Mapping Mortality: A Statistical Insight into COVID-19 Spread and Impact

Project Report

By
Shreshika Bommana

April 12, 2025

Contents

Abstract

This study is a comprehensive visual and interactive analysis of COVID-19 trends in the United States, using both static and dynamic visualizations to show meaningful patterns in confirmed cases and deaths. Using publicly available datasets, the analysis begins with preprocessing and time-series aggregation, where national-level daily data is refined to compute 7-day moving averages for both cases and deaths. The data span portions of 2020, with reporting periods ranging from early January to late August. A series of visualizations are generated to chart the evolution of total confirmed cases, total deaths, and their respective trends using smoothed averages. One key insight explored is the temporal behavior of the death-to-case ratio, showing shifts in lethality or reporting throughout the pandemic. To explore regional disparities, the analysis identifies the top 10 U.S. states by cumulative COVID-19 deaths, visualized using horizontal bar plots. All visualizations are saved and organized for accessibility. This work demonstrates a methodical approach to exploratory data analysis and interactive storytelling with COVID-19 data.

Introduction

The emergence of COVID-19 in late 2019 led to an unprecedented global health crisis. In the United States, the virus spread rapidly, resulting in millions of infections and hundreds of thousands of deaths. The need for timely and accurate data analysis became necessary to inform public health decisions, track virus transmission, and allocate medical resources efficiently. This study aims to analyze the pandemic's trajectory in the U.S. by using data visualization techniques that provide a comprehensive and accessible representation of the crisis. By using various statistical methods, the study seeks to create a narrative that shows the severity, spread, and impact of COVID-19 in the country.

The primary objective is to analyze the COVID-19 outbreak in the United States by looking at trends in confirmed cases and fatalities. By creating interactive visualizations and statistical models, this study highlights significant patterns in the data, such as surges in infections, mortality trends, and variations across different states. Additionally, this report aims to illustrate how factors like time and regional differences influenced the severity of the outbreak. Beyond descriptive analysis, the findings aim to provide insights into the effectiveness of public health measures and give a way to use data visualization in crisis management.

Dataset Description

The dataset used in this study is sourced from the publicly available repository titled "COVID-19 Analysis and Visualization" published on Kaggle by Subhojit Paul. It comprises multiple CSV files containing aggregated and time-series data related to the COVID-19 pandemic in the United States during the year 2020. This analysis specifically utilizes two components of the dataset: a time-series file tracking confirmed cases and deaths over time by state (covid.csv), and a file summarizing weekly death counts by jurisdiction (us_deaths_covid.csv).

The covid.csv file contains daily cumulative counts of confirmed COVID-19 cases and deaths across various U.S. states and territories. The earliest entry in this file is dated January 22, 2020, and the latest is July 27, 2020. Each row in the dataset corresponds to a U.S. state on a specific day and includes key fields such as the number of confirmed cases and deaths reported up to that date. This format enables the temporal analysis of case progression and fatality counts across states.

The us_deaths_covid.csv file complements the daily data by aggregating COVID-19 deaths into weekly intervals. This dataset begins on February 1, 2020, and ends on August 29, 2020. Each row represents a specific jurisdiction (usually a state) and week, with columns indicating the number of deaths reported during that period. The file includes fields such as the "Start Week," "End Week," "COVID-19 Deaths," and potentially total deaths, allowing for weekly trend analysis and cross-comparisons between jurisdictions.

Methodology

To prepare these datasets for analysis, several preprocessing steps were carried out using Python’s pandas, numpy, and datetime libraries. Date fields such as “Date”, “Start Week”, and “End Week” were explicitly converted to datetime objects to ensure temporal alignment. Null or missing values were handled through filtering or by excluding records with unresolved inconsistencies. Grouped aggregations were performed to compute daily and weekly summaries, allowing for trend analysis, such as identifying peaks in case counts and assessing the stability of death reporting over time.

Exploratory data analysis and visualization were conducted using matplotlib and seaborn to create clear and informative charts. To better understand long-term trends and mitigate daily fluctuations, moving averages were applied over rolling time windows (e.g., 7-day averages), showing underlying patterns that are obscured in raw data. Additional transformations included the calculation of case-to-death ratios, which were used to assess possible differences in healthcare strain, testing coverage, or reporting practices across states.

