COVID-19 Data Project

Stuart McAllister

2024-07-23

```
# Load libraries
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   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- 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
```

```
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
##
## The following object is masked from 'package:purrr':
##
##     set_names
##
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(lubridate)
```

Project Step 1: Import Relevant Dataset

To begin we will import the COVID19 global time series dataset from the John Hopkins University Github website. This data shows reported cases and deaths for reporting nations from Jan 22, 2020 until March 9, 2023. There is another optional file related to recovered cases which we will not include in this analysis.

This analysis will focus on the global data, specifically subsetting the information presented for the Central American countries. We will investigate the characteristics of both cases and deaths due to COVID-19 in the 7 Central American countries during the reported period of the provided dataset (Jan 22, 2020 - Mar 9, 2023) to see if we can identify any trends or opportunities for further research.

```r
# Import COVID-19 dataset from CSSEGISandData on github
url_in <- "https://github.com/CSSEGISandData/COVID-19/raw/master/csse_covid_19_data/csse_covid_19_time_s

# Identify individual COVID-19 csv files required
file_names <- c("time_series_covid19_confirmed_global.csv",
                "time_series_covid19_deaths_global.csv")
urls <- str_c(url_in, file_names)

# Import country data and clean for population values
uid_lookup_url <- "https://github.com/CSSEGISandData/COVID-19/raw/master/csse_covid_19_data/UID_ISO_FIPS
uid <- read_csv(uid_lookup_url) %>%
        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.
```

```r
# Assign individual data files to variables
global_cases <- read_csv(urls[1])
```

```
## 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.
```

```r
global_deaths <- 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.
```

Project Step 2: Tidy and Transform Data

```r
# Review the first rows of each dataset to understand what actions need
# to be taken to clean the data
head(global_cases)
```

```
## # A tibble: 6 x 1,147
##   `Province/State` `Country/Region`   Lat  Long `1/22/20` `1/23/20` `1/24/20`
##   <chr>            <chr>            <dbl> <dbl>     <dbl>     <dbl>     <dbl>
## 1 <NA>             Afghanistan       33.9 67.7          0         0         0
## 2 <NA>             Albania           41.2 20.2          0         0         0
## 3 <NA>             Algeria           28.0  1.66         0         0         0
## 4 <NA>             Andorra           42.5  1.52         0         0         0
## 5 <NA>             Angola           -11.2 17.9          0         0         0
## 6 <NA>             Antarctica       -71.9 23.3          0         0         0
## # i 1,140 more variables: `1/25/20` <dbl>, `1/26/20` <dbl>, `1/27/20` <dbl>,
## #   `1/28/20` <dbl>, `1/29/20` <dbl>, `1/30/20` <dbl>, `1/31/20` <dbl>,
## #   `2/1/20` <dbl>, `2/2/20` <dbl>, `2/3/20` <dbl>, `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>, ...
```

```
head(global_deaths)
```

```
## # A tibble: 6 x 1,147
##   `Province/State` `Country/Region`   Lat  Long `1/22/20` `1/23/20` `1/24/20`
##   <chr>            <chr>            <dbl> <dbl>     <dbl>     <dbl>     <dbl>
## 1 <NA>             Afghanistan       33.9 67.7          0         0         0
## 2 <NA>             Albania           41.2 20.2          0         0         0
## 3 <NA>             Algeria           28.0  1.66         0         0         0
## 4 <NA>             Andorra           42.5  1.52         0         0         0
## 5 <NA>             Angola           -11.2 17.9          0         0         0
## 6 <NA>             Antarctica       -71.9 23.3          0         0         0
## # i 1,140 more variables: `1/25/20` <dbl>, `1/26/20` <dbl>, `1/27/20` <dbl>,
## #   `1/28/20` <dbl>, `1/29/20` <dbl>, `1/30/20` <dbl>, `1/31/20` <dbl>,
## #   `2/1/20` <dbl>, `2/2/20` <dbl>, `2/3/20` <dbl>, `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>, ...
```

```
head(uid)
```

```
## # A tibble: 6 x 5
##     UID FIPS  Province_State Country_Region Population
##   <dbl> <chr> <chr>          <chr>               <dbl>
## 1     4 <NA>  <NA>           Afghanistan      38928341
## 2     8 <NA>  <NA>           Albania           2877800
## 3    10 <NA>  <NA>           Antarctica             NA
## 4    12 <NA>  <NA>           Algeria          43851043
## 5    20 <NA>  <NA>           Andorra             77265
## 6    24 <NA>  <NA>           Angola           32866268
```

The following steps are used to clean the datasets for our use in this analysis:

1. It appears that in both the global_cases and global_deaths datasets each date is given a column. We would like to convert these through the pivot_longer() function in order to see each date entry for each nation as a single entry.

