Data Tools

7. Complete Multiyear Arrest Records for Los Angeles, New York City, and Washington, DC

Charles Murray

American Enterprise Institute

June 2023

Samuel Johnson famously said that a man who is tired of London is tired of life. I feel analogously about the dataset for Data Tools #7. If you are a quantitative social policy analyst and cannot think of good questions to ask about this dataset, you need a new career.

The database, *LawanyArrests.csv*, contains the complete arrest records for arguably the three most important cities in America: Los Angeles, New York, and Washington, DC. The arrest data go back to 2013 for Washington, 2010 for Los Angeles, and 2006 for New York. They continue through calendar year 2022. In all, *LawanyArrests.csv* contains records on 7,122,725 arrests, including for 21,624 murders, 20,519 rapes, 215,378 robberies, 358,801 aggravated assaults, and 113,777 burglaries.

LawanyArrests.csv contains variables coded so they are comparable across the three cities and across years, including not just the arrest date and offense category but also the latitude and longitude of the offense and the arrestee's age, sex, and race. *LawanyArrests.csv* also contains variables with the original police characterizations of the arrests and the charges brought. Two variables that I created enable you to analyze arrests for the eight FBI index offenses and 40 other offense categories across the three cities.

Are you asking whether "broken windows" policing has changed over time? Several of the categories correspond to classic broken-windows offenses (e.g., "Graffiti/Deface Property" and "Damage/Destroy Property"). Do you want to test Steve Sailer's hypothesis that the sudden rise in African American deaths in traffic accidents after summer 2020 resulted from increased police unwillingness to pull over blacks for reckless driving? You can examine racial patterns of arrests

for traffic violations and drunk driving before and after that summer—by week or month as well as year.

LawanyArrests.csv also contains a variable that opens another world of possible analyses. I reverse geocoded the data on latitude and longitude shown in the arrest records to the ZIP codes where the arrests occurred. A supplemental database, *LawanyZips.csv*, contains information on each ZIP code, including its demographic characteristics and inhabitants' educational attainment, household income, and family structure.

You can compare crime in any neighborhood in these three cities with crime in any other neighborhood. You can compare levels and patterns of crime in rich and poor neighborhoods, crime in neighborhoods with high and low proportions of single mothers, and crime in neighborhoods with different levels of racial diversity. You can explore how demographic profiles, family structure profiles, and economic profiles interact in the radically different cultures in parts of Los Angeles, New York, and Washington.

These are a few examples from the long list of uses for *LawanyArrests.csv* and *LawanyZips.csv*. Have fun. As always, let me know of errors that need to be corrected by emailing me at cmurray@aei.org.

Data Tools

7. Documentation for the Los Angeles, New York City, and Washington, DC, Arrest Records

Charles Murray

American Enterprise Institute

June 2023

This document will help you understand and use *LawanyArrests.csv*. It has four sections. "Variables from the Police Files" gives the bare-bones list of the labels and codes of the variables that are unchanged from the publicly available datasets. "Created Variables" does the same thing for variables I created to permit analyses across the three datasets. "Coding Details and Issues" describes aspects of the database you should be aware of before using it. The concluding section describes *LawanyZips.csv*, a dataset with basic demographic and socioeconomic variables for the ZIP codes represented in Los Angeles, New York City, and Washington, DC.

Variables from the Police Files

PDID. The police department's identification number for each arrest, unique within cities but not across cities.

PDARRESTCAT. The police department's classification of the offense category.

PDCHARGE. The police department's description of the most serious charge.

PDLEGALCAT. The police department's coding of legal categories. The codes for each city are:

Los Angeles	**New York**	**Washington, DC**
Felony	Felony	Felony
Misdemeanor	Misdemeanor	Lesser Charge
Violation	Violation	
Infraction	Infraction	
Dependent		
Other		

LATITUDE and LONGITUDE. The geographic coordinates for determining the location of the offense.

AGE. Single-year age of the arrestee in Los Angeles and Washington.

AGEGP. For New York, the age group of the arrestee:

1 <18
2 18–24
3 25–44
4 45–64
5 65+

SEX. Sex of the arrestee, coded "Female" and "Male." Washington had 280 arrests for which its sex variable was coded "Unknown."

PDRACE. Each police department's racial categories. The codes for each city are:

Los Angeles	**New York**	**Washington, DC**
White	White	White
Black	Black	Black
Hispanic/Latin/Mexican	White Hispanic	Asian
Chinese	Black Hispanic	Multiple
Japanese	Asian/Pacific Islander	Other
Korean	American Indian/AK Native	Unknown
Cambodian	Other	
Laotian	Unknown	
Vietnamese		
Asian Indian		
Other Asian		
American Indian/AK Native		
Hawaiian		
Filipino		
Samoan		
Guamanian		
Other Pacific Islander		
Other		
Unknown		

PDLATINO. Only Washington had a separate variable for coding ethnicity. The codes are "Hispanic," "Not Hispanic," and "Unknown."

