HW 6: Forging Your Own Path Toward Understanding Attitudes Toward Scientists Among US Voters

due 7/27/2023

This assignment is something like a very exciting choose-your-own adventure novel. A choose-your-own-model adventure, if you will! Very fun. It's a chance for you to grapple with some of the topics in the course on your own. You are doing this in your projects as well, yes, but we also want to get a sense of how you can do on your own, at a slightly smaller scale.

This homework will also be for practicing justifying decisions made during analysis. In your justifications, refer back to the materials from the class.

Introduction

Click the Github Assignment link on Canvas, then open it in Posit Cloud as a new project from a GitHub repository.

Warm up

Let's warm up with some simple exercises. Open the `hw6.qmd` Quarto file. Update the YAML of your Quarto file with your information, render, commit, and push your changes. Make sure to commit with a meaningful commit message. Then, go to your repo on GitHub and confirm that your changes are visible in your `.qmd` file. If anything is missing, commit and push again.

Packages

You have some flexibility about the packages that you will use, as you will be picking the methods that you would to use. Look back through your previous assignments and the example codes for informaton on which packages are best suited for which method. Remember that the `tidyverse` suite of packages is useful for data wrangling and visualization. Recall that you can use `patchwork`to combine `ggplot2` plots.

```
#load R packages
```

Data and Objective

The data for this assignment, contained in the `data` folder of your repo, comes from the 2020 American National Election Studies survey. The ANES, which is carried out in 4-year cycles, asks voters a series of demographic, political, economic, and social questions *before* and *after* each presidential election.[1] The ANES is an extensive survey. You can find a codebook for The ANES includes a series of questions called *feeling thermometer* questions. In these questions, respondents are asked how warmly on a 0-100 scale, where 0 is cold adn 100 is very warm, they feel about a range of topics and concepts. For almost all respondents, the data contains responses to a similar set of questions before the election and after the election.[2]

The goal of this assignment will be to predict the **post-election** feeling thermometer toward *scientists*. In other words, to what extent do respondents feel warmly (approve) of scientists after the 2020 election.

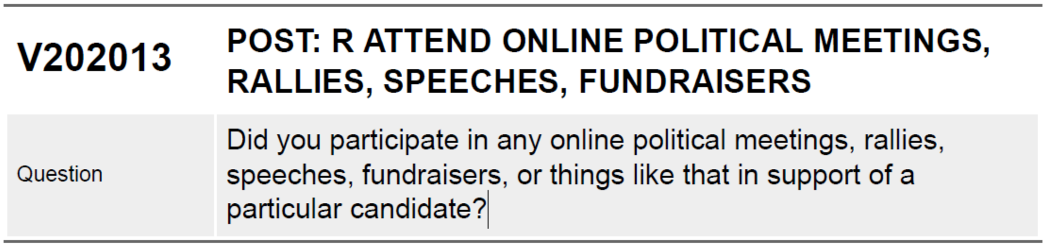
A few things to keep in mind:

1. All variables in the ANES have alphanumeric identifiers – you will have to look in the codebook to see which variable is connected to which alphanumeric code. For example, the outcome variable (post-election feeling thermometer) has alphanumeric identifier `V202173`.
2. Keep in mind that for each respondent there are pre- and post-election responses. **You should only use the post-election variables** (on pages 324 - 665 in the codebook in the `data` folder). This may seem like a daunting number of variables (and there are quite a lot!), but you will be asked to choose only a small subset of them. As shown in the image below, all **post** election variables are also clearly marked in the codebook.
3. Almost all variables are numeric in the `.csv` file in the `data` folder. **However, most variables are not actually numeric!** Instead, responses to categorical (limited choice) questions are given numerically. The picture below indicates how variable `V202013`

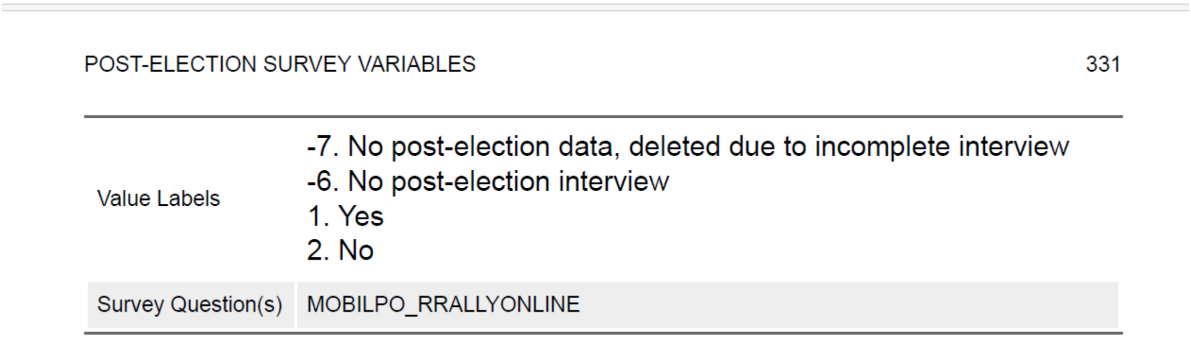
[1] Previous editions were also carried out during midterm elections.

