Exploration of COVID-19 tracking data from multiple resources

Wei Sun

2020-09-12

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

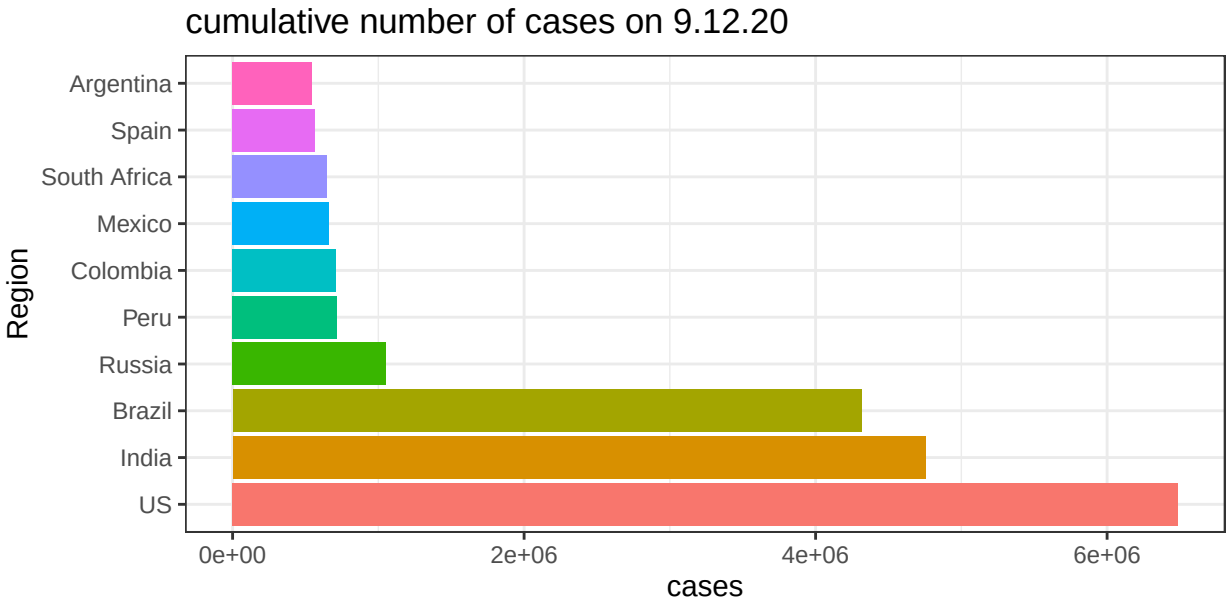
JHU

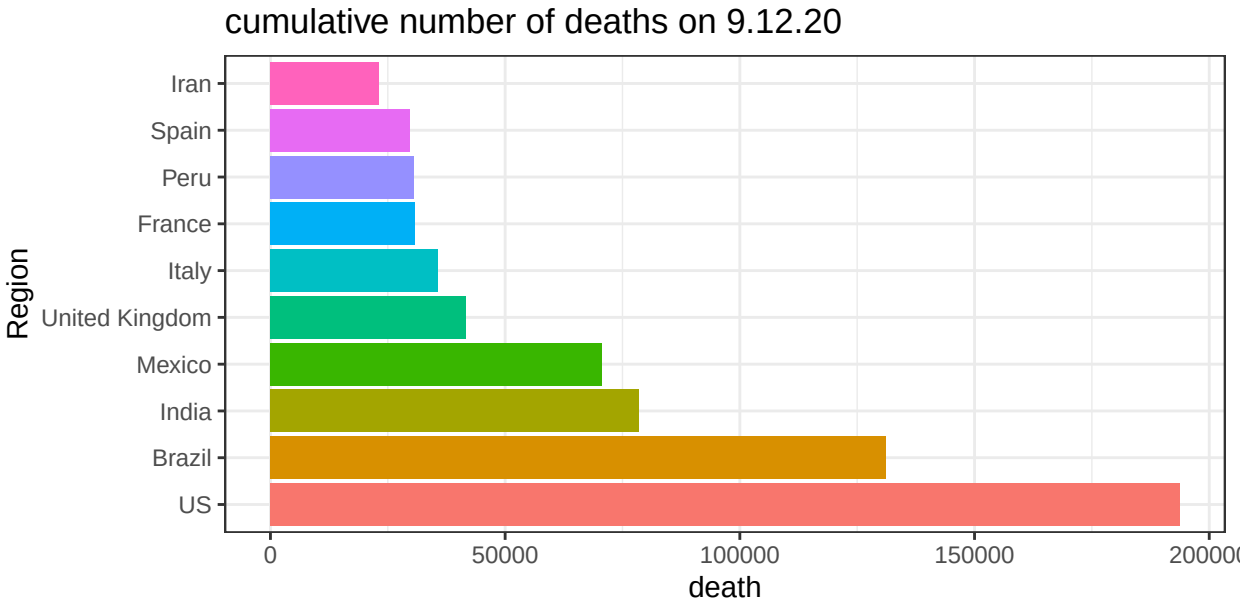
Assume you have cloned the JHU Github repository on your local machine at "../COVID-19".

time series data

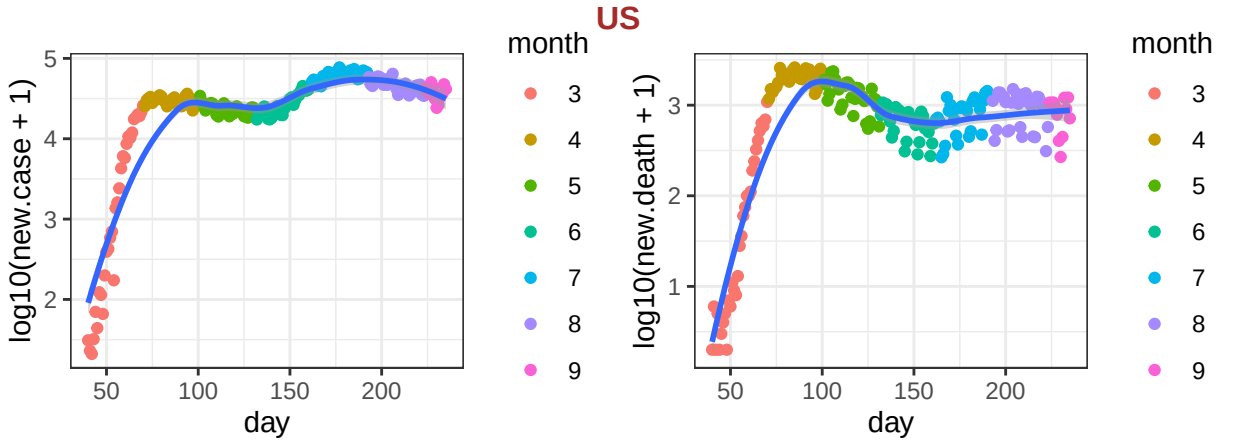
The time series provide counts (e.g., confirmed cases, deaths) starting from Jan 22nd, 2020 for 253 locations. Currently there is no data of individual US state in these time series data files.

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

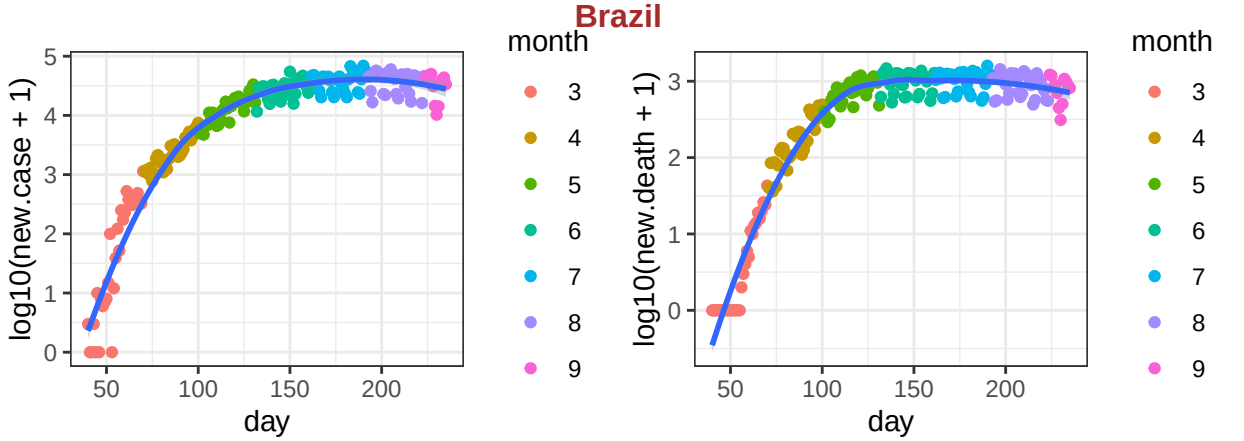




Next, I check for each country/region, what is the number of new cases/deaths? This data is important to understand what is the trend under different situations, e.g., population density, social distance policies etc. Here I checked the top 10 countries/regions with the highest number of deaths.

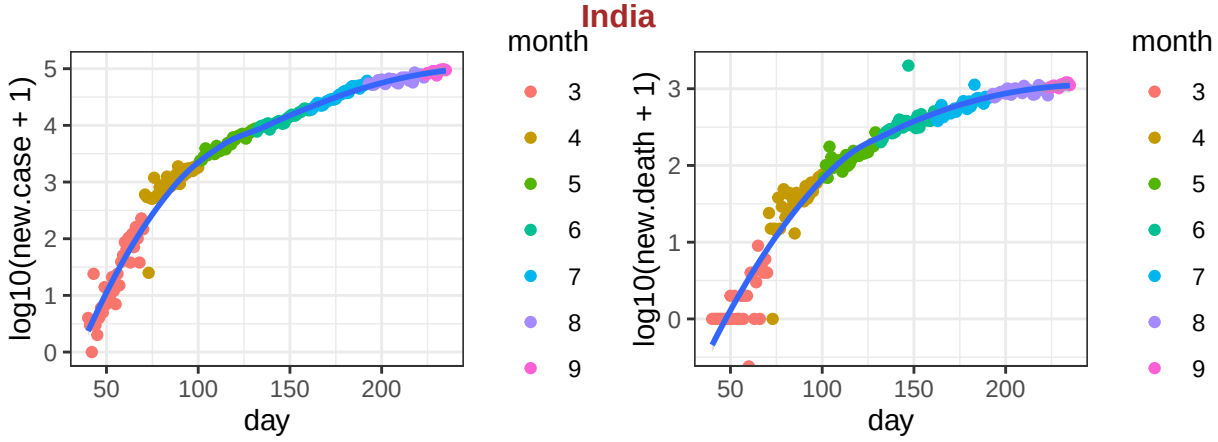


data source: https://github.com/CSSEGISandData/COVID−19



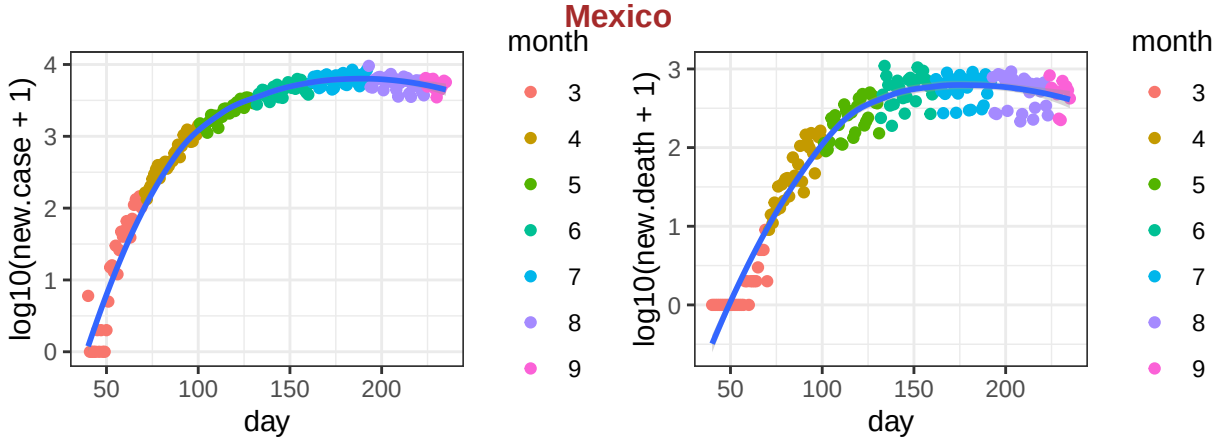
data source: https://github.com/CSSEGISandData/COVID−19

India



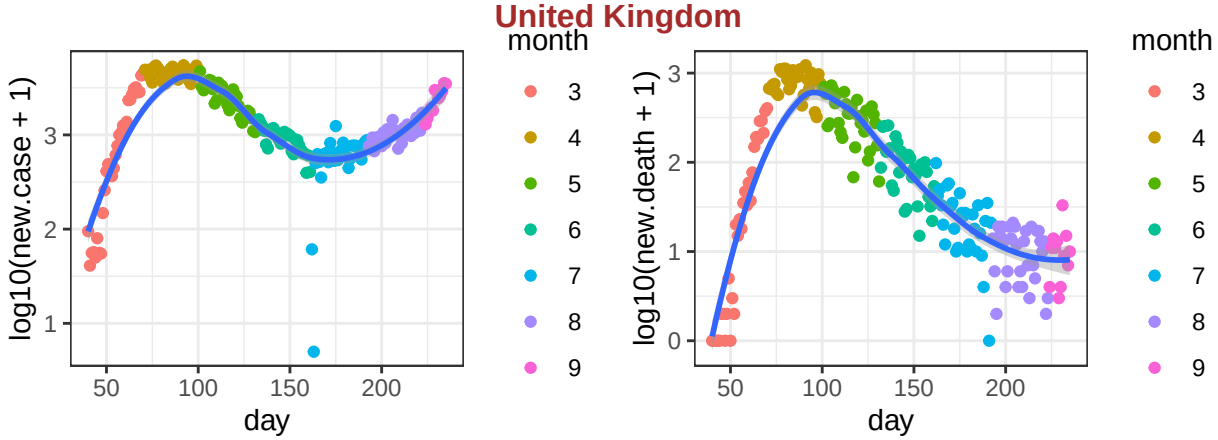
data source: https://github.com/CSSEGISandData/COVID-19

Mexico



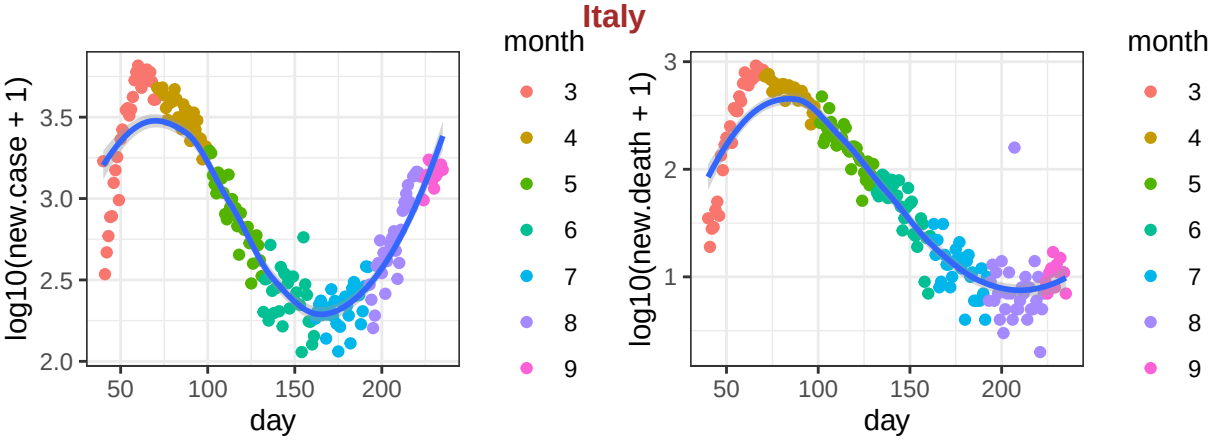
data source: https://github.com/CSSEGISandData/COVID-19

United Kingdom



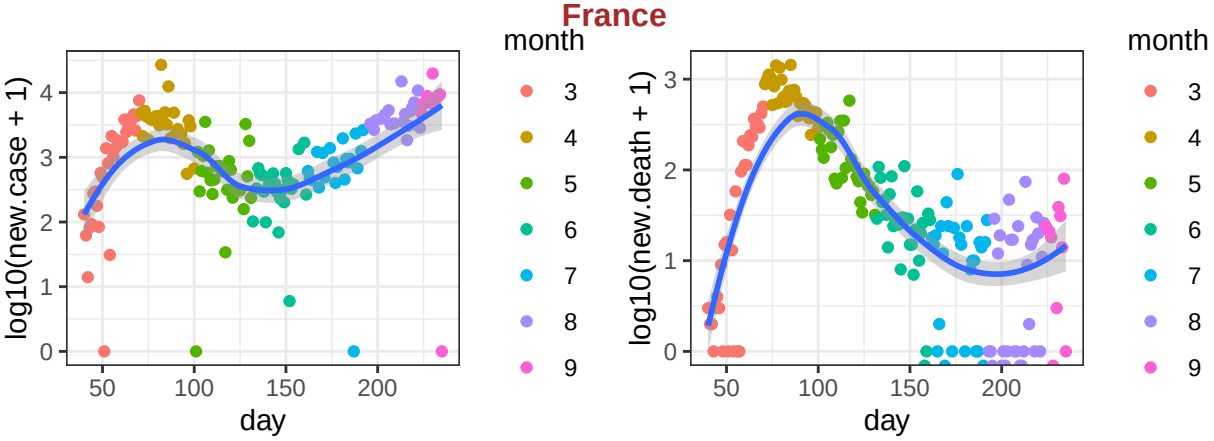
data source: https://github.com/CSSEGISandData/COVID-19

Italy



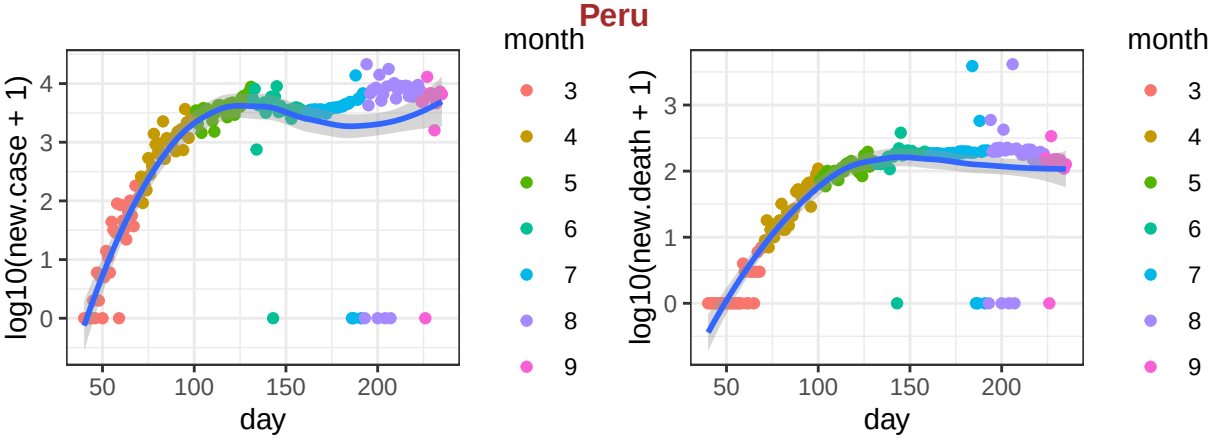
data source: https://github.com/CSSEGISandData/COVID−19

France

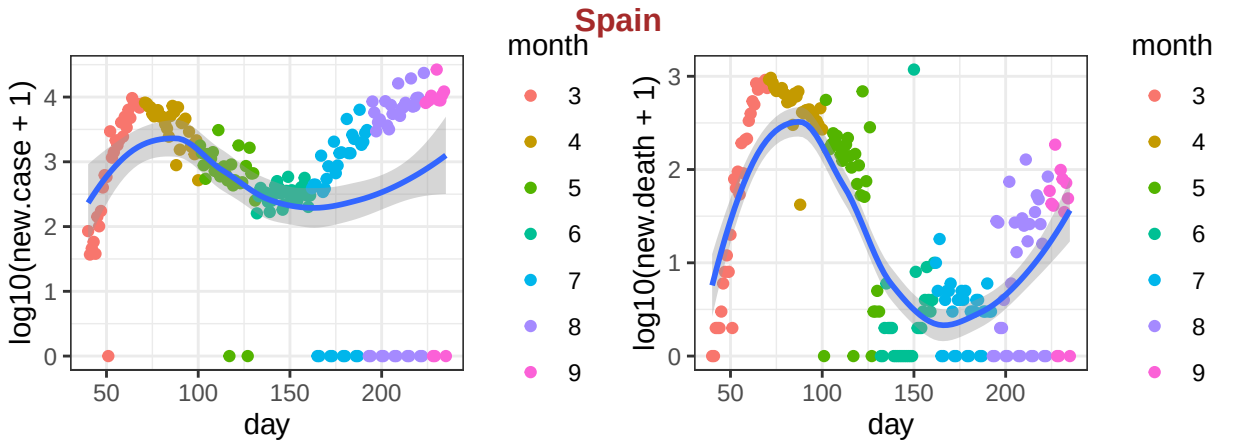


data source: https://github.com/CSSEGISandData/COVID−19

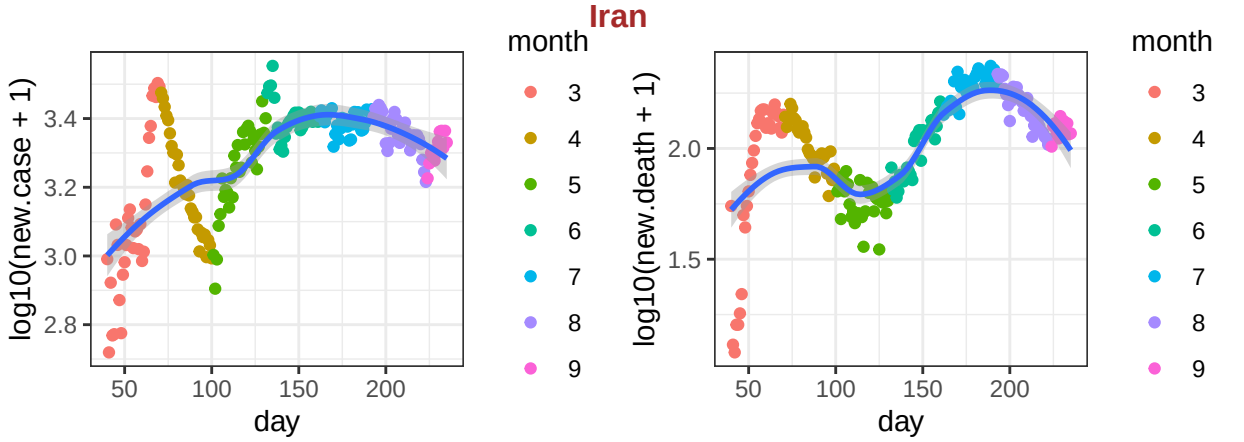
Peru



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



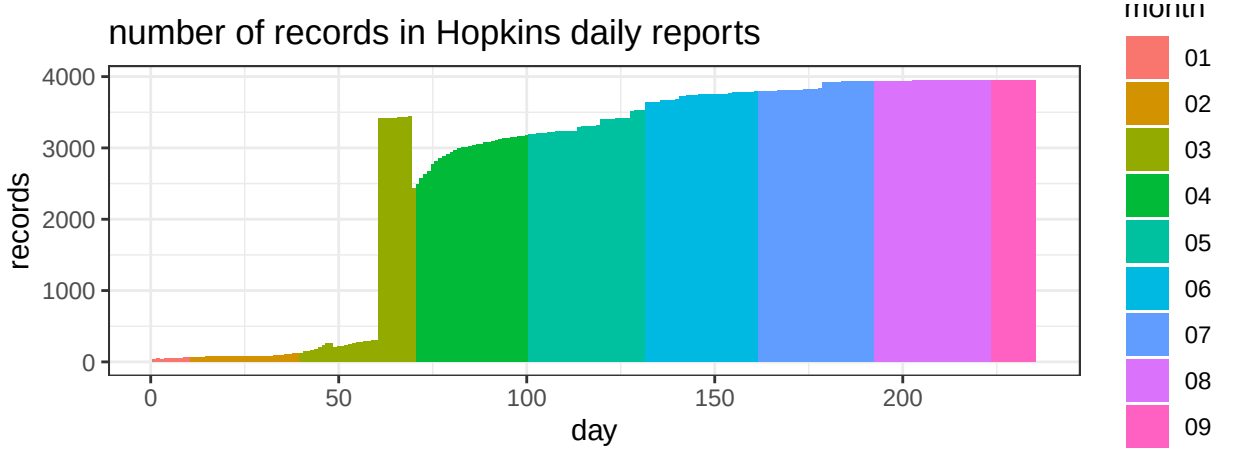
data source: https://github.com/CSSEGISandData/COVID−19



