Blarry Wang
Rebecca U. Thorpe
November 21, 2019

Lab 2 Report

Datasets and How to Find Them

Question 4 - Independent & Dependent Variables

My research focuses on the the relationship between the degree of political polarization and the viciousness of gubernatorial candidates' tweets. The independent variable is the degree of party polarization collected by Shor and McCarthy (2018) and popular polarization measured by Pew Research Center. The dependent variable is the average score of all candidate's tweets measured on the viciousness scale varied by states.

Independent Variable - Party Polarization

The Shor and McCarthy data gives me the degree of the polarization in state senates and houses. I have the data in Stata 13 format, and I'm able to load the data in R. There are no more actions needed to clean this data.

Independent Variable - Popular Polarization

The Pew Research Center releases survey data that measures popular polarization. I have downloaded the survey data but I'm not sure how to reconstruct the degree of polarization just yet. I'm planning on figuring out how to do so before the quarter ends.

Dependent Variable - Viciousness of Tweets

I have a rudimentary viciousness indicator (VI) devised. The evaluation of the scale is yet to be underway. In terms of data, I need to scrape the tweets from gubernatorial elections in 2016. I have to 1) collect, 2) clean, and 3) evaluate each tweet against VI. Step 1) and Step 3) are essentially parallel. For this quarter, I will be doing test runs on getting tweets from candidates, and in the Winter quarter I will be collecting these tweets and cleaning them.

Question 5 - Control Variables

There are several control variables I will include in the study.

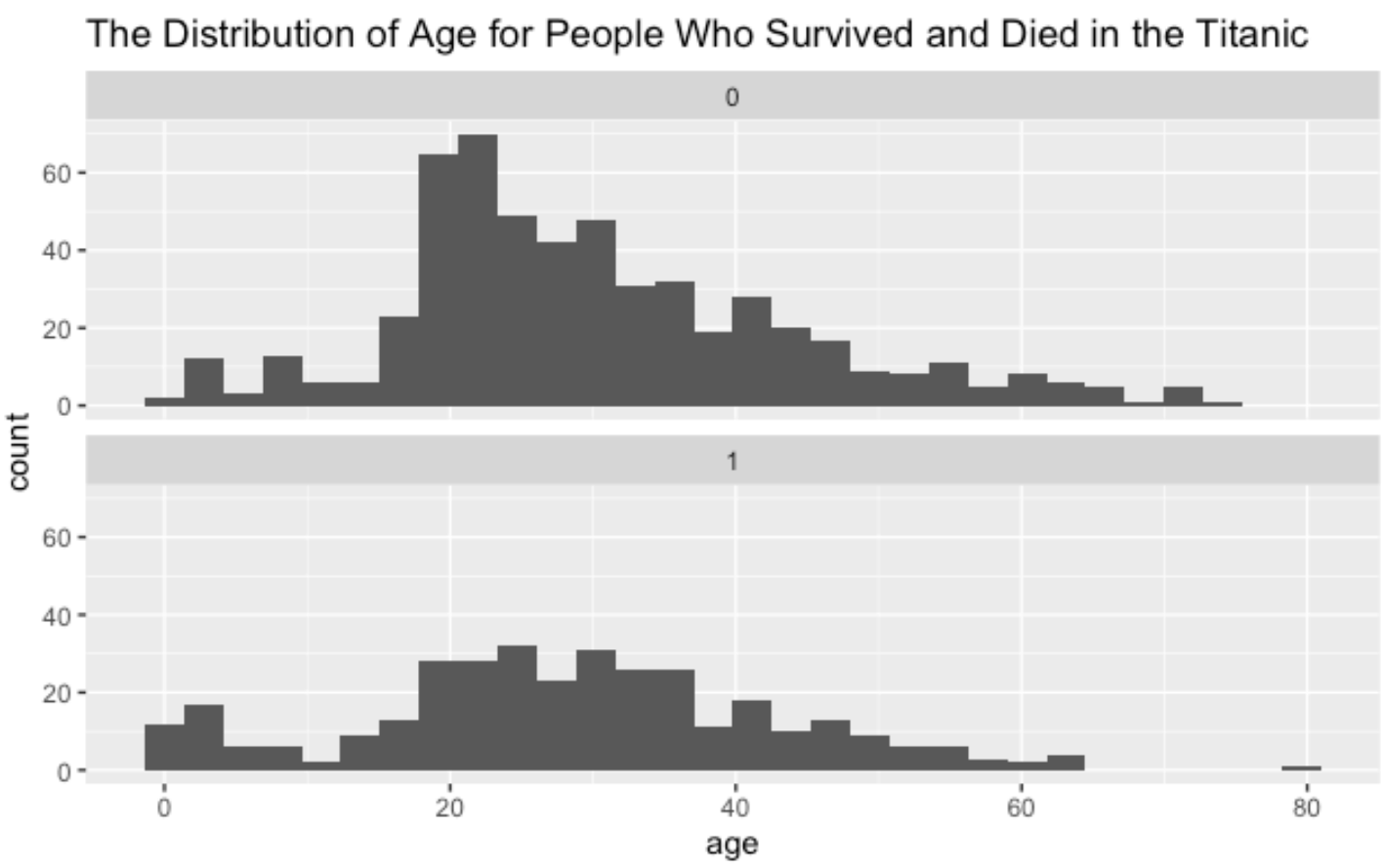
Control Variable	Description
state	the state which candidate is from
party	the party of candidate
incumbency	whether or not the candidate is incumbent
attacked	Existing literature points to a higher likelihood of attacking when attacked by their opponent. This variable shows whether the candidate has been attacked prior to the tweet.
competitiveness	Existing work shows that candidates are more likely to tweet in competitive races. This variable shows if the race is competitive.

