The Air War

Gov 1347: Election Analytics

Kiara Hernandez

October 6, 2022

Harvard University

Today's agenda

- **How to do district-level predictions with different types of data**
  -How to incorporate national-level data in district-level predictions
  -What to do in the absence of district-level data (polling)
  -Thinking through relevant outcome variables
  -How to deal with open seat contests
  -How to incorporate individual district-level incumbency
- **Advertising data mini-hackathon (20 minutes)**

- **Suggested blog extensions**
  - build your own model of ad effects
  - what's the deal with social media?

Reviewing district-level prediction

Reviewing district-level prediction

- Let's think through 4 scenarios we might be in when making district-level predictions:

Scenario (1) We have national-level data and no district-level data

Scenario (1) We have national-level data and no district-level data

-Example 1: Q8-Q7 GDP growth

-Best option: Append the data as-is — that is, assign all districts the national-level data
-Intuition: We might assume that all districts are experiencing if not the same, then similar, economic conditions (probably a bad assumption, but what else can we do?)

-Example 2: the generic congressional ballot

-Best option: Same as above - all districts are assigned the same D/R approval rating

Scenario (1) We have national-level data and no district-level data

-Example 1: Q8-Q7 GDP growth

-Best option: Append the data as-is — that is, assign all districts the national-level data
-Intuition: We might assume that all districts are experiencing if not the same, then similar, economic conditions (probably a bad assumption, but what else can we do?)

-Example 2: the generic congressional ballot

-Best option: Same as above - all districts are assigned the same D/R approval rating

Scenario (2a) We have good district-level data in some cases and bad district-level data in others

Scenario (2a) We have good district-level data in some cases and bad district-level data in others

-Example: District polls

-Best option: As we did in lab last week, we can borrow data from comparable districts

-Doing this on a case-by-case basis may be tedious, so we might consider coming up with a more generalized coding scheme: match on region, citizen voting age population (size), and lagged voteshare margin [1]
$\rightarrow$ For each district, code the region, CVAP, and lagged voteshare margin. If districts fall within the same region, fall within the same quantile of the voter distribution, and have a voteshare margin within 0.1% of one another, append the polling data if there is no existing data. Example below:

Scenario (2a) We have good district-level data in some cases and bad district-level data in others

-Example: District polls

-Best option: As we did in lab last week, we can borrow data from comparable districts

-Doing this on a case-by-case basis may be tedious, so we might consider coming up with a more generalized coding scheme: match on region, citizen voting age population (size), and lagged voteshare margin [1]
$\rightarrow$ For each district, code the region, CVAP, and lagged voteshare margin. If districts fall within the same region, fall within the same quantile of the voter distribution, and have a voteshare margin within 0.1% of one another, append the polling data if there is no existing data. Example below:

```
# example code (TO BE GIVEN, standby please)
# polls_df <- read_csv("dist_polls_2018-2022.csv")
```

Scenario (2b) We have good district-level data in some cases and bad district-level data in others

Scenario (2b) We have good district-level data in some cases and bad district-level data in others

-Example: Expert predictions

-Intuitively, why no expert predictions in the majority of districts? Because most districts will following historical voting patterns $\rightarrow$ What does this mean for our predictions?

-Option: create separate models for competitive districts and "safe" districts. In safe districts, you can model using only fundamentals + ... + (incumbency, polling, etc.). In competitive districts, you can then also incorporate expert predictions.

Scenario (2b) We have good district-level data in some cases and bad district-level data in others

-Example: Expert predictions

-Intuitively, why no expert predictions in the majority of districts? Because most districts will following historical voting patterns $\rightarrow$ What does this mean for our predictions?

-Option: create separate models for competitive districts and "safe" districts. In safe districts, you can model using only fundamentals + ... + (incumbency, polling, etc.). In competitive districts, you can then also incorporate expert predictions.

Scenario (3) How to incorporate individual district-level incumbency

Scenario (3) How to incorporate individual district-level incumbency

-For district-level predictions, we need to think about how to align our dependent variable and independent variable. We've thought a lot about this, and this our recommendation.

