Homework #2

In this homework assignment, you will be cleaning data, practicing your descriptive statistics, and running an ANOVA or regression. There are three datasets affiliated with this homework:

1. **Tweets from President Trump** (from January 20, 2020 to today)
2. **A COVID-19 Survey** (with questions about whether people trust information from politicians, news media, and health experts)
3. **Public Facebook posts using "defund the police", "abolish the police", and/or "reform the police"** (from March 1, 2020 to today)

Each of these datasets comes with its own challenges and "dirty" bits.

For your assignment, you will **choose two of the three datasets** (each dataset you work on is 50 points, and you can earn a maximum of 100 points). Then you will do the following:

1. Clean the data.
2. Run some descriptive analyses.
3. Run a regression or ANOVA (this is <u>intentionally</u> broad: test a hypothesis that you may be interested in)
4. At the bottom of your submission file, write a number from 1 to 10 for how easy (1) or difficult (10) this assignment was.

For this project, you will share a rmarkdown file. We will go over what a rmarkdown file is in class on Wednesday, or you can check out the tutorials posted on this week's page.

More specific instructions for each dataset can be found below. Data dictionaries (descriptions of each dataset's variables) for each dataset can be found here.

Data 1: Trump's Twitter

1. Clean the dataset. At minimum, you should:
    a. Fill the missing data for retweet_count and favorite_count (which should be interpreted as 0). Favorite_count has a few strings that need to be fixed into numerics as well!
    b. Make the timeofday variables all consistent in casing ("Afternoon" vs. "afternoon")
    c. Create a binary "link" variable (logic or factor; or fix the "urls" variable to be a binary variable) for whether the tweet contains a link.
    d. Revise the year variable into a numeric
    e. Make sure is_retweet is a binary or factor and contains only TRUE (1) or FALSE (0).
2. Run the following descriptive/answer the following questions:
    a. How many tweets are retweets?
    b. What is the average and standard deviation of retweets of Trump's tweets?
    c. What is the average and standard deviation of likes of Trump's retweets?

    d. How many times did Trump tweet in 2017, 2018, 2019, and 2020?

3. Run an ANOVA and Tukey’s Test. Summarize the results of your ANOVA and Tukey’s Test below your analysis.

Data 2: A COVID-19 Survey (Roper iPoll ID: 31117235)

1. Clean the dataset. At minimum, you should:
    a. Revise TRUDP105R, CNRNCMCV1, TRSTCV1A, TRSTCV1B, and TRSTCV1D from factor to numeric variables
    b. Clean the PARTYID variable. Group Independents and Other together. Collapse the missing values and the “Unsure-Blanks-Refused” group together.
2. Run the following descriptive/answer the following questions:
    a. Run contingency tables for (1) TRSTCV1A and PARTYID, (2) TRSTCV1B and PARTYID, and (3) TRSTCV1D and PARTYID
    b. What is the average and standard deviation of President Trump’s approval (TRUDP105R)?
        i. What is the median?
    c. Are older participants/age groups more or less likely to trust news media (TRSTCV1B)?
    d. Are college graduates more or less likely to trust news media?
        i. How about public health experts?
3. Run a regression where the dependent variable is either TRSTCV1A, TRSTCV1B, and TRSTCV1D. Summarize the results of your regression below your analysis.

Data 3: Facebook Data of Reform/Defund/Abolish the Police Discourse

1. Clean the data. At minimum, you should:
    a. Revise the names of the following variables Video Share Status, Total Interactions, BLM Page to replace dots with underlines and to undercase the name (e.g., “Video.Share.Status” → “video_share_status”)
    b. Revise the Haha and Care variable to replace strings with numerics
    c. Replace “N/A” (read as a character or factor in R) with “NA” (read as a missing value in R) in “Video.Share.Status”
    d. Create a binary “video” variable (logic or factor) for whether a Facebook post included a video or not.
    e. Create a binary “link” variable (logic or factor) for whether a Facebook post included a url or not.
    f. Revise the BLM.page variable to be a logic or factor variable.
    g. Revise the “Total.Interactions” variable to be a numeric
2. Run the following descriptive/answer the following questions:
    a. What is the average and standard deviations of the following reactions for each post?
        i. Likes
        ii. Love
        iii. Wow
        iv. Haha

v. Sad

vi. Angry

vii. Care

b. What percentage of posts contain links? What percentage of posts contains videos?

c. How many of the posts were posted on pages run by a self-identified BLM organization?

3. Run a regression where the dependent variable is either total_interactions or Likes. Summarize the results of your regression below your analysis

FAQ:

1) Can I work on a different dataset?

a. If there is another dataset you would like to clean aside from the ones below, please email me for approval (if approved, you will chose only one of the datasets below, as your second dataset will be the personally approved one). The alternate dataset cannot be a TidyTuesday dataset.

2) Help! I didn’t see how to do <thing> in the tutorials!

a. Do not fret! Each dataset comes with its own trials and tribulations. Part of learning data wrangling techniques is learning how to search for the right answer online. Here are some steps you can take:

i. Search for the answer online. It sometimes helps to use “in R” to ensure you get answers for the right language (e.g., “rename variables in r” Google search).

ii. Ask your fellow students on Slack! If you are struggling with something, it is likely that someone else is, too. Posting on the Slack channel will also give me an opportunity to help troubleshoot.

3) I noticed there are some other variables in the dataset that are not listed in the data dictionaries. Can I use those in my ANOVA/Regression?

a. Yes, absolutely! Some of them may require you to look more into different data types, such as dates (see R4DS Chapter 16) or strings (see R4DS Chapter 14).

4) Can I do more descriptive?

a. Always! As a rule of thumb, you should understand each variable individually before you include it in any statistical test. This is important not just for making sure your analysis runs smoothly, but because descriptive can be an important finding themselves!