Problem Set 5

Multivariate Visualization

[YOUR NAME]

Due Date: 2024-09-27

Getting Set Up

Open `RStudio` and create a new RMarkDown file (`.Rmd`) by going to `File -> New File -> R Markdown...` . Accept defaults and save this file as `[LAST NAME]_ps5.Rmd` to your `code` folder.

Copy and paste the contents of this `.Rmd` file into your `[LAST NAME]_ps5.Rmd` file. Then change the `author: [Your Name]` to your name.

We will be using two datasets this week. The first is the `Pres2020_PV.Rds` file from the course github page (https://github.com/jbisbee1/DS1000_S2024/blob/main/data/Pres2020_PV.Rds) and should be saved to an object called `pres` . The second is the `Pres2020_StatePolls.Rds` file from the course github page (https://github.com/jbisbee1/DS1000_S2024/blob/main/data/Pres2020_StatePolls.Rds) and should be saved to an object called `state` .

All of the following questions should be answered in this `.Rmd` file. There are code chunks with incomplete code that need to be filled in. To submit, compile (i.e., `knit`) the completed problem set and upload the PDF file to Brightspace on Friday by midnight. Instructions for how to compile the output as a PDF can be found in Problem Set 0 (https://github.com/jbisbee1/DS1000_F2024/blob/main/Psets/ds1000_pset_0.pdf) and in this gif tutorial (https://github.com/jbisbee1/DS1000_F2024/blob/main/Psets/save_as_pdf.gif).

This problem set is worth 5 total points, plus 1 extra credit point. The point values for each question are indicated in brackets below. To receive full credit, you must have the correct code. In addition, some questions ask you to provide a written response in addition to the code.

You will be deducted 1 point for each day late the problem set is submitted, and 1 point for failing to submit in the correct format.

You are free to rely on whatever resources you need to complete this problem set, including lecture notes, lecture presentations, Google, your classmates…you name it. However, the final submission must be complete by you. There are no group assignments.

Note that the TAs and professors will not respond to Campuswire posts after 5PM on Friday, so don't wait until the last minute to get started!

Good luck!

*Copy the link to ChatGPT you used here: _____

Question 0

Require `tidyverse` and load the `Pres2020_PV.Rds` data to an object called `pres` .

```
require(tidyverse)
```

```
## Loading required package: tidyverse
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
pres <- read_rds("https://github.com/jbisbee1/DS1000_F2024/raw/main/data/Pres2020_PV.Rds")
```

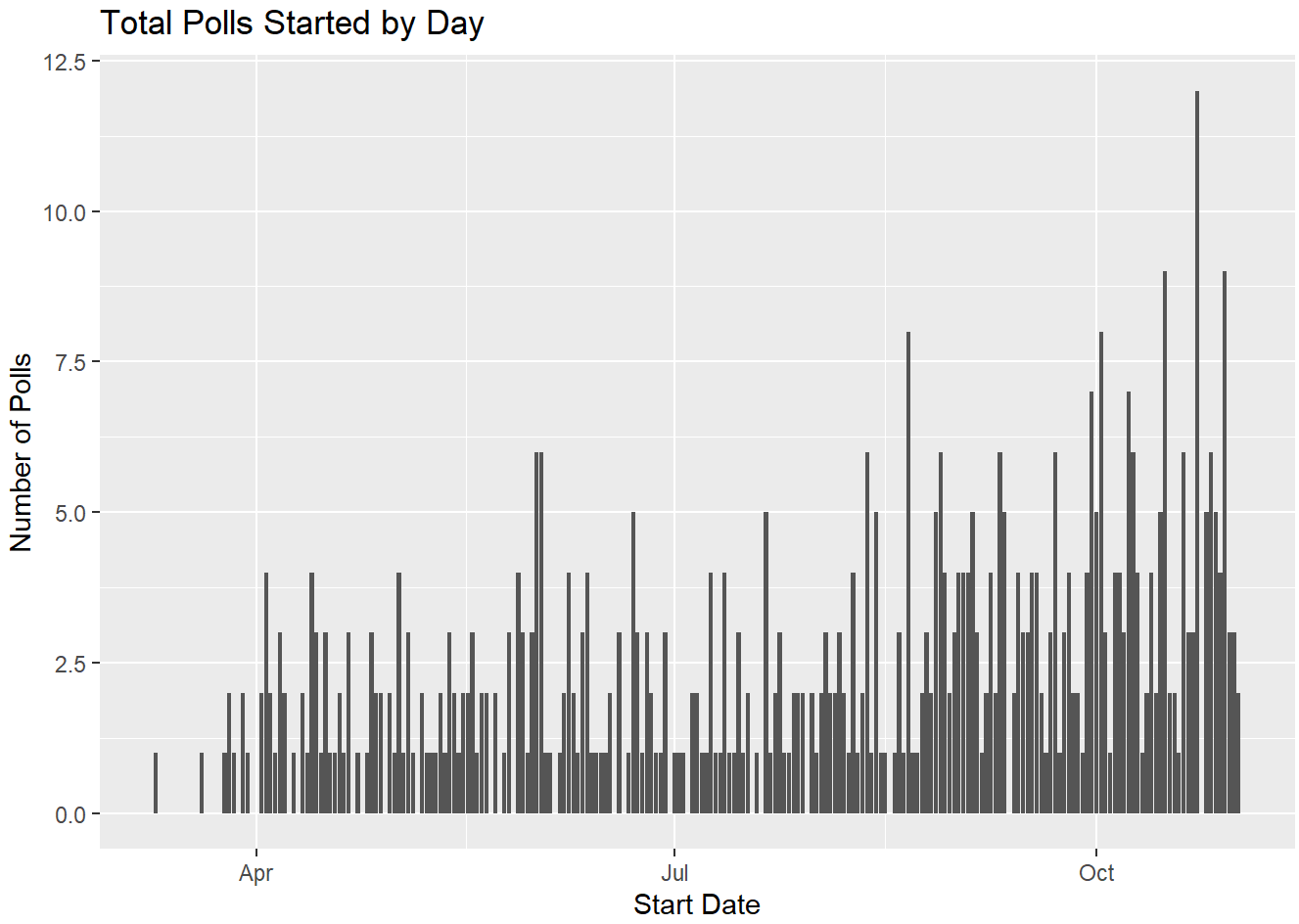
Question 1 [1 point]

