Covid Data Analysis

Nate Sema

2025-06-16

In this project, I analyzed COVID-19 data from 01-01-2020 to 04-30-2021 using datasets from Our World In Data: https://ourworldindata.org/covid-deaths. The goal was to look at COVID-19 cases, deaths, and vaccinations to answer key questions about the global impact of the pandemic. I explored how many total cases and deaths each country reported, examined daily new cases in the United States over time, and compared countries to see which had higher death rates relative to total cases. I also looked at the relationship between vaccinations and total cases to see if higher vaccination numbers were linked to fewer infections. Additionally, I compared daily new cases among the United States, India, and Brazil, and to account for population differences, I analyzed deaths per million people to provide a fairer comparison across countries of different sizes.

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(lubridate)
library(scales)
```

```
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

```r
library(tinytex)

#loading the datasets
deaths <- read_csv('CovidDeaths.csv')
```

```
## Rows: 85171 Columns: 26
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): iso_code, continent, location, date
## dbl (22): population, total_cases, new_cases, new_cases_smoothed, total_deat...
##
## 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.
```

```
vaccinations <- read_csv('CovidVaccinations.csv')
```

```
## Rows: 85171 Columns: 37
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (5): iso_code, continent, location, date, tests_units
## dbl (32): new_tests, total_tests, total_tests_per_thousand, new_tests_per_th...
##
## 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.
```

```
# Cleaning the data -------------------------------------------------------

#making sure all dates are converted from character to numerical
deaths$date <- mdy(deaths$date)
vaccinations$date <- mdy(vaccinations$date)
str(deaths$date)
```

```
##  Date[1:85171], format: "2020-02-24" "2020-02-25" "2020-02-26" "2020-02-27" "2020-02-28" ...
```

```
#checking for missing data by checking for TRUE/FALSE for each cell (is.na), then counting how many TRU
colSums(is.na(deaths))
```

```
##                           iso_code                          continent
##                                  0                               4111
##                           location                               date
##                                  0                                  0
##                         population                        total_cases
##                                549                               2099
##                          new_cases                 new_cases_smoothed
##                               2101                               3102
##                       total_deaths                         new_deaths
##                              11763                              11605
##                new_deaths_smoothed            total_cases_per_million
##                               3102                               2548
##              new_cases_per_million     new_cases_smoothed_per_million
##                               2550                               3546
##           total_deaths_per_million             new_deaths_per_million
##                              12199                              12041
##    new_deaths_smoothed_per_million                  reproduction_rate
##                               3546                              16229
##                       icu_patients           icu_patients_per_million
##                              76487                              76487
##                       hosp_patients          hosp_patients_per_million 
##                               74357                              74357 
##               weekly_icu_admissions  weekly_icu_admissions_per_million 
##                               84382                              84382 
##              weekly_hosp_admissions weekly_hosp_admissions_per_million 
##                               83876                              83876 
```

```
colSums(is.na(vaccinations))
```

```
##                                iso_code                              continent 
##                                       0                                   4111 
##                                location                                   date 
##                                       0                                      0 
##                               new_tests                            total_tests 
##                                   46226                                  46519 
##                total_tests_per_thousand                 new_tests_per_thousand 
##                                   46519                                  46226 
##                      new_tests_smoothed        new_tests_smoothed_per_thousand 
##                                   40546                                  40546 
##                           positive_rate                         tests_per_case 
##                                   42267                                  42860 
##                             tests_units                     total_vaccinations 
##                                   39092                                  75797 
##                       people_vaccinated                people_fully_vaccinated 
##                                   76427                                  78740 
##                        new_vaccinations              new_vaccinations_smoothed 
##                                   77217                                  70079 
##          total_vaccinations_per_hundred          people_vaccinated_per_hundred 
##                                   75797                                  76427 
##     people_fully_vaccinated_per_hundred new_vaccinations_smoothed_per_million 
##                                   78740                                  70079 
##                        stringency_index                     population_density 
##                                   12964                                   5897 
##                              median_age                          aged_65_older 
##                                    8465                                   9341 
##                           aged_70_older                         gdp_per_capita 
##                                    8895                                   8125 
##                         extreme_poverty                  cardiovasc_death_rate 
##                                   32722                                   7537 
##                     diabetes_prevalence                         female_smokers 
##                                    6392                                  24343 
##                            male_smokers                 handwashing_facilities 
##                                   25240                                  46164 
##              hospital_beds_per_thousand                        life_expectancy 
##                                   14324                                   4338 
##                 human_development_index 
##                                    7654
```

```
#Filtering out any rows of data where location AND date do not exist
deaths_clean <- deaths %>%
  filter(!is.na(location) & !is.na(date))

vaccinations_clean <- vaccinations %>%
```

```r
  filter(!is.na(location) & !is.na(date))

#joining the two data sets to see covid deaths and vaccinations together for each place and date.

covid_data <- inner_join(deaths_clean, vaccinations_clean,
                         by = c('location', 'date'))

head(covid_data)
```

```
## # A tibble: 6 x 61
##   iso_code.x continent.x location    date       population total_cases new_cases
##   <chr>      <chr>       <chr>       <date>          <dbl>       <dbl>     <dbl>
## 1 AFG        Asia        Afghanistan 2020-02-24   38928341           1         1
## 2 AFG        Asia        Afghanistan 2020-02-25   38928341           1         0
## 3 AFG        Asia        Afghanistan 2020-02-26   38928341           1         0
## 4 AFG        Asia        Afghanistan 2020-02-27   38928341           1         0
## 5 AFG        Asia        Afghanistan 2020-02-28   38928341           1         0
## 6 AFG        Asia        Afghanistan 2020-02-29   38928341           1         0
## # i 54 more variables: new_cases_smoothed <dbl>, total_deaths <dbl>,
## #   new_deaths <dbl>, new_deaths_smoothed <dbl>, total_cases_per_million <dbl>,
## #   new_cases_per_million <dbl>, new_cases_smoothed_per_million <dbl>,
## #   total_deaths_per_million <dbl>, new_deaths_per_million <dbl>,
## #   new_deaths_smoothed_per_million <dbl>, reproduction_rate <dbl>,
## #   icu_patients <dbl>, icu_patients_per_million <dbl>, hosp_patients <dbl>,
## #   hosp_patients_per_million <dbl>, weekly_icu_admissions <dbl>, ...
```

