Data Wrangling - Data Processing

Matthieu Cisel - CY Tech

September 2021

1 Goals

Data preprocessing might not be the most fascinating part of data science, but it is a necessary step to obtain reliable results. It often accounts for most of the time required by a proper data analysis. In this course, we teach the basics of data wrangling and cleansing. By data wrangling, we mean the ability to manipulate data in a quick and efficient way based on a compact code, to change the shape of a dataset according to one's needs using functions tailored for that objective. More specifically, we will explore the most important functions of packages like dplyr or tidyr, in R. Regarding Data preprocessing, we will teach students how to deal with missing values, to identify outliers and abnormalities often encountered in databases, in order to treat them in a relevant fashion. They apply the content of the course on a real-life dataset on PhD defenses drawn from theses.fr.

2 Introduction

The question of the open access or research work has been discussed extensively in academia over the past two decades, (Charton and Schöpfel, 2017) as the business model of scientific journals was increasingly being challenged by the research community. According to Suber (2007), open access literature (of all types) is 'digital, online, free of charge and free of most copyright restrictions'.

Higher education and research institutions have become reluctant to pay expensive subscriptions to editors who indirectly benefit from taxpayer money that subsidised research work, and funding organizations have pushed an open access agenda for years (Harnad, 2011). For instance, the Plan S, initially launched by Europe but joined by China soon after, aims on the long term at compelling publicly-funded research word to be published in open access reviews (Rabesandratana, 2019). While debates on open access have mostly focused on research articles, some scholars have pinpointed the importance of extending the question to PhD thesis manuscripts (Copeland et al., 2005). On average, such manuscripts do not bring any economic benefit to scientific editors. However, they are often funded by taxpayer's money and some authors argue that they

should therefore be accessible to the general audience (Moxley, 2001 ; Hawkins, Kimball and Ives, 2013).

While privately-owned websites like ProQuest have gained momentum in the United States of America, public online repositories avec developed over the years in other countries. For instance, France launched theses.fr in the early 2000s with its associated archive TEL (Schöpfel et al., 2014). International repositories have been launched, notably in Europe, but have remained at an embryonic stage (Moyle, 2008).

3 Instructions

In this learning unit on Data Wrangling, you will learn how to scrap from TEL to create a new database on PhD defenses. You will then proceed to clean the dataset, analyze the spread of missing data, detect issues in the variables, and produce your first research result. Your code should be submitted along with your report (in Latex), that should follow the following structure :

I. Missing data

II. Common issues

III. Outliers

IV. Preliminary Results

3.1 Scraping

Use Python to scrap data from the provided list of URLs (url theses) (Beautiful Soup, Scrapy or Selenium). The syntax is available in associated classes on Datacamp.

Follow Web Scraping in Python in Datacamp and submit the certificate on Teams.

https://learn.datacamp.com/courses/web-scraping-with-python The rest of the analysis is done with R. Use either tidyverse or dplyr to clean the data.

Follow Data Manipulation with dplyr (Datacamp) and submit in Teams

If the scraping takes too long or is unsucessful (prove that you have tried), then proceed with the dataset provided by the instructor.

