www.keepgoing.ai

DATA CLEANING

2022

Table of Contents

Introduction 0
A. The Data Cleaning Process 0
B. Sources of Error..1
C. First Things First 2
D. Screening Data 2
E. Diagnosing Data 4
F. Treatment of Data 4
G. Missing Values 5
H. Documenting Changes 6
I. Adapt Process 7
J. Recoding Variables 7
K. Quality Control Procedures 9
L. Data Integration 10
M. Key Principles for Data Cleaning 10
N. Tools and Tutorials for Data Cleaning 11
O. Sources and Background Readings 11
Annex 1 – Checklist for Data Cleaning 13

Introduction

No matter how data are collected (in face-to-face interviews, telephone interviews, self-administered questionnaires, etc.), there will be some level of error. Messy data" refers to data that is riddled with inconsistencies. While some of the discrepancies are legitimate as they reflect variation in the context, others will likely reflect a measurement or entry error. These can range from mistakes due to human error, poorly designed recording systems, or simply because there is incomplete control over the format and type of data imported from external data sources. Such discrepancies wreak havoc when trying to perform analysis with the data. Before processing the data for analysis, care should be taken to ensure data is as accurate and consistent as possible.

Used mainly when dealing with data stored in a database, the terms *data validation, data cleaning or data scrubbing* refers to the process of detecting, correcting, replacing, modifying or removing messy data from a record set, table, or database.

This document provides guidance for data analysts to find the right data cleaning strategy when dealing with needs assessment data. The guidance is applicable to both primary and secondary data. It covers situations where:

- Raw data is generated by assessment teams using a questionnaire.
- Data is obtained from secondary sources (displacement monitoring systems, food security data, census data, etc.)
- Secondary data is compared or merged with the data obtained from field assessments

This document complements the DR.KESKIN technical note on How to approach a dataset which specifically details data cleaning operations for primary data entered into an Excel spreadsheet during rapid assessments.

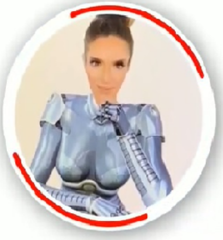
A. The Data Cleaning Process

Data cleaning consists primarily in implementing error prevention strategies before they occur (see data quality control procedures later in the document). However, error-prevention strategies can reduce but not eliminate common errors and many data errors will be detected incidentally during activities such as:

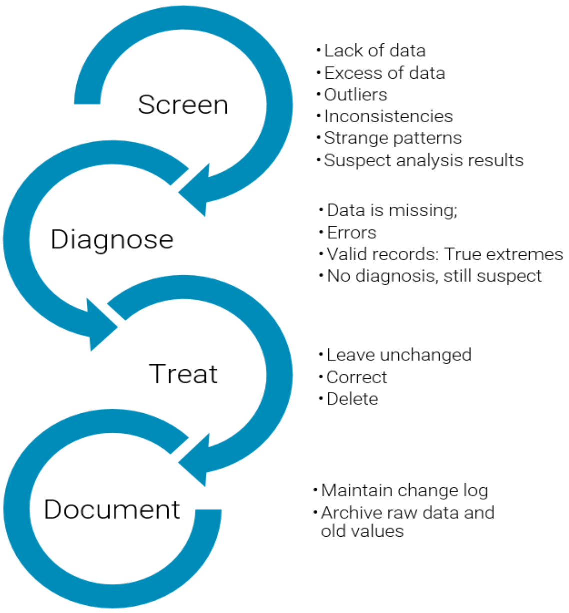
- When collecting or entering data
- When transforming/extracting/transferring data
- When exploring or analysing data
- When submitting the draft report for peer review

Even with the best error prevention strategies in place, there will still be a need for actively and systematically searching for, detecting and remedying errors/problems in a planned way.

Data cleaning involves repeated cycles of screening, diagnosing, treatment and documentation of this process. As patterns of errors are identified, data collection and entry procedures should be adapted to correct those patterns and reduce future errors.



The four steps of data cleaning:



Adapted from Van den Broeck J, Argeseanu Cunningham S, Eeckels R, Herbst K (2005) and Arthur D. Chapman

Screening involves systematically looking for suspect features in assessment questionnaires, databases, or analysis datasets.

The diagnosis (identifying the nature of the defective data) and treatment (deleting, editing or leaving the data as it is) phases of data cleaning requires an in depth understanding of all types and sources of errors possible during data collection and entry processes.

Documenting changes entails leaving an audit trail of errors detected, alterations, additions and error checking and will allow a return to the original value if required.

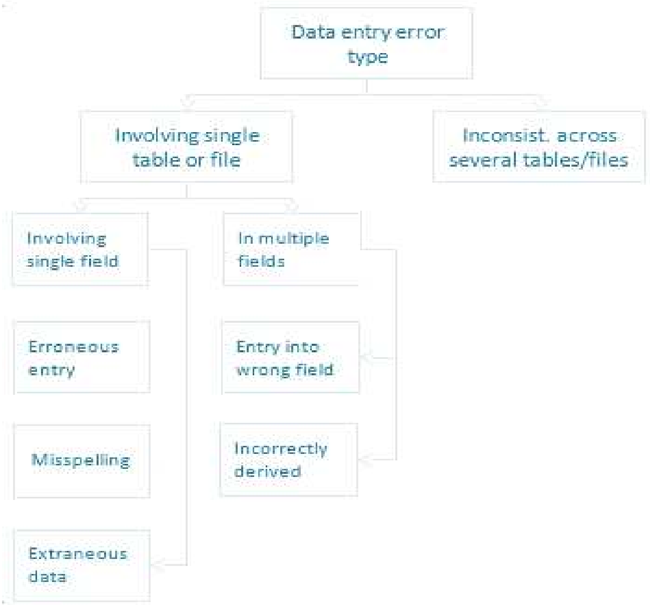
B. Sources of Error

After measurement, data are the object of a sequence of typical activities: they are entered into databases, extracted, transferred to other tables, edited, selected, transformed, summarized, and presented. It is important to realise that errors can occur at any stage of the data flow, including during data cleaning itself. Many of the sources of error in databases fall into one or more of the following categories:

Measurement errors: Data is generally intended to measure some physical process, subjects or objects, i.e. the waiting time at the water point, the size of a population, the incidence of diseases, etc. In some cases, these measurements are undertaken by human processes that can have systematic or random errors in their design (i.e., improper sampling strategies) and execution (i.e., misuse of instruments, bias, etc.). Identifying and solving such inconsistencies goes beyond the scope of this document. It is **recommended to refer to the DR.KESKIN Technical Brief** *How sure are you?* to get an understanding of how to deal with measurement errors.

