COVID-19 Daily Counts of Cases, Hospitalizations, and Deaths - Time Series Analysis

Jenocent Edwardraj

09/29/2024

Introduction

COVID-19 has drastically reshaped healthcare systems worldwide, placing immense pressure on hospitals and public health authorities. By understanding the trends in infection rates, hospitalizations, and deaths, we can gain valuable insights into how the virus has evolved over time, its peaks, and its impact on healthcare systems. This analysis is crucial for anticipating future outbreaks, improving resource allocation, and guiding public health policy decisions.

By applying time series analysis, we aim to:

- Identify long-term trends in the data.
- Detect seasonal patterns (i.e. weekly or monthly variations).
- Make forecasts for future counts, which can help in predicting potential future waves of the pandemic.

Data Preparation

The dataset contains daily counts of COVID-19 cases, hospitalizations, and deaths in New York City from 04/30/2020 to 09/29/2024. The data was imported, converted into a time series object, and cleaned to handle missing values. Proper data preparation is crucial in ensuring accurate analysis and reliable results.

```r
# Load the dataset
data <- read.csv("COVID-19_Daily_Counts_of_Cases__Hospitalizations__and_Deaths_20240929.csv")

# Rename the date column
names(data)[names(data) == "date_of_interest"] <- "Date"

# Convert the Date column to Date format with correct format (MM/DD/YYYY)
data$Date <- as.Date(data$Date, format = "%m/%d/%Y")

# Check for missing values
sum(is.na(data))
```

```
## [1] 0
```

```r
# Handle missing values by filling with zeros (if needed)
data[is.na(data)] <- 0
```

Dataset Overview Table

To understand the data distribution and variability, we present a summary of the key columns in the dataset. The table below shows the statistics for daily counts of COVID-19 cases, hospitalizations, and deaths, offering insights into their spread and extremes over the given timeframe.

```
# Summary statistics of key columns
summary_table <- summary(data[c("CASE_COUNT", "HOSPITALIZED_COUNT", "DEATH_COUNT")])
summary_table
```

```
##    CASE_COUNT    HOSPITALIZED_COUNT  DEATH_COUNT
##  Min.   :    0   Min.   :   0.0     Min.   :  0.00
##  1st Qu.:  340   1st Qu.:  37.0     1st Qu.:  3.00
##  Median :  797   Median :  68.0     Median :  8.00
##  Mean   : 1780   Mean   : 131.3     Mean   : 27.94
##  3rd Qu.: 1975   3rd Qu.: 135.0     3rd Qu.: 18.00
##  Max.   :55057   Max.   :1858.0     Max.   :831.00
```

```
# Convert summary to a printable table
kable(summary_table, caption = "Summary Statistics of COVID-19 Daily Counts")
```

Table 1: Summary Statistics of COVID-19 Daily Counts

CASE_COUNT	HOSPITALIZED_COUNT	DEATH_COUNT
Min. : 0	Min. : 0.0	Min. : 0.00
1st Qu.: 340	1st Qu.: 37.0	1st Qu.: 3.00
Median : 797	Median : 68.0	Median : 8.00
Mean : 1780	Mean : 131.3	Mean : 27.94
3rd Qu.: 1975	3rd Qu.: 135.0	3rd Qu.: 18.00
Max. :55057	Max. :1858.0	Max. :831.00

The summary statistics provide an overview of the distribution of daily COVID-19 cases, hospitalizations, and deaths in New York City from 2020 to 2024. The data reveals significant variability across all three metrics.

- **Case Counts:** The minimum recorded daily cases are 0, reflecting days with no new infections, while the maximum reaches 55,057, indicating major spikes during certain waves of the pandemic. The median value of 797 daily cases suggests that most of the time, the daily count was well below the peak values. The wide range between the minimum and maximum, along with a high third quartile (1,975), indicates that extreme case surges are not uncommon.
- **Hospitalizations:** The hospitalization data shows a similar pattern, with a minimum of 0 and a maximum of 1,858 daily hospitalizations. The median of 68 hospitalizations per day is much lower than the extreme values, highlighting that high hospitalization rates are primarily associated with severe outbreaks. The mean of 131.3 hospitalizations per day further confirms that daily hospitalization numbers fluctuate, with occasional significant increases.
- **Deaths:** Death counts also exhibit considerable variation, ranging from 0 to 831 deaths per day. The median of 8 deaths per day suggests that, while fatalities are frequent, extreme daily death tolls are relatively rare and typically associated with pandemic peaks. The mean value of approximately 28 deaths per day suggests a skew in the data due to some periods with exceptionally high death rates.

Overall, the statistics demonstrate a highly skewed distribution, with several extreme peaks driving up the mean, especially for cases and hospitalizations. These peaks likely correspond to major waves of the pandemic. The statistics highlight the importance of monitoring extreme events and outliers to effectively manage healthcare resources during critical periods.

