

# Untitled

Haocheng Zhang

2023-05-28

```
##Q1
#load the data
polls_data_2016 = read.csv("president_general_polls_sorted_end_date_2016.csv")
library(lubridate)

##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##   date, intersect, setdiff, union

##dimension of this dataset
dim(polls_data_2016)

## [1] 12621    29

##variable name of the data
names(polls_data_2016)

## [1] "cycle"          "branch"          "type"            "matchup"
## [5] "forecastdate"   "state"           "startdate"       "enddate"
## [9] "pollster"       "grade"           "samplesize"     "population"
## [13] "poll_wt"        "rawpoll_clinton" "rawpoll_trump"   "rawpoll_johnson"
## [17] "rawpoll_mcmullin" "adjpoll_clinton" "adjpoll_trump"   "adjpoll_johnson"
## [21] "adjpoll_mcmullin" "multiversions"   "url"            "poll_id"
## [25] "question_id"    "createddate"     "timestamp"       "total.clinton"
## [29] "total.trump"

#Filter the data
date_2016= mdy(polls_data_2016$createddate)
index_selected=which(date_2016>='2016-08-01' & date_2016<='2016-11-02')
polls_data_2016 = polls_data_2016[index_selected,]

# Part a
index_Mic=which(polls_data_2016$state=="Michigan")

n1=sum(polls_data_2016$total.clinton[index_Mic])
n2=sum(polls_data_2016$total.trump[index_Mic])

n1/(n1+n2)
```

```
## [1] 0.5182924
```

```
n2/(n1+n2)
```

```
## [1] 0.4817076
```

```
n1/(n1+n2)-n2/(n1+n2)
```

```
## [1] 0.0365849
```

```
#The final result is 0.0365849, which shows Clinton was ahead with 3.65849%
```

```
index_Geo=which(polls_data_2016$state=="Georgia")
```

```
n1=sum(polls_data_2016$total.clinton[index_Geo])
```

```
n2=sum(polls_data_2016$total.trump[index_Geo])
```

```
n1/(n1+n2)
```

```
## [1] 0.4664512
```

```
n2/(n1+n2)
```

```
## [1] 0.5335488
```

```
n1/(n1+n2)-n2/(n1+n2)
```

```
## [1] -0.0670975
```

```
#The final result is -0.0670975, which shows Trump was ahead with 6.70975%
```

```
index_Nor=which(polls_data_2016$state=="North Carolina")
```

```
n1=sum(polls_data_2016$total.clinton[index_Nor])
```

```
n2=sum(polls_data_2016$total.trump[index_Nor])
```

```
n1/(n1+n2)
```

```
## [1] 0.4978619
```

```
n2/(n1+n2)
```

```
## [1] 0.5021381
```

```
n1/(n1+n2)-n2/(n1+n2)
```

```
## [1] -0.004276271
```

*#The final result is -0.004276271, which shows Trump was ahead with 0.4276271%*

*# Part b*

polls\_data\_2016\$enddate

##	[1]	"1/12/16"	"1/12/16"	"1/12/16"	"2/16/16"	"2/16/16"	"2/16/16"
##	[7]	"2/16/16"	"2/16/16"	"2/16/16"	"3/17/16"	"3/17/16"	"3/17/16"
##	[13]	"3/17/16"	"3/17/16"	"3/17/16"	"3/21/16"	"3/21/16"	"3/21/16"
##	[19]	"3/24/16"	"3/24/16"	"3/24/16"	"6/20/16"	"6/20/16"	"6/20/16"
##	[25]	"6/23/16"	"6/23/16"	"6/23/16"	"7/5/16"	"7/5/16"	"7/5/16"
##	[31]	"7/5/16"	"7/5/16"	"7/5/16"	"7/6/16"	"7/6/16"	"7/6/16"
##	[37]	"7/14/16"	"7/14/16"	"7/14/16"	"7/14/16"	"7/14/16"	"7/14/16"
##	[43]	"7/15/16"	"7/15/16"	"7/15/16"	"7/16/16"	"7/16/16"	"7/16/16"
##	[49]	"7/17/16"	"7/17/16"	"7/17/16"	"7/17/16"	"7/17/16"	"7/17/16"
##	[55]	"7/18/16"	"7/18/16"	"7/18/16"	"7/19/16"	"7/19/16"	"7/19/16"
##	[61]	"7/19/16"	"7/19/16"	"7/19/16"	"7/20/16"	"7/20/16"	"7/20/16"
##	[67]	"7/21/16"	"7/21/16"	"7/21/16"	"7/22/16"	"7/22/16"	"7/22/16"
##	[73]	"7/23/16"	"7/23/16"	"7/23/16"	"7/24/16"	"7/24/16"	"7/24/16"
##	[79]	"7/24/16"	"7/24/16"	"7/24/16"	"7/25/16"	"7/25/16"	"7/25/16"
##	[85]	"7/26/16"	"7/26/16"	"7/26/16"	"7/26/16"	"7/26/16"	"7/26/16"
##	[91]	"7/27/16"	"7/27/16"	"7/27/16"	"7/27/16"	"7/27/16"	"7/27/16"
##	[97]	"7/27/16"	"7/27/16"	"7/27/16"	"7/28/16"	"7/28/16"	"7/28/16"
##	[103]	"7/29/16"	"7/29/16"	"7/29/16"	"7/30/16"	"7/30/16"	"7/30/16"
##	[109]	"7/30/16"	"7/30/16"	"7/30/16"	"7/31/16"	"7/31/16"	"7/31/16"
##	[115]	"7/31/16"	"7/31/16"	"7/31/16"	"7/31/16"	"7/31/16"	"7/31/16"
##	[121]	"7/31/16"	"7/31/16"	"7/31/16"	"7/31/16"	"7/31/16"	"7/31/16"
##	[127]	"7/31/16"	"7/31/16"	"7/31/16"	"7/31/16"	"7/31/16"	"7/31/16"
##	[133]	"7/31/16"	"7/31/16"	"7/31/16"	"7/31/16"	"7/31/16"	"7/31/16"
##	[139]	"7/31/16"	"7/31/16"	"7/31/16"	"8/1/16"	"8/1/16"	"8/1/16"
##	[145]	"8/1/16"	"8/1/16"	"8/1/16"	"8/1/16"	"8/1/16"	"8/1/16"
##	[151]	"8/1/16"	"8/1/16"	"8/1/16"	"8/1/16"	"8/1/16"	"8/1/16"
##	[157]	"8/1/16"	"8/1/16"	"8/1/16"	"8/1/16"	"8/1/16"	"8/1/16"
##	[163]	"8/1/16"	"8/1/16"	"8/1/16"	"8/1/16"	"8/1/16"	"8/1/16"
##	[169]	"8/1/16"	"8/1/16"	"8/1/16"	"8/1/16"	"8/1/16"	"8/1/16"
##	[175]	"8/2/16"	"8/2/16"	"8/2/16"	"8/2/16"	"8/2/16"	"8/2/16"
##	[181]	"8/2/16"	"8/2/16"	"8/2/16"	"8/2/16"	"8/2/16"	"8/2/16"
##	[187]	"8/2/16"	"8/2/16"	"8/2/16"	"8/2/16"	"8/2/16"	"8/2/16"
##	[193]	"8/2/16"	"8/2/16"	"8/2/16"	"8/3/16"	"8/3/16"	"8/3/16"
##	[199]	"8/3/16"	"8/3/16"	"8/3/16"	"8/3/16"	"8/3/16"	"8/3/16"
##	[205]	"8/3/16"	"8/3/16"	"8/3/16"	"8/3/16"	"8/3/16"	"8/3/16"
##	[211]	"8/3/16"	"8/3/16"	"8/3/16"	"8/3/16"	"8/3/16"	"8/3/16"
##	[217]	"8/3/16"	"8/3/16"	"8/3/16"	"8/3/16"	"8/3/16"	"8/3/16"
##	[223]	"8/4/16"	"8/4/16"	"8/4/16"	"8/4/16"	"8/4/16"	"8/4/16"
##	[229]	"8/4/16"	"8/4/16"	"8/4/16"	"8/4/16"	"8/4/16"	"8/4/16"
##	[235]	"8/4/16"	"8/4/16"	"8/4/16"	"8/4/16"	"8/4/16"	"8/4/16"
##	[241]	"8/4/16"	"8/4/16"	"8/4/16"	"8/4/16"	"8/4/16"	"8/4/16"
##	[247]	"8/4/16"	"8/4/16"	"8/4/16"	"8/4/16"	"8/4/16"	"8/4/16"
##	[253]	"8/4/16"	"8/4/16"	"8/4/16"	"8/4/16"	"8/4/16"	"8/4/16"
##	[259]	"8/4/16"	"8/4/16"	"8/4/16"	"8/5/16"	"8/5/16"	"8/5/16"
##	[265]	"8/5/16"	"8/5/16"	"8/5/16"	"8/5/16"	"8/5/16"	"8/5/16"
##	[271]	"8/5/16"	"8/5/16"	"8/5/16"	"8/5/16"	"8/5/16"	"8/5/16"
##	[277]	"8/5/16"	"8/5/16"	"8/5/16"	"8/5/16"	"8/5/16"	"8/5/16"
##	[283]	"8/5/16"	"8/5/16"	"8/5/16"	"8/6/16"	"8/6/16"	"8/6/16"

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]

```
## [8389] "11/1/16" "11/1/16" "11/1/16" "11/1/16" "11/1/16" "11/1/16" "11/1/16"
## [8395] "11/1/16" "11/1/16" "11/1/16" "11/1/16" "11/1/16" "11/1/16" "11/1/16"
## [8401] "11/1/16" "11/1/16" "11/1/16" "11/1/16" "11/1/16" "11/1/16" "11/1/16"
## [8407] "11/1/16" "11/1/16" "11/1/16" "11/1/16" "11/1/16" "11/1/16" "11/1/16"
## [8413] "11/1/16" "11/1/16" "11/1/16" "11/1/16" "11/1/16" "11/1/16" "11/1/16"
## [8419] "11/1/16" "11/1/16" "11/1/16" "11/1/16" "11/1/16" "11/1/16" "11/1/16"
```

```
t.test(polls_data_2016$total.clinton[index_Mic]-
       polls_data_2016$total.trump[index_Mic],alternative='greater')
```

```
##
## One Sample t-test
##
## data: polls_data_2016$total.clinton[index_Mic] - polls_data_2016$total.trump[index_Mic]
## t = 10.484, df = 167, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
## 22.52678 Inf
## sample estimates:
## mean of x
## 26.74654
```

*##The t-test result shows a t-value of 10.484 with 167 degrees of freedom.  
 #The test's p-value is less than 2.2e-16, which is effectively zero and therefore  
 #statistically significant. This means there's a statistically significant  
 #difference between the counts in polls for Clinton and Trump in favor of  
 #Clinton. The confidence interval does not include 0 (22.52678 to infinity),  
 #and the estimated mean difference is 26.74654, both of which further support  
 #Clinton being favored in this state.*

```
t.test(polls_data_2016$total.clinton[index_Geo]-
       polls_data_2016$total.trump[index_Geo],alternative='greater')
```

```
##
## One Sample t-test
##
## data: polls_data_2016$total.clinton[index_Geo] - polls_data_2016$total.trump[index_Geo]
## t = -19.191, df = 164, p-value = 1
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
## -56.48463 Inf
## sample estimates:
## mean of x
## -52.00212
```

*##Here, the t-value is -19.191 with 164 degrees of freedom.  
 #The test's p-value is 1, indicating that the result is not statistically  
 #significant under a traditional alpha level (such as 0.05). The confidence  
 #interval goes from -56.48463 to infinity and the mean difference is negative  
 #(-52.00212). This suggests that Trump was favored in this state, but because  
 #the p-value is not statistically significant, we cannot conclude that  
 #the observed difference is not due to chance.*

```
t.test(polls_data_2016$total.clinton[index_Nor]-
      polls_data_2016$total.trump[index_Nor], alternative='greater')
```

```
##
## One Sample t-test
##
## data:  polls_data_2016$total.clinton[index_Nor] - polls_data_2016$total.trump[index_Nor]
## t = -1.5907, df = 263, p-value = 0.9436
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  -6.27371      Inf
## sample estimates:
## mean of x
##  -3.07883
```

