Data visualization : Milestone 3

AJEGHRIR Mustapha
mustapha.ajeghrir@epfl.ch
(Sciper 333806)

Sorin-Sebastian Mircea
sorin.mircea@epfl.ch
(Sciper 306618)

Simon Dayer
simon.dayer@epfl.ch
(Sciper 271991)

I. Goal

The goal of this project is to create a website that will give the user a good understanding of his/her fitness progress through data collected via a smartwatch.

With this in mind, we first focused on constructing exploratory plots that will give a general summary over all the collected data (this includes both activities and sleep data).

Then, our focus shifted towards finding different correlations and patterns in the activities made by the user, more specifically we wanted to find out how sleep influences and is influenced by doing various sport activities.

Lastly, we explored more broad temporal patterns, i.e the time spent doing different types of sport during the year.

II. Overview of the user data

The first challenging idea was to define what a sport oriented person would be interested to see (from a data visualisation point of view). So, after doing the proper research we came to the conclusion that an overview of all the activities is a must have.

A. Initial geoplot

This is why we decide to first let the user see how its data looks on the world map. In our case, the data source of all visualizations have already been explained in the 1st milestone (it is collected by one of our team members).

Fig. 1: Activities locations

In this graph, we were able to experience what we learned in the Maps lecture.

In order to make this plot, we wrote a Python script that iterates through all the activities and fetches one location point per activity, then we appended these two extra columns (start_latitude and start_longitude) to the activities summary CSV that we already had exported from Strava.

This plot does not have any extra interactivity, also, focusing on only one geo point per activity is not really our goal (but a proof of concept), hence a more in-depth geoplot will be shown in the next sections.

B. Activity Time Analysis

Having an overview of the time spent doing sport activities is extremely important, thus with the bellow plot, inspired by the well known chart already existent on every GitHub user's profile page, we aim to let the user glance over how active was in the past years.

With a lighter green color are depicted days with little activities, and as the color intensifies it shows a more active day.

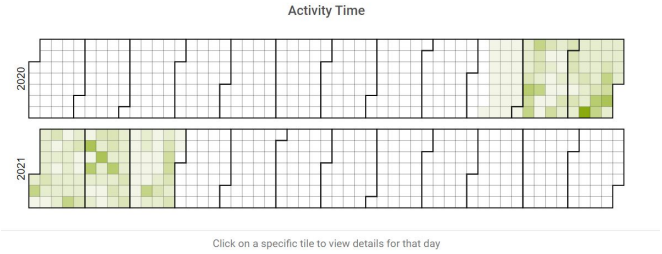


Fig. 2: Activity days

A pop-up appears by clicking on a specific day, showing the number of calorie consumed, the minutes of activities, number of steps and total distance traveled.

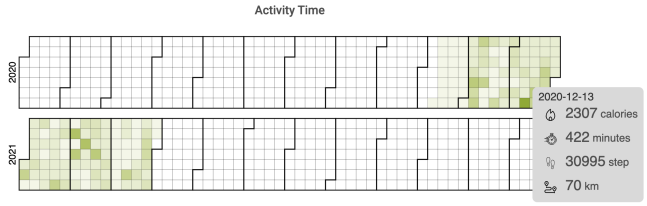


Fig. 3: Activity days - in depth

III. In depth analysis

A. Activities list and information

Each activity is shown in a table, along with important information like: activity type, average heart rate, calories, distance and time elapsed.

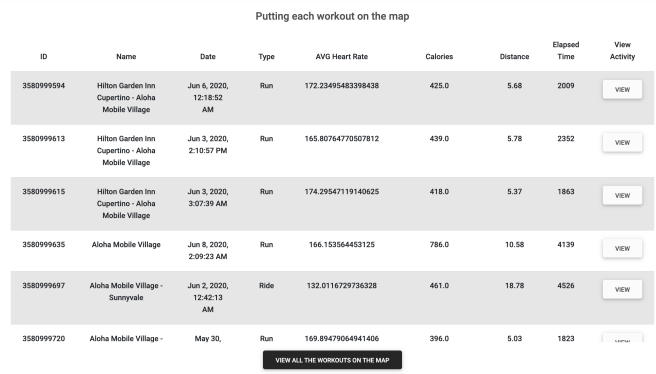


Fig. 4: All activities - table

From a technical point of view, we have generated this table in **d3**, thus, further improving and diversifying the learning experience (rendering html instead of rendering elements in a svg).

