Exploration of COVID-19 tracking data from multiple resources

Wei Sun

2020-05-16

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

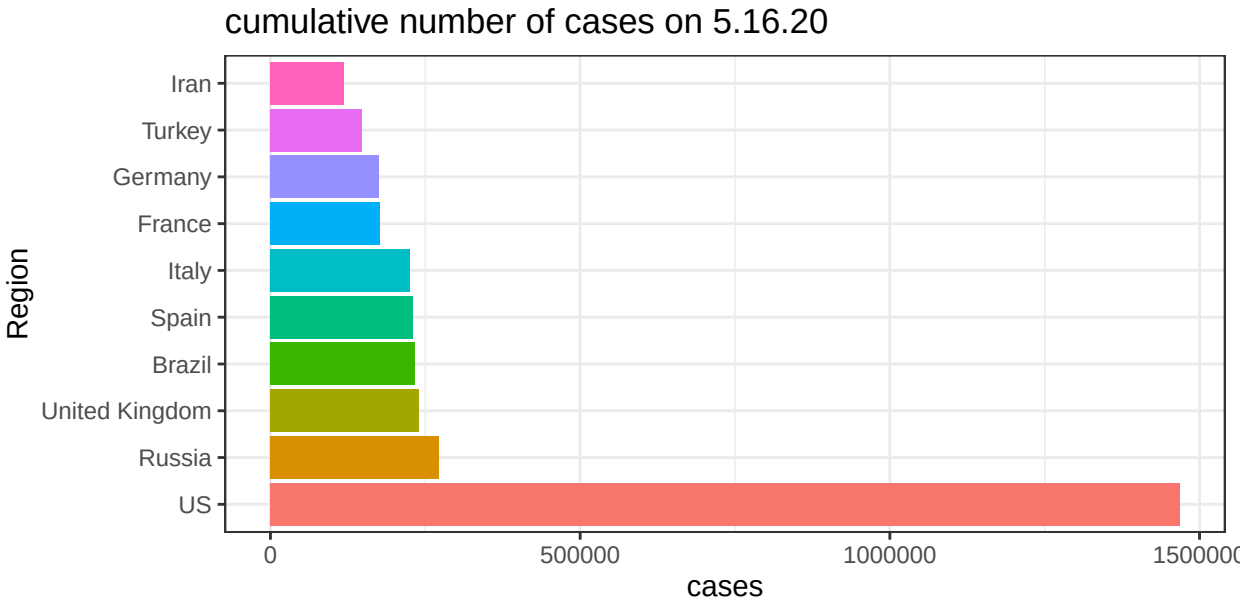
JHU

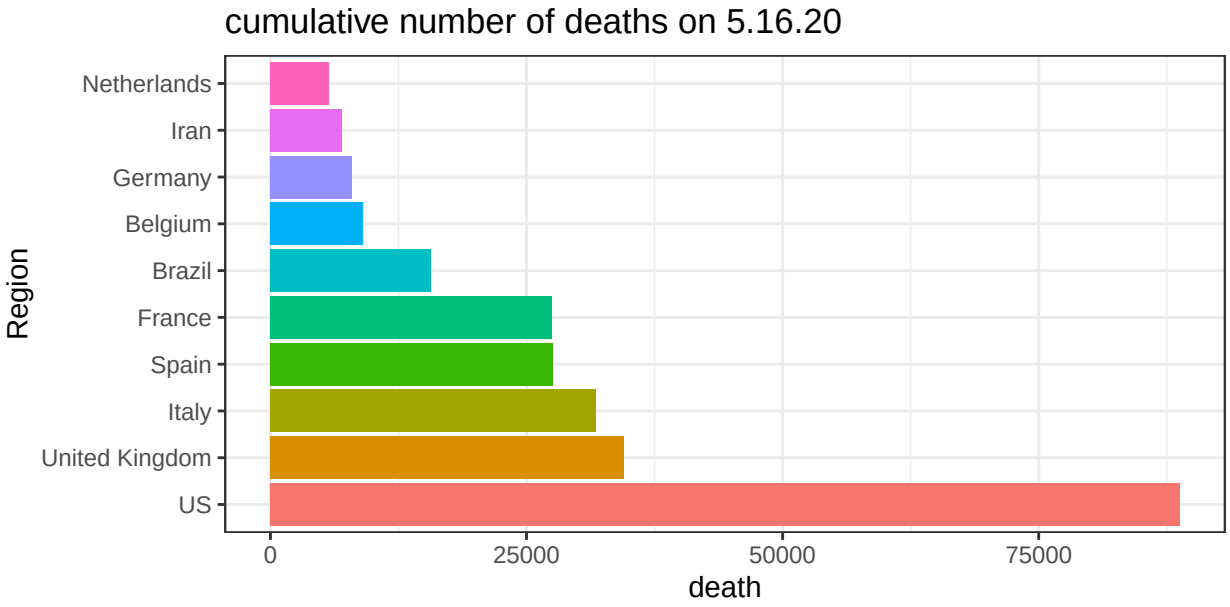
Assume you have cloned the JHU Github repository on your local machine at "../COVID-19".

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.

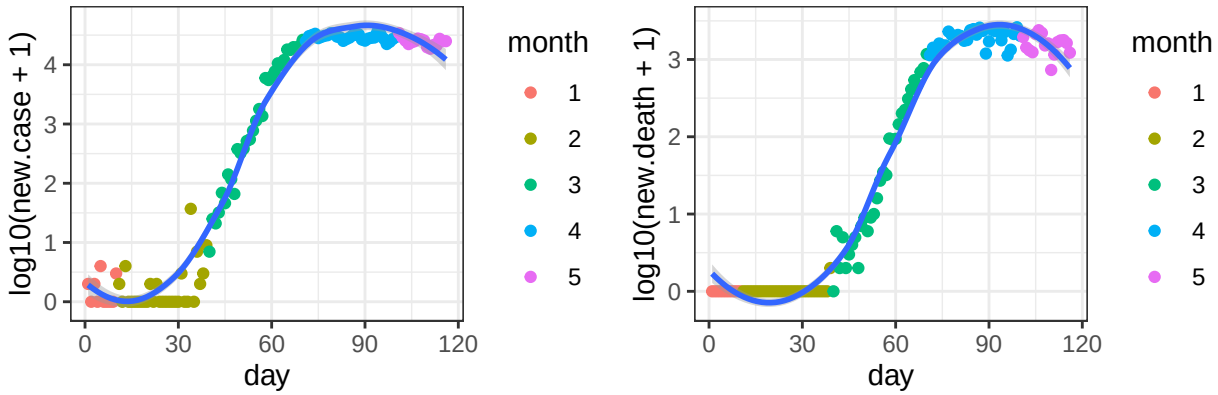
Here is the list of 10 records with the largest number of cases or deaths on the most recent date.





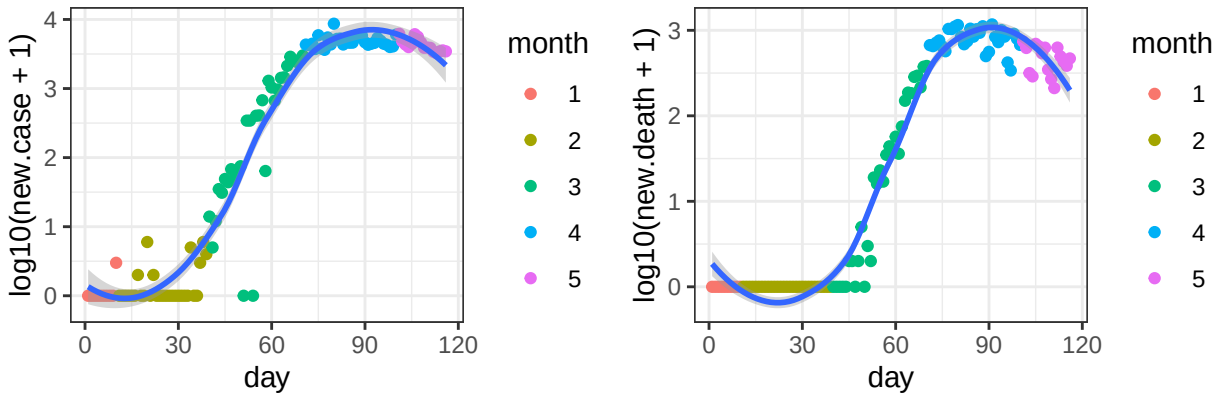
Next, I check for each country/region, what is the number of new cases/deaths? This data is important to understand what is the trend under different situations, e.g., population density, social distance policies etc. Here I checked the top 10 countries/regions with the highest number of deaths.

US



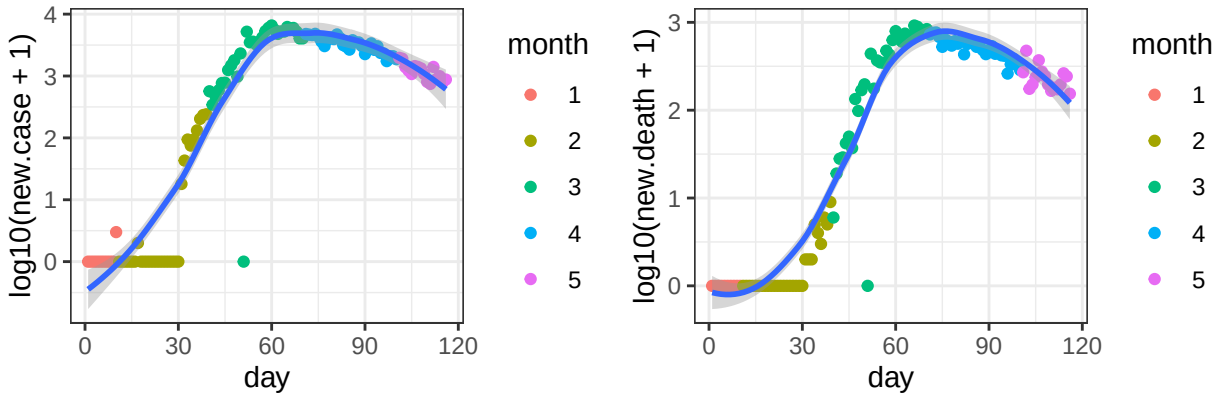
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

United Kingdom



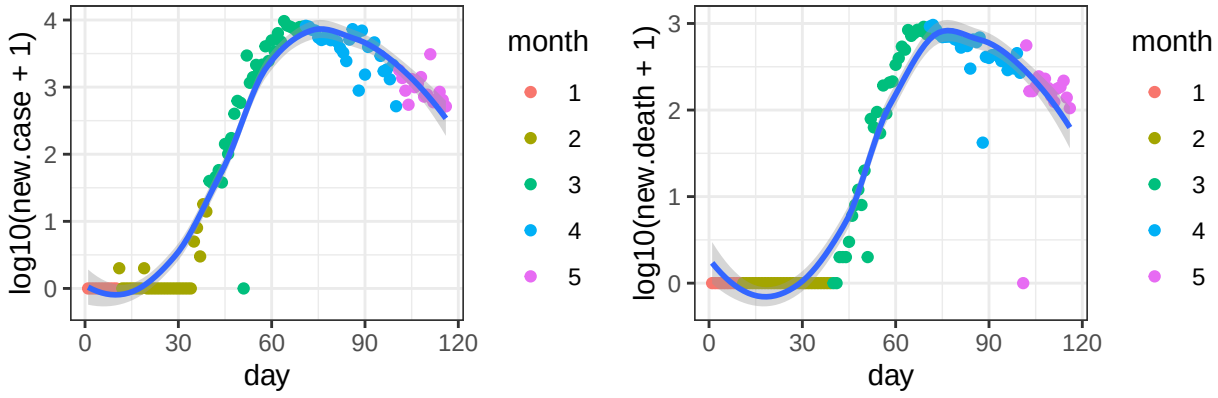
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Italy



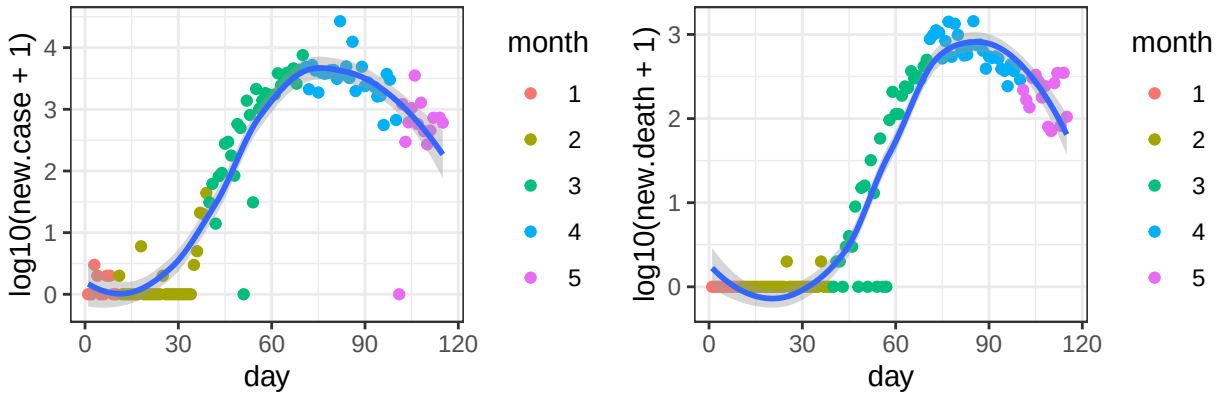
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Spain



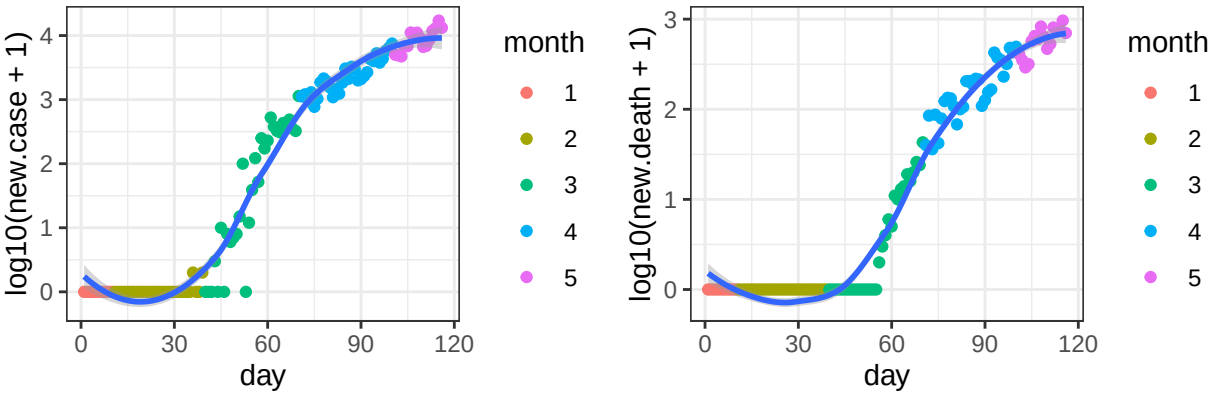
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

France



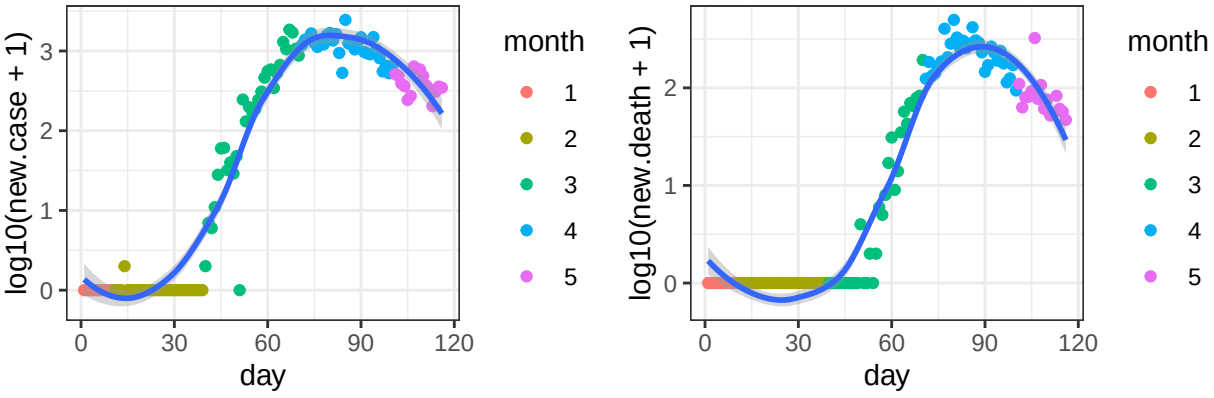
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Brazil



