The Journal of Open Source Education

Reproducible Data Science with Python: An Open Learning Resource

Valentin Danchev[1, 2]

1 Department of Sociology, University of Essex, United Kingdom **2** School of Business and Management, Queen Mary University of London, United Kingdom

DOI: 10.21105/jose.00156

Software

- Review
- Repository
- Archive

Submitted: 19 November 2021
Published: 24 October 2022

Summary

This paper describes a computational learning resource on Reproducible Data Science with Python. The resource provides an accessible, hands-on introduction to data science techniques, skills, and workflows necessary to perform open, reproducible, and ethical data analysis. By using research problems of real-world relevance (such as vaccine hesitancy and the impact of COVID-19 lockdown measures on human mobility) and real-world social data (including anonymised mobility data from digital sources and recent COVID-19 survey data), the resource encourages students to use open-source tools and coding to learn from diverse and large social data sources.

The learning resource aims to minimise barriers to entry for students from social sciences, public health, and related fields. With no software installation and setup requirements, students can start coding from their web browser using free and open-source software (FOSS), including the Python programming language, Jupyter notebook, and Markdown. Through real-world data applications, students are introduced to the open source Python ecosystem of libraries for data science—including `pandas` (McKinney, 2010), `seaborn` (Waskom, 2021), `scikit-learn` (Pedregosa et al., 2011), `statsmodels` (Seabold & Perktold, 2010), and `networkX` (Hagberg et al., 2008)—and learn about open and reproducible workflow, data wrangling, data exploration and visualization, pattern discovery (e.g., clustering), prediction and machine learning, causal inference, network analysis, and data ethics.

Statement of need

The importance of data science education for drawing conclusions from diverse data sources in a transparent, reproducible, and ethical manner and for building data literacy skills is now widely recognised (Adhikari et al., 2021; Arnold et al., 2019; Danyluk et al., 2021; National Academies of Sciences & Medicine, 2018; Stoudt, 2021). For example, the National Academies of Science, Engineering, and Medicine's 2018 'Data Science for Undergraduates' report (National Academies of Sciences & Medicine, 2018) identified the importance of workflow, reproducibility, and ethical problem solving in data science undergraduate curriculum. A key challenge ahead for the community of data science educators, and particularly those at the intersection of data science and social sciences, is how to democratise the knowledge and skills needed for conducting reproducible data science and to engage students—with little to no programming experience and from diverse social and academic backgrounds—in accessible, inclusive, and cross-disciplinary data science learning. To address this and related challenges, successful examples of data science curriculum (e.g., Adhikari et al., 2021; Timbers et al., 2022) have been developed. The learning resource presented in this paper addresses the data-science democratisation challenge through the use of down-to-earth research questions and real-world social data sets about the COVID-19 pandemic. Both real-world questions and data can encourage students to engage with hands-on computation, data science techniques, reproducible data analysis workflow, and data ethics. The democratisation of reproducible and ethical data science is important for empowering students (and citizens) to use, analyse, and learn from large data sets as well as to critically evaluate data, models, and their social impact.

Learning objectives

The learning resource aims to enable students to:

- Freely use open-source computational tools—e.g., Python, Jupyter Notebook, Markdown—for data analysis.
- Build a transparent and reproducible research workflow.
- Perform data analysis and critically interpret and openly communicate research process, objects, and results.
- Study research questions of societal importance by wrangling, exploring, visualising, and modeling a variety of real-world tabular and network data using Python libraries.
- Apply basic models for machine learning, causal inference, and network analysis to real-world data.
- Identify and deal with methodological issues of overfitting, selection bias, collider bias, and confounding.
- Articulate and address issues of data ethics and social impact of data science models.
- Write clean, reusable, and reproducible code in Python.

