Google Data Analytics Project Bellabeat Case Study

Clear Summary of Business Task

Company Background

Bellabeat (founded in 2013) is a tech-driven wellness company that produces smart products related to health. The company collects data on health activity and launched multiple products globally. The company has invested on both traditional advertising media and internet social media pages.

Task:

Analyze non-Bellabeat smart device usage data and gain insights of how customers are using the application that are not from Bellabeat, trends and insights from doing so

Problem to Solve

Bellabeat needs useful information for its marketing strategy and grow its potential in the global smart device market. Identifying how consumers are using software and smart device as their health activity is crucial as an opportunity to improve sales, brand awareness, and positive outcomes for user's health.

Key Stakeholder

Urška Sršen as the cofounder and Chief Creative Officer. She will approve and make decisions based on analysis results along with the executive team.

Sando Mur is also one of Bellabeat's cofounder and a Mathematician. As a key member of Bellabeat executive team, he will help Urška with the executive decisions the marketing strategy. Any findings in this task are directly reported to the executive team.

Bellabeat marketing analytics team will provide support and guidance necessary to complete the task including collecting, analyzing and reporting the data for Bellabeat's marketing strategy.

Description of all Data Sources Used

Data is publicly available on Kaggle website and stored with 18 csv files. Data is generated by 30 eligible and consenting Fitbit users by a third-party company called Amazon Mechanical Turk between the date of March 12-May 12 2016 (2 months duration).

Data is organized as long format. There are multiple user ids in a single column.

Limitation of this dataset (ROCCC):

- o Reliable – The data doesn't show high reliability since only 30 respondents are selected which indicates lower confidence level. Moreover, there are no information regarding the gender, occupation, and age of the respondents. There are no descriptions on how the participants are eligible or how they are selected.
- o Original – Third-party data has low originality
- o Comprehensive – Data is comprehensive enough and the parameters relate to Bellabeat's parameters for health activity

- o Current – Data is more than 7 years old as of 2023 and no longer relevant
- o Cited – unknown transparency of citation except for the third-party provider

This dataset is poor in quality and recommendations based on this data is very unsuitable. The anonymity and privacy of the users are verified. The licensing and accessibility are provided by the third party.

Integrity of the dataset is verified and the values matches the criteria for each field, all expected files can be opened, structure and metadata matches.

Data selection: There are 7 files selected for analysis.

1) dailyActivity_merged.csv
2) Heartrate_seconds_merged.csv
3) hourlyCalories_merged.csv
4) hourlySteps_merged.csv
5) hourlyIntensities_merged.csv
6) sleedDay_merged.csv
7) weightLogInfo_merged.csv

Documentation of any cleaning or manipulation of Data

Google Bigquery does not allow Data Manipulation Language (DML) in its free tier. Alternatively, spreadsheet can still handle data cleaning just as well.

Cleaning Steps:

1. Import CSV files to Google Sheets separately
2. Change of Date and time format into 24-Hour format
3. Trim whitespace and blank columns
4. Missing values are labelled as null

SQL Query: https://console.cloud.google.com/bigquery?sq=164068980688:664a3cd551b1467cb1bdd6e18d4e9008

USING SQL:

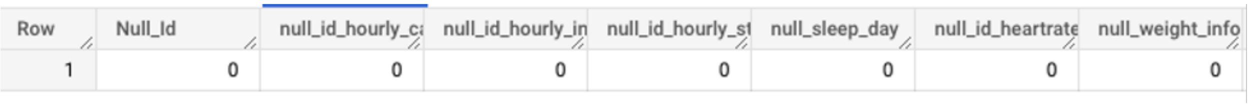
1. Count how many ID is unique from each table –

Exported as unique_id.csv

The result of this query shows that only 8 users out of 33 have used the weight tracking with their health device which is significantly low The result of this query also shows that 14 out of 33 respondents use their smart health device to track their sleep and 24 out of 33 users use it to track their heart rates.

2. Count how many null values in ID column –

Exported as count_non_null.csv



Row	Null_Id	null_id_hourly_ca	null_id_hourly_in	null_id_hourly_st	null_sleep_day	null_id_heartrate	null_weight_info
1	0	0	0	0	0	0	0

There are no null data in all the table's id column.

3. Average activity by all ID–

Exported as avg_non_null.csv and will be used in R.

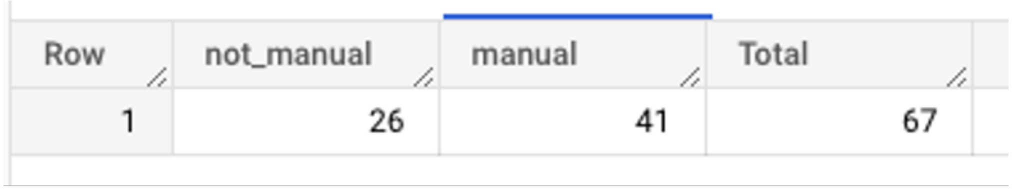
5. Average activity by date

Exported as avg_by_date.csv and will be used in R

6. Average activity by day of week

Exported as change_weight.csv and will be used in R.

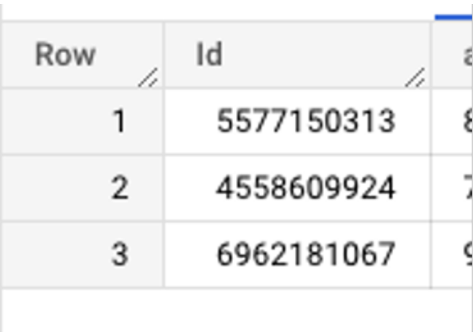
9. COUNT HOW MUCH WEIGHT RECORD IS MANUALLY REPORTED –



Row	not_manual	manual	Total
1	26	41	67

The results show that only 26 out of 67 weight records are automatically report by their smart health weighing device to Fitbit app.

10. AVERAGE BY ID WITH COMPLETE RECORD



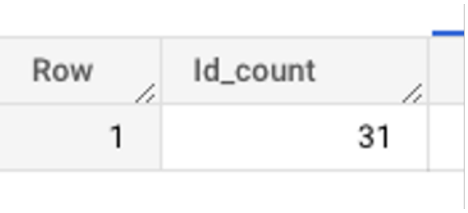
Row	Id	a
1	5577150313	8
2	4558609924	7
3	6962181067	9

Only 3 results are shown which means only 3 users with complete records (no null) including weight.

11. Datediff for every activity

The result is exported as date_diff.csv. It shows that 15 out of 33 respondents did not use the health activity device for 30 days for daily activity, hourly steps, hourly intensity and hourly calories. Only 6 out of 14 respondents use their smart device to tract their heart rates for 30 days. Only 7 out of 24 respondents use to track their sleep activity for 30 days.

12. Count How many Id are less than 30 for either activity and has null values



Row	Id_count
1	31

The result shows that only 2 Ids has 30 days of records and no null in either activity. 31 out of 33 users have either null values or less than 30 days of usage in their smart health devices

Setting up Packages and Importing Files

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

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

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

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

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

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
install.packages('dplyr')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
install.packages('tidyr')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
install.packages('ggpmisc')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

Load these packages

```
library(ggplot2)
library(lubridate)
```

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

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

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr   1.1.1     v stringr 1.5.0
## v forcats 1.0.0     v tibble  3.2.1
## v purrr   1.0.1     v tidyr   1.3.0
## v readr   2.1.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dplyr)
library(tidyr)
library(ggpmisc)
```

```
## Loading required package: ggpp
##
## Attaching package: 'ggpp'
##
## The following object is masked from 'package:ggplot2':
##
##     annotate
```

Import datasets

```
setwd("/cloud/project/SQL Results")
by_id <- read.csv("avg_by_id.csv")
by_time <-read.csv("Avg_by_time_day.csv")
by_day <-read.csv("Avg_by_day_of_week.csv")
by_date <-read.csv("avg_by_date.csv")
weight_change <- read.csv("change_weight.csv")
```

Use the head() functions to make sure everything is imported correctly

```
head(by_date)
```

```
##         Date avg_steps avg_distance avg_very_active_distance
## 1 2016-04-12  8236.848     5.982727                 1.826364
## 2 2016-04-13  7198.727     5.103333                 1.326667
## 3 2016-04-14  7743.576     5.599394                 1.509697
## 4 2016-04-15  7533.848     5.287879                 1.055758
## 5 2016-04-16  8679.156     6.291563                 1.993750
## 6 2016-04-17  6409.250     4.540625                 1.145312
##   avg_moderately_active_distance avg_light_distance avg_sedentary_distance
## 1                      0.3460606           3.410000           0.0003030303
## 2                      0.4200000           3.140909           0.0015151515
## 3                      0.5096970           3.568485           0.0021212121
## 4                      0.4039394           3.767273           0.0015151515
## 5                      0.7087500           3.450625           0.0015625000
## 6                      0.4975000           2.822188           0.0006250000
##   avg_very_active_minutes avg_fairly_active_minutes avg_lightly_active_minutes
## 1                22.30303                  7.848485                   199.0000
## 2                20.33333                 10.575758                   181.7576
## 3                20.93939                 12.393939                   201.0000
## 4                19.18182                  9.878788                   213.8485
## 5                27.84375                 15.125000                   193.8125
## 6                18.90625                 11.843750                   165.3437
##   avg_sedentary_minutes Avg_calories avg_heartrate avg_minutes_asleep
## 1             1026.2121     2390.697      79.70778           441.9231
## 2             1021.7879     2286.636      75.40048           430.4286
## 3             1010.0303     2356.394      76.15573           445.2308
## 4              961.0606     2355.182      79.98256           427.4706
## 5             1002.6563     2392.937      80.39683           391.7143
## 6             1049.9687     2230.969      78.22742           464.0833
##   avg_time_bed
## 1     479.6923
## 2     471.8571
## 3     480.2308
## 4     476.3529
## 5     433.0000
## 6     509.1667
```