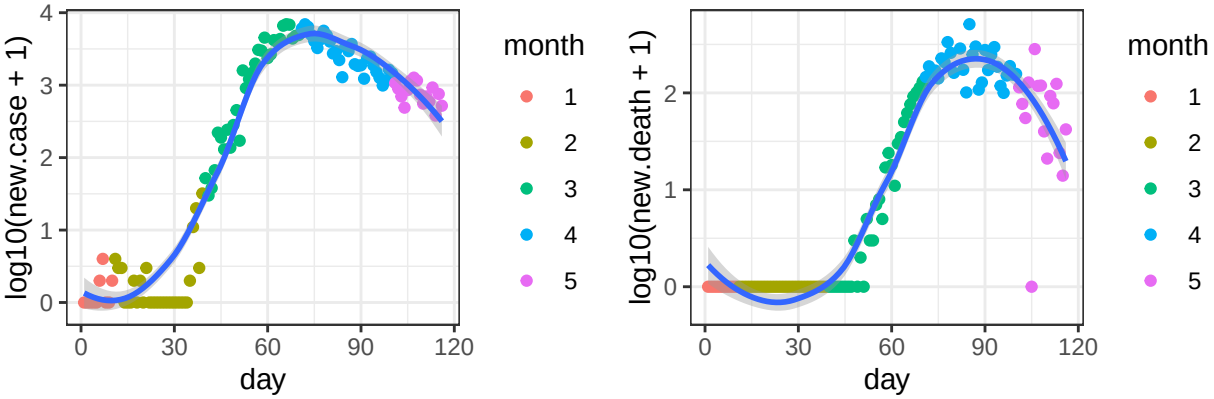
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Belgium

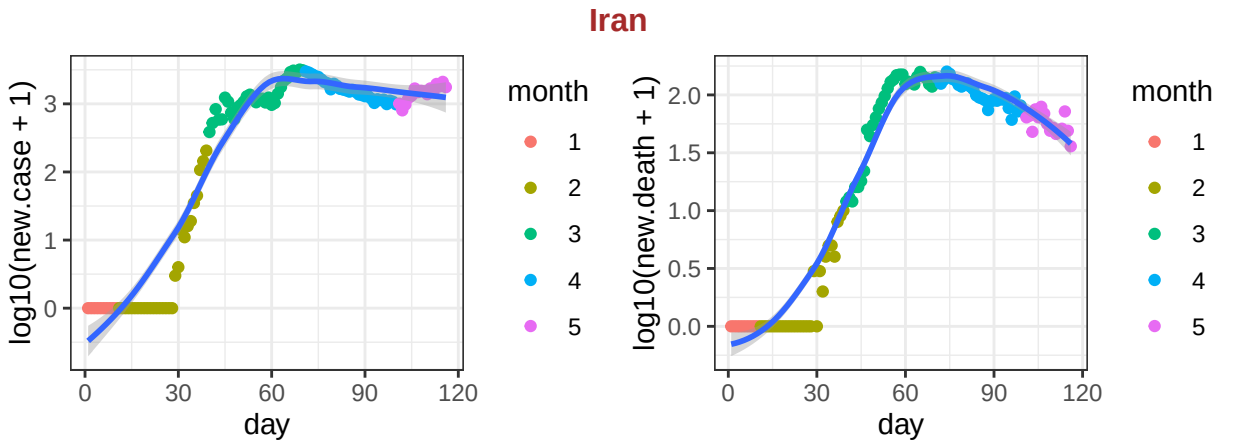


data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

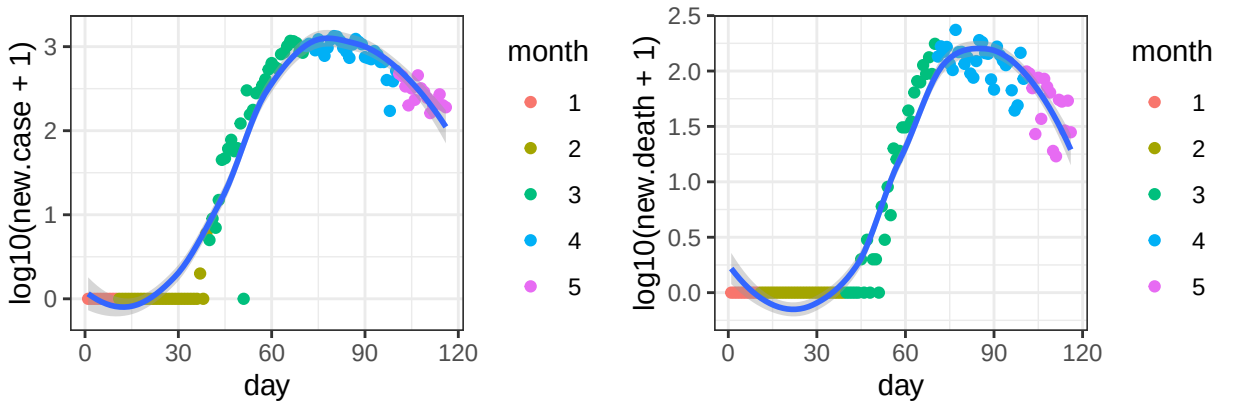
Germany



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



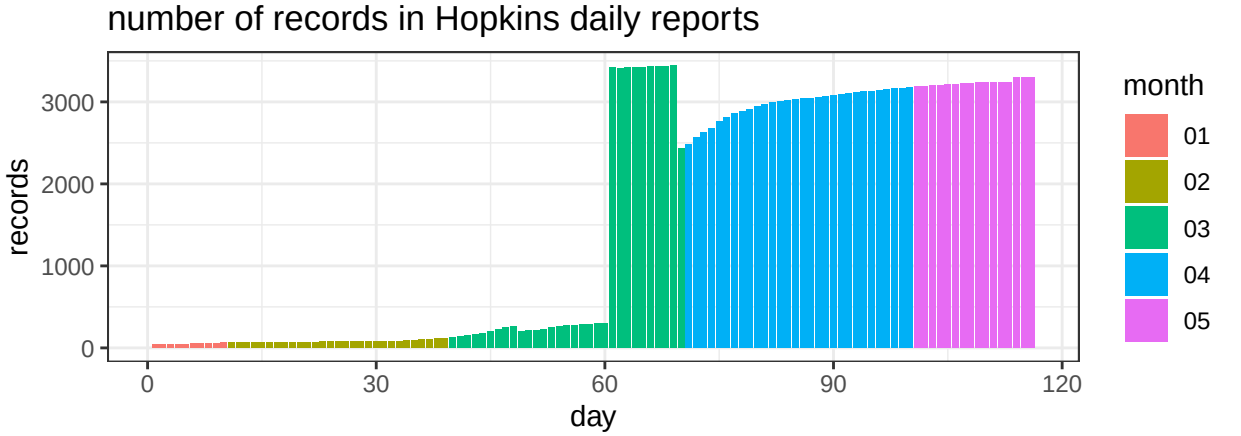
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020



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] "2020-05-15"
```

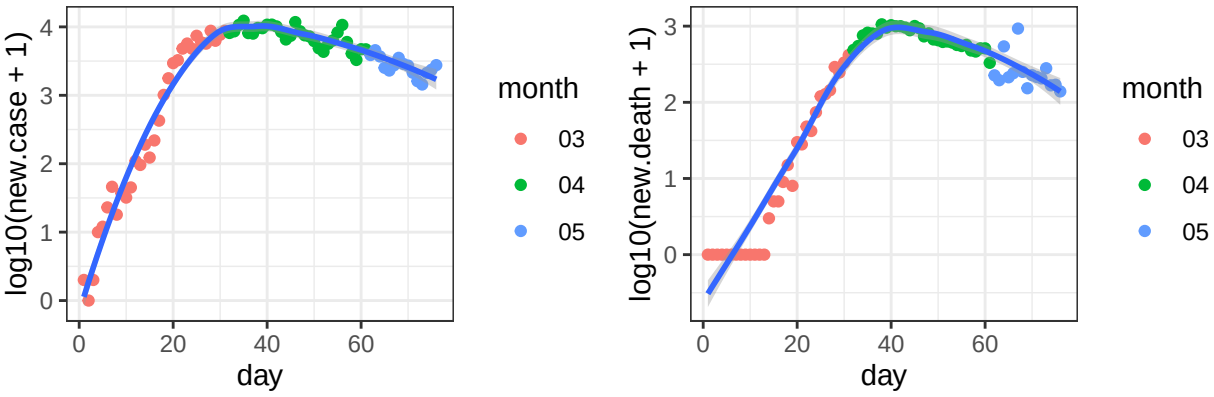
state level data

First check the 30 states with the largest number of deaths.

