Redacted Name
STA 141
28 October 2019

Visualizing Democratic Primary Polling Results

I – Introduction

Compared to other democracies, the United States has an unusually lengthy election cycle, resulting in billions of dollars of campaign expenditures and countless hours of political punditry. The merits of beginning election analysis over a year before a general election are debatable, but one exciting result of the American election system is that it provides ample opportunity to amass large quantities of polling data for visualization and analysis.

American Presidential elections are clearly consequential for the United States and, by extension, the world. However, the 2020 election will be particularly interesting given the increasingly polarized political landscape in which it is taking place. Furthermore, electoral outcomes are especially significant in the context of time-sensitive and existential challenges such as climate change. The largest field of Democratic candidates in American history has opted to rise to meet this occasion.

As a result, an unprecedented amount of attention is being dedicated to efforts to understand and predict the outcome of the Democratic primary election process, through which the Democratic nominee for the general election is selected. Since Jimmy Carter's unexpected Iowa caucus win in 1975, pollsters and political analysis have dedicated abundant resources to "the Olympics of pandering" by tracking voter sentiment in states with early caucuses and primaries (Malone 2016). This report takes advantage of the availability of early polling data to visualize levels of support for Democratic primary candidates, specifically Joe Biden, Elizabeth Warren, Bernie Sanders, Kamala Harris and, when adequate data exists, Pete Buttigieg.

II – Methods

Graphics and maps were produced using Democratic primary polling data from FiveThirtyEight, map data and insetting procedures for Alaska and Hawaii from Richard Carega on RPubs, and Census division data from Chris Halpert's GitHub page. From these sources, I created three data sets of Democratic primary polling results. The first data set includes head-to-head polls at the national level. The second data set includes all other poll types at the national level. The third data set includes all poll types except head-to-head polls at the Census division level.

To create the first data set, I filtered the FiveThirtyEight polling results to exclude state-level results, to include only head-to-head polls, and to exclude all candidates except for Biden, Warren, Sanders and Harris. This yielded a data set with 11 polls from May to October of this year, representing Biden in 11 polls, Warren in 11 polls, Sanders in 10 polls, and Harris in 7 polls. As Buttigieg was included in only 1 head-to-head poll, he was excluded from this data set. Results in the FiveThirtyEight data set were presented as pairs of rows, so I created a data set indicating the month associated with the start of the polling