```
head(by_day)
```

```
##   Day_number Day_name avg_steps avg_distance avg_very_active_distance
## 1          1      Mon  7758.163     5.532530                 1.525190
## 2          2      Tue  8137.068     5.840509                 1.614710
## 3          3      Wed  7593.228     5.514909                 1.654076
## 4          4      Thu  7928.447     5.690749                 1.522354
## 5          5      Fri  7455.470     5.300973                 1.268997
## 6          6      Sat  8248.486     5.924167                 1.551117
##   avg_moderately_active_distance avg_light_distance avg_sedentary_distance
## 1                      0.5756515           3.360338            0.002643288
## 2                      0.5850167           3.473245            0.001426508
## 3                      0.5238942           3.267079            0.001380626
## 4                      0.5565288           3.466289            0.002452219
## 5                      0.4754668           3.527893            0.001729710
## 6                      0.6781712           3.643202            0.001119397
##   avg_very_active_minutes avg_fairly_active_minutes avg_lightly_active_minutes
## 1                23.06648                  13.80194                   191.9777
## 2                22.91163                  14.13995                   197.7530
## 3                20.86586                  13.18596                   190.9383
## 4                21.37650                  13.24585                   197.7594
## 5                19.84298                  11.88119                   205.5227
## 6                22.32379                  15.25513                   208.2546
##   avg_sedentary_minutes Avg_calories avg_heartrate avg_minutes_asleep
## 1             1027.7448     2323.381      77.39010           418.1206
## 2             1008.9618     2360.501      77.22349           405.6543
## 3              996.3590     2314.677      76.40627           434.4235
## 4             1003.5901     2344.583      77.05641           398.9727
## 5              994.5040     2332.543      77.74210           406.9985
## 6              967.4028     2364.524      79.89676           418.7618
##   avg_time_bed
## 1     455.5441
## 2     444.3866
## 3     470.5638
## 4     433.0152
## 5     447.2778
## 6     459.3068
```

```
head(by_id)
```

```
##           Id avg_steps avg_distance avg_very_active_distance
## 1 1624580081  5743.903     3.914839                0.9393548
## 2 1644430081  7282.967     5.295333                0.7300000
## 3 2022484408 11370.645     8.084193                2.4216129
## 4 2347167796  9519.667     6.355556                1.0594444
## 5 3977333714 10984.567     7.517000                1.6150000
## 6 4319703577  7268.839     4.892258                0.2780645
##   avg_moderately_active_distance avg_light_distance avg_sedentary_distance
## 1                      0.3606452           2.606774            0.006129032
## 2                      0.9510000           3.609000            0.004000000
## 3                      0.7200000           4.942581            0.000000000
## 4                      1.0750000           4.221667            0.000000000
## 5                      2.7510000           3.134333            0.000000000
## 6                      0.5022581           3.768710            0.000000000
##   avg_very_active_minutes avg_fairly_active_minutes avg_lightly_active_minutes
## 1                8.677419                  5.806452                   153.4839
## 2                9.566667                 21.366667                   178.4667
## 3               36.290323                 19.354839                   257.4516
## 4               13.500000                 20.555556                   252.5000
## 5               18.900000                 61.266667                   174.7667
## 6                3.580645                 12.322581                   228.7742
##   avg_sedentary_minutes Avg_calories avg_heartrate avg_minutes_asleep
## 1             1257.7419     1483.355            NA                 NA
## 2             1161.8667     2811.300            NA           294.0000
## 3             1112.5806     2509.968      80.23686                 NA
## 4              687.1667     2043.444      76.72279           446.8000
## 5              707.5333     1513.667            NA           293.6429
## 6              735.8065     2037.677            NA           476.6538
##   avg_time_bed
## 1           NA
## 2     346.0000
## 3           NA
## 4     491.3333
## 5     461.1429
## 6     501.9615
```

```
head(by_time)
```

```
##   Time_of_Day avg_hourly_calories avg_hourly_intensity avg_hourly_step
## 1           0            72.04173            2.2447378       44.479095
## 2           1            70.07866            1.4012412       21.531939
## 3           2            69.13746            1.0476691       18.779369
## 4           3            67.45088            0.4160785        6.003563
## 5           4            68.00397            0.5831131       11.836522
## 6           5            81.34966            4.8722736       43.350994
##   avg_hourly_heartrate
## 1             65.16947
## 2             65.19585
## 3             63.55669
## 4             60.70934
## 5             61.66294
## 6             59.72444
```

```
head(weight_change)
```

```
##           Id                    Latest_Date Latest_BMI Latest_weight
## 1 4558609924 2016-05-09 23:59:59.000000 UTC      27.00          69.1
## 2 4319703577 2016-05-04 23:59:59.000000 UTC      27.38          72.3
## 3 1503960366 2016-05-03 23:59:59.000000 UTC      22.65          52.6
## 4 8877689391 2016-05-12 06:42:53.000000 UTC      25.14          84.0
## 5 2873212765 2016-05-12 23:59:59.000000 UTC      21.69          57.3
## 6 5577150313 2016-04-17 09:17:55.000000 UTC      28.00          90.7
##                     Initial_date initial_BMI initial_weight Datediff   BMI_DIFF
## 1 2016-04-18 23:59:59.000000 UTC       27.25           69.7       21 -0.2500000
## 2 2016-04-17 23:59:59.000000 UTC       27.45           72.4       17 -0.0700016
## 3 2016-05-02 23:59:59.000000 UTC       22.65           52.6        1  0.0000000
## 4 2016-04-12 06:47:11.000000 UTC       25.68           85.8       29 -0.5400009
## 5 2016-04-21 23:59:59.000000 UTC       21.45           56.7       21  0.2399998
## 6 2016-04-17 09:17:55.000000 UTC       28.00           90.7        0  0.0000000
##   WEIGHT_DIFF
## 1 -0.59999848
## 2 -0.09999848
## 3  0.00000000
## 4 -1.80000305
## 5  0.59999848
## 6  0.00000000
```

also Use tibble() functions

```
tibble(by_id)
```

```
## # A tibble: 33 x 15
##          Id avg_steps avg_distance avg_very_active_dist~1 avg_moderately_activ~2
##       <dbl>     <dbl>        <dbl>                  <dbl>                  <dbl>
##  1   1.62e9     5744.         3.91                  0.939                  0.361
##  2   1.64e9     7283.         5.30                  0.730                  0.951
##  3   2.02e9    11371.         8.08                  2.42                   0.720
##  4   2.35e9     9520.         6.36                  1.06                   1.07
##  5   3.98e9    10985.         7.52                  1.61                   2.75
##  6   4.32e9     7269.         4.89                  0.278                  0.502
##  7   4.39e9    10814.         8.39                  1.72                   0.902
##  8   4.70e9     8572.         6.96                  0.417                  1.30
##  9   5.58e9     8304.         6.21                  3.11                   0.658
## 10   6.78e9     2520.         1.81                  0.709                  0.384
## # i 23 more rows
## # i abbreviated names: 1: avg_very_active_distance,
## #   2: avg_moderately_active_distance
## # i 10 more variables: avg_light_distance <dbl>, avg_sedentary_distance <dbl>,
## #   avg_very_active_minutes <dbl>, avg_fairly_active_minutes <dbl>,
## #   avg_lightly_active_minutes <dbl>, avg_sedentary_minutes <dbl>,
## #   Avg_calories <dbl>, avg_heartrate <dbl>, avg_minutes_asleep <dbl>, ...
```

```
tibble(by_day)
```

```
## # A tibble: 7 x 16
##   Day_number Day_name avg_steps avg_distance avg_very_active_distance
##        <int> <chr>        <dbl>        <dbl>                    <dbl>
## 1          1 Mon          7758.         5.53                     1.53
## 2          2 Tue          8137.         5.84                     1.61
## 3          3 Wed          7593.         5.51                     1.65
## 4          4 Thu          7928.         5.69                     1.52
## 5          5 Fri          7455.         5.30                     1.27
## 6          6 Sat          8248.         5.92                     1.55
## 7          7 Sun          6953.         5.04                     1.50
## # i 11 more variables: avg_moderately_active_distance <dbl>,
## #   avg_light_distance <dbl>, avg_sedentary_distance <dbl>,
## #   avg_very_active_minutes <dbl>, avg_fairly_active_minutes <dbl>,
## #   avg_lightly_active_minutes <dbl>, avg_sedentary_minutes <dbl>,
## #   Avg_calories <dbl>, avg_heartrate <dbl>, avg_minutes_asleep <dbl>,
## #   avg_time_bed <dbl>
```

Check if imported date columns are correct

```
str(by_date)
```

```
## 'data.frame':    31 obs. of  15 variables:
```

```
##  $ Date                         : chr  "2016-04-12" "2016-04-13" "2016-04-14" "2016-04-15" ...
##  $ avg_steps                    : num  8237 7199 7744 7534 8679 ...
##  $ avg_distance                 : num  5.98 5.1 5.6 5.29 6.29 ...
##  $ avg_very_active_distance     : num  1.83 1.33 1.51 1.06 1.99 ...
##  $ avg_moderately_active_distance: num  0.346 0.42 0.51 0.404 0.709 ...
##  $ avg_light_distance           : num  3.41 3.14 3.57 3.77 3.45 ...
##  $ avg_sedentary_distance       : num  0.000303 0.001515 0.002121 0.001515 0.001562 ...
##  $ avg_very_active_minutes      : num  22.3 20.3 20.9 19.2 27.8 ...
##  $ avg_fairly_active_minutes    : num  7.85 10.58 12.39 9.88 15.12 ...
##  $ avg_lightly_active_minutes   : num  199 182 201 214 194 ...
##  $ avg_sedentary_minutes        : num  1026 1022 1010 961 1003 ...
##  $ Avg_calories                 : num  2391 2287 2356 2355 2393 ...
##  $ avg_heartrate                : num  79.7 75.4 76.2 80 80.4 ...
##  $ avg_minutes_asleep           : num  442 430 445 427 392 ...
##  $ avg_time_bed                 : num  480 472 480 476 433 ...
```

```
str(by_day)
```

```
## 'data.frame':    7 obs. of  16 variables:
##  $ Day_number                   : int  1 2 3 4 5 6 7
##  $ Day_name                     : chr  "Mon" "Tue" "Wed" "Thu" ...
##  $ avg_steps                    : num  7758 8137 7593 7928 7455 ...
##  $ avg_distance                 : num  5.53 5.84 5.51 5.69 5.3 ...
##  $ avg_very_active_distance     : num  1.53 1.61 1.65 1.52 1.27 ...
##  $ avg_moderately_active_distance: num  0.576 0.585 0.524 0.557 0.475 ...
##  $ avg_light_distance           : num  3.36 3.47 3.27 3.47 3.53 ...
##  $ avg_sedentary_distance       : num  0.00264 0.00143 0.00138 0.00245 0.00173 ...
##  $ avg_very_active_minutes      : num  23.1 22.9 20.9 21.4 19.8 ...
##  $ avg_fairly_active_minutes    : num  13.8 14.1 13.2 13.2 11.9 ...
##  $ avg_lightly_active_minutes   : num  192 198 191 198 206 ...
##  $ avg_sedentary_minutes        : num  1028 1009 996 1004 995 ...
##  $ Avg_calories                 : num  2323 2361 2315 2345 2333 ...
##  $ avg_heartrate                : num  77.4 77.2 76.4 77.1 77.7 ...
##  $ avg_minutes_asleep           : num  418 406 434 399 407 ...
##  $ avg_time_bed                 : num  456 444 471 433 447 ...
```

