Exploration of COVID-19 tracking data from multiple resources

Wei Sun

2020-09-26

Contents

Introduction 1

JHU 2

time series data 2

daily reports data 6

NY Times 7

state level data 7

county level data 18

COVID Trackng 36

Session information 39

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

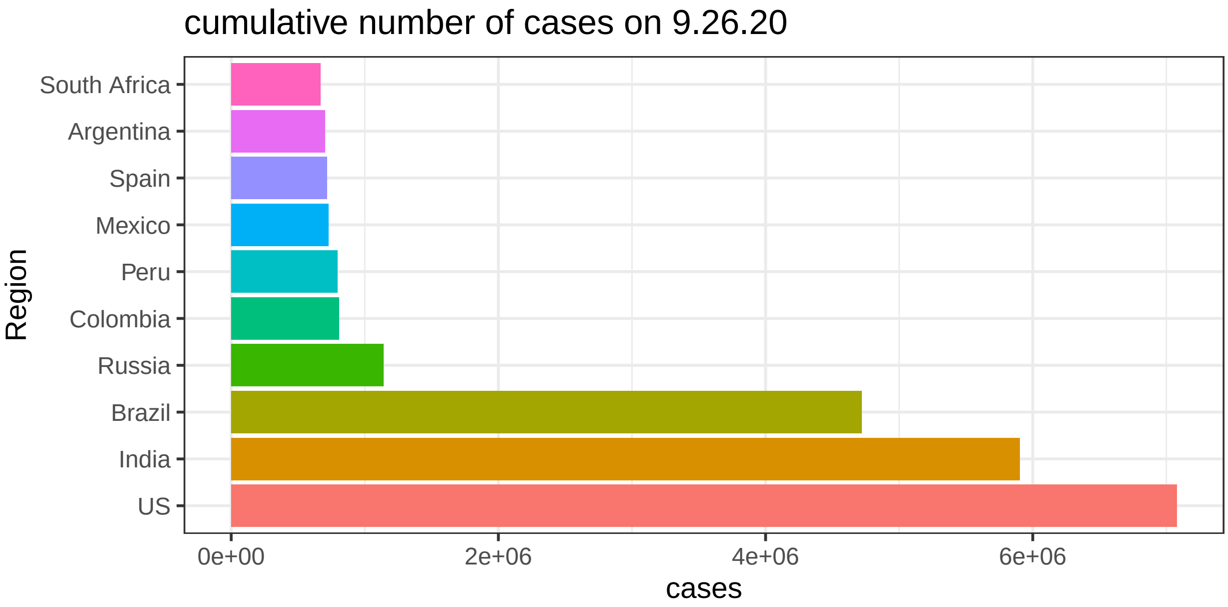
JHU

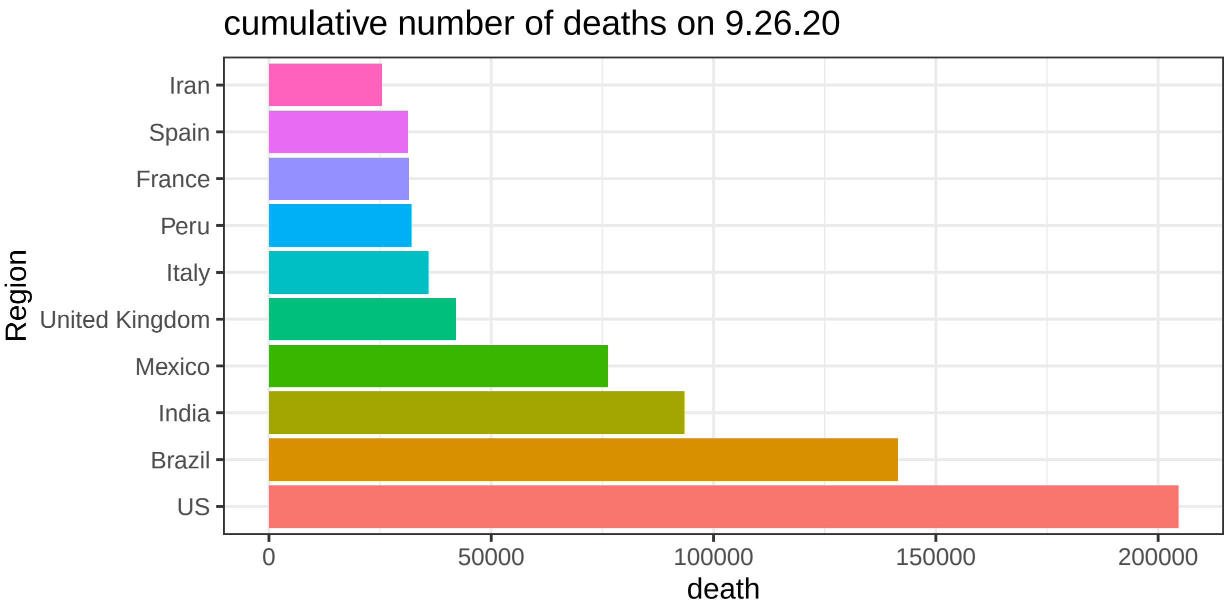
Assume you have cloned the JHU Github repository on your local machine at "../COVID-19".

time series data

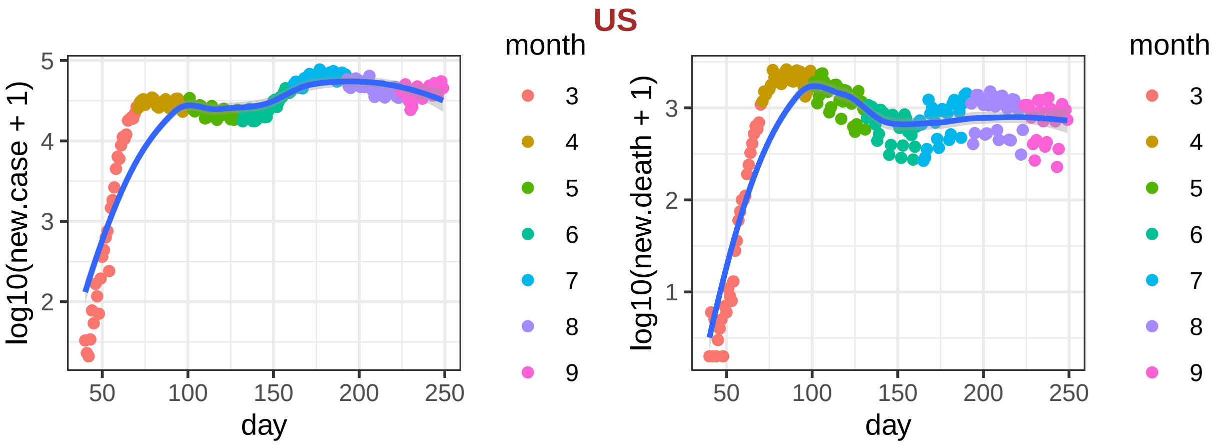
The time series provide counts (e.g., confirmed cases, deaths) starting from Jan 22nd, 2020 for 253 locations. Currently there is no data of individual US state in these time series data files.

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

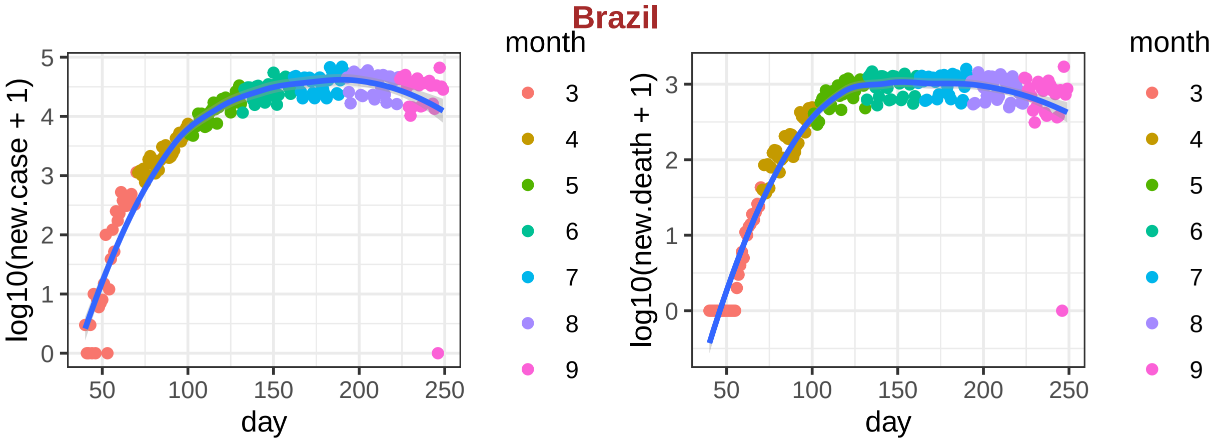




Next, I check for each country/region, what is the number of new cases/deaths? This data is important to understand what is the trend under different situations, e.g., population density, social distance policies etc. Here I checked the top 10 countries/regions with the highest number of deaths.

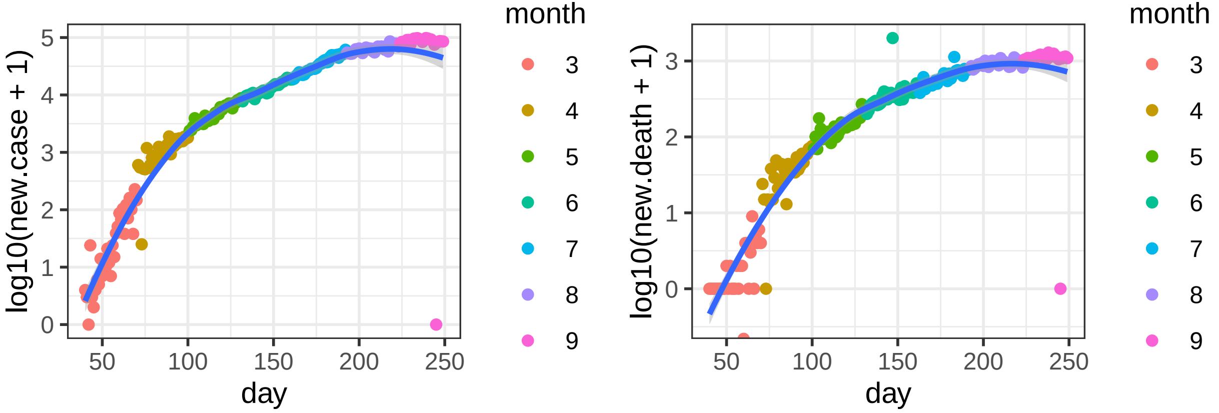


data source: https://github.com/CSSEGISandData/COVID−19



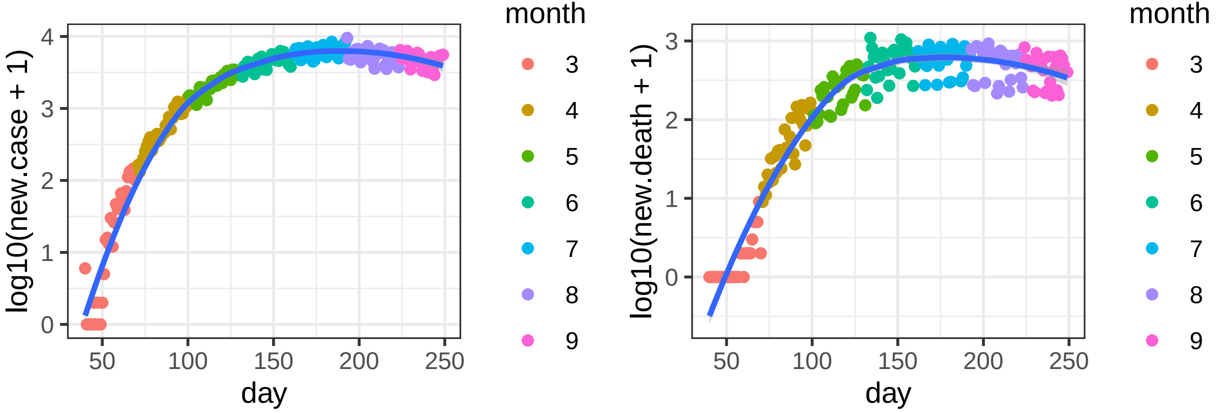
data source: https://github.com/CSSEGISandData/COVID−19

India



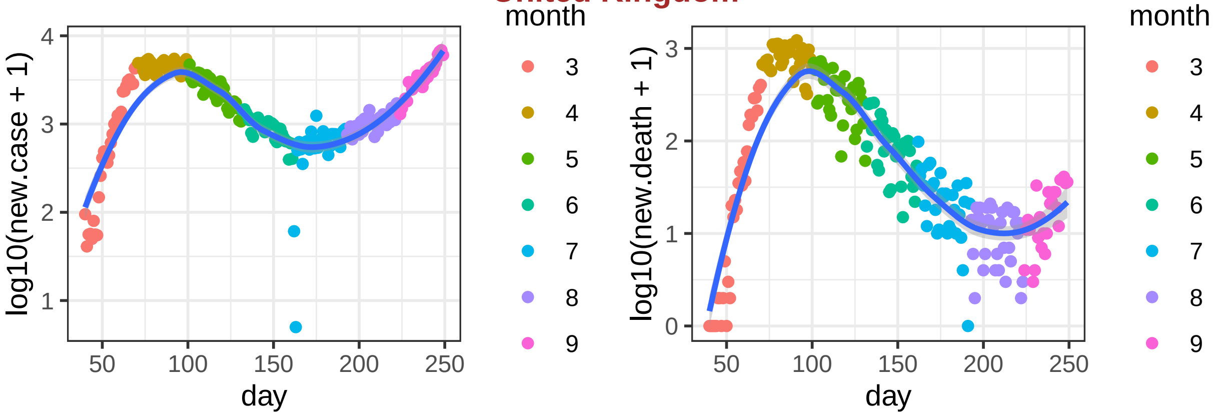
data source: https://github.com/CSSEGISandData/COVID−19

Mexico



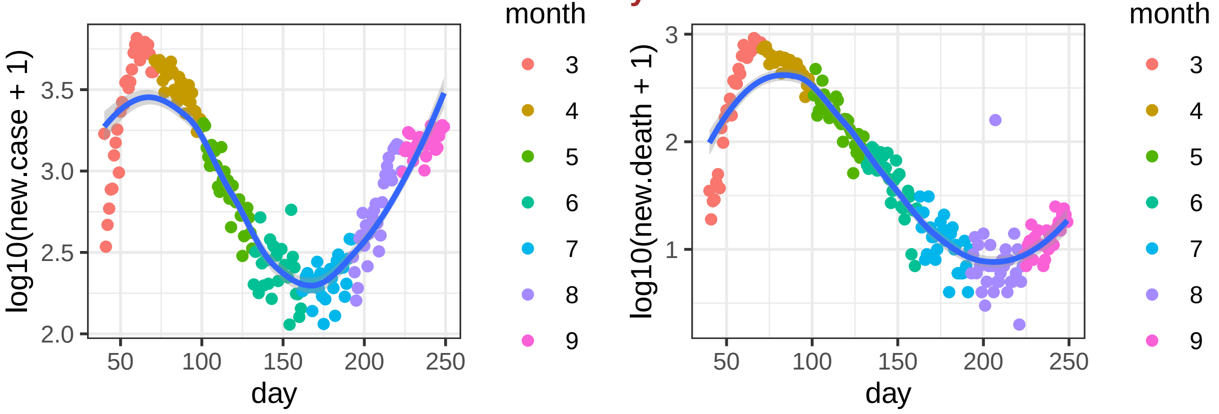
data source: https://github.com/CSSEGISandData/COVID−19

United Kingdom



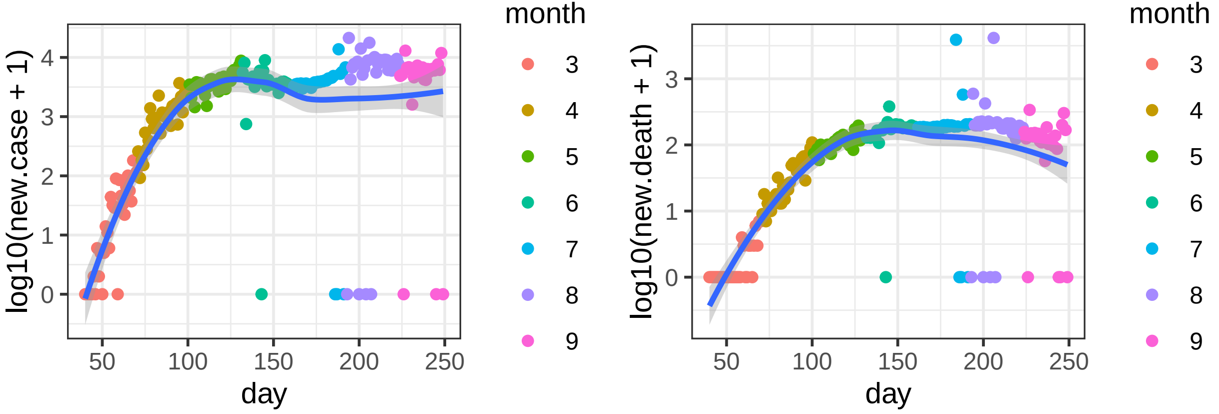
data source: https://github.com/CSSEGISandData/COVID−19

Italy



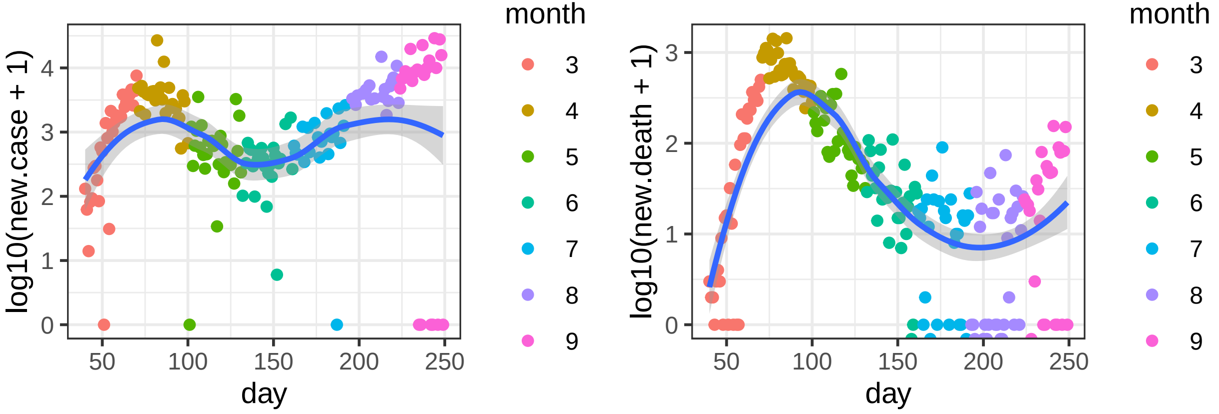
data source: https://github.com/CSSEGISandData/COVID−19

Peru

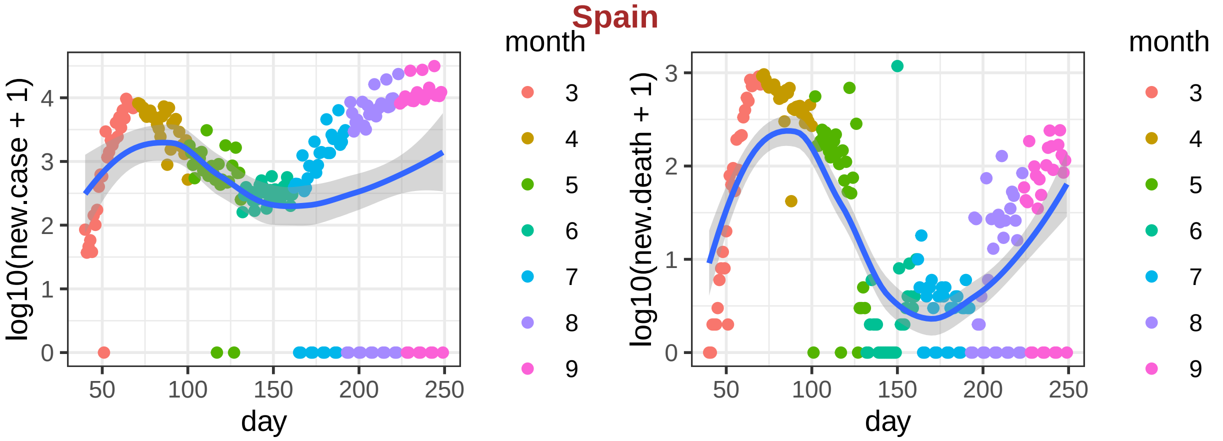


data source: https://github.com/CSSEGISandData/COVID−19

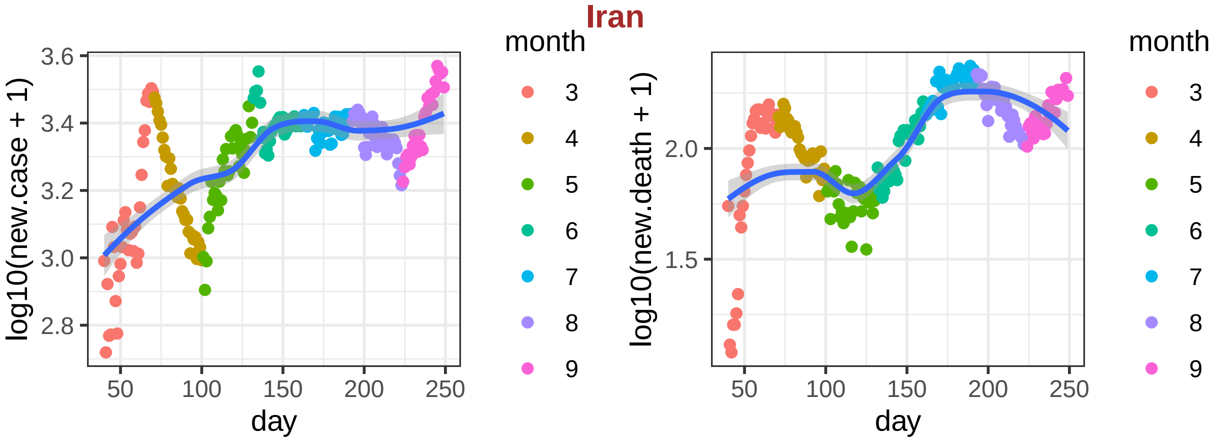
France



data source: https://github.com/CSSEGISandData/COVID−19



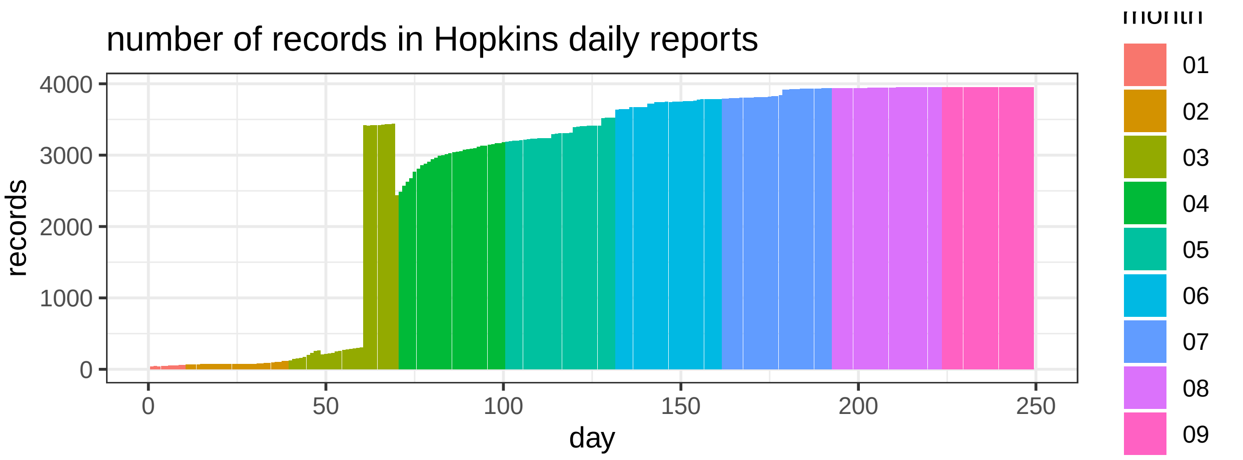
data source: https://github.com/CSSEGISandData/COVID−19



