Center for Policing Equity Data User Guide

Introduction

The data discussed in this use guide can be accessed at the following link: [Center for Policing Equity Data](#).

This user guide will provide an overview of 268 datasets compiled by the Center for Policing Equity related to police activity in American communities. Each folder of data in this collection corresponds to one of 12 metropolitan cities in the United States, including Los Angeles, Boston and Seattle.

In each city's folder, there are data files with information about that city's police department encounters over time. For the purposes of this guide, the term "police encounters" can be understood as any instance where a police officer from a city's police department has an interaction with a civilian. In most cases, these police encounter files are given as a single .csv file for each department. There is not a standardized data collection method for each American police department to track its encounters data, so the information contained in each of these files can range from traffic stops to shootings.

Each city's folder also contains standardized demographic information about that community taken from the U.S. Census' American Community Survey (ACS) 5-year estimates. The U.S. Census Bureau collects demographic and lifestyle data related to every United States community for the ACS and publishes its conclusions both annually and once every five years. Each of these folders also contains a sub-folder of files with geometric information about each police department's jurisdiction. The geometric data in these files maps police districts, divisions, precincts and/or sectors onto the physical boundaries of a given city. District is the most prevalent police jurisdiction mapping unit in these files.

By analyzing Census demographic data against police activity logs, those using this dataset can draw conclusions about racial disparities present in police departments across the United States. They can also map police activity in certain regions of a city against the demographic information collected about that area. The Center for Policing Equity looks to utilize this Census and police department deployment data to analyze factors that drive racial disparities in policing. The ultimate goal of this collection is to (1) analyze its contents and (2) inform police agencies of improvements they can make based on any racial disparities contained in its datasets, according to the center's website.

Table of Contents

Downloading the Data

These data files are freely available on Kaggle, a data science competition platform and online community of data scientists. Users must create a free account on Kaggle to access the data download.

When logged into Kaggle, users can click the "Download" button to access all 268 data files in a downloadable ZIP file. The data downloads to the user's computer as a file named "archive.zip." Users should double-click that ZIP file to "unzip" the compressed content and access it. Once the datasets are downloaded and unzipped, the user should be greeted with 13 folders and one variable descriptions file, as pictured below:

Name	Date Modified	Size	Kind
ACS_variable_descriptions.csv	Oct 4, 2019 at 3:30 AM	30 KB	CSV Document
> cpe-data	Today at 2:29 PM	--	Folder
> Dept_11-00091	Today at 2:29 PM	--	Folder
> Dept_23-00089	Today at 2:29 PM	--	Folder
> Dept_24-00013	Today at 2:29 PM	--	Folder
> Dept_24-00098	Today at 2:29 PM	--	Folder
> Dept_35-00016	Today at 2:29 PM	--	Folder
> Dept_35-00103	Today at 2:29 PM	--	Folder
> Dept_37-00027	Today at 2:29 PM	--	Folder
> Dept_37-00049	Today at 2:29 PM	--	Folder
> Dept_49-00009	Today at 2:29 PM	--	Folder
> Dept_49-00033	Today at 2:29 PM	--	Folder
> Dept_49-00035	Today at 2:29 PM	--	Folder
> Dept_49-00081	Today at 2:29 PM	--	Folder

The user should move these files to a directory of their choice to open them in a programming language, examine them and utilize them for analysis.

About the File Types

- CSV
  - There are 186 CSV files in this dataset.
  - Comma Separated Value files are versatile. They are among the most common file types that professionals work with when analyzing or displaying data.

- ○ A CSV file stores tabular data in plain text, and each line of the file typically represents one data record.
  - ○ CSVs can be imported into Python, R, SQL, Excel and most other data analysis software programs.
- SHP
  - ○ There are 14 shapefiles in this dataset.
  - ○ Shapefiles store geometric information about geographic features of a certain area. They are often used for mapping a given area.
  - ○ Geographic features in a shapefile can be represented by points, lines, or polygons.
  - ○ Shapefiles can be opened, visualized and manipulated using packages such as geopandas in Python and sf in R.
  - ○ There are three main components that must be included for shapefiles to be valid:
    - ■ SHP files contain the feature's geometry.
    - ■ SHX files contain the index of a feature's geometry.
    - ■ DBF files store a feature's attribute information.
- XML
  - ○ There are 6 xml files in this dataset.
  - ○ Extensible Markup Language describes the text in a document.

