Covid 19 Statistical Analysis

H. Justino

2024-12-07

I want to understand if the deaths and cases of Covid 19 are consistent across States

Note: I will provide a link to the github repo because I needed an absolute path to the datasets. Results will be reproducible.

Github Link: https://github.com/hermanjustino/Msc.-DS-Assignment-Repo/tree/main/Covid_19_Data_Analysis

Data is coming from Johns Hopkins University

Tidy global cases

These are US cases. It will include where the case was recorded, whether or not it was fatal.

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

tidy global deaths

First few rows of global cases. Country, status

```
## [1] "X1.22.20" "X1.23.20" "X1.24.20" "X1.25.20" "X1.26.20" "X1.27.20"
```

Join Cases with deaths

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

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

Tidy Us Cases

Tidy Us Deaths

Join us tables

```
## Joining with `by = join_by(Admin2, Province_State, Country_Region,
## Combined_Key, date)`
```

Filter us so only days with cases are visible

```
us <- us %>% filter(cases > 0)

us
```

```
## # A tibble: 3,474,292 x 8
##    Admin2 Province_State Country_Region Combined_Key date       cases Population
##    <chr>  <chr>          <chr>          <chr>        <date>     <int>      <int>
##  1 Autau~ Alabama        US             Autauga, Al~ 2020-03-24     1      55869
##  2 Autau~ Alabama        US             Autauga, Al~ 2020-03-25     5      55869
##  3 Autau~ Alabama        US             Autauga, Al~ 2020-03-26     6      55869
##  4 Autau~ Alabama        US             Autauga, Al~ 2020-03-27     6      55869
##  5 Autau~ Alabama        US             Autauga, Al~ 2020-03-28     6      55869
##  6 Autau~ Alabama        US             Autauga, Al~ 2020-03-29     6      55869
##  7 Autau~ Alabama        US             Autauga, Al~ 2020-03-30     8      55869
##  8 Autau~ Alabama        US             Autauga, Al~ 2020-03-31     8      55869
##  9 Autau~ Alabama        US             Autauga, Al~ 2020-04-01    10      55869
## 10 Autau~ Alabama        US             Autauga, Al~ 2020-04-02    12      55869
## # i 3,474,282 more rows
## # i 1 more variable: deaths <int>
```

Combine global

Filter global so only days with cases are visible

Format UID

Join global and uid

Summary

```
##     Admin2          Province_State     Country_Region     Combined_Key
##  Length:3474292     Length:3474292     Length:3474292     Length:3474292
##  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.   :      1   Min.   :       0   Min.   :    0.0
##  1st Qu.:2020-12-27   1st Qu.:    687   1st Qu.:   10953   1st Qu.:   10.0
##  Median :2021-09-20   Median :   2849   Median :   26248   Median :   47.0
##  Mean   :2021-09-19   Mean   :  15489   Mean   :  104502   Mean   :  205.1
##  3rd Qu.:2022-06-15   3rd Qu.:   9345   3rd Qu.:   68098   3rd Qu.:  137.0
##  Max.   :2023-03-09   Max.   :3710586   Max.   :10039107   Max.   :35545.0

##  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          Population        Combined_Key
##  Min.   :      0   Min.   :6.700e+01   Length:306827
##  1st Qu.:      7   1st Qu.:7.866e+05   Class :character
##  Median :    214   Median :6.948e+06   Mode  :character
##  Mean   :  14405   Mean   :2.890e+07
##  3rd Qu.:   3665   3rd Qu.:2.914e+07
##  Max.   :1123836   Max.   :1.380e+09
##                    NA's   :6729
```

