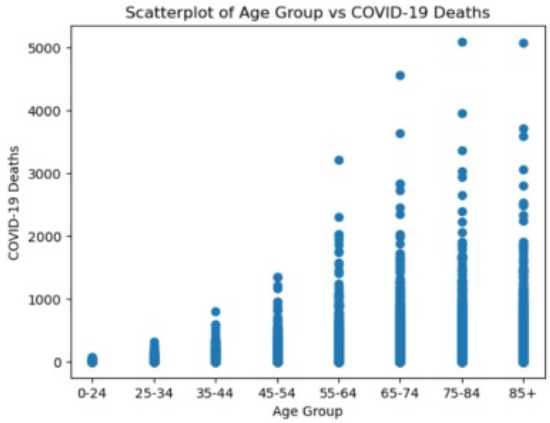
Analysis of COVID-19 Data

Analysis of Covid-19 Data

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



Introduction to the Problem

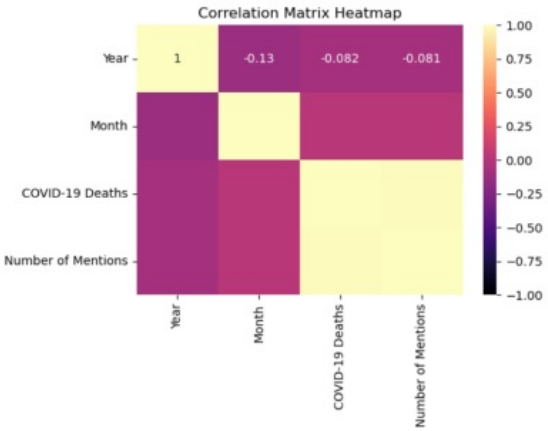
Research Focus: This study aims to analyze the factors contributing to COVID-19 deaths across various demographics and regions within the United States. By understanding these elements, public health officials and policymakers can better focus their interventions.

Importance: The COVID-19 pandemic has had a profound effect on global health. Gaining insights into the contributing conditions is crucial for guiding targeted health interventions and policy decisions, ultimately reducing mortality rates and enhancing public health outcomes.

Analysis of COVID-19 Data

Initial Data Exploration

The initial data exploration provides strong evidence of the relationships between COVID-19 deaths and COVID-19. The findings highlight the importance of considering demographic and condition-specific factors in understanding and managing the impact of COVID-19.



Demographic Analysis:

COVID-19 Deaths and Number of Mentions:

Strong Positive Correlation: Lighter color boxes represent a strong positive correlation between 'COVID-19 Deaths' and 'Number of Mentions'.

Year and Month:

No Significant Linear Relationship: 'Year' and 'Month' show minimal correlation with COVID-19 deaths and mentions. Temporal trends are complex, influenced by infection waves, reporting changes, or interventions.

Predictive Modeling:

COVID-19 Deaths and Number of Mentions:

Strong Predictive Potential: Mention counts predict COVID-19 deaths effectively. Incorporating this feature enhances mortality predictions, aiding real-time healthcare resource allocation.

Year and Month:

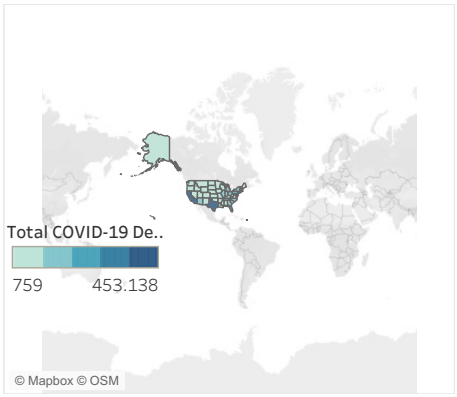
Non-linear Effects: While not directly correlated, 'Year' and 'Month' interact with other factors in predictive models. Advanced techniques (e.g., time series analysis, machine learning) capture their complex relationships.

..

Analysis of COVID-19 Data

Geographical Analysis

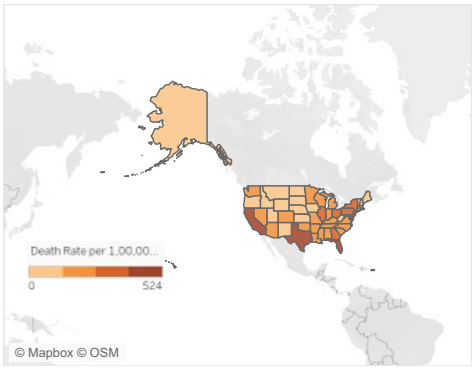
Total COVID- 19 Deaths



Total COVID-19 Deaths Map
High Total Deaths in Populous States:
States with significant COVID-19 death tolls typically coincide with those having large populations. This correlation arises due to the sheer size of their resident populations, leading to higher absolute numbers of fatalities attributed to the virus.

