analysis

```
1+1
```

```
[1] 2
```

```
# Load required libraries
library(ggplot2)
library(tidyverse)
```

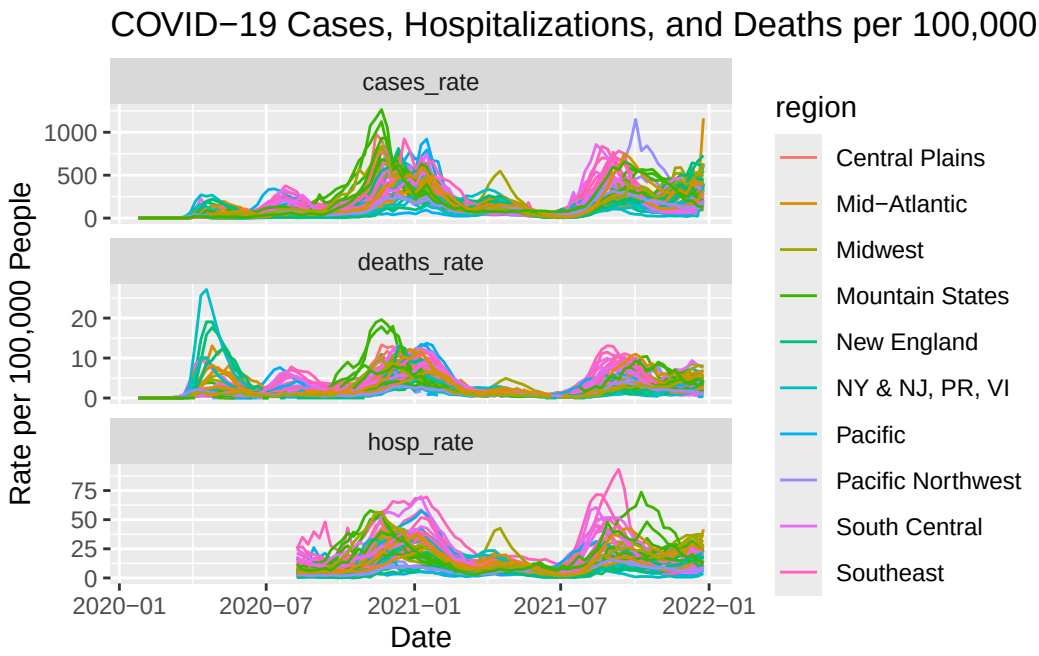
```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v lubridate 1.9.3     v tibble    3.2.1
v purrr     1.0.2     v tidyr     1.3.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(lubridate)
library(ggrepel)

# Load the final dataset
dat <- readRDS("../data/dat.rds")
```

Figure 1, Question 10, a trend plot for cases, hospitalizations and deaths.

```r
library(tidyr)
f1 <- dat |>
  mutate(cases_rate = cases/population * 100000,
         hosp_rate = (hosp / population) * 100000,
         deaths_rate = (deaths / population) * 100000)|>
  pivot_longer(
    cols = c(cases_rate, hosp_rate, deaths_rate),
    names_to = "metric",
    values_to = "rate"
  )|>
  filter(!is.na(metric)) |>
  ggplot(aes(x=date,y=rate, color = region_name, group = state))+
  geom_line()+
  facet_wrap(~metric,ncol=1, scales = "free_y")+
  labs(title = "COVID-19 Cases, Hospitalizations, and Deaths per 100,000 People",
       x = "Date",
       y = "Rate per 100,000 People",
       color = "region")
f1
```



