
## $ date                       <date> 2016-04-12, 2016-04-13, 2016-04-14, 2016-0…
## $ day                        <chr> "Tuesday", "Wednesday", "Thursday", "Friday…
## $ total_steps                <int> 13162, 10735, 10460, 9762, 12669, 9705, 130…
## $ total_distance             <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9…
## $ tracker_distance           <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9…
## $ logged_activities_distance <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ very_active_distance       <dbl> 1.88, 1.57, 2.44, 2.14, 2.71, 3.19, 3.25, 3…
## $ moderately_active_distance <dbl> 0.55, 0.69, 0.40, 1.26, 0.41, 0.78, 0.64, 1…
## $ light_active_distance      <dbl> 6.06, 4.71, 3.91, 2.83, 5.04, 2.51, 4.71, 5…
## $ sedentary_active_distance  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ very_active_minutes        <int> 25, 21, 30, 29, 36, 38, 42, 50, 28, 19, 66,…
## $ fairly_active_minutes      <int> 13, 19, 11, 34, 10, 20, 16, 31, 12, 8, 27, …
## $ lightly_active_minutes     <int> 328, 217, 181, 209, 221, 164, 233, 264, 205…
## $ sedentary_minutes          <int> 728, 776, 1218, 726, 773, 539, 1149, 775, 8…
## $ calories                   <int> 1985, 1797, 1776, 1745, 1863, 1728, 1921, 2…
## $ total_sleep_records        <int> 1, 2, NA, 1, 2, 1, NA, 1, 1, 1, NA, 1, 1, 1…
## $ total_minutes_asleep       <int> 327, 384, NA, 412, 340, 700, NA, 304, 360, …
## $ total_time_in_bed          <int> 346, 407, NA, 442, 367, 712, NA, 320, 377, …
## $ total_time_awake_in_bed    <int> 19, 23, NA, 30, 27, 12, NA, 16, 17, 39, NA,…
```

**-> Analyse Phase : Finding Trends in Data**

**Summary statistics of all our tables: To get brief insights of our data**

```
# daily table
summary(select(daily, total_steps, total_distance, sedentary_minutes, calories,
total_minutes_asleep, total_time_in_bed,   total_time_awake_in_bed))
```

```
##   total_steps    total_distance   sedentary_minutes    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  
##                                                                   
##  total_minutes_asleep total_time_in_bed total_time_awake_in_bed
##  Min.   : 58.0        Min.   : 61.0     Min.   :  0.00         
##  1st Qu.:361.0        1st Qu.:403.8     1st Qu.: 17.00         
##  Median :432.5        Median :463.0     Median : 25.50         
##  Mean   :419.2        Mean   :458.5     Mean   : 39.31         
##  3rd Qu.:490.0        3rd Qu.:526.0     3rd Qu.: 40.00         
##  Max.   :796.0        Max.   :961.0     Max.   :371.00         
##  NA's   :530          NA's   :530       NA's   :530
```

```
# hourly_intensity table
 summary(select(hourly_intensity, total_intensity))
```

```
##  total_intensity 
##  Min.   :  0.00  
##   1st Qu.:  0.00  
##   Median :  3.00  
##   Mean   : 12.04  
##   3rd Qu.: 16.00  
##   Max.   :180.00
```

From the above we can note that:

1.The **mean** of total_steps for about 34 participants is below 8,000 which is not sufficient to see maximum health benefits. This shows that most users of fitness devices would require some motivation to increase their activity levels.

2.The mean **sedantary minutes are a staggering 991.2 minutes (16.5 hours)**. This metric gives us the insight that most fitness device users are likely to be people who have long sitting hours and do a desk job- a great idea of target group for bellabeat marketing.

3.The analysis shows that **an average participant remains awake in bed for 21-30 minutes**, before they are finally able to sleep.

4.The mean Intensity per hour for the participants was merely **12 minutes**.

There are days for each participant when no information has been collected. Let us find out the total number of days for each participant when the data was collected.

```
days<- daily%>%
group_by(id)%>%
summarise(number_of_days_info_collected= n_distinct(date))%>%
arrange(number_of_days_info_collected)
days
```

```
## # A tibble: 33 × 2
##            id number_of_days_info_collected
##         <dbl>                         <int>
##  1 4057192912                             4
##  2 2347167796                            18
##  3 8253242879                            19
##  4 3372868164                            20
##  5 6775888955                            26
##  6 7007744171                            26
##  7 6117666160                            28
##  8 6290855005                            29
##  9 8792009665                            29
## 10 1644430081                            30
## # … with 23 more rows
```

As we see, information was **not collected for all the participants on all the days**. Why did the device did not collect information on few days? was the device not used on said dates? or perhaps the participants were involved in other activities like jump rope, swimming, bicycling etc that the device could not register?

**Finding out the target audience**

We can do this by analysing either the steps taken, the calories burned or the very_active_minutes against the days of the week.

