Matt Ring

I. Sources

A. 5-Year American Community Survey (ACS)

ACS subjects collected included race, commute, poverty, employment, healthcare, and internet access. Only data on DC, Maryland, and Virginia at a county level were collected.

B. Department of Housing & Urban Development (HUD) Homelessness Point-in-Time (PIT) Data

PIT data contains homelessness trends in various "Continuums of Care" (CoC), or metro areas.

C. Bureau of Transportation Statistics (BTS) Intermodal Passenger Connectivity Database (IPCD)

IPCD data contains location and mode of transit information on every transit hub in the United States.

D. Center for Medicare and Medicaid Services (CMS) Hospital Location & Rating (Kaggle)

Kaggle dataset of Center for Medicare and Medicaid Services (CMS) data on U.S. Hospitals.

E. Census New Residential Construction (NRC) Housing Permits Data

Census data on the number of new residential housing units authorized per place and county.

F. Other Attempts

Attempts were made to access data on school funding, school location, and hospital funding. Linking these data to Census tracts, lack of unique identifiers, and ease of download were the main difficulties. This Census API and this National Center for Education Statistics crosswalk may help for future attempts. Note also that future work may be done on the IPCD data to split data by year.

II. License

All data were collected from public sources, such as the Census and HUD. Works published by the U.S. Government are in the public domain, and thus permission is not required to use them.

III. Collection Methods

ACS data were downloaded using the R package "tidycensus", which accesses the Census API using an API key. All other data were downloaded from the sources linked above.

IV. Biases/Sampling

First, HUD PIT data is at a CoC level, but we are disaggregating to the county level. In doing so, we're assuming homelessness varies with a county's population within the CoC. Thus, intra-CoC homelessness variation only represents population variation. Second, HUD PIT data only exists for CoCs, which are often metropolitan areas. Finally, both transit and hospital data are for specific times, 2021 and 2017. Thus, no analysis of hospitals or transit hub change over time is possible.

V. Descriptive Statistics

All data are numerical, except for the county and state names. Distributions for each column were assessed in pandas. No tribal hospitals exist within these data and multifamily homes, ferries, train stations, and hospitals are very uncommon, with the median value for each feature being zero.

ACS data on healthcare coverage seemed off with a minimum proportion of 1.48, maximum of 1.95, and average of 1.78. This issue will be assessed in depth in the future.

Finally, there were obvious outliers due to the presence of cities. For example, the maximum county population was ~1 million, whereas the median population was ~30,000.

VI. Data Cleaning & Identified Issues

Cleaning the data required putting the data into a tidy format. This means that 1) all rows represent a single observation, 2) all columns represent a unique feature of that observation, and 3) each observation cannot be subdivided further.

For this analysis, data were stored at a county-year resolution. When provided tract-level FIPS, this was required removing the last 6 digits to get county-level FIPS. Crosswalks and APIs were used to convert CoC, latitude/longitude, and county names data to county FIPS for the housing, school, transit, and hospital data. Once formatted, merging required a left join in R.

The main concerns upon cleaning and merging were missing and duplicate data, as each county-year should be a unique value. First, missing data was analyzed using the missingno package.

