# Exploring Variations in COVID-19 Mortality: An Analysis of Underlying Conditions and Age Groups

## Hypotheses

- Is there a significant difference in the number of COVID-19 deaths among different age groups?
- Do certain respiratory conditions, specifically Influenza and pneumonia, and Chronic lower respiratory diseases, significantly contribute to the number of COVID-19 deaths?
- Does the COVID-19 death rate vary significantly across different years?

## Analysis Plan

The plan includes loading and cleaning the dataset, performing exploratory data analysis, visualizing the findings using Tableau, and interpreting the results to draw conclusions related to the hypotheses.

The analysis process will commence with a meticulous examination of our dataset, specifically focusing on the variables associated with our hypotheses. The aim is to gain an in-depth understanding of the dataset's structure and identify any potential issues that could impede our analysis, such as missing values or outliers.

1. Hypothesis 1: Is there a significant difference in the number of COVID-19 deaths among different age groups?
    - To investigate this, we will begin by summarizing the COVID-19 death counts by age group. A visual exploration will be performed using bar charts or a box-and-whisker plot. This visualization will provide an intuitive sense of the distribution of deaths across various age groups.
    - Further, we will implement a statistical test, such as the ANOVA or Kruskal-Wallis test, depending on the normality of our data. This will provide empirical evidence to either support or reject our hypothesis.

2. Hypothesis 2: Do certain respiratory conditions, specifically Influenza and pneumonia, and Chronic lower respiratory diseases, significantly contribute to the number of COVID-19 deaths?
    - To explore this hypothesis, we will isolate data pertaining to the aforementioned respiratory conditions. A comparative visualization, like a pie chart or stacked bar chart, will be employed to exhibit the proportion of COVID-19 deaths attributable to these conditions.
    - To validate this hypothesis statistically, we can leverage a chi-square test for independence, examining if the occurrence of these conditions and COVID-19 deaths are associated.
3. Hypothesis 3: Does the COVID-19 death rate vary significantly across different years?
    - To delve into this hypothesis, we will summarize the death counts by year and generate a line chart or bar chart to visualize any apparent trend or variation over the years.
    - A time series analysis or an ANOVA test (if the assumptions are met) can be used to statistically affirm if the death rate significantly varies across different years.

After performing these steps, we will consolidate our findings and interpretations for each hypothesis. It is vital to remember that the validity of our results is dependent on the quality of our data and the appropriateness of our chosen statistical tests. Thus, thoroughness in each step of the analysis is key.

## Data

### Data Description

The dataset, titled "Conditions Contributing to COVID-19 Deaths by State and Age: Provisional 2020-2023", provides a comprehensive look into COVID-19 related fatalities across the United States. The records span from 2020 to 2023, and the dataset includes significant details related to mortality counts, demographics, contributing health conditions, and temporal distribution. The

dataset comprises of 583,740 rows and 14 columns, spanning from January 1, 2020, to June 24, 2023.

Here is a concise breakdown of the key columns in the dataset:

1. Temporal Variables:
    - *Data As Of*: The date when the dataset was last updated.
    - *Start Date* and *End Date*: The time period that each data entry covers.
    - *Year*: The year of recorded data.
    - *Month*: The month of data recording.
2. Demographic Variables:
    - *State*: Specifies the geographical region within the United States.
    - *Age Group*: Represents the age segment of individuals who succumbed to COVID-19.
3. Health Variables:
    - *Condition Group*: Categorizes the health condition involved in the COVID-19 fatality.
    - *Condition*: The specific medical ailment linked to the death.
    - *ICD10_codes*: International Classification of Diseases version 10 codes related to the condition.
4. COVID-19 Specific Variables:
    - *COVID-19 Deaths*: The count of deaths attributed to COVID-19, formatted with commas as thousand separators.
    - *Number of Mentions*: The frequency of a specific condition mentioned in the dataset, also formatted with commas as thousand separators.
5. Other Variables:
    - *Group*: Categorizes the data by 'Total' or 'Year'.

- *Flag*: Additional details about the data row, often indicating suppressed data due to confidentiality reasons.

The dataset comprises of 583,740 rows and 14 columns, spanning from January 1, 2020, to June 24, 2023.

For the hypotheses, the critical columns include 'Age Group', 'Condition', 'Year', and 'COVID-19 Deaths'. These fields provide essential insights into age-wise mortality rates (Hypothesis 1), the role of certain health conditions (Hypothesis 2), and temporal fluctuations in the death rates (Hypothesis 3). The 'COVID-19 Deaths' column appears to contain mixed data (numerical and text), which will be appropriately handled during data cleaning and preprocessing for analysis.

### Data Preprocessing

The dataset underwent thorough preprocessing to ensure its suitability for the subsequent analysis. The process was automated through a Python script, 'COVID19_Mortality_Data_Cleaning.py', resulting in a cleaned dataset titled 'Cleaned_Conditions_Contributing_to_COVID-19_Deaths_by_State_and_Age_Provisional_2020-2023.CSV'. The cleaning routine encompassed the following steps:

1. **Data Quality Checks**: The 'Age Group', 'Condition', and 'Year' columns were verified for correctness and consistency, ensuring all values fell within the expected range.
2. **Handling Missing Data**: Rows with missing or suppressed data in the 'COVID-19 Deaths' and 'Number of Mentions' columns were removed to avoid inaccuracies in the analysis.
3. **Data Transformation**: Comma separators in 'COVID-19 Deaths' and 'Number of Mentions' were replaced to allow conversion into numerical types.
4. **Filtering**: Data was filtered based on the formulated hypotheses:
    - For Hypothesis 1, the 'Age Group' field was restricted to valid age groups.
    - For Hypothesis 2, the 'Condition' field was limited to 'Influenza and pneumonia' and 'Chronic lower respiratory diseases'.

