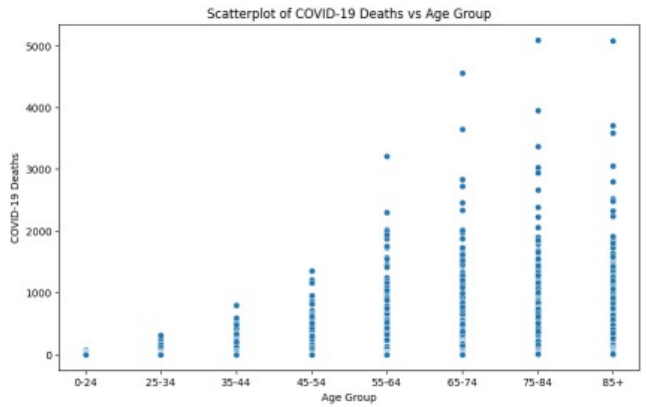
Analysis of COVID-19 Data

This presentation explores various aspects of COVID-19 data, focusing on conditions contributing to COVID-19 deaths, correlations between conditions, geographical analysis of death rates, and statistical and machine learning analyses. The data covers the period from 2020 to 2023 across different states in the USA.

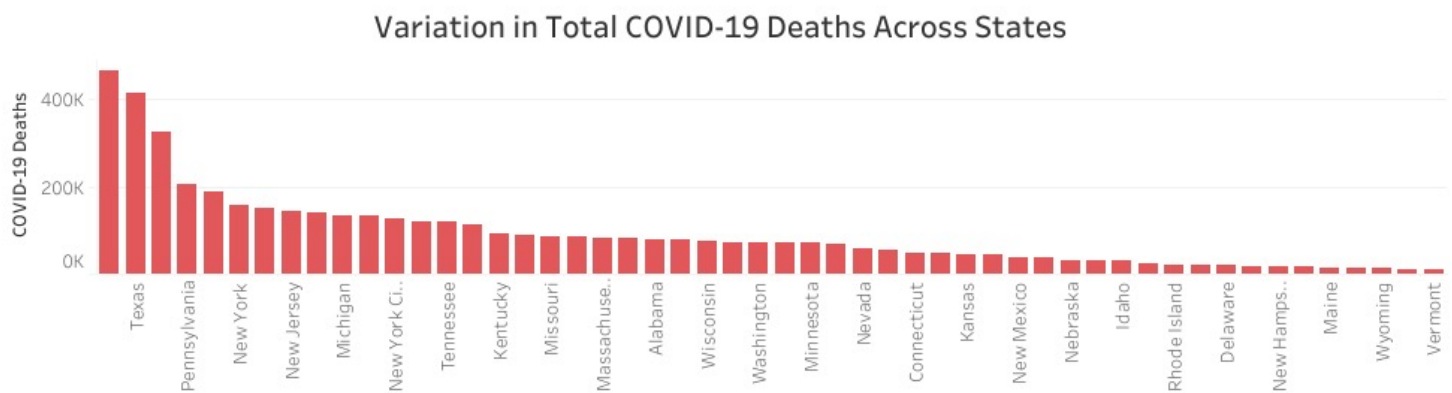


Introduction to the Problem

Research Focus: Analyzing conditions contributing to COVID-19 deaths across different demographics and locations in the United States. Understanding these factors helps public health officials and policymakers target interventions effectively.

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

Dataset Description:
Title: Conditions Contributing to COVID-19 Deaths, by State and Age, Provisional 2020-2023
Source: Centers for Disease Control and Prevention (CDC)
Variables: Date of death, state, condition group, specific condition, age group, number of COVID-19 deaths, number of mentions on death certificates.

