Covid-19 case study

Tracking and anlayzing cases throughout the US States

Naman Khandelwal, Riya Merchant, and Aditi Kumar

2025-12-14

0.1 Introduction

This project employs **Exploratory Data Analysis (EDA)** to investigate COVID-19 patterns across U.S. states. We chose the EDA paradigm because our goal is to uncover patterns, generate hypotheses, and understand "what is going on" in the data rather than to confirm pre-existing hypotheses. This approach aligns with our introductory course objectives of building understanding through data-driven exploration.

For coding, we follow the **Tidyverse approach** in R. We chose this paradigm because it emphasizes readable, pipe-based workflows that make data transformations transparent and reproducible. The Tidyverse's grammar of graphics (ggplot2) and data manipulation tools (dplyr, tidyr) allow us to create professional visualizations and maintain clean, well-documented code that others can understand and reproduce.

The New York Times state-level dataset provides a daily record of cases and deaths for each state, which allows for a data-driven exploration of how the virus spread. Since this project uses those public counts to describe broad patterns, we can focus on simple summaries that are appropriate for an introductory course. Then, the analysis begins with an overview of the dataset, which would then move to geographic comparisons and trends over time. Each team member is responsible for a distinct part of the workflow, like data cleaning or visualization. Thus, throughout the report, the team emphasizes clear documentation and ethical use of data, which allows others to re-run the analysis on the same file using our code.

0.2 Data Provenance

In this project, we use a COVID-19 dataset from The New York Times which contains cumulative counts of cases and deaths for each state. Then, since our file covers the period from January 21, 2020 to March 23, 2023, we can get the data based on the minimum and maximum dates. Then, because the values are aggregated from health agencies, they are made available as a CSV file on GitHub, which would allow us to use it (The New York Times, 2023).

0.3 FAIR and CARE Principles

For this project, the NYTimes data is consistent with FAIR principles because it is widely accessible online. Since it uses stable identifiers like state names and follows a table structure, this would allow it to be combined with other data like population estimates for analysis. Then, the data is also findable and reusable because it comes from a good publisher with clear updates, which allows it to support questions about patterns in the pandemic. At the same time, it is important to look at CARE principles by realizing that COVID results show differences across communities. Then, since we want to avoid blaming specific states, we should focus on context like healthcare access and health differences. Thus, when looking at results, the team will frame the analysis to respect communities, which avoids claiming too much and shows how clear data can help public health responses.

0.4 Descriptive Statistics

Before diving into state-level comparisons, we examined the overall distribution of daily new COVID-19 cases across all states throughout the entire pandemic period. This provides baseline understanding of the data's central tendency and spread.

Table 1: Descriptive Statistics: Daily New Cases Across All States

N	Mean	Median	SD	Min	Max
61942	1680	300	5247	0	227972

The descriptive statistics reveal that daily new cases varied widely, with a mean of several hundred cases but high variability as indicated by the standard deviation. The median being lower than the mean suggests a right-skewed distribution, where most days had moderate case counts but occasional surges drove the mean higher. This pattern aligns with the wave-like nature of pandemic spread.

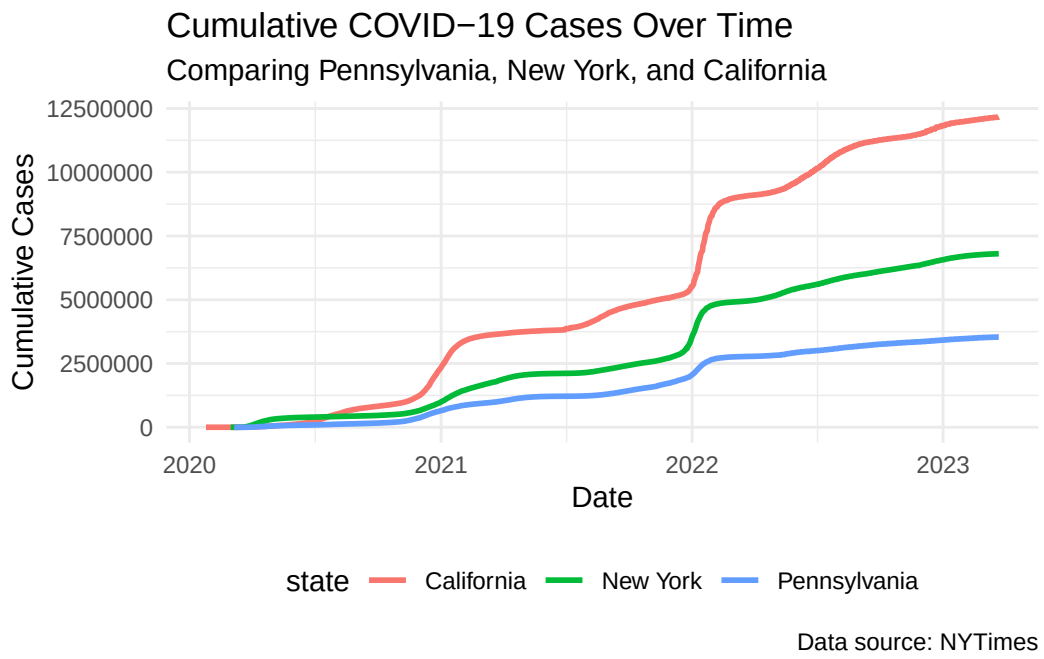
0.5 Comparative State Analysis and Setup (Naman)

