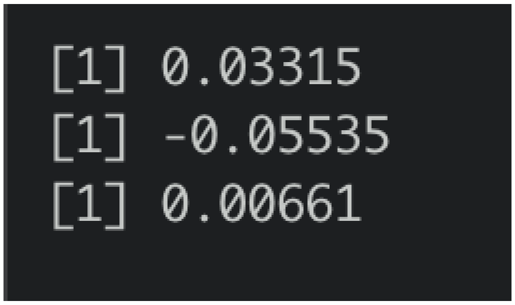
Data Analytic Report

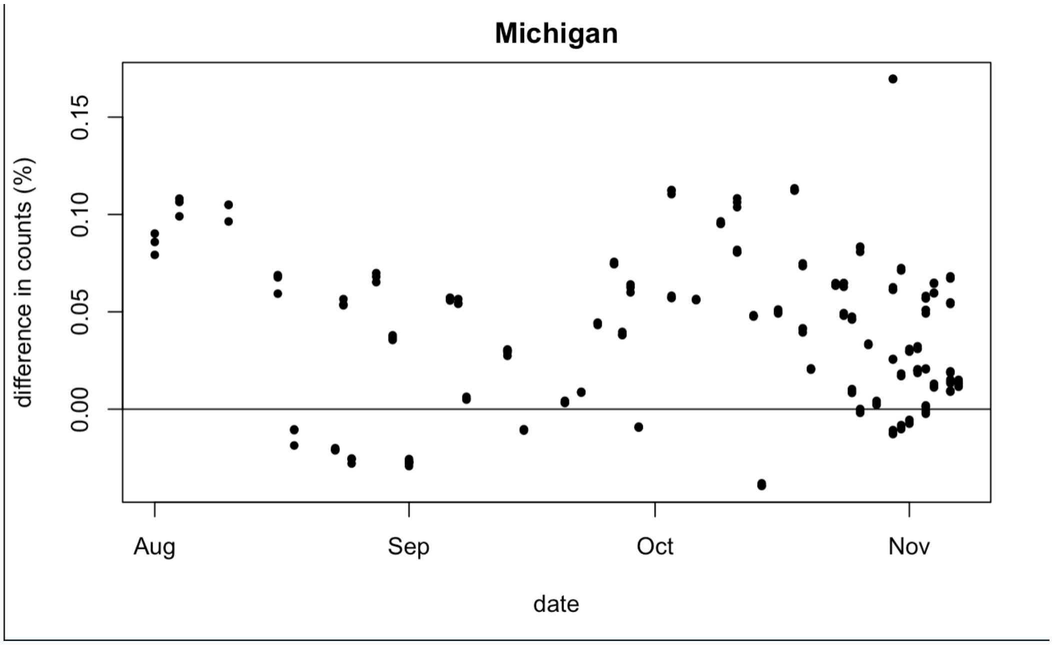
Sayan Andrews

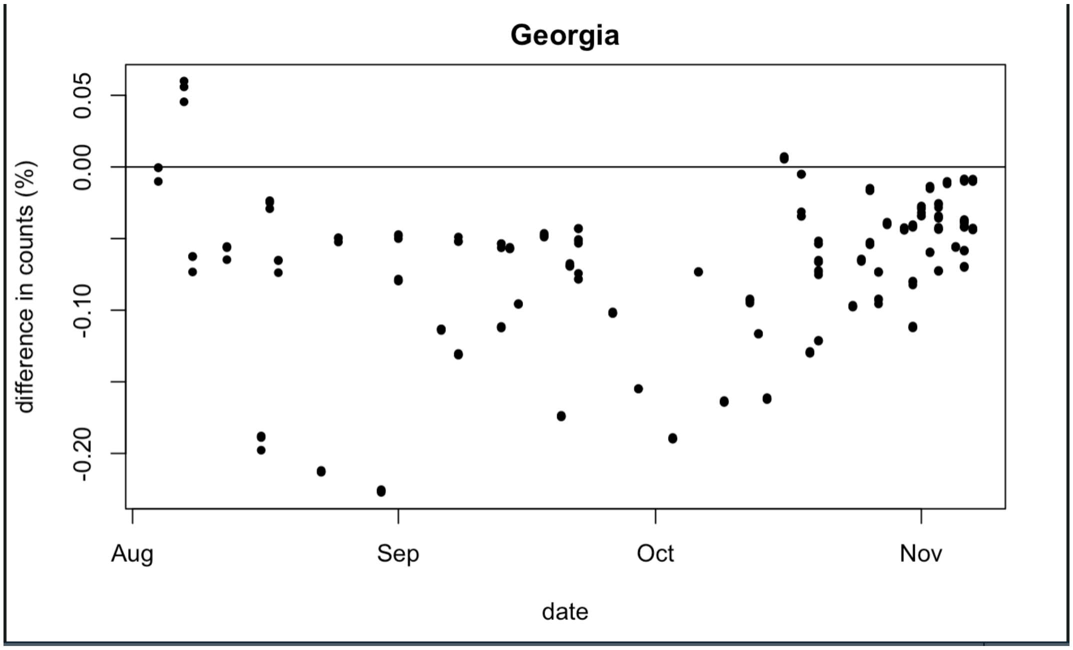
5/25/2023

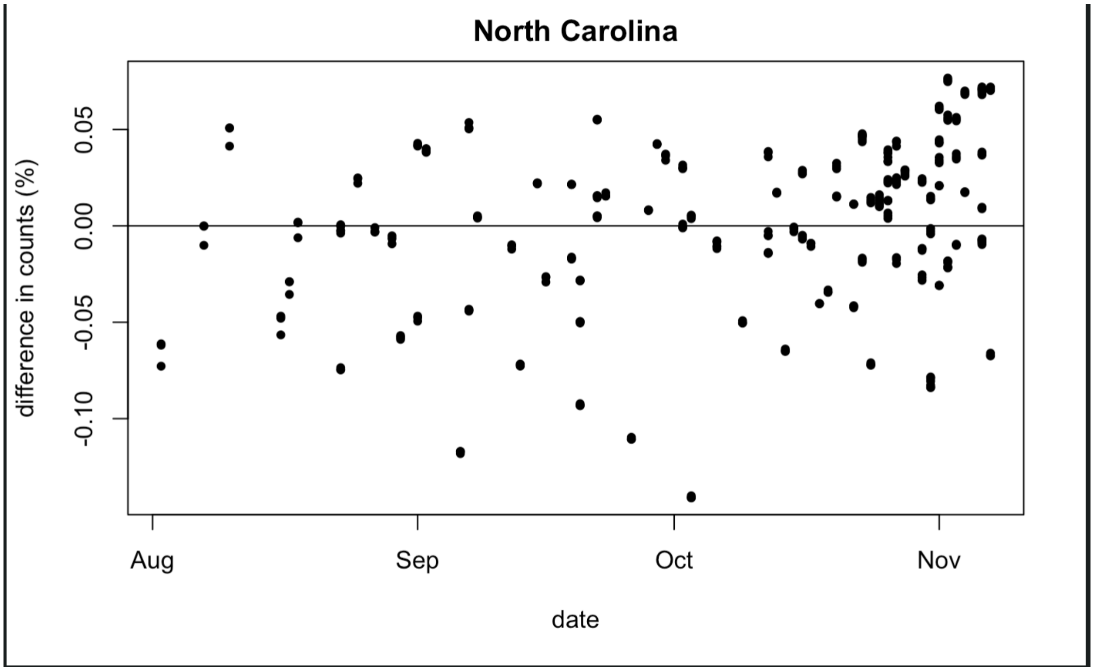
1. a) We can see that Clinton is ahead in the state of Michigan, by a percentage difference of 3.315%. Trump is ahead in the state of Georgia, by a percentage difference of 5.54%. Clinton is barely ahead in the state of North Carolina, by a percentage difference of 0.66%.



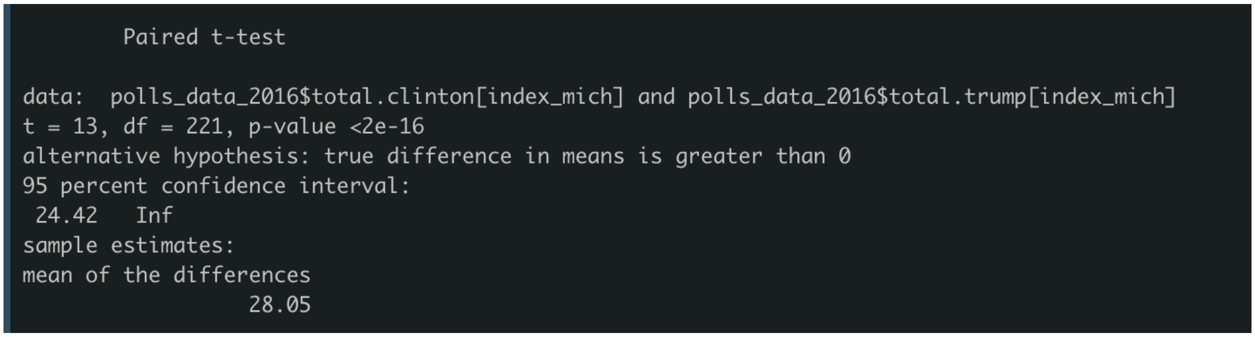
```
[1] 0.03315
[1] -0.05535
[1] 0.00661
```



