DATA 310 Final Project

2/15/2022

For the final project of this course you will present a short report that demonstrates your knowledge of the course material.

The project is due on **Wednesday, March 9th at midnight**. Because I must grade these by Saturday the 12th, I cannot accept any late projects. If you turn it in late you will receive a zero.

In this project you will be using data from the 2020 American National Election Study. Specifically, you will be looking at just over 8000 observations from the the pre-election component of the study, completed in October of 2020. I've reduced the number of variables in the dataset, and labels are included, but you will have to use the included codebook to clean the data in a way that is appropriate for analysis. (For example, many of the missing values are set to be numbers, not NA.) Note that `V200010a` is the survey weight you should be using in your analysis. You should use this weight when running regressions and performing hypothesis tests (i.e. when using `svyttest` and `lm`), but don't worry about weighting when producing visualizations or presenting summary statistics.

Each of you will use as your main dependent/outcome variable the "feeling thermometer" question[1] for Joe Biden:

```
summary(anes$V201151)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -9.00   15.00   50.00   47.81   85.00  998.00
```

The goal of the assignment is for you to form and test a hypothesis about what explains (or fails to explain) voters' feelings towards Joe Biden in 2020. For example: a report might investigate how religiosity and ideology interact to produce positive and negative feelings about Biden. You should make your singular hypothesis clear to the reader within the first 2 or 3 sentences of the report. While you are testing a single hypothesis, that hypothesis does not need to involve just one variable. That being said, you should limit yourself to using no more than 10 variables in your analysis.

You will hand in two files:

- A report with the results of your analysis. This report will be no more than 5 pages long, including figures and tables. This report should **not** include any R code, and should present your findings as if you were presenting them in a professional setting. You should focus on presenting a coherent argument about why your chosen factors explain (or don't explain) feelings towards Trump, not a travelogue of the process you used to get those findings. In other words, the report should not be, "First I did X. That didn't work so I tried Y". Once you have your results/figures/tables complete, you should write your report to present those in a clear and concise way.
- The R code that produced the findings in the paper. This code should be organized and well commented, such that I am able to understand each step you took to get from the raw dataset to the figures/tables/statistics presented in your paper. Each and every number, table, and figure that is in your report needs to be able to be generated from your code by me, starting with the raw data.

To satisfy these two requirements you can also hand in a .Rmd file and a corresponding PDF, but that PDF should not contain any R code.

[1] It should be clear from this summary that you will have to do some cleaning for this variable!

Focus your work in a way that demonstrates your knowledge of course materials. You must perform *at least* one regression analysis. Other things that you may wish to include (not mandatory, and not exhaustive):

- Visualizing data in appropriate ways.
- Presenting summary statistics of the main variables you make use of.
- Demonstrating knowledge of sampling and weights.
- Performing hypothesis tests on sample statistics.
- Considering the means of variables.
- Considering the difference in means between groups.
- Correctly interpreting the output of a regression.
- Assessing the fit of a regression.
- Adding variables to improve regression fit or to deal with omitted variable bias.
- Considering interactive hypotheses for regression.
- Considering the standard errors of your regression and if they can be improved (Week 7.)
- Considering the causal identification of your regression (Week 8, bonus if you include this).