County Buddy: A Companion Dataset for Socioeconomic Data Analysis and Exploration of U.S. Datasets

Colin Vu[ab], Leila Baniassad[a], Clio Andris[ab]

[a]College of Computing, Georgia Institute of Technology, Atlanta, Georgia, USA

[b]Corresponding authors: colinhvu@gmail.com, clio@gatech.edu

Keywords: Demographics, Geospatial Data, Incarceration, Prisons, Universities, Colleges, Students, Military, Native Americans, Geographic Information Systems (GIS), Data Visualization, Geospatial Mapping,

Introduction

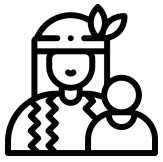
County Buddy is a new dataset that represents information on the presence, number of, and institution names of special populations at the U.S. county and census tract levels. These *distinguished populations* include the **incarcerated population**, **college student population**, **military population**, and **Native American population**.

The objective of this dataset is to provide geographic and demographic context to help explain variations in socio-economic indicators (such as life expectancy, household income, educational attainment, etc.). The data on distinguished populations can help provide a potential 'explainer' for why these values (e.g. high marriage rates) appear or provide a landscape to put the values into context.

Incarcerated Population

Student Population

Military Population



Native American Population

Although these four datasets are publicly available through the U.S. Census, National Center for Education Statistics (NCES), Department of Transportation, and Department of the Interior, we gathered them here because they can be difficult to merge, geolocate, and filter. These four populations were chosen as pilot populations that the project's researchers noted as overlapping with geographic anomalies they found on choropleth maps. Incarcerated populations accounted for anomalous concentrations of Black individuals in Pennsylvania and New York tracts. Native American populations were associated with the lowest life expectancies in U.S. counties, specifically in South Dakota. University populations correlated with higher income counties in Ohio, Virginia, and other locations. Military populations accounted for high in-migration in places such as Ft. Bragg, North Carolina.

The four variables, while limited, may be useful for researchers who are unable to aggregate and clean data from multiple sources, and automatically manipulate cut-off values. We also provide the names of each institution (e.g., Fort Drumm), which facilitates connections with other information and datasets.

The data can be used by researchers and developers who are studying causality and correlation in health and livelihood outcomes, as they can be easily integrated into statistical models (e.g., in R or STATA). For example, a researcher conducting statistical regression analysis may identify a census tract exhibiting an unusually high infection rate of a disease and use County Buddy to find that the value correlates with a high university or a prison population within the tract.

The data can also be used by developers looking to enhance their maps of U.S. thematic data (life expectancy, percent smokers, car accidents, number of people on Medicare) with tooltips and extra features. For example, a map designer may want to help users investigate why they see counties with a higher average income or higher birth rate than surrounding areas. By integrating County Buddy into their tooltips or adding County Buddy points on the map, the user can explore and learn that this area has a military base nearby.

Data Description and Methods

County Buddy is provided in the form of two CSV-format spreadsheets: one containing county-level data and the other containing census tract-level data. Each record corresponds to a distinct county or census tract. Within each record, demographic data categories (such as university student population, military population, etc.) are excluded or "zeroed out" if they do not meet the criteria defined in the default threshold we use. This threshold is defined as data where the percentage of the total population represented by a given demographic exceeds some number of standard deviations above the mean. For county data, two standard deviations are used. For tract data, three standard deviations are used as the data is more granular. A data snippet of the County-Level dataset is provided below (Table 3).

These thresholds can be considered subjective. Users can run code to aggregate these datasets and choose their own threshold values (e.g., counties with at least 10,000 students or tracts with at least 50% Native American population) (see below).

Except where noted, all data are from 2020. Prison populations use 2017 figures, and institutional population counts come from the 2020 Census (U.S. Census Bureau, 2020a). We apply category-specific rules to identify institutions in each region:

- **Prisons, Military Bases**: include any facility whose boundaries fall within the region.
- **Native American Reservations**: include any reservation overlapping the region.
- **Universities**: include only active non-profit colleges, excluding junior colleges, with more than 1,000 students. In the university dataset, each campus is represented by a single point; we list all points inside the region and, for tract-level data, also record any points within 1 km of each tract in a separate column.

Further details on these criteria appear in the Data Dictionary (Table 4). Below is a summary of the data holdings at the county level (Table 1) and at the tract level (Table 2).

