COVID19 Data Report

2023-11-18

Dataset Description

This COVID-19 dataset is from the Johns Hopkins Github site and contains daily time series summary tables, including confirmed, deaths, and recovered. The COVID-19 data repository is operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). Since January 21, 2020, this dataset has collected data from sources such as the World Health Organization (WHO), Los Angeles Times, and QQ News, etc. On March 10, 2023, the Johns Hopkins Coronavirus Resource Center ceased its collecting and reporting of global COVID-19 data.

(Please refer to https://github.com/CSSEGISandData/COVID-19 for additional information about this dataset.)

Step 0: Import Packages

```
library(tidyverse)
library(forecast)
```

Step 1: Import the Data

- Copy the link address of the csv file.

```
# Get the beginning part of the link address
url_in = "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covi

# Get the file names
file_names = c("time_series_covid19_confirmed_global.csv", "time_series_covid19_deaths_global.csv")

# Use `str_c()` to put those together
urls = str_c(url_in, file_names)
urls
```

```
## [1] "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_
## [2] "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_
```

- Use `read_csv()` to read in the data.

```
global_cases = read_csv(urls[1])
```

```
## # A tibble: 6 x 1,147
##   `Province/State` `Country/Region`   Lat  Long `1/22/20` `1/23/20` `1/24/20`
##   <chr>            <chr>            <dbl> <dbl>     <dbl>     <dbl>     <dbl>
## 1 <NA>             Afghanistan       33.9 67.7          0         0         0
## 2 <NA>             Albania           41.2 20.2          0         0         0
## 3 <NA>             Algeria           28.0  1.66         0         0         0
## 4 <NA>             Andorra           42.5  1.52         0         0         0
## 5 <NA>             Angola           -11.2 17.9          0         0         0
## 6 <NA>             Antarctica       -71.9 23.3          0         0         0
## # i 1,140 more variables: `1/25/20` <dbl>, `1/26/20` <dbl>, `1/27/20` <dbl>,
## #   `1/28/20` <dbl>, `1/29/20` <dbl>, `1/30/20` <dbl>, `1/31/20` <dbl>,
## #   `2/1/20` <dbl>, `2/2/20` <dbl>, `2/3/20` <dbl>, `2/4/20` <dbl>,
## #   `2/5/20` <dbl>, `2/6/20` <dbl>, `2/7/20` <dbl>, `2/8/20` <dbl>,
## #   `2/9/20` <dbl>, `2/10/20` <dbl>, `2/11/20` <dbl>, `2/12/20` <dbl>,
## #   `2/13/20` <dbl>, `2/14/20` <dbl>, `2/15/20` <dbl>, `2/16/20` <dbl>,
## #   `2/17/20` <dbl>, `2/18/20` <dbl>, `2/19/20` <dbl>, `2/20/20` <dbl>, ...
```

```
global_deaths = read_csv(urls[2])
```

```
## # A tibble: 6 x 1,147
##   `Province/State` `Country/Region`   Lat  Long `1/22/20` `1/23/20` `1/24/20`
##   <chr>            <chr>            <dbl> <dbl>     <dbl>     <dbl>     <dbl>
## 1 <NA>             Afghanistan       33.9 67.7          0         0         0
## 2 <NA>             Albania           41.2 20.2          0         0         0
## 3 <NA>             Algeria           28.0  1.66         0         0         0
## 4 <NA>             Andorra           42.5  1.52         0         0         0
## 5 <NA>             Angola           -11.2 17.9          0         0         0
## 6 <NA>             Antarctica       -71.9 23.3          0         0         0
## # i 1,140 more variables: `1/25/20` <dbl>, `1/26/20` <dbl>, `1/27/20` <dbl>,
## #   `1/28/20` <dbl>, `1/29/20` <dbl>, `1/30/20` <dbl>, `1/31/20` <dbl>,
## #   `2/1/20` <dbl>, `2/2/20` <dbl>, `2/3/20` <dbl>, `2/4/20` <dbl>,
## #   `2/5/20` <dbl>, `2/6/20` <dbl>, `2/7/20` <dbl>, `2/8/20` <dbl>,
## #   `2/9/20` <dbl>, `2/10/20` <dbl>, `2/11/20` <dbl>, `2/12/20` <dbl>,
## #   `2/13/20` <dbl>, `2/14/20` <dbl>, `2/15/20` <dbl>, `2/16/20` <dbl>,
## #   `2/17/20` <dbl>, `2/18/20` <dbl>, `2/19/20` <dbl>, `2/20/20` <dbl>, ...
```

