Final Project

Hongyi Ji

hj312

Introduce

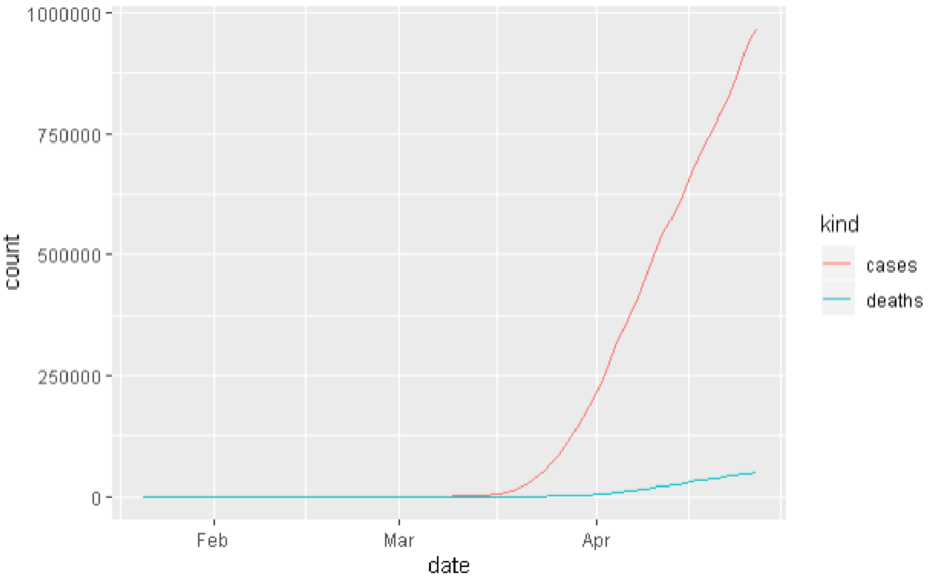
There are 2 datasets are uses in the report. The first one is the Coronavirus (Covid-19) Data in the United States. The data begins with the first reported coronavirus case in Washington State on Jan. 21, 2020 and ends on Apr. 26,2020. Data on cumulative coronavirus cases and deaths can be found in two files for states and counties. State-level data can be found in the states.csv file and County-level data can be found in the counties.csv file.

The second dataset is from the website of the US Census Bureau. It provides statistics for all states and counties, and for cities and towns with a population of 5,000 or more. There are many statistics such as population estimates, percent of female persons, percent of White alone, percent of foreign-born persons, etc.

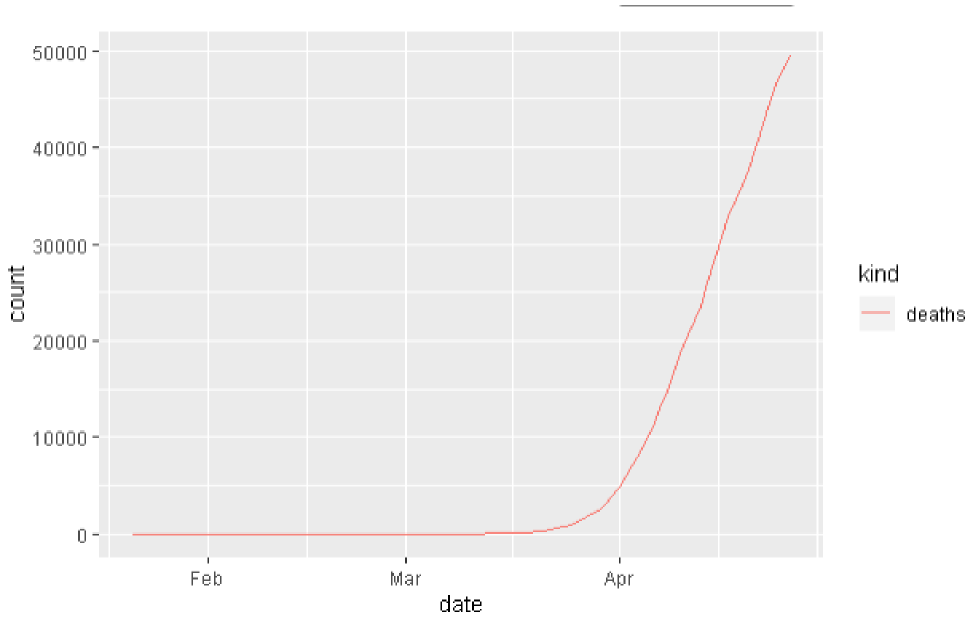
In this report, I will make some plots to show the cases and deaths trend of some states and counties at first. Then I will scrape and clean the data form the second source. After choosing and creating some indicators from the second source, I will try to show the relationships between the amounts of Covid-19 deaths and cases with those indicators and create some plots.

