# Project (preliminary title)

# U.S. COVID-19 Deaths: 2020 – 2023

The goal of this project is to explore multiple variables and their relationship to COVID-19 deaths throughout the United States from 2020 to 2023. Variables include age groups, geographical location, and conditions contributing to COVID-19 deaths (comorbidities). We should be able to first examine COVID-19 mortality rates followed by any trends that exist between these variables to better understand how COVID-19 affects people in the US. Once an understanding of trends is established the project will provide a forecast of COVID-19 deaths for the following year, 2024. Recommendations for how to minimize COVID-19 deaths will be included as well.

Key project questions and hypotheses are described in the last section of this document.

There are two data sets that will be used in the project:

**Primary set:** *Provisional COVID-19 Deaths by Sex and Age, 2020 – 2023*

**Secondary set:** *Conditions Contributing to COVID-19 Deaths by State and Age, Provisional 2020 – 2023*

## 1. Primary set: *Provisional COVID-19 Deaths by Sex and Age, 2020 – 2023*

### 1.1 Data description for the data set *Provisional COVID-19 Deaths by Sex and Age, 2020-2023*

| | |
|---|---|
| **Data Source** | Data was obtained from data.gov, a reliable source that handles government data. The data was published by the Centers for Disease Control and Prevention (CDC) and was originally collected by the National Center for Health Statistics (NCHS) National Vital Statistics System. The dataset is intended for public access and use.<br>Downloaded: 7/12/23<br>https://catalog.data.gov/dataset/provisional-covid-19-death-counts-by-sex-age-and-state |
| **Data Collection** | Mortality data such as this is collected by the National Vital Statistics System. The provisional counts include deaths that have |

| | |
|---|---|
| | been received and coded as of the date specified. It is important to note that it can take several weeks for death records to be submitted to National Center for Health Statistics (NCHS), processed, coded, and tabulated. Therefore, data downloaded may be incomplete, likely containing an underestimate of deaths that occurred during a given time period. Death counts for earlier weeks are continually revised as updated death certificate data is received by NCHS. Data is currently lagging by an average of 1-2 weeks.<br>https://www.cdc.gov/nchs/covid19/covid-19-mortality-data-files.htm |
| **Contents** | The data set includes deaths involving COVID-19, pneumonia, and influenza reported to NCHS by sex, age group, and jurisdiction of occurrence from 2020 to 2023.<br><br>The data set consists of 16 columns:<br>• **Data As Of:** date that data was last updated.<br>• **Start Date:** the beginning date of the time period for which the data is being reported or collected.<br>• **End Date:** the ending date of the time period for which the data is being reported or collected.<br>• **Group:** how the data is grouped, as in By Total, By Year, or By Month.<br>• **Year:** the year of the death total being reported.<br>• **Month:** the month of the death total being reported.<br>• **State:** the US state where the death total occurred.<br>• **Sex:** male/female/all sexes categories<br>• **Age Group:** the age group of the death totals being reported. All Ages, Under 1 year, 0-17 years, 1-4 years, 5-14 years, 15-24 years, 18-29 years, 25-34 years, 30-39 years, 35-44 years, 40- |

| | |
|---|---|
| | 49 years, 45-54 years, 50-64 years, 55-64 years, 65-74 years, 75-84 years, 85 years and over.<br>- **COVID-19 Deaths:** total deaths due directly to COVID-19.<br>- **Total Deaths:** total deaths altogether, no specific cause.<br>- **Pneumonia Deaths:** total deaths due directly to pneumonia.<br>- **Pneumonia and COVID-19 Deaths:** total deaths due directly to simultaneous pneumonia and COVID-19.<br>- **Influenza Deaths:** total deaths due directly to influenza.<br>- **Pneumonia, Influenza, or COVID-19 Deaths:** total deaths due to pneumonia, influenza, or COVID-19. This total can be obtained by subtracting **Pneumonia and COVID-19 Deaths** from **Pneumonia Deaths** and adding that result to the **COVID-19 Deaths** and the **Influenza Deaths**.<br>- **Footnote:** a note section, primarily used to document when “one or more data cells have counts between 1-9 and have been suppressed in accordance with NCHS confidentiality standards.” |
| **Data relevance** | - Provides data on mortality rates due to COVID-19 and related illnesses.<br>- Provides demographic data in relation to these mortality rates: geographic location, sex, age.<br>- Data is recent and being updated weekly.<br>- Collected by NCHS and distributed by CDC. |

