Using Exploratory Questions to Better Understanding Your Problem Unifying Data Science 2024

Nick Eubank

January 24, 2024

Due February 13th, 12pm

Over the course of the semester, you and your team will be asked to choose a problem you care about. You will then work to help solve that problem (you get to be your own stakeholders!) by specifying and answering a set of Exploratory Questions, a set of Passive-Prediction Questions, and a set of Causal Questions.

Your Assignment

Your team's first assignment has several components:

1. You must identify an *issue area* you care about.
2. Your team will be paired with a second team with similar issue interests (your Partner Team). You and your Partner Team will then be required to specify a problem that matters to you.
3. You must then write a report in which you:
   - Specify and answer (by analysis of real-world data) at least three Exploratory Questions.
   - Specify and answer at least one Exploratory Question using existing reports and analyses (i.e., you do *not* have to find and analyse raw data to answer this question — if someone's already answered a question, why reinvent the wheel?)
4. Your team will then trade reports with your Partner Team and provide feedback to one another.

Part 2: Choosing a Problem

The Problem you choose to address can come from any domain and can be an issue of global importance or personal interest (provided you can get the rest of your team and Partner Team to also agree to work on it). Examples of the types of problems and Exploratory Questions that you might wish to answer to help improve your understanding of the problem include:

- *Problem:* Too many people are killed in car accidents.
  Exploratory Questions:

  - What share of car-related fatalities is due to car-pedestrian, single-car, or multiple-car accidents?
  - What share of car-related fatalities occurs on freeways as opposed to in cities?
  - What share of car-related fatalities involves a driver under the influence of drugs or alcohol?
- *Problem:* Many states are adding bureaucratic hurdles to getting social services, but the effect of these hurdles is unclear, both in terms of their effect on reducing fraud and on deterring entitled recipients from getting aid.
  Exploratory Questions:
  - What states have changed their rules around social service provision (helpful if we want to do a pre-post analysis or a difference-in-difference analysis)?
  - Are there a lot of people who are entitled to social services who don't receive them? (Do we know?)
  - What social service programs that have had bureaucratic hurdles imposed serve the largest populations?
- *Problem:* Police shootings involving people with mental health issues are much too common, and it's not clear the police are appropriately trained to deal with people dealing with mental health crises.
  Exploratory Questions:
  - What states have changed their rules around social service provision (helpful if we want to do a pre-post analysis or a difference-in-difference analysis)?
  - Are there a lot of people who are entitled to social services who don't receive them? (Do we know?)
  - What social service programs that have had bureaucratic hurdles imposed serve the largest populations?

Note that while the problems in these examples are all "big" problems—in the sense of being societally important questions—your problem need not be of this nature. Past teams have done projects trying to figure out how to optimally train in tennis (by looking at whether playing more tennis improves or hinders subsequent tournament performance), how to improve AirBnB host profits (by looking at whether "super host" status improves AirBnB host revenues above and beyond the effect of just having the features that make one eligible to be a super-host), and how to minimize cell-phone user churn.

With that said, be aware that **you will need to find data about your problem**, and that *is* often easier to do when wrestling with public problems; companies tend to be quite protective of their commercially sensitive data.

Part 3: Specifying and Answering Questions

To help, you can find

Data That ISN'T Allowed:

One major rule about data: data from Kaggle or other competition sites is not allowed. Part of the goal of this class is to give you experience picking your own questions, and then having to find, sort through, and clean real-world data to answer those questions. Kaggle data and the like are pre-cleaned, and come with JUST the variables needed to answer specific pre-determined questions. So absolutely no pre-curated data from competitions.