Capstone Project Part 1: US COVID-19 Cases & Deaths - National Averages

Hannah Bravo De Rueda

20 March, 2024

Data Import The New York Times (the Times) has aggregated reported COVID-19 data from state and local governments and health departments since 2020 and provides public access through a repository on GitHub. One of the data sets provided by the Times is county-level data for cumulative cases and deaths each day.

```
# Import New York Times COVID-19 data
us_counties_2020 <-
  read_csv(
    "https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties-2020.csv")
```

```
## Rows: 884737 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): county, state, fips
## dbl  (2): cases, deaths
## date (1): date
##
## 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_counties_2021 <-
  read_csv(
    "https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties-2021.csv")
```

```
## Rows: 1185373 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): county, state, fips
## dbl  (2): cases, deaths
## date (1): date
##
## 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_counties_2022 <-
  read_csv(
    "https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties-2022.csv")
```

```
## Rows: 1188042 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): county, state, fips
## dbl  (2): cases, deaths
## date (1): date
##
## 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.
```

```
# Import Population Estimates from US Census Bureau
us_population_estimates <- read_csv("https://raw.githubusercontent.com/HannahBravo/COVID-19_US_National_
```

```
## Rows: 6286 Columns: 7
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): STNAME, CTYNAME
## dbl (5): fips, STATE, COUNTY, Year, Estimate
##
## 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.
```

```
# Importing a more current population estimate from US Census to include estimates for 2022
us_population_estimates_20_22 <- read_csv("https://raw.githubusercontent.com/HannahBravo/COVID-19_US_Na
```

```
## New names:
## Rows: 52 Columns: 4
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (1): table with row headers in column A and column headers in rows 3 thr... num
## (3): ...2, ...3, ...4
## 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`
## * `` -> `...3`
## * `` -> `...4`
```

I imported data for COVID-19 cases and deaths for each state, and their various counties, across the US, for each year of the pandemic: 2020, 2021, and 2022. Each row of data reports the cumulative number of cases and deaths for a specific county each day. Additionally, I imported population estimates from the US census bureau, unfortunately, the first dataset only contained estimates for 2020, and 2021. So, I chose to import a second dataset of US population estimates from 'census.gov' that contained estimates for all three years; 2020, 2021, and 2022. These county-level population estimates will be used to caluclate statistics per 100,000 people.

