Documentation of the individual project

2022/03/13

Ryu Sonoda

The purpose of this project:

Health is one of the most important aspects in our life especially in the current world with pandemic. There have been numerous researches to explore what factors are related to the community health outcome (N. Krieger1, D. R. Williams2, and N. E. Moss, 1997) and still going on. In this project, I tried to build a web application that can help the public health researchers, policy makers, or the general public could explore the current trend of the health-related outcome across the U.S, which could lead to the further research questions to investigate in the future. I implemented the function by utilizing choropleth map. Additionally, the simple linear regression feature to examine the correlation between two variable was also implemented.

Data Collection and Data Description:

The main dataset, 2020 County Health Rankings Data, was obtained from County Health Rankings & Roadmaps. This dataset compiles data on various health outcome measures by US county, collected from many different sources such as American Community Surveys, 5-year estimates, Behavioral Risk Factor Surveillance System, and Bureau of Labor Statistics, using telephone surveys and through accessing vital registration systems. I combined data on the area of each county from the Missouri Census Data Center with the population data from the 2020 County Health Rankings dataset to calculate the population density of each county. I identified 36 explanatory variables that seems to indicate the health outcome of the community or seems to influence the health outcome. Additionally, the primary geological data, cartographic boundary file, which contains the boundary information of the county across the U.S, was obtained from United States Census Bureau. In order to project the user entered county on the map, longitude and latitude of each county was added by merging United Stated Counties Database from simplemaps. The following is a brief description of the variables in the dataset

Primary variables

|  |  |  |
| --- | --- | --- |
| Variable Name | Description | Type of variable |
| State | State name | Categorical |
| NAME | County name | Categorical |
| lng | Longitude | Numerical |
| lat | Latitude | Numerical |
| % Fair or Poor Health | Percentage of adults that report fair or poor health | Numerical |
| Average Number of Physically Unhealthy Days | Average number of mentally unhealthy days reported in past 30 days (age-adjusted) | Numerical |
| Average Number of Mentally Unhealthy Days | Average number of physically unhealthy days reported in past 30 days (age-adjusted) | Numerical |
| % Low Birthweight | Percentage of live births with low birthweight (< 2,500 grams). | Numerical |
| % Smokers | Percentage of adults who are current smokers. | Numerical |
| % Adults with Obesity | Percentage of the adult population (age 20 and older) that reports a body mass index (BMI) greater than or equal to 30 kg/m2. | Numerical |
| Food Environment Index | Indicator of access to healthy foods - 0 is worst, 10 is best | Numerical |
| % Physically Inactive | Percentage of adults that report no leisure-time physical activity | Numerical |
| % With Access to Exercise Opportunities | Percentage of the population with access to places for physical activity | Numerical |
| % Excessive Drinking | Percentage of adults that report excessive drinking | Numerical |
| Teen Birth Rate | Births per 1,000 females ages 15-19 | Numerical |
| % Unemployed | Percentage of population ages 16+ unemployed and looking for work | Numerical |
| Income Ratio | Ratio of household income at the 80th percentile to income at the 20th percentile | Numerical |
| Social Association Rate | Associations per 10,000 population | Numerical |
| Violent Crime Rate | Violent crimes per 100,000 population | Numerical |
| % Severe Housing Problems | Percentage of households with at least 1 of 4 housing problems: overcrowding, high housing costs, or lack of kitchen or plumbing facilities | Numerical |
| % Drive Alone to Work | Percentage of workers who drive alone to work | Numerical |
| % Long Commute - Drives Alone | Among workers who commute in their car alone, the percentage that commute more than 30 minutes | Numerical |
| Life Expectancy | Average number of years a person can expect to live. | Numerical |
| Age-Adjusted Death Rate | Number of deaths among residents under age 75 per 100,000 population (age-adjusted). | Numerical |
| Child Mortality Rate | Number of deaths among children under age 18 per 100,000 population. | Numerical |
| Infant Mortality Rate | Number of all infant deaths (within 1 year), per 1,000 live births. | Numerical |
| % Food Insecure | Percentage of population who lack adequate access to food | Numerical |
| % Limited Access to Healthy Foods | Percentage of population who are low-income and do not live close to a grocery store | Numerical |
| Drug Overdose Mortality Rate | Indicator of number of drug poisoning deaths (x) per 100,000 population: very few (<10), few (10<x<12), moderate (12<x<43), or many (43<x) | Numerical |
| % Insufficient Sleep | Percentage of adults who report fewer than 7 hours of sleep on average | Numerical |
| % Uninsured | Percentage of people under age 65 without insurance | Numerical |
| Median Household Income | Income where half of households in a county earn more and half of households earn less | Numerical |
| Suicide Rate (Age-Adjusted) | Number of deaths due to suicide per 100,000 population. | Numerical |
| % Severe Housing Cost Burden | Percentage of households that spend 50% or more of their household income on housing | Numerical |
| Population | Number of population | Numerical |
| Total area (square miles) | Total area (square miles) | Numerical |
| Population density | Number of people per square mile | Numerical |
| Mental Health Providor Rate | Number of Mental Health Providers per 100,000 population | Numerical |
| Primary Care Physicians Rate | Primary Care Physicians per 100,000 population | Numerical |
| Dentist Rate | Dentist per 100,000 population | Numerical |

