# An analysis of COVID19 cases data

COVID dataset urls:

- Confirmed US: https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_US.csv (https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_US.csv)
- Confirmed Global: https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv (https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv)
- Deaths US: https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_US.csv (https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_US.csv)
- Deaths Global: https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_global.csv (https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_global.csv)
- Recovered Global: https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_recovered_global.csv (https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_recovered_global.csv)
- Data repo address: https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/ (https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_time_series/)

```
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
## Warning: package 'tidyr' was built under R version 4.3.2
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.5.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
to become errors
```

```
url_in <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_1
9_data/csse_covid_19_time_series/"

file_names <- c("time_series_covid19_confirmed_US.csv","time_series_covid19_confirmed_gl
obal.csv","time_series_covid19_deaths_US.csv", "time_series_covid19_deaths_global.csv")
urls <- str_c(url_in,file_names)
urls
```

```
## [1] "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_d
ata/csse_covid_19_time_series/time_series_covid19_confirmed_US.csv"
## [2] "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_d
ata/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv"
## [3] "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_d
ata/csse_covid_19_time_series/time_series_covid19_deaths_US.csv"
## [4] "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_d
ata/csse_covid_19_time_series/time_series_covid19_deaths_global.csv"
```

```
us_cases <- read_csv(urls[1])
```

```
## Rows: 3342 Columns: 1154
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr    (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (1148): UID, code3, FIPS, Lat, Long_, 1/22/20, 1/23/20, 1/24/20, 1/25/20...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
global_cases <- read_csv(urls[2])
```

```
## Rows: 289 Columns: 1147
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr    (2): Province/State, Country/Region
## dbl (1145): Lat, Long, 1/22/20, 1/23/20, 1/24/20, 1/25/20, 1/26/20, 1/27/20,...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
us_deaths <- read_csv(urls[3])
```

```
## Rows: 3342 Columns: 1155
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr    (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (1149): UID, code3, FIPS, Lat, Long_, Population, 1/22/20, 1/23/20, 1/24...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
global_deaths <- read_csv(urls[4])
```

```
## Rows: 289 Columns: 1147
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr    (2): Province/State, Country/Region
## dbl (1145): Lat, Long, 1/22/20, 1/23/20, 1/24/20, 1/25/20, 1/26/20, 1/27/20,...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Intially I will do some adjustment and manipulation to make the data easier to explore and analyze

```
global_cases_long <- global_cases %>%
  pivot_longer(cols = starts_with("1/"), names_to = "Date", values_to = "Cases")

