# Problem Set 9

Due April 11, 2023

### Instructions

- Read all of these instructions closely.
- This problem set is due Tuesday, April 11, 2023 at 4pm.
- Submit files via Github:
  1. the .Rmd (R Markdown) file
  2. the knitted .pdf file
  3. anything else the particular problem set might require
- Use a copy of this file, perhaps with your name or initials appended to the file name, to write your answers to the questions. You'll see there is a designated space where your answers should begin.
- Knitting the .Rmd file to a .pdf file *as you work* will ensure your code runs without errors and is working how you expect. Knit early and often. You've already read the instruction that a knitted .pdf is required when you submit.
- Per the syllabus, I will not accept any late work. Keep in mind the two lowest problem set scores are dropped. Turn in what you have.
- Clarification on the expectations for problem set submissions (posted in Slack, copied here):
  - Always print the output of the code I'm requesting.
    - Ex: If I want you to create a vector x with elements 1 through 10, print x after creating it so I can see it worked.
  - Write any written answers in the space outside the code chunk, not inside with an R comment.
    - R comments are great to clarify code, but not for answering the question.
  - Make sure any code or written content is not cut off in the pdf.
    - This really should only apply to code, because if you follow item 2 in this list, the pdf will compile your written answers nicely.

## Question 1

In problem set 4, you scraped the Notre Dame Political Science department faculty websites to create a database of their contact information, fields of study, etc. In this problem set, you'll use my copy of that dataset, and we'll practice merging two datasets together.

### 1a

To start, read in my version of the the `faculty_df` and `courses_df` objects, where `courses_df` contains information from the department website about the graduate classes being offered in Spring 2022.

### 1b

Complete a left join. Print the dimensions of the result. Explain the results of each join statement in terms of these data. Be very specific about why we got the resulting dimensions.

### 1c

Complete a full join. Explain the results of each join statement in terms of these data. Be very specific about why we got the resulting dimensions.

### 1d

Complete an inner join. Explain the results of each join statement in terms of these data. Be very specific about why we got the resulting dimensions.

## Question 2

Your task is to combine two datasets in order to observe how many endorsements each candidate received.

- Change the `endors` variable name `endorsee` to `candidate_name`
- Filter `polls` to only include the following 6 candidates: Amy Klobuchar, Bernard Sanders, Elizabeth Warren, Joseph R. Biden Jr., Michael Bloomberg, Pete Buttigieg. I've made it easy for you – this is exactly how they appear in the `polls` data without other variations.
- Subset `polls` to the following five variables: `candidate_name`, `sample_size`, `start_date`, `party`, `pct`
- Compare the candidate names in the two datasets and find instances where the a candidates name is spelled differently i.e. Bernard vs. Bernie. You'll need to make these variables comparable across `endors` and `polls` in order to merge.
- Now add poll-level information to the endorsement dataset. Specifically, we want to know the average polling numbers for each candidate from the `pct` variable. Defend the kind of join statement you used.

```
#install.packages("fivethirtyeight")
library(fivethirtyeight)
library(tidyverse)
polls <- read_csv("president_primary_polls_feb2020.csv")
endors <- endorsements_2020 # from the fiverthirtyeight package
```