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] "2020-09-12"
```

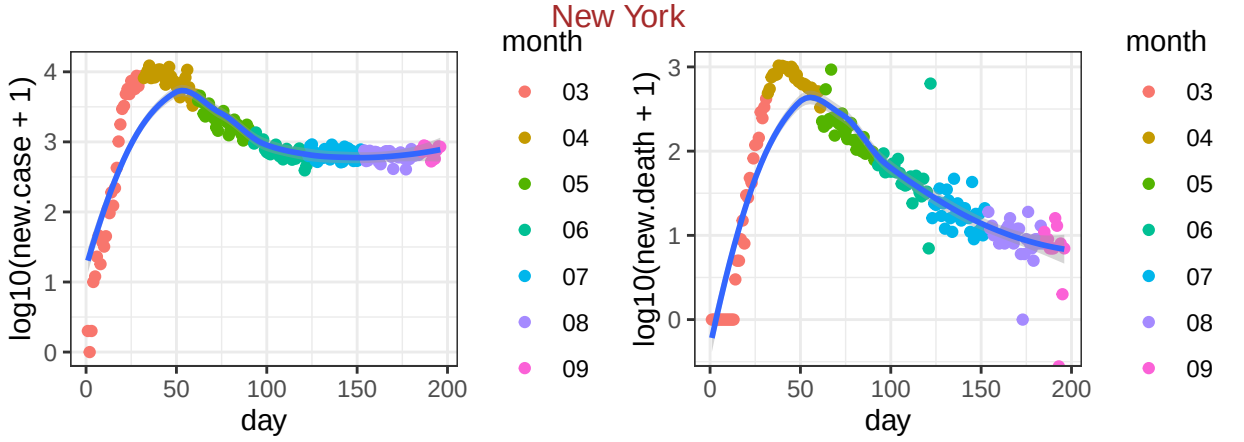
state level data

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

