Stat 27850/30850: Group project # 1

Data The data set for the first project is the Chicago Taxi Rides data set, which was obtained from `https://www.kaggle.com/datasets/chicago/chicago-taxi-trips-bq`. This data set aims to record all taxi rides in the city of Chicago from 2013 onwards. On Canvas, we provide a subset of the data, `taxidata.csv`, which contains 200,000 data points each from 2014 and 2019. These data points are randomly sampled from the set of all complete (i.e., no missing values) data points from each of those years. Each data point includes information on the taxi company, the start and end location of the ride, the ride duration in distance and in time, the total cost of the trip (split into: the original fare, the tip, any road tolls, and any extra charges), and the type of payment (cash/credit/etc). Note that you may need to do some data cleaning if you observe values in the data that are obviously incorrect or corrupted.

Assignment Your task is to study the following question: can you detect any *interesting changes in customer behavior* between the two years? For example:

- Are customers using taxis for late-night rides more frequently in 2019 as compared to 2014?
- Are customers less likely to tip for short rides in 2019 as compared to 2014?
- Are customers picked up on the Magnificent Mile less likely to use cash in 2019 as compared to 2014?

These are just examples—please design your own questions to find an interesting aspect of the problem to study. Note that we are not interested in changes that do not reflect customer behavior (e.g., due to changes in traffic patterns, increase in taxi price, etc), so you should be sure that your question does not reflect such changes. For example:

- Are customers leaving higher tips? This might just be due to increase in fares.
- Are customers taking rides with a longer time duration? This might just be due to increase in traffic.

Your goal is to find questions that involve interesting challenges of multiple testing. For instance, instead of asking whether tips have increased overall, we might look for specific locations / specific times of day / etc that show a change in tipping behavior. When designing your questions, be sure to choose topics for which the emphasis is on *inference* rather than on modeling and estimation.

For any question you ask, you will likely need to control for covariates/confounders: for example, if looking for an increase in tipping on a particular route, you may need to control for time of day / duration of the ride which reflects traffic / increase in fares / etc. You can consider strategies that incorporate tools such as matching, permutations, regression, etc—whatever tools you feel are appropriate. You are not required or expected to research methods beyond the scope of what's covered in the class, but you are welcome to use whatever statistical tools you choose.

Your final report should give a thoughtful discussion of the issues you have identified, and the strategies and methods you developed to address them. There might be confounding effects you identify that it's not possible to address with the limited data available (or simply due to time constraints); your report can also mention issues that you were not able to address, and assess to what extent you think it might affect the validity of the analysis. Depending on your approach, it may happen that you are or are not able to detect significant changes. Either outcome is fine, as long as your conclusions and methods are well explained and justified.

Guidelines Groups of size 2, 3, or 4 are allowed for the project. The extent of the project (e.g., the range of questions explored / methods tried / etc) should be proportional to the size of the group. What you hand in:

- Each group should hand in a written report and either include code throughout the report and/or include the code as an appendix. Please designate a single group member to submit everything on Gradescope, and add the other students in the team group members.

- For your code, it should be clearly organized and commented— for example, you may want to label sections of the code so that we can see which part of the report or which plot/table it corresponds to, add comments to explain steps where notation / variable names / nature of the calculation aren't obvious, etc.

- There are no page length or formatting requirements for the written report. Your report should describe the problems and questions you posed, the details of any methods you implemented / models fitted / hypotheses tested, describe your findings and show plots or numerical results as appropriate, and should discuss some interesting issues relating to inference (for example, multiple testing / appropriately controlling for confounding factors / reducing a high dimensional model to a manageable size / etc). You can also include a discussion of open questions and issues that were not addressed (due to time limitations and/or limitations of the available data).