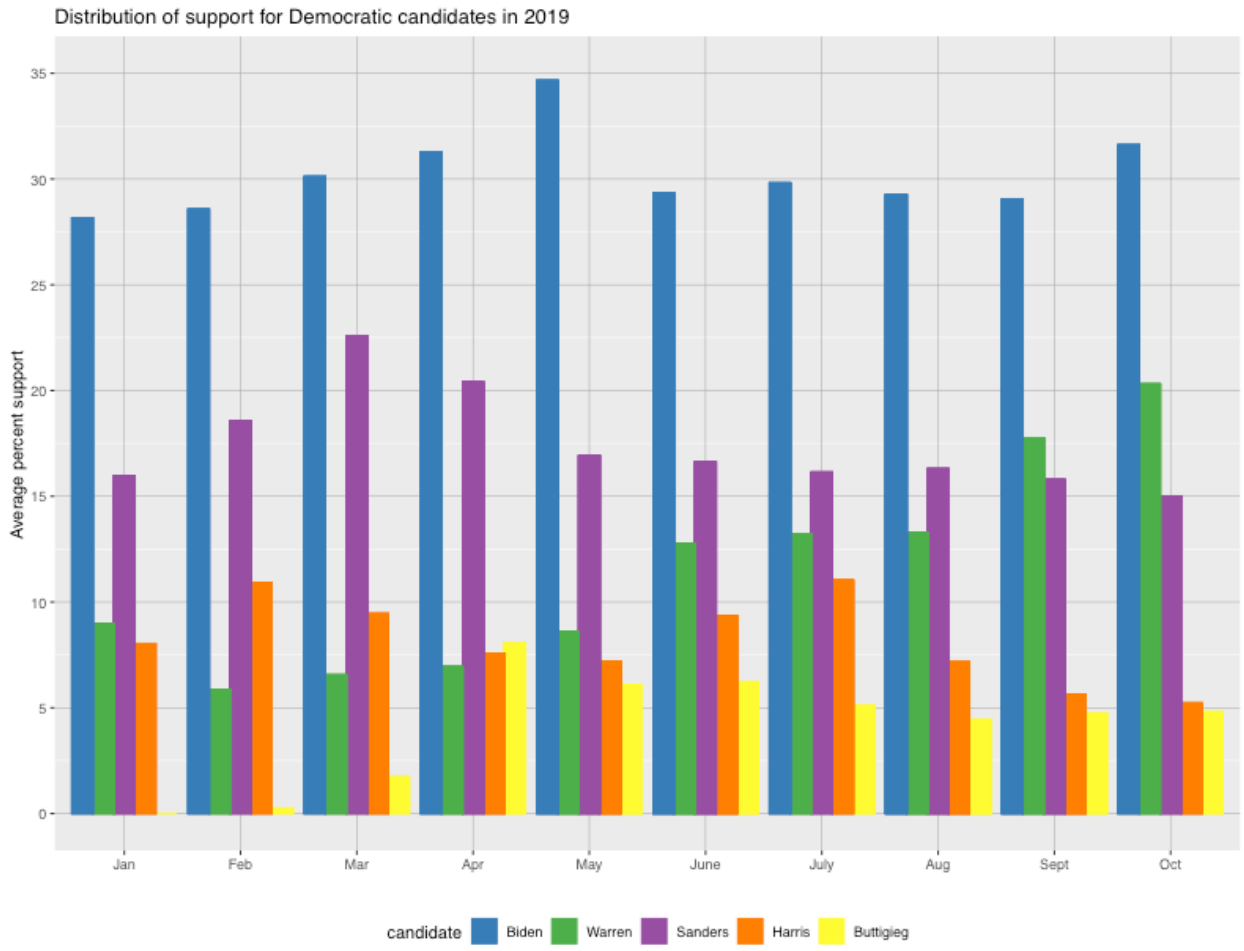
period, the two candidates being considered, and their respective polling percentages. Then, I created a data set for each candidate, which included the month, the competing candidate with whom they were being compared, the candidate's polling percentage, and the difference between the candidate and the competing candidate. The final step was transforming these data sets to reflect monthly averages.

The second data set similarly excluded state-level results, but it excluded head-to-head polls rather than including them, it included Buttigieg, and it excluded polling data from before 2019. This yielded a data set with 329 polls from January to October of 2019, in which all candidates were represented in all polls with the exception Buttigieg who was only included in 296 polls. For each candidate, monthly averages were calculated, as well as monthly averages by polling grade (re-factored as categories representing A's, B's, C's, and all lower grades or missing values), and by the population polled (registered voters, likely voters, voters, and adults).

The third data set included only polls for which a state value was supplied and considered Biden, Warren, Sanders, Harris and Buttigieg. Biden, Warren, Sanders and Harris appeared in all 207 such polls, and Buttigieg appeared in 196. Census division data was joined to these polls, and polling averages for each candidate were calculated by division so that there would be no missing values when the 9 divisions were mapped for each candidate. The South Atlantic division had 45 polls, the Mountain division had 40 polls, the West North Central division had 35 polls, the New England division had 30 polls, the Pacific and East North Central divisions had 25 polls, the East South Central and West South Central divisions had 20 polls, and the Middle Atlantic division had 15 polls. Filtering the state-level data to include only Arizona yielded 5 polls.

III – Results

To visualize the national-level monthly averages for each candidate considered, I created the following bar graph.