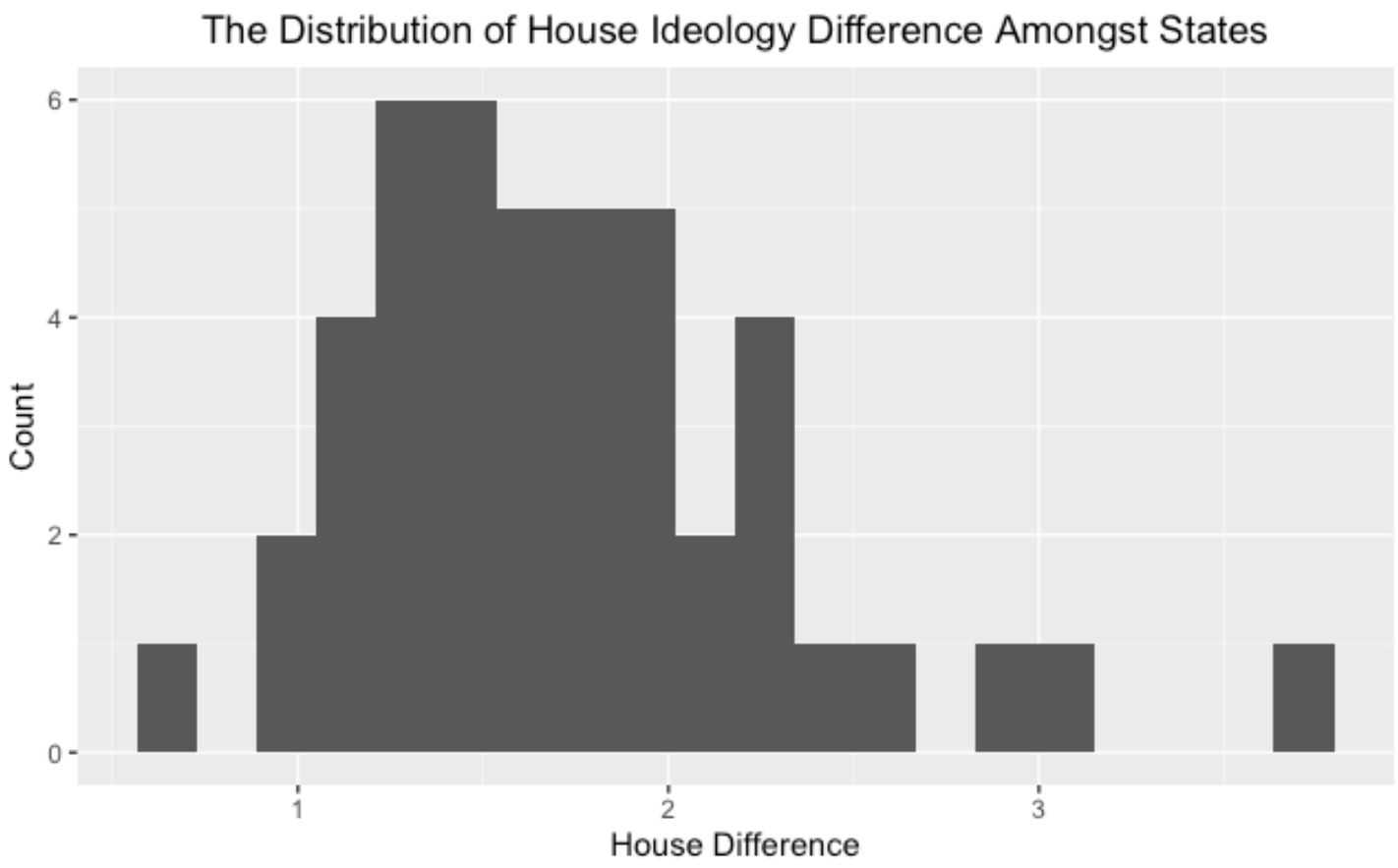
CAPPP Lab Homework 2

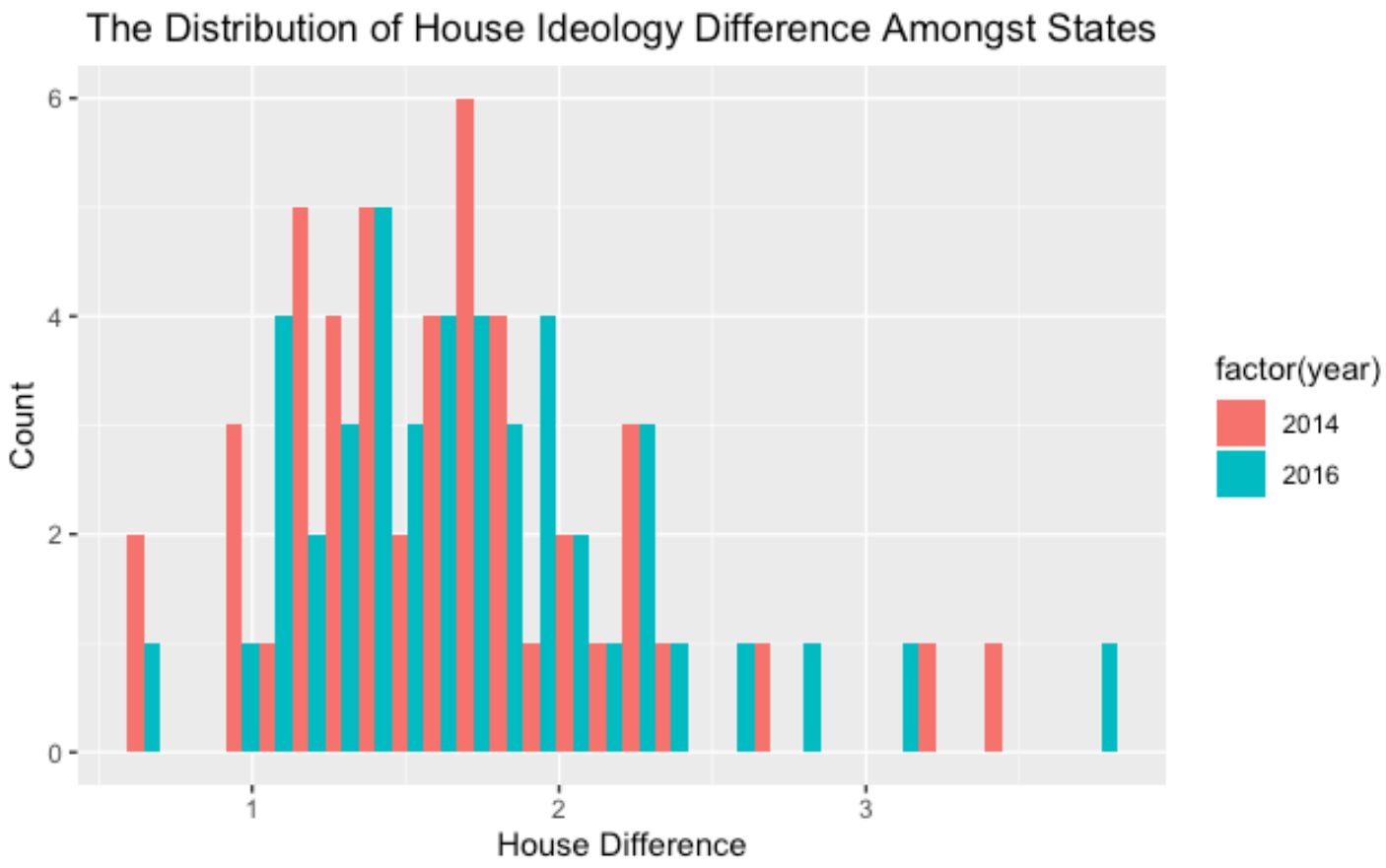
Collaboration with: Veronica, Kaley, Nikita, Batoul, Raina, and Maha

```r
library(dplyr)
library(tidyr)
library(ggplot2)
library(gapminder)

# Q1
data(gapminder)
data <- gapminder

data <- data %>% mutate(
  "econ" = rep_len(c("IMF_loan", "WTO_member"), nrow(data)),
  "econ_value" = rep_len(c(0, 1, 1), nrow(data)))

data <- data %>% spread(econ, econ_value)
data <- data %>% select(-c(WTO_member, IMF_loan))


#Q2
titanic <- read.csv("./Desktop/CAPPP/r_enviroment/excercises/ex4/lab5_titanic.csv", header = TRU
stats <- titanic %>% group_by(survived) %>%
  summarise(mean = mean(age), std = sd(age))
stats

titanic %>%
  ggplot() +
  geom_histogram(aes(age)) +
  facet_wrap(~survived, ncol = 1) +
  labs(title = "The Distribution of Age for People Who Survived and Died in the Titanic")

