The Quartz guide to bad data

An exhaustive reference to problems seen in real-world data along with suggestions on how to resolve them.

As a reporter your world is full of data. And those data are full of problems. This guide presents thorough descriptions and suggested solutions to many of the kinds of problems that you will encounter when working with data.

Most of these problems can be solved. Some of them can't be solved and that means you should not use the data. Others can't be solved, but with precautions you can continue using the data. In order to allow for these ambiguities, this guide is organized by who is best equipped to solve the problem: you, your source, an expert, etc. In the description of each problem you may also find suggestions for what to do if that person can't help you.

You cannot possibly review every dataset you encounter for all of these problems. If you try to do that you will never get anything published. However, by familiarizing yourself with the kinds of issues you are likely to encounter you will have a better chance of identifying an issue before it causes you to make a mistake.

If you have questions about this guide please email Chris[1]. Good luck!

Send your pull requests!

Index

Issues that your source should solve

Issues that you should solve

Issues a third-party expert should help you solve

Issues a programmer should help you solve

Detailed list of all problems

Issues that your source should solve

Values are missing

Beware blank or "null" values in any dataset unless you are certain you know what they mean. If the data are annual, was the value for that year never collected? If it is a survey, did a respondent refuse to answer the question?

Any time you're working with data that has missing values you should ask yourself: "Do I know what the absence of this value means?" If the answer is no, you should ask your source.

Zeros replace missing values

Worse than a missing value is when an arbitrary value is used instead. This can be the result of a human not thinking through the implications or it can happen as the result of automated processes that simply don't know how to handle null values. In any case, if you see zeros in a series of numbers you should ask yourself if those values are really the number `0` or if they instead means "nothing". (`-1` is also sometimes used this way.) If you aren't sure, ask your source.

The same caution should be exercised for other non-numerical values where a `0` may be represented in another way. For example a false `0` value for a date is often displayed as `1970-01-01T00:00:00Z` or `1969-12-31T24:59:59Z` which is the Unix epoch for timestamps. A false `0` for a location might be represented as `0°00′00.0″N+0°00′00.0″E` or simply `0°N 0°E` which is a point in the Atlantic Ocean just south of Ghana often referred to as Null Island[3].

See also:

Data are missing you know should be there

Sometimes data are missing and you can't tell from the dataset itself, but you can still know because you know what the data purports to be about. If you have a dataset covering the United States then you can check to ensure all 50 states represented. (And don't forget about the territories[4]—50 isn't the right number if the dataset includes Puerto Rico.) If you're dealing with a dataset of baseball players make sure it has the number of teams you expect. Verify that a few players who you know are included. Trust your intuition if something seems to missing and double-check with your source. The universe of your data might be smaller than you think.

Rows or values are duplicated

If the same row appears in your dataset more than once you should find out why. Sometimes it need not be a whole row. Some campaign finance data include "amendments" that use the same unique identifiers as the original transaction. If you didn't know that then any calculations you did with the data would be wrong. If something seems like it should be unique verify that it is. If you discover that it isn't, ask your source why.

Spelling is inconsistent

Spelling is one of the most obvious ways of telling if data have been compiled by hand. Don't just look at people's names—those are often the hardest place to detect spelling errors. Instead look for places where city names or states aren't consistent. (`Los Angelos` is one very common mistake.) If you find those, you can be pretty sure the data were compiled or edited by hand and that is always a reason to be skeptical of it. Data that have been edited by hand are the most likely to have mistakes. This doesn't mean you shouldn't use them but you may need to manually correct those mistakes or otherwise account for them in your reporting.

OpenRefine's[5] utility for text clustering[6] can help streamline the spelling correction process by suggesting close matches between inconsistent values within a column (for example, matching `Los Angelos` and `Los Angeles`). Be sure, however, to document the changes you make[7] so as to ensure good data provenance.