```
# by Calories burned
ta_calories<-daily%>%
group_by(day)%>%
summarise(sum_calories= sum(calories) )
ta_calories
```

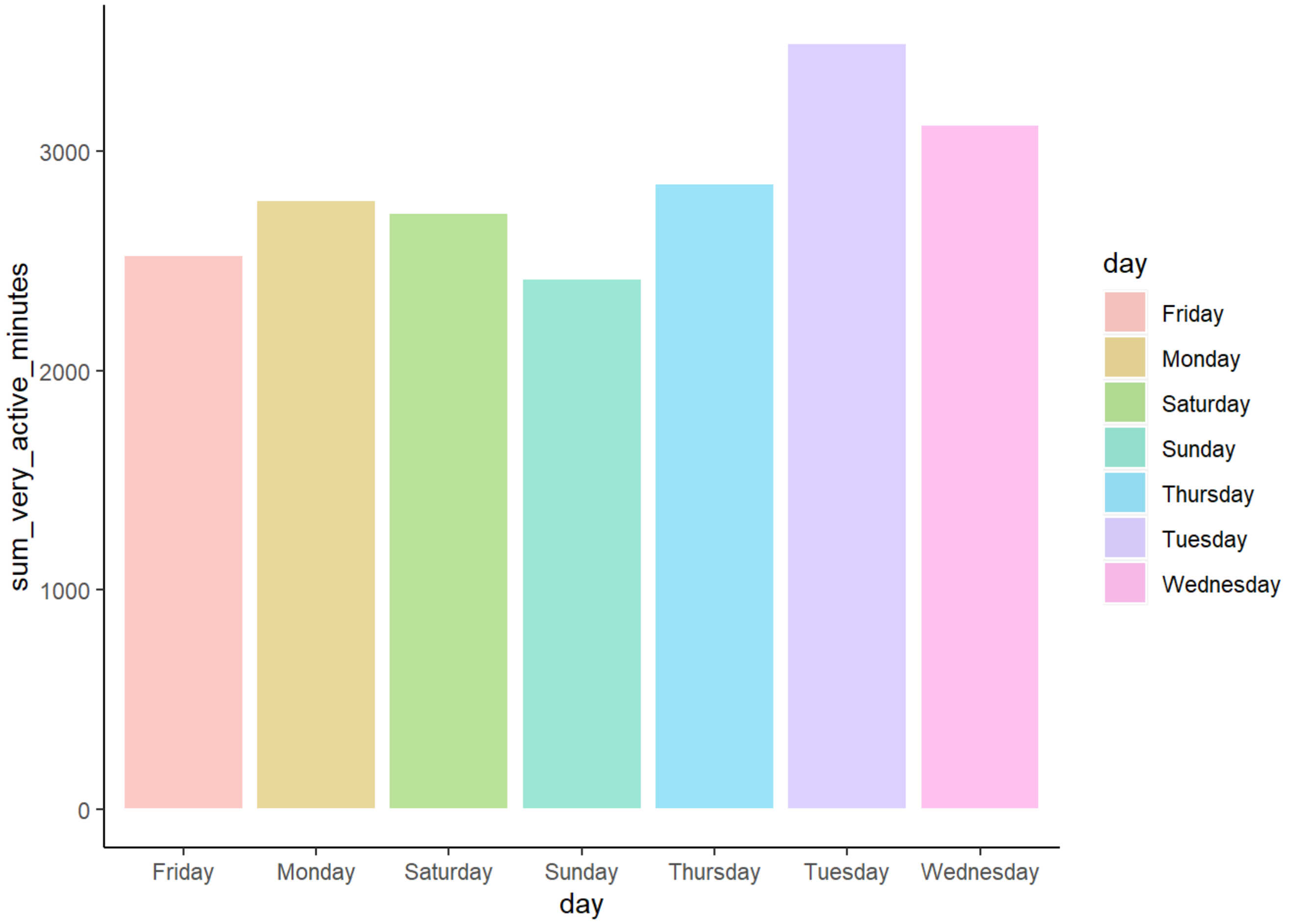
```
## # A tibble: 7 × 2
##   day       sum_calories
##   <chr>            <int>
## 1 Friday          293805
## 2 Monday          278905
## 3 Saturday        292016
## 4 Sunday          273823
## 5 Thursday        323337
## 6 Tuesday         358114
## 7 Wednesday       345393
```

```
# By very_active_minutes
ta_sum_very_active_mins<-daily%>%
group_by(day)%>%
summarise(sum_very_active_mins= sum(very_active_minutes) )
ta_sum_very_active_mins
```

```
## # A tibble: 7 × 2
##   day       sum_very_active_mins
##   <chr>                    <int>
## 1 Friday                    2527
## 2 Monday                    2773
## 3 Saturday                  2718
## 4 Sunday                    2418
## 5 Thursday                  2853
## 6 Tuesday                   3489
## 7 Wednesday                 3117
```

```
# Plotting the results down for better understanding, with geom_histogram and some fancy esthet
ics
ggplot(daily,mapping=aes(x=day, weight=very_active_minutes, fill=day))+geom_histogram(stat="cou
nt",
colour="white", alpha=0.4)+labs(y="sum_very_active_minutes")+theme(panel.background=element_bla
nk(), axis.line=element_line(colour="black"))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