Consider the following hypothesis: "Most Americans don't pay very much attention to politics, and don't know who they will vote for until very close to the election. Therefore polling predictions should be more accurate closer to the election." Based on this hypothesis and theoretical intuition, which variable is the X variable and which is the Y variable(s)?

- Based on this hypothesis, the X variable should be date the poll was fielded and the Y variable should be the prediction error. [RUBRIC: 0.25 points for correct answer. 0 points otherwise.]

Now let's first look at each variable by itself using univariate visualization. First, plot the total number of polls per start date in the data. NB: you will have convert `StartDate` to a `date` class with `as.Date()`. If you need help, see this post (https://www.r-bloggers.com/2013/08/date-formats-in-r/). Do you observe a pattern in the number of polls over time? Why do you think this is?

```
pres %>%
  mutate(StartDate = as.Date(StartDate,'%m/%d/%Y')) %>%
  ggplot(aes(x = StartDate)) +
  geom_bar(stat = 'count') +
  labs(title = 'Total Polls Started by Day',
       x = 'Start Date',
       y = 'Number of Polls')
```



There are more polls fielded the closer we get to the election. This is probably because there is more demand for information about the election the closer the election is, and media outlets want to make more money by selling polling data. [RUBRIC: 0.75 points for correct plot and answer. 0.5 points for correct plot but incoherent written answer. 0.25 points for incorrect plot and incoherent answer. 0 points for no attempt.]

Question 2 [1 point]

Next, let's look at the other variables. Calculate the **prediction error** *for Biden (call this variable* `demErr` *) and Trump (call this variable* `repErr` *) such that positive values mean that the poll* **overestimated** *the candidate's popular vote share (* `DemCertVote` *for Biden and* `RepCertVote` *for Trump).*

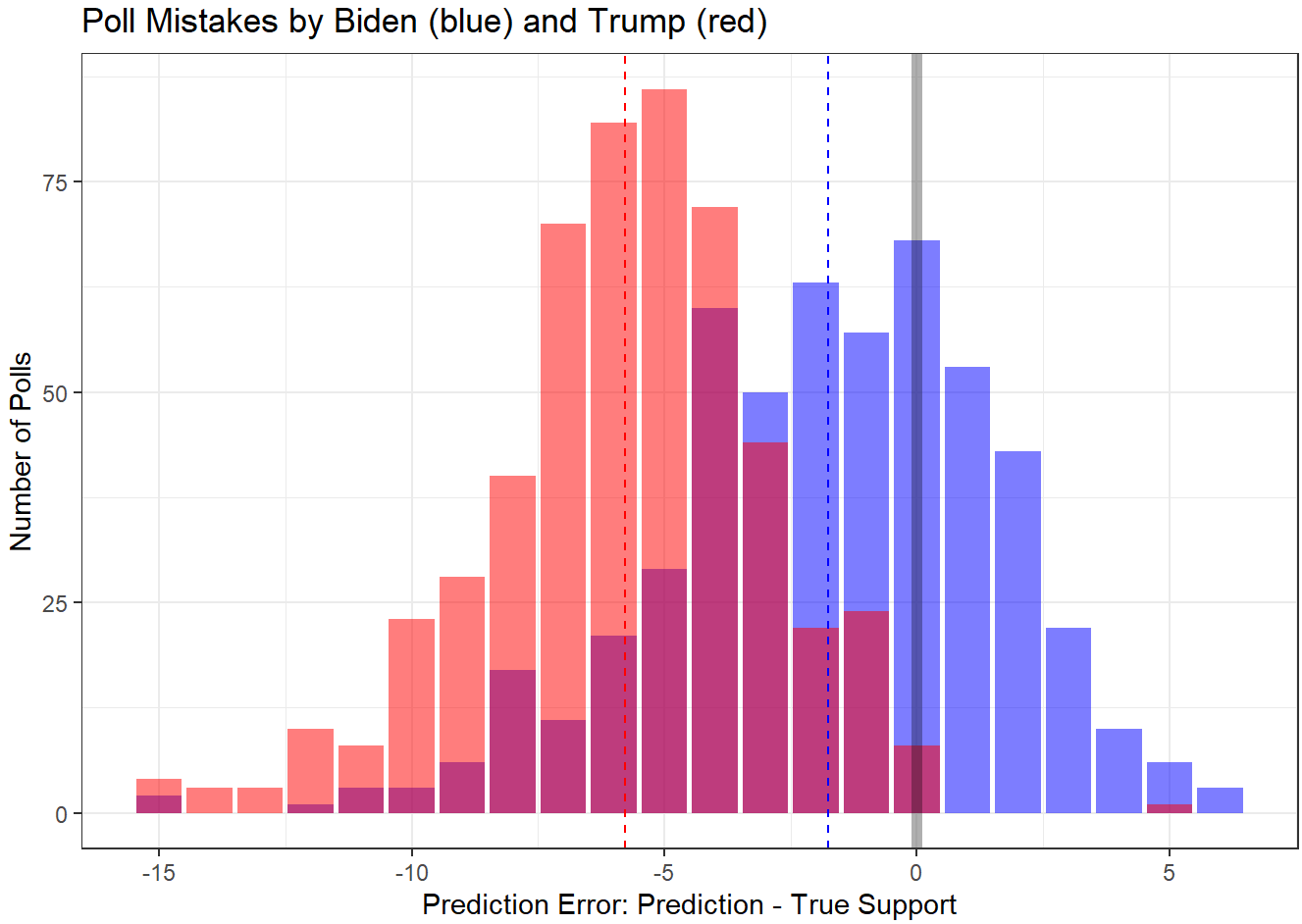
```
pres <- pres %>%
  mutate(demErr = Biden - DemCertVote,
         repErr = Trump - RepCertVote)
```