To take advantage of the gpx files (generated by the smart watch), it activity can be seen rendered on a google map.

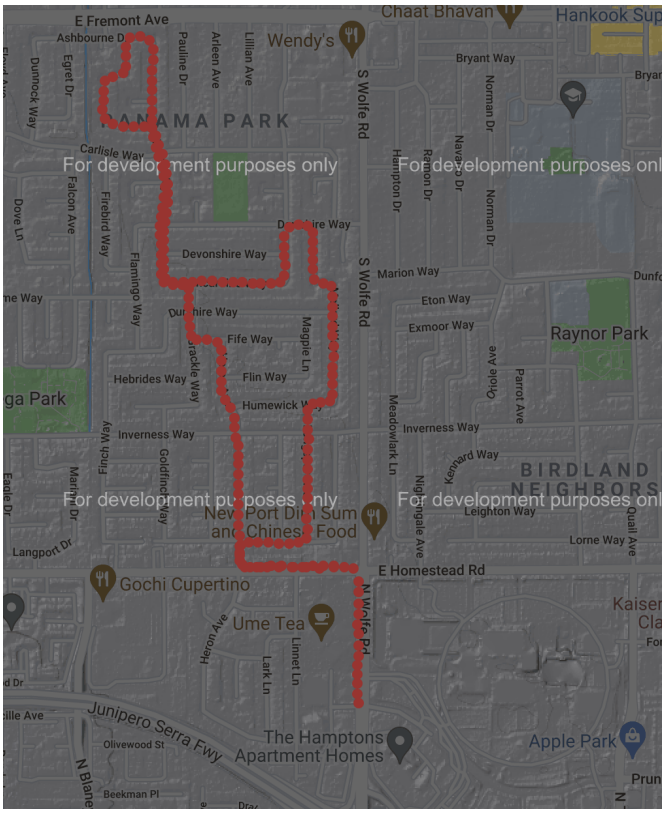


Fig. 5: Single Activity Geoplot

As we don't have a paid account for using Google Maps API, unfortunately we cannot get rid of the *for development purposes* banner, but this does not affect the functionality and the learning experience.

In order to plot the above visualization we have created a separate html page that takes an **activity_id** parameter that identifies the activity (then it takes the appropriate gps points from the json file).

The gps data for each workout is preprocessed beforehand in Python, we are transforming the **gpx** files into a json that can be more easily parsed with **d3**. In addition to this we also compress the amount of gps points by only saving every 5th point.

The plotting is done by superimposing the gps points (in d3) on the google map.