Step 2: Tidy and Transform the Data

1. Tidy the columns

- Put each variable (**date**, **cases**, and **deaths**) in their own column.
- Remove columns: **Lat** and **Long**.
- Rename columns: **Province/State** and **Country/Region**.
- Convert column **date** to date object.

```
# Use `pivot_longer()` to make each date on a separate row
tidy_cases = global_cases %>%
  pivot_longer(cols = -c(`Province/State`, `Country/Region`, Lat, Long), names_to = "date",
               values_to = "cases")
```

```
tidy_deaths = global_deaths %>%
  pivot_longer(cols = -c(`Province/State`, `Country/Region`, Lat, Long), names_to = "date",
               values_to = "deaths")
```

```
# Use `full_join()` to combine the datasets
global_tidy = tidy_cases  %>%
  full_join(tidy_deaths)  %>%
  select(-c(Lat, Long)) %>%
  rename(Country_Region = `Country/Region`, Province_State = `Province/State`) %>%
  mutate(date = mdy(date))
```

2. Tidy the rows

- Filter the rows of **Country_Region** of Taiwan*.

```
# Filter out the rows
# Because Taiwan* has no provinces or states, remove column Province_State
taiwan_tidy = global_tidy %>%
  filter(Country_Region == "Taiwan*") %>%
  select(-c(Province_State))
```

```
summary(taiwan_tidy)
```

```
##  Country_Region          date                cases              deaths
##  Length:1143        Min.   :2020-01-22   Min.   :      1   Min.   :    0
##  Class :character   1st Qu.:2020-11-02   1st Qu.:    565   1st Qu.:    7
##  Mode  :character   Median :2021-08-15   Median :  15852   Median :  821
##                     Mean   :2021-08-15   Mean   :1720368   Mean   : 3214
##                     3rd Qu.:2022-05-27   3rd Qu.:1775385   3rd Qu.: 1866
##                     Max.   :2023-03-09   Max.   :9970937   Max.   :17672
```

