Reproducible Report on COVID19 Data

Christopher Chery

2022-06-15

> Project Instructions: Import, tidy and analyze the COVID19 dataset from the Johns Hopkins github site. This is the same dataset I used in class. Feel free to repeat and reuse what I did if you want to. Be sure your project is reproducible and contains some visualization and analysis that is unique to your project. You may use the data to do any analysis that is of interest to you. You should include at least two visualizations and one model. Be sure to identify any bias possible in the data and in your analysis.

```
library(tidyverse)
library(lubridate)
library(ggplot2)
library(dplyr)
```

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

```
## Rows: 285 Columns: 885
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (2): Province/State, Country/Region
## dbl (883): 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.
## Rows: 285 Columns: 885
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (2): Province/State, Country/Region
## dbl (883): 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.
## Rows: 3342 Columns: 892
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (886): 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.
```

```
## Rows: 3342 Columns: 893
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (6): iso2, iso3, Admin2, Province_State, Country_Region, Combined_Key
## dbl (887): 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.
```

```
## # A tibble: 251,085 x 6
##    `Province/State` `Country/Region`   Lat  Long date    cases
##    <chr>            <chr>            <dbl> <dbl> <chr>   <dbl>
##  1 <NA>             Afghanistan       33.9  67.7 1/22/20     0
##  2 <NA>             Afghanistan       33.9  67.7 1/23/20     0
##  3 <NA>             Afghanistan       33.9  67.7 1/24/20     0
##  4 <NA>             Afghanistan       33.9  67.7 1/25/20     0
##  5 <NA>             Afghanistan       33.9  67.7 1/26/20     0
##  6 <NA>             Afghanistan       33.9  67.7 1/27/20     0
##  7 <NA>             Afghanistan       33.9  67.7 1/28/20     0
##  8 <NA>             Afghanistan       33.9  67.7 1/29/20     0
##  9 <NA>             Afghanistan       33.9  67.7 1/30/20     0
## 10 <NA>             Afghanistan       33.9  67.7 1/31/20     0
## # ... with 251,075 more rows
```

```
## # A tibble: 251,085 x 6
##    `Province/State` `Country/Region`   Lat  Long date    deaths
##    <chr>            <chr>            <dbl> <dbl> <chr>    <dbl>
##  1 <NA>             Afghanistan       33.9  67.7 1/22/20      0
##  2 <NA>             Afghanistan       33.9  67.7 1/23/20      0
##  3 <NA>             Afghanistan       33.9  67.7 1/24/20      0
##  4 <NA>             Afghanistan       33.9  67.7 1/25/20      0
##  5 <NA>             Afghanistan       33.9  67.7 1/26/20      0
##  6 <NA>             Afghanistan       33.9  67.7 1/27/20      0
##  7 <NA>             Afghanistan       33.9  67.7 1/28/20      0
##  8 <NA>             Afghanistan       33.9  67.7 1/29/20      0
##  9 <NA>             Afghanistan       33.9  67.7 1/30/20      0
## 10 <NA>             Afghanistan       33.9  67.7 1/31/20      0
## # ... with 251,075 more rows
```

```
## # A tibble: 2,944,302 x 6
##    Admin2  Province_State Country_Region Combined_Key         date       cases
##    <chr>   <chr>          <chr>          <chr>                <date>     <dbl>
##  1 Autauga Alabama        US             Autauga, Alabama, US 2020-01-22     0
##  2 Autauga Alabama        US             Autauga, Alabama, US 2020-01-23     0
##  3 Autauga Alabama        US             Autauga, Alabama, US 2020-01-24     0
##  4 Autauga Alabama        US             Autauga, Alabama, US 2020-01-25     0
##  5 Autauga Alabama        US             Autauga, Alabama, US 2020-01-26     0
##  6 Autauga Alabama        US             Autauga, Alabama, US 2020-01-27     0
##  7 Autauga Alabama        US             Autauga, Alabama, US 2020-01-28     0
##  8 Autauga Alabama        US             Autauga, Alabama, US 2020-01-29     0
##  9 Autauga Alabama        US             Autauga, Alabama, US 2020-01-30     0
## 10 Autauga Alabama        US             Autauga, Alabama, US 2020-01-31     0
## # ... with 2,944,292 more rows
```

```