Joe Biden leads the field in each month, with his highest level of support occurring in May, following the announcement of his candidacy in late April. His level of support has remained relatively consistent throughout 2019. Despite dramatic policy differences between Biden and Sanders, voters who prefer Biden or Sanders are most likely to have the other of those two candidates as their second choice (Rakich 2019). This helps explain the consistent dip in Sanders' poll numbers in the month following Biden's announcement.

Warren is the candidate that Sanders supporters are the second most likely to list as their second choice (Rakich 2019). Warren is steadily ascending in the polls, after a low point in February when she released the results of her DNA test in order to illustrate Cherokee heritage. Since Biden and Sanders' polling numbers have remained relatively consistent since June, it is reasonable to infer that Warren's increased level of support is in large part driven by voters abandoning lower-tier candidates in favor of Warren.

Similarly to Biden and Sanders, supporters of Harris or Warren tend to list the other of these two candidates as their second choice (Rakich 2019). Harris experienced a jump in the polls in July, following her strong showing in the first Democratic debate in late June. This was later followed by weaker debate performances and decreasing poll numbers. Many analysts question whether debate performances are significantly consequential in the long term, while others emphasize the role that early debates play in increasing name recognition of lesser-known candidates. Regardless of the specific dynamics at play, the

excitement surrounding Harris' candidacy has dissipated since July. Many of her supporters have probably turned to Warren instead, contributing to her increase in the polls.

Lastly, more of Buttigieg's supporters listed Harris as their second choice than any other candidate, and a significant share of Harris' supporters also support Buttigieg as a second choice (Rakich 2019). This helps explain the loose inverse relationship between their poll numbers starting in April. While Harris has typically led Buttigieg in the polls, they are now virtually tied.

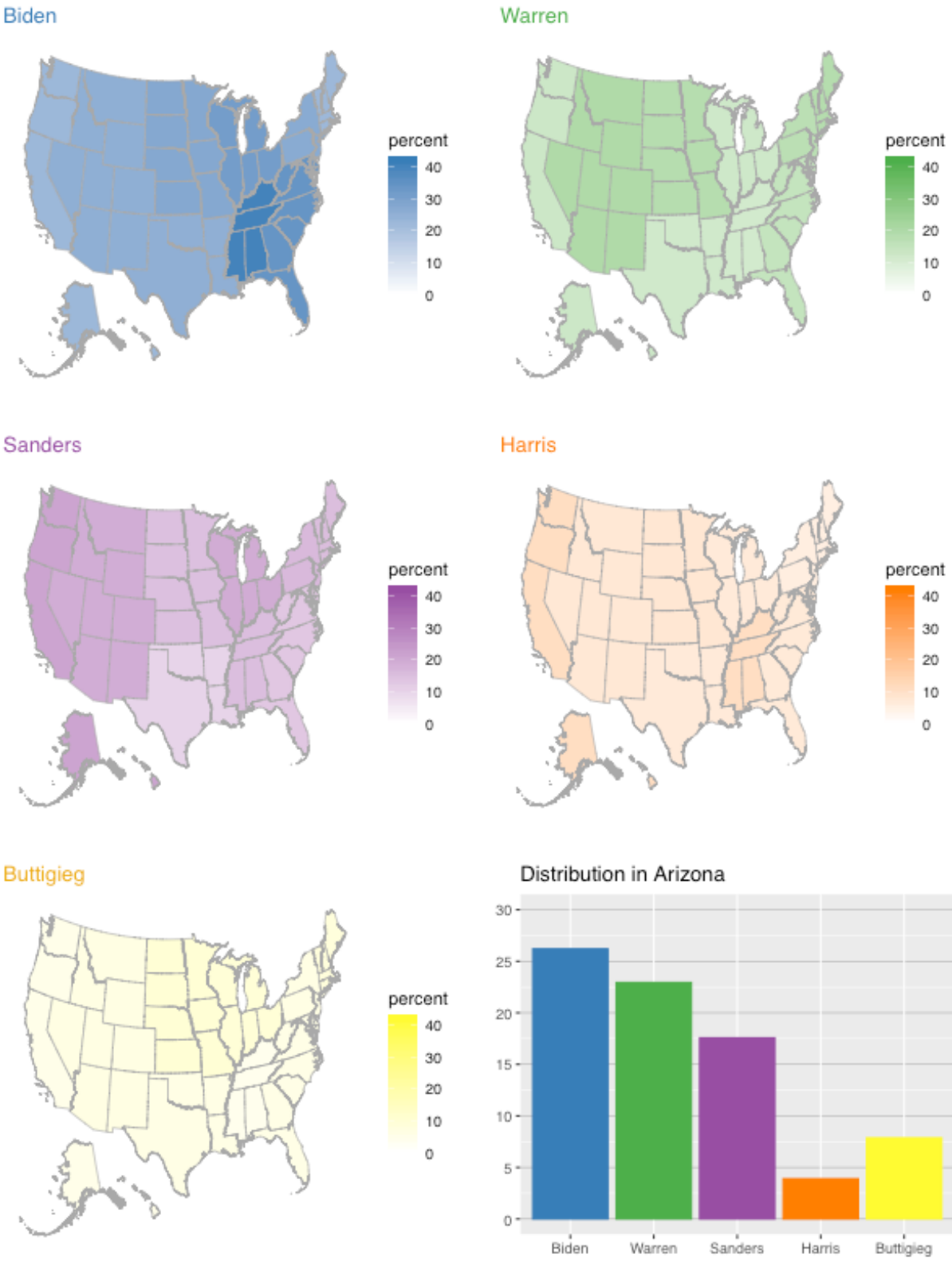
Head-to-head polling results provide additional details about relationships between support between pairs of candidates. However, the following graphic should be analyzed skeptically since it represents only 11 polls, with Biden and Warren being the only candidates represented in all of them.

