BELLABEAT CASE STUDY

Introduction

Bellabeat, is a high-tech manufacturer of health-focused products for women, and meet different characters and team members. Bellabeat, believes that analyzing smart device fitness data could help unlock new growth opportunities for the company. Marketing analytics team of Bellabeat has been asked to focus on one of Bellabeat's products and analyze smart device data to gain insight into how consumers are using their smart devices.

Business task

Analysis of how are smart devices being used by customers and how these trends can be used for future marketing strategy of Bellabeat.

Tast could be divided into 2 parts:

1. How are devices being used by customers?
2. What improvments could be implemented to increase their usage?

Data

Data used for analysis is stored on Kaggle. It is a public domain dataset made available through Mobius. (https://www.kaggle.com/arashnic/fitbit)

The dataset is:

- open-source,
- verified by it's metadata.

This dataset is generated by respondents to a distributed survey via Amazon Mechanical Turk between 03.12.2016-05.12.2016. It includes 33 persons, however in case study scenario is written 30 participants took part in the survey.

Data are from 2016 – i.e. 2 years old - thus it might be considered as out of date and irrelevant, but the purpose of the case study we will consider it as relevant data.

Data are stored in zip file and contains 18 csv files in both long and wide format.

Data Organization Process

For data processing I have used RStudio, Google Sheets and Tableau.

Works done in RStudio

First part of work is done in RStudio and I have started with packages installation.

```
install.packages("tidyverse")
install.packages("ggplot2")
install.packages("dplyr")
install.packages("skimr")
install.packages("janitor")

library(tidyverse)
library(ggplot2)
library(dplyr)
library(skimr)
library(janitor)
```

```
> install.packages("tidyverse")
Installing package into ‘/cloud/lib/x86_64-pc-linux-gnu-library/4.1’
(as ‘lib’ is unspecified)
trying URL 'http://rspm/default/__linux__/focal/latest/src/contrib/tidyverse_1
.3.1.tar.gz'
Content type 'application/x-gzip' length 424699 bytes (414 KB)
==================================================
downloaded 414 KB

* installing *binary* package ‘tidyverse’ ...
* DONE (tidyverse)

The downloaded source packages are in
        ‘/tmp/RtmpASiFn9/downloaded_packages’
> install.packages("ggplot2")
Installing package into ‘/cloud/lib/x86_64-pc-linux-gnu-library/4.1’
(as ‘lib’ is unspecified)
trying URL 'http://rspm/default/__linux__/focal/latest/src/contrib/ggplot2_3.3
.5.tar.gz'
Content type 'application/x-gzip' length 4113418 bytes (3.9 MB)
==================================================
downloaded 3.9 MB

* installing *binary* package ‘ggplot2’ ...
* DONE (ggplot2)

The downloaded source packages are in
        ‘/tmp/RtmpASiFn9/downloaded_packages’
> install.packages("dplyr")
Installing package into ‘/cloud/lib/x86_64-pc-linux-gnu-library/4.1’
(as ‘lib’ is unspecified)
trying URL 'http://rspm/default/__linux__/focal/latest/src/contrib/dplyr_1.0.7
.tar.gz'
Content type 'application/x-gzip' length 1251246 bytes (1.2 MB)
==================================================
downloaded 1.2 MB

* installing *binary* package ‘dplyr’ ...
* DONE (dplyr)
```

```
The downloaded source packages are in
        ‘/tmp/RtmpASiFn9/downloaded_packages’
> install.packages("skimr")
Installing package into ‘/cloud/lib/x86_64-pc-linux-gnu-library/4.1’
(as ‘lib’ is unspecified)
trying URL 'http://rspm/default/__linux__/focal/latest/src/contrib/skimr_2.1.3
.tar.gz'
Content type 'application/x-gzip' length 1224706 bytes (1.2 MB)
==================================================
downloaded 1.2 MB

* installing *binary* package ‘skimr’ ...
* DONE (skimr)

The downloaded source packages are in
        ‘/tmp/RtmpASiFn9/downloaded_packages’
> install.packages("janitor")
Installing package into ‘/cloud/lib/x86_64-pc-linux-gnu-library/4.1’
(as ‘lib’ is unspecified)
trying URL 'http://rspm/default/__linux__/focal/latest/src/contrib/janitor_2.1
.0.tar.gz'
Content type 'application/x-gzip' length 247530 bytes (241 KB)
==================================================
downloaded 241 KB

* installing *binary* package ‘janitor’ ...
* DONE (janitor)

The downloaded source packages are in
        ‘/tmp/RtmpASiFn9/downloaded_packages’
>
> library(tidyverse)
── Attaching packages ─────────────────────────────────────── tidyverse 1
.3.1 ──
✓ ggplot2 3.3.5     ✓ purrr   0.3.4
✓ tibble  3.1.6     ✓ dplyr   1.0.7
✓ tidyr   1.1.4     ✓ stringr 1.4.0
✓ readr   2.1.1     ✓ forcats 0.5.1
── Conflicts ────────────────────────────────────────── tidyverse_conflic
ts() ──
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
> library(ggplot2)
> library(dplyr)
> library(skimr)
> library(janitor)

Attaching package: ‘janitor’

The following objects are masked from ‘package:stats’:

    chisq.test, fisher.test
```