died <- titanic %>% filter(survived == 0)
lived <- titanic %>% filter(survived == 1)
diff <- mean(died$age) - mean(lived$age)

t.test(died$age, lived$age)

"
The p-value of there is a difference between the two age groups
is 0.07834. This fails to reject the null hypothesis, which means
there is likely no difference between the two means.

We are ~92% sure that we can reject the null.

Our confidence interval is (-0.1966269  3.6569075). The
margin of error is +/-1.9267672.

According to our data, there is probably not a relationship
between age and whether they will survive. There is 92% chance
that the fact we saw there is a difference is just a fluke.
"

#Q3
```

```r
library(readstata13)
state_pol <- read.dta13("./Desktop/CAPPP/data/polarization/Shor & McCarthy/shor_mc.sta")
state_pol %>%
  filter(year == 2016) %>%
  ggplot(aes(h_diffs)) +
  geom_histogram(bins = 20) +
  labs(
    x = "House Difference", y = "Count",
    title = "The Distribution of House Ideology Difference Amongst States") +
  theme(plot.title = element_text(hjust = 0.5))

state_pol %>%
  filter(year == 2016 | year == 2014) %>%
  ggplot() +
  geom_histogram(aes(h_diffs, fill = factor(year)), position = "dodge") +
  labs(
    x = "House Difference", y = "Count",
    title = "The Distribution of House Ideology Difference Amongst States") +
  theme(plot.title = element_text(hjust = 0.5))

```


The Distribution of Age for People Who Survived and Died in the Titanic
0
1
count
60
40
20
0
0
20
40
60
80
age


The Distribution of House Ideology Difference Amongst States
Count
6
4
2
0
1
2
3
House Difference


The Distribution of House Ideology Difference Amongst States
Count
0
2
4
6
1
2
3
House Difference
factor(year)
2014
2016