DSCI 5260 Section 501 - Business Process Analytics

Analyzing COVID-19 Mortality Factors: A Comprehensive Study Based on Age Groups and Underlying Health Conditions

Final Project Report

Venkateswara Reddy Lingareddygari

1. Introduction

The proliferation of infectious diseases poses a threat to health authorities worldwide. In the past few decades, the situation of several novel pathogens such as the Ebola, Zika, and avian influenza epidemics has shown how important preparedness and response strategies are in the fight against emerging infectious threats [1]. Amidst this background, the COVID-19 emergence has proved to be one of the most daunting challenges faced by the 21st century, having impacted health systems, economies, and societies of countries worldwide.

COVID-19 caused by the novel coronavirus SARS-CoV-2 has led to over 200 million confirmed cases and millions of deaths worldwide [2]. The rapid spread of the virus is due to transmission through respiratory droplets and aerosols. There is a great need to highlight and understand the epidemiology and transmission dynamics of the pathogen. Although efforts are made to implement infection controls and develop vaccines, the appearance of variants such as Delta and Omicron poses new difficulties which show that the virus is capable of rapid evolution and adaptation [3].

There is a growing realization that a holistic approach is necessary to minimize COVID-19's effect and to avoid similar outbreaks in the future. Epidemiological studies have contributed very much to understanding the transmission dynamics of the virus, features that make one more susceptible to severe disease, and disparities in health outcomes. Several studies have identified the role of advanced age, underlying health conditions, and socioeconomic factors as the most prominent determinants of COVID-19 severity and mortality. Meanwhile, genomic surveillance has monitored the evolution of SARS-CoV-2 as well as variants, informing public health responses and vaccine development. Through multidisciplinary learning and data analytics, researchers and policymakers can better understand and manage the complicated process involving many factors influencing the pandemics, which in turn will design the effective interventions needed to mitigate the impact

The pandemic has led to significant disruptions in various sectors, including global supply chains, with an estimated 94% of Fortune 1000 companies experiencing supply chain disruptions [4]. Economic downturns have been widespread, with global GDP contracting by 3.5% in 2020, representing the worst recession since the Great Depression [5]. Unemployment rates have surged,

with over 114 million people losing their jobs globally in 2020, according to the International Labour Organization. Lockdown measures and social distancing guidelines have forced businesses to transition to remote work setups, impacting productivity and revenue streams. Additionally, educational institutions have faced challenges in adapting to online learning, exacerbating existing inequalities in access to education. The mental health implications of prolonged isolation and uncertainty have been profound, with surveys indicating increased rates of anxiety, depression, and substance abuse worldwide [6].

Throughout the COVID-19 pandemic hospitals and clinics have had to deal with critical supply shortages and overcrowded intensive care units (ICUs). Frontline healthcare workers, who are at the frontline take the risk of burnout and exhaustion. Despite efforts to scale up testing and contact tracing, bottlenecks in testing capacity especially for those communities that are located far apart from the testing centers remain, with noted disparities in testing rates across regions. Vaccination campaigns have been very important in the fight to reduce the spread and severity of COVID-19 worldwide, over 10 billion vaccines have been administered as of January 2024 [7]. Nevertheless, some people have shown hesitation in using vaccines as well as unjust distribution of vaccines among low-income communities

Governments, international organizations, and research centers have deployed all their resources to develop vaccines, treatments, and public health interventions. Collaborative efforts have resulted in the acceleration of the development and rollout of many vaccines. mRNA vaccines have been proven with high efficacy rates in clinical trials. In light of these developments, differences in the distribution of vaccines have, however, remained, thus low-income demographics experience problems in accessing enough vaccines [8]. In terms of research into novel therapeutics like monoclonal antibodies and antiviral drugs, they are developing and can provide new tools for the management and treatment of COVID-19. The ongoing surveillance efforts, combined with research and innovation, are likely to be vital in managing the pandemic and the emerging challenges.

This study sought an in-depth analysis of the factors of COVID-19 mortality classifying them by state and age group within the U.S. Through utilization of data science approaches, we intend to look into all the components such as geographic and healthcare elements that impinge on the death

rates. Our data analysis comes from rigorous data analytics that include but are not limited to statistical models and trend analysis, which enables us to come out with hidden patterns, trends, and disparities. We attempt harmonious solutions by supplying practical solutions and evidence-based recommendations for decision-making and public health policies. In addition to the pandemic being in place, with the elderly and others vulnerable to the virus and new variants, our work is aiming at creating a complete understanding of the evolution of mortality rates. Healthcare disparity can be viewed through mortality rates across various demographic and geographical screens, to support evidence-led sourcing, improved healthcare provision, and inter-developmental disease impact reduction.

Research Questions

1. Can we identify states with the highest COVID-19 mortality rates?
2. What are the trends and patterns in COVID-19 deaths over the months and years?
3. Can we identify clusters of age groups and health conditions with similar patterns in covid 19?

4. Create a predictive model for the number of COVID-19 deaths based on age group, specific health conditions?

2. Literature review

The existing literature about the factors of COVID-19 mortality provides some clues to the multipronged issue of this pandemic. Several principal themes emerge from a critical analysis of relevant literature, which reveals the interconnectedness of various factors determining COVID-19 mortality. Numerous studies have investigated the relationship between age, health, geographic variations, and COVID-19. The studies have explored the links between age, health issue-s, geographical differences, and COVID-19 outcomes. Using different data analysis methods, researchers have conducted detailed studies to provide key details on how the pandemic affects different demographics and help make evidence-based decisions.

2.1 Age and health conditions

Age and underlying health conditions are the most important risk factors for the severity of COVID-19 infection and mortality [9]. Existing studies suggest that older people and those with underlying health conditions have the possibility of more severe outcomes associated with COVID-19. The elderly population and underlying diseases such as cardiovascular diseases, diabetes, and respiratory disorders have been consistently found to be associated with severe infection and death due to COVID-19 [10]. Many studies have pointed to how different ages can affect COVID-19 outcomes. For example, a study found that middle-aged and older people are more at risk of COVID-19 complications, often with high blood pressure and heart damage being cited as common issues [11].

Furthermore, the research on age's effect on COVID-19 complications has shown meaningful findings. For example, a New York-based study of 2634 COVID-19 patients found that 21% lost their lives and 14.2% were admitted to the ICU. 12.2% required mechanical breathing help, and 3.2% needed kidney replacement [12]. This highlights how severe the disease can be for particular age- demographics.

The age-related physiological decline and the impaired immune responses in individuals with comorbidities present the conditions of COVID-19-related complications, resulting in deadly mortality rates in these populations.

2.2 Geographic Variations

Geographical variations in COVID-19 mortality rates are related to different contextual factors including healthcare infrastructure, socioeconomic status, and public health interventions. Differences in healthcare access and quality of care contribute to the different outcomes in urban versus rural areas. Research has drawn upon health electronic surveillance networks to understand COVID-19 trends in a geographical context. For example, a study done across Saudi Arabia concerning COVID-19 complications in different regions showed varying results. While everyone- had a fever and cough, some are-as had more severe issues. More serious respiratory complications were identified to depend on population, healthcare infrastructure, and socioeconomic status [13]. Be-sides that, COVID-19 travel rules change the rate and direction of transmission. A study looked at travel patterns using data from Baidu migration and Google’s

Community Mobility Report. The data shows a big drop in transmissibility after the implementation of the travel restrictions [14].

Key socioeconomic factors shaping COVID-19 mortality patterns include income inequality, housing conditions, and access to healthcare services. Additionally, the implementation of public health measures which include testing, contact tracing, and vaccination campaigns, differ across different regions, which in turn causes a difference in mortality rates.

2.3 Data Science Techniques and Studies

Researchers have used a variety of data analysis techniques to study age, health conditions, and geographical differences. These techniques allow researchers and data scientists to conduct in-depth analyses of complex datasets. Researchers have used logistic regression models and machine learning to find indicators that can predict COVID-19 patient outcomes and the risk factors linked to these outcomes. In one study [15], logistic regression models were employed to develop a diagnostic model for COVID-19. The correlation of each diagnostic regressor with COVID-19 was calculated using the Chi-square method, and the important regressors were identified. The study computed the binary logistic regression model and yielded a specificity, sensitivity, and accuracy of 97.3%, 98.8%, and 98.2%, respectively.

2.4 Comprehensive Analyses

Although the majority of prior studies linked age, health status, and geographic characteristics with COVID-19 mortality, integrating the multiple determinants to understand the mortality patterns is essential. Focusing on the strategies that account for the synergistic impacts of demographic, clinical, and environmental factors is critical for the development of a comprehensive understanding of COVID-19 mortality dynamics.

2.5 Synthesis

The existing literature highlights the complexity that COVID-19 mortality faces, showing an intricate network of factors that shape the pandemic. Age and some diseases act as separate determinants, putting elderly people (advanced age) and populations with comorbidities at higher

risk of severe consequences. Moreover, it emphasizes the need to target intervention for a marginalized population. Additionally, geographical differences matter greatly, as the extent of infrastructure, economic conditions, and public health measures in different regions leads to variability in the levels of mortality rate.

The utilization of data science tools has revolutionized the knowledge about COVID-19 dynamics, allowing for the pinpointing of predictive markers and risk factors through the utilization of methods such as logistic regression and machine learning. Nevertheless, individual studies are invaluable in shedding light on specific aspects of COVID-19 mortality. Through the application of interdisciplinary approaches and data-focused approaches, researchers and policymakers can develop measures to reduce disparities, enhance healthcare services, and ultimately reduce the magnitude of the pandemic on society as a whole.

3. Methods and results

3.1 Data Acquisition and Preprocessing

3.1.1 Data Collection

For data acquisition, we will get COVID-19 mortality CSV datasets from https://catalog.data.gov/dataset/conditions-contributing-to-deaths-involving-coronavirus-disease-2019-covid-19-by-age-group. The dataset contains the conditions which led to death by state, age group, and several other related items. The tags consist of the period the data was gathered, how the groups were categorized, year, month, state, flag condition group and specific condition, ICD10 codes number of ill people and age group, COVID-19 deaths, and the number of mentions. Each dataset consists of 621,000 rows in 14 columns all featured. Here is a breakdown of the columns in the dataset:

1. "Data As Of": Date when the data was last updated.
2. "Start Date" and "End Date": The period during which the data was collected.
3. "Group": Categorization of data based on certain criteria.