*#The t-value is -1.5907 with 263 degrees of freedom. The p-value is 0.9436, suggesting that the result is not statistically significant at the usual alpha levels. The confidence interval ranges from -6.27371 to infinity, and the mean difference is negative (-3.07883). This implies that Trump has a higher count in polls, but as the p-value is not statistically significant, the observed difference could be due to chance.*

*#A potential problem with these tests could be the assumption of normality. The t-test assumes that the differences in the data are normally distributed. If this assumption is not met, the results may not be valid. Additionally, outliers could affect the results of the t-test, leading to biased results. Finally, these tests do not account for other factors that could influence the election results, such as demographic variables or the timing of the polls.*

*# Part c*

```
wilcox.test(polls_data_2016$total.clinton[index_Mic], polls_data_2016$total.trump[index_Mic], alternative='greater')
```

```
##
## Wilcoxon rank sum test with continuity correction
##
## data:  polls_data_2016$total.clinton[index_Mic] and polls_data_2016$total.trump[index_Mic]
## W = 16505, p-value = 0.003601
## alternative hypothesis: true location shift is greater than 0
```

*#The Wilcoxon rank sum test shows a W-value of 16505 and a p-value of 0.003601. Because the p-value is less than the traditional alpha level of 0.05, we can conclude that there is a statistically significant difference between the counts in polls for Clinton and Trump in favor of Clinton. This is because the test's alternative hypothesis is that the true location shift is greater than 0, which suggests a preference for Clinton.*

```
wilcox.test(polls_data_2016$total.clinton[index_Geo], polls_data_2016$total.trump[index_Geo], alternative='greater')
```

```
##
## Wilcoxon rank sum test with continuity correction
```

```
##
## data:  polls_data_2016$total.clinton[index_Geo] and polls_data_2016$total.trump[index_Geo]
## W = 9341, p-value = 1
## alternative hypothesis: true location shift is greater than 0
```

*#Here, the W-value is 9341, and the p-value is 1. Given that the p-value is not less than the usual alpha level of 0.05, we can't conclude that there is a statistically significant difference between the counts in polls for Clinton and Trump. Despite this, the alternative hypothesis suggests a preference for Clinton, but we can't confidently affirm this due to the lack of statistical significance.*

```
wilcox.test(polls_data_2016$total.clinton[index_Nor], polls_data_2016$total.trump[index_Nor], alternative="greater")
```

```
##
## Wilcoxon rank sum test with continuity correction
##
## data:  polls_data_2016$total.clinton[index_Nor] and polls_data_2016$total.trump[index_Nor]
## W = 33778, p-value = 0.7293
## alternative hypothesis: true location shift is greater than 0
```

*#The W-value is 33778, and the p-value is 0.7293. Since the p-value is not less than the standard alpha level of 0.05, the result is not statistically significant. Thus, there is no significant difference between the counts in polls for Clinton and Trump in this state. Although the alternative hypothesis implies a preference for Clinton, we can't confirm this because of the test's non-significant outcome.*

*##Potential issues with the Wilcoxon signed-rank test include the assumption of symmetry, i.e., the distribution of differences between the pairs should be symmetrical. If this assumption is not met, the results may not be valid. Furthermore, outliers can impact the Wilcoxon test as the rank of the outliers will be affected, thereby influencing the test results. As with the t-test, these tests do not account for other factors that could influence the election results, such as demographic variables or the timing of the polls.*

*# Part d*

```
date_Mic <- mdy(polls_data_2016$enddate[index_Mic])
date_Geo <- mdy(polls_data_2016$enddate[index_Geo])
date_Nor <- mdy(polls_data_2016$enddate[index_Nor])

counts_Mic_for_lm <- data.frame(
  data_date = date_Mic,
  percentage_diff = (polls_data_2016$total.clinton[index_Mic]-
    polls_data_2016$total.trump[index_Mic])/
    (polls_data_2016$total.clinton[index_Mic]+
    polls_data_2016$total.trump[index_Mic])
)
as.numeric(date_Mic)
```

```
## [1] 17014 17014 17014 17017 17017 17017 17023 17023 17023 17029 17029 17029
## [13] 17031 17031 17031 17036 17036 17036 17037 17037 17037 17038 17038 17038
```

```
## [25] 17041 17041 17041 17043 17043 17043 17045 17045 17045 17045 17045 17045
## [37] 17050 17050 17050 17051 17051 17051 17052 17052 17052 17057 17057 17057
## [49] 17057 17057 17057 17059 17059 17059 17064 17064 17064 17066 17066 17066
## [61] 17068 17068 17068 17070 17070 17070 17071 17071 17071 17072 17072 17072
## [73] 17073 17073 17073 17077 17077 17077 17077 17077 17077 17080 17080 17080
## [85] 17083 17083 17083 17085 17085 17085 17085 17085 17085 17087 17087 17087
## [97] 17088 17088 17088 17090 17090 17090 17092 17092 17092 17093 17093 17093
## [109] 17093 17093 17093 17094 17094 17094 17097 17097 17097 17098 17098 17098
## [121] 17098 17098 17098 17099 17099 17099 17099 17099 17099 17100 17100 17100
## [133] 17100 17100 17100 17101 17101 17101 17102 17102 17102 17104 17104 17104
## [145] 17104 17104 17104 17104 17104 17104 17104 17104 17104 17105 17105 17105
## [157] 17105 17105 17105 17105 17105 17105 17106 17106 17106 17106 17106 17106
```

```
lm_model_Mic=lm(percentage_diff~(data_date),data=counts_Mic_for_lm)
summary(lm_model_Mic)
```

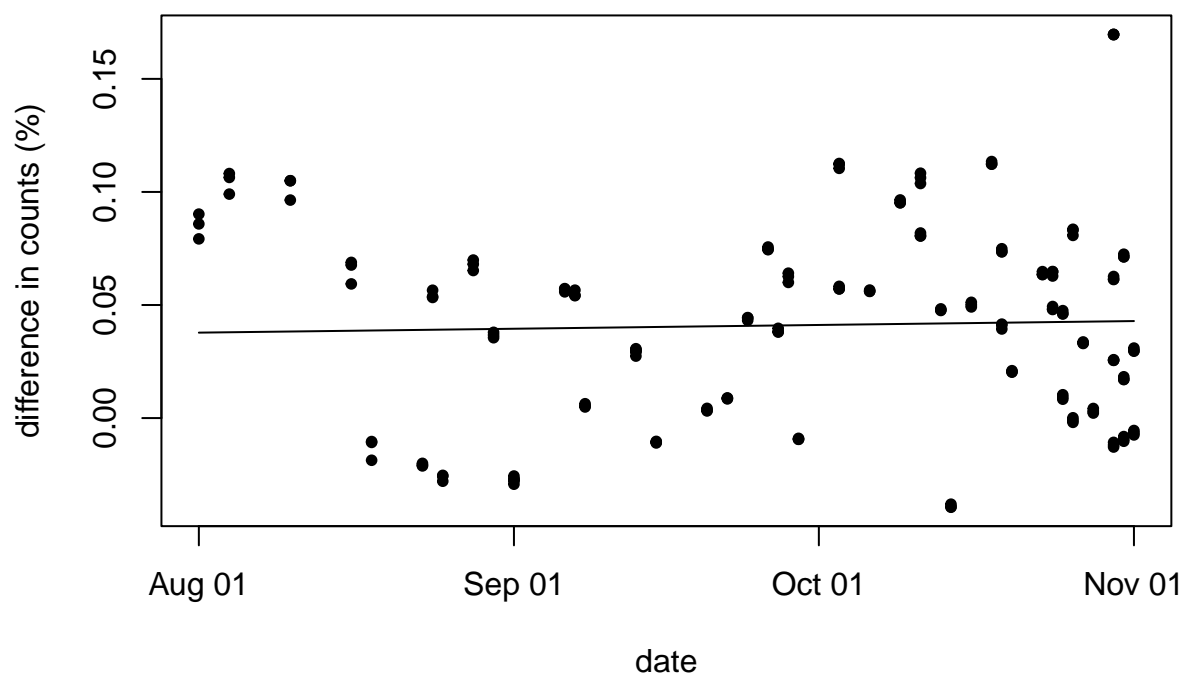
```
##
## Call:
## lm(formula = percentage_diff ~ (data_date), data = counts_Mic_for_lm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.081321 -0.035304  0.003581  0.028445  0.126897
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.087e-01  2.123e+00  -0.428   0.669
## data_date    5.563e-05  1.243e-04   0.447   0.655
##
## Residual standard error: 0.04336 on 166 degrees of freedom
## Multiple R-squared:  0.001205,    Adjusted R-squared:  -0.004812
## F-statistic: 0.2002 on 1 and 166 DF,  p-value: 0.6551
```

```
#The fitted linear model's coefficient for data_date (representing time) is
#5.563e-05, which is not statistically significant (p = 0.655).
#This means there is no statistically significant linear trend in the percentage
#difference in polls over time for this state. The Adjusted R-squared value is
#-0.004812, indicating the model explains very little of the variance in the data.
```

```
plot(counts_Mic_for_lm$data_date,counts_Mic_for_lm$percentage_diff,
     col='black',pch=20,type='p',xlab='date',ylab='difference in counts (%)',main='Michigan')
lines(counts_Mic_for_lm$data_date,lm_model_Mic$fitted.values,
     col='black',pch=20,type='l',xlab='date',ylab='difference in counts (%)',main='Michigan')
```



## Michigan



*#The plot is slightly increasing, showing that the winning rate for Clinton is  
#increasing slowly.*

```
counts_Geo_for_lm <- data.frame(
  data_date = date_Geo,
  percentage_diff = (polls_data_2016$total.clinton[index_Geo]-
    polls_data_2016$total.trump[index_Geo])/
    (polls_data_2016$total.clinton[index_Geo]+
    polls_data_2016$total.trump[index_Geo])
)
as.numeric(date_Geo)
```

```
## [1] 17013 17013 17013 17013 17013 17013 17017 17017 17017 17020 17020 17020
## [13] 17021 17021 17021 17025 17025 17025 17029 17029 17029 17030 17030 17030
## [25] 17031 17031 17031 17036 17036 17036 17038 17038 17038 17043 17043 17043
## [37] 17045 17045 17045 17045 17045 17045 17050 17050 17050 17052 17052 17052
## [49] 17052 17052 17052 17057 17057 17057 17057 17057 17057 17058 17058 17058
## [61] 17059 17059 17059 17062 17062 17062 17064 17064 17064 17065 17065 17065
## [73] 17066 17066 17066 17066 17066 17066 17066 17066 17066 17070 17070 17070
## [85] 17073 17073 17073 17077 17077 17077 17080 17080 17080 17083 17083 17083
## [97] 17086 17086 17086 17087 17087 17087 17088 17088 17088 17090 17090 17090
## [109] 17092 17092 17092 17092 17092 17092 17093 17093 17093 17094 17094 17094
## [121] 17094 17094 17094 17094 17094 17094 17094 17094 17094 17098 17098 17098
## [133] 17099 17099 17099 17100 17100 17100 17100 17100 17100 17101 17101 17101
## [145] 17101 17101 17101 17102 17102 17102 17104 17104 17104 17105 17105 17105
```

```
## [157] 17105 17105 17105 17105 17105 17105 17106 17106 17106
```

