Tenimba Coulibaly - 300086545
Ayaan Arora - 300217923
Eunice Konan - 300131563
Ilias Fousi - 300148739
Samantha Sénécal - 300214936

Data Visualization

Nowadays, data analysis plays an essential role in our understanding of the impact of different variables on our lives. Focusing on the context of the US election, our project aims to explore and visualize this polling data to provide meaningful insights into pollster performance, FiveThirtyEight's accuracy, and the factors or variables that contributed to give more precise poll results. This study therefore aims to provide answers to the following questions, based on data collected before the 2016 elections:

- Is there a significant difference / preference for a candidate per region of the US?
- Which polls were better at predicting the elections(date/population/grade)?
- How accurate were the pollsters regarding the prediction of the president of the US?
- How accurate was FiveThirtyEight regarding the prediction of the president of the US?

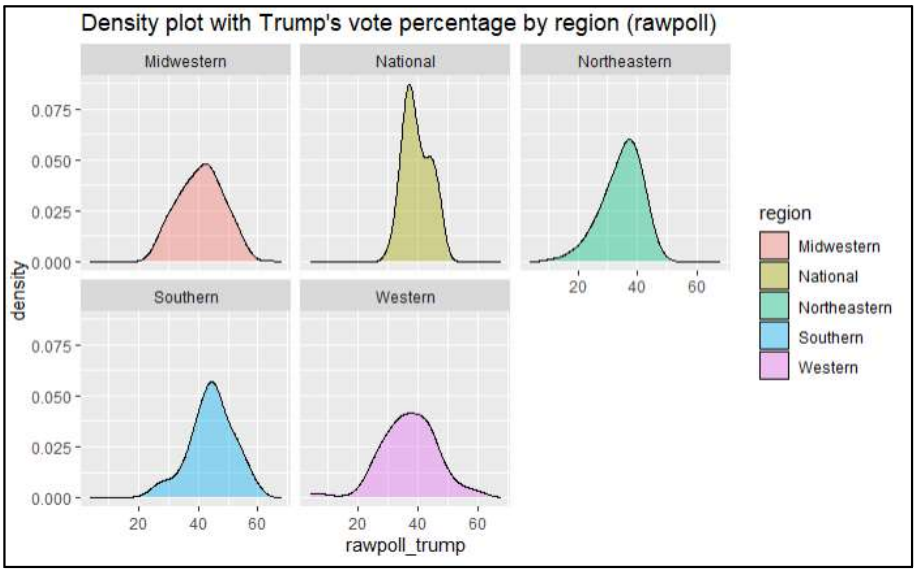
In this project, we will produce data visualizations for the 'polls_us_election_2016.csv' dataset which consists of a total of 15 variables. However, we have only made a data dictionary for the variables that we think are important for our analysis. The variable 'state' tells us in which state the poll was taken. Moreover, there is 'US' as a state variable which describes a national poll. For the sake of analysis, we have decided to give the US polls a national region and to group the states into 4 different regions as follows:

Northeastern	Western	Midwestern	Southern
Connecticut	Alaska	Illinois	Alabama
Maine	Arizona	Indiana	Arkansas
Maine CD-1	California	Iowa	Delaware
Maine CD-2	Colorado	Kansas	Florida
Massachusetts	District of Columbia	Michigan	Georgia
New Hampshire	Hawaii	Minnesota	Kentucky
New Jersey	Idaho	Missouri	Louisiana
New York	Montana	Nebraska	Maryland
Pennsylvania	Nevada	Nebraska CD-1	Mississippi
Rhode Island	New Mexico	Nebraska CD-2	North Carolina
Vermont	Oregon	Nebraska CD-3	Oklahoma
	Utah	North Dakota	South Carolina
	Washington	Ohio	Tennessee
	Wyoming	South Dakota	Texas
		Wisconsin	Virginia
			West Virginia