2. There are also columns for Latitude and Longitude values which we will not need for this investigation
3. We will then join the datasets for both deaths and cases together into one single dataset.
4. Dates are mutated using the lubridate package for ease of use in time-series analysis.
5. A Combined_Key variable is created in this new table which represents a combination of both Province_State and Country_Region. NA values are removed at this point as well.
6. The uid dataset contains country population values and will be joined with the now larger global dataset containing both COVID19 reported deaths and cases.

```
# Clean individual global datasets by using pivot_longer and removing
# the Latitude and Longitude values
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))

# Join the data from global_cases and global_deaths. Rename Country/Region
# and Province/State to maintain a consistent format of column names. Change
# date format using lubridate for consistency and ease of use
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)`
```

```
# Filter for data points with cases greater than 0
global <- global %>% filter(cases > 0)

# Create Combined_Key to match other datasets
global <- global %>%
            unite("Combined_Key", c(Province_State, Country_Region),
            sep = ", ",
            na.rm = TRUE,
            remove = FALSE)

# Join global data with uid dataset to include population values
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)
```

```
# Show first rows of cleaned global data
head(global)
```

```
## # A tibble: 6 x 7
##   Province_State Country_Region date       cases deaths Population Combined_Key
##   <chr>          <chr>          <date>     <dbl>  <dbl>      <dbl> <chr>
## 1 <NA>           Afghanistan    2020-02-24     5      0   38928341 Afghanistan
## 2 <NA>           Afghanistan    2020-02-25     5      0   38928341 Afghanistan
## 3 <NA>           Afghanistan    2020-02-26     5      0   38928341 Afghanistan
## 4 <NA>           Afghanistan    2020-02-27     5      0   38928341 Afghanistan
## 5 <NA>           Afghanistan    2020-02-28     5      0   38928341 Afghanistan
## 6 <NA>           Afghanistan    2020-02-29     5      0   38928341 Afghanistan
```

Project Step 3: Visualization and Analysis

To continue, we will filter the global data to focus only the Central American countries and their reported experiences during the COVID19 pandemic.

```
# Filter the global dataset to create a subset for all of the
# Central American countries
global.cam <- global %>%
                  filter(Country_Region == 'Panama' |
                         Country_Region == 'El Salvador' |
                         Country_Region == 'Costa Rica' |
                         Country_Region == 'Nicaragua' |
                         Country_Region == 'Honduras' |
                         Country_Region == 'Guatemala' |
                         Country_Region == 'Belize' )
head(global.cam)
```

