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.

Q1

We approached cleaning and analyzing the data systematically to ensure accuracy and usability. First, we loaded the population data from the provided Excel file and cleaned it by renaming columns, filtering out irrelevant rows, and removing unnecessary columns such as "Base_2020." We standardized the geographic area names and mapped state names to their respective abbreviations using conditional logic. Next, we reshaped the dataset from a wide to a long format to facilitate time-series analysis and integrated regional data from an external JSON source, mapping each state to its corresponding region. To analyze COVID-19 cases and deaths, we fetched data from an API, selected relevant columns, renamed them for clarity, and converted data types for numerical calculations and date handling. Anomalies such as negative values were identified and managed. We joined the cleaned population data with the case and mortality data to calculate metrics like cases and deaths per 100,000 people, enabling a comparative analysis across states and over time. We defined specific time periods, termed "waves," corresponding to COVID-19 surges and analyzed the peaks in cases and deaths for each wave. Using visualizations, we plotted trends in cases and deaths, with shaded

regions indicating distinct waves. By applying color coding and facet wrapping, we made the trends clearer, while combined plots provided a cohesive view of cases and mortality dynamics. This approach ensured that we effectively cleaned the data, integrated multiple sources, and presented insights through comprehensive visualizations.

Q2

We approached cleaning and analyzing the data by first ensuring that all necessary variables were properly formatted and any anomalies were addressed. For the mortality data, we converted the dates to a standardized format and replaced negative death rates with zeros to maintain logical consistency. We then categorized regions into broader major regions, such as "Northeast," "South," "West," and "Midwest," to facilitate regional analysis. Death rates were analyzed over time by plotting trends for each major region, highlighting the changes in mortality rates over the years.

To enhance interpretability, we used color-coded categories to distinguish between the "Top 5" (worst-performing states) and the "Bottom 5" (best-performing states) for each wave. These visualizations were organized in facets by wave, providing a clear and comparative view of state performance across different time periods. This approach ensured a comprehensive analysis by integrating data cleaning, regional classification, temporal segmentation, and detailed visualization to uncover meaningful trends and patterns in the data.

Q3

We began the cleaning and analysis process by defining the six distinct COVID-19 waves using specific start and end dates. We applied a custom function to assign each record in the cases and deaths datasets to its corresponding wave based on the date, ensuring accurate temporal segmentation. After this, we cleaned and harmonized the datasets by retaining essential columns, such as the date, wave, and rate values (e.g., cases or deaths per 100,000), and added a type label to differentiate between cases and deaths. This allowed us to unify the data structure for streamlined analysis.

We aggregated the cleaned data by wave to calculate total cases and deaths, and then computed the fatality ratio (deaths per 100 cases) for each wave as a percentage to assess the severity of outcomes. Finally, we visualized the trends of cases and deaths over time, segmented by wave, and plotted the fatality ratio to highlight variations in mortality risk across the waves. Our approach focused on accurate wave identification, precise data integration, and clear metric computation to analyze and present the pandemic's dynamics effectively.

Q4

We analyzed excess mortality trends using the **population dataset** (`nst-est2020int-pop.xlsx`) and the **death counts dataset** (`Weekly_Counts_of_Deaths_by_Jurisdiction_and_Age_20241220.csv`). Historical mortality data from 2015 to 2019 was utilized to predict future mortality patterns, which were then compared to observed data from 2020 onward. First, we enriched the death counts dataset by adding `week` and `year` columns derived from the date field, facilitating

temporal analysis. Using the enriched data, we developed a polynomial regression model for each state, predicting mortality rates (`deaths_per_100k`) as a function of week and year. This model, trained on pre-2020 historical data, was applied to forecast weekly mortality rates for each state from 2020 to 2024.

We validated these predictions by merging them with observed mortality data from the same death counts dataset for 2020 and beyond. Excess mortality was calculated as the difference between observed mortality rates and our predicted rates, providing insights into deviations during the pandemic period. The resulting data was visualized to highlight excess mortality trends across states, revealing patterns and variations.

To assess the relationship between excess mortality and COVID-related deaths, we calculated the Pearson correlation coefficient to quantify the linear association. Using the merged dataset, we constructed a linear regression model to evaluate how COVID-specific deaths predict excess mortality. The coefficients and confidence intervals from the model provided insights into the statistical significance and reliability of the findings. This analysis integrated the population and death counts datasets with predictive modeling, real-world validation, and statistical analysis to explore excess mortality trends and their link to COVID-related factors, while accounting for assumptions about population stability and addressing limitations such as data quality and external influences.

