Final Project of MSDS 597

Hui Wang

5/4/2020

(With Presentaion) Link in Github: https://github.com/Hui-Dora-Wang/Data-Wrangling-FinalProject

COVID-19, Political Standing, Education Level, Population Density, and GDP

During this hard time, COVID-19 is a terrifying yet unavoidable keyword in our life. In this project, when applying the skills learned from the course, I try to find if there is any direct relationship between the spread of COVID-19 and the political standing, education level, population density, and GDP of the states in the US.

To make the full dataset, I scraped five tables from different websites and did the data cleaning for four of them. Let's use the Political Standing table as a simple example; it provides the data of the 2016 presidential election. As you can see, to process the raw data scraped from the website, we need to split the strings and convert the data type to numerical. And since the names of the states are abbreviations, we need to assign the names of states by hand. It is easier after arranging them by state and we just need to pay attention to some different orders between the states, such as the District of Columbia is after Delaware, but D.C. was in front of Del in the original table.

```
##    State    Clinton    Trump     Other Rpt.
## 1   Ala. 34%Clinton 62%Trump 2%Johnson 100% NA
## 2 Alaska 37%Clinton 51%Trump 6%Johnson 100% NA
## 3  Ariz. 45%Clinton 48%Trump 4%Johnson 100% NA
## 4   Ark. 34%Clinton 61%Trump 3%Johnson 100% NA
## 5 Calif. 62%Clinton 32%Trump 3%Johnson 100% NA
## 6  Colo. 48%Clinton 43%Trump 5%Johnson 100% NA
```

```
##        state Clinton Trump Other_poli
## 1    alabama    0.34  0.62       0.02
## 2     alaska    0.37  0.51       0.06
## 3    arizona    0.45  0.48       0.04
## 4   arkansas    0.34  0.61       0.03
## 5 california    0.62  0.32       0.03
## 6   colorado    0.48  0.43       0.05
```

Full Table

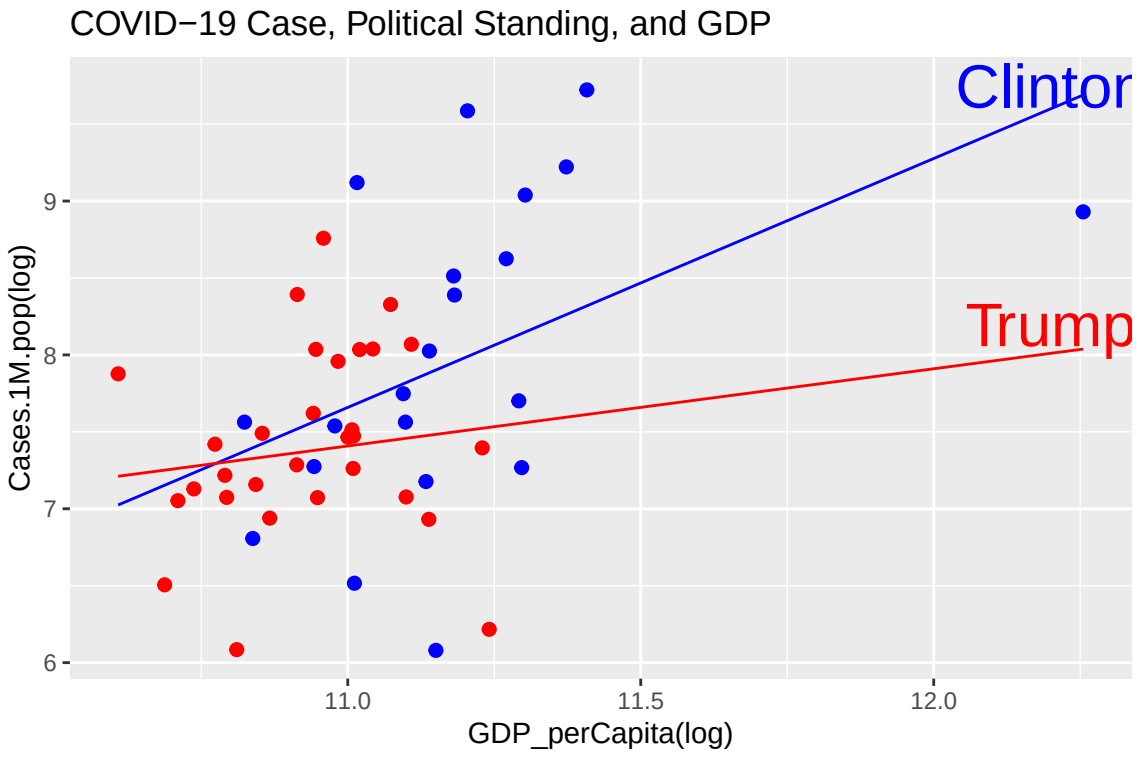
Well, after cleaning up all the tables and left join them by state, we can get the full table. It contains 50 states and DC in row and 18 variables. The names of columns are pretty straightforward.