Afterwards csv files have been uploaded. Following files have been chosen for my analysis:

daily_activity_2 <-read_csv("dailyActivity_merged.csv")
daily_calories <-read_csv("dailyCalories_merged.csv")
daily_intensities <-read_csv("dailyIntensities_merged.csv")
daily_steps <-read_csv("dailySteps_merged.csv")
daily_sleep <-read_csv("sleepDay_merged.csv")

Data review and cleaning

```
> head(daily_activity_2)
# A tibble: 6 × 15
          Id ActivityDate TotalSteps TotalDistance TrackerDistance LoggedActiv
itiesD…
       <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(daily_calories)
# A tibble: 6 × 3
          Id ActivityDay Calories
       <dbl> <chr>          <dbl>
1 1503960366 4/12/2016       1985
2 1503960366 4/13/2016       1797
3 1503960366 4/14/2016       1776
4 1503960366 4/15/2016       1745
5 1503960366 4/16/2016       1863
6 1503960366 4/17/2016       1728
```

```
> head(daily_intensities)
# A tibble: 6 × 10
          Id ActivityDay SedentaryMinutes LightlyActiveMinutes FairlyActiveMin
utes
```

```
       <dbl> <chr>                  <dbl>                <dbl>                <
dbl>
1 1503960366 4/12/2016                728                  328
13
2 1503960366 4/13/2016                776                  217
19
3 1503960366 4/14/2016               1218                  181
11
4 1503960366 4/15/2016                726                  209
34
5 1503960366 4/16/2016                773                  221
10
6 1503960366 4/17/2016                539                  164
20
# … with 5 more variables: VeryActiveMinutes <dbl>, SedentaryActiveDistance <d
bl>,
#   LightActiveDistance <dbl>, ModeratelyActiveDistance <dbl>,
#   VeryActiveDistance <dbl>


> head(daily_sleep)
# A tibble: 6 × 5
          Id SleepDay              TotalSleepRecords TotalMinutesAsle… TotalTi
meInBed
       <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(daily_steps)
# A tibble: 6 × 3
          Id ActivityDay StepTotal
       <dbl> <chr>           <dbl>
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

> glimpse(daily_activity_2)
Rows: 940
Columns: 15
$ Id                       <dbl> 1503960366, 1503960366, 1503960366, 150396036
6, 15…
```

```
$ ActivityDate             <chr> "4/12/2016", "4/13/2016", "4/14/2016", "4/15/
2016"…
$ TotalSteps               <dbl> 13162, 10735, 10460, 9762, 12669, 9705, 13019
, 155…
$ TotalDistance            <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9.8
8, 6.…
$ TrackerDistance          <dbl> 8.50, 6.97, 6.74, 6.28, 8.16, 6.48, 8.59, 9.8
8, 6.…
$ LoggedActivitiesDistance <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
$ VeryActiveDistance       <dbl> 1.88, 1.57, 2.44, 2.14, 2.71, 3.19, 3.25, 3.5
3, 1.…
