Covid Project

2025-03-22

Overview

Goal is to do a short data exploration of world wide Covid data. The analysis and modeling will focus on total Covid cases and deaths.

```
url_in <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/refs/heads/master/csse_covid_19_da

file_names = c("time_series_covid19_confirmed_global.csv", "time_series_covid19_deaths_global.csv", "ti

urls = str_c(url_in, file_names)
urls
```

```
## [1] "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/refs/heads/master/csse_covid_19_data/
## [2] "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/refs/heads/master/csse_covid_19_data/
## [3] "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/refs/heads/master/csse_covid_19_data/
## [4] "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/refs/heads/master/csse_covid_19_data/
```

```
global_cases = read_csv(urls[1])
global_deaths = read_csv(urls[2])
us_cases = read_csv(urls[3])
us_deaths = read_csv(urls[4])
```

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

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

```
global <- global_cases %>%
  full_join(global_deaths) %>%
  rename(Country_Region = 'Country/Region',
         Province_State = 'Province/State') %>%
  mutate(date = mdy(date))
```

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

```r
summary(global)
```

```
##  Province_State     Country_Region          date                 cases
##  Length:330327      Length:330327      Min.   :2020-01-22   Min.   :        0
##  Class :character   Class :character   1st Qu.:2020-11-02   1st Qu.:      680
##  Mode  :character   Mode  :character   Median :2021-08-15   Median :    14429
##                                        Mean   :2021-08-15   Mean   :   959384
##                                        3rd Qu.:2022-05-28   3rd Qu.:   228517
##                                        Max.   :2023-03-09   Max.   :103802702
##      deaths
##  Min.   :      0
##  1st Qu.:      3
##  Median :    150
##  Mean   :  13380
##  3rd Qu.:   3032
##  Max.   :1123836
```

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

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
##  Min.   :      0
##  1st Qu.:      7
##  Median :    214
##  Mean   :  14405
##  3rd Qu.:   3665
##  Max.   :1123836
```

```r
us_cases <- us_cases %>%
  pivot_longer(cols = -(UID:Combined_Key),
               names_to = "date",
               values_to = "cases") %>%
  select(Admin2:cases) %>%
  mutate(date = mdy(date)) %>%
  select(-c(Lat,Long_))
```

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

```r
us_deaths <- us_deaths %>%
  pivot_longer(cols = -(UID:Combined_Key),
               names_to = "date",
               values_to = "deaths") %>%
  select(Admin2:deaths) %>%
  mutate(date = mdy(date)) %>%
  select(-c(Lat,Long_))

US <-us_cases %>%
  full_join(us_deaths)
```

```
global <- global %>%
  unite("Combined_Key",
        c(Province_State, Country_Region),
        sep = ",",
        na.rm = TRUE,
        remove = FALSE)
```

```
uid_lookup_file = "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/refs/heads/master/csse_covi

uid = read_csv(uid_lookup_file)%>%
  select(-c(Lat, Long_, Combined_Key, code3, iso2, iso3, Admin2))
```

