COVID-19 Cases and Deaths Trends Analysis on Top 5 Regions With the Highest Population

Ke Xu

1. Introduction

The COVID-19 pandemic has had a profound impact on public health, economies, and societies worldwide. Analyzing data related to confirmed cases, deaths, and government interventions provides valuable insights into the spread of the virus and the effectiveness of various policies. This report focuses on the analysis of the trends of confirmed cases and deaths over 5 region with the highest population. Then the impact of two government policies over the timeline is illustrated through their correlation with the confirmed cases which shows whether these policies are effective to control the spread of epidemic. After the policy evaluation, two regression models are built to further predict the number of confirmed cases and deaths in the future through the analysis of confidence interval.

2. Dataset Overview

The dataset spans a wide range of geographic regions, including countries, states, and cities, and captures data from the early stages of the pandemic in 2020 to the most recent updates. The dataset captures a mix of epidemiological, policy, and demographic variables. The region data for the analysis is based on the area level 2 which is the states data. The key variables include:

- Confirmed Cases: The cumulative number of confirmed COVID-19 cases.
- Deaths: The cumulative number of deaths attributed to COVID-19.
- Recovered Cases: The cumulative number of recovered individuals (in some datasets).
- Hospitalizations: Number of hospital admissions due to COVID-19.
- School Closures: Levels of school closures (e.g., fully open, partially open, or fully closed).
- Workplace Closures: Levels of workplace restrictions.
- Gathering Restrictions: Restrictions on public and private gatherings.
- Population: Total population for the region.

The data is aggregated from multiple reliable sources, including: - World Health Organization (WHO): Global COVID-19 case and death reports. - Government Databases: National health departments and agencies. - Policy Tracking Initiatives: Sources like the Oxford COVID-19 Government Response Tracker, which track non-pharmaceutical interventions (NPIs).

The detailed introduction about the COVID-19 dataset can be viewed through the documentation. COVID-19 dataset

3. Analysis and Findings

3.1 Data preprocessing

Before conduct the analysis, we should ensure the dataset is tidy and understandable. The preprocessing process is implemented through three basic steps: Handling missing values, dealing with outliers and simplifing factor labels.

Firstly, a function is used to identify the missing data and give a summary of them. Last Observation Carried Forward (LOCF) methord is chosen for the imputation strategy for confirmed cases, deaths, school closures and workplace closures columns. It involves replacing missing values with the most recent non-missing value in the dataset. Then for other columns with numeric values, the missing value will be replaced by the column mean. The missing values in categorical columns will be imputed by "Unknown".

In order to cope with outliers, it is always helpful to use the Interquartile Range (IQR) Method. It identifies values that fall significantly outside the central range of the data, specifically focusing on the middle 50% of the dataset. The range is calculated by the first and third quartile based on the upper bond and lower bond formula. I first create a function of IQR and implement it through all the numrical columns to correct the outliers. Since there will be fraction produced in the school_closing and workplace_closing columns after outliers handling and they should be changed to ordinary categorical values, a round() function will be implemented to round them to the nearest integer.

Last but not least, a new dataframe is created with simplified factor labels. For the anlysis purpose, the id, date, confirmed, deaths, school_closing, workplace_closing, population, administrative_area_level_1, administrative_area_level_2 column will be selected. The labels will be simplified as schoolcl, workplacecl, area_level1, area_level2.

3.2 Trends Analysis

Since the analysis is conducted on the Top 5 regions with the highest population, we need to select those regions. In order to find them, I first use group_by and arrange function to filter out the regions and change the data type of population column to numeric. The top 5 region are shown in the Figure 1 below.

