data_report_1

Question 1

Question 1.a

Read polls data in 2016.

```
current_address=getwd()
polls_data_2016=read.csv(paste0(current_address,"/data/president_general_polls_sorted_end_date_2016.csv
```

Extract the data of Minnesota, Florida and North Carolina.

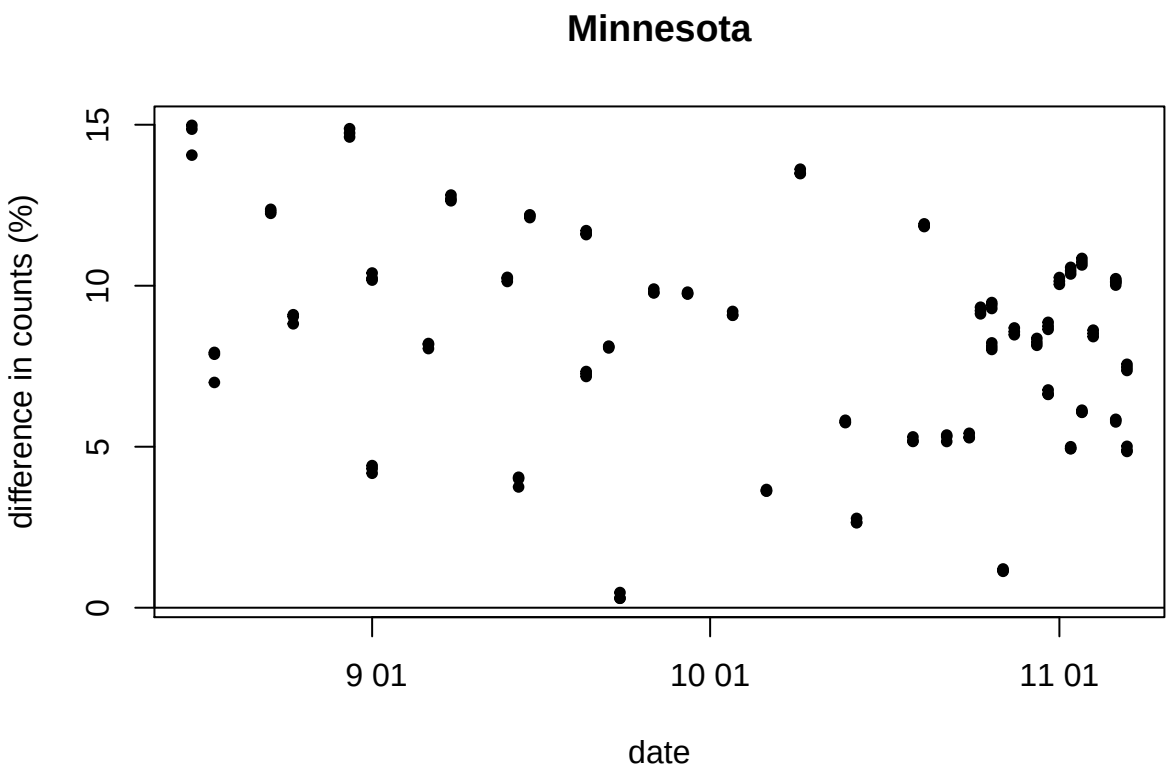
```
library(lubridate)
```

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

```
## The following object is masked from 'package:cowplot':
##
##     stamp
```

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

```
index_mn=which(polls_data_2016$state=="Minnesota")
date_mn <- mdy(polls_data_2016$enddate[index_mn])
index_mn = index_mn[date_mn > "2016-08-01"]
date_mn = date_mn[date_mn > "2016-08-01"]
percentage_diff_mn=(polls_data_2016$total.clinton[index_mn]-
                      polls_data_2016$total.trump[index_mn])/
  (polls_data_2016$total.clinton[index_mn]+polls_data_2016$total.trump[index_mn])
plot(date_mn,percentage_diff_mn * 100,
     col='black',pch=20,type='p',xlab='date',ylab='difference in counts (%)',main='Minnesota')
abline(a=0,b=0)
```



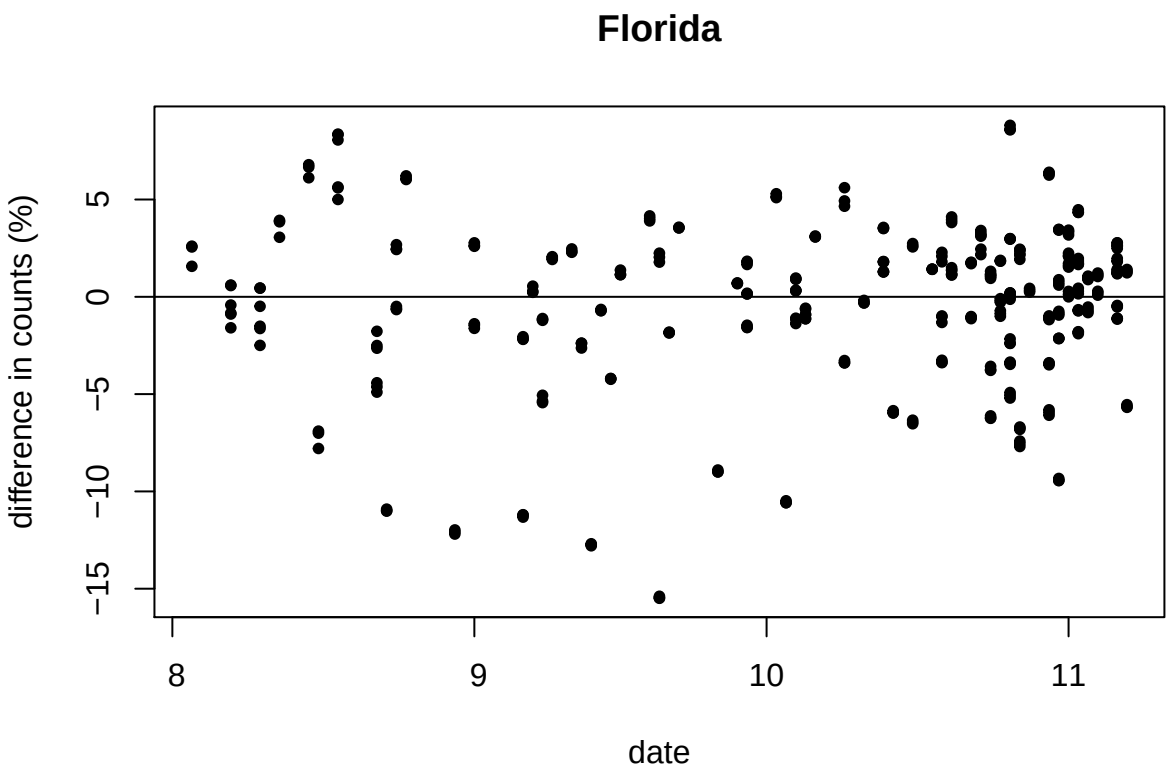
Clinton was ahead of Trump in Minnesota according to the polls.

```
print((sum(polls_data_2016$total.clinton[index_mn]) -
         sum(polls_data_2016$total.trump[index_mn])) /
        (sum(polls_data_2016$total.clinton[index_mn]) +
           sum(polls_data_2016$total.trump[index_mn])) * 100)
```

```
## [1] 7.87602
```

The mean lead for Clinton was 7.88%.

```
index_fl=which(polls_data_2016$state=="Florida")
date_fl <- mdy(polls_data_2016$enddate[index_fl])
index_fl = index_fl[date_fl > "2016-08-01"]
date_fl = date_fl[date_fl > "2016-08-01"]
percentage_diff_fl=(polls_data_2016$total.clinton[index_fl]-
                      polls_data_2016$total.trump[index_fl])/
  (polls_data_2016$total.clinton[index_fl]+
     polls_data_2016$total.trump[index_fl])
plot(date_fl,percentage_diff_fl * 100,
     col='black',pch=20,type='p',xlab='date',ylab='difference in counts (%)',main='Florida')
abline(a=0,b=0)
```



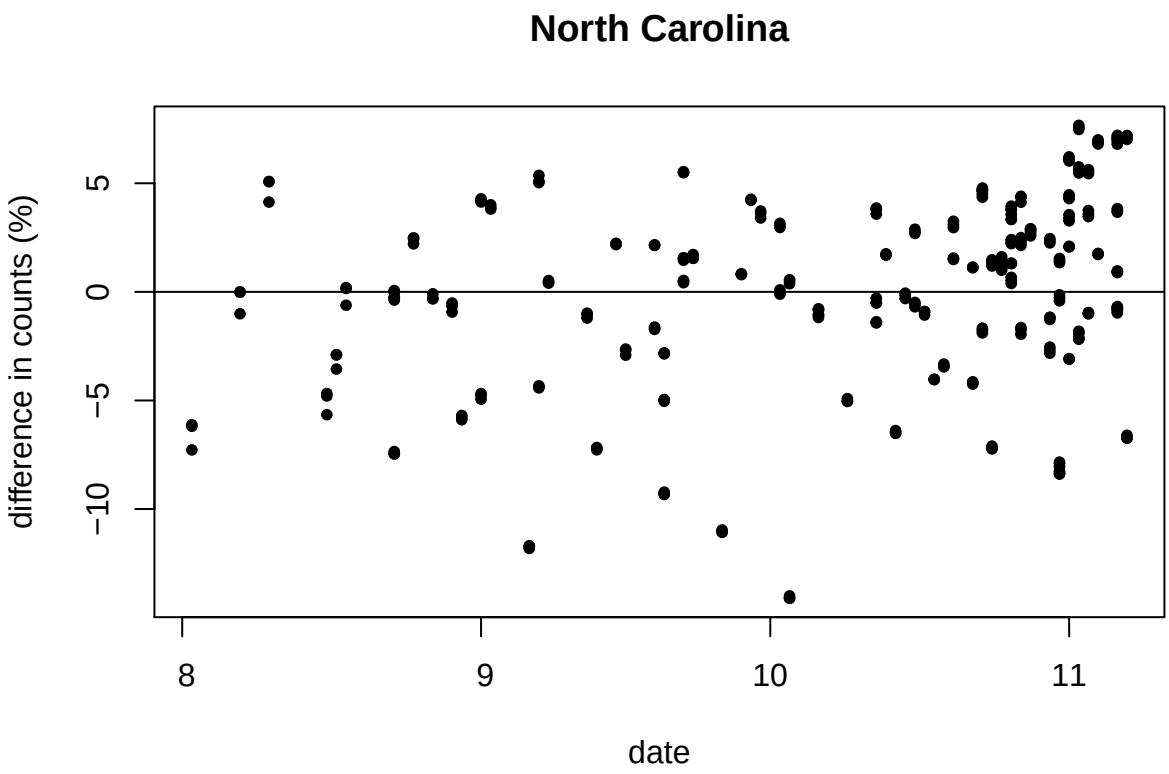
Nobody was significantly ahead in Florida according to the poll data.

```
print((sum(polls_data_2016$total.clinton[index_fl]) -
         sum(polls_data_2016$total.trump[index_fl])) /
        (sum(polls_data_2016$total.clinton[index_fl]) +
           sum(polls_data_2016$total.trump[index_fl])) * 100)
```

```
## [1] -0.5966367
```

The mean difference was -0.60%. So Trump was slightly ahead.

```
index_nc=which(polls_data_2016$state=="North Carolina")
date_nc <- mdy(polls_data_2016$enddate[index_nc])
index_nc = index_nc[date_nc > "2016-08-01"]
date_nc = date_nc[date_nc > "2016-08-01"]
percentage_diff_nc=(polls_data_2016$total.clinton[index_nc]-
                      polls_data_2016$total.trump[index_nc])/
  (polls_data_2016$total.clinton[index_nc]+
     polls_data_2016$total.trump[index_nc])
plot(date_nc,percentage_diff_nc * 100,
     col='black',pch=20,type='p',xlab='date',ylab='difference in counts (%)',main='North Carolina')
abline(a=0,b=0)
```



Simiarly, no one was significantly lead in North Carolina.