$ ModeratelyActiveDistance <dbl> 0.55, 0.69, 0.40, 1.26, 0.41, 0.78, 0.64, 1.3
2, 0.…
$ LightActiveDistance      <dbl> 6.06, 4.71, 3.91, 2.83, 5.04, 2.51, 4.71, 5.0
3, 4.…
$ SedentaryActiveDistance  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
$ VeryActiveMinutes        <dbl> 25, 21, 30, 29, 36, 38, 42, 50, 28, 19, 66, 4
1, 39…
$ FairlyActiveMinutes      <dbl> 13, 19, 11, 34, 10, 20, 16, 31, 12, 8, 27, 21
, 5, …
$ LightlyActiveMinutes     <dbl> 328, 217, 181, 209, 221, 164, 233, 264, 205,
211, …
$ SedentaryMinutes         <dbl> 728, 776, 1218, 726, 773, 539, 1149, 775, 818
, 838…
$ Calories                 <dbl> 1985, 1797, 1776, 1745, 1863, 1728, 1921, 203
5, 17…

> glimpse(daily_calories)
Rows: 940
Columns: 3
$ Id          <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 1503960366
, 150…
$ ActivityDay <chr> "4/12/2016", "4/13/2016", "4/14/2016", "4/15/2016", "4/16/
2016"…
$ Calories    <dbl> 1985, 1797, 1776, 1745, 1863, 1728, 1921, 2035, 1786, 1775
, 182…

> glimpse(daily_intensities)
Rows: 940
Columns: 10
$ Id                       <dbl> 1503960366, 1503960366, 1503960366, 150396036
6, 15…
$ ActivityDay              <chr> "4/12/2016", "4/13/2016", "4/14/2016", "4/15/
2016"…
$ SedentaryMinutes         <dbl> 728, 776, 1218, 726, 773, 539, 1149, 775, 818
, 838…
$ LightlyActiveMinutes     <dbl> 328, 217, 181, 209, 221, 164, 233, 264, 205,
211, …
$ FairlyActiveMinutes      <dbl> 13, 19, 11, 34, 10, 20, 16, 31, 12, 8, 27, 21
, 5, …
$ VeryActiveMinutes        <dbl> 25, 21, 30, 29, 36, 38, 42, 50, 28, 19, 66, 4
1, 39…
$ SedentaryActiveDistance  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
```

```
$ LightActiveDistance      <dbl> 6.06, 4.71, 3.91, 2.83, 5.04, 2.51, 4.71, 5.0
3, 4.…
$ ModeratelyActiveDistance <dbl> 0.55, 0.69, 0.40, 1.26, 0.41, 0.78, 0.64, 1.3
2, 0.…
$ VeryActiveDistance       <dbl> 1.88, 1.57, 2.44, 2.14, 2.71, 3.19, 3.25, 3.5
3, 1.…
```

```
> glimpse(daily_sleep)
Rows: 413
Columns: 5
$ Id                 <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 150
39603…
$ SleepDay           <chr> "4/12/2016 12:00:00 AM", "4/13/2016 12:00:00 AM", "
4/15/…
$ TotalSleepRecords  <dbl> 1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1,…
$ TotalMinutesAsleep <dbl> 327, 384, 412, 340, 700, 304, 360, 325, 361, 430, 2
77, 2…
$ TotalTimeInBed     <dbl> 346, 407, 442, 367, 712, 320, 377, 364, 384, 449, 3
23, 2…
```

```
> glimpse(daily_steps)
Rows: 940
Columns: 3
$ Id          <dbl> 1503960366, 1503960366, 1503960366, 1503960366, 1503960366
, 150…
$ ActivityDay <chr> "4/12/2016", "4/13/2016", "4/14/2016", "4/15/2016", "4/16/
2016"…
$ StepTotal   <dbl> 13162, 10735, 10460, 9762, 12669, 9705, 13019, 15506, 1054
4, 98…
```