Content

The learning resource is designed around an understanding of data science as the use of coding to draw conclusions from diverse data sets by solving five classes of tasks (Adhikari et al., 2021; Hernán et al., 2019):

- Data preprocessing—preparing data for analysis using techniques for data cleaning, data wrangling, and data transformation.
- Description—discovering patterns in data using exploratory data analysis, visualisation, and automated discovery techniques.
- Prediction—using information about outcomes we know to make informed guesses about unknown outcomes by applying techniques from simple regression to supervised machine learning.
- Inference—quantifying our degree of certainty to determine whether what we find in our data will hold among different scenarios using resampling methods and related techniques.
- Causal data analysis—studying cause-and-effect questions via the application of causal graphs, counterfactuals, and causal inference techniques.

The resource does not aim to cover a single data science task in detail but introduces students to these tasks with a focus on real-world data and applications, hands-on computation, and reproducible data analysis.

The content follows a typical data science lifecycle (Lau et al., 2021), in which students would begin with a research question, and then select their data set(s), preprocess the data, perform descriptive analysis to explore basic features of their data, and then model their data to predict an outcome or establish causal effect. Transparency of research process and computational reproducibility are embedded throughout the data science lifecycle.

Prerequisites

Prior knowledge of programming is not required as coding for data analysis is taught from first principles. Background in mathematics or statistics are not required beyond basic algebra and descriptive statistics.

Instructional design

The learning resource is provided in the form of Jupyter textbook that consists of ten chapters, each of which is an independent Jupyter notebook. Each notebook provides an introduction of the chapter's topic. To accommodate students' different styles of learning, each notebook also points to four categories of learning materials—i.e., articles, books, videos, and tutorials—that provide learners with background knowledge on the topic before immersing in hands-on coding. Coding and data analysis are motivated by data-centered research questions such as "How has human mobility differed across the three lockdowns in the United Kingdom during the COVID-19 pandemic?" and "Can we predict people who are unlikely to take a coronavirus vaccine from socio-demographic and health features." A variety of real-world (and daily updated) data sets are brought to bear on these research questions, including the COVID-19 Google Community Mobility Reports (Aktay et al., 2020) and the Understanding Society: COVID-19 Study (Burton et al., 2020). Methodologies (e.g., supervised machine learning, causal inference, network analysis) and techniques (e.g., cross-validation) are briefly described such that students are empowered to apply them without being overwhelmed by technical detail. Students can engage with the notebooks, interactively execute the code, examine the outputs, and work on the hands-on coding exercises in an active-learning process. The code in the learning resource is validated against the PEP 8 style guide for Python code by executing on all notebooks the automated code formatter Black and style guide checker flake8 via nbQA.

The resource is designed to scaffold a reproducible and transparent research workflow by integrating research questions, data inputs, computer code, documentation, data analysis, visualizations, and narrative text in a single document. This integration of various research objects is made possible by the open-source Jupyter Notebook (Kluyver et al., 2016)—a user-friendly, free, open-source, interactive web tool that implements the notion of "literate programming" (Knuth, 1984) and is widely used across research and education (Perkel, 2018). While the Jupyter computational notebook does not guarantee reproducibility (Guzharina, 2020), it fosters reproducible workflow and good research practices (Rule et al., 2019) by enabling students to describe the process of data analysis (not just the "final" results but also the "dead ends"), make transparent choices of parameter selection, and document research outputs, however "useful" or "(un)expected" they may seem.

Experience of use

The learning resource can be used for different modes of learning, including a semester-long module, training workshop, and independent study. So far, elements of the learning resource have been used in a 10-week module delivered in the Spring term of the academic years 2020–21 and 2021–22 to third year undergraduate students at the Department of Sociology, University of Essex. More recently, core elements of the resource have formed the backbone of a 1.5-hour hands-on training on reproducible workflow with dynamic documents as part of the Research Transparency and Reproducibility Training (RT2), August 23–September 3, 2021, organised by the Berkeley Initiative for Transparency in the Social Sciences. The Jupyter notebook used in the training is publicly available and can be accessed via the RT2's repository on the Open Science Framework (https://osf.io/5neky/) and via GitHub (https://github.com/valdanchev/dynamic-documents-with-jupyter-notebook).

