Data Analytics Projct

Alan Feria

2023-05-25

```
library(knitr)
hook_output = knit_hooks$get('output')
knit_hooks$set(output = function(x, options) {
  # this hook is used only when the linewidth option is not NULL
  if (!is.null(n <- options$linewidth)) {
    x = xfun::split_lines(x)
    # any lines wider than n should be wrapped
    if (any(nchar(x) > n)) x = strwrap(x, width = n)
    x = paste(x, collapse = '\n')
  }
  hook_output(x, options)
})
```

Please first review the demo code in polls election.zip. Please submit your report as a pdf or a word file. Please submit your R code in separate file(s). Please attach figures from R in your report to illustrate your answers. Please interpret your results.

Question 1

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.

Part A

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

Steps Taken:

1. Selecting respective state state
2. Isolating data from 08/01/16 - 11/02/2016
3. Sum of total votes for each candidate by state
4. Outputing results

Michigan

```
polls_data_2016$enddate=mdy(polls_data_2016$enddate)
```

```
start_date_2016='2016-08-01'
end_date_2016='2016-11-02'
```

```
index_selected_michigan_2016=which(polls_data_2016$state=="Michigan" & polls_data_2016$enddate >= start_
polls_mich_data_2016=polls_data_2016[index_selected_michigan_2016,]
```

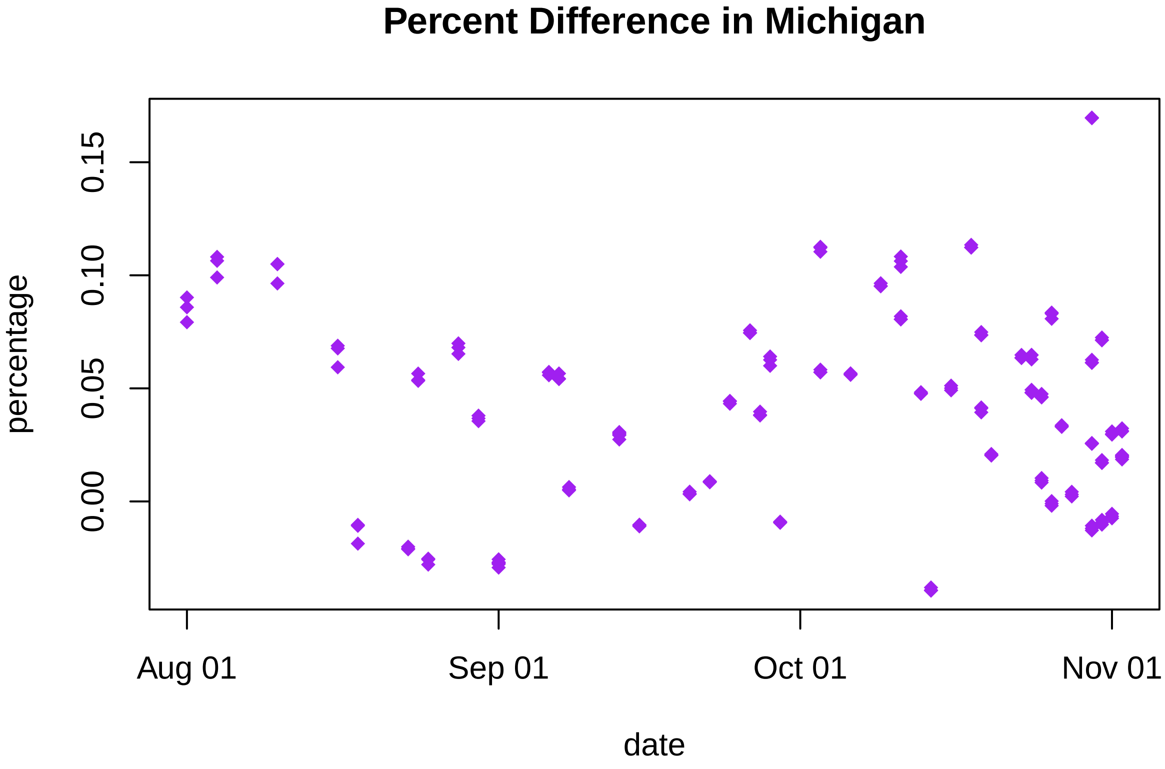
```
clinton_michigan_2016 = sum((polls_mich_data_2016$total.clinton))

trump_michigan_2016 = sum((polls_mich_data_2016$total.trump))

if (clinton_michigan_2016 > trump_michigan_2016){
  cat("Hillary Lead Michigan with ", clinton_michigan_2016, " votes.")
}else {
  cat("Trump Lead Michigan with ", trump_michigan_2016, " votes.")
}
```

Who is leading

```
## Hillary Lead Michigan with  68302.81  votes.
```

```
percentage_diff_michigan_2016=(
  polls_mich_data_2016$total.clinton-polls_mich_data_2016$total.trump)/
  (polls_mich_data_2016$total.clinton+polls_mich_data_2016$total.trump)
```

```
plot(polls_mich_data_2016$enddate,percentage_diff_michigan_2016,
     col='purple',pch=18,cex=1,type='p',xlab='date',
     ylab='percentage',main='Percent Difference in Michigan')
```



Difference

Georgia

```
index_selected_georgia_2016=which(polls_data_2016$state=="Georgia" & polls_data_2016$enddate >= start_da
polls_georg_data_2016=polls_data_2016[index_selected_georgia_2016,]
```

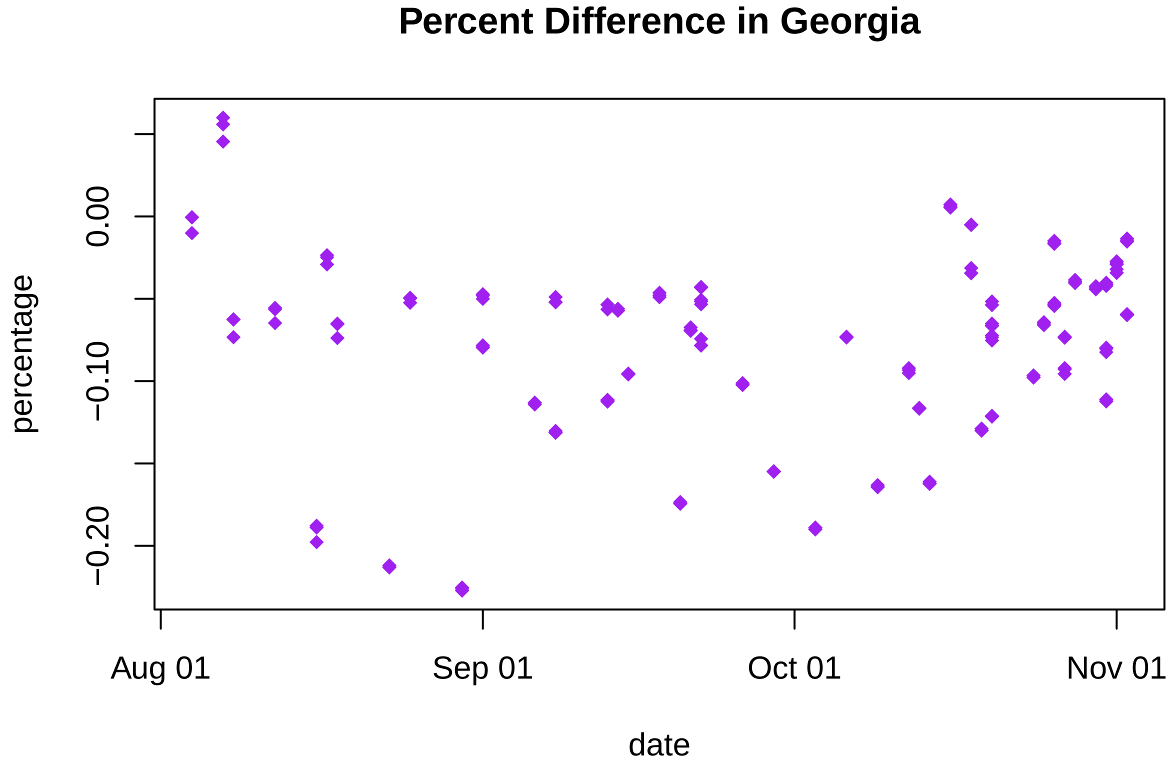
```r
clinton_georgia_2016 = sum((polls_georg_data_2016$total.clinton))

trump_georgia_2016 = sum((polls_georg_data_2016$total.trump))

if (clinton_georgia_2016 > trump_georgia_2016){
  cat("Hillary Lead Georgia with ", clinton_georgia_2016, " votes.")
}else {
  cat("Trump Lead Georgia with ", trump_georgia_2016, " votes.")
}
```

Who is leading

```
## Trump Lead Georgia with  71390.73  votes.
```

```r
percentage_diff_georgia_2016=(
  polls_georg_data_2016$total.clinton-polls_georg_data_2016$total.trump)/
  (polls_georg_data_2016$total.clinton+polls_georg_data_2016$total.trump)
```

```r
plot(polls_georg_data_2016$enddate,percentage_diff_georgia_2016,
     col='purple',pch=18,cex=1,type='p',xlab='date',
     ylab='percentage',main='Percent Difference in Georgia')
```



Difference

North Carolina

```r
index_selected_north_carolina_2016=which(polls_data_2016$state=="North Carolina" & polls_data_2016$endda
polls_north_car_data_2016=polls_data_2016[index_selected_north_carolina_2016,]
```

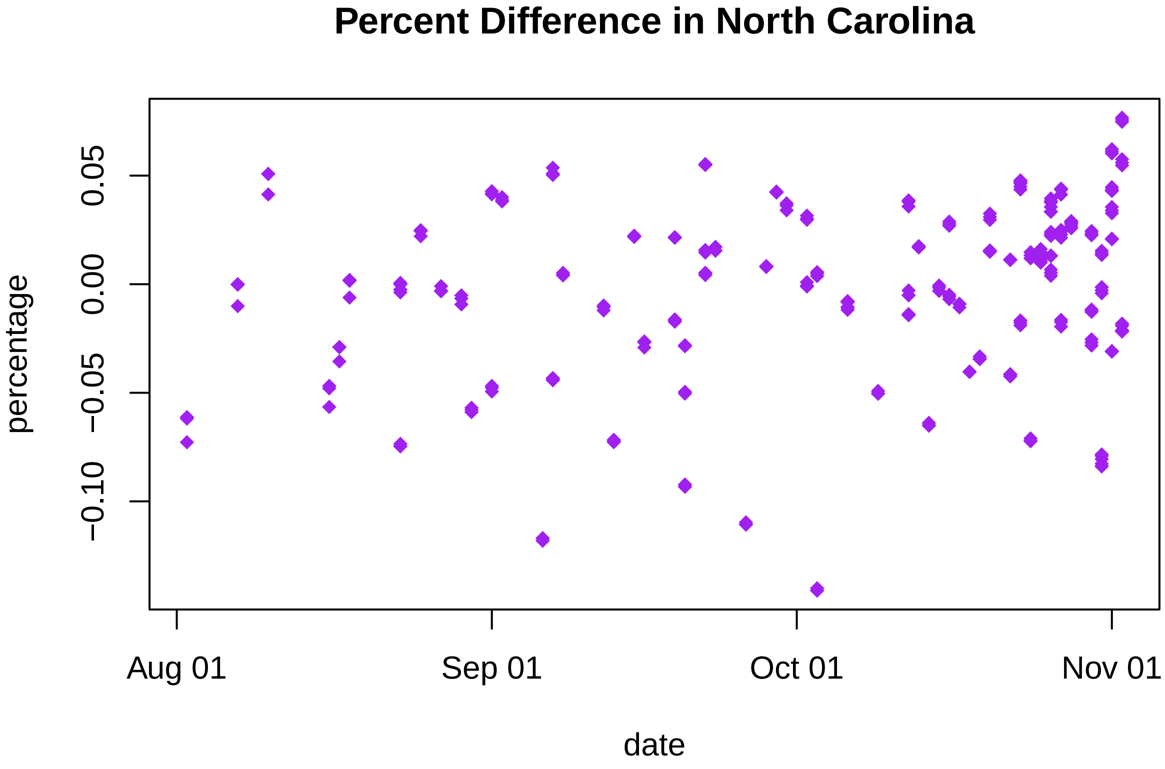
```
clinton_north_carolina_2016 = sum((polls_north_car_data_2016$total.clinton))

trump_north_carolina_2016 = sum((polls_north_car_data_2016$total.trump))

if (clinton_north_carolina_2016 > trump_north_carolina_2016){
  cat("Hillary Lead North Carolina with ", clinton_north_carolina_2016, " votes.")
}else {
  cat("Trump Lead North Carolina with ", trump_north_carolina_2016, " votes.")
}
```

