COVID Data Unveiled: An Exploratory Dive

This report presents an analysis of COVID-19 data for the United States, compiled from the states_daily.csv dataset provided by the COVID Tracking Project API (https://covidtracking.com/data/api). The dataset covers the period from January 13, 2020 to March 7, 2021. To enhance the analysis, population data from the year 2020 was obtained from https://state.1keydata.com/state-population-density.php and merged with the dataset.

Data Preparation and Challenges:

A significant challenge was the high number of missing values across several key columns. Due to the delicate nature of health data and the potential impact of imputation on the analysis, an exhaustive inspection was conducted.

This inspection involved:

- **Null** Analysis: Thorough examination of null patterns across states and time.
- **Outlier** Detection: Identification and analysis of outliers using statistical methods and visualizations.
- Data **Distribution** Analysis: Examination of data distributions to understand the spread and skewness of variables.
- **Imputation** Impact Assessment: Evaluation of the impact of various imputation techniques, including mean, median, hot-deck, KNN, regression, and interpolation, on the overall analysis and per-state results.
- **Scatterplot Analysis**: Visual examination of relationships between variables, per state, to inform imputation decisions.
- **State-Specific Variations**: Acknowledging the significant differences in data reporting and availability between states.

The approach taken was to carefully consider the implications of each imputation method and to prioritize accuracy and transparency. In many cases, imputation was avoided due to the potential for introducing bias, especially given the variability in reporting practices across states.

Key Findings:

National Impact:

- Approximately 28.7 million COVID-19 cases and 515,000 deaths.
- Over 363 million tests conducted.
- Significant daily fluctuations in cases, deaths, and tests.

State-Level Disparities:

- California, Texas, and New York had the highest death tolls.
- Northern Mariana Islands, Virgin Islands, and Vermont had the lowest.
- New Jersey and northeastern states had the highest death rates per population.
- Hawaii and western states had the lowest.

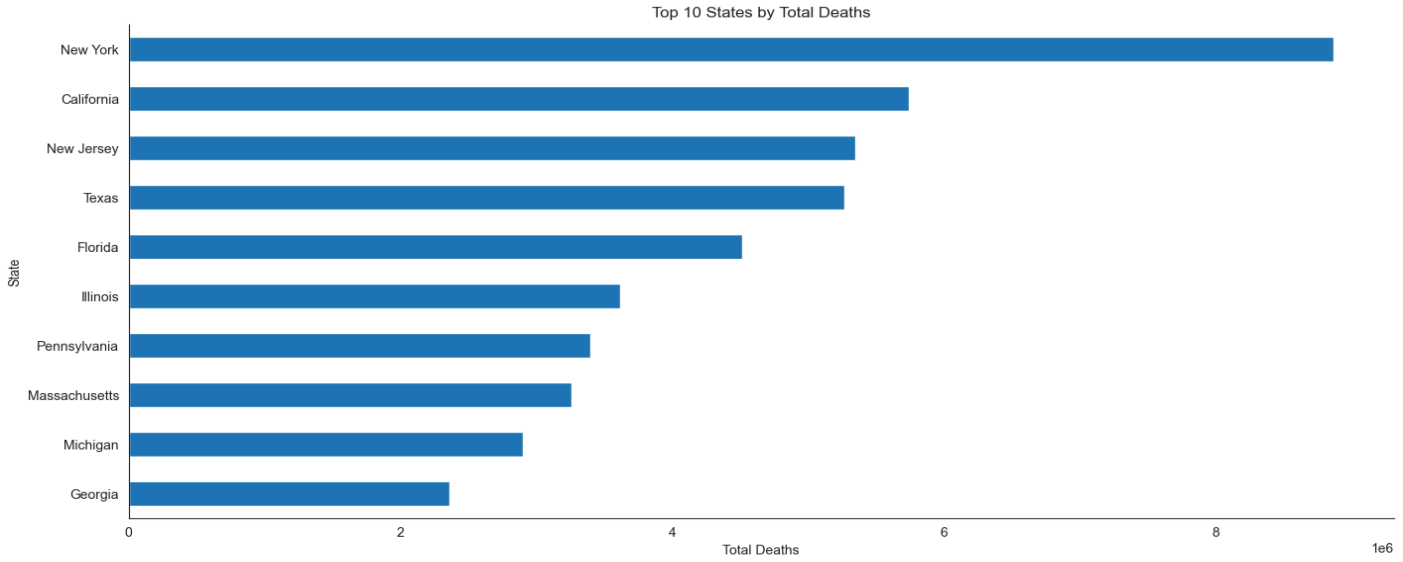
Data Characteristics:

- Death rate relative to cases: 1.79%.
- Case rate relative to tests: 7.90%.