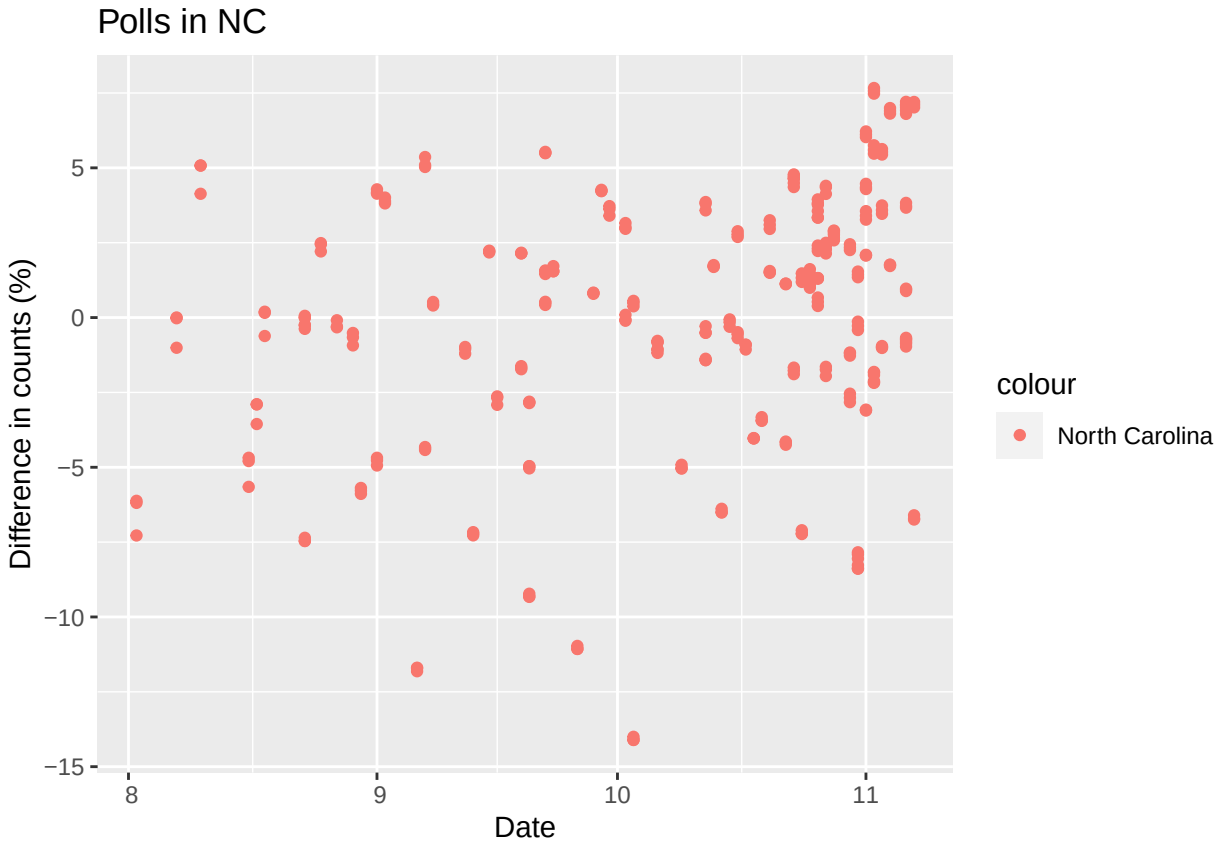
```
print((sum(polls_data_2016$total.clinton[index_nc]) -
         sum(polls_data_2016$total.trump[index_nc])) /
        (sum(polls_data_2016$total.clinton[index_nc]) +
           sum(polls_data_2016$total.trump[index_nc])) * 100)
```

```
## [1] 0.6609668
```

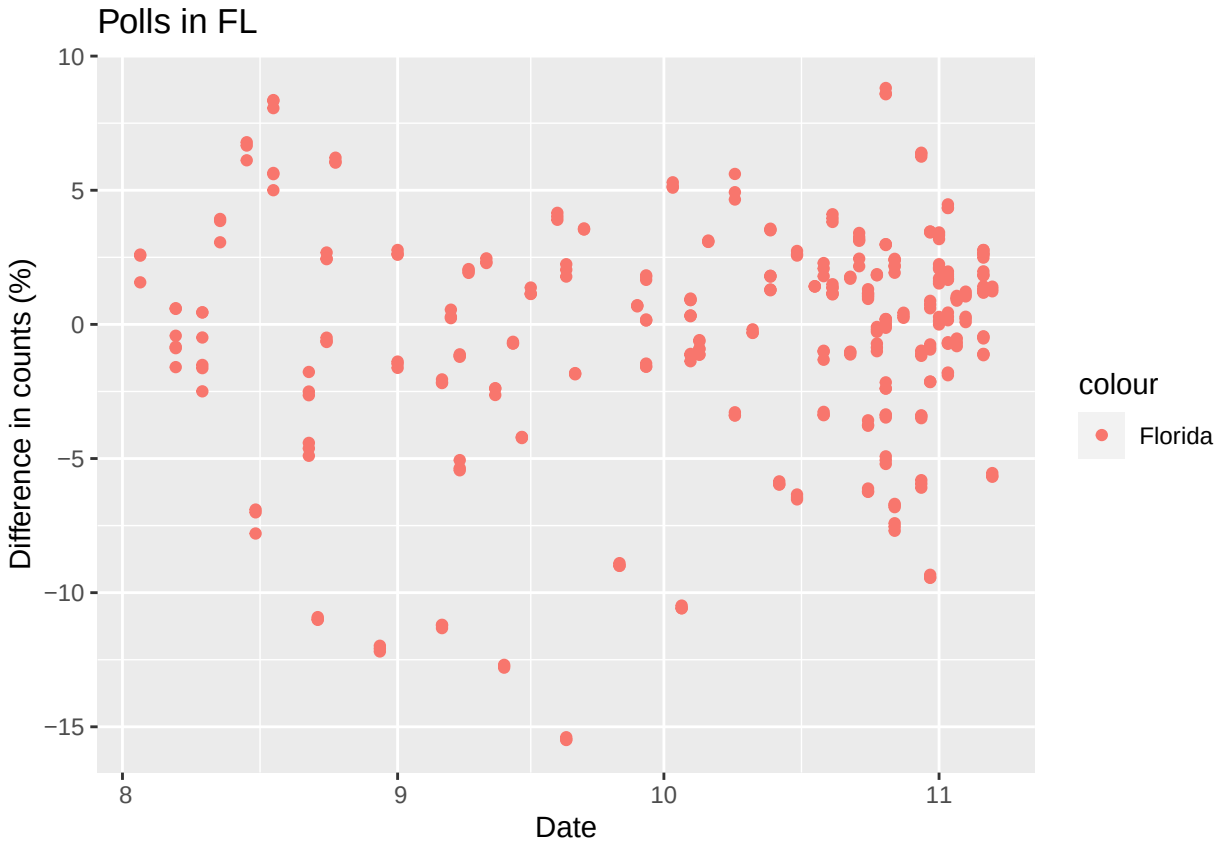
The mean difference was 0.66%. So Trump was slightly ahead too.

Question 1.b

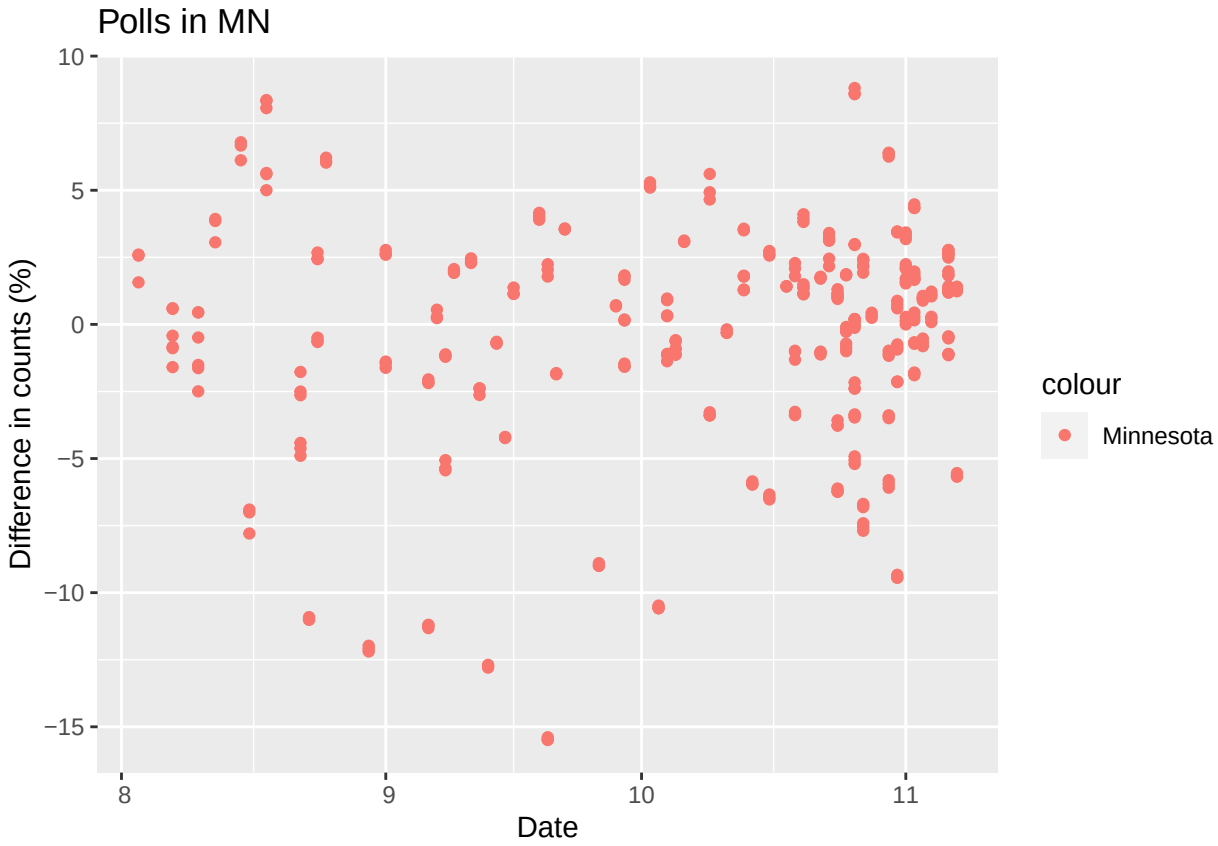
```
ggplot() + geom_point(aes(x = date_nc, y=percentage_diff_nc * 100, color="North Carolina")) +
  ggtitle("Polls in NC") + xlab("Date") + ylab("Difference in counts (%)")
```




```
ggplot() + geom_point(aes(x = date_fl, y=percentage_diff_fl * 100, color="Florida")) +
  ggtitle("Polls in FL") + xlab("Date") + ylab("Difference in counts (%)")
```




```
ggplot() + geom_point(aes(x = date_fl, y=percentage_diff_fl * 100, color="Minnesota")) +
  ggtitle("Polls in MN") + xlab("Date") + ylab("Difference in counts (%)")
```

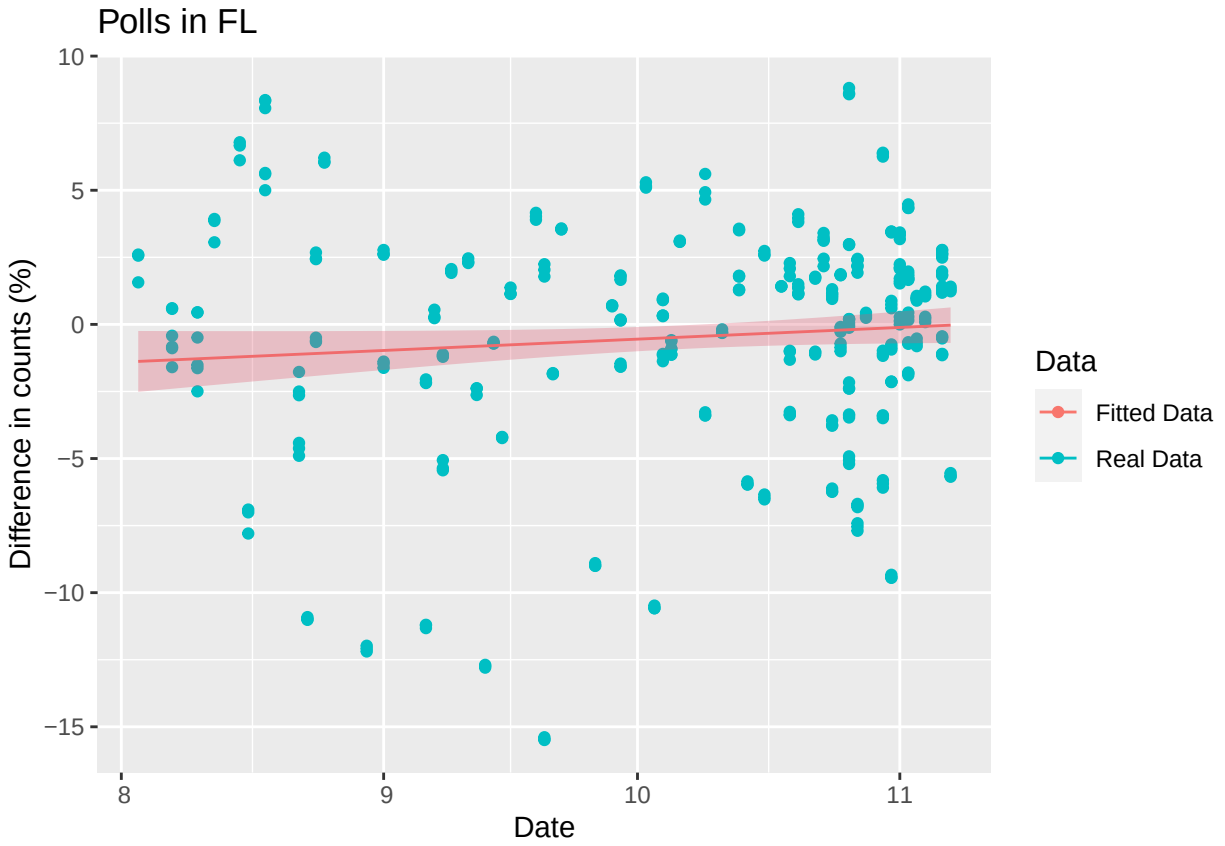


As we can see, the advantage of both candidates is not significant in the polls.I felt that there is no significant trend in the polls of Florida and Minnesota, and a slight increasing trend for Clinton in North Carolina.