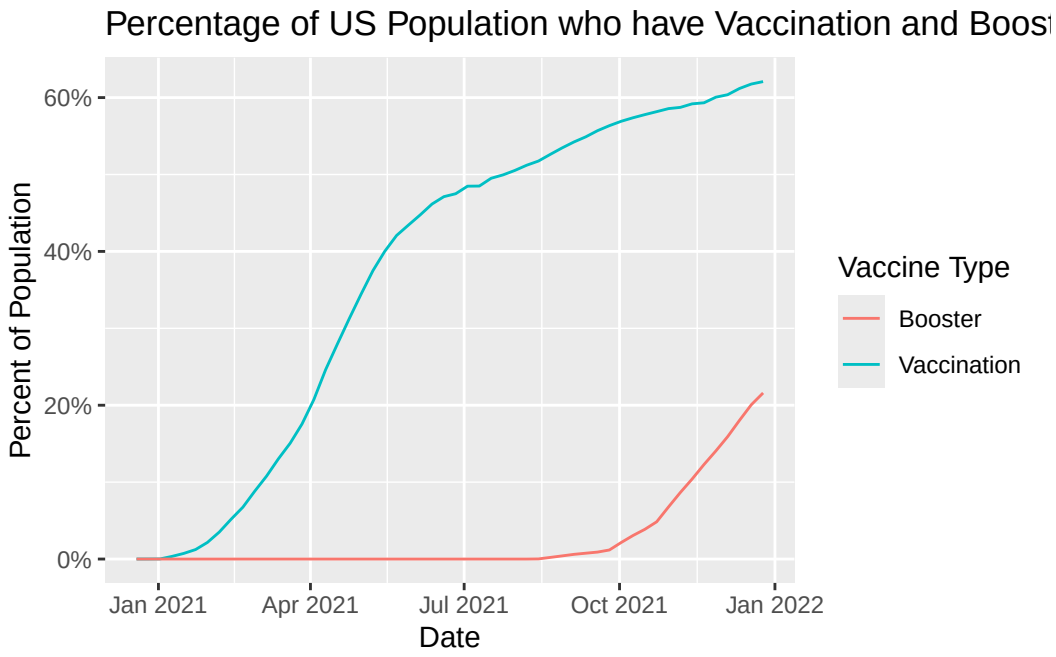
```
ggsave("../figs/figure1.png")
```

```
Saving 5.5 x 3.5 in image
```

Figure 2, Question 11, percentages of the US population who have vaccination by date/.

```
f2 <- dat |>
  drop_na() |>
  group_by(date) |>
  summarize(total_first_dose = sum(as.numeric(series_complete_cumulative_week), na.rm = TRUE
            total_booster = sum(as.numeric(booster_cumulative_week), na.rm = TRUE),
            total_population = sum(population, na.rm = TRUE)) |>
  mutate(percentages_dose = total_first_dose / total_population * 100,
         percentages_booster = total_booster / total_population * 100)|>
  ggplot(aes(x = date)) +
  geom_line(aes(y = percentages_dose, color = "Vaccination")) +
  geom_line(aes(y = percentages_booster, color = "Booster")) +
  labs(title = "Percentage of US Population who have Vaccination and Booster",
       x = "Date", y = "Percent of Population",
       color = "Vaccine Type") +
  scale_y_continuous(labels = scales::percent_format(scale = 1))

f2
```



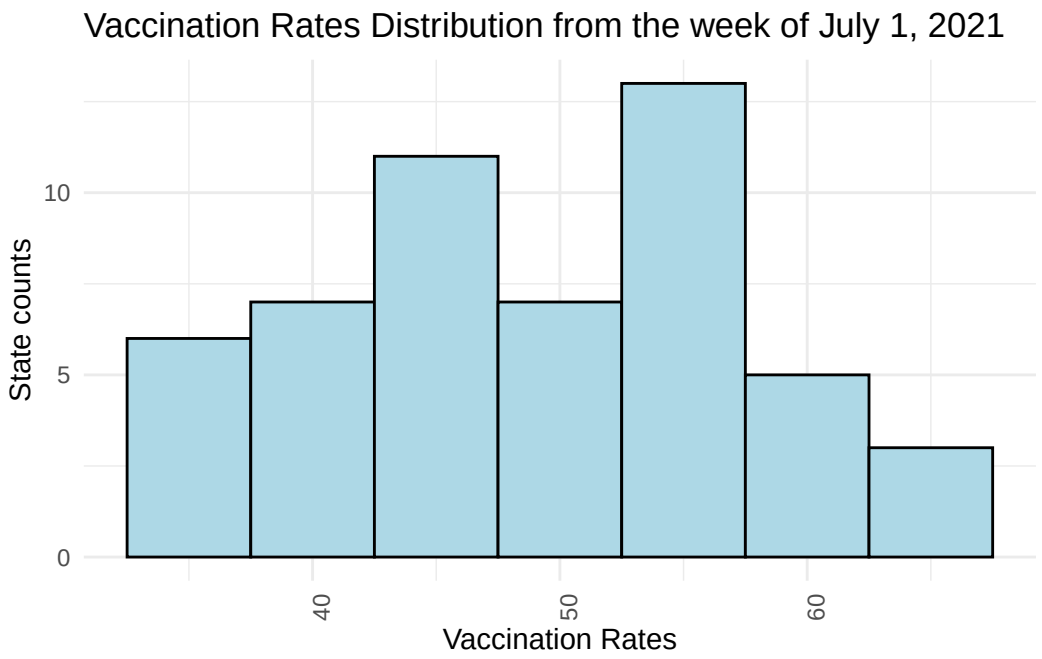

```
ggsave("../figs/figure2.png", plot = f2)
```

```
Saving 5.5 x 3.5 in image
```

