Data Cleaning Protocol

Table of Contents

Introduction

This data cleaning checklist entails all the data cleaning procedures conducted to ensure that the data collected, entered and submitted for analysis is correct, clean and timely. It comprises all data processing procedures for the three major projects namely: Dispensers for Safe Water (DSW).

Objectives

The process is focused at ensure that data that is finally analysed is:

1. Complete
2. Error free.
3. Consistent.

It is important therefore, to understand the tool/survey/form that was used for data collection before starting the actual data cleaning; in order to understand the data, check for skip patterns and check for any errors that might arise from programming e.g. a question appearing on SurveyCTO but not required resulting in it being skipped during data entry.

It is also important to check with the analysis team on what variables they would be focusing on during analysis so as to pay closer attention to such variables during cleaning.

1. Dispensers for Safe Water (DSW)

Dispensers for Safe Water collects two types of data. Operations and Monitoring. These will be discussed in detail later on in the document.

For all Forms and Surveys:

- Check how many form versions were used for a particular activity so that you can download complete data.
- Save the survey and raw data from the server in their respective program folders in Box.
- Go through the form/survey used for data collection to familiarize oneself, check for any omissions or errors e.g. a question not being required, errors in skip logic etc.
- Check that the form/survey has all the necessary constraints e.g. verification id constrained to 10 digits, compound id to 12 etc. Make necessary changes as required
- Check the surveyor names and id are entered correctly and make necessary changes as required
-
- Check that the data is complete i.e. makes sense with proper spelling e.g. no meaningless letters entered instead of meaningful responses e.g. enter G instead of a waterpoint name, or Chg instead of county name etc and make necessary changes as required
-
- Drop any observations in a data set that have entries recorded as test e.g. the surveyor name entered as *test* instead of an actual name.
- Replace any blanks in the main variable column with what has been entered in the "other" column for that variable e.g. if a *village name column* is blank and the name entered in "*other_village column*", replace the blank in the *village name column* with the response in the "other_village*column*"
- Establish a clean set of geographical identification variables for each program. Check that all the geographical locations are correct and also that they are spelled correctly and consistently e.g. Kisumu not Kisumo. A merge with the Master Waterpoints list/pass list will be useful in cleaning geographical identification variables. Feel free to reach out to the Area Coordinators where you are not sure about the geographical locations.
- Format the date to date/month/year format e.g. 26Dec14. Also check that the date is correct e.g. not having a 2015 date when the year is 2014.
- Format text responses to be either proper or upper and trim unnecessary spaces or punctuation. Avoid Mass replacement of a string character in all string variables e.g replacing "." With "" in all string variables.

- After appending a particular data set, double check that no variables have been displaced i.e. no variable has been renamed with a V followed by a number e.g. V25. When you notice this, rerun your do as you careful check where the error arose from. This will mostly happen when we have compounded quotes (") in .csv and it happens just after importing the csv. To resolve this, go back to the CSV file downloaded from SurveyCTO CTRL+F to get where the " is and replace the " with nothing and save the csv again. Also check if the " has caused data to shift in the raw .csv file. Remember to format any variable that might get deformed when re-saving. For example dates and variables with many digits like verification ID.

- Clean individual or appended data should not have blanks where the raw data had values. Browse your data after completing the cleaning process to ensure no observations were dropped.
- Do a count of observations after dropping true duplicates i.e. if they exist in a particular dataset. After importing data, check the number of observation in the data. Note that if the csv has compounded quotes in row 25 of data with 500 rows stata will only import 25 rows.
- NEVER use the keyword FORCE when appending data. Check for duplicates by instanceID after appending to make sure data has not been appended twice
- Check for outliers in variables (e.g yes/know question should not have other figures like 15, 18 or having string character in numeric variables etc

a. Operations Data

This includes data collected on **Forms** for the following activities:

I. **LSM (Local Stakeholders Meeting) –** To secure buy in from the local stake-holders.

II. **Waterpoint Verification –**Involves visiting all the waterpoints in a particular geographical area to assess if they qualify for a Chlorine Dispenser.

