Impact Data Cleaning Guidelines for Structured Data

1. Summary

Data cleaning processes are essential to the overall quality of our work at IMPACT for two main reasons:

1. Raw primary data is unavoidably messy. We can not expect data to be correct by default, with mistakes as the exception. Instead, raw data - especially in the sometimes difficult environments that we operate in - will almost inevitably be unusable unless data cleaning is implemented as an integral part of data collection.
2. For our research to be transparent, reproducible and in consequence trustworthy, we must ensure that changes to the raw data are traceable and justified. If the data cleaning does not follow a transparent process, it breaks the otherwise complete chain of reproducability. Thus it is important to have formalised processes to document data cleaning in place.

With the growing size of IMPACT it has become imperative to deploy organisation wide minimum standards as well as guidance for best practices across the research cycles; this applies to data cleaning as well. The Data Cleaning Guidelines are part of the general research cycle guidelines. They fit between the research design guidelines and the data analysis guidelines. They should be consulted at the beginning of, during and at the end of data collection. Their primary goal is to make it faster and easier for field teams to produce transparent, high quality data outputs by formalising processes and standards that have proven useful. To achieve that, the guidelines include the following chapters:

- **Procedures** clarifies who does what when: which responsibilities lie with the country team and which with HQ.
  - the country supervises the collection, transparently monitors and cleans incoming data and gives status updates
  - the global data unit reviews and validates the dataset before the analysis and before publication, and facilitates sharing of data cleaning tools and best practices between teams
- **Minimum standard checklists:** To advise the cleaning process and to establish transparent requirements to speed up validation. They are split into two sections, the *data cleaning checklist* and the *data formatting checklist.*
  - The minimum data cleaning checks relate mainly to technical errors such as record duplication, formal and logical consistency of the data as well as research design requirements
  - The data and cleaning log format standards adhere mostly to ODK/Kobo conventions
- **Details** on the minimum standard checks
  - Each group of common issues is layed out with an *explanation,* techniques for their *identification*, the common *issue sources* and the best practices regarding their *prevention.*
- **Issue treatment process:** Explains how to treat potential issues identified during the data checks. It includes guidance on:
  - *Timing*: The quality of data products and associated workload heavily depends on *when* data cleaning procedures are applied (the earlier the better).
  - how changes to the raw data should be *documented.*
  - *Investigating validity* of data entries flagged during cleaning checks
  - When and how to *feed learnings back* into ongoing data collection processes
  - *Correcting Errors*: when and how to keep, change or remove data
- **Anonymisation and Personal Identifiers:** Data cleaning includes a thorough check for any remaining personal identifiers in the data; This Chapter includes:
  - A *definition* of personal identifiers
  - Guidance on how they should be *processed*
- **Resources**: Links to tools and resources available for data cleaning:
  - *Templates* for standardised formats are essential to build and share tools between assessments
  - *Scripts and Tools* for R users, including an example data unit validation script

2. Procedures

The procedures associated with data cleaning and the corresponding responsibilites between field teams and HQ are outlined here: In short, the country team supervises the collection, monitors and cleans incoming data with a consistent record of all changes and gives status updates to the global team. The global team validates datasets before analysis and external sharing, provides advise throughout the process and facilitates tool sharing and collaboration between country teams.

General (applies throughout)

- **Everyone: Internal data sharing in line with ToRs and data protection guidelines**
    - **Details:** Any individual handling raw or sensitive data ensure data is shared only with people with the appropriate access rights according to the ToR as per IMPACT's data protection guidelines. Any individual handling raw or sensitive data is directly accountable to follow data protection procedures. The country coordinator is accountable to ensure that the necessary processes are in place and respected. This applies to field and HQ staff equally.
- **HQ: Support**
    - As the HQ data unit reviews all IMPACT data outputs and cleaning processes, they are usually well positioned to pull in organisation wide experiences and learnings. HQ is available to provide ad-hoc advice at any point before, during and after data collection. AO's are responsible to reach out for support when needed. The HQ data unit is responsible to be available throughout the data collection and cleaning process.
- **HQ: Sharing tools and best practices between teams**
- **Field AOs, CFPs: Communicating progress**
    - Provide regular updates to external partners and stakeholders as appropriate.
    - Inform the global team of start and end dates of data collection, including notification of delays.
- **Field AOs, CFPs: Follow Research Department communications tree**