[RUBRIC: 0.25 points for correct code. 0 points for incorrect code or no answer.]

Plot the Biden and Trump prediction errors on a single plot using `geom_bar()` *, with red indicating Trump and blue indicating Biden (make sure to set alpha to some value less than 1 to increase the transparency!). Add vertical lines for the average prediction error for both candidates (colored appropriately) as well as a vertical line indicating*

no prediction error.

```
pres %>%
  ggplot() +
  geom_bar(aes(x = demErr),fill = 'blue',alpha = .5) +
  geom_bar(aes(x = repErr),fill = 'red',alpha = .5) +
  labs(title = 'Poll Mistakes by Biden (blue) and Trump (red)',
       x = 'Prediction Error: Prediction - True Support',
       y = 'Number of Polls') +
  theme_bw() +
  geom_vline(xintercept = 0,lwd = 2,alpha = .3) +
  geom_vline(xintercept = mean(pres$demErr,na.rm=T),color = 'blue',linetype = 'dashed') +
  geom_vline(xintercept = mean(pres$repErr,na.rm=T),color = 'red',linetype = 'dashed')
```



[RUBRIC: 0.5 points for correct plot (be flexible on how identical the student's plot is compared to this). 0.25 points for an attempted plot but serious mistakes. 0 points for no attempt.]

Do you observe a systematic bias toward one candidate or the other?

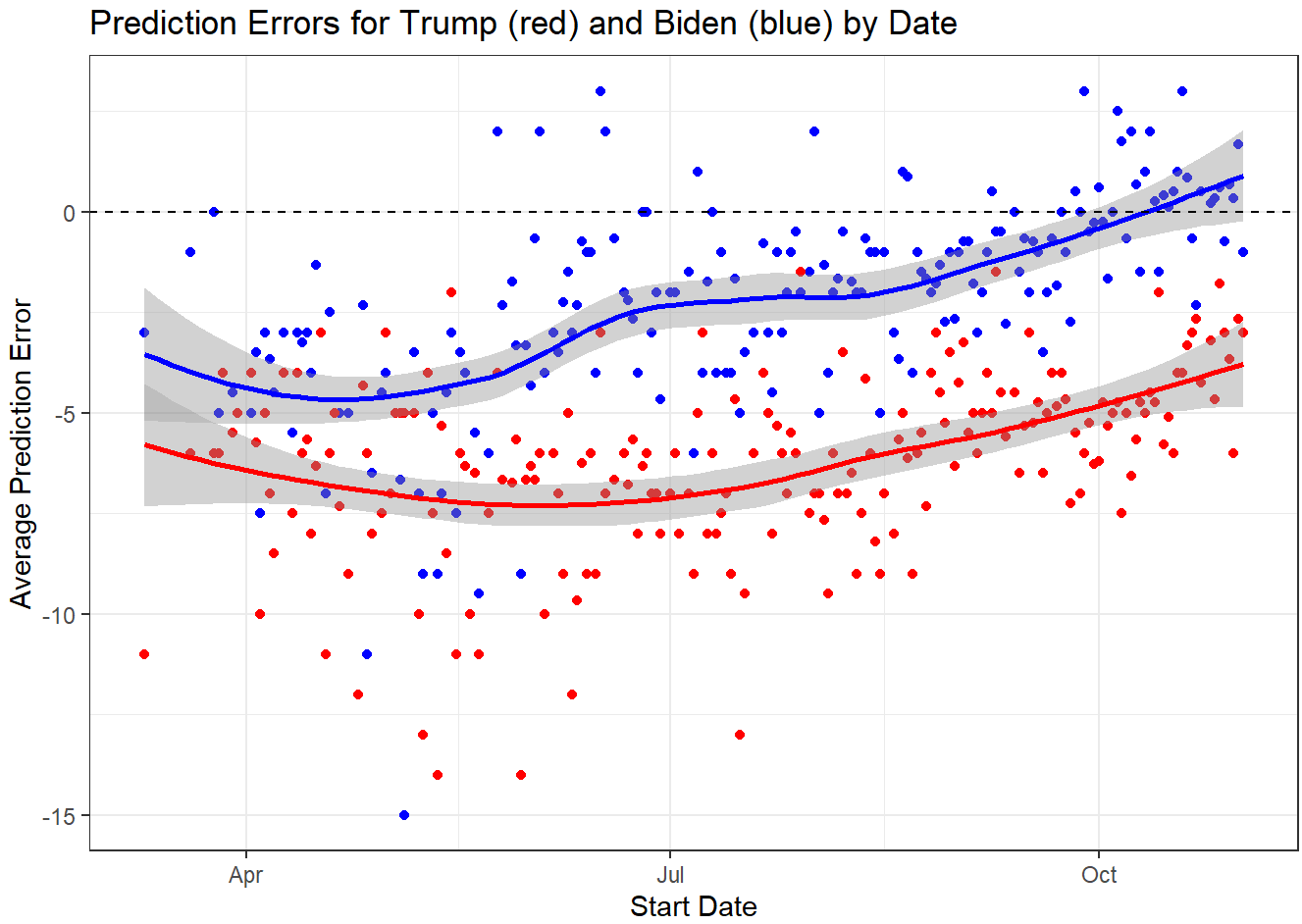
I observe a systematic bias against both candidates where the polls underestimate the amount of support for Biden and Trump. However, the magnitude of this bias against Trump is larger than the bias against Biden. [RUBRIC: 0.25 points for any coherent answer. 0 points for no attempt.]

