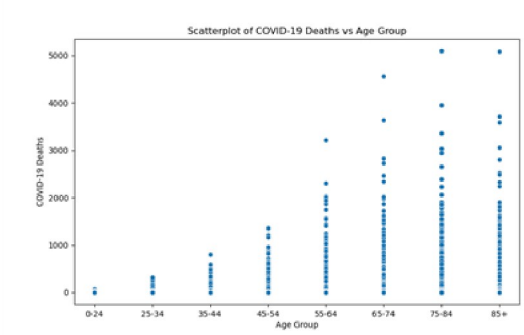
Analysis of COVID-19 Data

This dashboard will explore aspects of COVID-19, focusing on conditions contributing to COVID-19 deaths, any correlation between conditions, geographical analysis of death rates, and statistical analyses. This data covers the timeframe from 2020 through 2023 across the USA.

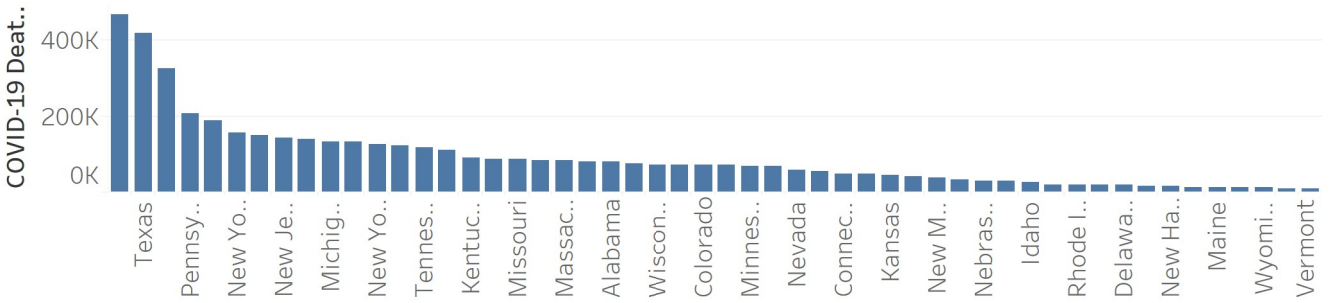


Introduction

Research Focus: Analyzing conditions contributing to COVID-19 deaths across different demographics and geographic regions in the United States. Developing an understanding of these factors could help physicians, public health officials, and lawmakers create effective interventions.

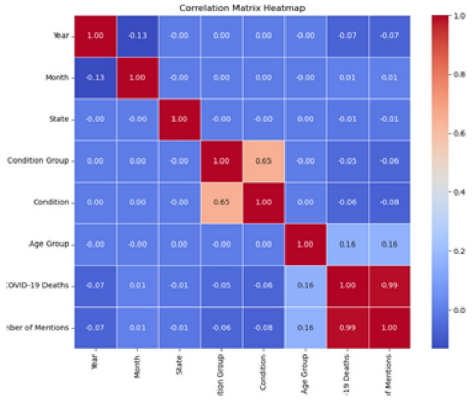
Importance: The COVID-19 pandemic significantly impacted global health. Insights into contributing conditions will guide interventions and policy decisions designed to reduce mortality and improve public health.

