# Exploration of COVID-19 tracking data from multiple resources

## Wei Sun

#### 2020-06-11

### Contents

Introduction	1
JHU time series data	
daily reports data	 6
NY Times state level data	
county level data	
COVID Trackng	36
Session information	36

## 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 Tracking (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

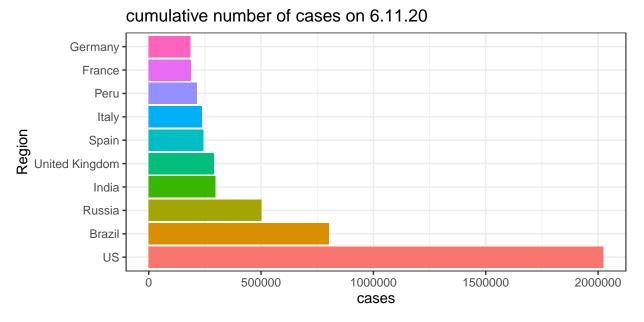
## JHU

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

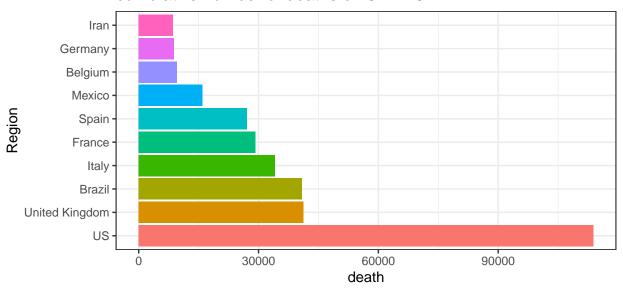
## time series data

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.

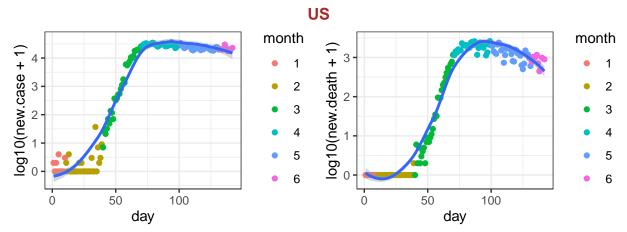
Here is the list of 10 records with the largest number of cases or deaths on the most recent date.



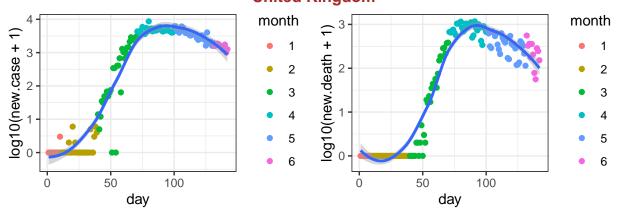
## cumulative number of deaths on 6.11.20



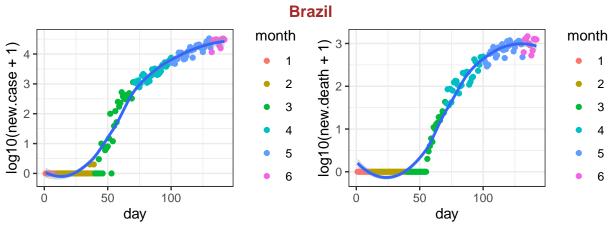
Next, I check for each country/region, what is the number of new cases/deaths? This data is important to understand what is the trend under different situations, e.g., population density, social distance policies etc. Here I checked the top 10 countries/regions with the highest number of deaths.



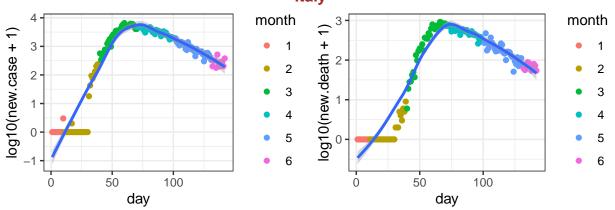
data source: https://github.com/CSSEGISandData/COVID-19, day 1 is 1/22/2020 **United Kingdom** 



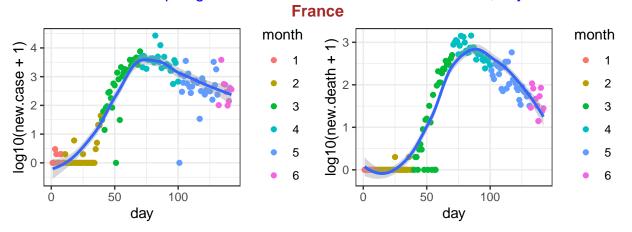
data source: https://github.com/CSSEGISandData/COVID-19, day 1 is 1/22/2020



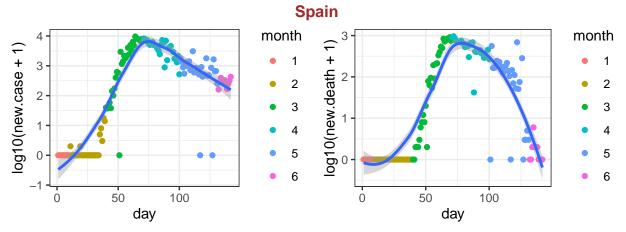
data source: https://github.com/CSSEGISandData/COVID-19, day 1 is 1/22/2020 **Italy** 



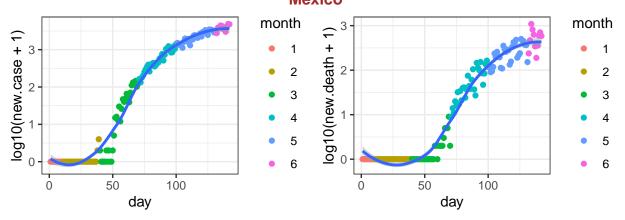
data source: https://github.com/CSSEGISandData/COVID-19, day 1 is 1/22/2020



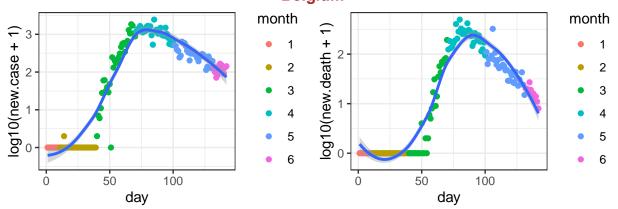
data source: https://github.com/CSSEGISandData/COVID-19, day 1 is 1/22/2020



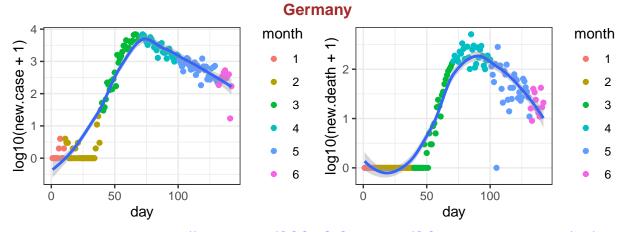
data source: https://github.com/CSSEGISandData/COVID-19, day 1 is 1/22/2020 **Mexico** 



