Understanding COVID-19 Dynamics: Analyzing Total Cases, Deaths, and Daily Incremental Changes

Oliviu Lazar

2024-04-24

Import libraries

Import necessary libraries

```
Sys.setenv("VROOM_CONNECTION_SIZE" = 5000000)

library(tidyverse)
library(ggplot2)
library(dplyr)
library(tidyr)
library(forecast)
library(tseries)
```

Data description

The data comes from the data repository for the 2019 Novel Corona virus Visual Dashboard operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). It includes data on COVID-19 cases, deaths, and recoveries up to March 10, 2023, when the Johns Hopkins Corona virus Resource Center ceased its collecting and reporting of global COVID-19 data.

Importing data for analysis

Import COVID-19 data by constructing URLs to access time series data files hosted on GitHub.

```
domain <- "https://raw.githubusercontent.com/"
subdir <- "CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/"

covid_19_url <- paste0(domain, subdir)

file_names <- c("time_series_covid19_confirmed_US.csv",
                "time_series_covid19_confirmed_global.csv",
                "time_series_covid19_deaths_US.csv",
                "time_series_covid19_deaths_global.csv")

urls <- str_c(covid_19_url, file_names)
```

```
us_cases_raw <- read_csv(urls[1], show_col_types = FALSE)
global_cases_raw <- read_csv(urls[2], show_col_types = FALSE)
us_deaths_raw <- read_csv(urls[3], show_col_types = FALSE)
global_deaths_raw <- read_csv(urls[4], show_col_types = FALSE)
```

Clean, pivot, and summarize data

Clean data by removing unwanted columns, and convert date strings to R Date objects. Pivot dates from wide to long format to organize data into columns for provinces/states, countries/regions, dates, and corresponding case or death counts. Next, I combine the US and Global cases and deaths into one data frame for analysis and add two additional columns for daily incremental cases/deaths. Finally, I summarize the total cases/deaths data frame to provide an overview of the data structure and content.

```
us_cases_pivot <- us_cases_raw %>%
  select(-c('UID', 'iso2', 'iso3', 'code3', 'FIPS', 'Admin2', 'Combined_Key', 'Lat', 'Long_')) %>%
  pivot_longer(cols = -c('Province_State', 'Country_Region'),
               names_to = "Date", values_to = "Cases")  %>%
  select(c(
    "Province/State" = "Province_State",
    "Country/Region" = "Country_Region",
    "Date",
    "Cases")) %>%
  mutate(
    Date = mdy(Date),
  )

global_cases_pivot <- global_cases_raw %>%
  select(-c('Lat', 'Long')) %>%
  pivot_longer(cols = -c('Province/State', 'Country/Region'),
               names_to = "Date", values_to = "Cases") %>%
  mutate(
    Date = mdy(Date),
  )

us_deaths_pivot <- us_deaths_raw %>%
  select(-c('UID', 'iso2', 'iso3', 'code3', 'FIPS', 'Admin2', 'Combined_Key', 'Lat', 'Long_', 'Populati
  pivot_longer(cols = -c('Province_State', 'Country_Region'),
               names_to = "Date", values_to = "Cases")  %>%
  select(c(
    "Province/State" = "Province_State",
    "Country/Region" = "Country_Region",
    "Date",
    "Cases")) %>%
  mutate(
    Date = mdy(Date),
  )

global_deaths_pivot <- global_deaths_raw %>%
  select(-c('Lat', 'Long')) %>%
  pivot_longer(cols = -c('Province/State', 'Country/Region'),
               names_to = "Date", values_to = "Cases") %>%
  mutate(
    Date = mdy(Date),
  )

cases <- rbind(us_cases_pivot, global_cases_pivot)
deaths<- rbind(us_deaths_pivot, global_deaths_pivot)

total_cases <- cases %>%
  group_by(Date) %>%
  summarise(Total_Cases = sum(Cases, na.rm = TRUE))

total_deaths <- deaths %>%
  group_by(Date) %>%
  summarise(Total_Deaths = sum(Cases, na.rm = TRUE))

# Combine the data frames into one to include cases and deaths
total_cases_deaths <- merge(total_cases, total_deaths, by = "Date", all = TRUE)

# At two columns for daily incremental cases
total_cases_deaths <- total_cases_deaths %>%
  mutate(Daily_Incremental_Cases = c(NA, diff(Total_Cases)),
         Daily_Incremental_Deaths = c(NA, diff(Total_Deaths)))

summary(total_cases_deaths)
```

