COVID-19 Deaths in the United States: A Data Analysis Perspective

Lucien Junior Jura

Department of Computer Science. Wentworth Institute of Technology, Data Science

This project explores the impact of COVID-19 on mortality across the United States using publicly available datasets. The analysis focuses on death counts over time and across states, considering different age groups and demographics. Data cleaning, transformation, and visualizations were applied to uncover trends, particularly focusing on total deaths per month and per state. This report provides insights into how the pandemic affected different regions and populations and offers a foundation for further study.

# COVID-19, Data Analysis, U.S. States, Mortality Trends, Visualization

# Introduction (*Heading 1*)

Begin your introduction by clearly presenting your topic and explaining its significance - why it is important or interesting. Instead of listing questions separately, weave them together into a cohesive narrative that naturally connects the topic, its relevance, and its context. Provide an overview of existing research and key findings in this area, incorporating necessary citations to support your discussion. Your goal is to create a compelling introduction that sets the stage for your report.

The COVID-19 pandemic had an unprecedented impact on global health, economies, and societies. In the United States alone, millions were infected, and hundreds of thousands lost their lives. Understanding the patterns and trends of COVID-related deaths is essential not only for historical documentation but also for planning and preparedness in the face of future pandemics.

This project focuses on analyzing COVID-19 mortality data in the United States. By examining data trends by month, state, and demographic attributes, the goal is to identify when and where the pandemic hit hardest. The analysis uses Python-based tools and publicly available datasets. Previous studies have emphasized factors such as age, comorbidities, and access to healthcare in determining outcomes, and this analysis attempts to explore such patterns on a macro level using aggregated data.

# Datasets

## Source of dataset (Heading 2)

In this section, introduce your dataset by explaining its source—where you obtained it and whether it is from a credible provider. Include details such as when the dataset was generated and how it was created by its original author. If you generated the dataset yourself, describe the methods and processes you used.

The dataset used in this project was sourced from Google Drive, referencing publicly available mortality data compiled by health organizations such as the CDC. It contains weekly death counts and includes detailed breakdowns by cause of death, state, gender, and age group.

## Character of the datasets

Describe the dataset’s format and size. Additionally, provide an overview of the dataset’s characteristics, including its features, size, structure, and any relevant attributes that are important for your analysis. Describe the dataset’s format and size, as well as its key features, including the parameters, columns, rows, and character attributes along with their respective units. Using a table to present this information is recommended for clarity. Explain whether you cleaned the data or converted any units, specifying the formulas or rules applied. If multiple datasets were combined, describe how they were merged. Additionally, mention if you created any new categories for analysis, detailing what they are and how they were generated. Providing this background ensures transparency and helps readers understand the reliability and relevance of your data.

The dataset is in CSV format and includes the following key columns:

* **State**: U.S. state name
* **COVID-19 Deaths**: Deaths explicitly attributed to COVID-19
* **Pneumonia Deaths, Flu Deaths**: Related causes of death for comparative context
* **Age Group, Sex**: Demographic info
* **Date Columns**: "Start Date", "End Date", and "Data As Of"

The dataset required cleaning to handle missing values and convert date strings into proper datetime format. We filtered for rows where “COVID-19 Deaths” was non-null and aggregated data by month and state to create meaningful visualizations. A summary table was created to ensure transparency of key variables.

# Methodology

In this part, you should give an introduction of the methods/model. First, what’s the method/model. What’s the assumption of this method/model. What’s the advantage/disadvantage of this method/model. Why did you choose it. What Python module or function do you apply to apply this method/model. Any optional input/extra work did you adjust to make the results better. If you have multiple methods, feel free to use subsection A., B. to separate them.

Example: Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections A-D below for more information on proofreading, spelling and grammar.

## Method A

Example: The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

*a**b* 

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1)”, not “Eq. (1)” or “equation (1)”, except at the beginning of a sentence: “Equation (1) is . . .”

## Method B

* Bulletin 1
* Bulletin 2.
* Bulletin 3

Identify applicable funding agency here. If none, delete this text box.

## Method C

Example: The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

*a**b* 

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1)”, not “Eq. (1)” or “equation (1)”, except at the beginning of a sentence: “Equation (1) is . . .”

An excellent style manual for science writers is [7].

# Results

In this section, present your findings using an appropriate method, such as equations, numerical summaries, or visualizations like charts and graphs. Clearly explain all results and provide guidance on how to interpret them. If any unexpected results arise, discuss possible reasons or contributing factors. To improve clarity and organization, consider using subsections (e.g., A, B) to separate different aspects of your results.