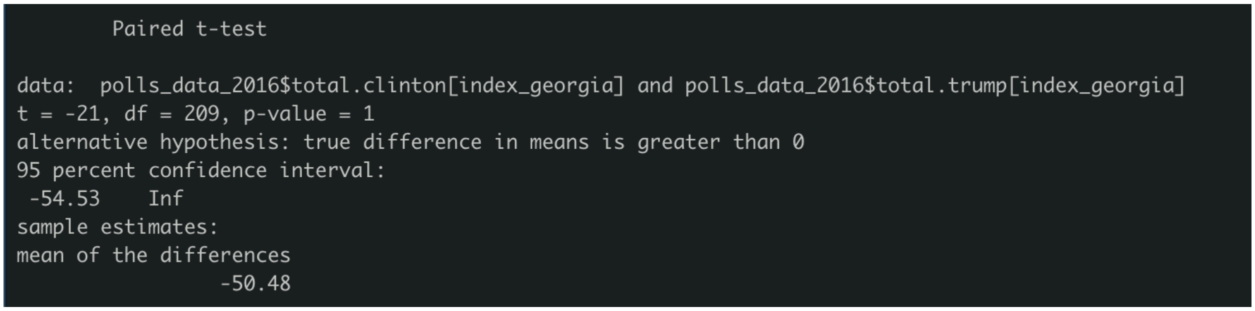




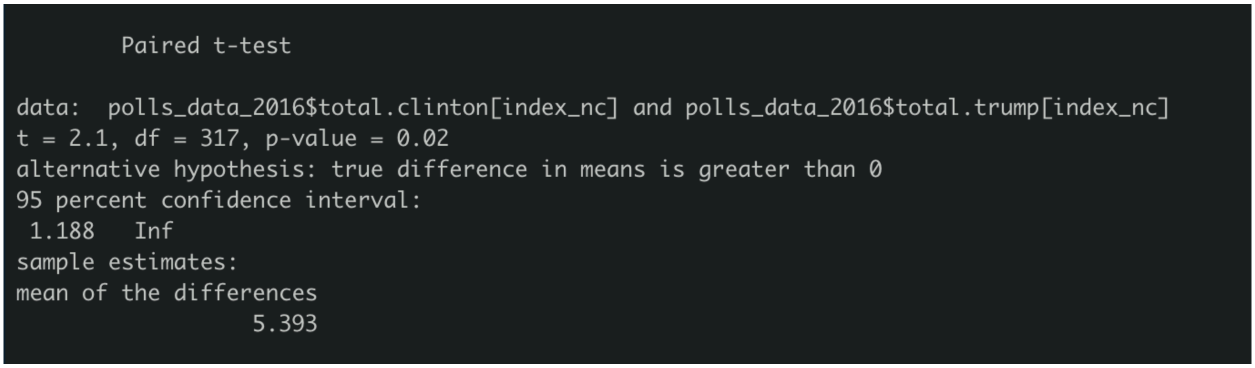
b) Based on these pair t-tests, the p-value for Michigan is <2e-16, a very small p-value, meaning we reject the null hypothesis, so Clinton is in favor of winning because the difference between her and Trump is greater than 0. The p-value for Georgia is 1, a very high p-value, so we fail to reject the null hypothesis, but this only tells us that the true difference in means is about 0, not that Trump is in favor of winning in Georgia. Here arises the problem with this test, in that it would need multiple alternatives, or the means to find a negative difference in order to rule Trump in favor of winning that particular state. Therefore, this test lacks significance because it only is able to factor in negligible difference or difference in favor of Clinton, given that our test statistic is Clinton-Trump. The p-value for North Carolina is 0.02, which would be lower than the standard significance level of 0.05, so we would reject the null and find Clinton ahead. But it is ambiguous as to if this makes sense for North Carolina, given that there was a very small percentage difference that we found for this state in part (a). Another potential problem with this test is that it assumes that all trials are independent, when that may not necessarily be the case.



```
	Paired t-test

data:  polls_data_2016$total.clinton[index_mich] and polls_data_2016$total.trump[index_mich]
t = 13, df = 221, p-value <2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 24.42   Inf
sample estimates:
mean of the differences
                  28.05
```

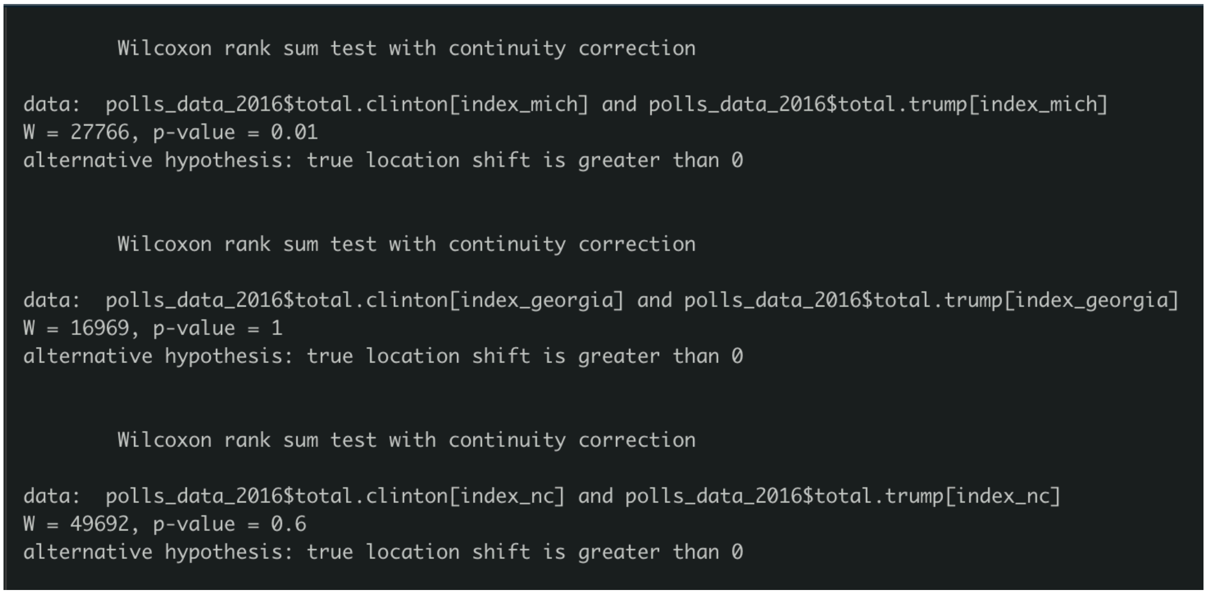


```
	Paired t-test

data:  polls_data_2016$total.clinton[index_georgia] and polls_data_2016$total.trump[index_georgia]
t = -21, df = 209, p-value = 1
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -54.53    Inf
sample estimates:
mean of the differences
                -50.48
```



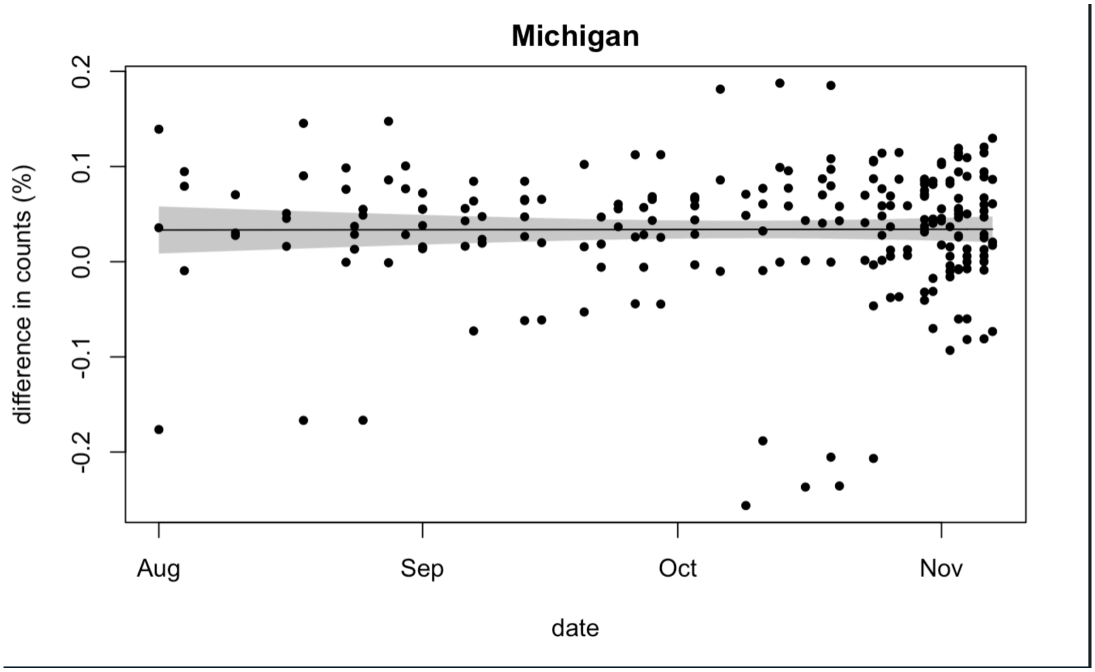
```
	Paired t-test

data:  polls_data_2016$total.clinton[index_nc] and polls_data_2016$total.trump[index_nc]
t = 2.1, df = 317, p-value = 0.02
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 1.188   Inf
sample estimates:
mean of the differences
                  5.393
```

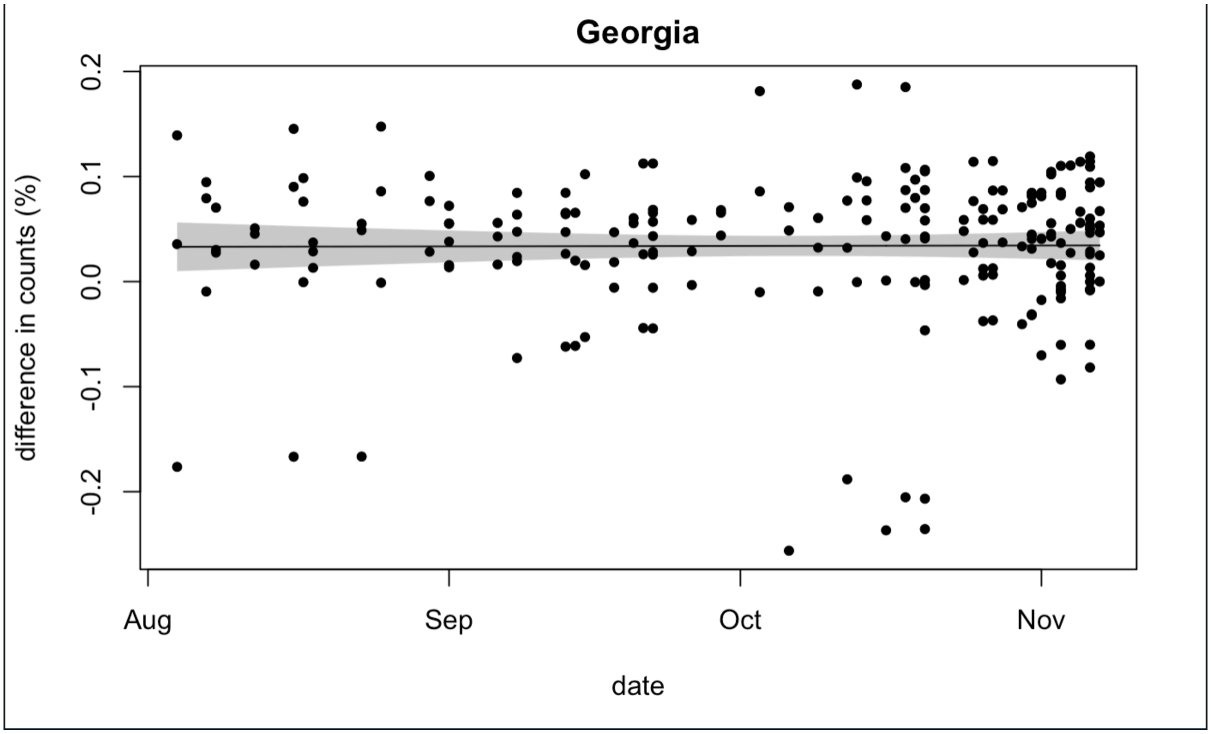
c) Based on these Wilcoxon Sign Rank tests, Clinton is in favor of winning again in Michigan with a p-value of 0.01, leading us to reject the null hypothesis that the location shift between her and Trump is 0, in favor of it being greater than 0. However, with p values of 1 and 0.6 for Georgia and North Carolina, we again run into the same problem of ambiguity as with the paired t-tests in part While these results lead us to fail to reject the null hypothesis, we are left with the notion that the true location shift is 0, meaning Clinton and Trump hold the same favor in those states, but this is simply not true. Given that in part (a), we were able to determine a statistically significant percentage difference in Georgia specifically in favor of Trump, we again have the same problem when there is no rejection of the null, we have no determination of whether or not Trump won favor in that state with these tests, as that would require a negative difference to be determined from this test, which is not possible under the current circumstances of hypotheses.

