Econ 5/Poli 5D: Final Project Description

Due before 11:00 AM on Friday, March 19th

Purpose

This project is designed to improve your data analytics skills by putting into practice the tools you've acquired during our class. You will explore data – either from suggested datasets or ones of your choosing (and approved by your instructor) – and present interesting findings both descriptively and visually. This project is intended to present challenges such as aggregating data, handling missing values, discovering relationships between variables, etc. that you might encounter working as a data analyst or research assistant. Outstanding projects can be included in your portfolio when you interview for jobs or apply to graduate school.

Requirements

While it's not expected that your project comprise a publishable peer-reviewed journal article, the best projects will present evidence that supports answers to specific questions about phenomena observed in society whose answers are not obvious. Here are some examples of good project ideas:

1. How is a candidate's ideological polarization related to her campaign funding sources and amounts?
2. How have the number of gun-related crimes changed over time by state and following passage of gun-related legislation?
3. What happens to the number of traffic fatalities following major sporting events, e.g. the Super Bowl?
4. What is the relationship between Trump's tweets about the Middle East and oil prices? About China and aluminum/steel prices? About X and Y?
5. How does weather (temperature, rain, wind) impact finishing times at the World Marathon Majors (or results at other sporting events)?
6. Is there any association between hospital density (hospitals per capita) and life expectancy? Medical cost?
7. How sensitive is 2018 congressional campaign spending in 2018 to how close the vote was in the 2016 presidential election?
8. What is the relationship between cost of tuition and earnings at one's first job after graduation? 10 years later? How does this vary by major, region, type of university, gender, race, etc.?
9. How has state-wide legalization of marijuana (for medicinal and/or recreational use) affected fatalities from marijuana? From alcohol?
10. How are the wage and employment gaps (between female and male workers) in a state/country related to educational and health outcomes of children in that state/country?

Here are some examples of not-so-good project ideas:

1. What are the impacts of terrorism? (Too vague – needs to be more specific.)
2. How is education related to wages? (It's well established that education increases earnings.)
3. What is the optimal prices at which to buy and sell Bitcoin? (Likely too ambitious.)

You must pick a dataset or multiple datasets from which your analysis can provide answers to your question of interest. Several example datasets have been provided, or you may choose your own and have them approved by your instructor.

After composing your question and choosing relevant datasets, you will analyze the data to thoroughly answer your question. Your analysis will consist of both descriptive and visual analysis:

- *Descriptive analysis* includes anything that describes the central tendency (i.e. mean, median, etc.) or spread (i.e. standard deviation, range, etc.) of the data.
- *Visual analysis* includes anything that illustrates data, generally in the form of plots.
- It is expected that you use methods taught during the quarter including using statistical functions, estimating linear models, and plotting.

Both forms of analysis help readers understand how the data are distributed, relationships between variables, differences across groups, and in general support the answer that you present. Your analysis should be written like a research paper or technical report for a scholarly or professional audience.

Grading

Your final project makes up 40% of your final grade in this course. The homeworks will include some questions related to your projects, but homework grades are not included in the 40%.

Your final project should include a thorough investigation of your question of interest. Evidence in support of your answer should include both descriptive evidence (e.g. differences in statistics/models across groups) and visual evidence (e.g. various plots). There is no minimum or maximum length – the project should be long enough to sufficiently defend your arguments. All code used to produce your complete project should be submitted along with your writeup. You will almost certainly use some commands not covered in class; learning how to use new commands (with help menus and StackExchange, for example) is a learning objective for this course.

Projects will be graded on the basis of completeness and degree to which it answers the question of interest. Outstanding, high quality work will receive As, good but not exceptional projects Bs, acceptable but insufficiently defended work Cs, and incomplete or insubstantial work Ds or Fs. Late assignments will lose one letter grade per day late.

Below is a guideline for what is considered outstanding, high quality work.

