Visualizing the COVID-19 Story in America

Jeremy McCormick

DS 745: Visualization and Unstructured Data Analysis

Fall Semester 2022

Instructor: Lauren Mauel

October 31, 2022

1 Introduction

COVID-19 caused unprecedented disruptions across the world, affecting nearly every aspect of people's lives. First reported in China around the beginning of January 2020, it quickly spread throughout the world and developed into the most serious and far reaching disease pandemic of modern times. Responses varied by country, with some responding quickly and effectively and others more slowly and haphazardly. In the United States, the federal system of government resulted in responses that varied widely by state. While there were two federally mandated lockdowns restricting most activity, states had large discretion in how they implemented these measures and when they were imposed and lifted. Failures within the U.S. medical and political systems to the pandemic have been widely acknowledged and reported in the mainstream media (Yong, 2020). The decentralized and heterogeneous nature of America's political and medical systems lead to initial deficiencies in testing, treatment, PPA use (mask-wearing), vaccine availability and so on, all of which varied from state-to-state. Demographics across the United States also vary depending on the state, with some groups being more vulnerable than others to the effects of the virus.

This visualization study attempts to tell the story of how coronavirus affected America at the state level. While there have been a wide variety of different statistics reported in this domain, three primary variables broken down by state will be used: cases, deaths, and vaccinations. These are normalized by population where appropriate from the raw values in order to facilitate inter-state comparisons. The data were gathered from several different sources and loaded into Tableau where aggregate calculations were performed. A number of different visualizations were used to provide insight into these statistics, and a Tableau story was authored including all of them along with descriptive captions.

2 The Data Sources

Data files were downloaded locally and then loaded into a Tableau workbook using the text file option. Three different sources were used: one containing both case and death counts, another with vaccination counts and rates, and then, finally, a table of U.S. population counts was included. A supplementary file of U.S. state abbreviations (not cited) was used for purposes of labeling data points by two-letter code rather than full state name. All of this data was gathered from the following sources around October 23, 2022.

2.1 Vaccinations

Vaccination statistics were retrieved in comma-separated value (CSV) format from a COVID-19 data repository maintained by a well-known organization which focuses on philanthropic causes (Mathieu & Ortiz-Ospina, 2021). Data files were downloaded from the organization's Github project page (Our World in Data, 2022a). The "Vaccinations against COVID-19" data used by this study were collected by them from official reports. The file *us_state_vaccinations.csv* was downloaded from the sub-page containing vaccination data (Our World in Data, 2022c). This file is updated daily based on information from the COVID Data Tracker of the United States Centers for Disease Control and Prevention (CDC, 2022). This analysis uses the following subset of columns from this dataset:

- **location:** The name of the state or federal entity
- **date:** The date of the observation in *YYYY-MM-DD* format
- **people_fully_vaccinated_per_hundred:** People fully vaccinated per 100 in the state population

Only data from the 50 U.S. states was used and the records from other entities such as territories (Guam, Puerto Rico, etc.) were filtered out. The *people_fully_vaccinated_per_hundred* field in this table is a running daily total at a state-by-state level of how many individuals out of a hundred have received "all doses prescribed by the initial vaccination protocol per 100 people in the total population" (Our World in Data, 2022b). Receiving just a single dose was not counted as a vaccination for purposes of this analysis. The vaccination rate for each state was retrieved from its corresponding record for October 12, 2022, the last date from which this information was available in the data repository. This value was used as the "final" vaccination rate value for analysis and visualization.

2.2 Cases and Deaths

Case and death counts were retrieved in CSV format from a data repository published by the New York Times (NYT, 2022). This is the same data that is used for its online maps and graphs that track the Coronavirus pandemic (NYT, 2022). The specific file was *us-states.csv* from the rolling averages datasets (NYT, 2022). The following columns were used in the analysis:

- **state:** Name of the U.S. state
- **date:** Date of the record in *YYYY-MM-DD* format
- **cases:** New coronavirus cases reported in the state on that day
- **deaths:** Deaths reported in the state on that day

Though this file has rolling averages of per capita statistics, it was decided to instead compute these for cases and deaths based on totalling all the measurements for a single state and then normalizing by the state population value from another data source.