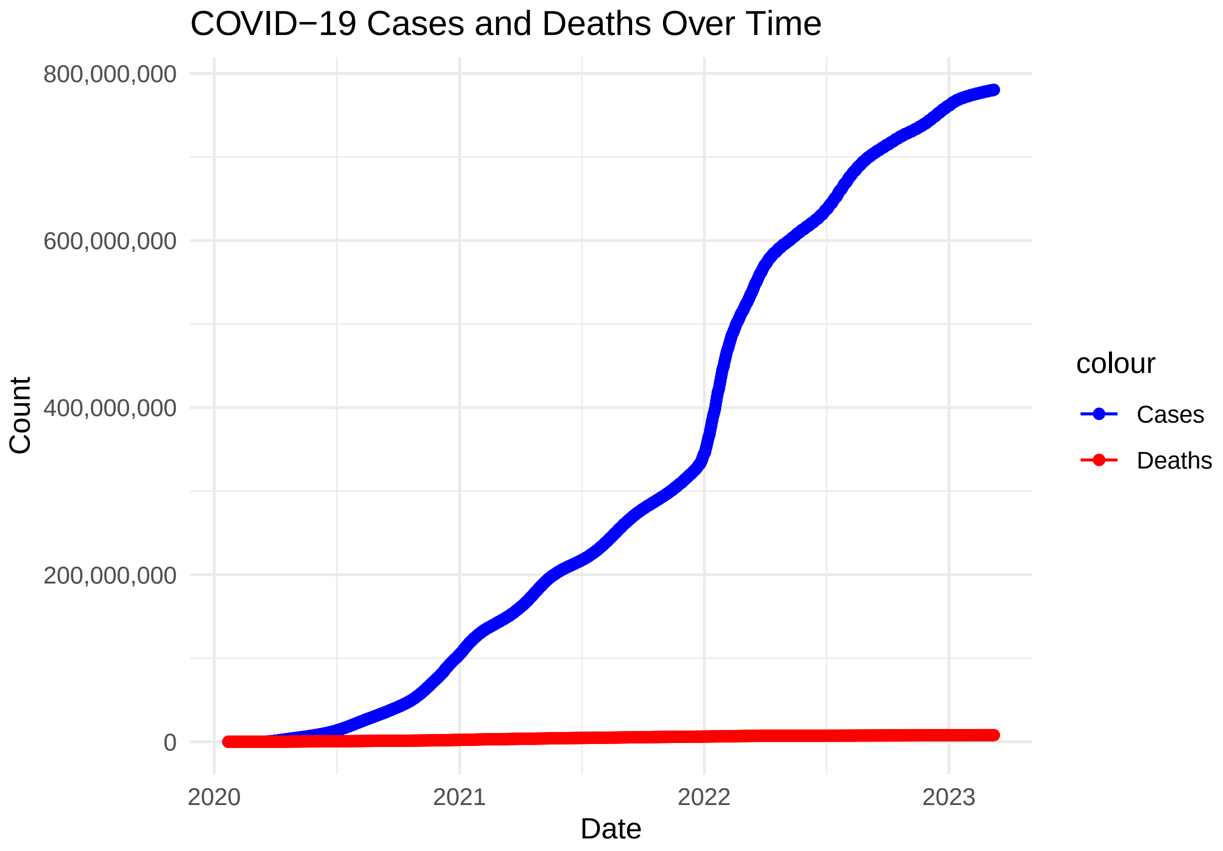
```
##       Date             Total_Cases          Total_Deaths
##  Min.   :2020-01-22   Min.   :      558   Min.   :     18
##  1st Qu.:2020-11-02   1st Qu.: 56827940   1st Qu.:1514979
##  Median :2021-08-15   Median :244661351   Median :5006729
##  Mean   :2021-08-15   Mean   :324342646   Mean   :4491419
##  3rd Qu.:2022-05-27   3rd Qu.:612914342   3rd Qu.:7319327
##  Max.   :2023-03-09   Max.   :780372851   Max.   :8005638
##
##  Daily_Incremental_Cases Daily_Incremental_Deaths
##  Min.   :    100         Min.   :    1
##  1st Qu.: 308072         1st Qu.: 2702
##  Median : 547477         Median : 6751
##  Mean   : 683338         Mean   : 7010
##  3rd Qu.: 811112         3rd Qu.:10202
##  Max.   :4993126         Max.   :60920
##  NA's   :1               NA's   :1
```

