Exploration of COVID-19 tracking data from multiple resources

Wei Sun

2020-07-14

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

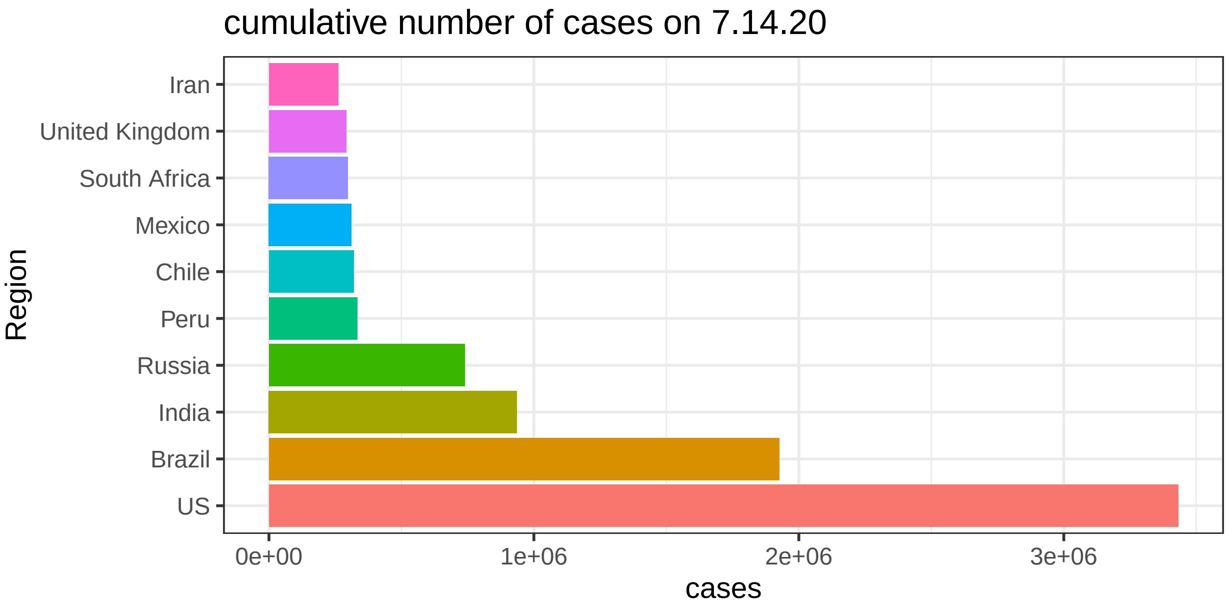
JHU

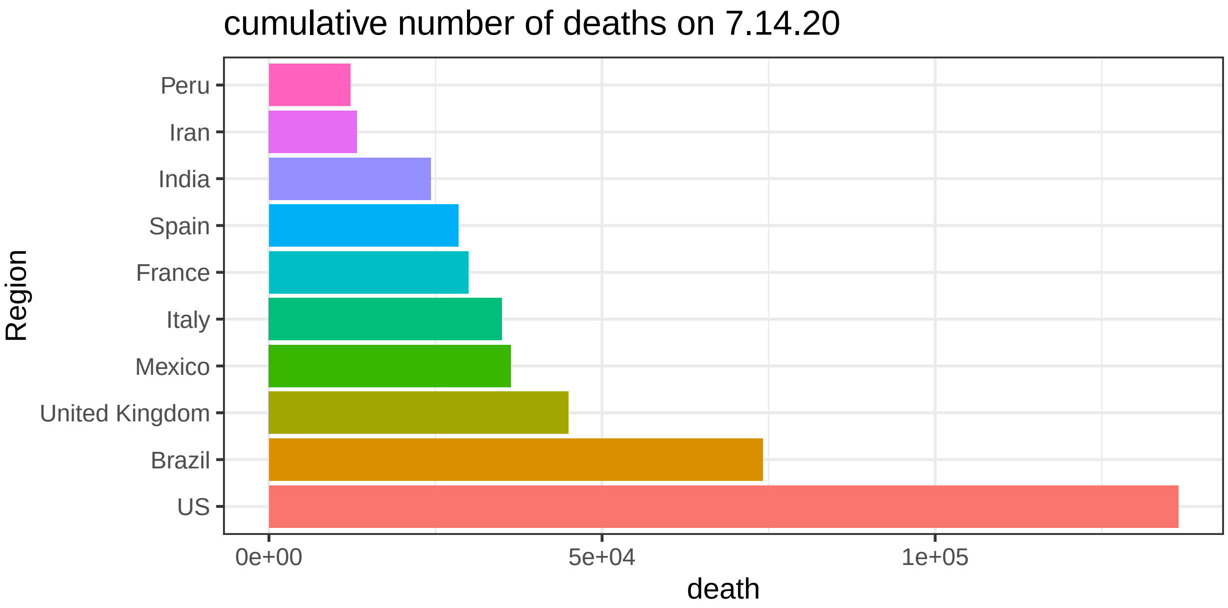
Assume you have cloned the JHU Github repository on your local machine at “../COVID-19”.

time series data

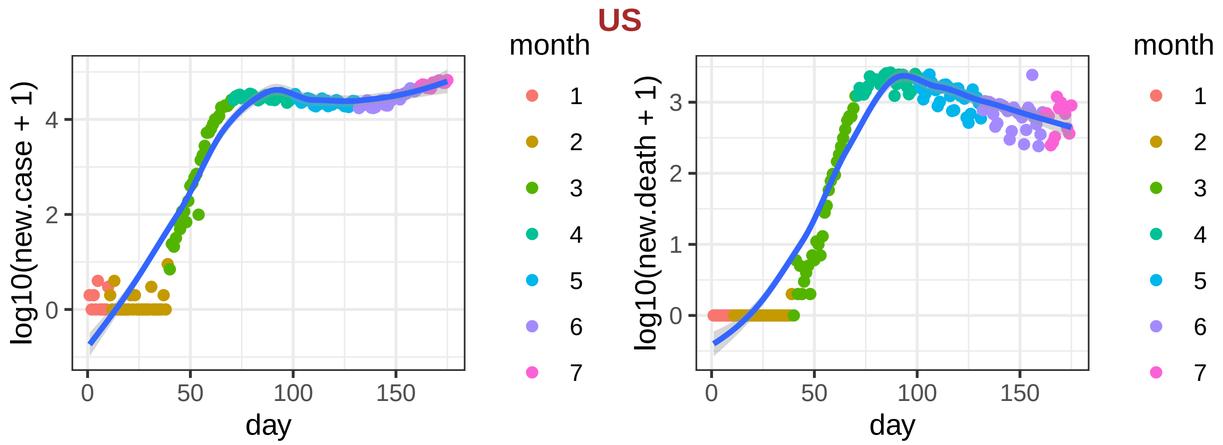
The time series provide counts (e.g., confirmed cases, deaths) starting from Jan 22nd, 2020 for 253 locations. Currently there is no data of individual US state in these time series data files.

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

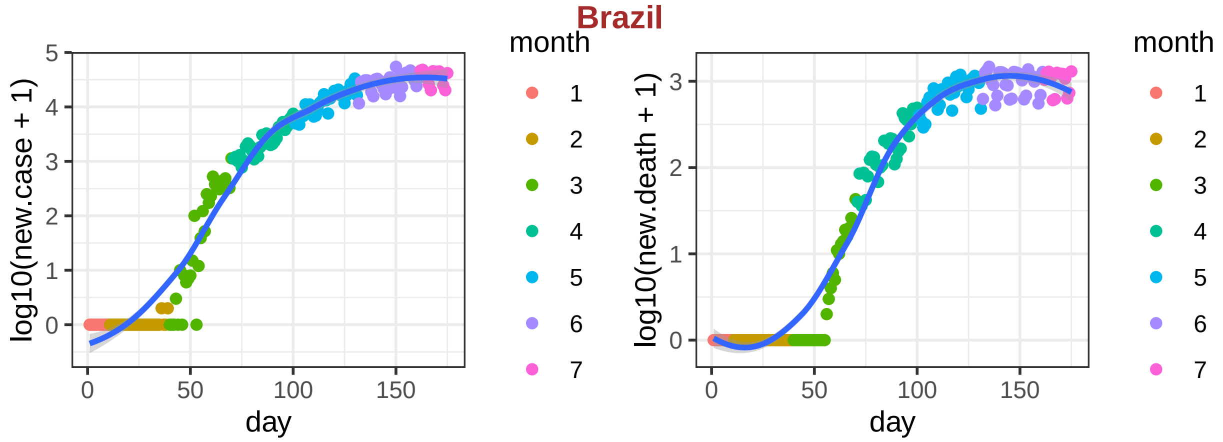




Next, I check for each country/region, what is the number of new cases/deaths? This data is important to understand what is the trend under different situations, e.g., population density, social distance policies etc. Here I checked the top 10 countries/regions with the highest number of deaths.

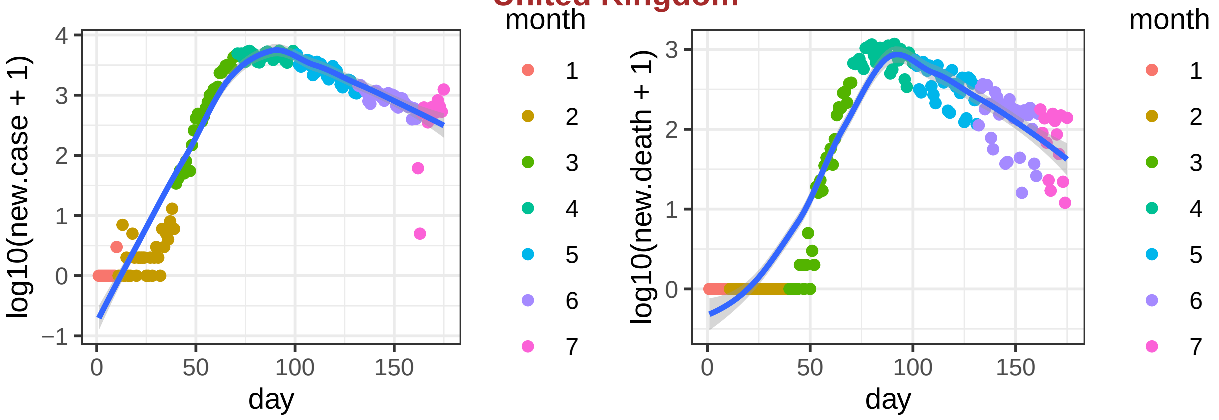


data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020



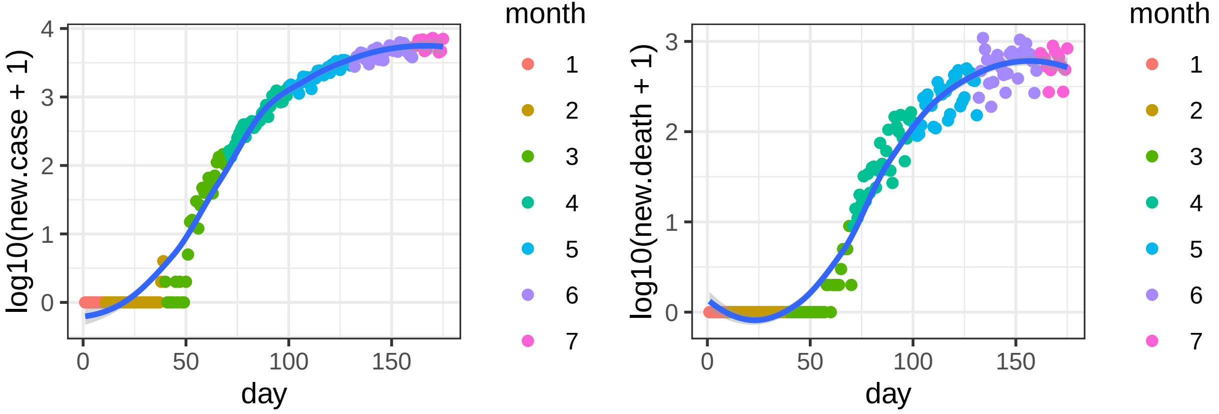
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

United Kingdom



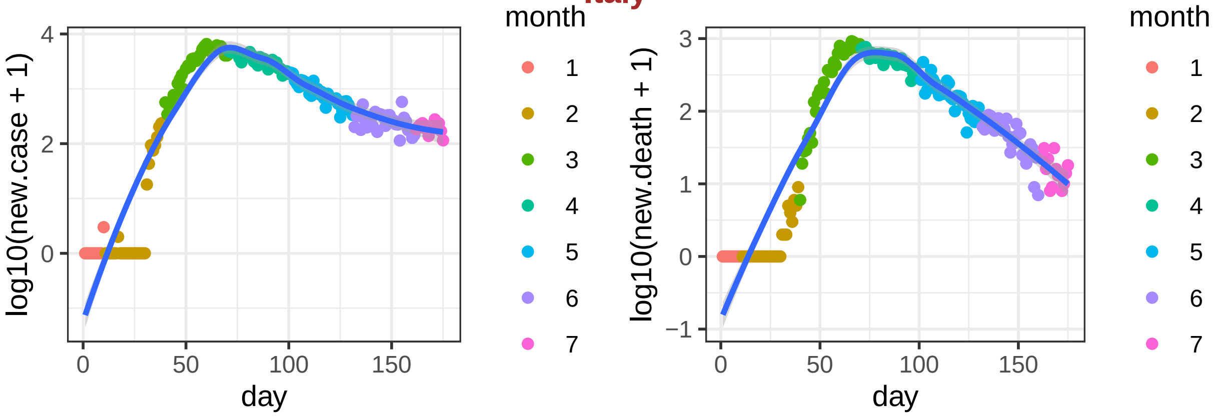
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Mexico



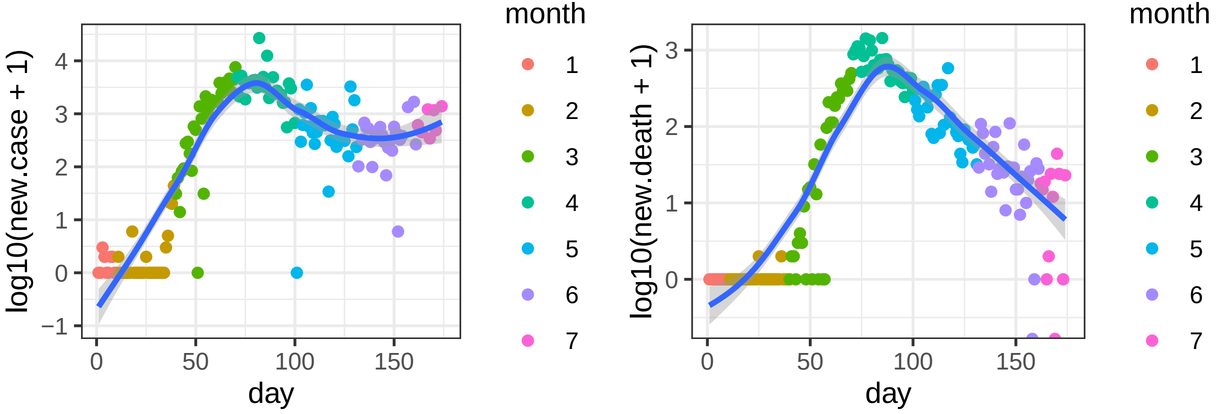
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Italy



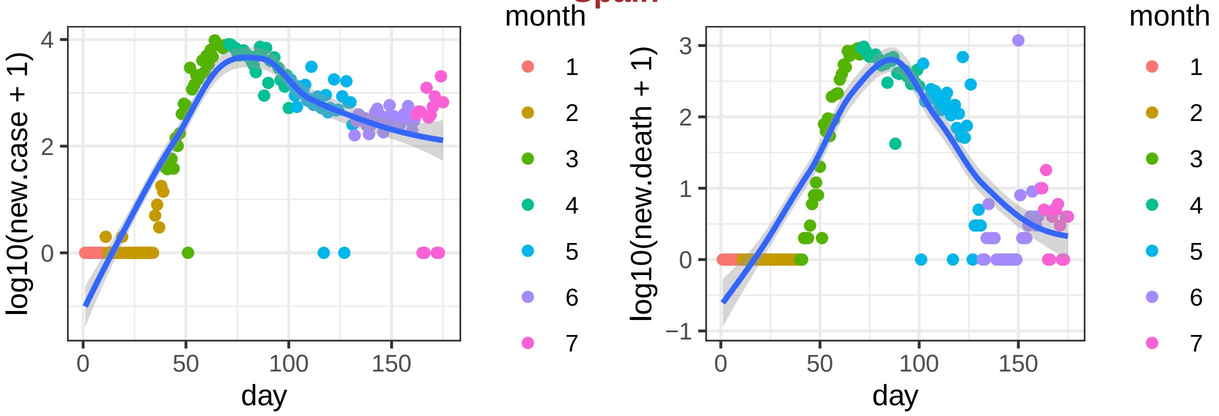
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

France



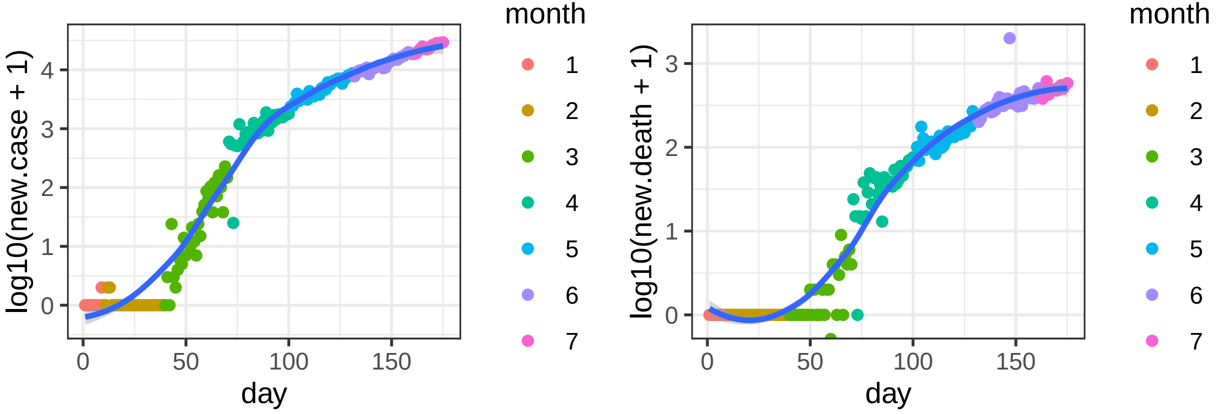
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