data source: https://github.com/CSSEGISandData/COVID−19

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] "2020-09-25"
```

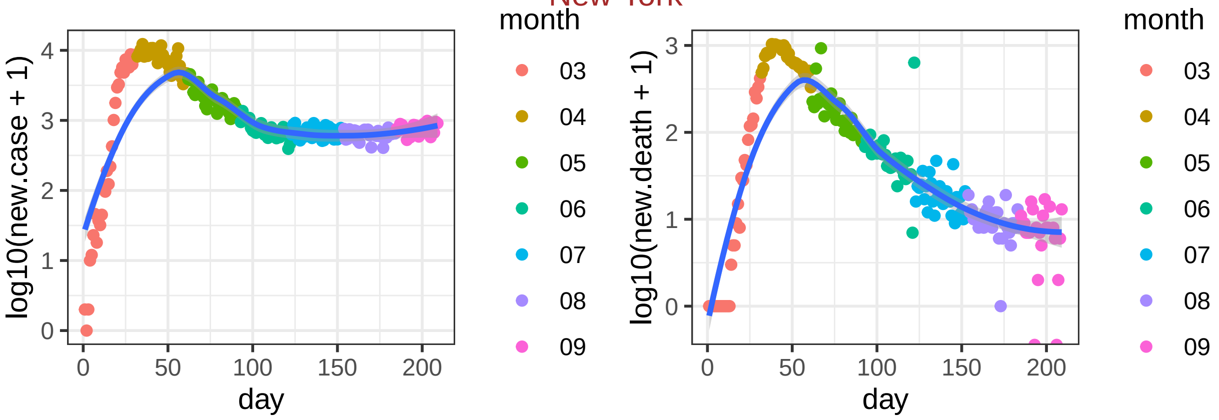
state level data

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