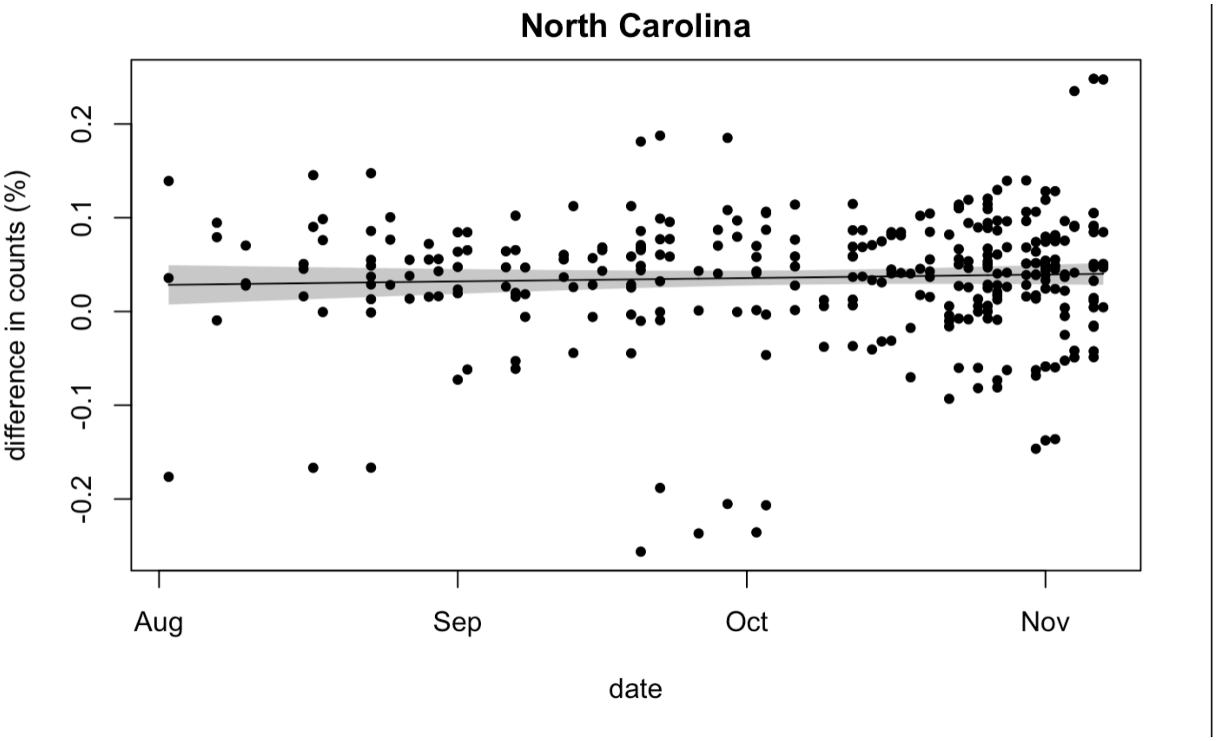


```
        Wilcoxon rank sum test with continuity correction

data:  polls_data_2016$total.clinton[index_mich] and polls_data_2016$total.trump[index_mich]
W = 27766, p-value = 0.01
alternative hypothesis: true location shift is greater than 0


        Wilcoxon rank sum test with continuity correction

data:  polls_data_2016$total.clinton[index_georgia] and polls_data_2016$total.trump[index_georgia]
W = 16969, p-value = 1
alternative hypothesis: true location shift is greater than 0


        Wilcoxon rank sum test with continuity correction

data:  polls_data_2016$total.clinton[index_nc] and polls_data_2016$total.trump[index_nc]
W = 49692, p-value = 0.6
alternative hypothesis: true location shift is greater than 0
```

(d) From the linear model and observations, North Carolina may have the closest election by percentage difference. Looking at all of the graphs below, North Carolina shows the best linear fit, with a cluster by the line at a 0.0% difference.



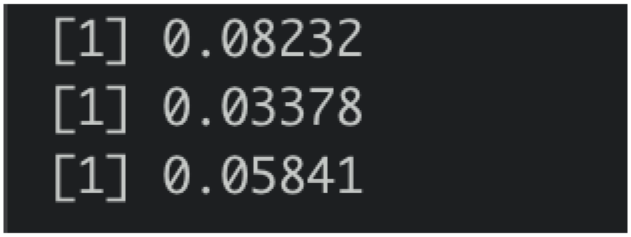




(e) From the real results of the 2016 election, Michigan had the smallest margin in terms of percentage difference, with a 0.3 percent difference. Two reasons for which the real results may differ from what the polls indicate are sampling and response bias, as well as methodological issues.

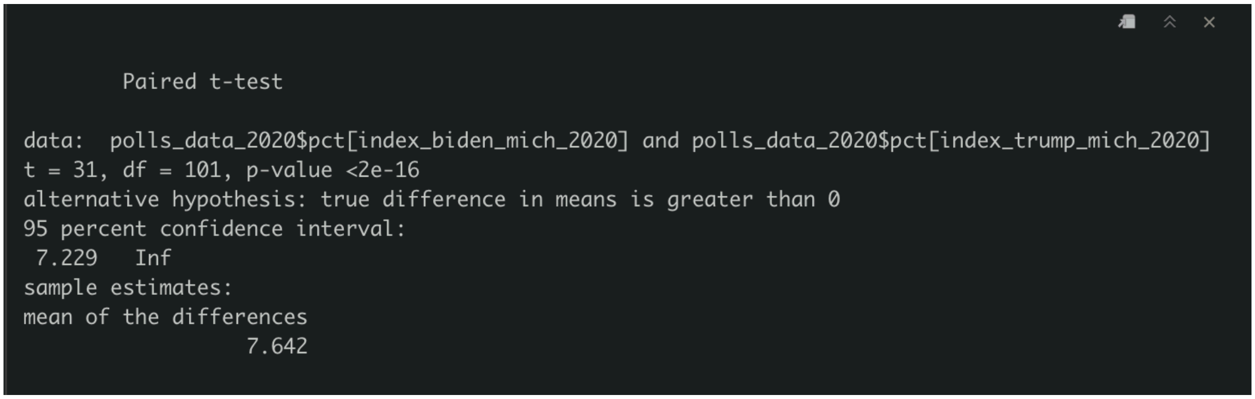
(f) Polls do not always correctly predict the candidate who wins each state in an election. There can be biases in polls that contribute to these inaccuracies. Some possible reasons for poll biases in predicting state winners include sampling bias and non-response bias.

2. a) From the data we can see that Biden was leading in Michigan by an 8.232% difference. Biden was leading Georgia by a 3.378% difference. Finally, Biden was leading North Carolina by a 5.841% difference.