The trend of cases and deaths

Firstly, I make a plot to visualize the trend of cases and deaths on the country level. The picture below is the plot.

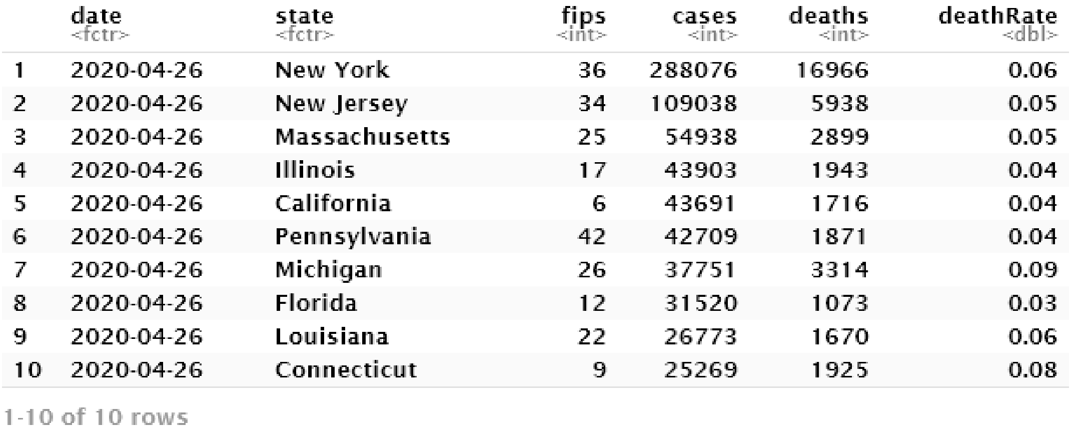


We can see that the count of cases starts to grow drastically from the middle of March. The counts of deaths begin to increase obviously from the beginning of April, a bit later than the count of cases. I also make another plot to show the increase of deaths more clearly. We can find that the counts of case and death both grow exponentially.



State level and counties level data

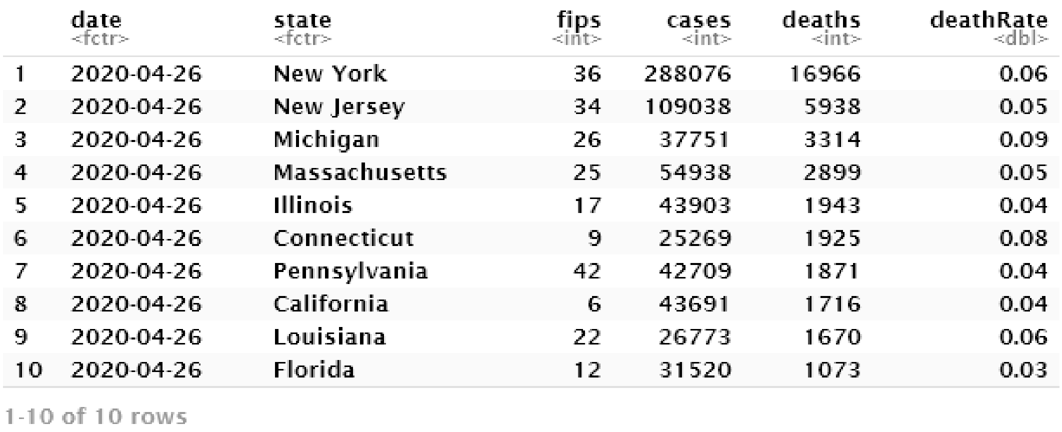
Then I focus on the data on the state level. I order the states by the counts of cases and deaths and then choose the top 10 states to make the data frame. The table below shows the states ordered by the cases.



	date <fctr>	state <fctr>	fips <int>	cases <int>	deaths <int>	deathRate <dbl>
1	2020-04-26	New York	36	288076	16966	0.06
2	2020-04-26	New Jersey	34	109038	5938	0.05
3	2020-04-26	Massachusetts	25	54938	2899	0.05
4	2020-04-26	Illinois	17	43903	1943	0.04
5	2020-04-26	California	6	43691	1716	0.04
6	2020-04-26	Pennsylvania	42	42709	1871	0.04
7	2020-04-26	Michigan	26	37751	3314	0.09
8	2020-04-26	Florida	12	31520	1073	0.03
9	2020-04-26	Louisiana	22	26773	1670	0.06
10	2020-04-26	Connecticut	9	25269	1925	0.08

