Project Overview

PPOL670 – Introduction to Data Science

Spring 2019

Contents

Overview

The following is an outline of the data science project that you and your assigned team will be responsible for completing over the course of the semester. The project aims to offer students an opportunity to apply the skills and tools that they've learned over the course of the semester in an applied setting.

The project will play out in four distinct parts comprising two memos, one report, and one group presentation. The memos outline the general plan of the project and will offer an opportunity for the professor and teaching assistant to provide feedback. The report and presentation constitute the final products of the project's analysis. Both will be due on the final day of class.

Each team will be responsible for creating a public repository on `Github` and tracking all contributions to that repository using Git/Github. Memos, reports, and presentations should be generated using (RMarkdown)[https://rmarkdown.rstudio.com/authoring_quick_tour.html] files. This will facilitate a text-based data approach, allowing for version control as your project develops. Students are encouraged to use commit messages to track changes and contributions, rather than creating seperate files with different versions.

Note that **each part of the project is due at the start of class on the assigned day**.

Team Assignment

Teams of 2 - 3 students will be randomly assigned. Students will receive a listing of those in their team on **January 28** (after the add/drop period has concluded). Team assignments are posted at the end of this document. It is the responsibility of each team to locate viable meeting times, find a mutual topic of interest, and coordinate the workload.[1] **Each team should select a name and will be responbile for building a `Github` repository to house the project's work product.**

Repository

Each team must create a `Github` repository to house their project. All members should be able to pull/push to the repository. The repository should be named after your team. Please have one team member send the repository url, the team name, and a list of all members in an email to the professor (with the TA cc'd) prior to the start of class on **February 4**.

Memos

Memo 1: Topic and Data

Due	Proportion of Grade	Length
February 4	10%	1 - 2 pages (singlespaced; 12pt font)

The first project memo asks that your team generate a 1 - 2 page (singlespaced; 12pt font) proposal statement that outlines the general direction of the project. This proposal should offer the following:

1. A high-level statement of the problem your team intends to address;
2. The data source(s) your team intends to use;
3. How your team plans to obtain that data; and
4. Any potential issues in either data extraction, quality, or availability that your team foresees (if any).

Please be detailed but *succinct* as possible when writing. Any material that exceeds page 2 will not be considered when grading. Note that there is no advantage/incentive to exceeding

[1] Note that any disagreements or grievances should be resolved internally.

the page limit — in fact, doing so can backfire as the memo will appear incomplete. Be sure to properly cite all referenced materials and packages (it is okay if your work cited runs onto a third page.)

Memo 2: Analysis Plan

Due	Proportion of Grade	Length
March 11	10%	1 - 2 pages (singlespaced; 12pt font)

The second project memo asks that your team outline the goals and products of your analysis. The memo should be 1 - 2 pages (singlespaced; 12pt font).

1. Succinctly restate the problem;
2. What are the testable hypotheses;
   - What are your expectations regarding the empirical relationships in your analysis (e.g. "We think that `x` will result in an *increase/decrease* in `y`.")?
   - What are the success metrics that you'll use to determine if there is an effect/result?
   - In general, how will you draw an answer with respect to the question from the data and analysis your team will employ?
3. Provide a description of the data analysis tools your team plan to use;
   - What tools are you aiming to use to clean, visualize, and process the data?
   - What, if any, machine learning or statistical techniques are you planning to use?
4. Outline what products you plan to build: visualizations, analyses/tables, interactive applications, ect.
   - Note that these will be the tables and graphics in your report. If your team decides to produce any additional material, such as a Shiny application or a website, please note them here.

Please be detailed but *succinct* as possible when writing. Any material that exceeds page 2 will not be considered when grading. Note that there is no advantage/incentive to exceeding the page limit — in fact, doing so can backfire as the memo will appear incomplete. Be sure to properly cite all referenced materials and packages (it is okay if your work cited runs onto a third page.)

Project Showcase

Project Report

Due	Proportion of Grade	Length
April 29	20%	6 - 10 pages (doublespaced; 12pt font)

The report is a formal and completed description of the project. The report should be 6-10 pages in length (doublespaced; 12 pt font) and cover the below bullet points. As with the memos, no written material will be considered beyond page 10 when reviewing the report. Note that the structure mirrors items that were touched on in the memos.