Questions to answer: 1: How many total cases and deaths were there per country? 2: What did the daily new cases look like over time for the United States? 3: Which countries had higher death rates relative to total cases? 4: Looking at vaccinations vs cases to see if countries with more vaccinations had fewer cases. 5: How does the US compare with India, and Brazil on daily new cases 6: Since total deaths alone can be misleading as big countries will have bigger totals, what would deaths per million look like?

```r
# Question 1: How many total cases and deaths were there per country?

total_cases <- covid_data %>%
  filter(!location %in% c("World", "Europe", "European Union",
                          "Asia", "South America", "North America",
                          "Africa", "Oceania")) %>% #filtering world/continent totals
  group_by(location) %>%
  summarise(
    total_cases = max(total_cases, na.rm = TRUE),
    total_deaths = max(total_deaths, na.rm = TRUE)
  ) %>%
  slice_max(total_cases, n = 10) #Displaying the top 10
```

```
## Warning: There were 48 warnings in 'summarise()'.
## The first warning was:
## i In argument: 'total_cases = max(total_cases, na.rm = TRUE)'.
## i In group 6: 'location = "Anguilla"'.
## Caused by warning in 'max()':
## ! no non-missing arguments to max; returning -Inf
## i Run 'dplyr::last_dplyr_warnings()' to see the 47 remaining warnings.
```

```r
print(total_cases)
```

