

Matthew Burgess
Capstone 1 Project
Data Wrangling Summary
7 March 2020

Data Wrangling Summary: Pollster Data

Form of Data:

I've saved the data to my github repository in both .xlsx and .csv formats. It is roughly 9000 rows by 25 columns of pollster data.

Feature Engineering:

In examining polls and their error, the sample size of the poll itself is most important as an indicator of whether the margin will be significantly large. However, an interesting question is to find the lowest threshold for sample size that will still produce a reliable outcome. Therefore, I've kept all of the polls regardless of sample size.

Other feature engineering that needed to happen is to convert all of the dates to DateTime entries, and to consolidate redundant columns. This includes getting rid of the "pollno" column and to sort based on type of race.

The data is organized such that the column "cand1_pct" is always a Democratic race for general presidential election polls and "cand2_pct" is always a Republican race for general presidential election polls. When the error is calculated as a difference between the two rows actual results respectively, a positive value for the column "margin_actual" indicates a Democratic victory, and a negative value for the same column indicates a Republican victory. Any sorting of rows based on type of race or party affiliation I will do in later data exploration. I have also deleted the "comment" column as well as removed all partisan pollsters as indicated by the "partisan" column.