4. "Year" and "Month": Timeframe of the data.
5. "State": Geographic location of occurrence.
6. "Condition_Group" and "Condition": Classifications and specific health conditions mentioned in conjunction with COVID-19 deaths.
7. "ICD10_codes": International Classification of Diseases (ICD-10) codes corresponding to the mentioned health conditions.
8. "Age Group": Grouping individuals based on their age.
9. "COVID-19 Deaths": Number of deaths involving COVID-19.
10. "Number of Mentions": Frequency of mentions of specific health conditions concerning COVID-19 deaths.

The data above gives useful hints about how COVID-19 had varied impacts across different states, age groups, and specific deaths contributing to COVID-19 deaths. It enables the production of complex cases and visualization of data trends along time, geography, and demographics.

3.1.2 Data Preprocessing

The analysis process is articulated here, where the imports of some key libraries, such as pandas, and seaborn, are done. After that, the COVID dataset is loaded into a data frame from a CSV file. Next, we derived the dataset information for every column and the count of non-null values. The last step was checking the total count of missing values for each column in the data frame. To avoid unnecessary entries, we are removing rows containing null values in the 'Year' column and the 'Month' column and replacing the missing values in the 'COVID-19 Deaths' and 'Number of Mentions' columns with the median.

Column	Non-Null Count	Dtype
Data As Of	621000	object
Start Date	621000	object
End Date	621000	object
Group	621000	object
Year	608580	float64
Month	558900	float64
State	621000	object
Condition Group	621000	object
Condition	621000	object
ICD10_codes	621000	object
Age Group	621000	object
COVID-19 Deaths	437551	float64
Number of Mentions	443423	float64
Flag	183449	object

Table 1: Dataset information

The next task was cleaning the data by dropping unnecessary columns like ‘Flag’, thus preparing the dataset for analysis. Also, the column sections with incompatible data formats are converted to improve the data consistency and accuracy in general.

We detected duplicates and calculated the sum of duplicate entries. Finally, comprehensive descriptive statistics are computed, presenting summary metrics such as count, mean, standard deviation, minimum, maximum, and quartiles for numeric variables within the dataset.

3.2 Exploratory Data Analysis (EDA)

The EDA stage proved to be extremely useful in highlighting important characteristics of the COVID-19 death data including the patterns and the distributions present in the data set. This section presents a comprehensive view of the vital aspects of EDA, and their consequent results, giving appropriate visualizations and statistical summaries.

3.2.1 Initial Data Exploration

Statistical Summaries

Summaries of key statistical measures were calculated for COVID-19 Deaths and Number of Mentions in the exploratory data analysis. These summaries were composed of descriptive statistics, like the mean, median, and mode, as well as measures of dispersion, such as the standard deviation and range. The examination of the data summary statistics enables us to know the central tendencies and the variation, which also informs us about the COVID-19 mortality rates.

	Year	**Month**	**COVID-19 Deaths**	**Number of Mentions**
count	558900	558900	558900	558900
mean	2021.4	6.2	30.716765	33.574384
std	1.0832	3.3506	483.371735	518.880351
min	2020	1	0.0	0.0
25%	2020	3	0.0	0.0
50%	2021	6	0.0	0.0
75%	2022	9	0.0	0.0
max	2023	12	105566.0	105566.0

Table 2: Summary statistics for COVID-19 deaths and mentions from January 2020 to December 2023.

Histograms for Each Group: A histogram was plotted for each group, which allowed the visualization of the distribution of COVID-19 deaths. These visualizations highlight the COVID-19 mortality disparity that exists between different populations and in various backgrounds. The histograms show the distribution of COVID-19 deaths for all ages. The x-axis shows the number of deaths, while the y-axis shows the frequency of deaths. To better visualize and capture the distributions in the histogram well, we handled outliers by plotting values based on a percentile threshold. From the histograms, the mortality rate is higher in older populations and lowest in younger populations. There is a significant increase in the number of deaths occurring, with the population ages 0-24 having a little to no frequency of deaths while ages 85+ have a frequency of 40 for around 250 deaths.