See also:

- Data were entered by humans

Name order is inconsistent

Does your data have Middle Eastern or East Asian names in it? Are you sure the surnames are always in the same place? Is it possible anyone in your dataset uses a mononym? These are the sorts of things that data makers habitually get wrong. If you're working with a list of ethnically diverse names—which is any list of names—then you should do at least a cursory review before assuming that joining the `first_name` and `last_name` columns will give you something that is appropriate to publish.

- Data were entered by humans

Date formats are inconsistent

Which date is in September:

- `10/9/15`
- `9/10/15`

If the first one was written by a European and the second one by an American then they both are[8]. Without knowing the history of the data you can't know for sure. Know where your data came from and be sure that it was all created by folks from the same continent.

Units are not specified

Neither `weight` nor `cost` conveys any information about the unit of measurement. Don't be too quick to assume that data produced within the United States are in units of pounds and dollars. Scientific data are often metric. Foreign prices may be specified in their local currency. If the data do not spell out their units, go back to your source and find out. Even if it does spell out its units always be wary of meanings that may have shifted over time. A dollar in 2010 is not a dollar today. And a ton[9] is not a ton[10] nor a tonne[11].

See also:

Categories are badly chosen

Watch out for values which purport to be only `true` or `false`, but really aren't. This is often the case with surveys where `refused` or `no answer` are also valid—and meaningful—values. Another common problem is the usage of any kind of `other` category. If the categories in a dataset are a bunch of countries and an `other`, what does that mean? Does it mean that the person collecting the data didn't know the right answer? Were they in international waters? Expatriates? Refugees?

Bad categories can also artificially exclude data. This frequently happens with crime statistics. The FBI has defined the crime of "rape" in a variety of different ways over time. In fact, they've done such a poor job of figuring out what rape is that many criminologists argue their statistics should not be used at all. A bad definition might mean a crime is counted in a different category than you expect or that it wasn't counted at all. Be exceptionally ware of this problem when working with topics where definitions tend to be arbitrary, such as `race` or `ethnicity`.

Field names are ambiguous

What is a `residence`? Is it where someone lives or where they pay taxes? Is it it a city or a county? Field names in data are never as specific as we would like, but particular concern should be applied to those that could obviously mean two or more things. Even if you correctly infer what the values are supposed to mean, that ambiguity could have easily caused the person collecting the data to enter the wrong value.

Provenance is not documented

Data are created by a variety of kinds of individuals and organizations including businesses, governments, nonprofits and nut-job conspiracy theorists. Data are gathered in many different ways including surveys, sensors and satellites. It may be typed, tapped or scribbled. Knowing where your data came from can give you a huge amount of insight into its limitations.

Survey data, for example, is rarely exhaustive. Sensors vary in their accuracy. Governments are often disinclined to give you unbiased information. Data from a war zone may have a strong geographical bias due to the danger of crossing battle lines. To make this situation worse, these different sources are often daisy-chained together. Policy analysts frequently redistribute data they got from the government. Data that were written by a doctor may be keyed by a nurse. Every stage in that chain is an opportunity for error. Know where your data came from.

See also:

- Units are not specified

Suspicious values are present

If you see any of these values in your data, treat them with an abundance of caution:

Numbers:

- `65,535`[12]
- `2,147,483,647`[13]
- `4,294,967,295`[14]
- `555-3485`[15]
- `99999` (or any other long sequence of 9's)
- `00000` (or any other sequence of 0's)

Dates:

Locations:

- `0°00’00.0”N+0°00’00.0”E`[16] or simply `0°N 0°E`[17]
- US zip code `12345` (Schenectady, New York)
- US zip code `90210` (Beverly Hills, CA)

Each of these numbers has an indication of a particular error made by either a human or a computer. If you see them, ensure they actually mean what you think they mean!

See also:

Data are too coarse

You've got states and you need counties. You've got employers and you need employees. They gave you years, but you want months. In many cases we get data that have been aggregated too much for our purposes.

