**538ers ETL Project Description**

The purpose of this project is to compare polling data for individual candidates for the Democratic nomination for the Presidency with each other and with approval polls for Donald Trump.

We expect to see an inverse relationship between overall support for Democratic candidates with approval of Trump's performance. We would also expect to see a stronger relationship between tracking polls for Trump and for more moderate Democratic candidates like Biden and Gabbard. Voters who are at the edge of supporting Trump (those few percent who vary between approval and disapproval), if they report to a pollster that they will probably vote for a Democratic should be more likely, the should be more likely to support the Democratic candidate they perceive as more moderate.

Data were drawn from two sources: fivethirtyeight.com and Rasmussen polling. Rasmussen publishes polls five days a week, excluding holidays, using daily polling into a rolling average model. Fivethirtyeight.com aggregates polling data from a variety of non-proprietary sources. These polls are usually published with a date range. For these polls, we use the last day in the sampling period as the poll date.

To extract the data, we are using python with beautiful soup to extract the html code from the web pages. With the html code extracted we then used loops to cycle through the tables and pull out particular columns.

Fivethirtyeight.com polling dates were published in a variety of formats, essentially whatever the source pollster used. These dates proved to be extremely difficult to convert and consumed most of our time.

The six polls that excluded Biden were dropped. Only one poll excluded Warren – but it had been dropped as it excluded Biden. The next two candidates, Sanders and Harris, were not excluded from any polls. Buttigieg, Yang, O'Rourke, and Booker were excluded from more polls. For the most part, 0 imputation was appropriate in these cases as the candidates were polling at or near zero in contemporaneous polls. A few (2-4) polls excluded Beto O'Rourke at times when he likely would have been the top choice of 4-8% of the respondents. These were coded to 0 as well.

Preferences for Democratic candidates varied over time. This was likely in response to how many and who were running, news cycles, and campaign performance. Approval for Trump varied little over the time studied – it took him about three months in office to harden positions on both sides. Subsequently, Trump approval hardly varied at all in response to outside events or news cycles. (Id est, support for Trump has little to do with what is happening in the real world.)

There was no relationship between approval for any candidate and for Trump. Disaffected and waffling Trump supports are generally not likely to vote in Democratic primary races. Most polls only included likely Democratic primary voters. This "LV" bias probably dampened any effect we would have otherwise seen. However, the "reality exclusion" effect is probably the most important explanation.