Table 1: County-Level Summary Statistics

Category	Total Number of Institutions	Total Population of Institution	Outlier Threshold (% of Total Pop.)	Counties Affected
Prisons	2,424	1,978,489	8.9%	149
Universities	1,289	2,794,201	4.4%	128
Military Institutions	743	328,615	1.4%	36
Native American Reservations	611	3,745,005	17.8%	72

Table 2: Tract-Level Summary Statistics

Category	Total Number of Institutions	Total Population of Institution	Outlier Threshold (% of Total Pop.)	Tracts Affected
Prisons	2,424	1,978,489	17.5%	886
Universities	1,289	2,794,201	19.2%	1055
Military Institutions	743	328,615	8.2%	197
Native American Reservations	611	3,745,005	15.1%	659

Table 3: Data Snippet of County-Level Incarceration Data

FIPS	State	County	Total Pop.	Prison Pop.	% Prison	Institutions	...
12043	Florida	Glades County	12,126	1,640	13.50%	Florida Environmental Institute; Moore Haven Correctional Facility; Glades County Detention Center	...
12045	Florida	Gulf County	14,192	0	0.00%		...
12047	Florida	Hamilton County	14,004	2,323	16.50%	Hamilton Work Camp; Panther Success Center; Hamilton Correctional Institution; Hamilton Correctional Institution Annex; Hamilton County Jail	...

Full code resources are available for users to update the dataset links with more recent data (or past data) and change the thresholds that we determined for this original dataset. This code is written in Python and is part of a Jupyter Notebook, available on GitHub (Vu, 2025).

Usage Example

Below is an example implementation of County Buddy in the Exploropleth suite of tools, created by Arpit Narechania (Narechania, Endert, and Andris, 2025). The map's county-level data has been joined with County Buddy's county-level data using their federal identification processing standards (FIPS) codes. County Buddy's data is shown in the tooltip (i.e., White Earth Reservation is shown with the Percent Native American given as well). In this example only Native American populations are shown in the tooltip because there were not substantial university, incarcerated, or military populations residing in that county.

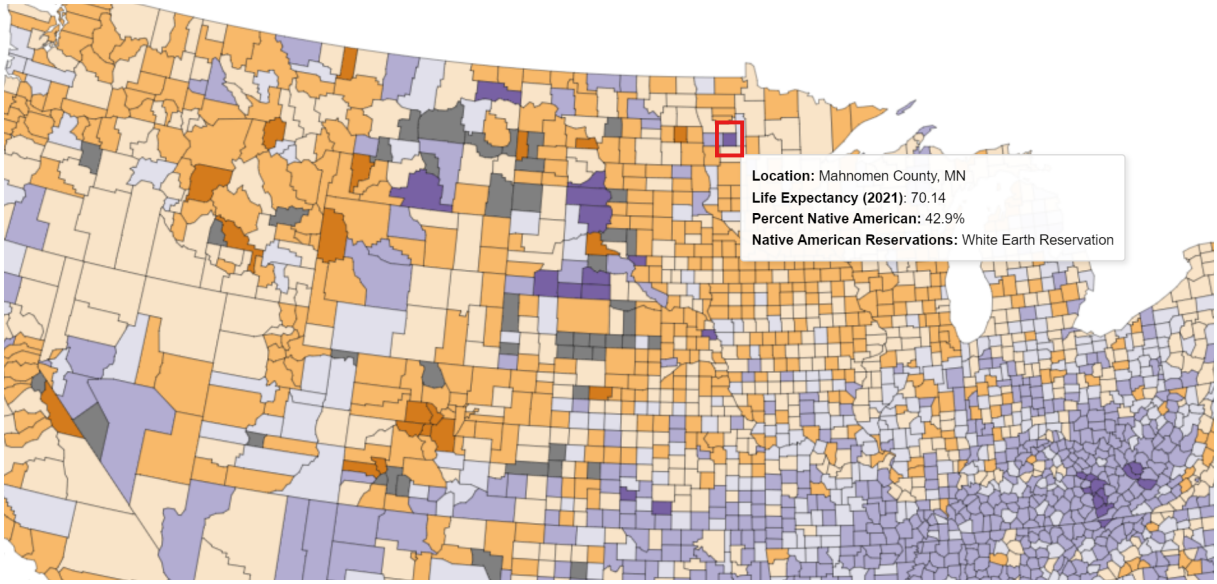


Figure 1: Example usage of County Buddy in the Exploropleth app. The figure shows Mahnomen County, MN has a much lower low life expectancy than surrounding counties. County Buddy informs the user that this is also the location of the White Earth Reservation, as the Native American population is about 43% of the total population in the county. The user can then investigate the systemic reasons why these two statistics align.

