STA 141A Project

Filip Wilhelm Sjostrand

2023-03-22

```
# Packages ----------
library(dplyr)
library(ggplot2)
library(ggthemes)
library(mapproj)
library(maps)
library(readr)
library(readxl)
library(viridis)
```

Introduction

In this report, we examine the trends in daily cases and death rates due to Covid-19 in the United States, analyzing the data at national, state, and county levels. The aim is to provide an in-depth understanding of the progression of the pandemic and to identify potential factors influencing these patterns.

Exercise I

a)

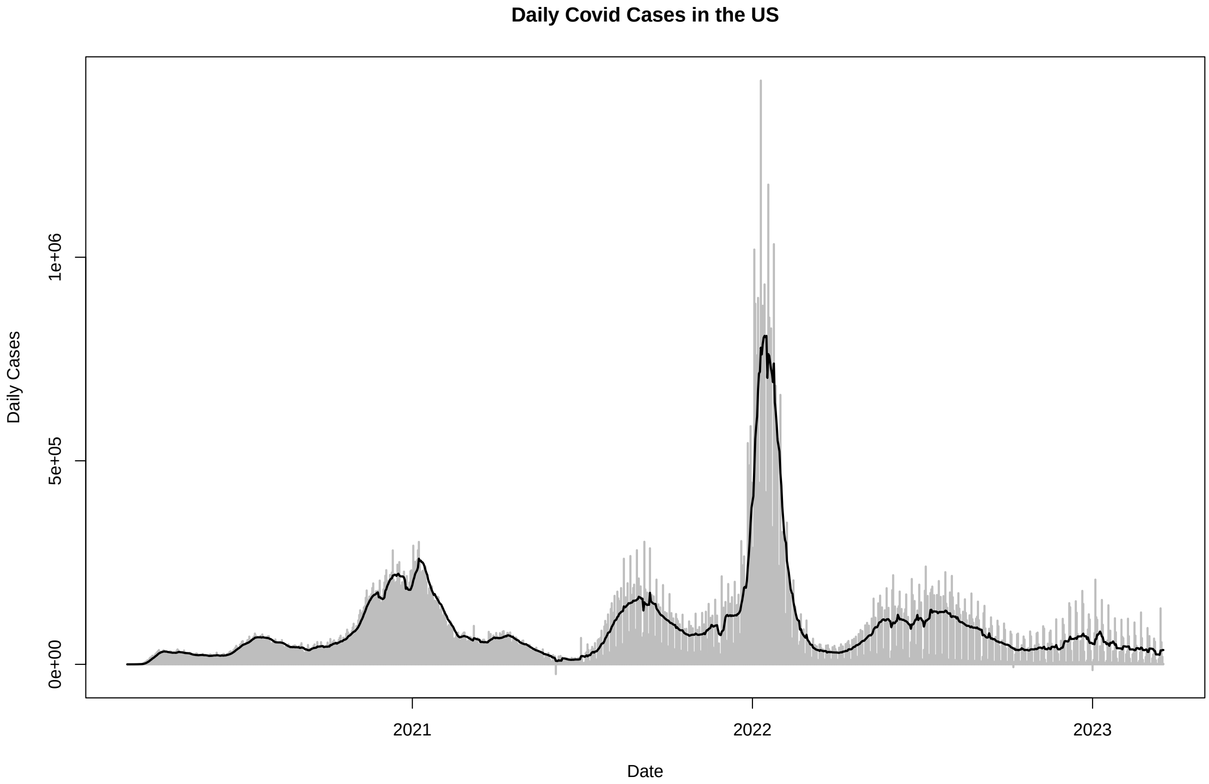
```
# Get Data ----------
us <- data.frame(read_csv(
  "~/Documents/UC Davis/Courses/STA 141A/Proj/us.csv",
  col_types = cols(
    date = col_date(format = "%Y-%m-%d"),
    cases = col_integer(),
    deaths = col_integer()
    )
  ))
```

```
# Daily and average values ----------
us["new_cases"] <- stats::filter(
  us$cases,
  c(1, -1),
  method = "convolution",
  sides = 1
  )
```

```r
us["new_deaths"] <- stats::filter(
  us$deaths, c(1, -1),
  method = "convolution",
  sides = 1
  )

us["avg_new_cases"] <- stats::filter(
  us$new_cases,
  rep(1/7, 7),
  method = "convolution",
  sides = 1
  )

us["avg_new_deaths"] <- stats::filter(
  us$new_deaths,
  rep(1/7, 7),
  method = "convolution",
  sides = 1
  )
```

