Data Analytic Report: Presidential Polls

Leslie Liu
PSTAT 120C, Spring 2023

Introduction

The 2016 Presidential election was a pivotal moment in U.S. history, and exposed deep differences between voters' views in the states and profound changes occurring within American society. The two main candidates were Democrat Hillary Clinton, the first woman to win a presidential nomination, and Republican Donald Trump, the first U.S. President with no prior military or government experience. The candidates were highly polarizing, with liberal states like California and conservative states like West Virginia receiving a vast majority of votes in one direction. However, certain swing states–states which had an almost equal split of voters in each political party–were crucial in determining the outcome of the election.

The competitiveness of the election came down to the candidates vying for electoral votes which were needed to secure victory, leading to a series of closely contested battleground states. The margins of victory in several key states were incredibly tight. Swing states such as Pennsylvania, Michigan, and Wisconsin became focal points. In fact, the election was decided by a few critical states, with razor-thin margins in some instances. Another factor was that polls were not always completely accurate, so there were discrepancies between polls and the actual votes cast. This tight race demonstrated the high stakes involved and highlighted the immense impact that even a small number of votes could have on determining the future course of the nation.

Section 1.

Michigan, Georgia and North Carolina emerged as three pivotal battleground states during the 2016 presidential election. Using R to load in a dataset called *polls_data_2016* which includes data from 12,621 polls, I filtered this information into three different datasets to conduct analysis. Looking specifically at the dates August 1 to November 2, the period right before the election, I found that:

1. Clinton had received 66,664 votes for 51.72% of Michigan followed by Trump with 62,240 votes (48.28%). The percentage difference favors Clinton by 3.43%.
2. In Georgia, Trump actually led with 53.22% and Clinton had 46.78%. The percentage difference favors Trump by 6.44%.
3. North Carolina had a tighter margin, with Trump at 50.09% and Clinton trailing closely at 49.91%. The percentage difference is in favor of Trump by 0.18%.

These percentages may seem very close; however, given the vast sample size of votes, they may indicate a statistically significant preference for a certain candidate. In this case, a paired t-test would be appropriate to test if there is a significant difference between the sample means of the two voter groups in each state. We don't know what the true population mean or variance is, but since the votes are in the ten thousands, we can assume normality. I used the function *t.test* and input the total number of votes for Clinton and total number of votes for Trump for each state. I tested the alternative hypothesis that the true mean number of votes for each candidate was different, against the null hypothesis that the true means were the same.

1. Michigan's t-statistic was -10.36 on 170 degrees of freedom with a p-value of <-2e-16. There is significant evidence to show that the true mean number of votes is different for the two candidates in Michigan.

```
data:  michigan_2016$total.trump and
michigan_2016$total.clinton
t = -10.36, df = 170, p-value < 2.2e-16
```

2. Georgia's t-statistic was 19.242 on 167 degrees of freedom with a p-value of <-2e-16. There is significant evidence to show that the true mean number of votes differs in Georgia.

```
data:  georgia_2016$total.trump and
georgia_2016$total.clinton
t = 19.242, df = 167, p-value < 2.2e-16
```

3. North Carolina's t-statistic was 0.64 on 275 degrees of freedom with a p-value of 0.2612, which is greater than the significance level of 0.05. There is not enough evidence to reject the null hypothesis.

```
data:  nc_2016$total.trump and nc_2016$total.clinton
t = 0.64049, df = 275, p-value = 0.2612
```

The t-test revealed that there was evidence to show that Michigan was in favor of Clinton and Georgia was in favor of Trump. The t-test is most robust with large sample sizes, which we have, and a normality assumption. However, one other requirement of

the t-test is that the samples are independent, which these are not. The votes highly depend on each other, as one vote for a candidate directly takes away a vote for the other. In addition, missing data highly skews the results for a paired t-test, and we must assume that the samples have equal variance. Given these caveats, the results of the t-test are not absolute.

However, we can further test the data using the Wilcoxon Signed-Rank test for dependent samples, reflected in R by *wilcox.test(.....alternative = "two.sided", paired=TRUE)* which tests the null hypothesis that the median difference between paired values in the polls is 0. Here is what I found from the tests:

1. Michigan `data:`
   `polls_data_2016$total.clinton[index_michigan] and`
   `polls_data_2016$total.trump[index_michigan]`

```
V = 30898, p-value < 2.2e-16
```

2. Georgia `data:`
   `polls_data_2016$total.clinton[index_georgia] and`
   `polls_data_2016$total.trump[index_georgia]`

```
V = 559, p-value < 2.2e-16
```

3. North Carolina `data:`
   `polls_data_2016$total.clinton[index_nc] and`
   `polls_data_2016$total.trump[index_nc]`

```
V = 36135, p-value = 0.6737
```

The Wilcoxon-Signed Rank test provided consistent results with the paired t-test: that is, Michigan was in favor of Clinton, Georgia was in favor of Trump and North Carolina did not have a statistically significant difference. The test is nonparametric but assumes that the differences between samples are symmetrically distributed. It is also sensitive to outliers. However, it is important to note that even given these considerations, we obtained results that aligned with those of the paired t-test.

