COVID Cases Project

Zeeshan Amjad

2024-08-19

R Markdown

This is an R Markdown document for NYPD shooting project for Data Science as Field course.

Loading packages

- tidyverse
- dplyr
- ggplot2

```
library('tidyverse')
library('dplyr')
library(lubridate)
install.packages('ggplot2')
```

Loading COVID Data

Read the data from the url

- https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_US.csv
- https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv
- https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_US.csv
- https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_global.csv
- https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_recovered_global.csv

```
us_confirmed_cases_url <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_
us_confirmed_cases <- read_csv(us_confirmed_cases_url)
```

```
## 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_confirmed_cases_url <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_co
global_confirmed_cases <- read_csv(global_confirmed_cases_url)
```

```
## 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_death_cases_url <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_da
us_death_cases <- read_csv(us_death_cases_url)
```

```
## 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_death_cases_url <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_
global_death_cases <- read_csv(global_death_cases_url)
```

```
## 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.
```

```
global_recovered_cases_url <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_co
global_recovered_cases <- read_csv(global_recovered_cases_url)
```

```
## Rows: 274 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.
```

Preparing the data

tidying the data

```
us_confirmed_cases <- us_confirmed_cases %>%
  select(-c('UID', 'iso2', 'iso3', 'code3',
            'FIPS', 'Admin2', 'Country_Region', 'Lat', 'Long_', 'Combined_Key')) %>%
  rename('State/Territory' = 'Province_State') %>%
  pivot_longer(col = -'State/Territory', names_to = 'date', values_to = 'cases') %>%
  mutate(date=mdy(date))

us_death_cases <- us_death_cases %>%
  select(-c('UID', 'iso2', 'iso3', 'code3',
            'FIPS', 'Admin2', 'Country_Region', 'Lat', 'Long_', 'Combined_Key')) %>%
  rename('State/Territory' = 'Province_State') %>%
  pivot_longer(col = -'State/Territory', names_to = 'date', values_to = 'deaths') %>%
  mutate(date=mdy(date))
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `date = mdy(date)`.
## Caused by warning:
## !  3342 failed to parse.
```

```
global_confirmed_cases <- global_confirmed_cases %>%
  pivot_longer(cols = -c('Province/State', 'Country/Region', Lat, Long),
               names_to = 'date', values_to = 'cases') %>%
  select(-c(Lat, Long))

global_death_cases <- global_death_cases %>%
  pivot_longer(cols = -c('Province/State', 'Country/Region', Lat, Long),
               names_to = 'date', values_to = 'deaths') %>%
  select(-c(Lat, Long))

global_recovered_cases <- global_recovered_cases %>%
  pivot_longer(cols = -c('Province/State', 'Country/Region', Lat, Long),
               names_to = 'date', values_to = 'recovered') %>%
  select(-c(Lat, Long))
```

transforming the data

```
global <- global_confirmed_cases %>%
  full_join(global_death_cases) %>%
  full_join(global_recovered_cases) %>%
  mutate(date=mdy(date))
```

```
## Joining with `by = join_by(`Province/State`, `Country/Region`, date)`
## Joining with `by = join_by(`Province/State`, `Country/Region`, date)`
```

filter the date

```
global <- global %>%
  filter(cases > 0)
```

summary the data

```
summary(global)
```

```
##  Province/State     Country/Region          date                cases
##  Length:306827      Length:306827      Min.   :2020-01-22   Min.   :        1
##  Class :character   Class :character   1st Qu.:2020-12-12   1st Qu.:     1316
##  Mode  :character   Mode  :character   Median :2021-09-16   Median :    20365
##                                        Mean   :2021-09-11   Mean   :  1032863
##                                        3rd Qu.:2022-06-15   3rd Qu.:   271281
##                                        Max.   :2023-03-09   Max.   :103802702
##
##      deaths          recovered
##  Min.   :      0   Min.   :      -1
##  1st Qu.:      7   1st Qu.:       0
##  Median :    214   Median :       0
##  Mean   :  14405   Mean   :   79865
##  3rd Qu.:   3665   3rd Qu.:    1235
##  Max.   :1123836   Max.   :30974748
##                    NA's   :16010
```

