PROJECT: GERRYMANDERING NAME:

Main Research Question

On November 4th 2025, Californians will cast their votes on Proposition 50, which adopts a congressional district map drawn by the state legislature (bill AB 604) until after the 2030 census.

Will the adoption of the new district map lead to an increase in partisan advantage relative to 2024? If so, by how much?

Materials Provided

See the Ed post for links to the assignment repo with the data and starter documents.

Context

- California Statewide Special Election (https://www.sos.ca.gov/elections/upcoming-elections/statewide-special-nov-4-2025): official website for the election from the CA Secretary of State.
- Statewide Database Precinct Data (https://statewidedatabase.org/d20/g24.html): source of precinct-level data from the 2024 General Election.
- 2024 Election Geographic Data (https://statewidedatabase.org/d20/g24_geo_conv.html): source of shapefiles for precinct maps from 2024 General Election.
- Proposed Congressional Map (https://aelc.assembly.ca.gov/proposed-congressional-map): source of shapefiles for district map proposed in AB 604.

Data

1. `bsr-statistics.xlsx`: Ballot return statistics for 2025 special election (source: https://www.sos.ca.gov/elections/upcoming-elections/statewide-special-nov-4-2025 10/29/25).
2. `g24_sov_by_g24_svprec.csv`: Voting results (Statement of Vote) at the voting precinct level for the 2024 general election (source: https://statewidedatabase.org/d20/g24.html).
3. `g24-results-by-district.xlsx`: Voting results at the district level, broken down by county.
4. `g24-candidates-by-district.csv`: Names of candidates running in each district race.

Part 1: Understanding the Data

Visit the website of the California Secretary of State for their information on the upcoming election and Proposition 50 and learn about the process, the documents that they make available, and the data that they make available.

Visit the website of the Statewide Database and learn about what data products are available from the 2024 General Election and any extra information they provide about how to interpret them. Also learn about who runs the database, where it is housed, and how it operates.

Part 2: Data Cleaning

Read in the vote count data from the 2024 General Election at the precinct level. Clean the data frame to ensure values are recorded correctly, then write to desk a cleaned version of the csv to use in the rest of the analysis. Record your code, with comments, in `data-cleaning.qmd`

Tips

- Read the website of the statewide database carefully for disclaimers about the state of the data.
- Inspect columns to ensure they're the correct type. If they're not, investigate why not and correct it.
- Inspect the values in each column to ensure they have the values that you would expect.
- Check that columns that you expect to have unique values indeed don't have replicates.

Part 3: Exploratory Data Analysis

Now that you have a clean data set for precinct-level vote totals from the 2024, election, now is your opporunity to scratch your curious itch and ask a few questions that can be answered with summary statistics or visualizations. Provide at least three such exploratory questions and answers in `exploratory-data-analysis.qmd`.

Examples include: How many precincts/counties/districts featured two candidates from the same party? Which precinct/county/district had the closest race? Do our district totals agree with those found in the district-level results excel file? What is the range of total sizes of the precincts? Which races were the closest? Which races were the biggest blowouts? What proportion of incumbents were re-elected? Be creative here!

Part 4: Calculating Gerrymandering

For the 2024 general election, calculate the mean-median score and efficiency gap associated with the 2024 map of congressional districts. You should have everything you need in your cleaned precinct-level data frame. Record your code and results in `gerrymandering-metrics.qmd`

Tip

- It may be helpful to write a function that takes a vector of votes across districts for party A and vector of votes for the party B and returns a vector of the number votes wasted by party A across districts.
- You may run across races that were not D vs R. Think carefully about how to treat these before calculating your statistics.

Part 5: Re-running the 2024 Election

Calculate what the election results would have been had the 2024 election been run using the district map proposed in AB 604. There are two approaches that can be taken. Both require that you use voting data collected at the level of "SR Precinct" instead of "SV Precinct" (the former unit is always geographical while the latter is not).

The data files listed below are not in the assignment repo; you'll be downloading them directly. Put csv files in `data` and unzip any shapefiles into a new subdirectory called `data/shapefiles`. Put your code for this part into a second section of `data-cleaning.qmd`.

Of these two approaches, Approach A is slightly more direct but induces a bit more error than Approach B. You can choose one to pursue. If you end up doing both, I'd be very interested in an analysis that compares the results. Please email me directly if you do this to share your work.

