STA 393 Data Science Education: From Tutorials to Assessment

Spring 2022 Reflection

Evan Dragich

Supervised by Dr. Mine Çetinkaya-Rundel, PhD.

The following document serves as the final artifact representing my work in STA393 this Spring 2022 semester. I met with Dr. Çetinkaya-Rundel weekly on Fridays 9-10am this semester, along with three other IS students.

Data Science Tutorials

While the {dsbox} package was not the primary focus of my work this semester, I still oriented myself with the project and contributed two main additions.

First, I resolved a pending issue calling attention to the ambiguous levels of a particular variable. I set out to determine which numeric level corresponded to which qualitative description, and found that the levels were indeed conflated as the tutorial stood. In the process, I also updated the URL for this data's source to a more permanent link, as I was unsuccessful using the given link as-is, removing some of the HTML portion of the URL.

Then, I added the missing documentation for the Denny's and LaQuinta datasets, resolving another pending issue. In the process, I learned much about {roxygen2} and the relationship between .Rd files, .R files, and the documentation that appears in the RStudio pane when you call `?`.

Finally, the culmination of my semester's work with {dsbox} was generating a tutorial using the datasets, which is the tenth in the package so far. Modeled after the Data Science In a Box labs associated with these datasets, the tutorial serves as a capstone to the previous tutorials, integrating previously-seen topics such as descriptive statistics, data wrangling, particularly `mutate()` and `*_join()` functions, as well as spatial mapping. Through this, I gained a significant confidence using {learnr} to set exercise code chunks and provide hints, as well as {gradethis} to automatically give feedback on inputted answers. I found it very enjoyable to consider the structure of the tutorial, as I sought to balance having one overarching narrative with a variety of topics and questions to get at many facets of the data science curriculum.

While this pull request has yet to be merged, all checks are passed. Since I began work on {dsbox}, I have also updated the `DESCRIPTION` file to include all currently-used packages and render using a new version of {roxygen2}. I believe that next steps for this project consist of user-testing, to encounter both small bugs or typos and the high-level flow and user experience navigating the tutorials. All data is up to date and all checks are currently passed, so if the current iteration meets the desired specifications for such a package, I think a CRAN release is definitely in the near future (especially because the documentation is somewhat inherently accounted for in the nature of the tutorials themselves).

Data Science Assessment

The bulk of my work this semester focused on the creation of a Data Science Assessment. When I began, the project consisted of a draft ~50 question assessment, to which I provided initial feedback These were then translated into pull requests, a new tool for me, but not before converting the entire assessment to a Quarto Book. I was completely new to Quarto, and enjoyed the practical motivation to learn another new tool, cleaning up the raw questions in the process. Some highlights of my cleaning work include: delineating items into 26 discrete passages, aligning the item numbers accordingly, ensuring that all content translated well to the HTML Book format, creating a download PDF link on the website and ensuring all content rendered well in PDF form, and adding code-styling to the pseudocode and table passage using HTML.

We were then ready to gather feedback on the polished assessment, and scheduled our first interview with Dr. Nick Horton, of Amherst College. The interview occurred on 4/21/2022 for 2 hours, and Dr. Çetinkaya-Rundel and I walked through the assessment with him, taking notes and eliciting interview responses, respectively. We had some immediate work after the interview, fixing a handful of "obvious" issues illuminated via the interview.

The interview audio recording and transcription file were saved to the cloud, and I downloaded into NVivo to begin coding. Following the stellar guidelines given in Reinhart et al., I believe that simply coding the "tags" that Dr. Horton gave each item as well as some general themes in the introductory and concluding open-ended questions will suffice for our rounds of informal faculty interviews. However, I have gotten good practice loading audio/text into NVivo, and playing with timestamps to draw out themes. I look forward to possibly using this tool to more closely analyze student data.

Currently, the next steps are to contact more professors to repeat the interview process, possibly pruning items beforehand. While these may not occur until the fall semester, there is a chance that work will continue over the summer to be able to gather student data as soon as possible.