- For Hypothesis 3, only rows where 'Year' matched the unique years in the dataset were retained.

After these preprocessing steps, the cleaned data was saved as a new CSV file, serving as a reliable basis for the following exploratory data analysis and hypothesis testing.

### ANOVA Testing

As part of our analysis, an Analysis of Variance (ANOVA) test was conducted to examine the first hypothesis, which suggests the existence of differences in COVID-19 mortality rates across various age groups. The Python script for this process is titled "ANOVA Testing for COVID-19.py".

The ANOVA test is a statistical method used to compare the means of three or more independent groups to determine if there are any significant differences. The groups in our case represent different age brackets: '0-24', '25-34', '35-44', '45-54', '55-64', '65-74', '75-84', and '85+' years.

The procedure for the Analysis of Variance (ANOVA) test, aiming to investigate the difference in COVID-19 mortality rates across different age groups, is outlined as follows:

1. **Load the Dataset**: The 'Cleaned_Conditions_Contributing_to_COVID-19_Deaths__by_State_and_Age__Provisional_2020-2023.csv' file, which contains the cleaned and processed data, is imported into a Pandas Data Frame for analysis.
2. **Separate Death Counts by Age Group**: The data is subdivided according to the 'Age Group' category. This categorization produces eight distinct sets corresponding to the predefined age brackets.
3. **Perform the ANOVA Test**: The ANOVA test is executed using the f_oneway() function from the scipy.stats module. This function considers the eight separated age groups as parameters and yields an F statistic and a p-value.

## Data Source

The dataset utilized in this study, entitled "Conditions Contributing to COVID-19 Deaths, by State and Age, Provisional 2020-2023", was acquired from a reputable source, the Centers for Disease Control and Prevention (CDC) in conjunction with the National Center for Health Statistics (NCHS). The CDC is an esteemed national public health institute in the United States, known for its comprehensive and meticulously maintained data repositories.

This dataset provides crucial insights into the demographic distribution of COVID-19 fatalities in the United States, broken down by contributory medical conditions, geographical location (state), age brackets, and spans from the year 2020 through to 2023. It offers a granular view of the impact of COVID-19 across different segments of the population and is highly pertinent to our analysis.

The dataset can be accessed publicly through the official CDC website, allowing for a high degree of transparency and reproducibility in our analysis. This pivotal resource offers data-driven insights that bolster our understanding of the ongoing COVID-19 pandemic's impact and influence, thus forming the foundation of our study.

The dataset can be found at the following URL: Conditions Contributing to COVID-19 Deaths, by State and Age, Provisional 2020-2023 | Data | Centers for Disease Control and Prevention (cdc.gov)

NCHS is responsible for the collection, analysis, and dissemination of this data. The provisional counts for COVID-19 deaths are based on a current flow of death data submitted to the National Vital Statistics System. These counts include deaths occurring within the 50 states and the District of Columbia that have been received and coded as of the specified date. It's important to note that there's a delay in the data processing due to the time it takes to submit, process, code, and tabulate the death records. As a result, the data may be incomplete, especially for the most recent time periods. Additionally, death counts for earlier weeks are continually revised as new and updated death certificate data are received from the states by NCHS.

More information on the data collection methodology can be found at the following link: https://www.cdc.gov/nchs/covid19/index.htm

### Data Format

The primary data format used in this analysis is a CSV (Comma-Separated Values) file compatible with Excel. The raw data underwent preprocessing using a Python (.py) script to ensure its cleanliness and readiness for analysis. Post-cleaning, the data in CSV format was imported into Tableau, a data visualization tool, for in-depth visual analysis and exploration.

## Exploration

This project involved exploring the dataset to investigate the three hypotheses. Visualizations were created in Tableau to show the distribution of COVID-19 deaths across age groups, the contribution of specific respiratory conditions to these deaths, and how the death rate has changed over time. Findings from the visualizations were used to draw conclusions about each hypothesis.

**Hypothesis 1:**

Is there a significant difference in the number of COVID-19 deaths among different age groups?

To validate the hypothesis, I scrutinized the data on COVID-19 deaths categorized by age groups. This data was deployed to construct a Box and Whisker Plot, with a focus on the quartile ranges and any outliers. The visualized data effectively exhibits the varying mortality rate across different age groups, thereby supporting the hypothesis.