Data Organization

There are 12 police departments included in this dataset. Each folder of data corresponds to a police department in a major metropolitan area in the United States. Here is what city is associated with each folder:

Folder Name	Jurisdiction
Dept_49-00033	Los Angeles, California
Dept_49-00035	Oakland, California
Dept_49-00081	San Francisco, California
Dept_23-00089	Indianapolis, Indiana
Dept_37-00049	Dallas, Texas
Dept_37-00027	Austin, Texas
Dept_11-00091	Boston, Massachusetts

| Dept_35-00103	Charlotte, North Carolina
Dept_24-00013	Minneapolis, Minnesota
Dept_24-00098	St. Paul, Minnesota
Dept_35-00016	Orlando, Florida
Dept_49-00009	Seattle, Washington

Each department folder assigns a unique ID to the city it pertains to. For example, Seattle's unique identifier is "49-00009." Each sub-folder of data within the Seattle folder also has this identifier of "49-00009" to signify its connection to the Seattle metropolitan area and/or the Seattle Police Department.

Within each city's police department folder, there are two sub-folders of data and one .csv file.
- The folder labeled with the department's unique ID and "ACS_data" contains data from the Census bureau's American Community Survey.
- The folder labeled with the department's unique ID and "Shapefiles" contains the geometric data needed to map the police department's districts, divisions and/or sectors.
  - Districts, divisions and sectors are all different ways to refer to a given police department's geographical jurisdiction.
  - Think of these like ZIP codes that pertain to each police department.
- The single .csv file is data submitted by the department that documents some version of their civilian encounters data.

ACS data organization

Here are the sub-folders of ACS and shapefile data located in the city of Oakland, California's folder:

📁 49-00035_ACS_data

📁 49-00035_Shapefiles

📄 49-00035_Incidents_2016.csv
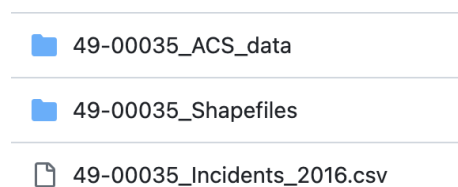
The "ACS_data" folder contains raw American Community Survey data from seven categories in .csv format. Each of these folders contains the ACS data related to each metro area's:
- Educational attainment over 25
- Overall educational attainment

- Employment
- Income
- Owner-occupied housing
- Poverty and
- Race/sex/age

The American Community Survey collects this information for every metropolitan area across the United States. The Center for Policing Equity pulled this data directly from the U.S. Census' ACS website.

Within the ACS data folder, there are always seven folders that correspond to the Census-designated metro area the police department in question resides in. For reference, here is what the ACS data folders look like for Oakland, designated by code "49-00035."

📁 49-00035_ACS_education-attainment-over-25

📁 49-00035_ACS_education-attainment

📁 49-00035_ACS_employment

📁 49-00035_ACS_income

📁 49-00035_ACS_owner-occupied-housing

📁 49-00035_ACS_poverty

📁 49-00035_ACS_race-sex-age

Each folder is structured the same way. The two files in each folder are named after the Census data category that corresponds to the folder title. For example, the ACS education attainment data can be universally accessed via the Census Bureau's unique ID "B15003". Therefore, the data files in the "ACS_education-attainment-over-25" folder are titled "ACS_15_5YR_B15003_xxx".

📄 ACS_15_5YR_B15003_metadata.csv

📄 ACS_15_5YR_B15003_with_ann.csv

Each of the seven ACS data folders contain one "metadata" file and one "with_ann" file. The file titled "with_ann" is the raw, Census-recorded data about the jurisdiction. The "metadata" file contains a data dictionary that defines each column name in the "with_ann" file.

For example, below is a subset of the "metadata" .csv file in the educational attainment over 25 folder. It contains definitions for each column name in the folder's "with_ann" .csv file. This file structure is uniform for every department's ACS data folder.

The city of Oakland's "metadata" file for educational attainment over 25:

HD01_VD03	Estimate; Total: – Nursery school
HD02_VD03	Margin of Error; Total: – Nursery school
HD01_VD04	Estimate; Total: – Kindergarten
HD02_VD04	Margin of Error; Total: – Kindergarten
HD01_VD05	Estimate; Total: – 1st grade
HD02_VD05	Margin of Error; Total: – 1st grade
HD01_VD06	Estimate; Total: – 2nd grade
HD02_VD06	Margin of Error; Total: – 2nd grade
HD01_VD07	Estimate; Total: – 3rd grade
HD02_VD07	Margin of Error; Total: – 3rd grade

The city of Oakland's "with_ann" file for educational attainment over 25:

HD01_VD02	HD02_VD02	HD01_VD03	HD02_VD03	HD01_VD04	HD02_VD04	HD01_VD05
Estimate; Total: – No schooling completed	Margin of Error; Total: – No schooling completed	Estimate; Total: – Nursery school	Margin of Error; Total: – Nursery school	Estimate; Total: – Kindergarten	Margin of Error; Total: – Kindergarten	Estimate; Total: – 1st grade
0	12	0	12	0	12	0
0	12	0	12	0	12	0
86	92	0	17	0	17	0
36	46	0	12	0	12	0
6	9	0	12	0	12	0
44	35	0	12	0	12	24
67	51	0	12	0	12	0
41	33	0	12	0	12	0
27	31	0	12	0	12	0
65	58	0	17	0	17	0

The first row of this "with_ann" file — and every other "with_ann" file — contains the same column name definition that users of this data can find in each "metadata" .csv file. But users would want to drop this row from a dataframe before using it for any substantive analysis, as it is a descriptor and does not contain any data about the jurisdiction to parse.

Shapefile data organization

Each department submitted slightly different data to the Center for Policing Equity depicting the map of their jurisdictions. As discussed previously in this guide, shapefiles need three components to be valid data sources for analysis: a .shp file, a .dbf file and a .shx file. Some departments included additional maps of their jurisdictions, such as .pdf illustrations. But the primary data that can be utilized for analysis, manipulation and visualization comes in the form of .shp files.

Most data folders, such as the Minneapolis, Minnesota, one below, contain more data files than the three that are required to open a valid shapefile. Minneapolis' data contains one .shp, one .dbf, one .shx, one .prj, one .cpg and one .xml file. These additional file types are common to see when handling shapefile data.

- 24-00013_Shapefiles
    - Minneapolis_Police_Precincts.cpg
    - Minneapolis_Police_Precincts.dbf
    - Minneapolis_Police_Precincts.prj
    - Minneapolis_Police_Precincts.shp
    - Minneapolis_Police_Precincts.shx
    - Minneapolis_Police_Precincts.xml

Many department folders also contain various shapefiles pertaining to police districts, divisions, precincts, zones and/or sectors. Each of these describes a different way of referring to a department's jurisdiction area. Users can examine each police department's individual website to determine the most appropriate unit of geographic measurement to use when analyzing department-level encounters data.

Department encounters data organization

The data each police department submitted to the Center for Policing Equity related to its civilian encounters, most often, comes in the form of one .csv file located outside of the ACS and shapefile data folders. See the below example of the Minneapolis folder of data:

- 24-00013_ACS_data
- 24-00013_Shapefiles
- 24-00013_UOF_2008-2017_prepped.csv

Understanding the Data

The Center for Policing Equity organized the files in this dataset for a competition it hosted on Kaggle in 2018. The files contained in this dataset are selected from a larger database of police encounters and demographic data maintained in the organization's National Justice Database. On

its website, the Center for Policing Equity touts the National Justice Database as the "the first and largest collection of standardized police behavioral data." Users can access this broader database and the conclusions drawn from it so far at this link.

For the Kaggle competition, users were tasked with finding factors that drive racial disparities in policing by analyzing these 268 demographic and deployment data files. The organization offered a total of $15,000 in prizes to users who participated in the competition and whose analysis revealed relevant conclusions about policing and racial disparities.

Entries were judged based on the following criteria, which users of this guide can also find here:
- *Accuracy:* Does the solution provide reliable and accurate analysis? How well does it match census-level demographics to police deployment areas?
- *Approachability:* The best solutions should use best coding practices and have useful comments. Plots and graphs should be self-explanatory. CPE might use your work to explain to stakeholders where to take action, so the results of your solution should be developed for an audience of law enforcement professionals and public officials.

In crafting this user guide, group members researched some participants' projects to understand how people have used the Center for Policing Equity's data in the past. If users are interested in seeing previous use cases for data analysis with these files, they can parse through many of the final reports participants generated for the competition at this link. Some of the best entries group members reviewed are linked below for convenience:

- [Geoprocessing Census and Center for Policing Equity Data by user Jared Knowles](#)
- [Solution Workflow for Science of Policing Equity by user Shivam Bansal](#)
- [Center for Policing Equity with Census Tracts Analysis by user Bukun](#)

Data Limitations

With the amount of data provided in the Center for Policing Equity's collection, there are several things to note if users wish to properly utilize each file.

