Analysis of Johns Hopkins University COVID-19 Data

John Creath, III

2025-12-11

Introduction

This report provides a detailed analysis of COVID-19 trends in Peoria County, Illinois, using data collected by Johns Hopkins University.

The purpose of this analysis is to understand both the temporal progression of the pandemic and the relationship between cases and deaths at the county level.

Specifically, we examine:

- **Daily and monthly new cases and deaths**, to identify spikes and trends over time.
- **Cumulative counts**, to understand the overall scale and impact of COVID-19 in the county.
- **Predicted deaths** using a linear model that accounts for cases and population, to compare observed outcomes with model estimates.

The analysis leverages **interactive visualizations** powered by `plotly`, which allow readers to explore data dynamically, hover for exact values, and examine key periods of heightened transmission or mortality.

By combining both raw counts and derived metrics, this report provides a comprehensive view of local COVID-19 dynamics that can inform public health insights and planning.

Source of Analysis

The primary data source for this report is the Johns Hopkins University COVID-19 Data repository, available on GitHub (click here for link).

These data provide **daily counts of confirmed COVID-19 cases and deaths** for the United States at the **county level**, beginning in early 2020. The source dataset includes county identifiers, geographic coordinates, population estimates, and time-series data in **wide format**, where each date is a separate column.

Why this dataset?

Using county-level time-series data allows us to:

- Track local **infection trends and mortality** over time.

- Aggregate daily counts into **monthly summaries** for clearer visualization.
- Compute **new cases and deaths**, correcting for any negative or missing values due to reporting adjustments.

The data are transformed into **long format** for analysis, making it easier to compute metrics such as **monthly totals, cumulative counts, and per-capita adjustments**, which are essential for trend analysis and modeling.

By sourcing the data from a reputable and continuously updated repository, this analysis ensures that findings are based on **accurate, high-quality information**, allowing for meaningful insights into Peoria County's pandemic experience.

Setup

All required packages are loaded here. plotly is used to create interactive plots with tooltips, while dplyr and tidyr handle data wrangling.

```
library(dplyr)
library(tidyr)
library(readr)
library(lubridate)
library(plotly)
library(curl)
```

Loading Data

Defines the base GitHub URL for the Johns Hopkins COVID-19 time-series data. Creates a list of the two file names for COVID cases and deaths. These file names are appended to the base URL to form complete download links. Each CSV file is downloaded locally as cases.csv and deaths.csv. The downloaded files are then read into R as two data frames: jhu_covid_cases and jhu_covid_deaths.

```
base_url <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_c

files <- c(
  cases  = "time_series_covid19_confirmed_US.csv",
  deaths = "time_series_covid19_deaths_US.csv"
)

urls <- setNames(paste0(base_url, files), names(files))

# Download
curl::curl_download(urls["cases"],  destfile = "cases.csv")
curl::curl_download(urls["deaths"], destfile = "deaths.csv")

# Read
jhu_covid_cases  <- read_csv("cases.csv")
jhu_covid_deaths <- read_csv("deaths.csv")
```

Transforming the Data

The wide-format data (columns for each date) are transformed into long-format, where each row corresponds to a single date for a single county. This makes the dataset suitable for time-series analysis and plotting. Finally, daily new cases and deaths are calculated - this is crucial for trend analysis. Any missing or negative values are corrected to prevent misleading spikes caused by data corrections.

```
cases_long <- jhu_covid_cases %>%
  pivot_longer(
    cols = matches("^\\d{1,2}/\\d{1,2}/\\d{2}$"),
    names_to = "date",
    values_to = "cases"
  ) %>%
  mutate(date = mdy(date))

deaths_long <- jhu_covid_deaths %>%
  pivot_longer(
    cols = matches("^\\d{1,2}/\\d{1,2}/\\d{2}$"),
    names_to = "date",
    values_to = "deaths"
  ) %>%
  mutate(date = mdy(date))

covid_long <- cases_long %>%
  left_join(deaths_long,
            by = c("UID", "FIPS", "Admin2", "Province_State",
                   "Country_Region", "Lat", "Long_",
                   "Combined_Key", "date"))

covid_clean <- covid_long %>%
  # Drop unwanted columns
  select(
    -iso2.x, -iso3.x,
    -iso2.y, -iso3.y,
    -code3.y
  ) %>%
  # Rename columns
  rename(
    Code      = code3.x,
    County    = Admin2,
    State     = Province_State,
    Country   = Country_Region,
    Latitude  = Lat,
    Longitude = Long_,
    Date      = date,
    Cases     = cases,
    Deaths    = deaths
  ) %>%
  # Reorder columns
  select(
    UID,
    Code,
    FIPS,
    County,
```

```
    State,
    Country,
    Latitude,
    Longitude,
    Combined_Key,
    Population,
    Date,
    Cases,
    Deaths
  )

covid_clean_grouped <- covid_clean %>%
  group_by(County, State) %>%
  arrange(Date) %>%
  mutate(
    New_Cases  = Cases  - lag(Cases),
    New_Deaths = Deaths - lag(Deaths)
  ) %>%
  ungroup() %>%
  mutate(
    New_Cases  = replace_na(New_Cases, 0),
    New_Deaths = replace_na(New_Deaths, 0)
  )
```

