Vera
INSTITUTE OF JUSTICE

Incarceration and Inequality Project

Economic Dataset (v1)

Codebook & Methodology

May 31, 2025

Table of Contents

Introduction

Involvement in the criminal legal system has a profound and persistent impact on the well-being of people who are directly affected and their loved ones, undercutting economic opportunity and trapping families and children in poverty. With almost 2 million people behind bars on any given day, mass incarceration is a primary driver of disparities in wealth and income, yet—until now—we have not had the tools to track its impact on communities around the country.

Building on the Vera Institute of Justice's (Vera's) existing research and data projects, Vera's Incarceration and Inequality Project (IIP) team is working to compile and analyze data that illuminates the nexus of inequality and incarceration as a national research and policy priority.

The team's priorities include conducting research and constructing datasets that describe the connections between incarceration and economic outcomes, developing an online visualization tool that can be used by wide-ranging audiences to elucidate these connections and designing materials to increase awareness and inform in-community and policy responses.

This dataset is the first version of a resource that Vera will maintain as a companion to the established Incarceration Trends data, which Vera has maintained since 2015. This iteration of the dataset presents the data effectively as-is from the sources, with quality control checks and processing to facilitate analysis between disparate data sources. Vera includes the full breadth of variables to enable the research community's exploration of the relationships between incarceration and economic inequality. The Vera team will release a second version of the dataset alongside a data visualization tool, in January 2026. The next version will offer enhancements to the initial dataset such as additional history, variables, and new metrics developed by Vera.

The dataset is primarily sourced from the U.S. Census Bureau and U.S. Bureau of Labor Statistics, in addition to Vera's Incarceration Trends dataset. It contains detailed demographic, economic, and incarceration data at the state and county levels, with core history covering 2009 to 2023. It contains additional history dating back to 1950 at the county level and 1970 at the state level, where available at the time of collection. Additional details are covered in the Data Sources section of this document.

Codebook

Data sources

This set of data tables combines data from the sources described in detail below. There are two primary groupings of tables based on the geographical units of analysis available: county and state. Vera sourced the incarceration data from Vera's existing Incarceration Trends project, with some additional processing described in the Methodology section. Vera sourced demographic data from the U.S. Census American Community Survey (ACS). The research team sourced economic data from the U.S. Census Small Area Income and Poverty Estimates (SAIPE), Census of Business Patterns (CBP), and ACS datasets, and the U.S. Bureau of Labor Statistics Local Area Unemployment Statistics (LAUS) and Alternative Measures of Labor Underutilization for States (STALT) datasets.

U.S. Census Bureau

American Community Survey[1]

ACS five-year estimates provide detailed demographic data for every year from 2009 through 2023. Vera collected data from the ACS at the county and state levels for the IIP data tables. Vera collected the following series of data from ACS to capture a breadth of demographic and economic information that can be analyzed to better understand the relationship between incarceration and economic inequality: educational attainment, household size, household relationships, household type, public assistance, health insurance status, public health insurance status, poverty status, employment status, median household income, median rent, median rent as a percentage of median income, housing tenure, occupation by class of worker, and overall population statistics (race, sex, and age). Vera provides estimates and margins of error for all series of data.

Small Area Income and Poverty Estimates[2]

The SAIPE program provides single-year estimates of income and poverty for all states and counties. Vera extracted data for all years available at the time of collection, which included 1995 through 2022. Vera's SAIPE dataset includes median household income and poverty estimates for three age groups: all ages, ages five to 17 in families, and under age 19. Vera includes this dataset because it contains additional years of data compared to the ACS; however it does not include race, sex, or more granular age breakdowns.

[1] U.S. Census Bureau, "American Community Survey 5-year Data (2009-2023)," December 12, 2024, https://www.census.gov/data/developers/data-sets/acs-5year.html.

[2] U.S. Census Bureau, "Small Area Income and Poverty Estimates Program, About" https://www.census.gov/programs-surveys/saipe/about.html

Census of Business Patterns[3]

CBP provides information about establishments with paid employees by industry on an annual basis. Vera sourced CBP data at the county level and aggregated across all sectors for which data was reported. For years prior to 2007, the U.S. Census Bureau applied complementary cell suppression to protect individual establishments from disclosure where necessary. Starting in 2007, the Census infused minimal noise (typically less than 5 percent) to accomplish the same goal. The Census primarily applied these methods to sub-sectors containing a small number of businesses, with suppressed or de-noised values being included in higher-level sector code aggregates, which minimizes the impact on county- and state-level aggregates.

U.S. Bureau of Labor Statistics

Local Area Unemployment Statistics[4]

The LAUS program provides monthly estimates of total employment and unemployment. Vera collected data at the annual and county level for this dataset. Vera produced county-level LAUS estimates using data from the Current Population Survey (CPS), the Current Employment Statistics (CES) survey, state unemployment insurance systems, and the ACS.

