COVID-19: Data Exploration

Blake Pappas

Sep. 22, 2022

COVID-19 Data

Load the data using the *COVID19* R package:

```
library(tidyverse)
library(ggplot2)
# install.packages("COVID19")
library(COVID19)

# Load State-Level Data
raw <- covid19(c("US"), level = 2, verbose = FALSE)
```

Question: Pick 4 variables from the data set *raw* and explain why you picked these 4.

Code:

```
COVID19_DATA <- select(raw, confirmed, deaths, date, people_fully_vaccinated)
```

Answer: I have selected the following four variables from the COVID19 data set: *confirmed*, *deaths*, *date*, and *people_fully_vaccinated*. I picked these four variables because I would like to conduct a time-series analyses of rolling full vaccinations by day, deaths by day, full vaccinations by day, and death rate by day. Visualizing full vaccinations by day requires the *date* and *people_fully_vaccinated* variables. Visualizing death rates by day requires the *date*, *confirmed* and *deaths* variables. Visualizing deaths by day requires the *date* and *deaths* variables. Visualizing confirmed cases by day requires the *date* and *confirmed* variables.

Question: Calculate summary statistics for these 4 variables.

Code:

```
summary(COVID19_DATA$confirmed)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##        1    70784   351496   889830  1043231 12169158    16048
```

```
summary(COVID19_DATA$deaths)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0    1171    5035   11779   14912  104277   16048
```

```r
summary(COVID19_DATA$date)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## "2020-01-01" "2021-01-29" "2022-01-12" "2022-01-11" "2022-12-26" "2023-12-15"
```

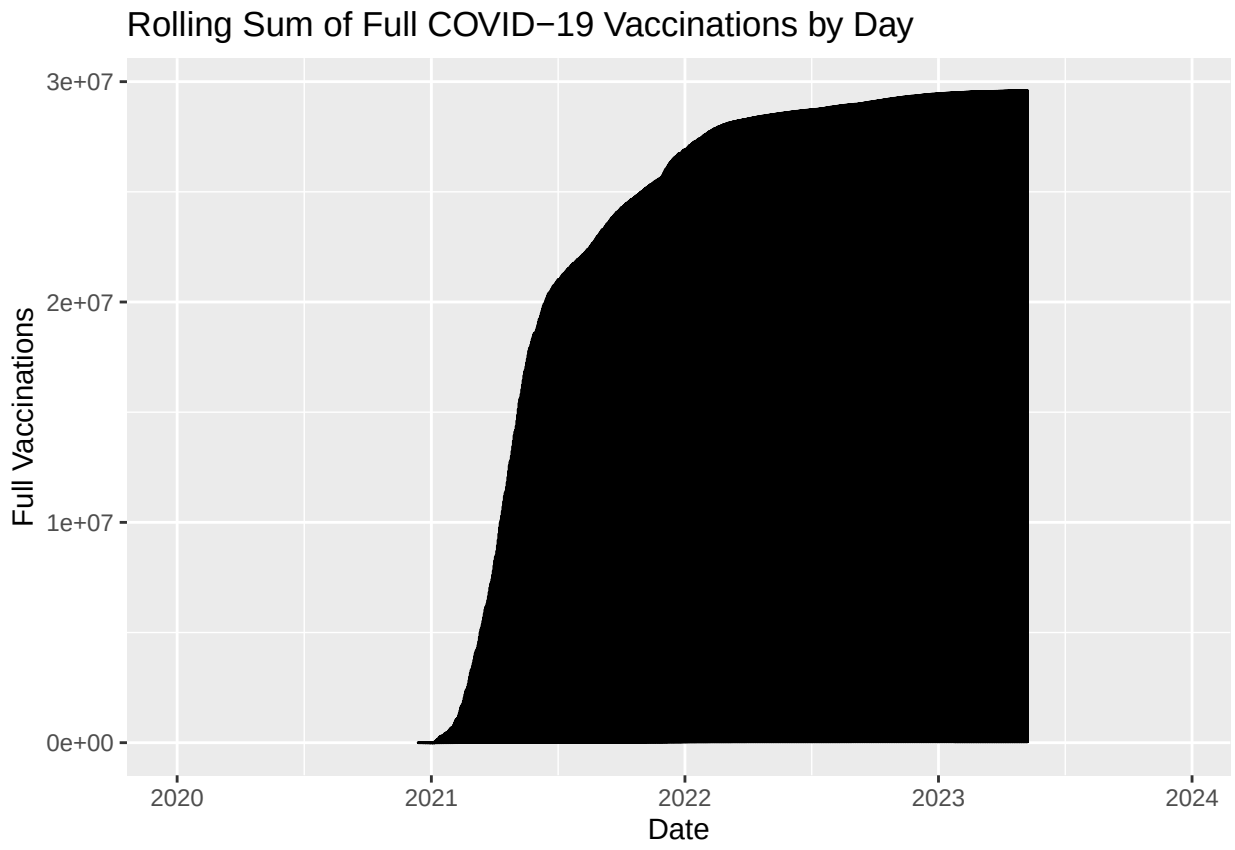
```r
summary(COVID19_DATA$people_fully_vaccinated)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##        0   545888  1777018  3228177  3978254 29588939    28822
```

