By MaryJo Webster
@MaryJoWebster
Mjwebster71@gmail.com

WRESTLING WITH DATA:

Once you have the data, you might still have another hurdle. Perhaps the data is not set up in a fashion that allows you to easily ask the question(s) that you want. This is a common scenario, but it can usually be remedied.

Let's say you want to figure out whether the city is meeting its goal of inspecting restaurants at least once a year. You find out that the city has a database of inspection results, which includes the date of the inspection. Although it has the date, it doesn't spell out specifically when the last inspection was (there isn't a field for that). So you have to come up with some creative analysis to compare the most recent inspection to the prior inspection record. But you've got hundreds of restaurants, so you're not going to do this by hand. This is not something that can be done with a basic Access query (and definitely can't be done in Excel), but it can be done with a little more advanced knowledge.

You will run into situations like this more often than not. The trick is to first focus on what pieces of information you need to answer your question, and whether or not those pieces are included somewhere in your data. In this case, the piece you need is the date of the last inspection – which presumably would be included in a separate record if you have multiple years of data. The next step then is to find out how to transform your data into a format that will allow you to do the analysis you want. For beginners, this might be the best time to consult experts on the NICAR listserv or colleagues.

Other times you might need a second dataset to help you. For example, Milwaukee last year did a story about high water usage in the poorer neighborhoods of the city. The city's water data provided the consumption figures, but the reporter had to use Census data to determine which areas of the city would be considered "poor" and which were not.

How do I know if what I'm seeing in the data is abnormal or newsworthy?

Data analysis USALLY involves making a judgment about whether or not something is "good" or "bad" or whether an agency is doing its job properly or whether a problem has gotten out of control.

The data itself isn't going to give you a big red flag saying "Whoa, there's a problem here!"

You're going to have to figure out what the proper benchmark is. Where is the line that determines something is good versus bad? Or what's the standard that the agency is supposed to be meeting in order to be considered "doing their job properly"?

This is where an expert is a necessity. And more than one expert is even better.

Talk to as many people in the field/industry you're writing about as possible. Ask them if there's a standard that's supposed to be followed? What are red flags? What's considered good? What's considered bad?

For example, fire departments have a national standard that they are expected to meet in terms of how quickly they respond to fires, on average. You would need to know this if you're going to try to figure out if your local fire department is responding to fires in a timely manner.

Once you know this benchmark, then you have an easier time asking the data questions and determining the focus of your story.

Some stories won't necessarily need a benchmark, but instead rely on comparing change over time or comparing one organization to another that is similar.

For example, the Pioneer Press looked at gun-related incidents in St. Paul. We compared the first five months of 2012 to the average for that same time period (the average was calculated using data from 2008-2010), to show that gun-related incidents were up 65 percent.

In the story we didn't need to say whether this was good or bad or anything like that. We just had to lay out the facts.

School test scores are one that could go either way — you could have a benchmark for determining which schools are doing well and which ones are not. Or you could simply compare the schools amongst each other (as long as you are making fair comparisons…for example you wouldn't want to compare an elementary school to a high school)

Dirty data:

What do I mean by dirty? The answer varies from dataset to dataset. But the bottom line is that something is wrong with the data in a way that makes it difficult, if not impossible, for you to ask your question(s) and get the right answer.

Sometimes there are inconsistencies in the content of the data (another way to phrase this is to say that the data is not standardized), other times there might be data missing (either it didn't get entered in the first place or you didn't receive it as part of the transfer), other times you have a formatting problem — such as a date field stored as text and appearing like "20120704" instead of 07/04/2012.

For example, campaign finance data lists the name of the person making the donation, but you'll often find that the person's name might be listed differently on one donation than on another donation. One might say "Todd B. Jones" and another might just say "Todd Jones". Another might say "Jones, Todd." If you want to find out how much in total each person gave (and who gave the most), these names would need to be standardized.

A computer program (regardless of which one you're using) will see those as three different people. In order to tally up who gave the most money, you first have to standardize all those variations.

You will encounter dirty data more often than not. Some of the cleanest data I run across is from the Minnesota Department of Education. They clearly do a great job making sure all the numbers are there. But they change the way they list the name of a school or a district from one year to another, so if I'm trying to analyze data across many years' worth of data, I have trouble matching up the schools (i.e. one year they say "St. Paul Public School District" and another year they say "St. Paul Public Schools")

When do I need to clean data?

