Worldwide COVID-19 Data Visualization

Yutong Wang

2022-11-07

1. Dataset Description

The data was obtained from The World Health Origanization (WHO) coronavirus (COVID-19) database, containing official daily counts of COVID-19 cases and deaths reported by 237 countries, territories and areas. from January 3rd, 2020 to November 3rd, 2022.

The data can be obtained by https://covid19.who.int/data

```
# Read datasets into WHO_COVID_19_global_data
WHO_COVID_19_global_data <- read_csv("C:/Users/Yutong/Desktop/WHO-COVID-19-global-data.csv")
```

```
## Rows: 245532 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): Country_code, Country, WHO_region
## dbl  (4): New_cases, Cumulative_cases, New_deaths, Cumulative_deaths
## date (1): Date_reported
##
## 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.
```

```
# view first couple of rows
head(WHO_COVID_19_global_data)
```

```
## # A tibble: 6 x 8
##   Date_reported Country_code Country     WHO_r~1 New_c~2 Cumul~3 New_d~4 Cumul~5
##   <date>        <chr>        <chr>       <chr>     <dbl>   <dbl>   <dbl>   <dbl>
## 1 2020-01-03    AF           Afghanistan EMRO          0       0       0       0
## 2 2020-01-04    AF           Afghanistan EMRO          0       0       0       0
## 3 2020-01-05    AF           Afghanistan EMRO          0       0       0       0
## 4 2020-01-06    AF           Afghanistan EMRO          0       0       0       0
## 5 2020-01-07    AF           Afghanistan EMRO          0       0       0       0
## 6 2020-01-08    AF           Afghanistan EMRO          0       0       0       0
## # ... with abbreviated variable names 1: WHO_region, 2: New_cases,
## #   3: Cumulative_cases, 4: New_deaths, 5: Cumulative_deaths
```

```
# Retrieve the full column specification for this data
readr::spec(WHO_COVID_19_global_data)
```

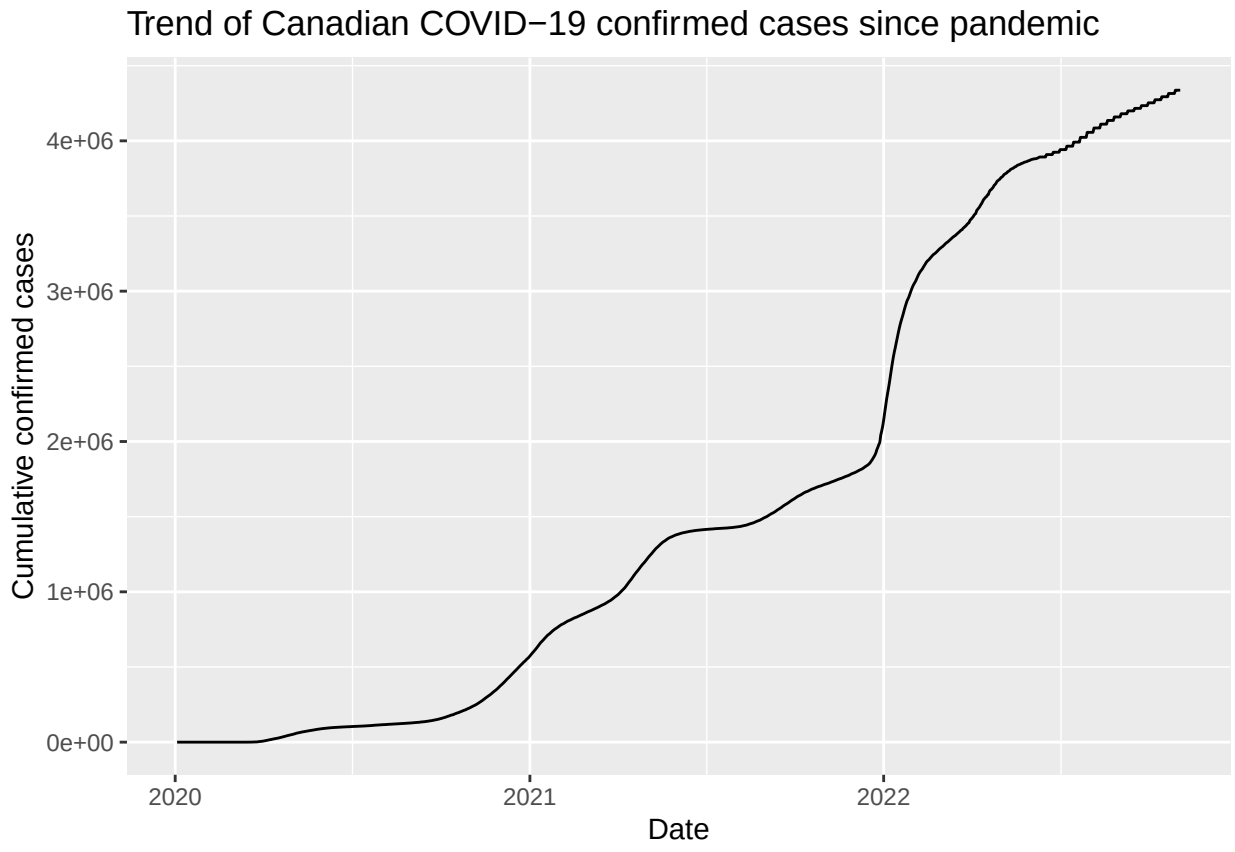
```
## cols(
```

```
##   Date_reported = col_date(format = ""),
##   Country_code = col_character(),
##   Country = col_character(),
##   WHO_region = col_character(),
##   New_cases = col_double(),
##   Cumulative_cases = col_double(),
##   New_deaths = col_double(),
##   Cumulative_deaths = col_double()
## )
```