Exploratory Data Analysis

An initial examination of the data focused on understanding the progression of COVID-19 cases in the US during the early stages of the pandemic. Case numbers were relatively low during the earlier months; for instance, fewer than 100 total cases were reported nationwide through February 2020. However, a sharp increase occurred beginning in March 2020, with the country surpassing 100,000 confirmed cases by late March. This growth continued rapidly, and by late July 2020, the cumulative case count had exceeded 4.2 million.

The data revealed several distinct surges in case numbers, with notable acceleration between mid-June and the end of July 2020. This period corresponded with broader reopenings and increased mobility across many states. In terms of deaths, a lagged but correlated pattern emerged. The national death toll crossed 100,000 in late May and climbed steadily through the summer months, reaching over 146,000 by the end of July.

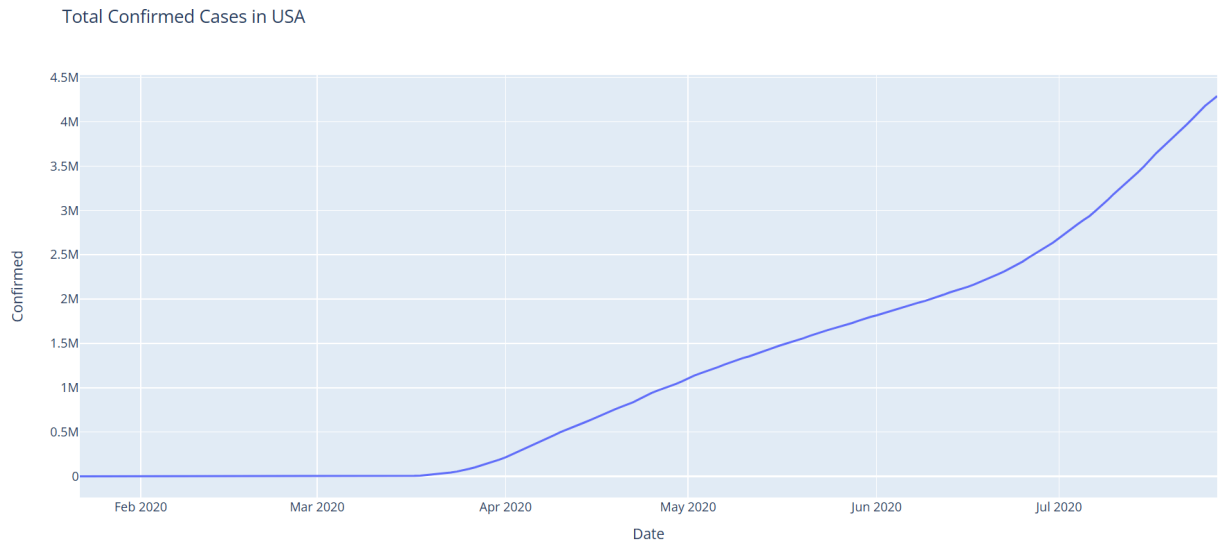

Figure 1: Total Confirmed COVID-19 Cases in the United States

An interactive line chart showing the cumulative increase in confirmed COVID-19 cases across the country, based on daily data from January 22 to July 27, 2020. (See: usa_total_cases.html for full interactive version)