Data usually cannot be disaggregated once they have been merged together. If you're given data that are too coarse, you'll need to ask your source for something more specific. They may not have it. If they do have it they may not be able or willing to give it to you. There are many federal datasets that can't be accessed at the local level to protect the privacy of individuals who might be uniquely identified by them. (For example, a single Somali national living in western Texas.) All you can do is ask.

One thing you should never ever do is divide a yearly value by 12 and call it the "average per month". Without knowing the distribution of the values, that number will be meaningless. (Maybe all instances occurred in one month or one season. Maybe the data follows an exponential trend instead of a linear one.) It's wrong. Don't do it.

See also:

Totals differ from published aggregates

Imagine that after a long FOIA fight you receive a "complete" list of incidents of police use-of-force. You open it up and discover it has 2,467 rows. Great, time to report it out. Not so fast. Before you publish anything from that dataset go find the last time that police chief went on the record about his department's use of force. You may find that in an interview six weeks earlier he said "less than 2,000 times" or that he named a specific number and it doesn't match your dataset.

These sorts of discrepancies between published statistics and raw data can be a very great source of leads. Often the answer will be simple. For instance, the data you were given may not cover the same time period he was speaking about. But sometimes you'll catch them in a lie. Either way, you should make sure the published numbers match the totals for the data you're given.

Spreadsheet has 65536 rows

The maximum number of rows an old-fashioned Excel spreadsheet was allowed to have was 65,536. If you receive a dataset with that number of rows you have almost certainly been given truncated data. Go back and ask for the rest. Newer versions of Excel allowed for 1,048,576 rows, so it's less likely you'll be working with data that hits the limit.

Spreadsheet has dates in 1900, 1904, 1969, or 1970

For reasons beyond obscure, Excel's default date from which it counts all other dates is `January 1st, 1900`, *unless* you're using Excel on a Mac, in which case it's `January 1st, 1904`. There are a variety of ways in which data in Excel can be entered or calculated incorrectly and end up as one of these two dates. If you spot them in your data, it's probably an issue.

Many databases and applications will often generate a date of `1970-01-01T00:00:00Z` or `1969-12-31T24:59:59Z` which is the Unix epoch for timestamps. In other words this is what happens when a system tries to display an empty value or a `0` value as a date.

Text has been converted to numbers

Not all numerals are numbers. For instance, the US Census Bureau uses "FIPS codes" to identify every place in the United States. These codes are of various lengths and are numeric. However, they are *not* numbers. `037` is the FIPS code for Los Angeles County. It is not the number `37`. The numerals `37` are, however, a valid FIPS code: for North Carolina. Excel and other spreadsheets will often make the mistake of assuming numerals are numbers and stripping the leading zeros. This can cause all kinds of problems if you try to convert it to another file format or merge it with another dataset. Watch out for data where this has happened before it was given to you.

Issues that you should solve

Text is garbled

All letters are represented by computers as numbers. Encoding problems are issues that arise when text is represented by a specific set of numbers (called an "encoding") and you don't know what it is. This leads to a phenomenon called mojibake[18] where the text in your data looks like garbage, or like this: ���.

In the vast majority of cases your text editor or spreadsheet application will figure out the correct encoding, however, if it screws it up you could publishing somebody's name with a weird character in the middle. Your source should be able to tell you what encoding your data are in. In the event they can't there are ways of guessing that are about fairly reliable. Ask a programmer.

Line endings are garbled

All text and "text data" files, such as CSV, use invisible characters to represent the ends of lines. Windows, Mac and Linux computers have historically disagreed about what these line ending characters should be. Attempting to open a file saved on one operating system from another operating system can sometimes cause Excel or other applications to fail to properly identify the line breaks.

Typically, this is easy to resolve by simply opening the file in any general-purpose text editor and re-saving it. If the file is exceptionally large you may need to consider using a command-line tool or enlisting the help of a programmer. You can read more about this issue here[19].

