Project 1 - The 2016 U.S. presidential election

Jiayu Tan, Elizabeth Shevchenko, Léo Séror, Daniel Verdon

2016: In the blue corner, the Democrats. In the red corner, the Republicans.

The 2016 U.S. presidential election revolved around Hillary Clinton and Donald Trump, both polarizing figures with strong opinions. This deepened the division between the Democrats and Republicans, which form the dominant political parties in the U.S. electoral system, which bisected the nation while third-party candidates like Gary Johnson and Evan McMullin struggled to gain attention. Notable aspects included Clinton being the first female major party nominee, while Trump ran primarily on anti-establishment policy and was viewed as a political outsider, especially considering his lack of political experience previous to his bid for the Republican nomination.

Many predicted a Clinton victory, citing factors like polling data showing her lead over Trump. For instance, an October 4 *NBC News|SurveyMonkey Weekly Election Tracking Poll* showed that Clinton held a 46 percent to 40 percent lead over Trump.[1] On October 23, the *ABC News poll* showed Clinton with a 12-point lead.[2]

And yet, after the long confrontation, Donald Trump reigned victorious, surprising many. This unpredictability underscores a recurrent theme when it comes to memorable U.S. presidential elections. Often, many pollsters' predictions miss their marks and end up with a totally unexpected result. And so, the question arises: Why were the predictions off and, could the outcome of the 2016 elections have been more accurately forecasted?

The analysis that follows will try to offer some insights into the accuracy of poll ratings, into the importance of battleground states and pollster grading.

Understanding the Electoral College (no, it is not a place).

Unfortunately, simply amassing the most votes will not win a candidate the presidency in the U.S. Instead, candidates need to secure the majority of electoral votes. Each U.S. state is allocated a number of electoral votes based on a census organized every 10 years. For example, as of 2023, California has 54 while Alaska has 3. A candidate who wins the popular vote in a state will typically receive all its electoral

[1] Hannah Hartig, John Lapinski and Stephanie Psyllos, "Poll: Hillary Clinton Holds National Lead Over Donald Trump," NBC news, October 4, 2016, https://www.nbcnews.com/storyline/data-points/poll-hillary-clinton-holds-national-lead-over-donald-trump-n658721.

[2] Eric Bradner, "New poll shows Clinton over Trump by double-digits," CNN politics, October 23, 2016, https://www.cnn.com/2016/10/23/politics/hillary-clinton-donald-trump-presidential-polls/index.html.

votes (with Maine and Nebraska as exceptions)[3]. The candidate who collects 270 or more electoral votes becomes president of the United States.

Candidates have sometimes won the popular vote, but lost the electoral votes, as seen with Al Gore in 2000 and, in our analysis case, with Hillary Clinton in 2016. Given its deviation from more conventional electoral processes worldwide (the direct voting systems), it's given that the Electoral College's unique structure plays a major role in sometimes unexpected polling results. However, is the electoral college system all to blame? Is there another aspect we are not seeing in our story?

Data Dictionary: Polls_us_election_2016

The given data set contains data relating to 4209 polls that were held in the U.S. during the 2016 presidential election. The data set includes the following information associated to each poll:

a) The state in which the poll was held.
b) The start date and end date of the poll (in other words, when polls started and ended).
c) The pollster who was responsible for the poll.
d) The grade of the pollster (a rating given to a pollster based on its historical accuracy, transparency and method).
e) The sample size of the votes, meaning the number of votes a poll ultimately obtained.
f) If the votes of the poll were submitted by registered voters (RV) or likely voters (LV).
g) The percentage of the votes (aka. the sample size) the four candidates (Hillary Clinton, Donald Trump, Gary Johnson and Evan McMullin) earned based on the raw votes.[4]
h) The percentage of the votes (aka. the sample size) the four candidates (Hillary Clinton, Donald Trump, Gary Johnson and Evan McMullin) earned based on the adjusted votes.[5]