## # A tibble: 2,944,302 x 7
##    Admin2  Province_State Country_Region Combined_Key      Population date
##    <chr>   <chr>          <chr>          <chr>                  <dbl> <date>
##  1 Autauga Alabama        US             Autauga, Alabama~      55869 2020-01-22
##  2 Autauga Alabama        US             Autauga, Alabama~      55869 2020-01-23
##  3 Autauga Alabama        US             Autauga, Alabama~      55869 2020-01-24
##  4 Autauga Alabama        US             Autauga, Alabama~      55869 2020-01-25
##  5 Autauga Alabama        US             Autauga, Alabama~      55869 2020-01-26
##  6 Autauga Alabama        US             Autauga, Alabama~      55869 2020-01-27
##  7 Autauga Alabama        US             Autauga, Alabama~      55869 2020-01-28
##  8 Autauga Alabama        US             Autauga, Alabama~      55869 2020-01-29
##  9 Autauga Alabama        US             Autauga, Alabama~      55869 2020-01-30
## 10 Autauga Alabama        US             Autauga, Alabama~      55869 2020-01-31
## # ... with 2,944,292 more rows, and 1 more variable: deaths <dbl>
```

```
## Joining, by = c("Province/State", "Country/Region", "Lat", "Long", "date")
```

```
## # A tibble: 251,085 x 7
##    Province_State Country_Region   Lat  Long date       cases deaths
##    <chr>          <chr>          <dbl> <dbl> <date>     <dbl>  <dbl>
##  1 <NA>           Afghanistan     33.9  67.7 2020-01-22     0      0
##  2 <NA>           Afghanistan     33.9  67.7 2020-01-23     0      0
##  3 <NA>           Afghanistan     33.9  67.7 2020-01-24     0      0
##  4 <NA>           Afghanistan     33.9  67.7 2020-01-25     0      0
##  5 <NA>           Afghanistan     33.9  67.7 2020-01-26     0      0
##  6 <NA>           Afghanistan     33.9  67.7 2020-01-27     0      0
##  7 <NA>           Afghanistan     33.9  67.7 2020-01-28     0      0
##  8 <NA>           Afghanistan     33.9  67.7 2020-01-29     0      0
##  9 <NA>           Afghanistan     33.9  67.7 2020-01-30     0      0
## 10 <NA>           Afghanistan     33.9  67.7 2020-01-31     0      0
## # ... with 251,075 more rows
```

```
## Joining, by = c("Admin2", "Province_State", "Country_Region", "Combined_Key",
## "date")
```

```
## # A tibble: 2,944,302 x 8
##    Admin2 Province_State Country_Region Combined_Key date       cases Population
##    <chr>  <chr>          <chr>          <chr>        <date>     <dbl>      <dbl>
##  1 Autau~ Alabama        US             Autauga, Al~ 2020-01-22     0      55869
##  2 Autau~ Alabama        US             Autauga, Al~ 2020-01-23     0      55869
##  3 Autau~ Alabama        US             Autauga, Al~ 2020-01-24     0      55869
##  4 Autau~ Alabama        US             Autauga, Al~ 2020-01-25     0      55869
##  5 Autau~ Alabama        US             Autauga, Al~ 2020-01-26     0      55869
##  6 Autau~ Alabama        US             Autauga, Al~ 2020-01-27     0      55869
##  7 Autau~ Alabama        US             Autauga, Al~ 2020-01-28     0      55869
##  8 Autau~ Alabama        US             Autauga, Al~ 2020-01-29     0      55869
##  9 Autau~ Alabama        US             Autauga, Al~ 2020-01-30     0      55869
## 10 Autau~ Alabama        US             Autauga, Al~ 2020-01-31     0      55869
## # ... with 2,944,292 more rows, and 1 more variable: deaths <dbl>
```

```
##  Province_State     Country_Region          Lat               Long
##  Length:251085      Length:251085      Min.   :-71.950   Min.   :-178.12
##  Class :character   Class :character   1st Qu.:  4.571   1st Qu.: -23.04
```

```
##  Mode  :character   Mode  :character   Median : 21.694   Median :  20.94
##                                        Mean   : 20.178   Mean   :  22.33
##                                        3rd Qu.: 41.113   3rd Qu.:  88.09
##                                        Max.   : 71.707   Max.   : 178.06
##                                        NA's   :1762      NA's   :1762
##       date                cases              deaths
##  Min.   :2020-01-22   Min.   :       0   Min.   :      0
##  1st Qu.:2020-08-29   1st Qu.:     321   1st Qu.:      2
##  Median :2021-04-06   Median :    7331   Median :     91
##  Mean   :2021-04-06   Mean   :  609161   Mean   :  10641
##  3rd Qu.:2021-11-12   3rd Qu.:  133256   3rd Qu.:   2043
##  Max.   :2022-06-20   Max.   :86297081   Max.   :1013493
##
```