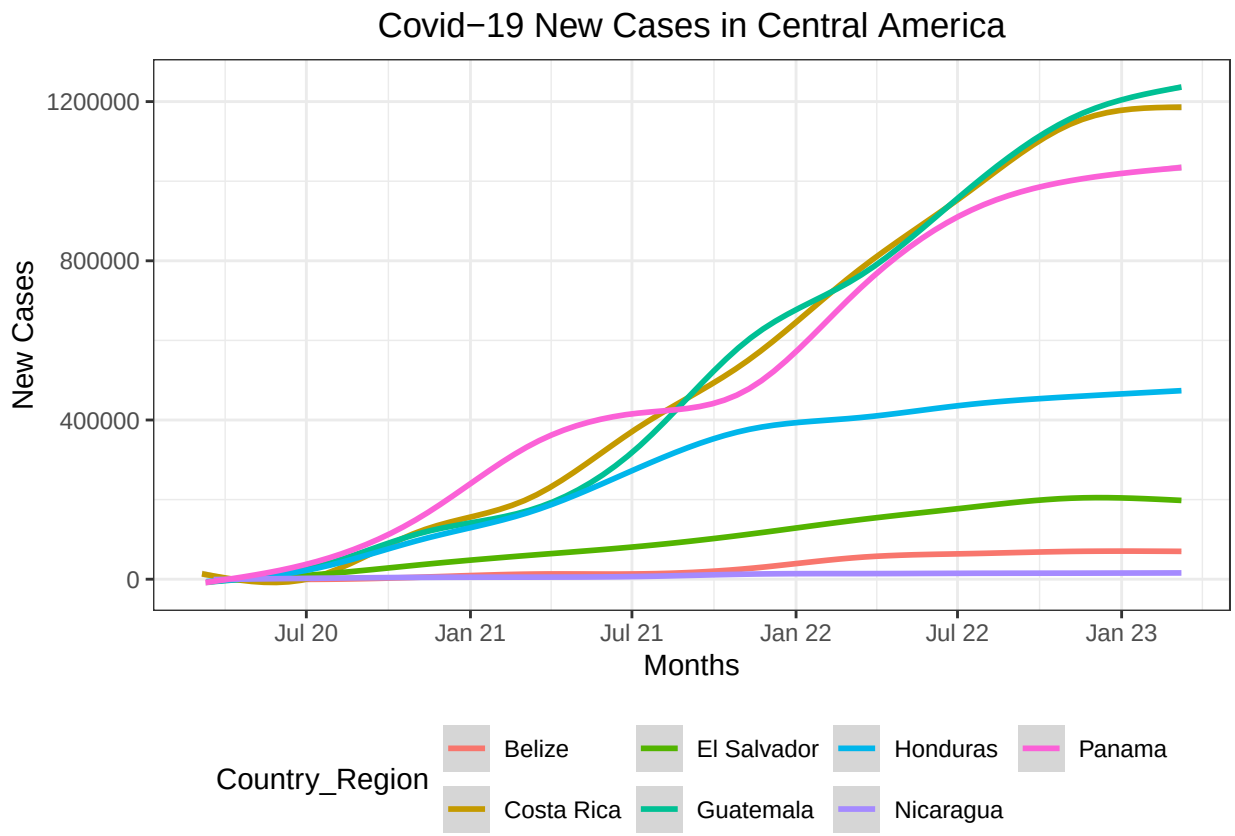
```
## # A tibble: 6 x 7
##   Province_State Country_Region date       cases deaths Population Combined_Key
##   <chr>          <chr>          <date>     <dbl>  <dbl>      <dbl> <chr>
## 1 <NA>           Belize         2020-03-23     1      0     397621 Belize
## 2 <NA>           Belize         2020-03-24     1      0     397621 Belize
## 3 <NA>           Belize         2020-03-25     2      0     397621 Belize
## 4 <NA>           Belize         2020-03-26     2      0     397621 Belize
## 5 <NA>           Belize         2020-03-27     2      0     397621 Belize
## 6 <NA>           Belize         2020-03-28     2      0     397621 Belize
```

Using this filtered dataset we will begin by plotting the number of cases per country in Central America over the time frame of the dataset. From this initial plot it can be seen that there is a significant difference in the number of cases, especially between Guatemala, Costa Rica, and Panama compared to the other four countries. In order to make these values more relatable between nations we will change these to per capita values, dividing by the population values shown in the second plot. Finally, we show the cases per capita values in the third plot for a comparative look at the tendency between countries.

```
# Create an initial plot of cases per country in Central America over
# the time series of the dataset
global.cam %>% ggplot(aes(x = date, y = (cases))) +
                geom_smooth(aes(col = Country_Region)) +
```

```
                labs(y = 'New Cases', x = 'Months') +
                scale_x_date(date_breaks = "6 months", date_labels = "%b %y") +
                theme_bw() +
                theme(legend.position = "bottom",
                      plot.title = element_text(hjust = 0.5)) +
                ggtitle("Covid-19 New Cases in Central America")
```

