Project Overview

PPOL670 – Introduction to Data Science

Spring 2020

Contents

Overview

The following provides an overview of the data science project that you will be responsible for completing by the end of the semester. The project is an opportunity to apply the skills and tools that you've learned throughout the course on an area of substantive interest to you.

The project is composed of three distinct parts: a proposal, a presentation, and a report. The proposal should outline the general plan for the project and will serve as an opportunity for the professor and teaching assistant to provide guidance on its feasibility. The presentation is an opportunity to present your work in mid-stream to receive verbal feedback from the professor, TA, and classmates. These comments will hopefully help you as you move forward with the final report. The report is the written analysis of the project in its entirety. The report will be due on May 7 @ 6pm (PPOL670's designated finals slot).

While completing your project, you will be responsible for creating and maintaining a public repository on `Github`, tracking all progress made on your project using the version control implementations discussed at the start of the semester. All work product should be reflected on the repository. The proposal, presentation materials, and the report should be generated using RMarkdown and should follow all reproducibility practices discussed in class.

Project Proposal

Due	Proportion of Grade	Length
March 25	5%	2 pages (single-spaced; 12pt font)

The project proposal asks that you sketch out a general 2 page (single-spaced; 12pt font) project proposal. The proposal should offer the following information:

1. A high-level statement of the problem you intend to address or the analysis you aim to generate;
2. The data source(s) you intend to use;
    - The data you choose to utilize should take some effort to compose (it shouldn't be entirely "off-the-shelf data", i.e. data easily downloaded online).
    - If using API's or other easily downloadable data sources (such as World Bank data) to collect your data, try joining other data to the downloaded data, change the unit of analysis, etc.
    - In general, one should have some wrangling component to get their data in order.
3. Your plan to obtain that data;
4. The methods (learned in class) that you aim to employ; and
    - The project must contain:
        - a data wrangling component
        - (multiple) data visualizations
        - a machine (statistical) learning component
5. A definition for what "success" means with respect to your project.
    - In your words, what would a successful project look like? How will you know that you solved the problem or accomplished your goal?
    - Four weeks isn't a long time to complete a project like this. Thinking serious about what a "finished" or "successful" project might look like. This will help you set realistic goals/expectations.

Please be detailed but *succinct* as possible when writing. *Any material that exceeds page 2 will not be considered when grading/reading* (I mean it!). There is no advantage/incentive to exceeding the page limit. Be sure to properly cite any referenced materials and/or packages (it is okay if your work cited runs onto a third page).

Project Presentation

Due	Proportion of Grade	Length
April 22	10%	7 minutes in length

Please prepare a slide presentation *using `R Markdown`*, which walks us through the progress you've made on your project. The presentation will offer you the opportunity to summarize your project and talk through your (preliminary) results. Moreover, it'll (a) provide an opportunity for the Professor/TA to provide constructive feedback, which you can incorporate into your final paper, and (b) motivate you to make significant progress on the project before the final deadline.

When preparing your talk, use the following format. You should plan on having with 5-10 slides in total. The layout of the presentation should take on the following form.

1. (1-3 slides) Problem statement and Background
2. (1-3 slides) Methods you explored or considered using.
3. (1-3 slides) The methods/tools you used, and the rationale for their use.
4. (2-4 slides) Results (however preliminary).
   - Show main visuals, analyses/tables, and/or any products built (interactive graphics, websites, etc.)
5. (1-2 slides) Lessons learned thus far and/or plans to mitigate challenges.

Keep in mind that 7 minutes passes quickly. The time will be *strictly* enforced (we need to make sure we get through everyone's presentation in time), so be sure to practice/polish your presentation prior to giving it. In addition, note that presentations will start promptly at the beginning of class. If you're presenting first, then you must be ready to go at the start of class. Failure to do so will only eat into your time.

Note again that the presentation should be rendered using a `.rmd` file and the slides should be rendered as either a `.html`, `.pdf`, or power point file. Students must submit both their slides *and* the `.rmd` file used to render the slides to CANVAS.

Order of the speakers will be randomly assigned and circulated on CANVAS a week prior to the in-class presentations.

Project Report

Due	Proportion of Grade	Length
May 7	30%	12 pages (double-spaced; 12pt font)

The report is a complete description of the project's analysis and results. The report should be 12 pages in length (double-spaced; 12 pt font) and cover the below bullet points. As with the proposals, no written material will be considered beyond page 12 when reviewing the report (a work cited can span onto a 13th page). Below I've outlined points that one should aim to discuss in each section. Note that paper should read as a cohesive report, so do not respond to these bullet points verbatim.

- **Introduction**

    - What is the aim of the project?
        * Summarize the problem
        * State your goals
    - What do you do in this report?
        * offer a roadmap of the project

- **Problem Statement and Background**
    - Give a clear and complete statement of the problem and/or aim of your analysis.
    - Include a brief summary of any related work that has tried to tackle a project similar to yours (i.e. a *light* literature review)

- **Data**
    - Where does the data come from?
    - What is the unit of observation?
    - What are the variables of interest?
    - What steps did you take to wrangle the data?

- **Analysis**
    - Describe the methods/tools you explored in your project.
    - Outline in detail our entire analysis.
        * Justify the tools/methods that you used.
        * Assume the reader is smart but doesn't know `R`/Machine Learning well. That is, be crystal clear about what you're doing and why.

- **Results**
    - Give a detailed summary of your results.
    - Present your results clearly and concisely.
    - Please use visualizations and tables whenever possible.

- **Discussion**
    - Speak on the "success" of your project (as defined in your proposal).
        * Did you achieve what you set out to do? If not why?
    - What tools/methods did you consider but *not* use in the final analysis?
    - How would you expand the analysis if given more time?

The reports must be submitted as a hardcopy (i.e. the `.rmd` notebook must be rendered as a `.pdf`) to CANVAS by 6PM on May 7th. ***The code for all the tables and visuals in the rendered document must be included in your Github repository***. The professor or TA should be able to clone the repository and run your code without issue. ***Note that given the page constraints, no `R` code should be visible in the rendered document.***