```
## Rows: 4321 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (7): iso2, iso3, FIPS, Admin2, Province_State, Country_Region, Combined_Key
## dbl (5): UID, code3, Lat, Long_, Population
##
## 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.
```

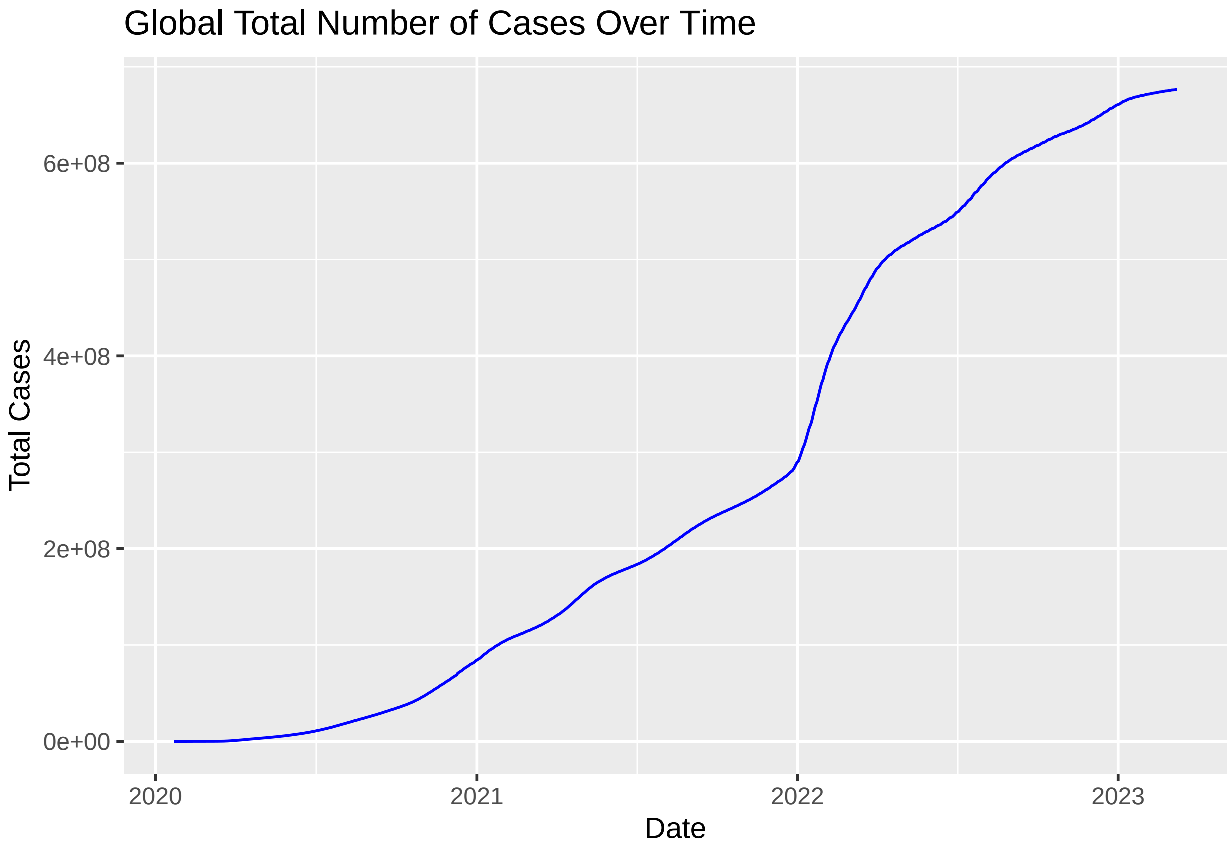
```
global <- global %>%
  left_join(uid, by = c("Province_State", "Country_Region")) %>%
  select(-c(UID, FIPS)) %>%
  select(Province_State, Country_Region, date, cases, deaths, Population, Combined_Key)
```