```
str(by_id)
```

```
## 'data.frame':    33 obs. of  15 variables:
##  $ Id                           : num  1.62e+09 1.64e+09 2.02e+09 2.35e+09 3.98e+09 ...
##  $ avg_steps                    : num  5744 7283 11371 9520 10985 ...
##  $ avg_distance                 : num  3.91 5.3 8.08 6.36 7.52 ...
##  $ avg_very_active_distance     : num  0.939 0.73 2.422 1.059 1.615 ...
##  $ avg_moderately_active_distance: num  0.361 0.951 0.72 1.075 2.751 ...
##  $ avg_light_distance           : num  2.61 3.61 4.94 4.22 3.13 ...
##  $ avg_sedentary_distance       : num  0.00613 0.004 0 0 0 ...
##  $ avg_very_active_minutes      : num  8.68 9.57 36.29 13.5 18.9 ...
##  $ avg_fairly_active_minutes    : num  5.81 21.37 19.35 20.56 61.27 ...
##  $ avg_lightly_active_minutes   : num  153 178 257 252 175 ...
##  $ avg_sedentary_minutes        : num  1258 1162 1113 687 708 ...
##  $ Avg_calories                 : num  1483 2811 2510 2043 1514 ...
##  $ avg_heartrate                : num  NA NA 80.2 76.7 NA ...
##  $ avg_minutes_asleep           : num  NA 294 NA 447 294 ...
##  $ avg_time_bed                 : num  NA 346 NA 491 461 ...
```

```
str(by_time)
```

```
## 'data.frame':    24 obs. of  5 variables:
##  $ Time_of_Day         : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ avg_hourly_calories : num  72 70.1 69.1 67.5 68 ...
##  $ avg_hourly_intensity: num  2.245 1.401 1.048 0.416 0.583 ...
##  $ avg_hourly_step     : num  44.5 21.5 18.8 6 11.8 ...
##  $ avg_hourly_heartrate: num  65.2 65.2 63.6 60.7 61.7 ...
```

Confirming date column to be date format using either

```
by_date$date <- format(by_date$Date, format = "%Y-%m-%d")
as.Date(by_date$Date, "%Y-%m-%d")
```

```
##  [1] "2016-04-12" "2016-04-13" "2016-04-14" "2016-04-15" "2016-04-16"
##  [6] "2016-04-17" "2016-04-18" "2016-04-19" "2016-04-20" "2016-04-21"
## [11] "2016-04-22" "2016-04-23" "2016-04-24" "2016-04-25" "2016-04-26"
## [16] "2016-04-27" "2016-04-28" "2016-04-29" "2016-04-30" "2016-05-01"
## [21] "2016-05-02" "2016-05-03" "2016-05-04" "2016-05-05" "2016-05-06"
## [26] "2016-05-07" "2016-05-08" "2016-05-09" "2016-05-10" "2016-05-11"
## [31] "2016-05-12"
```

Summary of Data

```
# Daily Activity
by_id %>%
  select(avg_steps,
         avg_distance,
         Avg_calories,
         avg_minutes_asleep,
         avg_time_bed,
         avg_very_active_minutes,
         avg_very_active_distance,
         avg_fairly_active_minutes,
         avg_moderately_active_distance,
         avg_lightly_active_minutes,
         avg_light_distance,
         avg_sedentary_minutes,
         avg_sedentary_distance) %>%
  summary()
```

```
##    avg_steps        avg_distance      Avg_calories  avg_minutes_asleep
##  Min.   :  916.1   Min.   : 0.6345   Min.   :1483   Min.   : 61.0
##  1st Qu.: 5566.9   1st Qu.: 3.4548   1st Qu.:1917   1st Qu.:336.3
##  Median : 7283.0   Median : 5.2953   Median :2132   Median :419.1
##  Mean   : 7519.3   Mean   : 5.3990   Mean   :2282   Mean   :377.6
##  3rd Qu.: 9519.7   3rd Qu.: 6.9135   3rd Qu.:2600   3rd Qu.:449.3
##  Max.   :16040.0   Max.   :13.2129   Max.   :3437   Max.   :652.0
##                                                     NA's   :9
##   avg_time_bed   avg_very_active_minutes avg_very_active_distance
##  Min.   : 69.0   Min.   : 0.09677        Min.   :0.006129
##  1st Qu.:377.1   1st Qu.: 3.58065        1st Qu.:0.142258
##  Median :447.9   Median :10.38710        Median :0.730000
##  Mean   :420.1   Mean   :20.30877        Mean   :1.449551
##  3rd Qu.:485.3   3rd Qu.:23.41935        3rd Qu.:2.214210
##  Max.   :961.0   Max.   :87.33333        Max.   :8.514839
##  NA's   :9
##  avg_fairly_active_minutes avg_moderately_active_distance
##  Min.   : 0.2581           Min.   :0.01129
##  1st Qu.: 4.0345           1st Qu.:0.12828
##  Median :12.3226           Median :0.50226
##  Mean   :13.2602           Mean   :0.55704
##  3rd Qu.:19.3548           3rd Qu.:0.77323
##  Max.   :61.2667           Max.   :2.75100
##
##  avg_lightly_active_minutes avg_light_distance avg_sedentary_minutes
##  Min.   : 38.58             Min.   :0.5071     Min.   : 662.3
##  1st Qu.:143.84             1st Qu.:2.6068     1st Qu.: 766.4
##  Median :206.19             Median :3.5045     Median :1077.5
##  Mean   :191.52             Mean   :3.3175     Mean   : 999.2
##  3rd Qu.:245.81             3rd Qu.:4.1435     3rd Qu.:1206.6
##  Max.   :327.90             Max.   :6.1887     Max.   :1317.4
##
##  avg_sedentary_distance
##  Min.   :0.0000000
##  1st Qu.:0.0000000
##  Median :0.0000000
##  Mean   :0.0016250
##  3rd Qu.:0.0007692
##  Max.   :0.0110000
##
```

The mean time asleep per day is 377 minutes or 6.3 hours which is below the recommended 8 hours by health experts. The average sedentary minutes is 999 minutes or 16 hours which is too much time.

The average sedentary minutes per day is very high at 999 or 16 hours which should be 10 hours at maximum.

The average steps taken per day is too low at 7,519 steps which should be aimed at around 10,000 steps per day according to medical health experts.

According to dietary guidelines, it is recommended for the average adult woman to burn around 1600 to 2400 calories per day, while for the average adult man to burn around 2000 to 3000 calories per day. The average here shows 2282 calories per day which is quite high for woman, but not too high for a man. There should also be other parameters such as calorie intake to measure calorie deficit and diet goal from BMI. Since there are no gender data and other parameters for calorie in the respondents therefore it is not sufficient to draw conclusions from the data.

```
#Weight Change Summary
weight_change %>%
  select(Datediff,
         initial_BMI,
         Latest_BMI,
         BMI_DIFF,
         WEIGHT_DIFF) %>%
  summary()
```

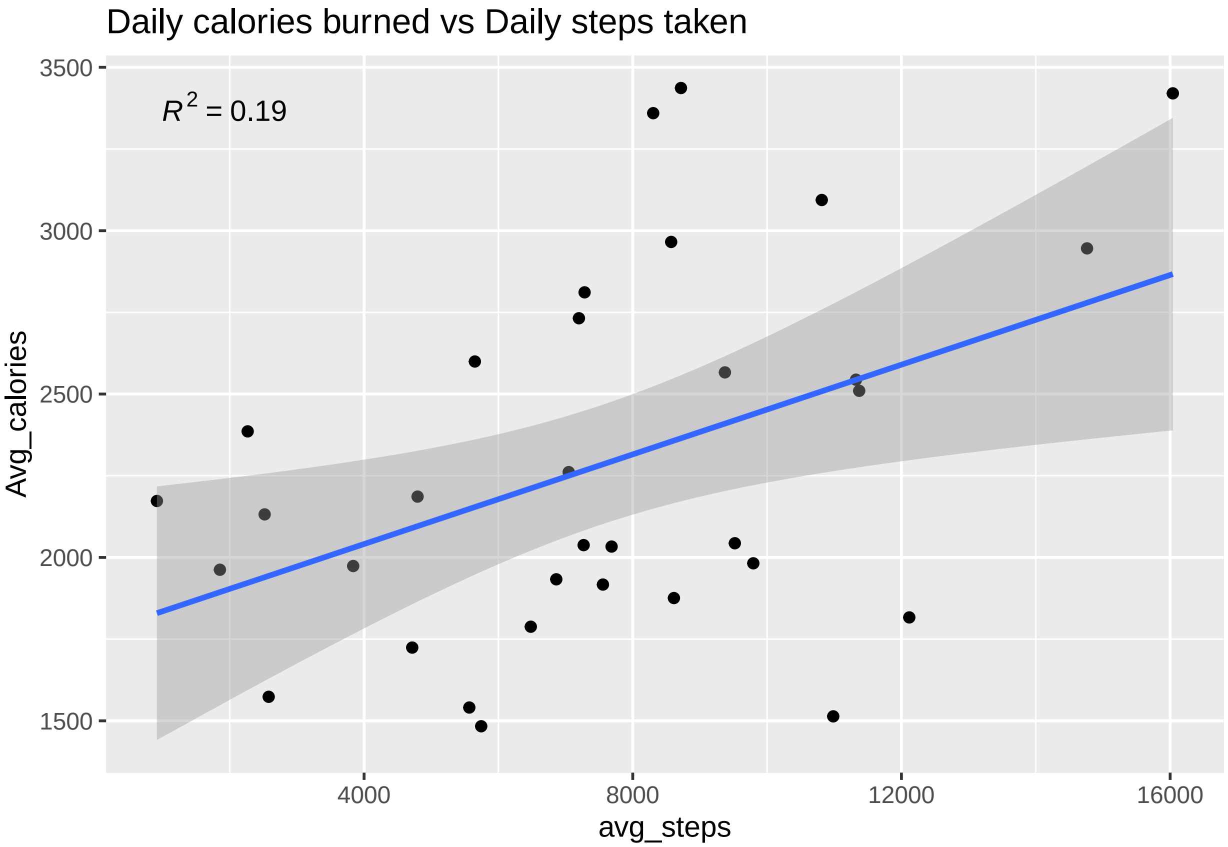
```
##     Datediff      initial_BMI      Latest_BMI       BMI_DIFF
##  Min.   : 0.00   Min.   :21.45   Min.   :21.69   Min.   :-0.5400
##  1st Qu.: 0.75   1st Qu.:23.95   1st Qu.:23.79   1st Qu.:-0.2275
##  Median :19.00   Median :26.46   Median :26.07   Median :-0.0350
##  Mean   :14.88   Mean   :28.05   Mean   :27.95   Mean   :-0.1050
##  3rd Qu.:23.00   3rd Qu.:27.59   3rd Qu.:27.54   3rd Qu.: 0.0000
##  Max.   :30.00   Max.   :47.54   Max.   :47.54   Max.   : 0.2400
##   WEIGHT_DIFF
##  Min.   :-1.8000
##  1st Qu.:-0.6000
##  Median :-0.0500
##  Mean   :-0.3125
##  3rd Qu.: 0.0000
##  Max.   : 0.6000
```

Despite small sample size, the respondents that track their weight data shows considerable weight change. According to CDC, the normal range of BMI is between 18.5 to 24.9 and a healthy range of weight loss is 0.4 to 0.8 kg per week. The data here shows an average of 0.3 kg decrease per 2 weeks. Their average BMI lowered by 0.1 from an average of 28. Although not much, Fitbit smart health device has helped respondents in making progress toward their weight loss journey.

Visualization