The detailed process of the data wrangling

First, I read the Excel data, County Health Rankings Data - v2.xlsx, which holds numerous health-related variables into R. Since it contains more than 100 variables, I chose 36 variables that are the most relevant or interesting to keep. Since the chosen variables were located on the different spreadsheet, I merged the two data frames by county and state name. Then, I read another CSV file, called dexter\_2129905758\_extract.csv, to create a new variable called, population density. I computed population density through dividing Population by newly added variable, Total Area. Since a few variables such as Mental Provider Ratio, Primary Care Physicians Ratio was storing the data as ratio (like 1:200), I separated the values into two columns and by dividing the first column with the second column, I created a new variable that holds the rate (like 0.05) in order to convert the values into numerical. Then, I completed compiling the health-related variables, and called the dataset as CSC324\_project\_data.csv. Additionally, in order to add the wrangled variables as new attributes to the shapefile, I merged the csv file and shapefile with FIPS of the county. To preserve the format of shapefile, I first created an intermediate that holds the health variables in the same order as the shapefile and then add the intermediate into the shapefile.

What-why-how?

Geometry: The dataset was generated based on the multiple attributes of health outcomes of counties such as mentally unhealthy days and % adults with obesity. With these attributes, shapefile contains boundary information of each county.

Why?

Since one of the purposes of users to investigate how different the health circumstances are across the U.S. After they identify some interesting particular variable, they can start investigating the correlation with other variables.

How?

The result was illustrated by the map. First, the system separates each observation based on location and juxtaposes them. Then, based on the values, they paint each county with different colors.

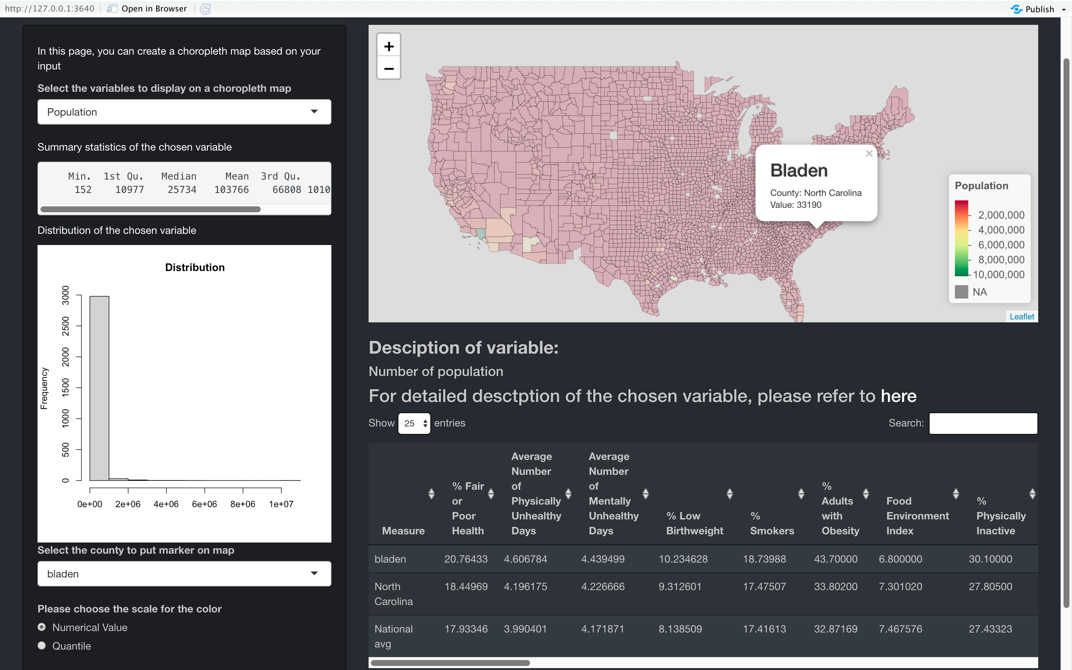
Primary feature

The application contains two pages.