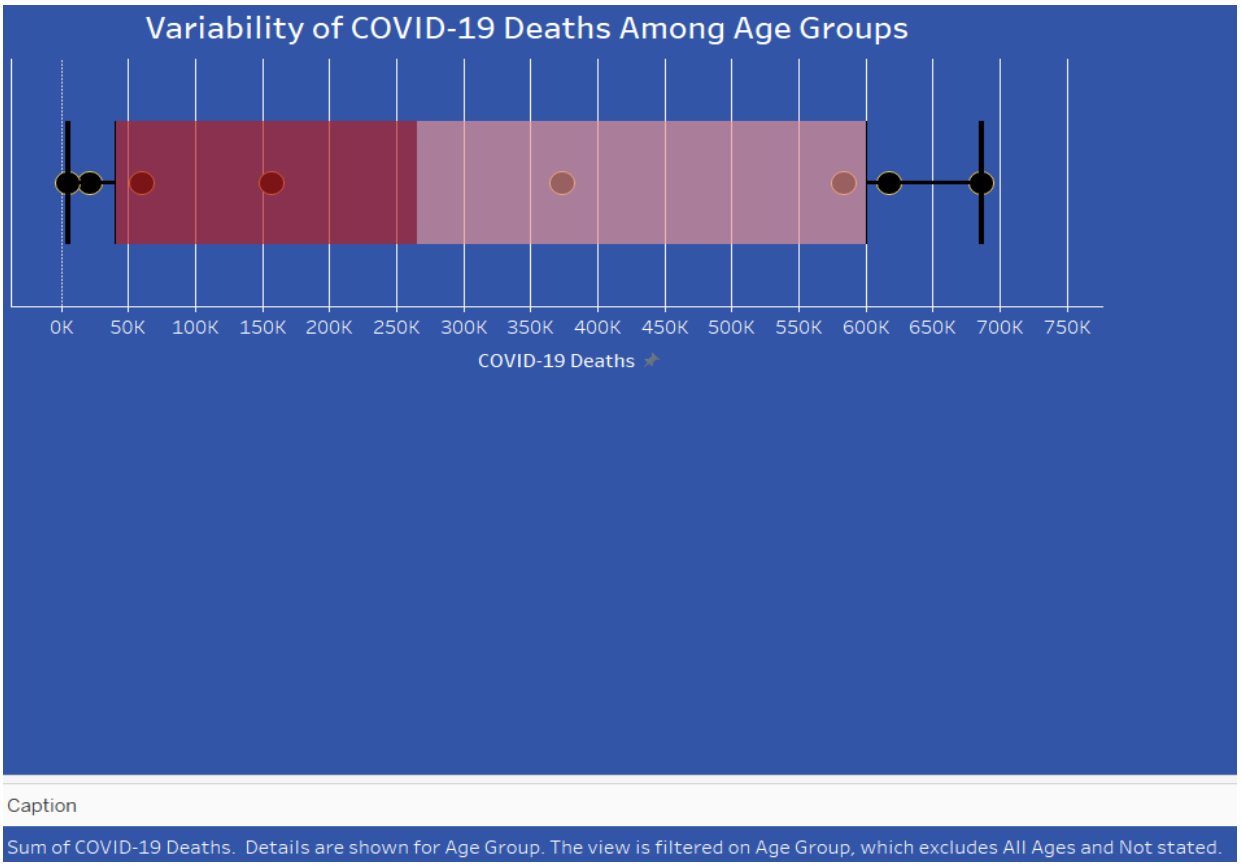
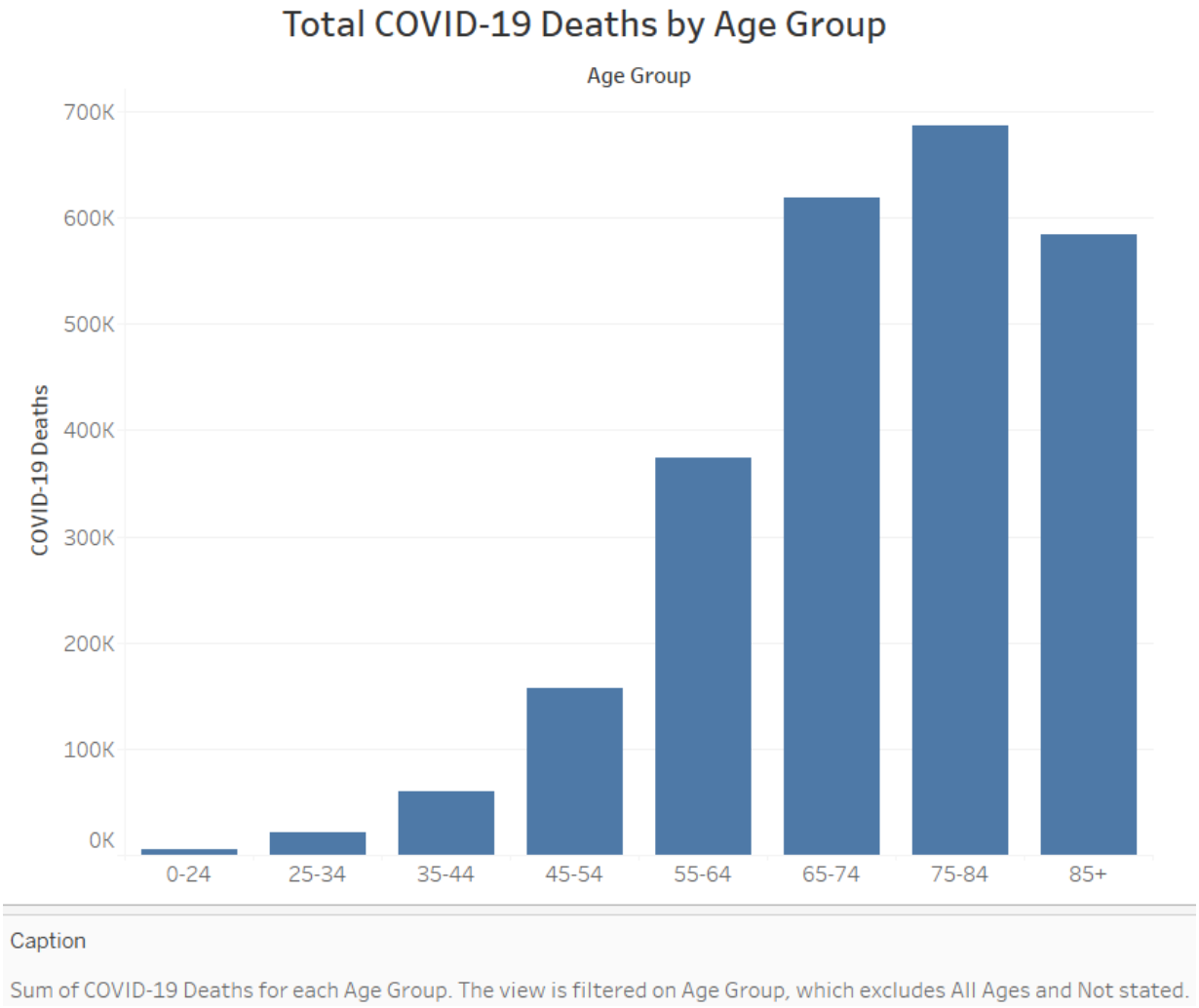