Data entry error: "Data entry" is the process of transferring information from the medium that records the response (traditionally responses written on printed questionnaires) to a computer application. Under time pressure, or for lack of proper supervision or control, data is often corrupted at entry time.

Main errors include



Adapted from Kim et Al, 2003; Aldo Benini 2013

- An erroneous entry happens if, e.g., age is mistyped as 26 instead of 25.
- Extraneous entries add correct, but unwanted information, e.g. name and title in a name-only field.
- Incorrectly derived value occurs when a function was incorrectly calculated for a derived field (i.e. error in the age derived from the date of birth).
- Inconsistencies across tables or files occur e.g. when the number of visited sites in the province table and the number of visited sites in the total sample table do not match.

A large part of the data entry errors can be prevented by using an electronic form (e.g. ODK) and conditional entry.

Processing errors: In many settings, raw data are pre-processed before they are entered into a database. This data processing is done for a variety of reasons:

to reduce the complexity or noise in the raw data, to aggregate the data at a higher level, and in some cases simply to reduce the volume of data being stored. All these processes have the potential to produce errors.

Data integration errors: It is rare for a database of significant size and age to contain data from a single source, collected and entered in the same way over time. Very often, a database contains information collected from multiple sources via multiple methods over time. An example is the tracking of the number of people affected throughout the crisis, where the definition of "affected" is being refined or changed over time. Moreover, in practice, many databases evolve by merging other pre-existing databases. This merging task almost always requires some attempt to resolve inconsistencies across the databases involving different data units, measurement periods, formats etc. Any procedure that integrates data from multiple sources can lead to errors. The merging of two or more databases will both identify errors (where there are differences between the two databases) and create new errors (i.e. duplicate records). Table 1 below illustrates some of the possible sources and types of errors in a large assessment, at three basic levels: When filling the questionnaire, when entering data into the database and when performing the analysis.

Table 1: Sources of data error

	Sources of error	
Stage	Lack or excess of data	Outliers and inconsistencies
Measurement	• Form missing • Form double, collected repeatedly • Answering box or options left blank • More than one option selected when not allowed	• Correct value filled out in the wrong box • Not readable • Writing error • Answer given is out of expected (conditional) range
Entry	• Lack or excess of data transferred from the questionnaire • Form of field not entered • Value entered in wrong field • Inadvertent deletion and duplication during database handling	• Outliers and inconsistencies carried over from questionnaire • Value incorrectly entered, misspelling • Value incorrectly changed during previous data cleaning • Transformation (programming) error
Processing and Analysis	• Lack or excess of data extracted from the database • Data extraction, coding or transfer error • Deletions or duplications by analyst	• Outliers and inconsistencies carried over from the database • Data extraction, coding or transfer error • Sorting errors (spreadsheets) • Data-cleaning errors

Adapted from Van den Broeck J, Argeseanu Cunningham S, Eeckels R, Herbst K (2005)

Inaccuracy of a single measurement and data point may be acceptable, and related to the inherent technical error of the measurement instrument. Hence, data cleaning should focus on those errors that are beyond small technical variations and that produce a major shift within or beyond the analysis. Similarly, and under time pressure, consider the diminishing marginal utility of cleaning more and more compared to other demanding tasks such as analysis, visual display and interpretation.

- Understand when and how errors are produced during the data collection and workflow.
- Resources for data cleaning are limited. Prioritisation of errors related to population numbers, geographic location, affected groups and date are particularly important because they contaminate derived variables and the final analysis.

The following sections of this document offer a step by step approach to data cleaning.

C. First Things First

The first thing to do is to make a copy of the original data in a separate workbook and name the sheets appropriately, or save in a new file.

ALWAYS keep the source files in a separate folder and change its attribute to READ-ONLY, to avoid modification of any of the files.

D. Screening Data

To prepare data for screening, tidy the dataset by transforming the data in an easy to use format.
Within a tidied dataset:

- Fonts have been harmonised
- Text is aligned to the left, numbers to the right
- Each variable has been turned into a column and each observation into a row.
- There are no blank rows
- Column headers are clear and visually distinct.
- Leading spaces have been deleted

Afterwards, examine data for the following possible errors:

- Spelling and formatting irregularities: are categorical variables written incorrectly? Is the date format consistent? For numeric fields, are all of the values numbers? Etc.
- Lack of data: Do some questions have far fewer answers compared to others?
- Excess of data: Are there duplicate entries or more answers than originally allowed?
- Outliers/inconsistencies: Are there values that are so far beyond the typical distribution that they seem potentially erroneous?
- Remarkable patterns: Are there patterns that suggest that the respondent or enumerator has not answered or recorded questions honestly? (i.e. several questionnaires with the exact same answers)?
- Suspect analysis results: Do the answers to some questions seem counterintuitive or extremely unlikely?

Common mistakes in needs assessments

- Misspelling of place names, particularly when translating between different alphabets (e.g. Arabic to English)
- Use of inconsistent date formats
- Totals differ from the results of disaggregate questions (e.g. total number of household members is not matching the aggregation of a different question where respondents are asked to list household members by age and gender)
- Values are outside of the acceptable range for values for that question, including negative values in fields that can only have positive values (e.g. price of bread)
- Unclear cause of missing data
- Merging of datasets with different units of measurement (e.g. different interpretations of the term household) or administrative boundaries.
- In case of multiple choice questions: selecting ‘other, please specify’ for a variable that is one of the multiple choice options.
- Malfunctioning skip patterns
- Overall lack of consistency within the answers provided by a respondent (e.g. the response to questions on main needs is not in line with sector specific questions).

Data cleaning can be partly automated through statistical software packages Descriptive statistic tools can for instance be used during the screening phase to predefine expectations, assumptions or criteria about normal ranges, distribution shapes, and strength of relationships. This can facilitate the flagging of dubious data, patterns, or results.

However, screening methods are not only statistical. Many outliers are detected by perceived non-conformity with prior expectations or the norm. This is for instance based on the analyst's experience, results from secondary data review, numerical constraints or plane common sense (weight cannot be negative, people cannot have more than 2 parents, women cannot bear 35 children, etc.).