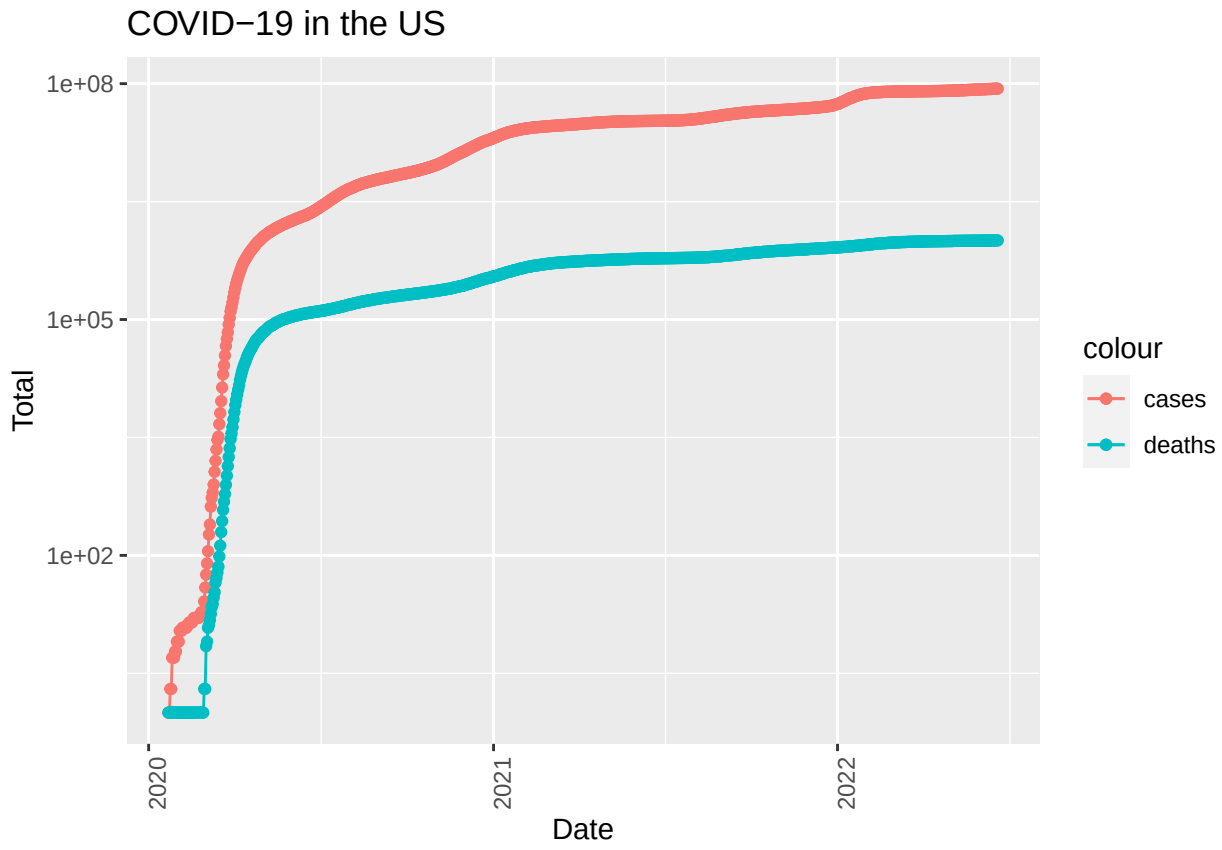
```
##     Admin2          Province_State     Country_Region     Combined_Key
##  Length:2944302     Length:2944302     Length:2944302     Length:2944302
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##       date                cases           Population           deaths
##  Min.   :2020-01-22   Min.   :  -3073   Min.   :       0   Min.   :  -82.0
##  1st Qu.:2020-08-29   1st Qu.:    140   1st Qu.:    9917   1st Qu.:    1.0
##  Median :2021-04-06   Median :   1414   Median :   24892   Median :   24.0
##  Mean   :2021-04-06   Mean   :   9636   Mean   :   99604   Mean   :  147.5
##  3rd Qu.:2021-11-12   3rd Qu.:   5393   3rd Qu.:   64979   3rd Qu.:   90.0
##  Max.   :2022-06-20   Max.   :3069037   Max.   :10039107   Max.   :32261.0
```