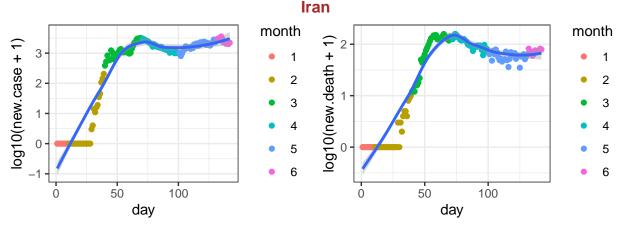
data source: https://github.com/CSSEGISandData/COVID-19, day 1 is 1/22/2020 **Belgium** 



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



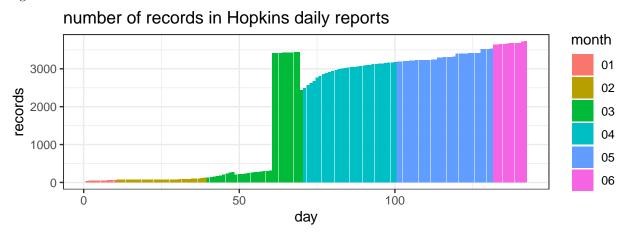
data source: https://github.com/CSSEGISandData/COVID-19, day 1 is 1/22/2020



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) include information from individual states of US or individual counties, as shown in the following figure. So I turn to NY Times data for information 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

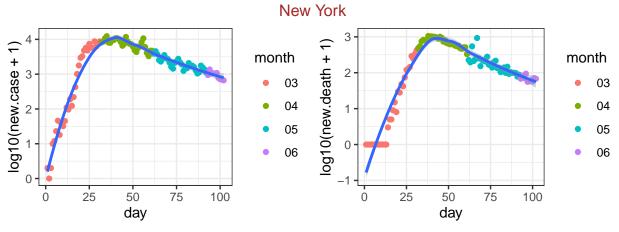
## [1] "2020-06-10"

## state level data

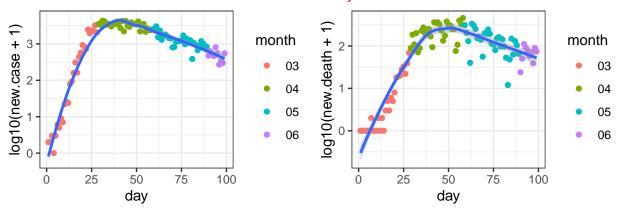
First check the 30 states with the largest number of deaths.

##		date	state	fips	cases	deaths
##	5493	2020-06-10	New York	36	384945	30376
##	5491	2020-06-10	New Jersey	34	165346	12377
##	5482	2020-06-10	Massachusetts	25	104156	7454
##	5474	2020-06-10	Illinois	17	130889	6302
##	5500	2020-06-10	Pennsylvania	42	81410	6143
##	5483	2020-06-10	Michigan	. 26	65377	5958
##	5464	2020-06-10	California	. 6	140123	4869
##	5466	2020-06-10	Connecticut	9	44347	4120
##	5479	2020-06-10	Louisiana	. 22	44143	2968
##	5481	2020-06-10	Maryland	. 24	60114	2844
##	5469	2020-06-10	Florida	. 12	67363	2800
##	5497	2020-06-10	Ohio	39	39575	2457
##	5475	2020-06-10	Indiana	. 18	39297	2355
##	5470	2020-06-10	Georgia	. 13	51465	2292
##	5506	2020-06-10	Texas	48	81771	1916
##	5465	2020-06-10	Colorado	8	28484	1573
##	5510	2020-06-10	Virginia	. 51	52177	1514
##	5484	2020-06-10	Minnesota	. 27	28900	1267
##	5511	2020-06-10	Washington	. 53	25940	1183
##	5462	2020-06-10	Arizona	. 4	29981	1100
##	5494	2020-06-10	North Carolina	. 37	38305	1096
##	5485	2020-06-10	Mississippi	28	18483	868
##	5486	2020-06-10	Missouri	29	15662	861
##	5502	2020-06-10	Rhode Island	. 44	15756	812
##	5460	2020-06-10	Alabama	. 1	21989	744
##	5513	2020-06-10	Wisconsin	. 55	21772	673
##	5476	2020-06-10	Iowa	. 19	22733	638
##	5503	2020-06-10	South Carolina	. 45	15759	575
##	5468	2020-06-10	District of Columbia	. 11	9537	499
##	5478	2020-06-10	Kentucky	21	12029	498

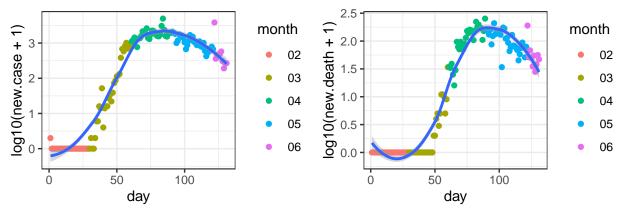
For these 20 states, I check the number of new cases and the number of new deaths. Part of the reason for such checking is to identify whether there is any similarity on such patterns. For example, could you use the pattern seen from Italy to predict what happen in an individual state, and what are the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



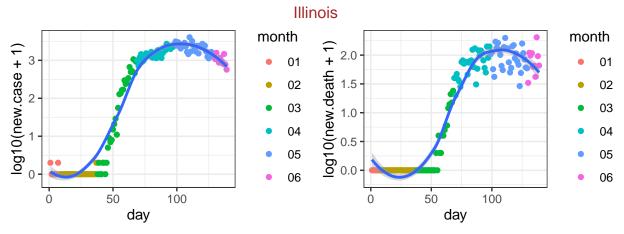
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-01
New Jersey



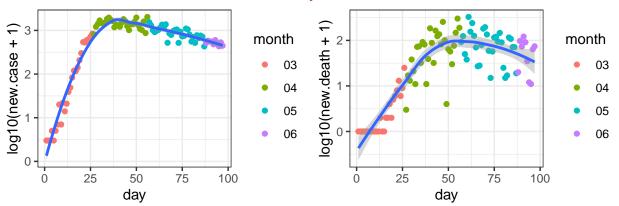
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-04
Massachusetts



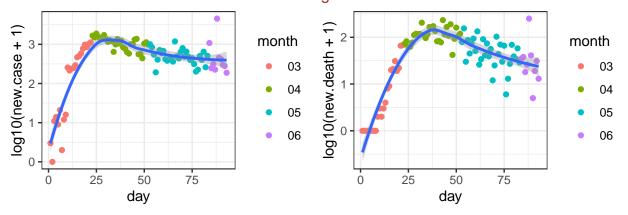
data source: https://github.com/nytimes/covid-19-data, day 1 is 02-01



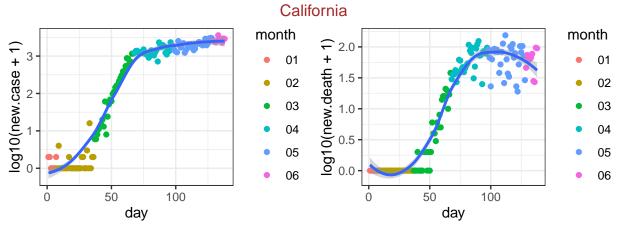
data source: https://github.com/nytimes/covid-19-data, day 1 is 01-24 Pennsylvania