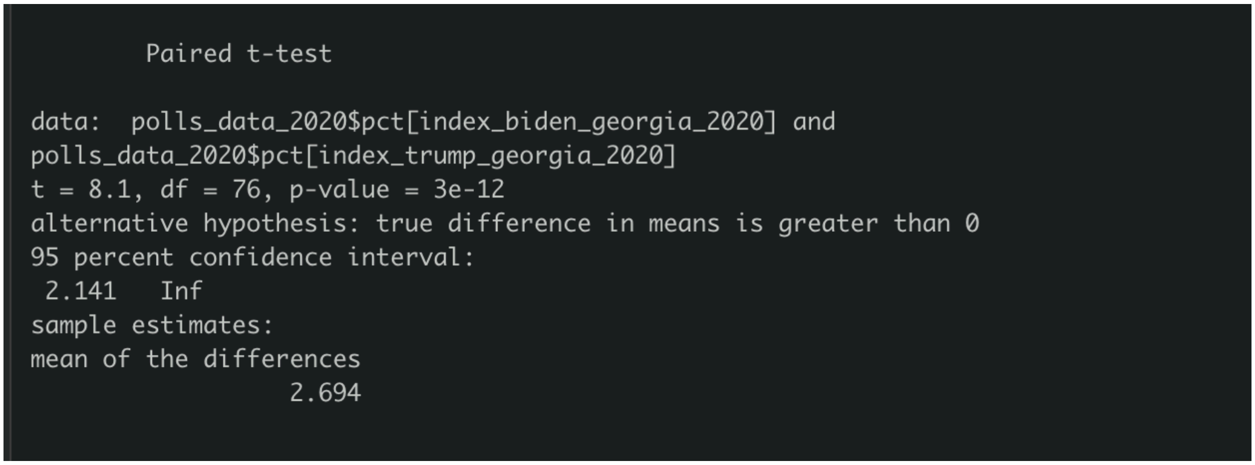
   

   ```
   [1] 0.08232
   [1] 0.03378
   [1] 0.05841
   ```

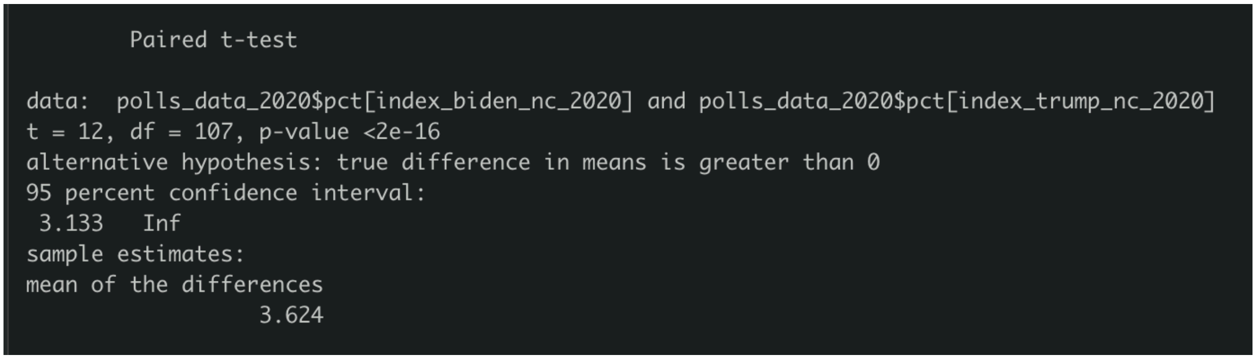
   b) Based on the t-test, the p-value for Michigan is <2e-16, a very small p-value, meaning we reject the null hypothesis in favor of the alternative, that the true mean of the difference in poll counts in Michigan is greater than 0. Therefore, Biden holds significant favor within Michigan, which is in line with our data from part (a). For the state of Georgia, we find a p-value of 3e-12, also very small, leading to the same conclusion. For North Carolina, we get another small p-value of <2e-16, which is again in line with part (a). This proves that this test is significant. However, a potential problem with this test is that it assumes that all trials are independent, when that may not necessarily be the case.



```
       Paired t-test

data:  polls_data_2020$pct[index_biden_mich_2020] and polls_data_2020$pct[index_trump_mich_2020]
t = 31, df = 101, p-value <2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 7.229   Inf
sample estimates:
mean of the differences
                  7.642
```

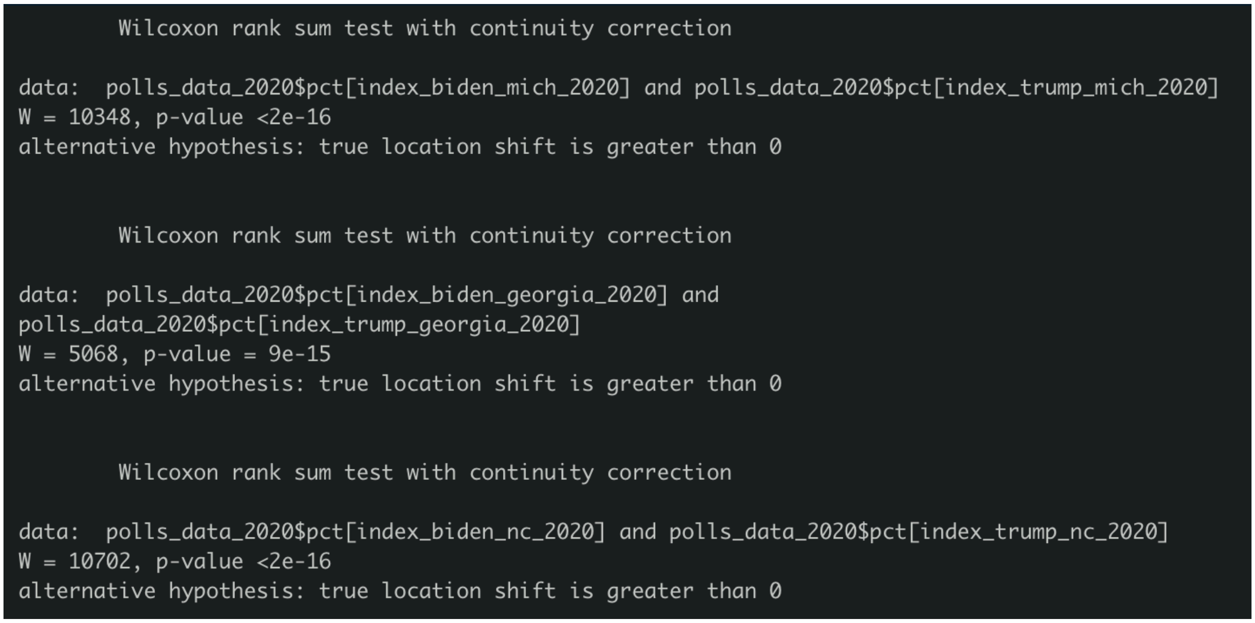


```
       Paired t-test

data:  polls_data_2020$pct[index_biden_georgia_2020] and
polls_data_2020$pct[index_trump_georgia_2020]
t = 8.1, df = 76, p-value = 3e-12
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 2.141   Inf
sample estimates:
mean of the differences
                  2.694
```



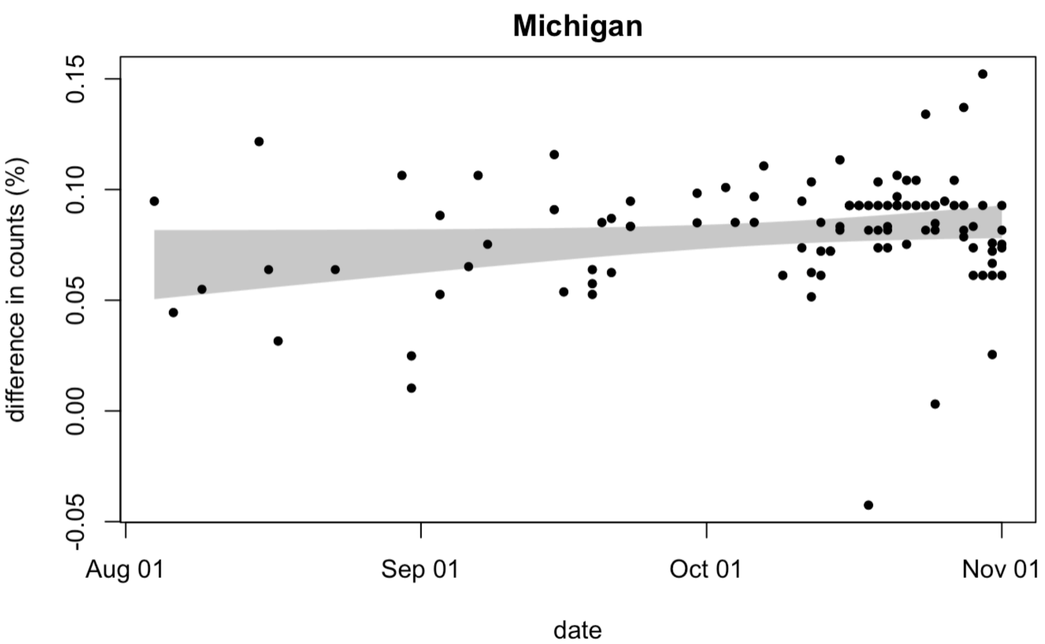
```
       Paired t-test

data:  polls_data_2020$pct[index_biden_nc_2020] and polls_data_2020$pct[index_trump_nc_2020]
t = 12, df = 107, p-value <2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 3.133   Inf
sample estimates:
mean of the differences
                  3.624
```

c) Based on the Wilcox rank test, the p-value for Michigan is <2e-16, meaning we reject the null hypothesis in favor of the location shift being greater than 0, meaning Biden holds favor in Michigan. This checks out with our previous data. The p-value for Georgia is small at 9e-15, below any reasonable significance level such as 0.05, meaning we reject the null hypothesis, and we believe the true difference to be greater than 0, which matches our notion that Biden held a significant lead in the state. We get a p-value of <2e-16 as well for North Carolina, which gives us the same conclusion. There seem to be no potential problems with this test.

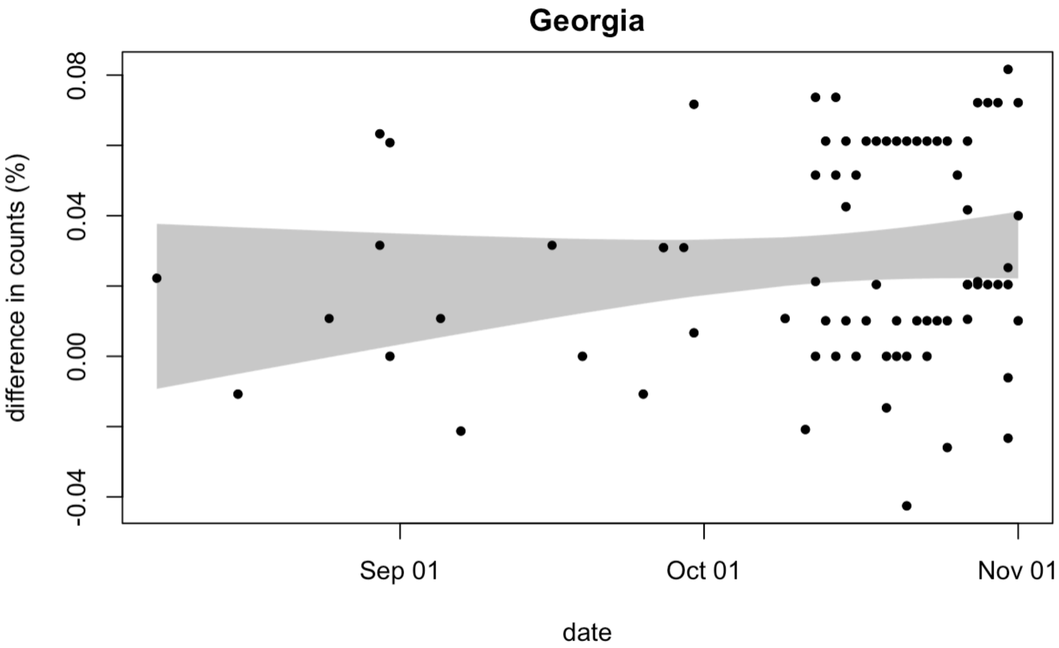


```
	Wilcoxon rank sum test with continuity correction

data:  polls_data_2020$pct[index_biden_mich_2020] and polls_data_2020$pct[index_trump_mich_2020]
W = 10348, p-value <2e-16
alternative hypothesis: true location shift is greater than 0


	Wilcoxon rank sum test with continuity correction

data:  polls_data_2020$pct[index_biden_georgia_2020] and
polls_data_2020$pct[index_trump_georgia_2020]
W = 5068, p-value = 9e-15
alternative hypothesis: true location shift is greater than 0


	Wilcoxon rank sum test with continuity correction

data:  polls_data_2020$pct[index_biden_nc_2020] and polls_data_2020$pct[index_trump_nc_2020]
W = 10702, p-value <2e-16
alternative hypothesis: true location shift is greater than 0
```