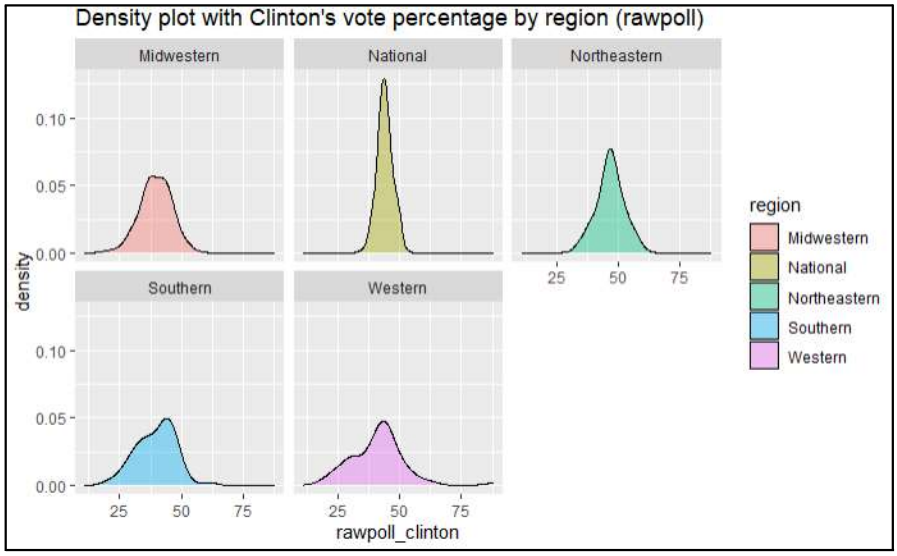
The variables ‘startdate’ and ‘enddate’ show when the poll’s started and ended respectively. For our analysis, we have opted to solely examine the start and end months of the polls conducted in 2016, and have grouped all 2015 polls to see if the ones conducted closer to the US elections date produced more accurate predictions in contrast to the ones conducted farther from said date.
There is also a grade variable which is associated with each poll (A+, A, A-, B+, B, B-, C+, C, C- and D) assigned by FiveThirtyEight.
FiveThirtyEight is an American website focusing on opinion poll analysis, politics, economics, and sports blogging in the United States. They calculate said grade by analyzing the historical accuracy of each polling organization’s polls along with its methodology. To analyze the grade versus poll accuracy, we created a new variable ‘adjgrade’ which groups all the A, B, C and D grades to simplify visualizations.
The variable ‘population’ gives us the type of population being polled and there are four types named ‘a’, ‘v’, ‘rv’, and ‘lv’. Type ‘a’ is adults meaning, United States residents of age 18 and older that are not necessarily eligible or intend on voting, type ‘v’ is voters refers to individuals who are eligible to vote(registered or not), type ‘rv’ is registered voters referring to individuals who have completed the necessary registration process to be eligible to vote, and type ‘lv’ is likely voters which are individuals who are not only registered to vote but also considered likely to participate in an upcoming election.
The remaining variables are the raw polls (rawpoll_x) and adjusted polls (adjpoll_x) for our four candidates (Trump, Clinton, Johnson and McMullin), where raw poll gives us the percentage of votes gathered from a poll and adjusted poll is the FiveThirtyEight’s adjusted poll numbers for a specific candidate.

Before we begin our analysis, you have to know that during the last century, none of the elected presidents of the United States were from another party other than the Republicans or the Democrats.The last time this happened was in 1969 with President Andrew Johnson head of the National Union party.
Nowadays, voting for Republicans or Democrats is like a family business, the parents usually transmit their political thoughts to their children and it’s usually how the tradition is carried on.
That being said, it leads us to our first goal: Try to determine if the rumor stating that there is a tendency to vote for the Republicans or the Democrats based on the region of the United States of America is true.

The following plots are a representation of the average raw poll score for Trump and Clinton in all five regions where the “National” data group is only giving us an idea of the average national vote and whether it is lower or higher than a specific region.