2.3 State Population

Finally, state population data was used to normalize the counts by state in order to make them comparable. Official U.S. census data for 2020 was downloaded in CSV format for this purpose (United States Census Bureau, 2021). The following fields were used from this data:

- **NAME:** The name of the U.S. state, matching those from the other datasets
- **POPESTIMATE2020:** The 2020 state population estimate

For simplicity's sake, only the 2020 populations numbers were used in the analysis. A more rigorous calculation would have tallied deaths by a specific year. Since 2022 population numbers were not available, it was decided to instead use the 2020 values as a baseline. This may introduce inaccuracy in the normalized values but most likely only at the 1-2% level, as year-to-year fluctuations tend to be small relative to the total population, and less than three years of data were used for the analysis, those being 2020, 2021 and part of 2022.

3 Data Preparation and Analysis

Tables were linked via their state and date fields. Some of the sources such as the vaccination data included records from non-state entities such as U.S. foreign territories. These were excluded from analysis using a filter on each worksheet. The case and death rates were computed within the workbook, whereas the vaccination rate was read from the newest record in the data for each state, as it was measured for every date.

These fields were used without significant transformation from the original data sources besides renaming them:

- **State Code:** Two-letter state abbreviation
- **State Name:** Full name of the state
- **Date (State Deaths):** The date on which a state death count was reported
- **State (Deaths):** The state in which a death count was reported
- **Cases:** The detailed number of cases by state and date
- **Deaths:** The detailed number of deaths by state and date
- **Population 2020:** Population of the state from the 2020 census data
- **Date (Vax):** Date of a vaccination record
- **Location (Vax):** Location of a vaccination record (e.g. U.S. full state name)
- **Daily Vaccinations:** The number of vaccinations performed by state and date
- **People Fully Vaccinated Per Hundred:** Running calculation of those fully vaccinated per hundred people in a state

The following calculated fields were derived from the initial data:

- **Cases Per 100k:** The total number of cases in a state multiplied by 100000 and then divided by its population
- **Deaths Per 100k:** The total number of deaths in a state multiplied by 100000 and then divided by its population
- **Max Vaccinations:** The maximum vaccination rate achieved by each state
- **Min Vaccinations:** The minimum vaccination rate achieved by each state
- **State With Max Cases:** The single state with the highest case rate
- **State With Min Cases:** The single state with the lowest case rate
- **State With Max Deaths:** The single state with the highest death rate
- **State With Min Deaths:** The single state with lowest death rate
- **State With Min Vaccinations:** The single state with the lowest vaccination rate
- **State with Max Vaccinations:** The single state with the highest vaccination rate
- **State Total Cases:** The total number of cases by state
- **State Total Deaths:** The total number of deaths by state