```
##             date          state fips  cases deaths
## 10663 2020-09-12       New York   36 448347  32625
## 10661 2020-09-12     New Jersey   34 198126  16027
## 10676 2020-09-12          Texas   48 685921  14384
## 10634 2020-09-12     California    6 760581  14333
## 10639 2020-09-12        Florida   12 661563  12599
## 10652 2020-09-12  Massachusetts   25 124540   9196
## 10644 2020-09-12       Illinois   17 262957   8546
## 10670 2020-09-12   Pennsylvania   42 148635   7915
## 10653 2020-09-12       Michigan   26 123048   6912
## 10640 2020-09-12        Georgia   13 276319   6144
## 10632 2020-09-12        Arizona    4 208128   5316
## 10649 2020-09-12      Louisiana   22 157109   5202
## 10636 2020-09-12    Connecticut    9  54326   4480
## 10667 2020-09-12           Ohio   39 136568   4411
## 10651 2020-09-12       Maryland   24 115876   3836
## 10645 2020-09-12        Indiana   18 106777   3437
## 10664 2020-09-12 North Carolina   37 184305   3073
## 10673 2020-09-12 South Carolina   45 129978   3040
## 10680 2020-09-12       Virginia   51 132940   2722
## 10655 2020-09-12    Mississippi   28  89620   2685
## 10630 2020-09-12        Alabama    1 137646   2350
## 10681 2020-09-12     Washington   53  82958   2080
## 10675 2020-09-12      Tennessee   47 168552   2040
## 10635 2020-09-12       Colorado    8  61311   1994
## 10654 2020-09-12      Minnesota   27  83640   1958
## 10656 2020-09-12       Missouri   29 102900   1793
## 10659 2020-09-12         Nevada   32  73291   1453
## 10646 2020-09-12           Iowa   19  74205   1218
## 10683 2020-09-12      Wisconsin   55  93102   1218
## 10648 2020-09-12       Kentucky   21  60187   1099
```

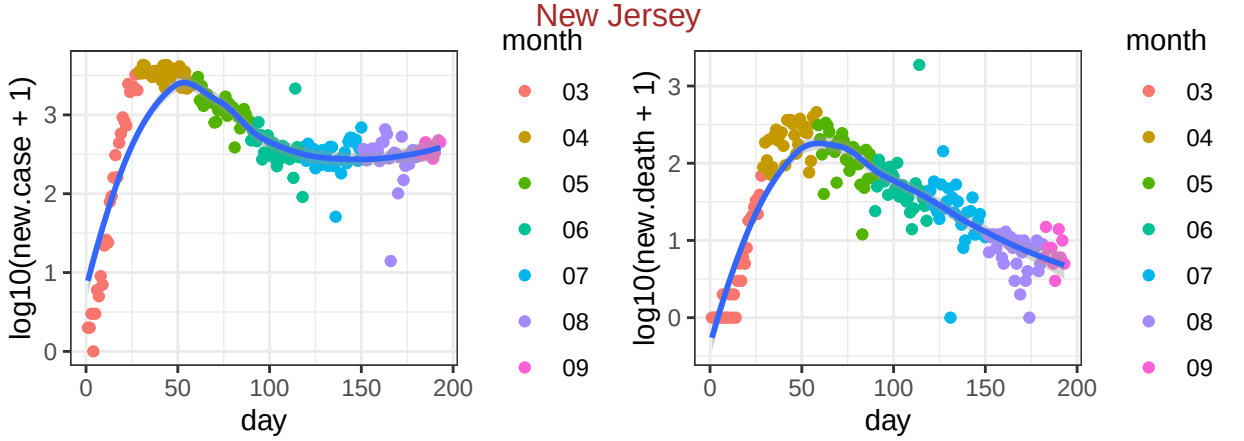
For these 30 states, I check the number of new cases and the number of new deaths. Part of the reason for such checking is to identify whether there is any similarity on such patterns. For example, could you use the pattern seen from Italy to predict what happen in an individual state, and what are the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

New York



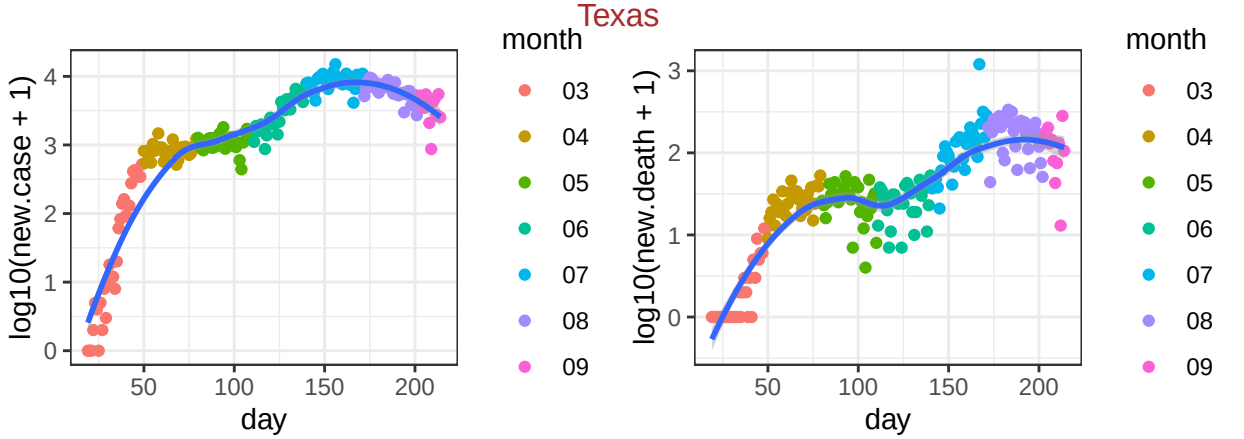
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

New Jersey



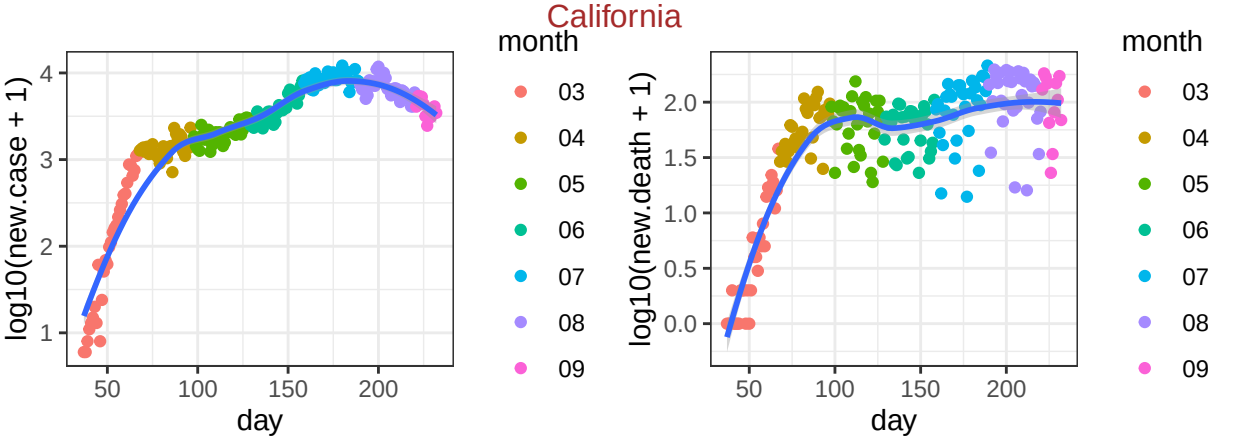
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Texas



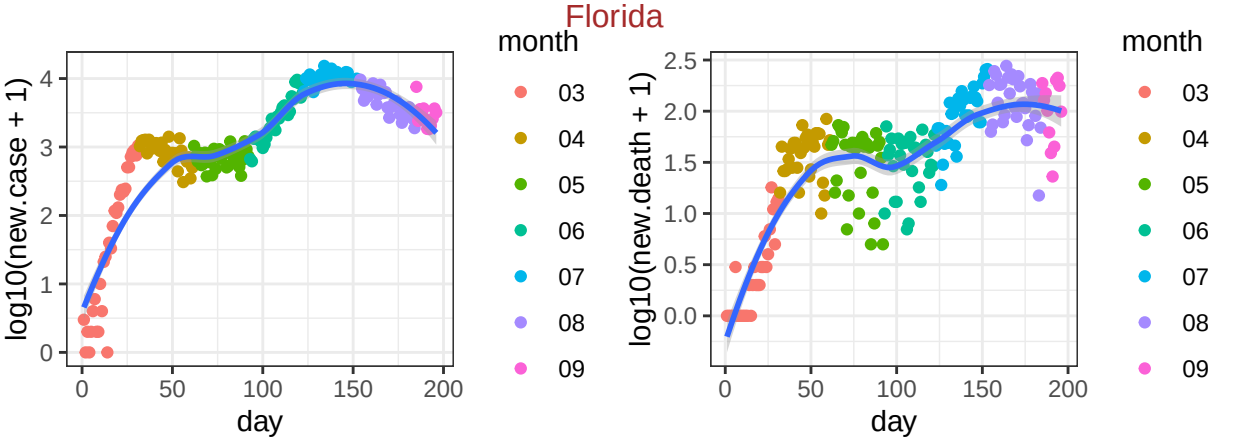
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

California



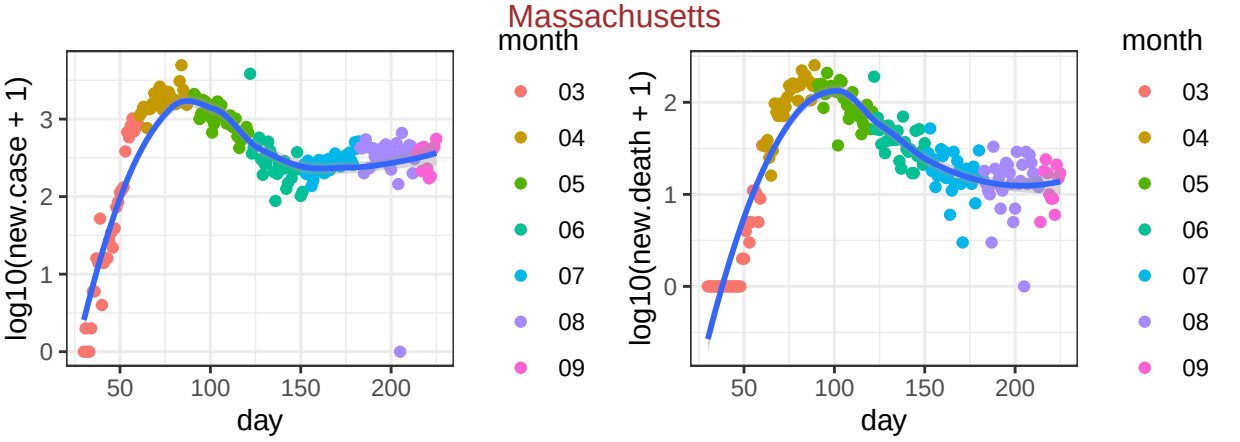
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Florida



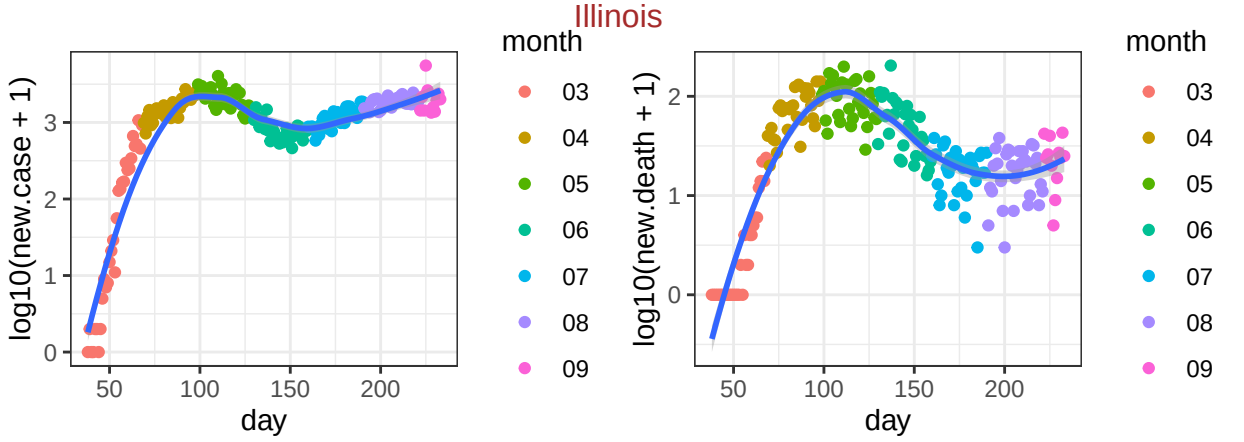
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Massachusetts



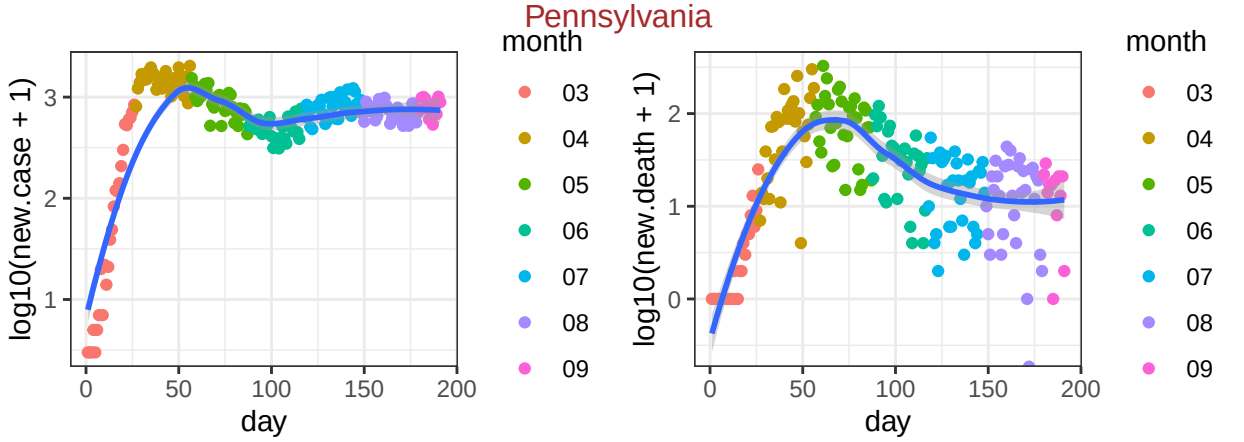
