**When do we need to consider data cleansing activities? What are the usual steps? Have you ever performed any of these actions? Is there ever a need to gather and/or use large datasets at least some of which is in "spreadsheet" form?**

Realistically we should always be doing data cleansing when working with new data. The first sentence for the Wikipedia entry says, “Data cleansing…is the process of *detecting*…inaccurate records.” Anytime prior to the data being used, some level of inspection should occur, either passively (executed by a script for instance) or actively (by a user manually sifting through the data). With the former, we can even set up a script to ensure the data is in the correct format as it gets written to the database.

Elite Data Science has an interesting entry on data cleaning (<https://elitedatascience.com/data-cleaning>). I recently had to perform many of these steps for a work project which required downloading weather data from NOAA and merging it with data generated by our building monitoring system. Both raw datasets had many more entries than I needed, as well as missing or incoherent entries. I had to filter out the irrelevant data and figure out how to handle the missing entries.

I can imagine many reasons why large datasets might show up in “spreadsheet” form. One that is relevant to my work experience is data that is collected from manufacturing production records. At our company, this information is still captured by hand (some companies are moving towards electronic batch records) which requires one individual to manually enter the information into Excel at the end of a campaign. If we want to perform data analysis on it, we would need to use the spreadsheets.