Title: COVID-19 Analysis 2020-2023

Name: Riddhi Mahesh Dange

CWID: 20012299

Introduction

Why I chose this topic?

The COVID-19 pandemic began in 2020, and despite widespread vaccination by 2023, its effects endure. As society returns to a pre-pandemic, in-person mode, I am driven to reassess its consequences across different geographical areas. In this comprehensive analysis project, I've delved into the critical issue of the COVID-19 pandemic. The choice of this topic is motivated by the ongoing challenges and concerns that have resurfaced after what appeared to be initial progress in controlling the virus. The pandemic, which first emerged in late 2019, has had a profound global impact, leading to millions of cases and deaths, societal disruptions, and economic consequences. In some regions, it seemed that the worst was behind us, but even now after 3 years, recent developments have prompted a reevaluation of the situation.

Importance of this topic:

The world has witnessed waves of infections, lockdowns, vaccination campaigns, and varying government responses. These complexities have raised questions about the effectiveness of public health measures, vaccine distribution, and the virus's evolving nature. To gain a comprehensive understanding and make informed decisions, it is crucial to analyze and visualize the available data.

Process:

This project employs the ggplot2 library in R to create insightful visualizations based on a comprehensive COVID-19 dataset. The dataset includes critical columns such as Date_reported, Country, WHO_region, New_cases, Cumulative_cases, New_deaths, Cumulative_deaths, and deaths. By exploring this data, we aim to assess the current state of the pandemic, compare it to past periods, and identify patterns and trends that can inform future strategies.

Outcome:

Through these visualizations, we seek to shed light on questions such as the resurgence of cases, regional disparities, the impact of vaccination campaigns, and the effectiveness of public health interventions. Our goal is not only to provide valuable insights into the current situation but also to contribute to the broader understanding of COVID-19, ultimately aiding decision-makers, healthcare professionals, and the public in navigating this ongoing crisis.

Data Explanation

This data set contains the following parameters of COVID-19 cases in the range of year 2019-2023:

Date_reported: This column records the date on which COVID-19 cases or deaths were officially reported. The timeline of reporting is essential for tracking the evolution of the pandemic over time, identifying peaks, and understanding trends in case and death reporting.

Country: This attribute indicates the specific country or region where the COVID-19 data was reported. Analyzing data by country helps assess the geographical spread of the virus and the varying impacts it has had on different regions.

WHO_region: The World Health Organization (WHO) region to which the country belongs. This categorization enables the analysis of regional differences in pandemic management, healthcare infrastructure, and responses to COVID-19.

New_cases: This column represents the number of new COVID-19 cases reported on a given date. Tracking daily new cases helps monitor infection rates, identify surges, and assess the effectiveness of containment measures.

Cumulative_cases: Cumulative_cases indicates the total number of COVID-19 cases reported up to a specific date. This metric provides insight into the overall scale of the pandemic's impact within a region or country.

New_deaths: New_deaths records the daily number of new COVID-19-related deaths. Monitoring daily deaths is crucial for understanding the virus's lethality and healthcare system strain.

Cumulative_deaths: Cumulative_deaths is the total number of COVID-19-related deaths reported up to a specific date. This metric helps assess the overall mortality associated with the virus in different regions.

Deaths: This column potentially includes detailed information about COVID-19-related deaths, such as age group, gender, and comorbidities. Such details are critical for understanding the demographics of those most affected by the virus and tailoring healthcare responses accordingly.

By comprehensively analyzing these data attributes, we can gain valuable insights into the pandemic's current status, compare it to past phases, and inform evidence-based decisions and policies for managing and mitigating the ongoing COVID-19 crisis.

***Graphs Plotted**

Box plot, Heat map, Bar graph, Line graph, Scatterplot graph

```
library(ggplot2)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(reshape2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(igraph)
```

```
##
## Attaching package: 'igraph'
```

```
## The following objects are masked from 'package:dplyr':
##
##     as_data_frame, groups, union
```

```
## The following objects are masked from 'package:lubridate':
##
##     %--%, union
```

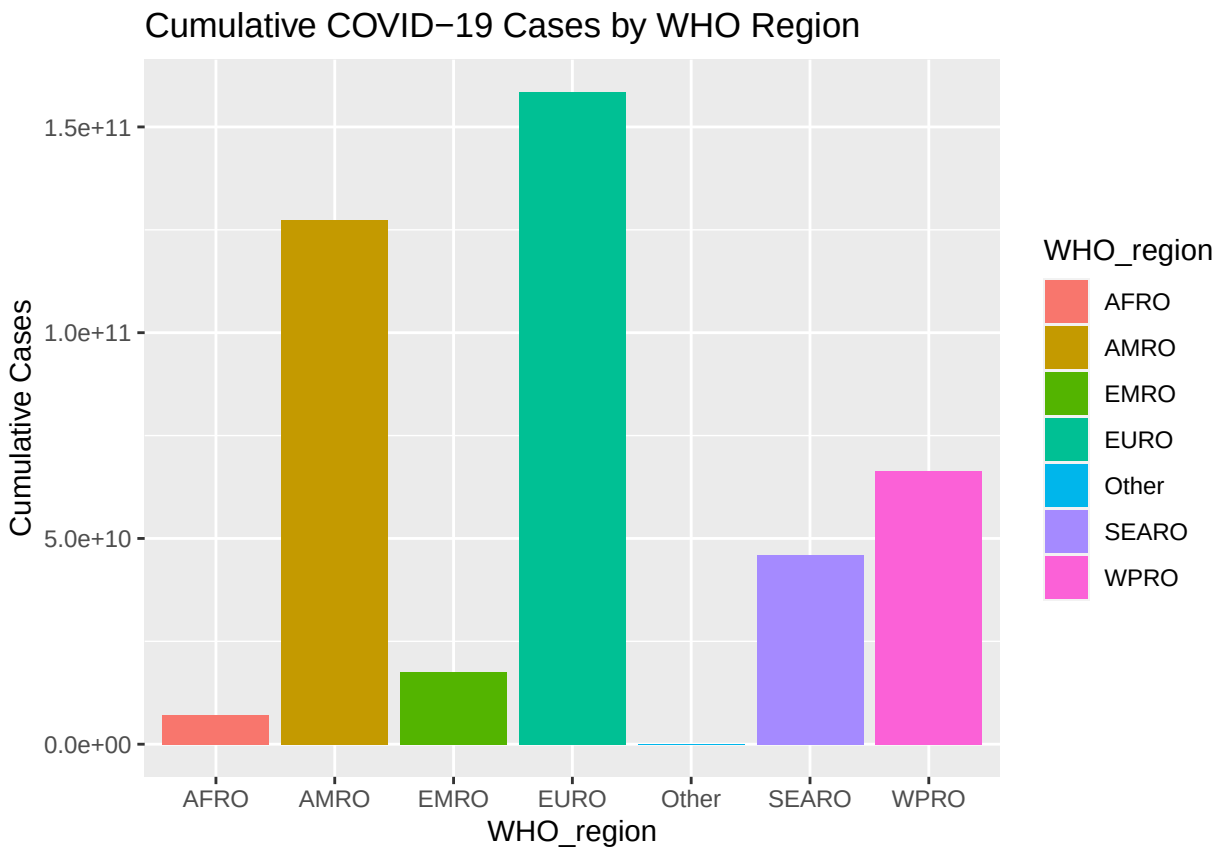
```
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum
```