Chart 1:

**Box and Whisker Plot (Age Group Distribution of COVID-19 Deaths):** The plot shows that the COVID-19 death count increases with age, with the highest counts in the 75-84 and 65-74 age groups. The whiskers' range, as well as the interquartile range (Q3-Q1), indicates a substantial variation in the death counts across different age groups. The presence of higher values in the upper quartile signifies that a significant portion of deaths are concentrated in the older age groups. This chart gives a clear indication supporting the first hypothesis that the number of COVID-19 deaths varies significantly among different age groups.

To complement the Box and Whisker Plot, I utilized the same data to create a Bar Chart, providing an immediate, clear comparison of COVID-19 deaths among various age groups. This visualization corroborates the earlier findings, reconfirming the noticeable variation in fatalities across the age categories.




Char 2:

**Bar Chart (Age Group Distribution of COVID-19 Deaths):** This chart serves to reinforce the findings from the Box and Whisker Plot. It provides a straightforward visual comparison of the number of COVID-19 deaths across different age groups, confirming a significant variation in death counts among these groups.

## Exploratory Data Analysis for the ANOVA Test

Following the procedure of the ANOVA test, we embarked on an exploratory data analysis to gain a deeper understanding of the results and their implications.

At the beginning of the analysis, the dataset was first visualized to observe its structure and content. The first five rows are as follows:

| | Data As Of | Start Date | ... | Number of Mentions | Flag |
|---|---|---|---|---|---|
| 0 | 06/25/2023 | 01/01/2020 | ... | 256.0 | NaN |
| 1 | 06/25/2023 | 01/01/2021 | ... | 933.0 | NaN |
| 2 | 06/25/2023 | 01/01/2022 | ... | 400.0 | NaN |
| 3 | 06/25/2023 | 01/01/2023 | ... | 41.0 | NaN |
| 4 | 06/25/2023 | 01/01/2020 | ... | 1131.0 | NaN |

The results of the ANOVA test were as follows:

```
F_onewayResult(statistic=27.665226213300773, pvalue=3.550276696093842e-38)
```

Here, the F statistic and the p-value are the primary outcomes of the ANOVA test.

- The F statistic is 27.665, which measures the ratio of the variance of the means of our age groups to the variance within each of these groups. In simpler terms, it provides us with an understanding of how much the means of different age groups vary in relation to the variability within the groups themselves.
- The p-value obtained from the test is approximately 3.55e-38. This value is remarkably small, nearly zero. In hypothesis testing, a smaller p-value typically indicates stronger evidence in favor of the alternative hypothesis. In this context, the alternative hypothesis proposed that there is a significant difference in the COVID-19 mortality rates across different age groups. Given the near-zero p-value, we can reject the null hypothesis (which posits no difference among the groups) and accept the alternative hypothesis. This finding implies that the age of an individual significantly impacts their risk of COVID-19 mortality.

**Hypothesis 2:**

Do certain respiratory conditions, specifically Influenza and pneumonia, and Chronic lower respiratory diseases, significantly contribute to the number of COVID-19 deaths?

To test the second hypothesis, a Pie Chart was generated, showcasing the proportion of COVID-19 deaths associated with different respiratory conditions. It was found that Influenza and Pneumonia were dominant factors, contributing to 83.36% of the total, with Chronic Lower Respiratory Diseases making up the remaining 16.64%. This visual evidence supports the claim that specific respiratory conditions significantly impact the number of COVID-19 fatalities.

