Exam 3 (Make-Up) - Data Science for the Social World

Michael G. Findley* Michael Denly†

Instructions

This is a two-hour, open-book, open-internet exam. The timer starts 3 minutes after we email you the exam. If you submit the exam late, you will be penalized by 1 point for each minute late that you submit. For example, if you submit 5 minutes late, you will lose 5 points on your overall exam grade.

Please email Theodore Charm, Mike Denly, and Professor Mike Findley once you are done. Your submission should contain: a PDF file or Word document corresponding to the output of your `R` Markdown file. Kindly note that we will not accept Google Docs, and that you must write your exam using `R` Markdown. Students who use a regular `R` script and then copy/paste outputs onto a Word Document or PDF file will receive a 20-point penalty.

The last question of the exam asks you to submit a link to a GitHub repo with your `.Rmd`, `.dta`, and PDF file/Word Document. Failure to submit a link with all of these files to a working GitHub repo will result in an additional 15-point penalty. You may name the repo anything that you would like, but maybe something like "exam3" would be appropriate. Since GitHub provides time stamps for everything, we will be able to discern if you modify the files outside your two-hour exam window. In short, please respect the two-hour window.

Please work independently. You may *not* consult anyone in the class or outside the class for help, and you may not post the exam questions on the Stack Exchange, Google Groups, or any similar website. However, you may visit these websites or others. Please also do not discuss the questions or answers over WhatsApp, GroupMe, text message, or any other platform, especially because everyone will be taking the exams at different times. We will be monitoring accordingly, and anyone who violates any one of these policies will receive a zero on the exam.

Please annotate your `R` code chunks in your `R` Markdown file with comments, or make sure that the text surrounding it sufficiently explains what you are doing. Essentially, your `R` Markdown file should mimic that notes files that we submit on Canvas to accompany the

*Professor, Department of Government, UT Austin, mikefindley@utexas.edu
†PhD Candidate, Department of Government, UT Austin, mdenly@utexas.edu

video lectures. We will remove points when you do not provide clear comments or explanation to tell us exactly what you are doing with your code.

We have endeavored to make the exam self-explanatory, but feel free to email the instructors and the TA if you have questions. At least one of us will be available over email for the entire exam period. However, please email all three of us if you have a question (i.e., do not email only one or two of us), because we will be taking shifts.

And one final hint: use your time wisely. If you can't answer one question, move on to the next one, and come back to it once you are done with the ones that you can answer more quickly. Good luck!

Questions

1. Clear the environment. [5 points]
2. Use the `WDI` package to download data on female labor force participation for all countries for the years 2010-2015. Save the data frame as `female_lfp`. (Hint: you may will need to Google the indicator.) [5 points]
3. Rename the female labor force participation variable `flfp`. [5 points]
4. Collapse `female_lfp` by the mean value for `flfp` for each country. When you do, keep the ISO-2 country code in your data frame as well as the country name. Name your resulting data frame `collapsed_flfp`. [5 points]
5. Use `R` to show which countries have average female force participation rates for the 2010-2015 period that are less than 15%. [5 points]
6. Use `R` to present a map of the world of using `collapsed_flfp`, using the viridis color scheme. Note: you have already the world border shape files from the training. However, there are a few different ways that you can present the map. [25 points]
7. Based on the map, which area of the world has, perhaps surprisingly, a cluster of yellow-colored average female labor force participation rate states, indicating the highest on the scale? [5 points]
8. Use `R` to show the same cluster of states referenced in the previous question. [5 points]
9. In a Shiny app, what are the three main components and their subcomponents? [5 points]
10. Pull this `.pdf` file from Mike Denly's webpage. It is a report that Mike Denly and Mike Findley prepared for the US Agency for International Development (USAID). [5 points]
11. Convert the text pulled from this `.pdf` file to a data frame, using the `, stringsAsFactors=FALSE` option. Call the data frame `armeniatext` [5 points].
12. Tokenize the data by word and then remove stop words. [5 points]

13. Figure out the top 5 most used word in the report. [5 points]

14. Load the Billboard Hot 100 webpage, which we explored in the course modules. Name the list object: `hot100exam` [5 points]

15. Use `rvest` to obtain identify all of the nodes in the webpage. [5 points]

16. Use Google Chrome developer to identify the necessary tags and pull the data on *Rank*, *Artist*, *Title*, and *Last Week*. HINT 1: In class we showed you how to get the first three of these. You simply need to add the *Last Week* ranking. HINT 2: You can navigate two ways. Hovering to find what you need or by doing `Cmd+F / Ctrl+F` and using actual data to find the location. HINT 3: You're looking to update the code based on the *way the information is in referenced*. Try out some different options and see what shows up in the environment. Keep trying until you see that you have a `chr [1:100]` with values that correspond to what is in the web page. [5 points]

17. Save all of the files (i.e. `.Rmd`, `.dta`, `.pdf`/Word Doc), push them to your GitHub repo, and provide us with the link to that repo. [no points; 15-point penalty for lack of submission (see above)]