COVID-19 Dataset Analysis

2024-08-19

COVID-19 Dataset Analysis

Importing required dependencies

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

Read in the cases and the deaths data from the given URLs

```
# Read in urls
UsCasesDataUrl <-
  "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_ti
UsDeathsDataUrl <-
  "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_ti

# Read in csv files into tibbles
usCasesData <- read_csv(UsCasesDataUrl)
usDeathsData <- read_csv(UsDeathsDataUrl)
```

Data Cleanup

In this section we will be organizing the data by removing un-necessary columns and restructuring the dataset where cases organized by date.

```
# Organize cases data by date
usCasesData <- usCasesData %>%
  pivot_longer(cols = -c(Province_State, Country_Region, UID, iso2, iso3, code3, FIPS, Admin2, Lat, Lon
# Remove un-necessary columns for our analysis
usCasesData <- usCasesData %>%
  select(-c(Lat, Long_, UID, FIPS, iso2, iso3, code3, Combined_Key))
```

Cleaning up COVID-19 cases data

```
# Organize deaths data by date
usDeathsData <-
  usDeathsData %>%
  pivot_longer(cols = -c(Province_State, Country_Region, UID, iso2, iso3, code3, FIPS, Admin2, Lat, Lon
# Remove unneeded columns
usDeathsData <-
  usDeathsData %>%
  select(-c(Lat, Long_, UID, FIPS, iso2, iso3, code3, Combined_Key))
```

Cleaning up COVID-19 deaths data

Seasonal Analysis

As the first step in our analysis, we will attempt to group COVID-19 cases and deaths by season. This summary allows us to compare the seasonal impact on COVID-19 cases and fatalities more effectively.

```
# Ensure that the date column is in Date format
usCasesData <- usCasesData %>%
  mutate(date = as.Date(date, format = "%m/%d/%y")) # Adjust the format if necessary

# Add a season column based on the correct date format
usCasesDataSeason <- usCasesData %>%
  mutate(season = case_when(
    month(date) %in% c(12, 1, 2) ~ "Winter",
    month(date) %in% c(3, 4, 5) ~ "Spring",
    month(date) %in% c(6, 7, 8) ~ "Summer",
    month(date) %in% c(9, 10, 11) ~ "Fall"
  ))

usCasesDataSeason <- usCasesDataSeason %>%
  drop_na(date)
```

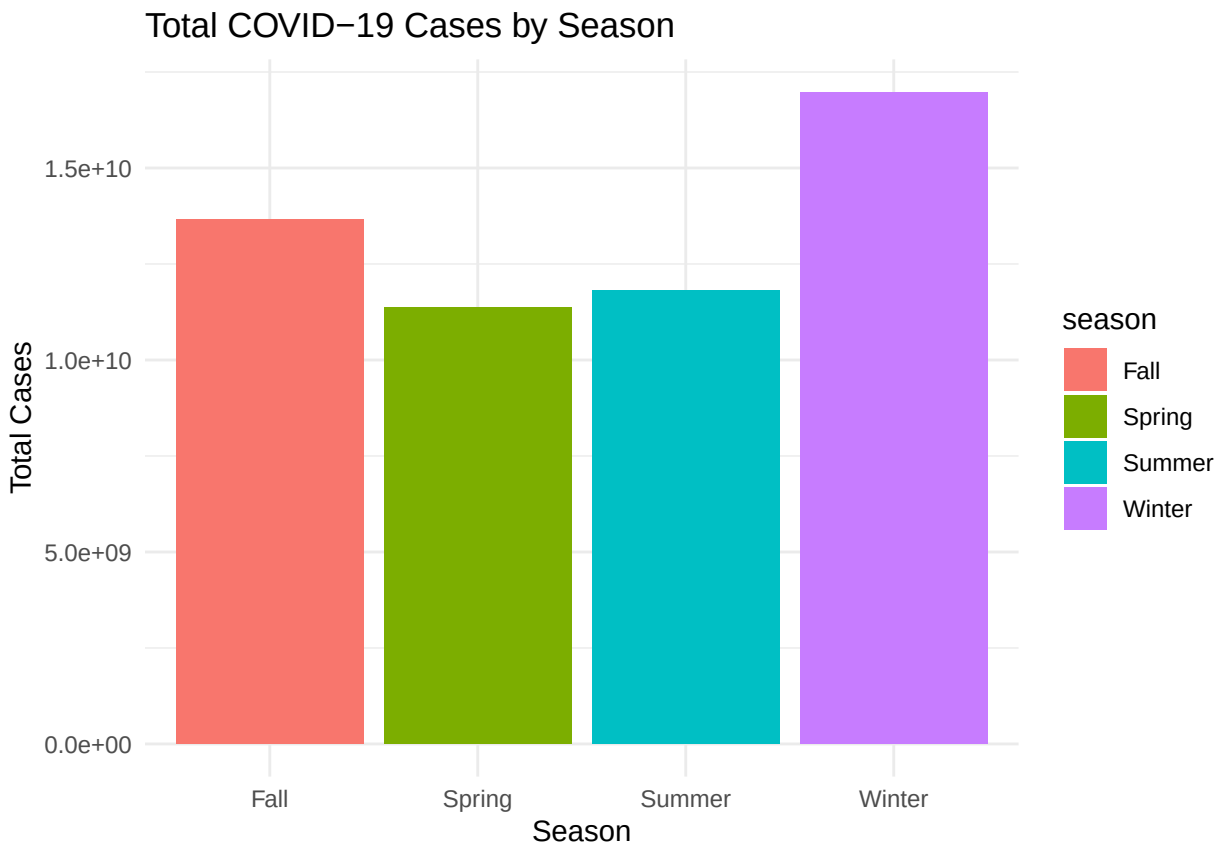
```
# Ensure that the date column is in Date format
usDeathsData <- usDeathsData %>%
  mutate(date = as.Date(date, format = "%m/%d/%y")) # Adjust the format if necessary

# Add a season column based on the correct date format
usDeathsDataSeason <- usDeathsData %>%
  mutate(season = case_when(
    month(date) %in% c(12, 1, 2) ~ "Winter",
    month(date) %in% c(3, 4, 5) ~ "Spring",
    month(date) %in% c(6, 7, 8) ~ "Summer",
    month(date) %in% c(9, 10, 11) ~ "Fall"
  ))

usDeathsDataSeason <- usDeathsDataSeason %>%
  drop_na(date)
```