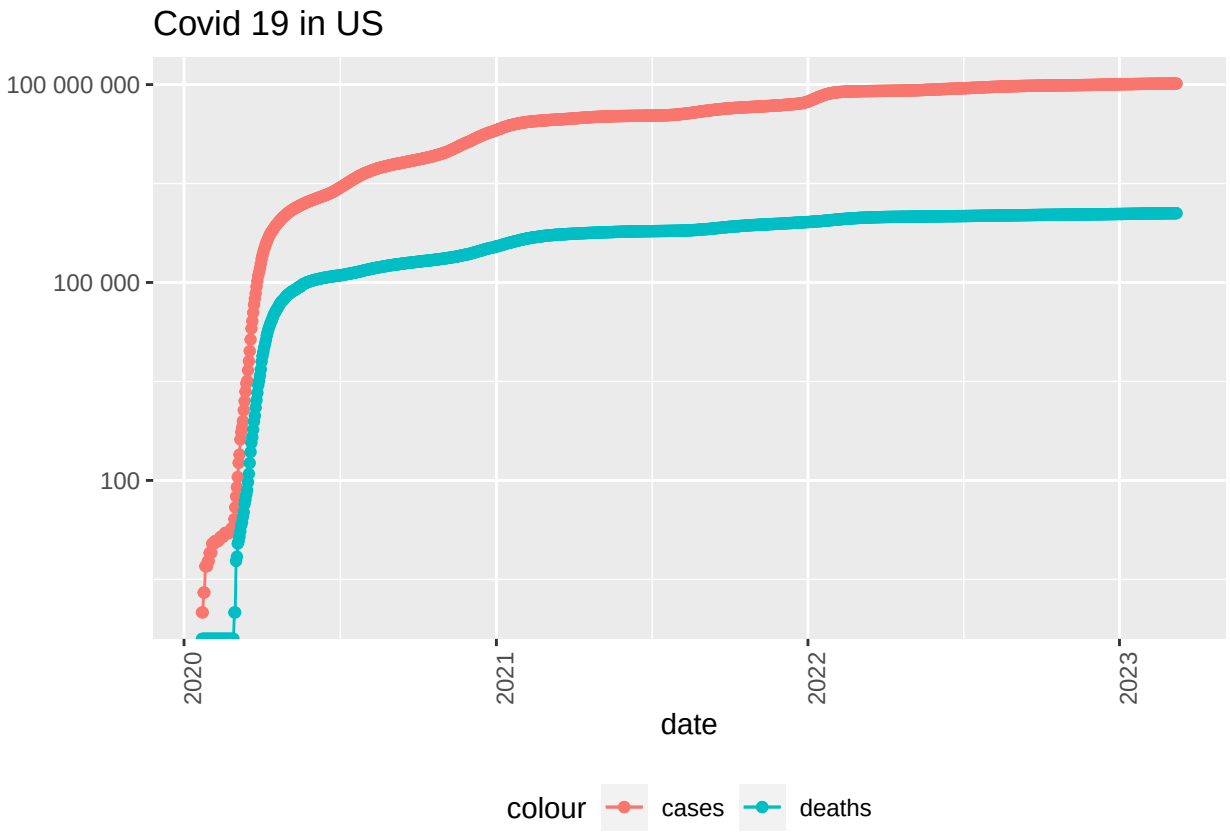
US By State

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

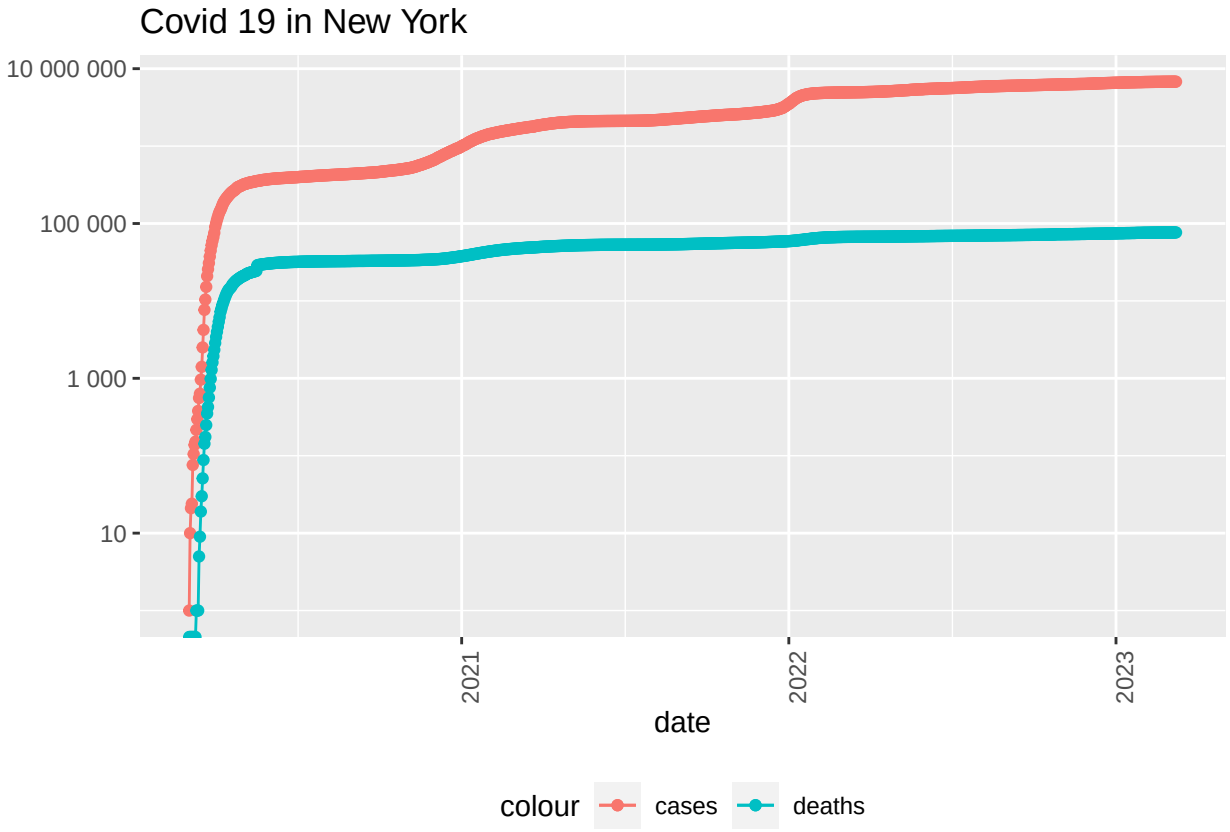
US totals

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

Filter US totals

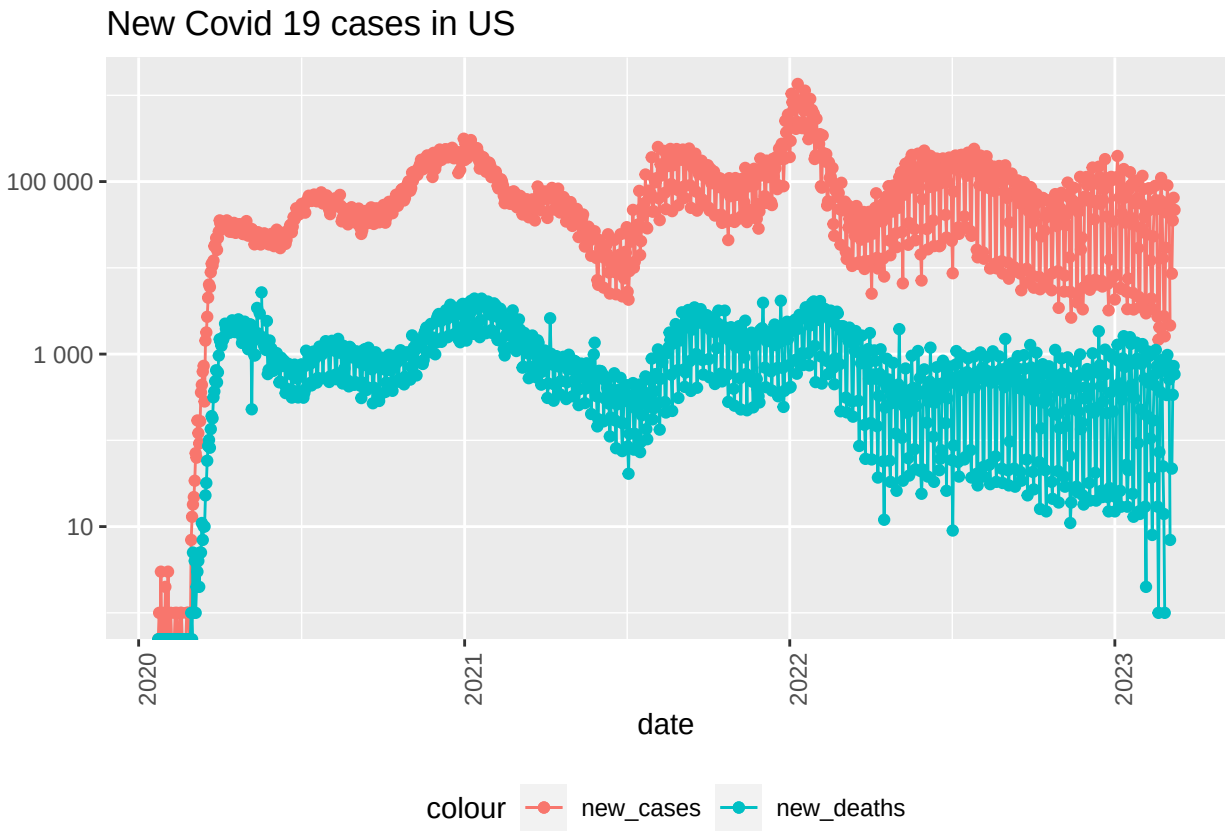


Cases By State



New Cases

Graph New Cases



Highest case rate in us

###Highest Rate

```
## # A tibble: 10 x 6
##    Province_State deaths   cases Population cases_per_thou deaths_per_thou
##    <chr>           <int>   <int>      <int>          <dbl>           <dbl>
##  1 Arizona         33102 2443514    7278717           336.            4.55
##  2 Oklahoma        17972 1290929    3956971           326.            4.54
##  3 Mississippi     13370  990756    2976149           333.            