Step 3: Add Visualizations and Analysis

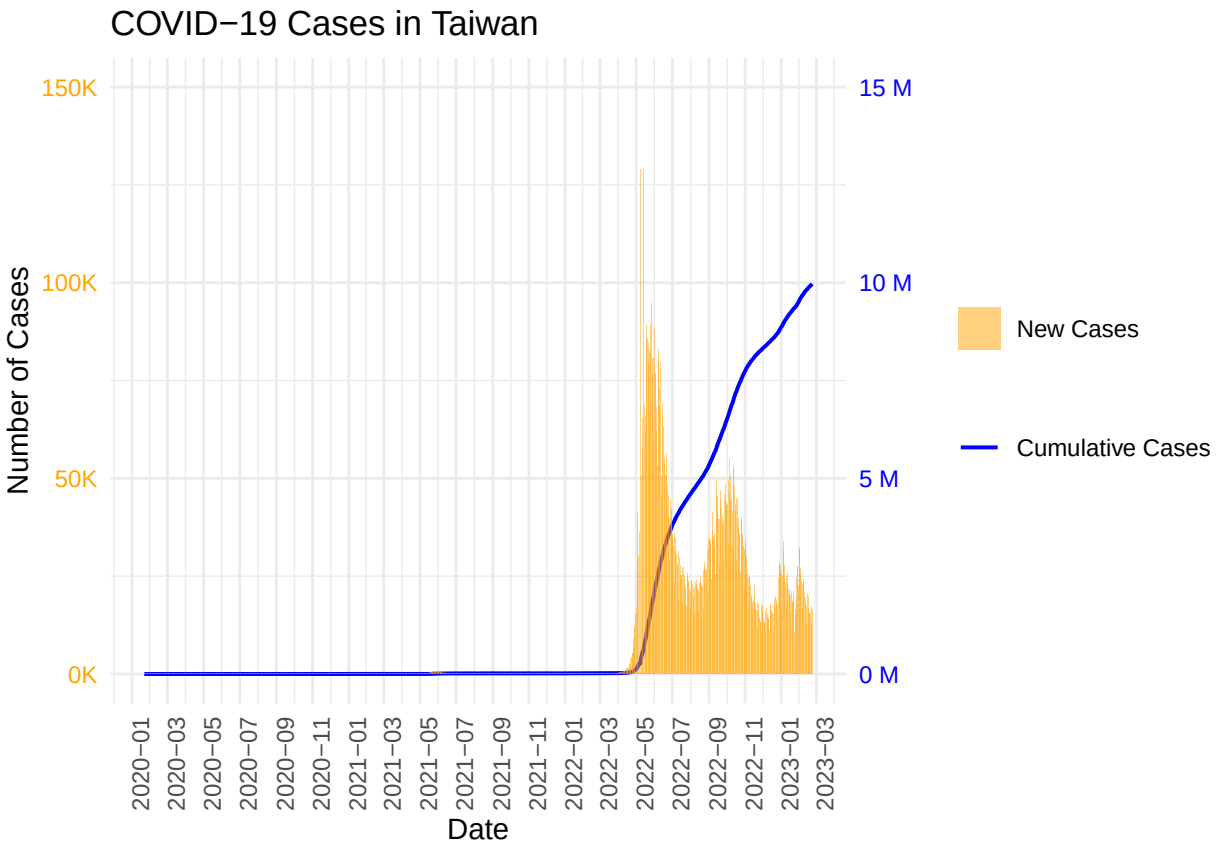
Question 1: What are the trends for daily cumulative confirmed cases and new confirmed cases of COVID-19 in Taiwan?

- Since the spread of the COVID-19 pandemic in **January 2020**, Taiwan did not see a significant surge in confirmed cases until **April 2022**.
- After the outbreak of a large-scale epidemic, there were higher cases of infections in **September 2022** and **January 2023**, but there has been an overall downward trend.

```
# Calculate the new cases
# Handle NA and negative values
taiwan_tidy = taiwan_tidy %>%
  mutate(new_cases = cases - lag(cases)) %>%
  mutate(new_cases = ifelse(is.na(new_cases) | new_cases < 0, 0, new_cases))

# Print and check the tail
tail(taiwan_tidy %>% select(new_cases, everything()))
```

```
## # A tibble: 6 x 5
##   new_cases Country_Region date         cases deaths
##       <dbl> <chr>          <date>       <dbl>  <dbl>
## 1         0 Taiwan*        2023-03-04 9970937  17672
## 2         0 Taiwan*        2023-03-05 9970937  17672
## 3         0 Taiwan*        2023-03-06 9970937  17672
## 4         0 Taiwan*        2023-03-07 9970937  17672
## 5         0 Taiwan*        2023-03-08 9970937  17672
## 6         0 Taiwan*        2023-03-09 9970937  17672
```

```r
