Google Data Analyitics Professional Certificate
Final Assignement

Vincenzo Junior Striano

Summer 2022

Abstract

In the summer of 2022, I earned a professional certification in Data Analytics from Google. At the end of that challenging course, I was required to complete analytics similar to those required in a real-world scenario. I investigated, therefore, the usage trends of a fitness tracker, providing insights about customers.

Contents

1 Data Collection and Tools

The FitBit Fitness Tracker Data from Mobius, on Kaggle contains fitness-related data FitBit tracker's users. Thirty eligible Fitbit users consented to the submission of personal tracker data, including minute-level output for physical activity, heart rate, and sleep monitoring. It includes information about daily activity, steps, and heart rate that can be used to explore users' habits.

I've decided to analyze the following data sets:

- weight, by day
- steps, by day
- heart rate, by second
- sleep hours, by day
- burnt calories, by day
- activities, by day

R studio will be used to complete the analysis.

2 Exploratory Data Analysis and Data Visualization

For this section, each dataset will be treated independently. Generally, there are at least two columns in each data set: one referring to the athlete id and one referring to the monitoring period. The other columns, peculiar to each dataset, contain information about the measured data.

2.1 Calories

Calories dataset shows the daily amount of calories burnt by 33 people for 31 days starting from the 12th of April, 2016.

- *Id*, the identification code of the person.
- *ActivityDay*, the day of the measurement.
- *Calories*, number of calories burnt.

Using the powerful command *skim_without_charts* by *skimr* package, we found out that the dataset contains 940 observations and 3 variables: as expected, 31 day, 0 missing values, and the following statistical distribution for the amount of calories:

Mean	sd	Q0	Q0.25	Q0.5	Q0.75	Q1
2304	718	0	1828	2134	2793	4900

Table 1: Statistical distribution of calories burnt

Unfortunately, it's not possible to realize meaningful analysis based on the dataset alone. Better ones can be performed merging datasets containing different information. That is because of different reasons:

- the Basal Metabolic Rate, or the number of calories burnt during a day spent without moving depends, depends on age (N/A), sex (N/A), height (N/A) and weight. Let's take for reference 1634kcal/day, as the BMR for a 40-year-old man, 175cm, 70kg using the Harris-Benedict formula. [1]
- To BMR, it's necessary to add the number of calories burnt for everyday activities which depend on the lifestyle of the person. Of course we can make an estimate, using *intensities* datasets we have.
- Losing or gaining weight (either fat or lean muscles) depends on the difference between the calories intake (N/A) and the calories burnt. If *net calories* < 0, they will lose weight; viceversa, they'll gain weight. We don't have information about weight goal and calories intake.

Given that, assume that:

- On average, BMR represents 70% of daily caloric expenditure for weight maintenance.[2]
- So, the Total Day Energy Expenditure (TDEE) is 2340kcal.
- All users have a TDEE of 2340kcal.
- All users want to lose weight, so they need to burn more than 2340kcal.

2.1.1 How they performed in losing weight? (see Figure 1)

The dashed line represent the TDEE, so, on average, people should burn calories above that line. Only few people accomplish that goal. What a bad! Let's greet *Mara, id = (...)391*, our second user from the right side of the barchart. It burns a lot of calories!

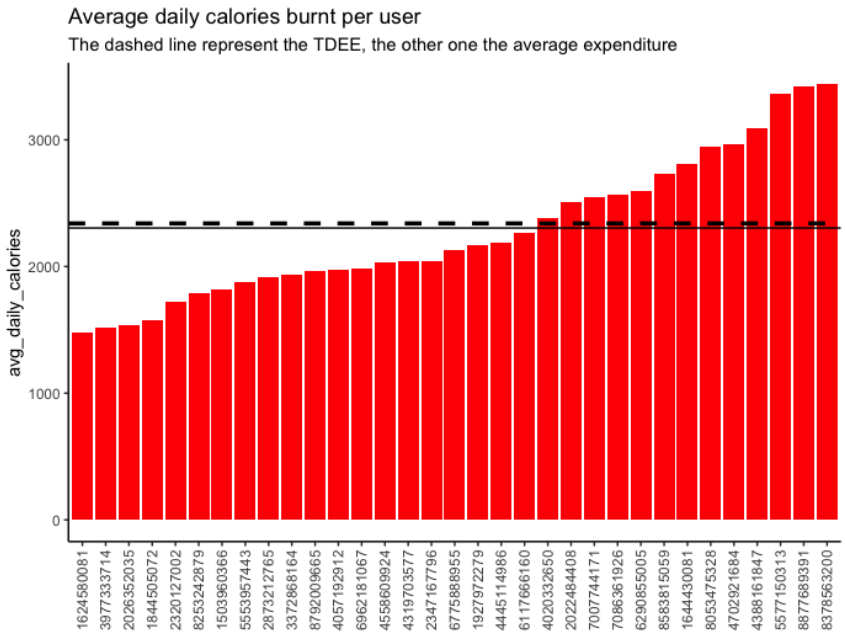


Figure 1: Average daily calories burnt per user

2.1.2 How calories expenditure changed over time? (see Figure 20)

. First thing first: the matrix is not sparse, so people use the calories expenditure function, which makes sense, because it doesn't require to be manually activated. The reddish, the more calories spent: again *Mara*, on the right side, is a pure athlete!

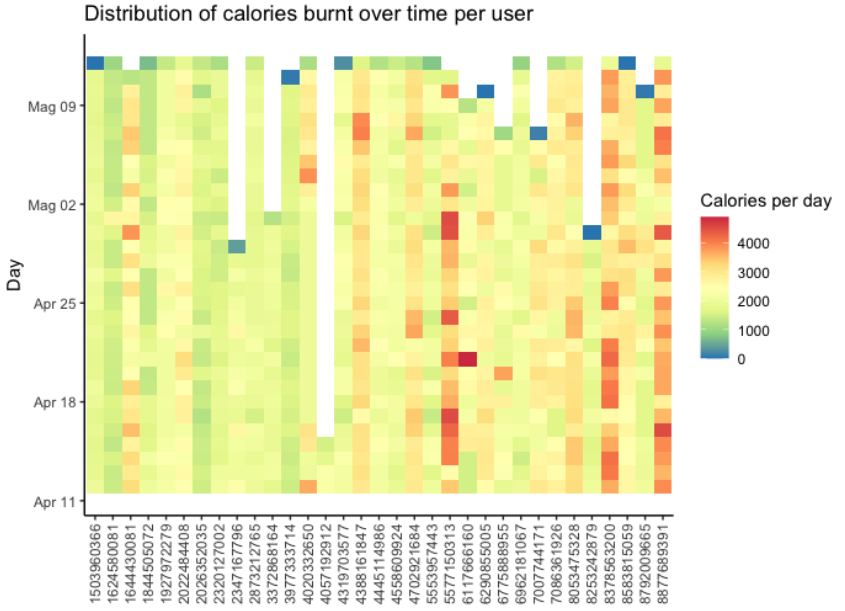


Figure 2: Calories consumption over time per user

2.1.3 How calories expenditure changed over the weeks? (see Figure 3)

. The trend is quite stable over the period and the weekdays. On Tuesday the 10th of May, a number people burnt a lot of calories. Maybe they attended a challenge, like a little marathon? People tend to rest on Sundays.

