Project 02: Homelessness Correlations

Emily Hillenbrand

10.02.2019

Problem

Homelessness is a complicated problem with many different related variables. By comparing different metrics across US states over several years, I hope to gain insights into the factors that contribute to the rate of homelessness in the US.

Goals

1. Show the relationship between homelessness and explored variables.
2. Create a model to predict rate of homelessness based on a given set of input values.

Milestones

I. Minimum Viable Product (MVP)

An MVP would show the relationship between the total rate of homelessness and income inequality. This would then be expanded to include housing cost, poverty rate, unemployment rate, and other variables shown in the Data table.

II. Stretch Goals

Other variables to consider and add to the MVP would be: rate of mental illness, number of people with criminal histories, and health insurance coverage. I would also want to break the HUD data down by city to get a better model for homelessness in major cities, and consider different demographics among homeless people to gain a better understanding of more specific contributing features.

Data

Independent Variable	Source	Type	Dimension
Income Inequality	census.gov	Continuous (float)	Gini Index
Poverty Rate	census.gov	Continuous (float)	People in poverty
Average Income	census.gov	Continuous (int)	Inflation-adjusted USD
CoCs*	hudexchange.info	Continuous (int)	count
Total Population	census.gov	Continuous (int)	individuals
Cost of Living	numbeo.com	Continuous (float)	index
Education Budget	https://www2.ed.gov/about/overview/budget/history/index.html	Continuous (int)	USD
Unemployment Rate	bls.gov/data/#api	Continuous (float)	Percentage
Dependent Variable	**Source**	**Type**	**Dimension**
Homelessness	hudexchange.info	Continuous (int)	Individuals

*CoC = Continuum of Care, which represents a community-level program to assist homeless persons, with the goal of long-term stability.

Data notes:

All data will be iterated by year and US state.

Known Unknowns

1. Defining interactions between different features may be beyond the scope of current resources.
2. Because there are different ranges of years between different data sources/features, the timescale may need to be limited to that for which there is the most overlap. Some features may need to be excluded for this reason.
3. Some variables (e.g. cost of living) are given according to specific cities instead of states. Consider whether it would be best to take an average of values present for each state, or save this feature for when model is recalculated in terms of cities.