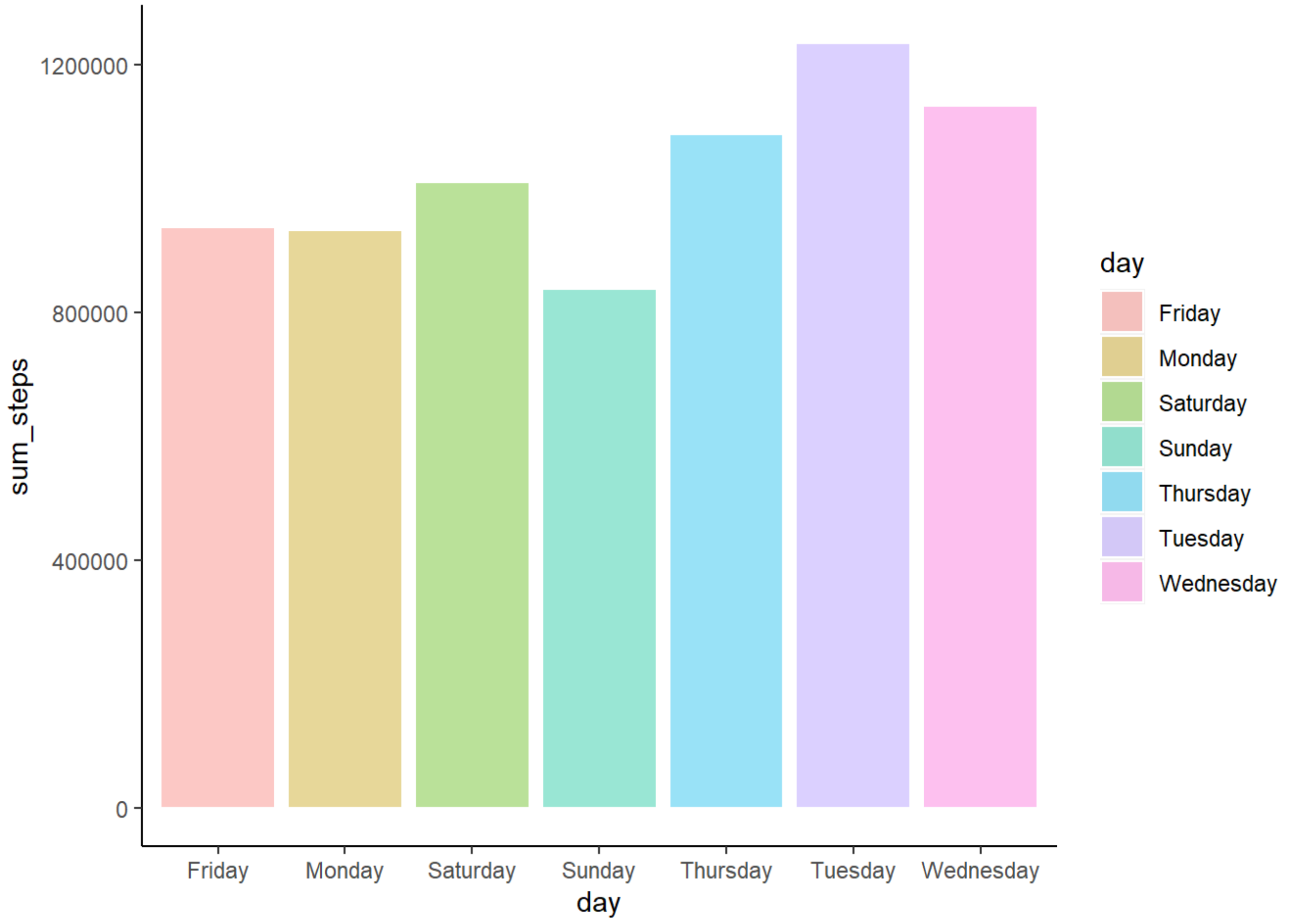


Let us confirm our analysis once, with similar calculations and plotting for **total_steps taken on each day**

```
# By total_steps taken on each day
ta_sum_steps<- daily%>%
group_by(day)%>%
summarise(sum_steps= sum(total_steps))
ta_sum_steps
```

```
## # A tibble: 7 × 2
##   day        sum_steps
##   <chr>          <int>
## 1 Friday        938477
## 2 Monday        933704
## 3 Saturday     1010969
## 4 Sunday        838921
## 5 Thursday     1088658
## 6 Tuesday      1235001
## 7 Wednesday    1133906
```

```
# Plotting the results with  geom_col
plot<-ggplot(ta_sum_steps, mapping=aes(x= day, y=sum_steps, fill=day))+geom_col(colour="white",
alpha= 0.4)
plot+theme(panel.background=element_blank(), axis.line=element_line(colour="black"))
```



from the above we achieve that:

The most active days for participants are **Tuesdays and Wednesdays**

The least active day for participants are **Sundays followed by Fridays**

So the target audience for Bellabeat is one that most likely:

aims to relax and rest on the beginning and end of weekends(Fridays and Sundays) but

follows a routine and remains active throughout the workdays(mon-thursdays) ,

while being most energetic and active during the mid-week period(Tuesdays and Wednesdays)

```
ta_very_active_dist_sum<-daily%>%
group_by(day)%>%
summarise(very_active_dist_sum= sum(very_active_distance))
ta_very_active_dist_sum
```

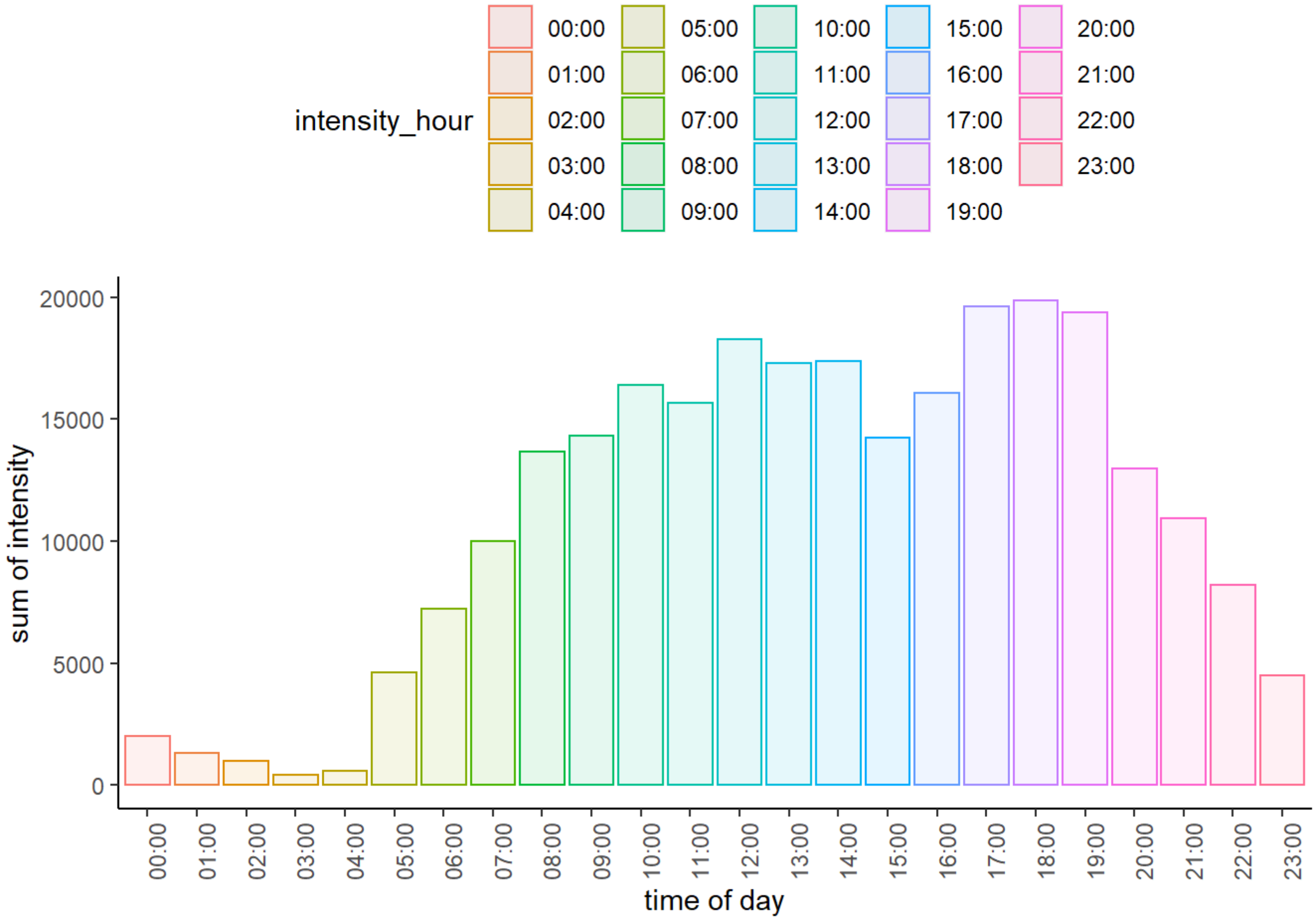
```
## # A tibble: 7 × 2
##   day       very_active_dist_sum
##   <chr>                    <dbl>
## 1 Friday                    165.
## 2 Monday                    184.
## 3 Saturday                  188.
## 4 Sunday                    180.
## 5 Thursday                  204.
## 6 Tuesday                   245.
## 7 Wednesday                 245.
```