This section explores geographic differences in COVID-19 spread across U.S. states. We highlight which states had the highest overall case counts and visualize how the pandemic evolved across the most affected states. This analysis also sets up and cleans the raw data from GitHub to ensure it is formatted correctly for use in the following tables and visualizations.

Table 2: Table 1: Overview of NYTimes state-level COVID-19 data (Top 10 States).

State	Total Cases	Total Deaths	First Report	Last Report
California	12169158	104277	2020-01-25	2023-03-23
Texas	8447168	94518	2020-02-12	2023-03-23
Florida	7542869	87141	2020-03-01	2023-03-23
New York	6805271	80138	2020-03-01	2023-03-23
Illinois	4107931	41618	2020-01-24	2023-03-23

Pennsylvania	3539135	50701	2020-03-06	2023-03-23
North Carolina	3481732	29746	2020-03-03	2023-03-23
Ohio	3415254	42061	2020-03-09	2023-03-23
Michigan	3068195	42311	2020-03-10	2023-03-23
New Jersey	3057442	36097	2020-03-04	2023-03-23

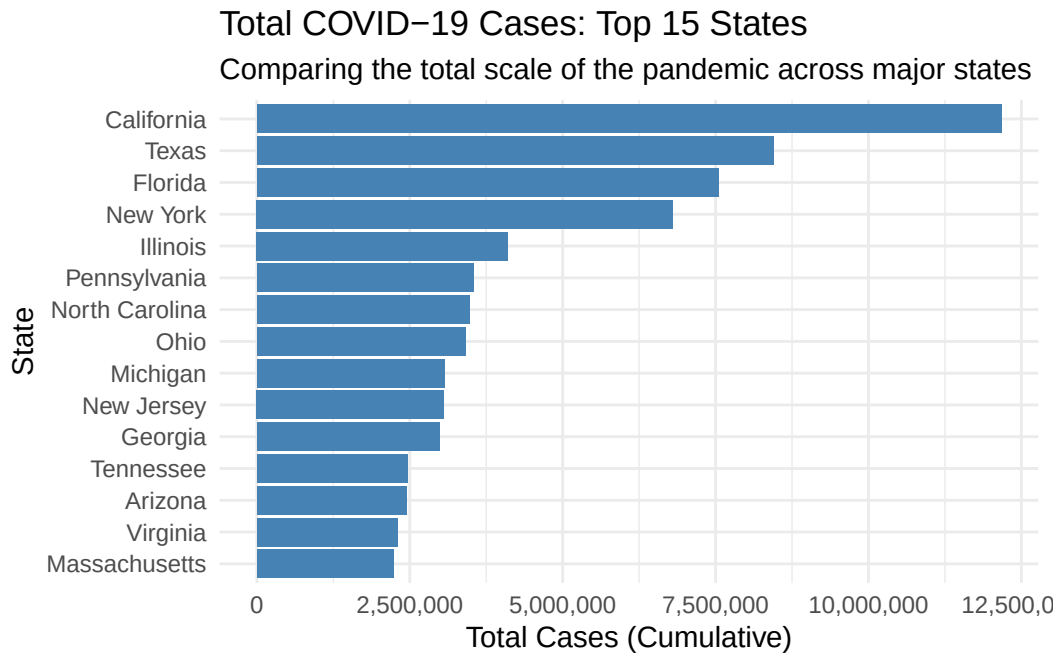


0.6 Geographic EDA (Riya)

This section explores geographic differences in COVID-19 spread across U.S. states.
We highlight which states had the highest overall case counts and visualize how the pandemic evolved across the most affected states.