2. Confirmed cases throughout Canada

In this section, we presented the visualization of the COVID-19 data for Canada by date

```
# Draw a line plot of cumulative cases vs. date
# Label the y-axis
options(repr.plot.width = 12, repr.plot.height = 8) # Image sizing
ggplot(WHO_COVID_19_global_data %>%
         filter(Country %in% c("Canada")), aes(Date_reported, Cumulative_cases)) +
  geom_line() + xlab("Date") +
  ylab("Cumulative confirmed cases")+
  ggtitle("Trend of Canadian COVID-19 confirmed cases since pandemic")
```

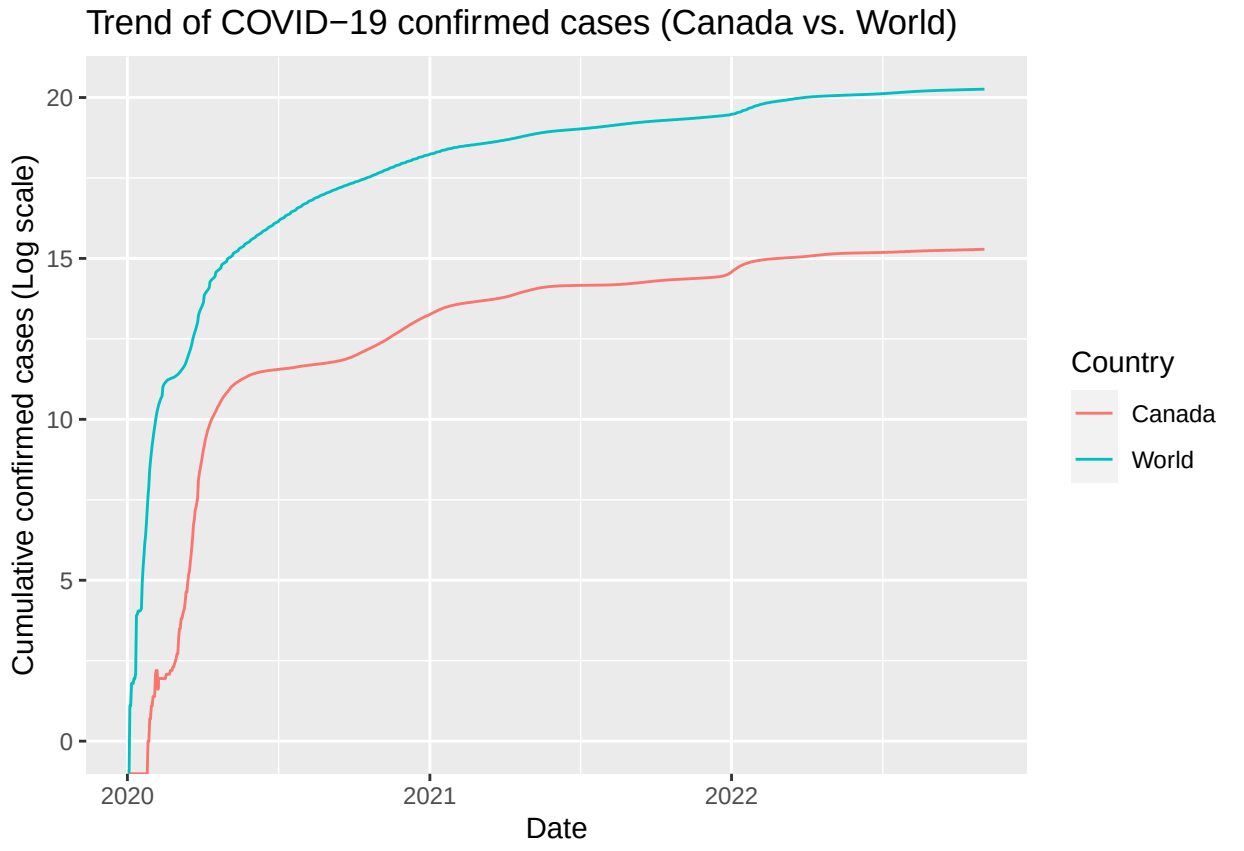


We further compared Canadian data to the rest of the world. Since Canada have much smaller number of confirmed cases compared to the rest of the world. The data was ploted in the log scale.

```r
Canada <- WHO_COVID_19_global_data %>%
                  filter(Country %in% c("Canada"))
Canada_cases_temp <- Canada[,c("Date_reported", "Country", "Cumulative_cases")]


WorldTotal_cases <- WHO_COVID_19_global_data %>%
  group_by(Date_reported) %>%                          # Specify group indicator
  summarise_at(vars(Cumulative_cases),                 # Specify column
               list(Cumulative_cases = sum))
WorldTotal_cases$Country <- rep("World", nrow(WorldTotal_cases))
plot.data <- rbind(WorldTotal_cases, Canada_cases_temp)

plt_cum_confirmed_cases_canada_vs_world <- ggplot(plot.data) +
  geom_line(aes(Date_reported, log(Cumulative_cases), color = Country)) +
  ylab("Cumulative confirmed cases (Log scale)") + xlab("Date") +
  ggtitle("Trend of COVID-19 confirmed cases (Canada vs. World)")
plt_cum_confirmed_cases_canada_vs_world
```