Self-guided learning

Learners can access the resource in its entirety on the dedicated website https://valdanchev.github.io/reproducible-data-science-python. The resource website is built using Jupyter Book and is deployed to GitHub Pages from the resource's public GitHub repository. To interactively work with the code, learners can access the interactive versions of the Jupyter notebooks via MyBinder and Colab with no setup or download requirements.

- MyBinder (Jupyter et al., 2018) is a free open-source online service that lets you open and execute Jupyter notebooks and work with the code interactively in your browser. MyBinder uses the `requirements.txt` file from the resource's public GitHub repository, which lists all the packages and package versions used in the resource, to build a live environment that includes the package dependencies and versions used in the original notebooks, enabling reproducibility and minimising possible errors due to package updates. MyBinder is suited for relatively short sessions—a user session can last up to 6 hours and will be shut down automatically after more than 10 minutes of inactivity. Notebooks launched on MyBinder are non-persistent—any changes will be lost after user session times out unless the user downloads the notebook. Regarding access and user privacy, MyBinder is public service that requires no log-in and does not keep track of user data. All code and data that are run during a session are destroyed once the session finishes. More information on how MyBinder ensures user privacy is provided by the MyBinder team in their Frequently Asked Questions (FAQs).
- Colab is an environment from Google Research that runs Jupyter notebooks on the Google Cloud, allowing interactive work with notebooks from the browser. Similar to MyBinder, Colab is free of charge but requires a Google account and a log-in (for more information, see Colab's Frequently Asked Questions). Colab notebooks likely run faster and have longer session lifetimes (up to 12 hours) compared to MyBinder. In comparison to MyBinder, which reproduces the computing environment and package dependencies used in the original notebooks, Colab opens notebooks in a new environment with preinstalled package dependencies. At the time of use, the packages and package versions preinstalled on Colab may differ from the packages and package versions used in the original notebooks, introducing possible errors. To install the original package dependencies, Colab users would need to run the `requirements.txt` file following the instructions in the README.md file on the resource's public GitHub repository.

Acknowledgements

I thank the participants at the 2021 National Workshop on Data Science Education (organised by UC Berkeley's Division of Computing, Data Science, and Society) and students who studied elements of the learning resource at the Research Transparency and Reproducibility Training (RT2) Virtual 2021 (organised by Berkeley Initiative for Transparency in the Social Sciences (BITSS)) and at the Department of Sociology, University of Essex, for helpful feedback. I also thank Kirils Makarovs and Hamid Nejadghorban for help with teaching earlier iterations of the learning resource. Finally, I would particularly like to thank the two JOSE reviewers, Tom Donoghue and Jens Lechtenbörger, for helpful comments and constructive suggestions.

References

Adhikari, A., DeNero, J., & Jordan, M. I. (2021). Interleaving Computational and Inferential Thinking: Data Science for Undergraduates at Berkeley. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.cb0fa8d2

Aktay, A., Bavadekar, S., Cossoul, G., Davis, J., Desfontaines, D., Fabrikant, A., Gabrilovich, E., Gadepalli, K., Gipson, B., Guevara, M., Kamath, C., Kansal, M., Lange, A., Mandayam, C., Oplinger, A., Pluntke, C., Roessler, T., Schlosberg, A., Shekel, T., … Wilson, R. J. (2020). *Google COVID-19 Community Mobility Reports: Anonymization Process Description (version 1.1)*. https://doi.org/10.48550/arXiv.2004.04145

Arnold, B., Bowler, L., Gibson, S., Herterich, P., Higman, R., Krystalli, A., Morley, A., O'Reilly, M., Whitaker, K., & others. (2019). The Turing Way: A Handbook for Reproducible Data Science. *Zenodo*. https://doi.org/10.5281/zenodo.3233853

Burton, J., Lynn, P., & Benzeval, M. (2020). How Understanding Society: The UK Household Longitudinal Study Adapted to the COVID-19 Pandemic. *Survey Research Methods*, *14*(2), 235–239. https://doi.org/10.18148/srm/2020.v14i2.7746