Who is leading

```
## Trump Lead North Carolina with  104162.4  votes.
```

```
percentage_diff_north_carolina_2016=(
  polls_north_car_data_2016$total.clinton-polls_north_car_data_2016$total.trump)/
  (polls_north_car_data_2016$total.clinton+polls_north_car_data_2016$total.trump)
```

```
plot(polls_north_car_data_2016$enddate,percentage_diff_north_carolina_2016,
     col='purple',pch=18,cex=1,type='p',xlab='date',
     ylab='percentage',main='Percent Difference in North Carolina')
```



Difference

Part B

```
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

```
##
##  One Sample t-test
##
## data:  polls_mich_data_2016$total.clinton - polls_mich_data_2016$total.trump
## t = 10.94, df = 176, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  22.57701      Inf
## sample estimates:
## mean of x
##  26.59718
```

Since **Non-paired** T-test and **paired** T-test have the same output meaning that the paired data being analyzed in the paired t-test exhibit no correlation or dependency between the paired observations. In other words, the two groups being compared in the non-paired t-test can be considered as independent and unrelated.

The p-value is less than 2.2e-16, which is essentially 0. This indicates strong evidence against the null hypothesis.

Based on this output, we can conclude that there is a statistically significant difference between the mean values of total.clinton and total.trump in Michigan. The positive mean difference suggests that, on average, the total.clinton count is higher than the total.trump count. From this we can predict that Hillary won Michigan

Georgia Paired T-Test

```
##
##  Paired t-test
##
## data:  polls_georg_data_2016$total.clinton and polls_georg_data_2016$total.trump
## t = -19.242, df = 167, p-value = 1
## alternative hypothesis: true mean difference is greater than 0
## 95 percent confidence interval:
##  -55.92167       Inf
## sample estimates:
## mean difference
##       -51.49507
```

Since **Non-paired** T-test and **paired** T-test have the same output meaning that the paired data being analyzed in the paired t-test exhibit no correlation or dependency between the paired observations. In other words, the two groups being compared in the non-paired t-test can be considered as independent and unrelated.

The p-value is 1, which is greater than the typical significance level of 0.05. This suggests that there is not enough evidence to reject the null hypothesis.

Based on this output, we cannot confidently claim a significant difference between the mean values of total.clinton and total.trump in the given dataset for the specific subset related to Georgia. The results suggest that there is no strong evidence to support the claim that either candidate is favored based on the available data.

North Carolina Paired T-Test

```
##
##  One Sample t-test
```

```
##
## data:  polls_north_car_data_2016$total.clinton - polls_north_car_data_2016$total.trump
## t = -0.76542, df = 278, p-value = 0.7777
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  -4.904571       Inf
## sample estimates:
## mean of x
## -1.553973
```

Since **Non-paired** T-test and **paired** T-test have the same output meaning that the paired data being analyzed in the paired t-test exhibit no correlation or dependency between the paired observations. In other words, the two groups being compared in the non-paired t-test can be considered as independent and unrelated.

The p-value is 0.2216, which is greater than the typical significance level of 0.05. This suggests that there is not enough evidence to reject the null hypothesis.

Based on this output, we cannot confidently claim a significant difference between the mean values of total.clinton and total.trump in the given dataset for the specific subset related to North Carolina. The results suggest that there is no strong evidence to support the claim that either candidate is favored based on the available data.

Part C

```
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

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  polls_mich_data_2016$total.clinton and polls_mich_data_2016$total.trump
## W = 18112, p-value = 0.005515
## alternative hypothesis: true location shift is greater than 0
```

The p-value is 0.002999, which is less than the typical significance level of 0.05. This suggests that there is strong evidence to reject the null hypothesis.

Based on this output, we can conclude that there is a significant difference in the medians of total.clinton and total.trump in the given dataset for the specific subset related to Michigan. The results suggest that there is a strong evidence to support the claim that one candidate is favored over the other based on the available data.

Georgia Wilcoxon signed-rank test

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  polls_georg_data_2016$total.clinton and polls_georg_data_2016$total.trump
## W = 9869, p-value = 1
## alternative hypothesis: true location shift is greater than 0
```

The p-value is 1, which is greater than the typical significance level of 0.05. This suggests that there is not enough evidence to reject the null hypothesis.

Based on this output, we cannot confidently claim a significant difference in the medians of total.clinton and total.trump in the given dataset for the specific subset related to Georgia. The results suggest that there is no strong evidence to support the claim that either candidate is favored based on the available data.

North Carolina Wilcoxon signed-rank test

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  polls_north_car_data_2016$total.clinton and polls_north_car_data_2016$total.trump
## W = 37939, p-value = 0.697
## alternative hypothesis: true location shift is greater than 0
```

The p-value is 0.6381, which is greater than the typical significance level of 0.05. This suggests that there is not enough evidence to reject the null hypothesis.

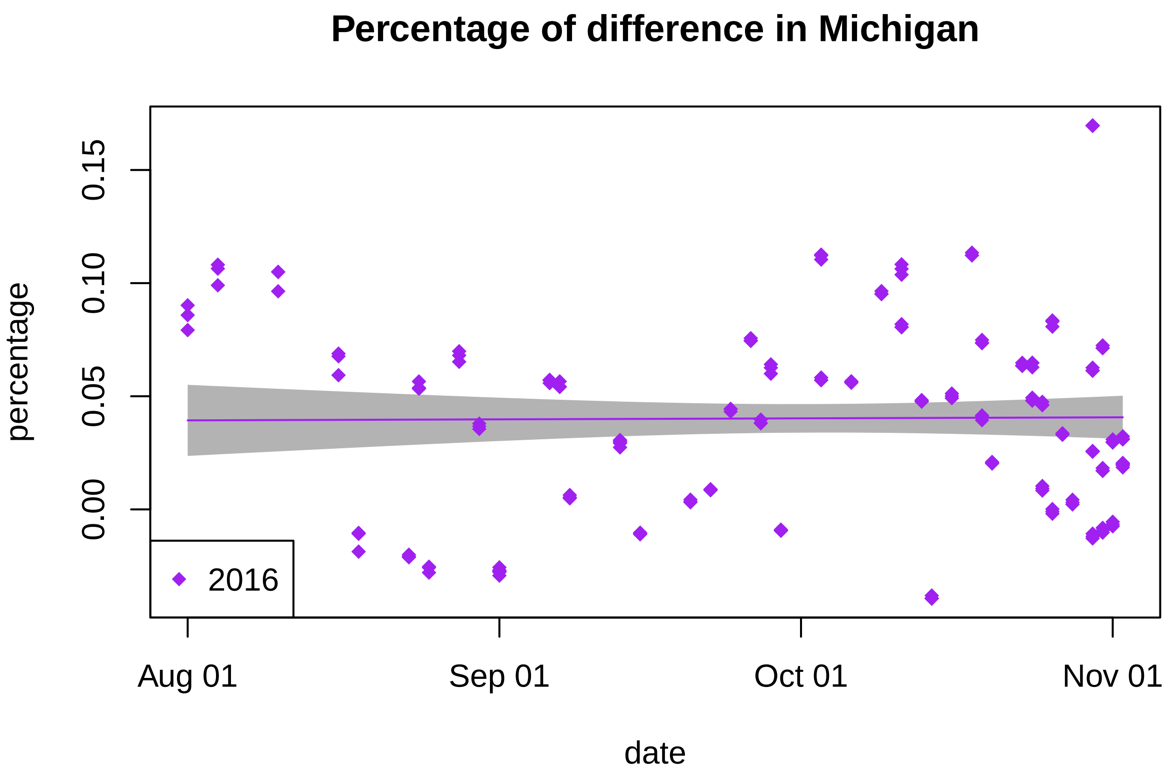
Based on this output, we cannot confidently claim a significant difference in the medians of total.clinton and total.trump in the given dataset for the specific subset related to North Carolina. The results suggest that there is no strong evidence to support the claim that either candidate is favored based on the available data.

Part D

```
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)?
```

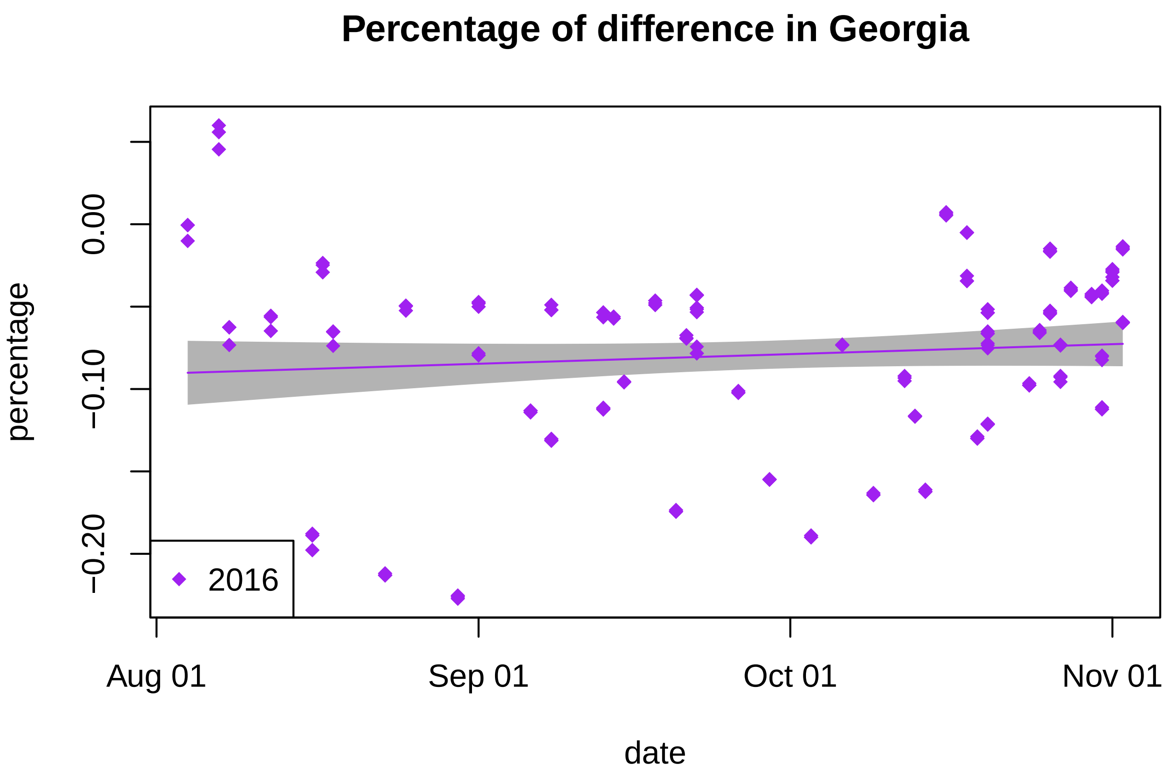
Michigan