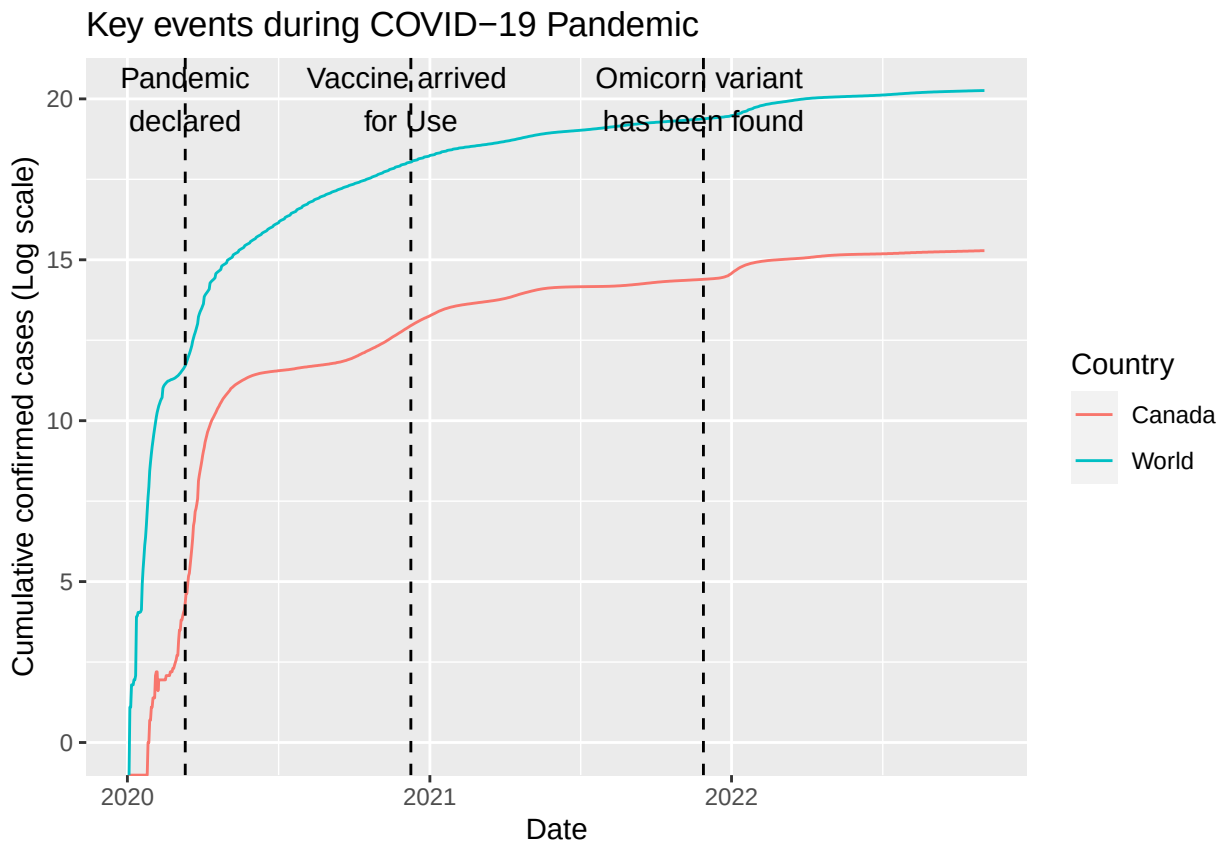


3.Annotating some key events

As we notice there were a couples of huge jumps form the graph in both of Canada & Worldwide line. Therefore, we can better interpret changes in the plot by annotating some key events. It will help us know the influences of landmark events that happened during the outbreak.

```r
who_events <- tribble(
  ~ date, ~ event,
  "2020-03-11", "Pandemic\ndeclared",
  "2020-12-09", "Vaccine arrived \nfor Use",
  "2021-11-28", "Omicorn variant \nhas been found"
) %>%
  mutate(date = as.Date(date))

# Using who_events, add vertical dashed lines with an x-intercept at date
# and text at date, ;labeled by event, and at 20 on the y-axis
plt_cum_confirmed_cases_canada_vs_world +
  geom_vline(aes(xintercept = date), data = who_events, linetype = "dashed") +
  geom_text(aes(date, label = event), data = who_events, y = 20) +
  ggtitle("Key events during COVID-19 Pandemic")
```