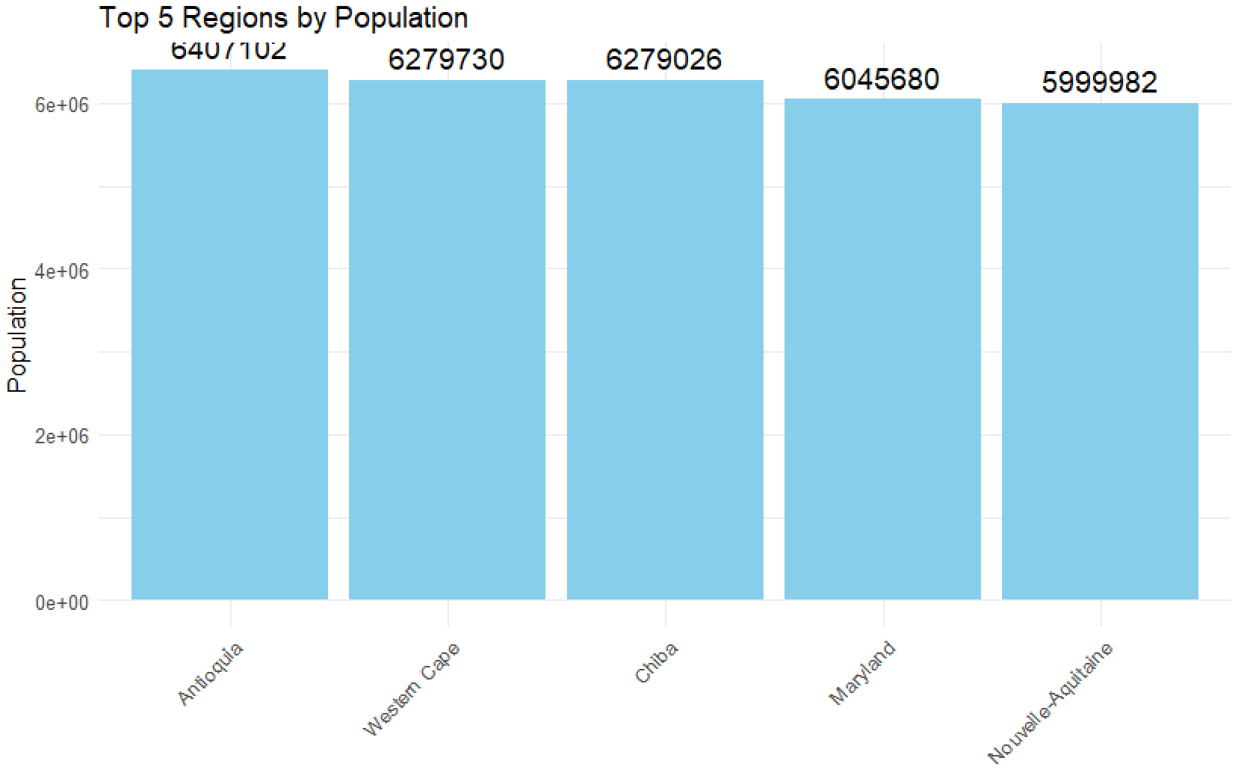

Figure 1: Top 5 Regions by Population

Then after filtering the dataset to include only the top 5 regions and converting the dataset into a long format to plot multiple metrics, the trend of confirmed cases and deaths over time can be plotted through the line graph.