Steps

- **Field AOs: Supervising data collection**
    - Supervising data collection includes:
        1. Briefing and de-briefing of enumerators
        2. Logging the collection progress
        3. Monitoring incoming structured data and loging potential errors, verifying these with the field and logging changes in the cleaning log.
- **Field AOs: Create a cleaned data set from the raw data and the cleaning log**
- **Field AOs / Any individual handling sensitive data: Anonymisation**
    - Any individual handling sensitive data is responsible and accountable to anonymise the data as specified in the ToR and according to IMPACT's data protection guidelines. This task is not delegable.
- **AOs: Submitting data for validation**
    - AOs send the cleaned and anonymised data with cleaning logs in standardised format to the data unit for validation before starting the analysis phase. Country Coordinators are accountable to ensure all datasets are submitted for validation.
- **HQ Data Unit: Validation**
    - Review and validate data cleaning log, raw and clean datasets before the data analysis phase.
    - Review and validate data cleaned before sharing with external parties / publication.
- **AOs / CFPs: External sharing only after validation**

- The field AOs and CFPs must await validation from HQ for the cleaned and anonymised data with cleaning logs in standardised format before publication / sharing of any data or products based on the data with external parties. (Country Coordinator is accountable)

3. Checklists

The Checklists are here to allow quick checking whether all minimum standards outlined in this document are fulfilled. For detailed explanations please consult the corresponding sections ("Details" and "The Data cleaning checklist should be consulted in the daily cleaning process. The Cleaning log and data format checklist should be consulted once before, once during and once after data collection.

3.1. Data Cleaning

Use this checklist during data cleaining (following the instructions in the "Issue Treatment Process" chapter) to ensure that the most basic data quality checks are performed. Each group has a corresponding section in the "Details" chapter.

3.1.1. Enumerator Metadata

[ ] Enumerator interview speed is reasonnable
[ ] Enumerator interview location is consistent with sampling targets
[ ] None of the enumerator with interview consistenly following the shortest questionnaire path

3.1.2. Spelling Consistency

[ ] No duplicate UUIDs. Duplicate entries are checked and removed.
[ ] Unique names for categorical variables.
[ ] Spelling of same categories is consistent, including capitalisation and blank spaces

3.1.3. Outliers and Inliers

[ ] Outliers are identified, investigated and corrected
[ ] Inliers are identified, investigated and corrected

3.1.4. Logical Coherence

[ ] Inconsistencies between questions found, investigated and corrected
[ ] Follow up questions are checked for coherence with top level questions
[ ] Within each variable, all data has the same unit in all rows

3.1.5. 'Other' Choices

[ ] "Other" responses that are similar to existing options are removed and added to choices.
[ ] "Other" responses that occur frequently are added to the questionaire an the data is transfered accordingly

3.1.6. Duplicates and Meta Information

[ ] All records have unique IDs or UUIDs
[ ] A backup of the unedited raw data is preserved
[ ] All cleaning log entries are applied in the cleaned data set

3.1.6. Sampling Strategy

[ ] The GPS locations match the sampling strategy locations
[ ] The number of records match the sampling strategy (considering the state of collection)

Links to Sampling Frame

[ ] There are variables whose values match exactly the strata names in the sampling frame (if applicable)
[ ] There is a variable whose values match exactly the cluster names in the sampling frame (if applicable)

3.2. Formatting

Data Format

[ ] Data in "tidy" format: A single row only as the data header; one row per record; one column per variable (exception: select_multiple choice columns)
[ ] Data is exported from kobo using xml values and headers (not labels)
[ ] Column headers are unique (see previous; kobo default when downloaded as xml values)
[ ] Column headers are exactly identical to names listed in the questionnaire (see previous; kobo default when downloaded as xml values)
[ ] Multiple choice questions have one column with the question name as a header, one column for each response option and one column for "other" responses (kobo default)
[ ] All columns for a multiple choice question are exactly next to each other, beginning with the one that correspond to no individual answer (kobo default)
[ ] All column headers for the multiple choice responses start with the exact question name followed by the response name, separated by a dot (".") (kobo default)
[ ] The values for multiple choice responses are exclusively "TRUE", "FALSE" or blank. (kobo default)
[ ] Column headers are exactly identical to names listed in the data analysis framework
[ ] Each column only contains a single datatype (numerical, text, TRUE/FALSE). E.g. numerical columns should not contain text in any of the fields (kobo default)
[ ] Missing data fields are left blank or replaced by NA. If for data collection other codes were introduced (ie.999 - not recommended), replace by blank or NA
[ ] No calculations are performed in the data sheet itself