```
# Correlation between calories and total steps
ggplot(data=by_id, aes(x=avg_steps, y=Avg_calories)) + geom_point() + geom_smooth(method = "lm") + stat
```

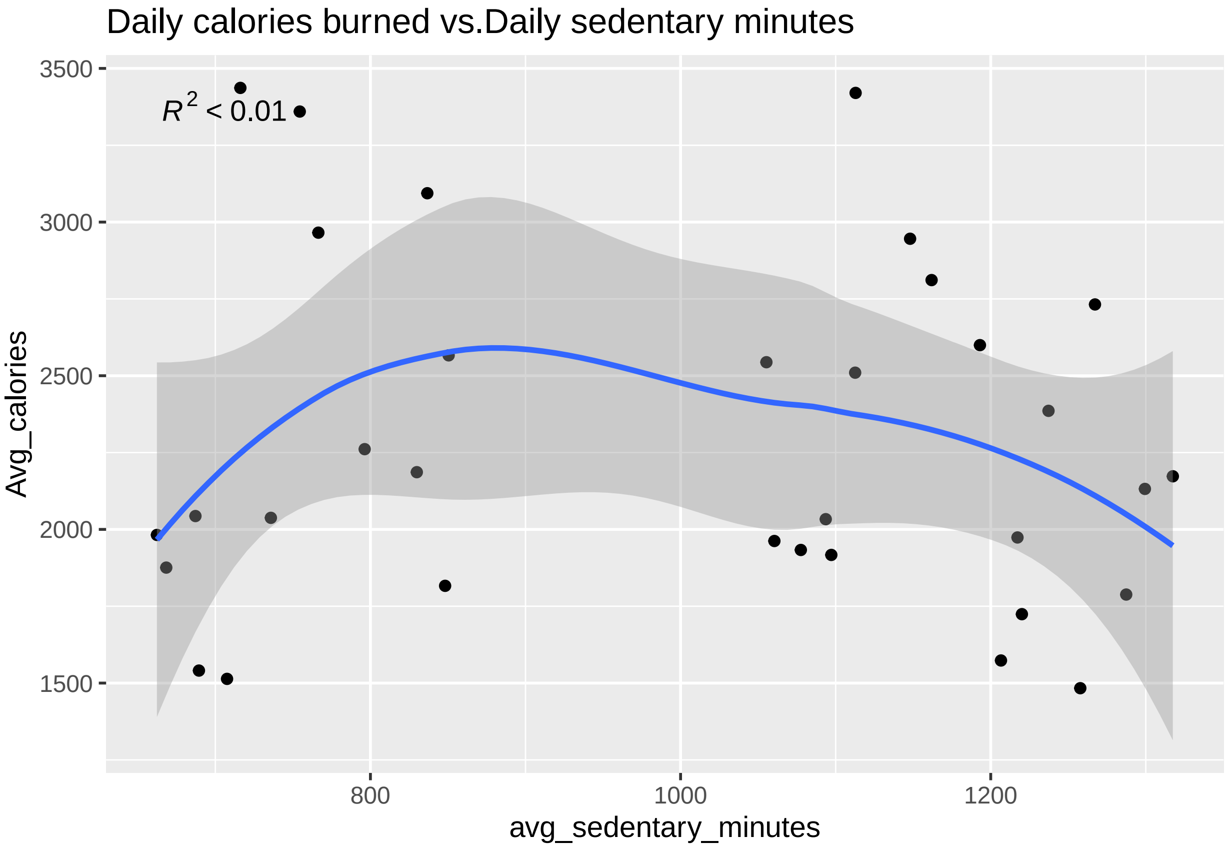
```
## `geom_smooth()` using formula = 'y ~ x'
```



The correlation between daily calories burned and daily steps taken is quite low since it is less than 0.4. This indicates that there are respondents who burn a lot of calories without taking many steps and there are respondents who burn less calories despite taking many steps. The prior is due to high intensity activity with little displacement such as exercising in the gym or lifting weights. The latter represents people who took more steps, but still burned less calorie. A probable reason for this is most steps taken are casual strolling with low intensity.

```
# Correlation between calories and sedentary minutes
ggplot(data=by_id, aes(x=avg_sedentary_minutes, y=Avg_calories)) + geom_point() + geom_smooth(method =
```

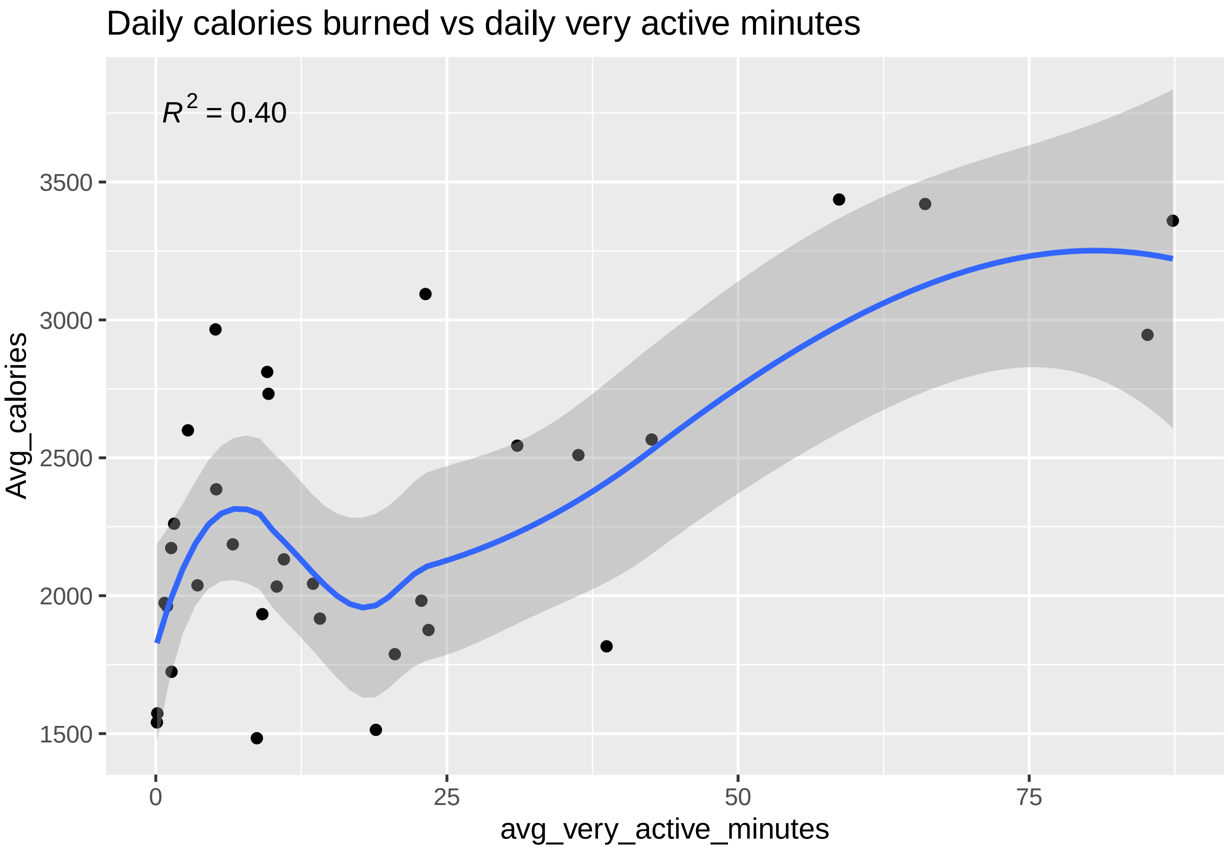
```
## `geom_smooth()` using formula = 'y ~ x'
```



There is a very insignificant negative correlation between sedentary minutes and calories burned. The chart below will explain further why.

```
# Correlation between calories and very active minutes
ggplot(data=by_id, aes(x=avg_very_active_minutes, y=Avg_calories)) + geom_point() + geom_smooth(method =
```

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



This chart shows that there is a stronger correlation in a positive direction between calories burned and very active minutes compared to calories and sedentary minutes. A plausible reason behind this is if someone does a high intensity activity, it does not matter if they take a longer rest. What matters more for calorie burning is the active minutes. Long sedentary minutes, however, could negatively impact health and wellness.

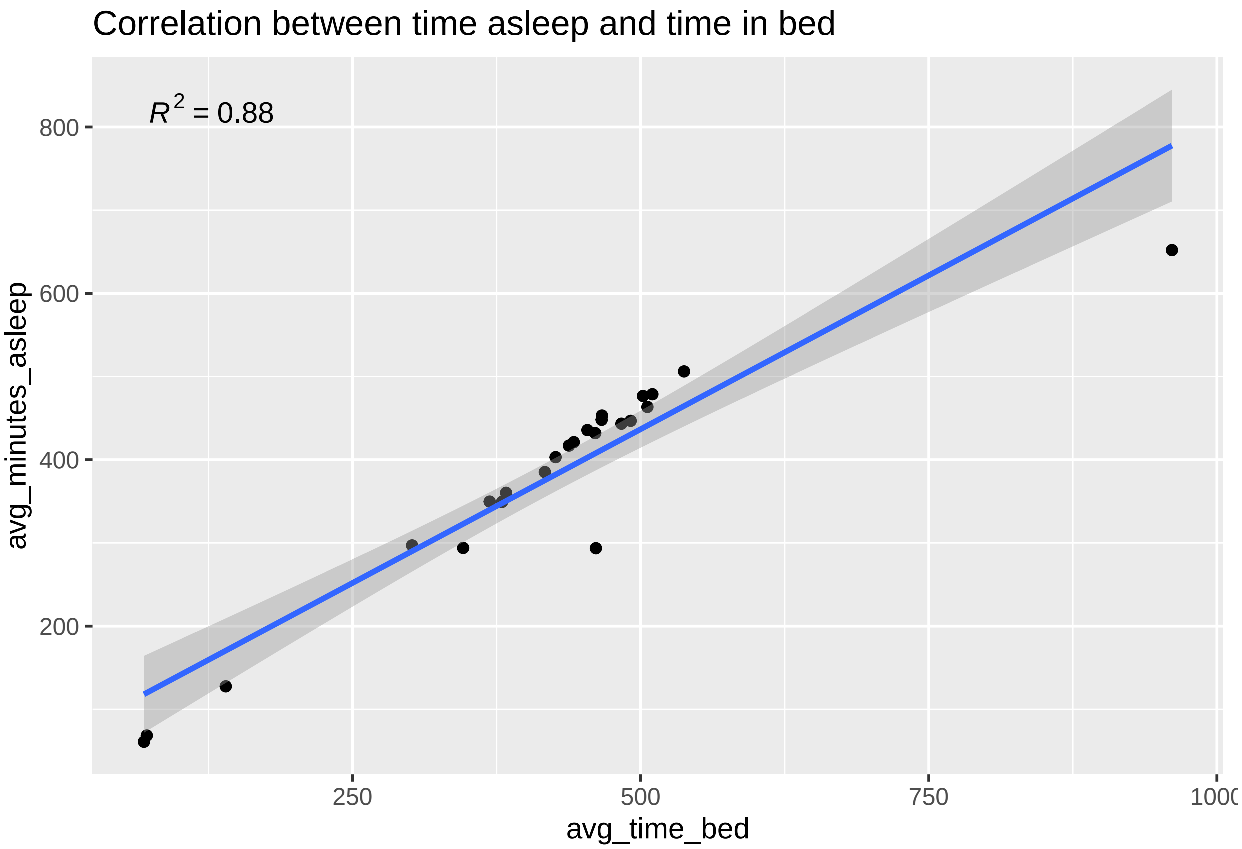
```
# Correlation between time in bed and minutes asleep
ggplot(data=by_id, aes(x=avg_time_bed, y=avg_minutes_asleep)) + geom_point() + geom_smooth(method = "lm
```

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