Question 3 [1 point]

Plot the average prediction error for Trump (red) and Biden (blue) by start date using `geom_point()` *and add a curvey line of best fit using* `geom_smooth()` *(allow it to be curved!).*

```
pres %>%
  mutate(StartDate = as.Date(StartDate,'%m/%d/%Y')) %>%
  group_by(StartDate) %>%
  summarise(demErr = mean(demErr),
            repErr = mean(repErr)) %>%
  ggplot() +
  geom_point(aes(x = StartDate,y = demErr),color = 'blue') +
  geom_point(aes(x = StartDate,y = repErr),color = 'red') +
  geom_smooth(aes(x = StartDate,y = demErr),color = 'blue') +
  geom_smooth(aes(x = StartDate,y = repErr),color = 'red') +
  labs(title = "Prediction Errors for Trump (red) and Biden (blue) by Date",
       x = "Start Date",
       y = "Average Prediction Error") +
  geom_vline(xintercept = 0,linetype = 'dashed') +
  geom_hline(yintercept = 0,linetype = 'dashed') +
  theme_bw()
```

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



[RUBRIC: 0.75 point for correct plot. 0.5 points for plots with straight lines or no lines, or other mistakes in coloring / general aesthetics. 0.25 points for no labels or more serious mistakes in the plot. 0 points for no attempt.]

What pattern do you observe over time, if any? Does this support the hypothesis presented in Question 1 above?

I observe a gradual decline in the prediction error over time, where polls underestimate both Trump and Biden less and less. However, polls still underestimated Trump by the time of the election, whereas they perfectly predicted Biden's support. Overall, the hypothesis is supported (prediction errors get smaller as we move closer to the election), but there is still evidence of an anti-Trump bias in polling in the 2020 election. [RUBRIC: 0.25 points for correct answer. 0.1 point for ignoring the worse accuracy for Trump than for Biden. 0 points if no attempt was made.]

Question 4 [1 point]

