Exploration of COVID-19 tracking data from multiple resources

Wei Sun

2020-04-22

Contents

Introduction

Coronavirus disease 2019 (COVID-19) is an infectious disease caused by a new type of coronavirus: severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The outbreak first started in Wuhan, China in December 2019. The first kown case of COVID-19 in the U.S. was confirmed on January 20, 2020, in a 35-year-old man who teturned to Washington State on January 15 after traveling to Wuhan. Starting around the end of Feburary, evidence emerge for community spread in the US.

We, as all of us, are indebted to the heros who fight COVID-19 across the whole world in different ways. For this data exploration, I am grateful to many data science groups who have collected detailed COVID-19 outbreak data, including the number of tests, confirmed cases, and deaths, across countries/regions, states/provnices (administrative division level 1, or admin1), and counties (admin2). Specifically, I used the data from these three resources:

- JHU (https://coronavirus.jhu.edu/)
  - The Center for Systems Science and Engineering (CSSE) at John Hopkins University.
  - World-wide counts of coronavirus cases, deaths, and recovered ones.
  - https://github.com/CSSEGISandData/COVID-19
- NY Times (https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html)
  - The New York Times
  - "cumulative counts of coronavirus cases in the United States, at the state and county level, over time"
  - https://github.com/nytimes/covid-19-data

- COVID Trackng (https://covidtracking.com/)
  - COVID Tracking Project
  - "collects information from 50 US states, the District of Columbia, and 5 other US territories to provide the most comprehensive testing data"
  - https://github.com/COVID19Tracking/covid-tracking-data

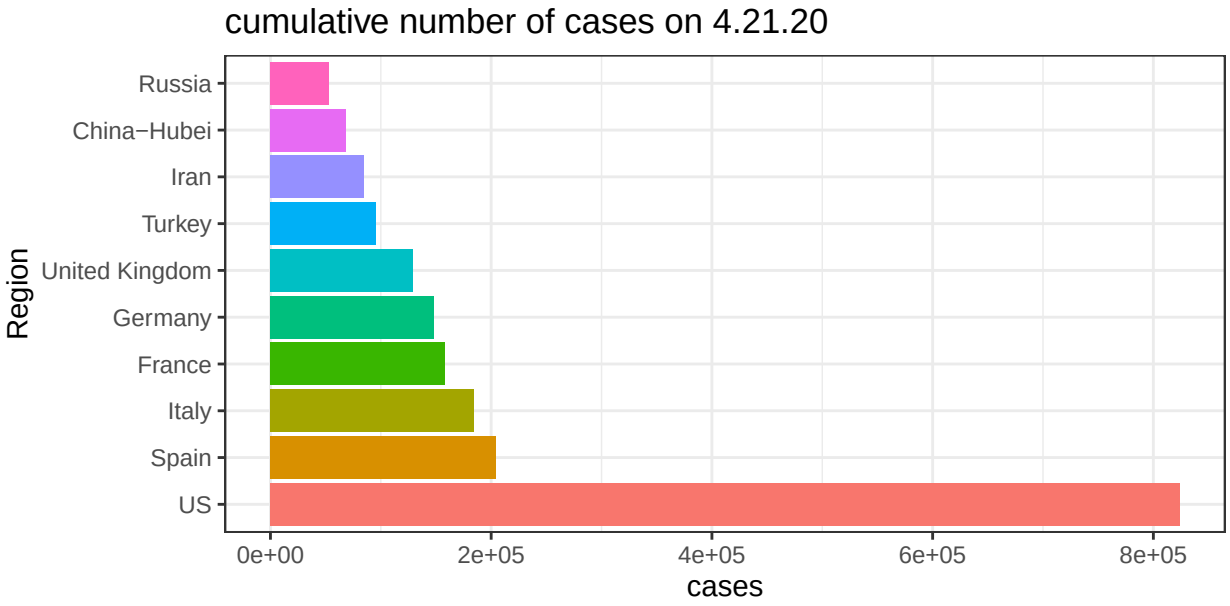
JHU

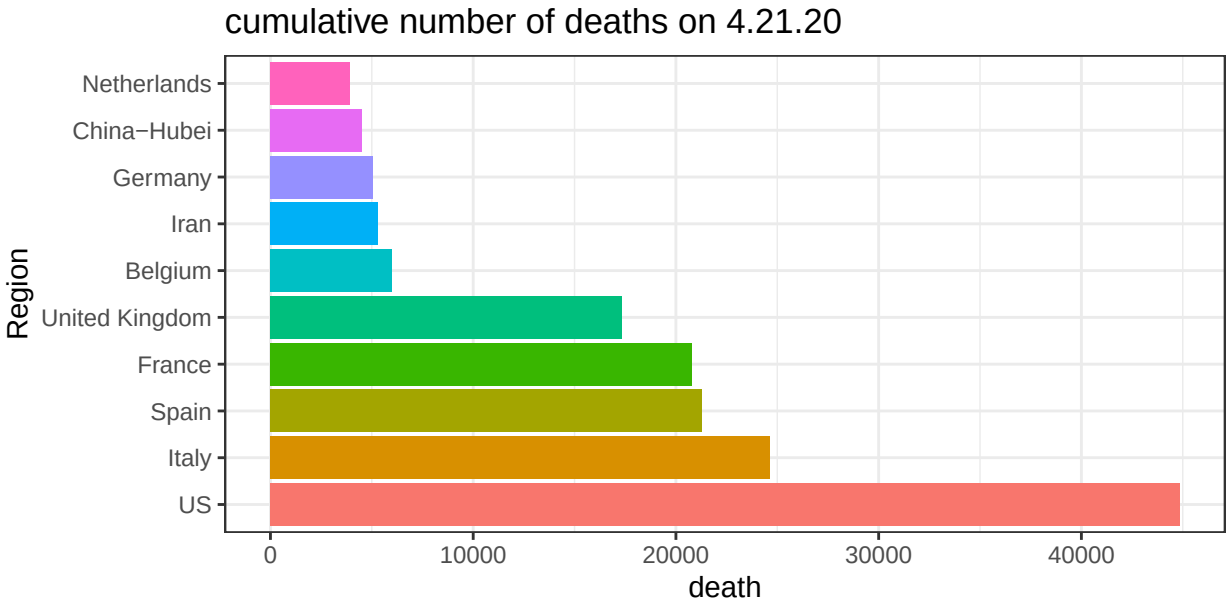
Assume you have cloned the JHU Github repository on your local machine at "../COVID-19".

time series data

The time series provide counts (e.g., confirmed cases, deaths) starting from Jan 22nd, 2020 for 253 locations. Currently there is no data of individual US state in these time series data files.

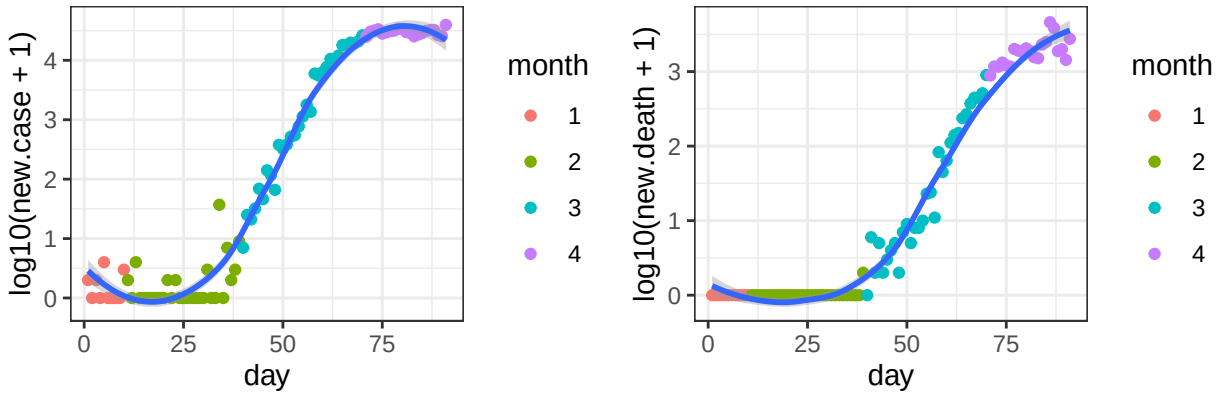
Here is the list of 10 records with the largest number of cases or deaths on the most recent date.





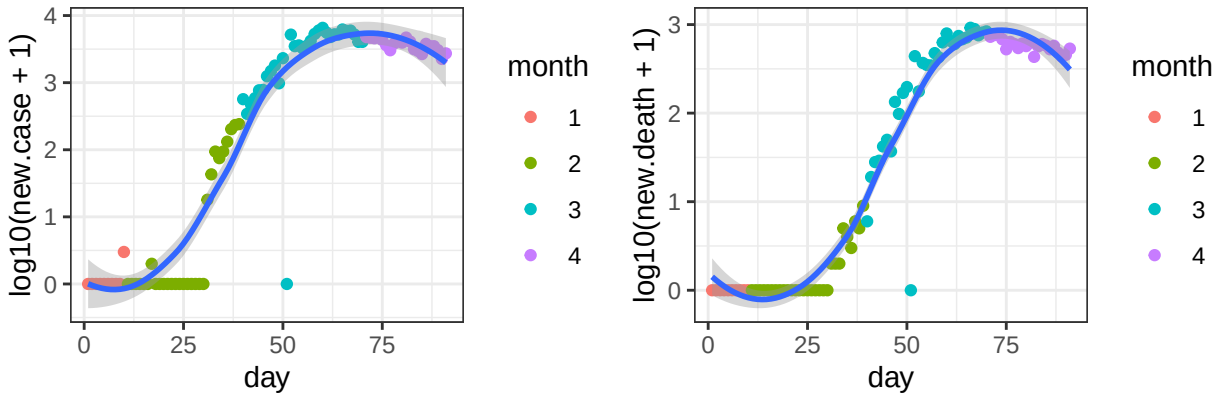
Next, I check for each country/region, what is the number of new cases/deaths? This data is important to understand what is the trend under different situations, e.g., population density, social distance policies etc. Here I checked the top 10 countries/regions with the highest number of deaths.

US



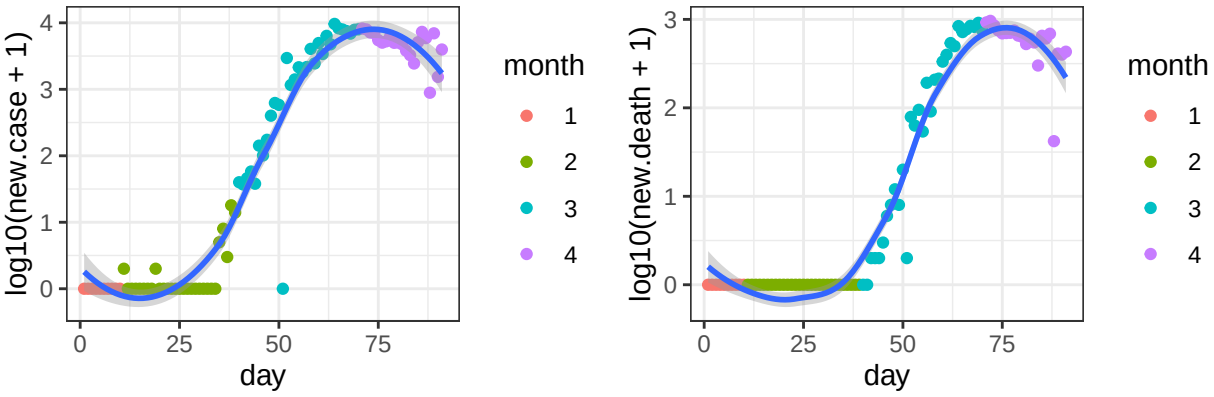
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Italy



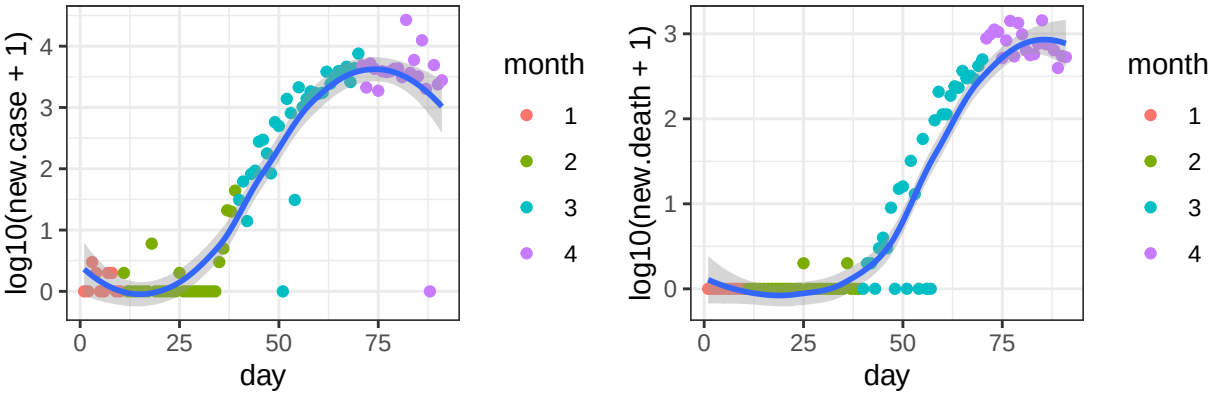
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Spain



