1. **Choice of Data**

We decided to employ three quantitative variables to create a plot displaying a multivariate time series that retains relational information among the variables. In that, we wanted to create a time series plot showing the relationship between weekly new COVID cases and weekly new vaccination doses in the United States.

There are certain shortcomings of existing plots:

1. Some of the time series graphs retain new cases only. They do not show any relationship between new cases and other variables.
2. Some other plots use facets to display new cases and vaccination doses, but they also need to tell the relationship between the variables. Also, some of these plots show separate lines or dots where relational elements are not prominent.
3. Other plots lack legends or essential aesthetic elements that could make the visualization more appealing to the audience.

We will incorporate a time series plot with geom_line and geom_rect visualization techniques to address these shortcomings. The geom_rect function could present information on vaccination doses without having to draw it separately. Therefore, by doing so, we could conclude the correlation between vaccination doses and new cases. Moreover, the geom_rect combines perfectly with scale_fill_gradient and legend elements, making the visualization aesthetically pleasant to the audience. In addition, the line will show a general trend in new cases, and we can conclude how vaccination help reduce the increase in new cases.

We will use data sets from CDC, where data regarding COVID cases and vaccination doses are presumed to be accurate.

## 2. Data Set Description

We felt that there was a need to explore the relationship between new cases and new vaccination doses since we were wondering if vaccination did help reduce the increase in cases.

Our data set covers COVID new cases and vaccinations between the start of COVID, which is around early 2020, and the present. It was retrieved from CDC's Data Tracker (*https://covid.cdc.gov/covid-data-tracker/#trends_weeklycases_totalvaccinesadministered_00*).

The data consists of three quantitative variables:

1. time (date)
2. new COVID cases per week
3. new vaccination doses per week

The data covers COVID new cases and vaccinations between the start of COVID, which is around early 2020, and the present. For ease of interpretation, we transformed our variables (new cases per week and new vaccination doses per week) into millions. New cases per week range from 0 to 6 million people, whereas new vaccination doses range from 0 to 23 million people. We would like to note that there is a large number of dates having 0 new vaccination records at the beginning of COVID.