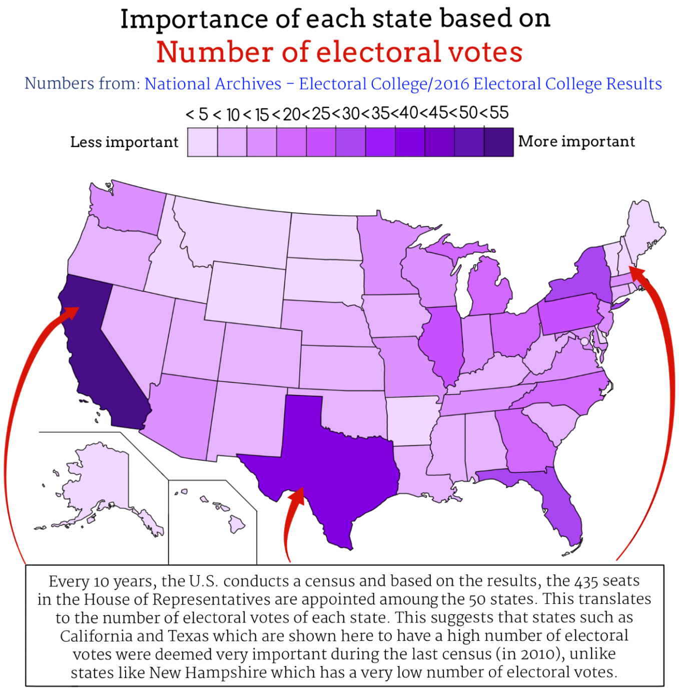
Which states are the most important ? Two ways to see.

In the Electoral College system, it is natural to think that the most important states for the candidates would be the one with the most electoral votes associated with them like Texas or California. The map below provides a visual of which states provide the most amount of electoral college votes.

[3] While all other states use what is called the "Winner-gets-all" system where a candidate who gets the most popular vote in a state wins all of its electoral votes, Maine and Nebraska don't follow that rule. They use the "congressional district method" where the state-wide popular vote winner gets two electoral votes. Then, the winner in each congressional district (2 in Maine, 3 in Nebraska) gets one electoral vote. This means these two states can split their electoral votes among candidates.

[4] The raw votes is the initial, unaltered data collected from voters.

[5] The adjusted votes takes the raw votes, and after considering known biases, demographic mismatches, and other factors, presents a "revised" or "corrected" version which is hoped to be closer to reality.



Importance of each state based on
Number of electoral votes

Numbers from: National Archives – Electoral College/2016 Electoral College Results

Every 10 years, the U.S. conducts a census and based on the results, the 435 seats in the House of Representatives are appointed amoung the 50 states. This translates to the number of electoral votes of each state. This suggests that states such as California and Texas which are shown here to have a high number of electoral votes were deemed very important during the last census (in 2010), unlike states like New Hampshire which has a very low number of electoral votes.

However, perhaps there is another way to allocate state importance to candidates. We wanted to see if there was any correlation between the quality (grade) of pollsters tracking polls and state importance for candidates. The process to visualize this hypothesis is as follows:

1. First, we assigned a corresponding weight to the different grades of pollsters.[6]
2. Then, we computed the importance of each state by the following equation

$$\sum_{i \in Grading} (W_i * \Sigma P_{ij}), \ \forall j \in state$$

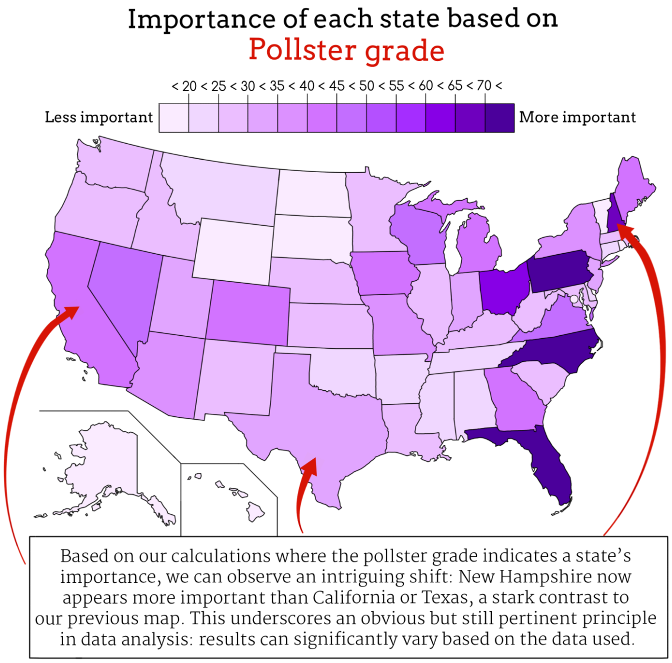
   where W is the weight we assigned to each grade and P is the pollster.
3. The bigger the number we obtain, the more important a state is due to it representing an overall high number of high grade pollsters.