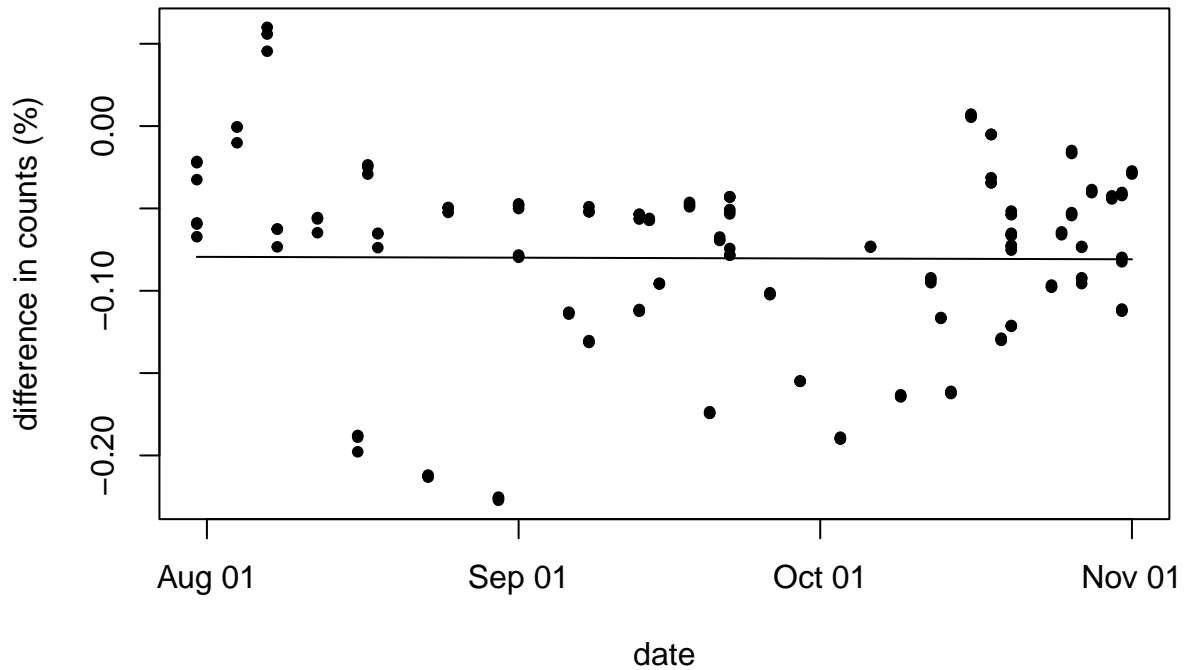
```
lm_model_Geo=lm(percentage_diff~(data_date),data=counts_Geo_for_lm)
summary(lm_model_Geo)
```

```
##
## Call:
## lm(formula = percentage_diff ~ (data_date), data = counts_Geo_for_lm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.14732 -0.03146  0.01223  0.03239  0.13947
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.920e-01  2.610e+00   0.074   0.941
## data_date   -1.595e-05  1.529e-04  -0.104   0.917
##
## Residual standard error: 0.05609 on 163 degrees of freedom
## Multiple R-squared:  6.675e-05, Adjusted R-squared:  -0.006068
## F-statistic: 0.01088 on 1 and 163 DF, p-value: 0.917
```

```
#Here, the coefficient for data_date is -1.595e-05, which is also not
#statistically significant (p = 0.917). This means there is no significant
#linear trend in the percentage difference in polls over time for this state.
#The Adjusted R-squared value is -0.006068, suggesting that the model doesn't
#explain the variance in the data well.
```

```
plot(counts_Geo_for_lm$data_date,counts_Geo_for_lm$percentage_diff,
      col='black',pch=20,type='p',xlab='date',ylab='difference in counts (%)',main='Georgia')
lines(counts_Geo_for_lm$data_date,lm_model_Geo$fitted.values,
      col='black',pch=20,type='l',xlab='date',ylab='difference in counts (%)',main='Georgia')
```

## Georgia



*#The plot is almost horizontal, showing that the winning rate for Clinton/Trump  
#is not significantly changing. So it would be the closest race.*

```
counts_Nor_for_lm <- data.frame(
  data_date = date_Nor,
  percentage_diff = (polls_data_2016$total.clinton[index_Nor]-
    polls_data_2016$total.trump[index_Nor])/
    (polls_data_2016$total.clinton[index_Nor]+
    polls_data_2016$total.trump[index_Nor])
)
as.numeric(date_Nor)
```

```
## [1] 17015 17015 17015 17020 17020 17020 17023 17023 17023 17029 17029 17029
## [13] 17030 17030 17030 17031 17031 17031 17036 17036 17036 17036 17036 17036
## [25] 17036 17036 17036 17038 17038 17038 17040 17040 17040 17042 17042 17042
## [37] 17043 17043 17043 17045 17045 17045 17045 17045 17045 17046 17046 17046
## [49] 17050 17050 17050 17051 17051 17051 17051 17051 17051 17052 17052 17052
## [61] 17056 17056 17056 17057 17057 17057 17059 17059 17059 17060 17060 17060
## [73] 17063 17063 17063 17063 17063 17063 17064 17064 17064 17064 17064 17064
## [85] 17064 17064 17064 17066 17066 17066 17066 17066 17066 17066 17066 17066
## [97] 17067 17067 17067 17070 17070 17070 17072 17072 17072 17073 17073 17073
## [109] 17074 17074 17074 17076 17076 17076 17076 17076 17076 17077 17077 17077
## [121] 17077 17077 17077 17080 17080 17080 17080 17080 17080 17083 17083 17083
## [133] 17086 17086 17086 17086 17086 17086 17086 17086 17086 17087 17087 17087
## [145] 17088 17088 17088 17089 17089 17089 17090 17090 17090 17090 17090 17090
```

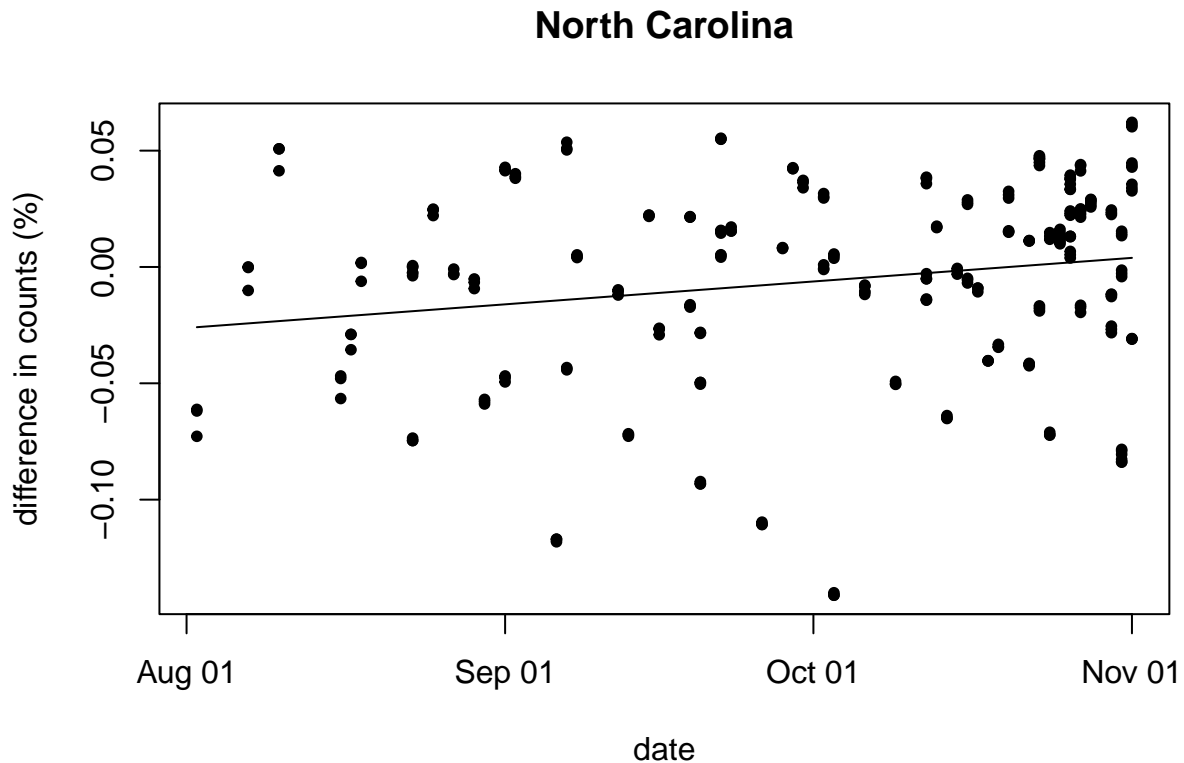
```
## [157] 17091 17091 17091 17092 17092 17092 17093 17093 17093 17094 17094 17094
## [169] 17094 17094 17094 17096 17096 17096 17096 17096 17096 17097 17097 17097
## [181] 17097 17097 17097 17097 17097 17097 17098 17098 17098 17098 17098 17098
## [193] 17099 17099 17099 17099 17099 17099 17100 17100 17100 17100 17100 17100
## [205] 17100 17100 17100 17100 17100 17100 17100 17100 17100 17101 17101 17101
## [217] 17101 17101 17101 17101 17101 17101 17101 17101 17101 17102 17102 17102
## [229] 17102 17102 17102 17104 17104 17104 17104 17104 17104 17104 17104 17104
## [241] 17105 17105 17105 17105 17105 17105 17105 17105 17105 17105 17105 17105
## [253] 17106 17106 17106 17106 17106 17106 17106 17106 17106 17106 17106 17106
```

```
lm_model_Nor=lm(percentage_diff~(data_date),data=counts_Nor_for_lm)
summary(lm_model_Nor)
```

```
##
## Call:
## lm(formula = percentage_diff ~ (data_date), data = counts_Nor_for_lm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.13551 -0.02274  0.01072  0.02690  0.07406
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.5960539   1.7205518   -3.252  0.00129 **
## data_date    0.0003274   0.0001008    3.249  0.00131 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04147 on 262 degrees of freedom
## Multiple R-squared:  0.03873,    Adjusted R-squared:  0.03506
## F-statistic: 10.56 on 1 and 262 DF,  p-value: 0.001309
```

```
#The coefficient for data_date is 0.0003274, which is statistically
#significant (p = 0.00131). This suggests there is a significant positive
#linear trend in the percentage difference in polls over time for this state.
#The Adjusted R-squared value is 0.03506, indicating that the model explains
#about 3.5% of the variance in the data, which is relatively low, but higher
#than the other two states.
```

```
plot(counts_Nor_for_lm$data_date,count_Nor_for_lm$percentage_diff,
     col='black',pch=20,type='p',xlab='date',ylab='difference in counts (%)',main='North Carolina')
lines(counts_Nor_for_lm$data_date,lm_model_Nor$fitted.values,
     col='black',pch=20,type='l',xlab='date',ylab='difference in counts (%)',main='North Carolina')
```



*#The plot is slightly increasing, showing that the winning rate for Clinton is  
#increasing slowly.*

#### **##Part e**

*#In reality North Carolina voted 49.83% Trump 46.17% Clinton,  
#Michigan voted 47.25% Trump 47.03% Clinton and Georgia voted 50.38%  
#Trump 45.29% Clinton. The smallest margin was in Michigan with 0.22%  
#Based on the provided real election data, indeed, the state with the smallest  
#margin in the 2016 election is Georgia, with a difference of only 0.22%.  
#However, if we compare these results with the predictions from the polls,  
#there might be some discrepancies. As per the polls, the state of Michigan was  
#expected to favor Clinton significantly, whereas the actual results show a rather  
#close election. Similarly, in Georgia and North Carolina, the margin of  
#difference between the two candidates in actual results appears to be  
#smaller than predicted by the polls.*

#### **##Potential problems:**

- #1. Sampling bias: Polls are based on a sample of voters rather than the entire  
#population of voters.*
- #2. Late decision-makers: Some voters might not have made up their minds when  
#they were polled but decided later on, and these late deciders might have  
#favored Trump over Biden. These changes would not have been captured in polls.*

#### **##Part f**

*#In Michigan, the polls indicated that Clinton would win (average of 26.7% more*

#votes than Trump). According to the data we got, Clinton won,  
#but the margin was only 2.78%. This indicates a significant overestimation of #Democratic support in t

#In Georgia, the polls showed Trump winning (average of 52% more votes than  
#Clinton). The actual result, however, was a narrow win for Clinton (0.23%  
#margin). This suggests a significant underestimation of Democratic support  
#in the polls.

#In North Carolina, the polls indicated a close race, but slightly favored Trump  
#(average of 3.08% more votes than Clinton). The actual results showed a  
#win for Clinton, albeit with a -1.34% margin.  
#This again suggests an underestimation of Democratic support in the polls.

#### ##Potential problems:

#1. differential turnout: The voters who actually turn up to vote on Election Day  
#may not be the same as those who responded to the polls. For example,  
#if Trump supporters were more motivated to vote than Biden supporters,  
#this could have led to Trump receiving a higher proportion of the votes than  
#expected.

#2. Shy Trump voters: Some research suggests that there may be a subset of Trump  
#voters who were reluctant to express their support for him in a polling  
#situation, but who did vote for him on Election Day. If this were the case,  
#it could explain why Trump's actual vote share was higher than the polls  
#predicted.

