COVID 19 Analysis – Part 1

Meisam Yousefi

2023-03-06

Required Packages

Part 1 - Basic Exploration of US Data

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. This will be your primary data set for the first two parts of your analysis.

County-level COVID data from 2020, 2021, and 2022 has been imported below. Each row of data reports the cumulative number of cases and deaths for a specific county each day. A FIPS code, a standard geographic identifier, is also provided which you will use in Part 2 to construct a map visualization at the county level for a state.

Additionally, county-level population estimates reported by the US Census Bureau has been imported as well. You will use these estimates to caluclate statistics per 100,000 people.

```
# Import New York Times COVID-19 data
# Import Population Estimates from US Census Bureau

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.
```

```
us_population_estimates <- read_csv("fips_population_estimates.csv")
```

```
## 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.
```

Question 1

Your first task is to combine and tidy the 2020, 2021, and 2022 COVID data sets and find the total deaths and cases for each day since March 15, 2020 (2020-03-15). The data sets provided from the NY Times also includes statistics from Puerto Rico, a US territory. You may remove these observations from the data as they will not be needed for your analysis. Once you have tidied the data, find the total COVID-19 cases and deaths since March 15, 2020. Write a sentence or two after the code block communicating your results. Use inline code to include the `max_date`, `us_total_cases`, and `us_total_deaths` variables. To write inline code use `r`.

Answer 1

```
# Combine the three 2020, 2021, and 2022 COVID data sets using rbind

us_counties_total <- rbind(us_counties_2020, us_counties_2021, us_counties_2022)

# Removing "Puerto Rico" from the states list, and removing dates prior to March 15th
2020

us_counties_total <- us_counties_total %>% dplyr::filter(state != "Puerto Rico",
                                                         date > "2020-03-14")

# Summarizing data for total death and cases, per day

us_combined_total <- us_counties_total %>%
  group_by(date) %>%
  summarise("total_deaths" = sum(deaths),
            "total_cases" = sum(cases))

# Calculating values for the communication part

max_date <- max(us_combined_total$date) # replace the quotes with your code to find t
he most recent date in the data set
us_total_cases <- us_combined_total$total_cases[us_combined_total$date == max_date]
us_total_deaths <- us_combined_total$total_deaths[us_combined_total$date == max_date]
```

Displaying the output final table:

```
us_combined_total
```

date <date>	**total_deaths** <dbl>	**total_cases** <dbl>
2020-03-15	68	3595
2020-03-16	91	4502
2020-03-17	117	5901
2020-03-18	162	8345
2020-03-19	212	12387
2020-03-20	277	17998
2020-03-21	359	24507
2020-03-22	457	33050
2020-03-23	577	43474
2020-03-24	783	53899

As of December 31, 2022, there has been a cumulative number of 9.9374764^{7} individuals in the US who

were diagnosed with COVID-19, and there has been 1.094296^{6} deaths reported.

In this analysis we used the data from NYTimes on the daily number of cases and deaths in each county, from the begining of the pandemic until 2022-12-31.