```
##             date          state fips  cases deaths
## 11378 2020-09-25       New York   36 458466  32708
## 11376 2020-09-25     New Jersey   34 203891  16097
## 11391 2020-09-25          Texas   48 761644  15637
## 11349 2020-09-25     California    6 805733  15533
## 11354 2020-09-25        Florida   12 695879  13914
## 11367 2020-09-25  Massachusetts   25 129481   9373
## 11359 2020-09-25       Illinois   17 287375   8826
## 11385 2020-09-25   Pennsylvania   42 159051   8157
## 11368 2020-09-25       Michigan   26 133443   7028
## 11355 2020-09-25        Georgia   13 296089   6717
## 11347 2020-09-25        Arizona    4 216367   5588
## 11364 2020-09-25      Louisiana   22 165152   5444
## 11382 2020-09-25           Ohio   39 148894   4734
## 11351 2020-09-25    Connecticut    9  56587   4501
## 11366 2020-09-25       Maryland   24 122850   3917
## 11360 2020-09-25        Indiana   18 117656   3566
## 11379 2020-09-25 North Carolina   37 204658   3437
## 11388 2020-09-25 South Carolina   45 143902   3297
## 11395 2020-09-25       Virginia   51 144433   3136
## 11370 2020-09-25    Mississippi   28  96032   2894
## 11345 2020-09-25        Alabama    1 150658   2491
## 11390 2020-09-25      Tennessee   47 186769   2326
## 11396 2020-09-25     Washington   53  89149   2193
## 11371 2020-09-25       Missouri   29 123168   2070
## 11369 2020-09-25      Minnesota   27  94241   2046
## 11350 2020-09-25       Colorado    8  68506   2045
## 11374 2020-09-25         Nevada   32  77930   1573
## 11361 2020-09-25           Iowa   19  85031   1312
## 11398 2020-09-25      Wisconsin   55 117355   1285
## 11348 2020-09-25       Arkansas    5  79946   1266
```

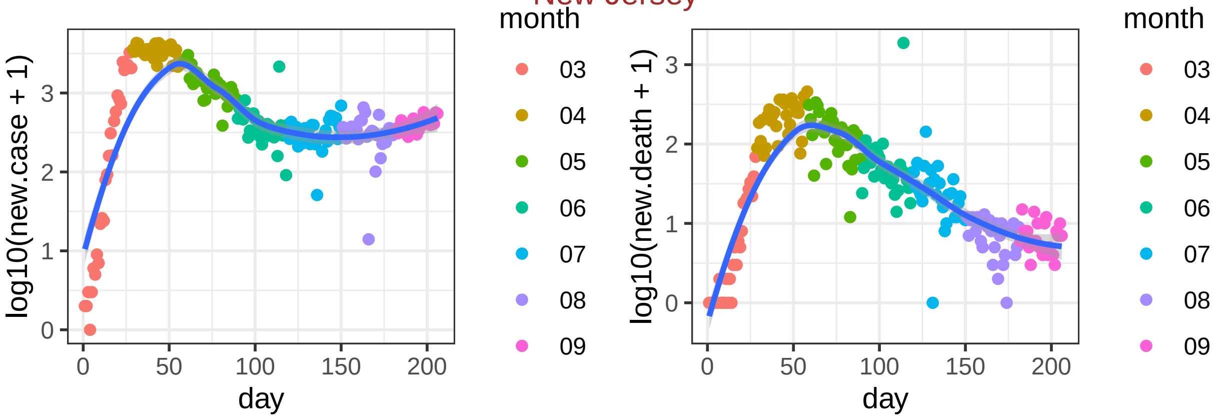
For these 30 states, I check the number of new cases and the number of new deaths. Part of the reason for such checking is to identify whether there is any similarity on such patterns. For example, could you use the pattern seen from Italy to predict what happen in an individual state, and what are the similarities and differences across states.

New York



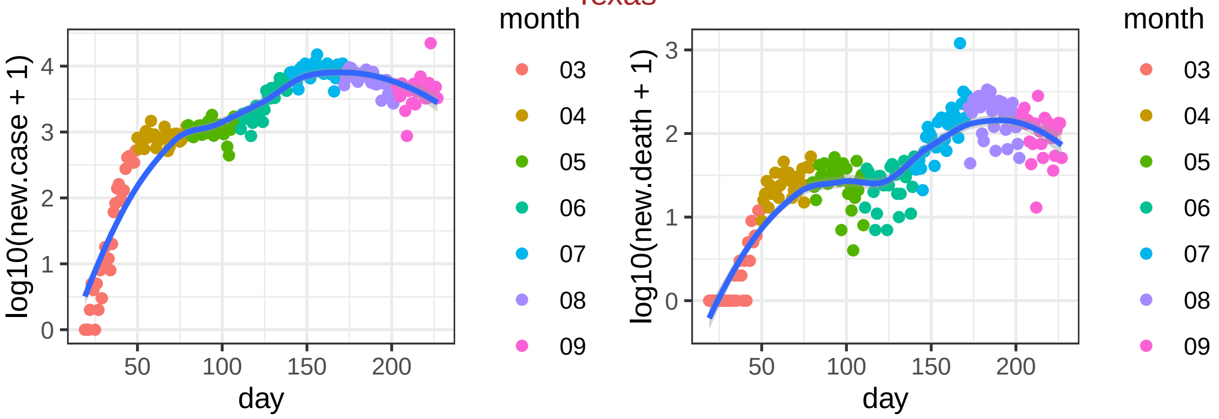
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

New Jersey

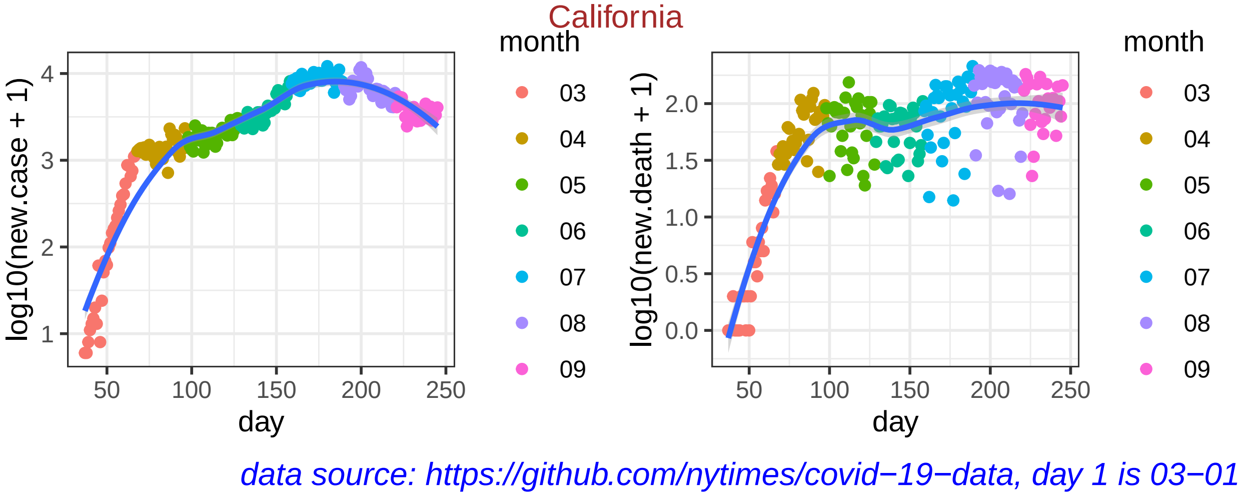


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Texas

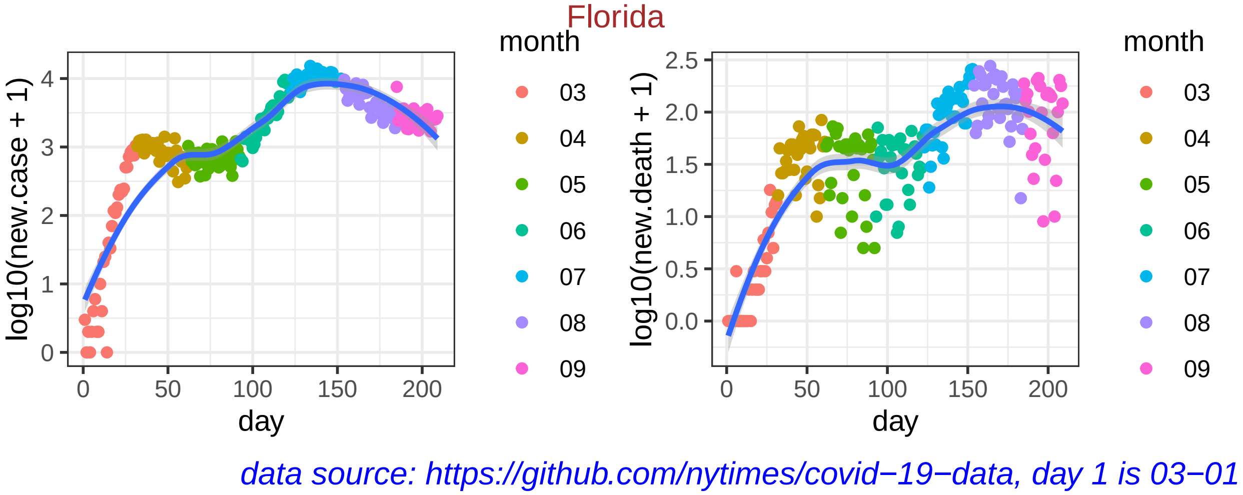


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



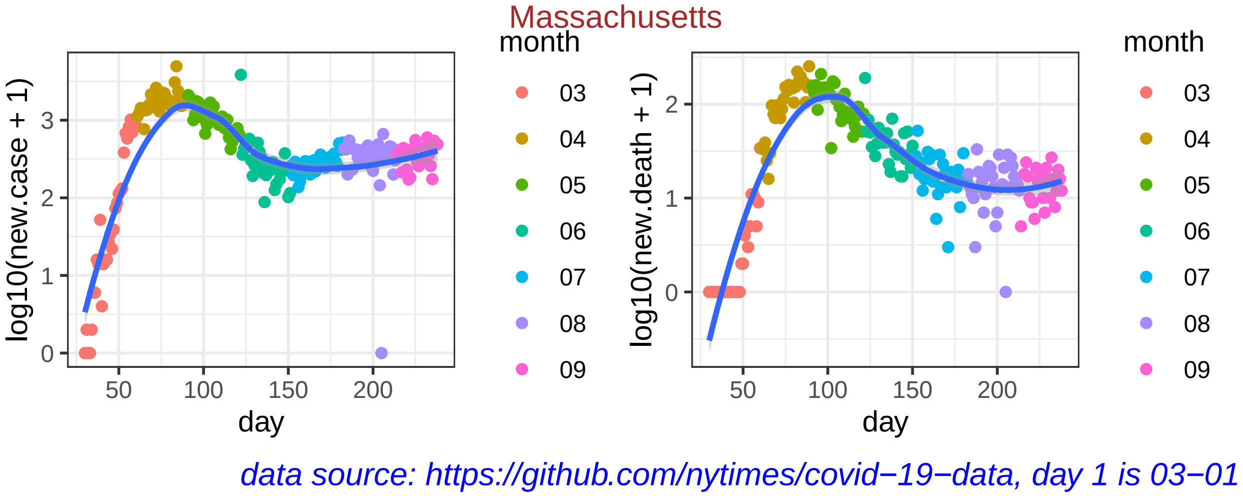
California

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



Florida

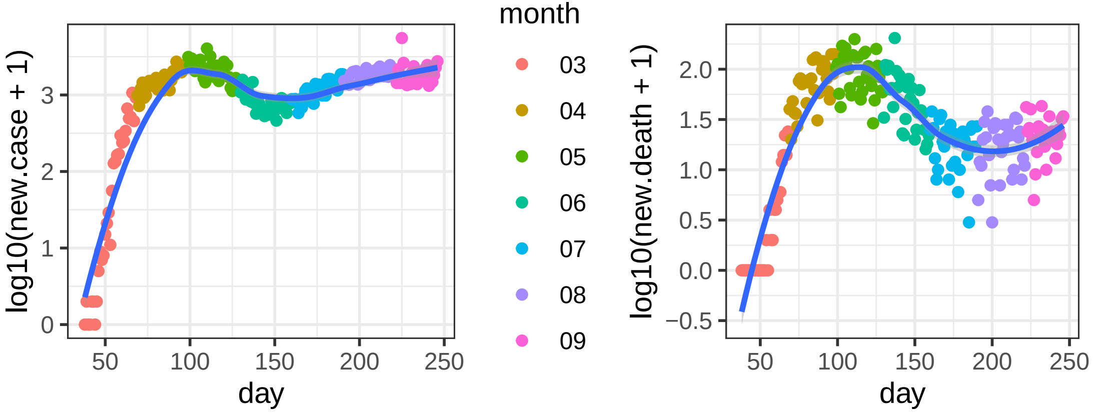
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



Massachusetts

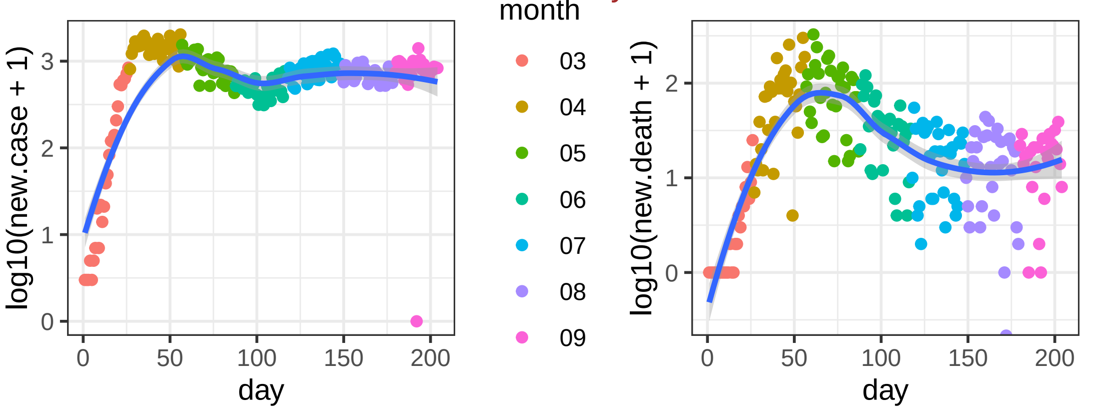
