Codes

Alice Zhang

Data

Load Packages

```r
library(tidyverse)
library(lubridate)
library(janitor)
library(stringr)
library(ggplot2)
library(httr2)
library(knitr)
```

Retrieve Data

For population data, only 2020-2021 population was available from pset 4 dataset.

```r
url <- "https://api.census.gov/data/2021/pep/population"
census_key = "7787d890d72636773b5bc1694aef2bff7ea56237"

request <- request(url) |>
  req_url_query(get = I("POP_2020,POP_2021,NAME"),
                `for` = I("state:*"),
                key = census_key)
response <- request |> req_perform()
pop <- response |> resp_body_json(simplifyVector = TRUE)
```

Get COVID data:

```
api <- "https://data.cdc.gov/resource/pwn4-m3yp.json"

cases_full <- request(api) |>
  req_url_query(`$limit` = 10000000000) |>
  req_perform() |>
  resp_body_json()|>
  map_df(~ as_tibble(.))
```

Wrangling

```
population <- pop |>
  row_to_names(row_number = 1) |>
  as_tibble() |>
  select(-state) |>
  rename(state_name = NAME) |>
  pivot_longer(-state_name, names_to = "year", values_to = "population") |>
  mutate(year = str_remove(year, "POP_")) |>
  mutate(across(-state_name, as.numeric)) |>
  mutate(state = case_when(state_name == "Puerto Rico" ~ "PR",
                           state_name == "District of Columbia" ~ "DC",
                           TRUE ~ state.abb[match(state_name, state.name)]))

cases_cleaned <- cases_full |>
  select(state, end_date, new_cases) |>
  mutate(date = as.Date(end_date, format = "%Y-%m-%d"),
         cases = as.numeric(new_cases)) |>
  select(-end_date, -new_cases)

head(cases_cleaned)
```

```
# A tibble: 6 x 3
  state date       cases
  <chr> <date>     <dbl>
1 AZ    2023-02-22  3716
2 LA    2022-12-21  4041
3 GA    2023-02-22  5298
4 LA    2023-03-29  2203
5 LA    2023-02-01  5725
6 LA    2023-03-22  1961
```

Question 1

Divide the pandemic period, January 2020 to December 2024 into waves. Justify your choice with data visualization.