```
plot(polls_mich_data_2016$enddate,percentage_diff_michigan_2016,
     col='purple',pch=18,cex=1,type='p',xlab='date',
     ylab='percentage',main='Percentage of difference in Michigan')
legend("bottomleft",col=c('purple'),pch=c(18,20),legend=c('2016'))
polygon(c(rev(polls_mich_data_2016$enddate), polls_mich_data_2016$enddate),
        c(rev(conf_interval_michigan_fitted_2016[,2]), conf_interval_michigan_fitted_2016[ ,3]), col =
lines(polls_mich_data_2016$enddate,lm_model_michigan_2016$fitted.values,
      col='purple',pch=20,type='l')

lines(polls_mich_data_2016$enddate,percentage_diff_michigan_2016,
     col='purple',pch=18,cex=1,type='p')
```



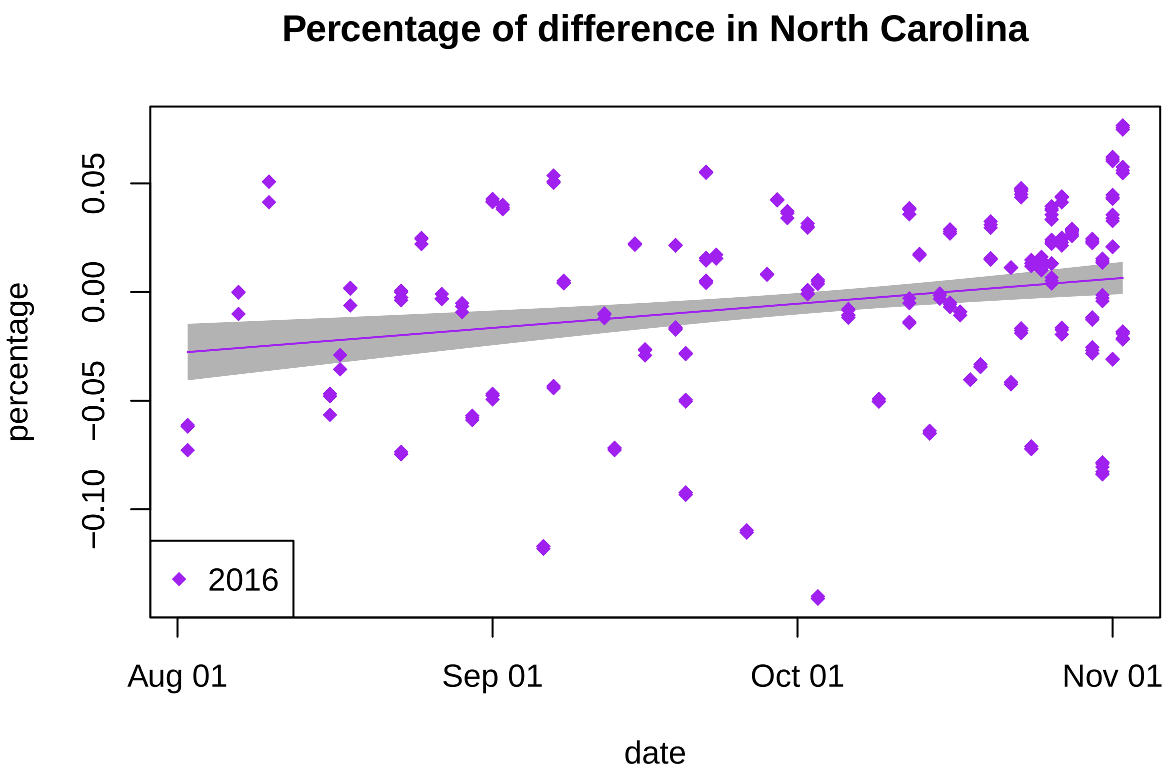
Georgia

```
plot(polls_georg_data_2016$enddate,percentage_diff_georgia_2016,
     col='purple',pch=18,cex=1,type='p',xlab='date',
     ylab='percentage',main='Percentage of difference in Georgia')
legend("bottomleft",col=c('purple'),pch=c(18,20),legend=c('2016'))
polygon(c(rev(polls_georg_data_2016$enddate), polls_georg_data_2016$enddate),
        c(rev(conf_interval_georgia_fitted_2016[,2]), conf_interval_georgia_fitted_2016[ ,3]), col = 'g
lines(polls_georg_data_2016$enddate,lm_model_georgia_2016$fitted.values,
      col='purple',pch=20,type='l')

lines(polls_georg_data_2016$enddate,percentage_diff_georgia_2016,
     col='purple',pch=18,cex=1,type='p')
```



North Carolina

```
plot(polls_north_car_data_2016$enddate,percentage_diff_north_carolina_2016,
     col='purple',pch=18,cex=1,type='p',xlab='date',
     ylab='percentage',main='Percentage of difference in North Carolina')
legend("bottomleft",col=c('purple'),pch=c(18,20),legend=c('2016'))
polygon(c(rev(polls_north_car_data_2016$enddate), polls_north_car_data_2016$enddate),
        c(rev(conf_interval_north_carolina_fitted_2016[,2]), conf_interval_north_carolina_fitted_2016[
lines(polls_north_car_data_2016$enddate,lm_model_north_carolina_2016$fitted.values,
      col='purple',pch=20,type='l')

lines(polls_north_car_data_2016$enddate,percentage_diff_north_carolina_2016,
     col='purple',pch=18,cex=1,type='p')
```



Based of the results from the linear model and confidence intervals, Michigan will have the closest running in terms of Trump winning or Hillary winning it.

Part E

```
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.)
```

```
min(percentage_diff_georgia_2016)*100
```

```
## [1] -22.7198
```

```
min(percentage_diff_michigan_2016)*100
```

```
## [1] -3.942341
```

```
min(percentage_diff_north_carolina_2016)*100
```

```
## [1] -14.1091
```

The election results showed that in North Carolina, Trump received 49.83% of the votes while Clinton received 46.17%. In Michigan, Trump received 47.25% while Clinton received 47.03%, and in Georgia, Trump received 50.38% while Clinton received 45.29%. The narrowest margin was observed in Michigan, with a difference of 3.94%. There are several potential explanations for this outcome.

Part F

```
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 polling results in Georgia and North Carolina were accurate, whereas Michigan's results were incorrect. It is possible that the discrepancy in Michigan can be attributed to the controversial nature of Trump's presidency. This particular election saw a surge in vocal individuals with strong opinions, who predominantly aligned with one political ideology. However, there were also a significant number of moderate voters who made their decisions late in the election cycle or altered their opinions. Additionally, polling lag could have played a role. The data collection process likely took some time, especially in Michigan, and as the election progressed, the gap between the candidates narrowed. Given a few more days of data, the polls might have indicated support for Trump.

Question 2

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.)

For the presidential poll in 2020, explore the poll in michigan, Georgia and North Carolina from August 1, 2020 to November 2 in 2020. Use the data to answer the following questions.

Part A

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

Steps Taken:

1. Selecting respective state state
2. Isolating data from 08/01/2020 - 11/02/2020
3. Sum of total votes for each candidate by state
4. Out puting results

```
date_2020= mdy(polls_data_2020$end_date)
date_2020_latest_day=date_2020[1]
index_selected=which(date_2020>='2020-08-01' & date_2020<='2020-11-01' )

polls_data_2020=polls_data_2020[index_selected,]

index_na=which(is.na(polls_data_2020$sample_size)==T)


polls_data_2020=polls_data_2020[which(polls_data_2020$answer=='Biden'|polls_data_2020$answer=='Trump'),]
# ##you may delete USC Dornsife/Los Angeles Times and Survey Monkey as their polls seem not disjoint
polls_data_2020=polls_data_2020[which(polls_data_2020$pollster_id!=1610&polls_data_2020$pollster_id!=119

##they do not match. Some poll has mistakes
length(which(polls_data_2020$answer=='Biden'))
```

```
## [1] 1743
```

```
length(which(polls_data_2020$answer=='Trump'))
```

```
## [1] 1739
```

```
##now let's delete those poll that only contains one candidate
polls_data_2020_question_id_num=unique(polls_data_2020$question_id)

for(i in 1:length(unique(polls_data_2020$question_id)) ){
  index_set=which(polls_data_2020$question_id==polls_data_2020_question_id_num[i])
  if(length(index_set)!=2){
    polls_data_2020=polls_data_2020[-index_set,]
  }
}

##now they match
length(which(polls_data_2020$answer=='Biden'))
```

```
## [1] 1739
```

```
length(which(polls_data_2020$answer=='Trump'))
```

```
## [1] 1739
```

```
###delete sample size NA
index_NA=which(is.na(polls_data_2020$sample_size)==T)
index_NA
```

```
## [1] 2583 2584
```

```
polls_data_2020=polls_data_2020[-index_NA,]

index_trump=which(polls_data_2020$answer=='Trump')
index_biden=which(polls_data_2020$answer=='Biden')

total_count_trump=sum(polls_data_2020$pct[index_trump]*polls_data_2020$sample_size[index_trump])
##total counts to biden
total_count_biden=sum(polls_data_2020$pct[index_biden]*polls_data_2020$sample_size[index_biden])

##the percentage may be more meaningful as some poll may be double counted
(total_count_biden-total_count_trump)/(total_count_biden+total_count_trump)
```

```
## [1] 0.07349577
```

```
##look at poll for Michigan
date_2020= mdy(polls_data_2020$end_date)
index_michigan_2020=which(polls_data_2020$state=="Michigan")
```

Michigan

```

##who is leading Michigan
index_biden_michigan_2020=which(polls_data_2020$answer=='Biden' & polls_data_2020$state=="Michigan")
index_trump_michigan_2020=which(polls_data_2020$answer=='Trump' & polls_data_2020$state=="Michigan")
```

```

counts_biden_michigan_2020=polls_data_2020$pct[index_biden_michigan_2020]*polls_data_2020$sample_size[i
counts_trump_michigan_2020=polls_data_2020$pct[index_trump_michigan_2020]*polls_data_2020$sample_size[i
```

```

biden_2020_michigan=sum(counts_biden_michigan_2020)
trump_2020_michigan=sum(counts_trump_michigan_2020)

biden_2020_michigan
```

How is leading

```
## [1] 4276738
```

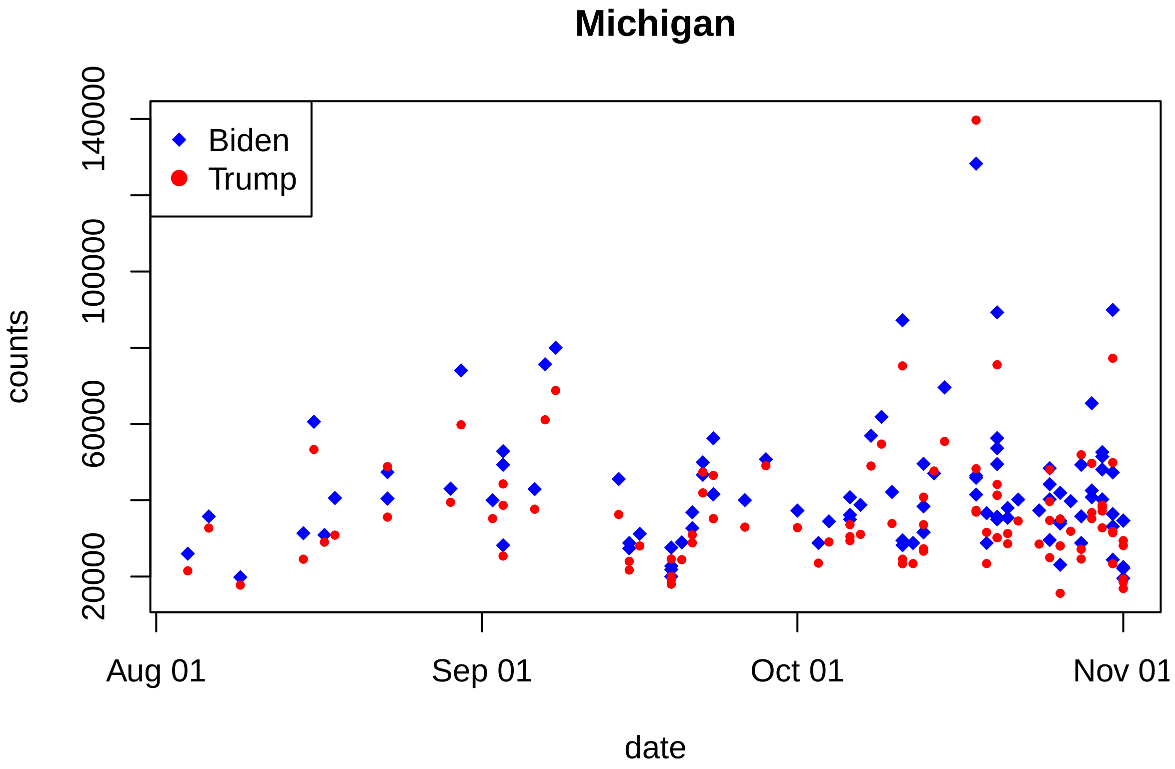
```
trump_2020_michigan
```

```
## [1] 3681786
```

