

KENYA HOUSEHOLD SURVEY DATA _ CLEANING LOG

	A	B	C	D	E	F
1	KENYA HOUSEHOLD SURVEY DATA _ CLEANING LOG					
2	log_id	column_name	issue_type	action_taken	records_affected	analyst_note
3	1	age	Missing values	Converted empty to null	123	Empty strings converted to NULL for consistency across all numeric columns (NULL represents values unknown to analyst, distinct from no income or 0 income)
4	2	age	Negative values	All converted to null	12	No valid age can be negative. Flagged as Invalid age - Negative Value in age_flag
5	3	age	Values above 100 (124 & 150)	All converted to null	259	Highly implausible for a primary household head. Flagged as Implausible above 100 in age_flag
6	4	age	Ages under 18(16 and 17)	Retained and flagged as emancipated minors	209	Emancipated minors heading their own households are legitimate respondents in the Kenyan context. Flagged as emancipated_minor_unconfirmed in age_flag
7	5	gender	Inconsistent coding - 6 variants found for 2 valid categories (Male, male, M, Female, female, F)	Standardised all variants to Male and Female using UPPER/LOWER case; Blanks converted to Unknown	5494	Unknown introduced as a category to preserve blank records - For consistency across all text columns
8	6	education_level	Inconsistent coding; 7 variants found for 4 valid categories	Standardised all variants to Primary, Secondary, University and College/Tvet. Blanks converted to Unknown	2966	Blanks converted to Unknown consistent approach for categorical variables across all text columns
9	7	income_monthly_kes	Missing values	Converted to NULL	1266	NULL retained to distinguish known missing from zero income
10	8	income_monthly_kes	Decimal precision	Rounded all valid values to 2 decimal places	9346	No substantive data change
11	9	income_monthly_kes	Negative values(-100)	All converted to null	47	Logically impossible for reported household income. Flagged as Invalid -Negative value
12	10	income_monthly_kes	Extreme outlier(9999999)	All converted to null	27	Almost certainly a system sentinel or data entry error. Flagged as Implausible - Extreme outlier in income_flag
13	11	household_size	Missing values	Converted to NULL	833	NULL represents unknown household size, consistent with missing values in numeric columns
14	12	household_size	Zero values	Converted to NULL	957	A household of zero members is logically impossible â€” the respondent alone constitutes at least one member. Zero values treated as data entry errors rather than valid responses, distinct from zero income which was retained as a valid value
15	13	water_source	Inconsistent coding; 10 variants found for 5 valid categories; Piped/piped water, River water/river, Rainwater/rain water, Borehole/bore hole, Water vendor/vendor	Standardised all variants to Rain Water, Borehole, Piped Water, Water Vendor and and River. Blanks converted to Unknown	5428	Blanks converted to Unknown consistent approach for categorical variables across all text columns
16	14	sanitation_type	Inconsistent coding; 8 variants for 4 categories	Standardized all variants to ;Open Defecation,Pit Latrine,VIP Latrine,Flush Toilet. Blanks converted to Unknown	4047	Blanks converted to Unknown consistent approach for categorical variables across all text columns

KENYA HOUSEHOLD SURVEY DATA _ CLEANING LOG

	A	B	C	D	E	F
1	KENYA HOUSEHOLD SURVEY DATA _ CLEANING LOG					
2	log_id	column_name	issue_type	action_taken	records_affected	analyst_note
17	15	interview_date	Inconsistent date formats; 4 distinct formats found: ISO (YYYY-MM-DD), Day-first (DD-MM-YYYY), Month/Day/Year numeric (MM/DD/YYYY), Month/Day/Year text (Month DD YYYY)	Normalised all dates to ISO 8601 standard (YYYY-MM-DD) using CASE-WHEN with regex pattern matching for each format variant	7087	Ambiguous numeric dates interpreted as DD/MM/YYYY based on Kenyan date conventions and dataset origin.
18	16	residence_type	Inconsistent coding; 8 variants for 3 categories;rural/Rural/RURAL, Peri-urban/peri urban, URBAN/Urban/urban	Standardized all variants to; Urban,Rural and Peri-Urban. Blanks converted to Unknown	2116	Blanks converted to Unknown consistent approach for categorical variables across all text columns
19	17	employment_status	Inconsistent coding â€” 16 variants found for 4 valid categories. Variants included case differences (EMPLOYED, Employed, employed), descriptive entries (Formally employed, Employed (Formal), Employed (Informal)), and agriculture-specific entries (Farmer, Smallholder Farmer, Subsistence Farmer, Pastoralist)	Standardised all variants to 4 categories: Employed, Unemployed, Self Employed, Casual Labour. Blanks converted to Unknown	8766	Agriculture-related entries (Farmer, Smallholder Farmer, Subsistence Farmer, Pastoralist) mapped to Unemployed - subsistence farming and pastoralism represent livelihood activities rather than formal or informal employment in the Kenyan context. GranularDetails of these activities are preserved in the livelihood_source column. Student and Not Working also mapped to Unemployed as neither represents active employment
20	18	livelihood_source	Inconsistent coding; 24 variants found across case differences, descriptive entries and overlapping categories. Variants included case differences (FARMING, Farming, farming), crop-specific entries (Crop farming, Maize farming, Tea farming), and business-related entries (Petty trade, Hawking, Business, Charcoal/firewood)	Standardised all variants. Consolidated into canonical categories: Business/Trade, Employment, Farming, Transport Services. Blanks converted to Unknown	5438	Consolidation decisions made: (1) Petty Trade, Hawking, Business and Charcoal/Firewood grouped into Business/Trade (They represent informal or semi-formal trade activities). (2) Formal Wage and Salaries grouped into Employment (Both represent formal employment arrangement. (3) Crop Farming, Maize Farming and Tea Farming grouped into Farming (Crop-specific entries retained as farming to avoid fragmentation (4) Bodaboda mapped to Transport Services(It is a distinct and common livelihood in Kenya warranting its own category). Remaining entries (Fishing, Remittances, Mixed Farming, Livestock, Agri-business, Domestic Work, Rental Income, Casual Farm Work, Multiple Sources) retained as-is as they represent sufficiently distinct livelihood categories
21	19	health_insurance	Inconsistent coding; 12 variants for 8 categories. Variants included case differences, (NHIF, nhif, NHIF lapsed, NHIF inactive, nhif-inactive, NHIF (lapsed)) and uninsured entries (No insurance, no, none)	Standardised to 8 categories: NHIF, NHIF - Inactive, SHA, Linda Mama, Employer-provided, Private insurance, Community-based, None. Blanks converted to Unknown	4079	NHIF retained as a distinct category - data collection spanned the NHIF to SHA transition; converting NHIF records to SHA cannot be done with certainty for all respondents. NHIF lapsed, NHIF inactive and all related variants consolidated into NHIF - Inactive, indicating households that did have insurance, but couldnt use it due to delayed payments, analytically distinct from None which represents households that have never been insured. No, none and No insurance all mapped to None

