Data Analytic Report

Tony Chan

2023-05-23

(NOTE: Differences are taken from Democrat(Clinton/Biden) - Republican (Trump) unless otherwised specified)

Question 1:

(20 points) For the presidential poll in 2016, explore the poll in Michigan, Georgia and North Carolina from August 1, 2016 to November 2 in 2016. Use the data to answer the following questions:

(NOTE: Differences are taken from Democrat(Clinton/Biden) - Republican (Trump) unless otherwised specified)

(a) Who is ahead in each of these three states? What is the percentage difference for each state?

```
#Michigan
mi.h = sum(poll.2016data$total.clinton[ind.m])
mi.t = sum(poll.2016data$total.trump[ind.m])
mip.h = mi.h/(mi.h+mi.t)
mip.t = mi.t/(mi.h+mi.t)
#Clinton ahead of Trump
m.diff = abs(mip.h-mip.t)
print(glue("Hillary is ahead of Trump by {m.diff*100}% in Michigan"))
```

```
## Hillary is ahead of Trump by 3.56920068670408% in Michigan
```

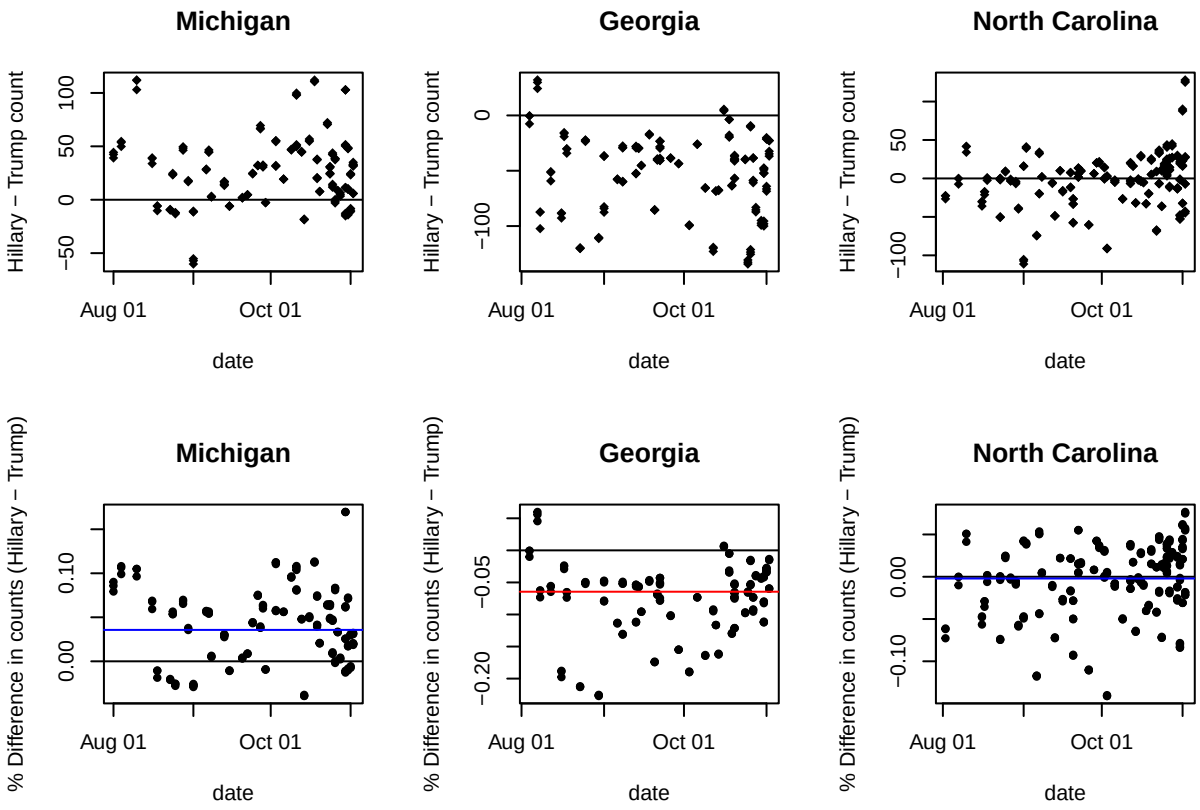
```
#Georgia
geo.h = sum(poll.2016data$total.clinton[ind.g])
geo.t = sum(poll.2016data$total.trump[ind.g])
geop.h = geo.h/(geo.h+geo.t)
geop.t = geo.t/(geo.h+geo.t)
#Trump ahead of Clinton
g.diff = abs(geop.h-geop.t)
print(glue("Trump is ahead of Clinton by {g.diff*100}% in Georgia"))
```

```
## Trump is ahead of Clinton by 6.44982754886921% in Georgia
```

```
#North Carolina
nc.h = sum(poll.2016data$total.clinton[ind.n])
nc.t = sum(poll.2016data$total.trump[ind.n])
ncp.h = nc.h/(nc.h+nc.t)
ncp.t = nc.t/(nc.h+nc.t)
#Clinton ahead of Trump
n.diff = abs(ncp.h-ncp.t)
print(glue("Hillary is ahead of Trump by {n.diff*100}% in North Carolina"))
```

```
## Hillary is ahead of Trump by 0.208550663077828% in North Carolina
```

Raw and Percent Difference between Hilary and Trump count



The figures and data above indicate that Hilary is ahead in Michigan (by 3.57%) and North Carolina (by 0.21%) (represented by positive % diff), while Trump is ahead in Georgia (by 6.45%) (represented by negative % diff). We can also see that her % lead in Michigan is larger then her lead in North Carolina. (% diff is seen by the colored line in the figures above)

(b) Run a paired t test of the counts in polls for each of the state. Who is in favor of winning based on the test? Is the test significant? Is there potential problem?

MICHIGAN: Paired t-test, with the null hypothesis being Clinton having less votes than Trump ($H_0 : \mu_C < \mu_T, H_a : \mu_C \not< \mu_T$).

```
##
##  Paired t-test
##
## data:  poll.2016data$total.clinton[ind.m] and poll.2016data$total.trump[ind.m]
## t = 10.94, df = 176, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  22.57701      Inf
## sample estimates:
## mean of the differences
##                26.59718
```

Paired t-test, with the null hypothesis being Clinton having more votes than Trump ($H_0 : \mu_C > \mu_T, H_a : \mu_C \not> \mu_T$).

```
##
##  Paired t-test
##
## data:  poll.2016data$total.clinton[ind.m] and poll.2016data$total.trump[ind.m]
## t = 10.94, df = 176, p-value = 1
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 30.61736
## sample estimates:
## mean of the differences
##                26.59718
```

GEORGIA:

Paired t-test, with the null hypothesis being Clinton having less votes than Trump ($H_0 : \mu_C < \mu_T, H_a : \mu_C \not< \mu_T$).

```
##
##  Paired t-test
##
## data:  poll.2016data$total.clinton[ind.g] and poll.2016data$total.trump[ind.g]
## t = -19.242, df = 167, p-value = 1
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -55.92167       Inf
## sample estimates:
## mean of the differences
##               -51.49507
```

Paired t-test, with the null hypothesis being Clinton having more votes than Trump ($H_0 : \mu_C > \mu_T, H_a : \mu_C \not> \mu_T$).

```
##
##  Paired t-test
##
## data:  poll.2016data$total.clinton[ind.g] and poll.2016data$total.trump[ind.g]
## t = -19.242, df = 167, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -47.06848
## sample estimates:
## mean of the differences
##               -51.49507
```

NORTH CAROLINA: Paired t-test, with the null hypothesis being Clinton having less votes than Trump ($H_0 : \mu_C < \mu_T, H_a : \mu_C \not> \mu_T$).

```
##
##  Paired t-test
##
## data:  poll.2016data$total.clinton[ind.n] and poll.2016data$total.trump[ind.n]
## t = -0.76542, df = 278, p-value = 0.7777
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -4.904571       Inf
## sample estimates:
## mean of the differences
##               -1.553973
```

Paired t-test, with the null hypothesis being Clinton having more votes than Trump ($H_0 : \mu_C > \mu_T, H_a : \mu_C \not< \mu_T$).

```
##
##  Paired t-test
##
## data:  poll.2016data$total.clinton[ind.n] and poll.2016data$total.trump[ind.n]
## t = -0.76542, df = 278, p-value = 0.2223
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 1.796625
## sample estimates:
## mean of the differences
##               -1.553973
```

Based on these paired t-tests the following conclusions can be made:

- In Michigan, the first test shows an `extremely low p-value (near-zero)` indicating that we should reject the null hypothesis of Trump having more votes. Thus, this test `is significant` and shows that **Clinton is in favor of winning in Michigan**.
- In Georgia, the second test shows an `extremely low p-value (near-zero)` indicating that we should reject the null hypothesis of Clinton having more votes. Thus, this test `is significant` and shows that **Trump is in favor of winning in Georgia**.

- In North Carolina, neither test has a `low enough p-value (less then .05)` to make a significant conclusion. The lowest p-value is the second test, thus it is more likely to reject the null hypothesis of Clinton having more votes. This indicates that **Trump is more favored to win in North Carolina**, however, the test `is NOT significant`.

It is important to note that these paired t-tests are testing the difference in (mean) vote counts between Clinton and Trump. This means that the margin between the vote counts of these two candidates is not being analyzed. Thus, although the tests indicate that Clinton is likely to win Michigan, and Trump is likely to win Georgia, the margin they win by can be small (making it possible for the elections to be very close).

(c) Run a Wilcoxon signed-rank test of the counts in polls for each of the state. Who is in favor of winning based on the test? Is the test significant? Is there potential problem of the test?

MICHIGAN: Wilcoxon signed-rank test for null hypothesis being Clinton having less votes than Trump ($H_0 : median_C < median_T$)

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  poll.2016data$total.clinton[ind.m] and poll.2016data$total.trump[ind.m]
## V = 14041, p-value < 2.2e-16
## alternative hypothesis: true location shift is greater than 0
```

Wilcoxon signed-rank test for null hypothesis being Clinton having more votes than Trump ($H_0 : median_C > median_T$)

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  poll.2016data$total.clinton[ind.m] and poll.2016data$total.trump[ind.m]
## V = 14041, p-value = 1
## alternative hypothesis: true location shift is less than 0
```

GEORGIA: Wilcoxon signed-rank test for null hypothesis being Clinton having less votes than Trump ($H_0 : median_C < median_T$)

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  poll.2016data$total.clinton[ind.g] and poll.2016data$total.trump[ind.g]
## V = 158, p-value = 1
## alternative hypothesis: true location shift is greater than 0
```

Wilcoxon signed-rank test for null hypothesis being Clinton having more votes than Trump ($H_0 : median_C > median_T$)

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  poll.2016data$total.clinton[ind.g] and poll.2016data$total.trump[ind.g]
## V = 158, p-value < 2.2e-16
## alternative hypothesis: true location shift is less than 0
```

NORTH CAROLINA: Wilcoxon signed-rank test for null hypothesis being Clinton having less votes than Trump ($H_0 : median_C < median_T$)

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  poll.2016data$total.clinton[ind.n] and poll.2016data$total.trump[ind.n]
## V = 19635, p-value = 0.4691
## alternative hypothesis: true location shift is greater than 0
```

Wilcoxon signed-rank test for null hypothesis being Clinton having more votes than Trump ($H_0 : median_C > median_T$)

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  poll.2016data$total.clinton[ind.n] and poll.2016data$total.trump[ind.n]
## V = 19635, p-value = 0.5312
## alternative hypothesis: true location shift is less than 0
```

Based on these Wilcoxon signed-rank tests the following conclusions can be made:

- In Michigan the first test shows a `near-zero p-value` indicating that we should reject the null hypothesis of Clinton having less votes than Trump. Thus the test `is significant` and shows that **Clinton is in favor of winning in Michigan**.
- In Georgia the second test shows a `near-zero p-value` indicating that we should reject the null hypothesis of Clinton having more votes than Trump. Thus the test `is significant` and shows that **Trump is in favor of winning in Georgia**.
- In North Carolina, `neither test has a low enough p-value (less then .05)` to make a significant conclusion. The lowest p-value is the first test, thus it is more likely to reject the null hypothesis of Clinton having less votes. This indicates that Clinton is more favored to win in North Carolina, however, the test is NOT significant.