Independent variable:
binary incumbent_party variable (1 if the incumbent party is D, 0 if the incumbent party is R)
dummy open_seat variable (1 if open contest, 0 if closed contest)

Outcome variable: dem_party_voteshare (or rep_party_voteshare; reverse code incumbent_party in this case)
-This gets around the problem of open seat contests, but you could imagine other ways in which this outcome variable is not ideal

Scenario (4) National-level predictions (voteshare or seatshare)

Scenario (4) National-level predictions (voteshare or seatshare)

This follows from the example code given two weeks ago.

-Model 1: All incumbents

-Model 2: All challengers

-Model 3: Open seats

Scenario (4) National-level predictions (voteshare or seatshare)

This follows from the example code given two weeks ago.

-Model 1: All incumbents

-Model 2: All challengers

-Model 3: Open seats

Other general notes

Consider coding all variables (**especially** expert predictions) as dummy variables (recall the conversation from Discussion this week) -This simplifies our lives because it allows us to more easily identify the association between expected and voteshare.

Other general notes

Consider coding all variables (**especially** expert predictions) as dummy variables (recall the conversation from Discussion this week) -This simplifies our lives because it allows us to more easily identify the association between expected and voteshare.

Advertising data mini-hackathon

Midterm ad campaigns from 2006-2018 (excluding 2008)

Midterm ad campaigns from 2006-2018 (excluding 2008)

`ads_2006-2018.csv`: data related to the design and content ("creative") and day-by-day ad spendings and airings in each district of candidate/PAC/party-run ads day-by-day in each campaign.

`ads_issues_2012-2018.csv`: the same as `ads_2006-2018.csv`, but in 2012, ad issue content and estimated ad cost became available. Look at all the variables: names(ads_2006_2018), names(ads_issues_2012-2018).

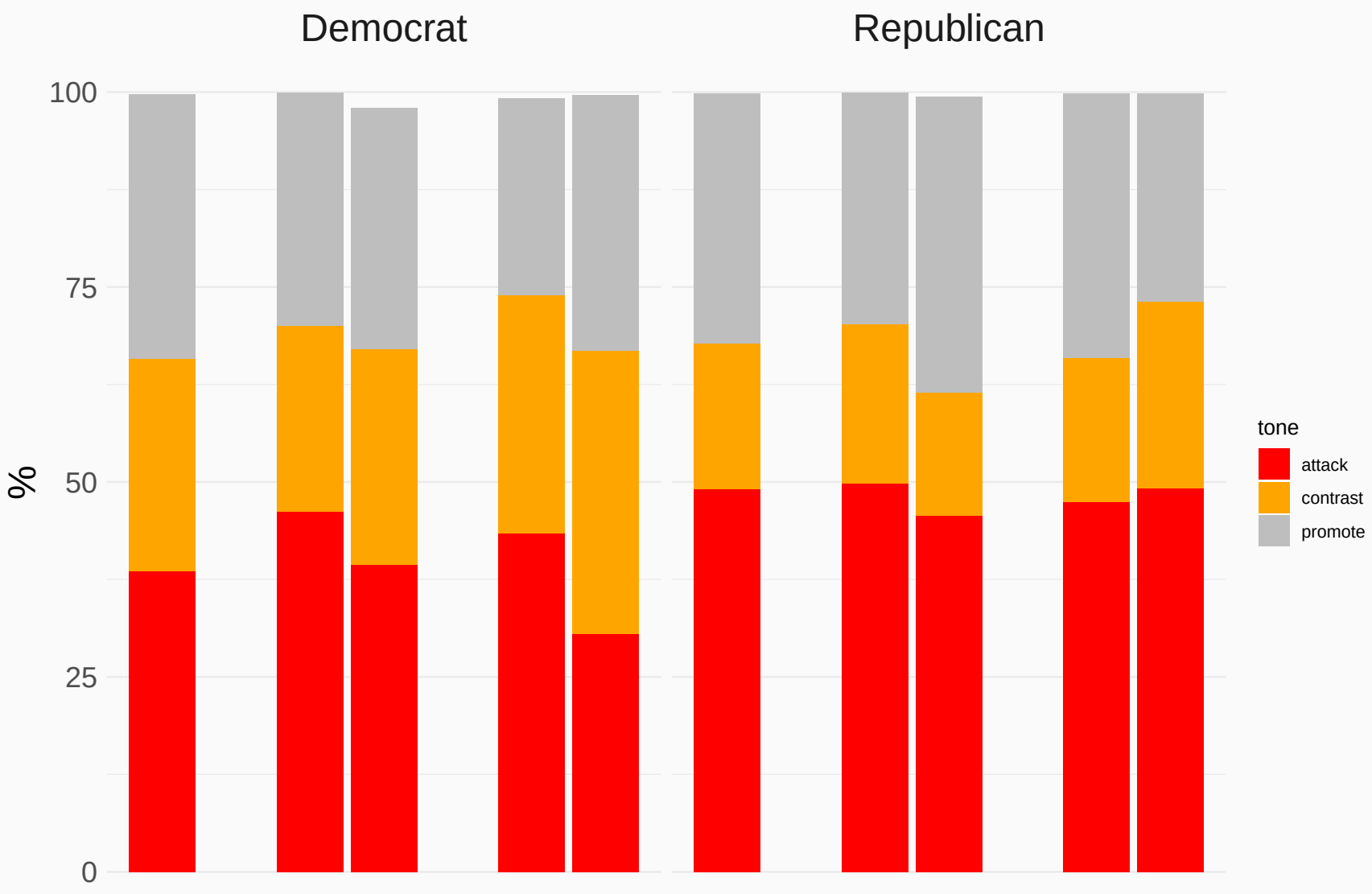
Advertising data mini-hackathon (20 minutes)