Question: Visualize these 4 variables.

Code:

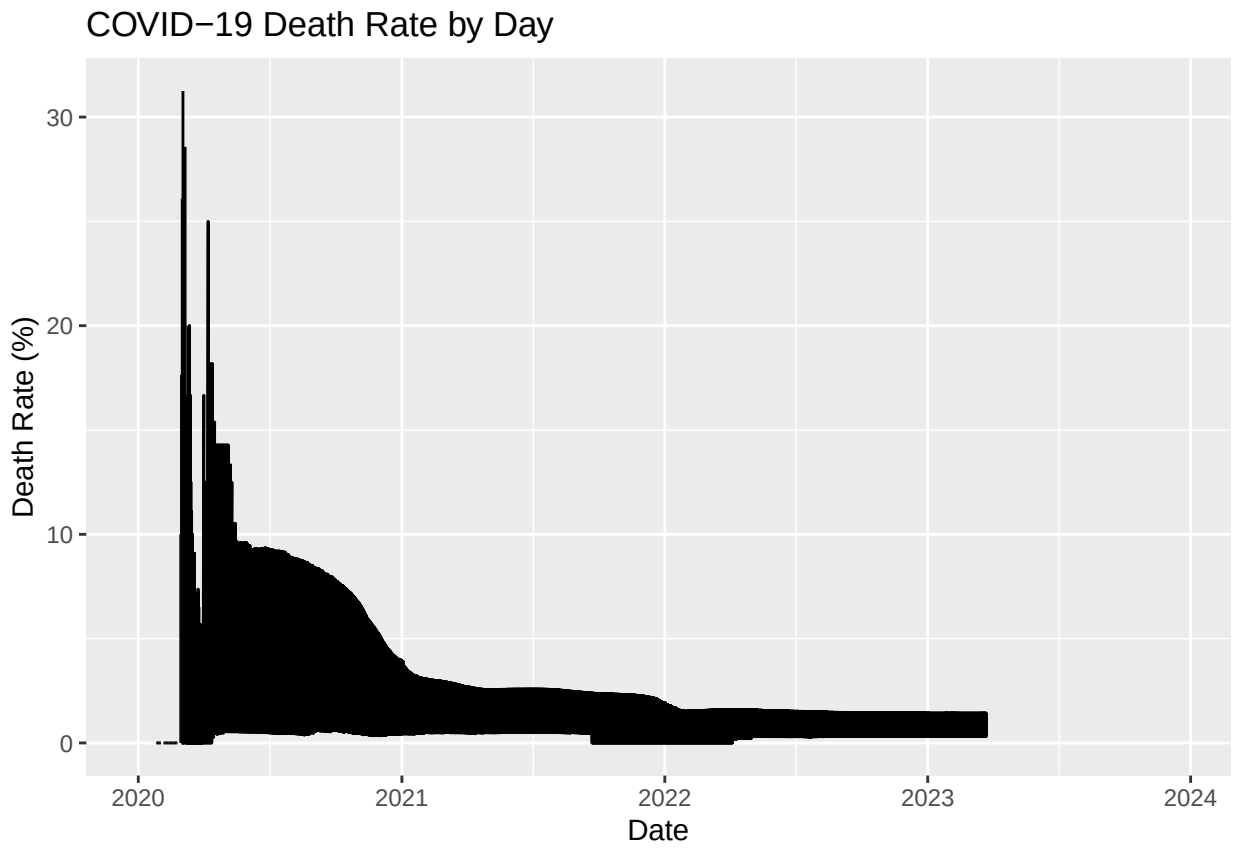
```r
# Rolling Sum of Full Vaccinations by Day
ggplot(data = COVID19_DATA, aes(x = date, y = people_fully_vaccinated)) +
               geom_line() +
  labs(title = 'Rolling Sum of Full COVID-19 Vaccinations by Day', x = 'Date', y = 'Full Vaccinations')
```