When you get a dataset, you'll most likely find that something is wrong with it. If you're lucky, it's just one thing. The good news is that — generally — you don't need to "clean" the whole thing until it's bright and sparkly.

You just need to deal with any problems that stand in the way of getting the answers you want from it.

For example, you might encounter some problems that make it impossible for you to import the data properly into Excel or Access. This, of course, would require attention. This is usually some sort of problem with the column delimiters (which tell the software where to split the data into columns) or row delimiters (which tell the software where a row ends and new one begins).

More common are problems with the content of the data itself. You want to make sure the information is standardized, at least for the fields that you will use for asking questions. So, if you have a database of NFL draftees and one of the fields is the state where he graduated from high school, you'll want to make sure that the names (or abbreviations) of the states are consistent. You don't want one player to say "MN" and another to say "Minn." and another to say "Minnesota", etc.

But if you're not planning to use that column as part of your analysis, there's no need to clean it up.

Here are some real-life examples of journalists "wrestling" with problematic data, for various reasons:

From John Perry (now at the Atlanta Journal Constitution):

When I was at The Oklahoman we did a convergence project with a local TV station on "Rape Hotspots." (It was sweeps week.) Turns out, the two big rape hotspots were at the downtown police station and the University of Oklahoma Health Science Center. Of course that meant that the cops taking the rape report, at the police station or at HSC, where victims were examined, and under location they wrote down where they were standing, not where the rape occurred. (Lesson: know how your data is collected.)

At the Center for Public Integrity, when we built the first congressional travel database from the paper records, we ran into a problem with a poorly designed form. The form had lines for lodging, meals and transportation expenses, and then there was a line labeled "Other Total." There was also a "Total" line. So in the "Other" field, some people entered only expenses that weren't lodging, meals or travel; (probably as intended). Others entered the total of lodging, meals, transportation and everything else. Some reports you could tell what they did by comparing the calculated total with the entered total. But for a large number, nothing added up like it was supposed to, and it was impossible to tell what they actually spent. (Lesson: the world is not perfect. Don't expect data to be perfect.)

In Atlanta, we were looking at daily summary report data from Atlanta Police 911 calls. Because the cops couldn't figure out how to measure response times for calls that spanned midnight, about 20 percent of the calls were missing from the summaries. (Lesson: cops can't do math.)

From Jamie Smith Hopkins - Baltimore Sun

Maryland's assessors track home sales and note whether the buyer is an owner-occupier or not. It's a useful way of seeing how many homes are being snapped up by investors. But one six-month period, a state programming error caused many Baltimore sales to show up as non-owner-occupied when they weren't. The dirty data made it seem like 90 percent of sales were going to investors, when actually it was 70 percent. (This was at the height of the housing bubble.)

Before I printed the 90 percent figure, I asked state officials, "Really? Are you sure this is right?" They assured me their dataset was good. And because I'd used earlier versions of the data several times before, and I knew the investor buying was truly at frenzy levels, I went with it. A few weeks later, a nonprofit researcher emailed to tip me off to the programming error.

Fortunately the story wasn't built around that number – it was a mention 10 grafs in. But still: Ouch. It was a good lesson that unbelievable numbers are probably unbelievable for a reason.

From Tom Torok, New York Times:

The New Jersey Attorney General's office gave us a statewide csv voter database with no text qualifiers. Thus, anything with a comma, like an address, would shift the data inappropriately. We asked the AG to rerun the data with text qualifiers. The AG's office refused, saying that the format was "almost standard." Which is pretty much like being "almost pregnant." So, we wrote a program to identify and extract any line with more than 42 commas (the number of fields). We then took all the 43s and pasted them into Excel, and reconciled the problems, which took care of the many students who had addresses like "Rutgers, The State University." We repeated same for 44s, 45s, etc.

A non-CAR story that I use to warn students about the validity of data: A friend of mine worked for a computer firm that was hit with an EEOC suit. A black worker complained that the company was not promoting or hiring blacks proportionately to the community. The company used its personnel database to demonstrate to the EEOC that it indeed was hiring and promoting many more blacks than it was required to and the EEOC dismissed the suit. When the company honchos told a board meeting about winning the suit, the secretary responsible for inputting the HR data turned ashen. When asked what was wrong, she explained that when inputting the data, the race field was a required field but there was nothing on the company paperwork to indicate race. So, she said, she looked at the name and guessed the race.

