DATA 606 Project Proposal

Keith Colella

2023-04-06

```
library(tidyverse)
library(httr)
library(jsonlite)
```

Data Preparation

I'll require three datasets that I've gathered myself. First, I'll load in a list of all ~4200 candidates running for positions in the House of Representatives in the 2022 election. I gathered the core data from the Federal Election Commission API (https://api.open.fec.gov/developers/ (https://api.open.fec.gov/developers/)), but I then enhanced the dataset with data from Ballotpedia, Politwoops, and the @unitedstates project. Of these ~4200 candidates, I was able to pair ~1600 with Twitter handles.

```
candidates <- read_csv('data/candidates2022_clean.csv')
```

```
## New names:
## Rows: 4231 Columns: 17
## ── Column specification
## ──────────────────────────────────────────────────────── Delimiter: "," chr
## (13): fec_id, name, state, party, office, incumbent_challenge, candidate... dbl
## (3): ...1, district, twitter_id lgl (1): twitter_status_id
## ℹ Use `spec()` to retrieve the full column specification for this data. ℹ
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## • `` -> `...1`
```

Second, I'll load in just under a million tweets from the above 1600 candidates, representing all of their tweets from the year 2022. I scraped these tweets using the `snscrape` package in Python.

```
# load data
kaggle <- jsonlite::read_json('data/kaggle.json')
username <- kaggle$username
authkey <- kaggle$key

url <- 'https://www.kaggle.com/api/v1/datasets/download/kac624/politicaltweets/candidate_tweets2
022_04.06.csv'
response <- GET(url, authenticate(username, authkey, type = 'basic'))
temp <- tempfile()
download.file(response$url, temp, mode = 'wb')
tweets <- read_csv(unz(temp, 'candidate_tweets2022_04.06.csv'))
unlink(temp)
```

