By MaryJo Webster
maryjo.webster@startribune.com
@MaryJoWebster

You've got a great story idea. You found the data you think you'll need. You requested it. And – perhaps much to your surprise – it arrived.

Now you're itching to finally start asking the data some questions.

But I will bet my paycheck that you won't be able to start the analysis just yet. More often than not data will show up on your doorstep with a problem – or two or three or four. In the data journalism world, we call this "dirty data."

In other words, something is wrong with the data in a way that makes it difficult, if not impossible, for you to ask your questions and/or get the right answers.

Usually it's not bad enough to stop your great idea. But first you need to find the problems. Trust me: you want to find the flaws in your data before they find you. Because they will.

Unfortunately, there isn't a definitive list of problems you might encounter because each dataset is unique.

But there are some that are all-too common that I'm going to highlight: inconsistent values, missing data, structure troubles and what I like to call "the agency screwed up" and "it's-too-good-to-be-true" problems. I've collected stories from some of my fellow data journalists, in addition to my own tales of woe, to help shed light on the wrestling match you'll likely have with your data.

Inconsistencies:

The term "dirty data" is most often associated with the most common problem you might encounter – values in your data that are inconsistent.

The classic example is campaign finance contribution data that have a separate record for each check written and the names of the donors are not consistent, even though the money might have come from the same person. You might have five records with names listed as: John Smith, John and Sally Smith, John J. Smith, John J Smith (no period), and "Smith, John." Through reporting you've confirmed this is the same guy. But no matter which computer software you use to tally up his total contributions, the computer will think those are five different people.

This poses a problem because it makes it impossible for you to summarize the data to find out who contributed the most money to a campaign.

Variations of other values or just outright misspellings can cause show up almost anywhere. I had another dataset that had at least 20 different spellings of Minneapolis. I really had no idea you could come up with that many variations. Another big city nearby me might show up in data as "St. Paul" (with a period), "St Paul" (without a period) or "Saint Paul." Occasionally I also find "St. Pual", of course.

Fields that have codes in them are supposed to be consistent (that's the whole point of using codes!), but you can't count on it. For example, you might have a field identifying a person's gender. The documentation says this should either be “M” or “F.” But in the data you have “M”, “F”, “O”, “N”, and lots of rows where this cell is blank. In my experience, the bad codes are either typos (most of the time) or relics from old coding systems that didn’t get translated to a new coding system. There could be all kinds of different reasons for the blanks (also known as NULL values).

Or you might have a field indicating a unit of measurement – such as a currency value for a salary rate – but maybe for one person it says “10” and another it says “2000.” But it doesn’t tell you the unit of measurement. Is this an hourly rate or a bi-weekly or an annual rate? And having it all in the same field makes it impossible to do any calculations, such as an average or median.

Outlier values are also red flags. Perhaps you have a date field and you expect the values to be somewhere in the 21st century, but a couple records show dates in the early 1900s. Are those bad dates that are supposed to be in the 2000s?

Finding inconsistencies is easy. Just do a summary of each field/column using a Pivot Table or a group by query or any other tool you would use to summarize data. You’ll quickly see the problems. Tracking down what the correct values should be is a bit harder and requires some traditional reporting.

Missing data:

Too often, you’ll discover that an agency doesn’t track what you want or doesn’t do a good job of tracking it.

Many years ago, I got a database of registered underground storage tanks in Kansas, including data on which ones had sprung a leak. The documentation showed there was a field indicating what had caused the leak and I was super excited because this was the focus of my work. But when I got the data, only a fraction of the records had values in this field. The agency told me they weren’t very good at filling in all the details. The field proved useless.

A few years ago, a colleague wanted to find out why foster parents were having their licenses suspended or revoked. Were they abusing the kids? Or neglecting them? Did they commit a crime? So, she requested Minnesota’s database of license revocations. Turns out the state didn’t put the reason for the revocation into the database; it was only stored in paper files.

For another project, I got data on police pursuits because we wanted to know how often bystanders – such as pedestrians or other drivers not involved in the pursuit – were injured or killed. A state law required every law enforcement agency to report details about each high-speed pursuit to a state agency. Our newspaper had written about many of these, so we pulled up the old stories and checked to see if those incidents were in the database. Turns out some of them weren’t. We also found cases where someone had died but the data didn’t indicate that. We learned that sometimes this was the fault of the state agency not putting the information into the database, and sometimes it was because law enforcement agencies didn’t submit anything.

Structure problems:

The main way to know you’ve got a structure problem is if you can’t figure out how to set up a Pivot Table or run a query that will get the answer you want. Your data table maybe has rows that should be columns. Or there are multiple rows for each thing that you want to count. Or it’s missing something that you need.

For example, you may want to know which neighborhood in the city had the most reported crime. But the crime data you have only has the address, not the name of the neighborhood. This is an example of where you need to “categorize” the data – creating a new field where you assign each record a category (in this case, a neighborhood name).