A particular problem is that of erroneous inliers, i.e., data points generated by error but falling within the expected range. Erroneous inliers will often escape detection. Detection approaches include:

- Viewing data in relation to other variables, using multivariate views, such as scatter plots or heatmap.
- More advanced and resource intensive techniques involves regression analysis, consistency/ plausibility checks (examining the history of each data point or comparing to a similar location) or by re-measurement. However, under time pressure, such examination is rarely feasible. Instead, one can examine and/or re-measure or do further inquiry about a sample of inliers to estimate an error rate.

Useful screening methods, from simpler to more complex, are:

- Screening of columns after sorting
- Use summary statistics
- Validated and/or double data entry
- Printouts of variables not passing range checks and of records not passing consistency checks
- Frequency distributions and cross-tabulations.
- Graphical exploration of distributions: box plots, histograms, and scatter plots using visual analysis software such as Tableau desktop
- Plots of repeated measurements on the same individual, i.e., growth curves
- Checking of questionnaires using fixed algorithms
- Statistical outlier detection.

- In many, if not most instances, data can only be cleaned effectively with some human involvement. Know (and/or train data cleaners) common mistakes and what errors to look for.
- Exploratory data analysis and data visualization are two main ways to detect data errors.
- Different types of errors call for different detection approaches – a spell check will recognise misspelled categorical variables while statistical outlier detection helps identification of extreme values.

E. Diagnosing Data

The identification or highlighting of an error is followed by diagnosis – finding the cause for this error. To clarify suspect data, review all of a respondent's answers to determine if the data makes sense within the context. Sometimes it is necessary to review a cross-section of different respondents' answers, to identify issues such as a skip pattern that was specified incorrectly.

There are a multitude of possible diagnoses for each suspected data point:

- Missing data: Answers omitted by the respondent (nonresponse), questions skipped by the enumerator or dropout
- Errors: Typos or answers that indicate the question was misunderstood.
- True extreme: An answer that seems high but can be justified by other answers (i.e., the respondent working 60 hours a week because they work a full-time job and a part-time job)
- True normal: A valid record.
- No diagnosis, still suspect: Make a judgment call on how to treat this data during the treatment phase.

Some data values are clearly logically or biologically impossible (men cannot be pregnant; the price of bread cannot be negative). Pre-defined cut-off points immediately help to detect this type of error. Sometimes the suspected value falls within the acceptable range and the diagnosis is less straightforward. In these cases, it is necessary to apply a combination of diagnostic procedures:

- Go back to previous stages of the data flow to see whether a value is consistently the same. This requires access to well-archived and documented data with justifications for any changes made at any stage.
- Look for information that could confirm the true extreme status of an outlying data point. For example, a very low score for weight-for-age (i.e., −6 Z-scores) might be due to errors in the measurement of age or weight, or the subject may be extremely malnourished, in which case other nutritional variables should also have extremely low values. This type of procedure requires insight into the coherence of the variables. This insight is usually available from experience or lessons learnt and can be used to plan and program data cleaning.
- Collect additional information, i.e., question the enumerator about what may have happened and, if possible or necessary, repeat the measurement. Such procedures can only happen if data cleaning starts soon after data collection.

The diagnostic phase is labour intensive and the budgetary, logistical, time and personnel requirements are typically underestimated or even neglected at the design stage. Fewer resources are required if conditional data entry is used (e.g. through electronic forms) and if it starts early in data collection process.

- Use common sense, experience, triangulation and lessons learnt to diagnose the type of error.
- Design your questionnaire form carefully to allow cross checks between questions.
- Consider the collection of additional information from enumerator to understand the cause of errors (i.e. debriefings)

F. Treatment of Data

After identification of missing values, errors, and true (extreme or normal) values, analysts must decide what to do with problematic observations:

- Leave it unchanged: The most conservative course of action is to accept the data as a valid response and make no change to it. The larger the sample size, the less one suspect response will affect the analysis; the smaller the sample size, the more difficult the decision.
- Correct the data: If the respondent's original intent can be determined, correct the answer (i.e. after discussing with the enumerator, it is clear that the respondent meant the lack of income instead of too much income).
- Delete the data? The data seems illogical and the value is so far from the norm that it will affect descriptive or inferential statistics. What to do? Delete just this response or delete the entire record? Remember that whenever data is deleted, there is a risk of consciously or subconsciously "cherry picking" the data to obtain the preferred results. To understand the impact of deleting a data point, a binary variable can be created (1=suspicious record, 0=not suspicious). This new variable can be used as a record filter in Pivot tables or in-table filtering to understand the impact of potentially erroneous data in the final results.
- If time and resources allow, re-measure the suspect or erroneous values.

There are some general rules to support a decision on how to treat the data:

- If the person undertaking data entry has entered values different from the ones in the questionnaire, the value should be changed to what was recorded in the questionnaire form. (I.e. the value in the questionnaire was 40,000 and the data entry operator keyed in 4,000 – a zero was left out).
- When variable values do not make sense, if there is no data entry error, and there are no notes to help determine where the error comes from, **leave the data as it is.** By changing the value into a more reasonable result, significant bias is introduced and there is no justification for changing it. The case should be listed as an outlier (i.e. by using conditional formatting for instance).
- When blank cells are found or the record was required even though key informants may not have that type of data or duplicate records were entered, then cases must be deleted from the data file.
- Impossible values are never left unchanged, but should be corrected if a correct value can be found, otherwise they should be deleted. For biological continuous variables, some within-subject variation or small measurement variation could be present. If a re-measurement is done very rapidly after the initial one and the two values are close enough to be explained by variation alone, take the average of both as the final value.
- With true extreme values and values that are still suspect after the diagnostic phase, the analyst should examine the influence of such data points, individually and as a group, on analysis results before deciding whether or not to leave the data unchanged.
- To limit the impact of outliers and extreme values analysts can decide to present the median. This is acceptable as long as clearly explained within the findings.
- Some authors have recommended that true extreme values should always stay in the analysis. In practice, many exceptions are made to that rule. The investigator may not want to consider the effect of true extreme values if they result from an unanticipated extraneous process. This becomes an "a posteriori" exclusion criterion. The data points should be reported as "excluded from analysis" in the methodology chapter of the final report.