For this calculation, we've combined observations from "Maine CD-1" and "Maine CD-2" into "Maine". And, we've combined "Nebraska CD-1", "Nebraska CD-2" and "Nebraska CD-3" into "Nebraska" for simplicity. We've also entirely removed the observations titled "US" in order to be able to concentrate on the states.

The following map shows the concentration of where higher grade pollsters like the *Washington Post* and *Selzer & Co.* conducted the majority of their polls:

[6]

A+	A	A-	B+	B	B-	C+	C	C-	D	NA
1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0



Based on our calculations where the pollster grade indicates a state's importance, we can observe an intriguing shift: New Hampshire now appears more important than California or Texas, a stark contrast to our previous map. This underscores an obvious but still pertinent principle in data analysis: results can significantly vary based on the data used.

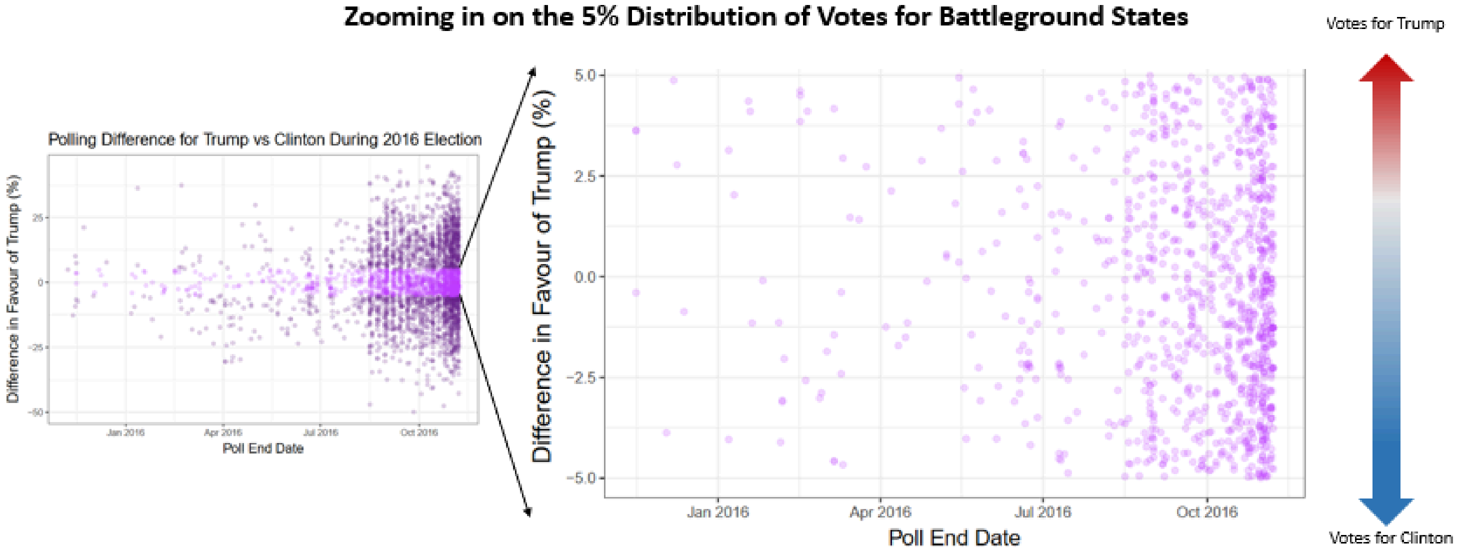
Comparing the two maps, we can tell that there is a discrepancy between high electoral vote importance states and high grade pollster importance states. Florida, North Carolina and Pennsylvania are now the top three most important states according to our pollster grade importance system. The states that are considered very important according to the latter system appear to reflect the battleground states.