Data are in a PDF

A tremendous amount of data—especially government data—are only available in PDF format. If you have real, textual data inside the PDF then there are several good options for extracting them. (If you've got scanned documents that's a different problem.) One excellent, free tool is Tabula[20]. However, if you have Adobe Creative Cloud then also have access to Acrobat Pro, which has an excellent feature for exporting tables in PDFs to Excel. Either solution should be able to extract most tabular data from a PDF.

See also:

- Data are in scanned documents

Data are too granular

This is the opposite of Data are too coarse. In this case you've got counties, but you want states or you've got months but you want years. Fortunately this is usually pretty straightforward.

Data can be aggregated by using the Pivot Table feature of Excel or Google Docs, by using a SQL database or by writing custom code. Pivot Tables are a fabulous tool that every reporter should learn, but they do have their limits. For exceptionally large datasets or aggregations to unusual groups you should ask a programmer and they can craft a solution that's easier to verify and reuse.

See also:

Data were entered by humans

Human data entry is such a common problem that symptoms of it are mentioned in at least 10 of the other issues described here. There is no worse way to screw up data than to let a single human type it in, without validation. For example, I once acquired the complete dog licensing database for Cook County, Illinois. Instead of requiring the person registering their dog to choose a breed from a list, the creators of the system had simply given them a text field to type into. As a result this database contained at least 250 spellings of `Chihuahua`. Even with the best tools available, data this messy can't be saved. They are effectively meaningless. This is not that important with dog data, but you don't want it happening with wounded soldiers or stock tickers. Beware human-entered data.

Data are intermingled with formatting and annotations

Complex representations of data such as HTML and XML allow for a clean separation between data and formatting, but this is not the case for common tabular representations of data such as a spreadsheet. Yet, people still try. A common problem with data provided as a spreadsheet is that the first few rows of data will actually be descriptions or notes about the data rather than column headings or data itself. A key or data dictionary may also be placed in the middle of the spreadsheet. Header rows may be repeated. Or the spreadsheet will include multiple tables (which may have different column headings) one after the other in the same sheet rather than separated into different sheets.

In all of these cases the main solution is simply to identify the problem. Obviously trying to perform any analysis on a spreadsheet that has these kinds of problems will fail, sometimes for non-obvious reasons. When looking at new data for the first time it's always a good idea to ensure there aren't extra header rows or other formatting characters inserted amongst the data.

Aggregations were computed on missing values

Imagine a dataset with 100 rows and a column called `cost`. In 50 of the rows the `cost` column is blank. What is the average of that column? Is it `sum_of_cost / 50` or `sum_of_cost / 100`? There is no one definitive answer. In general, if you're going to compute aggregates on columns that are missing data, you can safely do so by filtering out the missing rows first, but be careful not to compare aggregates from two different columns where different rows were missing values! In some cases the missing values might also be legitimately interpreted as `0`. If you're not sure, ask an expert or just don't do it.

This is an error you can make in your analysis, but it's also an error that others can make and pass on to you, so watch out for it if data comes to you with aggregates already computed.

See also:

Sample is not random

A non-random sampling error occurs when a survey or other sampled dataset either intentionally or accidentally fails to cover the entire population. This can happen for a variety of reasons ranging from time-of-day to the respondent's native language and is a common source of error in sociological research. It can also happen for less obvious reasons, such as when a researcher thinks they have a complete dataset and chooses to work with only part of it. If the original dataset was incomplete for any reason then any conclusions drawn from their sample will be incorrect. The only thing you can do to fix a non-random sample is avoid using that data.