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

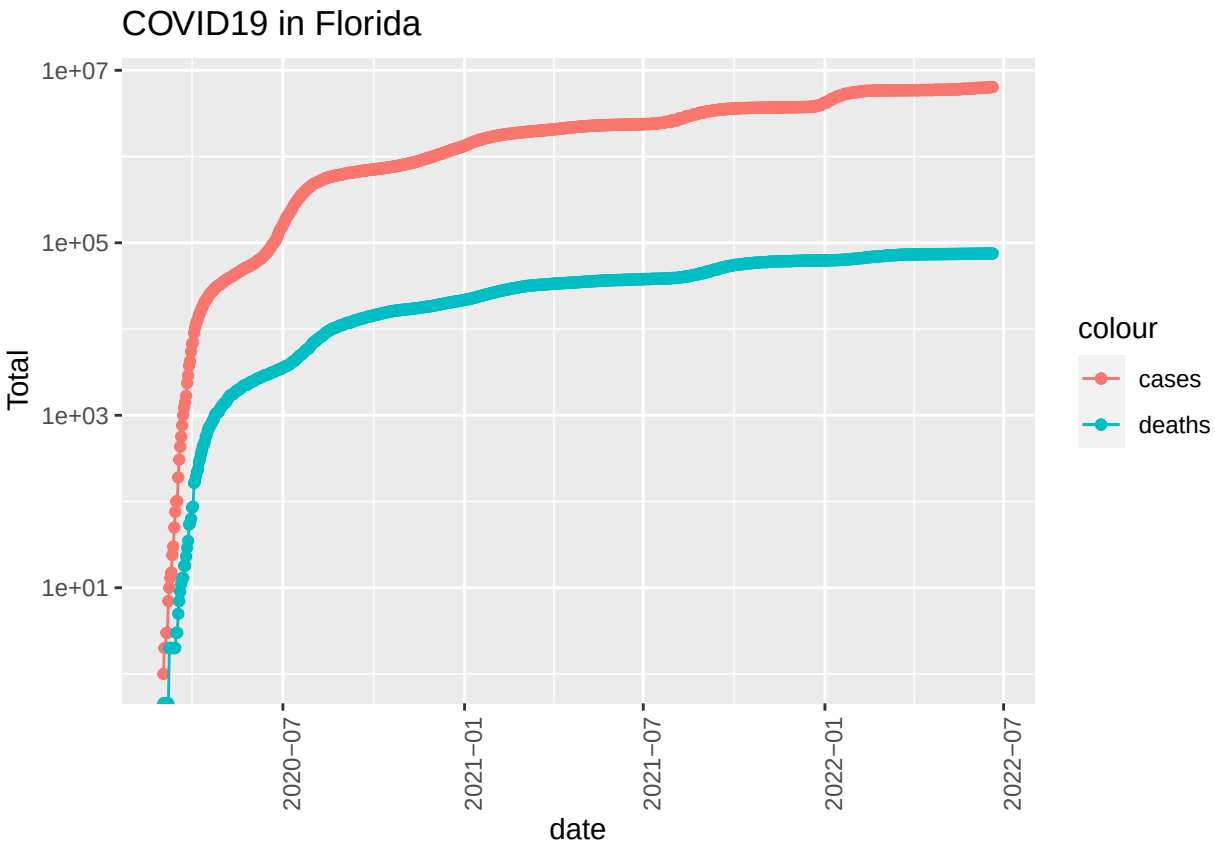
```
## # A tibble: 10 x 6
##    Province_State     deaths  cases population cases_per_thous~ death_per_thous~
##    <chr>               <dbl>  <dbl>      <dbl>            <dbl>            <dbl>
##  1 American Samoa         31 6.34e3      55641             114.            0.557
##  2 Northern Mariana ~     34 1.15e4      55144             208.            0.617
##  3 Hawaii               1474 2.98e5    1415872             210.            1.04
##  4 Vermont               679 1.34e5     623989             215.            1.09
##  5 Virgin Islands        118 2.06e4     107268             192.            1.10
##  6 Puerto Rico          4480 7.42e5    3754939             198.            1.19
##  7 Utah                 4806 9.69e5    3205958             302.            1.50
##  8 Washington          13115 1.63e6    7614893             214.            1.72
##  9 Alaska               1286 2.67e5     740995             360.            1.74
## 10 Maine                2408 2.68e5    1344212             199.            1.79
```

```
## # A tibble: 56 x 6
##    death_per_thousand cases_per_thousand Province_State deaths  cases population
##                 <dbl>              <dbl> <chr>          <dbl>  <dbl>      <dbl>
##  1               4.02               273. Alabama        19697 1.34e6    4903185
##  2               1.74               360. Alaska          1286 2.67e5     740995
```