4.49
##  4 West Virginia    7960  642760    1792147           359.            4.44
##  5 New Mexico       9061  670929    2096829           320.            4.32
##  6 Arkansas        13020 1006883    3017804           334.            4.31
##  7 Alabama         21032 1644533    4903185           335.            4.29
##  8 Tennessee       29263 2515130    6829174           368.            4.28
##  9 Michigan        42205 3064125    9986857           307.            4.23
## 10 Kentucky        18130 1718471    4467673           385.            4.06
```

Lowest Rate

```
## # A tibble: 10 x 6
```

```
##    Province_State        deaths  cases Population cases_per_thou deaths_per_thou
##    <chr>                  <int>  <int>      <int>          <dbl>           <dbl>
##  1 American Samoa            34 8.32e3      55641           150.           0.611
##  2 Northern Mariana Isl~     41 1.37e4      55144           248.           0.744
##  3 Virgin Islands           130 2.48e4     107268           231.           1.21
##  4 Hawaii                  1841 3.81e5    1415872           269.           1.30
##  5 Vermont                  929 1.53e5     623989           245.           1.49
##  6 Puerto Rico             5823 1.10e6    3754939           293.           1.55
##  7 Utah                    5298 1.09e6    2785478           391.           1.90
##  8 District of Columbia    1432 1.78e5     705749           252.           2.03
##  9 Alaska                  1486 3.08e5     728809           422.           2.04
## 10 Washington             15683 1.93e6    7614893           253.           2.06
```