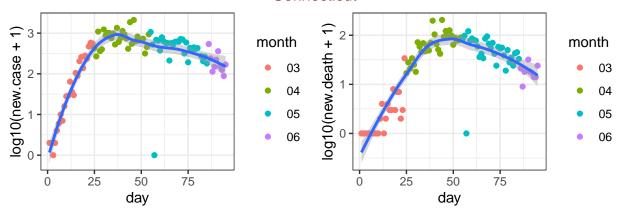
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-06 Michigan



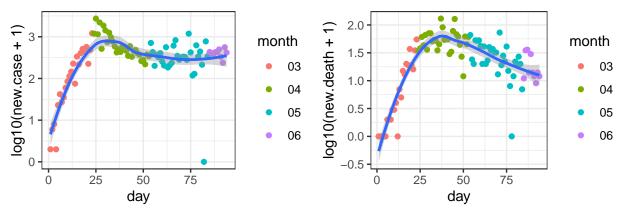
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-10



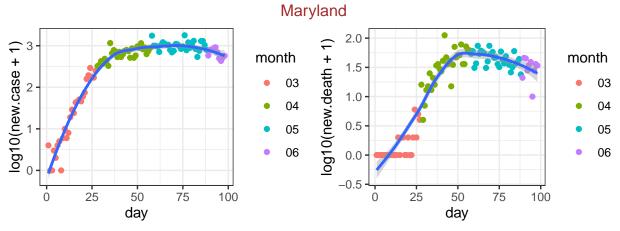
data source: https://github.com/nytimes/covid-19-data, day 1 is 01-25
Connecticut



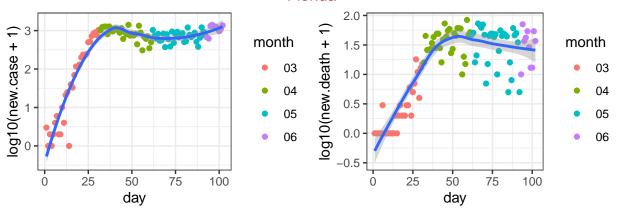
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-08 Louisiana



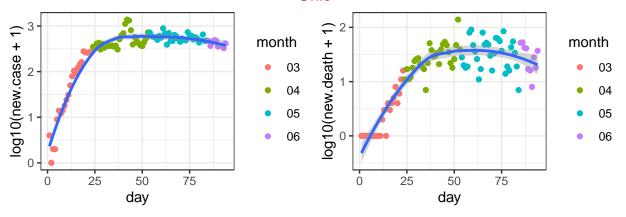
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-09



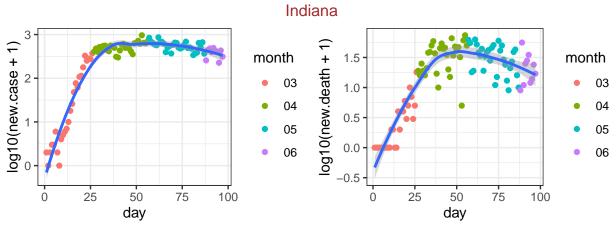
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-05 Florida



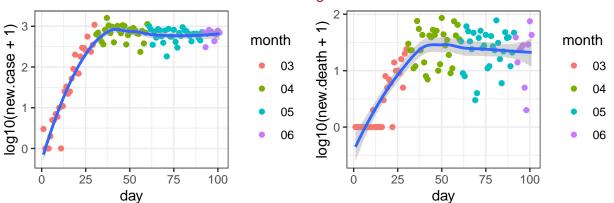
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-01
Ohio



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

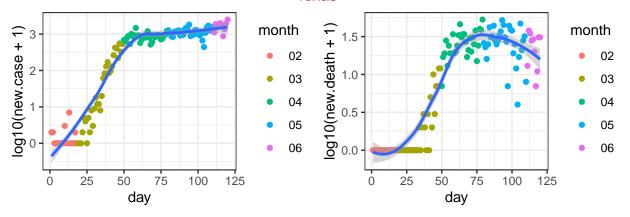


data source: https://github.com/nytimes/covid-19-data, day 1 is 03-06 Georgia

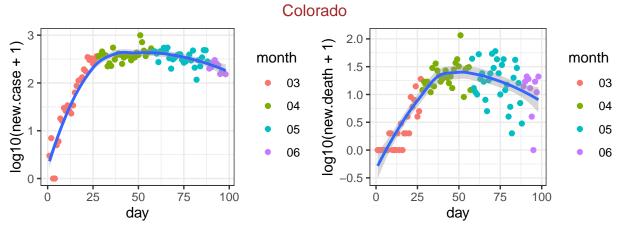


data source: https://github.com/nytimes/covid-19-data, day 1 is 03-02

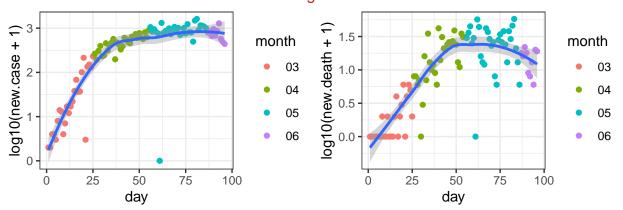
Texas



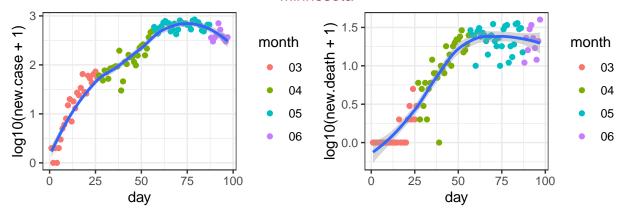
data source: https://github.com/nytimes/covid-19-data, day 1 is 02-12



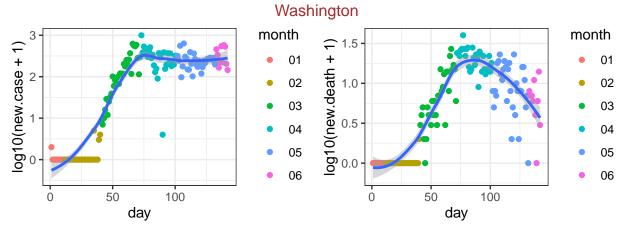
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-05 Virginia



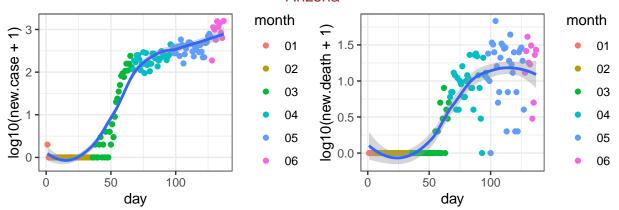
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-07
Minnesota



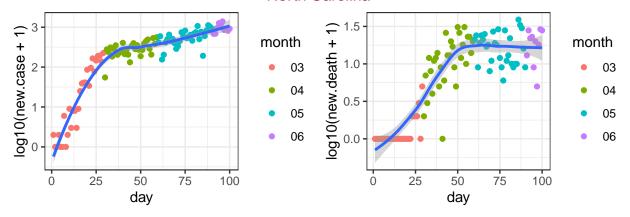
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-06