Created Variables

CITY. Codes are "DC," "LA," and "NY."

SUBCITY. For New York, the categories are "Bronx," "Brooklyn," "Manhattan," "Queens," and "Staten Island." For Washington, the categories are "Northwest," "Northeast," "Southwest," Southeast," and "Downtown." For Los Angeles, the categories are names for 60 of Los Angeles's officially defined neighborhoods.

DAY, MONTH, and YEAR. Three separate numeric variables for the date of the arrest, created from the variable for arrest dates in the original datasets.

STATION. Coded 1 if the geographic coordinates of an arrest locate it in the immediate vicinity of a police station; coded 0 otherwise.

RACE. Combined race and ethnicity codes for use across cities. Codes are:

1 N-L White
2 Black
3 Latino
4 Asian
5 Other
6 Unknown

ZIP. The five-digit ZIP code where the offense took place.

INDEXCRIMES. The FBI's Part I index offenses are classified consistently across cities based on the information in PDARRESTCAT, PDCHARGE, and PDLEGALCAT.

01 Murder
02 Rape
03 Robbery
04 Aggravated Assault
05 Burglary
06 Larceny
07 Motor Vehicle Theft
08 Arson

OTHERCRIMES. Classified consistently across cities based on PDARRESTCAT, PDCHARGE, and PDLEGALCAT. Codes are:

01 Assault, Other
02 Burglary-Related
03 Conspiracy
04 Contempt of Court
05 Damage/Destroy Property
06 Disorderly Conduct
07 Domestic/Family Offenses
08 Drug-Related
09 DUI
10 Failure to Appear
11 False Report
12 Fraud/Forgery
13 Fugitive from Justice
14 Gambling Law Violation
15 Graffiti/Deface Property
16 Juvenile Offenses
17 Kidnapping
18 Lewd/Obscene Offenses
19 Liquor Law Violation
20 Loitering
21 Manslaughter

22 Parole/Probation Violation
23 Prostitution & Solicitation
24 Public Drinking
25 Reckless Endangerment
26 Sex Offenses
27 Stolen Property
28 Tax Law Violation
29 Theft, Petty
30 Theft of Services
31 Threats/Menacing
32 Traffic Violation
33 Trespassing
34 Unauthorized Use of Vehicle
35 Unlawful Imprisonment
36 Vagrancy
37 Outstanding Warrant
38 Weapons Charges
39 Unclassified City/St Laws
40 Flash Incarceration
41 Uncoded Felony
42 Uncoded Non-Felony

Coding Details and Issues

The Universe

The data were downloaded directly from the datasets posted online for Los Angeles, New York, and Washington. All the available arrest data were included through 2022.[1] The Los Angeles arrests cover 2010 through 2022. The New York arrests cover 2006 through 2022. The Washington arrests cover 2013 through 2022.

LawanyArrests.csv includes all arrests that met two criteria: They had information on the nature of the offense and occurred within the police departments' formal legal jurisdiction. This meant deleting arrests that were missing data for PDARRESTCAT and PDCHARGE and arrests outside the police departments' legal jurisdiction.

Washington did not have missing data on both PDARRESTCAT and PDCHARGE for any of its arrests. New York had missing data for 9,159 arrests, 0.2 percent of the total. Los Angeles was the outlier, with missing data on both variables for 101,537 arrests, 6.8 percent of the total.

Arrests outside the department's legal jurisdiction involved only a handful of cases for Washington, a few dozen for New York, and more than 200 for Los Angeles. The greater number in Los Angeles results from its unusual situation: Many cities with their own police departments are contiguous with Los Angeles. The jurisdictions involved have reached understandings about the circumstances under which police can cross borders to make arrests, but the arrests by the Los

Angeles Police Department outside its legal boundaries constitute only a fraction of all the arrests in the ZIP codes of jurisdictions neighboring Los Angeles.

PDID

For New York and Los Angeles, PDID is the police department's identification number (labeled ARREST_KEY in the original New York database and REPORTID in the original Los Angeles database). The database released by Washington has two identification numbers but each is more than 60 characters long, too unwieldy to use easily. I created a unique numerical PDID. The first four digits are the year of the arrest, followed by the number of the observation in the Washington database for each year. For example, Washington reported 32,512 arrests in 2013. PDID for Washington arrests in 2013 runs from 20131 to 201335512.

