Exploration of COVID-19 tracking data from multiple resources

Wei Sun

2020-07-31

Contents

Introduction	1
JHU	2
time series data	2
daily reports data	6
NY Times	7
state level data	
county level data	18
COVID Tracking	36
Session information	37

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

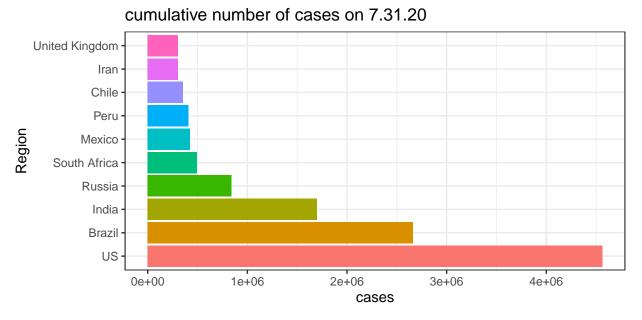
JHU

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

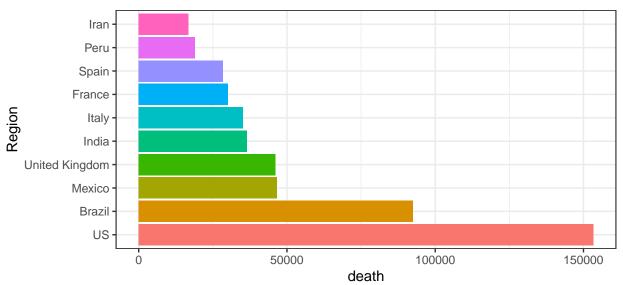
time series data

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.

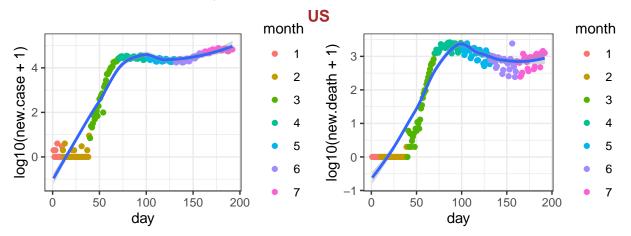
Here is the list of 10 records with the largest number of cases or deaths on the most recent date.



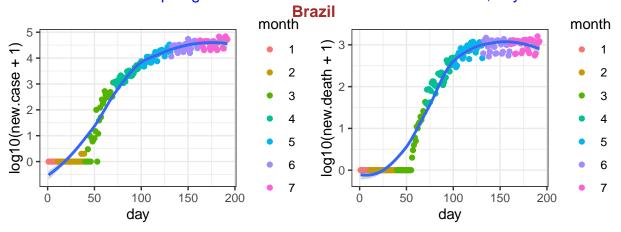
cumulative number of deaths on 7.31.20



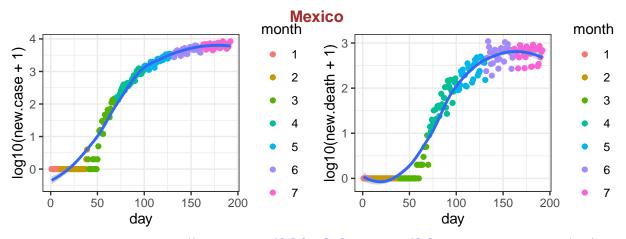
Next, I check for each country/region, what is the number of new cases/deaths? This data is important to understand what is the trend under different situations, e.g., population density, social distance policies etc. Here I checked the top 10 countries/regions with the highest number of deaths.



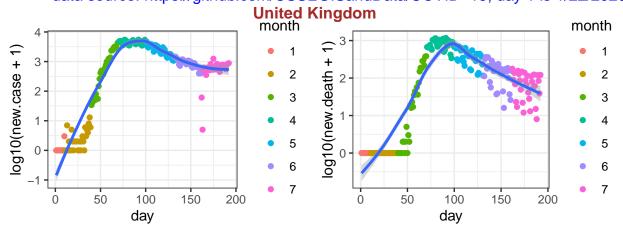
data source: https://github.com/CSSEGISandData/COVID-19, day 1 is 1/22/2020



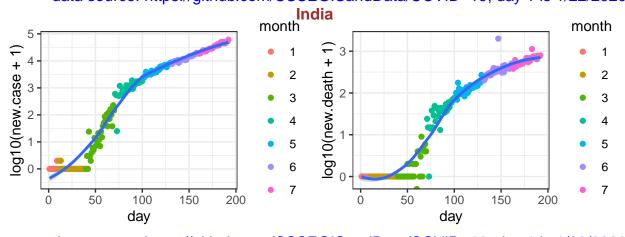
data source: https://github.com/CSSEGISandData/COVID-19, day 1 is 1/22/2020



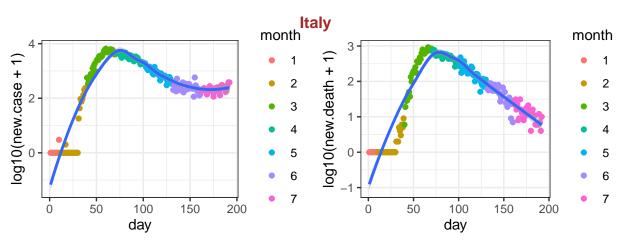
data source: https://github.com/CSSEGISandData/COVID-19, day 1 is 1/22/2020

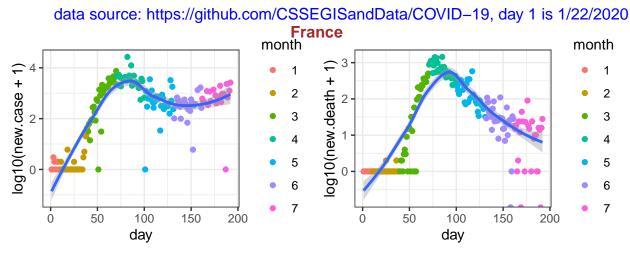


data source: https://github.com/CSSEGISandData/COVID-19, day 1 is 1/22/2020

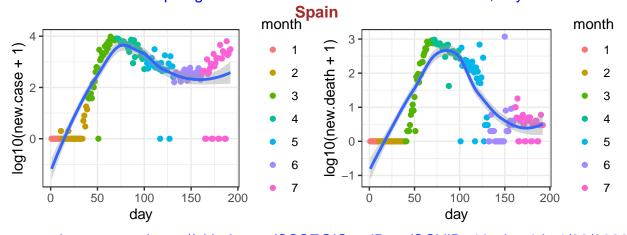


data source: https://github.com/CSSEGISandData/COVID-19, day 1 is 1/22/2020

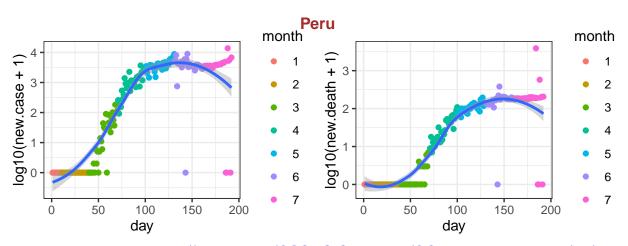




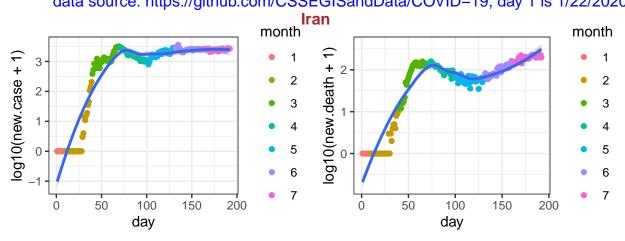
data source: https://github.com/CSSEGISandData/COVID-19, day 1 is 1/22/2020



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



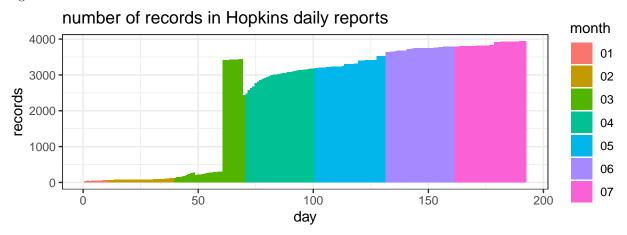
data source: https://github.com/CSSEGISandData/COVID-19, day 1 is 1/22/2020



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

daily reports data

The raw data from Hopkins are in the format of daily reports with one file per day. More recent files (since March 22nd) inlcude information from individual states of US or individual counties, as shown in the following figure. So I turn to NY Times data for informatoin of individual states or counties.