After categorizing each record by season, we then summarize the data by calculating the total number of cases and deaths for each season.

```
# Summarize cases and deaths by season
seasonal_cases <- usCasesDataSeason %>%
  group_by(season) %>%
  summarize(total_cases = sum(cases, na.rm = TRUE))

seasonal_deaths <- usDeathsDataSeason %>%
  group_by(season) %>%
  summarize(total_deaths = sum(cases, na.rm = TRUE))
```

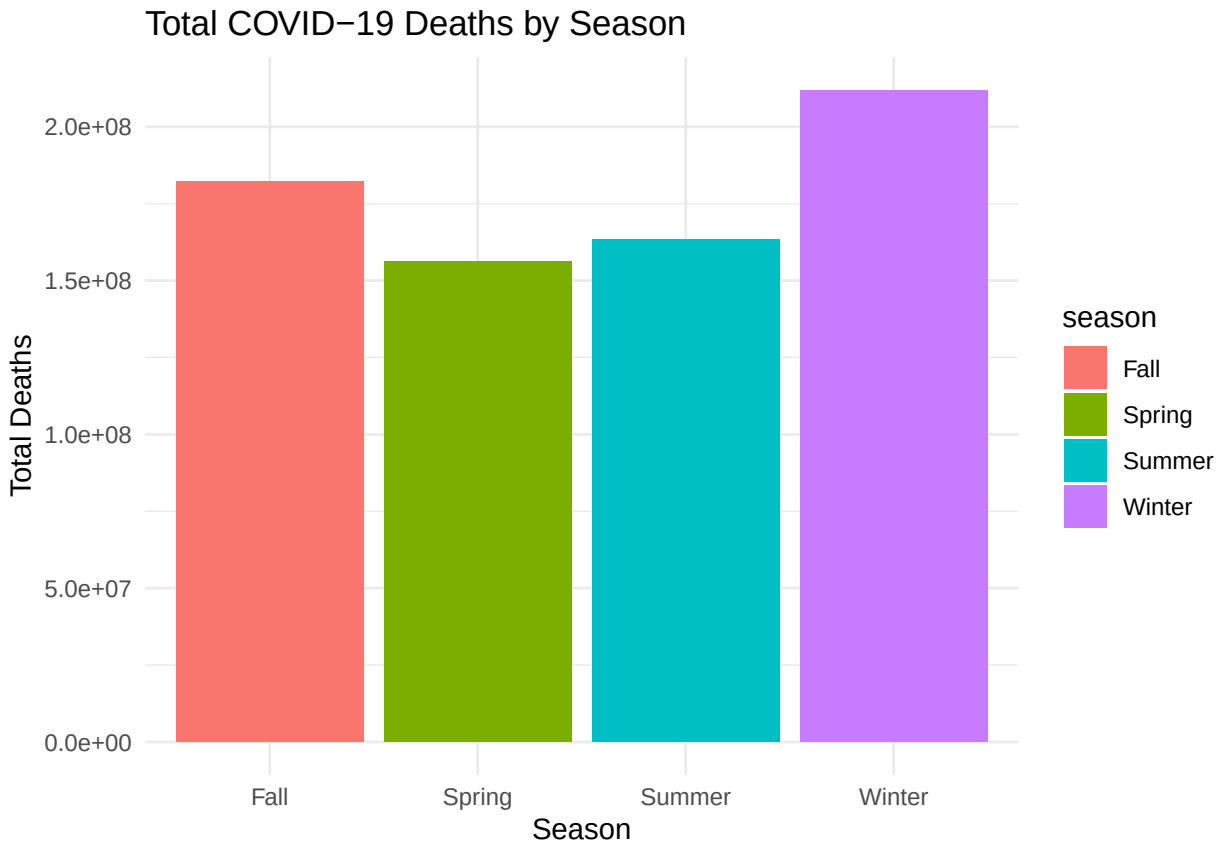
```
ggplot(seasonal_cases, aes(x = season, y = total_cases, fill = season)) +
  geom_bar(stat = "identity") +
  labs(title = "Total COVID-19 Cases by Season", x = "Season", y = "Total Cases") +
  theme_minimal()
```