These results are strong; however, it would be helpful to visualize the differences between votes for each state using linear models. I used the *lm* function in R to plot the percentage differences for each poll in the state on the y-axis for the same dates, August 1, 2016 to November 2, 2016 on the x-axis, along with fitted values and a confidence interval for the regression line.

Figure 1.1. A plot of the percentage difference between poll votes for Clinton and Trump in Michigan from August 1 to November 2. We observe a slight downward trend; however, the regression line remains above 0, meaning that there is still a positive favor of Clinton.

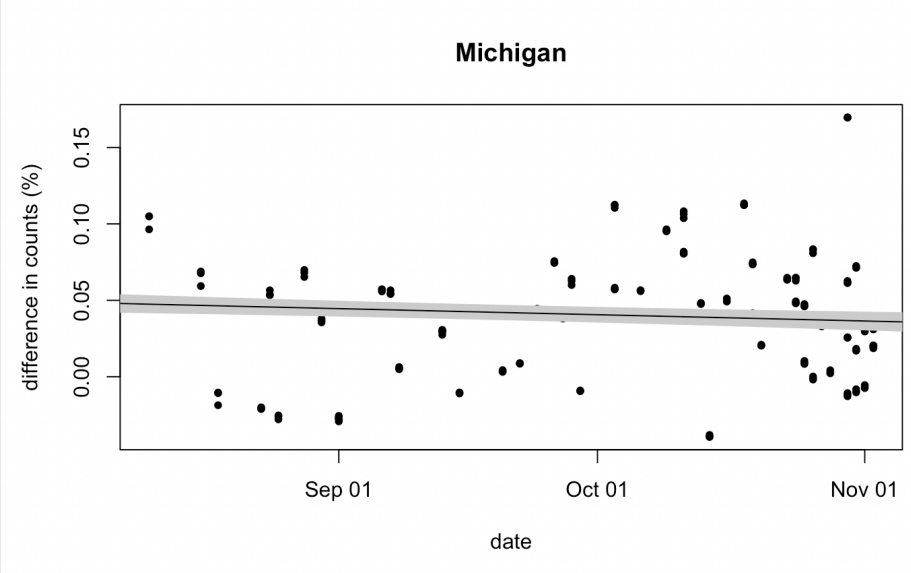


Figure 1.2. A plot of the percentage difference between votes in Georgia. As you can see, the regression line approaches 0, meaning that the votes for Clinton and Trump get closer.

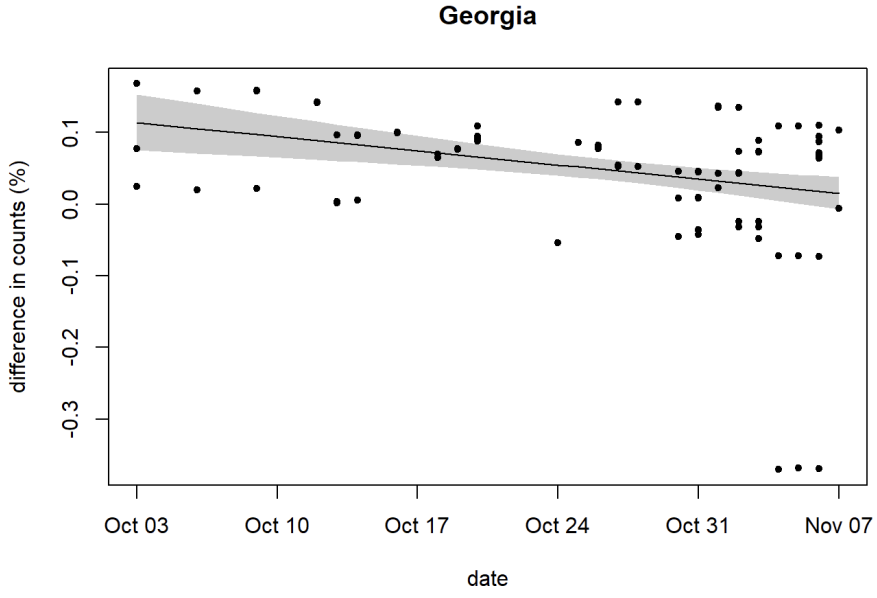
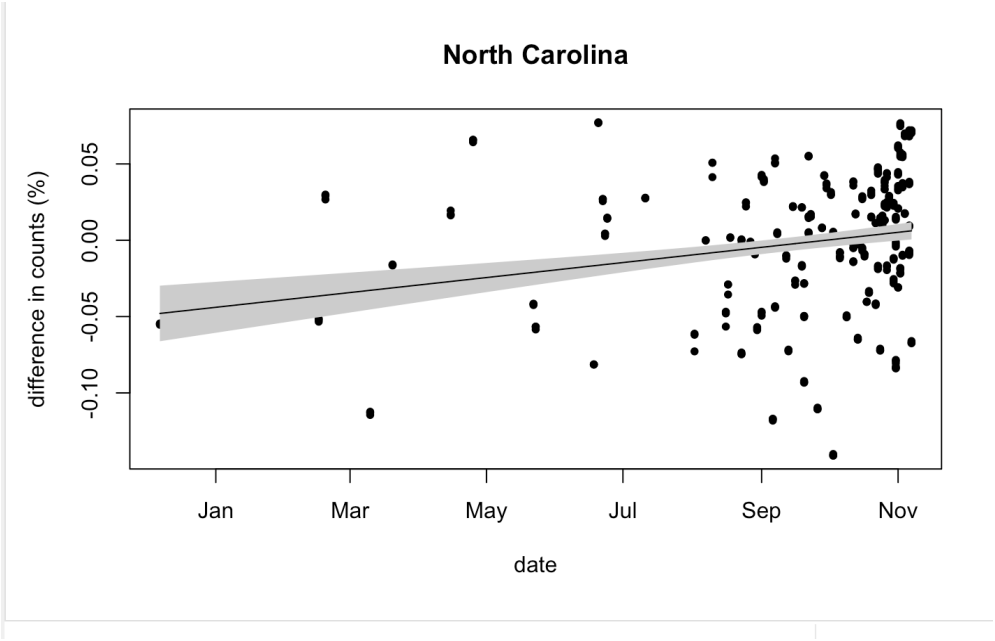