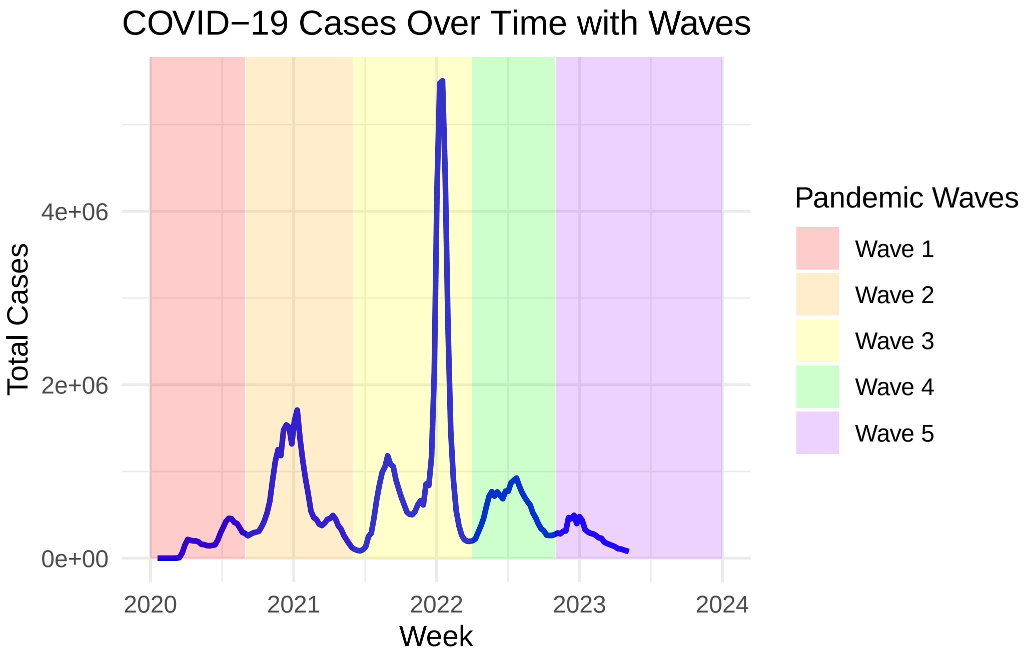
```r
cases_week <- cases_cleaned |>
  mutate(week = floor_date(date, "week")) |> # Aggregate by week
  group_by(state, week) |>
  summarise(total_cases = sum(cases, na.rm = TRUE),
            .groups = "drop")

# Summarize across states
us_cases <- cases_week |>
  group_by(week) |>
  summarise(total_cases = sum(total_cases, na.rm = TRUE),
            .groups = "drop")
```

```r
wave_periods <- data.frame(
  wave = c("Wave 1", "Wave 2", "Wave 3", "Wave 4", "Wave 5"),
  start = as.Date(c("2020-01-01", "2020-09-01", "2021-06-01",
                    "2022-04-01", "2022-11-01")),
  end = as.Date(c("2020-08-30", "2021-05-31", "2022-03-31",
                  "2022-10-31", "2023-12-31"))
)

ggplot(us_cases, aes(x = week, y = total_cases)) +
  geom_line(color = "blue", size = 1) +
  geom_rect(data = wave_periods,
            aes(xmin = start, xmax = end, ymin = 0, ymax = Inf, fill = wave),
            alpha = 0.2, inherit.aes = FALSE) +
  scale_fill_manual(values = c("Wave 1" = "red", "Wave 2" = "orange",
                               "Wave 3" = "yellow", "Wave 4" = "green",
                               "Wave 5" = "purple")) +
  labs(title = "COVID-19 Cases Over Time with Waves",
       x = "Week",
       y = "Total Cases",
       fill = "Pandemic Waves") +
  theme_minimal()
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



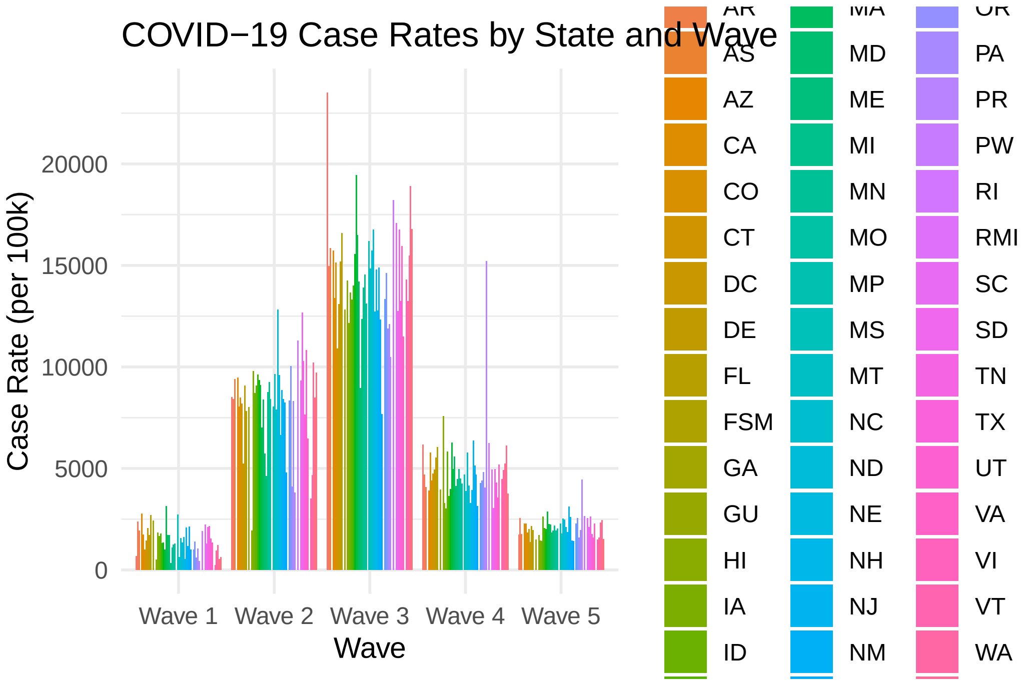
Question 2

For each period compute the deaths rates by state. Describe which states did better or worse during the different periods.

```
# Assign cases to waves
cases_by_wave <- cases_cleaned |>
  mutate(wave = case_when(
    date >= wave_periods$start[1] & date <= wave_periods$end[1] ~ "Wave 1",
    date >= wave_periods$start[2] & date <= wave_periods$end[2] ~ "Wave 2",
    date >= wave_periods$start[3] & date <= wave_periods$end[3] ~ "Wave 3",
    date >= wave_periods$start[4] & date <= wave_periods$end[4] ~ "Wave 4",
    date >= wave_periods$start[5] & date <= wave_periods$end[5] ~ "Wave 5",
    TRUE ~ NA_character_
  )) |>
  filter(!is.na(wave)) |>
  group_by(state, wave) |>
  summarise(total_cases = sum(cases, na.rm = TRUE), .groups = "drop")