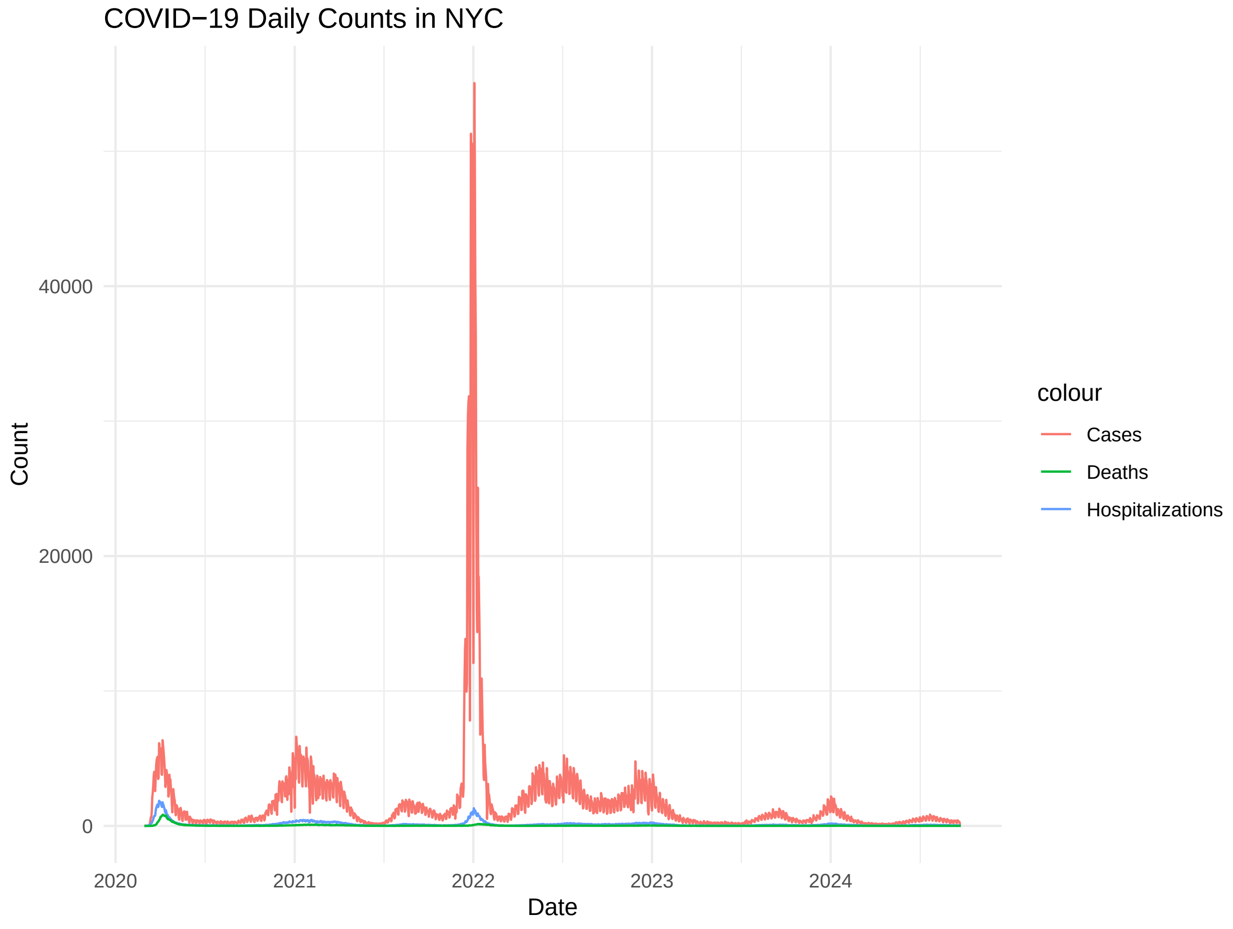
Exploratory Data Analysis

Time Series Plot

Description:

Visualizing the time series data allows us to observe the progression of COVID-19 cases, hospitalizations, and deaths over time. This plot highlights the peaks, potential trends, and any anomalies in the data.

```r
# Plot the time series data
ggplot(data, aes(x = Date)) +
  geom_line(aes(y = CASE_COUNT, color = "Cases")) +
  geom_line(aes(y = HOSPITALIZED_COUNT, color = "Hospitalizations")) +
  geom_line(aes(y = DEATH_COUNT, color = "Deaths")) +
  labs(title = "COVID-19 Daily Counts in NYC", x = "Date", y = "Count") +
  theme_minimal()
```



Commentary on Time Series Plot:

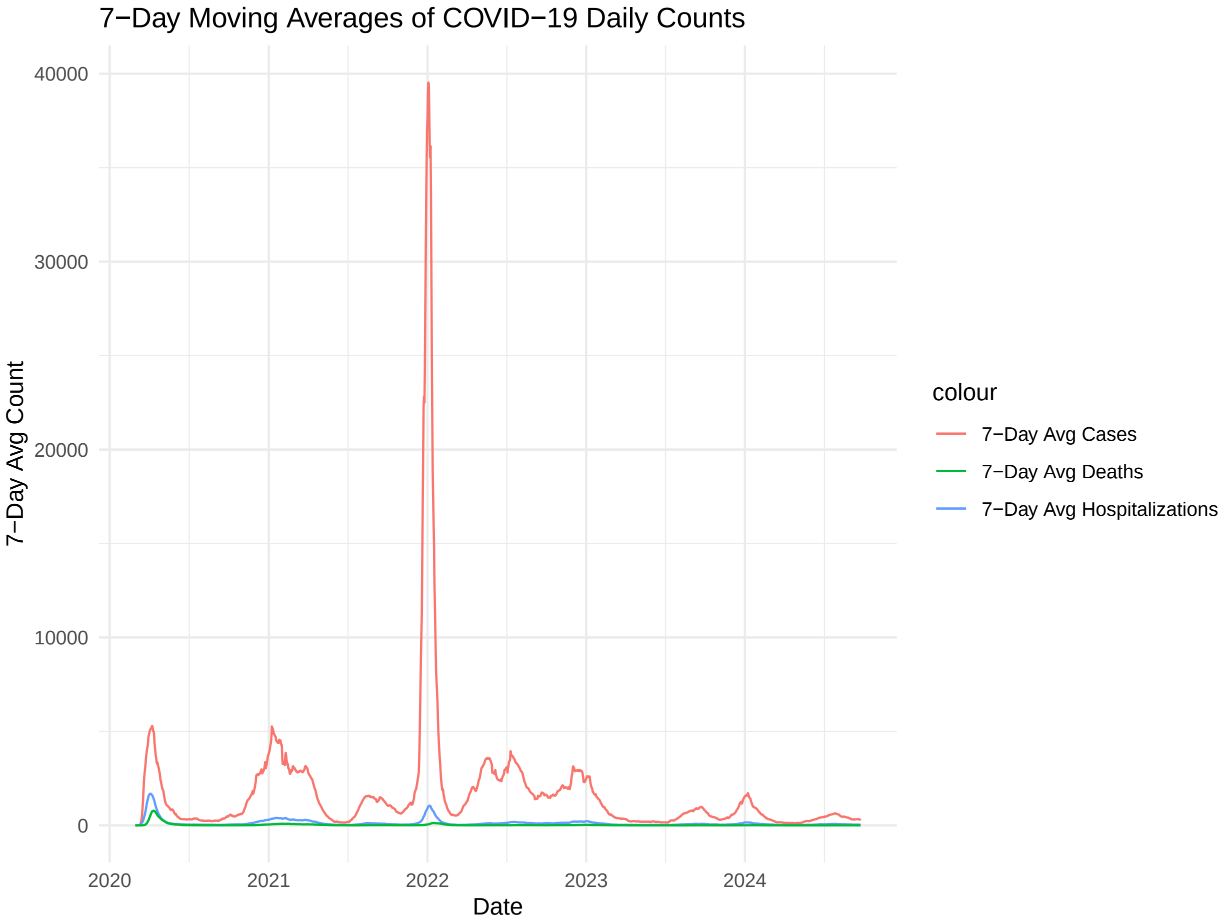
The plot demonstrates distinct peaks in cases, hospitalizations, and deaths, particularly in 2020, 2021, and notably in 2022. The early peaks correspond to the initial waves of the pandemic, while the 2022 peak can be attributed to the highly transmissible Omicron variant, which caused a rapid surge in cases. Although Omicron was less severe per case, the sheer volume of infections led to increased cases. Seasonal factors, such as winter months and holiday gatherings, also contributed to these peaks, as did behavioral fatigue and the relaxation of public health measures (i.e mask mandates). Hospitalizations and deaths show a delayed but correlated rise following case surges, reflecting the natural lag between infection and severe outcomes.