Date Ranges

Most of the department-level police encounter data collected by the Center for Policing Equity is updated through 2018. Some department-level datasets only have data through 2016 or 2017. Since the Center for Policing Equity last updated the competition page where users can download the data from more than five years ago, none of the data detailed in this guide should be used to draw modern conclusions about police use of force frequency or racial disparities.

The data provides accurate historical context for examining disparities in policing throughout the 20th century up to 2017 and 2018 and should be used in this context. The ACS and shapefile data contained in each folder is also only updated through 2018. There has been another ACS 5-year survey since 2018, so some of the demographic information about different U.S. metropolitan areas may have changed during that time period. With that said, the Census-level data contained in these files is also not the most current demographic data that exists about any given community, but serves as a good proxy for this information.

Data Collection Inconsistencies

If a user wants to draw a conclusion about this data from an individual police department, it could be helpful to reach out to the department to better understand its data collection methods. As discussed earlier in this guide, police departments across the United States do not have a standardized tracking system for use of force or civilian encounters. That means that each department dataset contained in the Center for Policing Equity's collection could contain slightly different data. For example, the St. Paul Police Department's data contains more than 700,000 police encounter incidents from 2001 to 2017. These incidents, based on analysis of the data and research into a handful of rows, pertain to police traffic stops. In contrast, the Orlando Police Department data includes 54 police use of force incidents from 2009 to 2017. These same discrepancies exist for many other department-level data files.

No police departments in the country have collected and submitted the exact same, standardized encounter data. Before comparing and contrasting departments, users should make sure they have an understanding of what the data shows and may need to standardize it (i.e., making sure that the data they compare between St. Paul and Orlando pertains to police killings rather than one pertaining to killing and one pertaining to traffic stops). Department-level data could pertain to police use of force, police shootings, police traffic stops and/or police killings. Users should keep this in mind as they explore the data for their own analysis and work to understand the differences in how the data has been collected from its sources.

Competitors in the Center for Policing Equity's Kaggle competition also had various ways of standardizing information from each police department to analyze and compare their data. These competition resources could also be useful for users to examine as they look into standardization methods and work to better understand how various departments collect their police encounters data.

Lack of Cleanliness

The sheer volume of data contained in this collection means there will be some messy properties beyond the date and department-level qualifications listed above.

Several columns across each department-level dataset are listed as "NA" values or completely empty. For instances where there are columns that contain very few values, users should omit this data from their analysis. Some department datasets have entire columns, such as age or race, that have more than 80% "NA" values. It would be unethical for users to engage with these largely empty columns in any of their department-level or broader analysis using this data. Data collected in columns with large numbers of "NA" values should instead be primarily used to contextualize individual rows of data for which these values do exist.

Census and Department-Level Data Discrepancies

Each department's ACS dataset pertains to a specific Census-designated metropolitan area in the United States. These Census-designated metro areas are usually made up of data from 1-3 major cities and their surrounding suburbs. In contrast, the police department-level data submitted for this project pertains to one single city and a police department's borders within it.

In many cases, the demographic information of a community does not change much within its Census-designated metro area. But this is still an important mismatch between the ACS and department-level data that users should keep in mind.

Take, for example, the Census-designated metro area for Oakland, California, which is defined as "San Francisco - Oakland - Fremont." Users looking to draw conclusions about the Oakland Police Department's encounters have ACS data that includes demographic survey results from San Francisco. That paints a vastly different picture of Oakland's socioeconomic and race/sex/age background. Users should make this qualification in their analyses. It may also be very useful for users of the Center for Policing Equity's data to refer to the data dictionary for Census designated metro areas to better understand what ACS jurisdictions they are working with as they analyze department-level data.

The ACS data is still the best proxy users of this police encounter data have for demographic information pertaining to a specific city. But this ACS data typically encompasses a larger area than one individual city's jurisdiction.

FAQ

- Where was this data collection found?

The group located this collection of data to document on Kaggle after searching for large databases that contained between 150 and 300 files to parse through. The Center for Policing Equity provided this collection of data for competitive use on Kaggle.

- What is the Center for Policing Equity?

The Center for Policing Equity is a research center based at Yale University that collects data to help law enforcement agencies improve their relationships with their communities. The nonprofit organization was founded in 2007 at the University of California, Los Angeles. The Center for Policing Equity looks to eliminate racism in public safety and make policing "less racist, less deadly and less omnipresent," according to the organization's website. Users can learn more about the Center for Policing Equity, its data collection and its goals by visiting their website.

- How were these data sets collected?