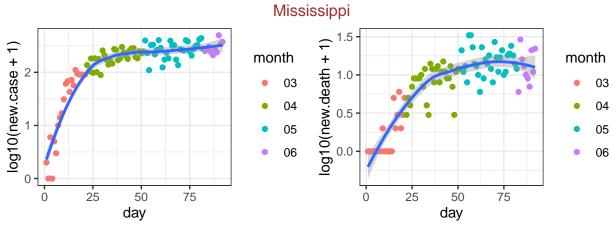
data source: https://github.com/nytimes/covid-19-data, day 1 is 01-21 Arizona



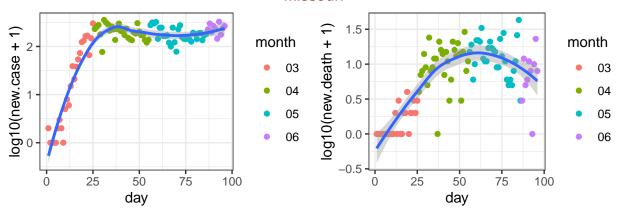
data source: https://github.com/nytimes/covid-19-data, day 1 is 01-26 North Carolina



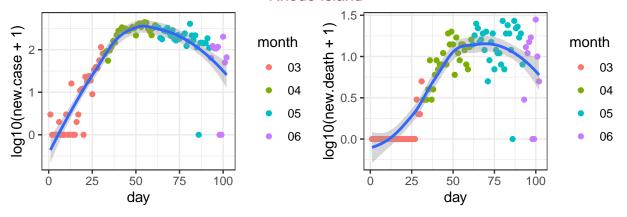
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-03



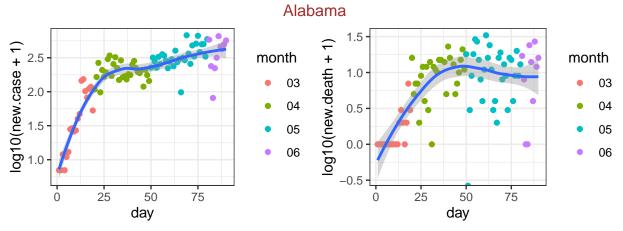
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-11
Missouri



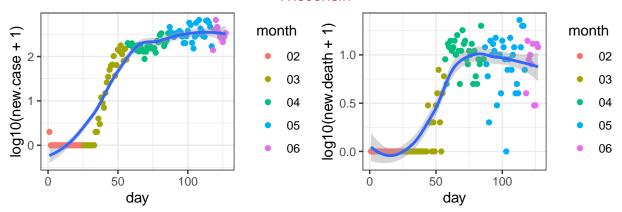
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-07 Rhode Island



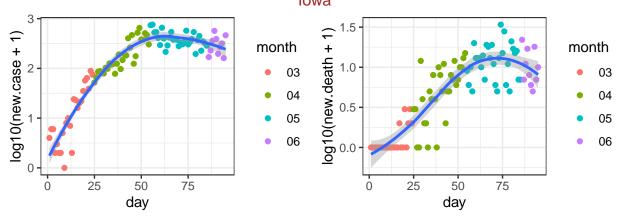
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-01



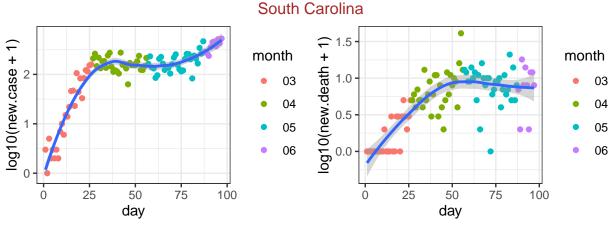
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-13 Wisconsin



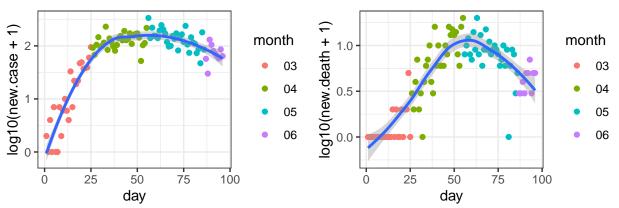
data source: https://github.com/nytimes/covid-19-data, day 1 is 02-05 lowa



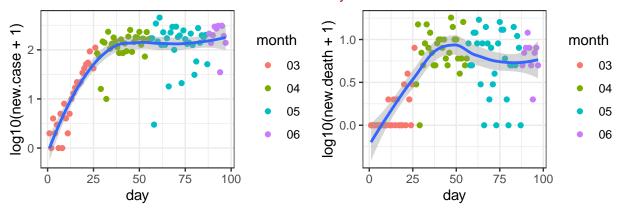
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-08



data source: https://github.com/nytimes/covid-19-data, day 1 is 03-06
District of Columbia

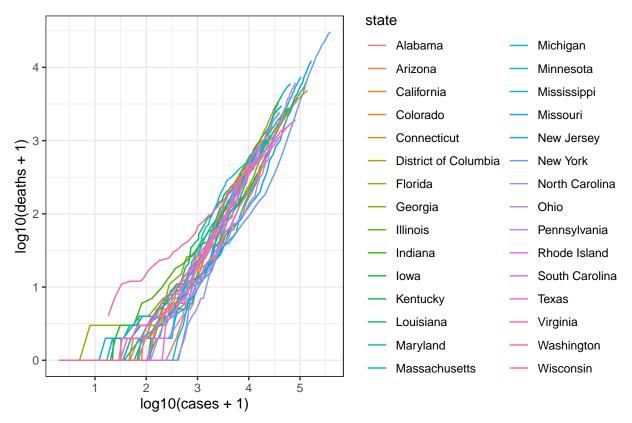


data source: https://github.com/nytimes/covid-19-data, day 1 is 03-07
Kentucky



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

Next I check the relation between the  $\mathbf{cumulative}$  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
##	223426	2020-06-10	New York City	New York	NA	212884	21436
##	222241	2020-06-10	Cook	Illinois	17031	83585	4053
##	221845	2020-06-10	Los Angeles	California	6037	67064	2768
##	222933	2020-06-10	Wayne	Michigan	26163	21570	2653
##	223425	2020-06-10	Nassau	New York	36059	41015	2653
##	223445	2020-06-10	Suffolk	New York	36103	40464	1990
##	222845	2020-06-10	Middlesex	Massachusetts	25017	22889	1725
##	223351	2020-06-10	Essex	New Jersey	34013	18206	1723
##	223346	2020-06-10	Bergen	New Jersey	34003	18667	1635
##	223453	2020-06-10	Westchester	New York	36119	34075	1530
##	223850	2020-06-10	Philadelphia	Pennsylvania	42101	23951	1454
##	221944	2020-06-10	Fairfield	Connecticut	9001	16134	1321
##	221945	2020-06-10	Hartford	Connecticut	9003	10924	1303
##	223353	2020-06-10	Hudson	New Jersey	34017	18647	1242
##	223364	2020-06-10	Union	New Jersey	34039	16317	1103
##	223356	2020-06-10	${ t Middlesex}$	New Jersey	34023	16288	1064
##	222914	2020-06-10	Oakland	Michigan	26125	11262	1058
##	221948	2020-06-10	New Haven	Connecticut	9009	11911	1024
##	222841	2020-06-10	Essex	Massachusetts	25009	15365	1024
##	223360	2020-06-10	Passaic	New Jersey	34031	16524	982
##	222849	2020-06-10	Suffolk	Massachusetts	25025	19099	936
##	222901	2020-06-10	Macomb	Michigan	26099	7000	876