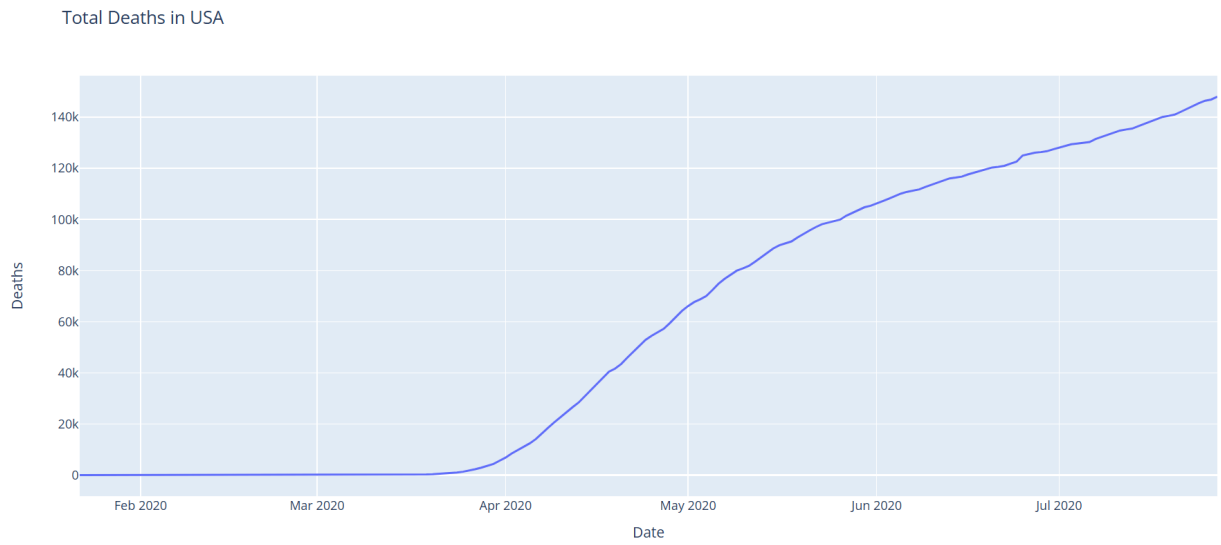

Figure 2: Total COVID-19 Deaths in the United States
An interactive line chart visualizing the cumulative death toll from COVID-19 in the country, from January 22 to July 27, 2020. (See: usa_total_deaths.html for full interactive version)

Moving Average and Trend Analysis

Moving averages and comparative trend visualizations were used to help decrease fluctuations in reporting and give a clearer picture of how infections and deaths evolved over time. The resulting visualizations support a broader narrative of spread, mortality, and the dynamic relationship between cases and deaths.

Smoothed Trends in Cases and Deaths

A 7-day moving average was applied to daily new confirmed cases and deaths to smooth short-term noise and reveal longer-term trends. The resulting chart (Figure 3) shows two distinct waves of case acceleration. The first rise begins in late March 2020 and peaks around April 10, coinciding with the initial national lockdown measures. This is followed by a decline through May. A second and more significant rise emerges in June, with daily new cases peaking in early July at over 60,000. This resurgence is closely linked to easing restrictions across several states.

In contrast, the trend in deaths follows a delayed but recognizable pattern, with the first peak in mid-April and a slower tapering afterward. The second wave of deaths begins to rise in July but does not reach the same steepness as the second wave in cases — a discrepancy that may point to improved treatment, younger infected populations, or changes in testing and reporting.

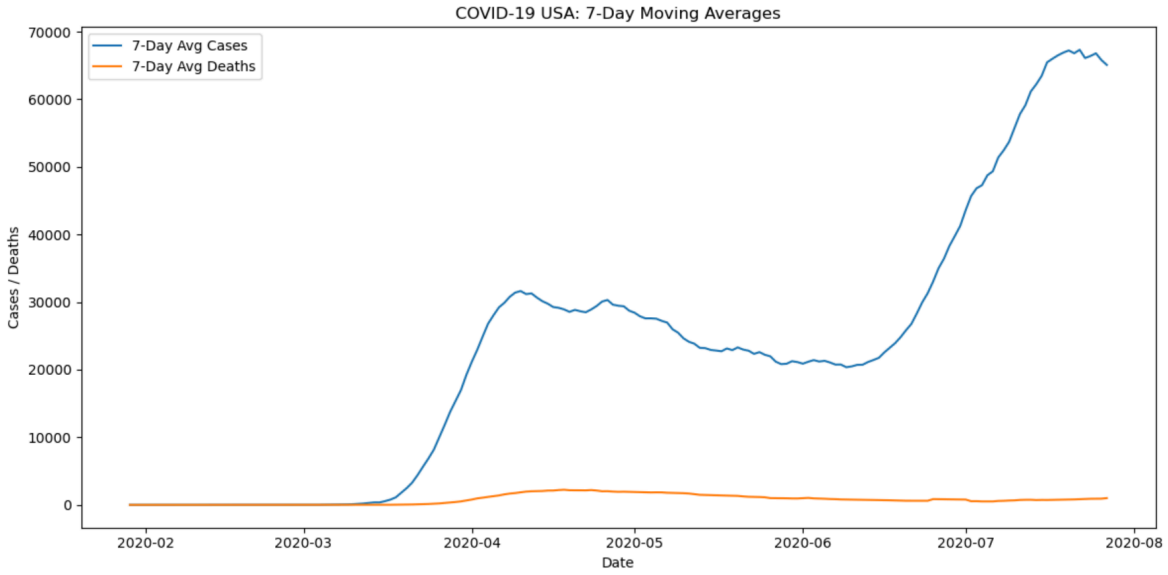

Figure 3: 7-Day Moving Averages of New COVID-19 Cases and Deaths in the US
A static line chart comparing the smoothed daily new cases and deaths using a 7-day moving average. Peaks in cases occur in early April and again in July, with deaths trailing behind. (See: moving_averages.png).

