More conditional relationships

Homework

Prof. Bisbee

Due Date: 2024-09-19

A new question

Suppose we were concerned with whether some polls might give different answers because of variation in who the poll is able to reach using that method. People who take polls via landline phones (do you even know what that is?) might differ from those who take surveys online. Or people contacted using randomly generated phone numbers (RDD) may differ from those contacted from a voter registration list that has had telephone numbers merged onto it.

Polls were done using lots of different methods in 2020.

Loading the data

```
require(tidyverse)
Pres2020.PV <- read_rds(file="https://github.com/jbisbee1/DS1000_F2024/raw/main/data/Pre
s2020_PV.Rds")
Pres2020.PV <- Pres2020.PV %>%
                mutate(Trump = Trump/100,
                      Biden = Biden/100,
                      margin = Biden - Trump)
```

```
Pres2020.PV %>%
  count(Mode)
```

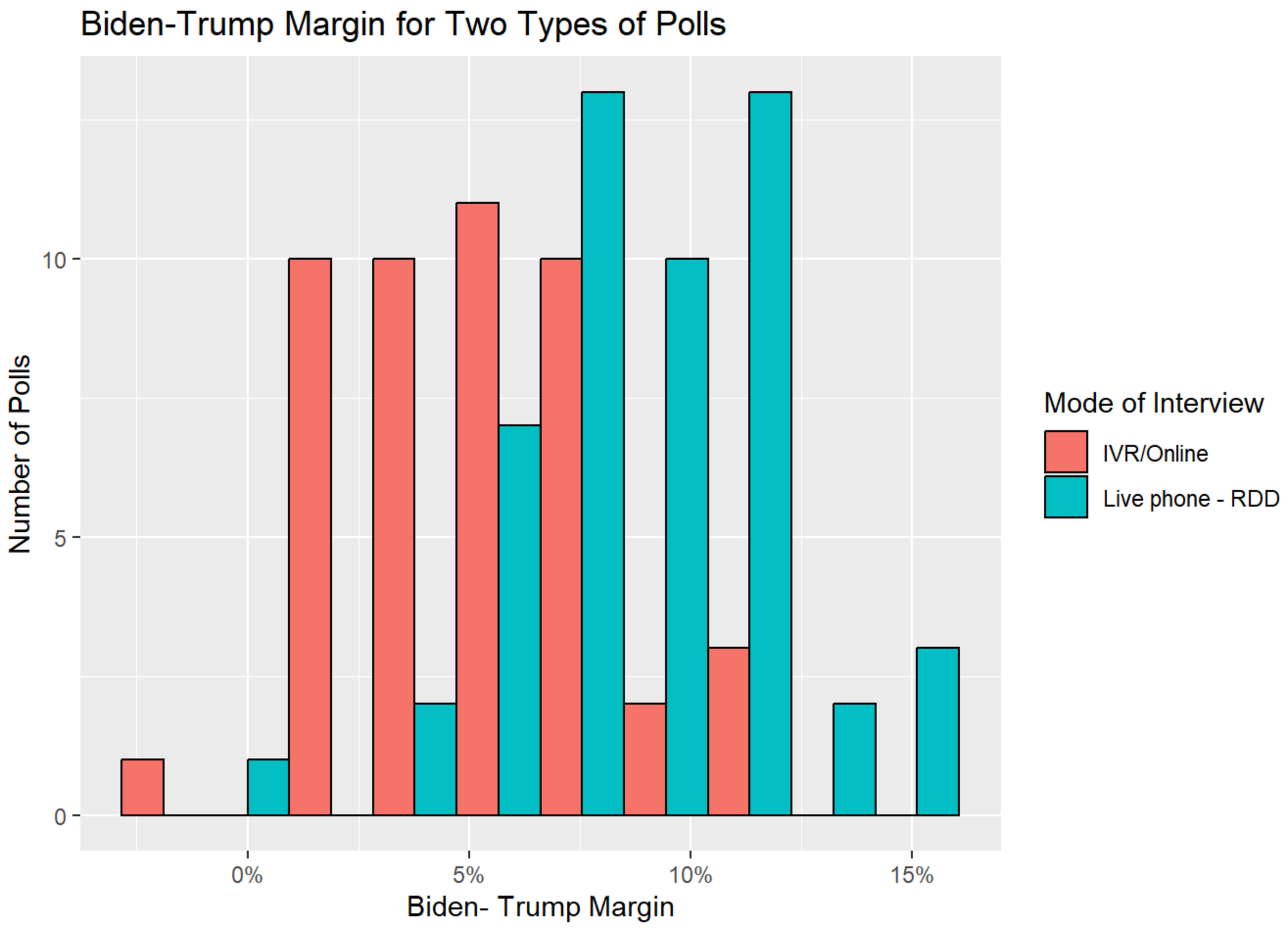
```
## # A tibble: 9 × 2
##   Mode                 n
##   <chr>            <int>
## 1 IVR                  1
## 2 IVR/Online          47
## 3 Live phone - RBS    13
## 4 Live phone - RDD    51
## 5 Online             366
## 6 Online/Text          1
## 7 Phone - unknown      1
## 8 Phone/Online        19
## 9 <NA>                29
```

This raises the question of – how do we visualization variation in a variable by another variable? More specifically, how can we visualize how the `margin` we get using one type of survey compares to the `margin` from another type of poll? (We cannot use a scatterplot because the data is from different observations (here polls).)

We could do this using earlier methods by `select` ing polls with a specific interview method (“mode”) and then plotting the `margin` (or `Trump` or `Biden`), but that will produce a bunch of separate plots that may be hard to directly compare. (In addition to having more things to look at we would want to make sure that the scale of the x-axis and y-axis are similar.)