Data Quality:

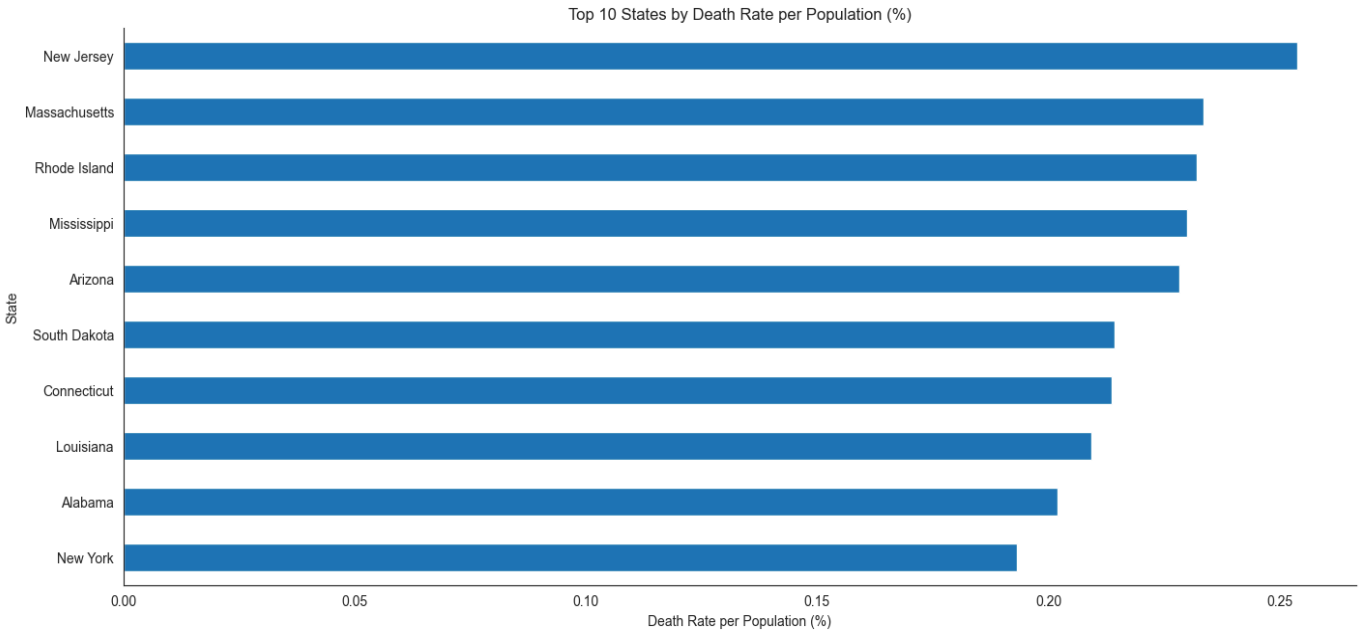
- Significant missing data points across multiple metrics.