G. Missing Values

Missing values require particular attention. The first thing is to decide which blank cells need to be filled with zeros (because they represent genuine negative observations, such as "no", "not present", "option not taken", etc.) and which to leave blank (if the convention is to use blanks for missing or N/A for "not applicable"). Some analysts replace blank cells with some explicit missing value code (e.g. using 999 to indicate a "do not know").

What to do with those cells remaining blank? Missing values can be classified as either random or non-random:

- Random missing values may occur because the subject inadvertently did not answer some questions. The assessment may be overly complex or too long, or the enumerator may be tired or not paying sufficient attention, and miss the question. Random missing values may also occur through data entry mistakes. If there are only a small number of missing values in the dataset (typically, less than 5%), then it is extremely likely to be a random missing value.
- Non-random missing values may occur because the key informant purposefully did not answer some question. This for instance occurs if the question is confusing, not appropriate or perceived as sensitive. The missing data is related to one or more characteristics of the respondent – e.g. if female are more likely to refuse a question on the level of income compared to male respondents.

The default option for dealing with missing values is to filter and exclude these values from analysis:

- Listwise / casewise deletion: All cases (e.g. a respondent) that have missing values are excluded. If only one variable is analysed, listwise deletion is simply analysing the existing data. When analysing multiple variables, then listwise deletion removes cases if there is a missing value on any of the variables. The disadvantage is the loss of data that occurs as all data is removed for a single case, even if some questions were answered.
- Pairwise deletion: Unlike listwise deletion which removes cases that have missing values on any of the variables under analysis, pairwise deletion only removes the specific missing values from the analysis (not the entire case). In other words, all available data is included. When conducting a correlation on multiple variables, this technique enables a bivariate correlation between all

available data points, and ignores only those missing values if they exist on some variables. In this case, pairwise deletion will result in different sample sizes for each variable. Pairwise deletion is useful when the sample size is small or if missing values are large because there are not many values to start with.

Try conducting the same test using both deletion methods to see how the outcome changes. Note that in these techniques, "deletion" means exclusion within a statistical procedure, not deletion (of variables or cases) from the dataset.

A second option is to delete all cases with missing values. Thus, you are left with complete data for all cases. The disadvantage to this approach is that the sample size of the data is reduced, resulting in a loss of statistical power and increased error in estimation (wider confidence intervals). It can also affect the representativeness of a sample: after removing the cases with non-random missing values from a small dataset, the sample size could be insufficient. In addition, results may be biased in case of non-random missing values. The characteristics of cases with missing values may be different than the cases without missing values.

Another option is imputation: to replace the missing values. This technique preserves all cases by replacing missing data with a probable value based on other available information. A simple procedure for imputation is to replace the missing value with the mean or median. Hot-deck imputation replaces missing values with the value for that same variable taken from a complete record for a similar person in the same dataset. Once all missing values have been imputed, the data set can then be analysed using standard techniques for complete data. However, this method can also bias results and p-values.

Under certain conditions, maximum likelihood approaches have also proven efficient to dealing with missing data. This method does not impute any data, but rather uses all the data available for the specific cases to compute maximum likelihood estimates.

Detailing technicalities, appropriateness and validity of each techniques goes beyond the scope of this document. Ultimately, choosing the right technique depends on how much data are missing, why this data is missing, patterns, randomness and distribution of missing values, the effects of the missing data and how the data will be used for analysis. It is strongly recommended to refer to a statistician in case of a small dataset with a large number of missing values.

Pragmatically, for needs assessment with few statistical resources, creating a copy of the variable and replacing missing values with the mean or median may often be enough and preferable to losing cases in multivariate analysis from small samples.

- There are several methods to deal with missing data, including deleting cases with missing values, imputing and the maximum likelihood approach. However, providing an explanation on why data are missing ("women could not be interviewed", “the last questionnaire section could not be filled due to lack of time”) may be much more informative to end user’s than a plethora of statistical fixes.
- Set up a dummy variable with value 0 for those who answered the question and value 1 for those who did not. Use this variable to show the impact of different methods.
- Look for meaning in non-random missing values. Maybe the respondents are indicating something important by not answering one of the questions.

H. Documenting Changes

Documentation of error, alterations, additions and error checking is essential to:

- Maintain data quality
- Avoid duplication of error checking by different data cleaners.
- Recover data cleaning errors
- Determine the fitness of the data for use.
- Inform users who may have used the data knowing what changes have been made since they last accessed the data

Create a change log within the workbook, where all information related to modified fields is sourced. This will serve as an audit trail showing any modifications, and will allow a return to the original value if required. Within the change log, store the following information:

- Table (if multiple tables are implemented)
- Column, Row
- Date changed
- Changed by
- Old value
- New value
- Comments

- Make sure to document what data cleaning steps and procedures were implemented or followed, by whom, how many responses were affected and for which questions.
- ALWAYS make this information available when sharing the dataset internally or externally (i.e. by enclosing the change log in a separate worksheet)

I. Adapt Process

Once errors have been identified, diagnosed, treated and documented and if data collection/entry is still ongoing, the person in charge of data cleaning should give instructions to enumerators or data entry operators to prevent further mistakes, especially if they are identified as non-random. Feedback will ensure common errors are not repeated and will improve the assessment validity and the precision of outcomes. Main recommendations or corrections can include:

- Programming of data capture, data transformations, and data extractions may need revision.
- Corrections of questions in the questionnaire form.
- Amendment of the assessment protocol, design, timing, enumerators training, data collection, and quality control procedures.
- In extreme cases, it may be necessary to re-conduct some field assessment (few sites) or contact again key informants or enumerators to ask additional information or more details or confirm some records.

- Data cleaning often leads to insight into the nature and severity of error-generating processes.
- Identify basic causes of errors detected and use that information to improve data collection and the data entry process to prevent those errors to re-occurring.
- Reconsider prior expectations and/or review or update quality control procedures.

J. Recoding Variables

Variables might need to be recoded to create new ones that fit within the analytic needs. The following recoding for instance is common:

- Formatting: date (day, month, and year), pre-fixes to create better sorting in tables
- Rounding continuous variables
- Syntax: Translation, language style and simplification.
- Recoding a categorical variable (e.g. ethnicity, occupation, an "other" category, spelling corrections, etc.).
- Recoding a continuous variable (e.g. age) into a categorical variable (e.g. age group).
- Combining the values of a variable into fewer categories (e.g. grouping all problems caused by access constraints).
- Combining several variables to create a new variable (e.g., the food consumption score, building an index based on a set of variables).
- Defining a condition based on certain cut-off points (e.g., population "at risk" vs. "at acute risk").
- Changing a level of measurement (e.g. from interval to ordinal scale).

