The Journal of Open Source Education

An Open-Source Active Learning Curriculum for Data Science in Engineering

Zachary del Rosario[1]

1 Assistant Professor of Engineering and Applied Statistics, Olin College of Engineering

DOI: 10.21105/jose.00117

Software

- Review
- Repository
- Archive

Submitted: 21 October 2020
Published: 13 March 2022

Summary

This work provides open-source content for an active learning curriculum in data science. The scope of the content is sufficient for a full-semester introduction to scientifically reproducible statistical computation, data wrangling, visualization, basic statistical literacy, and data-driven modeling. The content is broken into short **exercises** that introduce new concepts, and longer **challenges** that encourage students to develop those skills in an open-ended context.

Statement of Need

As of writing (Fall 2021), Data Science is an exciting new area of study: Many students seek to learn how to obtain, wrangle, and make sense of data, and instructors are working to both define and teach this field. An explosion of resources is available to both teachers and learners—to name just a few—published books (Grolemund & Wickham, 2018), course materials (Adhikari et al., 2021; Timbers et al., 2021), and YouTube channels such as StatQuest (Starmer, n.d.). Much of this content is excellent and informative, and the broader proliferation of open educational resources contributes positively to a more equitable learning environment.

However, one of the most profound findings from the science of learning is the importance of *active learning* for student outcomes; *active learning* is a pedagogical style that has students actively engaged in the process of learning, rather than simply reading a book, skimming a blog post, or watching a lecture. Notably, active learning has been shown to be superior to a purely passive form of instruction (Freeman et al., 2014). Passive resources are an essential component of self-study and are useful for reference, but curricular materials for active learning-based instruction are necessary for an instructor seeking to maximize student learning. This project seeks to provide active learning materials for an instructor seeking to teach data science *to engineers*.

The content in this project has been used to teach a Data Science course at Olin College of Engineering; a small engineering college outside Boston that emphasizes project-based and active learning pedagogies. Since this course caters to engineering students, it has a stronger emphasis on engineering problems than is typically found in data science materials. This includes analyzing datasets on material properties, basic reliability analysis, and using fitted models to make engineering predictions. This curriculum also makes use of concepts from statistics education to arm students with the conceptual tools to recognize and handle different forms of variability that typically arise in engineering problems (Aggarwal et al., 2021; Wild & Pfannkuch, 1999).

This work provides a two-tiered set of content for active learning of data science. This content is appropriate for both instructors seeking to teach data science using a flipped classroom model (Bishop et al., 2013), and for individual learners seeking a self-study course in data science. Additionally, since the materials herein are released under a permissive open-source license, other instructors are welcome (and encouraged) to rearrange and modify the content provided.

Story: How This Was Made

As a PhD student I took a course called *The Data Challenge Lab* (DCL) (Stanford Data Lab, n.d.). This was one of the most inspiring courses I've ever taken, both for the direct learning I gained and for the pedagogical innovations the course employed. The original DCL course focused on *exploratory data analysis* (EDA) (Tukey, 1977): a subfield of statistics that emphasizes hypothesis *generation* over hypothesis *testing*. Having taken several traditional statistics courses prior to the DCL, I found the skill set to be a beautiful complement to the largely mathematical treatment of statistical ideas I had encountered.

Pedagogically, the DCL was based on many sound principles from the learning sciences. There were essentially no lectures; the course leaned heavily into active learning (Freeman et al., 2014). Content was presented in a sequence of small, hands-on exercises and applied in more difficult challenges. Exercises were sequenced to rotate through topics, leveraging the benefits of interleaving (Lang, 2016). Students were assigned to small groups (4–5) to form "learning teams," which promoted motivation to learn through making the learning social (Lang, 2016). The basic design of the present materials is based heavily on this DCL design.

Now as a faculty member at Olin College, my teaching and research aims are to help engineers reason about uncertainty: I designed a Data Science course to give undergraduate engineers a solid foundation in EDA and statistical thinking. Since Olin students (and most engineers) only take a single course in statistics, I used the DCL design as a starting point and exchanged some of the data-manipulation content for statistical ideas. In the summer of 2020, I built a set of exercises and challenges inspired by—but distinct from—the DCL course. The result is a Data Science course that focuses less on a traditional introduction to probability theory, but instead gives students lots of practice studying real datasets, builds an "informal" understanding of probability to ground statistical inference (Moore, 1997), and contextualizes inferential ideas in real scenarios.

Pedagogical Design

The full set of desired learning outcomes is documented in the project repository, but at a high level the learning goals of this content are for students to develop:

1. The ability to set up and maintain a scientifically reproducible data science workflow,
2. The skills and self-efficacy to load, tidy, and access data,
3. The skills to visualize data, for both exploration and communication,
4. The tools and mindset to think statistically,
5. The skills and understanding to fit and interpret data-driven models, and
6. The skills to communicate results clearly and productively.