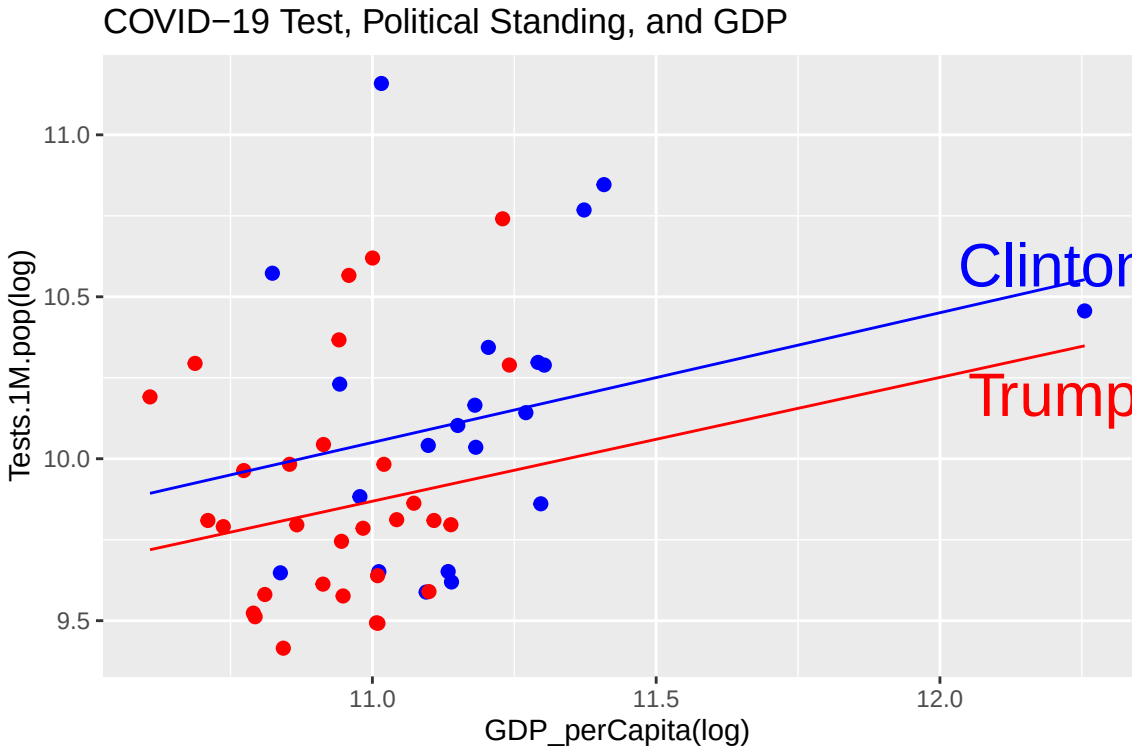
```
##        state TotalCases TotalDeaths Cases.1M.pop Deaths.1M.pop
## 1    alabama       8112         298         1668            61
```

