CovidData

A. Sorri

2024-08-17

Analyzing COVID-19 Cases and Deaths in Canada

Data Source: https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series/ (https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series/)

Description of Data: This data set comes from the Johns Hopkins Coronavirus Resource Center which has collected global COVID-19 data between 2020 and 2023. The data set includes information about the number of recorded COVID-19 infections and about the number of COVID related deaths for each country and state/province.

Libraries Used in Analysis: tidyverse, lubridate, ggplot2.

Step 1 and 2: Import Data and Tidy/Transform

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.4.4     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.0
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflic
ts to become errors
```

```
library(lubridate)
#Import and Read in the Data
url_in <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covi
d_19_data/csse_covid_19_time_series/"
file_names <- c("time_series_covid19_confirmed_global.csv","time_series_covid19_death
s_global.csv", "time_series_covid19_confirmed_US.csv", "time_series_covid19_deaths_U
S.csv")
urls <- str_c(url_in, file_names)
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.
```

```
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.
```

```
US_cases <- read_csv(urls[3])
```

```
## 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(urls[4])
```

```
## 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.
```

```
#Tidy each data file
US_cases <- US_cases %>% pivot_longer(cols= -(UID:Combined_Key), names_to = "date", v
alues_to = "cases") %>% select(Admin2:cases) %>% mutate(date=mdy(date)) %>% select(-
c(Lat,Long_))
US_deaths <- US_deaths %>% pivot_longer(cols=-(UID:Population), names_to = "date", va
lues_to = "deaths") %>% select(Admin2:deaths) %>% mutate(date=mdy(date)) %>% select(-
c(Lat, Long_))
global_cases <- global_cases %>% pivot_longer(cols = -c(`Province/State`, `Country/Re
gion`, 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, Lon
g))
global <- global_cases %>% full_join(global_deaths) %>% rename(Country_Region = `Coun
try/Region`, Province_State = `Province/State`) %>% mutate(date =mdy(date))
```

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

```
global
```

```
## # A tibble: 330,327 × 5
##    Province_State Country_Region date       cases deaths
##    <chr>          <chr>          <date>     <dbl>  <dbl>
##  1 <NA>           Afghanistan    2020-01-22     0      0
##  2 <NA>           Afghanistan    2020-01-23     0      0
##  3 <NA>           Afghanistan    2020-01-24     0      0
##  4 <NA>           Afghanistan    2020-01-25     0      0
##  5 <NA>           Afghanistan    2020-01-26     0      0
##  6 <NA>           Afghanistan    2020-01-27     0      0
##  7 <NA>           Afghanistan    2020-01-28     0      0
##  8 <NA>           Afghanistan    2020-01-29     0      0
##  9 <NA>           Afghanistan    2020-01-30     0      0
## 10 <NA>           Afghanistan    2020-01-31     0      0
## # ℹ 330,317 more rows
```

```
print(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
```

```
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
```

```
global <- global %>% unite("Combined_Key", c(Province_State, Country_Region),sep =
",",na.rm = TRUE, remove = FALSE)
uid_lookup_url <- "https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/c
sse_covid_19_data/UID_ISO_FIPS_LookUp_Table.csv"
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.
```

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