Data Exploration The 2020, 2021, and 2022 COVID data sets need to be combined and tidied, to find the total number of deaths and cases for each day since March 15, 2020 (2020-03-15).

```
# glance at the 2020 dataset to get a feel for it.
us_counties_2020 %>% head()
```

```
## # A tibble: 6 x 6
##   date       county    state      fips  cases deaths
##   <date>     <chr>     <chr>      <chr> <dbl>  <dbl>
## 1 2020-01-21 Snohomish Washington 53061     1      0
## 2 2020-01-22 Snohomish Washington 53061     1      0
## 3 2020-01-23 Snohomish Washington 53061     1      0
## 4 2020-01-24 Cook      Illinois   17031     1      0
## 5 2020-01-24 Snohomish Washington 53061     1      0
## 6 2020-01-25 Orange    California 06059     1      0
```

```
# Combine the 2020, 2021, and 2022 COVID data sets.
(us_counties <- us_counties_2020 %>%
  bind_rows(us_counties_2021) %>%
  bind_rows(us_counties_2022))
```

```
## # A tibble: 3,258,152 x 6
##    date       county      state      fips  cases deaths
##    <date>     <chr>       <chr>      <chr> <dbl>  <dbl>
##  1 2020-01-21 Snohomish   Washington 53061     1      0
##  2 2020-01-22 Snohomish   Washington 53061     1      0
##  3 2020-01-23 Snohomish   Washington 53061     1      0
##  4 2020-01-24 Cook        Illinois   17031     1      0
##  5 2020-01-24 Snohomish   Washington 53061     1      0
##  6 2020-01-25 Orange      California 06059     1      0
##  7 2020-01-25 Cook        Illinois   17031     1      0
##  8 2020-01-25 Snohomish   Washington 53061     1      0
##  9 2020-01-26 Maricopa    Arizona    04013     1      0
## 10 2020-01-26 Los Angeles California 06037     1      0
## # i 3,258,142 more rows
```

```
# The combined dataset looks tidy however it contains rows for Puerto Rico, a US territory, so I'll rem
(us_counties <- us_counties %>%
    filter(state != "Puerto Rico"))
```

```
## # A tibble: 3,181,427 x 6
##    date       county      state      fips  cases deaths
##    <date>     <chr>       <chr>      <chr> <dbl>  <dbl>
##  1 2020-01-21 Snohomish   Washington 53061     1      0
##  2 2020-01-22 Snohomish   Washington 53061     1      0
##  3 2020-01-23 Snohomish   Washington 53061     1      0
##  4 2020-01-24 Cook        Illinois   17031     1      0
##  5 2020-01-24 Snohomish   Washington 53061     1      0
##  6 2020-01-25 Orange      California 06059     1      0
##  7 2020-01-25 Cook        Illinois   17031     1      0
##  8 2020-01-25 Snohomish   Washington 53061     1      0
##  9 2020-01-26 Maricopa    Arizona    04013     1      0
## 10 2020-01-26 Los Angeles California 06037     1      0
## # i 3,181,417 more rows
```