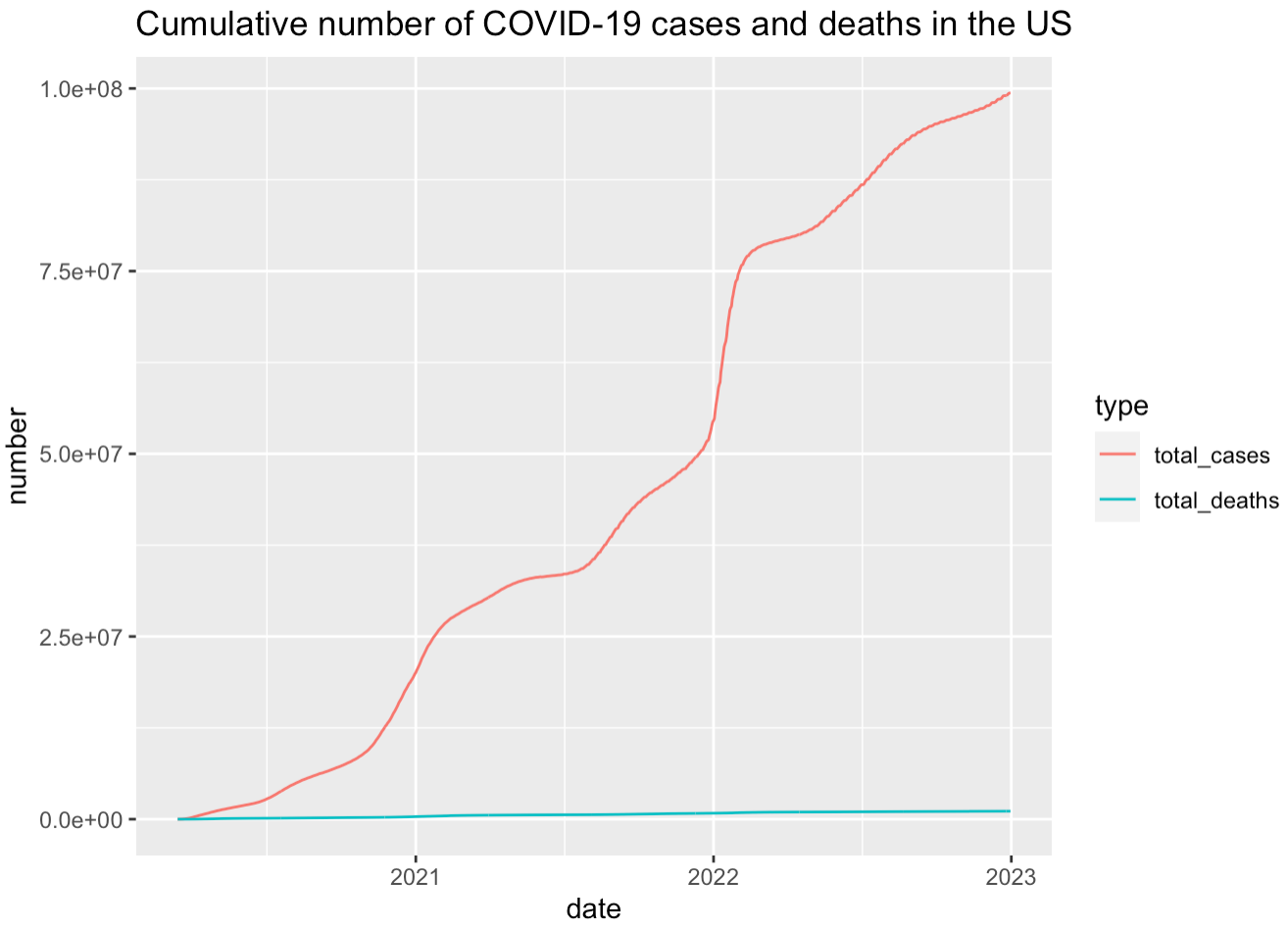
Question 2

Create a visualization for the total number of deaths and cases in the US since March 15, 2020. Before you create your visualization, review the types of plots you can create using the ggplot2 library and think about which plots would be effective in communicating your results. After you have created your visualization, write a few sentences describing your visualization. How could the plot be interpreted? Could it be misleading?

Answer 2

I'll present the data with a simple line-graph, with two separate lines one for the total number of deaths and the other for the total number of cases. To do so we might want to first pivot the table to the long format.