Visualize COVID-19 Cases by Season




```
ggplot(seasonal_deaths, aes(x = season, y = total_deaths, fill = season)) +
  geom_bar(stat = "identity") +
  labs(title = "Total COVID-19 Deaths by Season", x = "Season", y = "Total Deaths") +
  theme_minimal()
```

Visualize COVID-19 Deaths by Season



These visualizations indicate that the winter and fall seasons tend to have higher counts of COVID-19 cases, which in turn lead to an increase in deaths. This seasonal pattern suggests that colder months are associated with higher transmission rates and, consequently, more severe outcomes.

To understand the relationship between the frequency of cases and deaths on specific days of the week, we can visualize the average number of cases and deaths for each day of the week.

```
# Add a column for day of the week
usCasesDataSeason <- usCasesDataSeason %>%
  mutate(day_of_week = wday(date, label = TRUE))

usDeathsDataSeason <- usDeathsDataSeason %>%
  mutate(day_of_week = wday(date, label = TRUE))
```

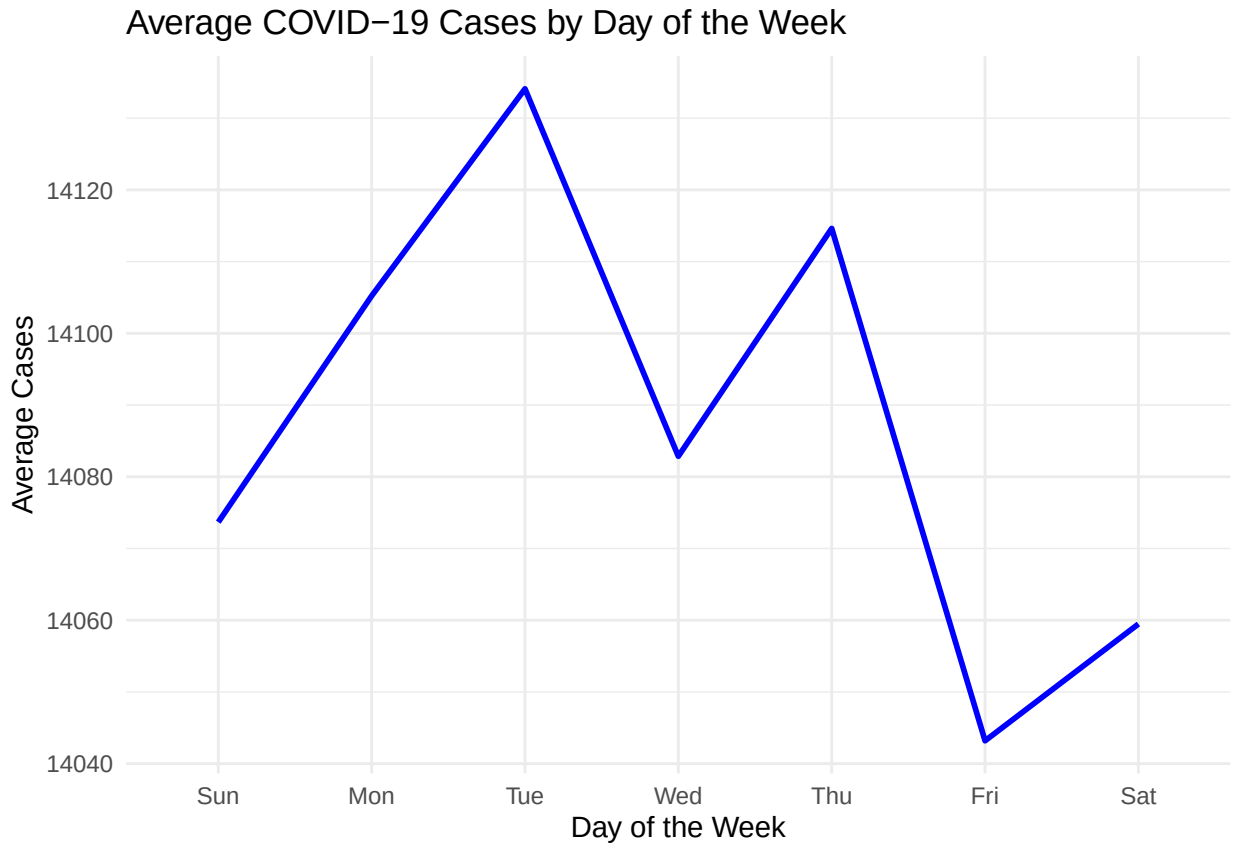
```
# Calculate the average cases by day of the week
average_weekly_cases <- usCasesDataSeason %>%
  group_by(day_of_week) %>%
  summarize(avg_cases = mean(cases, na.rm = TRUE))

# Calculate the average deaths by day of the week
average_weekly_deaths <- usDeathsDataSeason %>%
  group_by(day_of_week) %>%
  summarize(avg_deaths = mean(cases, na.rm = TRUE))
```

```
# Plot the smoothed average cases by day of the week
ggplot(average_weekly_cases, aes(x = day_of_week, y = avg_cases, group = 1)) +
  geom_line(color = "blue", size = 1) +
```