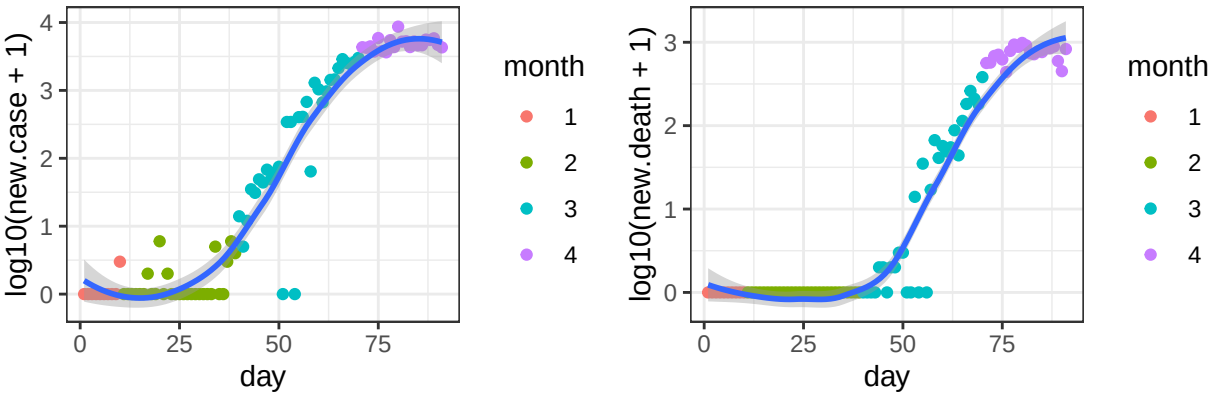
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

France



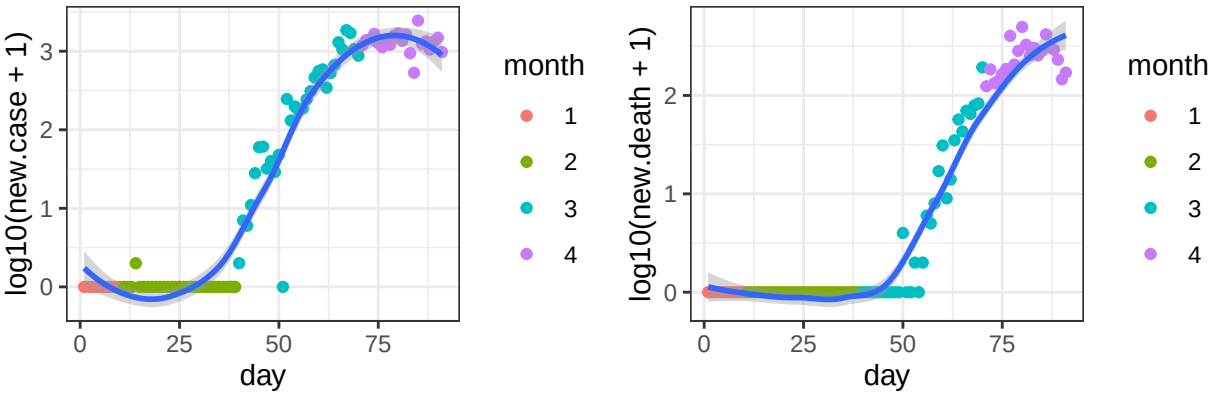
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

United Kingdom



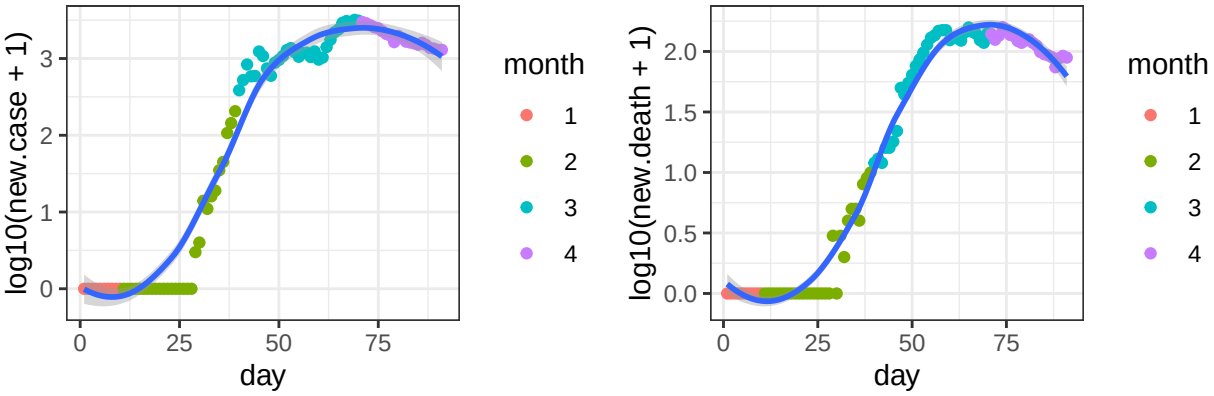
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Belgium



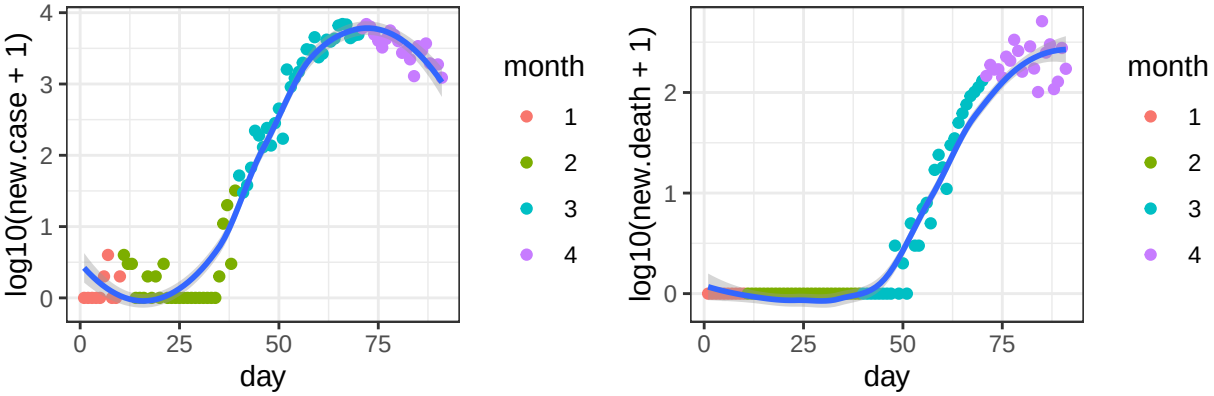
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Iran

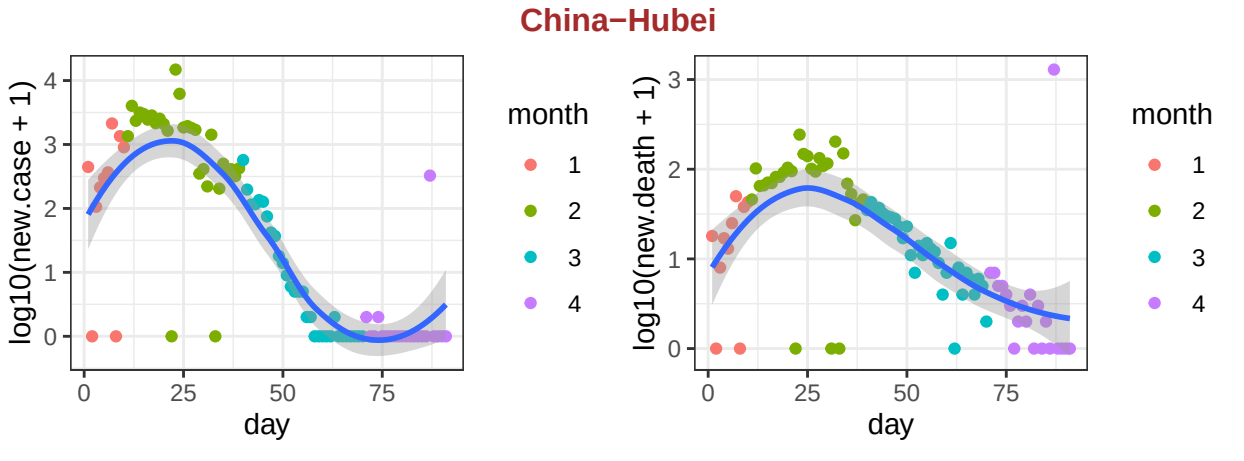


data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

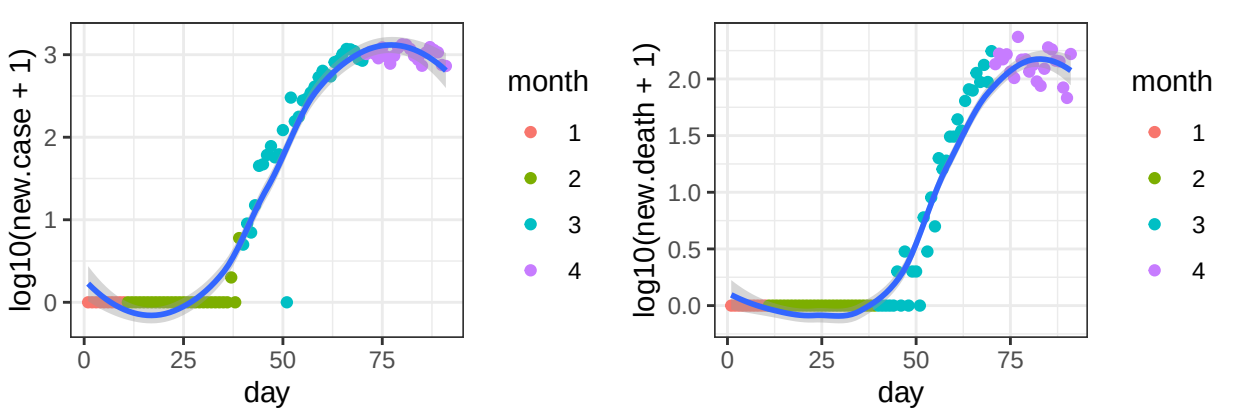
Germany



data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020



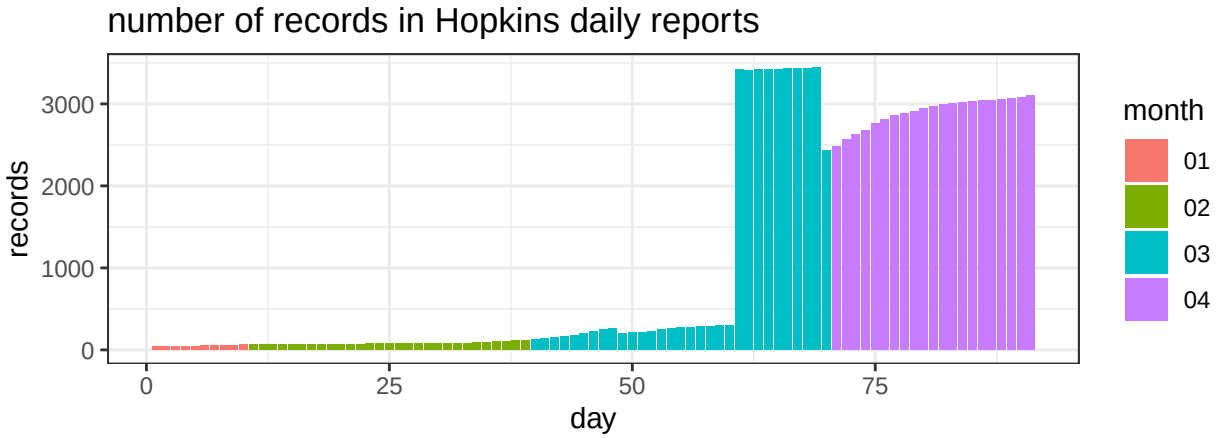
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020



data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

daily reports data

The raw data from Hopkins are in the format of daily reports with one file per day. More recent files (since March 22nd) inlcude information from individual states of US or individual counties, as shown in the following figure. So I turn to NY Times data for informatoin of individual states or counties.



data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

NY Times

The data from NY Times are saved in two text files, one for state level information and the other one for county level information.