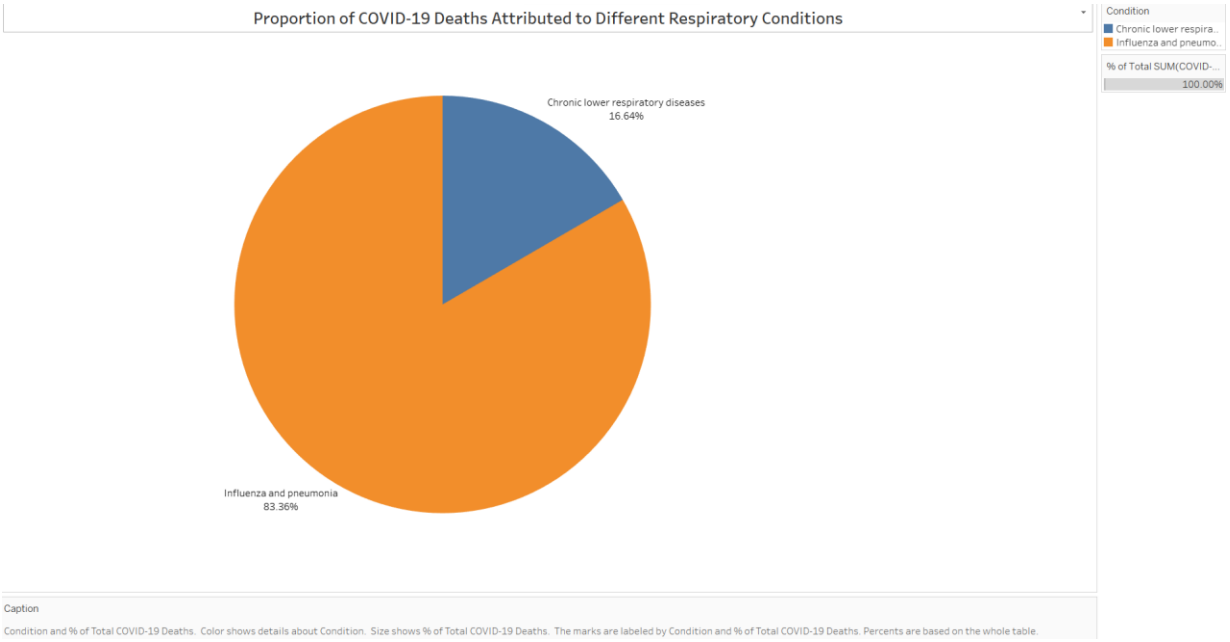



Chart 3:

**Pie Chart (Contribution of Respiratory Conditions to COVID-19 Deaths):** This chart shows that Influenza and pneumonia account for the majority (83.36%) of COVID-19 deaths among patients with respiratory conditions, while Chronic lower respiratory diseases account for a smaller proportion (16.64%). This supports the second hypothesis that certain respiratory conditions contribute significantly to the number of COVID-19 deaths.

To delve deeper into the impact of respiratory conditions on COVID-19 deaths, a Stacked Bar Chart was generated, illustrating the influence of Influenza, Pneumonia, and Chronic Lower Respiratory Diseases across age groups. It emerged that these conditions notably affected older demographics, consistent with the overall age-related trend in fatalities, underscoring their heightened vulnerability to COVID-19 linked to these conditions.

