Data Science in Action:
real-world final projects in an introductory course

Katherine Shoemaker, Assistant Professor of Statistics | University of Houston – Downtown

DETERMINED. DEDICATED. DOWNTOWN.

University and Program

The University of Houston – Downtown is a community serving four-year university that is both an HSI and MSI, located in the center of Houston, the fourth largest city in America. Many of our students are non-traditional, first generation, often working full time jobs, and there is a large variation in the amount of college preparation available to them before they start.

The BS in Data Science is the first of its kind in Texas and is designed to give students the tools and competencies that they need to immediately join the workforce as data literate employees, analysts, and data scientists; with this degree, the goal is for them to be able to combine skill sets from statistics, computer science, and technical communication.

Scan for more info about the Data Science Bachelors Degree at the University of Houston - Downtown

Prerequisites and Learning Objectives

Students often have little to no background in programming or statistics when they begin Data Science courses.

Data Science 2401 is the first course in a three-course series of DS specific classes. This course is designed to introduce the student to the "entire data science pipeline". Topics include:

- data collection and transformation;
- exploratory data analysis;
- introduction to statistical software;
- data visualization and presentation

The ***only*** prerequisite for the course is a precalculus course that has a trigonometry component.

Goals

The students need to learn the basics of communicating the results of their analyses in tandem with learning the tools that enable them to create the analysis and the report. We want to get them started early with good practices in reproducible research and version control!

Project Structure

After choosing their own data set, students are asked to put together a presentation and report that explores and summarizes their data. The report is done in RMarkdown, they are required to use version control via GitHub, and they get extra credit for using ioslides or slidy for their presentation instead of making a powerpoint.

Students immerse themselves in the entire reporting ecosystem alongside learning the DS toolset.

Students begin using R, GitHub, and RMarkdown at the beginning of the course. The project is introduced after the basics of R programming, but before the course begins to cover the tidyverse (primarily dplyr and ggplot2), about midway through the semester. The presentation days are the last 2 or 3 days of class, depending on class size, and the report is due during the finals time.

Tools

- R, RStudio, and the tidyverse for analysis
- RMarkdown for presentation
- Github and RProjects for version control
- In the last two semesters, added a brief introduction to APIs, plotly and Shiny, but the inclusion is optional.
- Various open source data sources are discussed as possible sources for data, but students can also provide their own data.

Significant Takeaways

- Students surprise themselves with what they're capable of producing, and they take pride in the work. Especially in cases where the student is using data they are personally invested in.
- Students seek out new tools on their own to accomplish something above and beyond the scope of the project.
- Peer feedback is often very insightful and useful. Occasionally, students will reflect on their own projects when giving feedback to others, saying they've learned something and want to try to implement it themselves.
- These projects are great for their resumes! The repository is public on GitHub, and they can showcase the project directly to potential employers.

Further Expansion and Changes

We work in an ever-changing field! A benefit of this type of activity is that it can be updated easily as the tools change and grow. I am currently considering moving to the new Quarto reports and adding on GitHub Pages.

Other faculty were inspired to create a similar project in the later DS courses. One of our group goals is to connect the pieces, with the students having the option to build on their prior work as they move through the program.

Cautions

- Students can be very overwhelmed by the limitless choices for topics. It's good to have a few topic suggestions on hand.
- Students often want to take on too much of a project! While it is great to see them be excited about data science, it is helpful to remind them of the scope of the project (while also hinting and forecasting that they'll be learning the ML tools to do those larger things later)

Conclusions

This project is designed to give introductory students a clear, immediate application of the new skills they are learning. Consulted and guided by the instructor, students produce professional grade reports in the same semester they are introduced to R and the data pipeline.

Materials

Scan for project description document, rubric, and details. Projects are available in the public Data2401 GitHub Organization.