```
## Warning: Removed 9 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 9 rows containing non-finite values (`stat_poly_eq()`).
```

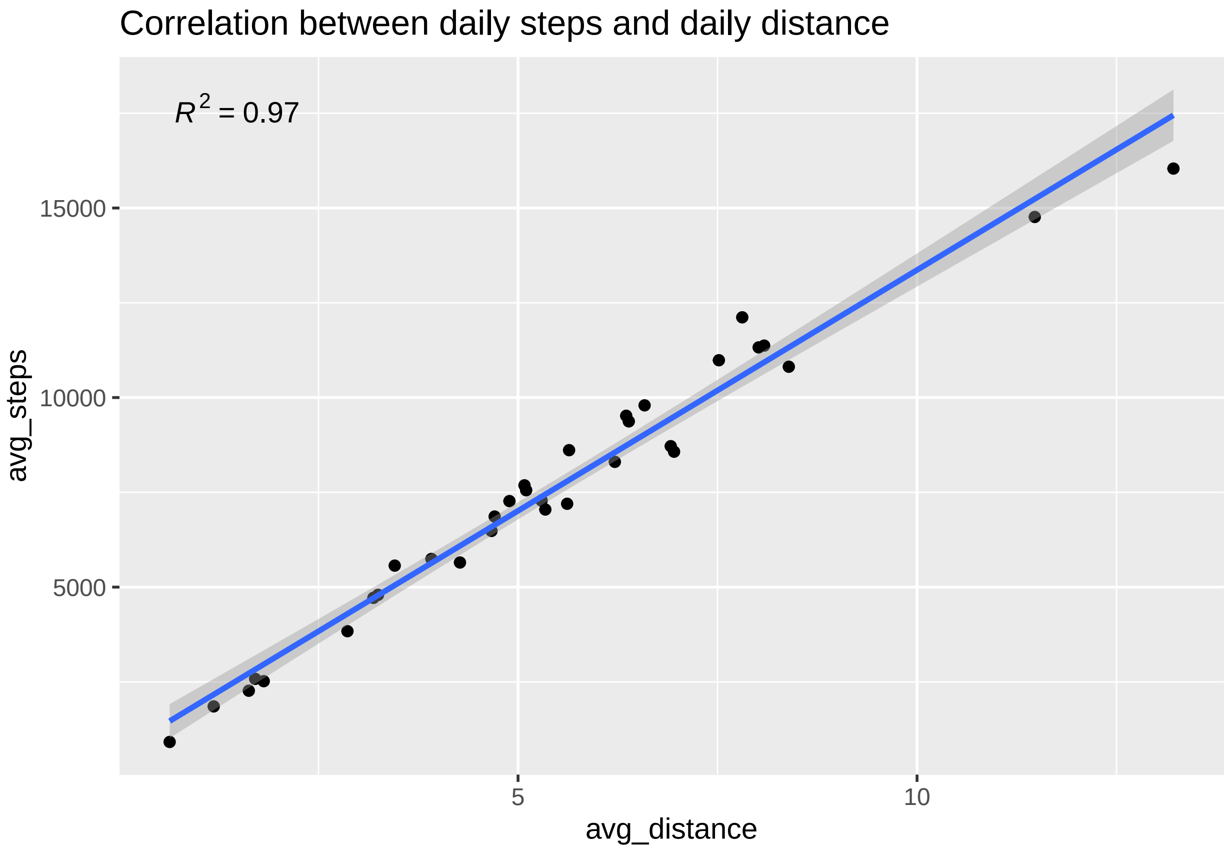
```
## Warning: Removed 9 rows containing missing values (`geom_point()`).
```



There is a very stong positive correlation between time spent in bed and minutes asleep.

```
# Correlation between daily steps taken and daily distance traveled
ggplot(data=by_id, aes(x=avg_distance, y=avg_steps)) + geom_point() + geom_smooth(method = "lm") + stat_
```

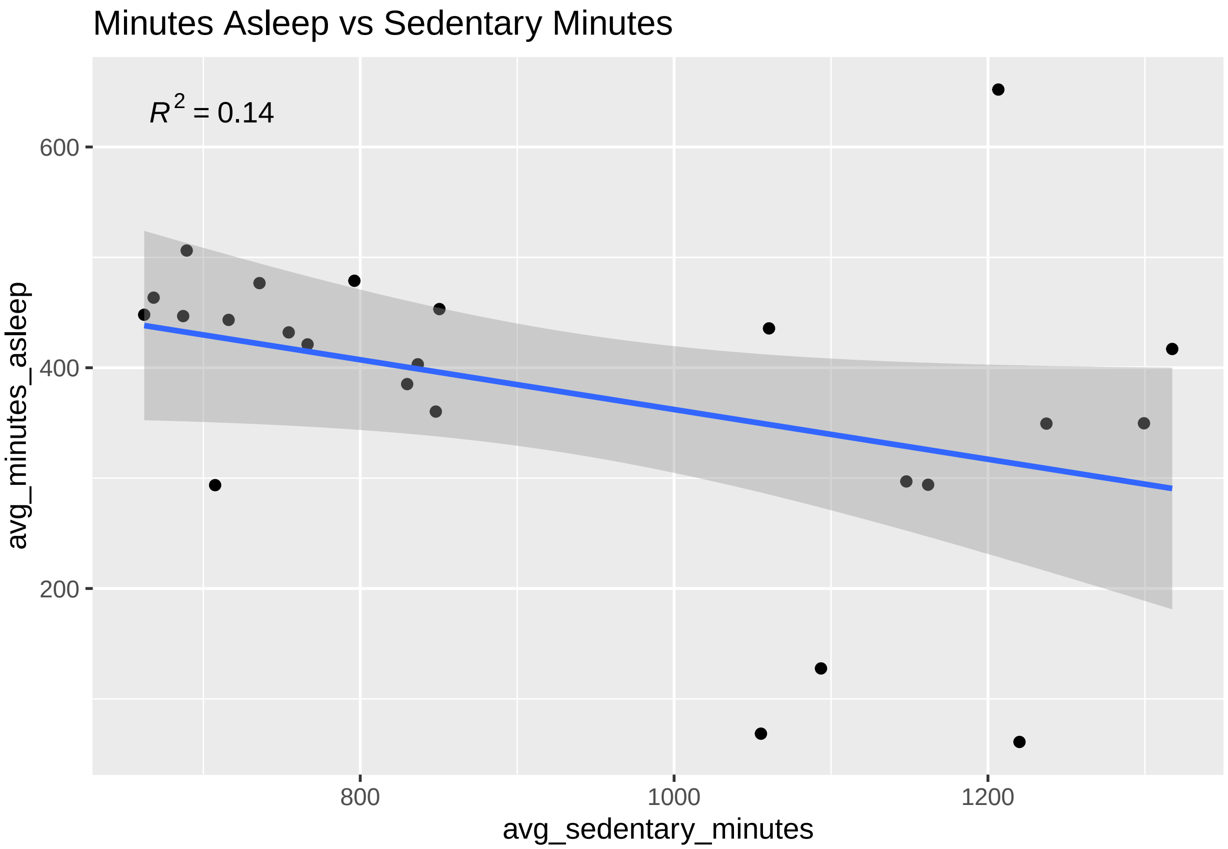
```
## `geom_smooth()` using formula = 'y ~ x'
```



It is clear that there is a strong positive correlation between daily steps and daily distance. The more steps taken the further the distance traveled.

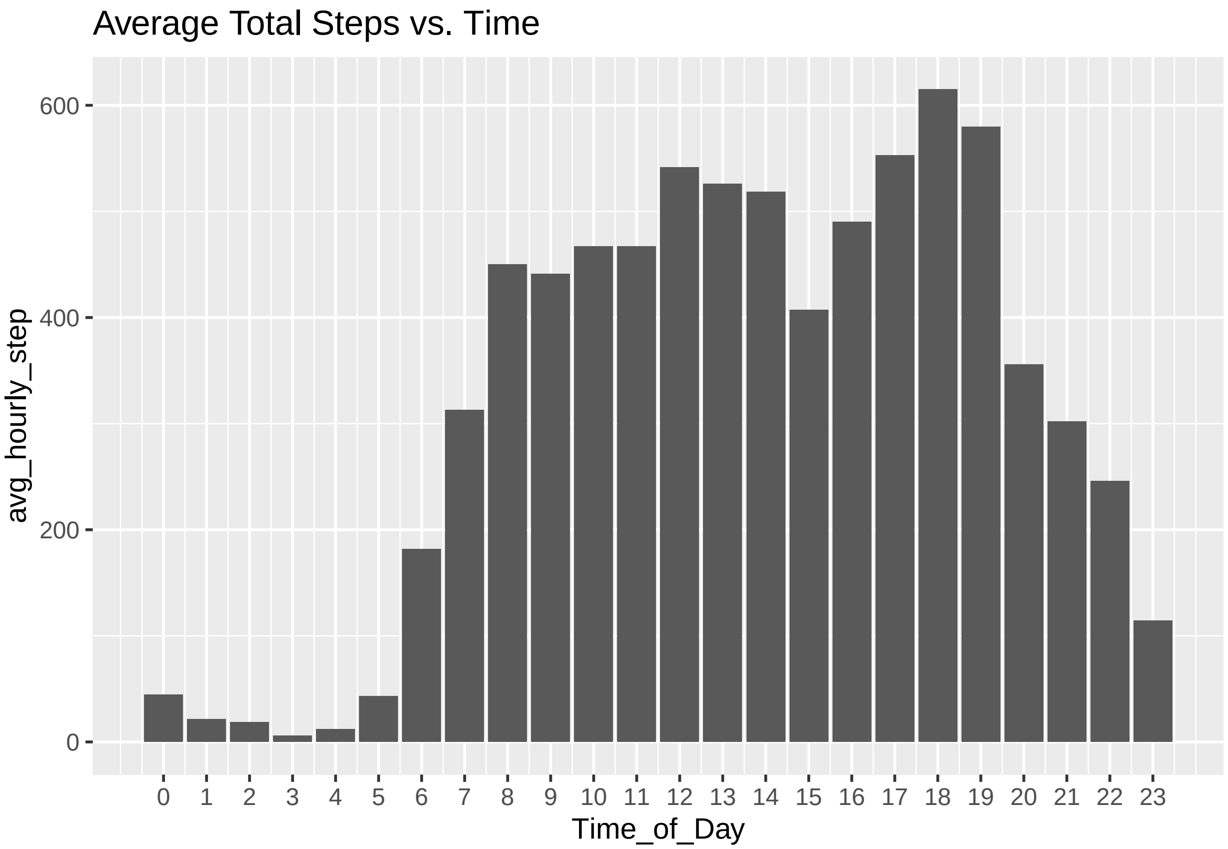
```
# Minutes Asleep vs . Sedentary Minutes
ggplot(data=by_id, aes(x=avg_sedentary_minutes, y=avg_minutes_asleep)) + geom_point() + stat_poly_eq() +
```