```
## # A tibble: 306,827 × 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
##  7 <NA>           Afghanistan    2020-03-01     5      0   38928341 Afghanistan
##  8 <NA>           Afghanistan    2020-03-02     5      0   38928341 Afghanistan
##  9 <NA>           Afghanistan    2020-03-03     5      0   38928341 Afghanistan
## 10 <NA>           Afghanistan    2020-03-04     5      0   38928341 Afghanistan
## # ℹ 306,817 more rows
```

Step 3 Visualize and Analyze the Data

How do the total number of COVID-19 cases and deaths change from 2020 and 2023? What does the trend look like?

Let's visualize the trajectory of the total number of COVID cases and the total number of deaths from COVID, in Canada, between 2020 and 2023.

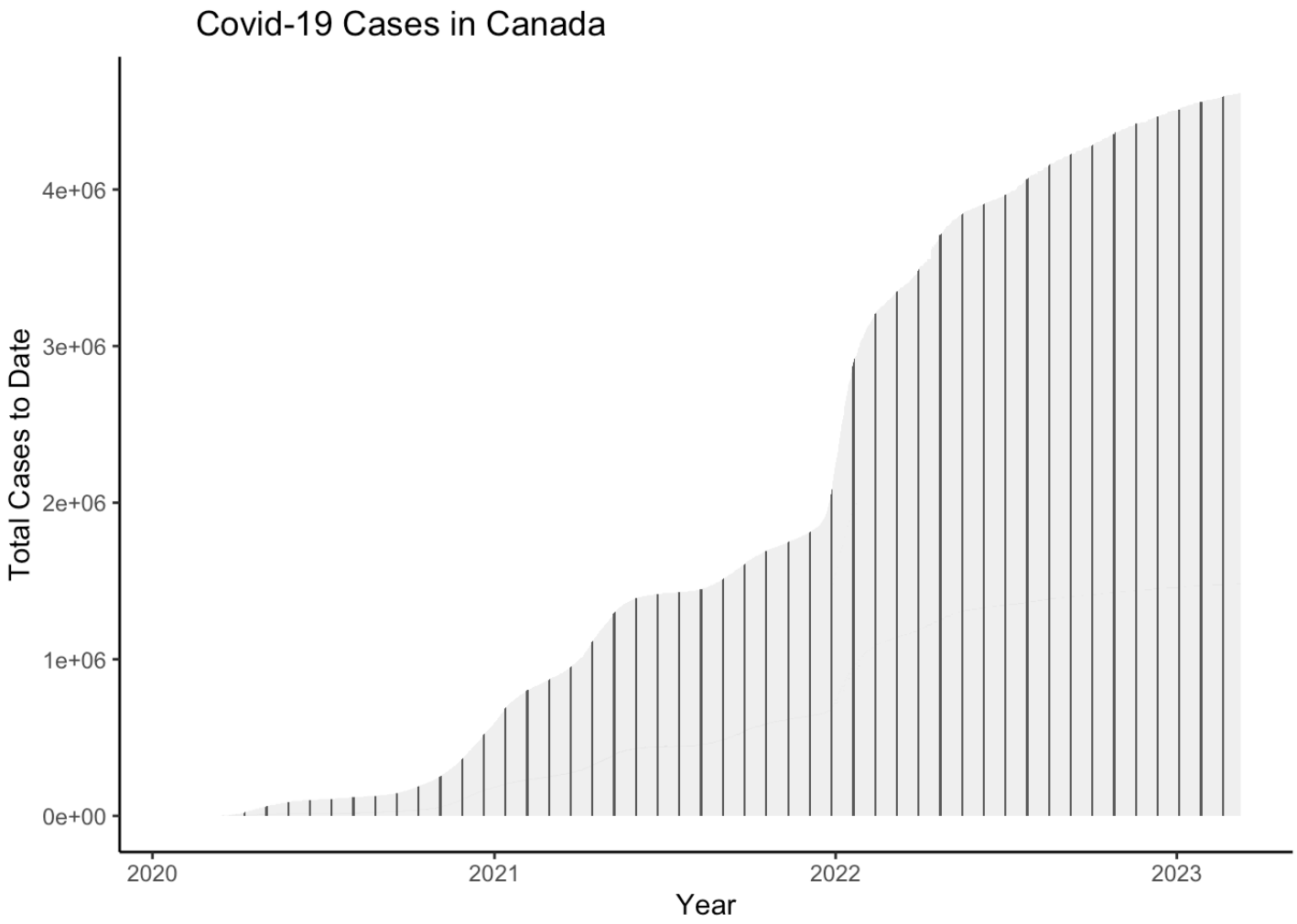
```
#Filter our global data to only the data for Canada.
library(ggplot2)
Cases_Canada <- global %>% group_by(Province_State, Country_Region, date) %>% group_b
y(year = lubridate::year(date)) %>% filter(Country_Region == "Canada")
Cases_Canada
```