## Cleaning

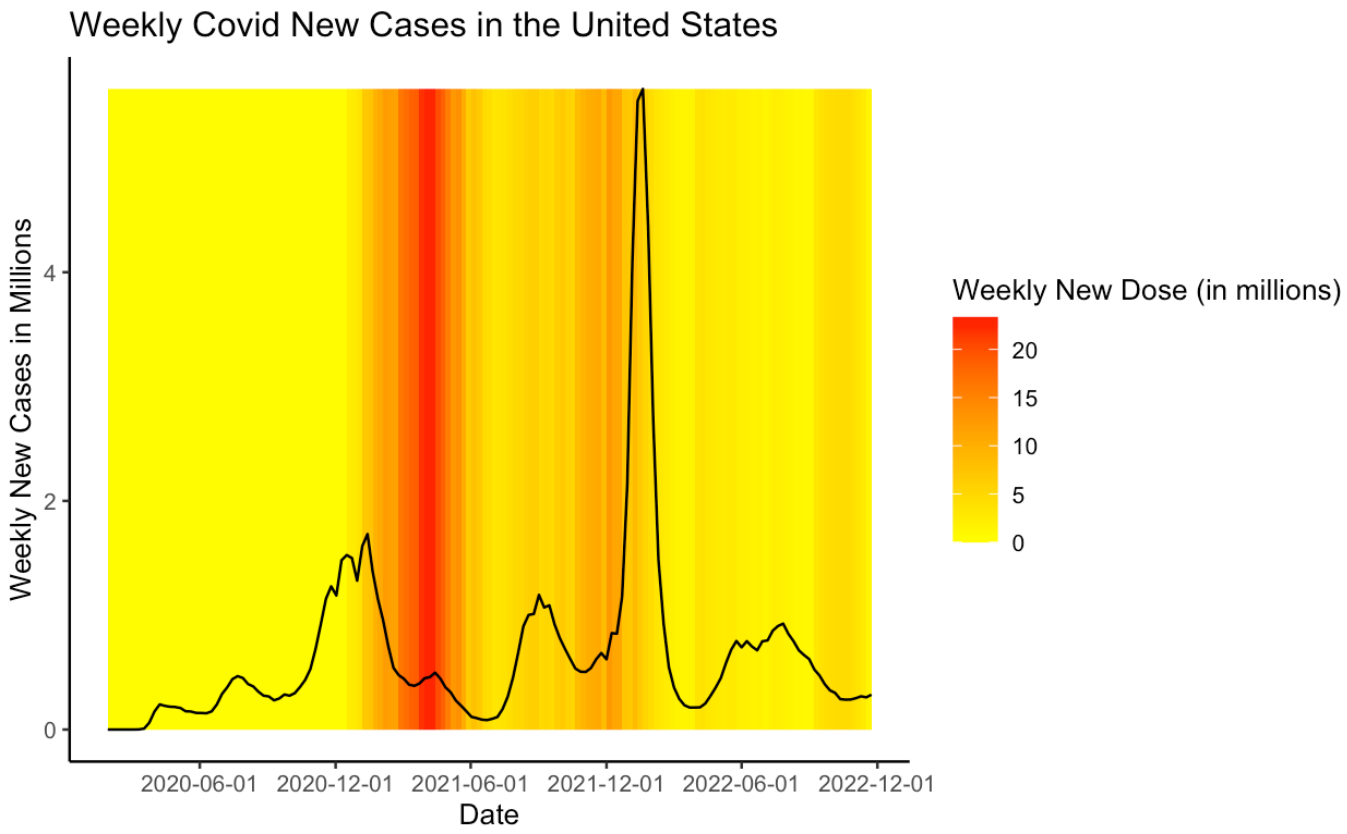
```{r}
library(ggplot2)
library(dplyr)
data <- read.csv("/Users/zach0422/Desktop/covid.csv")

data$Date <- gsub('  ',' ', data$Date)
data$Date <- gsub('Jan','01', data$Date)
data$Date <- gsub('Feb','02', data$Date)
data$Date <- gsub('Mar','03', data$Date)
data$Date <- gsub('Apr','04', data$Date)
data$Date <- gsub('May','05', data$Date)
data$Date <- gsub('Jun','06', data$Date)
data$Date <- gsub('Jul','07', data$Date)
data$Date <- gsub('Aug','08', data$Date)
data$Date <- gsub('Sep','09', data$Date)
data$Date <- gsub('Oct','10', data$Date)
data$Date <- gsub('Nov','11', data$Date)
data$Date <- gsub('Dec','12', data$Date)

data$Date <- gsub(' 1 ',' 01 ', data$Date)
data$Date <- gsub(' 2 ',' 02 ', data$Date)
data$Date <- gsub(' 3 ',' 03 ', data$Date)
data$Date <- gsub(' 4 ',' 04 ', data$Date)
data$Date <- gsub(' 5 ',' 05 ', data$Date)
data$Date <- gsub(' 6 ',' 06 ', data$Date)
data$Date <- gsub(' 7 ',' 07 ', data$Date)
data$Date <- gsub(' 8 ',' 08 ', data$Date)
data$Date <- gsub(' 9 ',' 09 ', data$Date)
data$Date <- gsub(' ','-', data$Date)
```

```
data$Date <- as.Date(data$Date, '%m-%d-%Y')
data <- data[order(data$Date),]
rownames(data) = seq(length=nrow(data))
data[data$Total.Doses.Administered=='N/A','Total.Doses.Administered'] <- 0
data$Total.Doses.Administered <- as.numeric(data$Total.Doses.Administered)
data$new_dose <- c(0, diff(data$Total.Doses.Administered))/1000000
data$Weekly.Cases <- data$Weekly.Cases/1000000
```

### 3. Plot




```
## Visualization
```{r}
my_theme <- theme_bw() +
  theme(axis.text = element_text(size = 10),
        axis.title = element_text(size = 10),
        legend.text = element_text(size = 10),
        legend.title = element_text(size = 12))

plot <- ggplot(data=data, aes(x = Date, y = Weekly.Cases)) +
        geom_rect(aes(xmin = lag(Date),xmax = Date,
                      ymin = 0, ymax = max(Weekly.Cases),
                      fill = new_dose)) +
        geom_line(aes(x = Date, y = Weekly.Cases)) +
        scale_fill_gradient('Weekly New Dose (in millions)',
                            low = 'yellow',
                            high ='red') +
        scale_x_date(date_breaks = '6 months') +
        labs(x = 'Date', y = 'Weekly New Cases in Millions',
             title = 'Weekly Covid New Cases in the United States') +
    theme_bw() +
theme(axis.line = element_line(colour = "black"),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank())

plot
```
```