PDARRESTCAT and PDCHARGE

Each city had separate variables for categorizing the arrest (PDARRESTCAT) and the charge (PDCHARGE). Sometimes, the charge is worded the same as the arrest category, but usually the description of the charge contains more detail. In the original Washington database, PDCHARGE reached up to 154 characters. I shortened the longest charge to a maximum of 70 characters without losing the essentials.

Except for shortening some of the Washington PDCHARGE text, PDARRESTCAT and PDCHARGE are shown as they appear in the downloaded datasets, complete with abbreviations, misspellings, and variant wording within cities for the same arrest category or charge. Be aware of these issues when you are breaking out specific types of offense using PDARRESTCAT and PDCHARGE.

PDLEGALCAT

New York and Los Angeles had a variable that explicitly coded for felonies versus misdemeanors and other categories of lesser charges. (The variable labels were LAW_CAT_CD in New York and ARRESTTYPECODE in Los Angeles.) Washington did not have such a variable. I created two categories for Washington ("Felony" and "Lesser Charge") combining information from PDARRESTCAT and PDCHARGE. The distinction should be assumed to be less precise for the Washington arrests than for New York and Los Angeles arrests.

AGE

The Washington dataset includes only adult arrests, excluding arrests of persons younger than 18.

LATITUDE, LONGITUDE, and STATION

All three cities entered values for latitude and longitude that enable you to associate a place with each arrest. Most places are the location of the offense, but a complication must be kept in mind: Sometimes the police entered the location of the police station where the arrestee was booked, not the place where the offense occurred. This happened seldom in Los Angeles and Washington but often in New York. This significantly contaminates analyses of crimes' geographic locations.

To permit analysts to deal with the problem, I created the variable STATION, coded 1 for all arrests with values of LATITUDE and LONGITUDE that place an arrest in the immediate vicinity of a police station. "Immediate vicinity" was defined as any arrest with a combination of LATITUDE and LONGITUDE within 0.001 decimal degrees of the police station's Google Maps coordinates. I then examined the clustering of arrests near each police station to identify arrests that were outside the ±0.001 limit (in the United States, equivalent to roughly 400–500 feet) but highly likely to reflect arrests for which the GPS reading was taken near the police station (e.g., in the parking lot) rather than at the location of the offense. Arrests coded STATION = 1 amounted to only 1.2 percent of all arrests in Los Angeles, 2.4 percent in Washington, and 16.5 percent in New York. Such arrests should be excluded from many analyses of geographic patterns of crime at the ZIP code level.

ZIP

ZIP was created by reverse geocoding LATITUDE and LONGITUDE. The initial values were obtained from batch reverse geocoding performed by Texas A&M University GeoServices. While usually accurate, batch reverse geocoding produces errors caused by irregularities in the geographic boundaries of ZIP codes. (ZIP code boundaries routinely have small bumps and hollows.) Batch reverse geocoding also produces many missing values when the coordinates identify locations such as parkland or highways that do not have houses or businesses associated with ZIP codes.

The results from the batch reverse geocoding thus had to be subjected to detailed review. I filled in the missing cases with searches on Google Maps using the nearest residential or business area to ascertain the ZIP code. I sought to catch and correct errors with a two-step process. First, I

sorted arrests by ZIP and then by LATITUDE and LONGITUDE, checking for accuracy using Google Maps when the value of ZIP changed. I then repeated the process but sorting first by LATITUDE and LONGITUDE and then by ZIP. I did not keep track, but I estimate that I individually checked 3,000 to 4,000 ZIP values.

These efforts appear to have been successful. In a random sample of 300 arrests (100 for each city) for which STATION = 0, ZIP gave the correct ZIP code for 292 of them—an accuracy rate of 97.3 percent.[2] In the remaining cases, the value of ZIP referred to the adjacent ZIP code, with the arrest usually located just across the border from the correct ZIP code.

A note about Washington: Several ZIP codes in the area of the Mall and Capitol Hill are specific to a few government buildings. I have recoded ZIP to reflect the geographic ZIP code that best characterizes the arrest's location. ZIP code 20310 is recoded as 20010, 20543 as 20002, 20540 as 20003, 20535 as 20004, and 20039 as 20011. ZIP codes 20228, 20230, 20242, 20560, 20565, and 20585 are all coded as 20024.

RACE

RACE is a created variable that allows a common characterization of the arrestee's race across the three cities, but complications associated with distinguishing between race and ethnicity necessarily involve some imprecision. The coding of "Black" includes Latino blacks, and the coding of "Asian" includes Latino Asians. The coding of "N-L White" (meaning non-Latino white) includes some false positives—people who are racially white but would self-identify as ethnically Latino. You should thus think of the category "N-L White" as having the prefix "Overwhelmingly But Not Completely." A few comments about each of these issues follow.