Ratio of Deaths to Cases Over Time

To understand the relationship between infections and mortality over time, the ratio of new deaths to new cases was calculated using a 7-day moving average. This ratio is particularly useful for understanding the burden on healthcare systems and the effectiveness of interventions over time. As shown in Figure 4, the death-to-case ratio begins at relatively high levels in February and March, frequently exceeding 10%, reflecting limited testing and severe case bias in early pandemic months. As testing expanded and mild/asymptomatic cases were more frequently detected, the ratio steadily declined.

By July 2020, the death-to-case ratio stabilized at under 4%, showing either a lower severity of illness among new cases or improved medical response and hospitalization protocols. The chart also shows points of unpredictability early in the pandemic when daily case and death counts were both low and highly variable.

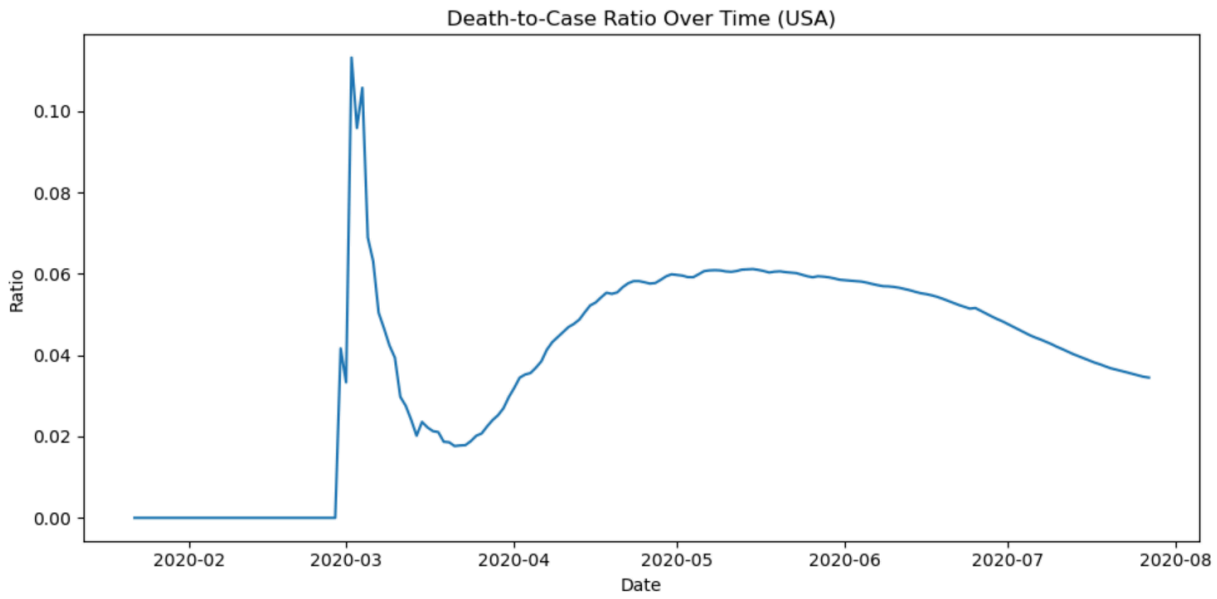

Figure 4: Ratio of New Deaths to New Cases (7-Day Average) in the US
A static line chart showing the evolving ratio of daily new deaths to new cases. Early values are high due to underreporting of cases; by July 2020, the ratio declines significantly. (See: death_to_case_ratio.png).

Comparison of Case and Death Growth

An interactive dual-line visualization (Figure 5) was used to directly compare cumulative confirmed cases and deaths over time. This chart gives temporal context to the rapid increase in both variables from mid-March onward. While cases rise steeply in both April and July, the trajectory of deaths rises sharply only in the first wave and flattens after, again showing the trend observed in later months.