Variation in Total COVID-19 Deaths across the United States

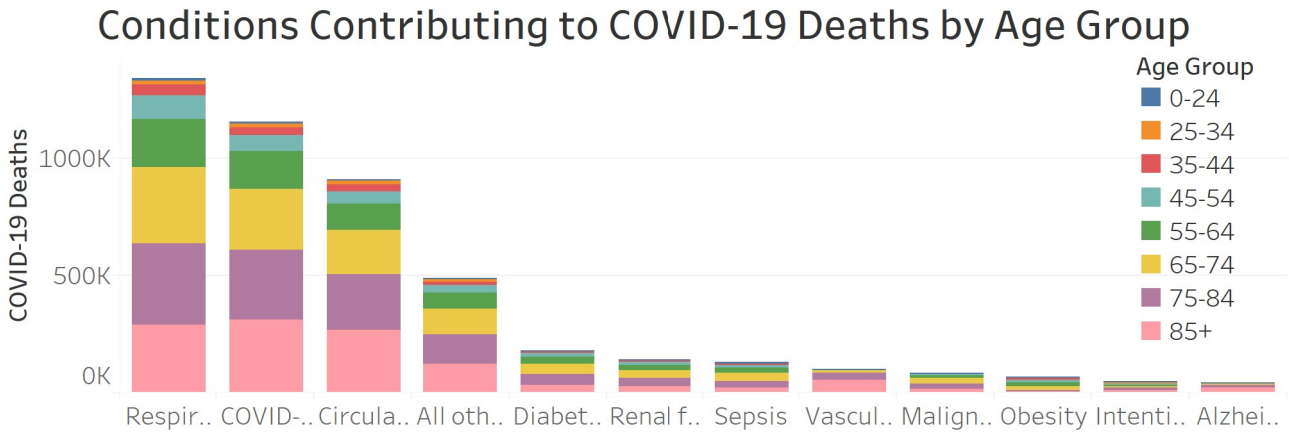


Initial Exploratory Analyses

The initial data exploration provides strong evidence of the relationship between COVID-19 deaths and specific conditions, with older age groups with certain high-risk conditions being particularly impacted. The findings highlight the importance of considering demographic and condition-specific factors in understanding and managing the impact of COVID-19.



Demographic Analysis:
Common medical conditions contributing to COVID-19 deaths include respiratory and cardiovascular diseases, especially in older age groups.
Older age groups show higher death counts for almost all conditions.

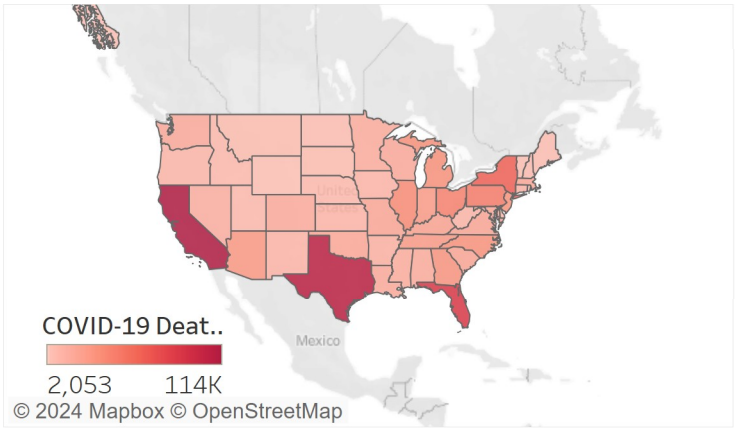


Geographical Analysis

Total COVID-19 Deaths by contributing condition map

The states with the highest total COVID-19 deaths are known for their large populations, which naturally results in higher absolute numbers of deaths.

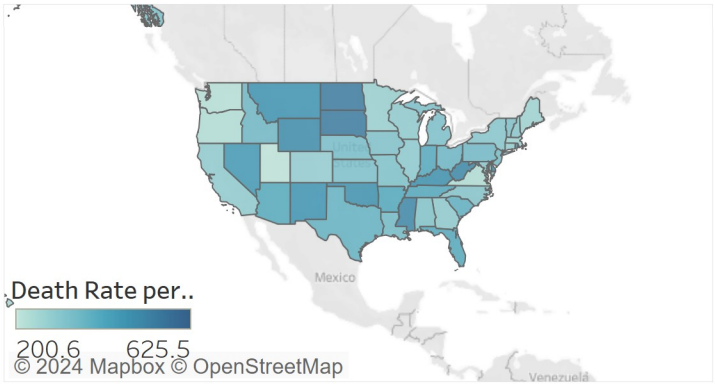
Total COVID-19 Deaths by Top Contributing Condition



COVID-19 Mortality by Contributing Condition per 100,000 Map

The highest death rates per 100,000 population are observed in states like North Dakota, South Dakota, and Mississippi. These high rates indicate a severe impact relative to the population size, suggesting significant outbreaks and possibly less effective containment measures in these areas.

