Salomon Villatoro

2/18/2020

Final Dataset Selection

**Topic (Revised from last time):** Covid cases, death, and vaccine distributions across the state of Texas by race and income. (I originally wanted to do something including misinformation and how it may affect minority and immigrant populations from getting vaccinated, but found it difficult to find datasets like that).

**Datasets to be used:** For this topic, I want to take a look at a couple publicly available datasets around the state of Texas at the county level (ideally would want to find at the Zip Code level). Below are the datasets I wish to use and the methodologies behind them.

**ACS 5-year data at the county level:** For this specific dataset, I will use the tidycensus R package to download tables **DP05: Demographic and Housing Estimates** and **S19103: Median HH Income in the past 12 months.** The ACS (American Community Survey) is a survey run the U.S. census bureau every year to provide estimates across a range of variables in the U.S. population. Random household are selected to fill out a questionnaire with a variety of questions about social and economic needs.

From these datasets, the variables I will look at are:

* Race (Hispanic/Latino, White, Black, Asian, AI/AN, PI/NH, Other, two or more races). Provides a racial breakdown at the county level for each state.
* Median Household Income and Median household income by race; Provides the median income at the county level for each state.

**TEXAS COVID-19 Vaccine Allocation and Distribution:** This dataset was downloaded from the Texas Department of State Health Services, the data comes from vaccine data uploaded from healthcare workers the previous night, it is updated every day by 4:00 PM. Whenever I plan on running this analysis, the data will be the most up to date one**. (ON A SIDE NOTE)** I think it would be interesting to make a cross-country dashboard looking at cases, deaths, and vaccine distribution but wonder if it makes more sense to keep it to one state for now.

From this dataset, the variables I will look at are:

* County Name: Name of the county in Texas
* Total Doses Allocated; Number of doses promises to the aforementioned county as of data of data downloaded.
* Vaccine Doses Administered; Number of vaccine doses administered (1st and 2nd doses).
* People Vaccinated with at least One Dose
* People fully vaccinated; Number of people who have been fully vaccinated as of date of data downloaded

**Covid Tracking Project County by Race cases and deaths:** The dataset selected was downloaded from the Atlantic’s COVID-19 tracking project. It has been made in collaboration with Dr. Ibram X. Kendi at Boston University by compiling race and ethnicity data from states reporting several COVID-19 data categories. The dataset is split at the county level and reports back the two largest races/ethnicities of said county.

From this dataset, the variables I will look at are:

* County; Name of county
* Total; Total county population
* largestRace1; largest racial population in the county
* largestRace2; second largest racial population in the county
* cases; number of covid-19 cases in the county
* deaths; number of covid deaths in the county
* casesPer100k; number of covid cases per 100,000 people in said county
* deathsPer100k; number of covid deaths per 100,000 people in said county