PPHA 30560 Problem Set 1

Michael Gorman

3/31/2021

```
data <- read_csv("hw1-data.csv") %>%
  arrange(date) %>%
  group_by(state) %>%
  mutate(new_cases = cases - lag(cases),
         new_deaths = deaths - lag(deaths),
         rolling_average_cases = rollmean(new_cases, 7, fill = NA_real_, align = "right"),
         rolling_average_deaths = rollmean(new_deaths, 7, fill = NA_real_, align = "right"),
         death_percent = new_deaths / new_cases,
         rolling_average_death_percent = rolling_average_deaths / rolling_average_cases)

# NOTE: new_cases and new_deaths each have a couple dozen negative numbers,
# which is likely a result of governments retroactively changing their reporting
# practices without retroactively updating the old observations.
#
# California only has one day with a negative new death count (September 20),
# and it's only -5, so I don't expect it to have much impact on the data.
# The US as a whole has no negative days, but I am not sure whether that's
# because the federal government had consistent reporting standards throughout
# the period, because the federal government's data were retroactively updated,
# or because there is such a large national population that any retroactive
# negative adjustments are masked by new cases.
#
# I am choosing to ignore this finding for the purpose of the assignment,
# but if there were a wider use case for the chart, we would want to take the
# time to understand the causes of these adjustments and try to correct for them.
```

Task 1

How many unique values are in the state variable?

```
data %>%
  group_by(state) %>%
  summarise() %>%
  nrow()
```

```
## [1] 57
```

Task 2

List the unique values in the state variable.

```r
data %>%
  group_by(state) %>%
  summarise() %>%
  arrange(state) %>%
  knitr::kable()
```

state
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
District of Columbia
Florida
Georgia
Guam
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New England
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Northern Mariana Islands
Ohio
Oklahoma
Oregon
Pennsylvania

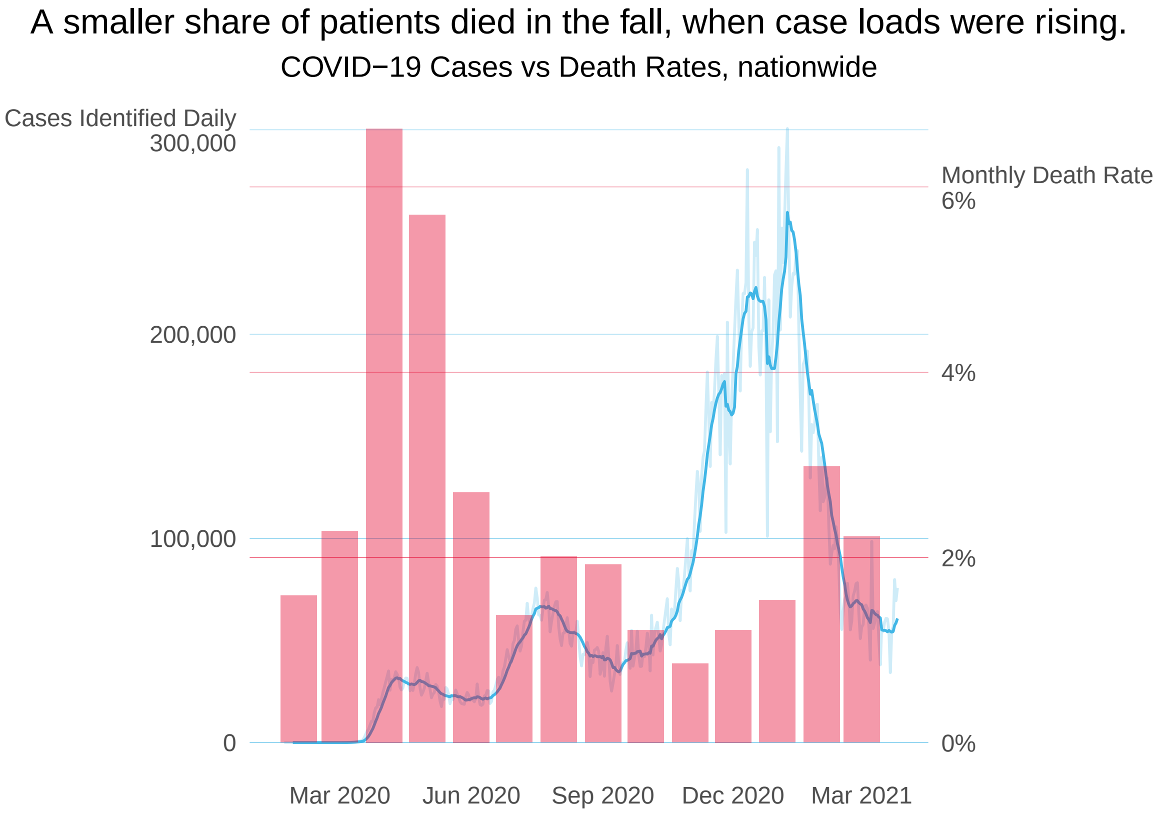
state
Puerto Rico
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
USA
Utah
Vermont
Virgin Islands
Virginia
Washington
West Virginia
Wisconsin
Wyoming