The currente date is

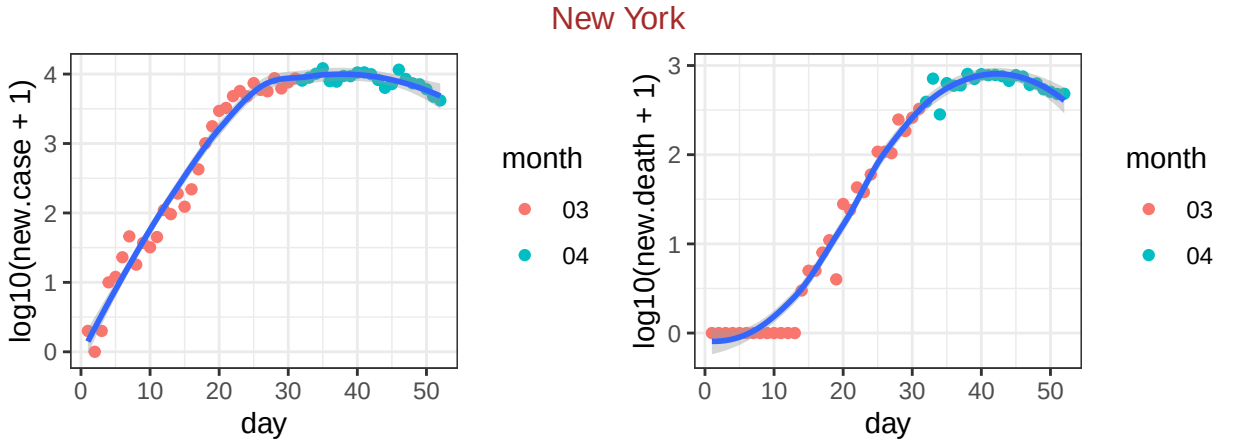
```
## [1] "2020-04-21"
```

state level data

First check the 20 states with the largest number of deaths.

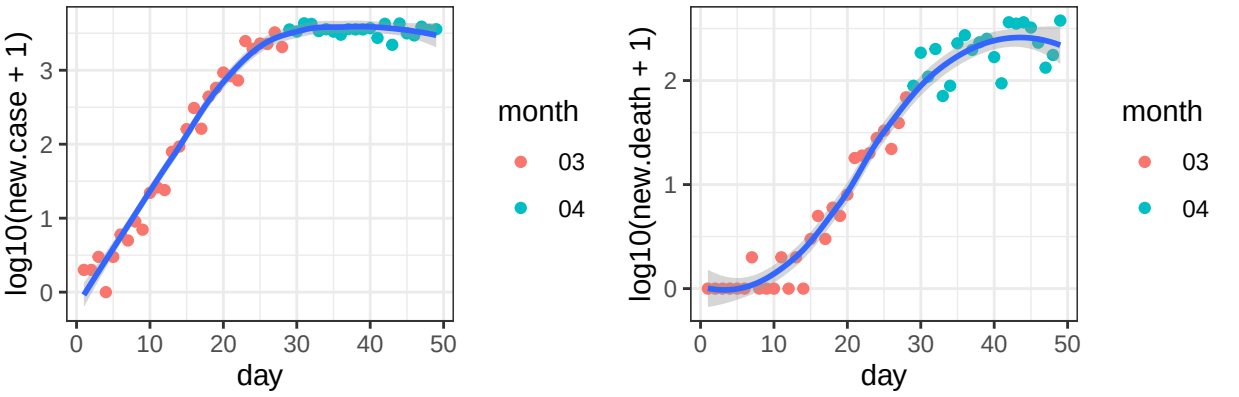
```
##            date          state fips  cases deaths
## 2756 2020-04-21       New York   36 251720  14828
## 2754 2020-04-21     New Jersey   34  92387   4753
## 2746 2020-04-21       Michigan   26  32935   2698
## 2745 2020-04-21  Massachusetts   25  41199   1961
## 2763 2020-04-21   Pennsylvania   42  35384   1620
## 2737 2020-04-21       Illinois   17  33059   1479
## 2729 2020-04-21    Connecticut    9  20360   1423
## 2742 2020-04-21      Louisiana   22  24854   1405
## 2727 2020-04-21     California    6  35844   1316
## 2732 2020-04-21        Florida   12  27861    866
## 2733 2020-04-21        Georgia   13  19189    810
## 2774 2020-04-21     Washington   53  12345    683
## 2738 2020-04-21        Indiana   18  12097    630
## 2744 2020-04-21       Maryland   24  14193    584
## 2760 2020-04-21           Ohio   39  13725    557
## 2769 2020-04-21          Texas   48  20949    552
## 2728 2020-04-21       Colorado    8  10447    484
## 2773 2020-04-21       Virginia   51   9630    325
## 2776 2020-04-21      Wisconsin   55   4620    243
## 2749 2020-04-21       Missouri   29   5941    221
```

For these 20 states, I check the number of new cases and the number of new deaths. Part of the reason for such checking is to identify whether there is any similarity on such patterns. For example, could you use the pattern seen from Italy to predict what happen in an individual state, and what are the similarities and differences across states.



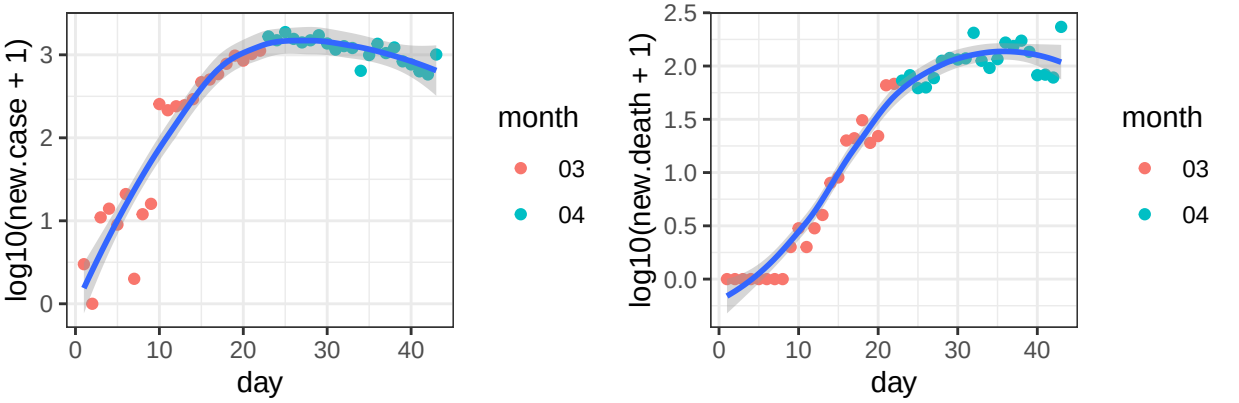
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

New Jersey



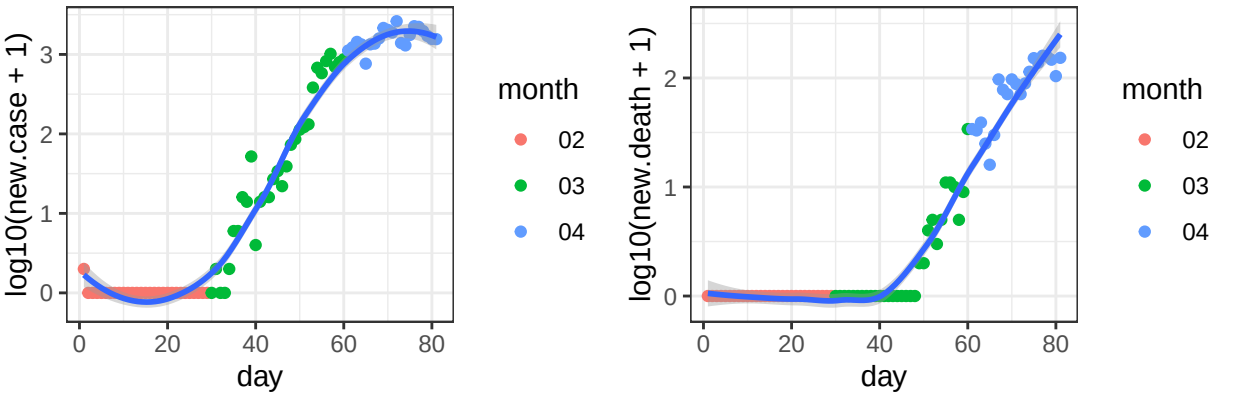
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Michigan



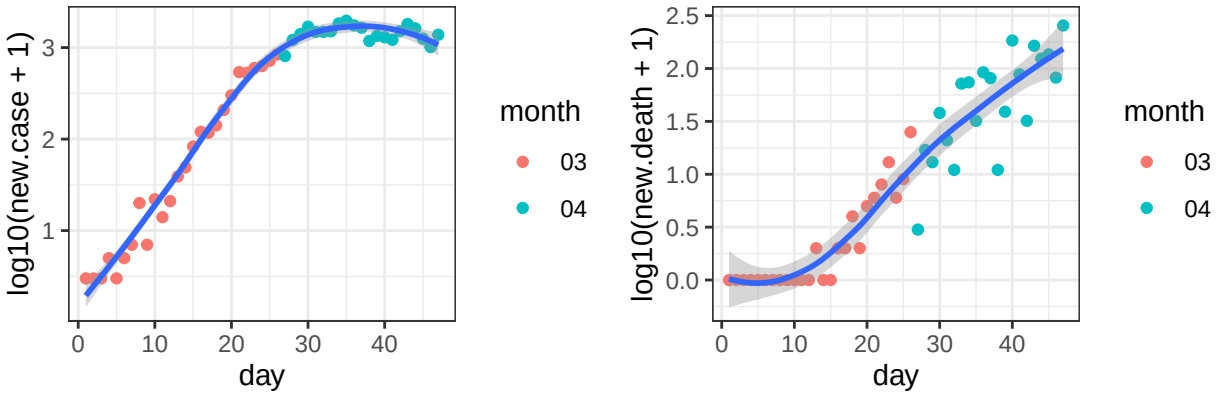
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Massachusetts



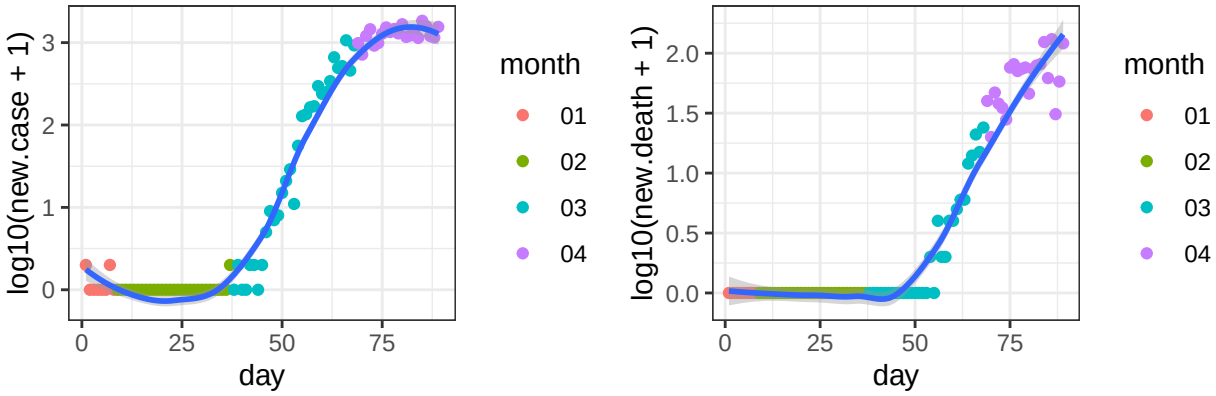
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01

Pennsylvania



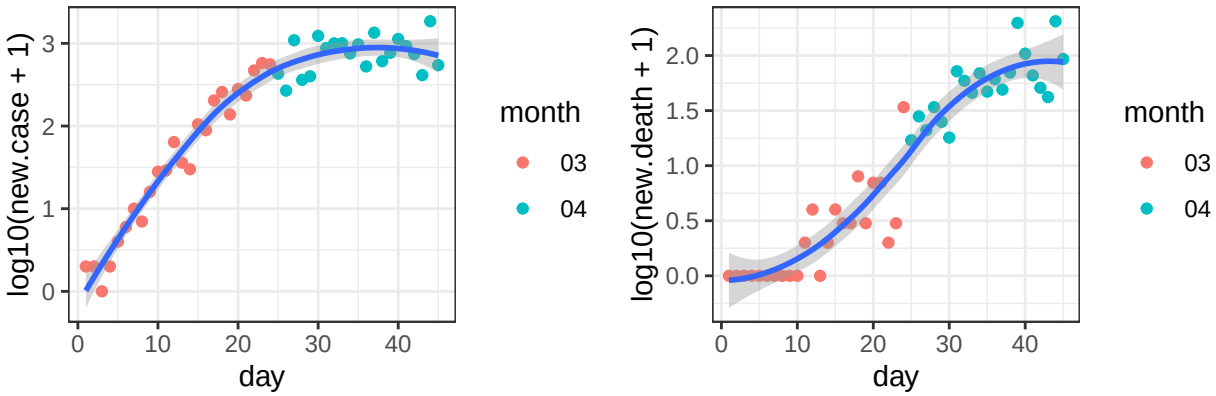
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Illinois