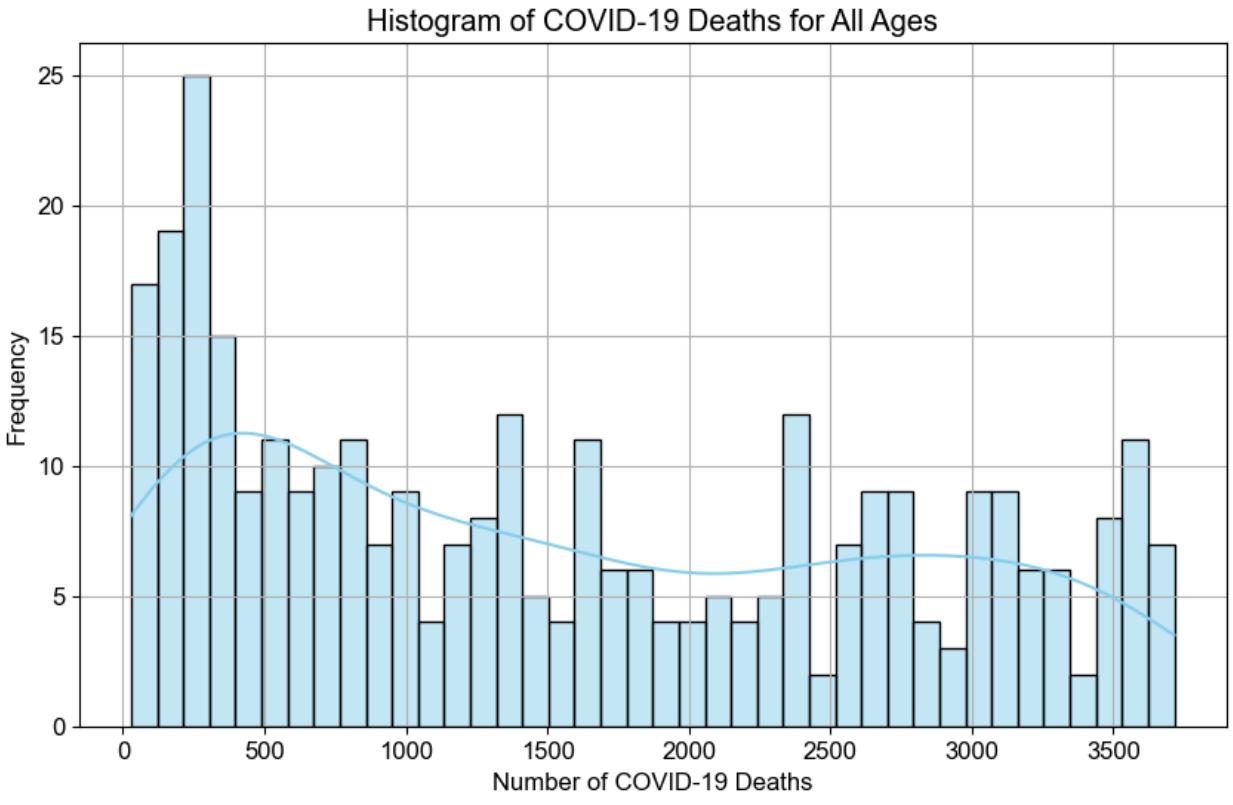


Figure 1: Histogram of COVID-19 deaths for all ages

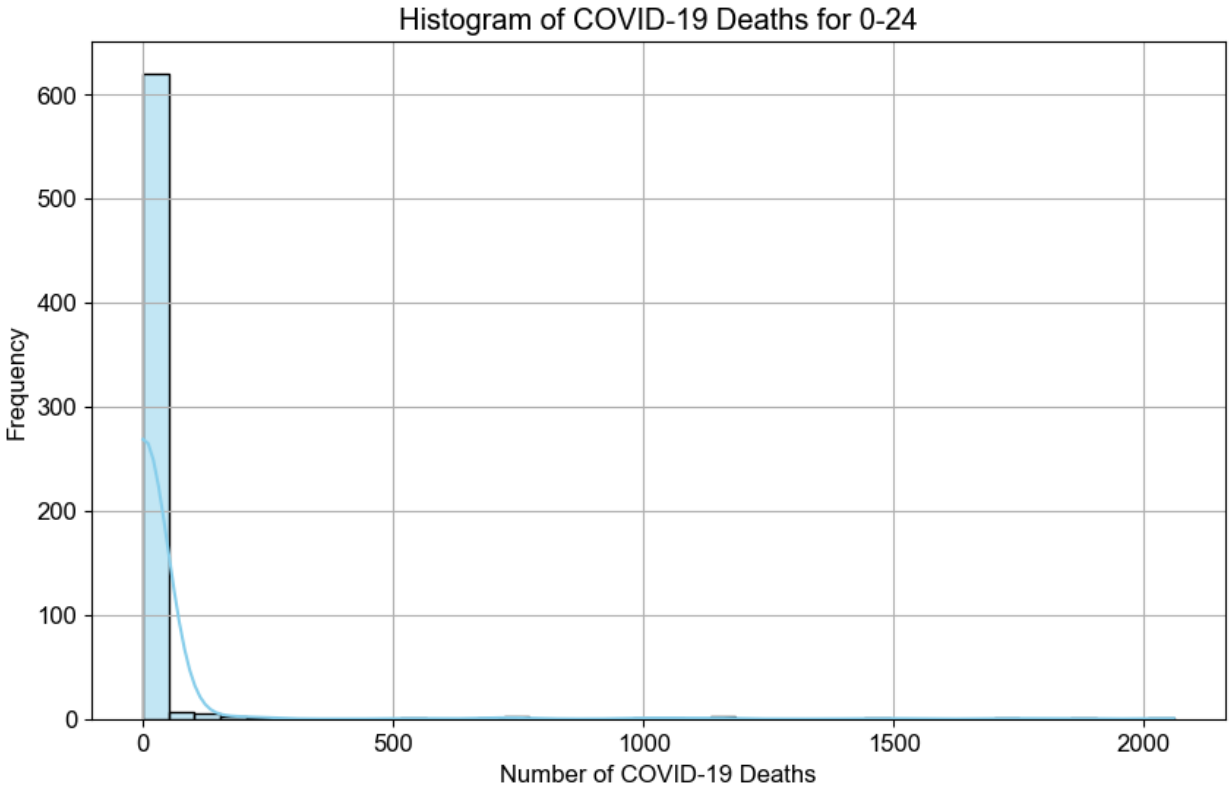


Figure 2: Histogram of COVID-19 deaths for ages 0-24

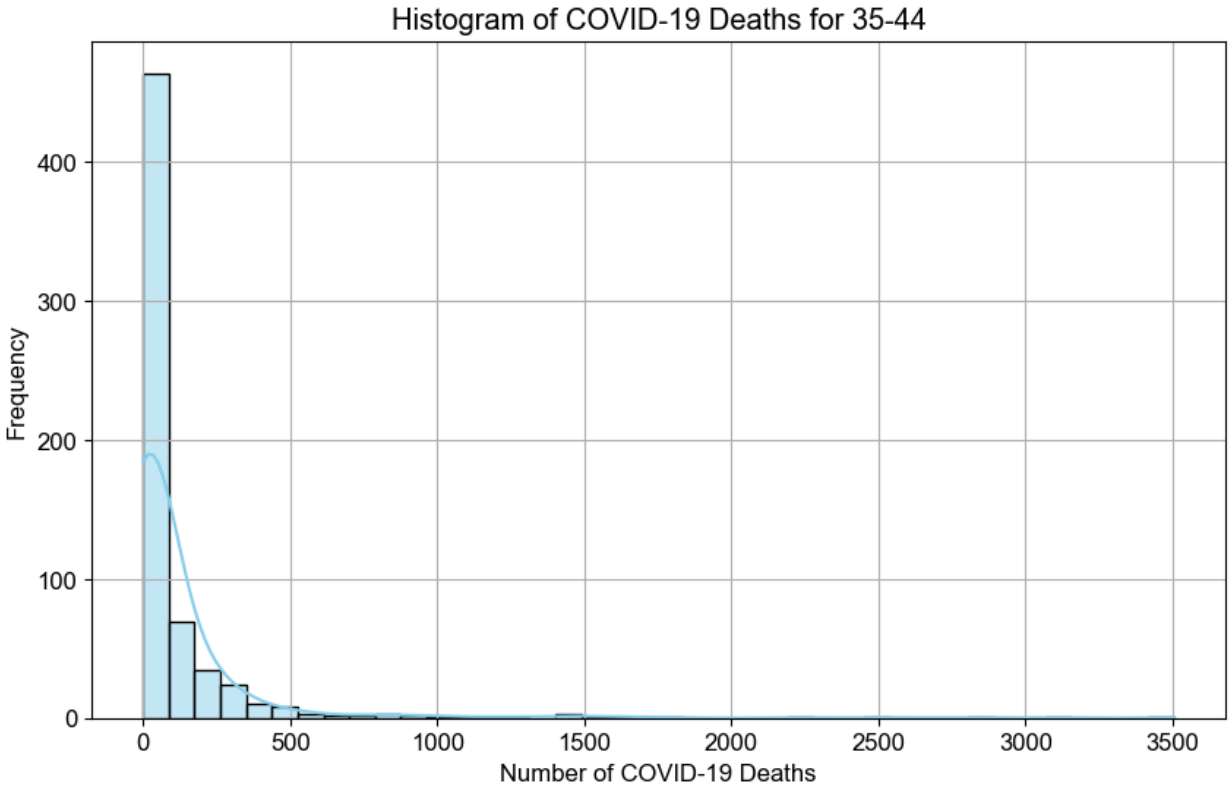


Figure 3: Histogram of COVID-19 deaths for ages 35-44

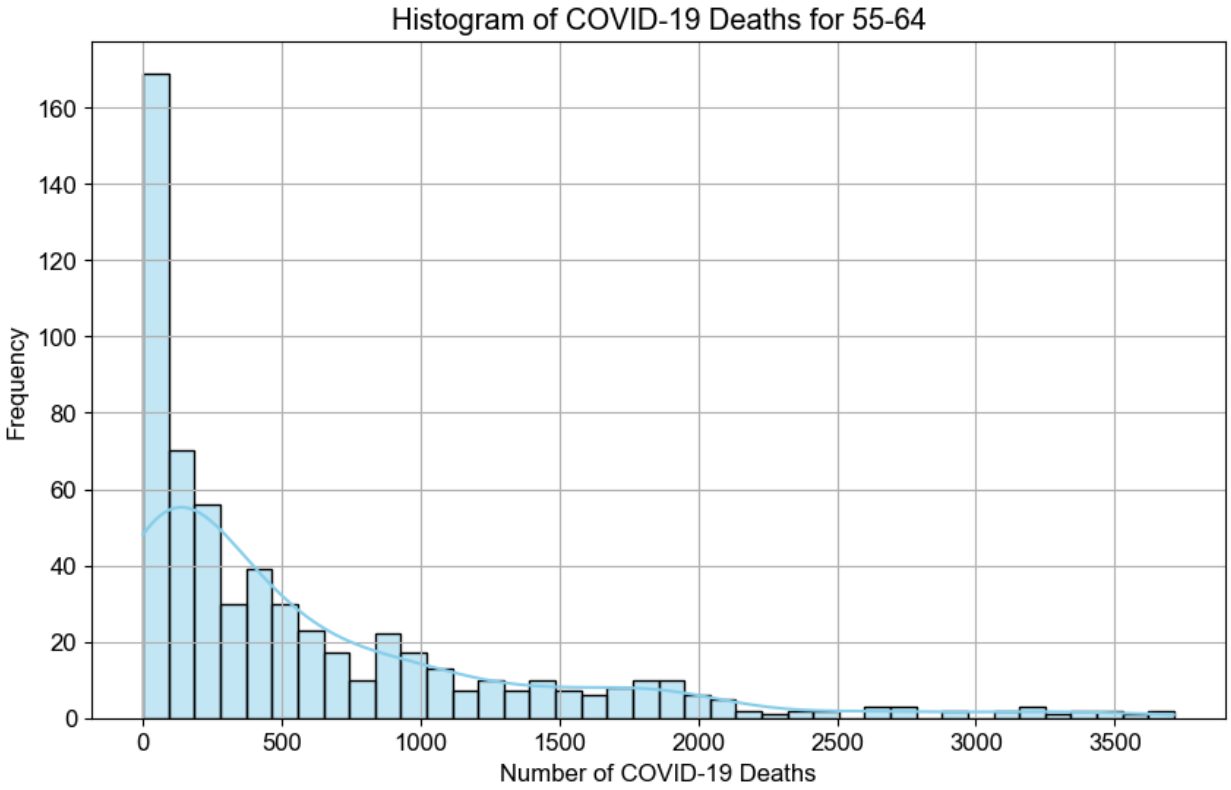


Figure 4: Histogram of COVID-19 deaths for ages 55-64

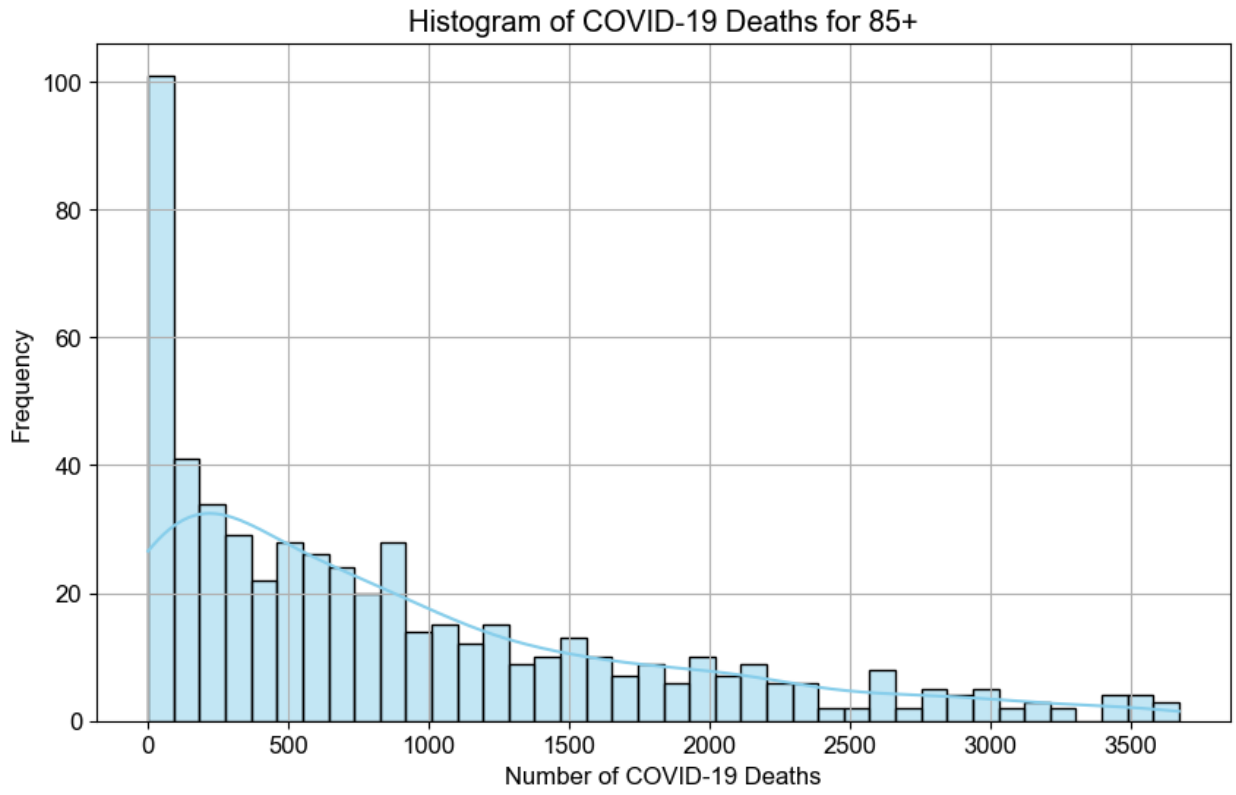


Figure 5: Histogram of COVID-19 deaths for ages 85+

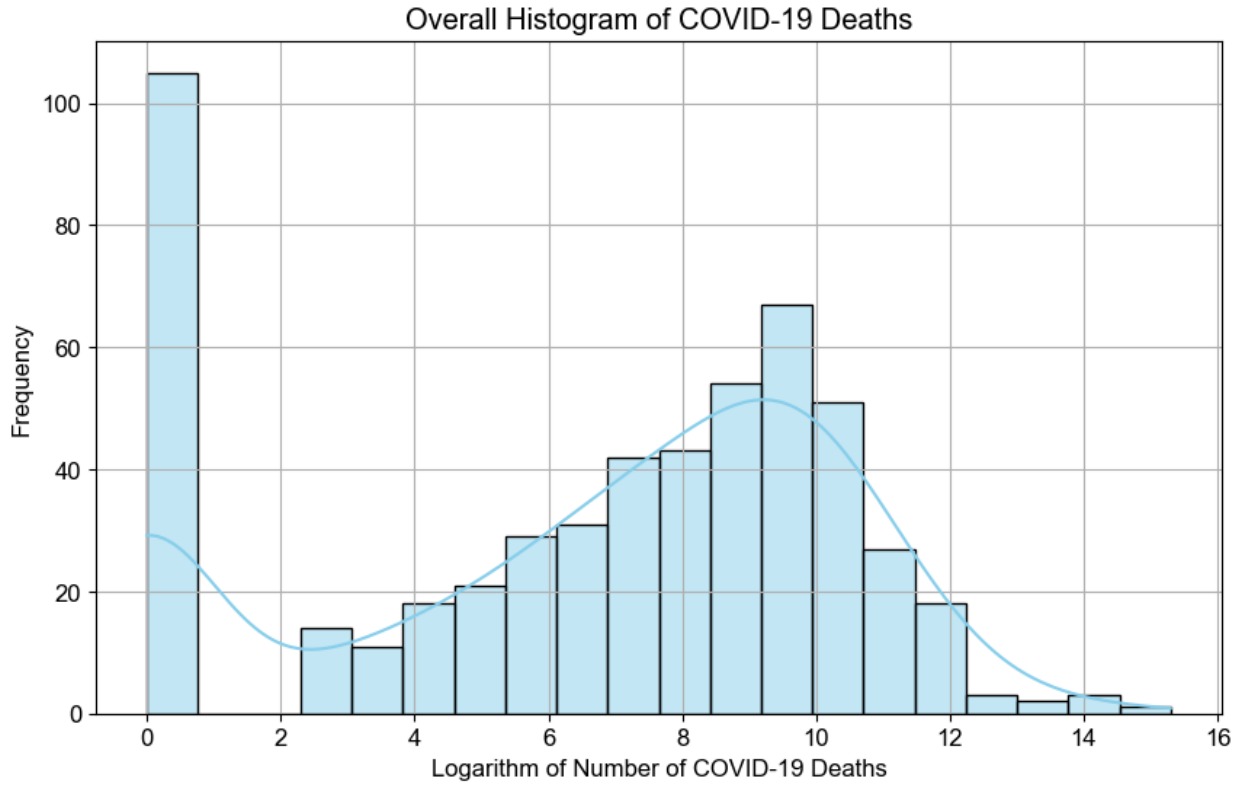


Figure 6: Histogram of COVID-19 deaths for ages 85+(Logarithmic)

3.2.2 Exploring Relationships

Temporal Trend Analysis: A line chart that reflects the current trend of COVID-19 deaths was plotted. These charts display variations in the trends of COVID-19-related metrics within the sample period. Temporal trends in COVID-19 deaths for specific conditions were also plotted.