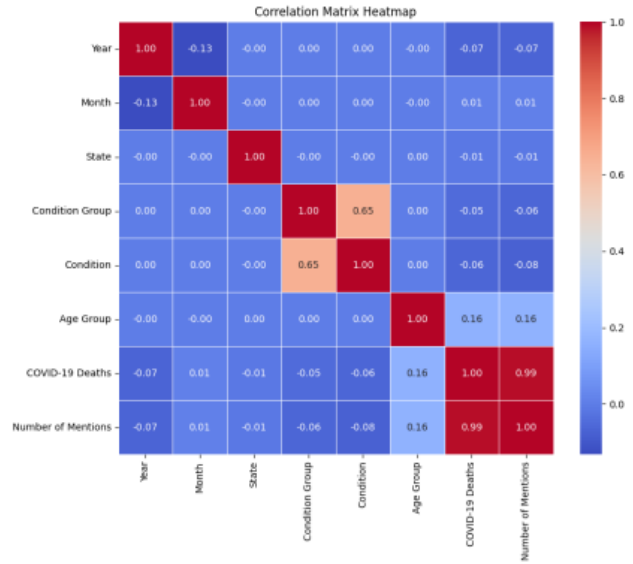


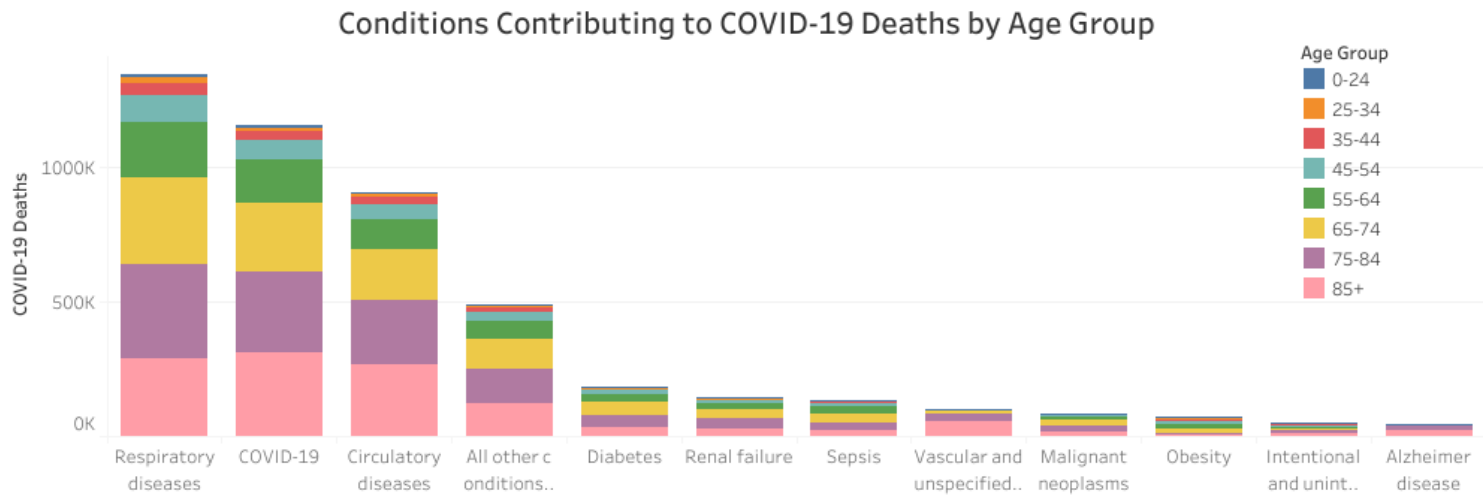
Initial Data Exploration

The initial data exploration provides strong evidence of the relationships between COVID-19 deaths and specific conditions, with older age groups and 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 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.

Predictive Modeling:
Strong correlations suggest potential for developing predictive models based on mentions of conditions, age group, and specific high-risk conditions.