```
Global_by_cntry <- global %>%
  group_by( Country_Region, date) %>%
  # add up counties and population
  summarize(cases = sum(cases),
            deaths = sum(deaths), Population = sum(Population)) %>%
  mutate(deaths_per_mill = deaths * 1000000 / Population) %>%
  select( Country_Region, date,
         cases, deaths, deaths_per_mill, Population) %>%
  ungroup()
```

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

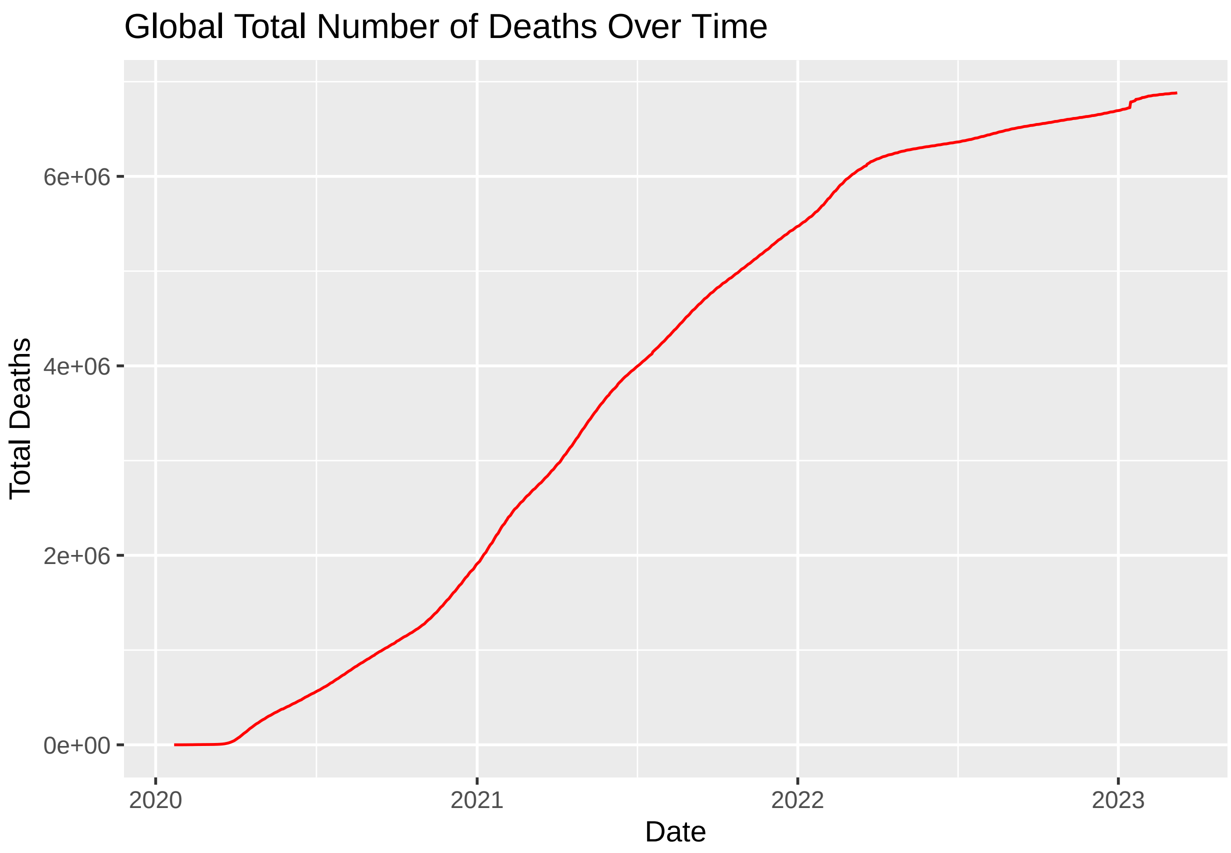
```
Global_total <- Global_by_cntry %>%
  group_by(date) %>%
  summarize(cases = sum(cases))
```