# Untitled

Haocheng Zhang

2023-05-28

```
##Q2
#load the data
polls_data_2020 = read.csv("president_polls_2020.csv")
library(lubridate)

##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##   date, intersect, setdiff, union

#Filter the data
date_2020= mdy(polls_data_2020$end_date)
index_selected=which(date_2020>='2020-08-01' & date_2020<='2020-11-02')
polls_data_2020 = polls_data_2020[index_selected,]

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,]
  }
}
index_NA=which(is.na(polls_data_2020$sample_size)==T)
index_NA

## [1] 6459 6460

polls_data_2020=polls_data_2020[-index_NA,]

# Part a
index_biden_Mic_2020=which(polls_data_2020$answer=='Biden' & polls_data_2020$state=="Michigan")
index_trump_Mic_2020=which(polls_data_2020$answer=='Trump' & polls_data_2020$state=="Michigan")

counts_biden_Mic_2020=polls_data_2020$pct[index_biden_Mic_2020]*
polls_data_2020$sample_size[index_biden_Mic_2020]
counts_trump_Mic_2020=polls_data_2020$pct[index_trump_Mic_2020]*
```

```
polls_data_2020$sample_size[index_trump_Mic_2020]
```

```
n1_2020_Mic=sum(counts_biden_Mic_2020)
```

```
n2_2020_Mic=sum(counts_trump_Mic_2020)
```

```
n1_2020_Mic
```

```
## [1] 15356023
```

```
n2_2020_Mic
```

```
## [1] 13020191
```

```
(n1_2020_Mic-n2_2020_Mic)/(n1_2020_Mic+n2_2020_Mic)
```

```
## [1] 0.08231654
```

```
#The final result is 0.08231654, which shows Biden was ahead with 8.231654%
```

```
index_biden_Geo_2020=which(polls_data_2020$answer=='Biden' & polls_data_2020$state=="Georgia")
```

```
index_trump_Geo_2020=which(polls_data_2020$answer=='Trump' & polls_data_2020$state=="Georgia")
```

```
counts_biden_Geo_2020=polls_data_2020$pct[index_biden_Geo_2020]*
```

```
polls_data_2020$sample_size[index_biden_Geo_2020]
```

```
counts_trump_Geo_2020=polls_data_2020$pct[index_trump_Geo_2020]*
```

```
polls_data_2020$sample_size[index_trump_Geo_2020]
```

```
n1_2020_Geo=sum(counts_biden_Geo_2020)
```

```
n2_2020_Geo=sum(counts_trump_Geo_2020)
```

```
n1_2020_Geo
```

```
## [1] 13307078
```

```
n2_2020_Geo
```

```
## [1] 12437490
```

```
(n1_2020_Geo-n2_2020_Geo)/(n1_2020_Geo+n2_2020_Geo)
```

```
## [1] 0.03377752
```

```
#The final result is 0.03377752, which shows Biden was ahead with 3.377752%
```

```
index_biden_Nor_2020=which(polls_data_2020$answer=='Biden' & polls_data_2020$state=="North Carolina")
```

```
index_trump_Nor_2020=which(polls_data_2020$answer=='Trump' & polls_data_2020$state=="North Carolina")
```

```
counts_biden_Nor_2020=polls_data_2020$pct[index_biden_Nor_2020]*
```



```
polls_data_2020$sample_size[index_biden_Nor_2020]
counts_trump_Nor_2020=polls_data_2020$pct[index_trump_Nor_2020]*
polls_data_2020$sample_size[index_trump_Nor_2020]
```

```
n1_2020_Nor=sum(counts_biden_Nor_2020)
n2_2020_Nor=sum(counts_trump_Nor_2020)
```

```
n1_2020_Nor
```

```
## [1] 16857852
```

```
n2_2020_Nor
```

```
## [1] 14997186
```

```
(n1_2020_Nor-n2_2020_Nor)/(n1_2020_Nor+n2_2020_Nor)
```

```
## [1] 0.05841043
```

```
#The final result is 0.05841043, which shows Biden was ahead with 5.841043%
```

```
# Part b
```

```
library(lubridate)
polls_data_2020$enddate
```

```
## NULL
```

```
t.test(polls_data_2020$pct[index_biden_Mic_2020]*polls_data_2020$
sample_size[index_biden_Mic_2020]-polls_data_2020$pct[index_trump_Mic_2020]*
polls_data_2020$sample_size[index_trump_Mic_2020],alternative='greater')
```

```
##
```

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

```
##
```

```
## data: polls_data_2020$pct[index_biden_Mic_2020] * polls_data_2020$sample_size[index_biden_Mic_2020]
```

```
## t = 10.147, df = 101, p-value < 2.2e-16
```

```
## alternative hypothesis: true mean is greater than 0
```

```
## 95 percent confidence interval:
```

```
## 19153.57 Inf
```

```
## sample estimates:
```

```
## mean of x
```

```
## 22900.31
```

```
##The t-test result shows a t-value of 10.147 with 101 degrees of freedom.
#The test's p-value is less than 2.2e-16, which is effectively zero and therefore
#statistically significant. This means there's a statistically significant
#difference between the counts in polls for Biden and Trump in favor of
#Biden. The confidence interval does not include 0 (22.52678 to infinity),
#and the estimated mean difference is 22900, both of which further support
#Biden being favored in this state.
```

```
t.test(polls_data_2020$pct[index_biden_Geo_2020]*polls_data_2020$
sample_size[index_biden_Geo_2020]-polls_data_2020$pct[index_trump_Geo_2020]*
polls_data_2020$sample_size[index_trump_Geo_2020],alternative='greater')
```

```
##
## One Sample t-test
##
## data:  polls_data_2020$pct[index_biden_Geo_2020] * polls_data_2020$sample_size[index_biden_Geo_2020]
## t = 6.8926, df = 76, p-value = 6.991e-10
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  8565.036      Inf
## sample estimates:
## mean of x
## 11293.34
```

*##The t-test result shows a t-value of 6.8926 with 76 degrees of freedom.  
 #The test's p-value is less than 6.991e-10, which is effectively zero and  
 #therefore statistically significant. This means there's a statistically  
 #significant difference between the counts in polls for Biden and Trump in favor  
 #Biden. The confidence interval does not include 0 (22.52678 to infinity),  
 #and the estimated mean difference is 11293.34, both of which further support  
 #Biden being favored in this state.*

```
t.test(polls_data_2020$pct[index_biden_Nor_2020]*polls_data_2020$
sample_size[index_biden_Nor_2020]-polls_data_2020$pct[index_trump_Nor_2020]*
polls_data_2020$sample_size[index_trump_Nor_2020],alternative='greater')
```

```
##
## One Sample t-test
##
## data:  polls_data_2020$pct[index_biden_Nor_2020] * polls_data_2020$sample_size[index_biden_Nor_2020]
## t = 8.4302, df = 107, p-value = 8.801e-14
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  13837.52      Inf
## sample estimates:
## mean of x
## 17228.39
```

*##The t-test result shows a t-value of 8.4302 with 107 degrees of freedom.  
 #The test's p-value is less than 8.801e-14, which is effectively zero and  
 #therefore statistically significant. This means there's a statistically  
 #significant difference between the counts in polls for Biden and Trump in favor  
 #Biden. The confidence interval does not include 0 (22.52678 to infinity),  
 #and the estimated mean difference is 17228.39, both of which further support  
 #Biden being favored in this state.*

*#A potential problem with these tests could be the assumption of normality.  
 #The t-test assumes that the differences in the data are normally distributed.  
 #If this assumption is not met, the results may not be valid. Additionally,  
 #outliers could affect the results of the t-test, leading to biased results.*

*#Finally, these tests do not account for other factors that could influence  
#the election results, such as demographic variables or the timing of the polls.*

*# Part c*

```
wilcox.test(polls_data_2020$pct[index_biden_Mic_2020]*polls_data_2020$  
            sample_size[index_biden_Mic_2020],polls_data_2020$  
            pct[index_trump_Mic_2020]*  
            polls_data_2020$sample_size[index_trump_Mic_2020],  
            alternative = "greater")
```

```
##  
## Wilcoxon rank sum test with continuity correction  
##  
## data:  polls_data_2020$pct[index_biden_Mic_2020] * polls_data_2020$sample_size[index_biden_Mic_2020]  
## W = 6042, p-value = 0.02322  
## alternative hypothesis: true location shift is greater than 0
```

*#The Wilcoxon rank sum test shows a W-value of 6042 and a p-value of 0.02322.  
#Because the p-value is less than the traditional alpha level of 0.05, we can  
#conclude that there is a statistically significant difference between the  
#counts in polls for Biden and Trump in favor of Clinton. This is because  
#the test's alternative hypothesis is that the true location shift is greater  
#than 0, which suggests a preference for Biden.*

```
wilcox.test(polls_data_2020$pct[index_biden_Geo_2020]*  
            polls_data_2020$sample_size[index_biden_Geo_2020],  
            polls_data_2020$pct[index_trump_Geo_2020]*  
            polls_data_2020$sample_size[index_trump_Geo_2020],  
            alternative='greater')
```

```
##  
## Wilcoxon rank sum test with continuity correction  
##  
## data:  polls_data_2020$pct[index_biden_Geo_2020] * polls_data_2020$sample_size[index_biden_Geo_2020]  
## W = 3203, p-value = 0.1949  
## alternative hypothesis: true location shift is greater than 0
```

*#The Wilcoxon rank sum test shows a W-value of 3203 and a p-value of 0.1949.  
#Because the p-value is greater than the traditional alpha level of 0.05,  
#we can't conclude that there is a  
#statistically significant difference between the counts in polls for Biden  
#and Trump. Despite this, the alternative hypothesis suggests a preference for  
#Biden, but we can't confidently affirm this due to the lack of statistical #significance.*

```
wilcox.test(polls_data_2020$pct[index_biden_Nor_2020]*polls_data_2020$  
            sample_size[index_biden_Nor_2020]-polls_data_2020$  
            pct[index_trump_Nor_2020]*polls_data_2020$  
            sample_size[index_trump_Nor_2020],  
            alternative='greater')
```

```
##
## Wilcoxon signed rank test with continuity correction
##
## data:  polls_data_2020$pct[index_biden_Nor_2020] * polls_data_2020$sample_size[index_biden_Nor_2020]
## V = 4999, p-value = 1.205e-15
## alternative hypothesis: true location is greater than 0
```

*#The Wilcoxon rank sum test shows a W-value of 4999 and a p-value of 1.205e-15.  
#Because the p-value is less than the traditional alpha level of 0.05, we can  
#conclude that there is a statistically significant difference between the  
#counts in polls for Biden and Trump in favor of Clinton. This is because  
#the test's alternative hypothesis is that the true location shift is greater  
#than 0, which suggests a preference for Biden.*

*##Potential issues with the Wilcoxon signed-rank test include the assumption of  
#symmetry, i.e., the distribution of differences between the pairs should be  
#symmetrical. If this assumption is not met, the results may not be valid.  
#Furthermore, outliers can impact the Wilcoxon test as the rank of the outliers  
#will be affected, thereby influencing the test results. As with the t-test,  
#these tests do not account for other factors that could influence the election  
#results, such as demographic variables or the timing of the polls.*

*# Part d*

```
counts_Mic_for_lm_2020 <- data.frame(
  data_date = date_2020[index_trump_Mic_2020],
  percentage_diff = (counts_biden_Mic_2020-counts_trump_Mic_2020)/
    (counts_biden_Mic_2020+counts_trump_Mic_2020)
)

lm_model_Mic_2020=lm(percentage_diff~(data_date),data=counts_Mic_for_lm_2020)
summary(lm_model_Mic_2020)
```