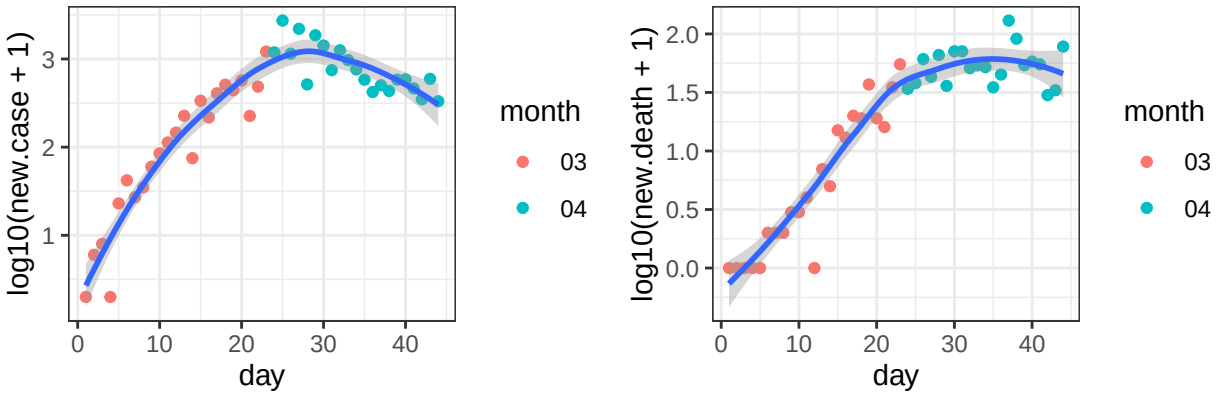
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

Connecticut



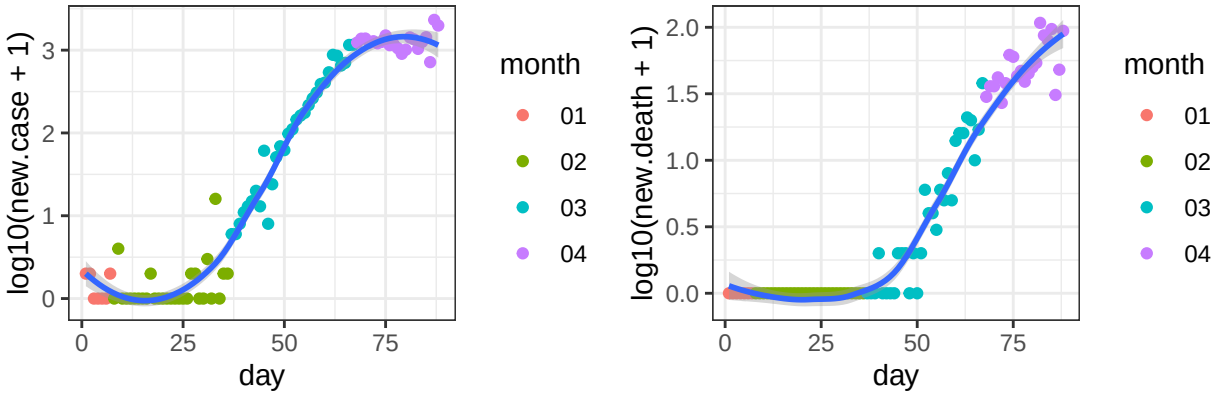
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Louisiana



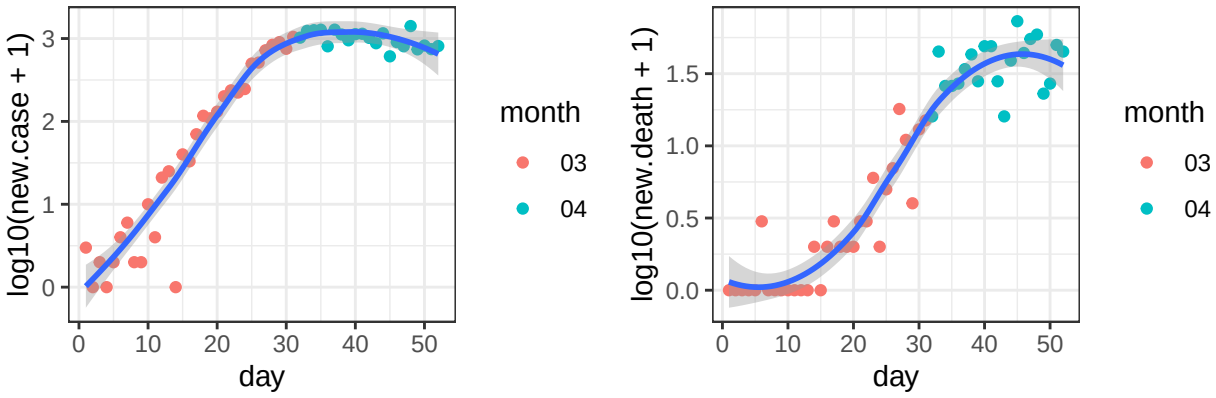
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

California



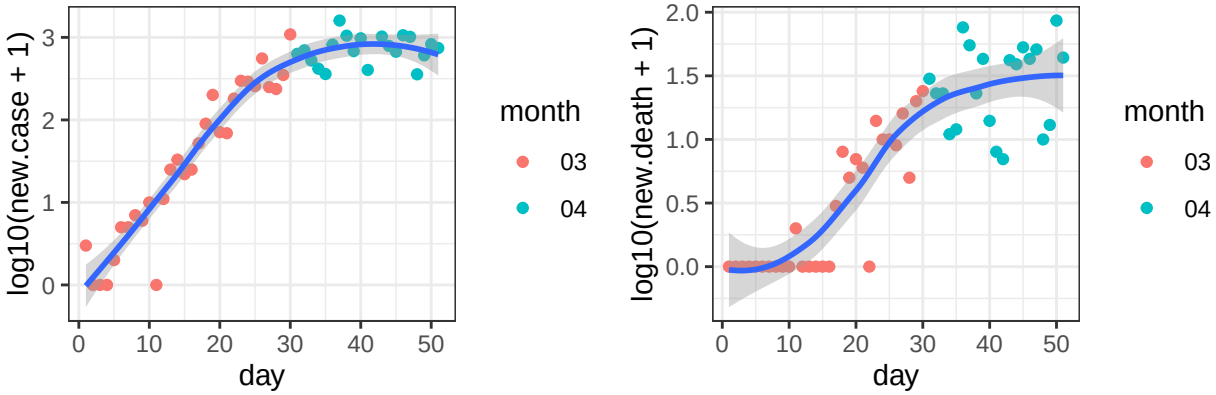
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−25

Florida



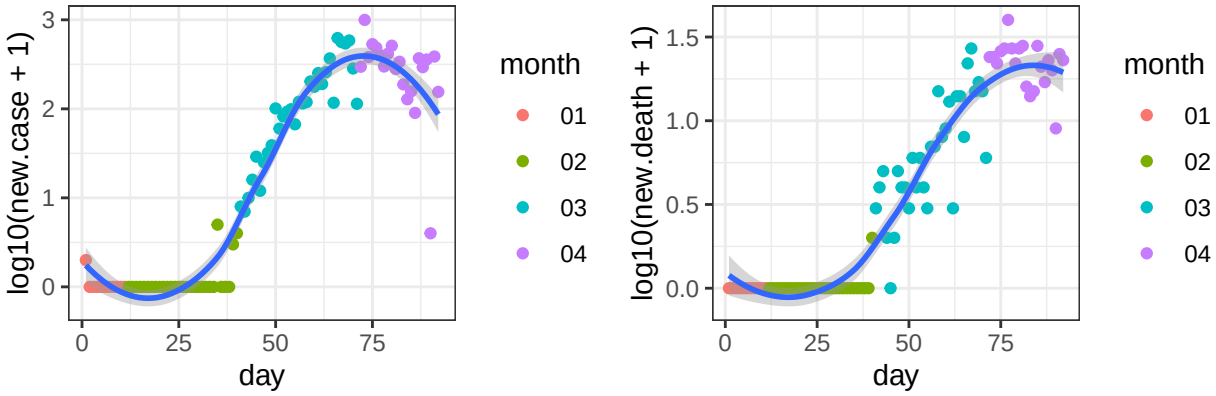
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Georgia



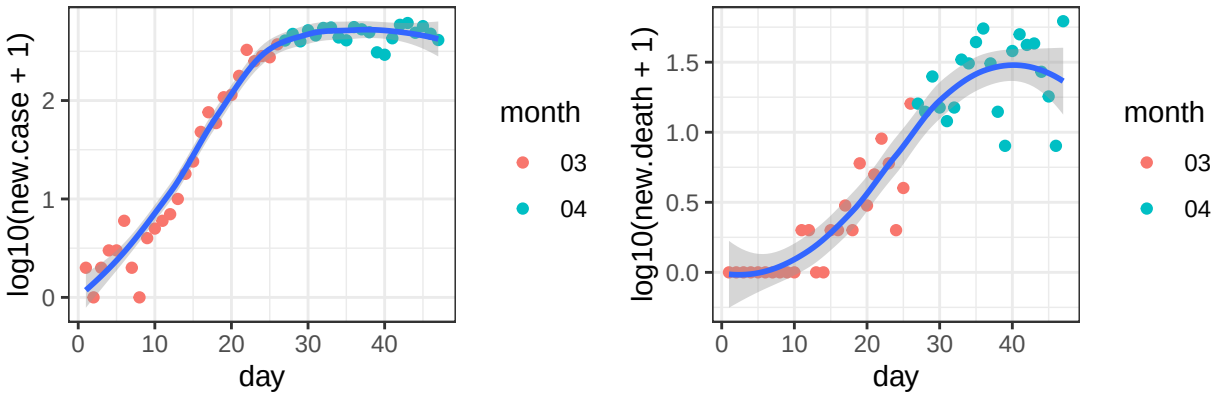
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Washington



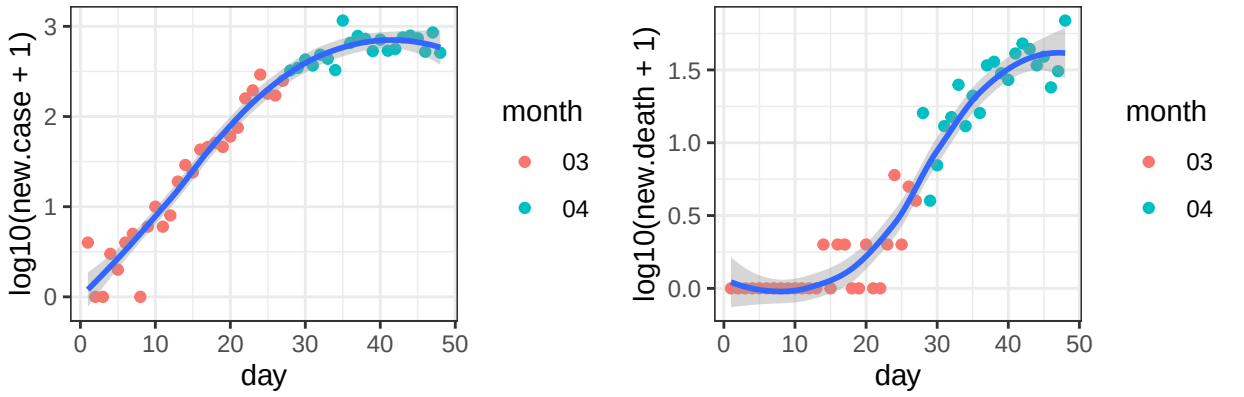
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−21