We can plot another “layer” of data in ` `ggplot` using the `fill` paramter. Previously we used it to make the graphs look nice by choosing a particular color. But if we set `fill` to be a variable in our `tibble` then `ggplot` will plot the data seperately for each unique value in the named variable. So if we want to plot the histogram of `margin` for two types of polls we can use the `fill` argument in `ggplot` to tell R to produce different fills depending on the value of that variable.

```
Pres2020.PV %>%
  filter(Mode == "IVR/Online" | Mode == "Live phone - RDD") %>%
    ggplot(aes(x= margin, fill = Mode)) +
  labs(y = "Number of Polls",
         x = "Biden- Trump Margin",
         title = "Biden-Trump Margin for Two Types of Polls",
        fill = "Mode of Interview") +
    geom_histogram(bins=10, color="black", position="dodge") +
    scale_x_continuous(breaks=seq(-.1,.2,by=.05),
                     labels= scales::percent_format(accuracy = 1))
```



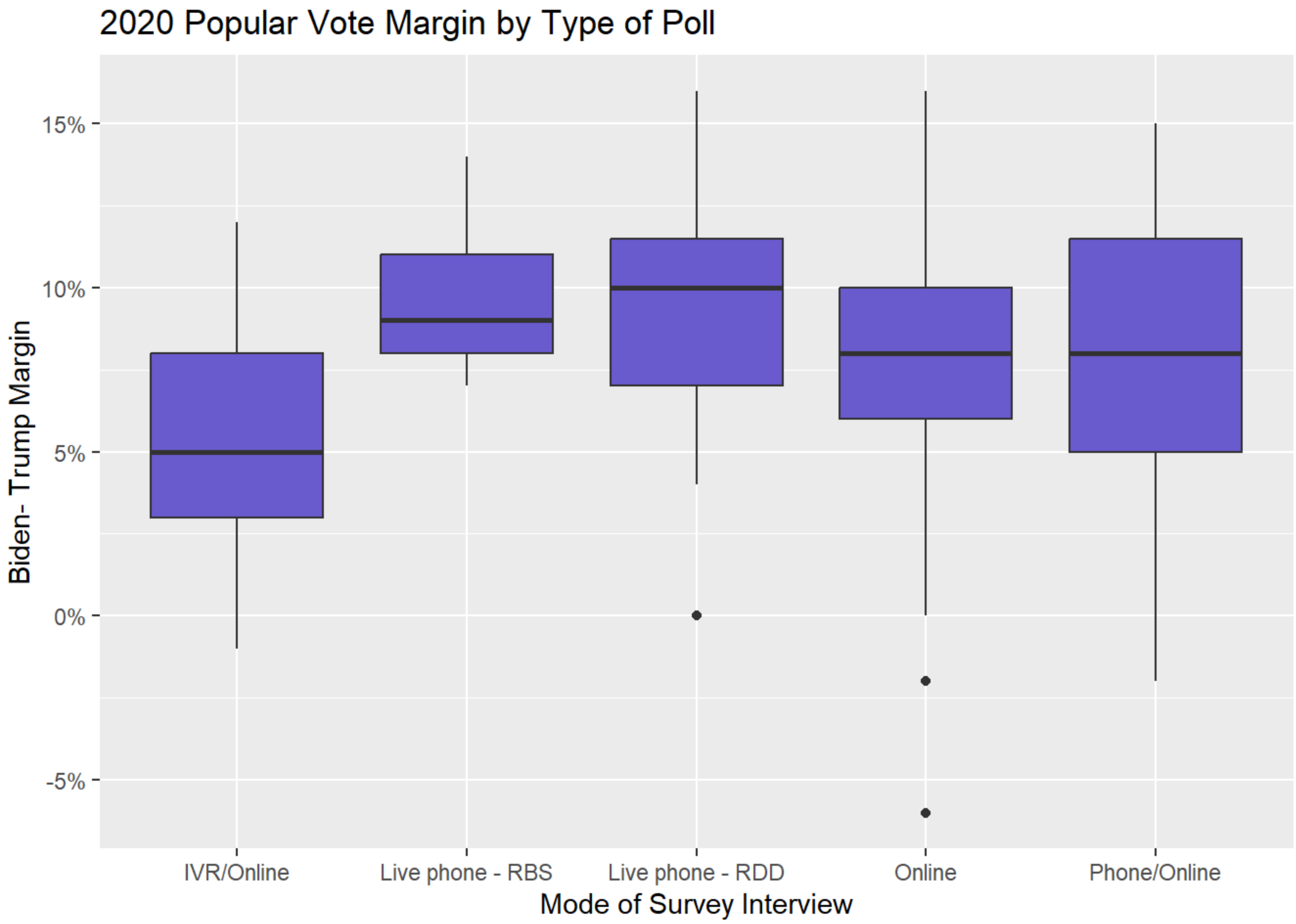
Quick Exercise Try running the code without the `filter` . What do you observe? How useful is this? Why or why not?

```
# INSERT CODE
```

While informative, it can be hard to compare the distribution of more than two categories using such methods. To compare the variation across more types of surveys we need to use a different visualization that summarizes the variation in the variable of interest a bit more. One common visualization is the `boxplot` which reports the mean, 25th percentile (i.e., the value of the data if we sort the data from lowest to highest and take the value of the observation that is 25% of the way through), the 75th percentile, the range of values, and notable outliers.

