PPOL-564-Project-Proposal

Merykokeb Belay

10/25/2020

Project Overview

According to research conducted using racial demographic data from the The COVID Racial Data Tracker, people of color in the United States are disproportionately affected by the COVID-19 pandemic: "nationally, Black, Hispanic and Native American cases and deaths exceed their share of population"(12). The goal of this project is to use race and ethnicity COVID-19 data to assess which states have the highest and lowest reported infections by race and ethnicity as a share of population. Additionally, using data that quantifies states' legislative actions to maintain the Affordable Care Act, this project will examine if there is a relationship between a states' effort to maintain the Affordable Care Act and the reported number of cases and deaths by race and ethnicity.

Data Sources

The COVID Racial Data Tracker is a collaborative project between The Atlantic's The COVID Tracking Project and Center for Antiracist Research at Boston University. The COVID Racial Data Tracker, launched on April 12, 2020, collects the most up-to-date racial demographic data on COVID-19 related metrics (10). In an effort to show the racial disparities in the number of cases and deaths in the United States, the people behind this Tracker have also created a Racial Data Dashboard, which provides racial demographic metrics as a percentage of the population for each state that is providing race and ethnicity figures for these metrics. While racial demographic data was not initially part of most state's reporting of COVID-19, several states have recently started providing this data. There are currently 51 states/territories reporting positive cases and 50 states/territories reporting deaths by race and ethnicity(10). This data behind the dashboard, which is freely available on the Tracking Project's website, will be the primary source of data for this project (9).

In addition to The COVID Racial Data Tracker, the project will make use of data tracking the extent to which states have exerted effort to maintain and expand the Affordable Care Act. Healthinsurance.org is a website that routinely provides up-to-date information on all things related to health insurance and health policy in the United States. One of their blog entries from June 2019 provides an in-depth analysis of what each state is doing to preserve the Affordable Care Act. As part of this blog, a PDF is freely available–this PDF works almost like a score-card, tracking which provisions of the Affordable Care Act each state has been working to preserve and providing a tally for each state (3). This will serve as the secondary source of data for this project.

Lastly, demographic population data from the American Community Survey, obtained from the Kaiser Family Foundation website, will be used to obtain and calculate population data by race and ethnicity for each state and territory in The COVID Racial Data Tracker (7).

Data Collection Plan

The raw data used to build and update The COVID Racial Data Tracker is available in CSV format (9). This data will be collected by simply downloading the CSV file. Conversely, the PDF on the healthinsurance.org

website will be obtained by scraping the table from the PDF in the blogpost (3). Lastly, data from the American Community Survey will be scraped from the Kaiser Family Foundation website (7).

Methods

1. Web scraping: Python's 'docx' and ' BeautifulSoup' packages will be used to scrape data from the Affordable Care Act PDF and the American Community Survey population data, respectively (8)(2). The scraped tables will be saved as a pandas dataframe for wrangling and analysis (5).

2. Data wrangling: The pandas data frame will be the primary tool used to manipulate the data sets into a version that can be accurately visualized and analyzed. The unit of observation in The COVID Race Data Tracker and the Affordable Care Act PDF will be changed using pandas' built-in 'groupby'/'pivot_table' functions (5). The unicode character (bullet point: u"Ž022") in the PDF will be changed to numeric tallies to create a row-level sum for each state (11). For the Data Tracker, each state's name will be converted from a two letter code to its full name using the "usstateabbrev.py" python dictionary (1). Once all three data sets have the same unit of observation (i.e., state), pandas 'merge' functions will be used to create a comprehensive table which will under-go further customization (ex: renaming column names; changing data types; creating new variables based on existing ones) to prepare for visualization and analysis.

3. Data visualization: Both plotline and matplotlib will be used to create compelling visualizations to depict the story behind the data collected and analyzed (4)(6). Graphs will depict aggregate total number of cases by race and ethnicity for all states that have provided data. Additionally there will be graphs depicting the share of population by race and ethnicity and share of positive COVID-19 cases by race and ethnicity.The Affordable Care Act data will also be visualized: graphs will depict commonalities and differences in the provisions states are working to maintain and expand.

4. Machine (statistical) learning component: This project will use the data compiled and prepared to create a model assessing the relationship between states' perceptions of Affordable Care Act and racial disparities in COVID-19 cases and deaths.

Defining Success

For this project, success is defined as getting sufficient experience applying the lessons from class to real-life data and creating meaningful visualization to depict health disparities in COVID-19 cases and deaths. This project will be considered complete when the following have been done: successfully obtaining and saving data from a csv file, a PDF, and a website into a pandas data frame; creating a comprehensive data frame using pandas/numpy built in functions (groupby/ pivot table; melting; stacking; merging; performing column and row operations; converting data types); creating compelling visualizations using both matplotlib and plotninie to depict the story behind the numbers; and developing a model to assess the relationship between the dependent and independent variable mentioned above.

References

1. Allen, Roger. 'Us_state_abbrev.py'. GitHub: https://gist.github.com/rogerallen/1583593. Accessed 29 October 2020.
2. "BeautifulSoup4". PyPi: https://pypi.org/project/beautifulsoup4/. Accessed 28 October 2020.
3. Gaba, Charles. "What's YOUR state doing to save the ACA?". Healthinsurance.org. https://www.healthinsurance.org/blog/2019/06/07/whats-your-state-doing-to-save-the-aca/. Accessed 27 October 2020.
4. "Matplotlib 3.3.2". PyPi: https://pypi.org/project/matplotlib/. Accessed 29 October 2020.
5. "pandas.Dataframe". pandas: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.htm. Accessed 28 October 2020.

6. “plotnine”. PyPi: https://pypi.org/project/plotnine/.. Accessed 29 October 2020.
7. “Population Distribution by Race and Ethnicity”. Kaiser Family Foundation. https://www.kff.org/other/state-indicator/distribution-by-raceethnicity/?dataView=1&currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D. Accessed 27 October 2020.
8. “Python-docx 0.8.010”. PyPi: https://pypi.org/project/python-docx/. Accessed 28 October 2020.
9. “Race Data Entry”. The COVID Tracking Project at The Atlantic: https://docs.google.com/spreadsheets/u/1/d/e/2PACX-1vR_xmYt4ACPDZCDJcY12kCiMiH0ODyx3E1ZvgOHB8ae1tRcjXbs_yWBOA4j4uoCEADVfC1PS2jYO68B/pubhtml#. Accessed 28 October 2020.
10. “The COVID-19 Racial Data Tracker”. The COVID Tracking Project at The Atlantic: https://covidtracking.com/race/about. Accessed 28 October 2020.
11. “Unicode Character ‘BULLET’ (U+2022)”. FileFormat.Info. https://www.fileformat.info/info/unicode/char/2022/index.htm.Accessed 29 October 2020.
12. Wood, Daniel. “As Pandemic Deaths Add Up, Racial Disparities Persist — And In Some Cases Worsen”. National Public Radio. https://www.npr.org/sections/health-shots/2020/09/23/914427907/as-pandemic-deaths-add-up-racial-disparities-persist-and-in-some-cases-worsen. Accessed 27 October 2020.