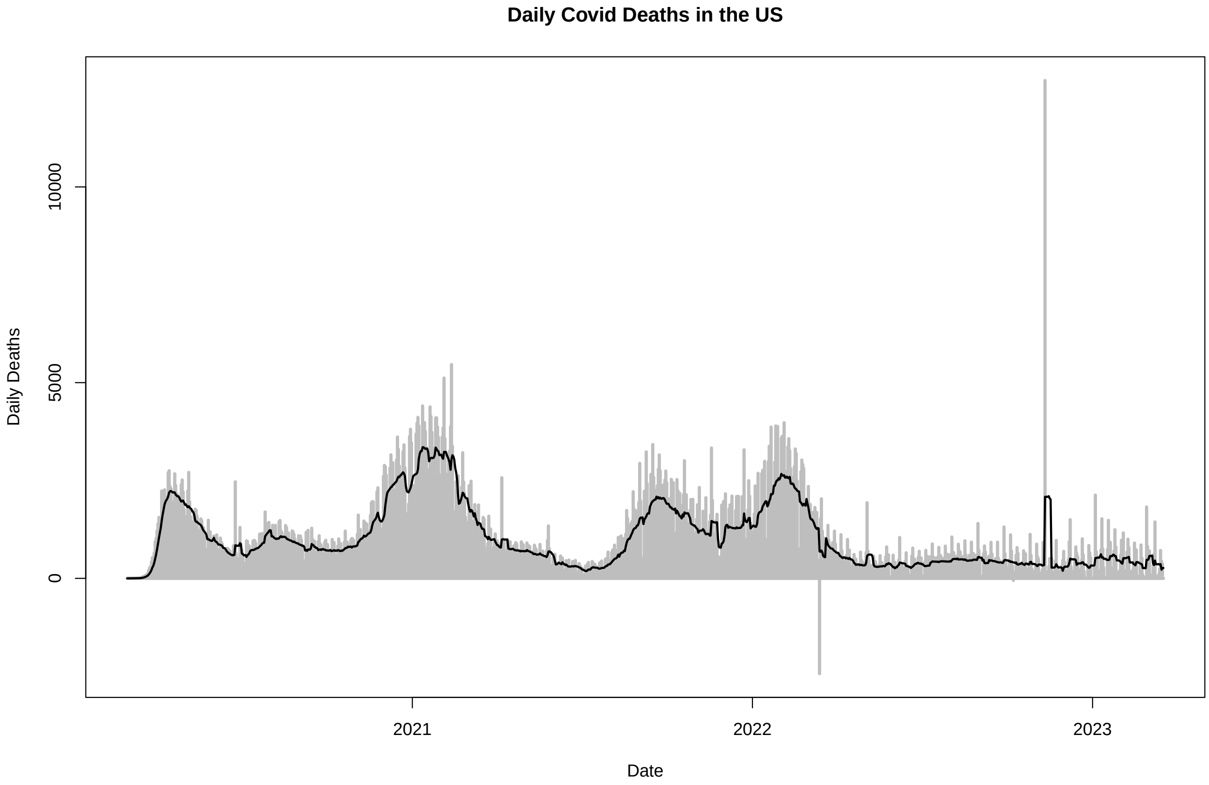
b)

```r
# Daily cases plot ----------
plot(
  new_cases ~ date,
  data = us,
  type = "h",
  col = "gray",
  subset = date >= "2020-03-01",
  lwd = 2,
  ylab = "Daily Cases",
  xlab = "Date",
  main = "Daily Covid Cases in the US"
  )
lines(avg_new_cases ~ date, data = us, subset = date >= "2020-03-01", lwd = 2)
```



Upon analyzing the plot, we can observe two more pronounced spikes in the number of daily cases: one in the transition from 2020 to 2021 and another from 2021 to 2022. The most significant spike occurred at the beginning of 2022, which could be attributed to factors such as the emergence of new variants, relaxation of preventive measures, and increased social interactions during the holiday season.

```r
# Daily death plot ----------
plot(
  new_deaths ~ date,
  data = us,
  type = "h",
  col = "gray",
  subset = date >= "2020-03-01",
  lwd = 3,
  ylab = "Daily Deaths",
  xlab = "Date",
  main = "Daily Covid Deaths in the US"
  )
lines(avg_new_deaths ~ date, data = us, subset = date >= "2020-03-01", lwd = 2)
```



When examining the plot of daily deaths, we notice a relatively constant trend, punctuated by three distinct peaks. The first peak occurred at the start of the pandemic, which is understandable as the most vulnerable individuals were affected before appropriate measures were implemented. The second and third peaks coincided with the first and second daily case peaks, respectively, which is expected given the correlation between the number of cases and the resulting fatalities.

Two significant anomalies warrant further investigation: a considerable negative value in early 2022 and an enormous positive value in late 2022. Since it is not possible to have negative deaths, and there is no indication of increased cases around the largely positive value, these anomalies likely resulted from sampling errors or data corrections.

