

30538 Final Project: Reproducible Research

Peter Ganong and Maggie Shi

2024-10-14

Project Description and Instructions

The goal of this project is to showcase your knowledge of Python by applying it to a research project about a **policy topic** you are interested in. You will be graded on **coding, writeup, and an in-class presentation**.

You may work on this project alone, or in groups of up to **three students**. **All groups must be formed declared in the Canvas proposal** before any work is done - it is not possible to join one after.

It is required that you use **GitHub**, and we may use your **past commits to understand your thought process for partial credit**. If you working in a group, note that as **we are grading we will be looking for multiple commits per individual throughout the project**. The **division of labor should be approximately evenly across both individuals**. While we will lean toward giving the same grade for all group members, it is possible that individuals may receive different grades based on the commit history.

If you choose to form a group, we recommend that you do so with other students in your section. You are allowed to have a group member from another section, **but all group members must be available to attend all the group members' lecture sessions**. If you are not present when your project is presented, you will not receive credit for the presentation.

Grading

Coding (70%)

The code for the project should have the following components:

1. **Data wrangling (25%)**
 - You must use a **minimum of two datasets**.

- All processing of the data should be handled by your code, including all merging and reshaping.

2. Plotting (25%)

- From that data, you will create a minimum of *two* static plots using `altair` or `geopandas`
- As well as one shiny app with one dynamic plot
 - You can also add additional dynamic plots into your app to substitute for a static plot. So, a shiny app with 3 dynamic plots will count for full credit.

3. Reproducibility (10%)

- The project and files should be structured and documented so that the TAs can clone your repository and reproduce your results (see “Final Repository” below) by knitting your `.qmd` and, if needed, downloading the dataset(s) you use using the link provided in the `.qmd` comments

4. Git (10%)

- You should submit your project as a Git repository.
- Create multiple branches as you work for different pieces of the analysis. Branches may correspond to work done by different partners or to different features if you are working alone.
- Your final repository should have one branch: `main`
- We reserve the right to check the git commit history to ensure that all members have contributed to the project.

5. Extra credit: text processing (up to 10%)

- Introduce some form of text analysis using natural language processing methods discussed in class.

Writeup (15%)

- You will then spend *no more than 3 pages* writing up your project.
- The primary purpose of this writeup is to inform us of what we are reading before we look at your code.
- You should describe your research question, then discuss the approach you took and the coding involved, including discussing any weaknesses or difficulties encountered.
- Display your static plots, and briefly describe them and your Shiny app. Discuss the policy implications of your findings.
- Finish with a discussion of directions for future work.
- The top of your writeup should include the names of all group members and Github user names.

Presentation (15%)

- On the day of the presentation, one of the group members will be *randomly selected* to give a *8-minute in-class presentation*. All group members must present.
- The presentation will be of slides that largely mirror the structure of the writeup, but will be more focused on discussing the research question and results as opposed to explaining the details of the coding.

Final Repository

Your final repository must contain the following:

- **Documentation and Meta-data**
 - A `requirements.txt` file
 - A `.gitignore` file that ignores unneeded files (e.g. `venv`)
- **Writeup**: a user should be able to knit your `.qmd` file and re-generate the HTML version of your writeup
 - The `.qmd` file associated with your write-up
 - An HTML and PDF'd version of your writeup
 - A folder named `pictures` that contains the files for any pictures required to knit your writeup
- **Data**
 - A folder named `data` that contains the **initial, unmodified dataframes you download and the final versions of the dataframe(s) you built**.
 - If the dataset is greater than 100MB, it can be hosted **on Drive or Dropbox and the link should be provided in your `.qmd` file as a comment**
- **Shiny app**
 - A folder named `shiny-app` that **contains the code and any additional files needed to deploy your app**
 - A user should be able to deploy your app directly from the command line within this folder

Key Dates

- By November 1
 - Proposal submitted to Canvas quiz

- (Optional) meeting with Professor Ganong, Professor Shi, or Head TA Ozzy Houck
- Sign up for presentation slot

- December 2- December 5: in-class presentations
- December 7, 5PM: final repository submitted via Gradescope