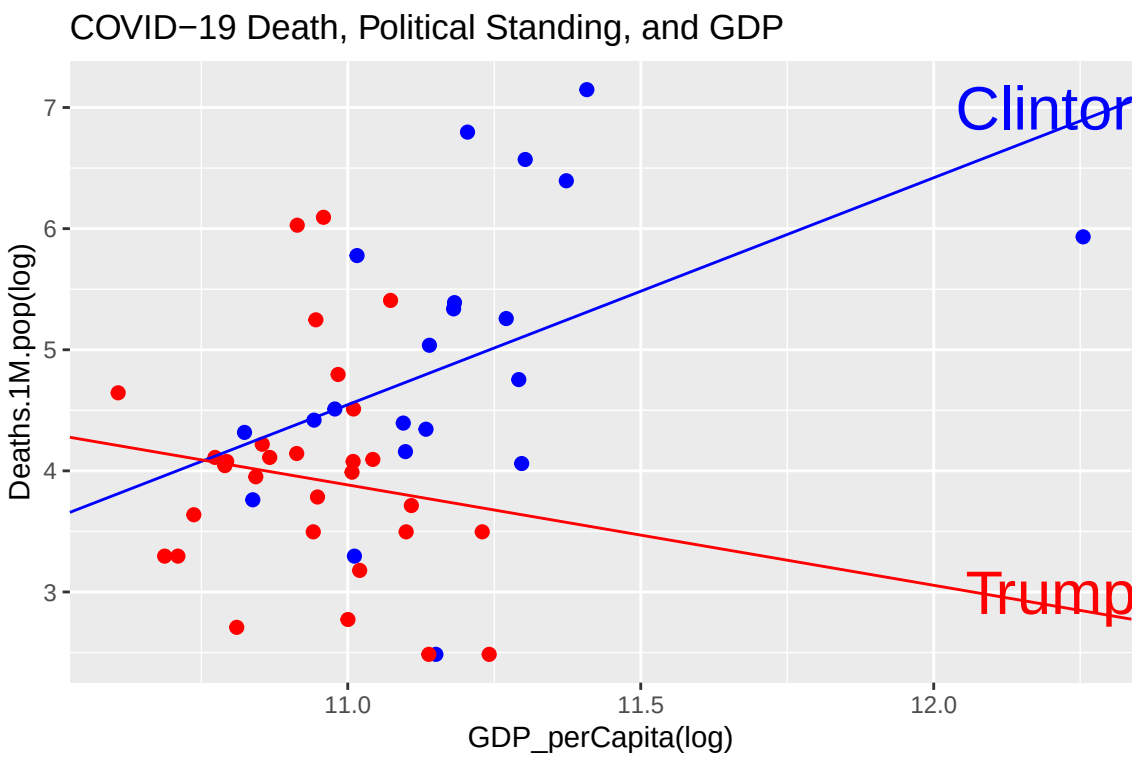
```
## 2     alaska        370           9          501            12
## 3    arizona       8919         362         1284            52
## 4   arkansas       3458          80         1156            27
## 5 california      56089        2283         1433            58
## 6   colorado      16907         851         3057           154
##   Tests.1M.pop Clinton Trump Other_poli Education Population   Land
## 1        21235    0.34  0.62       0.02     30.23    4833722  50645
## 2        29414    0.37  0.51       0.06     49.19     735132 570641
## 3        12272    0.45  0.48       0.04     42.61    6626624 113594
## 4        18204    0.34  0.61       0.03     27.90    2959373  52035
## 5        19165    0.62  0.32       0.03     50.03   38332521 155779
## 6        15054    0.48  0.43       0.05     67.97    5268367 103642
##   Density Q4_2019 GDP_portion GDP_perCapita  Area
## 1    95.4  234054         1.1         47735 South
## 2     1.3   55759         0.3         76220  West
## 3    58.3  372522         1.7         51179  West
## 4    56.9  135225         0.6         44808 South
## 5   246.1 3183251        14.6         80563  West
## 6    50.8  396367         1.8         68828  West
```

COVID-19, Political Standing, and GDP

Well, let's see the plots showing the relationship of the spread of COVID-19, the Political Standing, and GDP of the states in America. As mentioned before, the Political Standing stands for whether more popular votes went for Clinton or Trump in 2016. The GDP is the GDP per capita of the fourth season in 2019. In the plots, we have logged GDP per capita on the x-axis and logged COVID-19 case number, death number and test number on the y-axis. The two lines represent either the blue states, that voted for Clinton, and the red states, that voted for Trump. In the case plot, we can see that both lines go upward because both GDP and COVID-19 spread are positively correlated with urbanization: urban areas with high population density. The difference between the two slopes may be due to the different test rates. From the test plot, although the difference between slopes is subtle, the blue states have conducted more tests than the red states at any given GDP level. The death plot is interesting. We would expect places with higher GDP to be more prepared for COVID-19 and thus have a lower death rate. Indeed, the line for Trump states is what we expect. The one for Clinton does not appear so probably because there are too many confirmed cases in those big cities and the medical system breaks down.