The content is organized into two levels: *exercises* and *challenges*.

Exercises are designed to be small (1/2 to an hour long) hands-on introductions to particular topics; for instance, exercise `e-data01-isolate` introduces the concept of isolating rows and columns of a dataset. While many exercises point to an external reading, exercises mainly provide students with hands-on practice on new concepts; for instance `e-data01-isolate` has students work with a dataset of flights to select columns matching the pattern `"*_time"` and rows matching the destination `"LAX"`. The exercise solutions are available in a web-based solution manual, and are intended to be provided to students *from the beginning of the course*. Many of these exercises contain unit tests, which allow students to check the correctness of their work immediately—this provides instantaneous formative feedback.

Challenges are designed to be substantial (~3 hours or longer) hands-on elaborations of concepts learned in the exercises. Each challenge has students explore and answer questions about a dataset, with subsequent challenges increasing in complexity. For instance `c01-titanic` has students study the built-in Titanic dataset, while `c06-covid19` has students pull and join data from the New York Times and the US Census Bureau. While the challenge assignment files are openly available in the repository, the challenge solutions are withheld to discourage cheating—instructors may contact the author to obtain the challenge solution files. Providing the challenges without solutions is intended as an opportunity for summative feedback.

The exercises are provided in a recommended sequence that rotates through the broad learning outcomes listed above; this is to leverage the benefits of interleaving topics (Lang, 2016). For convenience, the exercises are renamed to follow this sequence and hosted in a build branch in the repository. However, tools are provided to re-sequence the materials as-desired (see 'Adapting These Materials').

Use of Materials in an Undergraduate Data Science Course

I have used this content to teach a semester-long Data Science course at Olin College for engineering undergraduates. This course satisfies their probability and statistics degree requirement, and is typically taken by students in their second or third year, with some senior students.

These materials are designed for use in a course model similar to a flipped-classroom (Bishop et al., 2013). Students complete exercises outside class and during unstructured class time, so their first introduction to course content includes an active learning component. During class time students continue work on exercises and challenges; members of the teaching team (the instructor and course assistants) serve as coaches to help students with assignments.

Students are assigned to small learning teams (4–5 members) to form social units: This was particularly helpful during remote teaching at the height of the covid-19 pandemic. Additionally, learning teams rotate a responsibility to facilitate a discussion about each challenge: Each team is assigned one challenge to present. The presenting team must give a short background on the challenge dataset, formulate a "central claim" about the data, and provide evidence supporting their claim. These conversations serve as an opportunity for students to practice productive communication skills; including dialogue (active and constructive responding) and visual communication. I have found these discussions to be extremely motivating for students; past teams have contacted the original authors for datasets to learn more, and often discover new and interesting insights into the datasets. I use these discussions to get students to practice talking about data analyses and asking good questions, and use each discussion as a segue into a mini-lecture to clarify course content in a specific context.

The challenges in the materials are an opportunity for summative assessment. However, in Data Science at Olin College I use challenges for additional formative feedback through a revision system. Feedback on student work is provided through a GitHub Issue, and students have an opportunity to edit and re-submit their work through their repository (similar to an academic peer review process). Anecdotally, I have found that this approach incentivizes students to read and engage with instructor feedback.

I have run this course with additional open-ended projects to encourage students to apply their skills to new areas. In previous offerings, I have assigned a project at the midpoint of the semester that builds off of c06-covid: Students are responsible for identifying an additional dataset of covid-19 data and studying another aspect of the pandemic. In the past students have investigated data disaggregated by demographic features to study racial gaps in disease outcomes, and economic data to see how the pandemic related to increases in demand for various products (e.g. meat, toilet paper).

Adapting These Materials

These materials can be used as-presented for a semester-long (14 week) Data Science course for engineering students. However, the modular nature of the content lends itself to remixing and adapting materials for other uses. The challenges are designed to highlight engineering topics; one could replace a few of the challenges to highlight topics from other disciplines. Alternatively, one could subset the materials to teach a shorter course, say on data and visualization basics by extracting the `data` and `vis` exercises.

The materials themselves can also be freely edited and adapted, as all materials are released under a CC-SA License. To edit the materials themselves, make changes to the `*-master.Rmd` files in the `exercises/` directory. Note that these source files include `begin|end` comment annotations to denote materials for either the `task|solution` files. Assignment and solution files can be generated using the `Makefile` in the `exercises/` directory.

