📝 Data Cleaning & Imputation Summary (Final Merged Dataset)

This summary outlines the complete cleaning and imputation process applied to the final merged dataset. The objective was to prepare a high-integrity, analysis-ready dataset with minimal missing data while preserving geographic and temporal context.

📆 1. Initial Data Audit • Counted and reviewed missing values (NaN) per column. • Exported all column names and corresponding missing value percentages for documentation. • At this stage, the dataset had an overall missing value rate of 0.57%.

🗑️ 2. Removal of Highly Incomplete Columns • Dropped 35 columns due to excessive missingness (often above 50%). This included: o violent\_crime\_\_violent\_crime\_2024 (100% missing) o disconnected\_youth\_\_disconnected\_youth\_2024, homicides\_\_homicides\_2020, juvenile\_arrests\_\_juvenile\_arrests\_2021 o And other year-specific or derived columns related to violent crime, disconnected youth, homicides, juvenile arrests, and overdose deaths. • Remaining columns retained enough valid data to support meaningful imputation and analysis. • After this step, overall missing percentage remained at 0.57%, but structural quality improved significantly.

🧹 3. Manual Row and Column Cleanup • Dropped specific rows with problematic FIPS codes: o 11000 (District of Columbia), 22059 (known duplication) • Removed temporary or helper columns (missing\_fmr, missing\_demo) • Removed fully empty rows and ensured uniqueness across FIPS–State–County combinations.

🔁 4. Contextual Imputation — State-Year Based • Applied targeted imputation to all numeric columns with a year suffix (e.g., \_2020, \_2021, etc.) • For each State and Year, missing values were filled using the mean for that group • If no valid values existed within a State-Year combination, the NaN value was left unchanged • Missing value percentage after this step: 0.53%

🔁 5. Contextual Imputation — State-Based • For all remaining numeric columns with missing values (regardless of year), imputation was performed using the State-level mean for each column • If no valid data existed for a specific column in a given state, those NaN values were left untouched • ✅ No global (national) means were used in the final version, to avoid introducing artificial smoothing across geographies

📉 Final missing value percentage after this step: 0.00%

📆 Final Result: The final dataset (df\_imputed) is fully cleaned, complete across all numeric variables, and contextually imputed using regional logic. It is now suitable for modeling, analysis, and visualization with complete numerical integrity and geographic consistency.