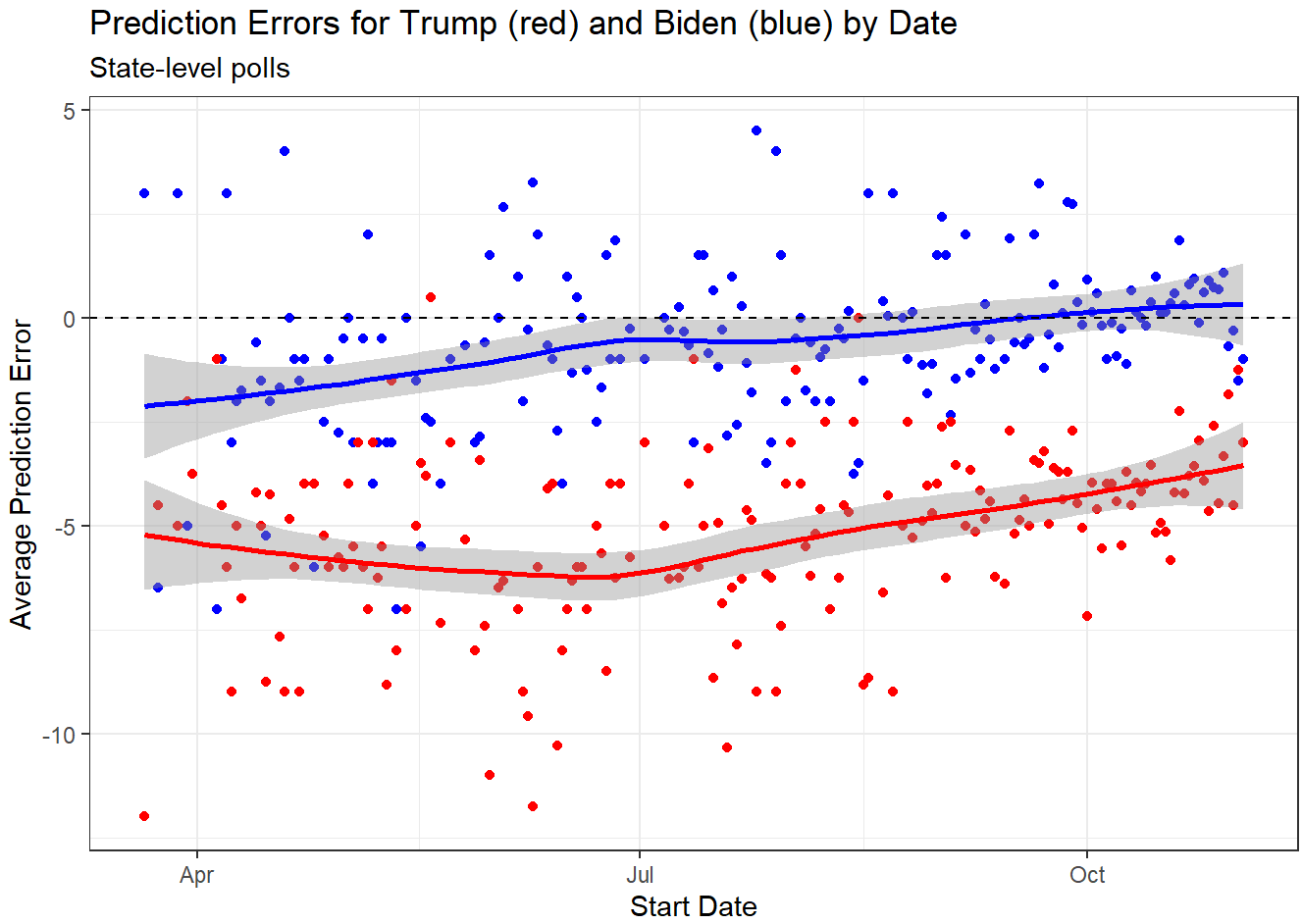
Can we do better by aggregating state-level polls? Load the [Pres2020_StatePolls.Rds] to an object called `state`*. First, create two new variables* `demErr` *and* `repErr` *just as you did in Question 2. Then recreate the same overtime plot comparing Biden and Trump prediction errors as you did in Question 3. What do you observe?*

```
state <- read_rds("https://github.com/jbisbee1/DS1000_F2024/raw/main/data/Pres2020_StatePolls.Rds")

state <- state %>%
  mutate(demErr = Biden - BidenCertVote,
         repErr = Trump - TrumpCertVote)

state %>%
  mutate(StartDate = as.Date(StartDate,'%m/%d/%Y')) %>%
  group_by(StartDate) %>%
  summarise(demErr = mean(demErr),
            repErr = mean(repErr)) %>%
  ggplot() +
  geom_point(aes(x = StartDate,y = demErr),color = 'blue') +
  geom_point(aes(x = StartDate,y = repErr),color = 'red') +
  geom_smooth(aes(x = StartDate,y = demErr),color = 'blue') +
  geom_smooth(aes(x = StartDate,y = repErr),color = 'red') +
  labs(title = "Prediction Errors for Trump (red) and Biden (blue) by Date",
       subtitle = 'State-level polls',
       x = "Start Date",
       y = "Average Prediction Error") +
  geom_vline(xintercept = 0,linetype = 'dashed') +
  geom_hline(yintercept = 0,linetype = 'dashed') +
  theme_bw()
```