The inclusion of Latino Asians in the "Asian" category is not important analytically. Based on the American Community Survey (ACS), fewer than 2 percent of self-identified Asians self-identify as Latino in any of the three cities.[3]

The inclusion of Latino blacks in the "Black" category is unimportant in Washington and Los Angeles. Again using ACS data, about 3 percent of self-identified blacks also identify as Latino in Washington. In Los Angeles, the comparable figure is about 2 percent.[4]

The comparable figure in New York City is larger, almost 9 percent, but still small enough to affect analyses by race only slightly. The New York dataset's advantage is that the raw number of arrests of Latino blacks is so large (444,551) that the arrest rates of non-Latino blacks, Latino

blacks, and non-black Latinos can be compared—an analysis that to my knowledge is not possible with any other publicly available dataset.

Unlike Asians and blacks, a large proportion of whites also identify as Latino, both in the general population and among arrestees in *LawanyArrests.csv*. In Washington, 33.2 percent of the 23,581 arrestees reported as white were also reported as Latino. In New York, 68.4 percent of the arrestees recorded as white were coded "White Hispanic" rather than simply "White." Add the fact that Latinos' arrest rates are about three times those of non-Latino whites, and it becomes essential to have a category in RACE that can be interpreted as non-Latino white.[5]

New York and Los Angeles pose no unusual problems. In both cities, the choice of how to classify an arrestee who might or might not be a self-identified Latino is forced upon the arresting officer by the lone variable for categorizing race and ethnicity—"White" versus "Hispanic White" in New York and "White" versus "Hispanic/Latin/Mexican" in Los Angeles. Hardly any arrests in either city were coded as "Unknown." There's no reason to think police judgments had high error rates, whether those judgments were based on the answer to a direct question or a Latino name, a Spanish accent (or fluent Spanish and broken English), or appearance. To the extent that errors occurred, they probably tilted toward incorrectly coding arrestees as non-Latino white—arrestees who had a Latino heritage but gave no visible clue through name, appearance, or Spanish-accented English.

The data for Washington posed a problem. Among arrestees coded "White" on PDRACE, 24.2 percent were coded "Unknown" on PDLATINO. How should they be treated when coding RACE?

One option was to code them as "Unknown," producing a sample of non-Latino white arrestees in Washington with a few false positives. But doing so would also have produced a downward bias in calculating the non-Latino white arrest rate if most of the arrestees coded as being of unknown ethnicity were actually non-Latino whites. And that is the most plausible expectation.

To see why, consider first how odd it is that 24 percent of arrestees have unknown ethnicity if we assume that the arresting officer asked the arrestee about his or her Latino ethnicity. (Few people are unable or unwilling to answer this question.) But in the real world, arresting officers do not always explicitly ask. When the time comes to fill out the paperwork, a code of "Unknown" for PDLATINO likely means that the arresting officer didn't ask the arrestee about ethnicity, and the

reason for not asking was that there was no reason to ask: The arrestee looked white, didn't have a Latino name, and spoke fluent English. If an arresting officer who didn't ask codes "Unknown" for PDLATINO, the odds are that the arrestee is non-Latino white.

I therefore treated Washington arrestees coded "White" in PDRACE and "Unknown" in PDLATINO as "N-L White" in RACE. Some Latino whites were surely misclassified as non-Latino whites in the Washington dataset, but the proportion is probably too small to have much statistical effect in analyses. To the extent it does have an effect, the misclassifications will produce an overestimate of the white crime rate, not an underestimate, and thereby minimize rather than exaggerate racial and ethnic differences in crime rates.

INDEXCRIMES and OTHERCRIMES

Each jurisdiction has its own way of entering arrest categories and charges, and arresting officers within the same city often use different abbreviations and punctuation. This made the process of creating INDEXCRIMES and OTHERCRIMES long and tedious.

The categories in INDEXCRIMES are reasonably straightforward. They are always coded as "Felony" in PDLEGALCAT, and the wording in PDCHARGE ordinarily makes clear the distinction between assault (often a misdemeanor) and aggravated assault (always a felony) and between murder (always a felony) and various categories of involuntary manslaughter. For the many ways of coding an arrest for theft, a PDLEGALCAT code of "Felony" usually distinguishes an arrest for larceny from an arrest for petty theft.