Exercise II

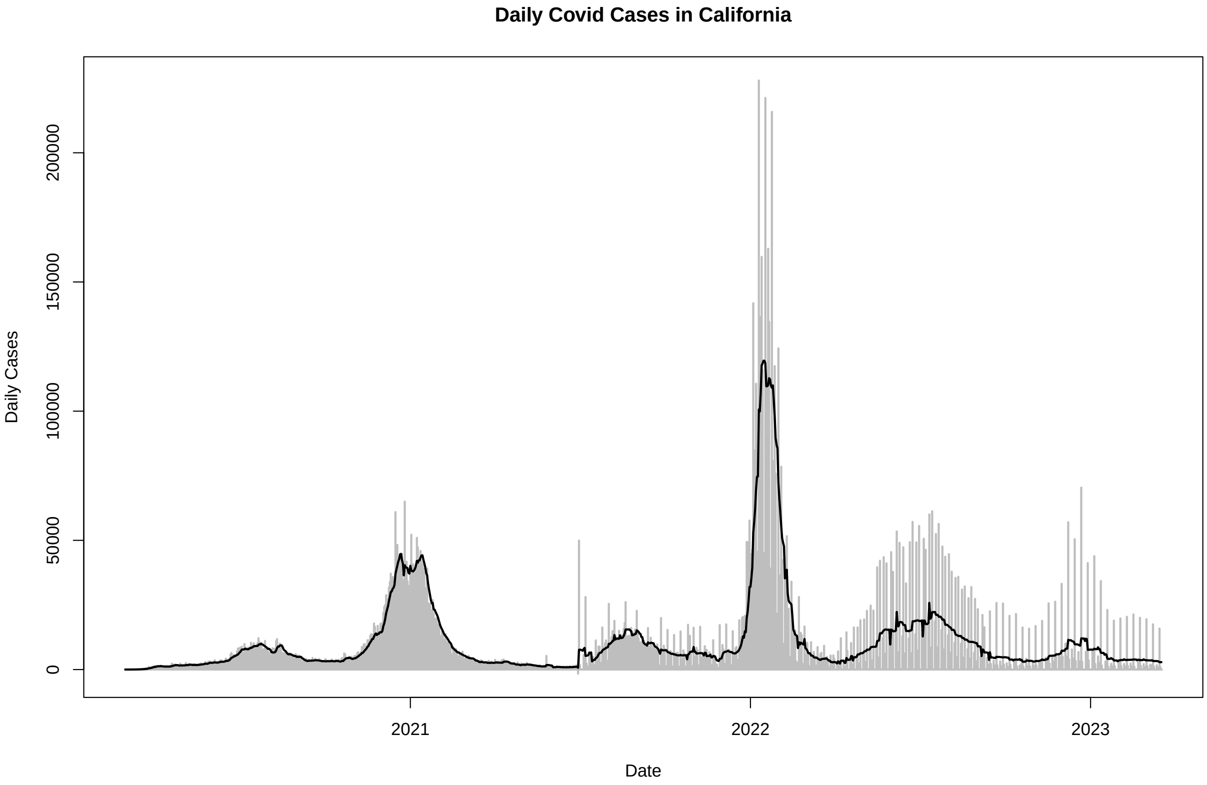
a)

```
# Get data ----------
us_states <- data.frame(read_csv(
  "~/Documents/UC Davis/Courses/STA 141A/Proj/us-states.csv",
  col_types = cols(
    date = col_date(format = "%Y-%m-%d"),
    cases = col_integer(),
    deaths = col_integer()
    )
  ))
```

b)

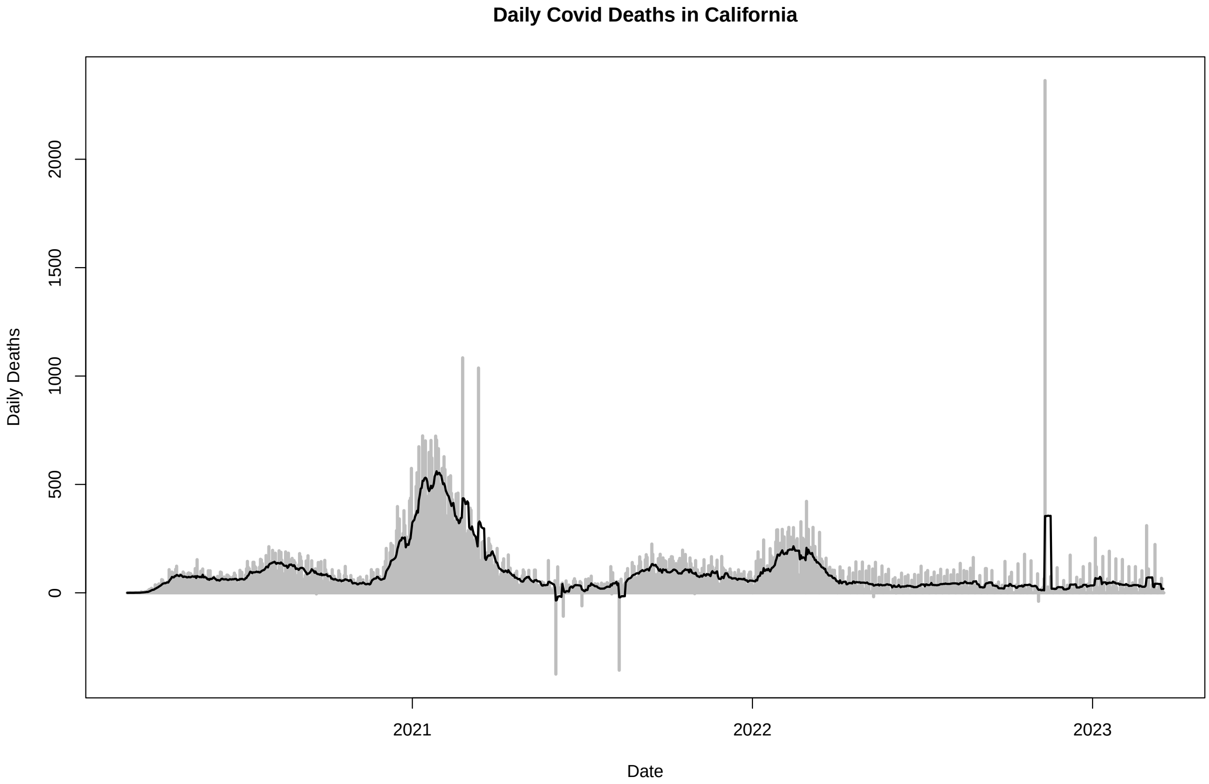
```r
# Californian cases and deaths ----------
california <- subset(us_states, state == "California")

# Daily and average values ----------
california["new_cases"] <- stats::filter(
  california$cases,
  c(1, -1),
  method = "convolution",
  sides = 1
  )

california["new_deaths"] <- stats::filter(
  california$deaths,
  c(1, -1),
  method = "convolution",
  sides = 1
  )

california["avg_new_cases"] <- stats::filter(
  california$new_cases,
  rep(1/7, 7),
  method = "convolution",
  sides = 1
  )

california["avg_new_deaths"] <- stats::filter(
  california$new_deaths,
  rep(1/7, 7),
  method = "convolution",
  sides = 1
  )
```

```r
# Daily cases plot ----------
plot(
  new_cases ~ date,
  data = california,
  type = "h",
  col = "gray",
  subset = date >= "2020-03-01",
  lwd = 2,
  ylab = "Daily Cases",
  xlab = "Date",
  main = "Daily Covid Cases in California"
  )
lines(
  avg_new_cases ~ date,
  data = california,
  subset = date >= "2020-03-01",
  lwd = 2
  )
```



