Hemant Sathian

Professor Muller & Professor Flinsch

Information Visualization 04:547:321 Section 02

1 May 2022

Final Project Report

For my Final Project Report, I wanted to dwell deeper into the COVID-19 response rates of U.S States/Provinces and U.S Counties. With the COVID-19 pandemic becoming a global issue over the past two years, much of the COVID-19 data that is available focuses solely on the summary statistics of an entire country as opposed to the specific jurisdictions that make up a country. Dwelling deeper into the COVID-19 response rates of specific U.S States/Provinces and U.S Counties would allow me to uncover particular trends that would not be seen when looking at the U.S's COVID-19 response as a whole. I will state the data sources I used for my report, the challenges I faced regarding the data curation process and analysis aspect of my report, and the key findings obtained from my report.

The data sources I used for my report included a Kaggle dataset, a dataset obtained through the data source, 'data.world', and a dataset obtained through the United States Census Bureau. The dataset obtained from Kaggle displays the daily number of COVID-19 cases and deaths by a specific U.S State and a specific U.S County. The Kaggle dataset was created by Sudalai Rajkumar, "considered one of the top data scientists in India by Analytics India Magazine and Jigsaw Academy". Given that Kaggle datasets are commonly used for university projects and the pedigree of Sudalai Rajkumar's work, I had no qualms regarding the reliability and ethicality of the dataset. Much like the Kaggle dataset by Rajkumar, the dataset obtained through the data source, 'data.world', also displays the number of COVID-19 cases and COVID-19 deaths for a specific U.S State and a specific U.S County, but by specific report dates. What I liked about the 'data.world' dataset was the geographic role certain columns in the dataset could play in Tableau. The geographic component of some of the columns in the 'data.world' dataset would enable me to create geographical heat maps in my report. The last dataset I would utilize in my report would be the 2020 Census Apportionment Results dataset obtained through the United States Census Bureau. The 2020 Census Apportionment Results dataset enabled me to obtain individual state populations in 2020. Utilizing state population data from 2020 would allow me to create calculations using columns of separate datasets. However, before I could use the dataset obtained through the United States Census Bureau, I would have to face the challenge of cleaning it.

When creating my report, the first challenge I encountered consisted of the data validation and data reconciliation efforts on the 2020 Census Apportionment Results dataset. Data validation efforts I made on the 2020 Census Apportionment Results dataset included removing the "NUMBER OF APPORTIONED REPRESENTATIVES BASED ON 2020 CENSUS" column and the "CHANGE FROM 2010 CENSUS APPORTIONMENT" column from the dataset. The only columns in the 2020 Census Apportionment Results dataset that would be useful in my analysis would be the "APPORTIONMENT POPULATION" column and the "STATE" column. Data reconciliation efforts on the 2020 Census Apportionment Results dataset consisted of removing any whitespaces or symbols from the "STATE" column. Ensuring the "STATE" column was clean in the 2020 Census Apportionment Results dataset would allow me to blend the dataset with Rajkumar's Kaggle dataset in Tableau to create aggregate calculations. Another challenge I encountered when creating my report included problems with the summation of the data across most of my charts. As mentioned by Professor Flinsch, certain columns for a dataset may be running totals for a variable up till a certain date as opposed to the total count for one day. To ensure I was utilizing the correct measure functions in Tableau for specific columns, I made an effort to research the nature of each column for the datasets I utilized. For instance, the "Cases" column in the Kaggle dataset by Rajkumar consisted of a summation of the number of positive cases by a certain date. I would use the 'MAX' Tableau function on the "Cases" column to ensure I was obtaining the actual total of positive COVID-19 cases. Having solved many of the issues that would have prevented me from obtaining accurate results in my report, I could then analyze the key findings I discovered.

