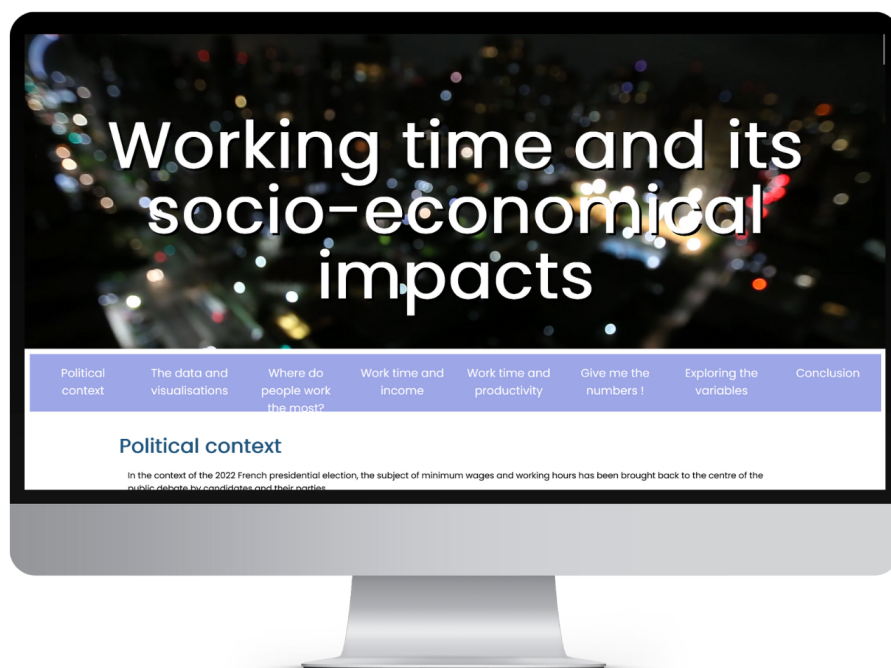


Data Visualisation: Process Book



Hannah Gelblat
Bastien Golomer
Clémentine Lévy-Fidel

EPFL

Spring 2022

Contents

1	Conception	2
1.1	The idea	2
1.2	Refining our dataset of interest	2
1.3	Understanding data behavior	2
2	What are we showing, exactly?	3
2.1	The map	3
2.2	Data at the cross-country level	4
2.3	Data at the individual country level	4
3	Correlation matrix	5
4	Global challenge: understanding the subject and the fear of misunderstandings	5
5	Peer assessment	6

1 Conception

1.1 The idea

What is hidden between working hours and socio-economical well-being?

To answer this question, our first step was to explore the large database of the OECD which provides a wide range of socio-economical data across several decades and countries. We started our investigation by diving in the database, exploring the variables it had to offer, where and for which period of time they were available. We identified a selection of variables that we found belonged to the subject – they could be classified into the following categories:

Working hours and incomes in the country	Impact on the productivity of the country	Impact on living standard
Average wages [1]	GDP per hour worked [2]	Income inequality [3]
Employee compensation by activity [4]	Labour productivity and utilisation [5]	Poverty rate [6]
Hours worked [7]		Working hours needed to exit poverty [8]
		Employment rate [9]

Table 1: Our first selection of variables from the OECD database

1.2 Refining our dataset of interest

Yet the selection could still be refined, for instance by questioning further the meaning and relevance of each variable. Variables like labour productivity or working hours needed to exit poverty were indeed secondary variables, i.e. variables computed from other variables, that already contain a form of interpretation *per se* and that we decided to discard so as to focus on a more specific question. As for productivity, we were still able to keep this concept with the variable GDP per hour worked.