Conceptually, a distinction is needed between:

- Activities related to recoding qualitative data: i.e. responses to open questions
- Activities that include transforming and deriving new values out of others, such as creating calculations (i.e. percentage), parsing, merging, etc. Here, the analyst is re-expressing what the data says (i.e. re-expressing deviation as a % change, weighted or moving average, etc.). The data has (normally) already gone through a cleaning stage before to be transformed.

For both types, recoding variables or values can serve both the purpose of cleaning dirty data and/or transforming clean data. This section focuses primarily on the former rather than the re-expression of values which will be tackled more extensively in another chapter of the data handbook on data transformation.

Recoding categorical variables starts with a full listing of all variants generated by a variable, together with their frequencies. The variant list can be copied into a new sheet, to create a table of variants and their desired replacements. ALWAYS keep a copy of the original values, and try out different recoding schemes before settling on a final one.

There are three ways to recode categorical data: 1. Collapse a categorical variable into fewer categories

What is the current occupation of the head of household?

Original categories	Recoded categories
Government employee	Agriculture
Agriculture and livestock	Non-Agriculture
Commerce	

Student
…etc.

Guidelines for collapsing data are as follows:

- Ordinal variables need to be collapsed in a method that preserves the ordering of categories.
- Combine only categories that go together. Do not combine two logically distinct categories just to eliminate categories with small numbers (e.g. lack of access due to lack of income and lack of access due to insecurity) as interpretation of data becomes difficult or meaningless.
- The way in which categories are collapsed can easily affect the significance level of statistical tests. Categories should be collapsed to avoid the criticism that the data were manipulated just to get a certain result. This does not mean there needs to be a decision before collecting the data (if this was the case, there would not be a need to collect separate categories).
- Do not oversimplify the data. Unnecessary reduction in the number of categories may reduce statistical power and obscure relationships in the data. As a general guideline, keep intact any categories that include 10% or more of your data (or 5 cases, for very small samples).

Breaking: There are several reasons for breaking a categorical variable into several smaller variables:

- Data was collected in a manner easy to collect to ease the burden of data collection on the subject. For example, it is easier for the key informant to provide a list of issues than to review a long list of problems.
- A variable may contain more than one "concept." For example, consider the ordinal variable "severity" below:

    1. There are no shortages
    2. A few people are facing shortages
    3. Many people are facing shortages
    4. Shortages are affecting everyone

This variable contains two concepts, "shortages" and "number of people affected". It is straightforward to code two new variables, shortages (0 = no shortages, 1 = shortage) and number of people (0 = no people, l= Few people, 2=Many people, 4= All of them).

Combining is the reverse process of breaking up, such as combining "shortages" and "number of people" back into the variable "severity".

Recoding variables can be tedious. The conceptual effort needed in order to produce a meaningfully recoded category set is often underestimated. Care must be taken to evaluate the combined category sets, to absorb excessive, incoherent or rarely used categories into broader ones, and to be clear about the rationale for the final number and content of distinct categories. Also, be aware that any recoding that reduces the number of categories entails some information loss. As in all stages of data analysis, analysts must be alert for errors.

Basic tips for effective recoding include:

- *Use distinct and easy to remember variable names.* Never use the same variable name to denote both the transformed and untransformed variable. For large data sets, a systematic way to name variables is desirable.
- *Pay attention to missing values.* When recoding is done, the number of cases with missing data should be the same as before recoding. A check that this is so will often be the first clue that recoding was in error. A safe procedure is to start the recoding process by setting the new variable to missing for all cases, and then changing missing values only for those with data on the initial variables to be recoded. For complicated recoding, check a few individual values by hand to make sure they were recoded properly, and check the distribution of values.
- *Use graphs to check the accuracy of recoding.* Recoding is a systematic translation of data values, so scatterplots of raw data *vs.* recoded data should show highly organized patterns reflecting the recoding system. Histograms can show whether your data is now more normally distributed.
- *Use variable codes consistently.* For example, with dichotomous "yes/no" variables, always use 0 = no and 1 = yes. For polychotomous variables, always make 0 the reference category/default option.
- *Keep a permanent record of your recoding.* For data entry errors, make the changes directly in the raw data file, because there is no need to maintain the erroneous data points. Recoding should be undertaken in a separate file, as there might be a need to review the initial data. Most statistical programs save data in a specially formatted file, and this file is the one to change. The recoding commands should all be put in one program (a *do* file) that can be executed again. The *do* file serves as a permanent record as well.

K. Quality Control Procedures

When deciding upon an approach to data cleaning, it is useful to consider the different types of errors which can be made, and to plan at what point in the data workflow prevention measures should be implemented. Best practices include:

Roles and responsibilities: Make sure the staff with responsibilities regarding data quality are aware of the cleaning protocols (see annex 1 for a complete checklist for needs assessments). Roles and responsibilities related to error detection and correction should be clearly defined and communicated as part of the job descriptions (see Annex 2), at each stage of the data collection, entry and processing.

Ensure that a second pair of eyes review and compare the original data to data entered. Data cleaning should start in the field (field editing) alongside data collection, as questionnaires are reviewed by supervisors or field editors on a daily basis. Similarly, during data entry, double checks should be mandatory, especially when:

- There is a process of translation at data entry, to ensure consistency/accuracy of translation.
- Data entry is distributed across various field locations and consolidation occurs in a different location.

At the data entry stage, computer-assisted quality control procedures should be used. Additional functionality can be added in the data entry software (i.e. Excel, Survey Gizmo, Access, SPSS, STATA, etc.) to highlight rule violations (null codes, conditional formatting, etc.) and prevent mistakes (i.e. drop down menus). The decision to include those rules in the database must be pragmatic, weighing up the merits of having errors detected and rectified by data entry staff, versus the time required to set this up and to quickly make necessary adjustments if the initial setup does not work as expected.