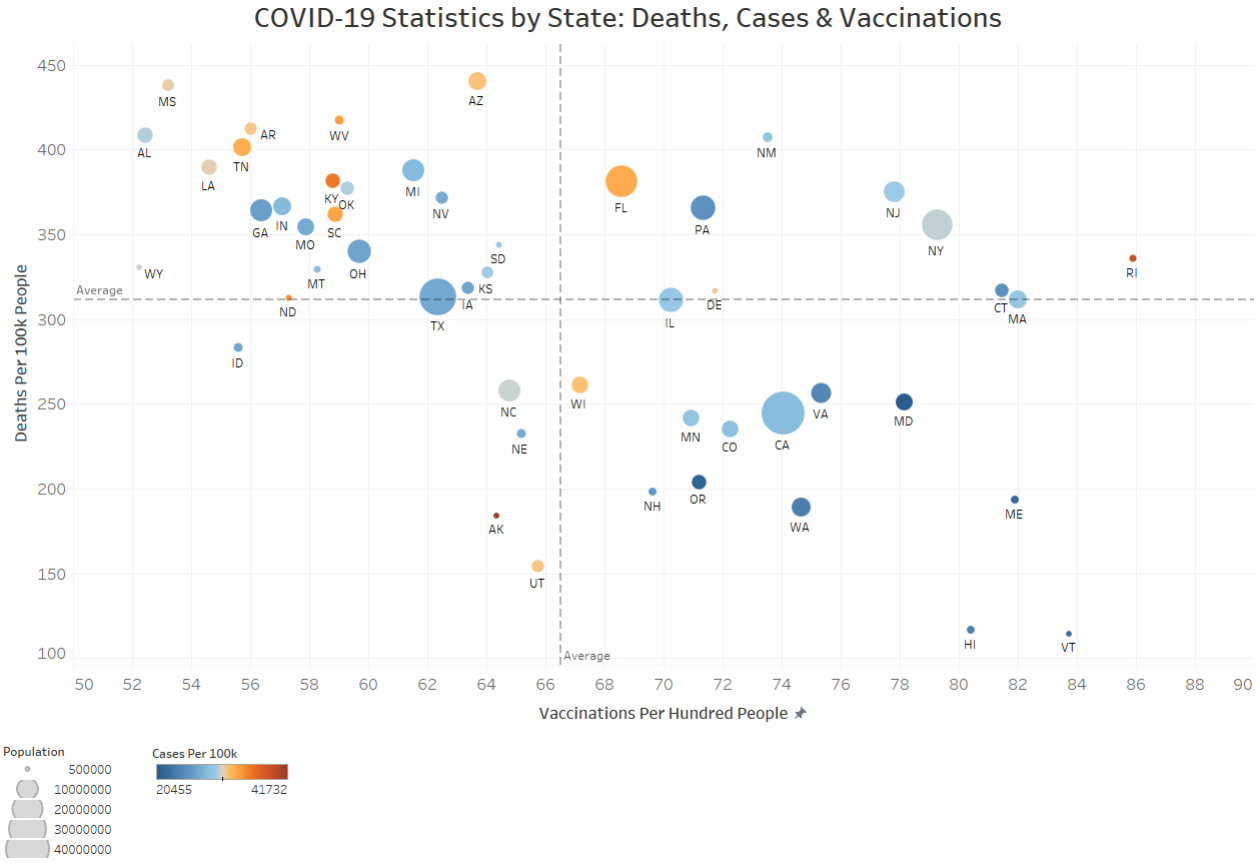
A filter called **US States Only** excluded those geographical records which did not pertain to the analysis.

4 Visualizations

There are four main visualizations comprising the project, plus a Tableau story including all of them in sequence. These were composed within a single Tableau workbook called *COVID Story.twbx* and are described in detail below.

4.1 Scatter Plot

Figure 1: Scatter Plot



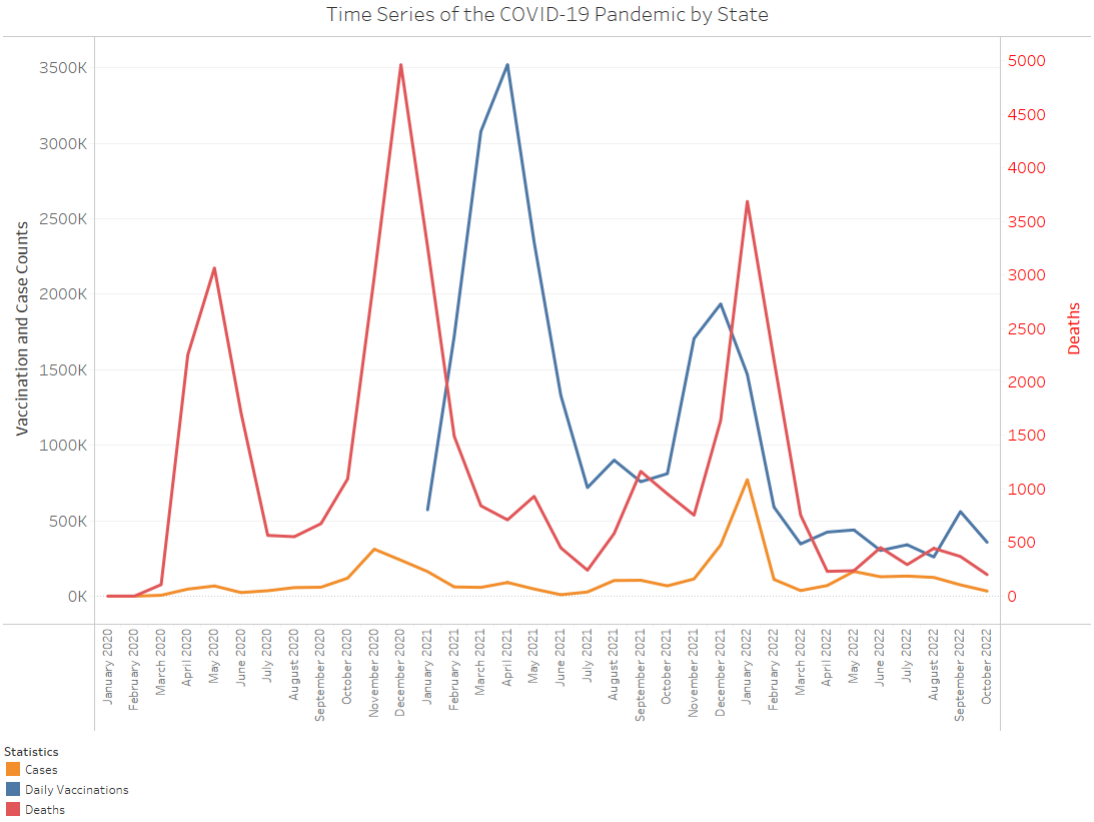
This plot shows rates of vaccination, cases, and deaths for each U.S. state, labeled with its two-letter abbreviation. Deaths and cases are normalized per one hundred thousand people while vaccinations are per one hundred. The mark size is proportional to the state's population, and the mark color denotes the case rate, which is likely not estimated with a great amount of accuracy due to different reporting requirements and systems by state. The data is based on accumulated totals from the beginning of the pandemic in early 2020 through an end date in October 2022. The state populations used for the death rate calculations