Spain

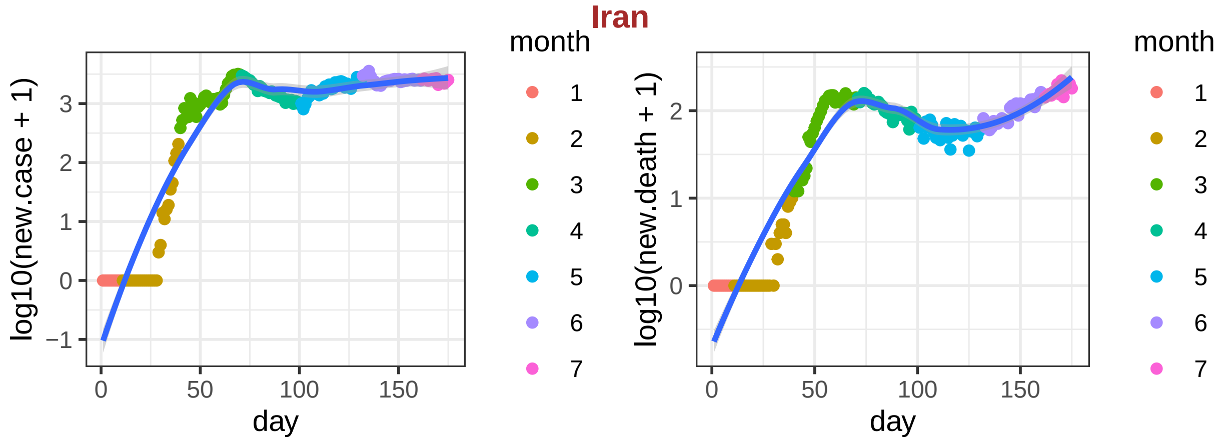


data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020

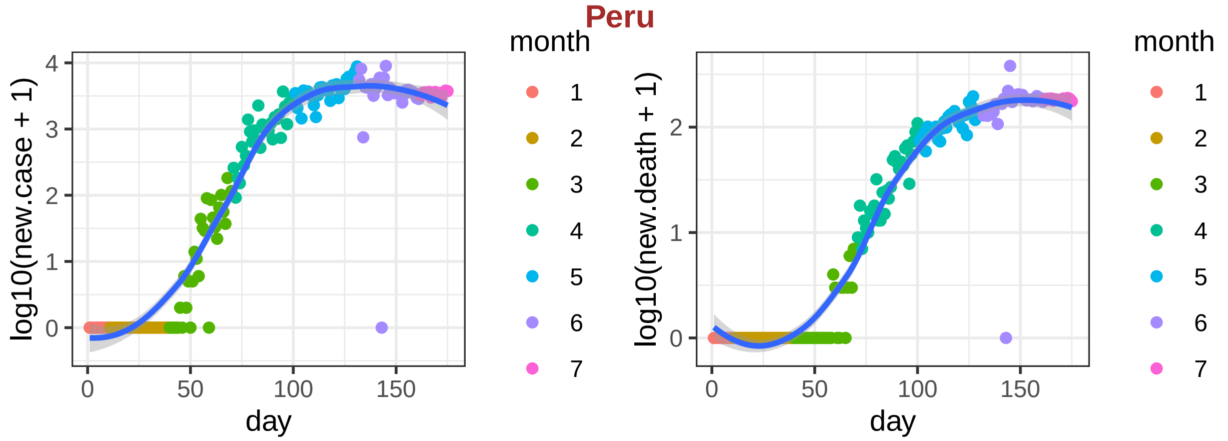
India



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



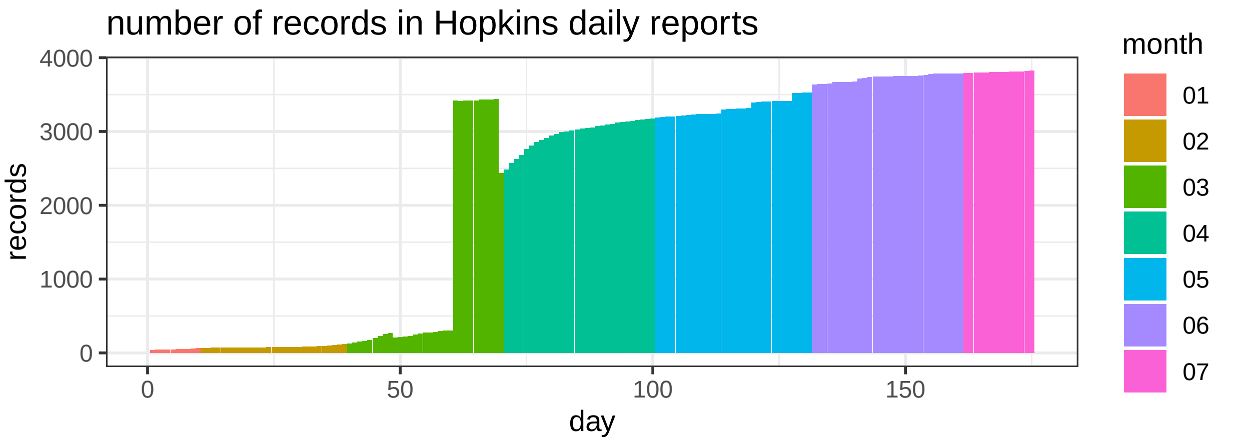
data source: https://github.com/CSSEGISandData/COVID−19, day 1 is 1/22/2020



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] "2020-07-13"
```

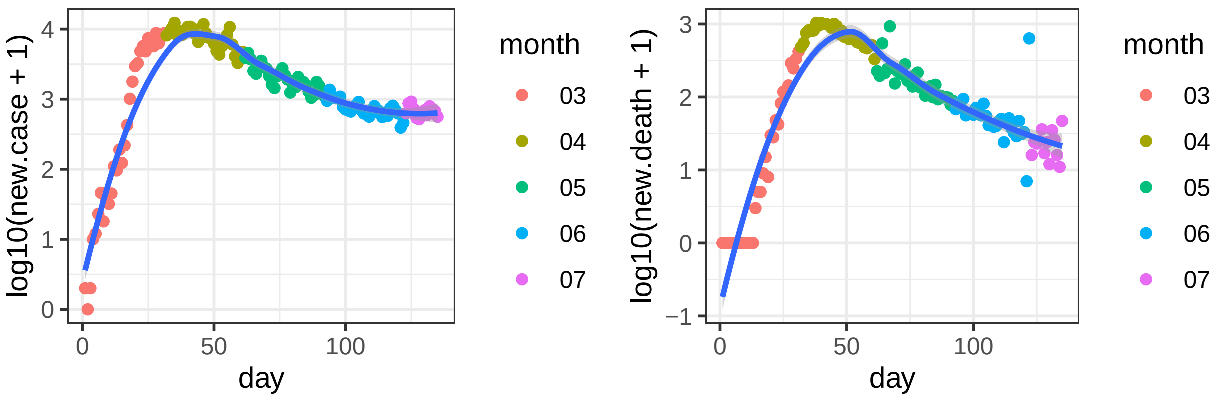
state level data

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

