# Causal and Predictive Analytics – Homework 1

Individual Assignment

This assignment will have you work directly with real-world experimental data. In this assignment, you will be working with a dataset from a landmark study on the effectiveness of online search advertising. eBay conducted an experimental study to determine the return of its advertising on search engines including Google, Bing, and Yahoo.

You will be graded both on your code, and the written answers you provide. When evaluating the code, the grader will take on the role of an eBay co-worker. Code will be evaluated both in terms of how correct and how clear it is. By correctness, I mean that the code fulfills the requirements of the question. By clarity, I mean that the grader should be able to understand what your code does within 30 seconds of reading it. As discussed in class, this is aided by clear comments, good variable names, proper indentation, and short lines.

The written portions will be evaluated based on the strength of the arguments presented, the use of data to support your statements, and the quality of the writing. Do not copy, share, or otherwise communicate your written answers to other students.

All answers should be submitted using the posted Rmarkdown template. Open the .rmd file in R-Studio, and follow the directions listed there.

## Assignment Materials for Download:

1. An Rmarkdown template titled 'Homework1Template.Rmd'
2. A data file in .csv format titled 'Homework 1 Data - 436R.csv'

## Submission Checklist:

To help us grade the assignments efficiently and correctly, we ask that you submit your assignments in a specific format. A complete submission for this assignment will submit the following via blackboard:

- A .rmd Rmarkdown file, based on the template for this assignment.
- A .html file, generated by knitting the .rmd file in RStudio.
- If you choose to use Generative AI, please use a single session of ChatGPT, and get a link to the chat through the "Share Link to Chat" button. Submit this on blackboard while submitting your assignment
- 
- Make sure your files have the appropriate extensions

- ***Zip*** all files into a single archive before submission. Do not use app

- Put your written answers in the textboxes in the submission link

## Data Dictionary:

- `date`: Date of advertising
- `DMA`: Designated Market Area Code. Basically a city
- `isTreatmentPeriod`, `isTreatmentGroup`: Dummy variables denoting whether date belonged to the treatment period, and DMA belonged to the treatment group
- `revenue`: Revenue for the DMA in dollars

## Part 1: Analysis (16 marks)

The study was conducted as follows. Users were categorized by their designated market area (DMA), which is given as a categorical variable in Column 1 of the dataset. DMAs were randomly selected to be in the treatment or the control group. The variable `isTreatmentGroup` indicates that the DMA was placed in the treatment group. After a certain date, the treatment period started, and the treatment group was no longer shown search ads from eBay. The variable `isTreatmentPeriod` indicates whether the treatment period had started.

This analysis follows part of the study. Please do these questions in order. Download the Homework 1 dataset and save it to your computer. You can complete this section using Boolean variables, and the `read.csv`, `lm`, `summary`, `log`, `subset`, `as.Date`, and `sort` functions. As a reference, you can consult the 'Interview Case' presented in class.

a) Load the dataset in R. *Hint: Use the* `read.csv` *function.*

b) Write code that will display the first 10 rows of the dataset in the console. *2 marks*.

c) Determine the date that started the treatment period. That is, write code to determine the earliest date in the treatment period. *Hint: Use the* `subset` *and* `sort` *functions*. *2 marks.*

d) The data contains a control group, which was shown search ads throughout the data, and a treatment group, which was only shown search ads before the treatment period. *4 marks*

i. Take a subset of all the data from the treatment group.
ii. Run a regression that compares *log(revenue)* of the treatment group in the pre-treatment period and the treatment period. *Hint: the independent variable should be* `isTreatmentPeriod`
iii. Display a `summary` of this regression

e) Now we will use the control group for a truly experimental approach. First, we will check to make sure that the randomization was done properly. *4 marks*
   i. Take a subset of all the data from the pre-treatment period
   ii. Run a regression that compares *log(revenue)* of the treatment group and the control group in the pre-treatment period.
   iii. Display a `summary` of this regression

f) Now, using the treatment period data, determine the effectiveness of eBay ads. Run a regression with *log(revenue)* as the dependent variable, and control for whether the DMA is in the treatment group. Display a summary of this regression. *4 marks*

# Part 2: Discussion (20 marks)

Please provide written answers to each of the following questions in the blackboard link. Answers will be judged on the accuracy, and correct spelling/grammar. Pay close attention to what each question is asking for and the course materials. As a reference, you can consult the 'Interview Case' presented in class. Each answer only requires a short response (**45 words max. Additional words will be deleted.**). Please use a spelling/grammar check before you submit.

a) In part *1d,* you ran the analysis without a control group. What do the resulting coefficient estimates say about the effectiveness of advertising? Use quantitative terms, and account for statistical uncertainty. *4 marks*

b) What is the purpose of the randomization check in part *1e*? What do the results of this analysis show? *4 marks*

c) In part *1f*, you ran the analysis with a control group. What do the resulting coefficient estimates say about the effectiveness of advertising? Use quantitative terms, and account for statistical uncertainty. *4 marks*

d) What does the control group allow us to control for? What specific omitted variables might have caused bias in part *1d*, but wouldn't in part *1f*? *4 marks*

e) Write down the R-Squared of the regression in part *1f* (you can see it with the `summary` function). Does this affect the interpretation or confidence in the estimate of the effectiveness of advertising? *4 marks*