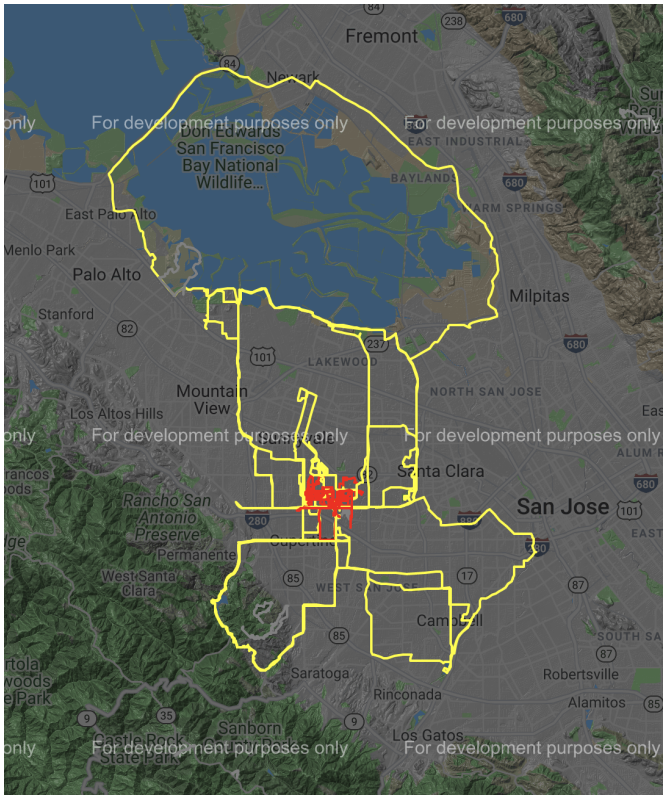


Fig. 6: All Activities Plot

We have tried plotting all the activities with d3, at once on top of google maps but because of performance issues we had to rely directly on the google maps API. Trying to further decrease the number of gps points would have made the chart usable but we still would not have solved the problem. Possibly, one way of tackling it would have been to transform all the points into a polygon and render it.

The type of activities are color coded, yellow being bike activities, gray is hiking and red is running; thus only from the color coding we see how much more distance was traversed using the bike (compared to running).

B. Activity types

Other information that will be of interest to the user is comparing different data points based on the type of activity: the time spent doing each type of activity, the consumed calories, relative effort and total distance.

In our case, we have a total of five activities (Ride, Run, Hike, Walk and swim).

Some of these charts are meaningful if the user tries to achieve certain goals, like loosing weight, case in which he/her would be interested to finding out by which sport he/she lost the most calories in the shortest amount of time.

Fig. 7: Activities type by calories

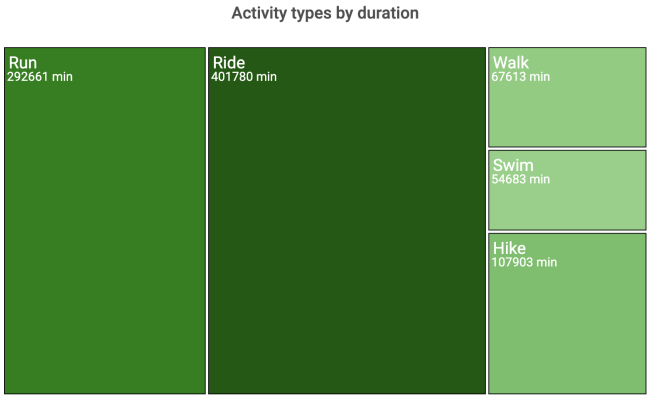


Fig. 8: Activities type by duration

In this graph, with the support of the lecture Mark, channels, we used a tree map graph and the size is adjusted in proportion with the observed quantity.

These are actually a group of charts that have the same backbone code, thus, we are focusing on a modular implementation that allows code reusing.

```
renderTreePlot('#treeplot_activities_type_duration', 'Elapsed Time', ' min')
renderTreePlot('#treeplot_activities_type_calories', 'Calories', ' cal')
renderTreePlot("#treeplot_activities_type_distance", "Distance", " km");
renderTreePlot("#treeplot_activities_type_effort", "Relative Effort", "");
```

Fig. 9: Coding style

C. Average speed in activities

D. Process

We wanted to make the website more interactive with the user and show simple and clear information. This is why we decide to plot the "average speed" which we find nice to know. Even if they shouldn't change too much between all the athletes. With this graph, the user could compare his progress in one of the three sport by reloading new data at another time of the year. The challenge encountered in this chart was mainly on how to use buttons in Javascript and to connect the svg with the backbone of the website.

First, we used a python notebook (preprocessing.ipynd) to get the average speed of each sport. Then, we create the graph with D3. [4]