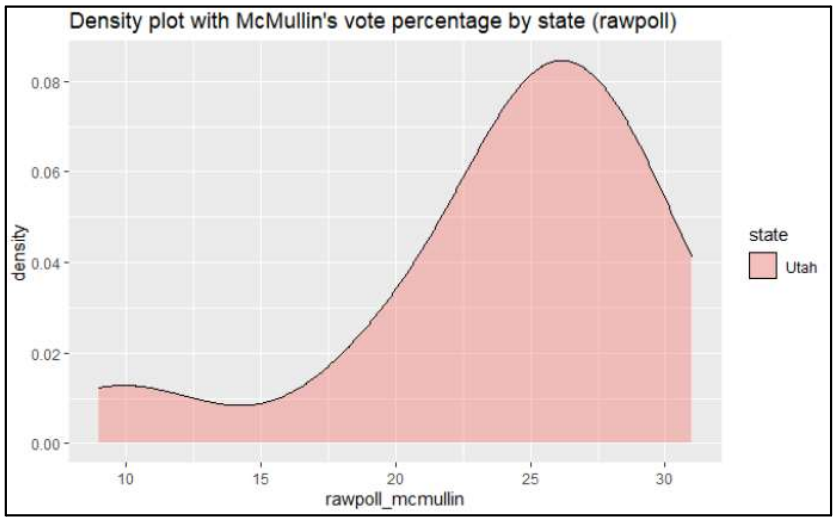
As we can see, there is a tendency for the southern region and the Midwestern region to vote for the Republicans (Trump), since the average score is respectively around 45% and 41%, which is higher than the national average score for trump (Around 36%).



For the Democrats we can see a tendency to vote for Clinton in the Northeastern region with an average of 48% (Higher than the national average).

The average vote for candidate Johnson is too low to actually make a regional difference, so our regional analysis isn't really relevant here.
There is an interesting insight given by the regional plot for McMullin, only the Western region is actually significantly voting for McMullin as we have no information about the other regions.

We discovered that more precisely, people voting for McMullin are from the state of Utah as the following plot shows it.

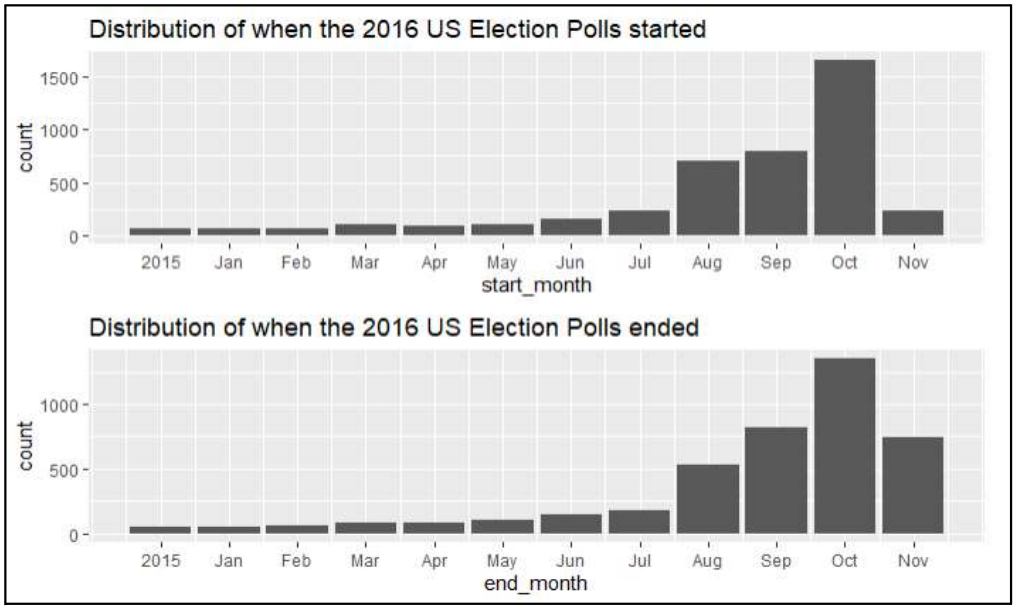


We can explain this tendency to vote for candidate McMullin only in Utah by the fact that it is the state where he was born and because his religion is “Mormon”, and slightly over half of all Utahns are Mormons.
It is important to focus on the results of the polls between the Republicans and the Democrats because of the bipartisanship. Therefore, the candidates that are likely to become president are Trump and Clinton.
You need to know that because of the electoral college, the state of Utah that is usually voting for Republicans is now voting for Mcmullin, which means that Trump will have 6 electoral college votes less if Utah votes for McMullin in majority. However, after looking at the results of the US elections, we found that Trump won the state of Utah and that is not the case.

