Midterm 2020

Instructions

1. The following questions should each be answered within an R script. Be sure to provide many comments in the script to facilitate grading.
2. This is open note, open book, open Internet test. But each keystroke in the assignment should be your own. Do not copy and paste from each other or from the Internet or even from the notes. And please do not discuss this work with your peers.
3. The test is designed to take approximately 5 hours. But you can spend up to eight hours to do it. Please turn it in by Friday, March 6 at 5pm. Just send me your R script in a private message on Slack. (Do not post it to a general channel.)
4. Part of your grade will be based on the readibility of your script. I will be running all of your code. If I don't understand what you are doing, I can't give partial credit. So pay attention to tabs, comments, and other tricks we have talked about for organizing your code.
5. As with all aspects of this class, there is not going to be a single "right" answer. That is not how real data analysis works. But I will try and provide you a lot of guidance. If you are confused, go ahead and ping me on slack. I can't promise I'll answer right away, but I will when I can.

Part 1: Presidential approval data

You are going to be using the following dataset: http://politicaldatascience.com/PDS/Datasets/Pres_Approval.csv

This is data includes approval data for US presidents Clinton, Bush, Obama, and Trump.

- The variable "Date" includes the date of the poll.
- The variable "Percent approve" is the percent of respondents who approve the current president.
- The variable "Percent disapprove" is the percent who disapprove of the president.
- The variable "Trump" includes the name of the person who was president.

---

1. Data cleaning

- Rename the variable "Trump" to be "Name"
- Make a new variable that called "Year" that just includes the year of the survey
- Remove the variable "No Opinion"
- Make a new variable called "NetApproval" that represents `Percent Approve - Percent Disapprove`

---

2. Some basic visualization

- Create a plot that shows approval by year.
- Now alter the plot to separate out this visualization by president.
- There is no "right" answer here, but here are a few things to consider, but I should be able to interpret the results clearly. This means clear labels on the axes, a legend, etc. Remember I am looking at the code, so useful comments are helpful here. Tell me what you are trying to do.

---

3. Making data play nice with other data

- The data below is "state of the union" addresses for each president.
- https://politicaldatascience.com/PDS/Datasets/State_Union.csv
- Using the year and the name of the president, you are going to merge in data into the state of the union data. For each speech, you will be be adding in variables that are the mean level of `Percent Approve` and mean `NetApproval`. So `NetApproval` would be the mean level of `NetApproval` for that president in the year the speech was given. Be careful here. Is this join going correctly? Check (and show me how you checked.)
- Make sure the new dataset *removes* all polls for years where there is no state of the union.
- Ask questions here if you are unsure. To be clear, the dataset should still be organized at the *speech* level when you are done.

---

4. Text

- Create variables that represents (a) the number of words in each speech, (b) the number of words that in in `ly` (a rough way to approximate the number of adverbs), and (c) references to the economy/economic growth/jobs. This is a **speech-level** dataset.

- Create a plot that shows the by-year relationship between `NetApproval` and mentions of jobs/economy.
- Make a new dataset so each row now represents a single sentence. It should still contain the polling data you merged in before.
- For each sentence, make a variable that indicates whether or not it references economy/economic growth/jobs.
- Now add back into the speech-level dataset a variable that indicates the *proportion* of sentences that are about the economy.