260_Final_Report

Abstract (150-200 words)

- **Purpose**: The abstract provides a concise summary of your project, including its objectives, key findings, and significance. Write this section last, after completing all other sections, to accurately reflect your project's focus and main results.
- **Guidelines**: Limit this section to 150-200 words. Briefly outline the purpose of your study, the approach you used, and the primary results and conclusions. The abstract should be clear, succinct, and give readers an immediate understanding of what your project entails.

Introduction (500-600 words)

The COVID-19 pandemic has profoundly influenced global health, economies, and societal structures. From January 2020 to December 2024, the United States experienced multiple waves of COVID-19 infections and deaths. These waves provide a unique opportunity to analyze the evolution of COVID-19's impact over time. By examining mortality data across states and time periods, this study aims to investigate patterns and trends to better understand the trajectory of this pandemic.

Recent studies have investigated and highlighted the importance of analyzing mortality trends to better understand the pandemic's impact. For instance, Chan et al. (2021) quantified the pandemic's effects on excess mortality and life expectancy (Chan et al., 2021). Moreover, Woolf et al. (2021) analyzed excess deaths in the United States and showed increased mortality from non-COVID-19 causes, such as heart disease and Alzheimer's disease during COVID-19 waves, demonstrating the pandemic's indirect effects on public health outcomes (Woolf et al., 2021). While these studies have provided valuable insights, there are limitations that our study intend to address. Most prior analyses focus on aggregated data or specific time periods and pay little attention to state-level variations and temporal shifts of the pandemic's impact. Our study tries to offer a much comprehensive and longitudinal analysis that encompasses the entire pandemic period in both national and state level.

In this study, the analysis focuses on several key aspects of the pandemic. First, we divide the timeline into distinct waves based on data visualizations of infection and mortality trends. This section allows for understanding the virus's evolution and impact over time. Second, we compute death rates for each state during these periods, identifying states that performed better or worse in terms of mortality outcomes. These findings could potentially identify the effectiveness of various public health strategies and problems within state-level healthcare systems. Third, we explore whether COVID-19 became more or less virulent across different periods by investigating changes in mortality rates and excess deaths over time. Furthermore, we extend the scope to estimate weekly excess mortality for each state and evaluate whether COVID-19 deaths explain these excesses.

By investigating those key aspects of the pandemic, this study seeks to provide a comprehensive understanding of the dynamics of COVID-19 mortality in the United States, potentially contributing to future pandemic analysis and serving as resources for public health officials to study healthcare systems and improve response strategies and allocation of resources across the nation.

Methods (600-700 words)

- **Purpose**: This section details the data sources, methods, and analytical techniques you used to conduct your analysis. It should be specific enough that someone else could replicate your study using the same resources and approach.
- **Guidelines**: Describe the dataset(s) you used, including information about data collection (e.g., sources, time frame). Outline your approach for cleaning and analyzing the data, including any statistical or computational methods applied. Clearly explain any assumptions or limitations in your approach.

Data Q123

1.Population data: Data source: NST-EST2024-POP.xlsx information about data collection source: data collected by U.S. Census Bureau. From website: https://www.census.gov/data/tables/time-series/demo/popest/2020s-state-total.html

data collection time frame: from 2020 to 2024. Provides annual state-level population estimates for the years 2020 to 2024. This dataset is used to calculate per capita COVID-19 case and death rates.

2.Regional Data: Source: JSON file hosted on GitHub. URL: https://github.com/datasciencelabs/2024/raw/refs Description: Defines regional groupings of U.S. states and territories (e.g., "NY & NJ & PR & VI") for aggregated regional analysis.

3.COVID-19 Cases and Death Data: Source: Centers for Disease Control and Prevention (CDC) API. data collection time frame: January 22, 2020, to May 10, 2023 API Endpoint:

https://data.cdc.gov/resource/pwn4-m3yp.json. Description: Provides daily records of new COVID-19 cases and deaths at the state level from 2020 to 2024.