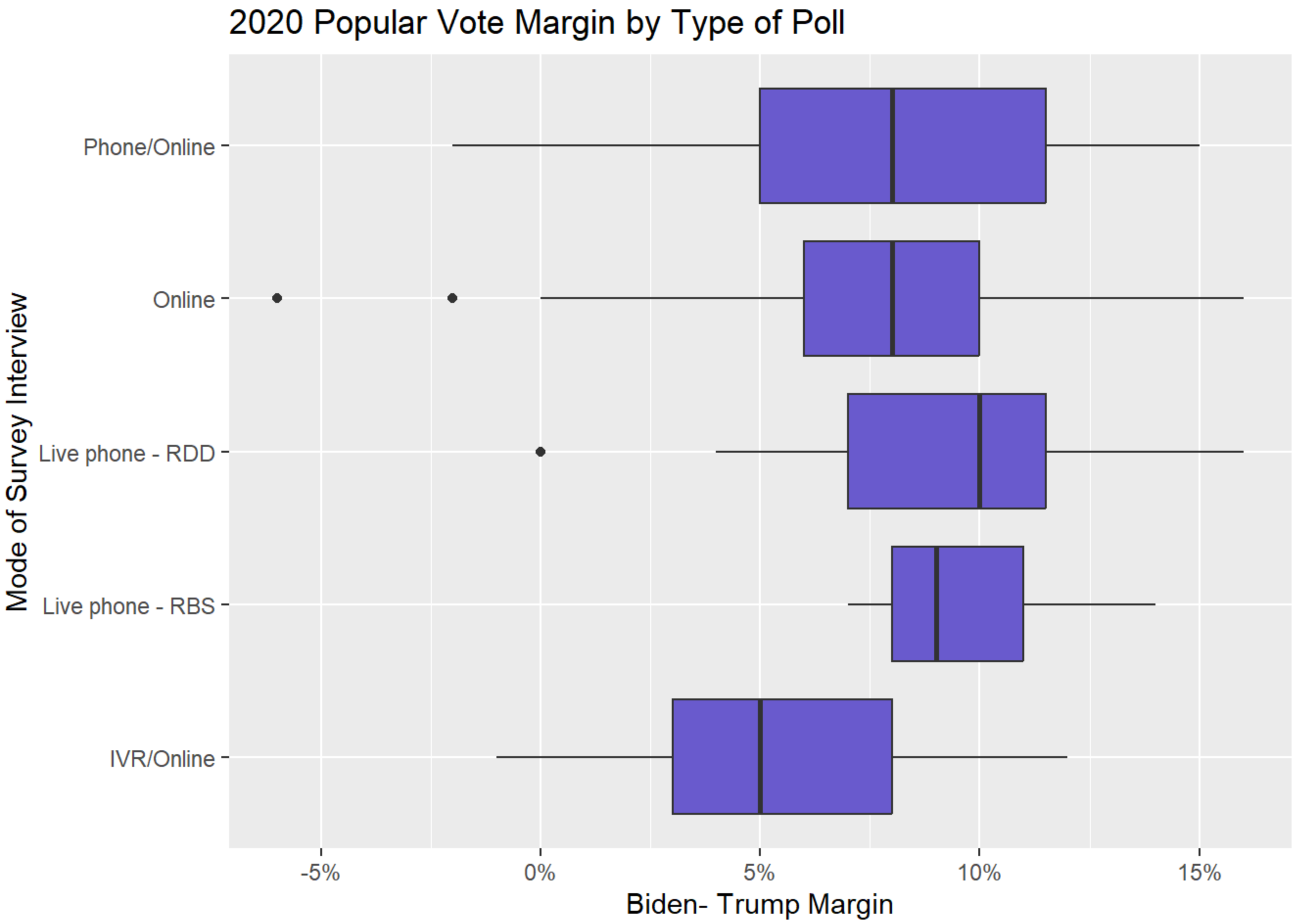
Let's see what the `boxplot` of survey mode looks like after we first drop surveys that were conducted using modes that were hardly used (or missing).

```
Pres2020.PV %>%
  filter(Mode != "IVR" & Mode != "Online/Text" & Mode != "Phone - unknown" & Mode != "N
A") %>%
  ggplot(aes(x = Mode, y = margin)) +
    labs(x = "Mode of Survey Interview",
         y = "Biden- Trump Margin",
         title = "2020 Popular Vote Margin by Type of Poll") +
    geom_boxplot(fill = "slateblue") +
    scale_y_continuous(breaks=seq(-.1,.2,by=.05),
                     labels= scales::percent_format(accuracy = 1))
```