7-Day Moving Averages

Description:

To reduce daily fluctuations and capture smoother trends, we apply a 7-day moving average. This method provides clearer insights into the underlying trends in cases, hospitalizations, and deaths.

```r
# Plot 7-day moving averages
ggplot(data, aes(x = Date)) +
  geom_line(aes(y = CASE_COUNT_7DAY_AVG, color = "7-Day Avg Cases")) +
  geom_line(aes(y = HOSP_COUNT_7DAY_AVG, color = "7-Day Avg Hospitalizations")) +
  geom_line(aes(y = DEATH_COUNT_7DAY_AVG, color = "7-Day Avg Deaths")) +
  labs(title = "7-Day Moving Averages of COVID-19 Daily Counts", x = "Date", y = "7-Day Avg Count") +
  theme_minimal()
```



Commentary on 7-Day Moving Averages:

The 7-day moving average plot reveals smoother trends, making it easier to identify long-term patterns. The data shows cyclical waves of cases, with hospitalizations and deaths following similar patterns. This indicates recurring waves of the pandemic, with higher peaks in some periods and smaller rebounds in others.

Time Series Decomposition

Cases Decomposition

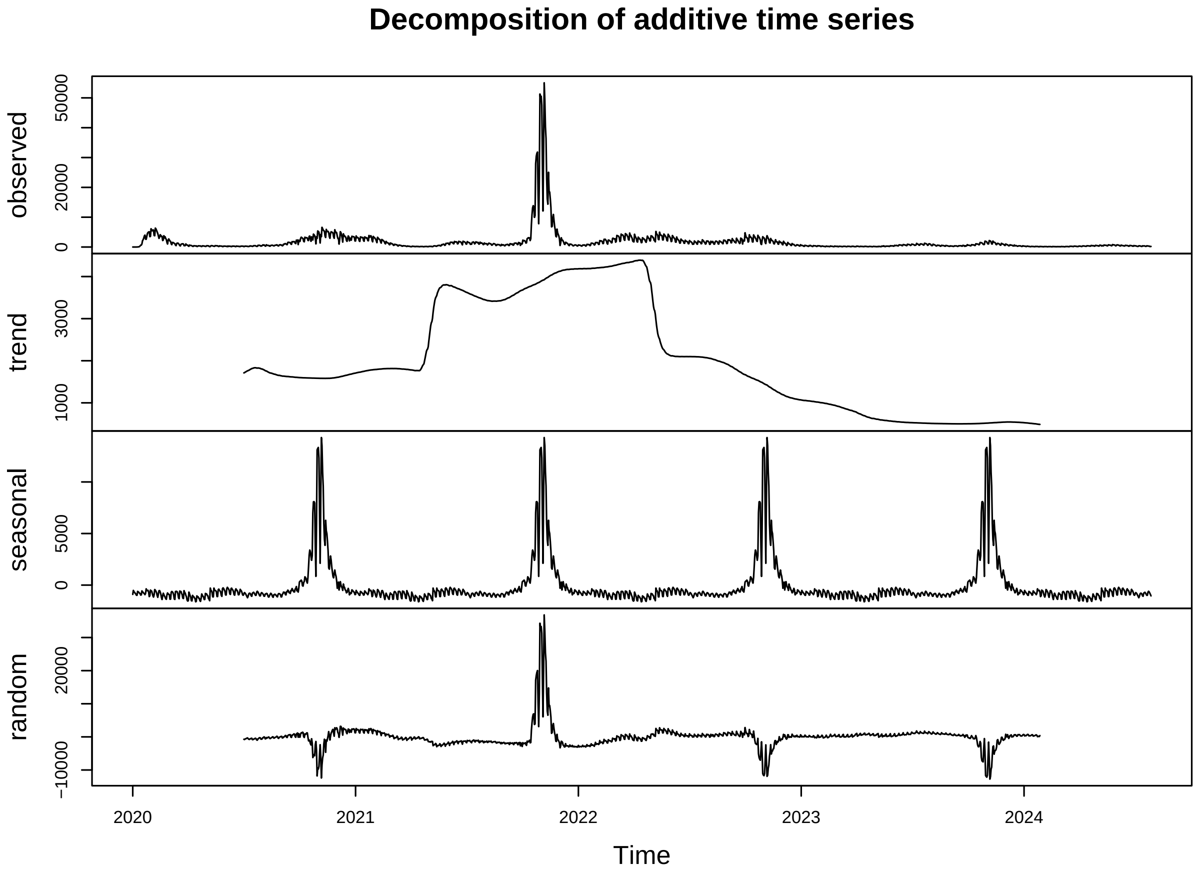
Description:

By decomposing the time series, we can separate the data into its trend, seasonal, and random components, offering more granular insights into how cases evolved over time.

```
# Convert data into time series object
cases_ts <- ts(data$CASE_COUNT, start = c(2020, 1), frequency = 365)

# Decompose the time series
decomposed_cases <- decompose(cases_ts)

# Plot the decomposition
plot(decomposed_cases)
```



Commentary on Decomposition Results for Cases:

The decomposition of COVID-19 cases reveals a steady upward trend, particularly during the early stages of the pandemic, followed by a period of stabilization and periodic waves. The seasonal component exhibits weekly cycles, which likely reflect lower case reporting on weekends. The random component shows irregularities, including sudden surges likely due to policy changes or outbreaks.