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

NY Times

The data from NY Times are saved in two text files, one for state level information and the other one for county level information.

The currente date is

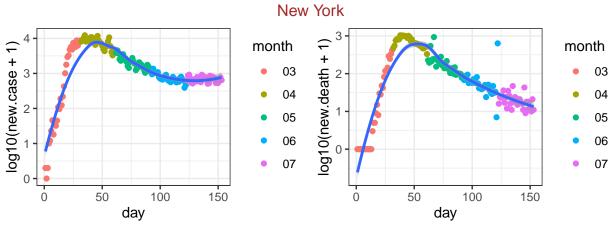
[1] "2020-07-31"

state level data

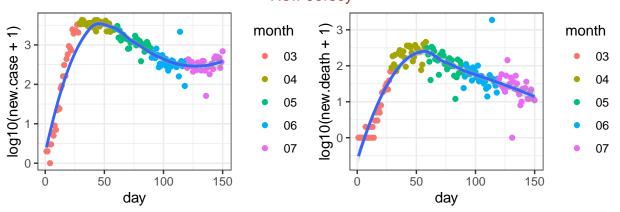
First check the 30 states with the largest number of deaths.

##		date	state		fips	cases	${\tt deaths}$
##	8298	2020-07-31		New York	36	419723	32372
##	8296	2020-07-31	Ne	ew Jersey	34	183535	15819
##	8269	2020-07-31	Ca	alifornia	6	502273	9222
##	8287	2020-07-31	Massa	Massachusetts		117612	8609
##	8279	2020-07-31		${\tt Illinois}$	17	180701	7703
##	8311	2020-07-31		Texas		441688	7265
##	8305	2020-07-31	Pennsylvania		42	116787	7261
##	8274	2020-07-31	Florida		12	470378	6842
##	8288	2020-07-31		Michigan	26	90752	6453
##	8271	2020-07-31	Cor	nnecticut	9	49810	4432
##	8284	2020-07-31	I	Louisiana	22	116394	3949
##	8267	2020-07-31		Arizona	4	174108	3695
##	8275	2020-07-31		Georgia	13	171342	3674
##	8286	2020-07-31		${\tt Maryland}$	24	88907	3493
##	8302	2020-07-31		Ohio	39	91159	3489
##	8280	2020-07-31		Indiana	18	67800	2965
##	8315	2020-07-31		Virginia	51	89888	2174
##	8299	2020-07-31	North	${\tt Carolina}$	37	122433	1947
##	8270	2020-07-31		${\tt Colorado}$	8	46948	1841
##	8308	2020-07-31	${\tt South}$	${\tt Carolina}$	45	89016	1712
##	8290	2020-07-31	Mississippi		28	58747	1663
##	8316	2020-07-31	Washington		53	58724	1654
##	8289	2020-07-31	Minnesota		27	54503	1640
##	8265	2020-07-31	Alabama		1	87723	1580
##	8291	2020-07-31	Missouri		29	51045	1305
##	8310	2020-07-31	-	Tennessee	47	103144	1047
##	8307	2020-07-31	Rho	de Island	44	19022	1007
##	8318	2020-07-31	I	Visconsin	55	57020	943
##	8281	2020-07-31		Iowa	19	44753	867
##	8294	2020-07-31	Nevada		32	48142	831

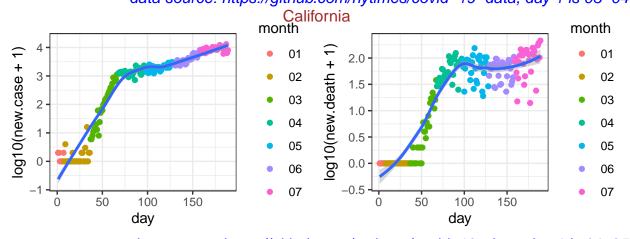
For these 20 states, I check the number of new cases and the number of new deaths. Part of the reason for such checking is to identify whether there is any similarity on such patterns. For example, could you use the pattern seen from Italy to predict what happen in an individual state, and what are the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



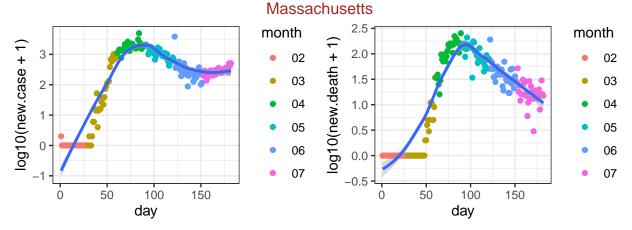
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-01
New Jersey



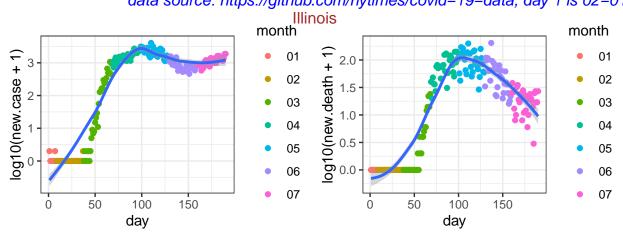
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-04



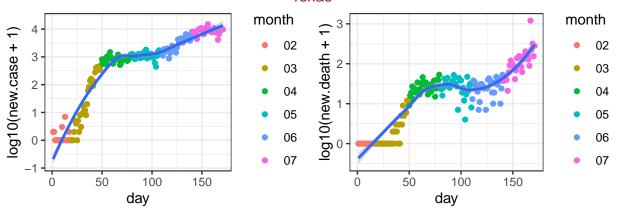
data source: https://github.com/nytimes/covid-19-data, day 1 is 01-25



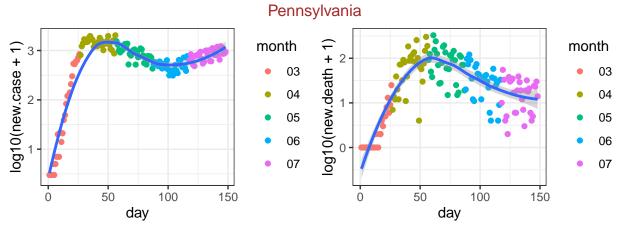
data source: https://github.com/nytimes/covid-19-data, day 1 is 02-01



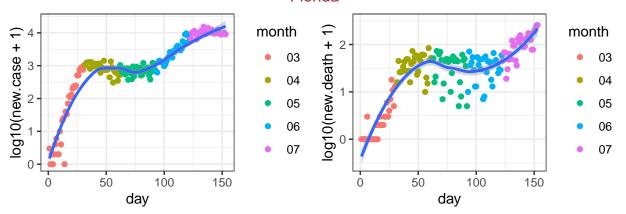
data source: https://github.com/nytimes/covid-19-data, day 1 is 01-24
Texas



