Codebook for Cleaned Indian Census Population Dataset

Madhav Singh

December 24, 2025

Table 1: Codebook for Cleaned Indian Census Population Dataset

Variable	Description	Unit of Observation	Type	Valid Range / Categories	Missingness	Notes
ent_id	Unique identifier for each administrative entity in the dataset	Administrative entity	String (Identifier)	See notes	0 (0.0\{}%)	Constructed identifier representing a single administrative entity (India, state/UT, district, or subdistrict); each entity appears multiple times across rows due to different urbanisation classes. Generated by concatenation of st_code, dist_code, and sbd_code; not part of original census coding.
unique_id	Row-level unique identifier for each administrative entity and its urbanisation component	Administrative entity × urbanisation class	String (Identifier)	See notes	0 (0.0\{}%)	Constructed identifier that uniquely identifies each row in the dataset. Generated by concatenation of st_code, dist_code, sbd_code, and the first letter of dem_class; not part of original census coding.

Continued on next page

Table 1: Codebook for Cleaned Indian Census Population Dataset (Continued)

Variable	Description	Unit of Observation	Type	Valid Range / Categories	Missingness	Notes
st_code	Official census state or union territory code	State / Union Territory	String (Identifier)	See notes	0 (0.0\{}%)	Highest-level administrative identifier in the census hierarchy. Encoded as a two-digit string; coded as '00' for aggregate India-level entries. Stable across all constituent districts and subdistricts.
dist_code	Official census district code within each state or union territory	District	String (Identifier)	See notes	0 (0.0\{}%)	Middle-level administrative identifier in the census hierarchy. Encoded as a three-digit string; coded as '000' for India-level and state/UT-level entries. Stable across all constituent subdistricts.
sbd_code	Official census subdistrict (tehsil/taluka) code within each district	Subdistrict	String (Identifier)	See notes	0 (0.0\{}%)	Lowest-level administrative identifier in the census dataset. Encoded as a five-digit string; coded as '00000' for India-level, state/UT-level, and district-level entries. Code '99999' denotes 'Area not under any Sub-district', used for geographic areas not formally categorized into standard subdistrict units.
ent_type	Administrative level of the entity represented in the row	Administrative entity	Categorical / String	DISTRICT, INDIA, STATE, SUB-DISTRICT	0 (0.0\{}%)	Categorical variable with values: 'INDIA', 'STATE', 'DISTRICT', and 'SUB-DISTRICT'. Indicates the hierarchical level of the administrative unit.
name	Official name of the administrative unit	Administrative entity	Categorical / String	See notes	0 (0.0\{}%)	Names standardized for formatting and whitespace consistency during cleaning.

Continued on next page

Table 1: Codebook for Cleaned Indian Census Population Dataset (Continued)

Variable	Description	Unit of Observation	Type	Valid Range / Categories	Missingness	Notes
dem_class	Urbanisation demographic classification of the administrative entity	Administrative entity × urbanisation class	Categorical / String	RURAL, TOTAL, URBAN	0 (0.0\{}%)	Categorical variable with values 'TOTAL', 'RURAL', and 'URBAN'. Each administrative entity is represented by up to three rows corresponding to these classes; 'TOTAL' represents the sum of rural and urban components where applicable and defines the structural logic of the dataset.
ivlg_count	Number of inhabited villages within the administrative unit	Administrative entity × urbanisation class	Integer	0 – 597608	0 (0.0\{}%)	Non-zero values occur only for 'RURAL' and 'TOTAL' entries. Always zero for 'URBAN' entries by definition. May be zero for 'RURAL' or 'TOTAL' entries if no rural component exists or for special administrative entities that function as neighbourhoods or commercial centres.
uvlg_count	Number of uninhabited villages within the administrative unit	Administrative entity × urbanisation class	Integer	0 – 43324	0 (0.0\{}%)	Non-zero values occur only for 'RURAL' and 'TOTAL' entries. Always zero for 'URBAN' entries. May be zero for rural entities with no uninhabited villages or for special administrative entities functioning as neighbourhoods or commercial centres.
town_count	Number of census or statutory towns within the administrative unit	Administrative entity × urbanisation class	Integer	0 – 7933	0 (0.0\{}%)	Non-zero values occur only for 'URBAN' and 'TOTAL' entries. Always zero for 'RURAL' entries by definition. May be zero for 'URBAN' or 'TOTAL' entries when no urban component exists or for special neighbourhood or commercial-centre entities.

Continued on next page

Table 1: Codebook for Cleaned Indian Census Population Dataset (Continued)

Variable	Description	Unit of Observation	Type	Valid Range / Categories	Missingness	Notes
hhld_count	Total number of households within the administrative unit	Administrative entity × urbanisation class	Integer	0 – 249501663	0 (0.0\{}%)	May be zero for 'RURAL' or 'URBAN' entries when that component does not exist within an administrative entity. Always non-zero for 'TOTAL' entries, including special neighbourhood or commercial-centre entities.
total_pop	Total resident population of the administrative unit	Administrative entity × urbanisation class	Integer	0 – 1210854977	0 (0.0\{}%)	May be zero for 'RURAL' or 'URBAN' entries when that component does not exist. Always non-zero for 'TOTAL' entries, indicating the absence of ghost administrative units.
male_count	Total male resident population of the administrative unit	Administrative entity × urbanisation class	Integer	0 – 623270258	0 (0.0\{}%)	May be zero for 'RURAL' or 'URBAN' entries when that component does not exist. Always non-zero for 'TOTAL' entries.
female_count	Total female resident population of the administrative unit	Administrative entity × urbanisation class	Integer	0 – 587584719	0 (0.0\{}%)	May be zero for 'RURAL' or 'URBAN' entries when that component does not exist. Always non-zero for 'TOTAL' entries.

Continued on next page

Table 1: Codebook for Cleaned Indian Census Population Dataset (Continued)

Variable	Description	Unit of Observation	Type	Valid Range / Categories	Missingness	Notes
area_sqkm	Geographic area of the administrative unit in square kilometers	Administrative entity × urbanisation class	Float	0.0 – 3287469.0	117 (0.59\{}%)	May be missing for certain entities due to census data limitations. For some 'TOTAL' entries, rural and urban areas may not sum perfectly due to very small floating-point discrepancies. May be zero for non-existent 'RURAL' or 'URBAN' components but never zero for 'TOTAL' entries.
pop_per_sqkm_old	Population density as reported in the raw census source	Administrative entity × urbanisation class	Float	0.0 – 89185.0	117 (0.59\{}%)	May be missing for the same entities where area data is unavailable. Retained for reference despite mixed integer and float values arising from inconsistent rounding in the raw census source. May be zero for non-existent 'RURAL' or 'URBAN' components but never zero for 'TOTAL' entries.
pop_per_sqkm_new	Recomputed population density based on cleaned population and area values	Administrative entity × urbanisation class	Float	0.0 – 89185.0216047336	117 (0.59\{}%)	May be missing for the same entities where area data is unavailable. Computed as total population divided by area; consistently stored as float. Equals zero when area is zero due to non-existence of a component. Never zero for 'TOTAL' entries.