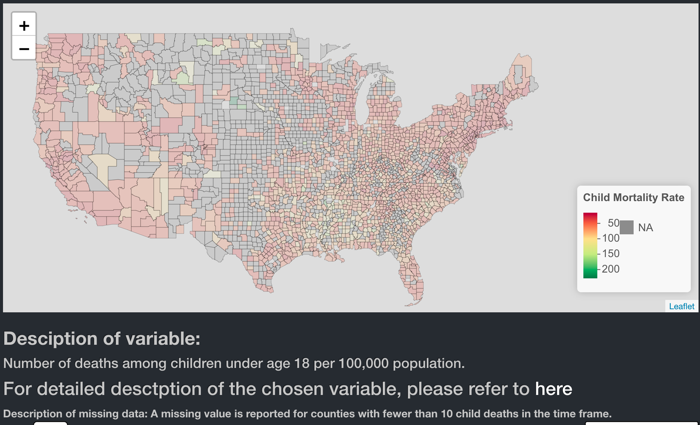
In the first page, you can explore the overall trend of each variable or investigate the characteristics of an individual county (The first screenshot below). On the sidebar, there are three widgets that take user input. The first bar allows users to choose the variable they want to explore. Based on their choice, the choropleth map is displayed. The users can choose either actual numerical value of quantile values of the values as a scalar of color by selecting the radio buttons located at the lowest part. Simultaneously, the summary statistics and the distribution of the values of the chosen variable are displayed on the side bar. Brief description of the variable is displayed main panel. For those who want to know how the data was collected or the definition of the variable, the hyperlink is imbedded on here. If the variable contains a lot of missing variables, the reason for that might be displayed as well (The second screenshot).

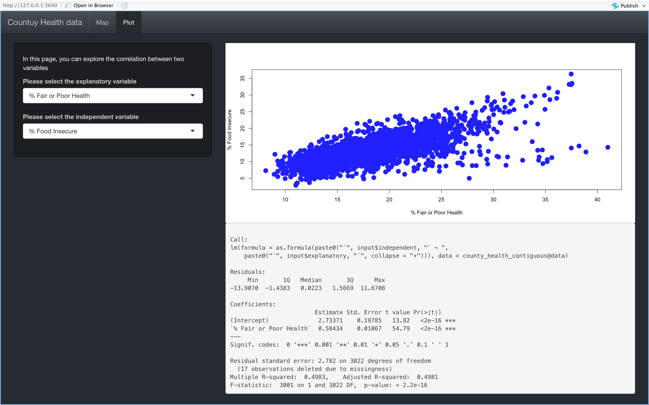
Moreover, you can search one specific county’s data by the input bar right below the histogram. The popup will appear on the county users selected with its name, state, and values. The table on the main panel shows the value of every variable for the county, the average of the state to which the county belongs and the national average so that users can compare a lot of characteristics at once.

In the second page, you can inquire the correlation between two variables. The plot helps users visually understand the strength of the correlation. The summary below the plot includes various information about the regression such as the p-value, coefficients, and the intercept (The third screenshot).



The layout of the first page

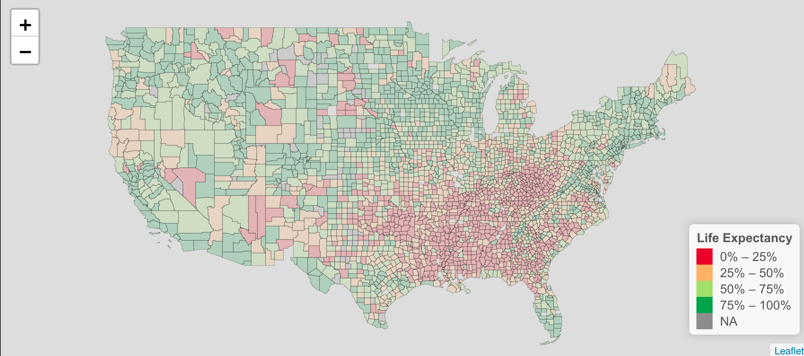
 The description when many N/A are found

