COVID in the US: Identifying outliers

Identifying social factors that play an important role in the transmission of COVID-19 using County transmission and demographic data.

Gabriel Cruz
University of Maryland, College Park, MD, USA
E-mail: gcruz12@umd.edu

Motivation

Apart from its effect on the health and well being of the general population, COVID-19 has been able to demonstrate the disparities that exist to access to healthcare services (Killerby, 2020). Additionally, recent studies are demonstrating that there are certain factors that can tell us a lot about the impact that COVID-19 can have on the people diagnosed with it. (Mukherji, 2020). For example, it can be noticed that some of the most vulnerable populations include minorities in low-income communities or communities with low socioeconomic status. These communities find themselves receiving the brunt of the pandemic as they have to balance difficulties accessing quality healthcare and socioeconomic factors that keep them from practicing social distancing and healthy habits as recommended by the CDC.

One of the reasons for pushback to CDC guidelines is known to be low health literacy in adult populations. Because of this, it is a priority to try to educate vulnerable populations and increase viral awareness (Lakhani et al., 2020). Resources are finite, however, and we must be able to quickly identify communities where the transmission of COVID-19 does not fit in the correlation trend of COVID-19 and associated socioeconomic factors. By doing so we might be able to point out communities that demonstrate that they are outliers. It is important that we do because outliers would help us identify other factors that are not in the data that are similar to factors found in order communities.

Approach

In order to complete this project, there will need to be a considerable amount of data processing and management. There will be two datasets needed to complete this project. One of these datasets contains county-level data for socioeconomic factors that are specific to each county (*County-Level Data Sets - Data.Gov*, n.d.). The other dataset contains county-level data for coronavirus cases and deaths since the pandemic has begun (*US Coronavirus Cases and Deaths*, 2020). This data will need to be merged together as one single data set and then aggregated and saved as a separate dataset so that further implementations can see users choosing to interact with county-level information or state-level information. All data processing will occur with Python.

This data will be prototyped using Tableau. The idea is to use Tableau to create maps that will demonstrate proof of concept prototypes but to also determine what visual channels can be used to convey the information needed. In this case, it is most likely that color will be the only method to convey outliers, however, it is entirely possible that during this phase the size of a state can convey data abnormalities.

Once the prototyping has been completed then the map will need to be implemented with d3.js in order to create the user interaction functionality that is needed for this project. Users should be able to click on states in order to get more information about state-level data (filtered to that particular state) and they should be able to click on counties to get more information about the counties that they are interested in analyzing. Users should also be able to identify two states or counties that they would like to compare and pick them to analyze their differences. The dashboard will display test labels representing the socioeconomic data related to each county.

Milestones

<u>Data Preparation/Preprocessing:</u>
The data needs to be cleaned and prepared so that it is ready to be used for further implementations. Specifically, all of the data needs to be merged on the county level so that the fields that will be kept as part of the data source can be determined. Any data that is not recent or duplicated will be discarded in favor of only necessary data. A version of the data will be grouped by the state so that the user can interact with county level data as well as state level data.

Alpha:
The Alpha version of the dashboard will have the basic map visualizations for the county level implementation. Users should be able to click on a county and view which fields do not correlate to other counties on the state and national level (meaning which metrics are outliers in that one county). Users should also be able to pick which metric they are looking for outliers in.

Beta:
The Beta version of the dashboard will allow users to access the same functionality as the alpha with the addition of being able to aggregate the data based on the state instead of the county. Users should be able to zoom in and out of the view to focus in on counties.

Final:
The final version of the dashboard will add the ability for users to displace multiple metrics at the same time and compare two states in relation to each other. This should happen either by preserving the county abstraction or by aggregating to the state level. This is determined by the user.

Extensions

Further work in this area could involve tying in live data so that the data is tied in with real live data sources that can verify past or present information. Additionally, dashboards can be connected with other socioeconomic data or even geographical or human activity data. Having information on mobility or business data could add a deeper understanding of the data and could make it easier to identify trends or potential problem spots.

Tools

- Python
    - Used to aggregate and manipulate data.
- Tableau
    - Used to prototype and create low fidelity visualizations.
- D3.js
    - Used to create the interactive visualizations and the dashboards.

References

County-level Data Sets—Data.gov. (n.d.). Retrieved September 22, 2020, from https://catalog.data.gov/dataset/county-level-data-sets

Killerby, M. E. (2020). Characteristics Associated with Hospitalization Among Patients with COVID-19—Metropolitan Atlanta, Georgia, March–April 2020. *MMWR. Morbidity and Mortality Weekly Report*, *69*. https://doi.org/10.15585/mmwr.mm6925e1

Lakhani, H. V., Pillai, S. S., Zehra, M., Sharma, I., & Sodhi, K. (2020). Systematic Review of Clinical Insights into Novel Coronavirus (CoVID-19) Pandemic: Persisting Challenges in U.S. Rural Population. *International Journal of Environmental Research and Public Health*, *17*(12), 4279. https://doi.org/10.3390/ijerph17124279

Mukherji, N. (2020). The Social and Economic Factors Underlying the Incidence of COVID-19 Cases and Deaths in US Counties. *MedRxiv*, 2020.05.04.20091041. https://doi.org/10.1101/2020.05.04.20091041

US Coronavirus Cases and Deaths. (2020, September 22). USAFacts.Org. /visualizations/coronavirus-covid-19-spread-map