

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.

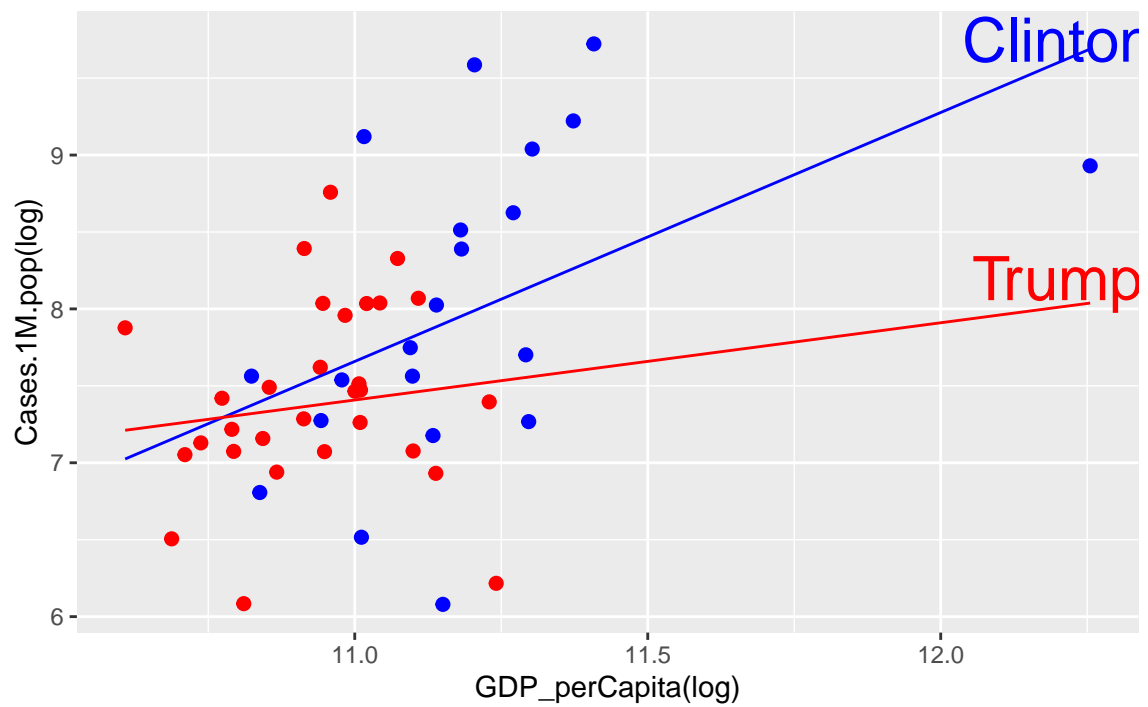
```
##      state TotalCases TotalDeaths Cases.1M.pop Deaths.1M.pop
## 1  alabama      8112         298      1668          61
```