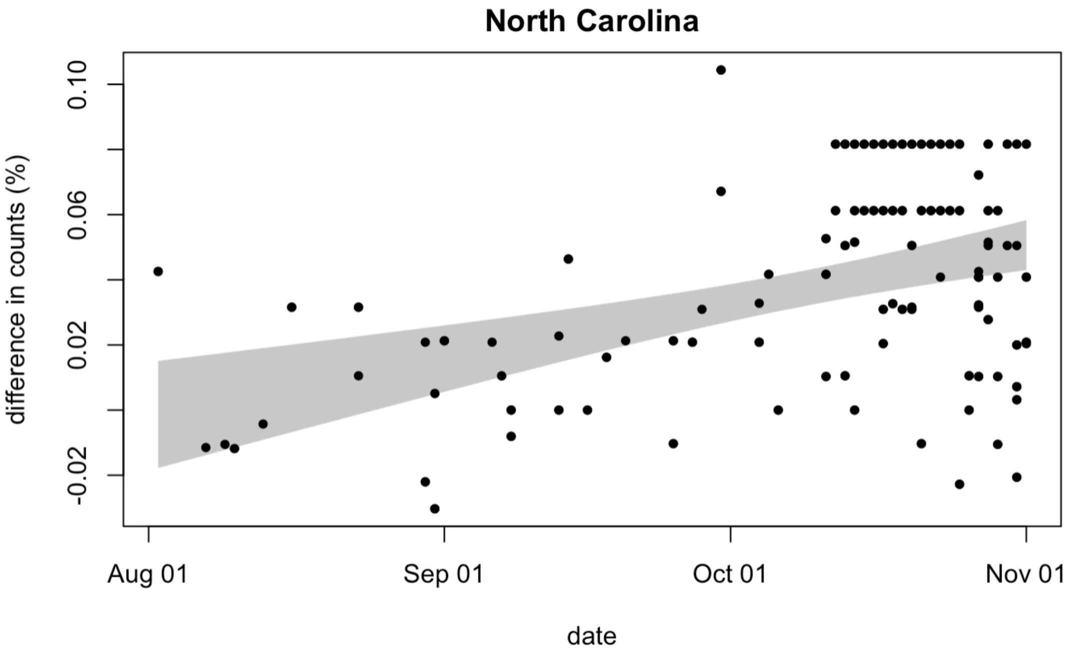
d) From the linear model and observations, Georgia looks to have the closest election by percentage difference. While both Georgia and North Carolina have a decent amount of clusters around the 0.0% difference mark in the later months we are examining, Georgia shows more regression to the means, though none of these states show a great linear fit in these graphs. Finally, Michigan has the smallest residual standard error, proving its fit to the mean, however, that line and mean regresses towards a larger percentage difference in the 0.05 to 0.1 areas.



Michigan



Georgia

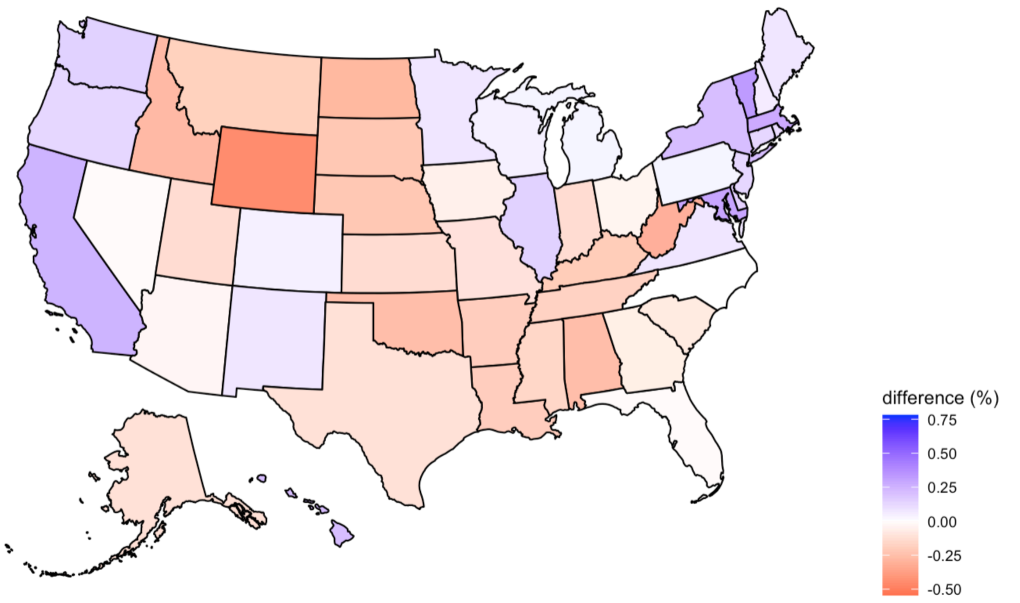


(e) From the real results of the 2020 election, Georgia had the smallest margin in terms of percentage difference, with a 0.2 percent difference. Two reasons for which the real results may differ from what the polls indicate are sampling and response bias, as well as methodological issues.

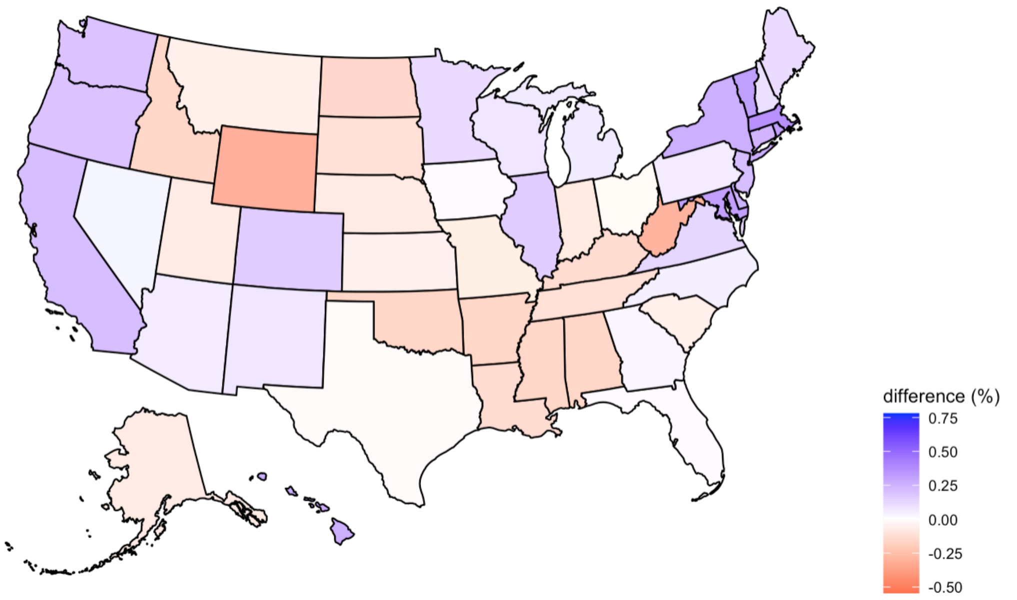
(f) Polls do not always correctly predict the candidate who wins each state in an election. There can be biases in polls that contribute to these inaccuracies. Some possible reasons for poll biases in predicting state winners include sampling bias and non-response bias.

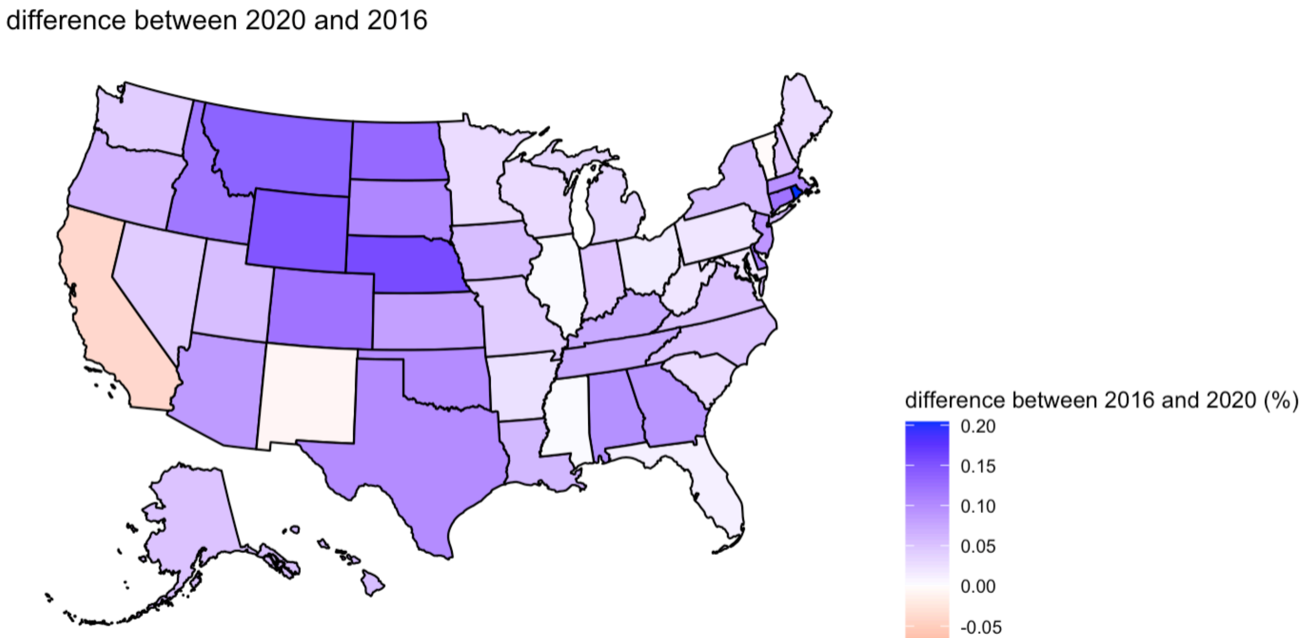
3. a) Looking at the difference in percentage between 2016 and 2020, we see mostly purple across the board, indicating most of the country had a great difference between those election years. The extremity of those differences in just 4 years suggests a great change in the political climate and voter satisfaction during the Trump presidency.