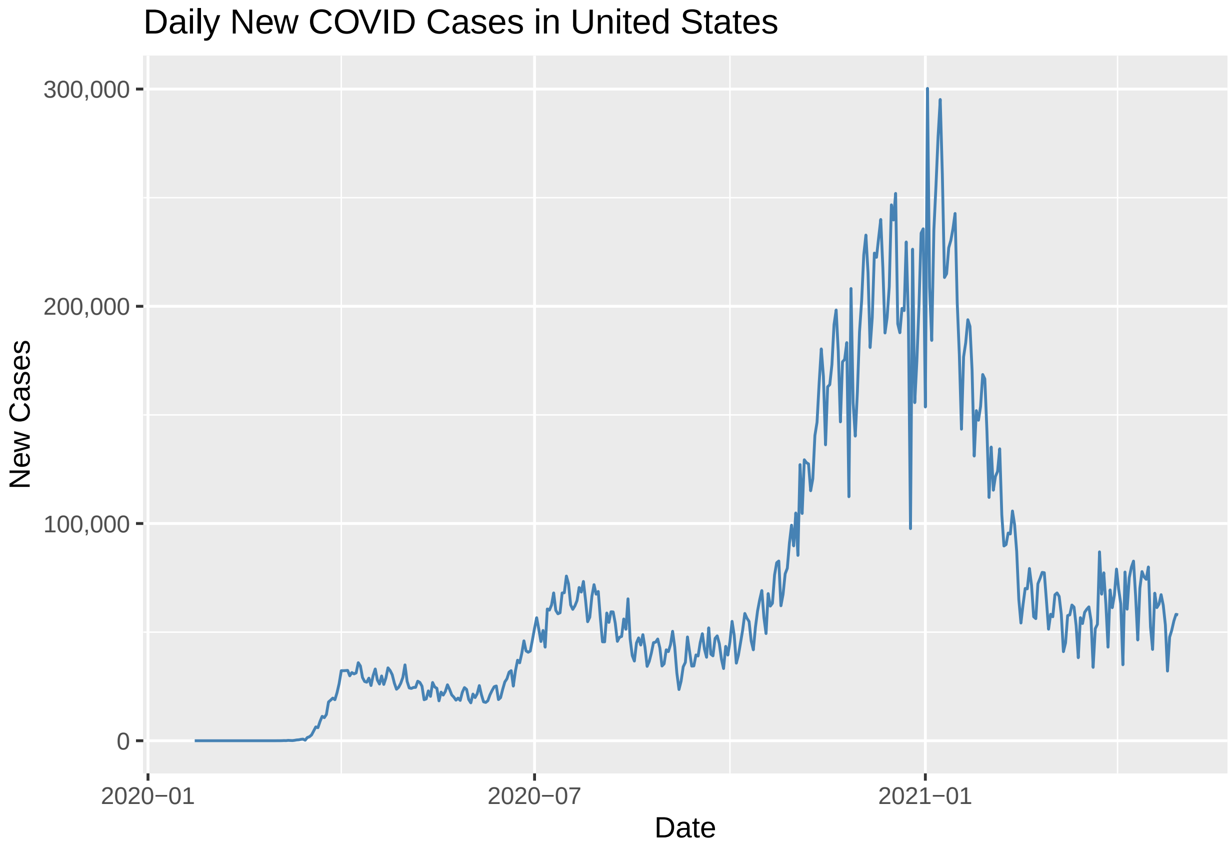
```
## # A tibble: 10 x 3
##    location       total_cases total_deaths
##    <chr>                <dbl>        <dbl>
##  1 United States     32346971       576232
##  2 India             19164969       211853
##  3 Brazil            14659011       403781
##  4 France             5677835       104675
##  5 Turkey             4820591        40131
##  6 Russia             4750755       108290
##  7 United Kingdom     4432246       127775
##  8 Italy              4022653       120807
##  9 Spain              3524077        78216
## 10 Germany            3405365        83097
```

Insight:

- At the time the data was collected, The United States had the highest cases (32,346,971) and the highest deaths (576,232).

```r
# Question 2: What did the daily new cases look like over time for the United States?

country_data <- deaths_clean %>%
  filter(location == "United States")

ggplot(country_data, aes(x = date, y = new_cases)) +
  geom_line(color = 'steelblue') +
  labs(title = 'Daily New COVID Cases in United States',
       x = 'Date',
       y = 'New Cases') +
  scale_y_continuous(labels = comma)
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (‘geom_line()‘).
```



Insight:

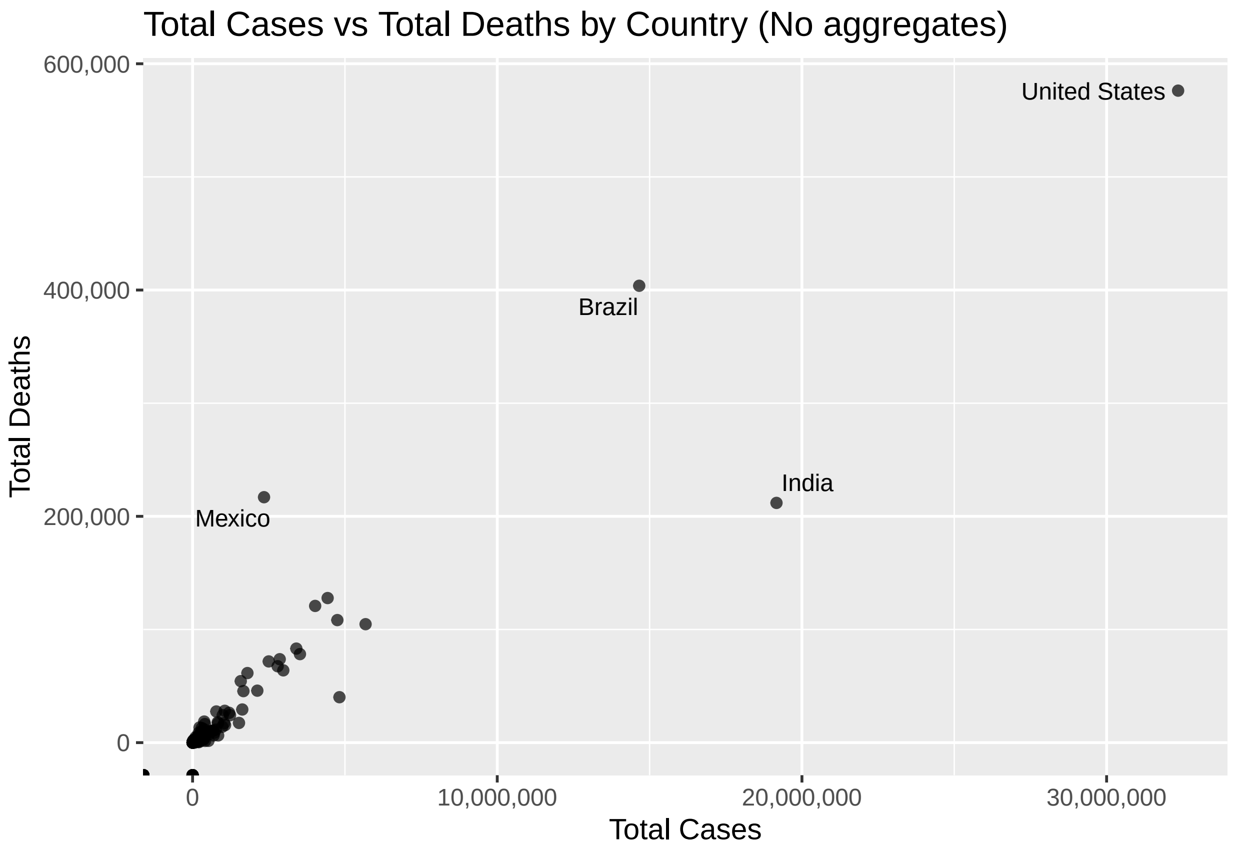
- Shows multiple waves of infection.
- The largest peak was during the winter months (Nov–Jan).
- After this peak, daily new cases decreased but did not drop to early-pandemic levels.

```
#Question 3: Which countries had higher death rates relative to total cases?

library(ggrepel)

country_totals <- deaths_clean %>%
  group_by(location) %>%
  summarise(total_cases = max(total_cases, na.rm = TRUE),
            total_deaths = max(total_deaths, na.rm = TRUE))
```

```
## Warning: There were 48 warnings in ‘summarise()‘.
## The first warning was:
## i In argument: ‘total_cases = max(total_cases, na.rm = TRUE)‘.
## i In group 7: ‘location = "Anguilla"‘.
## Caused by warning in ‘max()‘:
## ! no non-missing arguments to max; returning -Inf
## i Run ‘dplyr::last_dplyr_warnings()‘ to see the 47 remaining warnings.
```

```r
# filtering out continents/global totals to focus on countries

country_totals_filtered <- country_totals %>%
  filter(!location %in% c('World', 'Europe', 'European Union',
                         'Asia', 'South America', 'North America',
                         'Africa', 'Oceania'))

# plotting filtered data
ggplot(country_totals_filtered, aes(x = total_cases, y = total_deaths, label = location)) +
  geom_point(alpha = 0.7) +
  geom_text_repel(size = 3) +
  labs(title = "Total Cases vs Total Deaths by Country (No aggregates)",
       x = "Total Cases",
       y = "Total Deaths") +
  scale_x_continuous(labels = scales::comma) +
  scale_y_continuous(labels = scales::comma)
```

```
## Warning: ggrepel: 207 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



Insight:

- United States: Is on the far right and highest up indicating it had the highest total cases and highest total deaths.

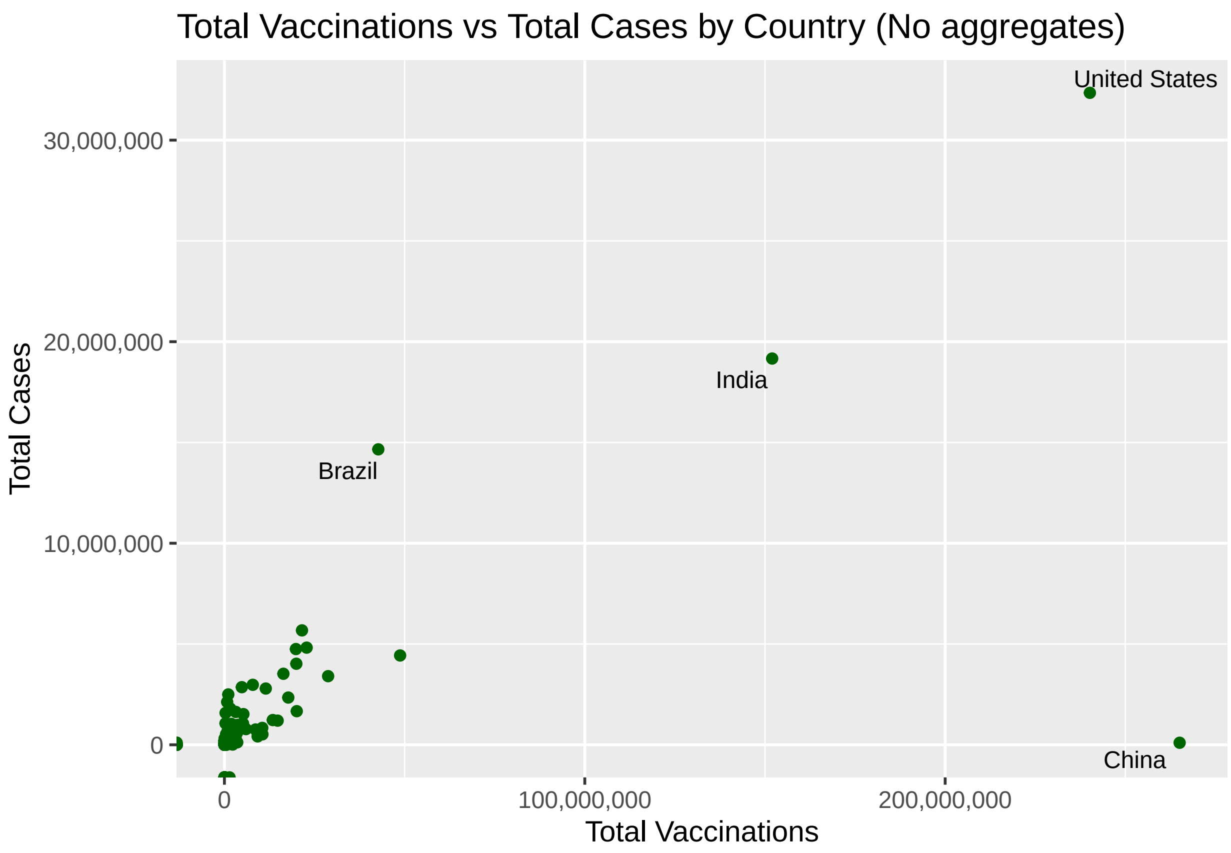
- Brazil: Has high deaths compared to its total cases, showing a high fatality count for its caseload.
- Mexico: Has fewer cases than Brazil or India but a high death count — suggesting a higher death rate per reported case.