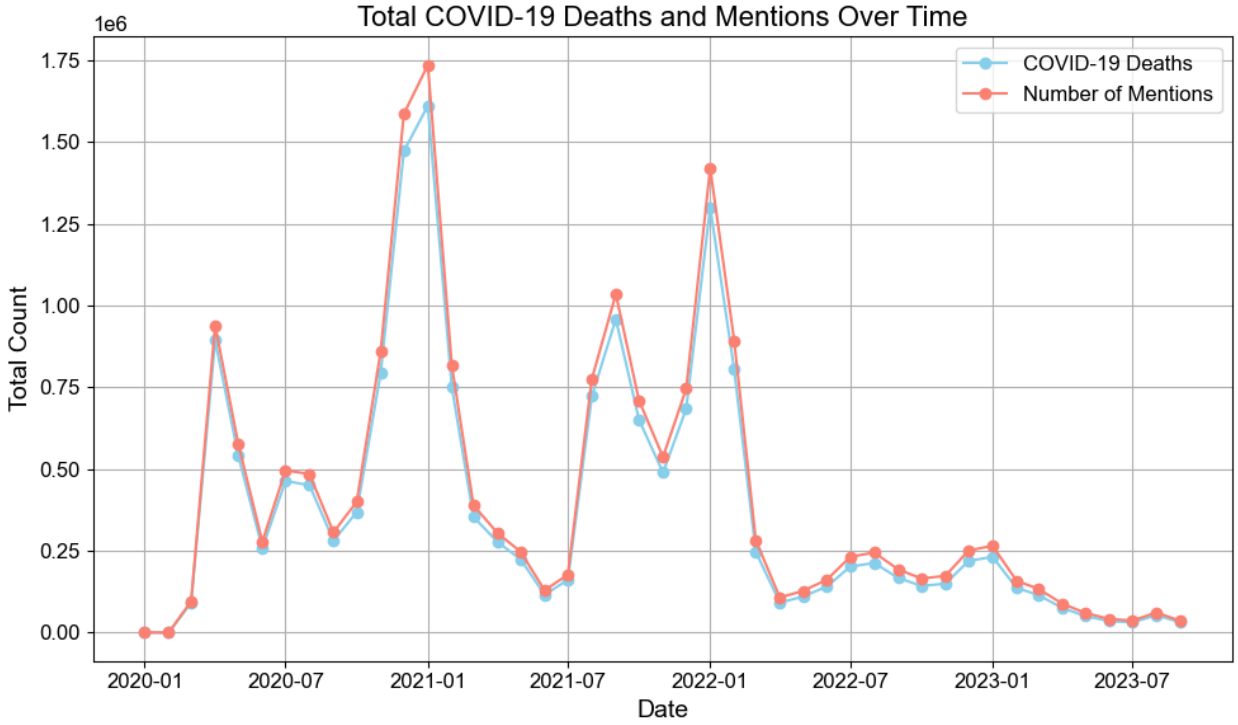


Figure 7: Total COVID-19 deaths and mentions over time

The graph shows the total number of COVID-19 deaths and the number of mentions of COVID-19 over time. The number of fatalities and mentions has decreased since early 2022. The number of deaths is in blue, while the number of mentions is in red. The graph uses a dual-axis scale. The peak of COVID-19 deaths was in 2021, with the lowest recorded deaths reported before 2020. There is a spike in covid deaths in 2022 followed by a steep decline to where the trend stabilizes from late 2022 to late 2023.

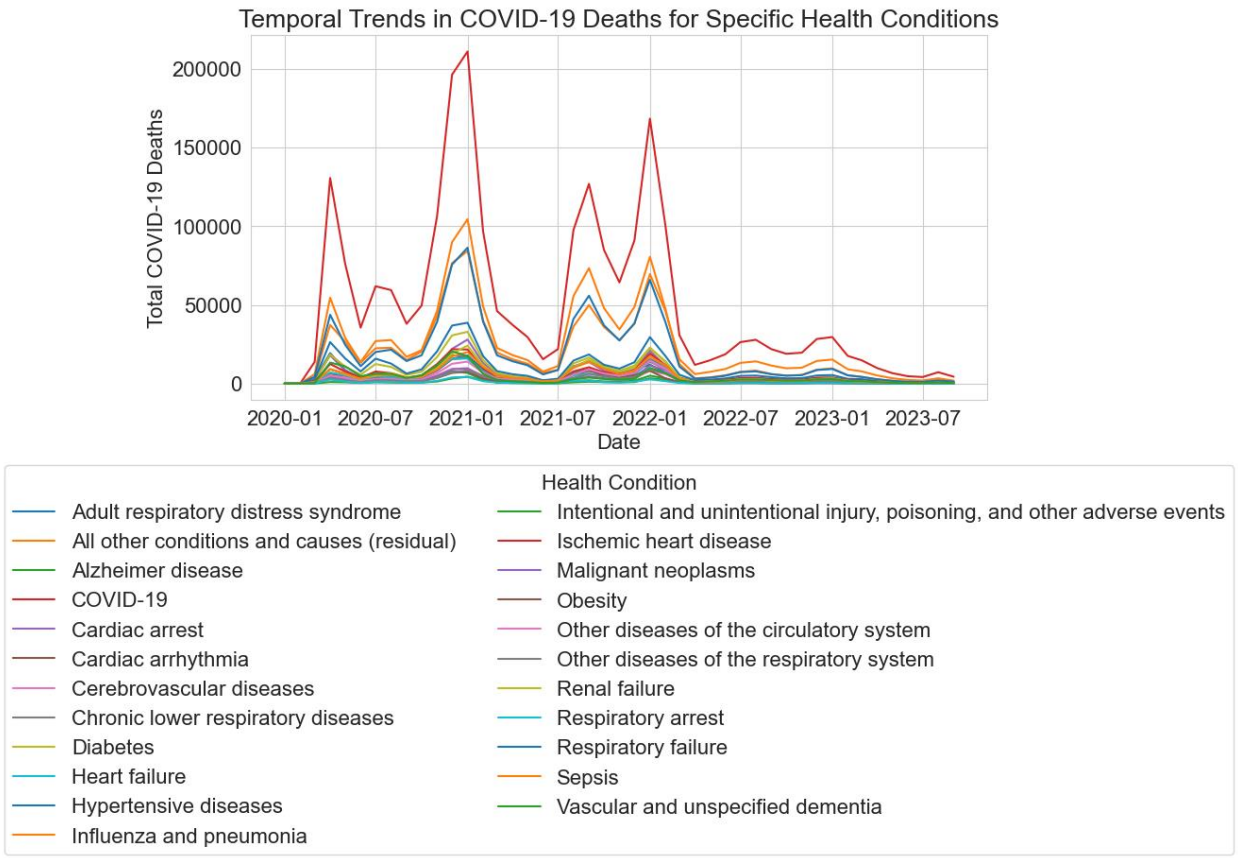
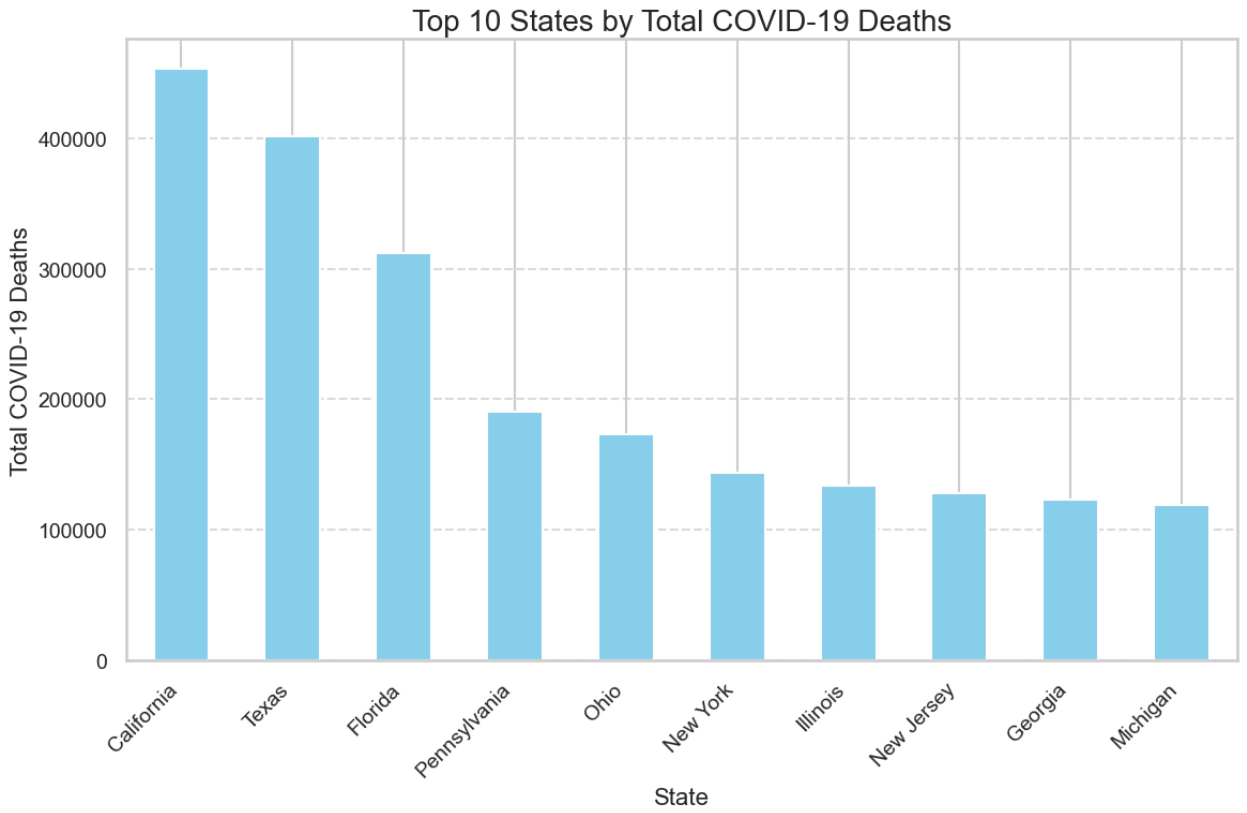