Q5

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

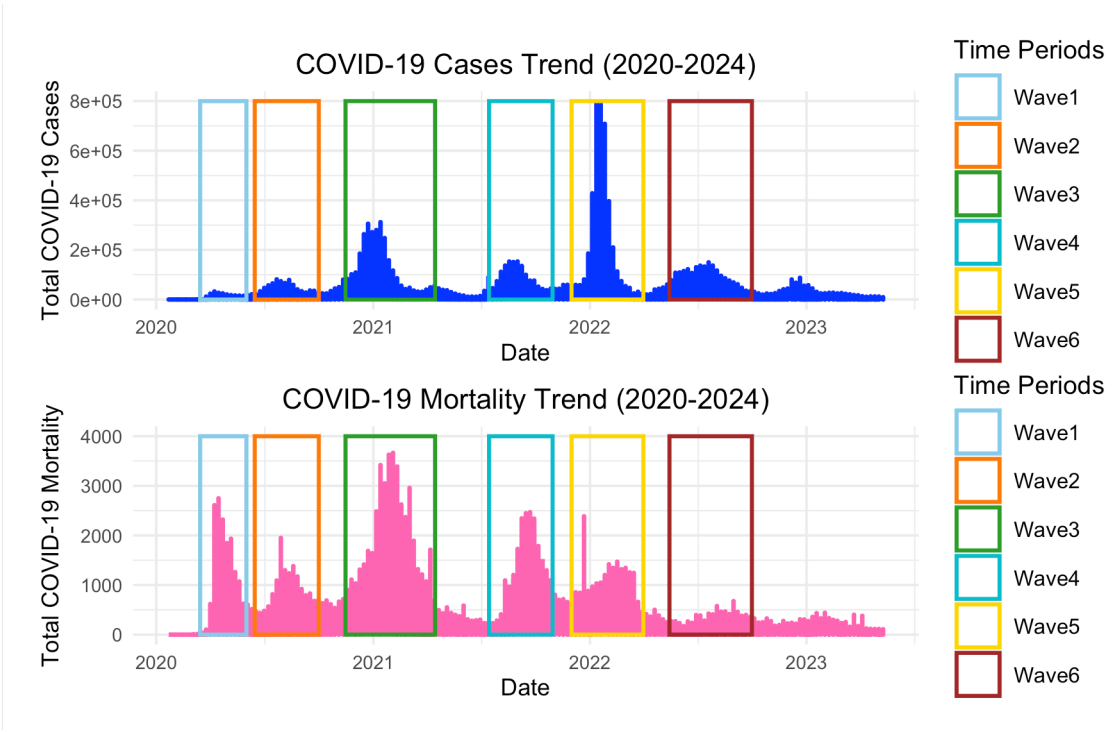
Divide the pandemic period, January 2020 to December 2024 into *waves*. Justify your choice with data visualization

```
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
```

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

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



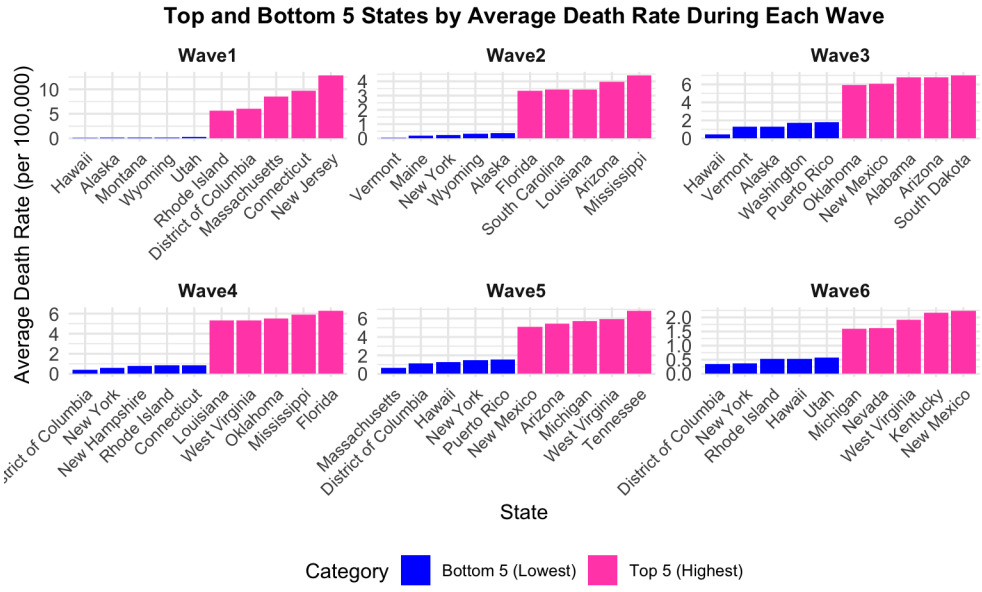
The COVID-19 case and mortality trends from 2020 to 2024 reveal six distinct waves, each varying in intensity and duration. Case trends show an initial gradual increase during Wave 1, followed by a sharp rise in Wave 2, peaking significantly in Wave 3, and then tapering into moderate peaks during Waves 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.

```
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

The fatality ratios (deaths per 100 cases) across the six COVID-19 waves show a decreasing trend over time. Wave 1 has the highest fatality ratio at 5.36%, followed by a significant drop in Wave 2 to 1.63%. Wave 3 records a slightly lower fatality ratio of 1.53%, while Wave 4 continues the decline to 1.08%. Waves 5 and 6 show further reductions, with fatality ratios of 0.64% and 0.42%, respectively. This table reflects the variation in fatality ratios across the waves.

