# ECNS 460/560 Term Project

Nick Hagerty

Fall 2023

Throughout the rest of the semester, you will pursue an independent project to analyze data in an area of your own interest. This is your chance to show off all the skills you've learned in this course – I expect you to challenge yourself! It will be different from a typical research project in economics – we won't focus quite as much on causal inference, but more on descriptive analysis, as well as on the process you follow to get there. Your code should be written so that someone else could easily pick up where you left off. The final outputs will consist of a GitHub code repository, a written report, and a short class presentation.

**Teams.** I strongly encourage you to work in teams of 2 students. The project may also be done alone, but expectations and grading standards will be the same. Teams of more than 2 students are not generally permitted. If you really want to work in a team of 3, you will need to meet with me to ask for special permission, and your team will be held to a higher standard. If you would like me to assign you to a partner, I am happy to play matchmaker – just email me.

## Stage 1: Topic and Data

*Due Thursday, October 12 by class time, on GitHub.*

**Agree on a topic** for your project with your partner. It can be anything relevant to economics, public policy, or business, but you should both find it interesting!

- You cannot choose the exact same topic as another team. If you hear that another team is working on a similar topic, talk with them and negotiate which team will work on which aspect of the topic (or just choose another topic).
- You also cannot use the exact same topic or datasets that any team members are using (or have used) for another class.

Another issue is that your data needs to feature cross-sectional relationships. It should not be primarily time series data, like macroeconomic indicators or prices of individual stocks. Analysis of time series data is important but not the focus of this class, so doing a good job tends to require tools you haven't learned here.

***(ECNS 560 only:)*** **Think of a few potential research questions.** Write down a few (2-4) well-defined research questions that you might be interested in answering in this project. A research question is a complete sentence that can be answered with the right data and assumptions. It is narrower and more precisely defined than a topic. Here are a few examples of good research questions:

1. (Descriptive) Has wealth inequality increased in the U.S. faster than in Europe since 2000?
2. (Causal) Does receiving Medicaid coverage reduce the risk of bankruptcy?
3. (Predictive) Can nighttime satellite imagery be used as a real-time indicator of economic activity?

**Find data** from at least two different sources that are related to your topic. Download them and make sure you're going to able to use them for analysis. The datasets should relate to each other in some way that will allow you to join (merge) them together, but you don't have to do that yet.

- All data you use should be publicly available on the internet (no confidential data from your summer internship).

- Data should come from the original (primary) source. They should not be pre-cleaned – a major objective of this project is for you to go through the process of downloading (or scraping!), importing, wrangling, and cleaning the data. You may not use datasets we've used in class, datasets that get automatically loaded in R packages, nor datasets posted on Kaggle or similar repositories.
- I may consider exceptions to these rules on a case-by-case basis. The bottom line is that your project will need to show skills you've learned in this course. If all the data you use is already nice and clean, then you need to find another way to make the project more challenging for yourself.

**Create a public repository on GitHub** for your project. *(You can always make your repo private after the semester if you prefer, but past projects on your GitHub account will look great to future employers.)* Add a document that describes your topic, motivate why it matters, and lists your research questions. Then, describe your datasets. Include the source URL, what kind of variables are contained in each dataset (no need for a full list), the timespan and spatial coverage of the datasets, and how the datasets are related to each other.

**Submit** the URL of your GitHub repo to D2L, so I know where to find it.

## Stage 2: Draft Exploratory Analysis

*Due Thursday, November 2 by class time, on GitHub.*

**Wrangle and clean:** Import your datasets into R, wrangle them into tidy format, and join them into a single data frame. Clean your data, keeping in mind our Data Cleaning Checklist. Your code should be heavily commented and follow the best practices that we discuss in class.

**Explore:** Conduct an exploratory analysis of your data. A good starting point is to focus on one outcome variable (or a small number of outcome variables) and investigate its unconditional distribution, then how it varies across geography, over time (or by month/week/day/hour), and/or according to other key categorical or quantitative variables in your data. (You may then want to show the distribution of and variation in those other key variables.) Apply transformations where appropriate, pay attention to missing data, and handle extreme values in a defensible way.

**Visualize and communicate:** Produce a set of descriptive findings about your data that could be interesting or useful to a policymaker or business executive, along with a set of compelling visualizations to show what's going on in your data.