Figure 1.3. A plot of the percentage difference of votes between Clinton and Trump in North Carolina.The line approaches 0, but has a positive slope, meaning that Trump

received more votes as the election deadline approached, but the margin was getting closer.



As aforementioned, the data from polls does not always accurately represent electoral outcomes. Taking a look at Wikipedia's results for the actual voting outcomes, Trump won Michigan narrowly by 0.22%. Georgia went in favor of Trump over Clinton by 5.16%, and North Carolina also in favor of Trump by 3.66%. These actual results agree with our tests of the polls; however, the margins are slightly different. This divergence between polling data and the final results can be attributed to various factors. One factor is the presence of shy voters, individuals who may hesitate to reveal their true voting intentions due to social desirability bias or concerns about backlash. Another factor is the difficulty in capturing the sentiment of certain demographic groups, particularly in states where the composition of the electorate is rapidly changing. Additionally, the role of undecided voters and late-deciders can significantly impact the accuracy of polls.

Despite these challenges, it is important to note that polls still provide valuable insights and serve as a tool for understanding public sentiment. While they may not always perfectly mirror the final election outcomes, they contribute to the broader understanding of voter preferences and can help shape campaign strategies and public discourse. The 2016 election served as a reminder of the inherent complexities involved in polling and highlighted the need for a comprehensive approach that takes into account the limitations and potential sources of error.

Section 2.

Given the closeness of the 2016 election, it was crucial to more closely monitor the polling data and how it could either reflect or deviate from the actual voting results in the next election in 2020. The 2020 election also mirrored incredibly polarizing results, with Republican Candidate Donald Trump seeking reelection, and former Vice President Joe Biden running for the Democratic Party. Provided with a dataset *president_polls_2020,* I conducted similar analysis to the 2016 polls in R for each candidate for the states of Michigan, Georgia and North Carolina.

Extracting the total number of votes for Biden and Trump for each of the three states, I calculated the percentage difference again for the 2020 election. In Michigan, I found that Biden led by 7.93%. Biden also led in Georgia by 2.04%, and in North Carolina by 3.25%.

To further test the data, I once again conducted a two-sample t-test with the 2020 data for these states, under the null hypothesis that there was no difference in the means of votes within the states.

1. Michigan's t-test was significant with a p-value of <2e-16. Reject the null hypothesis; there is statistically significant evidence to show that the mean number of votes is not the same for each candidate. `data: michigan_2020$Biden and michigan_2020$Trump`

```
t = 27.305, df = 144, p-value < 2.2e-16
```

2. Georgia's t-test was significant with a p-value of 1.08e-8. Reject the null hypothesis; there is statistically significant evidence to show that the mean number of votes is not the same for each candidate. `data: georgia_2020$Biden and georgia_2020$Trump`

```
t = 6.2306, df = 101, p-value = 1.08e-08
```

3. North Carolina's t-test revealed a similar result with a p-value of <2e-16. Reject the null hypothesis. `data: nc_2020$Biden and nc_2020$Trump`
   `t = 12.881, df = 154, p-value < 2.2e-16`

I then ran a Wilcoxon Signed-Rank test to see if our results align with that of the t-test. The process is the same as the one for the 2016 election, using *wilcox.test(...alternative= "two.sided", paired = TRUE).* I tested the null hypothesis that there is no difference between the median number of votes for Biden and Trump.

1. Michigan's Wilcoxon Signed-Rank test output a statistically significant p-value. Reject the null hypothesis. `data:  michigan_2020$Biden and michigan_2020$Trump`
   `V = 10530, p-value < 2.2e-16`
2. Georgia's Wilcoxon Signed-Rank test output a statistically significant p-value. Reject the null hypothesis. `data:  georgia_2020$Biden and georgia_2020$Trump`
   `V = 3230.5, p-value = 1.129e-07`
3. North Carolina's Wilcoxon Signed-Rank test output a statistically significant p-value. Reject the null hypothesis. `data:  nc_2020$Biden and nc_2020$Trump`
   `V = 10021, p-value < 2.2e-16`

The results from the Wilcoxon Signed-Rank test were consistent with the paired t-test results. However, these tests are not fully robust given the caveats discussed earlier. To visualize the percentage differences between each candidate for the state, I also created linear models with a confidence interval and fitted residuals.