```
# Question 4: Looking at vaccinations vs cases to see if countries with more vaccinations had fewer cas

vacc_cases <- covid_data %>%
  filter(!location %in% c("World", "Europe", "European Union",
                          "Asia", "South America", "North America",
                          "Africa", "Oceania")) %>% #filtering world/continent totals
  group_by(location) %>%
  summarise(
    total_vaccinations = max(total_vaccinations, na.rm = TRUE),
    total_cases = max(total_cases, na.rm = TRUE)
  )
```

```
## Warning: There were 40 warnings in ‘summarise()‘.
## The first warning was:
## i In argument: ‘total_vaccinations = max(total_vaccinations, na.rm = TRUE)‘.
## i In group 21: ‘location = "Benin"‘.
## Caused by warning in ‘max()‘:
## ! no non-missing arguments to max; returning -Inf
## i Run ‘dplyr::last_dplyr_warnings()‘ to see the 39 remaining warnings.
```

```
#plotting filtered data

ggplot(vacc_cases, aes(x = total_vaccinations, y = total_cases, label = location)) +
  geom_point(color = "darkgreen") +
  geom_text_repel(size = 3) +
  labs(title = "Total Vaccinations vs Total Cases by Country (No aggregates)",
       x = "Total Vaccinations",
       y = "Total Cases") +
  scale_x_continuous(labels = scales::comma) +
  scale_y_continuous(labels = scales::comma)
```