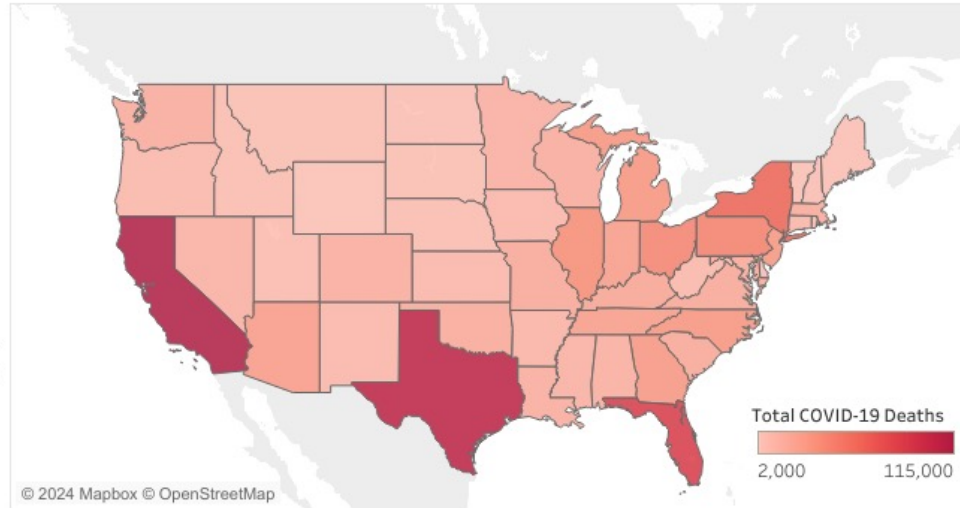




Geographical Analysis

Total COVID-19 Deaths

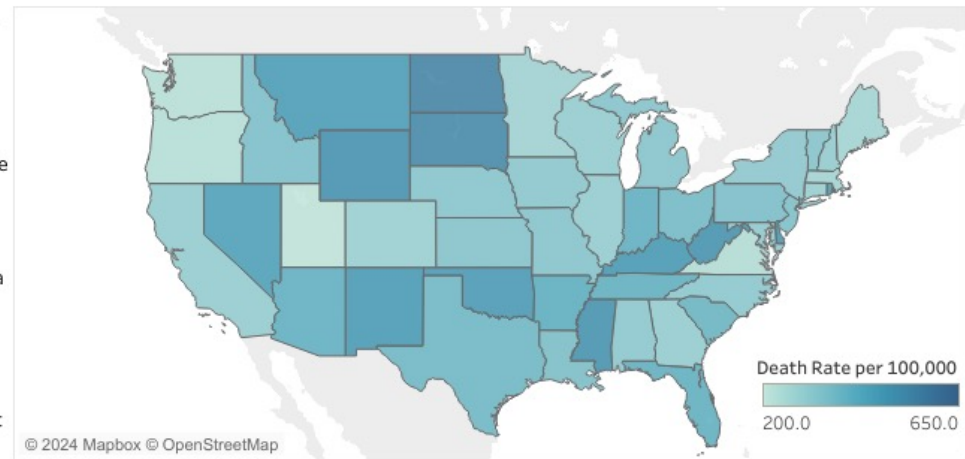
Total COVID-19 Deaths Map
High Total Deaths in Populous States:
The states with the highest total COVID-19 deaths are known for their large populations, which naturally results in higher absolute numbers of deaths.



Regional Variations:
Both maps together show that while populous states like California, Texas, and Florida have high total deaths, the death rates per 100,000 reveal more about the intensity of the outbreak relative to the population. This suggests that public health responses and healthcare capacity may have varied significantly across states.

COVID-19 Death Rates per 100,000

COVID-19 Death Rates per 100,000 Map
High Death Rates in the Midwest and South:
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.



The combination of 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.

Linear Regression Analysis

The objective of this analysis was to use 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 Introduction

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

Model Building

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

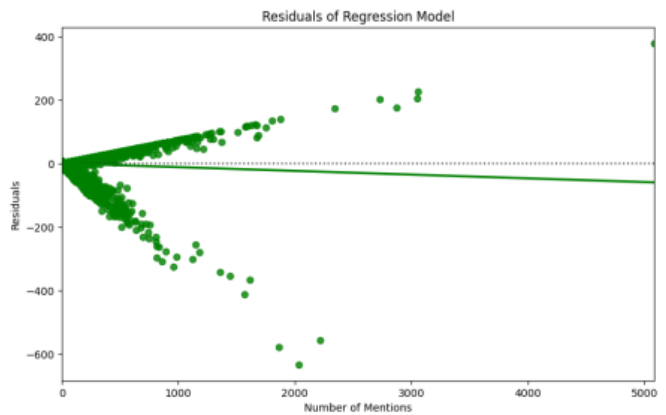
Split Data:
Split the data into training (70%) and test (30%) sets.

Run Linear Regression:
Initialized and fitted a linear regression model on the training data. Predicted on the test data.

Regression Line:
Created a plot showing the regression line on the test set, indicating a str..

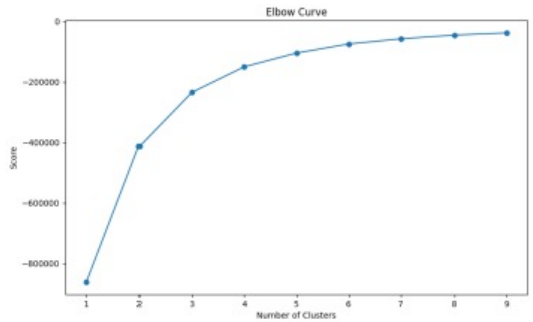
Model Performance

Performance Statistics:
Calculated the Mean Squared Error (MSE) and R-squared score:
MSE: 66.75
R2 Score: 0.975
A relatively 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.