Group by

```
group_by_country <- global %>%
  group_by(global$`Country/Region`) %>%
  summarise(cases = sum(cases),
            deaths = sum(deaths),
            recovered=sum(recovered))

group_by_date <- global %>%
  group_by(global$date) %>%
  summarise(cases = sum(cases, na.rm=TRUE),
            deaths = sum(deaths, na.rm=TRUE),
            recovered=sum(recovered, na.rm=TRUE))

us_confirmed_by_state <- us_confirmed_cases %>%
  group_by(`State/Territory`, date) %>%
  summarise(cases = sum(cases))
```

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

```
us_confirmed_by_state <- na.omit(us_confirmed_by_state)

us_death_by_state <- us_death_cases %>%
  group_by(`State/Territory`, date) %>%
  summarise(deaths = sum(deaths))
```

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

```
us_death_by_state <- na.omit(us_death_by_state)

us <- us_confirmed_by_state %>% full_join(us_death_by_state)
```

```
## Joining with `by = join_by(`State/Territory`, date)`
```

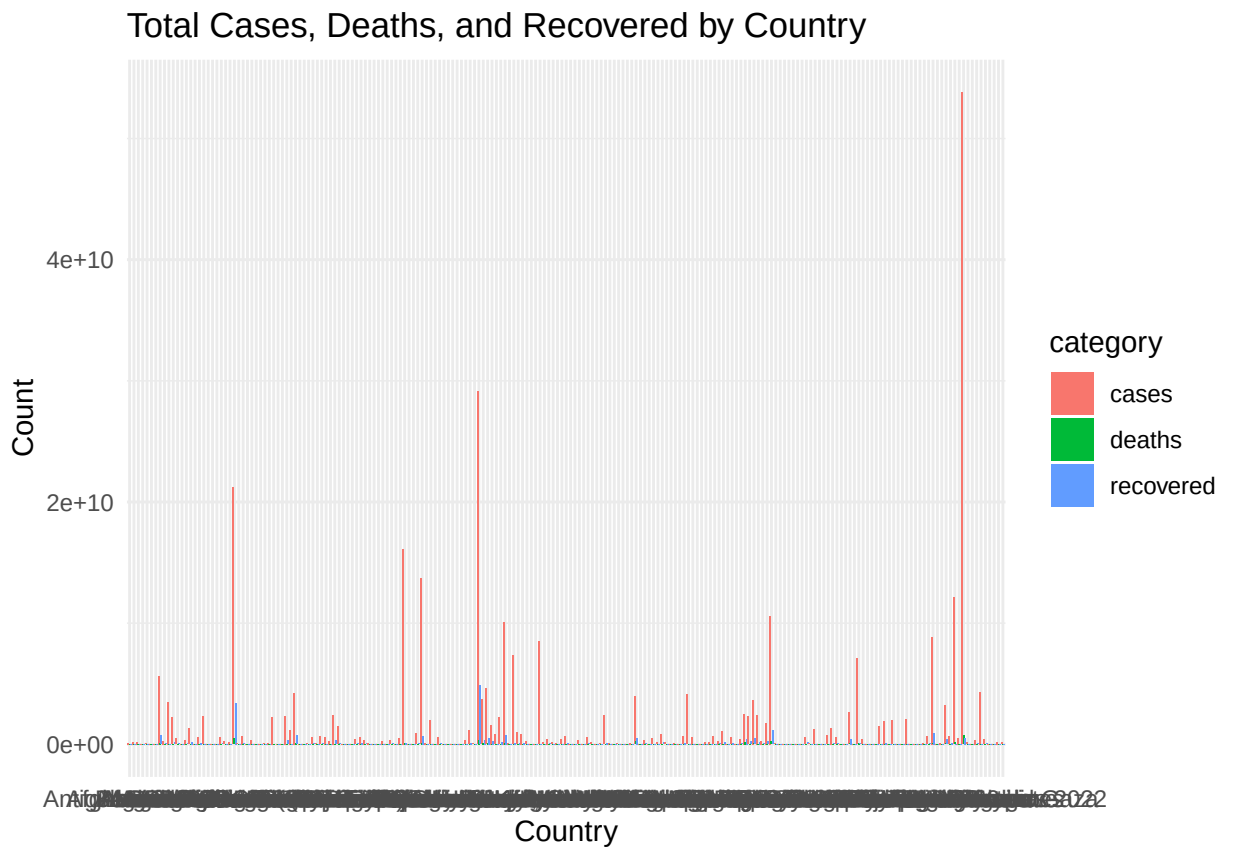
Visualization

```
summary_by_country <- group_by_country %>%
  pivot_longer(cols = c(cases, deaths, recovered),
               names_to = "category", values_to = "count")

ggplot(summary_by_country,
       aes(x = summary_by_country$`global$\`Country/Region\``,
           y = count, fill = category)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Total Cases, Deaths, and Recovered by Country",
       x = "Country", y = "Count") +
  theme_minimal()
```

```
## Warning: Use of `` summary_by_country$`global$\`Country/Region\`` `` is discouraged.
## i Use `` global$`Country/Region` `` instead.
```