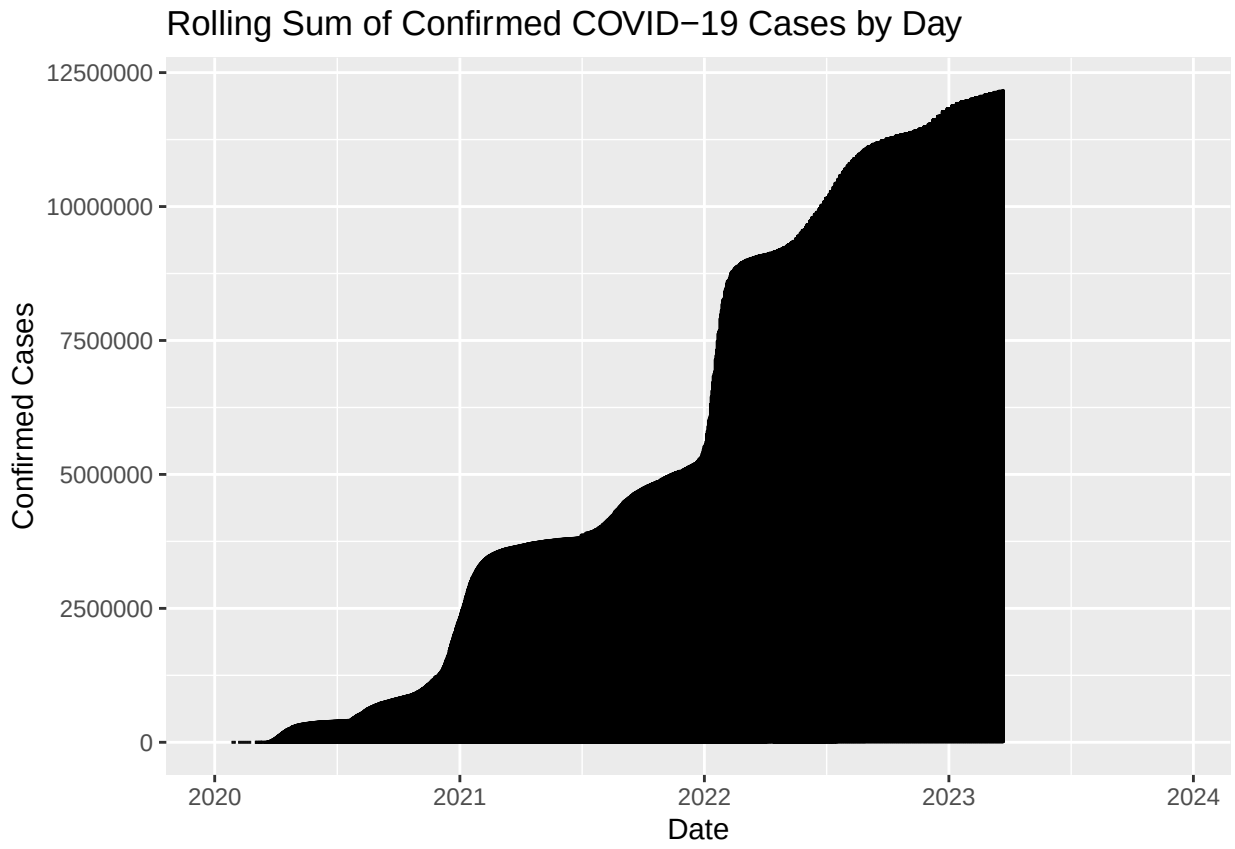

```r
# COVID-19 Death Rate by Day
COVID19_DATA <- COVID19_DATA %>%
  mutate(death_rate = (deaths / confirmed) * 100)

ggplot(data = COVID19_DATA, aes(x = date, y = death_rate)) +
```