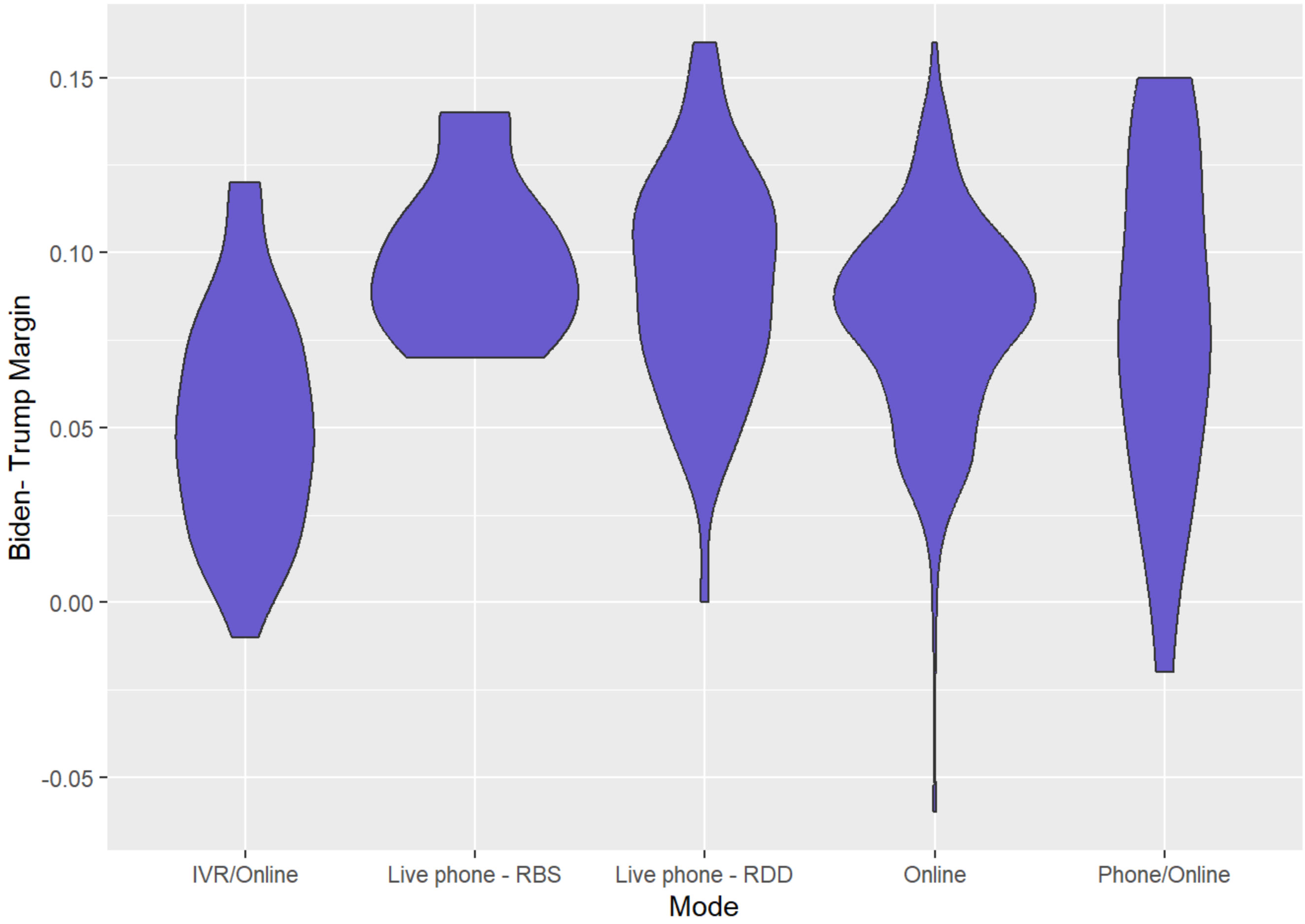
We can also flip the graph if we think it makes more sense to display it in a different orientation using `coord_flip`. (We could, of course, also redefine the x and y variables in the ` `ggplot` object, but it is useful to have a command to do this to help you determine which orientation is most useful).

```
Pres2020.PV %>%
  filter(Mode != "IVR" & Mode != "Online/Text" & Mode != "Phone - unknown" & Mode != "N
A") %>%
  ggplot(aes(x = Mode, y = margin)) +
    labs(x = "Mode of Survey Interview",
         y = "Biden- Trump Margin",
         title = "2020 Popular Vote Margin by Type of Poll") +
    geom_boxplot(fill = "slateblue") +
    scale_y_continuous(breaks=seq(-.1,.2,by=.05),
                     labels= scales::percent_format(accuracy = 1)) +
    coord_flip()
```



A downside of the boxplot is that it can be hard to tell how the data varies within each box. Is it equally spread out? How much data are contained in the lines (which are simply 1.5 times the height of the box)? To get a better handle on this we can use a “violin” plot that dispenses with a standard box and instead tries to plot the distribution of data within each category.

```
Pres2020.PV %>%
  filter(Mode != "IVR" & Mode != "Online/Text" & Mode != "Phone - unknown" & Mode != "NA") %>%
  ggplot(aes(x=Mode, y=margin)) +
    xlab("Mode") +
    ylab("Biden- Trump Margin") +
    geom_violin(fill="slateblue")
```



It is also hard to know **how much** data is being plotted. If some modes have 1000 polls and others have only 5 that seems relevant.

Quick Exercise We have looked at the difference in `margin` . How about differences in the percent who report supporting `Biden` and `Trump` ? What do you observe. Does this suggest that the different ways of contacting respondents may matter in terms of who responds? Is there something else that may explain the differences (i.e., what are we assuming when making this comparison)?

```
# INSERT CODE HERE
```

Quick Exercise Some claims have been made that polls that used multiple ways of contacting respondents were better than polls that used just one. Can you evaluate whether there were differences in so-called “mixed-mode” surveys compared to single-mode surveys? (This requires you to define a new variable based on `Mode` indicating whether survey is mixed-mode or not.)