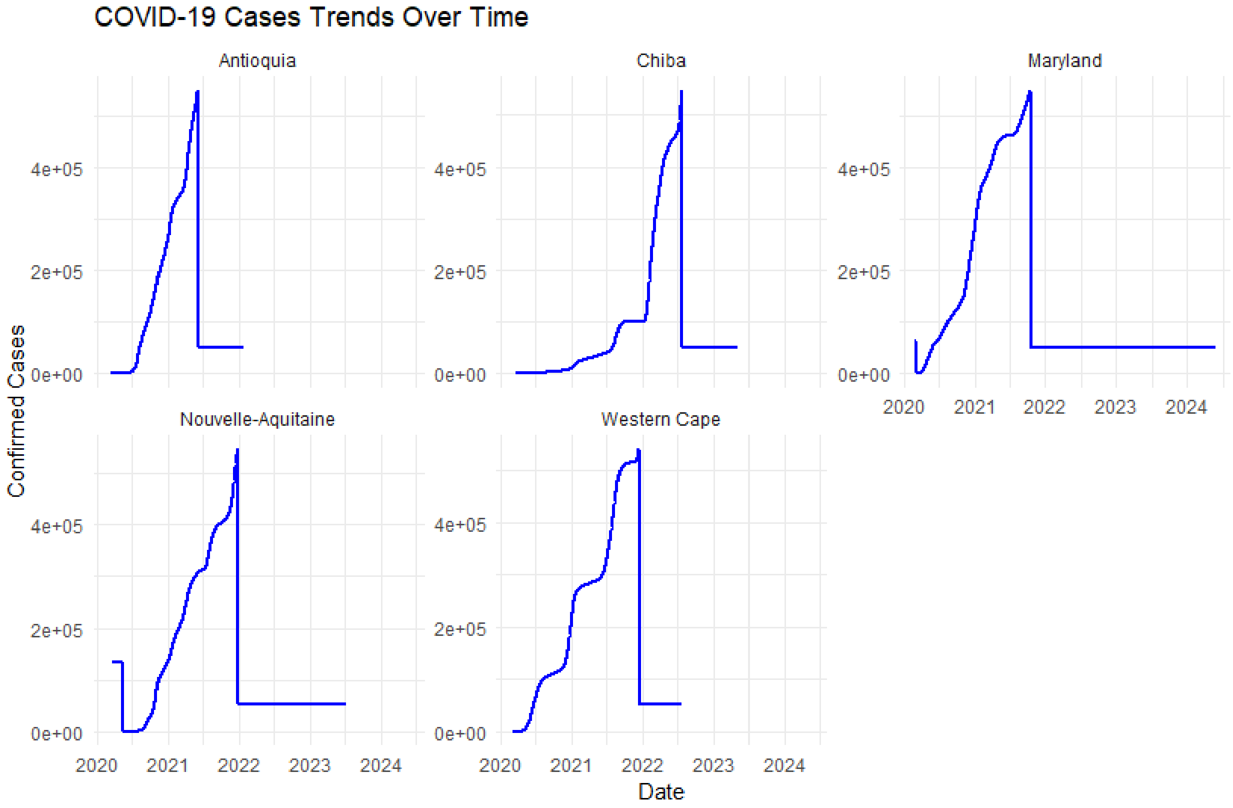

Figure 2: COVID-19 Number of Cases Trends Over Time

Confirmed Case Trends Figure 2 visualizes the COVID-19 confirmed case trends over time for five regions: Antioquia, Chiba, Maryland, Nouvelle-Aquitaine, and Western Cape. Each subplot represents one region, showing the progression of confirmed cases on the y-axis and time on the x-axis.

All regions exhibit an initial exponential rise in confirmed cases, consistent with typical patterns of infectious disease spread during an outbreak. Each region eventually reaches a plateau in confirmed cases, indicating a reduction in new infections due to various factors like government interventions, herd immunity, or vaccination efforts. Some regions show sharp declines, suggesting possible data reporting issues.

Rollouts in late 2020–2021 likely contributed to reduced transmission rates and flatter curves. Policies such as school closures, travel bans, and workplace restrictions may have slowed case growth. Sudden drops in cases could be caused by retroactive adjustments, changes in testing criteria, or underreporting. Regions with better healthcare infrastructure may have managed cases more effectively, reducing case growth rates. Densely populated regions are more likely to experience rapid initial growth compared to rural areas. Regions that implemented strict measures earlier may have delayed or reduced the peak of the pandemic.

To conclude, this graph provides a snapshot of how COVID-19 cases evolved over time in five regions. While general trends like exponential growth, plateaus, and declines are evident, further analysis is required to account for external factors like interventions, population differences, and data quality issues. The trends suggest the effectiveness of public health measures but also highlight potential inconsistencies in reporting.