Top 10 U.S. States with the Highest Total Number of Related Deaths



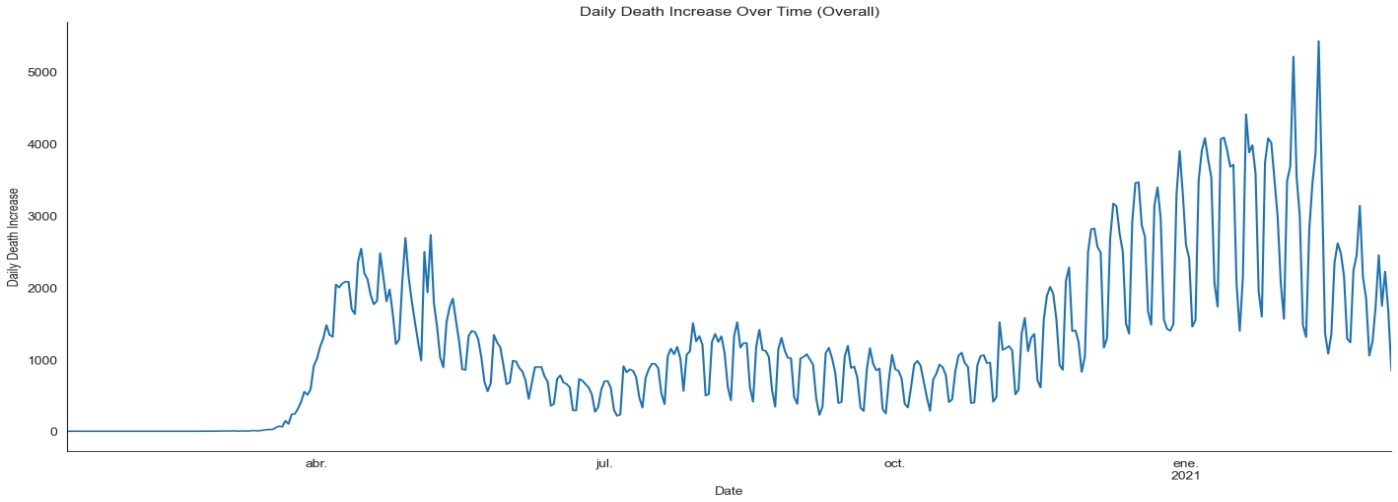
This chart shows the top 10 U.S. states by COVID-19 deaths. Notably, New York's death count is exceptionally high, significantly exceeding all other states. While California and New Jersey also show substantial deaths, the wide range across states highlights the varying impact of the pandemic. This prompts us to consider factors like population density and healthcare access, and to further investigate the public health measures that may have influenced these outcomes.

Top 10 States by Death Rate per Population (%)



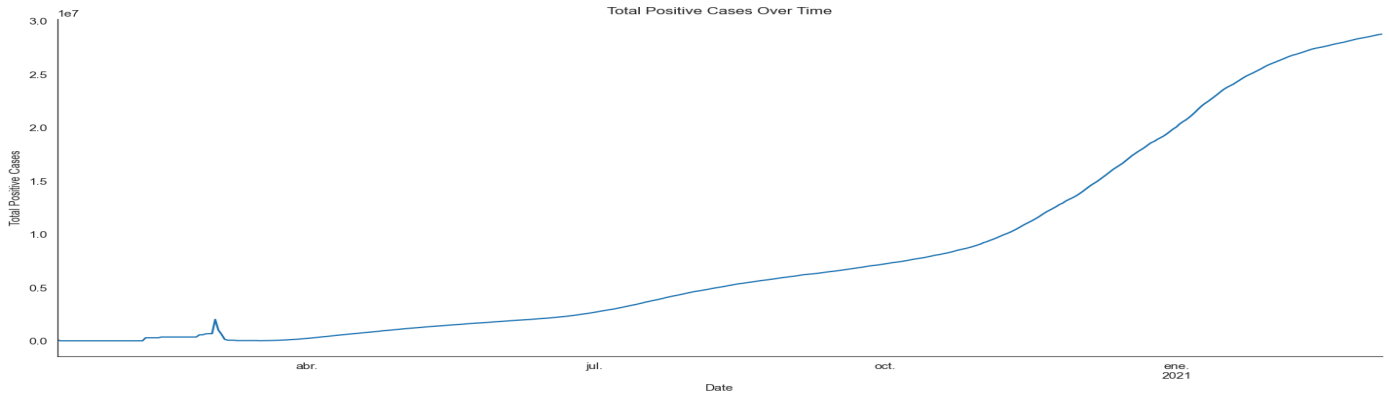
This chart ranks the top 10 U.S. states by COVID-19 death rate per population, showing the percentage of deaths relative to each state's population. New Jersey has the highest rate, followed closely by Massachusetts and Rhode Island. This highlights that population size doesn't directly correlate with total deaths; smaller states can have a proportionally higher impact. It prompts us to consider factors beyond sheer numbers, like healthcare capacity and population density, in understanding the pandemic's severity.