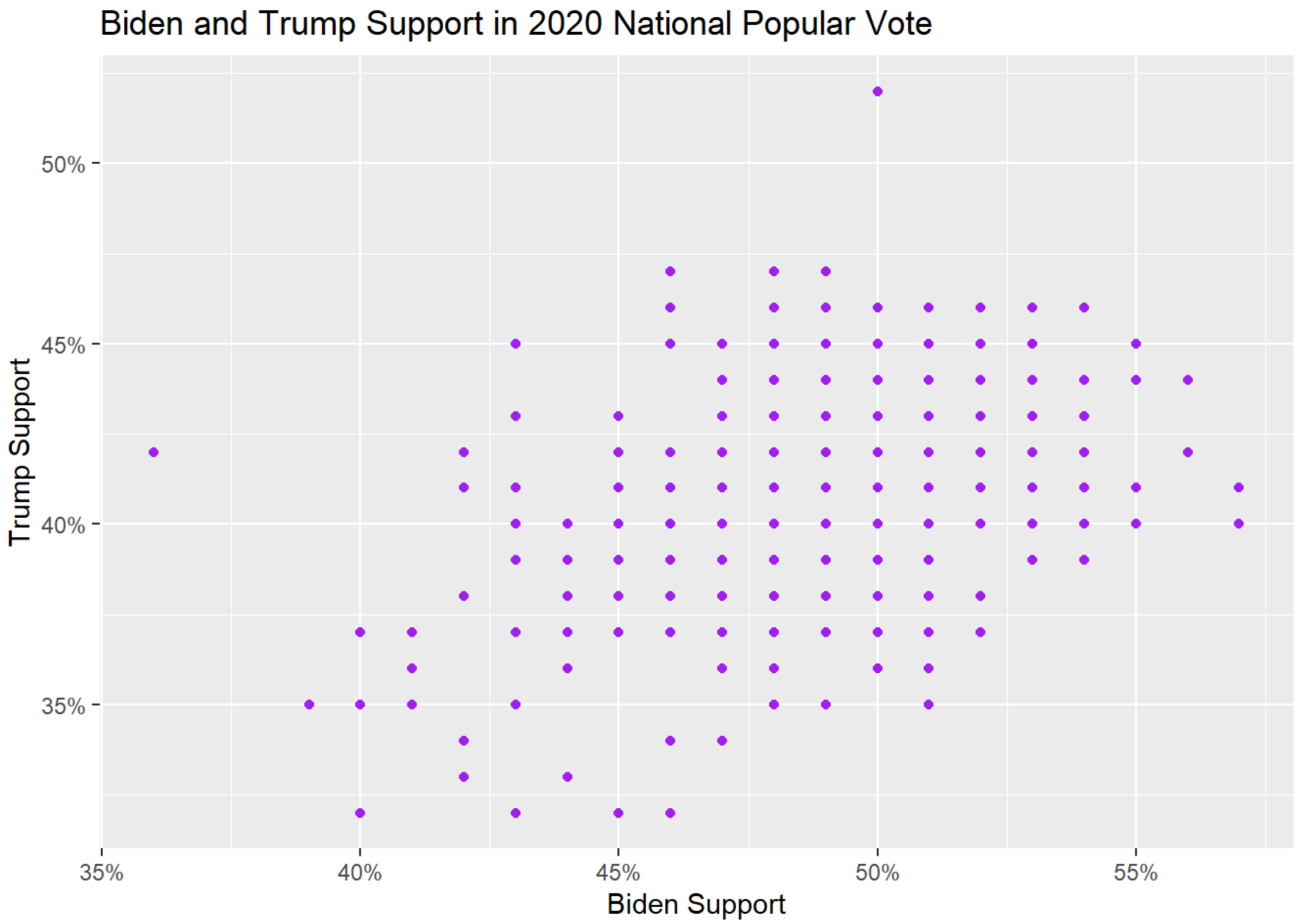
```
# INSERT CODE HERE
```

Continuous Variable By Continuous Variable (Scatterplot)

When we have two continuous variables we use a scatterplot to visualize the relationship. A scatterplot is simply a graph of every point in (x,y) where x is the value associated with the x-variable and y is the value associated with the y-variable. For example, we may want to see how support for Trump and Biden within a poll varies. So each observation is a poll of the national popular vote and we are going to plot the percentage of respondents in each poll supporting Biden against the percentage who support Trump.

To include two variables we are going to change our aesthetic to define both an x variable and a y variable – here `aes(x = Biden, y = Trump)` and we are going to label and scale the axes appropriately.

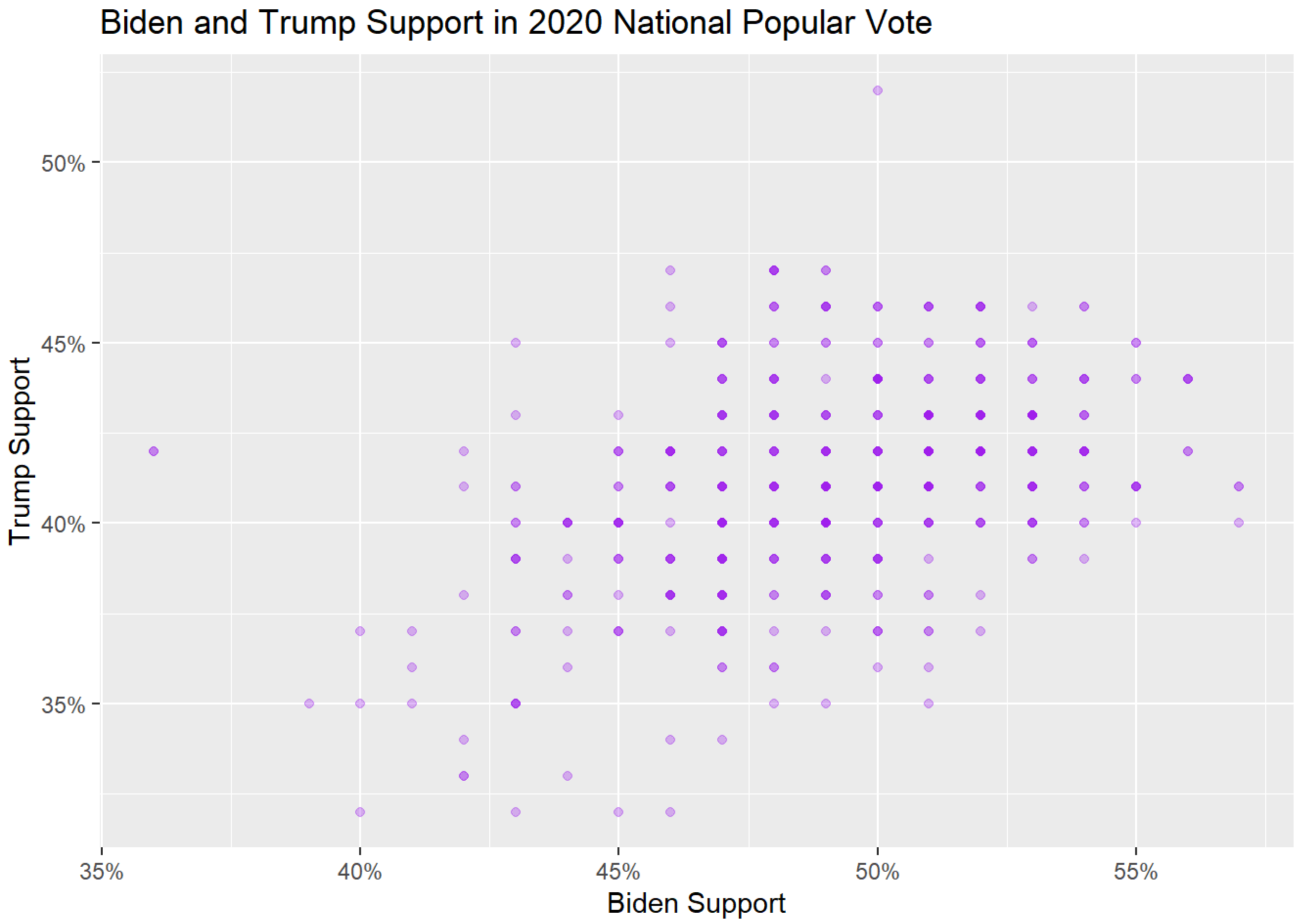
```
Pres2020.PV %>%
  ggplot(aes(x = Biden, y = Trump)) +
  labs(title="Biden and Trump Support in 2020 National Popular Vote",
       y = "Trump Support",
       x = "Biden Support") +
  geom_point(color="purple") +
    scale_y_continuous(breaks=seq(0,1,by=.05),
                     labels= scales::percent_format(accuracy = 1)) +
  scale_x_continuous(breaks=seq(0,1,by=.05),
                     labels= scales::percent_format(accuracy = 1))
```