Question 1.c

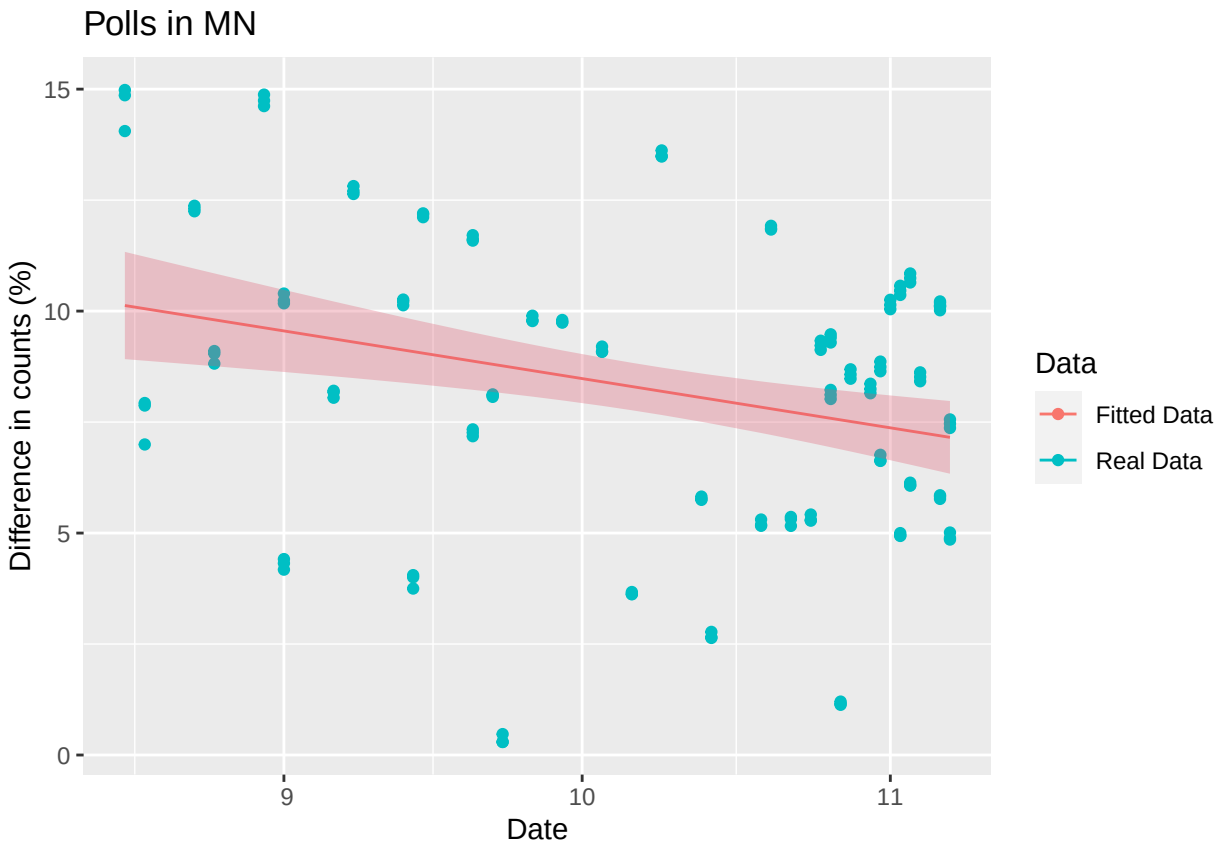
For Florida:

```
fit.fl <- lm(percentage_diff_fl~date_fl)
fitted.fl <- fitted(fit.fl)
conf <- predict(fit.fl, newdata = date_fl, interval = 'confidence')
ggplot() + geom_point(aes(x = date_fl, y=percentage_diff_fl * 100, color="Real Data")) +
  ggtitle("Polls in FL") + xlab("Date") + ylab("Difference in counts (%)") +
  geom_line(aes(x=date_fl, y=fitted.fl * 100, color="Fitted Data")) +
  geom_ribbon(aes(x=date_fl, ymin=conf[,2] * 100, ymax=conf[,3] * 100), alpha=0.3, fill=2) +
  guides(colour = guide_legend(title = "Data"))
```



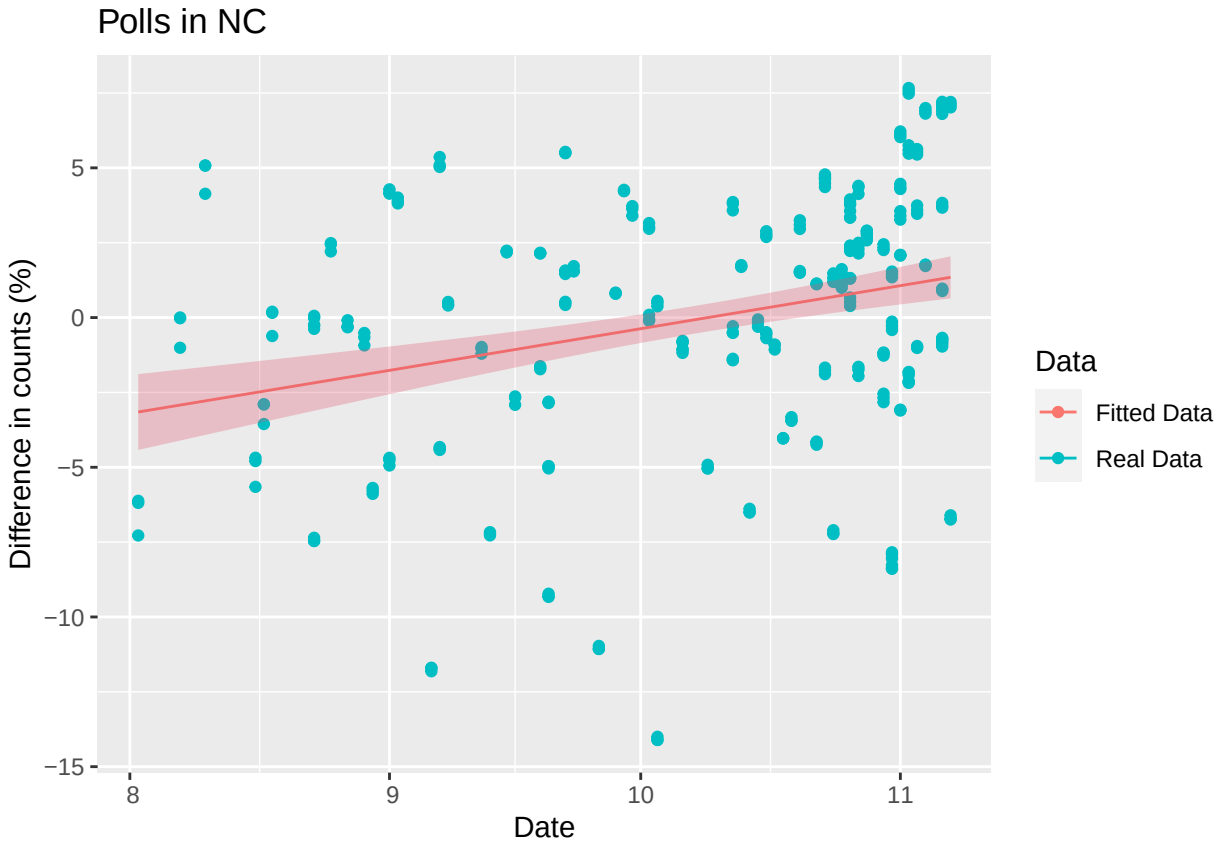
For Minnesota:

```
fit.mn <- lm(percentage_diff_mn~date_mn)
fitted.mn <- fitted(fit.mn)
conf <- predict(fit.mn, newdata = date_mn, interval = 'confidence')
ggplot() + geom_point(aes(x = date_mn, y=percentage_diff_mn * 100, color="Real Data")) +
  ggtitle("Polls in MN") + xlab("Date") + ylab("Difference in counts (%)") +
  geom_line(aes(x=date_mn, y=fitted.mn * 100, color="Fitted Data")) +
  geom_ribbon(aes(x=date_mn, ymin=conf[,2] * 100, ymax=conf[,3] * 100), alpha=0.3, fill=2) +
  guides(colour = guide_legend(title = "Data"))
```



For North Carolina:

```
fit.nc <- lm(percentage_diff_nc~date_nc)
fitted.nc <- fitted(fit.nc)
conf <- predict(fit.nc, newdata = date_nc, interval = 'confidence')
ggplot() + geom_point(aes(x = date_nc, y=percentage_diff_nc * 100, color="Real Data")) +
  ggtitle("Polls in NC") + xlab("Date") + ylab("Difference in counts (%)") +
  geom_line(aes(x=date_nc, y=fitted.nc * 100, color="Fitted Data")) +
  geom_ribbon(aes(x=date_nc, ymin=conf[,2] * 100, ymax=conf[,3] * 100), alpha=0.3, fill=2) +
  guides(colour = guide_legend(title = "Data"))
```



So Florida had the smallest margin.

Question 1.d

2016 results: Florida: R+1.2 North Carolina: R+3.7 Minnesota: D+1.52

Polls make correct prediction on Minneosta. The results in Florida in within the margin of error, so it is not statistically incorrect that this would happen. However the prediction for North Carolina was wrong. They failed to cover some people, or some Trump supporter does not respond to the poll.

So the two reasons are: 1. statistically margin of error; 2. undercoverage of some supporters.

