**METHODOLOGY REPORT**

**DEBT COLLECTION LAB**

August 2024

**Debt Collection Lab – Data Collection & Transformation Methodology**

The Debt Collection Lab has put together a repository of debt collection data from court systems across the United States. This document provides more information about the methods used to collect and transform the data for analysis. The Debt Collection Lawsuit Tracker is updated every six months.

**Identifying Debt Collection Lawsuits**

Each jurisdiction has different data collection practices, and court documents may not identify whether a case was a debt collection lawsuit. Instead, debt collection cases are often classified under more general categories, such as “Civil” or “Small Claims”. The Debt Collection Tracker statistics are calculated based on all lawsuits within those more general categories.

To restrict this dataset to just business-to-person debt collection lawsuits, we use an algorithm to identify plaintiffs that are businesses and defendants that are people. We then restrict the dataset to just those cases. Because we restrict these cases to just business-to-person cases, our counts may differ from state-reported numbers or numbers reported to the Court Statistics Project as “seller plaintiff” cases.

Below, we list information a bout how debt collection lawsuits were identified in each of the jurisdictions included in the Debt Collection Tracker. This table outlines what case type we looked at, what jurisdictions these are heard, how we collected the data, what court looked at, and the dollar limit of cases.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Location[[1]](#footnote-1) | Case Classification | Jurisdiction | Collection Method | Court | Dollar limit |
| Marion County (Indiana) | CC – Civil Collection | General jurisdiction | Provided directly by courts | County Civil | $8,000 to $50,000 |
| Lake County (Indiana) | CC – Civil Collection | General jurisdiction | Provided directly by courts | Superior Court | $8,000 to $50,000 |
| St. Joseph County (Indiana) | CC – Civil Collection | General jurisdiction | Provided directly by courts | Superior Court | $8,000 to $50,000 |
| Elkhart County (Indiana) | CC – Civil Collection | General jurisdiction | Provided directly by courts | Superior Court | $8,000 to $50,000 |
| Hamilton County (Indiana) | CC – Civil Collection | General jurisdiction | Provided directly by courts | Superior Court | $8,000 to $50,000 |
| St. Louis County (Missouri) | AC Suit on Account | General jurisdiction | Web scraping | Associate Civil | $5,000 to $25,000 |
| St. Louis City (Missouri) | AC Suit on Account | General jurisdiction | Web scraping | Associate Civil | $5,000 to $25,000 |
| Jackson County (Missouri) | AC Suit on Account | General jurisdiction | Web scraping | Associate Civil | $5,000 to $25,000 |
| St. Charles County (Missouri) | AC Suit on Account | General jurisdiction | Web scraping | Associate Civil | $5,000 to $25,000 |
| Clay County (Missouri) | AC Suit on Account | General jurisdiction | Web scraping | Associate Civil | $5,000 to $25,000 |
| North Dakota | Contract / Collection | General jurisdiction | Web scraping | District Civil | Above $15,000 |
| Harris County (Texas) | Debt Claim | Limited jurisdiction | Public data extract | Justice of the Peace | $20,000 or less |
| Hamilton County (Ohio) | Contract or Money | General jurisdiction | Web scraping | Municipal Civil Court | $500 to $15,000 |
| Philadelphia County (PA) | Contract – Civil  | Limited jurisdiction | Legal aid | Municipal Court | $12,000 |

**Data Collection Methods & Jurisdictions**

The data housed by the Debt Collection Lab are obtained through various methods, including data sharing agreements with the courts, public information act requests, publicly available data extracts, and web scraping.

Data obtained by data sharing agreements and public information act requests are delivered specifically for the Debt Collection Lab in a machine-readable file format. Data obtained through publicly available data extracts are also obtained in a machine-readable file format. Data obtained through web scraping were collected through the publicly accessible web interface for the court and received additionally quality assurance checks for accuracy and timeliness. This includes checking case counts collected against case courts reported to a state court administrative office that publishes annual statistical reports, as well as verifying totals with Legal Services Corporation and others.

Because each jurisdiction has different laws and legal procedures, the data do not match up perfectly from state to state and sometimes jurisdiction to jurisdiction. However, to be included in the tracker, every case in the Debt Collection Lab tracker needs to have common data elements as seen below.

|  |  |
| --- | --- |
| **Must Have** | **Nice to Have** |
| Case number | Judgment amount |
| Case status / disposition | Garnishment details |
| Case action / type | Execution details |
| Plaintiff and Defendant names | Attached case documents |
| Plaintiff and Defendant attorney names |  |
| Plaintiff and Defendant locations |  |

**Data Transformation**

After collecting the data from its source, we cleaned the defendant’s address by removing errant characters and trimming white space.