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



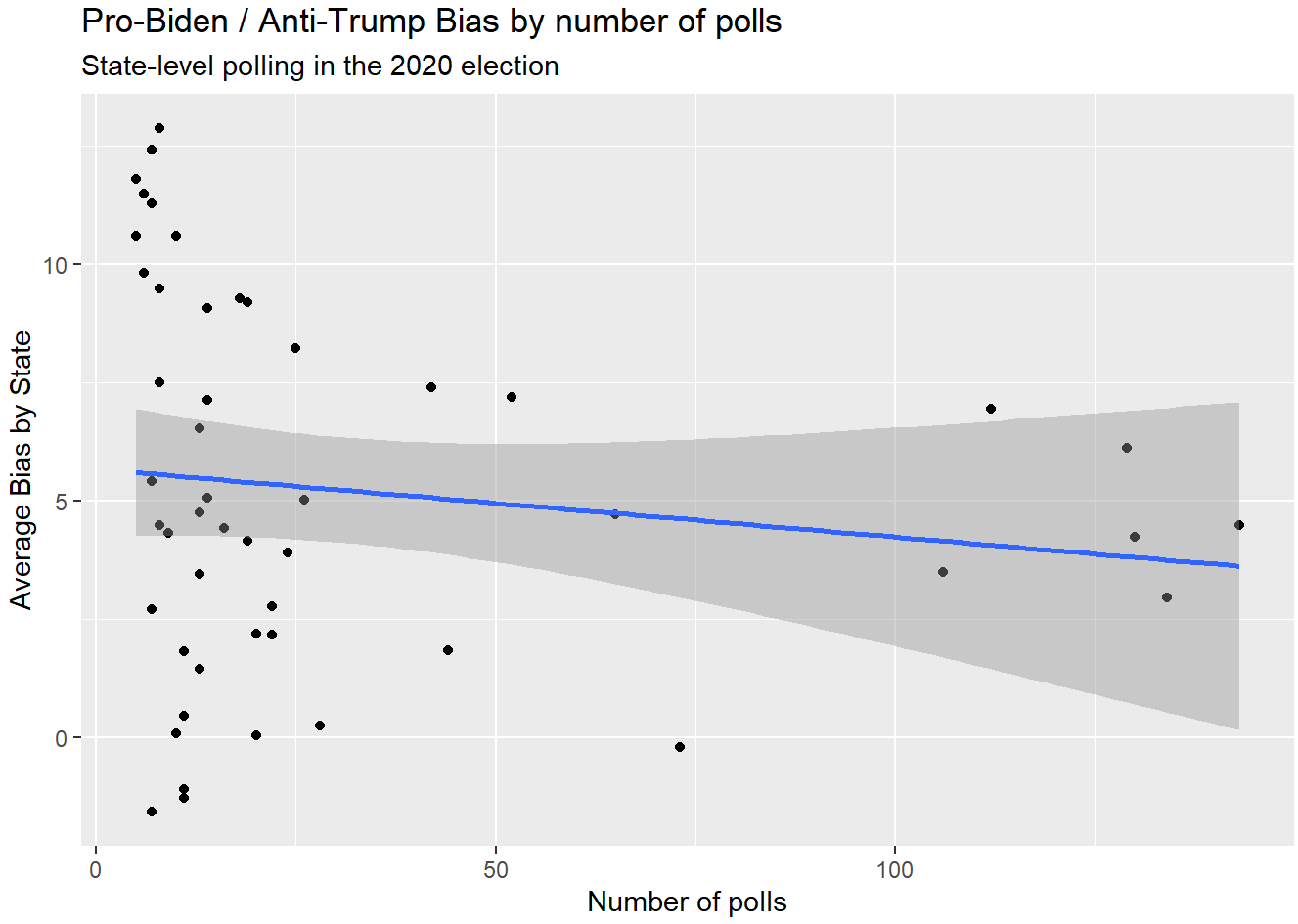
- I find the same overall pattern in the state data. First, there is evidence that the polls become more accurate the closer we get to the election. Second, there remains evidence of an anti-Trump bias in the state-level polls. [RUBRIC: Identical rubric to Q3.]

Question 5 [1 point]

One other explanation for inaccurate state polls is that some states do not have many polls run. Calculate the anti-Trump/pro-Biden bias for each state by subtracting the `repErr` *from the* `demErr` *(call this new variable* `bidenBias` *). Then calculate the average bias by state AND calculate the number of polls in that state. Finally, plot the relationship between the number of polls and the extent of bias. Does the data support the theory that states with more polls were predicted more accurately?*

```
state %>%
  mutate(bidenBias = demErr - repErr) %>%
  group_by(StateName) %>%
  summarise(bidenBias = mean(bidenBias,na.rm=T),
            nPolls = n()) %>%
  ungroup() %>%
  ggplot(aes(x = nPolls,
             y = bidenBias)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  labs(x = 'Number of polls',
       y = 'Average Bias by State',
       title = 'Pro-Biden / Anti-Trump Bias by number of polls',
       subtitle = 'State-level polling in the 2020 election')
```

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



[RUBRIC: 0.75 points for correct plot. 0.5 points for mistakes in calculating the number of polls and the average `bidenBias` by state using `group_by()` and `summarise()`. 0.25 points for no labels. 0 points for no attempt.]