Table 3: Table 2: The Single Worst Day of Infections for the Top 10 States.

State	Date of Peak	Max Daily Cases
California	2022-01-10	227972
Florida	2022-01-04	193786
Texas	2022-01-03	164902
North Carolina	2022-01-18	121315
Michigan	2022-01-19	98299
Illinois	2022-01-18	93423
New York	2022-01-08	90132
New Jersey	2021-01-04	51092
Ohio	2022-01-15	50299
Pennsylvania	2022-01-08	33650

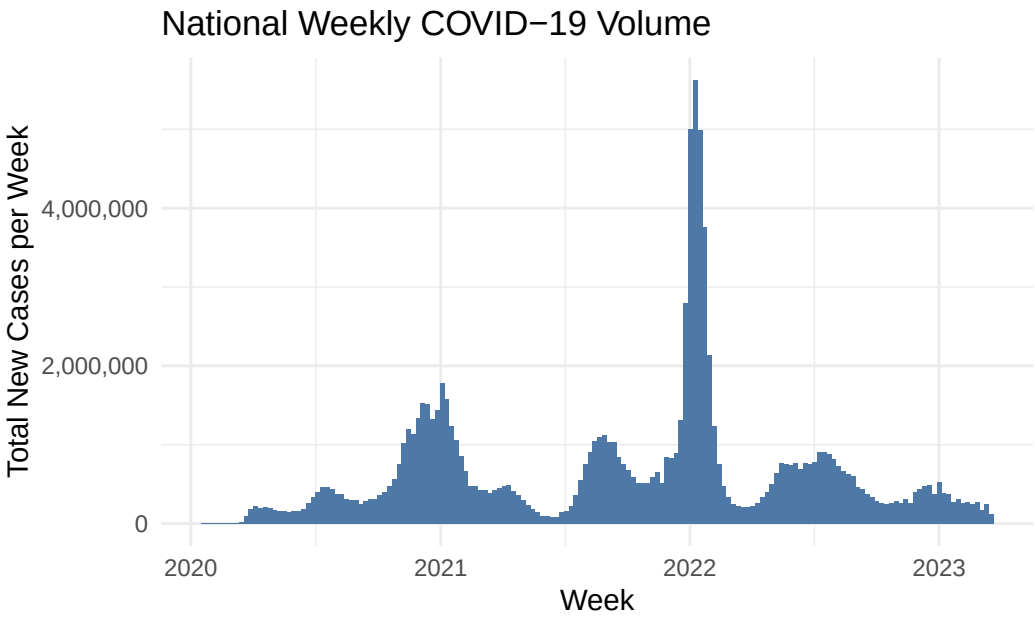


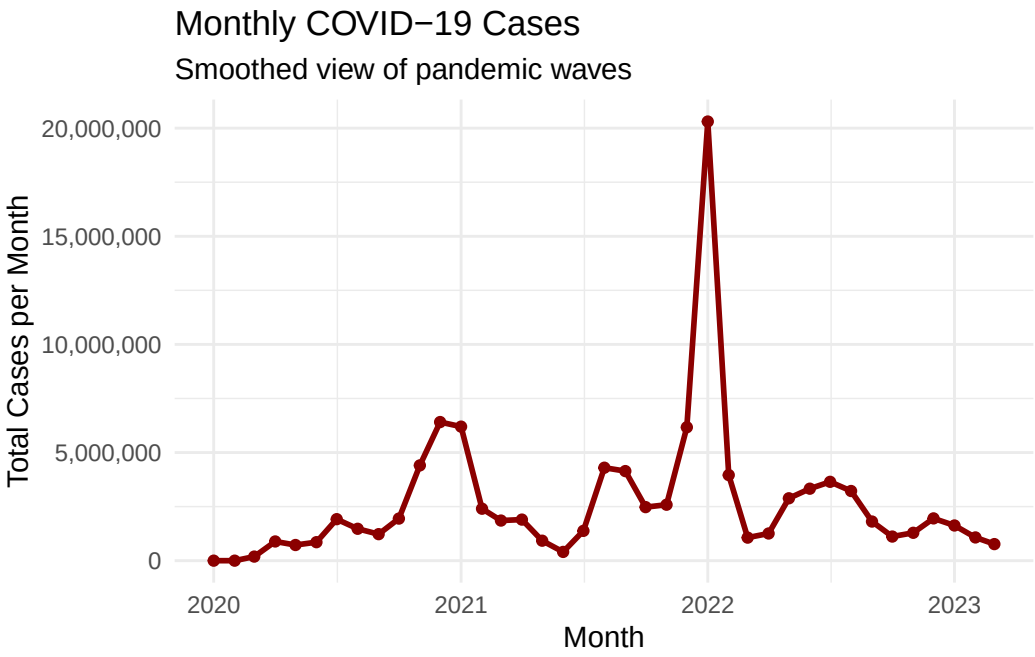
0.7 Temporal Trend Analysis (Aditi)