Data Q45

4.Weekly Counts of Deaths by Jurisdiction and Age Data: Data source: The dataset "Weekly Counts of Deaths by Jurisdiction and Age" is sourced from the National Center for Health Statistics (NCHS). Access URL: https://data.cdc.gov/NCHS/Weekly-Counts-of-Deaths-by-Jurisdiction-and-Age/y5bj-9g5w/about_data File used: Weekly_Counts_of_Deaths_by_Jurisdiction_and_A Time period: The dataset covers weekly counts of deaths from January 10, 2015, to September 16, 2023. Description: This dataset provides weekly counts of deaths categorized by jurisdiction and age groups across the United States. It spans multiple years and is used to analyze trends in mortality, including patterns during significant public health events such as the COVID-19 pandemic.

5. Intercensal Population Estimates (2010-2020): Data source: U.S. Census Bureau. Access URL: https://www.census.gov/data/tables/time-series/demo/popest/intercensal-2010-2020-national.html File used: nst-est2020int-pop.xlsx. Description: This dataset contains intercensal population estimates for the United States, bridging the gaps between decennial censuses (2010 and 2020). It offers annual national-level population figures for demographic analyses, allowing researchers to adjust for population changes over the decade. Both datasets are integral for demographic and mortality analysis, aiding in understanding trends over time and their implications on public health and population dynamics.

Results (500-600 words)

- **Purpose**: The results section presents the main findings of your analysis without interpretation. Organize the data logically to highlight key insights, using tables, figures, and charts to illustrate trends and comparisons.
- **Guidelines**: For each result, briefly describe it and refer to relevant visuals or tables where appropriate. Do not provide explanations or discuss implications in this section; focus only on presenting the findings clearly and accurately.

Q1

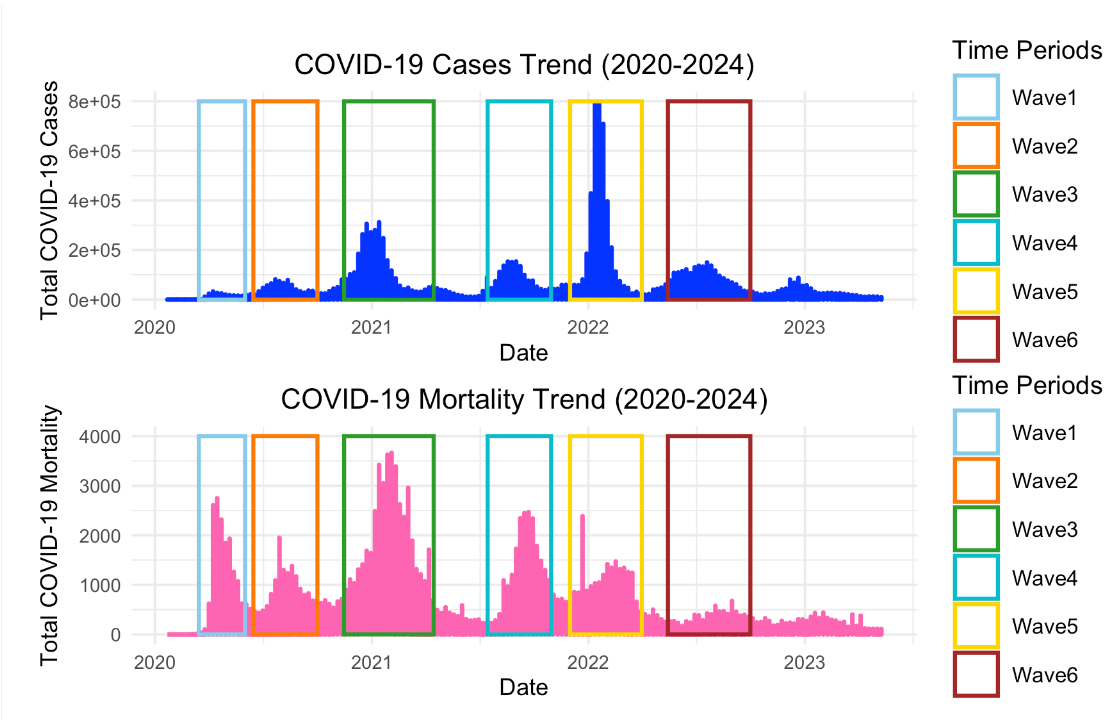
Divide the pandemic period, January 2020 to December 2024 into *waves*. Justify your choice with data visualization

```
library(magick)
```

```
Linking to ImageMagick 6.9.12.93
Enabled features: cairo, fontconfig, freetype, heic, lcms, pango, raw, rsvg, webp
Disabled features: fftw, ghostscript, x11
```

```r
img <- image_read("graph1.png")
print(img)
```

```
  format width height colorspace matte filesize density
1    PNG  1826   1132       sRGB  TRUE   268718   57x57
```