# Convert the Date column to Date type
global_cases_long$Date <- mdy(global_cases_long$Date)

# View the cleaned data
head(global_cases_long)
```

```
## # A tibble: 6 × 1,046
##   `Province/State` `Country/Region`   Lat  Long `2/1/20` `2/2/20` `2/3/20`
##   <chr>            <chr>            <dbl> <dbl>    <dbl>    <dbl>    <dbl>
## 1 <NA>             Afghanistan       33.9  67.7        0        0        0
## 2 <NA>             Afghanistan       33.9  67.7        0        0        0
## 3 <NA>             Afghanistan       33.9  67.7        0        0        0
## 4 <NA>             Afghanistan       33.9  67.7        0        0        0
## 5 <NA>             Afghanistan       33.9  67.7        0        0        0
## 6 <NA>             Afghanistan       33.9  67.7        0        0        0
## # ℹ 1,039 more variables: `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>, `2/21/20` <dbl>, `2/22/20` <dbl>,
## #   `2/23/20` <dbl>, `2/24/20` <dbl>, `2/25/20` <dbl>, `2/26/20` <dbl>,
## #   `2/27/20` <dbl>, `2/28/20` <dbl>, `2/29/20` <dbl>, `3/1/20` <dbl>, …
```

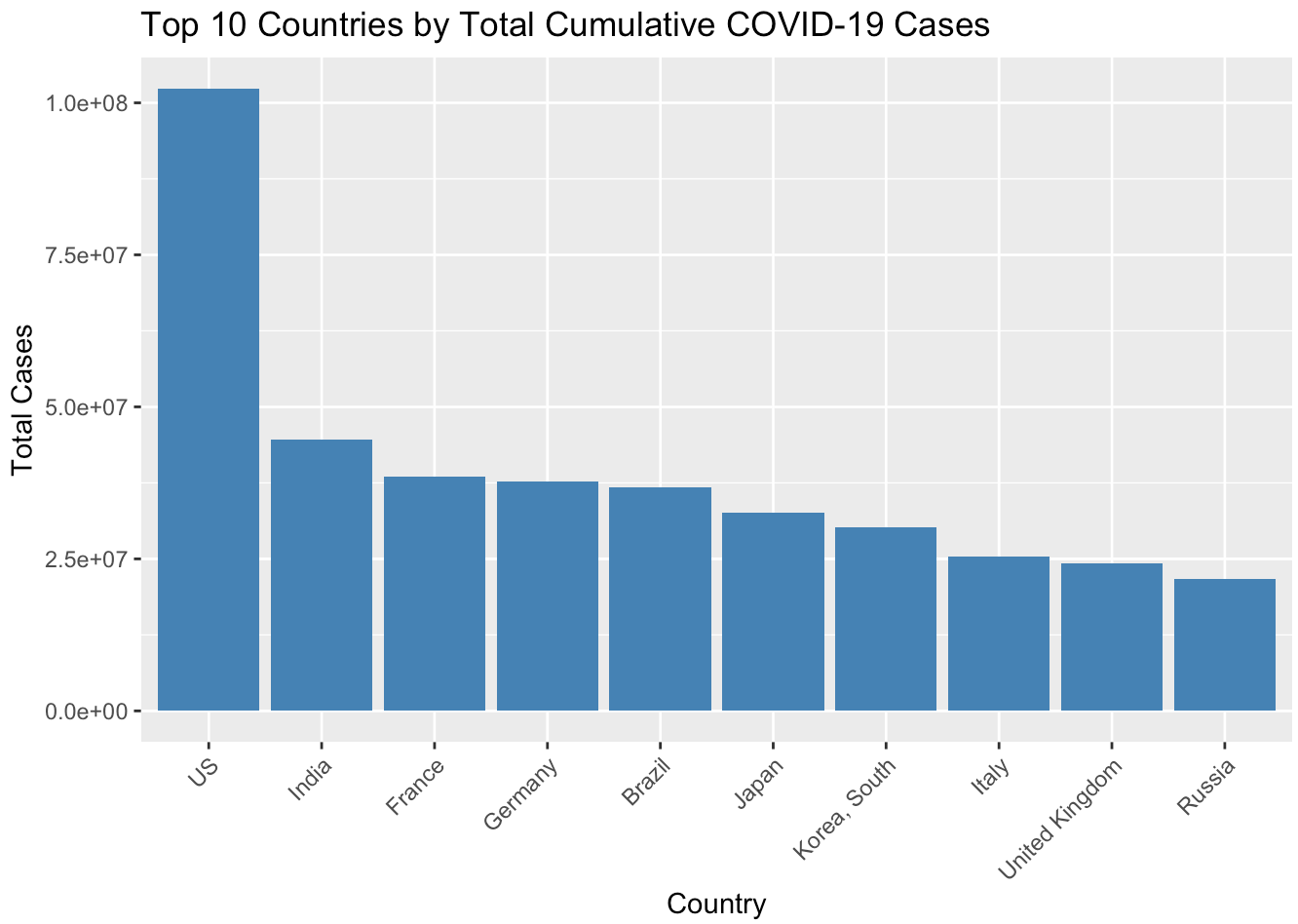
# Total number of cases

Since our data on cases seems to be cumilatives I'd like to get some stats on face value just to see what that cumulative country stats look like

```
total_cases_by_country <- global_cases_long %>%
  group_by(`Country/Region`) %>%
  summarize(Total_Cases = max(Cases, na.rm = TRUE)) %>%
  arrange(desc(Total_Cases))

# View the top 10 countries
head(total_cases_by_country, 10)
```

```
## # A tibble: 10 × 2
##    `Country/Region` Total_Cases
##    <chr>                  <dbl>
##  1 US                 102362870
##  2 India               44684120
##  3 France              38482878
##  4 Germany             37779833
##  5 Brazil              36824580
##  6 Japan               32555047
##  7 Korea, South        30197066
##  8 Italy               25453789
##  9 United Kingdom      24274357
## 10 Russia              21640952
```

```
ggplot(head(total_cases_by_country, 10), aes(x = reorder(`Country/Region`, -Total_Case
s), y = Total_Cases)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Top 10 Countries by Total Cumulative COVID-19 Cases",
       x = "Country",
       y = "Total Cases") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