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

```
#Reading the Data
data <- read.csv("COVID-19-global-data.csv")
data$Date_reported <- ymd(data$Date_reported)

summary(data)
```

```
##  Date_reported        Country_code         Country            WHO_region
##  Min.   :2020-01-03   Length:306678      Length:306678      Length:306678
##  1st Qu.:2020-11-21   Class :character   Class :character   Class :character
##  Median :2021-10-10   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2021-10-10
##  3rd Qu.:2022-08-30
##  Max.   :2023-07-19
##    New_cases         Cumulative_cases      New_deaths         Cumulative_deaths
##  Min.   :-1105466   Min.   :        0   Min.   :-120896.00   Min.   :      0
##  1st Qu.:       0   1st Qu.:     1939   1st Qu.:      0.00   1st Qu.:     15
##  Median :       1   Median :    33144   Median :      0.00   Median :    365
##  Mean   :    2505   Mean   :  1376457   Mean   :     22.67   Mean   :  17440
##  3rd Qu.:     191   3rd Qu.:   383204   3rd Qu.:      2.00   3rd Qu.:   5635
##  Max.   : 6966046   Max.   :103436829   Max.   : 120896.00   Max.   :1127152
```

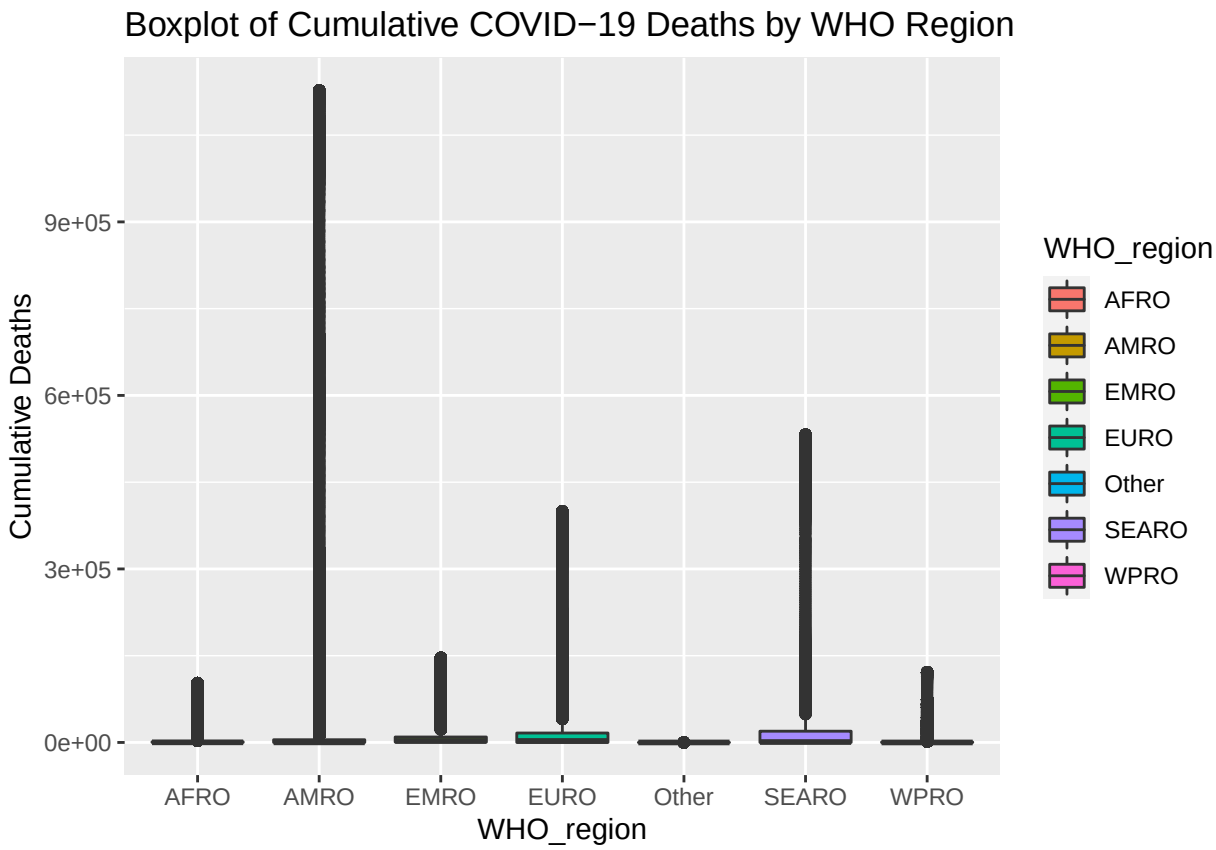
```
ggplot(data, aes(x = WHO_region, y = Cumulative_cases, fill = WHO_region)) +
  geom_bar(stat = "identity") +
  labs(title = "Cumulative COVID-19 Cases by WHO Region", y = "Cumulative Cases")
```



Interpretation of the Bar graph above

The following graph in a COVID-19 dataset provides insights into regional dynamic which aids resource allocation and policy planning. From the graph above, we understand that the WHO regions with the highest cumulative number of cases as of September 2023 are the Americas and Europe, followed by Southeast Asia, the Western Pacific, and Africa. The WHO region with the lowest cumulative number of cases is the Eastern Mediterranean.