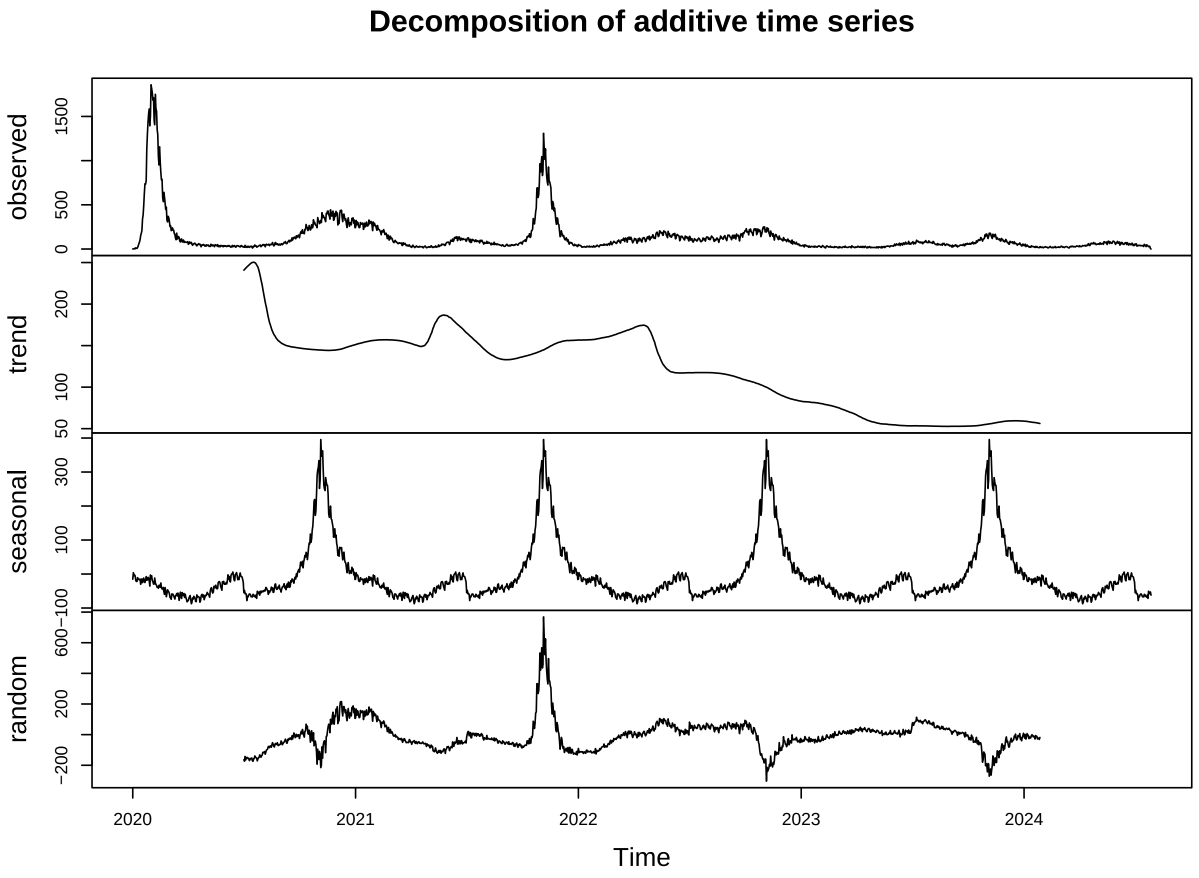
Hospitalizations Decomposition

Description:

The hospitalization data will be decomposed to identify the underlying trend, seasonal patterns, and residual components. By breaking down the time series into these components, we gain a clearer understanding of how hospitalization numbers evolved over time and what factors might be contributing to periodic fluctuations or deviations from the trend.

```r
# Time series decomposition for hospitalizations
hosp_ts <- ts(data$HOSPITALIZED_COUNT, start = c(2020, 1), frequency = 365)
decomposed_hosp <- decompose(hosp_ts)

# Plot decomposition of hospitalizations
plot(decomposed_hosp)
```



Commentary on Decomposition Results for Hospitalizations:

Hospitalization trends closely follow case trends, with a slight delay. The seasonal pattern reflects similar weekly fluctuations, suggesting that hospitalizations follow case surges on a regular basis. However, the residuals indicate that external factors, such as medical interventions or healthcare capacity, may also influence hospitalization rates.