Plot total cases and deaths

Plot the trends of COVID-19 cases and deaths over time. Although the number of reported cases shows a steady increase or periodic spikes, reported deaths remain relatively low compared to cases. Note that there are no missing values to handle.

```
ggplot(total_cases_deaths, aes(x = Date)) +
  geom_line(aes(y = Total_Cases, color = "Cases")) +
  geom_line(aes(y = Total_Deaths, color = "Deaths")) +
```

```r
  geom_point(aes(y = Total_Cases, color = "Cases")) +
  geom_point(aes(y = Total_Deaths, color = "Deaths")) +
  labs(x = "Date", y = "Count", title = "COVID-19 Cases and Deaths Over Time") +
  scale_color_manual(values = c("Cases" = "blue", "Deaths" = "red")) +
  scale_y_continuous(labels = scales::comma_format()) + # Disable scientific notation
  theme_minimal()
```

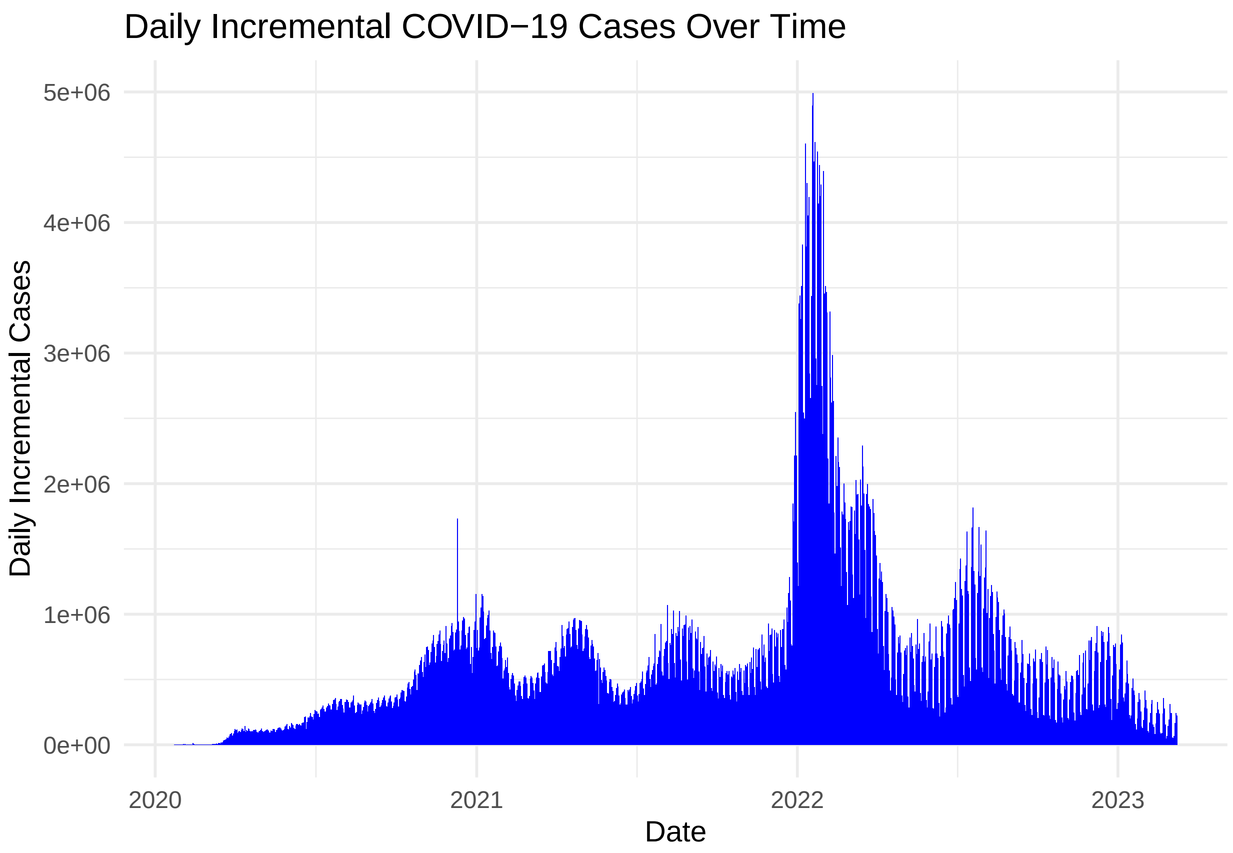


Plot daily incremental cases and deaths

Visualize the daily changes in cases and deaths to analyze the rate of change over time. The plots indicate that while daily incremental cases exhibit fluctuations and spikes, the corresponding increases in