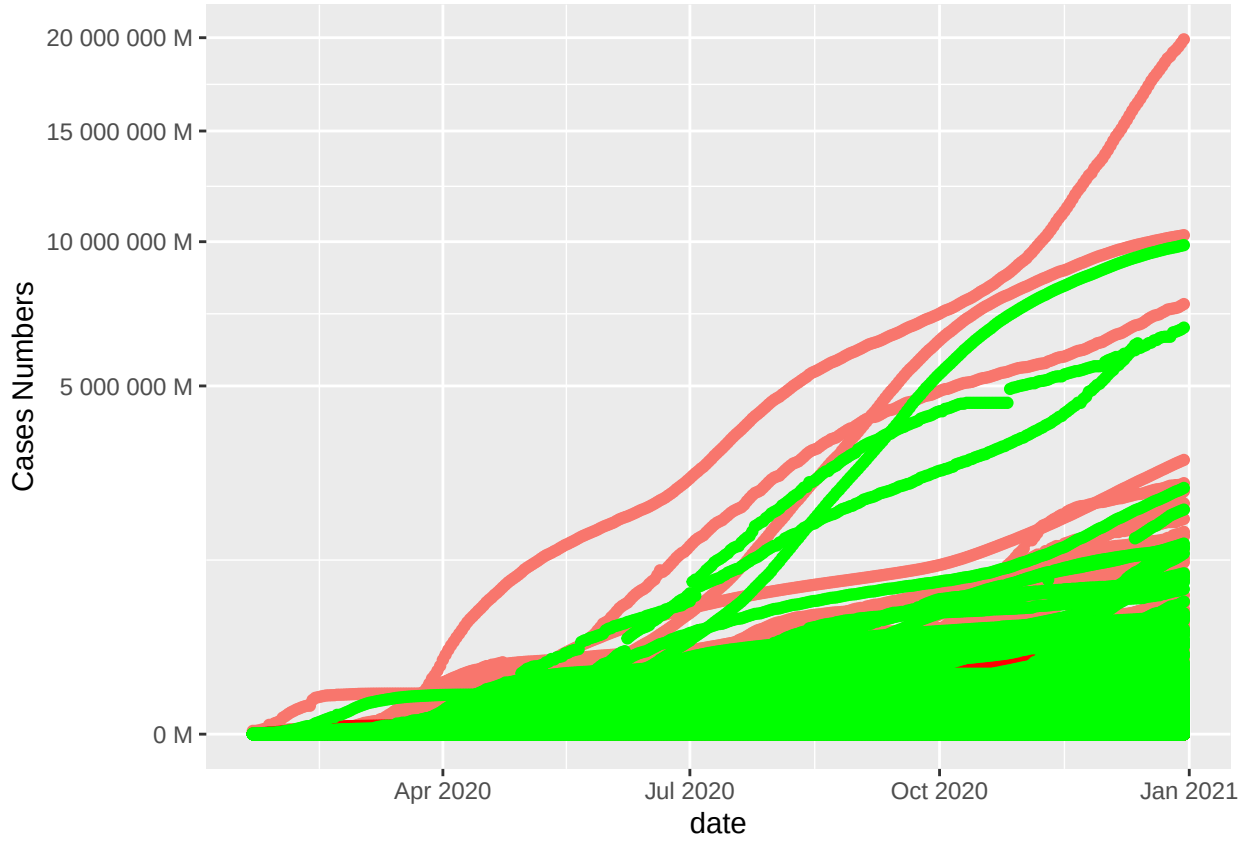
```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_bar()`).
```



We are trying to create a visualization by country, but it is not very useful. Now trying to visualize the data of the first year of COVID