The progression of Covid-19 cases in California generally mirrors the national trend, albeit with fewer peaks. The greater difference between daily cases and the 7-day average in California compared to the overall US plot suggests higher variability in the data.

```
# Daily death plot ----------
plot(
  new_deaths ~ date,
  data = california,
  type = "h",
  col = "gray",
  subset = date >= "2020-03-01",
  lwd = 3,
  ylab = "Daily Deaths",
  xlab = "Date",
  main = "Daily Covid Deaths in California"
  )
lines(
  avg_new_deaths ~ date,
  data = california,
  subset = date >= "2020-03-01",
  lwd = 2
  )
```



Interestingly, California experienced fewer deaths at the start of the pandemic and during the second peak compared to the national trend. This discrepancy may be attributed to a higher vaccination rate or better adherence to public health guidelines in the state. Similar to the national data, we observe large daily and negative death value anomalies in California, likely due to the reasons previously discussed.

Exercise III

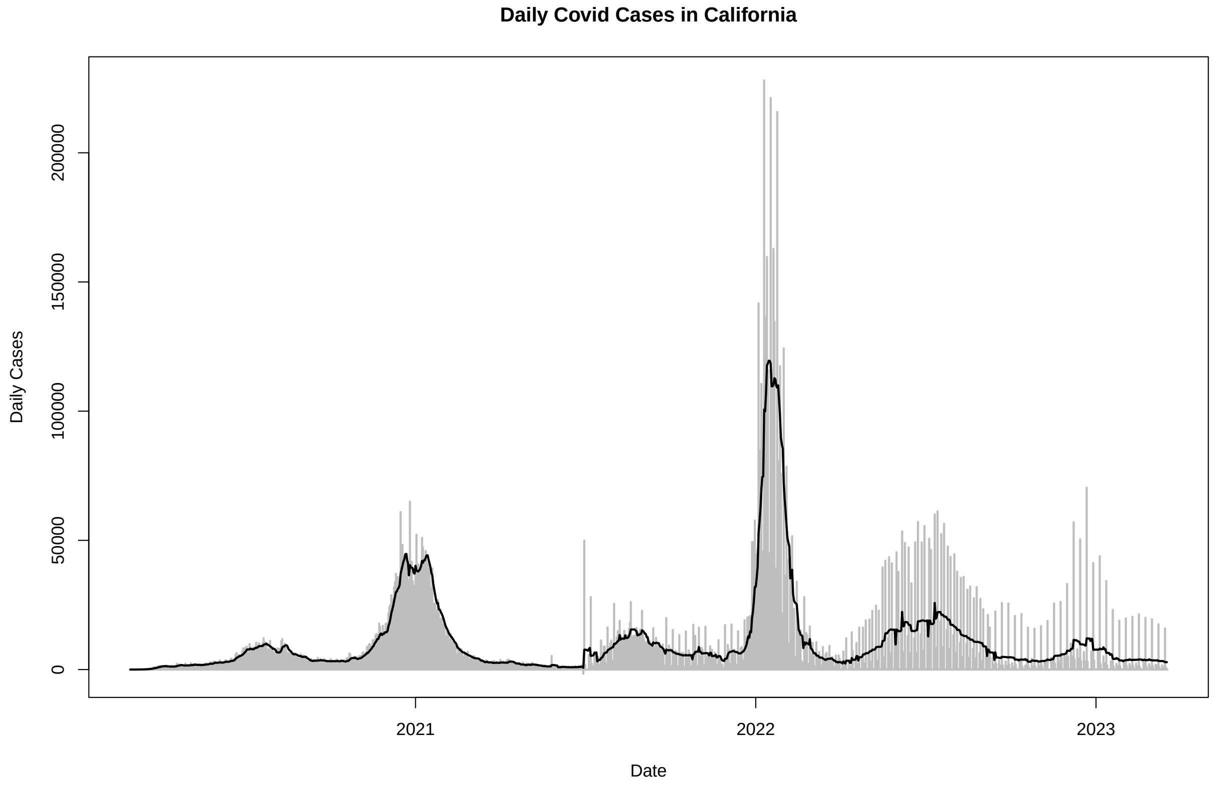
a)