3.3. Cleaning Log Format

[ ] Exactly one row for each individual data entry that was flagged during data the data checks
[ ] Headers are exactly: "uuid", "question.name", "old.value","new.value", "comment", "feedback" and "action.taken"
[ ] "question.name" values are exactly identical to the (XML) column headers in the data sheet
[ ] "new.value" column contains exactly and exclusively the new values as they appear in the data
[ ] Additional columns can be added to the cleaning log template

4. Details

4.1. Common Observable Data Errors

Here, details regarding standard data cleaning issues are layed out. Each section in the data cleaning checklist has a corresponding section here, plus an additional section on “Perfect Data”. The details relating to each group of common data cleaning issues with an *explanation* and how they can be *identified*; their most common *sources* are explained, followed by best practices in their *prevention.* Finally it is explained how they should be *treated.* The main types of common issues covered here are *enumerator metadata, spelling consistency, outliers, inliers, logical coherence, ‘Other’ choices, duplicates and meta information* as well as *perfect data.*

4.1.1. Enumerator metadata

explanation

We usually collect metadata regarding the interviewing process such as interview start and end date/time and GPS location. Further meta information can be inferred from the overall responses, such as whether enumerators always take the shorest path through the questionnaire.

Identification

Interview speed: Compare the start and end time of each interview to see if the interview was conducted too fast to be realistic. If necessary, give feedback to enumerators.
Location: Make sure that the GPS locations are coherent with selected location names in the questionnaire and in the assessment plan.
Shortest paths: Some answers lead to follow up questions and hence to slower interviews. It needs to be checked if enumerators tend to tick the fastest responses. That can be checked by counting the missing values by interview.

Common sources

These are usually indications of enumerators filling out forms without actually conducting interviews, or enumerators not following the sampling strategies. In rare cases interview speeds can vary due to security concerns of the enumerators, which should always be taken serious.

Prevention/Treatment

These issues can usually be prevented through close follow ups during the first days of data collection. They are often also reduced if enumerators understand the value and importance of their work.

4.1.2. Spelling Consistency

explanation

Choices in categorical data must be fully consistent in spelling (including blank spaces)

Identification

Crosstables / pivot tables can be used to identify varying spelling of the same categories. Calculating the Levenshtein distance between all answers can help identify similar responses.

Common sources

This is usually an issue when categorical variables are collected in free text entry fields. Inconsistent spelling can also occur when datasets are combined, when tools are changed during collection or through data cleaning efforts. Entries in the "other" category are a related issue (see the dedicated section on "Other" choices in these guidelines)

Prevention

Wherever possible, categorical variables should be set up as *select_one* or *select_multiple* questions with predefined choices. Where choices are unknown (for example location names in insufficienly mapped areas), it is still advisable to set it up as a "select" variable, and to clean and add choices on a daily basis.

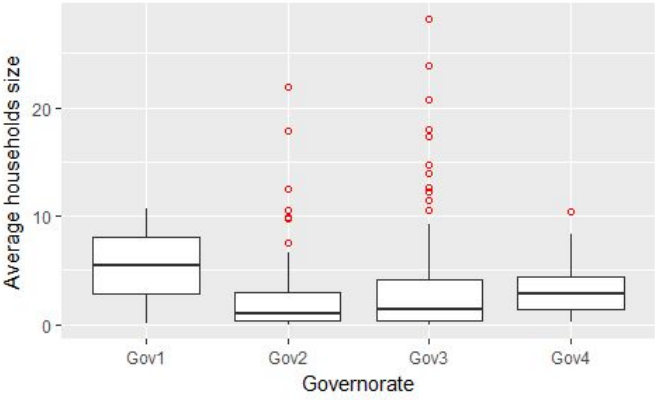
Treatment

All entries belonging to the same categories must be changed to consistent spelling (including spaces, dashes etc.). This can be partially automated / sped up using careful regex search / replace and by using the Levenshtein distance to identify similar entries.