The results from the paired t-test and Wilcoxon signed-rank test for North Carolina offer opposite conclusions. It is important to recognize that neither test has a low enough p-value to make any significant conclusion.

It is important to note that these Wilcoxon signed-rank tests are testing the difference in (median) vote counts between Clinton and Trump. This means that the margin between the vote counts of these two candidates is not being analyzed. Thus, although the tests indicate that Clinton is likely to win Michigan, and Trump is likely to win Georgia, the margin they win by can be small (making it possible for the elections to be very close).

(d) Fit a linear model of the percentage difference with respect to date of the polls separately for each of these states. Show a plot of the observations of the polls, fitted values and confidence interval of the fitted line for each of these state. From the linear model and observations, which state may have the closest election (in terms of percentage difference)?

```
## Warning in poll.2016data$total.clinton[ind.m] + poll.2016data$total.trump:
## longer object length is not a multiple of shorter object length
```

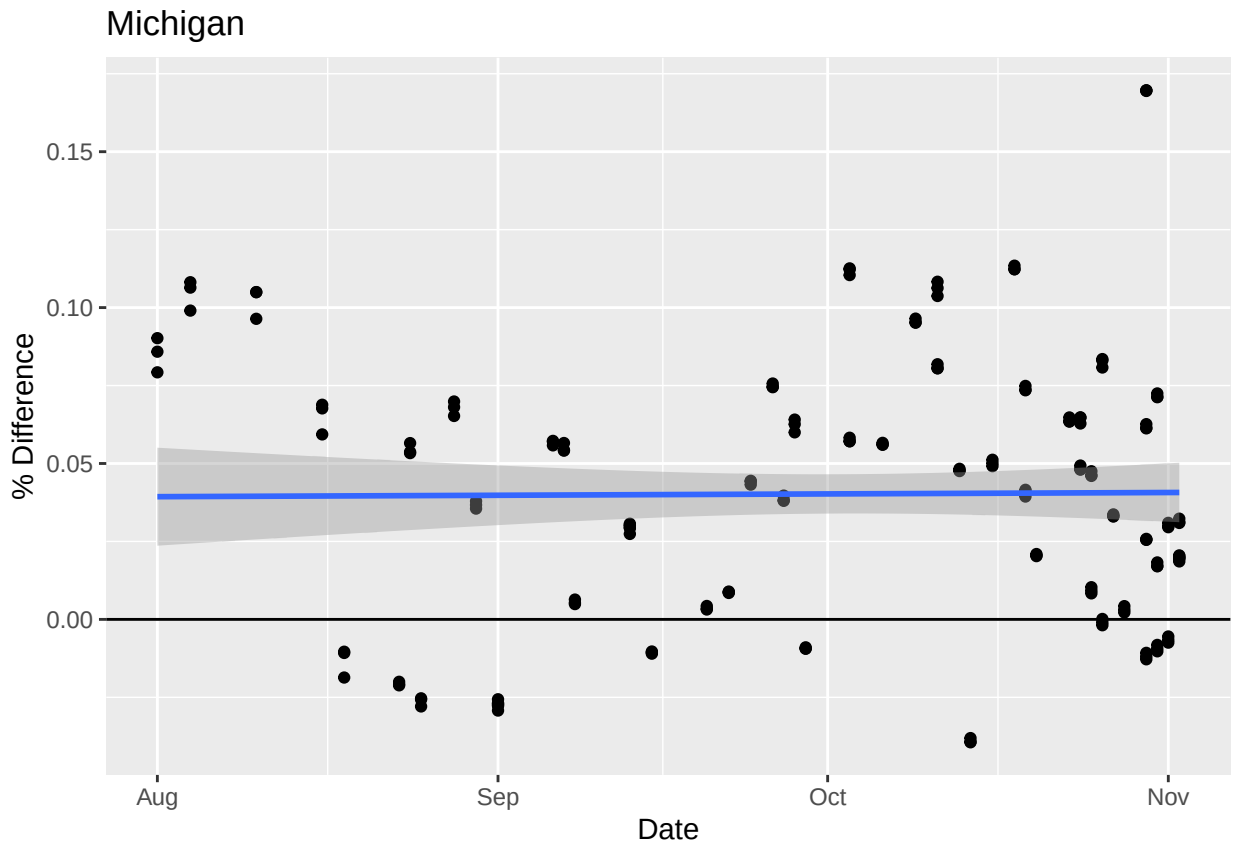
MICHIGAN:

```
m.dat <- dat %>% filter(state == 'Michigan')
mod <- lm(data = m.dat, pdiff ~ date)
summary(mod)
```

```
##
## Call:
## lm(formula = pdiff ~ date, data = m.dat)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.079852 -0.033555  0.000666  0.024237  0.129024
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.084e-01  2.005e+00  -0.104    0.917
## date         1.456e-05  1.174e-04   0.124    0.901
##
## Residual standard error: 0.04245 on 175 degrees of freedom
## Multiple R-squared:  8.784e-05,  Adjusted R-squared:  -0.005626
## F-statistic: 0.01537 on 1 and 175 DF,  p-value: 0.9015
```

```
ggplot(m.dat, aes(x = date, y = pdiff)) +
  geom_point() +
  stat_smooth(method = 'lm') +
  ggtitle("Michigan LM") +
  labs(x = 'Date', y = '% Difference', title = 'Michigan') +
  geom_hline(yintercept=0)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



GEORGIA:

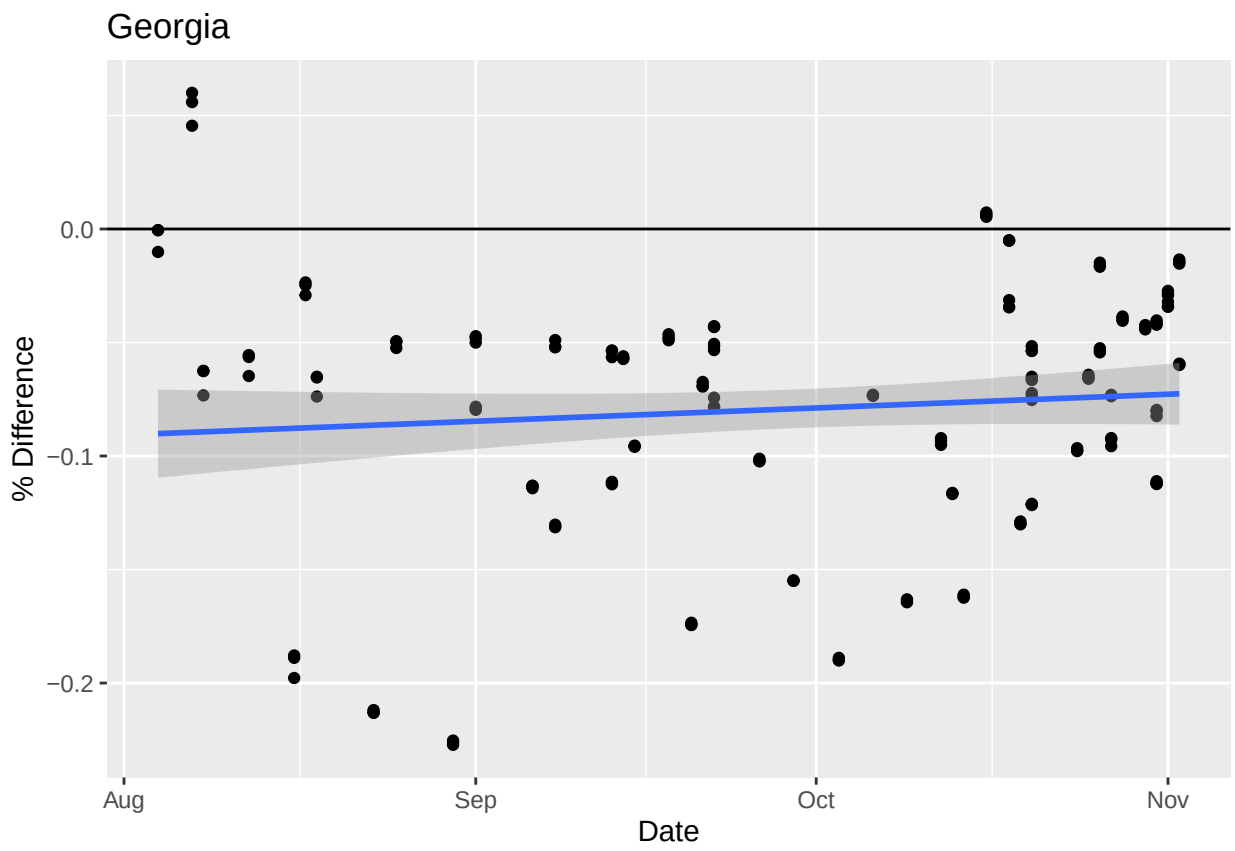
```
g.dat <- dat %>% filter(state == 'Georgia')
mod <- lm(data = g.dat, pdiff ~ date)
summary(mod)
```

```
##
## Call:
## lm(formula = pdiff ~ date, data = g.dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.14213 -0.02947  0.01069  0.03370  0.14951
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.4065220  2.6898032  -1.266    0.207
## date         0.0001949  0.0001575   1.237    0.218
##
## Residual standard error: 0.05589 on 166 degrees of freedom
## Multiple R-squared:  0.009134,   Adjusted R-squared:  0.003165
## F-statistic:  1.53 on 1 and 166 DF,  p-value: 0.2178
```

```
#plot(predict(mod))

ggplot(g.dat, aes(x = date, y = pdiff)) +
```

```
  geom_point() +
  stat_smooth(method = 'lm') +
  ggtitle("Georgia LM") +
  labs(x = 'Date', y = '% Difference', title = 'Georgia') +
  geom_hline(yintercept=0)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



NORTH CAROLINA:

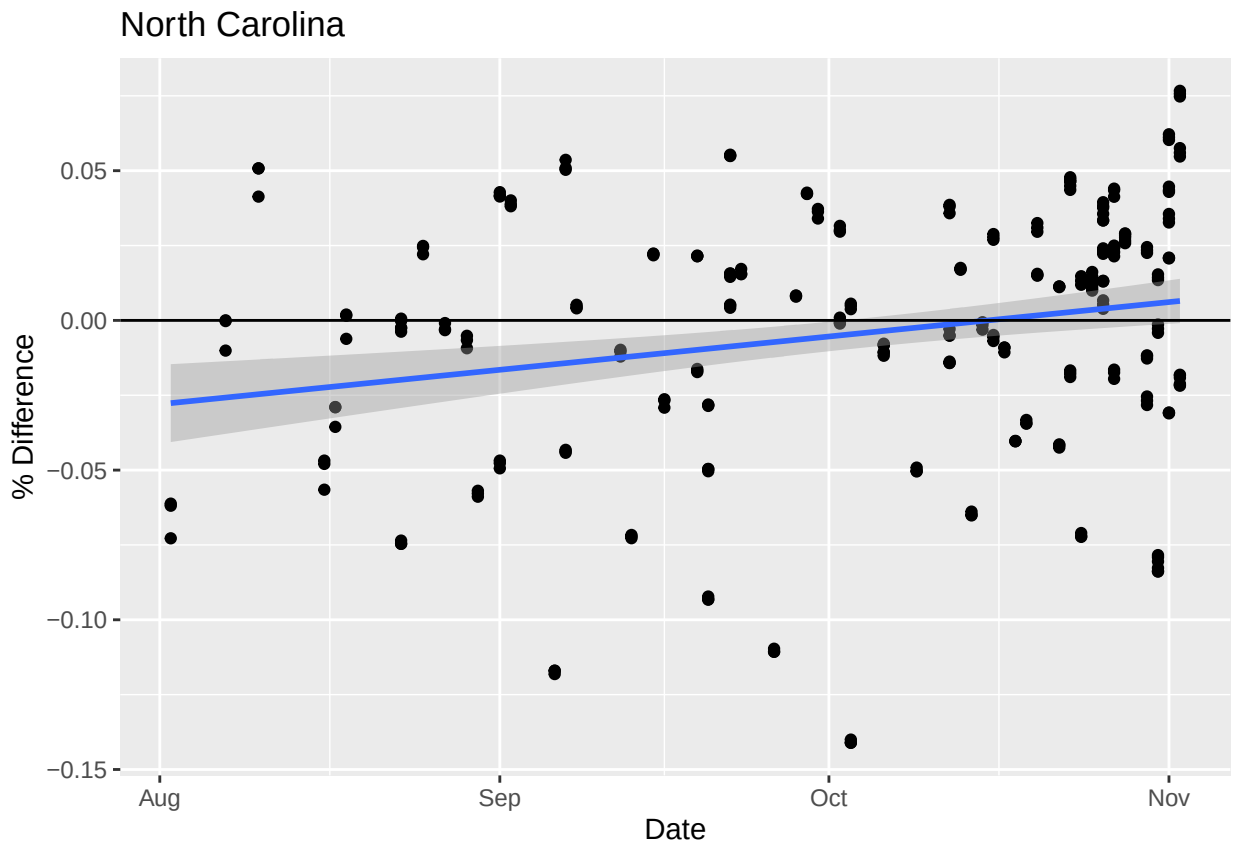
```
n.dat <- dat %>% filter(state == 'North Carolina')
mod <- lm(data = n.dat, pdiff ~ date)
summary(mod)
```

```
##
## Call:
## lm(formula = pdiff ~ date, data = n.dat)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.136454 -0.024750  0.009315  0.027180  0.075461
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.3366420  1.6599641  -3.817 0.000166 ***
```

```
## date         0.0003708  0.0000972   3.815 0.000168 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04153 on 277 degrees of freedom
## Multiple R-squared:  0.04991,    Adjusted R-squared:  0.04648
## F-statistic: 14.55 on 1 and 277 DF,  p-value: 0.0001681
```

```
ggplot(n.dat, aes(x = date, y = pdiff)) +
  geom_point() +
  stat_smooth(method = 'lm') +
  ggtitle("North Carolina") +
  labs(x = 'Date', y = '% Difference', title = 'North Carolina') +
  geom_hline(yintercept=0)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



In terms of the percentage difference, the linear models and their graphs indicates that North **Carolina may have the closest election**. This is due to the fact that the confidence interval for the fit includes 0, while the fitted values for Michigan and Georgia do not include 0. 0 represents a zero % difference between Clinton and Trump indicating that the polls show that the two candidates are very close contenders for the elections.

(e) From the real results of 2016 election, which state has the smallest margin (in terms of percentage difference)? Discuss at least two reasons that are different than what polls indicate. (You may check Wikipedia for 2016 US presidential election to find out the real voting results for each state.)

According to wikipedia:

```
##                Clinton % Trump % Real % diff Poll % diff
## Michigan           47.27   47.50       -0.23        3.75
## Georgia            45.64   49.83       -4.19       -5.56
## North Carolina     46.17   50.77       -4.60        0.22
```

Based on these results, the state with the smallest % difference in reality is Michigan. The difference in real % diff and the % diff seen from the polls can possibly be explained by flaws in the polling, and a flawed hypothesis test. The polls we based our test from may have been obtained poorly, they may not represent the population well due to bias, inconsistencies, and discrepancies in polling. Additionally, the hypothesis testing itself focused on who would have a greater amount of votes between Hillary Clinton and Donald Trump. This disregards the data in the polls relating to any other candidate and tests who has more votes when you can only choose candidates, which is not representative of reality. Another factor that may contribute to the failure of polls is that the elections is an ongoing process with constantly changing public opinion regarding candidates. Polls themselves are limited in how they can account for changing partisan activities, as well as the changing values of the poll participation.

(f) Do polls correctly predict the candidate who wins these states? Discuss the bias of polls in these states. Name a few possible reasons.

Clearly the 2016 polls do not correctly predict which candidates will win in which state. This is likely due to the fact that the polls do not properly represent the population of their respective states. This can arise from a plethora of reasons, ranging from the participants of the polls hiding their opinion, to the inconsistent variables of who is being polled or biased polls. For example, participation bias may result in difference between the sample and population because certain groups of people tend to participate more or avoid participating. Not only can the participants be biased, the sampling itself can be biased by being more likely to send polls to certain groups of individuals over others.

Question 2: (20 points) Redo Question 1 (a)-(f) for the same three states for the presidential polls in from August 1 to November 2 in 2020. (You may check Wikipedia for 2020 US presidential election to find out the real voting results for each state.)

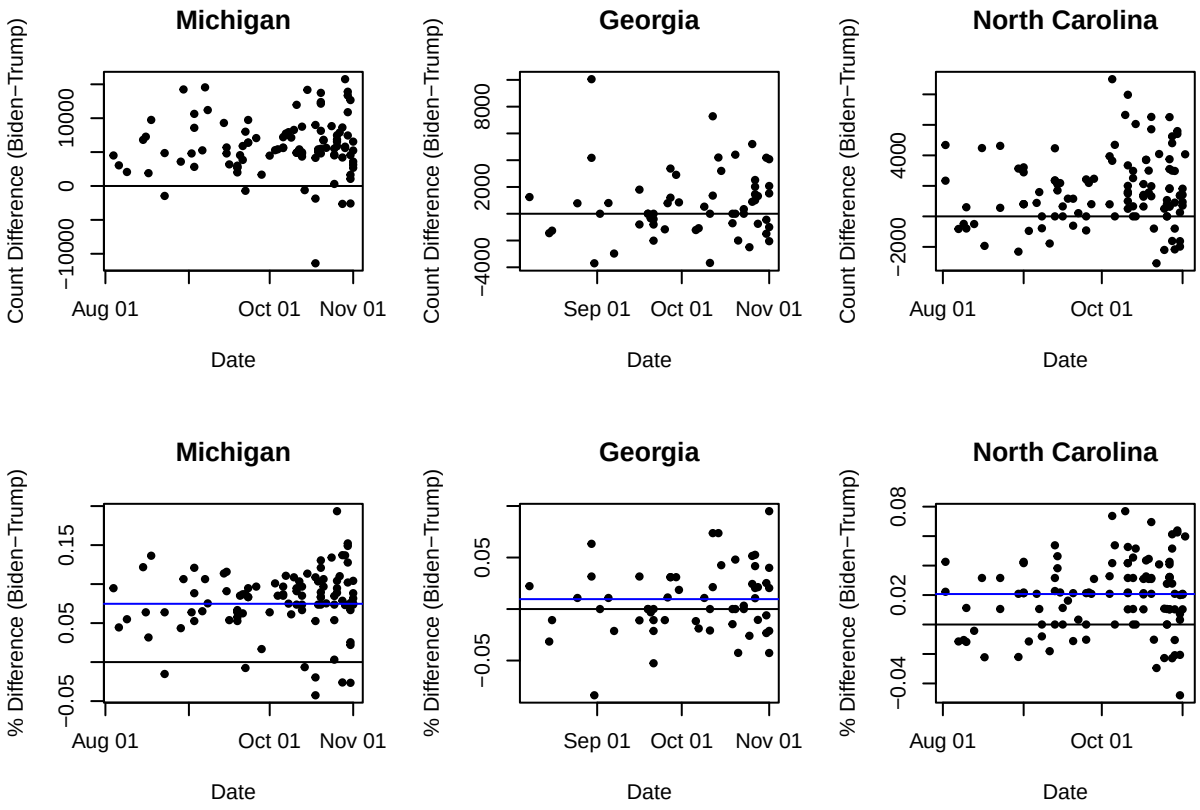
```
## [1] "\npoll.2016data <- poll.2016data %>% mutate(enddates = mdy(enddate))\npoll.2016data <- poll.201
```

(a) Who is ahead in each of these three states? What is the percentage difference for each state?

```
## Biden is ahead of trump by 7.4756484691515% in Michigan
```

```
## Biden is ahead of trump by 0.963109843582636% in Georgia
```

```
## Biden is ahead of trump by 2.06757157309794% in North Carolina

## integer(0)

## integer(0)

## integer(0)

## integer(0)

## integer(0)
```




```
## integer(0)
```

The figures and data above indicate that Biden is ahead in Michigan (by 7.48%), Georgia (by 0.96%), and North Carolina (by 2.01%) (positive % diff represents Biden being ahead).

(b) Run a paired t test of the counts in polls for each of the state. Who is in favor of winning based on the test? Is the test significant? Is there potential problem?

MICHIGAN: Paired t-test, with the null hypothesis being Biden having less votes than Trump ($H_0 : \mu_B < \mu_T, H_a : \mu_B \not< \mu_T$).

```
##
##  Paired t-test
##
## data:  b.m and t.m
## t = 14.133, df = 100, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  5198.639      Inf
## sample estimates:
## mean of the differences
##                5890.607
```

Paired t-test, with the null hypothesis being Biden having more votes than Trump ($H_0 : \mu_B > \mu_T, H_a : \mu_B \not> \mu_T$).