Indiana



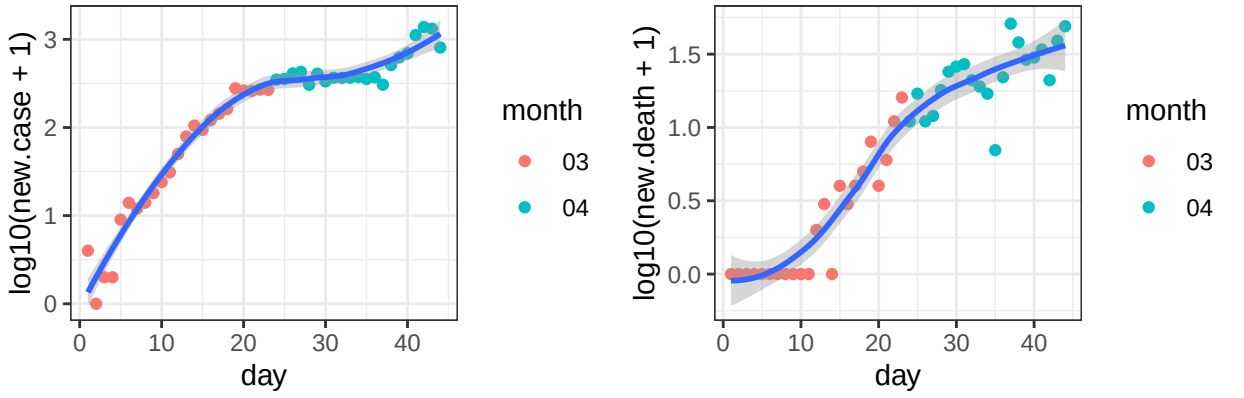
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Maryland



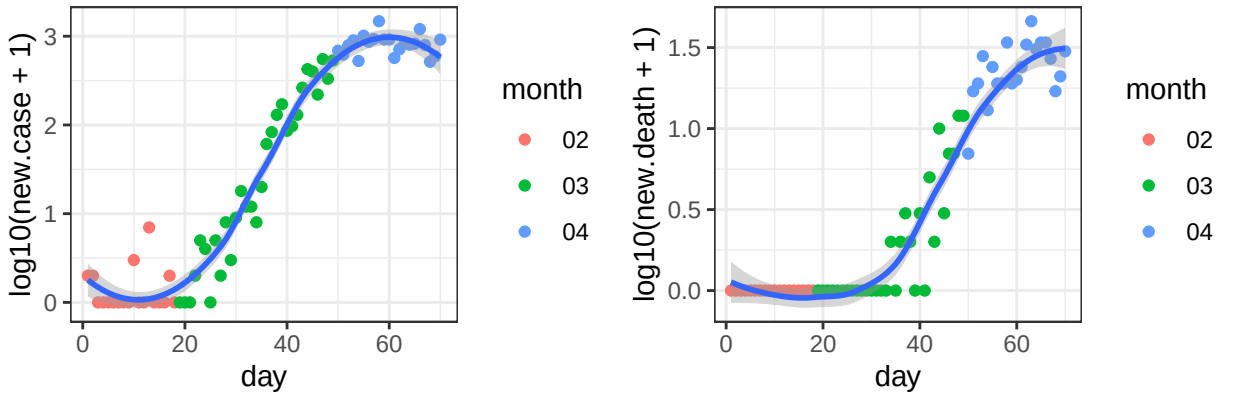
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Ohio



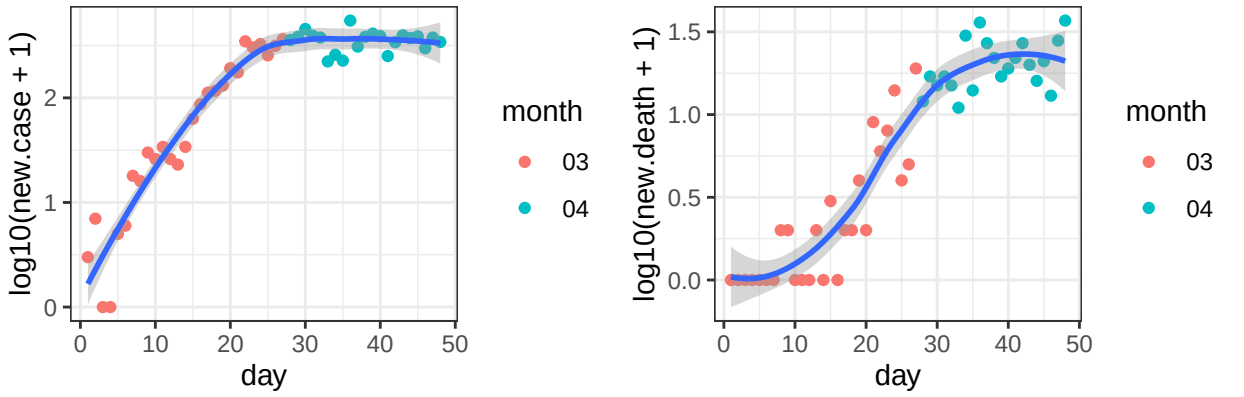
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Texas



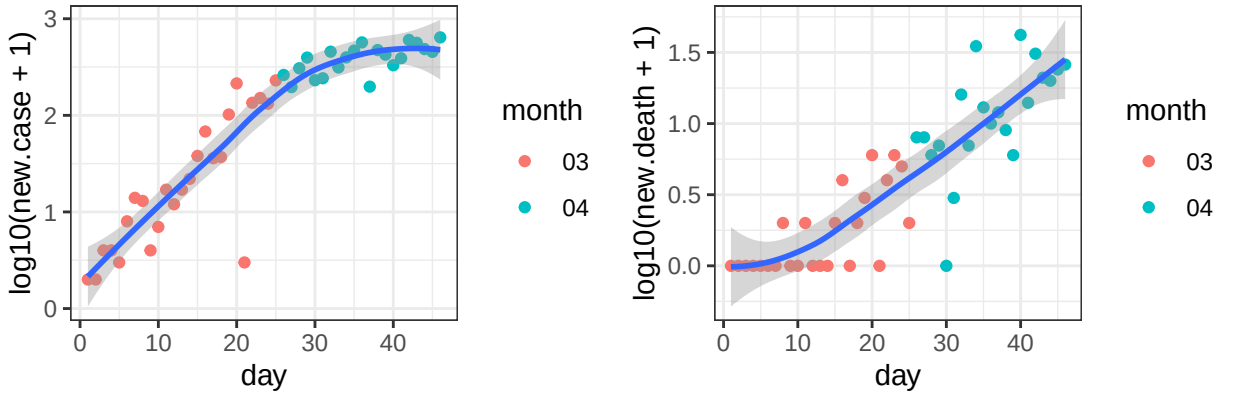
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−12

Colorado



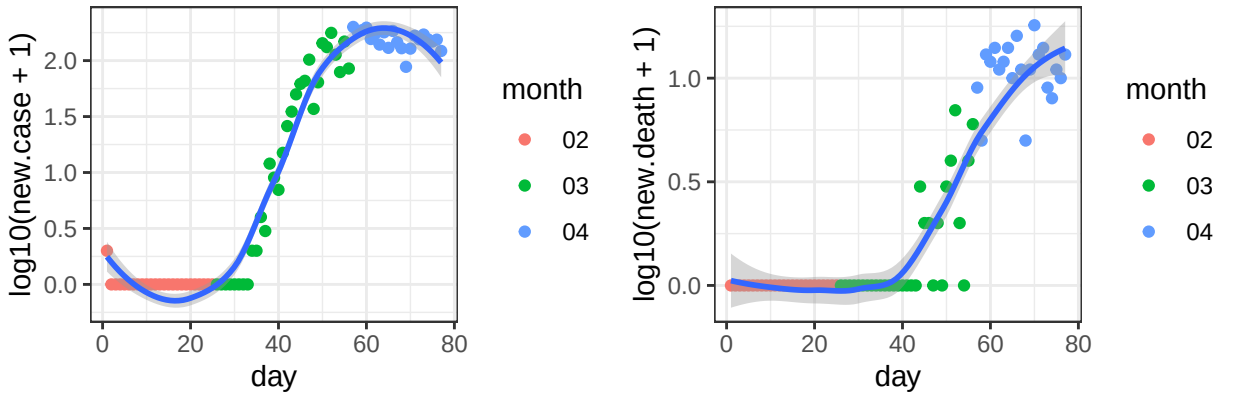
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Virginia

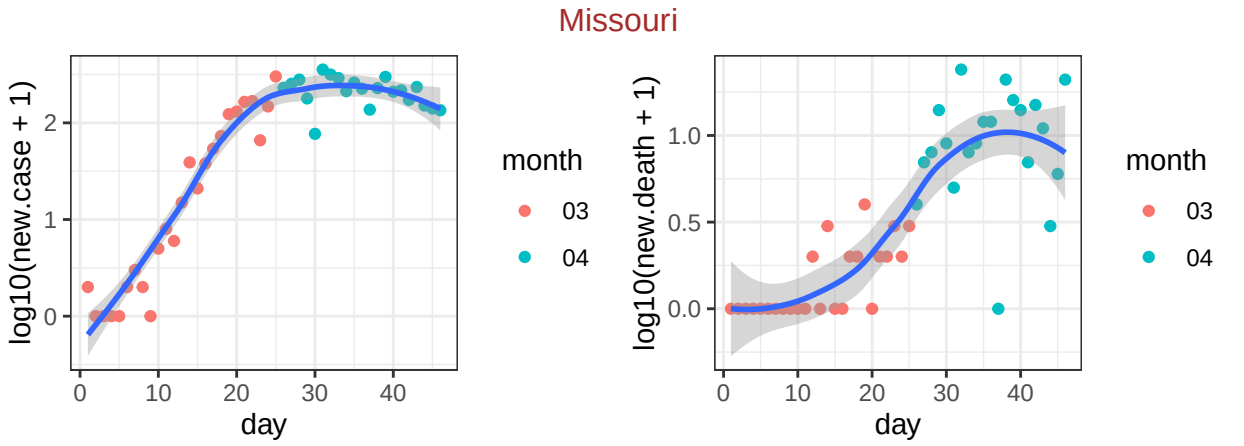


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Wisconsin

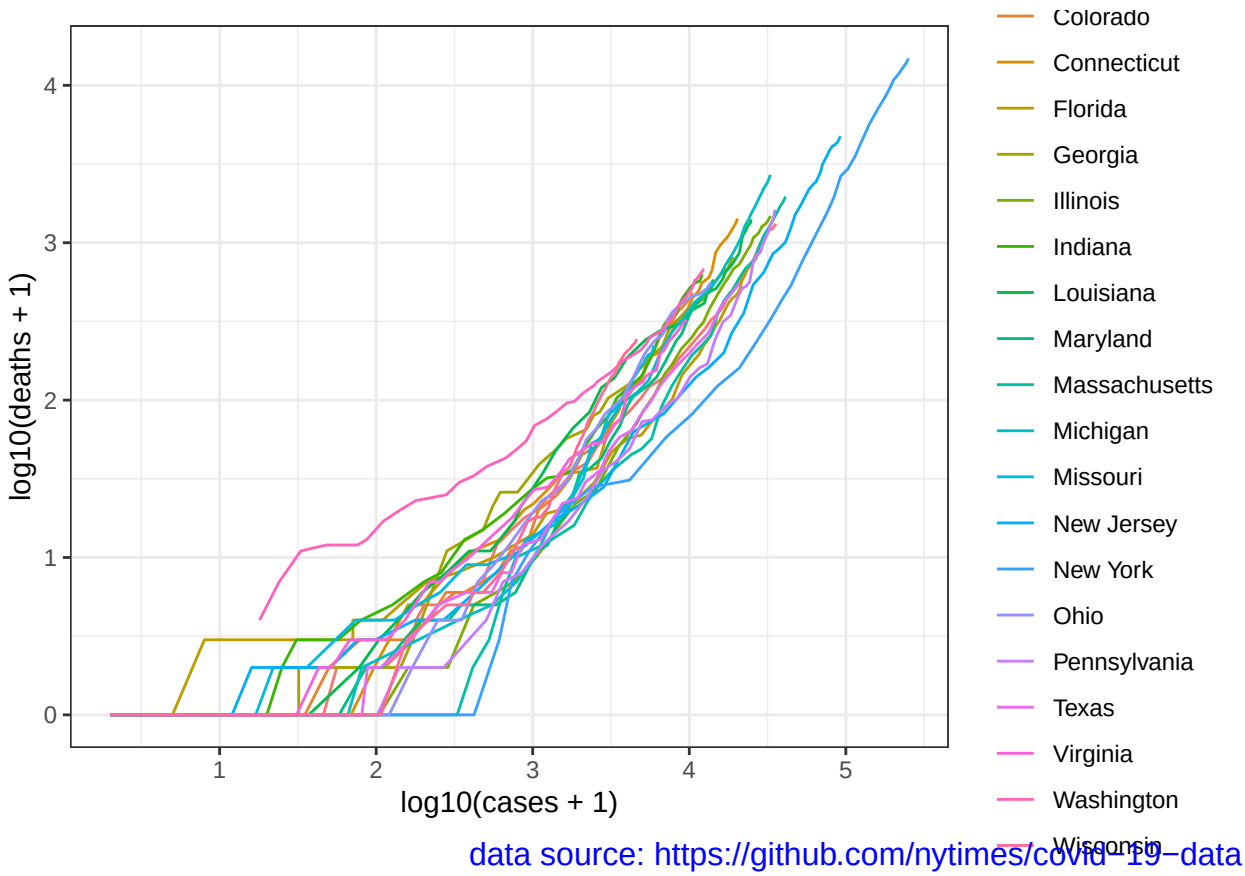


data source: https://github.com/nytimes/covid−19−data, day 1 is 02−05



data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Next I check the relation between the **cumulative** number of cases and deaths for these 10 states, starting on March