Finally, I'll load in data from the MIT Election Lab (https://electionlab.mit.edu/data (https://electionlab.mit.edu/data)) detailing district-level election results for historical house elections through 2020.

```
elections <- read_csv('data/1976-2020-house.csv')
```

```
## Rows: 31103 Columns: 20
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (8): state, state_po, office, district, stage, candidate, party, mode
## dbl (7): year, state_fips, state_cen, state_ic, candidatevotes, totalvotes, ...
## lgl (5): runoff, special, writein, unofficial, fusion_ticket
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Research question

You should phrase your research question in a way that matches up with the scope of inference your dataset allows for.

I want to examine the relationship between gerrymandering and polarization of political rhetoric. Specifically, I'd like to attempt to answer the following question: do less competitive districts typically have more ideologically extreme candidates? There is a prevailing narrative that gerrymandering tends to produce candidates closer to the political margins of their respective parties, but I'm not aware of strong empirical support for that narrative. While this study will be observation in nature, I'll attempt to evaluate whether such a relationship exists to a statistically significant degree.

I'll aim to analyze the data detailed in *Data Preparation* to extract two key insights for each candidate. First, I'll analyze the corpus of tweets produced by each candidate, and apply sentiment analysis with the goal of assigning each a "polarity score". This score will represent the degree to which they lean to the left or right of the US political spectrum. Second, I'll use historical election results to generate of measure of competitiveness for each congressional district. I'm still performing research to determine the best approach, but I'm most seriously considering using the "efficiency gap", which measures "wasted votes" (consisting of votes for a losing candidate + excess votes for a winning candidate) as a percentage of total votes.

With these two variables, I'll apply several statistical tests to examine their relationship. Specifically, I plan to apply a straightforward hypothesis test to assess whether the mean polarity score of candidates from less competitive districts is different from the mean polarity score fo candidates from more competitive districts. Depending on the results, I hope to fit a logistic regression model to assess the contribution of competitiveness to polarity scores.

I recognize this is a bit ambition, but I feel confident I can construct a cogent analysis on this topic. Of course, I look forward to hearing feedback!

Cases

What are the cases, and how many are there?

There are 1600 cases, each representing a 2022 Congressional election candidate. The primary variables of interest are polarity score and competitiveness score for the district in which the candidate ran in 2022.

Data collection

Describe the method of data collection.

I plan to use the data collection process as part of my final project for DATA607. Collection will be broken into several steps, each covered in a separate notebook.

1. Gather list of candidates from the FEC and identify each candidates twitter handle(s). Handles are gathered from Ballotpedia, Politwoops, and the @unitedstates project. Politwoops and @unitedstates project data have unique identifiers for each candidates (namely, the FEC and Bioguide IDs), but I'll need to rely on name-based fuzzy joins for mapping handles obtained from Ballotpedia.
2. Scrape tweets from Twitter. I'll use the `snscrape` package in Python.
3. Back in R, perform analysis of tweets to estimate polarity scores. Additionally, I'll perform analysis of historical election data to generate district-level competitiveness scores.

With this data, I'll be able to perform the statistical tests detailed above.

Type of study

What type of study is this (observational/experiment)?

This study will be observation. As such, conclusions will focus on correlation of the two variables, rather than a causal relationship.

Data Source

If you collected the data, state self-collected. If not, provide a citation/link.

Self-collected, though numerous pre-existing datasets were leverages. See the *Data Preparation* and *Data Collection* sections above for details.

Dependent Variable

What is the response variable? Is it quantitative or qualitative?

The target variable will be the "polarity score" obtained through analysis of candidates' tweets. While details are to be finalized, I expect the variable will an ordinal categorical variable, identifying a candidates level of ideological extremism of candidates on a scale from, say, 1 to 5. The exact details of the variable will be finalized as I lock down details on the analysis of candidate tweets (which I hope to do in the coming weeks).

Independent Variable(s)

The independent variable will be a measure of competitiveness for each district. Literature on gerrymandering frequently cites the "efficiency gap", which is a numerical percentage, calculated as $\frac{wastedvotes}{totalvotes}$.

Relevant summary statistics

Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

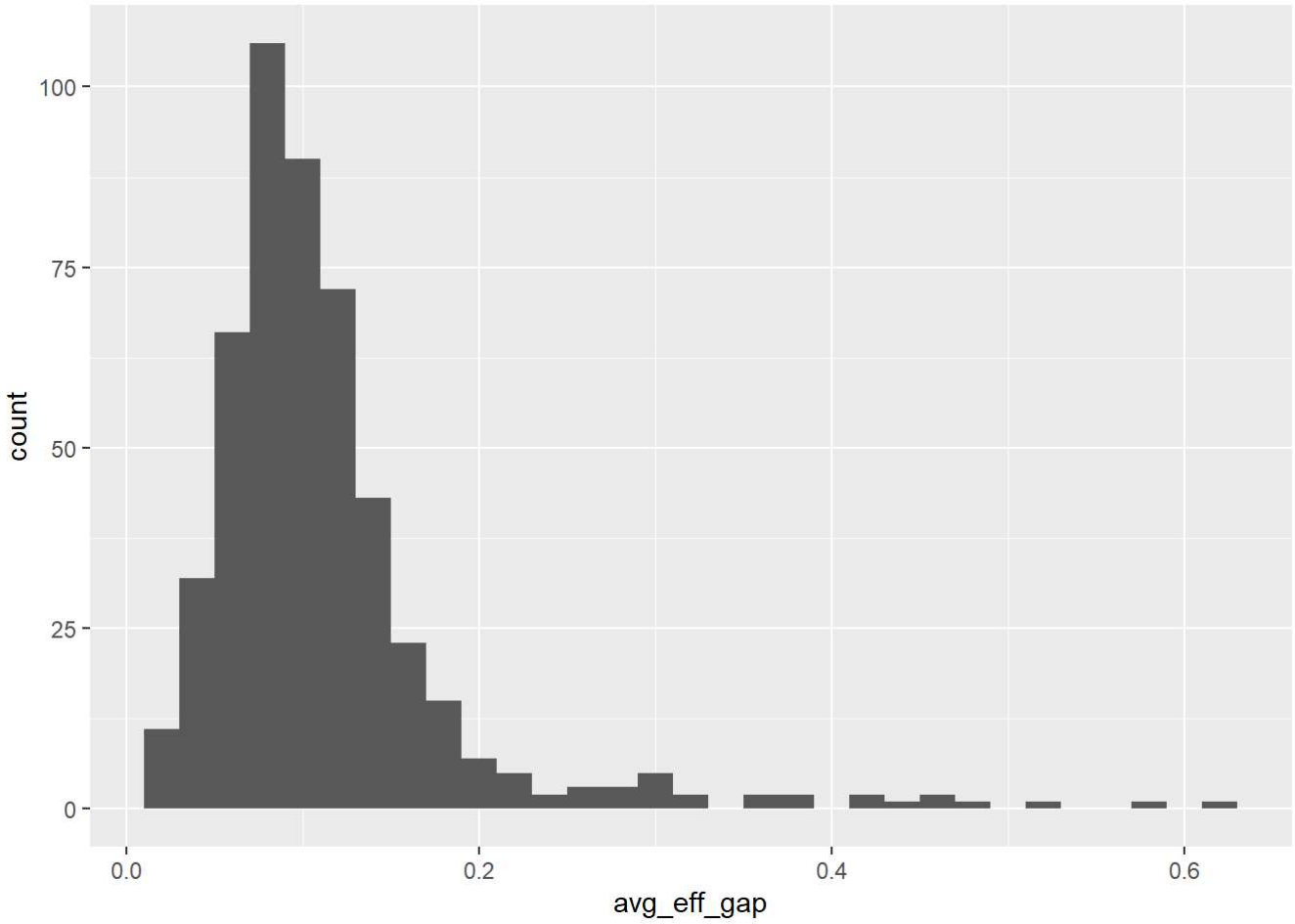
We can use the historical elections data to calculate the efficiency gap per district, based on the average over previous election years.