It does seem intresting that the top countries by cases are likely the top countries by population while very clearly missing china but we will look at this later when we bring in some population data

```
global_deaths_long <- global_deaths %>%
  pivot_longer(cols = starts_with("1/"), names_to = "Date", values_to = "Deaths")

# Convert the Date column to Date type
global_deaths_long$Date <- mdy(global_deaths_long$Date)

# View the cleaned data
head(global_deaths_long)
```

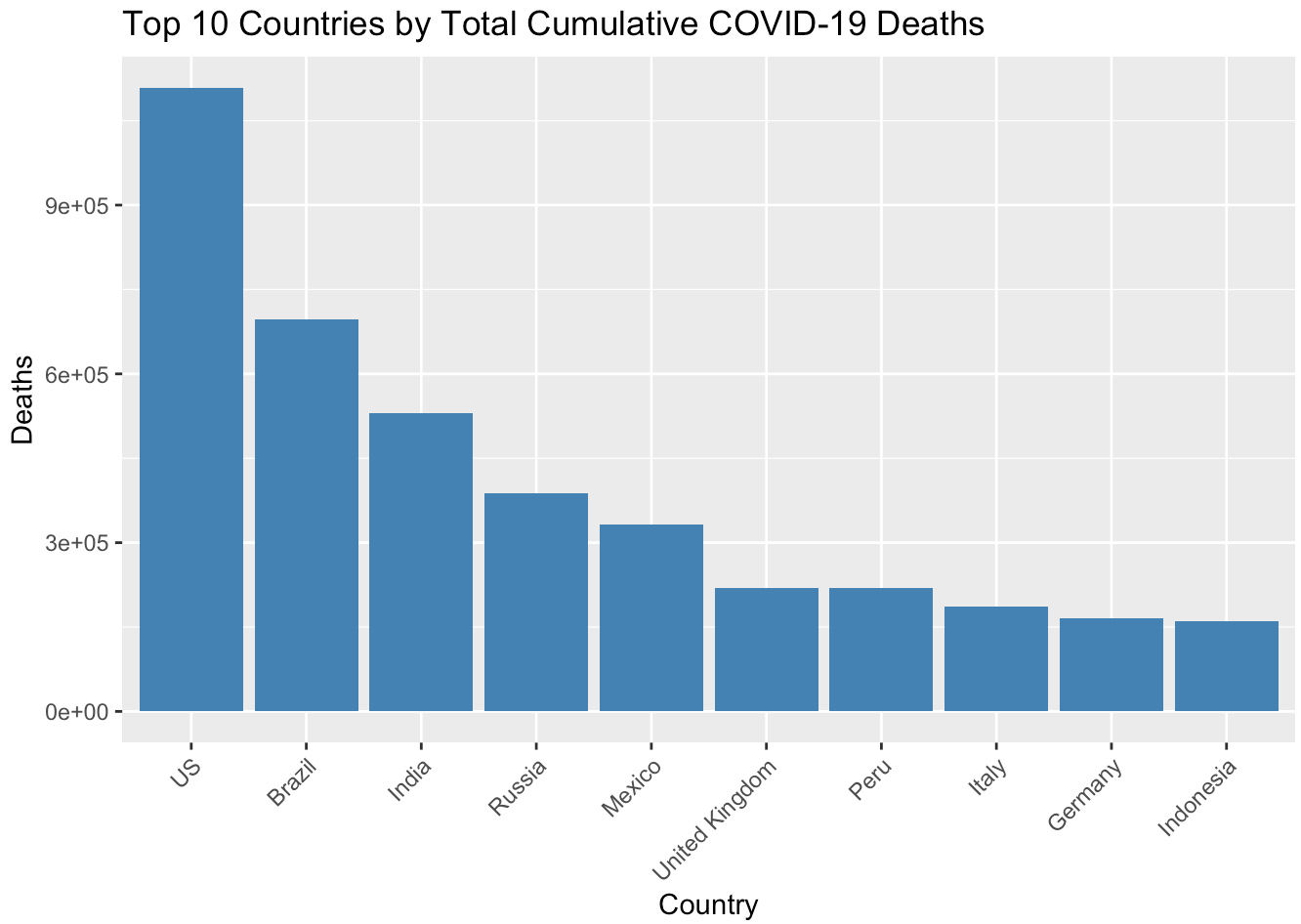
```
## # A tibble: 6 × 1,046
##   `Province/State` `Country/Region`   Lat  Long `2/1/20` `2/2/20` `2/3/20`
##   <chr>            <chr>            <dbl> <dbl>    <dbl>    <dbl>    <dbl>
## 1 <NA>             Afghanistan       33.9  67.7        0        0        0
## 2 <NA>             Afghanistan       33.9  67.7        0        0        0
## 3 <NA>             Afghanistan       33.9  67.7        0        0        0
## 4 <NA>             Afghanistan       33.9  67.7        0        0        0
## 5 <NA>             Afghanistan       33.9  67.7        0        0        0
## 6 <NA>             Afghanistan       33.9  67.7        0        0        0
## # ℹ 1,039 more variables: `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>, `2/21/20` <dbl>, `2/22/20` <dbl>,
## #   `2/23/20` <dbl>, `2/24/20` <dbl>, `2/25/20` <dbl>, `2/26/20` <dbl>,
## #   `2/27/20` <dbl>, `2/28/20` <dbl>, `2/29/20` <dbl>, `3/1/20` <dbl>, …
```