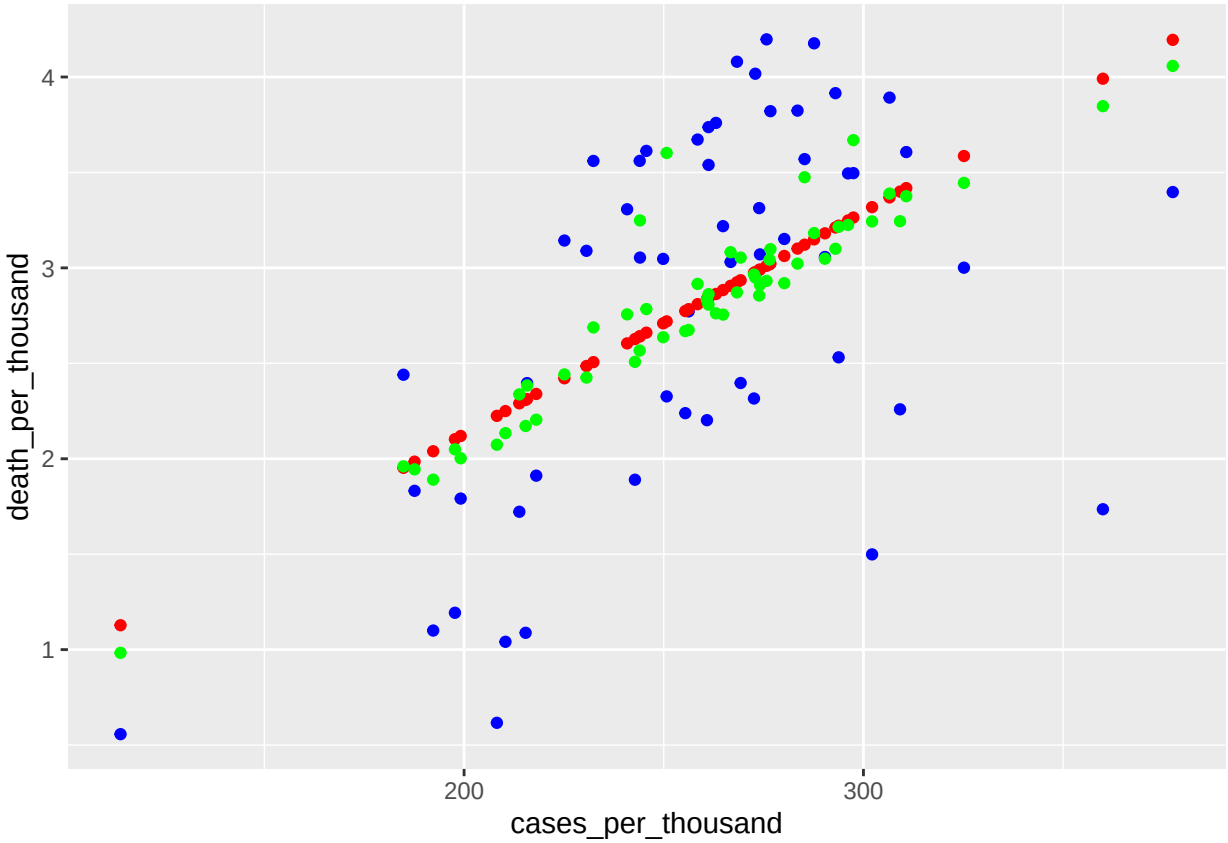
```
##  3              0.557              114. American Samoa     31 6.34e3      55641
##  4              4.18               288. Arizona         30400 2.09e6    7278717
##  5              3.82               283. Arkansas        11540 8.56e5    3017804
##  6              2.33               251. California      91933 9.91e6   39512223
##  7              2.20               261. Colorado        12682 1.50e6    5758736
##  8              3.09               231. Connecticut     11015 8.22e5    3565287
##  9              3.06               290. Delaware         2977 2.83e5     973764
## 10              1.91               218. District of C~   1349 1.54e5     705749
## # ... with 46 more rows
```




```
## Warning: Transformation introduced infinite values in continuous y-axis
## Transformation introduced infinite values in continuous y-axis
```




```
## # A tibble: 56 x 9
##    Province_State     deaths  cases population cases_per_thous~ death_per_thous~
##    <chr>               <dbl>  <dbl>      <dbl>            <dbl>            <dbl>
##  1 Alabama             19697 1.34e6    4903185             273.             4.02
##  2 Alaska               1286 2.67e5     740995             360.             1.74
##  3 American Samoa         31 6.34e3      55641             114.             0.557
##  4 Arizona             30400 2.09e6    7278717             288.             4.18
##  5 Arkansas            11540 8.56e5    3017804             283.             3.82
##  6 California          91933 9.91e6   39512223             251.             2.33
##  7 Colorado            12682 1.50e6    5758736             261.             2.20
##  8 Connecticut         11015 8.22e5    3565287             231.             3.09
##  9 Delaware             2977 2.83e5     973764             290.             3.06
## 10 District of Colum~   1349 1.54e5     705749             218.             1.91
## # ... with 46 more rows, and 3 more variables: pred <dbl>, pred2 <dbl>,
## #   std_ratio <dbl>
```