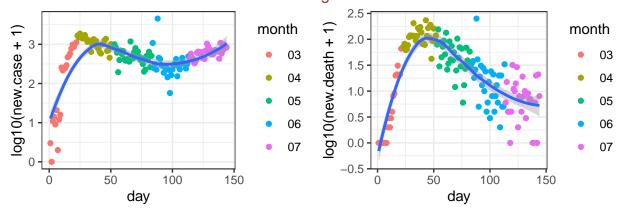
data source: https://github.com/nytimes/covid-19-data, day 1 is 02-12



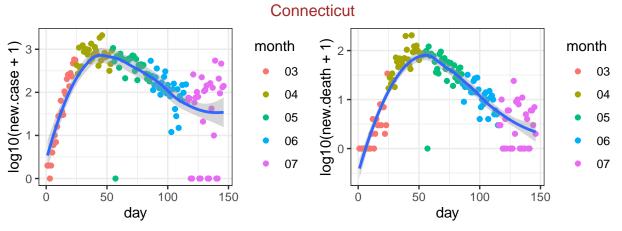
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-06 Florida



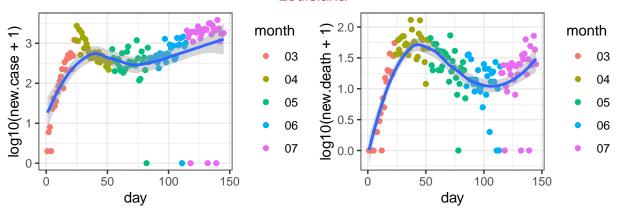
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-01 Michigan



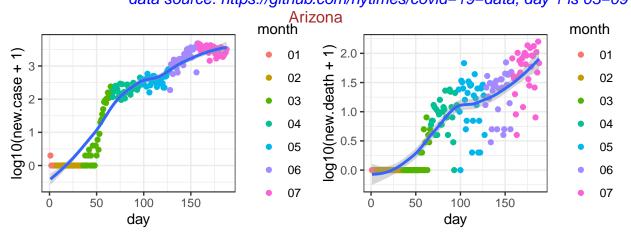
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-10



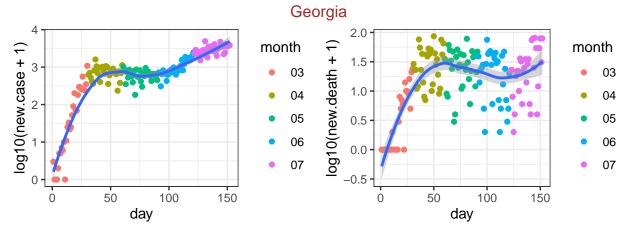
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-08 Louisiana



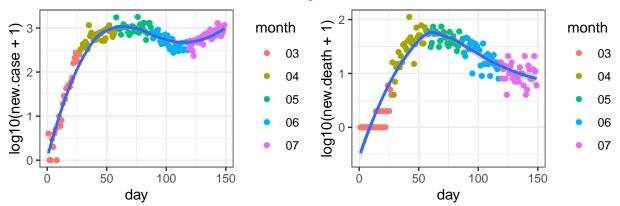
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-09



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

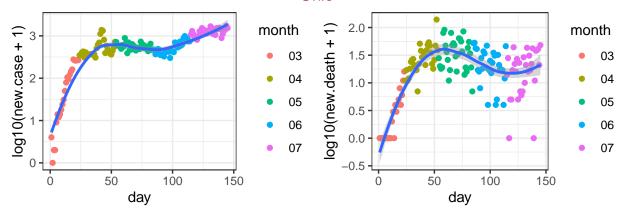


data source: https://github.com/nytimes/covid-19-data, day 1 is 03-02
Maryland

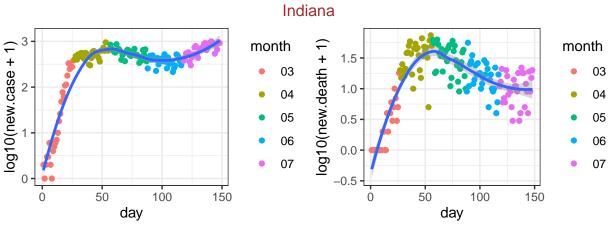


data source: https://github.com/nytimes/covid-19-data, day 1 is 03-05

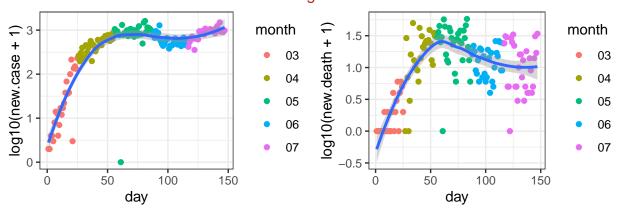
Ohio



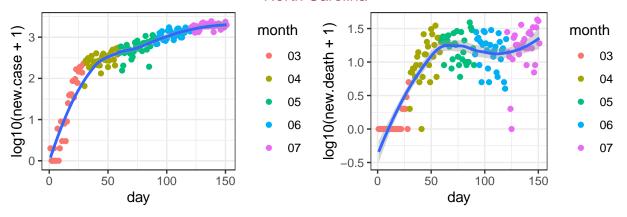
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-09



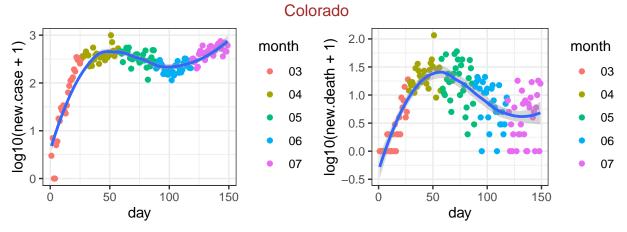
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-06 Virginia



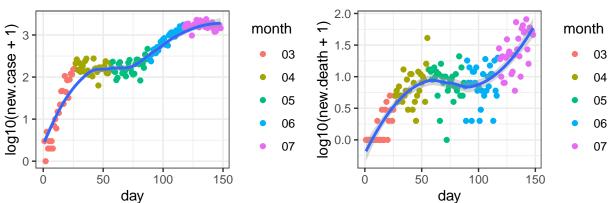
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-07 North Carolina



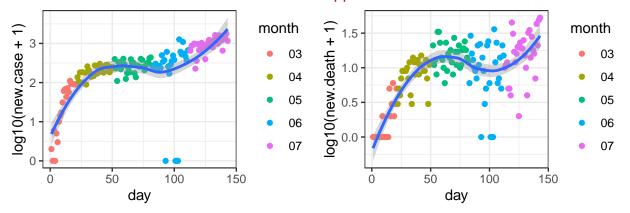
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-03



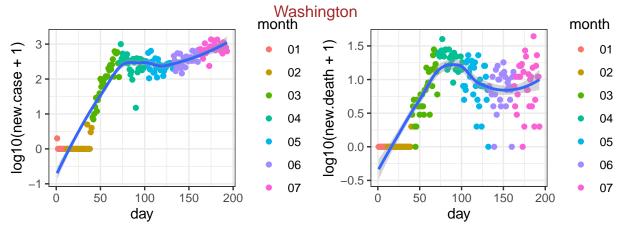
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-05 South Carolina



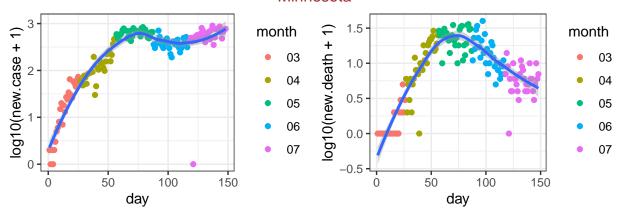
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-06 Mississippi



