Final Project

David Gerard

2019-09-26

Important Dates

- 10/09/2019, 5:30 PM:
  - You have emailed me the names of your group members.
  - You have listed your top three data sets.
  - For each dataset you list, you provide me with 3 initial hypotheses you could explore.
- 11/13/2019, 5:30 PM: The progress report is due. This should be submitted on Blackboard along with any code you've written so far.
- 12/04/2019: In class presentations on your project. Presentation materials should be submitted on Blackboard. You can use the feedback I provide in class to help you with your final project report.
- 12/11/2019, 5:30 PM: Final project report due, submitted on Blackboard. Also submit a zipped folder containing all of your analyses, scripts, and inputs. I should be able to run everything after unzipping the project folder.

Grade Breakdown

- (5 pts) Group formation, data requests, and hypotheses.
- (10 pts) Progress report.
- (15 pts) In-class presentation.
- (20 pts) Final project report.

Description

For the final project, you will explore a real dataset from a real study, present your results, and write up a report of your findings. You will work in **teams of 2 to 3 people** to complete this assignment.

The entire report may be no more than 10 pages (double spaced, 12 point font, 1 inch margins). The references do not count toward your page limits. The project report should be written in R Markdown and knitted to PDF.

You may only discuss your projects with your own group members (and me, of course). You may not discuss your projects with the members of other groups or people outside of the class.

The due date is **Wednesday, December 11th, 2019, at 5:30 pm**. Please submit your report on Blackboard by that time.

Your report should be **well written**. Your goal should be to tell a story with the data ("we thought we would see A, but we actually observed B, and we think we saw this because of C."). Please spell and grammar check your report.

Your report should be **reproducible** and **well-organized**. You will turn in a zipped folder containing all of your analyses, scripts, and inputs. I should be able to run everything after unzipping the project folder.

Your project report should include all of the following elements:

1. **Title Page**: Include the names of all team members.
2. **Introduction**: Provide as much information about what you know about the data as you can. Are the measures quantitative or categorical? What are the research questions? You should include a *brief* literature review. For the lit review, you should find peer-review articles related to the topic. Use a publication search engine such as Google Scholar. Include as many articles as individuals in your study team.
3. **Initial Hypotheses**: Before you look at the data, provide a list of *detailed* hypotheses. E.g. you might believe that a certain variable is associated with poverty — or that this association is only mediated through a third variable.
4. **Exploratory Data Analysis**: Describe the features of your data using numerical and graphical summaries. Do not just copy results from R. Select important graphs and numerical summaries.
5. **Data-driven Hypotheses**: What new hypotheses did you develop as you explored the data? Provide some description of things that you found interesting when you were looking at the data.
6. **Discussion**: A brief discussion of how your results may be placed in the context of the literature.
7. **References**: A list of references cited in your report. Use a standard format for references (such as APA or MLA)
8. **Appendix**: Additional R code not included in the main text.

Presentation

- Your group should also prepare a 10 minute presentation. The content of the presentation should be the same as the project report (background, hypotheses, data explorations, discussion).
- Each group member should talk for a proportionate amount of time during the presentation.
- No R code should be shown during the presentation (unless it is really cool R code). R warnings and R messages should not be present in the slides.

Progress Report

You will have to provide a progress report mid-way through the rest of the semester.

- It should be 2 pages long.
- Figures do not contribute to the page limits.
- Your report should be written in R Markdown and knitted to PDF. The names of all group members should be at the top of the page.
- You should provide the following details:

    1. What peer-reviewed articles are you including in your report? Summarize each in 1-2 sentences.
    2. Describe the work you've done so far.
        1. Provide evidence that you have downloaded the data you will work on. **You should submit any code you've written in a separate file.**
        2. For each submitted hypothesis, provide some evidence that you are making initial explorations. This could, for example, be in the form of a plot or a table of summary statistics.
    3. Write a paragraph on your next steps.

Data Sources:

Below are some possible sources for data. Each dataset may only be used by one group.

I will accept other sources as long as they are comparably large and complicated. For example, there are many open data portals from state agencies (https://opendata.dc.gov/).

I will not accept small curated datasets, or those that come from a paper where the data has already been analyzed.

- Federal bridge inspections from the Federal Highway Admin:
  https://www.fhwa.dot.gov/bridge/nbi/ascii.cfm
- Drinking water violations from the EPA:
  https://www.epa.gov/ground-water-and-drinking-water/drinking-water-data-and-reports
- Contracts or grants from USA Spending:
  https://www.usaspending.gov/#/
- DC Crime incident data from the DC police:
  https://mpdc.dc.gov/page/statistics-and-data
- SBA disaster loans from the Small Business Administration:
  https://www.sba.gov/offices/headquarters/oda/resources/1407821
- Fatal accidents from the NHTSA:
  https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars
- Baltimore crime data from the Baltimore police:
  https://www.baltimorepolice.org/crime-stats/open-data
- Aircraft animal strikes from the FAA:
  https://wildlife.faa.gov/databaseSearch.aspx
- Bank health data from the FDIC (also see banktracker at IRW):
  https://www.fdic.gov/bank/statistical/guide/data.html
- WMATA elevators and escalators from the WMATA:
  http://data.codefordc.org/dataset/wmata-escalators-elevators-hotcars
- USGS water quality data from the USGS:
  https://water.usgs.gov/owq/data.html
- Prison data:
  http://www.dc.state.fl.us/pub/obis_request.html
- College Scorecard:
  https://catalog.data.gov/dataset/college-scorecard
- Reporting Carrier On-Time Performance (1987-present) from the Bureau of Transportation Statistics:
  https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236
- The General Social Survey from the non-partisan and objective research organization at the University of Chicago (NORC):
  https://gssdataexplorer.norc.org/
- FBI National Incident-Based Reporting System (NIBRS)
  https://crime-data-explorer.fr.cloud.gov/downloads-and-docs