Data Wrangling and Preprocessing

Dr. Matthieu Cisel - Bachelor Data Science

Fall 2024

1 Introduction

In this learning unit, students focus on the practical applications of a set of data wrangling operations with R. Students learn these concepts and techniques through a hands-on approach. For the first activity, they are provided with a dataset on learning analytics drawn from a series of MOOCs. The focus of the analysis lays on learners' engagement in an online course. They will have the opportunity to compare their results with the ones that were obtained in this research work carried out in the mid-2010s. You will then focus on the PhD dataset. One of the focuses of the class is to learn how to design sound figures and captions, and describe, in your notebook, the results that are reported in said figures and tables.

2 The MOOC Dataset

2.1 Context

One of the most striking consequences of Massive Open Online Courses (MOOCs) openness is undoubtedly the high heterogeneity of their registrants, whether we think in terms of socioeconomic status, sociocultural background, motivations, or behaviors. Their engagement patterns are as heterogeneous as their profiles, and the monolithic distinction between completers and dropouts is not necessarily appropriate to describe the diversity of situations.

Most of the time, a large proportion of registrants could still represent a significant part of the course activity despite the fact that they do not complete the course. While these questions have attracted considerable attention from researchers and practitioners lately, few studies have focused on the long-term evolutions of these learning engagement patterns in a given course. Increasing attention is laid on the relationships between these engagement patterns, intentions, or sociodemographic variables. These questions are relevant to both course designers who would like to understand ongoing dynamics and wish to adapt course design accordingly, and to researchers who want to capture ongoing trends at a more global scale.

In this exercise, we propose you to analyze a MOOC that has been organized on Canvas before being on Coursera, the MOOC Effectuation. We addressed the question of the evolution of learners' profiles and course dynamics over time. To what extent have engagement patterns and registrants profiles evolved across iterations, and most importantly, how has the relationship between learners' behavior and profiles evolved over time?

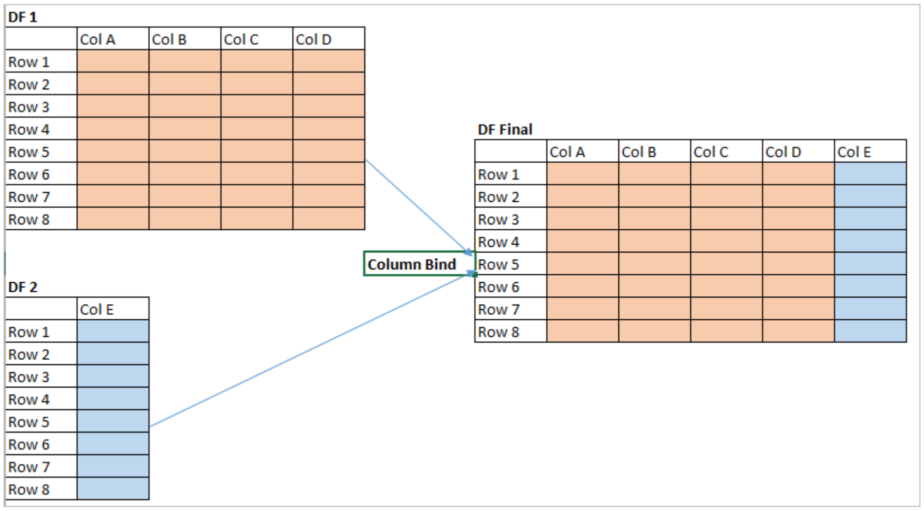
The case study you will analyze here is a five weeks long entrepreneurship course called Effectuation (Professor Philippe Silberzahn, EMLYON Business School), which will thereafter be referred to as MOOC1. It was hosted by a MOOC agency which used the open source LMS Canvas from Instructure.

It was necessary to submit a peer evaluated mid-term assignment and to pass an exam to earn the certificate. In both courses, new course material including quizzes and half a dozen of short videos was made available every week. Variations among iterations were minor. Course designers estimated that completing the course required fifteen to twenty-five hours. Student activity reports, gradebooks and survey responses were downloaded from the platform. Regarding video, consumption, we used a proxy as we considered that the video had been viewed when the page where it was embedded on was opened, regardless of the number of times this page was loaded. We manually removed from, subsequent analyses the videos that were not part of the course strictly speaking, such as weekly introductions or, tutorials. The global activity of the course was defined from the video perspective as the total number of views, without taking into account multiple views, and from the quiz perspective as the total number of submissions, without taking into account multiple submissions.

