*Sources, Licenses, and Concerns*

I plan to work with [538](https://data.fivethirtyeight.com/)’s [Election Denier](https://projects.fivethirtyeight.com/republicans-trump-election-fraud/) dataset primarily, combining it with election data, past and future. The goal of the project is to highlight the different constituencies demographics of election deniers, and show differences in how voters reacted to election denial. Several of the constituencies have at least one race without a denier and at least one with, which will allow me to generate some near graphics. To truly construct good visualizations, I’ll snag some district level data through Cook PVI or Dave’s redistricting (or both) if they allow it. 538 also has district data, but I’ve found their PVI metric to be deceptively generated (AL-4 is rated as [R+63](https://projects.fivethirtyeight.com/redistricting-2022-maps/alabama/), which is impossible), but they’ll work in a pinch since they have the clearest license. Data for the 2022 election results, which I hope to join to, is not out yet- I’ll have stand in data ready to build my visualization, and when the time comes will find a source that has the data published under a usable license.

538’s dataset is under Creative Commons Attribution 4.0 International, which allows me to share and adapt with proper attribution. Cook is on a [case-by-case clearance](https://www.cookpolitical.com/terms-and-conditions), so I’ve emailed them, and Dave’s Redistricting is [unclear](https://davesredistricting.org/TermsAndPrivacy.html), so I’ve reached out to them as well. If neither yield, I’ll investigate public records if I must.

The election data is collected via public records, combining geographic data with election results. The election denial data is sourced from a variety of campaign and personal publications of candidates, primarily speeches, twitter, and campaign websites. Because of 538’s reputation, thoroughness, and constant updates, I don’t believe they are biased in reporting on the issue. And overall, because the data is complete and factually sourced, there won’t be any sampling issues.

This data has no privacy concerns. All election data is at the district level at the smallest, and all denial data is either openly sourced from public officials, or listed as “No Comment.” Because the information isn’t gleaned deceptively or from private conversations, the names and statements of the candidates are clearly fair use.

*Data Cleaning and Diagnostics*

538’s data was incredibly clean to start with, and my personal copies of Cook/Dave’s datasets were as well, thanks to previous works. Because the PVI datasets I have State and District already in format, I simply had to join them together with left joins (just on state for statewide offices, and on state and district for house members). This required me to split the dataset into statewide and district candidates, merge with PVIs, and then rejoin to prevent overwriting. I also had to correct capitalization errors within 538’s data, to make sure categories were unified.

**Figure 1: Missing Data**

A picture containing text

Description automatically generated

There were some incomplete columns in the dataset. As seen in Figure 1, the Note, Source, and URL columns all had gaps in information. Note has a reason to be “NA” almost 100% of the time- it only records changes in rhetoric. For the other two, I did some exploration below.

Figure 2: Sources of Data without URL Figure 3: Stance of Data without Sources

Chart, bar chart

Description automatically generated Shape

Description automatically generated

Figure 2 demonstrates that the entries without associated URLs are predominately sourced from candidates emailing 538 with context on their stances. Other options include out of Congress during the roll call vote, or written stances to other sources not publicly available. Seeing no NAs, I now can certify that the incomplete URLs all make sense. Figure 3 shows a similar understandable missingness- all of the unsourced entries are due to a candidate making no comment on the 2020 election. Therefore, we can certify that the data is as complete as possible.

*Data and Distribution*

The dataset contains the 2020 election stances of 540 Republican officials running for national office; the bulk of the candidates are in House races, with 34 governors, 35 Senators, and a handful of other state officials. PVI has a wide distribution, although the races not covered also tend to be the most Democratic leaning (because they do not have a GOP Candidate). The most important distribution of the data is their stance on the authenticity of the 2020 election, featured below.

Figure 4: Stances on the 2020 election

Chart, pie chart

Description automatically generated

*Data Dictionary*

|  |  |
| --- | --- |
| Column | Description |
| `Candidate` | The name of the candidate in question. |
| `Incumbent` | Whether the candidate was the incumbent. |
| `State` | The state the candidate was running in. |
| `Office` | The office the candidate was running for. |
| `District` | The number of the congressional district the candidate was running for, if applicable. |
| `Stance` | The candidate's stance on the legitimacy of the 2020 election. |
| `Source` | The source from which FiveThirtyEight determined the candidate's stance. |
| `URL` | The URL of the source website from which FiveThirtyEight determined the candidate’s stance, if available. |
| `Note` | Whether the candidate changed his or her stance after their primary election. |
| `PVI` | A join to the Cook PVI of the constituency, showing the past relative presidential voting trends of the district |