Daily Death Increase Over Time (Overall)



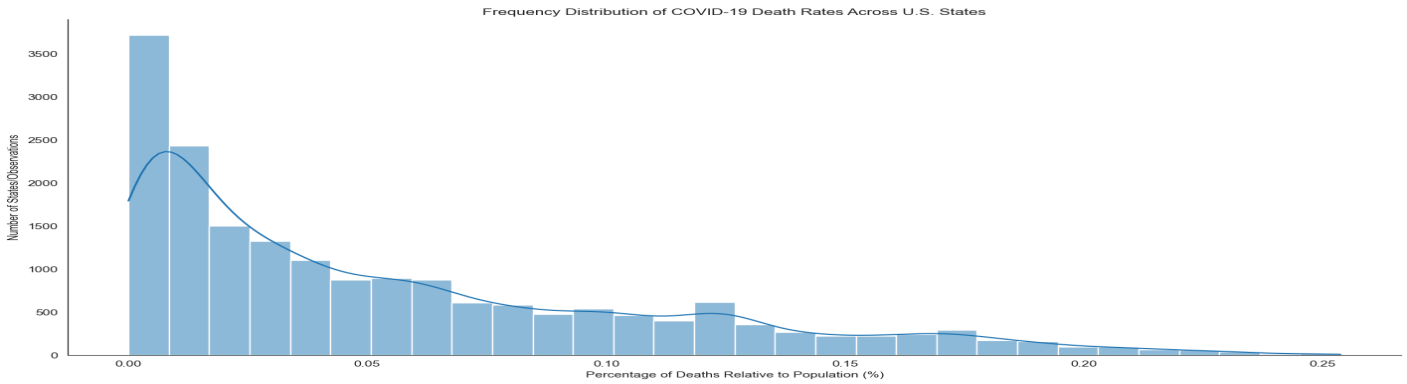
This line graph displays the daily increase in COVID-19 deaths across the U.S. over time. We observe two significant peaks: one in the spring of 2020 and a much larger surge in late 2020 and early 2021. The late 2020/early 2021 surge may reflect increased travel and gatherings during the Thanksgiving and Christmas holiday season, potentially leading to increased transmission. The fluctuations indicate the dynamic nature of the pandemic, with periods of rapid increase followed by relative plateaus. This visual highlights the importance of tracking daily data to understand the evolving severity of the pandemic.

Total Positive Cases Over Time



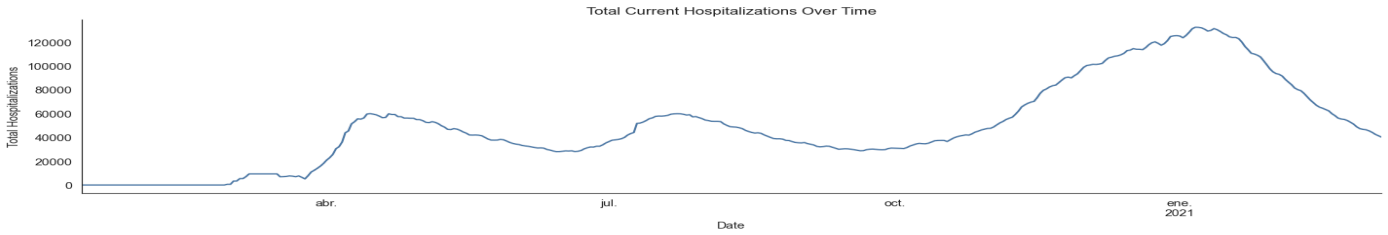
This line graph tracks the cumulative total of positive COVID-19 cases in the U.S. over the analysis period. We see a relatively slow initial increase, followed by a significant surge starting in late 2020 and continuing into early 2021. This rapid escalation indicates the accelerated spread of the virus during this time, highlighting the critical need for effective public health measures to control transmission.