## I'd like to look at the top deaths by countries same way we looked at the cases

```
total_deaths_by_country <- global_deaths_long %>%
  group_by(`Country/Region`) %>%
  summarize(Total_Deaths = max(Deaths, na.rm = TRUE)) %>%
  arrange(desc(Total_Deaths))

# view top 10 countries by deaths
head(total_deaths_by_country, 10)
```

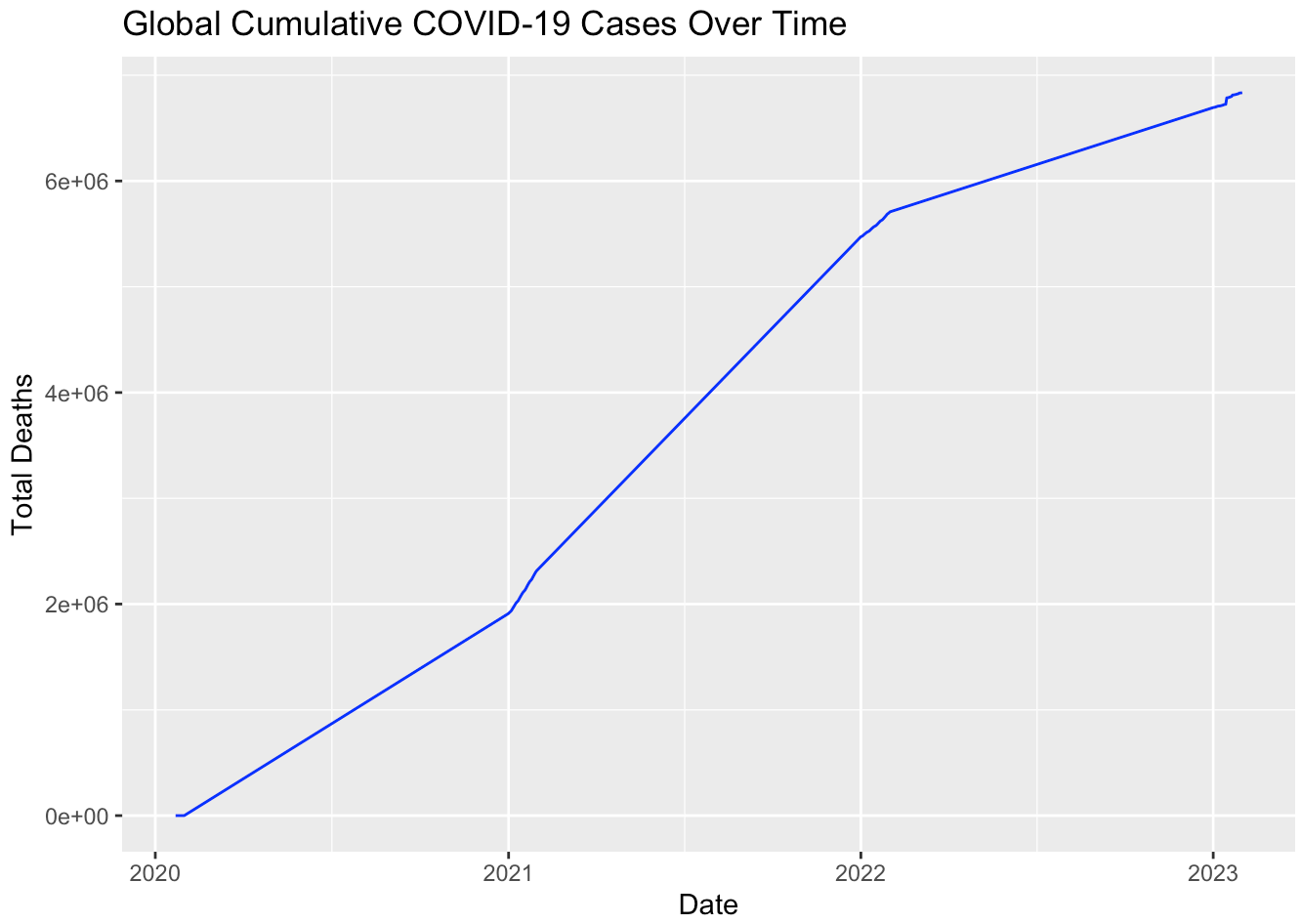
```
## # A tibble: 10 × 2
##    `Country/Region` Total_Deaths
##    <chr>                   <dbl>
##  1 US                    1108688
##  2 Brazil                 697074
##  3 India                  530740
##  4 Russia                 387113
##  5 Mexico                 332198
##  6 United Kingdom         219298
##  7 Peru                   218931
##  8 Italy                  186833
##  9 Germany                165711
## 10 Indonesia              160814
```

```
top_10_countries <- head(total_deaths_by_country, 10)
ggplot(top_10_countries, aes(x = reorder(`Country/Region`, -Total_Deaths), y = Total_Dea
ths)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Top 10 Countries by Total Cumulative COVID-19 Deaths",
       x = "Country",
       y = "Deaths") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




Roughly looking at the the deaths there also seems to be a population skew in the top couple of countries but there is also a clear effect from measures against covid

