Final Project: COVID-19 Death Analysis

Weixi Pan

2023-12-07

Introduction

Background

Coronavirus disease (COVID-19) is an infectious illness caused by the SARS-CoV-2 virus. It is well-established that the majority of individuals infected with COVID-19 will undergo a mild to moderate respiratory illness and recover without necessitating specialized treatment. However, a subset of individuals may experience severe illness or succumb to the disease, regardless of age.

The dataset utilized in this project is sourced from the Centers for Disease Control and Prevention (CDC) and provides information on deaths related to COVID-19, pneumonia, and influenza. These deaths are reported to the National Center for Health Statistics and are categorized by sex, age group, and jurisdiction of occurrence. The provisional counts for COVID-19 deaths are derived from the current flow of mortality data within the National Vital Statistics System. National provisional counts encompass deaths occurring in the 50 states and the District of Columbia that have been received and coded as of the specified date.

Data Description

The dataset comprises 137,700 observations and encompasses 16 variables. Each row delineates COVID-19 deaths categorized by sex, age, state, year, and month. The dataset is characterized by 8 character variables and 8 integer variables.

Given the incompleteness of the dataset for the year 2023, our analysis focuses exclusively on the years 2020, 2021, and 2022. The names and descriptions of these years are presented below.

Variable_Name	Description
Data As Of	Date of analysis
Start Date	First date of data period
End Date	Last date of data period
Group	Indicator of whether data measured by Month, by Year, or Total
Year	Year in which death occurred
Month	Month in which death occurred
State	Jurisdiction of occurrence
Sex	Sex
Age Group	Age group
COVID-19 Deaths	Deaths involving COVID-19
Total Deaths	Deaths from all causes of death
Pneumonia Deaths	Pneumonia Deaths
Pneumonia and COVID-19 Deaths	Deaths with Pneumonia and COVID-19

Variable_Name	Description
Influenza Deaths	Influenza Deaths
Pneumonia, Influenza, or COVID-19 Deaths	Deaths with Pneumonia, Influenza, or COVID-19
Footnote	Suppressed counts (1-9)

The pivotal variables utilized in this project encompass End Date, Year, Month, State, Sex, and Age Group. The primary output variable in our analysis is COVID-19 Deaths. This dataset provides a comprehensive record of the case numbers for COVID-19 deaths in the United States, categorized by state, gender, age, and other parameters, spanning from January 1, 2020, to December 31, 2022. Subsequent sections of the analysis will delve into more specific examinations and interpretations based on these key variables.

Abstract

As commonly understood, the majority of individuals infected with COVID-19 will undergo mild to moderate respiratory illness and recover without the need for specialized treatment. Nevertheless, a subset of individuals may experience severe illness or succumb to the virus, irrespective of age. The dataset utilized in this project originates from the Centers for Disease Control and Prevention (CDC) and pertains to provisional COVID-19 deaths in the United States, categorized by date, state, sex, and age groups.The primary objective of this project is to address the following questions: Does the incidence of COVID-19-related deaths exhibit a temporal decrease over time? To what extent do COVID-19-related deaths vary based on factors such as state, sex, and age group?

