Tips for building your own data:

- Make sure the data you want isn't already available in a structured format. Even a PDF with a table-like structure can be turned into a spreadsheet; and data stored on a web page can often be scraped.
- Don't rush into entering data! Take some time to think carefully about what you're going to include in your dataset, even down to seemingly minor details like what values will be used in each field.
- Even after you've carefully made your plan, give yourself some time to enter a few records and then assess whether your plan is working or not. Make changes, if necessary.
- Try to include something in your data that identifies the original source, in case you have to go back. For example, if it's a police report, list the name of the PDF and/or the case number. Putting down page numbers can also be useful if the information comes from a very large PDF or paper report.
- When figuring out what fields to include in your database, first think about what you hope to ultimately say in your story. For example, when we built a database of police reports on sex assault investigations, we knew from the start that we wanted to be able to say "Only X% of the cases were sent to prosecutors." So we knew we needed a field that said either "yes" or "no" regarding whether that case was sent for prosecution.
- Also think about what "buckets" you might want to put the data in for analysis purposes. This might mean you need to add something to your data that comes from an alternative source. Let's say you collected information from a variety of cities on some topic or another. Perhaps ultimately you want to say something like "We found X was true in cities outside the metro, but not as prominent in the metro area." Then you would want a field in your database that identifies the city as either "metro" or "non-metro".
- Try to limit how much typing is necessary in the data entry. Use fields with pull-down menus that limit the options wherever possible. This requires more upfront setup, but it will help ensure consistency from one record to another and reduce how much data cleanup you need to do at the end. You don't want one record to say "Minneapolis" and another to say "Mpls", for example.
- Each row should represent one thing and sometimes you'll need to think carefully about what that thing is. For example, let's say you're building a dataset of homicides in your city. You want to track information about the victims (race, age, gender, etc), details about the crime (weapon used, relationship of victim and perpetrator, location) and whether the case is solved, if anyone has been prosecuted/convicted, etc. This one is tricky because there might be multiple victims from the same incident. Should you have one row for each incident or one for each victim? There isn't a single perfect answer to this; it will mostly depend if you want to focus on the incident or on the victims. You

could go with each incident and have limited victim information. For example, instead of listing the race of each victim you might have one field that says either yes or no indicating whether one or more of the victims was a person of color. This would allow you to analyze whether homicides involving victims of color are less likely to be solved. You could have a separate sheet that has one record for each victim, using a case number to link your "incident" sheet to your "victim" sheet. The victim sheet would allow you to calculate things like – the percentage of victims who were black; or the percentage who were men versus women, etc. (This is also a perfect example of where a relational database software will be far more useful than a spreadsheet). But then you'd still have the incident-level data that you could use for things like where and when the incidents happened and how many are unsolved.

- Each column should only contain one value. For example, don't put more than one date in a field, like this: 4/3/2015 and 4/4/2015.
- Be prepared for needing to list a value as "unknown." It's better to type something in the field (such as "N/A" or "unknown") rather than leaving it blank. You (or your co-workers) have a hard time knowing that a blank field means you weren't able to get that information versus you simply forgot to fill it in.
- A good data entry form (such as Google Forms or Airtable) will let you restrict the format of data entered into a particular field. You can set a field to only accept a date formatted as mm/dd/yyyy; or only a phone number as xxx-xxx-xxxx. Or zip code field that only holds 5 digits (and nothing more). Take advantage of those features wherever possible to ensure standardization.
- If you have multiple people entering data, make a cheat sheet for how each field should be used to try to ensure consistency. After the data entry is finished, ASSUME that it won't be consistent and you'll need to do cleanup.

@MaryJoWebster
mjwebster71@gmail.com
August 2018