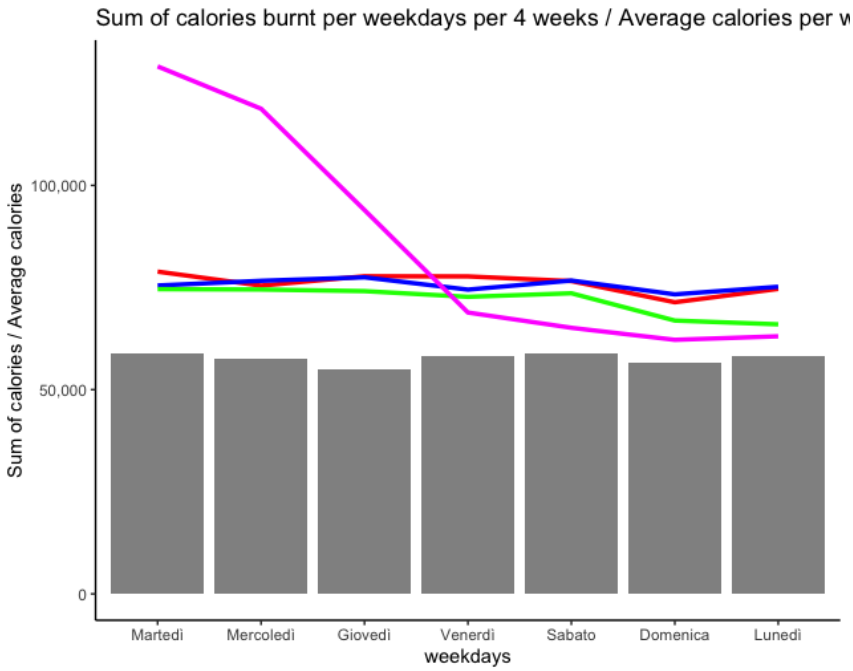


Figure 3: Calories consumption over weeks

2.1.4 Appendix: Profiling users basing on their performance

Profiling of users who subscribe to a service makes it possible to improve the targeting of advertising campaigns for the acquisition of new customers and to provide existing ones with services that are more appropriate to their characteristics.

We can tell the that most of users burns between [1586, 3022] calories during her activity day: I'll name this persona as *sporty*, mostly between 2000kcal and 3000kcal. Few people, burn more than 3022, those will be the *athletes*. The ones between 0 and 1828 Q0.25 will be the *couch potatoes*.

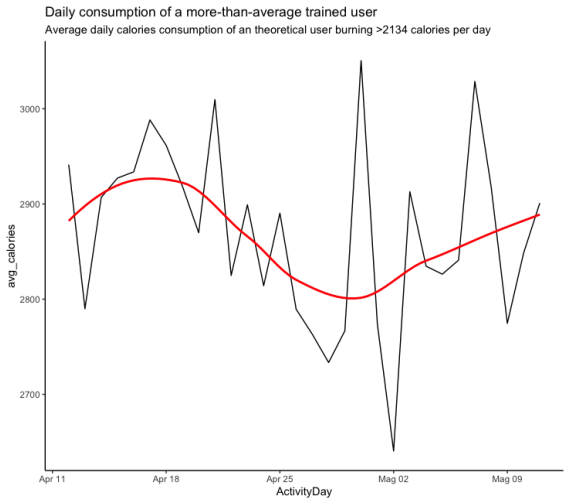


Figure 4: Daily average consumption of the *athlete* persona

The *athlete* persona shows a quite stable trend in the daily calories consumption over a one-month period of time, with pikes over 3000 and slightly less than 2700. That may tell that *athletes* are dedicated and consistent in their training. Moreover, their workouts are intensive.

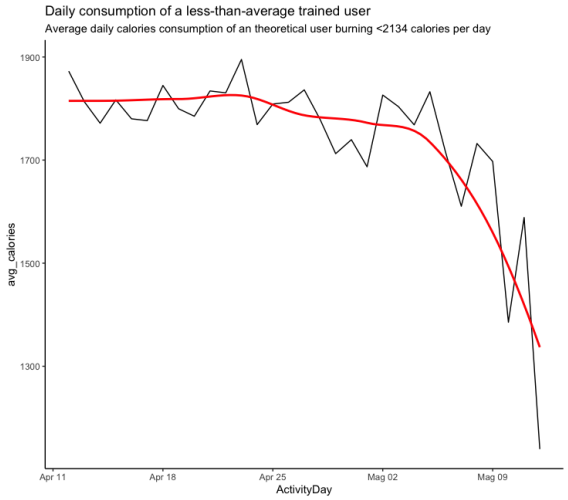


Figure 5: Daily average consumption of the *sporty* persona

The *sporty* persona shows an unstable trend with peaks of 1900kcal and less than 1300kcal. They may be inconsistent and workouts are easier.

2.2 Sleep

Sleep dataset contains information about sleep habits of 24 users in a month long, starting from the 12th of April.

- *Id*, the identification code of the person.
- *SleepDay*, the day of the measurement.
- *TotalTimeInBed*, number of minutes in bed, sleeping, reading, etc..
- *TotalMinutesAsleep*, minutes sleeping over the day and the night.
- *TotalSleepsRecord*, number of sleeps, naps included.

Using *skim_without_charts*, I found out that the dataset contains 413 observations and 5 columns: as expected, 31 day, 0 missing values, 24 unique users. Moreover, people sleep 419 minutes on average + 40 spent on bed.

Value	Mean	sd	Q0	Q0.25	Q0.5	Q0.75	Q1
TotalSleepRecords	1.12	N/A	1	1	1	1	3
TotalMinutesAsleep	419	1.18e+2	58	361	433	490	796
TotalTimeInBed	459	1.27e+2	61	3403	463	526	961
delta_minutes	39.2	4.66e+1	0	17	25	40	371

Table 2: Statistical distribution of sleeping minutes

Immediately, the variable delta_minutes was realized, given by the difference between total minutes spent in bed and actual sleep minutes: this is the number of minutes of inactivity spent on the cell phone, watching TV, reading, etc. The data are consistent because delta_minutes is never minor than 0. Also, the variable weekday, turns the date into numbers on the respective weekday.

2.2.1 How people's sleep changed over time? (see Figure 6)

Unfortunately, the matrix is quite sparse: people tend not to use the sleeping monitoring function. However, people seem to sleep enough.