4. Adding a trend line to Canada

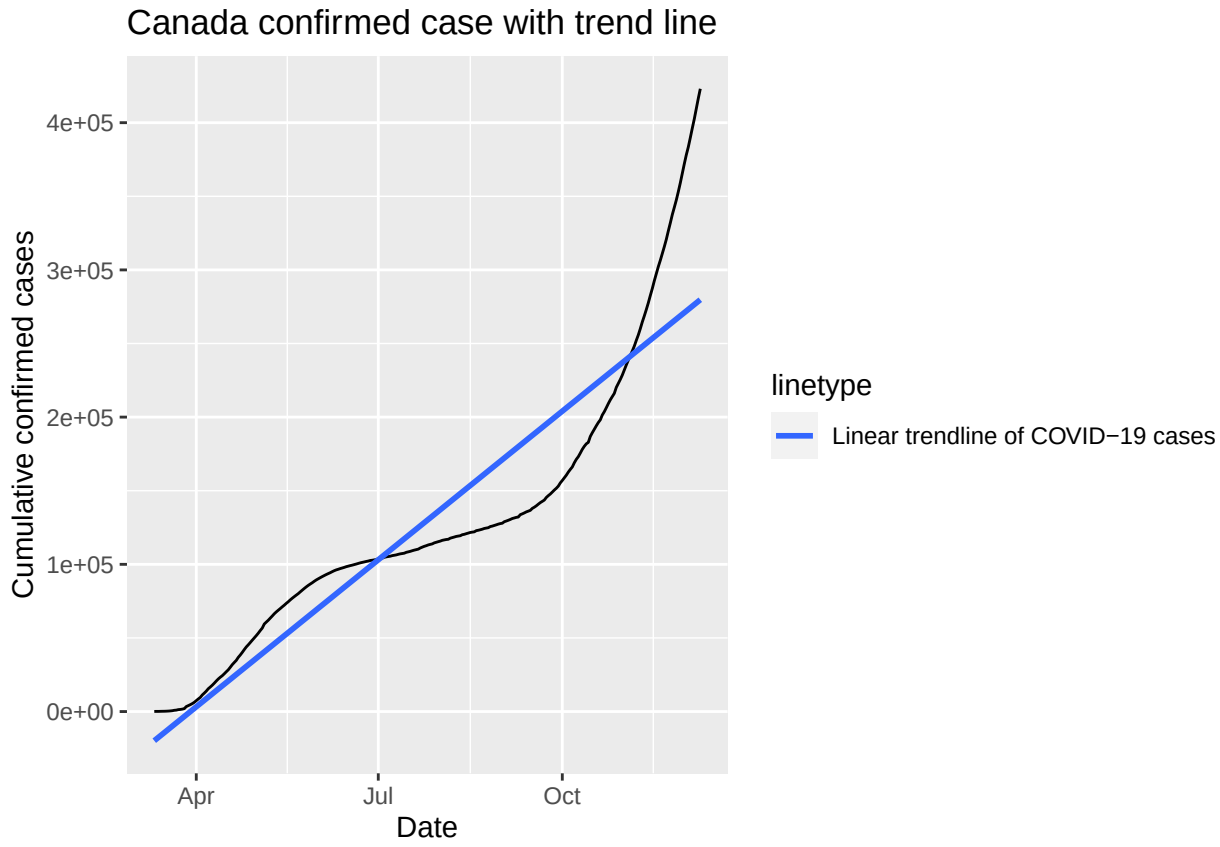
After WHO elevated COVID-19 to a pandemic on March 11th, 2020, we would like to see the bigger future of how fast the number of cases is growing within Canada in 2020. A good starting point is to see if the cases are growing faster or slower (as if the cases were grow linearly).

```r
# Filter for Canada, between Mar 11 2020 and Mar 9 2020
canada_after_mar11 <- Canada_cases_temp %>%
  filter(Country == "Canada" , Date_reported >= "2020-03-11" & Date_reported
```

```
         <="2020-12-09")