Alternative Measures of Labor Underutilization for States[5]

STALT data contains six alternative measures of labor underutilization, which are reported on a rolling quarterly average basis. The BLS uses CPS data to calculate these measures, with U-3 being the official concept of unemployment. The BLS defines these measures consistently across the state and national levels as follows:

- "U-1, persons unemployed 15 weeks or longer, as a percent of the civilian labor force;
- U-2, job losers and persons who completed temporary jobs, as a percent of the civilian labor force;
- U-3, total unemployed, as a percent of the civilian labor force;
- U-4, total unemployed plus discouraged workers, as a percent of the civilian labor force plus discouraged workers;

[3] U.S. Census Bureau, "County Business Patterns, About this Program," https://www.census.gov/programs-surveys/cbp/about.html

[4] U.S. Bureau of Labor Statistics, "Local Area Unemployment Statistics Overview," March 17, 2025, https://www.bls.gov/lau/lauov.htm

[5] U.S. Bureau of Labor Statistics, "Alternative Measures of Labor Underutilization for States, 2024 Annual Averages," January 31, 2025, https://www.bls.gov/lau/stalt.htm

- U-5, total unemployed, plus discouraged workers, plus all other marginally attached workers, as a percent of the civilian labor force plus all marginally attached workers; and
- U-6, total unemployed, plus all marginally attached workers, plus total employed part time for economic reasons, as a percent of the civilian labor force plus all marginally attached workers."[6]

Incarceration Trends

The state- and county-level incarceration data presented in this dataset are from Vera's Incarceration Trends project, which also includes a subsection of population data from the National Cancer Institute's Surveillance Epidemiology and End Results Program (SEER), which were created by the U.S. Census Bureau in partnership with the National Center on Health Statistics. Detailed documentation of the Incarceration Trends dataset can be found on the project's website and GitHub repository, and additional information about the processing conducted to assemble the data contained in this release can be found in the Methodology section of this document.

Variable descriptions

Due to the size of the dataset, Vera grouped the variables into a series of thematically organized tables in CSV files. A detailed table containing all variable names, descriptions, summary statistics, and other relevant information is available on the Incarceration and Inequality GitHub repository.

County-Level Tables	
Table Name	**Count of Variables**
Businesses	66
Education	70
Employment Status - Asian	40
Employment Status - Black	40
Employment Status - Hispanic or Latine	40
Employment Status - Multiple Races	40
Employment Status - Native American or Alaska Native	40

[6] Ibid.

Employment Status - Native Hawaiian or Other Pacific Islander	40
Employment Status - Other Race	40
Employment Status - Totals	131
Employment Status - White	40
Employment Status - White not Hispanic or Latine	40
Health Insurance Coverage - All Races	112
Health Insurance Coverage - All Sexes	144
Health Insurance Coverage - Totals	152
Household Relationships	76
Household Size	32
Household Type	180
Housing Tenure	60
Incarceration	122
Income	22
Occupation Type	42
Population - Asian	60
Population - Black	60
Population - Female	46
Population - Hispanic or Latine	60
Population - Male	46
Population - Multiple Races	60
Population - Native American or Alaska Native	60
Population - Native Hawaiian or Other Pacific Islander	60
Population - Other Race	60
Population - Totals	24
Population - White	60

Population - White not Hispanic or Latine	60
Poverty Status - Asian	112
Poverty Status - Black	112
Poverty Status - Female	56
Poverty Status - Hispanic or Latine	112
Poverty Status - Male	56
Poverty Status - Multiple Races	112
Poverty Status - Native American or Alaska Native	112
Poverty Status - Native Hawaiian or Other Pacific Islander	112
Poverty Status - Other Race	112
Poverty Status - Totals	75
Poverty Status - White	56
Poverty Status - White not Hispanic or Latine	168
Public Assistance	38
Rent	4

State-Level Tables	
Table Name	**Count of Variables**
Businesses	66
Education	70
Employment Status - Alternative Measures	14
Employment Status - Asian	40
Employment Status - Black	40
Employment Status - Hispanic or Latine	40
Employment Status - Multiple Races	40
Employment Status - Native American or Alaska Native	40

Employment Status - Native Hawaiian or Other Pacific Islander	40
Employment Status - Other Race	40
Employment Status - Totals	130
Employment Status - White	40
Employment Status - White not Hispanic or Latine	40
Health Insurance Coverage - All Races	112
Health Insurance Coverage - All Sexes	144
Health Insurance Coverage - Totals	152
Household Relationships	76
Household Size	32
Household Type	180
Housing Tenure	60
Incarceration	104
Income	21
Occupation Type	42
Population - Asian	60
Population - Black	60
Population - Female	46
Population - Hispanic or Latine	60
Population - Male	46
Population - Multiple Races	60
Population - Native American or Alaska Native	60
Population - Native Hawaiian or Other Pacific Islander	60
Population - Other Race	60
Population - Totals	24
Population - White	60