Methods

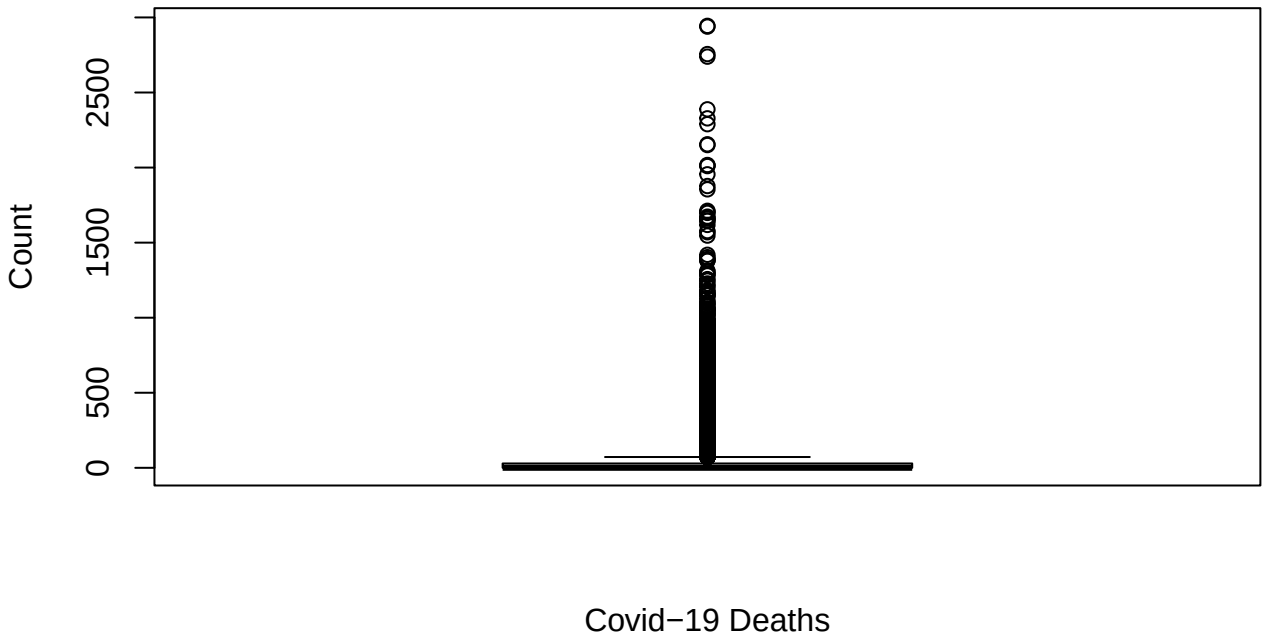
Data Preparation

The initial step in data cleaning involves obtaining a comprehensive overview of the entire dataset. Notably, each variable within this dataset features a "summary observation" designed to provide a condensed representation of the respective variable. For instance, in the "State" column, the table includes data for individual states along with observations labeled "United States," offering a summary of the total count aggregated across other variables. Similarly, the "Sex" variable incorporates a "All Sexes" summary, and the "Age Groups" variable includes a "All Ages" summary.

In preparation for subsequent analyses, it is imperative to selectively choose key variables for utilization and filter out these summary observations to eliminate redundancy. This strategic approach ensures a focused and streamlined dataset, enhancing the accuracy and precision of future analyses.

Following the exclusion of "summary rows" and the removal of missing values, the subsequent step involves scrutinizing the dataset for irregular values and subsequently eliminating them. Specifically focusing on the "COVID-19 Deaths" column in the refined dataset, a summary analysis reveals the absence of negative values. However, the maximum value is notably elevated, reaching 2944.

To further investigate and address potential outliers, a box plot analysis was conducted. This examination identified several exceptionally large values surpassing 2500. As a corrective measure, these outliers were filtered out, yielding refined results as detailed below.



End Date	Group	Year	Month	State	Sex	Age Group	COVID-19 Deaths	Total Deaths
01/31/2021	By Month	2021	1	California	Male	50-64 years	2739	5700
01/31/2021	By Month	2021	1	California	Male	65-74 years	2939	6026
01/31/2021	By Month	2021	1	California	Male	75-84 years	2944	6297
01/31/2021	By Month	2021	1	California	Female	85 years and over	2756	8248

Upon closer examination of the table, it is apparent that these larger values correspond to January 2021 in California. Given their meaningful nature, it would be inappropriate to classify them as irregular values warranting removal. In addition to the aforementioned steps, we have also undertaken the conversion of the date variable format. This conversion enhances the ease of subsequent analysis and visualization, streamlining the dataset for more effective interpretation.

Data Analysis

Now that we have obtained a clean dataset, it is opportune to proceed with data analysis. Our initial inquiry involves exploring the relationship between the number of deaths and dates. To address this, we begin by generating a summary table that delineates the data for different years.

Year	Total_Death_Cases	Avg_Death_Cases	Min_Death_Cases	Max_Death_Cases
2020	445344	39	0	2388
2021	588753	56	0	2944
2022	277651	30	0	955
2023	45349	6	0	286

The table above provides a comprehensive summary of total death cases, average COVID-19 death cases, as well as the minimum and maximum values for each year. A notable observation is that 2021 emerges as the year with the highest total death data, encompassing both average and maximum values. Conversely, all values for 2022 are comparatively the smallest.

To further investigate the distribution by months, a similar summary is conducted for all three years, and the results are arranged in descending order. However, the outcomes for the three years exhibit substantial differences, making it challenging to discern discernible patterns. Consequently, it can be inferred that the number of deaths exhibits little correlation with the month across these three years. To illustrate, the results for 2021 are presented in the table below, with interactive tables for all three years available on the project website.

Month	Total_Death_Cases	Avg_Death_Cases	Min_Death_Cases	Max_Death_Cases
1	124487	130	0	2944
9	89009	90	0	1645
8	67969	74	0	1664
12	59048	62	0	587
2	57708	66	0	1282
10	57366	59	0	808
11	41353	46	0	424
3	27857	33	0	456
4	23030	28	0	266
5	18299	24	0	219
7	13694	18	0	338
6	8933	12	0	139

To address the second question, a similar summary is conducted, stratifying the data by state and age groups. The outcomes are presented below.

State	Total_Death_Cases	Avg_Death_Cases
California	138673	134
Texas	136778	135
Florida	100201	102
Pennsylvania	61278	70
Ohio	58336	65
New York	48330	59
Georgia	45734	53
New York City	45720	57
Illinois	45643	53
Michigan	44038	51

In the context of states, we have chosen to display solely the top 10 results, while the full interactive dataset is available in the website Appendix. The table reveals that Texas, California, and Florida occupy the top three positions in terms of both total death cases and averages. Conversely, Vermont, Alaska, and Hawaii are the states with the lowest incidence of cases.