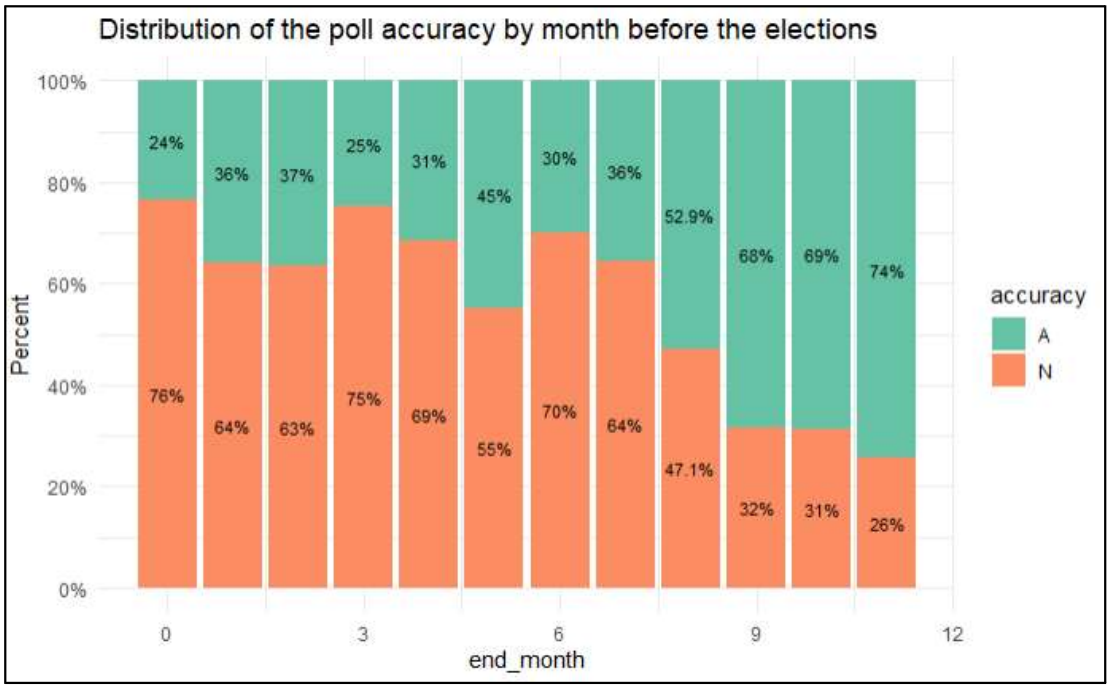
To conclude with our first question, we can confirm that there is a significant difference between the regions regarding their favorite candidate or party, indeed it seems that the Midwestern and Southern Regions residents are more likely to vote for Trump while the Northeastern region residents are more likely to choose Clinton.

Knowing tendencies between the polls is really interesting but you are probably wondering which variables can actually affect the accuracy of the polls that we analyzed.

To begin with, let's focus on the two variables we created, ‘start_month’ and ‘end_month’.
By looking at when each poll started and ended, we were able to create a bar graph by sorting each poll and grouping them by the month in which the poll started and ended in 2016 with all the data coming from 2015 grouped in one and only set.