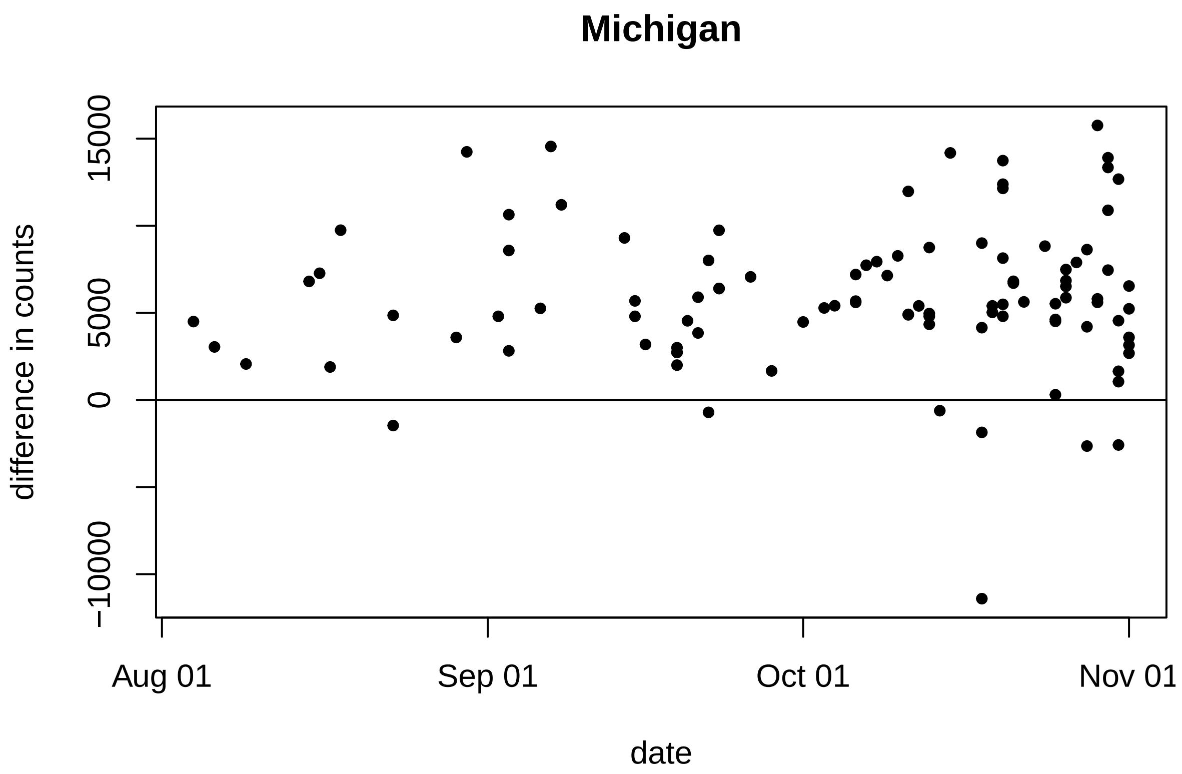
```
bidenperc <- (biden_2020_michigan-trump_2020_michigan)/(biden_2020_michigan+trump_2020_michigan)

trumpperc <- (trump_2020_michigan-biden_2020_michigan)/(biden_2020_michigan+trump_2020_michigan)

if (biden_2020_michigan> trump_2020_michigan){
  cat("Biden is leading the election in Michgan with a total of:",biden_2020_michigan,"votes or", bidenp
}else{
  cat("Trump is leading the election in Michgan with a total of:",trump_2020_michigan,"votes or", trumpp
}
```

```
## Biden is leading the election in Michgan with a total of: 4276738 votes or 7.475648 percent differen
```

```
#plot the poll by Biden and Trump over time in michigan
ylim_value=c(min(counts_biden_michigan_2020,counts_trump_michigan_2020),
             max(counts_biden_michigan_2020,counts_trump_michigan_2020))
plot(date_2020[index_biden_michigan_2020],counts_biden_michigan_2020,
     col='blue',pch=18,cex=1,type='p',xlab='date',ylab='counts',main='Michigan',ylim=ylim_value)
lines(date_2020[index_trump_michigan_2020],counts_trump_michigan_2020,col='red',pch=19,cex=.5,type='p')
legend("topleft",col=c('blue','red'),pch=c(18,19),legend=c('Biden','Trump'))
```



```
plot(date_2020[index_trump_michigan_2020],counts_biden_michigan_2020-counts_trump_michigan_2020,
     col='black',pch=20,type='p',xlab='date',ylab='difference in counts',main='Michigan')
abline(a=0,b=0)
```



Differnce

```
if (biden_2020_michigan> trump_2020_michigan){
  cat("Biden is leading the election in Michgan with a", bidenperc*100, "percent difference." )
}else{
  cat("Trump is leading the election in Michgan with", trumpperc*100, "percent difference.")
}
```

```
## Biden is leading the election in Michgan with a 7.475648 percent difference.
```

Georgia

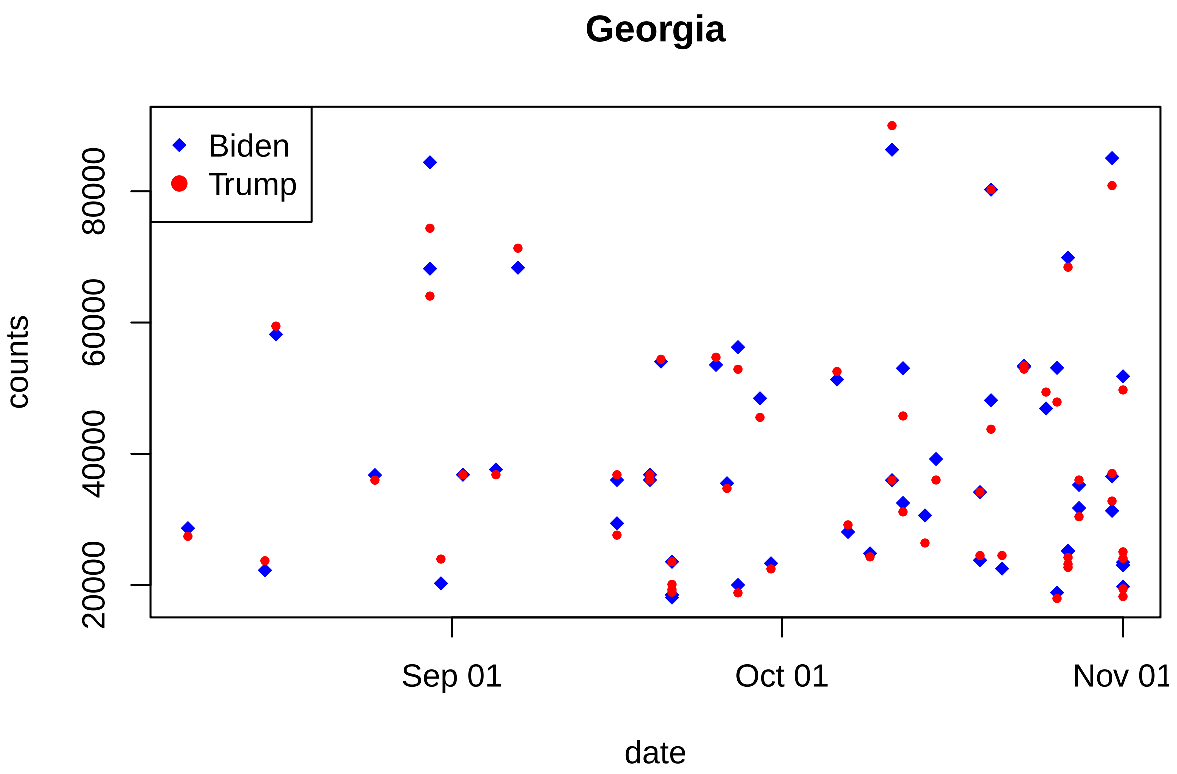
```

##who is leading Georgia
index_biden_georgia_2020=which(polls_data_2020$answer=='Biden' & polls_data_2020$state=="Georgia")
index_trump_georgia_2020=which(polls_data_2020$answer=='Trump' & polls_data_2020$state=="Georgia")
```

```

counts_biden_georgia_2020=polls_data_2020$pct[index_biden_georgia_2020]*polls_data_2020$sample_size[inde
counts_trump_georgia_2020=polls_data_2020$pct[index_trump_georgia_2020]*polls_data_2020$sample_size[inde
```

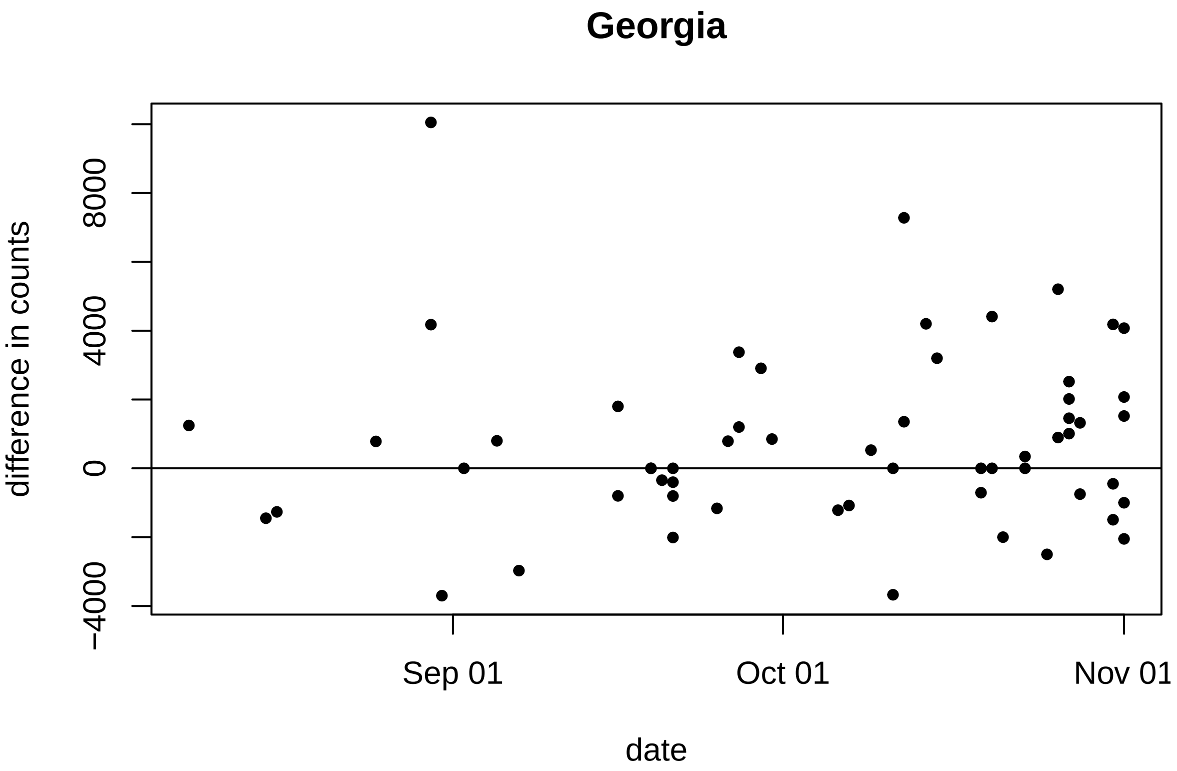
How is leading

```
## Biden is leading the election in Georgia with a total of: 2292043 votes or 0.9631098 percent differe
```

```
#plot the poll by Biden and Trump over time in georgia
ylim_value=c(min(counts_biden_georgia_2020,counts_trump_georgia_2020),
             max(counts_biden_georgia_2020,counts_trump_georgia_2020))
plot(date_2020[index_biden_georgia_2020],counts_biden_georgia_2020,
     col='blue',pch=18,cex=1,type='p',xlab='date',ylab='counts',main='Georgia',
     ylim=ylim_value)
lines(date_2020[index_trump_georgia_2020],counts_trump_georgia_2020,col='red',pch=19,cex=.5,type='p')
legend("topleft",col=c('blue','red'),pch=c(18,19),legend=c('Biden','Trump'))
```



```
plot(date_2020[index_trump_georgia_2020],
     counts_biden_georgia_2020-counts_trump_georgia_2020,
     col='black',pch=20,type='p',xlab='date',
     ylab='difference in counts',main='Georgia')
abline(a=0,b=0)
```