```
##            date          state fips  cases deaths
## 7308 2020-07-13       New York   36 406962  32075
## 7306 2020-07-13     New Jersey   34 177469  15560
## 7297 2020-07-13  Massachusetts   25 111827   8330
## 7289 2020-07-13       Illinois   17 156288   7398
## 7279 2020-07-13     California    6 336206   7086
## 7315 2020-07-13   Pennsylvania   42 100378   6955
## 7298 2020-07-13       Michigan   26  77354   6324
## 7281 2020-07-13    Connecticut    9  47510   4371
## 7284 2020-07-13        Florida   12 282427   4276
## 7294 2020-07-13      Louisiana   22  79935   3423
## 7296 2020-07-13       Maryland   24  74124   3325
## 7321 2020-07-13          Texas   48 273712   3313
## 7312 2020-07-13           Ohio   39  66853   3064
## 7285 2020-07-13        Georgia   13 111937   2972
## 7290 2020-07-13        Indiana   18  53327   2762
## 7277 2020-07-13        Arizona    4 123917   2250
## 7325 2020-07-13       Virginia   51  71642   1968
## 7280 2020-07-13       Colorado    8  37303   1728
## 7309 2020-07-13 North Carolina   37  87750   1547
## 7299 2020-07-13      Minnesota   27  42810   1542
## 7326 2020-07-13     Washington   53  43538   1439
## 7300 2020-07-13    Mississippi   28  36680   1250
## 7301 2020-07-13       Missouri   29  29781   1126
## 7275 2020-07-13        Alabama    1  55545   1124
## 7317 2020-07-13   Rhode Island   44  17487    984
## 7318 2020-07-13 South Carolina   45  58168    972
## 7328 2020-07-13      Wisconsin   55  40603    828
## 7291 2020-07-13           Iowa   19  35631    755
## 7320 2020-07-13      Tennessee   47  63615    738
## 7293 2020-07-13       Kentucky   21  20127    656
```

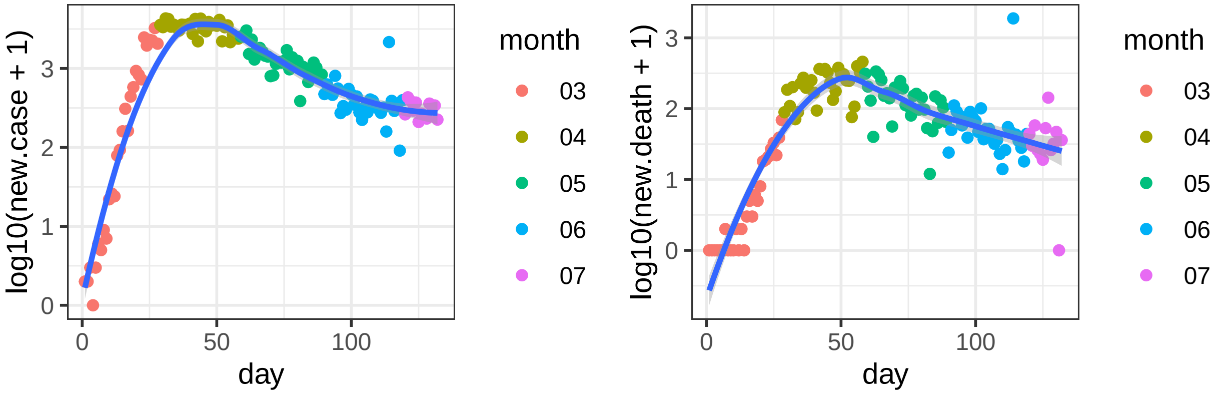
For these 20 states, I check the number of new cases and the number of new deaths. Part of the reason for such checking is to identify whether there is any similarity on such patterns. For example, could you use the pattern seen from Italy to predict what happen in an individual state, and what are the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

New York



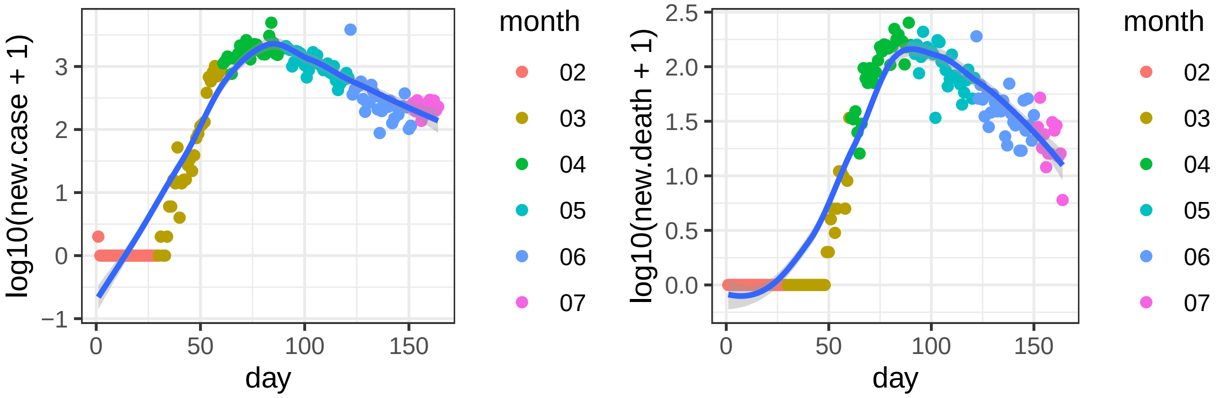
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

New Jersey



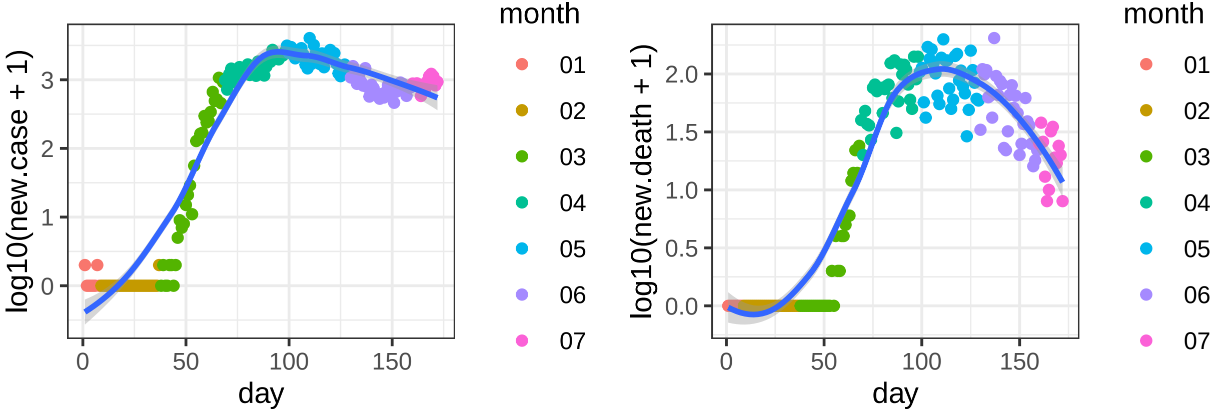
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Massachusetts



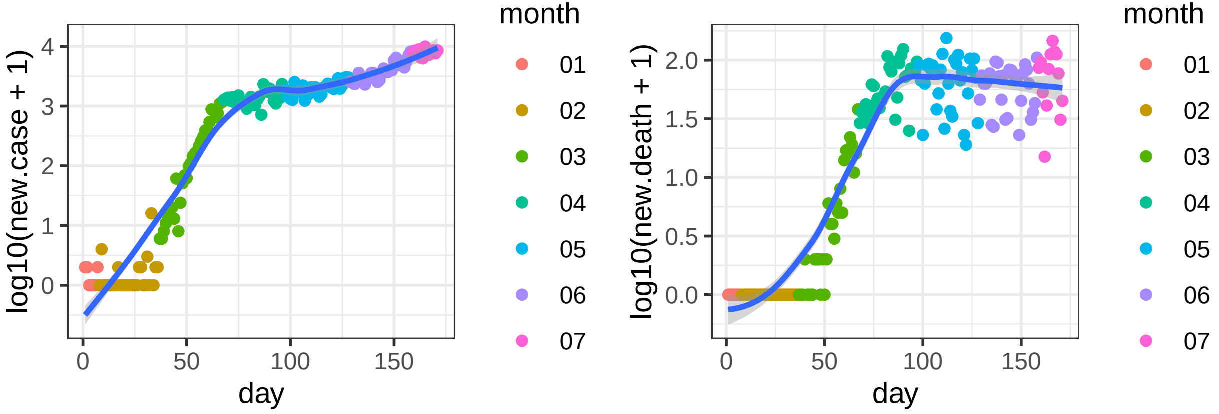
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01

Illinois



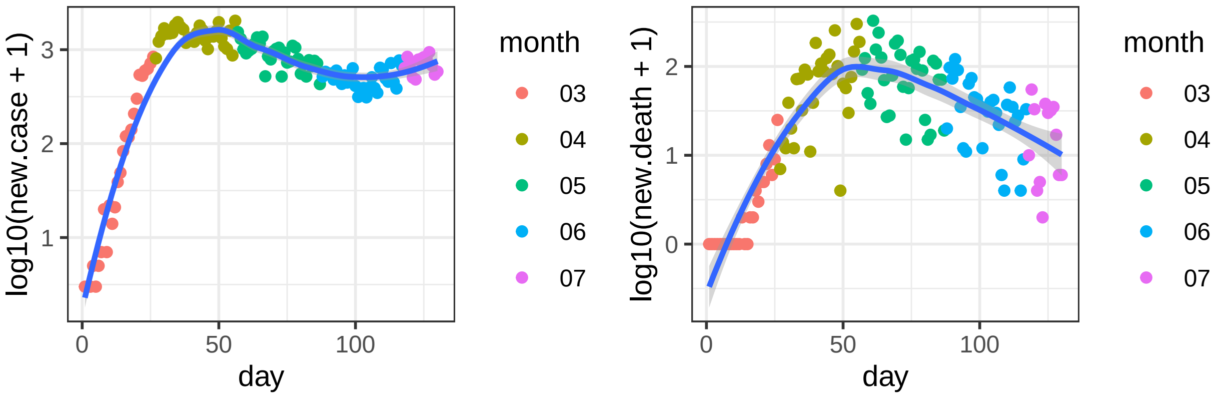
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

California



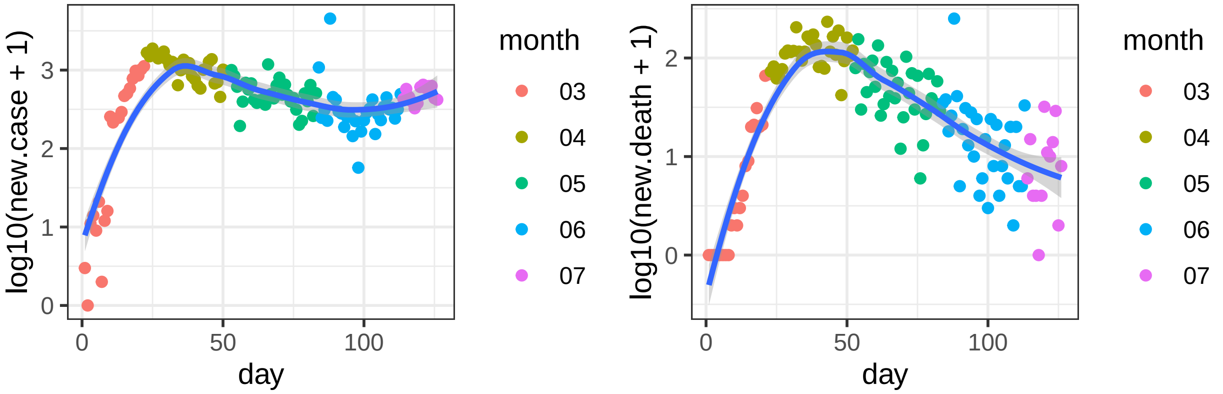
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−25

Pennsylvania



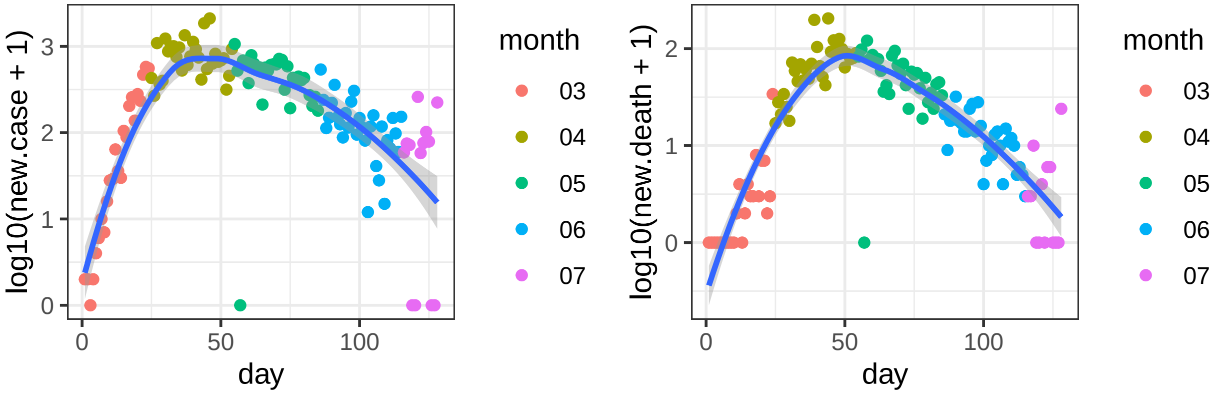
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Michigan