Population - White not Hispanic or Latine	60
Poverty Status - Asian	112
Poverty Status - Black	112
Poverty Status - Female	56
Poverty Status - Hispanic or Latine	112
Poverty Status - Male	56
Poverty Status - Multiple Races	112
Poverty Status - Native American or Alaska Native	112
Poverty Status - Native Hawaiian or Other Pacific Islander	112
Poverty Status - Other Race	112
Poverty Status - Totals	75
Poverty Status - White	56
Poverty Status - White not Hispanic or Latine	168
Public Assistance	38
Rent	4

For data from certain U.S. Census Bureau sources that report margins of error, Vera also includes that data in the same tables as their respective estimates—but Vera does not include them in the variables descriptions table. The variable descriptions table includes source variable names for variables that Vera did not calculate or otherwise modify from their original state.

Methodology

Data processing

Vera kept processing and manipulation of the data to the minimum required to produce a clean and well organized set of tables to allow researchers to apply further processing methods appropriate to their specific needs. Vera extracted all U.S. Census Bureau datasets (SAIPE, ACS, and CBP) via the Application Programming Interface (API), with only CBP data requiring significant processing prior to release. Vera collected the BLS datasets

(LAUS and STALT) through CSV downloads, and they similarly required very little processing.

Cleaning

Vera reviewed all datasets for completeness and quality and renamed variables for clarity (e.g., renaming the variable “B15002_019” as “educational_attainment_female”).

SAIPE

Though Vera collected SAIPE data for all years of data available at the time of extraction (1995 through 2022), years 1995 through 1997 were missing data across many variables. Vera excluded those years. Additionally, the series of variables in the SAIPE dataset pertaining to the under five age group were also missing a substantial number of records across all years, so Vera also excluded those variables.

CBP

Between 1997 and 1998, the set of codes used for classifying businesses in the CBP changed from the Standard Industrial Code (SIC) system to the North American Industry Classification System (NAICS). Theoretically, each detailed SIC code maps directly to a detailed NAICS code. However, there are part indicators across both sets of codes that are not present in CBP data, making it impossible to programmatically or manually map all possible SIC codes to NAICS codes consistently and accurately at any level, including the most general form of rolling up to two-digit sector codes. Additionally, the NAICS system splits certain concepts where SIC keeps them together, contributing to the difficult many-to-many relationship present in reconciling this dataset’s history. For example, the CBP often combines retail and wholesale trade within various sub-industries into a single three- or four-digit SIC code, while NAICS classifies retail and wholesale trade under completely separate two-digit sector codes.

For the purposes of this release, Vera aggregated CBP variables within two-digit sector codes. Vera removed records prior to 1998 to ensure the accuracy of these aggregations.

County tables

Vera collected data from all sources—with the exception of STALT—at the county level, so minimal reshaping was necessary for the creation of tables unique at the year-county level.

Incarceration Trends data

Vera collects county-level data for the Incarceration Trends project on a quarterly basis. Vera used linear interpolation to account for missing data between existing records in the quarterly data. To annualize the data, Vera calculated within-year averages across all numeric variables (i.e., counts and rates) except for admissions and discharges, which were cumulative counts across all quarters.

CBP

Unprocessed CBP data were not unique along any particular axis and included both detailed and aggregated summary rows. To aggregate the data within each two-digit NAICS sector, Vera summed rows across the relevant axes (sector codes and a series of flags relating to categorical breakdowns across different variables) or extracted them from an existing summary row. The original dataset included indicators for noise and data suppression, however CBP applies these within detailed NAICS codes to prevent the identification of businesses within smaller detailed sectors; they did not have a significant impact on two-digit sector-level summaries.

After aggregating data within each two-digit NAICS sector for the selected variables (count of establishments, number of employees, and total annual payroll), Vera reshaped the data for uniqueness at the year and county level.

State tables

Vera collected STALT and ACS data directly from the source at the state level, so did not require geographical aggregation. The BLS reports STALT data on a rolling quarterly basis, with each record containing an average of the previous four quarters of data. As such, Vera extracted the fourth quarter value for each year to generate an annual average. Similarly, Vera annualized quarterly state-level incarceration data by averaging quarterly data within each year.

Vera conducted the following aggregations and calculations to convert county to state level data for the SAIPE, CBP, and LAUS datasets: Vera researchers summed count values across all counties within a state; recalculated rates and percents using the summed numerators and denominators; and, recalculated margins of error using the appropriate formulas as recommended by the U.S. Census Bureau.[7]

Dataset compilation

All tables are unique at the year-county or year-state level, and can be joined using a combination of the `year` and `county_fips` or `state_fips` variables, depending on the unit of analysis being used.

[7] U.S. Census Bureau, *Understanding and Using American Community Survey Data: 8. Calculating Measures of Error for Derived Estimates* (Suitland, MD: U.S.Census Bureau, 2020), https://www.census.gov/content/dam/Census/library/publications/2020/acs/acs_general_handbook_2020_ch08.pdf