Five kind of checks can be automated:

- Range checks ensure that every variable in the survey contains only data within a limited domain of valid values. Categorical variables can have only one of the values predefined for them on the questionnaire (for example, gender can be coded only as “1” for males or “2” for females); chronological variables should contain valid dates, and numerical variables should lie within prescribed minimum and maximum values (such as 0 to 120 years for age and should always be expressed as integer of years, with rules for rounding up or down for infants).
- Reference data check are used when the data from two or more closely related fields can be checked against external reference tables, i.e. when the recorded values for height, weight and age are checked against the World Health Organization’s standard reference tables.
- Skip checks verify whether the skip patterns have been followed appropriately. For example, a simple check verifies that questions to be asked only of schoolchildren are not recorded for a child who answered "no" to an initial question on school enrolment.

- Consistency checks verify that values from one question are consistent with values from another question, for example, the date of birth and age of a given individual.
- Typographical checks limit, for instance, the transposition of digits like entering “14” rather than “41” in a numerical input. Such a mistake for age might be caught by consistency checks with marital status or family relation. Control totals, for instance, can significantly reduce typographical errors.

- Document the data quality rules to follow, where focus should be given, and how to solve errors/issues. Plan double checks.
- Communicate clear instructions to enumerators, team leaders, data entry clerk, at all relevant stages of the data flow.
- Ensure that data entry staff are familiar with the questionnaire filling procedures, so that mistakes can be identified early on and verified/rectified (i.e. rules such as ‘pick only three’ or ‘must add to 100 %’).
- Design a data-cleaning plan, including:
- Budget, timeframe and staff requirements.
- Screening tools.
- Diagnostic procedures used to discern errors (on going periodic basis and towards the end of the assessment).
- Instructions or training to enumerators and data entry staff in case of protocol violation and consistency check.
- Decision rules that will be applied in the editing phase.

L. Data Integration

A different set of problems can occur when datasets are integrated or merged with other data. Analysts do not always have control over the format and type of data that is imported from an external data source, such as a database, text file, or a Web page. Most common problems are

- Formats: Dates are especially problematic (26/02/1977, 26 February 1977, 26-02-1977, etc.). Analysts also need to be aware that different applications store dates internally in different ways. A simple copy-paste from one application to another will thus cause errors across the board.
- Units: Different metrics are used: litre, gallons, gourdes, etc.
- Ranges: Age intervals might differ from one survey to another. Only if the birth date is available, the age interval can be recreated.
- Inconsistency: When merging different data source, conflicting information can emerge. Analysts must choose between using both, using the most recently updated information, the most trusted source, investigate further or use neither. Duplicate records should generally not be deleted, but flagged so that they can be identified and excluded from analysis in cases where duplicate records may bias an analysis. While appearing to be duplicates, in many cases the records in the two databases may include some information that is unique to each. Just deleting one of the duplicates ("merge and purge") is consequently not always a good option as it can lead to valuable data loss.
- Spelling: Categorical variables and specifically place names, may have different spellings.
- Loss of bits of data: Some pieces of data, columns or rows are lost when extracted, i.e. when web scrapping or extracting from a pdf (good luck!).

- Data is dirty. Live with it. Analysts assuming that raw data comes clean and bypassing basic checks live dangerously.
- Check the dataset documentation available. If not available (even after request), DO NOT TRUST THE DATA, even if the source is generally reliable. Start checking for quality.
- Even under time pressure, take the time to screen the data for 15-30mn, focusing first on spelling and formatting of the datasets to be merged, then review outliers (use filters for quick visual detection). If no mistakes are spotted during this time interval, it is probably of good quality and usable as it is. If mistakes are detected, then proceed rigorously and methodically to screening, diagnosing and treatment.

M. Key Principles for Data Cleaning

Key principles for cleaning data in spreadsheets are as follows:

1. Create a backup copy of the original data in a separate workbook.
2. Regularly back-up the working file, at successive points during collating, cleaning and analysing. Save documents with file names that combine date and time (yymmdd-time prefixed allow for files to be sorted by order of creation).
3. When integrating or merging different datasets, ensure that the data is in a tabular format of rows and columns with: similar data in each column, all columns and rows visible, and no blank rows within the range. Check that there are no subtotals, totals or other calculated records down the columns. Calculated variables can remain on the right of the data.
4. Format the database for readability and easy navigation: Left align text, number right aligned, variable title is positioned horizontally, text variables fully visible, column separated by bold lines, header with background colours, numbers comma separated every 3 digits, etc.
5. Start with tasks that do not require column such as spell-checking or using the Find and Replace function.
6. Afterwards, undertake tasks that do require column manipulation. The general steps for manipulating a column are:
   - Insert a new column (B) next to the original column (A) that needs cleaning.
   - Transform the data in the column (B).
   - Remove the original column (A), which converts the new column from B to A.
7. Keep the questionnaire close. As each check is done, a list of issues will be produced. The questionnaires should be consulted to double check or identify the problems.
8. When checking for one type of problem for one site or key informant, verify that the data for the other variables for that case have been entered correctly.
9. Look at the values in all the variables and all the cases for that site, key informant or enumerator. Occasionally the data entry person will skip a variable or a key in the values from the previous variable or the subsequent variable, and all the

data that have been entered after will not be correct.

- Planning and budgeting for data cleaning is essential.
- Organizing data improves efficiency, i.e. by sorting data on location or records by enumerator.
- Prevention is better than cure. It is far more efficient to prevent an error than to have to find it and correct it later.
- The responsibility for generating clean data belongs to everyone, enumerators, custodian and users.
- Prioritisation reduces duplication. Concentrate on those records where extensive data can be cleaned at the lowest cost or that are of most value to end users.
- Feedback is a two-way street: data users or analyst will inevitably carry out error detection and must provide feedback to data custodians. Develop feedback mechanisms and encourage users to report errors.
- Education and training improve techniques: Poor training of enumerators and data entry operators is the cause of a large proportion of the errors. Train on quality requirements (readability, etc.) and documentation,
- Data cleaning processes need to be transparent and well documented with a good audit trail to reduce duplication and to ensure that once corrected, errors never re-occur.
- Documentation is the key to good data quality. Without good documentation, it is difficult for users to determine the appropriateness for use of the data and difficult for custodians to know what and by whom data quality checks have been carried out.