The results are intriguing! First the data seems like it falls along a grid. This is because of how poll results are reported in terms of percentage points and it highlights that even continuous variables may be reported in discrete values. This is consequential because it is hard to know how many polls are associated with each point on the graph. How many polls are at the point (Biden 50%, Trump 45%)? This matters for trying to determine what the relationship might be. Second, it is clear that there are some questions that need to be asked – why doesn't `Biden + Trump = 100\%` ?

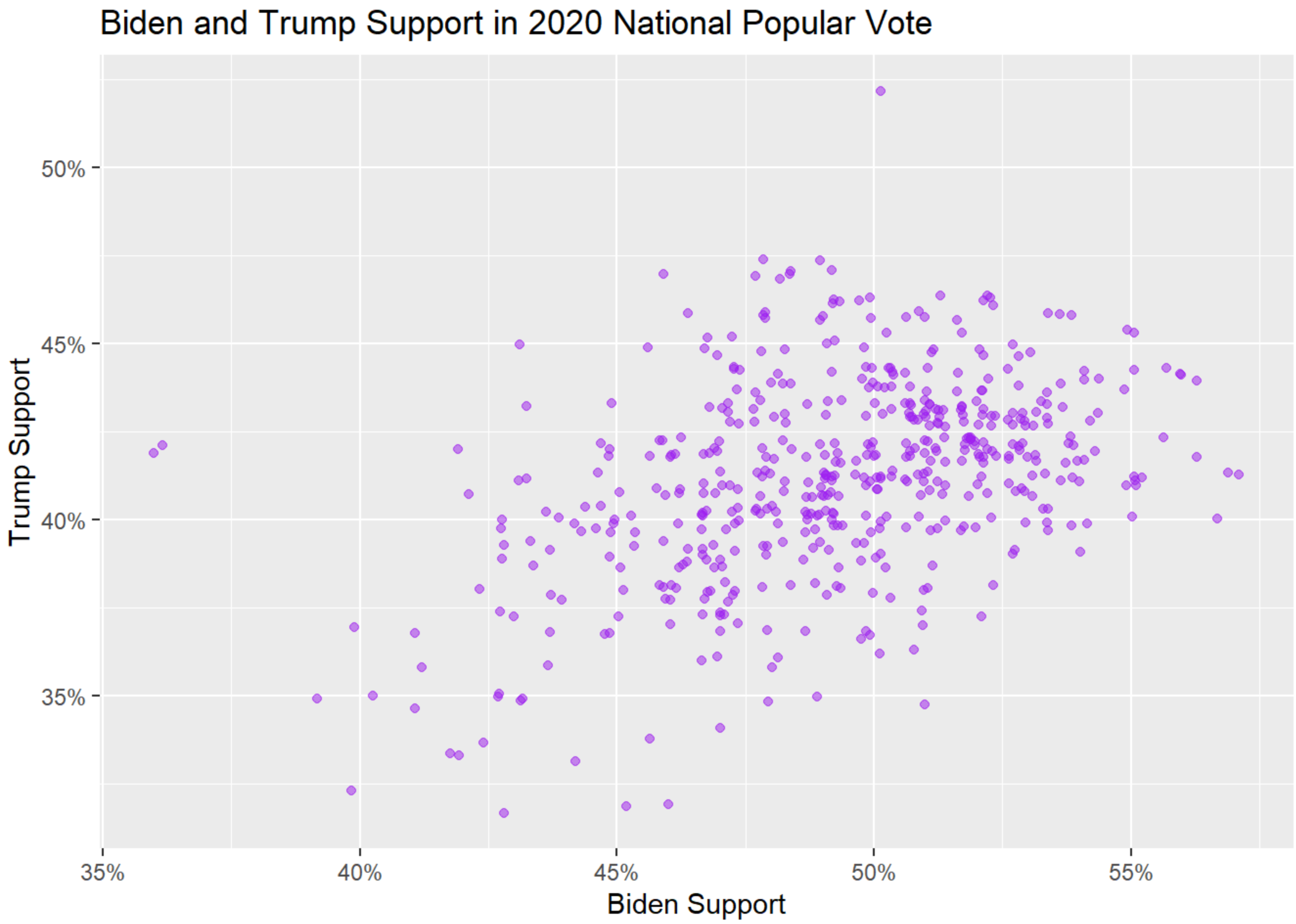
To try to display how many observations are located at each point we have two tools at our disposal. First, we can alter the "alpha transparency" by setting `alpha-.5` in the `geom_point` call. By setting a low level of transparency, this means that the point will become less transparent as more points occur at the same coordinate. Thus, a faint point indicates that only a single poll (observation) is located at a coordinate whereas a solid point indicates that there are many polls. When we apply this to the scatterplot you can immediately see that most of the polls are located in the neighborhood of Biden 50%, Trump 42%.

```
Pres2020.PV %>%
  ggplot(aes(x = Biden, y = Trump)) +
  labs(title="Biden and Trump Support in 2020 National Popular Vote",
       y = "Trump Support",
       x = "Biden Support") +
  geom_point(color="purple",alpha = .3) +
    scale_y_continuous(breaks=seq(0,1,by=.05),
                       labels= scales::percent_format(accuracy = 1)) +
  scale_x_continuous(breaks=seq(0,1,by=.05),
                       labels= scales::percent_format(accuracy = 1))
```