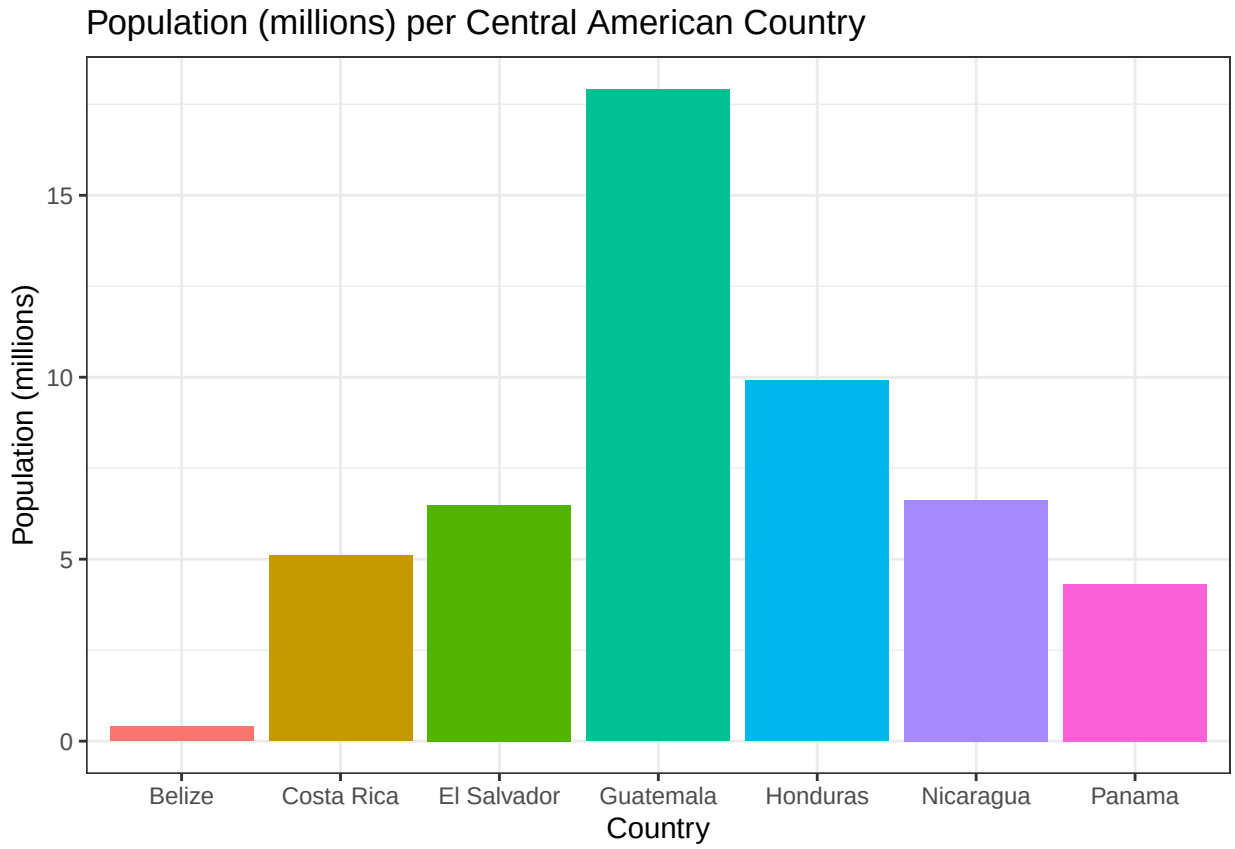
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```




```
# Group dataset by Country_Region and summarise with mean population
pop_sum <- global.cam %>%
                group_by(Country_Region) %>%
                summarise(Mean_Population = (mean(Population)/1000000))
pop_sum
```