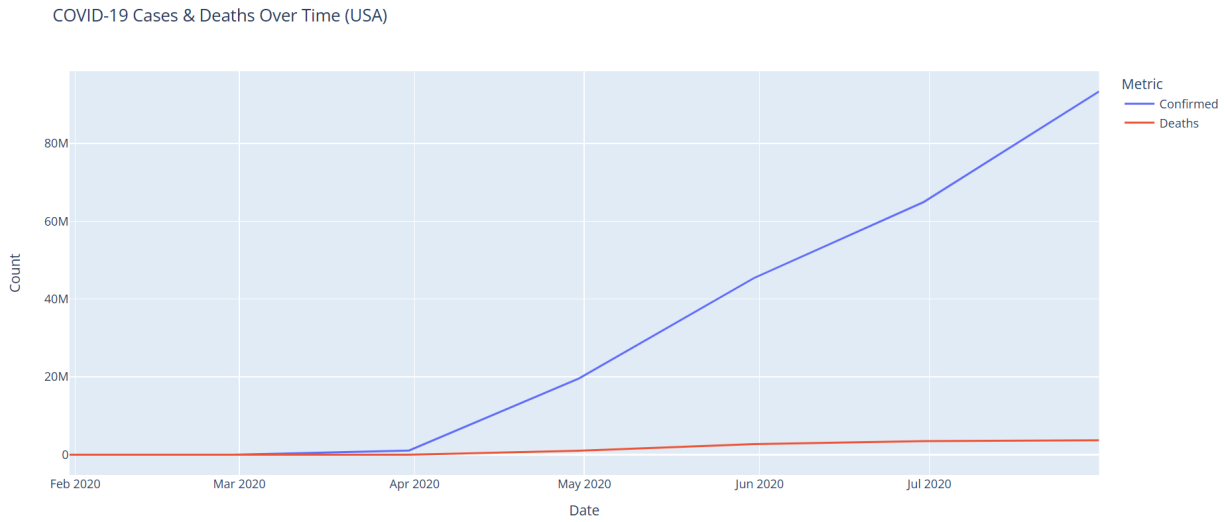

Figure 5: Cumulative Confirmed Cases and Deaths in the US
An interactive time series chart displaying the cumulative growth of cases and deaths, revealing steep early growth and a divergence pattern in later months. (See: cases_deaths_over_time.html for full interactive version).

Correlation Between Cases and Deaths

To further understand the association between the number of new cases and resulting deaths, a scatter plot with a regression line was used (Figure 6). The resulting plot shows a positive but non-linear correlation, with increasing scatter as the number of cases increases. This pattern reinforces the notion that while rising cases generally lead to more deaths, the strength of that relationship varies across time. Outliers may correspond to periods where deaths surged despite lower case counts or vice versa — likely due to testing lag, delayed mortality, or demographic differences in affected populations.