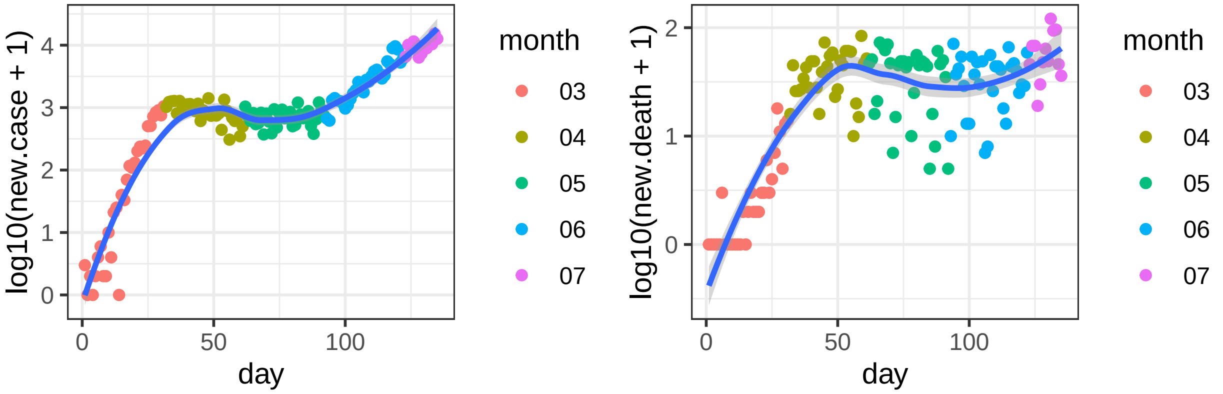
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Connecticut



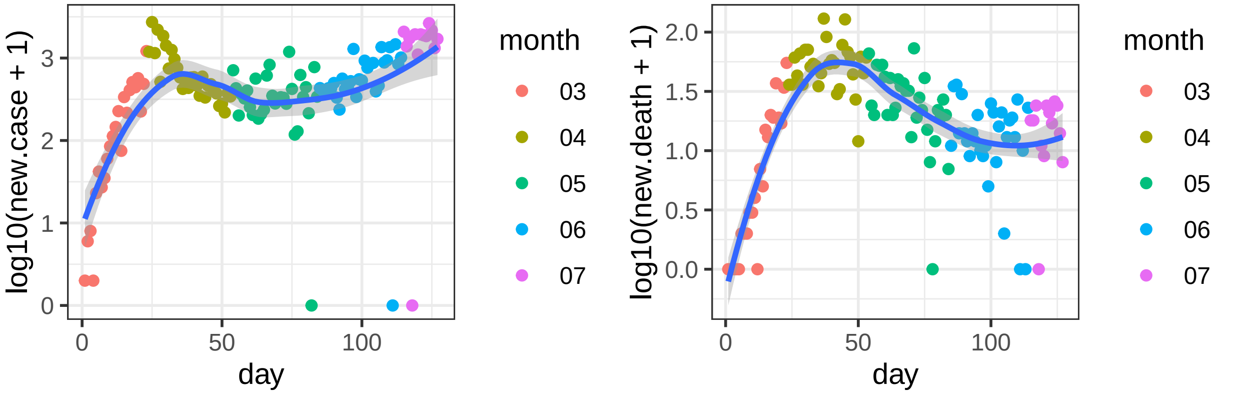
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Florida



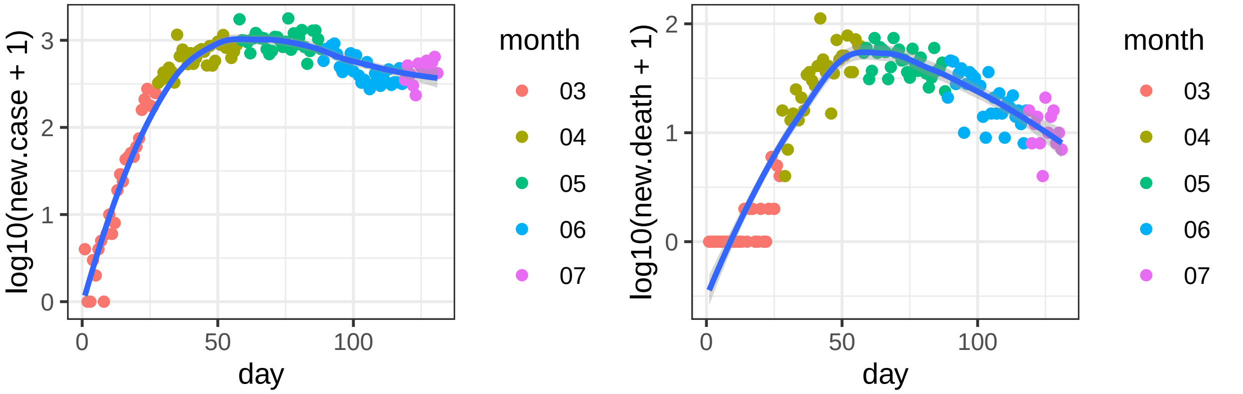
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Louisiana



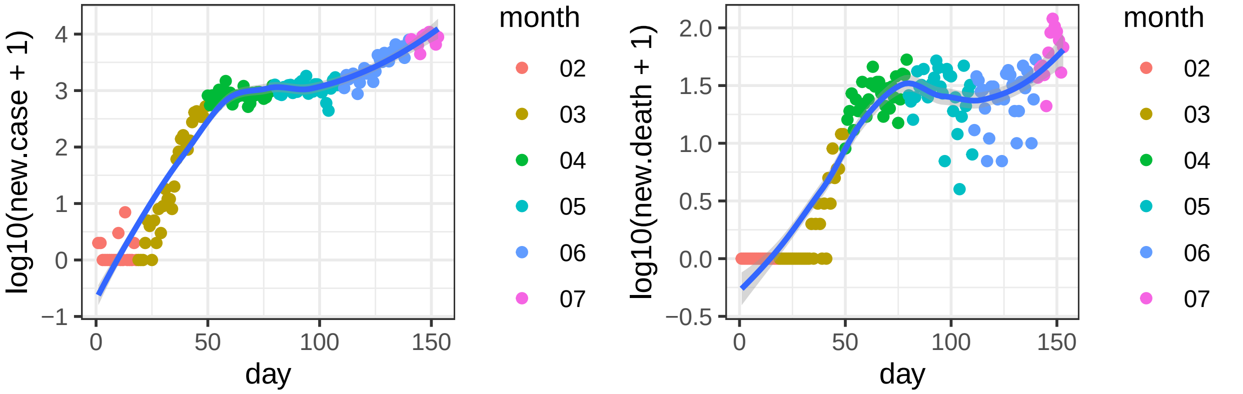
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Maryland



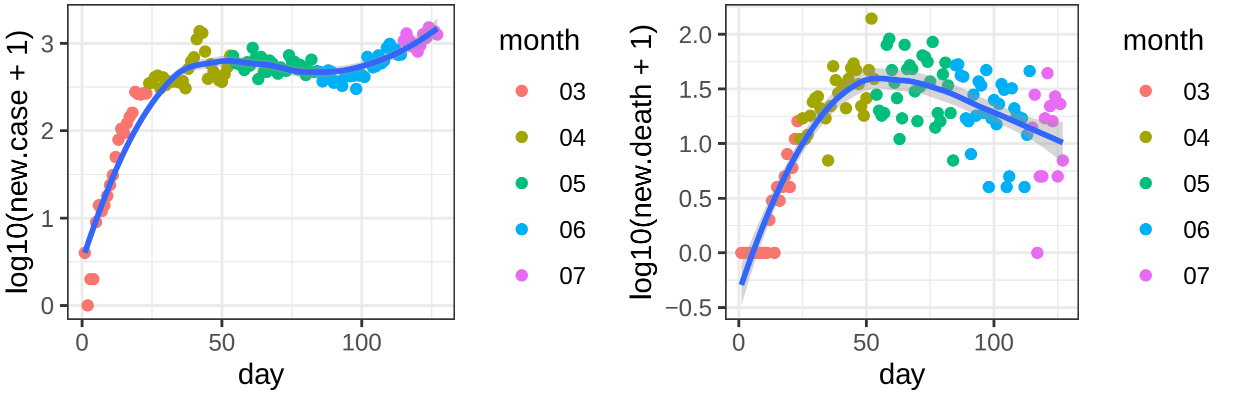
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Texas



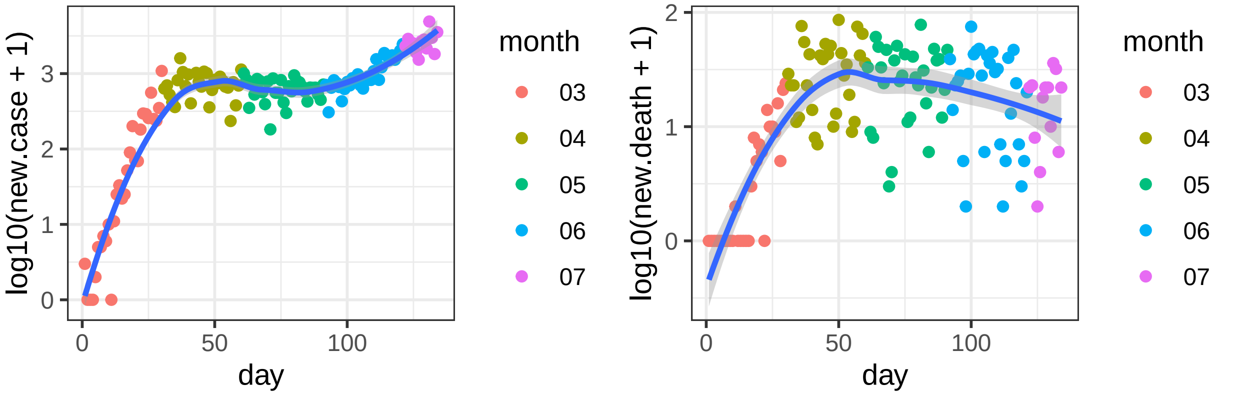
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−12

Ohio



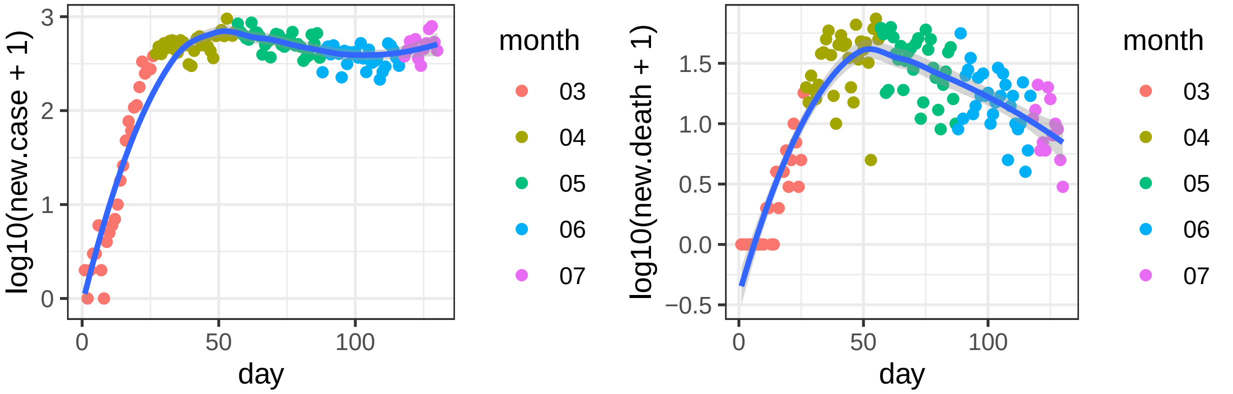
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Georgia



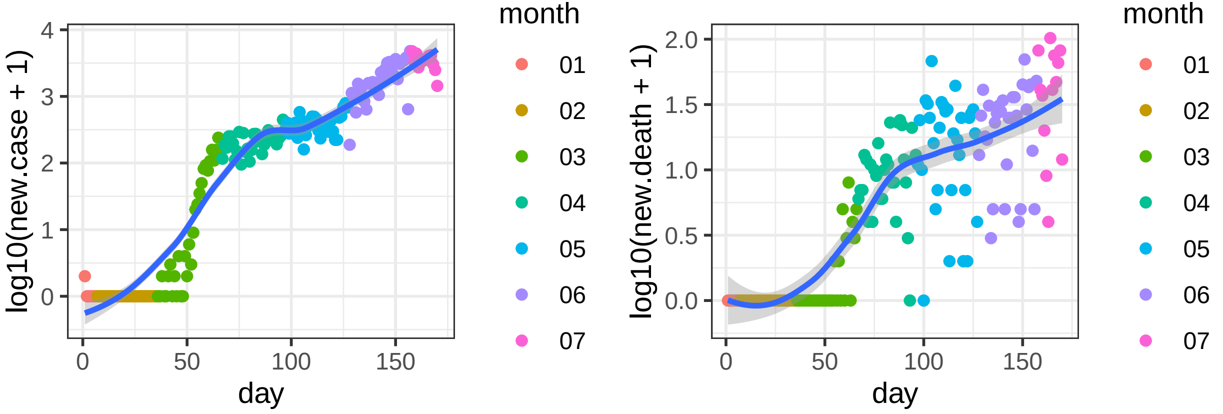
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Indiana



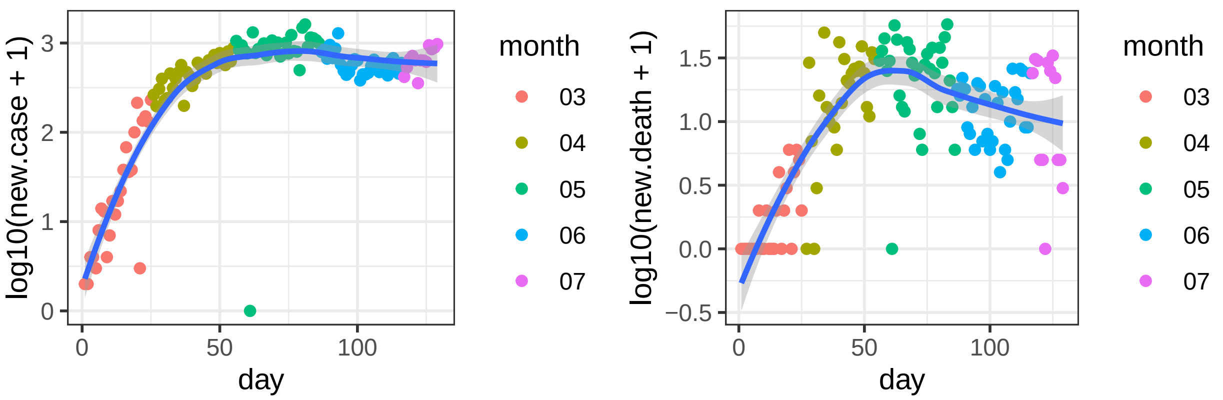
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Arizona



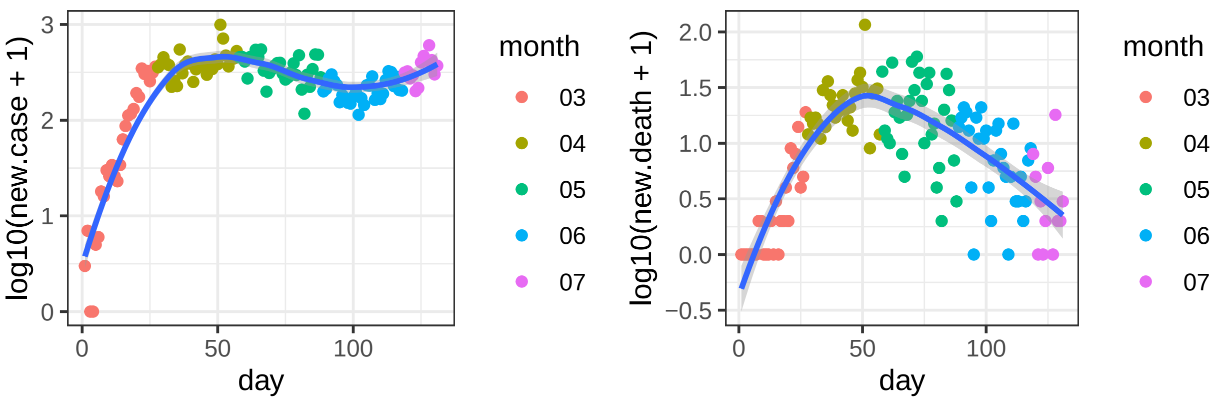
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Virginia



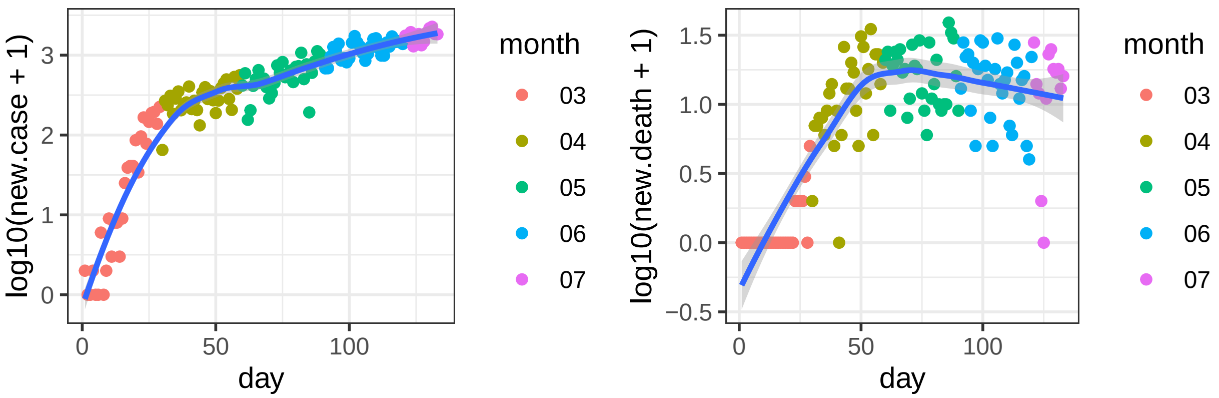
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Colorado



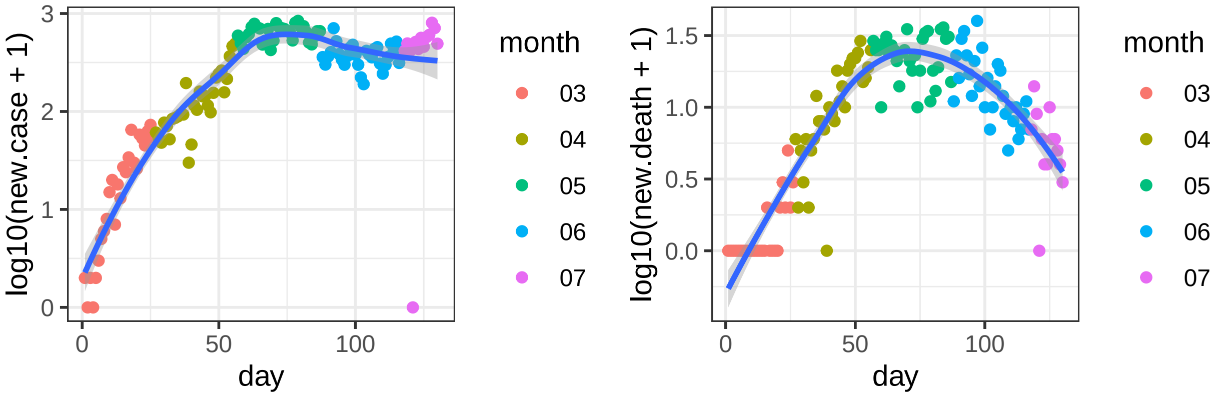
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

North Carolina



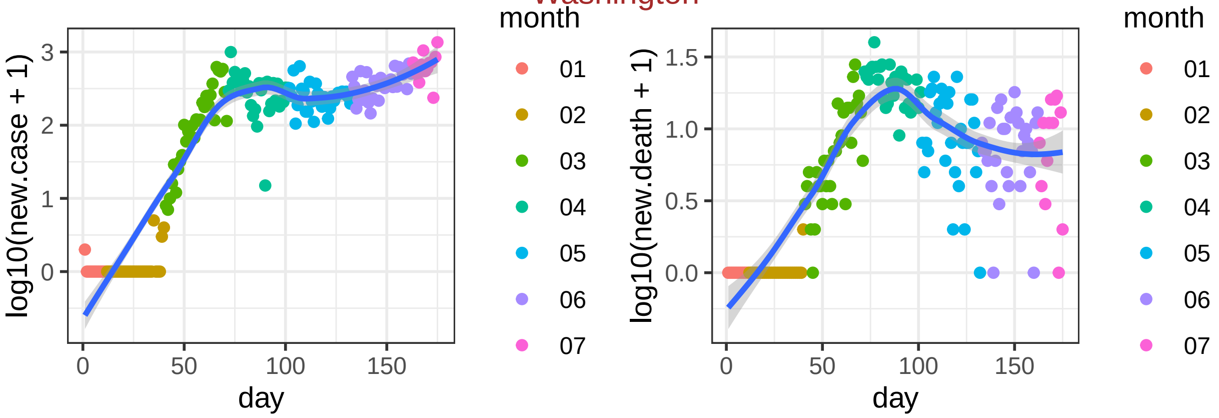
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−03

Minnesota



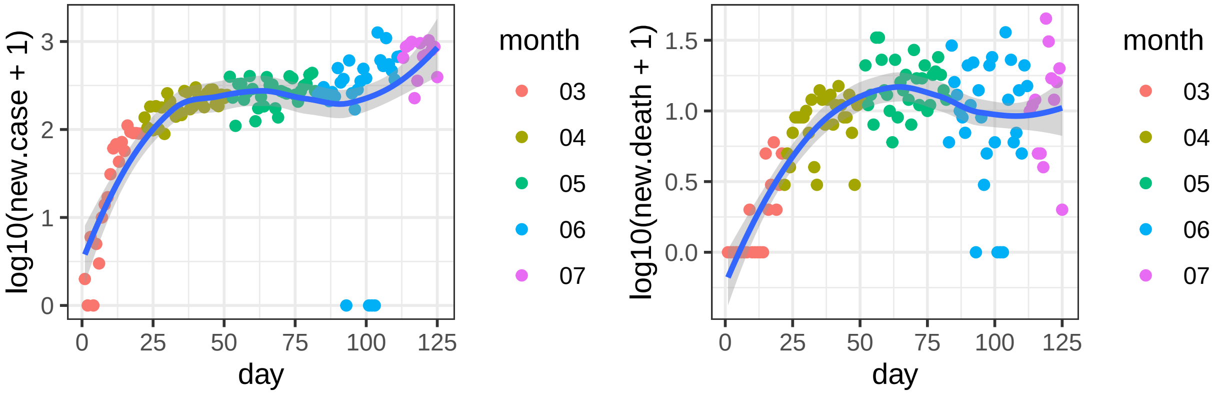
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Washington



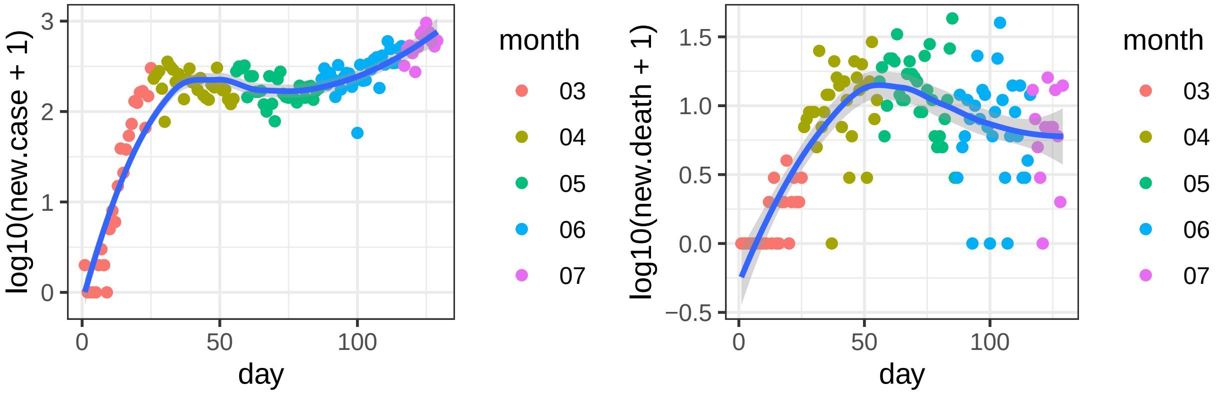
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−21

Mississippi



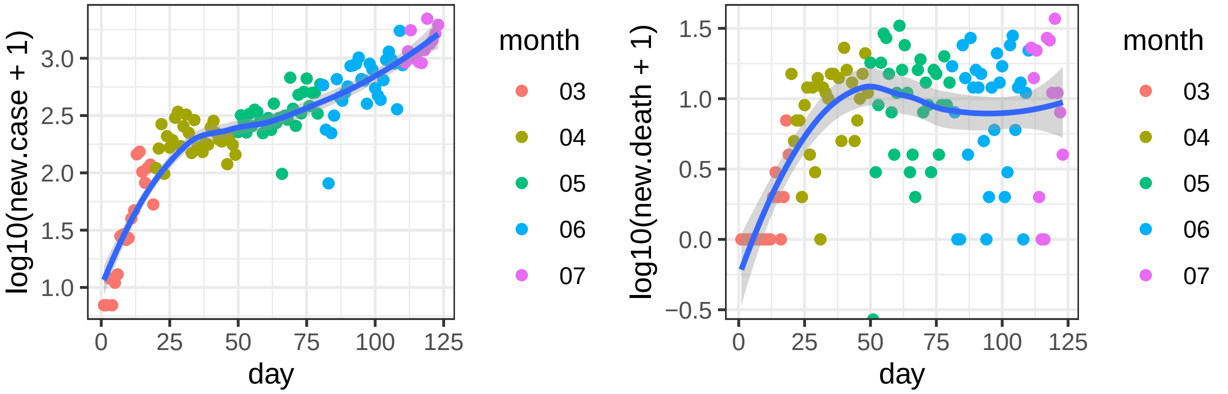
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Missouri



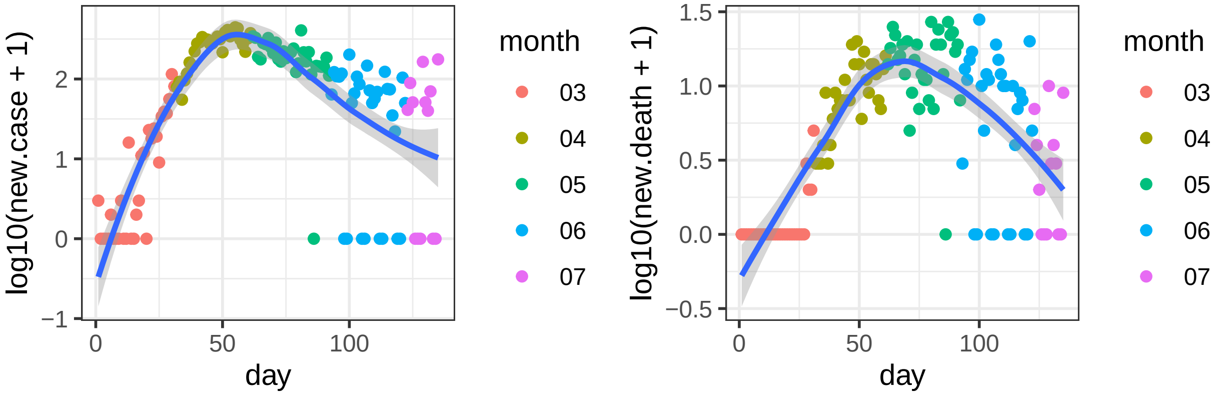
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Alabama



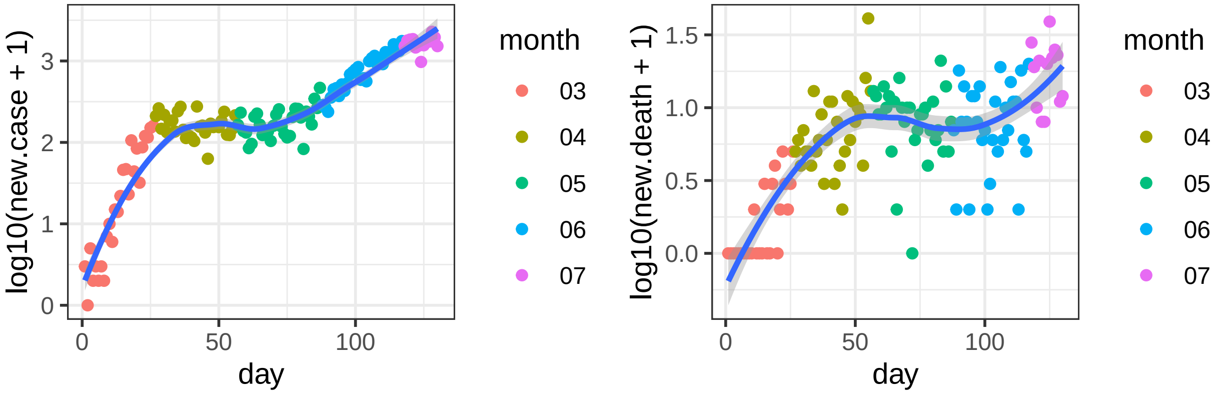
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Rhode Island



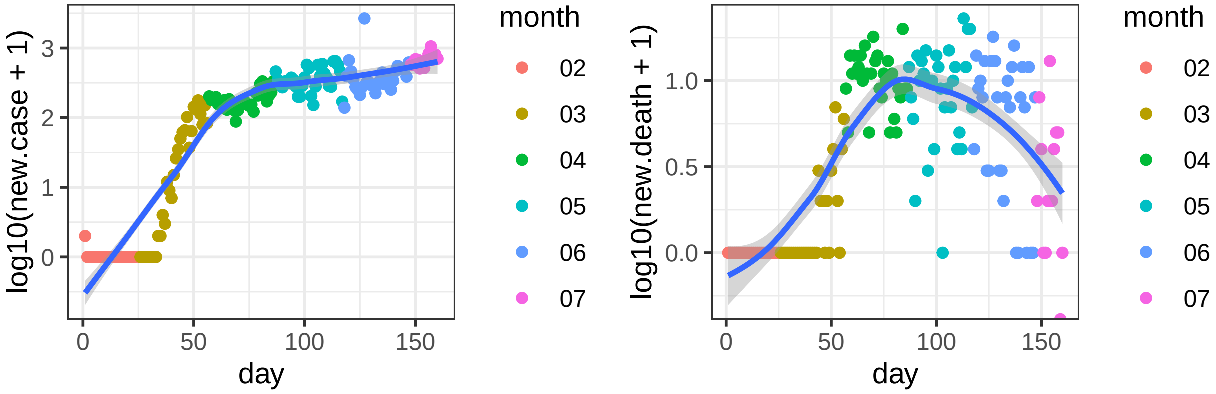
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

South Carolina

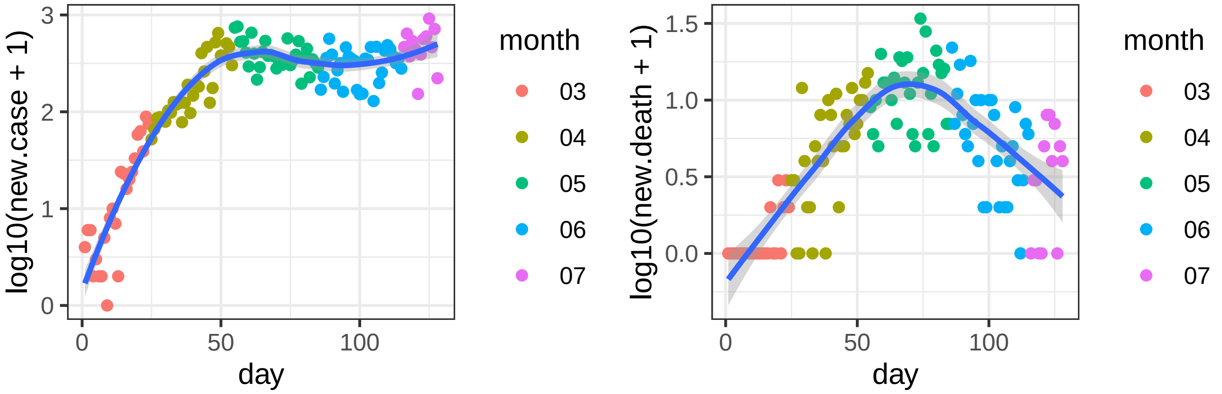


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Wisconsin

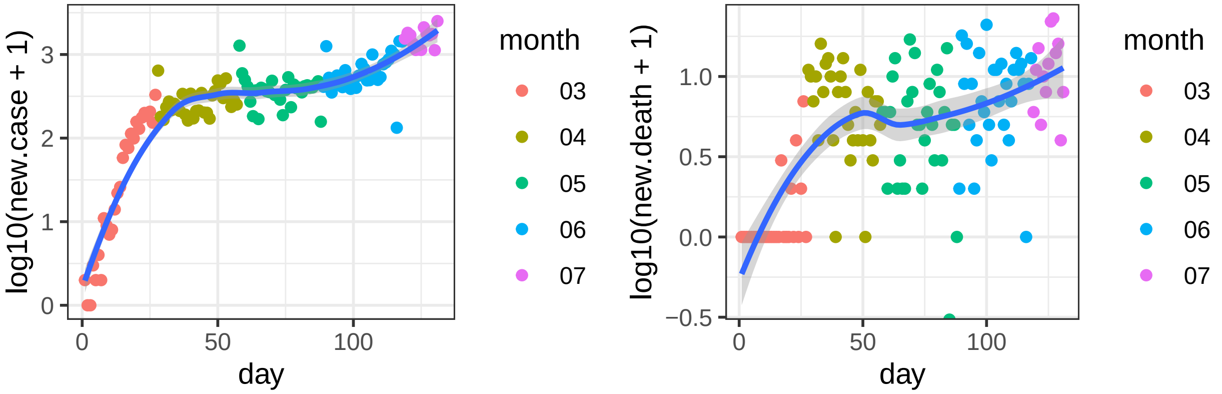


data source: https://github.com/nytimes/covid−19−data, day 1 is 02−05



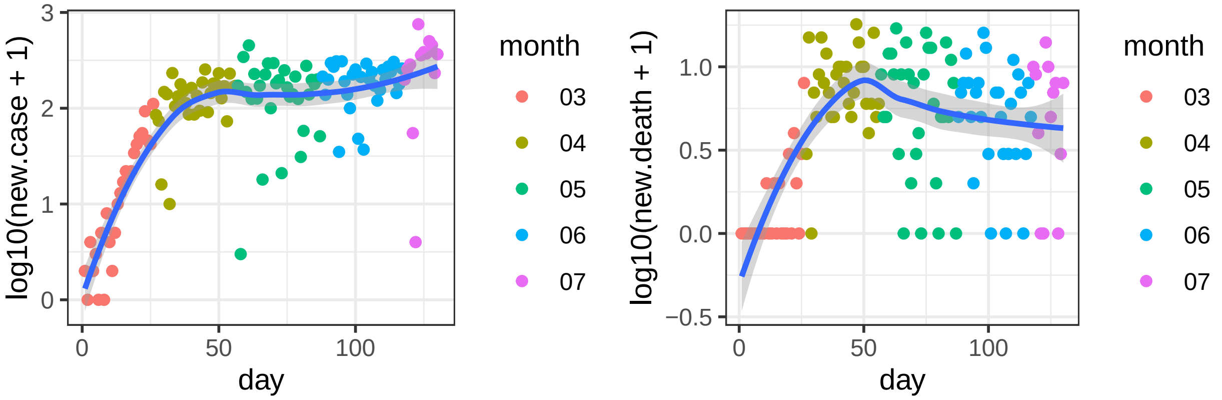
Iowa

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



Tennessee

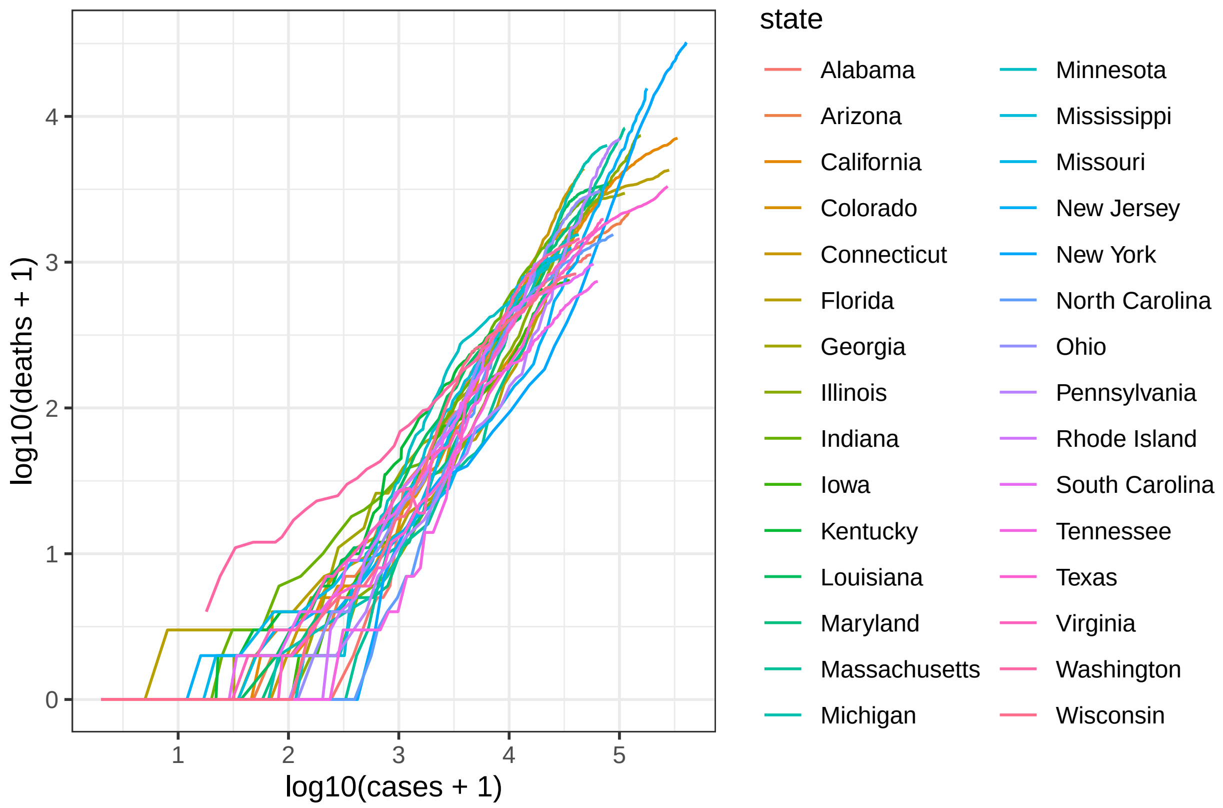
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05



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

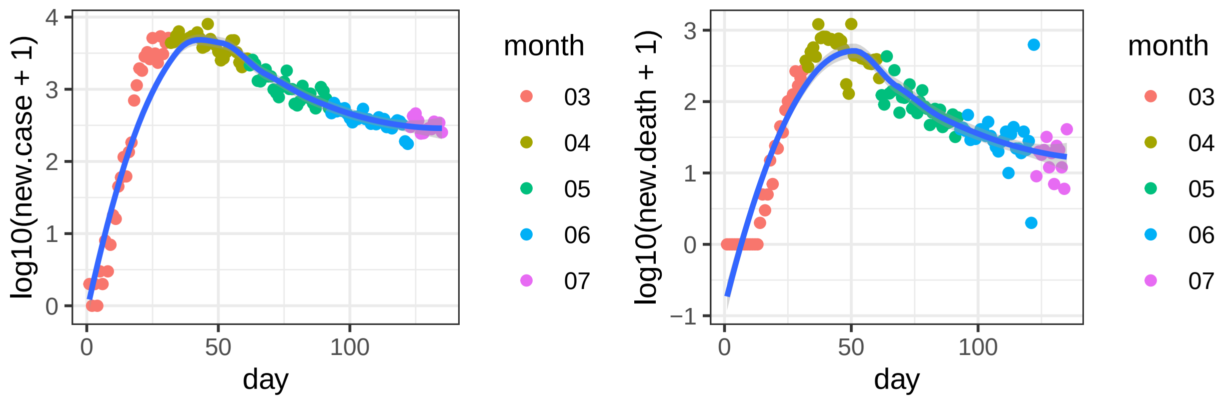
county level data

First check the 50 counties with the largest number of deaths.