Figure 2.1 shows the percentage difference between Biden and Trump In Michigan during the 2020 election between August 2 and November 2. The regression line appears to have a very slight positive slope, meaning that the percentage difference is increasing; however, it is important to note that the regression line stays above y=0 meaning that Biden maintains the lead.

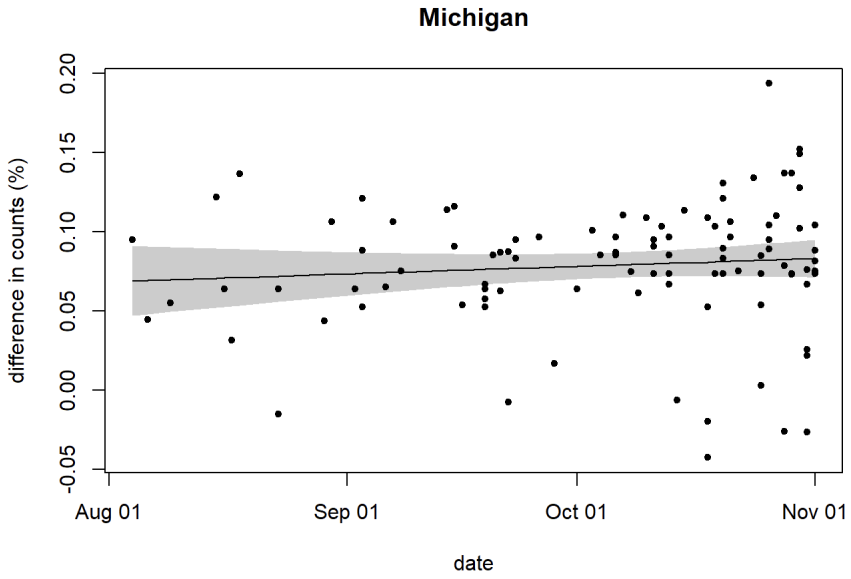


Figure 2.2 is a plot of the percentage difference between Biden and Trump in 2020. The regression line has a slight positive slope. The y-values begin in the negative and actually cross y=0 into the positive by November, meaning that the votes may have begun in favor of Trump but shifted in favor of Biden.

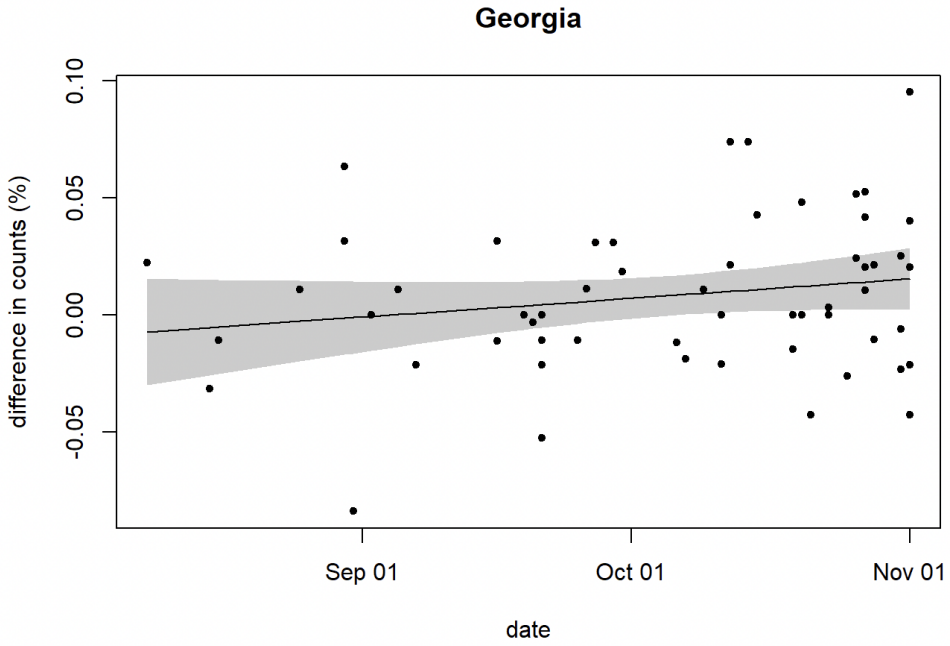
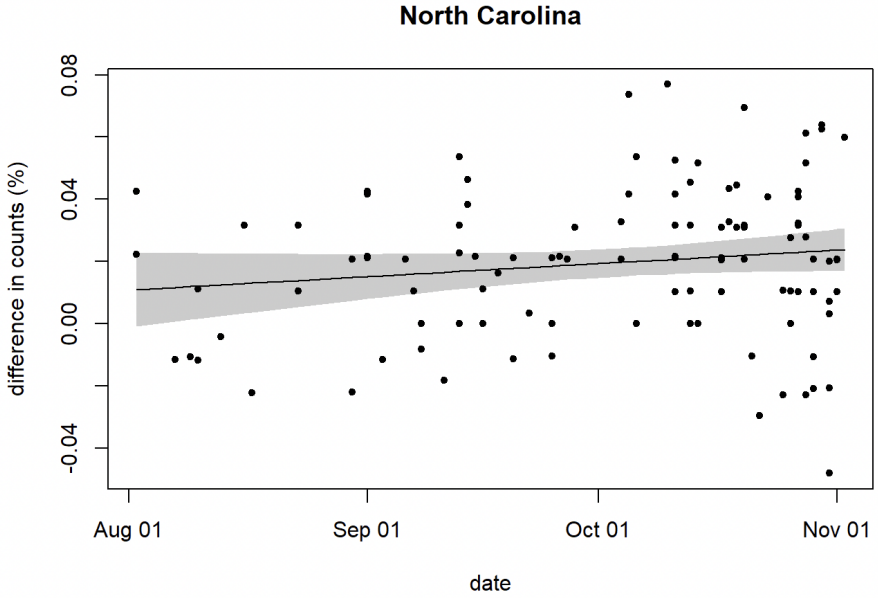