# Since there are no new cases updated, remove rows at the tail end that do not update
taiwan_tidy = taiwan_tidy %>%
  filter(!(cases == 9970937 & new_cases == 0))

# For charting purpose, convert cumulative and new cases into units of thousand
taiwan_tidy$cases_100k = taiwan_tidy$cases / 100000
taiwan_tidy$new_cases_k = taiwan_tidy$new_cases / 1000

# Create a chart
ggplot(taiwan_tidy, aes(x = date)) +
  geom_line(aes(y = cases_100k, color = "Cumulative Cases"), linewidth = 0.7) +
  geom_bar(aes(y = new_cases_k, fill = "New Cases"), stat = "identity", alpha = 0.5) +
  labs(x = "Date", y = "Number of Cases") +
  scale_color_manual(values = c("Cumulative Cases" = "blue")) +
  scale_fill_manual(values = c("New Cases" = "orange")) +
  ggtitle("COVID-19 Cases in Taiwan") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        axis.text.y.right = element_text(color = "blue"),
        axis.text.y.left = element_text(color = "orange"),
        axis.title.y.left = element_text(color = "black")) +
  scale_y_continuous(
    sec.axis = sec_axis(~.*100000, labels = scales::unit_format(unit = "M", scale = 1e-6)),
    limits = c(0, 150),
    breaks = seq(0, 150, by = 50),labels = function(x) paste0(x, "K")
    ) +
  scale_x_date(date_labels = "%Y-%m", date_breaks = "2 month") +
  guides(color = guide_legend(title = NULL), fill = guide_legend(title = NULL))
```