4.1.3. Outliers

explanation

Usually, data lies within a certain range. For example, people vary in height, but only within a certain range; measuring 2.5m or 1m for the height of a person would be an outlier. Most of the time, outliers are valid data points! By no means they shoud be removed just because they are outliers, since extreme values are often interesting and an important feature of the data. However, if there is a mistake during data collection, the resulting errors often end up being outliers as well - for example if someone types one too many zeros on a number, it will likely be much higher than the rest of the data.Identification



Identification

Part of the cleaning process is to find all outliers. They can be identified with the following:

- You can define common sense maximum and minimum limits. If we would measure people's heights, we could filter all data that is smaller 1.50m and bigger than 2.20m, and have a closer look at all cases beyond.
- For data that is normally distributed, 99.7% of the data lies within 3 standard deviations of the mean. Standard deviation is a measure of how spread out the data is. Values outside of that range should be double checked. For data that has a distribution closer to a log-normal distribution (money, population sizes), the same technique can be used on the log transformation of the data. The standard deviation based method is used in boxplots to display outliers.

Common sources

The most common sources for outliers are:

- valid extreme values
- placeholders such as “999” for a certain type of missing data (this technique is not recommended)
- Zeroes for missing values
- Typing errors (e.g. wrong number of zeros)
- Unclear or missunderstood units (e.g. kg instead of g)

Prevention

- The majority of outliers can be prevented entirely through upper and lower limit constraints in the kobo tool. The constraints must lie in the “impossible” value spectrum, and should never cut off “unlikely” values.
- Outliers in the “unlikely but possible” range can be avoided through notes in the kobo tool informing the enumerator to double check the value
- Questions must be formulated with no ambiguity regarding the units
- Missunderstandings with enumerators should be cleared up during the first days of data collection

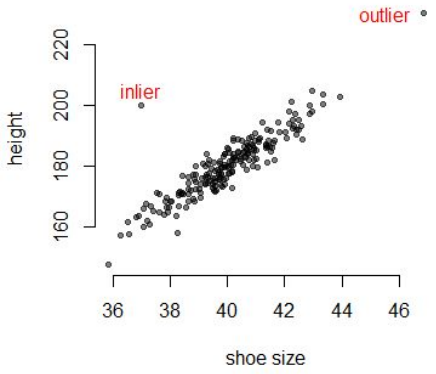
Treatment

Impossible values must be removed; which of the values is wrong can sometimes be derived from common sense or comparison with other variables, but usually requires follow ups with enumerators.
Unlikely but possible values that can not be confirmed nor falsified must be kept in the dataset.

4.1.4. Inliers

Explanation

Similar to outliers, there are “Inliers” between correlated variables. The values in a record can be within the expected range for two variables, but their combination may still be unlikely or impossible to occur. For example, both a shoe size of 36 and a body height of 2m might be within the expected range, but the combination of the two could hint at an error in the data.



Identification

These are more difficult to detect than outliers. They need to be investigated individually. We first make a list of interview questions whose answers should make sense together (for example a person’s height and their shoe size). Then, depending on the data type, we can:

- **between two numerical variables**: Make a scatterplot to compare numerical with other numerical data. Data points that are far away from the other data points are ‘inliers’ and should be checked.
- **between a numerical and a categorical variable:** Split the data by the category and follow identification of “outliers” for each group.
- **between two categorical variables:** Make a crosstable to compare categorical data with categorical data. Combinations with very small counts are inliers.

Prevention

Impossible combinations can usually be prevented entirely through carefully designed constrains and skip logic in the kobo questionnaire.
Unlikely but possible combinations must not be prevented in the kobo tool (this would skew the data), and need to be investigated on a case by case basis.

Treatment

Impossible combinations must be removed; which of the values is wrong can sometimes be derived from common sense or comparison with other variables, but usually requires follow ups with enumerators.
If it can not be identified which of the values causes the problem, both values must be removed.
Unlikely but possible values that can not be confirmed nor falsified must be kept in the dataset.

4.1.5. Logical Coherence

Explanation

Some variables have logical dependencies; if they are not fulfilled, there is most likely an error in the data. For example: Does the household age disaggregation add up to the total household size? Is there an expenditure reported on school fee when there is no school age children in the households? Further, answers to top level questions and follow up questions must be coherent. If the interviewee says 'There was damage to the shelter', and selects no type of shelter damage - or the other way around - that would be a mistake.

Identification

See Identification of "Inliers".

Common sources

Data entry errors; question missunderstandings; Lack of skip logic in the kobo tool, leading to questions being asked that do not apply (e.g. questions about children in households that have no children).