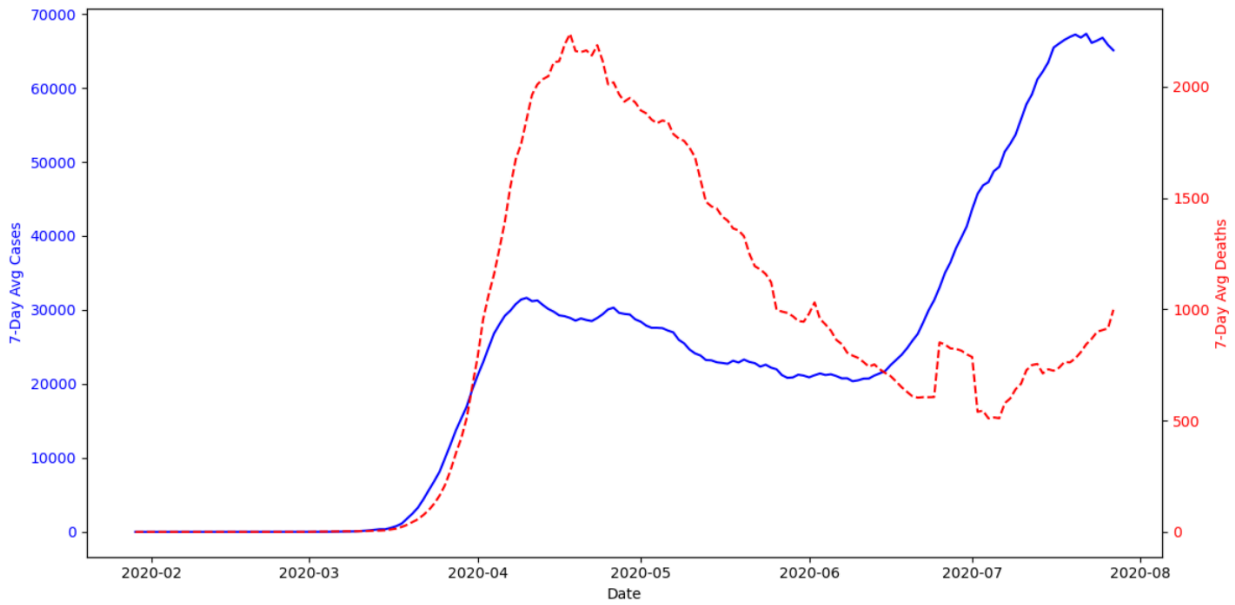

Figure 6: Daily New Deaths vs. Daily New Cases in the US

Scatter plot showing the relationship between new COVID-19 cases and deaths per day, with a regression line indicating a positive correlation (See: cases_vs_deaths.png).

Interactive Dash Visualizations

Interactive dashboards were developed to enable the manipulation of age group and condition filters dynamically, revealing complex patterns that would be difficult to distinguish from static charts alone. Each dashboard focuses on a different visual modality—choropleth mapping and bar plotting—to capture both geographic distribution and categorical comparisons.

Mortality by Comorbidity and Age Range Dashboard: Bar Plot by State

This dashboard enables the examination of COVID-19 mortality counts by underlying medical condition and age range. Two dropdown menus allow for simultaneous selection of an age range and one of 23 medical conditions, including "Diabetes", "Sepsis", "Ischemic heart disease", and "Obesity". Once selections are made, a bar chart updates to display COVID-19 death counts by state for individuals matching those criteria.

The dashboard captures the co-occurrence of COVID-19 with specific pre-existing conditions and reveals how this varies geographically. For instance, selecting "Obesity" and the "55–64" age group generates a distribution with visible spikes in Texas and Florida, all of which report relatively high death counts. When "Sepsis" is chosen for the "65–74" group, the bars are tallest in New Jersey, and California, and

Texas, reflecting areas hit hard in early 2020 when sepsis-related complications were common among hospitalized patients.

Another significant feature of this dashboard is its ability to highlight the interaction between aging and specific health conditions. Choosing "Heart failure" for the "85+" range produces an even more concentrated chart, dominated by states like New Jersey and Pennsylvania, indicating the risk among elderly populations with chronic cardiovascular issues. Conversely, choosing "Diabetes" and "35–44" yields fewer absolute deaths but gives insight into early-onset complications in relatively younger individuals.

The visual design makes comparisons across states intuitive, and the tight integration between age and condition filters improves hypotheses—for instance, identifying whether certain states systematically report higher mortality for obesity versus circulatory diseases across multiple age groups.

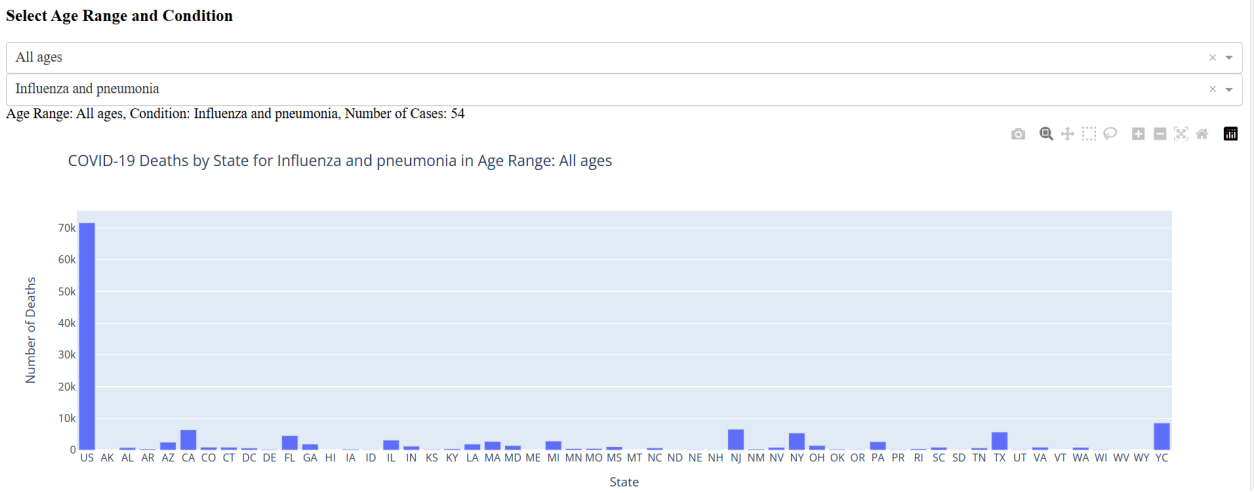

Figure 7: Bar Chart of COVID-19 Deaths by State, Filtered by Condition and Age Group
COVID-19 deaths by US state, filtered by underlying medical condition and age range (See: http://127.0.0.1:8050/ for full interactive dashboard).