```
##            date                state fips  cases deaths
## 4063 2020-05-15             New York   36 350951  27755
## 4061 2020-05-15           New Jersey   34 143905  10138
## 4052 2020-05-15        Massachusetts   25  83421   5592
## 4053 2020-05-15             Michigan   26  49982   4825
## 4070 2020-05-15         Pennsylvania   42  64178   4432
## 4044 2020-05-15             Illinois   17  90529   4075
## 4036 2020-05-15          Connecticut    9  36085   3285
## 4034 2020-05-15           California    6  77015   3192
## 4049 2020-05-15            Louisiana   22  33837   2382
## 4039 2020-05-15              Florida   12  44130   1916
## 4051 2020-05-15             Maryland   24  37105   1911
## 4045 2020-05-15              Indiana   18  27281   1691
## 4067 2020-05-15                 Ohio   39  26956   1581
## 4040 2020-05-15              Georgia   13  35242   1563
## 4076 2020-05-15                Texas   48  46987   1300
## 4035 2020-05-15             Colorado    8  21207   1150
## 4081 2020-05-15           Washington   53  19230   1008
## 4080 2020-05-15             Virginia   51  28672    977
## 4054 2020-05-15            Minnesota   27  14249    692
## 4064 2020-05-15       North Carolina   37  17190    660
## 4032 2020-05-15              Arizona    4  13169    651
## 4056 2020-05-15             Missouri   29  10567    581
## 4055 2020-05-15          Mississippi   28  10801    493
## 4030 2020-05-15              Alabama    1  11373    483
## 4072 2020-05-15         Rhode Island   44  12219    479
## 4083 2020-05-15            Wisconsin   55  11854    445
## 4073 2020-05-15       South Carolina   45   8407    380
## 4038 2020-05-15 District of Columbia   11   6871    368
## 4059 2020-05-15               Nevada   32   6744    345
## 4048 2020-05-15             Kentucky   21   7578    343
```

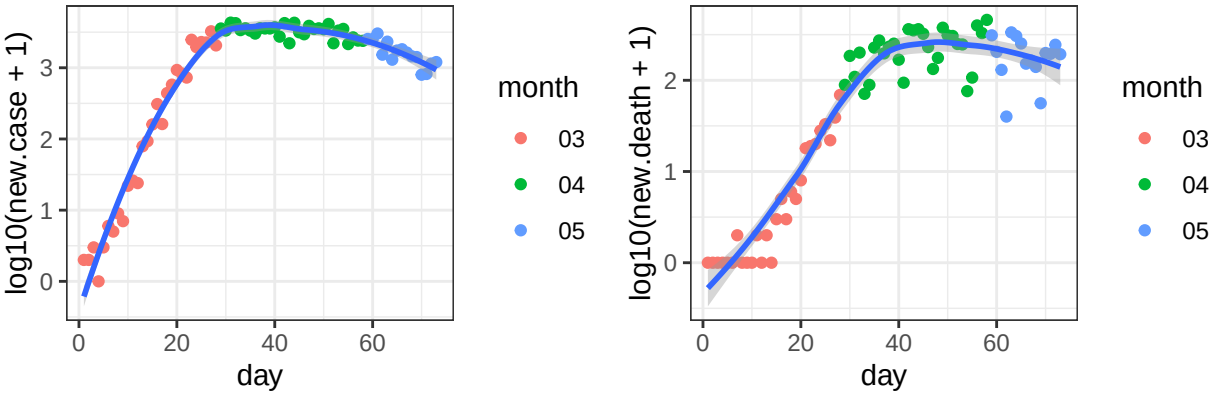
