COVID-19

2024-10-13

Importing Libraries

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

Loading Data

```
global_deaths_url <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/refs/heads/master/csse_covid_19_d
ata/csse_covid_19_time_series/time_series_covid19_deaths_global.csv"

global_deaths <- read.csv(global_deaths_url)
```

Introduction and Question of Interest

The data used in the report below describes a time series dataset regarding global COVID-19 related deaths from January 22, 2020, to March 9, 2023. This dataset breakdowns cases to the province/state level for a few countries, namely Australia, Canada, China, Denmark, France, and the United Kingdom.

This report will analyze the change in frequency of COVID-19 related deaths over time in the regions most affected by the diseases as described by the dataset. To this end some relevant visualization will provided, such as line charts and area charts showcasing total deaths over time and changes in month-to-month percentages of deaths.

Additionally, an ARIMA model will be used to predict the future trends of COVID-19 related deaths in the region with the highest total number of reported COVID-19 related deaths. This will provide insight into how accurately the data provided up to March 9, 2023 serve to predict the actual outcome.

Data Cleaning and Data Description For Visualizations

The visualization portion of this report will work with the time series data of describing global deaths, to that end the columns used in this analysis will be the following:

1. **Country/Region:** A column describing the Country/Region the data relates to.
2. **Province/State:** A column describing the Province/State the data relates to. These will be aggregated into a their relevant country.
3. **All Date Columns:** Each day is provided as a unique column since this data is formatted as a time series dataset.

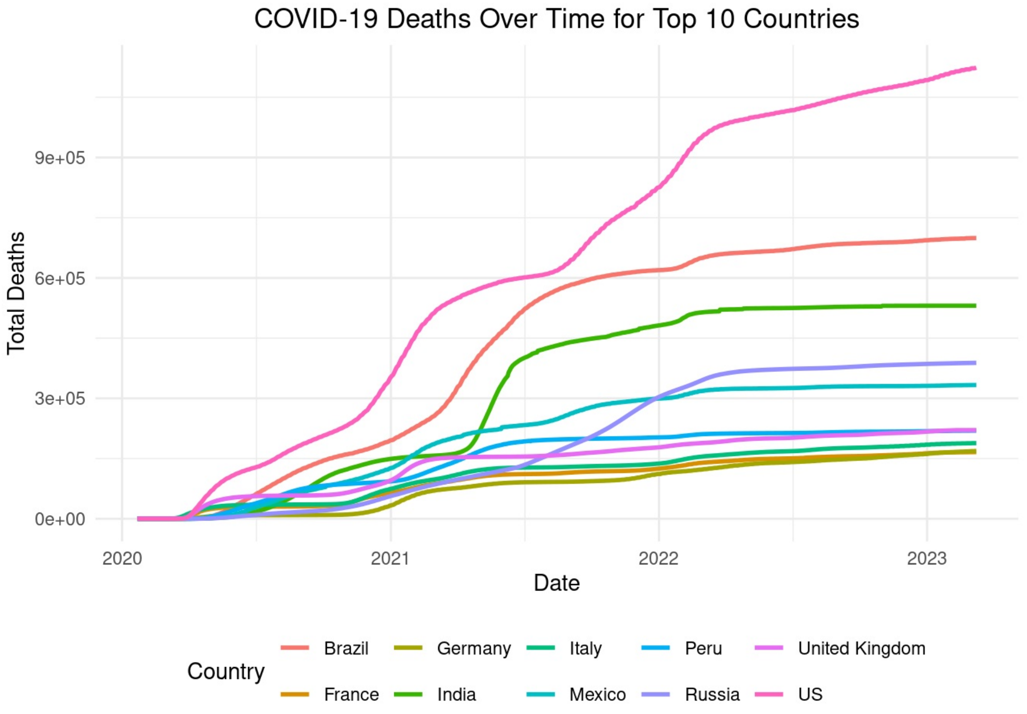
To compare all countries fairly we need to aggregate the countries that have been divided into state/provinces. Then the data will be pivoted to long format to facilitate working with the data.