# Using Canada_after_mar11, draw a line plot cumulative_cases vs. date
# Add a smooth trend line using linear regression, no error bars
ggplot(canada_after_mar11, aes(Date_reported, Cumulative_cases)) +
  geom_line() +
  geom_smooth(method = "lm", se = FALSE, aes(lty='Linear trendline of COVID-19 cases')) + xlab("Date") +
  ylab("Cumulative confirmed cases") +
  ggtitle("Canada confirmed case with trend line")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
options(repr.plot.width = 12, repr.plot.height = 8) # Image sizing
```

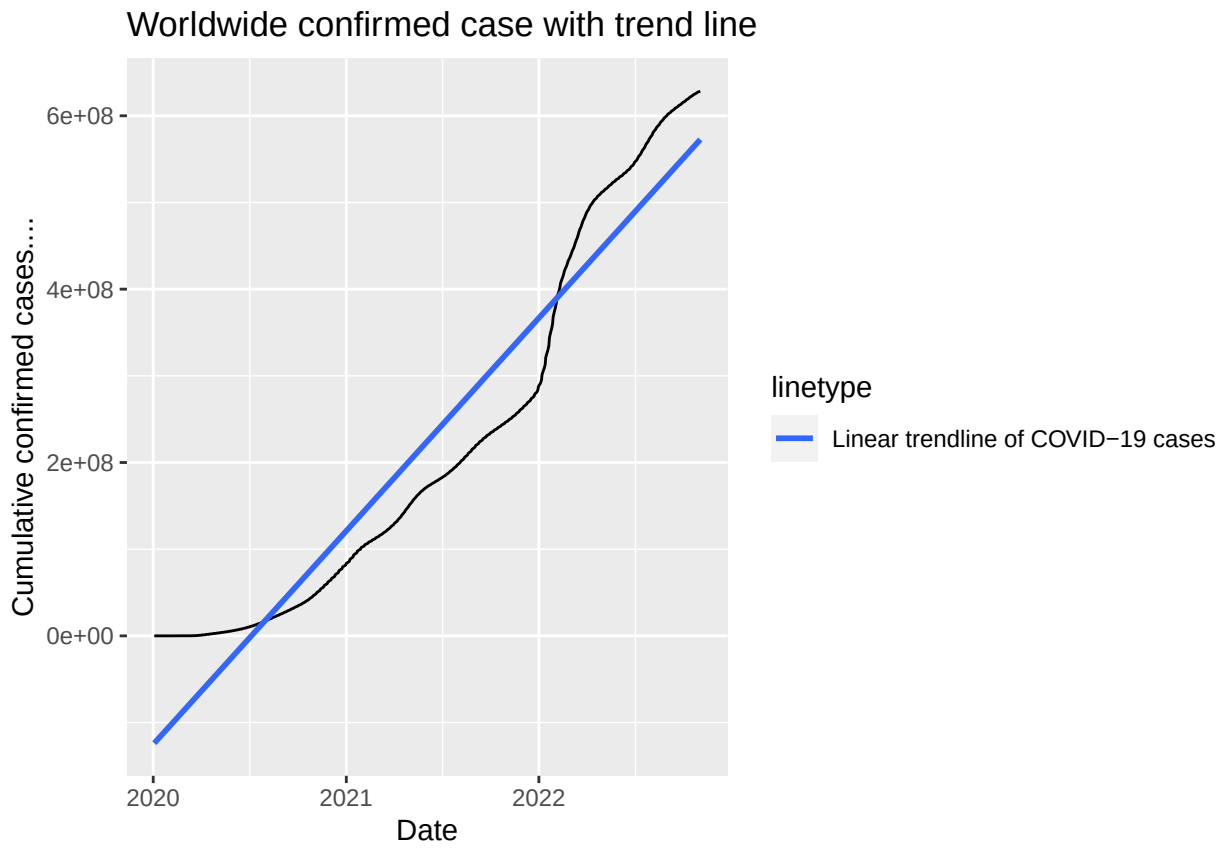
5. Adding a trend line to the rest of the world

