Probing the Relationship Between Education and Homlessness in the USA

Authors: Kim Horany, Roy Mojica, Lawrence Watson

1. Project Summary

Our topic:

We are exploring the relationship between homelessnes and education in the USA with an aim to identify trends that may predict the number of homeless people in a state.

Our Motivation:

The swift and unexpected onset of a global health crisis in 2020 led many states to issue stay-at-home orders to curtail the rate of infection. Without homes to retreat to, our homeless population was most vulnerable during the pandemic. We may not be able to predict or control the next devastating public health crisis, but is there an aspect to homelessness that we might govern to reduce the size of those affected? That is what we hope to discover.

Our Datasets:

Homelessness in the United States, 2007 to 2016

The raw data set contains Point-in-Time (PIT) estimates and national PIT estimates of homelessness as well as national estimates of homelessness by state and estimates of chronic homelessness from 2007 - 2016 .

U.S. Education Dataset:

The dataset is designed to bring together multiple facets of U.S. education data (K-12 financial, enrollment, and achievement data) from 1992 - 2017 into one CSV.

2. Exploring the Data

3. Analyzing the Data

Machine Learning Model

Data Preprocessing

Preprocessing the Homelessness Dataset

- Read raw data from CSV
- Drop unneeded columns – we decided that the shelter name was not relevant to the data
- Create a key column that will allow us to join to the education dataset
    - Extract year from Year column
    - Concatenate year with state, ex: 2007_AK
- Create bins for the Measures column to reduce the categories to either Sheltered, Unsheltered, or Other
- Drop original Measures column
- Group by State_Year, Year, State, Groups to aggregate the Count
- Flatten the table so that each year/state will have one record
    - Create 3 columns: Sheltered_Cnt, Unsheltered_Cnt, Other_Cnt and store count under appropriate column

Preprocessing the Education Dataset

- Read raw data from CSV
- Drop unneeded columns - columns with too many null values
- Create a key column that will allow us to join to homeless dataset
  - Convert long state name to abbreviated state: ALASKA --> AK
  - Concatenate year with state, ex: 2007_AK
- Replace NaN with zeros
- Convert floats to int
  - Merged dataset
- Join homeless dataset and education dataset on the key: state_year
- Drop duplicate columns
- Reorder columns so the count columns are at the end
- Encode State column with get_dummies
- Use Standard Scaler on the X features

Preliminary Engineering

Preliminary feature engineering includes reducing the 20+ Measures categories to just 3 categories, scaling the features using Standard Scaler, and encoding the categorical column (State) using get_dummies. We chose the features to be the state, year, total state revenue, total state expenditure, number of students in high school, and number of students in all grades to be the features. We wanted to see if this combination of columns could be used to accurately predict homeless counts for subsequent years.

Model Choice

- We chose a multivariate multiple regression model, because we had more than two independent variables, more than one dependent variable, and we wanted to predict a continuous value. The benefits of multivariate regression is that we can get a more realistic picture than when just observing one dependent variable. This technique can also provide a more powerful test of significance than typical multiple regression. A limitation of multivariate analysis is that you need to have large datasets to overcome high standard errors. Our dataset may not be large enough to overcome this limitation.

- To determine whether this model is a good fit, I got the R-squared score, which was 97.8%. This indicates a good fit. However, it may not be accurate, because our dataset might not be big enough to overcome the limitations of using a multivariate regression model. Despite this limitation we decided to use this statistic, because R-squared shows the fraction of the variance between values predicted and the value rather than the mean of the actual.

Database Integration

Visualization Blueprint

Visualizations

Heat Maps

Values: X -Years, Y - States, Heat Value - Total Homeless

Description: A heat map to display the homeless population in each state for every year 2007-2016. If a state's homeless population is trending downward over time, it will be represented.

Bubble Chart

Values: Y - Total Expenditure (10 year period), X - Grades All (10 year period), Bubble - Total Homeless per State

Description: Total homeless value as a percentage of State population is represented w/ dimensions of bubble labeled by State. Bubbles are plotted on a graph that shows how much money was spent on education and the aggregate k-12 grades. Will help identify which states are spending the most on education and whether or not it has an impact on the homeless population.

Interactive Element

Description: Map of the United States displaying where each state is colored with a heat value to represent either Homelessness, Educational Spending, or k-12 grade performance. A button will be used to select between the three values. A slider at the map's edge will be used to select the year 2006-2017.