ETC5512: Care and feeding of open data

Lecturer: Kate Saunders

Table of contents

Learning Objectives

- Reflect on the open data resources that you have seen in this unit. We'll consider factors that make this data easy or difficult to use.
- Place yourself in the role of data curator, based on your experiences as a user think about best practices for data sharing

For today's tutorial complete Tasks 1 and 2 in groups.

Task 1

This task is designed to mimic the questions on your assignment 4, and for you to consider this week's data curation lecture.

i. Pick one of Assignment 1, 2 or 3. What was difficult working with the data on your assignment? (Beyond the difficulties in compiling your report!)

- Reflect upon what aspects of **your** analysis involved data work that weren't sexy

Here is an example from Assignment 2.

The G17 census data is contained in three different spreadsheets. These are G17A, G17B and G17C. To work with this data, these spreadsheets first need to be read in and combined together.

The format of these spreadsheets is also not tidy. Column names encode information about gender, income brackets and age brackets (e.g. F_300_399_55_64_yrs). These columns need to be split so that there is a column for each of these separate variables.

The age brackets and income brackets also contain top and bottom coding. This requires a large amount of string handling to fix that is particularly tedious.

- Reflect upon any challenges **you** faced in your completing your analysis

Part of this analysis required filtering first preferences won by the non-major parties. This is trickier than what it appears.

Firstly there is an issue with data harmonisation. Party acronyms like ALP may be written with full stops, A.L.P, and this prevents exact matching.

There are also issues where the Labour party has slightly different party names and acronym variants for branches in different states. This means an exact matching with ALP or Labour does not work. Instead a fuzzy matching is needed.

And even with a fuzzy matching there were still edge cases that needed manual handling. This meant it was very difficult to automate this analysis.

If I was to approach this problem again I would create a reference data frame containing all variants of party names and abbreviations, along with a variable classification of major, minor or independent party.

- Reflect upon the imperfect aspects of **your** work or how you'd like to improve your analysis in the future.

Each SA1 region has slightly different characteristics. Some of these have zero median age, as no one lives there. I also found out some SA1 regions in WA have an area of 0 square metres. This is very odd! Finding artefacts like these makes me think I should search this data more for other types of peculiarities.

While this isn't strictly necessary, as taking a median to be robust to a few SA1 region with a 0 median age. I would still like to improve the screening of which SA1 regions were being including in my analysis.

In terms of future work, I'd also like to consider how changes in electoral boundaries impact voter demographics. This is very interesting and I'm almost certainly going to set this as the assignment question next year! Lucky ETC5512 students.

ii. What would you as data curator do to make the data drudgery easier, detective work or imperfect parts of your analysis easier? (Think about the next wild caught data student!)

- *I would create a data set taht contains information about all past parties that could be used to cross reference with the 2022 election data. This would help people understand whether a party was major, minor or independent. As part of this I could also include references to specfics that might be important for understanding the data and its context.*
- *I would create a set of rules about my expectations for SA1 regions. For example people must live there and there must be some area. This would allow me to screen for SA1 regions that might need investigation for use in a given analysis.*

Task 2

i. Brainstorm 3 questions you could try to answer on assignment 4. Also think about where you would need to source the data from to answer your questions. Remember brainstorming is about creativity, so don't censor your ideas yet. Just throw them out there.

- *How diversely do I read? An analysis of author nationalities in booklists.*
- *How long do people last on the TV series Alone?*
- *Is it true that most small businesses in Melbourne don't survive two years?*

ii. Share you ideas with your peers. Work together to evaluate if the questions would be easy, reasonable or hard to answer. In answering this, think about the scope of the assignment and challenges you may face with the data.

In this task have fun exploring your different ideas together and be respectful of each others ideas and process.

- *Refer to Lecture 10 to see why Question 1 on authors is hard to answer! There are definitely easier questions I can ask, and I should have stopped when I realised web-scraping wasn't going to provide an adequate data set for analysis as I had lots of missing data.*
- *Question 2 is much easier to answer! The data is formatted in nice tables on wikipedia. The challenge with this question will be showing appropriate depths of skills and analysis. I might need to find ways to challenge myself to get good marks.*
- *Question 3 seems approachable if I can find good data. There is an open data set on businesses in Victoria that could be used to answer this question. I may find filtering out the different types of businesses challenging, but this seems doable. It likely will also lead to interesting analysis when coupled with looking at trends in space and time. But I haven't tried - so that will be part of the adventure.*