Many of the aforementioned national-level trends are also evident in these head-to-head comparisons, but there are also unexpected trends that can most likely be explained

by a lack of data. For example, when Biden and Warren are compared, the margin by which voters prefer Biden fell from around 23 percentage points in May to around 3 percentage points in October. This reflects Warren's gains in the polls and the increasingly favorable view of her "electability." However, when Warren is compared to Sanders or to Harris, there is no clear trend, with her results against Sanders being particularly variable.

National-level polling results are important because they are one of the criteria for determining candidates' eligibility in Democratic debates, and because they ideally give an indication of the overall preferences of Democratic voters. However, state-level results are perhaps more important in determining the Democratic nominee because primary elections are not held by all states on the same day. Since early states like Iowa, New Hampshire, Nevada and South Carolina hold exaggerated significance, it is often important to understand state-level trends.

Furthermore, analyzing regional support for various candidates can help indicate whether a candidate has a stronger or weaker chance of winning swing states in a general election. Since polling data is not yet available for all states, average support for candidates in 2019 was considered by Census division in the map below. In the context of a general election, the electoral vote share for many of these divisions are a foregone conclusion. However, many states in the East North Central division, some states in the West North Central and Mountain divisions, and Pennsylvania and Iowa are swing states whose outcomes could drive the results of the Electoral College.