Figure 2.3 depicts the percentage difference in votes in North Carolina. As the regression line remains above y=0, the polls stay in favor of Biden and increase in favor of Biden by the election.



Interestingly, the plots suggested that the closest race was in Georgia, given that the favor shifted from Trump to Biden over time as the line crossed y=0 for a percentage difference. Looking at the actual results of the votes in 2020, Biden won Michigan at 50.62%, and Trump received 47.84%, with a percentage difference of 2.78%. In Georgia, Biden received 49.47% of the votes, while Trump received 49.24% of the votes, so Biden narrowly won by 0.23%. Finally, in North Carolina, Trump won with 49.9% of votes and Biden ended with 48.6% of votes for a difference of 1.3%. Clearly, Georgia had the closest race.

The graphs and tests predicted the outcome of Georgia's and Michigan's votes correctly; however, they were wrong about North Carolina. The polls seemed to suggest that the voters were more in favor of Biden, but Trump ended up winning the state's vote. Polling errors can occur due to the "shy Trump voter" phenomenon, where some Trump supporters may have been hesitant to reveal their true voting intentions to pollsters. This phenomenon can introduce bias and affect the accuracy of the polls, potentially leading to a discrepancy between projected results and actual outcomes.

Section 3.

How can we analyze the percentage differences between the states from 2016 to 2020? I set up a map in R that filled in the percentage differences per state on a gradient scale. On the higher end of the scale, I chose blue to represent the Democratic Party, and on the negative end of the scale I chose red to represent the Republican party.

Figure 3.1 The percentage difference between Trump and Clinton for 2016.