```
elections <- filter(elections, stage == 'GEN') %>%
  mutate(yearStateDist = paste0(year, state_po, district))

elections <- elections %>%
  group_by(yearStateDist) %>%
  mutate(result = case_when(candidatevotes == max(candidatevotes) ~ 'winner',
                            candidatevotes != max(candidatevotes) ~ 'loser')) %>%
  ungroup()

elections <- elections %>%
  mutate(wastedVotes = case_when(result == 'winner' ~ (candidatevotes - (totalvotes %/% 2)),
                                 result == 'loser' ~ candidatevotes),
         wastedVotes = case_when(party == 'DEMOCRAT' ~ -wastedVotes,
                                 party == 'REPUBLICAN' ~ wastedVotes,
                                 TRUE ~ 0))

eff_gap <- elections %>%
  group_by(year, state_po, district) %>%
  summarize(netWastedVotes = sum(wastedVotes),
            totalVotes = sum(totalvotes),
            eff_gap = abs(netWastedVotes / totalVotes),
            .groups = 'drop') %>%
  group_by(state_po, district) %>%
  summarize(avg_eff_gap = mean(eff_gap),
            .groups = 'keep')

summary(eff_gap)
```

```
##    state_po           district          avg_eff_gap
##  Length:499         Length:499         Min.   :0.01934
##  Class :character   Class :character   1st Qu.:0.07330
##  Mode  :character   Mode  :character   Median :0.09697
##                                        Mean   :0.11403
##                                        3rd Qu.:0.12927
##                                        Max.   :0.62512
##                                        NA's   :1
```

```
eff_gap %>%
  ggplot(aes(avg_eff_gap)) +
  geom_histogram(binwidth = 0.02)
```