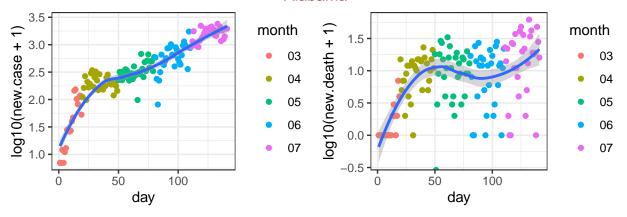
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-11



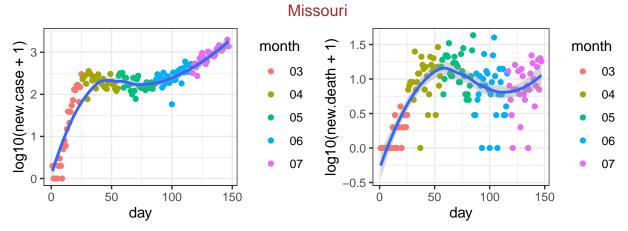
data source: https://github.com/nytimes/covid-19-data, day 1 is 01-21
Minnesota



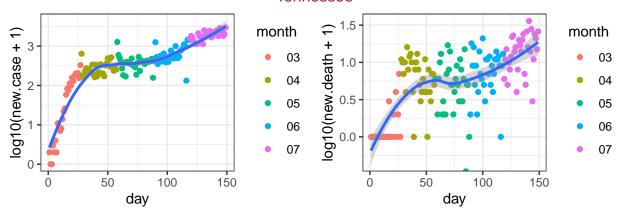
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-06
Alabama



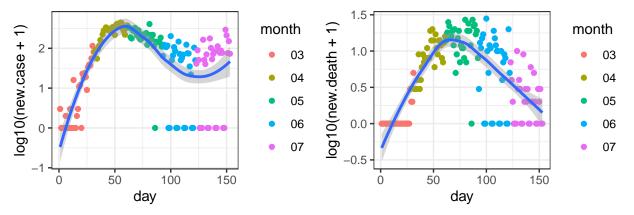
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-13



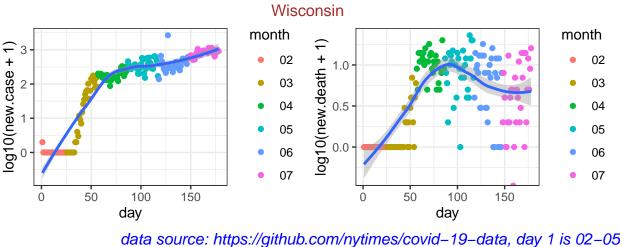
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-07
Tennessee



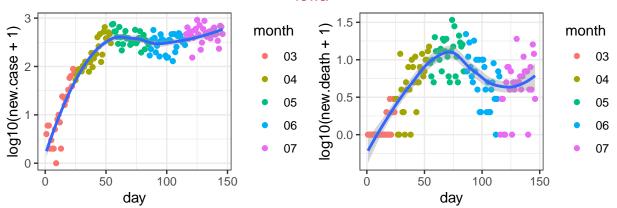
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-05 Rhode Island



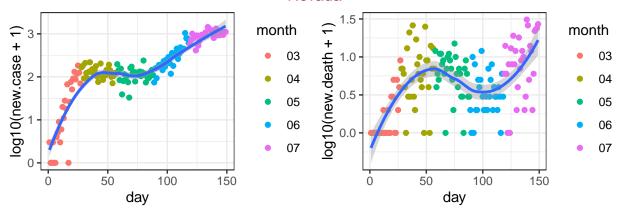
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-01



Iowa

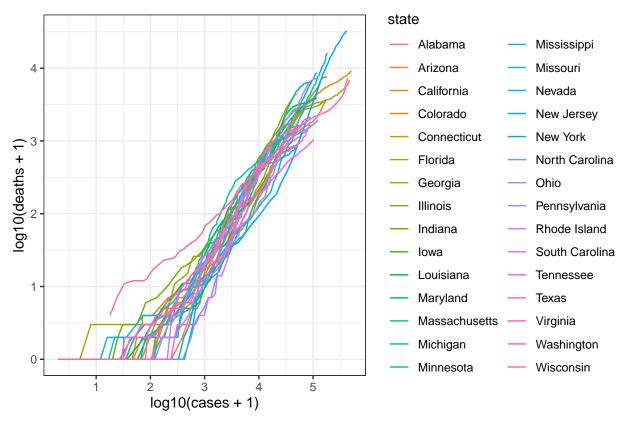


data source: https://github.com/nytimes/covid-19-data, day 1 is 03-08 Nevada



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

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	${\tt deaths}$
##	387837	2020-07-31	New York City	New York	NA	229834	23002
##	386596	2020-07-31	Cook	Illinois	17031	105493	4886
##	386189	2020-07-31	Los Angeles	California	6037	188481	4621
##	387303	2020-07-31	Wayne	Michigan	26163	27092	2804
##	387836	2020-07-31	Nassau	New York	36059	43203	2706
##	387761	2020-07-31	Essex	New Jersey	34013	19731	2102
##	386087	2020-07-31	Maricopa	Arizona	4013	117293	2054
##	387856	2020-07-31	Suffolk	New York	36103	43224	2044
##	387756	2020-07-31	Bergen	New Jersey	34003	20727	2040
##	387214	2020-07-31	Middlesex	${\tt Massachusetts}$	25017	25706	1978
##	388273	2020-07-31	Philadelphia	Pennsylvania	42101	30354	1690
##	386348	2020-07-31	Miami-Dade	Florida	12086	118461	1611
##	387864	2020-07-31	Westchester	New York	36119	35932	1578
##	387763	2020-07-31	Hudson	New Jersey	34017	19647	1500
##	386293	2020-07-31	Hartford	Connecticut	9003	12645	1412
##	386292	2020-07-31	Fairfield	Connecticut	9001	17793	1406
##	387766	2020-07-31	Middlesex	New Jersey	34023	17879	1404
##	387774	2020-07-31	Union	New Jersey	34039	16694	1347
##	388681	2020-07-31	Harris	Texas	48201	72964	1258
##	387770	2020-07-31	Passaic	New Jersey	34031	17580	1242
##	387210	2020-07-31	Essex	${\tt Massachusetts}$	25009	17232	1177
##	387283	2020-07-31	Oakland	Michigan	26125	14614	1126

```
## 386296 2020-07-31
                            New Haven
                                         Connecticut 9009
                                                             13041
                                                                     1101
  387218 2020-07-31
                              Suffolk Massachusetts 25025
                                                             21222
                                                                     1055
                                          New Jersey 34029
  387769 2020-07-31
                                                             10482
                                                                     1014
  387220 2020-07-31
                            Worcester Massachusetts 25027
                                                                      990
                                                             13322
   387216 2020-07-31
                              Norfolk Massachusetts 25021
                                                             10236
                                                                       984
  387270 2020-07-31
                               Macomb
                                            Michigan 26099
                                                              9689
                                                                       941
  387767 2020-07-31
                             Monmouth
                                          New Jersey 34025
                                                                       858
                                                             10157
## 388268 2020-07-31
                                        Pennsylvania 42091
                           Montgomery
                                                              9761
                                                                       848
   387768 2020-07-31
                               Morris
                                          New Jersey 34027
                                                              7287
                                                                       828
   386355 2020-07-31
                           Palm Beach
                                             Florida 12099
                                                                      815
                                                             33272
   387331 2020-07-31
                             Hennepin
                                           Minnesota 27053
                                                             17316
                                                                       813
## 388372 2020-07-31
                                                                       808
                           Providence
                                        Rhode Island 44007
                                                             14549
   387196 2020-07-31
                                            Maryland 24031
                                                                       789
                           Montgomery
                                                             17607
  386732 2020-07-31
                                             Indiana 18097
                                                             14575
                                                                       766
                               Marion
  387197 2020-07-31 Prince George's
                                            Maryland 24033
                                                             22869
                                                                       737
  388245 2020-07-31
                             Delaware
                                        Pennsylvania 42045
                                                              8669
                                                                       730
  387217 2020-07-31
                             Plymouth Massachusetts 25023
                                                              9086
                                                                       710
## 386311 2020-07-31
                              Broward
                                             Florida 12011
                                                             55411
                                                                       709
## 387212 2020-07-31
                              Hampden Massachusetts 25013
                                                              7404
                                                                       696
## 386203 2020-07-31
                            Riverside
                                          California 6065
                                                             37308
                                                                       695
  387730 2020-07-31
                                Clark
                                              Nevada 32003
                                                             41245
                                                                       687
  388637 2020-07-31
                               Dallas
                                               Texas 48113
                                                             49976
                                                                       671
## 389028 2020-07-31
                                          Washington 53033
                                                             15216
                                                                       668
                                 King
   387577 2020-07-31
                            St. Louis
                                            Missouri 29189
                                                             12880
                                                                       650
## 388688 2020-07-31
                              Hidalgo
                                               Texas 48215
                                                             16945
                                                                       644
                              Bristol Massachusetts 25005
## 387208 2020-07-31
                                                              9073
                                                                       623
## 387822 2020-07-31
                                 Erie
                                            New York 36029
                                                              8492
                                                                       621
## 386200 2020-07-31
                                          California 6059
                                                             36196
                                                                       618
                               Orange
```