are from the 2020 census. The case and death rates have a margin of error based on possible population changes from 2020 to 2022 which were not accounted for in the analysis. The dashed grey line represents the averages for the X and Y axes (vaccination and death rates, respectively). There is a notable cluster of states with high death rates in the upper lefthand quadrant which corresponds to below average vaccination rates. This visualization shows five different dimensions compactly within a single plot: vaccination rate, case rate, death rate, state, and population. Those states with above or below average death or vaccination rates can be easily identified by their relationship to the dotted grey line. The code annotations on the individual marks allow individual states to be easily identified.

As far as possible deficiencies, the chart may not be understand immediately or ”at a glance” since it encodes information using so many different methods. It may take some examination by the viewer to fully grasp how the information is being presented. The placement of deaths and vaccinations on the two axes was somewhat arbitrary, as this could have instead been case and death rates with vaccinations denoted by the color encoding, giving a different overall effect and interpretation of the data. The different sized marks denoting population are perhaps difficult to interpret very accurately visually, though it is clear which states have the highest and lowest populations. Determining the actual population accurately from the graph is not really possible compared with looking at values on an axis.

4.2 Time Series

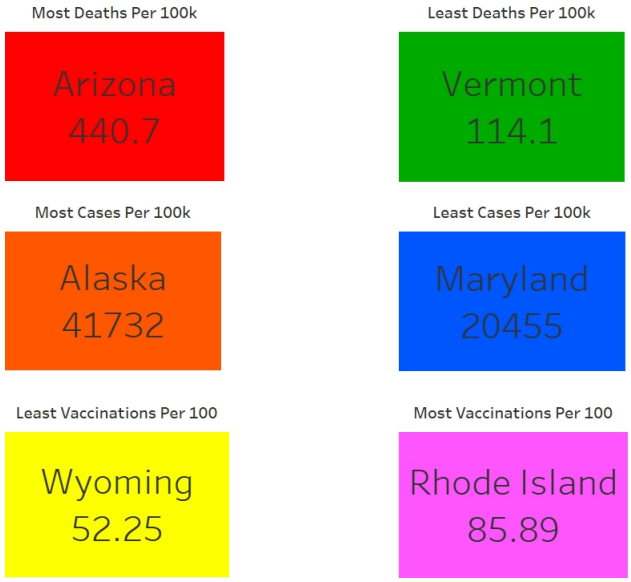
Figure 2: Time Series



This time series graph shows aggregated statistics by month for a selected state. Vaccinations and cases use one axis while deaths uses a separate one, which is colored in red to make this distinction obvious. The different variables can be distinguished by their color. Spikes in cases and deaths can be correlated with each other. The number of vaccinations starts in January 2021 and shows several peaks corresponding with major vaccination campaigns and follow-up boosters. The graph shows single vaccine doses rather than fully vaccinated people as shown on other plots. Not pictured on figure 2, but present in the workbook, is a drop-down menu for selecting a single state, which will update the graph to display only its data. There are several disadvantages of displaying the data in this fashion. Raw numbers may be somewhat difficult to compare between states. Instead, rates could be displayed by month. Having a separate axis for deaths means that this is displayed out of scale with the other two measurements and may seem distorted. Finally, this visualization scheme does not allow comparison of states side-by-side. Rather, the selection box has to be used in order to display only a single state's information.