- **Question**: the writer's question is specific, not obvious, and answerable within the timeframe of one quarter. She explains why her question is interesting and relevant to modern audiences.
- **Argument**: the writer's argument clearly and specifically answers the question of interest. She presents thorough evidence in favor of her argument.
- **Evidence**: the writer presents both descriptive and visual evidence that defends her argument. All of the evidence presented is concise and relevant to the question of interest. None of the presented evidence could have conveyed the same information in less space (that is, one plot instead of multiple plots, one column instead of multiple columns in a table, etc.). The writer presents a variety of evidence and does not draw upon the same kind of plot or descriptive statistics for all of her evidence.
- **Figures**: Figures clearly present evidence relevant to the research question. All figures have titles, labels, captions, and legends when appropriate. When multiple data series are included, different colors or patterns are used to distinguish each series. Formatting and axis scaling are chosen to make the figure easily understandable.
- **Discussion**: the writer explains how the presented evidence supports her argument. None of the plots or tables exist without references in the text of the paper that explain what the figures depict and why the reader should care. Limitations of the evidence, including potential sources of bias, are discussed.
- **Conclusion**: the writer leaves the reader with a clear takeaway given the findings presented. The writer explains how her argument should affect the decisions made by policymakers, business leaders, or other persons of influence.

Deliverables

You will submit two items:

1. **Report**: This contains the writeup of your results. Please submit your report in PDF format.
2. **Code**: All code used to generate the results in your report must be submitted at the same time your submit your report. Your code should include comments so that it is easy for the grader to follow your methodology, and it should run without errors. *All plots and statistics should be replicable by running the code you submit.*
    a. It is okay to use multiple scripts. For example, you may want to use a script to clean your raw data and another to generate summary statistics and plots. This is often a good strategy if one part of your workflow takes a long time to run.
    b. Please add comments at the top of the scripts that describe the code (i.e. the author, last date updated, what the code does). If you use multiple scripts, you should label the order that the scripts should be run.

You do not need to upload the data you use. However, if you use a dataset outside of the provided datasets, the instructor may ask you to share your data. Additionally, **please refrain from cleaning the raw data in a way that is not replicable** (i.e. do not open the data in Excel and make changes).

Example Structure

This serves as a rough guideline - you need not follow the structure outlined below exactly.

1. **Title page**: title of the report, author, institution, class, quarter/year, one paragraph abstract/summary of the report. This is generally on a page by itself.

2. **Introduction**: tell the reader what the question is and why anyone should care. If you reference other authors' work (including data you didn't collect yourself), you should give credit with a citation and list in the references section at the end of your report. This is also a good place to list some hypotheses that you will test and to give a preview of the key findings.

3. **Data**: tell the reader about your data. Where did they come from, who collected them, what are they representative of, any potential biases present. This would be a good place to include a summary table of your variables.

4. **Methods and Results**: describe the process by which you analyzed the data and what you found. Talk about the models you fit, the plots you built, and why you included them. Please do not talk about the code you used (no code in your report, please). All of your analysis should go in this section, but please add structure so that it flows naturally (as opposed to simply listing all plots and tables back to back without descriptions). Remember, even the best data analysis in the world is pointless if one cannot communicate it effectively!

5. **Discussion**: discuss the results you just found. Did the data confirm your hypothesis or present evidence that was counter to your hypothesis? What do the results suggest? Anything particularly surprising? Why might the relationships you found exist? It's okay to tell a story to give the reader some intuition for why the phenomenon exists. Any limitations to your methods? Why might the results be different than what other researchers found?

6. **Conclusion**: Repeat the key findings and why the reader should care about these findings. Leave the reader with some thoughts about future potential work that should be done to garner a better understanding of the phenomena you are exploring.

7. **References**: List all references here. Simply copy and pasting URLs is unacceptable. See the references page in any respectable peer-reviewed journal for examples of what's acceptable.

8. **Appendix**: Probably unnecessary for most students. This is where you put figures/tables that don't make the cut to show up in the main body of your report. Before you add plots to this section, see if there's a way to convey the same information more concisely and stick it in the body of your report.