Antonio Iubatti, Gianrocco Lazzari, Maxime Peschard, Ondine Chanon and Stefano Savarè

Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Objectives

The project's goal was to analyze data from *Datasport* [1], a company the provides data management and communication, for several sport event, mainly in Switzerland. The challenging objectives were:

- data parsing;
- exploratory data analysis, oriented to pattern discovery in sport habits, for instance across athletes' sex and age;
- creation of a website to host the results of our data analysis and to propose a user-friendly interface for browsing the data.

Data Parsing

The parsing was done in Python, with the help of BeautifulSoup (a python library) and Postman (a platform for API development). Here follow the relevant steps:

- **Global parsing**: for each Swiss race organized between 1999 and 2015, get its name, location, date and *Datasport* URL address, where to find the race results;
- **Ranking parsing**: from every URL retrieved, get all the information about every specific race, including the information on every runner: name, age, category, ranking, pace;
- **Weather parsing**: from the date and location of each event, find the corresponding temperature and weather, in order to do performance analysis with respect to the weather/temperature. Due to the API used, such information for races older than July 2008 is not available;
- **Gathering information**: merge the data on races and weather, together with the GPS coordinates of each location, when available.

The main challenge was to deal with **missing data, often formatted in an uncommon structure.**. Moreover, we found an **lack of consistency** in the data among races (names of the data fields are often different), partially (but not only) because of the different Swiss languages used. In general, the lack of an simple data structure (e.g. table), forced us to parse data often presented in long and heterogeneous strings.

Data Visualization

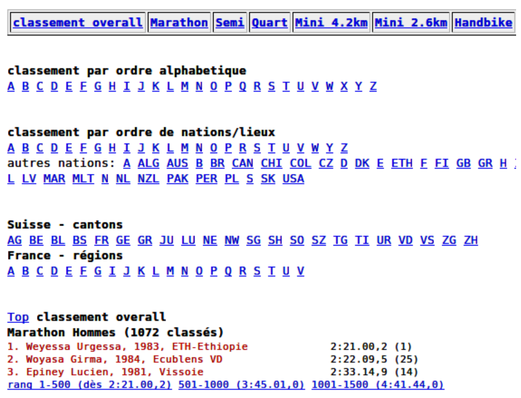

Figure 1: An original *Datasport* web page

The Jekyll-powered *Hop Suisse!* website hosted on GitHub displays the most widely accessible results of our data analysis (Fig.2), with an interactive and user-friendly interface. Several libraries and tools have been used such as C3.js, jQuery-Autocomplete, Leaflet, Google Fonts, Glyphicons Free and Weather Icons. The website contains:

- a page on the most reachable **exploratory statistics** on the entire dataset;
- a page showing a deep analysis of the **Lausanne marathon 2016**;
- the *Races* page where, through an auto-complete search field, the user can look for a specific race and find specific information about it;
- the *Runners* page where, similarly to the *Races* page, the user can **look for a specific runner** and find specific information about him/her;