Prevention / Treatment

See Prevention and Treatment for "Inliers"

4.1.6. 'Other' choices

explanation

Since we can not anticipate all responses that we can expect from a select_one or select_multiple choice variable, we include an option to specify "other" in almost all cases. Other choices are lost data if not handled correctly, but are essential to not miss unexpected but important results (For example lack of machine replacement parts as a main reason bakeries can not operate).

Identification

Cross tables over the specified 'others' can help identify frequent answers.

Common sources

- Important answers could not be anticipated at the research design stage
- Enumerators missunderstood how the form works (adding responses in the 'other' field although they are available as options)
- Enumerators try to record more granularity than the choices provide, the answer however falls into the available categories.

Prevention

Careful choice of options provided in the questionaire. Ideally these should be based on qualitative research and consultation with cluster experts prior to the assessment.

- Adding frequent choices during the pilot or the first days of data collection
- It is *not* an option to not provide an 'other' choice. It is also not advisable to use free text fields where select_one/select_multiple are possible (see 'categorical value consistency'')

Treatment

They need to be checked early and dealt with depending on the following cases.

'Other' choices that do not make sense: Other entries that appear to be results of the respondent or enumerator misunderstanding of the question should be removed. However, if the response is reported frequently, it needs to be investigated.

'Other' choices that are similar to existing categories: Often interviewers/interviewees give responses in 'Other' that are identical or very similar to existing choices. These should be removed from the 'Other' column. If it is a multiple choice question, the corresponding field in the response column should be set to 'TRUE', and the choice should be added to the column that has all selected answers as concatenated text.. If it is a single choice question, the corresponding field in the question column should be set to the name of the corresponding response.

Reoccuring choices for 'Other': If many people give the same 'Other' response for a question, this is usually because the questionnaire is missing a relevant answer in the actual context. If the answer options are not read to the interviewee by enumerators, the response should be added as an option to the questionnaire. All corresponding answers from before should be added accordingly.

4.1.7. Duplicates

explanation

Technical issues during data transmission between phones / servers as well as manual mistakes when digitising paper forms can lead to individual records appearing more than once in a raw data set.

Identification

Technical duplicates through issues with kobo server/upload structure can be identified through the “uuids” - a universally unique id that Kobo assigns to any interview.
Potential duplicate interviews with the same household can further be identified through unique identifiers in the data, and through large overlaps between the records’ values.

Common sources

Duplicate entries can arise because of technical issues, or data merging mistakes, but also by chance or due to overlaps in grid based / systematic sampling techniques.

Treatment

Duplicate records can be simply removed.

4.1.2. Sampling Strategy

explanation

The sampling strategy as defined in the ToRs needs to be understood and respected in the data collection. Without double checking with field managers, partners and enumerators, most likely the intended sampling strategy will not be respected during data collection. We also quite often find that some of the necessary assumptions made during research design were off, and so it might be impossible for enumerators to collect the data as intended. This can pose major and sometimes irreperable difficulties at the analysis stage. Noticing deviations early on allows us to either to get back to the intended sample, or at least make a deliberate choice on a revised strategy that matches better the situation encountered on the ground.

Identification

Mapping enumerator paths alongside intended sampling boundaries/centroids, as well as monitoring sample sizes per strata/cluster and location can help notice when the sampling strategy is not implemented as intended.

Common sources

This can occur through missunderstandings among field coordinators or enumerators; Sometimes the implications of changing the sampling strategies for operational purposes are underestimated / unclear. Often available population numbers are inaccurate or outdated, so without interference from the country team, enumerators are forced to improvise in order to reach their sample targets.

Treatment

Ususally this affects the dataset as a whole, so the usual process relating to individual values/records/variables does not apply. If a diversion from the sampling strategy is detected, the solution is highly dependent on the context. It is advised to get in touch with the data unit and the research design unit for support.
Generally, we would try the following:

- Assess if and how the original sampling probabilities can be restored in the subsequent sampling
- If this is not possible, it needs to be clarified whether the actual sampling probabilities can be estimated, and if/how the population of interest has changed

- In some cases, the data collected outside the intended sampling strategy needs to be discarded. Ideally we will try to post-stratify during the analysis stage to adjust for the disproportionate sampling probabilites, and transparently caveat / map any changes in the population of interest.