From the above plot, the growth rate in Canada is slower than linear during the summer months,and increasing rapidly after November.Let's compare the rest of world to linear growth.

```
# Filter confirmed_cases for not Canada
not_canada <- plot.data %>%
  filter(Country == "World")
```

```
# Using not_canada, draw a line plot cumulative_cases vs. date
# Add a smooth trend line using linear regression, no error bars
plt_not_canada_trend_line <- ggplot(not_canada, aes(Date_reported, Cumulative_cases)) +
  geom_line() +
  geom_smooth(method = "lm", se = FALSE,aes(lty='Linear trendline of COVID-19 cases')) + xlab("Date") +
  ylab("Cumulative confirmed cases....")+
  ggtitle("Worldwide confirmed case with trend line")
options(repr.plot.width = 12, repr.plot.height = 8) # Image sizing
# See the result
plt_not_canada_trend_line
```

```
## `geom_smooth()` using formula 'y ~ x'
```



6.Adding a logarithmic scale

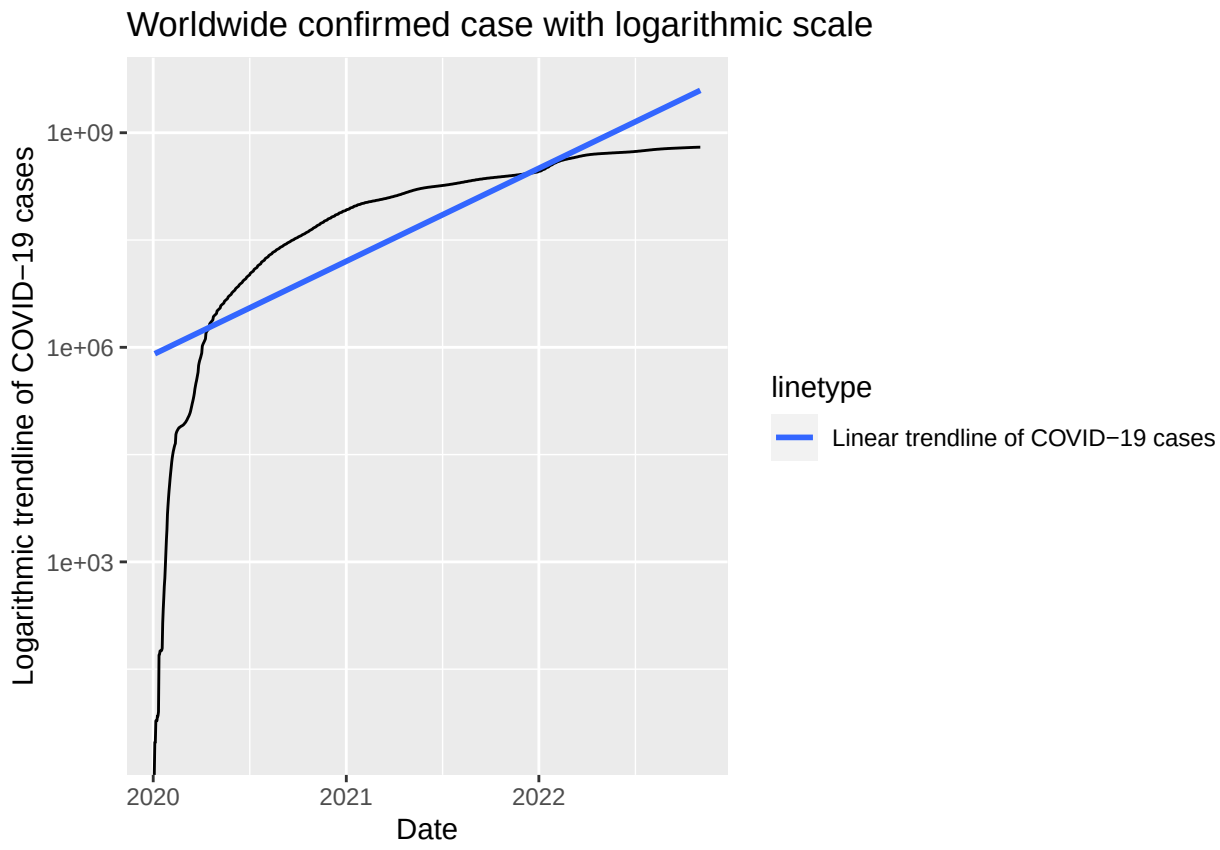
Plotting this graph in the logarithmic scale in case the straight line does not fit well.

```
# Modify the plot to use a logarithmic scale on the y-axis
plt_not_canada_trend_line +
  scale_y_log10(aes(lty='Logarithmic trendline of COVID-19 cases'))+
  ggtitle("Worldwide confirmed case with logarithmic scale")
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
## Transformation introduced infinite values in continuous y-axis
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```