Question 2

Question 2.a

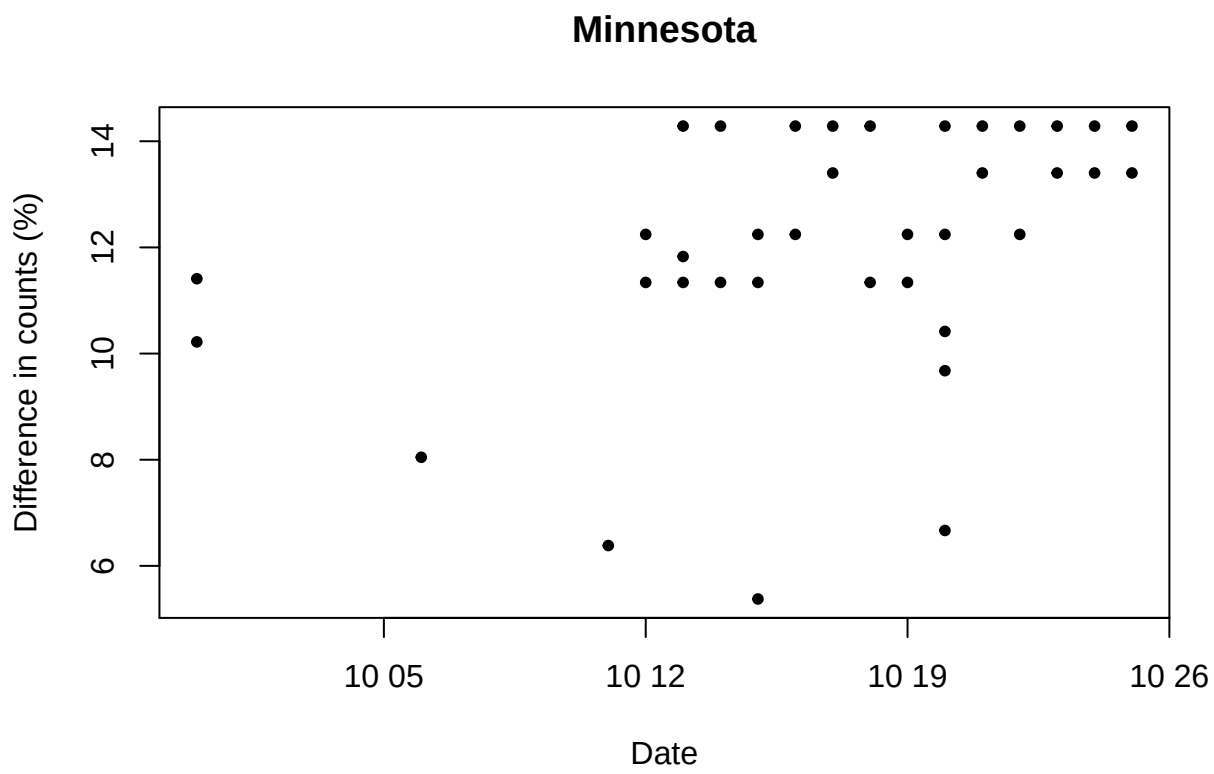
Read polls data in 2020 and then pre-process.

```
current_address=getwd()
polls_data_2020=read.csv(paste0(current_address,"/data/president_polls_2020.csv"))

date_2020= mdy(polls_data_2020$end_date)
date_2020_latest_day=date_2020[1]
index_selected=which(date_2020>='2020-09-26' & date_2020 <='2020-10-25')
polls_data_2020=polls_data_2020[index_selected,]
polls_data_2020=polls_data_2020[which(polls_data_2020$answer=='Biden'|polls_data_2020$answer=='Trump'),]
polls_data_2020_question_id_num=unique(polls_data_2020$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,]
  }
}

date_2020= mdy(polls_data_2020$end_date)
```

Extract the data of Minnesota, Florida and North Carolina.

```
index_mn_2020=which(polls_data_2020$state=="Minnesota")

index_biden_mn_2020=which(polls_data_2020$answer=='Biden' & polls_data_2020$state=="Minnesota")
index_trump_mn_2020=which(polls_data_2020$answer=='Trump' & polls_data_2020$state=="Minnesota")

counts_biden_mn_2020=polls_data_2020$pct[index_biden_mn_2020]*
  polls_data_2020$sample_size[index_biden_mn_2020]
counts_trump_mn_2020=polls_data_2020$pct[index_trump_mn_2020]*
  polls_data_2020$sample_size[index_trump_mn_2020]

##plot percentage
difference_mn = (counts_biden_mn_2020-counts_trump_mn_2020)/
  (counts_biden_mn_2020+counts_trump_mn_2020)
plot(date_2020[index_trump_mn_2020], difference_mn * 100,
     col='black',pch=20,type='p',xlab='Date',ylab='Difference in counts (%)',main='Minnesota')
abline(a=0,b=0)
```



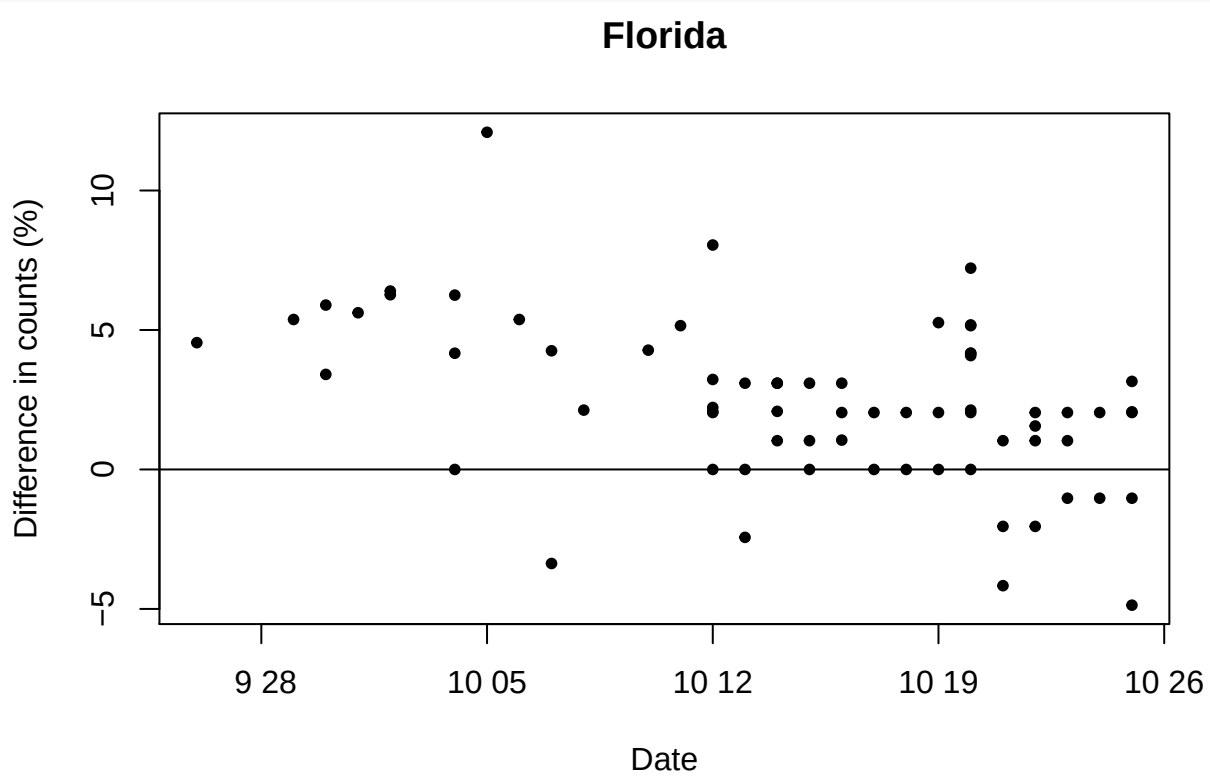
Biden was ahead of Trump in Minnesota according to the polls.

```r
mean(difference_mn * 100)
```

```
## [1] 12.04561
```

The mean lead for Biden was 12.05%.

```r
index_fl_2020=which(polls_data_2020$state=="Florida")

index_biden_fl_2020=which(polls_data_2020$answer=='Biden' & polls_data_2020$state=="Florida")
index_trump_fl_2020=which(polls_data_2020$answer=='Trump' & polls_data_2020$state=="Florida")

counts_biden_fl_2020=polls_data_2020$pct[index_biden_fl_2020]*
  polls_data_2020$sample_size[index_biden_fl_2020]
counts_trump_fl_2020=polls_data_2020$pct[index_trump_fl_2020]*
  polls_data_2020$sample_size[index_trump_fl_2020]

##plot percentage
difference_fl = (counts_biden_fl_2020-counts_trump_fl_2020)/
  (counts_biden_fl_2020+counts_trump_fl_2020)
plot(date_2020[index_trump_fl_2020], difference_fl * 100,
     col='black',pch=20,type='p',xlab='Date',ylab='Difference in counts (%)',main='Florida')
abline(a=0,b=0)
```



Biden was ahead in Florida according to the poll data.

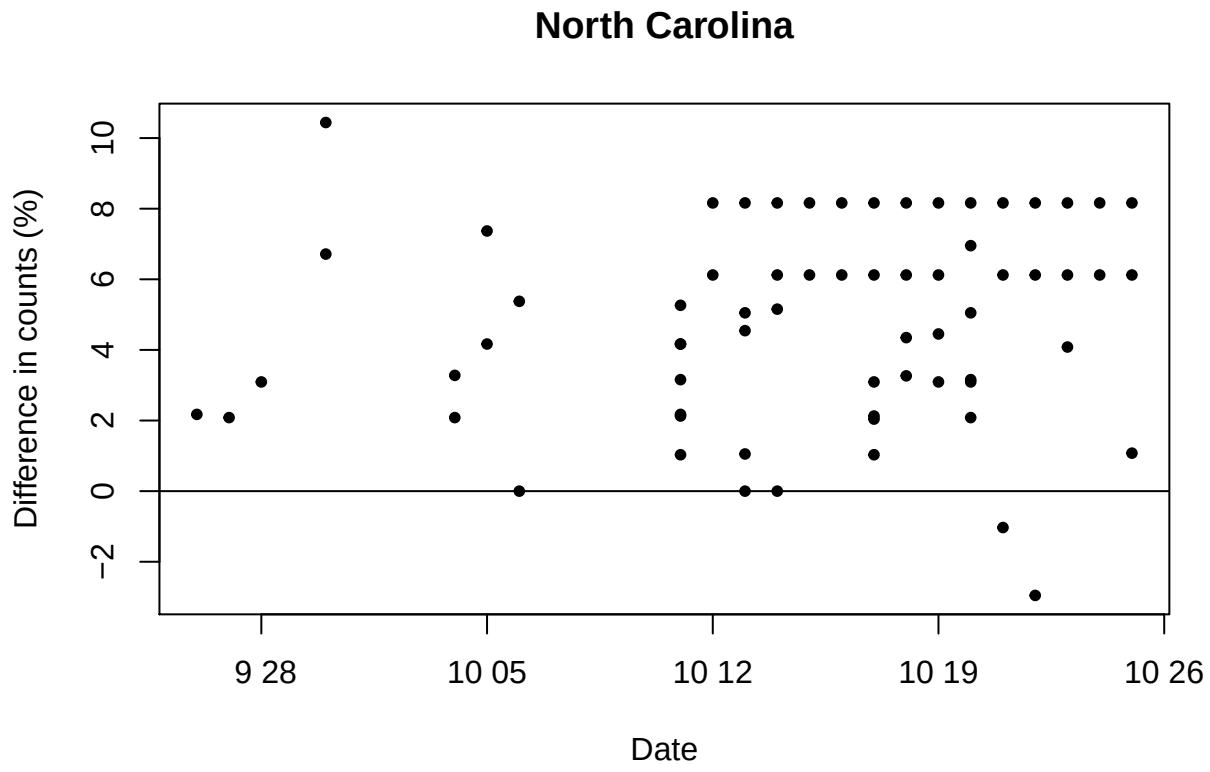
```r
mean(difference_fl * 100)
```

```
## [1] 2.363145
```

The mean difference was 2.36%. So Biden was ahead.

```r
index_nc_2020=which(polls_data_2020$state=="North Carolina")
```

```r
index_biden_nc_2020=which(polls_data_2020$answer=='Biden' & polls_data_2020$state=="North Carolina")
index_trump_nc_2020=which(polls_data_2020$answer=='Trump' & polls_data_2020$state=="North Carolina")

counts_biden_nc_2020=polls_data_2020$pct[index_biden_nc_2020]*
  polls_data_2020$sample_size[index_biden_nc_2020]
counts_trump_nc_2020=polls_data_2020$pct[index_trump_nc_2020]*
  polls_data_2020$sample_size[index_trump_nc_2020]

##plot percentage
difference_nc = (counts_biden_nc_2020-counts_trump_nc_2020)/
  (counts_biden_nc_2020+counts_trump_nc_2020)
plot(date_2020[index_trump_nc_2020], difference_nc * 100,
     col='black',pch=20,type='p',xlab='Date',ylab='Difference in counts (%)',main='North Carolina')
abline(a=0,b=0)
```



Simiarly, no one was significantly lead in North Carolina.