Model deaths vs cases

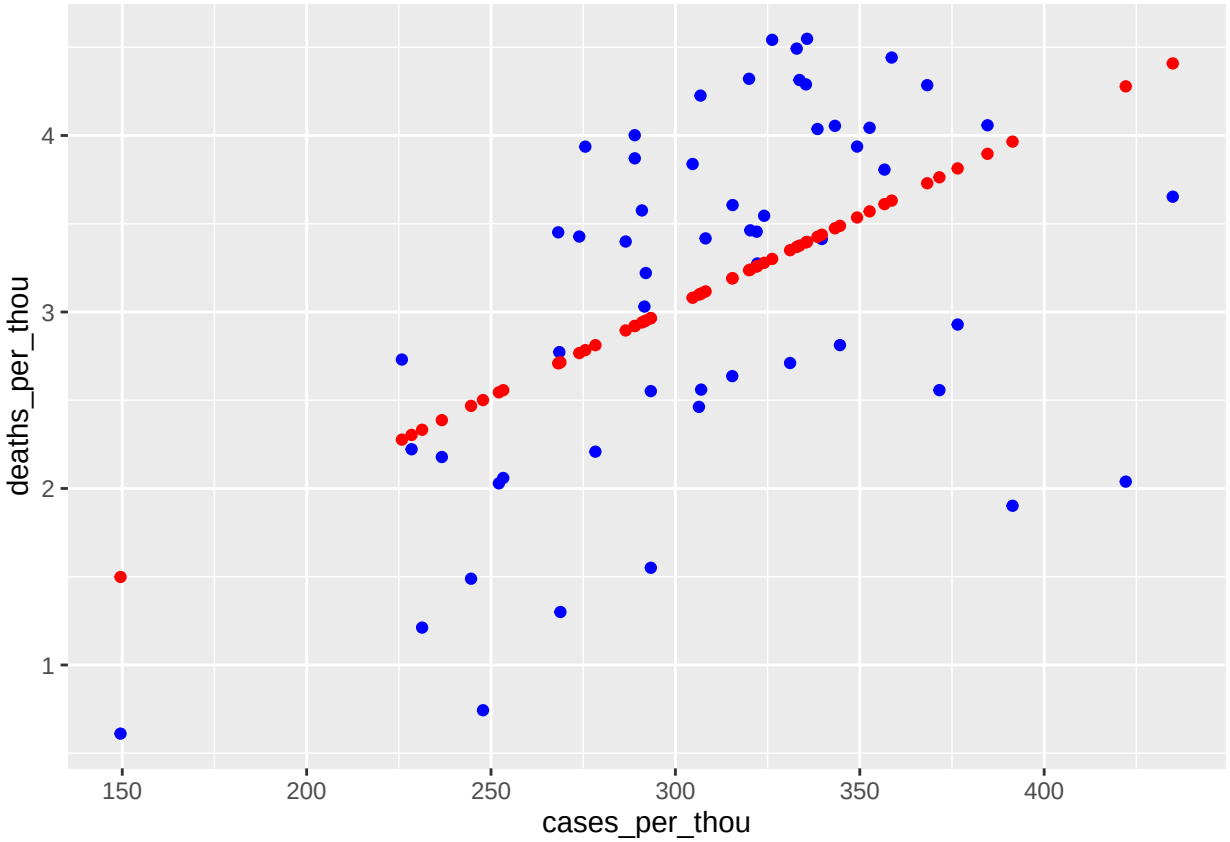
```
##
## Call:
## lm(formula = deaths_per_thou ~ cases_per_thou, data = us_state_totals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2394 -0.6114  0.1965  0.6413  1.2413
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.02599    0.72442  -0.036    0.972
## cases_per_thou  0.01020    0.00231   4.414 4.89e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8803 on 54 degrees of freedom
## Multiple R-squared:  0.2652, Adjusted R-squared:  0.2516
## F-statistic: 19.49 on 1 and 54 DF,  p-value: 4.894e-05
```

Prediction vs actual on plot

```
## # A tibble: 56 x 7
##    Province_State  deaths  cases Population cases_per_thou deaths_per_thou  pred
##    <chr>            <int>  <int>      <int>          <dbl>           <dbl> <dbl>
##  1 Alabama          21032 1.64e6    4903185           335.           4.29   3.39
##  2 Alaska            1486 3.08e5     728809           422.           2.04   4.28
##  3 American Samoa      34 8.32e3      55641           150.           0.611  1.50
##  4 Arizona          33102 2.44e6    7278717           336.           4.55   3.40
##  5 Arkansas         13020 1.01e6    3017804           334.           4.31   3.38
##  6 California      101159 1.21e7   39512223           307.           2.56   3.10
##  7 Colorado         14181 1.76e6    5758736           306.           2.46   3.10
##  8 Connecticut      12220 9.77e5    3565287           274.           3.43   2.77
##  9 Delaware          3324 3.31e5     973764           340.           3.41   3.44
## 10 District of Co~   1432 1.78e5     705749           252.           2.03   2.54
## # i 46 more rows
```

Create prediction table

Plot prediction vs actual



Conlusion

This is a great dataset because it offers endless possiblities for manipulation and analysis. One source of bias I could forsee is in reporting. As data scientists, we can only analyze the data that is provided to us. If different places are reporting cases differently, it would effect our results.