The Center for Policing Equity collected the ACS datasets directly from the U.S. Census website. The organization submitted records requests to police departments across the country to obtain the shapefiles of department geographies and each department-level police encounter dataset.

- How can I use this data?

There are a variety of ways this data can help users draw conclusions about 21st century policing in American communities. Users can overlay demographic information on a map depicting where police encounters have occurred most frequently in a department's jurisdiction. That can help them understand if over-policing is happening and, if so, what communities it impacts the most. Users can also utilize department-level data to draw conclusions about a city's police use of force, traffic stops, shootings and other information based on the data submitted by each department. After standardizing these individual datasets, users can also compare policing and racial equity trends across multiple American police departments. These are a few examples of the many ways in which users can utilize this data.

Sample Analysis and Visualization

Visualization #1

Different types of crimes reported to the Police Department in Oakland, California
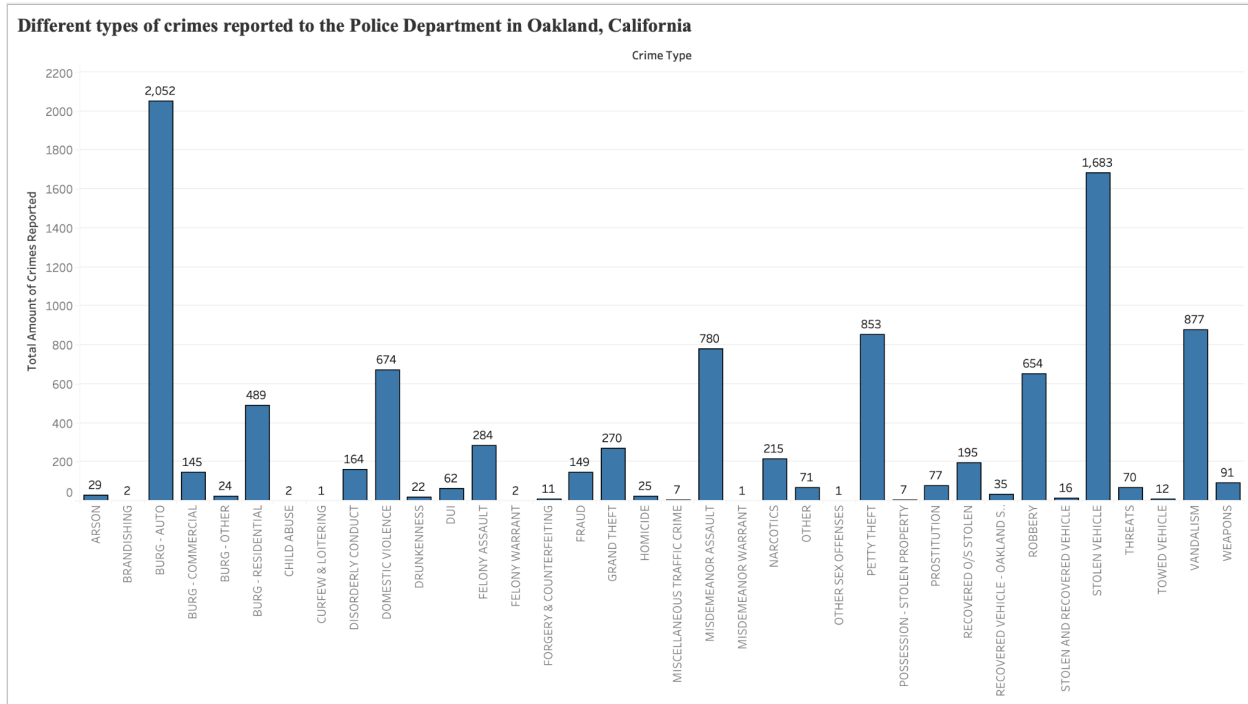
Research Question: What are the top 3 crimes reported in Oakland, California, according to the police department?

This visualization examines the number and type of crime reported to the Oakland Police Department in Oakland, California, in 2016 and 2017. The visualization came from the incident data file in the Dept 49-00035, which corresponds to the Oakland Police Department.

In creating this graph, users needed to manipulate the underlying data structure to ensure it was clear and accurate. At first, some labels on the visualization corresponded to "null" or "NA" values. They also corresponded to incorrect columns in the dataset, such as "incident type." This layout in the graph made the visualization very confusing and unclear. To solve this issue, the data structure needed to change to omit the null values. The labels related to irrelevant data categories were also dropped from the visualization.