```
## Warning: ggrepel: 207 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



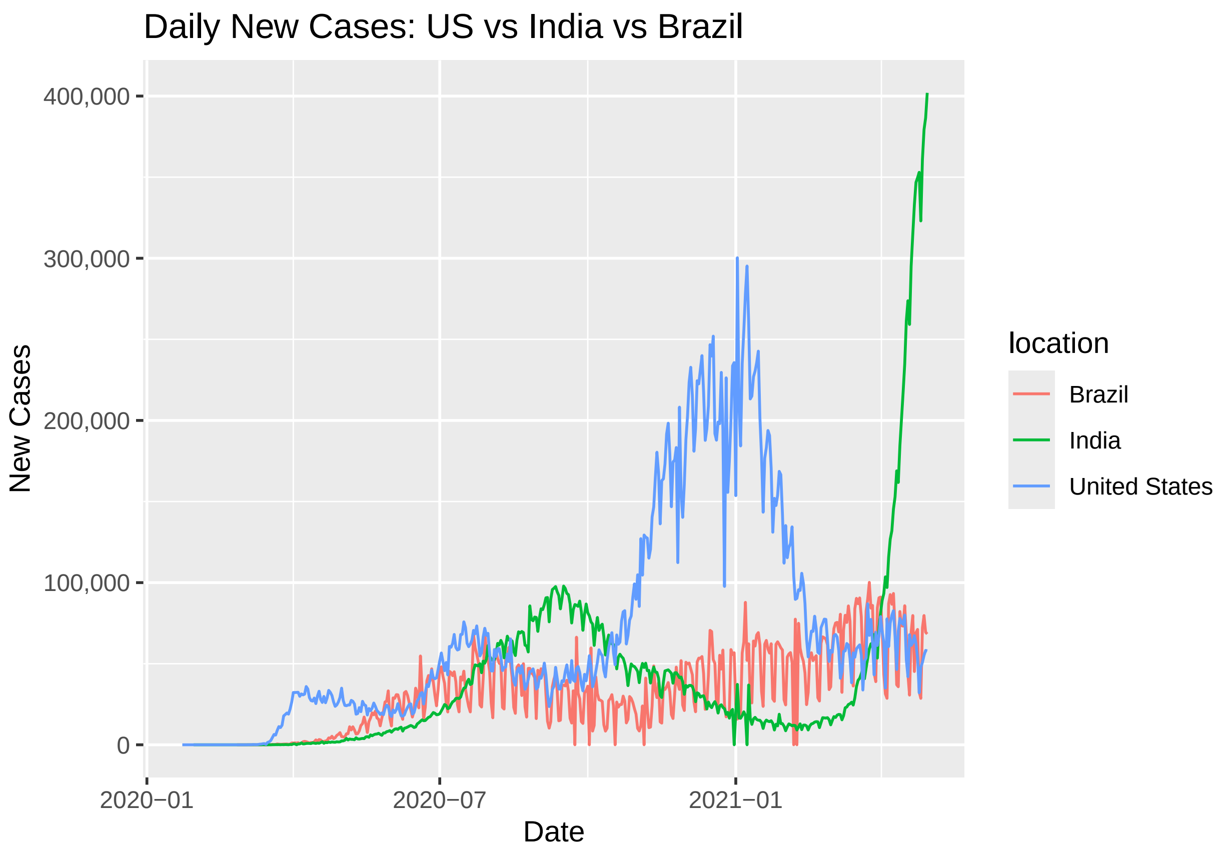
Countries to the right , have given more total vaccine doses while countires higher up have more total COVID cases over the whole pandemic. As expected, big population countries naturally had higher totals.

Insight:

- China: Far right which means they had the most doses given by far. This could be a result of their huge population but sicne they were not very high in total cases compared to the US/India, it suggests lower per-capita infection, tied to aggressive vaccination campaign and stricter control.
- US & India: Were high on both axes had very large outbreaks AND large vaccination campaigns.
- France, Brazil: Mid-range vaccinations but also high total cases.

```
# Question 5: How does the US compare with India, and Brazil on daily new cases

multi_country <- deaths_clean %>%
  filter(location %in% c('United States', 'India', 'Brazil'))

ggplot(multi_country, aes(x = date, y = new_cases, colour = location)) +
  geom_line() +
  labs(title = 'Daily New Cases: US vs India vs Brazil',
       x = 'Date',
       y = 'New Cases') +
  scale_y_continuous(labels = comma)
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (‘geom_line()‘).
```



Insight: - The US had the highest peaks earlier (winter 2020–2021) but India's spike at the end becomes very steep — showing how cases surged rapidly there.