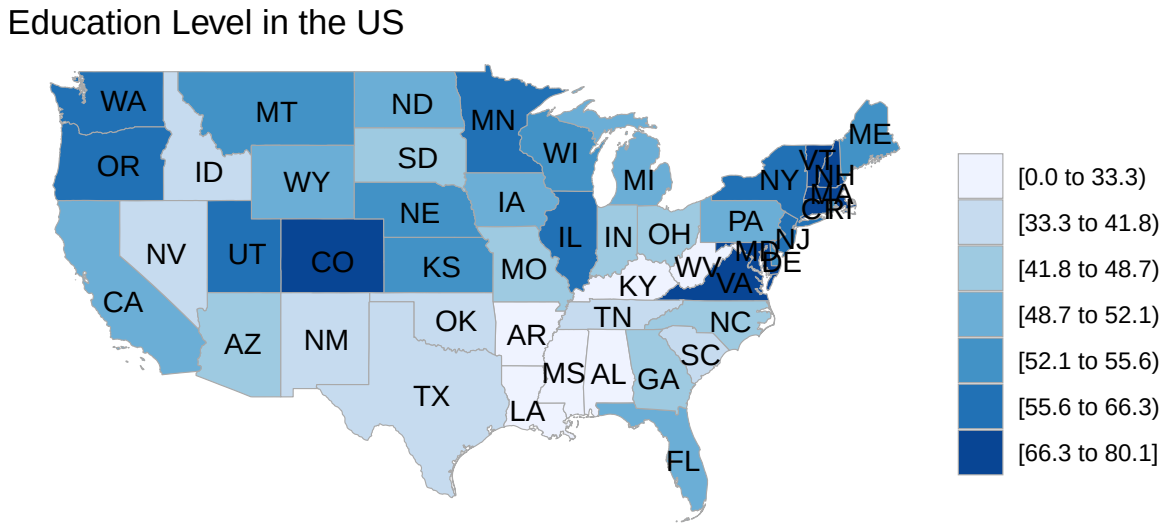

COVID−19 Case, Political Standing, and GDP
Clintor
Trump
Cases.1M.pop(log)
9
8
7
6
11.0
11.5
12.0
GDP_perCapita(log)

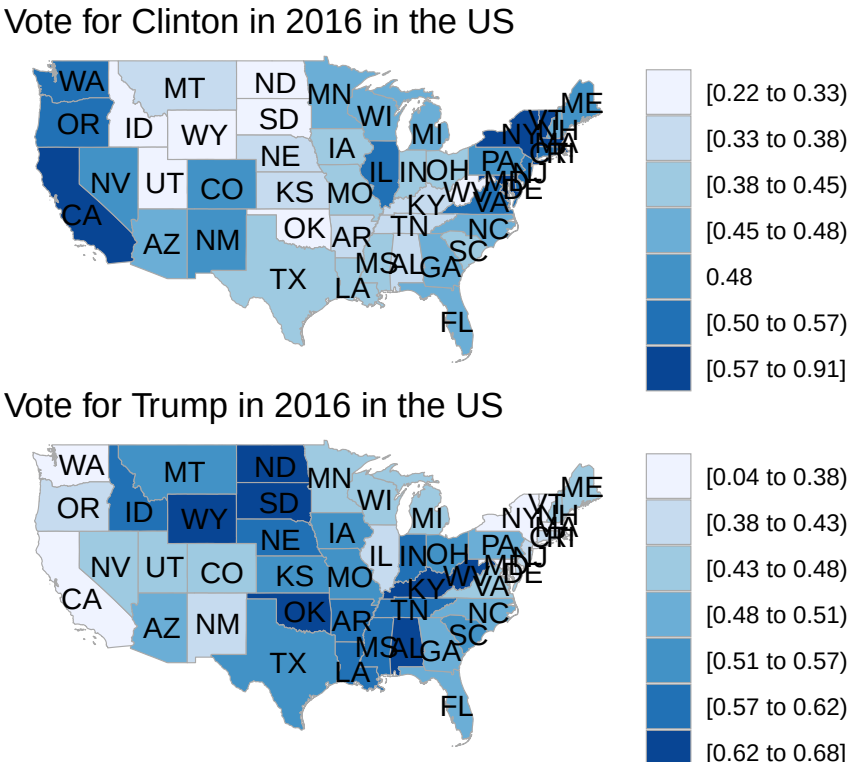

COVID−19 Test, Political Standing, and GDP
Clintor
Trump
Tests.1M.pop(log)
11.0
10.5
10.0
9.5
11.0
11.5
12.0
GDP_perCapita(log)