```
##                date               county                state  fips  cases deaths
## 329828 2020-07-13        New York City             New York    NA 223977  22795
## 328603 2020-07-13                 Cook             Illinois 17031  95884   4729
## 328200 2020-07-13          Los Angeles           California  6037 136129   3822
## 329309 2020-07-13                Wayne             Michigan 26163  24100   2763
## 329827 2020-07-13               Nassau             New York 36059  42354   2701
## 329752 2020-07-13                Essex           New Jersey 34013  19288   2077
## 329847 2020-07-13              Suffolk             New York 36103  42112   2039
## 329747 2020-07-13               Bergen           New Jersey 34003  20163   2030
## 329220 2020-07-13            Middlesex        Massachusetts 25017  24536   1921
## 330259 2020-07-13         Philadelphia         Pennsylvania 42101  27575   1640
## 329855 2020-07-13          Westchester             New York 36119  35326   1567
## 329754 2020-07-13               Hudson           New Jersey 34017  19314   1484
## 328301 2020-07-13             Hartford          Connecticut  9003  11949   1391
## 328300 2020-07-13            Fairfield          Connecticut  9001  16954   1385
## 329757 2020-07-13            Middlesex           New Jersey 34023  17265   1377
## 329765 2020-07-13                Union           New Jersey 34039  16703   1343
## 329761 2020-07-13              Passaic           New Jersey 34031  17238   1224
## 328356 2020-07-13           Miami-Dade              Florida 12086  67712   1143
## 328098 2020-07-13             Maricopa              Arizona  4013  81216   1140
## 329216 2020-07-13                Essex        Massachusetts 25009  16485   1136
## 329289 2020-07-13              Oakland             Michigan 26125  12752   1109
## 328304 2020-07-13            New Haven          Connecticut  9009  12612   1087
## 329224 2020-07-13              Suffolk        Massachusetts 25025  20342   1020
## 329760 2020-07-13                Ocean           New Jersey 34029   9901    997
## 329226 2020-07-13            Worcester        Massachusetts 25027  12679    959
## 329222 2020-07-13              Norfolk        Massachusetts 25021   9485    958
## 329276 2020-07-13               Macomb             Michigan 26099   8117    933
## 329758 2020-07-13             Monmouth           New Jersey 34025   9613    838
## 330254 2020-07-13           Montgomery         Pennsylvania 42091   8876    829
## 329759 2020-07-13               Morris           New Jersey 34027   7063    819
## 329337 2020-07-13             Hennepin            Minnesota 27053  13697    791
## 330358 2020-07-13           Providence         Rhode Island 44007  13144    774
## 329202 2020-07-13           Montgomery             Maryland 24031  15818    760
## 328739 2020-07-13               Marion              Indiana 18097  