### 1.2 Data Understanding

| Column | Qualitative/Quantitative | Discrete/Continuous | Nominal/Ordinal/Binary |
|---|---|---|---|
| **Data As Of** | Qualitative | | Nominal |
| **Start Date** | Qualitative | | Nominal |
| **End Date** | Qualitative | | Nominal |
| **Group** | Qualitative | | Nominal |
| **Year** | Quantitative | Discrete | Ordinal |
| **Month** | Quantitative | Discrete | Ordinal |
| **State** | Qualitative | | Nominal |
| **Sex** | Qualitative | | Nominal |
| **Age Group** | Qualitative | | Ordinal |
| **COVID-19 Deaths** | Quantitative | Discrete | |
| **Total Deaths** | Quantitative | Discrete | |
| **Pneumonia Deaths** | Quantitative | Discrete | |
| **Pneumonia and COVID-19 Deaths** | Quantitative | Discrete | |
| **Influenza Deaths** | Quantitative | Discrete | |
| **Pneumonia, Influenza, or COVID-19 Deaths** | Quantitative | Discrete | |
| **Footnote** | Qualitative | | Nominal |

### 1.3 Data cleaning and wrangling

**Cleaning (Integrity/Quality checks):**

| Problem | Description | Solution |
|---|---|---|
| Missing values | The 2754 null values in the 'Year' column are because it is not necessary to record the year for the "By Total" entries in this data frame since this refers to all three years 2020-2023. The 13770 null values in the 'Month' column are for the same reason - there are "By Total" and "By Year" entries included in this data frame that make the 'Month' column irrelevant. All of | I replaced these null values with random integers 1 – 9 because the data is already provisional and is constantly updated every week. We know that entries with this footnote had at least 1 death but are not disclosing this information due to privacy. In |

| | | |
|---|---|---|
| | the null values for 'COVID-19 Deaths', 'Total Deaths', 'Pneumonia Deaths', 'Pneumonia and COVID-19 Deaths', 'Influenza Deaths', 'Pneumonia, Influenza, or COVID-19 Deaths' are due to the following reason noted in the 'Footnote' column: "One or more data cells have counts between 1-9 and have been suppressed in accordance with NCHS confidentiality standards." The Footnote column has null values where there were no necessary comments such as this. | order to make our data as useful as possible I chose to input these random integers. I do not believe that this will cause us to have unrealistic data that is not worth studying. As the data is actively being updated, death counts in the data set may be lower than reality anyway. |
| Duplicates | N/A | |
| Outliers | There are 680 entries of total 38743 that are considered outliers for COVID-19 Death Counts. | We will leave these outliers alone since this data is sourced and collected reliably and does not indicate any error in the data. During a global pandemic I expect there to be quite a bit of variability in death counts. |

**Wrangling:**

| | |
|---|---|
| Deleted columns | Removed the following columns for sake of relevance: ‘Data As Of’, ‘Start Date’, ‘End Date’.<br>Removed ‘Group’ column after removing distinction between ‘By Total’, ‘By Year’ and ‘By Month’ because we are only interested in ‘By Month’ information and now that column is irrelevant.<br>Removed ‘Sex’ column after removing gender distinction because it is no longer needed. |
| Replaced Null Values | I think the missing values in the Death columns can be replaced with a random integer 1-9 because we know that at least one death occurred and can't be reported because of confidentiality and privacy issues, and based on how they update this data every week the newer counts in particular are likely an underestimate of the total deaths as they catch up with incoming death |

| | |
|---|---|
| | certificate data, so I do not believe that the data will be skewed irreparably by doing this. It is worth noting that null values make up 28%, 14%, 32%, 26%, 20%, and 32% of values in the 'COVID-19 Deaths', 'Total Deaths', 'Pneumonia Deaths', 'Pneumonia and COVID-19 Deaths', 'Influenza Deaths', 'Pneumonia, Influenza, or COVID-19 Deaths' respectively. |
| Remove rows | The ‘Group’ column has ‘By Total’, ‘By Year’ and ‘By Month’ classifications. We want to analyze this data by month and thus don’t need the other classifications as it is irrelevant information for this analysis. I removed all entries with ‘By Total’ and ‘By Year’ classifications.<br>The ’State’ column also has irrelevant information by giving us totals for the entire United States. Removed all rows with ‘United States’ classification so we can focus on a geographical analysis of COVID-19 deaths in the US.<br>The ’Sex’ column also has irrelevant information for the purpose of this analysis. We are not analyzing COVID-19 as it relates to gender. This will also help merge data with our other dataset that does not contain this distinction. |