Differnce

```
if (biden_2020_georgia> trump_2020_georgia){
  cat("Biden is leading the election in Georgia with a",
      bidenperc*100, "percent difference." )
}else{
  cat("Trump is leading the election in Georgia with",
      trumpperc*100, "percent difference.")
}
```

```
## Biden is leading the election in Georgia with a 0.9631098 percent difference.
```

North Carolina

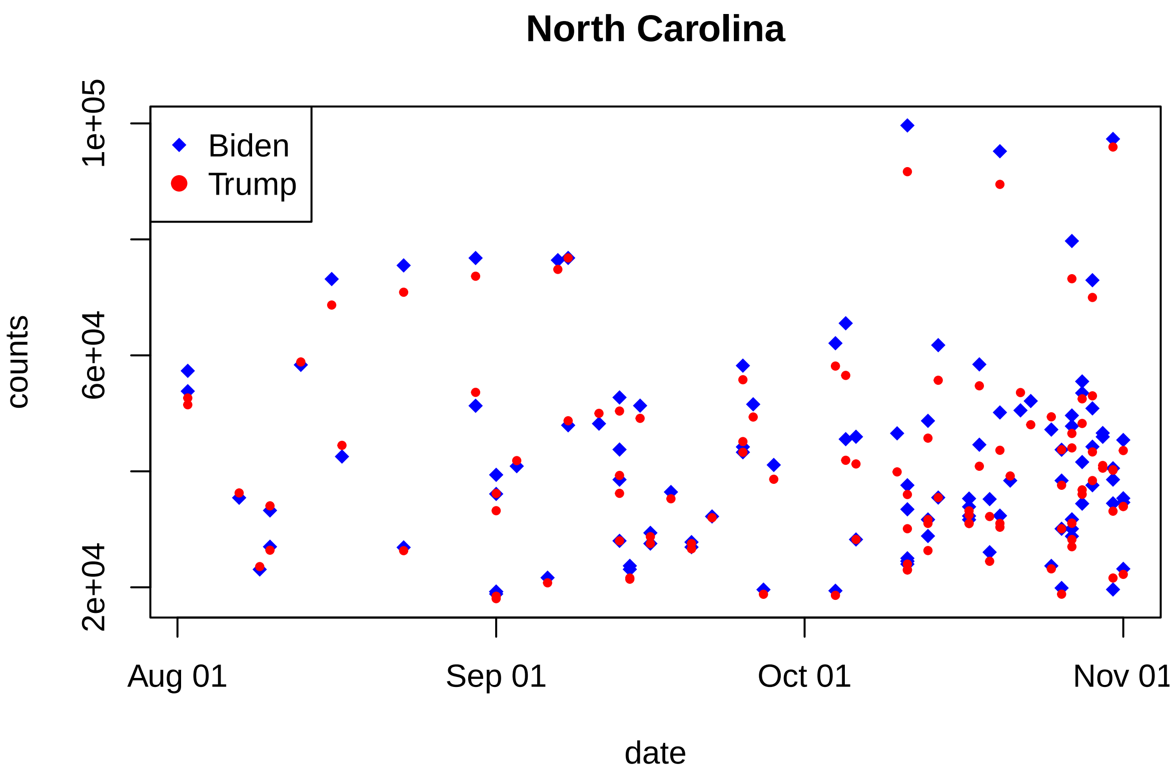
```
##who is leading North Carolina
index_biden_north_carolina_2020=which(polls_data_2020$answer=='Biden' & polls_data_2020$state=="North Ca
index_trump_north_carolina_2020=which(polls_data_2020$answer=='Trump' & polls_data_2020$state=="North Ca
```

```
counts_biden_north_carolina_2020=polls_data_2020$pct[index_biden_north_carolina_2020]*polls_data_2020$sa
counts_trump_north_carolina_2020=polls_data_2020$pct[index_trump_north_carolina_2020]*polls_data_2020$sa
```

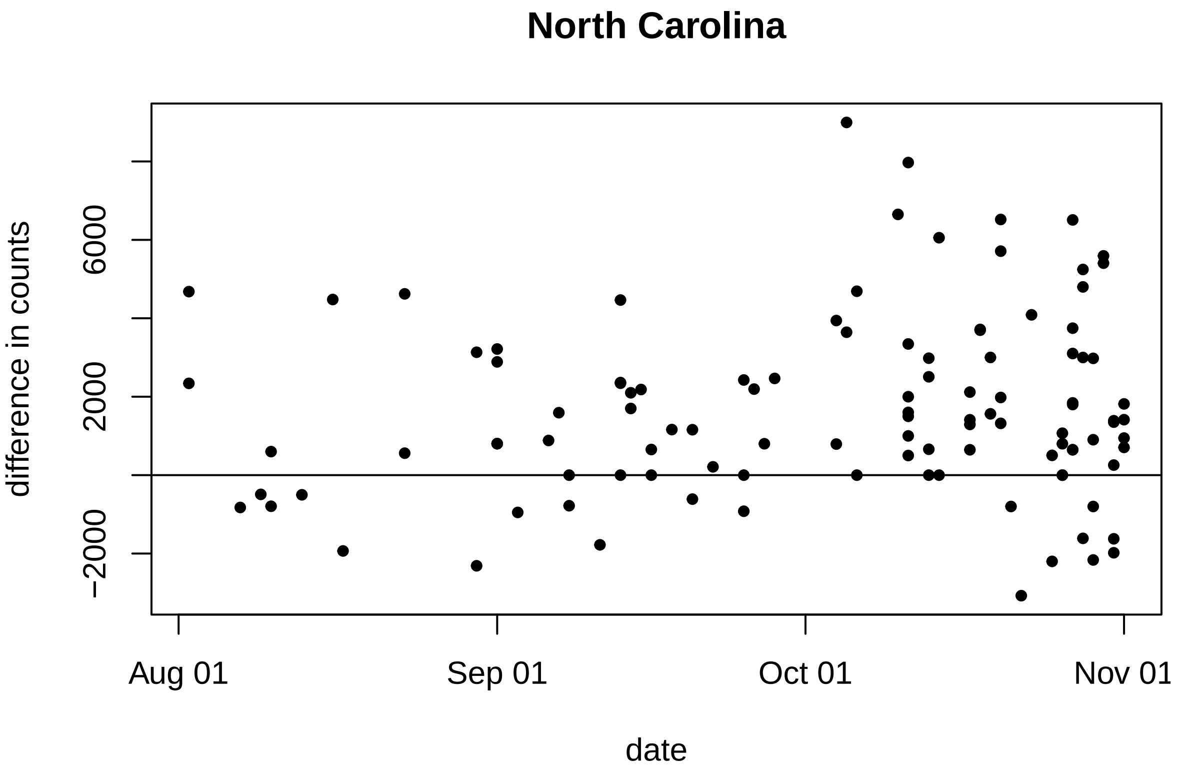
How is leading

```
## Biden is leading the election in North Carolina with a total of: 4602126 votes or 2.038051 percent di
```

```
#plot the poll by Biden and Trump over time in north_carolina
ylim_value=c(min(counts_biden_north_carolina_2020,counts_trump_north_carolina_2020),
             max(counts_biden_north_carolina_2020,counts_trump_north_carolina_2020))
plot(date_2020[index_biden_north_carolina_2020],counts_biden_north_carolina_2020,
     col='blue',pch=18,cex=1,type='p',xlab='date',ylab='counts',main='North Carolina',ylim=ylim_value)
lines(date_2020[index_trump_north_carolina_2020],counts_trump_north_carolina_2020,col='red',pch=19,cex=
legend("topleft",col=c('blue','red'),pch=c(18,19),legend=c('Biden','Trump'))
```



```
plot(date_2020[index_trump_north_carolina_2020],
     counts_biden_north_carolina_2020-counts_trump_north_carolina_2020,
     col='black',pch=20,type='p',xlab='date',
     ylab='difference in counts',main='North Carolina')
abline(a=0,b=0)
```



Differnce

```r
if (biden_2020_north_carolina> trump_2020_north_carolina){
  cat("Biden is leading the election in North Carolina with a", bidenperc*100, "percent difference diff
}else{
  cat("Trump is leading the election in North Carolina with", trumpperc*100, "percent difference differ
}
```

```
## Biden is leading the election in North Carolina with a 2.038051 percent difference difference.
```

Part B

```
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

```r
t.test(counts_biden_michigan_2020 -
         counts_trump_michigan_2020,
       alternative='greater')
```

```
##
##  One Sample t-test
##
## data:  counts_biden_michigan_2020 - counts_trump_michigan_2020
## t = 14.133, df = 100, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  5198.639      Inf
## sample estimates:
## mean of x
##  5890.607
```

```
t.test(counts_biden_michigan_2020,
         counts_trump_michigan_2020,paired=T,
       alternative='greater')
```

```
##
##  Paired t-test
##
## data:  counts_biden_michigan_2020 and counts_trump_michigan_2020
## t = 14.133, df = 100, p-value < 2.2e-16
## alternative hypothesis: true mean difference is greater than 0
## 95 percent confidence interval:
##  5198.639      Inf
## sample estimates:
## mean difference
##        5890.607
```

- The t-value of 14.133 indicates a significant difference between the means of the counts for Biden and Trump in Michigan.
- The extremely small p-value (< 2.2e-16) suggests strong evidence against the null hypothesis of no difference between the means.
- The alternative hypothesis states that the true difference in means is greater than 0, implying that Biden may have had a higher count than Trump in Michigan.
- The 95 percent confidence interval (5198.639 to Inf) indicates that we can be 95 percent confident that the true difference in means falls within this range.
- The sample estimate suggests that, on average, Biden had a higher count of votes than Trump in Michigan, with a mean difference of 5890.607.

Overall, these results suggest a significant and positive difference between the vote counts for Biden and Trump in Michigan, favoring Biden.

Since **Non-paired** T-test and **paired** T-test have the same output meaning that the paired data being analyzed in the paired t-test exhibit no correlation or dependency between the paired observations. In other words, the two groups being compared in the non-paired t-test can be considered as independent and unrelated.

Georgia Paired T-Test

```
t.test(counts_biden_georgia_2020 -
         counts_trump_georgia_2020,
       alternative='greater')
```

```
##
##  One Sample t-test
##
## data:  counts_biden_georgia_2020 - counts_trump_georgia_2020
## t = 2.257, df = 57, p-value = 0.01393
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  195.4117      Inf
## sample estimates:
## mean of x
##  753.9417
```

```
t.test(counts_biden_georgia_2020,
         counts_trump_georgia_2020,paired=T,
       alternative='greater')
```

```
##
##  Paired t-test
##
## data:  counts_biden_georgia_2020 and counts_trump_georgia_2020
## t = 2.257, df = 57, p-value = 0.01393
## alternative hypothesis: true mean difference is greater than 0
## 95 percent confidence interval:
##  195.4117      Inf
## sample estimates:
## mean difference
##        753.9417
```

- The t-value of 2.257 suggests a moderate difference between the means of the counts for Biden and Trump in Georgia.
- The p-value of 0.01393 indicates that there is statistical evidence to reject the null hypothesis of no difference between the means at a significance level of 0.05.
- The alternative hypothesis suggests that the true difference in means is greater than 0, implying that Biden may have had a higher count than Trump in Georgia.
- The 95 percent confidence interval (195.4117 to Inf) suggests that we can be 95 percent confident that the true difference in means falls within this range.
- The sample estimate suggests that, on average, Biden had a higher count of votes than Trump in Georgia, with a mean difference of 753.9417.

Overall, these results indicate a statistically significant difference between the vote counts for Biden and Trump in Georgia, favoring Biden.

Since **Non-paired** T-test and **paired** T-test have the same output meaning that the paired data being analyzed in the paired t-test exhibit no correlation or dependency between the paired observations. In other words, the two groups being compared in the non-paired t-test can be considered as independent and unrelated.

North Carolina Paired T-Test

```
t.test(counts_biden_north_carolina_2020 -
         counts_trump_north_carolina_2020,
       alternative='greater')
```