```
options(repr.plot.width = 12, repr.plot.height = 8) # Image sizing
```

7.Which countries outside of Canada have been hit hardest?

Since every country is being affected by COVID-19 not equally, the following graph describe

```
# Run this to get the data for each country
confirmed_cases_by_country <- read_csv("WHO-COVID-19-global-data.csv")
```

```
## Rows: 245532 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): Country_code, Country, WHO_region
## dbl  (4): New_cases, Cumulative_cases, New_deaths, Cumulative_deaths
## date (1): Date_reported
##
## 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.
```

```r
glimpse(confirmed_cases_by_country)
```

```
## Rows: 245,532
## Columns: 8
## $ Date_reported     <date> 2020-01-03, 2020-01-04, 2020-01-05, 2020-01-06, 202~
## $ Country_code      <chr> "AF", "AF", "AF", "AF", "AF", "AF", "AF", "AF", "AF"~
## $ Country           <chr> "Afghanistan", "Afghanistan", "Afghanistan", "Afghan~
## $ WHO_region        <chr> "EMRO", "EMRO", "EMRO", "EMRO", "EMRO", "EMRO", "EMR~
## $ New_cases         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ Cumulative_cases  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ New_deaths        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ Cumulative_deaths <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
```

```r
# Group by country, summarize to calculate total cases, find the top 10
top_countries_by_total_cases <- confirmed_cases_by_country %>%
  group_by(Country) %>%
  summarize(total_cases = max(New_cases)) %>%
  top_n(10, total_cases)