```
## Warning: Removed 9 rows containing non-finite values (`stat_poly_eq()`).

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

## Warning: Removed 9 rows containing non-finite values (`stat_smooth()`).

## Warning: Removed 9 rows containing missing values (`geom_point()`).
```



There is a insignificant low negative correlation between sedentary minutes and minutes asleep. We would expect that lower sedentary time means that higher minutes asleep since the users will have lower energy after being more active. Turns out there are other effects that makes falling asleep much more difficult such as smartphone usage or screen time or caffeine intake. This makes users more alert and decrease minutes asleep despite not being active.

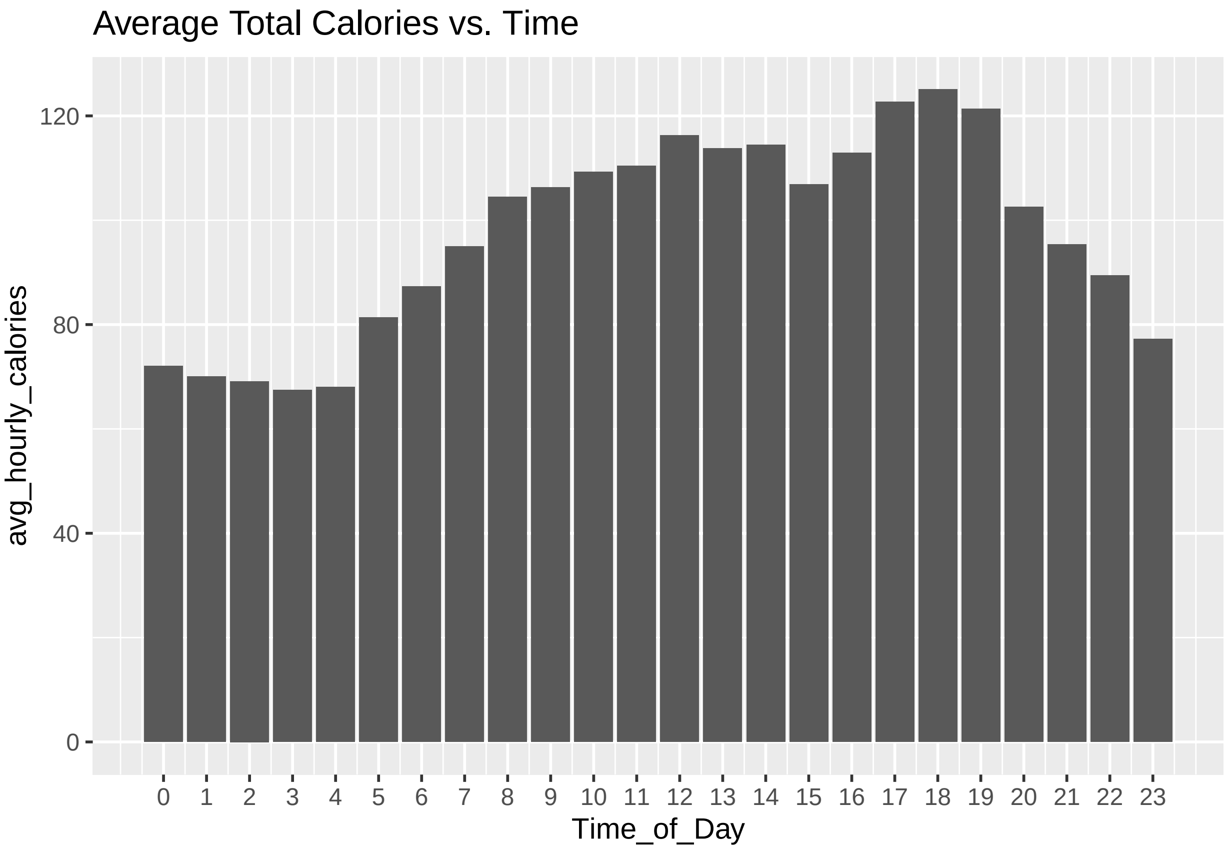
```r
# Average Total Steps vs. Time
ggplot(data = by_time, aes(x = Time_of_Day, y = avg_hourly_step)) + geom_col() + scale_x_continuous(brea
```



The bar chart shows that highest amount of steps taken is at 6-7 PM or 18:00 to 19:00. This makes sense due to respondents going back from work or just enjoying the evening walk since afternoon is too hot during the summer at the time data this was collected. The number of steps fall drastically at 8 PM or 20:00 as the users are preparing to go to bed.

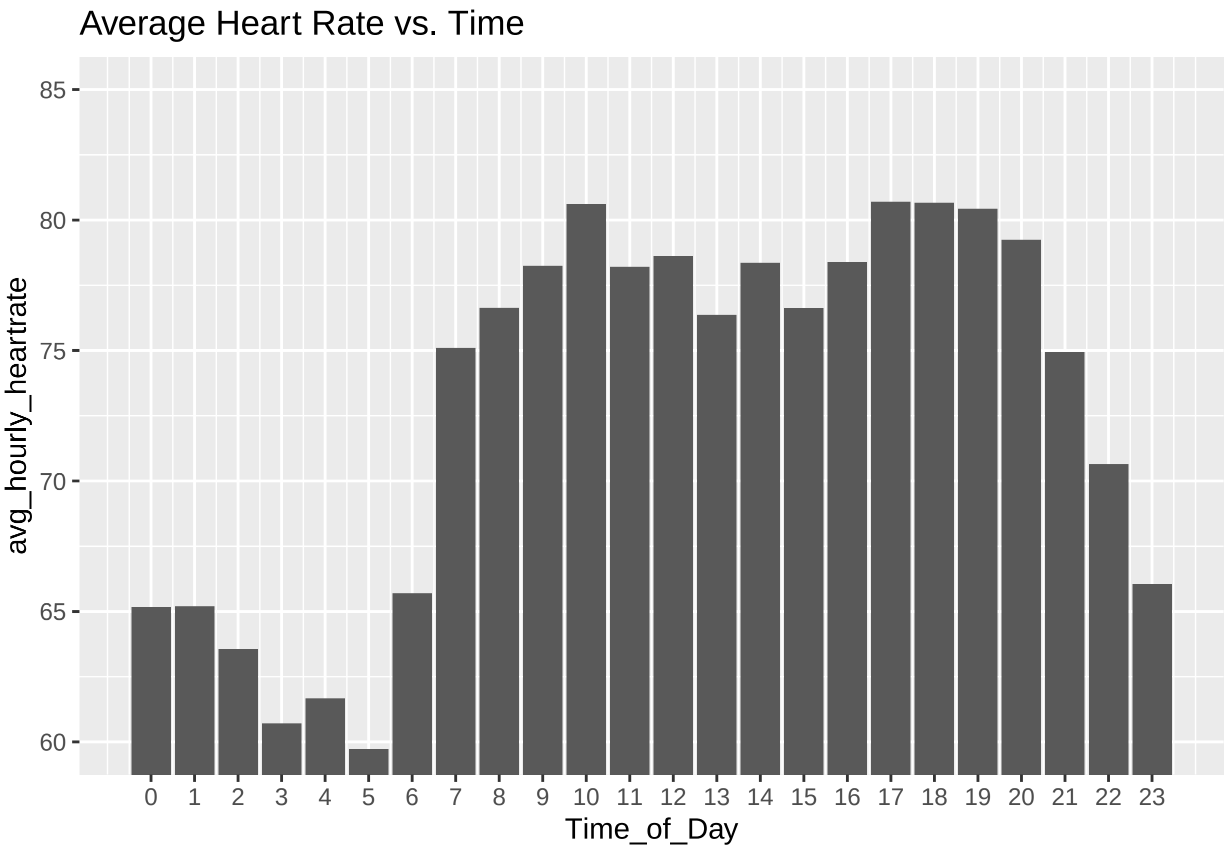
Between 00:00 to 05:00 the respondents are asleep so there is little to no steps unless needed to. At 6 AM, there is a significant increase which indicates respondents wake up around this time and it continues to increase until 8-11 AM which it remains relatively stable. At 12 PM, there is an increase which may be explained by users trying to find lunch.