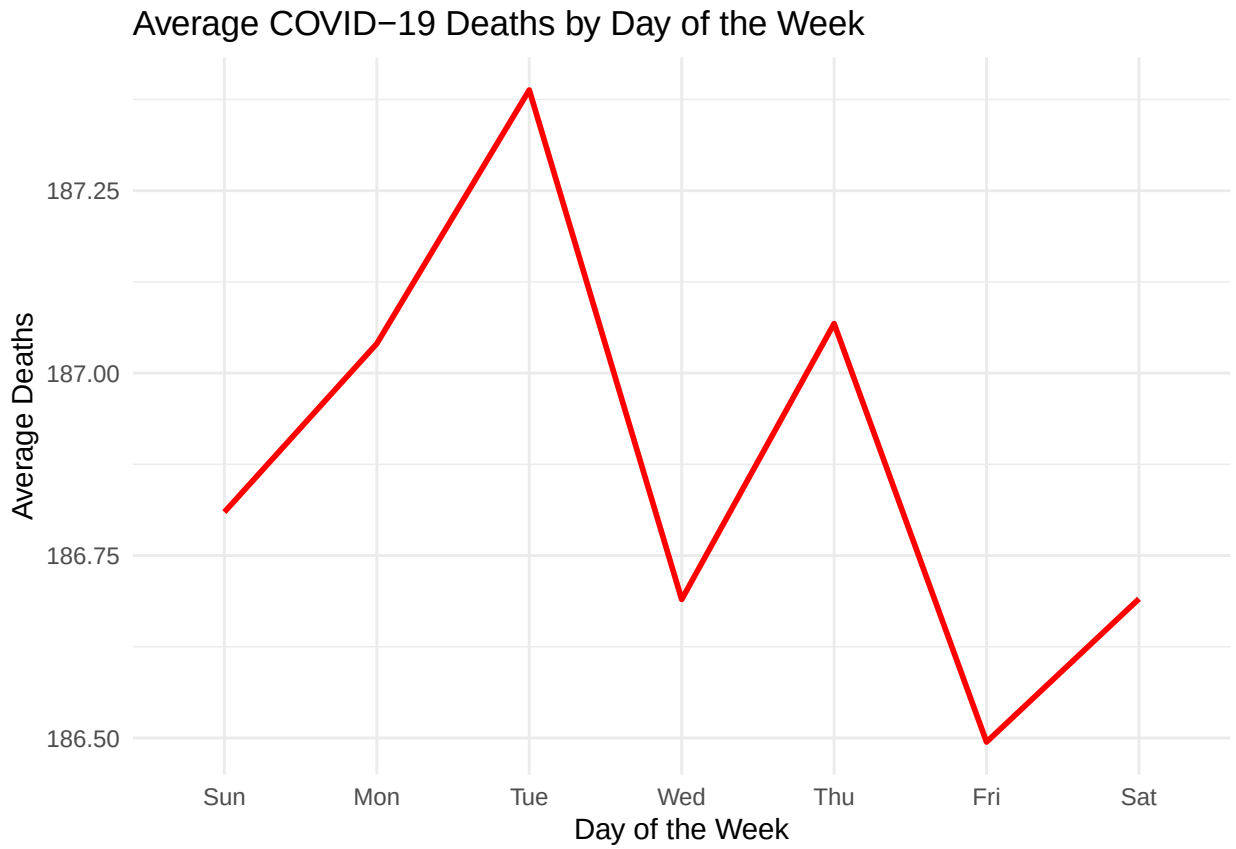
```
  labs(title = "Average COVID-19 Cases by Day of the Week", x = "Day of the Week", y = "Average Cases")
  theme_minimal()
```