However, the grid-like nature of the plot is still somewhat hard to interpret as it can be hard to discern variations in color gradient. Another tool is to add a tiny bit of randomness to the x and y values associated with each plot. Instead of values being constrained to vary by a full percentage point, for example, the jitter allows it to vary by less. To do so we replace `geom_point` with `geom_jitter`.

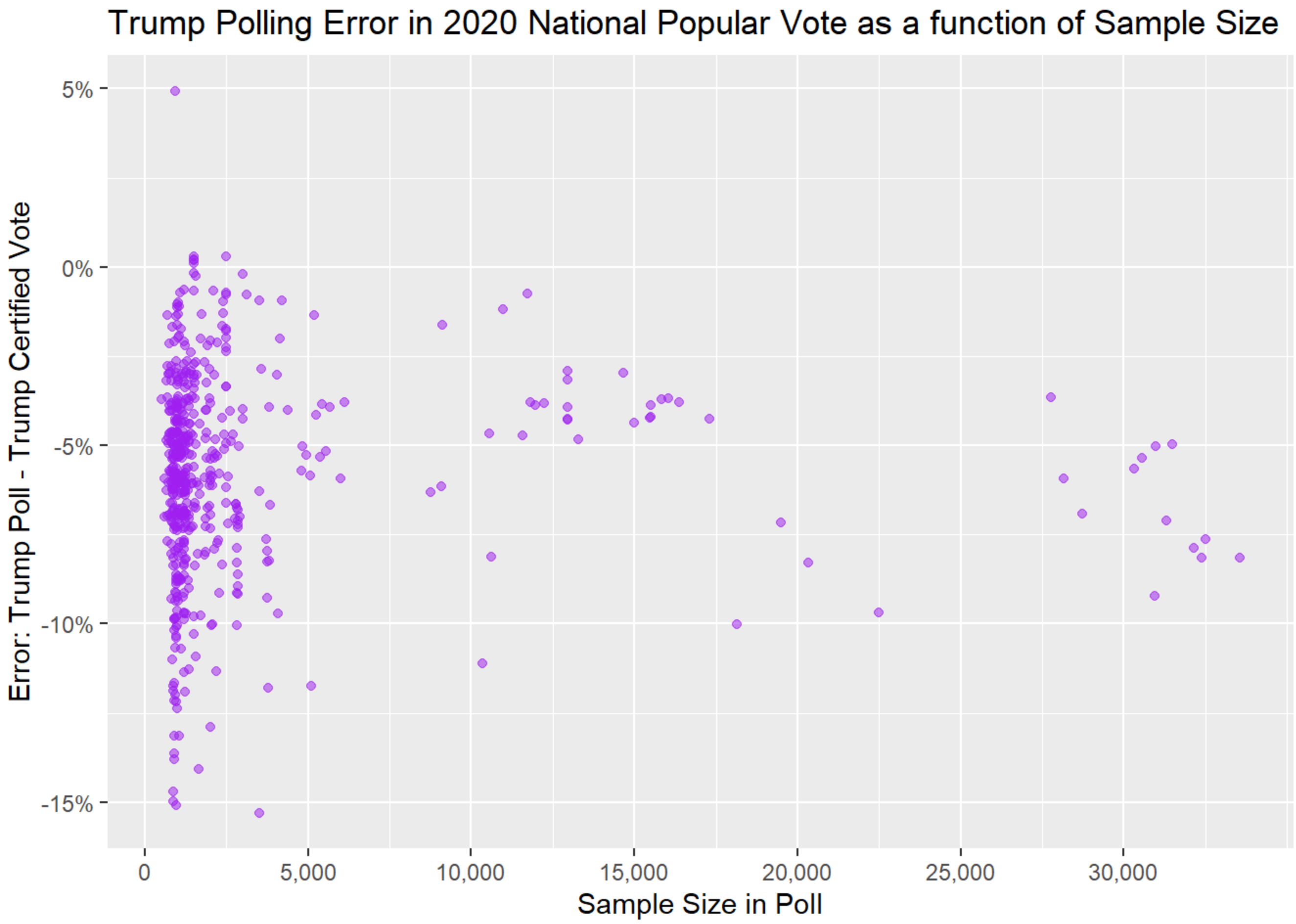
```
Pres2020.PV %>%
  ggplot(aes(x = Biden, y = Trump)) +
  labs(title="Biden and Trump Support in 2020 National Popular Vote",
       y = "Trump Support",
       x = "Biden Support") +
  geom_jitter(color="purple",alpha = .5) +
    scale_y_continuous(breaks=seq(0,1,by=.05),
                     labels= scales::percent_format(accuracy = 1)) +
  scale_x_continuous(breaks=seq(0,1,by=.05),
                     labels= scales::percent_format(accuracy = 1))
```



Note how much the visualization changes. Whereas before the eye was focused on – and arguably distracted by – the grid-like orientation imposed by the measurement, once we jitter the points we are immediately made aware of the relationship between the two variables. While we are indeed slightly changing our data by adding random noise, the payoff is that the visualization arguably better highlights the nature of the relationship. Insofar the goal of visualization is communication, this trade-off seems worthwhile in this instance. But here again is where data science is sometimes art as much as science. The decision of which visualization to use depends on what you think most effectively communicates the nature of the relationship to the reader.