1-10 of 10 rows

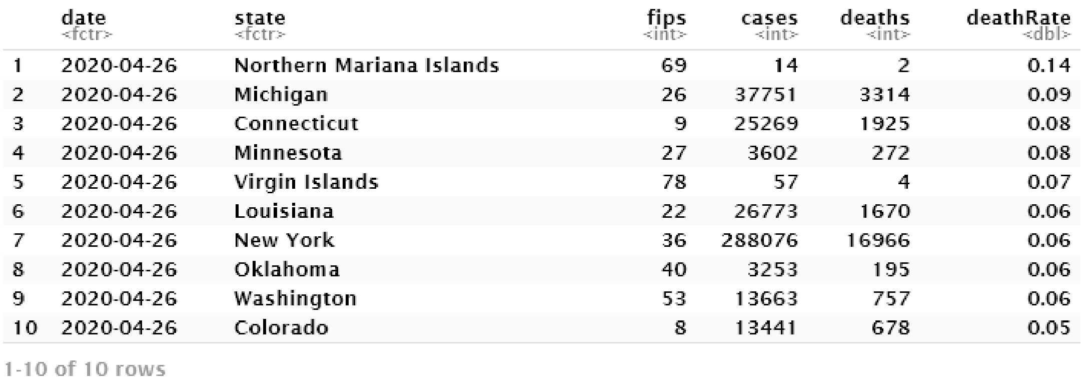
The table below shows the states ordered by the deaths.



	date <fctr>	state <fctr>	fips <int>	cases <int>	deaths <int>	deathRate <dbl>
1	2020-04-26	New York	36	288076	16966	0.06
2	2020-04-26	New Jersey	34	109038	5938	0.05
3	2020-04-26	Michigan	26	37751	3314	0.09
4	2020-04-26	Massachusetts	25	54938	2899	0.05
5	2020-04-26	Illinois	17	43903	1943	0.04
6	2020-04-26	Connecticut	9	25269	1925	0.08
7	2020-04-26	Pennsylvania	42	42709	1871	0.04
8	2020-04-26	California	6	43691	1716	0.04
9	2020-04-26	Louisiana	22	26773	1670	0.06
10	2020-04-26	Florida	12	31520	1073	0.03

1-10 of 10 rows

In the original dataset, only cases and deaths in included. To evaluate the medical condition of every state against the virus, I create a new column in the table. The new column is death rate and it equals deaths/cases. Then I order the states by their death rate. The result is shown below.

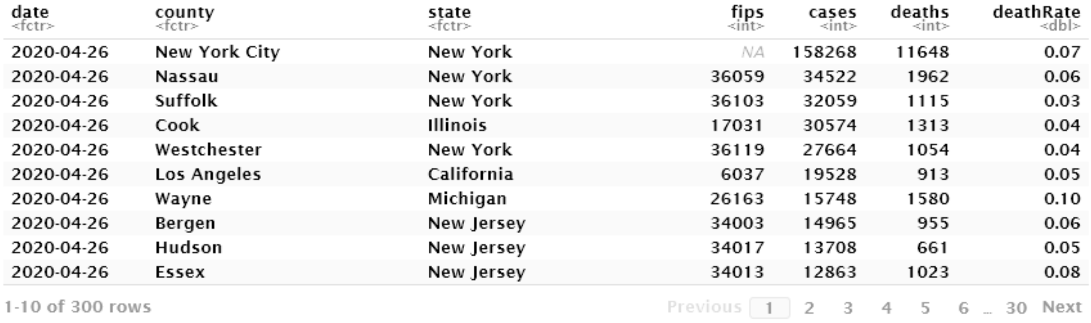


	date <fctr>	state <fctr>	fips <int>	cases <int>	deaths <int>	deathRate <dbl>
1	2020-04-26	Northern Mariana Islands	69	14	2	0.14
2	2020-04-26	Michigan	26	37751	3314	0.09
3	2020-04-26	Connecticut	9	25269	1925	0.08
4	2020-04-26	Minnesota	27	3602	272	0.08
5	2020-04-26	Virgin Islands	78	57	4	0.07
6	2020-04-26	Louisiana	22	26773	1670	0.06
7	2020-04-26	New York	36	288076	16966	0.06
8	2020-04-26	Oklahoma	40	3253	195	0.06
9	2020-04-26	Washington	53	13663	757	0.06
10	2020-04-26	Colorado	8	13441	678	0.05

1-10 of 10 rows

The death rate of Michigan, Connecticut and Minnesota is much higher than the other states'. Since the cases amount of North Mariana Islands and the Virgin Islands is small, the death rate of them is not so meaningful. However, the virus could be disastrous to the islands regions and countries if they do not have enough tests and fail to control the spread of the virus.