It is the year 2026 and you and a few of your Gov 1347 friends have landed a lucrative data science internship for the DNC. In 20 minutes, you all have a meeting with the chief campaign strategist of [your favorite soon-to-be congressional candidate] who wants answers to the following questions:

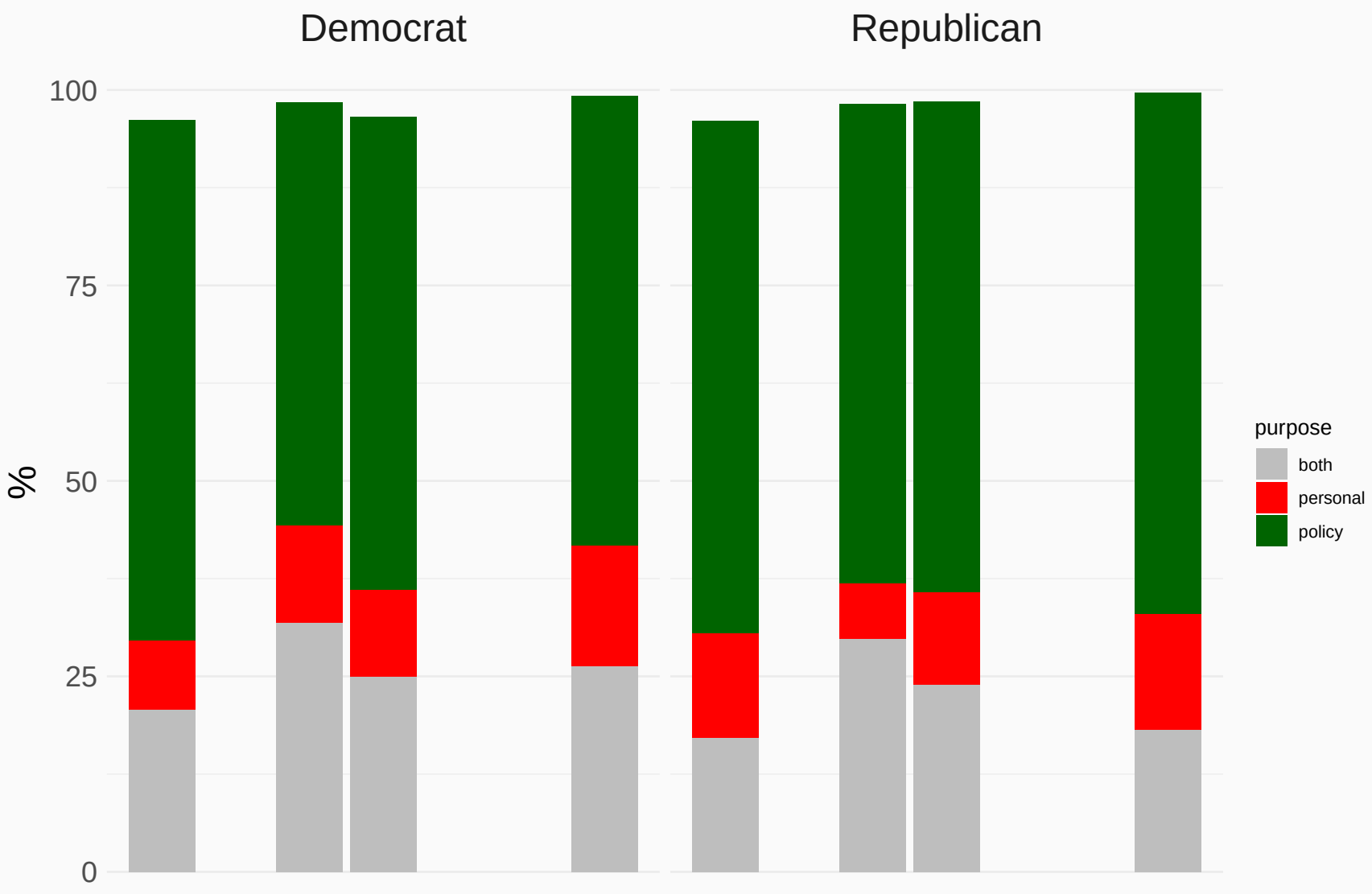
1. What are some trends in the tone and purpose of Democratic ads?
2. What are some trends in the issues mentioned in ads by each party?
3. How much money do winning challenger candidates spend on advertising and when during the campaign do they spend most of it?
4. Democrats did well in 2018... how much money did they spend on battleground states and which did they win?

