POLI381 FINAL PROJECT: DATA-DRIVEN RED TEAMING

This version: April 13, 2021

A major objective of Poli381 has been to develop your ability to be part of data analytical conversations: about what data to use, how to analyze them, and what conclusions to draw. The final project for this class asks you to put those skills to use to answer a substantive question in political science.

We will be engaged in what I call "red teaming." (I think the term comes from the CIA.) In a nutshell, a red team is a group that works independently from a main analysis team to provide a check on procedures and conclusions. For instance, if CIA analysts concluded that a foreign adversary had a secret weapons program, CIA leaders might form a red team to look at all the intelligence material again, and see if they reach the same conclusions. The procedure can help confirm that the analysis was done properly, and guard against motivated reasoning (e.g. an analyst who wants to reach a particular conclusion, because it will help his or her career).

For the final project, there will not be a "main" analysis team. Rather, each of several groups will be given a substantive question, and a dataset that might be used to answer it. You are to write a research paper that answers the question and explains how you answered it. But you are expressly prohibited from discussing your approach or conclusions with other groups during the project. At the end of the semester, all groups will share their procedures and conclusions, and we will see how much of a consensus was reached.

It bears emphasis that it's best not think of this as a *competitive* assignment. The point of having multiple groups answer the same question is not to see if one group can do "better" than the others. The assignments will be graded in absolute—not relative—terms. Instead, compartmentalizing the labor, as this exercise does, helps to identify different approaches to answering questions—and then we can have a conversation about what the pros and cons of each procedure might be. Working independently also helps guard against blind spots. But as with a CIA red team, we should all ultimately think of ourselves as being on the same "side."

The Final Project is due in three parts, on the dates noted on the syllabus:

- **First draft of Red Team Report**. This is a document that lays out your procedures and conclusions as concerns one of the substantive questions below. The document should follow all the structural guidelines—laid out below—for the final draft, except it does not need to have the "Response to Comments" section. Note that "first draft" is not a euphemism for "incomplete" or "sloppy." You should think of this as a formal, polished document. "First draft" merely signifies that, because the inferences have not been subjected to external scrutiny, they should be considered more tentative than in a final draft.

- **Comment on other teams' reports**. For this part of the final project, your group should read the reports of the other teams. (I'll tell you which ones to read.) Then, you should write a 1,200 word comment on them. (One 1,200-word document covering all the other reports.) This document will be shared with the report authors, for them to consider in revising their report. You can comment on any aspect of the other reports. However, the best comments 1) flag analytical choices that might lead a group to erroneous conclusions, explaining why, and 2) articulate any additional analytical steps to take (e.g. conducting an additional test to assess whether a problem exists, or proposing a correction to the underlying procedure).

- **Final draft of Red Team Report**. This document lays out your final analysis and conclusions. We will discuss expectations at greater length in class, but it should be organized as follows. (Word counts are approximate guidelines, not ironclad rules.)

- The report should start with an introduction, which should briefly discuss the substantive question you are answering, why it is important, and provide an overview of your procedures and conclusions. (500 words)

- Next should be a section that lays out, in full detail, the analyses you conducted and what you found. You should include all the information that an outsider would need to reproduce you results (even without having access to your R code). This section will be the longest. I expect it to include several tables and figures, and it very well might be organized into a series of subsections. (E.g. "Coding of the dependent variable," "Justification for regression model," "Considering alternative models," and "Conditionality in the main effect." The exact subsections will vary by group.) (3,000 words)

- For the final draft only, a response to the comments you received from the other red teams working on the same question as you. Here you should discuss what steps you took in response to the critiques you received, or explain why you disagree with one or more critiques. (800 words)

- A conclusion section that summarizes your steps, inferences, and discusses limitations. You should include a discussion of what additional information would be the highest priority to collect, and how you would incorporate this information into your analysis if it became available. (600 words)

- Separately, each team member must upload analytical files that produce all of the quantitative results in the paper. You can discuss the steps to be taken and discuss how to accomplish them in R. But you cannot copy-and-paste each other's code.

Citations in the paper should be parenthetical format and should look like this (Ryan 2019). There should be a works cited section at the end of the document that provides details on all the references. There is no minimum number of references. (You might need some, but given the nature of this class, our focus is more on analytical procedures more than deep engagement with a literature.)

Here is the substantive question we'll focus on for this project.

As you know if you have been awake in the past year, there has been a pandemic. In the United States, it became highly politicized. For instance, President Trump spoke derisively of masking. Public support for masking appeared to fall along political lines, too. Did this politicization have ramifications for county-level pandemic severity? In particular, did being more conservative cause counties to fare worse against the pandemic? If so, how much?

I am giving you some datasets and a very brief script that will get you started on answering this question. In particular, you will receive

- A dataset that shows cumulative case counts for many US counties. Case counts are pretty reliable and seem like a good way to operationalize "fare worse." In other words, you'll get your DV from this. However, I am not sure the right way to use this information. (Look at a particular date? Change over a range of dates? Counts versus proportions? I'm not sure.)

  This dataset comes from USA Facts.

- A dataset that shows county-level vote totals for presidential candidates in the United States, for several elections. Maybe this is useful for determining how conservative particular counties are; not sure.

This dataset comes from the MIT Election Lab.

- A dataset showing county populations, from the Census Bureau. I don't know if this will help or not, but you have it in case.

My script gives you **partial** guidance on organizing and merging these datasets. But you'll have more to do to get set up to do analysis.

There really is no single "right way" to do this project. Part of its purpose is to help you see how open-ended data analysis can be, and to have you go through the process of converting a big, broad question like the one above to something more concrete and tractable. Also, to induce you to have conversations with other groups about the tradeoffs among different approaches and what we might (or might not!) learn from each of them. I am not holding the "right" answer behind my back. This said, I do want you to practice and showcase particular analytical skills you have been developing. Here is some more specific guidance:

- I would like you to discuss what control variables might help you develop more credible estimates. What kinds of things might serve as a confounder for the effect of county conservatism on pandemic harm? When you have some ideas, I think you should see if you can find data related to one of them, merge the data into your dataset, and account for the confounder in your analysis in some way.

  One promising place to turn is the Census website, which has lots of information about county-level stuff—usually including FIPS codes, which you'll see are extremely helpful for merging.

  Experience tells me that I need to caution you against going overboard here. As you'll see, merging in new variables can be tricky. I do not expect you to develop highly elaborate models where you're accounting for each and every possible confounder that crosses your mind. One or two will be sufficient to meet expectations and earn a good grade.

- I'd like you to consider at least one *contingent* relationship: some circumstance where the effect of X on Y might depend on something else. (Think: interaction terms.) In your paper, you should describe one interactive relationship you wish to assess, and then tell us what you learned about it.

- You should employ your skills to make at least two visually appealing and informative data visualizations.

- You should give careful thought to the *functional form* of any relationships you estimate. What are your reasons for coding variables as you did? (Categorical variables, outliers, nonlinear relationships, and more are all relevant here.)

- You should discuss uncertainty surrounding your key estimates using the concepts acquired during Part 2 of class.

I'll close with some practical guidance. I recommend you begin by discussing with your group a meeting schedule, as well as a timeline for reaching various benchmarks related to this project. It is not the sort of thing that can come together in a day. And you should make wise use of our several in-class workshopping sessions. I especially recommend you get an early start on the "control variables" step above. Finding and merging datasets together can be more difficult than it sounds, and we might need to discuss how to do it. So get an early start!