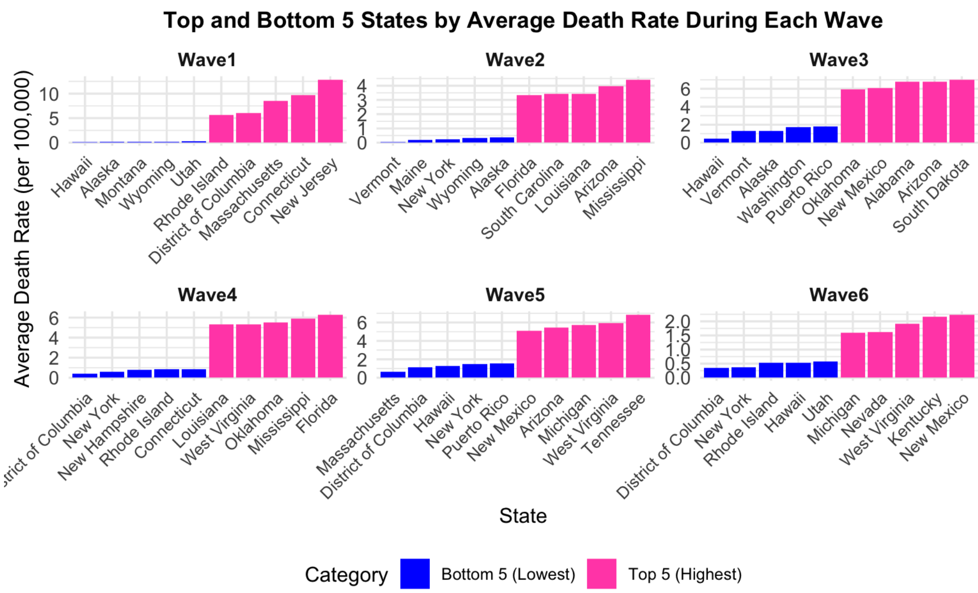
The COVID-19 case and mortality trends from 2020 to 2024 reveal six distinct waves, each varying in intensity and duration. Case trends show an initial gradual increase during Wave 1, followed by a sharp rise in Wave 2, peaking significantly in Wave 3, and then tapering into moderate peaks during Waves 4, 5, and 6. Mortality trends closely align with these waves, with the highest mortality observed in Wave 3, a noticeable decline in Wave 4, and relatively lower but steady levels in Waves 5 and 6. The provided charts highlight the temporal dynamics of the pandemic, illustrating the fluctuation of cases and mortality over time.

Q2

For each period compute the deaths rates by state. Describe which states did better or worse during the different periods.

```r
library(magick)
img <- image_read("graph2.png")
print(img)
```

```
  format width height colorspace matte filesize density
1    PNG  1540    951       sRGB FALSE   226266   72x72
```



During Wave 1, Northeastern states such as New Jersey, Connecticut, and Massachusetts had the highest death rates, while states like Hawaii, Alaska, and Montana, with smaller populations and lower densities, experienced the lowest rates. In Wave 2, Southern states including Mississippi, Arizona, and Louisiana performed the worst, whereas states like Vermont, Maine, and Wyoming fared better with the lowest death rates. Wave 3 saw high death rates in Midwest and Southern states such as South Dakota, Arizona, and Alabama, while coastal states like Hawaii, Vermont, and Washington reported the lowest rates. During Wave 4, states like Florida, Mississippi, and Oklahoma struggled the most, but Northeastern regions such as New York, New Hampshire, and the District of Columbia continued to maintain low death rates. In Wave 5, states in the Midwest and Mountain West, including Tennessee, West Virginia, and Michigan, had the highest death rates, while Massachusetts, Puerto Rico, and Hawaii were among the lowest. Finally, Wave 6 highlighted states such as New Mexico, Kentucky, and West Virginia with the worst outcomes, while states like New York, Hawaii, and Rhode Island continued to perform well with the lowest death rates.