This visualization can be replicated across other departmental datasets to understand more about the police response to crime in any given community. From this visualization, viewers can glean that the top three crimes reported were auto burglary, vehicle theft and vandalism. Users can also conclude how prevalent other types of crime, such as misdemeanor warrants or loitering, were in Oakland during this time.

If users wanted to further examine this visualization's topic, they could obtain more current data from the Oakland Police Department. The most recent data on the Oakland Police Department in the Center for Policing Equity's collection is from 2017. Even without this updated data, users

can use this visualization as the baseline for additional analysis. Users of this dataset could overlay race and demographic data on the Oakland crime report data to determine where police response was most frequent. They can also compare a graph like this to ones for other metropolitan areas to gain a sense of whether Oakland's vehicle crime rates are particularly high for a metro area or follow national trends.
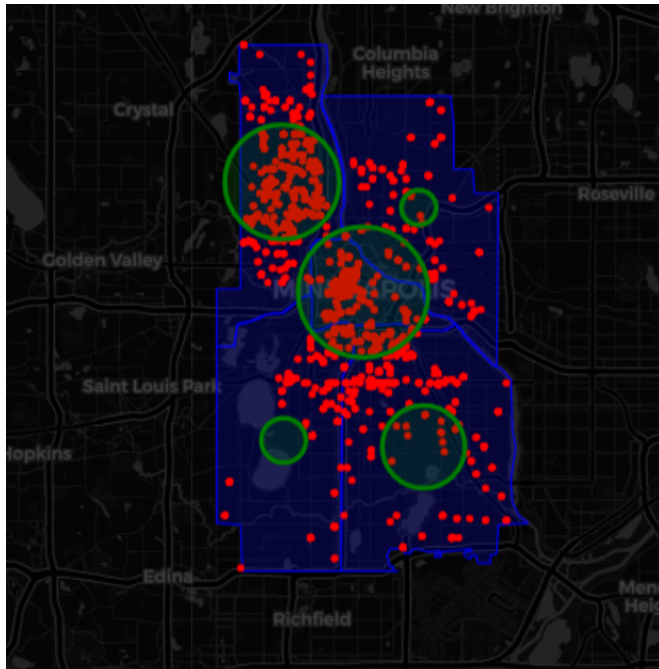
Visualization #2

Gender Heat Density Map
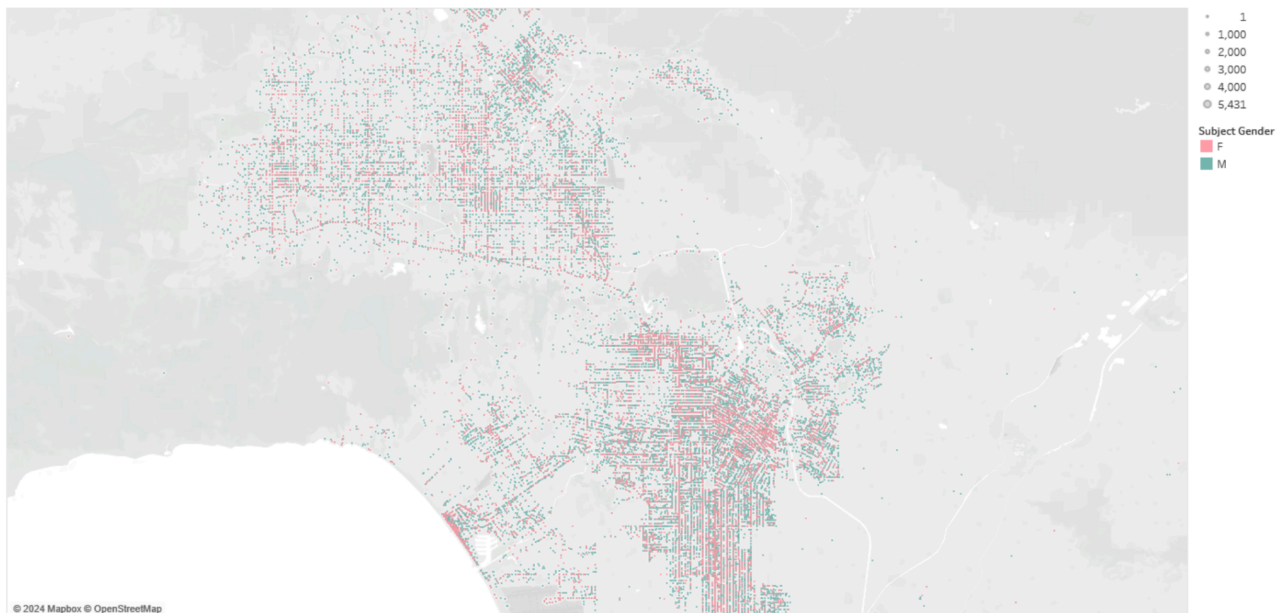
Research Question: How does the distribution of male and female populations vary geographically and affect crime across Los Angeles?