Visualize COVID-19 Cases and Deaths by day of the week

```
## 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.
```




```
ggplot(average_weekly_deaths, aes(x = day_of_week, y = avg_deaths, group = 1)) +
  geom_line(color = "red", size = 1) +
  labs(title = "Average COVID-19 Deaths by Day of the Week", x = "Day of the Week", y = "Average Deaths
  theme_minimal()
```



These visualizations reveal that the days of the week with the highest number of COVID-19 cases also correspond to the days with the highest recorded deaths.

Forecasting COVID-19 Deaths and Cases

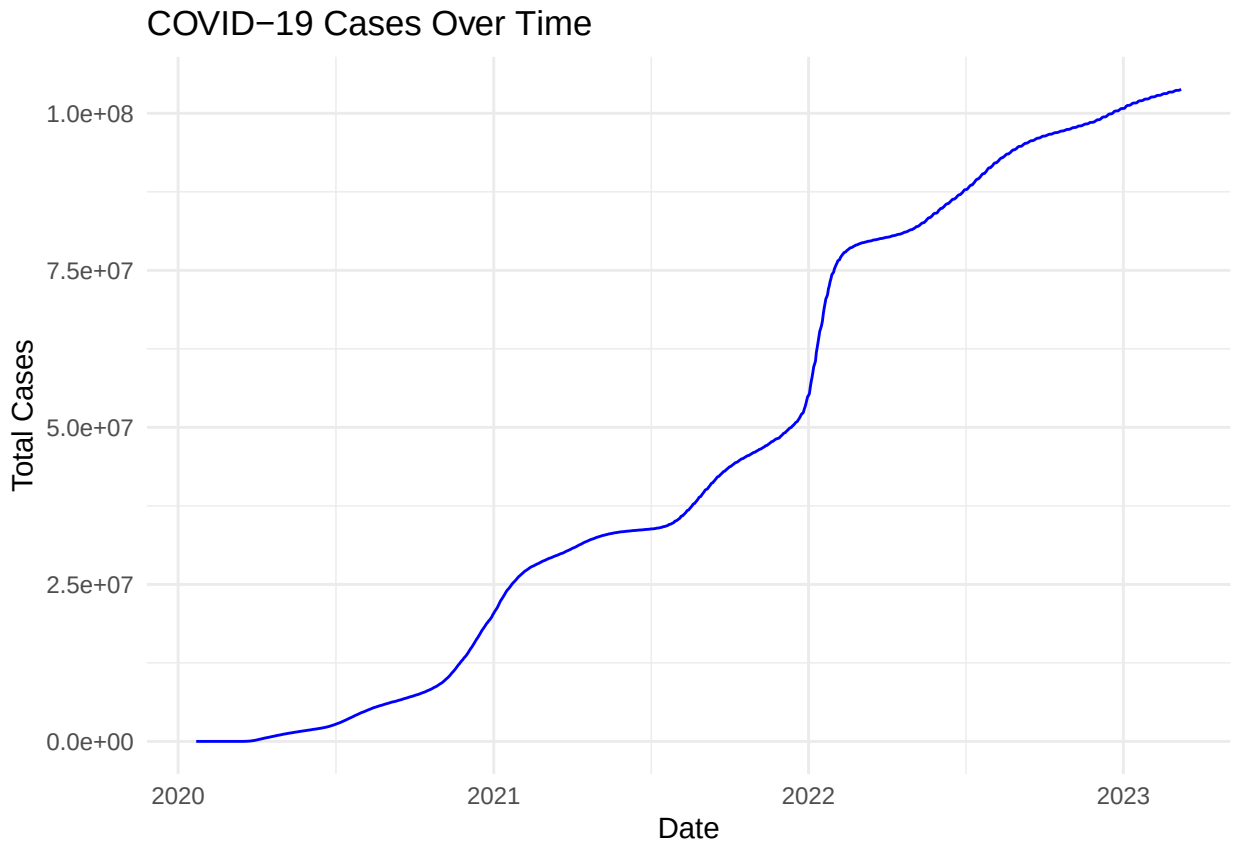
Observations

The historical data reveals a clear upward trend in COVID-19 cases and deaths. Notably, spikes in deaths typically follow spikes in cases, aligning with the expectation that a surge in cases leads to an increase in COVID-related fatalities. However, around early 2022, there is