Read in `ads_2006-2018.csv` and `ads_issues_2012-2018.csv` and answer these 4 questions with your team. If you're running short on time, guesstimate the remaining answers.

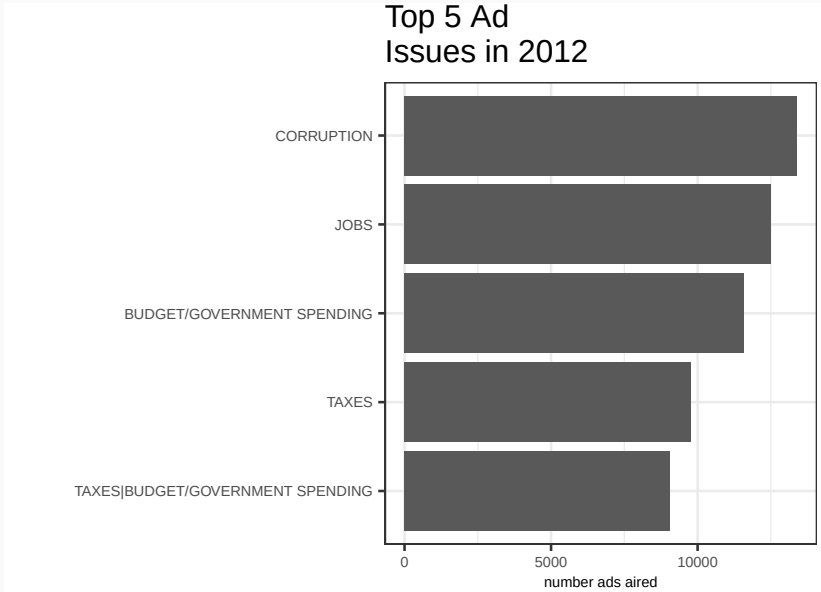
Tone and Political Ads

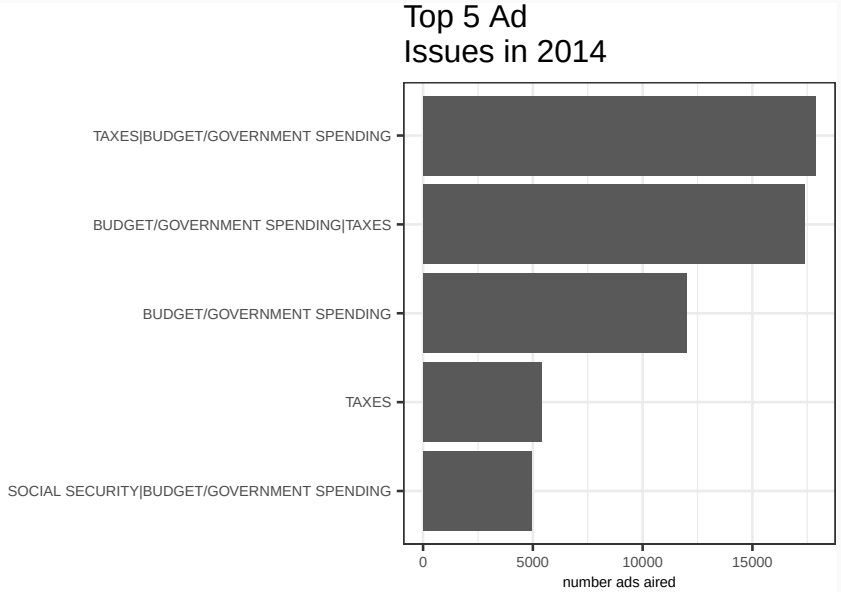


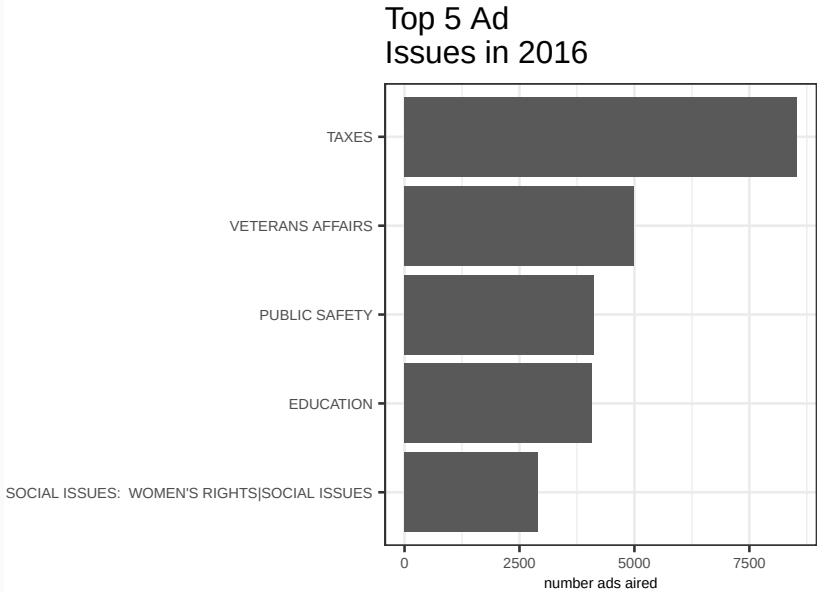
The Purpose of Political Ads

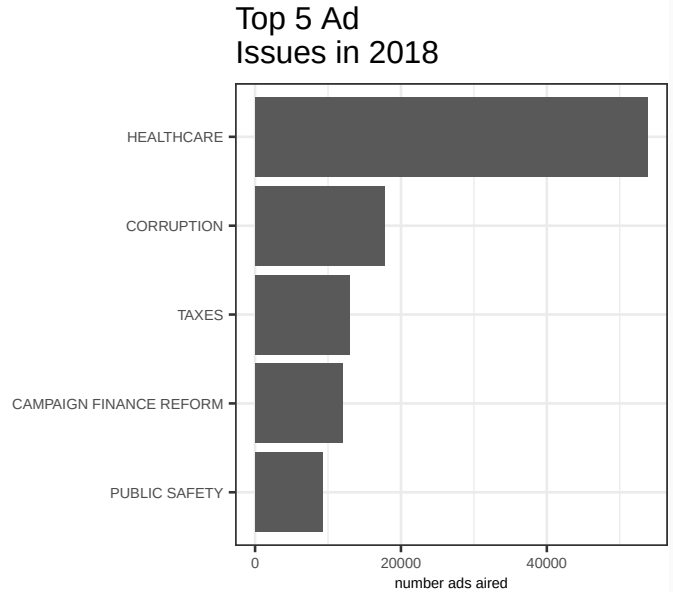