```r
ggplot(Global_total, aes(x = date, y = cases)) +
  geom_line(color = "blue") +
  labs(
    title = "Global Total Number of Cases Over Time",
    x = "Date",
    y = "Total Cases"
  )
```



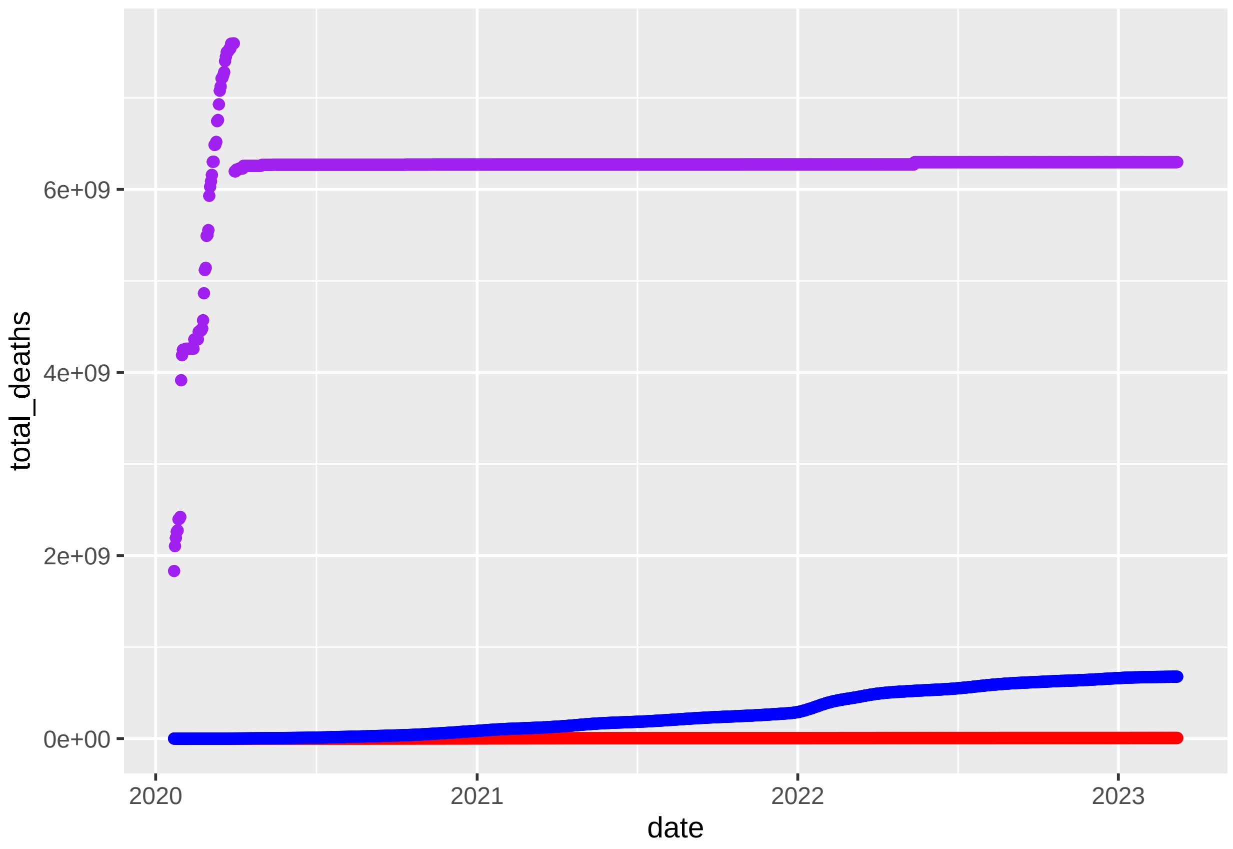
```r
Global_total_deaths <- Global_by_cntry %>%
  group_by(date) %>%
  summarize(deaths = sum(deaths))

ggplot(Global_total_deaths, aes(x = date, y = deaths)) +
  geom_line(color = "red") +
  labs(
    title = "Global Total Number of Deaths Over Time",
    x = "Date",
    y = "Total Deaths"
  )
```



```
global_totals <- Global_by_cntry %>%
     group_by(date) %>%
     summarize(
         total_cases = sum(cases, na.rm = TRUE),
         total_deaths = sum(deaths, na.rm = TRUE),
         total_population = sum(Population, na.rm = TRUE),
         .groups = "drop"
     )