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Illinois



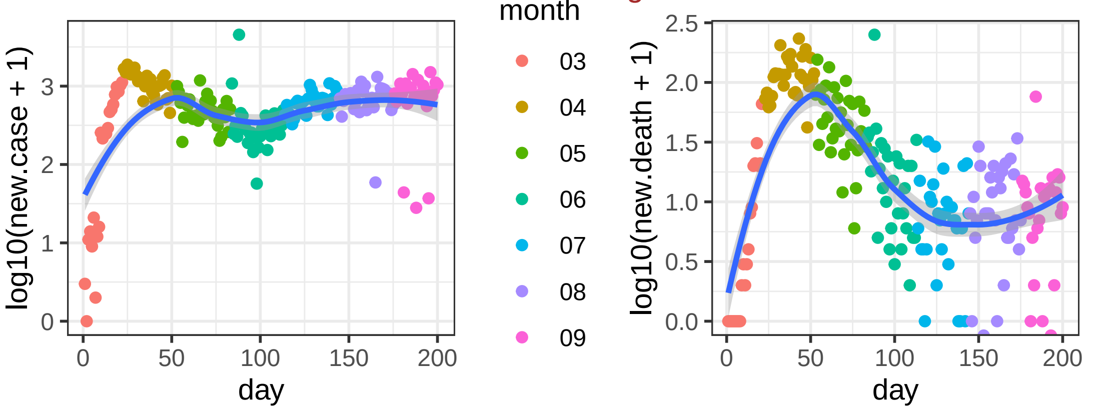
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Pennsylvania

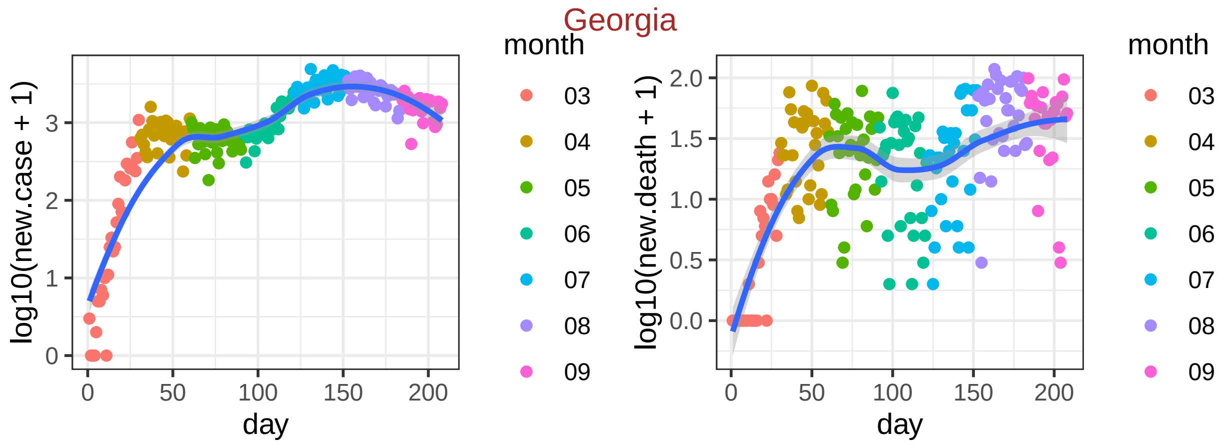


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

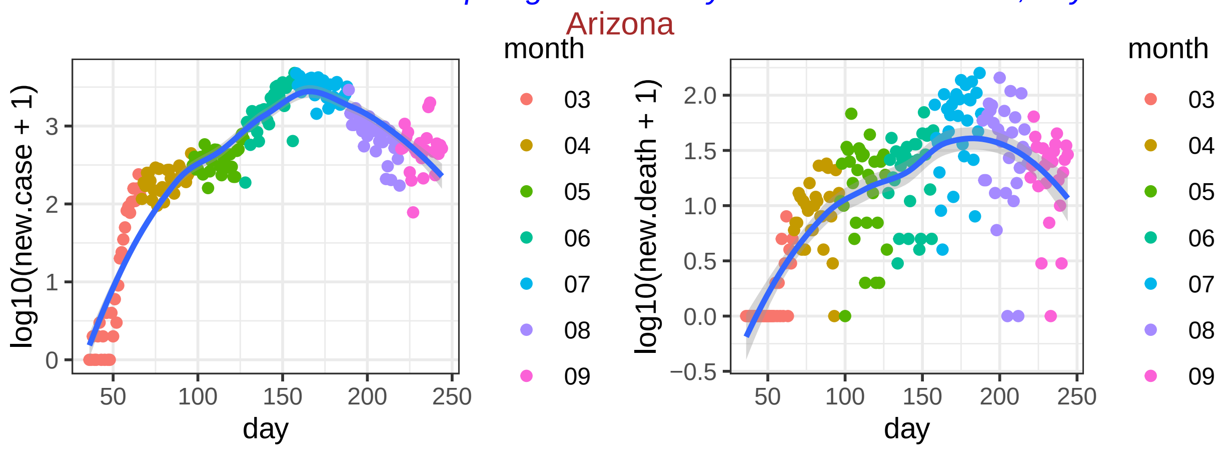
Michigan



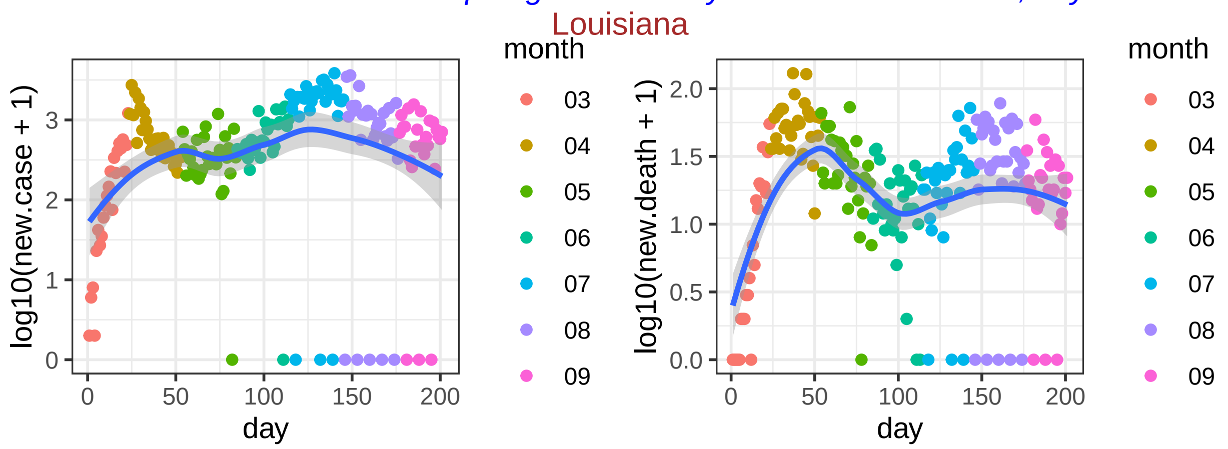
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10



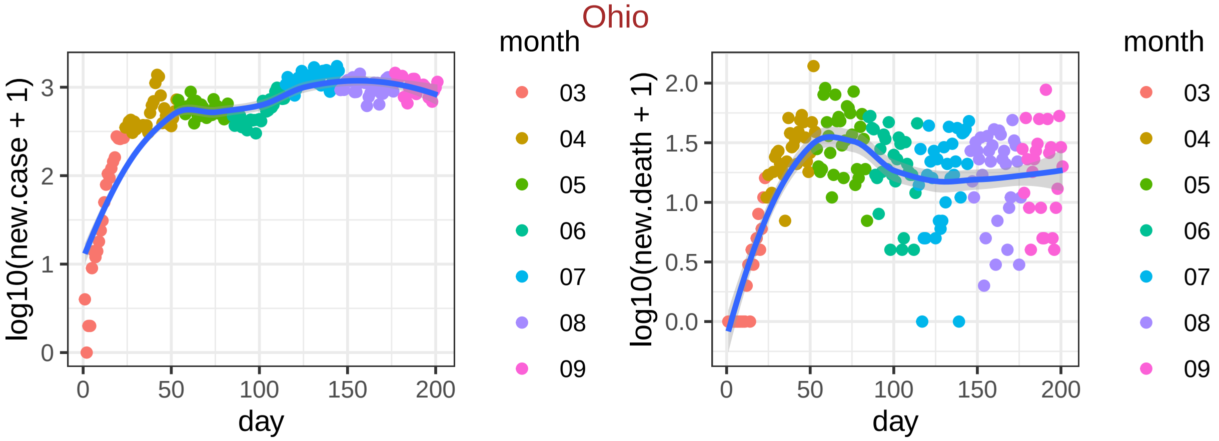
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02



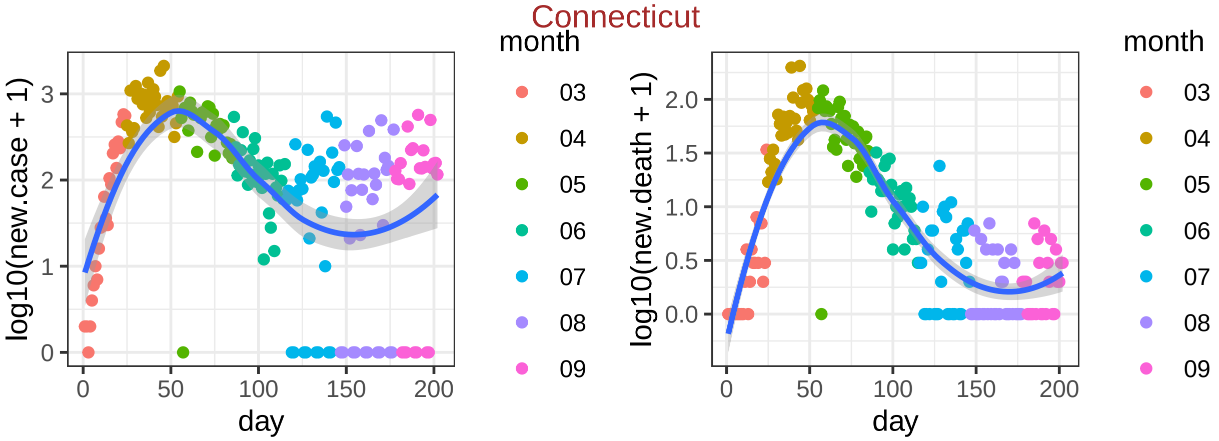
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



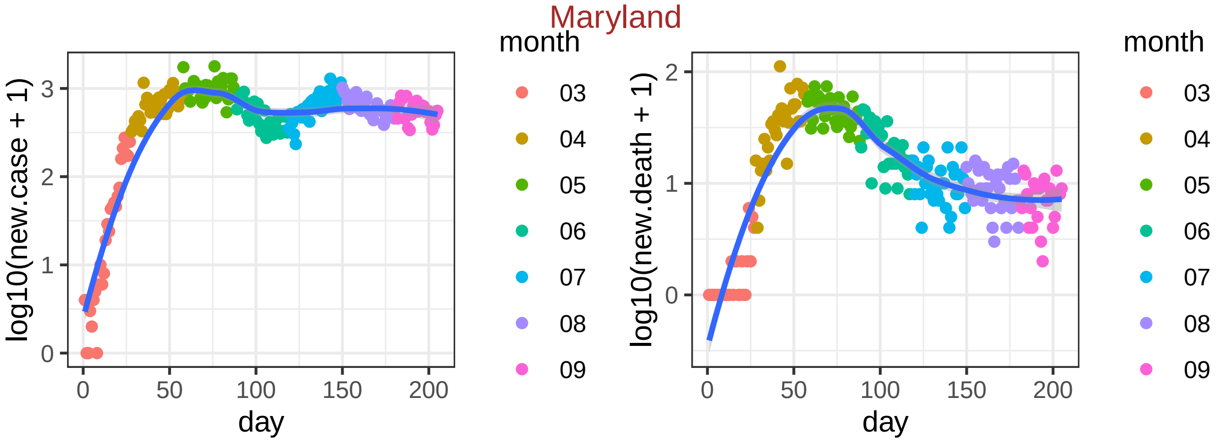
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09


Ohio

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

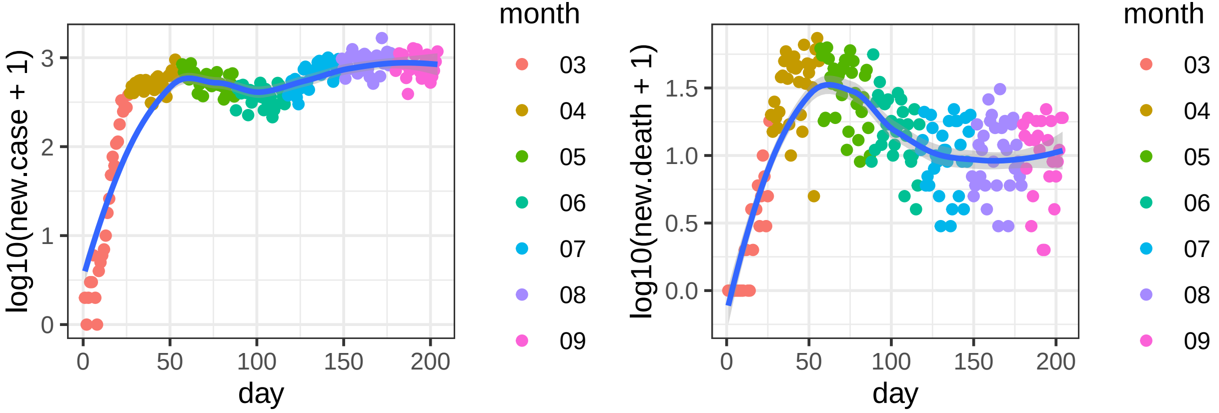

Connecticut

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


Maryland

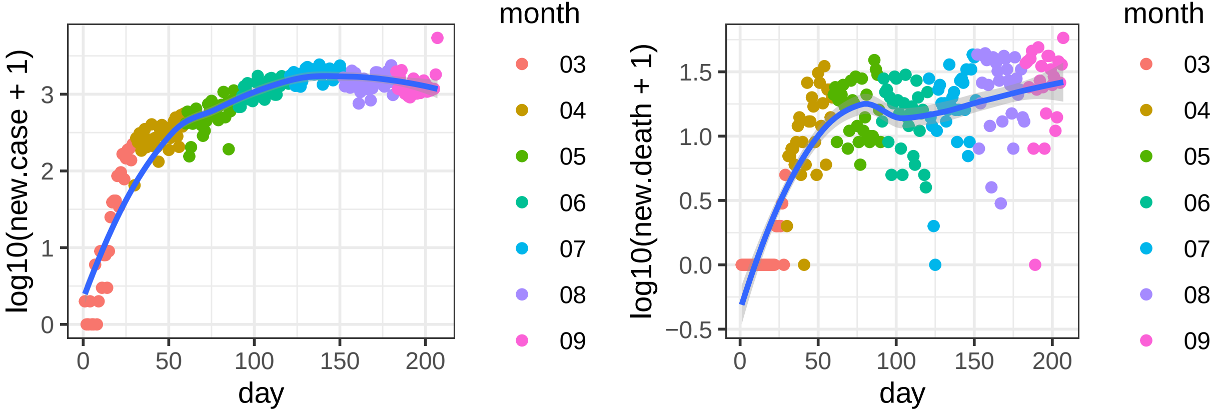
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Indiana



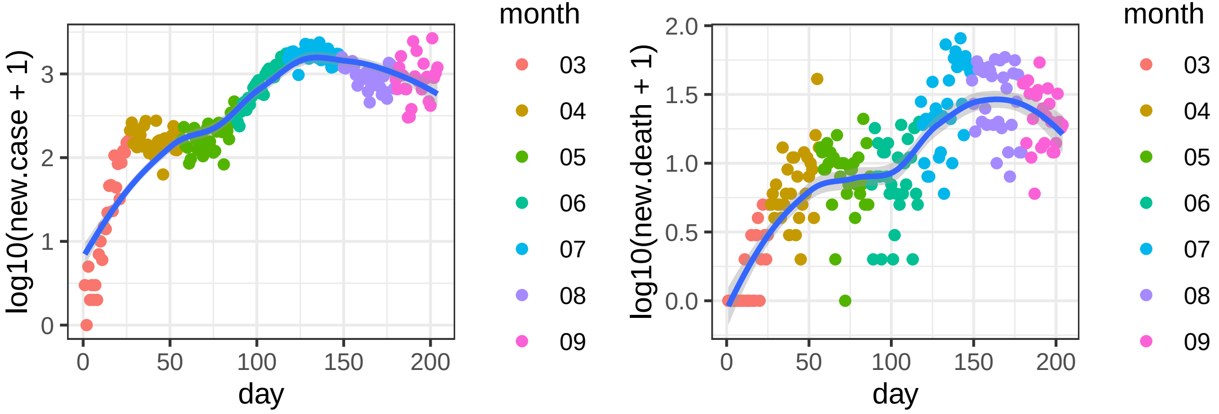
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

North Carolina



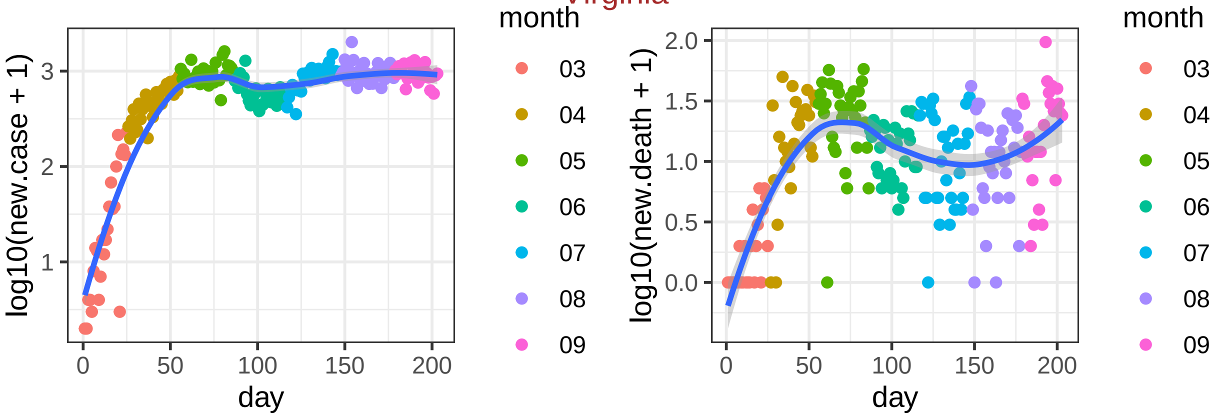
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−03

South Carolina



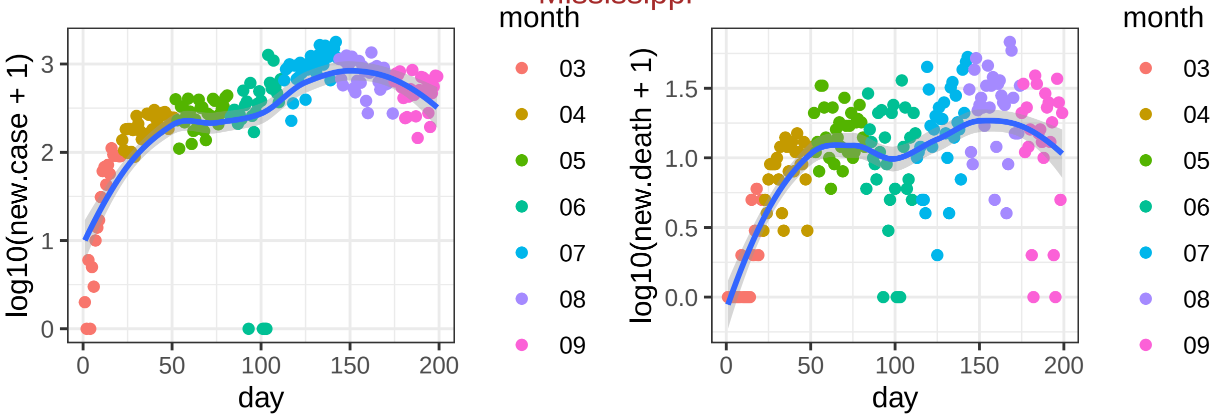
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Virginia



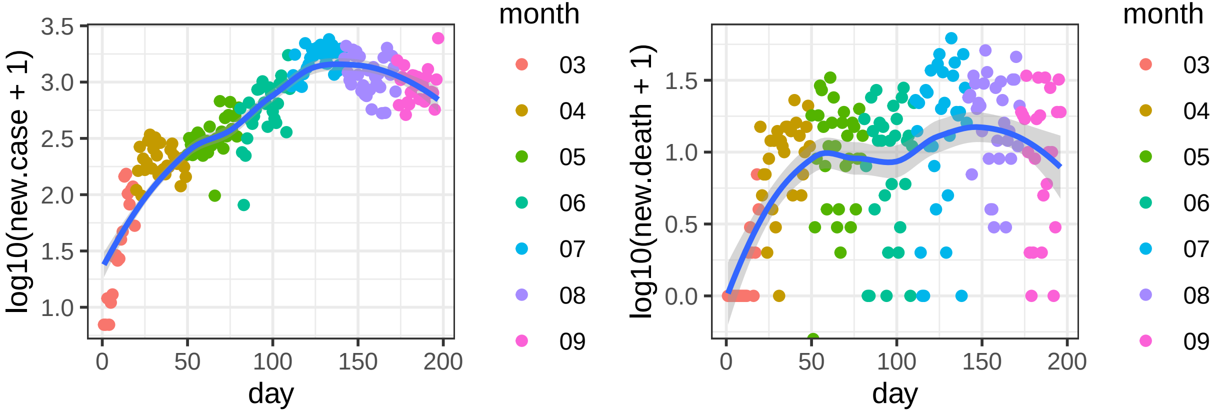
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Mississippi



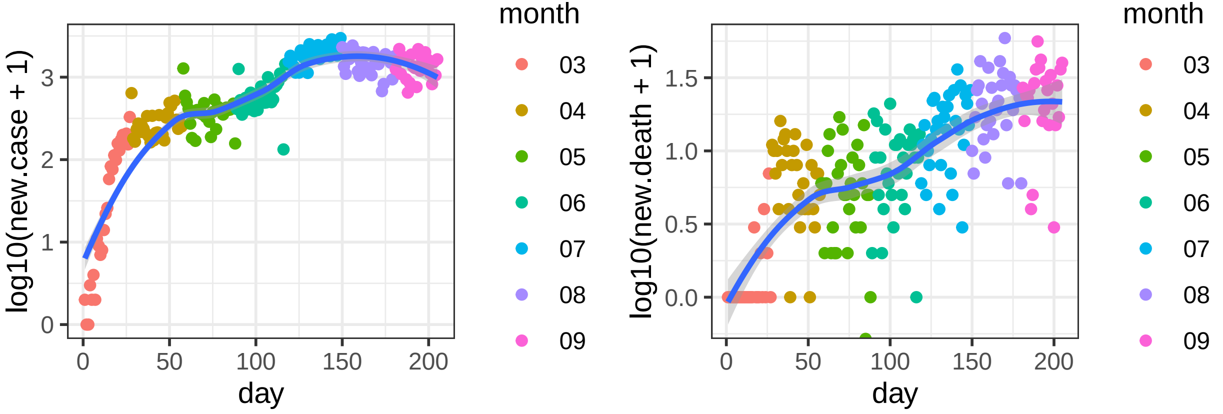
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Alabama



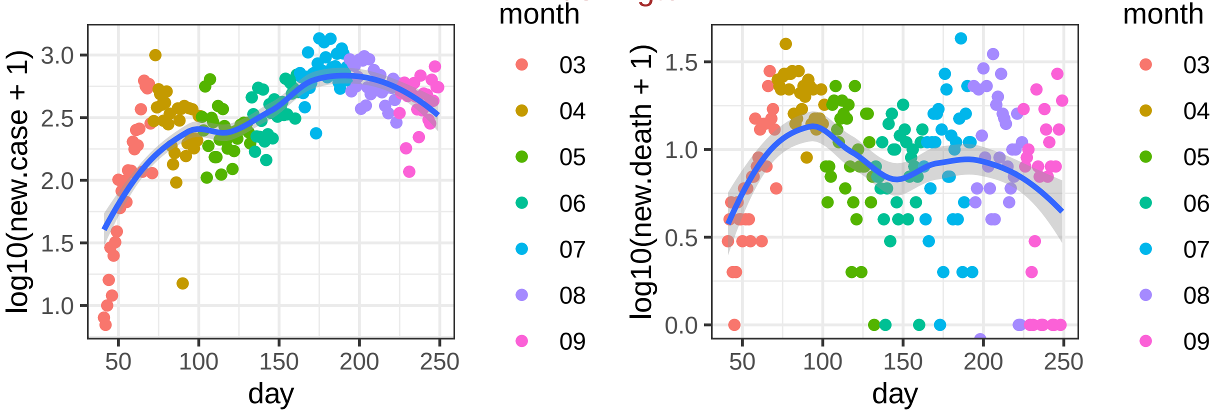
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Tennessee



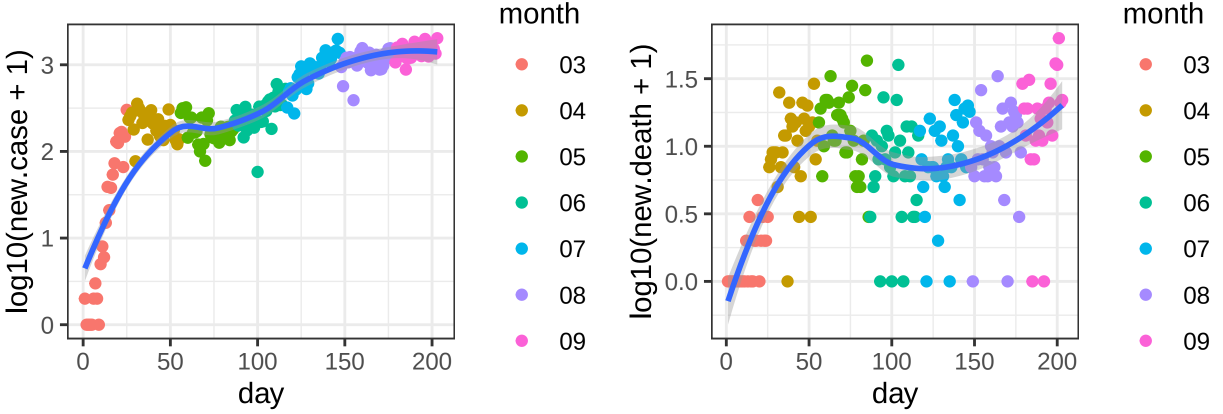
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Washington

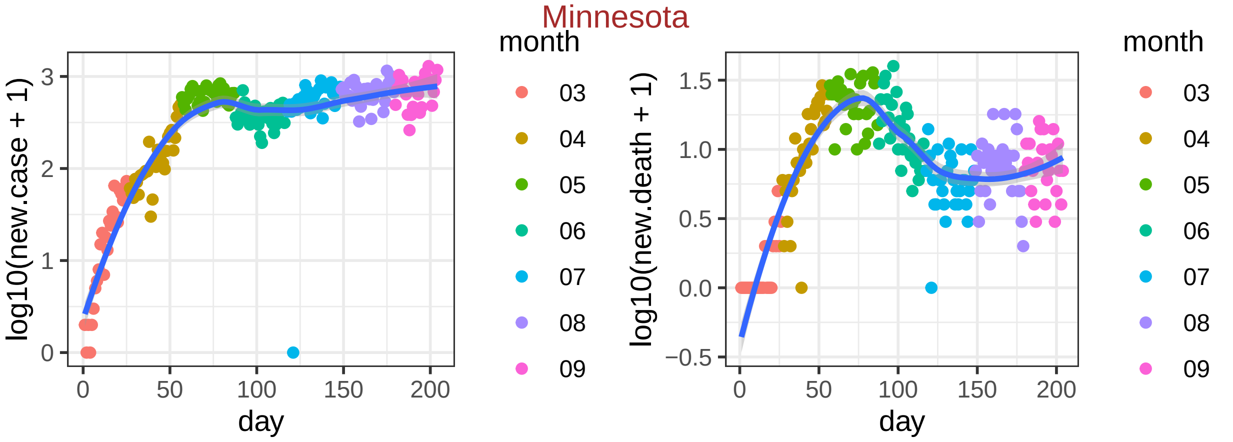


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

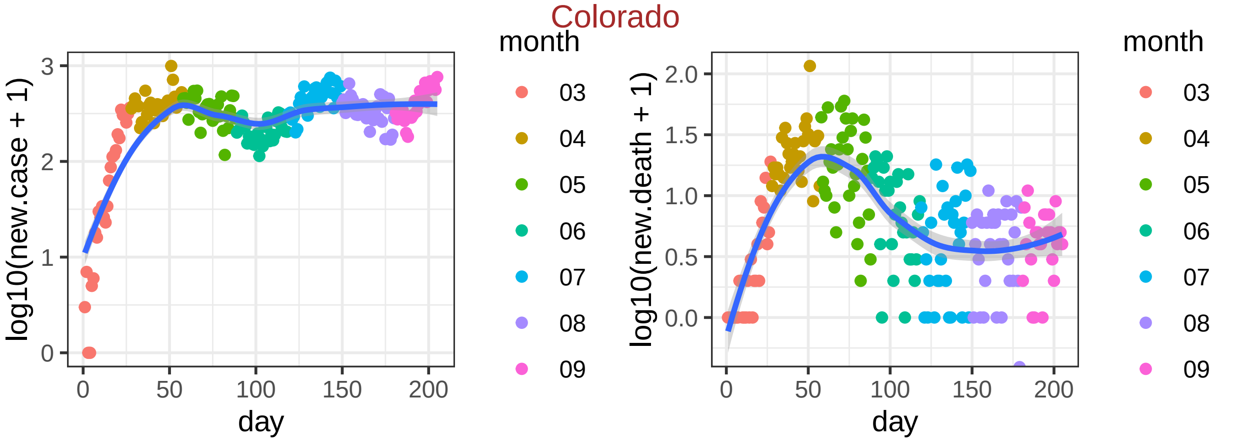
Missouri



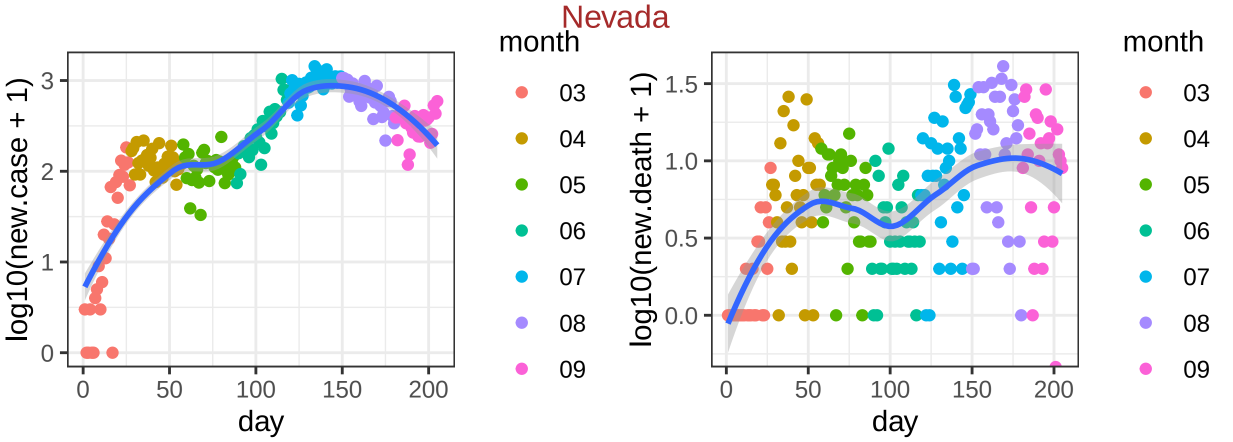
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07



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

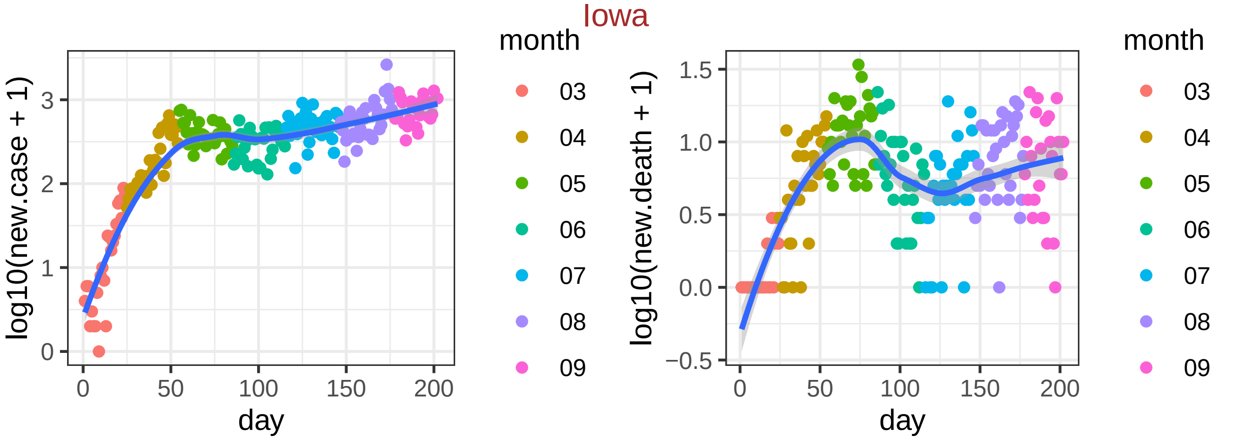


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05



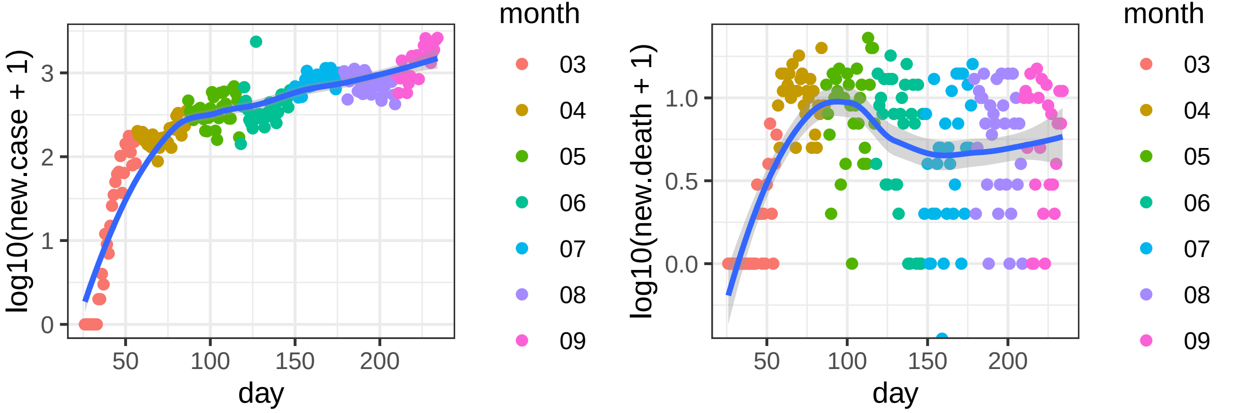
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Iowa



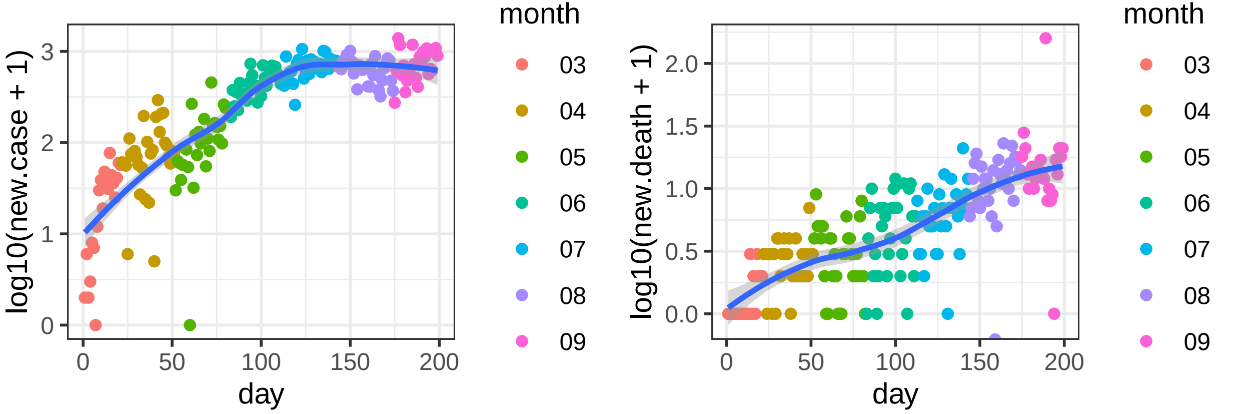
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Wisconsin



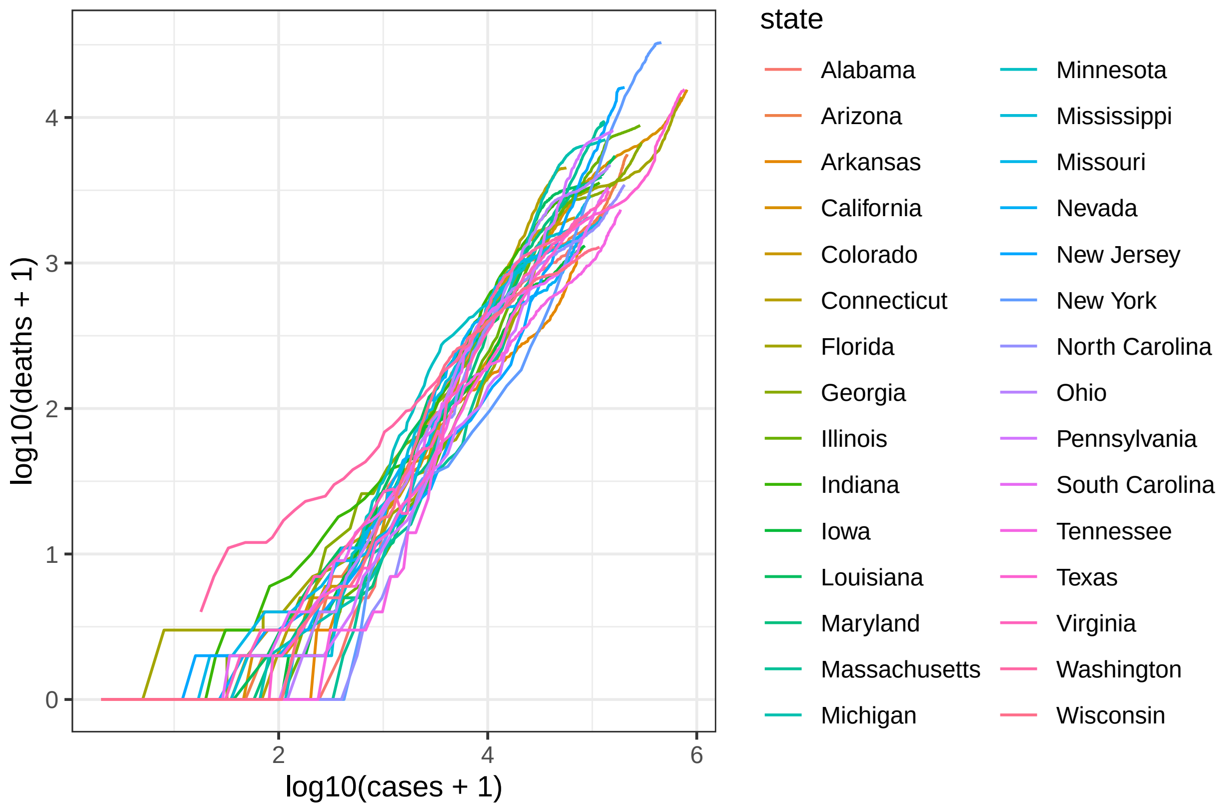
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Arkansas



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

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

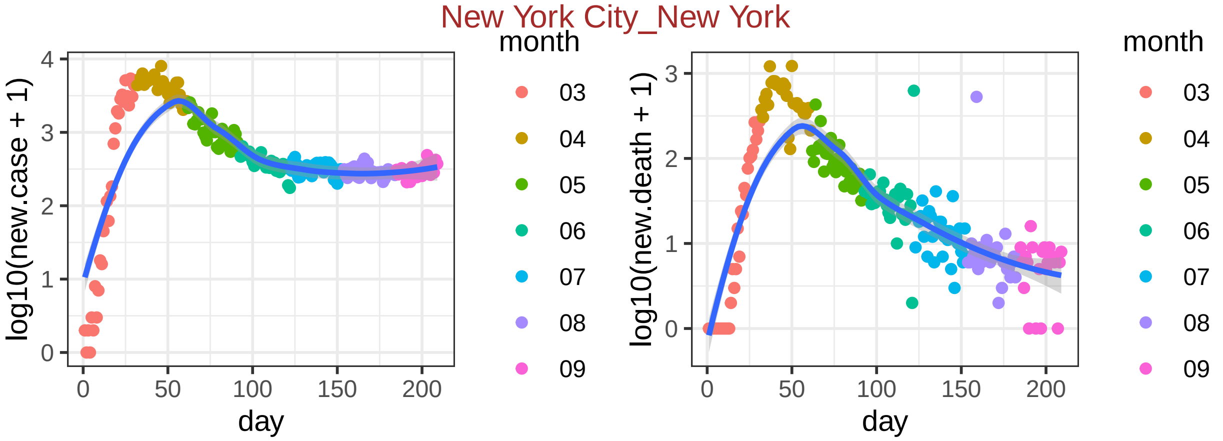
county level data

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