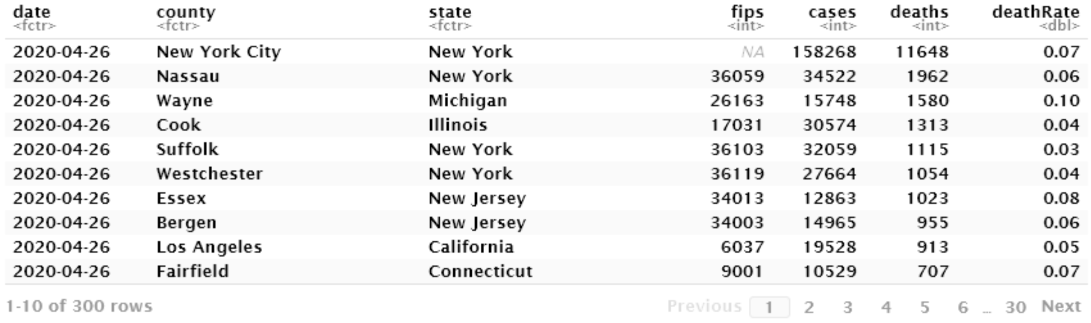
After the state level data, I focus on the county level data. To make the result be more representative, I choose the top 300 counties with the most cases. Firstly, I arrange the states by the amount of cases. The result is below.



date <fctr>	county <fctr>	state <fctr>	fips <int>	cases <int>	deaths <int>	deathRate <dbl>
2020-04-26	New York City	New York	NA	158268	11648	0.07
2020-04-26	Nassau	New York	36059	34522	1962	0.06
2020-04-26	Suffolk	New York	36103	32059	1115	0.03
2020-04-26	Cook	Illinois	17031	30574	1313	0.04
2020-04-26	Westchester	New York	36119	27664	1054	0.04
2020-04-26	Los Angeles	California	6037	19528	913	0.05
2020-04-26	Wayne	Michigan	26163	15748	1580	0.10
2020-04-26	Bergen	New Jersey	34003	14965	955	0.06
2020-04-26	Hudson	New Jersey	34017	13708	661	0.05
2020-04-26	Essex	New Jersey	34013	12863	1023	0.08

1-10 of 300 rows Previous 1 2 3 4 5 6 ... 30 Next

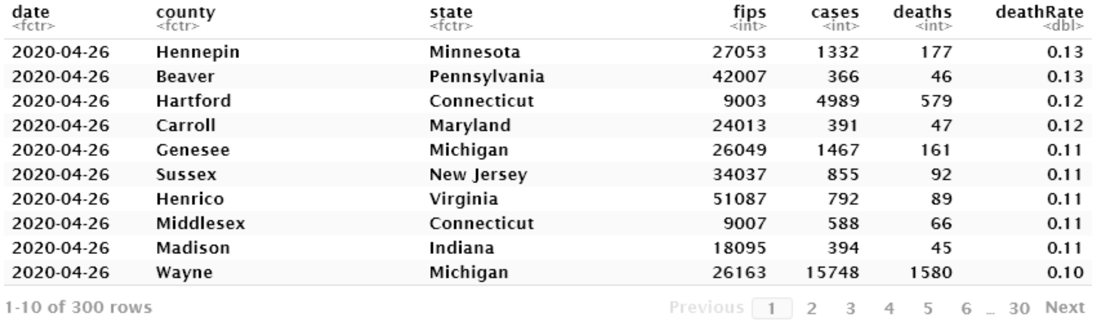
Secondly, I order the counties by the count of deaths. The result is below.



date <fctr>	county <fctr>	state <fctr>	fips <int>	cases <int>	deaths <int>	deathRate <dbl>
2020-04-26	New York City	New York	NA	158268	11648	0.07
2020-04-26	Nassau	New York	36059	34522	1962	0.06
2020-04-26	Wayne	Michigan	26163	15748	1580	0.10
2020-04-26	Cook	Illinois	17031	30574	1313	0.04
2020-04-26	Suffolk	New York	36103	32059	1115	0.03
2020-04-26	Westchester	New York	36119	27664	1054	0.04
2020-04-26	Essex	New Jersey	34013	12863	1023	0.08
2020-04-26	Bergen	New Jersey	34003	14965	955	0.06
2020-04-26	Los Angeles	California	6037	19528	913	0.05
2020-04-26	Fairfield	Connecticut	9001	10529	707	0.07

1-10 of 300 rows Previous 1 2 3 4 5 6 ... 30 Next

Finally, I sort the counties by their death rate.