For these 20 states, I check the number of new cases and the number of new deaths. Part of the reason for such checking is to identify whether there is any similarity on such patterns. For example, could you use the pattern seen from Italy to predict what happen in an individual state, and what are the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

New York



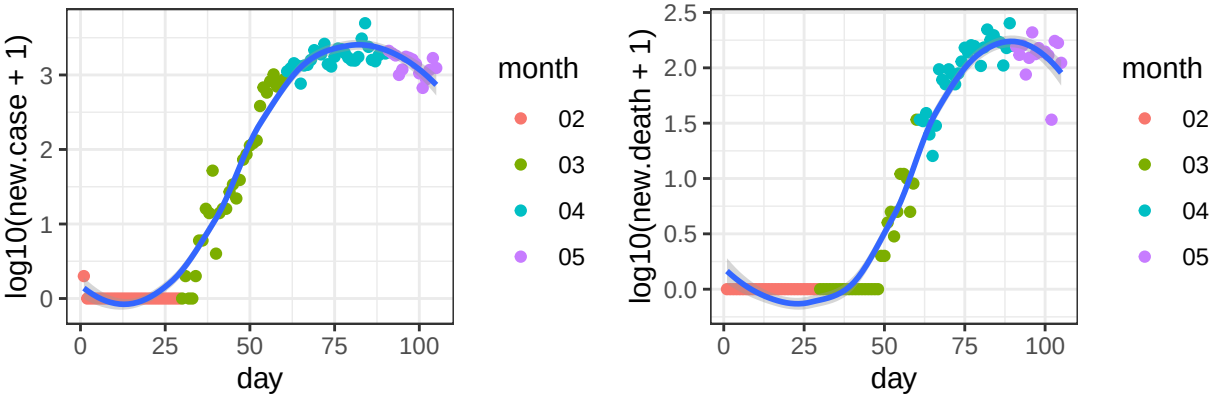
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

New Jersey



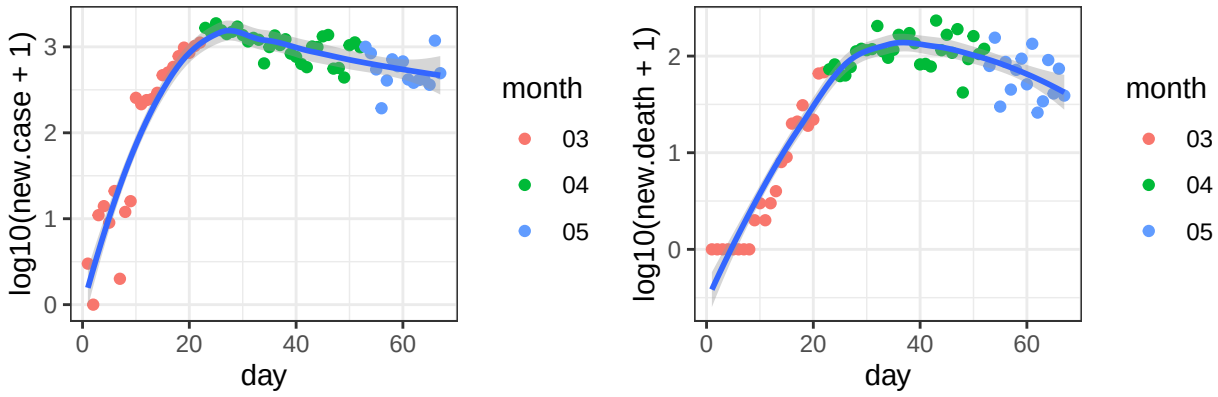
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Massachusetts



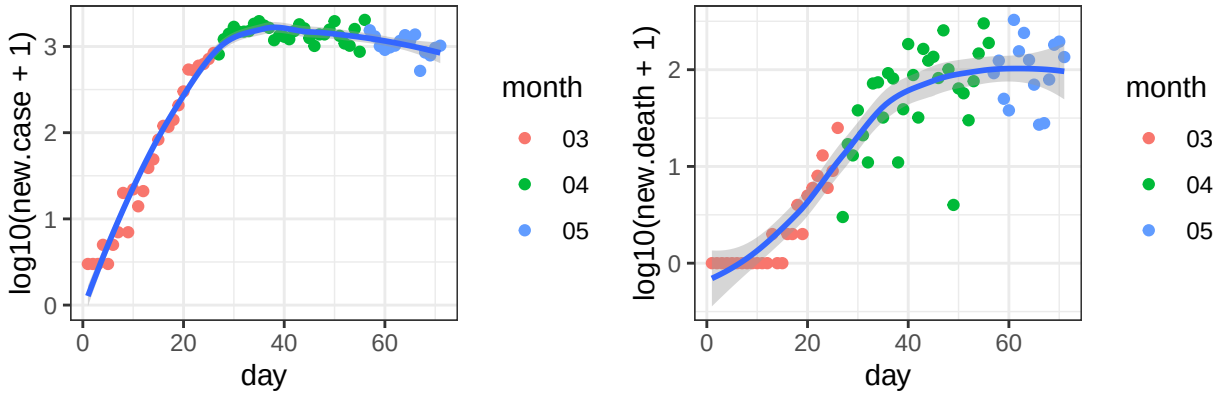
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01

Michigan



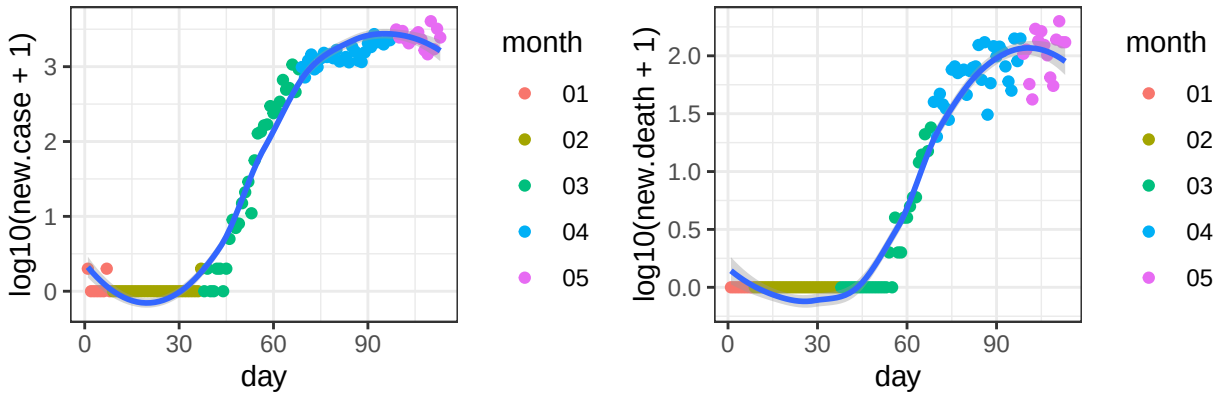
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Pennsylvania



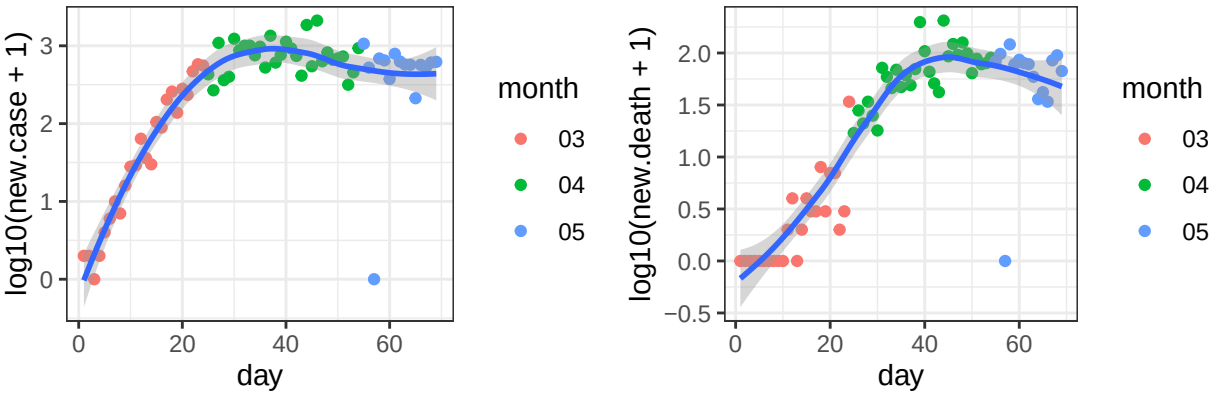
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Illinois



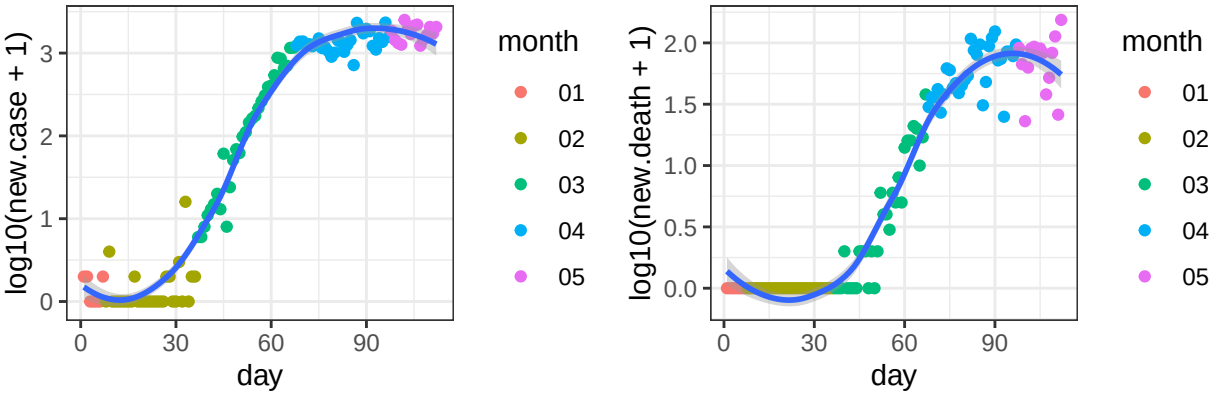
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

Connecticut



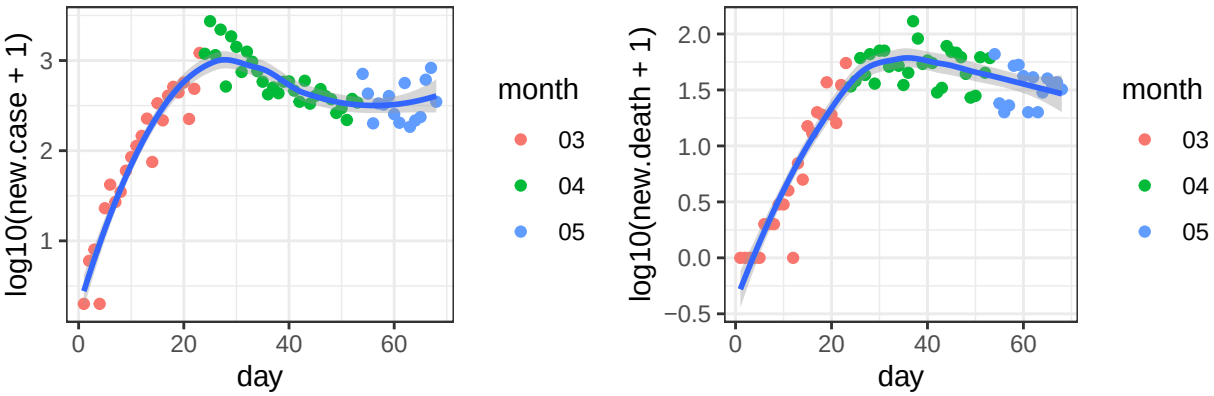
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

California



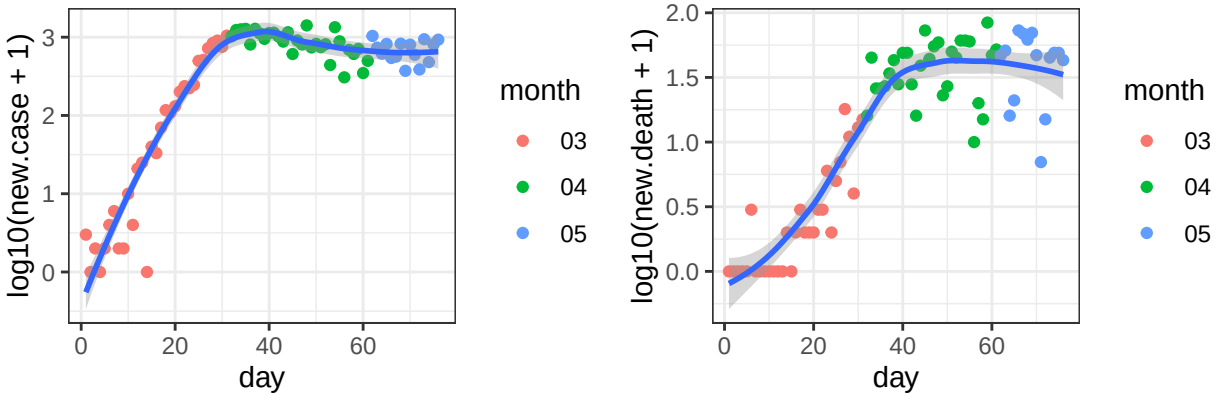
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−25

Louisiana



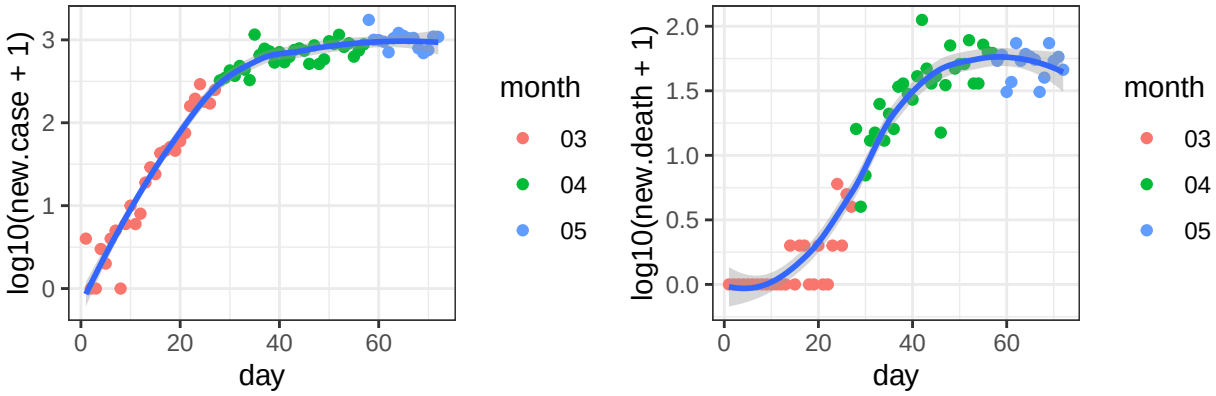
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Florida



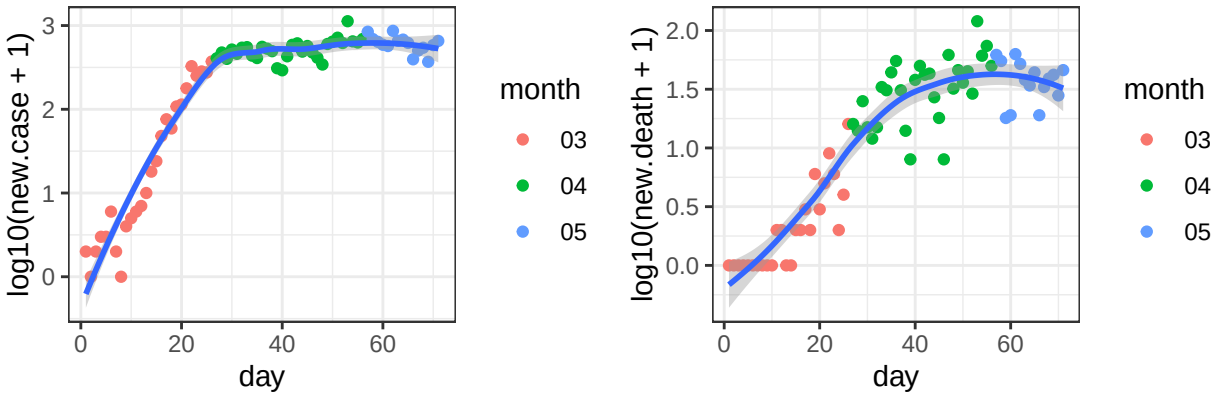
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Maryland



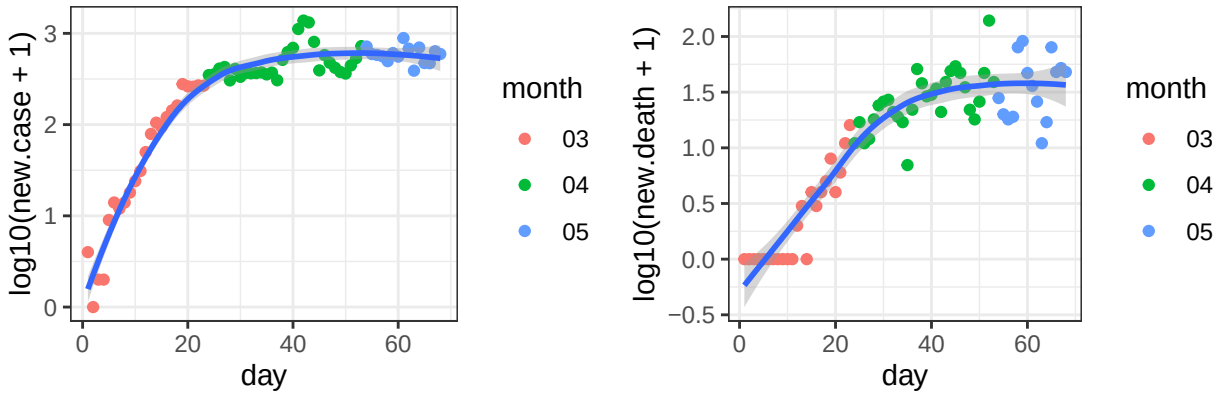
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Indiana



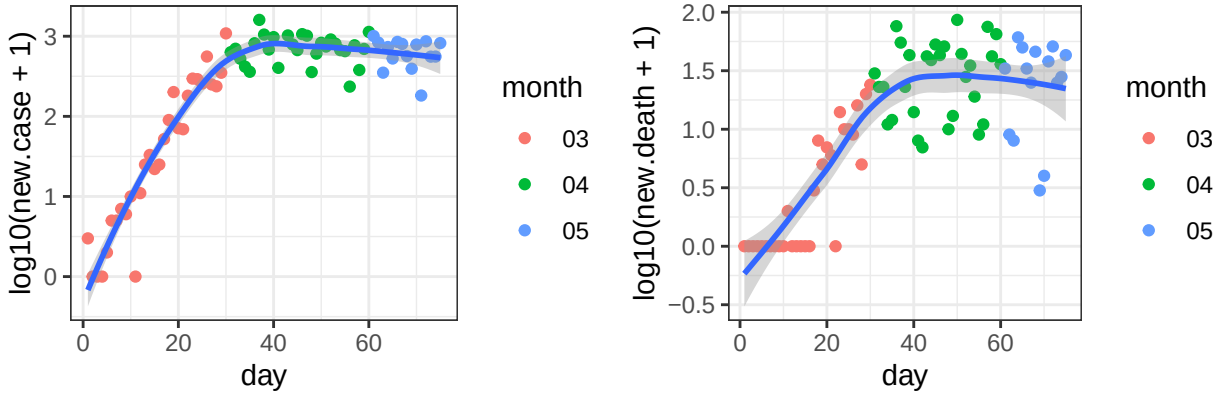
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Ohio



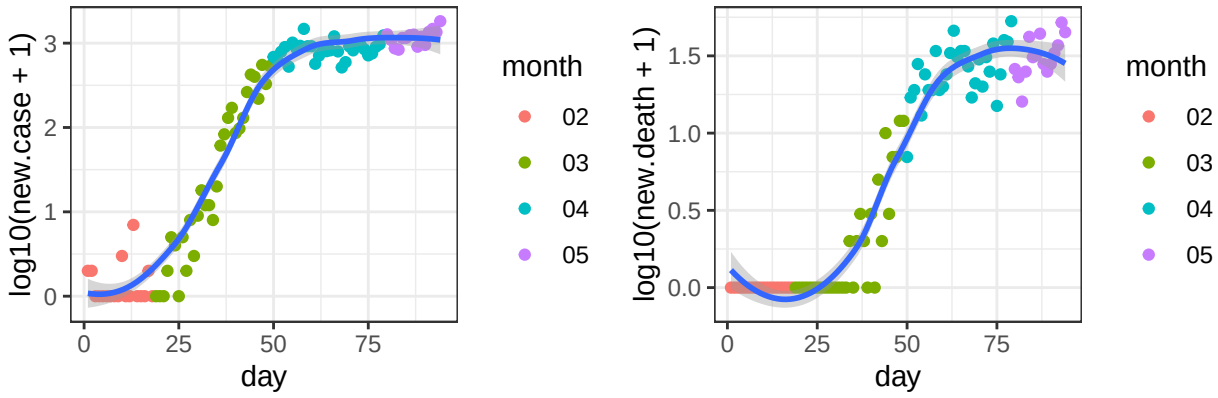
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Georgia



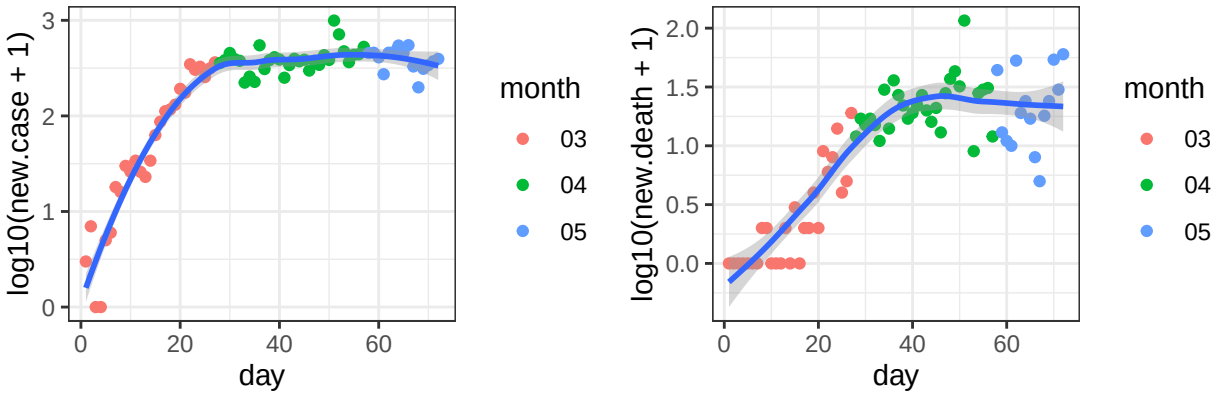
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Texas



