HAD 5744: Quant Methods I
Dr. Hoagland

Assignment #4

Late assignments will be penalized at the rate of 10% per day.
Note there is a hard cutoff for submission of this assignment on December 8, 2025

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. First, make a graph showing the popularity of "sourdough" changing over time (using the `hits` variable). Run a simple regression estimating the change in popularity between the pre-period and the post-period. Write down your regression specification (in an equation) and interpret your result.
   b. * Is this regression coefficient the causal effect of COVID on sourdough popularity? Why or why not? Ideally, what data would you need to pin down this effect?
   c. Now update your time series graph with the popularity of the other search terms in the data, with a separate line for `keyword`. Also add a vertical line for the "start of the pandemic" which we'll decide for our purposes is March 15, 2020. Describe what you find – does it lend support for a particular hypothesis? What kind of treatment effect are we looking at here?
   d. * Suppose that we wanted to use these `keywords` for 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?
   e. Formally test whether trends in popularity prior to COVID differed across your treatment and control group. What do you find? If you are concerned, make adjustments to your control group and defend them.
   f. * Update your regression equation from part (a), and estimate a difference-in-differences model (where the treatment occurred on March 15) using that specification. What are your two sets of dummy variables (called "fixed effects")? Report and interpret your results. Cluster your standard errors at the `keyword` level.
   g. What do you conclude about the effect of the COVID-19 pandemic on sourdough bread consumption?