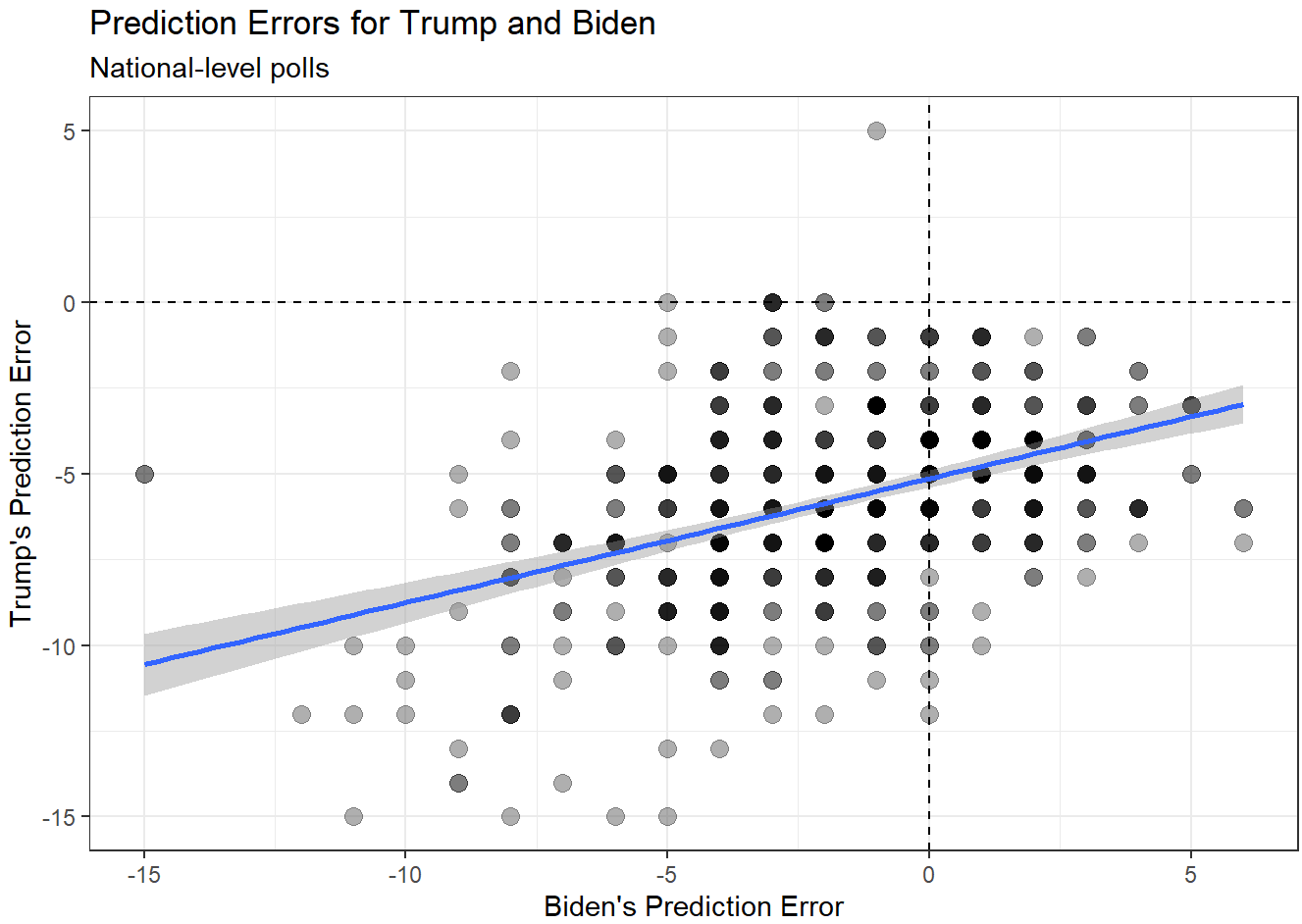
- Yes the plot supports the theory because the straight line of best fit is negative, tending toward zero. This means that states with more polls had smaller prediction errors, which is what the hypothesis predicted. However, the data is very noisy, meaning that this relationship is fairly weak. [RUBRIC:]

Extra Credit [1 point]

Do polls that underestimate Trump's support overestimate Biden's support? Investigate this question using both the national data (`pres`) and the state data (`state`). Use a scatterplot to test, combined with a (straight) line of best fit. Then, calculate the proportion of polls that (1) underestimate both Trump and Biden, (2) underestimate Trump and overestimate Biden, (3) overestimate Trump and underestimate Biden, (4) overestimate both candidates. In these analyses, define "overestimate" as prediction errors greater than or equal to zero, whereas "underestimate" should be prediction errors less than zero. What do you conclude? Is there any evidence of an anti-Trump bias in national polling? What about state polling?

```
# National scatterplot
pres %>%
  ggplot(aes(x = demErr,y = repErr)) +
  geom_point(size = 3,alpha = .3) +
  geom_smooth(method = 'lm') +
  labs(title = "Prediction Errors for Trump and Biden",
       subtitle = 'National-level polls',
       x = "Biden's Prediction Error",
       y = "Trump's Prediction Error") +
  geom_vline(xintercept = 0,linetype = 'dashed') +
  geom_hline(yintercept = 0,linetype = 'dashed') +
  theme_bw()
```