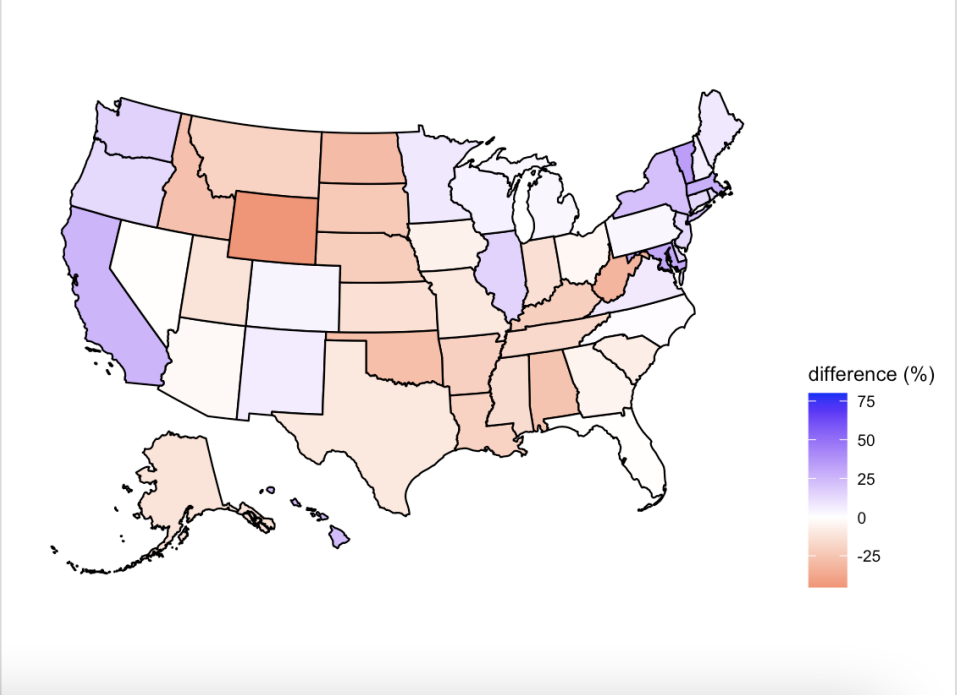
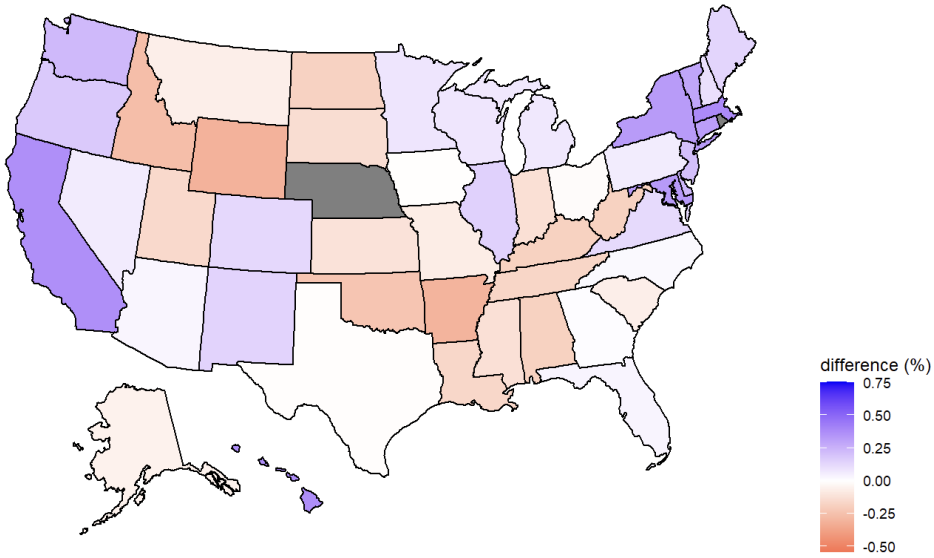
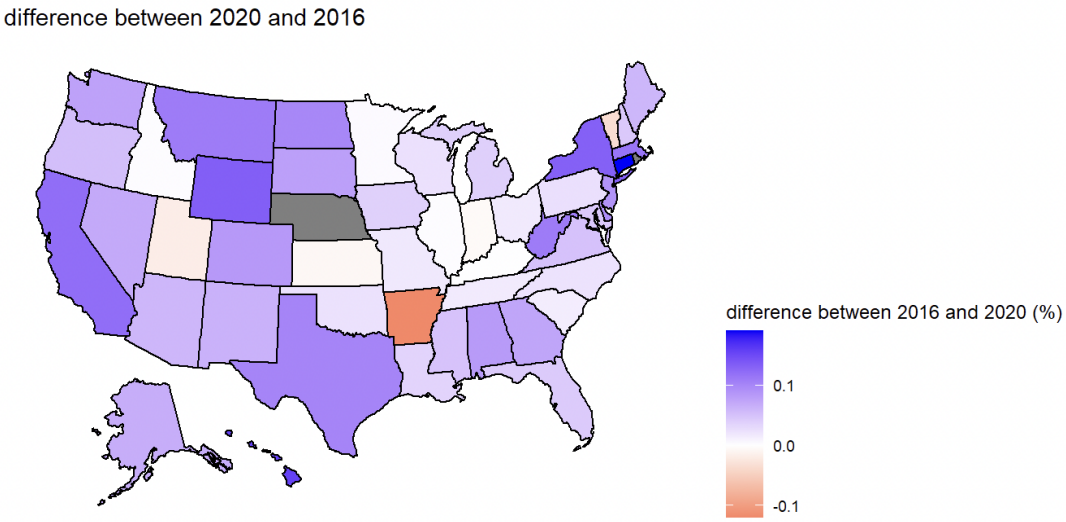


Figure 3.2 The percentage difference between Trump and Biden for 2020.



3.3 A plot showing the difference between 2020 and 2016 voting.



In general, from the graphs, it seems that the votes have shifted more democratically, given that the red states have transitioned to a lighter color on the scale. In particular, I noticed that Wyoming's vote split turned less conservative, and states such as California and Maine turned more liberal. This may be attributed to Joe Biden being a more moderate candidate than Hillary Clinton, who was also a woman, which could have secured more moderate votes from conservative states. In addition, in the 2020 election, Georgia, which has been historically considered a Republican stronghold, was won by the Democratic candidate Joe Biden for the first time since 1992. In addition, Wisconsin, Michigan and Pennsylvania, which were crucial blue states that Trump won in 2016, swung back in favor of the Democratic Party in 2020.

By arranging the percentage differences in 2020 by order from least to greatest, I was able to extract the 10 battleground states which had the closest margins.

```
##             state diff_percentage
## 1            Iowa    -0.005852330
## 2           Texas    -0.008706351
## 3         Georgia     0.009631098
## 4            Ohio    -0.014481256
## 5      Maine CD-2     0.017285552
## 6  North Carolina     0.020675716
## 7         Arizona     0.031183705
## 8         Florida     0.034448114
```

```
## 9          Alaska    -0.051229576
## 10  Nebraska CD-2     0.056377208
```

Based on these numbers, Iowa, Texas and Georgia had the closest elections with less than 1% of a difference between candidates. In 2020, Iowa had a close margin due to a combination of factors. The state has a predominantly rural and agricultural economy, and issues related to trade, agriculture policies, and rural development were of particular importance to Iowa voters. Texas, a traditionally Republican-leaning state, experienced a narrowing margin due to shifting demographics, including a growing population of younger and more diverse voters. Additionally, issues such as immigration, healthcare, and voter mobilization efforts played a role. Georgia, a reliably republican state, underwent a Democratic shift. Changing demographics, particularly the increased presence and influence of African American voters, contributed to a closer margin. Voter registration efforts and grassroots organizing played a significant role in mobilizing Democratic-leaning voters. Additionally, issues of racial justice, voting rights, and demographic shifts in suburban areas played a part in making Georgia a highly contested state.