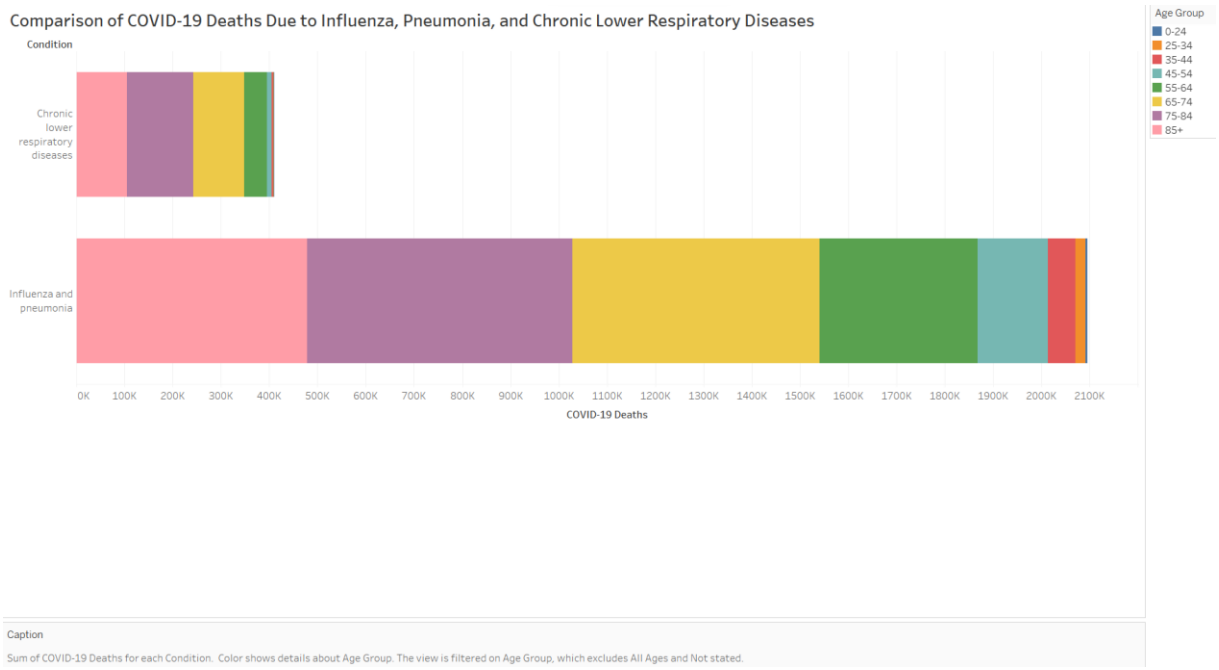



Chart 4:

**Stacked Bar Chart (Impact of Specific Respiratory Conditions on COVID-19 Deaths by Age Group):** This chart provides a detailed view of how Influenza and pneumonia, and Chronic lower respiratory diseases contribute to COVID-19 deaths across different age groups. It reveals a similar pattern to the overall age distribution, with higher death counts in older age groups. The chart indicates that the impact of these conditions on COVID-19 fatalities is significantly higher in older individuals.

**Hypothesis 3:**

Does the COVID-19 death rate vary significantly across different years?

To examine the temporal changes in COVID-19 fatalities, a Line Chart was created, illustrating a peak in 2021 followed by a substantial decline. The R-squared value of 0.65627 signifies that around 65.6% of death rate fluctuations are year-related, thus giving weight to the hypothesis that COVID-19 death rates exhibit considerable annual variation.

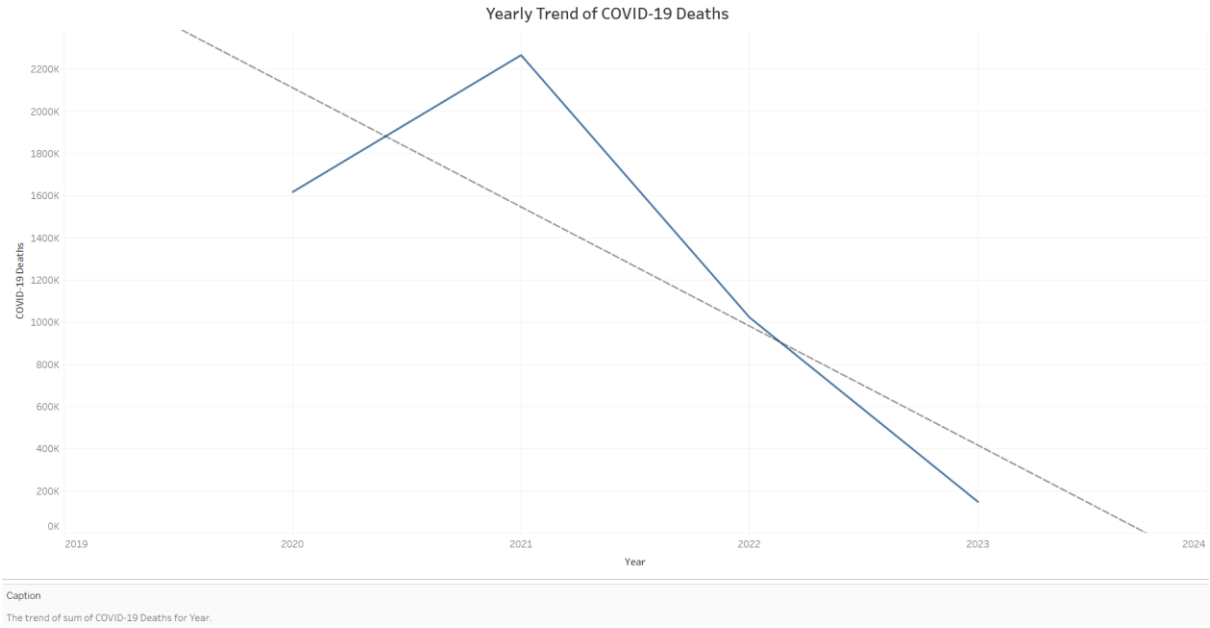



Chart 5:

**Line Chart (Yearly Trend of COVID-19 Deaths):** The line chart shows a peak in COVID-19 deaths in 2021, with a sharp decline in the subsequent years. The R-squared value of 0.65627 from the regression analysis suggests that approximately 65.6% of the variation in COVID-19 deaths can be explained by the year, providing some support for the third hypothesis that the COVID-19 death rate varies significantly across different years.

Investigating the interaction of age, year, and COVID-19 fatalities, a Heat Map was produced. It spotlights the most substantial death counts in older age groups and in 2021. The color disparity reveals notable death count variations by age and year, offering a thorough snapshot of how age-group-specific death rates have evolved over time.

