Data Dictionary

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. 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.

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, 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). If a pollster has fewer than 10 polls, 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 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 a 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 determined with the different categorical values, we did One-Hot Encoding 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 poll's **sample size** was reported, which we then smoothed by creating bins and placing values with the median of the bin's contents, 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 has 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.

Script

Introduction

[Opening Scene: Jenny enters a meeting room where Jim is waiting.]

Jim: Ah, Jenny, good to see you! Are you ready to present the results of our data analysis on pollster grades?

Jenny: Absolutely, Jim! I've got the insights you've been eagerly waiting for. Today, I'll unveil the story behind those grades and shed light on the factors that influenced them.

Jim: Excellent! I'm eager to delve into the details. Let's uncover the patterns and gain a deeper understanding of pollster performance during the 2016 US Presidential election.

Jenny: That's the spirit, Jim! I've crunched the numbers, examined the methodologies, and scrutinized the data. I'm confident our analysis will provide valuable insights for our decision-making processes.

Data Dictionary

[At the beginning of each section, Jim and Jenny are next to each other as if they are talking]

(1. The Relationship Between the Variables and Grade)

Jenny: First, though, I want to introduce you to some of the terminology you need to know. The most important term you need to know for this presentation is what a **grade** is. It is a letter value from A+ to D that reflects the accuracy of the polling organization as well as its predictive plus-minus score, which will be covered shortly. Each **pollster**, who is the polling organization that conducts the poll, has contributed the most to the methodology and execution of the poll in terms of intellectual property. These two terms are the focal points of our investigation today!

(3. NCPP)

Jenny: Throughout the next few minutes, we will be looking into the three following organizations:

- The National Council on Public Polls, or NCPP,
- The American Association Public Opinion Research Transparency Initiative, or AAPOR, and,
- The Roper Center data archive.

(5. Density Plot of Duration)

Jenny: As we moved forward in our investigation, we realized we wanted to look further into how the time length of the poll interacts with the pollster grades. We created a new variable, called **duration**, which is the length of the poll, essentially it is the end date minus the start date. This is a key variable throughout the next part of our analysis.

(6. Scandals)

Jenny: Jim, we now want to look at the data, but when it’s broken down by **month**.

(7. Plus-minus)

Jenny: Before we move into the last bit of our analysis, we need to introduce some essential terminology. You will need to know about:

- How a pollster’s average error compares to another, which is called the **advanced plus-minus** score,
- A projection of how accurate the poll will be in future elections, which is called the **predictive plus-minus** score,
- That 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, and
- That the **mean-reverted advanced plus-minus** score is the advanced plus-minus score that has been reverted to a mean of zero.

1. The Relationship Between the Variables and Grade

[Continuing with Jim and Jenny next to each other as in the Data Dictionary parts]

Jenny: Jim, before we dive into the different factors we looked into that impacted the pollster grades, we first want to tell you about our motives and why we looked into what we did.

Jim: Do tell!

[Switch to the graph with Jenny in the bottom right]

Jenny: Here we have a plot of each of the variables in our dataset and their correlation with the grades. This was our initial guide as to what we should be looking into. You'll see that the predictive and mean-reverted plus-minus scores as well as the start and end dates all had high correlations to the grade. Additionally, you can see that the variables NCPP/AAPOR/Roper and duration also had a relatively high correlation with the grade. These are all examples of the factors we will be looking into later and how they relate to the grade.

[Transition of Jenny and Jim side by side]

Jim: I see! So you looked at which factors had the highest correlation and analyzed them further to figure out why there was such a high correlation?

Jenny: Exactly! We wanted to understand why a pollster gets the grade that they do, and by looking at the factors that have the highest correlation to them, we hoped to gain some insight into our question.

2.1 States and Number of Polls (histogram)

(Bivariate/Univariate 1)

[Transition to displaying the histogram with Jenny in the bottom right]

Jenny: Jim, next I want to introduce you to the first part of our grade investigation… We began by looking at the number of polls and which state it was administered in. There were a few states that had high numbers of polls, namely Florida, Pennsylvania, and North Carolina, which all have a big role in the outcome of the election. Altogether, they made up over 20% of the electoral votes needed to sway the election and over 25% of the registered voters in the United States.

[Switch to Jenny and Jim side-by-side talking]

Jim: So you're saying there might be a correlation between the states and the number of polls administered to them?

Jenny: That's right! You'll also see that the majority of polls taken during this time were national polls. This skewed a lot of the results of our analysis, so we have decided to take a look at the state and national polls separately. Let's dive into this analysis a little deeper to see how the pollster grades come into play.

2.2 States and Their Grades/Bar Graph of National Polls' Grades (states/national)

(Bivariate/Univariate 2)

[Both circle graphs for states/grades and national/grades are on the same page, half and half]

Jenny: Jim, in the graphs in front of you, you'll see on the left that it is broken down by state, and on the right it is only the national polls. As in the previous graph, you can see that Florida, Pennsylvania, and North Carolina show a higher number of polls. What we can see with this graph is how the grades of the polls are distributed throughout the states. There are a lot of polls that had either an A- or a B for a grade, consistent across all grades.