What is a Battleground State?

Some states have a history of voting in a consistent manner for the Republicans or the Democrats. Meaning, one candidate consistently has a lead of at least 50% over the other. Hence, states like Wyoming whose electoral votes were gained by the Republican candidates since 1968 are called "red states", while states like Illinois whose electoral votes were gained by the Democratic candidate since 1992 are called "blue states". What this signifies is that over the years, for both a Republican and the Democratic candidate, during a U.S. presidential election, there are some states in which they have a nearly guaranteed victory in the popular votes.

However, there are some states that do not lean towards a given party. These are battleground, or swing states. Defined from a data perspective, a battleground state is those in which candidates are very closely tied, meaning candidates score poll ratings within 5% of each other.[7] Below is a snapshot of the election that excludes polls in which one candidate has a lead of at least 50% over the other. The following visualization gives us a closer look on the center where we can see the clustering of 5% difference votes more clearly.

[7] "What Are Battleground States?," Taegran Goddard's Electoral Vote Map, August 22, 2019, https://electoralvotemap.com/what-are-the-battleground-states/.

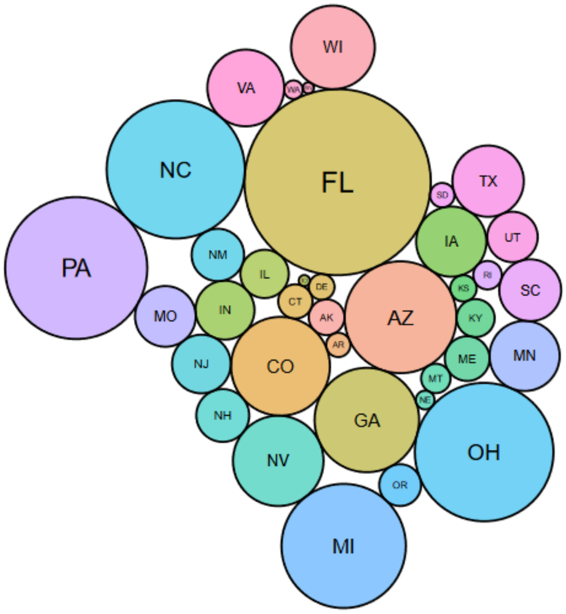


Prior to the Primaries, the Republican party saw a lot of eager votes as shown on the graph to the left by the outliers in late 2015 and early 2016

After Clinton and Trump were revealed to be the candidates at the primaries, the early votes appear to skew in Clinton's favor however, the battleground states vote differences are still incredibly clustered and close.

The final popular vote was taken by Clinton. The concentration in her favor seen in the graph to the right depicts how close the battle was however, the popular vote alone was not enough for her to win the election. Although winning over undecided battleground states may gain popularity votes, candidates need to also focus on the most important states in terms of Electoral College votes.