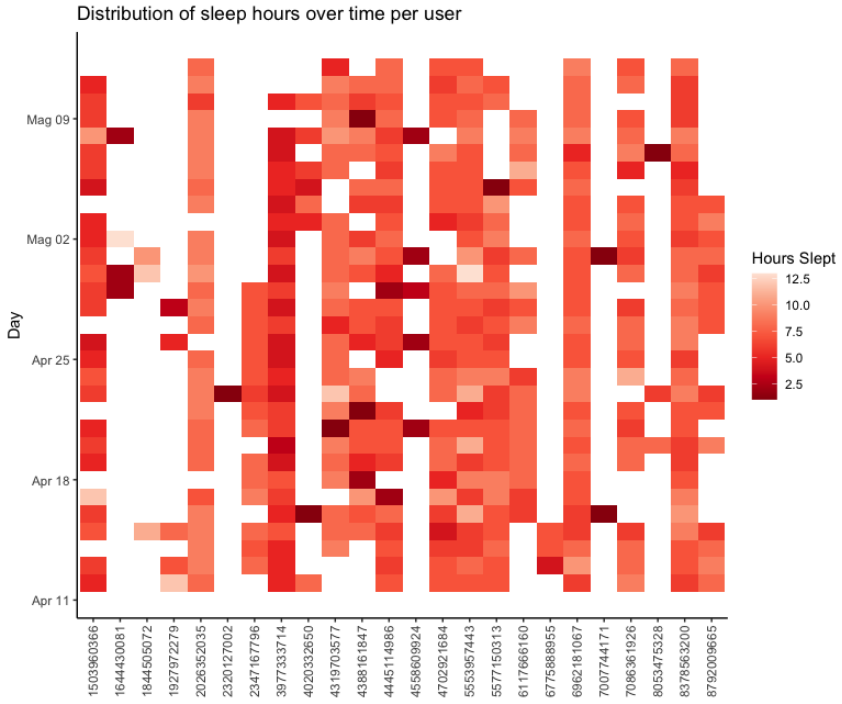


Figure 6: Hours slept per person over time

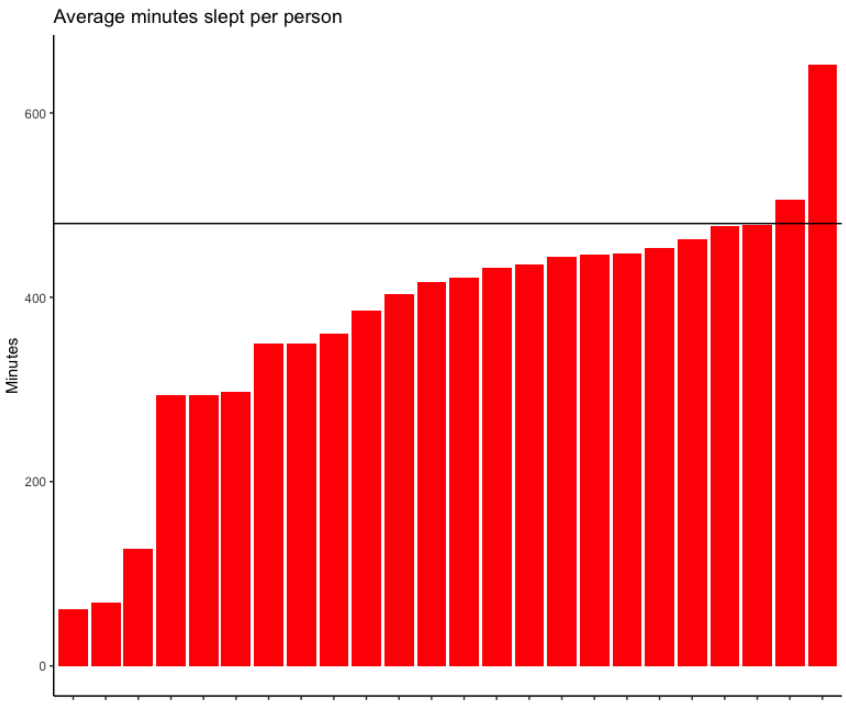


Figure 7: Average minutes slept per person

Next, the mean of minutes asleep daily per each user is computed: **is there any correlation between calories burnt and minutes slept?** This question will be answered later. Meanwhile, we can tell that a lot of people sleep about 420minutes per night. Few, less than 200, and few more than 500 (see Figure 7). **However, on average, most of people sleep less than the recommended 8 hours** (horizontal line)![3]

The difference is quite constant among the week (the bold line is the median for each box plot and the red box is related to the distribution of total time in bed, which is always higher than the time asleep). On Sunday people tend to sleep more (see Figure 8).

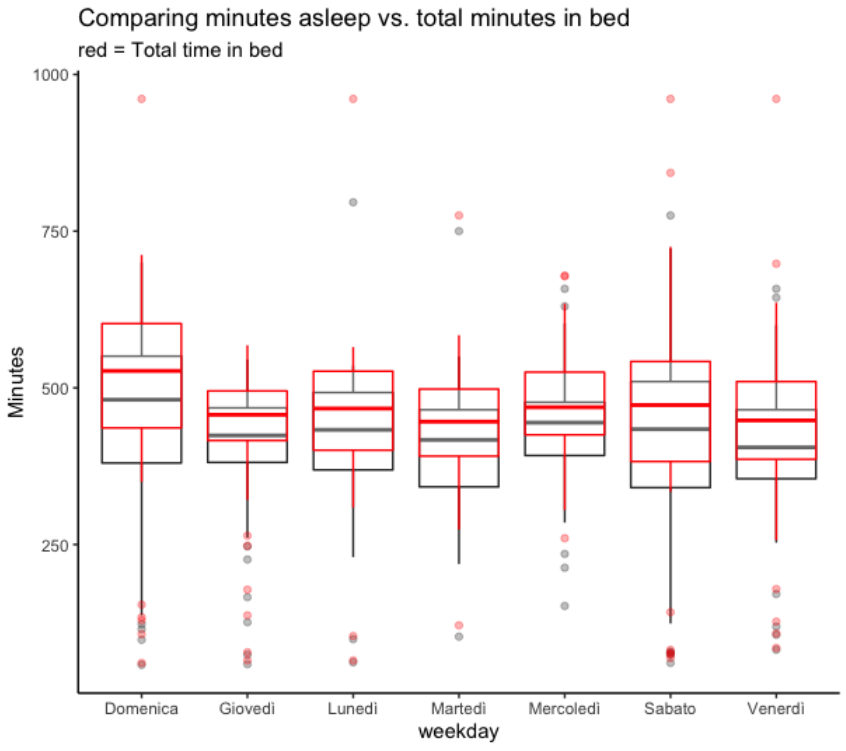


Figure 8: Sleeping minutes distribution among weekdays

2.2.2 How delta_minutes occurs among weekdays?

As already said, as mentioned, delta_minutes, in most cases, falls in the range (0-90 minutes), with an average of 40 minutes. On Sundays and Fridays, sporadically, delta_minutes even reaches up to 400 minutes. Someone may have run a marathon or had a hangover (see Figure 9).

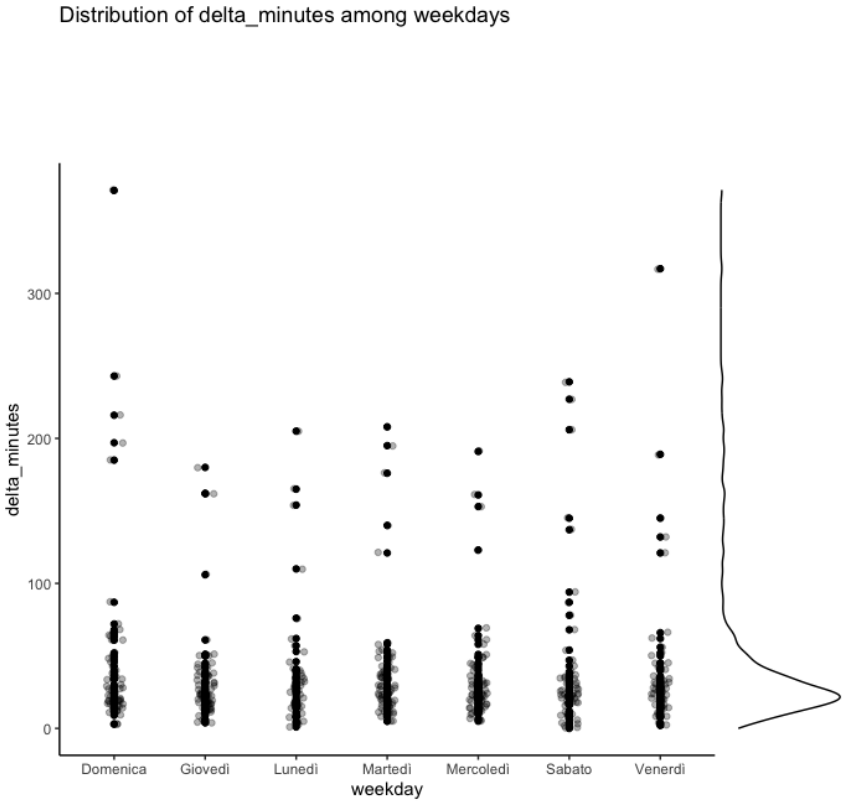


Figure 9: Distribution over the week of delta_minutes

The Pearson Correlation between the variable *TotalMinutesAsleep* and *TotalMinutesInBed* is 93%.

To avoid collinearity problems, even considering the low extent of delta_minutes, only the variable *TotalAmountAsleep* will be considered later.

2.3 Weight

The *Weight* dataset contains 67 records and 8 variables about daily weight measurements of 8 people in 31 days.

- *Id*, the identification code of the person.
- *Date*, measurement dates.
- *WeightKg*, the weight in kilograms.
- *WeightPounds*, the weight in pounds. **Eliminated** because perfectly correlated to *WeightKg*.
- *Fat*, the body fat percentage. **Eliminated** due to the poor number of measurements available.
- *BMI*, or Body Mass Index. **Eliminated**, because highly correlated to *WeightKg*.
- *IsManualReport*: *True* or 1, when the person manually inserted the weight in the FitBit app, *False* or 0 when the smart scale automatically updated the app via cloud.

