Analysis Challenge 5

Due Wednesday, April 3

For this analysis challenge, I want you to complete a series of sub-challenges. For each, you'll be confronted with a particular scenario and some **pre-treatment** data and you will have to decide, given the scenario and data, the best kind of **randomized controlled trial** (RCT) to implement to assess the effectiveness of a proposed policy change. In each scenario, the basic setup will be as follows:

1. There is a desired policy outcome with the probability of success represented as S_i and a proposed intervention to be implemented at random, represented as D_i.
2. The goal is to assess how the success rate differs based on the intervention. If we could observe all potential outcomes, our target estimand would be: $\bar{\delta} = \bar{S}_1 - \bar{S}_0$. That is, the difference in mean success with the policy versus without.

Certain factors about the data or the logistics of implementing the intervention will differ by scenario in ways that may bear on your decision about how best to randomize the intervention. In class, we discussed four basic kinds of randomization: (1) simple, (2) complete, (3) block, and (4) cluster.

For each scenario your task will be to do three things.

1. ***Specify what kind of randomization is the most appropriate to use and why.***
2. ***Write the code to implement that kind of randomization using*** `{randomizer}`.
3. ***Write the code to estimate the treatment effect based on your recommended randomization strategy.***

When you're done, render to a word or html document and submit to Canvas.

Here are the packages you'll need:

```
## packages you need for each scenario
library(tidyverse)
library(estimatr)
```

```r
library(randomizr)
```

Scenario 1

The local city government of Columbianapolis is considering redesigning their bike paths throughout the city to cut down on the number of serious and fatal accidents involving cars and cyclists. The city manager's office only has the budget to implement this change in a few neighborhoods in town, but if the redesign is successful the city council may approve more money to expand the redesign to the whole city. To help the city make the best evidence-based decision on whether to expand this redesign, how would you recommend they decide which neighborhoods to start with and then to assess the effect?

Here's the code to generate the data for this scenario:

```r
## scenario 1 data
tibble(
  treatment = NA,
  success = rbinom(1000, size = 1, prob = 0.2),
  street_number = 1:1000,
  neighborhood = sample(
    LETTERS[1:5],
    size = 1000,
    replace = T
  )
) -> data_1

## your code here to assign treatment


## your code here to estimate the ATE
```

Scenario 2

A foreign aid agency wants to know whether offering micro-loans of $1,000 dollars to small business owners in a developing country will help spur the growth of their business. They have only a limited budget for this program, and can only offer these loans to 900 individuals living and working in the developing country's capital city, which is a diverse metropolitan hub for the country with wide disparities in income depending on the neighborhood. What kind of randomized trial would be best to implement in this scenario?

Here's the code to simulate the data:

```
## scenario 2 data
tibble(
  treatment = NA,
  success = rbinom(10000, size = 1, prob = 0.2),
  owner_id = 1:10000,
  neighborhood = sample(
    LETTERS[1:10],
    size = 10000,
    replace = T
  )
) -> data_2

## your code here to assign treatment


## your code here to estimate the ATE
```

Scenario 3

You've decided to recruit some students on campus to participate in a study on cooperation using a game called the Prisoners' Dilemma (read more about it here: Investopedia). You want to figure out if playing the game for points without a material incentive versus with a material incentive changes students' willingness to cooperate with each other in the game. For your study, you'll randomize students into pairs and then have some of them play the game with the material incentive and some without. How will you determine which pairs get this incentive and which will not?

Here's the code to simulate the data:

```
## scenario 2 data
tibble(
  treatment = NA,
  success = rbinom(20, size = 1, prob = 0.2),
  pair_id = 1:20
) -> data_3

## your code here to assign treatment
```

```
## your code here to estimate the ATE
```