global_totals  %>%
     ggplot()+
     geom_point(aes(x=date, y=total_deaths), color = "red") +
     geom_point(aes(x=date, y=total_cases), color = "blue")+
     geom_point(aes(x=date, y=total_population), color = "purple")
```



Univariate Model using total cases

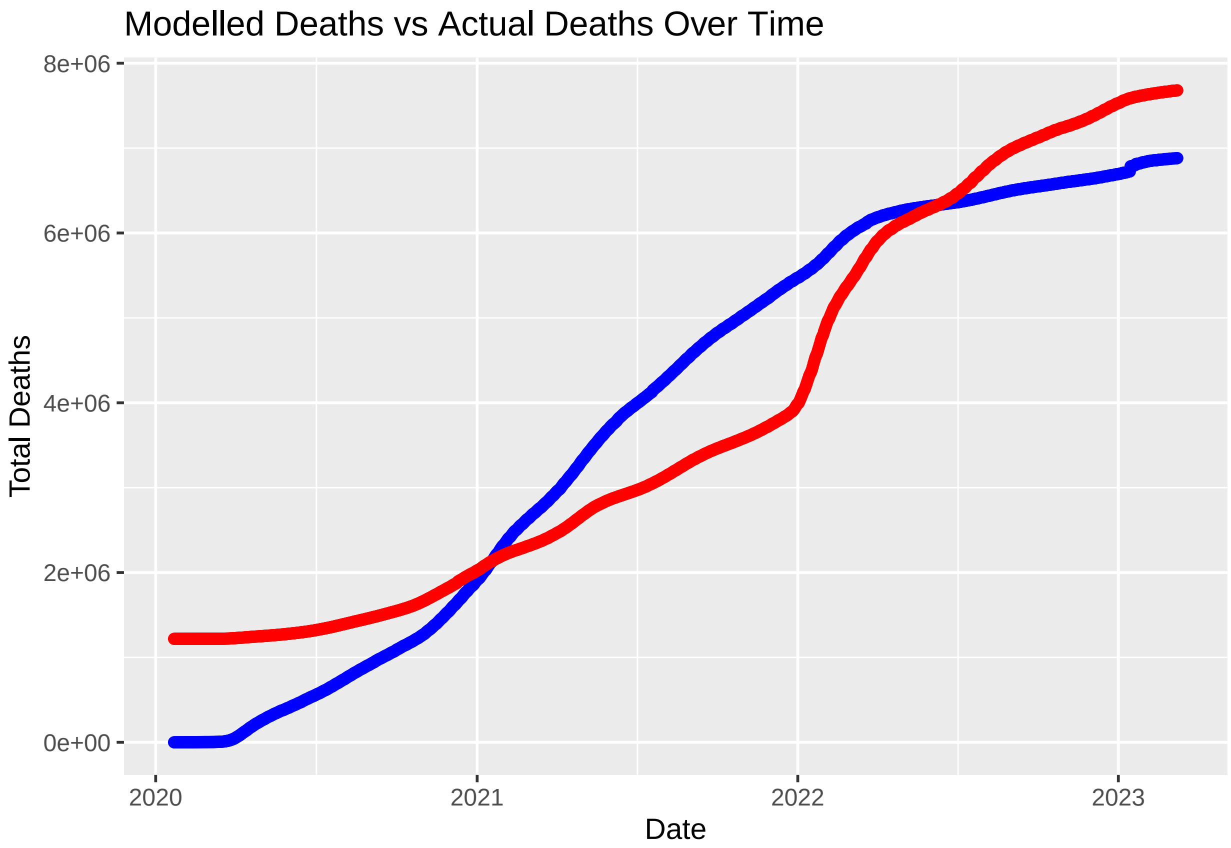
Build a quick linear model using only total cases to predict total deaths.

```
mod <-lm(data=global_totals, total_deaths ~ total_cases )
summary(mod)
```

```
##
## Call:
## lm(formula = total_deaths ~ total_cases, data = global_totals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1218838  -702698  -183706   674979  1541791
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.219e+06  3.769e+04   32.34   <2e-16 ***
## total_cases 9.551e-03  1.028e-04   92.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 833300 on 1141 degrees of freedom
## Multiple R-squared:  0.8832, Adjusted R-squared:  0.8831
## F-statistic:  8624 on 1 and 1141 DF,  p-value: < 2.2e-16
```

```
global_totals <- global_totals %>%
  mutate(predicted_deaths = predict(mod, newdata = global_totals))