a marked increase in cases without a corresponding rise in deaths. This deviation from earlier trends can likely be attributed to the widespread rollout of COVID-19 vaccines beginning in mid to late 2021. By 2022, a significant portion of the population had likely received at least one vaccine dose, which may have contributed to the reduction in death rates despite the rise in cases.

```
# Aggregate data by date
daily_cases <- usCasesDataSeason %>%
  group_by(date) %>%
  summarize(total_cases = sum(cases, na.rm = TRUE))

daily_deaths <- usDeathsDataSeason %>%
  group_by(date) %>%
  summarize(total_deaths = sum(cases, na.rm = TRUE))

# Plot cases and deaths over time
ggplot(daily_cases, aes(x = date, y = total_cases)) +
```

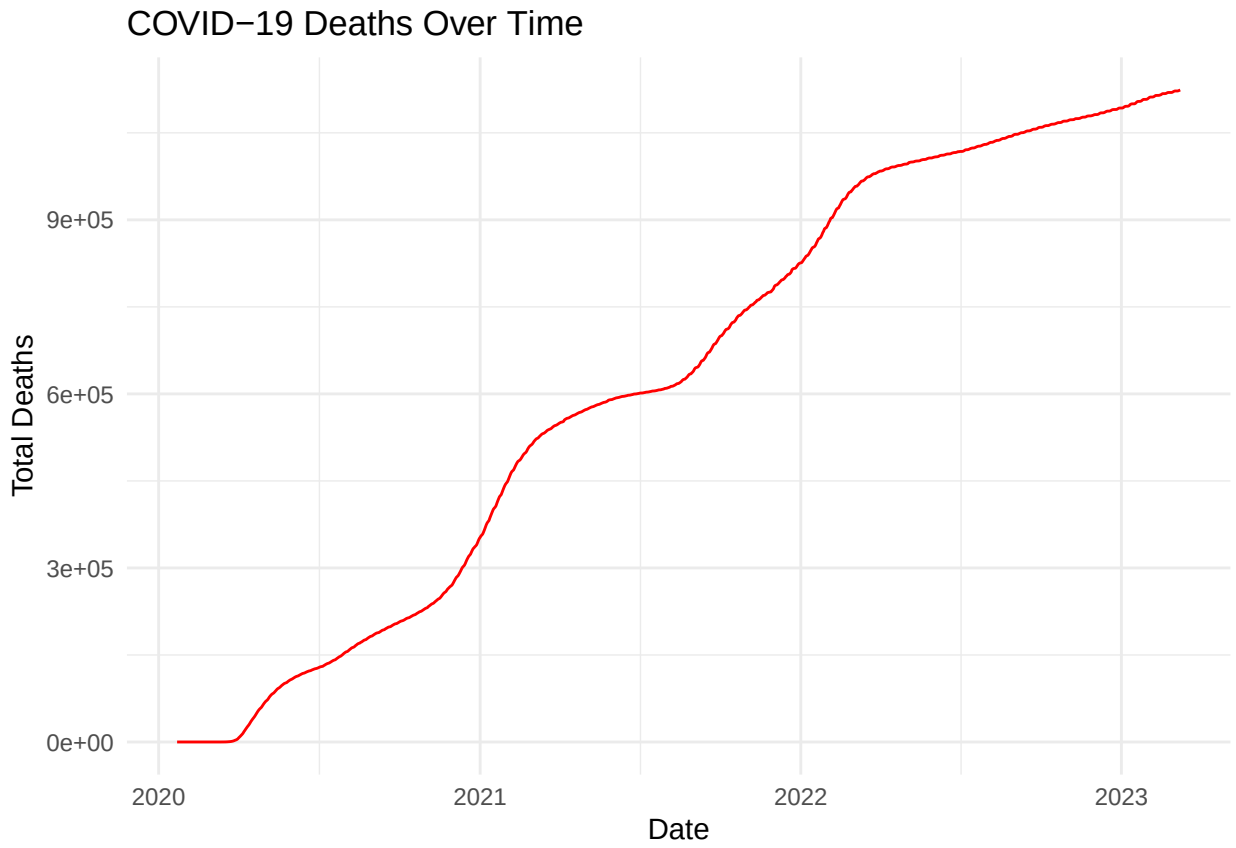
```
  geom_line(color = "blue") +
 labs(title = "COVID-19 Cases Over Time", x = "Date", y = "Total Cases") +
  theme_minimal()
```