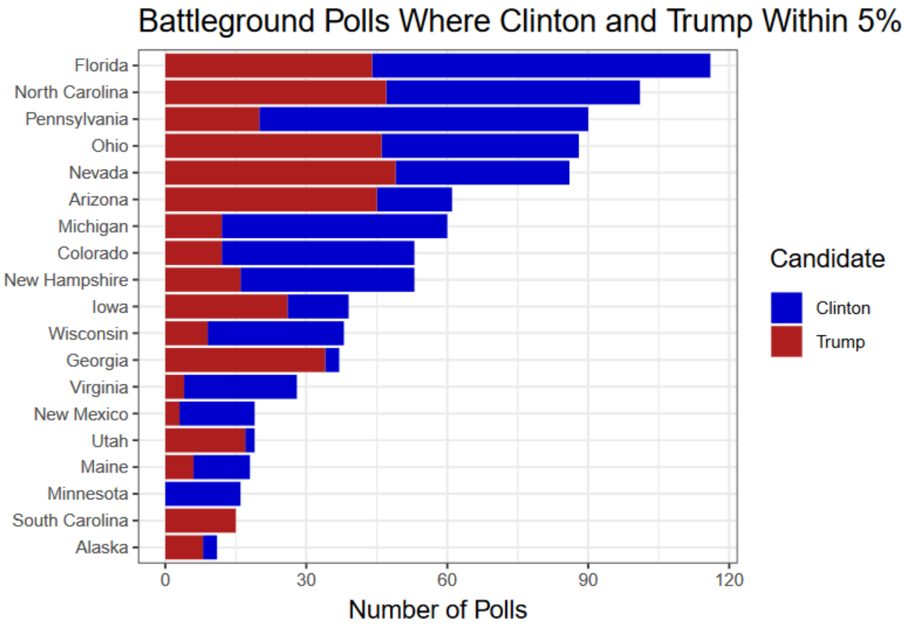
Through the use of electoral college votes and the percentage of polls within a 5% margin we can generate a general framework to identify a state's relative import to the election overall. The chart below, shows the relative importance of a state in winning the election given by the area of its associated bubble (where E is the number of electoral votes, and P_B/P_T is the percentage of polls where Trump and Clinton are within 5% of one another).



$$Area = \pi(W)^2$$

$$W = E\frac{P_B}{P_T}$$

First and foremost, we identify Florida as the most influential battleground state based on its polling data with an area of 23.51, followed by Pennsylvania with an area of 13.68. On the other end of the spectrum, we see Alaska with an area of 0.86, due to its low electoral votes (3) and relatively low percentage P_B/P_T (28.667%). Through the identification of these states, especially when considering past and current polling data, we can characterize the political leanings of a given state, and identify states where resources should be committed, such as the allocation of funds for advertising or rallies, or research into identifying key issues within a given state, and developing policy which would be viewed as favorable among a given state's voter base.



To develop a better understanding of the states depicted in the bubble diagram above, it is beneficial to take a closer look at a state's individual leaning over the course of the 2016 election cycle. First and foremost, it appears that the democrats have the upper hand when compared to republicans based on the number of polls where the difference between their vote percentages favor Clinton over Trump. An example of this advantage can be seen through the polling breakdown in the state of Florida (any percent advantage here is vital when considering its large amount of electoral votes), as well as Pennsylvania, Michigan, and Colorado (through the sheer numbers of polls that favor Clinton over Trump and their relative importance based on their areas in the bubble chart).

Furthermore, this breakdown is paramount when taking party status into consideration; a republican candidate's resources would be much better spent campaigning to secure Ohio's electoral college votes when compared to the states of New Mexico or Virginia. Likewise, a democratic rally in Alaska would most likely be a poor use of resources considering the distance between Alaska and the contiguous United States, as well as the republican lean of poll respondents.