```
# Find the total COVID-19 cases & deaths for each day since March 15, 2020

initial_date <- "2020-03-15" # Set demarcation date for beginning of pandemic
last_date <- "2022-12-31" # Set last date for pandemic data
```

```r
# us_total_cases & deaths per day from 3-15-2020 to 12-31-2022
(us_totals <- us_counties %>%
  filter(date >= initial_date) %>%
  group_by(date) %>%
  summarize(us_total_cases = sum(cases), us_total_deaths = sum(deaths)))
```

```
## # A tibble: 1,022 x 3
##    date       us_total_cases us_total_deaths
##    <date>              <dbl>           <dbl>
##  1 2020-03-15           3595              68
##  2 2020-03-16           4502              91
##  3 2020-03-17           5901             117
##  4 2020-03-18           8345             162
##  5 2020-03-19          12387             212
##  6 2020-03-20          17998             277
##  7 2020-03-21          24507             359
##  8 2020-03-22          33050             457
##  9 2020-03-23          43474             577
## 10 2020-03-24          53899             783
## # i 1,012 more rows
```

```r
# Total cases in US on 3-15-2020
(initial_date_cases <- us_counties %>%
  filter(date == initial_date) %>%
  summarize(us_total_cases = sum(cases)))
```

```
## # A tibble: 1 x 1
##   us_total_cases
##            <dbl>
## 1           3595
```

```r
# Total deaths in US on 3-15-2020
(initial_date_deaths <- us_counties %>%
  filter(date == initial_date) %>%
  summarize(us_total_deaths = sum(deaths)))
```

```
## # A tibble: 1 x 1
##   us_total_deaths
##             <dbl>
## 1              68
```

```r
# Total cases in US on 12-31-2022
(last_date_cases <- us_counties %>%
  filter(date == last_date) %>%
  summarize(us_total_cases = sum(cases)))
```

```
## # A tibble: 1 x 1
##   us_total_cases
##            <dbl>
## 1       99374764
```

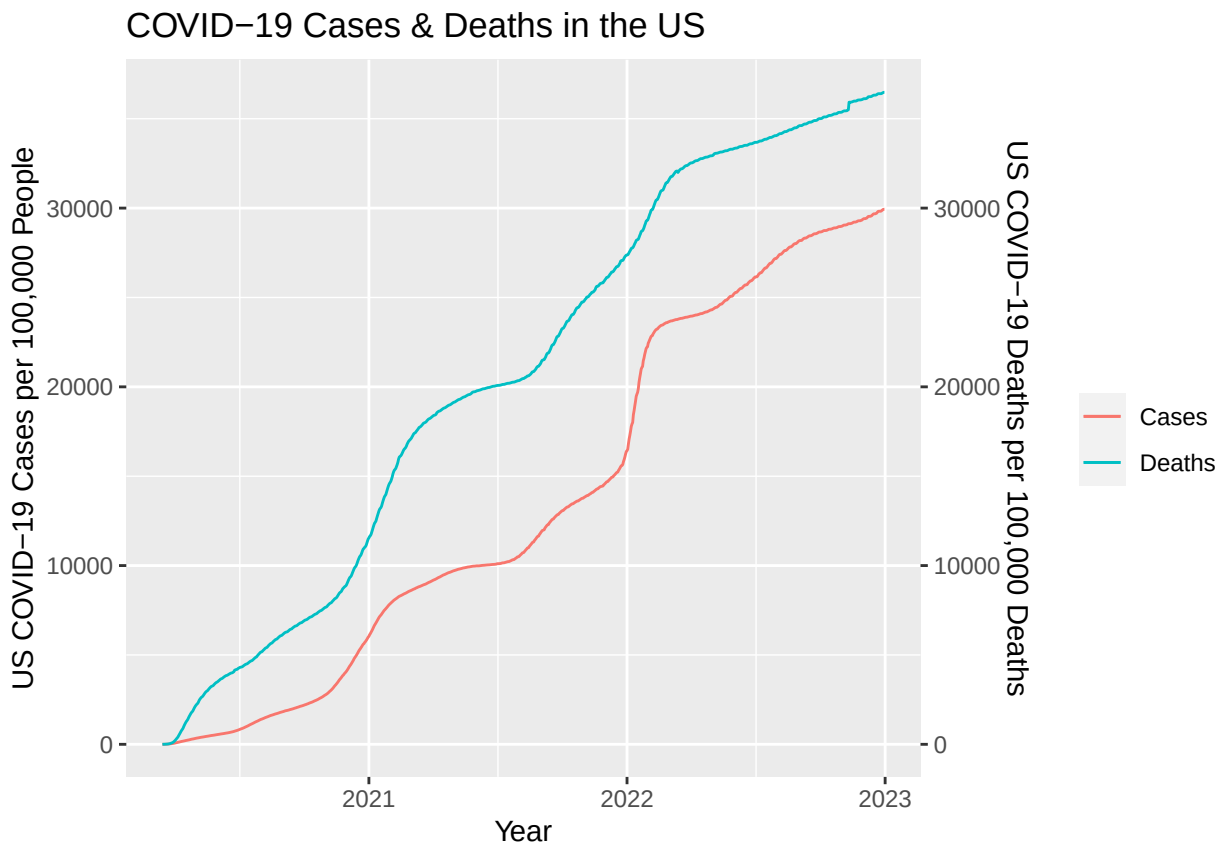
```r
# Total deaths in US on 12-31-2022
(last_date_deaths <- us_counties %>%
  filter(date == last_date) %>%
  summarize(us_total_deaths = sum(deaths)))
```

```
## # A tibble: 1 x 1
##   us_total_deaths
##             <dbl>
## 1         1094296
```