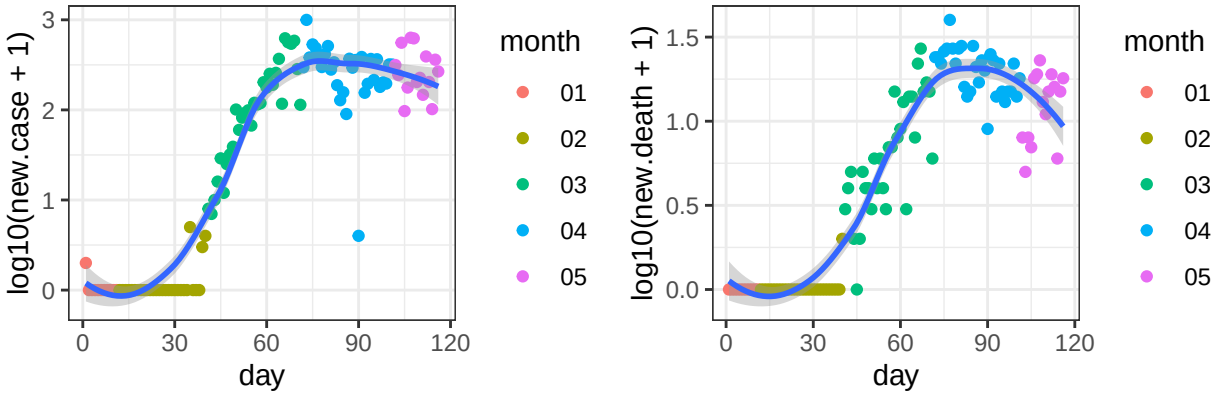
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−12

Colorado



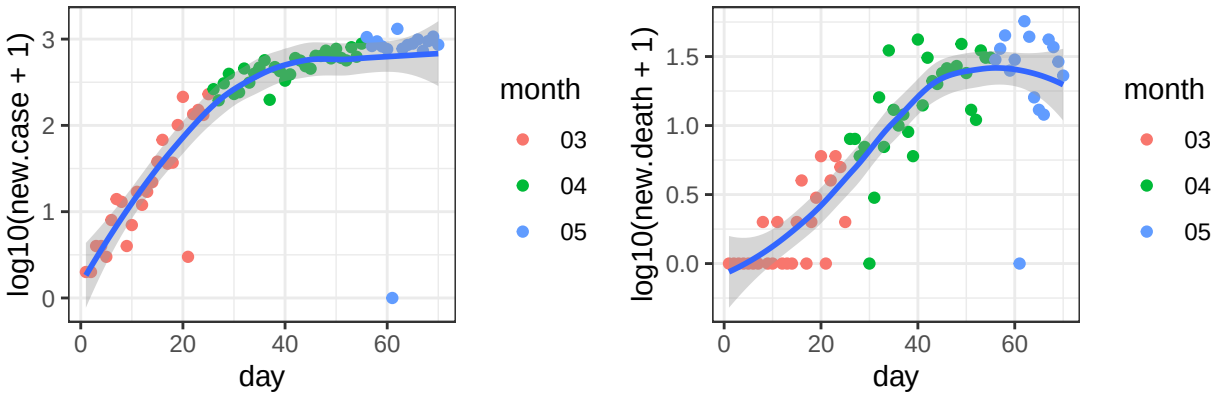
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Washington



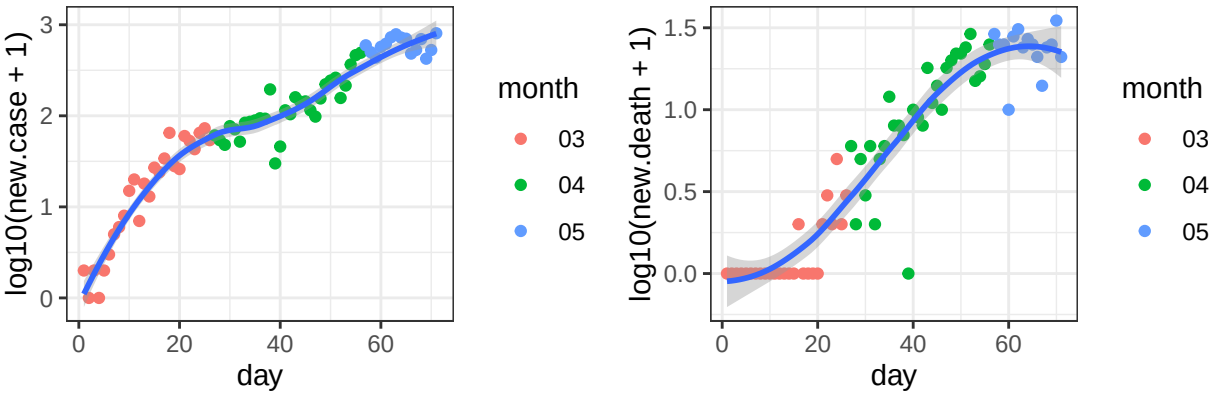
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−21

Virginia



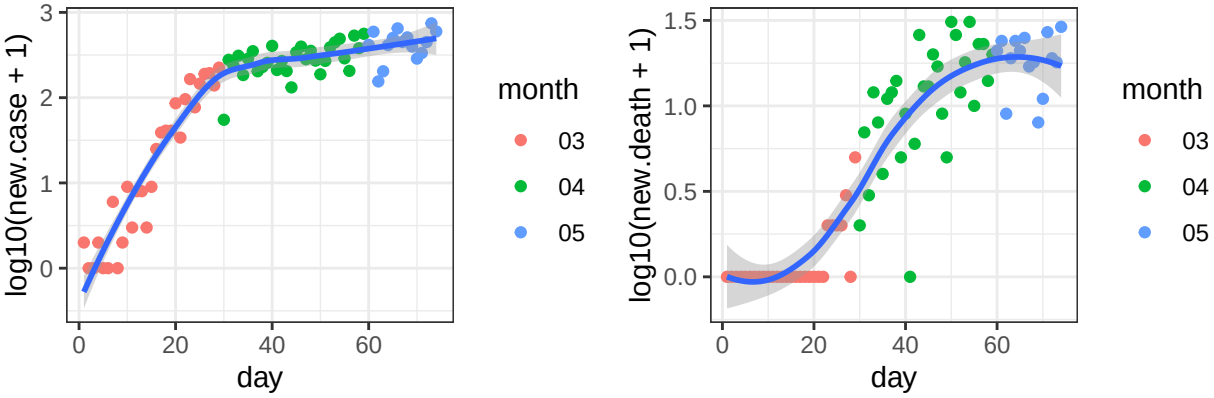
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Minnesota



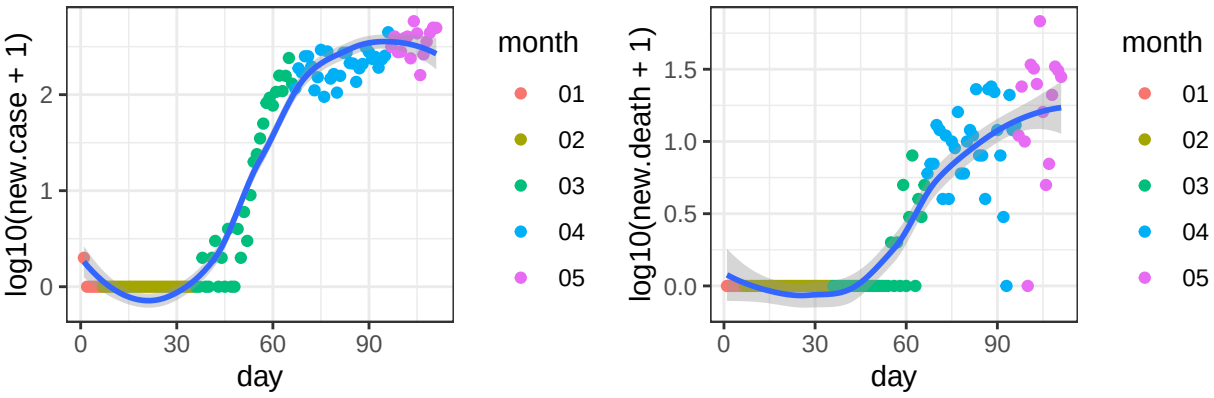
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

North Carolina



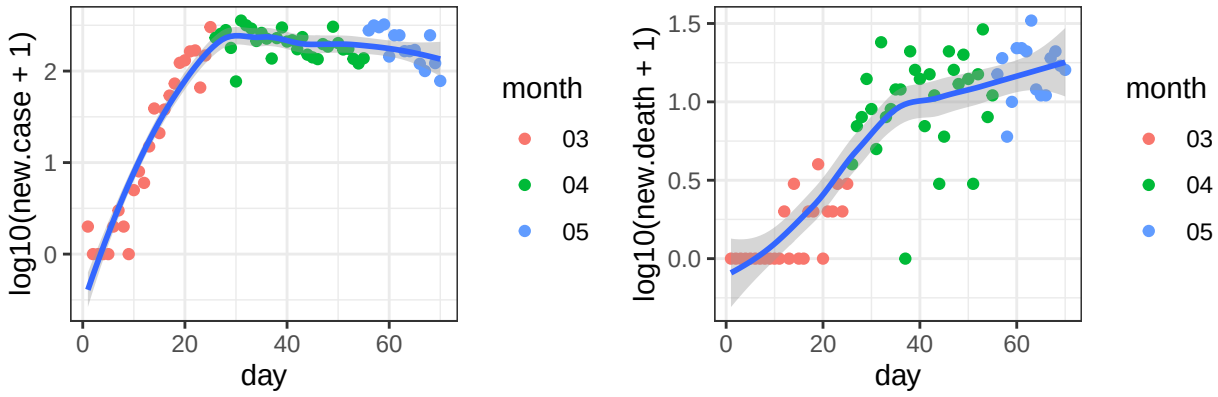
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−03

Arizona



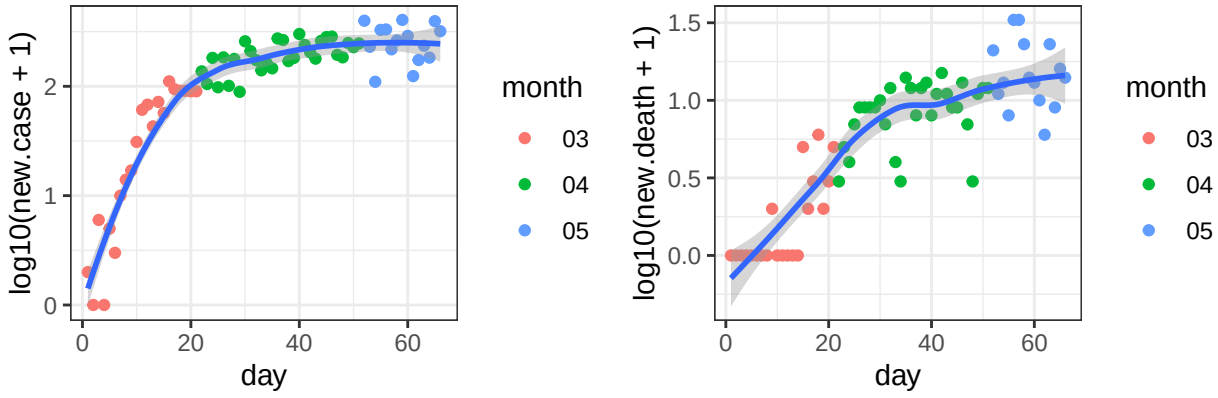
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Missouri



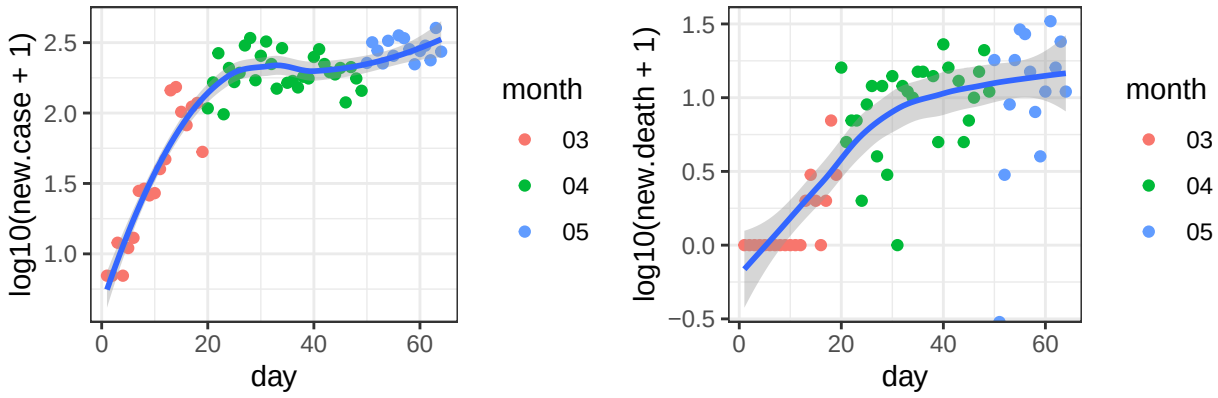
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Mississippi



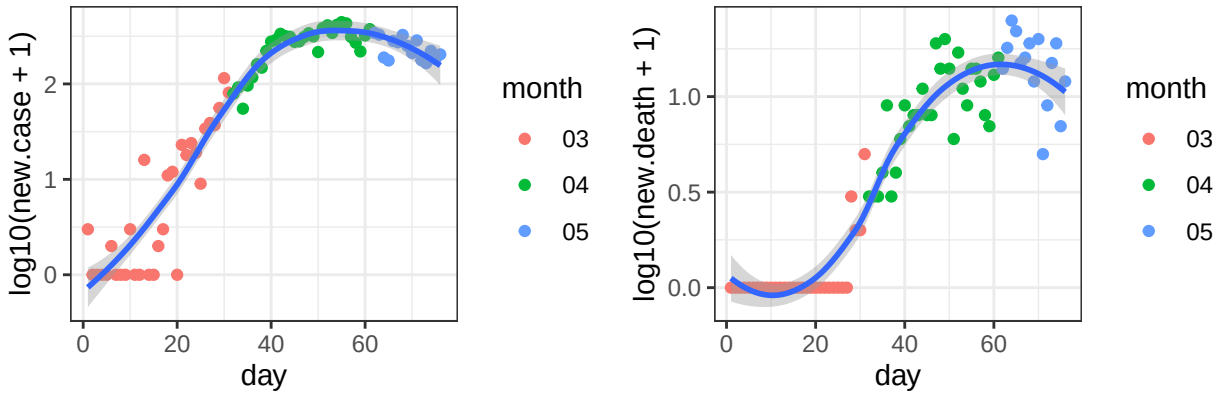
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Alabama



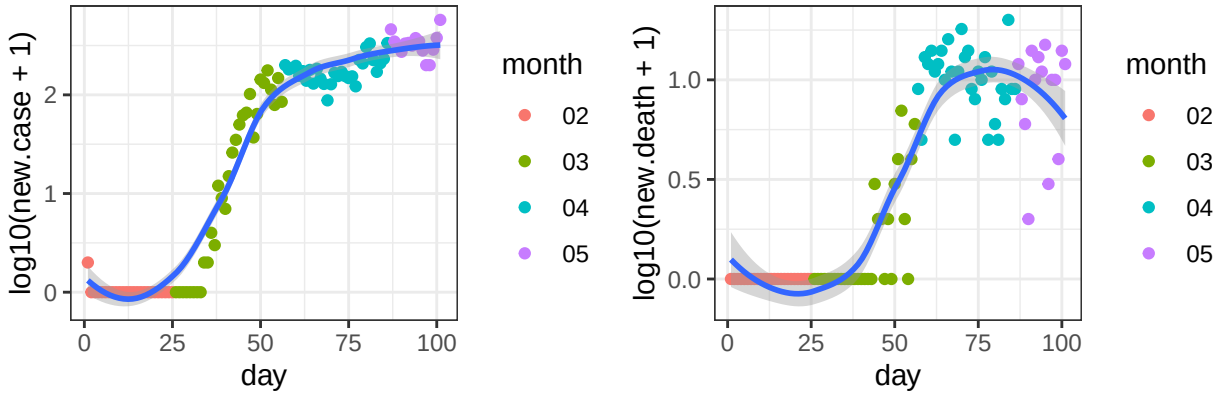
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Rhode Island



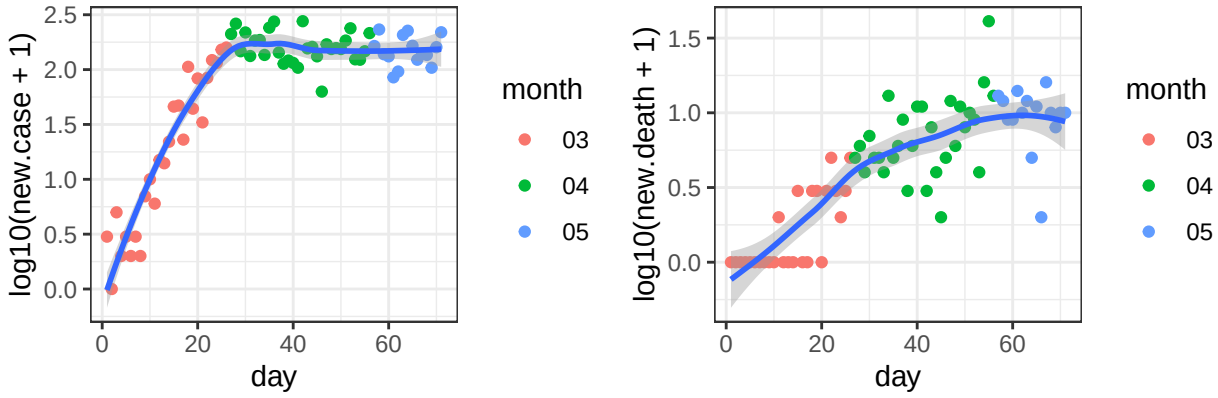
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Wisconsin



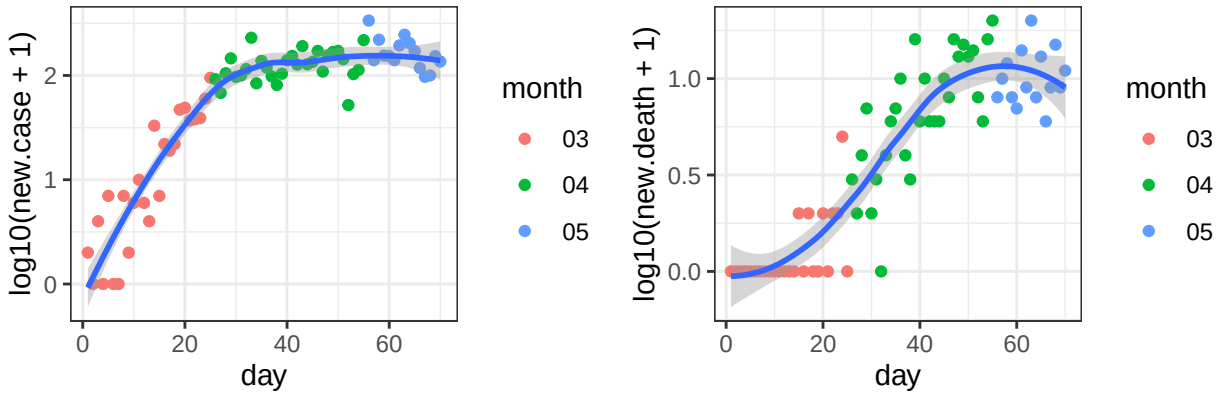
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−05

South Carolina



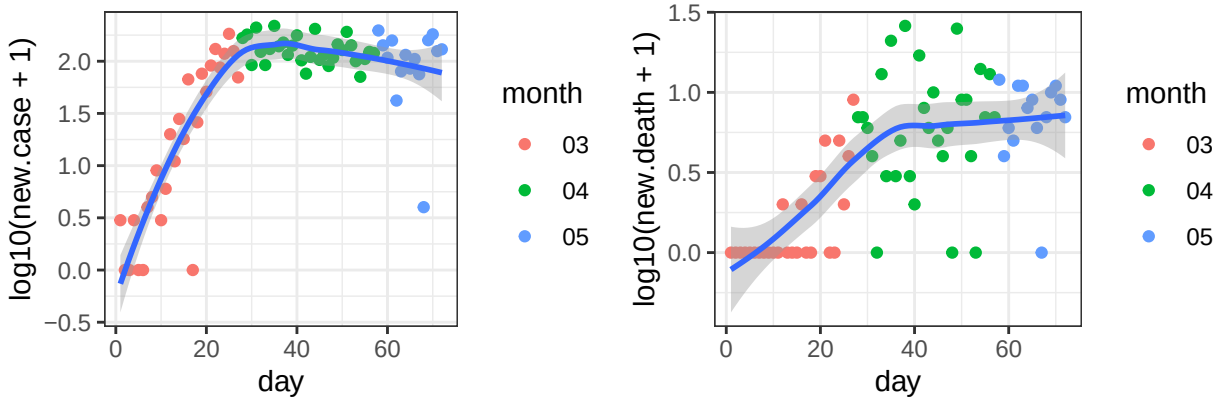
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

District of Columbia



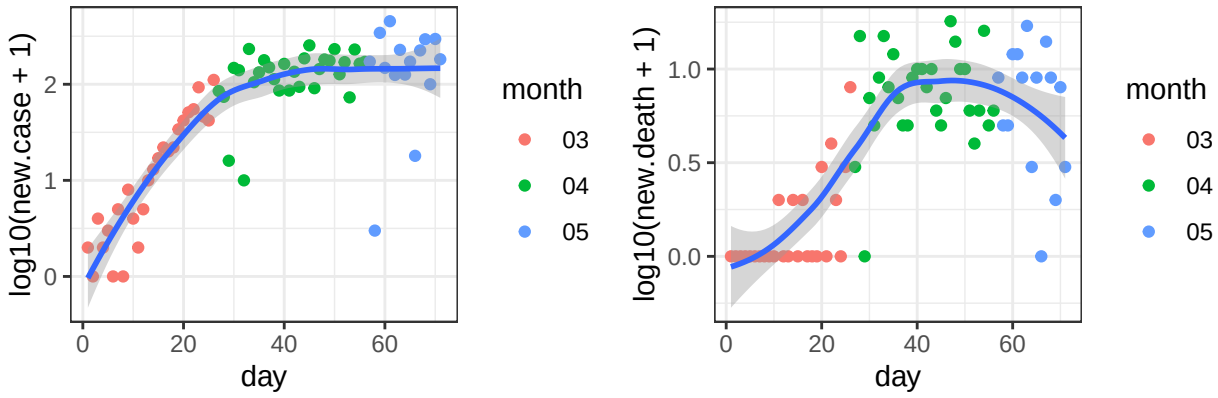
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Nevada



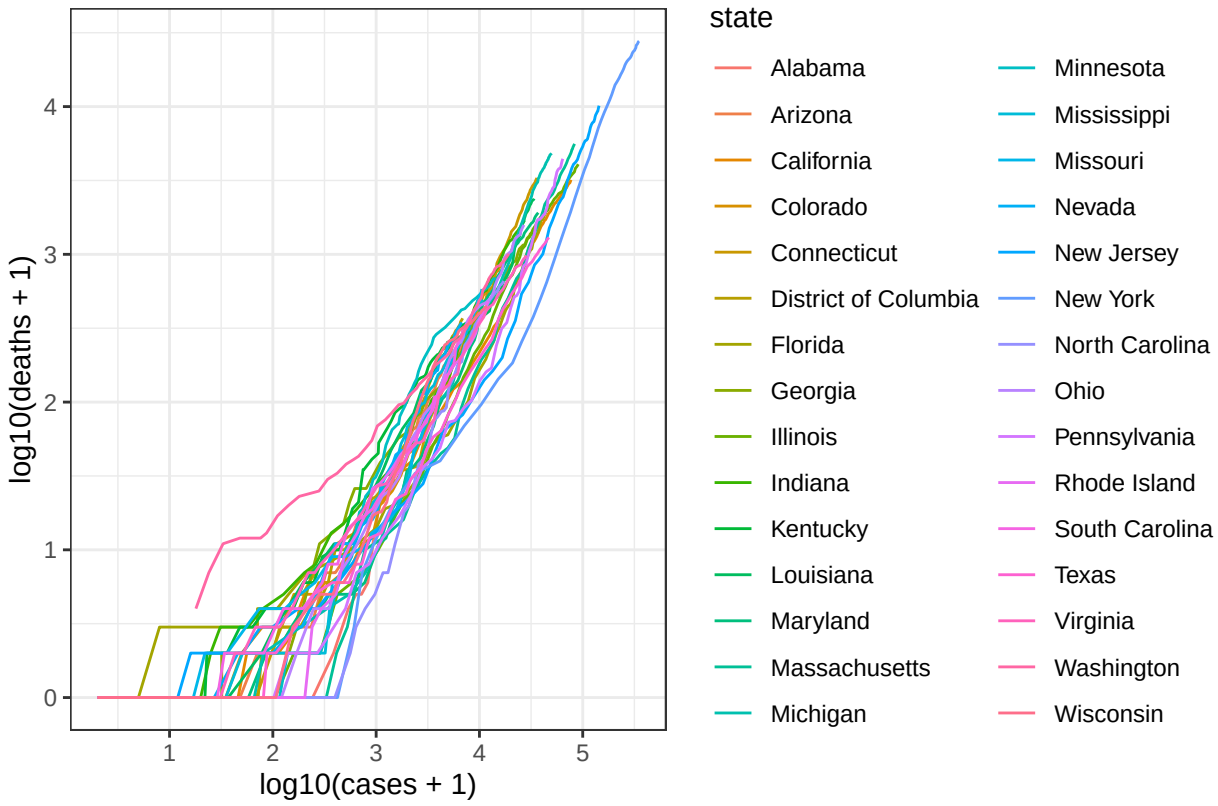
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Kentucky



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

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

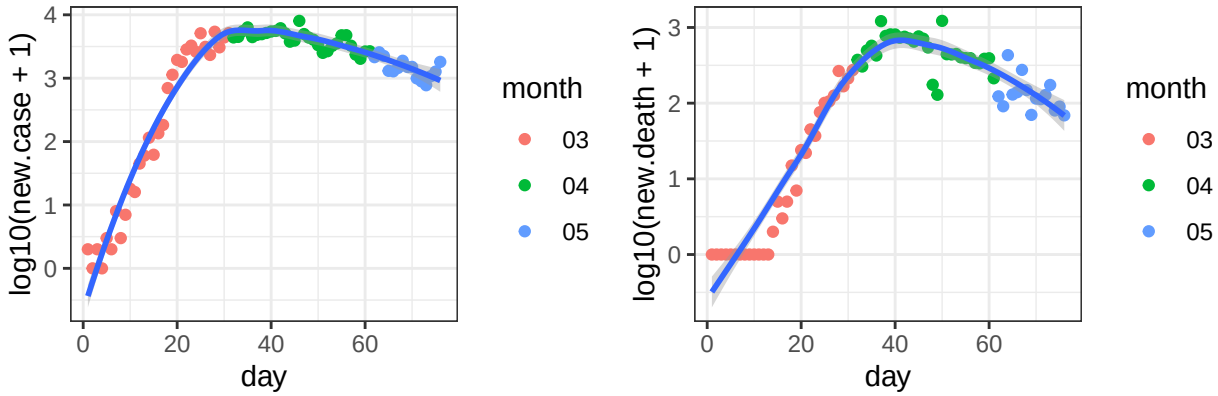
First check the 30 counties with the largest number of deaths.