Education & Political Standing

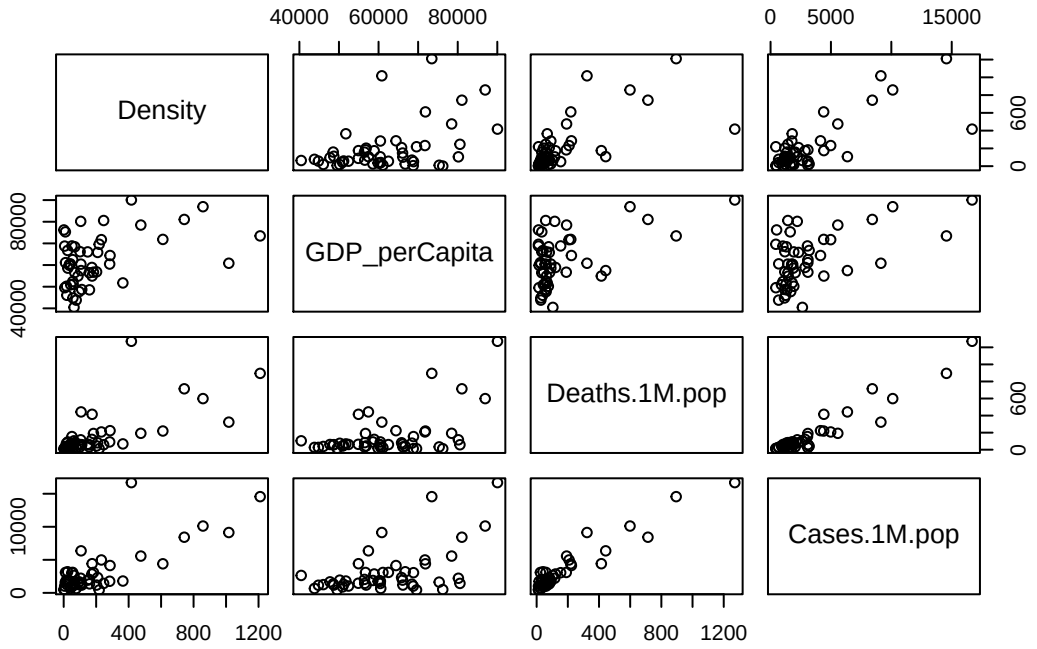
I also want to see the relationship between the education level and political standing. And it seems the states with higher education level are more likely to have voted for Clinton instead of Trump.

