HAD 5744: Quant Methods I
Dr. Hoagland

Assignment #4

Late assignments will be penalized at the rate of 10% per day.

Please submit a knitted R-Markdown file with text and code for this assignment. Don't forget to acknowledge your group members (although each of you must submit your own assignment) and to include your AI statement.

For the problems indicated with an asterisk (*), please submit a video of no more than 3 minutes explaining both how you got to your solution and what it means. Refer to the syllabus for complete instructions here.

1. **Difference-in-Differences: COVID-19 and Sourdough Consumption.** (Adapted from Nick Huntington-Klein and Peter Nencka.) During the early days of COVID-19, there was a brief craze for homemade sourdough bread, as stores were out of yeast (sourdough can be made at home using yeast from the air and does not require store-bought yeast). We will be estimating whether COVID lockdowns actually increased interest in sourdough bread.

   We will be measuring interest in sourdough bread using Google Trends data in the USA. The data is on the course website and is saved as "*a4_p1_sourdough_trends.csv*". Google Trends tracks the popularity of different search terms over time. This data has the popularity (measured in `hits`) of different search terms, including "sourdough" (defined in the `keyword` variable).

   a. Select a set of covariates that you think are relevant to this relationship. Clean the data, including selecting the appropriate measurement for each variable, any transformations needed, and any procedures to deal with missing data. Describe and justify your choices.
   b. Visualize the popularity of each keyword over time (with a separate trend for each keyword). In your visualization, add a vertical line for the "start of the pandemic" which we'll decide for our purposes is March 15, 2020. First, make a graph showing the popularity of "sourdough" changing over time (using the `hits` variable). Describe what you find – does it lend support for a particular hypothesis? What kind of treatment effect are we looking at here?
   c. Suppose that we wanted to use `keywords` other than sourdough in our control group. What assumptions are needed for that to be acceptable in a difference-in-differences context? What additional (if any) assumptions are needed to make this regression report the causal effect of COVID on sourdough popularity?
   d. * Determine what control group you will use and justify your choices, statistically if possible. Report your estimating equation and describe it as you would in a published paper. When doing so, don't forget to discuss the optimal standard error calculation.
   e. Estimate the difference-in-differences model given your setup. Report and interpret your results. What do you conclude about the effect of the COVID-19 pandemic on sourdough bread consumption?