## 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		

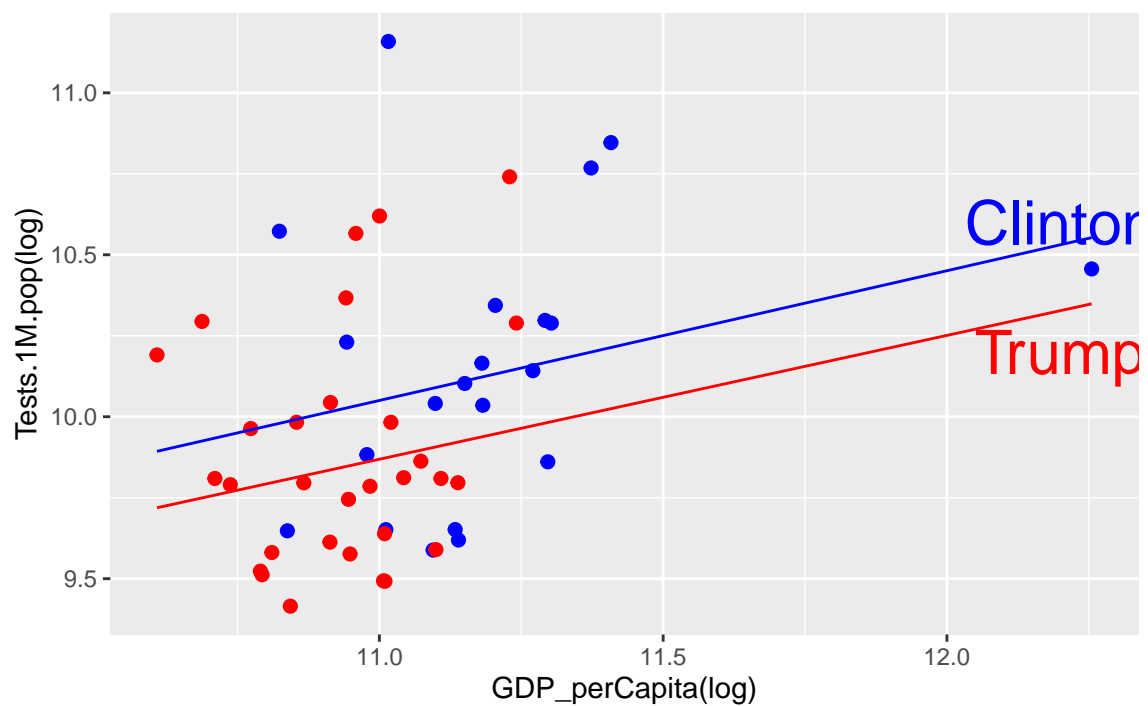
COVID-19, Political Standing, and GDP

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



COVID-19 Test, Political Standing, and GDP



The scatter plot displays the relationship between GDP_perCapita(log) on the x-axis and Deaths.1M.pop(log) on the y-axis. The data points are categorized into two groups: Clinton (blue) and Trump (red). The Clinton regression line (blue) shows a positive correlation, while the Trump regression line (red) shows a negative correlation.

Category	GDP_perCapita(log)	Deaths.1M.pop(log)
Clinton	11.0	5.8
Clinton	11.2	6.8
Clinton	11.4	7.2
Clinton	11.6	6.4
Clinton	11.8	5.9
Clinton	11.1	4.5
Clinton	11.3	5.3
Clinton	11.5	4.8
Clinton	11.7	4.1
Clinton	11.9	3.5
Clinton	12.1	3.2
Clinton	12.3	2.8
Clinton	12.5	2.5
Clinton	12.7	2.2
Clinton	12.9	1.9
Clinton	13.1	1.6
Clinton	13.3	1.3
Clinton	13.5	1.0
Clinton	13.7	0.7
Clinton	13.9	0.4
Clinton	14.1	0.1
Clinton	14.3	-0.2
Clinton	14.5	-0.5
Clinton	14.7	-0.8
Clinton	14.9	-1.1
Clinton	15.1	-1.4
Clinton	15.3	-1.7
Clinton	15.5	-2.0
Clinton	15.7	-2.3
Clinton	15.9	-2.6
Clinton	16.1	-2.9
Clinton	16.3	-3.2
Clinton	16.5	-3.5
Clinton	16.7	-3.8
Clinton	16.9	-4.1
Clinton	17.1	-4.4
Clinton	17.3	-4.7
Clinton	17.5	-5.0
Clinton	17.7	-5.3
Clinton	17.9	-5.6
Clinton	18.1	-5.9
Clinton	18.3	-6.2
Clinton	18.5	-6.5
Clinton	18.7	-6.8
Clinton	18.9	-7.1
Clinton	19.1	-7.4
Clinton	19.3	-7.7
Clinton	19.5	-8.0
Clinton	19.7	-8.3
Clinton	19.9	-8.6
Clinton	20.1	-8.9
Clinton	20.3	-9.2
Clinton	20.5	-9.5
Clinton	20.7	-9.8
Clinton	20.9	-10.1
Clinton	21.1	-10.4
Clinton	21.3	-10.7
Clinton	21.5	-11.0
Clinton	21.7	-11.3
Clinton	21.9	-11.6
Clinton	22.1	-11.9
Clinton	22.3	-12.2
Clinton	22.5	-12.5
Clinton	22.7	-12.8
Clinton	22.9	-13.1
Clinton	23.1	-13.4
Clinton	23.3	-13.7
Clinton	23.5	-14.0
Clinton	23.7	-14.3
Clinton	23.9	-14.6
Clinton	24.1	-14.9
Clinton	24.3	-15.2
Clinton	24.5	-15.5
Clinton	24.7	-15.8
Clinton	24.9	-16.1
Clinton	25.1	-16.4
Clinton	25.3	-16.7
Clinton	25.5	-17.0
Clinton	25.7	-17.3
Clinton	25.9	-17.6
Clinton	26.1	-17.9
Clinton	26.3	-18.2
Clinton	26.5	-18.5
Clinton	26.7	-18.8
Clinton	26.9	-19.1
Clinton	27.1	-19.4
Clinton	27.3	-19.7
Clinton	27.5	-20.0
Clinton	27.7	-20.3
Clinton	27.9	-20.6
Clinton	28.1	-20.9
Clinton	28.3	-21.2
Clinton	28.5	-21.5
Clinton	28.7	-21.8
Clinton	28.9	-22.1
Clinton	29.1	-22.4
Clinton	29.3	-22.7
Clinton	29.5	-23.0
Clinton	29.7	-23.3
Clinton	29.9	-23.6
Clinton	30.1	-23.9
Clinton	30.3	-24.2
Clinton	30.5	-24.5
Clinton	30.7	-24.8
Clinton	30.9	-25.1
Clinton	31.1	-25.4
Clinton	31.3	-25.7
Clinton	31.5	-26.0
Clinton	31.7	-26.3
Clinton	31.9	-26.6
Clinton	32.1	-26.9
Clinton	32.3	-27.2
Clinton	32.5	-27.5
Clinton	32.7	-27.8
Clinton	32.9	-28.1
Clinton	33.1	-28.4
Clinton	33.3	-