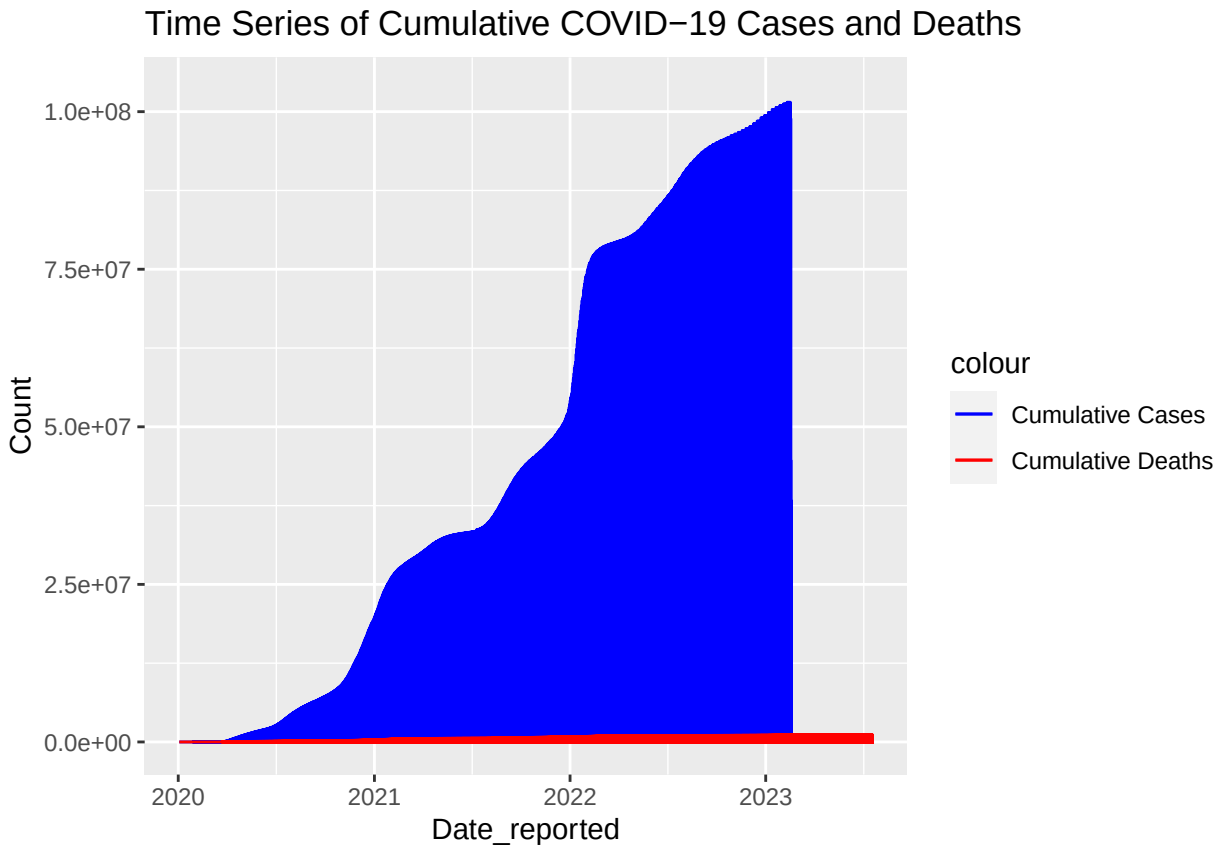
```
ggplot(data, aes(x = WHO_region, y = Cumulative_deaths, fill = WHO_region)) +
  geom_boxplot() +
  labs(title = "Boxplot of Cumulative COVID-19 Deaths by WHO Region", y = "Cumulative Deaths")
```



Interpretation of the Boxplot graph above

Boxplot with WHO Region on the X-axis and Cumulative Cases on the Y-axis for a COVID-19 dataset is valuable since it enables straightforward comparisons of case distributions across regions, identifies outliers indicating extreme values, assesses data variability, provides summary statistics, and offers insights into regional dynamics transparently. As we can see in the graph above, we identify that AMRO region contains the highest amount of outliers followed by SEARO and EURO; while AFRO observes the lowest.