Participants were asked to fill in a survey at the beginning of the course; response rates ranged from 40 percent, to 60 percent of enrollees. IP addresses were not collected; all available data on countries of residence come from these, surveys; the Human Development Index of these countries were retrieved from U.N data. In both courses, the students who could gain credits by completing the course were excluded from our analyses since they, were not strictly speaking following a self-directed learning approach. They represented a significant contingent, in the case of MOOC2. Participants were categorized based on their level of engagement: those who obtained a, certificate were called “completers”, those who submitted at least one quiz or assignment but did not complete, the course were referred to as “disengaging learners”; those who did not submit any quiz or assignment were, referred to as “auditing learners” if they had viewed at least 10 percent of available course videos, and bystanders, if they fell below this threshold.

2.2 Data wrangling, feature engineering

Patch together the information from usages.effec (logs from all three iterations) and the surveys. Use merge (base), or preferably full-join functions (dplyr library), and well as rbind, or rbind.fill if need be. Do not forget to keep the information on the iteration of the course when patching together the datasets (using the mutate variable from dplyr).



DF 1

	Col A	Col B	Col C	Col D
Row 1	🟧	🟧	🟧	🟧
Row 2	🟧	🟧	🟧	🟧
Row 3	🟧	🟧	🟧	🟧
Row 4	🟧	🟧	🟧	🟧
Row 5	🟧	🟧	🟧	🟧
Row 6	🟧	🟧	🟧	🟧
Row 7	🟧	🟧	🟧	🟧
Row 8	🟧	🟧	🟧	🟧

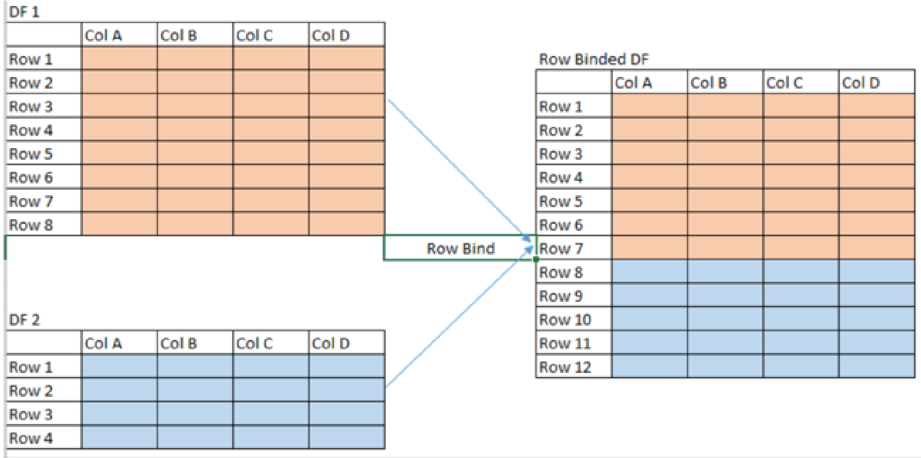
DF 2

	Col E
Row 1	🟦
Row 2	🟦
Row 3	🟦
Row 4	🟦
Row 5	🟦
Row 6	🟦
Row 7	🟦
Row 8	🟦

DF Final

	Col A	Col B	Col C	Col D	Col E
Row 1	🟧	🟧	🟧	🟧	🟦
Row 2	🟧	🟧	🟧	🟧	🟦
Row 3	🟧	🟧	🟧	🟧	🟦
Row 4	🟧	🟧	🟧	🟧	🟦
Row 5	🟧	🟧	🟧	🟧	🟦
Row 6	🟧	🟧	🟧	🟧	🟦
Row 7	🟧	🟧	🟧	🟧	🟦
Row 8	🟧	🟧	🟧	🟧	🟦

Figure 1: Column Bind



DF 1

	Col A	Col B	Col C	Col D
Row 1	🟧	🟧	🟧	🟧
Row 2	🟧	🟧	🟧	🟧
Row 3	🟧	🟧	🟧	🟧
Row 4	🟧	🟧	🟧	🟧
Row 5	🟧	🟧	🟧	🟧
Row 6	🟧	🟧	🟧	🟧
Row 7	🟧	🟧	🟧	🟧
Row 8	🟧	🟧	🟧	🟧