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

100

day

150

50

← 3 month month og10(new.case + 1) log10(new.death + 03 03 2 04 04 2 05 05 06 06 07 07

New York City_New York

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

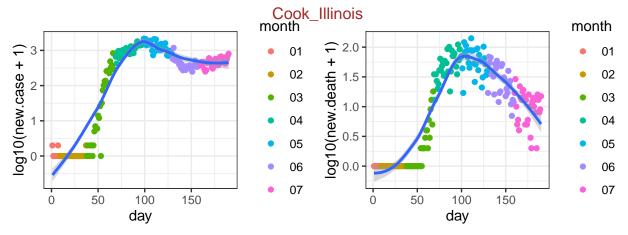
50

100

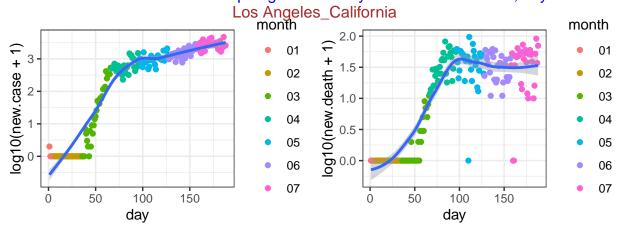
day

150

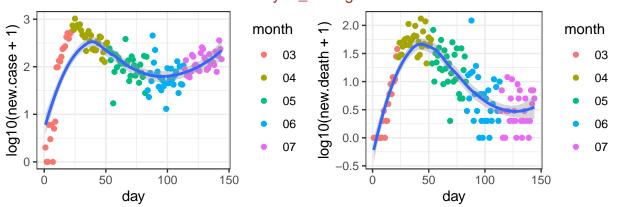
0



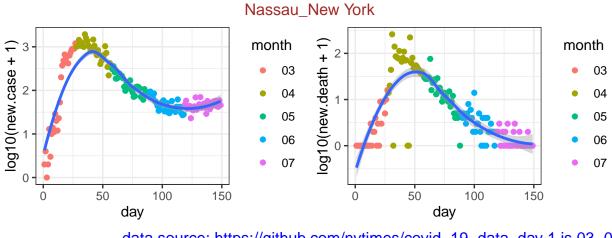
data source: https://github.com/nytimes/covid-19-data, day 1 is 01-24



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

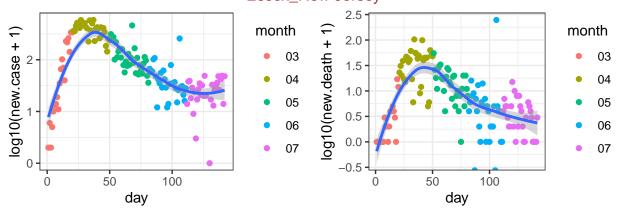


data source: https://github.com/nytimes/covid-19-data, day 1 is 03-10

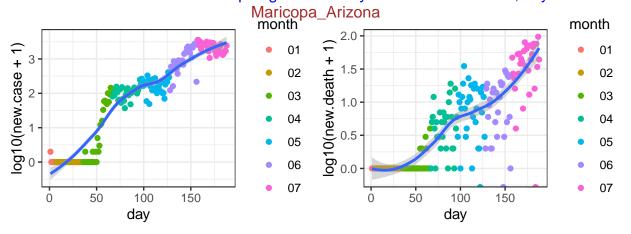


data source: https://github.com/nytimes/covid-19-data, day 1 is 03-05

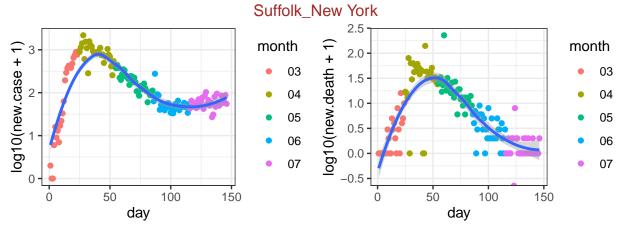
Essex_New Jersey



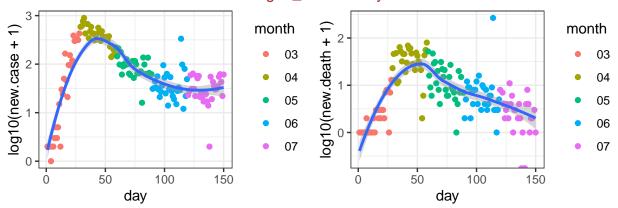
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-12



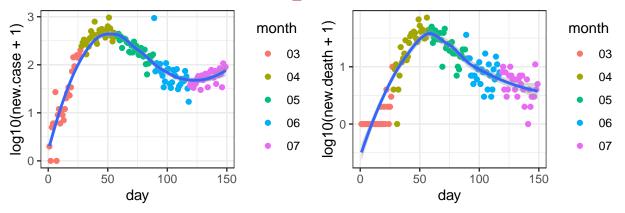
data source: https://github.com/nytimes/covid-19-data, day 1 is 01-26



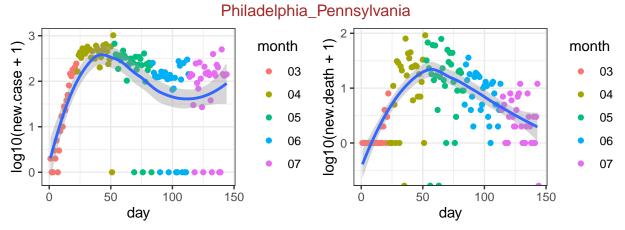
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-08
Bergen_New Jersey



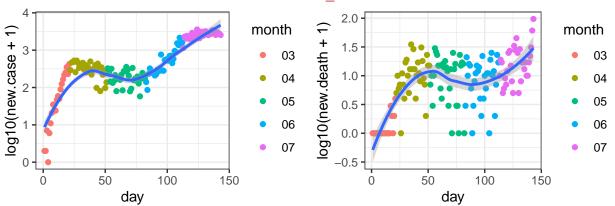
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-04 Middlesex_Massachusetts