```
##              date        county         state  fips  cases deaths
## 568725 2020-09-25 New York City      New York    NA 246570  23792
## 567059 2020-09-25   Los Angeles    California  6037 265775   6488
## 567469 2020-09-25          Cook      Illinois 17031 142215   5194
## 566957 2020-09-25      Maricopa       Arizona  4013 140753   3343
## 567219 2020-09-25    Miami-Dade       Florida 12086 168774   3202
## 568178 2020-09-25         Wayne      Michigan 26163  34928   2975
## 569573 2020-09-25        Harris         Texas 48201 140532   2548
## 568724 2020-09-25        Nassau      New York 36059  46505   2201
## 568089 2020-09-25     Middlesex Massachusetts 25017  27103   2139
## 568648 2020-09-25         Essex    New Jersey 34013  21184   2129
## 568643 2020-09-25        Bergen    New Jersey 34003  22574   2044
## 568744 2020-09-25       Suffolk      New York 36103  46293   2013
## 569163 2020-09-25  Philadelphia  Pennsylvania 42101  36187   1818
## 569580 2020-09-25       Hidalgo         Texas 48215  31502   1630
## 568650 2020-09-25        Hudson    New Jersey 34017  20756   1514
## 568752 2020-09-25   Westchester      New York 36119  38000   1456
## 567164 2020-09-25      Hartford   Connecticut  9003  14539   1434
## 568653 2020-09-25     Middlesex    New Jersey 34023  19521   1425
## 567163 2020-09-25     Fairfield   Connecticut  9001  20015   1422
## 568617 2020-09-25         Clark        Nevada 32003  65583   1368
## 567182 2020-09-25       Broward       Florida 12011  76520   1364
## 568661 2020-09-25         Union    New Jersey 34039  17731   1355
## 567226 2020-09-25      Palm Beach       Florida 12099  46021    1337
## 568085 2020-09-25           Essex Massachusetts 25009  19525    1275
## 569487 2020-09-25           Bexar         Texas 48029  54207    1271
## 568657 2020-09-25         Passaic    New Jersey 34031  19093    1253
## 567070 2020-09-25          Orange    California  6059  54328    1204
## 568158 2020-09-25         Oakland      Michigan 26125  20442    1196
## 567073 2020-09-25       Riverside    California  6065  58178    1189
## 568093 2020-09-25         Suffolk Massachusetts 25025  24089    1133
## 567167 2020-09-25       New Haven   Connecticut  9009  14340    1115
## 569529 2020-09-25          Dallas         Texas 48113  83304    1115
## 568095 2020-09-25       Worcester Massachusetts 25027  14252    1097
## 568091 2020-09-25         Norfolk Massachusetts 25021  10170    1054
## 568656 2020-09-25           Ocean    New Jersey 34029  12728    1047
## 568145 2020-09-25          Macomb      Michigan 26099  14644    1025
## 568206 2020-09-25        