```
##              date        county         state  fips  cases deaths
## 146045 2020-05-15 New York City      New York    NA 195472  19972
## 144901 2020-05-15          Cook      Illinois 17031  59905   2762
## 146044 2020-05-15        Nassau      New York 36059  38864   2499
## 145571 2020-05-15         Wayne      Michigan 26163  18882   2192
## 146064 2020-05-15       Suffolk      New York 36103  37719   1757
## 144507 2020-05-15   Los Angeles    California  6037  36259   1755
## 145971 2020-05-15         Essex    New Jersey 34013  15953   1510
## 145966 2020-05-15        Bergen    New Jersey 34003  17195   1443
## 146072 2020-05-15   Westchester      New York 36119  31942   1392
## 145486 2020-05-15     Middlesex Massachusetts 25017  18683   1347
## 144606 2020-05-15     Fairfield   Connecticut  9001  14009   1109
## 145973 2020-05-15        Hudson    New Jersey 34017  17237   1042
## 144607 2020-05-15      Hartford   Connecticut  9003   8126   1025
## 146457 2020-05-15  Philadelphia  Pennsylvania 42101  19349   1021
## 145984 2020-05-15         Union    New Jersey 34039  14492    939
## 145552 2020-05-15       Oakland      Michigan 26125   7994    896
## 145976 2020-05-15     Middlesex    New Jersey 34023  14429    865
## 145980 2020-05-15       Passaic    New Jersey 34031  14930    816
## 144610 2020-05-15     New Haven   Connecticut  9009   9881    783
## 145490 2020-05-15       Suffolk Massachusetts 25025  15996    768
## 145482 2020-05-15         Essex Massachusetts 25009  12131    751
## 145539 2020-05-15        Macomb      Michigan 26099   6274    729
```