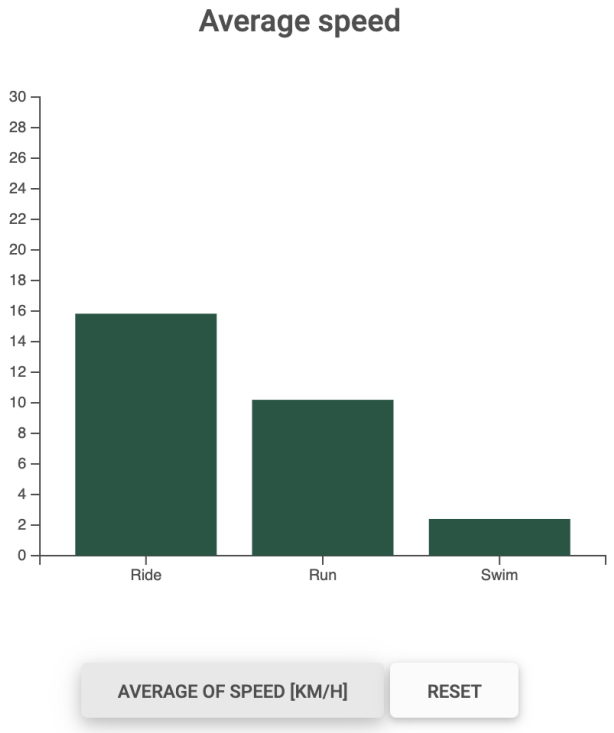


Fig. 10: Average speed in activities

The above chart depict the average speed for each activity type. The result is pretty obvious, the cycling ride being the fastest with an average of 16km/h and the swimming being the slowest.

IV. Sleep data analysis

A. Sleep duration histogram

We also want to present to the user an overview about his/her sleep, hence the next chart in which we have made a histogram of the sleep duration (hrs). The histogram concludes that in the most night, the sleep hours number is around 8.5 hours.

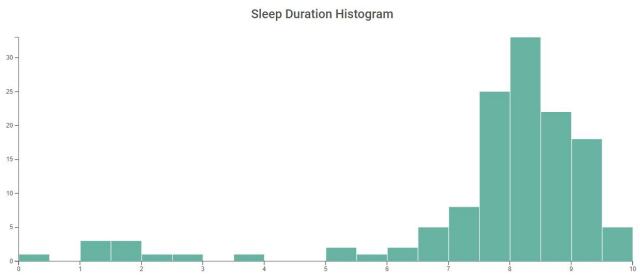


Fig. 11: Sleep duration

B. Sleep vs DeepSleep stacked bar plot

In healthy adults, it is normal to have about 13 to 23 percent of your sleep is deep sleep. Thus, we can conclude that the app we have been using is consistently showing more deep sleep.

Even though this makes us believe that we cannot fully rely on these values, even if the absolute values are incorrect, the differences (from night to night) and overall trends still have explinatory power.

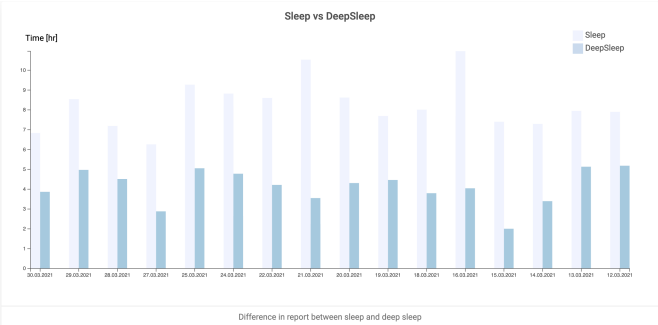


Fig. 12: Sleep vs DeepSleep stacked bar plot

C. Sleep, deep sleep duration; Sleep Cycles

Respecting the general code architecture, we have written a function to render barplots and then instantiated it multiple times in order to explore different sleep information

One can switch between barplots by clicking on the specific nav bar item.