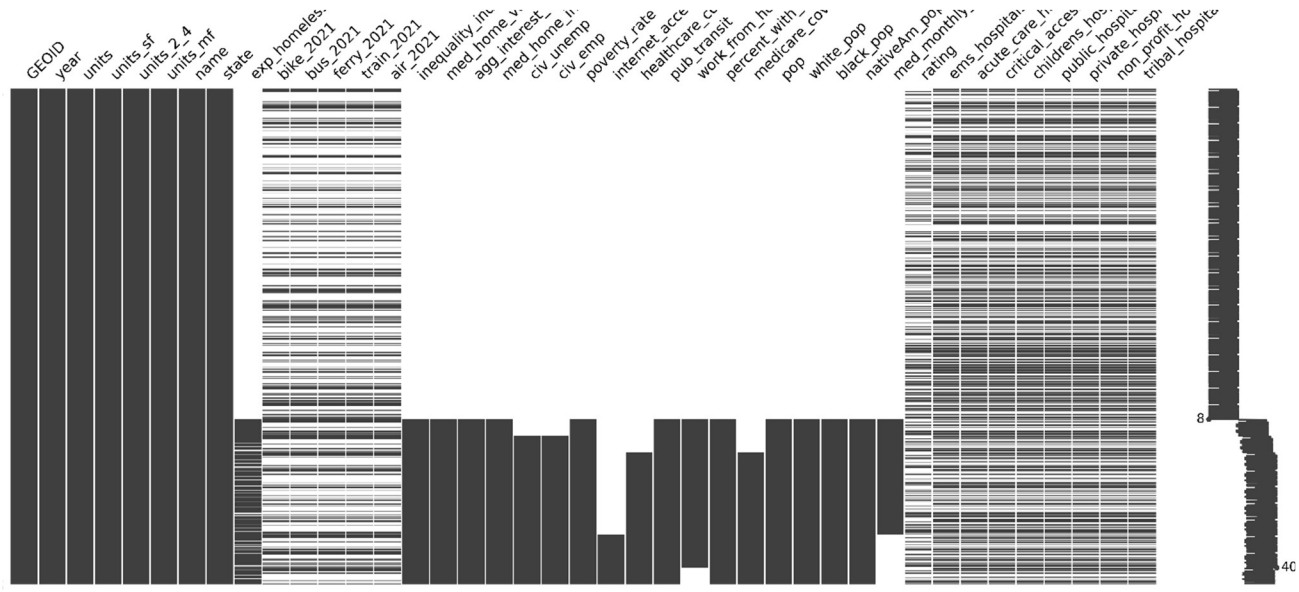

Figure 1: Missing data sorted by year then county.

As we can see in Figure 1, there are trends in the missing data. Most features are not present until around 2007-2009, when ACS and HUD homelessness data begin. Before that, only housing data is consistent. Transit and hospital data are more sporadic.

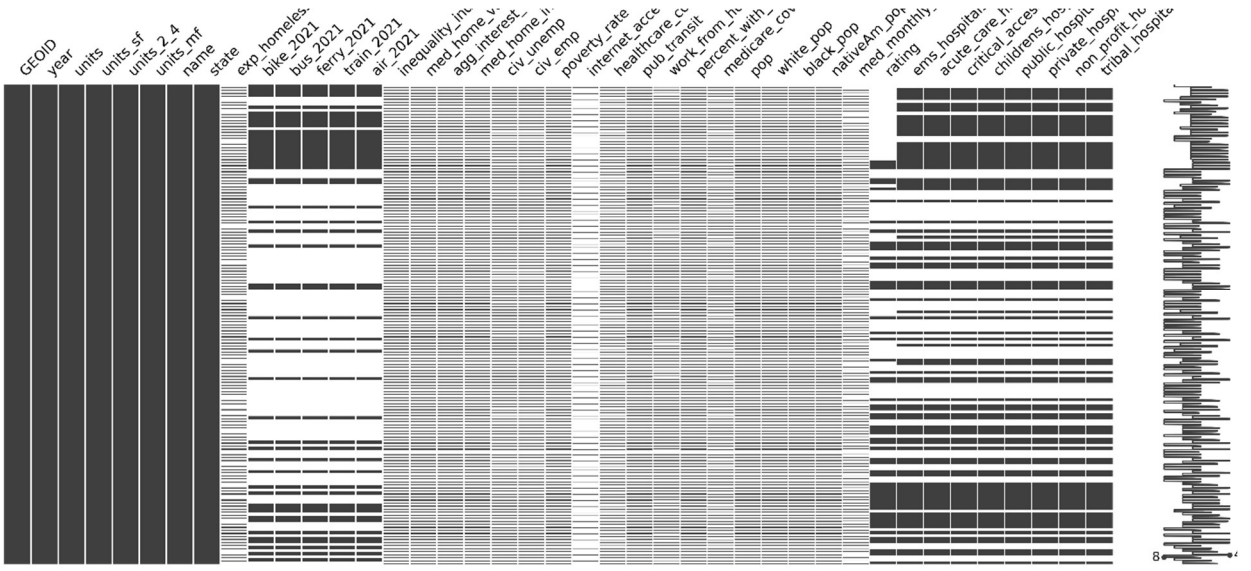

Figure 2: Missging data sorted by county then year.

Figure 2 highlights the fact that transit and hospital data are not randomly missing but absent for specific counties without any hospitals or transit hubs. We need to acknowledge these missing values, but they still provide meaningful information.

Finally, duplicate data can be detected using .drop_duplicates() on county (GEOID) and year in pandas. A quick check shows no duplicate county-year combinations. State labels were also assessed, and no duplicates, such as “VA” and “Virginia”, were found.

VII. Data Dictionary of Key Features

Category	Description	Example features
Identifiers	Used to identify the observation in the data and in reality.	GEOID, year, name, state
Housing	Number of different housing units permitted, such as single family (sf) homes.	units, units_sf, units_mf
Homelessness	Count of people experiencing homelessness.	exp_homelessness
Transit	Count of transit hubs by type in 2021. A single hub may support multiple modes of transit.	bike, bus, ferry, train, air
Hospital	Count and rating of hospitals. Hospital counts are disaggregated by type of hospital.	rating, public_hospitals
Wealth & Income	Metrics of wealth and income measured, such as salary, or calculated, such as inequality.	inequality_index, med_home_income
Demographic	Racial/ethnic population counts.	pop, black_pop, white_pop