# See the result
top_countries_by_total_cases
```

```
## # A tibble: 10 x 2
##    Country                  total_cases
##    <chr>                          <dbl>
##  1 Brazil                        298408
##  2 France                        500563
##  3 Germany                       307927
##  4 India                         414188
##  5 Japan                         326090
##  6 Netherlands                   391578
##  7 Republic of Korea             621328
##  8 Turkiye                       406322
##  9 United States of America     5528680
## 10 Viet Nam                      454212
```

8. Plotting hardest hit countries

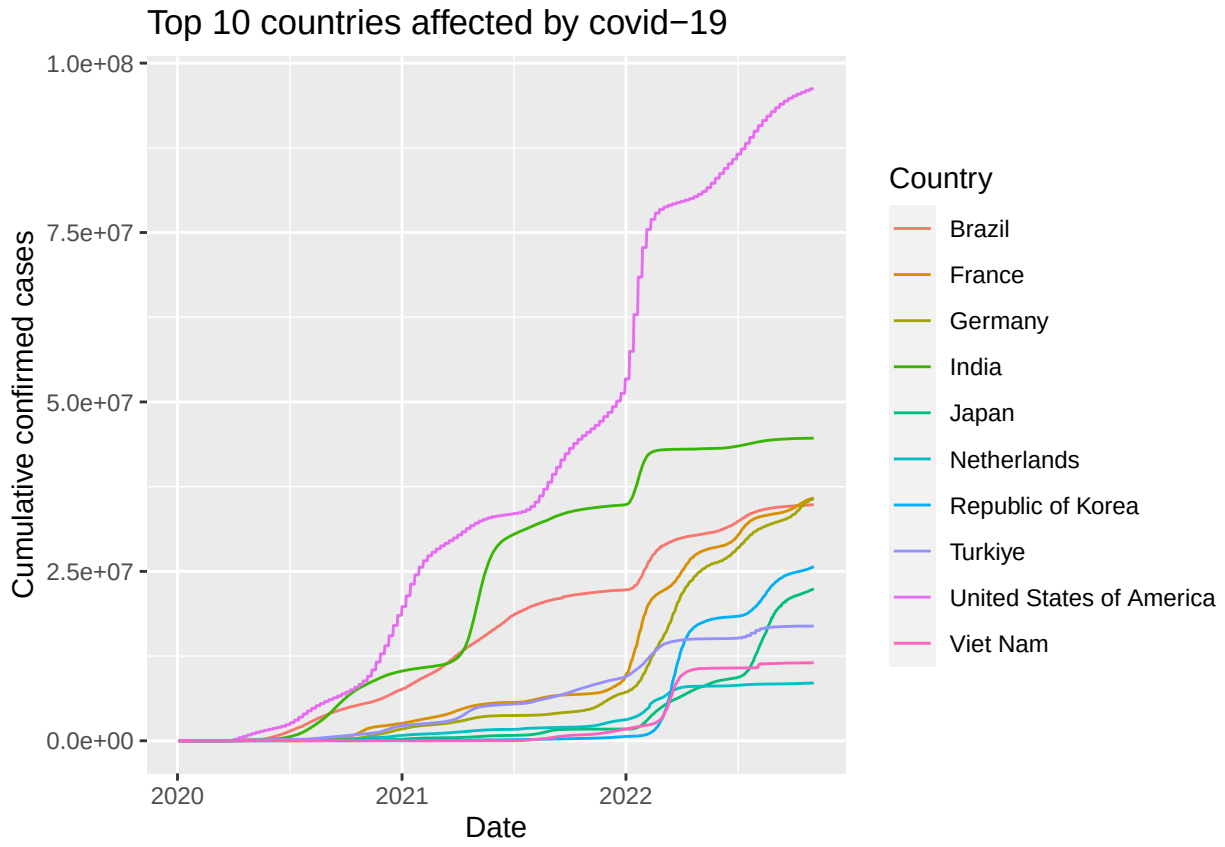
```r
# Read in the dataset from datasets/confirmed_cases_top10_outside_china.csv
confirmed_cases_top10 <-  WHO_COVID_19_global_data %>%
  filter(Country %in% top_countries_by_total_cases$Country)

# Glimpse at the contents of confirmed_cases_top10
glimpse(confirmed_cases_top10)
```

```
## Rows: 10,360
## Columns: 8
## $ Date_reported     <date> 2020-01-03, 2020-01-04, 2020-01-05, 2020-01-06, 202~
## $ Country_code      <chr> "BR", "BR", "BR", "BR", "BR", "BR", "BR", "BR", "BR"~
```

```
## $ Country           <chr> "Brazil", "Brazil", "Brazil", "Brazil", "Brazil", "B~
## $ WHO_region        <chr> "AMRO", "AMRO", "AMRO", "AMRO", "AMRO", "AMRO", "AMR~
## $ New_cases         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ Cumulative_cases  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ New_deaths        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ Cumulative_deaths <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
```

```
# Using confirmed_cases_top7, draw a line plot of
# cumulative_cases vs. date, colored by country
ggplot(confirmed_cases_top10, aes(Date_reported, Cumulative_cases, color = Country)) +
  geom_line() + xlab("Date") +
  ylab("Cumulative confirmed cases")+
  ggtitle("Top 10 countries affected by covid-19")
```




```
options(repr.plot.width = 12, repr.plot.height = 8) # Image sizing
```