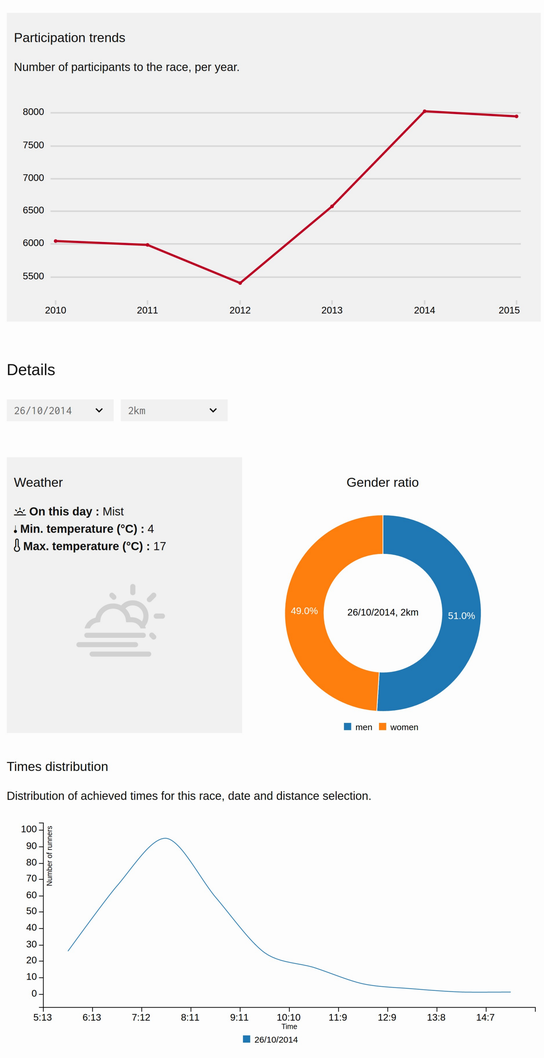

Figure 2: An interactive and user-friendly web page from *Hop Suisse!*

Exploratory Analysis

Here we present some of the results that follow our exploratory analysis, on the **full data set**:

- The number of runners has increased in time, for both sex (Fig.3). However, the number of participants to the marathon has decreased, while the one of the half-marathon has remained stable.

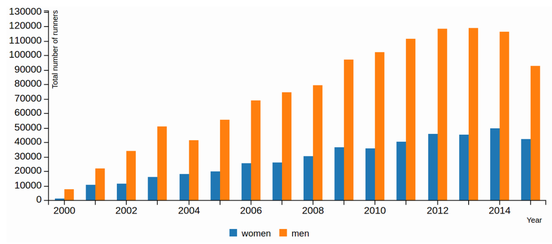

Figure 3: Evolution in time of the total number of runners, divided by sex.

- A U-shape relation is observed between the runners' median time to complete the race and their age, especially for long races (half or full marathon) with a large number of runners (Fig.4).

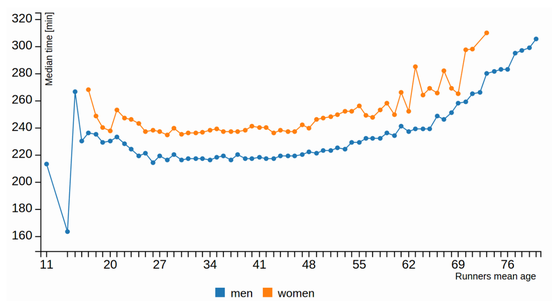

Figure 4: 20km de Lausanne, U-shaped relation between the runners' median time to complete the race and their age.

- The runners' mean age in some races significantly changed over time (i.e. Course de l'escalade seems to attract more and more young boys/girls).
- The distribution of the number-of-races/runner seems to follow a broad distribution (Fig. 5) - for more details see the related IPython notebook

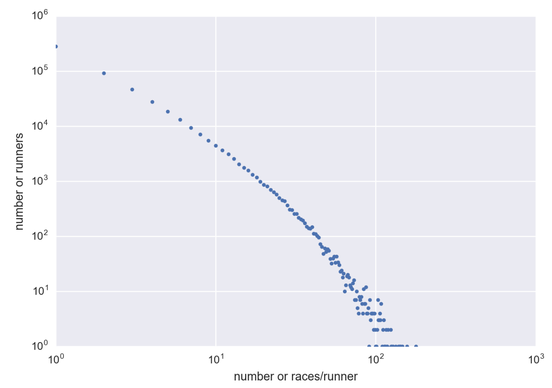

Figure 5: Distribution of the number of races per runner.

Lausanne Marathon 2016

This part of the analysis focused on the three main races: **10km** race, **half-marathon** and **full marathon**. After 2-sample Kolmogorov-Smirnov tests ($p < 0.05$), we conclude that:

- Women who run the *10km race* are significantly **younger** than the ones running the full marathon or the half marathon;
- Men who run the *full marathon* are significantly **older** than the ones running the half marathon or the 10km race;
- Women are significantly **younger** and **slower** than men in all the three chosen competitions;
- Men **pace increases** significantly with the distance of the races;
- Women **pace** is significantly **higher** only for the *marathon*.

How is a runner defined?

A unique runner is defined through a tuple containing [*first name, last name, birth year*]. We decided not to consider the *Living place*, as it is not reliable enough. Indeed, some athletes likely changed living place between two races. Counting this cases as different runners, would corrupt their longitudinal data.

References

[1] Datasport AG www.datasport.com, 2012.

Acknowledgements and Competing Interests

Hop Suisse! is a university project for the Applied Data Analysis course at EPFL. We thank Michele Catasta, our lecturer, for his involvement.
There are no conflicts of interest to disclose.

Supporting material

- Web: https://hopsuisse.github.io
- Video on runner's age: see QR code below