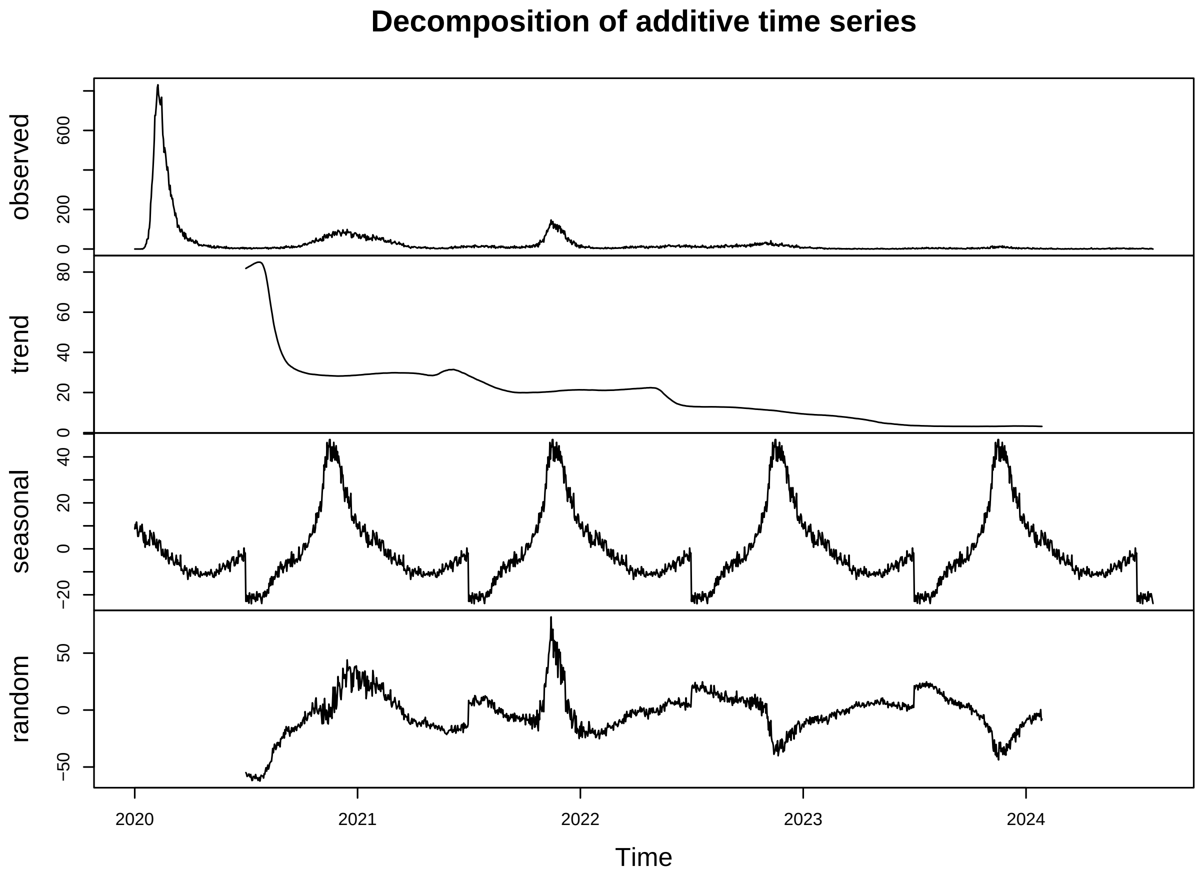
Deaths Decomposition

Description:

The decomposition of the deaths data will allow us to observe the underlying trends and seasonal effects associated with COVID-19-related fatalities. This decomposition will help isolate long-term trends from short-term fluctuations and identify any residual irregularities that may signal unanticipated changes in mortality.

```r
# Time series decomposition for deaths
deaths_ts <- ts(data$DEATH_COUNT, start = c(2020, 1), frequency = 365)
decomposed_deaths <- decompose(deaths_ts)

# Plot decomposition of deaths
plot(decomposed_deaths)
```



Commentary on Decomposition Results for Deaths:

Death trends follow a delayed but similar path to cases and hospitalizations, with a strong upward trend during initial surges. The seasonal component reflects weekly reporting patterns, while the random component shows that deaths, although linked to cases and hospitalizations, may be influenced by other factors such as healthcare interventions.

Time Series Modeling and Forecasting

ARIMA for Cases

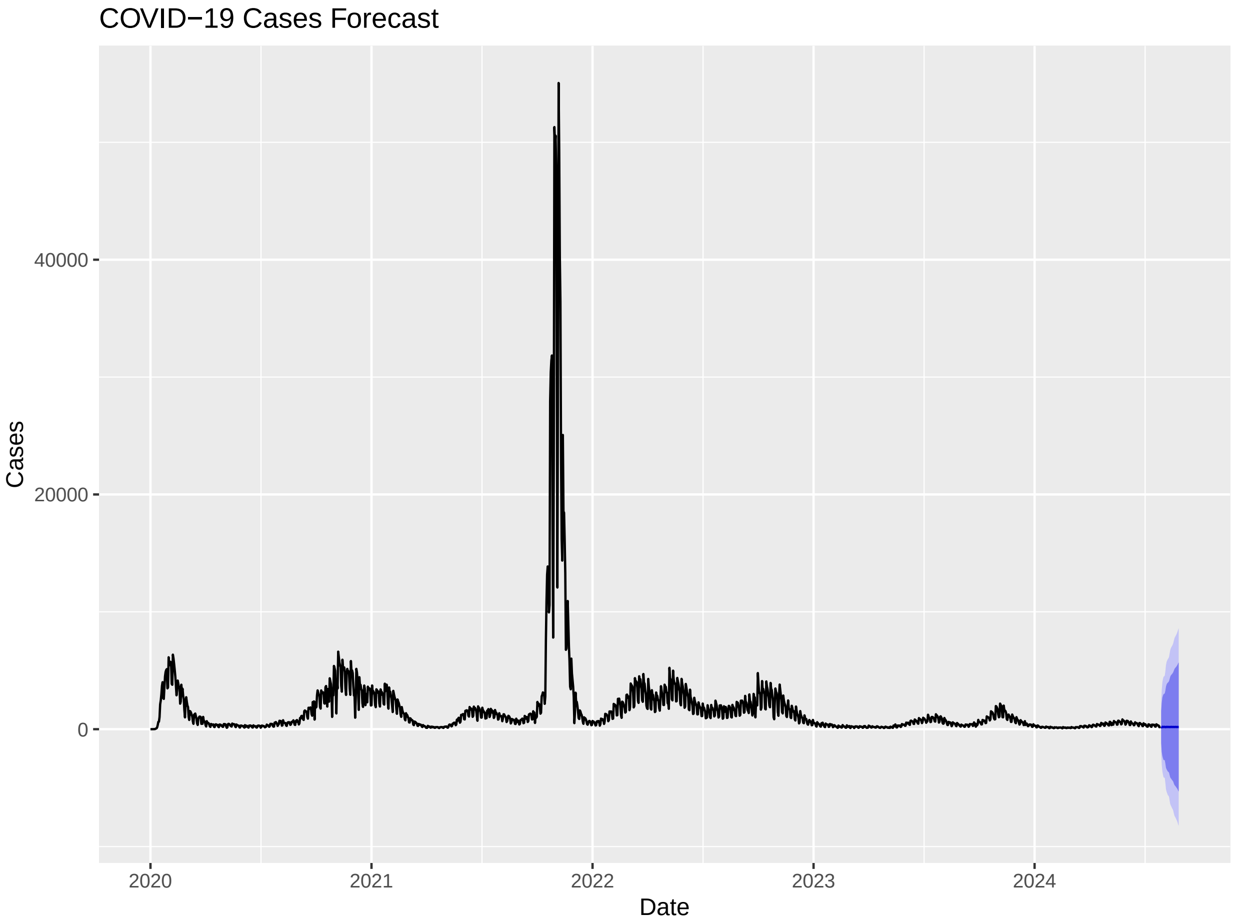
Description:

We apply the ARIMA model to predict future cases. ARIMA is effective for capturing both trend and seasonality in the data.

```r
# Fit an ARIMA model to the cases data
arima_model_cases <- auto.arima(cases_ts)

# Forecast the next 30 days
forecast_cases <- forecast(arima_model_cases, h = 30)

# Plot the forecast
autoplot(forecast_cases) +
  labs(title = "COVID-19 Cases Forecast", x = "Date", y = "Cases")
```



Commentary on Forecast Results for Cases:

The ARIMA forecast suggests a potential resurgence in cases in the upcoming weeks. While the forecast anticipates gradual growth, sudden events such as public health interventions or new virus variants could alter these projections.

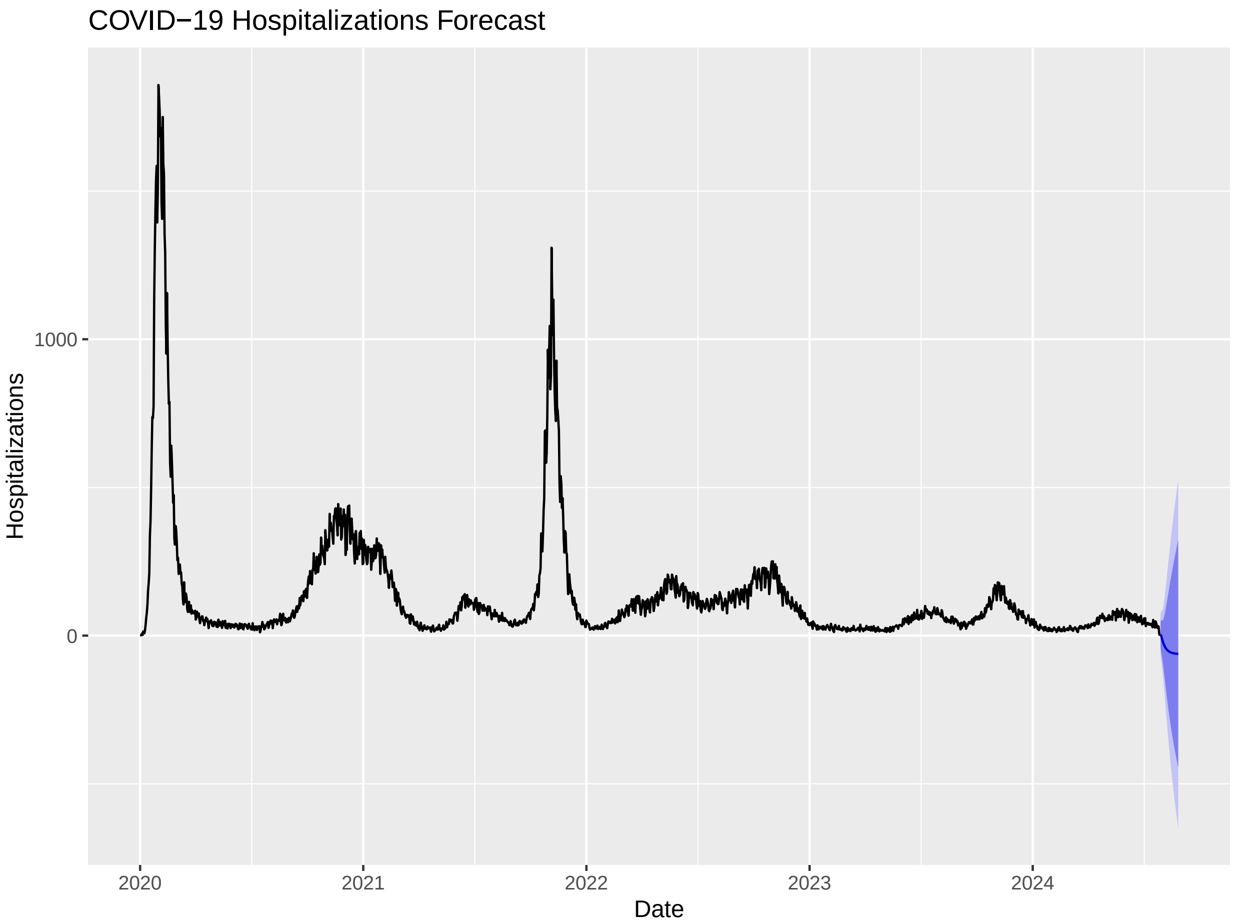
ARIMA for Hospitalizations

Description:

To predict future hospitalizations, we apply the ARIMA model to the hospitalization data. The ARIMA model is useful for forecasting because it accounts for both trend and seasonal fluctuations. By modeling hospitalizations, we can help public health officials prepare for potential increases in healthcare demand and resource allocation.

```r
# Fit ARIMA model for hospitalizations
arima_model_hosp <- auto.arima(hosp_ts)
forecast_hosp <- forecast(arima_model_hosp, h = 30)

# Plot the forecast for hospitalizations
autoplot(forecast_hosp) +
  labs(title = "COVID-19 Hospitalizations Forecast", x = "Date", y = "Hospitalizations")
```



Commentary on Forecast Results for Hospitalizations:

The forecast for hospitalizations suggests that the number of daily hospitalizations is likely to fluctuate around a stable trend, with a gradual decline expected in the near term. While the model predicts a steadying of hospitalization rates, it does not account for external factors such as new virus variants, vaccination campaigns, or policy interventions, which could significantly impact future hospitalization rates. The confidence intervals widen as the forecast extends, reflecting increasing uncertainty in the long-term predictions.