```
## 145488 2020-05-15       Norfolk Massachusetts 25021   7331    710
## 145979 2020-05-15         Ocean     New Jersey 34029   7829    610
## 146452 2020-05-15    Montgomery   Pennsylvania 42091   5697    608
## 145978 2020-05-15        Morris     New Jersey 34027   5990    550
## 144662 2020-05-15    Miami-Dade        Florida 12086  15010    548
## 145492 2020-05-15     Worcester Massachusetts 25027   8786    538
## 147079 2020-05-15          King     Washington 53033   7679    523
## 145034 2020-05-15        Marion        Indiana 18097   8082    500
```

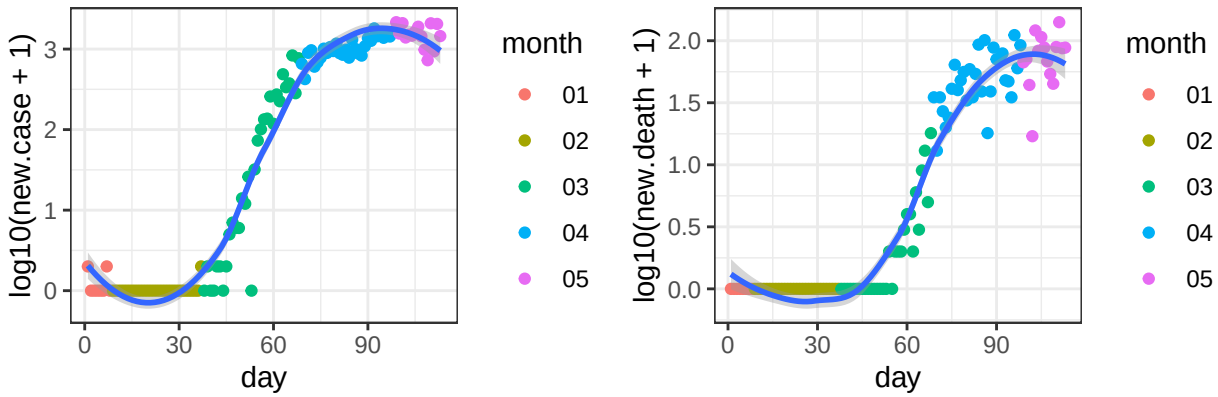
For these 30 counties, I check the number of new cases and the number of new deaths.

New York City_New York



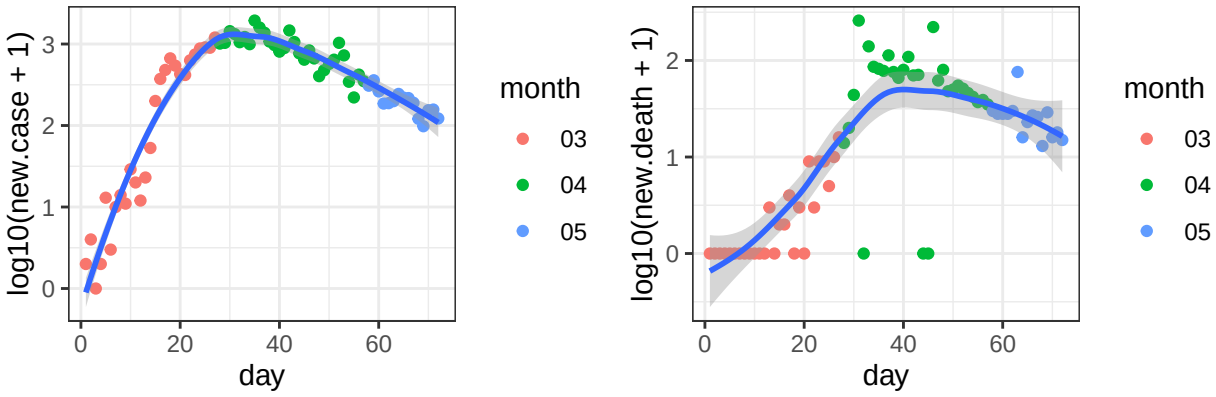
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Cook_Illinois



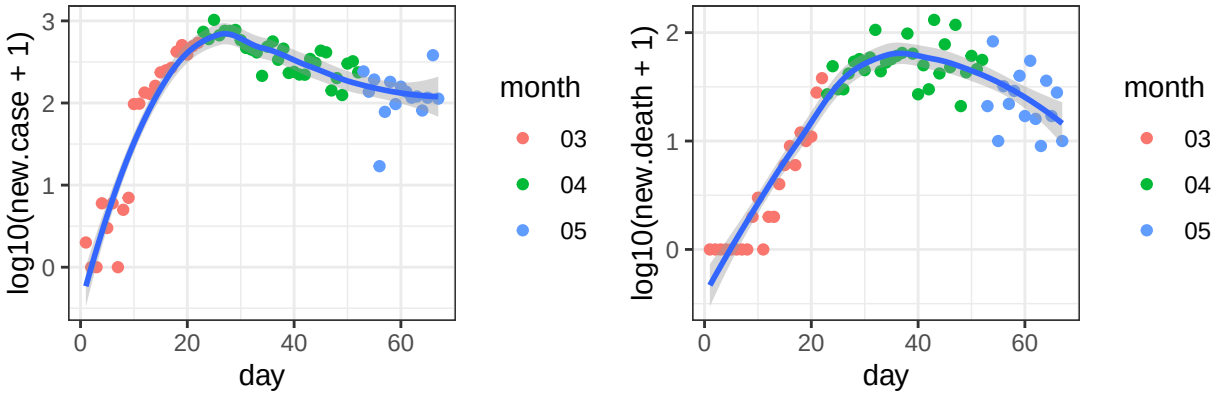
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

Nassau_New York



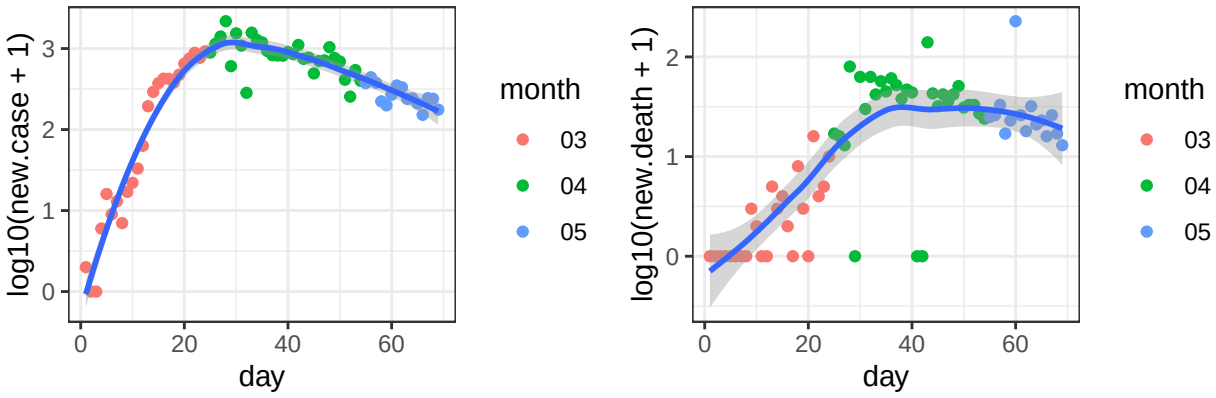
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Wayne_Michigan



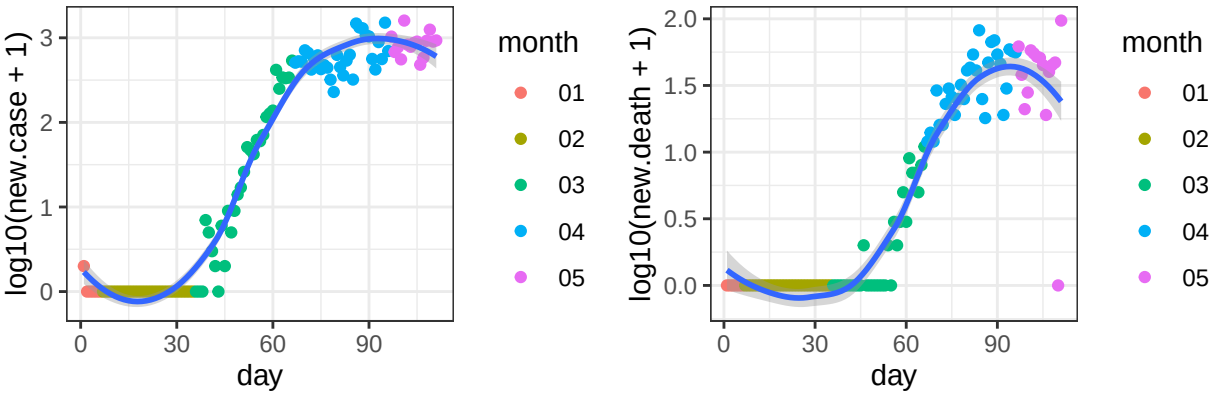
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Suffolk_New York



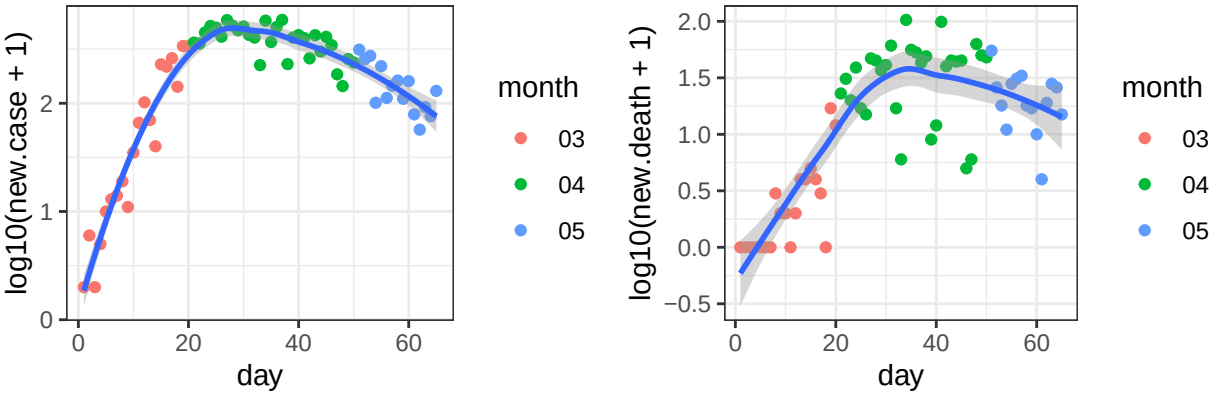
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Los Angeles_California



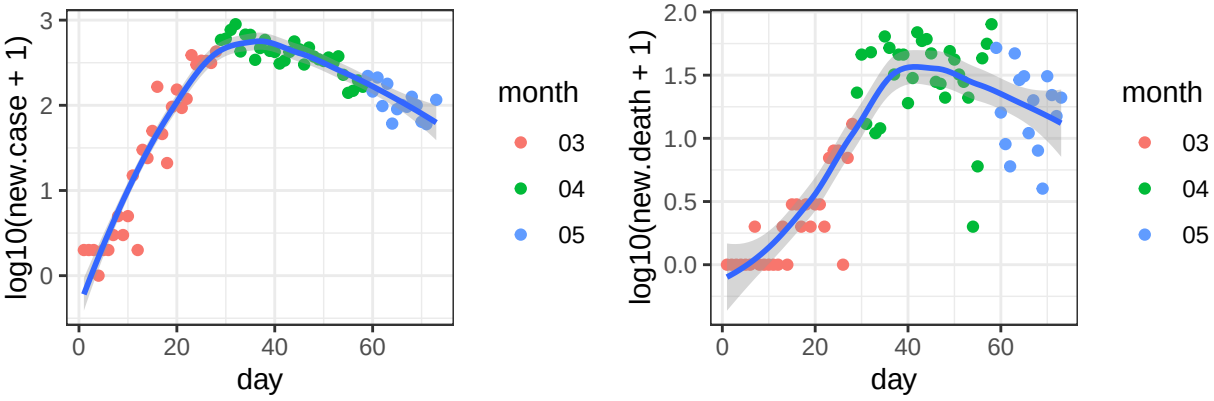
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Essex_New Jersey



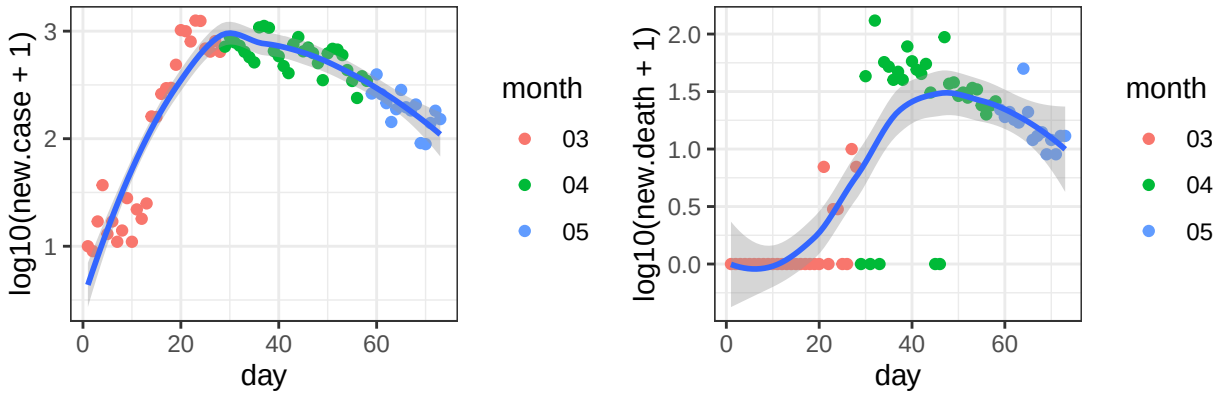
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Bergen_New Jersey



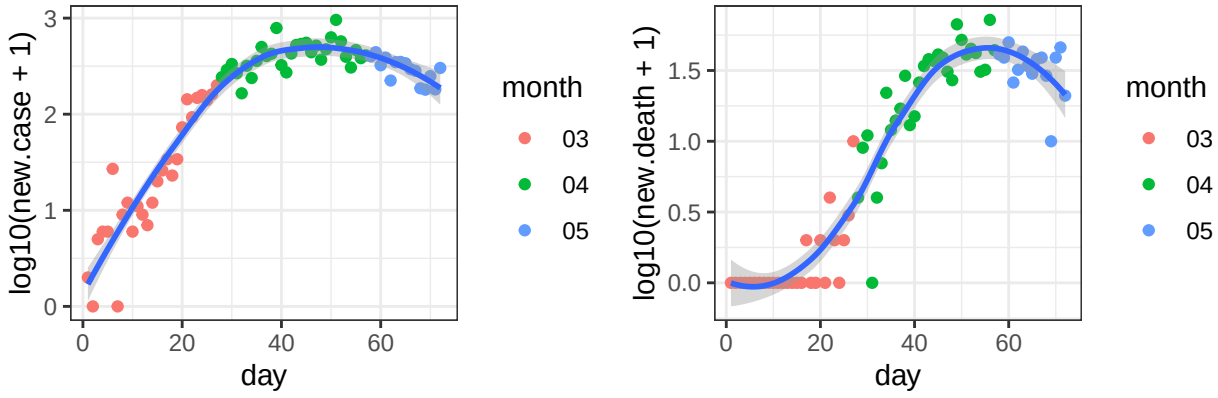
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Westchester_New York



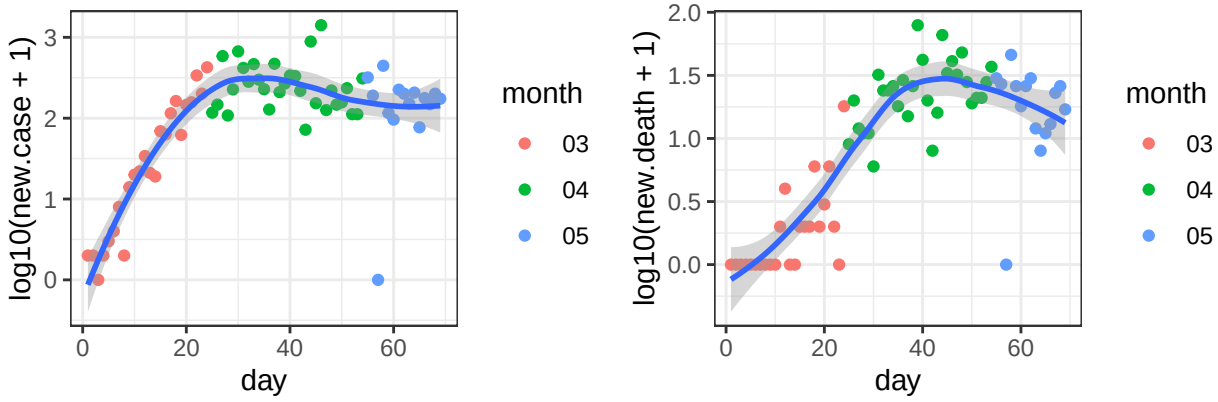
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Middlesex_Massachusetts



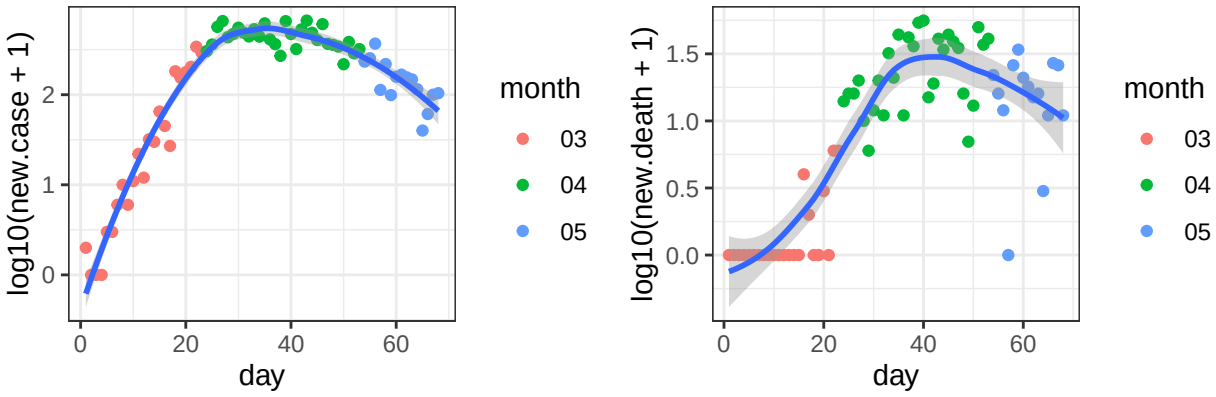
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Fairfield_Connecticut



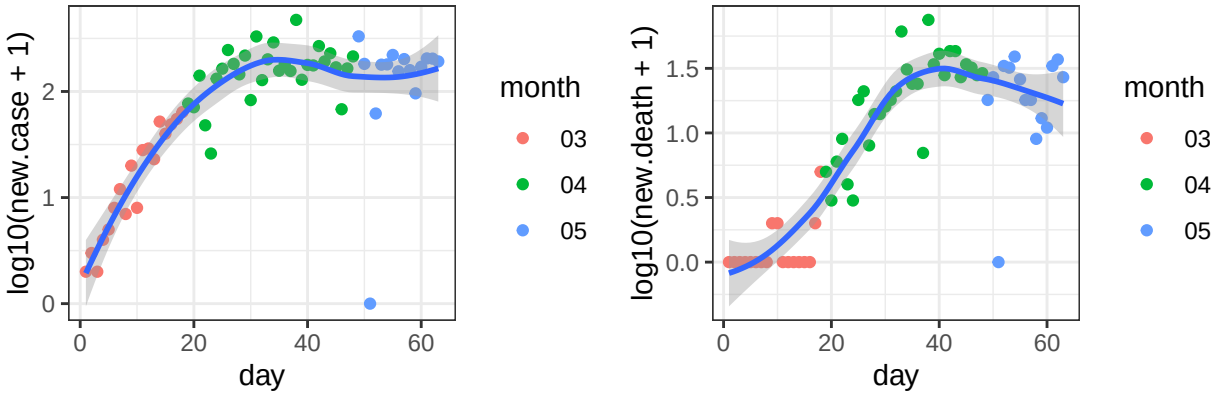
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Hudson_New Jersey



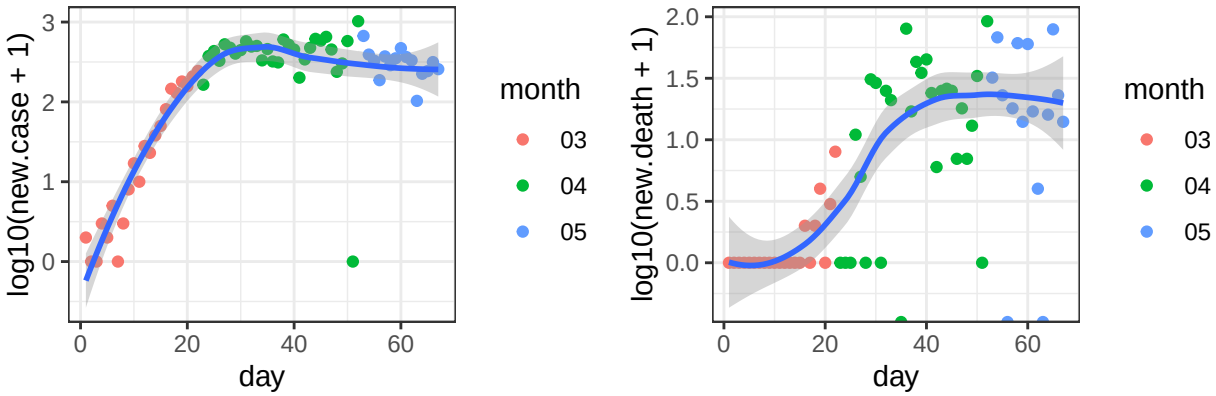
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Hartford_Connecticut



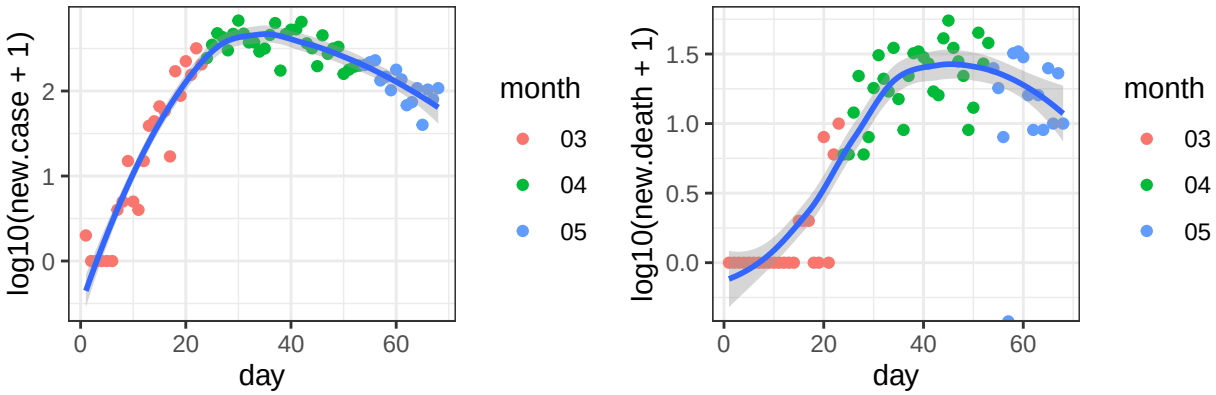