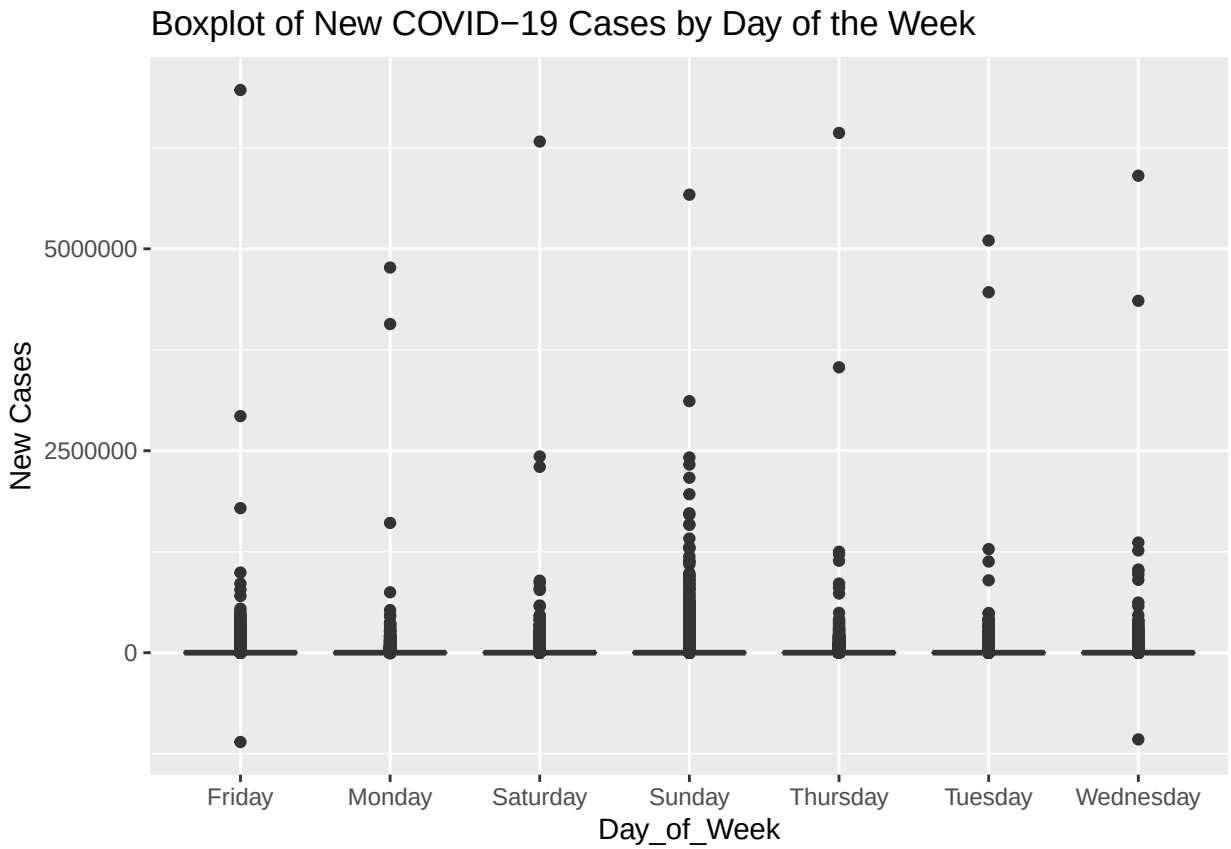
```
ggplot(data, aes(x = Date_reported)) +
  geom_line(aes(y = Cumulative_cases, color = "Cumulative Cases")) +
  geom_line(aes(y = Cumulative_deaths, color = "Cumulative Deaths")) +
  labs(title = "Time Series of Cumulative COVID-19 Cases and Deaths", y = "Count") +
  scale_color_manual(values = c("Cumulative Cases" = "blue", "Cumulative Deaths" = "red"))
```



Interpretation of the Time Series line graph above

A time series plot of cumulative COVID-19 cases and deaths is vital as it visually tracks the pandemic's evolution, identifies milestones, and assesses intervention effectiveness. From the graph above, we interpret that as we approach 2023, while the cases are increasing again; the deaths remain stagnant; which shows the impact of COVID-19 vaccination. The graph shows that the number of daily new cases and deaths has fluctuated over time, but has generally decreased since the peak of the pandemic in early 2022. However, there has been a recent uptick in new cases and deaths, which may be due to the emergence of new variants of the virus or changes in public health measures.

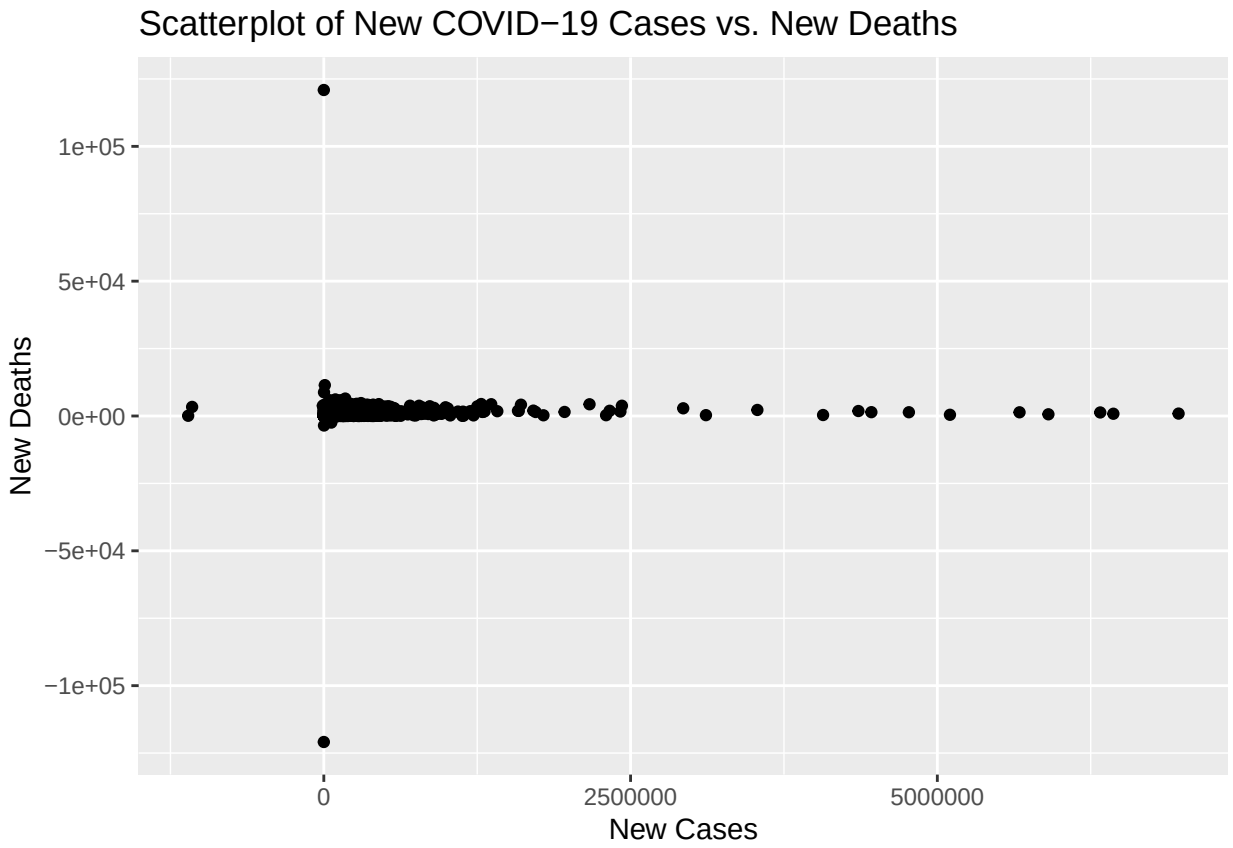
```
data$Day_of_Week <- weekdays(as.Date(data$Date_reported))

ggplot(data, aes(x = Day_of_Week, y = New_cases)) +
  geom_boxplot() +
  labs(title = "Boxplot of New COVID-19 Cases by Day of the Week", y = "New Cases")
```



Interpretation of the Boxplot graph above

Boxplot of New COVID-19 Cases by Day of the Week is important for understanding temporal patterns in the pandemic. It helps identify which days exhibit higher or lower case counts, aiding healthcare resource planning and informing public health interventions. We can observe that new COVID-19 cases are highest on Friday and Saturday, lowest on Sunday. There is a wide range in daily cases, with some days having over 100,000 and others having less than 10,000. The variability in daily cases is highest on Friday and Saturday.