```
## # A tibble: 16,010 × 8
## # Groups:   year [4]
##    Province_State Country_Region date       cases deaths Population Combined_Key
##    <chr>          <chr>          <date>     <dbl>  <dbl>      <dbl> <chr>
##  1 Alberta        Canada         2020-03-06     1      0    4442879 Alberta,Can…
##  2 Alberta        Canada         2020-03-07     2      0    4442879 Alberta,Can…
##  3 Alberta        Canada         2020-03-08     4      0    4442879 Alberta,Can…
##  4 Alberta        Canada         2020-03-09     7      0    4442879 Alberta,Can…
##  5 Alberta        Canada         2020-03-10     7      0    4442879 Alberta,Can…
##  6 Alberta        Canada         2020-03-11    19      0    4442879 Alberta,Can…
##  7 Alberta        Canada         2020-03-12    19      0    4442879 Alberta,Can…
##  8 Alberta        Canada         2020-03-13    29      0    4442879 Alberta,Can…
##  9 Alberta        Canada         2020-03-14    29      0    4442879 Alberta,Can…
## 10 Alberta        Canada         2020-03-15    39      0    4442879 Alberta,Can…
## # ℹ 16,000 more rows
## # ℹ 1 more variable: year <dbl>
```

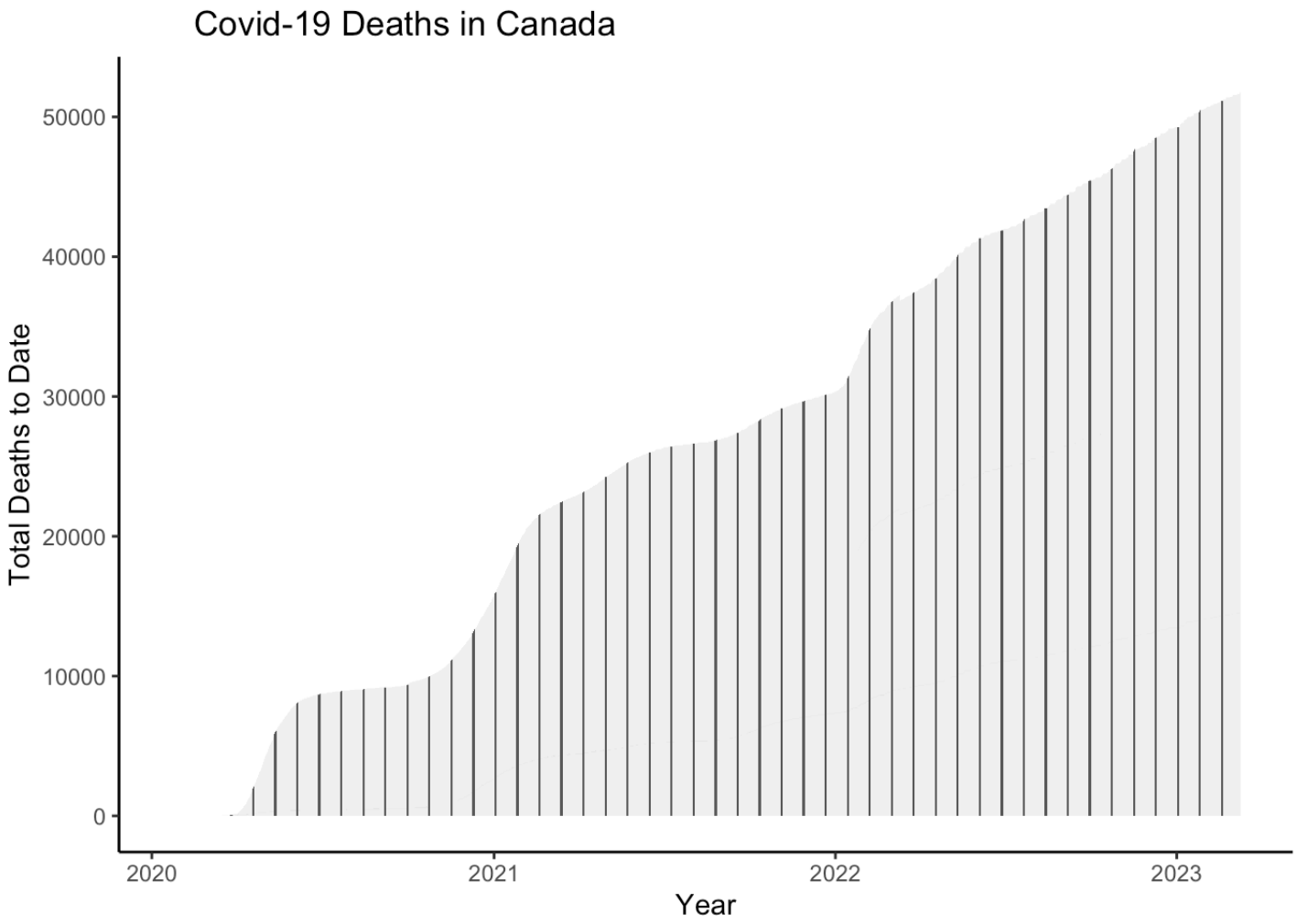
```
summary(Cases_Canada)
```

