RAM STORIES FROM THE FIELD

SOMALIA CO

SOMALIA CO DATA CLEANING PROCESS

Starting with raw data downloaded from MoDa, our data cleaning process begins by employing the **cleanR** package. This **toolkit** allows us to load the raw data and execute a series of essential data cleaning checks. These checks include survey completion time, identification of missing data points, handling of non-standard responses, detection of outliers, pinpointing inconsistencies and validation of GPS coordinates.

The outcome of this process is the generation of a comprehensive cleaning log file. The log file consists of 6 columns. It starts with a column named 'uuid', serving as the unique survey identifier, followed by a column named 'question name', which denotes the variable name. The third column in the log file is the 'issue' column, which provides details on the specific issue flagged. This is followed by the 'action' column, which provides guidance on the action to be taken to address the issue identified, the 'feedback' column records responses from the field teams on issues that have been flagged, the 'old.value' column retains the current value of the flagged variable. Finally, and when relevant and applicable, the `new.value` column documents the

updated accurate values, which replace the old values, ultimately resulting in the generation of clean data.

It is important to note that once an action has been identified, the CO Planning, Monitoring, Learning, and Evaluation (PMLE) team reviews the survey record that has been flagged and engages with the field team/enumerator(s), before deciding to either delete the record or update the value that has been flagged. It is also important to note that there are other activities involved in the overall data cleaning process employed by the Country Office and the sub-process mentioned here is but a part of a longer and more comprehensive data cleaning process.

Table 1 below shows an example of a log file. This log file serves a dual purpose. Firstly, it enables seamless communication with our field team by facilitating prompt corrective actions based on the issues discovered during data checks. Secondly, it aids in the documentation and standardization of the data cleaning process, ensuring that every alteration made to the raw data is both justified and documented to uphold transparency and accountability.

PMLE unit has adopted this approach successfully and applied it to multiple assessments, witnessing a remarkable improvement in the quality of data. This rigorous approach guarantees the reliability and integrity of our data which is a critical component of our ongoing efforts on improving data quality and utilization.

EXAMPLE OF CLEANING LOG FILE GENERATED FROM RAW DATASETS

uuid	question.name	issue	action	feedback	old.value	new.value
027a6261-832c-4d8f-85ef-7564d9	HHAsstWFPCBTRecYN	Beneficairy didn't receive assistance in the last two months, please delete the survey	delete		No	
43baf030-68ec-49b2-a81f-583d57	HHAsstWFPCBTRecYN	Beneficairy didn't receive assistance in the last two months, please delete the survey	delete		No	
62b8b5e3-cbe6-4768-837b-8d559	ExpFoodRM	Household monthly expenditure is below 20 USD, please confirm if this is correct	check		14	140
a106829b-d749-44e0-b48e-8a455	InfantNb	more than three children aged 0 to 23 months are reported, please confirm	check		5	2
7eacf02b-3272-4cb3-bb8b-4a2af8	interview_duration	The survey is short and took less than 35 minutes to complete	delete		12.95 minutes	