2016

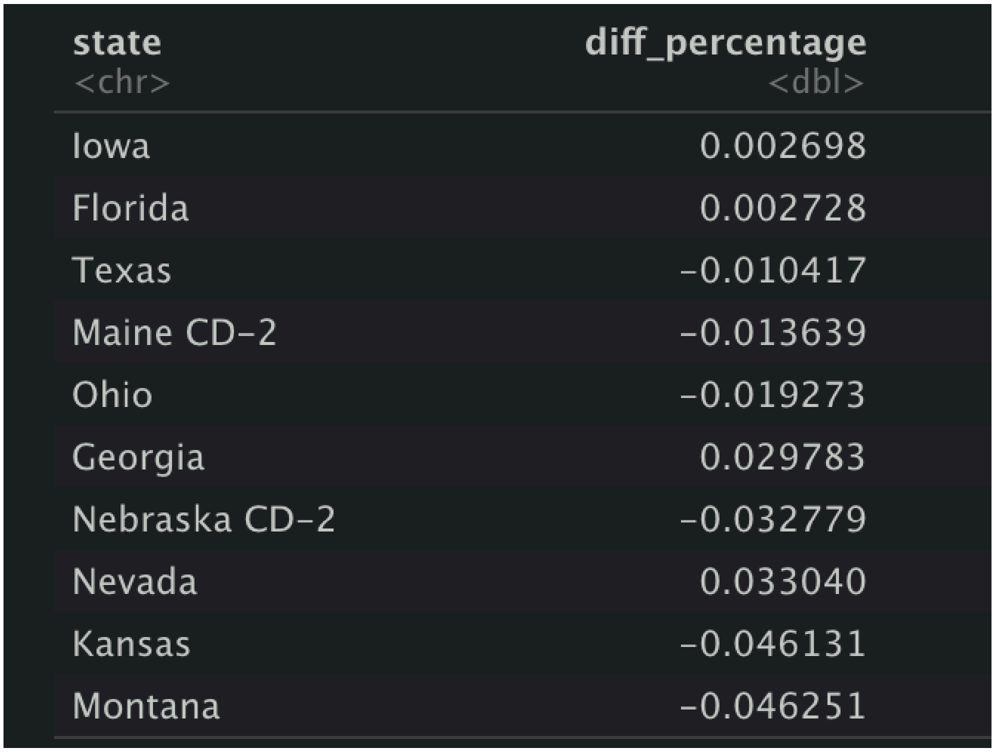

difference (%)
0.75
0.50
0.25
0.00
-0.25
-0.50

2020


difference (%)
0.75
0.50
0.25
0.00
-0.25
-0.50

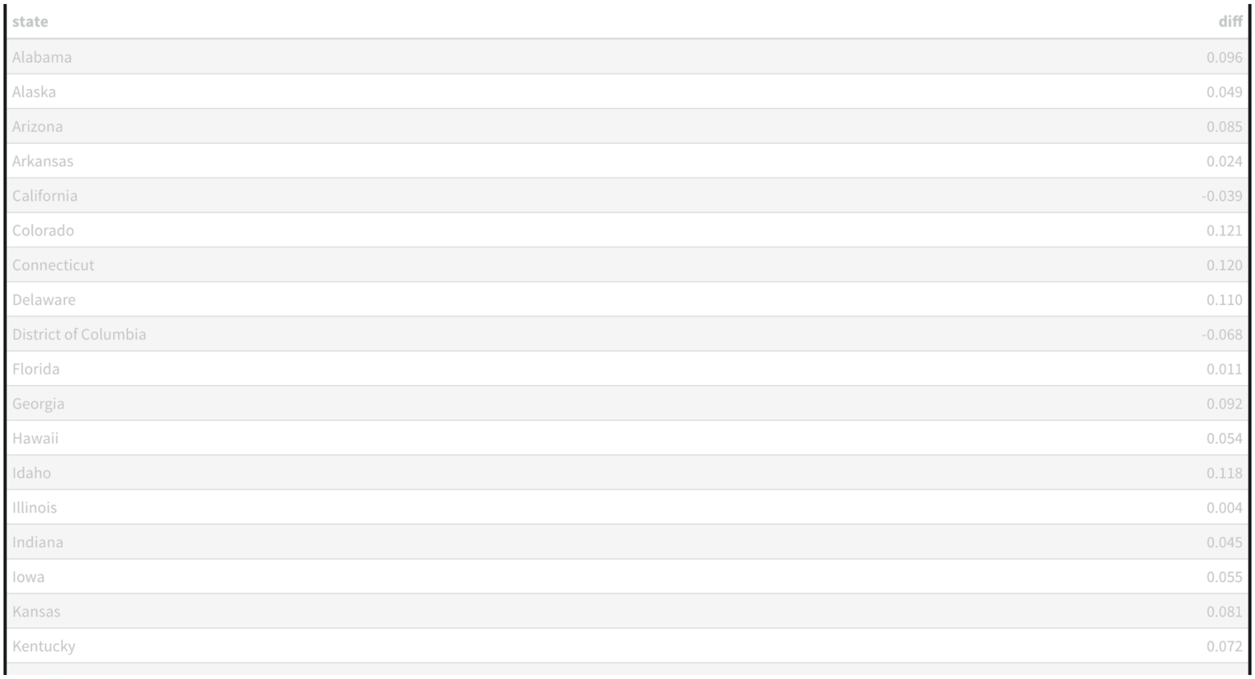

difference between 2020 and 2016
difference between 2016 and 2020 (%)
0.20
0.15
0.10
0.05
0.00
-0.05

b) Based on the plots for (a), Iowa, Florida, Texas, Maine, Ohio, Georgia, Nebraska, Nevada, Kansas, and Montana, were 10 battleground states in 2020. They had the smallest percentage differences when looking at the graphs.

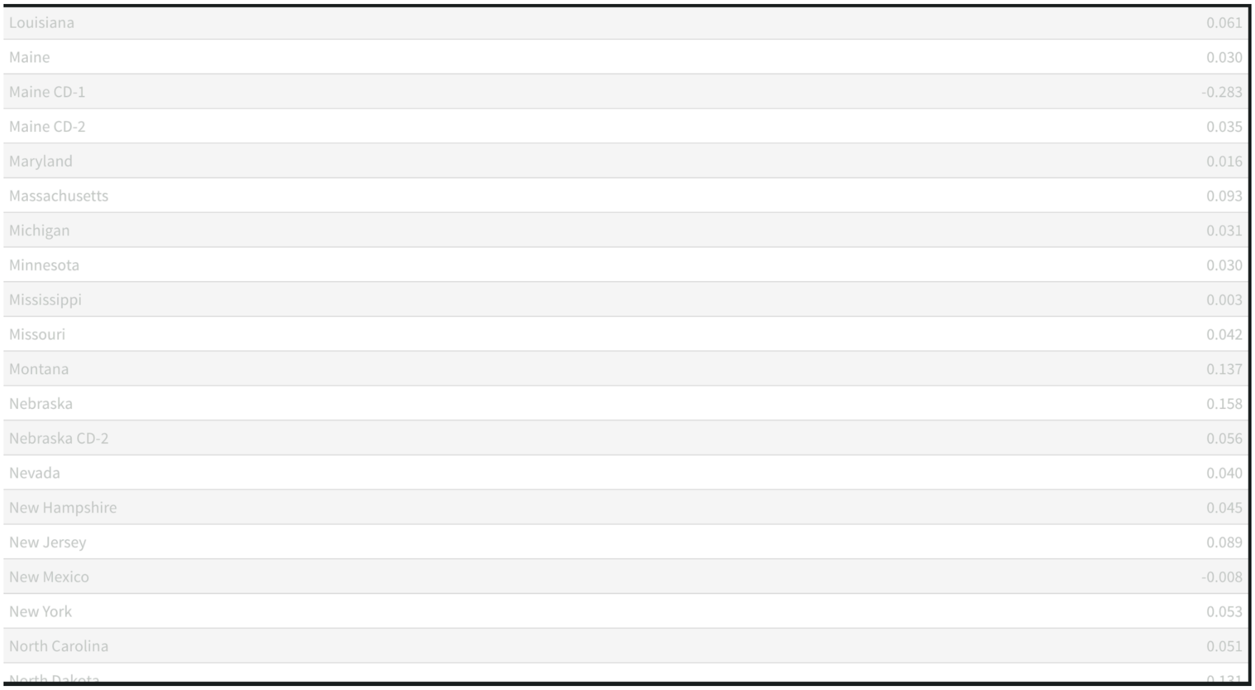


state <chr>	diff_percentage <dbl>
Iowa	0.002698
Florida	0.002728
Texas	-0.010417
Maine CD-2	-0.013639
Ohio	-0.019273
Georgia	0.029783
Nebraska CD-2	-0.032779
Nevada	0.033040
Kansas	-0.046131
Montana	-0.046251