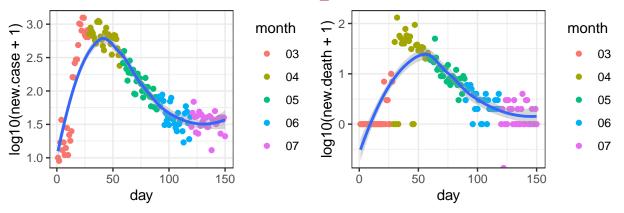
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-05



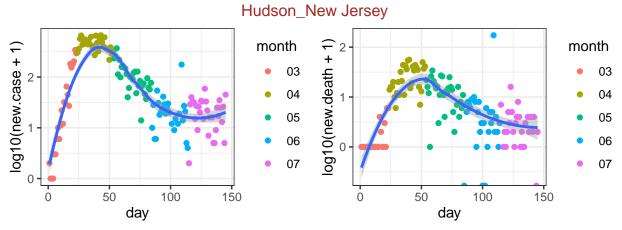
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-10 Miami-Dade_Florida



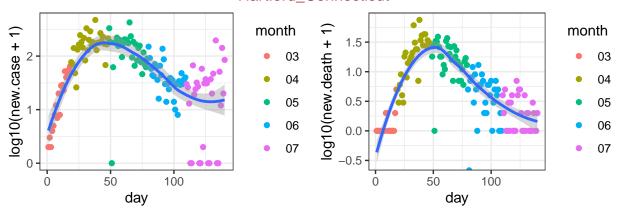
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-11 Westchester_New York



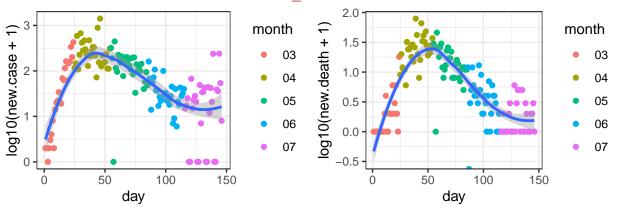
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-04



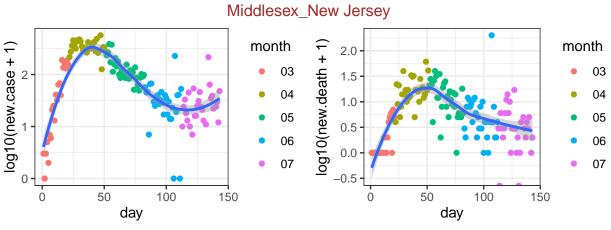
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-09
Hartford_Connecticut



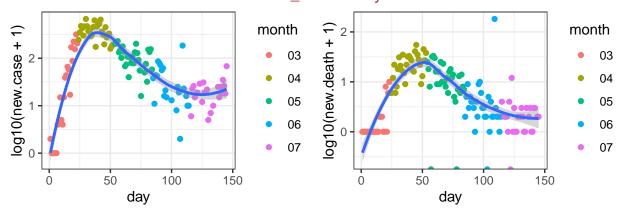
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-14 Fairfield_Connecticut



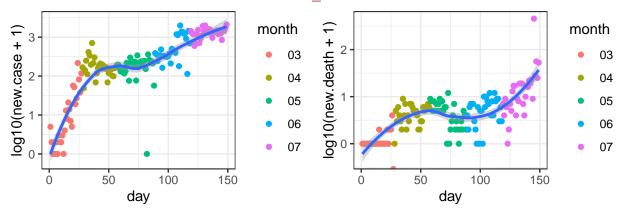
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-08



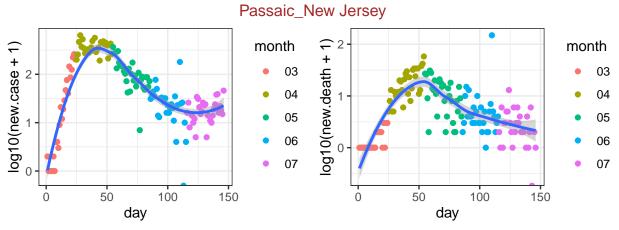
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-11 Union_New Jersey



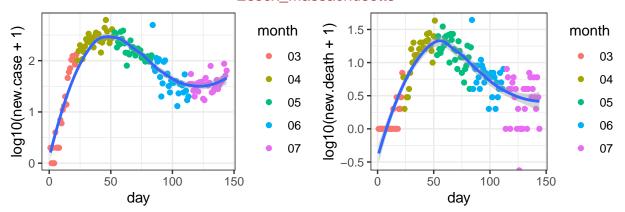
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-09
Harris_Texas



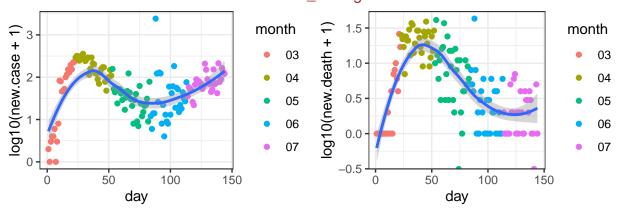
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-05



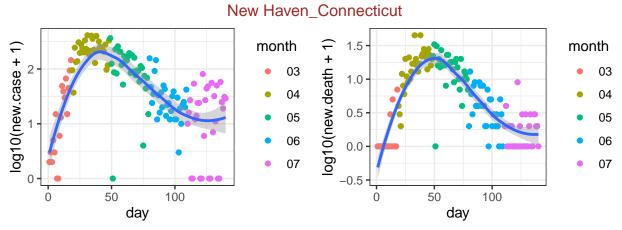
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-08
Essex_Massachusetts



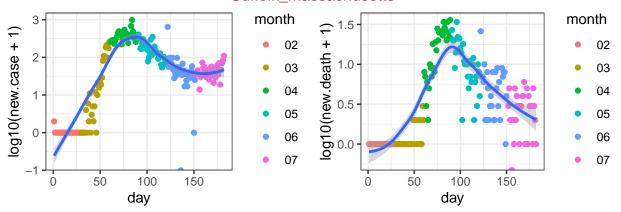
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-10 Oakland_Michigan



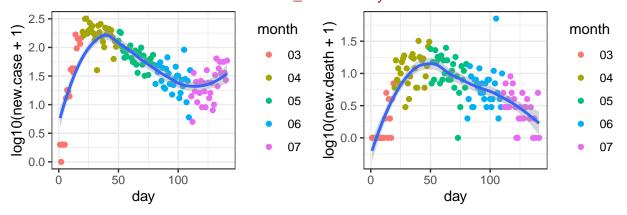
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-10



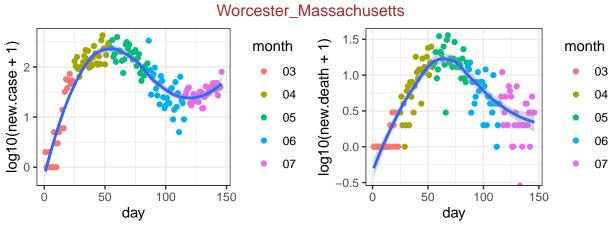
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-14 Suffolk_Massachusetts



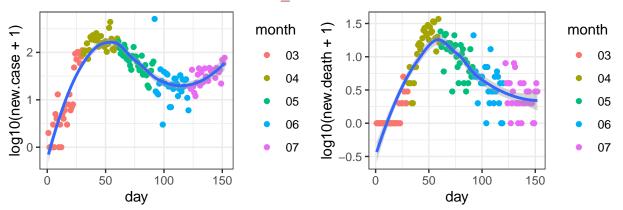
data source: https://github.com/nytimes/covid-19-data, day 1 is 02-01
Ocean_New Jersey



