Problem Set 6

NAME HERE

DATE

Due: October 12th

For this problem set we will work with fictional data comparing the efficacy of two interventions. The interventions took place across several states and cities, with slight variations in dates. The outcome is a continuous variable.

Question 1

There's missing data in this data set. Can you identify them? In the next question you will re-code these values to NA.

How many NAs did you find?

Are there other values you think may count as NA?

Question 2

For the other values you believe may also be NAs, re-code them as NA.

Question 3

At a glance, we can already see errors with city and state names. Let's first fix these entries to have uniform naming where cities are properly capitalized and states are capitalized. For example, we want to see "San Antonio" and "TX" rather than "san Antonio" and "tx". We want you to use distinct(), pull(), and case_when() for this question.

Question 4

Format the date column into a date format. Ominously, these interventions all occurred on the 25th day of the month.

Question 5

Now we want to fix the city information, but you may realize that we have two cities in California during the same date. We can't, at least from our data, distinguish the difference. Let's drop those rows with this inconsistency. One suggestion is to create a variable indicating whether to drop the row. If you performed this step correctly you should have 33 rows.

Question 6

We have one last issue: our interventions column has missing data. We have two interventions that occurred in these locations:

- Intervention 1: Hayward, Atlanta, San Antonio
- Intervention 2: Oakland, Atlanta, Austin

For all of the cities except Atlanta it's clear what intervention took place. Fix these clear instances. As for Atlanta, we are forced to throw out these observations since we cannot reliably determine which intervention occurred.

How many observations did you drop?

Question 7

We have a few NAs in the outcomes column. Our on-site researchers informed us that when a score of “0” was provided, the data collection team left the cell blank. Re-code the NAs to 0.

Challenge

Create a box plot comparing the two interventions and their outcome. The outcome is a continuous variable from 0 to 10. You may need to factor one of your variables.