Figure 8: Temporal trends for specific health conditions alongside COVID-19

From the line plot, we can see that COVID-19, followed by respiratory arrest and respiratory failure have been the most common ailments over the past years.

COVID-19 Mortalities by state: Here we extracted and visualized the states with the highest COVID-19 deaths over the entirety of our dataset. We can see that California and Texas have the highest total COVID-19 deaths.



Distribution by Age Group: Bar plots were constructed to display the distribution of COVID-19 deaths among various age groups. Through this, the visuals show a step-by-step breakdown of the pandemic by age group and display the disparities and hotspots in mortality rates, thus providing an overall picture of the pandemic's effect in the given geographical region.

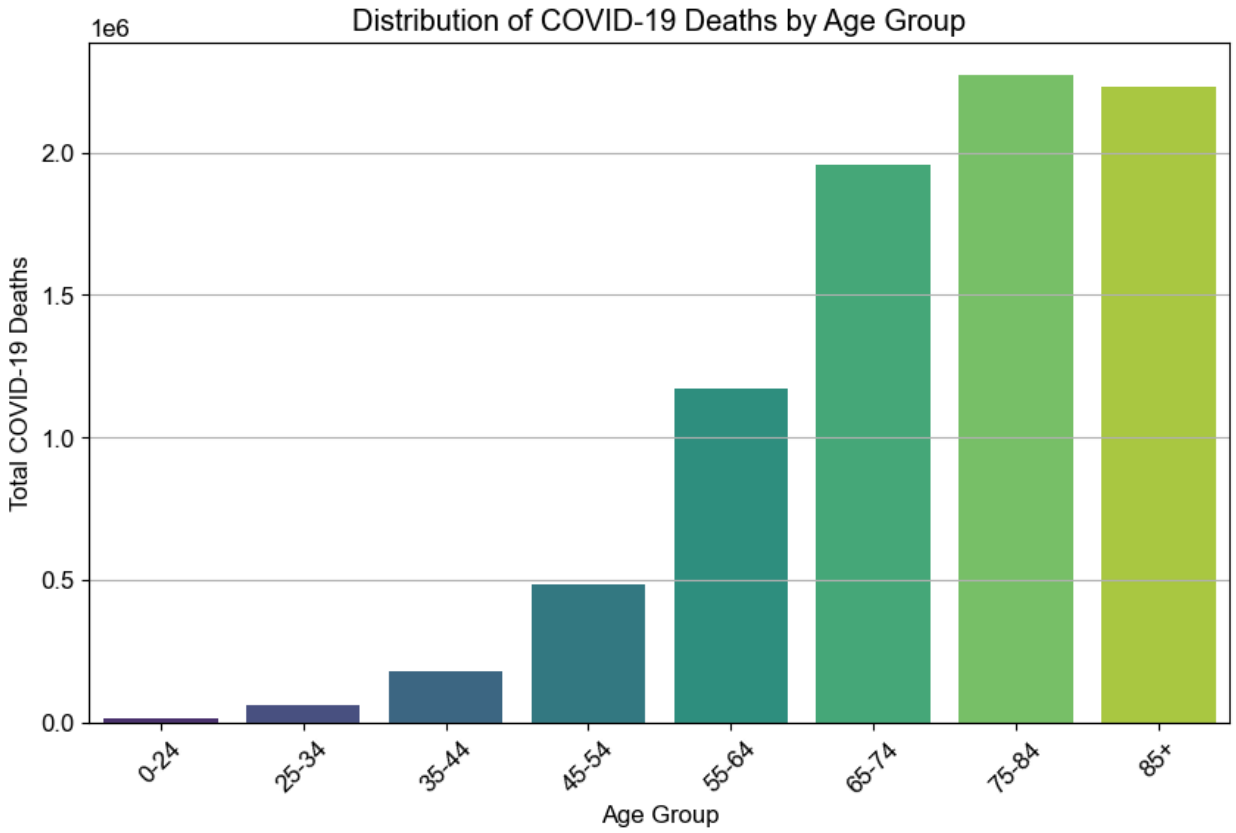


Figure 9: Distribution of COVID-19 deaths by age group

The age group 0-24 recorded the lowest number of COVID-19 deaths with a gradual increase with the increase in age. The mortality rate sped up from the age group of 45-64 to 75-84. The age group 75-84 recorded the most COVID-19 deaths. They were followed by a close peculiar second of the age group of 85 years and above.

3.2.3 Clustering Analysis

Clustering analysis is the process of finding natural groupings or clusters within a given dataset. In this study, clustering consists of searching for occurring trends in the death rate for different age groups and comorbidities. Particularly, the K-means algorithm was used to group the age groups and underlying comorbidities according to their mortality rates as related to COVID-19. This enabled the identification of patterns and disproportionate indices from where the causes of high mortality rates might be revealed.

Selection of Features: The cluster analysis was performed by making the total number of COVID-

19 deaths and the age group the main features of the analysis.

Clustering Algorithm: The K-mean algorithm we also used for decision-making. The K-means is an unsupervised learning algorithm that is quite popular and runs by partitioning data into 'k' clusters with a commonality metric.

Number of Clusters: During the first stage of our inspection, we tried various numbers of clusters and found three clusters were the optimum choice for making sense of the data.

Clustering Process: The first step was to initiate the K-means algorithm with the number of clusters desired (k=3), as it is for the case with the Modularity-based algorithm, according to the desired number.

The resulting clusters represent distinct groups of age demographics based on their COVID-19 mortality rates: The resulting clusters represent distinct groups of age demographics based on their COVID-19 mortality rates.

1. **Cluster 0 (Low Risk):** This demographic makes up the low-risk age groups with a lower mortality rate due to the disease.
2. **Cluster 1 (Medium Risk)**: People in this age cluster have a moderate risk of dying from COVID-19.
3. **Cluster 2 (High Risk):** The cluster includes segments of the population with the highest mortality rates caused by COVID-19.

Age Group	COVID-19 Deaths	Cluster
0-24	17031	0
25-34	64016	0

35-44	182292	0
45-54	486020	0
55-64	1173703	2
65-74	1954722	1
75-84	2270691	1
85+	2228145	1