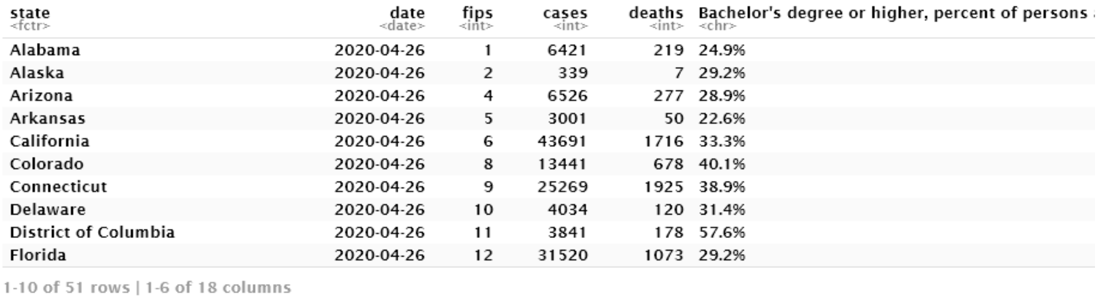
date <fctr>	county <fctr>	state <fctr>	fips <int>	cases <int>	deaths <int>	deathRate <dbl>
2020-04-26	Hennepin	Minnesota	27053	1332	177	0.13
2020-04-26	Beaver	Pennsylvania	42007	366	46	0.13
2020-04-26	Hartford	Connecticut	9003	4989	579	0.12
2020-04-26	Carroll	Maryland	24013	391	47	0.12
2020-04-26	Genesee	Michigan	26049	1467	161	0.11
2020-04-26	Sussex	New Jersey	34037	855	92	0.11
2020-04-26	Henrico	Virginia	51087	792	89	0.11
2020-04-26	Middlesex	Connecticut	9007	588	66	0.11
2020-04-26	Madison	Indiana	18095	394	45	0.11
2020-04-26	Wayne	Michigan	26163	15748	1580	0.10

1-10 of 300 rows Previous 1 2 3 4 5 6 ... 30 Next

US Census Bureau data

On the bureau data website, each table can only contain 6 states' information at most, so I need to download 9 csv files to get the whole information of 50 states and D.C.. Also, there are many indicators in each table and only part of them is needed. As a result, I have to do some data scraping and cleaning.

At first, I pick the indicators I need form the tables. Then I merge the 9 csv files into 1 data frame. Finally, I filter the rows that the date is 2020-04-26. The picture below is the sample of the result.



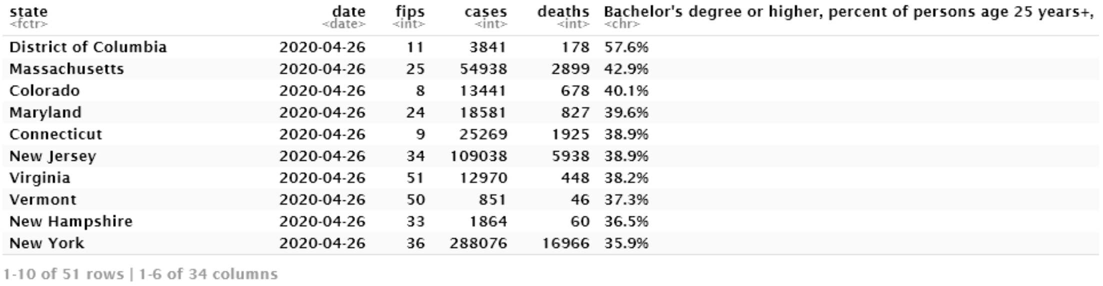
state	date	fips	cases	deaths	Bachelor's degree or higher, percent of persons
<fctr>	<date>	<int>	<int>	<int>	<chr>
Alabama	2020-04-26	1	6421	219	24.9%
Alaska	2020-04-26	2	339	7	29.2%
Arizona	2020-04-26	4	6526	277	28.9%
Arkansas	2020-04-26	5	3001	50	22.6%
California	2020-04-26	6	43691	1716	33.3%
Colorado	2020-04-26	8	13441	678	40.1%
Connecticut	2020-04-26	9	25269	1925	38.9%
Delaware	2020-04-26	10	4034	120	31.4%
District of Columbia	2020-04-26	11	3841	178	57.6%
Florida	2020-04-26	12	31520	1073	29.2%

1-10 of 51 rows | 1-6 of 18 columns

The indicators chosen from the original csv files are 'Population estimates, July 1, 2019, (V2019) ', 'Persons under 5 years, percent' , 'Persons under 18 years, percent' , 'Persons 65 years and over, percent' , 'Female persons, percent' , 'White alone, percent' , 'Black or African American alone, percent' , 'Hispanic or Latino, percent' , 'High school graduate or higher, percent of persons age 25 years+, 2014-2018' , 'Bachelor's degree or higher, percent of persons age 25 years+, 2014-2018' , 'Persons without health insurance, under age 65 years, percent' , 'Persons in poverty, percent' , 'Population per square mile, 2010'.

We can sort the states by the columns, For example, I am interested in the 'Bachelor's degree or higher, percent of persons age 25 years+, 2014-

2018' of every state. Here is the result.