- Brazil had a more consistent high baseline, indicating ongoing spread without sharp spikes or deep valleys.
- Each country's curve reflects different outbreak timings and possibly different containment and reporting patterns.

```
# Question 6: Since total deaths alone can be misleading as big countries will have bigger totals, what
# Deaths per-capita (million) = total deaths / population * 1000000

glimpse(deaths_clean)
```

```
## Rows: 85,171
## Columns: 26
## $ iso_code                           <chr> "AFG", "AFG", "AFG", "AFG", "AFG", ~
## $ continent                          <chr> "Asia", "Asia", "Asia", "Asia", "As~
## $ location                           <chr> "Afghanistan", "Afghanistan", "Afgh~
## $ date                               <date> 2020-02-24, 2020-02-25, 2020-02-26~
## $ population                         <dbl> 38928341, 38928341, 38928341, 38928~
## $ total_cases                        <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 2, 4, 4, 4,~
## $ new_cases                          <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 1, 2, 0, 0,~
## $ new_cases_smoothed                 <dbl> NA, NA, NA, NA, NA, 0.143, 0.143, 0~
## $ total_deaths                       <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ new_deaths                         <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ new_deaths_smoothed                <dbl> NA, NA, NA, NA, NA, 0, 0, 0, 0, 0, ~
## $ total_cases_per_million            <dbl> 0.026, 0.026, 0.026, 0.026, 0.026, ~
## $ new_cases_per_million              <dbl> 0.026, 0.000, 0.000, 0.000, 0.000, ~
## $ new_cases_smoothed_per_million     <dbl> NA, NA, NA, NA, NA, 0.004, 0.004, 0~
## $ total_deaths_per_million           <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ new_deaths_per_million             <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ new_deaths_smoothed_per_million    <dbl> NA, NA, NA, NA, NA, 0, 0, 0, 0, 0, ~
## $ reproduction_rate                  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ icu_patients                       <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ icu_patients_per_million           <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ hosp_patients                      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ hosp_patients_per_million          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ weekly_icu_admissions              <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ weekly_icu_admissions_per_million  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ weekly_hosp_admissions             <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ weekly_hosp_admissions_per_million <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,~
```

```
countries_per_capita <- deaths_clean %>%
  filter(!location %in% c("World", "Europe", "European Union",
                          "Asia", "South America", "North America",
                          "Africa", "Oceania")) %>% #filtering world/continent totals
  group_by(location) %>%
  summarise(
    total_deaths = max(total_deaths, na.rm = TRUE),
    population = max(population, na.rm = TRUE)
  ) %>%
  mutate(
    deaths_per_million = (total_deaths / population) *1000000
  )
```

```
## Warning: There were 30 warnings in ‘summarise()‘.
## The first warning was:
## i In argument: ‘total_deaths = max(total_deaths, na.rm = TRUE)‘.
## i In group 6: ‘location = "Anguilla"‘.
## Caused by warning in ‘max()‘:
## ! no non-missing arguments to max; returning -Inf
## i Run ‘dplyr::last_dplyr_warnings()‘ to see the 29 remaining warnings.
```

```
head(countries_per_capita)
```

```
## # A tibble: 6 x 4
##   location    total_deaths population deaths_per_million
##   <chr>              <dbl>      <dbl>              <dbl>
## 1 Afghanistan         2625   38928341               67.4
## 2 Albania             2394    2877800              832. 
## 3 Algeria             3253   43851043               74.2
## 4 Andorra              125      77265             1618. 
## 5 Angola               596   32866268               18.1
## 6 Anguilla            -Inf      15002             -Inf
```

```
#Plotting deaths per million in comparison to total cases
```

```r
country_per_capita <- deaths_clean %>%
  filter(!location %in% c("World", "Europe", "European Union",
                          "Asia", "South America", "North America",
                          "Africa", "Oceania")) %>% #filtering world/continent totals
  group_by(location) %>%
  summarise(
    total_cases = max(total_cases, na.rm = TRUE),
    total_deaths = max(total_deaths, na.rm = TRUE),
    population = max(population, na.rm = TRUE)
  ) %>%
  mutate(
    deaths_per_million = (total_deaths / population) * 1000000
  )
```