12243    739
## 329203 2020-07-13      Prince George's             Maryland 24033  20263    714
## 330231 2020-07-13             Delaware         Pennsylvania 42045   7569    709
## 329223 2020-07-13             Plymouth        Massachusetts 25023   8822    679
## 329218 2020-07-13              Hampden        Massachusetts 25013   7019    678
## 331007 2020-07-13                 King           Washington 53033  12101    639
## 328363 2020-07-13           Palm Beach              Florida 12099  21804    611
## 329582 2020-07-13            St. Louis             Missouri 29189   7897    603
## 329813 2020-07-13                 Erie             New York 36029   7766    603
## 329756 2020-07-13               Mercer           New Jersey 34021   7839    602
## 329214 2020-07-13              Bristol        Massachusetts 25005   8491    601
## 330217 2020-07-13                Bucks         Pennsylvania 42017   6105    572
## 328313 2020-07-13 District of Columbia District of Columbia 11001  10906    568
## 329749 2020-07-13               Camden           New Jersey 34007   7768    551
## 329763 2020-07-13             Somerset           New Jersey 34035   5104    551
## 328213 2020-07-13            Riverside           California  6065  26404    550
## 329140 2020-07-13              Orleans            Louisiana 22071   8745    540
```

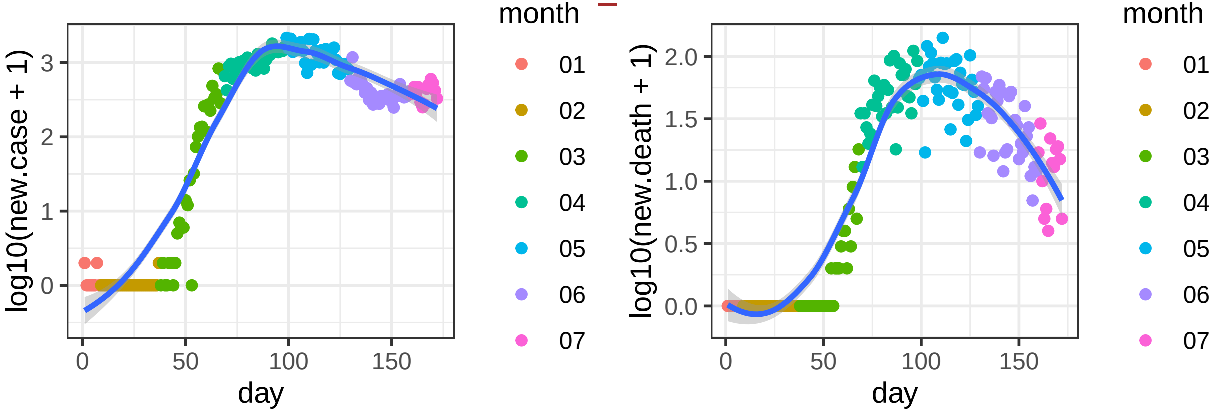
For these 50 counties, I check the number of new cases and the number of new deaths.

New York City_New York



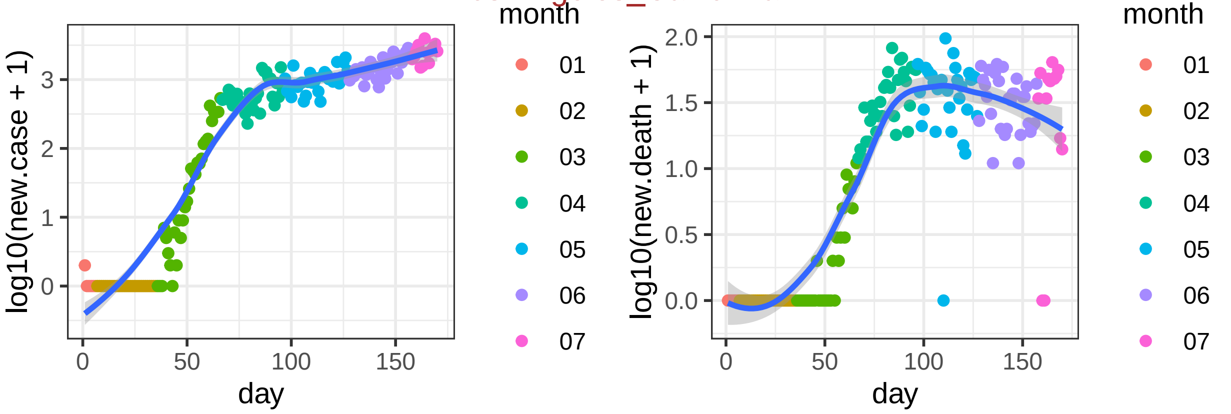
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Cook_Illinois



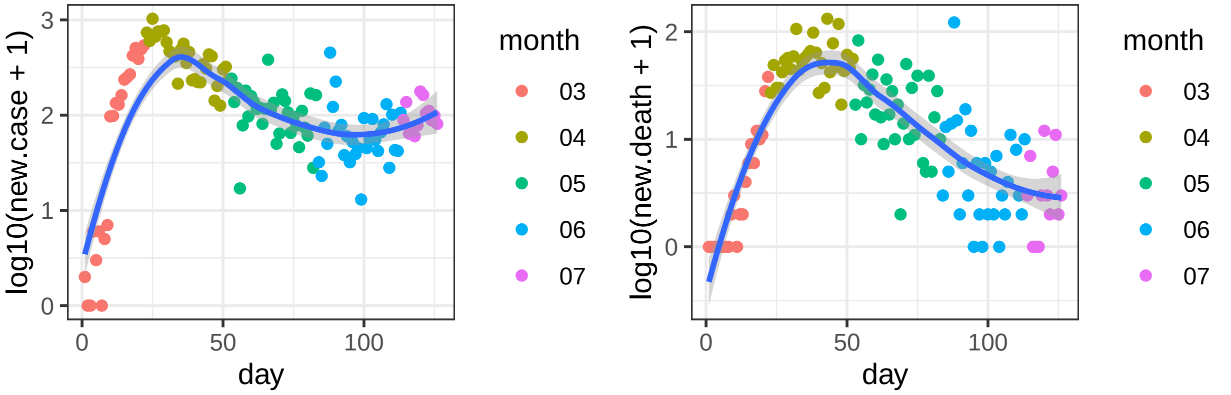
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−24

Los Angeles_California



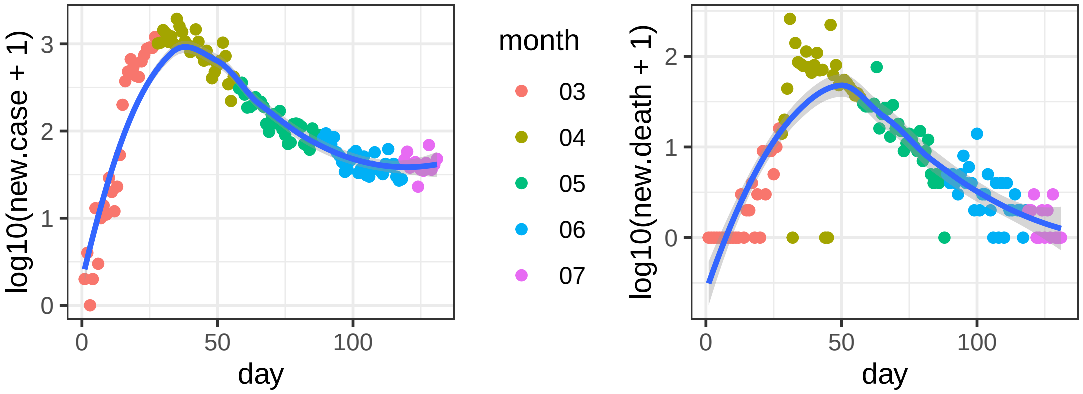
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Wayne_Michigan



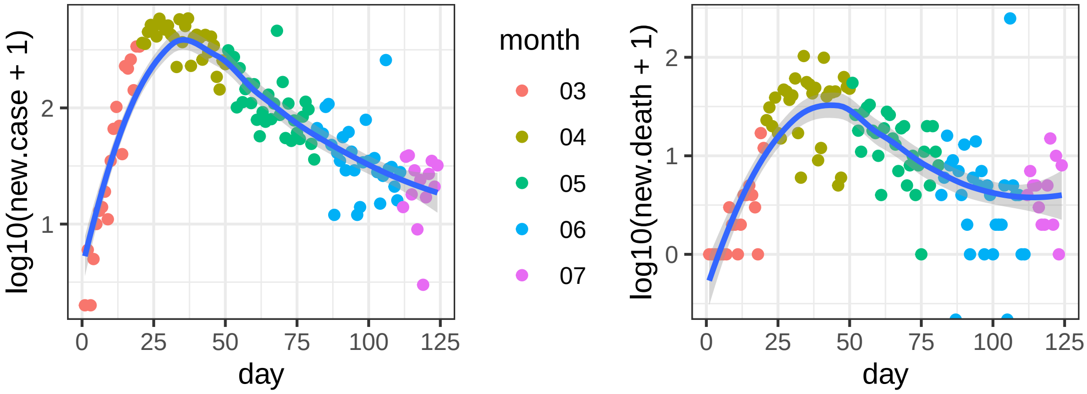
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Nassau_New York



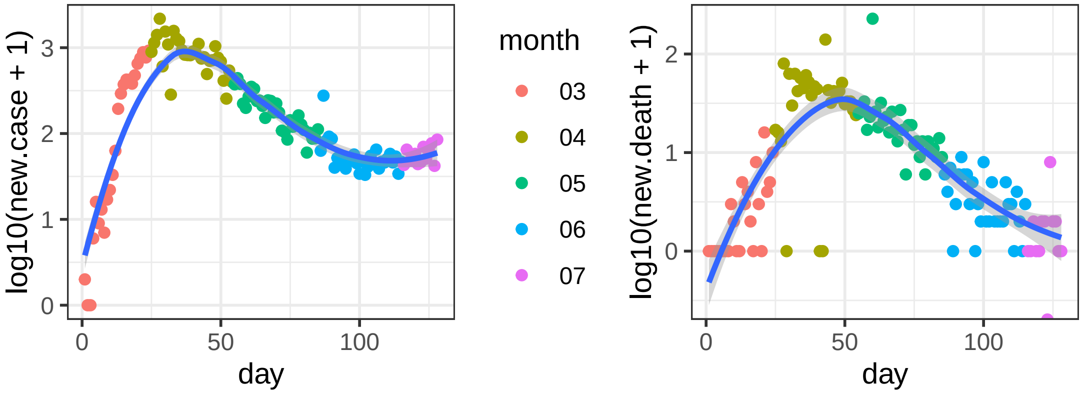
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Essex_New Jersey



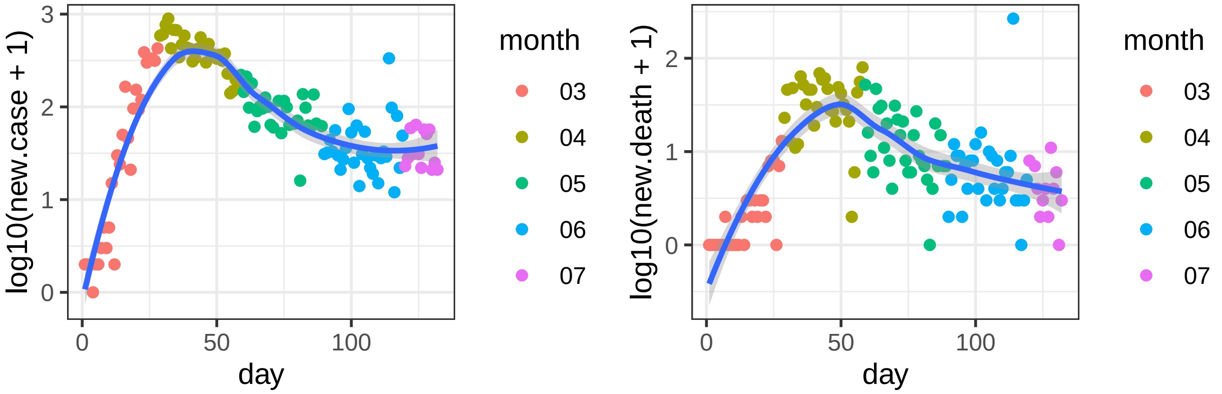
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Suffolk_New York



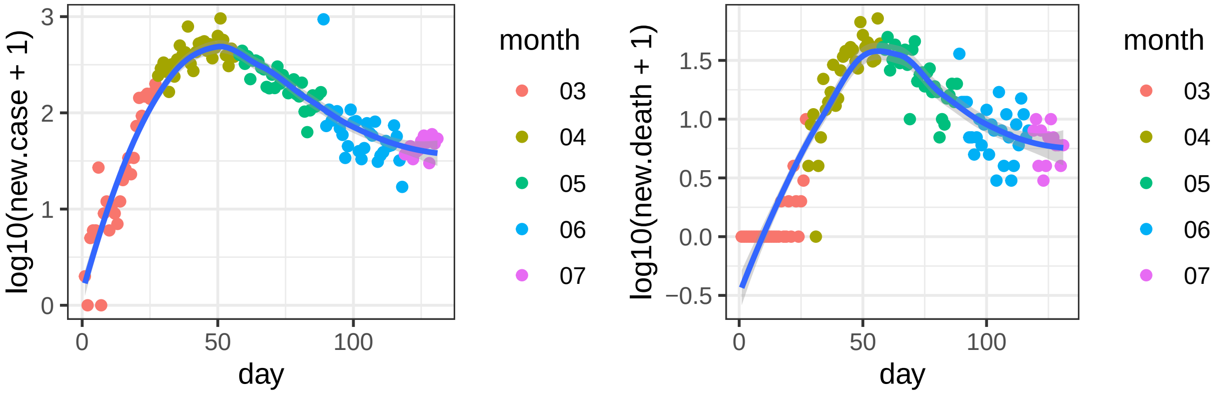
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Bergen_New Jersey



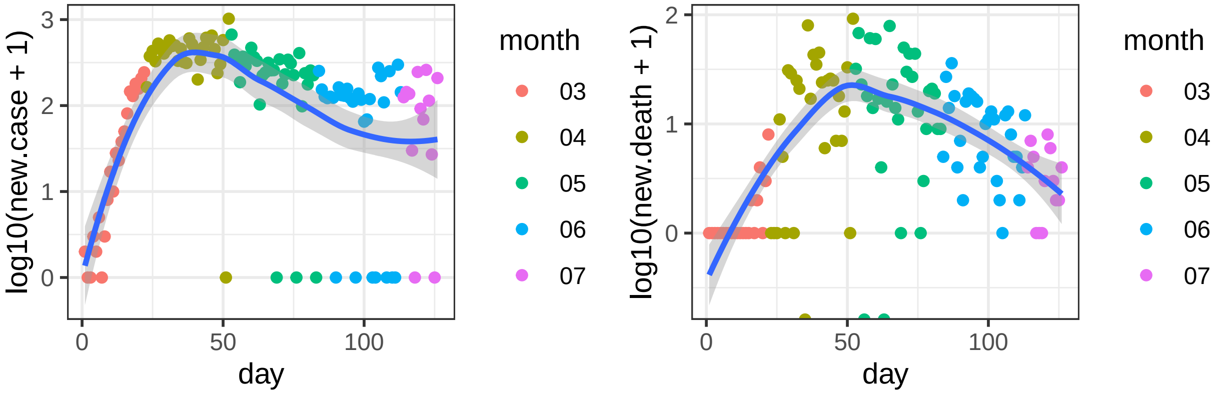
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Middlesex_Massachusetts

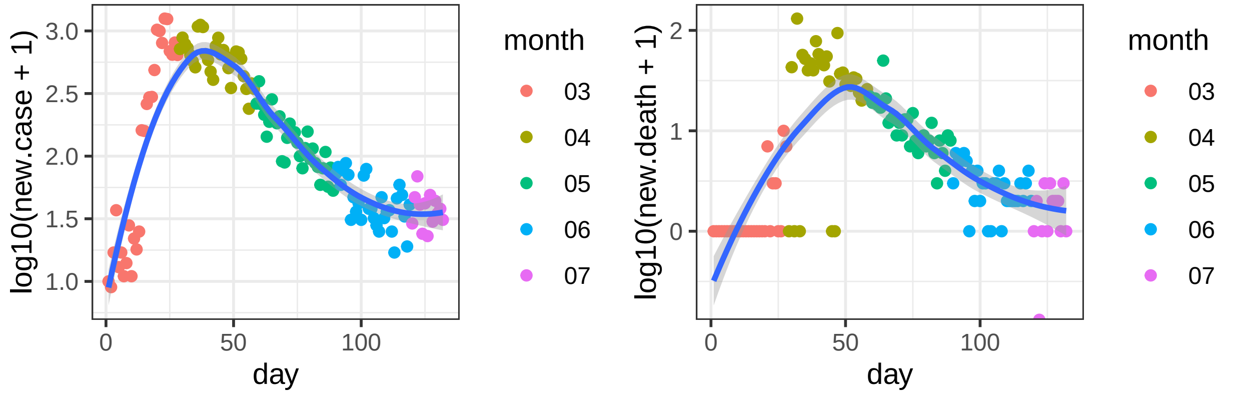


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Philadelphia_Pennsylvania

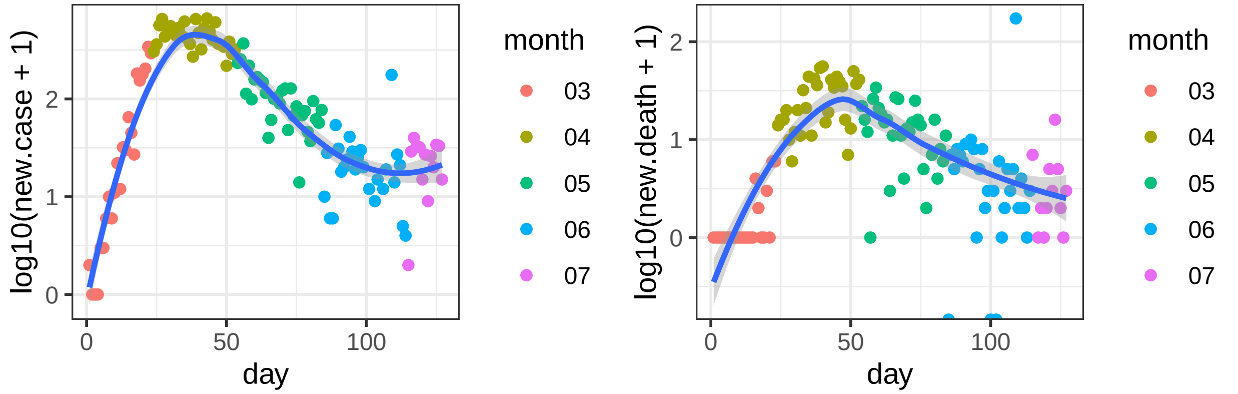


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10



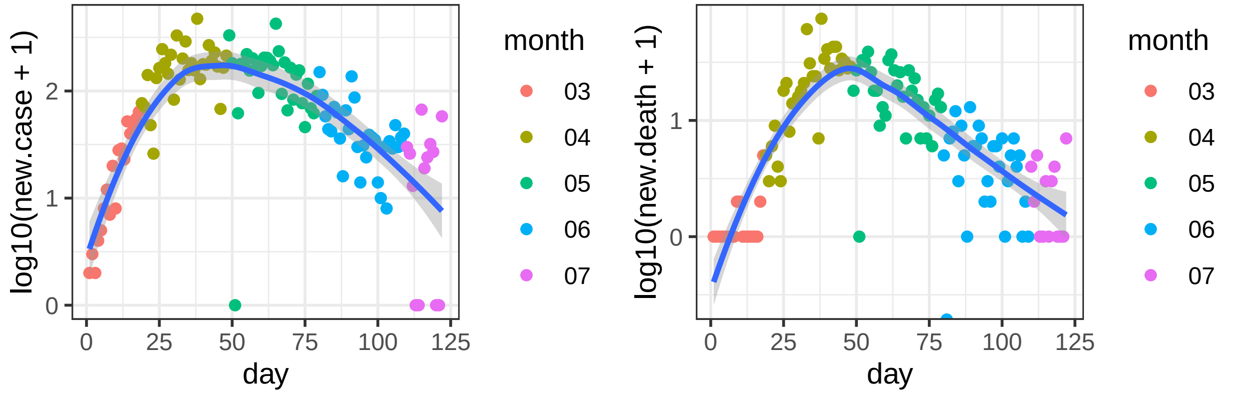
Westchester_New York

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



Hudson_New Jersey

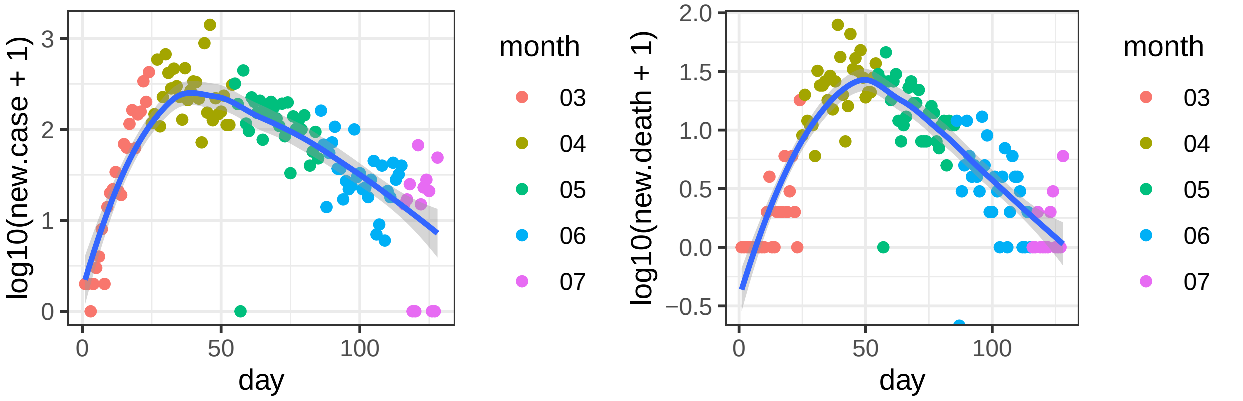
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09