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.

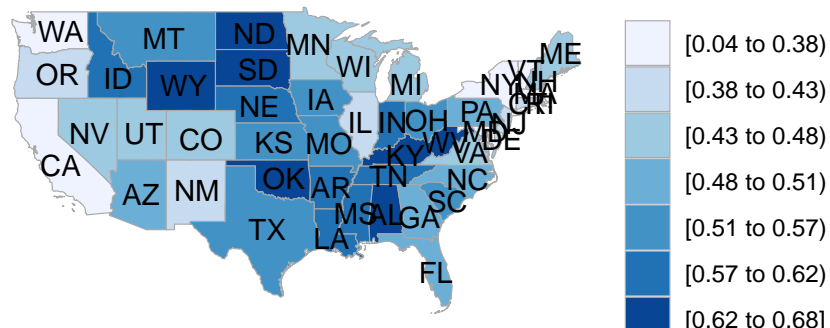
Legend:

- [0.0 to 33.3)
- [33.3 to 41.8)
- [41.8 to 48.7)
- [48.7 to 52.1)
- [52.1 to 55.6)
- [55.6 to 66.3)
- [66.3 to 80.1]

Vote for Clinton in 2016 in the US

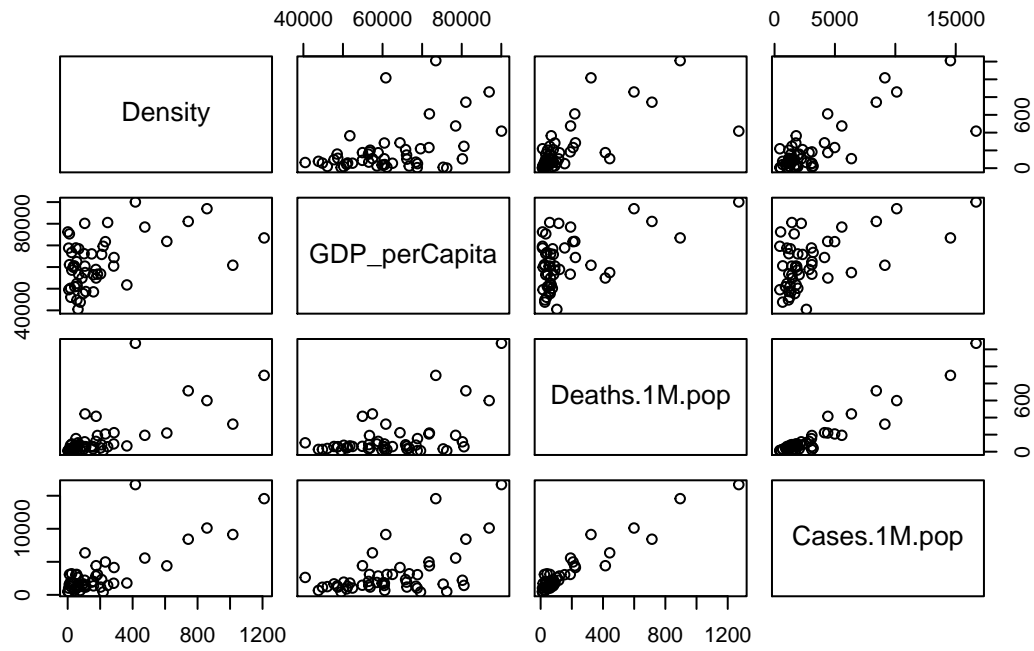


Vote for Trump in 2016 in the US



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.



COVID-19 Cases & Population Density