III. **VCS (Village Community Sensitization) –** To sensitive the community to the Dispensers for Safe Water program and the Chlorine Dispenser System.

IV. **Installation –** To collect information on all Chlorine Dispensers installed in the respective community.

V. **CEM (Community Education Meetings) –** To educate the waterpoint users where a Chlorine Dispenser has been installed on safe water and the Chlorine Dispenser System.

VI. **Chlorine Delivery form –** To keep track of delivered chlorine vs chlorine used.

This is information collected on all dispensers. Data should be downloaded and cleaned on a monthly basis. Reference should also be made to the respective tracker used for the particular activity so as to double check on the completeness of downloaded data and make the necessary follow up where there are discrepancies.

Waterpoint Verification data:

- Check for consistency in the data by first understanding the survey used to collect the data.
- Check that each waterpoint has a unique verification id that matches the village id. The verification id should be unique and 10 digits long, the first 8 comprising of the village id and the last 2 are unique to each waterpoint in a village.
- Check that each village ID is unique to each village and no two villages share the same village ID.
- Merge with the verification tracker when cleaning, to clean the waterpoint names and also do a count of all waterpoint so as to know whether the downloaded data has all the observations.
- Double check the total with Associate – Data Collection and Quality. Transmit a list of waterpoints missing in the SurveyCTO data and send to him/her to make a follow up with the respective field team to submit the missing data to the server.
- Check for duplicated waterpoints by duplicates tagging the Sublocation/parish, village and verification id.
- Repeat the process by duplicates tagging the Sublocation/parish, village and waterpoint name. Sort by waterpoint name because of slight errors in spelling that might cause the same waterpoint to appear as 2 different waterpoints.
- Drop true duplicates (confirm with supervisor if you are not certain).
- Kindly use the do file template to act as a guide during the cleaning.

Village &Community Sensitization data:

- Check that the questions on the VCS attendance roaster match the ones on SurveyCTO.
- Check for consistency in the data by first understanding the survey used to collect the data.
- Check that each village has a village id (it should be unique and 8 digits long).
- Merge with the VCS tracker when cleaning to clean geographies e.g. villages and also do a count of all villages so as to know whether the downloaded data has all the observations.
- Double check the total with Associate Data Collection and Quality. Transmit a list of villages missing in the SurveyCTO data and send to him/her to make a follow up with the respective field team to submit the missing data to the server.
- Check for duplicates by duplicates tagging the meeting date, sub-county/district, sublocation/parish and village id.
- Repeat the process, this time by duplicates tagging the meeting date, sub-county/district, sublocation/parish and village name. Note that the villages should be unique, only one VCS meeting is conducted per village. Sort by village name because of slight errors in spelling that might cause the same village to appear as 2 different villages.
- Drop true duplicates (confirm with supervisor if you are not certain).
- Kindly use the do file template to act as a guide during the cleaning.
- Confirm that sum of subtotal sums up to the total attendance (e.g. vcs202a_attendees_female+ vcs202b_attendees_male=vcs201_attendees_total). Also Check for Village population vcs total attendance. In most case attendance will be lower than village population but in cases where Attendance is higher than village pop this difference should not be too big

Use the checklist below to ensure that you have all the variables required for submission of datasets.

