Group 5

Data Wrangling Final Project Social Determinants of Health

Team Members
Jonathan Himes jonathan.himes@utah.edu U1367169
Michelle Kubicki michelle.kubicki@utah.edu U1370752
Yueqin Yang yueqin.yang@hsc.utah.edu U0737683
Jackie Bearnson u0763123@umail.utah.edu U0763123
Lucia Ranallo U1427138@umail.utah.edu U1427138

Project Description

Social determinants of health are defined by the World Health Organization (WHO) as nonmedical factors that influence health outcomes. Examples include the conditions in which people live and the broader set of influences creating conditions of daily life. Factors can include economic policies and systems, political systems, development, societal norms and policies, racism, and climate change. Knowledge of these factors and their incorporation into research, community, and medical practices could lead to improvements in health outcomes, as social determinants of health have been linked with health risks, life expectancy, and disease outcomes.

Social determinants of health are often not readily available or accessible on an individual level in electronic health records, making them difficult to incorporate into research and downstream practice. This information is, however, collected on a large scale via publicly available tools such as censuses. The ability to identify these factors and their effects could play an important role in improving health outcomes and reducing health disparities in disadvantaged populations by considering them in methods like policy change and community outreach.

The objective of this project is to compile a database of SDOH variables for Utah over multiple years to aid in SDOH comparison. The following is a summary of the Utah SDOH variables chosen and whether they were available for both 2018 and 2019.

SDOH items were chosen to be similar to those in the article:

- "FACETS: using open data to measure community social determinants of health"
  - by Michael N Cantor, Rajan Chandras, and Claudia Pulgarin.

This paper can be found at the following link:

- https://academic.oup.com/jamia/article/25/4/419/4569610

Extensive research was done in acquiring the data from the data sources. Some SDOH variables chosen by the authors of the above paper were not included in this project due to:

- Variables were New York Specific
- Variables cost money to obtain.
- Variables not available for designated years

Below is a table of the SDOH variables that were available for Utah for each year.

SDOH Variables Compiled

	Item	Description	Source	2018 Availability	2019 Availability
0	FIPS	US Census 11 digit FIPS code for tract	US Census	YES	YES
1	Urban	Urban/Rural Flag	USDA Food Access Research Atlas	NO	YES
2	Total_population	Total Population	ACS 2015 Estimates	YES	YES
3	P_WH	% White	ACS 2015 Estimates	YES	YES
4	P_AA	% African-American	ACS 2015 Estimates	YES	YES
5	P_AI	% American Indian	ACS 2015 Estimates	YES	YES
6	P_AS	% Asian	ACS 2015 Estimates	YES	YES
7	P_NH	% Native Hawaiian/Pacific Islander	ACS 2015 Estimates	YES	YES
8	P_OR	% Other Race	ACS 2015 Estimates	YES	YES
9	P_2R	% 2 or more Races	ACS 2015 Estimates	YES	YES
10	P_Latino	% Latino/Hispanic Ethnicity	ACS 2015 Estimates	YES	YES
11	P_native	% Native Born in US	ACS 2015 Estimates	YES	YES

	Item	Description	Source	2018 Availability	2019 Availability
12	P_FB	% Foreign Born	ACS 2015 Estimates	YES	YES
13	P_citizen	% US Citizen	ACS 2015 Estimates	YES	YES
14	P_non-citizen	% Non-citizen	ACS 2015 Estimates	YES	YES
15	P_NoSchool	% No schooling	ACS 2015 Estimates	YES	YES
16	P_HS_no_degree	% completed high school, no degree	ACS 2015 Estimates	YES	YES
17	P_HS_or_GED	% High school or GED degree	ACS 2015 Estimates	YES	YES
18	P_some_college	% Some college, no degree	ACS 2015 Estimates	YES	YES
19	P_college_degree	% AA or BA	ACS 2015 Estimates	YES	YES
20	P_Masters_prof_doc	% Masters, professional, doctorate	ACS 2015 Estimates	YES	YES
21	P_limited_eng	% Limited English proficiency	ACS 2015 Estimates	YES	YES
22	Poverty_rate	% in poverty	ACS 2015 Estimates	YES	YES
23	MED_HH_income	Median Household Income	ACS 2015 Estimates	YES	YES
24	UE_rate	Unemployment rate	ACS 2015 Estimates	YES	YES
25	P_UI	% Uninsured (health)	ACS 2015 Estimates	YES	YES
26	P_Insured	% Insured	ACS 2015 Estimates	YES	YES

	Item	Description	Source	2018 Availability	2019 Availability
27	P_UI_under_18	% Uninsured, <18 yo	ACS 2015 Estimates	YES	YES
28	P_UI_18-64	% Uninsured, age 18-64	ACS 2015 Estimates	YES	YES
29	P_UI_65_over	% Uninsured, over 65	ACS 2015 Estimates	YES	YES
30	P_any_private_ins	% Any private insurance, all ages	ACS 2015 Estimates	YES	YES
31	P_any_public_ins	% Any public insurance, all ages	ACS 2015 Estimates	YES	YES
32	P_Medicare_alone	% Medicare only	ACS 2015 Estimates	YES	YES
33	P_Medicaid_alone	% Medicaid only	ACS 2015 Estimates	YES	YES
34	Resp_HI	Respiratory Hazad Index	EPA National Air Toxics Assessment	YES	YES
35	Low_access	Low access to healthy food (1/2 mile)	USDA Food Access Research Atlas	NO	YES
36	Walkscore	Neighborhood walkability scale	Rundle- Columbia BEH	YES	NO
37	GINI	GINI inequality index	US Census	YES	YES
38	SVI_themes_total	Social Vulnerability index- total themes perce...	CDC SVI	YES	NO
39	SVI_flags	Social Vulnerability index- score for flags	CDC SVI	YES	NO

Must Have Features

1. A clear and concise description of the SDOH variables used in the analysis.

2. SDOH variables for every census tract in the state of Utah for 2018 and 2019

3. A cleaned and transformed dataset that is suitable for analysis.

4. Identification and treatment of missing data and outliers.

5. Exploratory Data Analysis (EDA) of the Data to understand distributions and skewness

6. Statistical Comparison of changes in SDOH between the 2 years

Methods

1. Dataset Creation
    1. Compile data from sources
    2. Ensure that all FIPS were accounted for each column
    3. Ensure no duplicate FIPS for each column
    4. Filter out only relevant data
    5. Ensure data is the correct data type
    6. Transform data as needed
    7. Ensure data is sorted by FIPS number so that the data harmonizes correctly
    8. Combine all columns into a dataset
2. Dataset Exploration
    1. Generate descriptive statistics for each variable (count, mean, std, min, etc.) and ensure that none of the data was out of the ordinary
    2. Generate histograms for each variable to see if anything looked out of the ordinary
    3. Check for missing data. If data was missing, it was replaced with a np.nan
3. Data Comparison
    1. View boxplots of the 2018 and 2019 data side by side for each variable to look for outliers and possible differences.
    2. Check if the data was normally distributed with the Shapiro Test
    3. Perform an unpaired t test comparing the 2018 and 2019 data for each variable

Conclusion

All objectives were met. We were able to generate two clean datasets of SDOH variables that were specific to Utah at a Census Tract level and show that they could be used for statistical comparison.