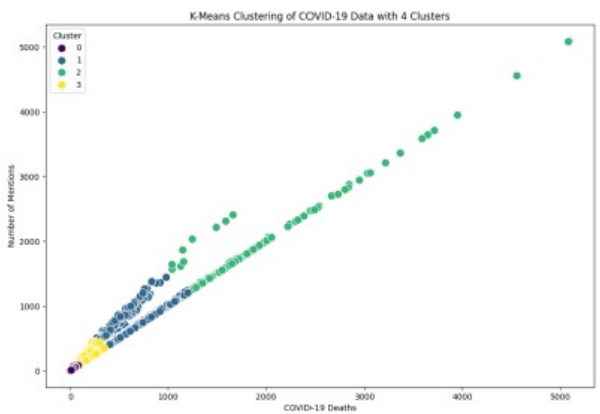


Cluster Analysis

The objective of this analysis was to apply K-means clustering to the dataset containing COVID-19 deaths and contributing conditions to identify meaningful groups within the data.



The Elbow Technique:
Applied the elbow technique to determine the optimal number of clusters.
Plotted the elbow curve, which suggested that the optimal number of clusters is 4 as the curve starts ..



Cluster Visualization:
Created scatterplots to visualize the clusters based on "COVID-19 Deaths" and "Number of Mentions."

Cluster Analysis

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

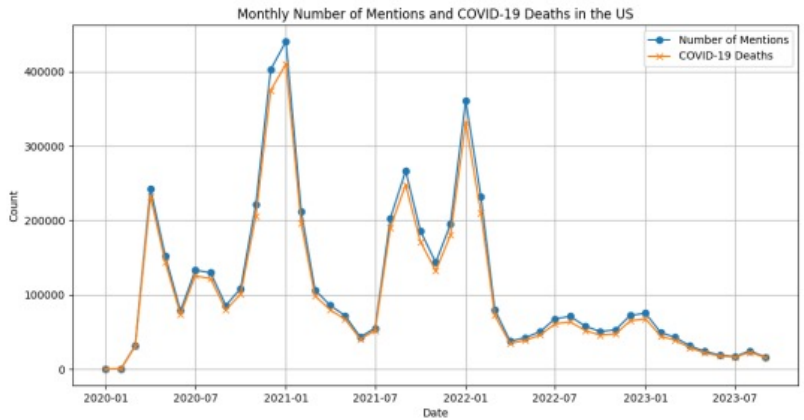
Descriptive Statistics:
Calculated descriptive statistics for each cluster.
Observed considerable variability within clusters, especially in Clusters 1 and 2.
The mean values of COVID-19 deaths and mentions increased progressively from Cluster 0 to Cluster 2, highlighting a gradient of COVID-19 impact.

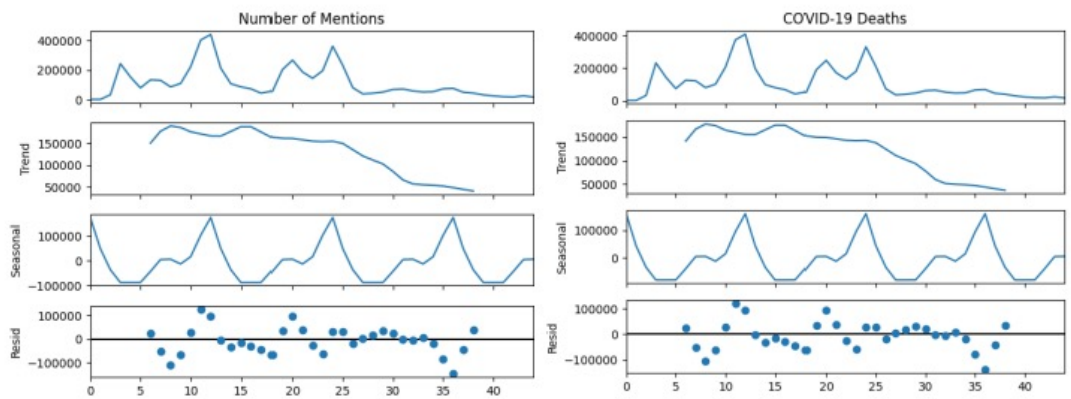
Future Use:
The K-means clustering results can be used in future steps of an analytics pipeline.
Cluster labels can be used as new features in feature engineering or incorporated into predictive modeling to improve accuracy and context-awareness.
Visualizations and reports based on cluster differences can enhance the communication of insights and support better-informed decision-..

