COVID-19 Data Analysis Final

HDaher

2024-04-28

Objective

This report aims to demonstrate effective methods for analyzing and visualizing COVID-19 data to understand global and US-specific trends in cases and deaths. Additionally, it aims to assess the accuracy of statistical modeling techniques in predicting trends in pandemic growth rates.

Data Source and Description

The COVID-19 data used in this report is sourced from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University.The dataset contains five time-series tables tracking COVID-19 cases and deaths. Two tables cover the United States at the county level, while the remaining three tables contain global data. Province/state-level data is included for certain countries, with country-level data for the rest.

```
library('tidyverse')
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library('lubridate')
library('choroplethr')
```

```
## Loading required package: acs
## Loading required package: XML
##
## Attaching package: 'acs'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
##
## The following object is masked from 'package:base':
##
##     apply
```

```
library('ggplot2')
library('dplyr')
```

```
# Get data
url_base <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_c
file_names <- c("time_series_covid19_confirmed_global.csv",
                "time_series_covid19_deaths_global.csv",
                "time_series_covid19_confirmed_US.csv",
                "time_series_covid19_deaths_US.csv")
urls <- str_c(url_base, file_names)

# Load and tidy global data
global_data <- read_csv(urls[1], show_col_types = FALSE) %>%
  select(-c('Province/State', 'Lat', 'Long')) %>%
  pivot_longer(cols = -c('Country/Region'), names_to = "date", values_to = "cases") %>%
  rename(country = 'Country/Region') %>%
  full_join(read_csv(urls[2], show_col_types = FALSE) %>%
            select(-c('Province/State', 'Lat', 'Long')) %>%
            pivot_longer(cols = -c('Country/Region'), names_to = "date", values_to = "deaths") %>%
            rename(country = 'Country/Region'),
            by = c("country", "date")) %>%
  mutate(date = mdy(date))
```

```
# Load and tidy us data
us_data <- read_csv(urls[3]) %>%
  select(-c('UID', 'iso2', 'iso3', 'code3', 'FIPS', 'Admin2', 'Lat', 'Long_', 'Combined_Key')) %>%
  pivot_longer(cols = -c('Province_State', 'Country_Region'), names_to = "date", values_to = "cases") %
  group_by(Province_State, date) %>%
  summarise(cases = sum(cases, na.rm = TRUE)) %>%
  mutate(date = mdy(date)) %>%
  full_join(read_csv(urls[4]) %>%
            select(-c('UID', 'iso2', 'iso3', 'code3', 'FIPS', 'Admin2', 'Lat', 'Long_', 'Combined_Key',
            pivot_longer(cols = -c('Province_State', 'Country_Region'), names_to = "date", values_to =
            group_by(Province_State, date) %>%
            summarise(deaths = sum(deaths, na.rm = TRUE)) %>%
            mutate(date = mdy(date)),
            by = c("Province_State", "date")) %>%
  arrange(Province_State, date) %>%
  group_by(Province_State) %>%
  mutate(new_deaths = deaths - lag(deaths),
         new_cases = cases - lag(cases))
```