As of 2020-03-15, at the beginning of the pandemic, the total number of COVID-19 cases in the US was 3595 and the total number of deaths was 68. Almost three years later, as of 2022-12-31, the total number of COVID-19 cases shot up to 99374764 and the total number of deaths rose to 1094296. That amounts to almost 30% of the US population contracting COVID-19 over the course of the pandemic, and accounting for 10% of deaths in the US each year of the pandemic. Making COVID-19 the third leading cause of death for residents in the US between 2020 and 2022.

Plotting Total Number of Cases & Deaths Now I'm going to create a visualization for the total number of deaths and cases in the US since March 15, 2020.

```r
# Time series for the total number of US cases and deaths since March 15, 2020.
us_totals %>%
  ggplot(aes(x = date)) + # Map 'date' on x-axis
  geom_line(aes(y = (us_total_cases/332000000*100000), color = "Cases")) +
  # map cases on left-hand y-axis
  # left-hand Y-axis scaled per 100,000 people, the average US population across 2020, 2021, and 2022 i
  geom_line(aes(y = (us_total_deaths/3000000*100000), color = "Deaths")) +
  # map deaths on right-hand y-axis
  # right-hand Y-axis scaled per 10,000 deaths, the average # of deaths in the US across 2020, 2021, an
  scale_y_continuous(
    name = "US COVID-19 Cases per 100,000 People", # Name for left-hand Y-axis
    sec.axis = sec_axis(trans = ~.*1, name = "US COVID-19 Deaths per 100,000 Deaths")) +
  # right-hand y-axis is already transformed, so I set 'trans' argument to multiplying the deaths data
  labs(x = "Year", title = "COVID-19 Cases & Deaths in the US", color = "") # Plot title, and x-axis la
```



Methodology: I created a time series plot comparing the number of COVID-19 cases and deaths across the three years of the pandemic: 2020, 2021, and 2022. I charted the number of COVID-19 cases on the left y-axis, and scaled it per 100,000 people by dividing the number of COVID-19 cases by the average US population, 332 million, for the three years of the pandemic. I then plotted the total number of COVID-19 deaths on the right y-axis and scaled it per 100,000 deaths; by dividing the total number of COVID-19 deaths by the average number of deaths per year in the US, for those three years, 3 million. Therefore, the COVID cases data is plotted relative to the total US population, whereas the COVID deaths data is plotted relative to the total US deaths for those three years.

Results & Interpretation: The chart shows the cumulative growth of both cases and deaths over the three years of the pandemic. Both, COVID-19 cases and deaths seem to follow a similar trend, without showing any outliers or major shifts in the data. The chart starts the beginning of the pandemic, which was early 2020, and goes through to the end of 2022. COVID-19 cases and deaths begins at or near zero, and over the three years, climbs to ~30,000 cases, per 100,000 people and ~35,000 deaths per 100,000 deaths.