Hartford_Connecticut

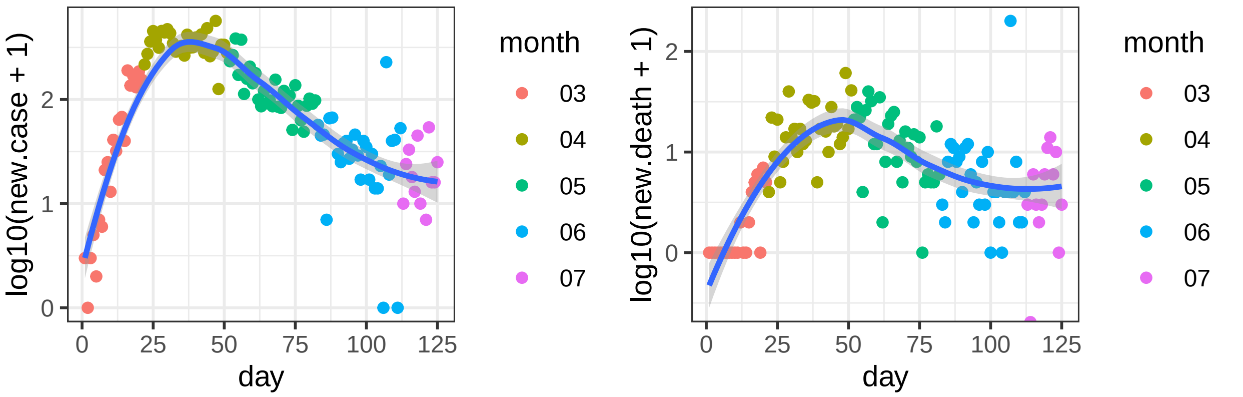
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Fairfield_Connecticut



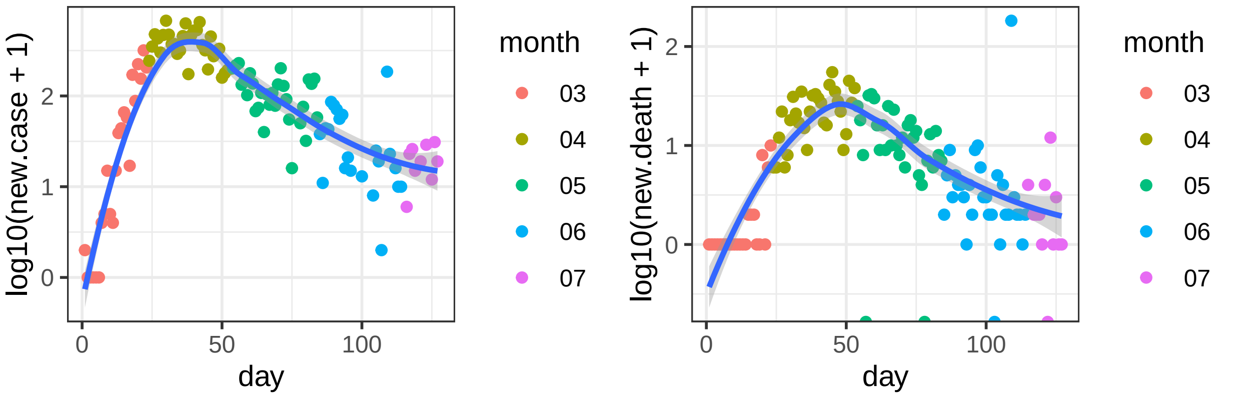
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Middlesex_New Jersey



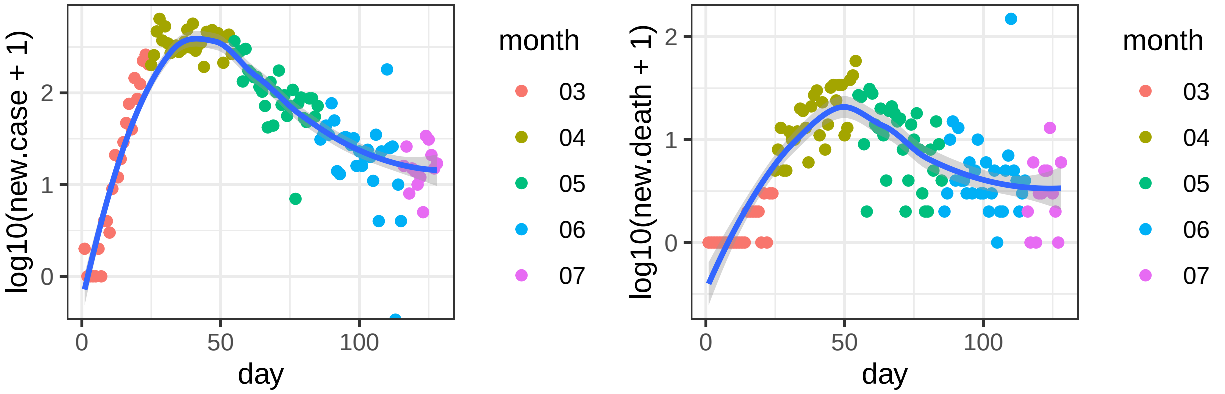
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Union_New Jersey



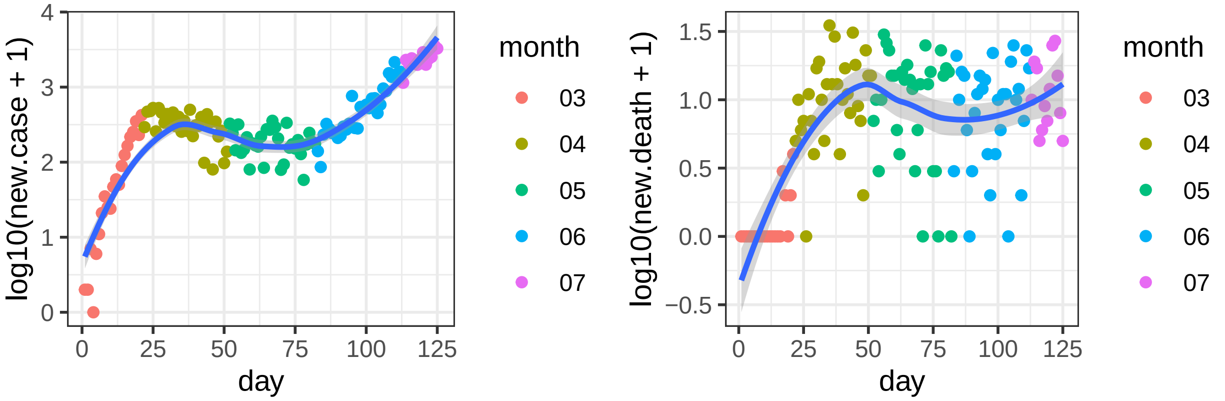
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Passaic_New Jersey



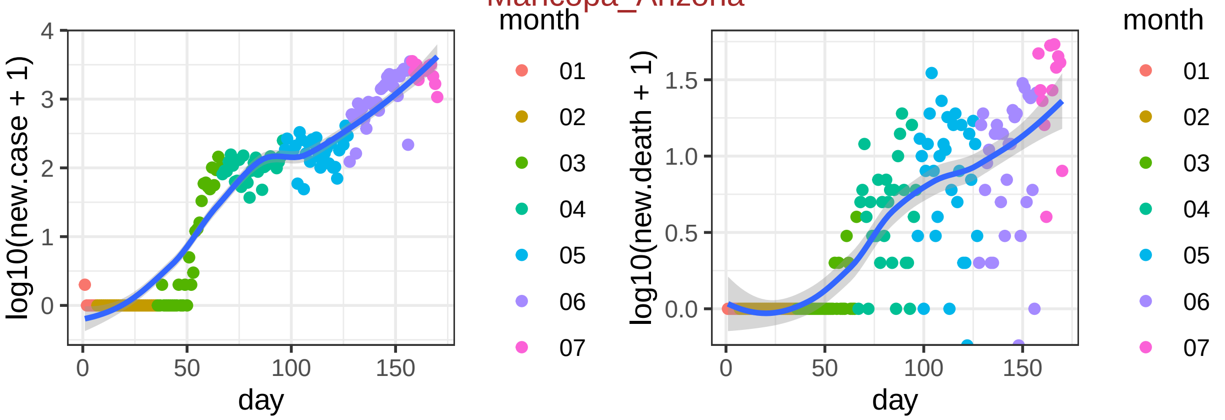
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Miami−Dade_Florida



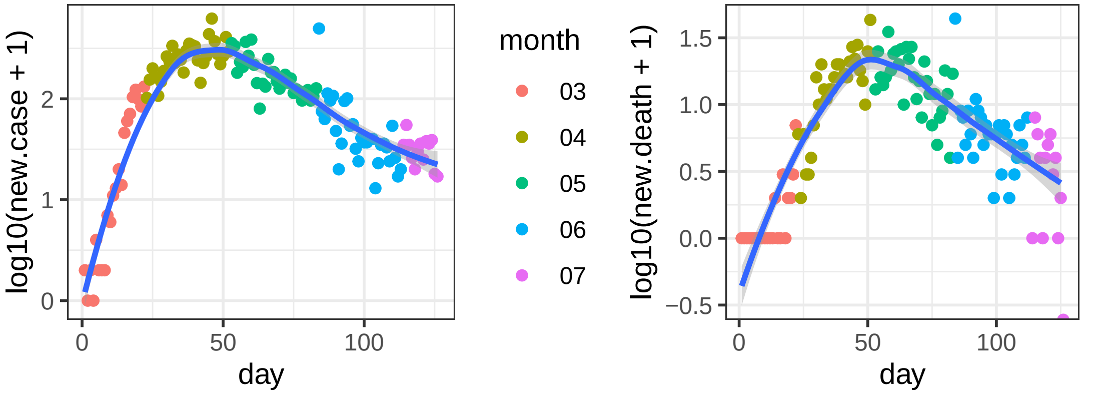
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Maricopa_Arizona



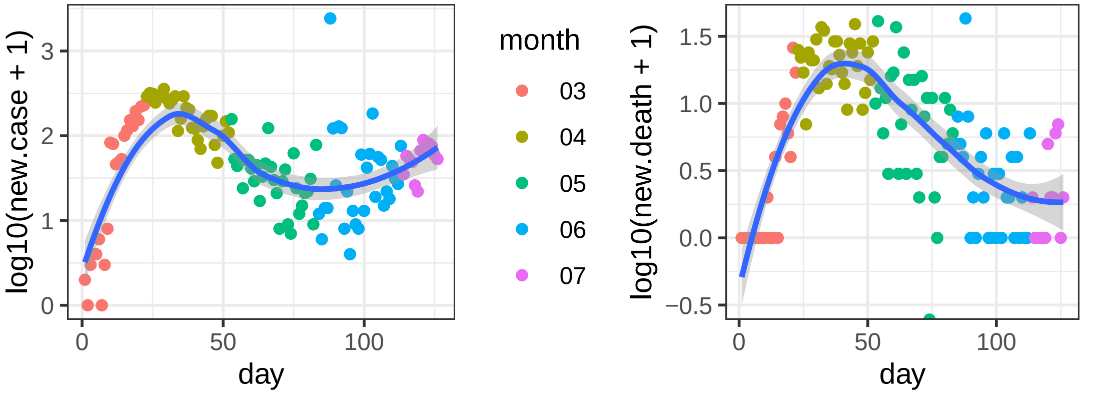
data source: https://github.com/nytimes/covid−19−data, day 1 is 01−26

Essex_Massachusetts



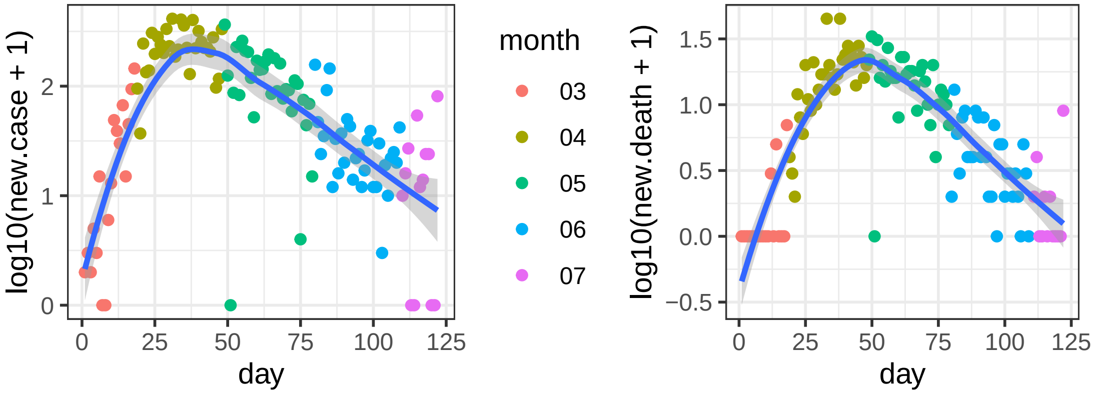
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Oakland_Michigan



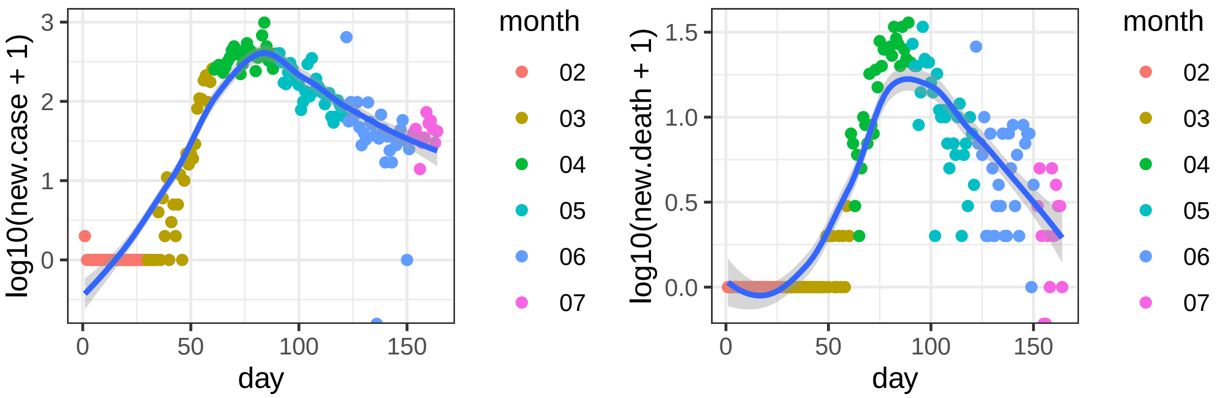
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

New Haven_Connecticut



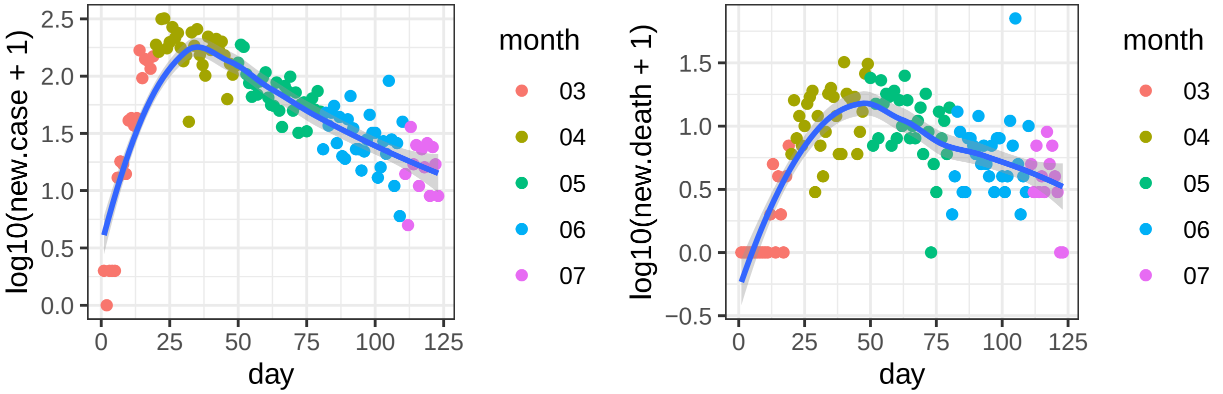
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Suffolk_Massachusetts



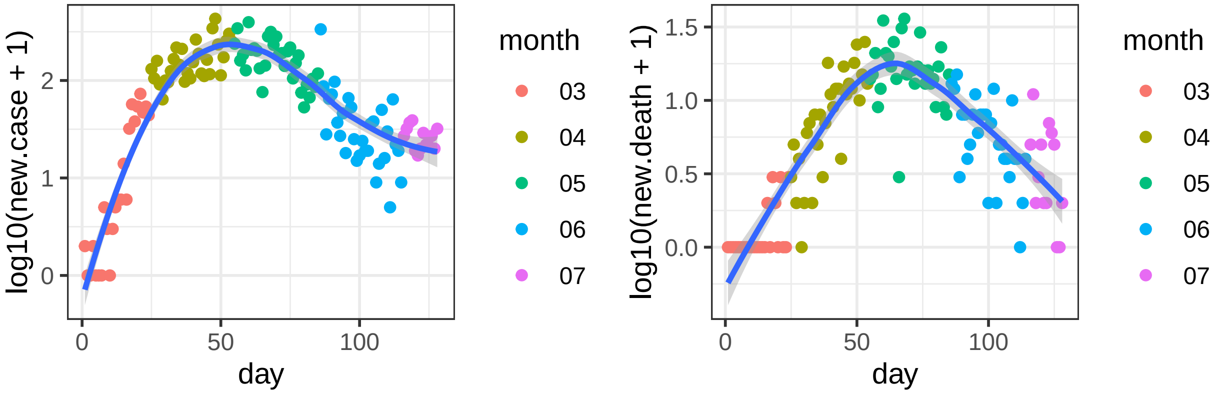
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−01

Ocean_New Jersey



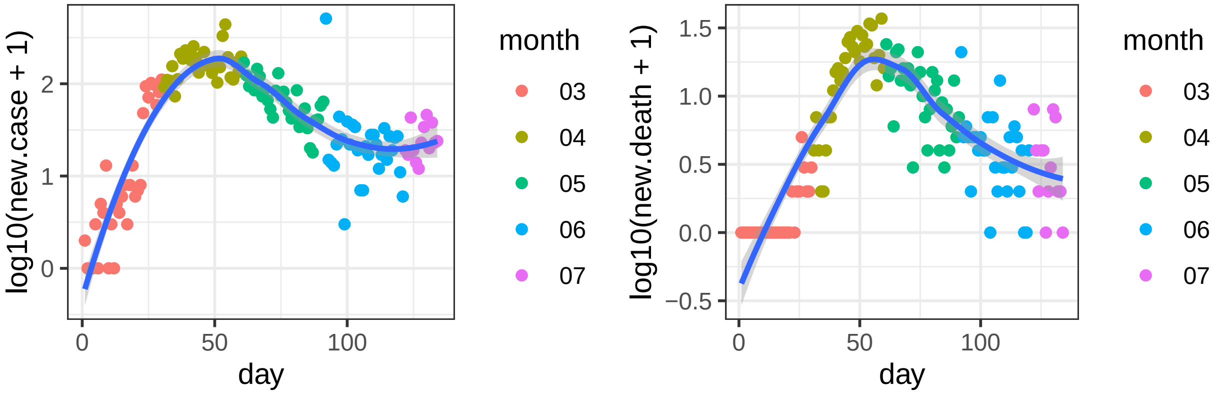
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Worcester_Massachusetts



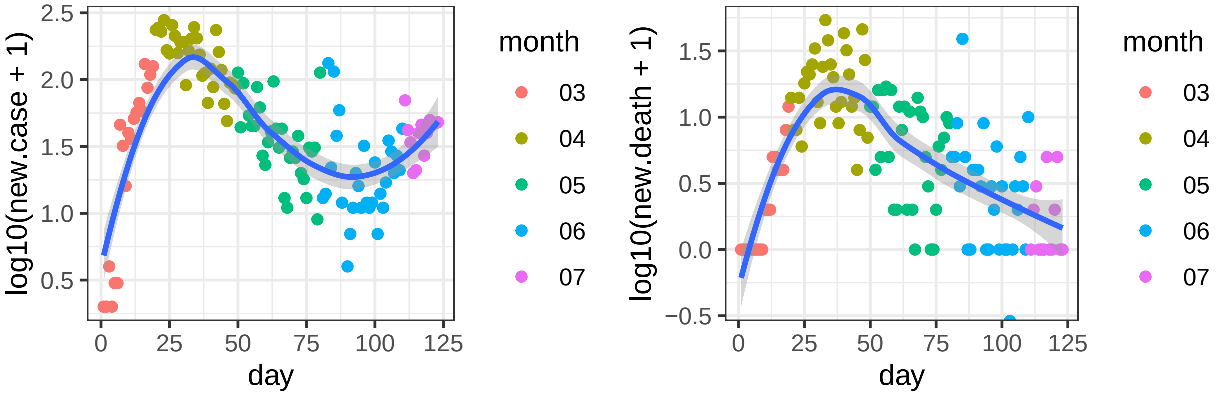
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Norfolk_Massachusetts



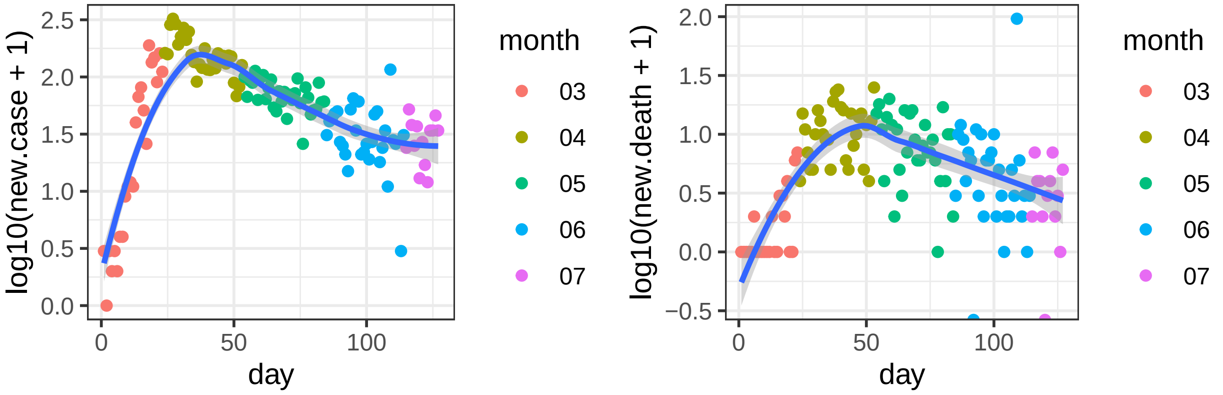
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Macomb_Michigan



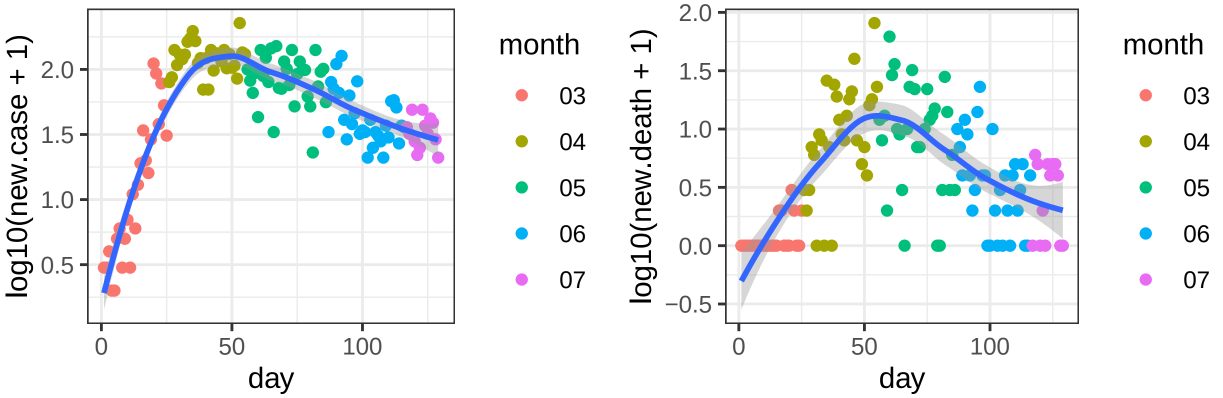
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Monmouth_New Jersey



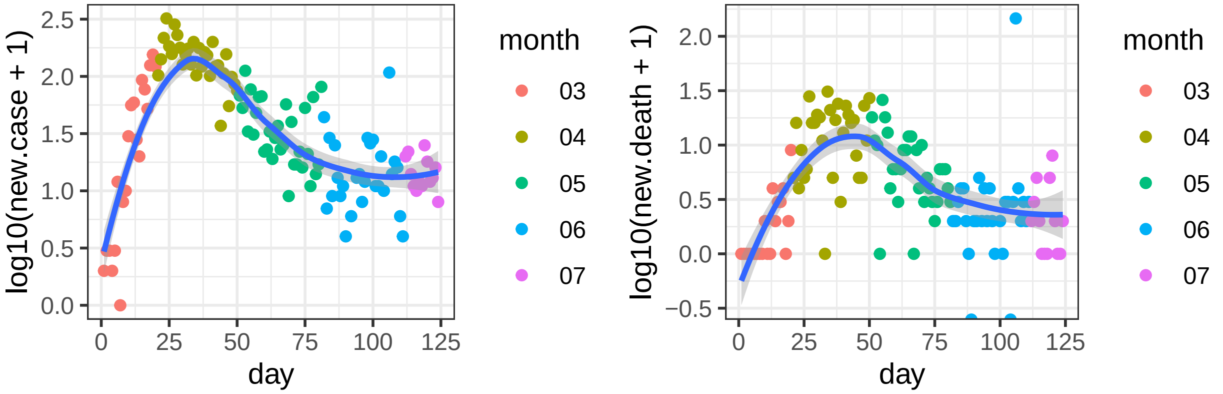
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Montgomery_Pennsylvania



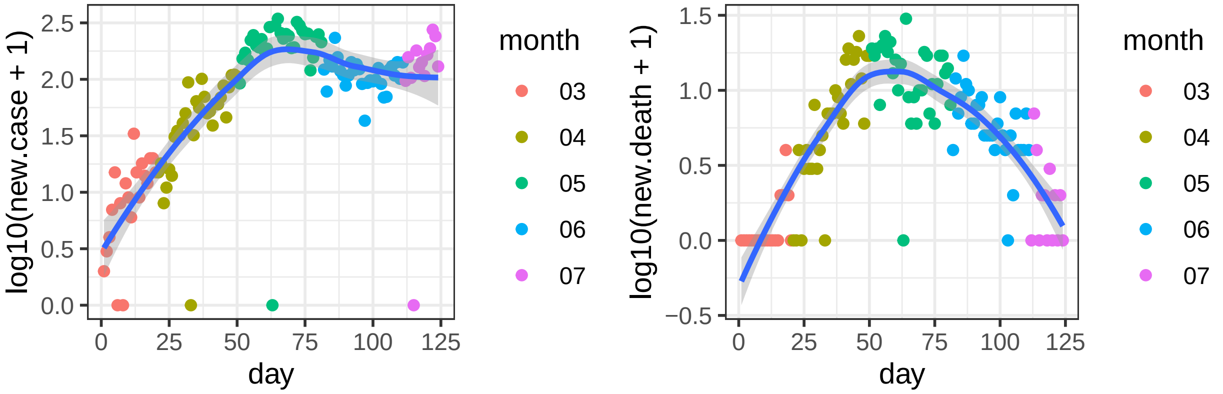
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Morris_New Jersey



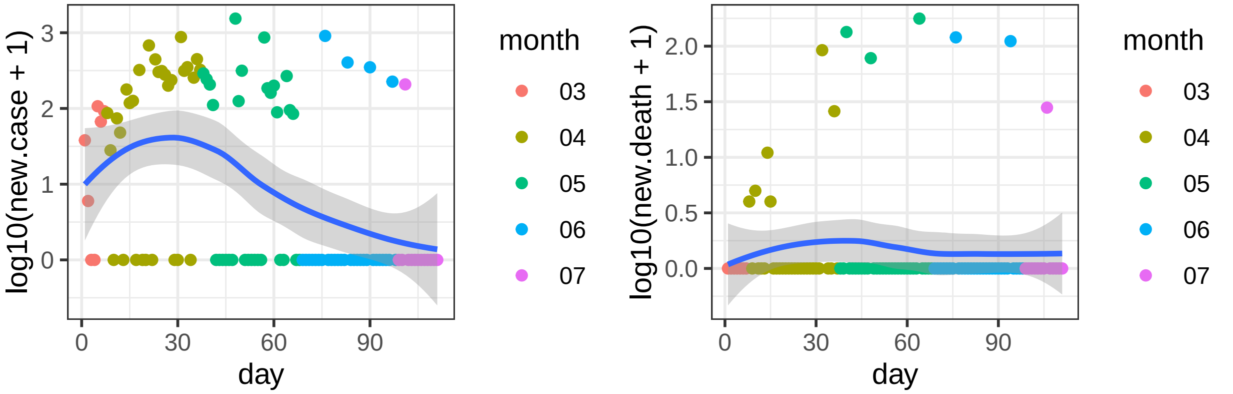
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Hennepin_Minnesota



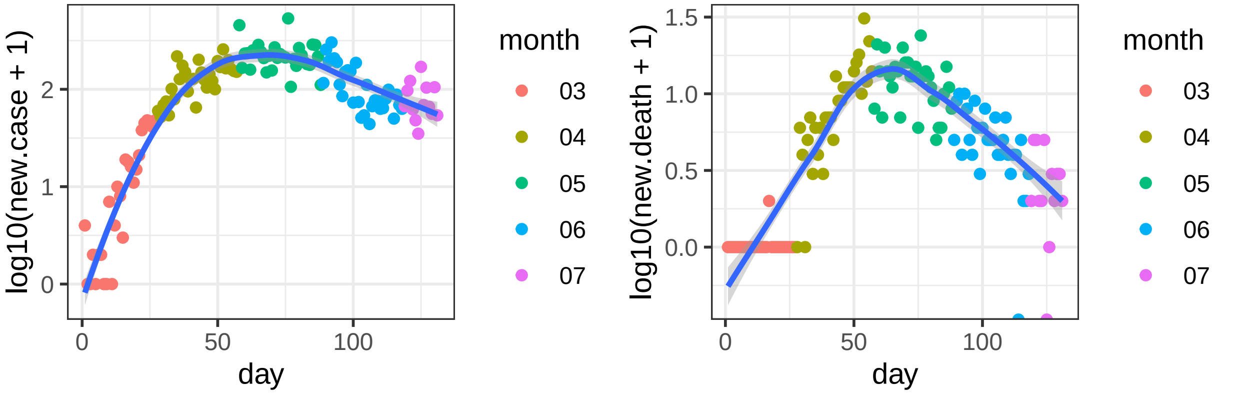
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Providence_Rhode Island



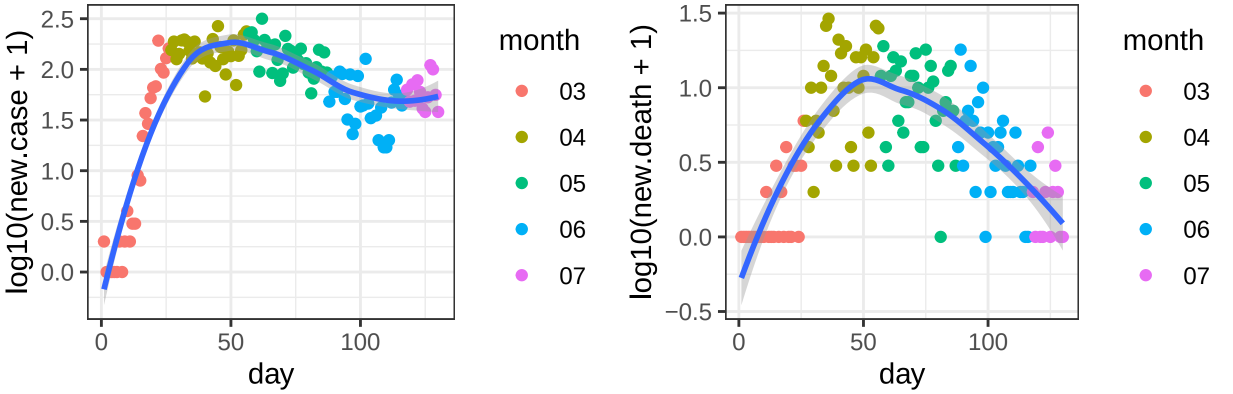
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−25

Montgomery_Maryland



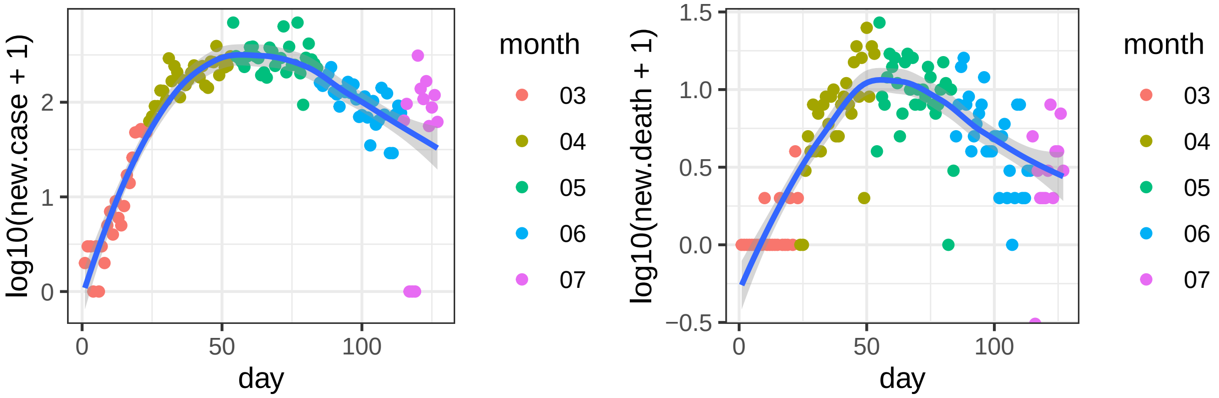
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Marion_Indiana



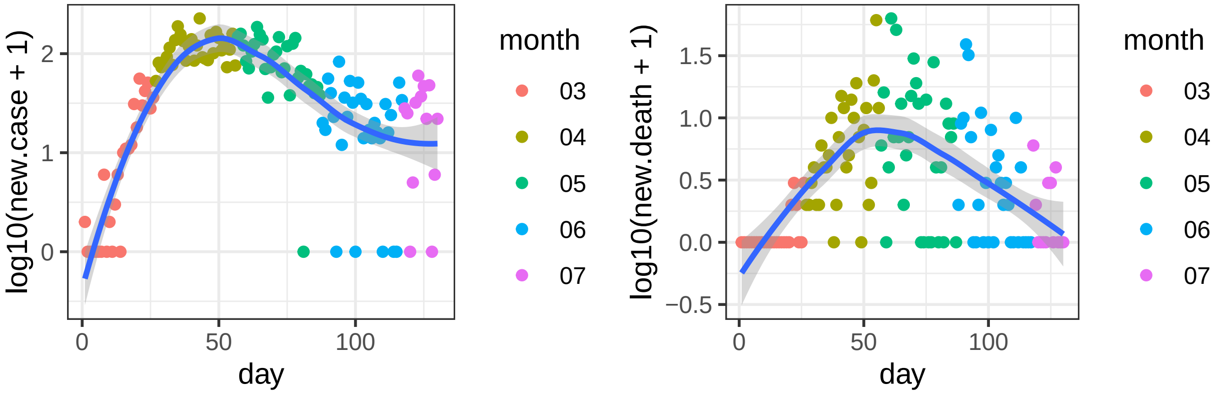
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Prince George's_Maryland



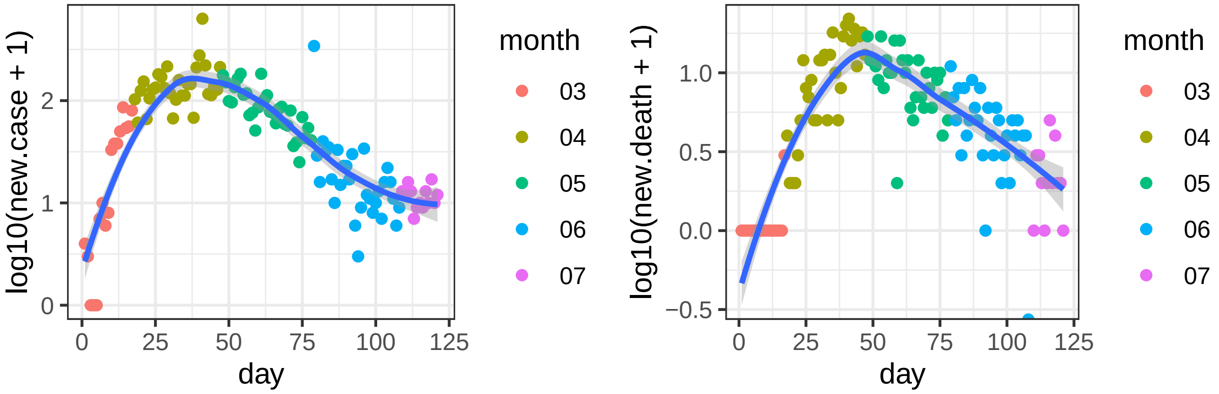
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Delaware_Pennsylvania



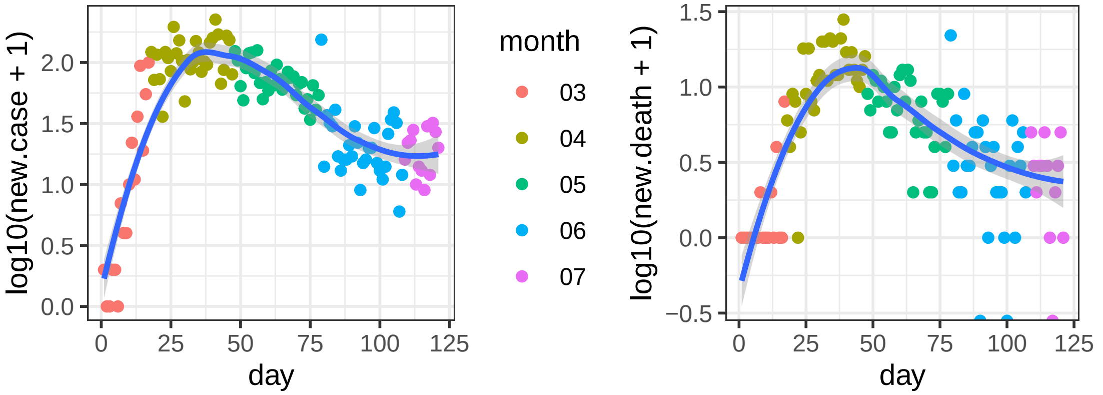
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Plymouth_Massachusetts



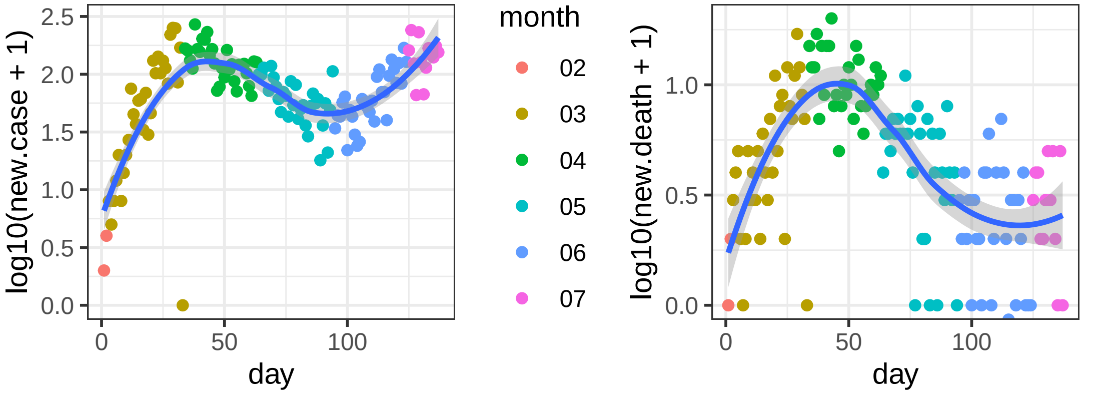