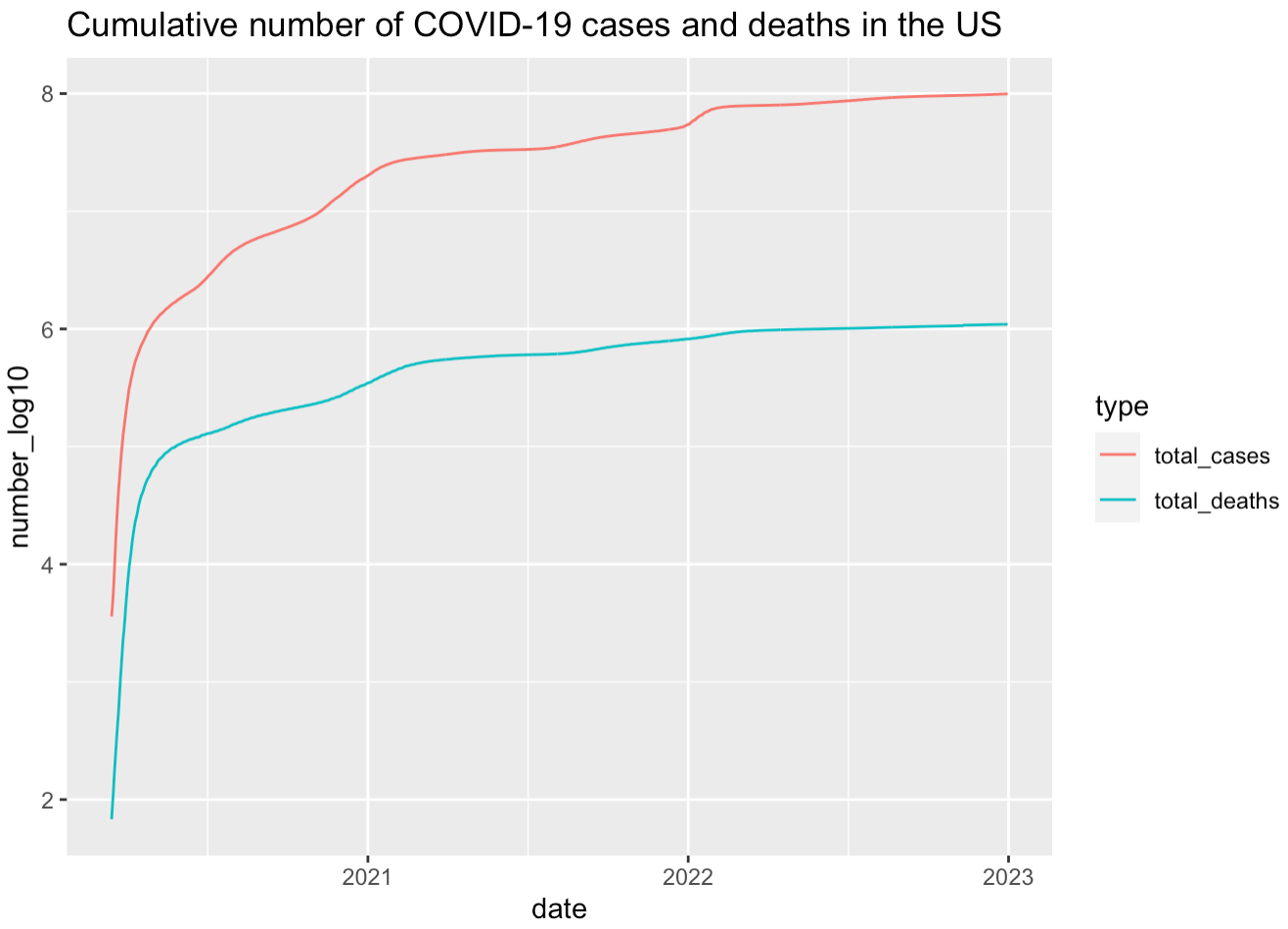
```
us_combined_total %>%
  pivot_longer(cols = -date, names_to = "type", values_to = "number") %>%
  ggplot(aes(x = date, y = number)) + geom_line(aes(color = type)) + labs(title = "Cu
mulative number of COVID-19 cases and deaths in the US")
```



The plot is effective in communicating the message and for the audience to get a grasp of the COVID-19 situation in the states until the end of the 2022, however there is probably one point which might be misleading: Since the actual number of the cases is orders of magnitude higher than the death, the death line seems to be static from this plot so the audience might think that the rate of the cases is getting higher than

the deaths. We can overcome this by converting all numbers to log transformed:

```
us_combined_total %>%
  pivot_longer(cols = -date, names_to = "type", values_to = "number") %>%
  mutate(number_log10 = log10(number)) %>%
  ggplot(aes(x = date, y = number_log10)) + geom_line(aes(color = type)) + labs(title
= "Cumulative number of COVID-19 cases and deaths in the US")
```



Now we can clearly see that the cumulative number of the deaths from COVID-19 is almost always 2 log lower than the total number of the cases, which brings us to an almost 1% chance of death from COVID-19 which was not growing as the pandemic progressed.

Question 3

While it is important to know the total deaths and cases throughout the COVID-19 pandemic, it is also important for local and state health officials to know the the number of new cases and deaths each day to understand how rapidly the virus is spreading. Using the table you created in Question 1, calculate the number of new deaths and cases each day and a seven-day average of new deaths and cases. Once you have organized your data, find the days that saw the largest number of new cases and deaths. Write a sentence or two after the code block communicating your results.