4.1.2. Links to Sampling Frame

explanation

In stratified and/or cluster samples, it is necessary to be able to link each record back to it's corresponding stratum/cluster in the sampling frame. If that information can not be obtained, analysis / representative aggregation may be impossible.

Identification

Can you automatically match every record with a corresponding row in the sampling frame?

Common sources

Cluster/Stratum names collected not at all, as free text or as choices that do not match the references in the sampling frame. Definitions used for population numbers (e.g. for IDPs / Returnees) do not match definitions used by enumerators

Prevention

Stratum/Cluster names should be recorded as a select_one variable, with the choices matching exactly the reference names in the sampling frame.

Treatment

Strata/Cluster names can sometimes be reconstructed from text entries or GPS locations.

4.1.8. Perfect Data

explanation

Perfect data should make you highly suspicious. For example two key informants will almost never give completely identical interviews; Even the same key informant is very unlikely to give the exact same responses twice. Furher, numerical distributions *should* produce a certain percentage of (valid) outliers.

Identification

When very few of the other issues are identified.
Through counting the number of disagreements between triangulated key informants.

Common sources

Key informant networks / enumerators / field coordinators may take note of the data review and feedback patterns, and try their best to deliver pre-triangulated information, avoid recording extreme values etc.

Prevention

Clarify with field teams that the review/feedback requests are a common procedure, and that it is normal (and necessary) to check back on many values that turn out to be fine. Explain that leaving out valid but extreme values can lead to biases in the data.

Depending on the context/situation, additional questions regarding the connection between and sources of different key informants can be added, to see if shared sources could be an explanation for the unexpected consistency, or if triangulation happens before we receive the data.

5. Issue Treatment Process

This chapter explains how to treat potential issues identified during the data cleaning checks. It includes guidance on how changes to *documented changes* to the raw data*,* how to *investigate validity* of data entries flagged during cleaning checks and advice on when to keep, change or remove data to *correct errors*.

5.1. Data Cleaning Timing

The quality of data products heavily depends on when data cleaning procedures are applied. Most issues can not be fixed after collection is over, leading to data loss and low quality information. If problems can be fixed early on in the collection process, a lot of work can be avoided down the line, and quality can be improved significantly.

Therefore, data cleaning ideally begins on Day 1 of data collection. For structured data, incoming data should be monitored for errors to ensure rapid field verification if needed. For semi-structured data, incoming data must also be monitored closely to identify where saturation is reached and a question can be dropped or data collection can be halted altogether, or where the question route needs to be amended to explore a topic in further detail.

The only way to reach high quality is to clean and check the data from the Day 1. Problems with questionnaires, misunderstanding of enumerators and data entry mistake can then be catch and adjusted early on, when the data is still being collected and the field team still available for questions.

This greatly decreases workload and time pressure towards the end, when other tasks (e.g. preliminary finding presentation, report writing) require attention.

We recommend:

- **Start of data collection / pilot and the first few days:** Daily data cleaning checks, logging of potential issues, and investigation of suspicious entries. Frequent issues are addressed with enumerators and careful adjustments to the questionnaire and kobo tool are made.
- **During data collection:** Data cleaning at least every other day. Frequent issues are identified and addressed with enumerators. Failling indicators are removed from the questionnaire.
- **End of data collection** (last few days): All remaining data screened for issues, logged and investigated.
- **After data collection:** Final review of potential issues across the dataset is conducted; misstrusted records and variables are removed. The cleaned dataset is created from the raw data and the cleaning log, including a readme sheet. The data is send to the Geneva Data Unit for validation.

5.2. Documenting Changes

In order to preserve reproducability and transparency, potential issues with data entries raised during data cleaning checks are never addressed directly in the raw data. Instead, they are listed in a cleaning log (see “Templates”), and their validity is investigated. The results of the investigation, the proposed action and new value are recorded in the cleaning log book. Then, decisions about removing entire records or variables are made and recorded and justified in the cleaning log book. Finally, a cleaned version of the dataset is created from the raw data and the cleaning log book, preserving the original raw dataset.

5.3. Investigating observation validity

For each outlier, inlier, and apparent incoherence in the data (see chapters C - Checklists and D - Details) we must decide which of the following categories the issue falls into:

- it is ann unusual but **valid observation**
- it is a genuine **error**
- will we have to accept that it will remain **unclear** whether it is an error or a valid observation

Common mistakes include: - typing errors - enumerator misunderstood the question - wrong units - errors introduced during data handling. Common reasons for valid data include: Extreme numeric values; uncommon combinations due to unexpected circumstances of the interviewees.
The main strategies to find out if it is a valid observation or an error, we can..