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Illinois



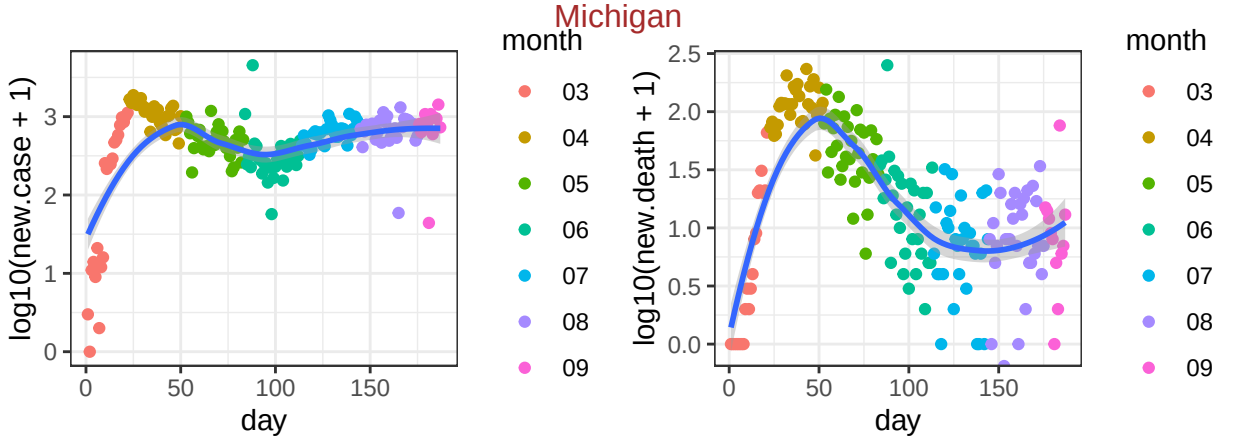
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Pennsylvania



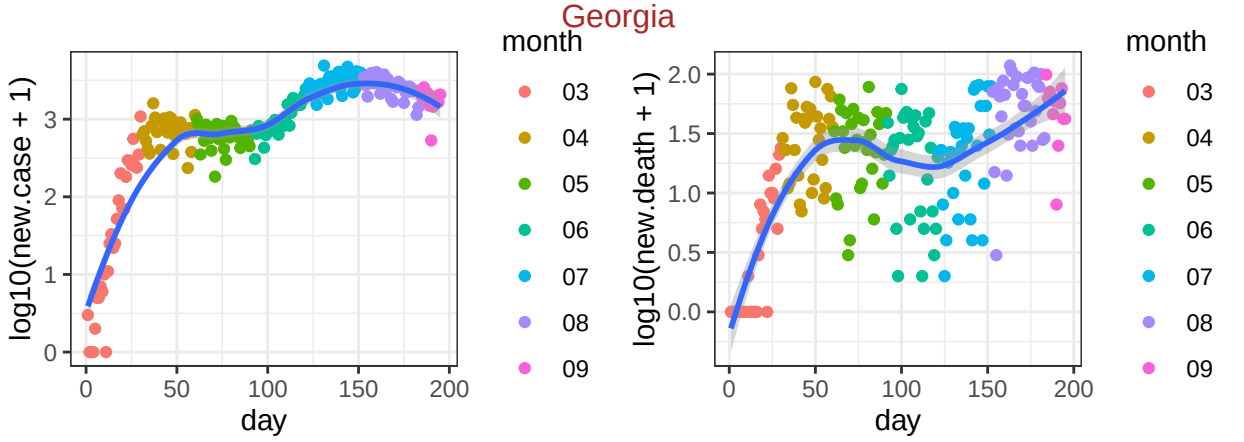
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Michigan



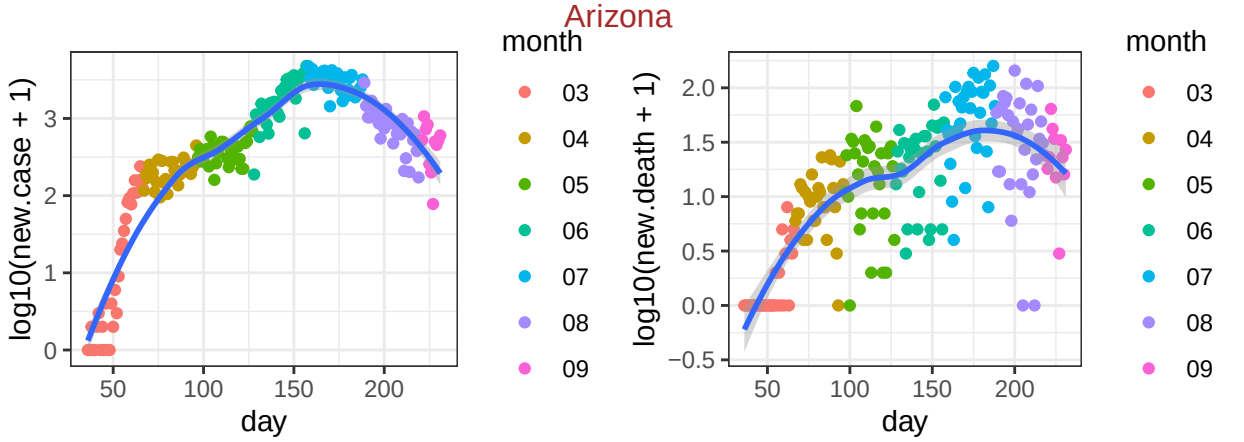
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Georgia



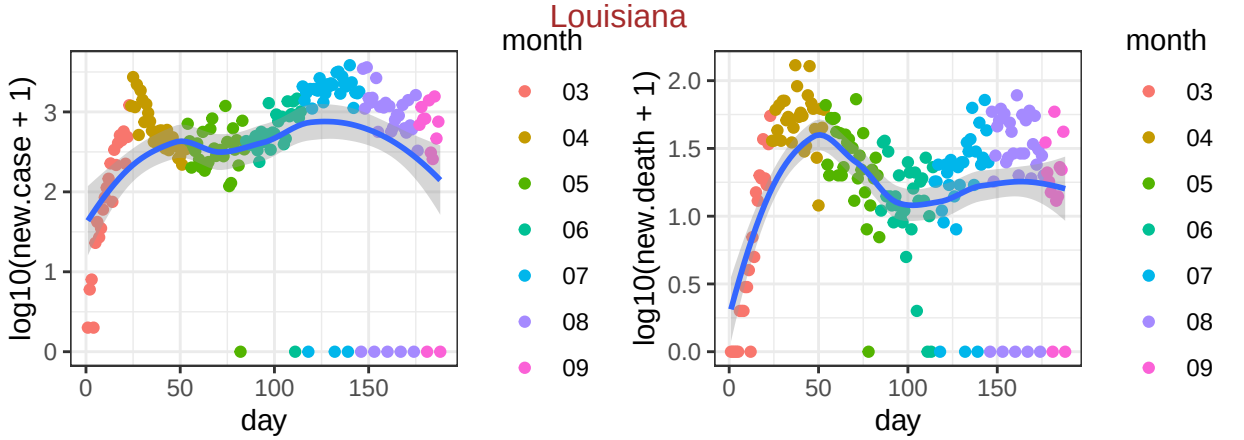
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Arizona



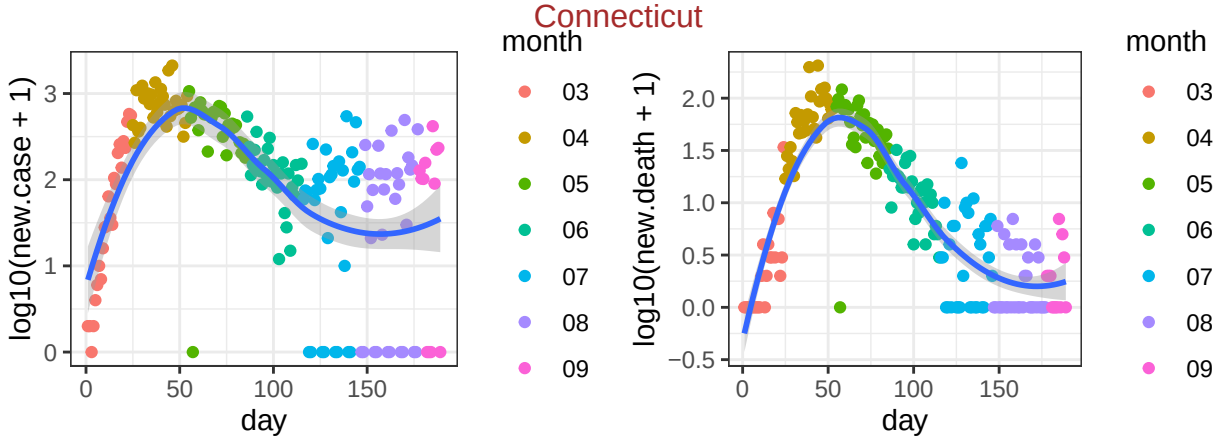
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Louisiana



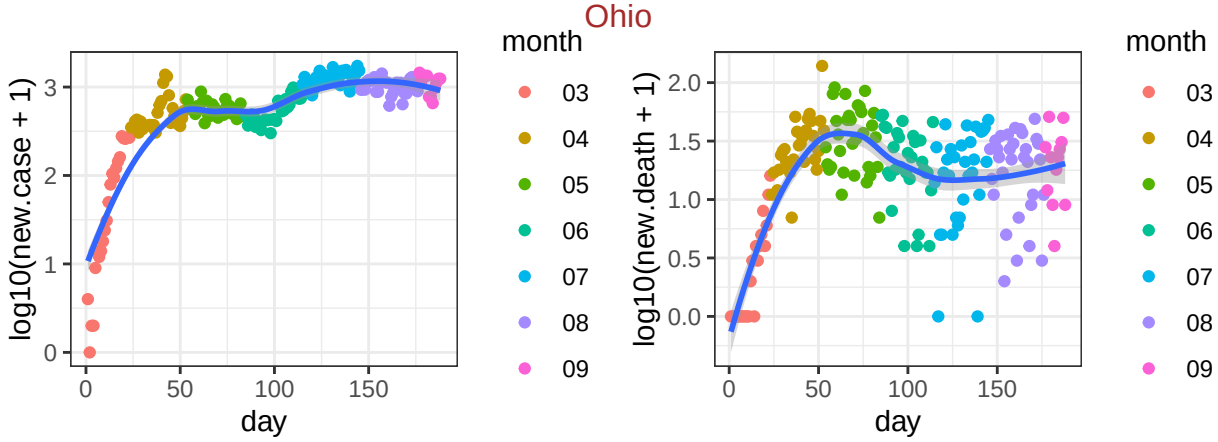
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Connecticut



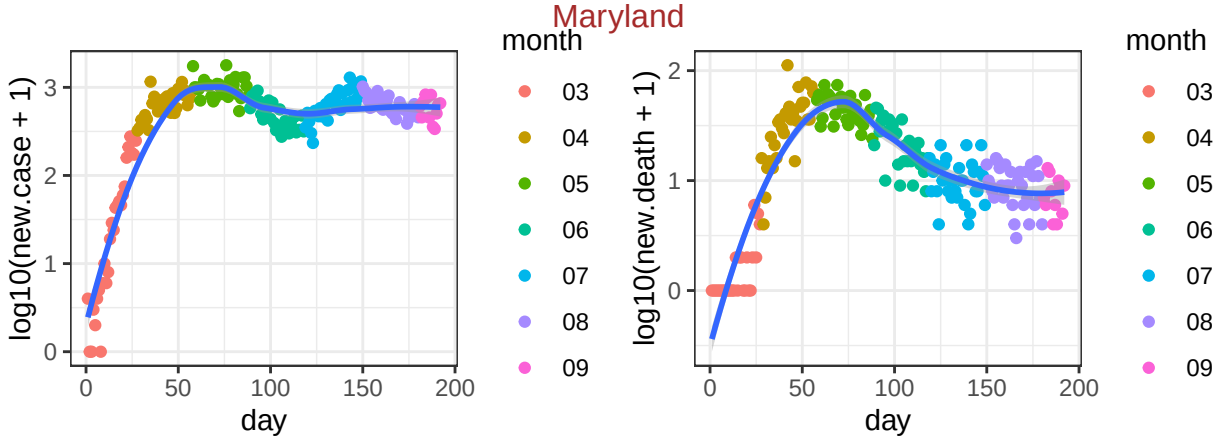
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Ohio



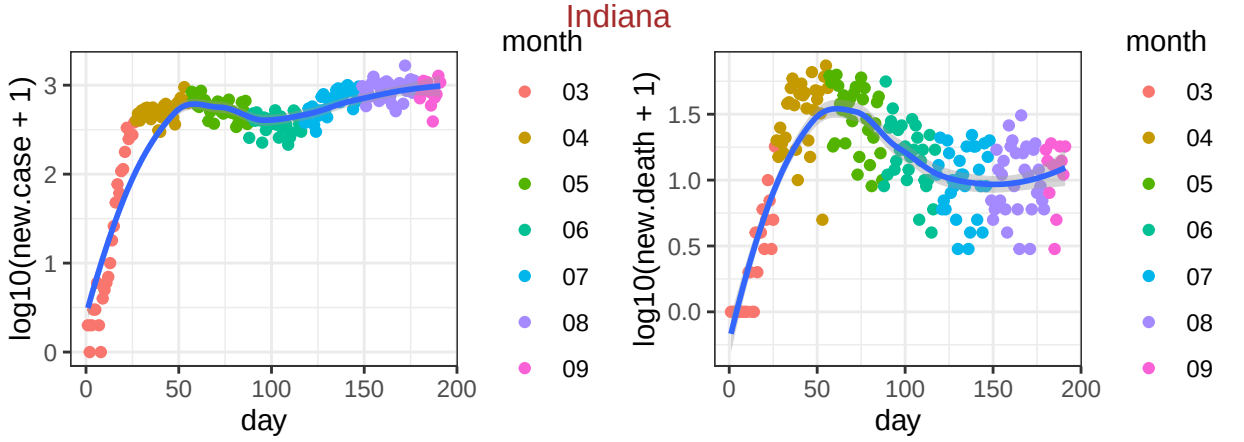
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Maryland



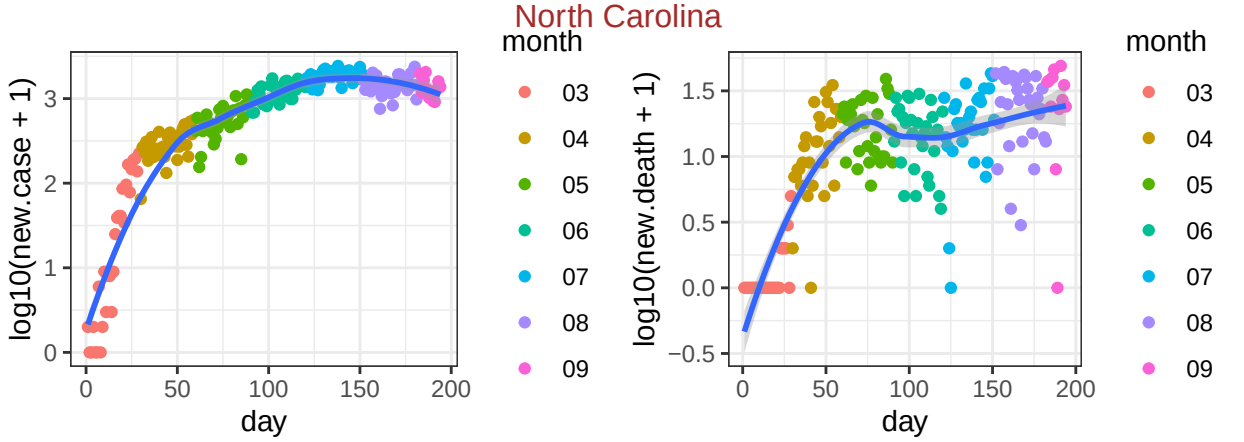
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Indiana



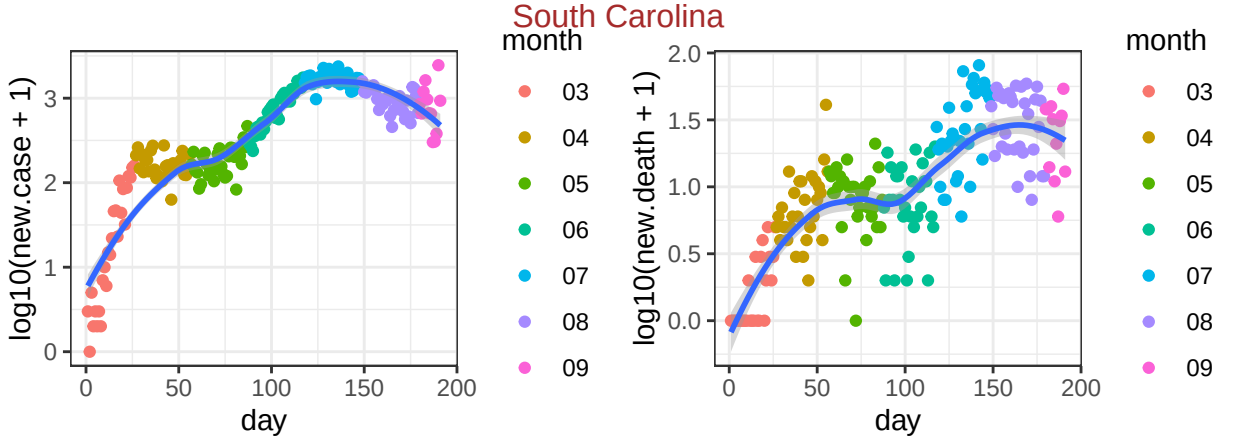
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

North Carolina



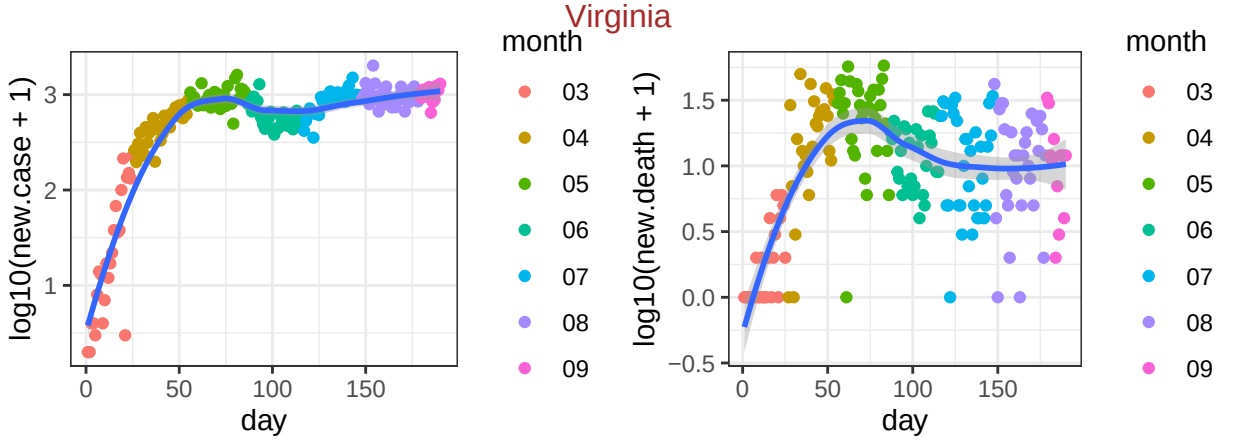
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−03

South Carolina



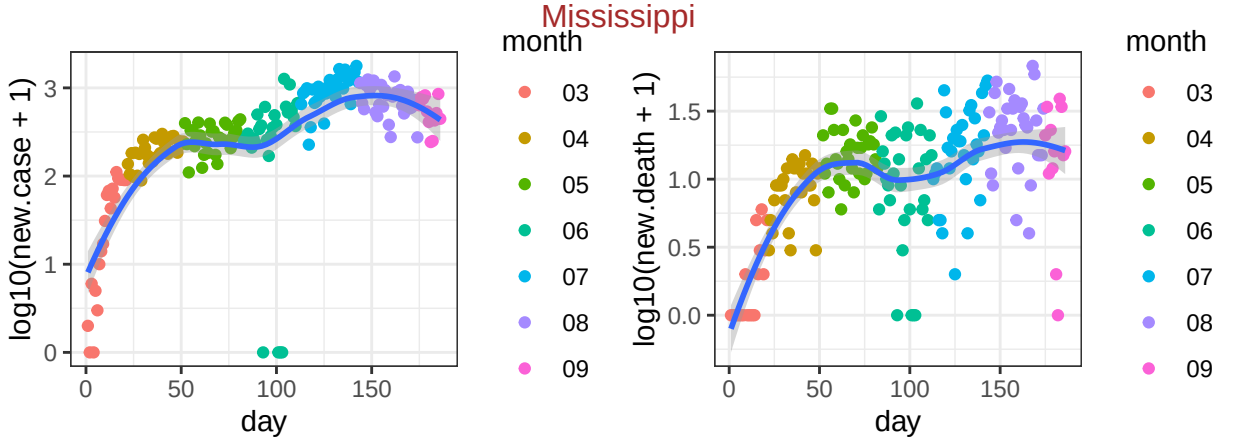
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Virginia



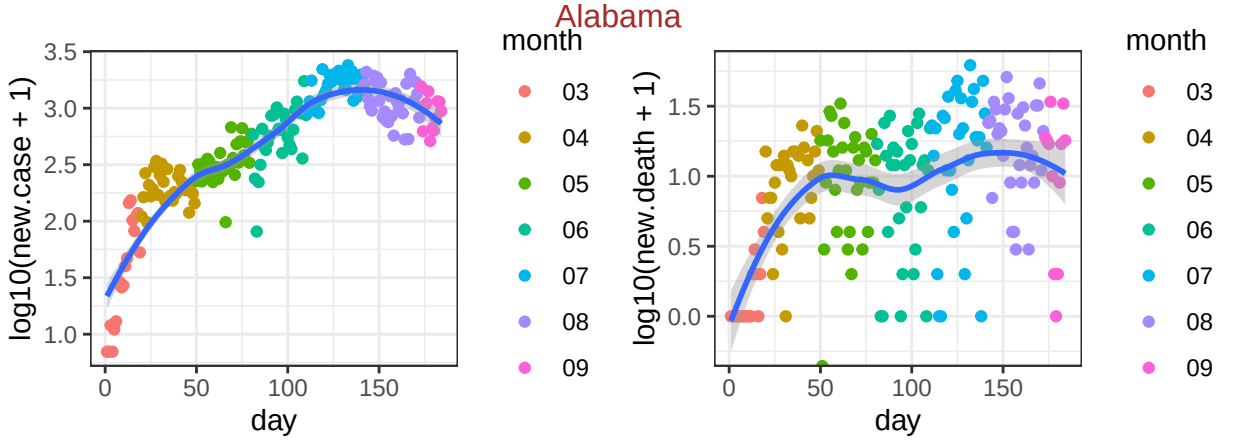
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Mississippi



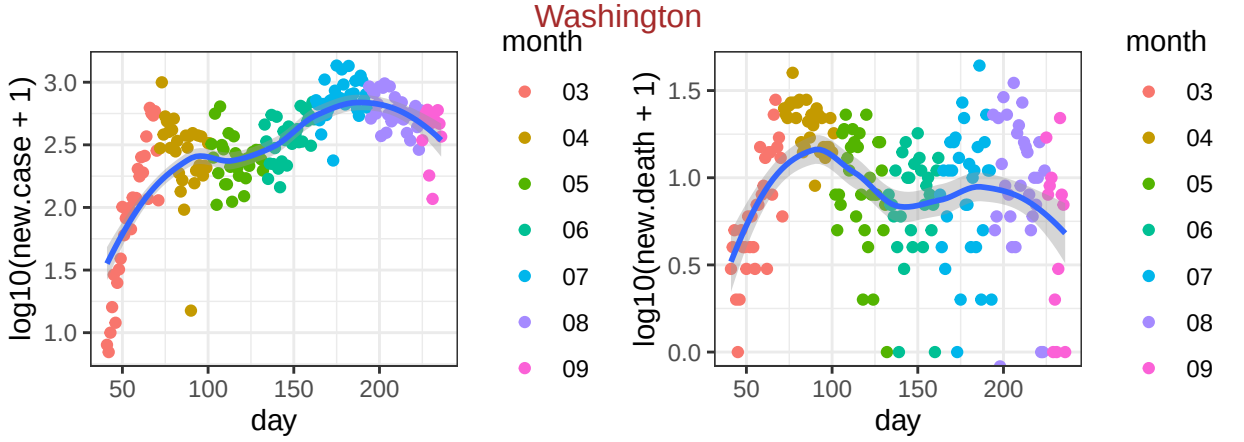
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Alabama



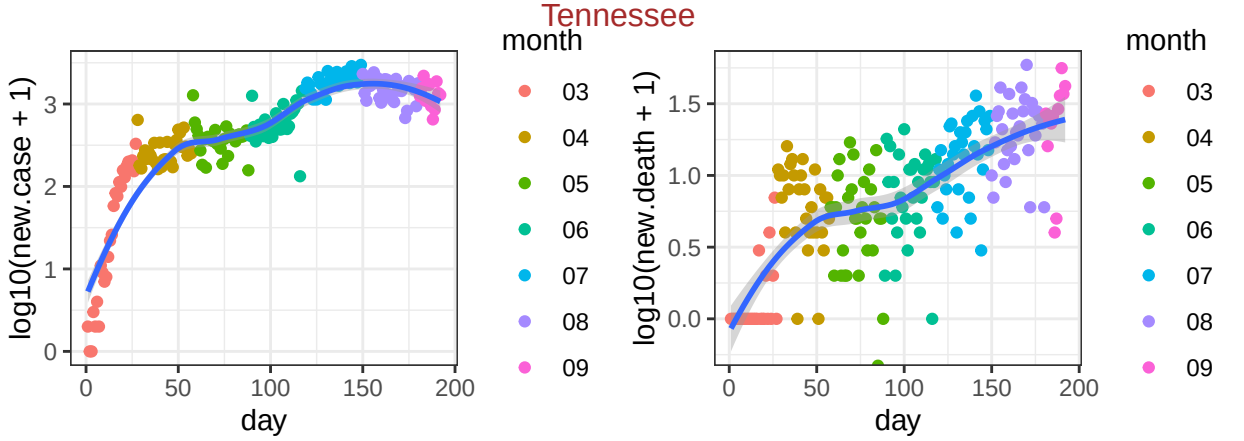
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Washington



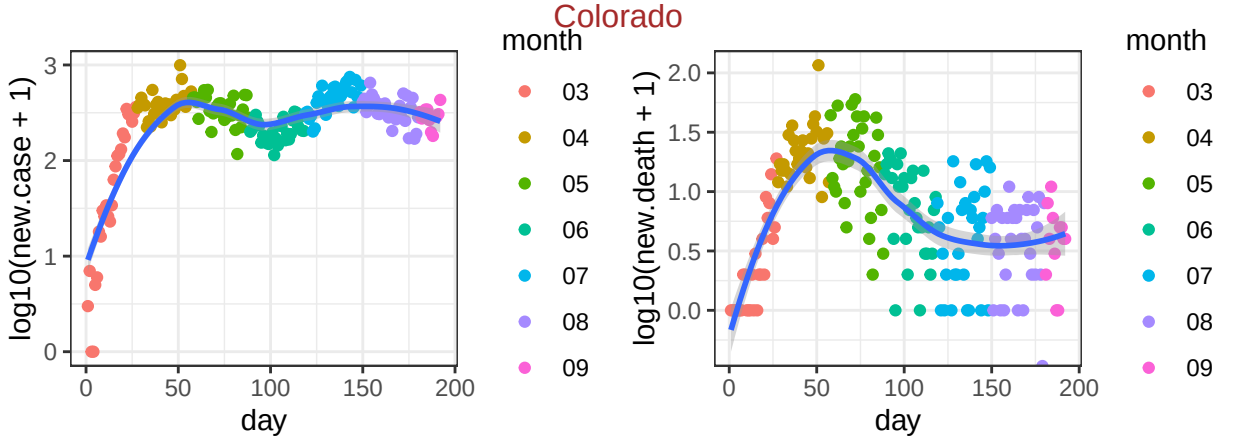
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Tennessee



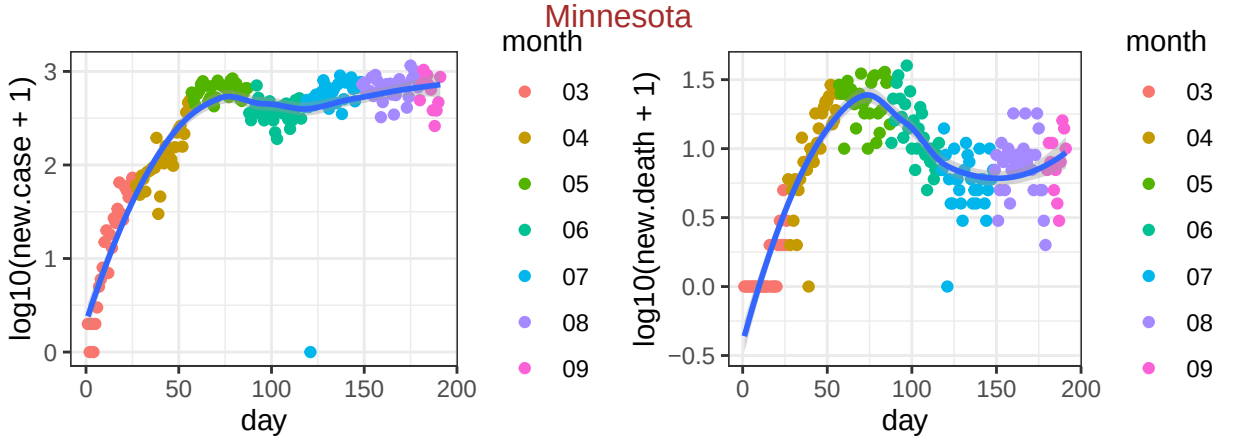
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Colorado



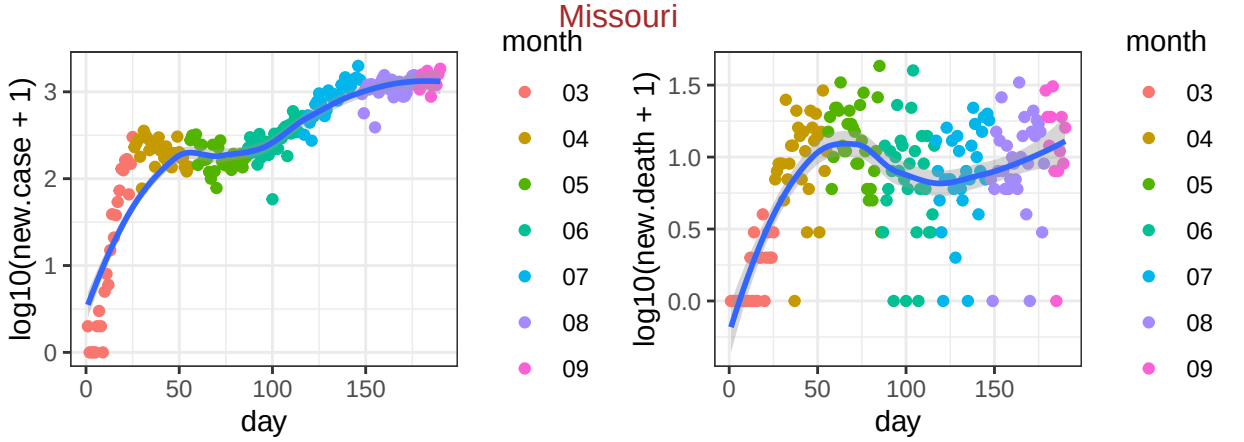
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Minnesota



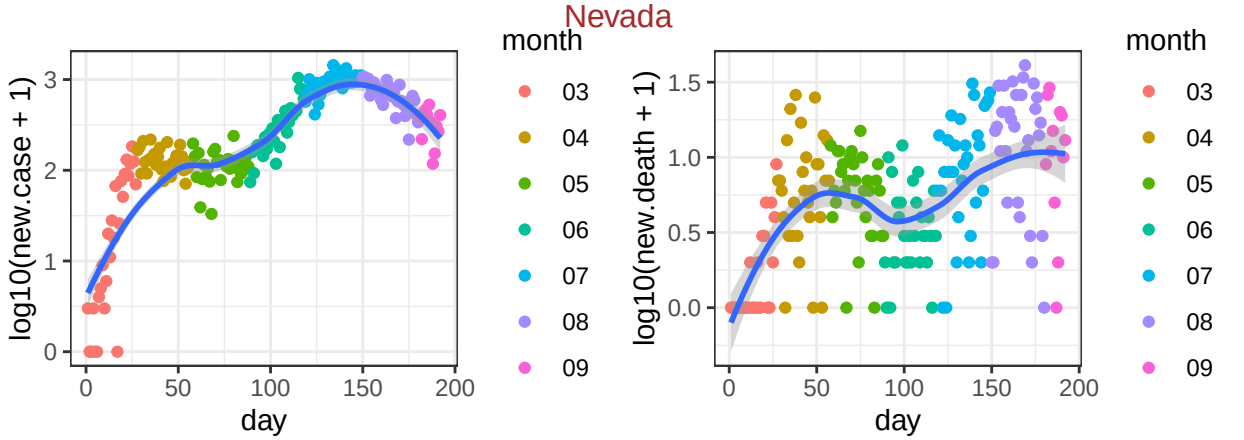
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Missouri



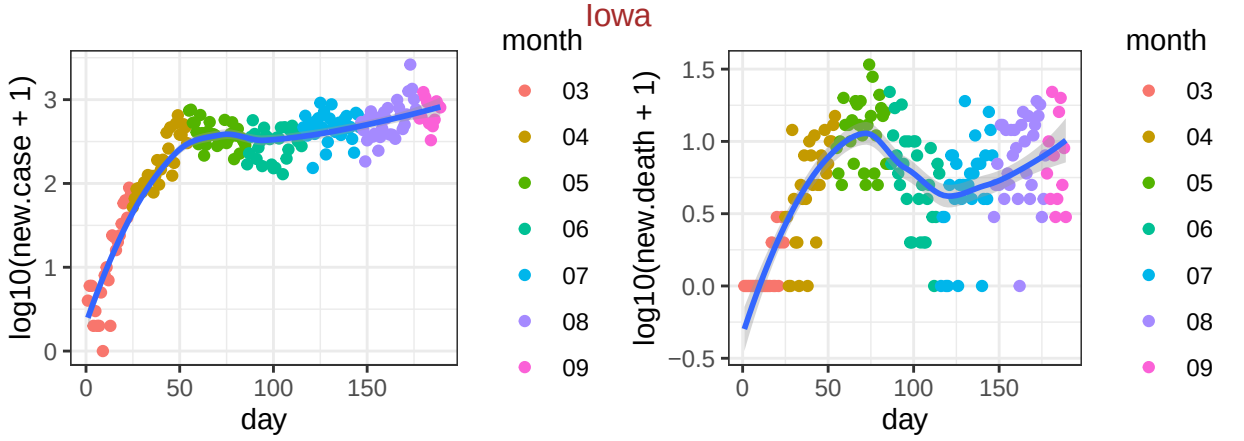
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Nevada



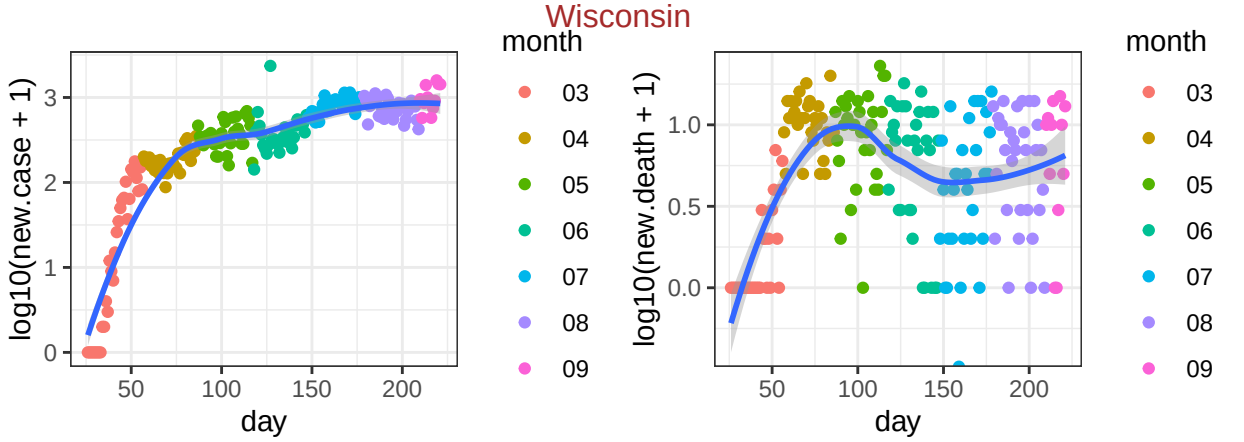
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Iowa



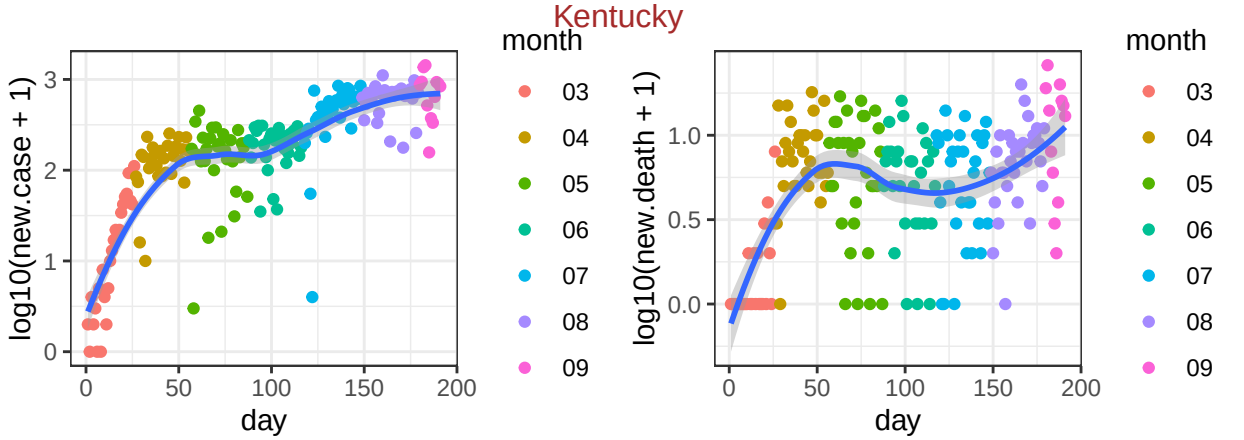
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Wisconsin



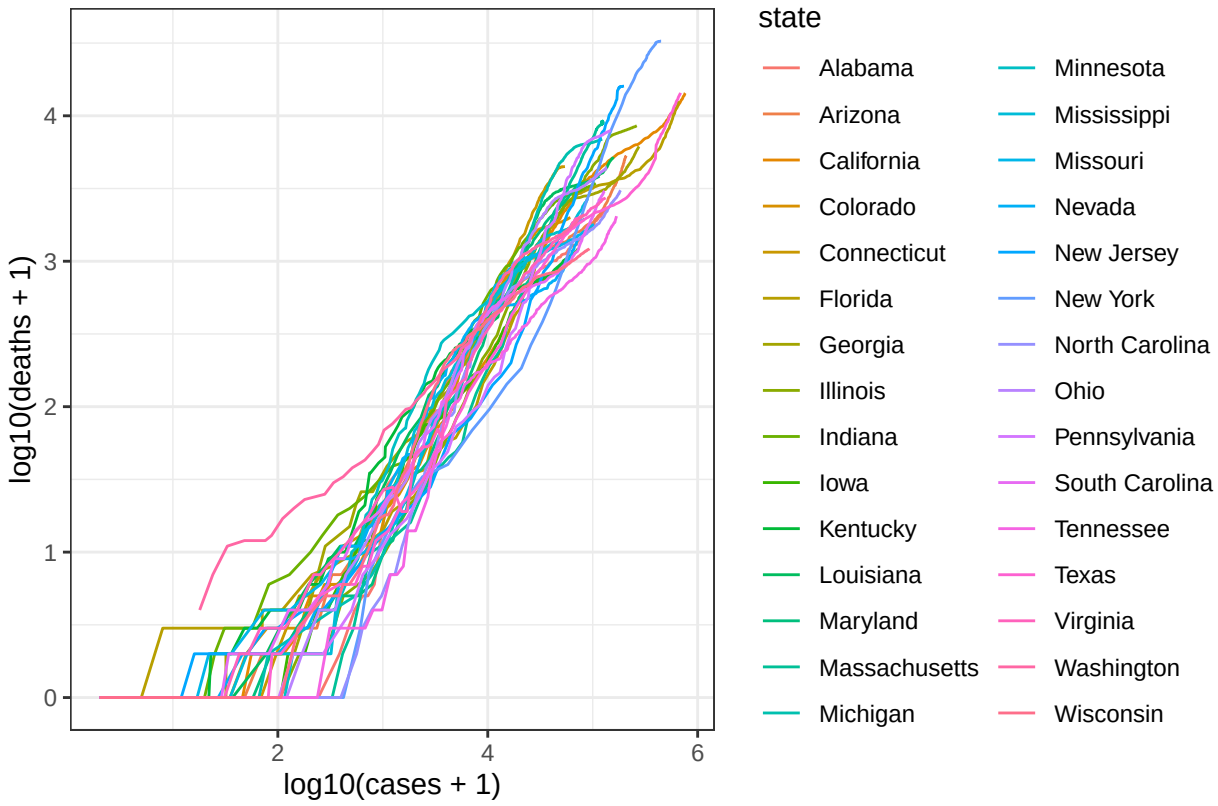
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

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

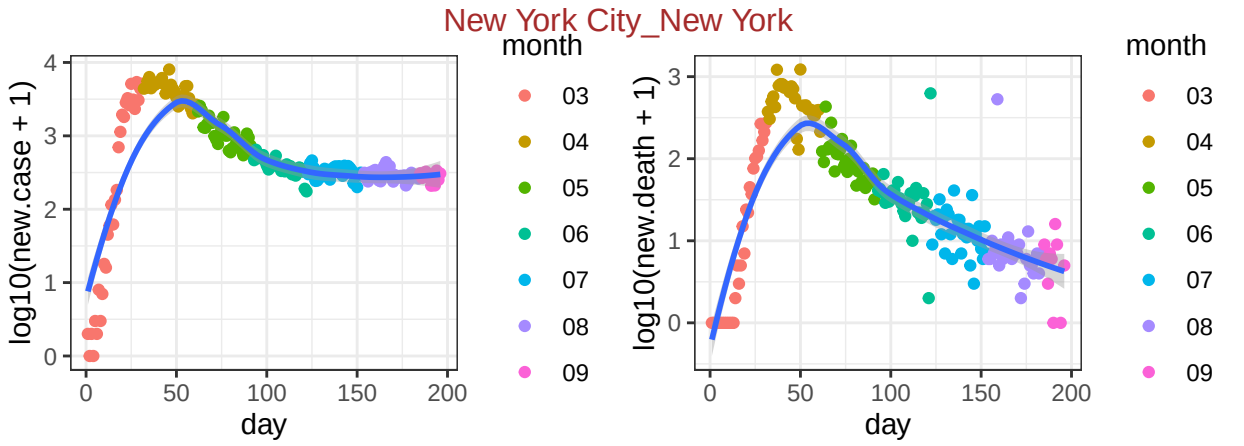
county level data

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