The following is an example of `R` comment tags.

```
## Code for both assignment and solution document

# task-begin
## Code for the assignment document
# task-end

# solution-begin
## Code for the solution document
# solution-end
```

The following is an example of markdown comment tags.

```
This text for inclusion in both the assignment and solution documents.

<!-- task-begin -->
- This text for inclusion in the assignment only.
<!-- task-end -->
<!-- solution-begin -->
- This text for inclusion in the solution only.
<!-- solution-end -->
```

An important adaptation that all instructors **should** make when using these materials is to update the setup instructions. Most importantly: `e-rep00-setup` covers instructions about course-specific tools (I use GitHub and Discord in place of a traditional learning management system), `e-setup00-install` lists the R packages necessary for the course, and `c00-diamonds` takes students through the process of submitting an assignment. Instructors should tailor these instructions to the particular workflow for their course, e.g. point to their institution's LMS (such as Canvas).

Tools are provided in the scripts directory to work with the materials. Most importantly, the notebook `make-schedule.Rmd` can be used to re-sequence the exercises; this is accomplished by changing the `day` column in the `df_schedule` tibble. Running this notebook produces a `schedule.csv`, which is then used by the `prepend.py` script in the root directory to produce the sequenced exercises. The build steps are automated with `Makefile`(s) in the various directories: To re-sequence the curriculum, one need only:

1. Modify the `df_schedule` tibble in `make-schedule.Rmd`.
2. Run the `make-schedule.Rmd` notebook to modify the `schedule.csv` data.
3. Run the root-level `Makefile` to update the sequenced exercise curriculum.

Inspiration and Dependencies

Both the content and structure of this curriculum are inspired by the (discontinued) Data Challenge Lab course at Stanford University (Stanford Data Lab, n.d.). The exercises make heavy use of the `Tidyverse` (Wickham et al., 2019) "metapackage," as well as other R packages (Bryan, 2017, 2020; Delignette-Muller & Dutang, 2015; Garnier, 2018; Genz et al., 2020; Ooms, 2019; Robinson et al., 2020; Slowikowski, 2020; Wickham, 2019, 2020).

References

Adhikari, A., DeNero, J., & Wagner, D. (2021). *Computational and inferential thinking: The foundations of data science.* https://inferentialthinking.com/chapters/intro.html

Aggarwal, R., Flynn, M., Daitzman, S., Lam, D., & del Rosario, Z. R. (2021). A qualitative study of engineering students' reasoning about statistical variability. *2021 Fall ASEE Middle Atlantic Section Meeting.*

Bishop, J. L., Verleger, M. A., & others. (2013). The flipped classroom: A survey of the research. *ASEE National Conference Proceedings, Atlanta, GA*, *30*, 1–18. https://doi.org/10.18260/1-2--22585

Bryan, J. (2017). *Gapminder: Data from gapminder.* https://CRAN.R-project.org/package=gapminder

Bryan, J. (2020). *googlesheets4: Access google sheets using the sheets API V4.* https://CRAN.R-project.org/package=googlesheets4

Delignette-Muller, M. L., & Dutang, C. (2015). fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, *64*(4), 1–34. https://doi.org/10.18637/jss.v064.i04

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, *111*(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111

Garnier, S. (2018). *Viridis: Default color maps from 'matplotlib'.* https://CRAN.R-project.org/package=viridis

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2020). *mvtnorm: Multivariate normal and t distributions.* https://CRAN.R-project.org/package=mvtnorm

Grolemund, G., & Wickham, H. (2018). *R for data science*. https://r4ds.had.co.nz/

Lang, J. M. (2016). *Small teaching: Everyday lessons from the science of learning*. John Wiley & Sons.

Moore, D. S. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review*, *65*(2), 123–137. https://doi.org/10.2307/1403333

Ooms, J. (2019). *Curl: A modern and flexible web client for r*. https://CRAN.R-project.org/package=curl

Robinson, D., Hayes, A., & Couch, S. (2020). *Broom: Convert statistical objects into tidy tibbles*. https://CRAN.R-project.org/package=broom

Slowikowski, K. (2020). *Ggrepel: Automatically position non-overlapping text labels with gplot2'*. https://CRAN.R-project.org/package=ggrepel

Stanford Data Lab. (n.d.). *Data challenge lab open content*. https://dcl-docs.stanford.edu/home/.

Starmer, J. (n.d.). *StatQuest*. https://www.youtube.com/user/joshstarmer.

Timbers, T.-A., Campbell, T., & Lee, M. (2021). *Data science: A first introduction*. https://ubc-dsci.github.io/introduction-to-datascience/

Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2). Reading, Mass.

Wickham, H. (2019). *nycflights13: Flights that departed NYC in 2013*. https://CRAN.R-project.org/package=nycflights13

Wickham, H. (2020). *Modelr: Modelling functions that work with the pipe*. https://CRAN.R-project.org/package=modelr

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223–248. https://doi.org/10.2307/1403699