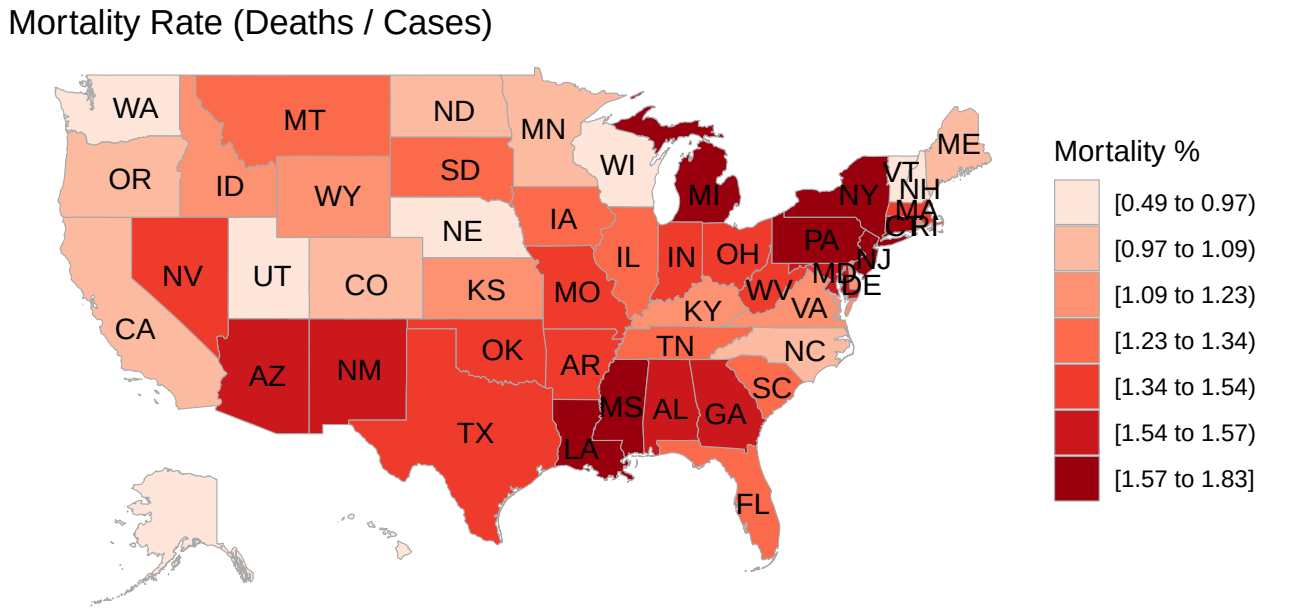
Visual 1: Mapping COVID-19 Severity Across US States

To help facilitate understanding of the severity of COVID-19 at the state level, a choropleth map was utilized to display mortality rates (deaths/cases) across US states. This visual representation offers a clear depiction of regional disparities in infection severity. It can be utilized to identify areas with higher severity

levels, enabling decision-making and resource allocation efforts, while also evaluating the effectiveness of interventions over time.

```
# Calculate US mortality rates
US_mortality_rate <- us_data %>%
  filter(!Province_State %in% c("American Samoa","Diamond Princess","Grand Princess","Guam",
                        "Northern Mariana Islands","Puerto Rico","Virgin Islands")) %>%
  group_by(Province_State) %>%
  summarize(totDeaths = sum(deaths),
            totConfirmed = sum(cases)) %>%
  mutate(DR = 100 * totDeaths/totConfirmed) %>%
  mutate(DR = round(DR,2)) %>%
  arrange(-DR) %>%
  select(Province_State, DR) %>%
  rename(region = 1, value = DR) %>%
  mutate(region = tolower(region))

# Plot US mortality rate map
choro <- StateChoropleth$new(US_mortality_rate)
choro$title <- "Mortality Rate (Deaths / Cases)"
choro$ggplot_scale <- scale_fill_brewer(name = "Mortality %", palette = "Reds", drop = FALSE)
choro$render()
```



The map indicates that the Northeast and South have the highest COVID-19 mortality rates, possibly due to high population density and early outbreak intensity. In contrast, the Midwest and Mountain West regions have the lowest rates, which could be attributed to lower population density or various public health interventions.

Visual 2: Trends in COVID-19 Cases and Deaths in the United States

Line graphs for new COVID-19 cases and deaths in the U.S. provide a straightforward view of how the virus has spread and its deadly impact over time. This clear format aids in identifying peak periods of infection, this can aid in identifying effectiveness of mitigative measures to reduce the spread and severity of the pandemic.