- use common sense regarding impossible values (a 2.3m tall person might be possible, a 20m tall person not.)
- check back with enumerators
- cross check with other parts of the interview that should be coherent (for example if a household has unusually high water usage, but also an unusual large number of members, the two values affirm each other)

5.4. Feeding learnings back into data collection

During data cleaning, we will see reoccuring issues especially during the pilot and the first days of data collection. We can prevent them from happening during the rest of the collection; examples include:

- **Enumerators missunderstand questions:** clarify with enumerators
- **Variable collected with varying units:** clarify with enumerators; add note/details to kobo tool
- **Inliers and Outliers:** Add a note to the kobo tool to ask the enumerator to double check on the spot on extreme values. Add constrains to the kobo tool to prevent absolutely impossible values (not recommended for variables related to money and population counts)
- **Inconsistencies**: Add constraints to the kobo tool
- **Frequent issues in certain variables:** get feedback from enumerators, double check translation and potentially ammend tool
- **Frequent issues in certain locations:** get feedback from enumerators on potential difficulties
- **Frequent issues from certain enumerators:** check in with enumerator
- **Issues with enumerator locations / interview times / shortest paths through interview**: clarify with / give feedback to enumerators to reaffirm their diligence

5.5. Correcting Errors

Once you made a judgement if a suspicious value is actually a **valid observation**, an **error**, or if the situation remains **unclear**, try to find find out the correct value. It is important not to remove valid extreme values. If we do that systematically, the whole data can become biased. Thus we must not automatically remove them, but make a case by case decision. If you can not find out if it is an error or not, the value should be kept:

- **valid observation** → keep: copy the old value to the new value column in the cleaning log
- **error:**
    - correct value found → *update*: add the new value to the cleaning log
    - correct value not found → *remove*: mark for removal in the cleaning log

- **unclear:**
    - observed value is unlikely but possible → *keep*: copy the old value to the new value column in the cleaning log
    - observed value is impossible → *remove*: mark for removal in the cleaning log

Once the individual values are dealt with, we need to look into systematic problems across variables/records. Depending on how errors are distributed, it can lead us to distrust a whole record, or even a whole variable:

	individual records	**Many records**
individual variables	keep/update/remove the individual values	Question may have been misunderstood: delete variable OR recollect the data for this variable
Many variables	Delete the whole record	Early in collection: Investigate and get feedback from enumerators; request support; After collection: It's too late!

If there are a lot of issues across the whole dataset, do not hesitate to request support from HQ / management. If this is noticed early in the assessment, ususally the sources can be identified and fixed.

6. Anonymisation and personal identifiers

In accordance with the data protection guidelines on personal identifies, anyone involved in data cleaning is accountable to onlz share data in adherence with the access rights defined in the ToR. Therefore anonymisation and removal of personal identifiers is a key step of the data cleaning process.

6.1. What is Personally Identifiable Information?

Personal identifiers are all variables that make it possible to single out an individual or a household - this also holds if identification is only possible in combination with other variables.
Personal Identifiers can relate to interviewees, key informants as well as enumerators. Direct identifiers often Include (but are not limited to) GPS coordinates, Names, Addresses and Phone Numbers. Indirect Identifiers - that allow singeling out individuals or households in combination with other variables - often include exact age, profession and location.

6.2. Process

The relevant variables must be identified at the research design stage, and the ToR should lay out which variables need to be deleted during data cleaning. They must be removed entirely from all files/sheets. All instances of values from variables to be deleted must be removed from all files/sheets, *including the cleaned data, the cleaning log, the kobo "choices" sheet, and even the*

raw data. This is the only exception where it is acceptable and imperative to ammend the orignal raw data. The only trace that these variable were collected in the first place should be references to the variable names (not the values) in the kobo tool questions and potentially in cleaning log entries.

For further details please consult the data protection guidelines regarding personally identifiable information.

7. Resources

7.1. Templates

In order to effectively share and reuse tools and processes between teams it is important to use standardised templates for data, cleaning logs and other processes. They can be downloaded from the online Toolbox.

7.2. Scripts and Tools

(Will be added incrementally)