Frequency Distribution of Death Rates Across U.S. States



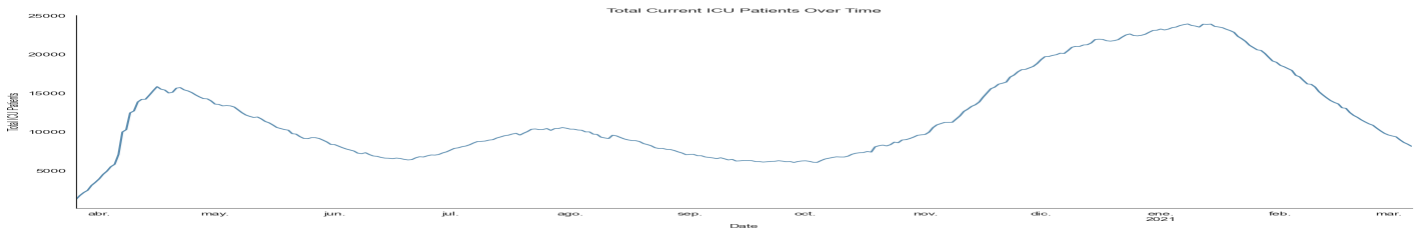
This histogram displays the distribution of COVID-19 death rates per population across all U.S. states. We observe a strong right skew, with a high concentration of states having very low death rates, and a long tail indicating a few states with significantly higher rates. This distribution highlights the variability in pandemic impact across the country, suggesting that factors like population density, healthcare access, and regional public health policies likely play a role in the observed differences.

Total Current Hospitalizations Over Time



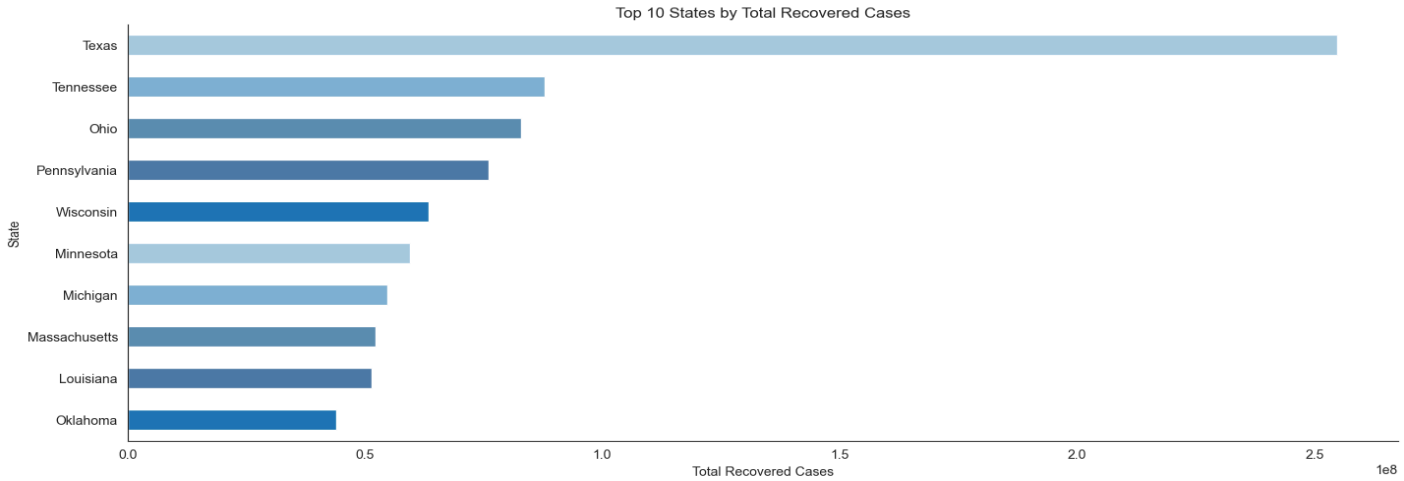
This line graph tracks the total number of patients currently hospitalized with COVID-19 in the U.S. over time. We observe distinct peaks, particularly in late 2020 and early 2021, mirroring the surge in cases seen in other charts. These peaks indicate periods of intense strain on the healthcare system. The fluctuations highlight the dynamic nature of hospitalizations, which are a critical metric for assessing the immediate impact of the pandemic on healthcare resources.