Hennepin     Minnesota 27053  26706     928
## 567076 2020-09-25  San Bernardino    California  6071  53669     922
## 569503 2020-09-25         Cameron         Texas 48061  22698     913
## 569262 2020-09-25      Providence  Rhode Island 44007  18437     882
## 569158 2020-09-25      Montgomery  Pennsylvania 42091  12084     878
## 568654 2020-09-25        Monmouth    New Jersey 34025  11818     867
## 568071 2020-09-25      Montgomery      Maryland 24031  22225     842
## 568655 2020-09-25          Morris    New Jersey 34027   7966     831
## 568072 2020-09-25 Prince George's      Maryland 24033  29365     825
## 567605 2020-09-25          Marion       Indiana 18097  21367     817
## 568452 2020-09-25       St. Louis      Missouri 29189  23688     796
## 569135 2020-09-25        Delaware  Pennsylvania 42045  11348     793
## 569922 2020-09-25            King    Washington 53033  21915     787
## 567077 2020-09-25       San Diego    California  6073  46064     775
```

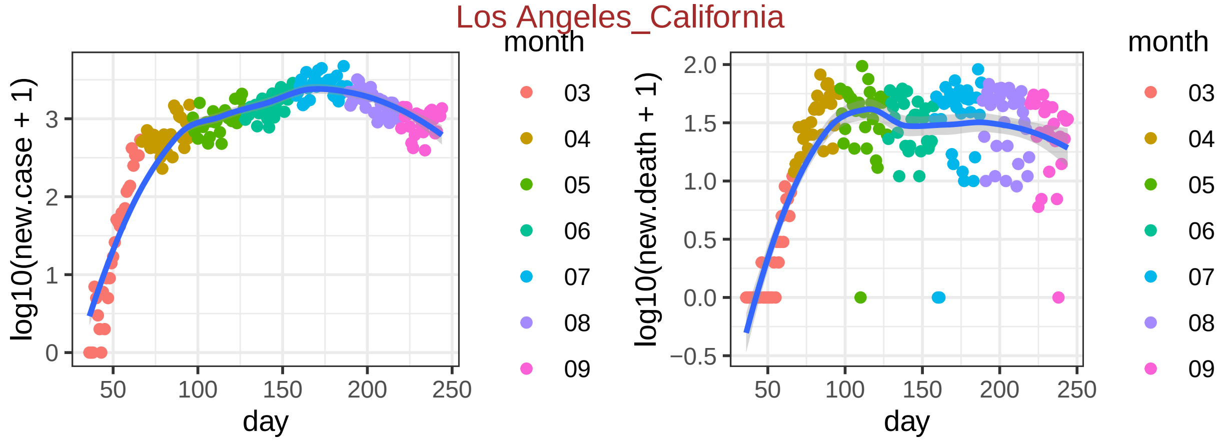
For these 50 counties, I check the number of new cases and the number of new deaths.

New York City_New York



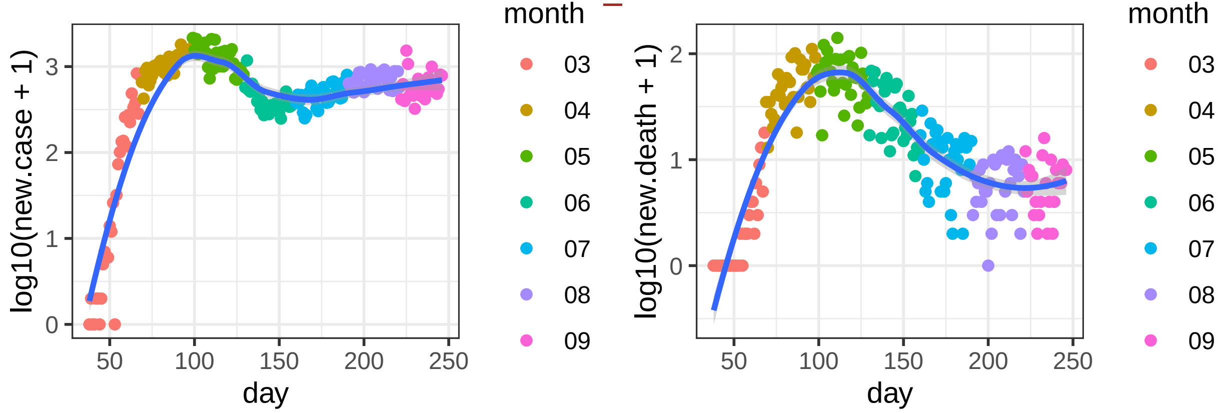
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Los Angeles_California



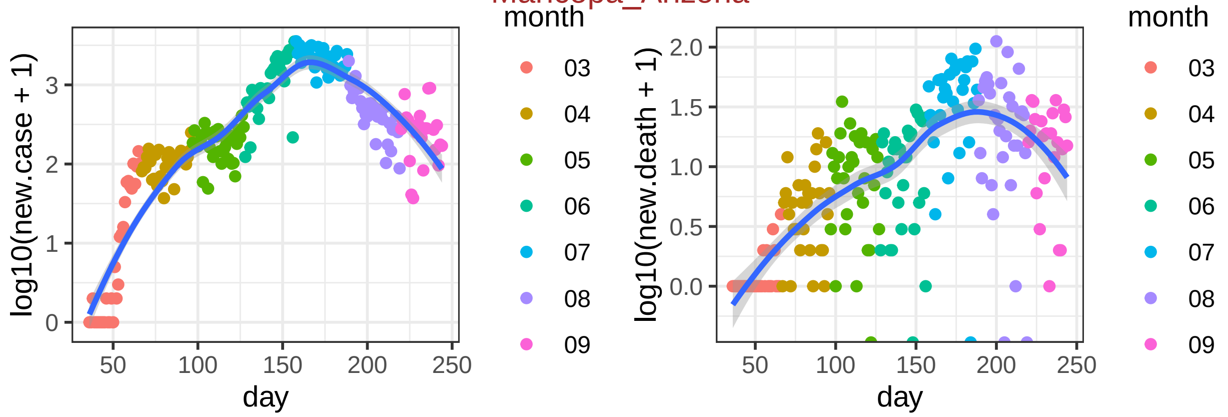
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Cook_Illinois



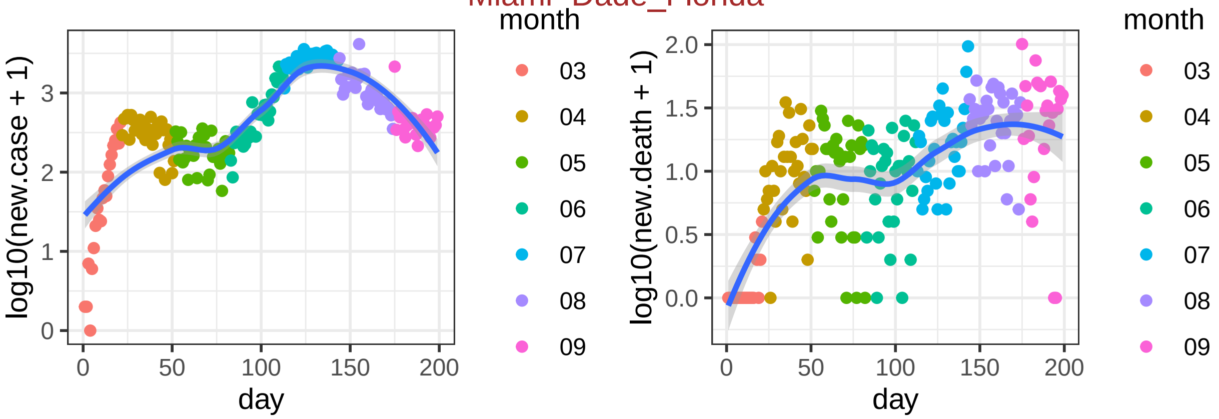
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Maricopa_Arizona



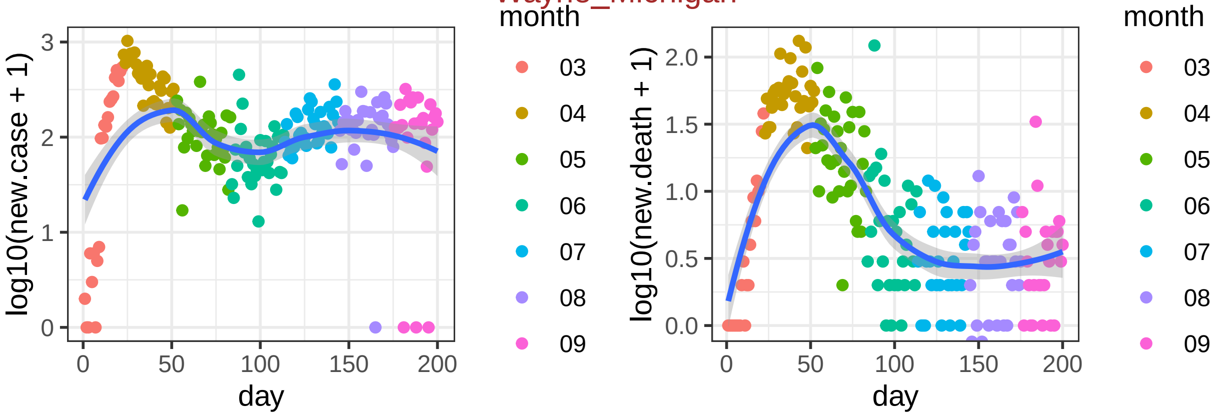
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Miami−Dade_Florida



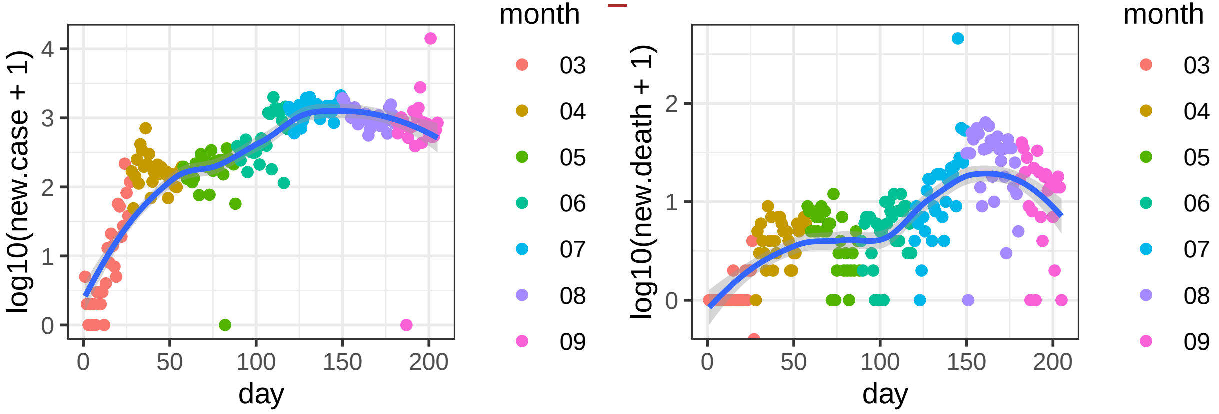
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Wayne_Michigan



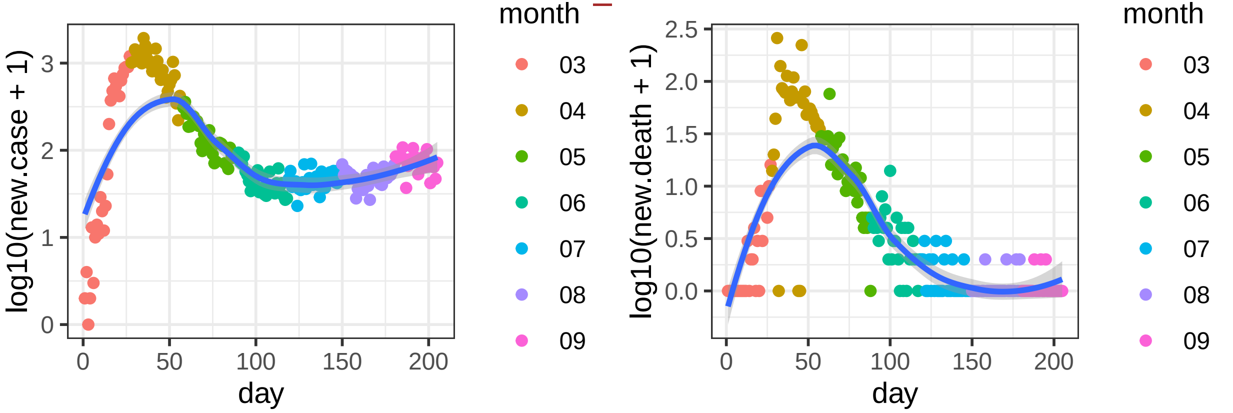
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Harris_Texas



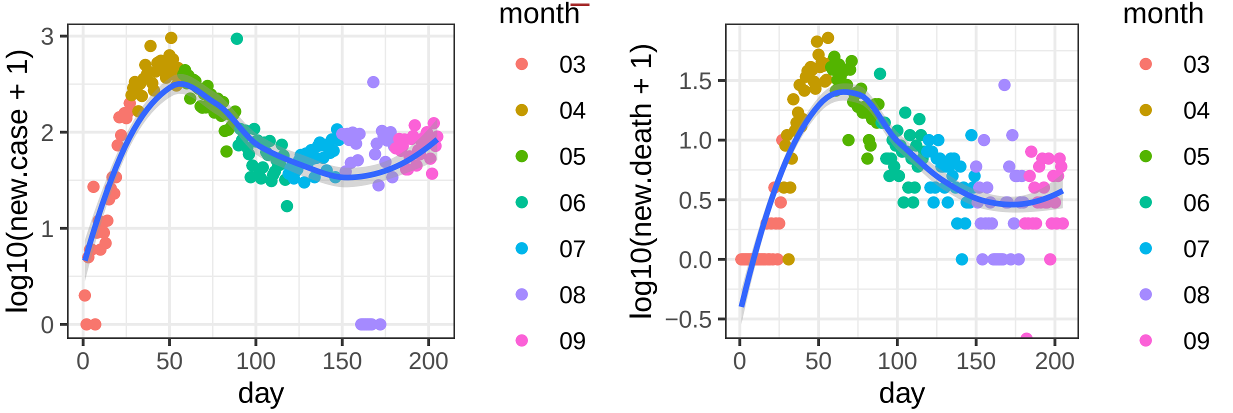
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Nassau_New York



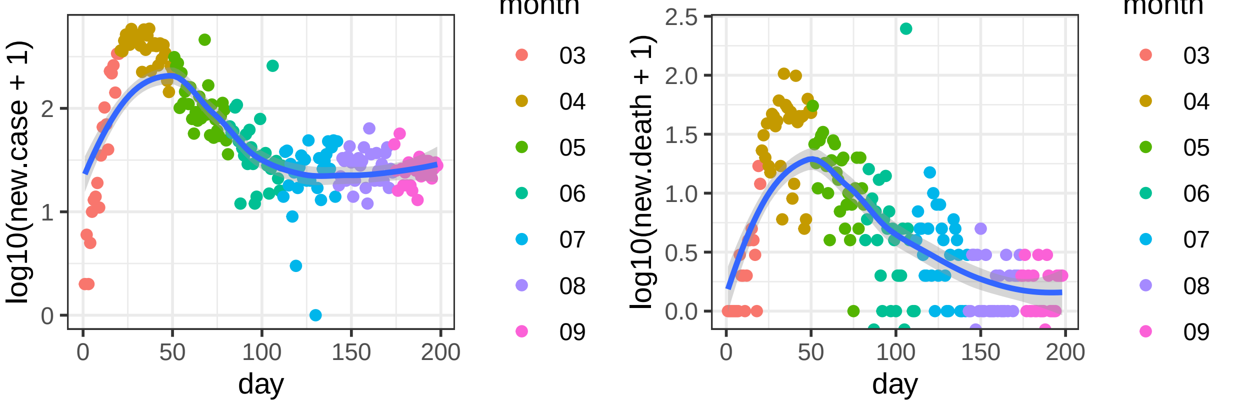
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Middlesex_Massachusetts

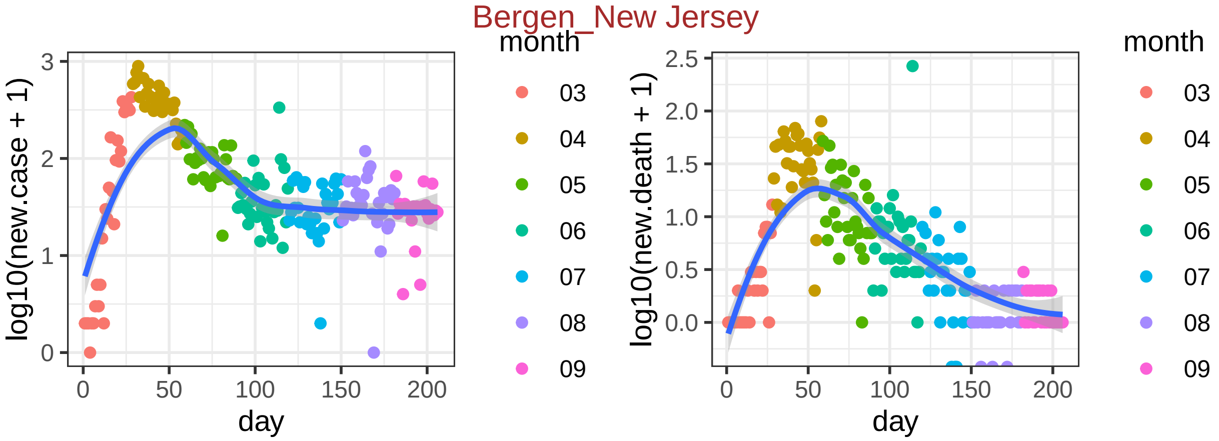


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Essex_New Jersey

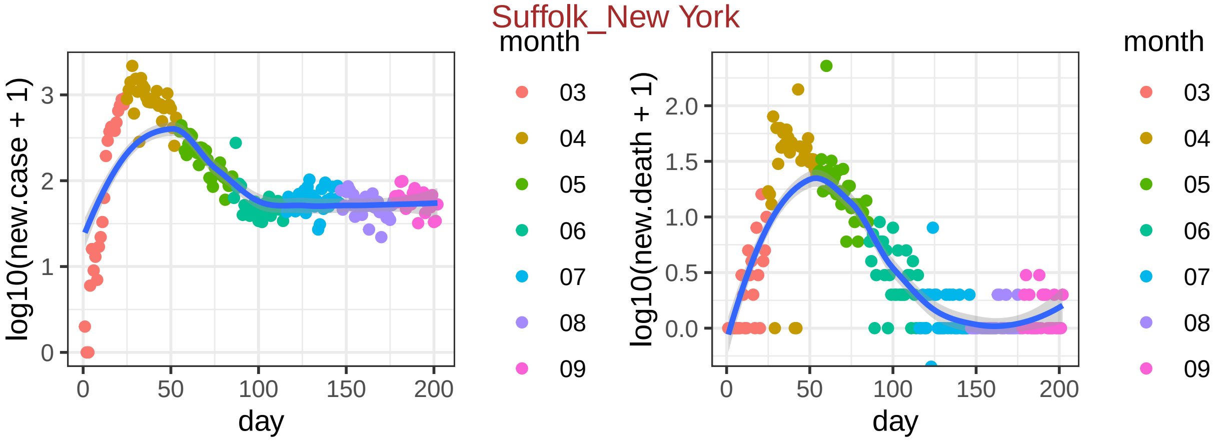


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12



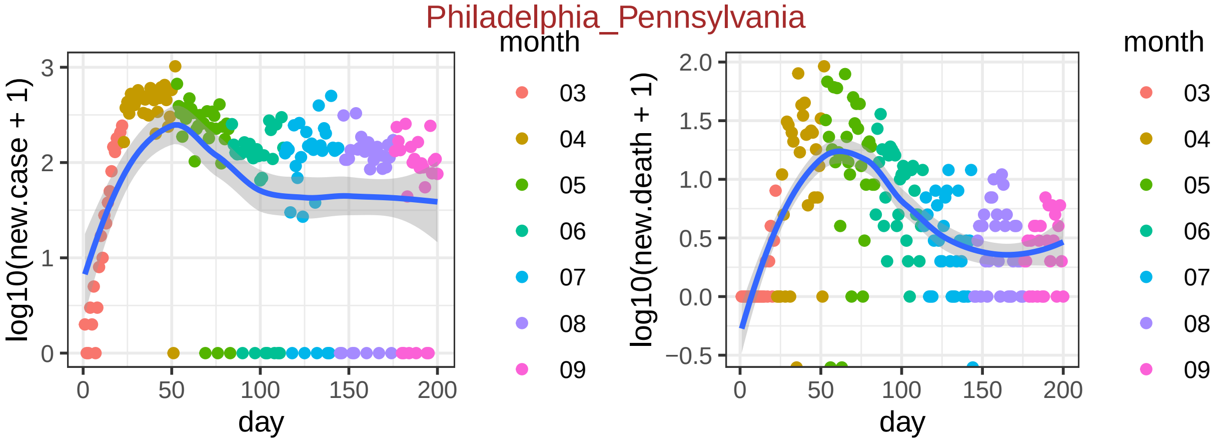
Bergen_New Jersey

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



Suffolk_New York

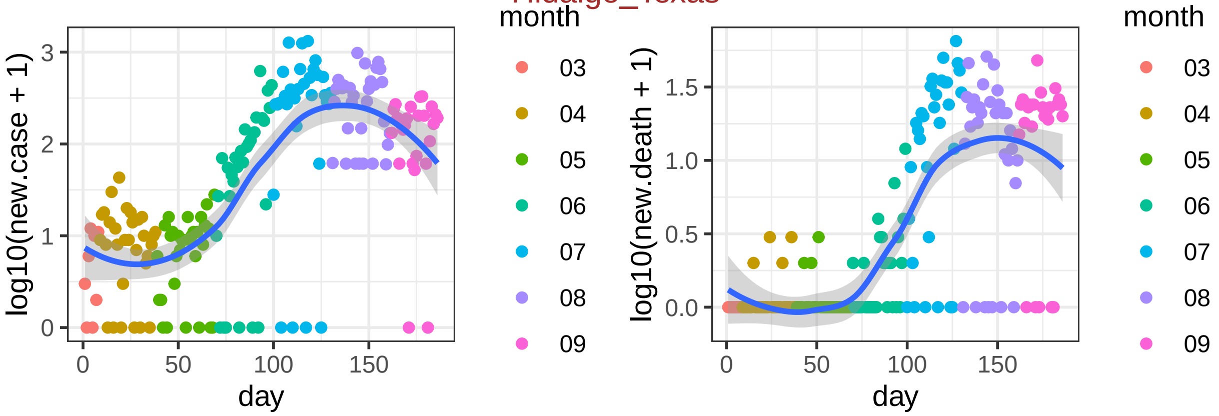
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08



Philadelphia_Pennsylvania

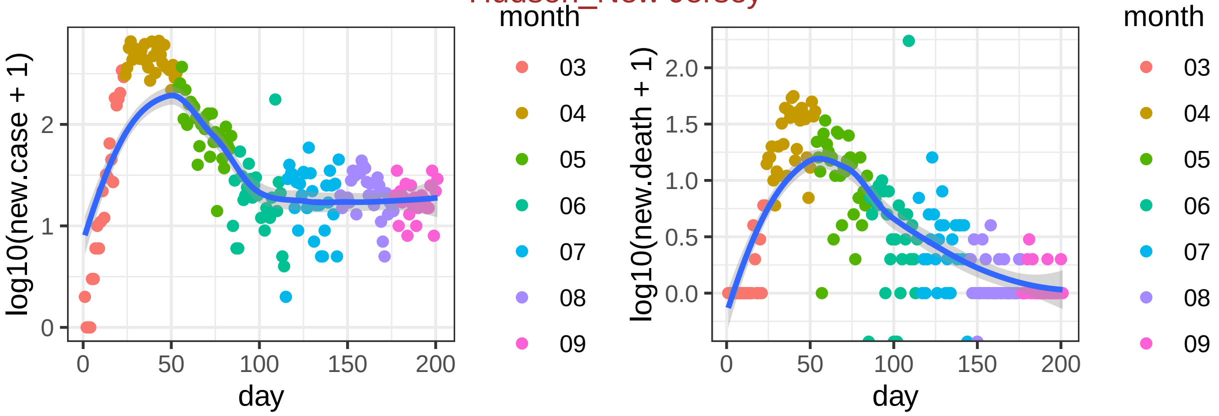
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Hidalgo_Texas