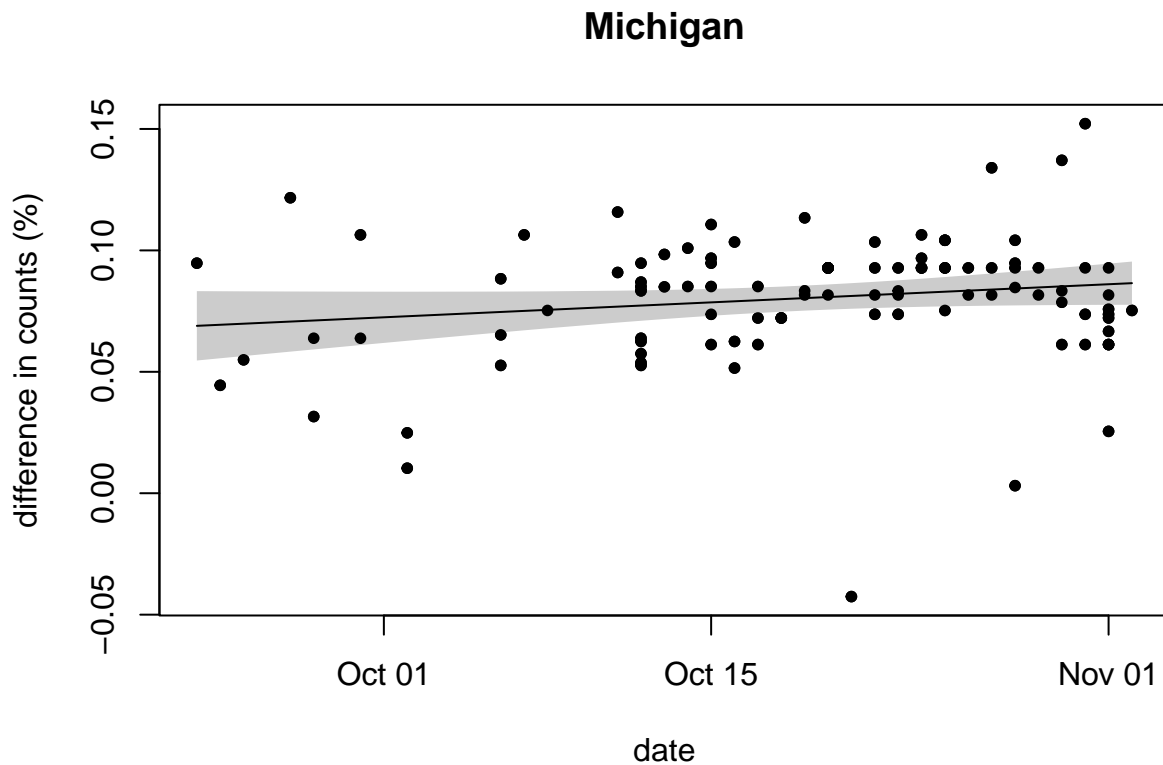
```
##
## Call:
## lm(formula = percentage_diff ~ (data_date), data = counts_Mic_for_lm_2020)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.123785 -0.011763  0.002129  0.011990  0.066550
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.0684484   4.8076060  -1.678   0.0964 .
## data_date    0.0004392   0.0002591   1.695   0.0932 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02624 on 100 degrees of freedom
## Multiple R-squared:  0.02793,    Adjusted R-squared:  0.01821
## F-statistic: 2.873 on 1 and 100 DF,  p-value: 0.09319
```

```

conf_interval_Mic_fitted_2020= predict(lm_model_Mic_2020, newdata=counts_Mic_for_lm_2020, interval="conf",
                                     level = 0.95)

plot(counts_Mic_for_lm_2020$data_date,counts_Mic_for_lm_2020$percentage_diff,
     col='black',pch=20,type='p',xlab='date',ylab='difference in counts (%)',main='Michigan')
polygon(c(rev(counts_Mic_for_lm_2020$data_date),
          counts_Mic_for_lm_2020$data_date),
        c(rev(conf_interval_Mic_fitted_2020[,2]),
          conf_interval_Mic_fitted_2020[,3]), col = 'grey80', border = NA)
lines(counts_Mic_for_lm_2020$data_date,lm_model_Mic_2020$fitted.values,
      col='black',pch=20,type='l',xlab='date',ylab='difference in counts (%)',main='Michigan')
lines(counts_Mic_for_lm_2020$data_date,counts_Mic_for_lm_2020$percentage_diff,
      col='black',pch=20,type='p',xlab='date',ylab='difference in counts (%)',main='Michigan')

```



*#The plot is slightly increasing, showing that the winning rate for Biden is  
#increasing slowly.*

```

counts_Geo_for_lm_2020 <- data.frame(
  data_date = date_2020[index_trump_Geo_2020],
  percentage_diff = (counts_biden_Geo_2020-counts_trump_Geo_2020)/(counts_biden_Geo_2020+counts_trump_Geo_2020)
)

lm_model_Geo_2020=lm(percentage_diff~(data_date),data=counts_Geo_for_lm_2020)
summary(lm_model_Geo_2020)

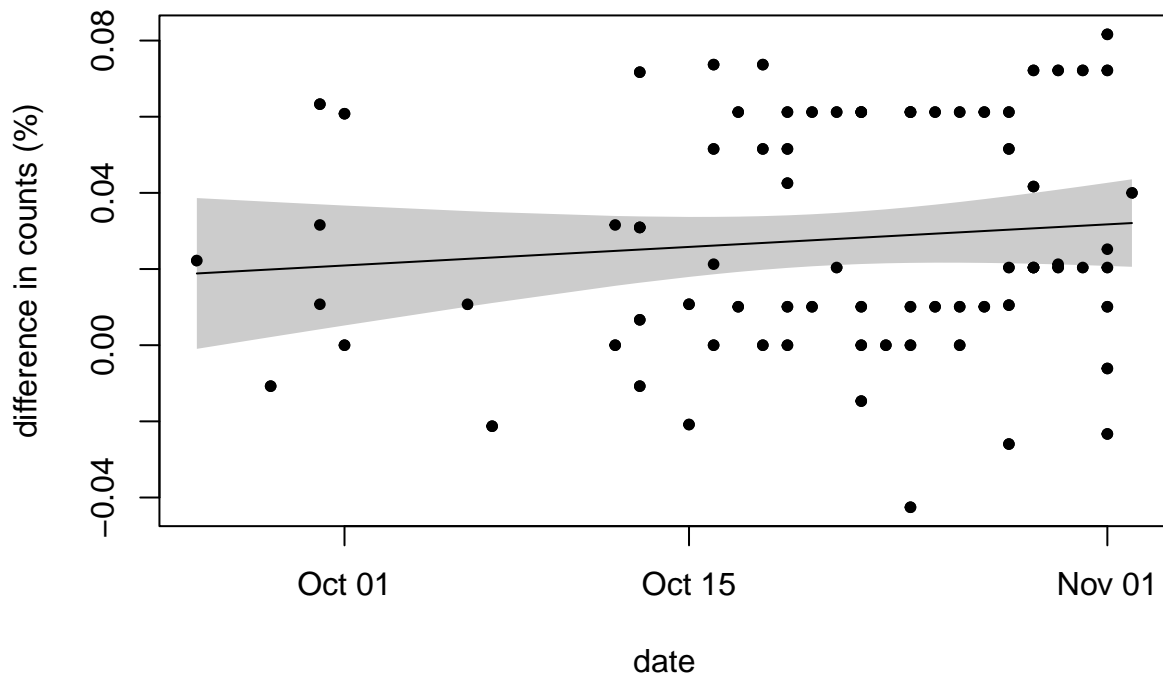
```

```
##
## Call:
## lm(formula = percentage_diff ~ (data_date), data = counts_Geo_for_lm_2020)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.071497 -0.019887 -0.009757  0.031584  0.049903
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.4355049   6.8214883   -0.943   0.348
## data_date    0.0003483   0.0003676    0.947   0.346
##
## Residual standard error: 0.03006 on 75 degrees of freedom
## Multiple R-squared:  0.01183,    Adjusted R-squared:  -0.001348
## F-statistic: 0.8977 on 1 and 75 DF,  p-value: 0.3464

conf_interval_Geo_fitted_2020= predict(lm_model_Geo_2020, newdata=counts_Geo_for_lm_2020, interval="conf",
                                     level = 0.95)

plot(counts_Geo_for_lm_2020$data_date,counts_Geo_for_lm_2020$percentage_diff,
     col='black',pch=20,type='p',xlab='date',ylab='difference in counts (%)',main='Georgia')
polygon(c(rev(counts_Geo_for_lm_2020$data_date), counts_Geo_for_lm_2020$data_date),
       c(rev(conf_interval_Geo_fitted_2020[,2]), conf_interval_Geo_fitted_2020[,3]), col = 'grey80', lty=2)
lines(counts_Geo_for_lm_2020$data_date,lm_model_Geo_2020$fitted.values,
     col='black',pch=20,type='l',xlab='date',ylab='difference in counts (%)',main='Georgia')
lines(counts_Geo_for_lm_2020$data_date,counts_Geo_for_lm_2020$percentage_diff,
     col='black',pch=20,type='p',xlab='date',ylab='difference in counts (%)',main='Georgia')
```

## Georgia



*#The plot is slightly increasing, showing that the winning rate for Biden is  
#increasing slowly.*

```
counts_Nor_for_lm_2020 <- data.frame(
  data_date = date_2020[index_trump_Nor_2020],
  percentage_diff = (counts_biden_Nor_2020-counts_trump_Nor_2020)/(counts_biden_Nor_2020+counts_trump_N
)

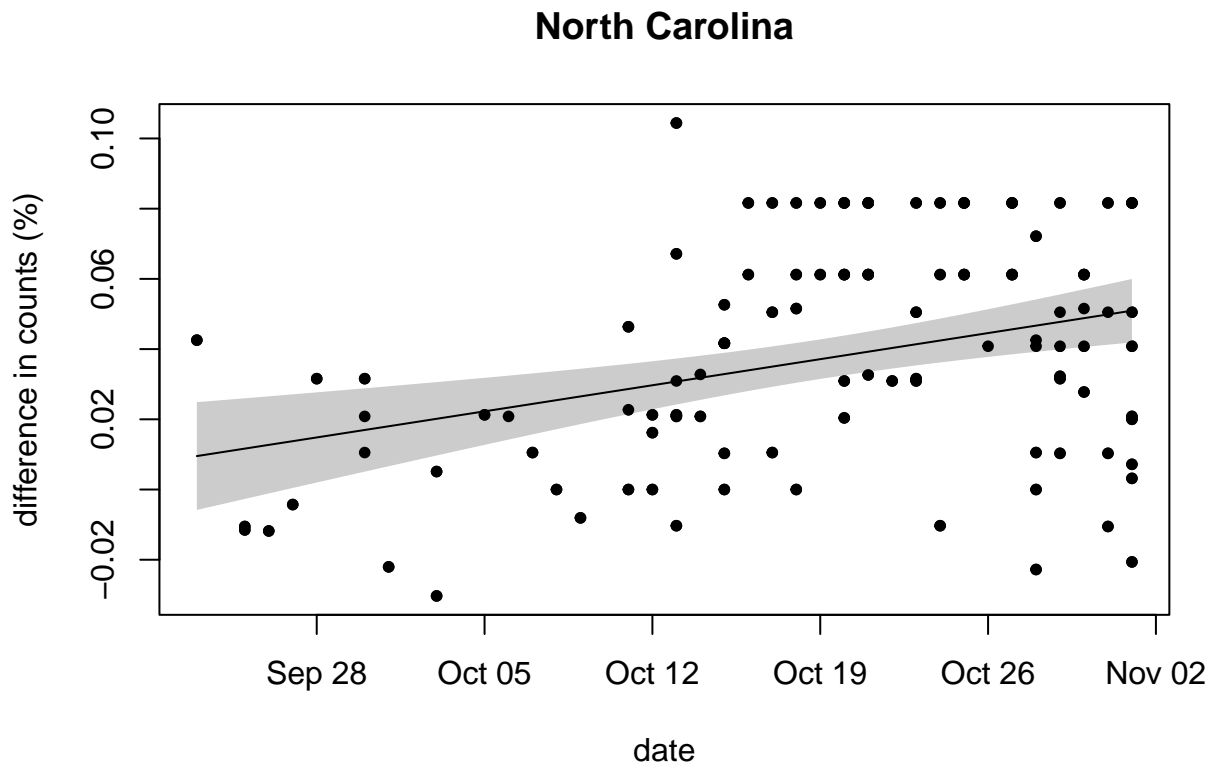
lm_model_Nor_2020=lm(percentage_diff~(data_date),data=counts_Nor_for_lm_2020)
summary(lm_model_Nor_2020)
```

```
##
## Call:
## lm(formula = percentage_diff ~ (data_date), data = counts_Nor_for_lm_2020)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.071556 -0.021322 -0.000993  0.022232  0.073634
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.966e+01  5.124e+00  -3.838 0.000211 ***
## data_date    1.062e-03  2.761e-04   3.845 0.000206 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.02954 on 106 degrees of freedom
## Multiple R-squared:  0.1224, Adjusted R-squared:  0.1141
## F-statistic: 14.78 on 1 and 106 DF,  p-value: 0.0002061

conf_interval_Nor_fitted_2020= predict(lm_model_Nor_2020, newdata=counts_Nor_for_lm_2020, interval="conf",
                                     level = 0.95)

plot(counts_Nor_for_lm_2020$data_date, counts_Nor_for_lm_2020$percentage_diff,
     col='black', pch=20, type='p', xlab='date', ylab='difference in counts (%)', main='North Carolina')
polygon(c(rev(counts_Nor_for_lm_2020$data_date), counts_Nor_for_lm_2020$data_date),
       c(rev(conf_interval_Nor_fitted_2020[,2]), conf_interval_Nor_fitted_2020[,3]), col = 'grey80', lty=1)
lines(counts_Nor_for_lm_2020$data_date, lm_model_Nor_2020$fitted.values,
     col='black', pch=20, type='l', xlab='date', ylab='difference in counts (%)', main='North Carolina')
lines(counts_Nor_for_lm_2020$data_date, counts_Nor_for_lm_2020$percentage_diff,
     col='black', pch=20, type='p', xlab='date', ylab='difference in counts (%)', main='North Carolina')
```