daily deaths are less pronounced. This observation suggests that measures aimed at managing the spread of COVID-19, such as vaccination campaigns, enhanced healthcare protocols, and public health interventions, may have effectively lowered mortality rates despite spikes in cases. Note that the first row is removed in the analysis since there is no incremental value for the first day.

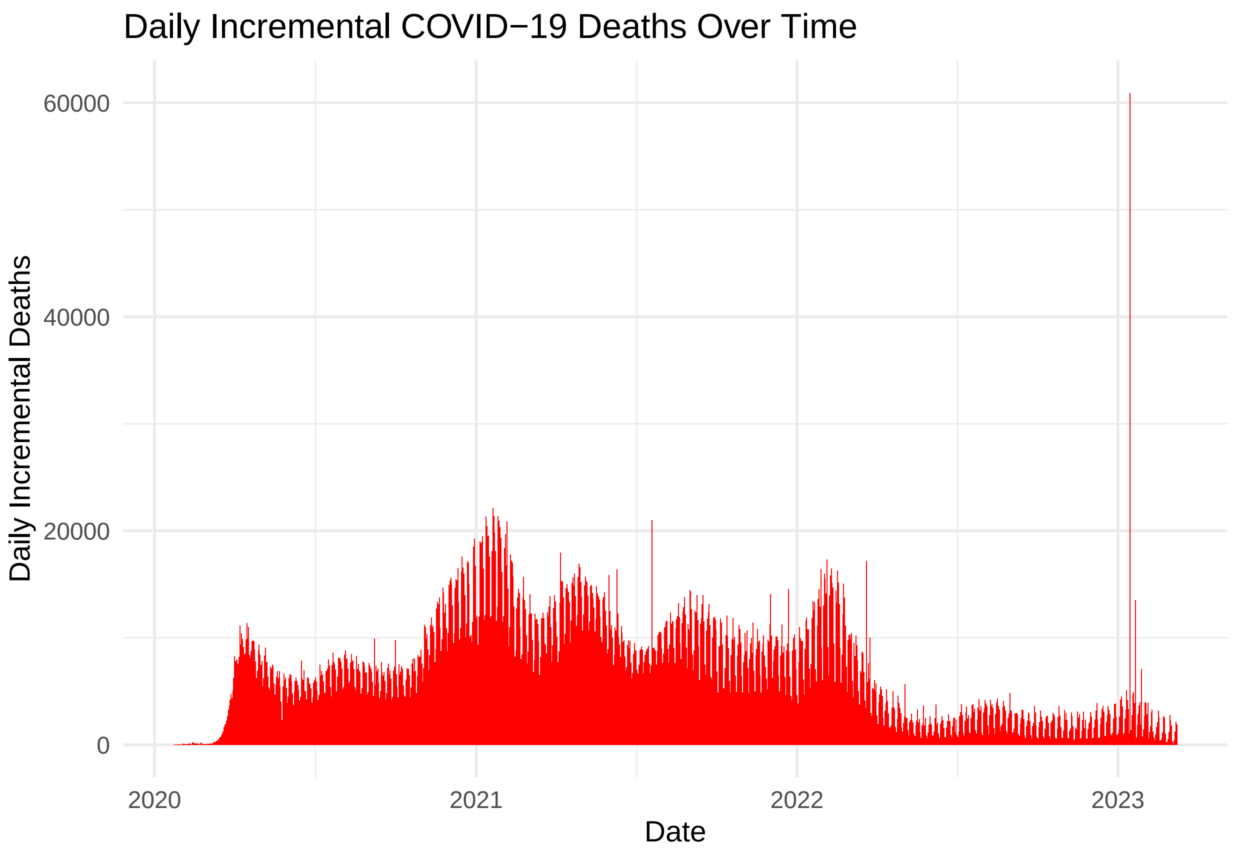
```r
ggplot(total_cases_deaths, aes(x = Date, y = Daily_Incremental_Cases)) +
  geom_bar(stat = "identity", fill = "blue") +
  labs(x = "Date", y = "Daily Incremental Cases", title = "Daily Incremental COVID-19 Cases Over Time")
  theme_minimal()
```

```
## Warning: Removed 1 rows containing missing values (‘position_stack()‘).
```



```
ggplot(total_cases_deaths, aes(x = Date, y = Daily_Incremental_Deaths)) +
  geom_bar(stat = "identity", fill = "red") +
  labs(x = "Date", y = "Daily Incremental Deaths", title = "Daily Incremental COVID-19 Deaths Over Time
  theme_minimal()
```

```
## Warning: Removed 1 rows containing missing values (‘position_stack()‘).
```