Calculating New Cases & Deaths, and 7-day Average Next I'm going to find the number of new cases and deaths each day, as well as a rolling 7-day average, to understand how rapidly the virus is spreading.

```
# Modify previous table of total cases & deaths, to calculate new cases & deaths
(us_totals <- us_totals %>% # Update us_totals table for new columns
    mutate(new_cases = us_total_cases - lag(us_total_cases, order_by = date,
                                            default = 0), # number of new cases each day
           new_deaths = us_total_deaths - lag(us_total_deaths, n = 1, order_by = date,
                                              default = 0)) %>% # number of new deaths each day
   filter(new_deaths >= "0") %>%
          filter(new_cases >= "0"))
```

```
## # A tibble: 1,018 x 5
##    date       us_total_cases us_total_deaths new_cases new_deaths
##    <date>              <dbl>           <dbl>     <dbl>      <dbl>
##  1 2020-03-15           3595              68      3595         68
##  2 2020-03-16           4502              91       907         23
##  3 2020-03-17           5901             117      1399         26
##  4 2020-03-18           8345             162      2444         45
##  5 2020-03-19          12387             212      4042         50
##  6 2020-03-20          17998             277      5611         65
##  7 2020-03-21          24507             359      6509         82
##  8 2020-03-22          33050             457      8543         98
##  9 2020-03-23          43474             577     10424        120
## 10 2020-03-24          53899             783     10425        206
## # i 1,008 more rows
```

```
# Modify new cases & deaths table to calculate a rolling 7-day average
(us_totals <- us_totals %>% # Update us_totals for weekly rolling averages
  mutate(wkly_avg_deaths = round(lag((lead(us_totals$us_total_deaths, n= 7) -
                                        us_total_deaths)/7, n= 7), 1),
         wkly_avg_cases = round(lag((lead(us_totals$us_total_cases, n= 7) -
                                       us_total_cases)/7, n= 7), 0)))
```

```
## # A tibble: 1,018 x 7
##    date       us_total_cases us_total_deaths new_cases new_deaths
##    <date>              <dbl>           <dbl>     <dbl>      <dbl>
##  1 2020-03-15           3595              68      3595         68
##  2 2020-03-16           4502              91       907         23
##  3 2020-03-17           5901             117      1399         26
##  4 2020-03-18           8345             162      2444         45
##  5 2020-03-19          12387             212      4042         50
##  6 2020-03-20          17998             277      5611         65
##  7 2020-03-21          24507             359      6509         82
##  8 2020-03-22          33050             457      8543         98
##  9 2020-03-23          43474             577     10424        120
## 10 2020-03-24          53899             783     10425        206
## # i 1,008 more rows
## # i 2 more variables: wkly_avg_deaths <dbl>, wkly_avg_cases <dbl>
```

```
# max number of new cases
(max_new_cases <- us_totals %>%
    slice(which.max(new_cases)) %>%
    select(new_cases))
```

```
## # A tibble: 1 x 1
##   new_cases
##       <dbl>
## 1   1427097
```

```
# 1-10-2022 with 1.4 million new cases in one day

# min number of new cases
(min_new_cases <- us_totals %>%
    slice(which.min(new_cases)) %>%
    select(new_cases))
```

```
## # A tibble: 1 x 1
##   new_cases
##       <dbl>
## 1       907
```

```
# 3-16-2020 with 907 new cases

# max number of new deaths
(max_new_deaths <- us_totals %>%
    slice(which.max(new_deaths)) %>%
    select(new_deaths))
```

```
## # A tibble: 1 x 1
##   new_deaths
##        <dbl>
## 1      12715
```

```
# 11-11-2022 with 12,715 new deaths in one day

# min number of new deaths
(min_new_deaths <- us_totals %>%
    slice(which.min(new_deaths)) %>%
    select(new_deaths))
```

```
## # A tibble: 1 x 1
##   new_deaths
##        <dbl>
## 1          0
```

```
# 9-4-2022 there were 0 new deaths, but was that the only day with 0 new deaths?

# How many days were there 0 new deaths?
us_totals %>%
  filter(new_deaths == "0")
```

```
## # A tibble: 13 x 7
##    date       us_total_cases us_total_deaths new_cases new_deaths
##    <date>              <dbl>           <dbl>     <dbl>      <dbl>
##  1 2022-09-04       93579541         1038406      4125          0
##  2 2022-09-25       94834344         1046981      5213          0
##  3 2022-10-02       95164638         1050162      5264          0
##  4 2022-10-23       95922687         1058147      2273          0
##  5 2022-10-30       96171860         1060707      2790          0
##  6 2022-11-12       96705038         1077712      1767          0
##  7 2022-11-13       96706992         1077712      1954          0
##  8 2022-11-20       97001230         1079622      2119          0
##  9 2022-11-26       97252291         1081543      1190          0
## 10 2022-12-04       97644392         1083606      2928          0
## 11 2022-12-11       98071928         1086968      3788          0
## 12 2022-12-18       98541102         1089601      3904          0
## 13 2022-12-25       99015890         1092033      3449          0
## # i 2 more variables: wkly_avg_deaths <dbl>, wkly_avg_cases <dbl>
```

```r
# 9-4-2022 through 12-18-2022 there were a total of 12 days with 0 new deaths

# Calculate the min & max of weekly averages for cases and deaths
us_totals %>%
  filter(!is.na(wkly_avg_deaths)) %>%
  summarize(wkly_c_min = min(wkly_avg_cases),  # 4208
            wkly_d_min = min(wkly_avg_deaths), # 55.6
            wkly_c_max = max(wkly_avg_cases),  # 798,663
            wkly_d_max = max(wkly_avg_deaths)) # 3340
```

```
## # A tibble: 1 x 4
##   wkly_c_min wkly_d_min wkly_c_max wkly_d_max
##        <dbl>      <dbl>      <dbl>      <dbl>
## 1       4208       55.6     798663      3340.
```

```r
# Let's find out what days those mins/maxs occurred
us_totals %>%
  filter(wkly_avg_cases == "4208" |     # 3-22-2020 - min wkly_avg_cases
           wkly_avg_deaths == "55.6" |  # 3-22-2020 - min wkly_avg_deaths
           wkly_avg_cases == "798663" | # 1-16-2022 - max wkly-avg_cases
           wkly_avg_deaths == "3339.9") # 1-12-2021 - max wkly-avg_deaths
```

```
## # A tibble: 3 x 7
##   date       us_total_cases us_total_deaths new_cases new_deaths wkly_avg_deaths
##   <date>              <dbl>           <dbl>     <dbl>      <dbl>           <dbl>
## 1 2020-03-22          33050             457      8543         98            55.6
## 2 2021-01-12       22787827          379249    229018       4403          3340.
## 3 2022-01-16       65368695          846998    332338        392          1949
## # i 1 more variable: wkly_avg_cases <dbl>
```

```r
# Calculate the mean and median for new cases/deaths
us_totals %>%
  summarise(mean_new_c = round(mean(new_cases), 2),    # 97338.4
            median_new_c = median(new_cases),          # 58195 a difference of 40%
            mean_new_d = round(mean(new_deaths), 2), # 1071.8
            median_new_d = median(new_deaths))         # 753 a difference of 30%
```

```
## # A tibble: 1 x 4
##   mean_new_c median_new_c mean_new_d median_new_d
##        <dbl>        <dbl>      <dbl>        <dbl>
## 1     97597.        58428      1077.          760
```

```r
# Calculate the mean and median for the 7-day average of cases/deaths
us_totals %>%
  filter(!is.na(wkly_avg_deaths)) %>%
  summarise(mean_avg_c = round(mean(wkly_avg_cases), 2), # 98119
            median_avg_c = median(wkly_avg_cases),         # 65523 a difference of about 40%
            mean_avg_d = round(mean(wkly_avg_deaths), 2),  # 1081.1
            median_avg_d = median(wkly_avg_deaths))        # 813.1 about a difference of about 30%
```

```
## # A tibble: 1 x 4
```

```
##   mean_avg_c median_avg_c mean_avg_d median_avg_d
##        <dbl>        <dbl>      <dbl>        <dbl>
## 1      98119        65523      1081.         813.
```