##	222847	2020-06-10	Norfolk	Massachusetts	25021	8774	873
##	222851	2020-06-10	Worcester	Massachusetts	25027	11820	844
##	223359	2020-06-10	Ocean	New Jersey	34029	9100	792
##	222000	2020-06-10	Miami-Dade	Florida	12086	20276	784
##	223845	2020-06-10	Montgomery	Pennsylvania	42091	7709	762
##	222960	2020-06-10	Hennepin	Minnesota	27053	9674	693
##	222376	2020-06-10	Marion	Indiana	18097	10581	680
##	222827	2020-06-10	Montgomery	Maryland	24031	13163	672
##	223822	2020-06-10	Delaware	Pennsylvania	42045	6811	662
##	223357	2020-06-10	Monmouth	New Jersey	34025	8563	652
##	223871	2020-06-10	Providence	Rhode Island	44007	11959	637
##	222843	2020-06-10	Hampden	Massachusetts	25013	6395	629
##	223358	2020-06-10	Morris	New Jersey	34027	6596	627
##	222828	2020-06-10	Prince George's	Maryland	24033	17305	613
##	222848	2020-06-10	Plymouth	Massachusetts	25023	8418	608
##	224496	2020-06-10	King	Washington	53033	8561	582
##	223411	2020-06-10	Erie	New York	36029	6616	563
##	223808	2020-06-10	Bucks	Pennsylvania	42017	5340	534
##	221744	2020-06-10	Maricopa	Arizona	4013	15282	519
##	222765	2020-06-10	Orleans	Louisiana	22071	7279	513
##	223355	2020-06-10	Mercer	New Jersey	34021	7245	510
##	221957	2020-06-10	District of Columbia	${\tt District\ of\ Columbia}$	11001	9537	499
##	222839	2020-06-10	Bristol	Massachusetts	25005	7754	487
##	223199	2020-06-10	St. Louis	Missouri	29189	5388	485
##	223437	2020-06-10	Rockland	New York	36087	13372	466
##	222755	2020-06-10	Jefferson	Louisiana	22051	7971	463
##	223362	2020-06-10	Somerset	New Jersey	34035	4698	431
##	224385	2020-06-10	Fairfax	Virginia	51059	12746	422

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

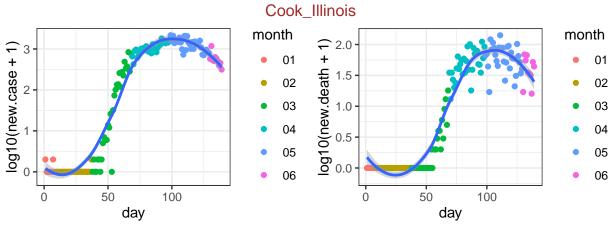
day

#### log10(new.death + 1) log10(new.case + 1) month month 2 -Ö Ö

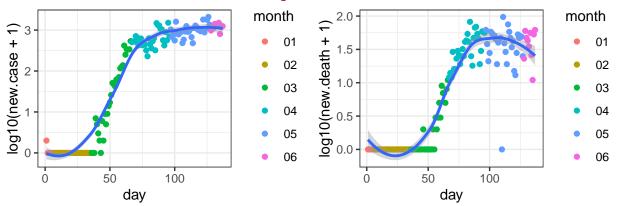
New York City\_New York

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

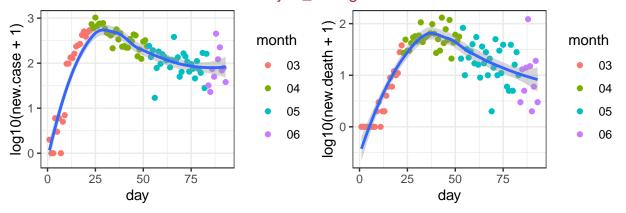
day



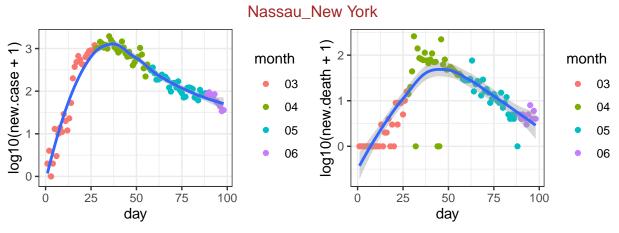
data source: https://github.com/nytimes/covid-19-data, day 1 is 01-24 Los Angeles\_California



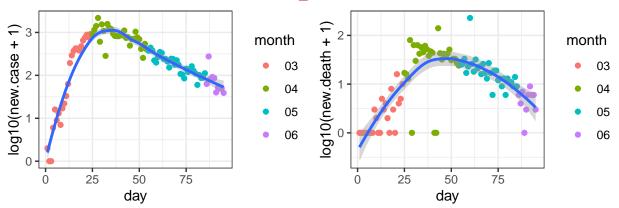
data source: https://github.com/nytimes/covid-19-data, day 1 is 01-26 Wayne\_Michigan



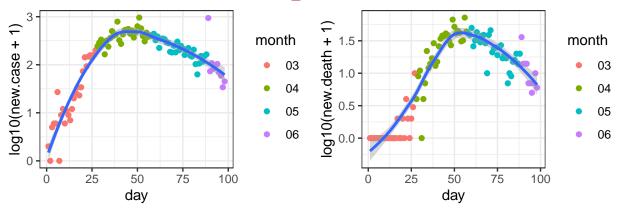
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-10



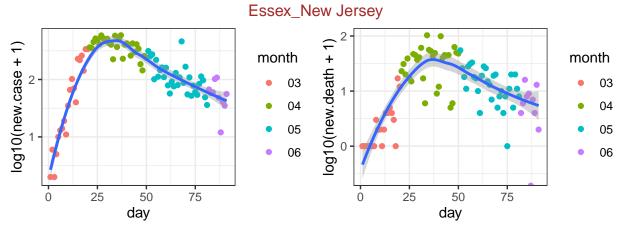
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-05 Suffolk\_New York



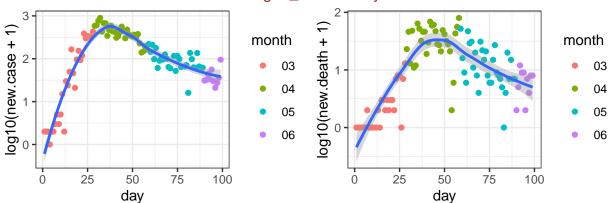
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-08
Middlesex\_Massachusetts



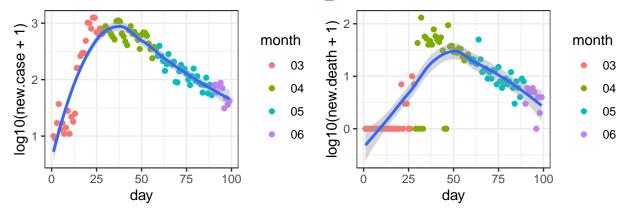
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-05