Table 3: Clustering age group based on mortality rates

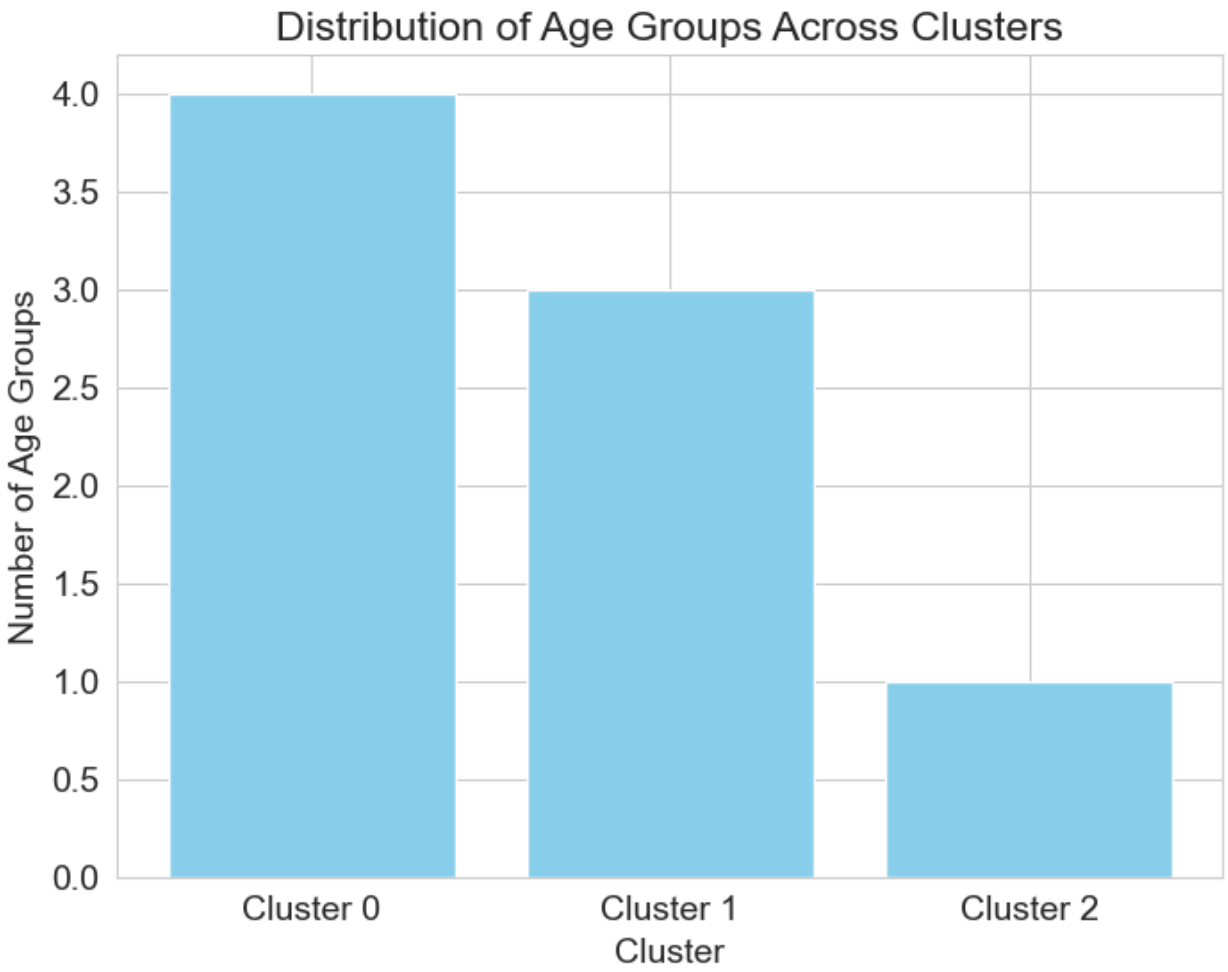


Figure 10a: Distribution of COVID-19 deaths across clusters(Bar chart)

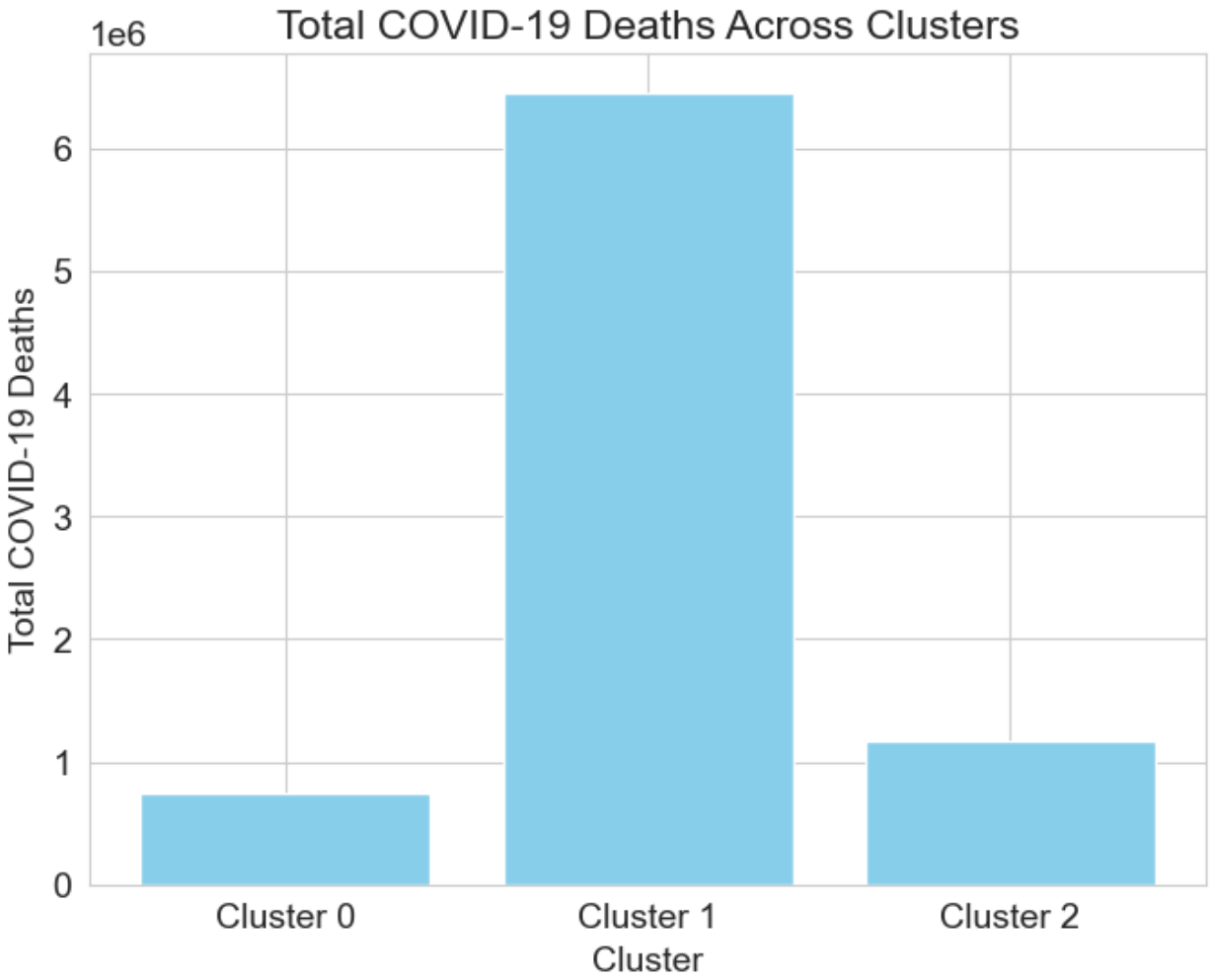


Figure 10a: Distribution of COVID-19 deaths by age group(Bar chart)

A clustering of k=5 is used for clustering by underlying conditions. This was done to account for all the various comorbidities present in the population. A similar approach was taken, where 0 meant the comorbidity with the lowest risk and 4 meant the comorbidity with the highest risk of having a fatal outcome.

Condition	***Cluster***
Adult respiratory distress syndrome	*4*

All other conditions and causes (residual)	*2*
Alzheimer disease	*3*
COVID-19	*1*
Cardiac arrest	*4*
Cardiac arrhythmia	*4*
Cerebrovascular diseases	*3*
Chronic lower respiratory diseases	*4*
Diabetes	*0*
Heart failure	*4*
Hypertensive diseases	*0*
Influenza and pneumonia	*2*
Intentional and unintentional injury, poisoning (X40-X49, X60-X84, Y10-Y19, Y35, Y87.0)	*3*
Ischemic heart disease	*4*
Malignant neoplasms	*3*
Obesity	*3*
Other diseases of the circulatory system	*4*
Other diseases of the respiratory system	*3*
Renal failure	*4*

Respiratory arrest	*3*
Respiratory failure	*2*
Sepsis	*4*
Vascular and unspecified dementia	*4*

Table 4: Clustering underlying health conditions based on COVID-19 deaths.

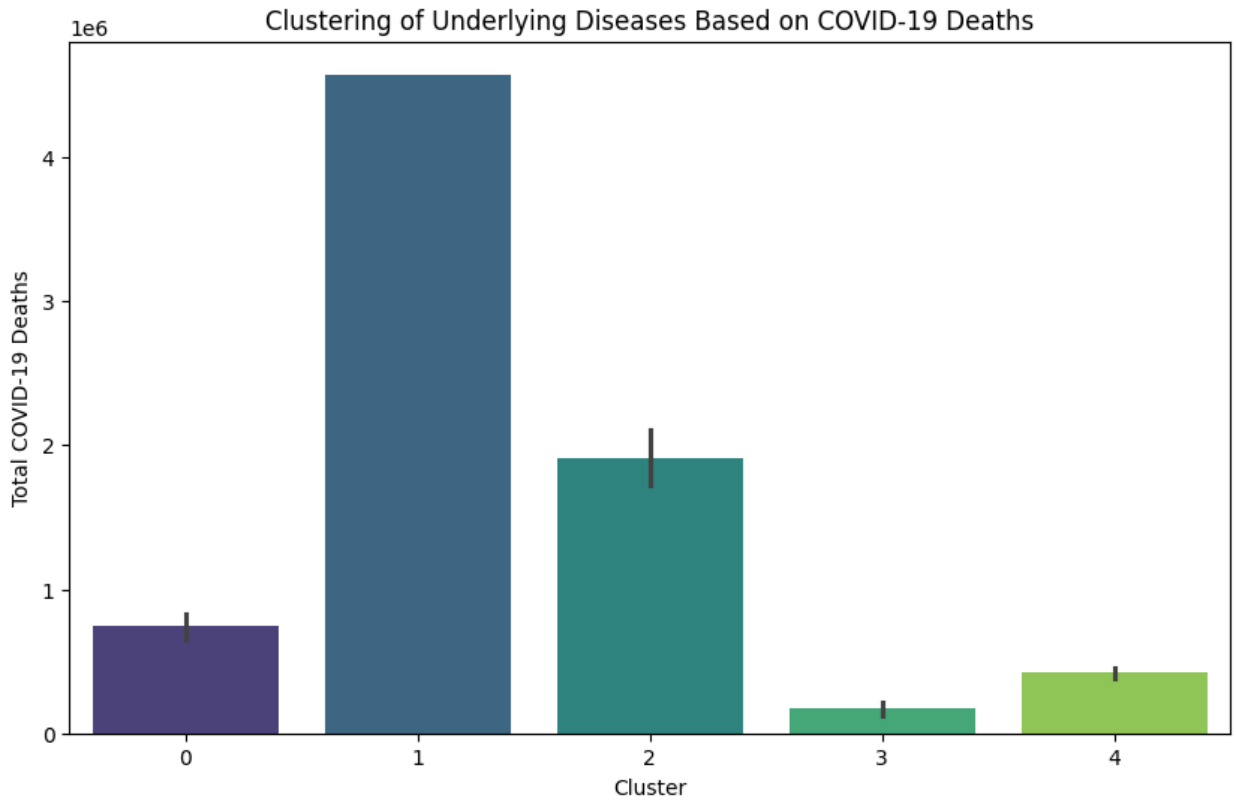


Figure 11: Clustering underlying health conditions based on COVID-19 deaths

Next, we perform clustering using two features namely age group and condition group. A

clustering of k=5 was chosen for this analysis. Cluster 1 with the highest mortality, mainly elderly ages (85+, 65-74, 75-84), was caused by the effect of COVID-19 on the elderly due to health conditions. Cluster 2, which is characterized by a lower mortality rate, still impacts older age groups but is associated with general underlying conditions rather than specific diseases. Cluster 0 and 3 correspond with moderate mortality of a broader age range, in which everyone had suffered from the general conditions.