Conclusion and Bias Identification

Covid 19 had a major impact on the world since 2020. A lot of panic ensued on how contagious the virus was and how the population of an area can be affected by the rising case count. With many different countries around the world implementing covid restrictions to combat the spread of the virus through lockdowns or mandatory quarantines there were signs that these actions slowed down the new case count. Remarkable in less than a year a vaccine was created that has ~ 90% efficacy against the virus and that became the start of life after the pandemic. The potential for bias in this data collection would be how states report there covid cases and deaths; many states were highlighted by the news on how they under reported cases/deaths to favor the next US election cycle of 2020. Another bias comes from when people are taking a covid test, where they having symptoms before or could they be asymptomatic? There was also a testing shortage early on the pandemic and around big holiday travel weekends which led to a lot of unknown positive test cases going unrecorded. There is also bias in my analysis of how I chose to focus on the US covid data. I felt more attached to the US covid data as that is my country of origin and had firsthand experiences with the testing shortages, rising case counts from my home state of Florida and going through the vaccine process.

R Session Info

```
## R version 4.2.0 (2022-04-22 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19043)
##
```

```
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.utf8
## [2] LC_CTYPE=English_United States.utf8
## [3] LC_MONETARY=English_United States.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.utf8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] lubridate_1.8.0 forcats_0.5.1   stringr_1.4.0   dplyr_1.0.9
##  [5] purrr_0.3.4     readr_2.1.2     tidyr_1.2.0     tibble_3.1.7
##  [9] ggplot2_3.3.6   tidyverse_1.3.1
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.1.2 xfun_0.31        haven_2.5.0      colorspace_2.0-3
##  [5] vctrs_0.4.1      generics_0.1.2   htmltools_0.5.2  yaml_2.3.5
##  [9] utf8_1.2.2       rlang_1.0.2      pillar_1.7.0     glue_1.6.2
## [13] withr_2.5.0      DBI_1.1.2        bit64_4.0.5      dbplyr_2.1.1
## [17] modelr_0.1.8     readxl_1.4.0     lifecycle_1.0.1  munsell_0.5.0
## [21] gtable_0.3.0     cellranger_1.1.0 rvest_1.0.2      evaluate_0.15
## [25] labeling_0.4.2   knitr_1.39       tzdb_0.3.0       fastmap_1.1.0
## [29] curl_4.3.2       parallel_4.2.0   fansi_1.0.3      highr_0.9
## [33] broom_0.8.0      backports_1.4.1  scales_1.2.0     vroom_1.5.7
## [37] jsonlite_1.8.0   farver_2.1.0     bit_4.0.4        fs_1.5.2
## [41] hms_1.1.1        digest_0.6.29    stringi_1.7.6    grid_4.2.0
## [45] cli_3.3.0        tools_4.2.0      magrittr_2.0.3   crayon_1.5.1
## [49] pkgconfig_2.0.3  ellipsis_0.3.2   xml2_1.3.3       reprex_2.0.1
## [53] assertthat_0.2.1 rmarkdown_2.14   httr_1.4.3       rstudioapi_0.13
## [57] R6_2.5.1         compiler_4.2.0
```