**Limitations:** As previously mentioned, the more recent data is provisional and the data overall is updated weekly, which means there is some allowable variation in death counts as the data is updated. We are limited by how behind NCHS is in collecting and coding the data, about 1-2 weeks. There were also many missing values that necessitated the use of value replacement.

**Ethical concerns**: I do not see any issues because NCHS and the CDC have already accounted for personal identification factors by not publicly distributing information on death counts of 9 or less.

## 2. **Secondary set: *Conditions Contributing to COVID-19 Deaths by State and Age, Provisional 2020 – 2023***

### **2.1 Data Description for the data set *Conditions Contributing to COVID-19 Deaths by State and Age, Provisional 2020 – 2023***

| | |
|---|---|
| Data source | Data was obtained from data.gov, a reliable source that handles government data. The data was published by the Centers for Disease Control and Prevention (CDC) and was originally collected by the National Center for Health Statistics (NCHS) National Vital Statistics System. The dataset is intended for public access and use.<br>Downloaded: 7/12/23<br>https://catalog.data.gov/dataset/conditions-contributing-to-deaths-involving-coronavirus-disease-2019-covid-19-by-age-group |
| Data collection | Mortality data such as this is collected by the National Vital Statistics System. The provisional counts include deaths that have been received and coded as of the date specified. It is important to note that it can take several weeks for death records to be submitted to National Center for Health Statistics (NCHS), processed, coded, and tabulated. Therefore, data downloaded may be incomplete, likely containing an underestimate of deaths that occurred during a given time period. Death counts for earlier weeks are continually revised as updated death certificate data is received by NCHS. Data is currently lagging by an average of 1-2 weeks.<br>https://www.cdc.gov/nchs/covid19/covid-19-mortality-data-files.htm |
| Contents | This dataset shows health conditions and contributing causes mentioned in conjunction with deaths involving coronavirus disease 2019 (COVID-19) by age group and jurisdiction of occurrence. 2022 and 2023 |

| | |
|---|---|
| | data are provisional. Estimates for 2020 and 2021 are based on final data.<br><br>The data set consists of 14 columns:<br>• **Data As Of:** date that data was last updated.<br>• **Start Date:** the beginning date of the time period for which the data is being reported or collected.<br>• **End Date:** the ending date of the time period for which the data is being reported or collected.<br>• **Group:** how the data is grouped, as in By Total, By Year, or By Month.<br>• **Year:** the year of the death total being reported.<br>• **Month:** the month of the death total being reported.<br>• **State:** the US state where the death total occurred.<br>• **Condition Group:** the more general group of conditions related to the COVID-19 death count.<br>• **Condition:** the specific condition within the condition group related to the COVID-19 death count.<br>• **ICD10_codes:** the International Classification of Diseases code that corresponds to the condition.<br>• **Age Group:** the age group of the death totals being reported. All Ages, 0-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75-84, 85+.<br>• **COVID-19 Deaths:** total deaths due directly to COVID-19.<br>• **Number of Mentions:** not explicitly clear what information is provided here. I’m inferring that this column includes not only COVID-19 deaths but also mentions of other cases related to the specific condition. May not be relevant for our analysis, especially since there is no |

| | |
|---|---|
| | information given by the source as to what this column documents.<br>- **Flag:** a note section, primarily used to document when "one or more data cells have counts between 1-9 and have been suppressed in accordance with NCHS confidentiality standards." |
| Data relevance | - Provides data on mortality rates due to COVID-19 and related illnesses.<br>○ More specific information on conditions mentioned as contributing causes of deaths involving COVID-19.<br>- Provides demographic data in relation to these mortality rates: geographic location, age.<br>- Data is recent and being updated monthly.<br>- Collected by NCHS and distributed by CDC. |