COVID-19 Mortality Rates by Contributing Condition per 100,000

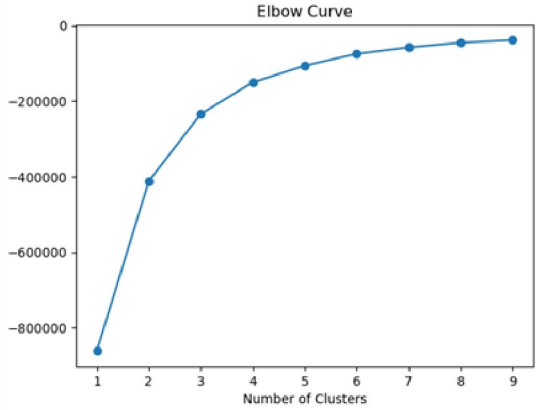


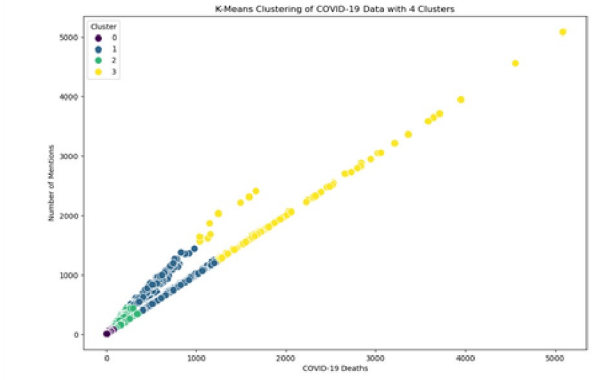
Regional Variations:

Both maps show that populous states like California, Texas, and Florida have high total deaths, the death rates per 100,000 reveal more about the outbreak relative to the population. This suggests that public health responses and healthcare capacity may have varied significantly across states. These two maps provides a comprehensive view of the COVID-19 impact across the United States. While total deaths highlight the absolute scale of the pandemic, death rates per 100,000 population offer critical insights into the relative severity and effectiveness of public health measures in different regions.

Cluster Analysis

This analysis applied K-means clustering to our data that contained COVID-19 deaths and contributing medical conditions to identify any meaningful groups within our data.





Analysis Interpretation

The Elbow technique was applied to determine the optimal number of clusters. After plotting the elbow curve, 4 clusters are optimal as the curve begins to flatten out after this point.

Cluster 0: Represents minimal impact, with low numbers of deaths and mentions.
Cluster 2: Represents moderate impact.
Cluster 1: Represents high impact.
Cluster 3: Represents very high impact.

There is considerable variability observed within the clusters, especially in Clusters 1 and 3.
The mean values of COVID-19 deaths and mentions increased progressively from Cluster 0 to Cluster 3, highlighting a gradient of COVID-19 impact.

We can use the K-means clustering results in future steps of an analytics pipeline. Cluster labels can be incorporated into future predictive models to improve accuracy and context-awareness.
Visualizations and reports based on cluster differences can be used to enhance communication and support more informed decisions.

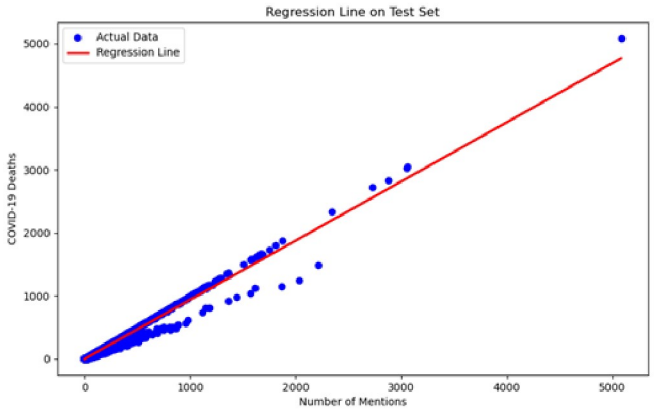
Linear Regression Analysis

This analysis used supervised machine learning, specifically linear regression, to explore the relationship between the Number of mentions of a condition on death certificates and COVID-19 deaths.

Hypothesis: If the number of mentions of a condition on death certificates is higher, then the COVID-19 death count will be significantly higher.

Building the Model: Defined the independent variable (X = Number of Mentions) and the dependent variable (y = COVID-19 Deaths).