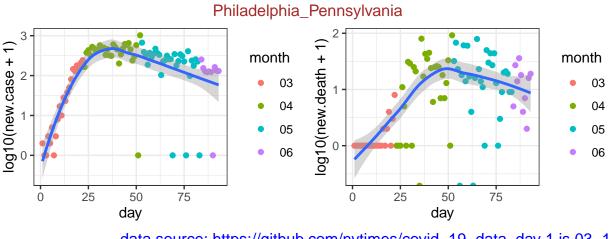
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-12
Bergen\_New Jersey



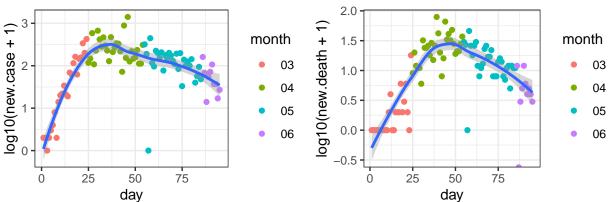
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-04 Westchester\_New York



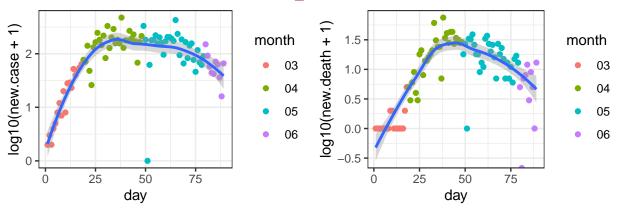
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-04



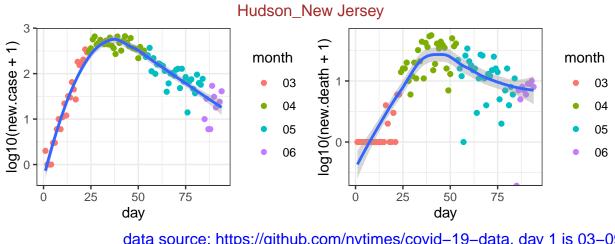
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-10 Fairfield\_Connecticut



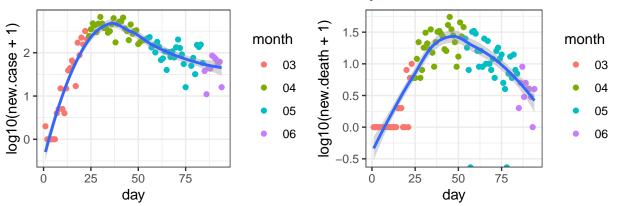
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-08
Hartford\_Connecticut



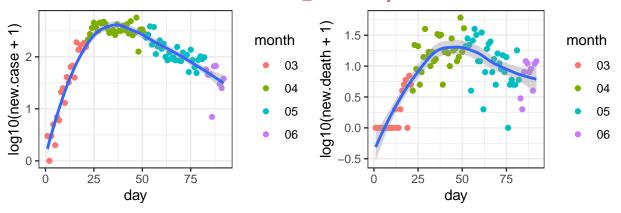
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-14



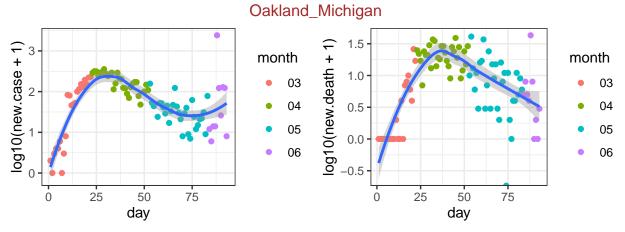
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-09
Union\_New Jersey



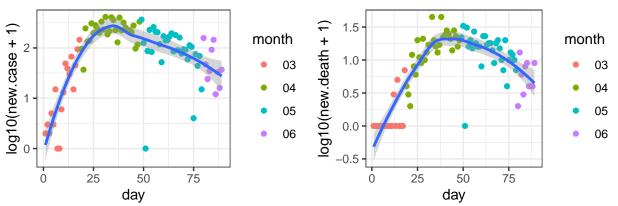
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-09
Middlesex\_New Jersey

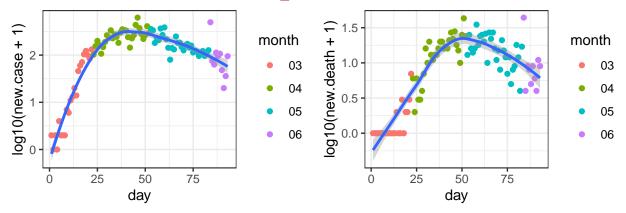


data source: https://github.com/nytimes/covid-19-data, day 1 is 03-11

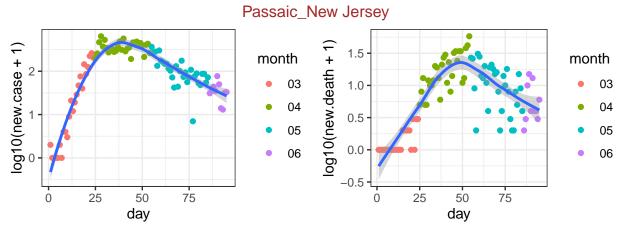


data source: https://github.com/nytimes/covid-19-data, day 1 is 03-10
New Haven\_Connecticut

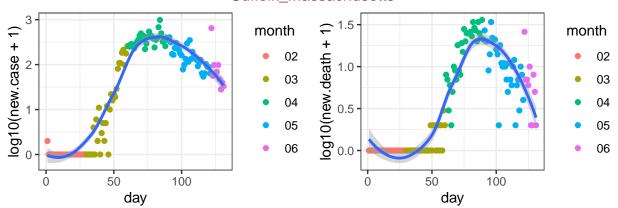




data source: https://github.com/nytimes/covid-19-data, day 1 is 03-10

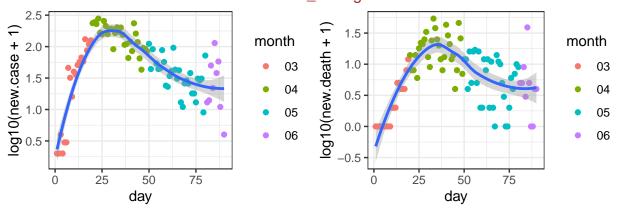


data source: https://github.com/nytimes/covid-19-data, day 1 is 03-08 Suffolk\_Massachusetts

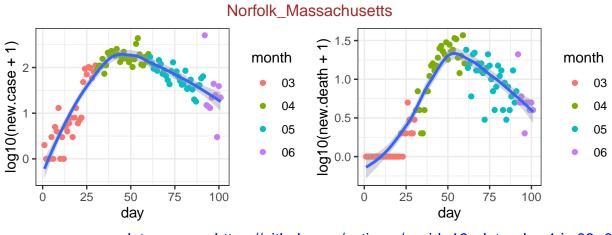


data source: https://github.com/nytimes/covid-19-data, day 1 is 02-01

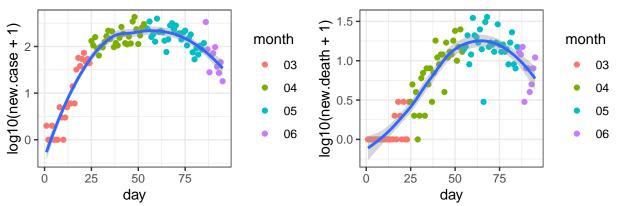
Macomb\_Michigan



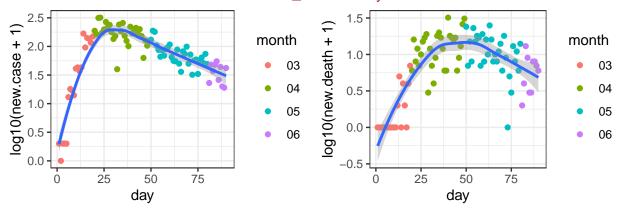
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-13