```
##              date          county         state  fips  cases deaths
## 526636 2020-09-12   New York City      New York    NA 242242  23743
## 524974 2020-09-12     Los Angeles    California  6037 253176   6197
## 525383 2020-09-12            Cook      Illinois 17031 134352   5128
## 524872 2020-09-12        Maricopa       Arizona  4013 137292   3158
## 526092 2020-09-12          Wayne      Michigan 26163  33276   2943
## 525133 2020-09-12      Miami-Dade       Florida 12086 163789   2882
## 527483 2020-09-12          Harris         Texas 48201 115149   2414
## 526635 2020-09-12          Nassau      New York 36059  45633   2200
## 526559 2020-09-12           Essex    New Jersey 34013  20862   2123
## 526003 2020-09-12       Middlesex Massachusetts 25017  26138   2102
## 526554 2020-09-12          Bergen    New Jersey 34003  22178   2041
## 526655 2020-09-12         Suffolk      New York 36103  45615   2008
## 527074 2020-09-12    Philadelphia  Pennsylvania 42101  35094   1784
## 526561 2020-09-12          Hudson    New Jersey 34017  20522   1512
## 526663 2020-09-12     Westchester      New York 36119  37485   1452
## 525078 2020-09-12        Hartford   Connecticut  9003  13886   1430
## 526564 2020-09-12       Middlesex    New Jersey 34023  19016   1422
## 525077 2020-09-12       Fairfield   Connecticut  9001  19360   1418
## 527490 2020-09-12         Hidalgo         Texas 48215  29335   1381
## 526572 2020-09-12           Union    New Jersey 34039  17465   1354
## 525096 2020-09-12         Broward       Florida 12011  74273   1279
## 526528 2020-09-12           Clark        Nevada 32003  62402   1259
```

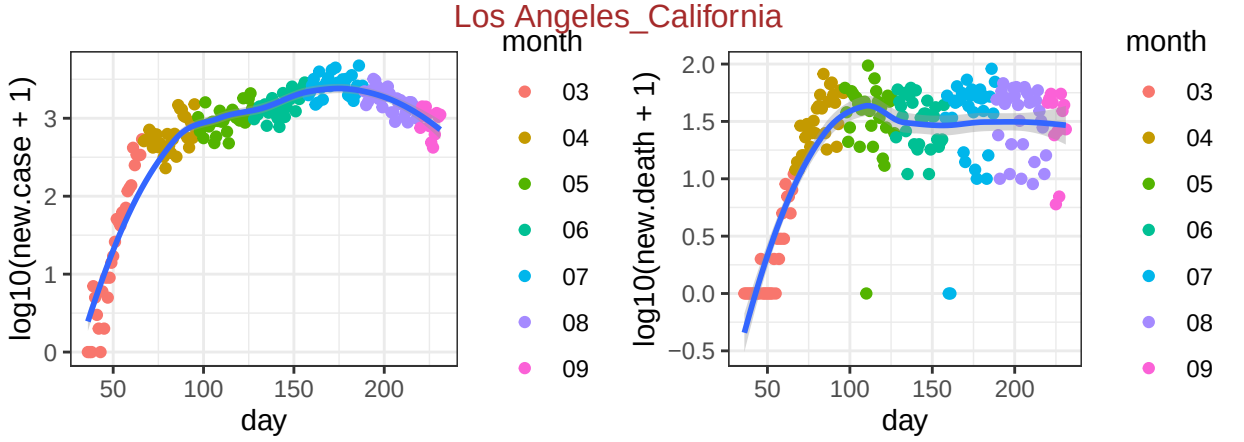
```
## 525999 2020-09-12           Essex Massachusetts 25009  18643   1258
## 526568 2020-09-12         Passaic    New Jersey 34031  18783   1249
## 527398 2020-09-12           Bexar         Texas 48029  48210   1200
## 525140 2020-09-12      Palm Beach       Florida 12099  43871   1196
## 526072 2020-09-12         Oakland      Michigan 26125  19348   1184
## 526007 2020-09-12         Suffolk Massachusetts 25025  23130   1116
## 525081 2020-09-12       New Haven   Connecticut  9009  13961   1110
## 524988 2020-09-12       Riverside    California  6065  55073   1103
## 524985 2020-09-12          Orange    California  6059  51936   1093
## 526009 2020-09-12       Worcester Massachusetts 25027  13807   1073
## 527439 2020-09-12          Dallas         Texas 48113  78511   1043
## 