The dataset contains observations for all 50 states, the District of Columbia, US territories, the New England region, and the nation as a whole.

Task 3

Make a coronavirus chart for the United States and California.

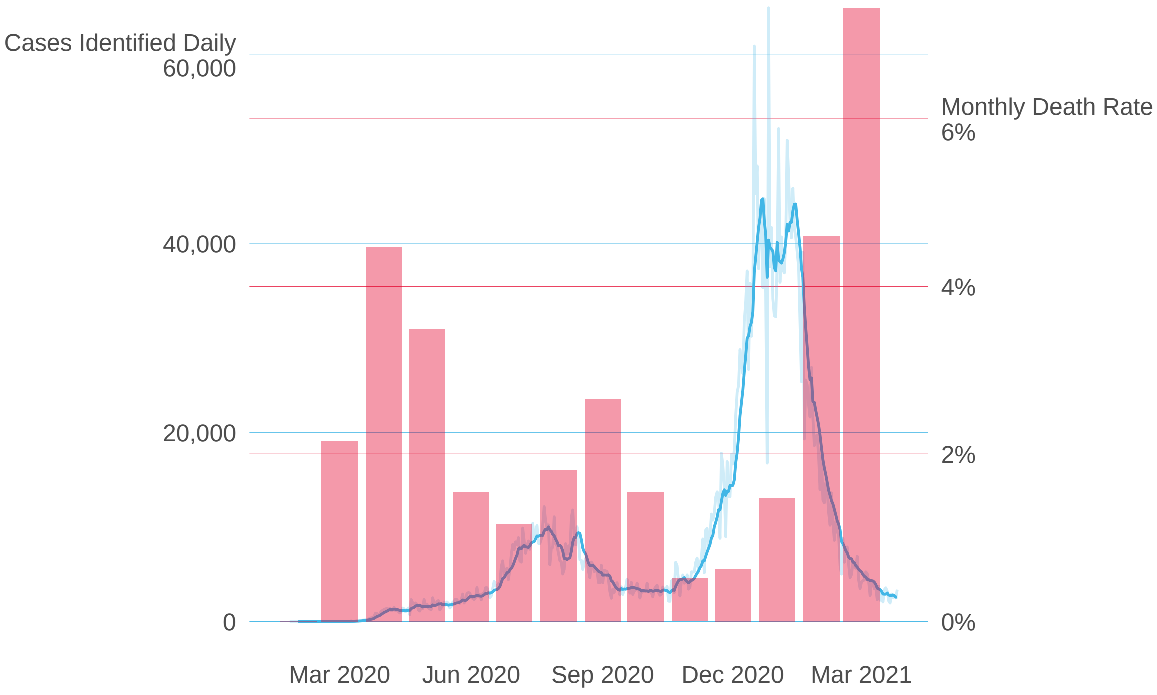
```r
us_data_cases <- data %>%
  filter(state == "USA")

us_data_deaths <- data %>%
  filter(state == "USA") %>%
  mutate(year = year(date),
         month = month(date)) %>%
  group_by(year, month) %>%
  summarise(death_percent = sum(new_deaths) / sum(new_cases)) %>%
  mutate(date = ymd(paste(year, month, "01", sep = "-")))

# dual Y axis solution inspired by: https://stackoverflow.com/a/51844068
us_scale_factor <- max(us_data_cases$new_cases, na.rm = TRUE) / max(us_data_deaths$death_percent, na.rm

us_data_cases %>%
  ggplot(mapping = aes(x = date)) +
  geom_line(mapping = aes(y = new_cases),
            color = "#41B6E6",
            alpha = 0.25) +
  geom_line(mapping = aes(y = rolling_average_cases),
            color = "#41B6E6") +
  geom_col(data = us_data_deaths,
           mapping = aes(y = death_percent * us_scale_factor),
           fill = "#E4002B",
           alpha = 0.4) +
  scale_x_date(name = "",
               date_breaks = "3 months",
               date_labels = "%b %Y") +
  scale_y_continuous(name = "",
                     breaks = c(0, 100000, 200000, 300000),
                     minor_breaks = c(0.0, 0.02, 0.04, 0.06) * us_scale_factor,
                     labels = c("0", "100,000", "200,000", "Cases Identified Daily\n300,000"),
                     sec.axis = sec_axis(trans = ~./us_scale_factor,
                                         name = "",
                                         breaks = c(0.0, 0.02, 0.04, 0.06),
                                         labels = c("0%", "2%", "4%", "Monthly Death Rate\n6%"))) +
  theme_minimal() +
  theme(panel.grid.major.y = element_line(color = "#41B6E6", size = 0.075),
        panel.grid.minor.y = element_line(color = "#E4002B", size = 0.075),
        panel.grid.major.x = element_blank(),
        panel.grid.minor.x = element_blank(),
        plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        plot.title.position = "plot") +
  labs(title = "A smaller share of patients died in the fall, when case loads were rising.",
       subtitle = "COVID-19 Cases vs Death Rates, nationwide")