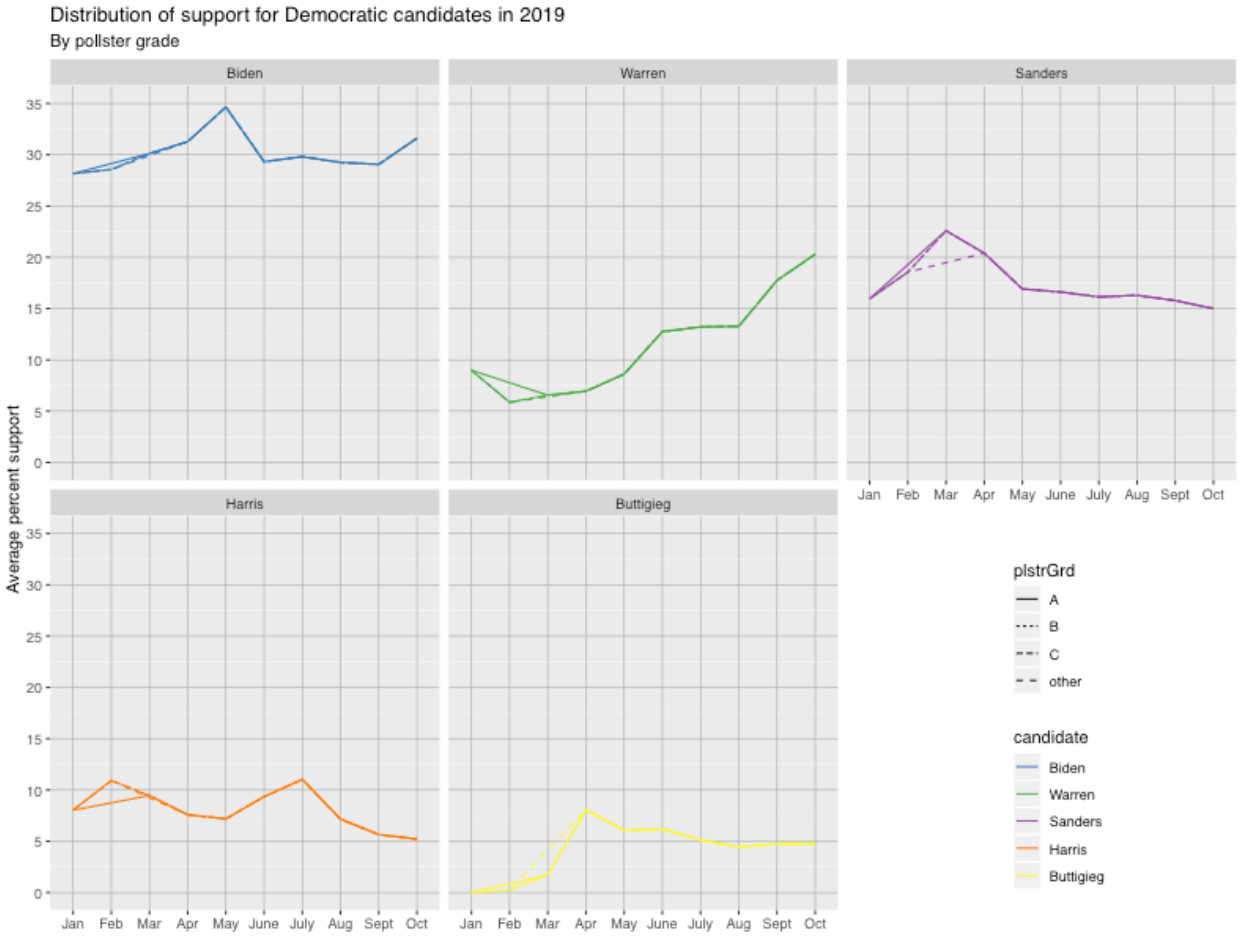


As illustrated by the maps, support for Biden is strongest in the South, the upper Midwest, and the Midwest. By contrast, Warren's supporters are disproportionately from

