JOHNS HOPKINS
KRIEGER SCHOOL
of ARTS & SCIENCES

CENTERS FOR
Civic Impact

What Works Cities
RESOURCE

Data Quality Problems and Solutions

Good data quality is an essential foundation for any plan to use data. Overall, quality refers to the data's suitability for the planned use. We must assess datasets to understand their level of quality, at the level of individual fields. Then we must take action to improve data quality. Strategies for improving data quality include audits and correction, front-end validation, and standards.

Problem	Problem Description	Solution
Inconsistent formatting	Data within the same field or column do not adhere to the same format. This makes the dataset difficult to organize, as attempts to sort or group the data will have problems. Also, calculations may not be able to handle the inconsistency. Addresses are a common source of this problem. For example: some ZIP codes are recorded in 5 digit format, and others in 5 + 4 format.	Adopt and enforce a data standard, specifying the format for each field. If your tool allows for autocompletion or validation at the time of data entry, this can prevent entry of bad data. To clean existing data, use a tool such as Openrefine, which enforces formatting rules specified by the user. For addresses, consider using an address ID instead of the actual address.
Duplicate values	Duplicate rows sometimes arise when datasets are combined or otherwise transformed. Duplicate rows pose a problem for analysis as it will inflate counts.	Many software tools have a function to find and handle duplicates. When duplicates are found, look at collection procedures to understand how the duplicate occurred.

Problem	Problem Description	Solution
Missing values	No data for a specified row-column pair. Also known as “blank” or “null” values. This may be expected or allowed, e.g. the second line of a street address may be blank. Sometimes missing values appear as values that make no sense in context, like “--” or “9999.” Unexpected missings are a problem for analysis, especially if they are not random. The analysis will suffer from not having a complete picture of reality.	The best solution is to do additional data collection to fill in the missings. When this is not possible, another solution for a numeric column is to fill the blank cells with the median or average value of all the values in the column. This technique is known as “imputation.” It may also be possible to interpolate the data: analyzing similar datapoints to calculate what the likely missing value would be.
Wrong data type	Software programs often specify a data type for each field or column. These types include numeric, text, and datetime. It is possible to enter data in the wrong type, such as numeric data that is entered as text. In these cases, analysis will suffer because the software will not be able to perform numeric calculations.	Many software tools have functions to convert a field’s data type. When using these, it is important to review all data in that field, to ensure that the conversion does not result in loss of data.
Accidental deletion	Users may accidentally delete or overwrite (replace) either entire data tables or individual datapoints.	Make back-ups of data, and perform data analysis operations in working copies, not in the source. Protect source data as read-only.
Invalid values	Numeric: data is out of range, such as a negative value for “day of the year.” Categorical: values not in the specified list, such as any text other than the seven days of the week, in a field for day of the week.	Perform additional data collection, or replace the incorrect value with a missing and treat it accordingly.

“Data Quality Problems and Solutions”