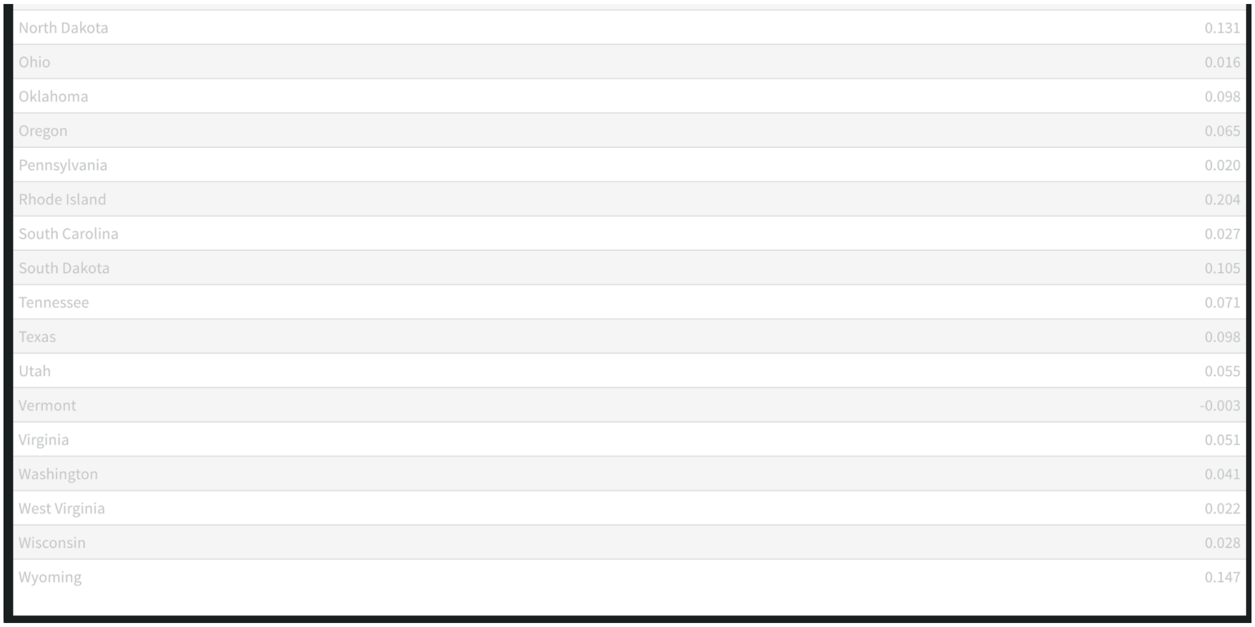
c) A lot of states became more leftist than rightist between 2016 and 2020.



state	diff
Alabama	0.096
Alaska	0.049
Arizona	0.085
Arkansas	0.024
California	-0.039
Colorado	0.121
Connecticut	0.120
Delaware	0.110
District of Columbia	-0.068
Florida	0.011
Georgia	0.092
Hawaii	0.054
Idaho	0.118
Illinois	0.004
Indiana	0.045
Iowa	0.055
Kansas	0.081
Kentucky	0.072



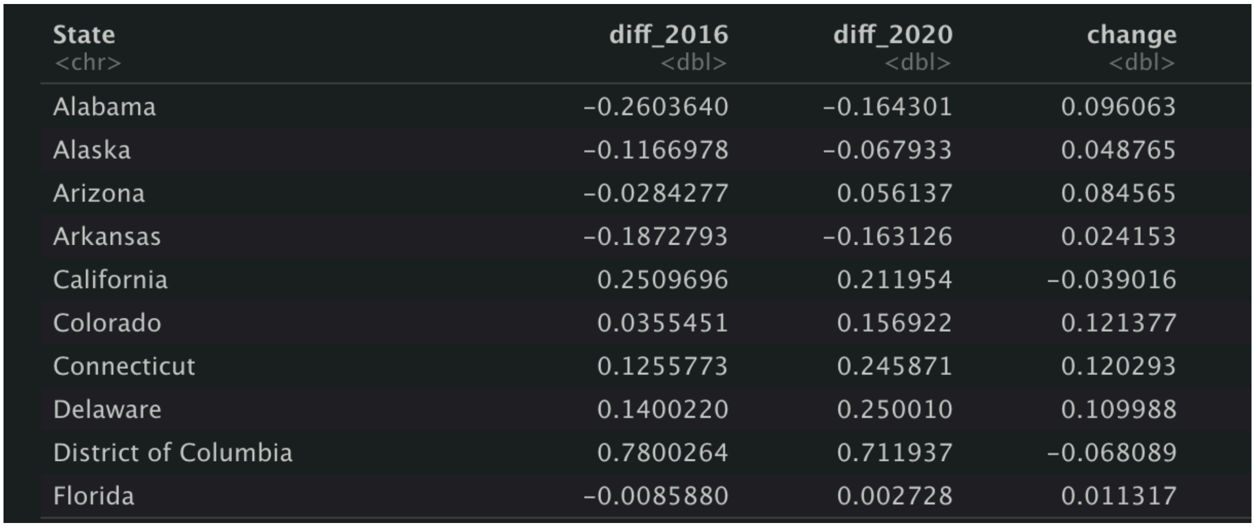
Louisiana	0.061
Maine	0.030
Maine CD-1	-0.283
Maine CD-2	0.035
Maryland	0.016
Massachusetts	0.093
Michigan	0.031
Minnesota	0.030
Mississippi	0.003
Missouri	0.042
Montana	0.137
Nebraska	0.158
Nebraska CD-2	0.056
Nevada	0.040
New Hampshire	0.045
New Jersey	0.089
New Mexico	-0.008
New York	0.053
North Carolina	0.051



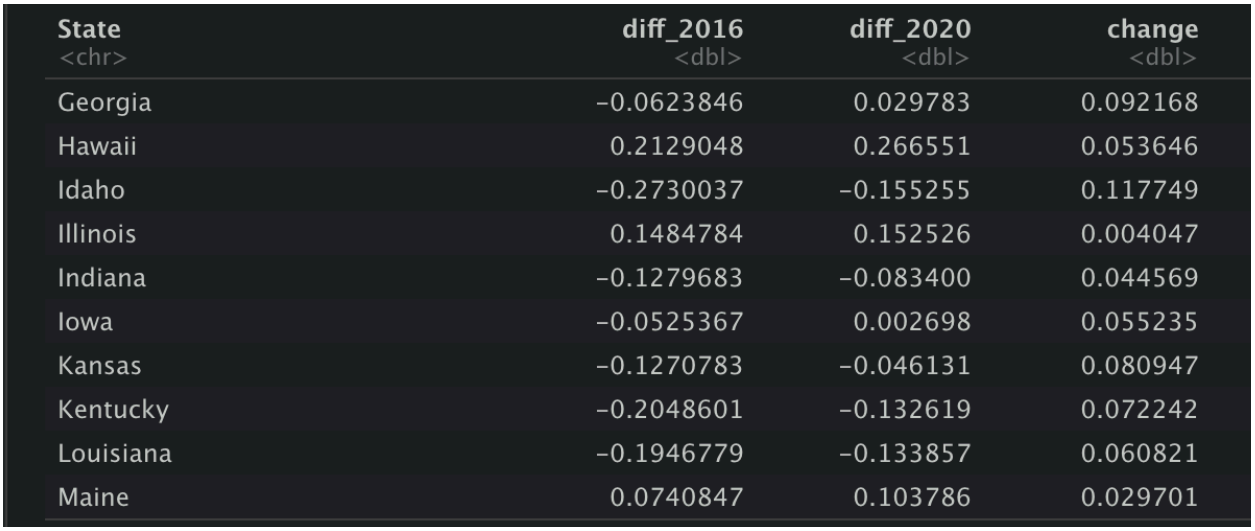
North Dakota	0.131
Ohio	0.016
Oklahoma	0.098
Oregon	0.065
Pennsylvania	0.020
Rhode Island	0.204
South Carolina	0.027
South Dakota	0.105
Tennessee	0.071
Texas	0.098
Utah	0.055
Vermont	-0.003
Virginia	0.051
Washington	0.041
West Virginia	0.022
Wisconsin	0.028
Wyoming	0.147

d) In both the 2016 and 2020 US presidential elections, polls did underestimate the percentage of the real votes received, especially for Donald Trump in both years. The bias in the polls can be explained by non-response bias, sampling and weighting issues, a differential in voter turnout, and volatile and late-deciding voters, to name a few.

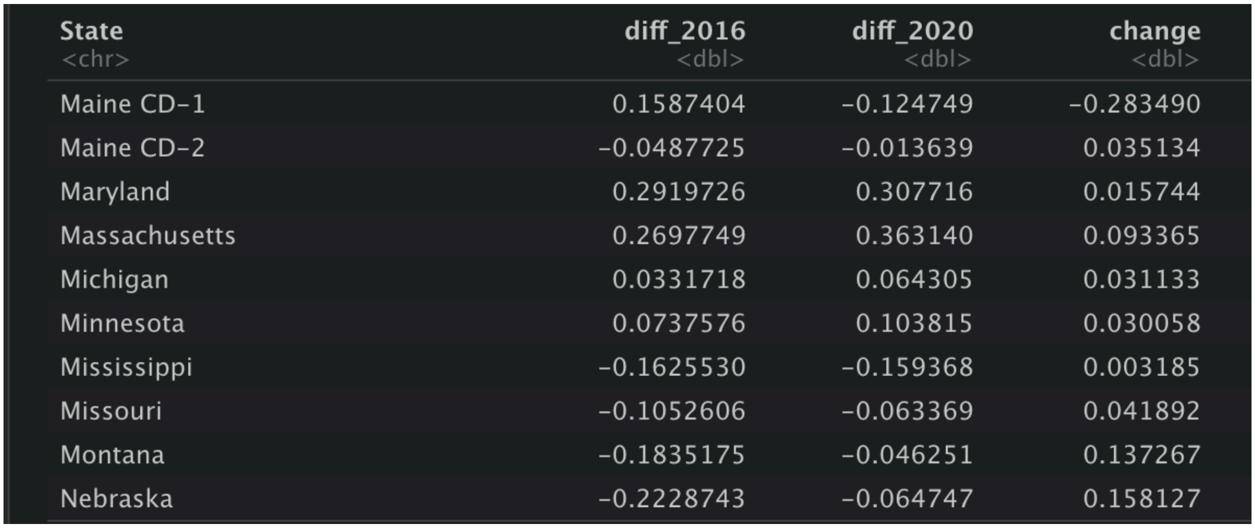
4. a)