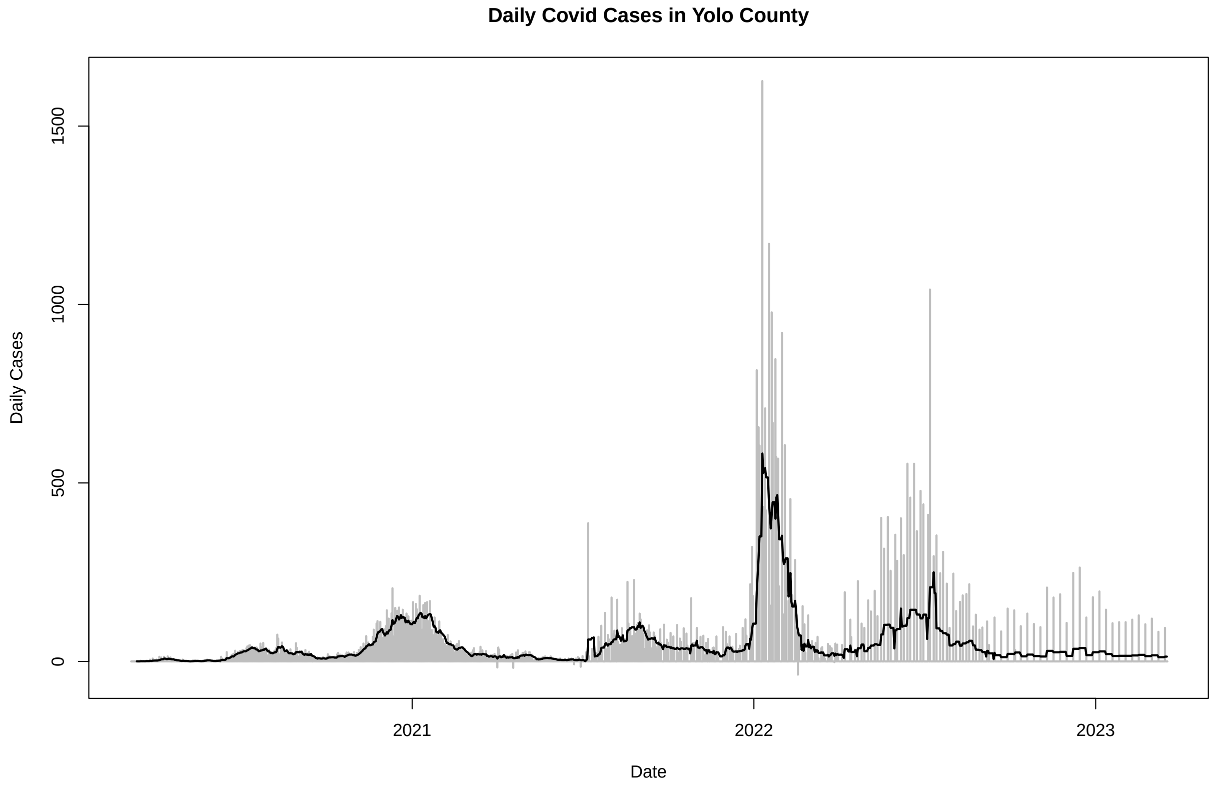
```
# Get data and tidy ---------
us_counties <- data.frame(read_csv(
  "~/Documents/UC Davis/Courses/STA 141A/Proj/us-counties.csv",
  col_types = cols(
    date = col_date(format = "%Y-%m-%d"),
    cases = col_integer(),
    deaths = col_integer())
  ))

us_counties_2022 <- data.frame(read_csv(
  "~/Documents/UC Davis/Courses/STA 141A/Proj/us-counties-2022.csv",
  col_types = cols(
    date = col_date(format = "%Y-%m-%d"),
    cases = col_integer(),
    deaths = col_integer())
  ))
```

```r
us_counties_2023 <- data.frame(read_csv(
  "~/Documents/UC Davis/Courses/STA 141A/Proj/us-counties-2023.csv",
  col_types = cols(
    date = col_date(format = "%Y-%m-%d"),
    cases = col_integer(),
    deaths = col_integer())
  ))

yolo <- subset(us_counties, county == "Yolo")
yolo2022 <- subset(us_counties_2022, county == "Yolo")
yolo2023 <- subset(us_counties_2023, county == "Yolo")

yolo <- yolo %>%
  rbind(yolo2022, yolo2023) %>%
  distinct(date, .keep_all = T)

# Daily and average values ----------
yolo["new_cases"] <- stats::filter(
  yolo$cases,
  c(1, -1),
  method = "convolution",
  sides = 1
  )

yolo["avg_new_cases"] <- stats::filter(
  yolo$new_cases,
  rep(1/7, 7),
  method = "convolution",
  sides = 1
  )
```