# Include population data
cases_with_population <- cases_by_wave |>
  left_join(population, by = c("state")) |>
  mutate(case_rate = total_cases / population * 100000)
```

```
ggplot(cases_with_population, aes(x = wave, y = case_rate, fill = state)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "COVID-19 Case Rates by State and Wave",
       x = "Wave",
       y = "Case Rate (per 100k)",
       fill = "State") +
  theme_minimal()
```



```
# Highlight states with highest and lowest case rates
rates = cases_with_population |>
  group_by(wave) |>
  summarise(
    max_rate_state = state[which.max(case_rate)],
    max_rate = max(case_rate, na.rm = TRUE),
    min_rate_state = state[which.min(case_rate)],
    min_rate = min(case_rate, na.rm = TRUE))

kable(rates)
```

wave	max_rate_state	max_rate	min_rate_state	min_rate
Wave 1	LA	3151.395	VT	241.1822
Wave 2	ND	12815.828	HI	1919.5391
Wave 3	AK	23516.160	NY	7543.5180
Wave 4	PR	15217.687	ID	2941.4658
Wave 5	PR	4440.486	DC	1332.4291

Question 3

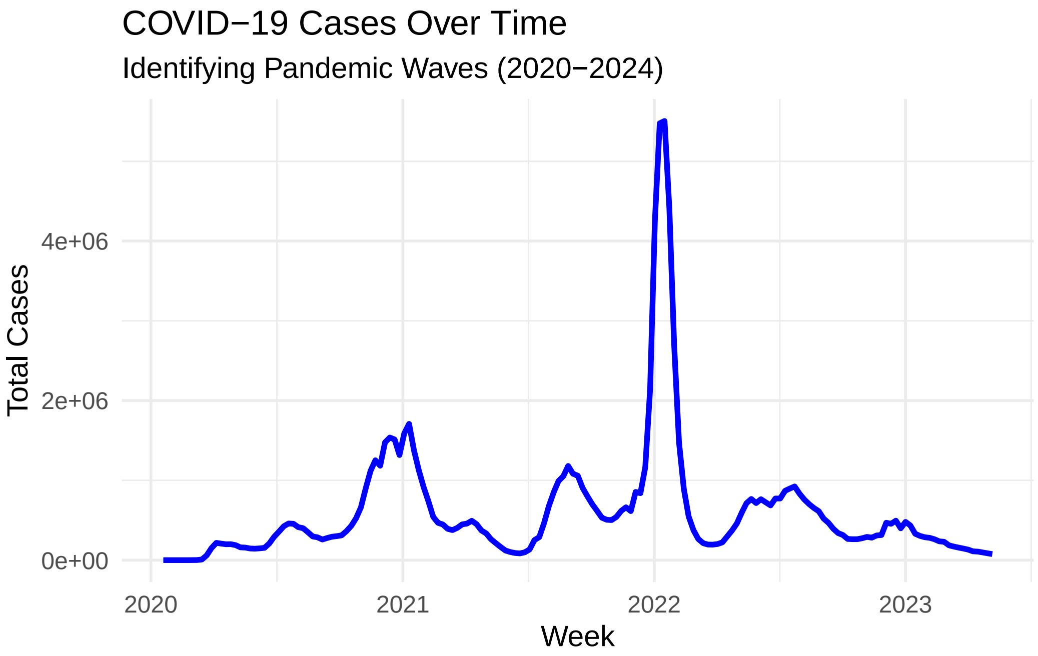
Describe if COVID-19 became less or more virulent across the different periods.

```r
# Summarize total cases per wave
cases_by_wave_summary <- cases_with_population |>
  group_by(wave) |>
  summarise(
    total_cases = sum(total_cases, na.rm = TRUE),
    avg_case_rate = mean(case_rate, na.rm = TRUE)
  )
```

Supplementary

Initial cases vs. time (week) plot for identifying waves.

```r
# Plot the entire US cases over time
ggplot(us_cases, aes(x = week, y = total_cases)) +
  geom_line(color = "blue", size = 1) +
  # geom_smooth(method = "loess", span = 0.2, color = "red", se = FALSE) +
  labs(title = "COVID-19 Cases Over Time",
       x = "Week",
       y = "Total Cases",
       subtitle = "Identifying Pandemic Waves (2020-2024)") +
  theme_minimal()
```


COVID−19 Cases Over Time
Identifying Pandemic Waves (2020−2024)
Total Cases
4e+06
2e+06
0e+00
2020
2021
2022
2023
Week