Q3

Describe if COVID-19 became less or more virulent across the different periods.

Discussion (600-700 words)

- **Purpose**: In the discussion, interpret the significance of your findings, explore potential implications, and relate the results back to your initial research questions or hypotheses. This section allows you to discuss any patterns, unexpected findings, or limitations and suggest possible future research.
- **Guidelines**: Analyze your results in the context of your research question, linking them back to the background information from the introduction. Consider what your findings reveal, any limitations they may have, and how they might impact future work or policy. End with a brief conclusion summarizing your main insights.

Q4 Estimate excess mortality for each week for each state. Do COVID-19 deaths explain the excess mortality?

COVID-19 deaths partially explain excess mortality, as evidenced by the statistically significant positive relationship between COVID-19 death rates and excess mortality in the regression analysis. The coefficient of 0.146 indicates that for every additional 10 COVID-19 deaths per 100,000 people, excess mortality increases by 1.46 per 100,000 on average. However, the low R2R^2R2 value (0.03076) suggests that COVID-19 deaths account for only a small fraction of the variation in excess mortality. This is because excess mortality is a complex phenomenon influenced by multiple factors beyond direct COVID-19 deaths. These include indirect deaths caused by delayed healthcare, mental health crises, and economic disruptions, as well as reductions in other types of deaths (e.g., traffic accidents) during lockdowns. Additionally, underreporting of COVID-19 deaths and regional differences in healthcare infrastructure and policy responses further contribute to the gap between COVID-19 deaths and excess mortality. While COVID-19 deaths are an important driver, other covariates such as demographic factors, socioeconomic conditions, and state-level interventions likely play a significant role in shaping excess mortality, underscoring the need for more comprehensive models to fully explain the observed trends.

Why is R2R^2R2 Low?

The low R2R^2R2 value (0.03076) indicates that only about **3% of the variance** in excess mortality is explained by the independent variable (deaths_per_100kdeaths_per_100kdeaths_per_100k). This is expected due to the following reasons:

(1) Complex Nature of Excess Mortality

- **Definition of Excess Mortality:** Excess mortality includes all deaths above the expected baseline during the pandemic, which encompasses both direct COVID-19 deaths and **indirect deaths** caused by other factors like:
  - Healthcare disruptions (e.g., reduced access to treatments).
  - Behavioral changes (e.g., increased substance use, stress-related deaths).
  - Reduced non-pandemic-related deaths (e.g., fewer traffic accidents due to lockdowns).

  Since indirect deaths may not be closely correlated with COVID-19 death rates, this introduces a significant amount of variability, reducing the model's explanatory power.

(2) Missing Covariates

- Excess mortality is influenced by **multiple factors** that are not captured in the current model, such as:
  - **Population demographics:** Age structure and pre-existing health conditions vary across states, significantly affecting baseline mortality and vulnerability to COVID-19.
  - **Healthcare infrastructure:** Availability of hospital beds, ICU capacity, and access to care affect mortality outcomes but are not accounted for in this model.
  - **Policy interventions:** Variations in state policies (e.g., lockdowns, mask mandates) influence both COVID-19 and non-COVID deaths.
  - **Economic and social disparities:** Socioeconomic factors, such as poverty and income inequality, have a strong impact on health outcomes.

(3) High Variability in Data

- The scatterplot shows **wide dispersion** of excess mortality values across different levels of COVID-19 deaths. This reflects noise in the data caused by regional differences in how deaths are reported, recorded, and classified.
- Some states may underreport COVID-19 deaths, inflating the indirect component of excess mortality and weakening the correlation.

References

Woolf, S. H., et al. (2021). "Excess Deaths from COVID-19 and Other Causes in the US, March 1, 2020, to January 2, 2021." JAMA. Retrieved from jamanetwork.com. https://jamanetwork.com/journals/jama/fullarticle/2778361

Chan, E. Y. S., et al. (2021). "Impact of COVID-19 on Excess Mortality, Life Expectancy, and Years of Life Lost in the United States." PLOS ONE. Retrieved from journals.plos.org. https://pubmed.ncbi.nlm.nih.gov/34469474/