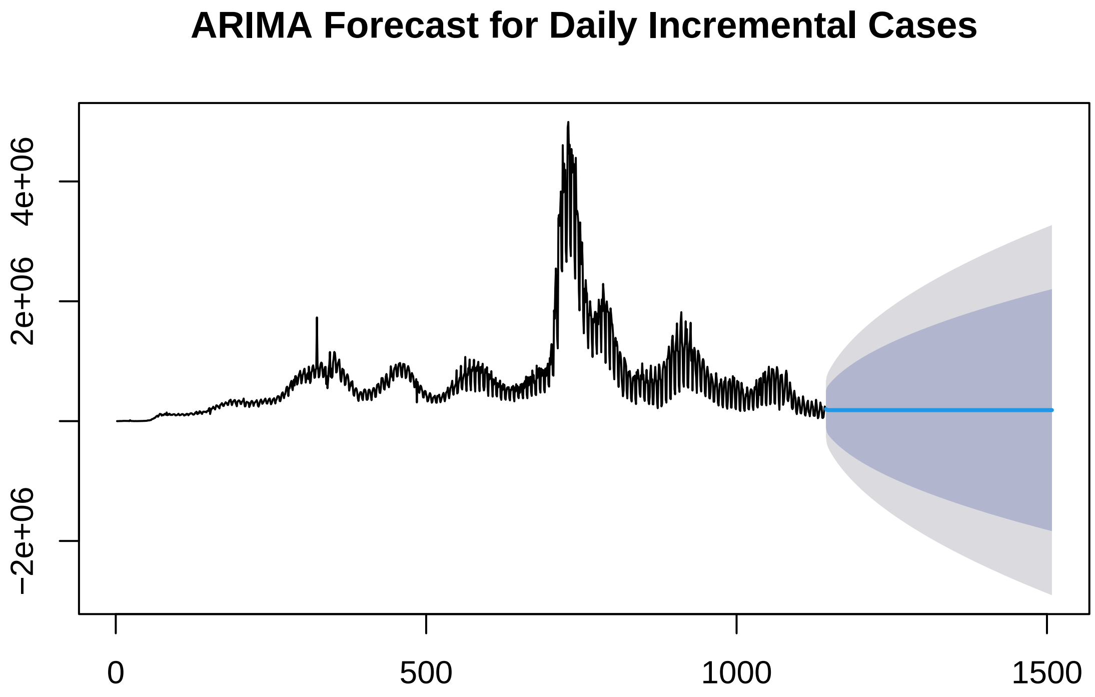
Arima time series model

Fit ARIMA model to the daily incremental cases and deaths data, generate forecasts for the next year (365 days) using these models, and then plot the forecasts.The ARIMA models provide additional validation trends and pattern mentioned above. Further the plots provide prediction intervals to help in forecasting future values for cases and deaths.