See also:

- Sample is biased

Margin-of-error is too large

I know of no other single issue that causes more reporting errors than the unreflective usage of numbers with very large margins-of-error. MOE is usually associated with survey data. The most likely place a reporter encounters it is when using polling data or the US Census Bureau's American Community Survey[21] data. The MOE is a measure of the range of possible true values. It may be expressed as a number (`400 +/- 80`) or as a percentage of the whole (`400 +/- 20%`). The smaller the relevant population, the larger the MOE will be. For example, according to the 2014 5-year ACS estimates, the number of Asians living in New York is `1,106,989 +/- 3,526` (0.3%). The number of Filipinos is `71,969 +/- 3,088` (4.3%). The number of Samoans is `203 +/- 144`. (71%)

The first two numbers are safe to report. The third number should never be used in published reporting. There is no one rule about when a number is not accurate enough to use, but as a rule of thumb, you should be cautious about using any number with a MOE over 10%.

See also:

- Margin-of-error is unknown

Margin-of-error is unknown

Sometimes the problem isn't that the margin of error is too large, it's that nobody ever bothered to figure out what it was in the first place. This is one problem with unscientific polls. Without computing a MOE, it is impossible to know how accurate the results are. As a general rule, anytime you have data that are from a survey you should ask for what the MOE is. If the source can't tell you, those data probably aren't worth using for any serious analysis.

See also:

- Margin-of-error is too large

Sample is biased

Like a sample that is not random, a biased sample results from a lack of care with how the sampling is executed. Or, from willfully misrepresenting it. A sample might be biased because it was conducted on the internet and poorer people don't use the internet as frequently as the rich. Surveys must be carefully weighted to ensure they cover proportional segments of any population that could skew the results. It's almost impossible to do this perfectly so it is often done wrong.

See also:

- Sample is not random

Data have been manually edited

Manual editing is almost the same as problem as data being entered by humans except that it happens after the fact. In fact, data are often manually edited in an attempt to fix data that were originally entered by humans. Problems start to creep in when the person doing the editing doesn't have complete knowledge of the original data. I once saw someone spontaneously "correct" a name in a dataset from `Smit` to `Smith`. Was that person's name really `Smith`? I don't know, but I do know that value is now a problem. Without a record of that change, it's impossible to verify what it should be.

Issues with manual editing are one reason why you always want to ensure your data have well-documented provenance. A lack of provenance can be a good indication that someone may have monkeyed with it. Academics and policy analysts often get data from the government, monkey with them and then redistribute them to journalists. Without any record of their changes it's impossible to know if the changes they made were justified. Whenever feasible always try to get the *primary source* or at least the earliest version you can and then do your own analysis from that.

See also:

Inflation skews the data

Currency inflation means that over time money changes in value. There is no way to tell if numbers have been "inflation adjusted" just by looking at them. If you get data and you aren't sure if they have been adjusted then check with your source. If they haven't you'll likely want to perform the adjustment. This inflation adjuster[22] is a good place to start.

See also:

- Natural/seasonal variation skews the data

Natural/seasonal variation skews the data

Many types of data fluctuate naturally due to some underlying forces. The best known example of this is employment fluctuating with the seasons. Economists have developed a variety of methods of compensating for this variation. The details of those methods aren't particularly important, but it is important that you know if the data you're using have been "seasonally adjusted". If they haven't and you want to compare employment from month to month you will probably want to get adjusted data from your source. (Adjusting them yourself is much harder than with inflation.)

See also:

- Inflation skews the data

Timeframe has been manipulated

A source can accidentally or intentionally misrepresent the world by giving you data that stops or starts at a specific time. For a potent example see 2015's widely reported "national crime wave". There was no crime wave. What there was was a series of spikes in particular cities when compared to just the last few years. Had journalists examined a wider timeframe they would have seen that violent crime was higher virtually everywhere in the US ten years before. And twenty years before it was nearly double.

If you have data that covers a limited timeframe try to avoid starting your calculations with the very first time period you have data for. If you start a few years (or months or days) into the data you can have confidence that you aren't making a comparison which would be invalidated by having a single additional data point.