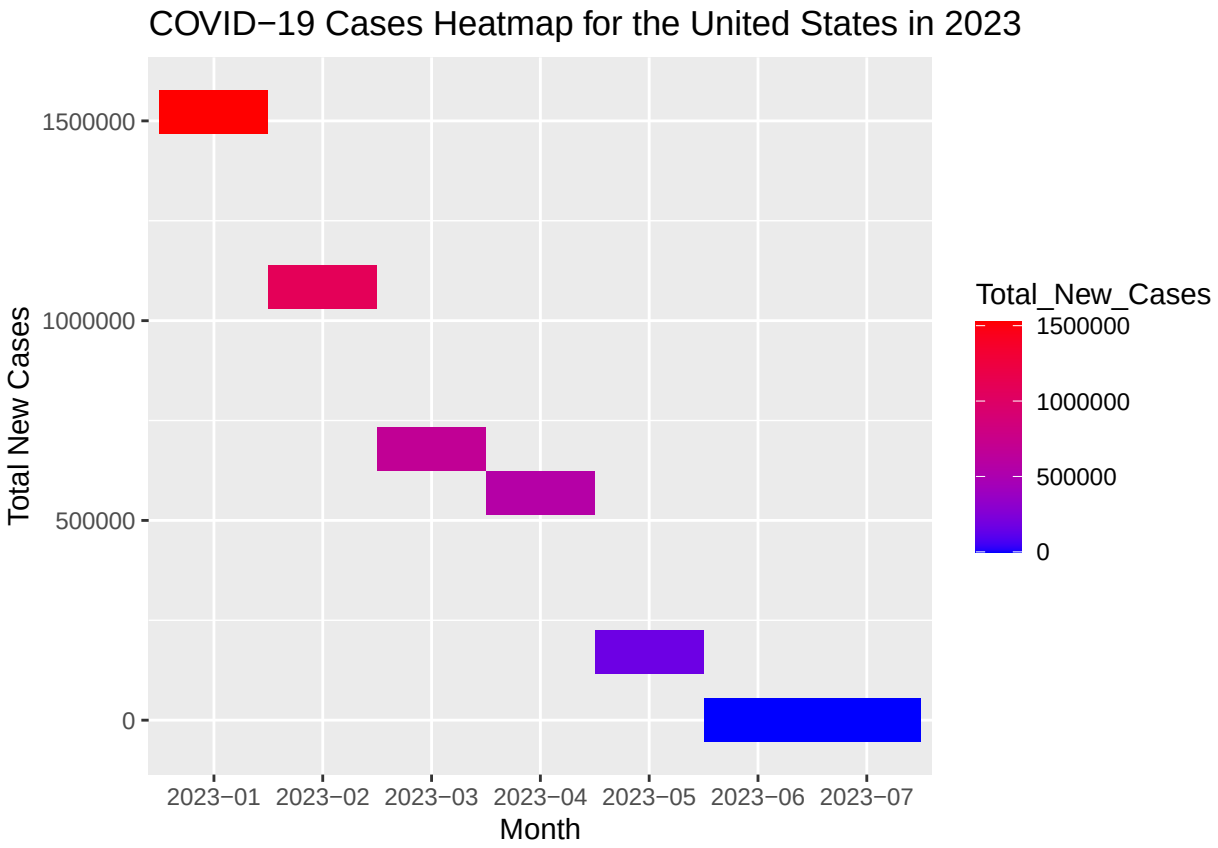
```
ggplot(data, aes(x = New_cases, y = New_deaths)) +
  geom_point() +
  labs(title = "Scatterplot of New COVID-19 Cases vs. New Deaths", x = "New Cases", y = "New Deaths")
```



Interpretation of the Scatterplot graph above

A scatterplot that displays New COVID-19 Cases against New Deaths is valuable as it visually assesses the relationship between case numbers and fatalities, identifies outliers, visualizes trends, and communicates insights transparently. We can see that as the number of new cases increases, the number of new deaths also tends to increase. However, as there are some data points that fall outside of the general trend; one possible explanation for this is that the number of deaths is not only influenced by the number of new cases, but also by other factors such as the severity of the disease, the age and health of the population, and the effectiveness of public health measures.

```
# Filter the data for the United States and the year 2023
data_us_2023 <- data %>%
  filter(Country == "United States of America", format(as.Date(Date_reported), "%Y") == "2023")

# Create a heatmap
heatmap_data <- data_us_2023 %>%
  mutate(Month = format(as.Date(Date_reported), "%Y-%m")) %>%
  group_by(Month) %>%
  summarise(Total_New_Cases = sum(New_cases))

ggplot(data = heatmap_data, aes(x = Month, y = Total_New_Cases, fill = Total_New_Cases)) +
  geom_tile() +
  scale_fill_gradient(low = "Blue", high = "red") +
  labs(title = "COVID-19 Cases Heatmap for the United States in 2023",
       x = "Month", y = "Total New Cases")
```