**Let us also find out the time of the day when the users are most active**

```
ta_hourly_intensity<- hourly_intensity%>%
group_by(intensity_hour)%>%
summarise(sum_total_intensity= sum(total_intensity))
ta_hourly_intensity
```

```
## # A tibble: 24 × 2
##    intensity_hour sum_total_intensity
##    <chr>                        <int>
##  1 00:00                         1989
##  2 01:00                         1324
##  3 02:00                          974
##  4 03:00                          414
##  5 04:00                          590
##  6 05:00                         4614
##  7 06:00                         7235
##  8 07:00                         9993
##  9 08:00                        13656
## 10 09:00                        14326
## # … with 14 more rows
```

```
# Visulaising the results
p<-ggplot(ta_hourly_intensity, aes(intensity_hour, sum_total_intensity, colour=intensity_hour,
fill=intensity_hour ))+ geom_col(alpha=0.1)
p+labs(x="time of day", y="sum of intensity")+ theme(axis.text.x= element_text(angle=90),
legend.position = "top", panel.background=element_blank(), axis.line=element_line(colour="blac
k"))
```



From the above we can note that:

The most active Time period for participants is **around 5pm to 7pm**

The least active timeframe for the participants is obviously the early morning hours of **2am to 4am**

So **the target audience for Bellabeat** is one that is most likely:

full time workers who focus on their physical health after work hours are finished (5pm to 7pm)

### Further Visual Analysis

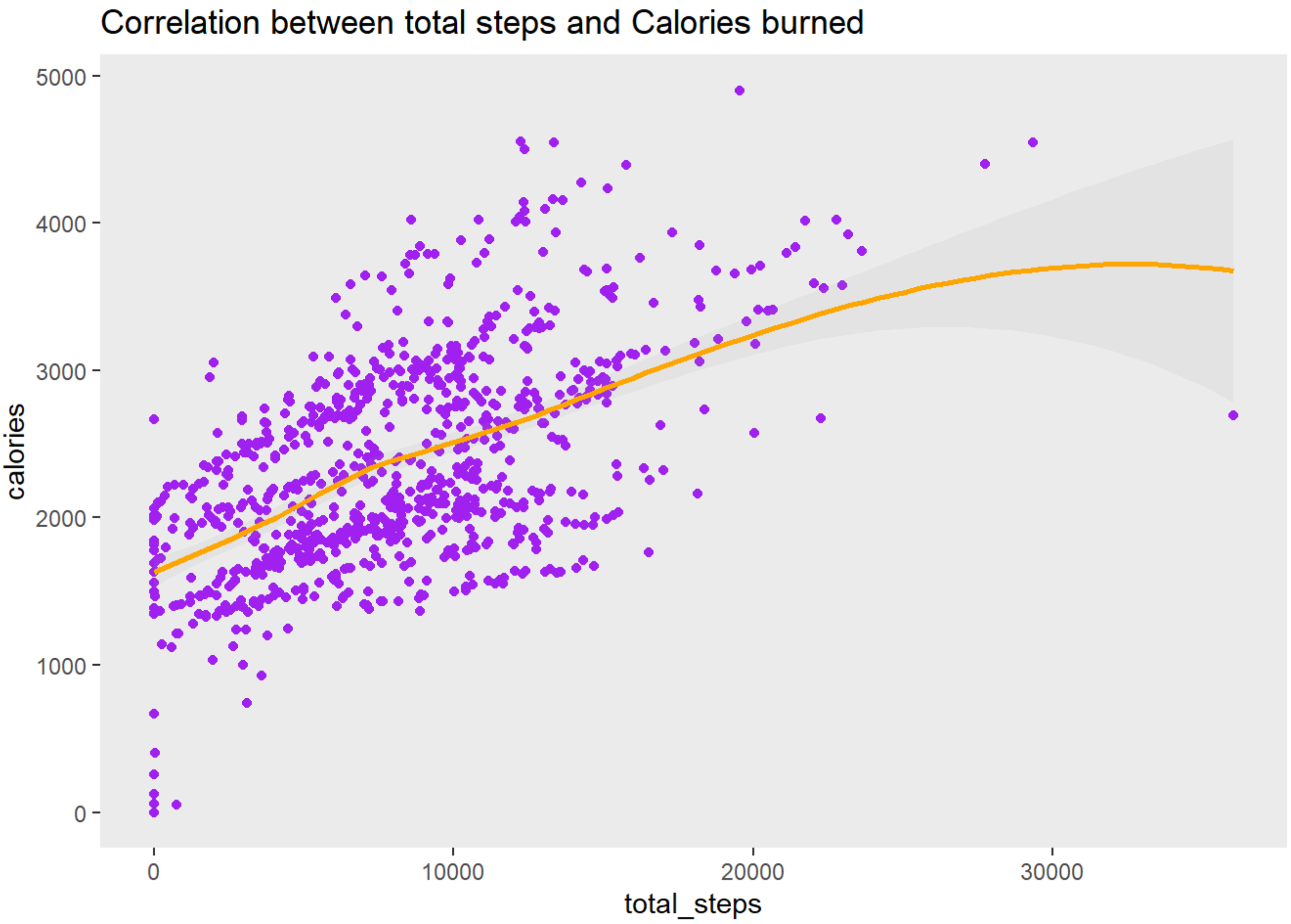
Understanding **the correlation between the calories burned and the total steps taken**

```
cor(x=daily$total_steps, y=daily$calories, method="pearson")
```

```
## [1] 0.5915681
```

```
ggplot(daily,aes(x= total_steps, y=calories))+geom_point(colour="purple")+ geom_smooth(alpha=0.
1, colour= "orange")+labs(title = "Correlation between total steps and Calories burned")+theme
(panel.grid.major=element_blank(), panel.grid.minor=element_blank())
```

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



The above analysis shows a **slightly positive relation between increase in daily number of steps and calories burned** i.e with every increase in the number of steps, the calories burned increases.