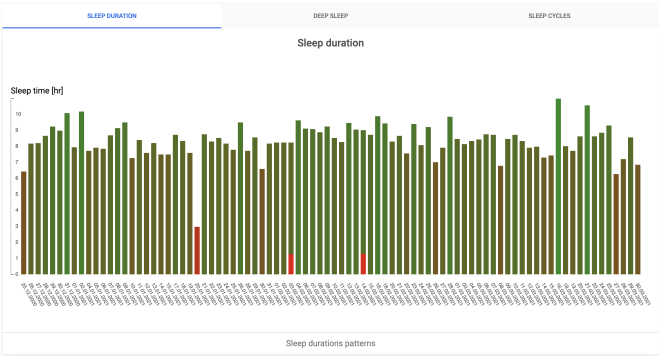


Fig. 13: Sleep Duration Barplot

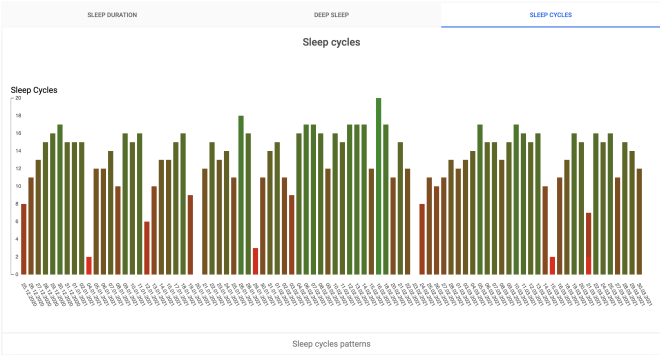


Fig. 14: Sleep Cycles Barplot

V. Motivation through the year

A. Process

After giving an overview of the general data recollected on the subject, we decide to deepen the study on the "Moving Time". In fact, this feature as describes before represents the time we spend on an activity. So, the motivation of a subject for each activity could be studied by plotting the evolution of the corresponding "Moving time". First, we had to preprocess our data on a Python script in order to form a dataset only composed by the monthly sum of each activity type and to plot the first version of this evolution. As the result was promising we decide to use a D3 stacked bar for visualization [4], so that we can clearly see the proportion made in each sport for every month. To facilitate the study of only one activity, we decide to use a "mouseover" on the legend, so that we can simply over on the legend and see only the corresponding area on the graph. Moreover, we decide to use metric of hour on the y axis because we thought that it will be easier for the reader to follow the graph because we use to speak about sport in hours. we all hear sentences like " How many hours are doing every week? , etc.". The challenges encountered during this visualization were mostly during the translation of examples from D3.js Graph Gallery in D3.v4 into the newest version of D3.v6. Indeed, very little documentation on the web is in v6 so we had a bit of a hard time dealing with formating the date from string to DateTime and also with how to treat the mouseover in v6.

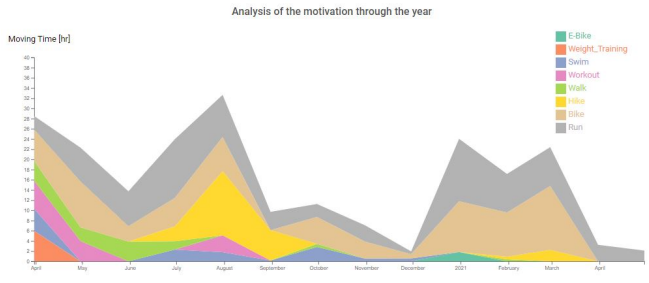


Fig. 15: Sport and motivation

B. Analysis

In the figure 15, we clearly see that a nonconstant evolution. Indeed, it is very interesting to see a drop in the motivation of our subject after September.This may be explained by the weather becoming colder, but the chart is mostly dedicated to being used by the same person which records the data. So, only the subject could recall why he has a drop in his motivation. Then, he may have a better understanding of what makes you want to do more sport. Moreover, if we observe activities one by one we can also observe that they are not constant over the year some activities are done at a certain time of the year as the "Hike" which is done only during summer due to the climatic condition. Moreover, we can also observe what could be the effect the COVID-19 on the weight training activity, because the activity disappear after May 2020.

VI. Athletes types by season

A. Process

Following the same idea, it will be interesting to have the profile of the athletes so that he could know more about the sport he loves to do in which season. In order to that, we also preprocessed our data on a Python file with panda and we regrouped our data used for the previous visualization in two grouped :

- Winter & Fall which regrouped the activity made from November to Mars
- Summer & Spring which regrouped the activity made from April to September

Then, we were able to get the weight of each activity on percent on the total moving time of each period. Finally, we were able to establish the profile of the athletes for the two period using a D3 spider chart. [3] The reason for the choice of this type of graph is mainly because it is rarely seen on other programs as python and that I find them very nice looking with an access to the information fast and simple To have better visualization and to gives a better experience to the user, we highlight the area of each period with a "mouse over" and we also show the corresponding values of each dot linked to his corresponding activities. Note that the