```
##
##  Paired t-test
##
## data:  b.m and t.m
## t = 14.133, df = 100, p-value = 1
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 6582.575
## sample estimates:
## mean of the differences
##                5890.607
```

GEORGIA: Paired t-test, with the null hypothesis being Biden having less votes than Trump ($H_0 : \mu_B < \mu_T, H_a : \mu_B \not< \mu_T$).

```
##
##  Paired t-test
##
## data:  b.g and t.g
## t = 2.257, df = 57, p-value = 0.01393
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  195.4117      Inf
## sample estimates:
## mean of the differences
##                753.9417
```

Paired t-test, with the null hypothesis being Biden having more votes than Trump ($H_0 : \mu_B > \mu_T, H_a : \mu_B \not> \mu_T$).

```
##
##  Paired t-test
##
## data:  b.g and t.g
## t = 2.257, df = 57, p-value = 0.9861
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
```

```
##      -Inf 1312.472
## sample estimates:
## mean of the differences
##                753.9417
```

NORTH CAROLINA: Paired t-test, with the null hypothesis being Biden having less votes than Trump ($H_0 : \mu_B < \mu_T, H_a : \mu_B \not< \mu_T$).

```
##
##  Paired t-test
##
## data:  b.n and t.n
## t = 7.7276, df = 110, p-value = 2.76e-12
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1329.498      Inf
## sample estimates:
## mean of the differences
##                1692.897
```

Paired t-test, with the null hypothesis being Biden having more votes than Trump ($H_0 : \mu_B > \mu_T, H_a : \mu_B \not> \mu_T$).

```
##
##  Paired t-test
##
## data:  b.n and t.n
## t = 7.7276, df = 110, p-value = 1
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf 2056.296
## sample estimates:
## mean of the differences
##                1692.897
```

Based on these paired t-tests the following conclusions can be made:

- In Michigan, the first test shows an `extremely low p-value (near-zero)` indicating that we should reject the null hypothesis of Trump having more votes. Thus, this test `is significant` and shows that **Biden is in favor of winning in Michigan**.
- In Georgia, the first test shows a `low p-value (less then 05)` indicating that we should reject the null hypothesis of Trump having more votes. Thus, this test `is significant` and shows that **Biden is in favor of winning in Georgia**.
- In North Carolina, the first test shows an `extremely low p-value (near-zero)` indicating that we should reject the null hypothesis of Trump having more votes. Thus, this test `is significant` and shows that **Biden is in favor of winning in North Carolina**.

It is important to note that these paired t-tests are testing the difference in (mean) vote counts between Biden and Trump. This means that the margin between the vote counts (%diff) of these two candidates is not being analyzed. Thus, although the tests indicate that Biden is likely to win Michigan, Georgia, and North Carolina the margin they win by can be small (making it possible for the elections to be very close).

(c) Run a Wilcoxon signed-rank test of the counts in polls for each of the state. Who is in favor of winning based on the test? Is the test significant? Is there potential problem of the test?

MICHIGAN: Wilcoxon signed-rank test for null hypothesis being Biden having less votes than Trump ($H_0 : median_C < median_T$)

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  b.m and t.m
## V = 5018, p-value < 2.2e-16
## alternative hypothesis: true location shift is greater than 0
```

Wilcoxon signed-rank test for null hypothesis being Biden having more votes than Trump ($H_0 : median_C > median_T$)

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  b.m and t.m
## V = 5018, p-value = 1
## alternative hypothesis: true location shift is less than 0
```

GEORGIA: Wilcoxon signed-rank test for null hypothesis being Biden having less votes than Trump ($H_0 : median_C < median_T$)

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  b.g and t.g
## V = 836.5, p-value = 0.02767
## alternative hypothesis: true location shift is greater than 0
```

Wilcoxon signed-rank test for null hypothesis being Biden having more votes than Trump ($H_0 : median_C > median_T$)

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  b.g and t.g
## V = 836.5, p-value = 0.9729
## alternative hypothesis: true location shift is less than 0
```

NORTH CAROLINA: Wilcoxon signed-rank test for null hypothesis being Biden having less votes than Trump ($H_0 : median_C < median_T$)

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  b.n and t.n
## V = 4566.5, p-value = 4.757e-11
## alternative hypothesis: true location shift is greater than 0
```

Wilcoxon signed-rank test for null hypothesis being Biden having more votes than Trump ($H_0 : median_C > median_T$)

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  b.n and t.n
## V = 4566.5, p-value = 1
## alternative hypothesis: true location shift is less than 0
```

Based on these Wilcoxon signed-rank tests the following conclusions can be made:

- In Michigan the first test shows a `near-zero p-value` indicating that we should reject the null hypothesis of Biden having less votes than Trump. Thus the test `is significant` and shows that **Biden is in favor of winning in Michigan**.
- In Georgia the first test shows a `low enough p-value (less then .05)` that we should reject the null hypothesis of Biden having less votes than Trump. Thus the test `is significant` and shows that **Biden is in favor of winning in Georgia**.
- In North Carolina, the first test shows a `near-zero p-value` indicating that we should reject the null hypothesis of Biden having less votes than Trump. Thus the test `is significant` and shows that **Biden is in favor of winning in Michigan**.

It is important to note that these Wilcoxon signed-rank tests are testing the difference in (median) vote counts between Clinton and Trump. This means that the margin between the vote counts of these two candidates is not being analyzed. Thus, although the tests indicate that Biden is likely to win Michigan, Georgia, and North Carolina, the margin Biden wins by can be small (making it possible for the elections to be very close).

(d) Fit a linear model of the percentage difference with respect to date of the polls separately for each of these states. Show a plot of the observations of the polls, fitted values and confidence interval of the fitted line for each of these state. From the linear model and observations, which state may have the closest election (in terms of percentage difference)?

MICHIGAN:

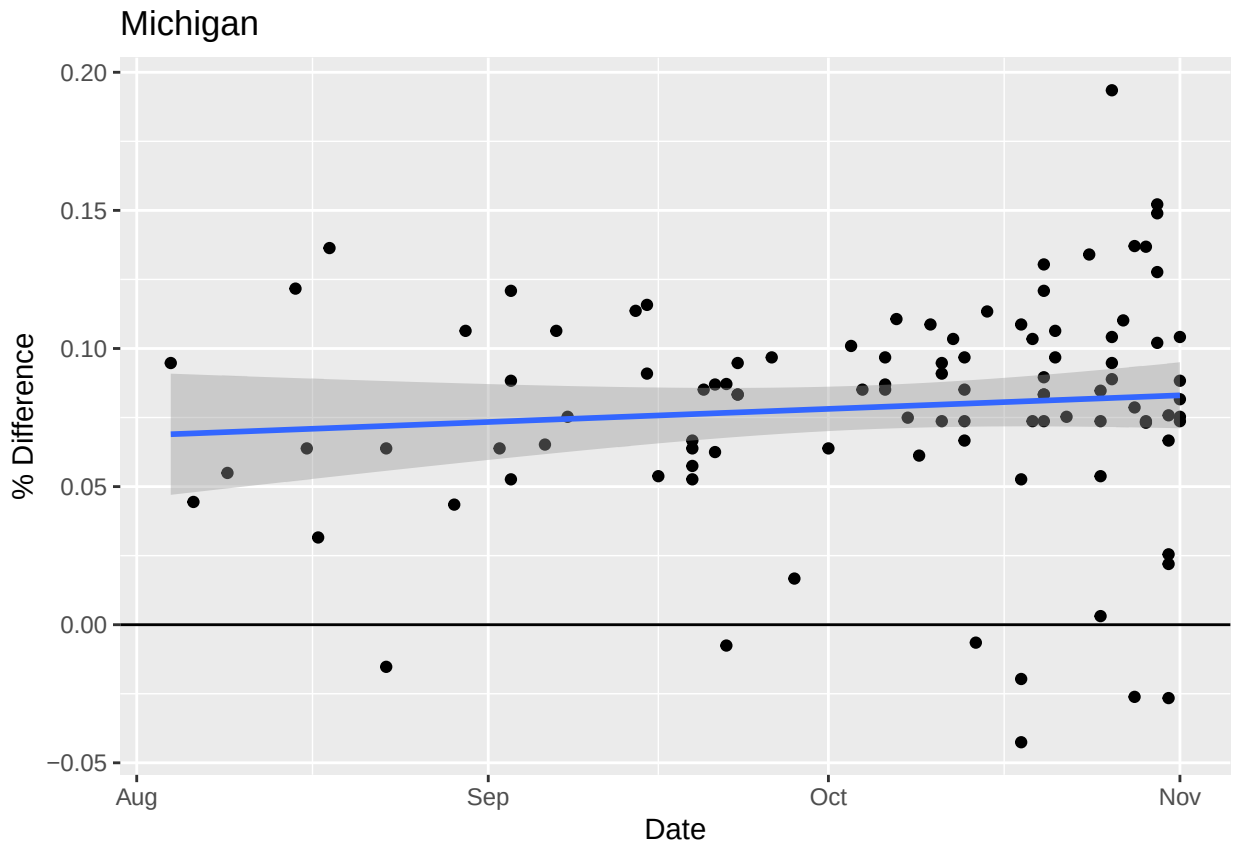
```
m.df <- data.frame(date = poll.2020data$end_date[ind.t.m] , pdiff = (b.m-t.m) / (b.m+t.m))
m.mod <- lm(data = m.df, pdiff ~ date)
summary(m.mod)
```

```
##
## Call:
## lm(formula = pdiff ~ date, data = m.df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.123375 -0.013364  0.005075  0.022468  0.111395
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.8548882  3.0804431  -0.927    0.356
## date         0.0001582  0.0001662   0.952    0.343
##
## Residual standard error: 0.04017 on 99 degrees of freedom
## Multiple R-squared:  0.009078,   Adjusted R-squared:  -0.0009312
## F-statistic: 0.907 on 1 and 99 DF,  p-value: 0.3432
```

```
ggplot(m.mod, aes(x = date, y = pdiff)) +
  geom_point() +
  stat_smooth(method = 'lm') +
  ggtitle("Michigan LM") +
  labs(x = 'Date', y = '% Difference', title = 'Michigan') +
  geom_hline(yintercept=0)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



GEORGIA:

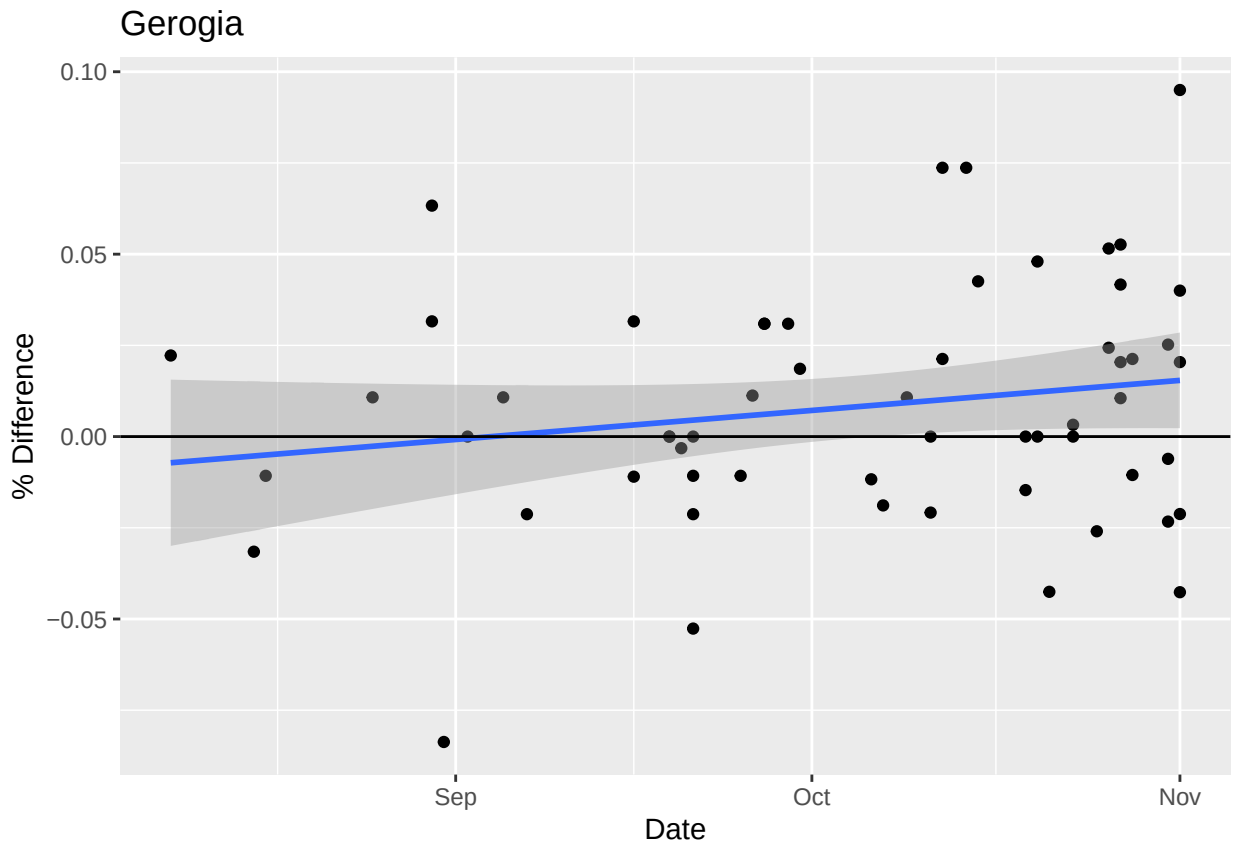
```
g.df <- data.frame(date = poll.2020data$end_date[ind.t.g] , pdiff = (b.g-t.g) / (b.g+t.g))
g.mod <- lm(data = g.df, pdiff ~ date)
summary(g.mod)
```

```
##
## Call:
```

```
## lm(formula = pdiff ~ date, data = g.df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.082642 -0.020989 -0.003764  0.024520  0.079591
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.9175032  3.3864614  -1.452    0.152
## date         0.0002657  0.0001827   1.455    0.151
##
## Residual standard error: 0.03237 on 56 degrees of freedom
## Multiple R-squared:  0.0364, Adjusted R-squared:  0.0192
## F-statistic: 2.116 on 1 and 56 DF,  p-value: 0.1514
```

```
ggplot(g.mod, aes(x = date, y = pdiff)) +
  geom_point() +
  stat_smooth(method = 'lm') +
  ggtitle("Georgia LM") +
  labs(x = 'Date', y = '% Difference', title = 'Gerogia') +
  geom_hline(yintercept=0)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



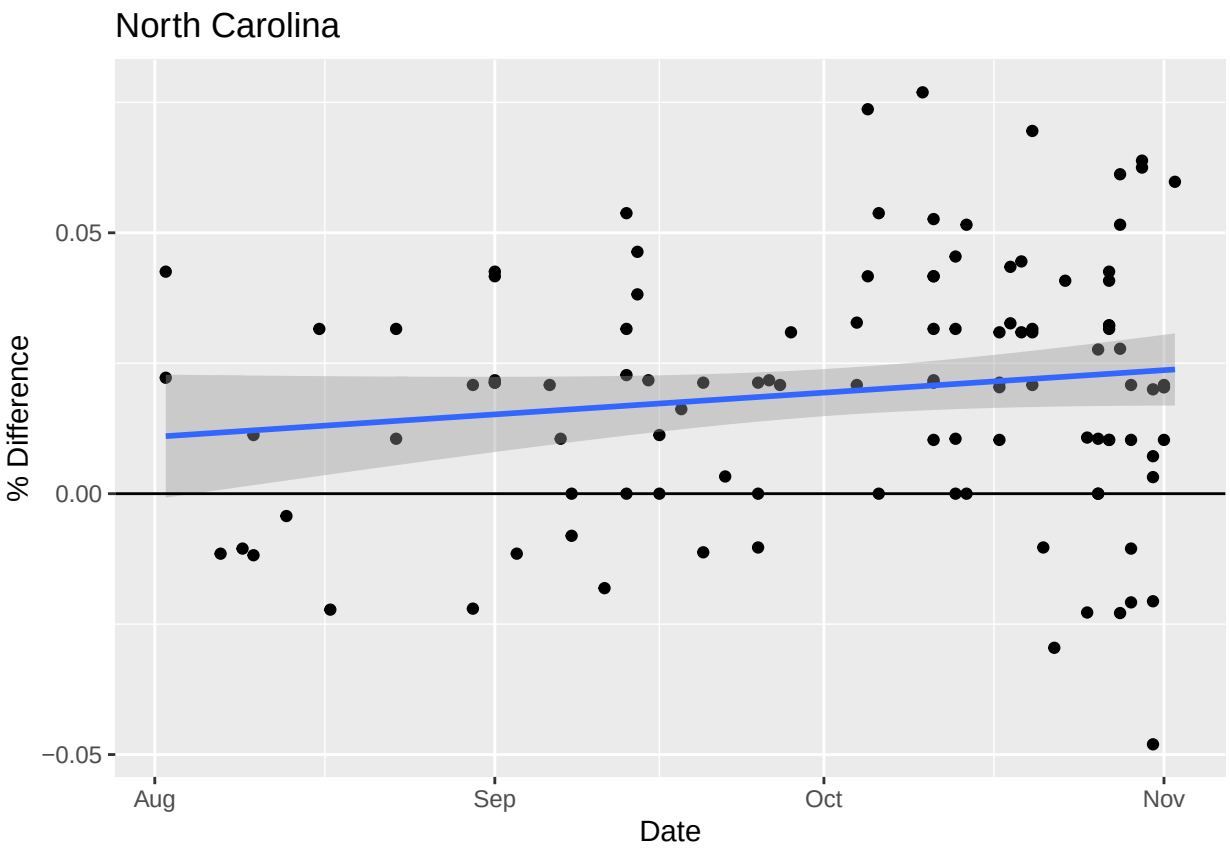
NORTH CAROLINA:

```
n.df <- data.frame(date = poll.2020data$end_date[ind.t.n] , pdiff = (b.n-t.n) / (b.n+t.n))
n.mod <- lm(data = n.df, pdiff ~ date)
summary(n.mod)
```

```
##
## Call:
## lm(formula = pdiff ~ date, data = n.df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.071566 -0.016594  0.001054  0.016176  0.056310
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.556e+00  1.641e+00  -1.558    0.122
## date         1.389e-04  8.852e-05   1.570    0.119
##
## Residual standard error: 0.02393 on 109 degrees of freedom
## Multiple R-squared:  0.0221, Adjusted R-squared:  0.01313
## F-statistic: 2.464 on 1 and 109 DF,  p-value: 0.1194
```

```
ggplot(n.mod, aes(x = date, y = pdiff)) +
  geom_point() +
  stat_smooth(method = 'lm') +
  ggtitle("North Carolina LM") +
  labs(x = 'Date', y = '% Difference', title = 'North Carolina') +
  geom_hline(yintercept=0)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



In terms of the percentage difference, the linear models and their graphs indicates that Georgia may have the closest election. This is due to the fact that the confidence interval for the fit includes 0, while the fitted values for Michigan does not include 0, and Georgia's CI includes zero to a much smaller degree. 0 represents a zero % difference between Biden and Trump indicating that the polls show that the two candidates are very close contenders for the election.

(e) From the real results of 2020 election, which state has the smallest margin (in terms of percentage difference)? Discuss at least two reasons that are different than what polls indicate. (You may check Wikipedia for 2020 US presidential election to find out the real voting results for each state.)

According to wikipedia:

```
##                Biden % Trump % Real % diff Poll % diff
## Michigan         50.62   47.84        2.78        7.48
## Georgia          49.47   49.24        0.23        0.96
## North Carolina   48.59   49.93       -1.34        2.07
```

Based on these results, the state with the smallest % difference is Georgia. The polls correctly predicted the state with the closest election. The polls were wrong in predicting the winner in North Carolina. The polls we based our test from may have been obtained poorly, they may not represent the population well due to bias, inconsistencies, and discrepancies in polling. Additionally, the hypothesis testing itself focused on who would have a greater amount of votes between Biden and Donald Trump. This disregards the data in the polls relating to any other candidate and tests who has more votes when you can only choose candidates, which is not representative of reality. Another factor that may contribute to the failure of polls is that the elections

is an ongoing process with constantly changing public opinion regarding candidates. Polls themselves are limited in how they can account for changing partisan activities, as well as the changing values of the poll participation.

(f) Do polls correctly predict the candidate who wins these states? Discuss the bias of polls in these states. Name a few possible reasons.

The 2020 Polls did a decent job at predicting which candidates would win in each state. There were still some mistakes which could be The polls we based our test from may have been obtained poorly, they may not represent the population well due to bias, inconsistencies, and discrepancies in polling. Additionally, the hypothesis testing itself focused on who would have a greater amount of votes between Hillary Clinton and Donald Trump. This disregards the data in the polls relating to any other candidate and tests who has more votes when you can only choose candidates, which is not representative of reality. Another factor that may contribute to the failure of polls is that the elections is an ongoing process with constantly changing public opinion regarding candidates. Polls themselves are limited in how they can account for changing partisan activities, as well as the changing values of the poll participation.

Q-3 Explore the poll data from September 1, 2016 to November 2, 2016 and September 1, 2020 to November 2, 2020 to answer the following questions.

