Minimalist Data Wrangling with Python [DRAFTv0.1]

Release [DRAFTv0.1]

Marek Gagolewski

2022-03-27T15:21:53+1100

Contents

Minimalist Data Wrangling with Python is a very-early-and-rough-draft of the forthcoming (ETA 2023) textbook by Marek Gagolewski[1]. It is distributed in the hope that it will be useful. If you detect any bugs or typos, please share them by email[2]. Although available online, this is a whole course, and should be read from the beginning to the end. In particular, refer to the Preface for general introductory remarks. Enjoy.

You can access this book at:

- https://datawranglingpy.gagolewski.com/ (a browser-friendly version)
- https://datawranglingpy.gagolewski.com/datawranglingpy.pdf (PDF)
- https://github.com/gagolews/datawranglingpy (source code)

[1] https://www.gagolewski.com/
[2] https://github.com/gagolews/datawranglingpy/blob/master/CODE_OF_CONDUCT.md
[3] https://www.gagolewski.com
[4] https://creativecommons.org/licenses/by-nc-nd/4.0/

0

Preface

0.1 The Art of Data Wrangling

The broadly-conceived data science aims at making sense of and generating predictions from data that have been collected in large quantities from various sources, e.g., physical sensors, files, databases, or (pseudo)random number generators. It can take different forms, e.g., vectors, matrices and other tensors, graphs, audio/video streams, text, etc. With the advent of the internet era, data have become ubiquitous.

Exercise 0.1 *Think of how much information you consume and generate when you interact with your social media or news feeds every day.*

Here are some application domains where data-driven decision making, modelling, and prediction has already proven itself very useful:

- financial services (banking, insurance, investment funds),
- real estate,
- pharmaceuticals,
- transportation,
- retail,
- healthcare,
- food production.

Okay, to be frank, the above list was generated by duckduckgoing the "biggest industries" query. That was a very easy task; data science (and its very different flavours, including statistics, operational research, machine learning, artificial intelligence, and so forth) is everywhere. Basically, wherever we have data and there is a need to improve some processes or discover new aspects about a problem domain, there is a place for data-driven solutions.

Of course, it's not all about business revenue (luckily). We can do a lot of great work for greater good; with the increased availability of open data, everyone can be a reporter, an engaged citizen that seeks truth. There are NGOs. Finally, there are researchers (remember that the main role of most universities is still to spread the advancement of knowledge and not to make money!) that need these methods to make new discoveries, e.g., in psychology, economics, sociology, agriculture, engineering, biotechnology, pharmacy, medicine, genetics, you name it.

Data rarely come in a *tidy* and *tamed* form. Performing accurate exploration and modelling heavily relies on **data wrangling**, which is the very broad process of appropriately preparing raw data for further analysis.

And thus, in this course, we are going to explore methods for:

- performing exploratory data analysis, including aggregating and visualising numerical and categorical data,
- working with different types of data (e.g., text, time series) gathered from structured and unstructured sources,
- cleaning data by identifying outliers,
- handling missing data,
- transforming, selecting, and extracting features, dimensionality reduction,
- identifying naturally occurring data clusters,
- applying sampling techniques,
- data modelling using basic machine learning algorithms,
- maintaining data privacy and exercising ethics in data manipulation.

0.2 Aims and Scope

Most of the time during the course of this course, we will be writing code in Python[5]. The 2021 StackOverflow Developer Survey[6] lists it as the 2nd most popular programming language used nowadays (slightly behind JavaScript).

Over the last years, Python has proven a very robust choice for learning and applying

[5] https://www.python.org/

[6] https://insights.stackoverflow.com/survey/2021#technology-most-popular-technologies

data wrangling techniques. This is possible thanks to the famous[7] high quality packages written by the devoted community of open source programmers, including but not limited to *numpy*[8], *scipy*[9], *pandas*[10], *matplotlib*[11], *seaborn*[12], and *sklearn*[13].

Important: Note that we will introduce the Python language from scratch and that we do not require any prior programming experience. Nevertheless, learning how to code is hard work; if you want to succeed, you will have to spend a decent amount of time getting your hands dirty by writing programs which solve the suggested problem sets, studying technical manuals, and so forth. One does not became a qualified and respected engineer by simply *reading* online tutorials or books.