*#The plot is significantly increasing, showing that the winning rate for Biden is  
#increasing significantly.*

```
##Part e
#In Michigan the true difference was 50.62% Biden - 47.84% Trump = 2.78%.
#In Georgia the difference was 49.47% Biden - 49.24% Trump = 0.23%.
#In North Carolina the difference was 49.93% Biden - 48.59% Trump = -1.34%.
#The state with the closest margin was Georgia. The polls were correct
```



*#that Georgia would be the closest race. However they were incorrect on the #prediction that Biden would win.  
#As per the polls, the state of Michigan was expected to favor Biden  
#significantly, whereas the actual results show a rather close election.*

#### **##Potential problems:**

*#1. Sampling bias: Polls are based on a sample of voters rather than the entire  
#population of voters.*

*#2. Late decision-makers: Some voters might not have made up their minds when  
#they were polled but decided later on, and these late deciders might have  
#favored Trump over Biden. These changes would not have been captured in polls.*

#### **##Part f**

*#In Michigan, the polls indicated that Biden would win  
#According to the data we got, Biden won,*

*#In Georgia, the polls indicated that Biden would win  
#According to the data we got, Biden won,*

*#In North Carolina, the polls indicated Biden would win, but  
#as a result, Trump won.*

#### **##Potential problems:**

*#1. Differential turnout: The voters who actually turn up to vote on Election Day  
#may not be the same as those who responded to the polls. For example,  
#if Trump supporters were more motivated to vote than Biden supporters,  
#this could have led to Trump receiving a higher proportion of the votes than  
#expected.*

*#2. Shy Trump voters: Some research suggests that there may be a subset of Trump  
#voters who were reluctant to express their support for him in a polling  
#situation, but who did vote for him on Election Day. If this were the case,  
#it could explain why Trump's actual vote share was higher than the polls  
#predicted.*

# Untitled

Haocheng Zhang

2023-05-28

```
##Q3
#Load the data
polls_data_2016 = read.csv("president_general_polls_sorted_end_date_2016.csv")
polls_data_2020 = read.csv("president_polls_2020.csv")
library(lubridate)

##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##   date, intersect, setdiff, union

library(usmap)
library(ggplot2)

date_2016= mdy(polls_data_2016$createddate)
index_selected=which(date_2016>='2016-09-01' & date_2016<='2016-11-02')
polls_data_2016 = polls_data_2016[index_selected,]

date_2020= mdy(polls_data_2020$end_date)

index_selected=which(date_2020>='2020-09-01' & date_2020<='2020-11-02')
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'),]

delete_index=which( (polls_data_2020$pollster_id==
                    1610|polls_data_2020$pollster_id==1193)&
                    mdy(polls_data_2020$end_date)!= "2020-11-01")
polls_data_2020=polls_data_2020[-delete_index,]

##now let's delete those poll that only contains one candidate
polls_data_2020_question_id_num=unique(polls_data_2020$question_id)
##clean the data
for(i in 1:length(unique(polls_data_2020$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,]
    }
  }

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

polls_data_2020=polls_data_2020[-index_NA,]

polls_data_2016_enddate=mdy(polls_data_2016$enddate)

poll_state_sum_clinton_2016=aggregate(polls_data_2016$total.clinton,
                                     by=list(State=polls_data_2016$state),
                                     FUN=sum)
poll_state_sum_trump_2016=aggregate(polls_data_2016$total.trump,
                                    by=list(State=polls_data_2016$state),
                                    FUN=sum)

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_state_sum_clinton_2016[,2]+poll_state_sum_trump_2016[,2])
delete_index=which(levels(poll_state_diff_percentage[,1])=='U.S.')
if(length(delete_index)>0){
  poll_state_diff_percentage=poll_state_diff_percentage[-delete_index,]
  poll_state_diff_percentage[,1]
}

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

index_biden_2020=which(polls_data_2020$answer=='Biden')
index_trump_2020=which(polls_data_2020$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]

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),
                                    FUN=sum)
poll_state_sum_trump_2020=aggregate(polls_data_2020$total.trump,
                                    by=list(State=polls_data_2020$state),
                                    FUN=sum)

```

```

poll_state_sum_biden_2020=poll_state_sum_biden_2020[-1,]
poll_state_sum_trump_2020=poll_state_sum_trump_2020[-1,]

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))

##difference between 2020 and 2016
##delete nebraska CD-1 and CD-3, US, 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"             "U.S."                "Utah"
## [52] "Vermont"           "Virginia"            "Washington"
## [55] "West Virginia"     "Wisconsin"           "Wyoming"

```

```
state_poll_2020$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-2"     "Nevada"              "New Hampshire"

```

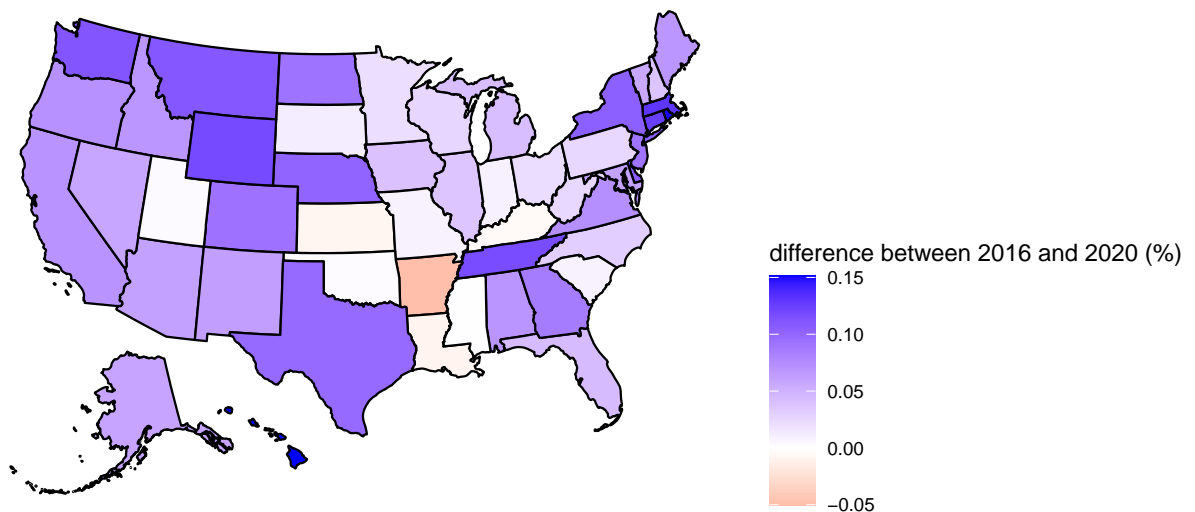
```
## [34] "New Jersey"      "New Mexico"      "New York"
## [37] "North Carolina"  "North Dakota"    "Ohio"
## [40] "Oklahoma"        "Oregon"          "Pennsylvania"
## [43] "Rhode Island"    "South Carolina"  "South Dakota"
## [46] "Tennessee"      "Texas"           "Utah"
## [49] "Vermont"         "Virginia"        "Washington"
## [52] "West Virginia"   "Wisconsin"       "Wyoming"

state_poll_2016=state_poll_2016[-c(31,33,50),]

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_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



*##The visual representation in these graphs suggests an increased polarization  
#between the two major political parties from 2016 to 2020, with an apparent  
#overall shift towards the Democratic party across the map. This shift could  
#potentially be attributed to dissatisfaction with the Republican response  
#to the pandemic. Additionally, it appears that discontentment with  
#President Trump's administration may have contributed to this trend.*

```
# Part b
sort_diff <- state_poll_2020_2016_diff[order(state_poll_2020_2016_diff$diff),]
sort_diff
```

##	state	diff
## 4	Arkansas	-0.0505050342
## 19	Louisiana	-0.0090943552
## 17	Kansas	-0.0072425960
## 18	Kentucky	-0.0038330859
## 27	Mississippi	0.0009961673
## 40	Oklahoma	0.0027909051
## 48	Utah	0.0031269857
## 28	Missouri	0.0082034611
## 44	South Carolina	0.0085093402
## 15	Indiana	0.0097141655
## 45	South Dakota	0.0124156189
## 26	Minnesota	0.0215032573
## 39	Ohio	0.0219281680
## 52	West Virginia	0.0255757019
## 42	Pennsylvania	0.0259456912
## 53	Wisconsin	0.0267207138
## 37	North Carolina	0.0310821930
## 14	Illinois	0.0373040776
## 33	New Hampshire	0.0376230501
## 16	Iowa	0.0392003319
## 25	Michigan	0.0419562005
## 10	Florida	0.0441124664
## 49	Vermont	0.0561286088
## 32	Nevada	0.0579893834
## 2	Alaska	0.0593615071
## 23	Maryland	0.0607202788
## 3	Arizona	0.0614151801
## 35	New Mexico	0.0623467550
## 22	Maine CD-2	0.0644133478
## 20	Maine	0.0685616033
## 13	Idaho	0.0689032507
## 1	Alabama	0.0692322291
## 5	California	0.0707219888
## 41	Oregon	0.0720323199
## 50	Virginia	0.0731736945
## 11	Georgia	0.0858615596
## 34	New Jersey	0.0911912704
## 38	North Dakota	0.0920670720
## 6	Colorado	0.0923346812
## 21	Maine CD-1	0.0936457976
## 47	Texas	0.0978514945

```
## 30          Nebraska 0.1018085818
## 36          New York 0.1046349413
## 8           Delaware 0.1065771992
## 29          Montana 0.1098752389
## 51          Washington 0.1114424912
## 46          Tennessee 0.1168463358
## 54          Wyoming 0.1185112001
## 7           Connecticut 0.1256462051
## 24          Massachusetts 0.1302587034
## 31          Nebraska CD-2 0.1455258422
## 9 District of Columbia 0.1467290357
## 43          Rhode Island 0.1499907313
## 12          Hawaii 0.1515343018
```

```
##Based on the results, the ten battleground states in 2020 --
```

```
#defined as those with the smallest percentage differences between the two major #candidates -- would b
```

```
#Arkansas
#Louisiana
#Kansas
#Kentucky
#Mississippi
#Oklahoma
#Utah
#Missouri
#South Carolina
#Indiana
```