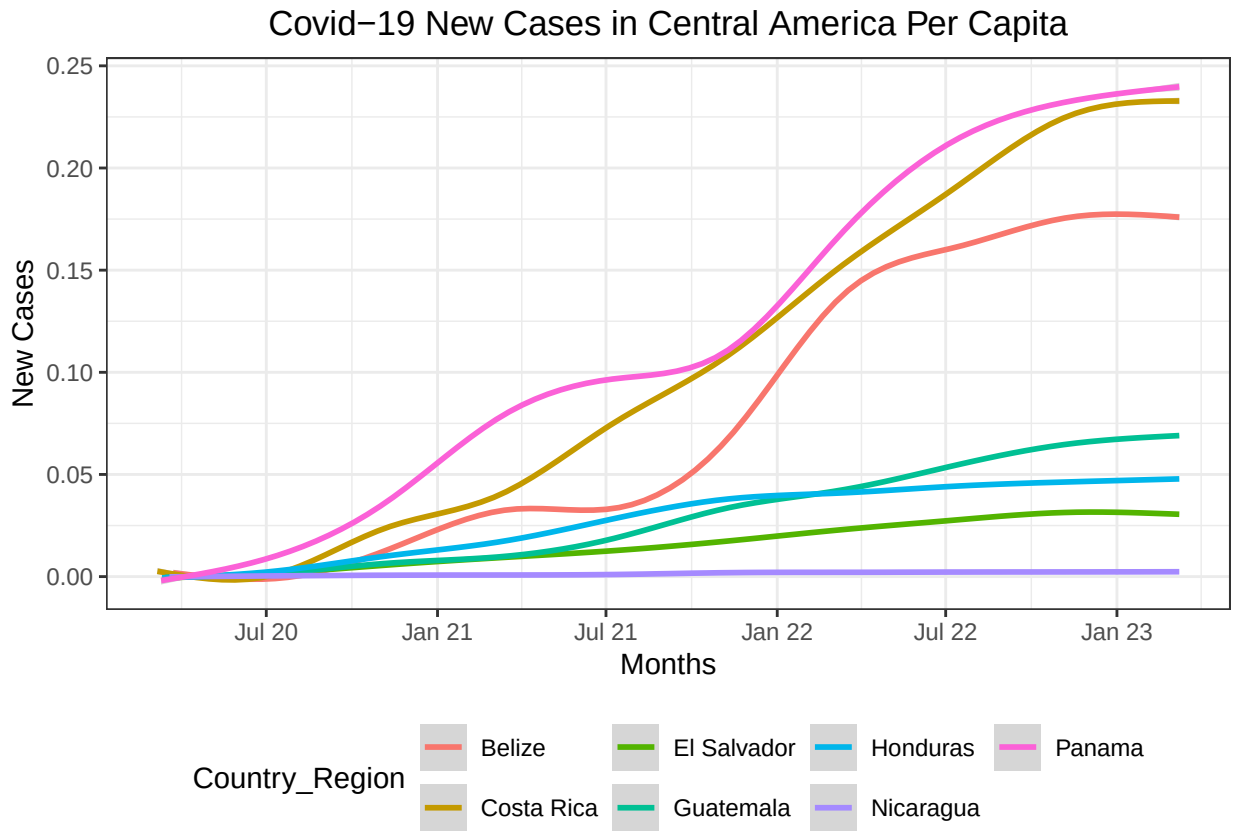
```
## # A tibble: 7 x 2
##   Country_Region Mean_Population
##   <chr>                    <dbl>
## 1 Belize                   0.398
## 2 Costa Rica               5.09
## 3 El Salvador              6.49
## 4 Guatemala               17.9
## 5 Honduras                 9.90
## 6 Nicaragua                6.62
## 7 Panama                   4.31
```

```r
# Bar chart of mean population per Central American country
pop_sum %>% ggplot(aes(x = Country_Region, y = Mean_Population,
              fill = Country_Region)) +
              geom_col(show.legend = FALSE) +
              labs(y = 'Population (millions)', x = "Country") +
              ggtitle("Population (millions) per Central American Country") +
              theme_bw()
```




```r
# Recreate country case data plot, now using cases per person
global.cam %>% ggplot(aes(x = date, y = (cases/Population))) +
                geom_smooth(aes(col = Country_Region)) +
                labs(y = 'New Cases', x = 'Months') +