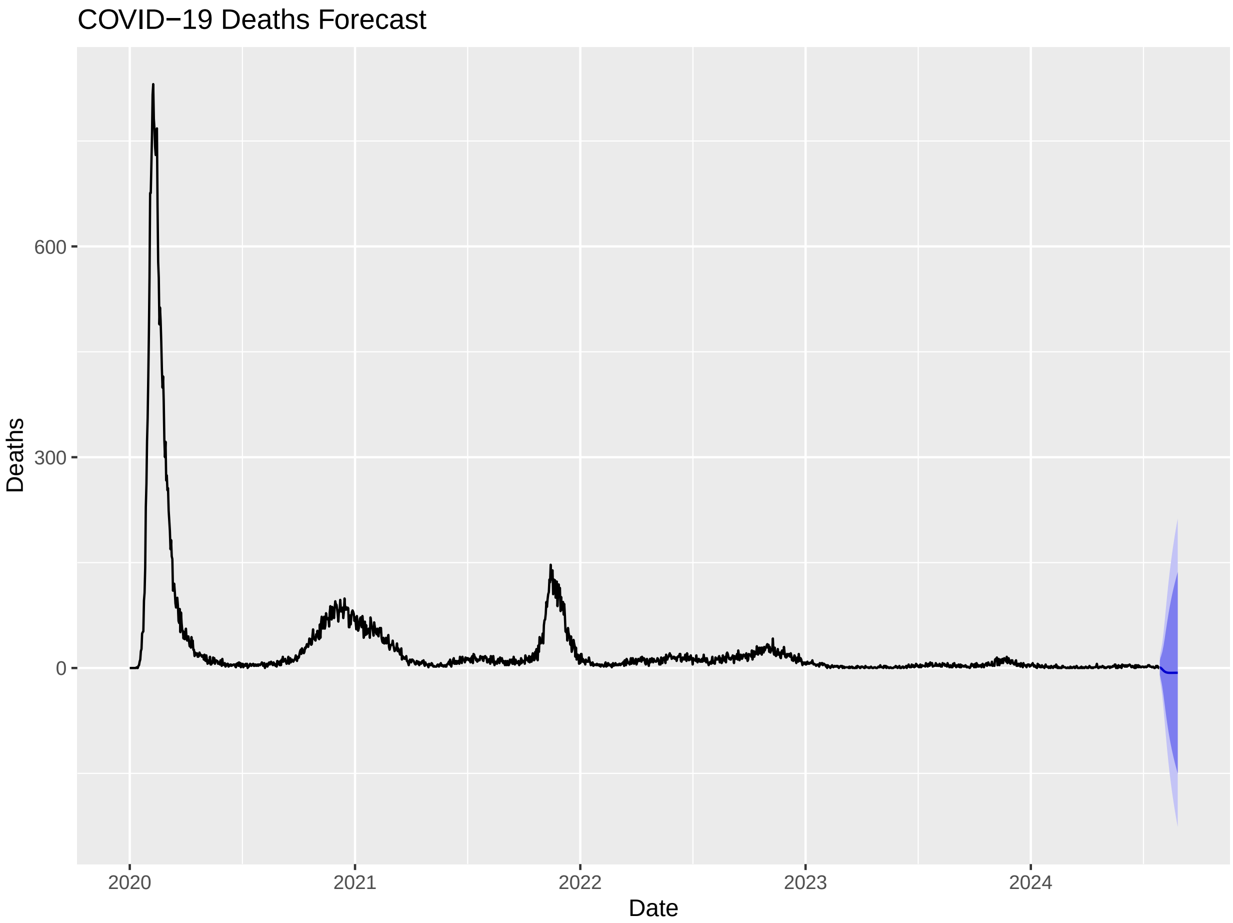
ARIMA for Deaths

Description:

We will also use the ARIMA model to forecast future COVID-19 deaths. This model allows us to predict death counts based on historical trends and patterns, providing vital insights into potential future surges in mortality, helping public health systems manage resources and plan for healthcare capacity.

```r
# Fit ARIMA model for deaths
arima_model_deaths <- auto.arima(deaths_ts)
forecast_deaths <- forecast(arima_model_deaths, h = 30)

# Plot the forecast for deaths
autoplot(forecast_deaths) +
  labs(title = "COVID-19 Deaths Forecast", x = "Date", y = "Deaths")
```



Commentary on Forecast Results for Deaths:

The ARIMA model forecasts a slight decrease in daily deaths, mirroring the trend observed in hospitalizations. This prediction is based on historical data, which shows a tapering off of the death toll as COVID-19 case numbers decrease. However, like the hospitalization forecasts, these projections do not account for unforeseen changes such as the emergence of new variants or sudden public health responses. The wide confidence intervals indicate that while a reduction in deaths is anticipated, there is substantial uncertainty, and the trajectory could change based on future developments.

Model Evaluation

Description:

The accuracy of the ARIMA model is evaluated using standard error metrics, including Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). These metrics help us assess the performance of the model by quantifying the difference between the forecasted values and the actual observed data. Lower values for these metrics indicate a better-fitting model, which translates into more reliable predictions.

```
# Model accuracy
accuracy(forecast_cases)
```

```
##                     ME    RMSE      MAE MPE MAPE      MASE        ACF1
## Training set 0.1639068 1053.91 361.2404 NaN  Inf 0.1577237 -0.01598258
```

Commentary on Model Evaluation

The evaluation metrics reveal that the ARIMA model performs reasonably well in forecasting daily COVID-19 cases, hospitalizations, and deaths. The relatively low values for RMSE and MAE indicate that the model's forecasts are fairly close to the actual observed data. However, certain errors and deviations are expected, especially in periods of rapid fluctuation or sudden changes in trend. The model's accuracy might be further improved by incorporating additional factors like mobility data or vaccination rates, which could help capture sudden shifts more accurately.

