Data visualization : Milestone 1

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

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

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

***Abstract*—The aim of this milestone is to provide to the reader a good understanding of the potential of our project. We will first look into the chosen data set and explain the problem that we want to investigate. Finally, we will go through related work and explains the first steps that we will follow.**

I. Dataset

As all the team is keen on sports and tracking as much data related to it as possible, we decided to direct our project towards those fields of studies. Moreover, one of the team members has been very dedicated to recording his physical activity and health related data during the past year. So based on those unique data points, we created a data-set with entries from around 150 runs and 90 bike rides collected with a "Samsung watch 3", we also have 3 months of sleep,walk ,calories and heart rate data. Based on these, we are looking forward to providing extensive visualisations with informative/statistical purpose and with the scope of providing correlations between different aspects (like how the amount of sport influences the quality of sleep) in the subject's life.

Basic prepossessing and data-cleaning are already done by Samsung, Strava and AutoSleep application, so our job is mainly to thoroughly explore all the data, aggregate and display it.

II. Problematic

The goal of the project is to provide a dashboard that aggregates multiple data sources (Samsung Health Data, Strava and AutoSleep) and gives a bird's eye view of the progress over an extended period of time.

A second goal is to use the different data points towards arriving to some correlations (like a possible link between quality of sleep and sport activities).

With this visualization, we want to mainly target the sports audience which records their data and would want a nice way to analyze them. Moreover, by going more deeply into the relationship between sleep and sports we can pretend to target a much wider audience.

III. Related work

Nike, Samsung, Strava and Garmin have developed diverse visualisation to have better design and provide better services to their customers. Indeed, some example of the "Nike" newly developed visualisation can be found on the website of Andre Salyer [1]. Others can be found on the well-known web service "Strava" which also have basic representation of running data.

IV. Inspiration

Strava and Galaxy Health are definitely the services that have inspired us towards choosing this project, we believe that these visualizations can be limiting in some aspects and through this project we want to impose our own view of providing meaningful visualizations.

V. Exploratory Data Analysis

The data comes in various formats: json, csv and gpx and it is time-based (meaning that every data point can be exactly placed on a time axis). The physical activities (runs and bike rides) have associated location data (GPS), this enables us to enter into the world of D3.js geoplots and heatmaps.

Strava provides an useful CSV that aggregates various aspects of the workout: *Activity Type, Elapsed Time, Distance, Relative Effort, Moving Time, Distance, Max Speed, Average Speed, Elevation Gain, Elevation Loss / Gain*. We wanted to augment this by adding to each row the latitude and longitude of the start of the workout, for this we had to join the CSV and GPX files and use a gpx python parser (ggps) to read the latitude and longitude.

As we wanted to make sure that the project is feasible, we have invested some time in designing the architecture of the code and implement a few prototype visualizations (to get an overview of how the data looks like).

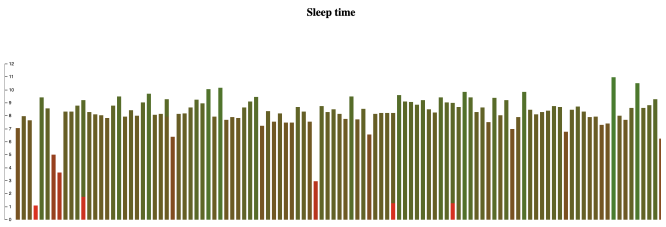

Fig. 1: Barplot with sleep duration

Regarding the sleep data, when processing it we observed that for each night we had multiple entries of different sleep segments, as some of them were overlapping we had to merge the intervals.

VI. Personal character of the data

We are aware that the data that we are going to use contains sensitive and personal information (locations, health data).As the owner of this data (Mircea Sorin-Sebastian) I fully agree to using it for this class and openly sharing all of the results. It is my intention to continue gathering more data points and publicly making them available on platforms like Kaggle.

Even more, as a team we believe that this aspect is going to provide a character of originality to the project and it is going to motivate us to provide meaningful visualizations and correlations.

REFERENCES

[1] Website, 2021. *Personal website of Andre Salyer* [online]. Found at this address: https://www.andresalyer.com/nike-heat-map/