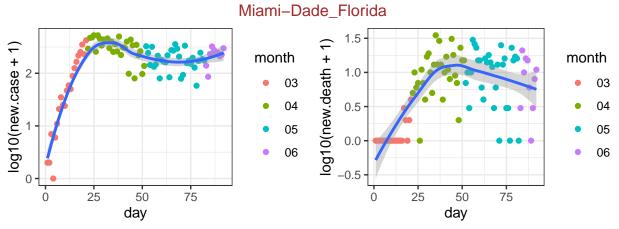
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-02 Worcester\_Massachusetts



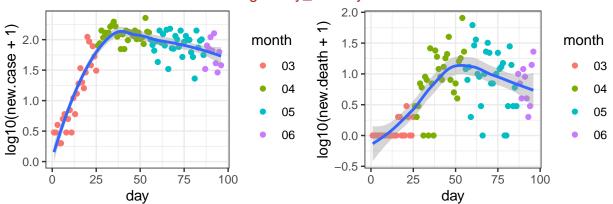
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-08
Ocean\_New Jersey



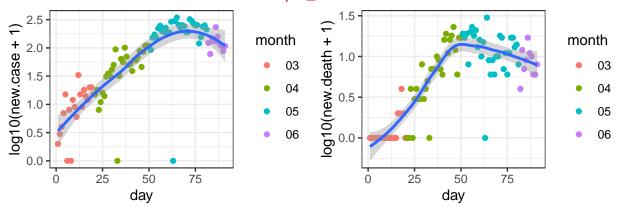
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-13



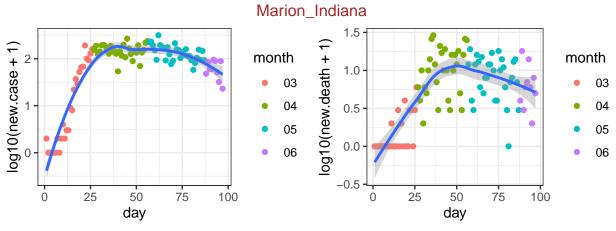
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-11
Montgomery\_Pennsylvania



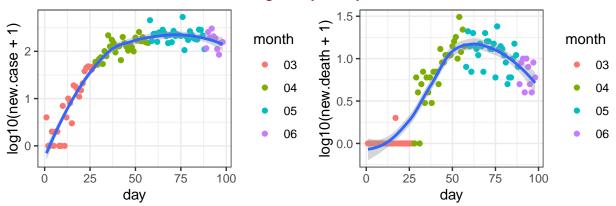
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-07
Hennepin\_Minnesota



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

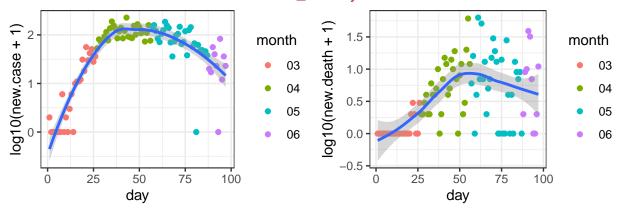


data source: https://github.com/nytimes/covid-19-data, day 1 is 03-06 Montgomery\_Maryland

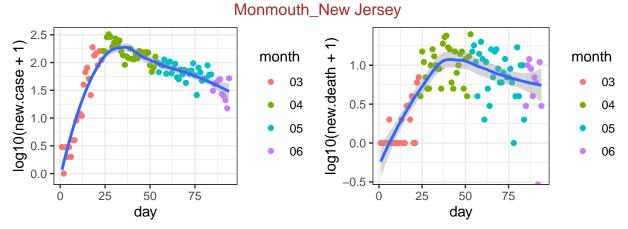


data source: https://github.com/nytimes/covid-19-data, day 1 is 03-05

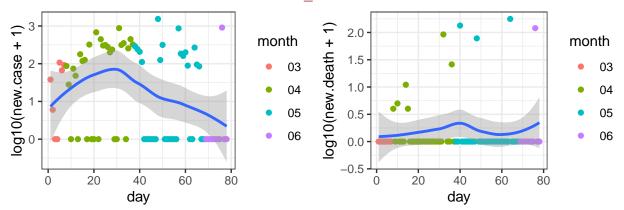
Delaware\_Pennsylvania



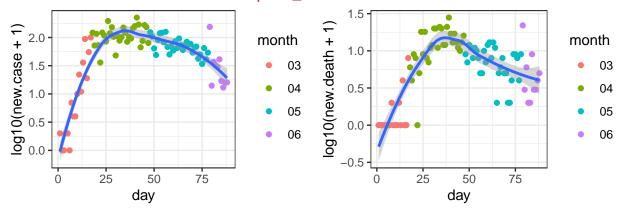
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-06



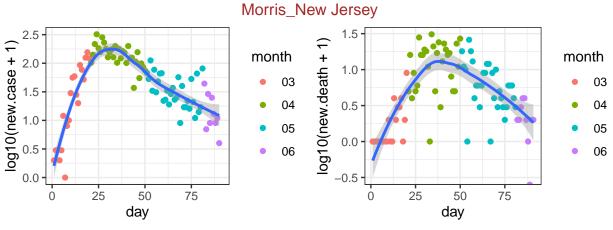
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-09 Providence\_Rhode Island



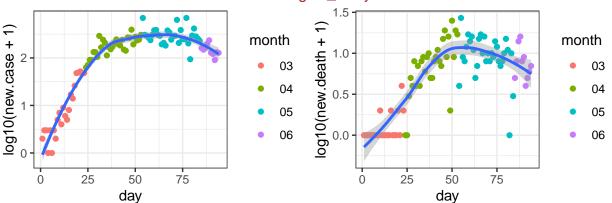
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-25
Hampden\_Massachusetts

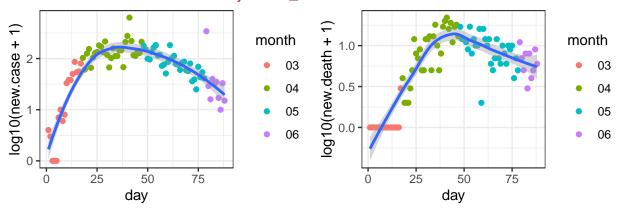


data source: https://github.com/nytimes/covid-19-data, day 1 is 03-15

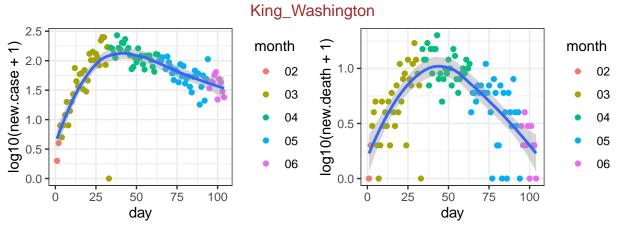


data source: https://github.com/nytimes/covid-19-data, day 1 is 03-12 Prince George's\_Maryland