- There should be a roughly one-for-one relationship between the findings and the visualizations: Each finding must be backed up by evidence you show, and each visualization should directly relate to the findings you discuss. Almost all of your visualizations should be built around a single, clear, focused point that you're trying to communicate.
- You might aim for around 4-5 findings and visualizations. I can't set a firm number because different visualizations can contain wildly varying amounts of information. Most of your evidence should be graphical; you can include one table of numbers if you must, but no more.
- Your visualizations should look professional, contain titles, labels, and captions, and follow the principles of good data visualization we discuss in class. There are limitless numbers of cool things you can do with `ggplot2` beyond what we cover in class. See if you can impress me!
- Be precise with language: Avoid implying causal relationships between variables at this stage, since you're merely describing your data. You don't want your audience to make high-stakes decisions based on unwarranted interpretations!

**Write up a report** in R Markdown, containing:

- Your topic, motivation, and the description of your data that you already submitted (revised as necessary to reflect any changes).
- A couple of paragraphs explaining how you processed your data and why you made the choices you did. Identify the unit of analysis in your cleaned, merged dataset.
- A couple of paragraphs explaining any transformations you applied, how you dealt with any missing data or extreme values, and why.

- A few pages that describe your findings and show the visualizations.

Focus on quality, not quantity. Endless pages of graphs or text are not impressive. Choose a limited number of findings, make the visualizations effective, and describe them thoroughly but succinctly. Your report should look neat and professional. Unlike the homework assignments, it should **not** contain R code or results. Use headings to organize the document, write clearly and straightforwardly but not informally, and check your spelling and grammar.

**Collaborate on GitHub:** If working with a partner, use GitHub as your platform for sharing files throughout the assignment. If working alone, still use it to store and back up your project as you work. **Each team member must make at least one substantive commit** to the repo. (The objective of this expectation is for each team member has some practice with GitHub. I will not examine the record of commits for other purposes, since the number of commits will not necessarily reflect the amount of effort and contributions of each team member.)

**Organize** your project repo, following the best practices we discussed in class, such as:

- Put raw data, code, cleaned data, and output in separate folders. Give all files and folders informative names.
- Thoroughly document your code with comments. Make it as easy as possible for a future collaborator or replicator to understand how to run all of your code and create the outputs.
- Do not use absolute filepaths in your scripts. Allow each user (including yourself) to define the location of the project folder on their own computer, by setting the working directory in the R console.
- Considering ordering your scripts with numbered filenames, or creating a "master" script that runs everything else with one click.

**Commit and push** a complete exploratory analysis to your GitHub repository. Make sure your repo includes:

- The R scripts you used to process your data and produce the visualizations. (These should be in standalone R scripts, not R Markdown documents. Do not include the code for everything you tried or looked at, only what's necessary to produce the results.)
- The final cleaned dataset as an object in a `.Rdata` file.
- The knitted report in **HTML or PDF format,** as well as the raw file(s) used to create it. (If you choose to submit it in HTML, make sure your repo includes all image or CSS files it depends on.)

## Stage 3: Final Exploratory Analysis

*Due Thursday, November 30 by class time, on GitHub.*

After you submit your draft exploratory analysis, I will read it, look at your repo, and give some feedback. You will then have another 2-3 weeks to revise and improve it before I assign final grades. All guidelines for the draft also apply to the final version.

## Stage 4: Econometric Analysis (ECNS 560 only)

*Also due Thursday, November 30 by class time, on GitHub.*

Once you have completed your exploratory analysis, you now understand your data well enough that you're ready for an econometric analysis. However, I do not expect this analysis to be a complete research project. Do not focus on finding a good identification strategy, or trying to convincingly answer a causal question (you will not be able to do so in the time allowed!). Instead, finalize a research question (or a small number of closely related questions) that your data can speak to. Then, conduct one of the following types of analysis:

1. **Descriptive regression analysis, as an end in itself.** If your research question(s) of interest is a descriptive question, use regression to help you quantify (and obtain standard errors for) some of the relationships that you found in the exploratory analysis.
    a. Write down 1-3 regression models that you could use to shed light on one or more of your research questions using your available data you have. No need to get fancy – a simple OLS or fixed effects

regression is fine. I won't prohibit you from trying something else you've learned in econometrics, but start with something simple. Explain all notation in your regressions and why they are appropriate to your situation.

   b. Run your regressions and display the estimate(s) of interest in a table. (A figure is optional.)
   c. Interpret your results. Put the coefficient(s) into complete sentences of English to explain what the regression told you. Explain what they mean. What new facts do we learn about the world from the relationships (or lack of relationships) you find? Since your research question is only descriptive, be *very* careful not to claim or even imply any causal interpretation of your results.