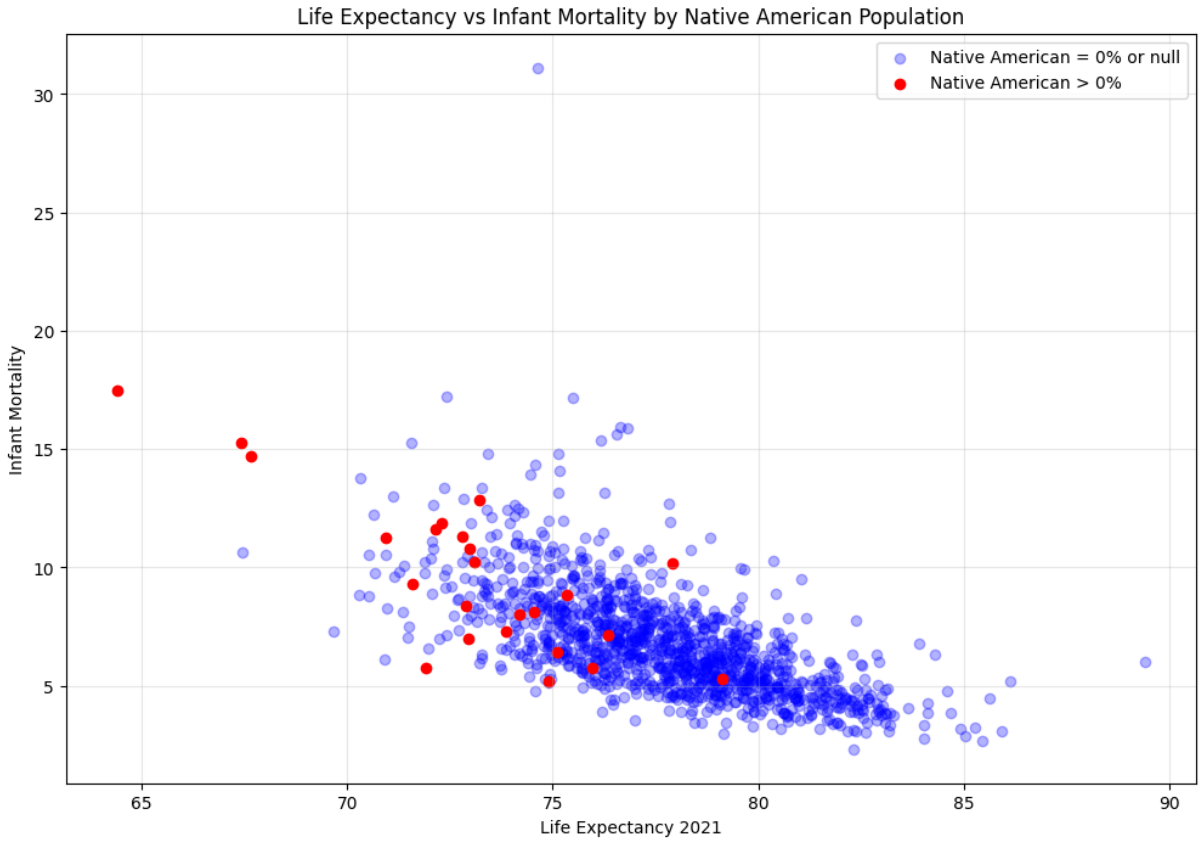


Figure 2: Data from https://ocular.cc.gatech.edu/resiliency-app plotted. Life Expectancy rates across counties are compared to Infant Mortality rates. Counties that meet County Buddy's Native American Population threshold are marked in red, whereas the rest are marked in blue.

Conclusion

In summary, this document describes a dataset created as a companion to traditional demographic and socio-economic analysis and data visualizations. While there are just four 'explainer' variables in this dataset, it was developed to facilitate and help support users who may not be able to aggregate and clean data from multiple sources. This data provides flexible thresholding and provides the names of institutions. Users of the data are encouraged to contribute to the public, open-source dataset and manipulate thresholds at their discretion.