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

Hampden_Massachusetts



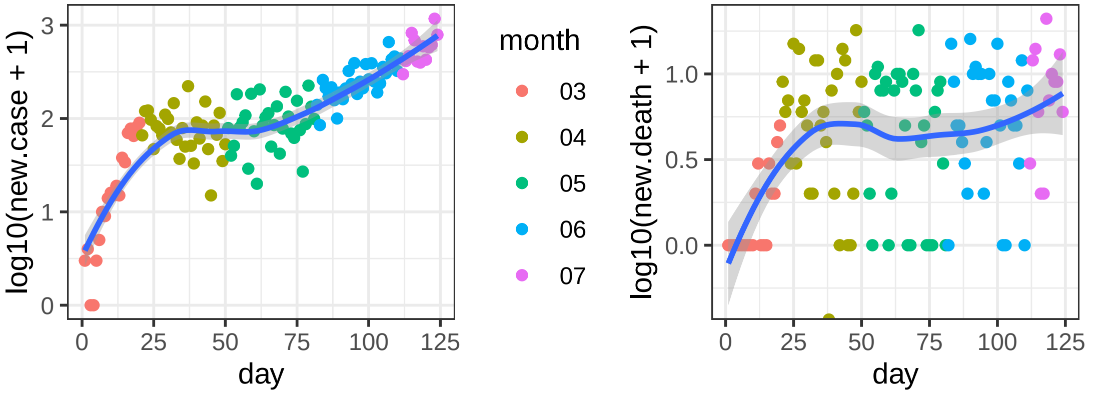
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

King_Washington



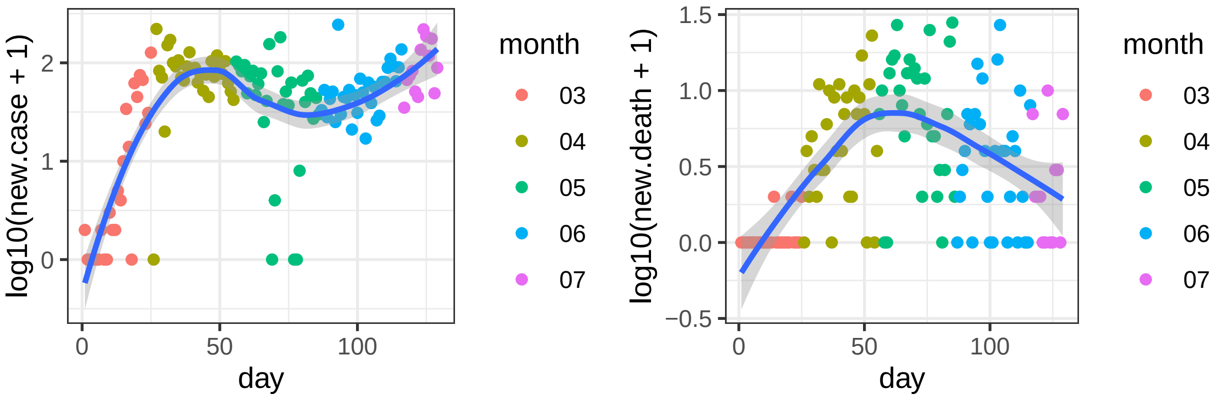
data source: https://github.com/nytimes/covid−19−data, day 1 is 02−28

Palm Beach_Florida



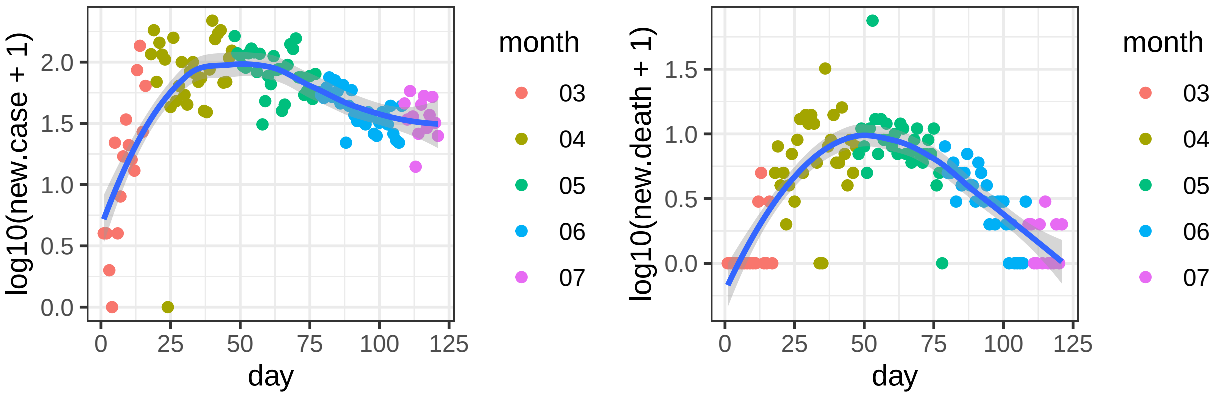
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

St. Louis_Missouri



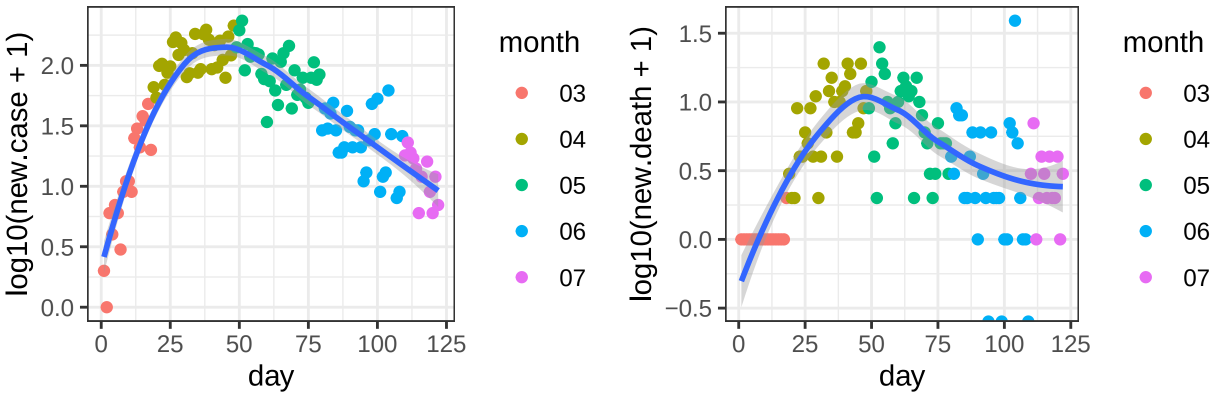
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Erie_New York



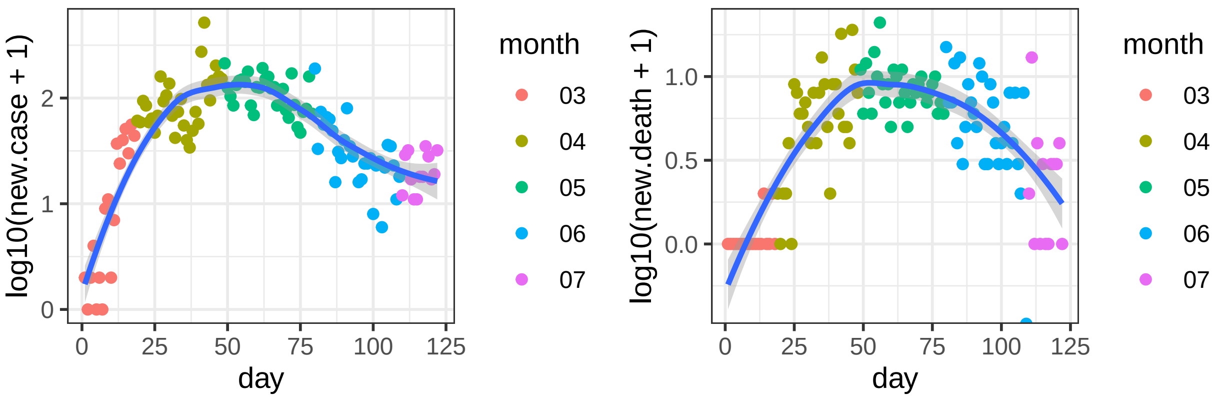
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

Mercer_New Jersey



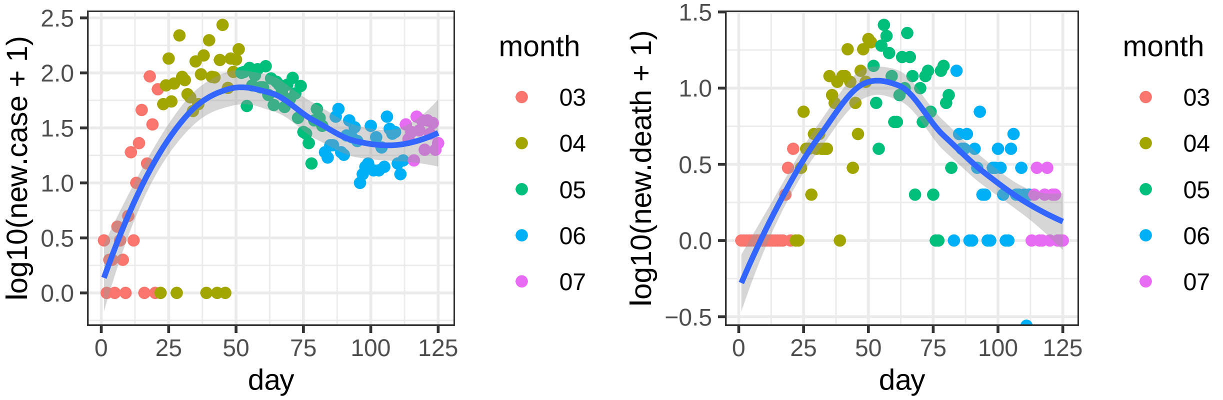
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Bristol_Massachusetts



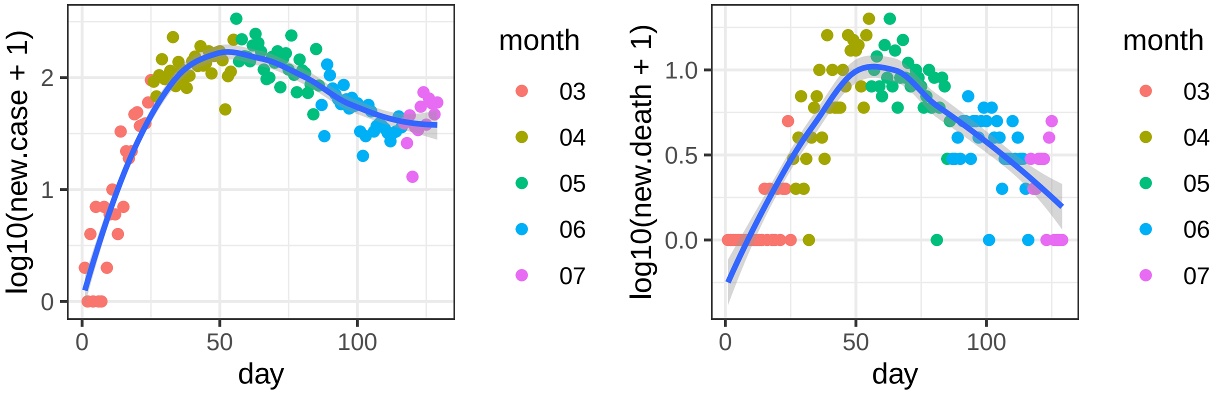
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Bucks_Pennsylvania



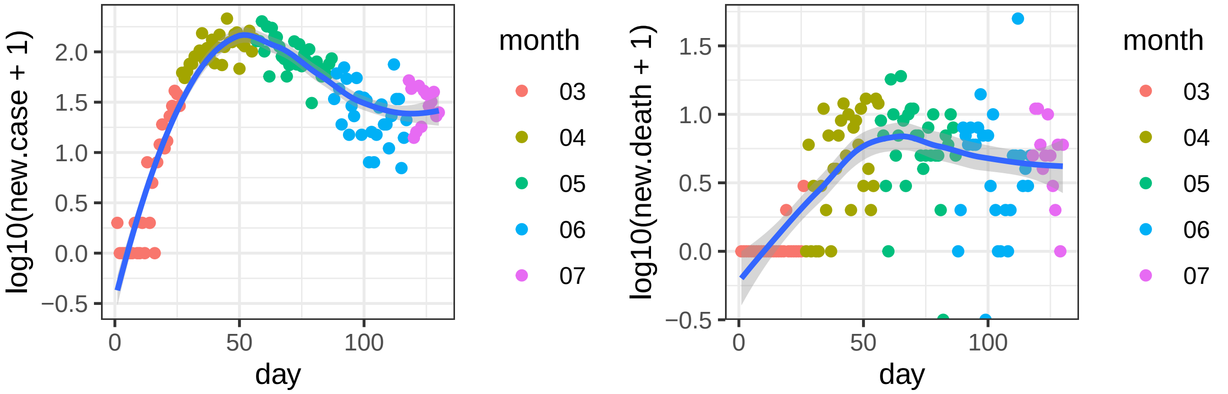
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

District of Columbia_District of Columbia



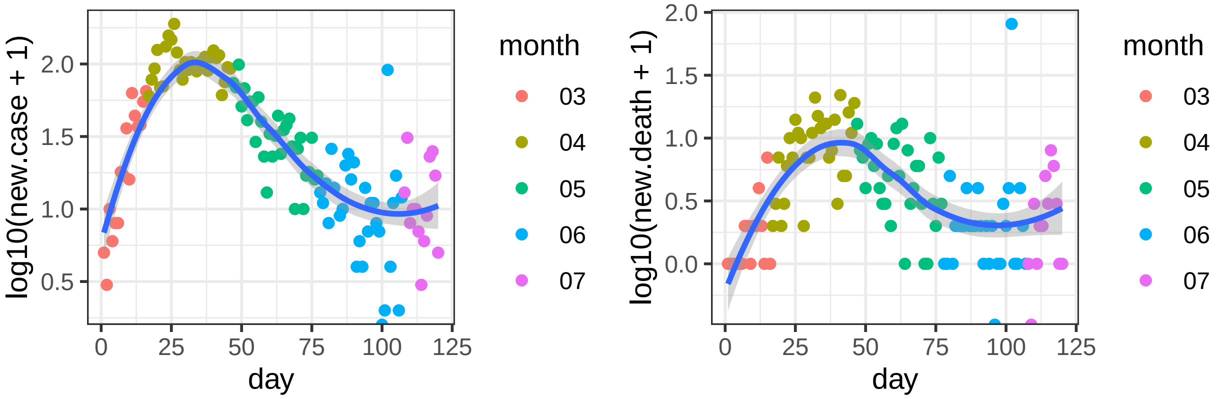
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Camden_New Jersey



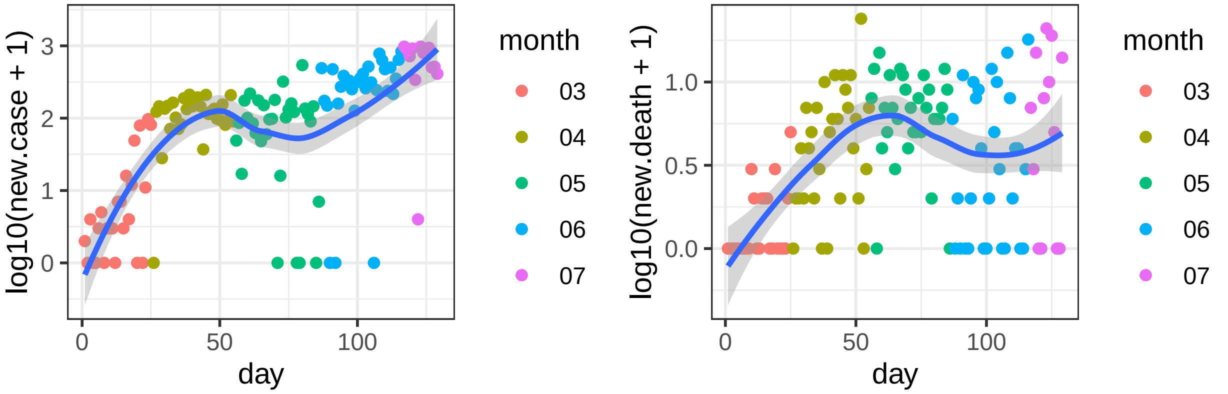
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Somerset_New Jersey

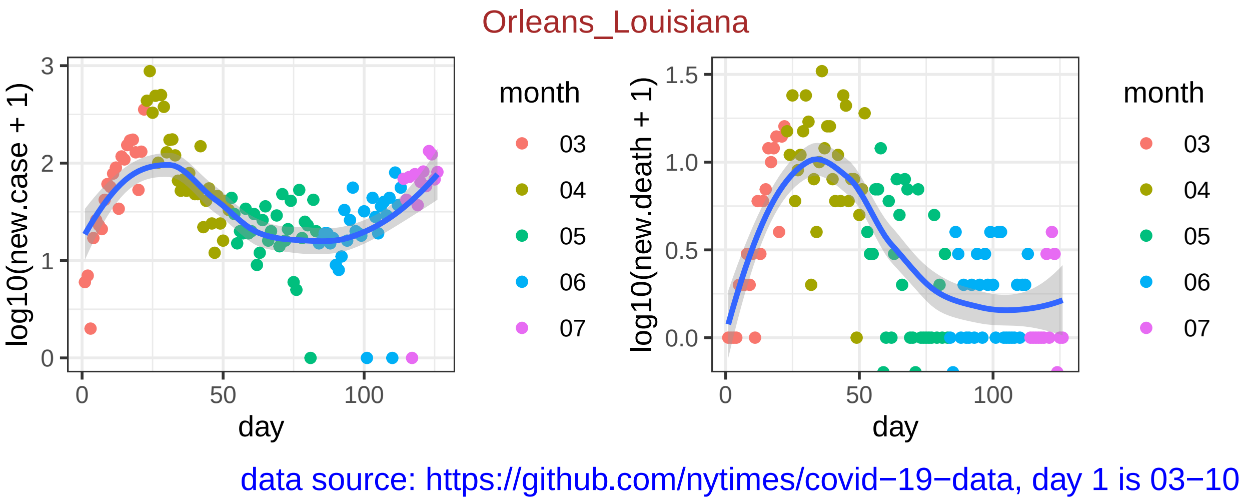


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−16

Riverside_California



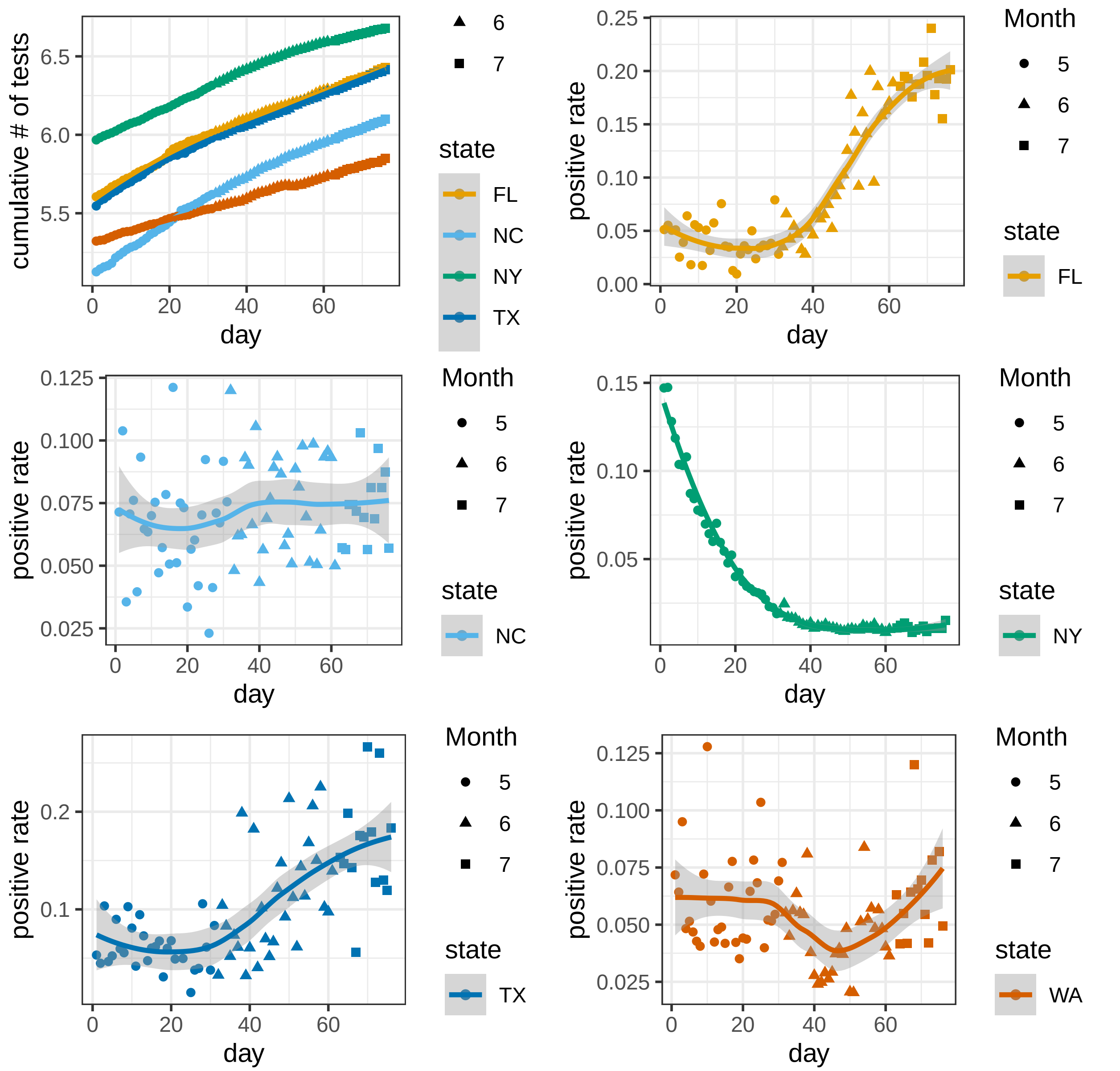
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07



COVID Trackng

The positive rates of testing can be an indicator on how much the COVID-19 has spread. However, they can be much 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.

The grades of the states may change over timea and I strongly recommend checking their webiste before puting serious interpretation on the following plot.



github.com/COVID19Tracking/, positive rate on 0714: 0.20(FL) 0.06(NC) 0.02(NY) 0.18(TX) 0.05(WA)

Session information

```
sessionInfo()
```

```
## R version 3.6.2 (2019-12-12)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Catalina 10.15.5
##
## 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_3.3.1
##
## 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    lattice_0.20-38  nlme_3.1-144     evaluate_0.14
##  [9] lifecycle_0.2.0  tibble_3.0.1     gtable_0.3.0     mgcv_1.8-31
## [13] pkgconfig_2.0.3  rlang_0.4.6      Matrix_1.2-18    yaml_2.2.1
## [17] xfun_0.12        gridExtra_2.3    withr_2.1.2      stringr_1.4.0
## [21] dplyr_0.8.4      knitr_1.28       vctrs_0.3.0      cowplot_1.0.0
## [25] grid_3.6.2       tidyselect_1.0.0 glue_1.3.1       R6_2.4.1
## [29] rmarkdown_2.1    purrr_0.3.3      farver_2.0.3     splines_3.6.2
## [33] scales_1.1.0     ellipsis_0.3.0   htmltools_0.4.0  assertthat_0.2.1
## [37] colorspace_1.4-1 ggsignif_0.6.0   labeling_0.3     stringi_1.4.5
## [41] munsell_0.5.0    crayon_1.3.4
```