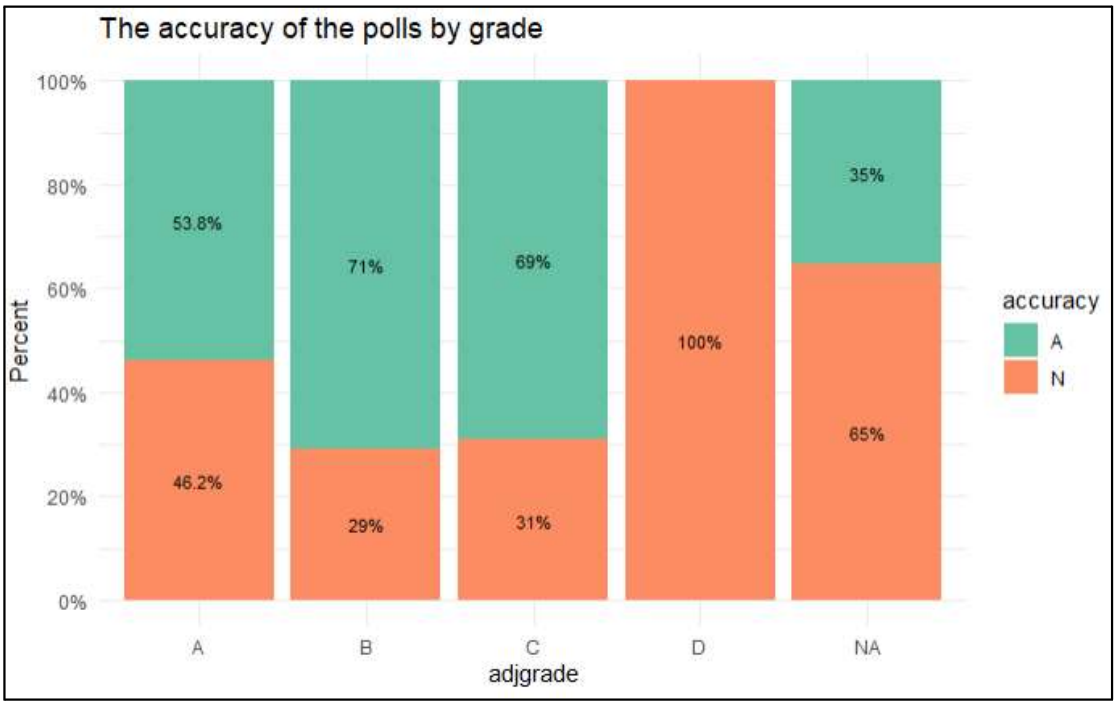
Taking a closer look at these bar graphs, we can see that most of the polls for US elections were held from August to November, with October being the month with most polls held.
We wanted to determine if there was correlation between the date and the accuracy of the poll, creating another column in the dataset to tell if the poll is accurate 'A' or inaccurate 'N':



We can see a clear relation between the end month of the polls and the accuracy, meaning that the closer the poll was to the election, the more accurate it was.
This can be explained by the fact that citizens tend to change their mind and start having a clearer idea of whom they will be voting for closer to the election date.

However there might be another way to check if the polls are accurate, indeed, one of the columns of our dataset called ‘grade’ is able to help us achieve this goal.
You can ask yourself, is the grade actually correlated with the effectiveness of the poll to bring valid information?
That is why we decided to give an accuracy score for each poll, grouping them by their grades.

	adjgrade	n	pct	lbl
1	A	1328	0.315589354	31.56%
2	B	1357	0.322480989	32.25%
3	C	1080	0.256653992	25.67%
4	D	14	0.003326996	0.33%
5	NA	429	0.101948669	10.19%



We can notice that the grade has an impact on the accuracy of the poll although it is not necessary that the poll assigned with a better grade gives us better results. For example, 32.25% of the polls were assigned grade B yet they were more accurate compared to the A grade polls, which made up for 31.56% and only had an accuracy of 53.8%. By looking at the graph it may seem that the polls assigned grade D were 100% inaccurate, however, only 0.33% of the polls were assigned grade D.

	population	n	pct	lbl
1	a	21	0.004990494	0.50%
2	lv	3727	0.885693916	88.57%
3	rv	418	0.099334601	9.93%
4	v	42	0.009980989	1.00%