The average weight measured is 72kg, which is below the average of a north american person. [5]

The variable *IsManualReport* shows also the percentage of people having a smartscale: 37.5%. A small number, a I'd say: Is the price too high?

Thanks to summarize, a new table is created:

Id	count	meankg	ismanualreport	weightdifference	delta_date	weight_over_days
6962181067	30	61.6	1	1.5	30 days	0.05
8877689391	24	85.1	0	1.8	30 days	0.06

Table 3: Summary table of the two best-performing people

- *count* tells how many records are available per Id.
- *meankg*, the average weight measured.
- *ismanualreport*, constant, per Id.
- *weightdifference*, the difference in kilograms between the highest weight measured and the lowest.
- *delta_date*, the difference in days between the last measurement day available and the first one.
- *weight_over_days*: the ratio between *weightdifference* and *delta*. The higher the ratio, the more effective the mealplan is (in case they're following a diet, either for gaining or losing weight).

2.3.1 Insights

The Pearson's correlation between *count* and *meankg* is -18%: there's a slight, inverse correlation between the weight and the number of times a person uses the scale. **The more you weigh the less you're interested in knowing it**.

The Pearson's correlation between *ismanualreport* and *meankg* is -80%: the leaner you are, the higher the chance to have a smartscale. That makes sense: **if leaner people are the most fit, they must and have interest to track precisely their parameters with advanced tools**.

The Pearson's correlation between *count* and *ismanualreport* is negligible: **having or not a scale does not affect the number of times you weigh yourself. Dedication only matters**.

2.3.2 Is it better to weigh yourself daily or weekly?

Many doctors recommend weighing yourself weekly, trying to recreate the same conditions each time to avoid fluctuations. Weight changes, therefore, will only be appreciable in the long run. [6]

Based on the available data, **I think it is more useful to weigh oneself every day and draw a trendline in the short run that projects in the long run what will happen: fluctuations will be eliminated statistically -and not randomly, as in the weekly approach- and the psychological reward for the athlete will be almost immediate**.

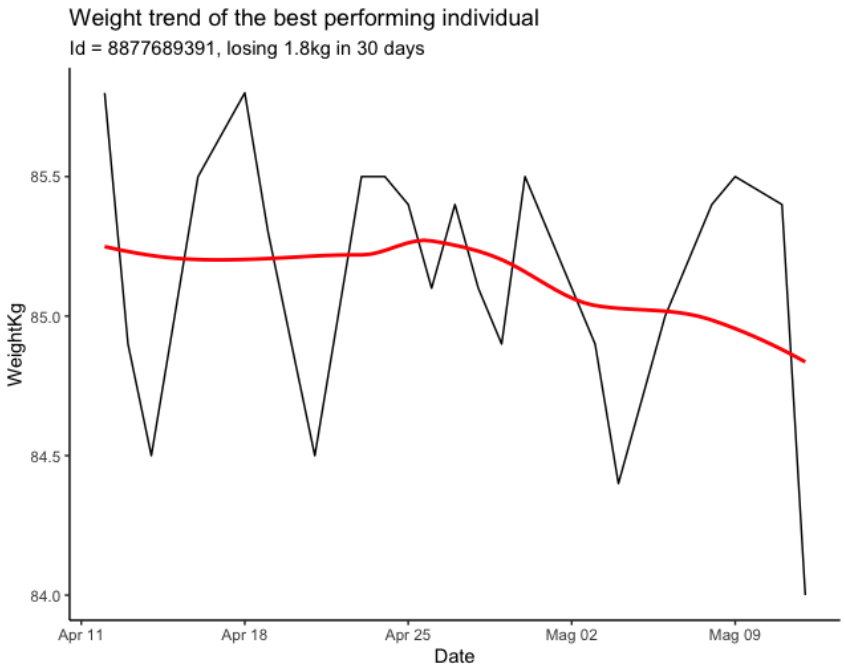


Figure 10: Weight trend of the best performing individual: -1.8kg in 30 days

2.4 Steps

The *Steps* dataset contains 940 records and 4 variables about daily steps measurements of 33 people in 31 days starting from the 12th of April, 2016.

- *Id*, the identification code of the person.
- *ActivityDay*, the day of the measurement.
- *StepTotal*, the number of steps.
- *weekdays*, a custom variable which "translate" the date from number to week day, like Monday, Tuesday etc.

On average, 7600 daily steps, users walk less then recommended by doctors to have for an healthy lifestyle [4].

Thanks to this *geom_tile*, the distribution of steps over time per user is displayed:

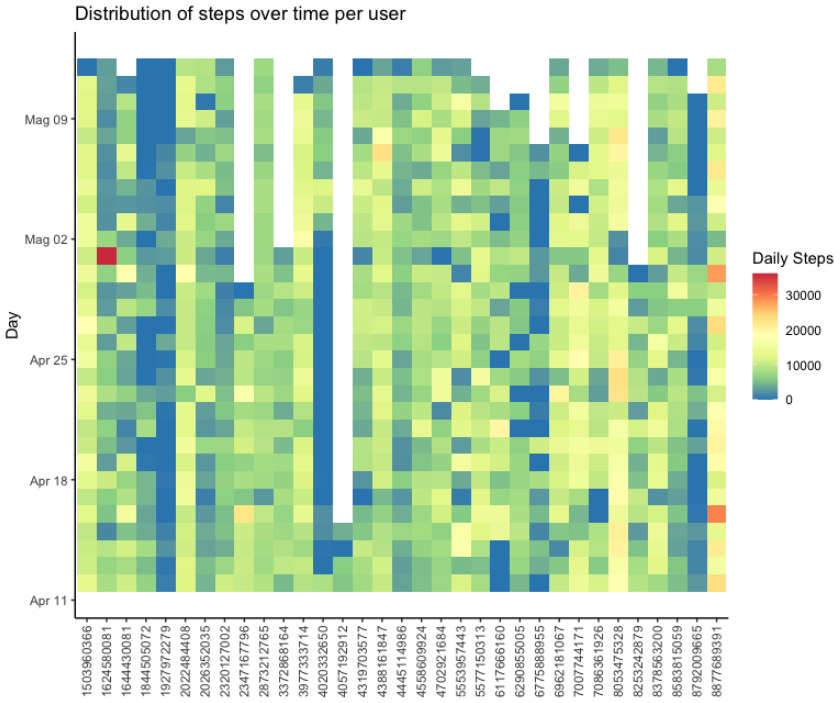


Figure 11: Distribution of steps over time per user is displayed.

2.4.1 Insights

White tiles are days when no records were available. Blue tiles are days where people walked 0 steps. What a bad! People with green, yellow and reddish tiles, are the most active, as the last user on the right of the plot, namely *Mara*.

Looking at the average number of daily steps per person (see Figure 12), *Mara* is also the one with the highest average steps number per day. Moreover, the dashed line represent the mean of the steps recorded, and the straight one the smallest number of steps associated with an healthy lifestyle: **People should definitely walk much more!**

2.4.2 When people usually walk?

Saturdays, Mondays, and Tuesdays. On Tuesday the 10th of May, a number people walked much more then usual: we can definitely see a positive correlation between steps and calories burnt. Each line represents the trend in a week, from the 12th of April to the 12th of May. The height of the line shows the total number of steps over the weeks. The columns represent the average number of steps per week day (see Figure 14).