```
#Part(C)
```

```
print(state_poll_2020_2016_diff)
```

```
##          state      diff
## 1      Alabama 0.0692322291
## 2      Alaska 0.0593615071
## 3      Arizona 0.0614151801
## 4      Arkansas -0.0505050342
## 5      California 0.0707219888
## 6      Colorado 0.0923346812
## 7      Connecticut 0.1256462051
## 8      Delaware 0.1065771992
## 9 District of Columbia 0.1467290357
## 10     Florida 0.0441124664
## 11     Georgia 0.0858615596
## 12     Hawaii 0.1515343018
## 13     Idaho 0.0689032507
## 14     Illinois 0.0373040776
## 15     Indiana 0.0097141655
## 16     Iowa 0.0392003319
## 17     Kansas -0.0072425960
## 18     Kentucky -0.0038330859
## 19     Louisiana -0.0090943552
## 20     Maine 0.0685616033
## 21     Maine CD-1 0.0936457976
```

## 22	Maine CD-2	0.0644133478
## 23	Maryland	0.0607202788
## 24	Massachusetts	0.1302587034
## 25	Michigan	0.0419562005
## 26	Minnesota	0.0215032573
## 27	Mississippi	0.0009961673
## 28	Missouri	0.0082034611
## 29	Montana	0.1098752389
## 30	Nebraska	0.1018085818
## 31	Nebraska CD-2	0.1455258422
## 32	Nevada	0.0579893834
## 33	New Hampshire	0.0376230501
## 34	New Jersey	0.0911912704
## 35	New Mexico	0.0623467550
## 36	New York	0.1046349413
## 37	North Carolina	0.0310821930
## 38	North Dakota	0.0920670720
## 39	Ohio	0.0219281680
## 40	Oklahoma	0.0027909051
## 41	Oregon	0.0720323199
## 42	Pennsylvania	0.0259456912
## 43	Rhode Island	0.1499907313
## 44	South Carolina	0.0085093402
## 45	South Dakota	0.0124156189
## 46	Tennessee	0.1168463358
## 47	Texas	0.0978514945
## 48	Utah	0.0031269857
## 49	Vermont	0.0561286088
## 50	Virginia	0.0731736945
## 51	Washington	0.1114424912
## 52	West Virginia	0.0255757019
## 53	Wisconsin	0.0267207138
## 54	Wyoming	0.1185112001

*##From the result in (C), in most of these states, polls may have underestimated  
#the percentage of real votes received by one candidate in both 2016 and 2020,  
#as most differences are positive.*

*#For example, in Alabama, the difference between the polling prediction and the  
#actual result was approximately 6.92% in 2016 and likely a similar value in  
#2020. This suggests that the candidate was predicted to get less support than  
#they actually received. However, in Arkansas, the polls overestimated the  
#support for the candidate by around 5.05%, as indicated by the negative  
#difference.*

*#There can be several reasons for this bias in the polls:  
#Nonresponse Bias: Some people are less likely to respond to polls,  
#and these people might be more likely to vote for one candidate.  
#If these individuals aren't adequately represented in the sample,  
#the poll can be biased towards the other candidate.*

*#Shy Voter Theory: Some voters might be hesitant to reveal their true voting  
#intention in a poll, especially if they feel their choice isn't the socially  
#desirable one. This can result in underestimation of support for a candidate.*



# Untitled

Haocheng Zhang

2023-05-28

## *#Problem 4*

### *#Load the data*

```
polls_data_2016 = read.csv("president_general_polls_sorted_end_date_2016.csv")
polls_data_2020 = read.csv("president_polls_2020.csv")
library(lubridate)
```

```
##
```

```
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
```

```
##
```

```
##      date, intersect, setdiff, union
```

```
library(usmap)
library(ggplot2)
```

```
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'),]
```

```
delete_index=which( (polls_data_2020$pollster_id==
                    1610|polls_data_2020$pollster_id==1193)&
                    mdy(polls_data_2020$end_date)!= "2020-11-01")
polls_data_2020=polls_data_2020[-delete_index,]
```

### *##now let's delete those poll that only contains one candidate*

```
polls_data_2020_question_id_num=unique(polls_data_2020$question_id)
```

### *##clean the data*

```
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,]
  }
}
```

```
index_NA=which(is.na(polls_data_2020$sample_size)==T)
```

```

polls_data_2020=polls_data_2020[-index_NA,]

polls_data_2016_enddate=mdy(polls_data_2016$enddate)

poll_state_sum_clinton_2016=aggregate(polls_data_2016$total.clinton, by=list(State=polls_data_2016$state),FUN=
poll_state_sum_trump_2016=aggregate(polls_data_2016$total.trump, by=list(State=polls_data_2016$state),FUN=

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_state_sum_clinton_2016[,2]+
                                                                poll_state_sum_trump_2016[,2])

delete_index=which(levels(poll_state_diff_percentage[,1])=='U.S.')
if(length(delete_index)>0){
  poll_state_diff_percentage=poll_state_diff_percentage[-delete_index,]
  poll_state_diff_percentage[,1]
}

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

index_biden_2020=which(polls_data_2020$answer=='Biden')
index_trump_2020=which(polls_data_2020$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]

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),FUN=
poll_state_sum_trump_2020=aggregate(polls_data_2020$total.trump, by=list(State=polls_data_2020$state),FUN=

poll_state_sum_biden_2020=poll_state_sum_biden_2020[-1,]
poll_state_sum_trump_2020=poll_state_sum_trump_2020[-1,]

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))

plot_usmap(data = state_poll_2016, values = "diff_percentage", color = "black") +
  scale_fill_gradient2(name = "difference (%)", low= "red",

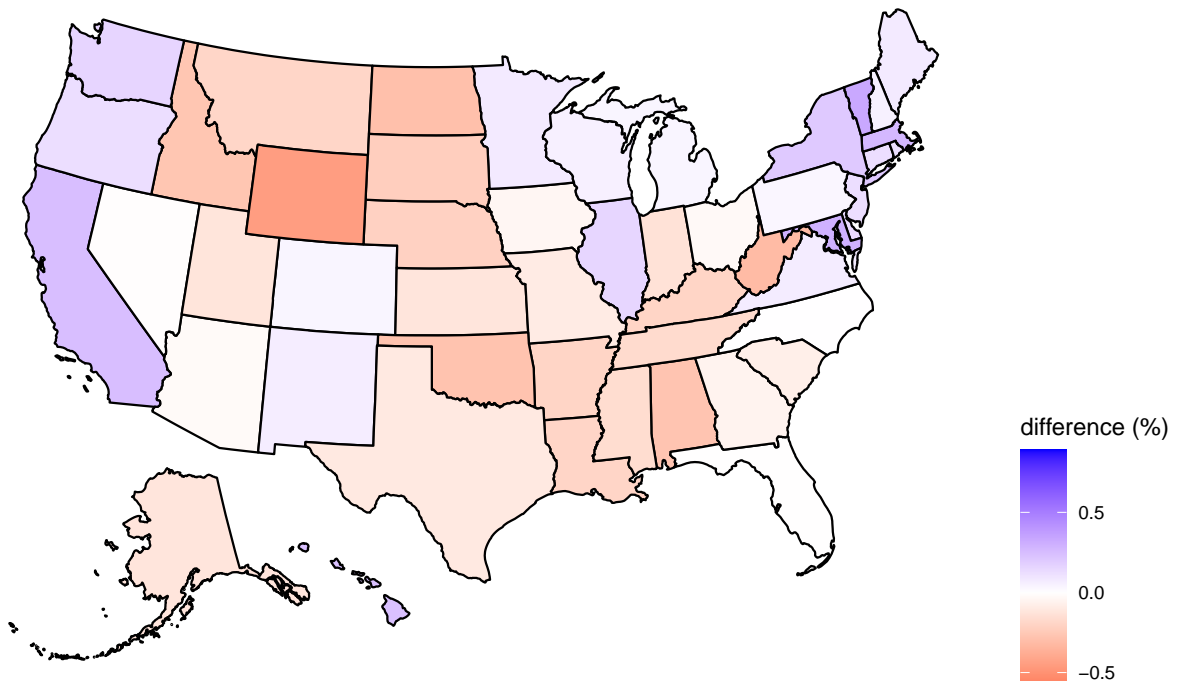
```

```

        mid = "white",
        high = "blue",
        midpoint = 0,limits=limit_val)+
theme(legend.position = "right")+
ggtitle("2016")

```

2016

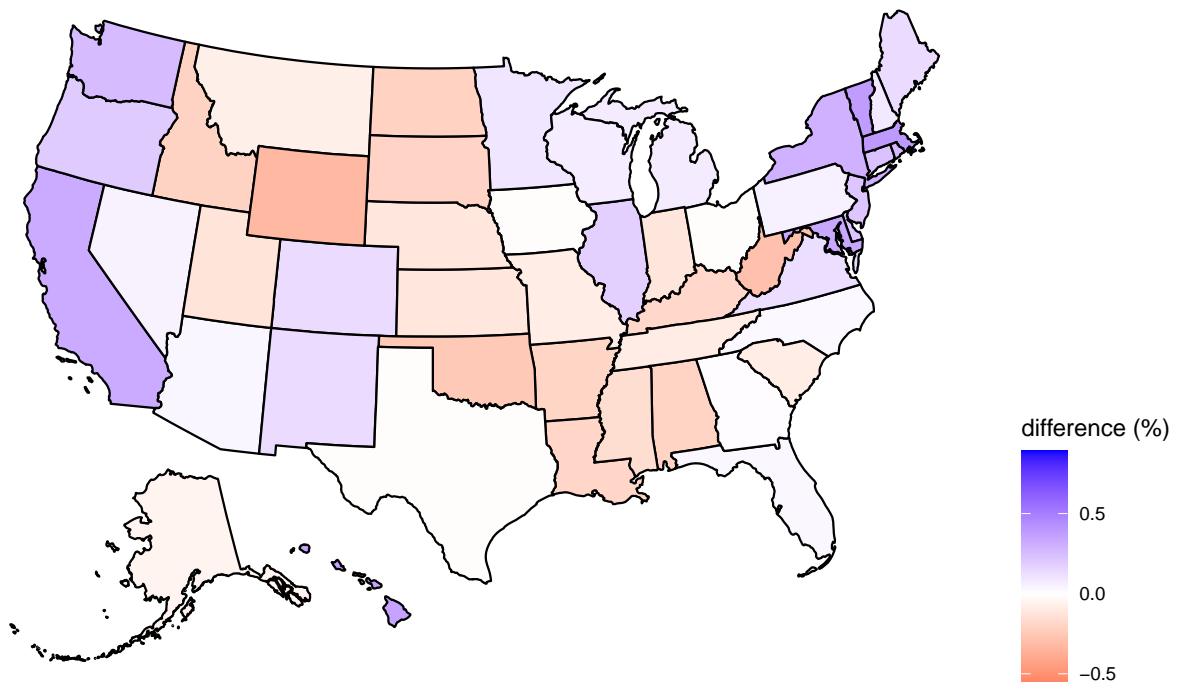


```

##2020
##it does not have poll from Nebraska. It does not plot the poll from
#congressional district
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")

```

2020



```
# Part b
state_poll_2016=state_poll_2016[-c(33,50),]

dif <- data.frame(
  state <- state_poll_2016$state,
  diff_2016 <- state_poll_2016$diff_percentage,
  diff_2020 <- state_poll_2020$diff_percentage
)
index_change = which(dif$diff_2016*dif$diff_2020<0)
dif$state[index_change]
```

```
## [1] "Arizona"      "Florida"      "Georgia"      "Maine CD-2"
## [5] "Nebraska CD-2" "Nevada"
```

```
##part(C)
```

*#No, there are not. The states I've listed that did not flip in reality --  
#Florida and Nevada -- could have been identified as potential flips due to  
#close polling results, demographic changes, or other state-specific factors. #However, many elements c  
#predict, such as voter turnout, campaign strategies, and major events.*

# Untitled

Haocheng Zhang

2023-05-28

```
##Q5
```

```
#Load the data
```

```
polls_data_2016 = read.csv("president_general_polls_sorted_end_date_2016.csv")
```

```
polls_data_2020 = read.csv("president_polls_2020.csv")
```

```
library(ggplot2)
```

```
#2016
```

```
index_2016_Flo=which(polls_data_2016$state=="Florida")
```

```
index_2016_Iow=which(polls_data_2016$state=="Iowa")
```

```
Fol_2016_real_clinton = 47.82;
```

```
Iow_2016_real_clinton = 41.74;
```

```
polls_data_2016[index_2016_Flo,]$rawpoll_clinton = polls_data_2016[index_2016_Flo,]$rawpoll_clinton - F
```