```



```
ca_data_cases <- data %>%
  filter(state == "California")

ca_data_deaths <- data %>%
  filter(state == "California") %>%
  mutate(year = year(date),
         month = month(date)) %>%
  group_by(year, month) %>%
  summarise(death_percent = sum(new_deaths) / sum(new_cases)) %>%
  mutate(date = ymd(paste(year, month, "01", sep = "-")))

# dual Y axis solution inspired by: https://stackoverflow.com/a/51844068
ca_scale_factor <- max(ca_data_cases$new_cases, na.rm = TRUE) / max(ca_data_deaths$death_percent, na.rm

ca_data_cases %>%
  ggplot(mapping = aes(x = date)) +
  geom_line(mapping = aes(y = new_cases),
            color = "#41B6E6",
            alpha = 0.25) +
  geom_line(mapping = aes(y = rolling_average_cases),
            color = "#41B6E6") +
  geom_col(data = ca_data_deaths,
           mapping = aes(y = death_percent * ca_scale_factor),
           fill = "#E4002B",
           alpha = 0.4) +
  scale_x_date(name = "",
               date_breaks = "3 months",
               date_labels = "%b %Y") +
  scale_y_continuous(name = "",
                     breaks = c(0, 20000, 40000, 60000),
                     minor_breaks = c(0.0, 0.02, 0.04, 0.06) * ca_scale_factor,
                     labels = c("0", "20,000", "40,000", "Cases Identified Daily\n60,000"),
                     sec.axis = sec_axis(trans = ~./ca_scale_factor,
                                         name = "",
                                         breaks = c(0.0, 0.02, 0.04, 0.06),
                                         labels = c("0%", "2%", "4%", "Monthly Death Rate\n6%"))) +
  theme_minimal() +
  theme(panel.grid.major.y = element_line(color = "#41B6E6", size = 0.075),
        panel.grid.minor.y = element_line(color = "#E4002B", size = 0.075),
        panel.grid.major.x = element_blank(),
        panel.grid.minor.x = element_blank(),
        plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        plot.title.position = "plot") +
  labs(title = "A higher share of patients are dying as the case load is decreasing.",
       subtitle = "COVID-19 Cases vs Death Rates, California alone")
```

A higher share of patients are dying as the case load is decreasing.

COVID−19 Cases vs Death Rates, California alone


Cases Identified Daily
Monthly Death Rate
Mar 2020
Jun 2020
Sep 2020
Dec 2020
Mar 2021