Dynamic Time Series Dashboard: Tracking State-Specific Trends Over Time

This dashboard presents a US state-level choropleth map that dynamically updates based on the selected age group. The map allows users to explore the total number of COVID-19 deaths reported in each state, aggregated by age. Users can isolate specific age ranges such as "0–24", "55–64", or "85+". This interactive functionality makes it possible to identify states where mortality was particularly concentrated within certain demographic brackets.

One critical observation enabled by this dashboard is the relatively low death count among individuals aged 0–24, regardless of state. In contrast, the "85+" group reveals stark geographic contrasts, with New York, New Jersey, and Massachusetts showing disproportionately high mortality totals. For this group, states in the Northeast show deep saturation in the choropleth, while large portions of the Midwest and South show lighter shading, indicating lower totals. Notably, filtering to the "55–64" or "65–74" ranges

shows a broader geographic spread, including surges in Texas, Florida, and California, likely reflecting the wave of infections during June and July 2020.

The choropleth visualization also supports comparative inference across regions. For example, even after accounting for population size, states such as Louisiana and Michigan show elevated death counts in multiple age brackets during the first half of 2020. This suggests an intersection of underlying vulnerability and early outbreak intensity. The interactivity helps draw attention to such patterns.

Figure 8: Choropleth Map of COVID-19 Deaths by Age Group
Display of total COVID-19 deaths by U.S. state, with dynamic filtering based on age group (See: http://127.0.0.1:8051/ for full interactive dashboard).

Contribution and Findings

This project addresses a critical gap in the public understanding and exploration of COVID-19 mortality data by developing an interactive visualization platform that reveals how underlying medical conditions and age groups intersect with state-level death counts. While national statistics provide broad overviews, they often obscure the subtle variations in mortality patterns across demographic and geographic dimensions. This work enables more precise insights by allowing users to isolate COVID-19 deaths based on specific comorbidities.

The primary question guiding this study was how different underlying conditions contributed to COVID-19 mortality across various segments of the population and geographic regions. The findings indicate that some conditions—particularly circulatory diseases, respiratory illnesses, and dementia—had a disproportionate presence in older populations and in certain states, suggesting regional health disparities and differing population vulnerabilities. Additionally, the inclusion of interactive dashboards allows users to observe these patterns dynamically rather than relying on static, aggregate figures. This functionality is especially important for public health officials, policymakers, and researchers who need granular data to inform interventions or resource planning. Ultimately, this project contributes a usable, data-driven tool that improves understanding of the heterogeneity in COVID-19 mortality. It gives evidence for targeted health communication and policy strategies by highlighting where and among whom mortality was most concentrated, based on both medical vulnerability and demographic risk. By focusing on the interaction between age, condition, and geography, this study elevates conversations beyond national averages and allows for a more informed response to pandemic health challenges.

Future Work

Several steps can build on the work completed in this project. First, a more nuanced temporal component can be added to the dashboards to allow users to explore how deaths associated with certain conditions changed over time—particularly across different COVID-19 waves or after the introduction of vaccines. This would transform the current cross-sectional analysis into a longitudinal one, adding value for understanding the evolution of the pandemic's impact on vulnerable populations.

Second, more robust statistical models (e.g., multivariate regression or geospatial clustering algorithms) could be introduced to test for significant associations between underlying conditions, state-level policies, demographic factors, and COVID-19 mortality. This would allow the dashboards to serve not only as descriptive tools but also as exploratory interfaces.

In a time when data overload and misinformation are common, this project contributes an interpretable, and evidence-based tool that clarifies complex mortality data into actionable insights. It shows the importance of multidimensional analysis—where age, condition, and geography intersect—to understand the true burden of a public health crisis. The tools and findings from this project serve as a basis for future work in pandemic preparedness, real-time health surveillance, and equitable health response planning.