### 2.2 Data understanding

| Column | Qualitative/Quantitative | Discrete/Continuous | Nominal/Ordinal/Binary |
|---|---|---|---|
| **Data As Of** | Qualitative | | Nominal |
| **Start Date** | Qualitative | | Nominal |
| **End Date** | Qualitative | | Nominal |
| **Group** | Qualitative | | Nominal |
| **Year** | Quantitative | Discrete | Ordinal |
| **Month** | Quantitative | Discrete | Ordinal |
| **State** | Qualitative | | Nominal |
| **Condition Group** | Qualitative | | Nominal |
| **Condition** | Qualitative | | Nominal |
| **ICD10_codes** | Qualitative | | Nominal |
| **Age Group** | Qualitative | | Ordinal |
| **COVID-19 Deaths** | Quantitative | Discrete | |
| **Number of Mentions** | Quantitative | Discrete | |
| **Flag** | Qualitative | | Nominal |

## 2.3 Data cleaning and wrangling

**Cleaning (integrity/quality checks):**

| Problem | Description | Solution |
|---|---|---|
| Missing values | The 159452 null values for 'COVID-19 Deaths' are due to the following reason noted in the 'Flag' column: "One or more data cells have counts between 1-9 and have been suppressed in accordance with NCHS confidentiality standards." The 'Flag' column has null values where there were no necessary comments such as this. | I replaced these null values with random integers 1 – 9 because the data is already provisional and is constantly updated every week. We know that entries with this footnote had at least 1 death but are not disclosing this information due to privacy. In order to make our data as useful as possible I chose to input these random integers. I do not believe that this will cause us to have unrealistic data that is not worth studying. As the data is actively being updated, death counts in the data set may be lower than reality anyway. |
| Duplicates | N/A | |
| Outliers | There are 5875 entries of total 511980 that are considered outliers for COVID-19 Death Counts. | We will leave these outliers alone since this data is sourced and collected reliably and does not indicate any error in the data. |

**Wrangling:**

| | |
|---|---|
| Deleted columns | Removed the following columns for sake of relevance: 'Data As Of', 'Start Date', 'End Date', 'ICD10_codes', 'Number of Mentions'. Removed 'Group' column after removing distinction between 'By Total', 'By Year' and 'By Month' because we are only interested in 'By Month' information and now that column is irrelevant. |
| Replaced Null Values | I think the missing values in the COVID-19 Death column can be replaced with a random |

| | |
|---|---|
| | integer 1-9 because we know that at least one death occurred and can't be reported because of confidentiality and privacy issues, and based on how they update this data every month the newer counts in particular are likely an underestimate of the total deaths as they catch up with incoming death certificate data, so I do not believe that the data will be skewed irreparably by doing this. It is worth noting that null values make up 31% of values in the 'COVID-19 Deaths' column. |
| Remove rows | The ‘Group’ column has ‘By Total’, ‘By Year’ and ‘By Month’ classifications. We want to analyze this data by month and thus don’t need the other classifications as it is irrelevant information for this analysis. I removed all entries with ‘By Total’ and ‘By Year’ classifications.<br>The ’State’ column also has irrelevant information by giving us totals for the entire United States. Removed all rows with ‘United States’ classification so we can focus on a geographical analysis of COVID-19 deaths in the US. |

**Limitations:** As previously mentioned, the more recent data is provisional and the data overall is updated monthly, which means there is some allowable variation in death counts as the data is updated. We are limited by how behind NCHS is in collecting and coding the data. There were also many missing values that necessitated the use of value replacement.

**Ethical concerns:** I do not see any issues because NCHS and the CDC have already accounted for personal identification factors by not publicly distributing information on death counts of 9 or less.

# Questions to Explore and Hypotheses

With my project, I would like to answer the following questions:

1. Which states have the most COVID-19 Deaths?
2. Which age groups have the most COVID-19 Deaths?
3. Is there a certain season (month range) that has increased COVID-19 Deaths?
4. How have COVID-19 deaths changed over time?
5. Which conditions have the most association with COVID-19 deaths?

Hypotheses:

- If a state is in the South, then the COVID-19 death count will be higher.
- If a person is 65 and older, then death due to COVID-19 is more likely.
- If a person has a respiratory disease, then death due to COVID-19 is more likely.
- COVID-19 deaths have overall decreased since 2020.
- More COVID-19 Deaths occur during Flu Season (Dec – March).