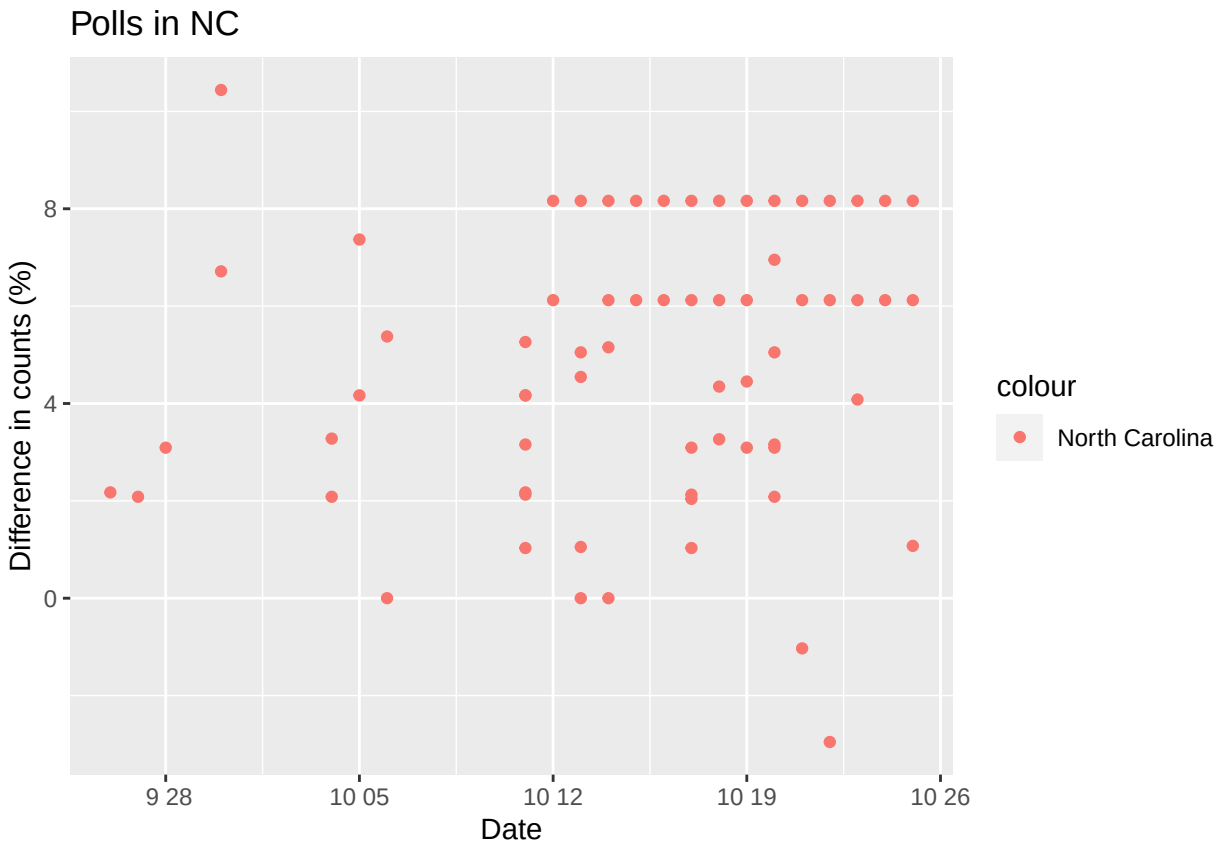
```r
mean(difference_nc * 100)
```

```
## [1] 4.737044
```

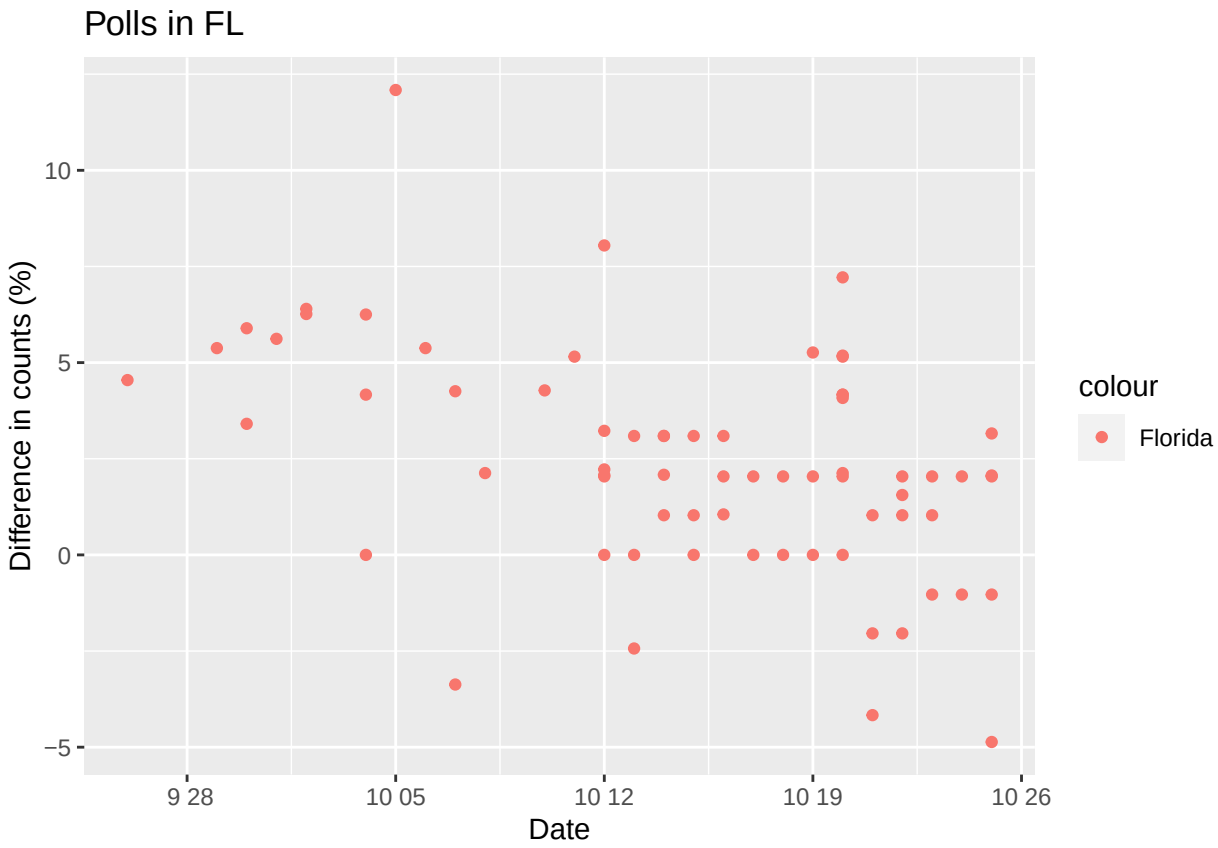
The mean difference was 4.74%. So Biden was ahead too.

Question 2.b

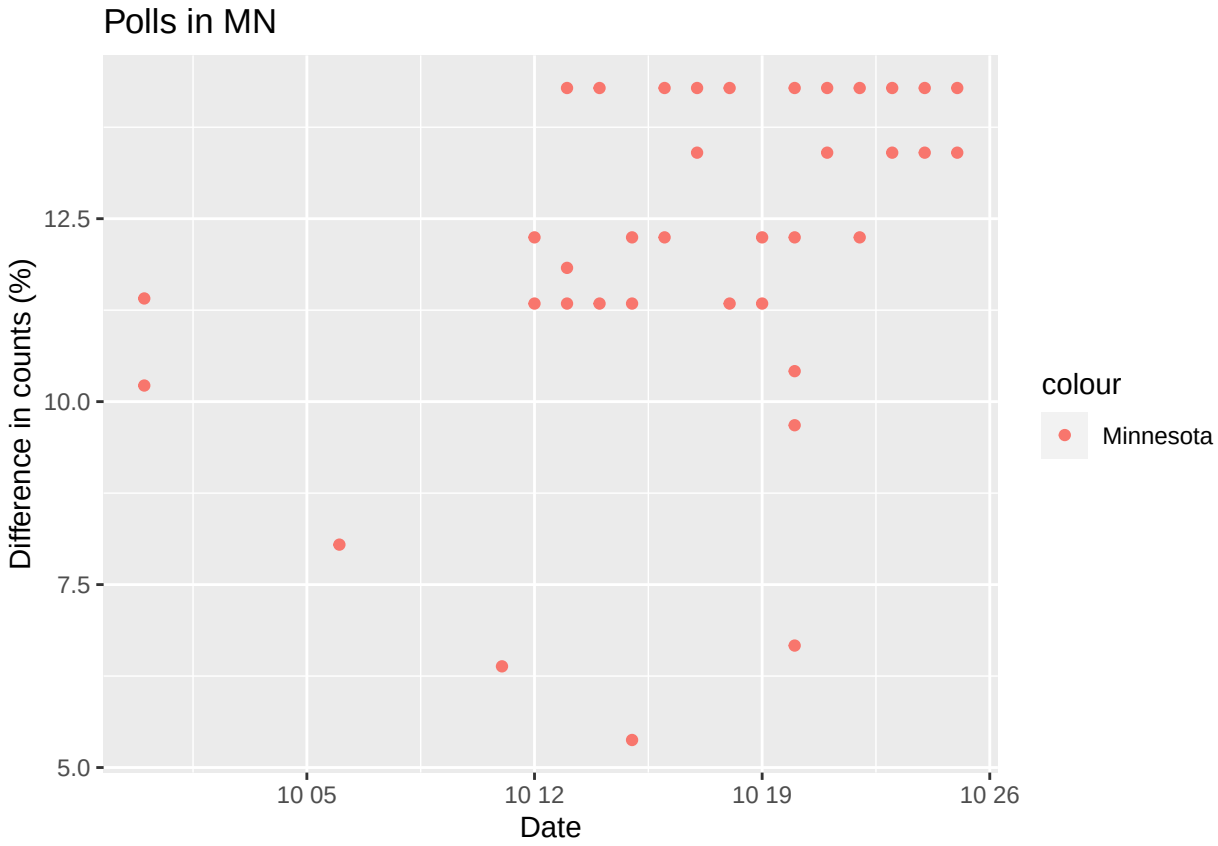
```r
ggplot() + geom_point(aes(x = date_2020[index_trump_nc_2020], y=difference_nc * 100, color="North Caroli
  ggtitle("Polls in NC") + xlab("Date") + ylab("Difference in counts (%)")
```




```
ggplot() + geom_point(aes(x = date_2020[index_trump_fl_2020], y=difference_fl * 100, color="Florida")) +
  ggtitle("Polls in FL") + xlab("Date") + ylab("Difference in counts (%)")
```




```
ggplot() + geom_point(aes(x = date_2020[index_trump_mn_2020], y=difference_mn * 100, color="Minnesota"))
  ggtitle("Polls in MN") + xlab("Date") + ylab("Difference in counts (%)")
```

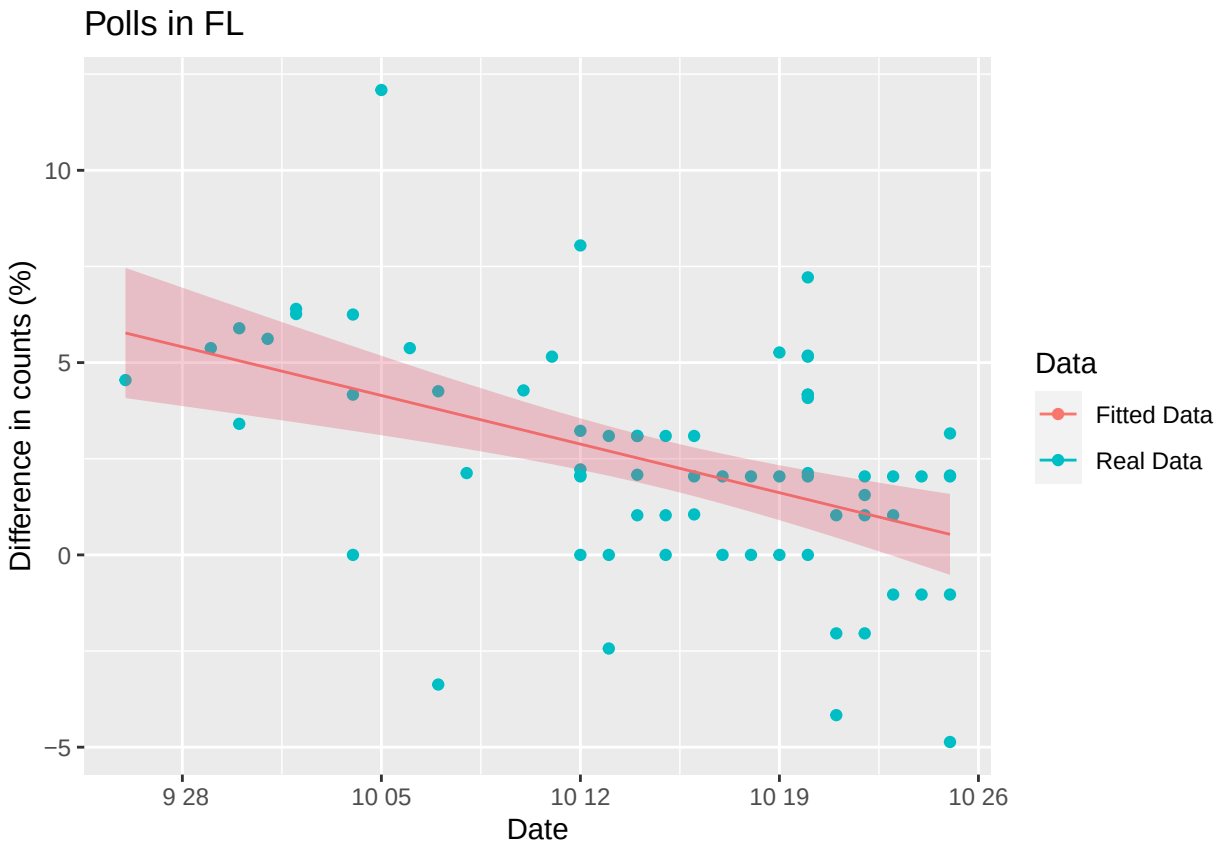


The lead in Minnesota for Biden went up, and went down in Florida. The trend was not clear in North Carolina.

