Data Wranging Final Spring 2020

Provide

- a pdf file with a somewhat detailed explanation of what you did and the code that you used and the answers to questions asked (typically the output of the scripts). I would rather the plots from R be included in the primary pdf file, but if you submit those as separate files I'll cope with that.
- a python script
- a database file
- an R script

I'm not expecting to receive a separate file of SQL commands; I expect you to interface with SQL through python and R (by way of the RSQlite package), so that any SQL commands you execute will be included in either the python code or the R script. I am expecting that a single python script could run, followed by a single R script; some of the work could be done in either, but it shouldn't be necessary to go back and forth between python and R.

It is recommended that you read all the way through the instructions before beginning your work; it is recommended that you begin your work well before the deadline. Contra the syllabus, I've extended the deadline to Tuesday morning, but your submission and your quality of life are likely to be better if you plan to finish by some reasonable time Monday evening. The web pages you're being asked to read and parse are less regular than those you've dealt with in the past.

The main exercise

Wikipedia has a webpage at

`https://en.wikipedia.org/wiki/1972_United_States_presidential_election`

that has, among other things, the state-by-state results of the general election; there are similar webpages for each election since then.[1] There is similarly a series of pages starting with `https://en.wikipedia.org/wiki/1970_United_States_Census` with state population data.

Create an SQLite database that contains population and two-party[2] voting totals. Use an SQL command to calculate, for each state, the fraction of the 1980 census population that voted for Ronald Reagan, giving the results in order by that fraction. (For example, Minnesota in 1980 had a population of 4075970, and Reagan received 873241 votes in Minnesota in 1980, so the relevant fraction for Minnesota is .2142.) (Again, report the output in your main pdf file, while also including the command in one of your scripts.)

Connecting to the database in RStudio, use ggplot to produce the following plots:

1. A scatter plot of states, with the Democrats' share of the two-party vote in that state in 1976 on one axis, and the Democrats' share of the two-party vote in that state in 2016 on the other. (E.g. West Virginia gave the Democrats .5807 of the two-party vote in 1976 and .2784 in 2016, so those would be the coordinates of the West Virginia dot.)

[1] There are also such pages for older elections, but the races in the 1960's produced some complications, and there are sufficient complications left for you to handle while starting in 1972.

[2] I.e. you may leave out candidates who aren't the nominees of the Republican or Democratic party.

2. A “cumulative electoral vote” graph

   (a) for the 1976 election

   (b) for the 2000 election

   (c) for the 2016 election

   in which the x-axis is two-party vote share and the y-axis is the total number of electoral votes in jurisdictions in which the party got less than that vote share. (It should look a bit like the graph of a cumulative distribution function.) For example, if you look at the Democrats' share, at a value of $x = 0.4$, the y value will be the total number of electoral votes in all those jurisdictions that gave the Democrats less than 2/5 of the two-party vote share. This will (like a cdf for a discrete distribution) be a line with steps in it; e.g. in the 2016 plot, where Florida had 29 electoral votes and the Democrats got a .4938 share of the two-party vote in Florida, there would be a step at $x = .4938$ with a vertical size of 29 that would correspond to Florida. Make sure there are grid lines such that $x = 0.5$ and $y = 269$ can be discerned on the graph.

Notes

For purposes of this exercise, Washington D.C. behaves like a state; where I've written “state”, include DC as well.

Maine (throughout this time period) and Nebraska (since 1996) allocate some electors on the basis of congressional districts; other states allocate them entirely on a winner-takes-all basis, at least since 1964. Some of the pages provide the vote totals in those states broken down by congressional district, but others don't. The 2016 data table, for example, has three rows for Maine, to denote the three different ways the four electors there were chosen: by plurality in the first congressional district, by plurality in the second district, and by plurality state-wide. Ideally these would be treated as separate jurisdictions for these exercises; where that's impossible (for lack of data) I'm certainly not expecting it, and you're free to make whatever choice you like where it is possible. (Either way you will need to cope with the fact that different pages handle it differently.)

Whether those electors actually voted for the candidate to which they were assigned is a separate issue. **For the purposes of this exercise, work with the number of electors awarded to a candidate, not the number of electoral votes ultimately won by that candidate.** Note that neither Nebraska nor Maine has had a "faithless elector" in this time period, so when those states split their electoral votes, as indicated in the tables (in 2008 and 2016 respectively), use those results, but when other states split their electoral votes (e.g. in Washington in 2016) award all electors to the candidate with the larger number of votes.

The "**two-party vote share**" is sometimes used by political scientists looking at races in which the winner is chosen by a plurality and there are two major parties who are overwhelmingly likely to win. This is simply the number of votes one of the two major parties received divided by the number of votes received in that race. In 1992 in Utah, for example, the Democrat received 183,429 votes, the Republican received 322,632, and one indepedent received 203,400 votes, with other candidates receiving 34,537 votes. Even though the Democrat finished third in this particular race, we consider the Democrat and the Republican to be the two parties of interest here, and the two-party vote shares are

$$\frac{183,429}{183,429+322,632} \approx .3625 \text{ and } \frac{322,632}{183,429+322,632} \approx .6375$$

respectively. (Arguably this should be treated differently, but this is a reasonable simplification. Note that all electors in this time period have been won by the Republican and Democratic parties, though as mentioned previously some of those electors have voted for other candidates.)

If you get stuck it's likely you would be unable to get answers to most of the questions, but you should still be able to provide the SQL commands and the R commands that you would use if you had been able to get everything to work. If you can only get the 1976 and 2016 election data into your database, that's enough to generate one of the interesting R plots, even without the rest of the data. If you can also get the 1980 census and election data and the 2000 election data, you can get full credit for everything that follows the creation of the database, even if you can't build a general scraper that gets all the data requested.