526567 2020-09-12           Ocean    New Jersey 34029  11777   1036
## 526005 2020-09-12         Norfolk Massachusetts 25021   9809   1030
## 526059 2020-09-12          Macomb      Michigan 26099  13917   1007
## 526120 2020-09-12        Hennepin     Minnesota 27053  24611    903
## 527069 2020-09-12      Montgomery  Pennsylvania 42091  11578    867
## 526565 2020-09-12        Monmouth    New Jersey 34025  11250    863
## 527173 2020-09-12      Providence  Rhode Island 44007  17500    854
## 527413 2020-09-12         Cameron         Texas 48061  21983    841
## 525985 2020-09-12      Montgomery      Maryland 24031  21129    834
## 526566 2020-09-12          Morris    New Jersey 34027   7775    831
## 524991 2020-09-12  San Bernardino    California  6071  50543    814
## 525986 2020-09-12 Prince George's      Maryland 24033  27812    810
## 525519 2020-09-12          Marion       Indiana 18097  19822    805
## 527046 2020-09-12        Delaware  Pennsylvania 42045  10857    780
## 526366 2020-09-12       St. Louis      Missouri 29189  21588    768
## 526001 2020-09-12         Hampden Massachusetts 25013   7846    767
## 527832 2020-09-12            King    Washington 53033  20819    763
```

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



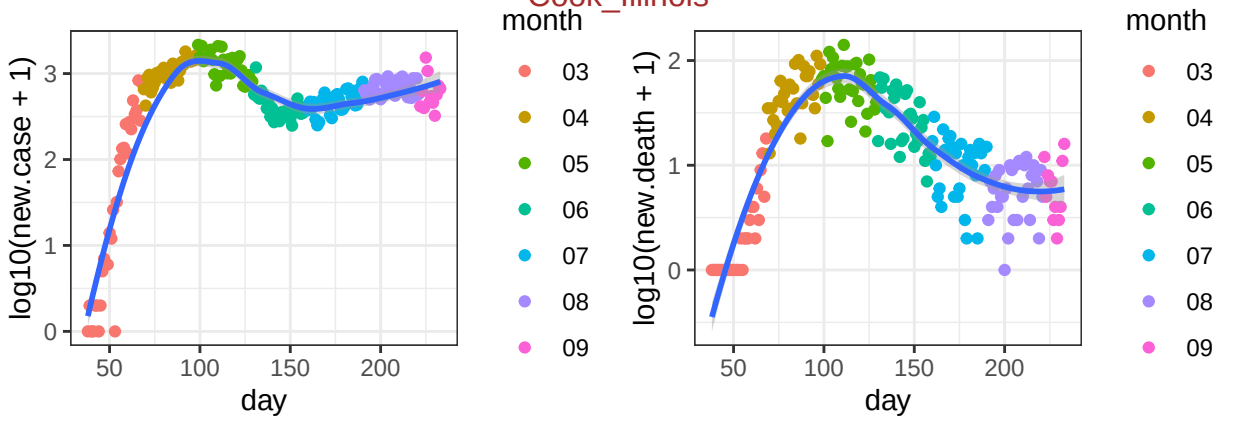
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Los Angeles_California



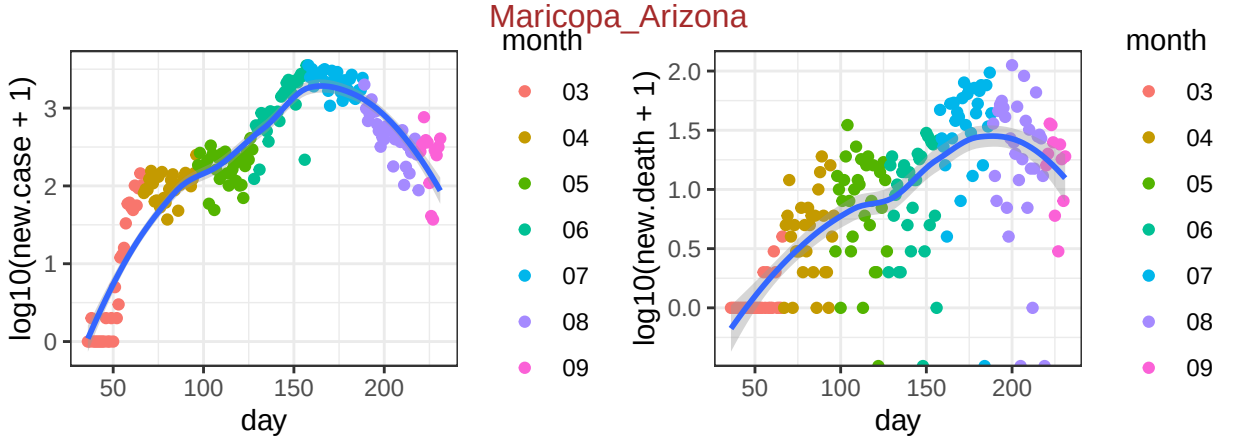
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Cook_Illinois



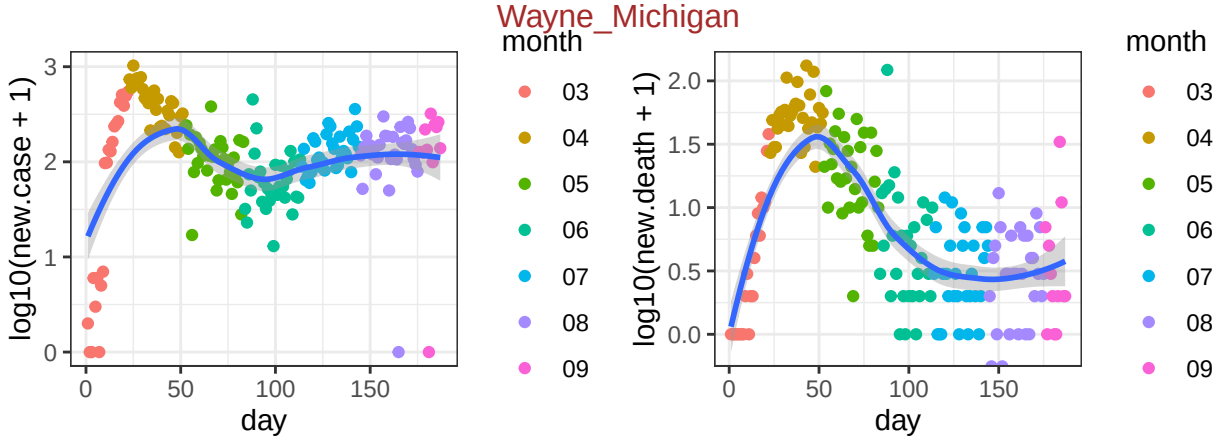
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Maricopa_Arizona



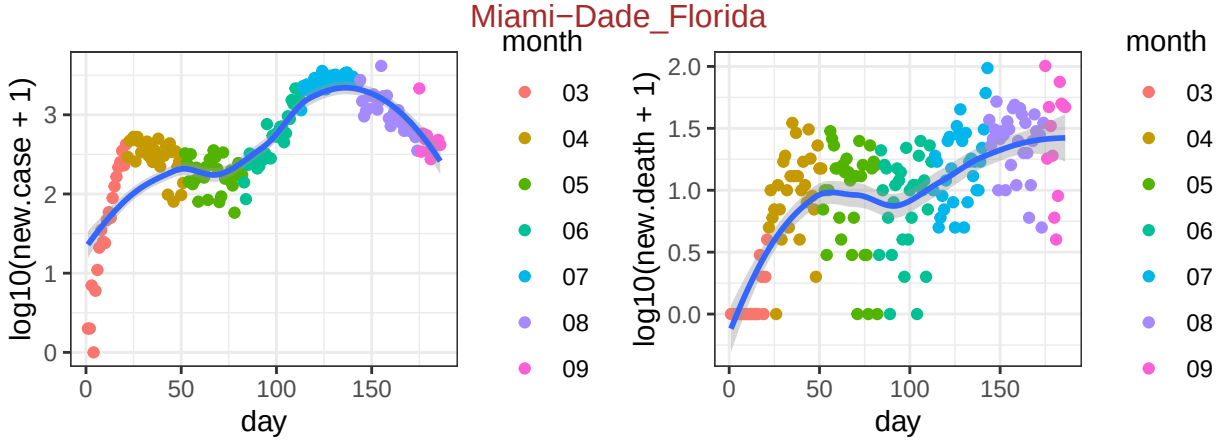
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Wayne_Michigan



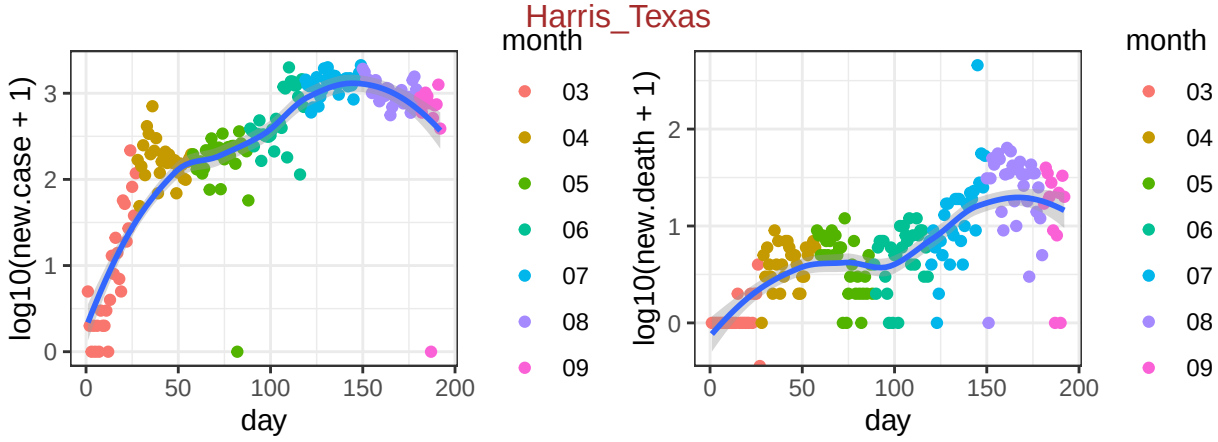
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Miami−Dade_Florida



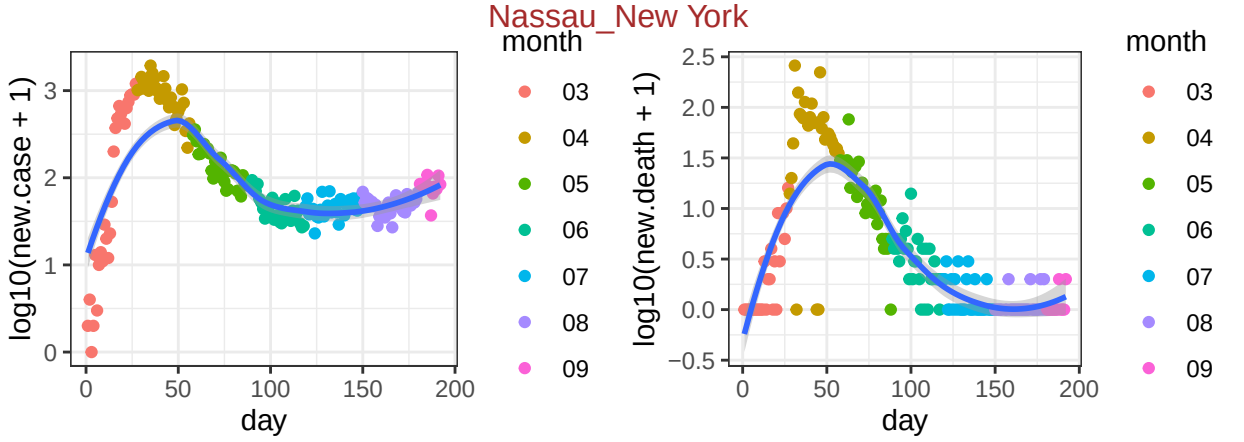
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Harris_Texas



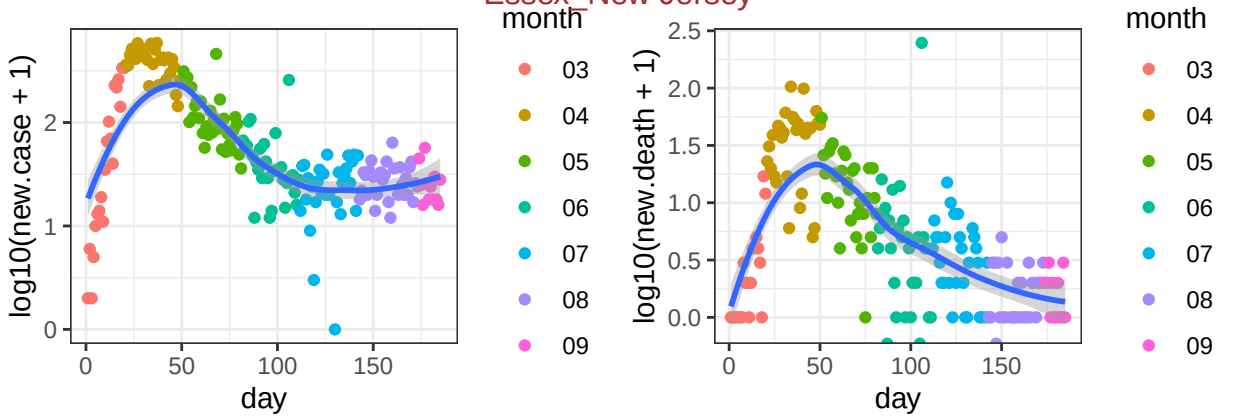
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Nassau_New York



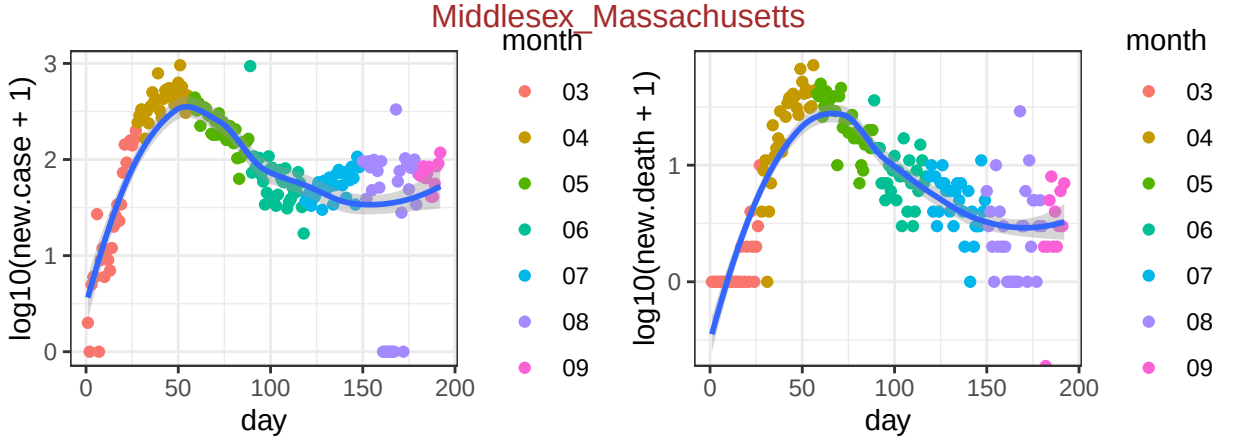
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Essex_New Jersey



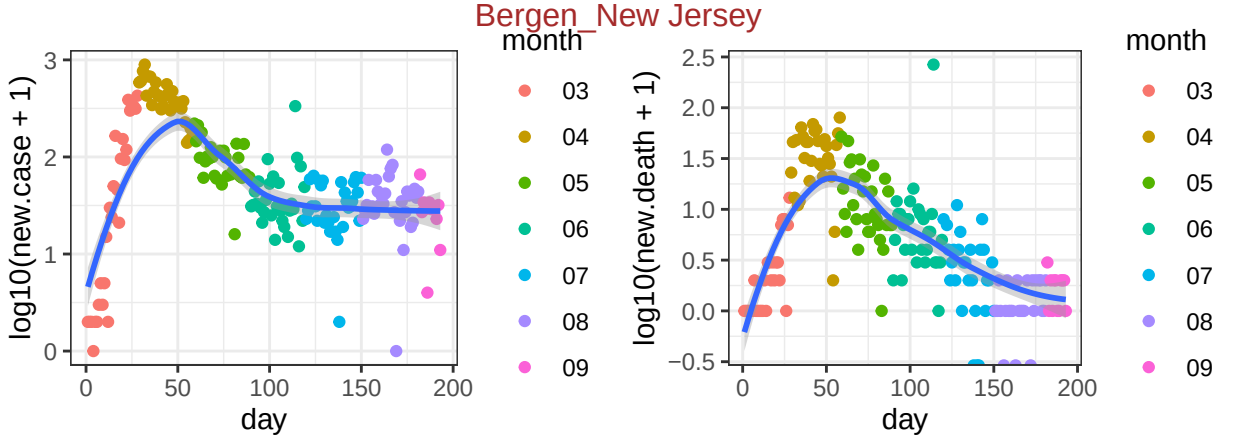
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Middlesex_Massachusetts



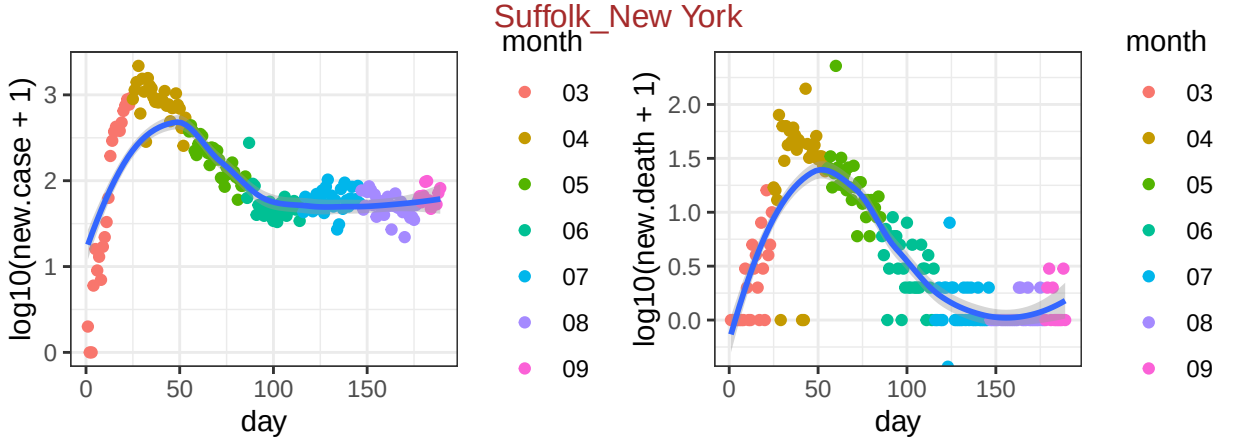
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Bergen_New Jersey



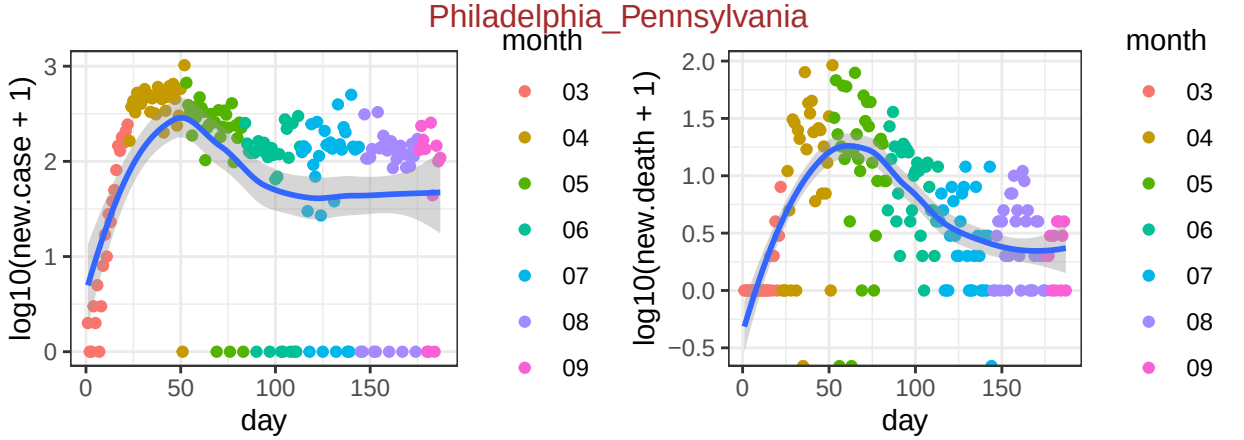