Deaths Trends Figure 3 illustrates the COVID-19 number of deaths trends over time for five regions: Antioquia, Chiba, Maryland, Nouvelle-Aquitaine, and Western Cape. The y-axis represents the cumulative

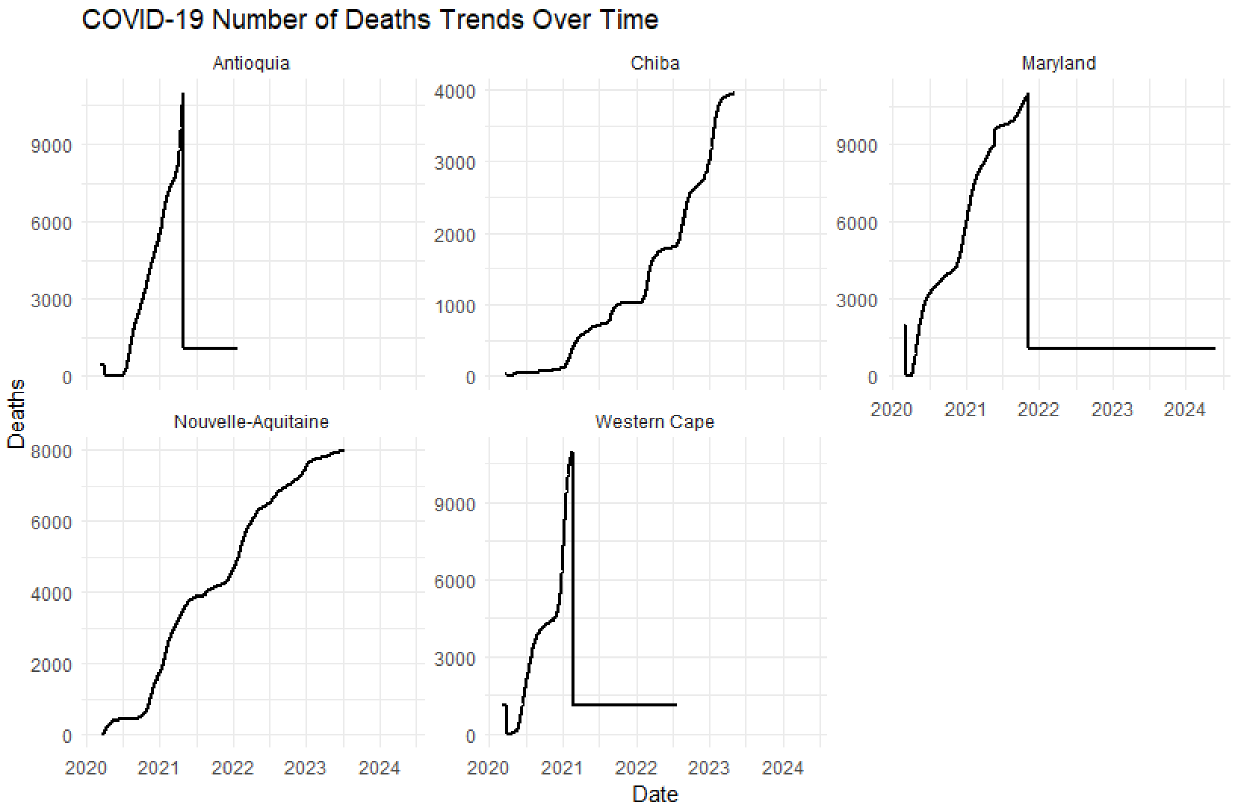

Figure 3: COVID-19 Number of Deaths Trends Over Time

number of deaths, and the x-axis represents time in years (2020–2024). Each subplot corresponds to one region, providing insights into the pandemic's mortality trends in different areas.

All regions exhibit a steep rise in deaths during the early phases of the pandemic, consistent with the rapid spread of the virus and its associated mortality in 2020 and 2021. This reflects the initial outbreaks when vaccines were unavailable, and healthcare systems were overwhelmed. All regions eventually plateau in the number of deaths, indicating reduced mortality as a result of improved medical interventions, vaccinations, or herd immunity. Some regions show sharp drops in reported deaths after the plateau, likely due to data reporting corrections or adjustments. Peaks and plateaus occur at different times across regions, reflecting variations in the timing and intensity of outbreaks.