Filtering and Aggregating for Discreet Use Case

Filters the cleaned dataset to Peoria County, Illinois. Aggregates data set of daily data for Peoria County into monthly totals smooths out noise and highlights long-term trends. Month_Label provides a human-readable format for plotting and tooltips.

```
peoria_monthly <- covid_clean_grouped %>%
  filter(County == "Peoria", State == "Illinois") %>%
  mutate(
    Month = floor_date(Date, "month"),
    # Remove negative corrections
    New_Cases  = if_else(New_Cases < 0, 0, New_Cases),
    New_Deaths = if_else(New_Deaths < 0, 0, New_Deaths)
  ) %>%
  group_by(Month) %>%
  summarize(
    Monthly_Cases  = sum(New_Cases, na.rm = TRUE),
    Monthly_Deaths = sum(New_Deaths, na.rm = TRUE)
  ) %>%
  ungroup()

peoria_monthly <- peoria_monthly %>%
  mutate(Month_Label = format(Month, "%B %Y"))
```

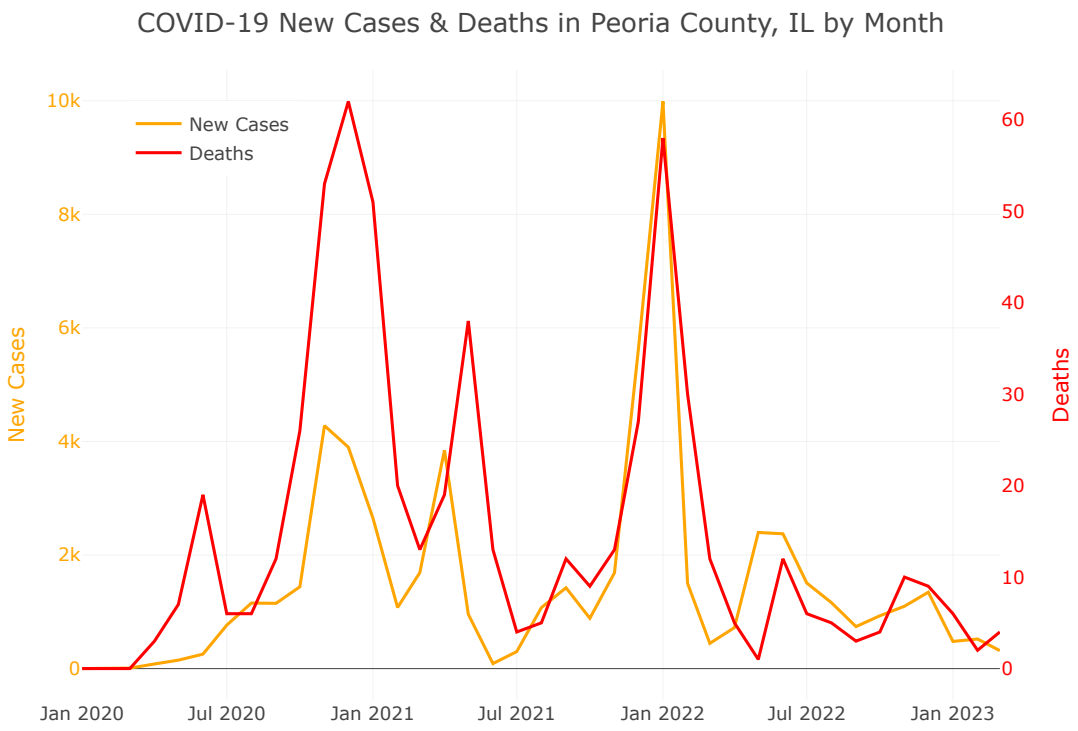
Plotting Discreet New Cases & Deaths

This interactive plot shows the monthly trajectory of new COVID-19 cases and deaths in Peoria County. The left axis represents new cases, while the right axis represents deaths. Axis title fonts and line colors are

synchronized for clarity so viewers can more easily intuit trends with relative scale. Peaks indicate periods of higher transmission or reporting.

```
plot_ly(peoria_monthly) %>%

  # Cases line
  add_lines(
    x = ~Month,
    y = ~Monthly_Cases,
    name = "New Cases",
    line = list(color = "orange"),
    hovertemplate = paste(