Jenny: On the right, we are showing how the grades are broken down for national polls that were administered. What is different with this graph, is that A- and C+ are the two grades with the highest number of polls. This is likely due to nongraded pollsters automatically receiving this grade and not a large number of polls having a proper C+ grade.

Jenny: Jim, what we get from these two graphs is a visual of how the grades are broken down by state and on a national level. We also get to see that A-, B, and C+ are the three grades with the highest number of polls, with each state receiving a relatively even distribution of each grade.

3. Number of Polls by Grade With and Without NCPP/AAPOR/Roper

(Definitive 1)

[Jim and Jenny are the focal points on the screen and they are talking to each other]

Jim: Jenny, could you tell us a little bit more about how the grading is affected by the organizations the pollsters belong to?

[change to the background of the NCPP/AAPOR/Roper visualization]

Jenny: Sure thing Jim! So we looked into how the grades are distributed when a pollster belongs or doesn't belong to NCPP, AAPOR, or Roper Center. As you can see in the visual, we've separated the polls into whether or not the corresponding pollster belongs to one of the organizations.

Jim: I see… and how does this relate to the pollster grades?

Jenny: Well that's the thing, Jim. We've broken up each group of polls based on their grades to see if there was a relationship between the grades and whether or not the pollster belongs to one of those organizations. We then plotted each of the grades based on the number of polls that had a pollster with the corresponding grade. The initial results were fascinating! We mentioned earlier that the two grades with the highest number of polls were an A- and B. Well, it looks like the majority of the A- polls belong to one of the organizations and the majority of the B polls don't belong to any of those organizations.

Jim: So what you're saying is there might be a correlation between higher grades and belonging to the organization?

Jenny: That's right Jim! All of the A and A+ pollsters belong to at least one of the organizations, and all of the C- and D pollsters don't belong to the organizations. Generally speaking, we can say that pollsters who belong to one of the three organizations will tend to have a higher grade than those who do not belong to one of them. It's a really interesting relationship between the two!

[Switch to Jenny and Jim talking side-by-side]

Jim: Good find, Jenny! I never would've thought to look at the organizations that the pollsters belong to.

Jenny: It makes a lot of sense that pollsters who belong to such renowned organizations would have a higher grade… it means they're more accredited and would thus receive a higher grade!

4. Scatter and Rug Chart of the Start and End Dates for the Polls

(Bivariate/Univariate 3)

[Jim and Jenny are next to each other as above]

Jenny: Now that we have looked at how the pollsters can be involved in their received grades, let's take a look at how the dates of the polls may have had an impact on the grades the pollsters received.

[Switch to chart on the screen]

Jenny: Since we wanted to look at how grades and start and end dates interact, we decided to do a scatter and rug chart of the polls. We plotted the start date against the end date and showed the grade of each poll being plotted. What we found was there is a linear relationship between the dates. We dove into this relationship further and found that all of the polls with an A+ took place between July and November of 2016. This could indicate there is a relationship between grade and when the poll was administered. You might even say the higher the grade the closer the poll to election day…

5. Density Plot of the Duration of Polls

(Bivariate/Univariate 4)

[Graph in the main part of the screen with Jenny in the bottom right]

Jenny: Here we have a density plot or rather a visual symphony that unravels the duration of polls during the 2016 US Presidential Election. Each peak and valley represents a different poll's duration, showcasing the rhythm and patterns of the election season.

[Put a circle on the graph on C- at (5,0.135)]

Jenny: For example, this peak has a density of data points that reveals a significant cluster of polls that lasted longer, capturing the attention and curiosity of voters, despite its lower grade of C-.

[Jim pops in next to Jenny]

Jim: That's quite interesting Jenny! Now how do grades of the polls change throughout the different levels of duration?

[Jim exits, just Jenny and the graph]

Jenny: Well Jim, the majority of the polls have a 0-10 day duration, but polls that spanned between 10-20 days often achieved an impressive grade of an A-. This is because these pollsters demonstrated a balance of precision and persistence, hitting the right notes with their predictions. However, if we extend the duration beyond the 30-day mark, the polls tend to achieve a solid B grade, right in the middle.

[Switch to Jim and Jenny talking next to each other]

Jenny: As we explored this density plot further, we discovered the ebb and flow of the election pulse. The graph painted a very vivid picture of how poll durations and grades varied throughout the electoral journey.

6. Number of Polls by Month and Grade (Scandals)

(Definitive 2)

[Keep Jim and Jenny talking to each other]

Jenny: Jim, now I want to switch gears a little bit and focus on what outside factors might have had an impact on the reported polls.

Jim: What kind of factors are you thinking of?

Jenny: I'm talking about external forces that had an impact on how many polls were administered and what might have caused any peaks or troughs.

Jim: Okay... I'm listening.

Jenny: We all know that the candidates have done or had some not-so-good things happen to them in the past. Well, I want to shine a light on the major scandals that arose during the year leading up to election day.