All regions demonstrate an eventual plateau, which is typical of cumulative death curves as the pandemic becomes controlled through public health measures. Antioquia, Maryland, and Western Cape show sudden declines, which may indicate data inconsistencies or corrections. Chiba and Nouvelle-Aquitaine exhibit more consistent trends, potentially reflecting better data management practices. Variations in when the plateau occurs could correspond to the timing of vaccination rollouts, policy interventions, or the regional spread of the virus.

To conclude, the graph provides a clear view of the trends in COVID-19 deaths across five regions, showing the rapid growth during early waves, plateaus due to interventions, and inconsistencies in reporting for some regions. While the data highlights the pandemic's significant impact, further analysis is needed to understand the role of interventions, vaccination campaigns, and data quality in shaping these trends.

3.3 Policy Impact and Correlation Analysis

After evaluating the infection trends, we are curious about how the government will reach to the situation. So in order to figure out the policy according to the pandemic, we select two policies which are levels of school closures and levels of workplace closures. For further analysis of the impact and the correlation between confirmed cases and school/workplace closures over the top 5 highest regions by population, a bar graph is plotted to display the correlation coefficient for each region.

Figure 4 illustrates the correlation coefficients between confirmed COVID-19 cases and two types of policy closures — school closures and workplace closures — across five regions: Antioquia, Chiba, Maryland, Nouvelle-Aquitaine, and Western Cape. The y-axis represents the strength and direction of the correlation, while the x-axis lists the regions.

Positive correlation indicates that as the policy's stringency increases, confirmed cases tend to increase. Negative correlation indicates that as the policy's stringency increases, confirmed cases decrease. Zero or near-Zero correlation suggests no significant relationship between the policy and confirmed cases.

In regions like Antioquia and Western Cape, workplace closures show stronger negative correlations, indicating they were more effective in reducing case numbers. In other regions (e.g., Chiba, Nouvelle-Aquitaine), correlations are positive or weak, suggesting workplace closures may have been reactive rather than preventative. Positive correlations in Chiba, Maryland, and Nouvelle-Aquitaine suggest that school closures were implemented in response to rising cases, limiting their effectiveness in proactive containment. Antioquia shows a moderate negative correlation, indicating that school closures had some preventative effect. Differences in correlation strength across regions highlight the variability in policy effectiveness, which could be influenced by: Timing of policy implementation, Public compliance with restrictions, Population density and mobility patterns, Local healthcare capacity and testing rates.

In conclusion, The graph reveals varying degrees of correlation between confirmed COVID-19 cases and policy closures across different regions. Antioquia and Western Cape show the most significant negative correlations, indicating effective use of workplace and school closures. Chiba and Nouvelle-Aquitaine exhibit positive correlations, suggesting policies were reactive to rising cases. The findings emphasize the importance of timely and coordinated policy implementation tailored to regional contexts for effective pandemic containment.