Answer 3

```
# Calculating the number of new deaths and cases each day and a seven day average of
new deaths and cases

us_combined_2 <- us_combined_total %>% mutate(
  delta_deaths_1 = total_deaths - lag(total_deaths),
  delta_cases_1 = total_cases - lag(total_cases),
  delta_deaths_7 = zoo::rollmean(delta_deaths_1, k = 7, fill = NA, align = "right"),
  delta_cases_7 = zoo::rollmean(delta_cases_1, k = 7, fill = NA, align = "right")
)

us_combined_2
```

date <date>	total_deaths <dbl>	total_cases <dbl>	delta_deaths_1 <dbl>	delta_cases_1 <dbl>	delta_deaths_7 <dbl>
2020-03-15	68	3595	NA	NA	NA
2020-03-16	91	4502	23	907	NA
2020-03-17	117	5901	26	1399	NA
2020-03-18	162	8345	45	2444	NA
2020-03-19	212	12387	50	4042	NA
2020-03-20	277	17998	65	5611	NA
2020-03-21	359	24507	82	6509	NA
2020-03-22	457	33050	98	8543	55.57143
2020-03-23	577	43474	120	10424	69.42857
2020-03-24	783	53899	206	10425	95.14286

1-10 of 1,022 rows | 1-6 of 7 columns Previous **1** 2 3 4 5 6 … 103 Next

```
# Finding the days with the highest number of new cases and deaths

max_new_cases_date <- us_combined_2$date[us_combined_2$delta_cases_1 == max(us_combin
ed_2$delta_cases_1, na.rm = T)]

max_new_deaths_date <- us_combined_2$date[us_combined_2$delta_deaths_1 == max(us_comb
ined_2$delta_deaths_1, na.rm = T)]
```

We can see that the pandemic has not been less severe in 2022, as the highest daily number of new confirmed cases belongs to 2022-01-10, and the highest number of deaths happened on NA, 2022-11-11 with 1.2715^{4} individuals died on that day.

Question 4

Create a new table, based on the table from Question 3, and calculate the number of new deaths and cases per 100,000 people each day and a seven day average of new deaths and cases per 100,000 people.

```
# Calculate the US total population in 2020, 2021 and 2022. Since the "population est
imates" were only available for 2020 and 2021, I extrapolated the total population in
2022 assuming a linear rate

us_population_total <- us_population_estimates %>% group_by(Year) %>% summarise(Popul
ation = sum(Estimate))

us_population_total <- rbind(us_population_total, c(2022, 2*(us_population_total$Popu
lation[2]) - us_population_total$Population[1]))

# Dividing each statistics by the total population and then multiplying by 100,000

us_combined_3 <- us_combined_2 %>%
  mutate(across(-date, ~ case_when(date < "2021-01-01" ~ (.x*100000)/us_population_to
tal$Population[us_population_total$Year == "2020"],
                                                   date >= "2021-01-01" &  date < "202
2-01-01" ~ (.x*100000)/us_population_total$Population[us_population_total$Year == "20
21"],
                                                   date >= "2022-01-01" ~ (.x*100000)/
us_population_total$Population[us_population_total$Year == "2022"])
                ))

us_combined_3
```

date <date>	total_deaths <dbl>	total_cases <dbl>	delta_deaths_1 <dbl>	delta_cases_1 <dbl>	delta_deaths_7 <dbl>
2020-03-15	0.02051275	1.084461	NA	NA	NA
2020-03-16	0.02745089	1.358065	0.0069381373	0.2736039	NA
2020-03-17	0.03529400	1.780085	0.0078431117	0.4220197	NA
2020-03-18	0.04886862	2.517337	0.0135746164	0.7372525	NA
2020-03-19	0.06395153	3.736639	0.0150829071	1.2193022	NA
2020-03-20	0.08355931	5.429243	0.0196077793	1.6926038	NA
2020-03-21	0.10829527	7.392736	0.0247359677	1.9634928	NA
2020-03-22	0.13785777	9.969802	0.0295624980	2.5770655	0.01676357
2020-03-23	0.17405675	13.114286	0.0361989771	3.1444845	0.02094369
2020-03-24	0.23619833	16.259072	0.0621415773	3.1447861	0.02870062