```
global_deaths_over_time <- global_deaths_long %>%
  group_by(Date) %>%
  summarize(Total_Deaths = sum(Deaths, na.rm = TRUE))

# plot global cumulative cases over time
ggplot(global_deaths_over_time, aes(x = Date, y = Total_Deaths)) +
  geom_line(color = "blue") +
  labs(title = "Global Cumulative COVID-19 Cases Over Time", x = "Date", y = "Total Deat
hs")
```




It does look like the most aggressive growth was in the 2021 -> 2022 season. Tappering off around february 2022, which is likely a slighlty lagging indicator after the two quarters of 2022 (summer and fall) where vaccine adminsitration was very high world wide at least based on this chart
https://ourworldindata.org/grapher/cumulative-covid-vaccinations
(https://ourworldindata.org/grapher/cumulative-covid-vaccinations)

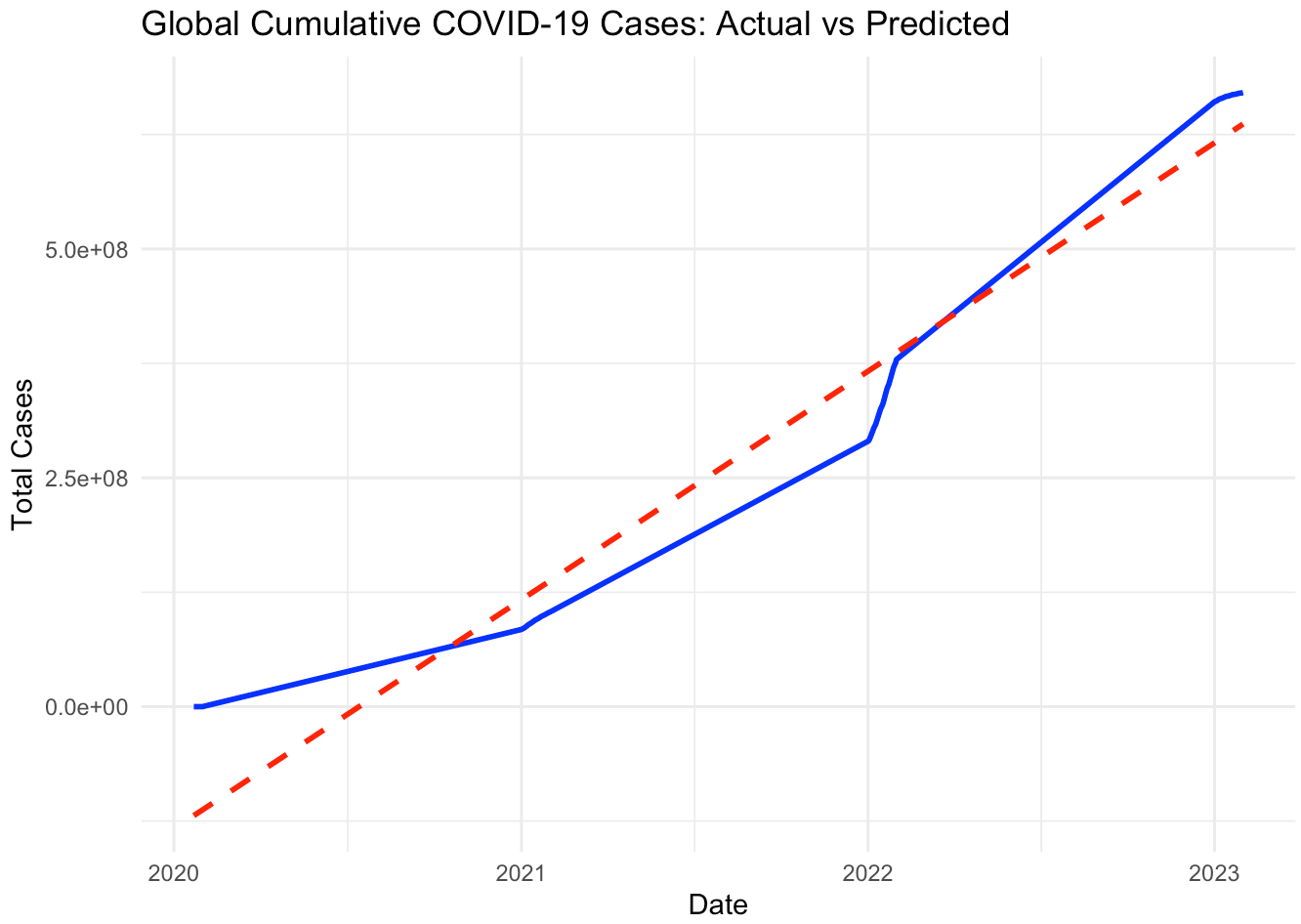
Linear Regression Model: Predicting Cumulative Cases Over Time