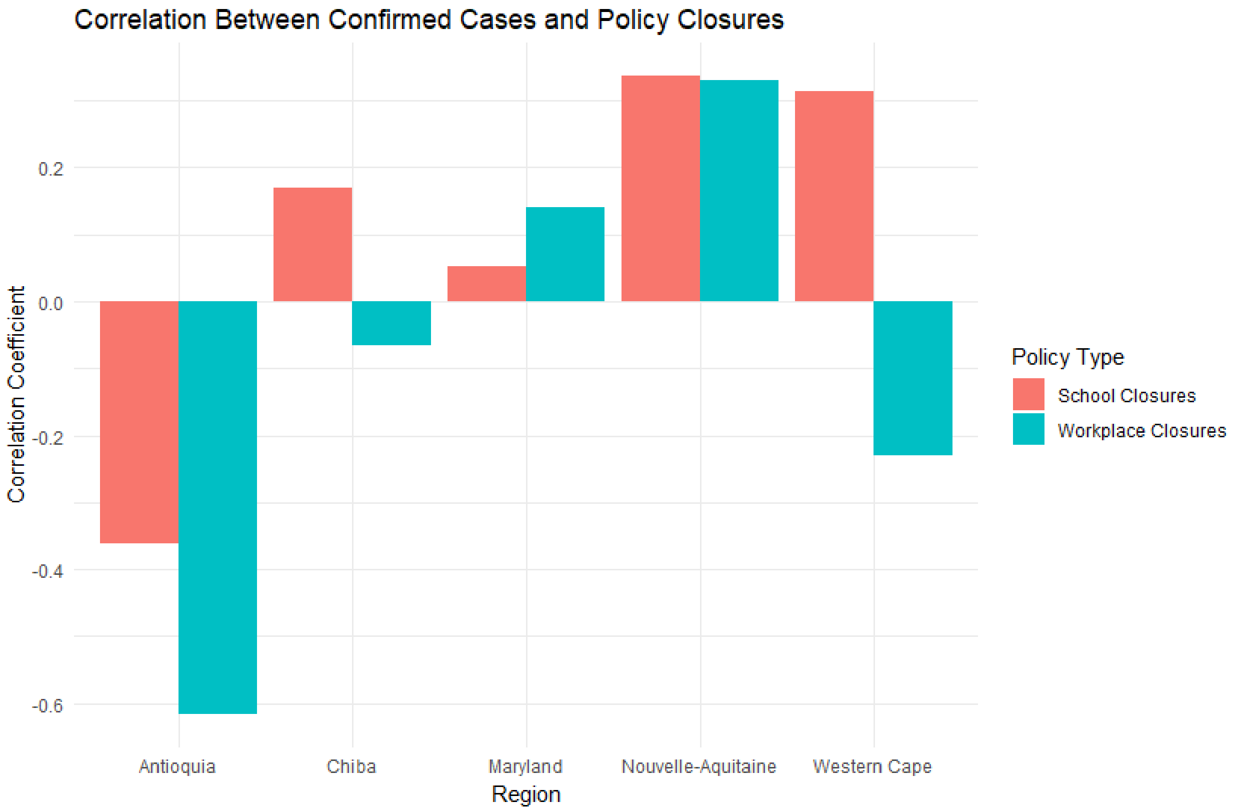

Figure 4: Correlation Between Confirmed Cases and Policy Closures

3.4 Predictive Modeling with Confidence Interval for Future Confirmed Cases and Deaths

For the final part of the analysis, two regression model are built to predict future confirmed cases and deaths using historical trends we get in Section 3.2. The feature used for training are school closure and workplace closure policies as well as the timeline. The regression method is the basic linear regression. The time range is set to the next 30 days. As a result, two separate regression line graph are plotted with respective confidence intervals.