 The layout of the second page

My questions, insights and answers

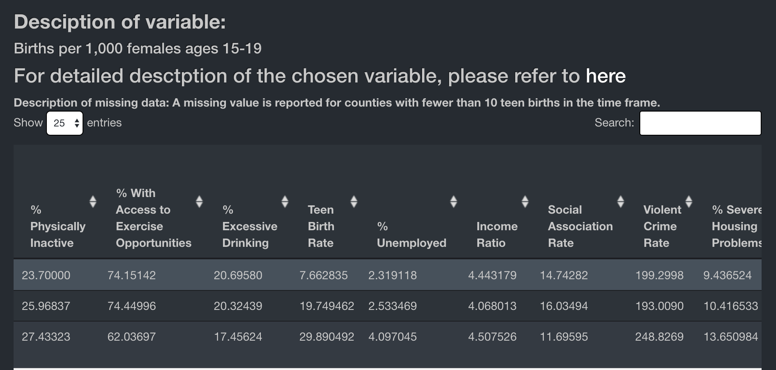
Is there any geographical trend about the life expectancy?

As the map below shows counties located on the south east part seems to have relatively shorter life expectancy. This kind of tendency was seen in other variables such as % adults with obesity, so the area might be slightly unhealthy compared to others due to cultural or geological reason.



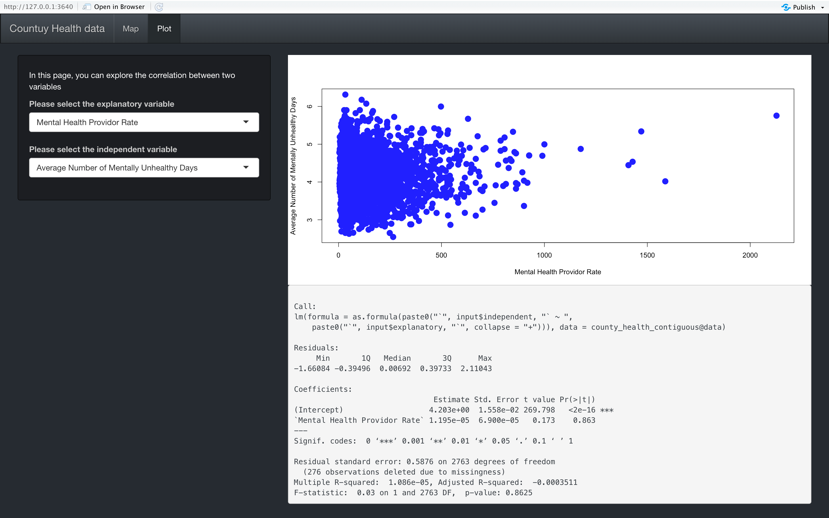
Does our county, Poweshiek, have any special characteristics in terms of health?

There did not seem to be any particular extreme values. However, teen birth rate in Poweshiek (7.66 births per 1000 females) is much lower than the state average (19.75) or national average (29.89).



Does the number of mentally unhealthy day decrease if there are more mental health care provider?

There seems to be not strong correlation since both the p-value (0.8625) is high and the adjusted R-square (0.0003) is very low. The finding is very interesting since I expected that the if there are more resources, the mental health situation might be better. However, we cannot conclude that those are completely unrelated, since it could have impact on mental health through interaction with other variables.



Improvement

Overall, I am satisfied with the application. However, there are several things to work on in the future. First, when using the function to search one specific county, there is a bug in the created table, if there are multiple counties with the same name. In order to avoid it, I should add another key to search county such as state. Second, can only do a simple linear regression with only one explanatory variable for now. A linear regression is not enough if the variables have complex relationships, so I will implement the more options in the future such as multilinear regression. Moreover, since the user can only explore 2020’s data, I would like to include datasets from previous years and allow users explore the change over the years,

References

MCDC Data Applications. Missouri Census Data Center,

<https://mcdc.missouri.edu/cgibin/broker?_PROGRAM=utils.uex2dex.sas&path=%2Fdata%2Fpl942020&dset=uscounties&view=0>

“Explore health rankings: Measures & data sources.” County Health Rankings & Roadmaps. <https://www.countyhealthrankings.org/explore-health-rankings/measures-data-sources>

United States Counties Database, Simplemaps, <https://simplemaps.com/data/us-counties>

Cartographic Boundary Files, United States Census Bureau, <https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html>