```
##
##  One Sample t-test
##
## data:  counts_biden_north_carolina_2020 - counts_trump_north_carolina_2020
## t = 7.597, df = 109, p-value = 5.58e-12
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  1306.321      Inf
## sample estimates:
## mean of x
##  1671.278
```

```r
t.test(counts_biden_north_carolina_2020,
         counts_trump_north_carolina_2020,paired=T,
       alternative='greater')
```

```
##
##  Paired t-test
##
## data:  counts_biden_north_carolina_2020 and counts_trump_north_carolina_2020
## t = 7.597, df = 109, p-value = 5.58e-12
## alternative hypothesis: true mean difference is greater than 0
## 95 percent confidence interval:
##  1306.321      Inf
## sample estimates:
## mean difference
##        1671.278
```

- The t-value of 7.7276 indicates a significant difference between the observed mean difference in vote counts for Biden and Trump in North Carolina and the hypothesized mean of 0.
- The p-value of 2.76e-12 is extremely small, providing strong evidence to reject the null hypothesis of no difference between the means.
- The alternative hypothesis suggests that the true mean of the difference in vote counts is greater than 0, indicating that Biden likely had a higher count than Trump in North Carolina.
- The 95 percent confidence interval (1329.498 to Inf) suggests that we can be 95 percent confident that the true mean difference falls within this range.
- The sample estimate indicates that the mean difference in vote counts between Biden and Trump in North Carolina is 1692.897

Overall, these results indicate a statistically significant difference between the vote counts for Biden and Trump in North Carolina, favoring Biden. The confidence interval also supports the notion that the true mean difference is likely positive.

Since **Non-paired** T-test and **paired** T-test have the same output meaning that the paired data being analyzed in the paired t-test exhibit no correlation or dependency between the paired observations. In other words, the two groups being compared in the non-paired t-test can be considered as independent and unrelated.

Part C

```
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

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  counts_biden_michigan_2020 and counts_trump_michigan_2020
## W = 6481, p-value = 0.0004468
## alternative hypothesis: true location shift is greater than 0
```

- The W value of 6481 indicates the sum of ranks assigned to the observations in the two groups.
- The p-value of 0.0004468 is below the conventional significance level of 0.05, providing strong evidence to reject the null hypothesis.

- The alternative hypothesis suggests that there is a true location shift, indicating that there is a difference in the distribution of vote counts between Biden and Trump in Michigan.
- The small p-value indicates that the difference in vote counts between the two candidates in Michigan is statistically significant.

Georgia Wilcoxon signed-rank test

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  counts_biden_georgia_2020 and counts_trump_georgia_2020
## W = 1711, p-value = 0.4375
## alternative hypothesis: true location shift is greater than 0
```

- The W value of 1711 indicates the sum of ranks assigned to the observations in the two groups.
- The p-value of 0.4375 is greater than the conventional significance level of 0.05, indicating that there is not enough evidence to reject the null hypothesis.
- The alternative hypothesis suggests that there is a true location shift, meaning a difference in the distribution of vote counts between Biden and Trump in Georgia.
- The relatively large p-value suggests that the difference in vote counts between the two candidates in Georgia is not statistically significant.

Overall, these results indicate that there is insufficient evidence to conclude a significant difference in vote counts between Biden and Trump in Georgia based on the Wilcoxon rank sum test. The test does not provide strong evidence of a location shift favoring one candidate over the other in Georgia.

North Carolina Wilcoxon signed-rank test

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  counts_biden_north_carolina_2020 and counts_trump_north_carolina_2020
## W = 6419, p-value = 0.2175
## alternative hypothesis: true location shift is greater than 0
```

- The W value of 6547 indicates the sum of ranks assigned to the observations in the two groups.
- The p-value of 0.2099 is greater than the conventional significance level of 0.05, indicating that there is not enough evidence to reject the null hypothesis.
- The alternative hypothesis suggests that there is a true location shift, meaning a difference in the distribution of vote counts between Biden and Trump in North Carolina.
- The relatively large p-value suggests that the difference in vote counts between the two candidates in North Carolina is not statistically significant.

Overall, these results indicate that there is insufficient evidence to conclude a significant difference in vote counts between Biden and Trump in North Carolina based on the Wilcoxon rank sum test. The test does not provide strong evidence of a location shift favoring one candidate over the other in North Carolina.

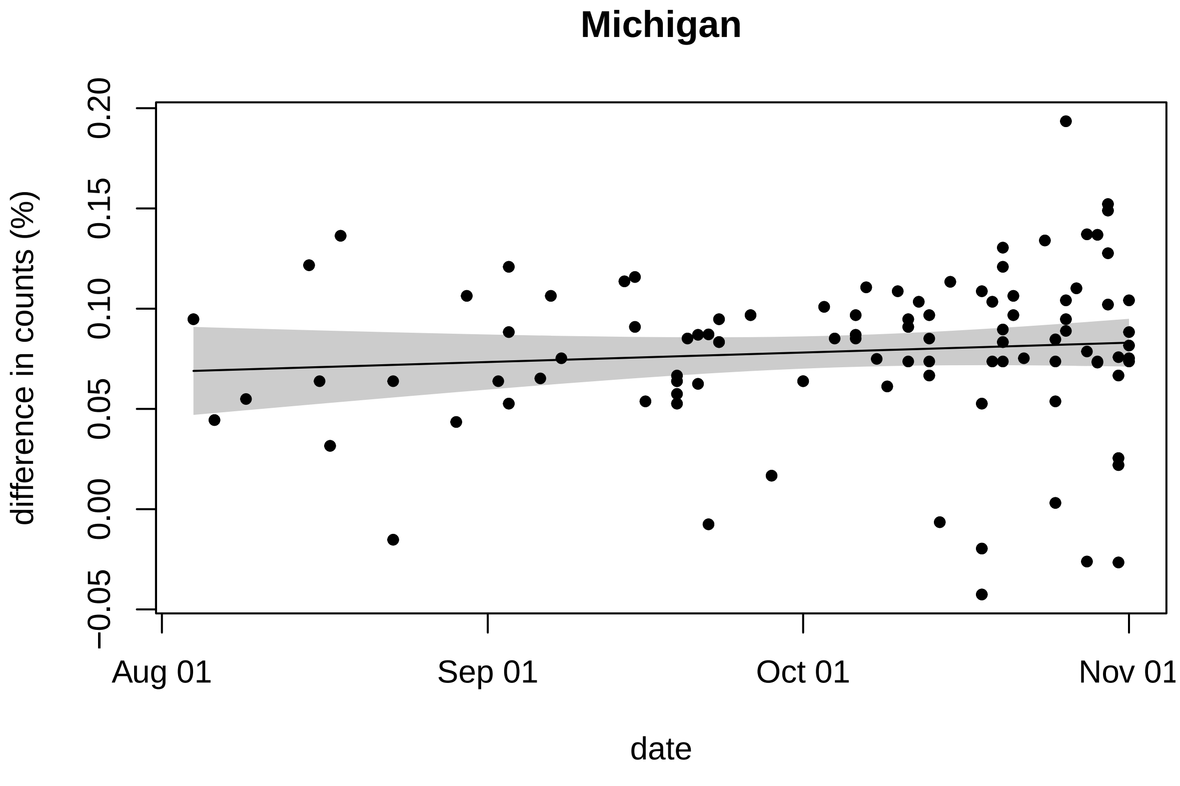
Part D

```
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)?
```

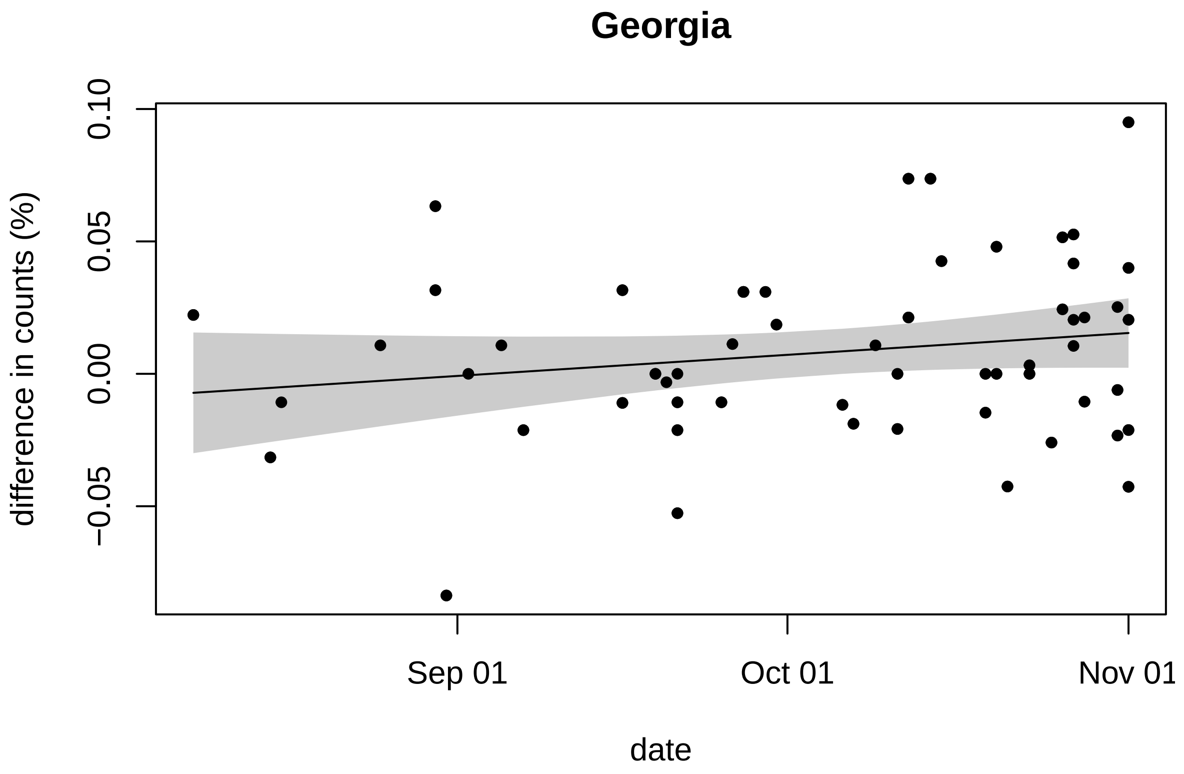
Michigan

```
plot(counts_michigan_for_lm_2020$data_date,
     counts_michigan_for_lm_2020$percentage_diff,
     col='black',pch=20,type='p',xlab='date',
     ylab='difference in counts (%)',main='Michigan')
polygon(c(rev(counts_michigan_for_lm_2020$data_date),
          counts_michigan_for_lm_2020$data_date),
        c(rev(conf_interval_michigan_fitted_2020[,2]),
          conf_interval_michigan_fitted_2020[ ,3]), col = 'grey80', border = NA)
lines(counts_michigan_for_lm_2020$data_date,
      lm_model_michigan_2020$fitted.values,
      col='black',pch=20,type='l',xlab='date',
      ylab='difference in counts (%)',main='Michigan')
lines(counts_michigan_for_lm_2020$data_date,
      counts_michigan_for_lm_2020$percentage_diff,
     col='black',pch=20,type='p',xlab='date',
     ylab='difference in counts (%)',main='Michigan')
```



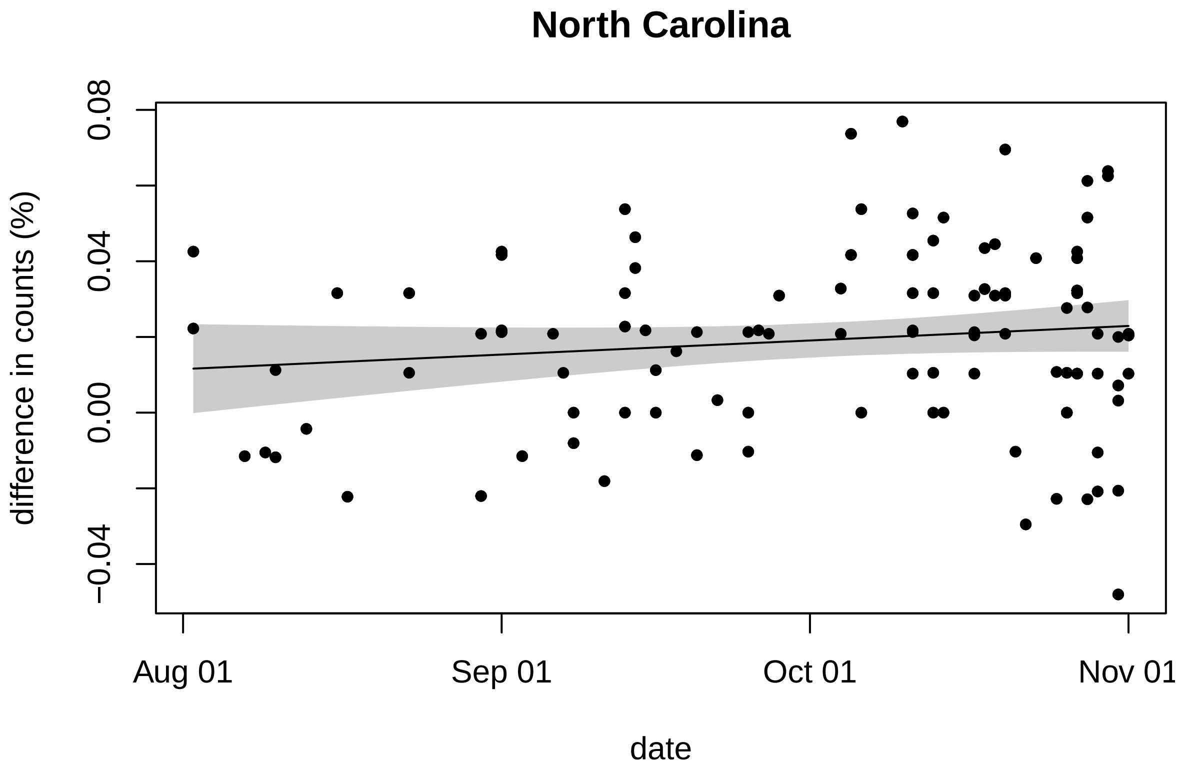
Georgia