      "<b>%{text}</b><br>",
      "New Cases: %{y:d}<extra></extra>"
    ),
    text = ~Month_Label,
    yaxis = "y"  # left axis
  ) %>%

  # Deaths line
  add_lines(
    x = ~Month,
    y = ~Monthly_Deaths,
    name = "Deaths",
    line = list(color = "red"),
    hovertemplate = paste(
      "<b>%{text}</b><br>",
      "Deaths: %{y:d}<extra></extra>"
    ),
    text = ~Month_Label,
    yaxis = "y2"  # right axis
  ) %>%

  layout(
    title = "COVID-19 New Cases & Deaths in Peoria County, IL by Month",
    xaxis = list(title = ""),  # hides the x-axis title
    yaxis = list(
      title = "New Cases",
      titlefont = list(color = "orange"),
      tickfont = list(color = "orange")
    ),
    yaxis2 = list(
      title = "Deaths",
      titlefont = list(color = "red"),
      tickfont = list(color = "red"),
      overlaying = "y",
      side = "right",
      showgrid = FALSE
    ),
    legend = list(x = 0.05, y = 0.95),
    margin = list(
      t = 50,
      r = 60
    )
```

```
)
```


COVID-19 New Cases & Deaths in Peoria County, IL by Month
New Cases
Deaths
10k
8k
6k
4k
2k
0
60
50
40
30
20
10
0
Jan 2020
Jul 2020
Jan 2021
Jul 2021
Jan 2022
Jul 2022
Jan 2023

Filtering and Aggregating for Cumulative Use Case

Filters the cleaned dataset to Peoria County, Illinois, sets any negative case or death corrections to zero, and groups total cases and deaths by month. Computes cumulative cases and cumulative deaths and also adds a formatted month label for display.