Age Group	Total_Death_Cases	Avg_Death_Cases
85 years and over	308590	84
75-84 years	296543	83
65-74 years	251901	76
50-64 years	197344	64
55-64 years	153098	53
45-54 years	63855	26
40-49 years	38857	17
35-44 years	22985	11
30-39 years	13396	7
25-34 years	7061	4
18-29 years	2866	1
15-24 years	547	0
0-17 years	54	0
1-4 years	0	0
5-14 years	0	0
Under 1 year	0	0

The table above provides a summary of death cases categorized by age group. It is evident from the table that the number of death cases exhibits a discernible increasing trend with advancing age.

Results and Discussion

Following a series of cleaning, summarization, and exploration of the dataset, several visualization results for the posed questions have been derived. These results will be presented sequentially, and an interactive version of these plots will be available on the project website.

Covid-19 Deaths by Date

The initial chart comprises a line plot depicting COVID-19 death cases over the date range from January 2020 to November 2022. Each distinct color within the plot corresponds to a different state, allowing for an examination of the distribution of cases across states over time. Notably, the chart reveals a discernible trend in the overall number of deaths from 2020 to 2022, exhibiting an initial increase followed by a subsequent decline. The data peaks in early 2021, with more recent death data (end of 2022) significantly lower than the levels observed during the initial spread of the virus in 2020. This suggests a substantial reduction in the death rate from COVID-19 over the specified period.

When considering individual states, the region with the highest number of deaths in 2020 is New York City, as indicated by the elevated blue lines, reflecting the proximity to New York State. Post-2020, California's data experienced a rapid ascent, reaching its peak in the early part of 2021. Towards the end of 2021, there is a subsequent decline in California, with Florida and Texas emerging as states with the highest peaks. This nuanced analysis provides insights into the temporal and regional dynamics of COVID-19 death cases, offering a comprehensive view of the evolving trends across states.

Covid-19 Deaths by Ages

The chart reveals a conspicuous upward trend in the number of death cases with advancing age. Notably, the bars of various colors in the figure denote different genders, with blue representing males and red representing females. Upon closer examination, it becomes evident that, except for the age groups 55-64 and those aged over 85, the majority of other age groups exhibit a higher incidence of male deaths compared to females.

This gender-related disparity is clearly illustrated in the graph, highlighting a nuanced demographic pattern within the distribution of COVID-19-related fatalities across age and gender categories.

Covid-19 Deaths by Total Deaths

This plot illustrates the relationship between the number of COVID-19 deaths and the total number of deaths. Notably, a discernible correlation exists between these two variables. As the total number of deaths increases, the count of COVID-19 deaths exhibits a pattern of close-to-linear growth. This observation suggests a proportional relationship between the overall mortality rate and the specific impact of COVID-19.

Examining the plot from a state-specific perspective further highlights the utility of this interactive visualization in identifying outlier data points. These distinctive observations contribute to a nuanced understanding of the variability in COVID-19-related fatalities across different states, emphasizing the significance of considering both total and COVID-19-specific mortality rates in comprehensive analyses.

Covid-19 Deaths Counts under 250

In the analysis of sex groups, the subsequent figure presents a histogram of death cases differentiated by gender. To enhance clarity, the figure exclusively displays data with death case counts below 250, as the majority of observations in the dataset fall within this range. This approach facilitates a more intuitive indication of the gender distribution in the dataset. The histogram unmistakably illustrates that, for death cases below 250, there is a prevalent predominance of males over females.

By synthesizing the insights garnered from the visualizations above, it becomes evident that, over the three-year period spanning from 2020 to 2022, the dataset reflects a higher incidence of COVID-19-related deaths among males than females. This comprehensive observation underscores the importance of considering gender disparities when examining the demographic patterns of COVID-19 fatalities.

Conclusion

Over the course of the past three years, the number of COVID-19 deaths has exhibited fluctuations, indicating an upward trajectory until the inception of 2021, followed by a discernible decline. An analysis of individual states reveals that California, Texas, and Florida consistently occupy the top three positions in terms of total COVID-19 death data. Notably, New York City, California, and Florida have all experienced peaks in COVID-19 deaths at different junctures.

The distribution of COVID-19 death cases reveals an age-related increase, with higher age groups demonstrating a higher incidence of fatalities. Furthermore, when considering gender, a predominant pattern emerges wherein the total number and proportion of COVID-19 deaths within the male demographic surpass those within the female demographic. These multifaceted observations offer valuable insights into the temporal, geographical, age-related, and gender-related dynamics of COVID-19 fatalities over the specified three-year period.

Reference

[1]Centers for Disease Control and Prevention, https://www.cdc.gov/nchs/covid19/covid-19-mortality-data-files.htm

[2]Covid-19, non-Covid-19 and excess mortality rates not comparable across countries, https://pubmed.ncbi.nlm.nih.gov/34338184