Project 02 Summary: Homelessness Correlations

Emily Hillenbrand

10.13.2019

Project Design

My goal was to find factors that correlate with homelessness in the United States. While I expected this problem to be complicated, ultimately I found that it's not the kind of problem that can be examined in the straightforward way that I had tried.

I had originally planned on aggregating information for each state, but after some exploratory data analysis I found that the homelessness counts from the Housing and Urban Development (HUD) site was inconsistent with the data it provided and seemed to have counts for individual cities more reliably than for all cities in a state.

One of the biggest challenges with the data was to figure out how to connect information from different sources. I ended up taking each Continuum of Care (CoC) and scraping the HUD website for its address, then taking that information and matching it up with population and land area. This involved repeatedly splitting address columns into individual city and state columns and coming up with a consistent way to match them to other dataframes. In the process of cleaning that final dataframe, I dropped rows where data was missing and ended up losing about 40 rows.

I was able to find that average homelessness counts for any particular year were highly correlated with those of the year before, but I was not able to find a higher correlation with any of the other features I had considered. Due to time constraints, I was not able to collect data to examine any other features, but I would be very interested in an opportunity to look at this problem differently.

Tools

1. Requests and BeautifulSoup for web scraping
2. Pandas, numpy, pickling for data importing, cleaning, and storage
3. SciKitLearn for data analysis:
    a. Train test split
    b. Cross-validation
    c. Linear regression

    d. Ridge

4. Google docs, Microsoft Powerpoint

Data

Variable	Source	Dimension	Method
Overall Homeless	hudexchange.info	count	csv
Addresses of CoCs	hudexchange.info	People in poverty	Web scraping
City Population	census.gov	count	csv
City land area	census.gov	Square miles	csv

Future Work

I. Data Stitching

In the process of stitching dataframes together from different sources, I needed to drop about 10% of the rows that I had started with because of missing information. The remaining rows were of dubious veracity, but were functional enough for the purposes of an initial analysis. In the future, I would like to find a more reliable way to connect the dataframes and make sure they contain correct information.

II. Density Change

While my initial hypothesis was that a sharp increase in population density has related to an increase in homelessness in Seattle, I have since amended that with the assertion that the inverse is not necessarily true; a decrease in population density does not necessarily cause a decrease in homelessness. In future analyses of this relationship, I will examine only the cities with a drastic increase in population density over the relevant time period.

III. Additional Features

I had originally chosen population density as a starting point because it corresponds to other changes that can lead to homelessness, including increased rent, cost of living, and income disparity. I am interested to know if these can be identified as leading indicators of increased homelessness.