Assignment 02

SIAS, Krea University (AY 2025-26)

Measuring Impact with A/B Testing (Course Code: **ECON345/BUSI333**)

Instructions and Advice:

- This assignment accounts for 10 percent of your final grades.
- The soft deadline for submission is Saturday (22 Nov 2025) 11.59pm.
- The hard deadline for submission is Monday (24 Nov 2025) 11.59pm.
- The objective of this assignment is threefold:
  - Write clean R script.
  - Produce pretty graphs using `ggplot`.
  - Generate a balance table.
- Please submit your R script, and a PDF with all the answers.
- The format of your files must be Assignment02_xxxx.R, Assignment02_xxxx.pdf where xxxx is your four-digit roll-number.
- I will appreciate it if you can highlight your main answers.
- I am okay with you using LLM, but this usage must be fair. If you use an LLM for the coding problems/issues, please ensure that you have added the prompt that got you the code.

Question 01 (30 Points)

Access to transport can dramatically change the distribution of firms in an economy. The economist Wei You analysed historical data for Boston to measure the impact of the switch from horse-drawn streetcar to electric rail. The comparison is being made between Charlestown and East Boston. While Charlestown was connected to central Boston by bridge, there was no direct connectivity between East Boston and the city center. In this question, you will replicate Panel A of Figure 3 from the paper.

The dataset is here. Please store the file to your data subfolder, and read it in R. After having read the data, filter the data for 1880 and 1885. Once you are done with this step, please perform the following steps:

(a) (**2**) Based on the column `Pshr`, create a new column called `ctgr` with the following rule:

- Food if Pshr = "GRO", "BAKER", "FISH", "FRUIT", "PROV", "PROD", "CON", "LIQ"
- Clothing if Pshr = "CLO", "B.S.", "DRYGOODS", "HAT", "MENFURN", "MILLINER", "TAILOR"
- Other Products

(b) (**2**) Create a column for distance band `dist` based on the column `DistToCBD`.

- $0 - 1km$ if the distance is less than 1
- $1 - 3km$ if the distance is ≥ 1 and less than 3
- $> 3km$ if the distance is ≥ 3

(c) (**1**) Change the rows within the column `treat_band` by `2-treat_band`.

(d) (**5**) Calculate the average `Sole` by `year, ctgr, dist, treat_band` and call it `mSole`. Drop if the number of observations within any group is less than 20.

(e) (**4**) Reshape the data wide by `ctgr` and `year`. Create a new column called `mSole_change` which measures the difference between average sole proprietorship in 1885 and 1880.

(f) (**16**) Create the graph. Please ensure that the graph appears in the PDF.

Question 02 (30 points)

A new study assesses the effectiveness of AI-led approaches to replicate research in social sciences. This study finds that:

> human teams matched the reproducibility success rates of teams using AI assistance, while both groups substantially outperformed AI-led approaches (with human teams achieving 57 percentage points higher success rates than AI-led teams, $p < 0.001$). Human teams were particularly effective at identifying serious problems in the analysis: they found significantly more major errors compared to both AI-assisted teams (0.7 more errors per team, $p = 0.017$) and AI-led teams (1.1 more errors per team, $p < 0.001$).

This was an experiment that divided teams of researchers into three groups. Our aim, in this question, is to recreate the balance table using R. Please pick any five randomly selected variables from the list of columns and reproduce the balance table for these variables. The data is available on Canvas as a csv file called 'a02_q02.csv'. The treatment variable is called 'branch' and the description of the rest of the variables is stored in an excel file. Please make sure that the final table appears in the PDF.