Regional Variations:
When comparing both maps, it becomes evident that states with large populations such as California, Texas, and Florida exhibit high total COVID-19 deaths. However, the death rates per 100,000 people provide a clearer indication of the outbreak's severity relative to each state's population size. This disparity suggests considerable differences in public health responses and healthcare capacities among states, underscoring the varying impacts of the pandemic across different region..

COVID- 19 Death Rates per 1,00,000 Population



COVID-19 Death Rates per 1,00,000 Map
High Death Rates in the Midwest and South:
The highest death rates per 1,00,000 population are observed in states like Washington, Colorado, and Minnesota. These high rates indicate a severe impact relative to the population size, suggesting significant outbreaks and possibly less effective containment measures

These two maps provide 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 1,00,000 population offer critical insights into the relative severity and effectiveness of public health measures in different regions.

Analysis of COVID-19 Data

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

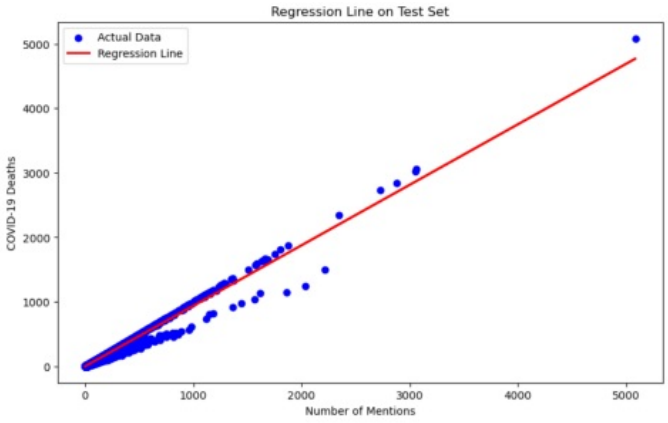
Hypothesis:
A higher number of mentions of a condition on death certificates will correlate with a significantly higher COVID-19 death count.

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

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

Run Linear Regression:
Initialized and trained a linear regression model on the training data, then made predictions on the test data.

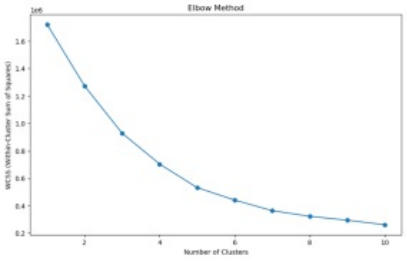
Regression Line:
Generated a plot showing the regression line on the test set, demonstrating a positive relationship between the "Number of Mentions" and "COVID-19 Deaths".



Analysis of COVID-19 Data

Cluster Analysis

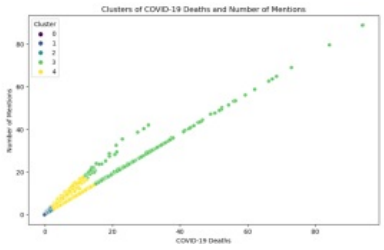
The objective of this analysis was to apply K-means clustering to the dataset containing COVID-19 deaths and contributing conditions to identify meaningful groups within the data.



The Elbow Technique:

Applied the elbow technique to determine the optimal number of clusters.
Converted the scores to positive values for plotting.

Plotted the elbow curve, which suggested that the optimal number of clusters is 5 as the curve starts to flatten at this ..



Cluster Visualization:

Created scatterplots to visualize the clusters based on the 'COVID-19 Deaths' and 'Number of Mentions'.

Cluster Analysis

Cluster interpretation:
Cluster 0: Represents minimal impact, with low numbers of deaths and mentions.
Cluster 4: 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.