                scale_x_date(date_breaks = "6 months", date_labels = "%b %y") +
                theme_bw() +
                theme(legend.position = "bottom",
                      plot.title = element_text(hjust = 0.5)) +
                ggtitle("Covid-19 New Cases in Central America Per Capita")
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



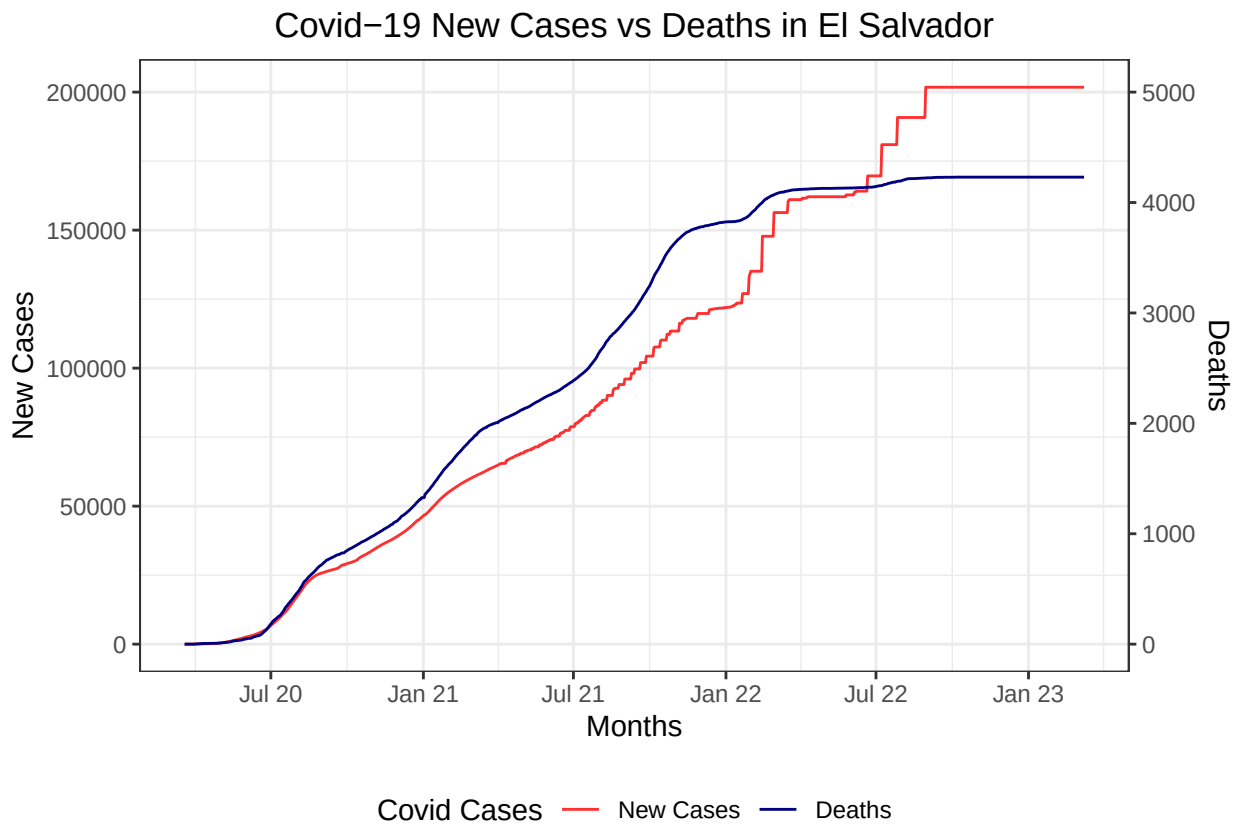
It is interesting that although some of the positions changed in the per capita plot there is still a wide gap between the top three nations (Panama, Costa Rica, and Belize) and the other four nations.

We will now dive into the information for two of the countries, beginning with El Salvador, which is represented in the group with a lower per capita case rate. The follow this with the information for Cost Rica, which is in the higher case per capita group.

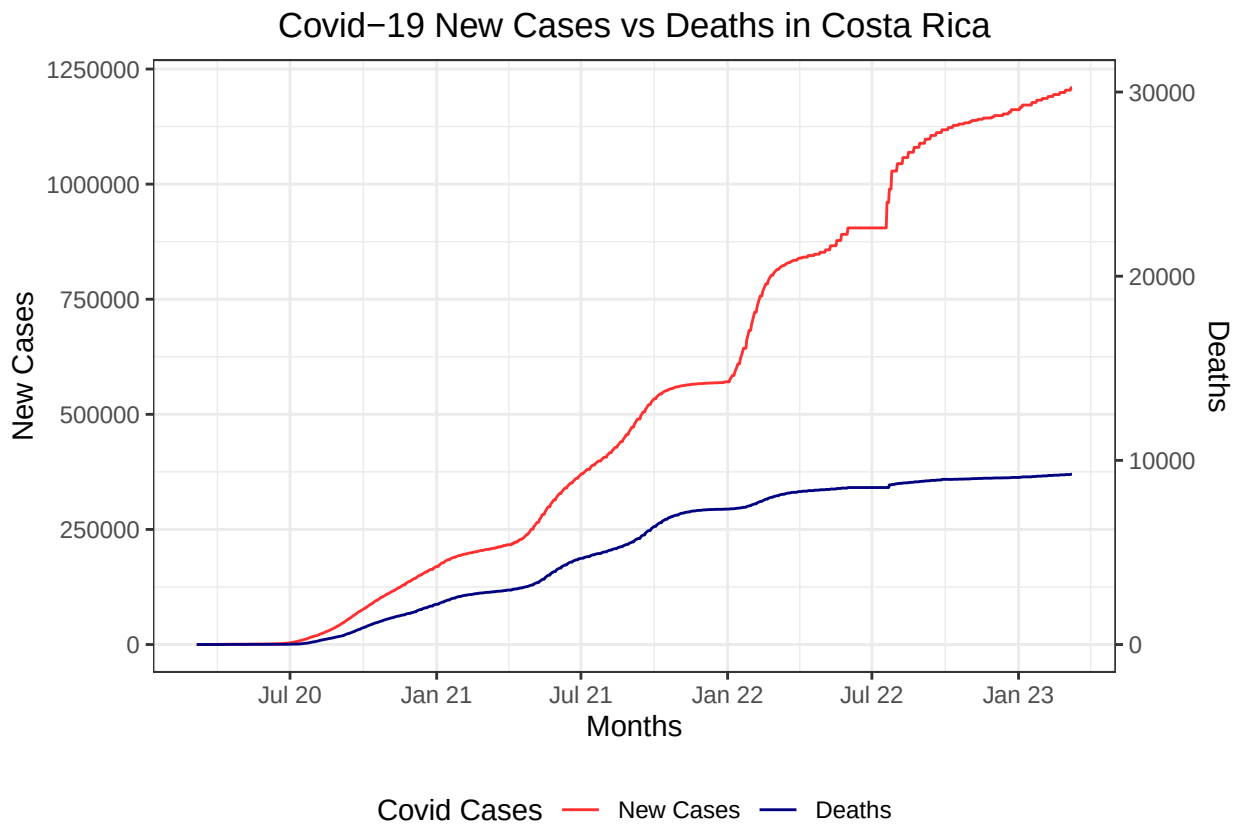
For each we will plot cases and deaths over time to see the trends that developped.