˜ # Figure 3, Question 12, describe the distribution of vaccination rates on July 1, 2021.

```
f3<- dat |>
  group_by(state)|>
  filter(mmwr_week == epiweek(as.Date("2021-07-01"))
         & mmwr_year == epiyear(as.Date("2021-07-01")))|># Filter data for July 1, 2021
  summarize(vaccination_rate = (series_complete_cumulative_week / population) * 100)|>
  ggplot(aes(x = vaccination_rate)) +
  geom_histogram(binwidth = 5, fill = "lightblue", color = "black") +
  labs(title = "Vaccination Rates Distribution from the week of July 1, 2021",
       x = "Vaccination Rates",
       y = "State counts") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

f3
```




```
ggsave("../figs/figure3.png", plot = f3)
```

```
Saving 5.5 x 3.5 in image
```

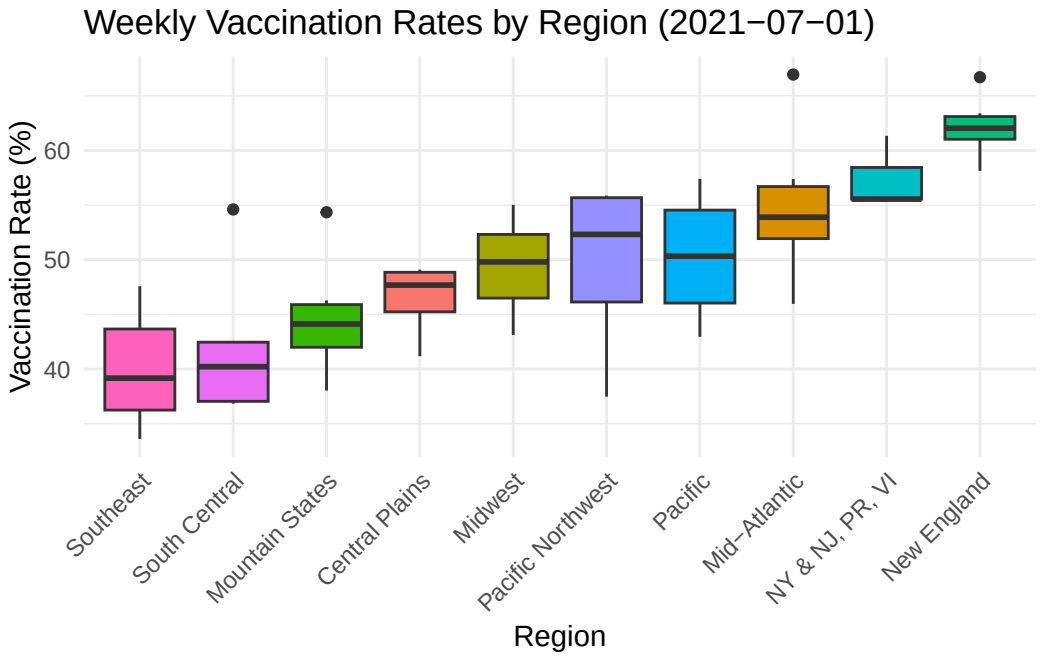
The vaccination histogram shows a bimodal distribution (two peaks), indicating that there are two clusters of states, One clustering around 45% and the other clustering around 55% in the week of July 1, 2021. Also, the distribution is a bit right-skewed, which indicates that there were relatively few states with vaccination rates surpassing 60%. In other words, still many states are under-vaccinated.The range of vaccination rates is moderate, with most states having vaccination rates between 35% and 65%.

Figure 4, Question 13, difference across region

```
f4<- dat |>
  filter(mmwr_year == epiyear(as.Date("2021-07-01")),
         mmwr_week==epiweek(as.Date("2021-07-01"))) |>
  mutate(vaccination_rate = (series_complete_cumulative_week / population) * 100) |>
  ggplot(aes(x = reorder(region_name, vaccination_rate),
             y = vaccination_rate,
```

```
            fill = region_name)) +
  geom_boxplot(show.legend = FALSE) +
  labs(title = "Weekly Vaccination Rates by Region (2021-07-01)",
       x = "Region",
       y = "Vaccination Rate (%)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

f4
```




```
ggsave("../figs/figure4.png", plot = f4)
```

```
Saving 5.5 x 3.5 in image
```

The boxplot shows regional great disparities across regions. For example, New England and NY & NJ, PR, VI have significantly higher vaccination rates with high medianvaccination rate and low variation, while regions like the Southeast and South Central are lagging behind with lower vaccination rates and high variation.

