DSC540

Term Project Milestone 1

Holly Figueroa

Project Subject Area:

Relationships to Nursing Home Qualities in Care

Data Sources:

Flat File (County Characteristics)
Sourced from the USDA's Economic Research Service. This dataset includes a range of county level data variables regarding distribution of population race, age group, and education attainment. It also includes measures for population density, unemployment, and more. Data contains measures across years to measure change/growth. Data was last updated as of 2021.File contains over 100 column variables and over 3K rows.
https://www.ers.usda.gov/data-products/atlas-of-rural-and-small-town-america/download-the-data/

API (Nursing Home Data)
Sourced from the CMS (Centers for Medicare and Medicaid Services) website. This dataset includes a range of information available on all individual, active nursing homes in the United States. Details about each provider's location, business name, ownership type, insurance acceptance, capacity, and more are provided. The data set also includes variables for quality of staff, quality of care, as well as recorded incidents, complaints, and fines. Data reflects provider information from 2021. File includes over 80 available column variables and over 15K rows.
https://data.cms.gov/provider-data/dataset/4pq5-n9py

Website (State Funding Choices)
Sourced from the KFF (Kaiser Family Foundation) website. This dataset details the distribution of state expenditures. Categories include levels of education, Medicaid, public assistance, corrections, and transportation. Data reflects state spending in the year 2020. File includes 9 column variables and roughly 50 rows.
https://www.kff.org/other/state-indicator/distribution-of-state-spending/?dataView=0&currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D

Relationships

All data sources share a "State" variable regarding location in the US. Each set can be aggregated to the state level to explore relationships between nursing homes qualities, state spending, and various state population attributes. Larger sets from the CMS and USDA share a county variable. Relationships from these sets can be explored between county characteristics, and nursing home qualities and Medicare spending per beneficiary (also provided by CMS data).

Approach & Challenges

My hope is to approach each milestone in a practical manner given the real and practical nature of the datasets. However, my personal goal for each milestone will be to include transformations that are practical and transformations that are unfamiliar to me for the purpose of practicing those skills. The datasets I have reviewed so far for this project will require a great deal of reduction. So, while my data will require many columns to be removed or subsetted, I will first explore their potential for unfamiliar transformations.

I will then explore transformations to be used to create new, relevant variables. Variables will have to be created through aggregation to combine data sets at the state level and at the county level. I may also explore ways to aggregate data at the provider chain level, if I find many to be identifiable through their shared strings (i.e., “Eldercare of Omaha”, “Eldercare of Palm Rio”).

Once variables of interest have been settled for each level of interest (State, County, Provider), I will combine and merge them respective to each level. Visualizations will be used to highlight any potential patterns between quality of care. Quality care ratings at both extremes may be filtered to get clearer visualization of their relationship to other variables. If these transformations work well, then the resulting data may be used to identify factors for consideration and concern when choosing a nursing home provider. Considerations for provider location, state policies, and state practices with funding may also be identifiable.