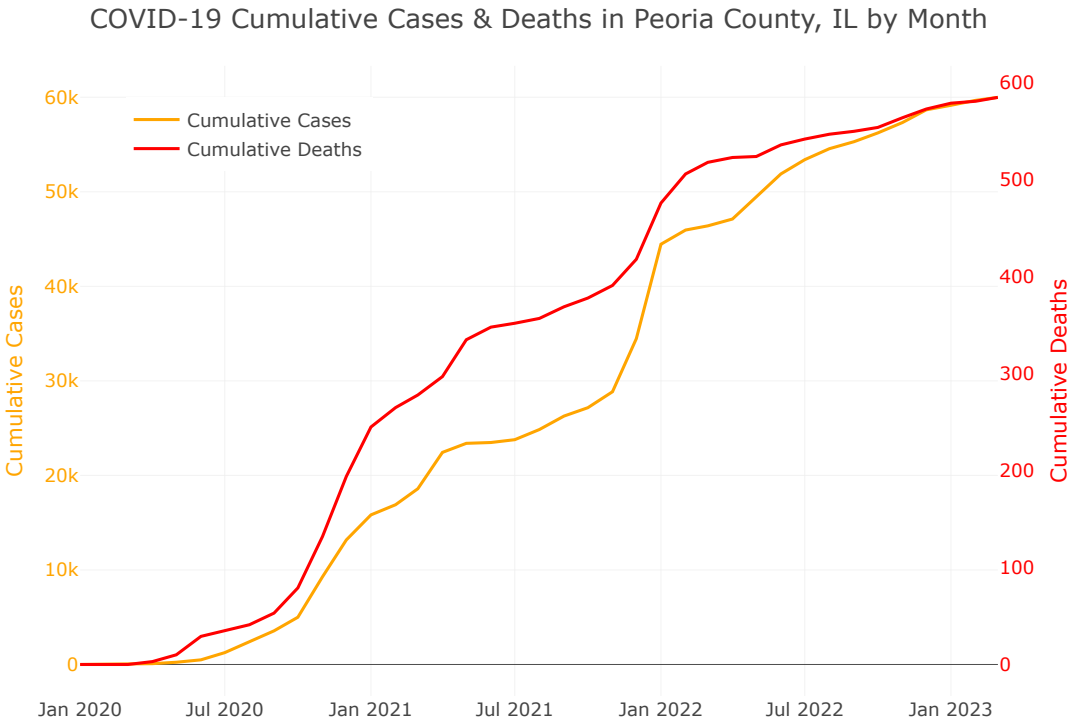
```
peoria_monthly_cum <- covid_clean_grouped %>%
  filter(County == "Peoria", State == "Illinois") %>%
  mutate(
    Month = floor_date(Date, "month"),
    # Remove negative corrections
    New_Cases  = if_else(New_Cases < 0, 0, New_Cases),
    New_Deaths = if_else(New_Deaths < 0, 0, New_Deaths)
  ) %>%
  group_by(Month) %>%
  summarize(
    Monthly_Cases  = sum(New_Cases, na.rm = TRUE),
    Monthly_Deaths = sum(New_Deaths, na.rm = TRUE)
  ) %>%
  ungroup() %>%
  arrange(Month) %>%
  mutate(
    Cumulative_Cases  = cumsum(Monthly_Cases),
    Cumulative_Deaths = cumsum(Monthly_Deaths),
    Month_Label = format(Month, "%B %Y")
  )
```

Plotting Cumulative New Cases & Deaths

Cumulative counts provide a sense of the overall scale and progression of the pandemic. They are particularly useful for comparing the total impact across regions or time periods. This cumulative plot clearly illustrates the overall growth of COVID-19 in Peoria County. The separation of cases and deaths helps assess mortality relative to total cases over time. Hover tooltips display the precise month and value for the viewer.

```
plot_ly(peoria_monthly_cum) %>%

  # Cases line
  add_lines(
    x = ~Month,
    y = ~Cumulative_Cases,
    name = "Cumulative Cases",
    line = list(color = "orange"),
    hovertemplate = paste(
      "<b>%{text}</b><br>",
      "Cumulative Cases: %{y:d}<extra></extra>"
    ),
    text = ~Month_Label,
    yaxis = "y"  # left axis
  ) %>%

  # Deaths line
  add_lines(
    x = ~Month,
```

```
    y = ~Cumulative_Deaths,
    name = "Cumulative Deaths",
    line = list(color = "red"),
    hovertemplate = paste(
      "<b>%{text}</b><br>",
      "Cumulative Deaths: %{y:d}<extra></extra>"
    ),
    text = ~Month_Label,
    yaxis = "y2"  # right axis
  ) %>%

  layout(
    title = "COVID-19 Cumulative Cases & Deaths in Peoria County, IL by Month",
    xaxis = list(title = ""),  # ← hides the x-axis title
    yaxis = list(
      title = "Cumulative Cases",
      titlefont = list(color = "orange"),
      tickfont = list(color = "orange")
    ),
    yaxis2 = list(
      title = "Cumulative Deaths",
      titlefont = list(color = "red"),
      tickfont = list(color = "red"),
      overlaying = "y",
      side = "right",
      showgrid = FALSE
    ),
    legend = list(x = 0.05, y = 0.95),
    margin = list(
      t = 50,
      r = 60
    )
  )
```


COVID-19 Cumulative Cases & Deaths in Peoria County, IL by Month
Cumulative Cases
Cumulative Deaths
60k
50k
40k
30k
20k
10k
0
600
500
400
300
200
100
0
Jan 2020
Jul 2020
Jan 2021
Jul 2021
Jan 2022
Jul 2022
Jan 2023