This information can be used to market features that involve setting and attaining goals to burn more calories by increasing the daily steps.

Understanding **the correlation between Calories burned and the sleeplessness period in bed**

```
daily_without_na<-daily%>%
drop_na()
glimpse(daily_without_na)
```

```
## Rows: 410
## Columns: 20
## $ id                         <dbl> 1503960366, 1503960366, 1503960366, 1503960…
## $ date                       <date> 2016-04-12, 2016-04-13, 2016-04-15, 2016-0…
## $ day                        <chr> "Tuesday", "Wednesday", "Friday", "Saturday…
## $ total_steps                <int> 13162, 10735, 9762, 12669, 9705, 15506, 105…
## $ total_distance             <dbl> 8.50, 6.97, 6.28, 8.16, 6.48, 9.88, 6.68, 6…
## $ tracker_distance           <dbl> 8.50, 6.97, 6.28, 8.16, 6.48, 9.88, 6.68, 6…
## $ logged_activities_distance <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ very_active_distance       <dbl> 1.88, 1.57, 2.14, 2.71, 3.19, 3.53, 1.96, 1…
## $ moderately_active_distance <dbl> 0.55, 0.69, 1.26, 0.41, 0.78, 1.32, 0.48, 0…
## $ light_active_distance      <dbl> 6.06, 4.71, 2.83, 5.04, 2.51, 5.03, 4.24, 4…
## $ sedentary_active_distance  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
## $ very_active_minutes        <int> 25, 21, 29, 36, 38, 50, 28, 19, 41, 39, 73,…
## $ fairly_active_minutes      <int> 13, 19, 34, 10, 20, 31, 12, 8, 21, 5, 14, 2…
## $ lightly_active_minutes     <int> 328, 217, 209, 221, 164, 264, 205, 211, 262…
## $ sedentary_minutes          <int> 728, 776, 726, 773, 539, 775, 818, 838, 732…
## $ calories                   <int> 1985, 1797, 1745, 1863, 1728, 2035, 1786, 1…
## $ total_sleep_records        <int> 1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ total_minutes_asleep       <int> 327, 384, 412, 340, 700, 304, 360, 325, 361…
## $ total_time_in_bed          <int> 346, 407, 442, 367, 712, 320, 377, 364, 384…
## $ total_time_awake_in_bed    <int> 19, 23, 30, 27, 12, 16, 17, 39, 23, 19, 46,…
```

```
cor(x=daily_without_na$total_time_awake_in_bed, y=daily_without_na$calories, method="pearson")
```

```
## [1] -0.2873181
```

```
ggplot(daily_without_na, mapping=aes(total_time_awake_in_bed, calories))+geom_point(colour="pur
ple")+facet_wrap(~total_sleep_records) +geom_smooth(colour="orange")+ theme(panel.background=el
ement_blank(), axis.line=element_line(colour="black"))
```