Some additional data transformations will be needed to compare the month to month growth rates. Since the initial months skew the data too significantly to analyze visually, the growth rates analyzed will start on January 1, 2021.

```
# Aggregating by country
deaths_agg <- global_deaths %>%
  # Group by Country/Region to sum rows for the same country
  group_by(`Country.Region`) %>%
  # Summarize (sum) all columns that represent daily deaths/cases
  summarise(across(where(is.numeric), sum, na.rm = TRUE)) %>%
  ungroup()

# Selecting top 10
top_10_countries <- deaths_agg %>%
  arrange(desc(`X3.9.23`)) %>%
  slice(1:10)

# Pivoting the data to long format and exclusing the Lat and Long columns
deaths_agg <- top_10_countries %>%
  pivot_longer(cols = -c(`Country.Region`, Lat, Long),
               names_to = "date",
               values_to = "deaths") %>%
  # Removing the "X" prefix from the date column and convert to Date format
  mutate(date = gsub("X", "", date),
         date = mdy(date))

##### Used only for growth rate

# Filtering data to 2021 for growth rates
deaths_agg_2021 <- deaths_agg %>%
  filter(date >= as.Date("2021-01-01"))

# Groups data by year and month
deaths_top10_monthly <- deaths_agg_2021 %>%
  mutate(year_month = floor_date(date, "month")) %>%
  group_by(`Country.Region`, year_month) %>%
  summarise(total_deaths = sum(deaths)) %>%
  arrange(`Country.Region`, year_month)

# Calculates percentage growth for each month per country
deaths_top10_growth <- deaths_top10_monthly %>%
  group_by(`Country.Region`) %>%
  mutate(perc_growth = (total_deaths - lag(total_deaths)) / lag(total_deaths) * 100) %>%
  filter(!is.na(perc_growth))
```

Visualizations and Analysis

```
# Line graph
ggplot(deaths_agg, aes(x = date, y = deaths, color = `Country.Region`, group = `Country.Region`)) +
  geom_line(size = 1) +
  labs(title = "COVID-19 Deaths Over Time for Top 10 Countries",
       x = "Date",
       y = "Total Deaths",
       color = "Country") +
  theme_minimal() +
  theme(legend.position = "bottom",
        plot.title = element_text(hjust = 0.5))
```



This line chart shows the cumulative number of COVID-19 deaths over time for the top 10 countries. Each line represents a country, and the y-axis reflects the total number of deaths over time.

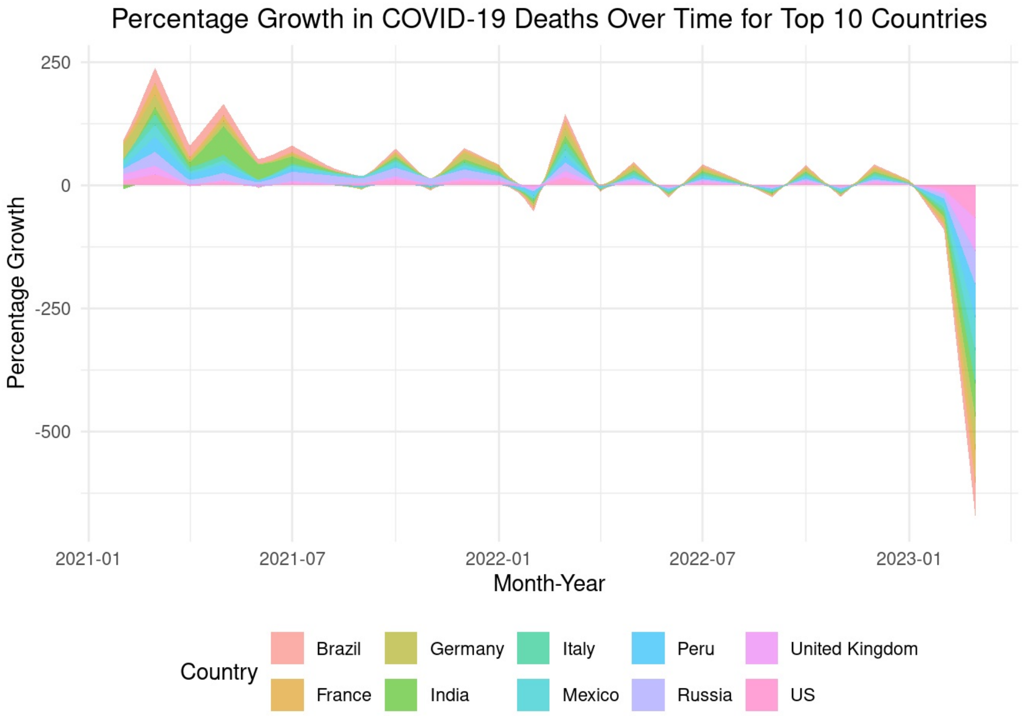
The United States exhibits a significant and steady increase in deaths, particularly in early 2021. The steep rise indicates that the country was dealing with a large number of COVID-19 deaths at the beginning of the year, possibly linked to the surge following the holiday season and delays in vaccine distribution.