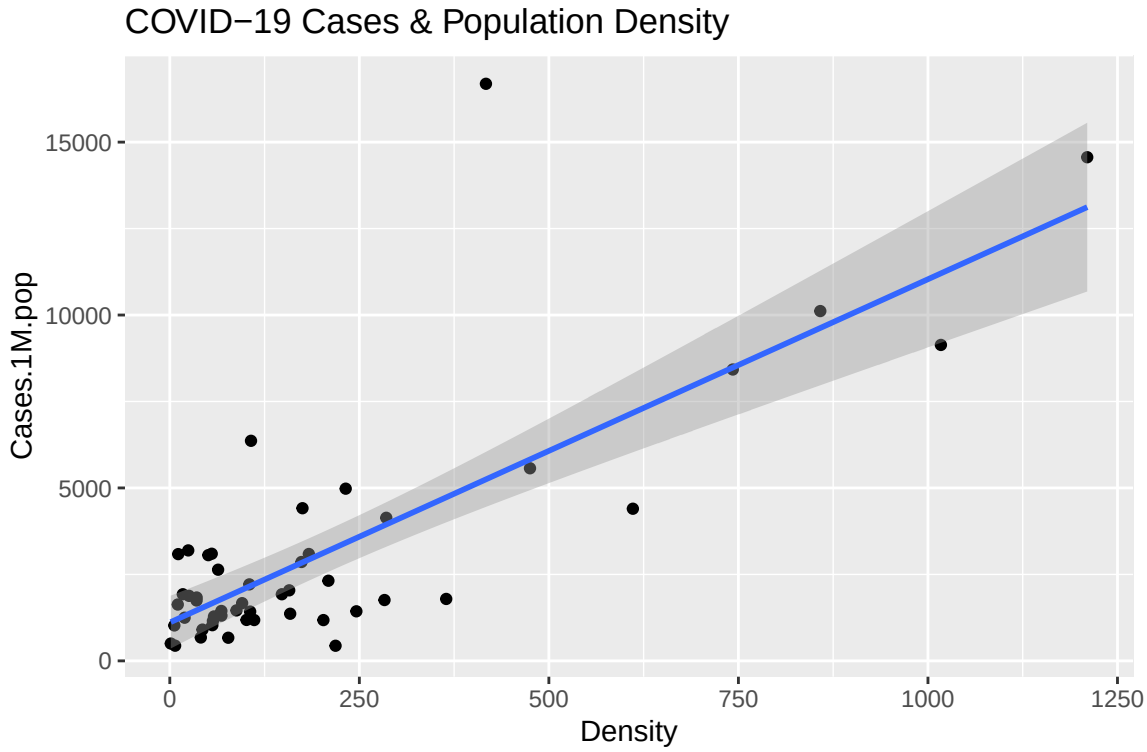


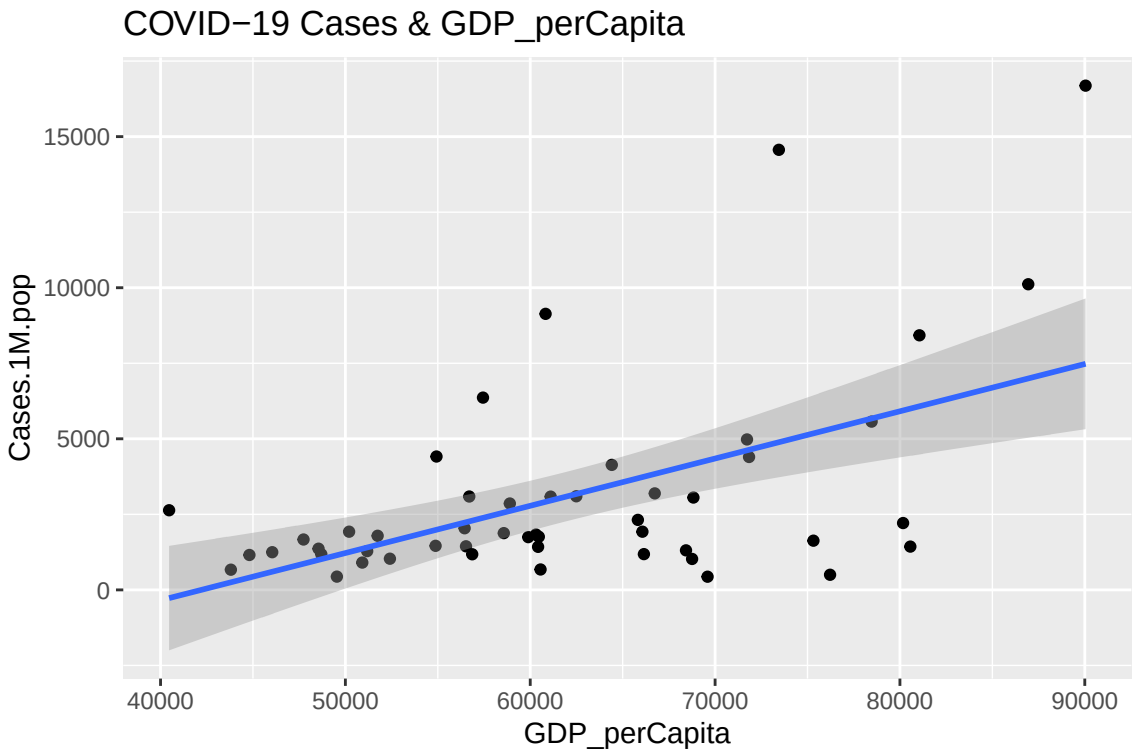


COVID-19, Population Density, and GDP

Last, about the COVID-19, Population Density, and GDP, I delete the outlier, District of Columbia. Seeing the pairs plot, I think it would be interesting to compare the relationship between cases on one hand and population density or the GDP per capita on the other. And from the linear plots, we can see the model of population density in the middle plot fits better than the one of GDP per capita on the right plot. So, the GDP per capita is not that much related to the cases as population density is.


Density
GDP_perCapita
Deaths.1M.pop
Cases.1M.pop
40000
60000
80000
0
5000
15000
600
10000
400
800
1200


COVID−19 Cases & Population Density
Cases.1M.pop
Density
15000
10000
5000
0
250
500
750
1000
1250


COVID−19 Cases & GDP_perCapita
Cases.1M.pop
15000
10000
5000
0
40000
50000
60000
70000
80000
90000
GDP_perCapita