Methodology: I created a new table, from the combined US county totals table, to calculate the number of new cases and deaths each day, beginning March 15, 2020. I was then able to calculate a 7-day rolling average of both. From there, I determined the maximum and minimum for new cases and new deaths comparing that to the maximum and minimum of weekly averages. Then, I calculated the mean and median for new cases and deaths comparing that to the mean and median for the weekly averages.

Results: The minimum number of new cases in one day was at the very beginning of the pandemic in 2020, on March 16th with 907 new cases, while the maximum number was at the beginning of 2022 on January 10th, with 1427097. As for deaths, the minimum number of new deaths possible is 0 and there were 12 days in 2022 that had no new deaths between September 4th and December 18th. Whereas the maximum number of new deaths occurred in 2022 on November 11th, with 12715. The minimum weekly average for cases and deaths occurred in the first week of the pandemic, in 2020 on March 22nd with 4208 cases, and 55.6 deaths. The maximum average of weekly cases occurred in 2022, on January 16th with 798663 cases; and the maximum average of weekly deaths occurred in 2021, on January 12th with 3339.9 deaths.

There's a significant difference between the mean and median for both new cases and new deaths, suggesting the data is skewed and possibly influenced by outliers. The mean for new COVID cases is: 97597.35 and the median: 58428; while the mean for new COVID deaths is 1077.26 while the median is 760. The mean for for weekly average cases is 98119 while the median is 65523. The mean for weekly average deaths is 1081.05 and the median is 813.1

Interpretation: There is no significant difference between the mean of new cases and the mean of the 7-day average of new case, or between the median of the two. There is also no significant difference between the mean of new deaths and the mean of the 7-day average of new deaths, or between the median of the two. There is however, a significant difference between the mean and median of new cases, as well as a significant difference between the mean and median of new deaths. This difference suggests that there are possible outliers that are skewing the data. Looking at the median of new cases compared to the maximum and minimum of new cases, we see that the lower 50% of new case data lies within a difference of about 58,000; whereas the upper 50% of new case data lies within a difference of about 1.3 million. Meaning that the upper 50% of data for new cases accounts for 96% of the range of new cases. When we look at the median of new deaths, we see the lower 50% of data occurs within a difference of about 700, while the upper 50% of new deaths data occurs within a difference of about 12,000. Meaning that the upper 50% of data for new deaths accounts for 95% of the range of new deaths. This not only reinforces that the data is skewed, it also suggests how skewed the data is, and in which direction. We can conclude that the data for new cases and new deaths, as well as the weekly averages of both is not normally distributed, and is skewed to the right. Now, looking at when the maximums and minimums occurred, there were a number of days in the fall of 2022 where there were no new deaths, which is interesting since the maximum number of new deaths also occurred in the fall of 2022, meaning that both the record high and low for new COVID deaths occurred in the fall of 2022. Additionally, the maximum number of new cases occurred in January of 2022. Suggesting that the last year of the pandemic, 2022, was the most volatile in terms of record highs and lows. Which is surprising since the vaccine was released at the beginning of 2021, however we also saw the virus adapt with different variants. It would be interested to add event lines to a time series plot denoting when the vaccine was released, and when COVID variants were discovered relative to the inflection points of new cases and deaths.

