Interviewing data

Common questions that journalists ask:

How often has this happened in the past?

How does this compare to....?

What's the worst...?

What's the best....?

Where is there the most...?

Where is there the least...?

Is this higher or lower than last year?

How has this changed since...?

We spend a lot of time seeking to put things or events or people into context -- basically trying to figure out "is this unusual?". Data can be a big help with that but it's not always obvious what data you would need.

Sometimes the answers to these questions have already been compiled by somebody else. But when that's not the case (or if the answers are outdated or problematic in some other way), then you would need to analyze data to get the answers.

Like interviewing a human source, sometimes you can find the right data source and ask your question and be done. But other times, you need to have a longer conversation. The good news is that data – unlike a human – will sit there and let you ask questions all day (and even night, if necessary!)

What is "interviewing data" ?

--It's NOT seeking a single number or summarized stats compiled by someone else

--It requires you to manipulate structured data in some way to get answers to your questions; the answers are not obvious in the data

--Usually this will involve working with the most detailed level of data available to you. Each row/record likely represents one thing or person, etc.

--Sometimes this analysis will be fairly minimal – a formula or two and sorting a table in Excel – or it might require extensive hours and lots of bandwidth for a large dataset.

--Sometimes this will require merging two datasets together (i.e. I did a story on why the Twins couldn't hit homeruns in their first couple years at Target Field. We used weather and hitting data to show that wind speed and direction played a huge role)

--Most of the data interviewing that journalists do can be accomplished in a spreadsheet software. But sometimes the data is too large or requires geographical analysis or advanced statistical analysis and you need to move to something more sophisticated. Or you might want to use something like a coding language (R and Python are two big examples) so that you can replicate your work. Each piece of software has a different syntax or different buttons to push, but generally you will be doing the same kinds of things regardless of the tool.

What's the advantage to doing the analysis myself?

If somebody else does the analysis, you are basically getting your results from an interpreter. If you've ever tried to talk to someone else via a language interpreter, you will recognize the similarities. It's really hard to have a deep conversation and ask detailed questions. Also the results you get will be summarized; you won't ever get to see the details. Sometimes you won't figure out what questions to ask until you start going through the details.

What if you wanted to write a story about all the people in your state who have died from opioid overdoses? Of course, you'd want to be able to say in your story how many people died. Let's say you ask the state health department for that number. Or maybe you ask them for how many people died each year going back the last 10 years. In both cases, you have minimal information. What if you asked for all the death certificate data for everybody who died from an overdose? This would include the names, date of death, where they lived, their age, gender, race and all kinds of other details. Maybe you won't analyze everything in here, but you've just found people who might be that "face" you need to put on your story.

How do I figure out what data to use?

Reporting. Use your reporting skills to find people who are knowledgeable about this topic. They will likely either have data or be able to point you to the right place.

Ask your sources about their sources: When a source tells you something like: "Last year we racked up the most overtime in the history of this agency…." ask them what data they used to come up with that statistic.

Government and non-profit reports/publications are good places to figure out what data an agency might have. Look at the fine print. What sources are cited for the summary stats in the report? Is there a methodology section of the report?

Get ideas from other news organizations. Even if the story was done in another state, it's plausible that similar data exists here or that the data came from a federal agency.

It's quite common that the exact data you want doesn't exist, but you can generate it from other sources. For example, let's say you want to measure whether schools in your area are segregated or not. You learn that this is not something the state education department measures. But in talking to experts on the topic, you learn that there are ways that others have measured segregation – i.e. 75% or more of the students are white or non-white – and that helps point you to what data the education department has (racial breakdowns of each school) that will help you answer your question.

Keep in mind that it is possible you might end up with data that won't work for what you're trying to do. Sometimes agencies don't store the RIGHT pieces of information that you would need for the questions you want to ask. There are many ways you might encounter problems that either limit what questions you can ask or possibly even prohibit you from doing any analysis.

You might have to build your own dataset, pulling pieces from elsewhere. For example, a Strib reporter noticed that summary stats showed the number of people killed in farming accidents had gone up. He wanted to get details to find out why. But nobody keeps details on those accidents. The state just kept a running count. Then he realized that all of these people should be in the state's death certificate database. He used an injury description field to find words like "farm", "tractor", etc. And then he also searched Minnesota newspapers for stories and obituaries. Then he put all these details into a database that he built himself.

How do I figure out what questions to ask?

You probably have an over-arching question you want to answer, but usually you need to ask a bunch of questions to not only get to know your data, but also to put your "big" answer into context or even just to check its validity.

Start out by getting to know your source. By exploring the fields that are included in the dataset and understanding what each row/record represents you can get a better idea of the breadth of this source's knowledge. For example, if the dataset includes only NFL games played in outdoor stadiums, then you know that it's knowledge is limited to only outdoor games (you can't ask any questions about all games). Or if a dataset of foster parent license revocations does NOT include a field indicating why the family lost its foster care license, then you know you can't ask any questions about why the licenses were revoked.

Ask basic questions on every field in your dataset – i.e. if there's a field indicating the neighborhood where the crime occurred, put that in a Pivot Tale (or do a summary query in another software) to find out if all neighborhoods show up in here and which one has the most reported crimes. This is also a really good way to figure out if you have "dirty" or inconsistent data.

Use filters in a spreadsheet to narrow down to subsets of your data as another way to explore what's in your data. Think about what you expect to find in here and go looking for it. Also think about what you don't expect to find and go looking for that.

Talk to other people who are already familiar with this source. Again, back to good ol' reporting. Talk to the people who gave you the data (preferably someone who works with the data, not the PR flak), or find academics or other experts on the topic. Tell them what you're hoping to figure out from this data

and get their opinions about whether the data can help you, and any suggestions they might have for how to do that.

Ask some questions then run the results past your sources. This will likely lead you to questions you forgot to ask or slightly different versions of the questions you already asked.

How do I know my answers are correct?

Back to good ol' reporting. Run your results past your sources. If anyone gives you any hint of "that seems a little off" or "I didn't really expect that" kind of attitude, assume that you did something wrong and back track. Ask those sources to give you as much feedback as possible about why they are questioning it. Did they see other analysis that came up with different results?

If your results seem "too good to be true" to either you or anyone else around you, assume that is correct. This usually means you did something wrong in your analysis or maybe you had bad data. Case study: A reporter at the Strib had data on chronic absenteeism in Minneapolis public schools. She got the data from the district. Her analysis showed one middle school had dropped its absenteeism rate dramatically from one year to the next. So she planned a story featuring that school. Her editor wanted her to include data on absenteeism rates in other districts, so she talked to the state department of education. It wasn't going to be easy because Minneapolis calculated their absenteeism rate slightly differently than the state did. (being absent X percentage of days over a certain period of time). And then they discovered that the state's data didn't show such a big drop as the Minneapolis data did. The reporter continued pressing her sources at both the state and the district about the discrepancies. Eventually a data person at the district dug into it and discovered their data was incorrect. The reporter caught this literally a day before publishing a glowing story about a school that really wasn't having that much success reversing chronic absenteeism.

If there are summary statistics compiled by someone else (especially the agency that gave you the data), try to generate the same stats from your data.

If possible, spot check your data against paper records or other sources. When I was at the Pioneer Press, we got data from the state on all police chases. The reporter went back through old stories and pulled up chases that she had written about and she checked to see if they were in the data. Not all of them were.

By MaryJo Webster
@MaryJoWebster
mjwebste@umn.edu
Last updated: January 2018