## 1 MA1 Instructions

Election fraud is a contentious issue in U.S. politics at the moment. Can you use data to cut through the noise and show whether fraud played a role in the 2020 election? For your first main assignment, you’ll answer this question by using a dataset that has state-level (+D.C.) election returns from the 2016 and 2020 U.S. presidential elections, along with measures produced by the conservative think tank, the Heritage Foundation, of instances of confirmed election fraud and measures of state-level election integrity. Your mission is to probe the data to find evidence (or a lack thereof) that states with higher historical frequency of fraud and worse election integrity either show unusual deviations in election returns between 2016 and 2020, or else that fraud and fraud risk tend to occur in states that lean toward one or the other party. *Note that I’m not looking for an answer one way or the other. You look at the data and tell me what your own answer is.*

You’ll consider these issues by producing a series of data visualizations. How you choose to look at the data will be largely up to you, but I have a few minimum requirements. First, you must produce at least ***four data visualizations***. Second, ***at least one visualization must be a map of the U.S. (either a true map or a geo-faceted map) that shows the distribution of one variable of your choosing***.

I also have some requirements for how you structure your submission. You must adhere to the following outline:

1. **Introduction**: In *one* paragraph, clearly state the question/problem you will address, and talk about why it matters.
2. **Data and Design**: In two paragraphs, describe the data you will use in your analysis and how you will use it. In describing the data you should mention: a) *what the unit of analysis is (e.g., “We use a statel-level dataset…”)*; b) *a brief summary of the variables you’ll use (e.g., “Our data is populated with counts of ballots cast for the major parties in the U.S. along with counts of cumulative fraud cases and election integrity scores.”)*; and c) *a quick statement about where the variables in the data came from (e.g., “Election outcomes come from the MIT Election Lab, while fraud and election integrity measures come from the Heritage Foundation.”).* To describe how you’ll use the data, you might say something like: “We plan to cross-reference how the Democratic share of the vote deviates in 2020 from what would be expected given the Democratic vote share in 2016 with a state’s level of election integrity.”
3. **Analysis**: In this section you’ll include your data visualizations and your text walking your audience through them. *You should summarize each data visualization you make in a paragraph that appears before the graph itself.* Don’t just start your paragraph by saying “Figure 1 shows…” Instead, be more creative and try to tell a story with the data. Also, your analysis should consist entirely of data visualizations. *Do NOT report raw R output, data tables, the results from statistical analysis like correlations or regression models, etc.* I will dock points on your submission if you do this. This is a class on data visualization, and this assignment is about honing your ability to make good data visualizations and then to use them to drive a story you want to tell with data. Statistics and models aren’t bad, but this is not the place for them.
4. **Conclusion**: Wrap up your submission with one or two paragraphs that discuss the implications of your findings. Don’t just summarize your findings. Think big and make potentially bold claims about what your findings mean for your audience and society.

Other rules:

* I would prefer it if you use the \_ma\_template.qmd file to format a Quarto document that you then render to a .docx file for your submission. But I understand that this approach might generate distracting issues and force you to troubleshoot rather than focus on your data viz and your writing. It might also make collaboration with your peers a problem. I’ll therefore accept word or pdf documents that you’ve produced the “old fashioned” way. As long as the formatting is 12pt font and double-spaced, this will be fine.
* If you cite external sources, you should cite them by hyperlinking your text. You don’t need to include a formal “References” or “Works Cited” section.
* Make sure your submission includes your names and a title.
* You should title your data visualizations using a convention like “Figure 1: Helpful summary.” Do not use unnecessary designations like “Figure 1a” or “Figure 1b” if you have related figures. They might be related, but they are separate and deserve their own unique number.

## 2 Details about the data

To access the data for this assignment, simply run the following code:

library(tidyverse)  
data\_url <- "https://tinyurl.com/2s3sum5s"  
ma1\_data <- read\_csv(data\_url)

The data contains 51 rows (one for each U.S. state and the District of Columbia). For each observation in the data, there are 13 columns with the corresponding names and definitions:

* state: State names.
* state\_po: State abbreviations.
* dem\_2016 and dem\_2020: Counts of ballots cast for the Democratic presidential candidate in 2016 and 2020, respectively.
* rep\_2016 and rep\_2020: Counts of ballots cast for the Republican presidential candidate in 2016 and 2020, respectively.
* other\_2016 and other\_2020: Counts of ballots cast for third-party presidential candidates in 2016 and 2020, respectively.
* total\_2016 and tota\_2020: Total counts of ballots cast for presidential candidates in 2016 and 2020, respectively.
* fraud\_rank: Election integrity ranking based on Heritage Foundation election integrity scores. Higher scores indicate that a state has worse election integrity.
* fraud\_score: Election integrity scores as measured by the Heritage Foundation. Scores can theoretically range from 0 to 100 with 100 indicating the highest level of integrity and 0 the lowest.
* cum\_fruad\_82\_24: A count of the cumulative number of discovered and confirmed voter fraud cases identified by the Heritage Foundation from 1982 to 2024. Higher numbers indicate more instances of proven fraud.