# Cleaning a Public-Health Incident Dataset

*Tracey Atieno Onyango-Data cleaning project in Microsoft Excel*

## About the dataset

I wanted a practice dataset that was genuinely messy rather than something already tidy, so I used the "Attacks on Public Health Incident Data" file published by Insecurity Insight, a humanitarian organisation that pulls together agency and open-source reports on events affecting public-health programmes such as COVID-19, polio and Ebola campaigns. Here is the link to the data files: [https://data.humdata.org/dataset/attacks-on-vaccination-campaigns](https://data.humdata.org/dataset/attacks-on-vaccination-campaigns). The raw file had 199 records across 28 columns, compiled from several different partner sources, which is exactly the kind of data where small inconsistencies creep in. It also sits close to the monitoring and programme data I have worked with, so it felt like a realistic test.

## What I found in the raw data

Before changing anything, I went through the file column by column to see what I was dealing with. A few things stood out. The country field mixed full names with abbreviations — "CAR", "DRC", "OPT", "UK" and "USA" all appeared alongside properly spelled-out names — although the ISO code column next to it was clean and consistent. Three columns (Event Description, Latitude and Longitude) were completely empty across all 199 rows. The six disease columns were coded in an awkward way: each one held the disease name as text where it applied and was simply left blank otherwise, which makes them hard to count or chart. The "killed" and "kidnapped" count columns were mostly blank, with no explicit zeros anywhere. And one detail I nearly missed — the file is titled 2016 to 2026, but the earliest record is actually from January 2020.

## How I cleaned it, and the decisions along the way

For the country names, rather than fixing each abbreviation by hand, I used the ISO code column as my anchor since it was already reliable. I built a small lookup table mapping each ISO code to its full country name and used VLOOKUP to bring the corrected names across. Repairing the messy column using the clean one felt more defensible than editing entries one by one. "OPT" was the one naming decision I thought about most — I went with "Occupied Palestinian Territory" because that is the label tied to the ISO 3166 standard, which kept the choice neutral rather than something I decided myself.

The three empty columns were easy — they held nothing, so I removed them. What I didn't expect was that my lookup values had introduced trailing spaces on a few country names, so "Kenya" came through as "Kenya ". That would have shown up as two separate countries in a pivot, which is a bit ironic for a cleaning project, so I trimmed them out once I spotted it.

The disease columns took the most thought. My first instinct was to collapse all six into a single "disease context" column, but before doing that I ran a COUNTA check across the six columns to see whether any incident was tagged to more than one disease. It turned out several were linked to two or three at once, so a single column would have quietly thrown away the extra ones. I kept a separate 0/1 flag for each disease instead, which preserves every tag.

Finally, for the count columns, I treated the blanks as zeros using a helper column, on the assumption that an unreported count means nothing happened rather than that the information is missing. I've noted this as an assumption rather than a fact, since it's the kind of thing worth being explicit about.

## Checking the cleaned data actually worked

Once everything was cleaned, I built three quick summaries to confirm the data was ready to work with — things that would have been awkward or misleading on the raw file. Grouping incidents by country showed Pakistan and Afghanistan accounting for most of the records, with all 34 country names now clean and consistent. Plotting incidents by year made the earlier metadata point visible: the line starts in 2020, not 2016, and the lower figure for 2026 simply reflects a partial year, since the data only runs to July, rather than any real drop. The disease chart came from summing the individual flags; because some incidents involved more than one disease, those totals overlap on purpose and shouldn’t be read as neat mutually exclusive shares.

## Where it ended up

The cleaned dataset is now consistent, correctly typed and ready for analysis, and I kept the original data on its own sheet so the before-and-after is easy to follow. I did this in Excel to keep every step visible and easy to check, though the same process could just as easily be scripted in R or Python if reproducibility mattered more than transparency. Source: Insecurity Insight (insecurityinsight.org).
Link to Excel file:

Data Cleaning Project -Public Health Incident.xlsx