Installation data:

- Check for consistency in the data by first understanding the survey used to collect the data.
- Check that each waterpoint has a correct waterpoint id the matches the pass list (it should be unique and 8 digits long).
- Merge with the installation tracker when cleaning to check for spellings and also do a count of all waterpoint so as to know whether the downloaded data has all the observations. Also compare the Installation dates in SurveyCTO data with those in the tracker to check for major differences in installation datesDouble check the total with Associate Data Collection and Quality. Transmit a list of waterpoints missing in the SurveyCTO data and send to him/her to make a follow up with the respective field team to submit the missing data to the server.
- Each waterpoint/dispenser should have a barcode and the barcode should be unique.
- Check for duplicate waterpoint ids and barcodes by duplicates tagging the sublocation/parish, village and waterpoint id.
- Repeat the exercise this time duplicates tagging the sublocation/parish, village and barcode.Note that each waterpoint should have a unique waterpoint id and barcode. Only one dispenser is installed at a waterpoint at a time.
- Drop true duplicates (confirm with supervisor if you are not certain). Use i102b_disp_installed to check/affirm true duplicates
- Check installation dates are within month and year when installation were done for that particular office/region
- Kindly use the do file template to act as a guide during the cleaning.

Community Education Meetings data:

- Check that the dates are within month and year when CEM were done for that particular office/region
- Check for consistency in the data by first understanding the survey used to collect the data.
- Check that the questions on the CEM attendance roaster match the ones on SurveyCTO.
- Check that each waterpoint has a waterpoint id that matches the passlist (it should be unique and 8 digits long).
- Merge with the CEM tracker when cleaning to check for spellings and also do a count of all waterpoints so as to know whether the downloaded data has all the observations. Double check the total with Associate Data Collection and Quality.Transmit a list of waterpoints missing in the SurveyCTO data and send to him/her to make a follow up with the respective field team to submit the missing data to the server.
- Check for duplicate waterpoint ids by duplicates tagging the sublocation/parish, village and waterpoint id. Note that each waterpoint should have a unique waterpoint id. Ideally, only one CEM is conducted at a particular waterpoint.
- Drop true duplicates (confirm with supervisor if you are not certain).
- Confirm that sum of subtotal sums up to the total attendance (e.g. attendees_female+ attendees_male=attendees_total)
- Kindly use the do file template to act as a guide during the cleaning.

Chlorine Delivery Data

- Check for consistency in the data by first understanding the survey used to collect the data.
- Deliveries are done on an estimate, after 10 weeks so a waterpoint may have more than one entry.
- Merge with the chlorine delivery tracker when cleaning to check for spellings and also do a count of all waterpoints so as to know whether the downloaded data has all the observations. Double check the total with Associate Data Collection and Quality. Transmit a list of waterpoints missing in the SurveyCTO data and send to him/her to make a follow up with a field team to submit the missing data to the server.
- Check for duplicate waterpoint ids by duplicates tagging the date of delivery and waterpoint id. Note that each waterpoint should have a unique waterpoint id. Ideally, only one delivery should be done to a particular waterpoint in a day.
- Once you get duplicates, double check by this time duplicates tagging the sublocation/parish, village and waterpoint name, waterpoint id and date of delivery. Drop true duplicates (confirm with supervisor if you are not certain).
- During cleaning, check for outliers in number of Jerricans delivered and jerricans from last stork. E.g. we should not have huge figures, 12 , 45 etc. Also check that these variables are not blank or have negatives and ensure that no deliveries are done before date of CEM
- Kindly use the do file template to act as a guide during the cleaning.

b. Monitoring Data

This includes data collected on **Surveys** for the following activities:

I. **Community Survey –** Done to assess chorine adoption in communities with the chlorine dispenser.
II. **Promoter Survey –** To assess promoter performance.
III. **Spotcheck –** To check on the condition and functionality of the dispenser.

This is information collected from randomly selected waterpoints on a monthly basis. Reference should also be made to the respective tracker used for the particular activity so as to double check on the completeness of downloaded data and make the necessary follow up where there are discrepancies. Note that data from the Community, promoter and Spotcheck surveys MUST be a perfect merge. If information is missing on either of them, clarification should be sought from the Associate – Data Collection and Quality. Ideally there should be 1 promoter survey, 1 Spotcheck survey and at least 8 Community surveys at a particular waterpoint within a month.