```
plot(counts_georgia_for_lm_2020$data_date,
     counts_georgia_for_lm_2020$percentage_diff,
     col='black',pch=20,type='p',xlab='date',
     ylab='difference in counts (%)',main='Georgia')
polygon(c(rev(counts_georgia_for_lm_2020$data_date),
          counts_georgia_for_lm_2020$data_date),
        c(rev(conf_interval_georgia_fitted_2020[,2]),
          conf_interval_georgia_fitted_2020[ ,3]),
        col = 'grey80', border = NA)
lines(counts_georgia_for_lm_2020$data_date,
      lm_model_georgia_2020$fitted.values,
      col='black',pch=20,type='l',xlab='date',
      ylab='difference in counts (%)',main='Georgia')
lines(counts_georgia_for_lm_2020$data_date,
      counts_georgia_for_lm_2020$percentage_diff,
     col='black',pch=20,type='p',xlab='date',
     ylab='difference in counts (%)',main='Georgia')
```



North Carolina

```
plot(counts_north_carolina_for_lm_2020$data_date,
     counts_north_carolina_for_lm_2020$percentage_diff,
     col='black',pch=20,type='p',xlab='date',
     ylab='difference in counts (%)',main='North Carolina')
polygon(c(rev(counts_north_carolina_for_lm_2020$data_date),
          counts_north_carolina_for_lm_2020$data_date),
        c(rev(conf_interval_north_carolina_fitted_2020[,2]),
          conf_interval_north_carolina_fitted_2020[ ,3]),
        col = 'grey80',
        border = NA)
lines(counts_north_carolina_for_lm_2020$data_date,
      lm_model_north_carolina_2020$fitted.values,
      col='black',pch=20,type='l',xlab='date',
      ylab='difference in counts (%)',main='North Carolina')
lines(counts_north_carolina_for_lm_2020$data_date,
     counts_north_carolina_for_lm_2020$percentage_diff,
    col='black',pch=20,type='p',
    xlab='date',
    ylab='difference in counts (%)',main='North Carolina')
```



Part E

```
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.)
```

```
min(counts_michigan_for_lm_2020$percentage_diff)*100
```

```
## [1] -4.255319
```

```
min(counts_georgia_for_lm_2020$percentage_diff)*100
```

```
## [1] -8.371041
```

```
min(counts_north_carolina_for_lm_2020$percentage_diff)*100
```

```
## [1] -4.803493
```

The election results showed that in North Carolina, Trump received 50.1% of the votes while Joe received 48.7%. In Michigan, Trump received 47.8% while Joe received 50.6%, and in Georgia, Trump received 49.3% while Joe received 49.5%. The narrowest margin was observed in Michigan, with a difference of 4.26%. There are several potential explanations for this outcome.

Part F

```
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 poll results in Michigan and Georgia accurately predicted the election outcome, but they overestimated the margin between the candidates. This discrepancy could be attributed to a potential bias in the polling process, which might have favored gathering data from suburban areas more heavily.

Question 3

3. (20 points). 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.

```
polls_data_2016=read.csv(paste0("~/ucsb_pstat120c-s23/president_general_polls_sorted_end_date_2016.csv"

library(usmap)
library(ggplot2)
##compare average from Sep 1 to Nov 01
start_date_2016='2016-08-01'
end_date_2016='2016-11-02'

polls_data_2016_enddate=mdy(polls_data_2016$enddate)
polls_data_2016_after_sep=polls_data_2016[which(polls_data_2016_enddate>=start_date_2016&polls_data_201

poll_state_sum_clinton_2016=aggregate(polls_data_2016_after_sep$total.clinton, by=list(State=polls_data_
poll_state_sum_trump_2016=aggregate(polls_data_2016_after_sep$total.trump, by=list(State=polls_data_201

poll_state_diff_percentage=poll_state_sum_clinton_2016
poll_state_diff_percentage[,2]=(poll_state_sum_clinton_2016[,2]-poll_state_sum_trump_2016[,2])/(poll_st
delete_index=which((poll_state_diff_percentage[,1])=='U.S.')
poll_state_diff_percentage=poll_state_diff_percentage[-delete_index,]
poll_state_diff_percentage[,1]
```

```
##  [1] "Alabama"              "Alaska"               "Arizona"
##  [4] "Arkansas"             "California"           "Colorado"
##  [7] "Connecticut"          "Delaware"             "District of Columbia"
## [10] "Florida"              "Georgia"              "Hawaii"
## [13] "Idaho"                "Illinois"             "Indiana"
## [16] "Iowa"                 "Kansas"               "Kentucky"
## [19] "Louisiana"            "Maine"                "Maine CD-1"
## [22] "Maine CD-2"           "Maryland"             "Massachusetts"
## [25] "Michigan"             "Minnesota"            "Mississippi"
## [28] "Missouri"             "Montana"              "Nebraska"
## [31] "Nebraska CD-1"        "Nebraska CD-2"        "Nebraska CD-3"
## [34] "Nevada"               "New Hampshire"        "New Jersey"
## [37] "New Mexico"           "New York"             "North Carolina"
## [40] "North Dakota"         "Ohio"                 "Oklahoma"
## [43] "Oregon"               "Pennsylvania"         "Rhode Island"
## [46] "South Carolina"       "South Dakota"         "Tennessee"
## [49] "Texas"                "Utah"                 "Vermont"
## [52] "Virginia"             "Washington"           "West Virginia"
## [55] "Wisconsin"            "Wyoming"
```

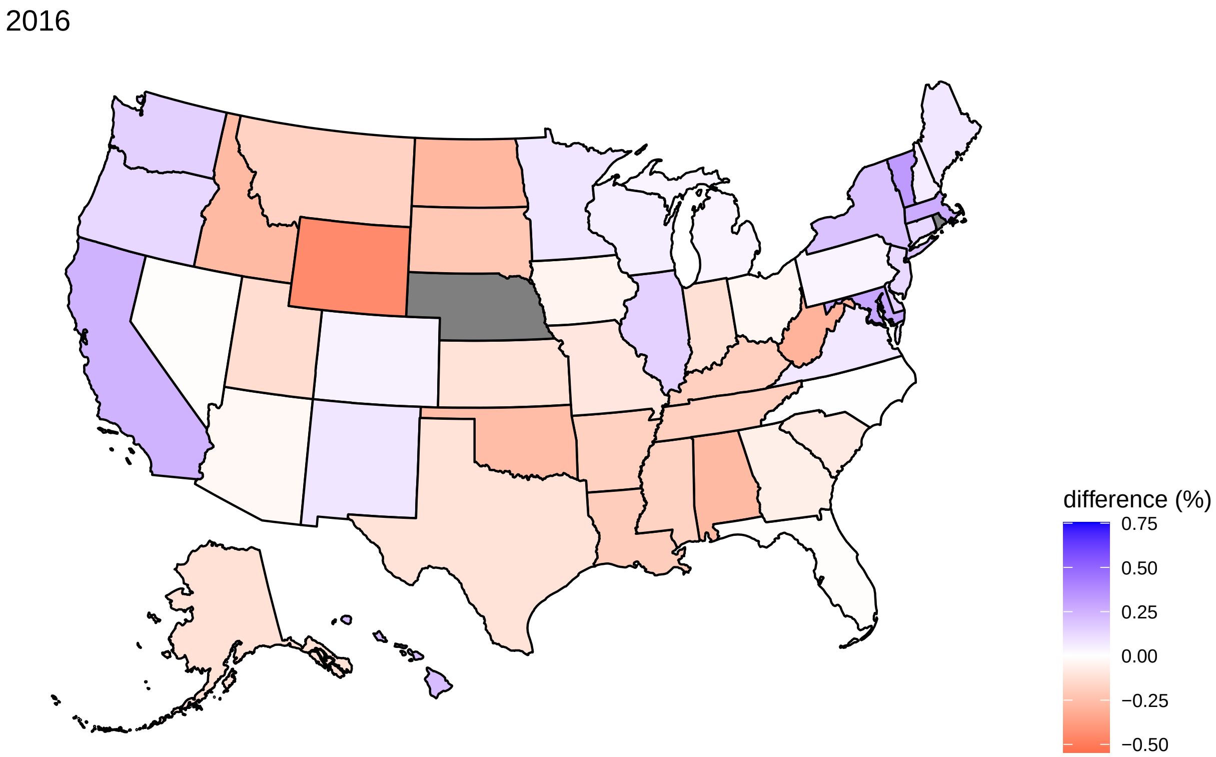
```
state_poll_2016 <- data.frame(
  state =poll_state_diff_percentage[,1],
  diff_percentage=poll_state_diff_percentage[,2]
)

##let's look at 2020
index_selected=which(date_2020>='2020-08-01' & date_2020<='2020-11-02')
polls_data_2020_after_sep=polls_data_2020[index_selected,]  ###only work on the poll after Aug 1


polls_data_2020_after_sep=polls_data_2020_after_sep[which(polls_data_2020$answer=='Biden'|polls_data_20

index_biden_2020=which(polls_data_2020_after_sep$answer=='Biden')
index_trump_2020=which(polls_data_2020_after_sep$answer=='Trump' )


counts_biden_2020=polls_data_2020$pct[index_biden_2020]*polls_data_2020$sample_size[index_biden_2020]
counts_trump_2020=polls_data_2020$pct[index_trump_2020]*polls_data_2020$sample_size[index_trump_2020]

##add two column
polls_data_2020$total.biden=rep(0,dim(polls_data_2020)[1])
polls_data_2020$total.trump=rep(0,dim(polls_data_2020)[1])

polls_data_2020$total.biden[index_biden_2020]=counts_biden_2020
polls_data_2020$total.trump[index_trump_2020]=counts_trump_2020

poll_state_sum_biden_2020=aggregate(polls_data_2020$total.biden, by=list(State=polls_data_2020$state),F
poll_state_sum_trump_2020=aggregate(polls_data_2020$total.trump, by=list(State=polls_data_2020$state),F
#delete the one with NA
poll_state_sum_biden_2020=poll_state_sum_biden_2020[-1,]
poll_state_sum_trump_2020=poll_state_sum_trump_2020[-1,]

##create a data frame
state_poll_2020 <- data.frame(
  state =poll_state_sum_biden_2020[,1],
  diff_percentage=(
    poll_state_sum_biden_2020[,2]-
      poll_state_sum_trump_2020[,2])/
    (poll_state_sum_biden_2020[,2]+
       poll_state_sum_trump_2020[,2])
)



limit_val=c(min(state_poll_2016$diff_percentage,state_poll_2020$diff_percentage),
            max(state_poll_2016$diff_percentage,state_poll_2020$diff_percentage))
##it does not have poll from Nebraska. It does not plot the poll from congressional district

##difference between 2020 and 2016
##delete nebrask CD-1 and CD-3, as 2020 does not have it
state_poll_2016$state
```

```
##  [1] "Alabama"              "Alaska"               "Arizona"
##  [4] "Arkansas"             "California"           "Colorado"
##  [7] "Connecticut"          "Delaware"             "District of Columbia"
## [10] "Florida"              "Georgia"              "Hawaii"
## [13] "Idaho"                "Illinois"             "Indiana"
## [16] "Iowa"                 "Kansas"               "Kentucky"
## [19] "Louisiana"            "Maine"                "Maine CD-1"
## [22] "Maine CD-2"           "Maryland"             "Massachusetts"
## [25] "Michigan"             "Minnesota"            "Mississippi"
## [28] "Missouri"             "Montana"              "Nebraska"
## [31] "Nebraska CD-1"        "Nebraska CD-2"        "Nebraska CD-3"
## [34] "Nevada"               "New Hampshire"        "New Jersey"
## [37] "New Mexico"           "New York"             "North Carolina"
## [40] "North Dakota"         "Ohio"                 "Oklahoma"
## [43] "Oregon"               "Pennsylvania"         "Rhode Island"
## [46] "South Carolina"       "South Dakota"         "Tennessee"
## [49] "Texas"                "Utah"                 "Vermont"
## [52] "Virginia"             "Washington"           "West Virginia"
## [55] "Wisconsin"            "Wyoming"
```