In this project, the primary method used was **data aggregation and visualization** to analyze trends in COVID-19-related deaths across U.S. states. Python libraries such as **Pandas**, **NumPy**, and **Matplotlib** were utilized to clean, group, and visualize the data effectively. Our approach was exploratory in nature, aiming to uncover time- and location-based patterns in the data.

Result A

The dataset contained death counts attributed to COVID-19 and other causes, along with demographic and geographic information. The following steps were taken:

* **Removed irrelevant rows** (e.g., "United States" aggregate rows).
* **Filtered for non-null values** in the "COVID-19 Deaths" column.
* **Converted date columns** (e.g., "Start Date", "End Date") into datetime format using pd.to\_datetime().
* **Created new columns** for month and year using dt.month\_name() and dt.year.

## To analyze trends over time and by state, the dataset was grouped using groupby() to compute total deaths per month and per state.

## Results C

*We used* ***Matplotlib*** *to generate bar charts and scatter plots that represent:*

*Monthly COVID-19 deaths across all states.*

*Total deaths per state, ranked from highest to lowest.*

*Advantages of this method:*

*Clearly reveals time-based trends.*

*Helps identify geographic disparities.*

*Intuitive for communicating findings to non-]technical audiences.*

*Limitations:*

*Visualizations alone cannot establish causality.*

*May mask underlying demographic patterns without further stratification.*

*Optional adjustments made:*

*Added grid lines and axis labels for clarity.*

*Sorted bar plots to emphasize high-death states (e.g., New York, California).*

1. Table Type Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

1. Sample of a Table footnote. (*Table footnote*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

# Discussion

Every method/project has its shortage or weakness. Please discuss the unsatisfied results in your project. And discuss the feasible suggestions of future work to revise/improve your result.

While our analysis revealed valuable insights into COVID-19 death trends across time and geography, there were several limitations and areas for improvement:

1. **Data Reporting Lag**: Many states report deaths at different rates, leading to potential inconsistencies or underreporting, especially in the most recent months of the dataset.
2. **Lack of Granularity**: The dataset aggregated deaths by state and time period, but lacked detailed demographic breakdowns (e.g., age, race, vaccination status), which could reveal disparities in outcomes.
3. **No Direct Causal Inference**: Our visualizations and summaries are descriptive; they do not explain *why* deaths rose or fell during specific periods.
4. **Assumption of Consistent Measurement**: The analysis assumes all states reported deaths using the same criteria, which may not hold true.

**Future Work Suggestions:**

* **Incorporate external variables** such as vaccination rates, mobility data, or public health policies to examine possible causes behind spikes in deaths.
* **Use time-series forecasting models** (e.g., ARIMA, Prophet) to predict future trends.
* **Apply statistical tests or machine learning models** for more in-depth inference or classification.
* **Conduct demographic stratification** to uncover hidden disparities within states.

Conclusion

In this part, you should summarize your project. What important results did you find for your topic, and what’s the effect of this result on the real world?

Example: In this project, we analyzed U.S. COVID-19 death data using Python libraries like Pandas and Matplotlib. We processed, grouped, and visualized trends in deaths over time and by state. Key takeaways include:

* COVID-19 deaths peaked in winter months (e.g., January 2021), aligning with known waves.
* States like California, Texas, and New York consistently reported higher death totals.
* Our findings highlight the importance of consistent data reporting and visual analytics in understanding pandemic dynamics.

**Impact:**  
Our visual approach supports public health decision-making by identifying critical time periods and geographic hotspots. It also sets the foundation for more advanced forecasting and policy analysis tools.

##### Acknowledgment *(Heading 5)*

##### We would like to thank **Prof. Pang Weijie** for guidance throughout this project. We also acknowledge the contributions of **Jacob, Joelle, and Junior** as team members who collaborated on data cleaning, analysis, and visualization.

##### References

Use the IEEE format for the citation. The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...” Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

1. [1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” *Phil. Trans. Roy. Soc. London*, vol. A247, pp. 529–551, April 1955.
2. [2] J. Clerk Maxwell, *A Treatise on Electricity and Magnetism*, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp. 68–73.
3. [3] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in *Magnetism*, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
4. [4] K. Elissa, “Title of paper if known,” unpublished.
5. [5] R. Nicole, “Title of paper with only first word capitalized,” *J. Name Stand. Abbrev.*, in press.
6. [6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” *IEEE Transl. J. Magn. Japan*, vol. 2, pp. 740–741, Aug. 1987.
7. [7] M. Young, *The Technical Writer’s Handbook*. Mill Valley, CA: University Science, 1989.

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord “Format” pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.