```
##  Province_State     Country_Region          date                cases
##  Length:16010       Length:16010       Min.   :2020-01-23   Min.   :      1.0
##  Class :character   Class :character   1st Qu.:2020-12-29   1st Qu.:    256.2
##  Mode  :character   Mode  :character   Median :2021-09-22   Median :   6725.5
##                                        Mean   :2021-09-18   Mean   : 141266.9
##                                        3rd Qu.:2022-06-16   3rd Qu.: 127849.8
##                                        Max.   :2023-03-09   Max.   :1601325.0
##
##      deaths         Population       Combined_Key            year
##  Min.   :    0   Min.   :   39403   Length:16010       Min.   :2020
##  1st Qu.:    1   1st Qu.:  164318   Class :character   1st Qu.:2020
##  Median :   61   Median :  992055   Mode  :character   Median :2021
##  Mean   : 1890   Mean   : 3053736                      Mean   :2021
##  3rd Qu.: 1662   3rd Qu.: 4442879                      3rd Qu.:2022
##  Max.   :18160   Max.   :14826276                      Max.   :2023
##                  NA's   :1960
```

```
#A ggplot is created to visualize the trend in covid cases and deaths. The following plots show how the total number of covid-19 cases and the total number of deaths cont inue to steadily rise from 2020 until 2023. The total number of COVID-19 cases in Can ada is well over 4 million by the beginning of 2023. And there had been over 50,000 c ovid related deaths by the beginning of 2023. The exact totals will be counted next.
ggplot(Cases_Canada, aes(x=date, y=cases)) +
  geom_bar(stat="identity", width=0.1) +
  theme_classic() +
  labs(title = "Covid-19 Cases in Canada", x= "Year", y= "Total Cases to Date") +
  theme(plot.title = element_text(hjust = 0.1))
```




```
ggplot(Cases_Canada, aes(x=date, y=deaths)) +
  geom_bar(stat="identity", width=0.1) +
  theme_classic() +
  labs(title = "Covid-19 Deaths in Canada", x= "Year", y= "Total Deaths to Date") +
  theme(plot.title = element_text(hjust = 0.1))
```



What is the max (total) number of covid related deaths and covid cases in Canada? What was the precise date range of the data recored in Canada? Is there a significant relationship or correlation between the total number of covid cases and the total number of covid deaths?

```
#Let's analyze the max (total) number of covid related deaths and covid cases in Cana
da. According to this data set the max number of deaths is 51719 and the max number o
f covid cases is 4617095. These numbers correspond to what we saw in the plots above.
This indicates that there were total of 4,617,095 covid cases and a total of 51,719 c
ovid deaths recorded.
Canada_totals <- Cases_Canada %>% group_by(Country_Region, date) %>% summarize(cases=
sum(cases), deaths=sum(deaths)) %>% select(Country_Region, date, cases, deaths) %>% u
ngroup()
```

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

```
Canada_totals
```

```
## # A tibble: 1,142 × 4
##    Country_Region date       cases deaths
##    <chr>          <date>     <dbl>  <dbl>
##  1 Canada         2020-01-23     2      0
##  2 Canada         2020-01-24     3      0
##  3 Canada         2020-01-25     3      0
##  4 Canada         2020-01-26     3      0
##  5 Canada         2020-01-27     3      0
##  6 Canada         2020-01-28     4      0
##  7 Canada         2020-01-29     4      0
##  8 Canada         2020-01-30     4      0
##  9 Canada         2020-01-31     4      0
## 10 Canada         2020-02-01     4      0
## # ℹ 1,132 more rows
```

```
print(max(Canada_totals$cases))
```

```
## [1] 4617095
```

```
print(max(Canada_totals$deaths))
```

```
## [1] 51719
```

```
#Then we can determine the precise date range for all of the data collected about the COVID cases and deaths in Canada. Data was recorded between 2020-01-23 and 2023-03-09.
print(min(Canada_totals$date))
```