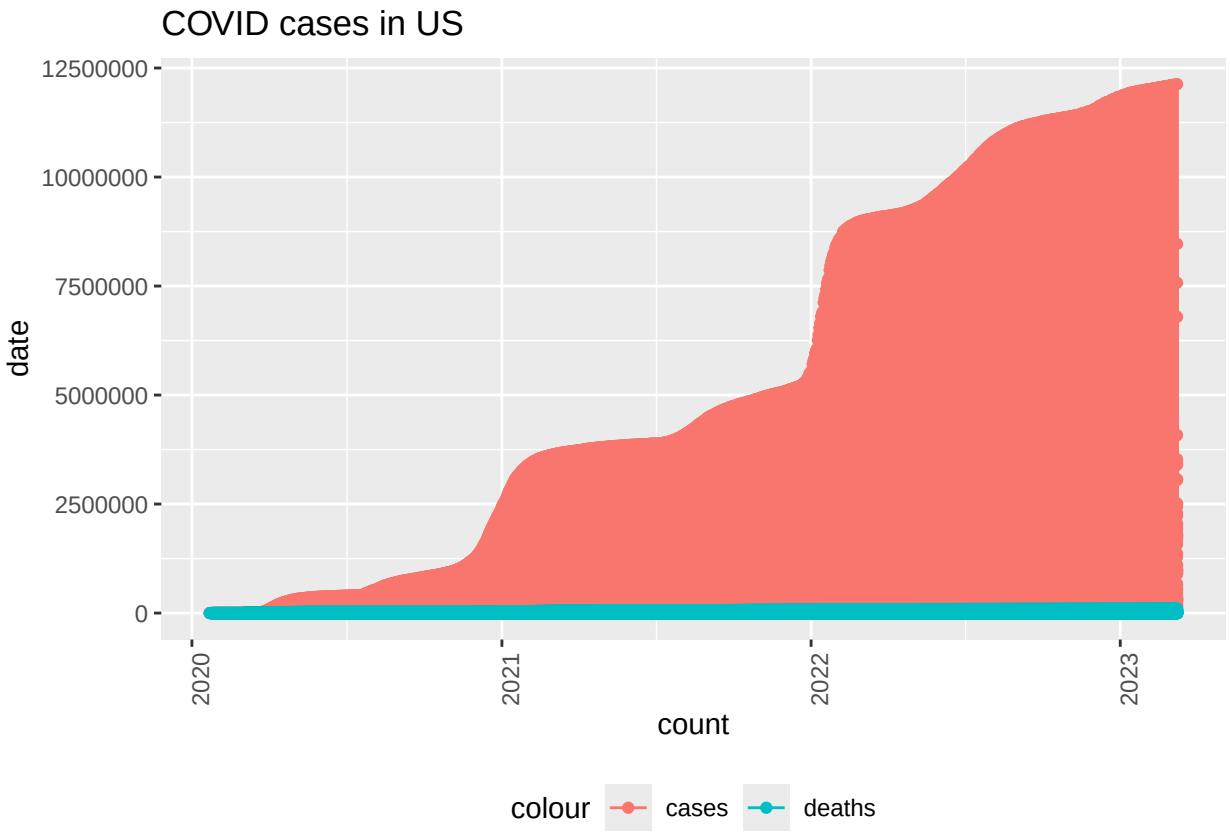
```
global %>%
  filter(date < '2020-12-31') %>%
  ggplot() +
  ylab("Cases Numbers") +
  theme(legend.position = "none") +
  scale_y_sqrt(labels = scales::unit_format(unit = "M")) +
  geom_point(aes(date, cases, colour = 'Blue')) +
  geom_point(aes(date, deaths), colour = 'Red') +
  geom_point(aes(date, recovered), colour = 'Green')
```

```
## Warning: Removed 4025 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Now trying to visualize the cases in the US

```
us %>%
  filter(cases > 0) %>%
  ggplot(aes(x = date, y = cases)) +
  geom_line(aes(color = "cases")) +
  geom_point(aes(color = "cases")) +
  geom_line(aes(y = deaths, color = "deaths")) +
  geom_point(aes(y = deaths, color = "deaths")) +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 90)) +
  labs(title = "COVID cases in US", x="count", y="date")
```



Model

Now creating a model for us

```
us_model <- lm(us$cases ~ us$date, us)

summary(us_model)
```

```
##
## Call:
## lm(formula = us$cases ~ us$date, data = us)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1882029  -599074  -105176   163026 10247621
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.453e+07  2.842e+05  -121.5   <2e-16 ***
## us$date      1.874e+03  1.507e+01   124.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1280000 on 66292 degrees of freedom
```

```
## Multiple R-squared:  0.1892, Adjusted R-squared:  0.1892
## F-statistic: 1.547e+04 on 1 and 66292 DF,  p-value: < 2.2e-16
```

Now creating a model for global data