Community Data

- Check for consistency in the data by first understanding the survey used to collect the data.
- An average of 8 households are visited at a particular waterpoint in a month. An average of 10 waterpoints are visited in a particular month but may vary depending on location and time.
- Merge with the community tracker when cleaning to check for spellings and also do a count of all observations so as to know whether the downloaded data has all the observations. Double check the total with Associate Data Collection and Quality. Transmit a list of waterpoints missing in the SurveyCTO data and send to him/her to make a follow up with the respective field team to submit the missing data to the server.
- Check for duplicate compound ids by duplicates tagging the compound ids and waterpoint id. Note that each household should have a unique compound id and is 12 digits long. The first 8 are the waterpoint id and the last 4 come from the randomization list.
- Send the list of compound ids that are wrong or duplicated to the Associate Data Collection and Quality so that they can follow up with the respective field team to send the correct values.
- Once you get duplicates, double check by this time duplicates tagging the sublocation/parish, village and respondent name, waterpoint id and compound id.
- Check that if a chlorine test was done, the TCR and FCR values are present after cleaning i.e. no values are missing/have been dropped.
- Drop true duplicates (confirm with supervisor if you are not certain).
- Check if data for all waterpoint that was to be visited in a particular month is present in the server. (Also applies for Promoter data). Also check number of households visited per waterpoint and confirm with DCT in cases where we have very few hh per waterpoint visited. e.g. 1, 2 or 3 hh per waterpoint
- Kindly use the do file template to act as a guide during the cleaning.

Promoter Data

- Check for consistency in the data by first understanding the survey used to collect the data.
- The promoter at a particular randomly selected water point should be visited once a month when the FO is conducting the continuous evaluations. On average, if 10 water points are visited in a month, expect 10 promoter surveys.
- Merge with the promoter tracker when cleaning to check for spellings and also do a count of all observations so as to know whether the downloaded data has all the observations. Double check the total with Associate Data Collection and Quality. Transmit a list of water points missing in the SurveyCTO data and send to him/her to make a follow up with the respective field team to submit the missing data to the server.
- Check for duplicate water point ids by duplicates tagging the date of the survey and water point id. Note that each promoter should have a unique water point id. Once you get duplicates, double check by this time duplicates tagging the sublocation/parish, village and respondent name, water point id and date of survey.
- Drop true duplicates (confirm with supervisor if you are not certain).
- Kindly use the do file template to act as a guide during the cleaning.

Spotcheck Data

- Check for consistency in the data by first understanding the survey used to collect the data.
- The spotcheck at a particular randomly selected waterpoint should be visited once a month when the FO is conducting the continuous evaluations. On average, if 10 waterpoints are visited in a month, expect 10 spotcheck surveys.
- Merge with the spotcheck tracker when cleaning to check for spellings and also do a count of all observations so as to know whether the downloaded data has all the observations. Double check the total with Associate Data Collection and Quality. Transmit a list of waterpoints missing in the SurveyCTO data and send to him/her to make a follow up with the respective field team to submit the missing data to the server.
- Check for duplicate waterpoint ids by duplicates tagging the sublocation/parish, village and respondent name, waterpoint id and date of survey. Note that each spotcheck should have a unique waterpoint id.
- Once you get duplicates, double check by this time duplicates tagging the sublocation/parish, village and respondent name, Barcode id and date of survey.
- Drop true duplicates (confirm with supervisor if you are not certain).
- Kindly use the do file template to act as a guide during the cleaning.

Appending data

- Ensure that each data set is clean and complete i.e. no values have been dropped.
- Use the loop in the do file to append the clean data. NOTE: (It is important to include the origin file name to the append loop so that in case you notice a problem after appending many files you can easily check which data set had the problem. This is however dropped after confirming all is well and saving the final appended data)
- Browse after appending to check that no values have been dropped e.g. dates, TCR reading etc.
- Double check your data for duplicates i.e. in case you append the same data twice. Isid instance id and any unique identifier.
- Do a final count to make sure it’s a sum of the individual counts per program.