Printing out the difference in votes for all the states in the US, I received the following output:

```
state     diff
Alabama   0.083
Alaska    0.067
Arizona   0.059
Arkansas  -0.120
California     0.121
Colorado  0.085
Connecticut    0.191
Delaware  0.093
Florida   0.043
Georgia   0.074
Hawaii    0.157
Idaho     0.003
Illinois  0.002
Indiana   -0.005
Iowa 0.038
Kansas    -0.008
Kentucky  0.001
Louisiana 0.035
Maine     0.060
```

```
Maine CD-1     0.098
Maine CD-2     0.060
Maryland  0.044
Massachusetts  0.115
Michigan  0.039
Minnesota 0.004
Mississippi    0.049
Missouri  0.018
Montana   0.110
Nebraska CD-2  0.146
Nevada    0.071
New Hampshire  0.044
New Jersey     0.090
New Mexico     0.064
New York  0.132
North Carolina 0.023
North Dakota   0.099
Ohio 0.017
Oklahoma  0.024
Oregon    0.053
Pennsylvania   0.025
South Carolina 0.015
South Dakota   0.079
Tennessee 0.016
Texas     0.101
Utah -0.018
Vermont   -0.035
Virginia  0.050
Washington     0.077
West Virginia  0.109
Wisconsin 0.024
Wyoming   0.134
```

Seeing that most of these values were positive, that means most of the states voted Democratically in 2020 compared to 2016. Democrats often perform well in suburban and urban areas, where there is typically a concentration of Democratic-leaning voters. In recent years, suburban areas have seen a shift towards the Democratic Party, driven by factors such as changing attitudes on social issues, education, and the economy. This shift was evident in some states in the 2020 election, as suburban voters played a crucial role in determining the outcome.

Polls in the 2016 presidential election exhibited a tendency to underestimate the percentage of votes received by Donald Trump, resulting in a polling error. This bias can be attributed to various factors. Firstly, sampling issues played a significant role, as polls rely on sampling techniques that may not accurately represent the broader electorate. If the sampling fails to capture the diversity of the population, the results may be skewed. Additionally, non-response bias could have influenced the accuracy of the polls. When certain individuals or groups decline to participate in the survey, their political preferences may be underrepresented, leading to a biased estimation of support. Furthermore, the presence of a social desirability bias may have affected respondents' willingness to disclose their support for Trump, particularly in cases where his candidacy was met with social stigma. It was unprecedented for a candidate to have no previous government or military experience, as well as some of his statements that were met with controversy.

Consequently, some Trump supporters may have been reluctant to reveal their true voting intentions to pollsters, resulting in an underestimation of their support. Methodological challenges, including variations in question wording, survey design, and data collection modes, also contributed to potential biases in poll results. While efforts were made to address these issues in the 2020 election, polls still exhibited some deviations from the actual outcomes. Polling, as an evolving field, continuously strives to refine methodologies and mitigate biases; however, the complexities of measuring public opinion present inherent uncertainties that can contribute to discrepancies between poll predictions and election results.

Section 4.

I explored the changes in voting behavior by each state from the 2016 to 2020 election, and sorted by the greatest percentage difference.

Figure 4.1 shows the voting difference between the Democratic and Republican parties from 2016 to 2020, ranked from greatest to least by absolute value. A positive change would be in favor of the Democratic Party, and a negative change indicates a shift toward the Republican candidate.

State	diff_2016	diff_2020	change

Connecticut	0.127	0.318	0.191
Hawaii	0.214	0.371	0.157
Nebraska CD-2	-0.089	0.056	0.146
Wyoming	-0.445	-0.311	0.134
New York	0.196	0.328	0.132
California	0.247	0.368	0.121
Arkansas	-0.183	-0.303	-0.120
Massachusetts	0.267	0.382	0.115
Montana	-0.176	-0.066	0.110
West Virginia	-0.296	-0.187	0.109
Texas	-0.110	-0.009	0.101
North Dakota	-0.283	-0.183	0.099
Maine CD-1	0.158	0.256	0.098
Delaware	0.138	0.232	0.093
New Jersey	0.121	0.211	0.090
Colorado	0.040	0.125	0.085
Alabama	-0.270	-0.187	0.083

South Dakota	-0.216	-0.137	0.079
Washington	0.152	0.229	0.077
Georgia	-0.064	0.010	0.074
Nevada	-0.008	0.063	0.071
Alaska	-0.118	-0.051	0.067
New Mexico	0.080	0.144	0.064
Maine	0.078	0.138	0.060
Maine CD-2	-0.043	0.017	0.060
Arizona	-0.028	0.031	0.059
Oregon	0.124	0.178	0.053
Virginia	0.073	0.123	0.050
Mississippi	-0.175	-0.126	0.049
Maryland	0.287	0.331	0.044
New Hampshire	0.059	0.103	0.044
Florida	-0.009	0.034	0.043
Michigan	0.036	0.075	0.039
Iowa	-0.044	-0.006	0.038