Approach A: Area-weighted Interpolation

Premise: if we assume the voters are spread evenly across their precincts, we can allocate a precinct's votes to a congressional district in proportion to the percentage of the area of that precinct that is in the district. To this you will need three files:

1. 2024 Votes by SR Precinct: Voting results (Statement of Vote) at the registration precinct level (SR precinct) for the 2024 general election (source: https://statewidedatabase.org/d20/g24.html).
2. SR Precinct Shapefiles: Shapefiles for the SR precincts (source: https://statewidedatabase.org/d20/g24_geo_conv.html).
3. Proposed Congressional Map: source of shapefiles for district map proposed in AB 604 (source: https://aelc.assembly.ca.gov/proposed-congressional-map).

General approach:

1. Clean SR precinct voting data and join their geometries from the shape file.
2. Load geometries of the proposed congressional map.
3. Ensure both geometries are valid and use the same projection (I recommend `st_transform(crs = 3310)`).
4. Use area-weighted interpolation to populate the vote counts of each district in the proposed map.

Note: the people at the Statewide Database are working to correct some errors in the geometries of the SR precincts. `sf` can do a quick job of it using the following:

```r
sr_shp <- sr_shp |>
  st_transform(3310) |>          # switch to equal area projection first
  st_set_precision(1) |>         # snap points to 1 m grid
  st_make_valid() |>             # fix self-intersections/bow-ties
  st_collection_extract("POLYGON")
```

Approach B: Proportional Allocation

Premise: we can improve on the uniform distribution assumption by working with a smaller geographic unit, census blocks. While we don't know the number of votes for each candidate at the census block level, we do know the proportion of registered voters at each census block. We can allocate the votes from the SR precincts across the blocks relative to the proportion of registered voters that live on that block.

1. 2024 Votes by SR Precinct: Voting results (Statement of Vote) at the registration precinct level (SR precinct) for the 2024 general election (source: https://statewidedatabase.org/d20/g24.html).
2. SR Precinct to Block Correspondence: Number of registered voters by (partial) block (source: https://statewidedatabase.org/d20/g24_geo_conv.html).

3. Block membership in AB 604 districts: A list of the census block keys and their new congressional district as of AB 604 (source: https://aelc.assembly.ca.gov/proposed-congressional-map).

General approach:

1. Load and clean SR precinct voting data.
2. Working from the census block data, add SR voting data to each (fractional) block in proportion to the percentage of registered voters in the precinct that live on that block.
3. Identify which new district each block is in.
4. Tally up the votes for each new district.

Note: there are some census blocks that are split between multiple SR precincts. Accordingly, each row of the SR Precinct - Block correspondence file is actually a potential fractional block: most rows represent an entire block, but some represent the portion of a block that is in a particular precinct. `BLKTOTREG` is the total number of registered voters in the full block; `SRTOTREG` is the total number of registered voters in the SR precinct. `PCTSRPREC` is the percent of the total SR registrants that live on the fractional block. `BLKREG` is the number of registered voters on the fractional block.

Part 6: Calculating Gerrymanding Again

Return to `gerrymandering-metrics.qmd` and calculate the same two metrics using hypothetical re-run of the 2024 election using the new district map.

Part 7: Dashboard

Your final deliverable is a data dashboard, written in `dashboard.qmd`. It should at minimum include the following:

1. Makes it easy to compare the 2024 congressional results with a hypothetical re-run of the election using the new district map.
2. Presents two maps, one using the 2024 districts and the other using the new districts. You can use any mapping library for this (ggplot2, base R, leaflet, etc.). You can find the shapefiles for the 2024 districts at https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2024&layergroup=Congressional+Districts+%28119%29.
3. For both the observed 2024 election results and the hypothetical re-run, presents at least three summary statistics:
   - Number or proportion of seats won by Democrats (or Republicans)
   - Mean-median score
   - Efficiency Gap
4. Includes a "Methodology" section that describes the data source, the data cleaning procedure, the estimation process, and caveats about inferring gerrymandering from single statistics.

If there is another visualization or statistic that you find insightful and effective, you're welcome to add it.

Take care in presenting the information in a manner that is clear, easy to navigate, and facilitates the most important comparisons. You can you column-wise or row-wise layouts and as many pages as you like.

I recommend reading in your cleaned data at the top of the dashboard file, creating all of your data products, then creating your dashboard layout beneath that, as in the demo-dashboard example.

How to Submit

Once you're happy with your work, render each of the Quarto documents one more time and commit both the .qmd files and the .pdf files that are produced. Then push these to your repo for your GSI to read. The dashboard is interactive, so you'll just be able to submit the .qmd file that created it. However…

Optional: Use `quarto publish` to publish your dashboard on the web through quarto pub. If you do, please add a `README.md` file to your repo before you submit it to point your GSI to the URL!