Question 2.c

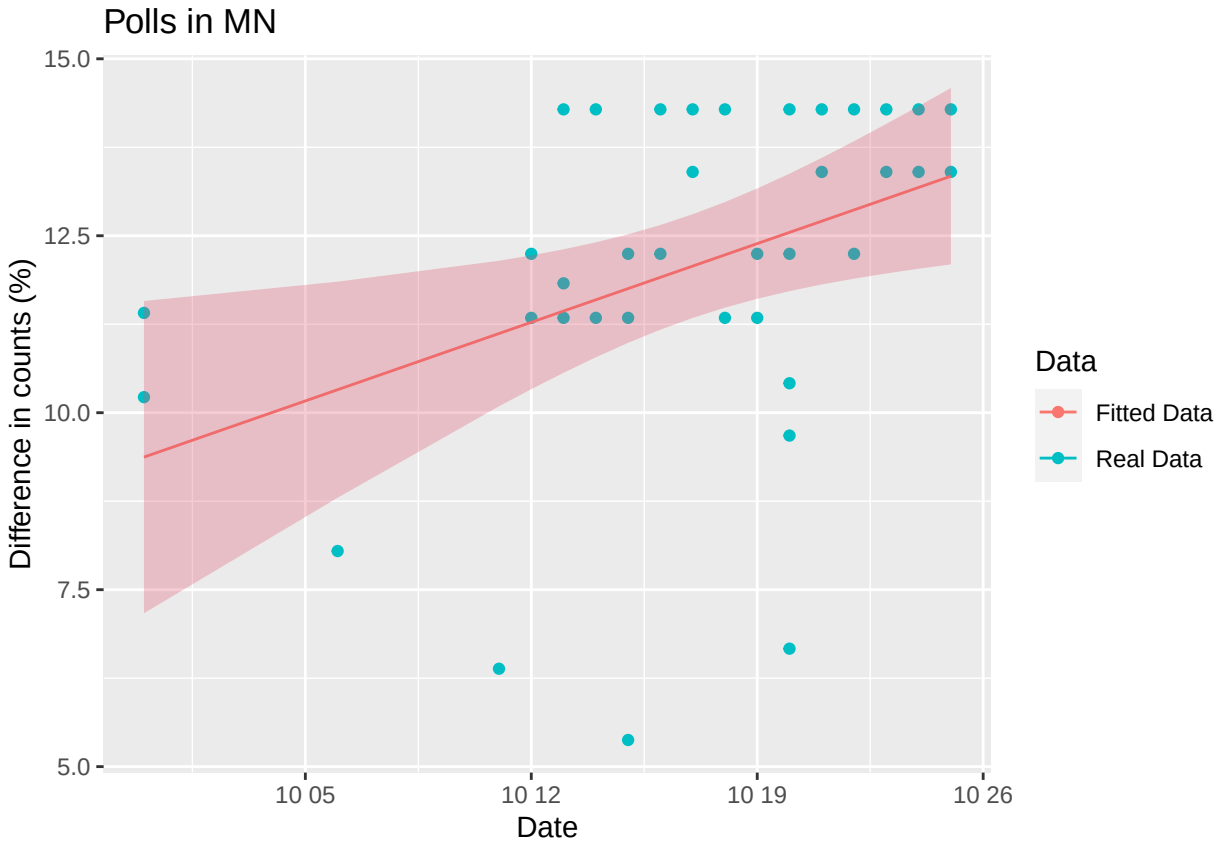
For Florida:

```
fit.fl <- lm(difference_fl~date_2020[index_trump_fl_2020])
fitted.fl <- fitted(fit.fl)
conf <- predict(fit.fl, newdata = date_2020[index_trump_fl_2020], interval = 'confidence')
ggplot() + geom_point(aes(x = date_2020[index_trump_fl_2020], y=difference_fl * 100, color="Real Data"))
  xlab("Date") + ylab("Difference in counts (%)") +
  geom_line(aes(x=date_2020[index_trump_fl_2020], y=fitted.fl * 100, color="Fitted Data")) +
  geom_ribbon(aes(x=date_2020[index_trump_fl_2020], ymin=conf[,2] * 100, ymax=conf[,3] * 100), alpha=0.
  guides(colour = guide_legend(title = "Data"))
```



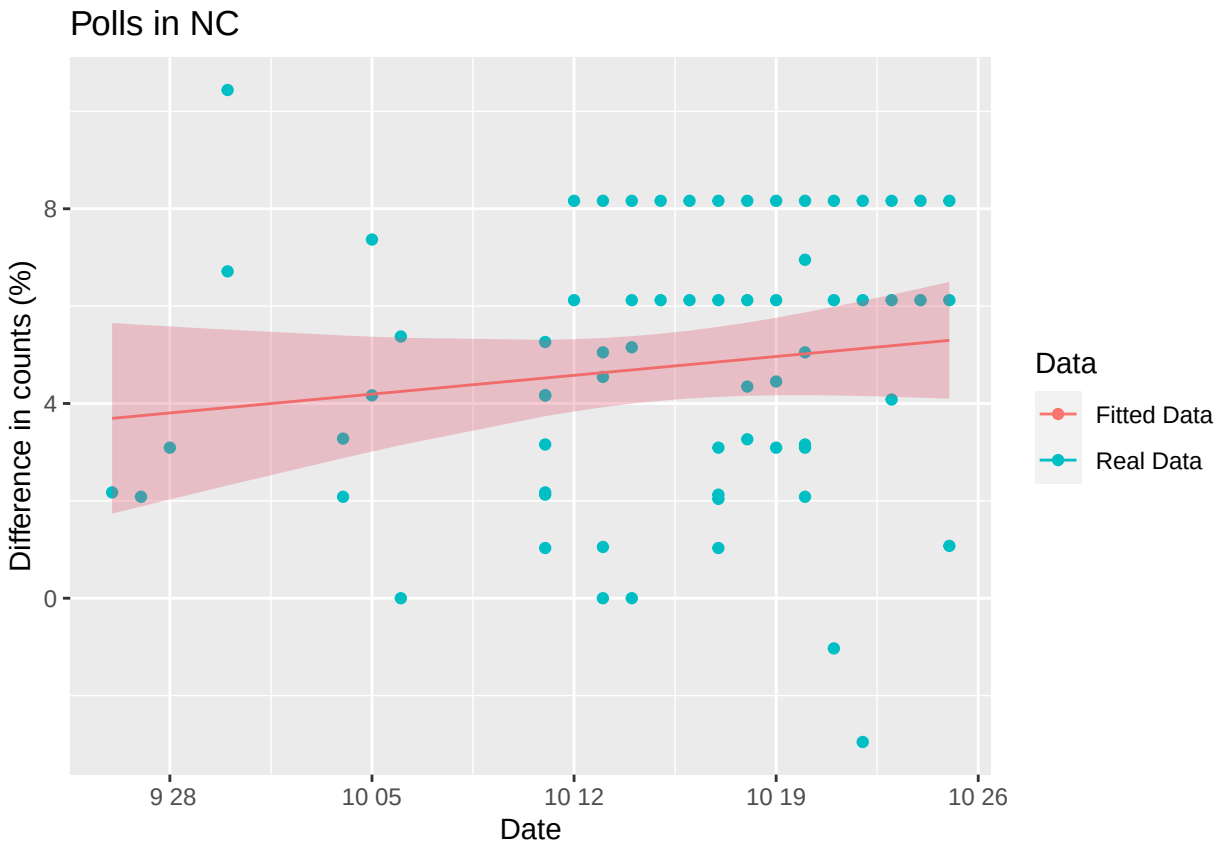
For Minnesota:

```
fit.mn <- lm(difference_mn~date_2020[index_trump_mn_2020])
fitted.mn <- fitted(fit.mn)
conf <- predict(fit.mn, newdata = date_2020[index_trump_mn_2020], interval = 'confidence')
ggplot() + geom_point(aes(x = date_2020[index_trump_mn_2020], y=difference_mn * 100, color="Real Data"))
  ggtitle("Polls in MN") + xlab("Date") + ylab("Difference in counts (%)") +
  geom_line(aes(x=date_2020[index_trump_mn_2020], y=fitted.mn * 100, color="Fitted Data")) +
  geom_ribbon(aes(x=date_2020[index_trump_mn_2020], ymin=conf[,2] * 100, ymax=conf[,3] * 100), alpha=0.
  guides(colour = guide_legend(title = "Data"))
```



For North Carolina:

```
fit.nc <- lm(difference_nc~date_2020[index_trump_nc_2020])
fitted.nc <- fitted(fit.nc)
conf <- predict(fit.nc, newdata = date_2020[index_trump_nc_2020], interval = 'confidence')
ggplot() + geom_point(aes(x = date_2020[index_trump_nc_2020], y=difference_nc * 100, color="Real Data"))
  ggtitle("Polls in NC") + xlab("Date") + ylab("Difference in counts (%)") +
  geom_line(aes(x=date_2020[index_trump_nc_2020], y=fitted.nc * 100, color="Fitted Data")) +
  geom_ribbon(aes(x=date_2020[index_trump_nc_2020], ymin=conf[,2] * 100, ymax=conf[,3] * 100), alpha=0.
  guides(colour = guide_legend(title = "Data"))
```

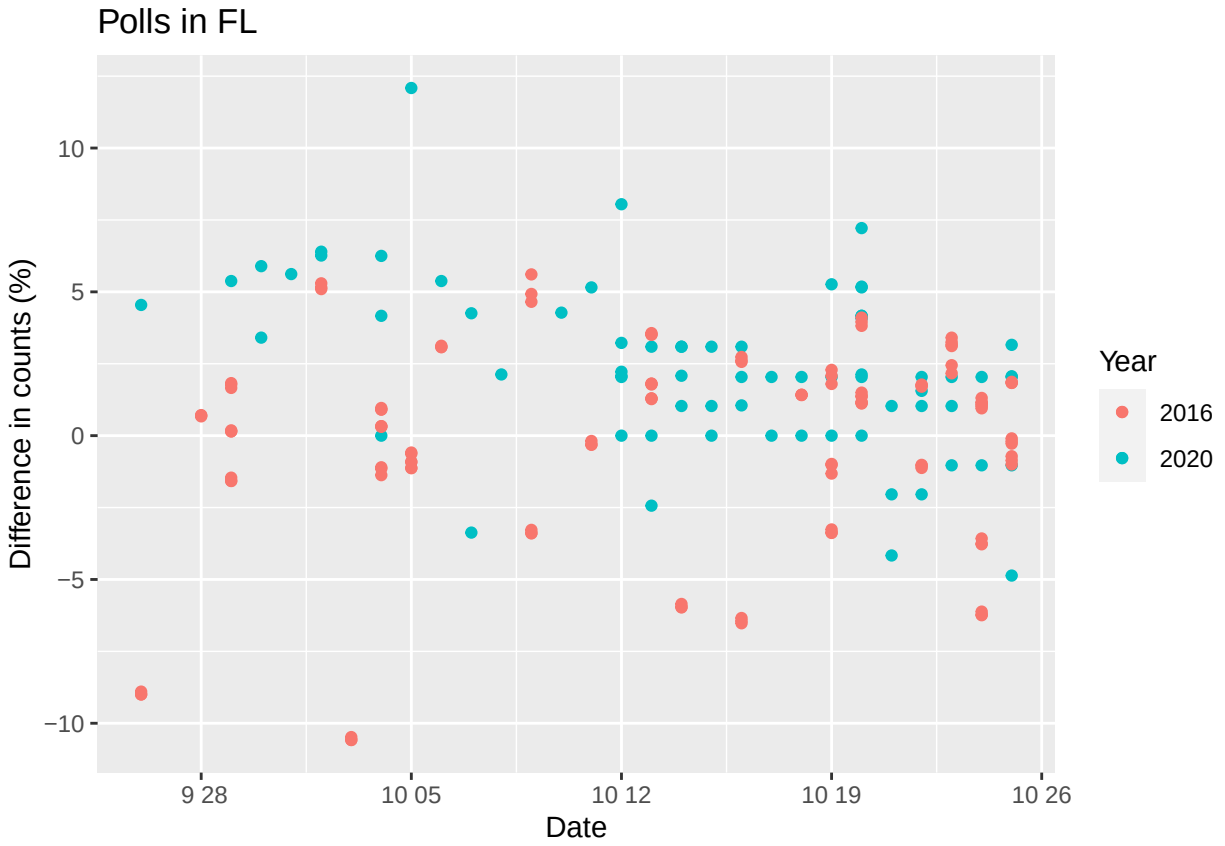


Again, Florida had the smallest margin.