**Addressing cleaning.** We clean the defendant’s address using the above methods, which is then added to the dataset as latitude and longitude fields. This allows us to determine the 2020 Census tract GEOIDs of the coordinates and include those IDs in the dataset.

**Plaintiff names.** In court data, the same plaintiff may appear with multiple spelling variations. For example, without cleaning and standardization, “Portfolio Recovery Assoc.” and “Portfolio Recovery Associates” would appear as separate plaintiffs. In order to clean and standardize the names, we performed a fuzzy matching [algorithm](https://github.com/Debt-Collection-Lab/algorithms.git) on each state’s data, selecting the most common spelling for a given plaintiff name as the standard across all variations. Additionally, we developed a specialized [algorithm](https://github.com/Debt-Collection-Lab/algorithms.git) to detect plaintiffs that are businesses (as opposed to individuals filing lawsuits against other individuals) and restricted the final datasets to only business filing against individuals. As a final step, we standardized the names of national debt buyers that file consistently across multiple states. This way, “Portfolio Recovery Associates” is spelled consistently across jurisdictions and allows for comparison, regardless of what the top spelling is by state.

**Calculating Rates in Tracts with Small Populations**

The Census Bureau denotes special areas (Census codes 9800 through 9899) that contain little or no population, in addition to representing a relatively large land use area such as industrial or business parks, parks, or airports, to name a few. Because these small population denominators make these census tracts appear to have much higher rates than others, we manually review each of these areas before choosing to exclude them. The table below lists tracts we currently exclude.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| GEOID | State | County | Lawsuits | Total Population |
| 29095989100 | Missouri | Jackson County | 6 | 21 |
| 29095988300 | Missouri | Jackson County | 4 | 173 |
| 29095989200 | Missouri | Jackson County | 1 | 0 |
| 48201980200 | Texas | Harris County | 7 | 9712 |
| 48201980700 | Texas | Harris County | 118 | 1427 |
| 48201980000 | Texas | Harris County | 4 | 15 |
| 48201980100 | Texas | Harris County | 3 | 0 |
| 48201980400 | Texas | Harris County | 2 | 4134 |
| 48201980300 | Texas | Harris County | 1 | 2270 |
| 42101980901 | Pennsylvania | Philadelphia County | 32 | 0 |
| 42101980701 | Pennsylvania | Philadelphia County | 1 | 0 |
| 42101980800 | Pennsylvania | Philadelphia County | 6 | 0 |
| 42101980002 | Pennsylvania | Philadelphia County | 1 | 23 |
| 42101989300 | Pennsylvania | Philadelphia County | 11 | 112 |
| 42101980300 | Pennsylvania | Philadelphia County | 1 | 0 |
| 42101980600 | Pennsylvania | Philadelphia County | 1 | 0 |
| 42101989100 | Pennsylvania | Philadelphia County | 2 | 1749 |
| 42101980200 | Pennsylvania | Philadelphia County | 2 | 392 |
| 42101980903 | Pennsylvania | Philadelphia County | 1 | 0 |
| 42101980702 | Pennsylvania | Philadelphia County | 5 | 0 |

**Calculating Racial Disparities in Filings**

The tracker uses a not-equal sign (≠) for counties in which there is a disproportionate number of lawsuits filed in census tracts that are majority-minority. We use 5-year ACS data from 2021 to determine the racial majority of a tract/zip. Then, for every county in the jurisdiction, we determine the share of total countywide lawsuits recorded in each Census tract. If lawsuits were evenly distributed across Census tracts, we would expect tot see that in a county where 20% of the Census tracts are majority Black, for example, those Census tracts would be receiving 20% of the lawsuits. For example, in the tracker report for [Philadelphia County](https://debtcollectionlab.org/lawsuit-tracker/pennsylvania/philadelphia-county/) where we highlight that there are racial disparities, 36% of the neighborhoods are majority White but they only receive 24.8% of lawsuits; by contrast, 36.3% of the neighborhoods are majority Black and they receive 46.5% of the lawsuits. About a fifth of the Philadelphia neighborhoods have no ethnic-racial majority and they receive about a fifth of the lawsuits.