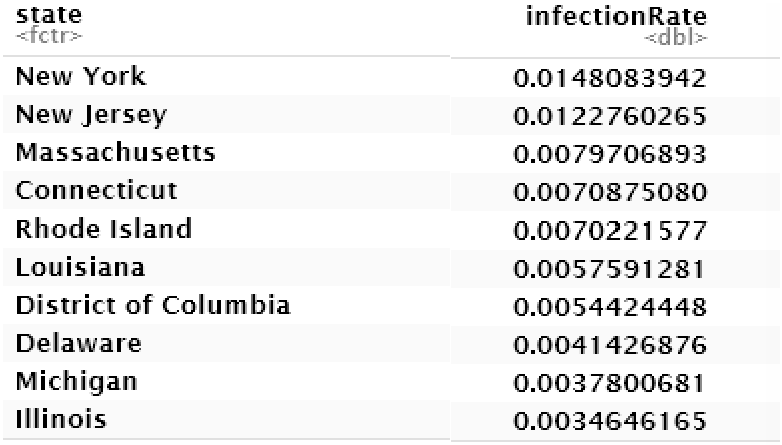


state <fctr>	date <date>	fips <int>	cases <int>	deaths <int>	Bachelor's degree or higher, percent of persons age 25 years+, <chr>
District of Columbia	2020-04-26	11	3841	178	57.6%
Massachusetts	2020-04-26	25	54938	2899	42.9%
Colorado	2020-04-26	8	13441	678	40.1%
Maryland	2020-04-26	24	18581	827	39.6%
Connecticut	2020-04-26	9	25269	1925	38.9%
New Jersey	2020-04-26	34	109038	5938	38.9%
Virginia	2020-04-26	51	12970	448	38.2%
Vermont	2020-04-26	50	851	46	37.3%
New Hampshire	2020-04-26	33	1864	60	36.5%
New York	2020-04-26	36	288076	16966	35.9%

1-10 of 51 rows | 1-6 of 34 columns

In this picture, the data of column 'date', 'fips', 'cases' and 'deaths' all come from the first dataset. I have merged it with the US Census Bureau dataset.

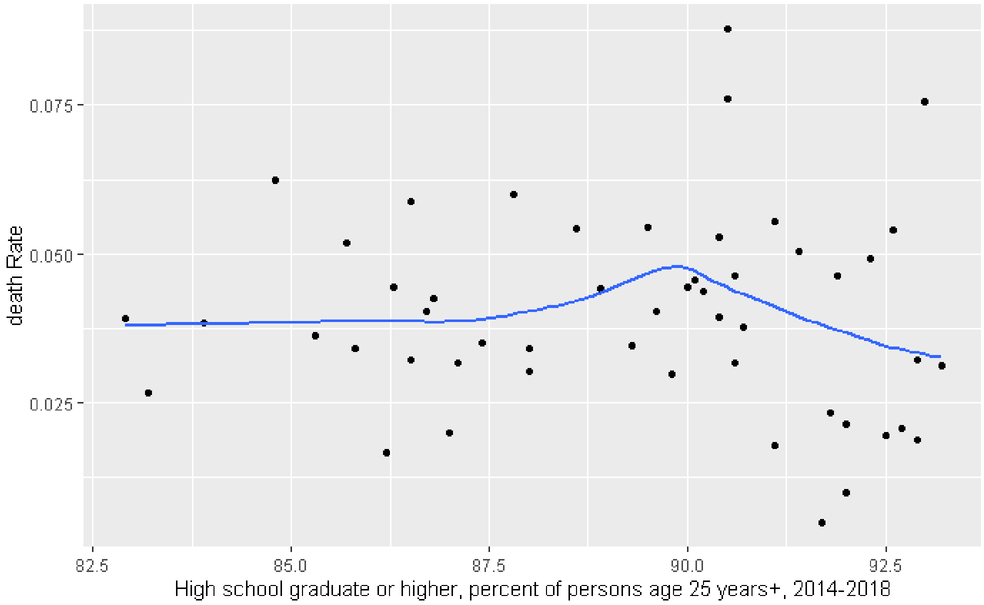
I add a new columns into the data frame which is 'infenctionRate'. It equals to state's cases/ state's population. The population is already included in the table. The new created indicator can give me a better understanding of the circumstances of every state. The table below shows the top10 states with the highest infection rate.