The first part of my analysis consisted of analyzing U.S states/provinces with the highest amount of COVID-19 deaths in 2020. From a geographical heat map, I deducted that I should primarily focus on the states California, Texas, New York, and Florida. Through a horizontal bar chart, I discovered that California had the highest amount of COVID-19 deaths in 2020, with Texas, Florida, and New York following, respectively. While it was interesting that California led all states in COVID-19 deaths, this trend inherently suggested that California had the highest population of any other state as opposed to "it was the most affected state by COVID-19". To answer which of the ten states was the most affected by the COVID-19 pandemic, I would create another horizontal bar chart that would utilize a calculation developed by blending the 2020 Census Apportionment Results dataset with the Kaggle dataset by Rajkumar. The calculation created from blending the two datasets would give me the death rate for the ten states with the most COVID-19 deaths. Having implemented the calculation in a horizontal bar chart, I discovered that New York had the highest death rate of any of the ten states, with nearly 12.1 percent of its population dying due to COVID-19. However, what surprised me the most was California's relatively low death rate in comparison to the amount of its COVID-19 deaths, with only 2 percent of its population having died due to COVID-19. The next step in my analysis was to analyze ten U.S counties with the most positive COVID-19 cases in 2020 and the ten U.S counties with the most COVID-19 deaths in 2020 with horizontal bar charts. Much like the story portrayed by the horizontal bar chart that visualized the ten U.S states/provinces with the most COVID-19 deaths, Los Angeles County and New York City County (counties in California state and New York state, respectively) held the most positive COVID-19 cases. Interestingly enough, despite having more positive COVID-19 cases than New York City Country by a large margin, a horizontal bar chart displaying ten U.S counties with the most COVID-19 deaths would show Los Angeles County having substantially fewer COVID-19 deaths in comparison to New York City County. To analyze the discrepancy in COVID-19 cases and COVID-19 deaths between Los Angeles County and New York City County, I would utilize area charts to implement time-series analyses. The area chart depicting the sum of COVID-19 cases by day for Los Angeles County and New York City County would show New York City County having substantially more daily positive COVID-19 cases than Los Angeles County from approximately March 23rd, 2020 until approximately August 28th, 2020. Los Angeles County would eventually exceed New York City County's amount of positive COVID-19 Cases from approximately August 28th, 2020, through December 1st, 2020. Due to New York City County having substantially more daily positive cases than Los Angeles County from approximately March 23rd, 2020, until approximately August 28th, 2020, that could be a possible reason as to why New York City County leads Los Angeles County in COVID-19 deaths by a considerable margin. Justifications for the results I found in my report could be made by John Woolfolk, a reporter for the Bay Area-based newspaper, The Mercury News. In the article, "COVID: What do California, Texas, New York and Florida have in common? Stunningly low infection rates", Woolfolk provides reasons as to why California may have low COVID-19 deaths totals. One reason for California's low COVID-19 death totals could be California having a high vaccination rate, with nearly 54 percent of its total population having received at least one shot (Woolfolk, 2021). Additionally, in June 2020, California required face masks for everyone down to age 2 in most settings outside the home (Woolfolk, 2021). California has also been one of the slowest states to reopen schools fully, with less than one in four districts operating entirely in-person rather than partly or wholly remote as of May 2021 (Woolfolk, 2021). California's hesitance in scaling down preventative action against the COVID-19 pandemic may be the reason as to why Los Angeles County and California state as a whole has not suffered as many COVID-19 deaths as New York state and New York City County.

To summarize, the final project enabled me to discover key trends regarding the COVID-19 pandemic at the U.S state/province and U.S county level. The data sources I utilized allowed me to uncover key trends regarding COVID-19 cases and COVID-19 deaths in U.S states/provinces and U.S counties and allowed me to create calculations in order to reveal trends not directly evident in the dataset. The challenges I faced in the data validation, data reconciliation, and data analysis processes allowed me to obtain new knowledge regarding data visualization creation in Tableau. By utilizing informative datasets and meeting the challenges of my data analysis, I uncovered differences between COVID-19 cases and COVID-19 deaths in New York state and California state, as well as in New York City County and Los Angeles County.

References

Coronavirus (COVID-19) Data Hub. (2020). COVID-19 Activity [Data file]. Retrieved May 1, 2022, from <https://data.world/covid-19-data-resource-hub/covid-19-case-counts/workspace/data-dictionary>

Rajkumar, S. (2020). COVID-19 in USA (Version 102) [Data file]. Retrieved May 1, 2022, from <https://www.kaggle.com/datasets/sudalairajkumar/covid19-in-usa?select=us_counties_covid19_daily.csv>

United States Census Bureau. (2021). Table 1. Apportionment Population and Number of Representatives [Data file]. Retrieved May 1, 2022, from <https://www.census.gov/data/tables/2020/dec/2020-apportionment-data.html?utm_medium=email&utm_source=govdelivery>

Woolfolk, J. (2021, May 22). COVID: What do California, Texas, New York and Florida have in common? Stunningly low infection rates. *The Mercury News*. Retrieved May 1, 2022, from <https://www.mercurynews.com/2021/05/22/covid-what-do-california-texas-new-york-and-florida-have-in-common-stunningly-low-infection-rates/>