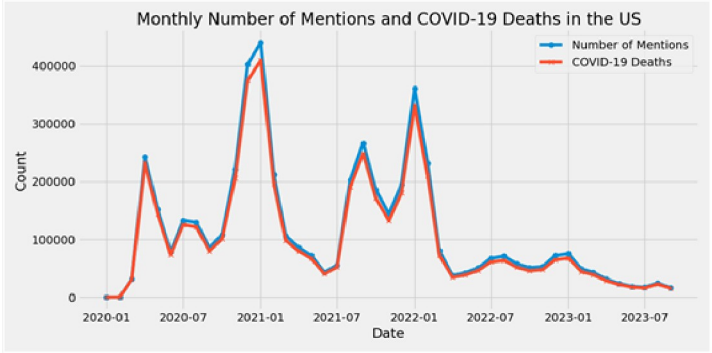
Model Performance Statistics: Calculated the Mean Squared Error (MSE) and R-squared score:
MSE: 66.75
R-squared Score: 0.975
A low MSE and a high R-squared score indicate that the model's predictions are close to the actual values and that a significant proportion of the variance in COVID-19 deaths can be explained by the number of mentions.

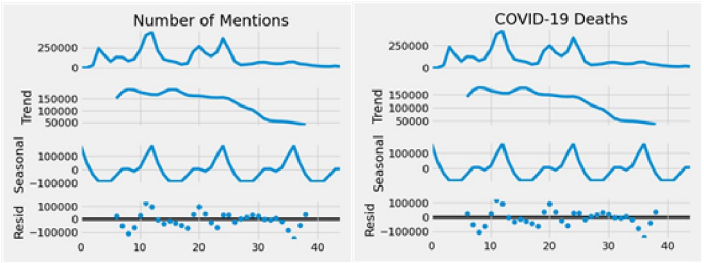


Analyzing Time-Series Data

This analysis explores the time series data of COVID-19 mortality and the number of mentions of medical conditions on death certificates to understand trends, seasonality, and attempt to make future forecasts.

The line chart displays monthly data for Number of mentions and COVID-19 deaths. Both exhibit significant fluctuations over time, with peaks corresponding to the waves of the pandemic.
Notable peaks occur during the winter months, indicating seasonal surges in COVID-19 diagnoses and mortality.





Decomposition Visualizations

The top panel shows the variabled observed values, mirroring the line chart above.
The second panel highlights a downward trend over time, indicating a decrease in the number of mentions and deaths as the pandemic progressed.
The third panel highlights periodic fluctuations, confirming a seasonal pattern in the data with regular increases during certain months.
The bottom panel shows residuals (random noise), which appear to be relatively random, indicating the trend and seasonal components capture most of the variability in the data.

Summary of Findings

Key Insights

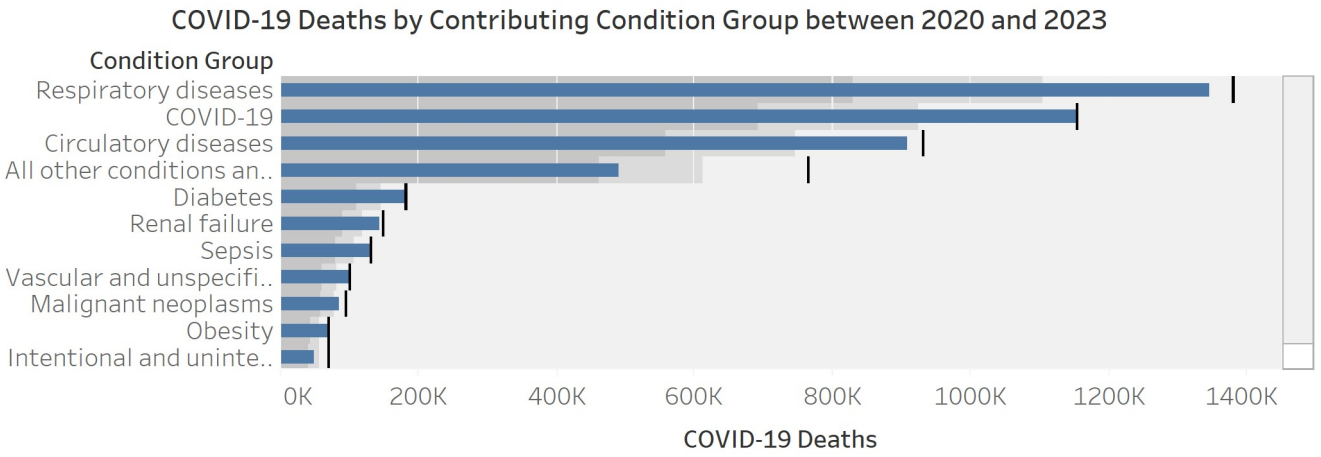
There is a strong positive correlation between COVID-19 deaths and the number of mentions of specific medical conditions on death certificates. Conditions like Influenza and pneumonia, vascular and unspecified dementia, diabetes, and ischemic heart disease show significant correlations with COVID-19 deaths.