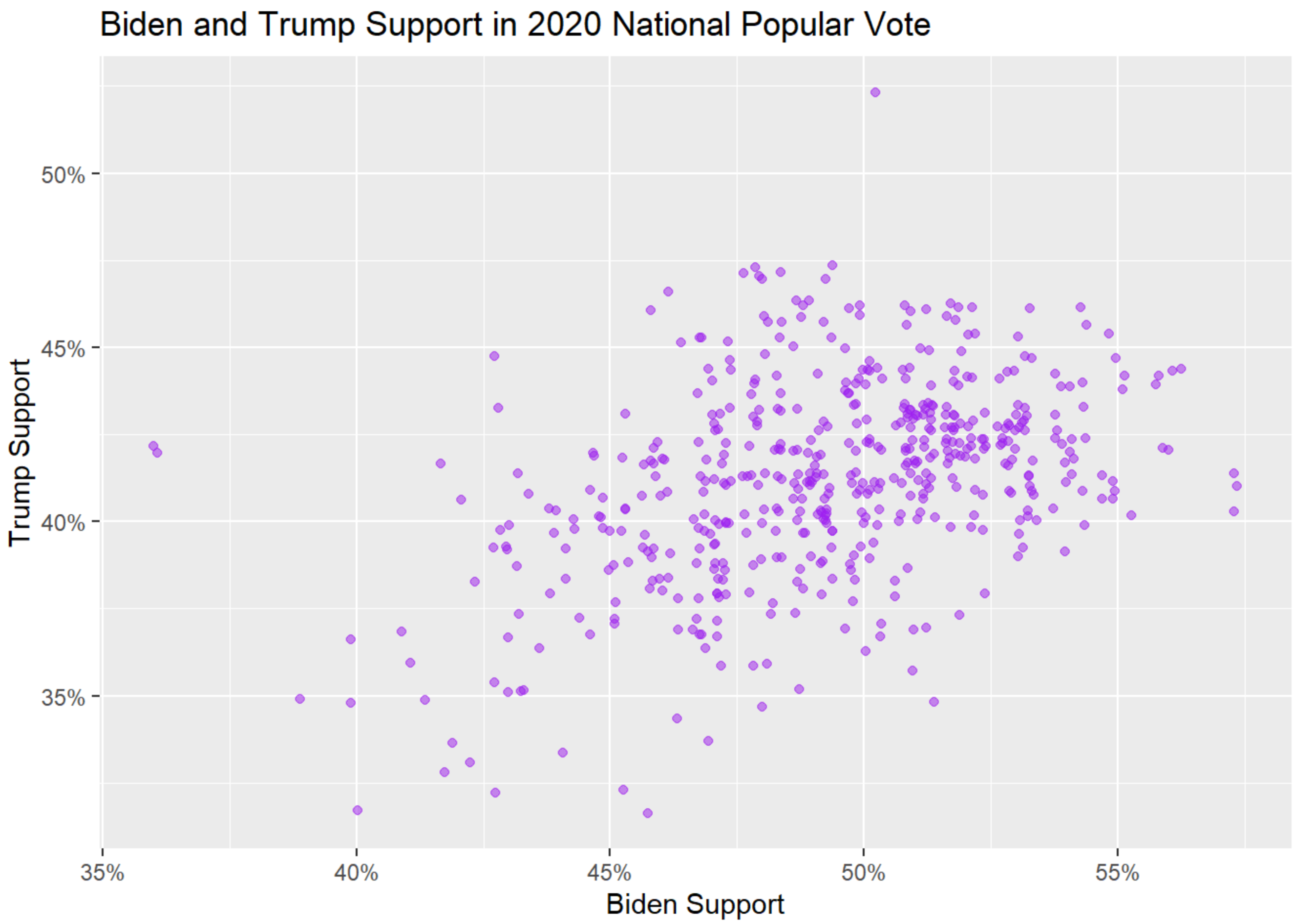
We can also look at the accuracy of a poll as a function of the sample size. This is also a relationship between two continuous variables – hence a scatterplot! Are polls with more respondents more accurate? There is one poll with nearly 80,000 respondents that we will filter out to me able to show a reasonable scale. Note that we are going to use `labels = scales::comma` when plotting the x-axis to report numbers with commas for readability.

```
Pres2020.PV %>%
  filter(SampleSize < 50000) %>%
  mutate(TrumpError = Trump - RepCertVote/100,
         BidenError = Biden - DemCertVote/100) %>%
  ggplot(aes(x = SampleSize, y = TrumpError)) +
  labs(title="Trump Polling Error in 2020 National Popular Vote as a function of Sample
Size",
       y = "Error: Trump Poll - Trump Certified Vote",
       x = "Sample Size in Poll") +
  geom_jitter(color="purple",alpha = .5) +
  scale_y_continuous(breaks=seq(-.2,1,by=.05),
                     labels= scales::percent_format(accuracy = 1)) +
  scale_x_continuous(breaks=seq(0,30000,by=5000),
                     labels= scales::comma)
```



In sum, we have tested Trump's theory that the MSM was biased against him. We found that polls that underpredicted Trump **also** underpredicted Biden. This is not what we would expect if the polls favored one candidate over another.

```
Pres2020.PV %>%
  ggplot(aes(x = Biden, y = Trump)) +
  labs(title="Biden and Trump Support in 2020 National Popular Vote",
       y = "Trump Support",
       x = "Biden Support") +
  geom_jitter(color="purple",alpha = .5) +
    scale_y_continuous(breaks=seq(0,1,by=.05),
                     labels= scales::percent_format(accuracy = 1)) +
  scale_x_continuous(breaks=seq(0,1,by=.05),
                     labels= scales::percent_format(accuracy = 1))
```



What is an alternative explanation for these patterns? Why would polls underpredict *both* Trump and Biden?

Perhaps they were fielded earlier in the year, when more people were interested in third party candidates, or hadn’t made up their mind. We’ll turn to testing this theory next time!