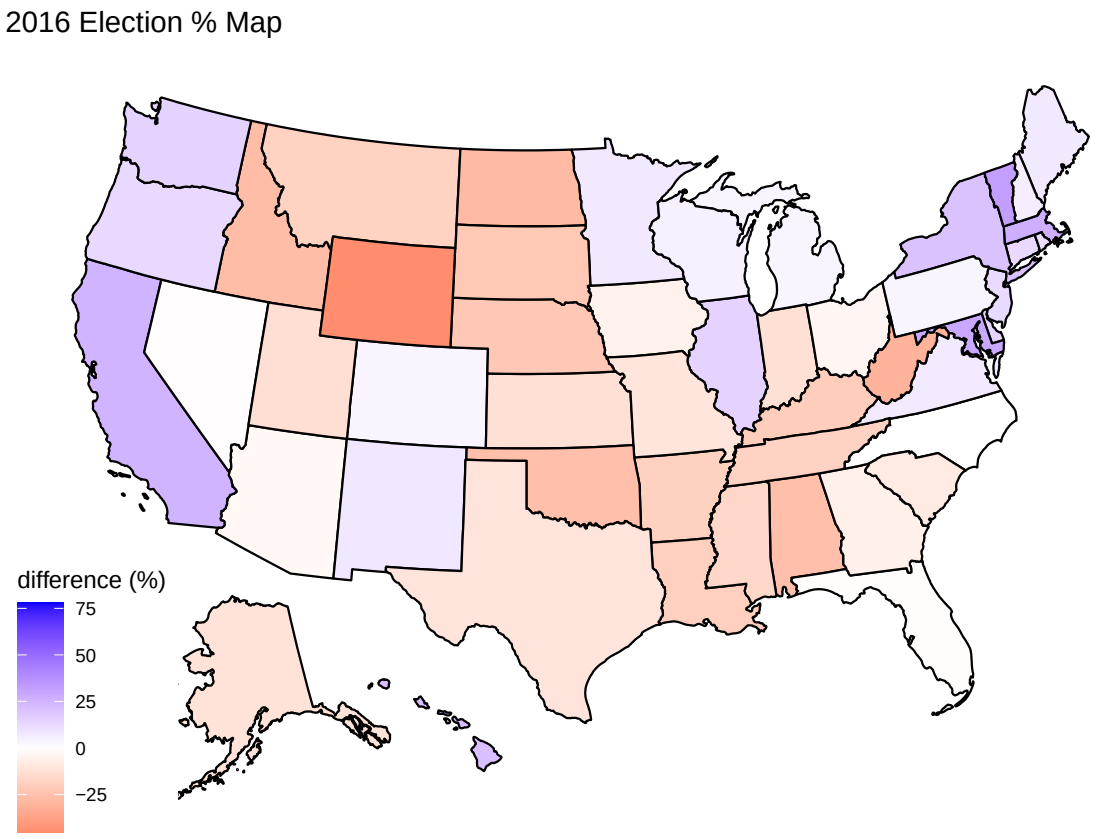
(a) Graph the percentage difference of polls in each state of US for 2016 and 2020. Compare the difference.

\begin{center} \textbf{2016-2020 State % Difference Table} \end{center}

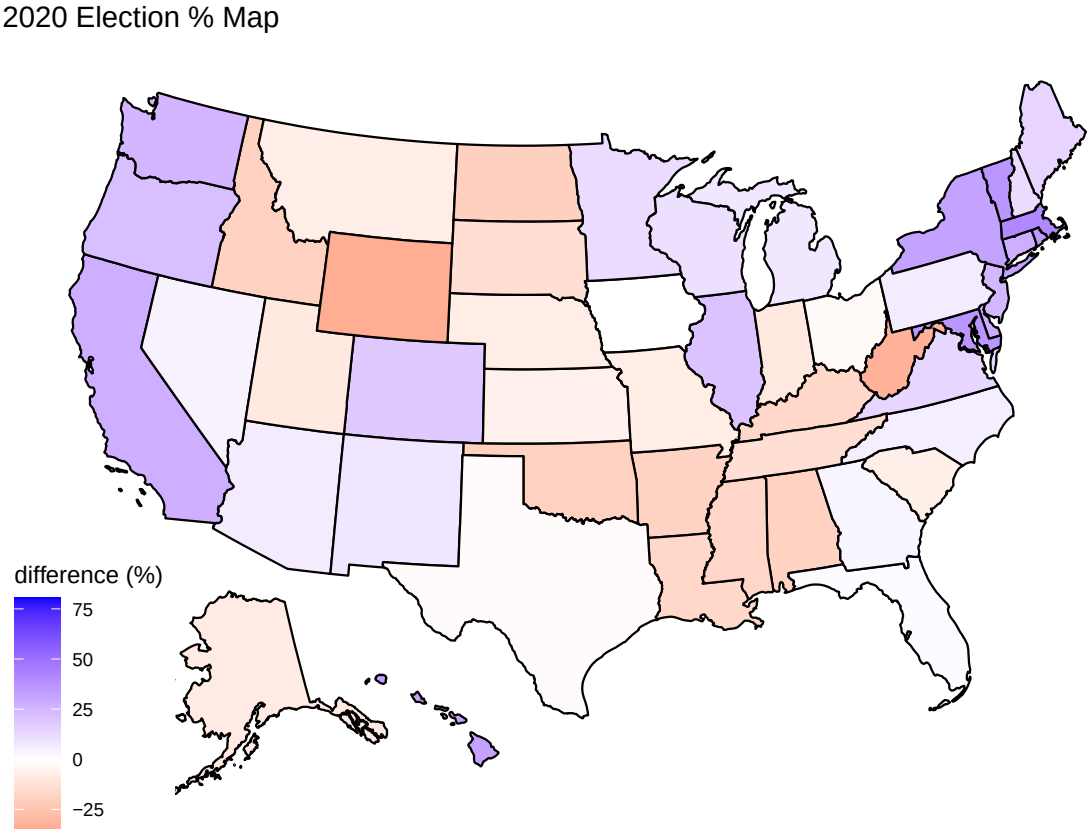
	State	2016 % diff	2020 % diff	2016 to 2020 change
1	Alabama	-26.03640239	-18.967223	-7.06917911
2	Alaska	-11.66978248	-8.466445	-3.20333725
3	Arizona	-2.84276819	6.722131	-9.56489901
4	Arkansas	-18.72792611	-18.692669	-0.03525737
5	California	25.09696425	27.453997	-2.35703266
6	Colorado	3.55450565	18.445174	-14.89066820
7	Connecticut	12.55773444	27.864740	-15.30700568
8	Delaware	14.00220485	28.107698	-14.10549316
9	District of Columbia	78.00263765	80.382676	-2.38003823
10	Florida	-0.85880347	1.261647	-2.12045014
11	Georgia	-6.23845603	3.224546	-9.46300180
12	Hawaii	21.29048167	31.985467	-10.69498543
13	Idaho	-27.30037104	-19.275627	-8.02474443
14	Illinois	14.84783958	20.575001	-5.72716096
15	Indiana	-12.79683401	-9.746841	-3.04999310
16	Iowa	-5.25366811	-0.129796	-5.12387208
17	Kansas	-12.70783255	-6.269227	-6.43860579
18	Kentucky	-20.48601335	-16.442821	-4.04319213
19	Louisiana	-19.46778640	-16.558465	-2.90932128
20	Maine	7.40846988	14.261571	-6.85310146
23	Maryland	29.19725687	36.241738	-7.04448076
24	Massachusetts	26.97748661	41.271098	-14.29361153
25	Michigan	3.31718461	8.304549	-4.98736475
26	Minnesota	7.37575565	12.464149	-5.08839307
27	Mississippi	-16.25529595	-17.043113	0.78781693
28	Missouri	-10.52605877	-7.564304	-2.96175485
29	Montana	-18.35174544	-7.447978	-10.90376734
30	Nebraska	-22.28743243	-7.633873	-14.65355899
34	Nevada	-0.68398929	4.457050	-5.14103959
35	New Hampshire	5.54344853	12.316213	-6.77276425
36	New Jersey	12.35516052	24.822972	-12.46781150
37	New Mexico	7.85285326	8.537469	-0.68461553
38	New York	20.31379551	32.175511	-11.86171582
39	North Carolina	-0.02762169	5.760768	-5.78838949
40	North Dakota	-28.17286598	-19.737535	-8.43533106
41	Ohio	-3.54393443	-2.464757	-1.07917712
42	Oklahoma	-26.09803877	-19.385119	-6.71291945
43	Oregon	12.59674612	22.022765	-9.42601854
44	Pennsylvania	3.32186471	6.698433	-3.37656856
45	Rhode Island	11.16691429	35.097337	-23.93042257
46	South Carolina	-8.28035285	-6.863415	-1.41693764
47	South Dakota	-22.07251049	-13.385534	-8.68697658

```
## 48            Tennessee -18.27335302  -13.282921         -4.99043186
## 49                Texas -10.86512125   -1.484116         -9.38100491
## 51                 Utah -12.88484691   -9.366719         -3.51812811
## 52              Vermont  32.38876188   36.322657         -3.93389538
## 53             Virginia   7.04663931   14.049153         -7.00251377
## 54           Washington  15.06049987   24.984948         -9.92444824
## 55        West Virginia -30.87353849  -32.425424          1.55188574
## 56            Wisconsin   5.08410223   10.703718         -5.61961535
## 57              Wyoming -45.39784179  -34.142599        -11.25524293
```

- Note: percentages are calculated with Biden-Trump (+% means in favor of Clinton/Biden/-% means in favor of Trump)




```
## List of 1
##  $ legend.position: chr "right"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```



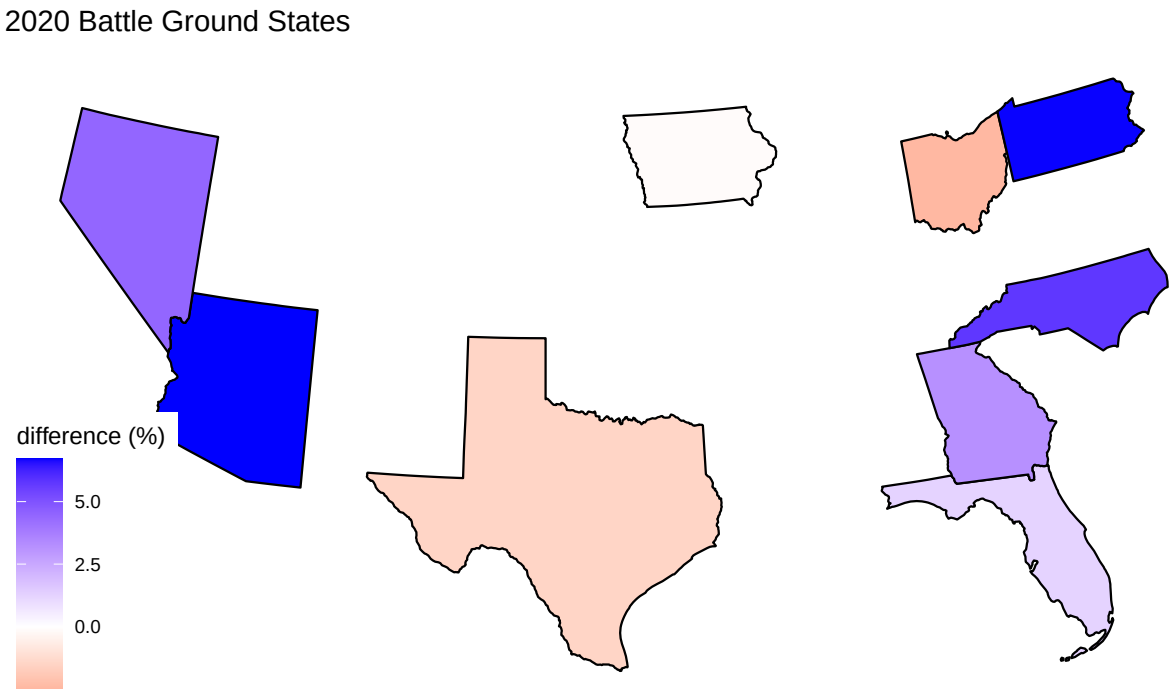

```
## List of 1
##  $ legend.position: chr "right"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

Comparing these two maps of the % diff based on polls from 2016 and 2020 we see that: * Texas went from a red to a white state (closer elections) * Nevada and Arizona shifted from white to more blue (favor blue more) * Colorado shifted even more blue * Wyoming went shifted closer to white from red * North Dakota, South Dakota, Nebraska, Kansas, and Oklahoma shifted closer to white from red * Virginia shifted more blue * New York, Maine, Massachusetts, New Jersey, Connecticut, Pennsylvania, shifted more blue

Red represents Republican (2016/2020-Trump) Blue represents Democrat (2016-Clinton / 2020-Biden) *White represents a "nuetral/battleground state" (close elections between two candidates)* Shifts to a certain color represents movement in votes towards a particular side according to polls

(b) Name 10 battleground states (states with closest percentage difference between two candidates) in 2020 based on the plots for (a). Explain your reasoning.

