Programming and Data Management Final Project/Exam

The American National Election Study is a gold-standard survey that runs every two years. It is one of the most important sources of data for the study of American politics.

In this final exam/project, you will use these data to investigate the attitudes of Americans.

This project is based on the ANES Time Series Cumulative Data File, which aggregates many questions that have been included in the survey since 1948. When you work with survey data like these, you generally work with a data file that includes responses, a codebook file that documents the questions asked, and a variable guide that lists in abbreviated form the questions on the survey and the variable name (column) of the data associated with each question. You'll need to use all these files to perform your analysis.

For these questions, do not worry about survey weights. Similarly, you should discard any responses where the interviewee reported "Don't know" or "refused to answer" or similar responses. You can just recode these kinds of responses as missing data, and you should not report or include these in your answers/analysis. (In real life, you probably would not adopt this approach, but for our purposes, it's ok for now).

Part One includes a series of guided questions. This is worth 40 points.

In Part Two, you explore questions of interest to you. Part Two is worth 20 points.

You'll work in a combination of R and Python, so you'll have some code chunks in R and some code chunks in Python.

Create a .qmd file that answers the questions in Part 1 and describes and performs your investigation in Part 2. The .qmd file should be attractive and well-formatted when rendered - for example, points will be reduced if you dump huge amounts of console output into the rendered document. You want the document to readable and meaningful. Display the code chunks as you go to show your work. Render an html file with your report and submit it along with the .qmd file.

When you import the ANES data into Python, you might encounter some issues with data types. To avoid these problems, use the following code to make sure that pandas coerces all

numbers into numeric format. Note that I renamed the actual .csv to `ANES.csv` here for clarity - the actual name when you download the document is longer.

```
anes = pd.read_csv("ANES.csv")

anes2 = anes.apply(pd.to_numeric,errors="coerce")
```

This code forces pandas to parse the data as numeric, avoiding problems with mixed types of data in columns (something is weird about ANES original .csv file - this is the kind of thing that happens with data analysis). If it occurs that pandas cannot coerce the data in a cell into a number, it will replace whatever value is there with a `NaN`. **Note:** there are a few columns in this data set where this transformation will cause a problem, like VCF0901b, which is the state postal abbreviation. If you want to do any state level analysis in Python, you'll need to account for this. I would just re-import the original .csv and then bind that unaltered state code column back to the transformed as a new column. In R, you shouldn't encounter any of these issues.

Part One: Guided Questions (40 points)

Question 1 (2 points)

R

Create a tibble that shows how many respondents are in each wave of the survey.

Question 2 (2 points)

Python

How are survey respondents distributed across the major geographic regions of the US in the 1996 wave of the survey? (i.e., how many respondents per region)

Question 3 (4 points)

R

Considering the 2008 wave and subsequent waves, what percent of these interviews in each wave were partially or entirely translated to Spanish? Don't forget to account for both pre and post election interviews (ANES surveys include pre or post election interviews).

Question 4 (8 points)

Python

One of the questions on this survey has the interviewer read a list of words and phrases that people use to describe political figures. Then, the interviewer asks the interviewee to think about a given political figure, and the interviewer asks whether a given phrase describes that political figure extremely well, quite well, not too well or not well at all.

So, for example, the interviewer might say, "Think about Ronald Reagan. In your opinion, does the phrase or word 'intelligent' describe Ronald Reagan extremely well, quite well, not too well, or not well at all?"

Based on the survey data between 1980 and 2008, which president did women under the age of 40 think was the most **knowledgeable?** Which president was the least knowledgeable in the eyes of this group? You can average across all surveys during a president's term. Some presidents will be included in more waves than others - that's fine, use the average regardless of the number of terms.

Question 5 (8 points)

R

These days, the evidence suggests that higher levels of education are associated with more liberal political attitudes, as measured on a traditional seven-point ideology scale.

Track this pattern over time. Use respondents from 1980, 1992, 2000, and 2020. What is the average political ideology of survey respondents with a college degree or greater vs. the political ideology of respondents without a college degree? (Note: some college doesn't count)

In addition, repeat this, but compare how this breaks down on along racial lines. Is the pattern the same for whites and non-whites?

Question 6 (8 points)

Python

Several questions on this survey are related to social trust. I'm talking here about questions VCF0619-VCF0621. Let's just look at the 2004 survey responses.

Construct a scale that adds together the responses to these three questions so that higher values indicate greater social trust. Set the scale so it runs from zero (the minimum amount of trust).

Now, consider how this scale relates to respondents' partisan identity (strong Democrat to strong Republican). Do you see any evidence that greater social trust is associated with partisan identity?

Question 7 (8 points)

R

A common type of question on political surveys is the "feeling thermometer" where respondents are asked how warm/cold they feel about certain topics or political groups or figures.

It is widely believed that political polarization today is worse than in the past - Republicans have more negative feelings about Democrats today than they did in years past, and vice versa.

Use these survey data to assess this claim.

Part 2: Exploration (20 points)

For the remainder of this project, you will further explore the ANES data and then create and answer some questions that are interesting to you about American politics.

Answer four different questions or test four different claims, or conduct some other analysis of your choice, in the style of what you did in Part 1. In each question, introduce at least one new variable into your analysis.

If you have any concerns about what makes a good question, please reach out for help.

You'll probably need about 500 words of text (1.5-2 pages) to describe what you are looking at/what analysis you are performing, and of course you'll need to include the associated code that provides the numerical/statistical evidence for your claims.

Use Python for at least two of these explorations.

You have flexibility to pursue questions that interest you. But, you need to leverage the data in a meaningful way. For example, "The mean response to question X was higher than the mean response to question Y" is not good enough. Use the temporal dimension of these data, the demographics of survey respondents, or the responses to different questions to create interesting comparisons or to focus on populations of particular interest to you. Let your creativity flow!

Good luck and have fun!