# Data Dictionary

We will be establishing a list of variables that are crucial to understanding the dataset.

**General Information**

Throughout the dataset, we will be using the poll’s **state** as a classification method. This variable includes all of the states in the United States, as well as an additional category ‘U.S.’ which refers to a national poll. Each poll has an associated **start date** and **end date** which correspond to when the poll occurred. Additionally, we have added a column for **duration**, or how long the poll was open for. The **month** in which the end of the poll occurred was also recorded. For example, if a poll started in October 2016 but ended in November 2016, its month would be November 2016.

There are many ways that a poll could be administered – online, live caller with cellphones, in person, etc. We looked into **internet** polls, which are polls that were administered at least partially through the internet. If a poll was (at least partially) administered online, then it would have a ‘1’ in the corresponding column. If not, then there would be a ‘0’. We also looked into **live caller with cellphone** polls, which indicates whether or not the polling firm is routinely placing calls to cellphones (and landlines) in its polls, excluding polls administered via text. A pollster would receive a ‘1’ if it does include polls over the phone and a ‘0’ otherwise.

**Pollsters**

We were given a dataset of around 4200 different polls that took place within the year leading up to the 2016 US Presidential Election (November 2015 to November 2016). We looked at 196 different **pollsters**, who are the organizations that have contributed the most to the methodology and execution of the poll in terms of intellectual property. Each pollster has a corresponding name of a polling organization, not to the organization that paid for or sponsored the poll. Organizations that have similar methodologies or often collaborate on polls may be grouped together, as per FiveThirtyEight’s discretion.

Each pollster received a corresponding **538 grade** or **grade**, which is a letter value from A+ to F or a normalized numerical representation between 0 and 1 that reflects the accuracy of the polling organization as well as its predictive plus-minus score (see ‘Plus-Minus Scoring’ below). For the 538 grades, if a firm is **banned by 538** then it automatically receives an ‘F’ (these firms were excluded from our dataset). If a pollster has fewer than 10 polls, then it receives a provisional grade of ‘A/B’, ‘B/C’, or ‘C/D’ depending on the strength of its initial performance. This grade is calculated through an analysis of the historical accuracy of each polling organization in the past as well as their methodology. For the **grade**, we manually assigned a numerical value to each of the polls based on their 538 grade. In our dataset, the lowest grade was a ‘D’, which corresponds to a ‘0’, and the highest grade was an ‘A+’, corresponding to a ‘1’. For example, if a pollster received a grade of ‘C+’, then its associated numerical grade would be a 0.333. Pollsters with fewer than 10 polls automatically received a provisional grade of 0.333. These grades will be the focus of our analysis.

If a polling firm was signatory to either the National Council on Public Polls (**NCPP**) or the American Association Public Opinion Research (**AAPOR**) Transparency Initiative, or was a contributor to the **Roper** Center data archive, then it would have a ‘1’ in the corresponding column. If not, there would be a ‘0’ instead. This classification was used as a proxy for methodological quality.

**Polling**

There are two different groups of polls in the dataset, namely the **raw polls** and the **adjusted polls**. The raw polls are the percentages that voted for each of the four candidates in the poll. The adjusted polls are the FiveThirtyEight adjusted percentages that voted for each candidate. For Johnson, if there was a missing value it was replaced with the global mean of the column, i.e. the global mean of all of his votes throughout all the polls. For McMullin, if there was a missing value it was replaced by 0 since there were too few entries to justify taking the global mean of the column. Each poll has an associated **polling weight**, which is based largely on how long before the election the data was collected, with the sample size and pollster grade contributing to the weight as well. Generally speaking, the older the poll the lower the weight and thus very older surveys will have weights close to or equal to 0. The polls that did not have a corresponding weight received a ‘0’ in the associated column.

**Population**

We will be working with different groups of populations, namely **likely voters, registered voters, voters,** and **adults**. Since not much information could be determine with the different categorical values, we did One-Hot Encoding in order to create four additional columns, **population type**, where if the original column had, for example, likely voters then the corresponding population type would have a ‘1’ in the column and the other three columns would have ‘0’. Additionally, the **sample size** of the poll was reported which we then smoothed by creating bins and placing values with the median of the contents of the bin, using the Fisher-Jenks Algorithm. Additionally, we reported the **number of polls** for each pollster in the FiveThirtyEight database that covered polls that have been conducted by the House, Senate, gubernatorial, and presidential general election campaigns since 1998, as well as the polls in presidential primaries and caucuses since 2000, all in the final three weeks leading up to the elections.

**Plus-Minus Scoring**

How a pollster’s average error compares to another is called the **advanced plus-**minus score. This score considers the type of election polled, the number of days until the election, the poll’s sample size, the competitiveness of the race, and the number of polls being done by other pollsters on the same election. A more recent poll has a heavier weight in the calculation of the score. A projection of how accurate the poll will be in future elections is called the **predictive plus-minus** score. This is calculated by reverting the pollster’s advanced plus-minus score to a mean based on FiveThirtyEight’s proxies for methodological quality and then applying penalties for herding. This is the basis for the polling weight defined above. A negative predictive or advanced plus-minus score is favourable and indicates above average quality for the pollster. The **historical advanced plus-minus** score refers to the average accuracy of past polls for a particular pollster compared to other pollsters’ results in the same surveyed race. This also helps to measure how closely the polling number of previous races have aligned with the elections in the past. An additional plus-minus score that we looked into was the **mean-reverted advanced plus-minus**, which is the advanced plus-minus score that has been reverted to a mean of zero. These four scores can help show the accuracy of pollsters in the past and present.

**Bias**

A pollster can report a **bias**, which is its historical average statistical bias towards a certain political party, namely if they leaned towards the Democratic or Republican candidate. For example, if a pollster has a bias score of “R +0.2” or -0.2, this indicates that the pollster, historically, has projected a margin 0.2 percentage points more favourable to the Republican candidate than the actual results of the races it surveyed. Positive numerical values correspond to percentage points more favourable to the Democratic candidate. The **number of polls for bias analysis** is the number of polls when the top two finishers of the poll was a Democrat and a Republican. This number excludes the polls of the single-party primaries and in the case that an independent candidate finishes in the top two. The **mean-reverted bias** is the pollster’s bias score that has been reverted to a mean of zero. Alternatively, it is the pollster’s historical average statistical bias towards one of the Democratic or Republican candidates that has been reverted to a mean of zero based on the number of polls in the database. For example, if a pollster has a score of “D +2”, this indicates that the pollster has historically overrated the performance of the Democratic candidate, possibly indicating a strong bias on average towards democrats. The average amount in which the firm’s polls have favourable a Democratic or Republican candidate in comparison to the other surveyors of the race is called the **house effect**. This is calculated by subtracting the average number of percentage points of the other pollsters in the race from the specific pollsters’ number of percentage points. In the situation where there was only one pollster that surveyed the race, it is not included in the calculation of the house effect. In addition to the house effect, a **races called correctly** category was reported on which shows the percentage of polls in which the pollster was able to correctly identify the winner of the race. If a tie was indicated and one of the candidates in the tie won the race, the pollster would receive a half credit.