This visualization shows the distribution of police encounters by gender in Los Angeles, California using data from the folder Dept_49-00033. The dataset for this visualization needed to be slightly manipulated to categorize the data and quantify the total crime count per gender per geographic area. The visualization highlights how many police encounters occurred for each gender in metropolitan Los Angeles. The dot size indicates how many encounters happened in that radius and the color indicates the gender of the person impacted by the police encounter. The heatmap indicates that police encounters are most prevalent in the southeastern area of Los Angeles. For example, areas with a high density of the pink color show a higher concentration of female subjects, while those with a lot of blue suggest a predominance of male subjects. The map can help in identifying patterns of gender distribution in where police encounters occur most.
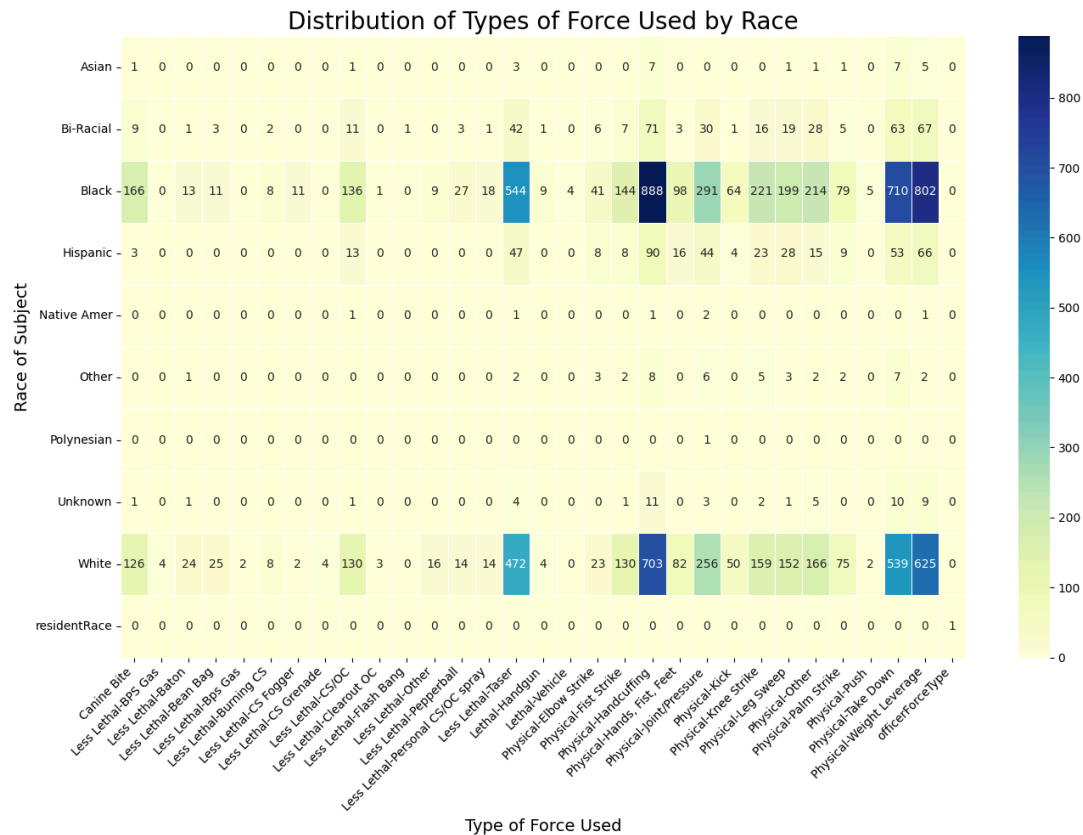
Visualization #3