New Cases & Deaths Per 100,000 People Building on the previous table, I'm going to calculate the number of new deaths and cases per 100,000 people each day as well as the 7-day average for new deaths and cases per 100,000 people.

```r
# View the population estimates data
us_population_estimates_20_22 # Everything looks readable and tidy
```

```
## # A tibble: 52 x 4
##    table with row headers in column A and column headers ~1   ...2   ...3   ...4
##    <chr>                                                     <dbl>  <dbl>  <dbl>
##  1 State                                                   2.02e3 2.02e3 2.02e3
##  2 .Alabama                                                5.03e6 5.05e6 5.07e6
##  3 .Alaska                                                 7.33e5 7.34e5 7.34e5
##  4 .Arizona                                                7.18e6 7.26e6 7.36e6
##  5 .Arkansas                                               3.01e6 3.03e6 3.05e6
##  6 .California                                             3.95e7 3.91e7 3.90e7
##  7 .Colorado                                               5.78e6 5.81e6 5.84e6
##  8 .Connecticut                                            3.60e6 3.62e6 3.63e6
##  9 .Delaware                                               9.92e5 1.00e6 1.02e6
## 10 .District of Columbia                                   6.71e5 6.69e5 6.72e5
## # i 42 more rows
## # i abbreviated name:
## #   1: `table with row headers in column A and column headers in rows 3 through 4. (leading dots indi
```

```r
# To calculate per 100,000 people, I'll need to calculate the US population in 2020, 2021, and 2022
# then divide each statistic by the estimated population and finally multiply by 100,000.

# Calculate the total US population for 2020, 2021, and 2022
us_pop_est <- us_population_estimates_20_22 %>%
  rename(state = "table with row headers in column A and column headers in rows 3 through 4. (leading d
         pop_est_2020 = "...2",
         pop_est_2021 = "...3",
         pop_est_2022 = "...4") %>%
  summarise(us_pop_est_2020 = sum(pop_est_2020),
            us_pop_est_2021 = sum(pop_est_2021),
            us_pop_est_2022 = sum(pop_est_2022)) %>%
    pivot_longer(cols = us_pop_est_2020:us_pop_est_2022, names_to = "year", values_to = "us_pop_est") %
    mutate(across("year", str_replace_all, "[us_pop_est_]", "")) %>%
    transform(year = as.numeric(year))

# Update 'us_totals' to extract year from the date column to be able to join the population estimates b
us_totals_per <- us_totals %>%
    mutate(year = year(date)) %>%
    relocate(year, .before = 2)

# Calculate
us_totals_per <- us_totals_per %>%
  left_join(us_pop_est, join_by(year == year)) %>%
  mutate(us_total_cases = case_when(year == "2020" ~ round(us_total_cases/us_pop_est*100000, 2),
                                    year == "2021" ~ round(us_total_cases/us_pop_est*100000, 2),
                                    year == "2022" ~ round(us_total_cases/us_pop_est*100000, 2)),
         us_total_deaths = case_when(year == "2020" ~ round(us_total_deaths/us_pop_est*100000, 3),
                                    year == "2021" ~ round(us_total_deaths/us_pop_est*100000, 3),
                                    year == "2022" ~ round(us_total_deaths/us_pop_est*100000, 3)),
         new_cases = case_when(year == "2020" ~ round(new_cases/us_pop_est*100000, 3),
                                    year == "2021" ~ round(new_cases/us_pop_est*100000, 3),
                                    year == "2022" ~ round(new_cases/us_pop_est*100000, 3)),
```

```
        new_deaths = case_when(year == "2020" ~ round(new_deaths/us_pop_est*100000, 4),
                               year == "2021" ~ round(new_deaths/us_pop_est*100000, 4),
                               year == "2022" ~ round(new_deaths/us_pop_est*100000, 4)),
        wkly_avg_cases = case_when(year == "2020" ~ round(wkly_avg_cases/us_pop_est*100000, 2),
                               year == "2021" ~ round(wkly_avg_cases/us_pop_est*100000, 2),
                               year == "2022" ~ round(wkly_avg_cases/us_pop_est*100000, 2)),
        wkly_avg_deaths = case_when(year == "2020" ~ round(wkly_avg_deaths/us_pop_est*100000, 4),
                               year == "2021" ~ round(wkly_avg_deaths/us_pop_est*100000, 4),
                               year == "2022" ~ round(wkly_avg_deaths/us_pop_est*100000, 4)))

# Remove the population estimate & year columns that were used for calculations
us_totals_per <- us_totals_per %>%
  select(date, us_total_cases, us_total_deaths, new_cases, new_deaths, wkly_avg_cases, wkly_avg_deaths)
us_totals_per
```

```
## # A tibble: 1,018 x 7
##    date       us_total_cases us_total_deaths new_cases new_deaths wkly_avg_cases
##    <date>              <dbl>           <dbl>     <dbl>      <dbl>          <dbl>
##  1 2020-03-15           1.08           0.021     1.08      0.0205          NA
##  2 2020-03-16           1.36           0.027     0.274     0.0069          NA
##  3 2020-03-17           1.78           0.035     0.422     0.0078          NA
##  4 2020-03-18           2.52           0.049     0.737     0.0136          NA
##  5 2020-03-19           3.74           0.064     1.22      0.0151          NA
##  6 2020-03-20           5.43           0.084     1.69      0.0196          NA
##  7 2020-03-21           7.39           0.108     1.96      0.0247          NA
##  8 2020-03-22           9.97           0.138     2.58      0.0296           1.27
##  9 2020-03-23          13.1            0.174     3.14      0.0362           1.68
## 10 2020-03-24          16.3            0.236     3.14      0.0621           2.07
## # i 1,008 more rows
## # i 1 more variable: wkly_avg_deaths <dbl>
```

```
# Calculate the max and min of new cases and deaths
us_totals_per %>%
  summarise(min_new_cases = min(new_cases), # matches the same date as the previous table
            max_new_cases = max(new_cases), # matches the same date as the previous table
            min_new_deaths = min(new_deaths), # matches the same date as the previous table
            max_new_deaths = max(new_deaths)) # matches the same date as the previous table
```

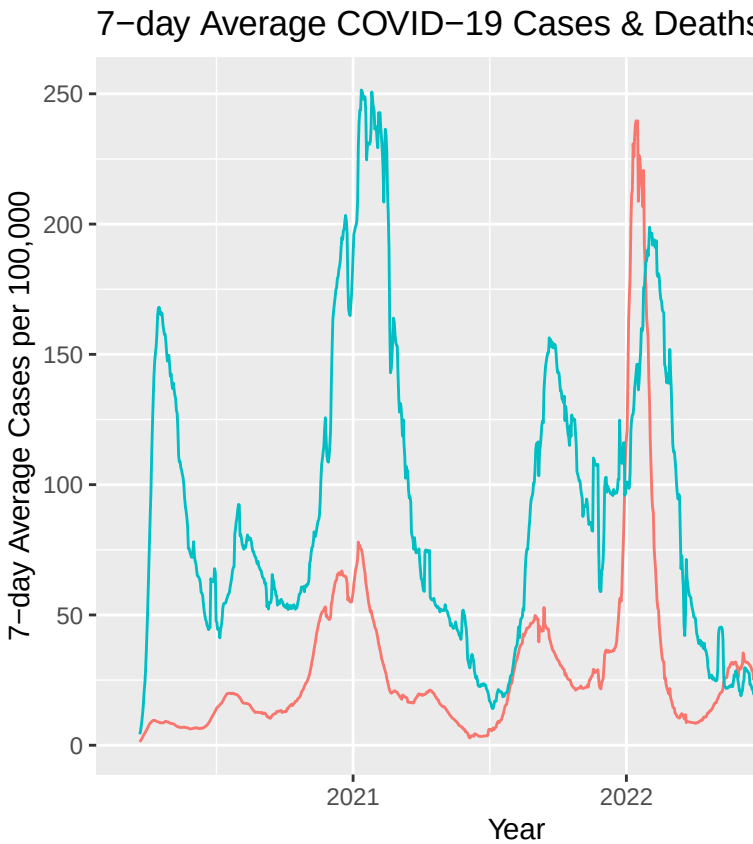
```
## # A tibble: 1 x 4
##   min_new_cases max_new_cases min_new_deaths max_new_deaths
##           <dbl>         <dbl>          <dbl>          <dbl>
## 1         0.274          428.              0           3.82
```