See also:

- Frame of reference has been manipulated

Frame of reference has been manipulated

Crime statistics are often manipulated for political purposes by comparing to a year when crime was very high. This can expressed either as a change (down `60%` since 2004) or via an index (`40`, where 2004 = 100). In either of these cases, 2004 may or may not be an appropriate year for comparison. It could have been an unusually high crime year.

This also happens when comparing places. If I want to make one country look bad, I simply express the data about it relative to whichever country is doing the best.

This problem tends to crop up in subjects where people have a strong confirmation bias. ("Just as I thought, crime is up!") Whenever possible try comparing rates from several different starting points to see how the numbers shift. And whatever you do, *don't use this technique yourself* to make a point you think is important. That's inexcusable.

See also:

- Timeframe has been manipulated

Issues a third-party expert should help you solve

Author is untrustworthy

Sometimes the only data you have are from a source you would rather not rely on. In some situations that's just fine. The only people who know how many guns are made are gun manufacturers. However, if you have data from a questionable maker always check it with another expert. Better yet, check it with two or three. Don't publish data from a biased source unless you have substantial corroborating evidence.

Collection process is opaque

It's very easy for false assumptions, errors or outright falsehoods to be introduced into these data collection processes. For this reason it's important that methods used be transparent. It's rare that you'll know exactly how a dataset was gathered, but indications of a problem can include numbers that assert unrealistic precision and data that are too good to be true.

Sometimes the origin story may just be fishy: did such-and-such academic really interview 50 active gang members from the south side of Chicago? If the way the data were gathered seems questionable and your source can't offer you ironcald provenance then you should always verify with another expert that the data could reasonably have been collected in the way that was described.

See also:

Data assert unrealistic precision

Outside of hard science, few things are routinely measured with more than two decimal places of accuracy. If a dataset lands on your desk that purports to show a factory's emissions to the 7th decimal place that is a dead giveaway that it was estimated from other values. That in and of itself may not be a problem, but it's important to be transparent about estimates. They are often wrong.

There are inexplicable outliers

I recently created a dataset of how long it takes for messages to reach different destinations over the internet. All of the times were in the range from `0.05` to `0.8` seconds, except for three. The other three were all over `5,000` seconds. This is a major red flag that something has gone wrong in the production of the data. In this particular case an error in the code I wrote caused some failures to continue counting while all other messages were being sent and received.

Outliers such as these can dramatically screw up your statistics—especially if you're using averages. (You should probably be using medians.) Whenever you have a new dataset it is a good idea to take a look at the largest and smallest values and ensure they are in a reasonable range. If the data justifies it you may also want to do a more statistically rigorous analysis using standard deviations[23] or median deviations[24].

As a side-benefit of doing this work, outliers are often a great way to find story leads. If there really was one country where it took 5,000 times as long to send a message over the internet, that would be a great story.

An index masks underlying variation

Analysts who want to follow the trend of an issue often create indices of various values to track progress. There is nothing intrinsically wrong with using an index. They can have great explanatory power. However, it's important to be cautious of indices that combine disparate measures.

For example, the United Nations Gender Inequality Index[25] combines several measures related to women's progress toward equality. One of the measures used in the GII is "representation of women in parliament". Two countries in the world have laws mandating gender representation in their parliaments: China and Pakistan. As a result these two countries perform far better in the index than countries that are similar in all other ways. Is this fair? It doesn't really matter, because it is confusing to anyone who doesn't know about this factor. The GII and similar indices should always be used with careful analysis to ensure their underlying variables don't swing the index in unexpected ways.

Results have been p-hacked

P-hacking is intentionally altering the data, changing the statistical analysis, or selectively reporting results in order to have statistically significant findings. Examples of this include: stop collecting data once you have a significant results, remove observations to get a significant result, or perform many analyses and only report the few that are significant. There has been some good reporting[26] on this problem.

If you're going to publish the results of a study you need to understand what the p-value is, what that means and then make an educated decision about whether the results are worth using. Lots and lots of garbage study results make it into major publications because journalists don't understand p-values.

