

DATA QUALITY ASSESSMENT REPORT

Kenya Household Survey: A Socioeconomic & Wellbeing Study (2022-2026)

Prepared By: Sharon Nyabuto - Research Data Analyst

Report Date

June 2026

Organisation: Community Insights Research Centre

Assessment Status

**PRE-CLEANING REQUIRED BEFORE
ANALYSIS**

1. Overview

This report documents findings of an assessment of the data quality of a raw Household Survey dataset prior to cleaning. The data was collected between 2022 and 2026, across 12 Kenyan counties, and the dataset contains 9,500 household records in 22 columns. This assessment is aimed at identifying quality issues in the raw data and define a cleaning plan to produce an analysis-ready dataset.

The assessment covers six dimensions:

- Uniqueness (duplicate records)
- Completeness (missing values)
- Consistency (categorical coding)
- Validity (numeric outliers and impossible values)
- Conformity (data types and formats)
- Integrity (cross-field logic).

3. Key Issues

3.1 Missing Values

Three columns exceed the 20% threshold at which missing data can meaningfully skew results:

- *education_level: 2,165 (22.8%) missing records*
This is the highest missing rate, likely a result of respondent refusal, exercising their freedom to willingly withhold information they are uncomfortable sharing.
- *dependency_ratio_category: 2,110 (22.2%) missing records*
Missing values are distributed across all household size values, confirming that household size is not the cause of missingness, and that missing values are not random. This could suggest a data capture gap in the dependant field during collection.
- *health_insurance: 2,012 (21.2%) missing records*
Given the sensitive nature of insurance-related questions, non-response may reflect respondent reluctance to disclose information or inconsistencies in questionnaire administration.

6 more columns fall between 10% and 20% missing values and must be explicitly handled: sanitation_type (19.5%), dwelling_type (18.1%), receives_remittances (17.9%), mobile_money_access (17.8%), land_ownership (17.1%), and food_security_status (16.6%).

3.2 Duplicate Records

80 rows are involved in duplication, distributed across three types requiring different resolutions:

- **50 exact duplicate pairs (Same household_id and respondent_id)**
This indicates a likely case of double submissions. Assign row numbers using Row_number() and drop the latest entry, retain one record per pair.
- **20 rows (Same household_id, different respondent_id)**
Indicates a possible re-interview in the household or a data entry error. Flag for review before dropping.
- **10 rows (Same respondent_id, different household_id)**
A respondent is expected to belong to only one household. Therefore, occurrences where the same respondent_id is linked to multiple household_ids indicate a possible data quality issue.

3.3 Inconsistent Categorical Coding

14 columns contain the same categories recorded in multiple ways; mixing capitalisation, abbreviations, and full names. None of these columns can be used for grouping or counting without standardisation first.

Column	Issue Type	Example
Gender	6 variants for 2 categories	(Male, male, M, Female, female, F)
Education level	9 variants for 4 categories	(PRIM, Primary, primary school, Uni, university, sec, Secondary, college, College/TVET)
water source	10 variants for 5 categories	Piped/piped water, River water/river, Rainwater/rain water, Borehole/bore hole, Water vendor/vendor

sanitation type	7 variants for 4 categories	Flush/flush toilet, Pit/pit latrine, Open/open defecation, VIP latrine
residence type	8 variants for 3 categories	(Urban/urban/URBAN, Rural/rural/RURAL, Peri-urban/peri urban)
employment status	17 variants for ~6 categories	e.g. Employed (Formal)/EMPLOYED/Formally employed; farmer/Smallholder Farmer/subsistence farmer
livelihood source	25 variants for ~14 categories	e.g. Business/BUSINESS, Crop farming/FARMING, Self-employed/self employed
health insurance	12 variants	NHIF /NHIF lapsed/NHIF inactive/nhif/nhif-inactive/NHIF (lapsed); No insurance/no/none
mobile money access	11 variants for ~6 categories	(M-Pesa/Mpesa/MPESA/M-PESA/AIRTEL/airtel money)
food security status	12 variants for 4 categories	Moderately Food Insecure/ moderately insecure; Food Secure/ FOOD SECURE/secure; Severely Food Insecure/SEVERELY INSECURE/chronically food insecure
dependency ratio category	6 variants for 3 categories	High (>1.0)/HIGH; Medium (0.5-1.0)/medium; Low (0-0.5)/low
receives remittances	8 variants for 3 categories	(Yes/yes/YES, No/no/NO, Occasionally/occasionally)
dwelling type	12 variants for ~7 categories	Permanent/permanent/PERMANENT; Semi-permanent; mud house/Mud/wattle; single room inconsistently categorised
land ownership	11 variants for 5 categories	Owned/owned/OWNED; Rented/rented; communal/Communal; no land/No land

3.4 Numerical Outliers and Implausible Values

Column	Issue Type	Example
Age	12 negative values 259 values above 100 209 values under 18	-19/-20/-23/-25/-27 124/150 16/17
income_monthly_kes (74 records with impossible or implausible values)	47 negative values 27 values exceed KES 9,000,000	min = -100 KES Potential outliers. The mean (62,721) is substantially higher than the median (21,869), confirming a positively skewed distribution
household_size	957 records with value = 0	A household cannot have zero members

3.5 Date Format Inconsistency

interview_date has no missing values but contains 4 distinct formats across 5,512 unique dates. All must be normalised to ISO 8601 (YYYY-MM-DD) before any time-based analysis:

- **ISO format:** 2023-04-22
- **Day-first:** 23-11-2024
- **Month/Day/Year:** 04/25/2024
- **Month/Day/Year:** June 09 2022

3.6 Mixed Data Types

203 cells in the **distance_to_health_facility_km** variable contain the string 'Don't know' in an otherwise numeric column. This prevents all distance-based calculations.