data source: https://github.com/nytimes/covid−19−data

county level data

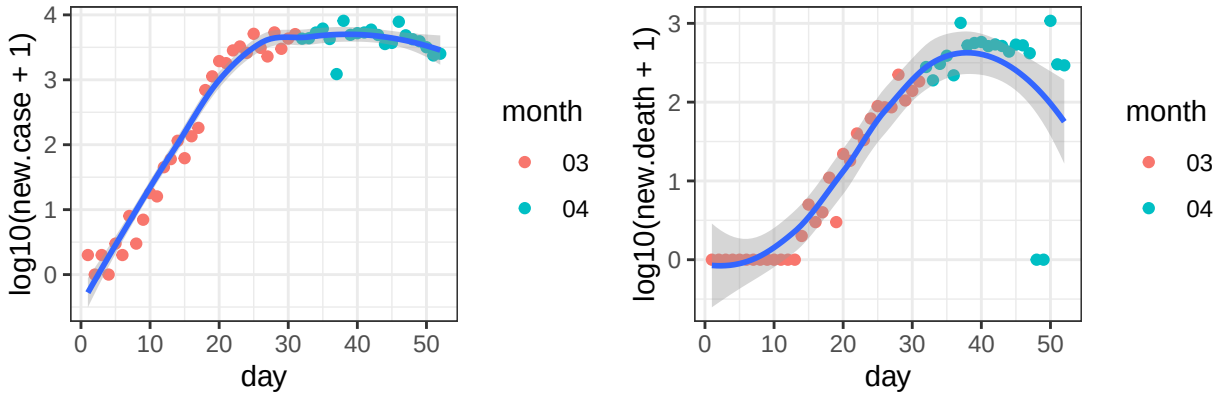
First check the 20 counties with the largest number of deaths.

```
##             date        county         state  fips  cases deaths
## 77415 2020-04-21 New York City      New York    NA 139335  10301
## 77414 2020-04-21        Nassau      New York 36059  31079   1717
## 76973 2020-04-21         Wayne      Michigan 26163  14255   1278
```

```
## 76334 2020-04-21          Cook      Illinois 17031  23181   1002
## 77434 2020-04-21       Suffolk      New York 36103  28154    918
## 77442 2020-04-21   Westchester      New York 36119  24655    904
## 77344 2020-04-21         Essex    New Jersey 34013  11128    849
## 77339 2020-04-21        Bergen    New Jersey 34003  13356    835
## 75950 2020-04-21   Los Angeles    California  6037  15140    663
## 76043 2020-04-21     Fairfield   Connecticut  9001   8472    544
## 77346 2020-04-21        Hudson    New Jersey 34017  11636    525
## 76954 2020-04-21       Oakland      Michigan 26125   6306    506
## 76941 2020-04-21        Macomb      Michigan 26099   4544    445
## 76888 2020-04-21     Middlesex Massachusetts 25017   9621    428
## 77357 2020-04-21         Union    New Jersey 34039  10289    427
## 76044 2020-04-21       Hartford   Connecticut  9003   3951    402
## 77808 2020-04-21   Philadelphia  Pennsylvania 42101  10028    394
## 78392 2020-04-21           King    Washington 53033   5381    374
## 77349 2020-04-21     Middlesex    New Jersey 34023   8767    360
## 76808 2020-04-21       Orleans     Louisiana 22071   6169    344
```

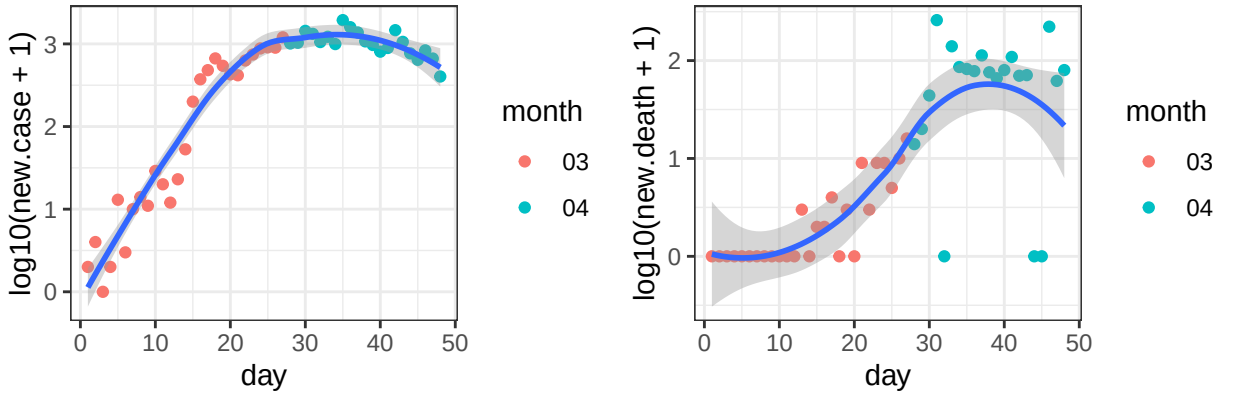
For these 20 counties, I check the number of new cases and the number of new deaths.

New York City_New York



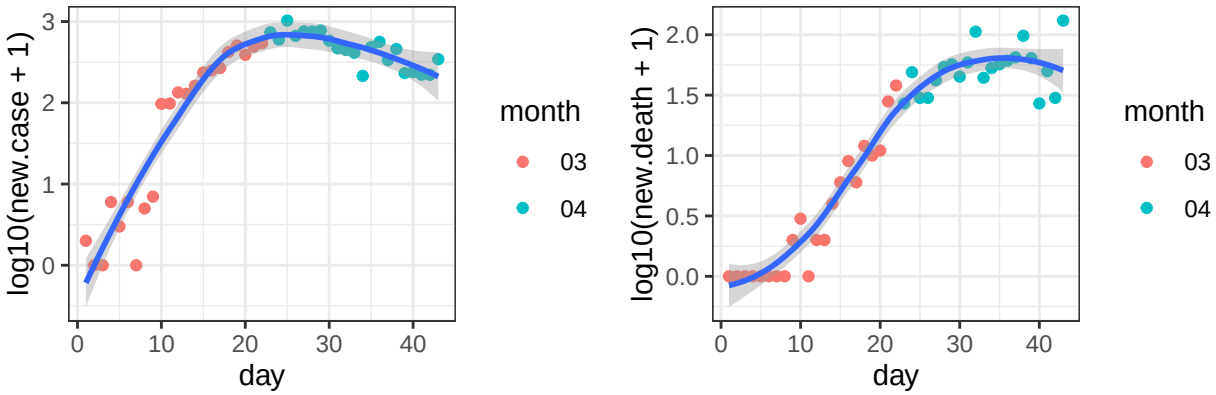
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Nassau_New York



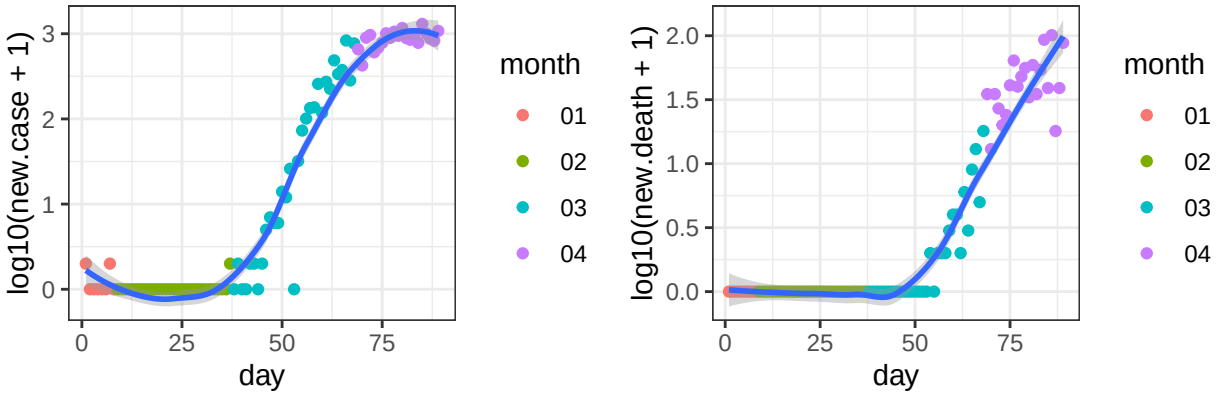
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Wayne_Michigan



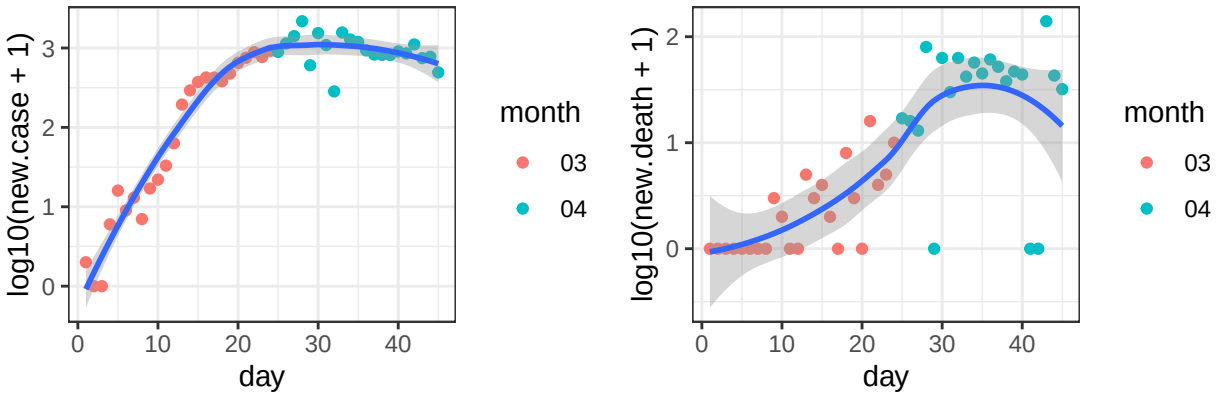
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Cook_Illinois



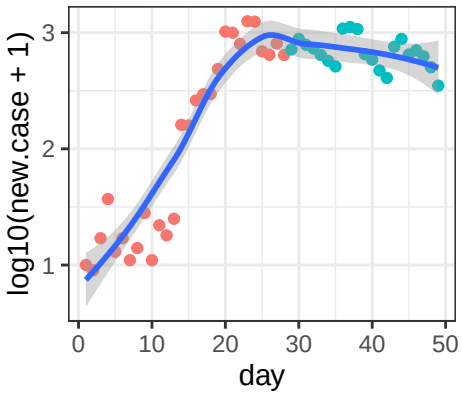
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