```
# Average Calories Burned vs. Time
ggplot(data = by_time, aes(x = Time_of_Day, y = avg_hourly_calories)) + geom_col() + scale_x_continuous
```



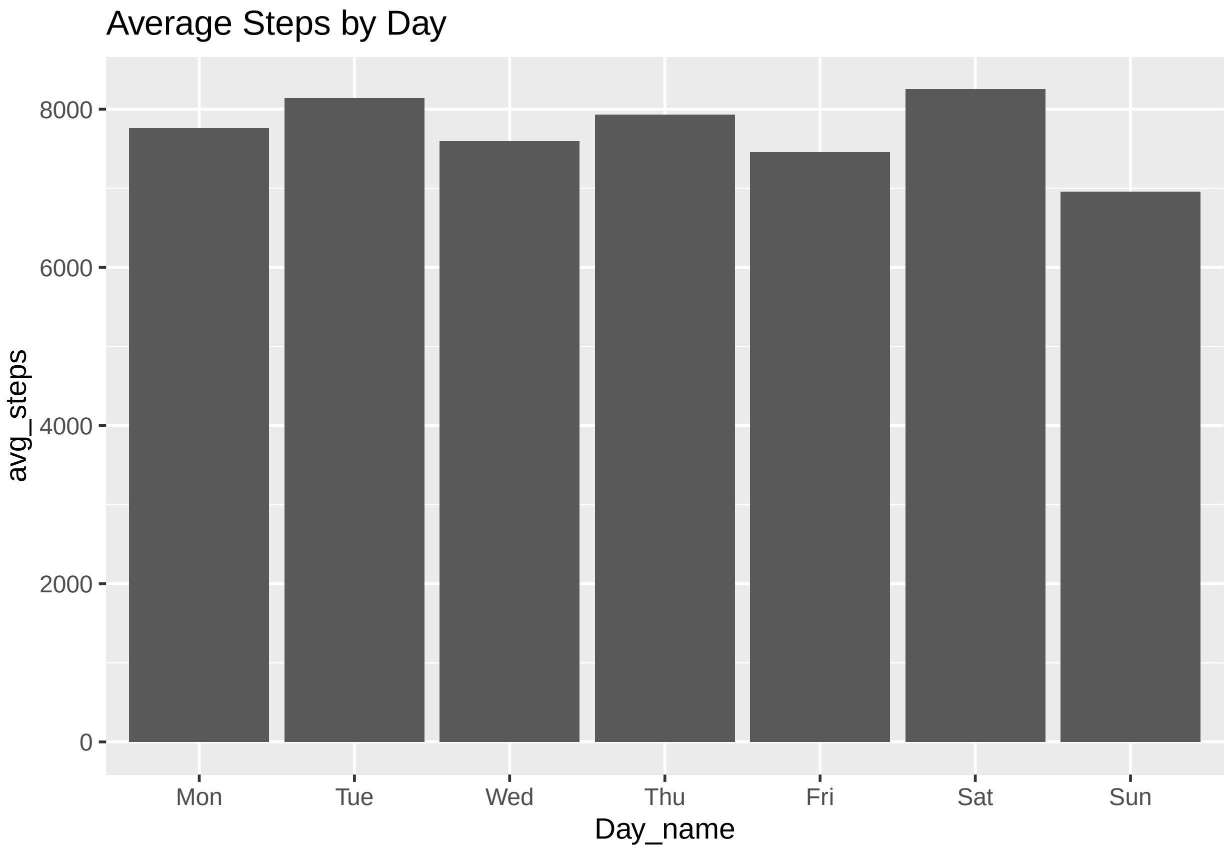
The time at which calories burned correlates to steps taken. At 18:00 where steps taken are the highest, calories burned are also highest. Calories burned least during sleeping time 23:00-04:00 as the body's metabolism slows down. While the calories burned increase, as the user wakes up and begins activity from 07:00-16:00.

```
# Average Heartrate vs. Time
ggplot(data = by_time, aes(x = Time_of_Day, y = avg_hourly_heartrate)) + coord_cartesian(ylim = c(60,85
```



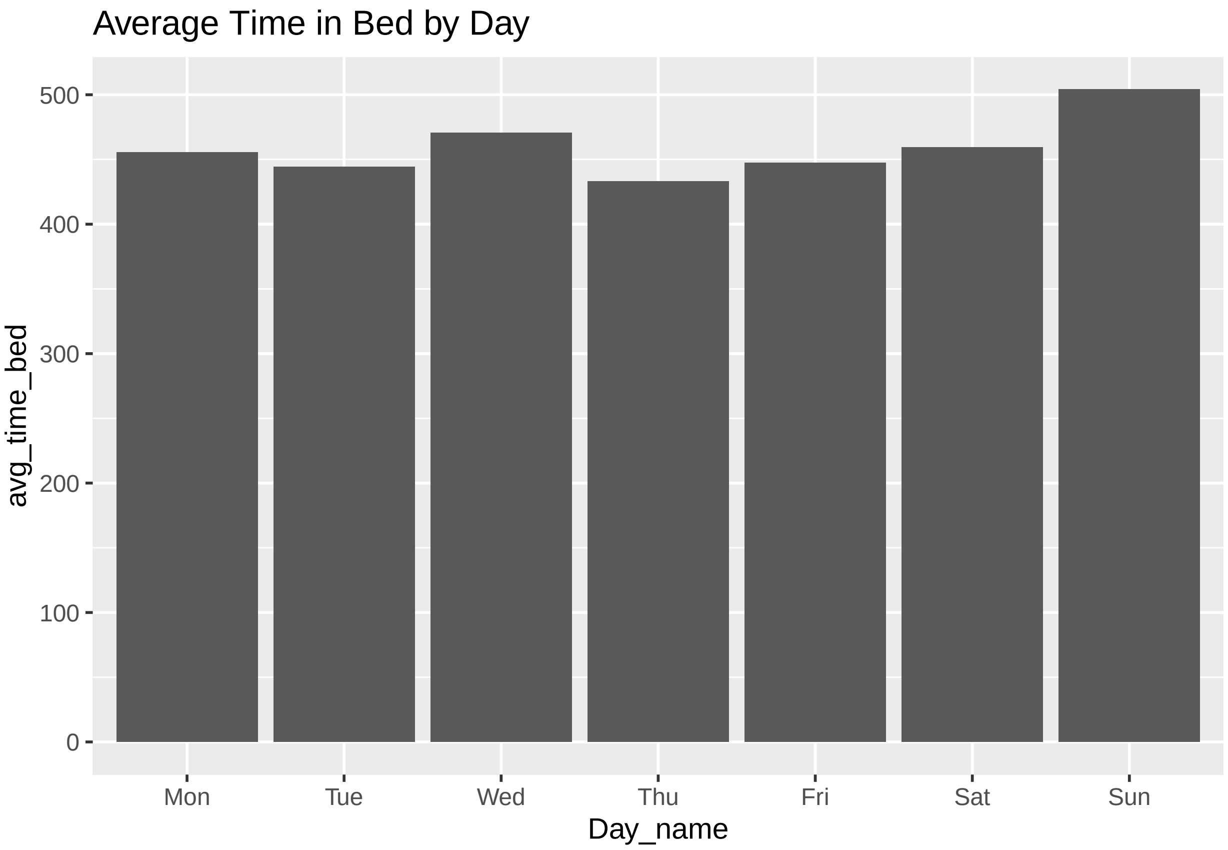
The normal heart rate is 60-100 bpm. Heart rate increases during high activity or stress response. Thus, it is lower during sleep and high during more intense activity. The anomaly from this chart is at 10 AM. At this time, the steps taken are lowm, but the heart rate is as high as 5-7 PM when the steps taken is highest. This is probably due to work related-stress, caffeine effect from the morning or other social factors which there isn't sufficient data collected to draw conclusion.

```
# Steps taken throughout the Week
ggplot(data = by_day, aes(x = Day_name, y = avg_steps)) + geom_col() + scale_x_discrete(limits = c("Mon
```



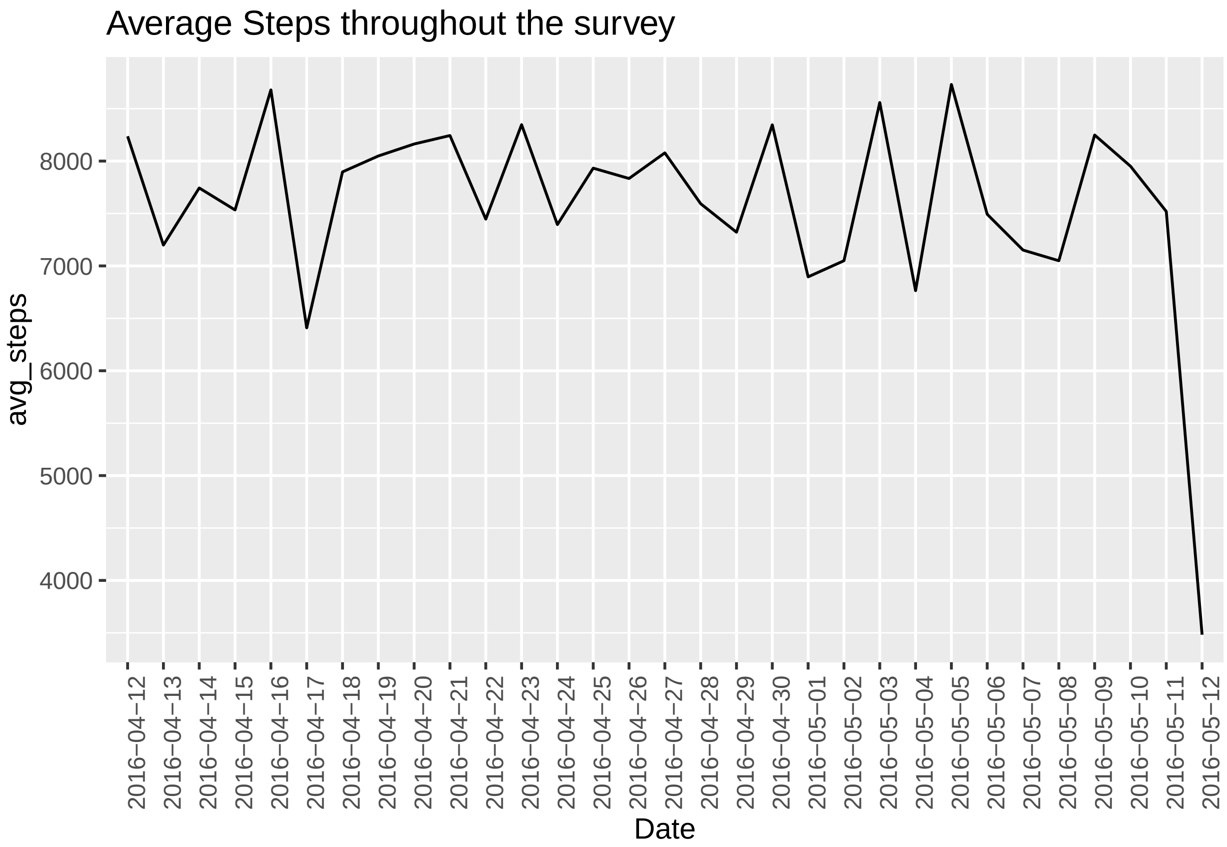
There is little differences between average steps taken throughout the weekdays from Monday-Friday. Saturday has the highest average steps and Sunday has the lowest. Users tend to relax more on Sundays to prepare for the work day. On Saturday, they plan to go out more either excercise or doing leisure activities.