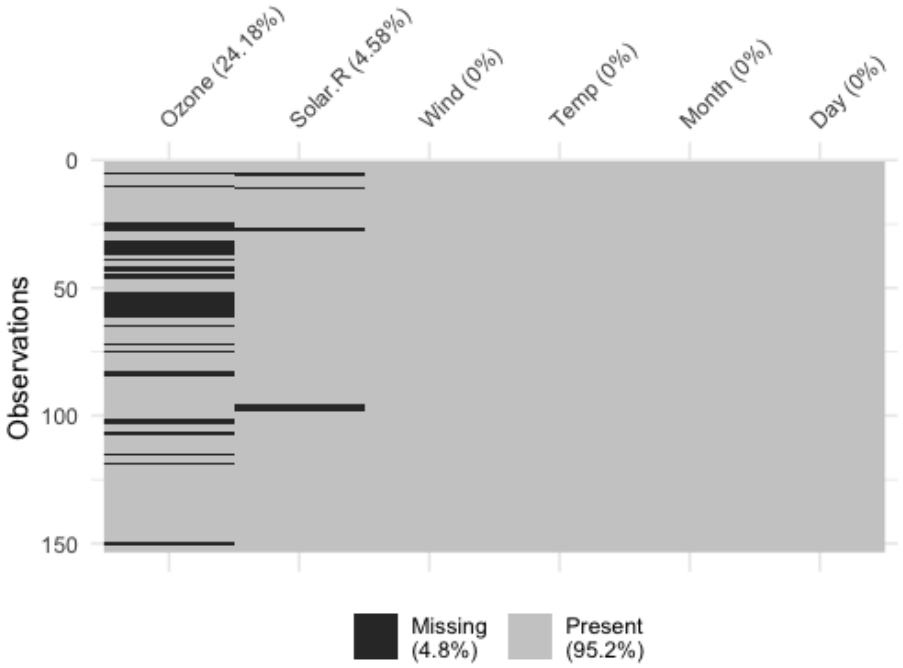

Figure 1: Example of the visualization of missing data

3.2 Missing Data

Follow Dealing With Missing Data in R in Datacamp. Complete the class, submit the certificate in Teams. Assess the quality of the dataset. Are you sure that all the data are visible in you coding environment ?

Create a graph and a table representing how missing data are spread across variables. It should look like something similiar to Figure 1.

What do you notice. For instance, how do you explain, with regards to the issue of missing data, the defence date vs. beginning date pattern ?

Create a variable (n.pages, for number of pages) with a normal distribution of 200 pages and a standard deviation of 50 pages, for a sample of 80 percent of the dataset. Complete missing values (the 20 percent that remain) using an imputation technique.

3.3 Common issues

Check for issues in the defence data. How common are the defences on the first of January. In R, use filter, groupby and count function from the dplyr

library, and the lubridate library, and answer the following question : How did the proportion of defences at the first of january evolve over the years ?

How do you explain it ? In the Author name, how common are homonyms ? Check for Cécile Martin. Investigate her case and try to figure out what happened.

Check for issues in the supervisor's ID. In what context does the issue appear ? How common is this issue ? Provide a quantitative analysis of it.

There is a sudden drop in the number of PhD defended in 2019 and 2020. Propose 3 hypotheses to explain this phenomenon.

3.4 Outliers

Find supervisors who have mentored a surprisingly large number of PhD candidates. Find a way to check whether it is an outlier or a mistake in the data. Explain how you did.

Try to identify early career scientists who have defended more than one PhD. How common is this phenomenon ? How do you differentiate multiple defences and homonymy ?

3.5 Preliminary Results

Recode languages using the dplyr library so that there are only four levels left (English, French, Bilingual, Other)

Find a way to process the defence date and show how the choice of the language of the manuscript evolved over the past decades. Do one graph with ggplot2, another with plotly.

By the way, at what period of the year do PhD candidates tend to defend ? Remove unreliable data and make a graph before answering.

You are provided with

Find a library in Python to derive the probable gender from the PhD candidate's first name. Plot the evolution of gender among PhD candidates over the past decades.

Propose another research result that we could derive from this dataset, from the top of your head, which has not been suggested before. All students must have different ideas (discuss with the instructor prior to writing it). +3 bonus on the report if you act on it and display a plot.

4 References

Copeland, S., Penman, A., Milne, R. (2005). Electronic theses: The turning point. Program, 39(3), 185 197. https://doi.org/10.1108/00330330510610546

ElSabry, E. (2017). Who needs access to research? Exploring the societal impact of open access. Revue Française Des Sciences de l'information et de La Communication, 11, Article 11. https://doi.org/10.4000/rfsic.3271

Harnad, S. (2011). Open Access to Research. Changing Researcher Behavior Through University and Funder Mandates. JeDEM - EJournal of EDemocracy and Open Government, 3(1), 33 41. https://doi.org/10.29379/jedem.v3i1.54

Martin, I. (2015). Le signalement des thèses de doctorat. I2D - Information, donnees documents, Volume 52(1), 46 47.

Moxley, J. M. (2001). American universities should require electronic theses and dissertations. Educause Quarterly, (3), 61.

Moyle, M. (2008). Improving access to European e-theses: the DART-Europe Programme. Liber Quarterly, 18(3-4).

Park, E. G., Richard, M. (2011). Metadata assessment in e-theses and dissertations of Canadian institutional repositories. The Electronic Library, 29(3), 394 407. https://doi.org/10.1108/02640471111141124

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., Farley, A., West, J., Haustein, S. (2018). The state of OA: A large-scale analysis of the prevalence and impact of Open Access articles. PeerJ, 6, e4375. https://doi.org/10.7717/peerj.4375

Rabesandratana, T. (2019). The world debates open-access mandates. Science, 363(6422), 11–12. https://doi.org/10.1126/science.363.6422.11

Stanton, K. V., Liew, C. L. (2011). Open Access Theses in Institutional Repositories: An Exploratory Study of the Perceptions of Doctoral Students. Information Research: An International Electronic Journal, 16(4).