```
## Warning: Removed 1 rows containing non-finite values (`stat_bin()`).
```

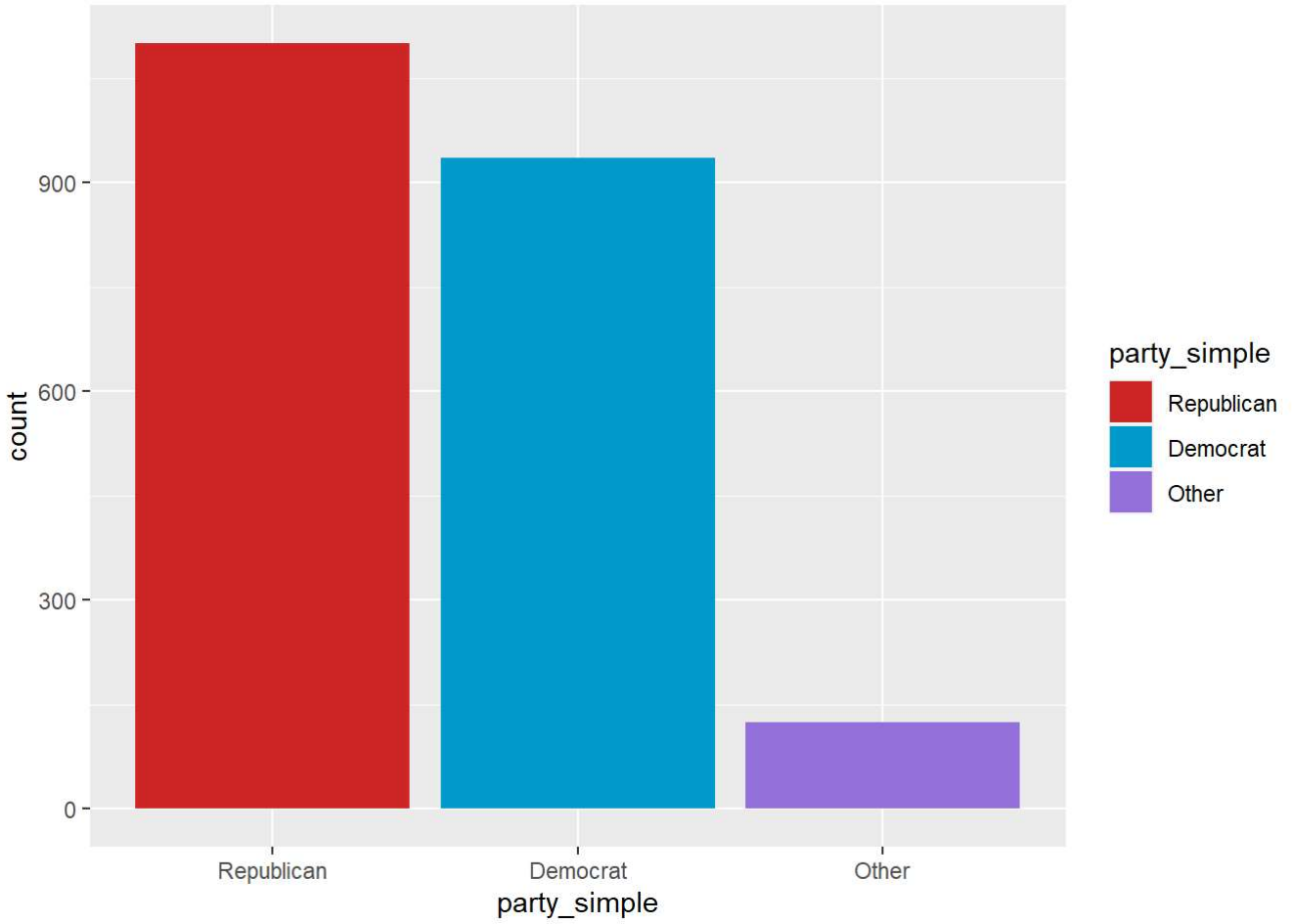


The above distribution contains significant outliers. Because the efficiency gap is meant to serve as a proxy for competitiveness, the exact numerical relationship is not especially important. So, I may consider data transformations or even conversion to a categorical variable.

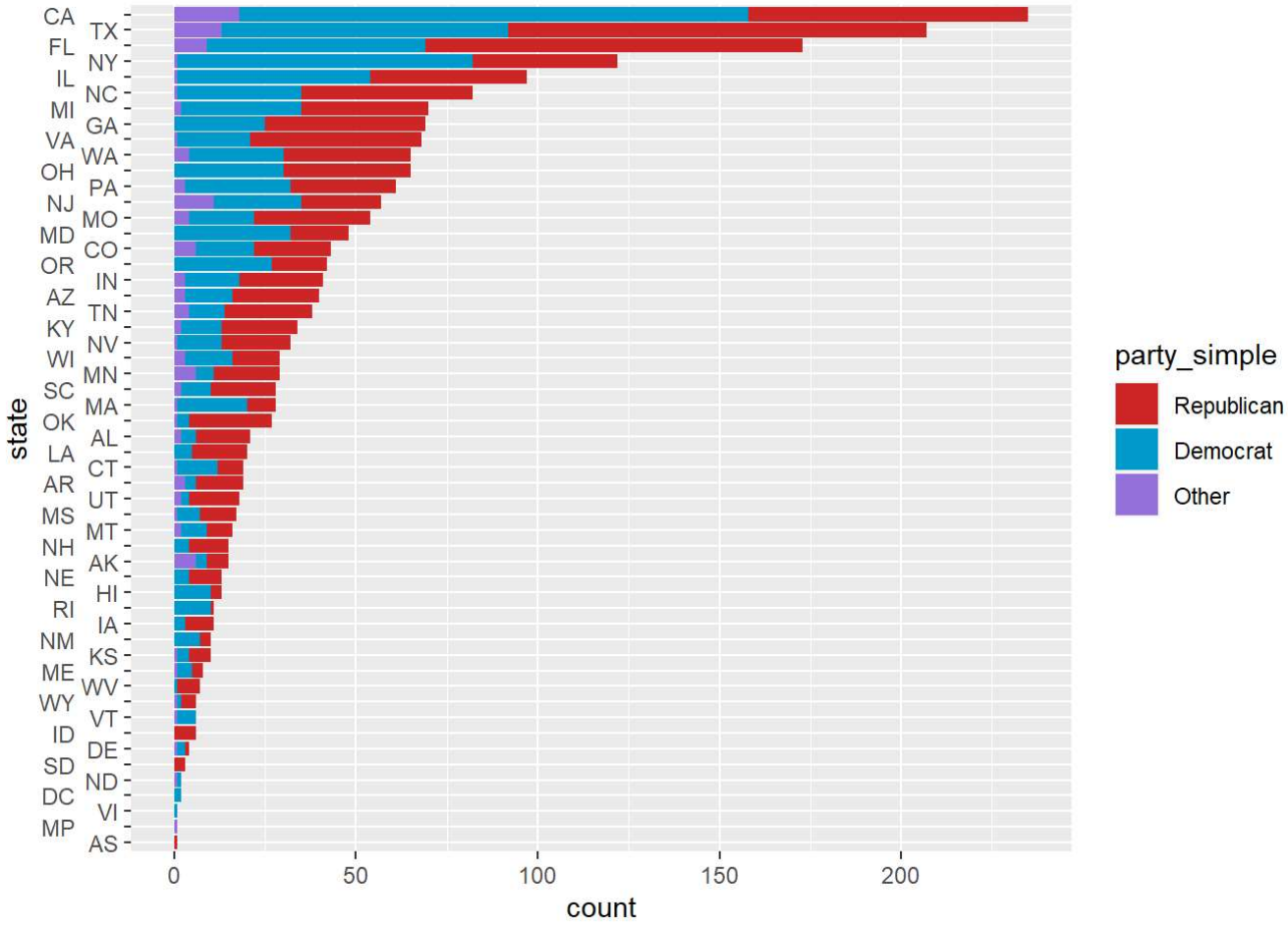
Regarding the independent variable (ideological polarity score), I'll need to finish the analysis and classification of candidate tweets before I can provide summary statistics. To get a sense, however please refer to https://rpubs.com/kac624/1023581 (https://rpubs.com/kac624/1023581), which provides a straightforward positive/negative sentiment analysis of a subset of the ~1mm tweets I now have gathered.

I can, however, provide some visualizations to give a sense of the distributions of target candidates across parties and geographies.