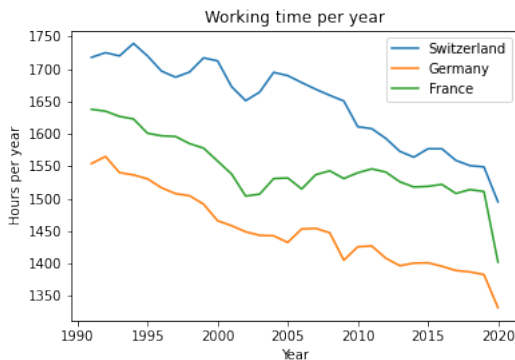
Working hours and incomes in the country	Impact on the productivity of the country	Impact on living standard
Average wages [1]	GDP per hour worked [2]	Income inequality [3]
Hours worked [7]		Poverty rate [6]
		Employment rate [9]

Table 2: Refined selection of variables from the OECD database

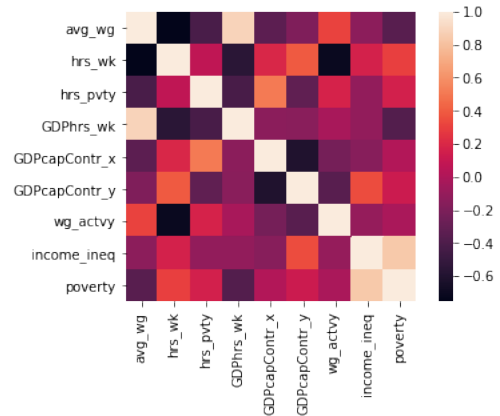
1.3 Understanding data behavior

Our final selection of variables still counted several socio-economical factors, each complex to understand and interpret. Our idea was to allow the user to have as broad an overlook as possible of these variables. This is achievable by providing our data story with simple yet straightforward representations of the distributions of the variables. Distributions are thus shown across time, and across countries, allowing users to visualize temporal evolution, compare between countries, analyse and reflect on what they observe.

Our first exploration was made with Matplotlib and Seaborn, both Python libraries:



(a) Simple distribution representation with Matplotlib on Python



(b) Correlation matrix with Seaborn on Python of the variables in Table 1

Figure 1: Initial plots

We turned the representation into multi-variable, multi-country and time-adapting interactive visualisations using VegaLite grammar. VegaLite offers a wide range of interactive options. Although it is less bespoke than D3, it offers more interactive widgets by default. We considered it an attractive alternative for our scatter plots or distribution plots.

As for cross-country comparison, another useful visualisation was to include a map. We chose to create it with D3.

2 What are we showing, exactly?

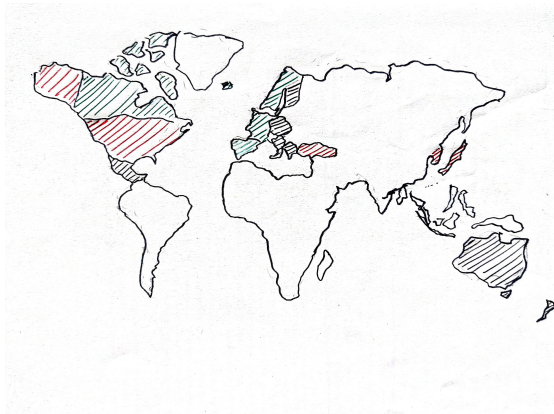
2.1 The map

We wanted to put forward our data in a non-specific way to give the user an overview of it. As our crucial variable is the number of working hours, it seemed essential to us to allow the user to visualize the distribution of working time and its usual values in the countries we consider.

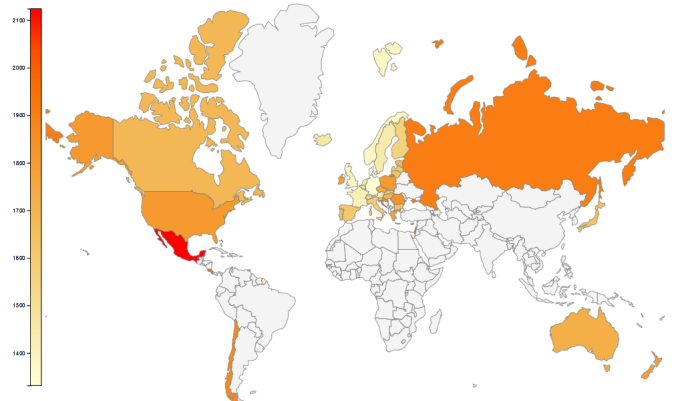
It also allows the user to get answers to hypothesis they can have, for example...