Total Current ICU Patients Over Time



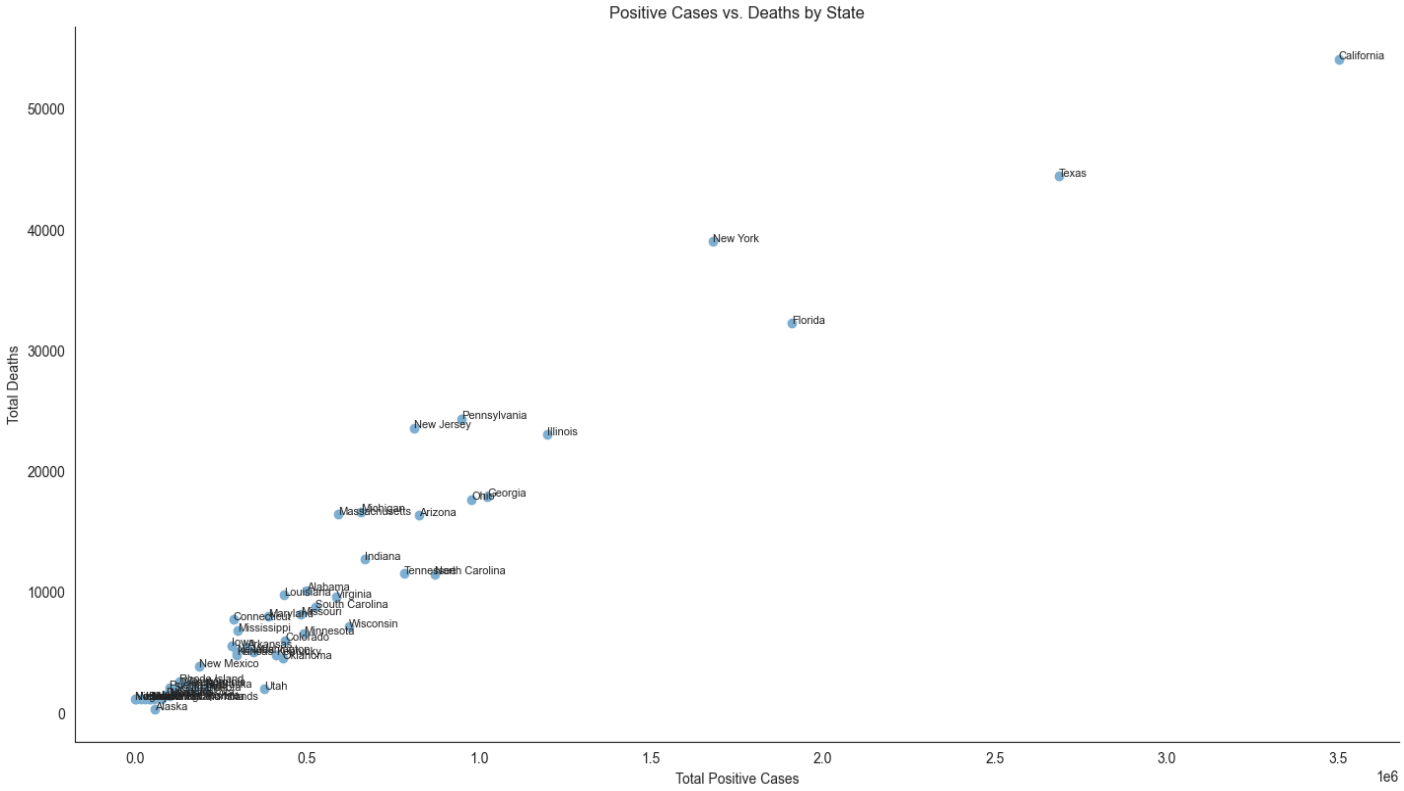
This line graph tracks the total number of patients currently in intensive care units (ICUs) due to COVID-19 in the U.S. over time. Similar to the hospitalization chart, we see significant peaks, especially in late 2020 and early 2021. The ICU data reflects the most severe cases of COVID-19, and these peaks indicate periods when healthcare systems faced extreme pressure in terms of critical care capacity. This visual underscores the strain on ICUs during peak infection periods.
Note: percentage of null in recovery is 40%. This is not taking into account the states with no information.