Question 3

Comparison in Florida:

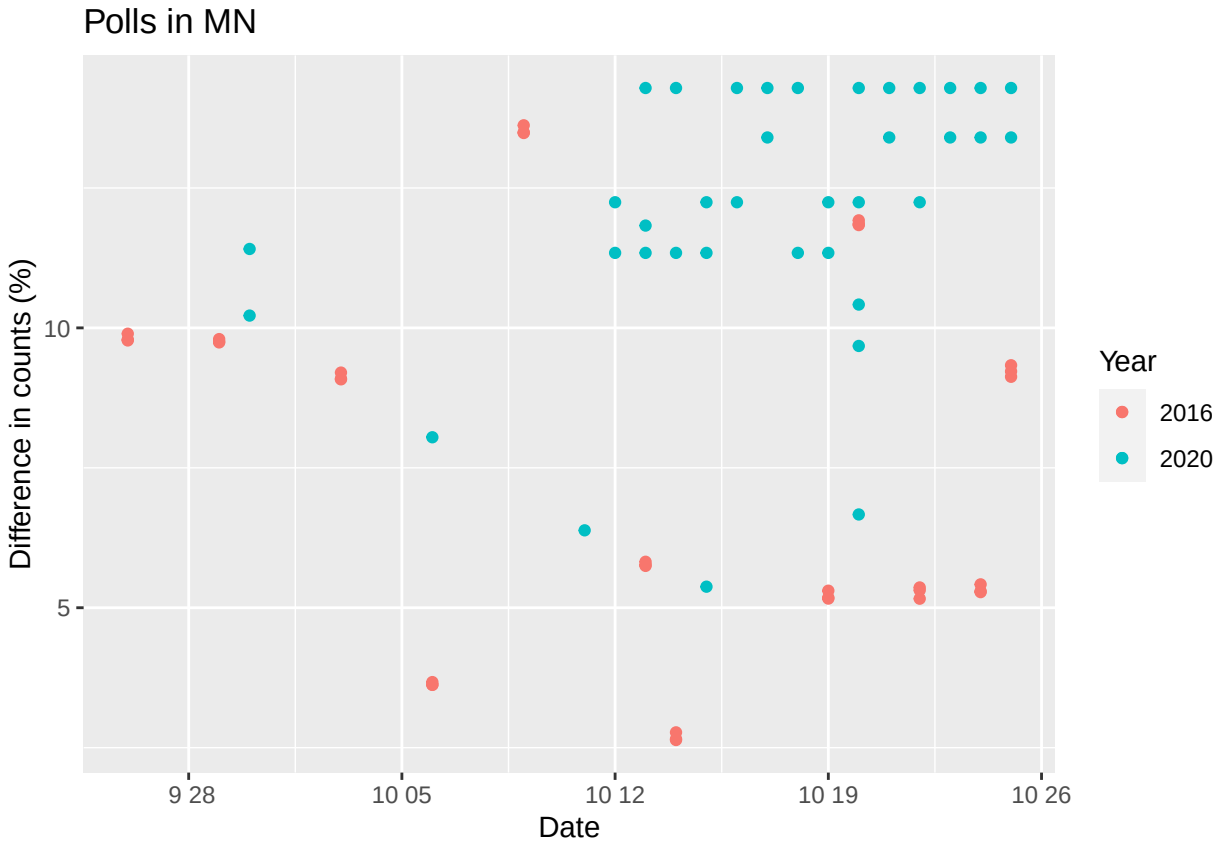
```
index_fl_2016 = which(polls_data_2016$state=="Florida")
date_fl_2016 <- mdy(polls_data_2016$enddate[index_fl_2016])
index_fl_2016 = index_fl_2016[date_fl_2016 >= "2016-09-26" & date_fl_2016 <= "2016-10-25"]
date_fl_2016 = date_fl_2016[date_fl_2016 >= "2016-09-26" & date_fl_2016 <= "2016-10-25"]

difference_fl_2016=(polls_data_2016$total.clinton[index_fl_2016]-polls_data_2016$total.trump[index_fl_2

for (i in 1:length(date_fl_2016)) year(date_fl_2016[i]) <- 2020
ggplot() + geom_point(aes(x = date_2020[index_trump_fl_2020], y=difference_fl * 100, color="2020")) +
  ggtitle("Polls in FL") + xlab("Date") + ylab("Difference in counts (%)") +
  geom_point(aes(x = date_fl_2016, y = difference_fl_2016 * 100, color="2016")) +
  guides(colour = guide_legend(title = "Year"))
```



Similarities: The lead was not significant for both candidates. The result is very uncertain. Differences: In this year, Biden took a lead mostly and in average. There are more polls in 2020 too.

Comparison in Minnesota:

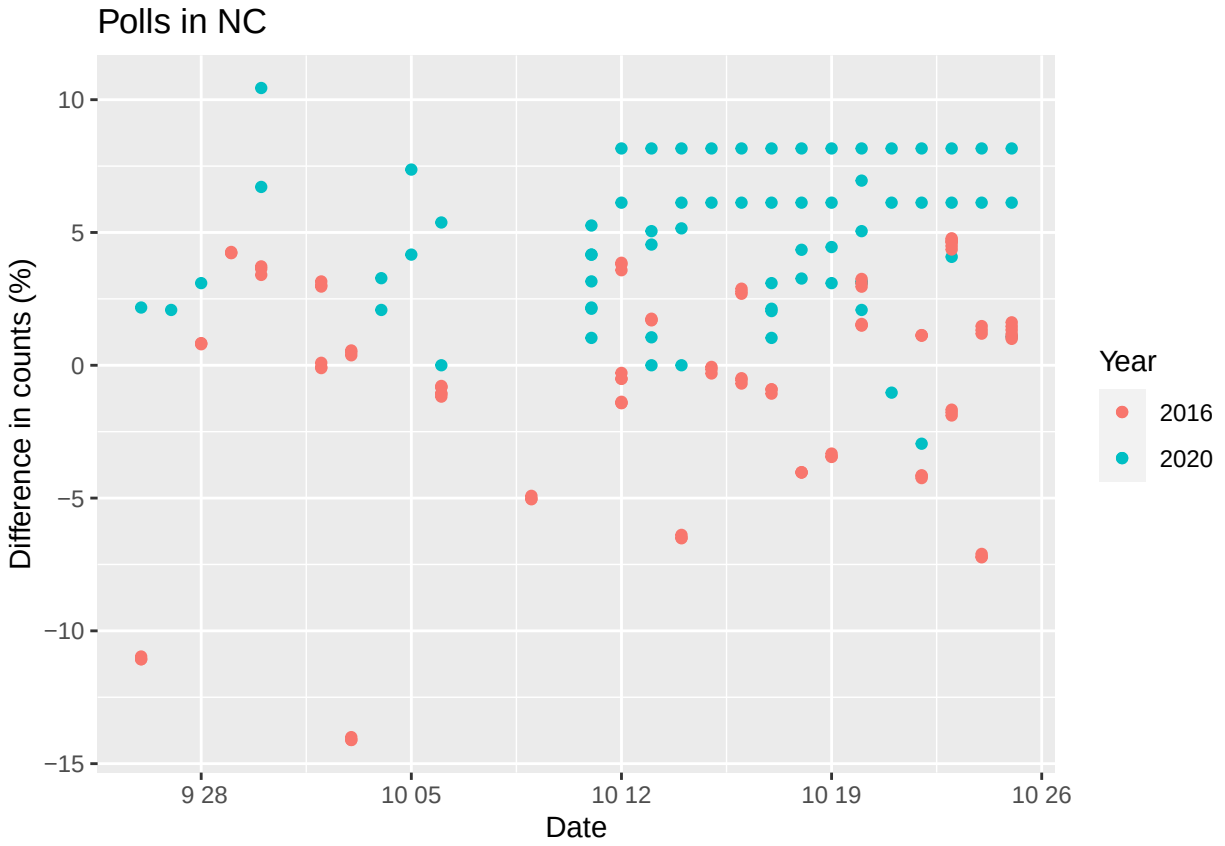
```
index_mn_2016 = which(polls_data_2016$state=="Minnesota")
date_mn_2016 <- mdy(polls_data_2016$enddate[index_mn_2016])
index_mn_2016 = index_mn_2016[date_mn_2016 >= "2016-09-26" & date_mn_2016 <= "2016-10-25"]
date_mn_2016 = date_mn_2016[date_mn_2016 >= "2016-09-26" & date_mn_2016 <= "2016-10-25"]

difference_mn_2016=(polls_data_2016$total.clinton[index_mn_2016]-
                      polls_data_2016$total.trump[index_mn_2016])/
  (polls_data_2016$total.clinton[index_mn_2016]+
     polls_data_2016$total.trump[index_mn_2016])

for (i in 1:length(date_mn_2016)) year(date_mn_2016[i]) <- 2020
ggplot() + geom_point(aes(x = date_2020[index_trump_mn_2020], y=difference_mn * 100, color="2020")) +
  ggtitle("Polls in MN") + xlab("Date") + ylab("Difference in counts (%)") +
  geom_point(aes(x = date_mn_2016, y = difference_mn_2016 * 100, color="2016")) +
  guides(colour = guide_legend(title = "Year"))
```



Similarities: All the polls show preference to one candidate.
Differences:There are more polls in 2020 too, and the lead was more significant.

Comparison in North Carolina:

```
index_nc_2016 = which(polls_data_2016$state=="North Carolina")
date_nc_2016 <- mdy(polls_data_2016$enddate[index_nc_2016])
index_nc_2016 = index_nc_2016[date_nc_2016 >= "2016-09-26" & date_nc_2016 <= "2016-10-25"]
date_nc_2016 = date_nc_2016[date_nc_2016 >= "2016-09-26" & date_nc_2016 <= "2016-10-25"]

difference_nc_2016=(polls_data_2016$total.clinton[index_nc_2016]-polls_data_2016$total.trump[index_nc_2
  (polls_data_2016$total.clinton[index_nc_2016]+polls_data_2016$total.trump[index_nc_2016])

for (i in 1:length(date_nc_2016)) year(date_nc_2016[i]) <- 2020
ggplot() + geom_point(aes(x = date_2020[index_trump_nc_2020], y=difference_nc * 100, color="2020")) +
  ggtitle("Polls in NC") + xlab("Date") + ylab("Difference in counts (%)") +
  geom_point(aes(x = date_nc_2016, y = difference_nc_2016 * 100, color="2016")) +
  guides(colour = guide_legend(title = "Year"))
```



Similarities: There was no obvious linear trend in these polls. Differences:There are more polls in 2020 too, and the lead was more significant. The lead seemed more consistent.

Other comments: This year, polls showed a stronger advantage for Biden.