India experienced a sharp and sudden increase in deaths starting around April-May 2021, corresponding to the devastating second wave driven by the Delta variant. This sudden rise sharply contrasts with the relatively steady growth seen in other countries up to that point.

Brazil shows a consistent increase in deaths throughout the period, with less fluctuation compared to India, but consistently high. This reflects the ongoing struggle in Brazil with multiple COVID-19 waves and challenges in managing the outbreak.

These countries show similar patterns, with high initial death tolls early in the pandemic followed by slower increases later in 2021. The stabilization in deaths likely reflects improved healthcare responses and vaccine rollouts.

```
# Area chart
ggplot(deaths_top10_growth, aes(x = year_month, y = perc_growth, fill = `Country.Region`, group = `Country.Region
`)) +
  geom_area(alpha = 0.6) +
  labs(title = "Percentage Growth in COVID-19 Deaths Over Time for Top 10 Countries",
       x = "Month-Year",
       y = "Percentage Growth",
       fill = "Country") +
  theme_minimal() +
  theme(legend.position = "bottom",
        plot.title = element_text(hjust = 0.5))
```



This area chart represents the cumulative percentage growth of COVID-19 deaths across the top 10 countries, using filled areas to highlight the contribution of each country over time.

Some countries, such as the USA, Brazil, and India, have more pronounced areas during certain months, reflecting periods when they contributed significantly to the global percentage growth in deaths. For example, the rapid growth seen in India during mid-2021 visually dominates the chart, corresponding to the Delta wave.

The chart shows how different countries' contributions to overall percentage growth shifted. For instance, the United States dominates early 2021, but its share decreases as countries like India and Brazil see sharp growth later in the year.

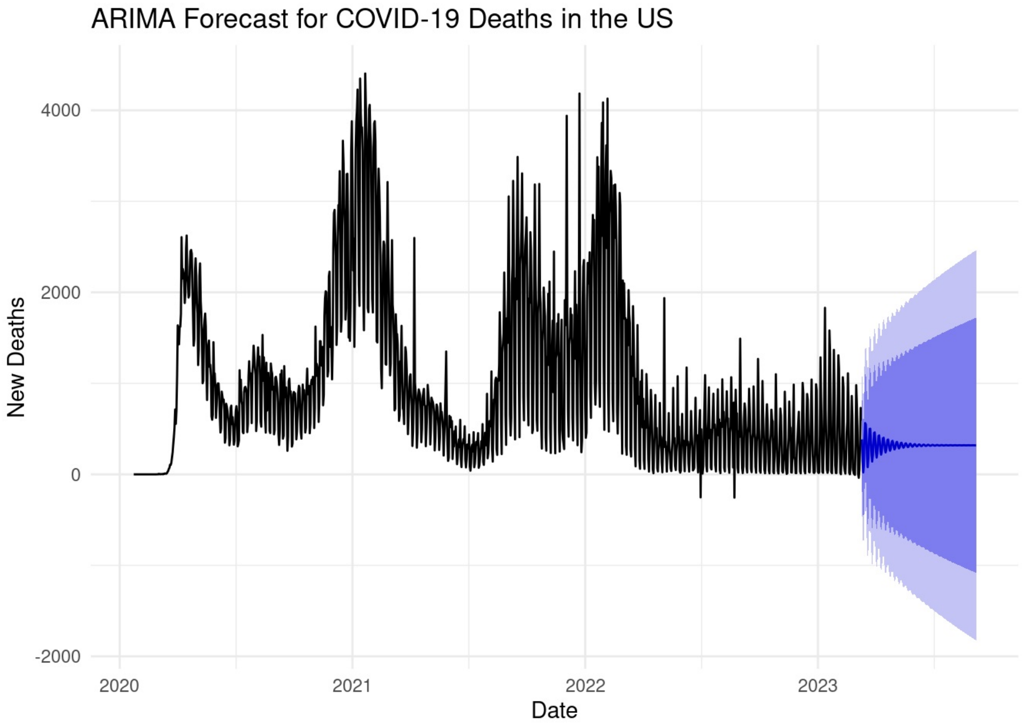
The rising and falling areas indicate seasonal or wave-like patterns in the pandemic. India's sharp increase in early 2021 and Brazil's sustained contributions suggest regional outbreaks that significantly influenced global trends.