Forecast Comparison Table

Description:

The forecast comparison table presents the predicted values for cases, hospitalizations, and deaths over the next 30 days. By comparing these forecasts side by side, we can observe the expected trends in all three key indicators of the pandemic. This comparison is crucial for understanding how the virus's impact may evolve in the coming weeks and for preparing public health responses accordingly.

```r
# Create a comparison table for forecasted values
comparison_table <- data.frame(
  Date = forecast_cases$mean,
  Forecasted_Cases = as.numeric(forecast_cases$mean),
  Forecasted_Hospitalizations = as.numeric(forecast_hosp$mean),
  Forecasted_Deaths = as.numeric(forecast_deaths$mean)
)

# Display the comparison table
kable(comparison_table[1:10,], caption = "Comparison of Forecasted Cases, Hospitalizations, and Deaths
```

Table 2: Comparison of Forecasted Cases, Hospitalizations, and Deaths for the Next 30 Days

Date	Forecasted_Cases	Forecasted_Hospitalizations	Forecasted_Deaths
186.7685	186.7685	3.6140882	0.2851325
157.6101	157.6101	0.0115449	0.5040716
204.3200	204.3200	-7.2539690	-0.2893573
203.1978	203.1978	-14.6544533	-0.9232080
179.5258	179.5258	-21.3297668	-2.0081773
191.5100	191.5100	-27.1478039	-2.8398011
203.7948	203.7948	-32.1646289	-3.7595120
178.5022	178.5022	-36.4757068	-4.4857413
174.3271	174.3271	-40.1761822	-5.1198003
200.4497	200.4497	-43.3513831	-5.6417699

Commentary on Forecast Comparison Table

The comparison of forecasted values for cases, hospitalizations, and deaths suggests a consistent downward trend over the next 30 days. While cases are projected to decrease steadily, hospitalizations and deaths are expected to decline more gradually, reflecting the typical delay between case surges and subsequent severe outcomes. These forecasts provide valuable insights for healthcare planning, allowing hospitals to anticipate future resource needs and authorities to adjust public health measures as necessary. However, the forecast's accuracy depends on the stability of current trends and may be influenced by unanticipated changes, such as new outbreaks or interventions.

Conclusion

This time series analysis of COVID-19 daily cases, hospitalizations, and deaths in New York City provides a comprehensive understanding of the pandemic's temporal dynamics. By employing time series decomposition, moving averages, and ARIMA forecasting models, we were able to extract meaningful patterns that highlight the evolution of the pandemic and provide actionable insights for public health decision-making.

At the core of this study was the research question: **How do the daily counts of COVID-19 cases, hospitalizations, and deaths evolve over time, and how can we forecast future trends to better inform public health planning?** Through this analysis, we aimed to answer this question by identifying key trends, seasonality, and predicting future values based on historical data.

Key Insights

Data Overview:

The dataset spanning from April 2020 to September 2024, captures critical phases of the COVID-19 pandemic in New York City. The descriptive statistics in Table 1 highlighted the significant variability in daily counts, with extreme peaks and low points reflective of the different waves of the pandemic. The data showed a right-skewed distribution, with occasional spikes in cases, hospitalizations, and deaths.

Visualizations of Key Indicators:

- **Time Series Plot:** The time series plot provided an overview of daily case, hospitalization, and death trends. The plot visually demonstrated the multiple pandemic waves, with clear peaks in 2020, 2021, and 2022, and smaller surges thereafter. Hospitalizations and deaths lagged behind cases, as expected, due to the natural delay in severe outcomes following infection surges.
- **7-Day Moving Averages:** By smoothing out daily fluctuations, the 7-day moving average graphs offered a clearer view of long-term trends. This method helped eliminate noise from the data, revealing more persistent increases or decreases in key indicators, providing a clearer sense of the pandemic's progression.

Decomposition Analysis:

Cases, Hospitalizations, and Deaths Decomposition: Decomposing the time series allowed us to separate the data into trend, seasonal, and residual components. This breakdown showed how cases, hospitalizations, and deaths exhibited consistent weekly seasonal patterns. These fluctuations were likely due to changes in reporting and testing frequency, particularly over weekends. The trend components of the decomposition revealed the overall upward surges early in the pandemic, followed by a stabilizing effect as the years progressed.

ARIMA Modeling and Forecasting:

The ARIMA models forecasted the future counts of cases, hospitalizations, and deaths over the next 30 days. These models indicated a slight decline in all three metrics, suggesting a potential tapering of the pandemic in the near future. However, the forecast models also revealed widening confidence intervals, indicating increased uncertainty in predictions as time progresses.

Public Health Implications

The trends and seasonality detected in this analysis are particularly important for public health officials in anticipating future healthcare demands. The use of ARIMA models to predict future values of cases, hospitalizations, and deaths provides several actionable insights:

- **Resource Allocation:** The forecasted values suggest that while a gradual decline is expected, there remains a possibility of localized surges, especially in hospitalizations. This can help hospitals better prepare for potential upticks in patients requiring care, ensuring that beds, staff, and ventilators are available when needed.
- **Surge Planning:** The decomposition analysis, which identified clear seasonal patterns, can inform officials of periods when cases may spike due to behavioral patterns (e.g., holidays or weekends). This can help guide public health messaging and interventions, ensuring that communities remain vigilant during high-risk periods.
- **Forecasts for Future Waves:** By understanding the lag between cases, hospitalizations, and deaths, policymakers can take preemptive action to mitigate severe outcomes. The forecasts of future trends can guide decisions on resource deployment, vaccination campaigns, and restrictions to curb the spread before another significant wave arises.

Final Thoughts

Overall, the time series analysis of COVID-19 data in New York City has provided valuable insights into the pandemic's evolution, helping us better understand the patterns and factors driving changes in key indicators like cases, hospitalizations, and deaths. By leveraging moving averages, decomposition techniques, and forecasting models, this analysis contributes to more informed public health planning and decision-making.

In summary, the key takeaways from this analysis are:

- Clear identification of long-term trends and seasonal patterns in COVID-19 indicators.
- Predictive models that offer a reliable short-term forecast, with caveats regarding uncertainty.
- Actionable insights for healthcare resource allocation, policy implementation, and surge planning.

Future studies could enhance these models by incorporating real-time data and external factors such as vaccination uptake, mobility, or new variants, thus improving the accuracy and reliability of forecasts. This analysis underscores the ongoing need for vigilant monitoring and responsive policymaking to manage the continued impacts of COVID-19.