```
# Aggregate global cases over time
global_cases_over_time <- global_cases_long %>%
  group_by(Date) %>%
  summarize(Total_Cases = sum(Cases, na.rm = TRUE))

# Fit a linear regression model
model <- lm(Total_Cases ~ Date, data = global_cases_over_time)

# Summary of the model
summary(model)
```

```
##
## Call:
## lm(formula = Total_Cases ~ Date, data = global_cases_over_time)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -76460835 -33290857 -32227672  40831144 118981193
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.262e+10  2.852e+08  -44.24   <2e-16 ***
## Date         6.836e+05  1.506e+04   45.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54190000 on 101 degrees of freedom
## Multiple R-squared:  0.9533, Adjusted R-squared:  0.9528
## F-statistic:  2062 on 1 and 101 DF,  p-value: < 2.2e-16
```

```
# Predicting the cases using the model
global_cases_over_time <- global_cases_over_time %>%
  mutate(Predicted_Cases = predict(model, newdata = global_cases_over_time))

# Plot the actual vs predicted cases
ggplot(global_cases_over_time, aes(x = Date)) +
  geom_line(aes(y = Total_Cases), color = "blue", size = 1) +
  geom_line(aes(y = Predicted_Cases), color = "red", linetype = "dashed", size = 1) +
  labs(title = "Global Cumulative COVID-19 Cases: Actual vs Predicted",
       x = "Date",
       y = "Total Cases") +
  theme_minimal()
```

```
## 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.
```




# Analysis of model

as we can see we can create a very rough linear model to predict very basic growth of global covid cases. There are some very important things to note about this model though, it only represents a pattern that is exhibited in a short and unusual window of time. If we were to expand this model to look forward past the historical data I would say it is extremely unlikely that the model would be accurate or representative for a few reasons. 1. We don't test and publish case data as rigorously as we did during COVID (2019-2022) so that is likely to bias. The other thing that is likely to change is how social distancing, masking and boosters are not rigorously enforced/used as they were, making the fundamentals that shape the pattern very different. For these reasons, I would say this is fine for what it is used for here. The last part is that I believe the relationship over time is fundamentally not linear, I think it is likely closer to a sigmoid growth curve where over time the rate slows down as more people are vaccinated and have already gotten covid.