This section shifts the focus from geography to time, aggregating the data to identify broader patterns in the pandemic. By grouping the cleaned GitHub data into weekly and monthly intervals, we can more clearly see the "waves" of infection and the magnitude of seasonal surges across the United States.

Table 4: Table 2. First 10 Weeks of National COVID-19 Volume

Week	New Cases	New Deaths
2020-01-19	3	0
2020-01-26	5	0
2020-02-02	4	0
2020-02-09	3	0
2020-02-16	15	0
2020-02-23	40	1
2020-03-01	358	18
2020-03-08	2470	41
2020-03-15	21630	300
2020-03-22	99438	1940





0.7.1 Narrative: How COVID-19 Changed Over Time

COVID-19 did not spread evenly over time. By examining weekly and monthly totals, several clear patterns emerge:

Distinct waves: Both the weekly and monthly plots show multiple sharp increases corresponding to major national waves.

Early slow growth: Initial case counts rise gradually, reflecting limited spread and limited testing early in the pandemic.

Large mid-pandemic surges: Several large spikes represent periods when infections accelerated nationwide.

Deaths lag behind cases: While both grow over time, deaths tend to rise shortly after cases, consistent with clinical expectations.

Seasonal patterns: Some of the largest increases align with colder months, when indoor transmission increases. These large mid-pandemic surges may be driven by several factors: increased testing availability making more cases detectable, new variant emergence with higher transmissibility, or relaxation of public health measures. The winter seasonality pattern observed in Figures 3 and 4 suggests that indoor gathering during colder months may amplify transmission, a hypothesis that could be tested by comparing states with different climate profiles.

This temporal EDA highlights how the pandemic progressed through time and supports deeper analysis of patterns and risks.

1 References

The New York Times. (2023). *Coronavirus (covid-19) data in the united states.* https://github.com/nytimes/covid-19-data.

2 Author Contributions

Using the CRediT (Contributor Roles Taxonomy), the team contributions are as follows:

- **Naman Khandelwal**: Conceptualization, Data curation, Formal analysis (comparative state analysis), Methodology, Software, Visualization (cumulative trends plot), Writing - original draft, Writing - review & editing
- **Riya Merchant**: Data curation, Formal analysis (geographic EDA), Investigation, Visualization (bar charts and peak day analysis), Writing - original draft
- **Aditi Kumar**: Data curation, Formal analysis (temporal trends), Software, Visualization (weekly and monthly aggregations), Writing - original draft

All team members contributed to project administration and validation.