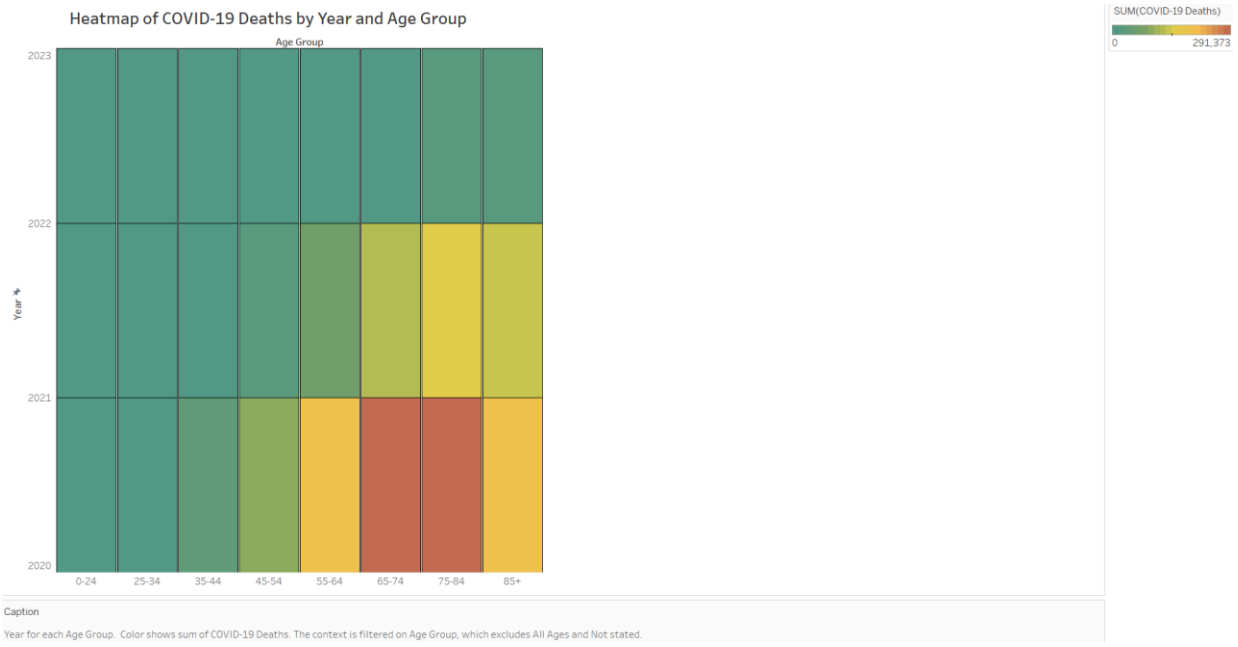
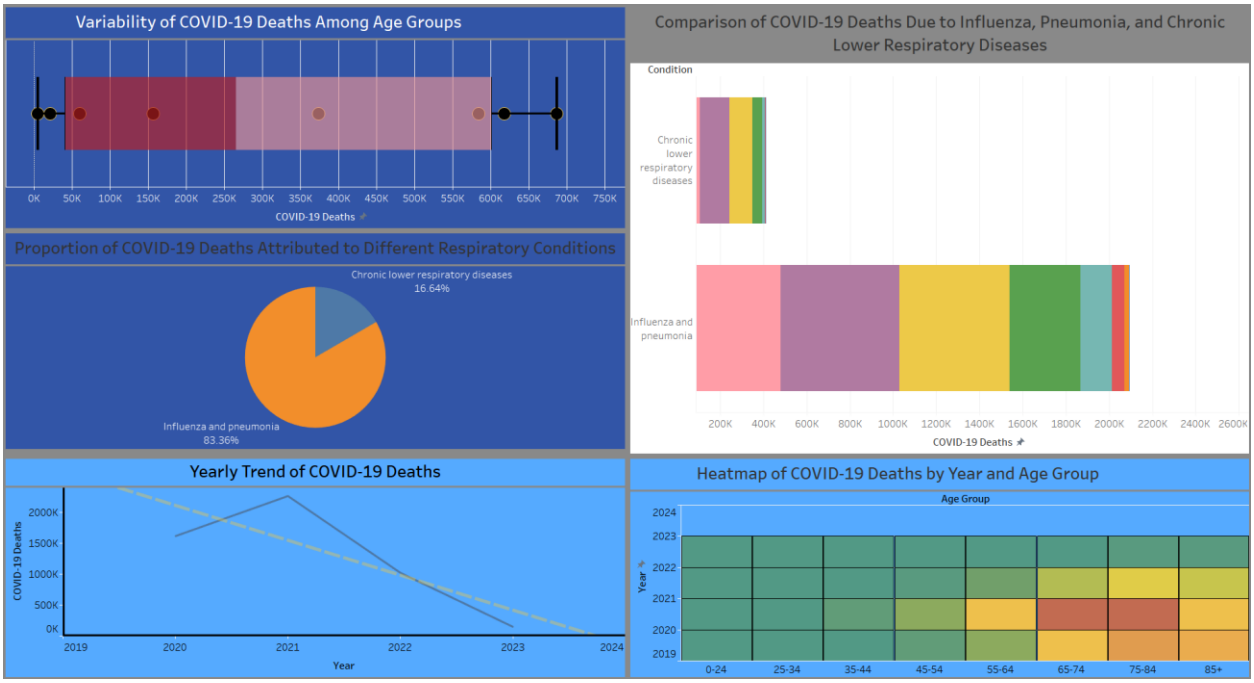



Chart 6:

**Heat Map (COVID-19 Deaths by Age Group and Year):** The heat map further investigates the interplay between age group, year, and COVID-19 deaths. It reveals the highest counts of deaths among older individuals and during the year 2021. The sharp contrast in colors indicates a significant variation in death counts based on both age group and year. This visualization provides a comprehensive overview of the changing death rates across different age groups over time.