b)

```r
# Daily cases plot ----------
par(mfrow=c(2,1))
plot(
  new_cases ~ date,
  data = california,
  type = "h",
  col = "gray",
  subset = date >= "2020-03-01",
  lwd = 2,
  ylab = "Daily Cases",
  xlab = "Date",
  main = "Daily Covid Cases in California"
  )
lines(
  avg_new_cases ~ date,
  data = california,
  subset = date >= "2020-03-01",
  lwd = 2
  )

plot(
  new_cases ~ date,
  data = yolo,
  type = "h",
  col = "gray",
  subset = date >= "2020-03-01",
  lwd = 2,
  ylab = "Daily Cases",
  xlab = "Date",
  main = "Daily Covid Cases in Yolo County"
  )
lines(avg_new_cases ~ date, data = yolo, subset = date >= "2020-03-01", lwd = 2)
```


Daily Covid Cases in California
Daily Cases
0
50000
100000
150000
200000
2021
2022
2023
Date


Daily Covid Cases in Yolo County
Daily Cases
0
500
1000
1500
2021
2022
2023
Date

Question

What do you notice when comparing the plot of daily new cases in Yolo county to the analogous plot for the state of California as a whole? What might explain what you are seeing?

When comparing the plot of daily new cases in Yolo County to the analogous plot for the state of California as a whole, it is evident that Yolo County follows a similar trend as California overall. Both California and Yolo County display a relatively flat curve until the end of 2020, experience a peak in early 2021, and then witness a decline back to a flatter curve in spring 2021. The initial flat curve can likely be attributed to widespread adherence to public health guidelines such as staying home, practicing social distancing, and using face masks. Additionally, the virus's seasonal nature, which makes it more prone to infect and spread during colder months, contributes to the higher peak observed in the winter. The introduction of vaccinations in April in the US also played a significant role in the prolonged flattening of the curve until early 2022.

However, some discrepancies between the two plots emerge in the winter period of late 2021 to early 2022. While California experienced a strong peak, Yolo County witnessed a less pronounced increase in cases. The strong peak in California could be attributed to people becoming less cautious about restrictions as vaccinations were rolling out and a sense of normalcy returned. However, with low fully vaccinated (68%) and booster levels (34%) (The New York times, 2023) coupled with the virus's seasonal potency, California faced its highest daily case rates.

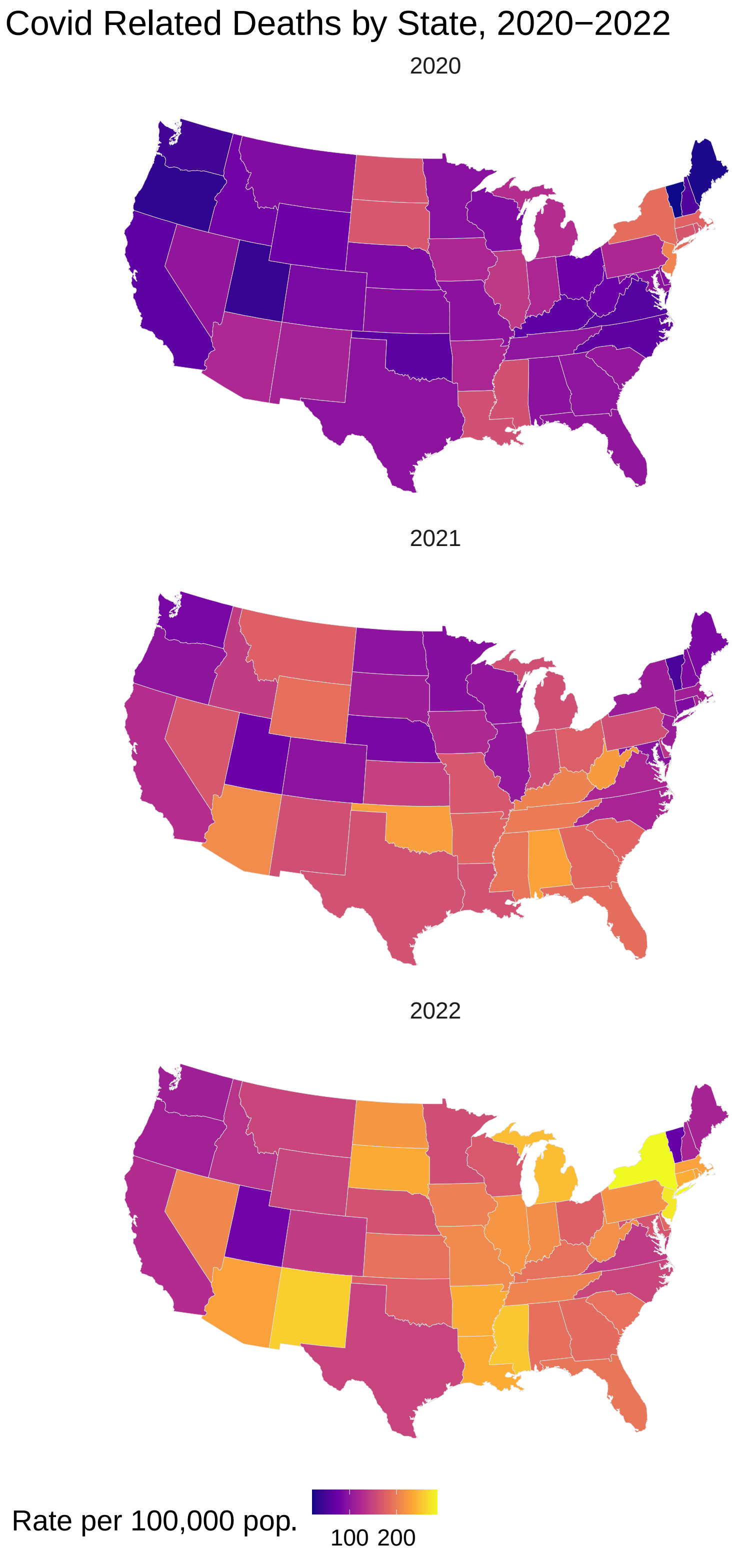
In contrast, Yolo County's less pronounced peak can be explained by its unique demographic and geographic characteristics. With a population of approximately 220,000 (U.S. Census Bureau, 2022), of which around 40,000 are university students (EdData, 2022), the migration patterns of students significantly impact the county's statistics. Two interrelated factors contribute to Yolo County's lower peak: its predominantly rural nature, which facilitates systematic social distancing, and the fact that most students are not originally from Yolo County. As a result, many students left their college towns during the early pandemic and only returned when campuses reopened. Consequently, Yolo County's peak is less pronounced than California's, likely due to the delayed return of many students to the area.