Using glimpse function I have identified that files used for analysis contains 940 rows except of one file **daily_sleeps** which contains only 413 rows.

Further I have checked whether the files includes data for 33 participants to ensure about completness of data. For this purpose n_distinct function was used.

```
> n_distinct(daily_activity$Id)
[1] 940
> n_distinct(daily_calories$Id)
[1] 33
> n_distinct(daily_intensities$Id)
[1] 33
> n_distinct(daily_sleep$Id)
[1] 24
> n_distinct(daily_steps$Id)
[1] 33
```

This check proved that daily_sleep file contains data only for 24 participants. Thus those data are not complete and could lead to misleading conclusions.

```
> clean_names(daily_activity_2)
> clean_names(daily_calories)
> clean_names(daily_intensities)
> clean_names(daily_sleep)
```

```
> clean_names(daily_steps)

> clean_names(daily_activity_2)
# A tibble: 940 × 15
        id activity_date total_steps total_distance tracker_distance logged_ac
tiviti…
     <dbl> <chr>               <dbl>          <dbl>            <dbl>
<dbl>
 1  1.50e9 4/12/2016           13162           8.5              8.5
0
 2  1.50e9 4/13/2016           10735           6.97             6.97
0
 3  1.50e9 4/14/2016           10460           6.74             6.74
0
 4  1.50e9 4/15/2016            9762           6.28             6.28
0
 5  1.50e9 4/16/2016           12669           8.16             8.16
0
 6  1.50e9 4/17/2016            9705           6.48             6.48
0
 7  1.50e9 4/18/2016           13019           8.59             8.59
0
 8  1.50e9 4/19/2016           15506           9.88             9.88
0
 9  1.50e9 4/20/2016           10544           6.68             6.68
0
10  1.50e9 4/21/2016            9819           6.34             6.34
0
# … with 930 more rows, and 9 more variables: very_active_distance <dbl>,
#   moderately_active_distance <dbl>, light_active_distance <dbl>,
#   sedentary_active_distance <dbl>, very_active_minutes <dbl>,
#   fairly_active_minutes <dbl>, lightly_active_minutes <dbl>,
#   sedentary_minutes <dbl>, calories <dbl>


> clean_names(daily_calories)
# A tibble: 940 × 3
           id activity_day calories
        <dbl> <chr>           <dbl>
 1 1503960366 4/12/2016        1985
 2 1503960366 4/13/2016        1797
 3 1503960366 4/14/2016        1776
 4 1503960366 4/15/2016        1745
 5 1503960366 4/16/2016        1863
 6 1503960366 4/17/2016        1728
 7 1503960366 4/18/2016        1921
 8 1503960366 4/19/2016        2035
 9 1503960366 4/20/2016        1786
10 1503960366 4/21/2016        1775
# … with 930 more rows

> clean_names(daily_intensities)
# A tibble: 940 × 10
           id activity_day sedentary_minutes lightly_active_minutes fairly_act
ive_mi…
```

```
        <dbl> <chr>                    <dbl>                  <dbl>
<dbl>
 1 1503960366 4/12/2016                 728                    328
13
 2 1503960366 4/13/2016                 776                    217
19
 3 1503960366 4/14/2016                1218                    181
11
 4 1503960366 4/15/2016                 726                    209
34
 5 1503960366 4/16/2016                 773                    221
10
 6 1503960366 4/17/2016                 539                    164
20
 7 1503960366 4/18/2016                1149                    233
16
 8 1503960366 4/19/2016                 775                    264
31
 9 1503960366 4/20/2016                 818                    205
12
10 1503960366 4/21/2016                 838                    211
8
# … with 930 more rows, and 5 more variables: very_active_minutes <dbl>,
#   sedentary_active_distance <dbl>, light_active_distance <dbl>,
#   moderately_active_distance <dbl>, very_active_distance <dbl>

> clean_names(daily_sleep)
# A tibble: 413 × 5
           id sleep_day        total_sleep_recor… total_minutes_asl… total_tim
e_in_b…
        <dbl> <chr>                         <dbl>              <dbl>
<dbl>
 1 1503960366 4/12/2016 12:00…                 1                327
346
 2 1503960366 4/13/2016 12:00…                 2                384
407
 3 1503960366 4/15/2016 12:00…                 1                412
442
 4 1503960366 4/16/2016 12:00…                 2                340
367
 5 1503960366 4/17/2016 12:00…                 1                700
712
 6 1503960366 4/19/2016 12:00…                 1                304
320
 7 1503960366 4/20/2016 12:00…                 1                360
377
 8 1503960366 4/21/2016 12:00…                 1                325
364
 9 1503960366 4/23/2016 12:00…                 1                361
384
10 1503960366 4/24/2016 12:00…                 1                430
449
# … with 403 more rows

> clean_names(daily_steps)
# A tibble: 940 × 3
```

```
          id activity_day step_total
       <dbl> <chr>             <dbl>
 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
 7 1503960366 4/18/2016         13019
 8 1503960366 4/19/2016         15506
 9 1503960366 4/20/2016         10544
10 1503960366 4/21/2016          9819
# … with 930 more rows
```

Works done in Google Sheets

File used and analyzed was dailyActivity_merged.csv, this was then saved as daily_activity.xlsx and sleepDay_merged.csv saved as sleep_day_v1.xlsx.

daily_activity.xlsx

Id	ActivityDate	TotalDistance	Tracker Distance	LoggedActivitiesDistance	VeryActiveDistance	ModeratelyActiveDistance	LightActiveDistance	Sedentary ActiveDistance	VeryActiveMinutes	FairlyActiveMinutes	LightlyActiveMinutes	SedentaryMinutes	Calories
1503960366	4.12.2016	8.5	8.5	0	1.879999995231 63	0.550000011920929	6.059999942779 54	0	25	13	328	728	1985
1503960366	4.13.2016	6.969999790191 65	6.969999790191 65	0	1.570000052452 09	0.689999997615814	4.710000038146 97	0	21	19	217	776	1797
1503960366	4.14.2016	6.739999771118 16	6.739999771118 16	0	2.440000057220 46	0.400000005960464	3.910000085830 69	0	30	11	181	1218	1776

Data have been sorted based on Id column in ascending order.
Column ActivityDate have been formated to Date (mm.dd.yyyy).
Column Day No. Have been added to cound how many days have participants took part in the survey.