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

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : span too small. fewer data values than degrees of freedom.
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 24.88
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 18.12
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 0
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 37.454
```

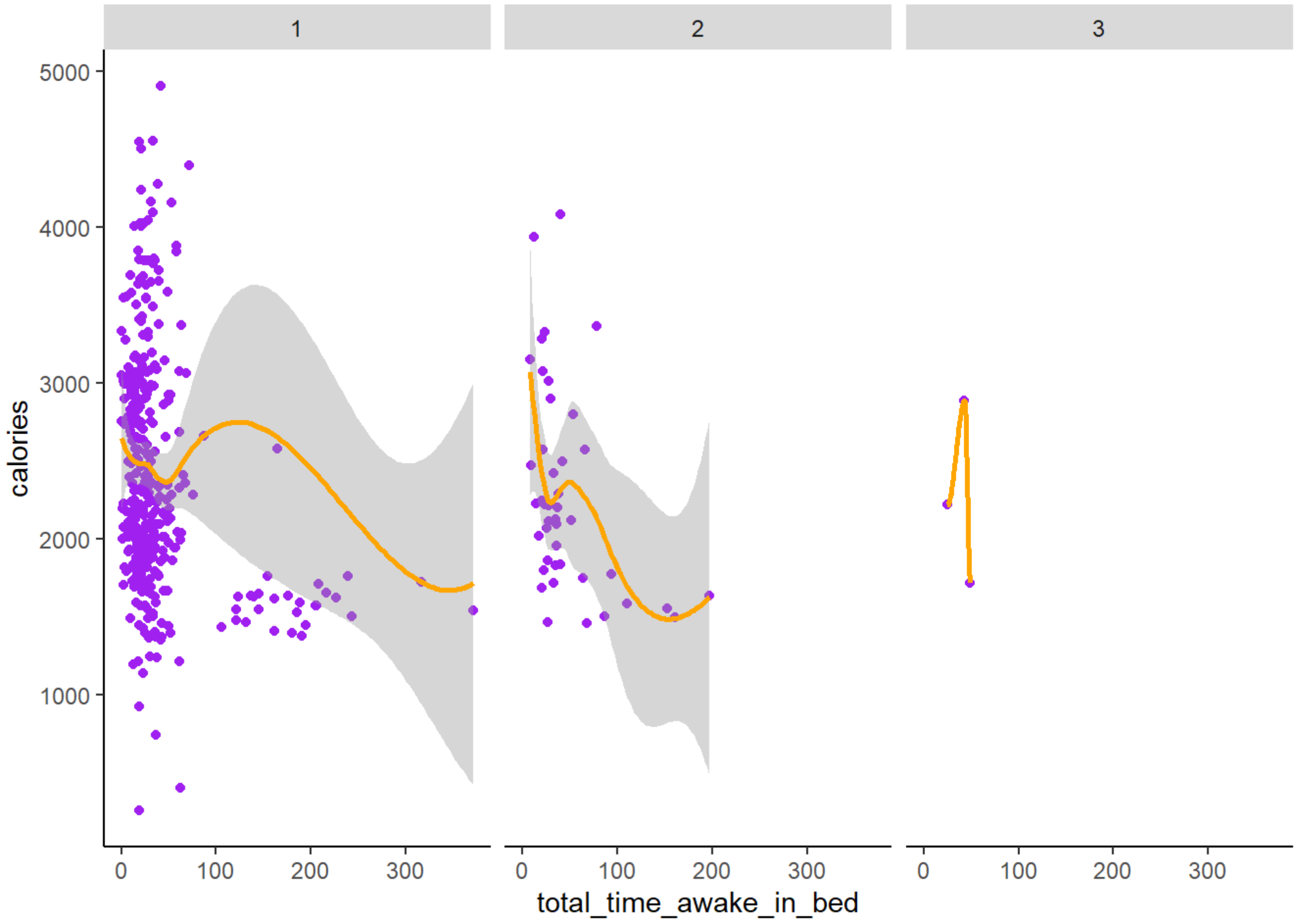
```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : span too small. fewer
## data values than degrees of freedom.
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used at
## 24.88
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood radius
## 18.12
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : reciprocal condition
## number 0
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : There are other near
## singularities as well. 37.454
```

```
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf
```



The analysis is depicting **slightly negative correlation between energy expenditure throughout the day and sleeplessness while in bed**. This means that an increased period of activity is associated with less time spend in bed before finally sleeping.

Let us also analyse **the relationship between sedantry minutes and sleep duration**

```
cor(x=daily_without_na$total_minutes_asleep, y=daily_without_na$sedentary_minutes, method="pear
son")
```

```
## [1] -0.6010731
```

```
#plotting the same
ggplot(daily_without_na, mapping=aes(total_minutes_asleep,sedentary_minutes))+geom_point(colour
="purple")+
  facet_wrap(~total_sleep_records)+geom_smooth(colour="orange")+theme(panel.background=element_
blank(), axis.line=element_line(colour="black"))
```

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

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : span too small. fewer data values than degrees of freedom.
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at 551.01
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 78.99
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 0
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 14639
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : span too small. fewer
## data values than degrees of freedom.
```

