Covid19 Time Series Data Analysis

Alana Hodge

2023-05-02

1 Import Data

```
library(tidyverse)
library(lubridate)
```

Importing John Hopkins COVID-19 time series Data:

- Global COVID-19 Cases
- Global COVID-19 Deaths
- US-specific COVID-19 Cases
- US-specific COVID-19 Deaths

```
global_cases = read_csv("https://github.com/CSSEGISandData/COVID-19/raw/master/csse_covid_19_data/csse_
```

```
## 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_deaths = read_csv("https://github.com/CSSEGISandData/COVID-19/raw/master/csse_covid_19_data/csse_
```

```
## 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_cases = read_csv("https://github.com/CSSEGISandData/COVID-19/raw/master/csse_covid_19_data/csse_covi
## 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.
```

```
us_deaths = read_csv("https://github.com/CSSEGISandData/COVID-19/raw/master/csse_covid_19_data/csse_covi
```

```
## 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.
```

2. Tidy And Transform The Data

Remove unnecessary columns

We most likely don't need the Latitude and Longitude Coordinates in any of the datasets. We can probably get rid of "iso2", "iso3", "code2" and "FIPS" columns in the US datasets as well.

```
global_cases = global_cases <- subset (global_cases, select = -Long)
global_cases = global_cases <- subset (global_cases, select = -Lat)

global_deaths = global_deaths <- subset (global_deaths, select = -Long)
global_deaths = global_deaths <- subset (global_deaths, select = -Lat)

us_cases = us_cases <- subset (us_cases, select = -iso2)
us_cases = us_cases <- subset (us_cases, select = -iso3)
us_cases = us_cases <- subset (us_cases, select = -code3)
us_cases = us_cases <- subset (us_cases, select = -FIPS)
us_cases = us_cases <- subset (us_cases, select = -Lat)
us_cases = us_cases <- subset (us_cases, select = -Long_)


us_deaths = us_deaths <- subset (us_deaths, select = -iso2)
us_deaths = us_deaths <- subset (us_deaths, select = -iso3)
us_deaths = us_deaths <- subset (us_deaths, select = -code3)
us_deaths = us_deaths <- subset (us_deaths, select = -FIPS)
us_deaths = us_deaths <- subset (us_deaths, select = -Lat)
us_deaths = us_deaths <- subset (us_deaths, select = -Long_)
```

Handle Missing Data

Next, we handle missing data, by removing any entries that have missing values in the remaining columns.

```
us_cases <- na.omit(us_cases)
us_deaths <- na.omit(us_deaths)
global_deaths <- na.omit(global_deaths)
global_cases <- na.omit(global_cases)
```

3. Visualizations and Analysis

Question: How many people died from COVID-19 during the pandemic years in North American countries?

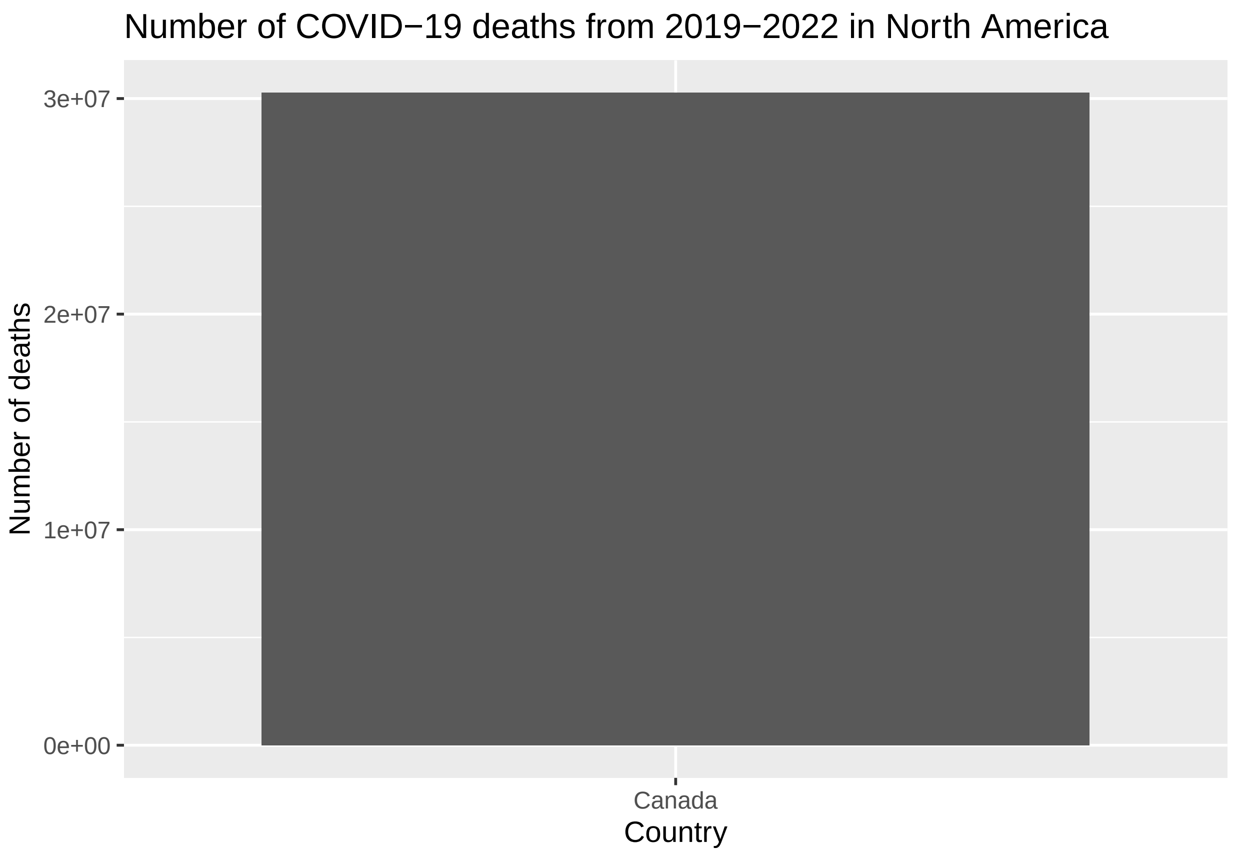
```
q1_data = global_deaths %>% filter(`Country/Region` == "Canada" |
                                     `Country/Region` == "US" |
                                     `Country/Region` == "Mexico")

q1_df1 = q1_data

q1_df1 <- subset (q1_data, select = -`Province/State`)
q1_df1 <- subset (q1_df1, select = -`Country/Region`)
q1_df1$Total_Deaths = rowSums(q1_df1)

q1_data$Total_Deaths = q1_df1$Total_Deaths

q1_plot <- ggplot(q1_data, aes(x = `Country/Region`, y = Total_Deaths)) +
  geom_bar(stat='identity') +
  labs(title = "Number of COVID-19 deaths from 2019-2022 in North America",
       x = "Country",
       y = "Number of deaths")
q1_plot
```