Per Capita Model

Creates a monthly version of the dataset by also converting each date to its month and grouping by state, county, and month. For each group, the model calculates total monthly cases and deaths and carries forward the population value. Fits a linear model predicting monthly deaths from monthly cases and population, then displays the model summary.

```
covid_monthly <- covid_clean_grouped %>%
  mutate(
    Month = floor_date(Date, "month")
  ) %>%
  group_by(State, County, Month) %>%
  summarize(
    Monthly_Cases = sum(New_Cases, na.rm = TRUE),
    Monthly_Deaths = sum(New_Deaths, na.rm = TRUE),
    Population = first(Population),
    .groups = "drop"
  )

model <- lm(
  Monthly_Deaths ~ Monthly_Cases + Population,
  data = covid_monthly
)

summary(model)
```

```
##
## Call:
## lm(formula = Monthly_Deaths ~ Monthly_Cases + Population, data = covid_monthly)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24074.6     -2.4     -1.0      0.4  10537.8
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.240e-01  2.547e-01   2.057   0.0397 *
## Monthly_Cases 3.318e-03  5.589e-05  59.376   <2e-16 ***
## Population    5.477e-05  8.865e-07  61.790   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 87.89 on 130335 degrees of freedom
## Multiple R-squared:  0.1072, Adjusted R-squared:  0.1072
## F-statistic:  7828 on 2 and 130335 DF,  p-value: < 2.2e-16
```

The model captures the basic idea that more cases → more deaths and larger population → more deaths, but it is far too simplistic for COVID data and produces poor predictive performance, especially for counties with large outbreaks or large populations. Precise observations:

1. The predictors matter, but the model barely explains the variation. Both Monthly_Cases and Population are extremely statistically significant ($p < 2e\text{-}16$), but the R^2 is only 0.107, meaning the model explains about 10.7% of the variation in monthly deaths. In a dataset this large, statistical significance is easy to get; explanatory power is what matters, and here it is low.

2. The effect sizes are tiny. Monthly Cases coefficient (0.0033) means roughly 1 additional death per 300 additional cases, on average. Population coefficient (5.5e-05) suggests deaths scale with population, but the magnitude is modest. These effects are plausible directionally, but they don't capture the true structure of COVID mortality, which is far more complex.

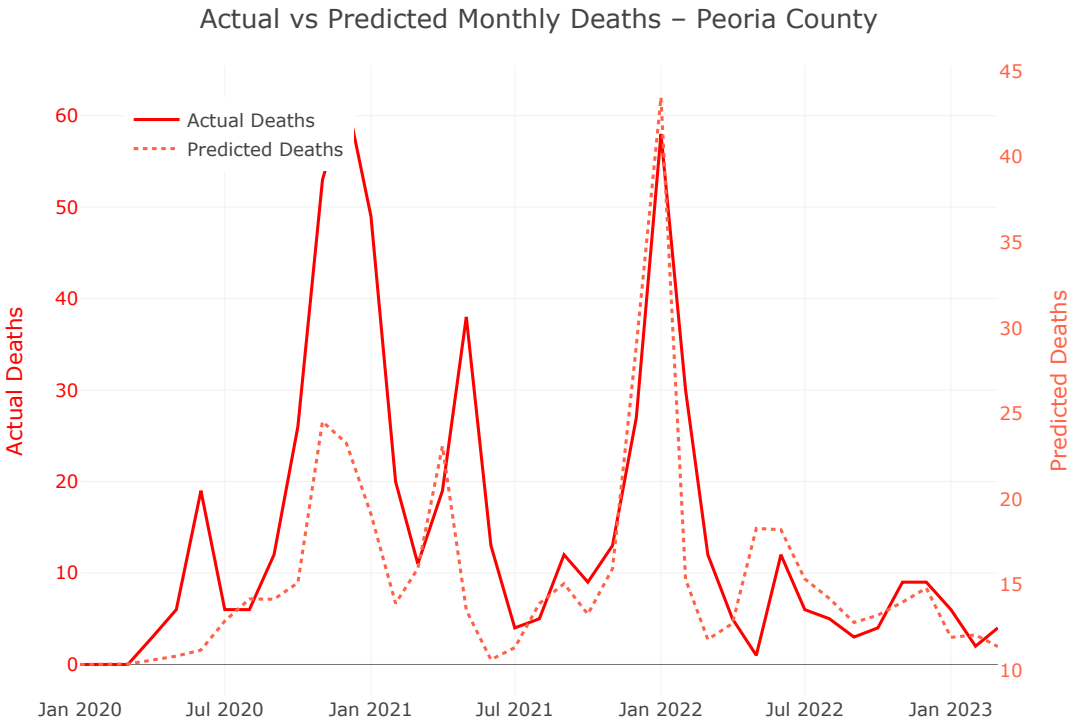
3. The residuals give away a deeper problem. Residuals range from –24,000 to +10,500 deaths, which appears absurd for a "monthly county deaths" model. However this is explained by the reality that counties vary wildly in population (for instance Los Angeles County vs. rural counties) so the model tries to fit both with the same linear structure, which creates gigantic residuals for large counties. That's a symptom of extreme outliers and a model structure that is far too simple for a national scale.