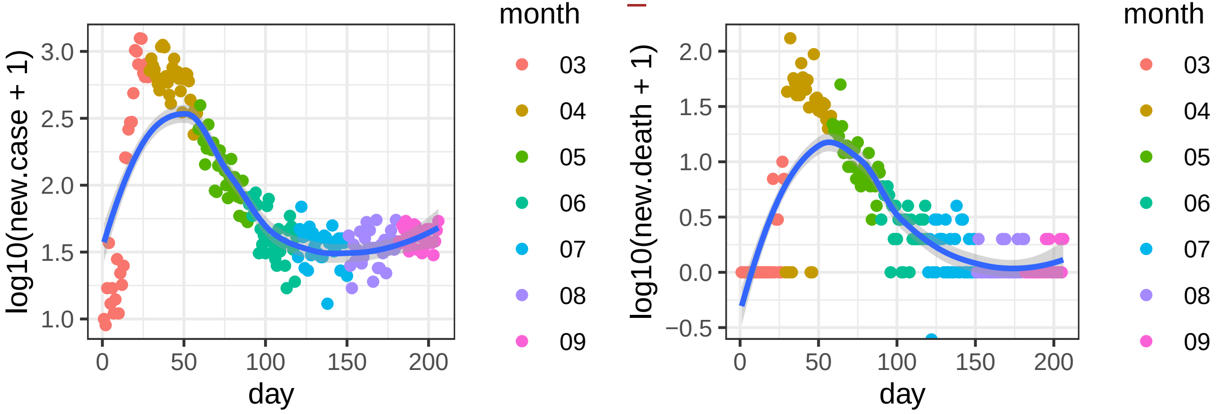
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−24

Hudson_New Jersey



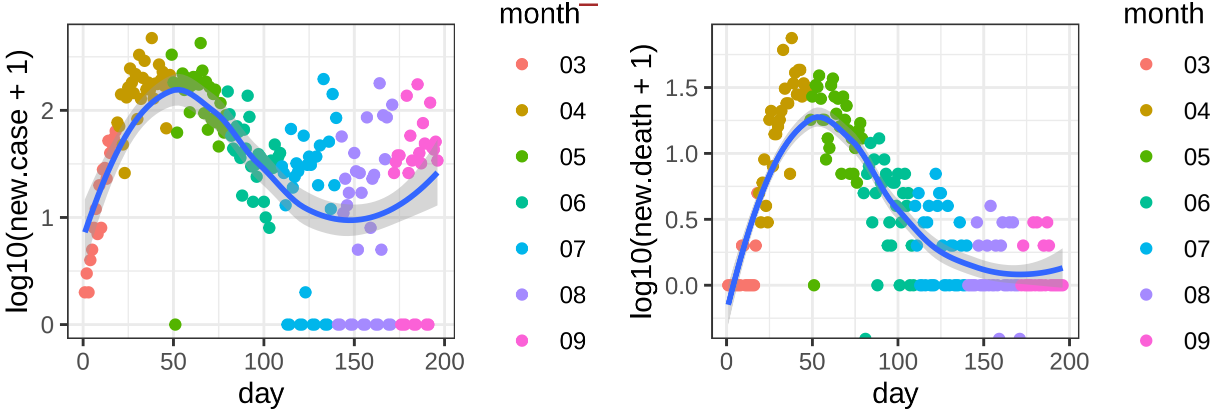
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Westchester_New York



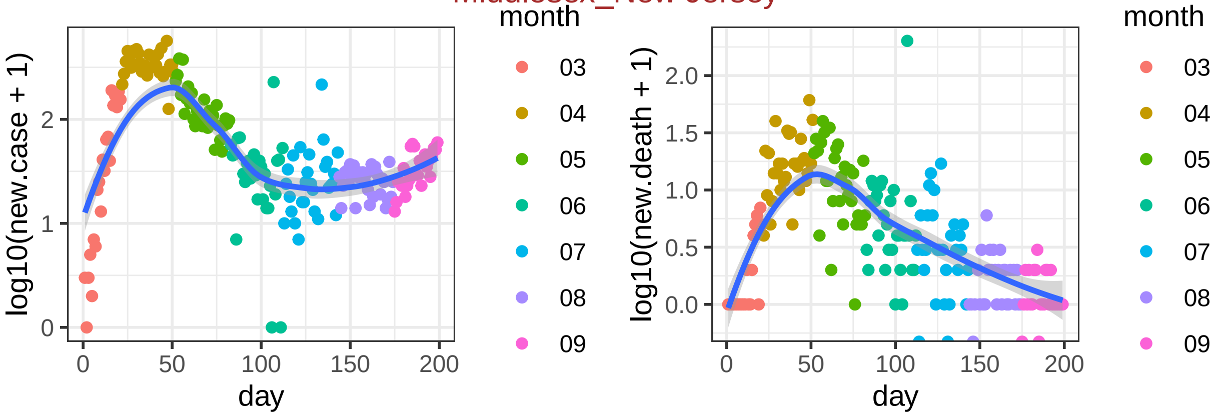
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Hartford_Connecticut



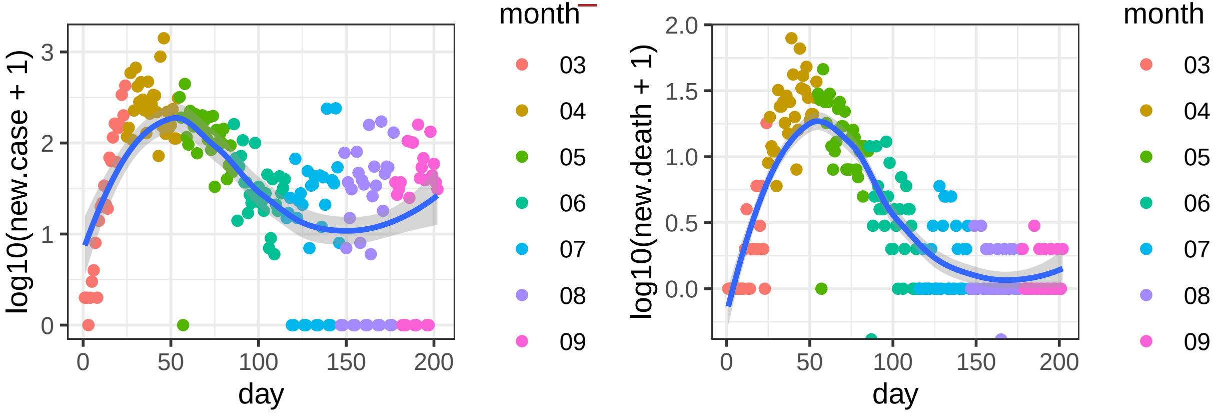
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Middlesex_New Jersey



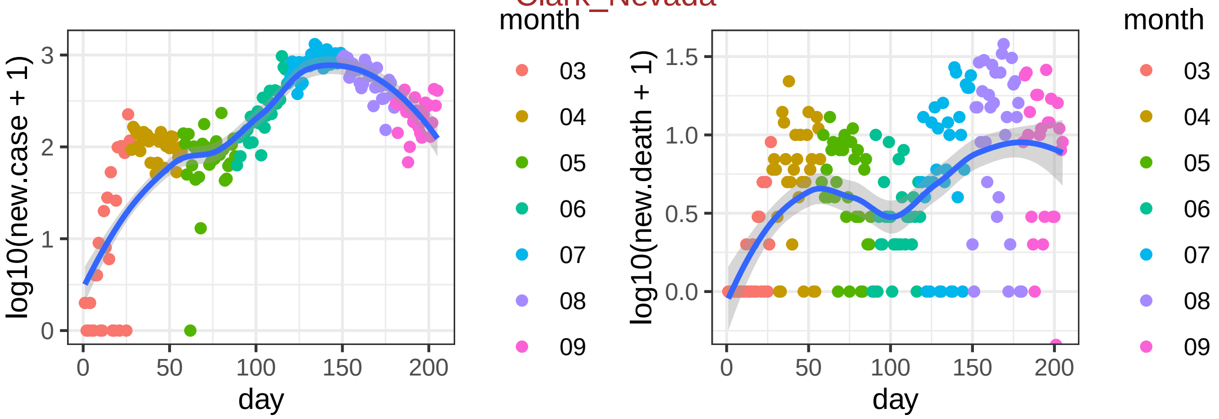
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Fairfield_Connecticut



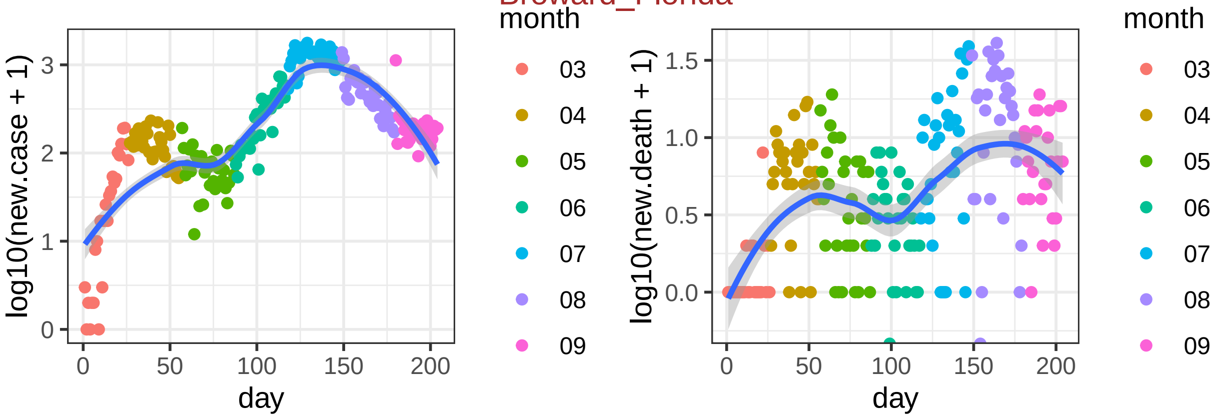
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Clark_Nevada



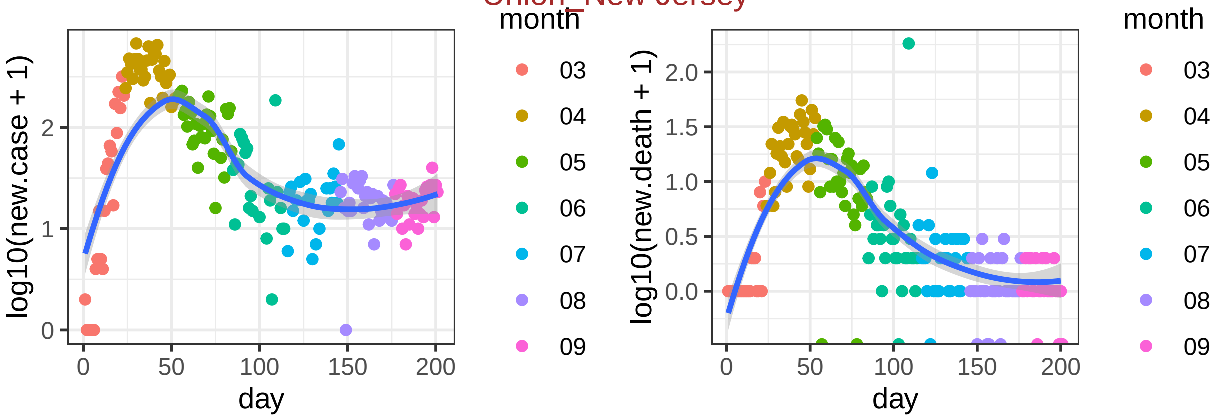
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Broward_Florida



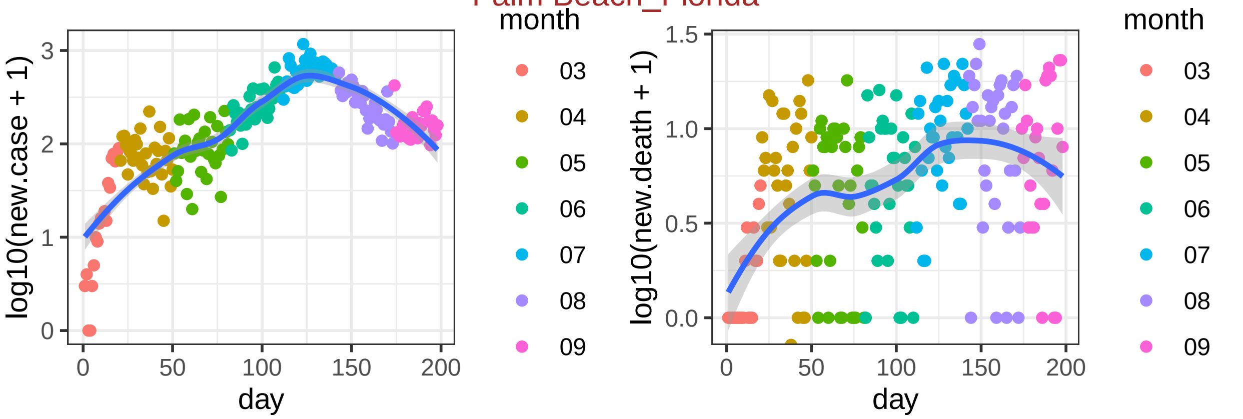
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Union_New Jersey



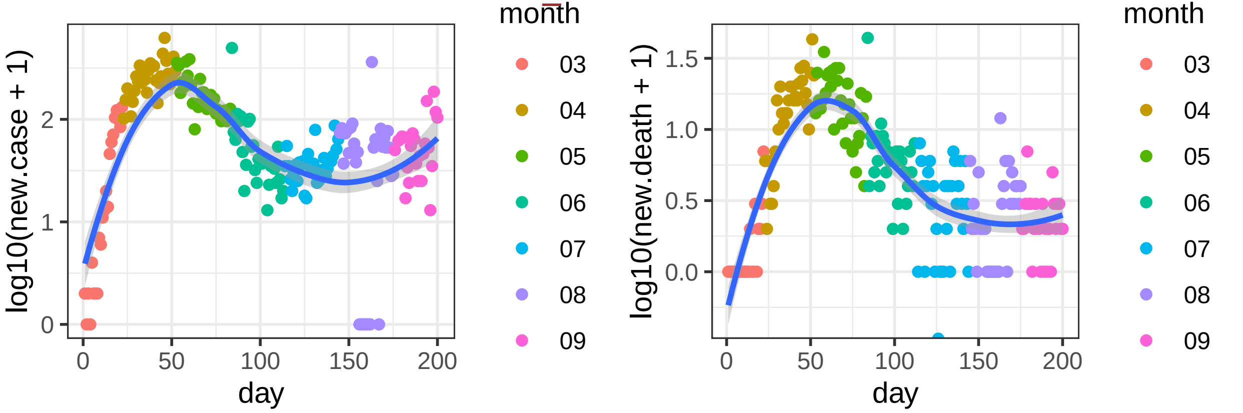
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Palm Beach_Florida



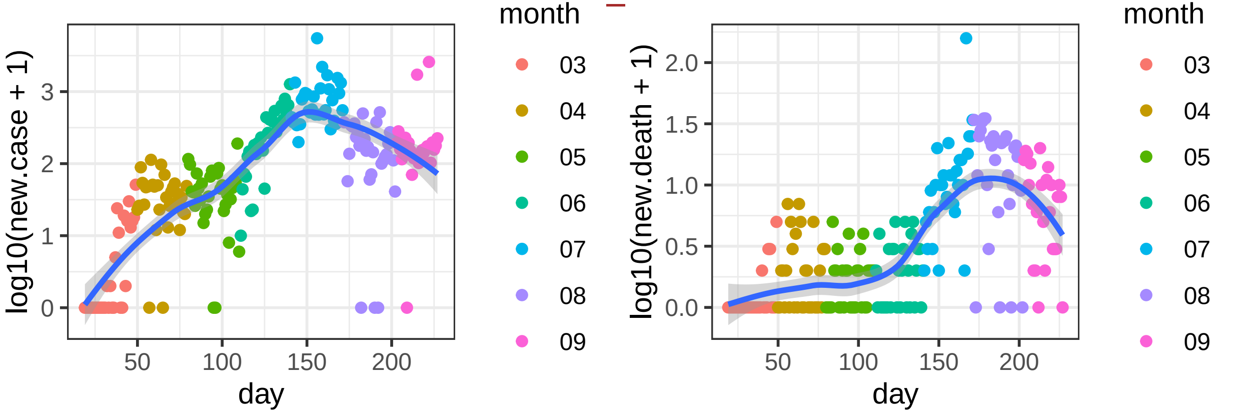
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Essex_Massachusetts



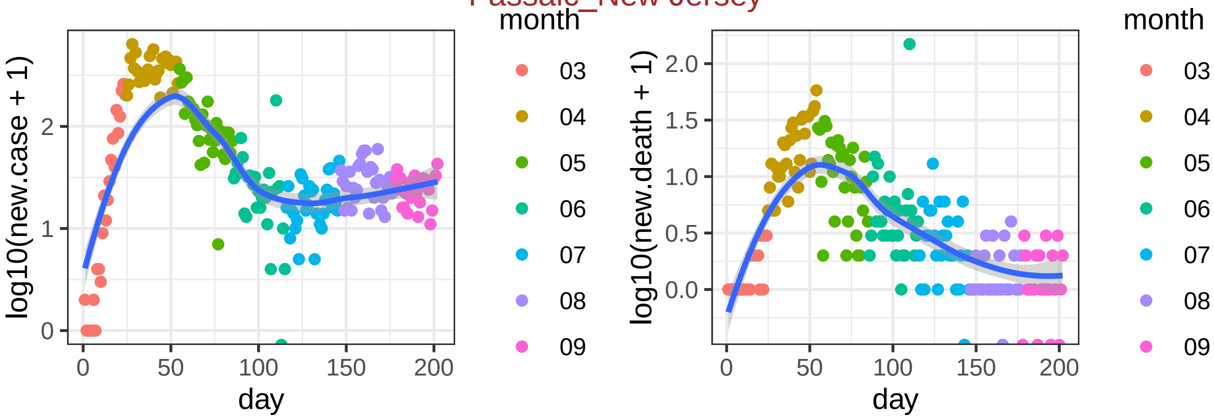
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Bexar_Texas



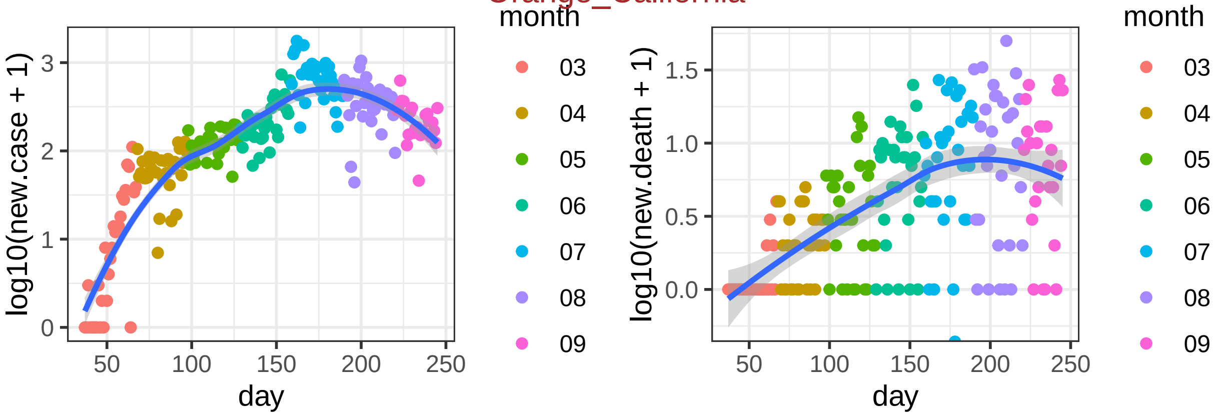
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Passaic_New Jersey



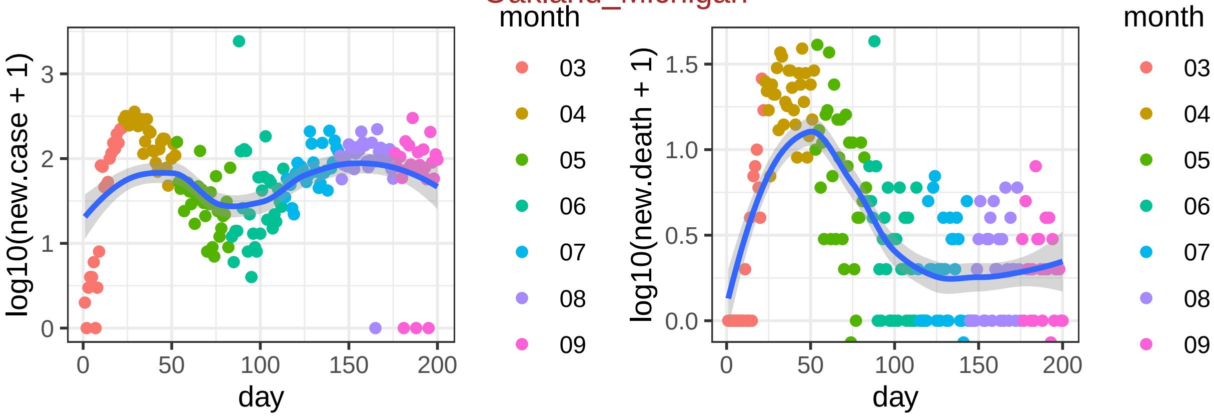
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Orange_California

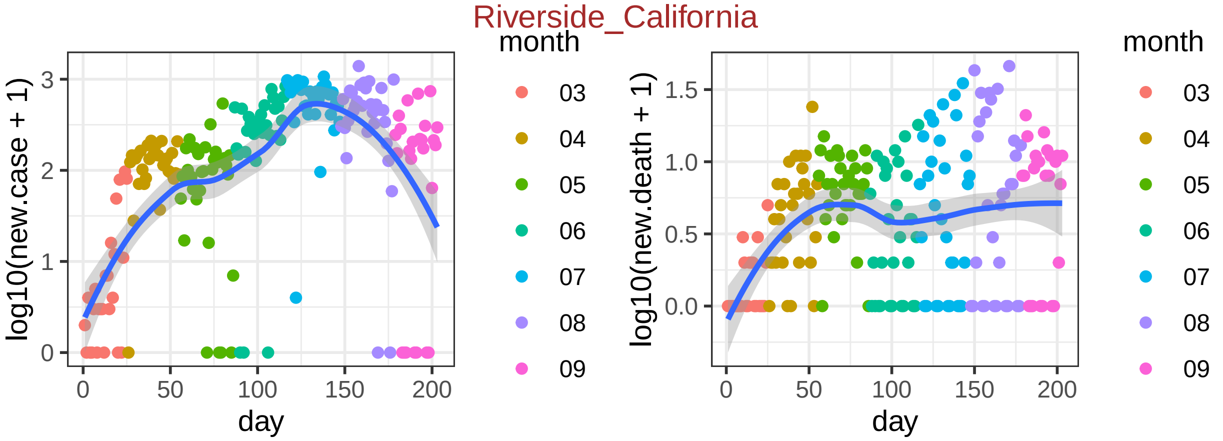


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

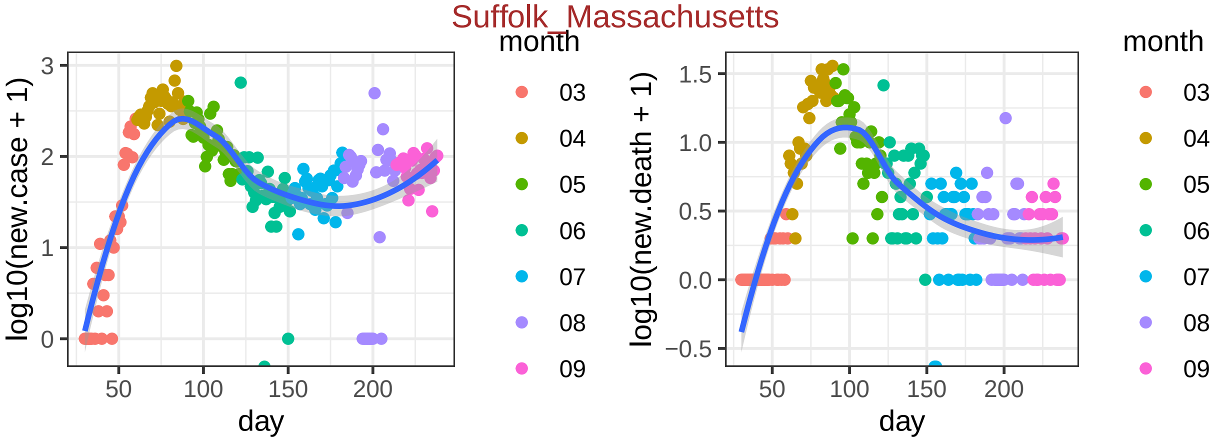
Oakland_Michigan



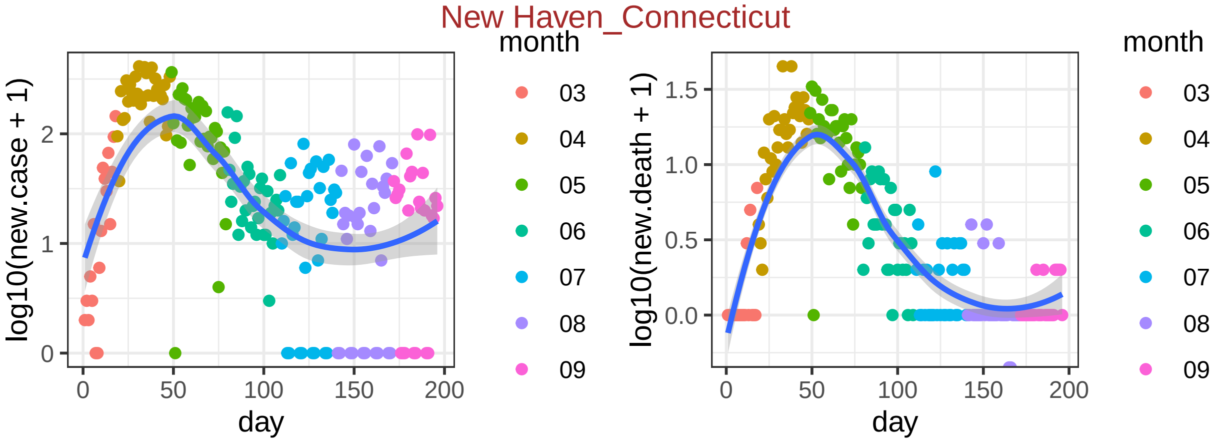
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10



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

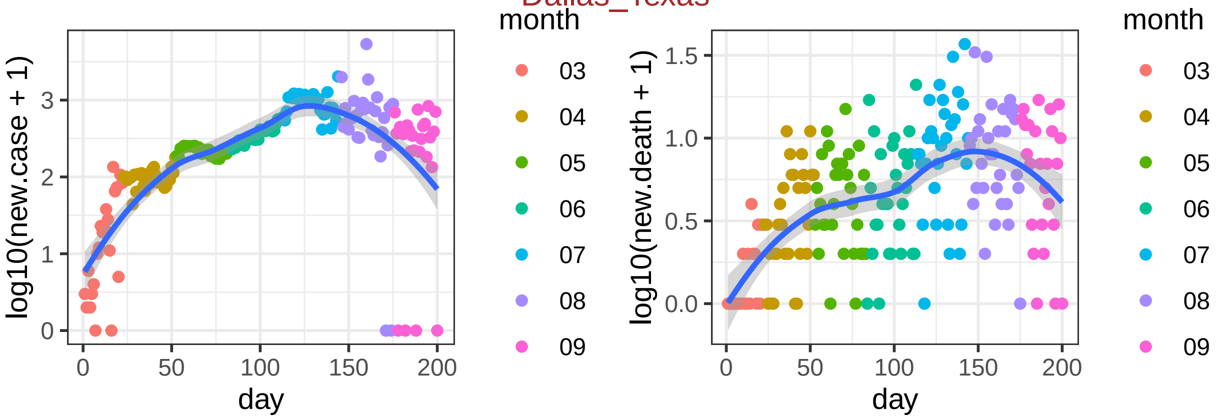


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01



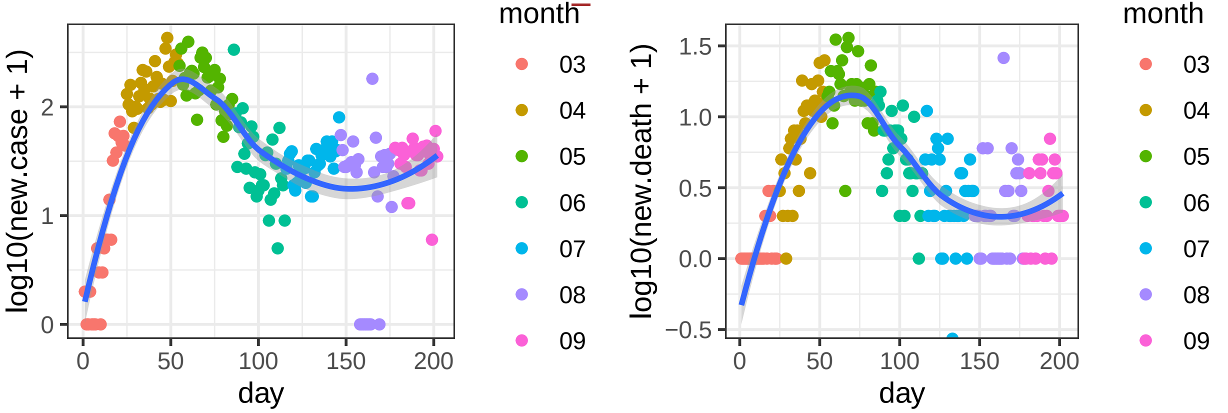
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Dallas_Texas



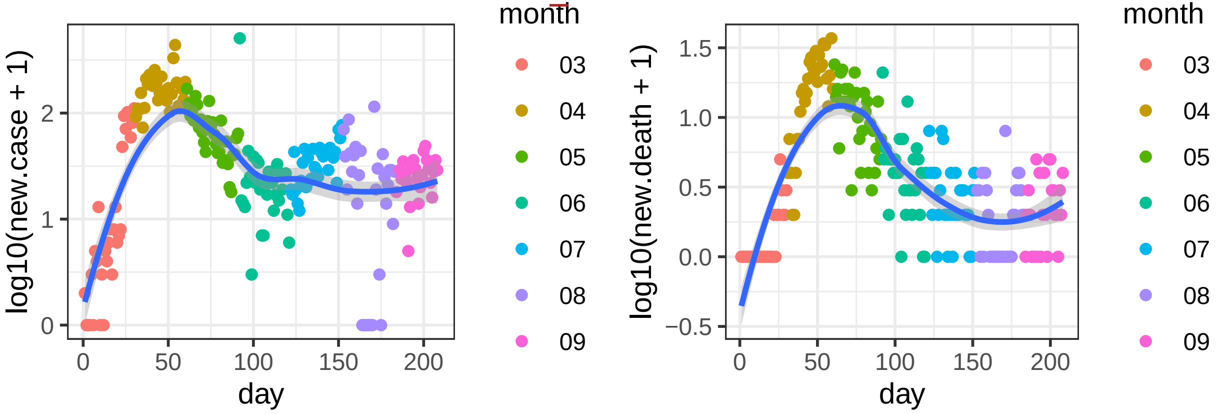
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Worcester_Massachusetts



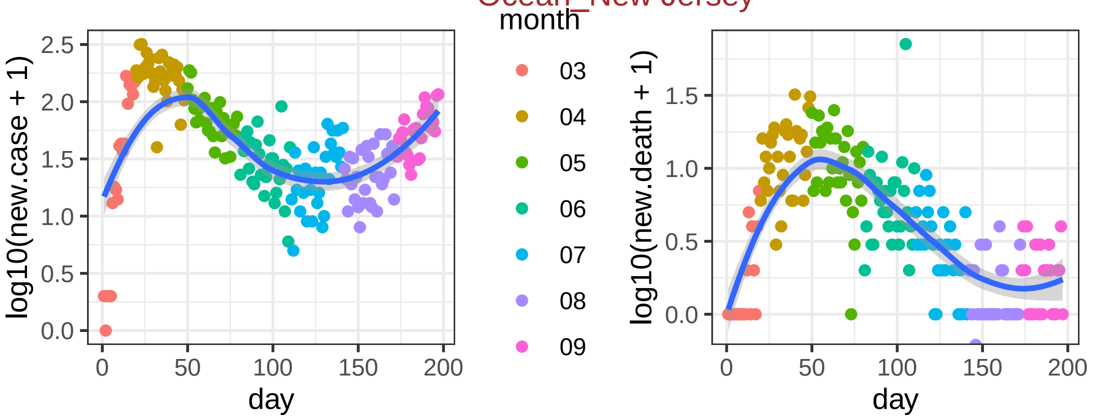
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Norfolk_Massachusetts



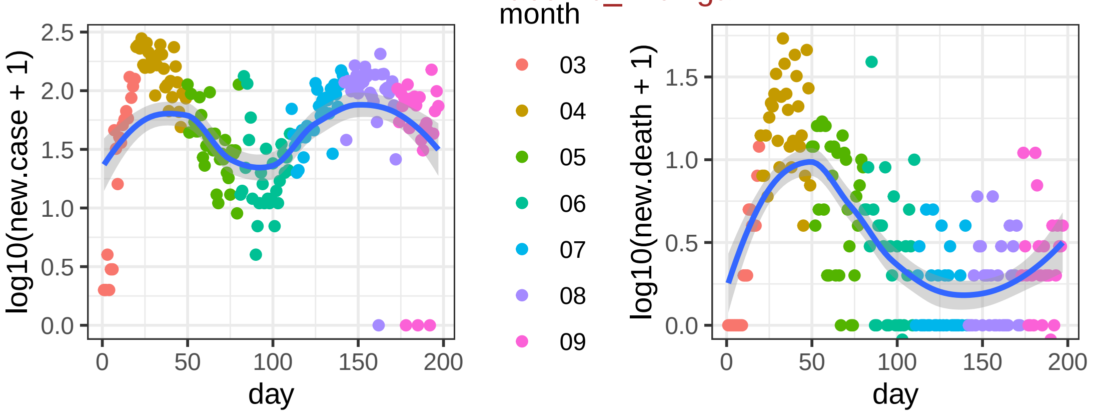
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Ocean_New Jersey



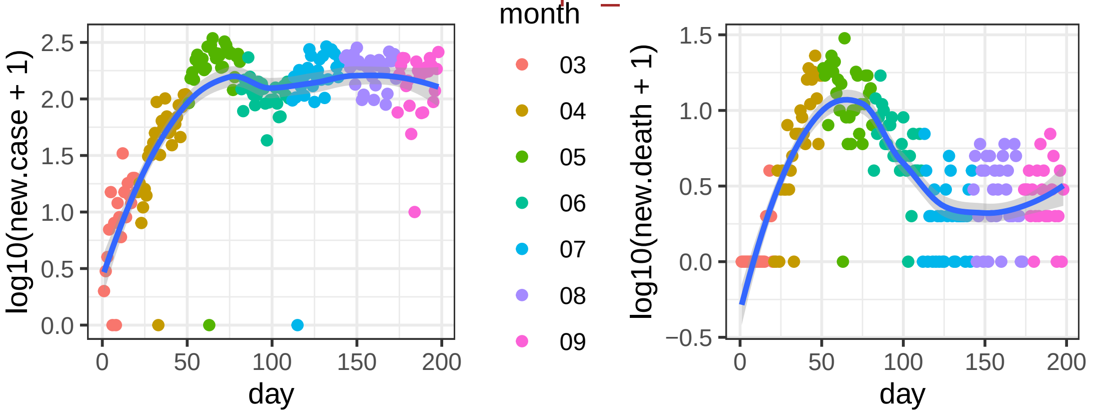
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Macomb_Michigan



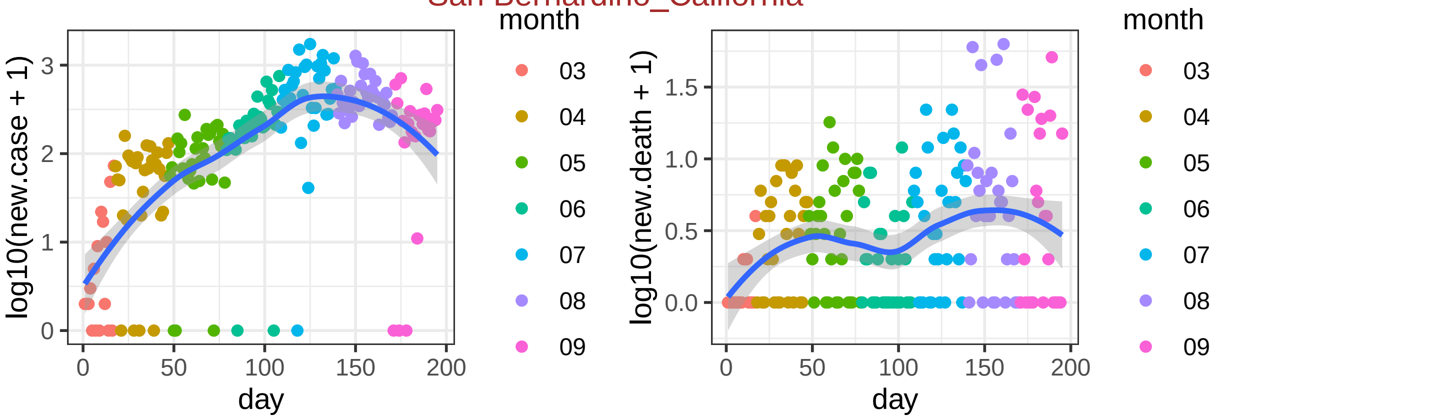
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Hennepin_Minnesota