The Elections and Their Issues

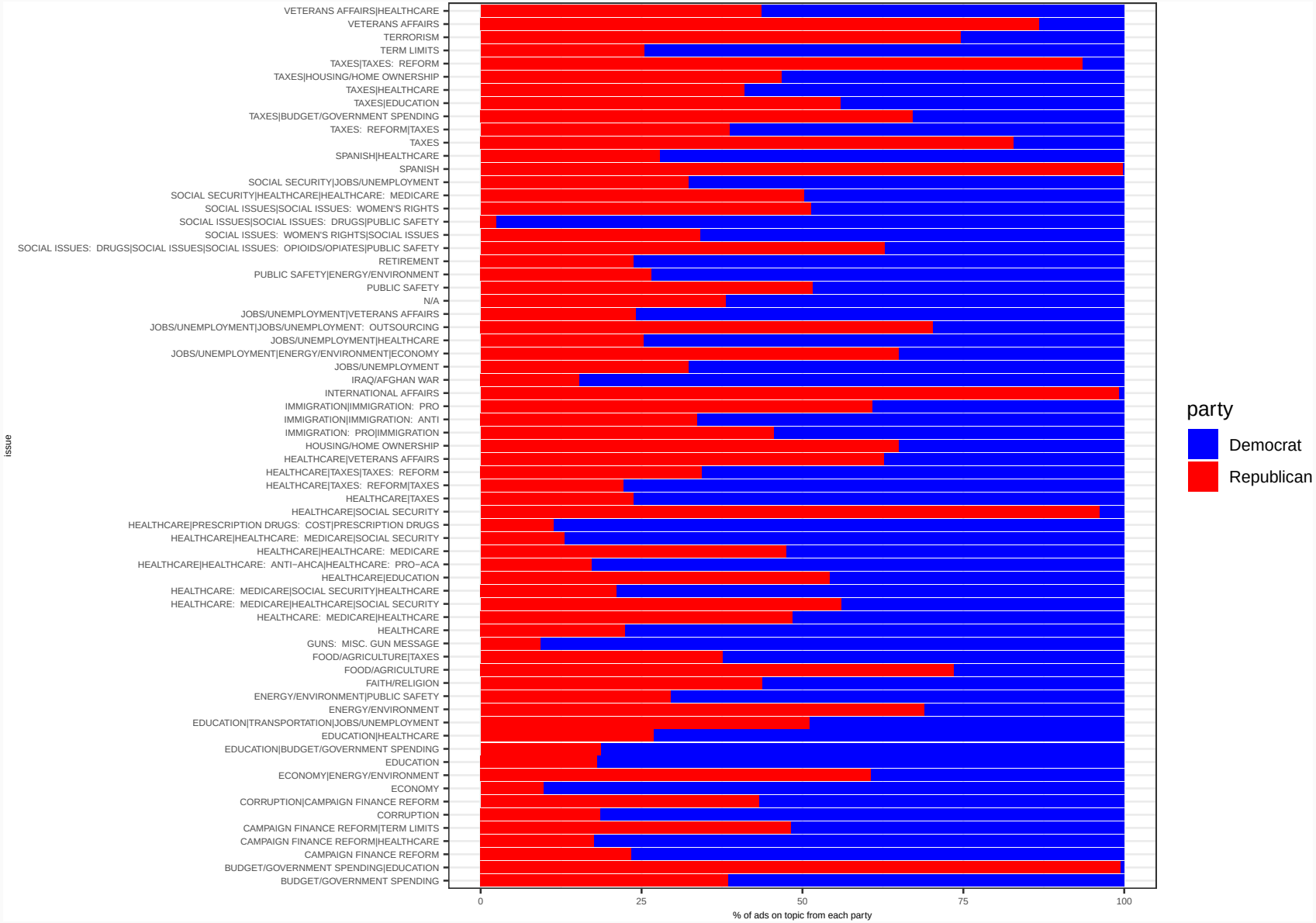




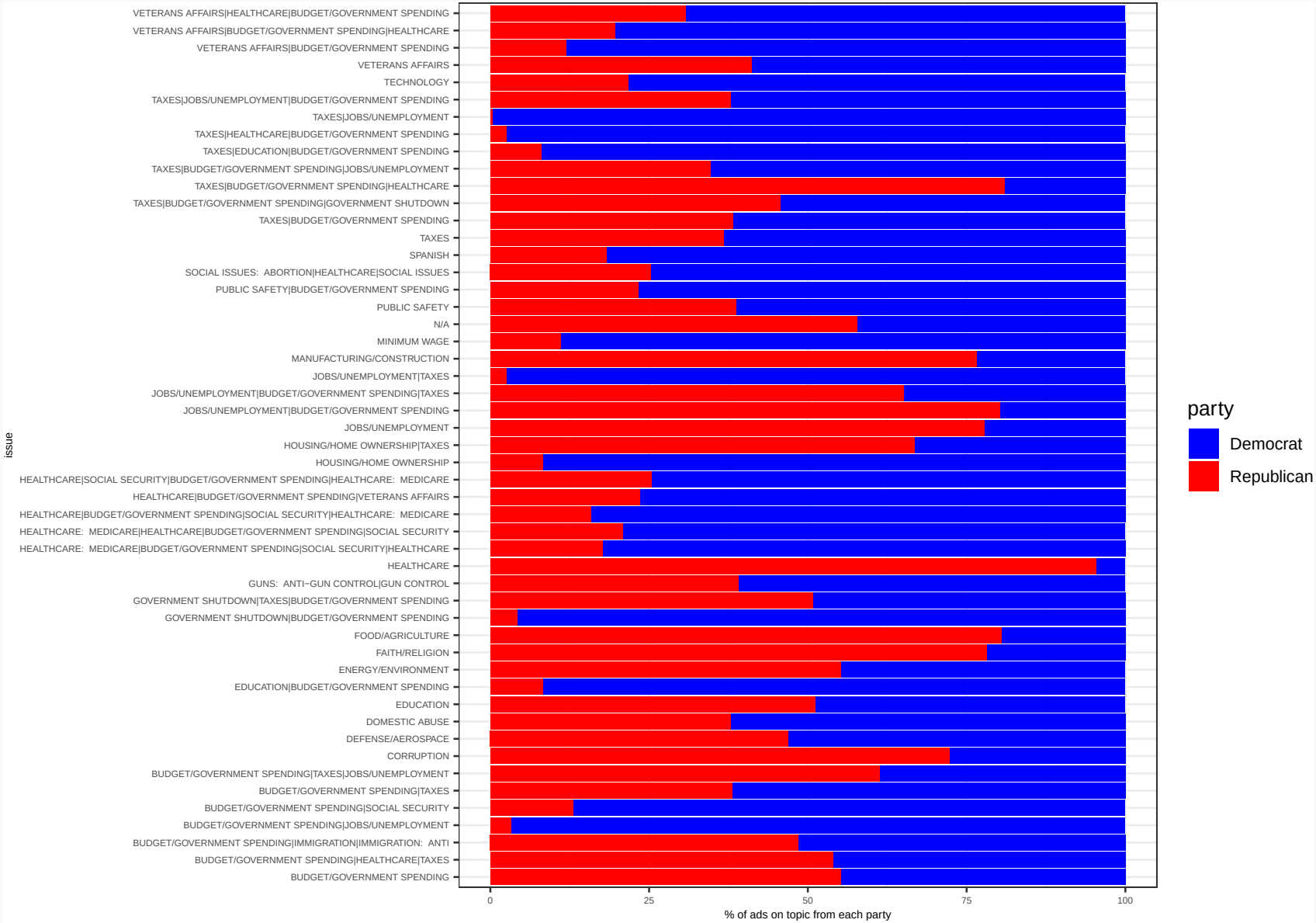




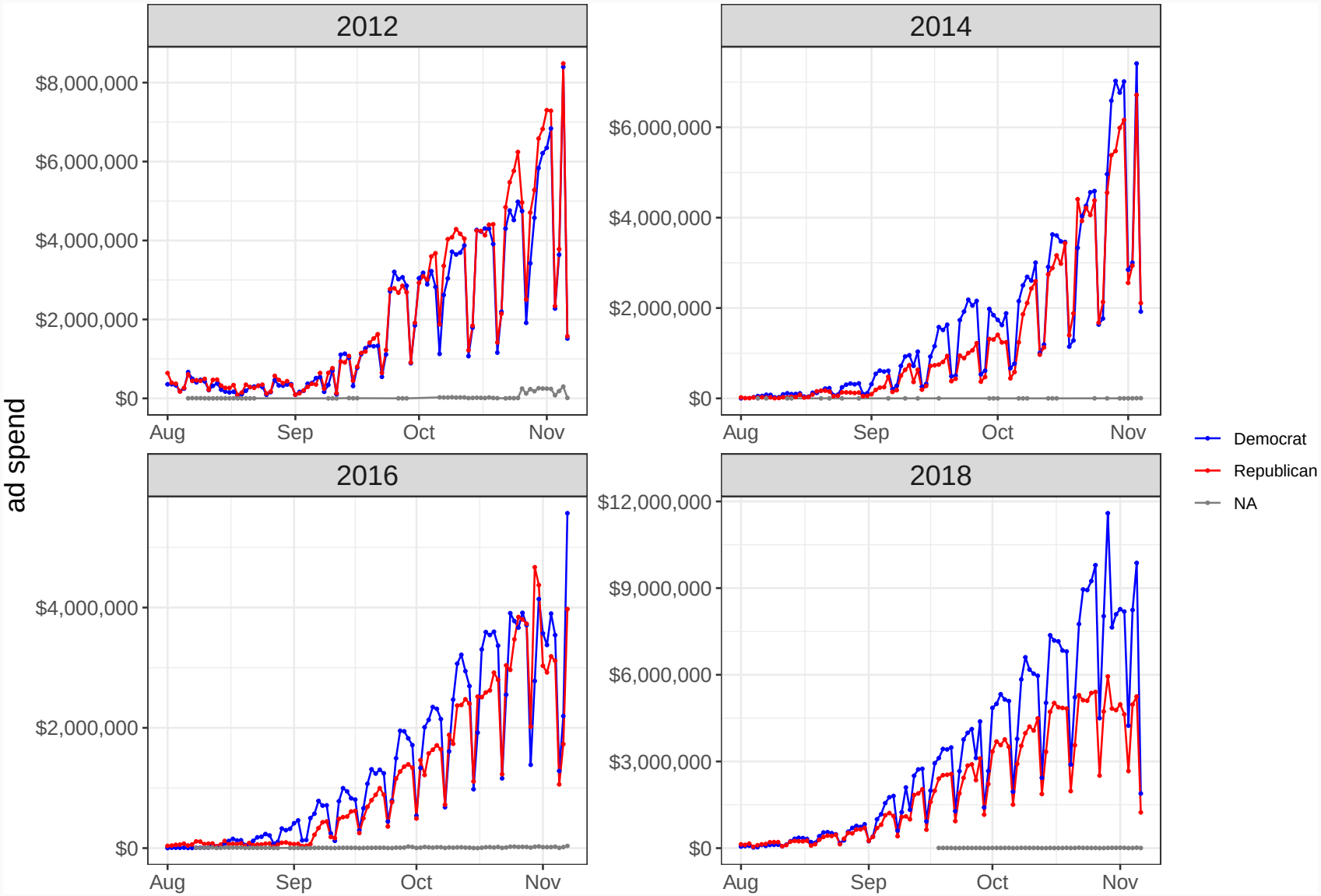
Campaign Ads Aired By Issue and Party: 2018



Campaign Ads Aired By Issue and Party: 2014



When to Buy Ads?



Tone in Political Ads

Tone in Political Ads

Blog Extensions (DUE ON TUESDAY 10/11)

Using 2018 Ads Data. Using the data from this section (and incorporating useful data from previous weeks) fit a model and predict 2022 district-level voteshare in relevants districts given existing data on ad spending in 2018.[2]. What are the limitations of your model?

Social Media. How much do campaigns spend on social media ads? Does social media influence election outcomes? How will it influence 2022? In the raw WMP files, you will find data on social media ads from 2012-2018. Reference the codebooks for more details about how these variables are coded.

Reviewing GPR. Review the readings for this week (Gerber et al. 2011, Huber and Arceneaux 2007). How do they quantify Gross Point Rating? Provide a descriptive/theoretical response to how you would approximate advertising effects at the national-level or for a specific district in which ads were run. Next week, we'll look more closely at how this can be done (Hint: Simulations based on voter population)

---

[2]Hint: to incorporate 2022 ad spending data, you will have to find that data yourself or use informed guesses about each campaign's spending in the relevant time period. You can use the most recent data as a stand-in (2018 data).