Battleground states are states in which the elections are close, or neither candidate is ahead of one another by a significant margin. Such states are represented by white states in the map and smallest |% diff| in the table. Ten battleground states in 2020 based on the map and table are **Texas, Florida, Iowa, Kansas, Ohio, Georgia, Nevada, North Carolina, Pennsylvania, Arizona**. They can be seen by the states below (stronger color shows which side the state is leaning to).




```
## List of 1
##  $ legend.position: chr "right"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

(c) Compare the difference of the polls in 2016 and in 2020 for states in US.

\begin{center} \textbf{2016-2020 State % Difference Table} \end{center}

```
##                   State  2016 % diff 2020 % diff 2016 to 2020 change
## 1               Alabama -26.03640239  -18.967223         -7.06917911
## 2                Alaska -11.66978248   -8.466445         -3.20333725
## 3               Arizona  -2.84276819    6.722131         -9.56489901
## 4              Arkansas -18.72792611  -18.692669         -0.03525737
## 5            California  25.09696425   27.453997         -2.35703266
## 6              Colorado   3.55450565   18.445174        -14.89066820
## 7           Connecticut  12.55773444   27.864740        -15.30700568
## 8              Delaware  14.00220485   28.107698        -14.10549316
## 9  District of Columbia  78.00263765   80.382676         -2.38003823
## 10              Florida  -0.85880347    1.261647         -2.12045014
## 11              Georgia  -6.23845603    3.224546         -9.46300180
## 12               Hawaii  21.29048167   31.985467        -10.69498543
## 13                Idaho -27.30037104  -19.275627         -8.02474443
## 14             Illinois  14.84783958   20.575001         -5.72716096
```

```
## 15              Indiana -12.79683401   -9.746841         -3.04999310
## 16                 Iowa  -5.25366811   -0.129796         -5.12387208
## 17               Kansas -12.70783255   -6.269227         -6.43860579
## 18             Kentucky -20.48601335  -16.442821         -4.04319213
## 19            Louisiana -19.46778640  -16.558465         -2.90932128
## 20                Maine   7.40846988   14.261571         -6.85310146
## 23             Maryland  29.19725687   36.241738         -7.04448076
## 24        Massachusetts  26.97748661   41.271098        -14.29361153
## 25             Michigan   3.31718461    8.304549         -4.98736475
## 26            Minnesota   7.37575565   12.464149         -5.08839307
## 27          Mississippi -16.25529595  -17.043113          0.78781693
## 28             Missouri -10.52605877   -7.564304         -2.96175485
## 29              Montana -18.35174544   -7.447978        -10.90376734
## 30              Nebraska -22.28743243   -7.633873        -14.65355899
## 34                Nevada  -0.68398929    4.457050         -5.14103959
## 35        New Hampshire   5.54344853   12.316213         -6.77276425
## 36            New Jersey  12.35516052   24.822972        -12.46781150
## 37            New Mexico   7.85285326    8.537469         -0.68461553
## 38              New York  20.31379551   32.175511        -11.86171582
## 39        North Carolina  -0.02762169    5.760768         -5.78838949
## 40          North Dakota -28.17286598  -19.737535         -8.43533106
## 41                  Ohio  -3.54393443   -2.464757         -1.07917712
## 42              Oklahoma -26.09803877  -19.385119         -6.71291945
## 43                Oregon  12.59674612   22.022765         -9.42601854
## 44          Pennsylvania   3.32186471    6.698433         -3.37656856
## 45          Rhode Island  11.16691429   35.097337        -23.93042257
## 46        South Carolina  -8.28035285   -6.863415         -1.41693764
## 47          South Dakota -22.07251049  -13.385534         -8.68697658
## 48             Tennessee -18.27335302  -13.282921         -4.99043186
## 49                 Texas -10.86512125   -1.484116         -9.38100491
## 51                  Utah -12.88484691   -9.366719         -3.51812811
## 52               Vermont  32.38876188   36.322657         -3.93389538
## 53              Virginia   7.04663931   14.049153         -7.00251377
## 54            Washington  15.06049987   24.984948         -9.92444824
## 55         West Virginia -30.87353849  -32.425424          1.55188574
## 56             Wisconsin   5.08410223   10.703718         -5.61961535
## 57               Wyoming -45.39784179  -34.142599        -11.25524293
```

The polls from 2016 to 2020 are different in their format. The 2020 polls are formatted in such a way that a single candidate is focused and the percent of people who voted for them and the sample size of the poll is provided. The 2016 polls gives you the count of total votes for a particular candidate from the poll directly. Additionally the 2020 polls provide more information like the grading of the poll, the candidate's party, and the methodology of the poll.

The difference between 2016 and 2020 polls in the data can be seen in the % change from 2016 to 2020 from the table. A negative change in % diff represents a shift towards the democratic side in terms of votes. Based off the chart, we see that 49/51 states showed more support for democrats in 2020 then in 2016. Analyzing the amount if red and blue states from the poll data from 2016 to 2020 we see that in 2016 there was 28 red states and 23 blue states, and in 2020 there were 23 red states and 28 blue states. The polls indicate a shifting of favor away from the republican party to the democratic party as a whole.

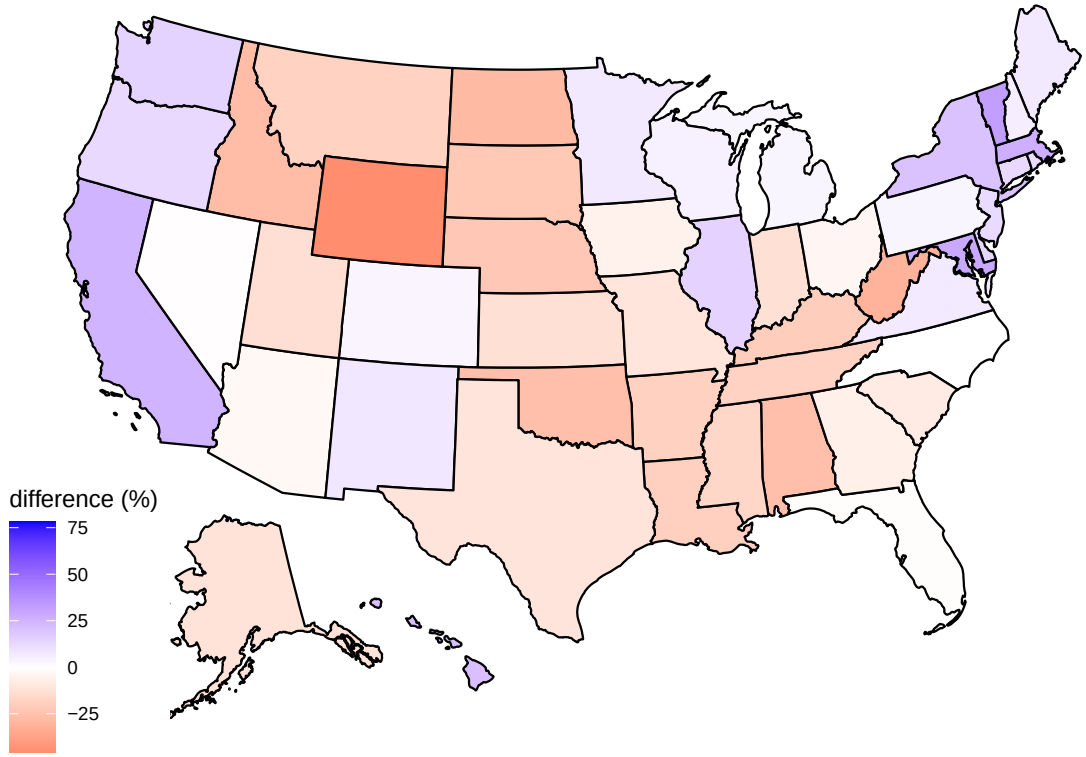
(d) Do polls underestimate the percentage of the real votes (in terms of percentage) received from one candidate in 2016? How about 2020? Discuss some reasons that may explain the bias in polls.

The polls appear to be misleading in 2016 in predicting that Hillary Clinton would win. To be fair, she did get a high percentage vote. Polls seem to predict that Biden would've won by a larger margin in 2020 then what really happened. These discrepancies between reality and the polls arise from the bad sampling of the population, unforeseen circumstances, and things that are generally out of the poll's controls. An important part of making predictions is making sure the sample which is being polled properly represents the population. In the practice of polling for elections it is difficult to properly guage what groups to send the polls to and to properly access/ensure that the polls themselves are done properly. Polling itself as a data gathering method is not very useful for elections because the public's opinion is very fickle and any of the candidates actions can have tremendous effects on their supporters.

Question 4: (20 points). Use data to explore states may change their electoral votes to another candidate from a different party and answer the following questions.

(a) Use figures or tables to compare the state level polls in 2016 and 2020.