2. **Descriptive regression analysis, as a first step toward causal inference.** If your research question is a causal question, do steps a-c above, plus a couple more things:
   d. Interpret your results through the lens of causal inference. How plausible would it be to interpret your results causally? Describe the limitations to doing so. What are some likely sources of bias, in what direction are they likely to push, and how concerned should we be about them? It's often helpful to describe what the ideal experiment would be, and then compare your situation to that experiment.
   e. Say a few words about how you might proceed from here, if you were going to more convincingly answer your causal research question.
3. **Predictive analysis.** If your research question is a predictive question, try using machine learning tools to conduct a predictive analysis.
   a. Choose one of the variables in your cleaned dataset to be the target of prediction. You should be able to explain why a policymaker, business executive, or another member of the general public might find it useful to be able to predict this outcome variable.
   b. Use other variables in your cleaned dataset to predict the outcome variable as best you can. Follow all the correct ML practices: Set aside a test set; train and predict using one or more ML algorithms; tune your model(s) using appropriate cross-validation, and finally test your performance with the held-out data. (Don't cheat!!! Really, don't touch the test set until you are done tweaking your model. If you change your model after seeing how it does in the test set, then you are reporting dishonest performance numbers.)
   c. Explain your methods: What learning method did you use? What parameters did you tune? What method did you use for tuning?
   d. Interpret your results: How did you measure performance/success? How did your model(s) perform? What do you think limited the performance? What did you learn in the process? It may be effective to create visualizations to accompany your interpretation.
4. **Prediction in service of causal inference.** If you want to go all out, you can try answering a causal research question using double/debiased machine learning (which I won't cover until November 14, but I can provide you materials in advance if you ask). The idea is that you can use predictive models to reduce omitted variables bias in a regression that you want to interpret causally. If you choose to pursue this, you will need to follow all applicable steps from each of the other types of analysis.

Then:

**Add a section to your report** describing your econometric analysis and the results. Use headers like Motivation, Methods, Results. Include all information required above for the type of analysis you chose. Update the first section of your report so that it describes the research question you landed upon and why it's important or interesting. Be sure to explain why you chose the type of analysis you did: What makes your research question either a descriptive, causal, or predictive question, and not either of the other two?

**Commit and push** your project to GitHub. Continue to follow best practices for organizing your project. Ensure your final repo includes all scripts, inputs, and outputs involved in the econometric analysis, as well as the updated report.

## Stage 5: Presentation

*Slides due Thursday, December 14 by 8 am. (Class meets 8:00-9:50 on Dec. 13.) Upload link will be provided.*

During finals week, you will give a short presentation on your project. Oral communication is an absolutely

crucial part of being a data scientist or economist. Presentations are the main way knowledge is transmitted in both corporate environments and academic research (surprising but true!). However, the presentation is a small part of the project grade – I'm hoping it will be a fun chance to show your classmates what you've been working on and in turn to see what they've been doing.

Your team's presentation should be 8 to 10 minutes if you are in ECNS 560, and 5 to 7 minutes if you are in ECNS 460. Prepare a short set of slides briefly describing your topic and motivation; research question (if applicable); data; any unusual methods you used for acquiring, processing, or analyzing data; and results. Your slides should be sparse and easy to read – don't try to include all the information from your report!

## Stage 6: Team Evaluation

*Due Thursday, December 14 by 8 am, on D2L.*

Unless you worked individually, submit a short evaluation of whether you and your team member(s) distributed work equally or if the work was unequal. If it was unequal, describe the contributions of each team member and whether you believe you should receive a different grade than the other team member(s).

## AI Disclosure

If you choose to use ChatGPT, GitHub Copilot, or any other chatbot or AI technology to help you complete your project, you must write an explanation to the end of your project disclosing that fact, naming the tool you used, and explaining in depth how you used AI and for what parts of the project. This is for the same reason we cite other kinds of sources – I will consider it a violation of academic integrity if you turn in any AI output and fail to disclose it. It is also because I am curious to learn the ways you are finding AI tools to be useful!

## Grading

Project grades will be based on the following categories:

- Code and data (60 points).
  - Style, coding practices, file organization.
  - Data importing and wrangling.
  - Data cleaning.
- Exploratory analysis (60 points).
  - Data exploration.
  - Visualizations.
  - Findings and communication.
- Econometric analysis (ECNS 560 only; 40 points).
  - Technical execution.
  - Findings and communication.
- Submitting two versions (10 points).
  - Draft shows a good-faith effort.
  - Revisions incorporate feedback.
- Presentation (10 points).
  - Slides.
  - Delivery.