Customized Covid-19 news

Ian D. Gow

2020-06-09

I recently read a newspaper headline that read "California and Some Other States See Coronavirus Cases Rise" and thought "I already knew that!" This was because I had been dabbling with the data to make my own plots.

An interesting aspect of the COVID-19 pandemic is that data-based reporting has moved to the fore. Nonetheless, I often find the reporting inadequate, not because it's bad, but because what I am interested in understanding isn't necessarily what is being reported.

But, with a modicum of data skills, it is easy to do your own reporting. For example, the New York Times provides COVID-19 data by state here. And *Our World in Data* has extensive data on COVID-19 around the world.

The two key data variables are *cases* and *deaths*. As a measure of progress of the pandemic, cases are a more timely statistic than deaths (and also a little less morbid), but there are by-now-well-understood issues with cases, such as under-reporting due to lack of symptoms, or symptoms insufficiently serious to lead to hospitalization or testing. As pointed out in the *WSJ* article linked to above, as testing capacity increases, we might expect to see a rise in reported cases, even if the underlying number of cases is flat.

Getting the data

Getting the data is quite easy.

```
library(readr)
library(lubridate)
library(dplyr, warn.conflicts = FALSE)
library(ggplot2)
library(RcppRoll)

raw <- read_csv(paste0("https://raw.githubusercontent.com/",
                       "nytimes/covid-19-data/master/us-states.csv"))

covid_world_raw <- read_csv(paste0("https://covid.ourworldindata.org",
                                   "/data/owid-covid-data.csv"),
                            col_types = cols(.default = col_guess(),
                                             new_tests = col_double(),
                                             new_tests_smoothed = col_double(),
                                             new_tests_smoothed_per_thousand = col_double(),
                                             tests_per_case =  col_double(),
                                             positive_rate =  col_double(),
                                             total_tests = col_double(),
                                             total_tests_per_thousand = col_double(),
                                             new_tests_per_thousand = col_double(),
                                             weekly_hosp_admissions = col_double(),
                                             weekly_hosp_admissions_per_million = col_double(),
```

```
                                                weekly_icu_admissions = col_double(),
                                                weekly_icu_admissions_per_million = col_double(),
                                                icu_patients = col_double(),
                                                icu_patients_per_million = col_double(),
                                                hosp_patients = col_double(),
                                                hosp_patients_per_million = col_double(),
                                                tests_units = col_character()))

covid_states <-
  raw %>%
  group_by(state) %>%
  arrange(date) %>%
  mutate(new_cases = cases - lag(cases),
         new_deaths = deaths - lag(deaths)) %>%
  rename(total_deaths = deaths,
         total_cases = cases)
```

