February 2018 – Data Analyst Performance Task

Your team has been asked to examine the effect of a school intervention in State X. The intervention consists of teacher peer-coaching and is administered by school districts. More specifically, each participating district identified a team of peer coaches to visit and work with teachers at treated schools – all elementary schools – over the course of the school year. The peer-coaching program was developed by a team of researchers who provided all materials and funding to support the program, but again, districts were the ones who implemented it.

You have a dataset of districts in the state that includes two variables: (1) the state's unique district identifier (labeled "corp1") and (2) the treatment status of the district. These data are provided by the research team. All schools in each treated district received treatment. Although the researchers obviously could not perfectly monitor implementation fidelity, the program was designed so that each school would receive the same treatment.

You also have a dataset provided by the state education agency (SEA X) that includes two outcome variables: (1) the annual test score gain in math (this is the average student-level gain at each school, standardized across schools) and (2) an indicator variable for a "positive work environment" at the school as assessed by SEA X using survey data (you only have the summary 0/1 variable of the SEA's assessment – not the underlying survey data).

The SEA data additionally include basic enrollment and demographic information about the students in each school, and at the district level, a measure of the local-area education level from the US Census (the measure at hand is the percent of residents age 25 and older in the district's zip code who do not have at least a high school diploma).

Here is a brief data key:

Variable	**Description**
district	District identifier
schl1	School identifier
enrollment	Total school enrollment
Asian_pct	Percent of students at school coded as Asian
black_pct	Percent of students at school coded as black
Hispanic_pct	Percent of students at school coded as Hispanic
white_pct	Percent of students at school coded as white
pct_frl	Percent of students at school eligible for free or reduced-price meals
ed_lesshs	Percent of district local-area population with less than a HS diploma
positive_env	Indicator for a positive work environment, as coded by SEA
mathscore_gain_std	Average math test score gain

Your team is at a very preliminary stage in investigating the intervention. **In preparation for the investigation, your team lead has asked that you take a first cut at exploring the data descriptively by completing the following tasks:**

1. **Combine the two data sets. Are there any missing data or other anomalies that your team will need to deal with when conducting the investigation of the interventions effects?**

2. **Using the data at your disposal, describe the characteristics of the districts in which the peer-coaching intervention takes place. Use one or more data displays or visualizations to share the results with your team lead. Briefly explain what the displays or visualizations demonstrate and why they provide helpful context for investigating the intervention.**

3. **Run a standard multivariate regression to assess the predictive importance of the available school and district characteristics for average math test score gain in state X. Which school and district characteristics are significant predictors of math test score gain in state X and which are not? Briefly explain your results.**

4. **The research team has told SEA X that the intervention was randomly assigned. Use the data at your disposal to support or refute this claim. Prepare a data display to share the results of this investigation with your team lead. Briefly explain what the visualization demonstrates about the success or failure of the random assignment process.**