Northeastern regions like New England and Mid-Atlantic have very high vaccination rates, indicating more proactive vaccination policies. On the other hand, Western regions like the Pacific Northwest and Pacific also show relatively high vaccination rates. However, in the

Southern regions, the vaccination rates are noticeably lower than those in other regions, showing relatively passive vaccination actions.

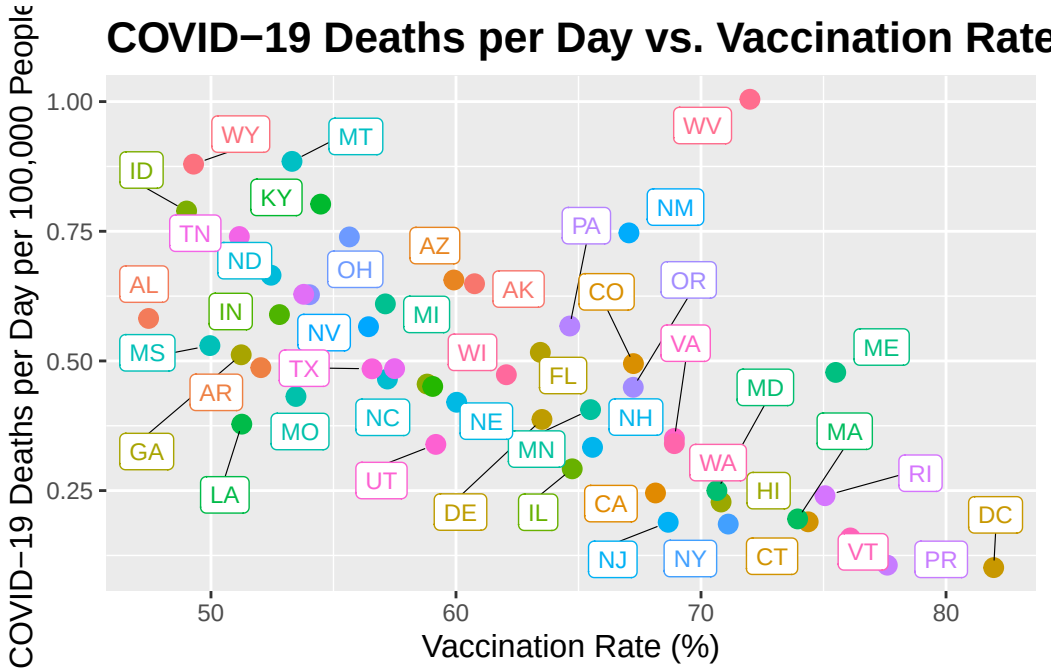
Figure 5, Question 14, identify 1 time period that satisfies the criteria for vaccination rate

```
start_date <- as.Date("2021-09-01")
end_date <- as.Date("2021-12-31")
num_days <- as.numeric(difftime(end_date, start_date, units = "days")) + 1


f5<- dat |>
  drop_na() |>
  filter(date >= start_date & date <= end_date)|>
  group_by(state) |>
  summarise(deaths_per_100k = (sum(deaths) / unique(population)) * 100000/num_days,
            vaccination_rate =
              (max(series_complete_cumulative_week) / unique(population)) * 100 )|>
  ggplot( aes(x = vaccination_rate, y = deaths_per_100k, label = state)) +
  geom_point(aes(color = state), size = 3) +  # Scatter plot points
  geom_label_repel(aes(fill = state, color = state),  # Color text labels
                   fill = "white",
                   size = 3,
                   box.padding = 0.35,
                   point.padding = 0.5,
                   segment.color = "black",
                   segment.size = 0.2
  ) +
  labs(
    title = "COVID-19 Deaths per Day vs. Vaccination Rate by State",
    x = "Vaccination Rate (%)",
    y = "COVID-19 Deaths per Day per 100,000 People"
  ) +
  theme(
    legend.position = "none",
    plot.title = element_text(size = 16, face = "bold"),
    axis.title = element_text(size = 12)
  )

f5
```

```
Warning: ggrepel: 5 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```




