**📌 Purpose**

This dataset consolidates data from multiple authoritative sources at the U.S. county level, enabling comprehensive cross-sector analysis. It includes:

* **Social, economic, and health indicators**
* **Population and demographic statistics**
* **Fair Market Rent (FMR) values from HUD**

The objective is to produce a single, standardized dataset with consistent coverage across all U.S. counties (excluding territories) for use in analytics, dashboarding (e.g., Tableau), and policy research.

**📁 Data Sources & Merged Files**

**1. merged\_df\_1**

Primary dataset with 50+ indicators covering healthcare, education, crime, housing, environment, and access.

**2. merged\_df\_2**

Demographic and population data from the Population and CO-EST2024-POP sheets.

**3. fmr\_df**

HUD Fair Market Rent values (2020–2024), segmented by unit type and county.

**🔄 Merge Methodology**

* **Order of merge:**
  1. merged\_df\_1 (base)
  2. Left join with merged\_df\_2
  3. Left join with fmr\_df
* **Merge keys:** FIPS, State, County
* **Join type:** Left join (to preserve base structure)
* All unmatched rows were retained as NaN — no data was dropped.

**🧹 Data Cleaning & Standardization**

**🔢 FIPS Code Formatting**

* All FIPS codes were padded to 5 characters (str.zfill(5))
* Verified format: 2-digit state code + 3-digit county code
* Inflated FIPS-like values (e.g., 2300177800) were truncated to valid 5-digit codes.

**🗃️ Deduplication**

* **Step 1:** Removed exact duplicates across all columns
* **Step 2:** Identified partial duplicates (based on FIPS, State, County) and aggregated numeric values using **median**
* Ensured only one row per county

**🌎 Geographic Scope**

* Excluded U.S. territories:
  + AS (American Samoa), GU (Guam), MP (Northern Mariana Islands), PR (Puerto Rico), VI (U.S. Virgin Islands)
* Retained **DC (District of Columbia)** as a valid county-equivalent.

**🧯 FIPS Corrections**

* Created a manual FIPS override dictionary for known mismatches in fmr\_df
* Final FIPS\_final column was renamed back to FIPS
* Column order standardized with FIPS as the third column for consistency

**🧼 Missing Value Handling**

* Final dataset retains NaN values for transparency
* In a special cleaned version (FMR-only), missing numeric values were imputed using the **state-level mean for Alaska (AK)** to preserve comparability
* This imputed version contains **no missing values**

**❗ Missing Data Overview**

* All **3,198 counties** are represented, though many have missing values in specific indicators
* **64 counties** had no match in FMR data
* **75 counties** lacked demographic data in merged\_df\_2

**High-Missingness Columns (for awareness)**

Indicators with widespread missing data (often crime-related):

* disconnected\_youth
* drug\_overdose\_deaths
* homicides
* juvenile\_arrests
* violent\_crime (especially for 2023–2024, likely due to reporting delays)

**📤 Output Files**

* final\_merged\_dataset.xlsx
* final\_merged\_dataset.csv

**📌 Notes for Future Use**

* All missing values are clearly marked as NaN
* Manual FIPS corrections are saved separately for transparency and reproducibility
* Dataset is structured and formatted for direct use in modeling, mapping, or dashboard development