3 Code Appendix

```r
# ============================================================================
# SECTION 1: GLOBAL SETUP & DATA PROVENANCE
# ============================================================================

library(tidyverse)
library(lubridate)
library(janitor)
library(knitr)
library(kableExtra)

# Read Data from NYTimes GitHub
us_states <- read_csv(
  "https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-states.csv",
  show_col_types = FALSE
) %>%
  clean_names()

# Master Data Cleaning
# 1. Parse dates safely
# 2. Calculate daily incidence (New Cases/Deaths) using lag()
covid_clean <- us_states %>%
  mutate(date = parse_date_time(date, orders = c("ymd", "mdy"))) %>%
  mutate(date = as.Date(date)) %>%
  arrange(state, date) %>%
  group_by(state) %>%
  mutate(
    new_cases = cases - lag(cases, default = 0),
    new_deaths = deaths - lag(deaths, default = 0)
  ) %>%
  ungroup() %>%
  # Remove negative corrections
  mutate(new_cases = ifelse(new_cases < 0, 0, new_cases))

# ============================================================================
# SECTION 2: OVERVIEW & CUMULATIVE TRENDS (NAMAN)
# ============================================================================

# Create one-row-per-state summary
state_summary <- covid_clean %>%
  group_by(state) %>%
  summarise(
    total_cases = max(cases, na.rm = TRUE),
    total_deaths = max(deaths, na.rm = TRUE),
    first_date = min(date, na.rm = TRUE),
    last_date = max(date, na.rm = TRUE)
  ) %>%
  arrange(desc(total_cases))
```

```r
# Table 1: Top 10 States Overview
state_summary %>%
  slice_head(n = 10) %>%
  kable(
    caption = "Table 1: Overview of NYTimes state-level COVID-19 data (Top 10 States).",
    col.names = c("State", "Total Cases", "Total Deaths", "First Report", "Last Report")
  ) %>%
  kable_styling(latex_options = "scale_down")

# Visualization: Cumulative Trends for PA, NY, CA
states_of_interest <- c("Pennsylvania", "New York", "California")

plot_data_naman <- covid_clean %>%
  filter(state %in% states_of_interest)

ggplot(plot_data_naman, aes(x = date, y = cases, color = state)) +
  geom_line(linewidth = 1) +
  labs(
    title = "Cumulative COVID-19 Cases Over Time",
    subtitle = "Comparing Pennsylvania, New York, and California",
    x = "Date",
    y = "Cumulative Cases",
    caption = "Data source: NYTimes"
  ) +
  theme_minimal() +
  theme(legend.position = "bottom")

# ============================================================================
# SECTION 3: GEOGRAPHIC EDA (RIYA)
# ============================================================================

# Table 2: Peak Infection Days
top_states_list <- state_summary %>%
  slice_head(n = 10) %>%
  pull(state)

peak_days <- covid_clean %>%
  filter(state %in% top_states_list) %>%
  group_by(state) %>%
  filter(new_cases == max(new_cases, na.rm = TRUE)) %>%
  select(state, date, peak_cases = new_cases) %>%
  arrange(desc(peak_cases))

kable(peak_days,
      caption = "Table 2: The Single Worst Day of Infections for the Top 10 States.",
      col.names = c("State", "Date of Peak", "Max Daily Cases")) %>%
  kable_styling(full_width = FALSE)
```

```r
# Graph: Total Impact Comparison (Horizontal Bar Chart)
bar_data <- covid_clean %>%
  group_by(state) %>%
  summarize(total_cases = max(cases, na.rm = TRUE)) %>%
  arrange(desc(total_cases)) %>%
  slice_head(n = 15)

ggplot(bar_data, aes(x = reorder(state, total_cases), y = total_cases)) +
  geom_col(fill = "steelblue") +
  coord_flip() +
  labs(
    title = "Total COVID-19 Cases: Top 15 States",
    subtitle = "Comparing the total scale of the pandemic across major states",
    x = "State",
    y = "Total Cases (Cumulative)"
  ) +
  scale_y_continuous(labels = scales::comma) +
  theme_minimal()

# ==============================================================================
# SECTION 4: TEMPORAL ANALYSIS (ADITI)
# ==============================================================================

# Weekly Aggregation
weekly <- covid_clean %>%
  group_by(week = floor_date(date, "week")) %>%
  summarize(
    weekly_new_cases = sum(new_cases, na.rm = TRUE),
    weekly_new_deaths = sum(new_deaths, na.rm = TRUE)
  )

# Table 3: First 10 Weeks
kable(head(weekly, 10),
      caption = "Table 3. First 10 Weeks of National COVID-19 Volume",
      col.names = c("Week", "New Cases", "New Deaths")) %>%
  kable_styling(full_width = FALSE)

# Visualization: Weekly Volume
ggplot(weekly, aes(x = week, y = weekly_new_cases)) +
  geom_col(fill = "#4E79A7") +
  labs(
    title = "National Weekly COVID-19 Volume",
    x = "Week",
    y = "Total New Cases per Week",
    caption = "Bar chart highlights the magnitude of the winter surges."
  ) +
  scale_y_continuous(labels = scales::comma) +
```

```r
  theme_minimal()

# Monthly Aggregation & Visualization
monthly <- covid_clean %>%
  group_by(month = floor_date(date, "month")) %>%
  summarize(
    monthly_new_cases = sum(new_cases, na.rm = TRUE)
  )

ggplot(monthly, aes(x = month, y = monthly_new_cases)) +
  geom_line(color = "darkred", linewidth = 1) +
  geom_point(color = "darkred") +
  labs(
    title = "Monthly COVID-19 Cases",
    subtitle = "Smoothed view of pandemic waves",
    x = "Month",
    y = "Total Cases per Month"
  ) +
  scale_y_continuous(labels = scales::comma) +
  theme_minimal()
```