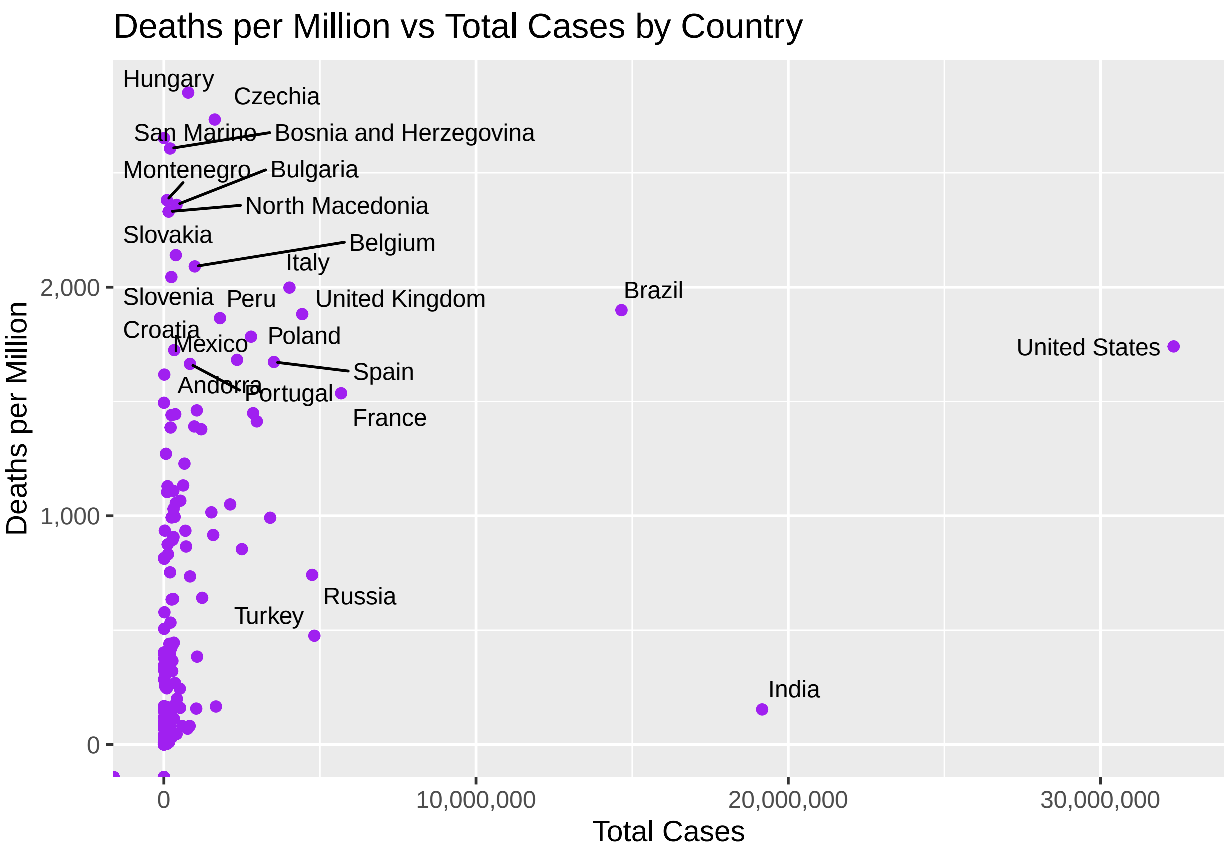
```
## Warning: There were 50 warnings in ‘summarise()‘.
## The first warning was:
## i In argument: ‘total_cases = max(total_cases, na.rm = TRUE)‘.
## i In group 6: ‘location = "Anguilla"‘.
## Caused by warning in ‘max()‘:
## ! no non-missing arguments to max; returning -Inf
## i Run ‘dplyr::last_dplyr_warnings()‘ to see the 49 remaining warnings.
```

```r
ggplot(country_per_capita, aes(x = total_cases, y = deaths_per_million, label = location)) +
  geom_point(color = "purple") +
  geom_text_repel(size = 3) +
  labs(
    title = "Deaths per Million vs Total Cases by Country",
    x = "Total Cases",
    y = "Deaths per Million"
  ) +
  scale_x_continuous(labels = scales::comma) +
  scale_y_continuous(labels = scales::comma)
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (‘geom_point()‘).
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (‘geom_text_repel()‘).
```

```
## Warning: ggrepel: 185 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



Insight:

- Countries like Hungary and Czechia stand out for having very high deaths per million despite not having the largest total case counts indicating high severity.
- Large-population countries like India report very high case numbers but appear low on deaths per million, which could reflect underreporting,or other factors.
- The US and Brazil show both large outbreaks and high per-capita death tolls, indicating significant health impacts.