Danyluk, A., Leidig, P., McGettrick, A., Cassel, L., Doyle, M., Servin, C., Schmitt, K., & Stefik, A. (2021). Computing Competencies for Undergraduate Data Science Programs: An ACM Task Force Final Report. *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, 1119–1120. https://doi.org/10.1145/3408877.3432586

Guzharina, A. (2020). We Downloaded 10,000,000 Jupyter Notebooks from GitHub — This is What we Learned. In *The JetBrains Datalore Blog*. https://blog.jetbrains.com/datalore/2020/12/17/we-downloaded-10-000-000-jupyter-notebooks-from-github&-this-is-what-we-learned/

Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). *Exploring Network Structure, Dynamics, and Function using NetworkX*. 11–15. http://conference.scipy.org/proceedings/SciPy2008/paper_2

Hernán, M. A., Hsu, J., & Healy, B. (2019). A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks. *CHANCE*, *32*(1), 42–49. https://doi.org/10.1080/09332480.2019.1579578

Jupyter, Project, Bussonnier, Matthias, Forde, Jessica, Freeman, Jeremy, Granger, Brian, Head, Tim, Holdgraf, Chris, Kelley, Kyle, Nalvarte, Gladys, Osheroff, Andrew, Pacer, M., Panda, Yuvi, Perez, Fernando, Ragan-Kelley, Benjamin, & Willing, Carol. (2018). Binder 2.0 — Reproducible, Interactive, Sharable Environments for Science at Scale. In Fatih Akici, David Lippa, Dillon Niederhut, & M. Pacer (Eds.), *Proceedings of* *the 17th Python in Science Conference* (pp. 113–120). https://doi.org/10.25080/Majora-4af1f417-011

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., & Willing, C. (2016). *Jupyter Notebooks — A Publishing Format for Reproducible Computational Workflows* (F. Loizides & B. Schmidt, Eds.; pp. 87–90). IOS Press. https://doi.org/10.3233/978-1-61499-649-1-87

Knuth, D. E. (1984). Literate Programming. *The Computer Journal*, *27*(2), 97–111. https://doi.org/10.1093/comjnl/27.2.97

Lau, S., Gonzalez, J., & Nolan, D. (2021). *Principles and Techniques of Data Science*. http://www.textbook.ds100.org

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 51–56). https://doi.org/10.25080/Majora-92bf1922-00a

National Academies of Sciences, Engineering, & Medicine. (2018). *Data Science for Undergraduates: Opportunities and Options*. The National Academies Press. https://doi.org/10.17226/25104

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., & others. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://doi.org/10.48550/arXiv.1201.0490

Perkel, J. M. (2018). Why Jupyter is Data Scientists' Computational Notebook of Choice. *Nature*, *563*(7732), 145–147. https://doi.org/10.1038/d41586-018-07196-1

Rule, A., Birmingham, A., Zuniga, C., Altintas, I., Huang, S.-C., Knight, R., Moshiri, N., Nguyen, M. H., Rosenthal, S. B., Pérez, F., & Rose, P. W. (2019). Ten Simple Rules for Writing and Sharing Computational Analyses in Jupyter Notebooks. *PLOS Computational Biology*, *15*(7), 1–8. https://doi.org/10.1371/journal.pcbi.1007007

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. *Proceedings of the 9th Python in Science Conference*. https://doi.org/10.25080/Majora-92bf1922-011

Stoudt, V. N. A. M., Sara AND Vásquez. (2021). Principles for Data Analysis Workflows. *PLOS Computational Biology*, *17*(3), 1–26. https://doi.org/10.1371/journal.pcbi.1008770

Timbers, T. A., Campbell, T., & Lee, M. (2022). *Data Science: A First Introduction*. Chapman & Hall/CRC. https://doi.org/10.1201/9781003080978

Waskom, M. L. (2021). Seaborn: Statistical Data Visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021