In-class activity: Review of Week 1 to Week 7

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

If needed, review Chapter 1 to 4 in the Art of Data Science book by Peng and Matsui, which covers in more details each of these phases.

Final outputs

You will submit an R script as a group + your data visualizations as images 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 needs to be reproducible.

Work plan

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

- Set up a shared folder to work on your code and outputs as a group (box, google drive…).
- Develop a data cleaning plan
- Develop a few questions that you would like to explore – what is your question? Do data allow you to answer this question? What data do you need?
- Refine your questions
- Sketch out your data visualization

You can split activities after today

but the final product should be one well-organized R script.

Data

These are the dataset 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

See codebook for more details about the data.

1. **Data cleaning protocol**

Open the three datasets in Excel or R. Develop a plan of action on how to clean the datasets before the data analysis.

- Write down your plan of action (or data cleaning protocol). This is a detailed plan of both problems that you see in the dataset and your actions to address them.

Conclude this section by checking that data are correctly organized (e.g., look at the first few rows, check rows and column numbers as you move to one step to the other, and so on).

A couple of notes:

- Make sure to identify the unit of analysis for each dataset
- You can eliminate all columns including geographical coordinate information as we won't need them
- Crimes_2019 contains a few observations from 2020. Eliminate them and keep only observations from 2019.
- You will probably update your data cleaning section as you move forward

2. Formulate your question

Brainstorm with your group a set of questions that you would like to answer by looking at the dataset.

Consider descriptive and at least one exploratory questions – remember: an exploratory question should lead to an analysis of the data to see if there are patterns, trends, or relationships between variables.

A good question is (Peng & Matsui, 2018):

- *Interesting to your audience* (make sure to 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
- *Specificity* – 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.

3. Exploratory data analysis

There are three goals to exploratory data analysis (Peng & Matsui, 2018):

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.

Step 3.1 - Familiarize yourself with the dataset. Explore your variables both by looking at summary statistics and visualizing them on exploratory plots (e.g., histograms, density plots, bar plots, boxplot…). You should focus on variables that are of interest to you based on your question.

At each time, provide short comments of your results and observations. You should check if your expectations are met at each point (i.e., does the income distribution make sense?).

Step 3.2 – Conduct some more in-depth analysis. Address your questions and refine them. Do you have the right question(s)? Step 2 and 3 are iterative. At this stage, you might need to refine your question to make it more specific. You might also find other questions that are more interesting to your audience.

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 (Peng & Matsui, 2018):

- 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. Make sure to provide enough text to explain what you are visualizing (e.g., your questions), how your audience should read your plot, and your main results.

Remember (Peng & Matsui, 2018):

1. Keep your audience in mind
2. Be focused and concise so that your audience will understand your message
3. Avoid jargon and technical terms
4. Your goal should be to get inputs onto your work rather than defend your analysis and question