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Suffolk_New York



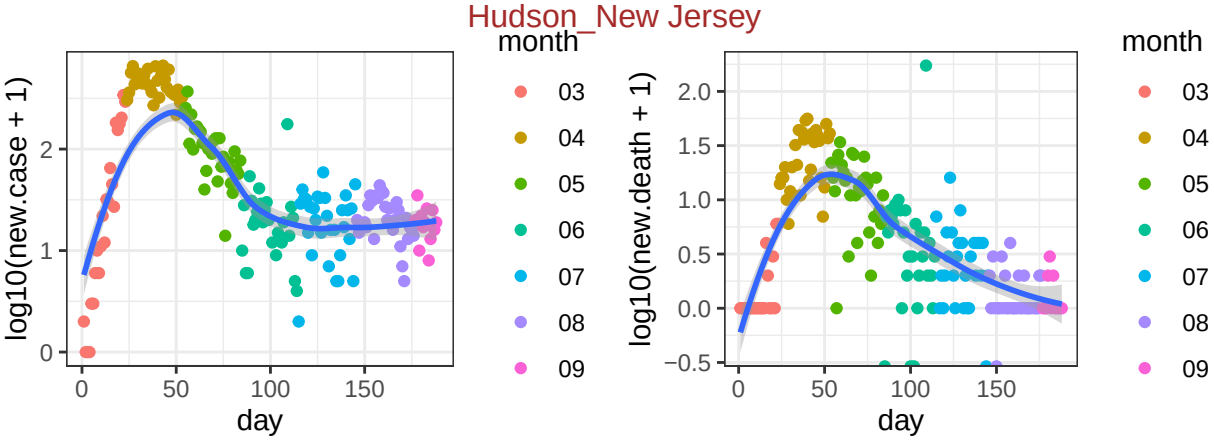
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Philadelphia_Pennsylvania



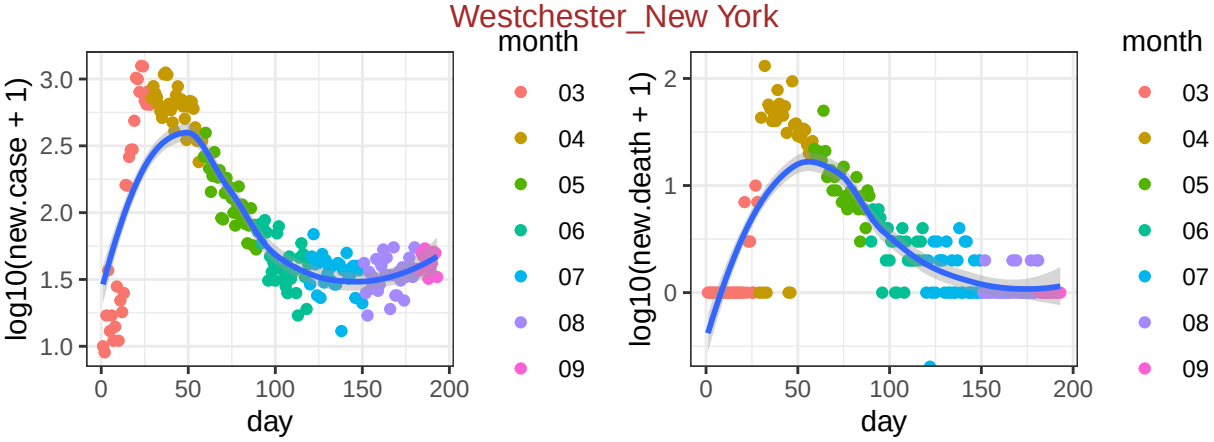
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Hudson_New Jersey



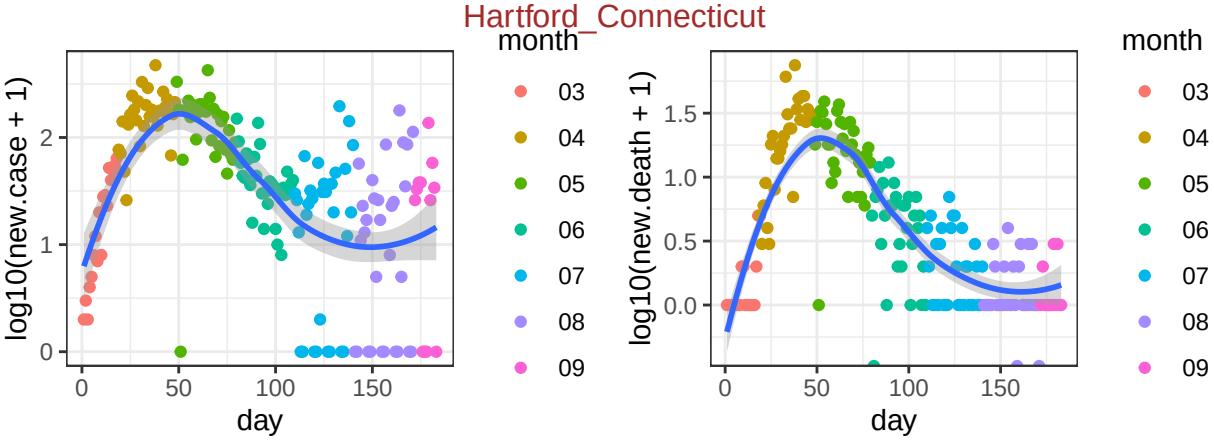
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Westchester_New York



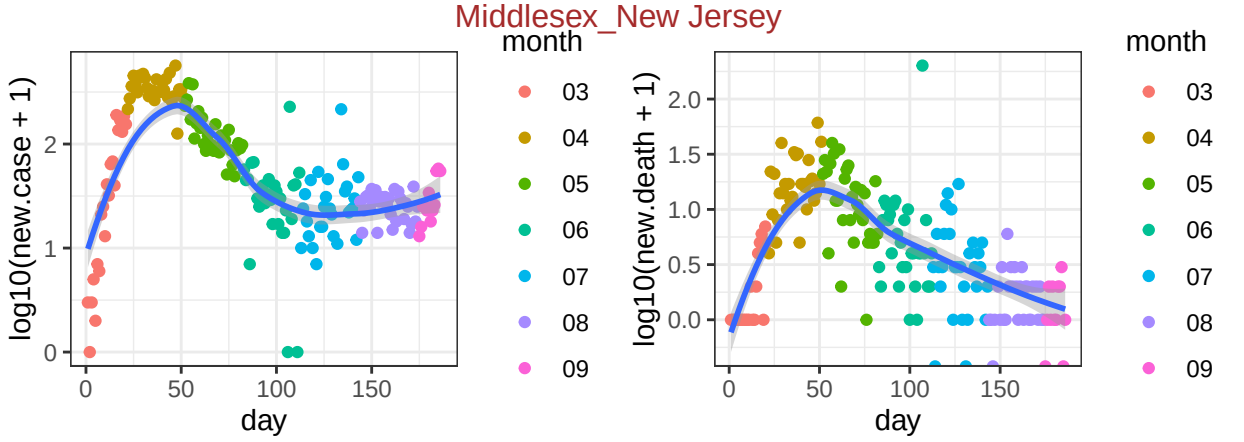
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−04

Hartford_Connecticut



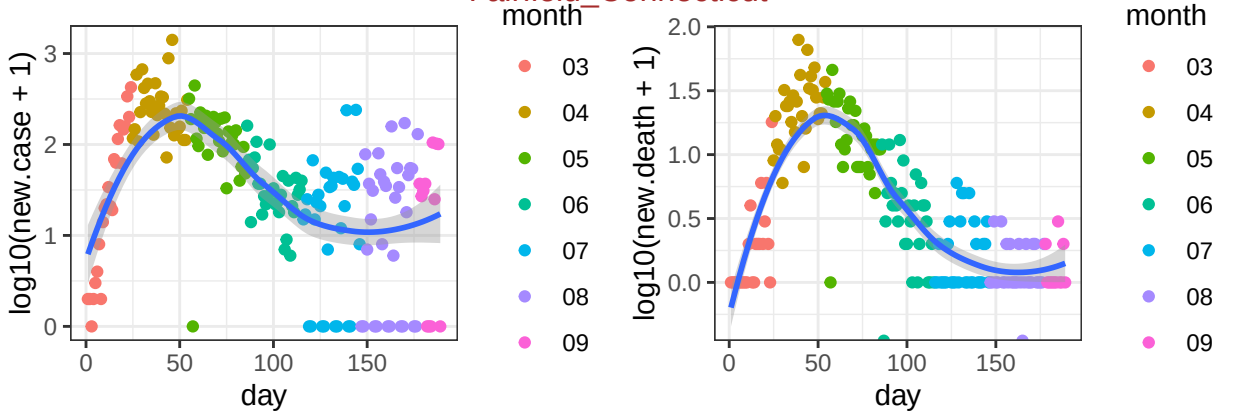
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Middlesex_New Jersey



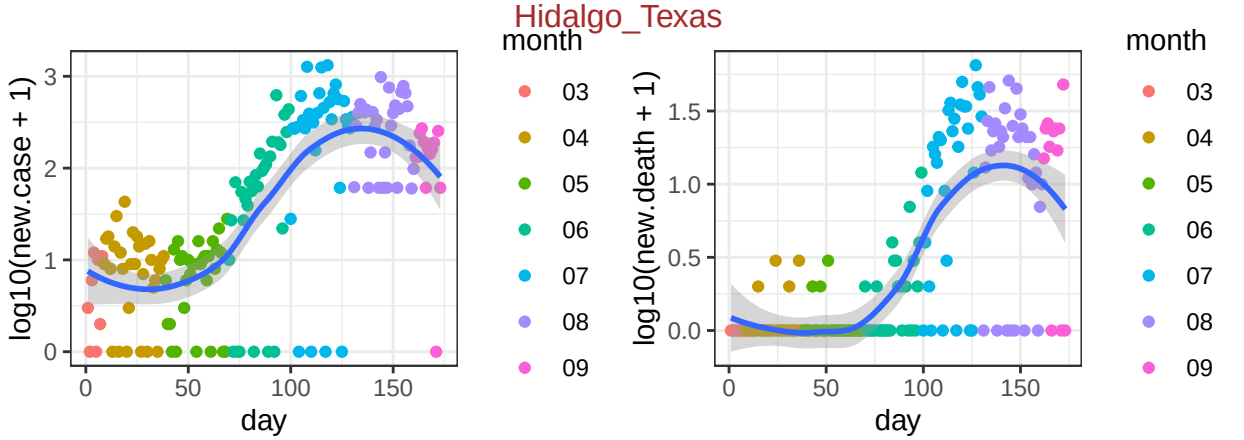
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−11

Fairfield_Connecticut



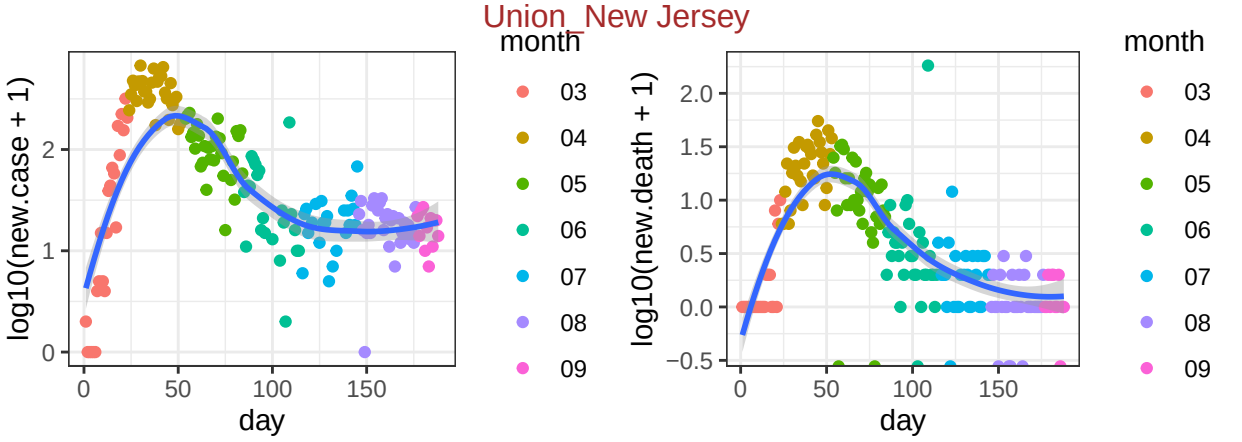
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Hidalgo_Texas



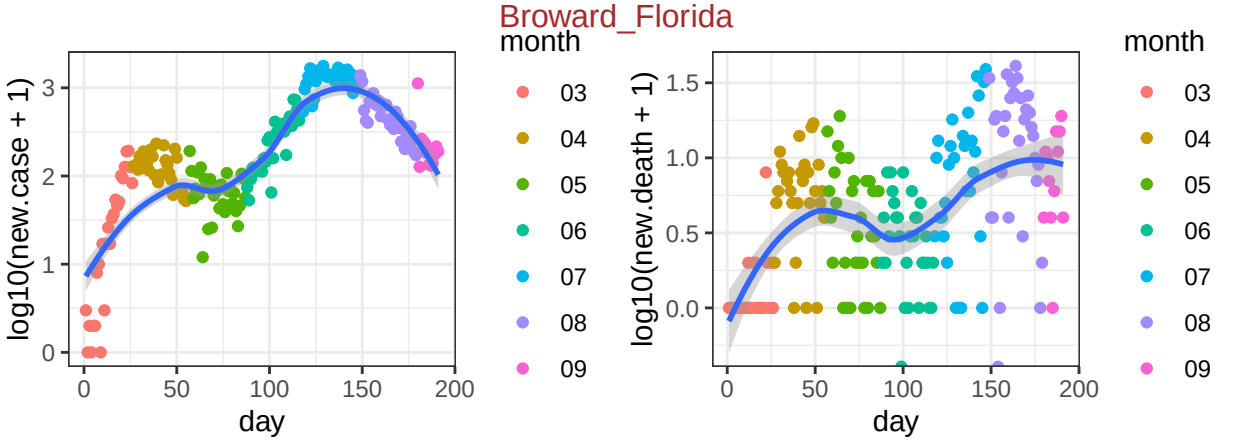
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−24

Union_New Jersey



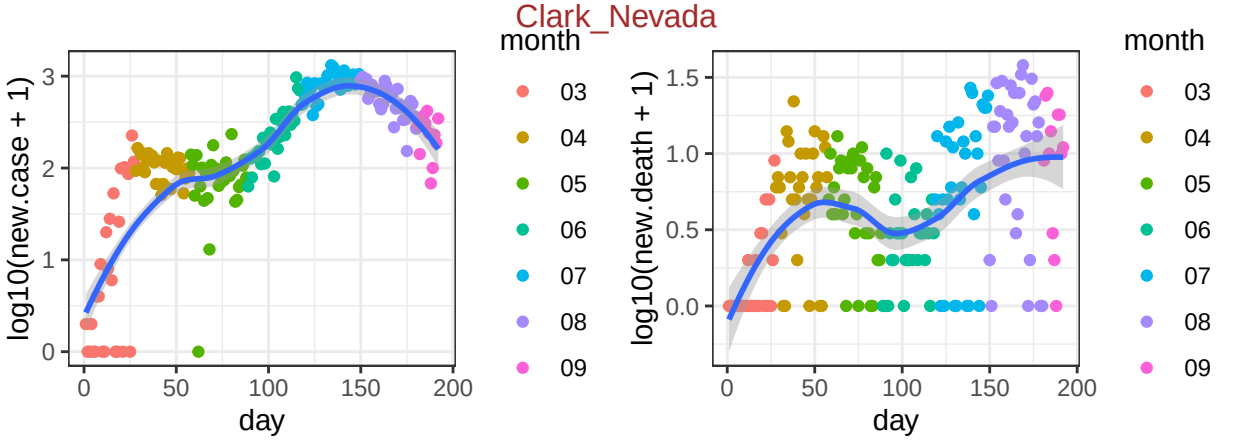
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Broward_Florida



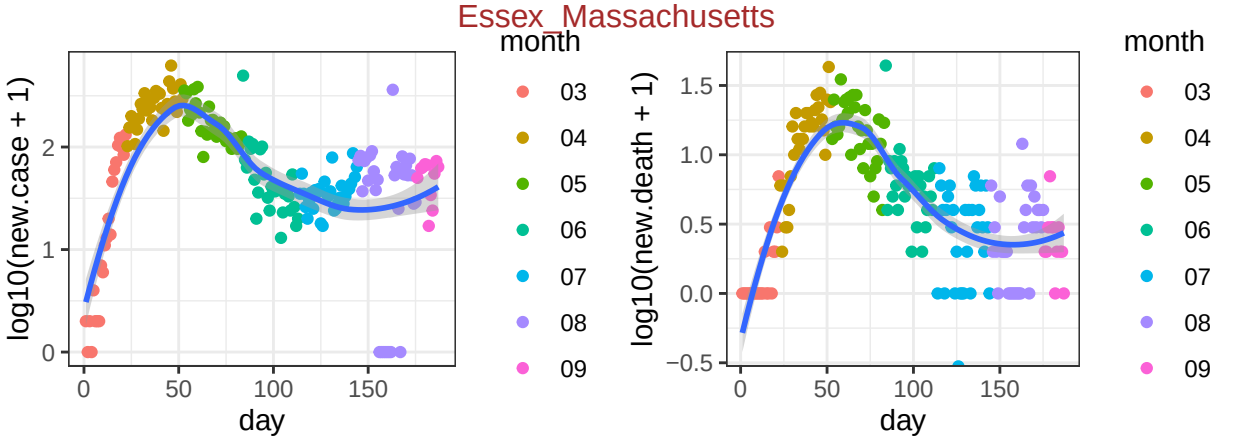
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Clark_Nevada



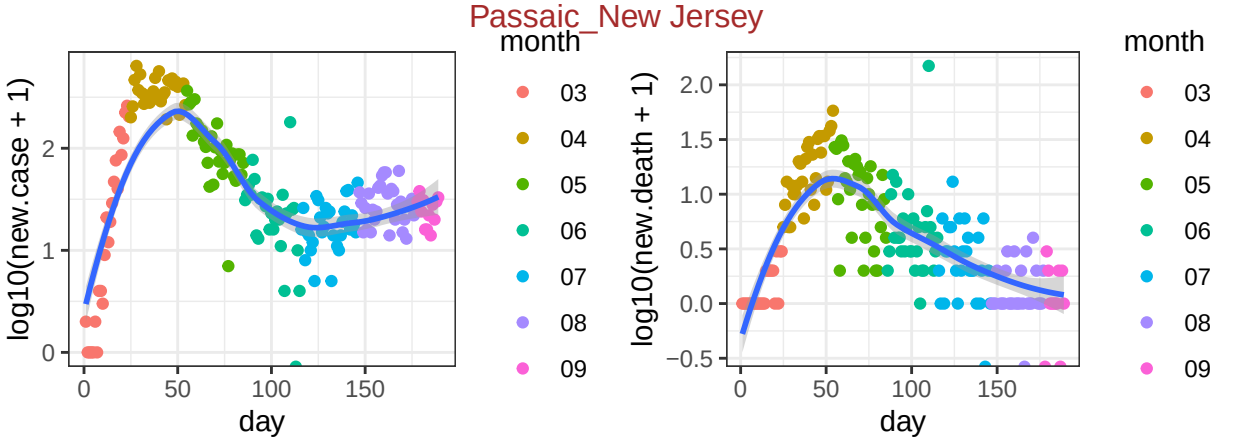
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Essex_Massachusetts



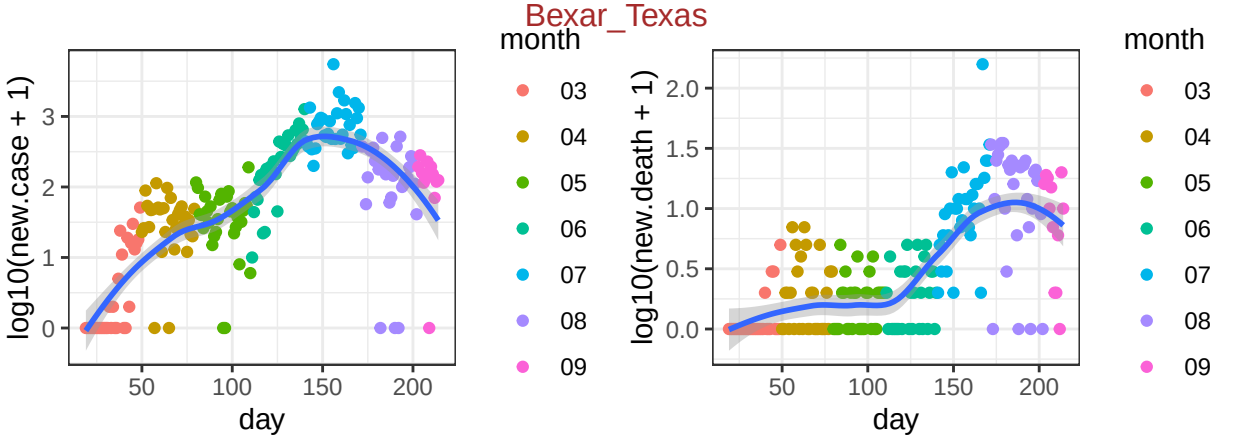
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Passaic_New Jersey



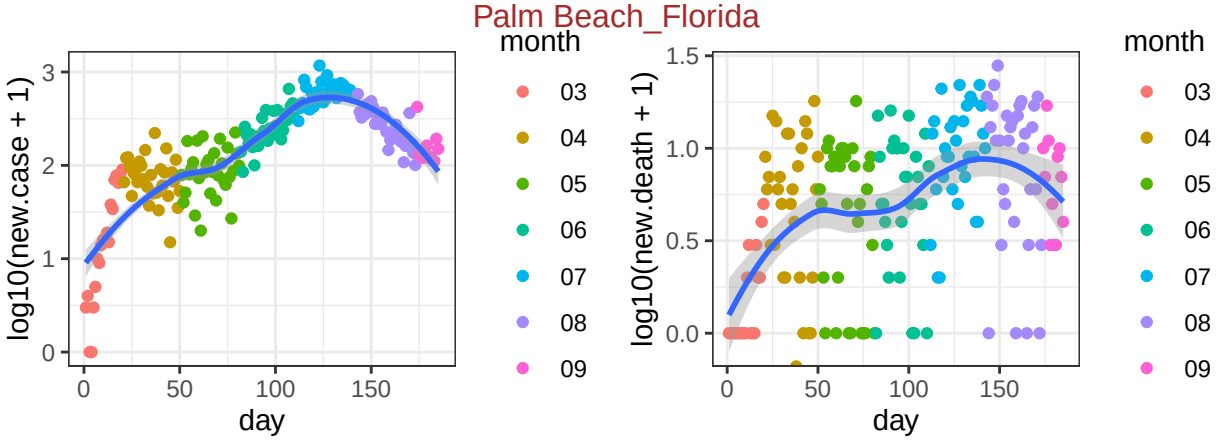
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Bexar_Texas



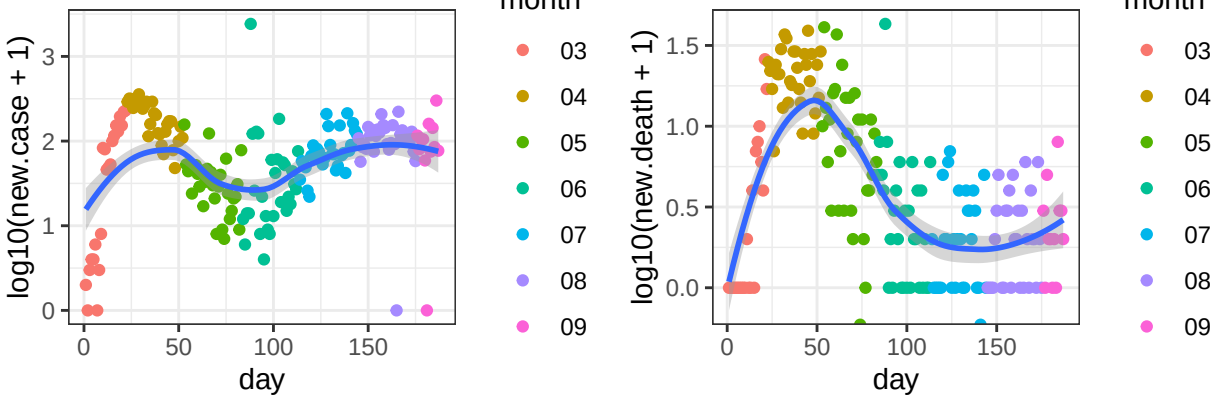
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Palm Beach_Florida



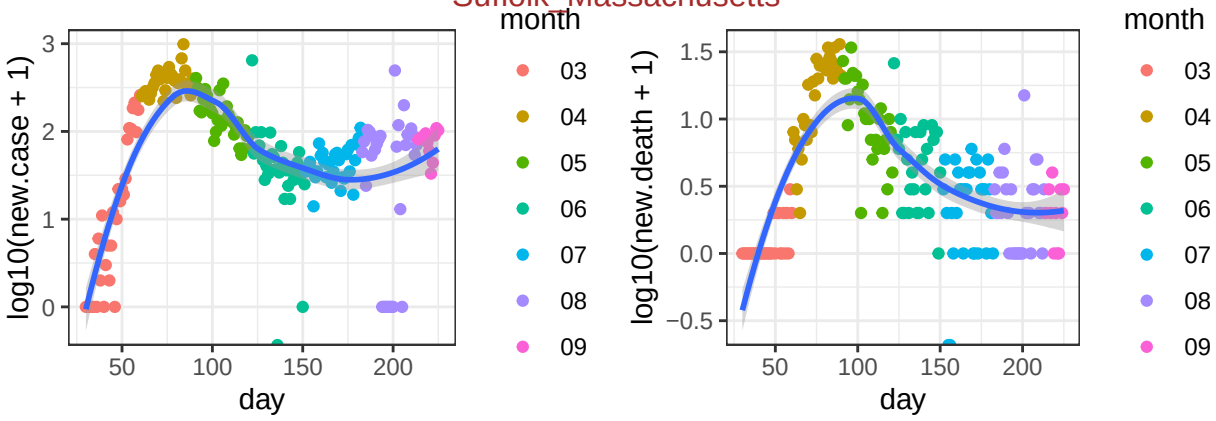
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Oakland_Michigan



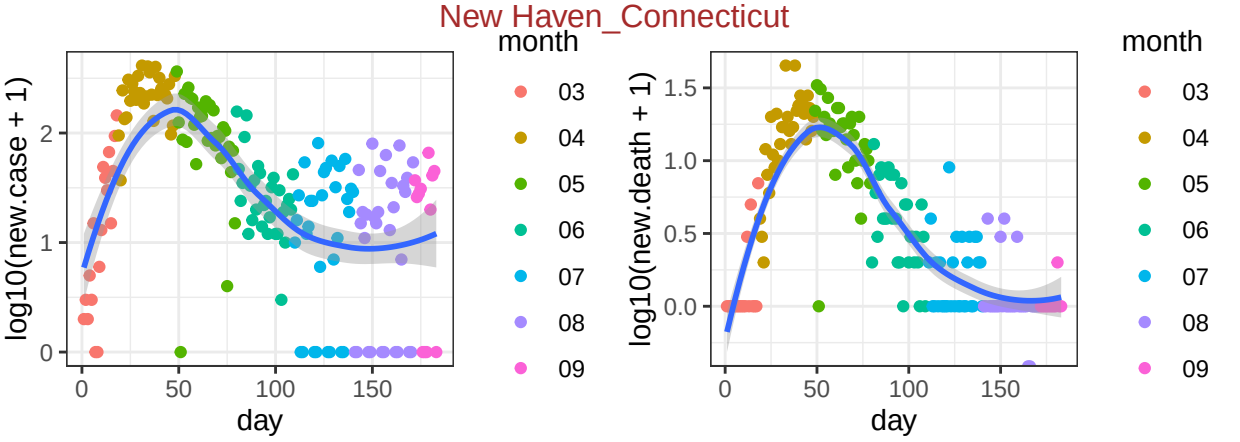
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Suffolk_Massachusetts



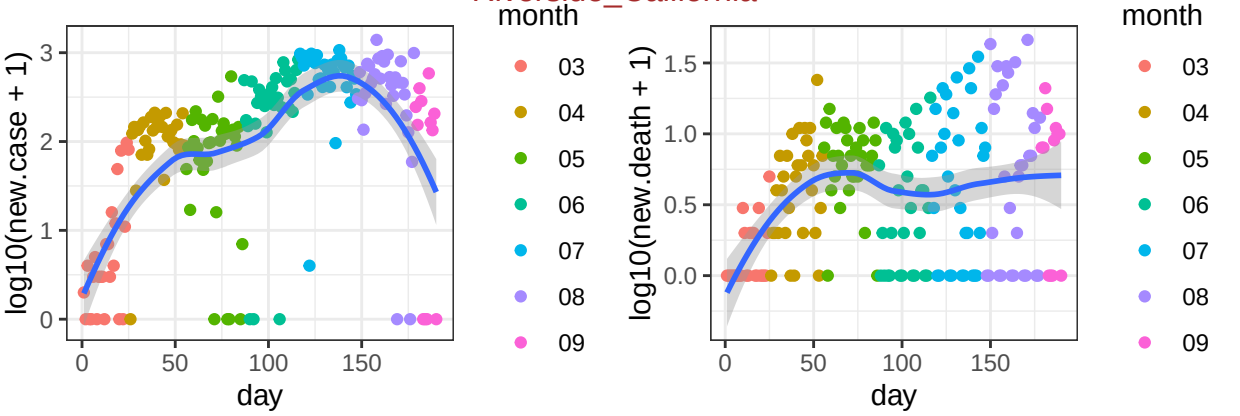
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

New Haven_Connecticut



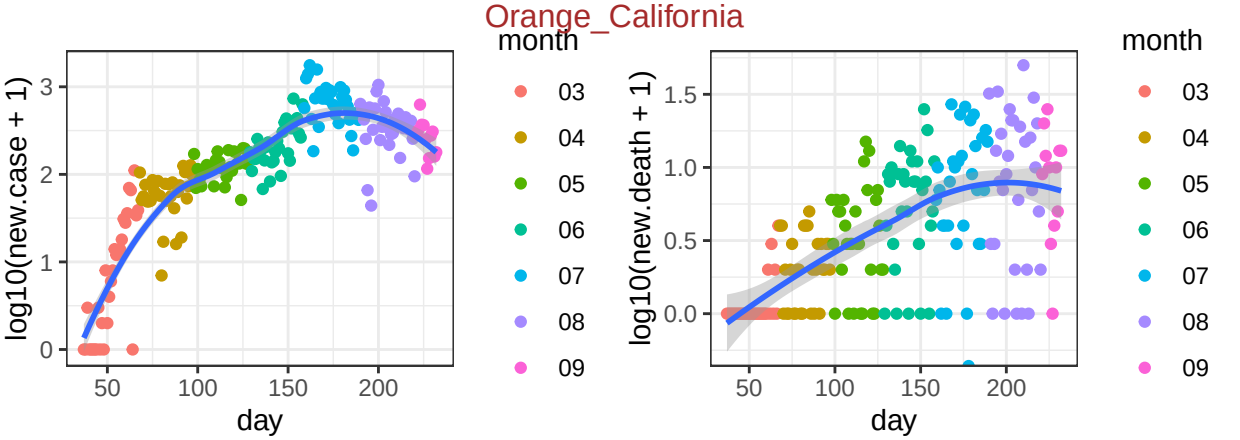
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−14

Riverside_California



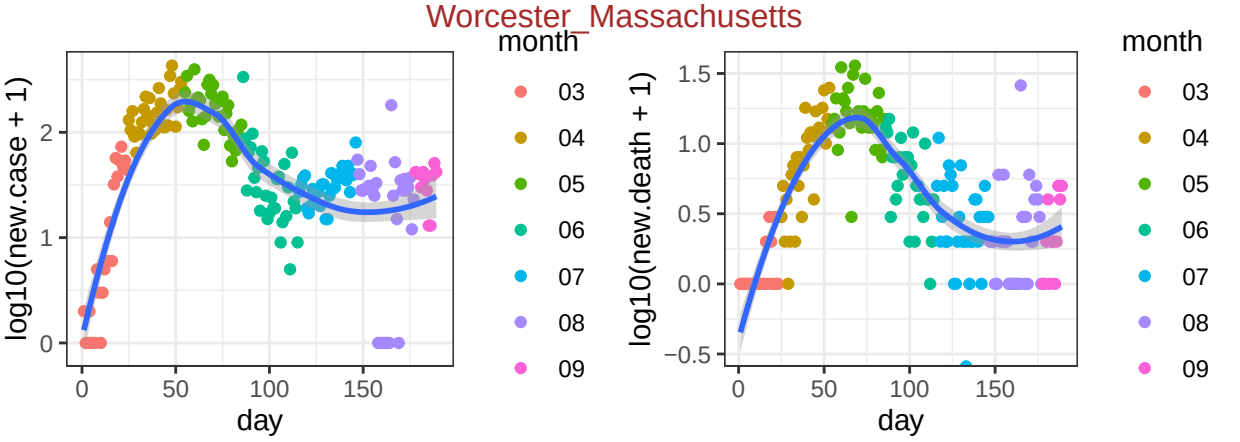
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Orange_California



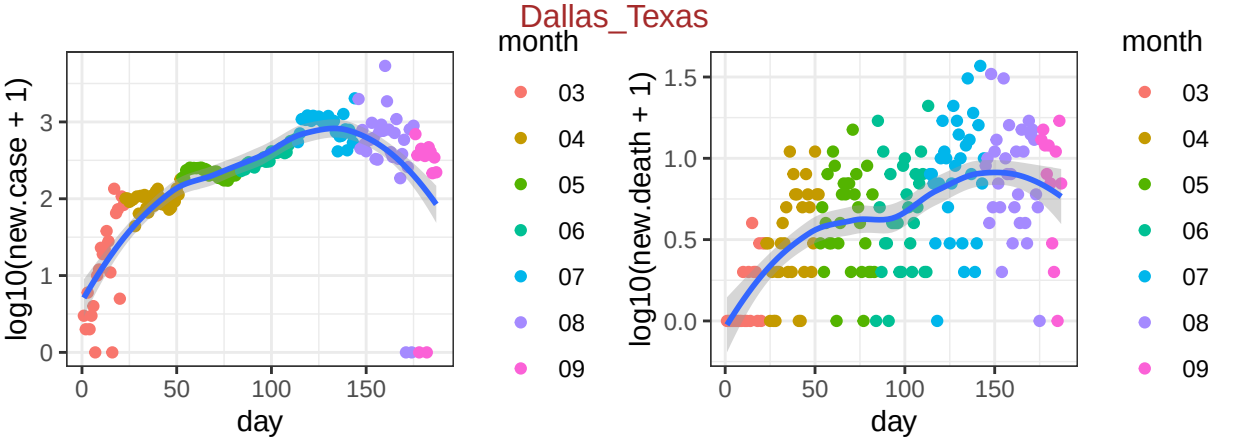
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

Worcester_Massachusetts



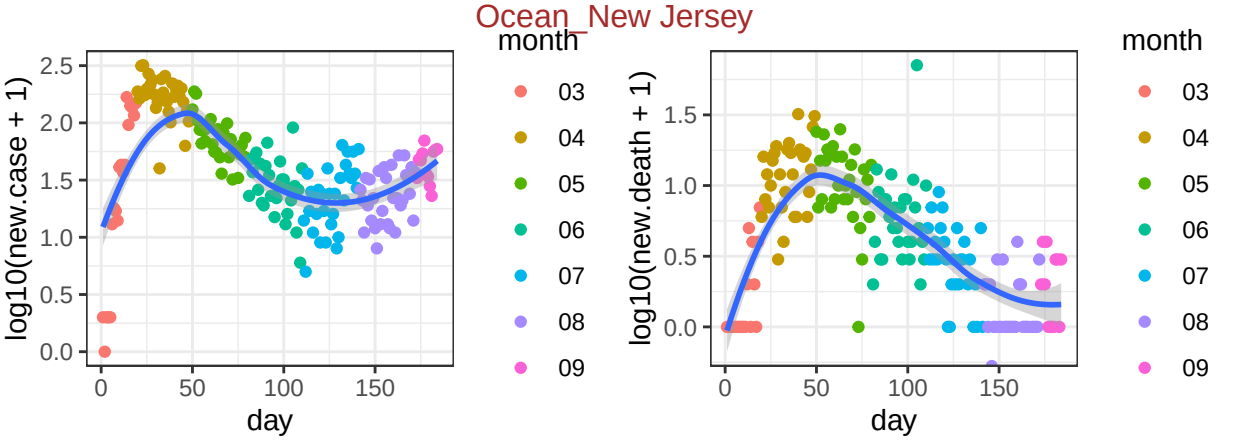