The codes for OTHERCRIMES required judgment. For example, New York made 74,000 arrests for "harassment," while Los Angeles and Washington almost never used that charge. Is the legal definition of "harassment" in New York closer to the definitions of "assault," "threats," or "menacing" in the other two cities? I chose to code "harassment" under the OTHERCRIMES category of "Threats/Menacing." As a general principle when analyzing OTHERCRIMES, I recommend that you incorporate the distinction between "Felony" and all the other codes in PDLEGALCAT to focus on more serious or less serious versions of a PDCHARGE code with ambiguous wording.

You can recover the coding methods used to create the categories for INDEXCRIMES and OTHERCRIMES by asking your statistical package to tabulate PDARRESTCAT or PDCHARGE for a given category. You will find that 126,058 arrests are uncoded. It sounds like a lot, but they amount to just 1.8 percent of arrests. Most of the uncoded arrests are for charges that occurred only a handful

of times. They could be coded for OTHERCRIMES given enough time and patience, but doing so would not substantively affect the results of analyses.

LawanyZips.csv

One of the most obvious uses of *LawanyArrests.csv* is to aggregate the arrest data by ZIP code and analyze the results against each ZIP code's demographic and socioeconomic characteristics. *LawanyZips.csv* enables you to do so for the basic variables. The data were downloaded from Social Explorer using the combined 2017–21 ACS. The dataset consists of 307 lines, one for each ZIP code represented in *LawanyArrests.csv.* The variables are as follows.

ZIP. Five-digit US Postal Service ZIP code.

TOTALPOP. Total population of the ZIP code.

HH. Number of households in the ZIP code.

PNLWHITE. Percentage of non-Latino whites.

PBLACK. Percentage of blacks. Includes black Latinos, consistent with the coding of race in *LawanyArrests.csv*.

PASIAN. Percentage of Asians. Includes Asian Latinos, consistent with the coding of race in *LawanyArrests.csv*.

PLATINO. Percentage of all other Latinos (white Latinos, American Indian Latinos, Hawaiian and Pacific Islander Latinos, Latinos of other races, and mixed-race Latinos).

PNLOTHER. Percentage of all other non-Latinos.

PHSMINUS. Percentage of persons age 25 and older with only a high school diploma or less.

PBAPLUS. Percentage of persons age 25 and older with a bachelor's or higher degree.

PDROPOUT. Percentage of persons age 16–19 who are not attending school.

MEDHHINC. Median household income expressed in 2021 dollars.

PHHSSI. Percentage of households receiving Supplemental Security Income.

PHHPUBASS. Percentage of households receiving public assistance income.

PFAMSINPOV. Percentage of families below the poverty level.

PMARRCOUPLES. Percentage of households headed by a married couple.

PFEMALEHHNSP. Percentage of households headed by a woman with no spouse present.

PCHILD1PARENT. Percentage of households with a child under 18 headed by a single parent.

Age Variables

Analyses of crime based on the arrestee's age may focus on a variety of age groups, so *LawanyZips.csv* has a variety of age variables that permit you to define the denominator of the percentages you want to calculate. Following the Census Bureau's definition, the prefixes "under" and "over" should be interpreted as *not* including the subsequent number. For example, OVER25 means age 26 and older. Age variables such as ALL2529 refer to all persons age 25–29, including the lowest and highest years. The age variables are OVER15, OVER18, OVER25, OVER55, UNDER35, UNDER40, UNDER45, ALL1014, ALL1517, ALL1819, ALL2021, ALL2224, ALL2529, FEMALES1834, and MALES1834.

---

[1] In 2022, the Washington Open Data website substituted an "Adult Arrests" file covering 2013–21 for the earlier year-by-year arrest databases. For unknown reasons, that file does not show any arrests for narcotics or kidnapping for 2013–17. All the other arrest numbers are identical to those in the earlier files. I have used the known numbers for narcotics and kidnapping from the earlier files.

[2] I used a random number generator for arrests with STATION = 0 and chose the 100 smallest random numbers in each city for the sample. The number of correct ZIP codes in New York, Los Angeles, and Washington, DC, were 96, 97, and 99, respectively.

[3] In the combined 2014–18 American Community Survey (ACS), the proportion of self-identified Asians who also self-identified as Latino in Los Angeles, New York City, and Washington were 1.3 percent, 0.9 percent, and 1.6 percent, respectively. (Author's analysis.)

[4] In the combined 2014–18 ACS, the proportion of self-identified blacks who also self-identified as Latino in Los Angeles, New York City, and Washington were 1.8 percent, 8.7 percent, and 3.2 percent, respectively. (Author's analysis.)

[5] Charles Murray, *Facing Reality: Two Truths About Race in America* (New York: Encounter Books, 2021), 50–52.