Interpretation of the Heatmap graph above

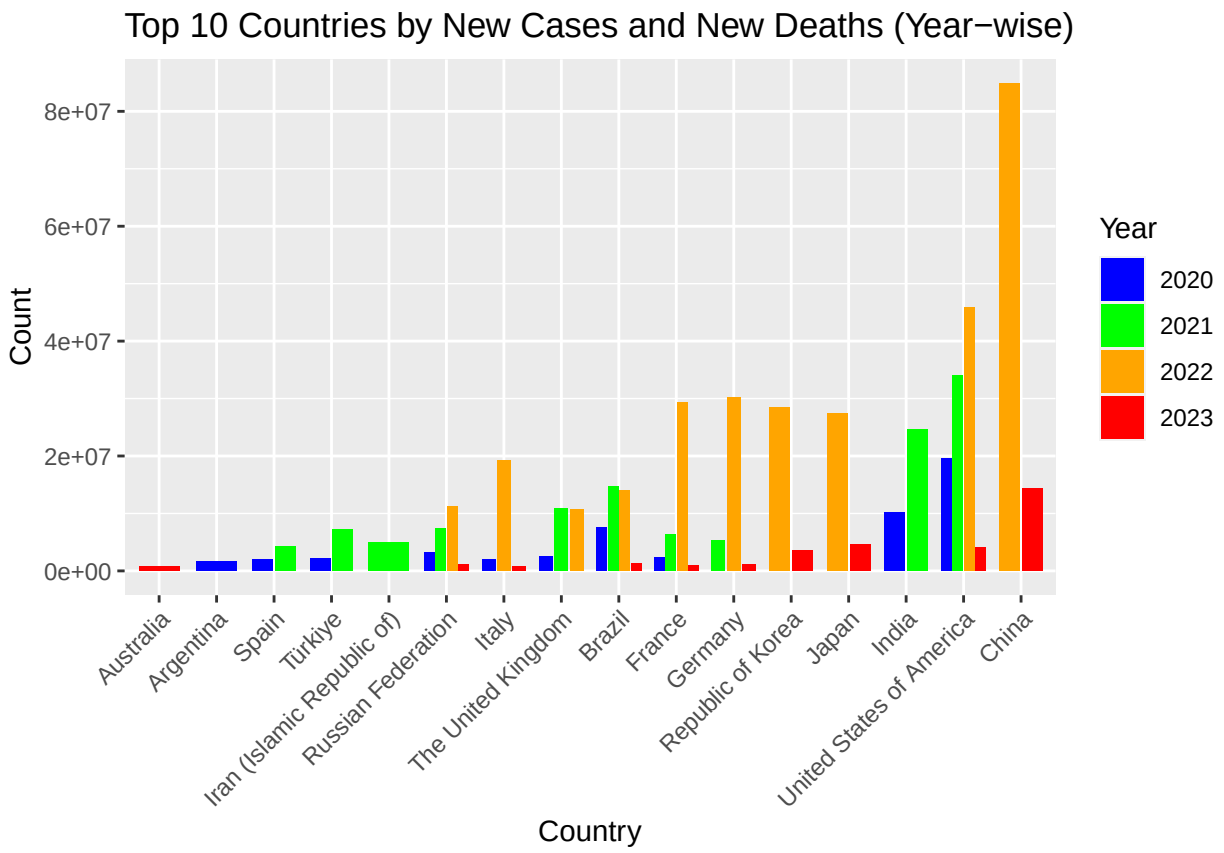
Creating a COVID-19 Cases Heatmap for the United States in 2023 with "Month" on the X-axis and "Total New Cases" on the Y-axis has several advantages. It visually tracks case trends, identifies seasonal patterns, and helps allocate resources efficiently. The visual appeal aids in public awareness.The heatmap shows that the number of new COVID-19 cases in each year in the United States in 2023. The darker the color, the higher the number of new cases. We can see how the number of new cases were highest in January of 2023, and are approaching to 0 by the month of July.

```
# Extract the year from the Date_reported column
data <- data %>%
  mutate(Year = format(as.Date(Date_reported), "%Y"))

# Calculate total new cases and new deaths for each country and year
top_countries <- data %>%
  group_by(Year, Country) %>%
  summarise(Total_New_Cases = sum(New_cases), Total_New_Deaths = sum(New_deaths)) %>%
  arrange(Year, desc(Total_New_Cases)) %>%
  group_by(Year) %>%
  slice(1:10)  # Select the top 10 countries for each year
```

```
## `summarise()` has grouped output by 'Year'. You can override using the
## `.groups` argument.
```

```
# Create a bar graph
ggplot(top_countries, aes(x = reorder(Country, Total_New_Cases),
                          y = Total_New_Cases,
                          fill = Year)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.8), width = 0.7) +
```