4.3 Scorecard

Figure 3: Scorecard

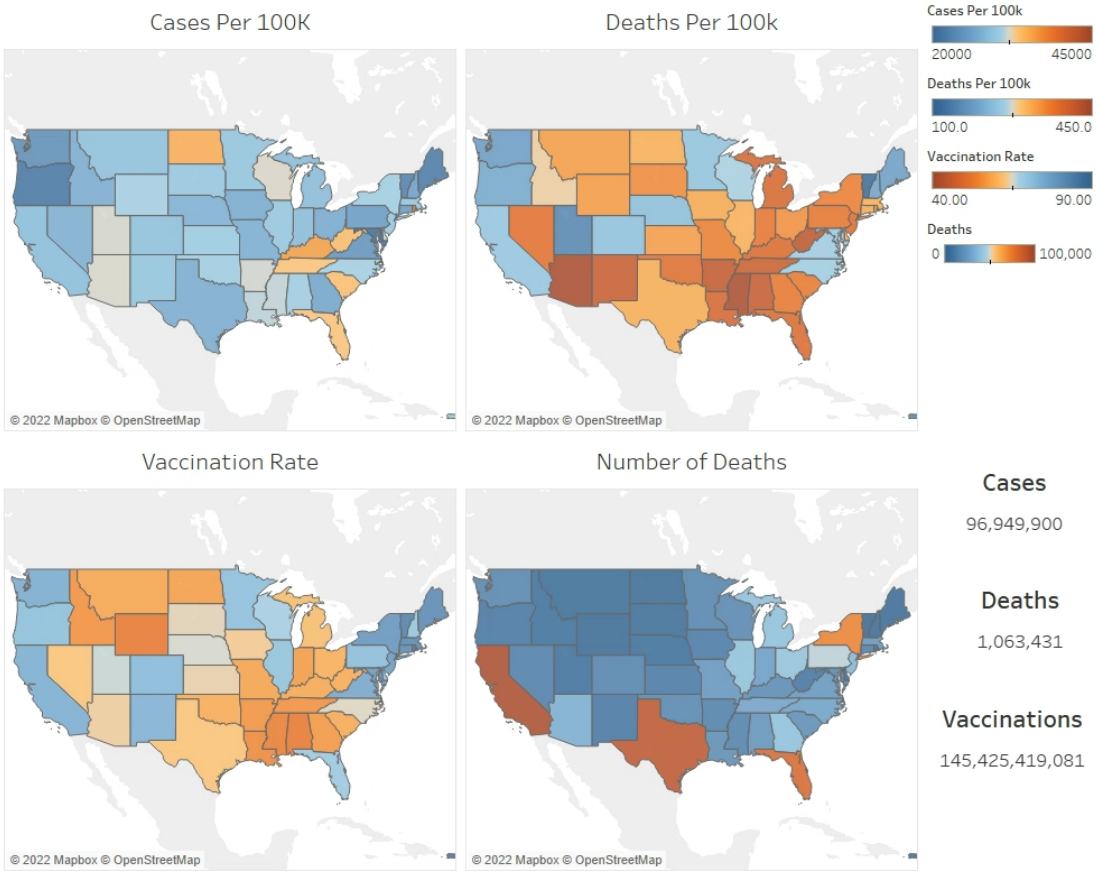


The scorecard shows the best and worse states judged by specific metrics. These are displayed as rates per some amount of population as a way of normalizing the comparisons. This scheme is a simple but powerful way of showing the greatest contrasts between states, as these are the minimum and maximum values. For instance, it is evidence that Arizona's death rate was over three times that of Vermont, the state with the least deaths. Similarly, Alaska reported twice as many cases per one hundred thousand people as Maryland, and Wyoming had a vaccination rate of around 52% while Rhode Island's was almost 86%. These contrasts tell an interesting story about the differences in effects and responses during the pandemic.