KENYA HOUSEHOLD SURVEY DATA _ CLEANING LOG

	A	B	C	D	E	F
1	KENYA HOUSEHOLD SURVEY DATA _ CLEANING LOG					
2	log_id	column_name	issue_type	action_taken	records_affected	analyst_note
22	20	mobile_money_access	Inconsistent coding; 11 variants for 5 categories. Variants included case differences (M-Pesa, Mpesa, MPESA, M-PESA) and platform name inconsistencies (AIRTEL, Airtel Money)	Standardised to 5 categories: M-Pesa, Airtel Money, T-Kash, M-Pesa & Airtel Money, No. Blanks converted to Unknown	7427	Platform names standardised to official branding. No represents households with no mobile money access
23	21	food_security_status	Inconsistent coding; 12 variants for 6 categories. Variants included case differences (FOOD SECURE, Food Secure, secure) and descriptive inconsistencies (Moderately Food Insecure, moderately insecure, Severely Food Insecure, chronically food insecure)	Standardised to 6 categories: Food Secure, Mildly Insecure, Moderately Insecure, Severely Insecure, Chronically Insecure, Food Insecure. Blanks converted to Unknown	8550	Chronically food insecure retained as distinct from Severely Insecure (chronic insecurity implies persistent/ongoing food deprivation)
24	22	distance_to_health_fac	Mixed data types; text entries of Don't Know and blank entries in a numeric column	Don't Know entries and blanks replaced with NULL	599	Don't Know represents a genuine non-response and cannot be inferred or estimated. Blanks treated as missing for the same reason, explaining why they are all nulled
25	23	dependency_ratio_cate	6 variants for 3 categories; mixed category and dependency ratio range	Standardized and retained 3 main categories(low, medium,high); Separated range and category; Blank entries converted to unknown.	7057	Ratio ranges extracted into a separate column named dependency ratio, to preserve the numeric detail without corrupting the categorical field. Column renamed to dependency category to reflect content
26	24	receives_remittances	Inconsistent coding; 8 variants for 3 categories. Variants were purely case differences (Yes/yes/YES, No/no/NO, Occasionally/occasionally)	Standardised to 3 categories: Yes, No, Occasionally. Blanks converted to Unknown	7795	Occasionally retained as a distinct category as it indicates partial remittance receipt
27	25	dwelling_type	Responses contained 12 variations in terminology, capitalization, and housing descriptions	Standardized into four categories: Permanent House, Apartment/Rental, Semi-permanent House, and Traditional House.	9500	Similar housing types were grouped into standard categories. Mabati houses were classified as Semi-permanent due to regional construction norms; single rooms and bedsitters were considered rentals due to limited structural information in the data
28	26	land_ownership	11 variants for 5 categories	Consolidated to canonocal values, to have 5 distinct categories	9500	Borrowed/caretaker retained as a category to capture informal arrangements such as caretaker or borrowed land access,and does not represent formal ownership or rental agreements
29	27	household_id	Duplicate entries	Used row_number() partitioning by household_id and respondent_id pairs to retain unique entries per pair	50	Duplicate records were identified based on repeated combinations of household_id and Respondent_id
30	28	respondent_id	Duplicate respondent_ids associated with different household ids, but identical records across al other columns	Removed duplicate & retained single entry per record, after confirming all other fields matched	10	Records were determined to be duplicate entries rather than distinct observations due to identical respondent information and interview dates

KENYA HOUSEHOLD SURVEY DATA _ CLEANING LOG

	A	B	C	D	E	F
1	KENYA HOUSEHOLD SURVEY DATA _ CLEANING LOG					
2	log_id	column_name	issue_type	action_taken	records_affected	analyst_note
31	29	Household_id	Duplicate Household ids with identical records across all columns	Removed duplicate records and retained one observation per duplicated Household ID	20	Upon inspection of other variables, records were determined to be duplicates, and were deleted to retain unique entries, maintaining household level uniqueness
32	30	Entire dataset	non defined data types for value fields	converted fields to appropriate data types, and set primary keys	All columns	Household id made primary key, while respondent id was made a unique key. Numeric and categorical columns defined appropriately based on their datatypes