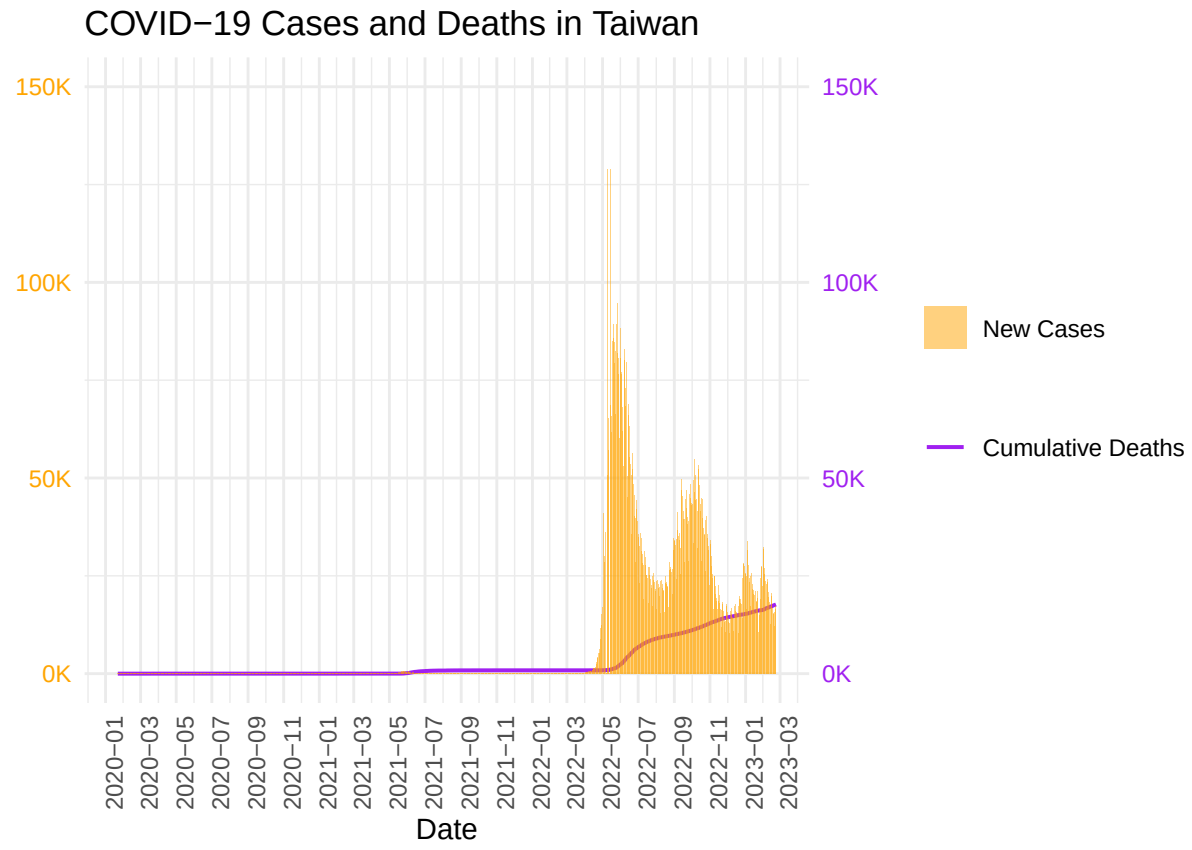


Question 2: What are the trends for daily cumulative deaths and new confirmed cases of COVID-19 in Taiwan?

- The number of deaths has been increasing alongside the rising number of confirmed cases.

```
# Convert cumulative deaths into units of thousand
taiwan_tidy$deaths_k = taiwan_tidy$deaths / 1000

# Create a chart
ggplot(taiwan_tidy, aes(x = date)) +
  geom_line(aes(y = deaths_k, color = "Cumulative Deaths"), linewidth = 0.7) +
  geom_bar(aes(y = new_cases_k, fill = "New Cases"), stat = "identity", alpha = 0.5) +
  labs(x = "Date", y = " ") +
  scale_color_manual(values = c("Cumulative Deaths" = "purple")) +
  scale_fill_manual(values = c("New Cases" = "orange")) +
  ggtitle("COVID-19 Cases and Deaths in Taiwan") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),
        axis.text.y.right = element_text(color = "purple"),
        axis.text.y.left = element_text(color = "orange")) +
  scale_y_continuous(
    sec.axis = sec_axis(~., labels = function(x) paste0(x, "K")),
    limits = c(0, 150),
    breaks = seq(0, 150, by = 50),labels = function(x) paste0(x, "K")
    ) +
```

```
  scale_x_date(date_labels = "%Y-%m", date_breaks = "2 month") +
  guides(color = guide_legend(title = NULL), fill = guide_legend(title = NULL))
```