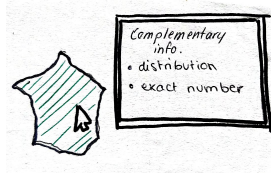
Here (Fig. 4a) is the sketch we had planned for the card - and what we obtained at the end on Fig. 2b We used series 8 on Maps of the course by adapting it to our project. We used a json file from around the world and adapted our data from the OECD to the necessary functions to display the corresponding working colors and scale on the left.



(a) Working hours among OECD countries in 2018. Data from the OECD ([7])



(b) Our final map



(c) Detail of the sketch - window appearing when the cursor passes

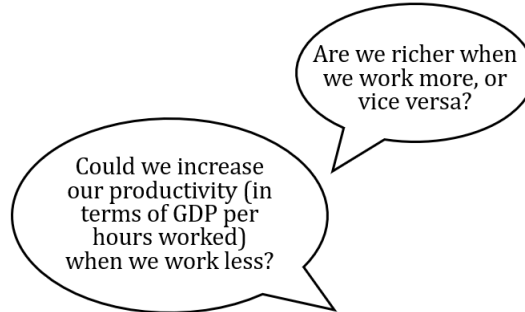
Figure 2: Visualising data on a map

For this visualization, the challenge was in fitting our data to the *Exercise serie 8* method while adapting scale and color choice to the data so the visualisation is more understandable. The map could be implemented

further and include features we initially wanted to add to the map - for example the additional information that would appear when clicking on one specific country (sketch on figure 2c).

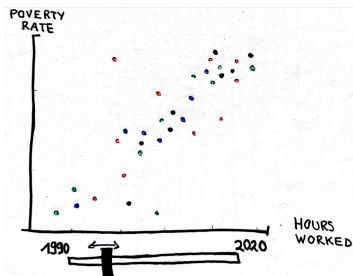
2.2 Data at the cross-country level

We explored the links between the variables of interest. We wanted to answer some basic questions asked in popular discussions about working time, for instance the following questions:

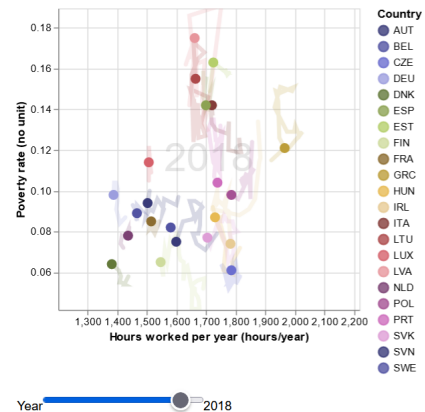


Our two main scatter plots to answer these tricky questions compare, on one hand, the hours worked and the poverty rate of the country, and on the other hand, the hours worked and the GDP per hours worked in the country.

We were able to follow closely what we expected to do for this visualisation. Our scatter plot can be adapted to the year of interest thanks to a cursor, a useful widget for interactivity. We separated European Union (EU) countries and non-EU countries to highlight any pattern that would differentiate those two groups. The example on fig. 3b shows one variable displayed for EU countries. By hovering the mouse over the data points, a small window appears with details about the country and the exact values.



(a) Comparison between variables in the countries of our dataset. Data from the OECD [7].