Showing the data in this way could be considered overly simplistic, as it boils down the ideas of best and worst to a single number. So there is possibly a lack of background information or nuance. This dashboard does not attempt to explain why the numbers have these values. Arizona, for instance, is a popular state for retirees and the elderly were at considerably elevated risk of death from coronavirus compared with the general population (CDC, 2021). Also, from figure 1, it is evident that Arizona has a well-below average vaccination rate, which may have contributed to an elevated number of deaths. Due to the lack of context, this particular visualization does not tell a thorough story. The case numbers may also not have been uniformly reported in all states, making that particular statistic inaccurate.

4.4 Color Maps

Figure 4: Color Maps



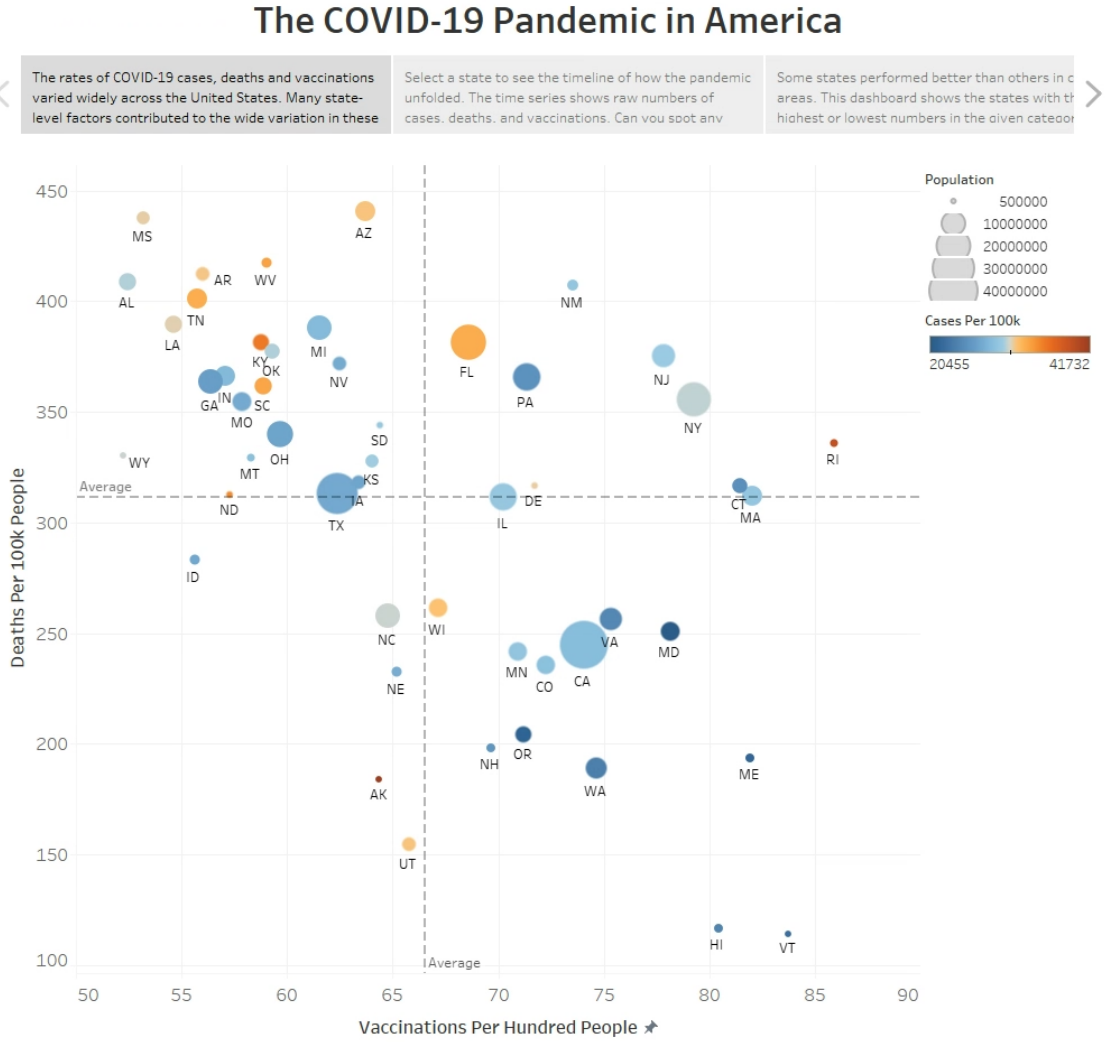
This dashboard shows several different coronavirus statistics mapped to colors on a U.S. map by state. A divergent, blue-orange color palette was used in order to highlight contrasting information. Values in the mid range are greyish; low ones are shades of blue; high ones are shades of orange. The statistics displayed on each of the four maps are cases and deaths per one hundred thousand people, people fully vaccinated per hundred (similar to a vaccination rate percentage), and, finally, the raw number of deaths to illustrate the magnitude of the pandemic's worse effects on the country. A legend for each map is displayed in the upper right. In the lower right-hand corner, summary statistics are displayed for numbers of cases, deaths, and vaccinations. Hovering over a specific state shows its value for that variable. These maps are a powerful and compact way of showing the impact of the pandemic, as well as illustrating contrasts between states via the color scaling.