```
## [1] "2020-01-23"
```

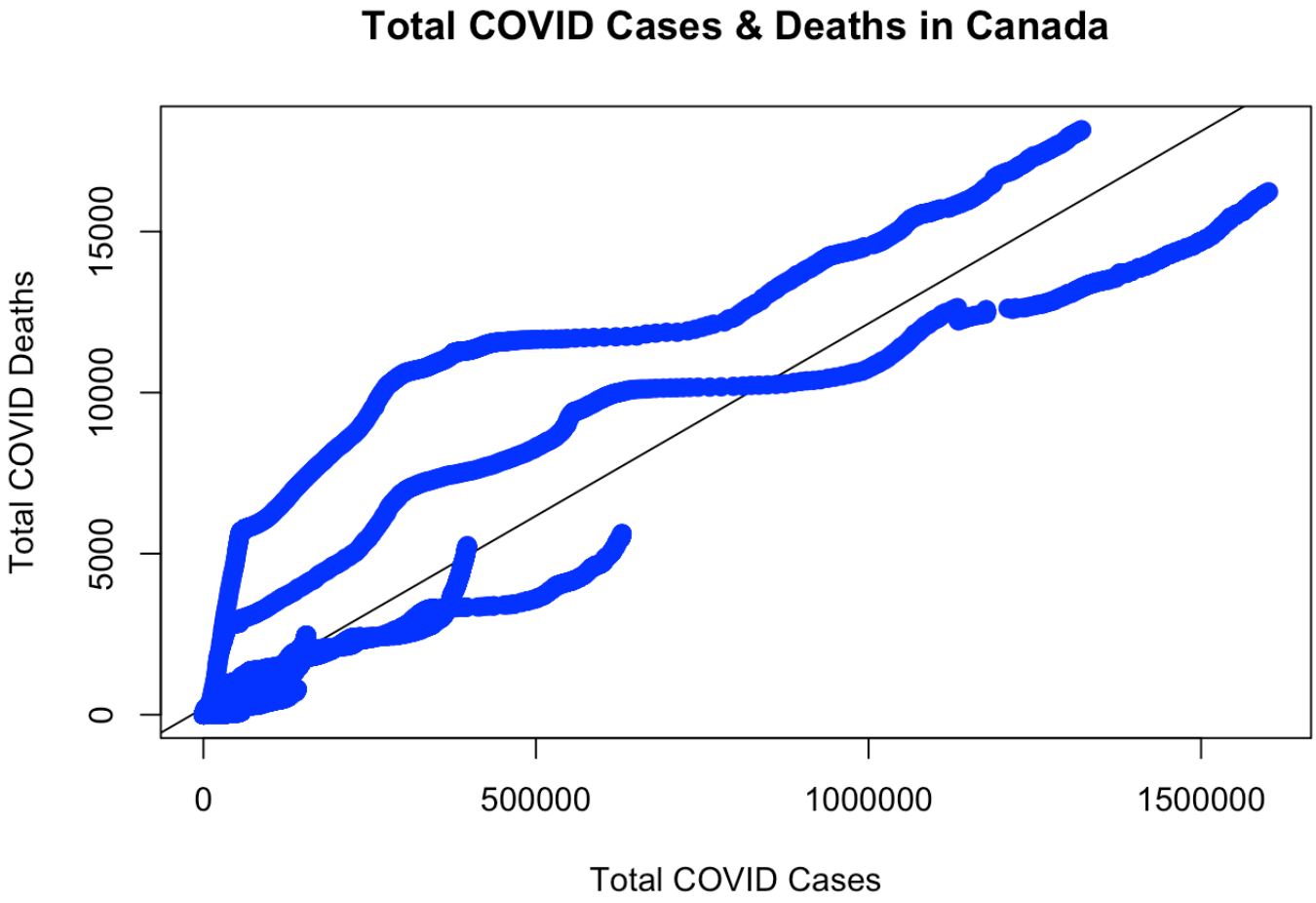
```
print(max(Canada_totals$date))
```

```
## [1] "2023-03-09"
```

```
#Is there a statistically significant relationship/correlation between the number of deaths and covid cases in Canada? Looking at this linear model, assuming a p-value of .05, there is statistically significant correlation between the two variables since t he F-stat is high and the p-value is lower than .05, the p-value is: < 2.2e-16. I hav e also plotted the linear regression model to visualize this correlation.
model <- lm(deaths~cases, data = Cases_Canada)
summary(model)
```