Id	ActivityDate	Day No.	TotalSteps	TotalDistance	Tracker Distance	LoggedActivitiesDistance	VeryActiveDistance	ModeratelyActiveDistance	LightActiveDistance	Sedentary ActiveDistance	VeryActiveMinutes	FairlyActiveMinutes	LightlyActiveMinutes	SedentaryMinutes	Calories
1503960366	4.12.2016	1	13162	8.5	8.5	0	1.879999995231 63	0.550000011920929	6.059999942779 54	0	25	13	328	728	1985
1503960366	4.13.2016	2	10735	6.969999790191 65	6.969999790191 65	0	1.570000052452 09	0.689999997615814	4.710000038146 97	0	21	19	217	776	1797
1503960366	4.14.2016	3	10460	6.739999771118 16	6.739999771118 16	0	2.440000057220 46	0.400000005960464	3.910000085830 69	0	30	11	181	1218	1776

sleep_day_v1.xlsx

Id	SleepDay	TotalSleepRecords	TotalMinutesAsleep	TotalTimeInBed
1503960366	4/12/16 0:00	1	327	346
1503960366	4.13.2016 12:00:00 AM	2	384	407
1503960366	4.15.2016 12:00:00 AM	1	412	442
1503960366	4.16.2016 12:00:00 AM	2	340	367
1503960366	4.17.2016 12:00:00 AM	1	700	712

Report of modifications

1. formating date column Sleep Day, slash replaced by point
2. added new column Day No.
3. removed line 382 - duplication identified

8378563200	4/25/2016 12:00:00 AM	1	388	402	14
8378563200	4/25/2016 12:00:00 AM	1	388	402	15

Id	SleepDay	TotalSleepRecords	TotalMinutesAsleep	TotalTimeInBed	Day No.
1503960366	4.12.2016 0:00:00	1	327	346	1
1503960366	4.13.2016 12:00:00 AM	2	384	407	2
1503960366	4.15.2016 12:00:00 AM	1	412	442	3
1503960366	4.16.2016 12:00:00 AM	2	340	367	4
1503960366	4.17.2016 12:00:00 AM	1	700	712	5

Analyze data

Maximum, minimum and average values have been reviewed in the following files – daily_calories, daily_steps and daily_sleep.

```
> max(daily_calories$Calories)
[1] 4900
> min(daily_calories$Calories)
[1] 0
> mean(daily_calories$Calories)
[1] 2303.61
>
> max(daily_steps$StepTotal)
[1] 36019
> min(daily_steps$StepTotal)
[1] 0
> mean(daily_steps$StepTotal)
[1] 7637.911
>
> max(daily_sleep$TotalMinutesAsleep)
[1] 796
> min(daily_sleep$TotalMinutesAsleep)
[1] 58
> mean(daily_sleep$TotalMinutesAsleep)
[1] 419.4673
```

```
> daily_activity_2 %>%
+   group_by(Id) %>%
+   drop_na() %>%
+   summarize(max_total_distance=max(TotalDistance),min_total_distance=min(Tot
alDistance),mean_total_distance=mean(TotalDistance))
# A tibble: 33 × 4
           Id max_total_distance min_total_distance mean_total_distance
        <dbl>              <dbl>              <dbl>               <dbl>
 1 1503960366              12.2               0                   7.81
 2 1624580081              28.0               0.980               3.91
 3 1644430081              13.2               0.890               5.30
 4 1844505072               5.32              0                   1.71
 5 1927972279               2.62              0                   0.635
 6 2022484408              12.9               2.31                8.08
 7 2026352035               7.71              0.160               3.45
 8 2320127002               7.49              0.520               3.19
 9 2347167796              15.1               0.0300              6.36
10 2873212765               6.65              1.70                5.10
# … with 23 more rows


> daily_calories %>%
+   group_by(Id) %>%
+   drop_na() %>%
+   summarize(max_calories=max(Calories),min_calories=min(Calories),mean_calor
ies = mean(Calories))
# A tibble: 33 × 4
           Id max_calories min_calories mean_calories
        <dbl>        <dbl>        <dbl>         <dbl>
 1 1503960366         2159            0         1816.
 2 1624580081         2690         1002         1483.
 3 1644430081         3846         1276         2811.
 4 1844505072         2130          665         1573.
 5 1927972279         2638         1383         2173.
 6 2022484408         3158         1848         2510.
 7 2026352035         1926         1141         1541.
 8 2320127002         2124         1125         1724.
 9 2347167796         2670          403         2043.
10 2873212765         2241         1431         1917.
# … with 23 more rows


> daily_steps %>%
+   group_by(Id) %>%
+   drop_na() %>%
+   summarize(max_steps=max(StepTotal),min_steps=min(StepTotal),mean_steps = m
ean(StepTotal))
# A tibble: 33 × 4
           Id max_steps min_steps mean_steps
        <dbl>     <dbl>     <dbl>      <dbl>
 1 1503960366     18134         0     12117.
 2 1624580081     36019      1510      5744.
 3 1644430081     18213      1223      7283.
 4 1844505072      8054         0      2580.
```

```
 5 1927972279      3790         0       916.
 6 2022484408     18387      3292     11371.
 7 2026352035     12357       254      5567.
 8 2320127002     10725       772      4717.
 9 2347167796     22244        42      9520.
10 2873212765      9685      2524      7556.
# … with 23 more rows
```