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

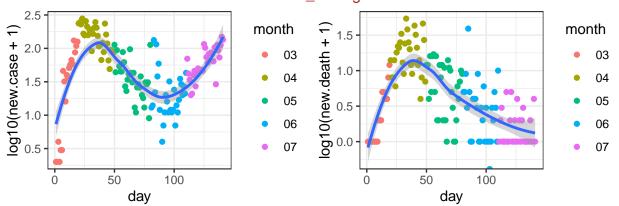


data source: https://github.com/nytimes/covid-19-data, day 1 is 03-08 Norfolk_Massachusetts

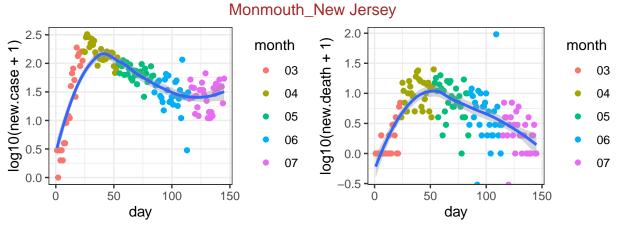


data source: https://github.com/nytimes/covid-19-data, day 1 is 03-02

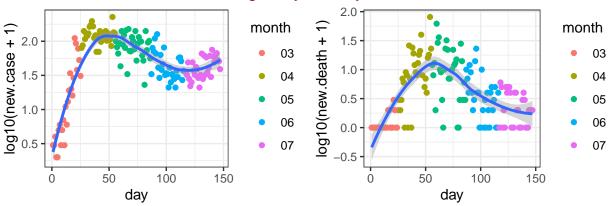
Macomb_Michigan



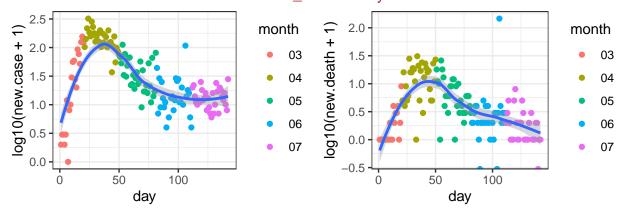
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-13



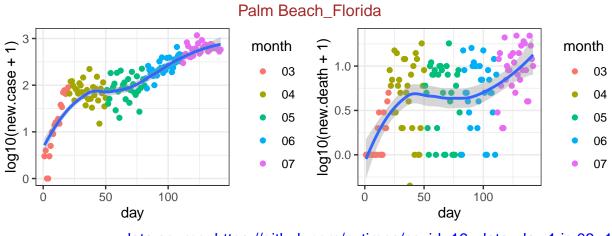
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-09 Montgomery_Pennsylvania



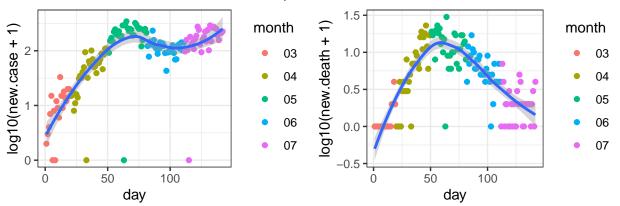
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-07
Morris_New Jersey



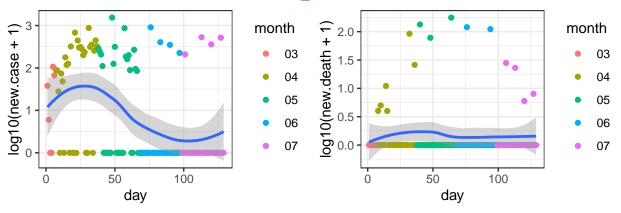
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-12



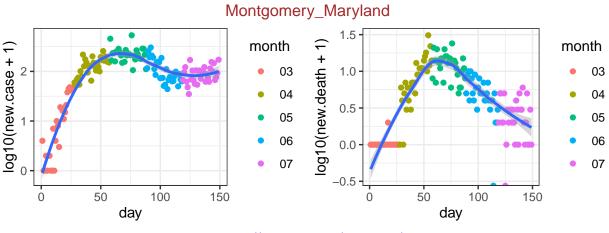
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-12
Hennepin_Minnesota



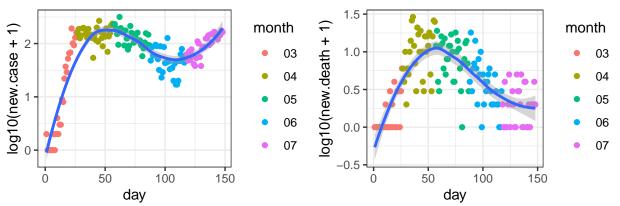
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-12 Providence_Rhode Island



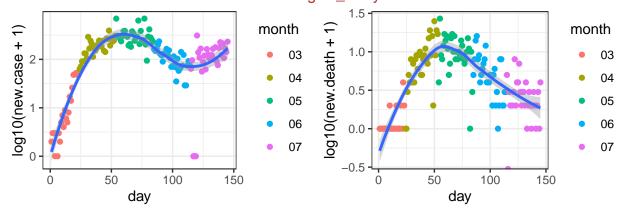
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-25



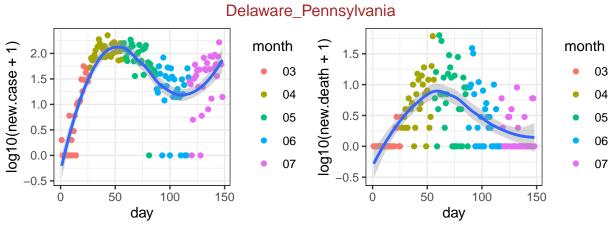
data source: https://github.com/nytimes/covid–19–data, day 1 is 03–05
Marion_Indiana



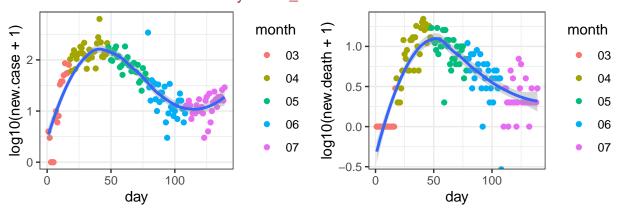
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-06 Prince George's_Maryland



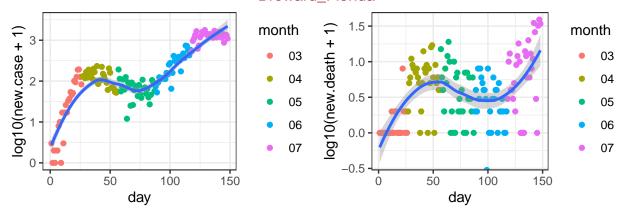
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-09



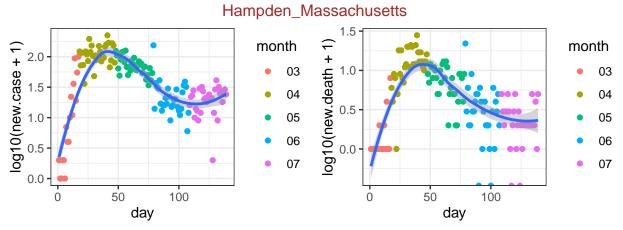
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-06 Plymouth_Massachusetts



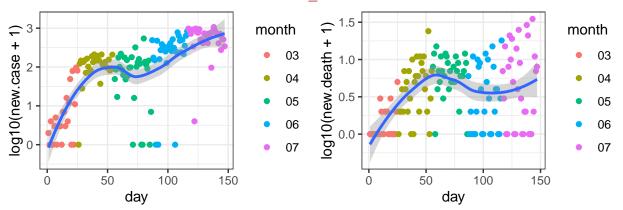
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-15
Broward_Florida