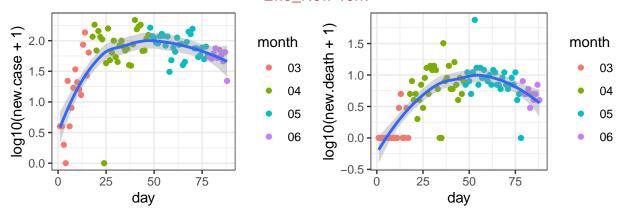




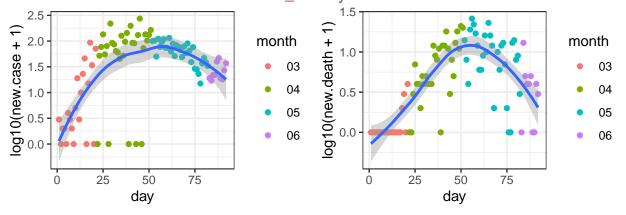
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-15



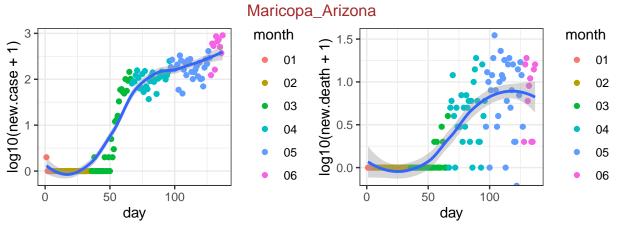
data source: https://github.com/nytimes/covid–19–data, day 1 is 02–28
Erie\_New York



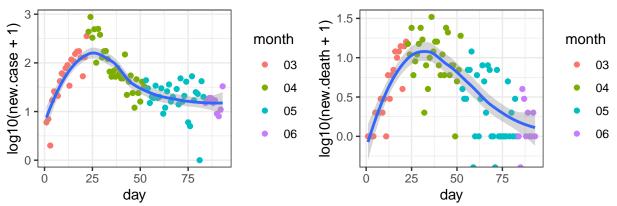
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-15
Bucks\_Pennsylvania



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

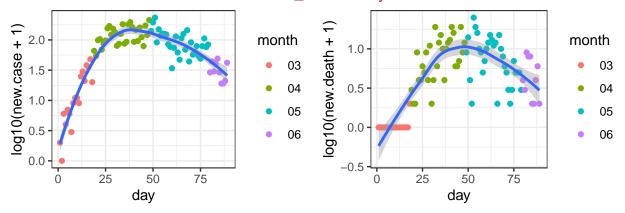


data source: https://github.com/nytimes/covid-19-data, day 1 is 01-26
Orleans\_Louisiana



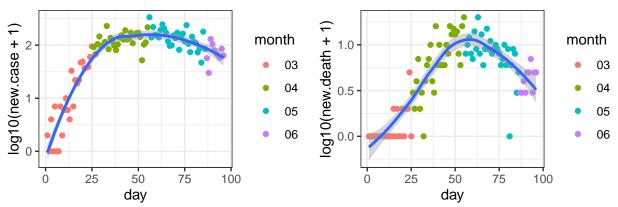
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-10

Mercer\_New Jersey

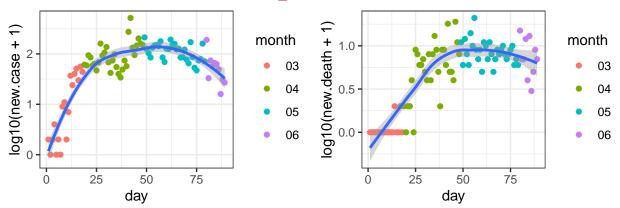


data source: https://github.com/nytimes/covid-19-data, day 1 is 03-14

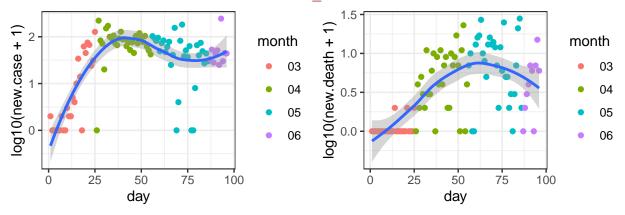




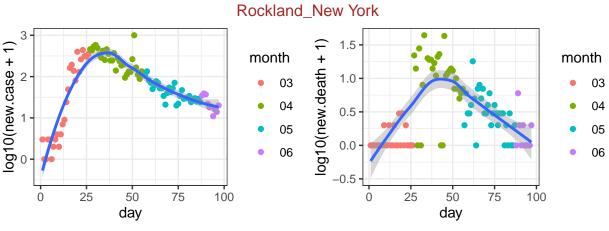
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-07
Bristol\_Massachusetts



data source: https://github.com/nytimes/covid-19-data, day 1 is 03-14 St. Louis\_Missouri

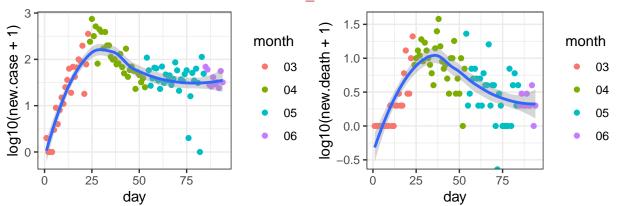


data source: https://github.com/nytimes/covid-19-data, day 1 is 03-07

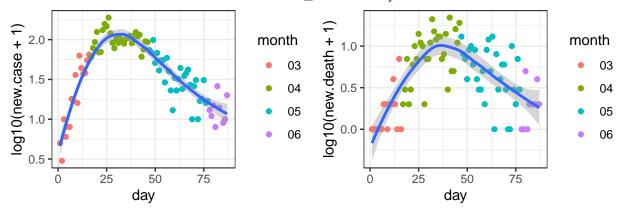


data source: https://github.com/nytimes/covid-19-data, day 1 is 03-06

Jefferson\_Louisiana



data source: https://github.com/nytimes/covid-19-data, day 1 is 03-09 Somerset\_New Jersey



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

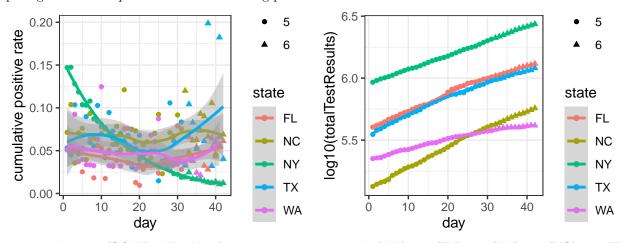
#### Fairfax\_Virginia log10(new.death + 1) log10(new.case + 1) month month 03 03 04 05 05 06 06 . 25 50 . 75 25 . 75 100 50 100 0 day day

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

## **COVID Tracking**

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 results 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 available 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 0611: NA(WA) 0.04(TX) 0.01(NY) 0.07(NC) 0.06(FL)

## 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
         /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
                                                           cowplot_1.0.0
## [21] dplyr_0.8.4
                        knitr_1.28
                                          vctrs_0.3.0
## [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
```