STAT 29000 Project 1

Please submit your STAT 29000 Project 1 at this URL: https://classroom.github.com/a/PBr7CuA6 using the instructions found in the GitHub Classroom instructions folder on Blackboard.

Topics: Reminders about Using UNIX on Scholar

Before getting started with this project, read the Motivation / Context / Scope discussion, called "Why are we here?", in the Blackboard site.

ThinLinc

Remember that you can either log onto Scholar via this website: https://desktop.scholar.rcac.purdue.edu

or you can install the ThinLinc Client onto your computer. The advantage of using the ThinLinc Client on your computer is that it is stable and allows copy-and-paste to work with your computer. At the start of the academic year, it is worthwhile to download the newest version of the ThinLinc Client:

Download the ThinLinc Client files:

- For Windows users, the installer is here:

https://www.cendio.com/downloads/clients/tl-4.10.0-client-windows.exe

- For Mac users, the installer is here:

https://www.cendio.com/downloads/clients/tl-4.10.0_6068-client-macos.iso

- For UNIX users, choose the appropriate installer:

https://www.cendio.com/thinlinc/download

Reminder about setting up the ThinLinc Client:

Here are the settings to use inside the ThinLinc Client:

- Server: `desktop.scholar.rcac.purdue.edu`
- Username: your Purdue username without the @purdue.edu – for instance, Dr Ward's username is: `mdw`
- Password: your regular Purdue Career account password. (This is not BoilerKey.)
- Options: Click on "Options" and when the window opens, choose the "Screen" tab, and do the following:
  - choose the "Resize remote session to the local window"
  - do NOT choose Full screen mode, and do NOT enable full screen mode over all monitors
- If you do accidentally get stuck in full screen mode, the F8 key will help you to escape.
- NOTE: The very first time that you log onto Scholar, you will have an option of "use default config" or "one empty panel". PLEASE choose the "use default config".

Motivation/Goal

Last year, we studied some of the data contributed by individual donors in a specific election cycle. To kickoff this academic year, we remind ourselves about some of the basic techniques that we have learned. This is intended to bring back some fond memories!

In this project we will learn about way that contributions to elections have changed over the years.

Scope: The Election Data

Consider the data about contributions by individuals in each of the 21 (two-year) election cycles from 1979 to 2020. The data for those election cycles are given online here:

https://www.fec.gov/data/browse-data/?tab=bulk-data

under the heading "contributions by individuals".

The data is documented online here:

https://www.fec.gov/campaign-finance-data/contributions-individuals-file-description/

Of course, we emphasize that the method of reporting has changed somewhat over the years, so we cannot easily capture the contribution patterns with 1 graph... but we will make a first attempt to understand such patterns, by making a graph.

For your convenience, the data from the 21 election cycles (including a header on each file) is already stored on Scholar here:

`/class/datamine/data/election`

Notice that we expanded the amount of data available about election donations, and we moved the data to a new location. It was previously stored at: `/depot/statclass/data/election2018` and `/depot/statclass/data/election2016` but now we have 21 election cycles of data stored on Scholar.

Question 1: Number of Donations per Election Cycle

1a. Using the terminal, find the *numbers* of donations in *each* of the 21 election cycles.

Note: It is OK if your count includes the header, i.e., if you are off by 1 in each count.

Background reading: Although you can run 21 separate commands to answer question 1a, it is also possible to do this with just 1 simple line of UNIX. Please feel welcome to brush-up on your UNIX skills if needed. Two potential ways to do this are to check out Chapters 1, 3, 4, 5 in this book:

http://proquestcombo.safaribooksonline.com.ezproxy.lib.purdue.edu/0596002610

or Chapters 1, 2, 4, 6 in this book:

http://linuxcommand.org/tlcl.php

(Chapter 20 is worthwhile as a reminder of some ideas too, but not needed for this question)

Question 2: Size of the Data (and Comparison with other Data Sets)

2a. The data for the elections is big, but it is not massive data. How many bytes are stored (altogether) in the 21 files?

2b. As a side note, just for comparison, how many bytes are stored (altogether) in the yellow taxi cab data? This data is located in the directory:

`/class/datamine/data/taxi/yellow`

Question 3: Trends in Campaign Donations

3a. How does the trend in the number of donations look? Take the 21 points of data about the 21 election cycles, and make a plot in R that shows the trend. The y-axis should show the number of donations in each of the 21 election cycles, and the x-axis should show the years (1980, 1982, 1984, ..., 2020). Reminder: the "seq" command in R could easily be used to create the data for the x-axis. It is OK to create a new variable for year (manually). It is not necessary to extract the year out of the file name.

3b. Since the data in recent years is so large, it tends to swamp the data from a few years ago, and it is difficult to understand the trend. Modify the plot in question 3a, so that the y-axis shows the logarithm of the number of donations in each election cycle.