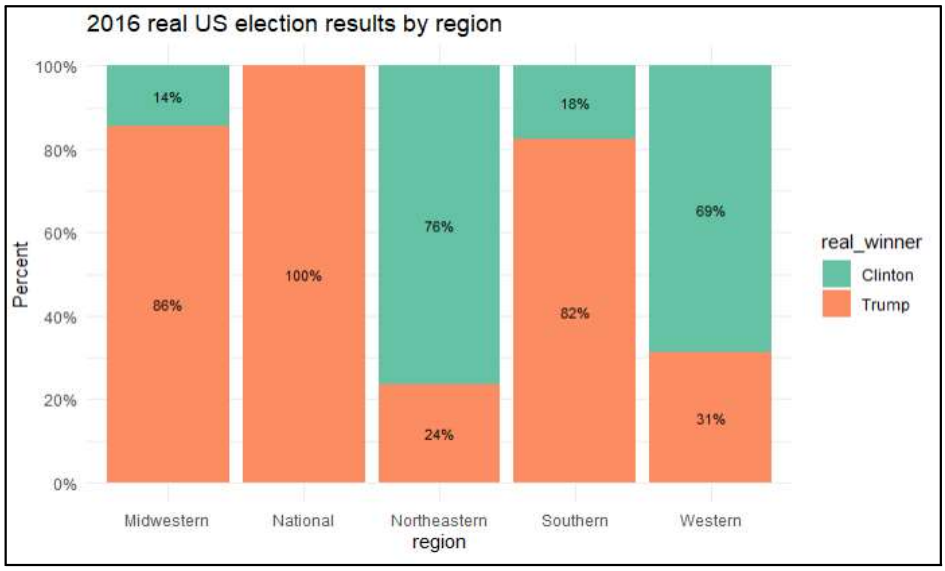
When considering the population, the polls of likely voters should be the best at predicting the results of the presidential election, considering they are deemed the most likely to vote on the election date. Thus, we can see that 88.57% of the polls were for likely voters which will have a positive impact on the accuracy of the poll winners.

Furthermore, creating a column giving the accuracy of the polls was the climax of the analysis because it opened multiple gates to explore.
We actually grouped the states where Clinton won and the states where Trump won giving us a winner per state and we compared the polls information with the actual results of the elections.

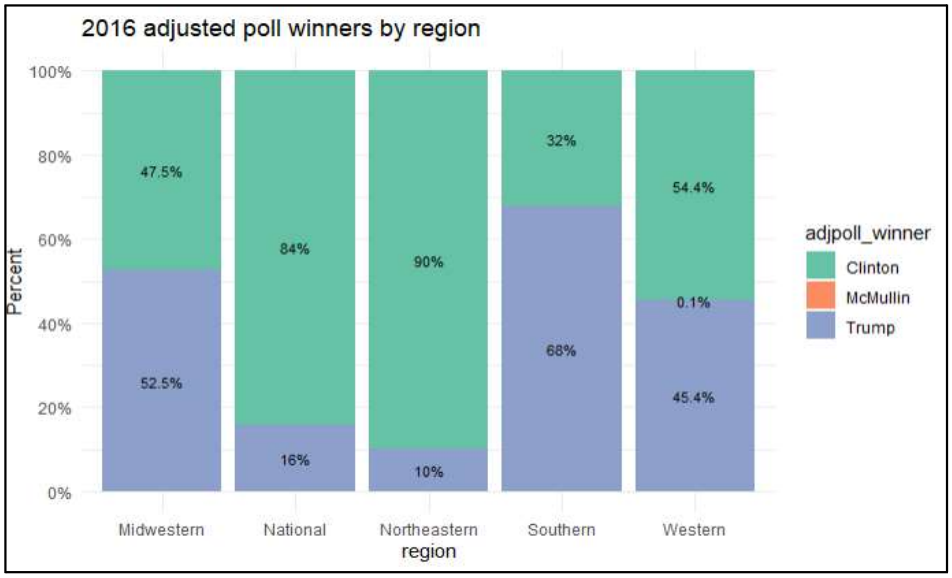
States (Districts) where Clinton won	Connecticut ,Maine, Maine CD-1,Massachusetts, New Hampshire, New Jersey, New York, Rhode Island, Vermont , California, Colorado, Hawaii, Nevada , New Mexico, Oregon, Utah, Washington, Illinois, Minnesota, Delaware, Maryland, District of Columbia, Virginia
States (Districts) where Trump won	Maine CD-2 , Pennsylvania, Alaska, Arizona, Idaho, Montana, Indiana, Iowa, Kansas, Michigan, Wyoming, Missouri, Nebraska, Nebraska CD-1, Nebraska CD-2, Nebraska CD-3, North Dakota, Ohio, South Dakota, Wisconsin, Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, West Virginia