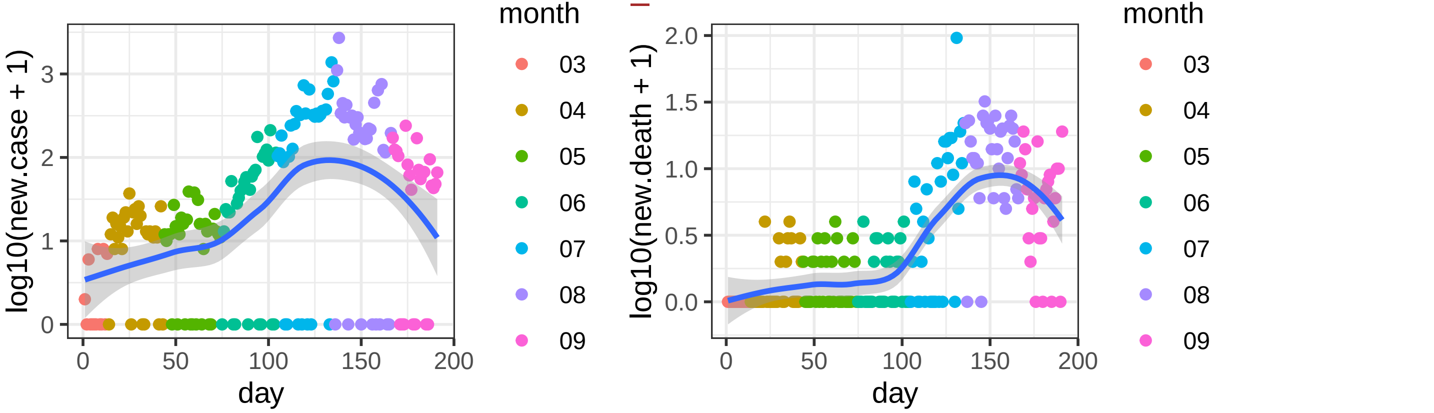
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

San Bernardino_California



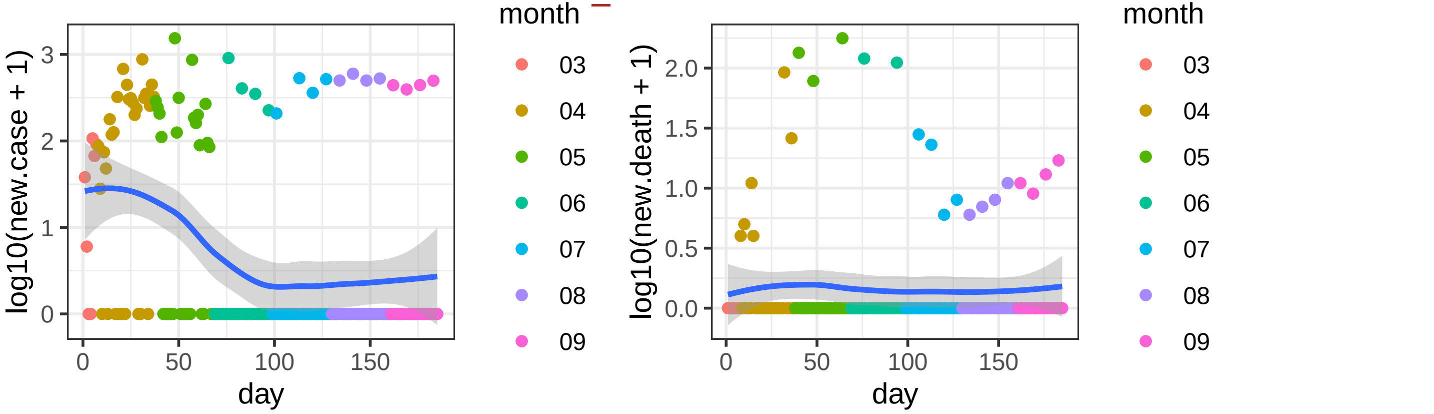
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

Cameron_Texas



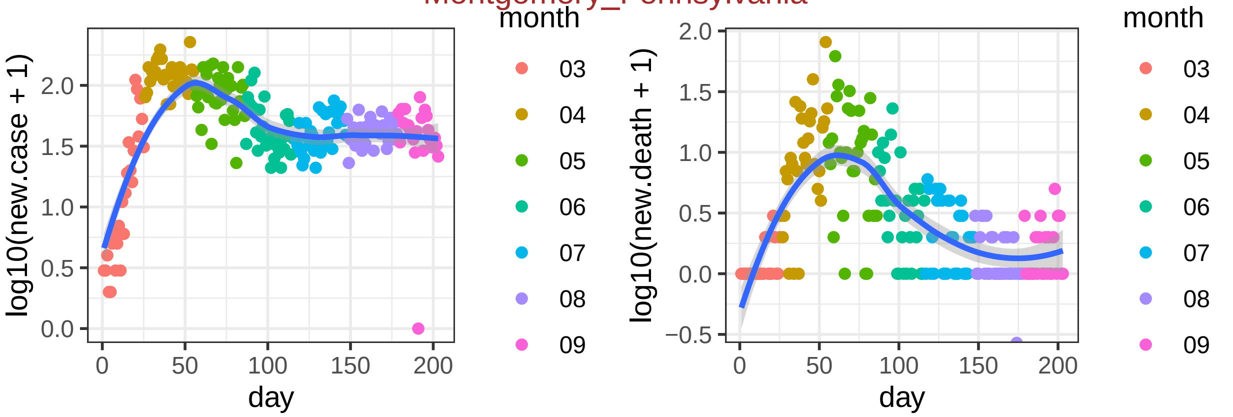
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−19

Providence_Rhode Island



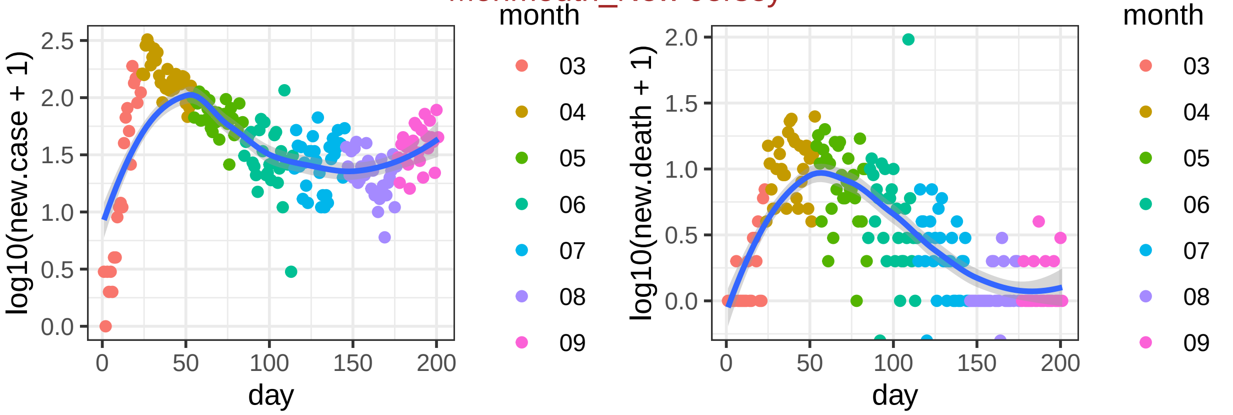
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−25

Montgomery_Pennsylvania



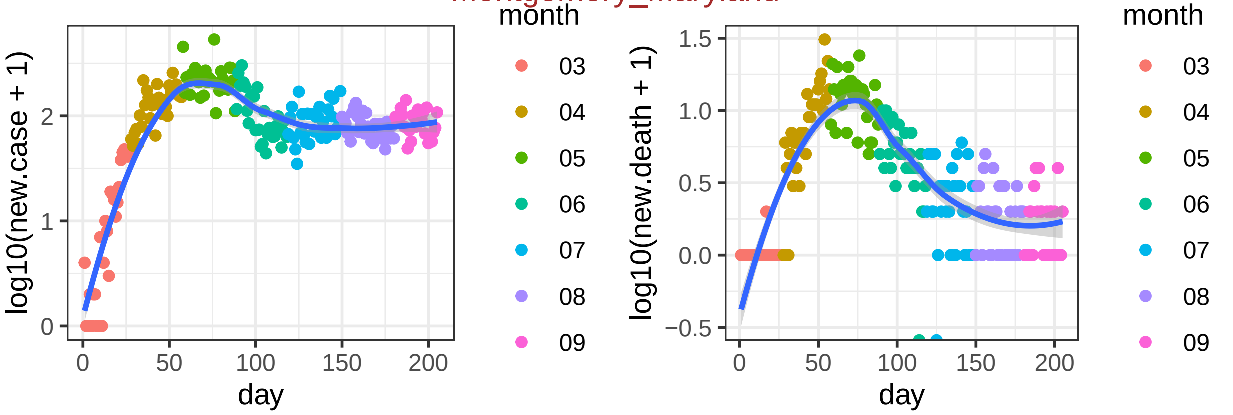
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Monmouth_New Jersey



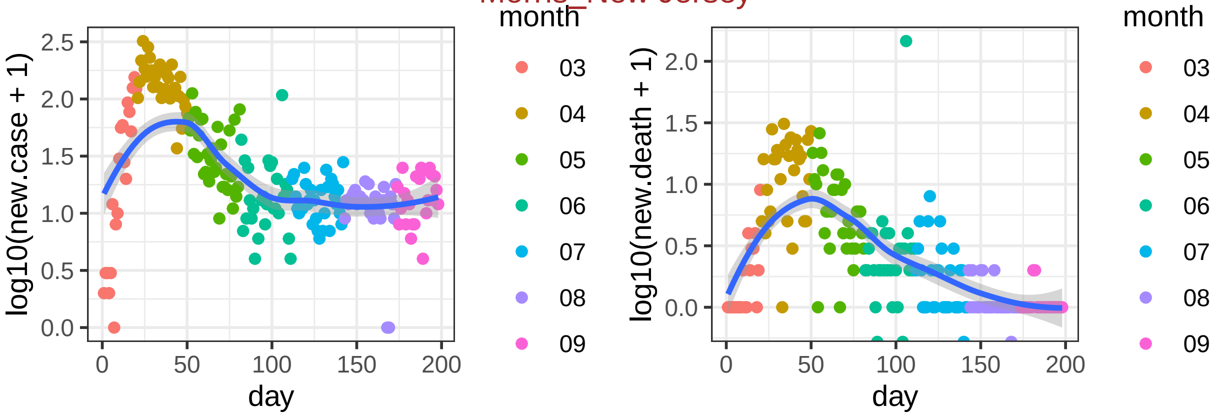
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Montgomery_Maryland



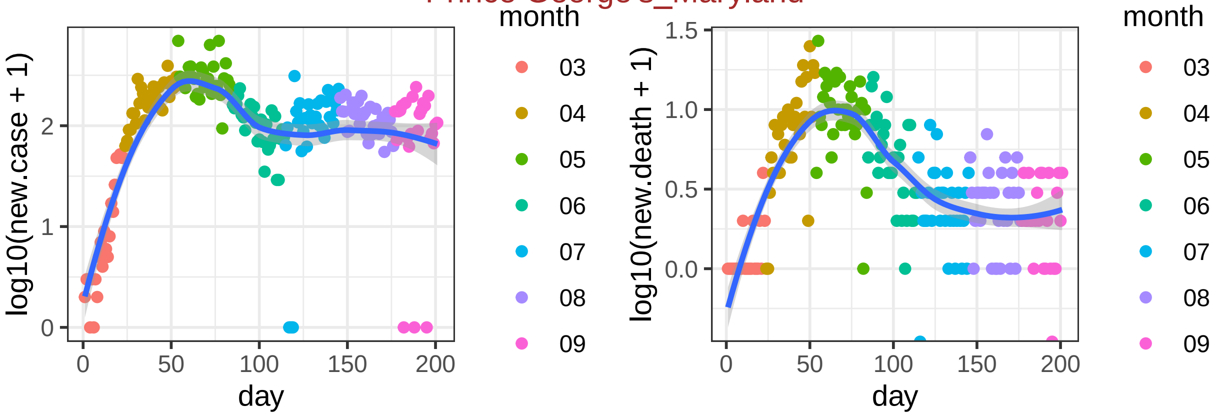
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Morris_New Jersey



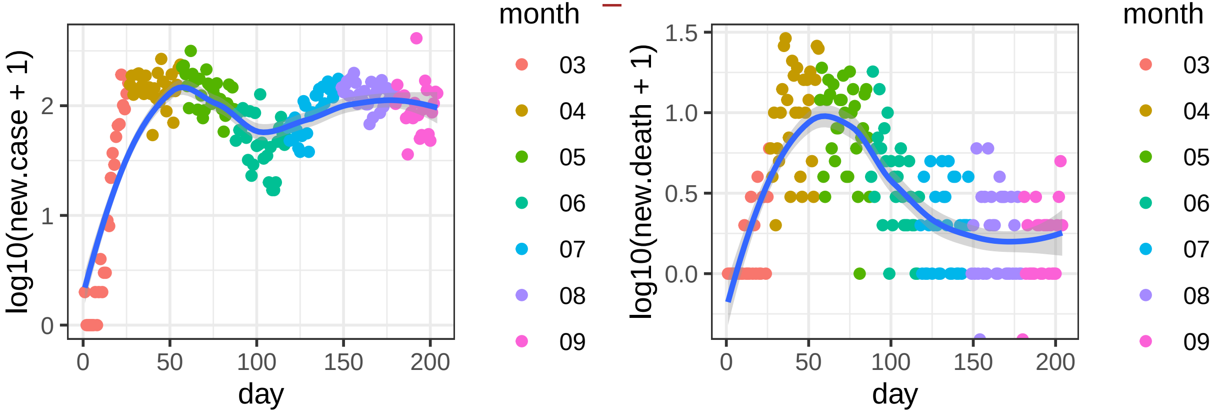
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Prince George's_Maryland



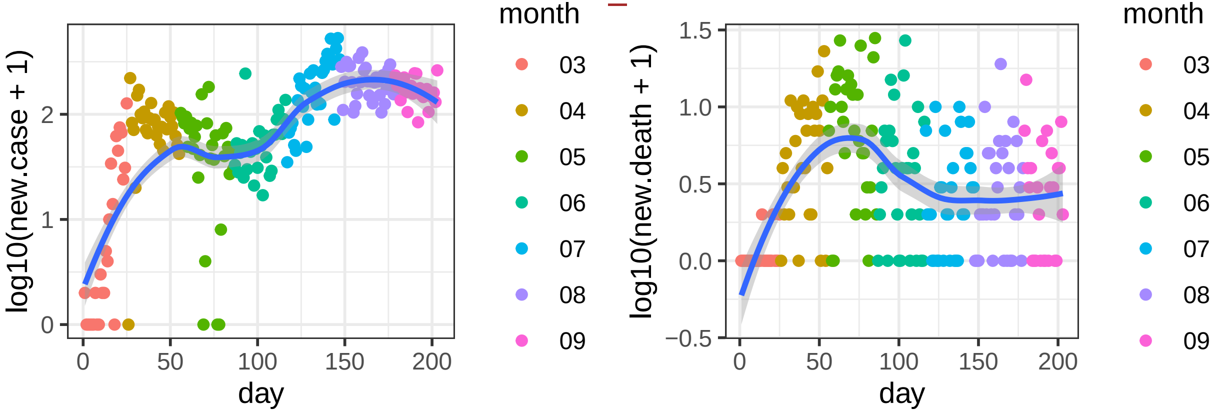
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Marion_Indiana



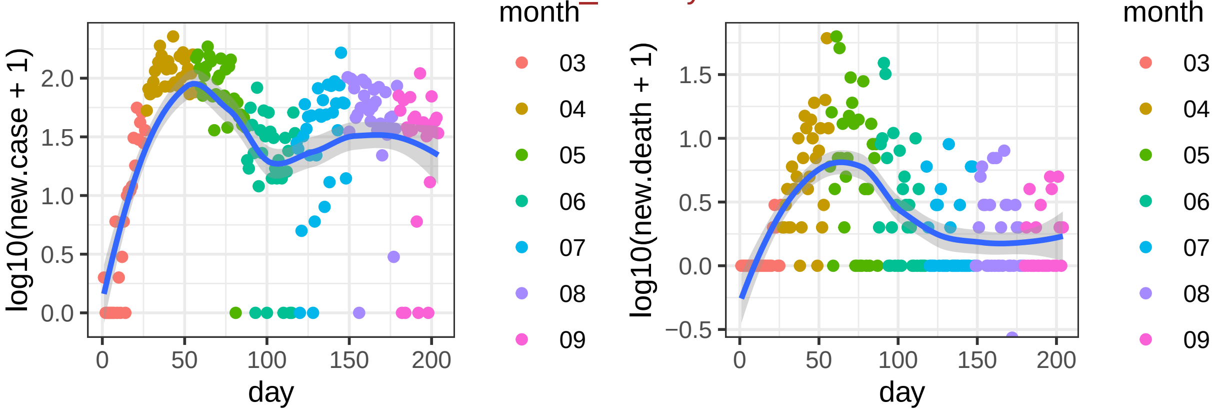
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

St. Louis_Missouri



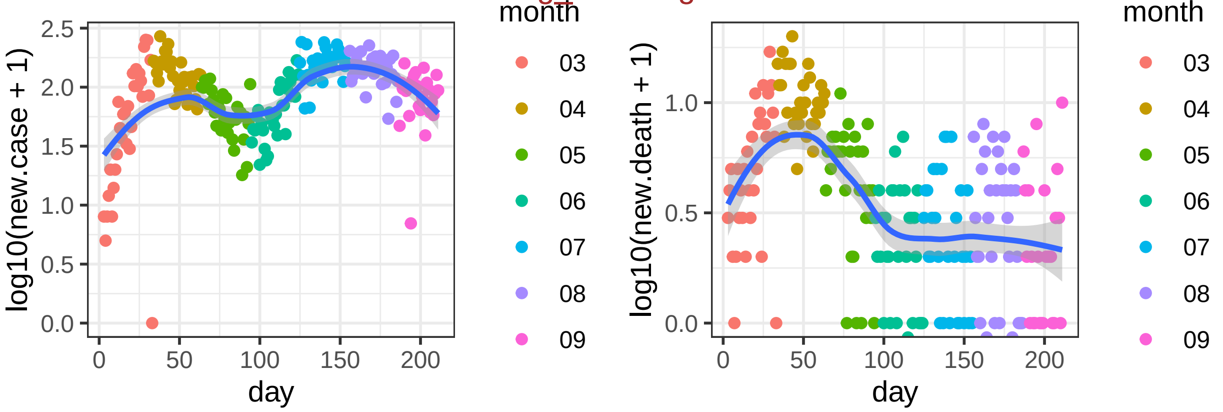
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Delaware_Pennsylvania

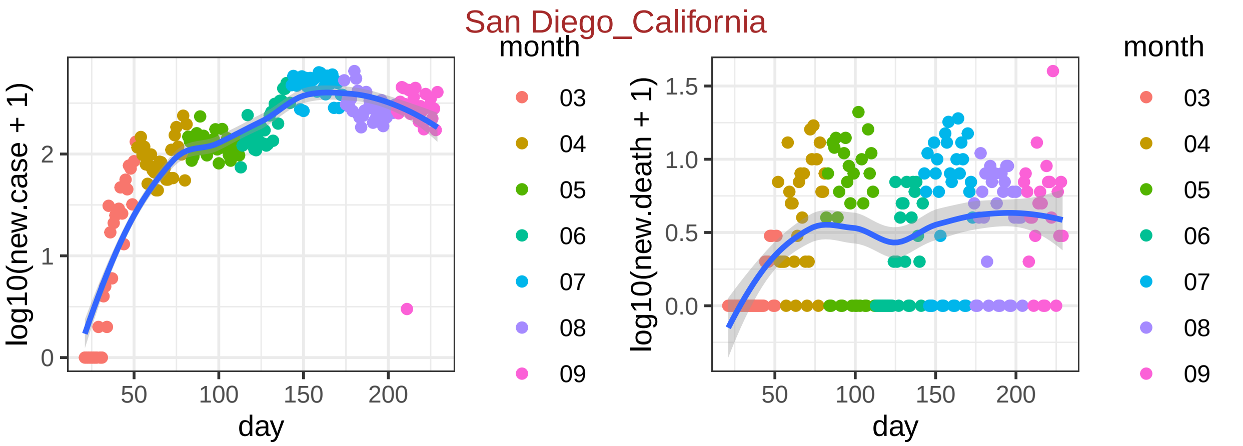


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

King_Washington



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

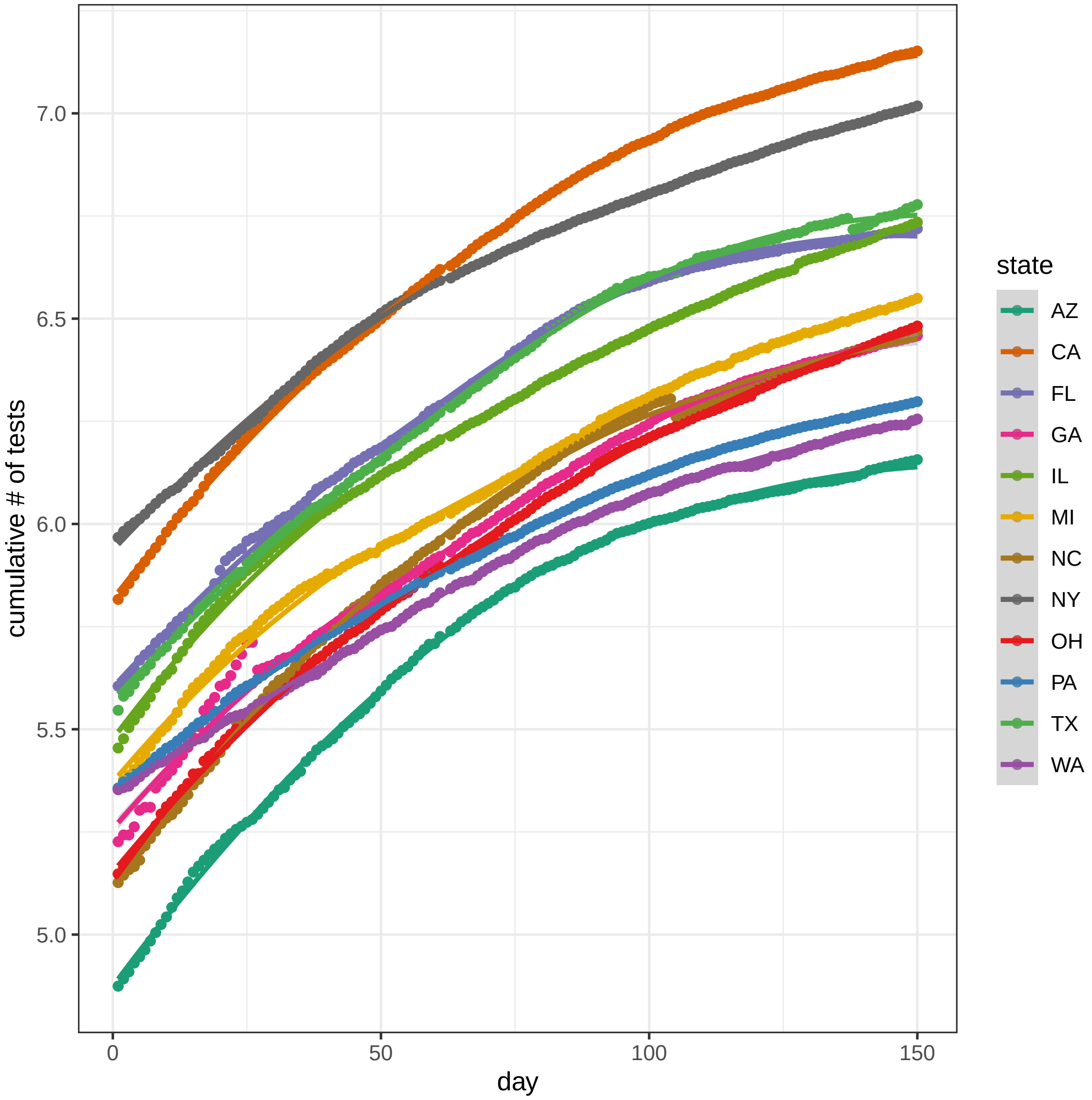


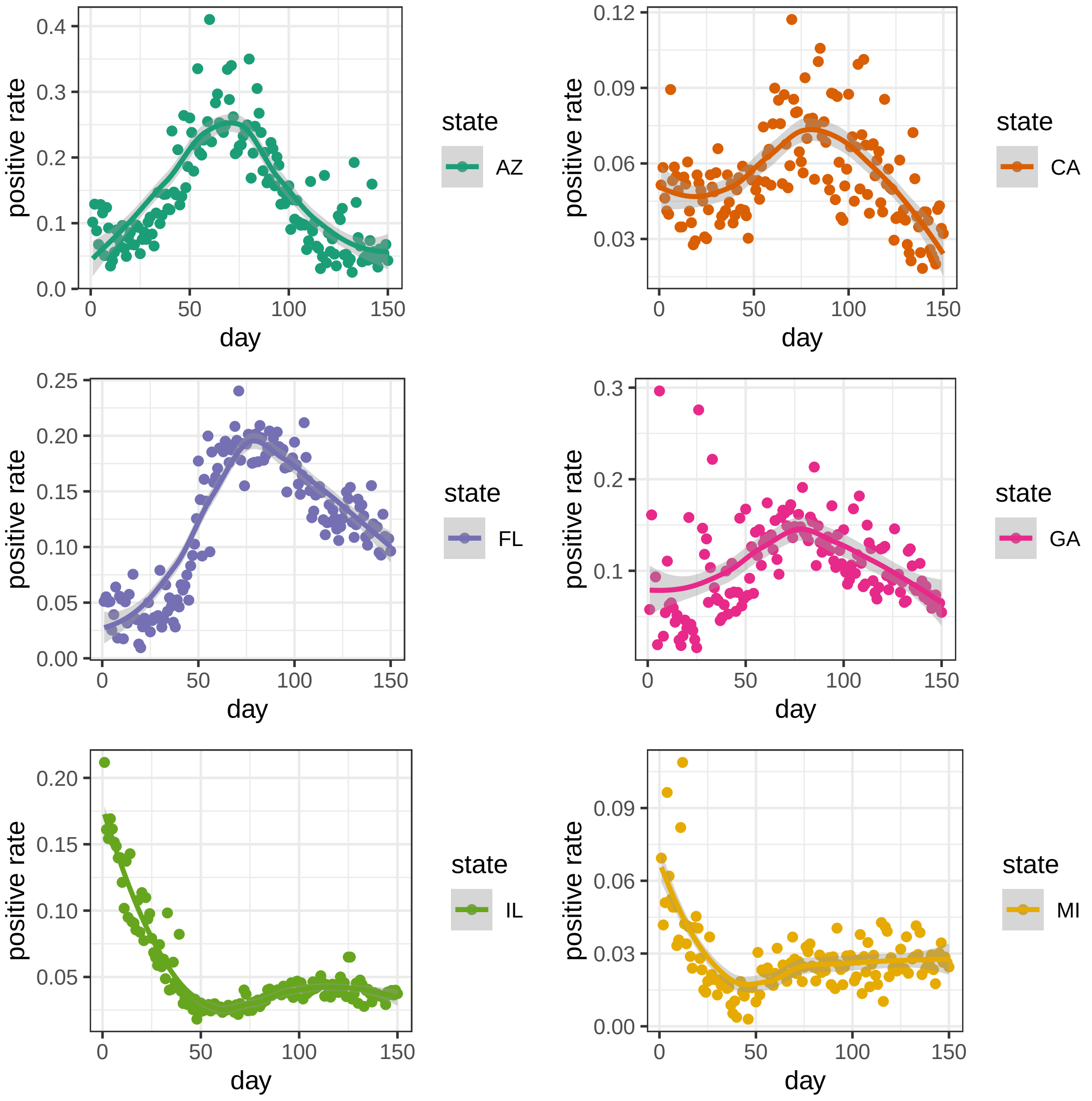
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

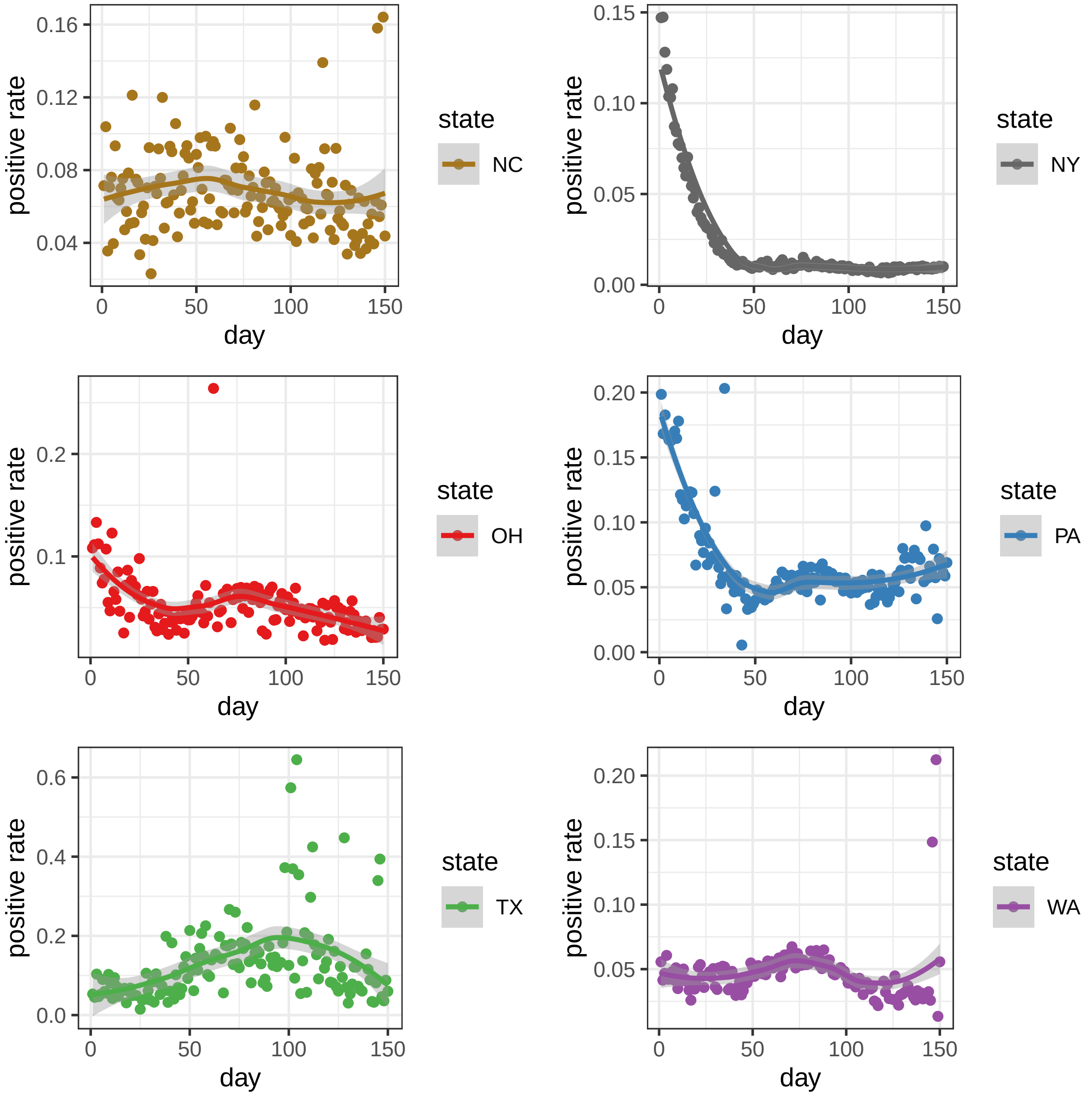
COVID Trackng

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.


cumulative # of tests
7.0
6.5
6.0
5.5
5.0
0
50
100
150
day
state
AZ
CA
FL
GA
IL
MI
NC
NY
OH
PA
TX
WA


positive rate
0.4
0.3
0.2
0.1
0.0
0
50
100
150
day
state
AZ
positive rate
0.12
0.09
0.06
0.03
0
50
100
150
day
state
CA
positive rate
0.25
0.20
0.15
0.10
0.05
0.00
0
50
100
150
day
state
FL
positive rate
0.3
0.2
0.1
0
50
100
150
day
state
GA
positive rate
0.20
0.15
0.10
0.05
0
50
100
150
day
state
IL
positive rate
0.09
0.06
0.03
0.00
0
50
100
150
day
state
MI



Session information

```
sessionInfo()
```

```
## R version 3.6.2 (2019-12-12)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Catalina 10.15.6
##
## 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] RColorBrewer_1.1-2 httr_1.4.1         ggpubr_0.2.5       magrittr_1.5
## [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    farver_2.0.3     purrr_0.3.3      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
```