```
##
## Call:
## lm(formula = deaths ~ cases, data = Cases_Canada)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3440.1  -296.6  -203.2  -185.0  6794.2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.031e+02  1.272e+01   15.97   <2e-16 ***
## cases       1.194e-02  3.813e-05  313.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1458 on 16008 degrees of freedom
## Multiple R-squared:  0.8597, Adjusted R-squared:  0.8597
## F-statistic: 9.81e+04 on 1 and 16008 DF,  p-value: < 2.2e-16
```

```
plot(Cases_Canada$cases, Cases_Canada$deaths, col = "blue",
     main = "Total COVID Cases & Deaths in Canada",
     abline(model), cex = 1.3, pch = 16,
     xlab = "Total COVID Cases", ylab = "Total COVID Deaths")
```



Step 4 Conclusion and Bias

Bias: There could be potential bias in the data set depending on how covid-19 case and death data are collected and recorded. If there were multiple sources for the collection of covid data within a certain region it may be possible that a single covid case could be counted more than once unless data collectors were mindful in verifying that each covid case recorded was a unique person/case. This could lead to an overestimation in the counts of covid infections and deaths. It's also likely that not all covid infections were recorded as some people didn't know if they were infected or they did not all cases would officially be reported. In addition, not all potentially infected people would have easy access to covid testing to confirm an infection. Therefore, there is potential bias in covid-19 data due to the potential difficulties or hurdles in accurately testing and reporting infections. These challenges could have lead to under counted covid cases and covid related deaths.

Conclusion: This analysis showed that covid-19 cases and deaths in Canada continued to rise between 2020 and 2023. The analysis indicates that there were total of 4,617,095 million covid cases and a total of 51,719 covid deaths recorded. GGplots were created to visualize the trend in total cases and deaths by year. The data recorded for Canada was collected between 01/23/2020 and 03/09/2023. Then a linear model was created to determine if there was a statistically significant relationship between the total number of covid cases to date and the total number of covid deaths to date. The model had a corresponding p-value of < 2.2e-16, which is much less than .05, therefore there is a statistically significant relationship or correlation between the two variables. A plot for the linear regression model was created to visualize the strength of the correlation.

```
sessionInfo()
```

```
## R version 4.3.2 (2023-10-31)
## Platform: aarch64-apple-darwin20 (64-bit)
## Running under: macOS Sonoma 14.5
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRblas.
0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRlapac
k.dylib;  LAPACK version 3.11.0
##
## 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/New_York
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] lubridate_1.9.3 forcats_1.0.0   stringr_1.5.1   dplyr_1.1.4
##  [5] purrr_1.0.2     readr_2.1.4     tidyr_1.3.0     tibble_3.2.1
##  [9] ggplot2_3.4.4   tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] sass_0.4.8        utf8_1.2.4        generics_0.1.3    stringi_1.8.3
##  [5] hms_1.1.3         digest_0.6.33     magrittr_2.0.3    evaluate_0.23
##  [9] grid_4.3.2        timechange_0.2.0  fastmap_1.1.1     jsonlite_1.8.8
## [13] fansi_1.0.6       scales_1.3.0      jquerylib_0.1.4   cli_3.6.2
## [17] rlang_1.1.2       crayon_1.5.2      bit64_4.0.5       munsell_0.5.0
## [21] withr_2.5.2       cachem_1.0.8      yaml_2.3.8        tools_4.3.2
## [25] parallel_4.3.2    tzdb_0.4.0        colorspace_2.1-0  curl_5.2.0
## [29] vctrs_0.6.5       R6_2.5.1          lifecycle_1.0.4   bit_4.0.5
## [33] vroom_1.6.5       pkgconfig_2.0.3   pillar_1.9.0      bslib_0.6.1
## [37] gtable_0.3.4      glue_1.6.2        highr_0.10        xfun_0.41
## [41] tidyselect_1.2.0  rstudioapi_0.15.0 knitr_1.45        farver_2.1.1
## [45] htmltools_0.5.7   rmarkdown_2.25    labeling_0.4.3    compiler_4.3.2
```