You can notice that Maine appears in both Clinton's and Trump's case, this is because this state separates its vote into 2 distincts congressional districts that voted in majority for two different candidates.
The winner of the US elections is determined by his electoral college votes, each state has a predetermined number of electoral college votes, if a candidate is the "winner" of a state, it gives all its electoral college votes to him, but two states are exceptions : Maine and Nebraska, each congressional district has a number of electoral college votes in these states.

Now that we know the winner of each state or CD's, we can try to answer the third question which was : How accurate were the pollsters regarding the prediction of the president of the US?
We kept the idea of grouping the states per region, but first of all let's take a look at the distribution of the real winners per region.



Donald Trump won the election and was the favorite in the Midwestern and Southern region.
Now let's take a look at the adjusted polls distribution and try to tell if the polls agrees with the real results:



Considering the polls, Clinton seems to be the favorite and is receiving 84% of the votes for the polls that don't specify the state. It doesn't mean that the polls are wrong because Clinton did in

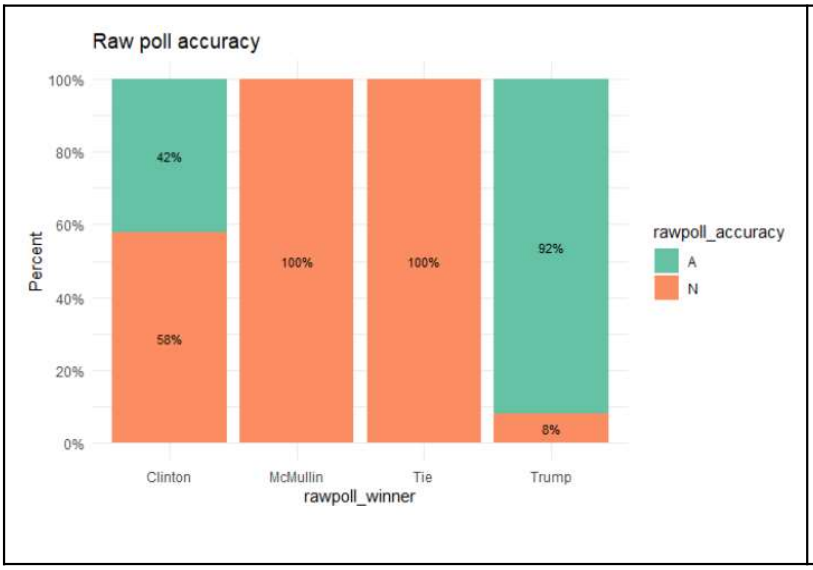
fact win the popular vote. Clinton, however, did not become President of the United States in 2016 as the system is slightly different with the electoral college as explained earlier.
We still notice a great difference between the polls results and the actual results implying that the prediction of the next president of the US elections remains uncertain. People either changed their mind or the polls were not studying a representative sample. Now let's jump right in our last question : How accurate was FiveThirtyEight regarding the prediction of the president of the US?