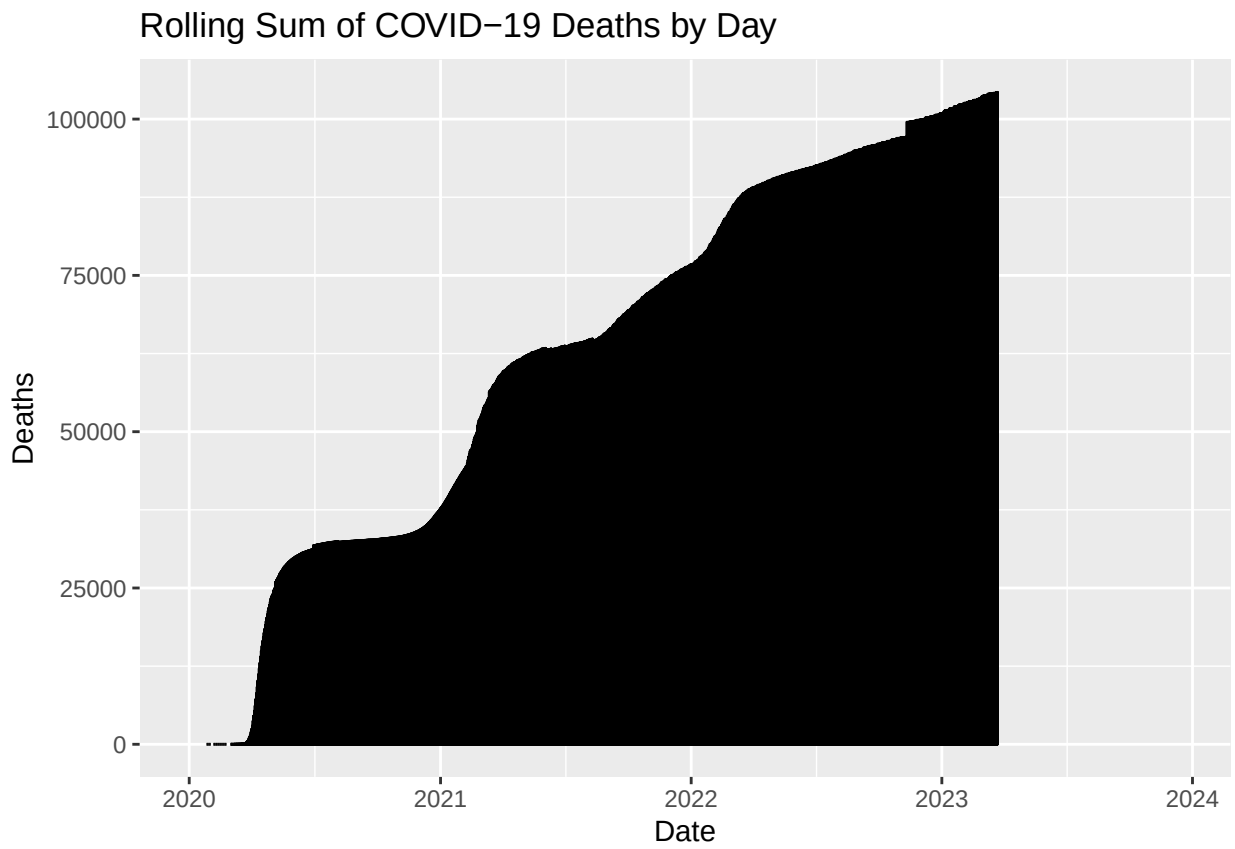
```
              geom_line() +
  labs(title = 'COVID-19 Death Rate by Day', x = 'Date', y = 'Death Rate (%)')
```




```
# Rolling Sum of Confirmed Cases by Day
ggplot(data = COVID19_DATA, aes(x = date, y = confirmed)) +
              geom_line() +
  labs(title = 'Rolling Sum of Confirmed COVID-19 Cases by Day', x = 'Date', y = 'Confirmed Cases')
```




```
# Rolling Sum of Deaths by Day
ggplot(data = COVID19_DATA, aes(x = date, y = deaths)) +
               geom_line() +
  labs(title = 'Rolling Sum of COVID-19 Deaths by Day', x = 'Date', y = 'Deaths')
```



Question: Briefly describe what you have learned from this data set.

Answer: Since the first half of 2020, the death rate from COVID-19 has decreased significantly. The rolling sum of full vaccinations increased exponentially in the year 2021 but has leveled off in 2022. There has been a gradual increase in COVID-19 deaths throughout the course of the pandemic. Throughout the pandemic, confirmed COVID-19 cases have increased gradually during the spring, summer, and fall months but significantly during the winter months.