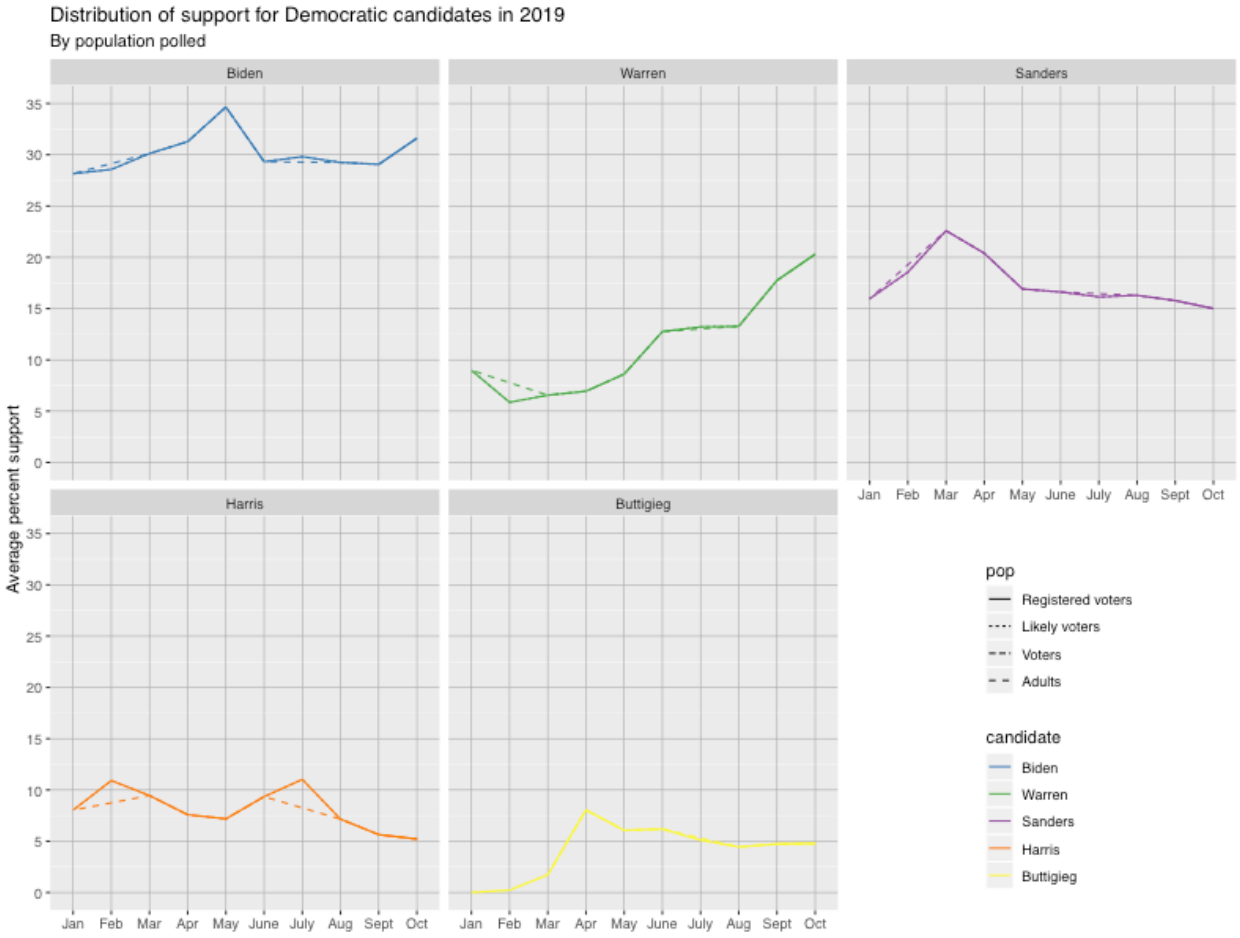
the Mountain region, the eastern Midwest, the Mid-Atlantic region and New England. Sanders has high levels of support on the West coast, in the upper Midwest, in the Mountain region. Harris performs best on the West coast, which includes her home state of California, and also performs well in the South. Buttigieg's map also reflects an advantage in the divisions containing and surrounding his home state, with his highest levels of support being in the Midwest.

Finally, graphs of national-level polling results grouped by pollster grade and by the population polled were also produced. The results between groups were extremely similar, with most seeming discrepancies simply being the result of months with missing data. The graphs grouped by pollster grades include 42 "A" level pollsters, 147 "B" level pollsters, 142 "C" level pollsters, and 18 pollsters with unknown grades or grades lower than a "C."