(b) One of our final scatter plot, 2018 data from OECD [7] and [6]

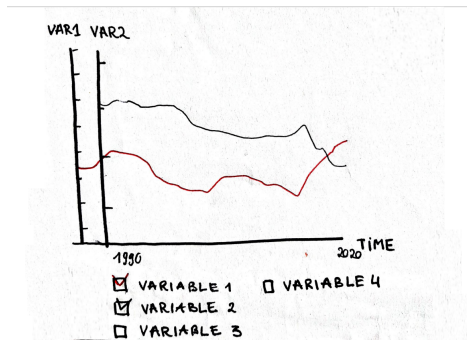
Figure 3: Visualising data at the cross-country level

2.3 Data at the individual country level

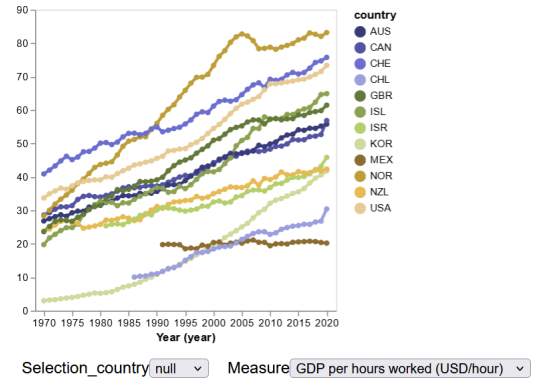
This visualisation has been adapted and the initial sketch (fig. 4a) is different from the final visualisation on several aspects (fig. 4b). It turned out that it is technically difficult to display different variables across time on one plot. Instead, we chose to display one variable at a time so that the user could choose which one to display thanks to a button. To allow for comparison, we added the possibility to observe one specific country or all at the same time. Again, we separated EU and non-EU countries to observe any pattern that would differ between the two groups.

Our goal here was to nuance the conclusions that a user could draw when looking at the scatter plots. These plots (fig. 3b for example, as well as other scatter plots from our website) show some correlation between working less hours, having a greater GDP and GDP per hours worked, having a lower poverty rate. We found important to highlight the fact that these are correlations and not causalities. For example, being a higher-income country is an external factor that impacts all the variables studied. At the stage of our study it still

is impossible to answer with certainty at the questions asked previously. Yet by enabling users to explore the variables, we still give them the opportunity to see the complexity of these relations.



(a) Multiple variable comparison over time for a specific country

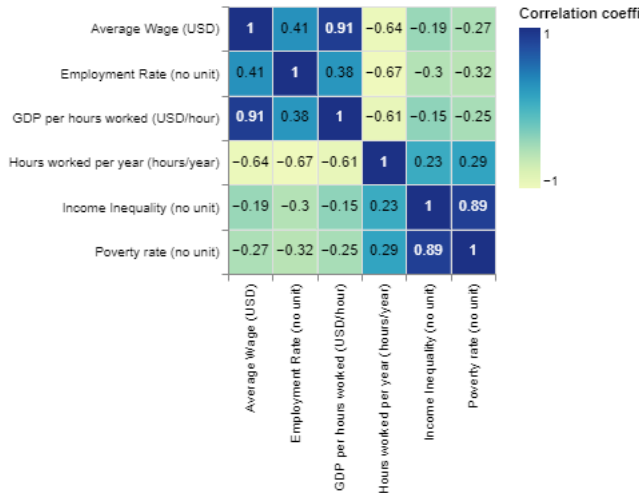


(b) Final visualisation: comparison of a variable of interest (here, GDP per hours worked) in non-EU countries.

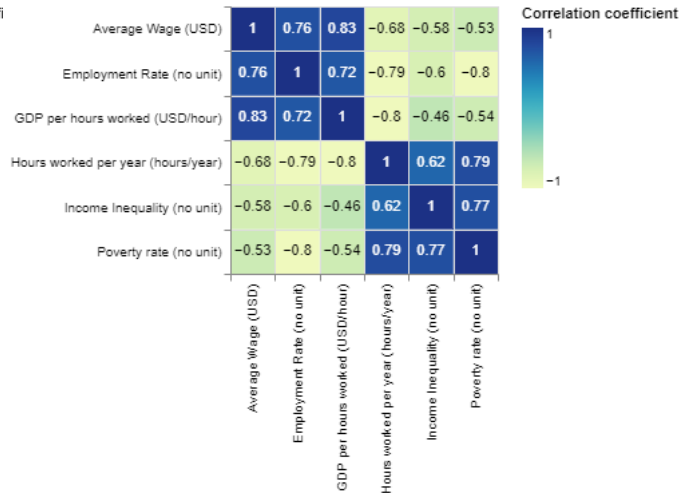
Figure 4: Visualising data at the individual country level