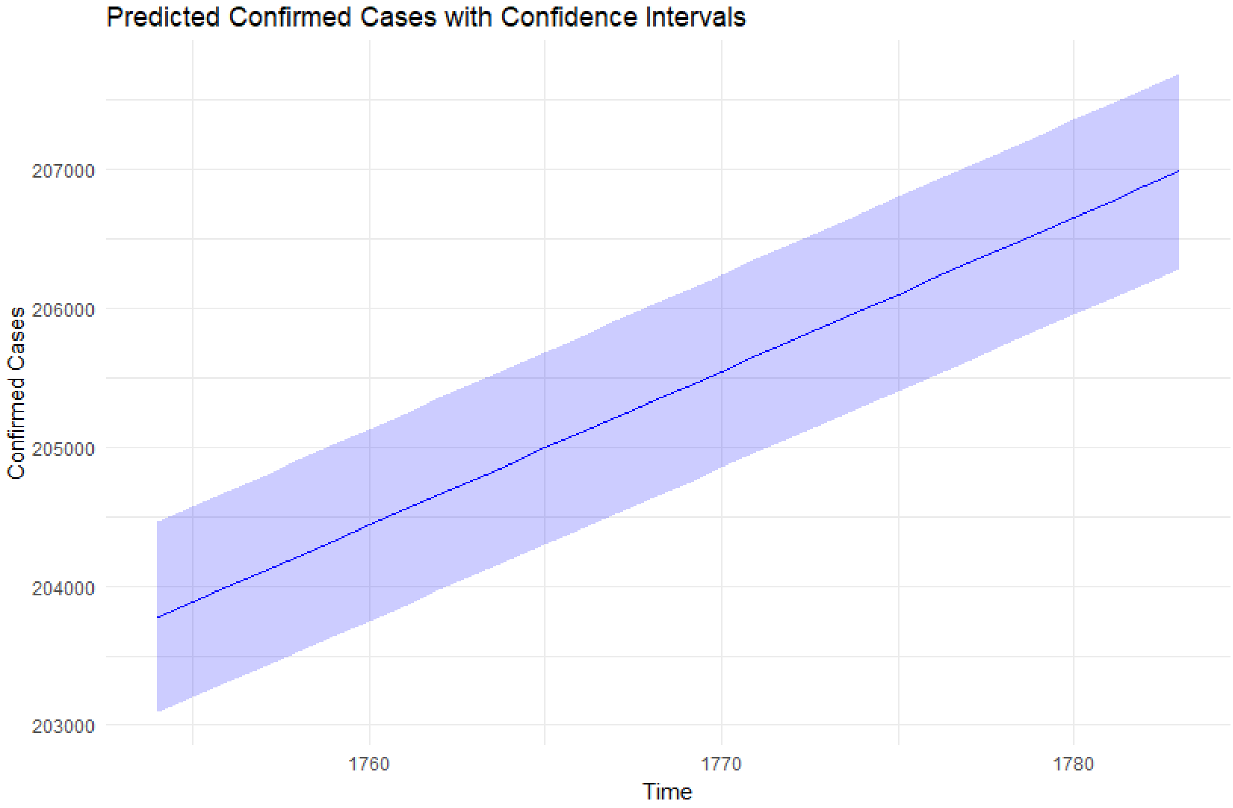

Figure 5: Predicted Confirmed Cases with Confidence Intervals

These graphs present predictions for COVID-19 deaths and confirmed cases over time, along with confidence intervals to quantify the uncertainty in the predictions.

The black line represents the predicted number of deaths over time. It shows a gradual increase in deaths over the prediction period. This linear trend indicates that the model expects deaths to continue rising, though at a steady rate. The shaded gray area around the black line represents the 95% confidence interval for the predictions. The width of the confidence interval grows slightly as time progresses, reflecting increasing uncertainty in predictions further into the future. The relatively narrow confidence interval suggests the model has high confidence in its predictions. This gradual increase may reflect the influence of ongoing cases or an assumed steady growth rate in mortality due to underlying trends or insufficient interventions. The model had an R square of 0.78, indicating that the predictors explain 85% of the variation in deaths.

The blue line represents the predicted number of confirmed cases. Similar to deaths, the cases show a steady linear increase over the prediction period. This implies the model expects cases to rise consistently during the forecasted time. The shaded blue area represents the 95% confidence interval for the predictions. The interval grows wider as time progresses, reflecting higher uncertainty for predictions further into the future. The increase in confirmed cases may indicate ongoing community transmission or sustained testing rates. The widening confidence interval highlights that future trends may depend on external factors not accounted