Top 10 States by Total Recovered Cases



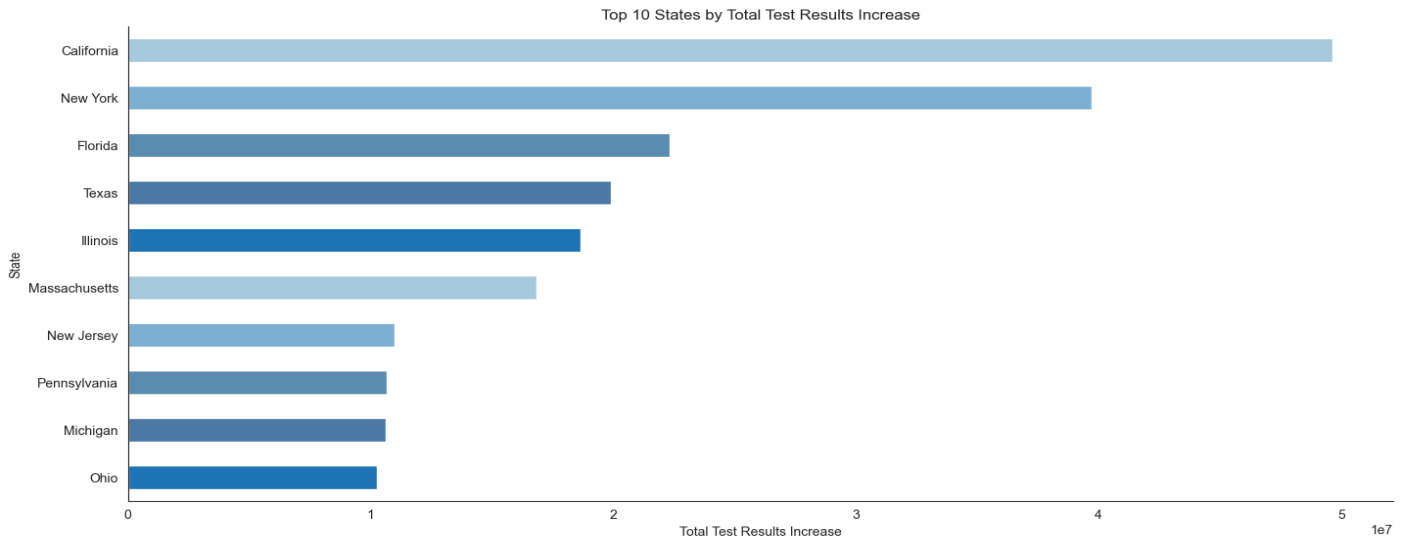
This horizontal bar chart displays the top 10 U.S. states with the highest total number of recovered COVID-19 cases. Texas shows a significantly higher number of recoveries compared to other states. This chart highlights the recovery aspect of the pandemic and the varying degrees of recovery across different states. It's important to note that recovery data can be complex and influenced by reporting differences, which may impact the direct comparison between states.
Note: percentage of null in recovery is 40%. This is not taking into account the states with no information.

Positive Cases vs. Deaths by State



This scatter plot illustrates the relationship between total COVID-19 positive cases and deaths across U.S. states, revealing a clear positive correlation where higher case counts generally correspond to higher death tolls. States like California, in the upper right quadrant, demonstrate the most severe impact with high values in both metrics, while others deviate, suggesting factors beyond case numbers, such as healthcare capacity or demographics, influence mortality. This visualization provides a broad overview, highlighting high-impact states and prompting further analysis into the varying pandemic experiences across the country.

States with the Highest Increase in COVID-19 Tests



This horizontal bar chart displays the top 10 U.S. states with the highest total increase in COVID-19 test results. California leads significantly, followed by New York and Florida. This visualization highlights the states with the most extensive *expansion* in testing efforts, which is crucial for tracking and controlling the spread of the virus. The variations in test increases across states likely reflect differences in population, available resources, and public health strategies.

C o n c l u s i o n

This analysis highlights the uneven and complex impact of COVID-19 across the U.S., with significant state-level disparities in cases, deaths, and testing. While large states like California and Texas recorded the highest absolute numbers, northeastern states faced the highest death rates per population, reflecting differences in healthcare capacity, population density, and response strategies.

A major challenge in this study was data quality, particularly the high proportion of missing values in key metrics like ICU admissions and recoveries. These gaps affected visualizations and emphasize the need for transparent and standardized data collection in future health crises.

Beyond case numbers, the pandemic's impact was shaped by socioeconomic factors, public health policies, and healthcare access. To fully understand these effects, further research should examine long-term consequences, vaccination impact, and economic recovery trends. Expanding data sources to include mobility patterns and healthcare infrastructure could offer a more comprehensive perspective on the crisis.

Ultimately, this study underscores the need for continuous analysis and improved data practices to draw meaningful lessons from the pandemic. These insights will be crucial for shaping future public health responses and enhancing preparedness for global health emergencies.