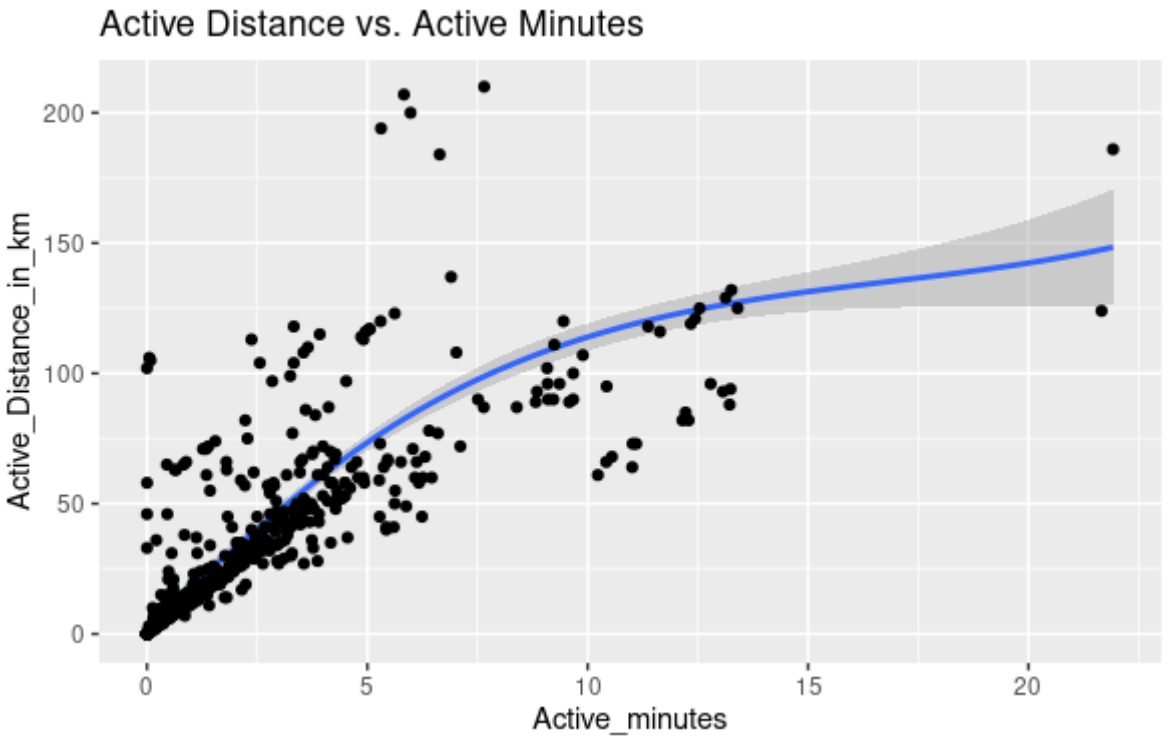
```
> daily_sleep %>%
+   group_by(Id) %>%
+   drop_na() %>%
+   summarize(max_sleep_time=max(TotalMinutesAsleep),min_sleep_time=min(TotalM
inutesAsleep),mean_sleep_time = mean(TotalMinutesAsleep))
# A tibble: 24 × 4
           Id max_sleep_time min_sleep_time mean_sleep_time
        <dbl>          <dbl>          <dbl>           <dbl>
 1 1503960366            700            245            360.
 2 1644430081            796            119            294
 3 1844505072            722            590            652
 4 1927972279            750            166            417
 5 2026352035            573            357            506.
 6 2320127002             61             61             61
 7 2347167796            556            374            447.
 8 3977333714            424            152            294.
 9 4020332650            501             77            349.
10 4319703577            692             59            477.
# … with 14 more rows
```

