General Assembly

STUDY GUIDE

CLEANING YOUR DATA

Dirty Data

Dirty data is data that contains problematic values. These can be illegal, incorrect, duplicate, or missing. Extreme values, or outliers, can also be problematic.

The most common types of dirty data are:

- **Illegal Values**: Values that do not conform to the logic of the system.
- **Incorrect Values**: Values that are simply not correct.
- **Missing Values**: Values that aren't present when they should be.
- **Duplicates Values**: Values that appear more than once when they should not.
- **Clean Data**: Data that are free of errors and therefore analyzable.
- **Dirty Data**: Data that contain errors that will impede your analysis.

Data Cleaning Tips

1. Always make a copy of your data set *before* attempting any cleaning or analysis. This could be a duplicate file or a copy of the data set in a new worksheet within the same file.
2. It's best practice to record every step you take when cleaning your data. You can use this as a template: "I did x because y and now have z."
3. Ask questions about the source of the data, the units of the data, and the methods of data collection.

Data Entry Methods

- **Manual: Free-form**: Expect misspellings, illegal values, and many variations of the same value (e.g., New York City and NYC). Ask yourself: Is this field mandatory to your analysis? If so, you'll likely need to perform a lot of cleaning.
- **Manual: Selected from a list**: Expect better integrity, as individuals had to select from a predetermined list of options. As yourself: Was an answer required? Was there an opt-out option like "not applicable"?
- **System- or software-generated**: Expect higher data integrity. This data is at the mercy of the software or system.

Human errors vs. machine errors

- Human errors are those caused by human interaction in the data collection process.
- Machine errors are those caused by machines in the data collection process.

Individual/one-off errors vs. systematic errors

- Individual errors can be harder to find and fix because they don't fit into a pattern.
- Systematic errors are patterns of error within a data set.