```
par(mfrow = c(1,2))
plot_usmap(data = state.2016, values = "pdiff", color = "black") +
  scale_fill_gradient2(name = "difference (%)",   low= "red",
                       mid = "white",
                       high = "blue",
                       midpoint = 0)+
  labs(title = '2016 Election % Map')
```

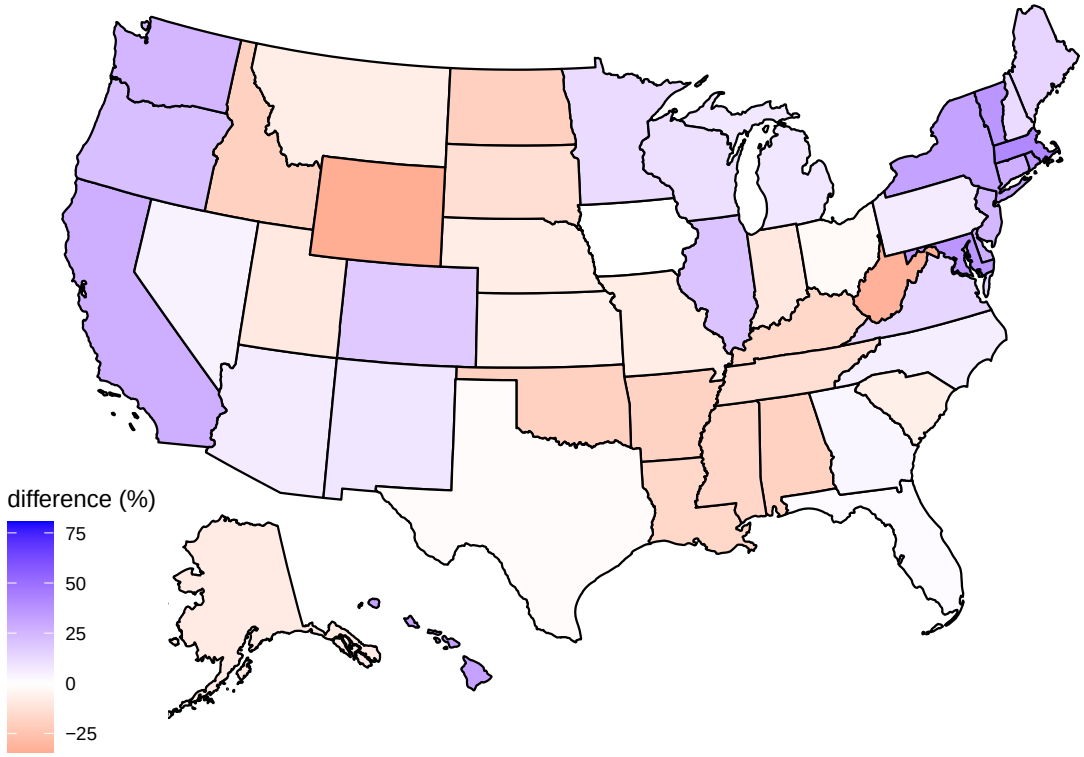



```
  theme(legend.position = "right")
```

```
## List of 1
##  $ legend.position: chr "right"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

```
plot_usmap(data = pstat.2020, values = "pdiff", color = "black") +
  scale_fill_gradient2(name = "difference (%)",   low= "red",
                       mid = "white",
                       high = "blue",
                       midpoint = 0)+
  labs(title = '2020 Election % Map')
```

2020 Election % Map




```
  theme(legend.position = "right")
```

```
## List of 1
##  $ legend.position: chr "right"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

```
pchart
```

```
##                   State  2016 % diff 2020 % diff 2016 to 2020 change
## 1               Alabama -26.03640239  -18.967223         -7.06917911
## 2                Alaska -11.66978248   -8.466445         -3.20333725
## 3               Arizona  -2.84276819    6.722131         -9.56489901
## 4              Arkansas -18.72792611  -18.692669         -0.03525737
## 5            California  25.09696425   27.453997         -2.35703266
## 6              Colorado   3.55450565   18.445174        -14.89066820
## 7           Connecticut  12.55773444   27.864740        -15.30700568
## 8              Delaware  14.00220485   28.107698        -14.10549316
## 9  District of Columbia  78.00263765   80.382676         -2.38003823
## 10              Florida  -0.85880347    1.261647         -2.12045014
## 11              Georgia  -6.23845603    3.224546         -9.46300180
## 12               Hawaii  21.29048167   31.985467        -10.69498543
## 13                Idaho -27.30037104  -19.275627         -8.02474443
## 14             Illinois  14.84783958   20.575001         -5.72716096
```

```
## 15              Indiana -12.79683401   -9.746841         -3.04999310
## 16                 Iowa  -5.25366811   -0.129796         -5.12387208
## 17               Kansas -12.70783255   -6.269227         -6.43860579
## 18              Kentucky -20.48601335  -16.442821         -4.04319213
## 19             Louisiana -19.46778640  -16.558465         -2.90932128
## 20                 Maine   7.40846988   14.261571         -6.85310146
## 23              Maryland  29.19725687   36.241738         -7.04448076
## 24         Massachusetts  26.97748661   41.271098        -14.29361153
## 25              Michigan   3.31718461    8.304549         -4.98736475
## 26             Minnesota   7.37575565   12.464149         -5.08839307
## 27           Mississippi -16.25529595  -17.043113          0.78781693
## 28              Missouri -10.52605877   -7.564304         -2.96175485
## 29               Montana -18.35174544   -7.447978        -10.90376734
## 30              Nebraska -22.28743243   -7.633873        -14.65355899
## 34                Nevada  -0.68398929    4.457050         -5.14103959
## 35         New Hampshire   5.54344853   12.316213         -6.77276425
## 36            New Jersey  12.35516052   24.822972        -12.46781150
## 37            New Mexico   7.85285326    8.537469         -0.68461553
## 38              New York  20.31379551   32.175511        -11.86171582
## 39        North Carolina  -0.02762169    5.760768         -5.78838949
## 40          North Dakota -28.17286598  -19.737535         -8.43533106
## 41                  Ohio  -3.54393443   -2.464757         -1.07917712
## 42              Oklahoma -26.09803877  -19.385119         -6.71291945
## 43                Oregon  12.59674612   22.022765         -9.42601854
## 44          Pennsylvania   3.32186471    6.698433         -3.37656856
## 45          Rhode Island  11.16691429   35.097337        -23.93042257
## 46        South Carolina  -8.28035285   -6.863415         -1.41693764
## 47          South Dakota -22.07251049  -13.385534         -8.68697658
## 48             Tennessee -18.27335302  -13.282921         -4.99043186
## 49                 Texas -10.86512125   -1.484116         -9.38100491
## 51                  Utah -12.88484691   -9.366719         -3.51812811
## 52               Vermont  32.38876188   36.322657         -3.93389538
## 53              Virginia   7.04663931   14.049153         -7.00251377
## 54            Washington  15.06049987   24.984948         -9.92444824
## 55         West Virginia -30.87353849  -32.425424          1.55188574
## 56             Wisconsin   5.08410223   10.703718         -5.61961535
## 57               Wyoming -45.39784179  -34.142599        -11.25524293
```

(b) Draw your conclusion and name 5 states that may change their electoral votes in 2020.

```
pchart[c(3,10,11,34,39),,]
```

```
##             State 2016 % diff 2020 % diff 2016 to 2020 change
## 3         Arizona -2.84276819    6.722131           -9.564899
## 10        Florida -0.85880347    1.261647           -2.120450
## 11        Georgia -6.23845603    3.224546           -9.463002
## 39 North Carolina -0.02762169    5.760768           -5.788389
## 44   Pennsylvania  3.32186471    6.698433           -3.376569
```

The `2016 to 2020 change` in the table above highlights the poll's prediction in which states would switch votes (party-wise) from 2016 to 2020. A negative value indicates that that state's favor has shifted for

Democrats, while a positive value represents a shift in favor of Republicans. This shift does not necessarily predict the state will vote diferently, instead we need to look at an actual sign change in the % diff (democrat-republican) from 2016 to 2020. States where the polls indicate a change in the sign of the % diff are Arizona (R->D), Florida (R->D), Georgia (R->D), Nevada (R->D), and North Carolina (R->D).

(c) Are these 5 states Arizona, Georgia, Michigan, Pennsylvania and Wisconsin (which elected another candidate from a different party)? If not, please give your reasons. If so, based on the polls, name one or two other states that may elect another candidate from a different party in 2020 as well but did not happen in reality. Explain the reason.

The states which the polls predicted would elect a different party's candidate from 2016 to 2020 are different from Arizona, Georgia, Michigan, and Pennsylvania, and Winconsin. The states the polls predicted would switch up are Arizona, Florida, Georgia, Nevada, and North Carolina. Some reasons for this is that these states could be swing states, states which elect a different party then predicted. Another potential reason would be that the sample which was polled poorly represents the population. This could possibly happen because the choosing of who to poll was biased (location) or the organization which conducted the polling was biased as well. The states elections are very hard to guage in practice, espcially because of the polarizing actions of the candidates in recent years which create a very volatile environment of predictions that could change on a whim.

Question 5: (20 points). Compare the polls in Florida and Iowa in 2016 and 2020.

(a) Are most of the polls in these two states accurate to predict the elected candidates? If not, please give some reasons.

```
## [1] "2016 Polls:"
```

```
## Florida: Correct Polls= 195 / Total Polls= 444 / Correct %= 43.9189189189189%
```

```
## Iowa: Correct Polls= 145 / Total Polls= 210 / Correct %= 69.0476190476191%
```

```
## [1] "2020 Polls:"
```

```
## Florida: Correct Polls= 23 / Total Polls- 158 / Correct %= 14.5569620253165%
```

```
## Iowa: Correct Polls= 33 / Total Polls- 56 / Correct %= 58.9285714285714%
```

Based on these results we see that most of the polls in Florida failed to predict Trump winning, and most of the polls in Iowa succeeded in predicting Trump's victory. Florida's polls may not be as accurate because Florida wasn't very consistent on a particular party. This indicates it is a battleground state which would worsen the chances that the polls properly represent the population as it requires more rigorous sampling to properly represent the population.

(b) For Iowa, is there a poll that approximately correct for the final outcome of the election in Iowa? What is the name of this poll? You may search the internet to know more some information about this pollster.

```
## The poll most approximately correct for the election in 2016 Iowa has poll-id = 48036,
## and is called SurveyMonkey
```

```
## The poll most approximately correct for the election in Iowa 2020 has poll-id = 437,
## and is called Selzer & Co.
```

The poll which is most correct for the final outcome of the election in Iowa would most match the % Trump won by: 2016-9.41% / 2020-8.2%. The pollers corresponding with these poll id's are `SurveyMonkey` for 2016 and `Selzer & Co.` for 2020. SurveyMonkey is an online poll creating company (Rated C by fivethirtyeight. Selzer & Co. is an A+ rated polling company (by fivethirtyeight) that focuses on political polls in Iowa.

(c) Name a few possible reasons that account for the bias in polls for these two states.

Florida and Iowa both have a history of voting for both democrats and republicans. This could account for the difficulty polls have in predicting the outcomes of the elections in these states. Additionally voting fraud and corrupt (biased) polling for politics can play a role int he bias of polls in these two states. Another reason for the bias is that geogrpahy of these two states feature countryside residents which may be difficult to access for polling.

(d) Discuss some possible ways to improve polls for political election.

A possible way to improve polling would be the utilization of the web to access pollers. This could be accompanied with verification to increase the sampling capacity of polls as well as the security of the information obtained. Another way to improve polls is to focus the questions to more stable conversations. For example focusing on specific topics that won't easily swayed by the candidate's actions on their campaign. A general approach to improving political polling is to ensure that the population is accurately represented. This would involve countermeasures to voting fraud, corruption, ability to garner responses, and a general understanding of sampling practices.