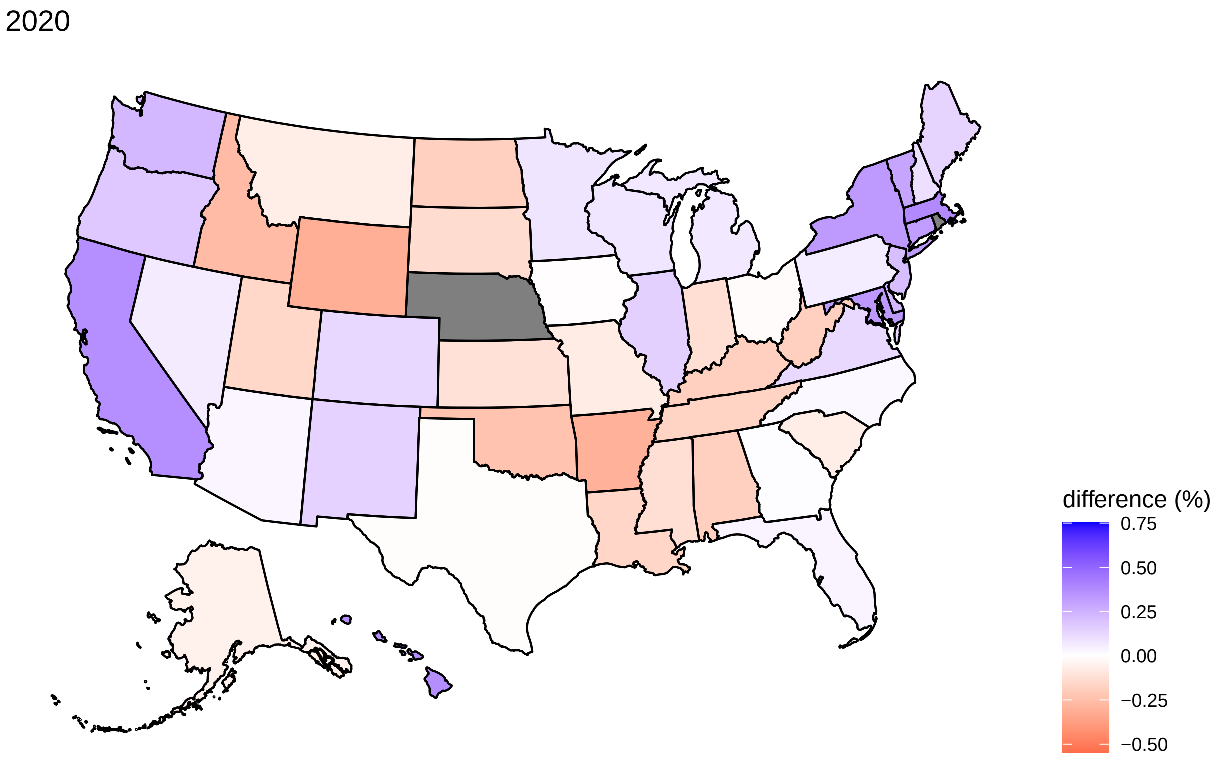
```r
state_poll_2020$state
```

```
##  [1] "Alabama"        "Alaska"         "Arizona"        "Arkansas"
##  [5] "California"     "Colorado"       "Connecticut"    "Delaware"
##  [9] "Florida"        "Georgia"        "Hawaii"         "Idaho"
## [13] "Illinois"       "Indiana"        "Iowa"           "Kansas"
## [17] "Kentucky"       "Louisiana"      "Maine"          "Maine CD-1"
## [21] "Maine CD-2"     "Maryland"       "Massachusetts"  "Michigan"
## [25] "Minnesota"      "Mississippi"    "Missouri"       "Montana"
## [29] "Nebraska CD-2"  "Nevada"         "New Hampshire"  "New Jersey"
## [33] "New Mexico"     "New York"       "North Carolina" "North Dakota"
## [37] "Ohio"           "Oklahoma"       "Oregon"         "Pennsylvania"
## [41] "South Carolina" "South Dakota"   "Tennessee"      "Texas"
## [45] "Utah"           "Vermont"        "Virginia"       "Washington"
## [49] "West Virginia"  "Wisconsin"      "Wyoming"
```

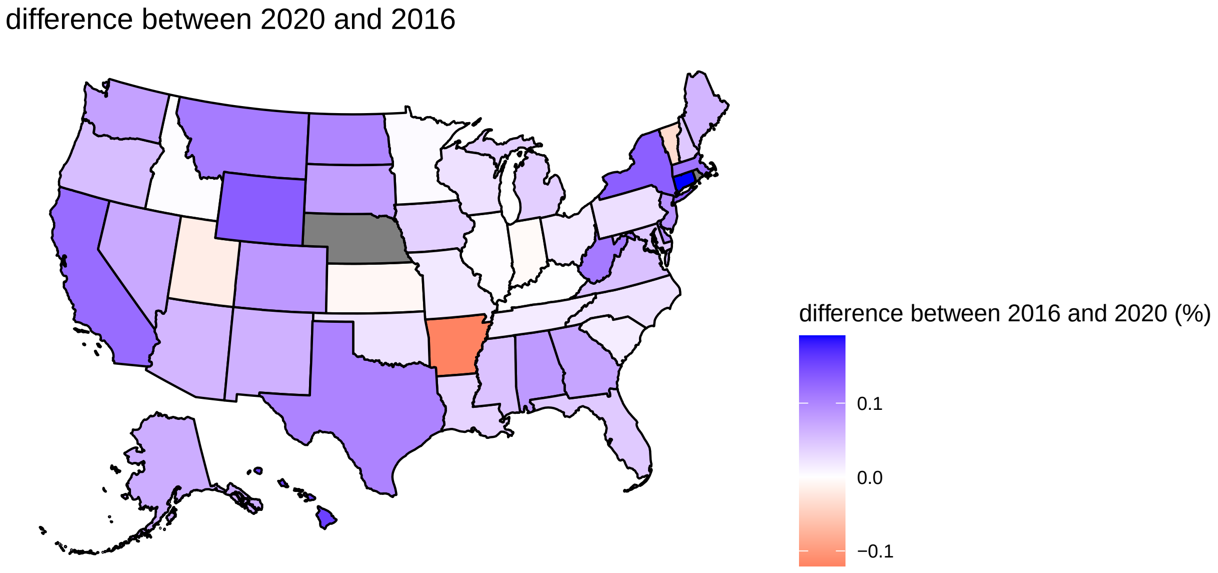
```r
state_poll_2016=state_poll_2016[-c(9,30,31,33,45),]


state_poll_2020_2016_diff <- data.frame(
  state =state_poll_2020$state,
  diff=state_poll_2020$diff_percentage-state_poll_2016$diff_percentage
)

plot_usmap(data = state_poll_2016, values = "diff_percentage", color = "black") +
  scale_fill_gradient2(name = "difference (%)",   low= "red",
                     mid = "white",
                     high = "blue",
                     midpoint = 0,limits=limit_val)+
  theme(legend.position = "right")+
ggtitle("2016")
```



```
plot_usmap(data = state_poll_2020, values = "diff_percentage", color = "black") +
  scale_fill_gradient2(name = "difference (%)",   low= "red",
                       mid = "white",
                       high = "blue",
                       midpoint = 0,limits=limit_val)+
  theme(legend.position = "right")+
ggtitle("2020")
```



```
plot_usmap(data = state_poll_2020_2016_diff, values = "diff", color = "black") +
  scale_fill_gradient2(name = "difference between 2016 and 2020 (%)",   low= "red",
                       mid = "white",
                       high = "blue",
                       midpoint = 0)+
  theme(legend.position = "right")+
  ggtitle("difference between 2020 and 2016")
```


difference between 2020 and 2016
difference between 2016 and 2020 (%)
0.1
0.0
−0.1

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

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.

Part A

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

```
state_poll_2020_2016_diff %>% kbl(digits = 3) %>% kable_styling(
  bootstrap_options = c("striped", "hover","condensed"))
```

```
merged_state = data.frame(State = state_poll_2016$state,
                          diff_2016 = state_poll_2016$diff_percentage,
                          diff_2020 = state_poll_2020$diff_percentage,
                          change = state_poll_2020_2016_diff$diff)

meged_final <- merged_state %>%
  arrange(desc(abs(change))) %>%
  kbl(digits = 3) %>%
  kable_styling(
    bootstrap_options =
      c("striped", "hover","condensed")
    )
```

```
meged_final
```

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	0.037
Kansas	-0.008
Kentucky	0.001
Louisiana	0.035
Maine	0.060
Maine CD-1	0.101
Maine CD-2	0.062
Maryland	0.044
Massachusetts	0.115
Michigan	0.039
Minnesota	0.004
Mississippi	0.049
Missouri	0.018
Montana	0.108
Nebraska CD-2	0.152
Nevada	0.071
New Hampshire	0.044
New Jersey	0.090
New Mexico	0.064
New York	0.132
North Carolina	0.022
North Dakota	0.099
Ohio	0.017
Oklahoma	0.024
Oregon	0.053
Pennsylvania	0.026
South Carolina	0.015
South Dakota	0.079
Tennessee	0.016
Texas	0.102
Utah	-0.018
Vermont	-0.035
Virginia	0.050
Washington	0.077
West Virginia	0.109
Wisconsin	0.024
Wyoming	0.134

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	0.063	0.152
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
West Virginia	-0.296	-0.187	0.109
Montana	-0.176	-0.067	0.108
Texas	-0.110	-0.008	0.102
Maine CD-1	0.158	0.259	0.101
North Dakota	-0.283	-0.183	0.099
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 CD-2	-0.043	0.019	0.062
Maine	0.078	0.138	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	0.037
Louisiana	-0.198	-0.162	0.035
Vermont	0.327	0.292	-0.035
Pennsylvania	0.035	0.060	0.026
Wisconsin	0.054	0.078	0.024
Oklahoma	-0.260	-0.236	0.024
North Carolina	-0.002	0.020	0.022
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

The table is organized based on the variance in percentages between the two elections, with larger changes appearing higher in the table. Positive values indicate counties that have shifted towards favoring the Democratic party, while negative values indicate counties that have shifted towards favoring the Republican party.

Part B

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

Based on the table, my prediction is that states with a minimal difference in 2020 but a substantial change value are likely to have the potential to flip in the near future. I believe Florida, Texas, Georgia, Arizona, and Michigan are among the possible states that could experience such a shift in the future.

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.

Florida and Texas have a chance of flipping but not Georgia, Michigan, Pennsylvania and Wisconsin not as much.

Question 5

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.

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.

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

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