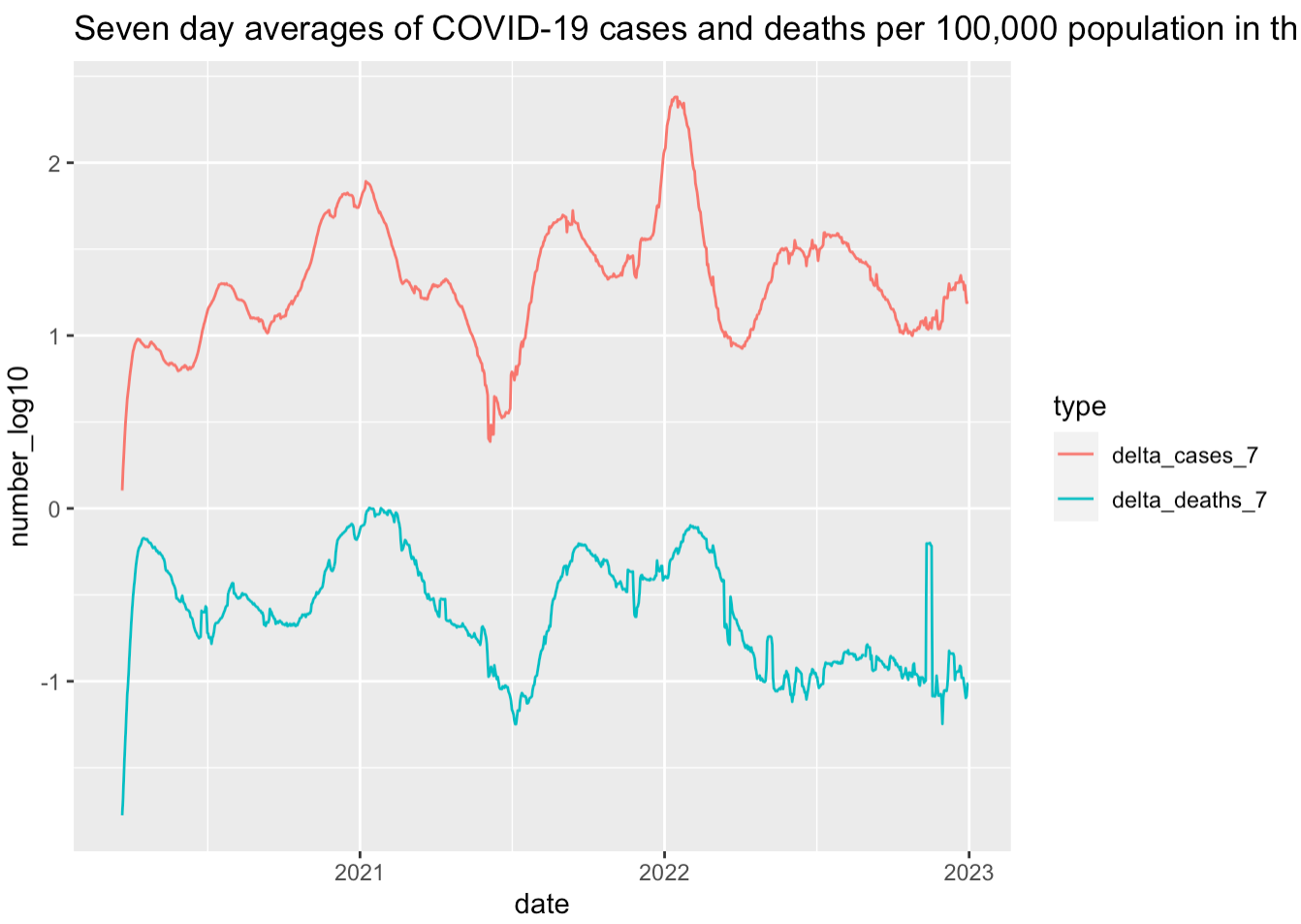
Question 5

Create a visualization to compare the seven-day average cases and deaths per 100,000 people

Answer 5

```
us_combined_3 %>%
  pivot_longer(cols = ends_with("7"), names_to = "type", values_to = "number") %>%
  mutate(number_log10 = log10(number)) %>%
  ggplot(aes(x = date, y = number_log10)) + geom_line(aes(color = type)) + labs(title
= "Seven day averages of COVID-19 cases and deaths per 100,000 population in the US")
```

```
## Warning: Removed 14 rows containing missing values (`geom_line()`).
```



By this we can see that again, the pattern of the COVID-19 cases and deaths are similar, with several waves of the pandemic from early 2020 until the end of 2022. However, notably, while the highest weekly average of case diagnosis is in early 2022, we see that the pick deaths belong to the early 2021 which probably shows the impact of vaccination programs on us national COVID-19 burden.

End of part 1