Final Project

Data Science for Public Policy

Alena Stern and Aaron R. Williams - Georgetown University

Deliverables:

- An individual project proposal (2 points)
- A team project proposal (2 points)
- A project check-in (4 points)
- A final project (19 points)
  - A link to a GitHub repository
  - A link to GitHub pages included in the GitHub repository README
  - A *brief* presentation

Final Project Requirements

The objective of this project is to demonstrate an analysis relevant to governance, policy, or social sciences using the tools learned in this course. Students will clearly articulate a problem, find novel data related to the problem, and use one or more of the following sets of tools to meaningfully inform the policy debate around that issue.

- accessing data through web APIs
- data visualization
- geospatial analysis
- supervised machine learning
- unsupervised machine learning
- text analysis
- data imputation

You will submit a GitHub Page that contains your analysis and addresses the following sections:

1. Background and Literature Review: Describes your research question, why the question is important to governance, policy, or social sciences, and a brief review of the literature that is relevant to your research question. Note that it will likely not be feasible to conduct a thorough literature review for this project, but I expect you to show that you have at least engaged with the existing research in your focus area.
2. Data Sources: This section should describe each of your data sources that you are using for your analysis and include your code to access and read in the data. If the data access includes any manual steps that are not captured in the code (e.g. downloading files in a point-and-click manner) please document those steps here.
3. Data Wrangling and Exploratory Data Analysis: This section should include your code to perform any data cleaning and new variable creation. You should also thoroughly explore your data, including assessing the presence of outliers/unexpected values and identifying and appropriately addressing missingness in any key variables.
4. Data Analysis: This section should include the code to conduct analysis to answer your question of interest. It should include writing explaining why the tools you selected are a good fit for the research question and a justification of key analytic decisions. If you are using machine learning methods, this section should include model evaluation using the methods discussed in class.
5. Discussion of Results: This section should include the interpretation of your results. Please discuss what your results suggest about the answer to your research question of interest. Please also discuss any limitations of your analysis and areas for potential future research.

Your GitHub page will be created from Quarto document (index.qmd) that you render to HTML (index.html). All of your analysis should be conducted in your Quarto document, or R scripts that you source in your Quarto document. You will push both the index.qmd and index.html file to GitHub (along with any R scripts used if applicable) to create the GitHub page for submission as outlined in the 05_reproducible-research-with-git note set.

Deliverables

1. **Individual proposal:** Every student must submit a brief proposal on Canvas. The proposal should only be a few sentences and include a brief description of the motivating question, all data sources that will be used, and a description of the approach. Project proposals will be posted in a shared Google doc so students with shared interests can develop teams. **Late submissions will immediately receive a grade of zero.**
2. **Team proposal:** Based on step one, students will form teams of 3-4 people. All teams need to submit a project proposal on Canvas. Only one proposal is needed per team. The proposal should be 2-3 paragraphs and include the names on the team, a brief description of the motivating question, all data sources that will be used, a description of the approach, and the two or three of the biggest anticipated technical hurdles in the project. Please cite any projects that you are emulating or building upon from other

classes or other researchers. **Late submissions will immediately receive a grade of zero.**

3. **A public GitHub repository:** All project materials should be created using .R, .md, and .qmd. Microsoft Word and Google docs will not be accepted. All files should be added to a GitHub repository during the project. A detailed README and clear commit messages are required. Projects that are created and then hastily added to Git/GitHub will be penalized.
4. **GitHub pages with all project materials:** The final "paper" and supporting documentation will be posted on GitHub pages. Your project README should include a link to all pages. There is no length requirement. The volume and quality of materials depends on the number of people on a team, the ambition of the parts of the project (data collection, methods, etc.), and method of discussion. For example, a project with many effective data visualizations may need fewer paragraphs. A project with many example simulations may need fewer paragraphs. The project proposal and mid-project check-in are opportunities to get feedback about the thoroughness of a project. The "final" paper should demonstrate exploring topics beyond the material in the course through citations and implementation. Be sure to include a bibliography.
5. **A brief presentation** All teams will be required to briefly share their projects.
6. **Team evaluation:** Teams will be given an opportunity to evaluate the contributions of team members.

Rubric

- The ambition of the project exceeds examples from class notes or exercises in assignments.
- The project includes data wrangling and exploratory data analysis using the methods covered in class.
- The project demonstrates at least three of the following methods covered in the course: data visualization, geospatial analysis, supervised machine learning, unsupervised machine learning, text analysis, and text modeling. If you prefer, you can also go deeper on fewer methods and we will expect that the project include content beyond what we covered in class (e.g. if you want to focus on supervised machine learning alone, you could implement new algorithms, explore new evaluation metrics, analyze model bias, etc.).
- The execution of methods incorporates practices demonstrated during the course (e.g. data visualizations use active titles and other best practices outlined in assignment 04 and predictive models are developed with resampling methods).
- Analytic decisions are justified.
- Code is clearly documented.
- Git and GitHub are appropriately used. We expect to see commits from all team members throughout the project - not just on the date of submission!
- The GitHub pages clearly communicate the project.

- All data sources and supporting materials (excluding class notes) are clearly cited.

Note: Learning and data science are both collaborative practices. We encourage you to discuss project topics and with each other. However, the work you submit must be your own. Copying-and-pasting text or ideas without citation will be considered a violation. Copying-and-pasting excessive code will be considered a violation. Please attend office hours or contact one of the instructors if you need help or clarification. Plagiarism on projects will be dealt with to the full extent allowed by Georgetown policy (http://honorcouncil.georgetown.edu).

Final note

A project doesn't need to work to be successful. For example, a team could effectively try many approaches to creating a predictive model for a meaningful application and never achieve an error rate low enough to justify implementation. Be ambitious. This project is an opportunity to push yourself. Hopefully every student will leave with a skill/idea to inform their dissertation or a project that can be explained/shared on the job market.