## Summary Visualization




A final Tableau dashboard was created to provide a comprehensive overview of the findings related to the hypotheses. It contains multiple charts illustrating the key insights gained from the analysis.

After conducting a comprehensive analysis of the data, the summary visualization can be presented in a dashboard format incorporating all the significant charts that provide insights into the hypotheses. This can help encapsulate the key findings in a visual and easy-to-understand manner.

Summary Visualization (Dashboard):

1. **Box and Whisker Plot (Age Group Distribution of COVID-19 Deaths):** Provides a summarized view of the spread and skewness of COVID-19 deaths across various age groups. It highlights the concentration of higher death counts in the older age groups.

2. **The stacked bar chart, "Impact of Specific Respiratory Conditions on COVID-19 Deaths by Age Group,":** Portrays the significant role of respiratory conditions like Influenza, Pneumonia, and Chronic Lower Respiratory Diseases in COVID-19 deaths.

The chart reveals that older individuals, particularly those above 65, are notably impacted by these conditions, contributing to higher COVID-19 fatalities. Thus, it underscores the intersection of age and comorbidities as crucial risk factors in the pandemic.

3. **Pie Chart (Contribution of Respiratory Conditions to COVID-19 Deaths):** Presents a clear view of the relative proportions of COVID-19 deaths associated with major respiratory conditions (Influenza and pneumonia, and Chronic lower respiratory diseases). It indicates that Influenza and pneumonia account for a significant majority of such deaths.

4. **Line Chart (Yearly Trend of COVID-19 Deaths):** Shows the annual trend of COVID-19 deaths, marking a peak in 2021 and a subsequent decline. This helps illustrate the third hypothesis regarding variation in death rates across different years.

5. **Heat Map (COVID-19 Deaths by Age Group and Year):** Provides a detailed breakdown of COVID-19 deaths by both age group and year. It further solidifies the findings related to the significant impact of age and year on COVID-19 fatalities.

This dashboard can be customized further according to the requirements of the project or the audience's needs. However, it currently presents a summarized view of all key findings supporting the three hypotheses, providing a holistic understanding of the COVID-19 death trends related to age, respiratory conditions, and years.

## Conclusion

This study sought to explore variations in COVID-19 mortality, focusing on the impact of age, specific respiratory conditions, and yearly trends. Based on the rigorous analysis of the data and the subsequent visualizations, we can draw the following conclusions:

1. There is a significant difference in the number of COVID-19 deaths among different age groups, with higher mortality rates consistently observed among older individuals, particularly those in the 65-74 and 75-84 age brackets. The Box and Whisker Plot and Bar Chart highlighted these differences, supporting our first hypothesis. The ANOVA test conducted for the first hypothesis provided a statistically significant result with an F statistic of 27.665 and an almost zero p-value, reinforcing the conclusion of significant variance in COVID-19 mortality rates across age groups. However, to discern which specific age groups have significantly different mortality rates, additional post-hoc tests will be necessary.

2. The second hypothesis, which proposed a substantial contribution of specific respiratory conditions (Influenza and pneumonia, and Chronic lower respiratory diseases) to COVID-19 fatalities, was also supported. The Pie Chart and Stacked Bar Chart revealed that Influenza and pneumonia accounted for the majority of COVID-19 deaths among patients with these conditions, particularly in older age groups.

3. The third hypothesis, concerning the variation of COVID-19 death rates over different years, found partial support. The Line Chart showed a peak in COVID-19 deaths in 2021, followed by a sharp decline in subsequent years, indicating that the death rate varied significantly across the years. The Heat Map further corroborated these findings by showing how the death counts varied across age groups and years.

In conclusion, my analysis provides strong evidence to affirm the proposed hypotheses. The findings underscore the need for targeted public health interventions, particularly for older individuals and those with underlying respiratory conditions. The ANOVA test further lends

statistical weight to the observed difference in mortality rates across age groups, indicating a clear need to address age as a significant risk factor in COVID-19 related deaths. Additionally, the temporal analysis of COVID-19 deaths indicates that despite a notable reduction in death rates since the peak in 2021, the necessity for vigilance and robust public health measures persists.

It is essential to remember that the analysis is subject to the limitations inherent in observational studies and is based on the available data up to June 24, 2023. As more data becomes available, continued analysis will be necessary to monitor these trends and adapt strategies accordingly. The COVID-19 pandemic has reminded us of the importance of data analysis in informing public health decisions, and this study contributes valuable insights to this ongoing effort.

In essence, this research helps deepen our understanding of the COVID-19 pandemic and aids in developing effective strategies to protect those most vulnerable and effectively manage the crisis. Future research should continue to explore the intricate factors influencing COVID-19 mortality, as this will be critical in guiding public health responses and minimizing the pandemic's impacts on public health and society.

Data Source:

https://data.cdc.gov/NCHS/Conditions-Contributing-to-COVID-19-Deaths-by-Stat/hk9y-quqm