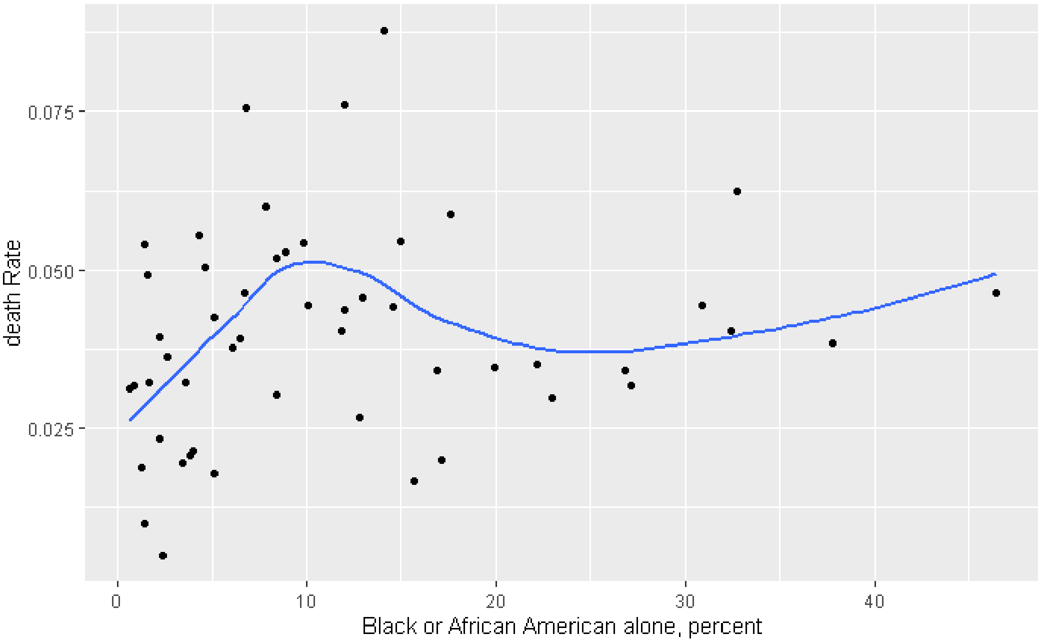


state <fctr>	infectionRate <dbl>
New York	0.0148083942
New Jersey	0.0122760265
Massachusetts	0.0079706893
Connecticut	0.0070875080
Rhode Island	0.0070221577
Louisiana	0.0057591281
District of Columbia	0.0054424448
Delaware	0.0041426876
Michigan	0.0037800681
Illinois	0.0034646165

Then I make several plots and try to find the relationship between the death rate and the chosen indicators on the states level. I choose some of the plots to show in the report.

The picture below shows there is no obvious relationship between the death rate and the High school graduate or higher percent.