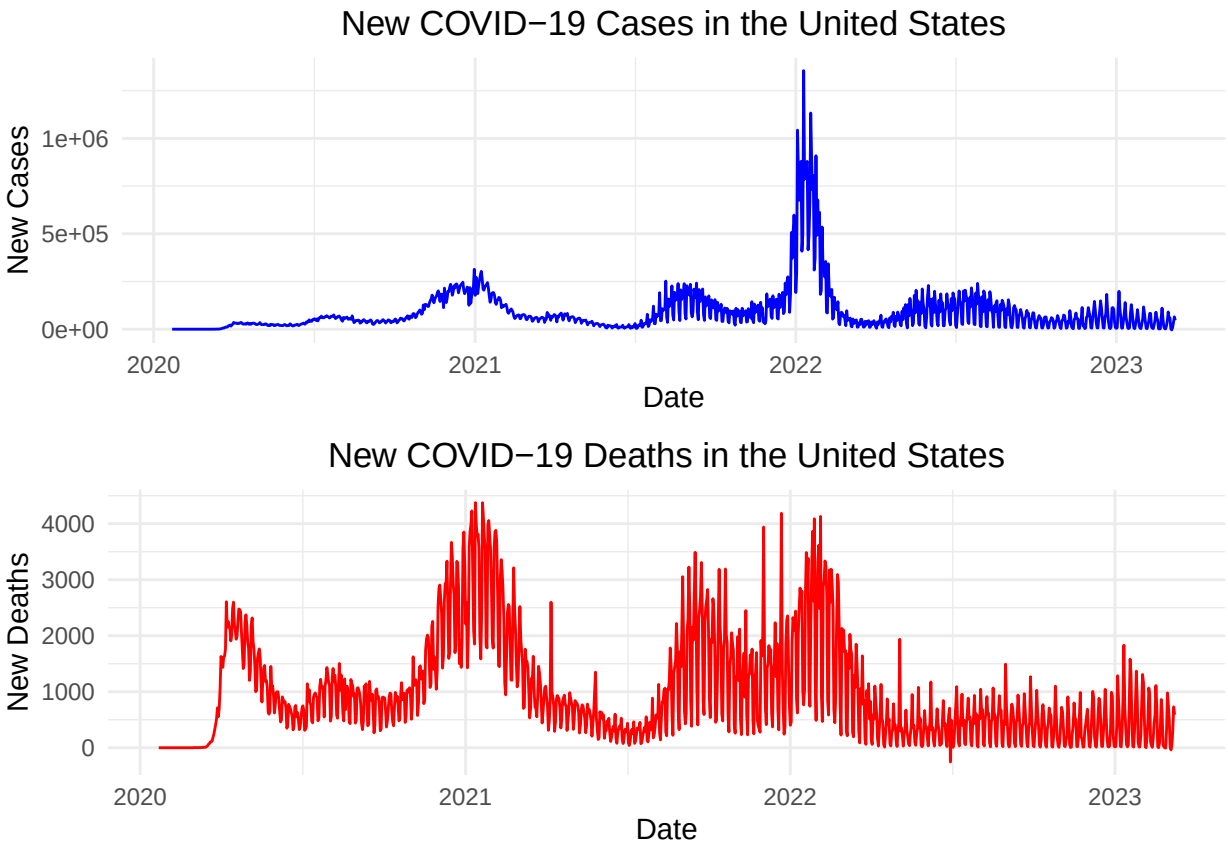
```
# Combine US data
us_aggregated_data <- us_data %>%
  group_by(date) %>%
  summarise(new_cases = sum(new_cases, na.rm = TRUE),
            new_deaths = sum(new_deaths, na.rm = TRUE))

# Plot new cases
cases_plot <- ggplot(us_aggregated_data, aes(x = date, y = new_cases)) +
  geom_line(colour = "blue", size = 0.5) +
  labs(title = 'New COVID-19 Cases in the United States',
       x = 'Date', y = 'New Cases') +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