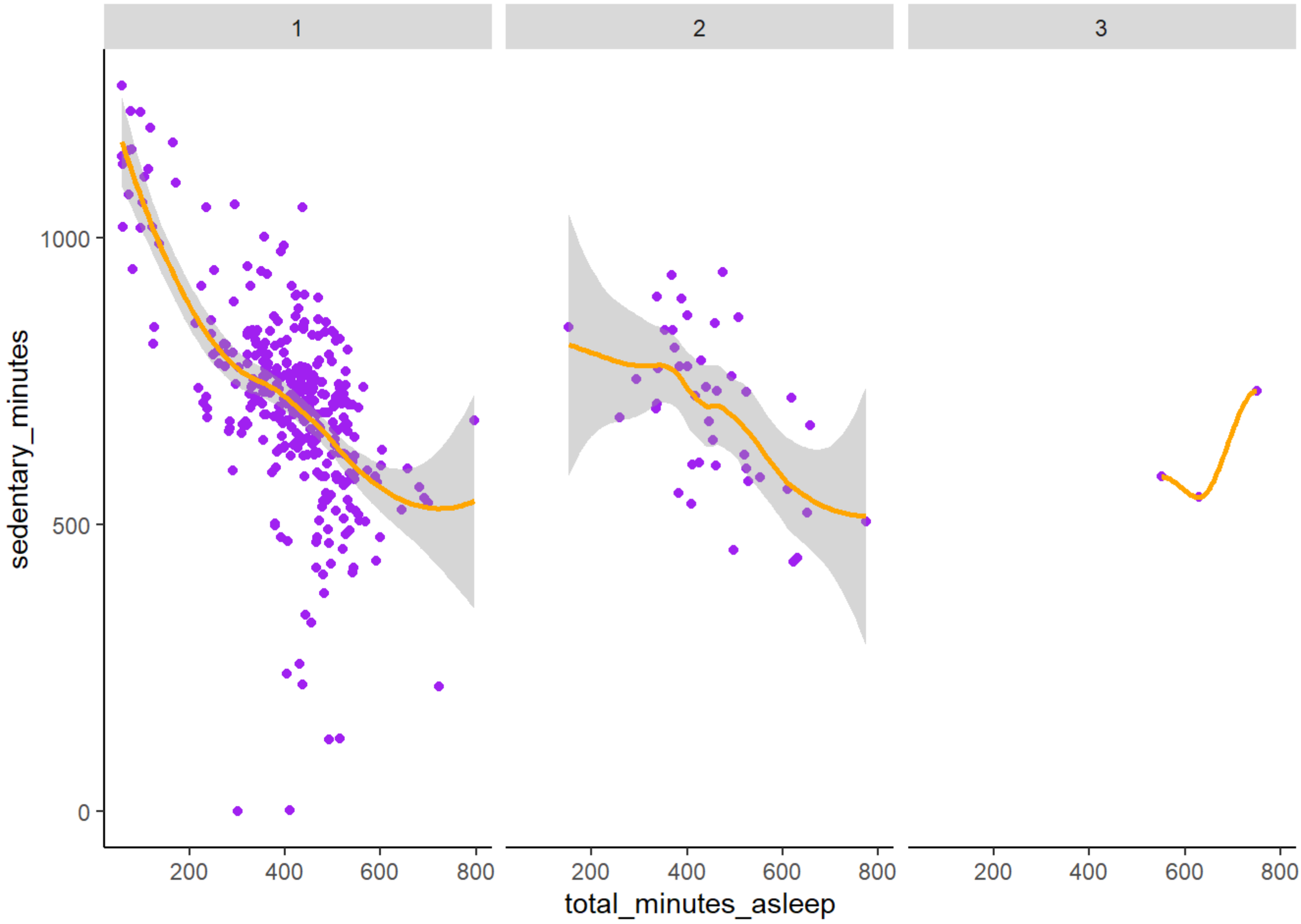
```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used at
## 551.01
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood radius
## 78.99
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : reciprocal condition
## number 0
```

```
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : There are other near
## singularities as well. 14639
```

```
## Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
## -Inf
```



As seen in both of the above analysis:

Having increased activity throughout the day is related to **less awake time before sleeping**.

Likewise increased periods of inactivity(sedantary minutes) is associated with poor amount of sleep.

These insights can be used to **market the Bellabeat for saliant feature of improved sleep with regular physical activity**.

**-> Share Phase : Sharing Recommendations**

**Summary of our Target Audience:**

Bellabeat's marketing team needs to focus their marketing towards the user segment that are:

1.Working adults that mostly does a routine(9-5) desk jobs. These users indulge in some light_activity to maintain their health but they definitely need some motivation **to increase their activity levels to reap maximum health benefits**.

2.These users also have a **sleepless phase of 20-30 minutes in bed** before they are finally able sleep.

3.They remain active throughout the **workdays(Mon-Thurs)** but considers relaxing on the weekends.

**Recommendations for the data**

1.The sample size of the data is too small, it needs to be expanded to draw any strong conclusions.

2.The number of participants should be equal for all the parameters, lack of it presented many difficulties while analysis. Some parameters have really few participants(weight logs and heart rate per second) which makes them unsuitable for analysis.

3.The data for the of values for VeryActiveDistance, ModeratelyActiveDistance and LightActiveDistance needs to be reassesed as their sum does not add upto the total distance for daily activity, this renders any analysis that can be performed on sum of these parameters as inaccurate.

Also, some information on the demographics of the users such as gender, age, and height , would provide deep insights for developing strategies, keeping in mind Bellabeat's vision of products curation especially for women.

**Recommendations for the app** Along with tracking the active movements, Bellabeat can have enhanced features such as a water log so that users can track their hydration status and maintain their overall health.

The app can also include sublte features of notifications or alarms for sleep schedules or going early to bed to enhance the user’s overall experience.

Lastly, the users have a number of days when no activity has been logged. One possibility is that the app did not register other forms of physical activity like biking, swimming, or playing a sport, muscle strengthning exercises etc. The app should include metrics to register these activities as well to provide a wholesome user experience.

**-> Act Phase : Future marketting startegy**

Most users struggle to remain highly active thrughout the day(Overall, the duration of “light active minutes” is much higher than “very active minutes”). This can be used to market some high intensity, short duration workouts so that the customers reap more health benefits.

There were very few users who logged in their weight details every day because it is a manual task. Bellabeat can utilize this information to promote features like weight log notifications and even daily alarms, reminding the users about the scheduled time for their daily physical activity.

Since the app generates a lot of health data, this information can be leveraged to develop and sell personalised goals and activity suggestions.

Finally, Many Participants experience a sleepless period of about 20-30 minutes in bed, Bellabeat can develop and promote sleep assisting features such as sleep inducing music, sleep journals, etc.

**Final note** The analysis strategies have been seen from some data analyst’s methods and tried in those ways. Thanks for going through the steps.