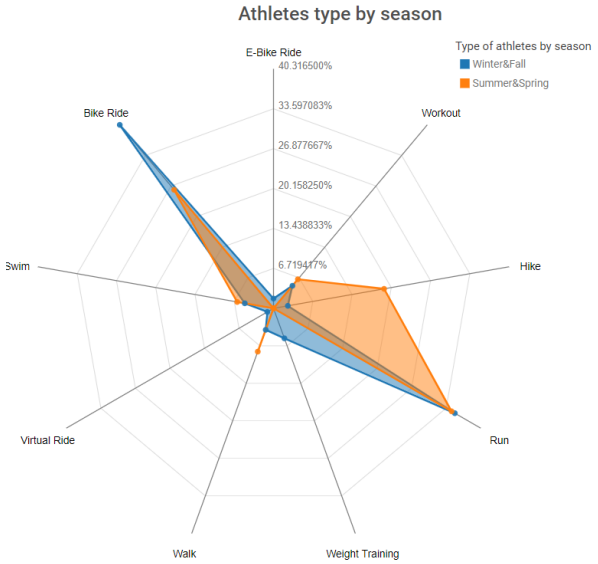


Fig. 16: Activity days

B. Analysis

In the figure 16, we observe interesting characteristics of the athletes between the two groups. In fact, we easily observe the difference between the sport made more during summer and spring and the one made during winter and fall. The two biggest differences are on the percentage of Moving time spend "Hiking" which is much bigger in summer than during winter, which is perfectly understandable and coherent with the figure 15. A surprising result is that our athlete spends much more percentage of his time Biking during winter and fall than during other time of the year. That could be explained by the fact that as hiking is not really possible during the winter due to the snow he will spend more time doing other sport like "Bike Ride". Finally, we observe that the athletes spend the same percentage of time running without taking into account the season.

VII. Study of correlation between sleep and activities

A. Process

As explained in the last milestones, the goal of this project was to compare activity and sleep. We first wanted to explore a possible correlation between sleep and late runs, but as we don't have enough night runs we left out this idea. Nevertheless, we study the correlation between the amount of deep sleep / the number of sleep cycles versus the quantity of physical activities done on that day (and other similar metrics). During the last milestone, we provided a sketch **??** of a 2d density plot and we manage to create it based on this idea develop in the sketch. The first idea was to have a discrete x-axis with 0 if the athletes have done an activity before his night and 1 else. Moreover, we decide to also study two continuous quantities as "Calories" and "Moving time" that have been recorded the day before. Note that for the calories graph, the data are only for nights where we have done an activity so

that we can observe the effect of spending more calories on sleep.

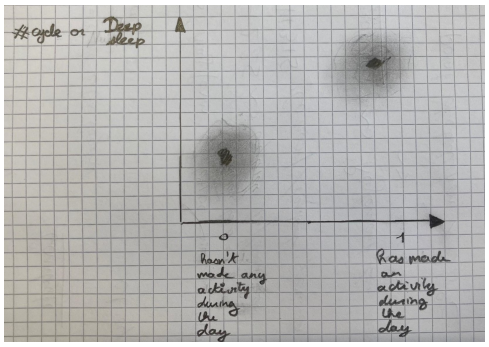
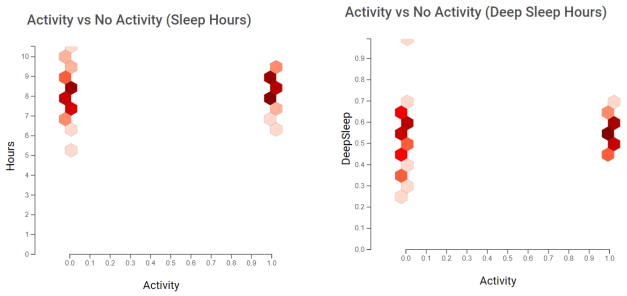
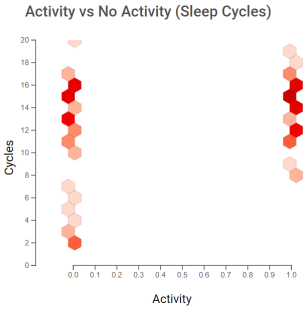