Data Visualisation

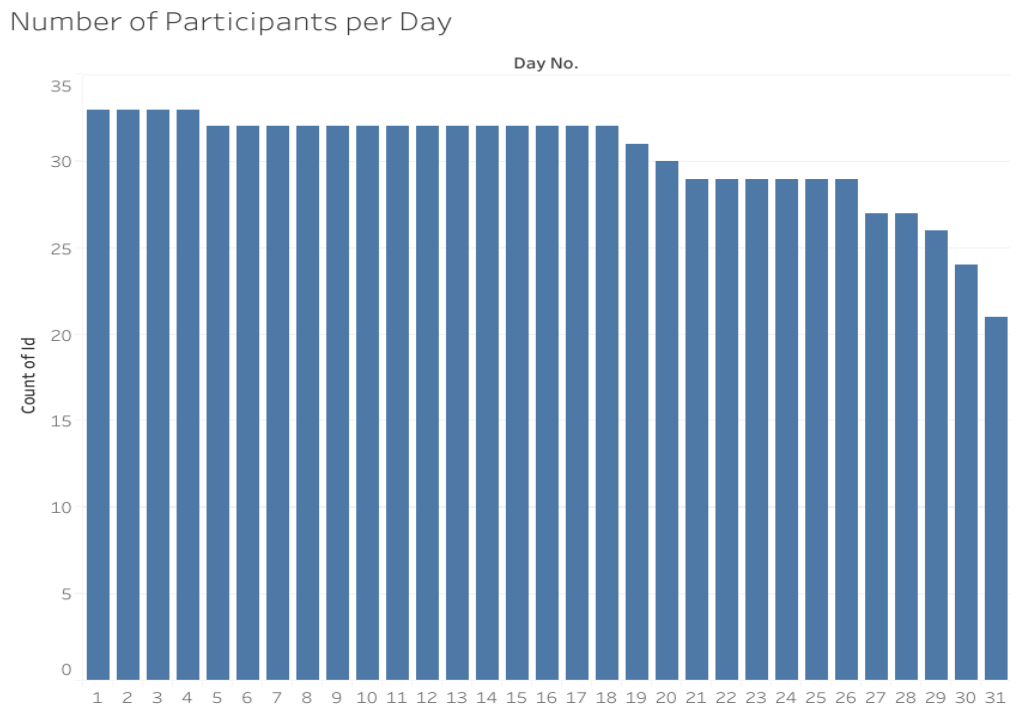
I have focused on different relationships between data.