```
# Plot new deaths
deaths_plot <- ggplot(us_aggregated_data, aes(x = date, y = new_deaths)) +
  geom_line(colour = "red", size = 0.5) +
  labs(title = 'New COVID-19 Deaths in the United States',
       x = 'Date', y = 'New Deaths') +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))


gridExtra::grid.arrange(cases_plot, deaths_plot, ncol = 1)
```



The line graphs show that the pandemic was initially more severe, with a pronounced spike in deaths early on despite the relatively low case count. Interestingly, despite a dramatic surge in cases in early 2022, the death rates did not show a corresponding increase, suggesting that interventions such as vaccinations may have played a crucial role in reducing mortality even when case numbers were high. This indicates that while infections remained a challenge, the healthcare responses, including vaccine distribution, were effective in mitigating the fatality of the virus.

Model: Logistic Growth Curve Analysis

A logistic growth curve was utilized to model global COVID-19 case data to understand and forecast the pandemic's progression. This model is beneficial for epidemiological studies as it accounts for the initial exponential increase in cases followed by a plateau, reflecting constraints like herd immunity or the impact of public health interventions.

```
library(minpack.lm)

# Combine global data by date
global_aggregated_data <- global_data %>%
  group_by(date) %>%
  summarise(total_cases = sum(cases, na.rm = TRUE))

# Convert dates to a days since the first date
global_aggregated_data$date_num <- as.numeric(global_aggregated_data$date - min(global_aggregated_data$

# Logistic growth model function
```

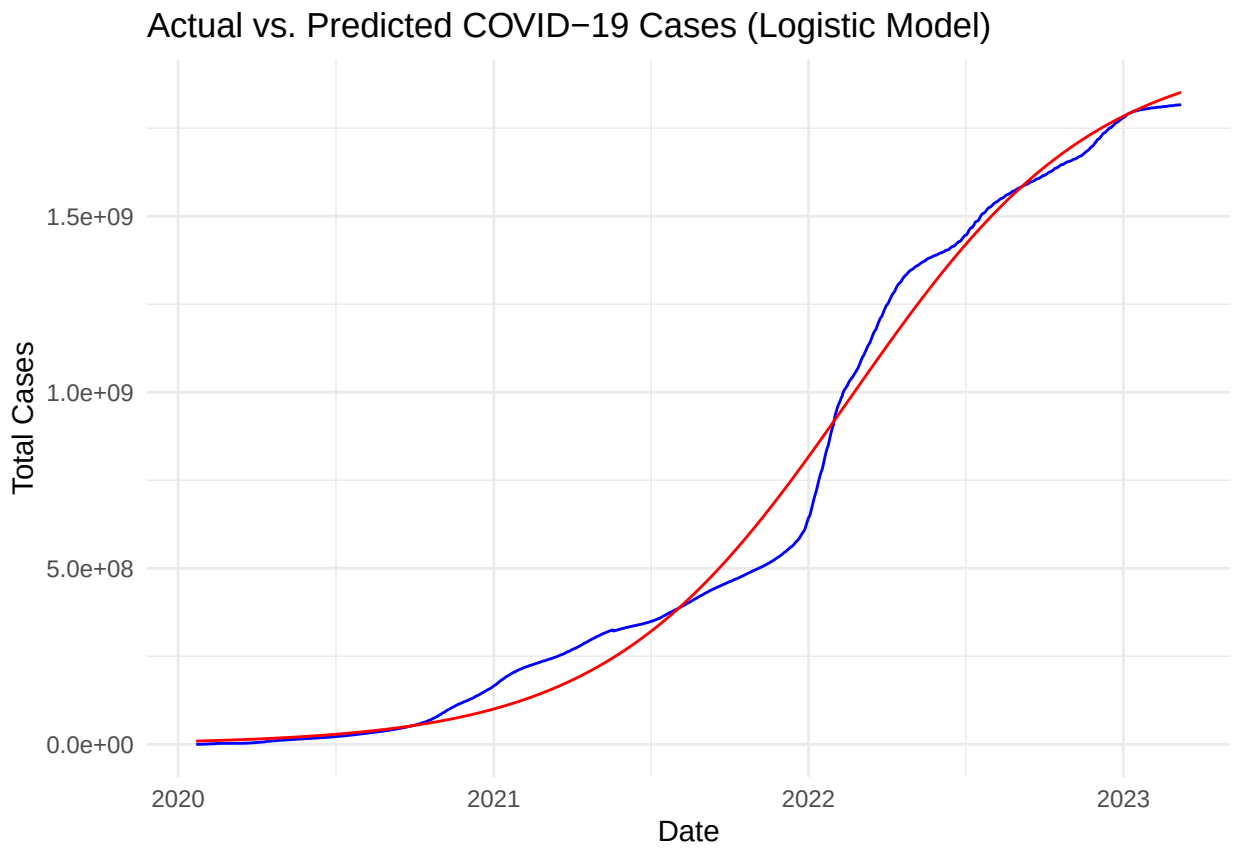
```r
logistic_growth <- function(x, K, r, x0) {
  K / (1 + ((K - x0) / x0) * exp(-r * x))
}