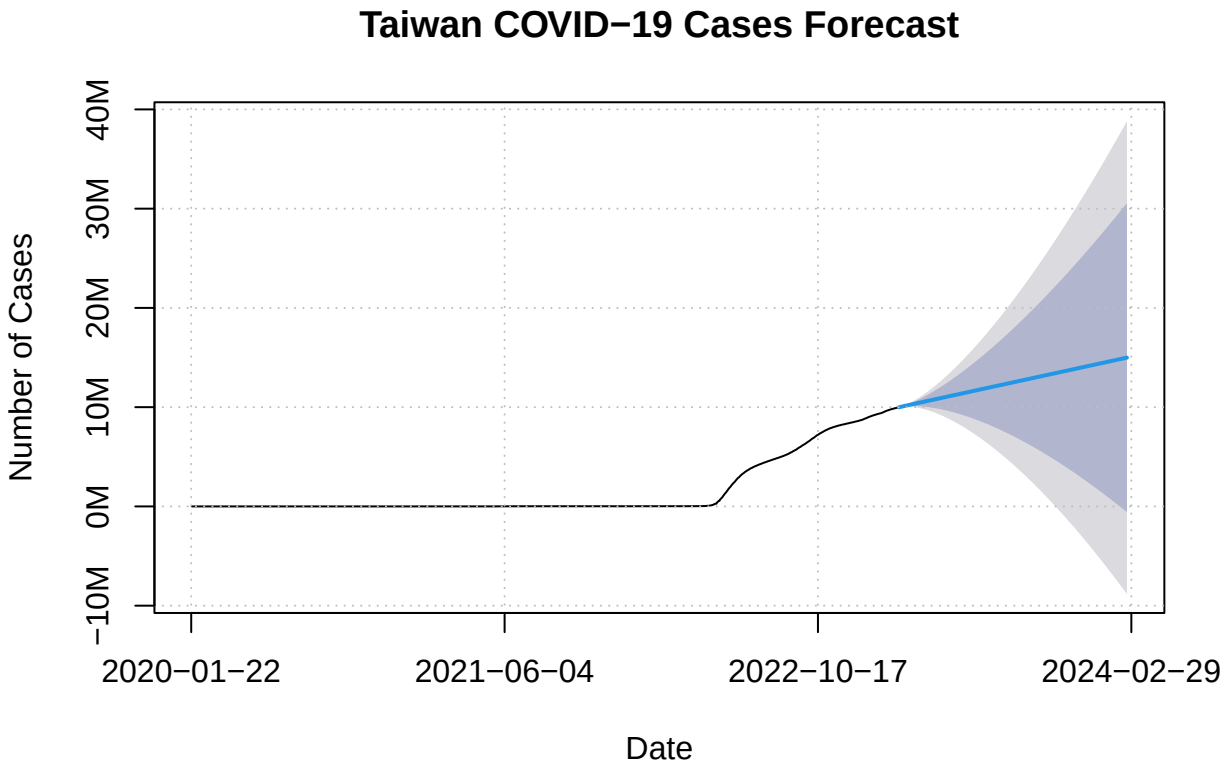


Question 3: Can we predict the future number of confirmed cases in Taiwan?

- **Purpose**: Predict the future number of COVID-19 confirmed cases in Taiwan for the upcoming year based on the data collected by JHU CSSE.
- **Methods**: Use a **ARIMA model** to model and forecast.
  - Use `auto.arima()` to build a time series model.
  - Use `forecast()` to predict future data.

```
# Use `ts()` convert data into a time series object
ts_cases = ts(taiwan_tidy$cases)

# ARIMA model
arima_model = auto.arima(ts_cases)

# Make predictions using the established ARIMA model
future_forecast = forecast(arima_model, h = 365)

# Create a chart
plot(future_forecast, main = "Taiwan COVID-19 Cases Forecast", yaxt = "n", xaxt = "n")
grid(lty = "dotted", col = "gray")
```

```r
# Draw the y-axis labels
y_labels = c(-10, 0, 10, 20, 30, 40) * 1e6
axis(2, at = y_labels, labels = paste0(y_labels / 1e6, "M"))

# Draw the x-axis labels
x_labels = c(0, 500, 1000, 1500)
x_labels_dates = c(taiwan_tidy$date[1], taiwan_tidy$date[1] + 499, taiwan_tidy$date[1] + 999,
                   taiwan_tidy$date[1] + 1499)
axis(1, at = x_labels, labels = paste0(x_labels_dates))

# Add text to x-axis and y-axis
mtext("Date", side = 1, line = 3)
mtext("Number of Cases", side = 2, line = 3)
```



Step 4: Add Bias Identification

1. Personal bias

- **Before analysis**: Due to having previously observed the global COVID-19 trends, I might have assumed that the situation in Taiwan had also eased. However, the data shows that due to Taiwan's later onset of a severe outbreak, while the global situation has been improving, Taiwan's situation hasn't followed suit yet.
- **After analysis**: Due to my limited experience, it might lead to overlooking crucial data or neglecting alternative explanations when interpreting results.

2. Other bias

- **Reporting Bias**: Reporting systems can vary across different regions, including differences in reporting times, methods, and accuracy. Some areas might report data more promptly, while others could experience delays or underreporting in their data.
- **Temporal Bias**: The pandemic evolves over time, and data from different stages can be influenced by factors such as implementation of measures, improved testing capabilities, or changes in societal behavior.

Conclusion

In summary, the confirmed cases and fatalities of COVID-19 in Taiwan have increased over time, but there's a declining trend in daily new confirmed cases. This report analyzed COVID-19 data from January 2020 to March 2023, presenting two data visualizations and one model for predictive purposes, offering insights into the future trajectory of the pandemic.