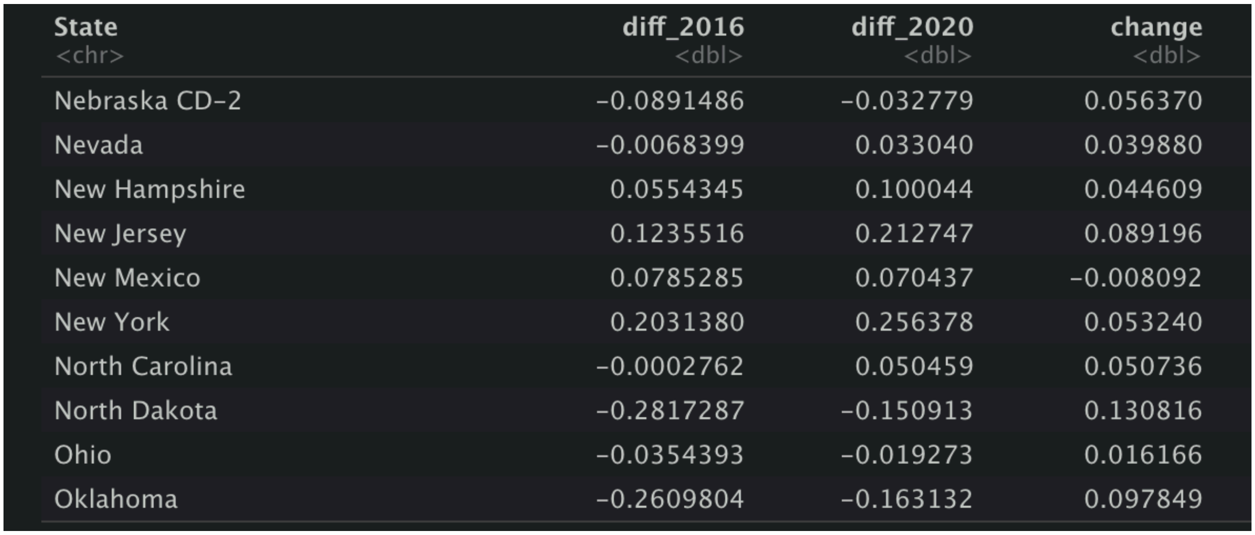
State	diff_2016	diff_2020	change
<chr>	<dbl>	<dbl>	<dbl>
Alabama	-0.2603640	-0.164301	0.096063
Alaska	-0.1166978	-0.067933	0.048765
Arizona	-0.0284277	0.056137	0.084565
Arkansas	-0.1872793	-0.163126	0.024153
California	0.2509696	0.211954	-0.039016
Colorado	0.0355451	0.156922	0.121377
Connecticut	0.1255773	0.245871	0.120293
Delaware	0.1400220	0.250010	0.109988
District of Columbia	0.7800264	0.711937	-0.068089
Florida	-0.0085880	0.002728	0.011317



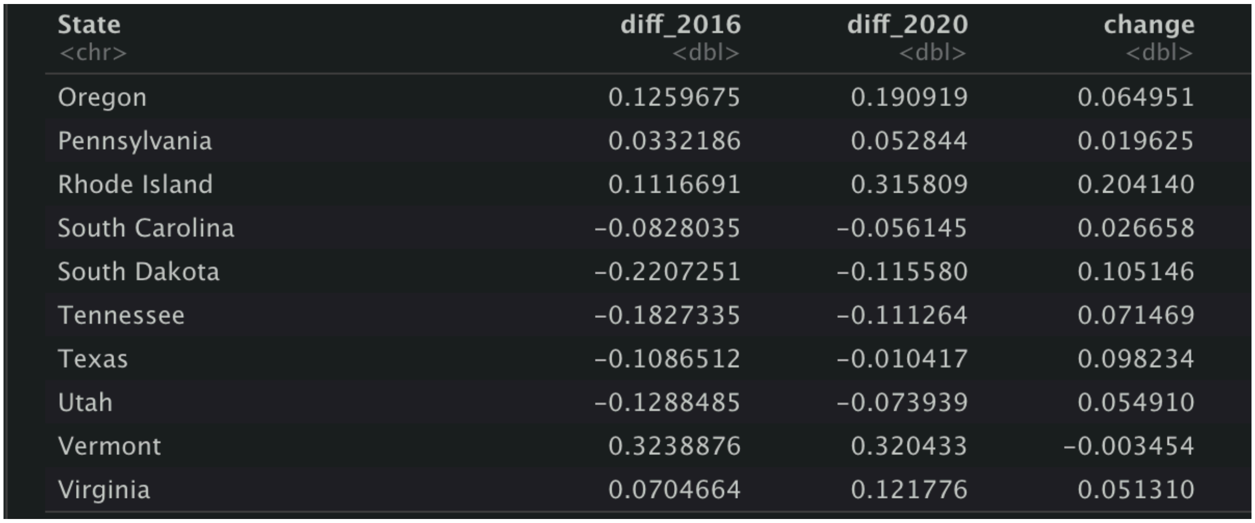
State	diff_2016	diff_2020	change
<chr>	<dbl>	<dbl>	<dbl>
Georgia	-0.0623846	0.029783	0.092168
Hawaii	0.2129048	0.266551	0.053646
Idaho	-0.2730037	-0.155255	0.117749
Illinois	0.1484784	0.152526	0.004047
Indiana	-0.1279683	-0.083400	0.044569
Iowa	-0.0525367	0.002698	0.055235
Kansas	-0.1270783	-0.046131	0.080947
Kentucky	-0.2048601	-0.132619	0.072242
Louisiana	-0.1946779	-0.133857	0.060821
Maine	0.0740847	0.103786	0.029701



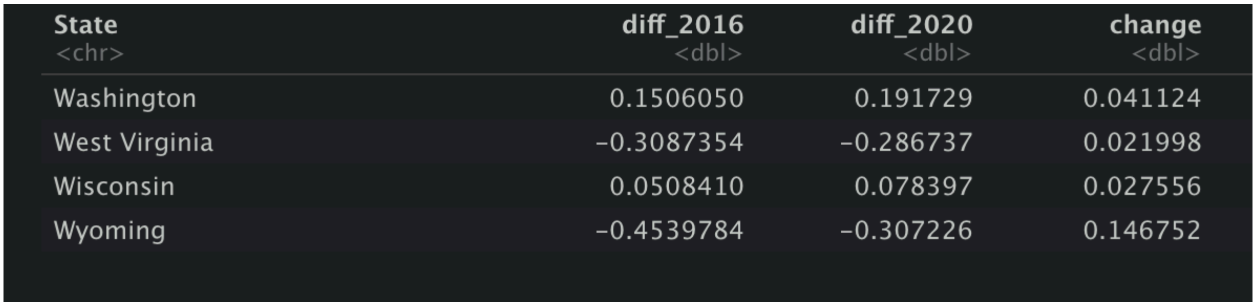
State	diff_2016	diff_2020	change
<chr>	<dbl>	<dbl>	<dbl>
Maine CD−1	0.1587404	−0.124749	−0.283490
Maine CD−2	−0.0487725	−0.013639	0.035134
Maryland	0.2919726	0.307716	0.015744
Massachusetts	0.2697749	0.363140	0.093365
Michigan	0.0331718	0.064305	0.031133
Minnesota	0.0737576	0.103815	0.030058
Mississippi	−0.1625530	−0.159368	0.003185
Missouri	−0.1052606	−0.063369	0.041892
Montana	−0.1835175	−0.046251	0.137267
Nebraska	−0.2228743	−0.064747	0.158127



State	diff_2016	diff_2020	change
<chr>	<dbl>	<dbl>	<dbl>
Nebraska CD−2	−0.0891486	−0.032779	0.056370
Nevada	−0.0068399	0.033040	0.039880
New Hampshire	0.0554345	0.100044	0.044609
New Jersey	0.1235516	0.212747	0.089196
New Mexico	0.0785285	0.070437	−0.008092
New York	0.2031380	0.256378	0.053240
North Carolina	−0.0002762	0.050459	0.050736
North Dakota	−0.2817287	−0.150913	0.130816
Ohio	−0.0354393	−0.019273	0.016166
Oklahoma	−0.2609804	−0.163132	0.097849



State	diff_2016	diff_2020	change
<chr>	<dbl>	<dbl>	<dbl>
Oregon	0.1259675	0.190919	0.064951
Pennsylvania	0.0332186	0.052844	0.019625
Rhode Island	0.1116691	0.315809	0.204140
South Carolina	−0.0828035	−0.056145	0.026658
South Dakota	−0.2207251	−0.115580	0.105146
Tennessee	−0.1827335	−0.111264	0.071469
Texas	−0.1086512	−0.010417	0.098234
Utah	−0.1288485	−0.073939	0.054910
Vermont	0.3238876	0.320433	−0.003454
Virginia	0.0704664	0.121776	0.051310

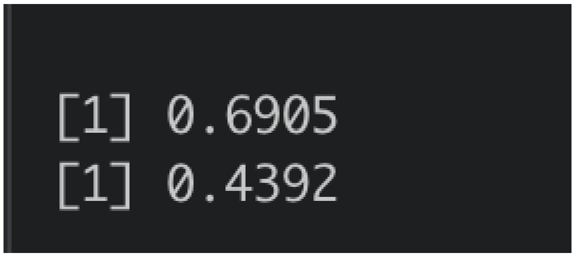


State <chr>	diff_2016 <dbl>	diff_2020 <dbl>	change <dbl>
Washington	0.1506050	0.191729	0.041124
West Virginia	-0.3087354	-0.286737	0.021998
Wisconsin	0.0508410	0.078397	0.027556
Wyoming	-0.4539784	-0.307226	0.146752

b) 5 states that may change their electoral votes in 2020 would be Nebraska, Montana, Georgia, Kansas, and Texas because they had a small 2020 differential but a large change from 2016 to 2020.

(c) As mentioned previously, Georgia would be likely to change their electoral vote. Arizona could flip in a future election due to a low 2020 differential and a still substantial change between the election years. The same is true for Michigan for the same reason. However, Wisconsin and Pennsylvania would be unlikely to flip due to higher 2020 differentials and very low changes between election years. Two other states that could flip in future elections include Kansas and Texas, for the same reasons as stated before.

5. a) Most of the polls in Iowa were accurate to predict the elected candidates (69%), but that was not the case in Florida (44%). Some potential reasons for this inaccuracy in Florida include sampling and weighting issues, as well as non-response bias.



```
[1] 0.6905
[1] 0.4392
```

b) For Iowa, one poll that approximately predicted the final outcome of the election was the Des Moines Register Poll. This poll has a reputation for accurately capturing the preferences of Iowa voters. It is widely recognized as a reliable source of information on Iowa elections.

(c) Some possible reasons accounting for the bias in polls for these two states include demographic composition, differential/variations in voter turnout, as well as volatile and late-deciding voters.

(d) Some possible ways to improve polls for political elections include refining sampling methodologies to ensure a more representative sample of the population. In addition, non-response bias needs to be reduced. Efforts can be made to increase response rates in polls, particularly among groups that are historically less likely to participate. Finally, encouraging collaboration and peer review among polling organizations can promote the sharing of best practices and the identification of potential biases or methodological shortcomings. This can enhance the accuracy of overall predictions.