DF 2

	Col A	Col B	Col C	Col D
Row 1	🟦	🟦	🟦	🟦
Row 2	🟦	🟦	🟦	🟦
Row 3	🟦	🟦	🟦	🟦
Row 4	🟦	🟦	🟦	🟦

Row Binded DF

	Col A	Col B	Col C	Col D
Row 1	🟧	🟧	🟧	🟧
Row 2	🟧	🟧	🟧	🟧
Row 3	🟧	🟧	🟧	🟧
Row 4	🟧	🟧	🟧	🟧
Row 5	🟧	🟧	🟧	🟧
Row 6	🟧	🟧	🟧	🟧
Row 7	🟧	🟧	🟧	🟧
Row 8	🟦	🟦	🟦	🟦
Row 9	🟦	🟦	🟦	🟦
Row 10	🟦	🟦	🟦	🟦
Row 11	🟦	🟦	🟦	🟦
Row 12	🟦	🟦	🟦	🟦

Figure 2: Row Bind

You must deal with the fact that all datasets do not have the same number of columns. In R (dplyr), use the select command to keep, in the surveys, only the variables that you are going to use (HDI, gender, socioeconomic status).

You will need to use an rbind, a cbind, and merge data following the scheme provided in the figure below.

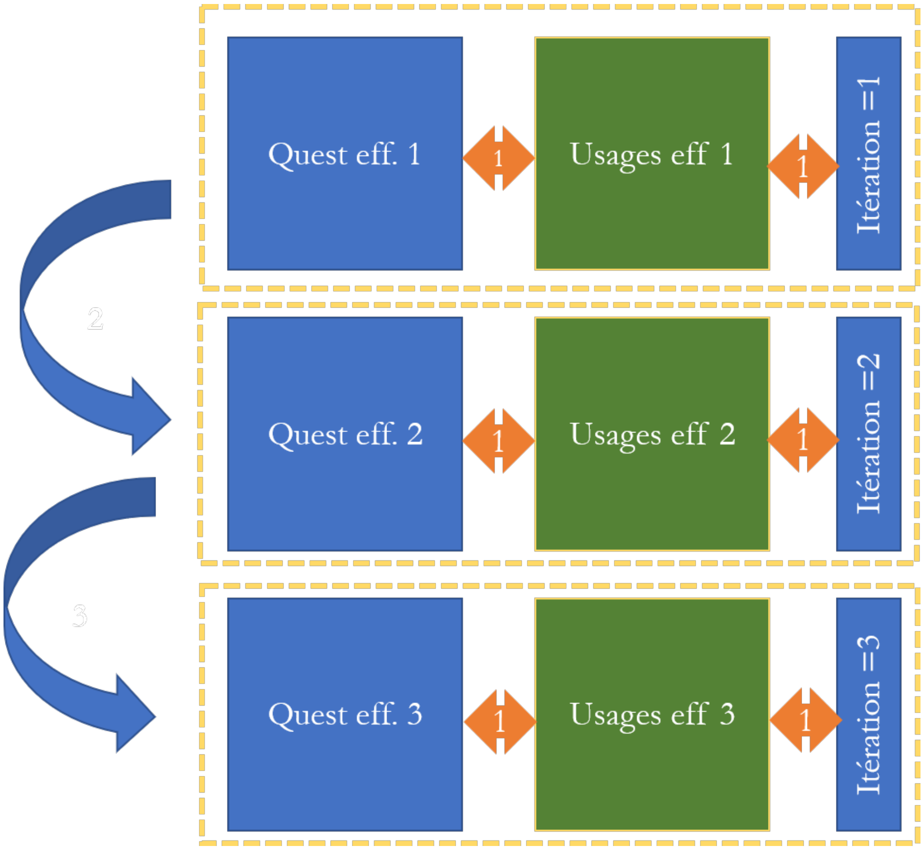


Figure 3: Merging the dataset

2.3 Describing behavior in the courses

Participants were categorized based on their level of engagement: those who obtained a certificate were called "completers", those who submitted at least one quiz or assignment but did not complete the course were referred to as "disengaging learners"; those who did not submit any quiz or assignment were referred to as "auditing learners" if they had viewed more than 10 percent of available videos, and bystanders (Anderson et al. 2014) if their fell below this threshold.

Present the proportion of Disengaging learners, auditing learners, bystanders, and completers. This typology of learners is inspired by a paper from Kizilcec et al. (2013) : Deconstructing Disengagement Analyzing Learner Subpopulations in Massive Open Online Courses.