3 Correlation matrix

To detail the relationships between the variables at stake and support the correlations observed in the interactive scatter plots further, we wanted to show the user the correlations calculated for the first milestone. We adapted our initial Python-generated correlation matrix with VegaLite to fit more nicely in our web site. The result is shown in fig. 5a and 5b.



(a) Correlation matrix for European countries



(b) Correlation matrix for non-European countries

4 Global challenge: understanding the subject and the fear of misunderstandings

The study of working hours, inequality, poverty and their links with the country's productivity are extremely vast subjects that require a systemic understanding of the economy of the countries concerned.

This socio-economic topic is of great interest to us and we wanted to allow users to visualize it globally. However, we found difficult to grasp the complexity of the subject – indeed, it is probably impossible to cover all its aspects. By showing correlations without highlighting the complexity of the relationships or reminding the user that the results are not causal, we risked conveying false or partial information to our users, which could be misleading.

To overcome this problem, we hence chose to remain impartial. Our solution was to show the correlations between variables at the country level, but to allow the user to explore the evolution of the variables over time and to compare them with each other and with the different countries.

Another part of the workload was to check for socio-economical aspects of our analysis, by checking for example that the GDP data took the inflation into account, or exploring what being part of the OECD really meant.

5 Peer assessment

We worked together for most of the steps of this project. The choice of the topic, the exploration of variables and the precise definition of our angle of study are the result of a common brainstorming. We faced the objectivity problems mentioned in section 4 together.

For the coding of the visualizations we initially divided them up, each one having a visualization to code and display on the site (Hannah Gelblat - HG - for the map, Clémentine Levy-Fidel - CLF - for the scatter plots and Bastien Golomer - BG - for the variables over time and the correlation matrices), and solved the problems together, each one helping the other when they got stuck on something. The writing of the process book was divided in the same way, each one explaining their part and the problems encountered. CLF took care of the preparation of the skeleton of the website, its structure and visual aspect while HG supervised the consistence of the data story. BG added some features to it and explored VegaLite to be able to use it in some visualizations, HG designed the process book outline, wrote common parts and the script of the screencast.

The three of us took part in filming and making the screencast.

References

- [1] “Average wages.” <https://data.oecd.org/earnwage/average-wages.htm>. Accessed: 2022-06-04.
- [2] “Gdp per hour.” <https://data.oecd.org/lprdt/gdp-per-hour-worked.htm>. Accessed: 2022-06-04.
- [3] “Income inequalities.” <https://data.oecd.org/inequality/poverty-rate.htm>. Accessed: 2022-06-04.
- [4] “Employee compensation by activity.” <https://data.oecd.org/earnwage/employee-compensation-by-activity.htm#indicator-chart>. Accessed: 2022-06-04.
- [5] “Labour productivity and utilisation.” <https://data.oecd.org/lprdt/labour-productivity-and-utilisation.htm#indicator-chart>. Accessed: 2022-06-04.
- [6] “Poverty rate.” <https://data.oecd.org/inequality/poverty-rate.htm>. Accessed: 2022-06-04.
- [7] “Hours worked.” <https://data.oecd.org/emp/hours-worked.htm>. Accessed: 2022-06-04.
- [8] “Working hours needed to exit poverty.” <https://data.oecd.org/benwage/working-hours-needed-to-exit-poverty.htm>. Accessed: 2022-06-04.
- [9] “Employment rate.” <https://data.oecd.org/emp/employment-rate.htm>. Accessed: 2022-06-04.