Data Cleaning For Model

```
# Filtering the data for the US, removing unused columns and calculating new deaths each day
us_data <- deaths_agg %>%
  filter(`Country.Region` == "US") %>%
  mutate(new_deaths = deaths - lag(deaths))%>%
  select(-c(`Country.Region`, Lat, Long,deaths))

# Creates time series objects
us_deaths_ts <- ts(us_data$new_deaths, start = c(year(min(us_data$date)), yday(min(us_data$date))), frequency = 3
65)
```

Model Set-up and Evaluation

```
# Fitting the ARIMA model
us_arima_model <- auto.arima(us_deaths_ts)

# Forecasting the next 30 days
us_forecast <- forecast(us_arima_model, h = 180)

# Plot the forecast with the actual data
autoplot(us_forecast) +
  labs(title = "ARIMA Forecast for COVID-19 Deaths in the US",
       x = "Date",
       y = "New Deaths") +
  theme_minimal()
```



The area in blue is the data predicted by the model for the six months following the end of the historical data. The model predicts a decreasing level number of new incidents initially, and then it plateaus at an even lower value nearing zero.

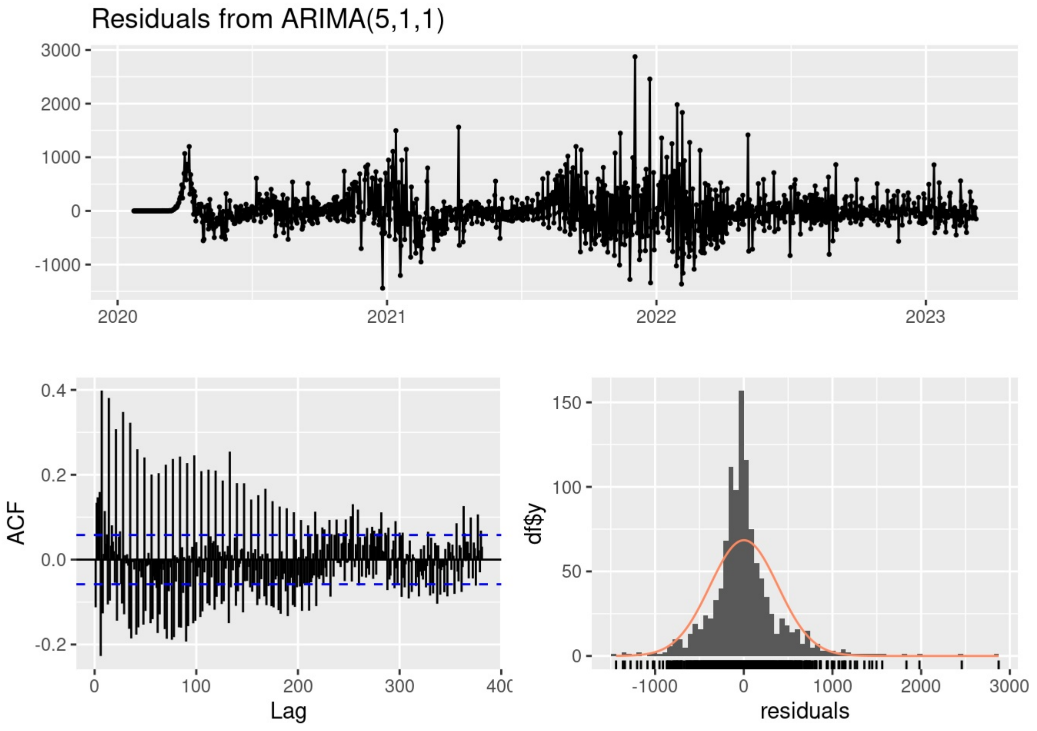
The shaded regions represent the prediction intervals, with darker blue indicating the 80% confidence interval and the lighter blue representing the 95% confidence interval.