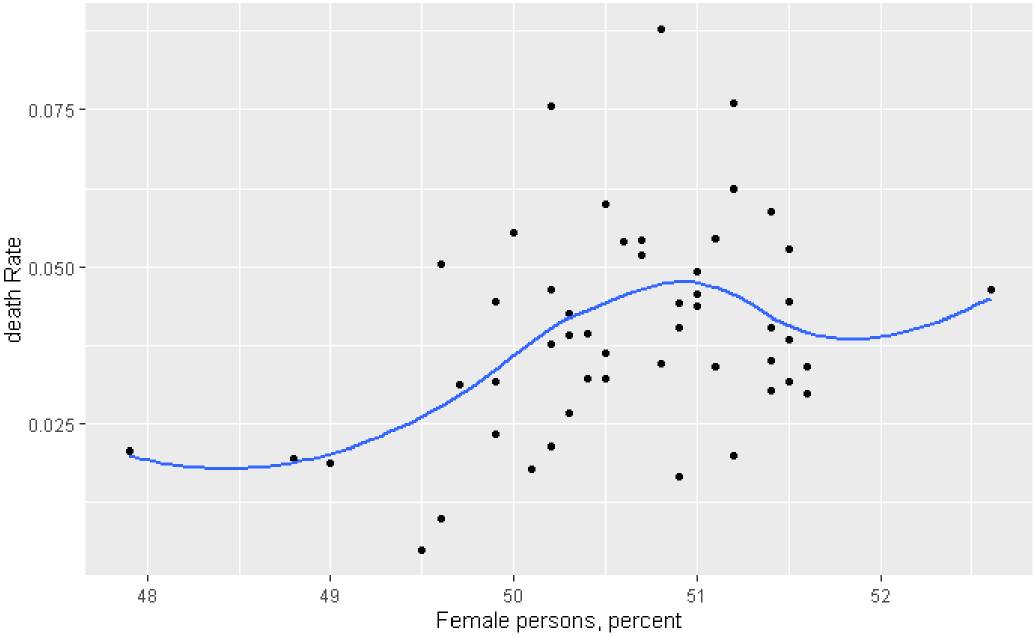


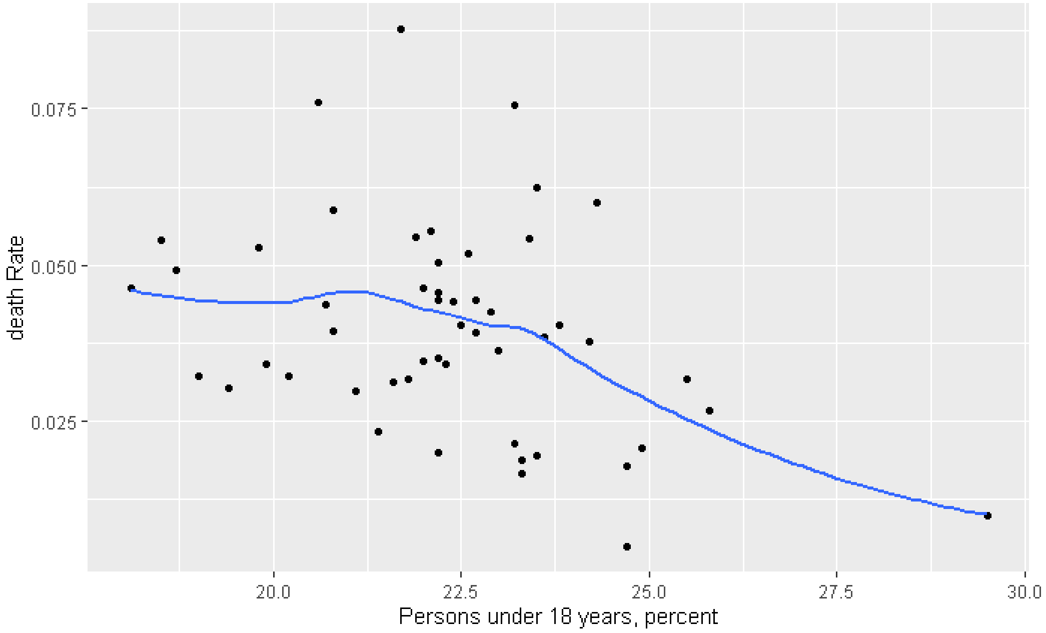


The picture above is the plot of death rate and African American percent.

The death rate grows when the percent is smaller than 10%.

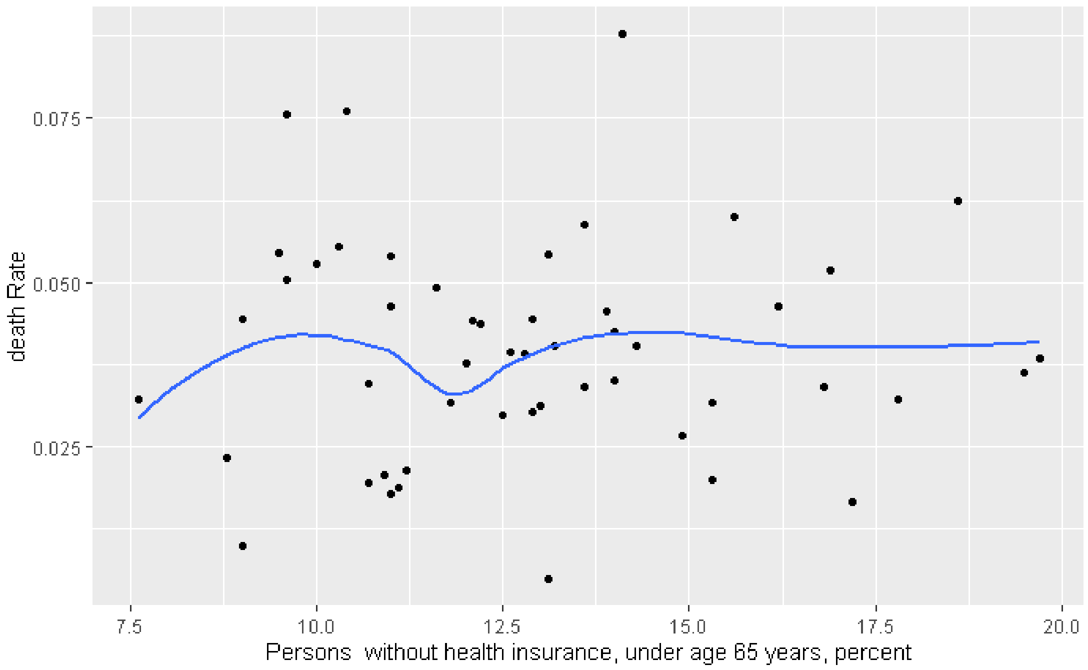
The photo below shows the relationship between death rate and Female person, percent. The trend of death rate is increasing while the female person percent is growing.





The plot above is about the death rate and persons under 18 years percent. It is obvious that the death rate is decreasing while the percent is growing.

The picture below is about the death rate and the person without health insurance. It is surprise to me because it is hard to find relationship in the plot.



There are much more plots in the code. However, it is hard to find relationship among them. Perhaps the reason is that there is not enough data on the states level. The result could be better if we explore on the county level.