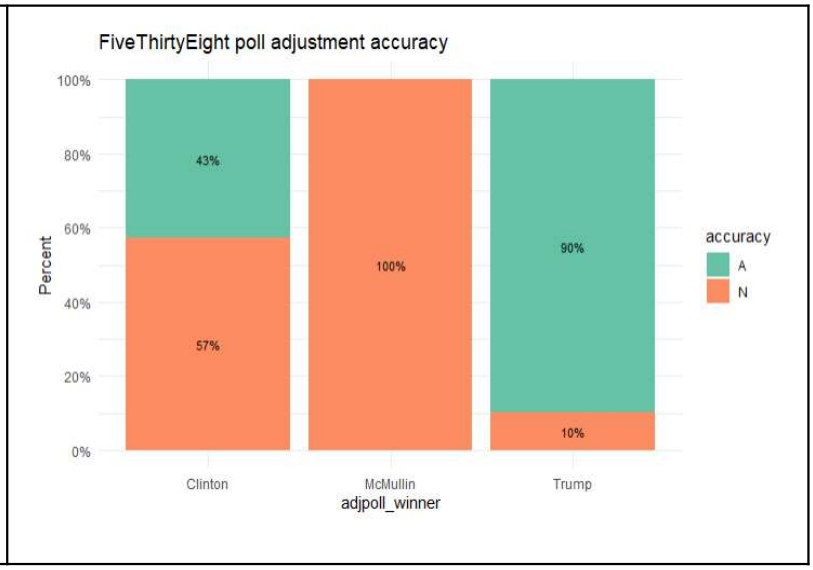
FiveThirtyEight has been adjusting the US elections polls by slightly changing the results creating a new column in our dataset 'adjpoll'. There is no major change in the results but this adjustment can actually change the candidate having the majority in each state.

poll_winner	adjpoll_winner	real_winner	accuracy
Tie	Trump	Trump	A
Tie	Trump	Clinton	N
Tie	Trump	Trump	A
Tie	Clinton	Trump	N
Tie	Clinton	Clinton	A

We have multiple examples of rawpolls not providing a winner per state because of the possibility of getting a tie, but if we look at the adjusted version of these polls, they usually assign a winner to one or the other without having the possibility of the same score for two candidates.

Now let's compare the accuracy of the adjusted polls with the raw polls:





As you can notice from the two visualizations, the rawpoll winner is often different from the adjpoll winner. We do not know how FiveThirtyEight adjusts the percentages but it is an important insight that it is able to have an impact on the accuracy of the polls.
Even if the adjusted polls were predicting 43% of the time that Clinton would be winning correctly, while 42% for the raw polls, the analysis for Trump is slightly different.
The adjusted polls got 90% of the winner prediction correct while the raw polls are having a higher performance with 92% of accuracy regarding the prediction of Trump being the winner.
It is important to note that these differences observed between the raw poll and the adjusted polls are not significant but we learned that the adjustment is supposed to bring clearer and unbiased data even if it is claimed to be inaccurate at times.

To conclude, we discovered through our analysis that it would indeed seem that the Midwestern and Southern Regions residents are more likely to vote for Trump while the Northeastern region residents are more likely to choose Clinton and it confirmed the popular thought that people are voting for different parties depending on their geographic location in the US.

We've also seen that the date, population and grade of a poll can drastically change the accuracy of the U.S. president prediction. Visualizations lead us to surprising results such as the fact that the grade of a poll is not in cohesion with the accuracy. As for the date, we found that the closer we are to the elections, the less the results are likely to differ. This is mostly due to the fact that voters gain newer information closer to the elections and develop strong opinions on the candidates. Polls with likely voters are the best way to predict the outcome of the election since they contain less outliers (people who are not eligible/do not intend to vote).

In addition, the polls indeed predicted that certain candidates would win certain regions. However, we discovered that there is a great difference between the polls results and the actual results implying that the prediction of the next president of the US elections remains uncertain. Each state has a different weight when it comes to elections so even if a region has the majority voting for Clinton, it doesn't mean that that region will have a big impact. Hence, the national polls did not do a good job at predicting the election outcome.

Moreover, the FiveThirtyEight adjusted polls did not show significant differences with the raw polls as their impact on the percentages were minimal and led to similar outcomes. Also, we do not know how FiveThirtyEight adjusts the percentages and assigns a winner when there is a tie for a poll winner. It would be interesting to know how the process of adjusting the polls is conducted.

Finally, the important insight about all this analysis is to keep in mind that data analysis is more than just looking at numbers, as we have multiple variables that are able to give a totally different aspect to the questions we tried to answer with the data. Therefore, it would be wise to ask the following question : How biased are we when it comes to evaluating the veracity of information ?