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



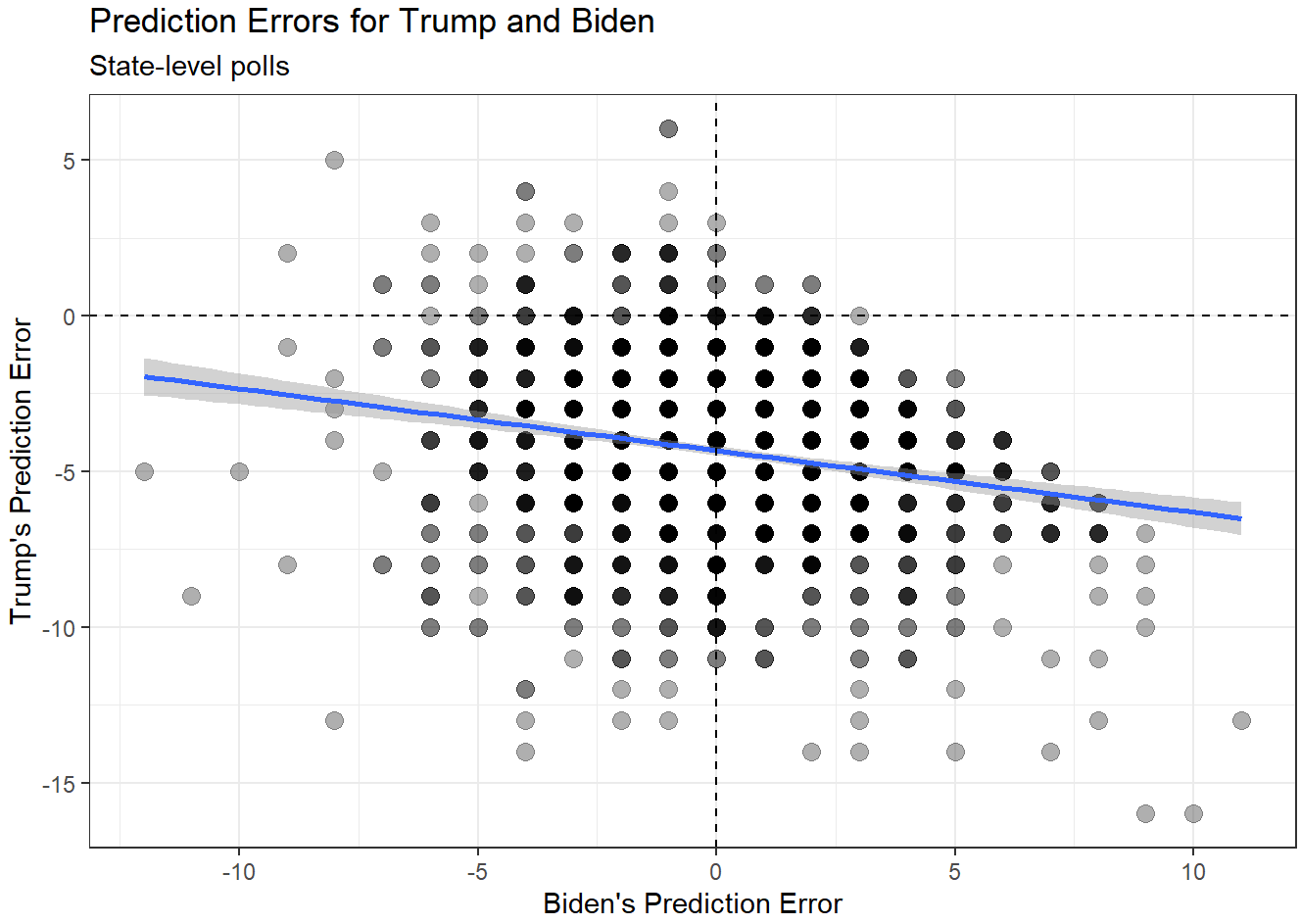

```
pres %>%
  summarise(UNboth = mean(demErr < 0 & repErr < 0),
            UNTrOVBi = mean(demErr >= 0 & repErr < 0),
            OVTrUNBi = mean(demErr < 0 & repErr >= 0),
            OVboth = mean(demErr >= 0 & repErr >= 0))
```

```
## # A tibble: 1 × 4
##   UNboth UNTrOVBi OVTrUNBi OVboth
##    <dbl>    <dbl>    <dbl>  <dbl>
## 1  0.595    0.388   0.0170      0
```

```
# State level
state <- state %>%
  mutate(demErr = Biden - BidenCertVote,
         repErr = Trump - TrumpCertVote)

state %>%
  ggplot(aes(x = demErr,y = repErr)) +
  geom_point(size = 3,alpha = .3) +
  geom_smooth(method = 'lm') +
  labs(title = "Prediction Errors for Trump and Biden",
       subtitle = 'State-level polls',
       x = "Biden's Prediction Error",
       y = "Trump's Prediction Error") +
  geom_vline(xintercept = 0,linetype = 'dashed') +
  geom_hline(yintercept = 0,linetype = 'dashed') +
  theme_bw()
```

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



```
state %>%
  summarise(UNboth = mean(demErr < 0 & repErr < 0),
            UNTrOVBi = mean(demErr >= 0 & repErr < 0),
            OVTrUNBi = mean(demErr < 0 & repErr >= 0),
            OVboth = mean(demErr >= 0 & repErr >= 0))
```

```
## # A tibble: 1 × 4
##   UNboth UNTrOVBi OVTrUNBi OVboth
##    <dbl>    <dbl>    <dbl>  <dbl>
## 1  0.342    0.590   0.0479 0.0207
```

The results from the national data show that the polls which underestimate Trump's support also underestimated Biden's support, as indicated by the positive slope. Almost 60% of the polls underpredicted both Biden and Trump, whereas 0% of the polls overpredicted both. Nevertheless, there is still some evidence of an anti-Trump bias, since 39% of polls underpredicted Trump and overpredicted Biden, while only 1.7% of polls overpredicted Trump and underpredicted Biden. In the state data, the evidence is even stronger. Here we find a negative relationship, meaning that polls which overestimated Biden's support ALSO underestimated Trump's support. Furthermore, 59% of polls overpredicted Biden and underpredicted Trump, compared to only 5% of polls which overpredicted Trump and underpredicted Biden. In summary, I would conclude there is stronger evidence of an anti-Trump bias in the state polls, compared to the national polls. [RUBRIC: 0.75 points for not using `alpha` or `size` or `geom_jitter` to account for multiple polls on the same points. 0.75 points for incorrectly calculating the four types of polls for both the state and national-level data. 0.5 points for missing labels. 0.25 points if written response did not acknowledge that the patterns found in the national data do not appear in the state-level data (i.e., the line of best fit is positive in the national data, but negative in the state data). 0 points if no attempt was made.]