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−08

Dallas_Texas



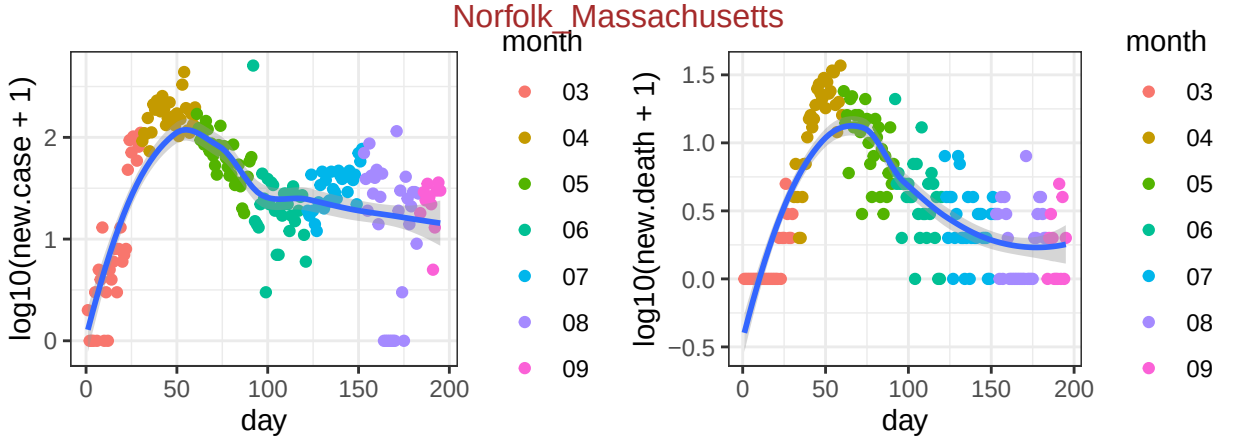
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−10

Ocean_New Jersey



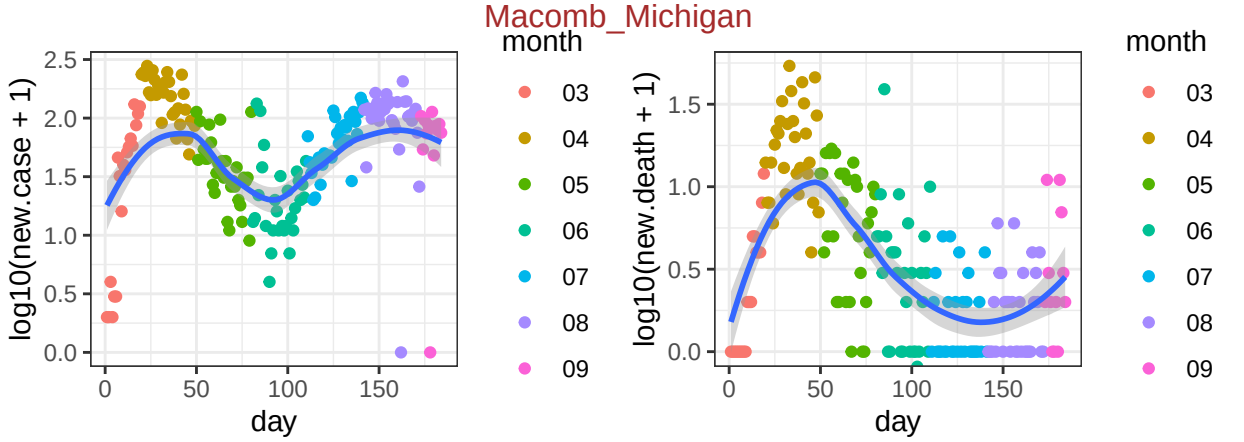
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Norfolk_Massachusetts



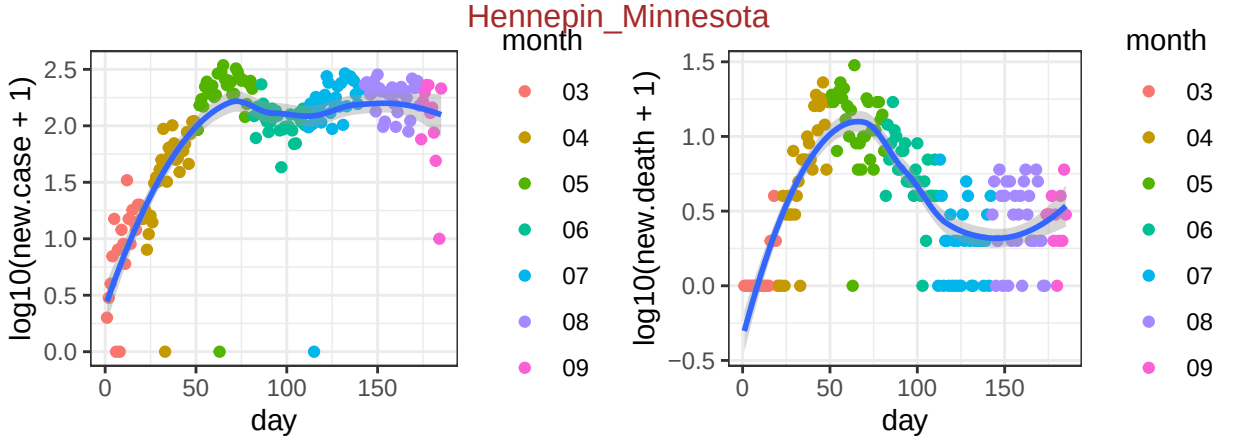
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−02

Macomb_Michigan



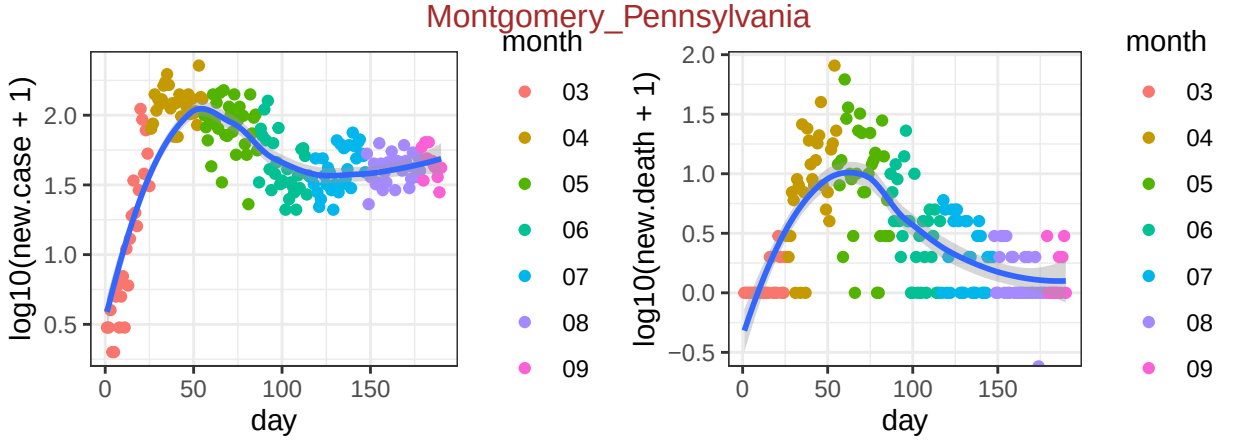
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−13

Hennepin_Minnesota



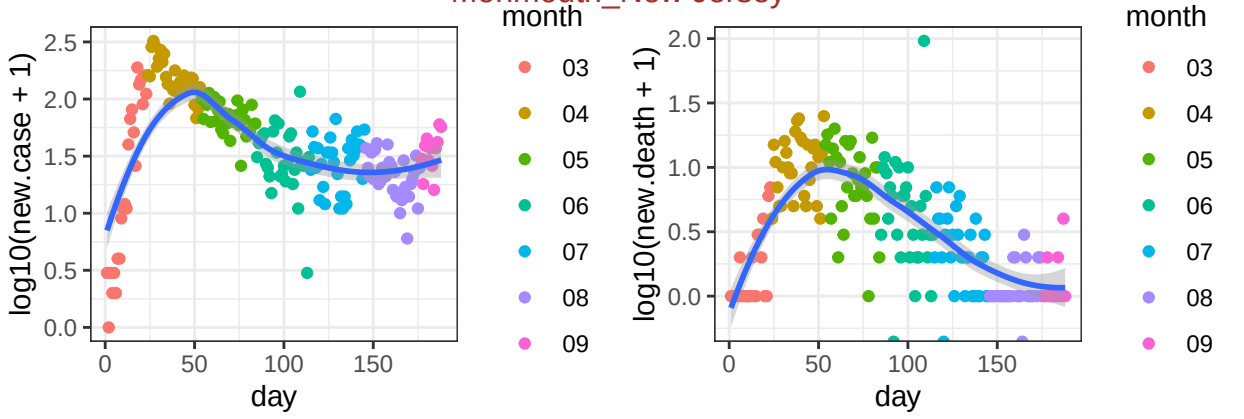
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

Montgomery_Pennsylvania



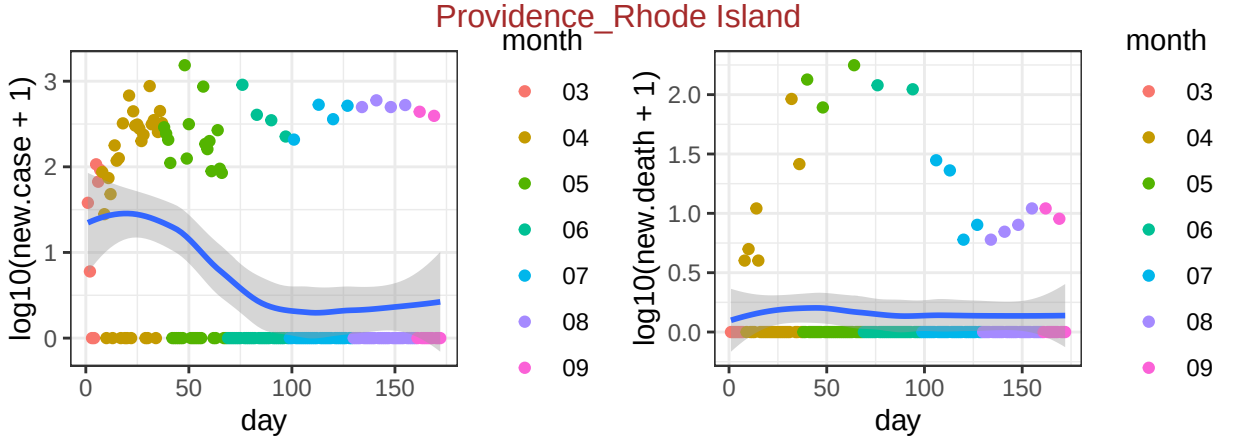
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Monmouth_New Jersey



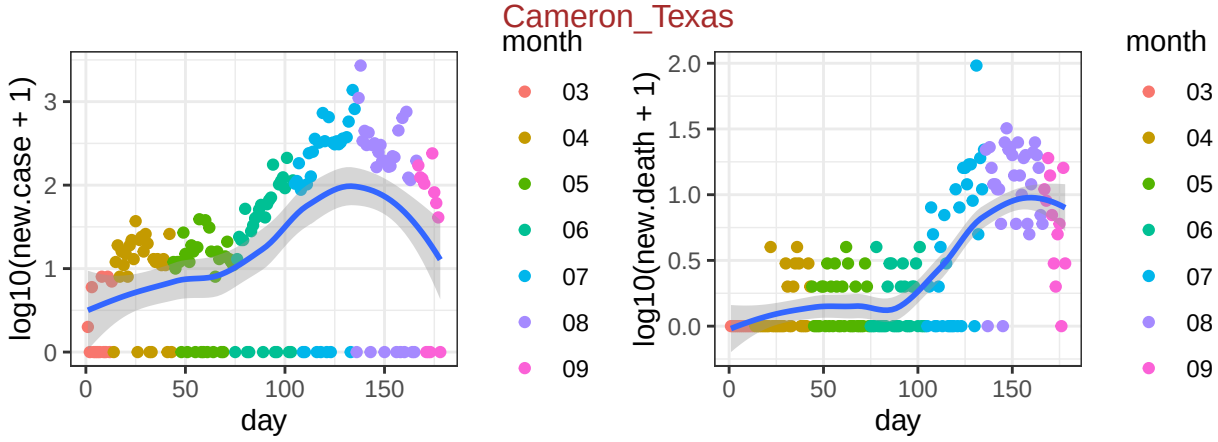
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Providence_Rhode Island



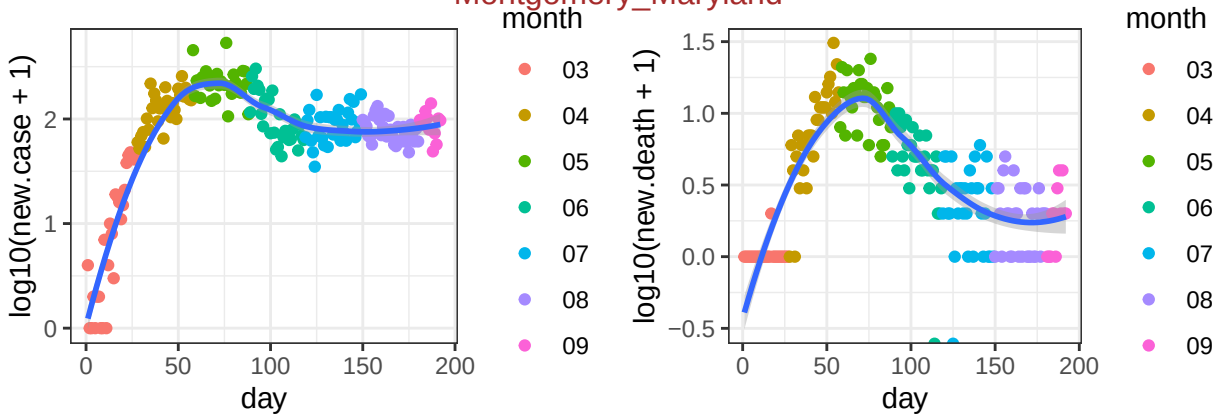
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−25

Cameron_Texas



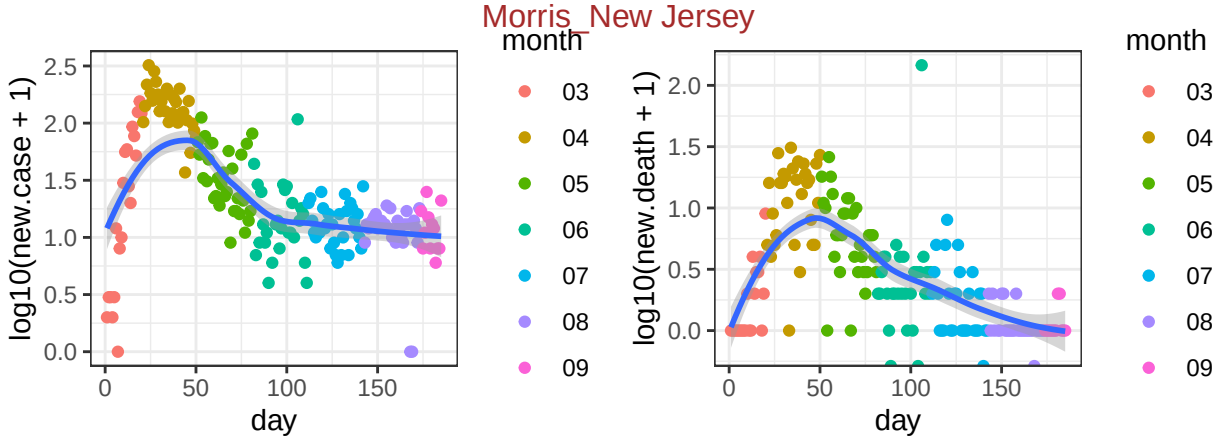
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−19

Montgomery_Maryland



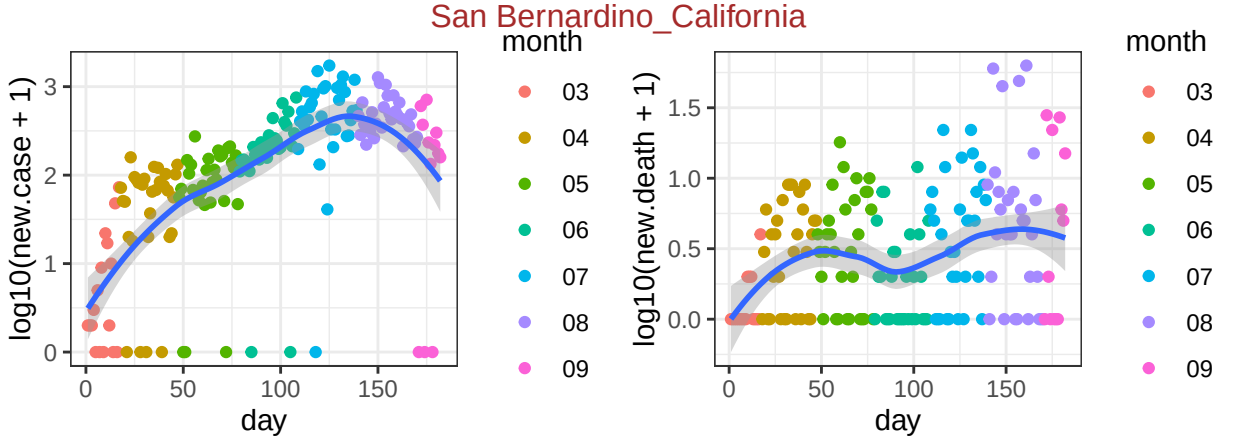
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−05

Morris_New Jersey



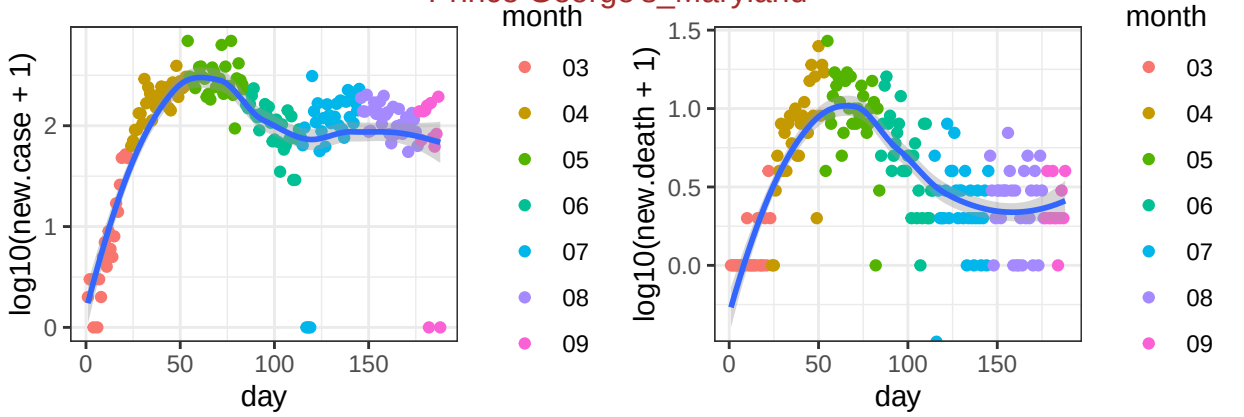
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−12

San Bernardino_California



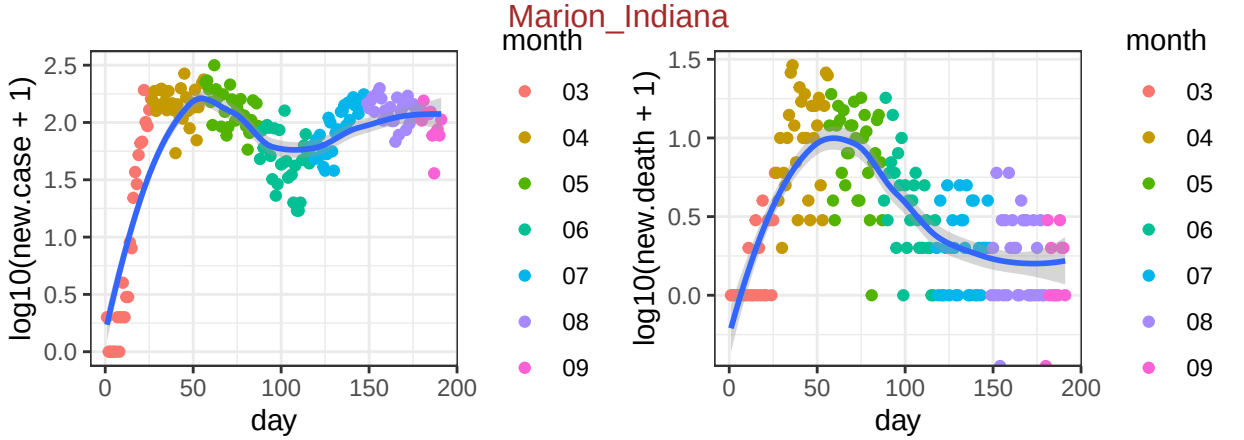
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−15

Prince George's_Maryland



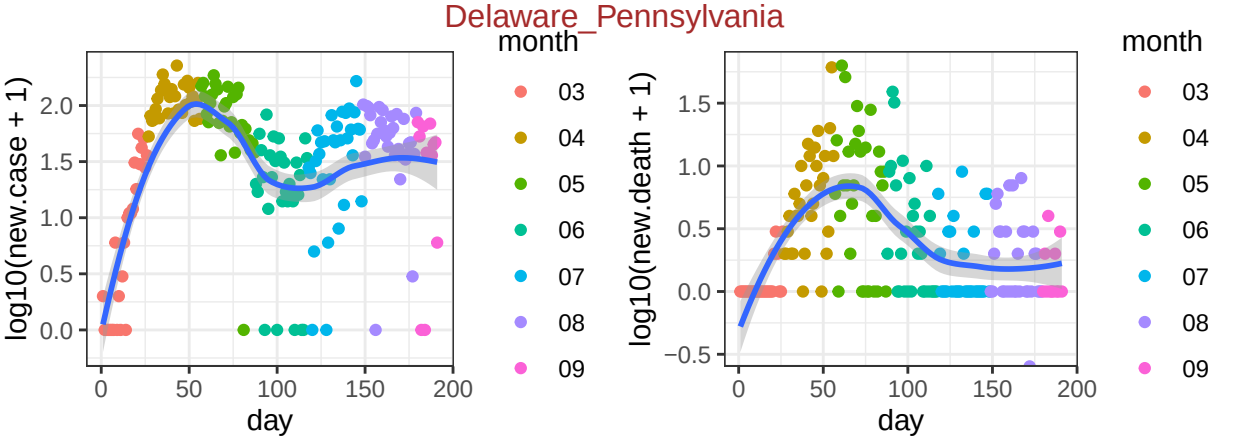
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−09

Marion_Indiana



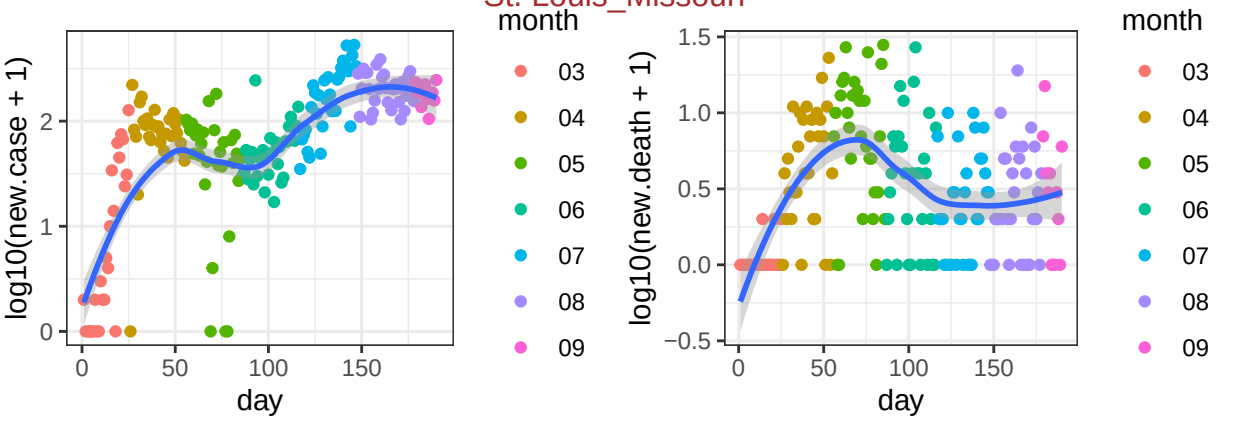
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

Delaware_Pennsylvania



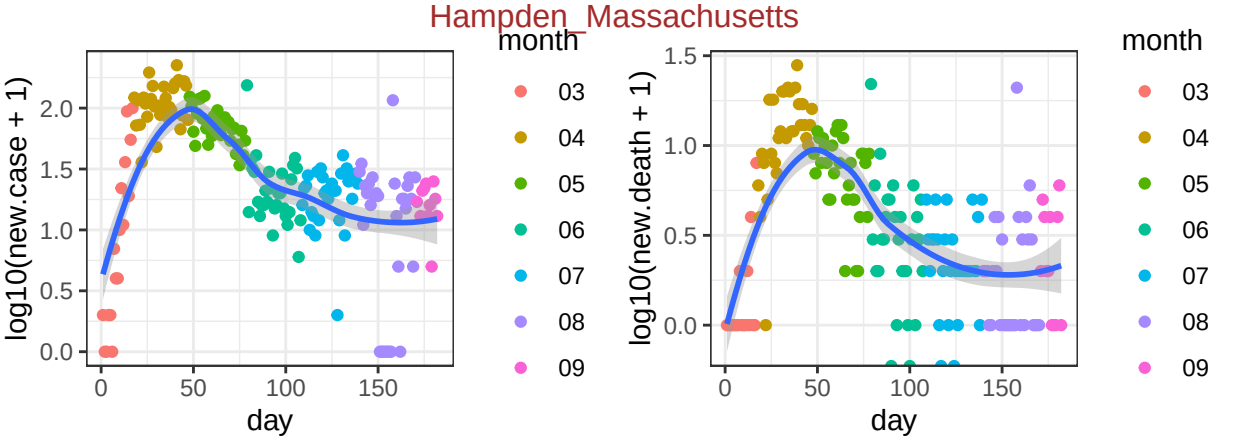
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−06

St. Louis_Missouri

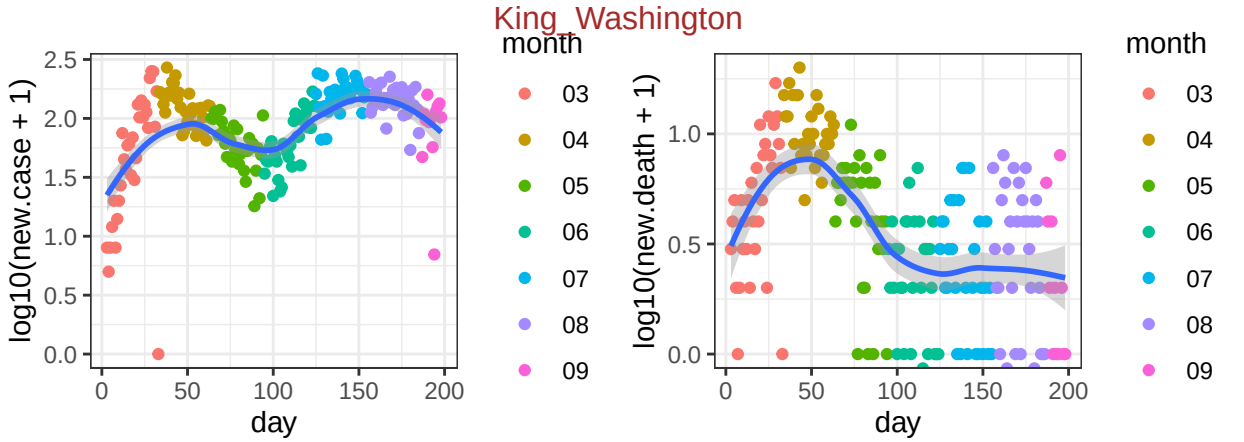


data source: https://github.com/nytimes/covid−19−data, day 1 is 03−07

Hampden_Massachusetts



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

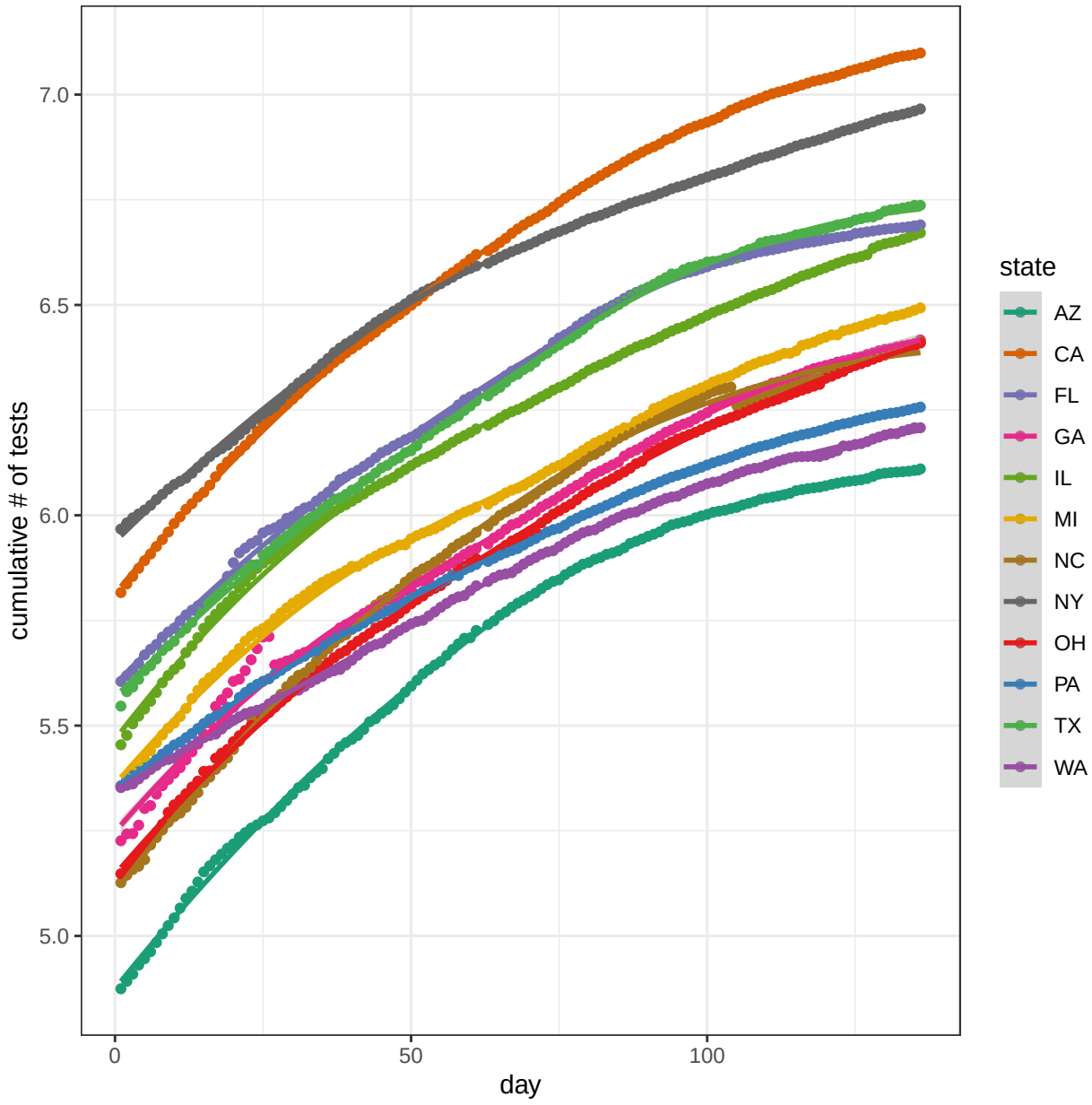


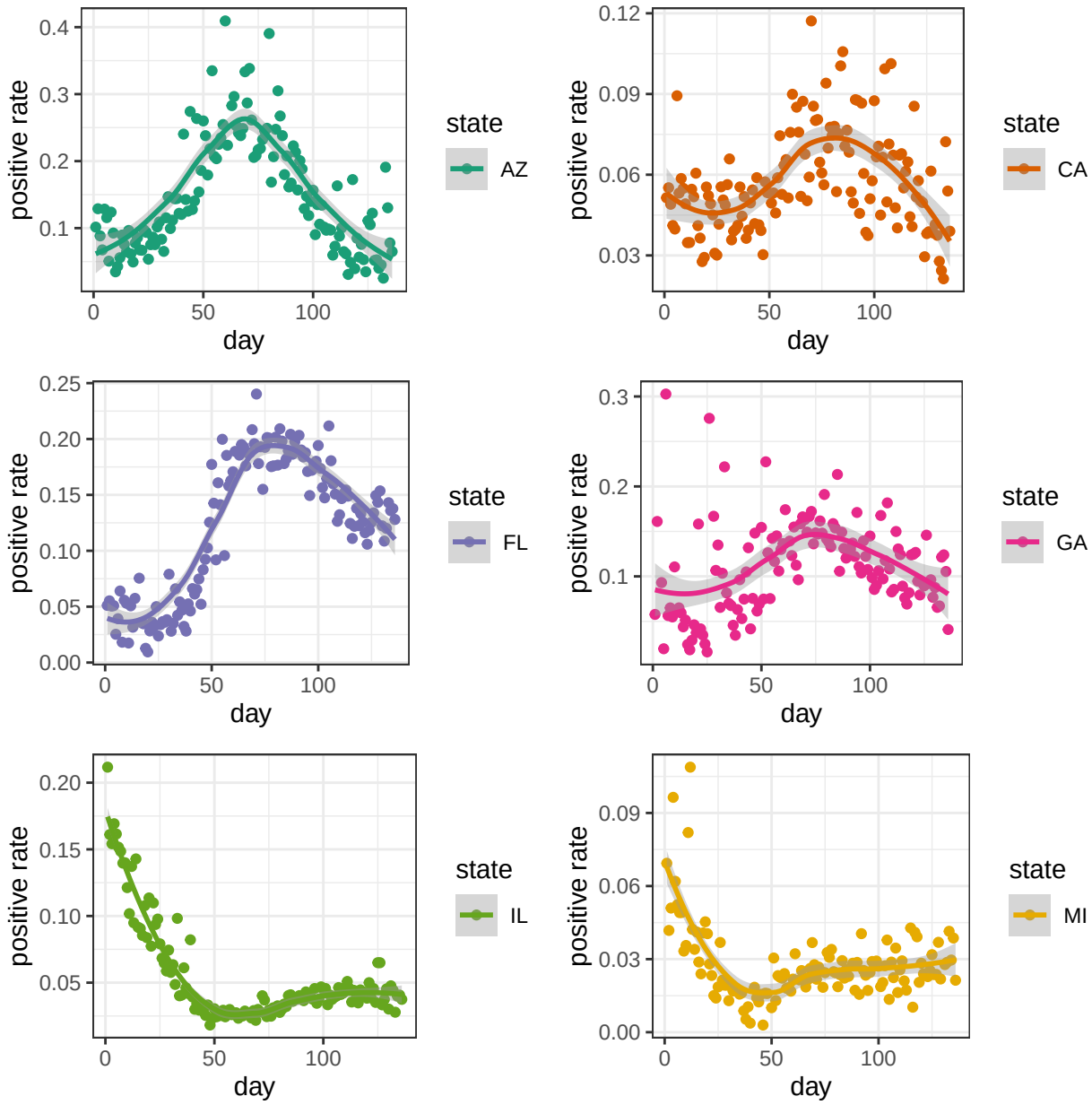
data source: https://github.com/nytimes/covid−19−data, day 1 is 03−01

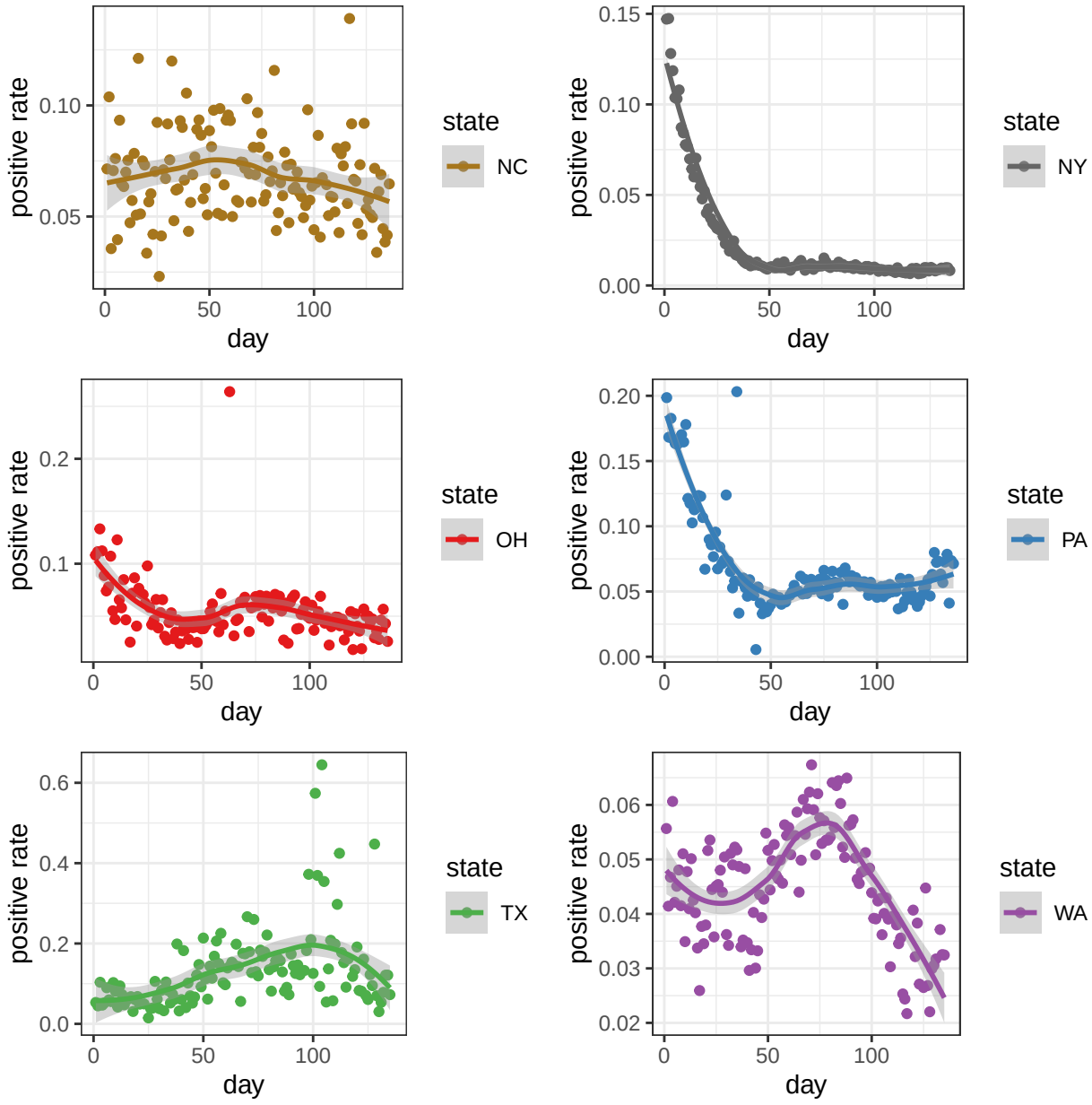
COVID Trackng

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.


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


positive rate
day
state
AZ
CA
FL
GA
IL
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
```