[2] There was some attrition between the two waves, annd not all respondents were able to be interviewed again after the election.

appears in the codebook; the chunk afterwards prints out how many time each value in the variable appears in the data. You will see that the category *values* are stored in the data, not the labels. Keep this in mind when working on this assignment.



V202013	**POST: R ATTEND ONLINE POLITICAL MEETINGS, RALLIES, SPEECHES, FUNDRAISERS**
Question	Did you participate in any online political meetings, rallies, speeches, fundraisers, or things like that in support of a particular candidate?



POST-ELECTION SURVEY VARIABLES 331

Value Labels	-7. No post-election data, deleted due to incomplete interview -6. No post-election interview 1. Yes 2. No
Survey Question(s)	MOBILPO_RRALLYONLINE

```
# first 10 values
head(anes2020$V202013, n = 10)
```

```
[1] 2 1 2 2 2 2 2 2 2 2
```

```
# last 10 values
tail(anes2020$V202013, n = 10)
```

```
[1]  2 -6  2  2  2  2  2  2  2  2
```

```
# counting up how many times each value/category appears in the data
anes2020 %>%
  count(V202013)
```

```
# A tibble: 4 x 2
  V202013     n
    <dbl> <int>
1      -7    77
2      -6   754
3       1   988
4       2  6461
```

4. In the image and code output above, you will notice that the variable contains *negative* values. These represent two different forms of *missing* data. For our purposes, we can treat these all the same: it just means that we don't have information for that question for that respondent. But be aware: *some questions have more missing value codes than others!* In addition, for some questions **very high values** (relative to the possible values of that variable) can *also* mean different kinds of missing values. Always check to see what the different values associated with a variable mean.

Exercises

```
# import data
```

Variable Selection

Q1

Choose at least 10 variables (they can be be functions of existing variables, like if you want to turn a multi-category variable in a binary one) as predictors for a model predicting feelings toward scientists. In 4-5 sentences, explain why you chose the variables that you did.

Data Visualization

Q2 - Outcome

Make an appropriate plot for the outcome variable.

In 4-5 sentences, explain what the plot shows and why you went with the plot you did. Remember that the plot has to look nice and interpretable.

Q3 - Predictor and Outcome

Choose one of the predictors that you are interested in and make an appropriate plot that helps you identify how it is related with the outcome.

In 4-5 sentences, explain what the plot shows and why you went with the plot you did. Remember that the plot has to look nice and interpretable.

Q4 - Two Predictors

Choose two of the predictors that you are interested in and make an appropriate plot that helps you identify how they are related.

In 4-5 sentences, explain what the plot shows and why you went with the plot you did. Remember that the plot has to look nice and interpretable.

Modeling

Q5 - Choosing Models

Choose two modeling approaches that you think would do well for this problem and your chosen variables. One of the two modeling approaches must be one of the tree based methods you learned about during Modeul 7 (single tree, random forest, gradient boosting machine, BART).

Explain in 4-5 sentences why you chose these two modeling strategies.

Q6 - Fitting Models

Fit the two models. In each case, if it involves tuning hyperparameters, explain how you tuned them and why. If you did not tune but chose specific tuning parameter values, explain why.

Q7 - Interpreting Models

For each of the models, what do they tell you about the relationship between the predictor you picked for Q4 and the outcome? Write a 3-4 sentence summary about the possible relationship implied by the two modeling approaches.

Q8 - Comparing Models

Compare the two models in an appropriate way. Explain which model fits better in 3-4 sentences and how you know. In a further 3-4 sentences, explain why you chose to compare them in the way you did.

Final Steps

When you have made your final edits, remember to render, stage the changed files, then commit, and push to the repo.

Don't forget to submit your pdf on Canvas.

Grading

! A note on how we will grade this assignment

We will **not** grade this assignment based on how well your model does or how complicated it is. We prefer that you use simpler models that you understand rather than more complicated approaches that you do not fully understand.
We **will** grade this assignment on whether the code is reasonable (and it should render without errors, of course).
But most of all, we will grade this assignment on the quality of your written justifications. You can use substantive- or data-science-inspired justifications for your decisions.

Please put your name on your answer document and add the date. You won't get points for doing this, but you will *lose* points for not doing so.

Remember to comment your code. You don't get points for commenting, but you will *lose* points for not commenting your code.

For clarity's sake, you should label your code chunks (this helps make debugging easier). You will *lose* points for not naming chunks.

If an exercise asks a question, you should answer it in narrative form and not just rely on the code. You will be *penalized* otherwise.

Please suppress warnings and messages. You will *lose* points if you do not suppress warnings and messages.

If you are asked to make a visualization, make sure that it is presentable, professional in appearance, has a title, and has properly labeled axes. You will *lose* points if you do not do this.

Points Breakdown

- Q1 - 2 pts
- Q2 - 3 pts
- Q3 - 3 pts
- Q4 - 3 pts
- Q5 - 4 pts
- Q6 - 6 pts
- Q7 - 4 pts
- Q8 - 4 pts
- Workflow (includes making sufficient commits and labeling chunks): 4 pts

Total: 33 pts