Another example: Recently, I worked on a project about police officers who had been convicted of criminal offenses. The court conviction data came to us with one record for each charge. For example, if an officer was charged with driving under the influence and reckless driving for the same incident, there would be two records even if one of the charges had later been dismissed. We wanted to count up how many convictions there had been, and by this, we really meant incidents or cases (not charges). I realized that this counting would be much easier if the data table had one record for each case, so I spent some time re-arranging to make a new table with one record that identified the number of charges and the highest level of conviction (such as felony, gross misdemeanor or misdemeanor), plus some other pertinent details.

I could write a whole separate article just about structure problems. Bottom line is that it happens often and is something you should assess at the outset.

Skewed data:

These are situations where flaws end up in the data because of the data collection or inputting process by the agency. If you’re lucky, the data keepers might tip you off. If you’re unlucky, you won’t notice it until you end up with strange results in your analysis. To be safe, the best thing to do is to ask a lot of questions about how the data ends up in the database.

Janet Roberts, now head of the Reuters data team, told me this tale about data she worked with while at the New York Times: For a project about the influence of drug company payola, Janet requested Minnesota Medicaid prescriptions data. They hoped to identify the top doctors who prescribed antipsychotic medications to children and compare that list to another database listing drug company payments to doctors.

After they’d had the data for a while, a flack for the Division of Human Resources sent an email: “By the way,” she wrote, “you should know that we estimate that 20 percent of the prescriber names are inaccurate.” When Janet asked how they arrived at that figure, they were unable to provide any valid methodology.

Janet started examining the data and found really strange things: for example, an ophthalmologist in northern Minnesota prescribing antipsychotics to a kid who was hundreds of miles away in Minneapolis, and the same patient getting prescriptions for the same drug from different doctors in very erratic patterns. Janet started inquiring how the data were collected. The prescriber information got entered at the pharmacy level, most likely by the clerks who take your prescription. They got the name of the physician from whatever was scribbled on the signature line. You can imagine how many times they got it wrong. Janet said they were unable to quantify the level of error in that crucial field, so they had to abandon the idea of identifying the top prescribers.

John Perry, now at the Atlanta Journal Constitution, told me about using crime data in Oklahoma City for a joint project between the city’s newspaper and a local TV station. The plan was to identify the “rape hotspots” in the city. Turns out, the two big hotspots were at the downtown police station and the

University of Oklahoma Health Science Center. In other words, the location field in the data was sometimes where the cops were standing when taking the rape report – not where the crime occurred.

Tom Torok, a now-retired data journalist, told me a story that he used to warn students about the validity of data. A friend of his worked for a computer firm that was hit with a lawsuit alleging racial discrimination in hiring and promotions. The company used its personnel database to demonstrate that it was hiring and promoting many more blacks than it was required to; the lawsuit was dismissed. When the company honchos announced the lawsuit victory at a board meeting, the secretary responsible for inputting the 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.

Agency screw-ups:

Sometimes you just get bad data handed to you.

My worst nightmare came true when working with data from the St. Paul Public School District several years ago. A colleague and I had requested all the payments they had made – essentially the equivalent of their checkbook register. We thought it would be fairly straightforward, and when we ran into questions or problems we did exactly what we were supposed to do – asked the school district for help.

The first red flag was that there were records that appeared to be duplicates – hundreds of them – but they weren't exactly the same. The recipient would be the same; the date was the same; the description of the expenditure was the same; the codes for which pot of revenue it came from were the same. But one check would say something like $5,000 and the other would be twice that amount.

We asked the school district to explain this and they said it was all legitimate. They assured us there weren't any problems. We shouldn't have listened to them.

Turns out they included voided checks, but they didn't include a field indicating which records were voided checks and which were not. To make matters worse, the way they handled voided checks was to simply double the dollar amount on the voided record. So, when we tallied up totals, we were way off.

Another schools story by a Star Tribune colleague narrowly avoided disaster. The reporter saved the day because she kept asking questions. Her data showed that a Minneapolis school had dropped its chronic absenteeism rate astonishingly fast. At first, she and her editor thought this would be a great feature story about this school solving a big problem.

Her editor asked her to include a chart with the story showing how this school district compared to others in the area. Problem was that the data she had was only for the Minneapolis school district. Data for the other districts was at the state education agency, and it turns out the two datasets used different methods for measuring chronic absenteeism.

And the state's data for that standout school showed they had not dropped their chronic absenteeism rate dramatically. The reporter started asking both the school district and the state agency to clarify. She got stonewalled for weeks. Nobody had a good answer – they just kept saying the two measurements were different because they used different definitions.

Still, that astonishingly fast drop in the rate seemed “too good to be true” to the reporter. Turns out it was. Just a few days before publication, the school district finally dug into their own data and discovered the numbers for that one school were incorrect.

The common thread you can take away from all these tales is that you really need to get to know your data and try to do as much of that research as possible before making a request. How does the information end up in the database? What fields did the agency transfer (or not transfer) to you? How is each field used? What codes are supposed to be in there (or not)?

And then you need to spend time looking for inconsistencies or missing data that you expected to be there. If possible, cross check your data against paper records or other sources. Aggregate your data to try to make it match published reports.

Finally, the method of last resort is always watching out for the “it’s-too-good-to-be-true” results.

For another listing of potential problems and how to solve them, I’d recommend you check out the Quartz Guide to Bad Data -- https://github.com/Quartz/bad-data-guide