These maps may be at a somewhat coarse level, geographically. Similar online visualizations produced by journalistic organizations often show data at the county level. Urban and rural areas had varying re-

sponses to the pandemic, and showing data at the state level obscures some of these differences, especially for states that have large population centers as well as rural areas. The ”Number of Deaths” map uses the raw number of deaths by state, so there is a strong correlation with population, e.g. California, Texas, Florida, and New York are highlighted in orange. But they are also the most populous states. The color scaling does not work particularly well since outside of these states, the rest are all colored blue.

4.5 The COVID-19 Story

Figure 5: The COVID-19 Story



Finally, all four visualizations are brought together into a Tableau story. These are shown in the order that they were presented above. Each visualization has a caption describing its purpose, which prompts the viewer to ask questions about what they are seeing. This follows the single-page paradigm that is often

used as a guideline for dashboard design. Whatever flaws are present in these visualizations are obviously carried over into this story.

5 Conclusion

Coronavirus had wide-reaching and devastating effects on the United States which varied state-to-state based on many factors. These visualizations attempt to tell just part of this story by highlighting differences in case, death, and vaccination statistics using various techniques. Hopefully, these visualizations help the viewer gain insight into this monumental event and spark their curiosity for engaging in further investigation and research in this area.

References

CDC. (2021). *Covid-19 risks and vaccine information for older adults.* Retrieved 2022-10-26, from https://www.cdc.gov/aging/covid19/covid19-older-adults.html

CDC. (2022). *Covid data tracker.* Retrieved 2022-10-25, from https://covid.cdc.gov/covid-data-tracker/

Mathieu, E., & Ortiz-Ospina, H. (2021). A global database of covid-19 vaccinations. *Nat Hum Behav*. Retrieved from https://doi.org/10.1038/s41562-021-01122-8

NYT. (2022). *Cases and deaths rolling averages and anomalous days.* Retrieved 2022-10-26, from https://github.com/nytimes/covid-19-data/tree/master/rolling-averages

NYT. (2022). *Coronavirus (covid-19) data in the united states.* Retrieved 2022-10-23, from https://github.com/nytimes/covid-19-data

NYT. (2022). *Coronavirus in the u.s.: Latest map and case count.* Retrieved 2022-10-26, from https://www.nytimes.com/interactive/2021/us/covid-cases.html

Our World in Data. (2022a). *Covid-19 dataset by our world in data.* Retrieved 2022-10-23, from https://github.com/owid/covid-19-data

Our World in Data. (2022b). *Data on covid-19 (coronavirus) by our world in data.* Retrieved 2022-10-25, from https://github.com/owid/covid-19-data/tree/master/public/data

Our World in Data. (2022c). *Data on covid-19 (coronavirus) vaccinations by our world in data.* Retrieved 2022-10-25, from https://github.com/owid/covid-19-data/tree/master/public/data/vaccinations

United States Census Bureau. (2021). *State population totals and components of change: 2020-2021.* Retrieved 2022-10-23, from https://www.census.gov/data/tables/time-series/demo/popest/2020s-state-total.html

Yong, E. (2020). How the pandemic defeated america. *The Atlantic*. Retrieved from https://www.theatlantic.com/magazine/archive/2020/09/coronavirus-american-failure/614191/