[Switch to the Scandals graph being in the top left corner (about 75% of the page) and Jenny and Jim being in the bottom right corner, still talking]

Jenny: One of the biggest turns of events that came out of the 2016 Presidential election was Clinton's email server scandal. What we looked into was any major scandal that may have had an impact on the polls.

[Switch to just Jenny and the graph]

Jenny: As you can see in the visual beside me, we looked into the number of polls per month where the monthly count is broken down by its distribution of grades. We then graphed the major scandals for each of the candidates on the same timeline below the polls, indicating their start and stop months and whether or not they were resolved within the timeline of our data.

Jenny: You'll see that within the month leading up to the election, there were over 1,300 polls administered and within the first 6 days of November, there were over 700 polls administered - that means that November polls made up over 55% of October polls in only a quarter of the time! If November had continued at the rate it was going, it can be expected that over 3,400 polls would have been administered - that's almost triple the amount in October!

Jenny: Looking at the scandals below, you'll see that there is a high concentration of scandals that came forward within the last two months of the election. This was likely due to the candidates bringing out the worst on their opponents at the last minute to try to skew the results. It's quite interesting to see how the months with the most amount of polls administered also corresponded to the months with the biggest cluster of major scandals.

[Switch back to Jenny and Jim side-by-side]

Jim: That's quite an interesting take on the distribution of polls. Now how does this relate to the grades of the pollsters?

Jenny: That's a good question, Jim.

[Switch back to the graph with Jenny in the bottom right]

Jenny: Between August and November 2016, you can see in the visual that a large portion of the polls had a pollster with either a grade of an A-, B, or C+. Looking just into the A- and B grades as C+ grades had an influx of ungraded pollsters, these two pollster grades had the highest corresponding number of polls administered, with their density increasing in those last few months as well. This could indicate a possible relationship between the grades of the pollsters administering the extra polls and the high concentration of scandals occurring in the last few months of the election.

[Switch to Jenny and Jim side-by-side]

Jim: So when each candidate's campaign published about another candidate's scandals, this had an impact on the polls and how many were administered?

Jenny: There's a possible correlation there, Jim!

7. Predictive and Historical Plus/Minus Score

(Bivariate/Univariate 5)

[Opening Scene: Jenny and Jim are sitting in a meeting room, transitioning from discussing the number of polls by months to the scatter plot.]

Jenny: Now, here we have a scatter plot with an array of colours that represent different grades assigned to the pollsters while also observing the relationship between the predictive plus-minus and historical advanced plus-minus. But here's the twist: the connection between predictive plus-minus and historical advanced plus-minus appears to have a weak linear relationship.

Jim: Interesting! So, the predictive plus-minus score alone can't solely rely on the historical advanced plus-minus to predict the grades accurately.

Jenny: Precisely, Jim! It's a subtle reminder that there are factors beyond historical data that influence pollster performance.

8. Pollster's Predictive Plus/Minus Score From the 2016 US Election

(Definitive 3)

Jenny: And now, behold the grand finale of our data visualization journey—the mesmerizing bubble chart!

[The screen transitions to the final bubble chart, showcasing vibrant bubbles and a timeline of predictive plus/minus scores.]

Jenny: Here's what we discovered:

- Higher-graded pollsters have lower plus/minus scores, highlighting the importance of accuracy.
- Colours in the chart indicate grade categories, revealing a clear relationship between grades and plus/minus scores.
- The most frequent grade is a C+, and the most frequent plus/minus score is 0.77, serving as common benchmarks.
- Pollsters with high frequencies of polls have plus/minus scores centred around 0.45, ranging from -0.4 to 1.3.
- We estimated missing pollsters' data based on their grade's average range.
- Low and high plus/minus scores can be influenced by pollster credibility, past predictions, and target audience.

Jim: Fascinating insights, Jenny! This bubble chart really brings the data to life. By the way, I noticed that the bubbles vary in size. What does that represent?

Jenny: Great observation, Jim! The size of each bubble represents the number of polls administered by the respective pollsters. It gives us an indication of their sample size and overall influence on the dataset.

Jim: Ah, that's interesting. So, we can see not only their performance but also the scale of their polling efforts. Is there anything else we should consider based on this chart?

Jenny: Absolutely, Jim. The bubble chart gives us a comprehensive view of pollster performance, considering their grades, plus/minus scores, and the volume of polls conducted. This helps us evaluate their overall credibility and expertise in predicting election outcomes.

Conclusion

Jenny: And that concludes our data analysis journey! Throughout our exploration of pollster grades, we've unravelled fascinating insights into the variables that impact their performance.

We found an inverse relationship between grades and predictive plus/minus scores, national and state level relationships with a grade, and the influential role of pollster frequency represented by bubble size.

By shedding light on these crucial variables, we gain a clearer understanding of the intricacies behind accurate election predictions.

Thank you, for joining me on this enlightening adventure through the world of data analysis. Stay curious, stay analytical, and keep exploring the world of numbers!