Visualize COVID-19 Cases over time




```
# Plot cases and deaths over time
ggplot(daily_deaths, aes(x = date, y = total_deaths)) +
  geom_line(color = "red") +
 labs(title = "COVID-19 Deaths Over Time", x = "Date", y = "Total Deaths") +
  theme_minimal()
```

Visualize COVID-19 Deaths over time



Forecasting

This section attempts to forecast COVID-19 cases and deaths 120 days into the future based on the historical data. For this I will be using the ARIMA (Auto-regressive Integrated Moving Average) model.

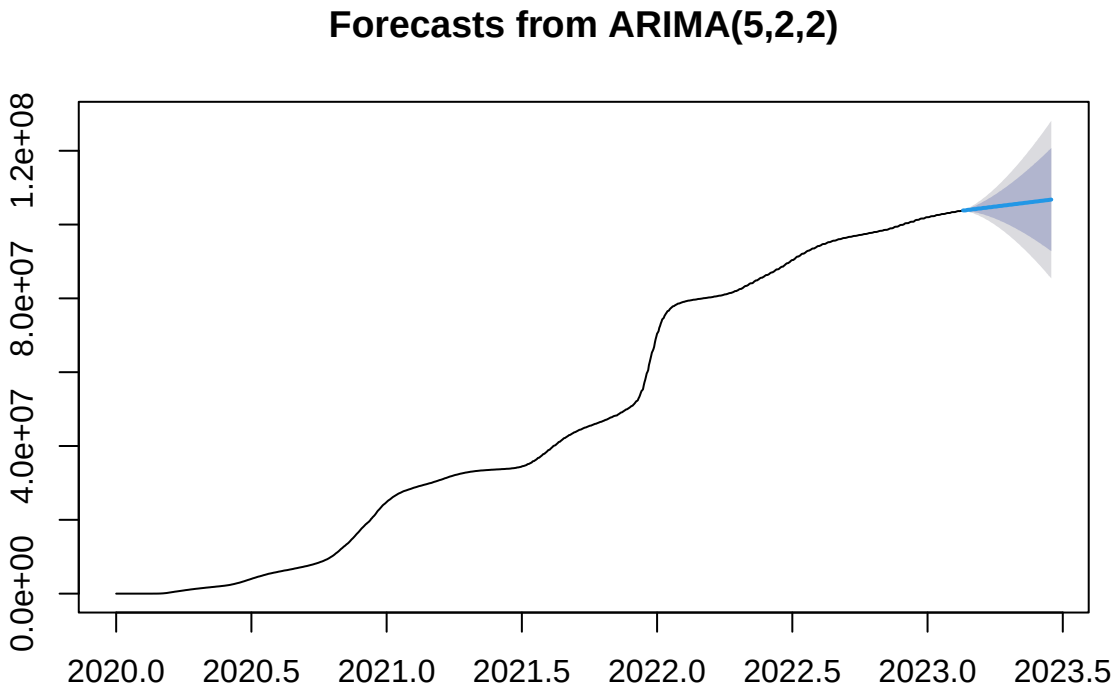
The ARIMA forecasting model assumes a degree of regularity in the trends of cases and deaths. However, COVID-19 waves can be driven by unexpected external factors such as new variants, changes in public behavior, or policy changes. This could result in overly optimistic or pessimistic forecasts.