Fig. 17: Sketch heatmap

B. Analysis



(a) Sleep Hours (b) Deep Sleep Hours



(c) Sleep Cycle

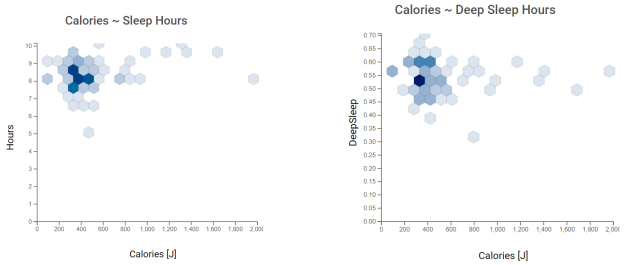
Fig. 18: Activity vs No Activity

In order to make the sketch real, we first had to preprocess our data in Python to clean and merge the two raw data sets (sleep and activities) that were available. The challenging part was to access the right data. More precisely, as it is possible to go to sleep and wake up the same day we had to use some hypothesis. So, we decide to merge the two data set on the "waking up date" of the athletes and always consider the day before this date. Moreover, we filter the sleep with too low cycles and consider them as naps. Note that we also balanced the data in order to have the same amount of night with and without activities in figure 18.
In the second part of the process, we visualize with D3 library [5]. More precisely, hexbin library (d3-hexbin.v0.2.min.js). To add interactivity with the user, a slider has been add which allows the user to change the radius of each hexbin. This way, the user will be able to understand better how the hexbin graph is constructed if he doesn't know them or he will be able to choose the size of the interval that he wants to study. The slider is set up with a default radius of 10 and can be changed between 2 and 30. Note that if the slider is set to 2 the graph will be close to a scatter plot. Finally, we also add a movement to the graph each time we change the radius to make transitions more dynamics.
The challenge encountered for those visualizations was first as the one before the translation from examples in D3.v4 into the newest version of D3.v6. Secondly, we add a hard time dealing with the slider so that it reloads the same image. Before it was adding a new image each time.

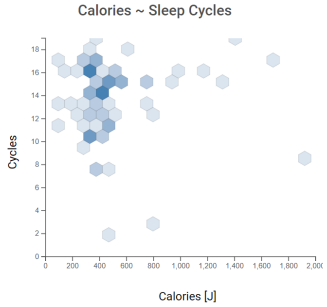
As you can see on the graph above, the result isn't what we were showing on the sketch 17. It's important to notice that the darker the hexbin is, the more nights have been recorded with those parameter quantities. Thus, we observe that when an athlete has done an activity before going to sleep we have fewer small values for Sleep Cycle and Deep Sleep Hours. For the hours versus Activity, where we observe that the result for both discrete variables is very small. This can be explained because the hours we sleep are linked to many other variables like what time are we working. In conclusion, no major correlation between a day with and without activity can be really extracted from those graph, but we propose some hypothesis to explain those result:

- This could be due to a lack of data and by adding a lot more data we will reach better results.
- We could question how the data were collected and therefore the precision of the connected watch.
- The existence of other factors influencing sleep could disturb or result
- The last hypothesis is the fact that activity and sleep aren't related

As explain before, we decide to analyse other continuous variable like the link between Sleep and the Calories spend during the day.



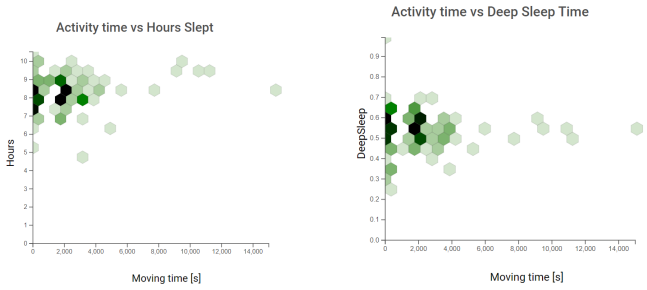
(a) Sleep Hours (b) Deep Sleep Hours



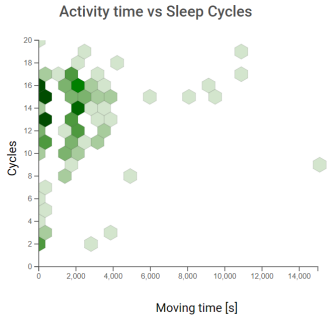
(c) Sleep Cycle

Fig. 19: Calories

Here we observe no meaningful correlation in our result this can be explained by the same hypothesis stated before. Then we can observe our last analysis which compare the effect of moving time on our sleep.