```
polls_data_2016[index_2016_Iow,]$rawpoll_clinton = polls_data_2016[index_2016_Iow,]$rawpoll_clinton - I
```

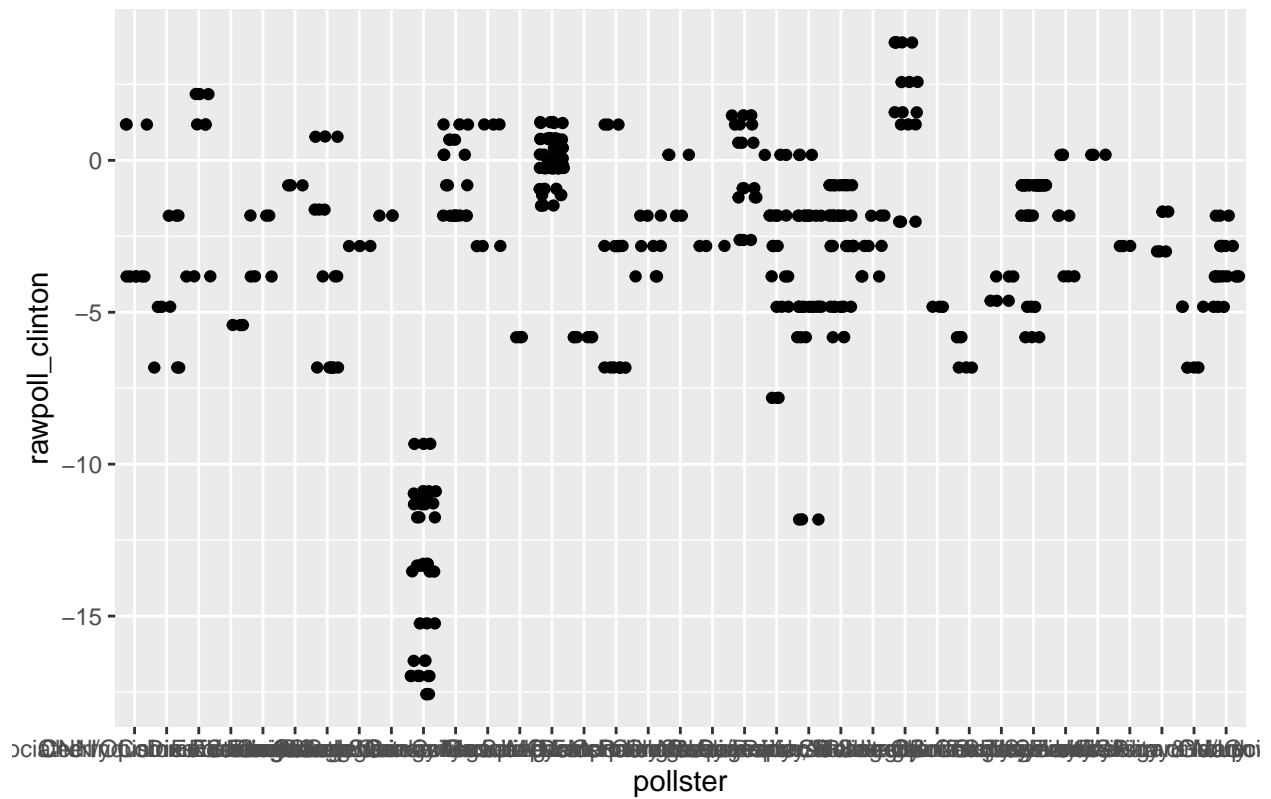
```
ggplot(data = polls_data_2016[index_2016_Flo,]) +
```

```
  aes(y = rawpoll_clinton, x = pollster) +
```

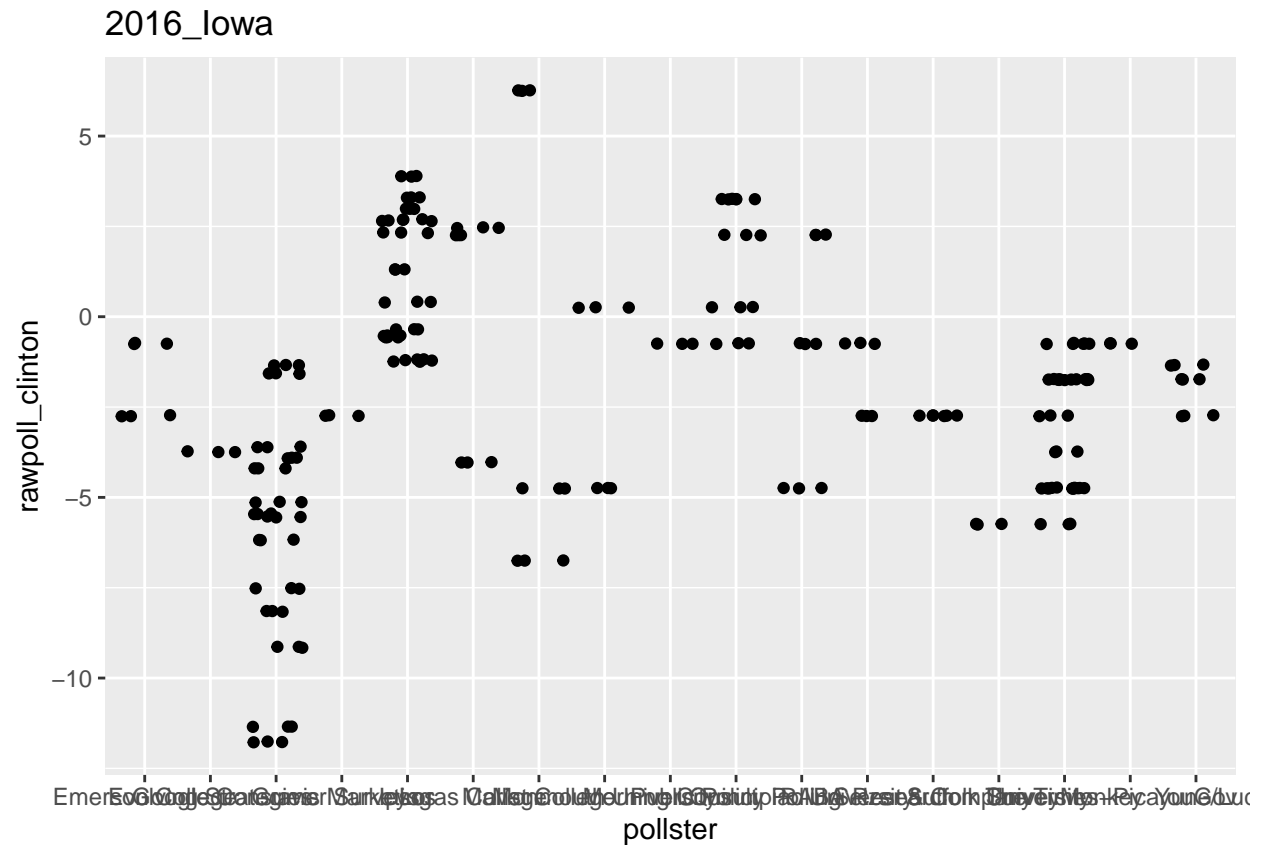
```
  geom_jitter()+
```

```
  ggtitle("2016_Florida")
```

2016\_Florida



```
ggplot(data = polls_data_2016[index_2016_Iowa,]) +
  aes(y = rawpoll_clinton, x = pollster) +
  geom_jitter()+
  ggtitle("2016_Iowa")
```



```
#2020
```

```
Flo_2020_real_biden = 47.86;
```

```
Iow_2020_real_biden = 44.89;
```

```
index_biden_Flo_2020=which(polls_data_2020$answer=='Biden' & polls_data_2020$state=="Florida")
```

```
index_biden_Iow_2020=which(polls_data_2020$answer=='Biden' & polls_data_2020$state=="Iowa")
```

```
polls_data_2020[index_biden_Flo_2020,]$pct = polls_data_2020[index_biden_Flo_2020,]$pct - Flo_2020_real_biden
```

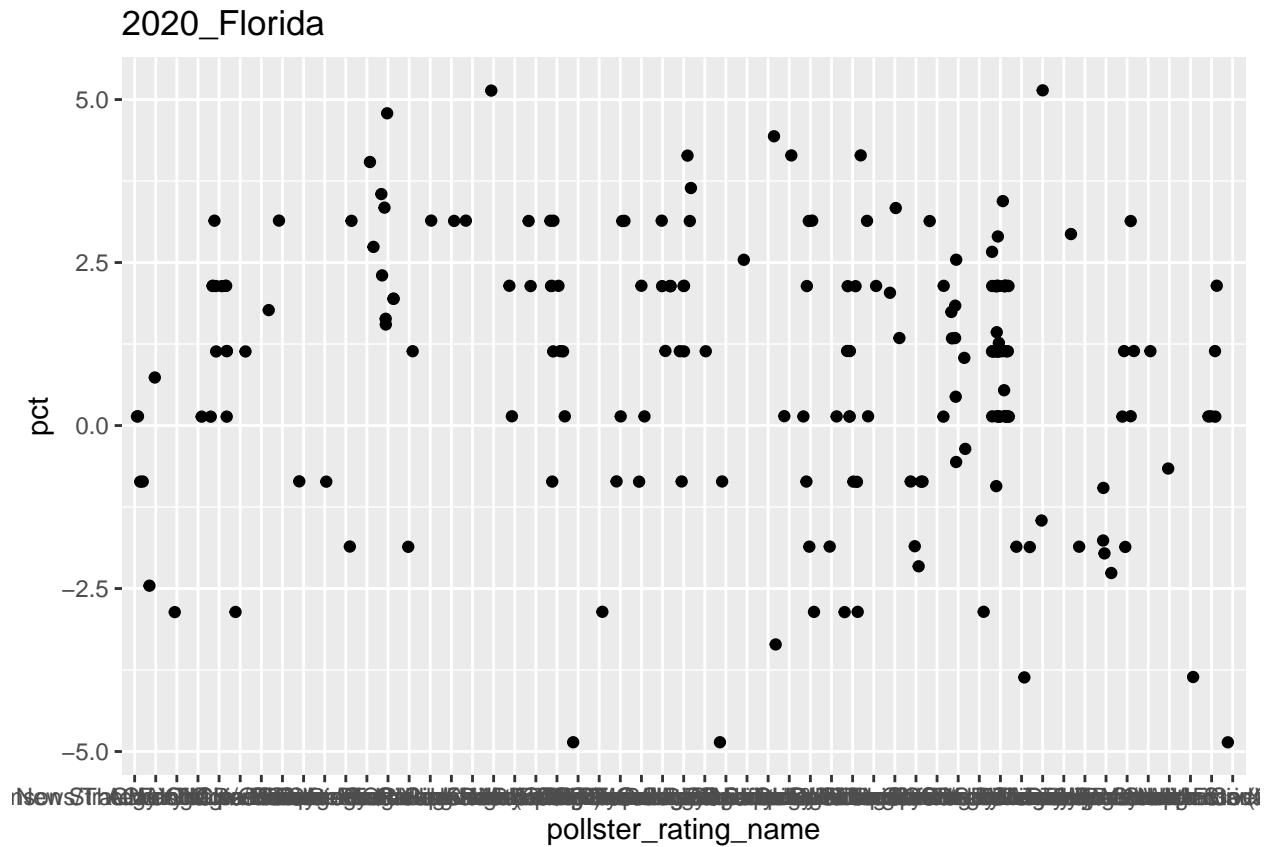
```
polls_data_2020[index_biden_Iow_2020,]$pct = polls_data_2020[index_biden_Iow_2020,]$pct - Iow_2020_real_biden
```

```
ggplot(data = polls_data_2020[index_biden_Flo_2020,]) +
```

```
  aes(y = pct, x = pollster_rating_name) +
```

```
  geom_jitter()+
```

```
  ggtitle("2020_Florida")
```



```
ggplot(data = polls_data_2020[index_biden_Iow_2020,]) +
  aes(y = pct, x = pollster_rating_name) +
  geom_jitter()+
  ggtitle("2020_Iowa")
```





**##part(d)**

*#Better Sampling: Ensuring that the sample is representative of the voting  
#population is crucial. This could mean taking into account factors like age,  
#race, education level, and geographic location.*

*#Weighting: If the sample isn't perfectly representative, statisticians can  
#assign weights to responses from underrepresented groups to ensure they have  
#the correct influence on the final results.*

*#Tracking Late Changes: Conducting polls close to the election day can help  
#capture late changes in public opinion.*