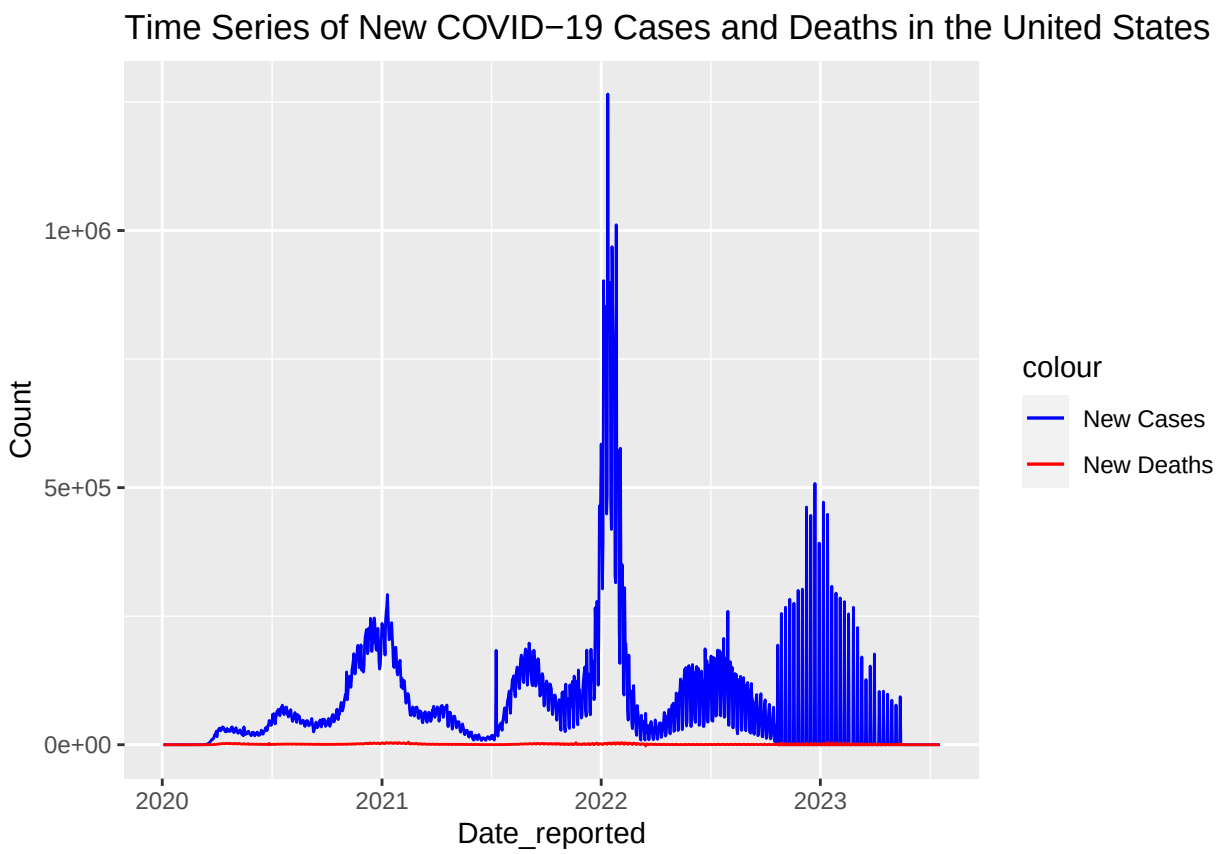
```
  geom_bar(aes(y = Total_New_Deaths), stat = "identity", position = position_dodge(width = 0.8), width =
  labs(title = "Top 10 Countries by New Cases and New Deaths (Year-wise)",
       x = "Country", y = "Count") +
  scale_fill_manual(values = c("2020" = "blue", "2021" = "green", "2022" = "orange", "2023" = "red")) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Interpretation of the Bar graph above

Creating a bar graph featuring the Top 10 Countries by New Cases and New Deaths year by year has significant benefits as it allows comparative analysis, highlights global hotspots, tracks trends, aids resource allocation, and communicates effectively. The graph shows that the United States has consistently had the highest number of new cases and new deaths throughout the pandemic. In 2023, the United States is still the country with the highest number of new cases and new deaths, followed by Canada, Mexico, and the United Kingdom. Other countries in the top 10 include Brazil, France, Germany, Italy, Russia, and India. However, the number of new cases and new deaths in these countries has fluctuated over time. For example, India had a high number of new cases and new deaths in 2021, but the number has since decreased.

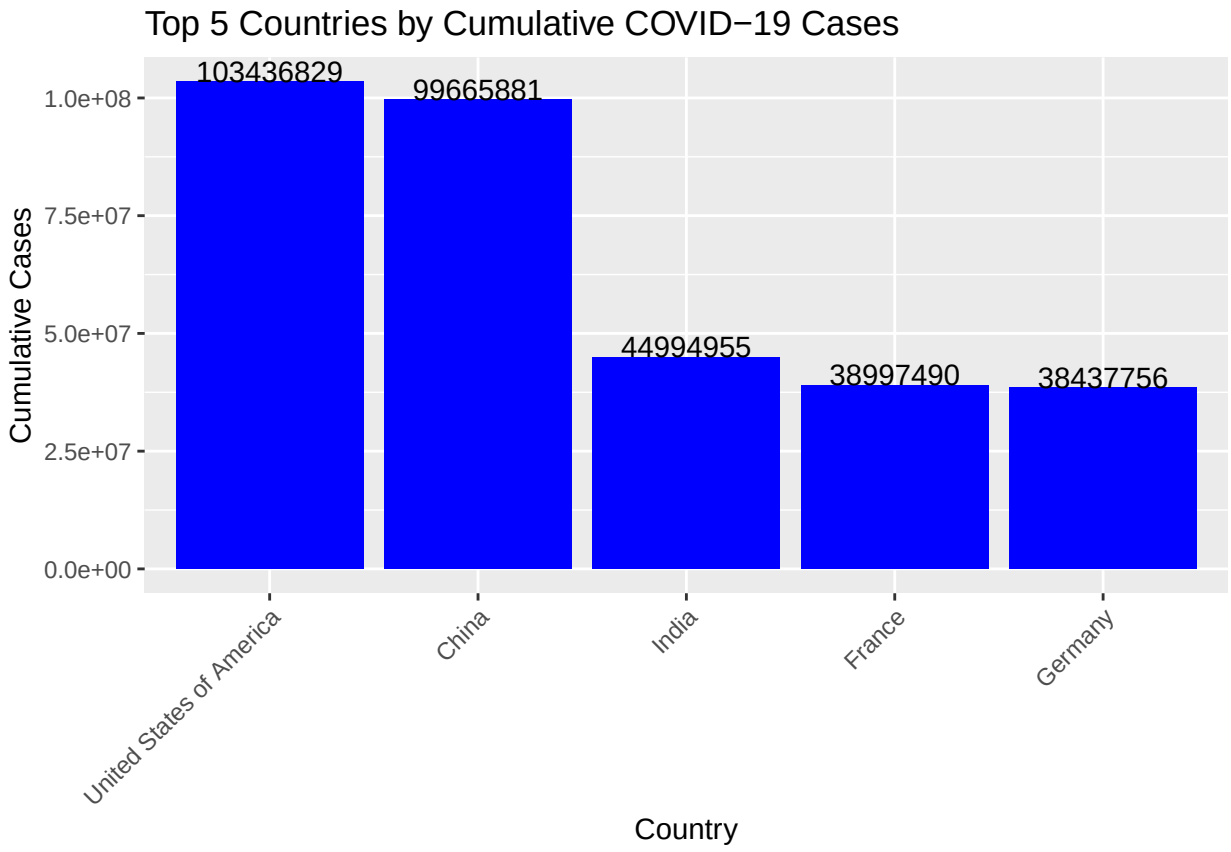
```
data_us <- data %>%
  filter(Country == "United States of America")