Condition Group	Age Group	COVID-19 Deaths	Cluster
All other conditions and causes (residual)	0-24	2870	0
All other conditions and causes (residual)	25-34	7862	0
All other conditions and causes (residual)	35-44	20986	0
All other conditions and causes (residual)	45-54	54135	0
All other conditions and causes (residual)	55-64	132728	4
...continuation	...continuation	...continuation	...
Vascular and unspecified dementia	45-54	98	0
Vascular and unspecified dementia	55-64	1321	0

Vascular and unspecified dementia	65-74	12157	0
Vascular and unspecified dementia	75-84	51639	0
Vascular and unspecified dementia	85+	107461	4

Table 5: Clustering by age group and condition group (Sample)

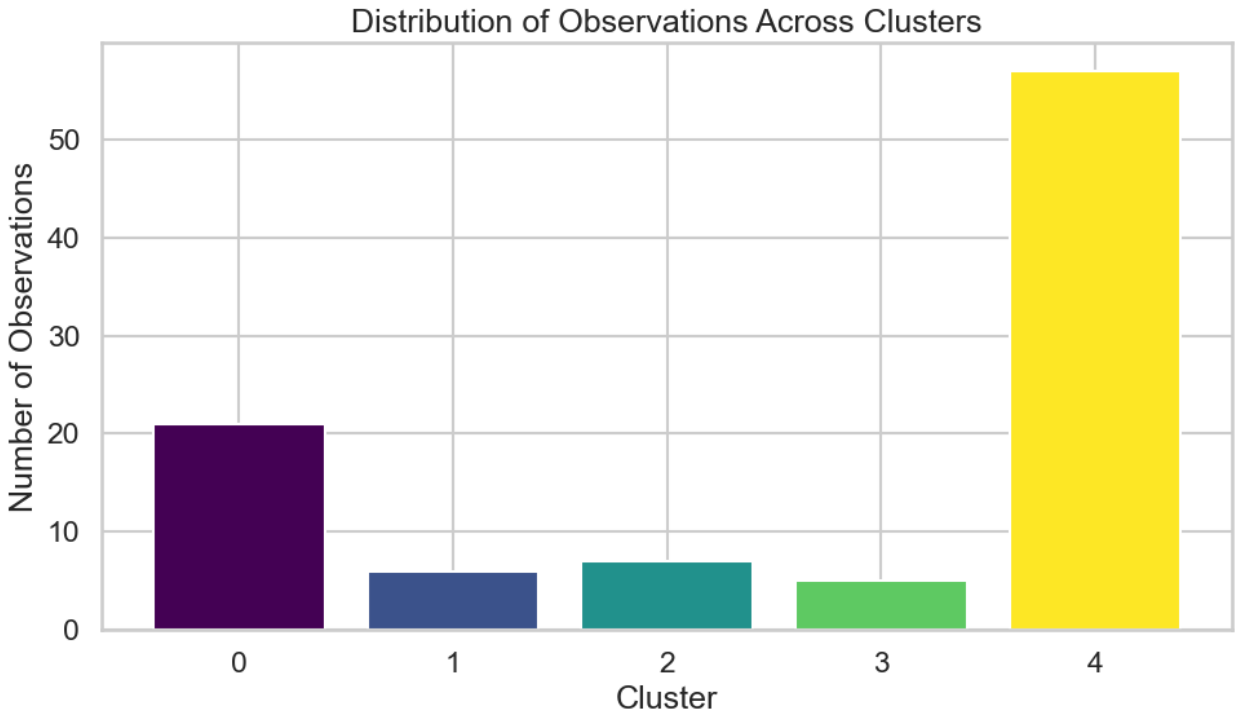


Figure 11a: Distribution of Observations Across Clusters

3.3 Predictive Modeling

In our next section, we will perform predictive analysis. The predictive model aims to forecast COVID-19 mortality rates and identify the hidden patterns behind the data. We use a step-by-step strategy to come up with model selection and development.

3.3.1 Model Selection

A few models were decided to be experimented upon. These were linear regression, gradient boosting regression, and decision tree regression. The dataset has variables that involve condition groups, the state, month, age group, and COVID-19 death. The "Age Group" column's presence of Nan values was solved by dropping off the corresponding rows. The categorical variables in the 'State' and 'Condition Group' columns were replaced by label encoding values i.e., they were transformed into numerical representations. The 'Age Group' data was then processed by one-hot encoding to enable the machine learning algorithms to function with categorical formats.

3.3.2 Model Training and Evaluation

Linear regression

The data set was split into training and test sets with the training set being 80% and the testing set being 20%. A Linear Regression model was initialized and trained using the variables in the training data set. The trained model was subsequently evaluated using the testing set to obtain a metric of its forecasting accuracy.

Mean Squared Error (MSE): 44096.94069065316

Mean Absolute Error (MAE): 32.9264676540234

R-squared (R2) Score: 0.007844049423549881

Cross-Validation

Cross-validation was also carried out on the training data to finetune the model output. In order to determine the accuracy of the mean squared error, the 5-fold cross-validation method was

implemented. Thus, the RMSE that was on average calculated across the training subsets was determined to be 231.88 which represented the average error of the model.

Residual Analysis

The model's residual plot was generated using predictions on the training set which helped in visualizing the accuracy of the model. The plot depicts the relationship between the deaths of COVID-19 as predicted against the observed value.

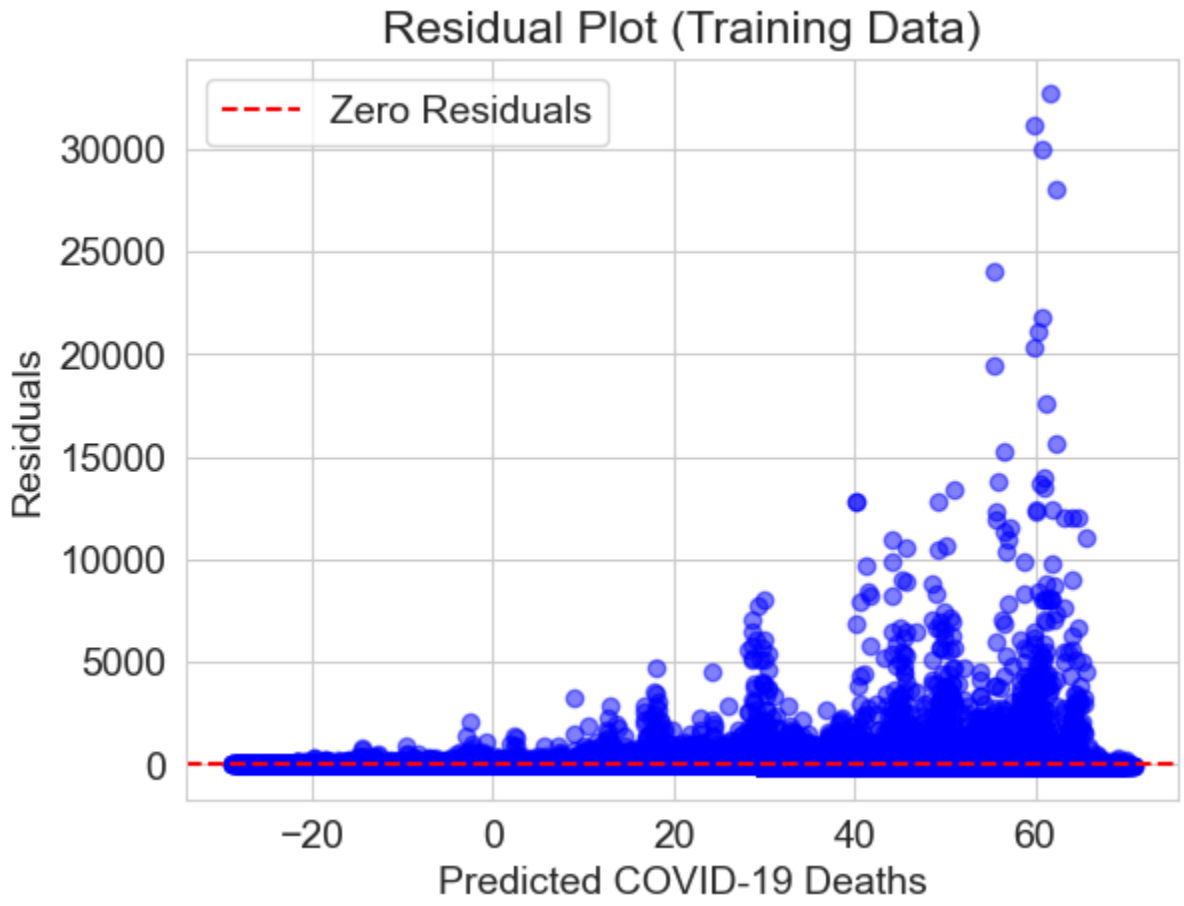


Figure 12: Residual plot for Linear Regression

Decision-tree regression

A similar preprocessing approach was employed in the decision-tree regression model. Missing values and numerical and categorical features were handled appropriately before training. These

were the results of the evaluation.

Mean Squared Error (MSE): 42602.05519319567

Mean Absolute Error (MAE): 19.690103929307913

R-squared (R2) Score: 0.04147811832954229

Gradient Boosting Regressor

A similar preprocessing approach was employed in the decision-tree regression model. Missing values and numerical and categorical features were handled appropriately before training.

Mean Squared Error (MSE): 34287.10890679057

Mean Absolute Error (MAE): 26.605539123882846

R-squared (R2) Score: 0.22855965522466304

4. Findings

4.1 Exploratory analysis

- The mean number of COVID-19 deaths was 30.72 with a standard deviation of 483.37 per observation in the dataset.
- COVID-19 deaths range from 0 to 105,566 per observation.
- The histograms help visualize the distribution of COVID-19 deaths across different age groups.
- In the age group of elderly people, death rates are higher than in young people.
- Age groups 0-24 registered the lowest fatalities while ages 85 and above had the highest fatalities, with about 250 recorded deaths.