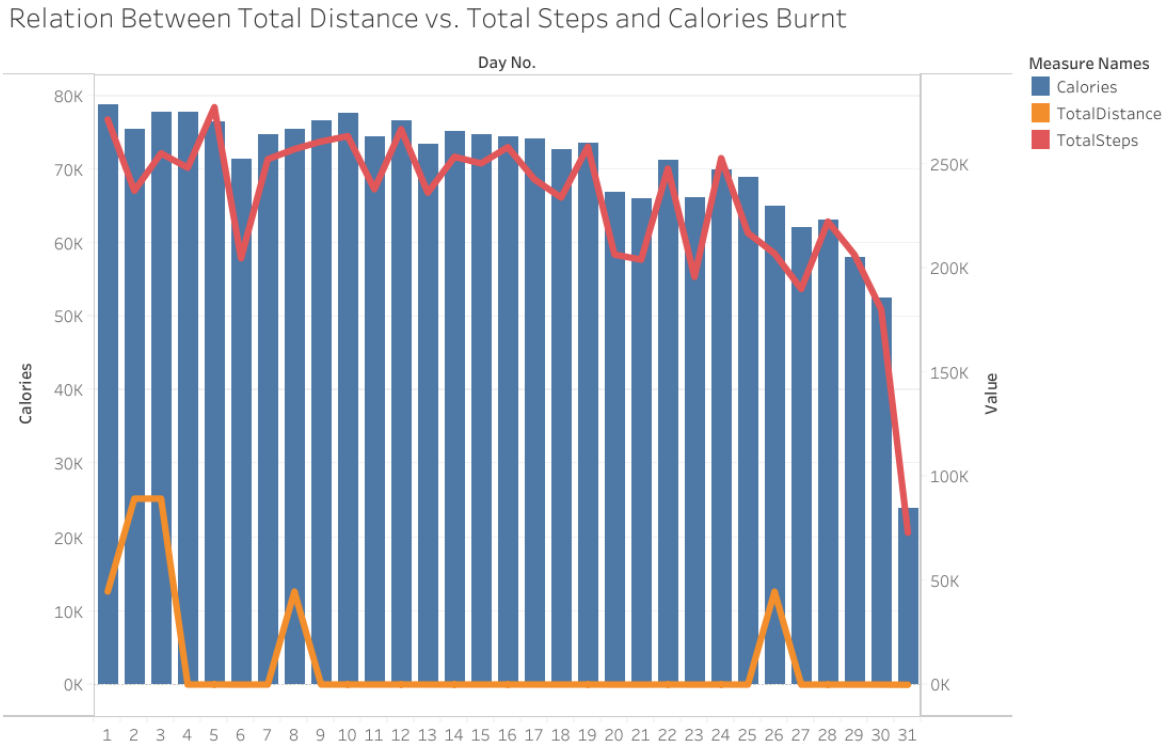
```
> ggplot(data=daily_activity_2)+
+   geom_smooth(mapping=aes(x=Calories,y=TotalSteps))+
+   geom_point(mapping=aes(x=Calories,y=TotalSteps),color="green")+
+   labs(title="Relation Between Steps Made and Calories Burnt",y="Total_Steps
")
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

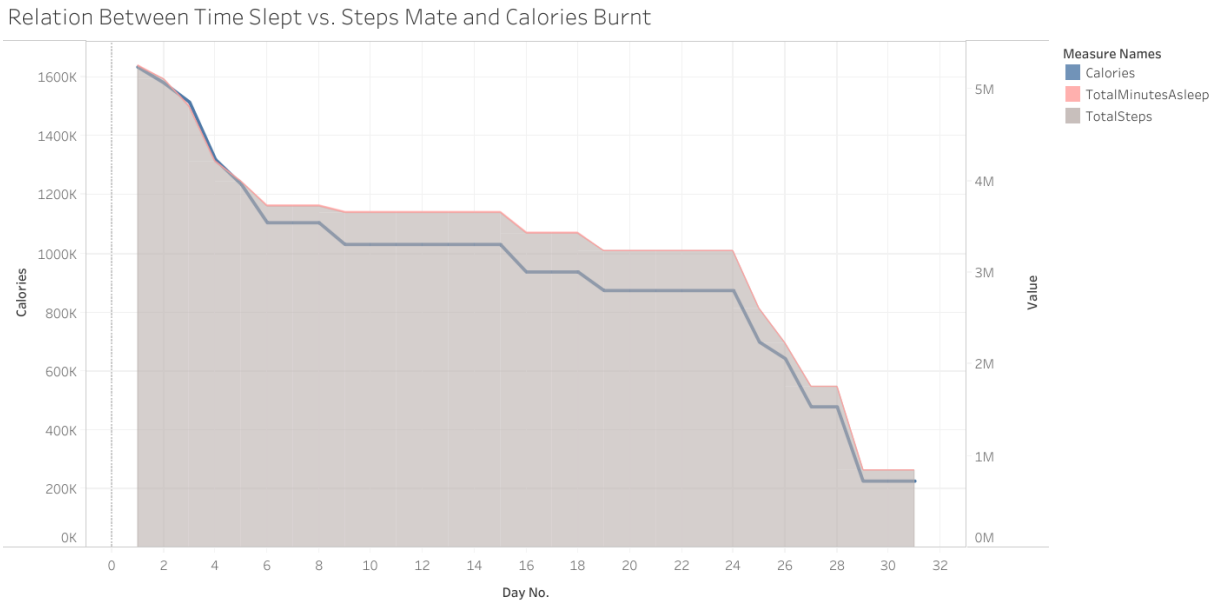

```
> ggplot(data=daily_activity_2)+
+   geom_smooth(mapping=aes(x=VeryActiveDistance,y=VeryActiveMinutes))+
+   geom_point(mapping=aes(x=VeryActiveDistance,y=VeryActiveMinutes),color="gr
een")+
+   labs(title="Active Distance vs. Active Minutes",y="Active_Distance_in_km",
x="Active_minutes")
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Visualisation prepared in Tableau







Findings

Based on analysis performed following trends have been identified:

1. in average the more steps/distance was made the more calories were burnt
2. in general participants were not really active as the Active Distance vs. Active minutes chart shows as the spots are concentrated in the left down corner

There were identified also some limitations:

1. sampling bias – survey included only 33 participants this and there was no information about age, location, etc.
2. 33 participants is not large enough to perform an analysis and base decision on such analysis. Even more if there are available only data from 21 participants for the whole period of survey.

Recommendations

1. Focus on highlighting of calories burnt in marketing campaings because each woman is interested in losing weight and thus monitoring how many calories they burnt
2. Implementation of app which would remind the users to switch on their device when going to sleep because only 24 participants tracked their sleep and only 3 of them track their sleep for the whole period of survey.