Research Question: To what extent do socio-economic factors, such as income and education, impact the frequency of incidents within Minneapolis?
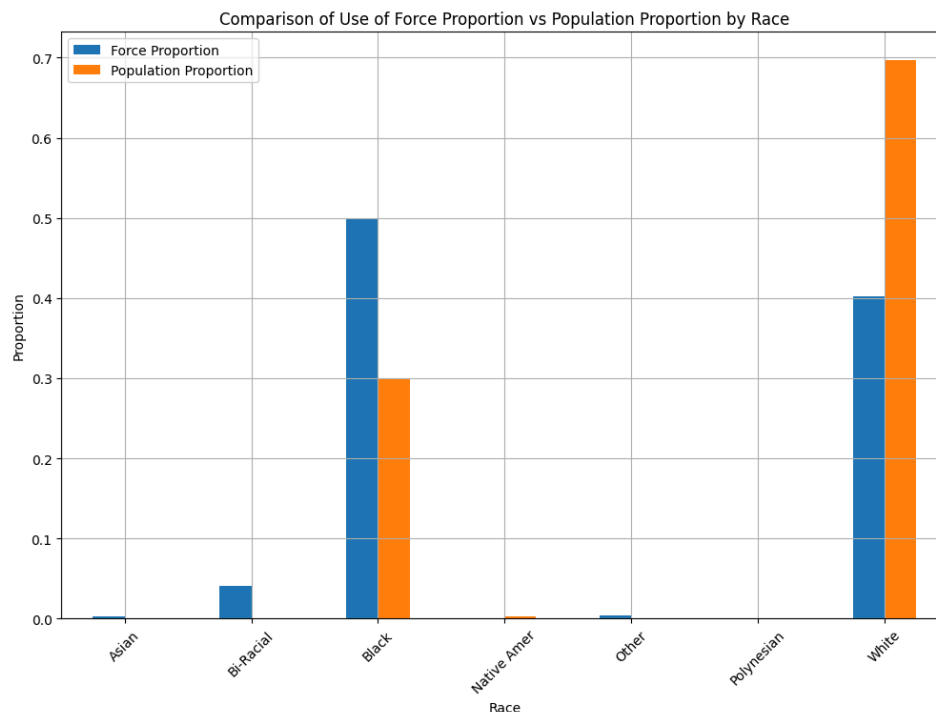
The dataset also allows for comparisons across various departments and cities and supports geographical analysis within city districts using shapefiles. Various visualizations can be generated to identify patterns or 'hotspots' of incidents, as demonstrated in the visualization based on the Dept_24-00013 dataset from the Minneapolis police department spanning from 2012 to 2015.

The map visualization derived from the Minneapolis dataset highlights aggregated total incidents (represented by green circles), use of force incidents (represented by red dots), and departmental districts (represented by blue polygons). As the visualization focuses on a single department, no data manipulation was necessary for standardization. With the combination of the visualization produced from shapefiles and the bar graphs or heatmaps to analyze the ACS data, it is possible to look into the effects of various variables, such as socio-economic facts, on the frequency of incidents on a geographical level.

To delve deeper for further potential research, users can conduct cross-departmental comparisons by creating similar visuals for other departments. Furthermore, exploring the dataset averages or introducing additional variables can provide further insights.

Visualization #4

Distribution of Types of Force Used by Race

Comparison of Use of Force Proportion vs Population Proportion by Race

Research Question: In Indianapolis, how does the use of force by law enforcement correlate with the racial composition of a population, and are there disparities in the proportion of force used relative to the population distribution of different racial groups?

The dataset utilized for these two visualizations is found in the folder corresponding to Dept_23-00089. This contains police encounters and demographic information related to the Indianapolis Police Department in Indianapolis, Indiana. The dataset was not directly manipulated to perform analysis, but it was crucial identifying and replacing column names from code to common expression was necessary to provide better visualizations.

The first visualization presents a heatmap detailing the distribution of types of force used by law enforcement across different racial categories. The heatmap indicates a higher frequency of certain forces being used on Black civilians, such as tasing and physical restraint. In contrast, other racial groups show fewer incidents across all types of force used.

The second graph is a bar chart comparing the proportion of force used against the population proportion by race. This chart reveals a larger discrepancy in the Black category, where the proportion of force used surpasses the proportion of Indianapolis' population that identifies as Black population. The proportions are more closely aligned for use of force data pertaining to white civilians.

These charts indicate variances in the application of force by race, with the Indianapolis Police Department's data suggesting a disproportionate use of force on Black civilians — especially when examining this proportion in relation to their demographic representation.

To further research this topic, users can replicate the analysis for other police departments across the country. It would also be helpful to include other types of data in these visualizations, such as the reason for use of force by a police officer. Some department-level datasets contain this information for analysis. There are many more complex and comprehensive conclusions that can be drawn when examining the correlation between the types of force used and demographics.

Version History

Version	Month and Year
1.0	January 2024
2.0	May 2024
3.0	November 2024