Analyzing Time Series Data

The objective of this analysis was to explore and model the time series data of COVID-19 deaths and the number of mentions of conditions on death certificates to understand trends, seasonality, and make future forecasts.

The line chart displays monthly data for "Number of Mentions" and "COVID-19 Deaths".
Both metrics exhibit significant fluctuations over time, with multiple peaks corresponding to the waves of the pandemic.
Notable peaks occur around the winter months, indicating seasonal surges in COVID-19 cases and deaths.





Decomposition of the Variables
Observed: The top panel shows the observed values of the variables, mirroring the peaks and troughs seen in the line chart.
Trend: The second panel highlights a downward trend over time, suggesting a decrease in the number of mentions and deaths as the pandemic progresses.
Seasonal: The third panel illustrates periodic fluctuations, confirming a seasonal pattern in the data with regular increases during certain months.
Residual: The bottom panel shows residuals (random noise), which appear to be relatively random, indicating the trend and seasonal components capture ..

Summary of Insights

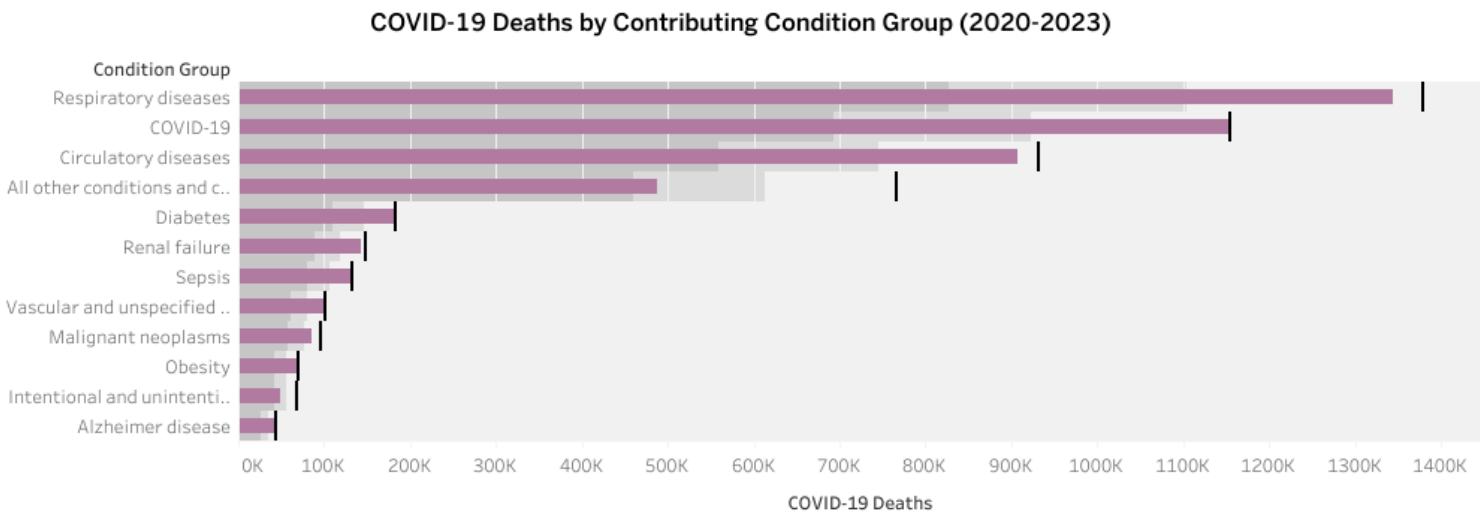
Key Insights

Correlation Analysis:
Strong positive correlation between COVID-19 deaths and the number of mentions of specific 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.

Geographical Analysis:
High total COVID-19 deaths in populous states like California, Texas, and Florida.
Highest death rates per 100,000 population in states like North Dakota, South Dakota, and Mississippi, indicating severe impacts relative to their population sizes.

Linear Regression Analysis:
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.

Cluster Analysis:
K-means clustering identified four clusters representing varying levels of COVID-19 impact.
Clusters ranged from minimal to very high impact, providing a structured way to interpret the data that can aid in targeted public health responses.
..



Limitations and Potential Biases

Limitations

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

Reporting Delays:
Reporting delays can range from 1 week to 8 weeks or more, meaning the data for recent periods may be incomplete. However, data for 2020 and 2021 are based on final data.

Inconsistent Reporting Standards:
Different states may have varying standards for reporting COVID-19 deaths and contributing conditions, which can make comparisons across states less reliable.