See also:

- Margin-of-error is too large

Benford's Law fails

Benford's Law[27] is a theory which states that small digits (1, 2, 3) appear at the beginning of numbers much more frequently than large digits (7, 8, 9). In theory Benford's Law can be used to detect anomalies in accounting practices or election results, though in practice it can easily be misapplied. If you suspect a dataset has been created or modified to deceive, Benford's Law is an excellent first test, but you should always verify your results with an expert before concluding your data have been manipulated.

Too good to be true

There is no global dataset of public opinion. Nobody knows the exact number of people living in Siberia. Crime statistics aren't comparable across borders. The US government is not going to tell you how much fissile material it keeps on hand.

Beware any data that purport to represent something that you could not possibly know. It's not data. It's somebody's estimate and it's probably wrong. Then again... it could be a story, so ask an expert to check it out.

Issues a programmer should help you solve

Data are aggregated to the wrong categories or geographies

Sometimes your data are at about the right level of detail (neither too coarse nor too granular), but they have been aggregated to different grouping than you want. This classic example of this is data that are aggregated by zip codes that you would prefer to have by city neighborhoods. In many cases this is an impossible problem to solve without getting more granular data from your source, but sometimes the data can be proportionally mapped from one group to another. This must be undertaken only with careful understanding of the margin-of-error that may be introduced in the process. If you've got data aggregated to the wrong groups, ask a programmer if it is possible to re-aggregate it.

See also:

Data are in scanned documents

Thanks to FOIA laws it is frequently the case that governments are required to give you data—even though they really don't want to. A very common tactic in these cases is for them to give you scans or photographs of the pages. These may be actual image files or, more likely, they will be gathered up into a PDF.

It is possible to extract text from images and turn it back into data. This is done through a process called optical-character recognition (OCR). Modern OCR can often be almost 100% accurate, but it very much depends on the nature of the document. Anytime you use OCR to extract data you will want to have a process for validating the results match the original.

There are many websites you can upload a document to for OCR, but there are also free tools that a programmer may be able to tune for your specific documents. Ask them what the best strategy is for the particular documents you have.

See also:

- Data are in a PDF

Links

1. mailto:c@qz.com

2. http://creativecommons.org/licenses/by-nc/4.0/
3. https://en.wikipedia.org/wiki/Null_Island
4. https://en.wikipedia.org/wiki/Territories_of_the_United_States
5. http://openrefine.org/
6. https://github.com/OpenRefine/OpenRefine/wiki/Clustering
7. https://github.com/OpenRefine/OpenRefine/wiki/Exporters
8. https://en.wikipedia.org/wiki/Date_format_by_country
9. https://en.wikipedia.org/wiki/Short_ton
10. https://en.wikipedia.org/wiki/Long_ton
11. https://en.wikipedia.org/wiki/Tonne
12. https://en.wikipedia.org/wiki/65535_%28number%29
13. https://en.wikipedia.org/wiki/2147483647_%28number%29
14. https://en.wikipedia.org/wiki/4294967295
15. https://en.wikipedia.org/wiki/555_%28telephone_number%29
16. https://en.wikipedia.org/wiki/Null_Island
17. https://en.wikipedia.org/wiki/Null_Island
18. https://en.wikipedia.org/wiki/Mojibake
19. https://nicercode.github.io/blog/2013-04-30-excel-and-line-endings/
20. http://tabula.technology/
21. https://www.census.gov/programs-surveys/acs/
22. http://inflation-adjust.herokuapp.com/
23. https://en.wikipedia.org/wiki/Standard_deviation
24. https://en.wikipedia.org/wiki/Median_absolute_deviation
25. https://github.com/Quartz/bad-data-guide/blob/master/p.org/en/content/gender-inequality-index-gii
26. http://fivethirtyeight.com/features/science-isnt-broken
27. https://en.wikipedia.org/wiki/Benford's_law