Question 4

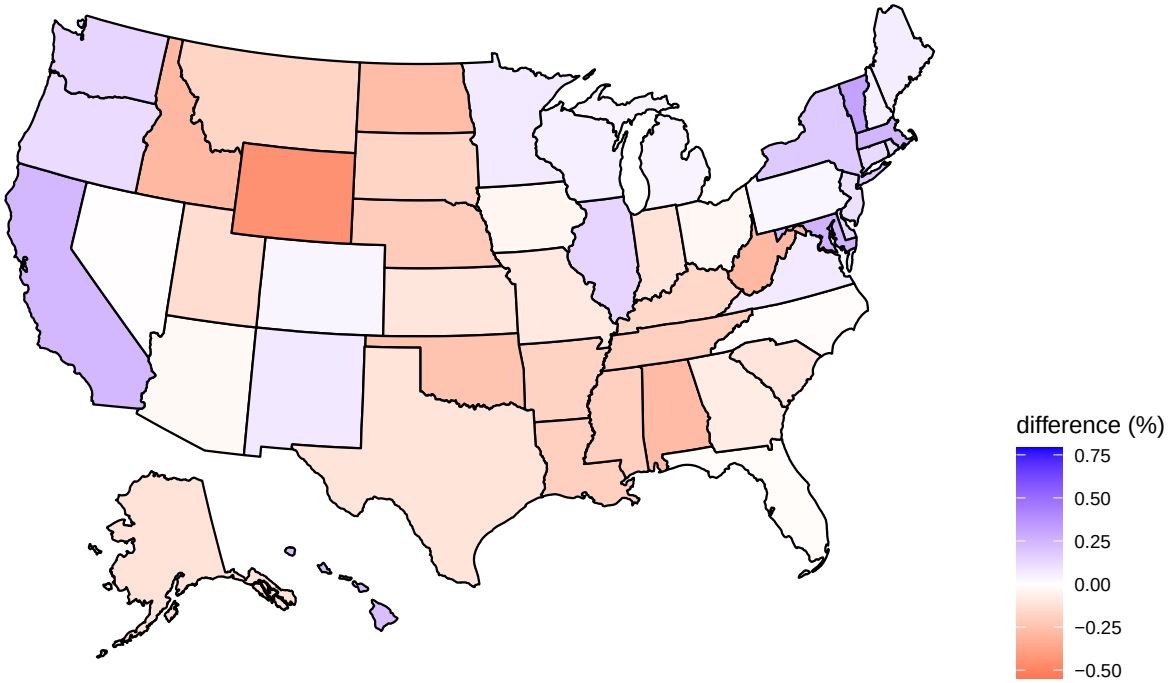
Question 4.a

```
polls_data_2020=read.csv(paste0(current_address,"/data/president_polls_2020.csv"))

date_2020= mdy(polls_data_2020$end_date)
date_2020_latest_day=date_2020[1]
index_selected=which(date_2020>='2020-08-31' & date_2020 <='2020-10-25')
polls_data_2020=polls_data_2020[index_selected,]
polls_data_2020=polls_data_2020[which(polls_data_2020$answer=='Biden'|polls_data_2020$answer=='Trump'),]
polls_data_2020_question_id_num=unique(polls_data_2020$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,]
  }
}

date_2020= mdy(polls_data_2020$end_date)
```

```
polls_data_2016_enddate=mdy(polls_data_2016$enddate)
polls_data_2016_after_sep=polls_data_2016[which(polls_data_2016_enddate>="2016-08-31"&polls_data_2016_en

poll_state_sum_clinton_2016=
  aggregate(polls_data_2016_after_sep$total.clinton,
            by=list(State=polls_data_2016_after_sep$state),FUN=sum)
poll_state_sum_trump_2016=
  aggregate(polls_data_2016_after_sep$total.trump,
            by=list(State=polls_data_2016_after_sep$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])
poll_state_diff_percentage=poll_state_diff_percentage[poll_state_diff_percentage[,1]!='U.S.',]

library(usmap)
library(ggplot2)

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


index_selected=which(date_2020>='2020-08-31' & date_2020<='2020-10-25' )
polls_data_2020_after_sep=polls_data_2020[index_selected,]

polls_data_2020_after_sep=polls_data_2020_after_sep[which(polls_data_2020$answer=='Biden'|polls_data_202

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),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))
##2016
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")
```

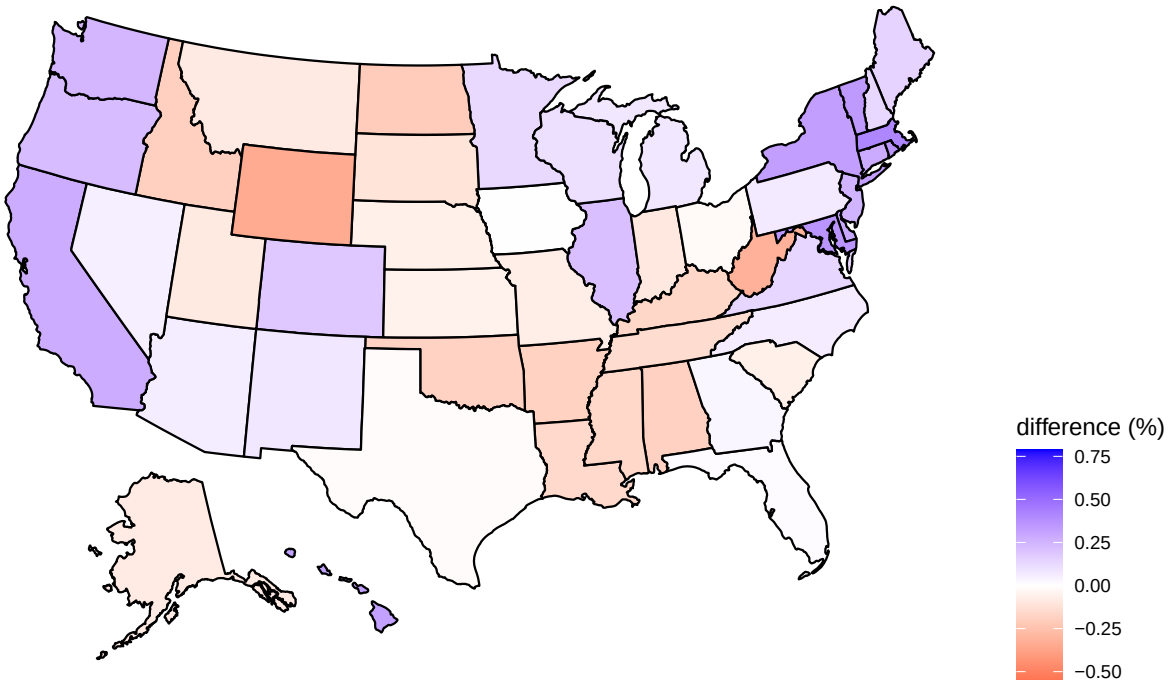
2016



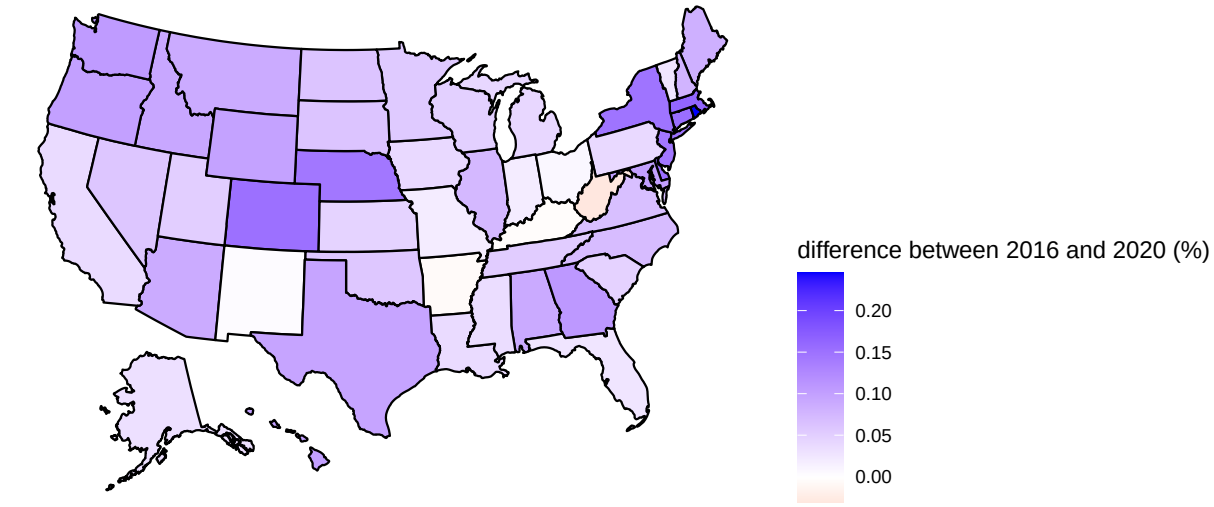

```
##2020
plot_usmap(data = state_poll_2020, values = "diff_percentage", color = "black") +
  scale_fill_gradient2(name = "difference (%)",   low= "red",
                       mid = "white",
                       high = "blue",
                       midpoint = 0,limits=limit_val)+
```

```
  theme(legend.position = "right")+
ggtitle("2020")
```

2020




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


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")
```



In these states, Biden is doing better than Clinton: South Carlina, Virginia, Colorado, Nevada, New Hampshire, Montana, North Carolina, Arizona, Alabama, Minnesota, New York, Texas, Georgia, New Jersey, Maryland, Massachusettes, Connecticut, Idaho, Oregon, Tennessee, Washington, Delaware, Rhode Island, North Dakota, Hawaii, South Dakota, DC, Nebraska.

In these states, Biden is doing just as well as Clinton: Wisconsin, Michigan, Iowa, Florida, Utah, Illinois, Alaska, Pennsylvania, Ohio, Kansas, Missouri, Maine, California, Mississippi, Indiana, Oklahoma, Louisiana, Arkansas, Vermont, Wyoming, New Mexico, Kentucky. In these states, Biden is worse than Clinton: West Virginia.

We can see that, there are more polls in this year, especially when close to the Election day. ## Question 4.b Based on the plots, if no more has a significant lead like consectively lead in 10 days or lead by over 5% in average, or they were red in 2016 but showed signs of turning blue, then I will call these states "battleground" states.

So battleground states are, Iowa, Wisconsin, Nevada, Florida, North Carolina, Michigan, Pennsylvania, Texas, Ohio, Georgia.

Question 4.c

```
battleground = c("Iowa", "Wisconsin", "Nevada", "Florida", "North Carolina", "Michigan", "Pennsylvania"
for (state in battleground) {
  cat(paste("Percentage Difference in", state, "is", state_poll_2020_2016_diff$diff[state_poll_2020_201
}
```

```
## Percentage Difference in Iowa is 0.039304715359788
## Percentage Difference in Wisconsin is 0.0497442571680234
## Percentage Difference in Nevada is 0.0596269090391235
## Percentage Difference in Florida is 0.027551947661719
## Percentage Difference in North Carolina is 0.0700092788301306
## Percentage Difference in Michigan is 0.0415971556006832
## Percentage Difference in Pennsylvania is 0.0397121138402842
## Percentage Difference in Texas is 0.0957358911438733
## Percentage Difference in Ohio is 0.00841241244846511
## Percentage Difference in Georgia is 0.10896572896696
```

Question 4.d

1. If we are going to make prediction on the presidential result only, we should more focus on the swing states (MI, WI, NC, etc.) rather than the national poll. A lead in the national poll does not necessarily mean winning the election.
2. Take the margin of error into account. Some polls had a large margin of error like 3%, so a lead by 2% can make mistakes.
3. Pay attention to the "shy" voters that had never responsed to polls. There are a large amount of shy voters among the states in the rust belt like Wisconsin, Michigan and North Carolina.

Question 5

5 states: Michigan, Arizona, Wisconsin, Pennsylvania, North Carolina

Take Wisconsin for an example, Trump won Wisconsin in 2016 by only 0.7 percentage, showing an approximately 6% of "shy" voters. This year, the margin is over 6% so it is likely that Wisconsin will turn blue. Other states share the same situation.

Question 6

Results: Iowa: Trump Texas: Biden Ohio: Biden Georgia: Biden North Carolina: Biden Arizon: Biden Florida: Biden Wisconsin: Biden Pennsylvania: Biden Michigan: Biden Minnesota: Biden Nevada: Biden

Reason: Texas and Florida has a very high turnout compared with 2016. It is likely that they become blue. I am optimistic so I think Biden can win them. The rest are judged by polls average and early vote data.