Methodology: I created a new table 'us_totals_per' based off the table in Question 3, 'us_totals', which scales the calculations from the previous table to be per 100,000 people. In order to scale those calculations, I first needed to calculate the total US population for each year of the pandemic; 2020, 2021, and 2022; using the US population estimates data. Once the population estimate was totaled for each year, I then joined those totals to the 'previous 'us_totals' table by year. I was then able to update each column in the 'us_totals' table by dividing the data in each column by the corresponding year's population estimate, and then multiplying by 100,000.

Results & Interpretation: The dimensions of the resulting table didn't change, but the data itself is now much easier to digest and interpret. Scaling the data per 100,000 people standardizes the numbers and

calculations across each year of the pandemic, controlling for changes in a population over the course of a year. However, the maximums and minimums for new cases and deaths still occurred on the same dates as those from the non-scaled data.

```
# Create a visualization to compare the seven-day average cases and deaths per 100,000 people.
us_totals_per %>%
  ggplot(aes(x = date)) + # Map 'date' on x-axis
  geom_line(aes(y = wkly_avg_cases, color = "Cases")) +
  # map cases on left-hand y-axis
  # left-hand Y-axis scaled per 100,000 people, the average US population across 2020, 2021, and 2022 i
  geom_line(aes(y = wkly_avg_deaths*250, color = "Deaths")) +
  # map deaths on right-hand y-axis
  # right-hand Y-axis scaled per 10,000 deaths, the average # of deaths in the US across 2020, 2021, an
  scale_y_continuous(
    name = "7-day Average Cases per 100,000", # Name for left-hand Y-axis
    sec.axis = sec_axis(trans = ~./250, name = "7-day Average Deaths per 100,000")) +
  # right-hand y-axis is already transformed, so I set 'trans' argument to multiplying the deaths data
  labs(x = "Year", title = "7-day Average COVID-19 Cases & Deaths in the US", color = "") # Plot title,
```



Plot to Compare 7-day Average of Cases & Deaths

Methodology: I created a dual axis time series plot comparing the seven-day average of COVID-19 cases per 100,000 people and COVID-19 deaths per 100,000 people, across the three years of the pandemic. I placed the 7-day average of COVID cases on the left y-axis and the 7-day average of COVID deaths on the right y-axis. The weekly average data used for this plot was already scaled to be per 100,000 people in the previous exercise.

Results: The overall trend of weekly averages for both cases and deaths follow a similar cyclical pattern, however there are a few time points on the graph where their patterns don't seem proportionate. In January of 2022, the weekly average of cases reaches an all time high almost three times greater than previous spikes, while the weekly average of deaths also displays a spike, it's only it's second highest. . Then again in the summer of 2022, an increase in the weekly average deaths causes it to surpass a decrease in the weekly average cases, after which it falls below once more. Additionally, in the initial months of the pandemic, there is a spike in weekly average cases without an accompanying spike in weekly average deaths. All of these spikes and dips are gradual, except for a

Interpretation: The time series plot shows that in April 2020 there was a significant spike in COVID-19 related deaths, but no corresponding spike in COVID-19 cases. After that, we see the COVID cases line proportionally mimic the COVID deaths line for the remainder of 2020, through 2021, up until 2022. In January of 2022 there is a significant spike in COVID cases, without a proportionate spike in deaths. After which, they both drop back off, until the summer of 2022, where there is another increase in COVID cases, while COVID related deaths displays a continued drop. COVID cases seems to drop back down again, but then in November of 2022, there is a very steep and dramatic spike in deaths that seems to last a matter of days before dropping right back down to its previous level, while COVID cases experiences a continued drop. Given how gradual the over spikes in COVID deaths were, and this spike accounting for the highest number of new COVID deaths, I suspect it may be an error. However, more information and statistical testing would need to be done to verify.