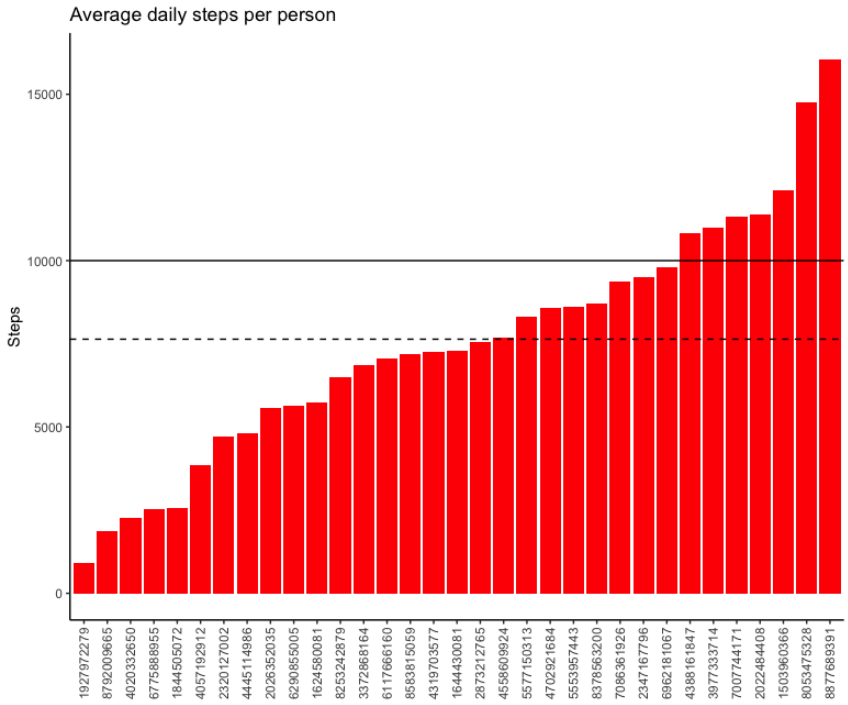


Figure 12: Average daily steps per person.

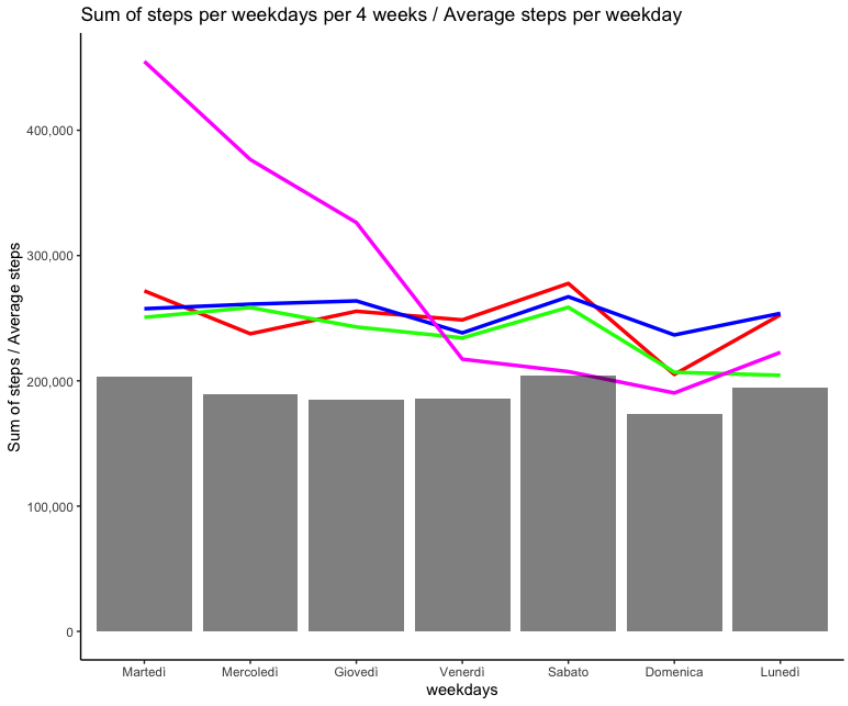


Figure 13: Steps over weekdays and average steps per weekday.

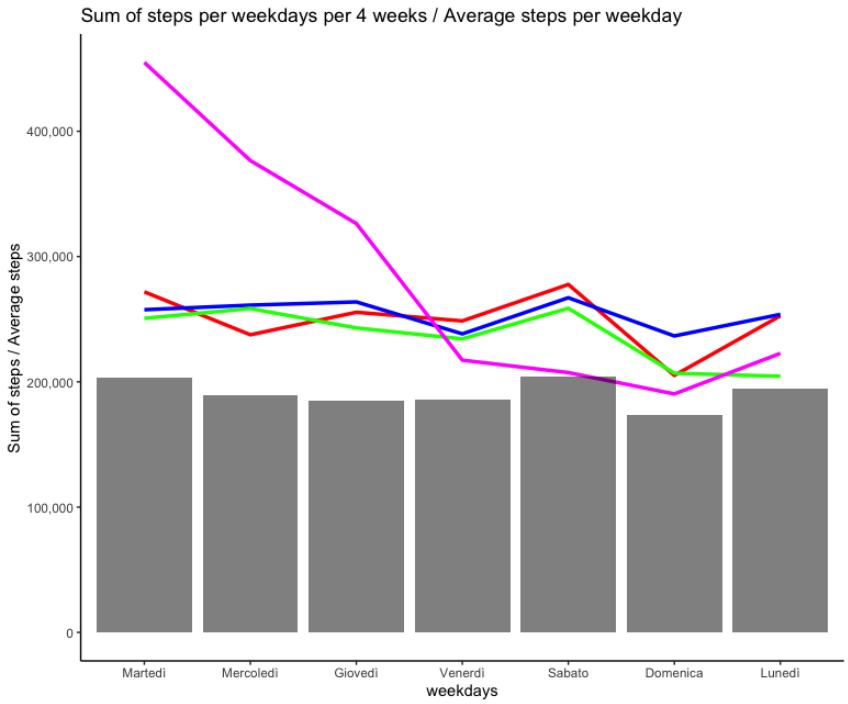


Figure 14: Steps over weekdays and average steps per weekday.

2.5 Heart Rate

The heart rate dataset contains heart rate measurements every 5/10 seconds for each person throughout a month, starting from the 12th of April, 2016. It contains several columns:

- *Id*, the identification code of the person.
- *Time*, date and time of the measurement.
- *Value*, BPM measured.
- *hours_brief*, hour of measurement, **custom created** (0,1,2,3...23)
- *weekdays*, weekday of the measurement, **custom created** (*Monday*, *Tuesday* etc.)
- *Time_of_day*, Based on the time of the measurement, it tells in which part of the day the measurement occurred (*Morning*, *Evening* etc.) **custom created**.

Thanks to *skim_without_charts()* we can tell that the dataset contains 2483658 observations with 3 columns (+3, custom created). 14 people shared their data from the 12th of April 2016, for 31 days. The average heart rate measured is 77.3BPM.

2.5.1 Addressing the nighttime bias:

First, **it must be said that the dataset is unbalanced because the nighttime measurements are about half as large as the daytime once: BPM will be negatively skewed so especially the medical considerations will be biased. I've investigated the reason of this bias looking at the next plot**:

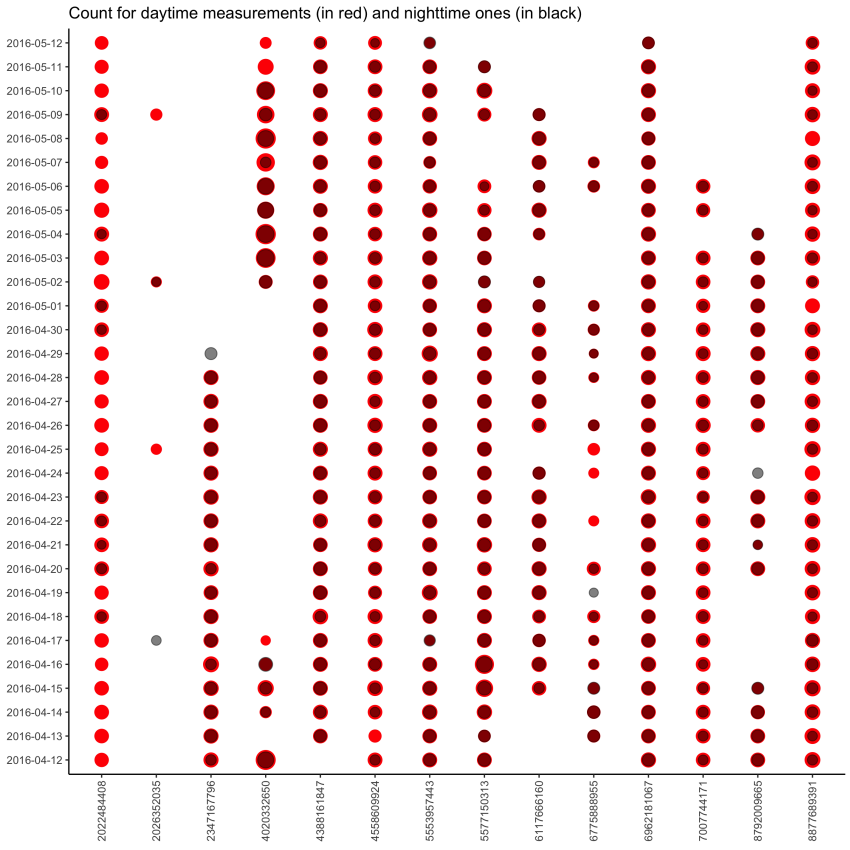


Figure 15: Count for daytime measurements (in red) and nighttime ones (in black)

Each column represent an Id, and each row, a day. Red bubbles relate to daytime measurements (7-22); black ones, to nighttime measurements (22-7). Their size, shows the count.

During the month, **most people wore their FitBit almost every day and night**: you can see that there are few days when there is no circle (in which case, no measurements would be available that day) and the circles often overlap perfectly. However, the red circles are always larger than the black ones: this means that the **FitBit measures beats less often at night**

(for power-saving reasons, probably), *see user -7443 and -1067*. In the case of the first user on the left *-4408*, it wears the FitBit only during the day and sporadically at night. We cannot yet draw any conclusions about the second user from the left *-2035*, of whom we do not know whether she occasionally wears the FitBit or does not take advantage of the heart rate monitoring function.

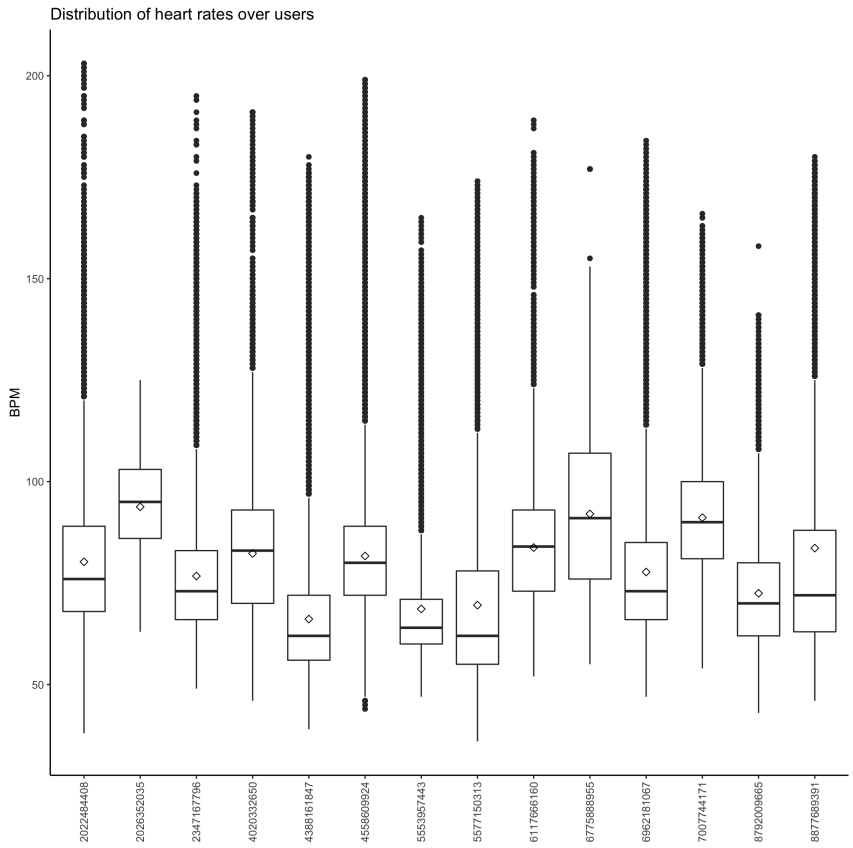


Figure 16: Heart rates distribution over users

2.5.2 Are people active and healthy? (see Figure 16)

Each boxplot represents the distribution of BPM for each user. The median of the BPM is the bold dash in the center of the box, the small rhombus is the average BPM. The cluster of dots above each boxplot are the outliers.

Outliers are numerous, suggesting that people are very active. However, given the peaks of more than 200BPM, people should consult a doctor to make sure they can heavy exercise without health risks.

All people show an average of beats between 60 and 100 per minute, which does not suggest cardiological pathology.[7]

2.5.3 When people train? (see Figure 17)

The *geom_tile* shows the average BPM per hour, per weekday. During the night, the average is about 60BPM, because people sleep. During the day, the mean is 75, which is a reasonable average heart rate. Between 4PM to 7PM, the average moves up to 85BPM, during all the weekdays: that may suggest that **people usually train after work**.

On Saturday night, BPM are higher than usual because people usually go out or dance.

On Saturday at 1PM, the Average BPM is higher because a lot of people trained those days at that time, especially the 23rd of April. *Mara*, always, confirms herself as one of the top athlete.

Our hypotheses are confirmed by the plot showing the number of measurements ¿120 (which I arbitrarily considered as the "training" threshold) taken during the 24-hour period: there are peaks at 9, 12 and 18. (see Figure 18)