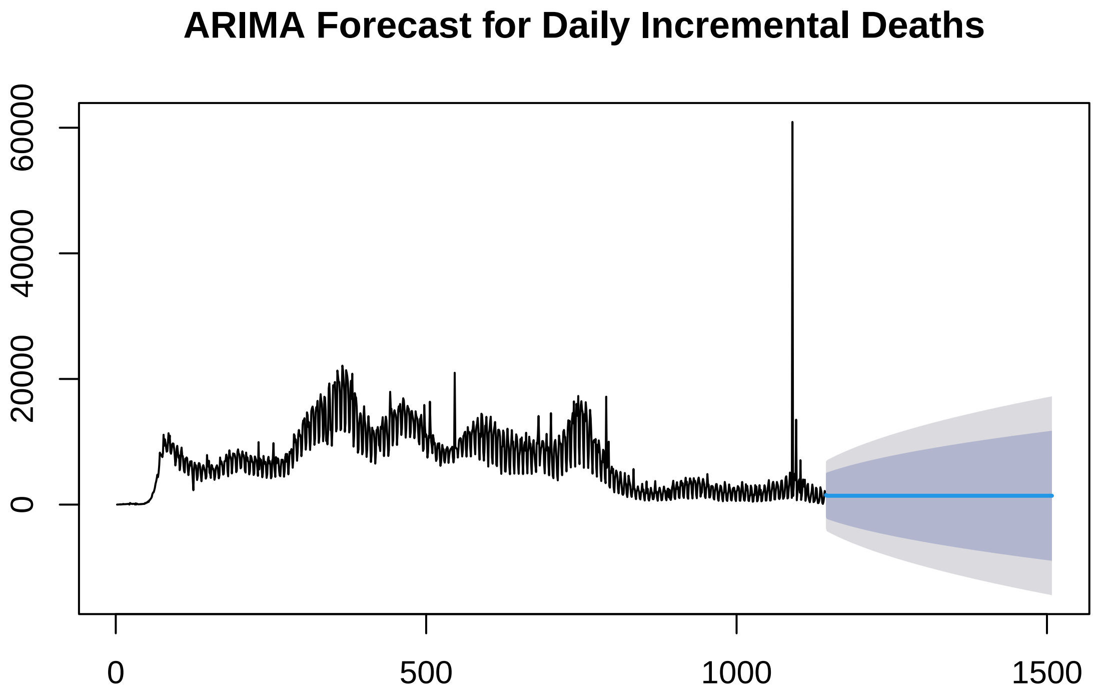
```
arima_model_cases <- arima(total_cases_deaths$Daily_Incremental_Cases, order = c(1,1,1))
arima_model_deaths <- arima(total_cases_deaths$Daily_Incremental_Deaths, order = c(1,1,1))

# Generate forecast for the next year (365 days)
arima_model_cases <- forecast(arima_model_cases, h = 365)
arima_model_deaths <- forecast(arima_model_deaths, h = 365)

# Plot forecast
plot(arima_model_cases, main = "ARIMA Forecast for Daily Incremental Cases")
```



```
plot(arima_model_deaths, main = "ARIMA Forecast for Daily Incremental Deaths")
```



Sources of Bias

Below are some potential sources of bias. Recognizing and addressing these sources of bias is critical for conducting rigorous and unbiased COVID-19 data analysis.

Selection Bias: Certain groups may be over represented in COVID-19 data due to differential access to testing and healthcare services.

Sampling Bias: Biases in sample selection may distort estimates of COVID-19 prevalence, transmission rates, and outcomes.

Social Bias: Societal norms, cultural values, and political ideologies can influence data collection, reporting, and interpretation practices.

Measurement Bias: Errors in diagnostic testing, misclassification of cases or deaths, and variations in data collection procedures can introduce inaccuracies into COVID-19 data.

Personal Bias: Individual beliefs, attitudes, or experiences can shape judgment and decision-making in COVID-19 data analysis, potentially introducing subjective analysis and interpretations.

Conclusion

The analysis of COVID-19 data has revealed insights into the trends of total cases and deaths, as well as the patterns in daily incremental changes. By examining the combined US and Global data, we've observed fluctuations and spikes in daily incremental cases over time, while corresponding increases in daily incremental deaths are comparatively less pronounced. This suggests potential effectiveness of measures aimed at managing the spread of the virus, such as vaccination campaigns and improved healthcare protocols, in reducing mortality rates even during periods of increased cases.