Usually negative values should be impossible in this context, since they imply someone being brought back from death Due to this the bottom portion of the confidence intervals should be disregarded.

It is worth noting that four negative values were present in the original data but this was likely due to them being corrections improper reporting.

The value forecasted by the model at the end of the prediction is 318.89684, with the high 80% confidence interval being 1720.9181 and the high 90% confidence interval being 2463.1034.

```
checkresiduals(us_forecast)
```



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(5,1,1)
## Q* = 3540.2, df = 223, p-value < 2.2e-16
##
## Model df: 6.   Total lags used: 229
```

Residual diagnostics are used to evaluate how well the ARIMA model fits the data and whether the residuals behave like white noise.

The top plot shows the residuals over time. Ideally, residuals should fluctuate randomly around zero, without showing any patterns. It shows that most residuals are close to zero, which is a good sign that the model is not systematically under- or over-predicting values. There are periods of increased volatility, particularly in early to mid-2021, where the residuals show large deviations from zero. This suggests that the model had difficulty predicting certain events during this period.

The bottom left plot shows the autocorrelation of the residuals at various lags. Autocorrelation measures how much the residuals are correlated with themselves over time. Ideally, there should be no significant. In the early lags, the ACF shows significant positive autocorrelations that exceed the confidence bounds. Autocorrelations gradually decay but remain outside the confidence bounds for many lags. The presence of significant autocorrelations suggests that the model may not have fully captured the time-dependent structure of the data. It looks like the is model leaving some temporal structure unexplained, and more adjustments may be needed.

The bottom right plot shows the distribution of residuals, with a normal distribution curve overlaid. Ideally, residuals should follow a normal distribution, as ARIMA models assume normally distributed errors. These residuals do not perfectly follow the normal distribution. The histogram shows that there are more extreme residuals than would be expected under a normal distribution. There is also a clear peak near zero, which is expected, but the presence of outliers indicates some extreme events that the model struggled to predict. The residuals are not perfectly normally distributed, suggesting that the model's error distribution is skewed and has heavy tails. This is often a sign that the model has difficulty predicting rare, extreme events such as COVID-19 death surges due to new variants.

Biases

Regarding biases, there are a few areas to consider regarding the visualizations and the model.

In the visualizations, percentage growth charts can mislead by showing high growth in countries with initially lower death counts. Time windows and external factors, such as healthcare capacity or public policy changes, are not always visible in the charts, which can result in misinterpretation of growth patterns. Including context, such as per capita adjustments or annotations for major events, could help mitigate these biases.

For the ARIMA model, potential biases stem from the assumption that past trends will continue, which may not hold true if significant events like new variants or policy shifts occur. The model may not capture all time-based patterns, as indicated by residual autocorrelations, and struggles to predict extreme events like sudden surges in deaths.

Conclusion

This report analyzed the change in total deaths overtime of the 10 countries with most total reported COVID-19 deaths, and the percentage growth in COVID-19 deaths over time.

The visualizations showed that the United States experienced a significant and steady rise in deaths, likely due to post-holiday surges and delayed vaccine distribution. India saw a sharp increase in deaths during April-May 2021, driven by the Delta variant, contrasting with the more gradual growth in other countries. Brazil had consistently high deaths throughout the period, reflecting ongoing struggles with the pandemic.

It is also worth noting countries overall did mirror each other in both visualizations. Over time deaths stabilized across all countries, likely due to improved healthcare and vaccination efforts.

Regarding the ARIMA model, it provides a useful short-term forecast for daily COVID-19 deaths in the US, suggesting that deaths are likely to stabilize or decrease in the near future. However, the model exhibits some limitations in the long term, where increasing uncertainty and residual issues indicate the need for further refinement. Incorporating external factors such as vaccination rates and policy measures could improve the model's accuracy, especially during periods of rapid change or extreme events.

Resources:

- https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series (https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series)