Multiple Conditions:
On average, there are 4 additional conditions per death, which may complicate the analysis.

Double Counting Risk:
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.

Population Data:
The population data used for analysis is from 2020, which may not accurate..

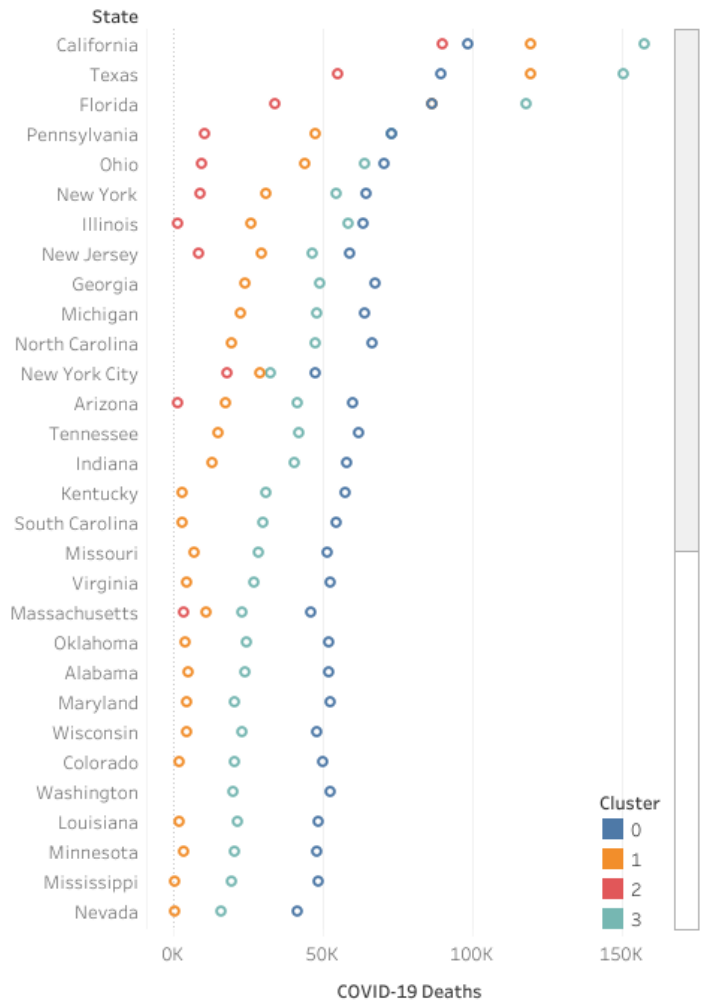
Potential Biases in the Dataset

Reporting Bias:
Inconsistent Standards: Varying state standards for reporting COVID-19 deaths and conditions can introduce biases.
Data Suppression: Suppressed counts for confidentiality can affect the completeness of the data.

Selection Bias:
Certain demographic groups may be underrepresented, impacting the accuracy of the analysis.
Differences in urban vs. rural reporting can lead to geographic biases.

Measurement Bias:
Multiple Conditions Reporting: Deaths with multiple conditions are counted in each category, potentially overestimating condition prevalence.
Non-Summation Rule: Summing conditions across categories n..

COVID-19 Deaths by State and Cluster Analysis



Recommendations and Next Steps

Recommendations

Public Health Preparedness:
Use insights from the analysis to improve preparedness for future pandemics, focusing on states that showed high death rates and were severely impacted.

Healthcare Resource Allocation:
Allocate healthcare resources and develop infrastructure based on the understanding of conditions that significantly contributed to COVID-19 deaths.
Prioritize healthcare facilities in regions with high death rates to ensure better preparedness for future health crises.

Ongoing Monitoring:
Maintain and update health data repositories to enable continuous analysis and readiness for unexpected health events.

Next Steps for Continuing the Analysis

Model Implementation:
Implement the ARIMA model with identified parameters for both mentions and deaths. Evaluate and refine the model using Mean Squared Error (MSE) and other metrics.
Explore seasonal ARIMA (SARIMA) models to account for strong seasonal patterns.

Enhanced Clustering:
Perform clustering analyses on additional conditions and demographics to uncover more granular patterns and insights.

Predictive Modeling:
Develop predictive models to identify high-risk populations based on the presence of certain conditions, demographics, and geographic data.
Identify factors most predictive of COVID-19 death rates to enhance public health strategies.

Bubble Chart of Deaths vs. Number of Mentions