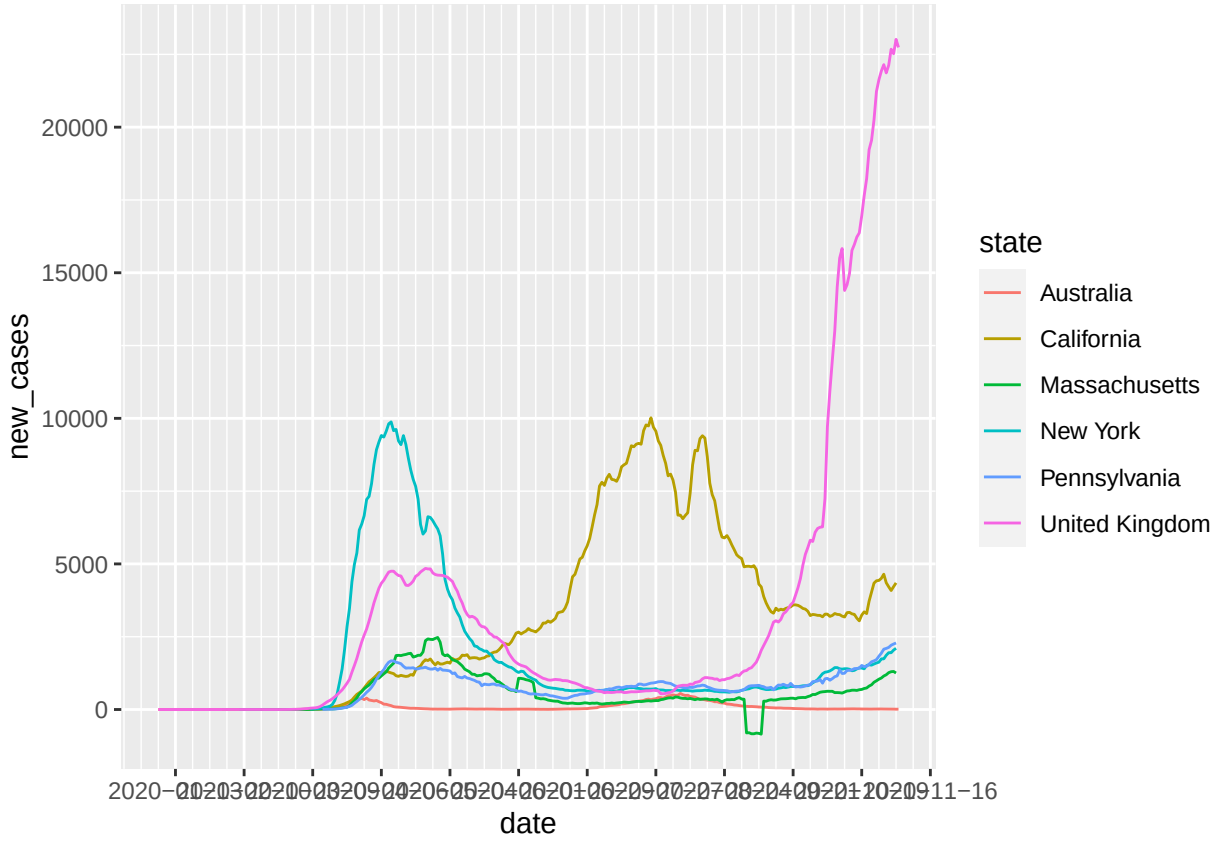
Initially, I wanted to consider Australia as a US state for the purpose of comparison. For the US, I focused on a few states of interest: New York was the worst-hit state, Massachusetts is where I am now, and California is the most populous state. Pennsylvania provides an interesting comparison for Massachusetts. Apart from Australia, I also considered the United Kingdom, which was hit at about the same time as New York.

```
selected_states <- c("Massachusetts", "New York", "Pennsylvania", "California")
selected_countries <- c("AUS", "GBR")

covid_aus_usa <-
  covid_world_raw %>%
  filter(iso_code %in% selected_countries) %>%
  select(location:new_deaths) %>%
  rename(state = location)

covid_selected <-
  covid_states %>%
  filter(state %in% selected_states) %>%
  union_all(covid_aus_usa)
```