Louisiana	-0.198	-0.162	0.035
Vermont	0.327	0.292	-0.035
Pennsylvania	0.035	0.060	0.025
Wisconsin	0.054	0.078	0.024
Oklahoma	-0.260	-0.236	0.024
North Carolina	-0.002	0.021	0.023
Missouri	-0.093	-0.076	0.018
Utah	-0.135	-0.153	-0.018
Ohio	-0.032	-0.014	0.017
Tennessee	-0.186	-0.170	0.016
South Carolina	-0.082	-0.068	0.015
Kansas	-0.109	-0.117	-0.008
Indiana	-0.123	-0.128	-0.005
Minnesota	0.078	0.082	0.004
Idaho	-0.270	-0.267	0.003
Illinois	0.147	0.149	0.002
Kentucky	-0.188	-0.187	0.001

Based on these numbers, if I were to predict which states would change their votes in 2020, I would choose the states with the lowest percentage difference in 2016, which may indicate that they will potentially flip. I would predict Florida, Nevada, Maine and Wisconsin. In reality, the states that flipped were Arizona, Georgia, Michigan and Pennsylvania. The fluctuating electoral landscape reflected the changing attitudes of the voters, and was crucial in determining the outcome of the 2020 election.

Section 5.

Iowa and Florida were both critical swing states which had a low percentage difference. Comparing the polls to the actual outcomes in R, I received a value of 0.69 for Iowa and 0.439 for Florida.

The relatively high accuracy rate of 69% in Iowa's polls can be associated with several factors. First, Iowa has a long-standing tradition of active political engagement and is considered a state of significance during elections. This has led to the development and utilization of well-established polling methods by organizations conducting surveys in the state. These methods encompass robust sampling techniques, effective data collection strategies, and meticulous analysis, all of which contribute to more accurate results.

Second, the representative sampling employed in Iowa polls plays a crucial role in their accuracy. Pollsters strive to ensure that the selected sample reflects the diverse demographic, geographic, and political characteristics of the state's population. By capturing the nuances and diversity of voter preferences, the polling results become more reliable and representative.

Furthermore, Iowa's political landscape, characterized by competitiveness and close margins between candidates, prompts pollsters to dedicate greater attention to the state. The recognition of Iowa's significance in the electoral process motivates pollsters to employ meticulous techniques and methodologies, resulting in more accurate predictions. With the state's involvement in the presidential nominating process, including the influential Iowa caucuses, there is a heightened level of political awareness and participation among its residents. This engaged electorate facilitates the collection of reliable and informed data, enabling pollsters to gauge public sentiment more effectively.

The general consensus is that the Des Moines Register is the most accurate poll in Iowa. Its collaboration with Mediacom, a local cable television provider, has resulted in

the Iowa Poll becoming a respected and influential source of political polling data in the state. The Des Moines Register has gained recognition for its rigorous sampling techniques, comprehensive analysis, and efforts to capture the nuances of voter sentiment in Iowa.

Florida's low polling percentage accuracy of 0.439 can be attributed to several factors. One is the unique and diverse nature of Florida's electorate. The state is home to a large population with varied demographic and ideological backgrounds, making it challenging for pollsters to capture the full range of voter preferences accurately. Florida also has historically been a swing state with closely contested elections, leading to volatile and unpredictable outcomes. The presence of a large number of undecided or late-deciding voters further complicates the accuracy of polls. Moreover, Florida's population includes a significant number of older voters, who may be less likely to participate in polls or have different voting patterns than anticipated. These complexities and inherent uncertainties make accurately predicting Florida's election results particularly challenging, resulting in a lower polling accuracy percentage.

In these two states, the polling was not completely accurate. There are several possible reasons that can account for biased polls. One reason could be the challenges of accurately capturing the diverse and complex demographics of these states. Both Florida and Iowa have populations with diverse political ideologies, age groups, ethnicities, and regional variations, making it difficult for pollsters to create representative samples. Additionally, non-response bias may contribute to the polling bias, as certain individuals or groups may be less likely to participate in surveys, leading to underrepresentation of their views.The presence of "hidden" or "shy" voters, who may not openly disclose their true preferences, can introduce additional bias. Finally, methodological issues, such as question wording, survey design, and data collection methods, can also impact the accuracy of polls. These factors combined contribute to the bias observed in the polls for Florida and Iowa.

Improving political election polls is a multifaceted endeavor that involves implementing a range of strategies and considerations. One crucial aspect is ensuring representative sampling. By employing rigorous sampling techniques such as random or stratified sampling, pollsters can obtain a diverse and unbiased sample that accurately reflects the electorate. Efforts should be made to include individuals from different demographic groups, geographic regions, and political affiliations to minimize biases.

Another way to improve polls is to address non-response bias, which occurs when certain individuals or groups are less likely to participate in surveys, leading to a skewed sample. Pollsters can mitigate this bias by implementing strategies to encourage

participation, such as offering multiple modes of data collection (online, telephone, etc.), providing incentives, and addressing privacy concerns. Techniques such as multilevel modeling, post-stratification, and machine learning algorithms can help account for complex relationships and identify potential biases in the data. These advanced approaches allow for more nuanced analysis and prediction.

Transparency is essential in enhancing the credibility of polls. Pollsters should disclose information about their methodologies, including details about the sampling process, survey design, and weighting techniques used. Transparency allows for greater scrutiny and understanding of the poll's limitations and strengths. It is important to acknowledge that predicting human behavior and election outcomes will always have inherent uncertainties. External factors, shifting political dynamics, and evolving voter preferences make it an ongoing challenge.