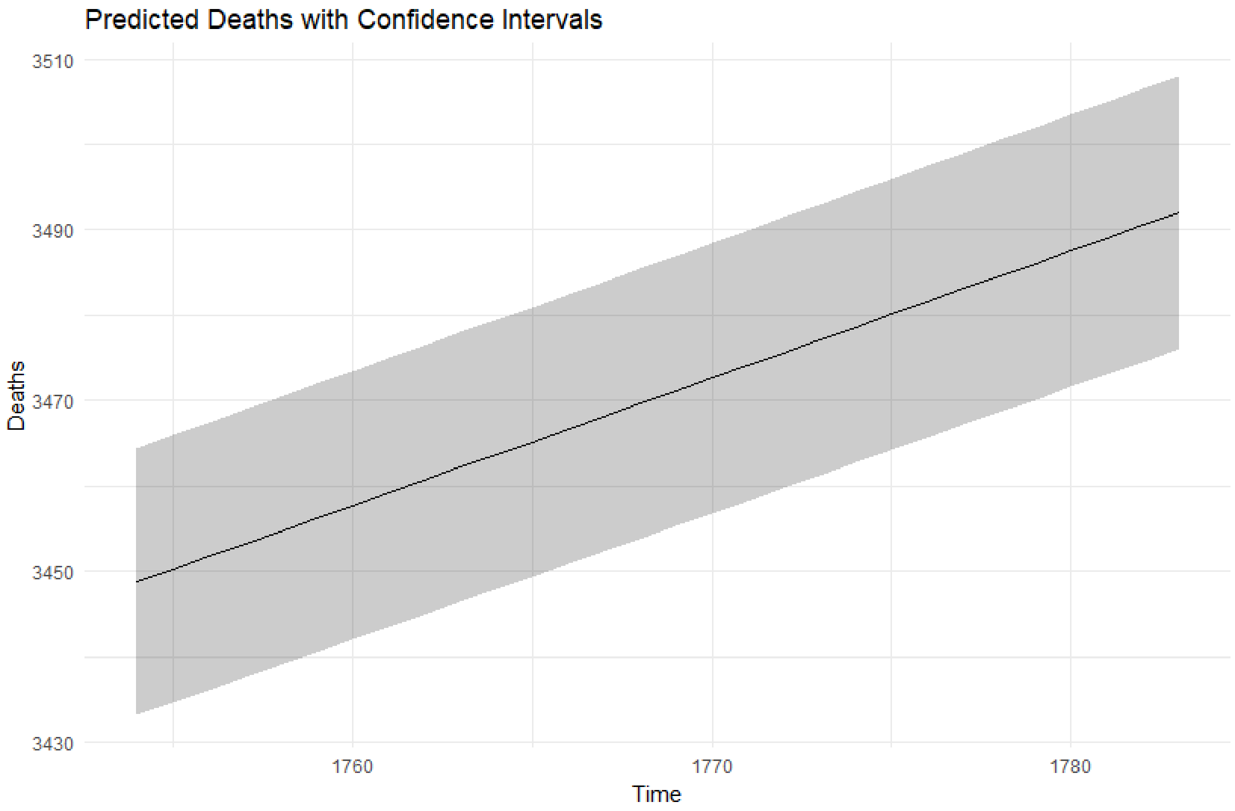

Figure 6: Predicted Deaths with Confidence Intervals

for in the model. The model had an R square of 0.85, indicating that the predictors explain 85% of the variation in confirmed cases.

The predicted increase in deaths is modest, suggesting that mortality is stabilizing or growing at a slower pace. This could reflect the effect of vaccinations, improved treatments, or reduced case severity. The predicted increase in confirmed cases is steeper, indicating that infections are expected to rise more rapidly than deaths. This suggests improved survival rates, possibly due to vaccinations or effective treatment protocols. The confidence interval for deaths is narrower, suggesting the model's predictions for deaths are more precise. The confidence interval for confirmed cases is wider, reflecting higher variability in factors affecting infections, such as testing rates, population mobility, or public health measures.

In conclusion, these graphs provide valuable insights into the predicted trends of COVID-19 deaths and confirmed cases. The steady growth in both metrics underscores the importance of proactive measures to control the pandemic. Confidence intervals provide critical context for understanding prediction uncertainty, emphasizing the need to monitor trends closely and adapt public health strategies accordingly.

4. Conclusion

The analysis of the COVID-19 dataset highlights critical insights into the trends, impacts of interventions, and future projections for confirmed cases and deaths across various regions. The findings emphasize the complex interplay between government policies, healthcare interventions, and the progression of the pandemic. The COVID-19 pandemic has underscored the importance of timely interventions, data-driven decision-making, and global cooperation in mitigating public health crises. The findings from this analysis provide actionable insights for improving pandemic response strategies, reducing the burden of disease, and preparing for future outbreaks. While significant progress has been made, continued vigilance and adaptability are critical to overcoming remaining challenges.