N. Tools and Tutorials for Data Cleaning

Spreadsheets such as Excel offer the capability to easily sort data, calculate new columns, move and delete columns, and aggregate data. For data cleaning of humanitarian assessment data, ACAPS developed a specific Technical Note providing a step by step approach in Excel and detailing cleansing operations,

For generic instructions about how to use excel formulas, functionalities or options to clean data, several Microsoft office guidance notes are available:

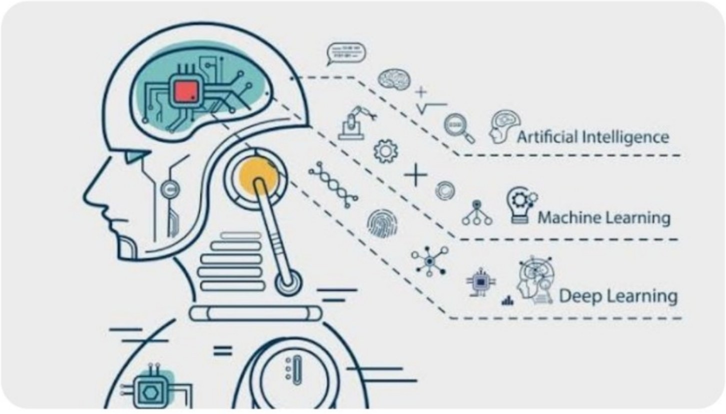
- Spell checking
- Removing duplicate rows
- Finding and replacing text
- Changing the case of text
- Removing spaces and nonprinting characters from text
- Fixing numbers and number signs
- Fixing dates and times
- Merging and splitting columns
- Transforming and rearranging columns and rows
- Reconciling table data by joining or matching
- Third-party providers

Openrefine (ex-Google Refine) and LODRefine are powerful tools for working with messy data, cleaning it, or transforming it from one format into another. Videos and tutorials are available to learn about the different functionalities offered by this software. The facets function is particularly useful as it can very efficiently and quickly gives a feel for the range of variation contained within the dataset.

Detailed data cleansing tutorials and courses are also available at the school of data:

- http://schoolofdata.org/handbook/recipes/cleaning-data-with-spreadsheets/
- http://schoolofdata.org/handbook/courses/data-cleaning/

Two specialized tools to accomplish many of these tasks are used in ACAPS. The first one is Trifacta Wrangler, the new version of Data Wrangler by the Stanford Visualization Group. Trifacta Wrangler is a user-friendly tool that can automatically find patterns in the data based on things selected, and automatically makes suggestions of what to do with those patterns. Beautiful and useful. The other cleaning star software is Data monarch from Datawatch, integrating a lot of wrangling, cleaning and enrichment functionalities that can take hours in Microsoft excel.



Sources and Background Readings

- Benini, A.. 2011. *Efficient Survey Data Entry – A Template for Development NGOs*. Friends in Village Development Bangladesh (FIVDB). http://aldo-benini.org/Level2/HumanitData/FIVDB_Benini_EfficientDataEntry_110314.pdf

- Buchner, D. M. *Research in Physical Medicine and Rehabilitation*. http://c.ymcdn.com/sites/www.physiatry.org/resource/resmgr/pdfs/pmr-viii.pdf

- Chapman, A. D. 2005. *Principles and Methods of Data Cleaning – Primary Species and Species-Occurrence Data*. http://www.gbif.org/orc/?doc_id=1262

- Den Broeck, J. V., Cunningham, S. A., Eeckels, R., Herbst, K. 2005. *Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities*. South Africa, Africa Centre for Health and Population Studies.

- Dr. Limaye, N. 2005. *Clinical Data Management – Data Cleaning*.

- Henning, J. 2009. *Data Cleaning*. http://blog.vovici.com/blog/bid/19211/Data-Cleaning

- Joint IDP profiling Service (JIPS). Retrieved July 2013. *Manual Data Entry Staff*. http://jet.jips.org/pages/view/toolmap

- Kassoff, M. 2003. *Data Cleaning*. http://logic.stanford.edu/classes/cs246/lectures/lecture13.pdf

- Kim et Al. 2003. *A Taxonomy of Dirty Data*. http://sci2s.ugr.es/docencia/m1/KimTaxonomy03.pdf

- Michigan State University. 2012. *Data Cleaning Guidelines (SPSS and STATA)*. 1st Edition. http://fsg.afre.msu.edu/survey/Data_Cleaning_Guidelines_SPSS_Stata_1stVer.pdf

- Munoz, J. 2005. *A Guide for Data Management of Household Surveys*. Santiago, Chile, Household Sample Surveys in Developing and Transition Countries. http://unstats.un.org/unsd/hhsurveys/

- Osborne, J. W. 2013. *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data*. California, SAGE.

- Psychwiki. Retrieved 7 September 2009. *Identifying Missing Data*. http://www.psychwiki.com/wiki/Identifying_Missing_Data

- Psychwiki. Retrieved 11 September 2009. *Dealing with Missing Data*. http://www.psychwiki.com/wiki/Dealing_with_Missing_Data

- Psychwiki. Retrieved 7 September 2009. *Missing Values*. http://www.psychwiki.com/wiki/Missing_Values

- Sana, M., Weinreb, A. A. 2008. *Insiders, Outsiders, and the Editing of Inconsistent Survey Data*. Sociological Methods & Research, Volume 36, Number 4, SAGE Publications. http://www.academia.edu/1256179/Insiders_Outsiders_and_the_Editing_of_Inconsistent_Survey_Data

- The Analysis Institute. 2013. *Effectively Dealing with Missing Data without Biasing Your Results*. http://theanalysisinstitute.com/missing-data-workshop/

- Wikipedia. Retrieved 31 July 2013. *Data Cleansing*. http://en.wikipedia.org/wiki/Data_cleansing

Annex 1 – Checklist for Data Cleaning

Prepare for data cleaning

Planning is essential. Make sure tools, material and contacts for cleaning data are available:

- The questionnaire forms
- The contacts of team leaders or enumerators, in case they need to be contacted for questions
- The original database
- A translator, if necessary
- Visual analysis software (i.e. tableau public)
- Spreadsheet (excel) or database (Access, Stata, etc.) software.
- Some would add coffee and music, and a place without noise and disturbance.

Identify the data custodian. He/she will generally be responsible for managing and storing the data, as well as for the supervision of the data cleaning, the consolidation of the changes and the update and maintenance of the change log.