**Standardized Variables**

For the lawsuit tracker, we standardize the data against this data dictionary.

|  |
| --- |
| **Debt Collection Data Dictionary** |
| This data dictionary explains the data columns contained within the CSV file that will be used for lawsuit related data in the tracker |
| **File name** | lawsuit\_data.csv |
| **Data Levels** | Census Tract |
|  |  |
| **Column** | **Description** |
| case\_number | Case number assigned by original court. |
| GEOID\_acs | FIPS code for the census tract based on current ACS boundaries (e.g. 09001100100 for Census Tract 1001, Fairfield County, Connecticut). |
| GEOID\_census | FIPS code for the census tract based on Census 2020 boundaries. |
| plaintiff | Name of the plaintiff (debt collector) based on original court records. |
| date | Date of the lawsuit (%Y-%m-%d). |
| case\_completed | 1 or 0 based on if the lawsuit is completed (1 for yes, 0 for no) |
| default\_judgement | 1 or 0 based on if the lawsuit resulted in a default judgment (1 for yes, 0 for no) |
| has\_representation | 1 or 0 based on if the defendant has legal representation (1 for yes, 0 for no) |
| amount | The amount of the judgment (decimal format, no currency formatting) (e.g. 156.42) |
| court | The name of the jurisdiction data source. |
|  |  |
|  |  |
| **Demographics Data Dictionary** |
| This data dictionary explains the data columns contained within the CSV file that will be used to provide demographic data for census tracts within the tracker |
| **File names** | demographic\_data\_acs.csvdemographic\_data\_census.csv |
| **Data Levels** | Census Tract |
|  |  |
| **Column** | **Description** |
| GEOID | FIPS identifier for the census tract (appended \_acs or \_census) |
| total\_pop | Total population for the census tract |
| non\_hisp | Total non-Hispanic population |
| white | White population |
| black | Black population |
| amer\_ind | American Indian / Native Alaskan population |
| asian | Asian population |
| haw\_pi | Hawaiian / Pacific Islander population |
| other | Population that does not fit prior categories |
| two\_or\_more | Population of two or more races |
| hisp\_total | Total Hispanic population |
| median\_hhi | census data for median household income for the census tract (only ACS) |
| asian\_pct | census data for % Asian for the census tract |
| black\_pct | census data for % Black for the census tract |
| hisp\_pct | census data for % Latine for the census tract |
| white\_pct | census data for % white for the census tract |
| other\_pct | census data for % that do not fit within the above categories for the census tract |
| sum\_of\_pct | Sum of previous percents; used for quality checking |
|  |  |
|  |  |
| **Moratorium Data Dictionary** |
| This data dictionary explains the data columns contained within the CSV file that will be used to display moratorium information for locations |
| **File name** | moratorium\_data.csv |
| **Data Levels** | State or county |
|  |
|  |  |
| **Column** | **Description** |
| id | FIPS identifier for the location the data corresponds to (state or county) |
| moratorium | A text string describing the debt collection moratoriums and any relevant information. |
| moratorium\_start | the start date of the moratorium (mm/dd/yyyy) |
| moratorium\_end | the end date of the moratorium (mm/dd/yyyy) |

In addition to the lawsuit tracker, we also offer CSV files with aggregated case counts. You can access those [here](https://debtcollectionlab.org/get-the-data).

|  |
| --- |
| **Debt Collection Aggregate Data Dictionary** |
| This data dictionary explains the data columns contained within the CSV file that will be used for lawsuit related data in the tracker |
| collector | Name of plaintiff |
| count | Aggregate cases by collector for all years available |
| count21 | Aggregate cases by collector for 2021 |
| count20 | Aggregate cases by collector for 2020 |
| percent\_total | Percentage of total cases by collector |
| percent\_total21 | Percentage of total cases in 2021 by collector |
| percent\_total20 | Percentage of total cases in 2020 by collector |
| dollar\_amount | Total dollar amount by collector across all cases for all years available |
| dollar\_amount21 | Total dollar amount by collector for 2021 |
| dollar\_amount20 | Total dollar amount by collector for 2020 |
| dollar\_percent\_total | Total dollar percent across all cases for all years available |
| dollar\_percent\_total21 | Total dollar percent of cases in 2021 by collector |
| dollar\_percent\_total20 | Total dollar percent of cases in 2020 by collector |
| percent\_with\_representation | Percent of cases with defendant representation by collector |
| percent\_with\_representation21 | Percent of cases with defendant representation in 2021 by collector |
| percent\_with\_representation20 | Percent of cases with defendant representation in 2020 by collector |
| percent\_default\_judgement | Percent of aggregate cases with a default judgment by collector |
| percent\_default\_judgement21 | Percent of cases with a default judgment in 2021 by collector |
| percent\_default\_judgement20 | Percent of cases with a default judgment in 2020 by collector |
| percent\_case\_completed | Percent of cases completed for all years by collector |
| percent\_case\_completed21 | Percent of cases completed in 2021 by collector |
| percent\_case\_completed20 | Percent of cases completed in 2020 by collector |

1. Before Q2 2024, the tracker had data on Connecticut that had been scraped directly from the court. This data was missing cases that the court removed from the public website due to a court rule, which undercounted the total number of cases. We expect to add Connecticut data to the tracker in Q1 2025 using data directly from the court. [↑](#footnote-ref-1)