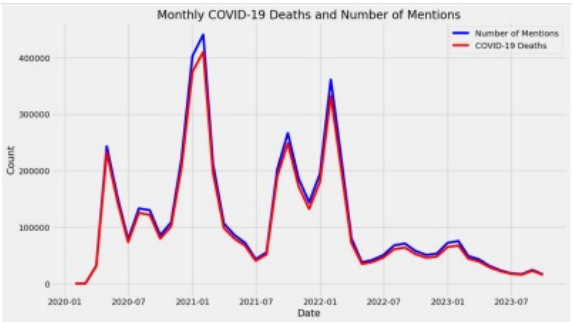
Future Use:
The K-means clustering results can be used in future steps of an analytics pipeline.
Cluster labels can be used as new features in feature engineering or incorporated into predictive modeling to improve accuracy and context awareness.
Visualizations and reports based on cluster

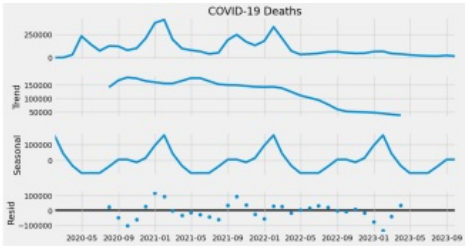
Analysis of COVID-19 Data

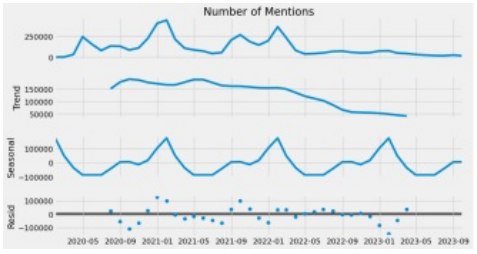
Analysing Time-Series Data

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, and 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 seen in the line chart.
Trends: 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 a regular increase during certain months.
Residual: The bottom panel shows residuals (

Analysis of COVID-19 Data

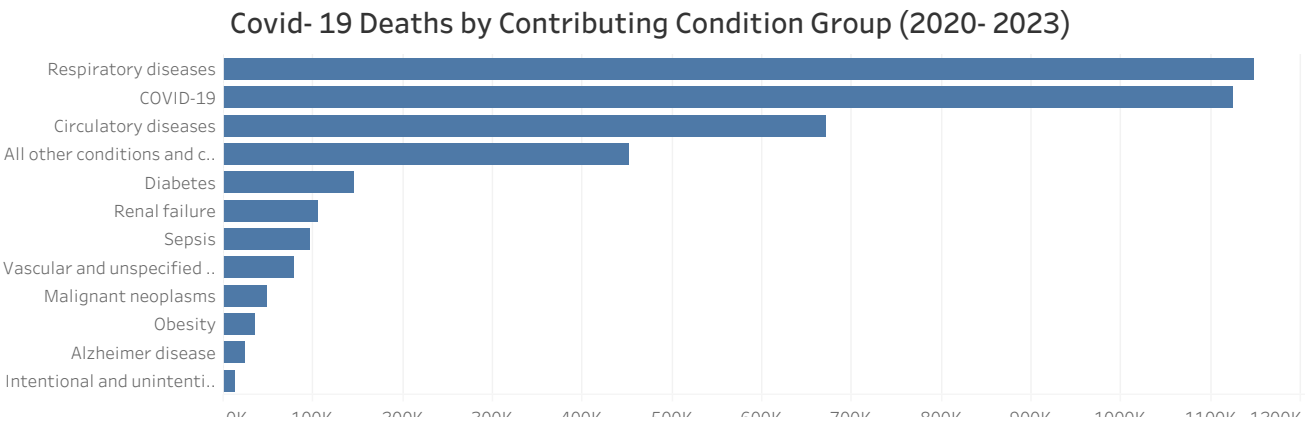
Summary and Insights

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 correlation with COVID-19 deaths.

Geographical Analysis:
High total COVID-19 deaths in populous states like California, Texas, and Florida.
Highest death rates per 1,00,000 population in states like Washington, Colorado, and Minnesota, indicating severe impacts relative to their population sizes.

Cluster Analysis:
K-means clustering identified 5 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



Analysis of COVID-19 Data

Limitations and Potential Biases in the Dat..

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 -8 weeks or more, meaning the data for recent periods may be incomplete. However, data for 20202 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.

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 need to be avoided to prevent overestimation.

Analysis of COVID-19 Data

Recommendations and Next Steps

Recommendations

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

Healthcare Resources 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 preparations 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 the Mean Squared Error (MSE) and other metrics.
Explore seasonal ARIMA (SARIMA) models to account for strong seasonal patterns.

Enhanced Clustering:
Perform clustering analyses of the 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 geographical data.
Identify factors that are most predictive of COVID-19 death rates to enhance public health strategies.