```
# Filter dataset to see only values for El Salvador
global.sv <- global.cam %>%
                  filter(Country_Region == 'El Salvador')

# Plot COVID-19 new cases and deaths over time for El Salvador
global.sv %>% ggplot(aes(x = date)) +
                geom_line(aes(y = cases, color = 'New Cases')) +
                labs(y = 'New Cases', x = 'Months') +
                geom_line(aes(y = deaths/0.025, color = 'Deaths')) +
                scale_x_date(date_breaks = "6 months", date_labels = "%b %y") +
                scale_y_continuous(sec.axis = sec_axis(~.*0.025,
                                                       name = "Deaths")) +
                theme_bw() +
                theme(legend.position = "bottom",
                      plot.title = element_text(hjust = 0.5)) +
                scale_color_manual(name = "Covid Cases",
                                   values = c("firebrick1", "navy"),
                                   breaks = c("New Cases", "Deaths")) +
```

```
                ggtitle("Covid-19 New Cases vs Deaths in El Salvador")
```




```
# Filter dataset to see only values for Costa Rica
global.cr <- global.cam %>%
                 filter(Country_Region == 'Costa Rica')

# Plot COVID-19 new cases and deaths over time for Costa Rica
global.cr %>% ggplot(aes(x = date)) +
                geom_line(aes(y = cases, color = 'New Cases')) +
                labs(y = 'New Cases', x = 'Months') +
                geom_line(aes(y = deaths/0.025, color = 'Deaths')) +
                scale_x_date(date_breaks = "6 months", date_labels = "%b %y") +
                scale_y_continuous(sec.axis = sec_axis(~.*0.025,
                                                       name = "Deaths")) +
                theme_bw() +
                theme(legend.position = "bottom",
                      plot.title = element_text(hjust = 0.5)) +
                scale_color_manual(name = "Covid Cases",
                                   values = c("firebrick1", "navy"),
                                   breaks = c("New Cases", "Deaths")) +
                ggtitle("Covid-19 New Cases vs Deaths in Costa Rica")
```



Finally, we will create linear models for each of the datasets for El Salvador and Costa Rica in order to see how deaths (response variable) are related to the cases reported and time (predictor variables).

```
# Create a linear regression model with deaths as a response variable
# and cases and date as the predictor variables for El Salvador
lm_sv <- lm(deaths ~ cases + date, data = global.sv)
summary(lm_sv)
```

```
##
## Call:
## lm(formula = deaths ~ cases + date, data = global.sv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -796.19 -324.12  -16.39  297.89  775.86
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.104e+04  6.470e+03  -7.890 7.38e-15 ***
## cases        8.652e-03  1.592e-03   5.433 6.83e-08 ***
## date         2.797e+00  3.514e-01   7.959 4.36e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 407.7 on 1083 degrees of freedom
## Multiple R-squared:  0.929,  Adjusted R-squared:  0.9289
```

```
## F-statistic:  7090 on 2 and 1083 DF,  p-value: < 2.2e-16
```

```r
# Create a linear regression model with deaths as a response variable
# and cases and date as the predictor variables for Costa Rica
lm_cr <- lm(deaths ~ cases + date, data = global.cr)
summary(lm_cr)
```

