Practice Midterm

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. Undocumented code will not be graded.
2. For the real midterm you may NOT work in teams. So you should practice coding on your own. But once you are done you are invited to compare your answers with peers to better learn.
3. 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.
4. The test is designed to take approximately 5 hours. But you can spend up to eight hours to do it.
5. 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.
6. 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.
7. Ping me on slack if you have questions. Feel free even to send examples of plots/code for clarifications as needed. I can't promise to always be available but I'll try and respond.

Part 1: State policy data

You are going to be using the following dataset: http://ippsr.msu.edu/sites/default/files/correlatesofstatepolicyprojectv2_1.csv

This is data from the *correlates of state policy* database. You can read more about that here: http://ippsr.msu.edu/public-policy/correlates-state-policy

> The Correlates of State Policy Project aims to compile, disseminate, and encourage the use of data relevant to U.S. state policy research, tracking policy differences across the 50 states and changes over time. We have gathered more than 900 variables from various sources and assembled them into one large, useful dataset. We hope this project will become a "one-stop shop" for academics, policy analysts, students, and researchers looking for variables germane to the study of state policies and politics.

Here is the complete codebook: https://ippsr.msu.edu/sites/default/files/CorrelatesCodebook.pdf

1. There is one observation in this dataset for each state for each year.
2. The dataset contains 118 years, but most variables are only available for a few years.

---

1. Data cleaning

- Remove Washington, DC from the data
- Remove all data from before 1940
- Add a variable to the data that indicates the decade (e.g., 80s, 90s, etc)
- Reduce the data down to the following columns: `incshare_top10`, `region`, `poptotal`, `undocumented_immigrants`, and the new decade variable you made before. Take a minute to read the codebook for these variables.
- Create a new variable that is the percent of each state that is an undocumented immigrant.

---

2. Some basic visualization

- Create a plot that shows the trend in the `inshare_top10` by region over time. The x-axis should be time and the y-axis should be `inshare_top10`. I want all 50 states to appear on the plot. There should be 50 lines. But there should be a different facet for each region.
- 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

- Now read in the data about US Mayors: https://raw.githubusercontent.com/jmontgomery/jmontgomery.github.io/master/PDS/Datasets/Mayors.csv"
- Using the state name, join these two datasets. For each mayor, you will be be adding in a variable that is the mean level of `inshare_top10`, percent undocumented, and the modal value of region. Be careful here. Is this join going correctly? Check (and show me how you checked.)
- Now join the resulting dataset (including these new variables) to the mayoral twitter dataset used in class.
- Make sure the new daset *removes* all mayors who do not have twitter accounts. And the resulting dataset should be organized at the tweet level.

---

4. Text

- Create a set of search terms to identify tweets related to immigrants, undocumented workers, etc.. It doesn't need to be too complex, but should involve more than one term.
- Create a set of search terms to identify tweets about inequality.
- Reorganize the data to the state level, where each row is a state and the main variables of interest are: the average number of mayoral tweets from that state on immigration, the average number of mayoral tweets in that state on inequality, percent undocumented, and `inshare_top10`.
- Create a plot that shows the relationship between the immigration tweets variable and percent undocumented.
- Create a plot that shows the relationship between the inequality tweets variable and `inshare_top10`.

---

5. Functions

- Go back to the *raw* state policy data.
- Write a function that takes in four arguments: (1) the dataset, (2) the variable that indicates time, (3) the variable that indicates state, and (4) a variable of interest.
- The function should output the mean level of the "variable of interest" for that state across all years. Try this on different variables and be sure it handles missing data correctly. So the output of the function should be a dataset organized at the state level (one row per state) with two variables: (a) state and (b) the mean value of "variable of interest" for that state.
- Implement this function using for loops and another function implementing this using other tools such as `map`. I don't care which one, but no loops. So you will end up with two functions.
- Compare the speed of these two functions.