4. Modeling

5. Bias Implications

Regarding biases identified in the data set:

Reporting Bias

Not all countries have tested and measured for Covid-19 related cases in deaths in the same way. Specifically for countries with rural populations and/or poor infrastructure, measuring the actual impact of Covid-19 among the population is not equitably done across all nations.

Moreover, it's possible that not all cases of Covid-19 are reported at all. Especially towards the end of the pandemic, where many people treated their symptoms in isolation, it's inaccurate to assume that all cases of Covid-19 in any given country/region were reported.

Personal Bias

The effect of Covid-19 from my own perspective is drastically different from others. With debates on proper treatment, vaccintaion and policies around wearing masks, it's reasonable to assume that not everyone reading this report or utilizing this data has the same perspective on Covid-19 and the pandemic. This by itself is a form of personal bias that everyone can have towards this topic, including myself.

Session info for reproducibility

```
sessionInfo()
```

```
## R version 4.3.0 (2023-04-21 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 11 x64 (build 22621)
##
## Matrix products: default
##
##
## locale:
## [1] LC_COLLATE=English_Canada.utf8  LC_CTYPE=English_Canada.utf8
## [3] LC_MONETARY=English_Canada.utf8 LC_NUMERIC=C
## [5] LC_TIME=English_Canada.utf8
##
## time zone: America/Toronto
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] lubridate_1.9.2 forcats_1.0.0   stringr_1.5.0   dplyr_1.1.2
##  [5] purrr_1.0.1     readr_2.1.4     tidyr_1.3.0     tibble_3.2.1
##  [9] ggplot2_3.4.2   tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] bit_4.0.5         gtable_0.3.3      highr_0.10        crayon_1.5.2
##  [5] compiler_4.3.0    tidyselect_1.2.0  parallel_4.3.0    scales_1.2.1
##  [9] yaml_2.3.7        fastmap_1.1.1     R6_2.5.1          labeling_0.4.2
## [13] generics_0.1.3    curl_5.0.0        knitr_1.42        munsell_0.5.0
## [17] pillar_1.9.0      tzdb_0.3.0        rlang_1.1.1       utf8_1.2.3
## [21] stringi_1.7.12    xfun_0.39         bit64_4.0.5       timechange_0.2.0
## [25] cli_3.6.1         withr_2.5.0       magrittr_2.0.3    digest_0.6.31
## [29] grid_4.3.0        vroom_1.6.3       rstudioapi_0.14   hms_1.1.3
## [33] lifecycle_1.0.3   vctrs_0.6.2       evaluate_0.20     glue_1.6.2
## [37] farver_2.1.1      fansi_1.0.4       colorspace_2.1-0  rmarkdown_2.21
## [41] tools_4.3.0       pkgconfig_2.0.3   htmltools_0.5.5
```