```
##
## Call:
## lm(formula = deaths ~ cases + date, data = global.cr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1673.7  -520.0  -113.6   576.5  1371.9
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.008e+05  8.027e+03 -25.016   <2e-16 ***
## cases       -2.820e-04  3.310e-04  -0.852    0.394
## date         1.092e+01  4.343e-01  25.135   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 732.1 on 1096 degrees of freedom
## Multiple R-squared:  0.9545, Adjusted R-squared:  0.9544
## F-statistic: 1.149e+04 on 2 and 1096 DF,  p-value: < 2.2e-16
```

According to these simple linear regression models, both date and number of reported cases are significant predictors to the model with a response variable of deaths. This can be seen in the very small p-values (significantly less than an $\alpha = 0.05$ which would indicate relevancy to the model). On the other hand, for Cost Rica the predictor cases has a p-value of 0.394 which would indicate that it is not significant to the proposed linear model.

Project Step 4: Conclusion and Bias Identification

Although it is difficult to come to many firm conclusions from this limited analysis, more importantly we have identified some areas for future study. Our data visualizations and models lead me to ask the following questions for further investigation:

1. What are the political factors that have contributed to the different levels of cases and reporting in the 7 Central American countries? These nations share many similar characteristics in terms of climate, geography, race, and heritage, yet there are stark contrasts in the level of per capita cases in each of the nations. This is especially apparent in Nicaragua where the per capita rate of cases remained close to zero for most of the pandemic. Is this due to better policy, or politically motivated reporting, or lack of infrastructure to report, or are there other contributing factors?

2. Why was there a much higher death rate in El Salvador compared to Costa Rica in relation to the number of cases reported? Is this due to poor reporting policies, or the lack of adequate medical response, or are there other factors that have contributed to this difference? Why, in the case of Costa Rica, does it appear that deaths are more relevantly explained by time than by number of cases? Does indicate better or worse management of the COVID19 outbreak in the two countries?

3. We could also look at the remaining nations in Central America and/or compare them to other geograpahical regions of the world to see if there are differences in the number and velocity of COVID19 cases and deaths.

Bias Identification

It is important to recognize that because we all bring our own personalities, experiences, and characteristics with us when we study, we can easily find ourselves with a biased perspective on the information that we are analyzing. In this case I can see a few personal biases that are important to be conscious of when preparing, analyzing or drawing conclusions from the dataset:

1. Probably most important is that we all lived in a certain context during the COVID19 pandemic. In my case I resided in El Salvador during that time period. This has the potential for me to question the motives, whether political, social, or economic for each government's response to the pandemic. Although it is a good thing to not always take data at face value, and question its veracity, there is a fine balance between questioning and reading in one's own experiences into projecting causal relationships.
2. There is also often a general sense that the economic global powers can make better decisions and better use of their resources in times of crisis, but we need to be careful that we don't use that as a filter when working with a region that has a much smaller GDP per capita than many of the Western nations. Even with the region itself, countries like Costa Rica tends to have a higher standard of living than countries like Nicaragua or El Salvador, so was there COVID response better because of these economic capabilities? It is important not to bias our information or our searching with these preconceived notions.
3. A final source of basis is really to ensure that we don't draw conclusions too quickly with the limited information at hand. These issues are complex and intertwined with many other social, political, and economic factors thoughout the world, so we need to be certain that it will be difficult to come to exact conclusions. Rather we may narrow focus and propose avenues of furhter exploration.

```
sessionInfo()
```

```
## R version 4.4.0 (2024-04-24)
## Platform: aarch64-apple-darwin20
## Running under: macOS Sonoma 14.5
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRlapack.dylib;  LAPACK v
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: America/El_Salvador
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] magrittr_2.0.3  lubridate_1.9.3 forcats_1.0.0   stringr_1.5.1
##  [5] dplyr_1.1.4     purrr_1.0.2     readr_2.1.5     tidyr_1.3.1
##  [9] tibble_3.2.1    ggplot2_3.5.1   tidyverse_2.0.0
```

```
##
## loaded via a namespace (and not attached):
##  [1] utf8_1.2.4        generics_0.1.3    stringi_1.8.4     lattice_0.22-6
##  [5] hms_1.1.3         digest_0.6.35     evaluate_0.23     grid_4.4.0
##  [9] timechange_0.3.0  fastmap_1.1.1     Matrix_1.7-0      mgcv_1.9-1
## [13] fansi_1.0.6       scales_1.3.0      cli_3.6.2         rlang_1.1.3
## [17] crayon_1.5.2      bit64_4.0.5       munsell_0.5.1     splines_4.4.0
## [21] withr_3.0.0       yaml_2.3.8        tools_4.4.0       parallel_4.4.0
## [25] tzdb_0.4.0        colorspace_2.1-0  curl_5.2.1        vctrs_0.6.5
## [29] R6_2.5.1          lifecycle_1.0.4   bit_4.0.5         vroom_1.6.5
## [33] pkgconfig_2.0.3   pillar_1.9.0      gtable_0.3.5      glue_1.7.0
## [37] highr_0.10        xfun_0.43         tidyselect_1.2.1  rstudioapi_0.16.0
## [41] knitr_1.46        farver_2.1.1      htmltools_0.5.8.1 nlme_3.1-164
## [45] labeling_0.4.3    rmarkdown_2.26    compiler_4.4.0
```