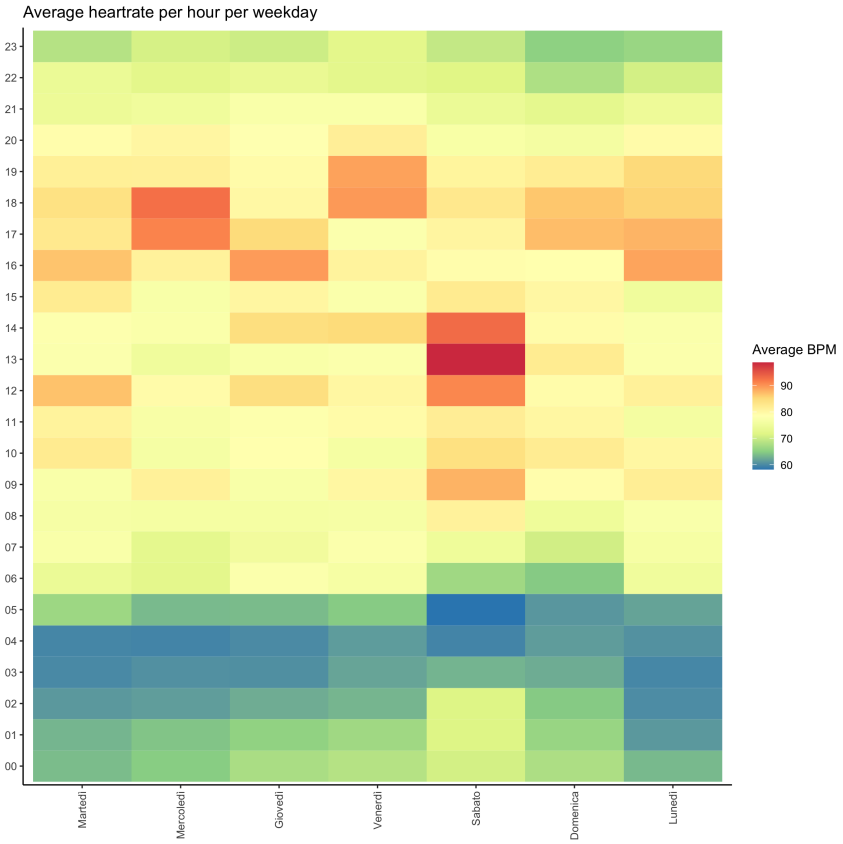


Figure 17: Heart rates distribution over users

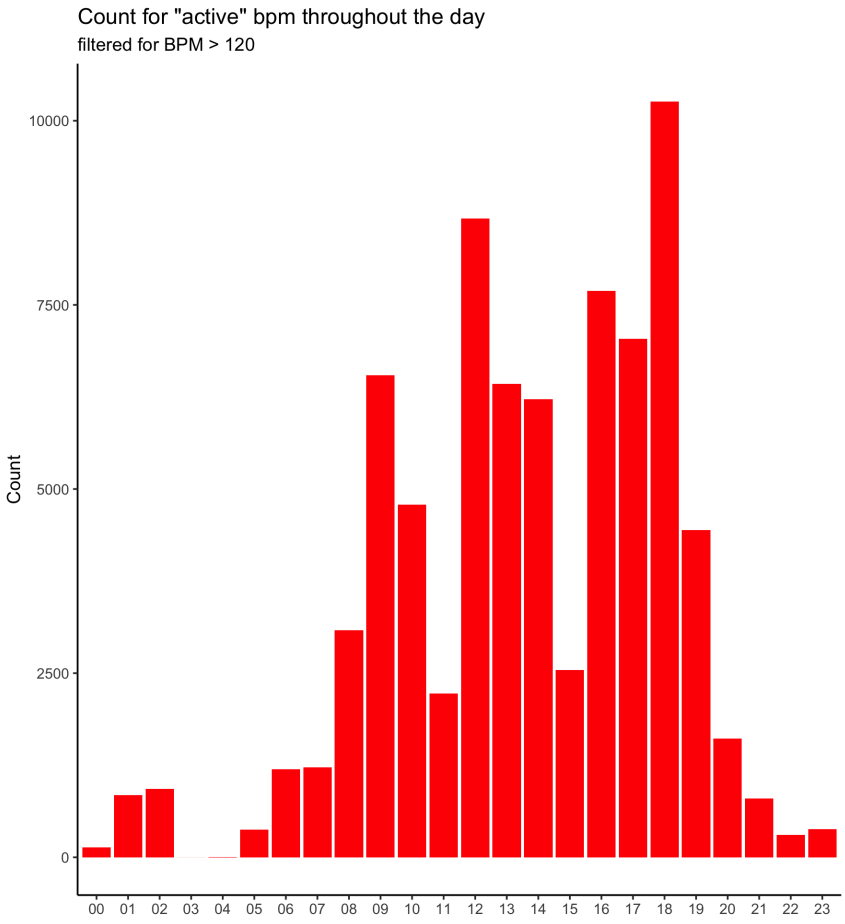
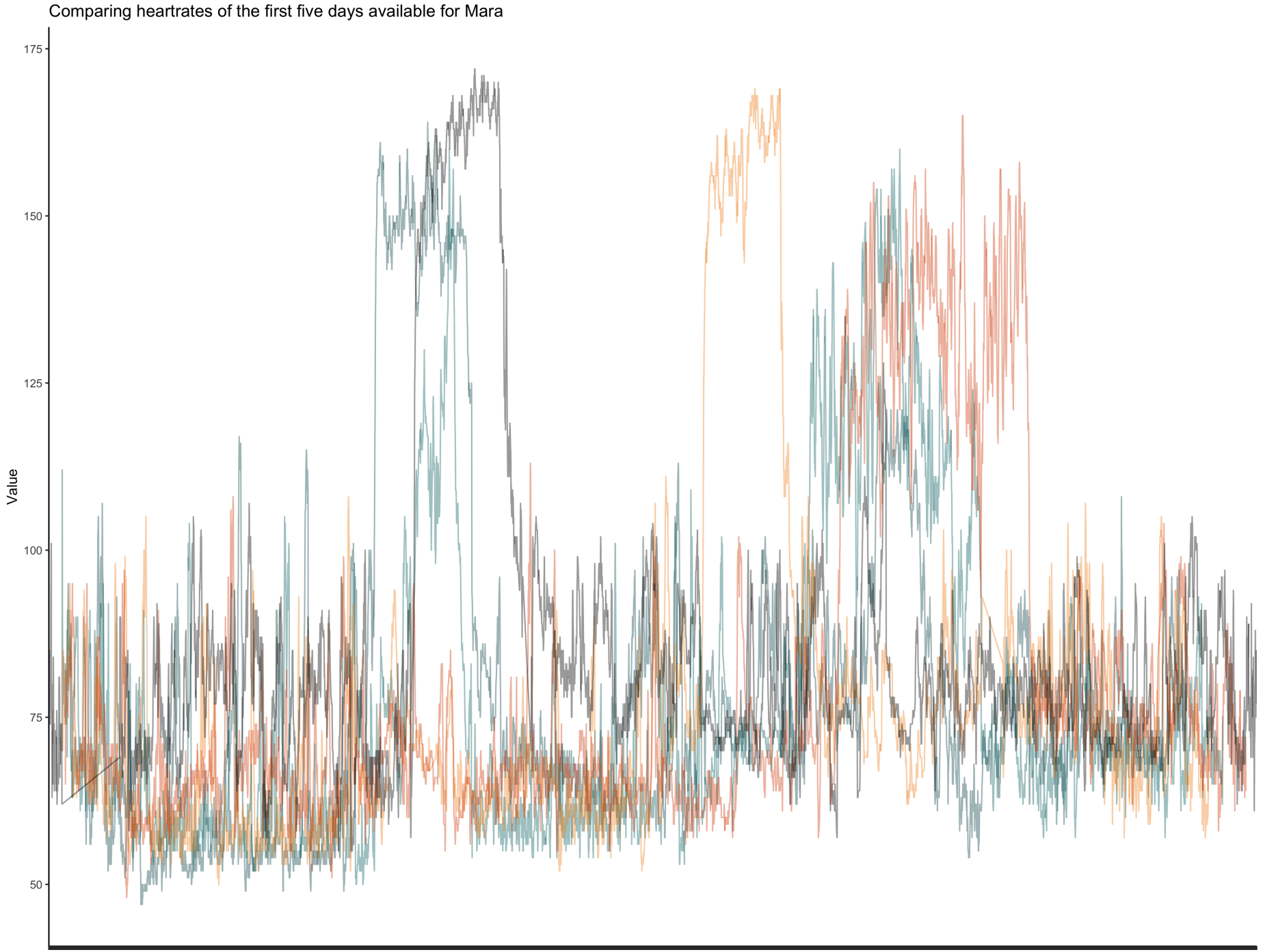


Figure 18: Count for "active" BPM throughout the day

2.5.4 Mara heart rate (see Figure 2.5.4)

Below, you can see the heart rates of *Mara* throughout the 24 hours of the first 5 days of measurements available. Days are represented by lines of different colors. Peaks occur slight before the midday and during the afternoon and the average is around 75 BPM, both in line with the average.


Comparing heartrates of the first five days available for Mara
Value
175
150
125
100
75
50

2.6 Activities

The activities dataset reports information on how physical activities were performed day by day. It contains:

- *Id*, the identification code of the person.
- *ActivityDate*, date of the measurement.
- *TotalSteps*, number of steps.
- *TotalDistance*, number of kilometers recorded (active recording + passive recording).
- *TrackerDistance*, number of kilometers recorded (active recording).
- *LoggedActivitiesDistance*, unexplained.
- *VeryActiveDistance*, number of kilometers recorded (active + passive recs.) *where* BPM > 160 *assumption.
- *ModeratelyActiveDistance*, number of kilometers recorded (active + passive recs.) *where* 120 < BPM <= 160 *assumption.
- *LightActiveDistance*, number of kilometers recorded (active + passive recs.) *where* 85 < BPM <= 120 *assumption.
- *SedentaryActiveDistance*, number of kilometers recorded (active + passive recs.) *where* *assumption.
- *VeryActiveMinutes*, number of minutes recorded (active + passive recs.) *where* BPM > 160 *assumption.
- *FairlyActiveMinutes*, number of minutes recorded (active + passive recs.) *where* 120 < BPM <= 160 *assumption.
- *LightlyActiveMinutes*, number of minutes recorded (active + passive recs.) *where* 85 < BPM <= 120 *assumption.
- *SedentaryMinutes*, number of minutes recorded (active + passive recs.) *where* BPM <= 85 *assumption.
- *Calories*, number of calories burnt.