```
candidates_with_tweets <- candidates %>%
  filter(!is.na(twitter_name)) %>%
  mutate(party_simple = factor(party_simple,
                               levels = c('Republican', 'Democrat', 'Other'),
                               ordered = TRUE))

candidates_with_tweets %>%
  ggplot(aes(party_simple, fill = party_simple)) +
  geom_bar() +
  scale_fill_manual(values = c('firebrick3','deepskyblue3','mediumpurple'))
```




```
candidates_with_tweets %>%
  mutate(state = factor(state,
                        levels = names(sort(table(state))))) %>%
  ungroup() %>%
  ggplot(aes(state, fill = party_simple)) +
  geom_bar() +
  scale_fill_manual(values = c('firebrick3','deepskyblue3','mediumpurple')) +
  scale_x_discrete(guide = guide_axis(n.dodge = 2)) +
  coord_flip()
```


state
CA
TX
FL
NY
IL
NC
MI
GA
VA
WA
OH
PA
NJ
MO
MD
CO
OR
IN
AZ
TN
KY
NV
WI
MN
SC
MA
OK
AL
LA
CT
AR
UT
MS
MT
NH
AK
NE
HI
RI
IA
NM
KS
ME
WV
WY
VT
ID
DE
SD
ND
DC
VI
MP
AS
0
50
100
150
200
count
party_simple
Republican
Democrat
Other