Represent the proportions of learners as a pie chart and/or as barcharts. The db is called usages.effec1 for the first iteration. When describing the results, use numbers and sentences.

2.4 Datacamp classes

To complete the tasks at hand, you must complete the following classes in Datacamp. Certificates will be required.

1. Web scraping in Python
2. Introduction to Data Manipulation with dplyr
3. Joining Data with dplyr
4. Introduction to Data Visualization with ggplot2
5. Data Manipulation in R assessment (score ¿ 150)

3 Exploring the PhD dataset

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. 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 work to be published in open access reviews. 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. 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. 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.

France launched theses.fr and its associated archive TEL. Research based on metadata in online archives or repositories is deeply constrained by the nature and the quality of the information collected on manuscripts. In this learning unit, you must first scrap the website to obtain a significant amount of entries (at least 10.000) to build the first draft of the dataset. In a second step, you are provided with the PhD dataset, which comprises around 480.000 entries of students who defended their PhD in France from 1985 to 2020. You will be required to identify relevant issues during the exploration phase of the dataset. We provide you with a limited amount of instructions; your goal is to provide graphs and statistics in order to explain what the issues are (missing or unreliable data).

You will hand out an ipynb version of your notebook. This is a speed run. The goal is to assess how fast you are in detecting and explaining issues, and how rigourous your thinking is.

3.1 Scraping

Use beautifulsoup, (Scrapy or selenium only if you are advanced) to scrap all relevant information for 1000 theses, from the theses.fr website. You must deliver a database in csv called your_name_PhD1000.csv. You can use the first chapter of the class called Intermediate Importing Data in Python in Datacamp. The completion of this section will depend upon the time left after the work on MOOCs.

3.2 Defence dates - unreliable data and missing data

There are missing data in defence dates. After a classic graph on missing data at the scale of the databases, find patterns with missing data (overall), and produce a specific graph that explains why defence data can be missing.

Next, create a graph to visualize the distribution of missing data within the dataset.

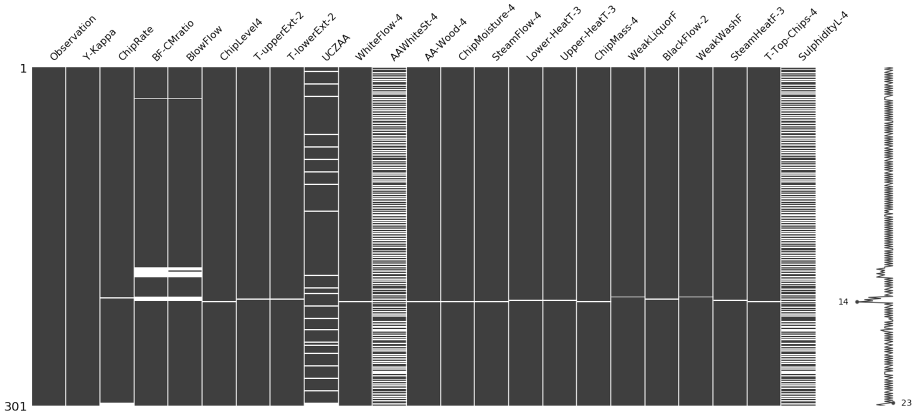


Figure 4: Example of missing data visualization

Generate an initial heatmap to represent the co-occurrence of missing data.