global_totals  %>%
     ggplot()+
     geom_point(aes(x=date, y=total_deaths), color = "blue") +
     geom_point(aes(x=date, y=predicted_deaths), color = "red") +
    labs(
    title = "Modelled Deaths vs Actual Deaths Over Time",
    x = "Date",
    y = "Total Deaths"
  )
```



Model starting with July 2020

This enables adding population as in input to the model, since prior the population metric didn't look correct in early 2020.

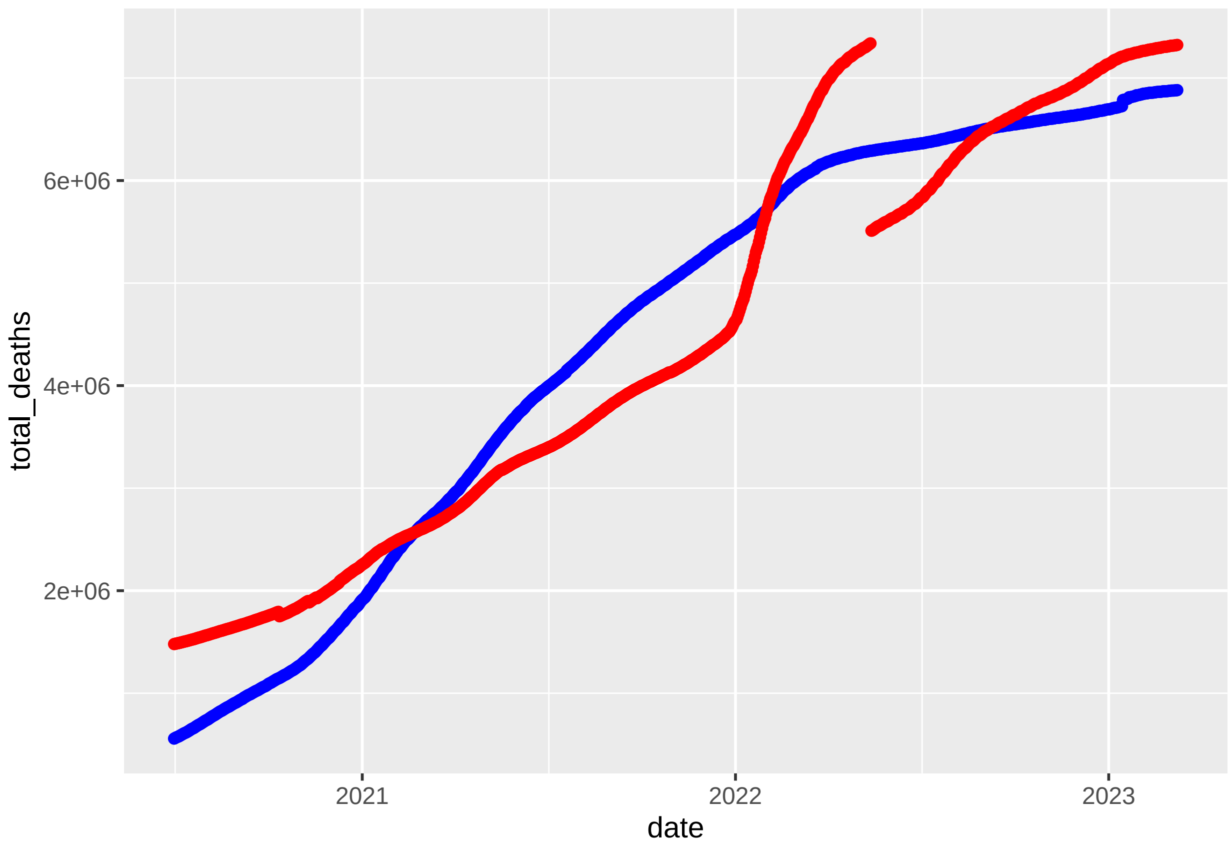
```
 global_totals <- global_totals %>%
     filter(date >= as.Date("2020-07-01"))

mod <-lm(data=global_totals, total_deaths ~ total_cases + total_population )
summary(mod)
```

```
##
## Call:
## lm(formula = total_deaths ~ total_cases + total_population, data = global_totals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1047828  -438772   -64715   572249   932697
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.471e+08  1.853e+07   24.12   <2e-16 ***
## total_cases       1.170e-02  1.559e-04   75.05   <2e-16 ***
## total_population -7.108e-02  2.958e-03  -24.03   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 577900 on 979 degrees of freedom
## Multiple R-squared:  0.9228, Adjusted R-squared:  0.9227
## F-statistic:  5854 on 2 and 979 DF,  p-value: < 2.2e-16
```

```
global_totals <- global_totals %>%
  mutate(predicted_deaths = predict(mod, newdata = global_totals))

global_totals  %>%
     ggplot()+
     geom_point(aes(x=date, y=total_deaths), color = "blue") +
     geom_point(aes(x=date, y=predicted_deaths), color = "red")
```



Conclusion

Based on the models, population and Covid cases are good predictors for Covid deaths. The second model tried to work around the inaccurate early 2020 population data, but is still impacted by a small population difference in mid-2022. With more time, it would be interesting to do a similar analysis by continents. Decided not to, since the country ISO look up data didn't contain a continent variable.