A checklist of questions for data analysis

Online datasets should always have companion notes that explain the content. Think about a footnote or endnote that reveals the source of the information or provides context. For datasets, these notes come in the form or so-called "readme" files, or data dictionaries. If these sources of information are unavailable, ask for them. But you should also be asking yourself a lot of questions about the data before diving too deeply into analysis. These is by no means an exhaustive list but should get you started.

1. **Who produced the data?** In an age when governments at all levels are turning to third-party organizations to handle administrative tasks such as data collection and storage, its essential to know who's in charge.
2. **Why is the government releasing the data?** This, too, might not be apparent. For instance, we know why Statistics Canada releases census information every five years, or the Ontario government releases the so-called "sunshine" list of public sector salaries over $100,000. However, if the intent is unclear, don't be afraid to ask.
3. **When was the data collected?** Knowing the answer to this question allows you to judge the dataset's currency. If the data is a few years' old, ask if there's anything more recent.
4. **How frequently is the dataset updated?** A good open-data site should indicate the frequency of updates. If it doesn't, ask.
5. **Could there be errors?** There's an ungrammatical saying among data journalists that "all data is dirty." Not because faceless bureaucrats are trying to deceive us, but because those bureaucrats are human beings who may simply make mistakes when keying in all those numbers and bits of text. Names on the sunshine list are inevitably spelled differently. John C. Smith may become John C Smith. He's the same person, without or without the period after his middle initial.
6. **Do the numbers like grand totals reflect the actual sum of their parts?** Even Statistics Canada must correct data after the fact.
7. **Is there missing data?** At first glance, you'll have no way of knowing. But follow your logic. In a salary disclosure database like Ontario's, it makes sense to know the employee's name. Not so with Alberta salary and severance disclosure for government employees. On this preview list we can see the employee name. However, we click on the "Alberta Open Government Portal" link to preview the version available for download, we see the names seem to have disappeared. Confused? Join the club.
8. **Has data been removed?** In the previous example, it's obvious what's missing or has been removed. The goal of this question is to remind you that most, if not all datasets, on public sites are subsets of much larger data. For instance, this medical device extract is a subset of Health Canada's so-called relational database containing hundreds of tables full of information about medical devices. I know because I negotiated for the data for Implant Files, the last data-driven investigation of which I was a part before eventually leaving the CBC. The extract the department is now posting on its website, is what we made available as part of the investigation.
9. **Has the data been cropped, filtered, or aggregated?** The answer to this question is typically, yes. For instance, Statistics Canada aggregates information in datasets such as the labour

force survey to protect privacy. So, we know that young women usually have lower unemployment rates in jurisdictions such as Ontario. But that's about it. Or a police force may tell us in which neighbourhood block a crime occurred, but not the address. Or the Office of the Superintendent of Bankruptcy will provide the first three letters of a postal code, the forward sortation area (FSA), but not the actual location of the individual who has run into serious money problems. There's nothing wrong with aggregating data to protect privacy. Still, it's worth knowing when this is the case, which helps your analysis.

10. **How does the institution use the data?** Journalists and bureaucrats use information from datasets differently. Take medical devices. For Implant Files, among other things, we wanted to know which medical devices were the most dangerous. For its Power Gap series, The Globe and Mail wanted to look at the wage disparity between male and female executives. In both these instances, the federal and provincial institutions, respectively, may simply use the datasets to find specific information about a particular medical device or bureaucrat's salary.
11. **How representative is the dataset?** This is an essential question. While a sunshine list may cover ALL public sector workers in provinces such as Ontario and Alberta, Health Canada's medical device database only covers a fraction of actual incidents, anywhere from one to 10 per cent, according to the experts we interviewed. Indeed, underreporting is a huge problem, not only with medical devices or prescription drug adverse reactions, but any kind of incident-based reporting system. Knowing the extent of the underreporting keeps us from overstating the importance of our data analysis.
12. **What period does the dataset cover?** Statistics Canada's Labour Force Survey dates all the way back to 1976. And if your computer's hard drive has enough space, you can download the entire dataset for significant analysis. However, many datasets other federal, provincial, territorial or municipal datasets may ONLY cover a few years. Remember, departments make choices about what to make public, based on factors such as the need to protect privacy, or safeguard data consistency which may be impossible by merging old datasets to newer versions. Limiting the number of years may be part of that decision-making process. So, it's worth asking.

Source: The Data Journalist