```
ggsave("../figs/figure5.png", plot = f5)
```

```
Saving 5.5 x 3.5 in image
```

```
Warning: ggrepel: 5 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```

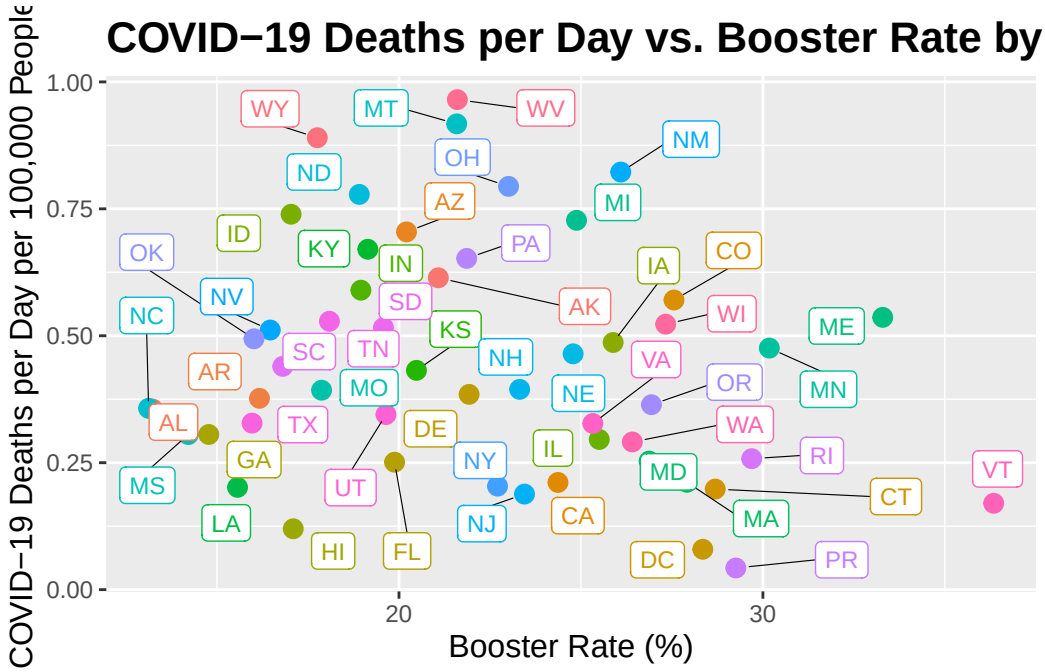
I picked the period from 2021-09-01 to 2021-12-31. There is a negative correlation between the vaccination rate and the number of COVID-19 deaths per day per 100,000 people. This is, the states with higher vaccination rates tend to have lower death rates. For example, States like NY and CT have high vaccination rates (above 70%) and low death rates (below 0.25 deaths per 100,000 people). In contrast, States like WY and ID show low vaccination rates (below 50%) and much higher death rates (higher than 0.75 death per 100,000 people). This plot indicates that vaccines play a significant role in reducing COVID deaths.

Figure 6, Question 15, identify 1 time period that satisfies the criteria for booster rate

```
start_date <- as.Date("2021-10-01")
end_date <- as.Date("2021-12-31")
num_days <- as.numeric(difftime(end_date, start_date, units = "days")) + 1


f6 <- dat |>
  drop_na() |>
  filter(date >= start_date & date <= end_date)|>
  group_by(state) |>
  summarise(deaths_per_100k = (sum(deaths) / unique(population)) * 100000/num_days,
            booster_rate = (max(booster_cumulative_week) / unique(population)) * 100 )|>
  ggplot( aes(x = booster_rate, y = deaths_per_100k, label = state)) +
  geom_point(aes(color = state), size = 3) +
  geom_label_repel(aes(fill = state, color = state),
                   fill = "white",
                   size = 3,
                   box.padding = 0.35,
                   point.padding = 0.5,
                   segment.color = "black",
                   segment.size = 0.2
  ) +
  labs(
    title = "COVID-19 Deaths per Day vs. Booster Rate by State",
    x = "Booster Rate (%)",
    y = "COVID-19 Deaths per Day per 100,000 People"
  ) +
  theme(
    legend.position = "none",
    plot.title = element_text(size = 16, face = "bold"),
    axis.title = element_text(size = 12)
  )

f6
```




```
ggsave("../figs/figure6.png", plot = f6)
```

```
Saving 5.5 x 3.5 in image
```

I picked the period from 2021-10-01 to 2021-12-31. There is not a clear pattern/ linear relationship between the booster rate and the number of COVID-19 deaths per day per 100,000 people. For example, VT and RI stand out as states with high booster rates (above 30%) and very low death rates. On the other hand, WY has a low booster rate and a very high death rate. However, there are some states with relatively high booster rates but also relatively high death rates, such as CO and NM.