There are a high number of total COVID-19 deaths in populous states like California, Texas, and Florida. The highest death rates per 100,000 population in states like North Dakota, South Dakota, and Mississippi, indicating a severe impact relative to their population size.

Linear regression model shows a strong fit with an R-squared value of 0.975, indicating that 97.5% of the variance in COVID-19 deaths can be explained by the number of mentions of conditions.

K-means clustering identified four clusters representing varying levels of COVID-19 impact. The clusters ranged from minimal impact to very high impact.

Clear seasonal trends with peaks around winter months and a downward trend over time in both mentions and deaths.



Limitations and Potential Bias

Data limitations

Data is provisional and conclusions based on this data may need revision as finalized data becomes available.

Reporting delays can range from 1 to 8 weeks or more, which means the data for recent timeframes may be incomplete.Data for 2020 and 2021 are based on final data.

Different states may have differing standards for reporting COVID-19 deaths and contributing medical conditions, which can make comparisons across states less accurate and reliable.

On average according to the official death certificates, there are an additional four medical conditions per death, which could complicate the analysis.

Deaths involving multiple conditions are counted in each relevant category, so numbers for different conditions should not be summed to avoid counting the same death multiple times.

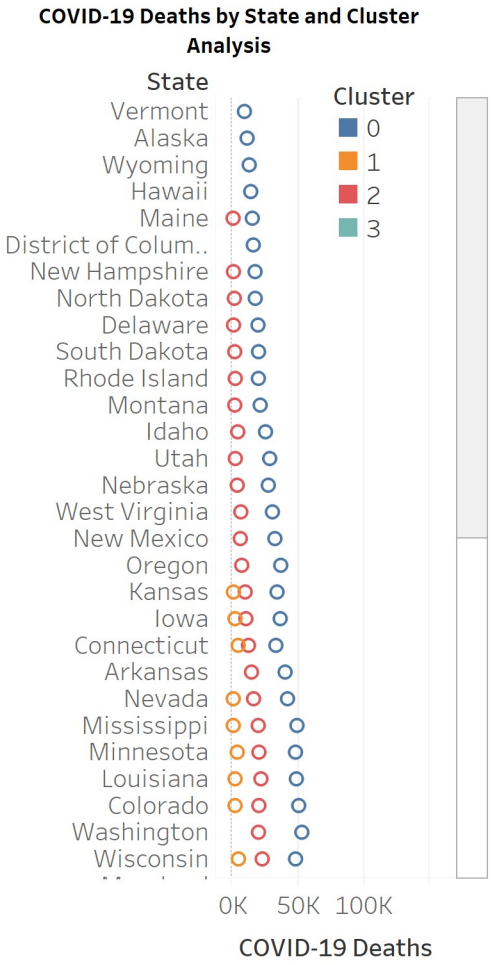
The population data used for analysis is from 2020, which may not accurately reflect changes in population over the period of the study.

Potential Sources of Bias

Inconsistent standards of reporting vary from state to state standards for reporting COVID-19 deaths and conditions can introduce biases.
Some of the data has been suppressed confidentiality which can affect the completeness of the data.

Certain demographic groups may be underrepresented, impacting the accuracy of the analysis along with differences in urban vs. rural reporting can lead to geographic biases.

Deaths with multiple comorbid conditions are counted in each category, which can lead to overestimating the prevelence of certain conditions.
Summing of conditions across categories must be avoided to prevent overestimation or results.



Recommendations and Ne..

Recommendations

Use insights gained from this analysis to improve readiness for future pandemics, with a focus on those states with high death rates.

Prioritize facilities in regions with high mortality rates to ensure better preparedness for future health crises that may develop. Also, ensure proper infrastructure and healthcare resources are allocated based on the insights obtained those specific medical conditions th..

Next Steps:

Implement the ARIMA model with identified parameters for both number of mentions and Covid-19 deaths. Continue to evaluate and refine our model using Mean Squared Error (MSE) and other metrics.

Perform clustering analyses on additional medical conditions and demographics to uncover more deeper trends and patterns.

..

Bubble Chart of Deaths versus Number of Mentions