(a) Sleep Hours (b) Deep Sleep Hours



(c) Sleep Cycle

Fig. 20: Activity time

In fact, we observe again no meaningful correlation so we conclude that doing more sport(moving time) doesn't necessarily make us sleep better. Nevertheless, we observe the moving interval from 6000 [s] to 12000 [s] we tend to have over 15 cycles and a deep sleep around 0.55 [%]. Under this interval for the sleep cycle, the values are more chaotic. In conclusion, to support the hypothesis on the fact that the connected watch may be imprecise we observe that in all the graphs we have a range of Sleep cycles between 2 and 20. Nevertheless, a normal human being in a typical night goes through four to six sleep cycles [6]. So the quality of the measurement with a watch could be discussed. To have more precise data, we should collect the same data with more high-tech material as it is done for similar studies in hospitals.

VIII. Technologies

The technology stack is kept as light as possible. We use an HTML/CSS design framework called mdbootstrap (a material design variation of the well-known bootstrap framework). Besides this d3-v6 along with a geo-projection and color scale plugin.

There are various data visualization plugins built on top of d3, in order to respect the didactic purpose of this project he hasn't use them.

IX. Individual Contributions

A. Mircea Sorin-Sebastian

Have built the website html/css skeleton (on top of mdbootstrap). Mainly focused on the data exploration charts, built:

- The overview github like activity chart - without the pop-up part which was added by Ajeghrir Mustapha
- The overview locations geoplot chart
- The table that lists all activities as well as the view activity / view activities functionality (the google maps views that plot the workouts paths) along with the python scripts that preprocess the needed data
- The treemap that plots activity types by duration, calories distance and relative effort
- The sleep exploratory plots: the barplots, sleep/deepsleep stacked bar plot and the sleep duration histogram

B. AJEGHRIR Mustapha

Mainly focused on reviewing the work for my two colleagues. Have implemented some features like the pop-up of the activity chart and card animations.

C. Simon Dayer

Mainly focused on first the preprocess of our data. author of the jupyter notebook (preprocessing.ipynb) and learn the tool in class at the same time (Html /JS / D3). Then, I studied a possible correlation between (Sleep vs Activity / Sleep vs calories / Sleep vs Moving time) with an implementation of different radius with a slider, try to express the motivation of the athletes with a graph , showing average speed on a bar plot with animation and create athletes types by seasons. Finally, I could integrated all the graph I made (more precision below) to the backbone of the website and reorganise them:

- The average speed bar plot

- The Sport and motivation stacked bat plot
- The activity type by season radar chart
- The hexbin chart with (could learn how to use mdbootstrap) Activity vs No activity
- The hexbin chart with (could learn how to use mdbootstrap) Calories
- The hexbin chart with (could learn how to use mdbootstrap) Moving Time

Finally, I created the Readme and clean the repository of the unused files.

X. Conclusion

In conclusion, the visualizations that were plotted in this project had the goal of encouraging people to continue their activities by looking in depth into their data. This project could also give some insight on correlations between sport-type/season or sleep-quality/activities for interested specialists. Note that it will be interesting to go further in this analysis by using more modern method to collect data during his sleep (Similar experience are made in hospital with modern method than simply a watch).
Finally, this website could also be a share-point where friends could share or compare their activities with loved ones.

References

[1] D3Js examples website: https://observablehq.com/@d3/gallery
[2] Activity days chart template https://github.com/DKirwan/calendar-heatmap
[3] Radar chart template https://gist.github.com/nbremer/6506614
[4] D3Js examples website graph gallery: https://www.d3-graph-gallery.com/index.html
[5] Hexbin chart : https://github.com/d3/d3-hexbin
[6] Website, 2021. *Sleep Fondation A OneCare Media Company* [online]. Found at this address: https://www.sleepfoundation.org/how-sleep-works/stages-of-sleep#:~:text=Sleep%20is%20not%20uniform.%20Instead%2C%20over%20the%20course,on%20average%20they%20last%20about%2090%20minutes%20each.