Excercise III: Bonus

By State

```r
# Get and tidy data ----------
population <- read_excel(
  "~/Documents/UC Davis/Courses/STA 141A/Proj/NST-EST2022-POP.xlsx",
  skip = 3
  )
population$state <-  sub('.', '', population$state)

# Yearly state deaths ----------
us_states["year"] <- format(as.Date(us_states$date, format="%Y-%m-%d"),"%Y")

cumulative_deaths <- us_states %>%
  group_by(year, state) %>%
  summarise(deaths = max(deaths))

deaths_2020 <- cumulative_deaths %>%
  filter(year == 2020) %>%
  left_join((population %>% select(state, `2020`))) %>%
  rename(population = `2020`) %>%
  mutate(death_capita = deaths/population*100000)
```

```r
deaths_2021 <- cumulative_deaths %>%
  filter(year == 2021) %>%
  left_join((deaths_2020 %>% select(state, deaths)), by = "state") %>%
  transmute(year = year.x, state = state, deaths = deaths.x - deaths.y) %>%
  left_join((population %>% select(state, `2021`))) %>%
  rename(population = `2021`) %>%
  mutate(death_capita = deaths/population*100000)

deaths_2022 <- cumulative_deaths %>%
  filter(year == 2022) %>%
  left_join((deaths_2021 %>% dplyr::select(state, deaths)), by = "state") %>%
  transmute(year = year, state = state, deaths = deaths.x - deaths.y) %>%
  left_join((population %>% select(state, `2022`))) %>%
  rename(population = `2022`) %>%
  mutate(death_capita = deaths/population*100000)

deaths <- rbind(deaths_2020, deaths_2021, deaths_2022)

# State Map ----------
states <- map_data("state")
deaths$region <- tolower(deaths$state)
deaths_map <- left_join(states, deaths, by = "region")

p0 <- ggplot(
  data = deaths_map,
  mapping = aes(
    x = long,
    y = lat,
    group = group,
    fill = death_capita
    )
  )

p1 <- p0 +
  geom_polygon(color = "gray90", size = 0.05) +
  coord_map(projection = "albers", lat0 = 39, lat1 = 45)

p2 <- p1 + scale_fill_viridis_c(option = "plasma")

p2 +
  theme_map() +
  facet_wrap(~ year, nrow = 3) +
  theme(
    legend.position = "bottom",
    strip.background = element_blank(),
    text = element_text(size = 20)
    ) +
  labs(
    fill = "Rate per 100,000 pop.",
    title = "Covid Related Deaths by State, 2020-2022"
    )
```