Of course, Python and third-party packages written therein is amongst many software tools which can help gain new knowledge from data. Other open source choices include, e.g., R[14] and Julia[15]. There are also some commercial solutions available on the market, but we believe that ultimately all software should be free[16].

Note: We put great emphasis on developing *transferable skills* so that all that we learn here can be then applied quite easily in other environments. In other words, this is a course on data wrangling (*with* Python), and not *on* Python (with examples in data wrangling). This is not a book of recipes. We are aiming for understanding and becoming independent learners.

The skills we are going to develop in this course are *fundamental* for the success in the numerous jobs available in our industry all over the world. And data engineers, data scientists, machine learning specialists, statisticians, and business analysts are amongst the most well-paid specialists[17]. Money does not bring joy, but luckily it is a very interesting domain anyway!

All that we shall learn can be used for improving different processes, in research, helping NGOs, debunking false news or wishful thinking, maintaining quality of various

[7] https://insights.stackoverflow.com/survey/2021#other-frameworks-and-libraries
[8] https://numpy.org/
[9] https://scipy.org/
[10] https://pandas.pydata.org/
[11] https://matplotlib.org/
[12] https://seaborn.pydata.org/
[13] https://scikit-learn.org/
[14] https://www.r-project.org/
[15] https://julialang.org/
[16] https://www.gnu.org/philosophy/free-sw.en.html
[17] https://insights.stackoverflow.com/survey/2021#other-frameworks-and-libraries

industrial processes, and doing any other good deeds for the advancement of humanity.

We're going to study many methods and algorithms that stood the test of time and that continue to inspire the researchers and practitioners. After all, many "complex" algorithms are merely variations on or clever combinations of the most basic ones. You might not see it now, but this will become evident as we progress.

Most importantly, however, we will get to know their limitations, which they are many. Being sceptical and cautious is one of the traits of a good scientist!

Note that we will definitely not be avoiding mathematical notation so as to maintain a healthy level of generality. Mathematics is both a universal tool and a language for describing the methods for processing various data structures and analysing the properties thereof. The people fluent in mathematics are those who have invented or derived most of the methods discussed herein, we should thus be too.

0.3 About the Author

I, Marek Gagolewski[18] (pronounced like Mark Gaggle-Eve-Ski), am currently a Senior Lecturer in Applied AI at Deakin University in Melbourne, VIC, Australia and an Associate Professor in Data Science (on long-term leave) at Faculty of Mathematics and Information Science, Warsaw University of Technology, Poland.

I'm actively involved in developing *usable* free (libre, independent) and open source software, with particular focus on data science and machine learning. He is the main author and maintainer of stringi[19] – one of the most often downloaded R packages that aims at natural language and string processing as well as the Python and R package genieclust[20] implementing the fast and robust hierarchical clustering algorithm *Genie* with noise point detection.

I'm an author of over 80 publications on machine learning and optimisation algorithms, data aggregation and clustering, statistical modelling, and scientific computing. I taught various courses related to R and Python programming, algorithms, data science, and machine learning in Australia, Poland, and Germany.

Minimalist Data Wrangling with Python bases on my experience as an author of a quite successful textbook *Przetwarzanie i analiza danych w języku Python* (Data Processing and

[18] https://www.gagolewski.com

[19] https://stringi.gagolewski.com

[20] https://genieclust.gagolewski.com

Analysis in Python), [[GBC16]] that I have written with my former (successful) PhD students Maciej Bartoszuk and Anna Cena (in Polish, 2016, published by PWN). The current one is a completely different work, however its predecessor served as a great testbed for many ideas conveyed here. They have also been battle-tested at Warsaw University of Technology, Data Science Retreat (Berlin), and Deakin University (Melbourne).

0.4 Acknowledgements

This book has been prepared with TeX (XeLaTeX) and Sphinx. Python code chunks have been processed with the R package *knitr*. A little help of Makefiles and custom shell scripts dotted the *j*'s and crossed the *f*'s.

Bibliography

[GBC16] Marek Gagolewski, Maciej Bartoszuk, and Anna Cena. *Przetwarzanie i analiza danych w języku Python (Data Processing and Analysis in Python)*. Wydawnictwo Naukowe PWN, Warsaw, Poland, 2016. ISBN 978-83-01-18940-2. in Polish. URL: https://github.com/gagolews/Analiza_danych_w_jezyku_Python.