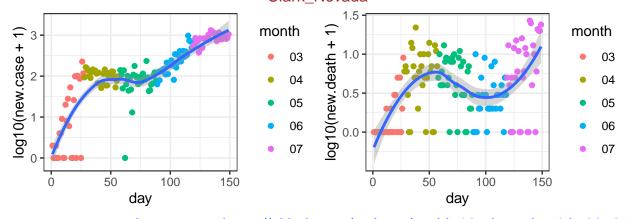


data source: https://github.com/nytimes/covid-19-data, day 1 is 03-06

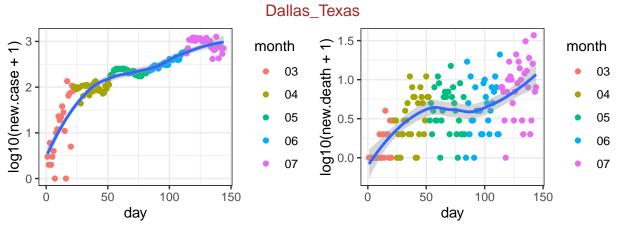


data source: https://github.com/nytimes/covid-19-data, day 1 is 03-15 Riverside_California

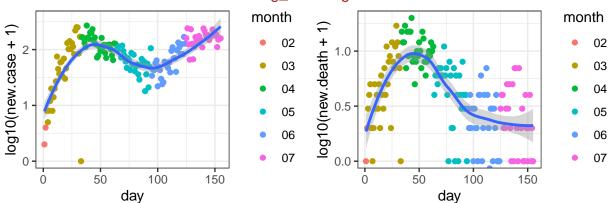




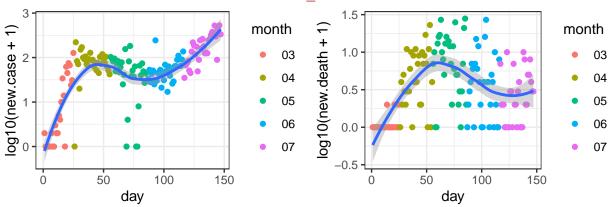
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-05



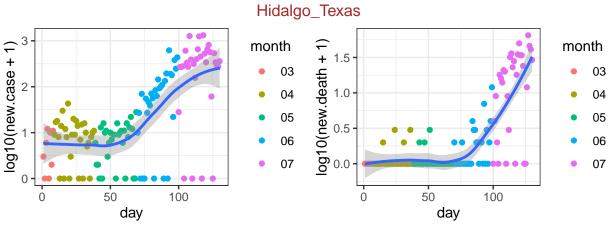
data source: https://github.com/nytimes/covid-19-data, day 1 is 03-10 King_Washington



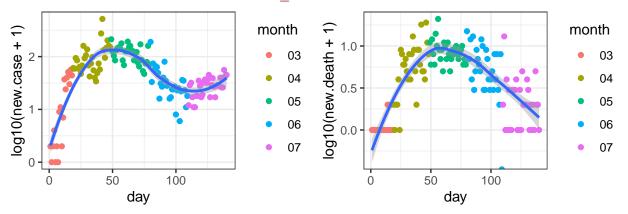
data source: https://github.com/nytimes/covid-19-data, day 1 is 02-28 St. Louis_Missouri

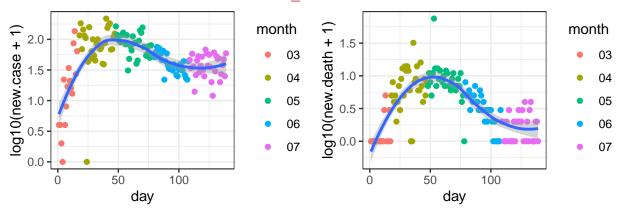


data source: https://github.com/nytimes/covid-19-data, day 1 is 03-07

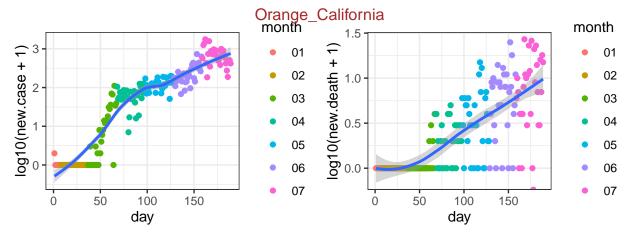


data source: https://github.com/nytimes/covid-19-data, day 1 is 03-24
Bristol_Massachusetts





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

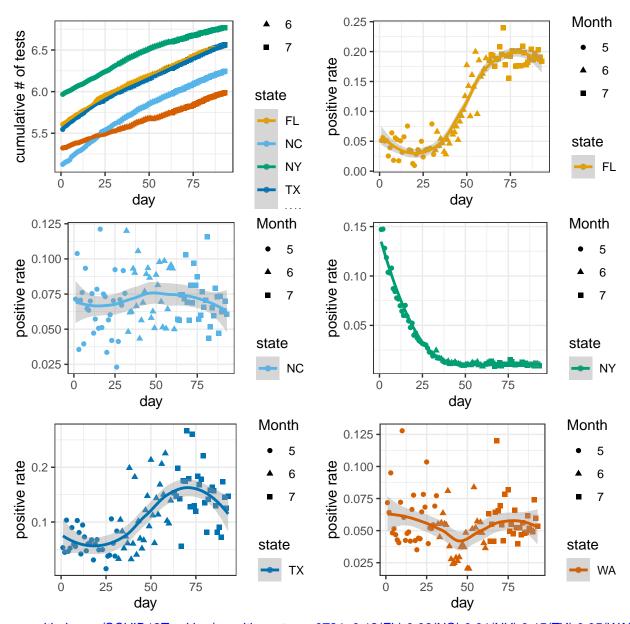


data source: https://github.com/nytimes/covid-19-data, day 1 is 01-25

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 0731: 0.18(FL) 0.06(NC) 0.01(NY) 0.15(TX) 0.05(WA)

Session information

```
sessionInfo()
```

```
## R version 3.6.2 (2019-12-12)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS Catalina 10.15.5
##
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib
##
## locale:
```

```
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/c/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats
                graphics grDevices utils
                                              datasets methods
                                                                  base
## other attached packages:
## [1] httr 1.4.1
                    ggpubr_0.2.5 magrittr_1.5 ggplot2_3.3.1
##
## loaded via a namespace (and not attached):
## [1] Rcpp_1.0.3
                        pillar_1.4.3
                                          compiler_3.6.2
                                                          tools_3.6.2
## [5] digest_0.6.23
                        lattice_0.20-38
                                         nlme_3.1-144
                                                          evaluate_0.14
## [9] lifecycle_0.2.0 tibble_3.0.1
                                          gtable_0.3.0
                                                          mgcv_1.8-31
## [13] pkgconfig_2.0.3 rlang_0.4.6
                                         Matrix_1.2-18
                                                          yaml_2.2.1
## [17] xfun_0.12
                        gridExtra_2.3
                                         withr_2.1.2
                                                          stringr_1.4.0
## [21] dplyr_0.8.4
                        knitr_1.28
                                          vctrs_0.3.0
                                                          cowplot_1.0.0
## [25] grid_3.6.2
                        tidyselect_1.0.0 glue_1.3.1
                                                          R6_2.4.1
## [29] rmarkdown_2.1
                        purrr_0.3.3
                                         farver_2.0.3
                                                          splines_3.6.2
                                                          assertthat_0.2.1
## [33] scales 1.1.0
                        ellipsis_0.3.0
                                         htmltools_0.4.0
## [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
```