Project Title

Fondren Library Data Repository for Data Science Education and Experiential Learning, Phase I

Project Mentor(s)

Su Chen, Assistant Teaching Professor; Anna Xiong, Government Information Coordinator; Catherine Barber, Data Services Specialist

Project Summary

This project will pilot a process for creating a repository of interesting, real-world government datasets that are easy to access, beginner-friendly, and suitable for educational use. Fellows will gain experience with the data science life cycle while creating a resource that will facilitate experiential learning in data science.

Project Description

Background

Data science is an interdisciplinary field that has quickly become one of the most popular and highly demanded in the job market. Interest in data science at Rice is evident in the expansion of the data science minor and the newly-launched Master of Data Science program. Data science education relies on real-world datasets to increase student engagement and to improve learning outcomes.

However, incorporating real data when teaching data science poses major challenges for instructors. Most well-suited datasets tend to be overused and outdated, while recent datasets often require extra cleaning and wrangling that is beyond the scope of an introductory course. Another challenge is that widely available public datasets, such as the US Census and American Community Survey, can be intimidating without sufficient instruction and detailed data documentation. This project proposes to address the need for ready-to-use, open datasets through a collaboration among Fondren Library's Kelley Center, Research Data Services, and the D2K Lab.

Fellows participating in this project will identify appropriate government datasets, process and clean the data to be optimally useful in an educational context, develop sample research questions guided by their own interests, write a brief supporting literature review, and conduct data analysis and data visualization to demonstrate how the datasets can be used. This work will result in a curated collection of open government datasets and data science case studies that will be housed in an open-source repository, such as the Rice Digital Scholarship Archive or GitHub. In addition, Fellows will document the barriers encountered throughout the process, solutions identified, and unanticipated discoveries made. These reflections will shape the next phase of the project.

Impact This project will address a growing need for a curated dataset and data science case study repository. This repository will benefit students at Rice who take the foundational data science introductory course, pursue a data science minor, and/or engage in data science experiential learning. The repository will also potentially benefit other universities, institutions, and individuals who have similar needs in data science education. Feasibility Dr. Chen's previous experience with teaching D2K classes has demonstrated that students can accomplish most of the data life-cycle tasks within one semester. The additional semester provides time to complete the remaining project tasks (e.g., data storage and sharing, testing, and publishing). Mentors will provide or arrange for the necessary training, and we will use open-source software that is readily available. In addition, our selection criteria will ensure that Fellows are sufficiently prepared to begin working on key tasks with appropriate support.

We expect to hire three Fellows. Each Fellow will be paired with a primary mentor who will serve as the main support person. Each Fellow will select a government database and develop research questions, literature review, and data analysis to suit their interests. The project Fellows, faculty, and staff will work together as a team to create, pilot-test, and evaluate the repository. We expect Fellows to work 5-10 hours per week during the semester.

Key Tasks for Fellow(s)

The following is an approximate schedule of key tasks, subject to change:

Fall semester:
Obtain training on data management and digital repositories.
Identify the general area of interest within government databases (e.g., education, health, census) and draft research questions.
Conduct a brief literature review relating to the research questions.
Identify a publicly available government dataset that is appropriate for beginners.
Create a brief introduction of the data source and type of variables contained.
Create an easy-to-access version of the dataset if not already available; for example, convert specific data format to standard csv files.
Fall deliverables: brief literature review, introduction to data source and variables, dataset in standard format.

Spring semester:
Design a data science pipeline to conduct necessary data wrangling and cleaning in order to analyze data using standard software such as R or Python. This pipeline should be reproducible and reusable, easy to incorporate with new data, and well-documented.
Develop a case study demonstration to perform meaningful analysis and illustrate how to tell stories with the dataset.
Select and set up the repository.
Share datasets and case study demonstrations in the repository.
As a team, produce a written summary of the project that incorporates reflection.
Spring deliverables: case study, repository, written team summary.

Ongoing:
Team meetings and reflection

Qualifications

The ideal candidate will have taken at least one course in data science, demonstrate passion for working with real data, and possess sufficient technical skills such as Python or R programming for basic data analysis and visualization.

Previous experience working with government data sets for research is preferred but not essential.

Previous data management experience is preferred, but strong interest in learning about data management is acceptable.

Learning Outcomes

The Fondren Fellows will gain firsthand research experience working with real data that applies methods and techniques to derive insights from data. They will hone their programming skills and obtain in-depth instruction on working with governmental data. They will also learn best practices in data management and sharing.

In addition to acquiring these skills, we anticipate that Fellows will benefit from the opportunity to work with a team that includes faculty, library staff, and peers. The team aspect of this project will promote effective interpersonal communication, self-reflection, and collaboration, all of which will positively impact their future academic performance and career development.