# Fit the model
# x0 is the initial number of cases, K is the carrying capacity, r is the growth rate
fit <- nlsLM(total_cases ~ logistic_growth(date_num, K, r, x0),
             data = global_aggregated_data,
             start = list(K = max(global_aggregated_data$total_cases),
                          r = 0.1, x0 = min(global_aggregated_data$total_cases)),
             control = nls.lm.control(maxiter = 100))


summary(fit)
```

```
##
## Formula: total_cases ~ logistic_growth(date_num, K, r, x0)
##
## Parameters:
##     Estimate Std. Error t value Pr(>|t|)
## K  1.976e+09  1.157e+07  170.78   <2e-16 ***
## r  7.064e-03  7.634e-05   92.53   <2e-16 ***
## x0 9.201e+06  4.632e+05   19.86   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 66570000 on 1140 degrees of freedom
##
## Number of iterations to convergence: 18
## Achieved convergence tolerance: 1.49e-08
```

```r
# Plot model fit
global_aggregated_data$predicted_cases <- predict(fit, newdata = global_aggregated_data)
ggplot(global_aggregated_data, aes(x = date)) +
  geom_line(aes(y = total_cases), colour = "blue") +
  geom_line(aes(y = predicted_cases), colour = "red") +
  labs(title = "Actual vs. Predicted COVID-19 Cases (Logistic Model)",
       y = "Total Cases",
       x = "Date") +
  theme_minimal()
```



The logistic growth model applied to COVID-19 case data demonstrates a statistically significant fit, with key parameters such as carrying capacity, growth rate, and initial cases showing strong relevance. Although there is some variation between predicted and actual cases, indicated by the residual standard error, the model achieves reliable convergence, effectively capturing the general trend of the pandemic. This suggests that the model is robust, despite inherent complexities like reporting variances and epidemiological factors that might cause deviations.

Conclusion

Effective methods for analyzing and interpreting COVID-19 data were demonstrated, showing how the virus spread and impacted different regions. Choropleth maps identified regions in the Northeast and South having the highest mortality rates due to the virus, likely influenced by factors such as high population density and early outbreak intensity. Conversely, the Midwest and Mountain West exhibited lower rates. Visualizations of the spread of the virus over time revealed initial severe impacts of the pandemic with sharp spikes in deaths, followed by a significant surge in cases in 2022 where, notably, death rates did not correspondingly increase, suggesting the efficacy of vaccinations. The predicted trend from the logistic growth model accurately followed the actual progression of the pandemic, effectively capturing the initial rapid increase in cases and the subsequent leveling off. Overall, these analyses underscored the importance of data-driven decision-making in managing health crises.

Possible Sources of Bias: Variations in data reporting can lead to inconsistencies, as regions differ in their methods and capabilities for tracking COVID-19 cases and deaths. The availability of testing also varies significantly, influencing reported case numbers; areas with more comprehensive testing might report higher numbers, but this doesn't necessarily indicate a greater spread of the virus. Additionally, disparities in healthcare quality can skew mortality rates, with better-equipped regions potentially showing

fewer deaths.Awareness of these potential biases is critical for ensuring accurate and reliable interpretations of the data.