Explanation of the visualizations

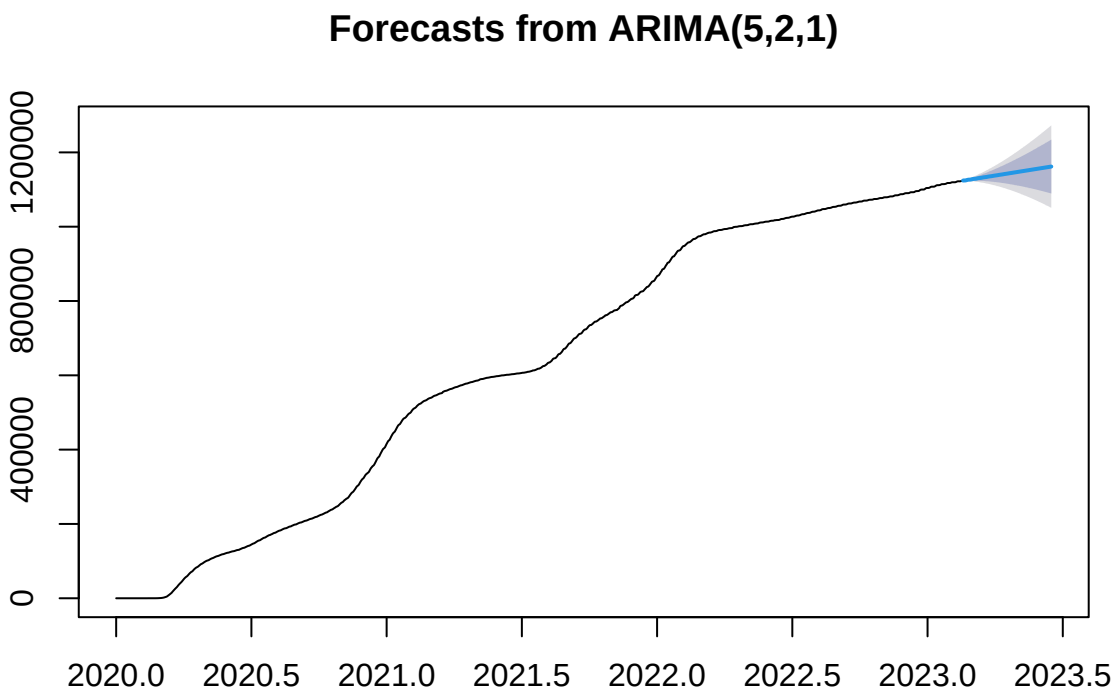
- The black line in the visualizations represent the values of COVID-19 cases (or deaths) up to the point where the forecast begins.
- The blue line represents the predicted future values for the COVID-19 cases or deaths.
- The shaded area around the blue line represents the confidence interval (usually 80 - 95 %) of the forecast

```
# Prepare time-series data
ts_cases <- ts(daily_cases$total_cases, frequency = 365, start = c(2020, 1))

# Fit an ARIMA model
arima_model <- auto.arima(ts_cases)

# Forecast the next 120 days
forecasted_values <- forecast(arima_model, h = 120)

# Plot the forecast
```

```r
plot(forecasted_values)
```




```r
# Prepare time-series data
ts_deaths <- ts(daily_deaths$total_deaths, frequency = 365, start = c(2020, 1))

# Fit an ARIMA model
arima_model <- auto.arima(ts_deaths)

# Forecast the next 120 days
forecasted_values <- forecast(arima_model, h = 120)

# Plot the forecast
plot(forecasted_values)
```



Observations When examining the two forecast visualizations, we observe that the forecasts for both cases and deaths follow similar trends with no significant deviations. Although it is typical for the confidence interval to widen over time in forecasting models, the confidence interval for cases is noticeably broader than that for deaths. This is likely due to the historical data for COVID-19 cases showing much more variability compared to the relatively stable trends seen in COVID-19 deaths.

Bias Identification

In this analysis, several potential biases could affect the results and the interpretations. The dataset spans multiple years, and the behavior of the virus and public response evolved over this period. Variations in lockdown measures, mask mandates, and vaccination rates across time could skew the results, especially if they disproportionately impacted different seasons or days of the week.

The dataset also includes the entire United States, but the severity of COVID-19, public health responses, and weather patterns vary significantly across states and even cities within the same state. A national-level analysis may mask important regional trends that could influence seasonality and day-of-the-week effects.

As we somewhat observed during the analysis, the vaccination campaign significantly altered the severity of COVID-19 outcomes. The introduction and widespread administration of vaccines, especially post 2021, may introduce a bias in trend analysis since post-vaccination data may show reduced death rates even with higher case counts.

Reporting practices for cases and deaths may vary by state, day of the week, and across time. For instance, weekends and holidays might show lower reported counts due to delays in data aggregation. This could artificially create patterns that aren't driven by actual case or death surges.

Conclusion

This analysis provides an overview of how COVID-19 cases and deaths have varied by season and day of the week across the United States. The findings show us there were higher transmission and mortality rates during the colder months, particularly in winter and fall. Additionally, the day-of-week analysis highlights that specific days, particularly earlier in the week, tend to see higher average cases and deaths—likely due to delays in reporting over weekends. The forecasting model predict that while cases may continue to fluctuate, deaths are expected to remain more stable, highlighting the impact of vaccination and improved treatment protocols.

Overall, the study re-emphasizes the importance of timely interventions during high-risk seasons and provides insights into how public health measures and policies might be adapted to mitigate future COVID-19 waves.