This was certainly the most challenging dataset because the data seemed insignificant for analysis. However, the findings will be laid out below:

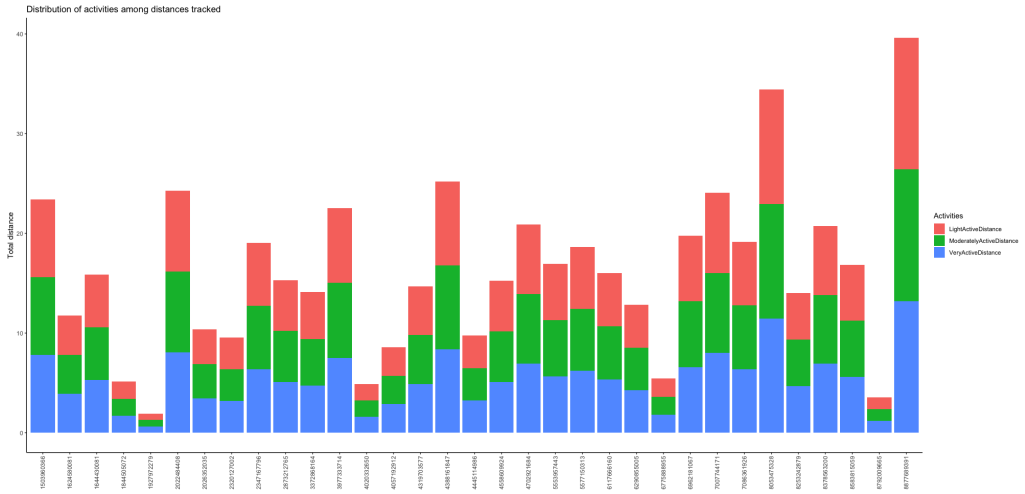


Figure 19: Distribution of *intensity* in the average distance walked

Mara, on the rightmost position, walks a lot with an equal distribution of exercise intensities. On average, during a day, people walk 62% of the total distance covered: 10%, doing moderate exercise and 27% running (see Figure 19).

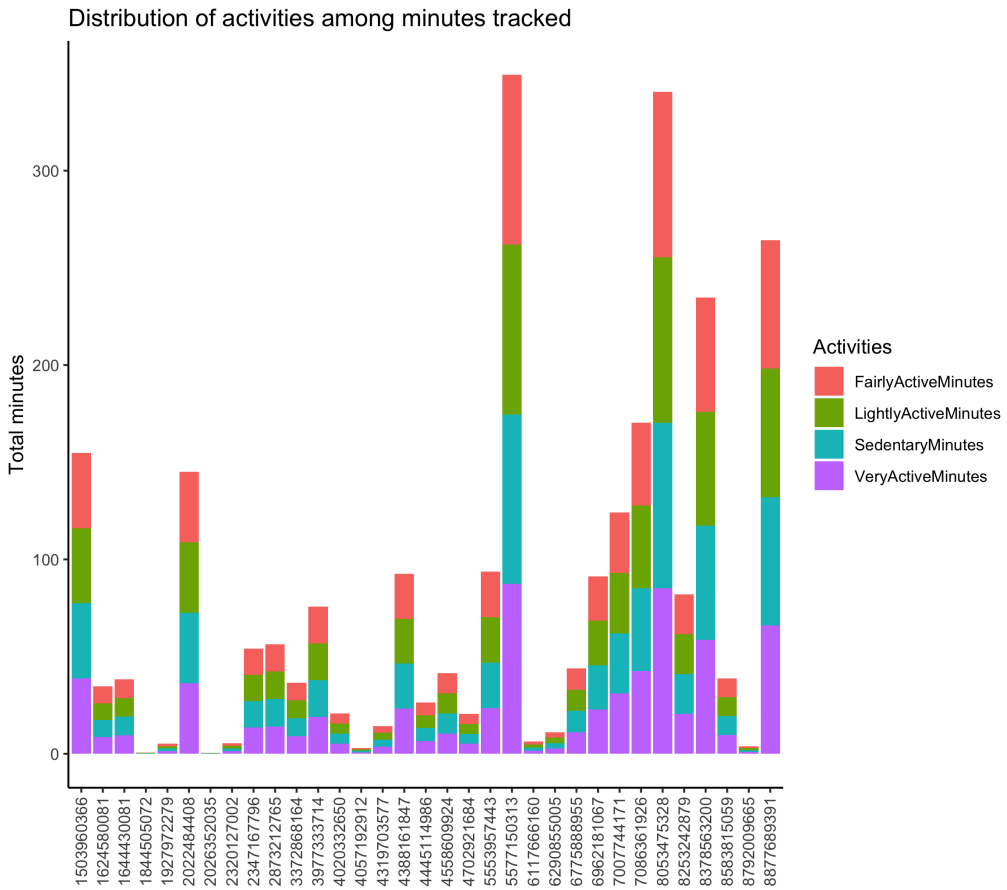


Figure 20: Distribution of *intensity* in the minutes tracked

On average, over the 1226 minutes tracked, only during the 1% a *very active* workout has been performed. Same said for the *FairlyActiveMinutes*. An average exercise is accomplished 15% of the time tracked. A sedentary activity, 81% of the time.

The correlation between *Calories* and *TotalSteps* is 59%, meaning that part of the calories must be burned through a steady activity, such as the gym.

The correlation between *TotalDistance* and *TotalSteps* is 98.53%: **FitBit's athletes go running and don't use a treadmill**.

3 Conclusions

- People tend to use their FitBit Tracker during all the day and night.
- Most of people using their FitBit usually walk and run. Someone also trains at gym.
- People tend to train after work.
- People only buy their FitBit Tracker and not other smart gear.
- On average, they have an healthy lifestyle, based on number of calories burnt, steps, heart rate and hours slept. However, they can improve their lifestyle walking much more.
- Only few people are real athletes, with noticeable performance.

4 Spin-offs

In order to obtain even more interesting insights, it may be necessary to merge multiple datasets, based on the dates and IDs of the athletes.

From a single unified dataset, it will be possible to cluster your clients based on their common characteristics, implementing, for example, programs specific to their level, or consultations with specialists who are informed of their progresses thanks to smart devices.

In addition, by knowing the typical characteristics of one's customers, it will be possible to carry out marketing campaigns targeting potential new users who have the same characteristics.

References

[1] Bmr calculator. https://www.omnicalculator.com/health/bmr. Accessed: 2022-06-23.

[2] Estimation of basal metabolic rate in chinese: are the current prediction equations applicable? https://nutritionj.biomedcentral.com/articles/10.1186/s12937-016-0197-2. Accessed: 2022-06-23.

[3] How many hours of sleep do you really need? https://www.healthline.com/nutrition/how-much-sleep-you-need. Accessed: 2022-06-23.

[4] How many steps should people take per day? https://www.medicalnewstoday.com/articles/how-many-steps-should-you-take-a-day. Accessed: 2022-06-22.

[5] Quanto pesa la popolazione mondiale? https://www.focus.it/cultura/curiosita/quanto-pesa-la-popolazione-mondiale. Accessed: 2022-06-21.

[6] Weigh yourself guidelines #1. https://www.healthline.com/health/fitness-exercises/weigh-yourself-guidelines#1.-Weigh-yourself-once-a-week. Accessed: 2022-06-22.

[7] What your heart rate is telling you. https://www.health.harvard.edu/heart-health/what-your-heart-rate-is-telling-you. Accessed: 2022-06-24.