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Philadelphia_Pennsylvania



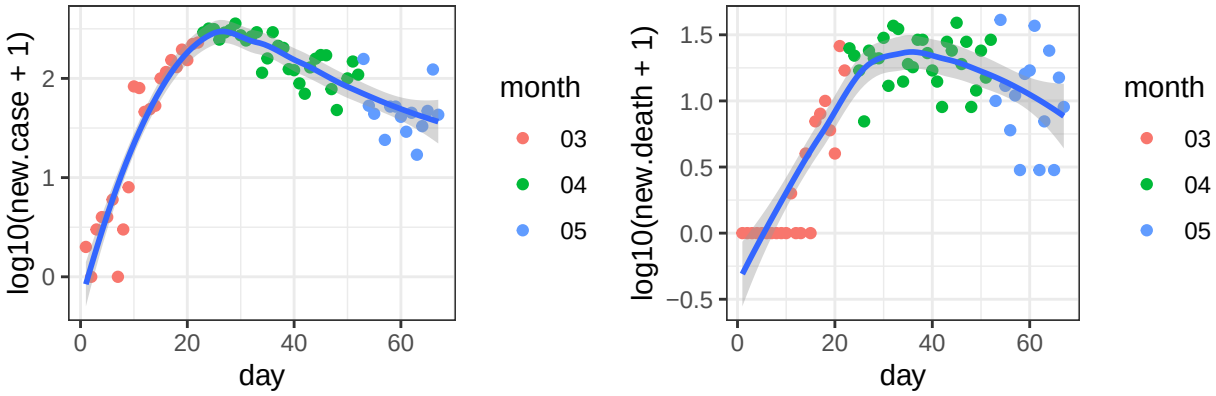
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Union_New Jersey



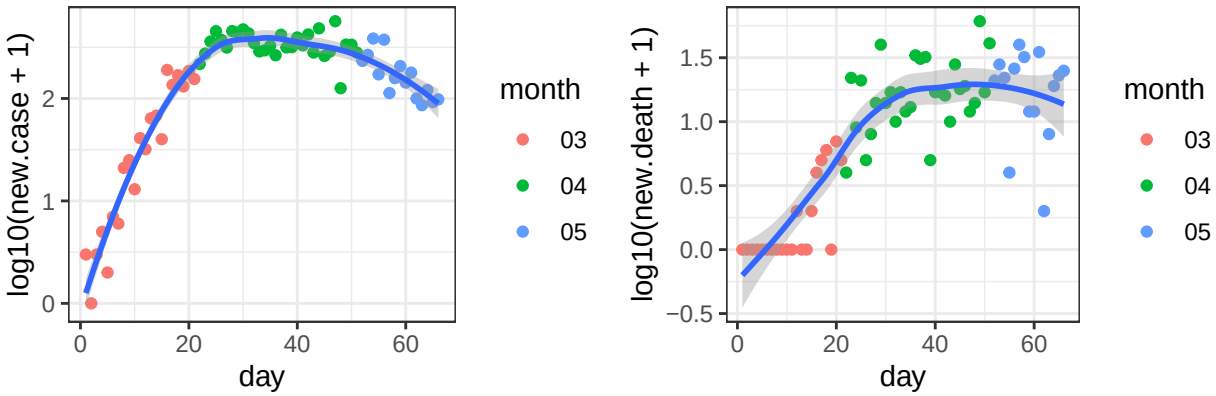
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Oakland_Michigan



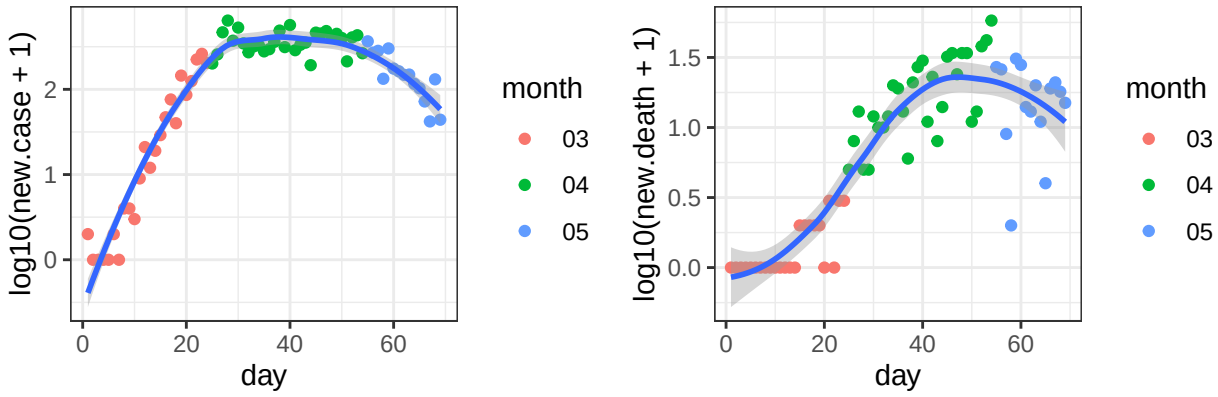
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Middlesex_New Jersey



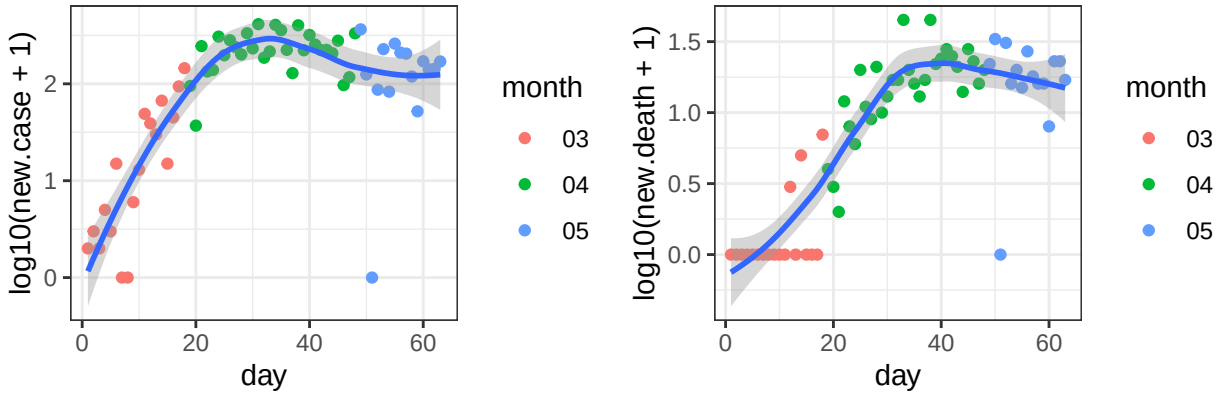
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Passaic_New Jersey



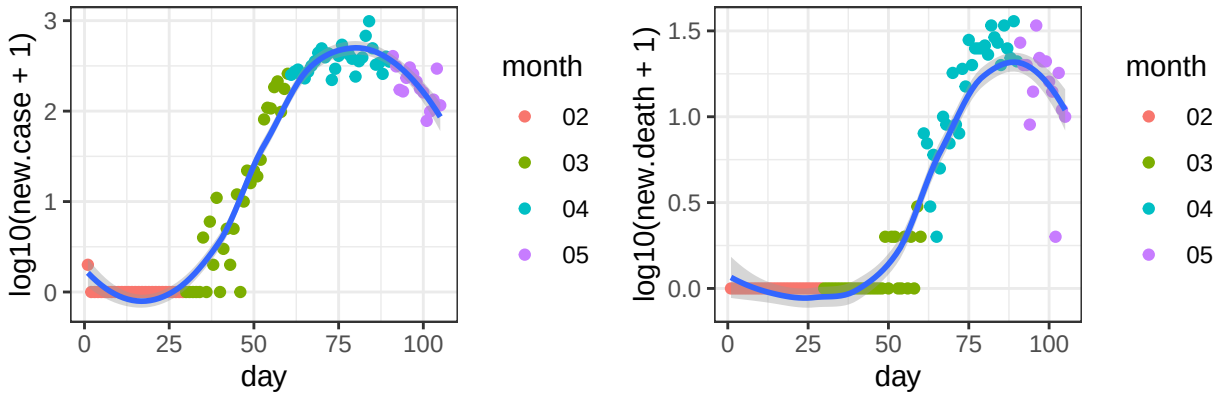
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

New Haven_Connecticut



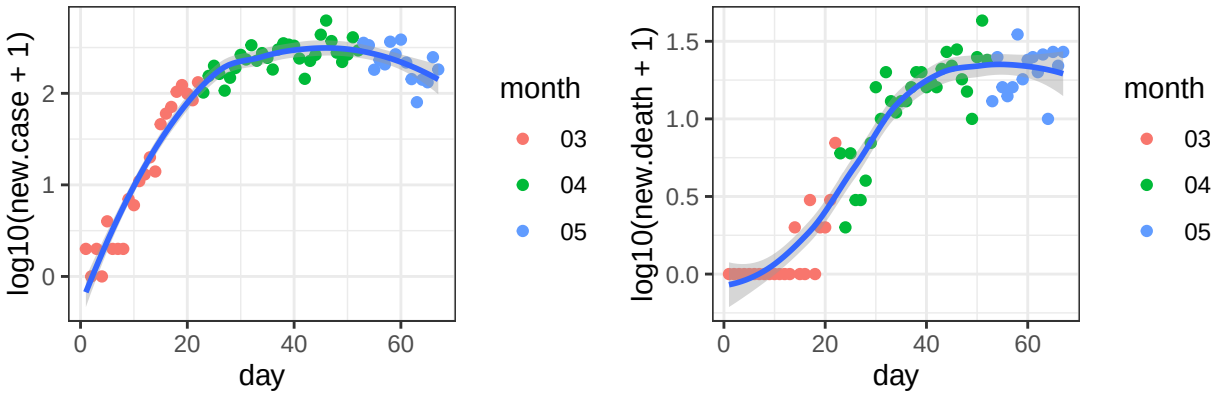
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Suffolk_Massachusetts



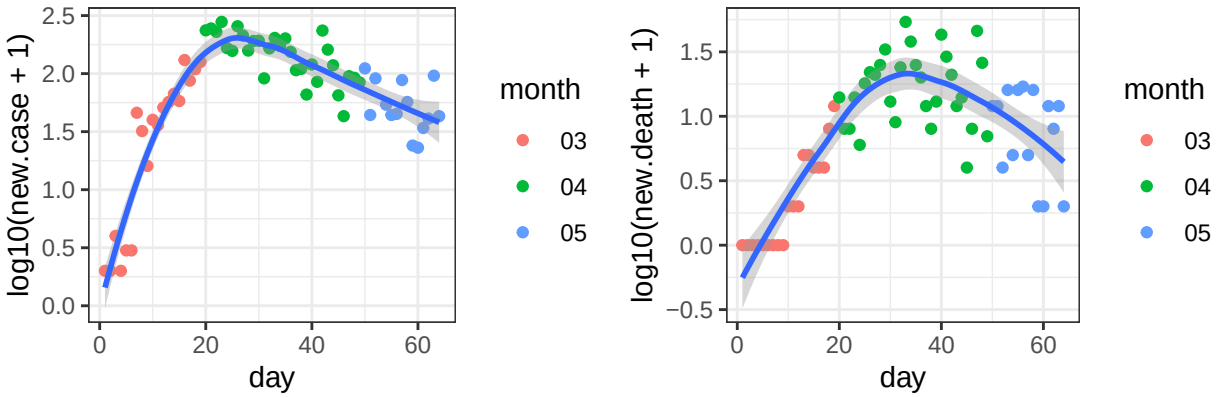
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01

Essex_Massachusetts



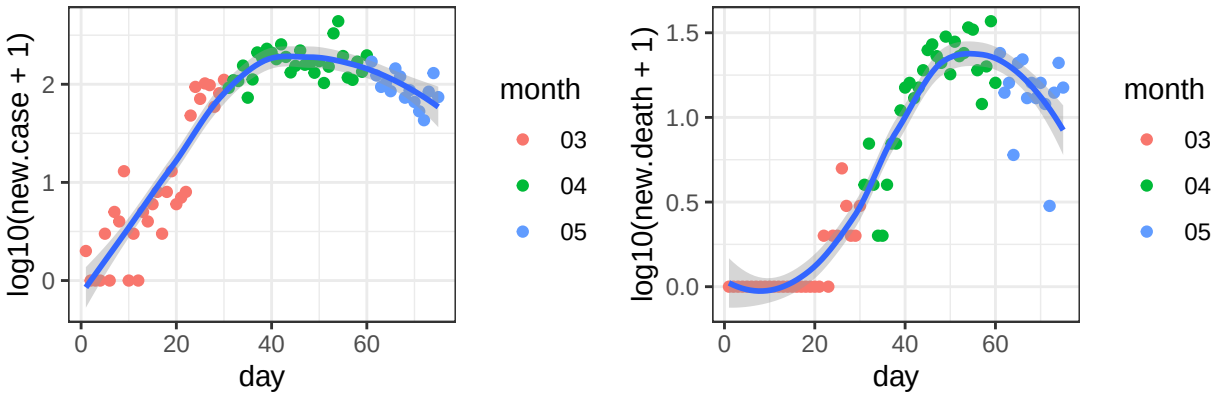
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Macomb_Michigan



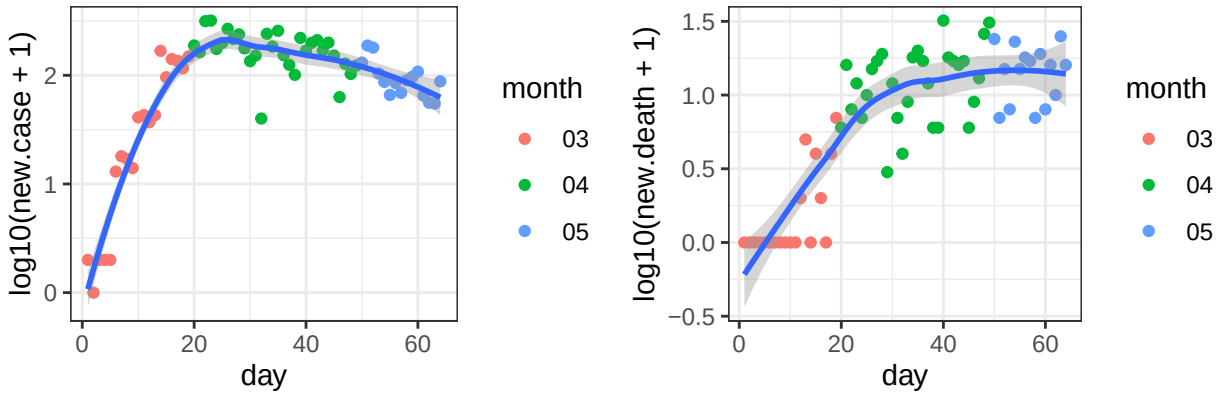
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Norfolk_Massachusetts



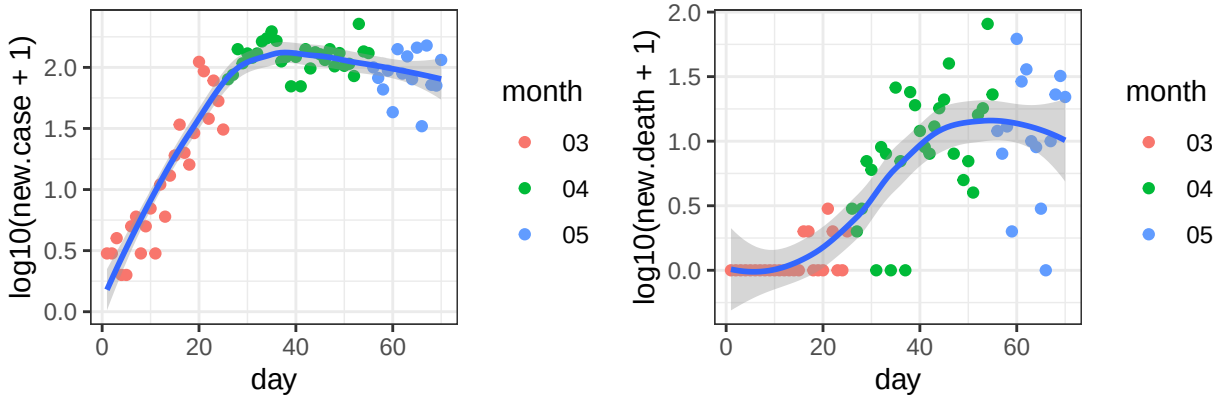
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Ocean_New Jersey



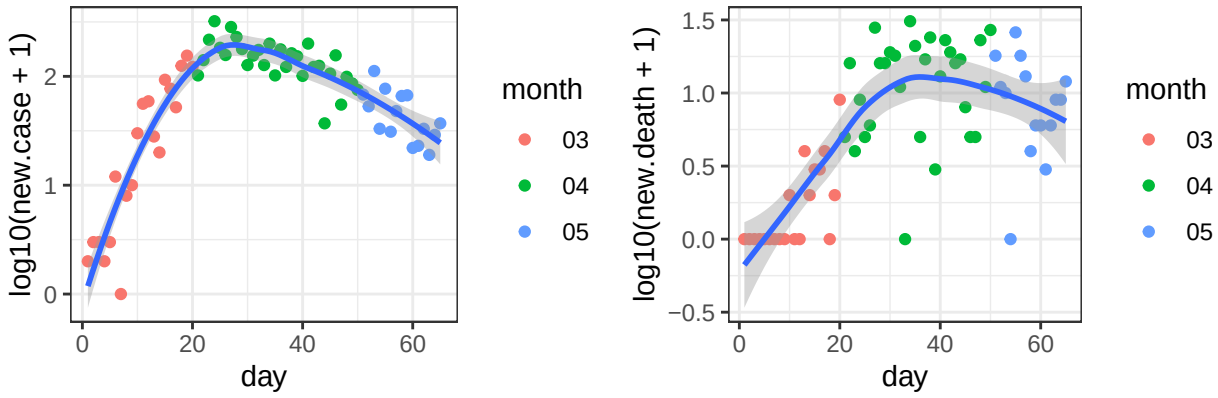
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Montgomery_Pennsylvania



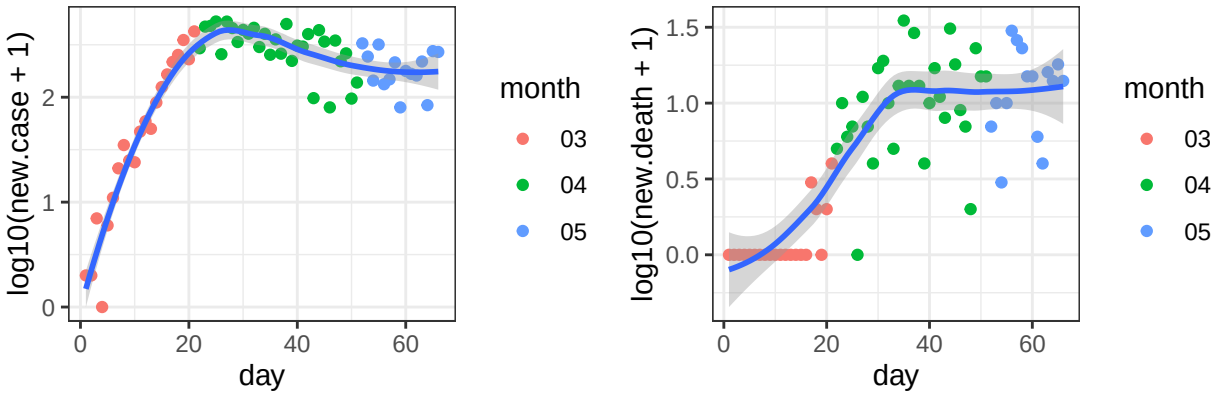
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Morris_New Jersey



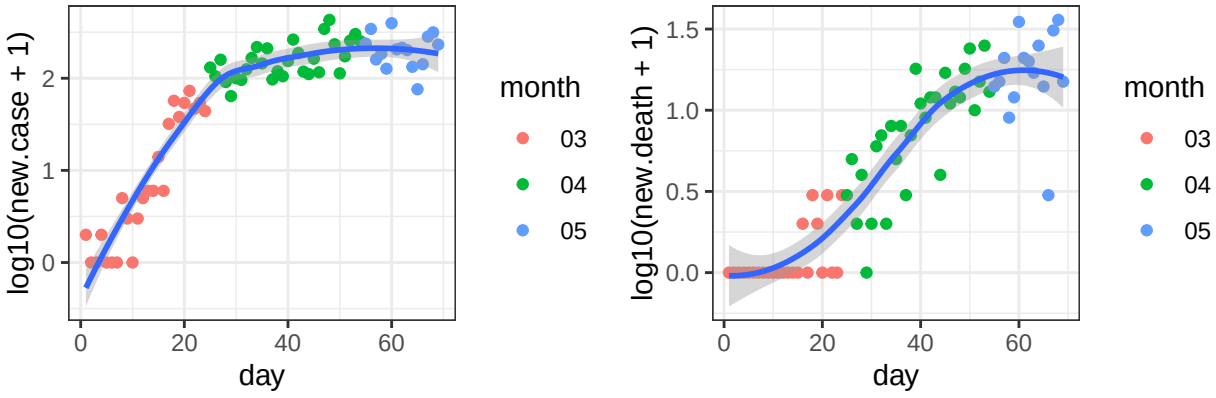
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Miami−Dade_Florida



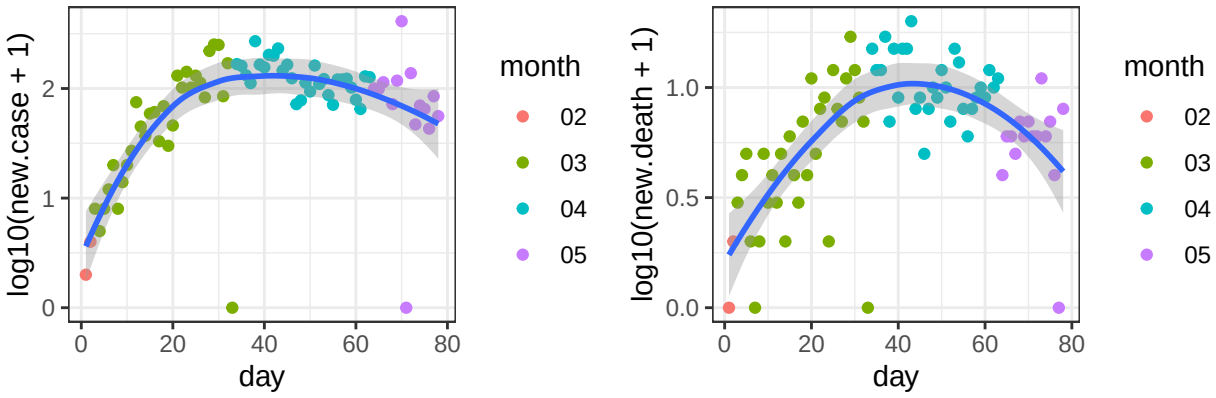
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Worcester_Massachusetts



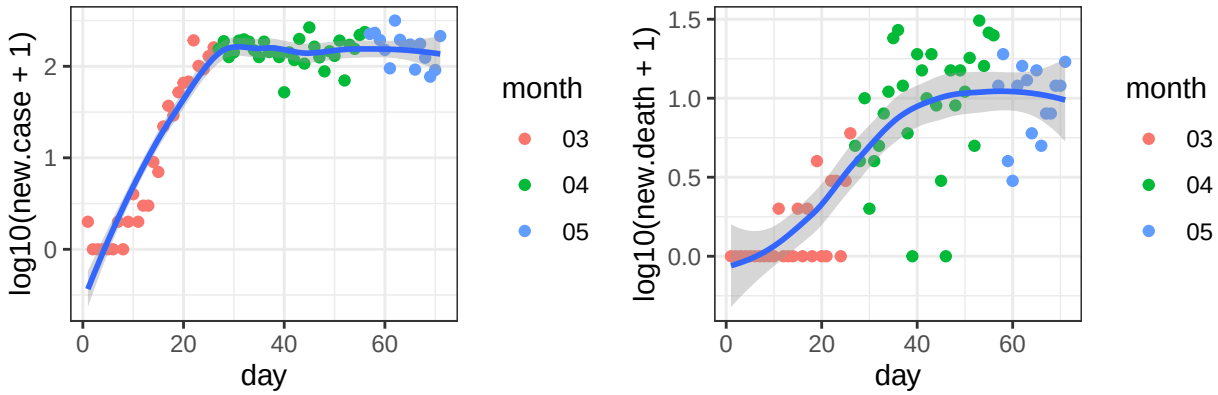
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

King_Washington



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

Marion_Indiana

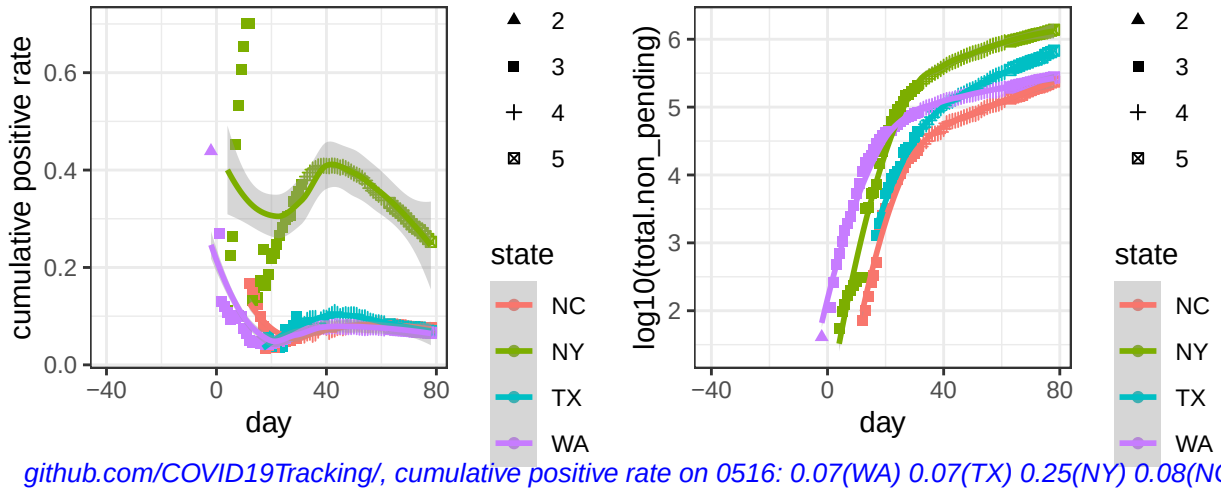


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

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 0516: 0.07(WA) 0.07(TX) 0.25(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
```