Though these averages are quite similar across groups, there are often considerable differences between specific polls. For example, two polls released within this past week produced dramatically contradictory results, with a Quinnipiac poll placing Warren ahead of Biden by 7 percentage points while a CNN poll had Biden leading Warren by 15 percentage points (Bump 2019). Such variation is not necessarily a bad thing; in fact, too little variation may indicate "herding," which describes "the tendency of polling firms to produce results that closely match one another, especially toward the end of a campaign" (Silver 2019). Demonstrating a tendency to engage in herding is a characteristic that results in a lower FiveThirtyEight pollster grade.



One pollster with an A+ grade, Selzer & Co., is considered by FiveThirtyEight to be the "best pollster in politics" (Malone 2016). Ann Selzer is known for making minimal assumptions about whether registered voters being polled are likely to vote in a caucus or primary, which allows for more flexibility when candidates generate higher levels of excitement among voters without a strong history of electoral participation. This was one of the reasons that she was able to correctly predict Obama's Iowa Caucus win in 2008 (Malone 2016). As a result, it was possible that there would be interesting variation in candidates' national averages by population polled. However, there was very little such variation. The graphs of average monthly candidate results below represent 183 polls of likely voters, 189 polls of registered voters, 1 poll of voters, and 16 polls of adults. As a result, the lines are only complete for registered voters and likely voters.



IV – Discussion

Visualizations of poll results will become more meaningful and appropriate for deeper analysis as more polling data becomes available. For example, once all states have some minimum number of polls, it will be possible to conduct a more nuanced analysis of geographic trends since candidate results may then be mapped by state rather than by Census division. Fortunately, the quantity of available polling data increases at a rather rapid pace, with 13 additional polls already having become available on FiveThirtyEight since this data for this report was pulled on the 19th of October. Additionally, FiveThirtyEight pollster grades will be updated in the next several days, allowing for higher quality analysis (Druke 2019).

V – References

Bump, P. (2019, October 24). Poll: Biden has a big lead. Another poll: Warren does. Washington Post. Retrieved from https://www.washingtonpost.com/politics/2019/10/24/poll-biden-has-big-lead-another-poll-warren-does/

Careaga, R. (2015, July 21). Thematic Mapping in R without the Tears, Insets and Outsets.

Retrieved from https://rpubs.com/technocrat/thematic-alaska-hawaii.
Careaga, R. (2015, July 20). Thematic Mapping in R without the Tears, Walkthrough. Retrieved from https://rpubs.com/technocrat/thematic-walkthrough.
Druke, G. What should Republicans do? FiveThirtyEight Politics [Audio podcast]. (2019, October 28). Retrieved from https://fivethirtyeight.com/features/politics-podcast-what-should-senate-republicans-do-about-impeachment/
Halpert, C. (2014, June 23). us census bureau regions and divisions [Data file]. Retrieved from https://github.com/cphalpert/census-regions/blob/master/us%20census%20bureau%20regions%20and%20divisions.csv
FiveThirtyEight. (2019, October 19). Presidential primary polls [Data file]. Retrieved from https://github.com/fivethirtyeight/data/tree/master/polls
Malone, C. (2016, January 27). Ann Selzer Is The Best Pollster In Politics. Retrieved from https://fivethirtyeight.com/features/selzer/.
Rakich, N. (2019, July 10). Lanes Are Starting To Emerge In The 2020 Democratic Primary. Retrieved from https://fivethirtyeight.com/features/lanes-are-starting-to-emerge-in-the-2020-democratic-primary/.
Silver, N. (2014, November 14). Here's Proof Some Pollsters Are Putting A Thumb On The Scale. Retrieved from https://fivethirtyeight.com/features/heres-proof-some-pollsters-are-putting-a-thumb-on-the-scale/.