Data Dictionary

Table 4: County Buddy Data Dictionary

Field	Description / Notes
FIPS code	The county/tract's geographic identifier. For county data, this is 5 digits. For tract data, this is 11 digits.
State Name	Name of the state where the county/tract is located.
County Name	Name of the county where the county/tract is located.
Tract Name (only for tract data)	Name of the tract.
Total Population	2020 population of the county/tract (U.S. Census Bureau, 2020a).
Prison Population	Residential population of the county/tract registered as incarcerated by the Census Bureau's 2020 Group Quarters survey (U.S. Census Bureau, 2020c).
Prison Population as Percent of Total Population	Percent of the total population that is registered as incarcerated, 2020.
Prisons	Prisons whose physical boundaries are contained within the county/tract. Only prisons with a capacity of 200 or greater in 2017 are included (Oak Ridge National Laboratory, 2017).
University Student Population	Residential population of the county/tract registered as a university student by the Census Bureau's 2020 Group Quarters survey (U.S. Census Bureau, 2020c).
University Student Population as Percent of Total Population	Percent of the total population that is registered as a university student, 2020.
Universities	Universities whose physical boundaries are contained within the county/tract. Only active public and private non-profit colleges (designated as type 1 or 2 by IPEDS) that are not considered Junior Colleges by NAICS classification and have a student population of more than or equal to 1,000 as of 2020 are considered (National Center for Education Statistics, 2019).
Nearby Universities	For tract data only. Universities whose physical boundaries are contained within the tract, or are within 1 kilometer of the border of the tract. Only active public and private non-profit colleges (designated as type 1 or 2 by IPEDS) that are not considered Junior Colleges by NAICS classification and have a student population of more than 1,000 as of 2020 are considered (National Center for Education Statistics, 2019).
Military Population	Residential population of the county/tract registered as active military by the Census Bureau's 2020 Group Quarters survey (U.S. Census Bureau, 2020c).
Military Population as a Percent of Total Population	Percent of the total population that is registered as in military, 2020.
Military Institutions	All Military Institutions whose borders contain some area within the region (U.S. Department of Transportation, 2023).
Native American Population	Residential population of the county/tract listed as Native American by the 2020 Census's demographic survey (U.S. Census Bureau, 2020b).
Native American Population as a Percent of Total Population	Percent of the total population that is surveyed as Native American, 2020.
Native American Reservations	All Native American Reservations whose borders contain some area within the region (United States Department of the Interior, 2021).

References

Arpit Narechania, Alex Endert, and Clio Andris. Exploropleth: Exploratory Analysis of Data Binning Methods in Choropleth Maps. *Cartography and Geographic Information Science*, 2025.

National Center for Education Statistics. Integrated postsecondary education data system - homeland infrastructure foundation-level data. https://web.archive.org/web/20250310062140/https://public.opendatasoft.com/explore/dataset/us-colleges-and-universities/table/, 2019. Data and geographic locations for Colleges and Universities.

Oak Ridge National Laboratory. Prison boundaries. https://www.arcgis.com/home/item.html?id=2d6109d4127d458eaf0958e4c5296b67, 2017. Data and geographic locations for secure detention facilities.

United States Department of the Interior. Indian reservations. https://www.arcgis.com/home/item.html?id=8fded139728f48b3b374a5dbf41dd4ec, 2021. Data and geographic locations for Native American Reservations.

U.S. Census Bureau. Cartographic boundary files. https://www.census.gov/geographies/mapping-files/time-series/geo/cartographic-boundary.html, 2020a. County level and tract level data and physical boundaries from the U.S. Census Bureau.

U.S. Census Bureau. 2020 census demographic and housing characteristics file (dhc). https://data.census.gov/, 2020b. County and tract level demographic data from the U.S. Census Bureau.

U.S. Census Bureau. Group quarters population by major group quarter type. https://data.census.gov/, 2020c. County and tract level group quarters data from the U.S. Census Bureau.

U.S. Department of Transportation. Military bases. https://catalog.data.gov/dataset/military-bases1, 2023. Data and geographic locations for Military Bases.

Colin Vu. County buddy code reference. https://github.com/ColinVu/CountyBuddy, 2025. Codebase for compiling outlier data.

Note: Disclaimer about images in this document. The researchers who created this document are not members of the incarcerated, military, or Native American population. If someone from one of these communities would like to adjust language in this document or suggest a graphic that better represents their population, please reach out.