```
global_model <- lm(global$cases ~ global$deaths, global)

summary(global_model)
```

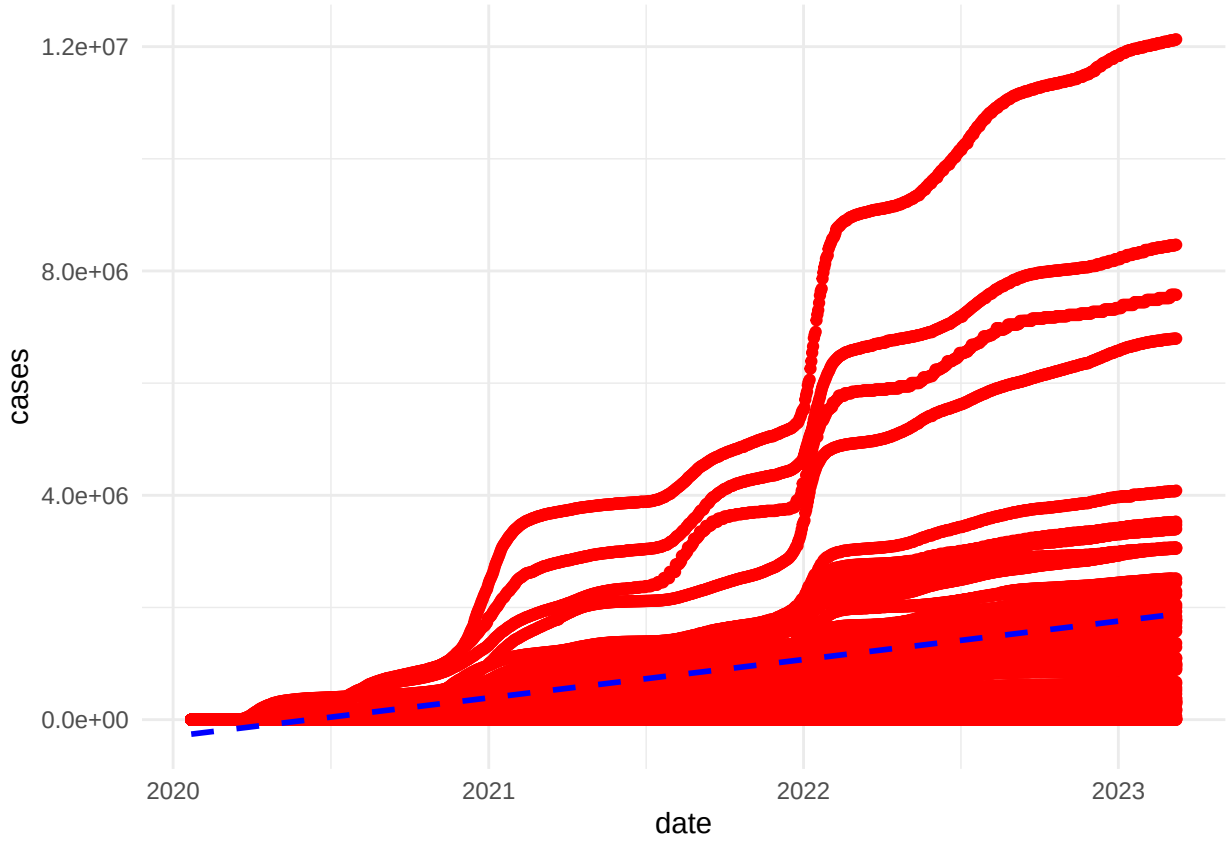
```
##
## Call:
## lm(formula = global$cases ~ global$deaths, data = global)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -19653026    -65327    -58762    -45469  28326024
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.891e+04  4.291e+03   13.73   <2e-16 ***
## global$deaths 6.761e+01  6.335e-02 1067.35   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2322000 on 306825 degrees of freedom
## Multiple R-squared:  0.7878, Adjusted R-squared:  0.7878
## F-statistic: 1.139e+06 on 1 and 306825 DF,  p-value: < 2.2e-16
```

plotting the model

```
us$predict <- predict(us_model)

ggplot(us, aes(x = date, y = cases)) +
  geom_point(color = "red") +
  geom_smooth(method = "lm", se = FALSE, linetype = "dashed", color = "blue") +
  theme_minimal()
```

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


1.2e+07
8.0e+06
4.0e+06
0.0e+00
cases
2020
2021
2022
2023
date