- The number of fatalities and mentions has decreased since early 2022.
- The peak of COVID-19 deaths was in 2021, with the lowest recorded deaths reported before 2020.
- There is a spike in covid deaths in 2022 followed by a steep decline to where the trend stabilizes from late 2022 to late 2023.

- From the line plot, we can see that COVID-19, followed by respiratory arrest and respiratory failure have been the most common ailments over the past years.

4.2 K means clustering

Low-Risk Cluster (Cluster 0):

- The breakdown of age groups for this category includes 0-24, 25-34, and 35-44.
- This age group has the lower mortality rate from COVID-19 disease among other groups.
- COVID-19 deaths among those infected account for 17031 in this cluster.

Medium-Risk Cluster (Cluster 1):

- Represents the mortality risk experienced by people of all ages in general.

High-Risk Cluster (Cluster 2):

- Includes ages; 65-74, 75-84, and 85+.
- These age bands are the most likely to succumb to the disease.

Low-Risk Conditions (Cluster 0):

- Diseases like diabetes and hypertensive diseases are in this category.

Moderate-Risk Conditions (Cluster 1):

- Illnesses such as flu and pneumonia fall under this cluster.

Moderate to High-Risk Conditions (Cluster 2):

- Respiratory failure and respiratory arrest are in this cluster.

High-Risk Conditions (Cluster 3 and Cluster 4):

- Alzheimer's disease, obesity, and chronic lower respiratory diseases fall under this category.

Clustering by age group and condition

Cluster	***COVID-19 Deaths***	***Top 3 Most Frequent Age Groups***	***Top 3 Most Frequent Condition Groups***
0	*56726.095238*	*[85+, 75-84, 65-74]]*	*[Diabetes, Renal failure, Sepsis]*
1	*591004.833333*	*[65-74, 75-84, 85+]*	*[COVID-19, Respiratory diseases]*
2	*187550.1428*	*[55-64, 45-54, 65-*	*[All other conditions and causes (residual),*

	57	*74]*	*COVID-19, Circulatory diseases]*
3	*386094.8*	*[55-64, 65-74, 75-84]*	*[Circulatory diseases, COVID-19, Respiratory diseases]*
4	*6947.684211*	*[0-24, 25-34, 35-44]*	*[Intentional and unintentional injury, poisoning, and other adverse events, Obesity, Alzheimer disease]*

Table 6:: Table showing the top 3 age groups and condition groups in each cluster

4.3 Predictive modeling

The predictive modeling analysis involved exploring three distinct regression models: simple linear regression, decision tree regression, and gradient boosting regression.

Linear regression

Mean Squared Error (MSE): 44096.94069065316

Mean Absolute Error (MAE): 32.9264676540234

R-squared (R2) Score: 0.007844049423549881

Decision-tree regression

Mean Squared Error (MSE): 42602.05519319567

Mean Absolute Error (MAE): 19.690103929307913

R-squared (R2) Score: 0.04147811832954229

Gradient Boosting Regressor

Mean Squared Error (MSE): 34287.10890679057

Mean Absolute Error (MAE): 26.605539123882846

R-squared (R2) Score: 0.22855965522466304

5. Discussion

From the results and extrapolated findings disparities are clear in COVID-19 mortality rates. Methods like clustering and visualizing different plots proved useful in extrapolating findings and insights from our data. The results suggest that age and underlying conditions are the major determining factors when it comes to COVID-19 being fatal. Given that older populations are more likely to succumb to the COVID-19 pandemic, mitigation measures should be targeted toward the elderly. Special attention and care should be given to the elderly during the pandemic especially if admitted to hospital. Underlying health conditions are also a major factor in determining COVID-19 mortality rates. Our results show that the mortality rate increases dramatically when an individual has preexisting health conditions especially if they are lung-related. Individuals with respiratory failure, lower respiratory diseases, and respiratory arrest should be given special medical attention during the pandemic. Alzheimer's disease was found to be highly correlated with COVID-19 deaths, but that could be because Alzheimer's disease is a disease frequently found in the elderly, who also have the highest mortality rates.

Temporal trends showed that there was a significant decline in COVID-19 mortalities, which later stabilized in 2023. This means that preventative and control measures had a positive impact. Measures like social distancing, the introduction of vaccines, and hand sanitization had a positive effect on controlling the virus. According to our analysis, California and Texas were the top states with the highest coronavirus death rates. The COVID-19 pandemic brought with it a heavy burden to human life. For California it could be explained by the fact that it is mostly comprised of densely populated urban centers with large tourist spots. For Texas, it is because it is a large state with

varying demographics.

K-means clustering which is an unsupervised machine language was useful in clustering according to age and comorbidities. It works by iteratively assigning data points to clusters according to similar characteristics. It offered in-depth insights on how they are related to the number of deaths. The predictive models, namely linear regression, gradient boosting regression, and decision trees, offered insights on how machine learning models can be used to predict factors that influence COVID-19 mortality rates. The decision trees and the gradient-boosting regressor outperformed the linear regression model. This was determined using evaluation metrics. Mean Absolute Error measures the difference between the actual and predicted values. Mean Squared Error measures the squared differences between the prediction and actual values. The R-squared score measures the variance the model can be able to fit with 0 meaning no variation is accounted for, and 1 meaning 100% of the dataset's variation has fit.

The nature of the COVID-19 pandemic can be seen in the model evaluation outcomes. Its high unpredictability due to new variants and varying peaks means that it can be difficult for models to capture such complex relationships. Even the best-performing model (gradient booster regressor) in terms of capturing variance only captured 0.2(20%) of the variance. This presents as one of the significant limitations of this study. More research and model refinement will be needed to get better predictive models.

This research highlights the necessity of exploring machine learning models and data analytics to gain insights that will inform policymakers and healthcare practitioners on the best ways to address the pandemic.

6. Recommendations

1. Targeted mitigation strategies for elderly people.
2. Enhanced monitoring and care for individuals with underlying health conditions
3. Continued vigilance and adherence to preventative measures
4. Integration of machine learning models in public health strategies:
5. Investment in research and model refinement.
6. Community engagement and sensitization.

7. Resilience building and preparedness.

6. Conclusion

Our study has shown how the complex relationship between demographic conditions, underlying health conditions, and COVID-19 mortality rate are connected. The use of clustering algorithms and predictive modeling gave way to revealing important and variable patterns and trends in the data. Moreover, the undeniable decrease in death rates underscores the impact of preventative and control strategies. While viral strain mutation continues to pose a great risk, it requires constant alertness and adaptation to stay on our toes regarding response efforts. With the complex and ever-morphing nature of the pandemic, the combination of using modern analytical tools and collaborations across different disciplines will be crucial in helping evidence-based decision-making and improving the resilience of our society against future pandemics.

7. References

[1] Zhang, Q., Gao, J., Wu, J. T., Cao, Z., & Zeng, D. (2021a). Data science approaches to confronting the COVID-19 pandemic: a narrative review. Philosophical Transactions of the Royal Society A, 380(2214). https://doi.org/10.1098/rsta.2021.0127

[2] Sharma, A., Tiwari, S., Deb, M. K., & Marty, J. L. (2020). Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2): a global pandemic and treatment strategies. International Journal of Antimicrobial Agents, 56(2), 106054. https://doi.org/10.1016/j.ijantimicag.2020.106054

[3] Chavda, V. P., Bezbaruah, R., Deka, K., Nongrang, L., & Kalita, T. (2022). The delta and Omicron variants of SARS-COV-2: What we know so far. Vaccines, 10(11), 1926. https://doi.org/10.3390/vaccines10111926

[4] Chavda, V. P., Bezbaruah, R., Deka, K., Nongrang, L., & Kalita, T. (2022). The delta and Omicron variants of SARS-COV-2: What we know so far. Vaccines, 10(11), 1926. https://doi.org/10.3390/vaccines10111926

[5] Magableh, G. M. (2021). Supply Chains and the COVID-19 Pandemic: A Comprehensive framework. European Management Review, 18(3), 363–382. https://doi.org/10.1111/emre.12449

[6] Magableh, G. M. (2021). Supply Chains and the COVID-19 Pandemic: A Comprehensive framework. European Management Review, 18(3), 363–382. https://doi.org/10.1111/emre.12449

[7] The Great Lockdown: worst economic downturn since the Great Depression. (2020, April 14). IMF. https://www.imf.org/en/Blogs/Articles/2020/04/14/blog-weo-the-great-lockdown-worst-economic-downturn-since-the-great-depression

[8] ONE. (2024, January 11). ONE Africa COVID-19 Tracker - ONE Data & Analysis. ONE Data & Analysis. https://data.one.org/data-dives/covid-19-response/

[9] Wilkialis, L., Rodrigues, N. B., Danielle, S., Siegel, A., Majeed, A., Lui, L. M., Tamura, J. K., Gill, B., Teopiz, K. M., & McIntyre, R. S. (2021). Social Isolation, Loneliness, and Generalized Anxiety: Implications and Associations during the COVID-19 Quarantine. Brain Sciences, 11(12), 1620. https://doi.org/10.3390/brainsci11121620

[10] Bayati, M., Noroozi, R., Ghanbari-Jahromi, M., & Jalali, F. S. (2022). Inequality in the distribution of COVID-19 vaccine: a systematic review. International Journal for Equity in Health, 21(1). https://doi.org/10.1186/s12939-022-01729-x

[11] Mueller, A. L., McNamara, M. S., & Sinclair, D. (2020). Why does COVID-19 disproportionately affect older people? Aging, 12(10), 9959–9981. https://doi.org/10.18632/aging.103344

[12] Mérad, M., & Martin, J. C. (2020). Pathological inflammation in patients with COVID-19: a key role for monocytes and macrophages. Nature Reviews Immunology, 20(6), 355–362. https://doi.org/10.1038/s41577-020-0331-4

[13] Zhang, Q., Gao, J., Wu, J. T., Cao, Z., & Zeng, D. (2021b).

[14] Data science approaches to confronting the COVID-19 pandemic: a narrative review. Philosophical Transactions of the Royal Society A, 380(2214). https://doi.org/10.1098/rsta.2021.0127

[15] Nopour, R., Shanbehzadeh, M., & Kazemi-Arpanahi, H. (2022). Using logistic regression to develop a diagnostic model for COVID-19: A single-center study. Journal of Education and Health Promotion, 11(1), 153. https://doi.org/10.4103/jehp.jehp_1017_21