```
# Time in Bed taken throughout the Week
ggplot(data = by_day, aes(x = Day_name, y = avg_time_bed)) + geom_col() + scale_x_discrete(limits = c("
```



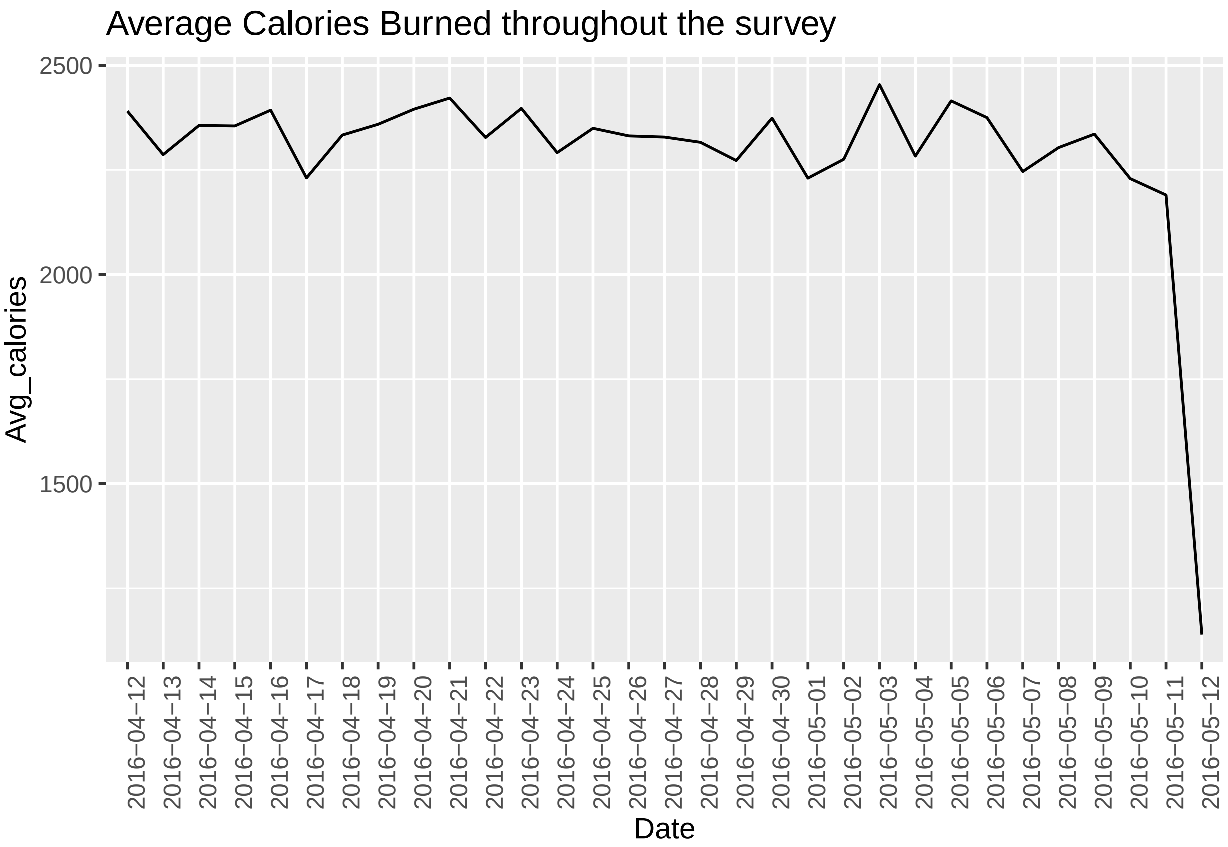
Users spend almost 500 minutes or 8.3 hours in bed on Sundays (the highest) followed by Wednesday. Despite Saturday being a weekend, time spent in Bed resting is lower than Wednesday. A probable cause of this is that Wednesday is the middle of the weekday where people tend to get exhausted from work.

```
# Average Progress of Steps
ggplot(data = by_date, aes(x = Date, y = avg_steps, group = 1)) + geom_line() + theme(axis.text.x=elemen
```



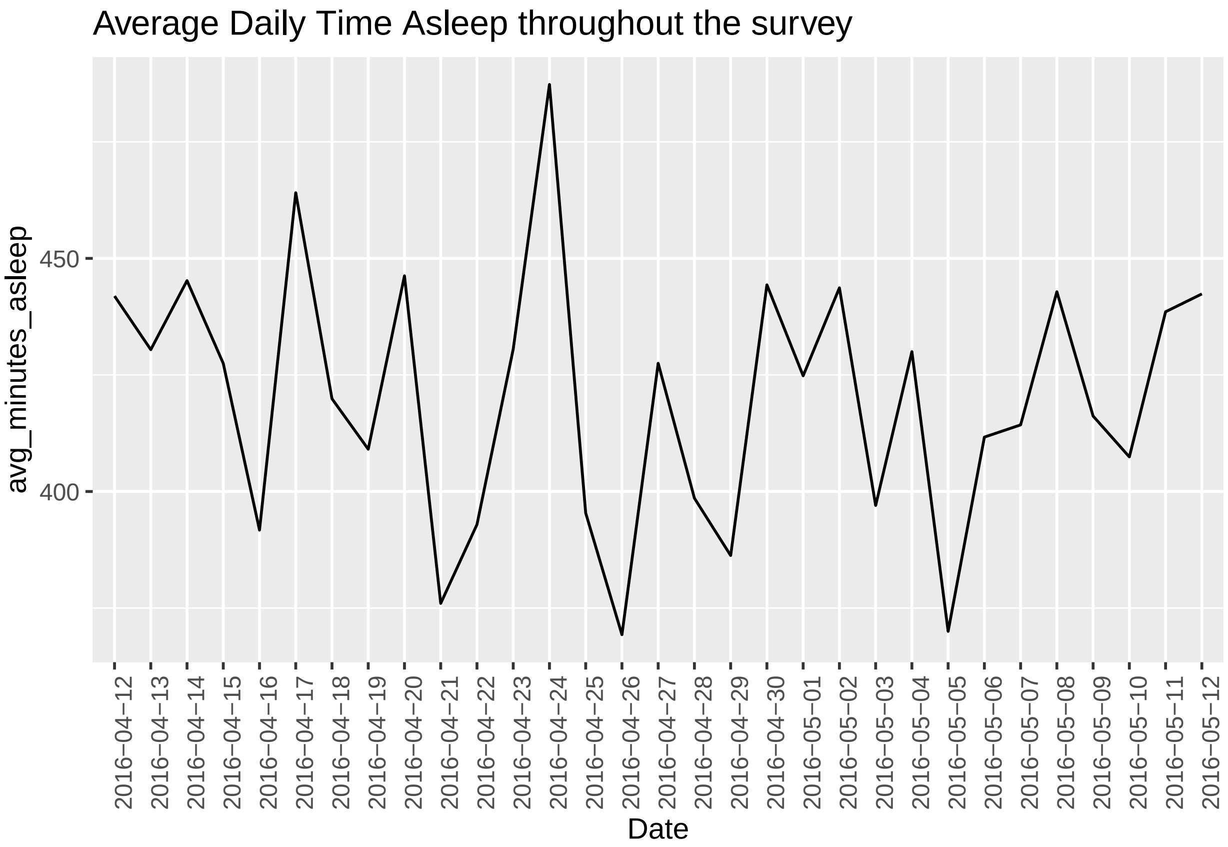
During the duration of the data collection period, there are small fluctuations in average steps taken. Throughout the 1 month, the last day is the lowest. A reason for this might be that users did not track their steps before submitting the data or the data was collected mid-day.

```
# Average Progress of Calories Burned
ggplot(data = by_date, aes(x = Date, y = Avg_calories, group = 1)) + geom_line() + theme(axis.text.x=ele
```



There is little fluctuation in calories burned. Calories burned drops drastically in the final day of data collection. The most likely reason is same as above which is the last day of data collection did not track full day.

```
# Average Progress of Daily Time Asleep
ggplot(data = by_date, aes(x = Date, y = avg_minutes_asleep, group = 1)) + geom_line() + theme(axis.tex
```



The average time asleep fluctuates more and most of the time below 450 minute except for 04/23 and 04/17. The remaining are lower than the recommended daily sleep time of 8 hours or 480 minutes. There is little to no progress in improving sleep time since the beginning of this data collection period. On the last day, a probable reason to why there is no drastic reduction in minutes asleep due to data submission is done during the day. Thus, only affecting calories and steps

Recommendations

Based on the the Fitbit health activity data, there are few recommendations to be applied to Bellabeat smart device:

Distinguish the user's sex

It is important to market differently toward man and woman. All the activity metrics that women do are different for men biologically. This should be implemented in the as a feature and and as an option Bellabeat app. As for the device, the Bellabeat can custom Leaf and Time to have larger diameter to accommodate for men's size.

Age input is important

Age is an important indicator for accuracy in health activity. Someone from 50-60 years old would have different heart rate, steps, calories burned and highly active minutes compared to someone from 20-30 years old. This is done to be more inclusive toward every generation.

Encourage more users to use the weight tracking

Despite small sample size, there is a considerable progress. Using a smart health device is helpful for people who wish to achieve their body weight goal. A notification in the app and a progress chart could motivate users to change their habits for a better body. Ensure privacy and protection of data, so that users are comfortable sharing their weight data.

Notification for too much sedentary minutes

Based on the data collected, users are most likely to be sedentary. High sedentary minutes is not only unhealthy, but also detrimental to mental wellness and linked to increase stress and poor sleep. If daily sedentary exceeds the recommended minutes, notify the user to take a walk outside, breathe fresh air and do some activities.

Reminders based on time

5-7 PM is the highest activity since it is the time when people go home from work. Based on the trends identified, Saturday has the highest activity. During these periods, Bellabeat can give Reminders to go for a run or walk.

Calories intake

Calories burned is not enough measurement if the intake offsets it. Track what the user is eating by making presets of food menu and their estimated calories in the app. Calories intake will be subtracted by calories burned to get net calorie. Users who wish to reduce their weight should have a calorie deficit.

Steps Goal and Intensity

The average total steps is lower than recommended and Bellabeat can motivate users to achieve their goal especially during weekend. However, there is likely that users prefer gym or sports with high intensity but lower displacement. Bellabeat Leaf or Time could calibrate users' heart rate, basal metabolic rate, accelerometer to give accurate judgement so that the users won't be notified to reach more steps if they have exercised.

Emphasize Sleep Pattern

The data shows that users get less than 8 hours of sleep. Encouraging the users to get more sleep when it is night time is important. Notify the users if the device detects that the users are in bed, but have not fallen asleep.

Incorporate weather and Climate Data

Hot weather and cold weather may impact health activity. E.g. during snowy days, users should not be expected to take a run outside.

Reconduct Survey

The data used is poor quality and has many limitations. Making decisions based on this data alone is not recommended. Conducting further survey with larger samples, longer time span, more information collected such as age, sex and occupation, climate region may improve reliability and relevancy of data.

Thank you

I hope this insight and recommendation will help in making executive decisions for Bellabeat's marketing strategy.