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

(COMP3125 Individual Project)

Lucien Junior Jura
Wentworth Data Science

***Abstract*— 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 state. This report provides insights into how the pandemic affected different regions and populations and offers a foundation for further study.**

I. KEYWORDS— COVID-19, Data Analysis, U.S. States, Mortality Trends, Visualization

Introduction (Heading 1)

The COVID-19 pandemic has 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.

II. DATASETS

A. Source of dataset (Heading 2)

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.

B. Character of the datasets

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 the 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.

III. METHODOLOGY

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.

A. Method 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.

B. Method B

To analyze trends over time and by state, the dataset was grouped using the group by() function to compute the total number of deaths per month and per state.

C. Method C

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

- Monthly COVID-19 deaths across all states.

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

- 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).

IV. Results

After applying the methodology above, several important trends emerged from the data analysis. Visualizations and aggregated statistics helped uncover patterns across time and geography.

A. Result A

The number of COVID-19 deaths varied significantly month by month. Peaks were observed during winter months (e.g., December 2020 and January 2021), reflecting known pandemic waves.

Fig. 1. Total COVID-19 deaths per month in the U.S.

Interpretation:

- Sharp increases correspond to nationwide surges.
- Declines align with the rollout of vaccines and public health interventions.

B. Results B

States such as **California**, **Texas**, and **New York** reported the highest total death counts, consistent with population size and early outbreak severity.

Fig. 2. Total COVID-19 deaths by state.

Interpretation:

- Population size and urban density likely contributed to higher death counts.
- Variability in state-level public health policies may also explain the spread.

C. Results C

To give context to the raw data, we included preview tables showing selected rows and summaries. df[["state", "COVID-19 Deaths", "Start Date", "End Date"]].head()

V. Discussion

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

VI. Conclusion

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 want 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

[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.

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.