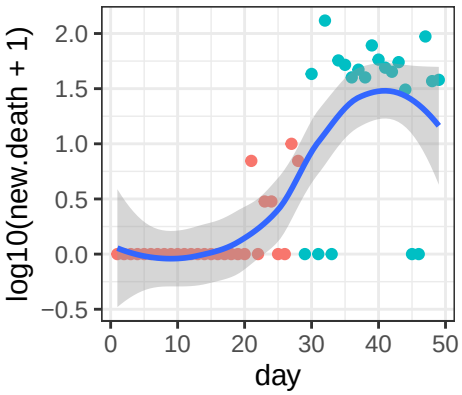
Suffolk_New York



data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

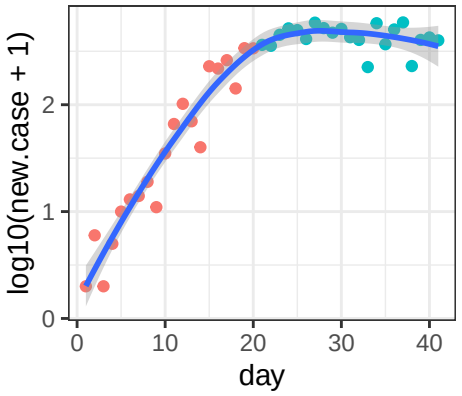
Westchester_New York

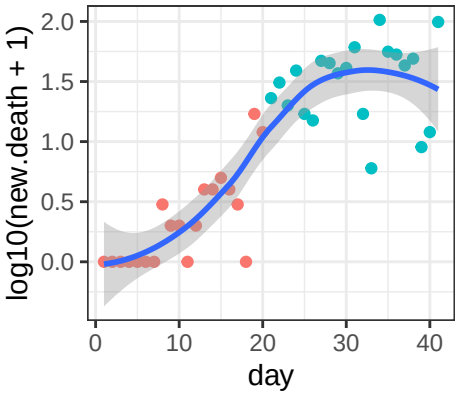




data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

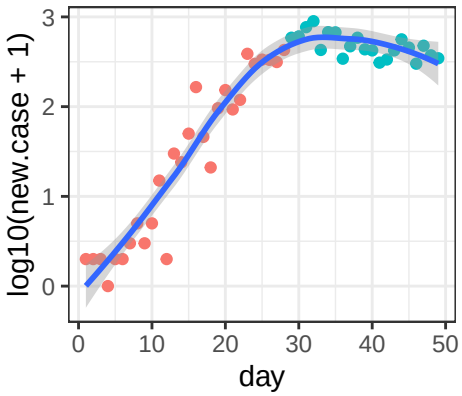
Essex_New Jersey

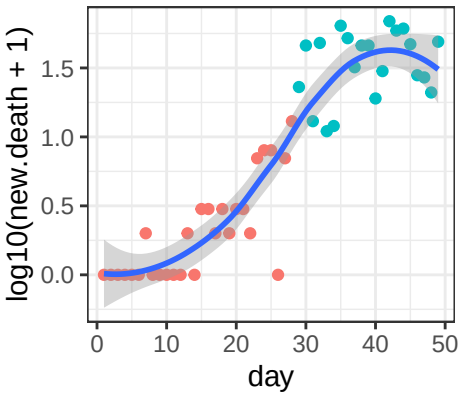




data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

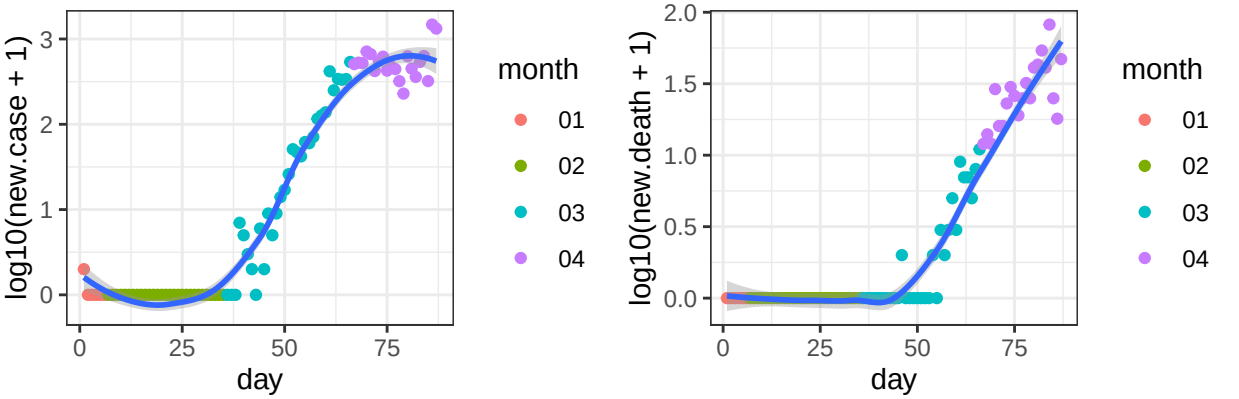
Bergen_New Jersey





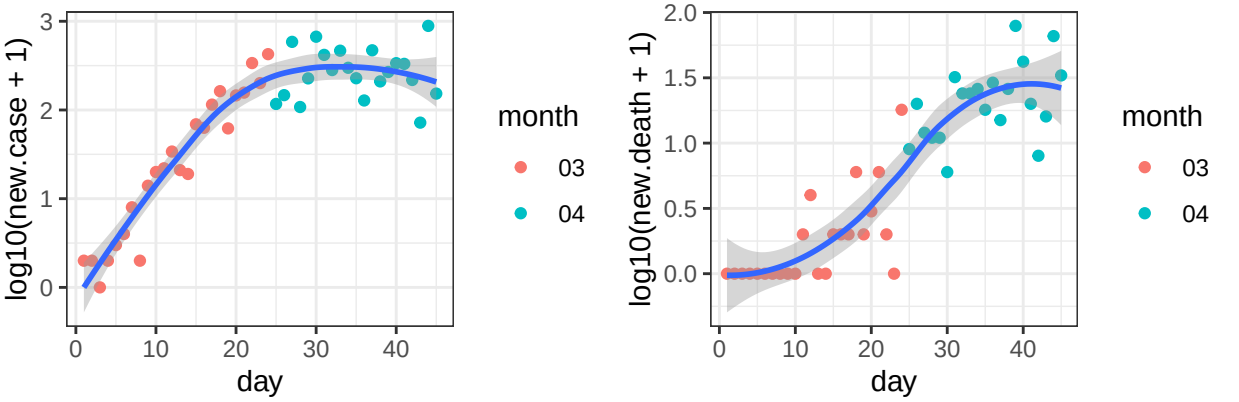
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Los Angeles_California



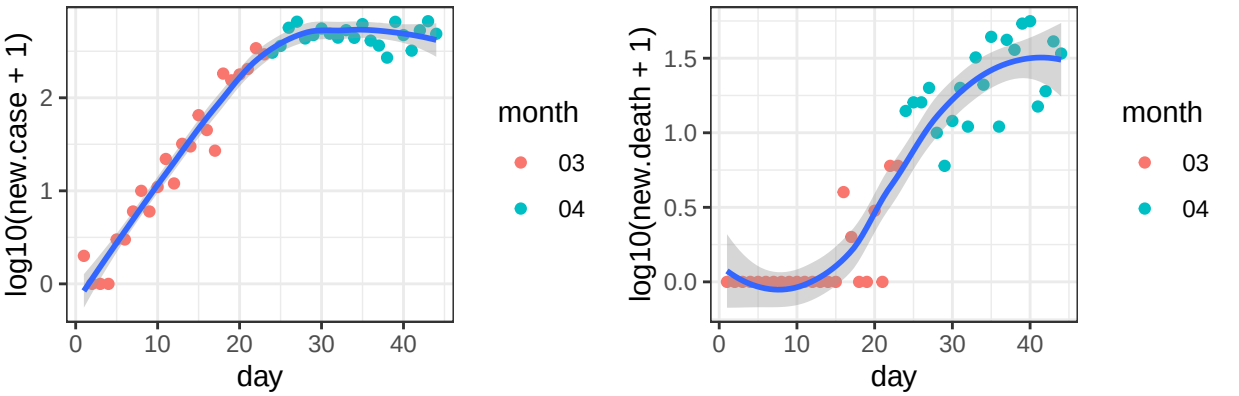
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Fairfield_Connecticut



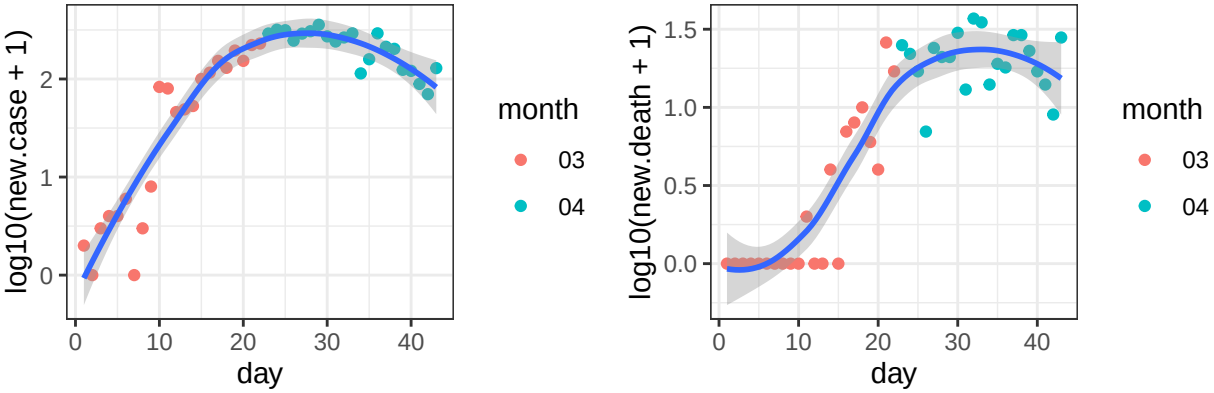
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Hudson_New Jersey



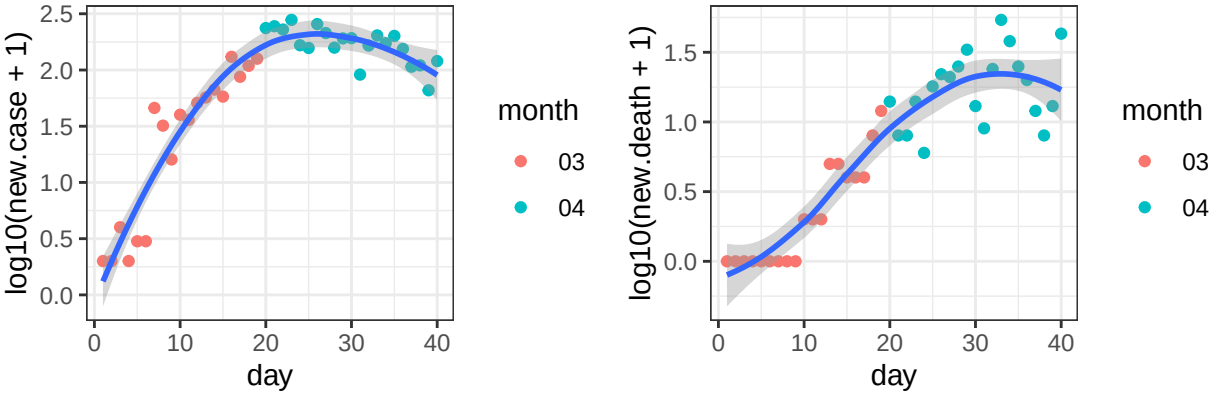
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Oakland_Michigan



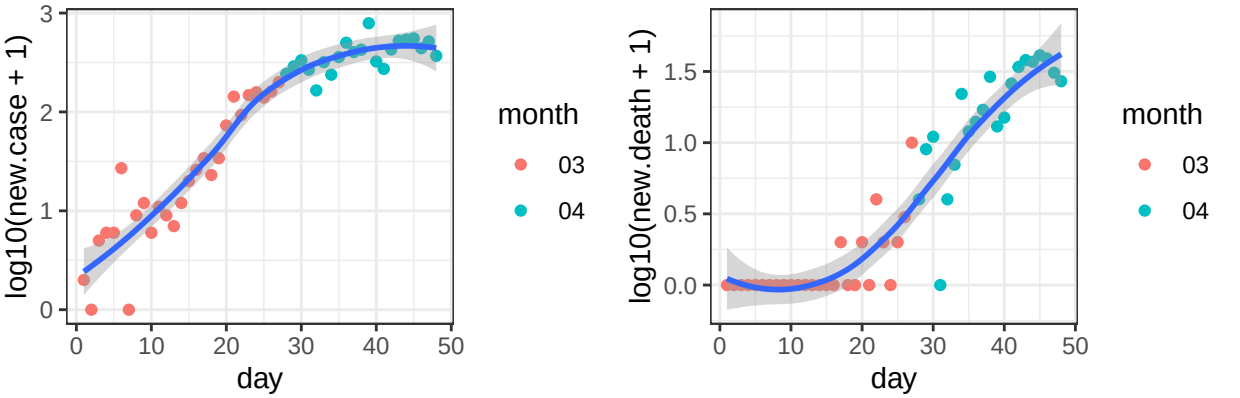
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Macomb_Michigan



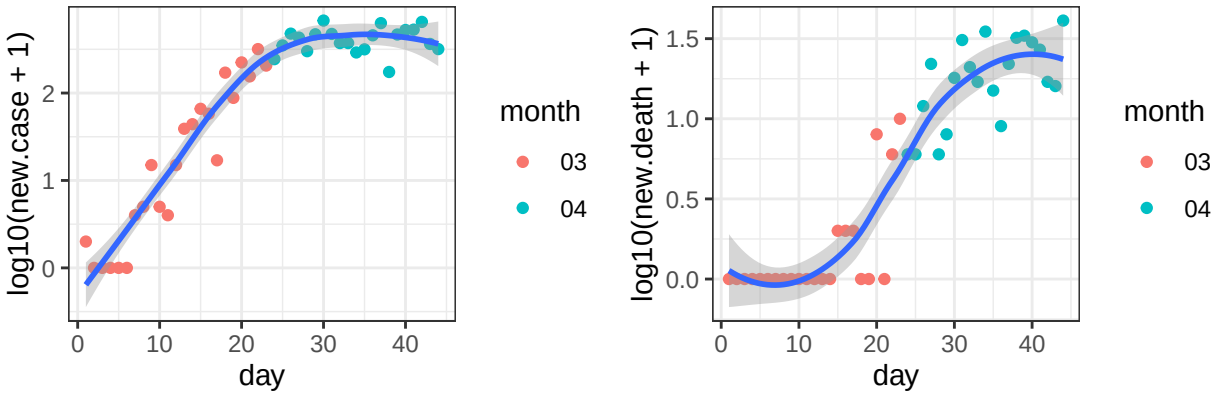
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Middlesex_Massachusetts