Establish, document and communicate

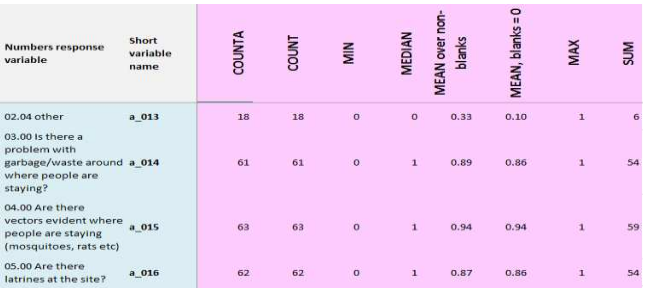
- Train the data entry operators on the how the questionnaire is populated. Explain the instructions given to enumerators. If possible, include data entry staff in the data collectors training to facilitate internal communication.
- Establish *decision rules* for when to change a value and when NOT.
- Establish procedures to document data that was modified or not collected, i.e. “missing”, or “not collected”.
- Explain how to use the change log file.
- Communicate to data entry operators or analysts the procedures to be followed and who to inform of detected errors.
- Establish communication channels for communicating detected errors. Written communication is recommended.
- For rapid assessments where data analysis, mapping and visualization generally coincides with data entry and cleaning, communicate regularly to analysts, GIS officers and graphic designers which parts of the datasets are clean and usable. Establish clear reporting procedures in case additional errors are identified. Plan with the team which variables are a priority for cleaning.

Review records

- If a sampling strategy was used, the records must be verified first. Verify if all the sites have been entered, including those where the assessment was not completed (this is not relevant in case of purposive sampling). Compare records with assessment teams field trip reports or the spreadsheet where you tracked the visited locations.
- Assign and check a unique ID for each site or household).
- Check for duplicate cases as a regular routine for each of the data rows. Remove any blank cases where the key variables have been entered but there are no data in any of the variables. Verify first that the blank cases should be removed and how this could affect other data in the row.

Screen, diagnose and treat data

- First clean filter questions, i.e. when the population is asked if they did or had a particular activity based on a response (yes/no). In that case there should be data in the following table in the questionnaire (or column in the database) if the response is “yes” or there should be no data if the response is “no”.
- Review the skip rules within the questionnaire and run the checks in the database to look for invalid or missing values in variables based on the skip rules.
- Clean questions with min or max response values (“tick three options only”, what are the top three priorities among the 5 following choice”, etc.).
- Inspect the remaining variables sequentially and as they are recorded in the data file. Create a general summary table of descriptive statistics, where for each variable the min, max, mean, median, sum and count are available.



Numbers response variable	Short variable name	COUNTA	COUNT	MIN	MEDIAN	MEAN over non-blanks	MEAN, blanks = 0	MAX	SUM
02.04 other	a_013	18	18	0	0	0.33	0.10	1	6
03.00 Is there a problem with garbage/waste around where people are staying?	a_014	61	61	0	1	0.89	0.86	1	54
04.00 Are there vectors evident where people are staying (mosquitoes, rats etc)	a_015	63	63	0	1	0.94	0.94	1	59
05.00 Are there latrines at the site?	a_016	62	62	0	1	0.87	0.86	1	54

Screenshot of summary statistics table from Aldo ***Benini, Dr.Keskin Technical note on how to approach a*** *dataset, preparation*

- If the variable is a categorical/qualitative variable, check if spelling is consistent and run a frequency count:

  - Look at the counts to see if those are reasonable for the sample – is the set of data complete?
  - All values should have labels if the variable is categorical. Check for the range of values.
- If the variable is a continuous/interval variable, run descriptive statistics such as min, max, mode, mean and median.
  - Look at minimum and maximum values. Are they reasonable? Look especially if “0” are really “0” and not missing values.
  - Is the mean and median as expected?
- Inspect data for missing values (blanks, explicit missing-value codes). Decide:
  - Which blank cells need to be filled with zeros (because they represent genuine negative observations, such as ("no", "not present", "option not taken", etc.)
  - Which to leave blank (if the convention is to use blanks for missing or not applicable)
  - Which to replace with some explicit missing value code (if we want all missing to be explicitly coded).
- Verify that in binary variables (yes/no), the positive value is coded as “1”, the negative as “0”.
- Check for the distribution of the values (use box plots if available). Look at the extremes and check them against the questionnaire even if the value is possible and may seem reasonable. If it is an extreme, other variable may be incorrect as well. Look out for the 5 smallest/largest values.
- Compare the data between two or more variables within the same case to check for logical issues. I.e., can the head of the household be less than 17 years old? Compare age with marital status. Is the person too young to have been married? Do the proportions sum up to 100%?
- Where there are questions asking about a “unit”, the data must standardized to a specific unit, i.e. when a response is collected using the unit specified by the respondent. For instance, units for area can be acre, hectare and square meters. To standardize the area unit, a lookup table can be used to merge in the conversion value to convert all areas to hectares.
- Check for consistencies within a set of cases: If there is a spouse, it is expected the spouse will be a different gender. The child of the head of household is not expected to be older than the head. The parent of the head cannot be younger than the head.
- Recode variables. Replace unhelpful entries (e.g. misspellings, verbose descriptions, category “others”, etc.) with more suitable variants, in consistent manner. Reasons for recoding are: spelling corrections, date (day, month, year) formatting, translation, language style and simplification, clustering, pre-fixes to create better sorting in tables, combination (in categorical variables), rounding (in continuous variables), and possibly others.
- Sort the file in various ways (by individual variables or groups of variables) to see if data errors that were not found previously can be identified.

Final considerations

- If the data are cleaned by more than one person, then the final step is to merge all the spreadsheets together so that there is only one database. The comments or change logs that are made as the cleaning progresses should be compiled into one document. Problem data should be discussed in the documentation file.
- Update cleaning procedures, change log and data documentation file as the cleaning progress. Provide feedbacks to enumerators, team leaders or data entry operators if the data collection and entry process is still ongoing. If the same mistakes are made by one team or enumerators, make sure to inform the culprit.
- Be prepared. Data cleaning is a continued process. Some problems cannot be identified until analysis has begun. Errors are discovered as the data is being manipulated by analysts, and several stages of cleaning are generally required as inconsistencies are discovered. In rapid assessments, it is very common that errors are detected even during the peer review process.

THANK YOU