Cases

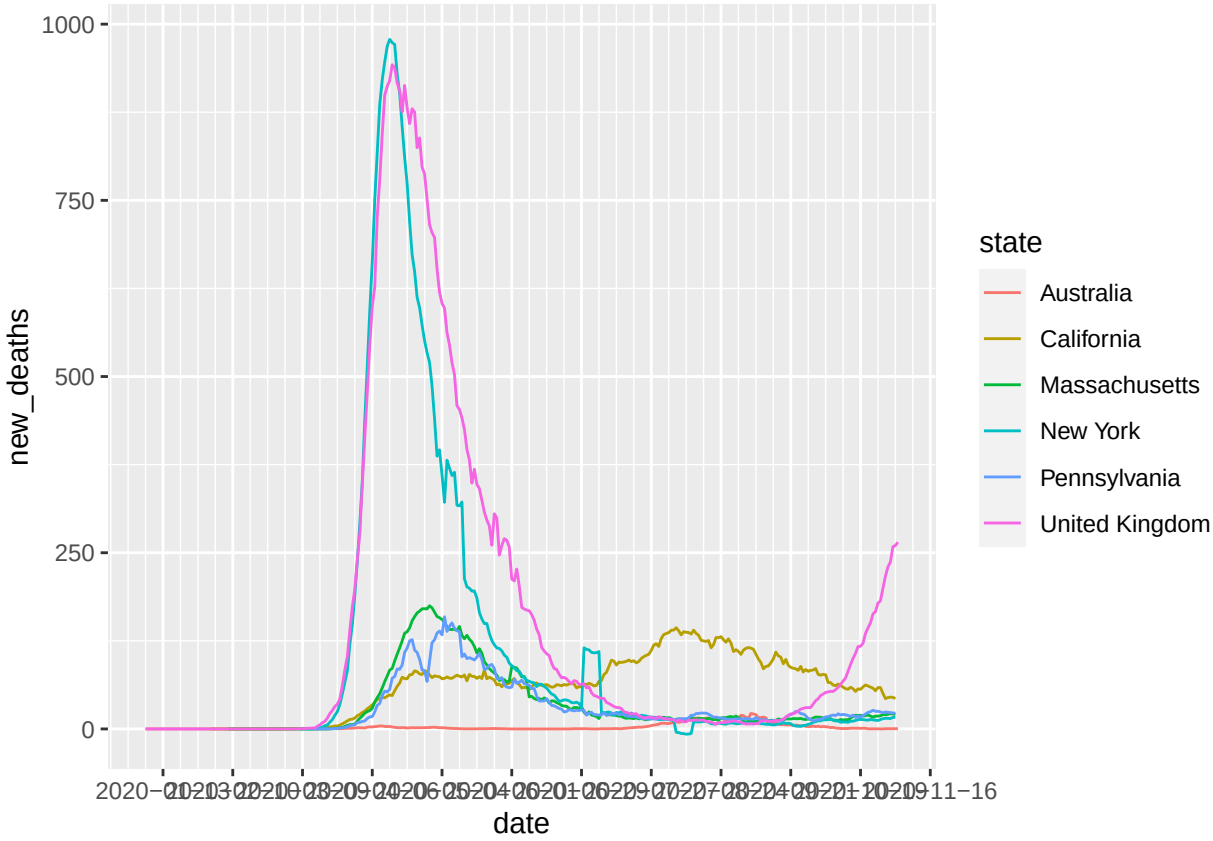
One thing I noticed initially was a definite lumpiness to the data (e.g., many more deaths on Tuesdays rather than Sundays in Pennsylvania), which I assume is down to reporting rather than actual events. Initially, I used a four-day moving average, but here I use a seven-day moving average (as the four-day one still showed clean peaks and valleys).

```
covid_selected %>%
  group_by(state) %>%
  arrange(date) %>%
  mutate(new_cases = roll_meanr(new_cases, n = 7, fill = NA)) %>%
  filter(!is.na(new_cases)) %>%
  ggplot(aes(x = date, y = new_cases, color = state)) +
  geom_line() +
  scale_x_date(breaks = "4 weeks", date_minor_breaks = "1 week")
```



Deaths

```
# So use a seven-day rolling average
covid_selected %>%
  group_by(state) %>%
  arrange(date) %>%
  mutate(new_deaths = roll_meanr(new_deaths, n = 7, fill = NA)) %>%
  filter(!is.na(new_deaths)) %>%
  ggplot(aes(x = date, y = new_deaths, color = state)) +
  geom_line() +
  scale_x_date(breaks = "4 weeks", date_minor_breaks = "1 week")
```


1000
750
500
250
0
new_deaths
2020−11−16
date
state
Australia
California
Massachusetts
New York
Pennsylvania
United Kingdom