From Janet Roberts, New York Times:

I worked on a project here at The Times about the influence of drug company payola on doctors. We got the Minnesota Medicaid prescriptions data (patient names redacted). We had hoped to identify the top doctors who prescribe antipsychotic medications to children and to compare that list to another database listing drug company payments to doctors. After we'd had the data for a while, the flack for the Division of Human Resources sent an e-mail: "By the way," she wrote, "you should know that we estimate that 20 percent of the prescriber names are inaccurate." When we asked how they arrived at that figure, they were unable to provide any valid methodology. We started examining the data ourselves and found really strange things: e.g., an ophthalmologist in Duluth prescribing antipsychotics to a kid in Minneapolis and the same patient getting scripts for the same drug from different doctors in very erratic patterns. It was clear we had dirty data.

I started inquiring about how the data are collected. The prescriber information gets entered at the pharmacy level, most likely by those pharmacy clerks who take your prescription form when you go to the counter. They get the name of the prescriber from whatever is scribbled on the signature line. You can imagine how many times they get it wrong.

At the end of the day, we were unable to quantify the level of error in that key field, so we had to abandon the idea of identifying the top prescribers.

From Dave Gulliver (formerly Dayton Daily News & other news orgs):
my first big data story was on campaign contributions in the Ohio legislature. I had just started working at the Dayton paper and was working off a database someone had acquired. the contribution numbers were big — a lot bigger than I expected — but they were coming from unions and interest groups, and so I figured, sure, they have deep pockets. but the state house reporter, a real veteran, questioned them, and so I went back to the database, and realized that the decimal point portion of the amount was being imported as if it were dollars — in other words, $300.00 became $30,000.

the data had been imported from a 9-track tape and used an old-school field format — 999PIC99, I think, or something like that. so the “amount” field wasn’t what it appeared to be. lesson learned: always go back to the documentation and be sure you understand every last detail.

second one:
I was doing all the data import and set-up for an investigation of military plane crashes.

the reporter had done the data FOIAs, and was famous for asking for a “data dump” and not specifying the format. So the Air Force, I think, gave us a CSV file (comma-delimited). Which would have been fine — except it had altitude and velocity fields that, yep, contained commas. So every time an event happened at more than 999 feet, (eg 30,000 feet) it split the field into two pieces — “30” and “000” for example. and that in turn threw off the number of fields in a row, and pushed everything down a field.

I did a couple of different things. in small tables, I kludged it and fixed the rows by hand. I don’t think I could figure a safe way to tell the text commas from the delimiters, so I ginned up a SQL program that checked the fields where the problem occurred and then concatenated and re-divided everything after the occurrence. or something like that. it was a nightmare.

From MaryJo Webster:

When I was at the Pioneer Press, I worked with an education reporter on the school district’s spending database. It was essentially their checkbook register --- every payment they made. We thought it would be fairly straightforward and, when we ran into questions or problems, we asked the district for help. We did everything you're supposed to do. And we still ended up running a correction. Turns out that when the district's finance office gave us the data, they included all the voided records – but they left off the field that indicated whether a record was a void or not. To make matters worse, the way they handled voids was to simply double the dollar amount. So there would be the voided record with the doubled amount, then sometimes the actual check that was cut -- both going to the same place, for the same reason, but with different amounts. So when we tallied up totals, we had the wrong numbers. Read more about that here.

OTHER PITFALLS:

Numbers don't always tell the whole story. A good simple example is that when you look at Census data you will find pockets where there are huge percentages of single men. But you have to be wary ….most of those pockets are prisons.

Crime data has been under fire recently for not being truly reflective in comparing whether crime is worse in one place or another. Recently some academics have done studies to see if other factors might play a role: these include poverty, racial makeup, divorce rates and availability of good trauma medical care. One study found that while the number of incidents of violent crime have gone up, the survival rates for victims have also improved so that the overall "homicide rate" had actually declined. There also have been so many news organizations that uncovered the fact that police in one city report crimes differently than police in other cities, making it almost impossible to come crime data that is reported to the FBI.

Some examples:
--Nobody knows how many Americans the police kill each year
--Reports on college crime deceptively inaccurate
--Scope of nationwide heroin 'epidemic' unknown

When reporting on ANY numbers – whether you are doing the analysis yourself or if you work with information provided by another source – it's crucial that you investigate all the possible pitfalls or missing pieces of information so that your reporting can compensate for that.

A good read on how to "Distrust your data"

And another on "Common Misconceptions and how to avoid them"