- **Problem Statement and Background**
  - Give a clear and complete statement of the problem. Don't describe methods or tools yet. Where does the data come from, what are its characteristics?
  - Include background material as appropriate:
    - who cares about this problem,
    - what impact it has,
    - what implications better solutions might have.
  - Included a brief summary of any related work you know about.
- **Methods**
  - Describe the methods your team explored. Justify your team's methods in terms of the problem statement. What methods did you consider but *not* use?
  - Include every method you tried, even if it did not "work". When describing methods that didn't work, make clear how they failed and any evaluation metrics you used to decide so.
- **Tools**
  - Describe the tools that your team used and why. Justify the tools used in terms of the problem itself and the methods your team was aiming to utilize.
    - Tools can include anything from packages used for data wrangling and visualization to machine learning and statistical processing.
  - How did you employ the tools used? What features worked well and what did not?
  - Describe any tools that you tried and ended up not using. What was the problem? Briefly, what could be improved in these packages to make them more functional?
- **Results**
  - Give a detailed summary of the results of your work. Here is where you specify the exact performance measures you used. Usually there will be some kind of accuracy or quality measure. There may also be a performance (runtime or throughput) measure.
  - Please use visualizations and tables whenever possible. Include links to interactive visualizations or websites if you built them.

The reports must be submitted as a hardcopy (i.e. the `.rmd` notebook must be rendered as a `.pdf` or `.docx`, *printed*, and stabled as one document) at the start of class on **April 29**. The code for all the tables and visuals in the rendered document must be included in the `.rmd` (e.g. the notebook that the report is generated on should read in the necessary data and render the graphics and tables accordingly). Note that given the page constraints, all `R` code should *not* be visible in the rendered document. Specifically, for each code chunk, `echo=FALSE` or globally with `knitr::opts_chunk$set(echo=FALSE)`. The team will submit the final project with all accompanying data necessary for rendering the `.rmd`, the `.rmd` itself (with all prose and code incorporated), and a rendered `.pdf`/`.docx`/`.html`. The document should render without any adjustment of the script (i.e. there should be no specific path references or any other material/files that are not included).

Project Presentation

Due	**Proportion of Grade**	**Length**
April 29	10%	Approx. 12 minutes in length

Please prepare a slide presentation using `R Markdown`, which summarizes your teams's project and outcomes so far. The presentation will offer your team the opportunity to summarize the project and their results.

Use the following format. You should end up with 6-15 slides in total, not including the title slide. Please identify all members of your team on a title slide. The layout of the presentation should mirror the report. Keep in mind that 12 minutes passes quickly.

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 and/or plans to mitigate challenges.

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 that used to render the slides to CANVAS.

One or more students in the group can be responsible for presenting. All group members must stand in front of the room when presenting. If time allows, we'll leave some time for questions from the audience. All group members are responsible for participating in Q & A.

Peer Evaluations

Due	**Proportion of Grade**	**Length**
April 29	5%	5 minutes (end of class)

Students will be given a brief survey regarding the contributions of the members in their group. Students will be asked to evaluate each member on a 1 (barely contributed) to 5 (contributed greatly) scale, and to provide comments about each members' contributions.

These evaluations will be taken in concert with the empirical record contained within the team's git logs. **Note that freeriding is observable!** The professor reserves the right to penalize students who fail to pull their weight by reducing their total showcase grade. (That is, freeriders will only get partial credit on the group's final project).

Groups

Below are the group assignments for the Spring 2019 semester. All pairings were randomly assigned.[2] Given the number of people in the course, all groups are composed of 3 members, except 2 groups of 2 members. Please reach out to the professor or TA on Slack or email if there are any questions/concerns about the assignments.

Spring 2019 Project Group Assignments

All groups were randomly assigned.

	Student Email
Group No. 1	
Harrison, Connor A.	cah296@georgetown.edu
Sun, Haorui	hs914@georgetown.edu
Lee, Dong Hoon	dl988@georgetown.edu
Group No. 2	
Taylor, Jamie	jt1183@georgetown.edu
Wang, Yihan	yw524@georgetown.edu
Alarcon, Natalia G.	nga13@georgetown.edu
Group No. 3	
Liu, Qianying	ql94@georgetown.edu
Sang, Terry	ts1171@georgetown.edu
Xiang, Yuchen	yx150@georgetown.edu
Group No. 4	

[2] Replication code for the random assignment is available for those interested.

Zhang, Fangwen	fz97@georgetown.edu
Monticello, Benjamin A.	bam131@georgetown.edu
Durrani, Humera K.	hkd15@georgetown.edu
Group No. 5	
Hernandez Heimpel, Hector H.	hh744@georgetown.edu
Meng, Tingjie	tm1305@georgetown.edu
Chen, Liumin	lc1077@georgetown.edu
Group No. 6	
Mangelsdorf, Steffi	sm3227@georgetown.edu
Contreras Gomez, Rafael E.	rec2148@georgetown.edu
Wang, Diya	dw810@georgetown.edu
Group No. 7	
Stringer, Ann M.	ams670@georgetown.edu
Denk, Erich	ed719@georgetown.edu
Deng, Ivy	xd66@georgetown.edu
Group No. 8	
Lourme, Nick L.	nl471@georgetown.edu
Xie, Ruilian	rx21@georgetown.edu
Group No. 9	
Deng, Xinran	xd73@georgetown.edu
Pan, Wen	wp252@georgetown.edu