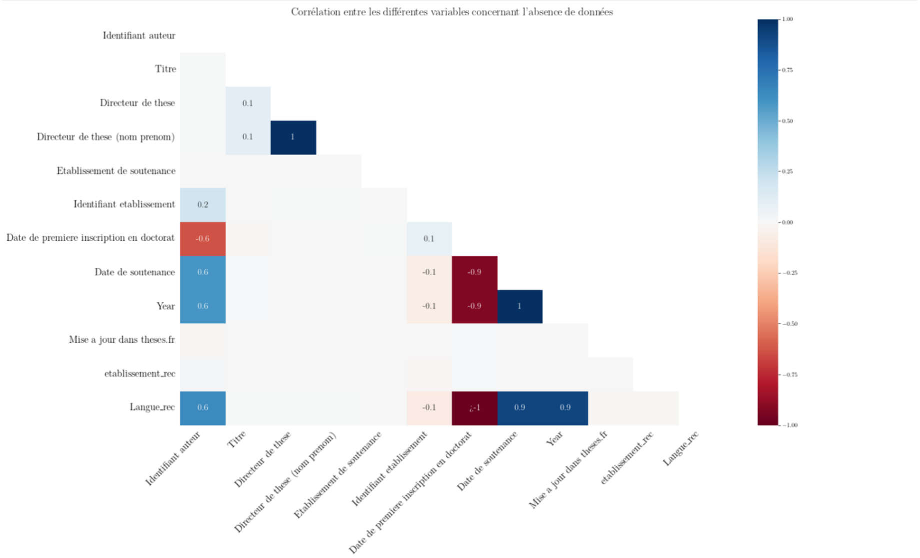


Figure 5: Visualization of patterns in missing data

Create a second heatmap through dplyr manipulations (with color depending on the percentage of missing data) while selecting ”status” as the variable on the x-axis. This way, you can contrast the levels ”enCours” and ”soutenue.” Choose a subset of variables on the y-axis that you find relevant for the analysis.

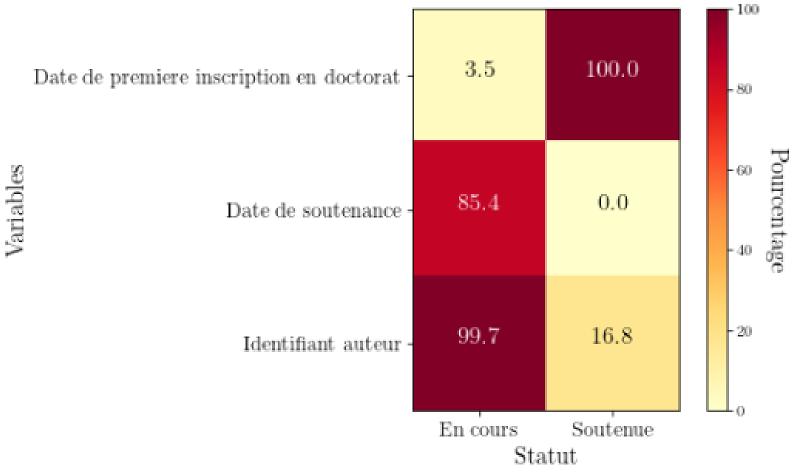
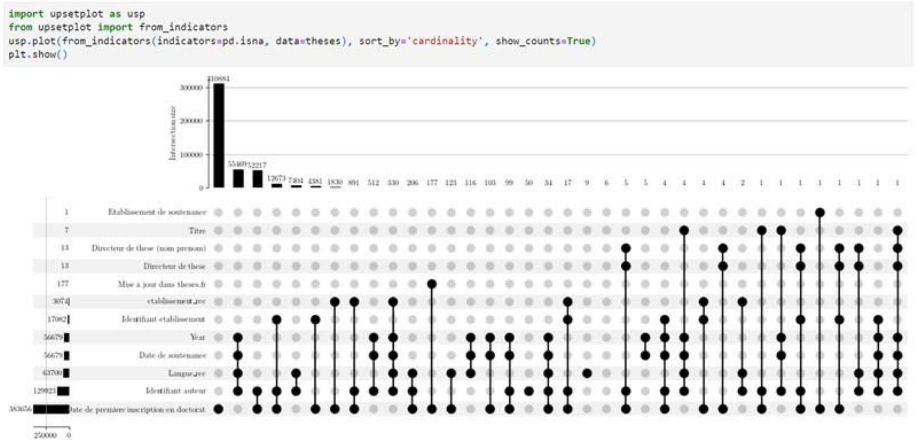


Figure 6: Visualization of patterns in missing data - zoom on categories

Do you observe any regularities in the missing data pattern? Try to visualize them. For example, is there a connection between the thesis defense date and the thesis launch date? How could you explain this pattern?

Lastly, analyze the missing data patterns using the UpSet package. Interpret your findings.



```
import upsetplot as usp
from upsetplot import from_indicators
usp.plot(from_indicators(indicators=pd.isna, data=theses), sort_by='cardinality', show_counts=True)
plt.show()
```

Figure 7: Example of missing data visualization

Use the column ”Year” to show how the number of defences evolves over time. How do you explain the dip in 2020 ?

At what period of the year do PhD candidates tend to defend ? There is an issue with defence dates. Find the best way to represent this issue, and then find the correct way of showing which is the favourite month to defend the thesis (after taking this issue into account).