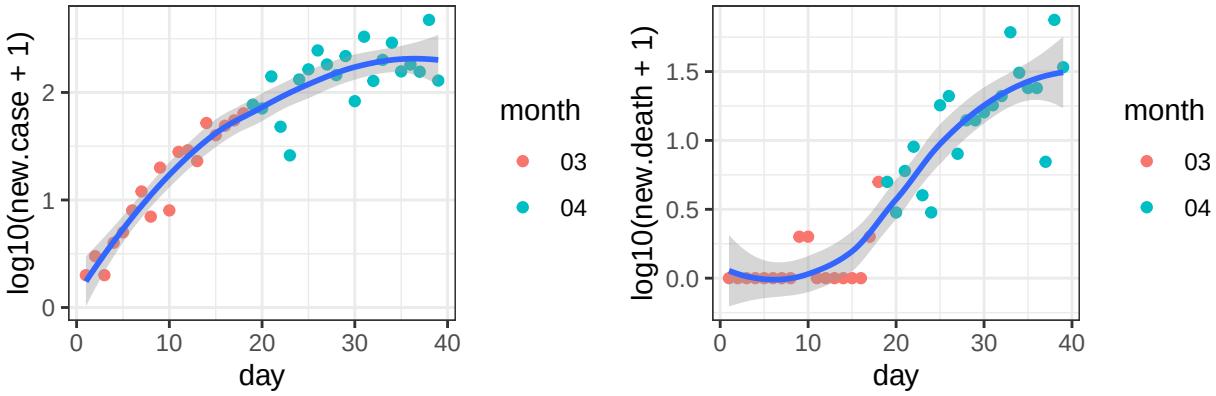
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Union_New Jersey



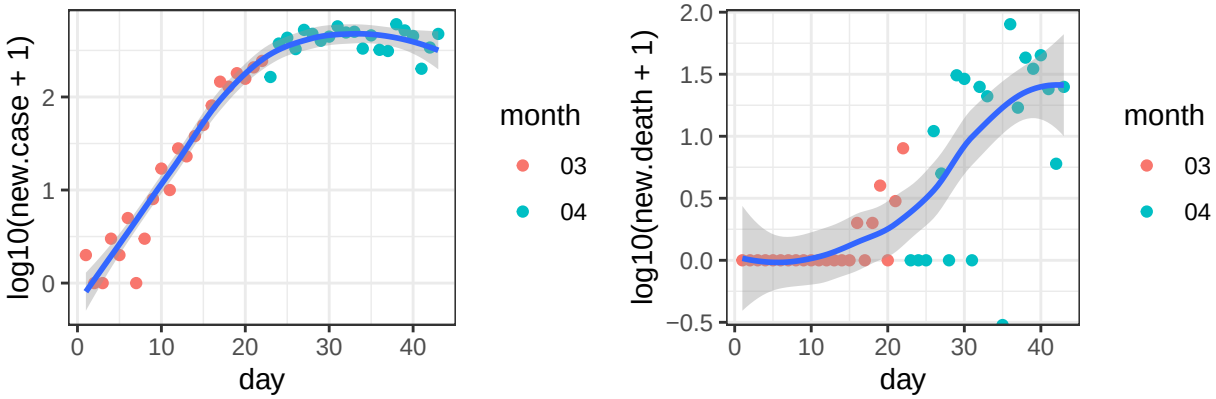
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Hartford_Connecticut



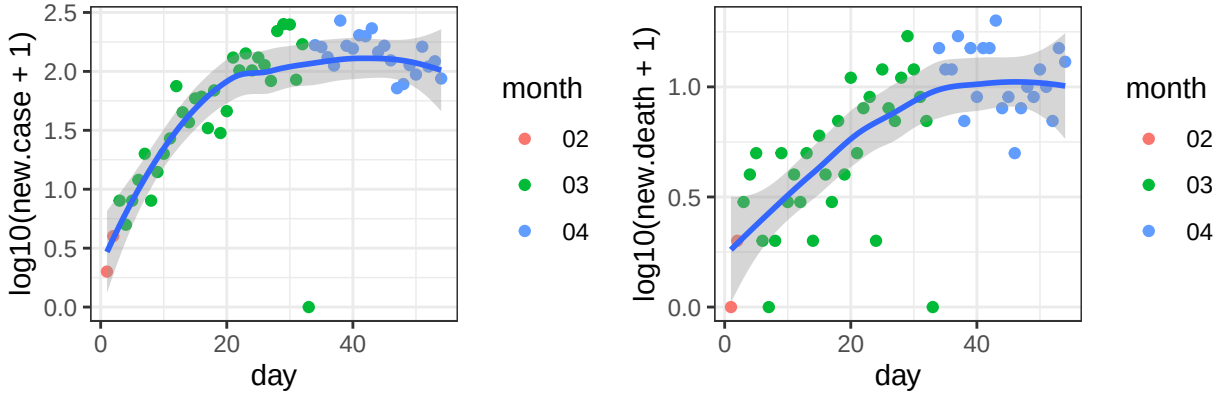
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Philadelphia_Pennsylvania



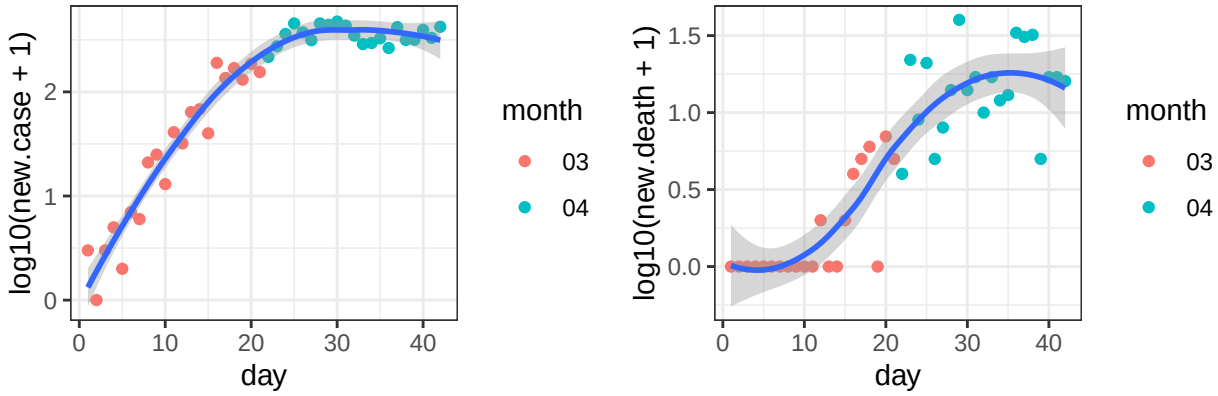
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

King_Washington



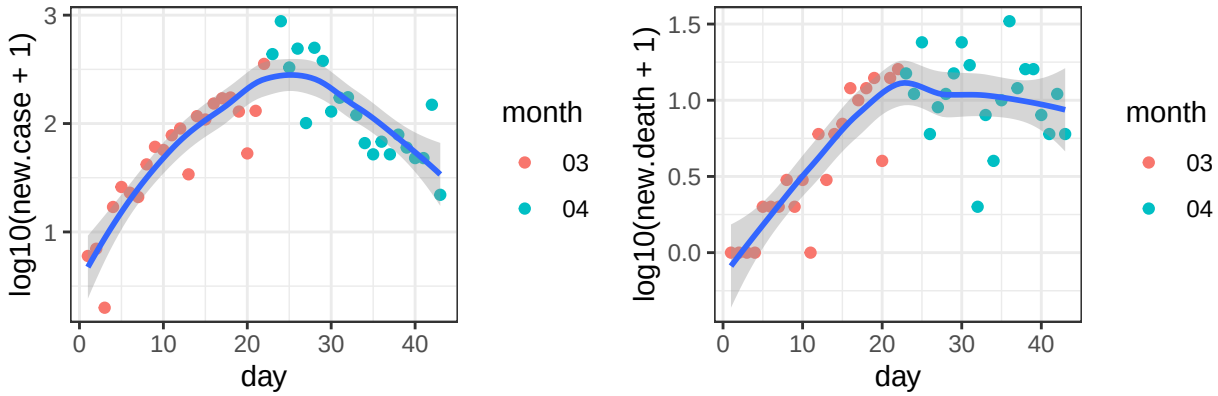
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−28

Middlesex_New Jersey



data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Orleans_Louisiana



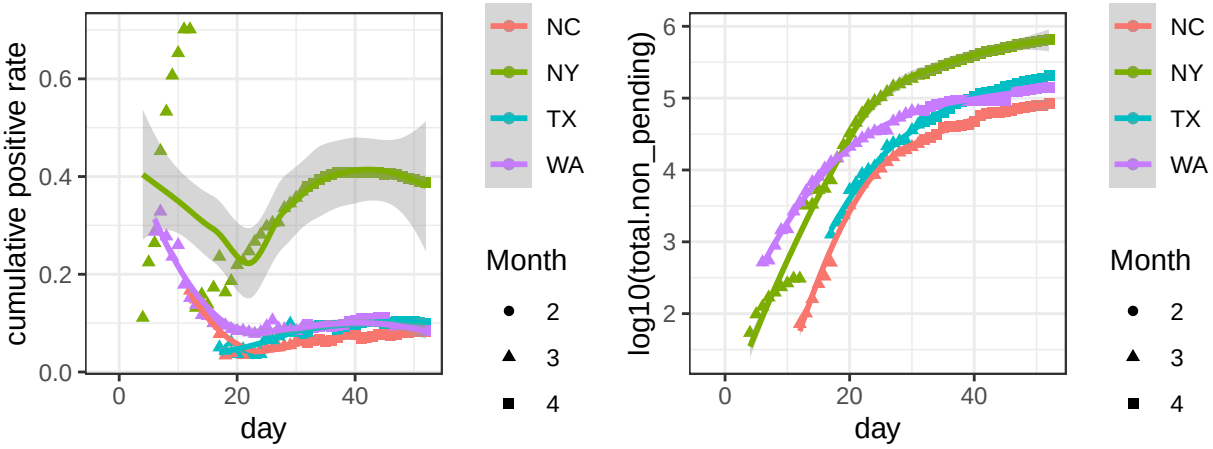
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

COVID Trackng

The positive rates of testing can be an indicator on how much the COVID-19 has spread. However, they are more noisy data since the negative testing resutls are often not reported and the tests are almost surely taken on a non-representative random sample of the population. The COVID traking project proides a grade per state: “If you are calculating positive rates, it should only be with states that have an A grade. And be

careful going back in time because almost all the states have changed their level of reporting at different times." (https://covidtracking.com/about-tracker/). The data are also availalbe for both counties and states, here I only look at state level data.

Since the daily postive rate can fluctuate a lot, here I only illustrae the cumulative positave rate across time, for four states with grade A data. Of course since this is an R markdown file, you can modify the source code and check for other states.



github.com/COVID19Tracking/, cumulative positive rate on 0421: 0.09(WA) 0.10(TX) 0.39(NY) 0.08(NC)

Session information

```
sessionInfo()
```

```
## R version 3.6.2 (2019-12-12)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Catalina 10.15.4
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] httr_1.4.1    ggpubr_0.2.5  magrittr_1.5  ggplot2_3.2.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.3       pillar_1.4.3     compiler_3.6.2   tools_3.6.2
##  [5] digest_0.6.23    evaluate_0.14    lifecycle_0.1.0  tibble_2.1.3
##  [9] gtable_0.3.0     pkgconfig_2.0.3  rlang_0.4.4      yaml_2.2.1
## [13] xfun_0.12        gridExtra_2.3    withr_2.1.2      dplyr_0.8.4
## [17] stringr_1.4.0    knitr_1.28       grid_3.6.2       tidyselect_1.0.0
## [21] cowplot_1.0.0    glue_1.3.1       R6_2.4.1         rmarkdown_2.1
## [25] purrr_0.3.3      farver_2.0.3     scales_1.1.0     htmltools_0.4.0
## [29] assertthat_0.2.1 colorspace_1.4-1 ggsignif_0.6.0   labeling_0.3
```

```
## [33] stringi_1.4.5    lazyeval_0.2.2   munsell_0.5.0    crayon_1.3.4
```