ggplot(data_us, aes(x = Date_reported)) +
  geom_line(aes(y = New_cases, color = "New Cases")) +
  geom_line(aes(y = New_deaths, color = "New Deaths")) +
  labs(title = "Time Series of New COVID-19 Cases and Deaths in the United States",
       y = "Count") +
  scale_color_manual(values = c("New Cases" = "blue", "New Deaths" = "red"))
```



Interpretation of the Time series line graph above

Creating a time series of New COVID-19 Cases and Deaths in the United States tracks case and death trends, aids in detecting patterns, and informs resource allocation.The graph shows that the number of new cases and deaths has fluctuated over time, but has generally decreased since the peak of the pandemic in early 2022. However, there has been a recent uptick in new cases and deaths, which may be due to the emergence of new variants of the virus or changes in public health measures.

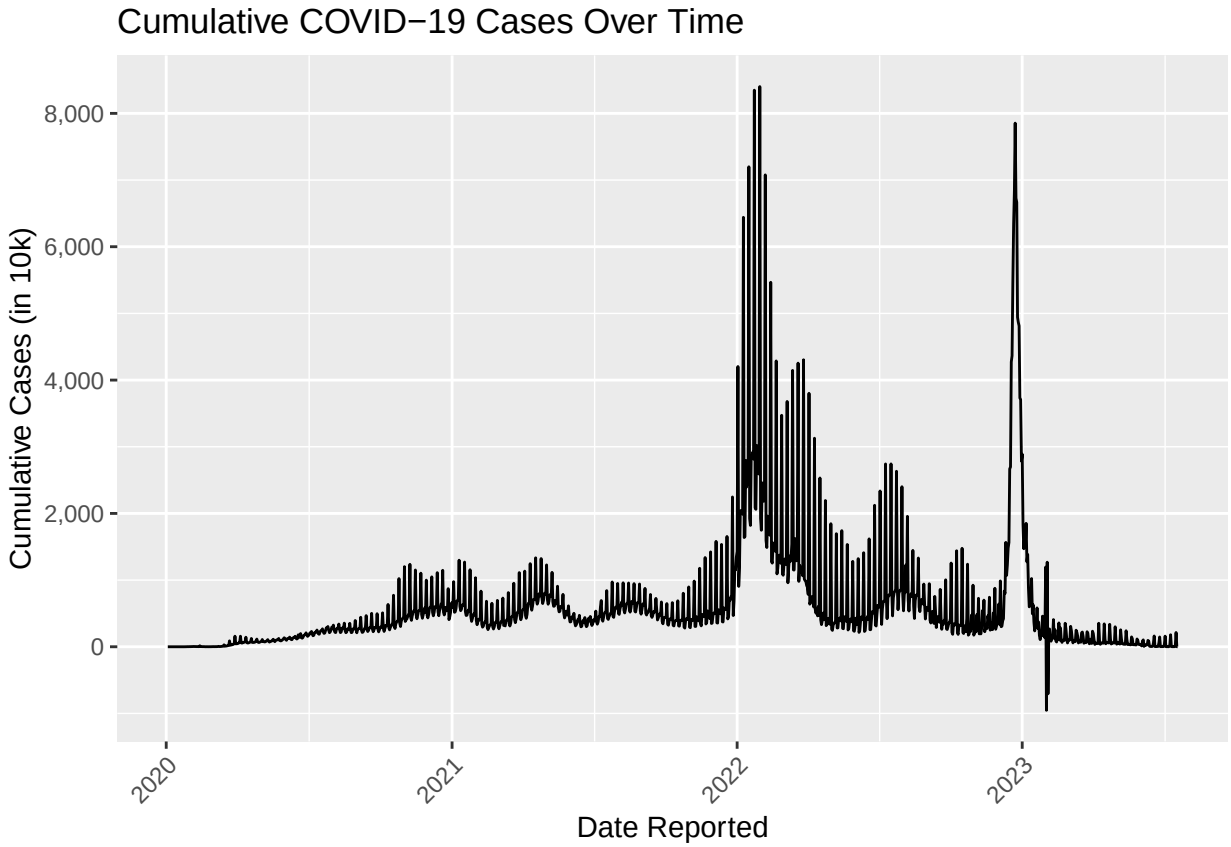
```
# Calculate the total cumulative cases for each country
top_countries <- data %>%
  group_by(Country) %>%
  summarise(Total_Cumulative_Cases = max(Cumulative_cases)) %>%
  arrange(desc(Total_Cumulative_Cases)) %>%
  slice(1:5)  # Select the top 5 countries with the highest cumulative cases

# Create a bar graph with counts on top of each bar
ggplot(top_countries, aes(x = reorder(Country, -Total_Cumulative_Cases),
                          y = Total_Cumulative_Cases)) +
  geom_bar(stat = "identity", fill = "blue") +
  geom_text(aes(label = Total_Cumulative_Cases), vjust = -0.0) +
  labs(title = "Top 5 Countries by Cumulative COVID-19 Cases",
       x = "Country", y = "Cumulative Cases") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Interpretation of the Bar graph above

Creating a bar graph of the Top 5 Countries by Cumulative COVID-19 Cases offers a straightforward comparison of case numbers, tracks changes in rankings, and provides insight into the global impact of the pandemic. The bar chart shows that the United States has the highest number of cumulative COVID-19 cases, followed by China, India, France, and Germany.

```
# Calculate the cumulative sum of cases for each date
cumulative_data <- data %>%
  group_by(Date_reported) %>%
  summarise(Cumulative_Cases = sum(New_cases))

# Create a line graph
ggplot(cumulative_data, aes(x = Date_reported, y = Cumulative_Cases)) +
  geom_line() +
  labs(title = "Cumulative COVID-19 Cases Over Time",
       x = "Date Reported", y = "Cumulative Cases (in 10k)") +
  scale_y_continuous(labels = scales::comma_format(scale = 1e-3)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Interpretation of the Line graph above

Creating a line graph for Cumulative COVID-19 Cases Over Time tracks case trends, detects patterns, and aids resource planning.The graph shows that the number of cumulative COVID-19 cases has increased steadily over time, with a few major peaks. The first peak occurred in April 2020, followed by a second peak in January 2021, and a third peak in August 2021. The graph also shows that there has been a recent uptick in cumulative cases, possibly due to the emergence of new variants of the virus or changes in public health measures.