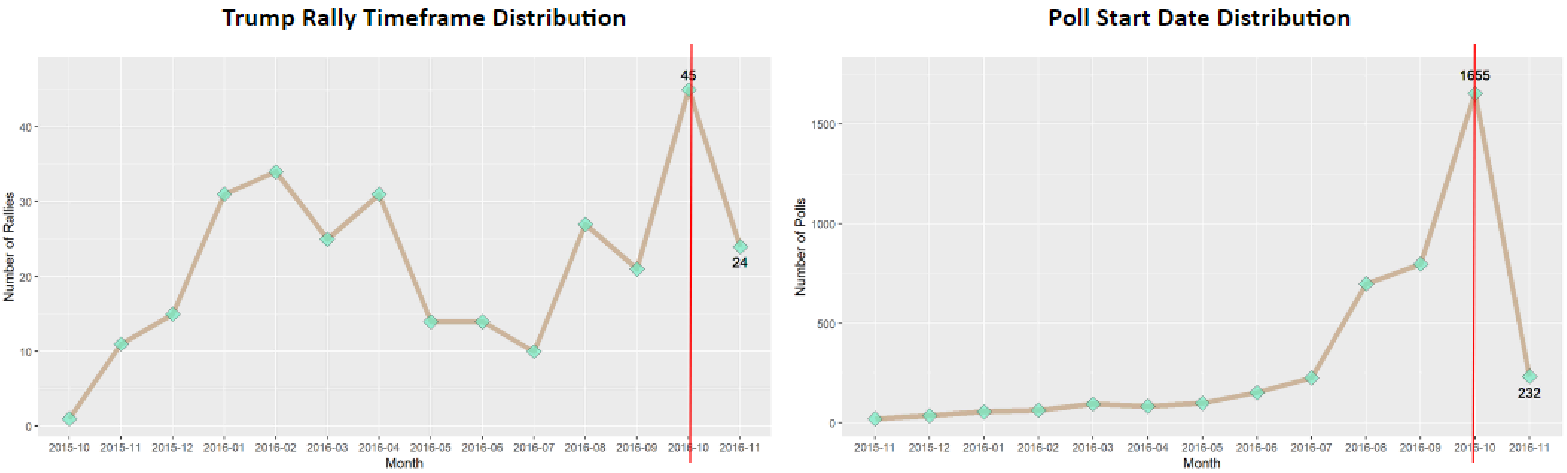
When taking this information in concert with the influence assigned to each given state above, candidates can generate a general path forward towards securing the 270 electoral votes needed to win the presidency.

Playing the cards at the right time.

To support our hypothesis about the importance of the pollster grade and the link it has with battleground states, we can also observe the timing of the rallies done by Donald Trump during the 2016 presidential election.[8] A rally is, put simply, a big presentation done in front of a large audience, sharing opinions and support for a political party with the goal to gain support from the present audience. If timed right, a rally is able to give a candidate a significant advantage over its contestant by empowering audiences to think one way or another right before a poll starts.

Comparison of Trump Rallies and Poll Start During the Election Season

Both graphs peak the month before the U.S. election. Although Trump's rally frequency fluctuated throughout the year, we hypothesize that Trump won the election due to him conducting more rallies in both the highly important rated battleground states as well as the importantly rated electoral college states closer to the elections as opposed to Clinton. A larger time frame would have allowed to dive deeper into an analysis of which states both candidates profited each candidate most.

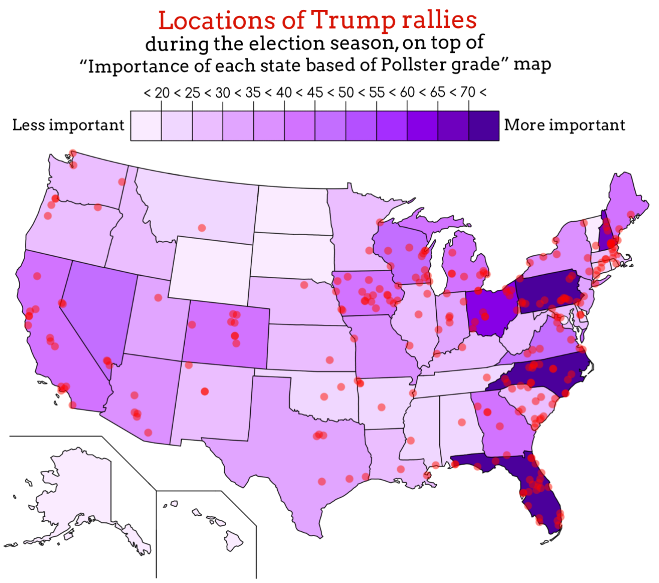


Trump Rally Timeframe Distribution

Poll Start Date Distribution

The graphs above demonstrate how rallies can be powerful tools for candidates particularly in swinging battleground states in their favor. If we add to our map representing the *Importance of each state based on Pollster grade* the locations where Trump did a rally during the 2016 presidential election, we end up with the figure below:

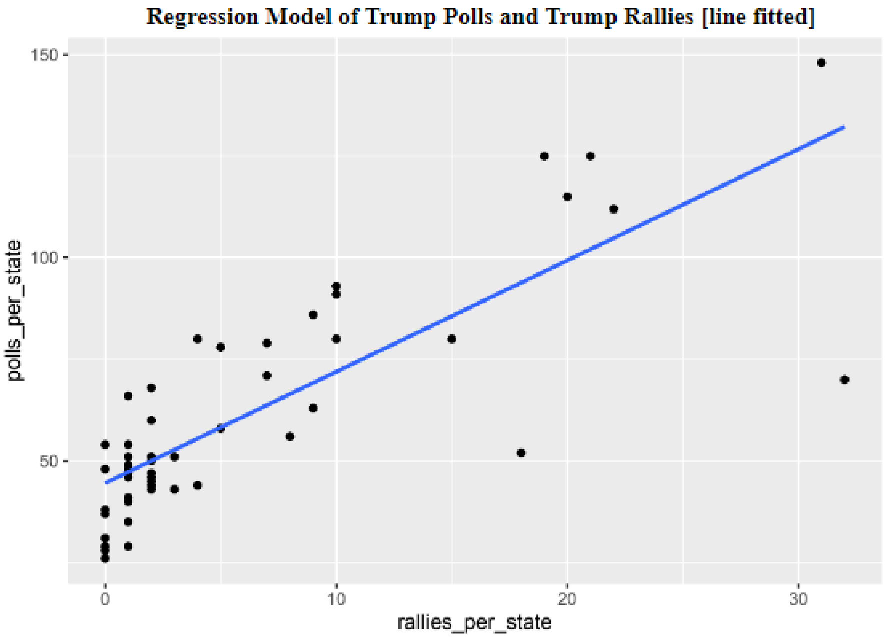
[8] "List of rallies for the 2016 Donald Trump presidential campaign," Wikipedia, August 20, 2023, List of rallies for the 2016 Donald Trump presidential campaign - Wikipedia. -Wikipedia sources checked for data accuracy


Locations of Trump rallies
during the election season, on top of
“Importance of each state based of Pollster grade” map

We observe more red dots, representing Trump rally locations, on the right side of the map compared to the left. This underscores the significance of battleground states or those in which we can find highly graded pollsters, since most of them are located on the right.

It is clearly evident that Trump held a lot more rallies in Florida and New Hampshire (both of which are prominent swing states) than historically blue Michigan or historically red Idaho.

To take the analysis between the importance of states based on pollster grade allocation (keeping in mind that these are also battleground states) in relation to the Trump rally timeline further, we verified the relationship by using regression analysis.


Regression Model of Trump Polls and Trump Rallies [line fitted]

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X	1	23973	23972.6	88.769	1.4e-12
Residuals	49	13233	270.1		

Not only do the dots, representing the states, in the graph show a somewhat linear relationship to the line, but by comparing the Residuals and the P-value in the table above, we notice that the residuals are much bigger than the P-value (49 13233 270.1 > 1.4e-12). This demonstrates a strong correlation between where Trump conducted his rallies and the importance of the states, supporting the idea that the more important pollster grade states are valuable to be rallied in.

To repeat and conclude this section, the importance of the states during a presidential election varies depending on what you base on. In our case, it seems that basing on the pollster grade is very accurate to real life,

Conclusion

When considering the allocation of a candidate's resources, it's vital to consider not only a state's past political climate (to generate a state's general political leanings), but to assess the level of the undecidedness of a state's voter block, which is highlighted through the battleground state analysis in this report. From the analysis of pollster grading in a given state, one can recognize influential states at a glance through pollster interest which has a strong positive correlation to the importance of that state during the election cycle. Furthermore, battleground states once identified can be analyzed further and characterized by where the central tendency lies. In this case, a democratic or republican character in their undecidedness, which allows for more effective strategy planning for either party. As demonstrated in this analysis, it would seem that the number of rallies is a good predictor of the pollster's interest in a given state, which has a strong correlation to the importance of that state.

It is clear that further analysis is necessary to fully answer the questions asked in our introduction. What our analysis was successful in demonstrating was the importance of a given state to the major parties through analysis of the polling data of the 2016 election cycle. The election's outcome is not something that is easy to predict so, of course, the process wouldn't be as well. It's a simple, obvious yet crucial notion that is the key to any data analysis.