```
library(magick)
img <- image_read("table1.png")
print(img)
```

```
  format width height colorspace matte filesize density
1    PNG  1088    406       sRGB  TRUE    49057   57x57
```

wave <chr>	fatality_ratio <dbl>
Wave1	5.3559447
Wave2	1.6330246
Wave3	1.5284469
Wave4	1.0846909
Wave5	0.6384178
Wave6	0.4228565

Q4

The results reveal a moderate positive correlation between COVID-specific deaths per 100,000 and excess mortality, with a Pearson correlation coefficient of 0.677. A linear regression model further confirms this relationship, showing that each additional COVID-specific death per 100,000 is associated with an increase of 1.642 in excess mortality per 100,000 ($p < 2e\text{-}16$), with an intercept of 1.533 ($p < 2e\text{-}16$). The model explains 45.78% of the variance in excess mortality, as indicated by an adjusted R-squared value of 0.4578, and has a residual standard error of 4.245. The 95% confidence interval for the `deaths_per_100k` coefficient ranges from 1.605 to 1.680, confirming the precision of the estimate. Supporting visuals include scatterplots illustrating the relationship between COVID-specific deaths and excess mortality, with a regression line depicting the trend, and weekly excess mortality trends highlighting temporal and geographic variability.

```
library(magick)
img <- image_read("table2.png")
print(img)
```

```
  format width height colorspace matte filesize density
1    PNG  1510    908       sRGB  TRUE   160302   57x57
```

```
Call:
lm(formula = excess_mortality ~ deaths_per_100k, data = merged_death_2020_after_prediction)

Residuals:
    Min      1Q  Median      3Q     Max
-70.147  -2.268  -0.323   1.904  43.631

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.53360    0.05664   27.07   <2e-16 ***
deaths_per_100k  1.64205    0.01890   86.90   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.245 on 8942 degrees of freedom
Multiple R-squared:  0.4579,    Adjusted R-squared:  0.4578
F-statistic:  7552 on 1 and 8942 DF,  p-value: < 2.2e-16

    (Intercept) deaths_per_100k
       1.533604        1.642054
                    2.5 %   97.5 %
(Intercept)     1.422572 1.644635
deaths_per_100k 1.605014 1.679093
```

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.

Q1

The results highlight the evolving dynamics of the COVID-19 pandemic, shaped by public health measures, vaccination efforts, and viral mutations. The sharp rise in cases and mortality during Wave 3 reflects a critical period when the healthcare system was overwhelmed and vaccination efforts were still ramping up. The subsequent decline in mortality in Waves 4, 5, and 6, despite continued case fluctuations, underscores the effectiveness of vaccines, improved treatments, and increased immunity. These trends illustrate the transition from a high-mortality pandemic phase to a more controlled but persistent presence of the virus, influenced by policy decisions and population behavior

Q2

The results reflect how geographic, demographic, and systemic factors influenced the distribution of COVID-19 death rates across different regions and waves. In early waves, densely populated Northeastern states experienced the highest death rates, likely due to the rapid initial spread of the virus and overwhelmed healthcare systems. By Wave 2, the focus shifted to the South, where structural healthcare disparities and resistance to public health measures may have exacerbated outcomes. Wave 3 marked a turning point, with Midwest and Southern states bearing the brunt, potentially due to a combination of surging variants, winter conditions favoring transmission, and vaccine rollout delays.

As vaccination efforts expanded and treatment protocols improved, later waves showed a notable shift. Death rates in traditionally high-risk regions like the Northeast declined significantly, suggesting the impact of better healthcare access and higher vaccine uptake. Conversely, states in the South, Midwest, and Mountain West continued to struggle in Waves 4 through 6, where structural healthcare weaknesses and vaccine hesitancy likely played a role. Overall, the data underscores how public health infrastructure, demographic factors, and behavioral responses influenced regional outcomes, with states demonstrating varied capacities to adapt to the evolving challenges of the pandemic.

Q3

The decreasing fatality ratios across the six COVID-19 waves indicate that the virus became less virulent over time, with Wave 1 showing the highest fatality ratio (5.36%) and Wave 6 the lowest (0.42%). This trend likely reflects advancements in treatments, widespread vaccination, and increased population immunity. It may also suggest that later variants, while more transmissible, caused less severe disease. However, variations in testing rates, healthcare access, and reporting practices could have influenced these results. These findings emphasize the importance of continued public health efforts and research to understand and respond to evolving disease dynamics.

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.

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/