Plotting Predictive Model for Peoria County

Filters the monthly dataset to Peoria County, Illinois. For each month, computes predicted deaths using the fitted model and adds a formatted month label. Despite the model's low explanatory power, its predictions for Peoria appear closer to the observed values later in the pandemic. This reflects greater stability in the underlying data rather than improved model performance.

```
peoria_predictions <- covid_monthly %>%
  filter(County == "Peoria", State == "Illinois") %>%
  mutate(
    Predicted_Deaths = predict(model, newdata = .),
    Month_Label = format(Month, "%B %Y")
  ) %>%
  select(Month, Month_Label, Monthly_Cases, Monthly_Deaths, Predicted_Deaths)


plot_ly(peoria_predictions) %>%

  # Actual deaths (solid red)
  add_lines(
    x = ~Month,
    y = ~Monthly_Deaths,
    name = "Actual Deaths",
    line = list(color = "red", width = 2),
    hovertemplate = paste(
      "<b>%{text}</b><br>",
      "Actual Deaths: %{y:d}<extra></extra>"
    ),
    text = ~Month_Label,
    yaxis = "y"    # left axis
  ) %>%

  # Predicted deaths (dotted, lighter red)
  add_lines(
    x = ~Month,
    y = ~Predicted_Deaths,
    name = "Predicted Deaths",
    line = list(color = "tomato", width = 2, dash = "dot"),
    hovertemplate = paste(
      "<b>%{text}</b><br>",
      "Predicted Deaths: %{y:d}<extra></extra>"
    ),
```

```
  text = ~Month_Label,
  yaxis = "y2"  # right axis
) %>%

layout(
  title = "Actual vs Predicted Monthly Deaths - Peoria County",

  xaxis = list(title = ""),

  # Left axis → actual
  yaxis = list(
    title = "Actual Deaths",
    titlefont = list(color = "red"),
    tickfont  = list(color = "red")
  ),

  # Right axis → predicted
  yaxis2 = list(
    title = "Predicted Deaths",
    titlefont = list(color = "tomato"),
    tickfont  = list(color = "tomato"),
    overlaying = "y",
    side = "right",
    showgrid = FALSE
  ),

  legend = list(x = 0.05, y = 0.95),

  margin = list(
    t = 50,
    r = 60
  )
)
```


Actual vs Predicted Monthly Deaths – Peoria County
Actual Deaths
Predicted Deaths
Actual Deaths
Predicted Deaths
0
10
20
30
40
50
60
10
15
20
25
30
35
40
45
Jan 2020
Jul 2020
Jan 2021
Jul 2021
Jan 2022
Jul 2022
Jan 2023

Conclusion

Trends in Peoria County: Peaks in monthly new cases correspond to periods of increased transmission, likely linked to local outbreaks or broader pandemic waves.

- Mortality trends: Deaths lag behind new cases, as expected, and cumulative deaths provide a sense of the pandemic's severity.
- Cumulative impact: The cumulative plots highlight the overall burden, showing the growth of infections and deaths over time.
- Interactive exploration: These plots can be used to explore specific months and identify periods of concern, aiding in understanding local pandemic dynamics.

This analysis framework can be applied to other counties or states for comparative studies. Using a combination of daily, monthly, and cumulative metrics provides a comprehensive view of COVID-19 trends.