In-class activity

Class 8

Federica Fusi

MSCA, UIC

Updated: 2021-03-02

Before we start

Quick announcements

I have a few revisions on the syllabus for the second part of the class.

- In the short term: We are temporarly eliminating the "Relational datasets" session and we will cover "Reproducibility and communication" next week instead (e.g., R markdown) - Class 9

- Class 10 will cover indicators, variables, and scales and will introduce packages to work with strings and factors, to be continued in Class 11.

I will upload an updated version of the syllabus.

There is no change in assignment due dates nor evaluation criteria.

I will talk more about the final project next week.

In-class activity

Goals

By working on this activity, you will be able to:

1. Review content discussed in the past weeks
2. Conduct the different stages of a data analysis (data cleaning, exploration, visualization)
3. Practice your coding (and ask questions in real-time)
4. Develop good questions for data analysis

Final ouputs

You will submit an R script as a group + images of your data visualizations by Monday, March 8th at midnight.

The R script should be well annotated and provide adequate comments to each phase of the analysis.

The R script should be reproducible.

The activity is worth 10% of your final grade.

Data

There are four datasets available to you

1. Crime data in 2010 and 2019
2. Demographic data on each Chicago ward in 2010 and 2016
3. Income data on each Chicago ward in 2019

Four steps

You are expected to complete the following activities:

1. Data cleaning
2. Define your questions
3. Exploratory data analysis
4. Data visualization for your audience

1. Data cleaning

In this phase, you need to make your data ready for analysis - this includes preparing a tidy datasets, clean up column names, check for missing data, ...

Develop a data cleaning plan before starting. What needs to be done? What issues do you need to address to make your data ready for the analysis?

Write your plan down before moving onto the next phase.

2. Define your questions

Brainstorm a few questions that you would like to answer as a group. A "good" question is:

- Interesting to your audience (define your audience: policymakers, board of a nonprofit, citizens...)
- Plausible – the question should make sense according to prior research and knowledge of the topic
- Answerable – you have data available to answer your question
- Specific – the question should be precise in what is asking. A specific question helps to make your plan clearer. Moreover, the answer you will get is more interpretable and can lead to policy implications. (“is crime high?” vs “is crime higher than the city average?”)

2. Define your question

There are also several types of questions. We generally focus on two types of questions:

Descriptive questions seek to summarize a characteristics of a set of data (e.g., which wards have the highest number of crimes?)

Exploratory questions analyze the data to see if there are patterns, trends, or relationships between variables. These types of questions are called "hypothesis-generating" analyses because rather than testing a hypothesis, you are looking for patterns that would support proposing the hypothesis.

You need to have at least one exploratory question

2. Define your question

Other questions include:

Inferential questions are statements of a hypothesis which is then tested on a representative sample (statistics focuses on these questions)

Causal questions test the cause-effect relationships stated in the hypothesis. You need appropriate statistical tools to establish causality (experimental methods focus on these questions)

Predictive questions seek to understand individuals, organizations, and so on that are more likely to do something – you are not interested in explaining why something happen but only to predict whether it will happen (machine learning focuses on these questions)

Mechanistic questions focus on how a certain cause leads to certain effect (several qualitative methodologies apply here).

3. Exploratory data analysis

Three goals:

1. To determine if there are any problems with your dataset
2. To determine whether the question you are asking can be answered by the data you have
3. To develop a sketch of the answer to your question

By the end of this step, you should have an expectation of what your dataset looks like and whether your question can be answered by the data you have.

3. Exploratory data analysis

3.1. Check out descriptive statistics and distributions of main variables - start from the one that you are interested in.

3.2. Explore and assess your questions

- Do you have the "right" question?
- Your question might not be plausible, specific enough, or answerable. It might also happen that results are not as interesting as expected (e.g., there is no variation)
- Step 2 (Define your question) and 3 (exploratory data analysis) are iterative: you might go back and forth between them.

4. Data visualization

Data visualization is here used to communicate with your audience. There are several reasons why you might want to communicate your findings to your audience:

- Show your results
- Get help to work through some puzzling results
- Get general impressions and feedback on your work

Create at least two data visualizations that will help you achieve those goals. Along with the final plot, provide enough text to explain what you are visualizing (e.g., your question), how the audience should read your plot, and your main findigs.

Suggested work plan

This is a group activity and you have approximately two hours of class time to start working on it.

Things you should do by the end of the class:

- Set up a shared folder to work on your code and outputs as a group (box, google drive...).
- Develop a data cleaning plan
- Develop questions that you would like to address and assess their feasibility
- Sketch of the explanatory plots

After today, you can split these activities but you will have to submit one unique R script.