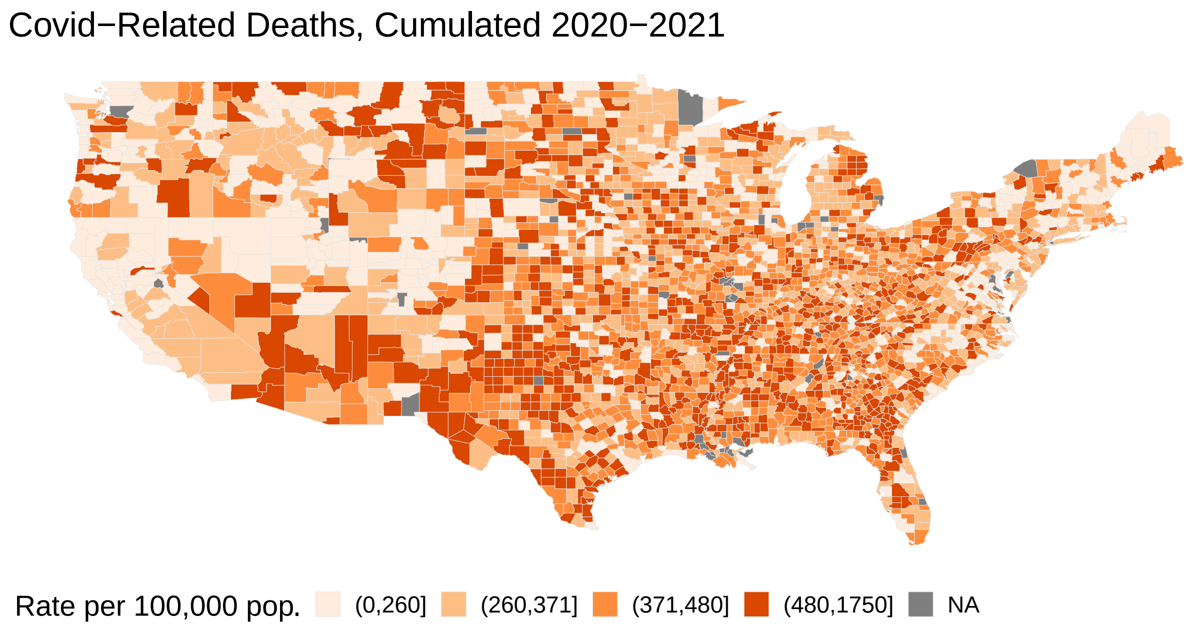


The map displays the progression of Covid-related deaths from 2020 to 2022 across the United States, highlighting that each state experienced an increase in deaths. However, some states fared better than others. For example, Utah maintained a relatively low death rate compared to the rest of the country, whereas states like New York and New Jersey saw significant increases. A probable cause for these disparities could be attributed to factors such as population density, public health infrastructure, and adherence to preventive measures. Utah, being a more rural state, might have experienced fewer cases and deaths due to natural social distancing, while densely populated states like New York and New Jersey faced greater challenges in controlling the spread of the virus.

By County

```r
# Add county population ----------
county_population <- read_csv(
  "~/Documents/UC Davis/Courses/STA 141A/Proj/co-est2021-alldata.csv"
  )

pops <- county_population %>%
  select(STATE, COUNTY, STNAME, CTYNAME, POPESTIMATE2021) %>%
  mutate(fips = paste(STATE, COUNTY, sep="")) %>%
  select(fips, POPESTIMATE2021)

county_data <- us_counties %>%
  group_by(fips, county) %>%
  summarise(cumulated_deaths = max(deaths)) %>%
  left_join(pops, by = "fips") %>%
  mutate(death_capita = cumulated_deaths/POPESTIMATE2021*100000) %>%
  mutate(subregion = tolower(county))

# Create Factor ----------
county_data$interval <- as.factor(cut(
  county_data$death_capita,
  quantile(county_data$death_capita, na.rm = T)))

fltr <- levels(county_data$interval)=="(480,1.75e+03]"
levels(county_data$interval)[fltr] <- "(480,1750]"

# Map
county_map <- map_data("county")
county_full <- left_join(county_map, county_data,  by = "subregion")

death_p <- ggplot(
  data = county_full,
  mapping = aes(
    x = long,
    y = lat,
    fill = interval,
    group = group
    )
  )

death_p1 <- death_p +
  geom_polygon(color = "gray90", size = 0.05) +
  coord_equal()

death_p2 <- death_p1 +
  scale_fill_manual(values = RColorBrewer::brewer.pal(n = 4, name = "Oranges"))
```

```
death_p2 +
  labs(
    title = "Covid-Related Deaths, Cumulated 2020-2021",
    fill = "Rate per 100,000 pop.") +
  theme_map() +
  theme(
    legend.position = "bottom",
    text = element_text(size = 20)
    )
```



This map tells a similar story to the state-level map but provides a more granular perspective on the death rates in each county. While it is difficult to discern an overarching pattern, certain regions, such as the west coast, the area stretching from California to Utah, and the northeast coast, appear to have lower death rates than the rest of the country. The explanation for these patterns may be similar to those mentioned for the state-level map, involving factors like population density, public health measures, and individual behaviors.

However, this map also highlights the importance of understanding the relative contribution of a county to the state's overall death rate. For instance, some northern states appear to have relatively low death rates at the state level, but at the county level, they exhibit darker patches, indicating higher death rates. Without the context provided by the state-level map, one might incorrectly assume that these states have higher death rates compared to the rest of the country, while in reality, their death rates are relatively low.

Conclusion

Analyzing daily cases and death rates at the national, state, and county levels offers valuable insights into the progression of the Covid-19 pandemic in the United States. While certain trends and patterns are consistent across different geographic scales, it is crucial to consider the unique factors and circumstances in each area to gain a comprehensive understanding of the pandemic's impact. Furthermore, recognizing the limitations and potential anomalies in the data is essential for drawing accurate conclusions and informing effective public health strategies.

Sources

EdData. (2022). County Summary - Yolo, Available Online: https://www.ed-data.org/county/yolo/ [Accessed 22 March 2023]

The New York times. (2023). Covid in the U.S.: Latest Maps, Case and Death Counts, Available Online: https://www.nytimes.com/interactive/2021/us/covid-cases.html [Accessed 22 March 2023]

U.S. Census Bureau. (2022). QuickFacts: Yolo County, California, Available Online: https://www.census.gov/quickfacts/yolocountycalifornia [Accessed 22 March 2023]